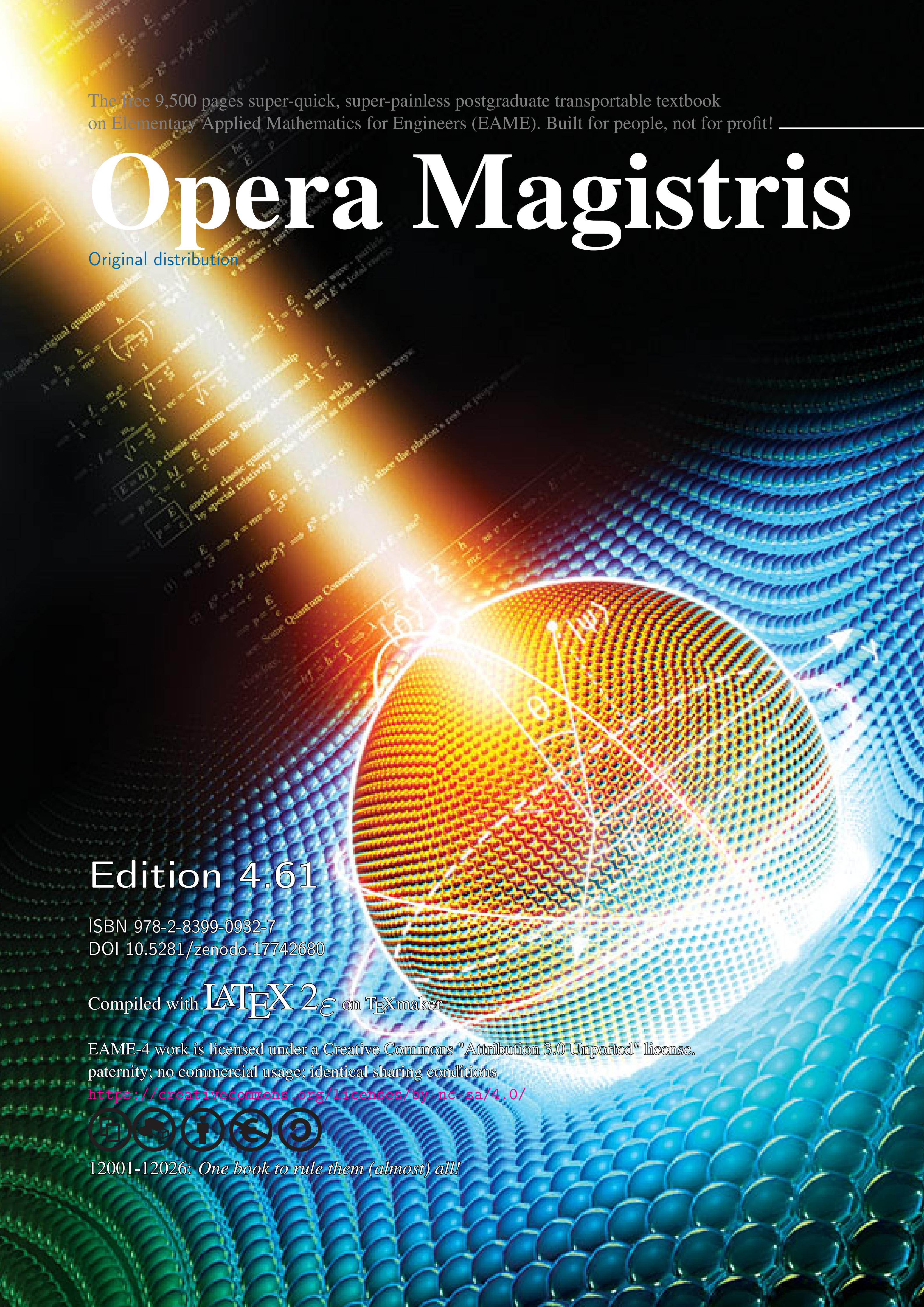


The free 9,500 pages super-quick, super-painless postgraduate transportable textbook on Elementary Applied Mathematics for Engineers (EAME). Built for people, not for profit!

Opera Magistris

Original distribution

Edition 4.61

ISBN 978-2-8399-0932-7
DOI 10.5281/zenodo.17742680

Compiled with LaTeX 2ε on TeXmaker

EAME-4 work is licensed under a Creative Commons "Attribution 3.0 Unported" license.
paternity; no commercial usage; identical sharing conditions
https://creativecommons.org/licenses/by-nc-sa/4.0/

12001-12026: *One book to rule them (almost) all!*

Science reveals to us the origin of the Universe and its mechanisms. It shows us the condition of humans (and other species of the animal world to which we belong), why they have so many cognitive and biological (genetic) defects and limitations, how their condition can be improved (or vice versa!) and it also shows us the fate of those who lose themselves in ignorance and in the beliefs of those who base themselves on the facts and the real multifactorial complexity of our environment. Science is not dogmatic, it self-corrects and continually questions itself, it admits to being imperfect by definition, seeks to find fault and establishes that there is very likely no "absolute truths" and that complex systems are rarely dichotomously categorizable. Its precepts are constraining, its stories are verifiable by multiple independent sources and its decisions are always revocable because only the idiots are full of certainties and the wise full of doubts. Study reproducible, peer reviewed Science published in non-predatory journals to gain wisdom, follow its evidence to accept the inevitability of probabilities and causal chains, and apply what it teaches to live a more serene life, be more tolerant of differences and unforeseen events and other probabilistic fatalities. Science contains the light that will guide you in the darkness and comfort you during the manipulation attempts of multiple religions and pseudo-sciences and other charlatans who play on rhetoric monologues and with your feelings to sell you goods, services or ask you for donations and who take advantage of your intellectual weaknesses. Science is the universal traveller's guide, the pilgrim's staff, the sailor's compass, the rationalist's weapon and the sage's charter. Through it, life and its unpredictable, sometimes painful events are relativized and accepted and obscurantism returns to obscurity.

Science is the main subject of this textbook, our good is its main purpose, and the glory of Nature its aim. It should fill our memory and guide our steps through the only, fragile and short life we have. Study it carefully, often, in a spirit of understanding and endless curiosity about complexity. It is a mine of wealth, a subject of exchanges and a river of intellectual delights. It is given to you during your short lifespan and you will take it with you to do as little harm as possible around you until your last moment, hoping that you will transmit its values to your descendants and entourage for the best possible living together. It involves the highest responsibility, will reward the greatest efforts and will condemn to ignorance, intolerance and stupidity all those who will make fun of it.

Offered to: ______________________________

By: ______________________________

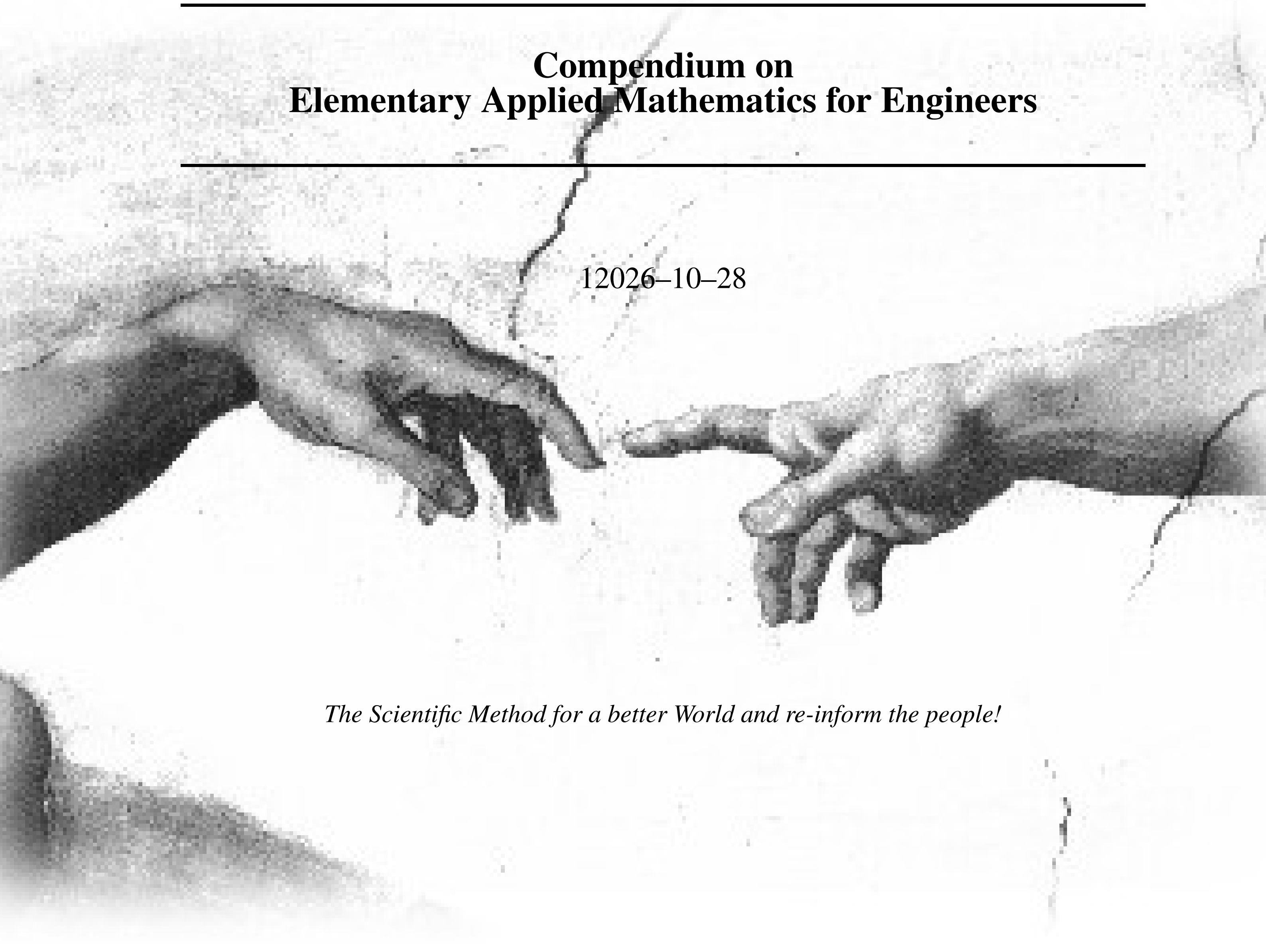
Reason: ______________________________

Date: ______________________________

ORIGINAL DISTRIBUTION

OPERA MAGISTRIS

4TH EDITION



Compendium on Elementary Applied Mathematics for Engineers

12026–10–28

The Scientific Method for a better World and re-inform the people!

1

Revision History

The reader will notice that the years are given on the basis of the Holocene calendar for obvious scientific reasons.

Version	Date	Redactor(s)	Description
4.6	12026-06-01	VI	New topics and some small corrections (see changelog at page 9347)
4.5	12025-10-17	VI	New topics and some small corrections (see changelog at page 9347)
4.4	12025-06-06	VI	New topics and some small corrections (see changelog at page 9347)
4.3	12025-05-01	VI	New topics and some small corrections (see changelog at page 9347)
4.2	12024-07-10	VI	New topics and changes in headers and footers (see changelog at page 9347)
4.1	12024-04-21	VI	Some new topics added and a few corrections of typing/translation errors and new illustrations made with TikZ (see changelog at page 9347)
4.0	12023-12-23	VI	Again a few new small topics added and a few corrections of typing/translation errors (see changelog at page 9347)
3.11	12018-06-10	VI, DK	New topics added and a few corrections of typing/translation errors (see changelog at page 9347)
3.10	12018-01-20	VI, DK	New topics added and a few corrections of typing/translation errors (see changelog at page 9347)
3.9	12017-11-13	VI, DK	Some minor updates for the free version of the book (see changelog at page 9347)
3.8	12017-08-01	VI, DK	Some major and critical updates and corrections (see changelog page 9347)
3.7	12017-06-01	VI, DK	French 3^{rd} Edition translation into English finished!
3.6	12017-05-01	VI, DK	French 3^{rd} Edition translation into English. Translation progress $\sim 99\%$

3.5	12017-04-01	VI, DK	French 3rd Edition translation into English. Translation progress $\sim 96\%$
3.4	12017-02-11	VI, DK	French 3rd Edition translation into English. Translation progress $\sim 96\%$
3.3	12016-01-27	VI	French 3rd Edition translation into English. Translation progress $\sim 96\%$
3.0	12014-01-22	VI, RR, LG, HL	French 3rd Edition published on the Internet as PDF ($4,888$ pages).
2.0	12005-03-10	VI, RR	French 2nd Edition published on the Internet as PDF ($2,001$ pages).
1.0	12002-05-01	VI	French 1st Edition published on the Internet as HTML only.

quote 1.0:iv.#1

> ” *If there’s a book that you want to read, but it hasn’t been written yet, then you must write it.* ”
>
> Toni Morrison

! Warning

It is forbidden to sell this textbook! It’s completely free and must remain free!

Contents

1 Revision History iii

2 Warnings 2

3 Acknowledgements 30

4 Introduction 34

5 Arithmetic 125

6 Algebra 1344

7 Analysis 2168

8 Geometry 2499

9 Mechanics 3063

10 Electromagnetism 3964

11 Atomistic 4612

12 Cosmology 5212

13 Chemistry 6122

14 Theoretical Computing 6267

15 Social Sciences 7292

16 Engineering 8234

17 Epilogue 9062

18 Biographies 9064

19 Chronology 9166

20 Humor 9211

21 Links 9315

22 Quotes 9330

23 Changelog 9347

24 Nomenclature 9372

List of Figures 9379

List of Tables 9430

List of Algorithms 9440

Bibliography 9443

Index 9478

25 Donate 9556

Table of Contents

1 Revision History **iii**

2 Warnings **2**

2.1 Impressum 4
- 2.1.1 Use of content 4
- 2.1.2 How to use this textbook 5
  - 2.1.2.1 Style guide 6
  - 2.1.2.2 Credits 7
  - 2.1.2.3 Videos and Animations 7
  - 2.1.2.4 Ancillaries 8
- 2.1.3 Data Protection 10
- 2.1.4 Use of data 10
- 2.1.5 Data transmission 10
- 2.1.6 Agreement 10
- 2.1.7 Errata 10

2.2 License 12
- 2.2.1 Preamble 12
- 2.2.2 Applicability and Definitions 13
- 2.2.3 Verbatim Copying 14
- 2.2.4 Copying in Quantity 14
- 2.2.5 Modifications 15
- 2.2.6 Combining Documents 16
- 2.2.7 Collections of Documents 17
- 2.2.8 Aggregation with independent Works 17
- 2.2.9 Generating this document 17
- 2.2.10 Translation 18
- 2.2.11 Permissions and Copyrights 18
- 2.2.12 Termination 18
- 2.2.13 Future revisions of this License 18

2.3 Safety Guides 19

2.4 Roadmap (To Do) 21

3 Acknowledgements **30**

4 Introduction **34**

4.1 Forewords 35

4.1.1 Motivation and goals . . . 41
4.1.2 Who this textbook is intended for . . . 47
4.1.3 How is this textbook structured . . . 49
4.2 Methods . . . 50
4.2.1 Descartes' Method . . . 56
4.2.1.1 Blind studies . . . 58
4.2.2 Research Integrity and Engineering/Scientific Ethics . . . 66
4.2.2.1 Singapore Statement on Research Integrity . . . 66
4.2.2.2 European Code of Conduct for Research Integrity . . . 67
4.2.2.3 Archimedean Oath . . . 68
4.2.3 Scientific Publication Rules (SPR) . . . 70
4.2.4 Scientific Mainstream Media communication . . . 79
4.2.4.1 Social Networks . . . 81
4.2.4.2 Expert Opinions . . . 85
4.3 Vocabulary . . . 88
4.3.1 On Sciences . . . 89
4.3.2 Terminology . . . 99
4.4 Science and Faith . . . 103
4.4.1 Baloney detection kit . . . 109
4.5 Scientific communication backfire . . . 115
4.6 Is Science dogmatic? . . . 120

5 Arithmetic 125

5.1 Proof Theory . . . 131
5.1.1 Foundations Crisis . . . 134
5.1.2 Paradoxes . . . 138
5.1.2.1 Hypothetical-Deductive Reasoning . . . 141
5.1.3 Propositional Calculus . . . 142
5.1.3.1 Propositions (premises) . . . 144
5.1.3.2 Connectors . . . 146
5.1.3.3 Decision procedures . . . 157
5.1.3.3.1 Non-axiomatic procedural decisions . . . 158
5.1.3.3.2 Axiomatic procedural decisions . . . 159
5.1.3.4 Quantifiers . . . 163
5.1.4 Predicate Calculus . . . 167
5.1.4.1 Grammar . . . 168
5.1.4.2 Languages . . . 169
5.1.4.2.1 Symbols . . . 169
5.1.4.2.2 Terms . . . 170
5.1.4.2.3 Formulas . . . 172
5.1.5 Proofs . . . 176
5.1.5.1 Rules of Proofs . . . 177
5.1.6 Laws of thought . . . 187
5.1.6.1 Some famous logical fallacies . . . 199
5.1.6.2 Is logic a science? . . . 204

5.2 Numbers . . . 209
5.2.1 Digital Bases . . . 212
5.2.2 Type of Numbers . . . 215
5.2.2.1 Natural Integer Numbers . . . 215
5.2.2.1.1 Peano axioms . . . 218
5.2.2.1.2 Odd, Even and Perfect Numbers . . . 219
5.2.2.1.3 Prime Numbers . . . 220
5.2.2.2 Relative Integer Numbers . . . 222
5.2.2.3 Rational Numbers . . . 223
5.2.2.4 Irrational Numbers . . . 228
5.2.2.5 Real Numbers . . . 230
5.2.2.5.1 Percentages . . . 234
5.2.2.6 Transfinite Numbers . . . 239
5.2.2.7 Complex Numbers . . . 243
5.2.2.7.1 Geometric Interpretation of Complex Numbers . . . 249
5.2.2.7.1.1 Fresnel Vectors (phasors) . . . 255
5.2.2.7.2 Transformation in the plane . . . 257
5.2.2.8 Quaternion Numbers . . . 261
5.2.2.8.1 Matrix Interpretation of Quaternions . . . 268
5.2.2.8.2 Rotations with Quaternions . . . 269
5.2.2.9 Algebraic and Transcendental Numbers . . . 276
5.2.2.10 Universe Numbers (normal numbers) . . . 279
5.2.2.11 Abstract Numbers (variables) . . . 280
5.2.2.11.1 Domain of a Variable . . . 281
5.3 Arithmetic Operators . . . 285
5.3.1 Binary Relations . . . 286
5.3.1.1 Equalities . . . 287
5.3.1.2 Comparators . . . 288
5.3.2 Fundamental Arithmetic Laws . . . 296
5.3.2.1 Addition . . . 296
5.3.2.2 Subtraction . . . 303
5.3.2.3 Multiplication . . . 308
5.3.2.4 Division . . . 313
5.3.2.5 n-root . . . 320
5.3.3 Arithmetic Polynomials . . . 323
5.3.4 Absolute Value . . . 324
5.3.5 Calculation Rules (operators priorities) . . . 328
5.4 Number Theory . . . 333
5.4.1 Principle of good order . . . 333
5.4.2 Induction Principle . . . 334
5.4.3 Divisibility . . . 336
5.4.3.1 Euclidean Division . . . 338
5.4.3.1.1 Greatest common divisor . . . 340
5.4.3.2 Euclidean Algorithm . . . 343
5.4.3.3 Least Common Multiple . . . 347

5.4.3.4 Fundamental Theorem of Arithmetic . . . 352
5.4.3.5 Congruences (modular arithmetic) . . . 356
5.4.3.5.1 Congruence Class . . . 362
5.4.3.5.2 Complete set of residues . . . 365
5.4.3.5.3 Chinese remainder theorem . . . 366
5.4.3.6 Continued fractions . . . 369
5.5 Set Theory . . . 378
5.5.1 Zermelo-Fraenkel Axiomatic . . . 382
5.5.1.1 Cardinals . . . 388
5.5.1.2 Cartesian Product . . . 391
5.5.1.3 Intervals . . . 392
5.5.2 Set Operations . . . 393
5.5.2.1 Inclusion . . . 394
5.5.2.2 Intersection . . . 394
5.5.2.3 Union . . . 396
5.5.2.4 Difference . . . 397
5.5.2.5 Symmetric Difference . . . 398
5.5.2.6 Product . . . 399
5.5.2.7 Complementarity . . . 399
5.5.3 Functions and Applications . . . 403
5.5.3.1 Cantor-Bernstein Theorem . . . 410
5.5.4 Structures . . . 416
5.5.4.1 Magma . . . 418
5.5.4.2 Monoid . . . 419
5.5.4.3 Groups . . . 421
5.5.4.4 Ring . . . 426
5.5.4.4.1 Sub-ring . . . 429
5.5.4.5 Field . . . 429
5.5.4.6 Vector Spaces . . . 431
5.5.4.7 C-algebra A . . . 433
5.5.4.8 Summary . . . 435
5.5.5 Morphisms . . . 439
5.5.5.1 Ideal . . . 444
5.6 Probabilities . . . 450
5.6.1 Event Universe . . . 452
5.6.1.1 Infinite monkey theorem (Borel's law) . . . 455
5.6.1.1.1 Second Borel-Cantelli lemma . . . 456
5.6.1.2 Do probabilities refute complex structures? . . . 459
5.6.2 Kolmogorov's Axioms . . . 476
5.6.3 Conditional Probabilities . . . 485
5.6.3.1 Conditional Expectation and Variance . . . 499
5.6.3.2 Bayesian Networks . . . 504
5.6.4 Geometric Probabilities . . . 516
5.6.5 Martingales . . . 518
5.6.5.1 Gambling . . . 520

5.6.6 Combinatorial Analysis . . . 522
5.6.6.1 Simple Arrangements with Repetitions . . . 523
5.6.6.2 Simple Permutations without Repetitions . . . 525
5.6.6.3 Simple Permutations with Repetitions . . . 526
5.6.6.4 Simple Arrangements without Repetitions . . . 527
5.6.6.5 Simple Combinations without Repetitions . . . 528
5.6.6.6 Simple Combinations with Repetitions . . . 530
5.6.7 Markov Chains . . . 537
5.7 Statistics . . . 542
5.7.1 Samples . . . 545
5.7.2 Averages . . . 546
5.7.2.1 Laplace Smoothing . . . 570
5.7.2.2 Means and Averages properties . . . 571
5.7.2.2.1 Jensen Inequality . . . 576
5.7.3 Type of variables . . . 580
5.7.3.1 Discrete Variables and Moments . . . 582
5.7.3.1.1 Mean and Variance of Discrete Random Variables . . . 585
5.7.3.1.1.1 Semivariance . . . 598
5.7.3.1.2 Discrete Covariance . . . 599
5.7.3.1.2.1 Quasi-correlation matrix . . . 607
5.7.3.1.2.2 Anscombe's famous quartet . . . 609
5.7.3.1.3 Mean and Variance of the Average . . . 611
5.7.3.1.4 Coefficient of Correlation . . . 613
5.7.3.1.4.1 Geometric interpretation . . . 622
5.7.3.1.4.2 Point biserial correlation coefficient . . . 622
5.7.3.1.4.3 Partial correlation and semipartial correlation . . . 625
5.7.3.1.4.4 Intraclass coefficient of correlation . . . 628
5.7.3.2 Continuous Variables and Moments . . . 633
5.7.3.2.1 Skewness and Kurtosis . . . 636
5.7.4 Fundamental postulate of statistics . . . 643
5.7.5 Diversity Index . . . 644
5.7.6 Distribution Functions (probabilities laws) . . . 647
5.7.6.1 Discrete Uniform Distribution . . . 649
5.7.6.2 Bernoulli Distribution . . . 652
5.7.6.3 Geometric Distribution . . . 654
5.7.6.4 Binomial Distribution . . . 658
5.7.6.5 Negative Binomial Distribution . . . 667
5.7.6.6 Hypergeometric Distribution . . . 672
5.7.6.7 Multinomial Distribution . . . 679
5.7.6.8 Poisson Distribution . . . 686
5.7.6.9 Normal & Gauss-Laplace Distribution . . . 692
5.7.6.9.1 Sum of two random Normal variables . . . 709
5.7.6.9.2 Product of two random Normal variables . . . 711
5.7.6.9.3 Bivariate Normal Distribution . . . 713
5.7.6.9.4 Multivariate Normal distribution . . . 720

5.7.6.9.4.1 Marginal and conditional distributions of multivariate Normal distributions . . . 723
5.7.6.9.5 Normal Reduced Centered Distribution . . . 725
5.7.6.9.6 Henry's Line . . . 726
5.7.6.9.7 Q-Q plot . . . 731
5.7.6.10 Log-Normal Distribution . . . 734
5.7.6.11 Continuous Uniform Distribution . . . 738
5.7.6.12 Triangular Distribution . . . 743
5.7.6.13 Pareto Distribution . . . 746
5.7.6.14 Exponential Distribution . . . 751
5.7.6.15 Cauchy Distribution . . . 755
5.7.6.16 Beta Distribution . . . 760
5.7.6.16.1 Regularized Incomplete beta function integral . . . 764
5.7.6.17 Gamma Distribution . . . 767
5.7.6.17.1 Generalized Gamma Distribution . . . 773
5.7.6.18 Chi-Square (Pearson) Distribution . . . 773
5.7.6.18.1 Noncentral chi-square distribution . . . 780
5.7.6.19 Student Distribution . . . 783
5.7.6.19.1 Multivariate Student T-distribution . . . 789
5.7.6.20 Fisher Distribution . . . 790
5.7.6.21 General Folded Normal Distribution . . . 795
5.7.6.21.1 Half-Normal distribution . . . 796
5.7.6.22 Laplace Distribution . . . 800
5.7.6.23 Benford Distribution . . . 803
5.7.6.24 ex-Gaussian distribution . . . 806
5.7.7 Likelihood Estimators . . . 818
5.7.7.1 Normal Distribution MLE . . . 819
5.7.7.2 Poisson Distribution MLE . . . 825
5.7.7.3 Binomial Distribution MLE . . . 826
5.7.7.4 Geometric Distribution MLE . . . 826
5.7.7.5 Weibull Distribution MLE . . . 827
5.7.7.6 Gamma Distribution MLE . . . 830
5.7.7.7 Pareto Distribution MLE . . . 831
5.7.7.8 Censored data MLE . . . 832
5.7.7.9 Fisher Information Matrix . . . 835
5.7.8 Finite Population Correction Factor . . . 843
5.7.9 Confidence Intervals (inference) . . . 848
5.7.9.1 Frequentist Inference . . . 848
5.7.9.1.1 C.I. on the Mean with known theoretical Variance . . . 852
5.7.9.1.2 C.I. on the Variance with known theoretical Mean . . . 857
5.7.9.1.3 C.I. on the Variance with known empirical Mean . . . 862
5.7.9.1.4 C.I. on the Mean with known empirical Variance . . . 864
5.7.9.1.5 C.I. on the Geometric Mean (for known or unknown Variance) . . . 868
5.7.9.1.6 Binomial exact Test . . . 869
5.7.9.1.7 C.I. for a Proportion . . . 872
5.7.9.1.7.1 Variance stabilizing transformation . . . 878

5.7.9.1.7.2 Test of equality (or difference) of two Proportions . . . 880
5.7.9.1.8 Sign Test . . . 883
5.7.9.1.9 Mood's Median Test . . . 885
5.7.9.1.10 Poisson Test (1 sample) . . . 886
5.7.9.1.11 Poisson Test (2 samples) . . . 889
5.7.9.1.12 Confidence/Tolerance/Prediction/Credibility Intervals . . . 892
5.7.9.2 Bayesian Inference . . . 894
5.7.9.2.1 Beta prior conjugate to Binomial likelihood (beta-binomial distribution) . . . 898
5.7.9.2.1.1 Uninformative prior . . . 901
5.7.9.2.2 Normal prior conjugate to Normal likelihood . . . 904
5.7.9.2.3 Binomial prior conjugate to Binomial likelihood . . . 907
5.7.9.2.4 Credibility Intervals . . . 910
5.7.9.2.5 Bayes Factor (BF) . . . 914
5.7.10 Weak Law of Large Numbers (Khinchin's law) . . . 920
5.7.11 Characteristic Function . . . 926
5.7.11.1 Moment generating function and cumulants . . . 931
5.7.12 Central Limit Theorem . . . 938
5.7.13 Univariate Hypothesis and Adequation Tests (NHST) . . . 944
5.7.13.1 Direction of hypothesis tests and p-values . . . 948
5.7.13.2 Fisher's method for multiple p-values . . . 963
5.7.13.2.1 Simpson's Paradox (sophism) . . . 967
5.7.13.3 Power of a Test . . . 972
5.7.13.3.1 Evidence ("proof") of a negative . . . 974
5.7.13.3.1.1 Two-One-Sided T-test procedure (TOST) . . . 977
5.7.13.3.2 Power of the one sample Z-test and T-test . . . 980
5.7.13.3.3 Power of the two sample Z-test and T-test . . . 983
5.7.13.3.4 Power of the one and two samples P-test . . . 986
5.7.13.3.5 Power calculation based on simulations . . . 990
5.7.13.4 A/B Tests . . . 991
5.7.13.4.1 Naive proportion approach . . . 991
5.7.13.4.2 Fieller's test (ratio of two means) . . . 992
5.7.13.4.3 Bayesian A/B testing for binary outcomes . . . 996
5.7.13.4.4 Bayesian A/B testing for count datas . . . 999
5.7.13.5 Analysis Of VAriance (ANOVA) . . . 1001
5.7.13.5.1 ANOVA with one fixed factor . . . 1004
5.7.13.5.1.1 Welch Test . . . 1020
5.7.13.5.1.2 Contrasts . . . 1024
5.7.13.5.2 ANOVA with two fixed factors without repetitions . . . 1027
5.7.13.5.3 ANOVA with two fixed factors with repetitions . . . 1045
5.7.13.5.4 Multifactor ANOVA with Repeated Measures . . . 1051
5.7.13.5.4.1 Interlude on CRD, RBD, RCBD, IBD, GRBD... . . . 1052
5.7.13.5.4.2 Interlude on Type I, Type II, Type III ANOVAs . . . 1059
5.7.13.5.5 Hierarchical (Nested) ANOVA (HANOVA) . . . 1065
5.7.13.5.6 ANOVA with blocks . . . 1073
5.7.13.5.7 Latin Square ANOVA without repetitions . . . 1074

5.7.13.5.8 Graeco-Latin Square ANOVA 1082
5.7.13.5.9 Multivariate ANOVA (MANOVA) 1091
5.7.13.5.9.1 Hotelling's T-squared Test 1096
5.7.13.6 Equivalence Tests 1101
5.7.13.7 Cochran C-test 1104
5.7.13.8 Family-wise error rate (adjusted p-values) 1106
5.7.13.8.1 Holm-Bonferroni Correction 1109
5.7.13.8.2 Holm–Šidák Correction 1111
5.7.13.9 Adequation Tests (Goodness of Fit tests) 1111
5.7.13.9.1 Pearson's Chi-squared GoF Test 1111
5.7.13.9.2 G-test GoF 1117
5.7.13.9.3 Kolmogorov-Smirnov GoF Test 1119
5.7.13.9.4 Ryan-Joiner GoF Test 1126
5.7.13.9.5 Anderson-Darling GoF Test 1130
5.7.13.9.6 Cramér–von Mises GoF Test 1140
5.7.13.10 Kernel density estimation 1143
5.7.13.11 Likelihood-ratio Tests 1150
5.7.14 Robustness and Nonparametric Statistics 1153
5.7.14.1 Rank Statistics 1156
5.7.14.1.1 L-Statistics 1156
5.7.14.1.2 Ranks Distribution law 1157
5.7.14.1.3 Distribution free confidence intervals for percentiles 1158
5.7.14.1.4 Wilcoxon Rank Sum Test 1160
5.7.14.1.5 Mann-Whitney Rank Sum Test 1170
5.7.14.1.5.1 Treatment of equal ranks in Ranks based tests 1176
5.7.14.1.6 One sample Wilcoxon rank sum signed Test 1178
5.7.14.1.7 Wilcoxon rank sum signed test for two paired samples 1182
5.7.14.1.8 Kruskal-Wallis Test 1185
5.7.14.1.9 Friedman Test 1189
5.7.14.1.10 Spearman's rank correlation coefficient 1193
5.7.14.1.11 Kendall's tau correlation Coefficient 1196
5.7.14.2 Range Statistics 1203
5.7.14.2.1 Tukey's Range Test 1207
5.7.14.3 Extreme Value Theory 1211
5.7.15 Multivariate Statistics 1212
5.7.15.1 Principal Component Analysis 1213
5.7.15.1.1 SVD and PCA 1230
5.7.15.2 Chi-square Correspondence Factor Analysis (C2CFA) 1233
5.7.15.3 Chi-squared Test of Independence 1238
5.7.15.3.1 Equivalence of Z-test and χ^2 test for proportions 1243
5.7.15.3.2 Cramér's V 1247
5.7.15.4 Pearson's phi coefficient 1250
5.7.15.5 Exact Fisher's Test 1254
5.7.15.6 Cohen's kappa agreement 1260
5.7.15.7 McNemar's test 1264

5.7.15.7.1 Derivation for Cochran-Armitage's trend test for the 2 × 3 genotype table 1268
5.7.16 Survival Statistics 1271
5.7.16.1 Kaplan-Meier Survival Rate 1273
5.7.16.2 Cochran–Mantel–Haenszel Tests 1283
5.7.16.3 Cox (Proportional Hazard) Model 1291
5.7.16.3.1 Stratified Cox Model 1305
5.7.17 Propagation of Errors (experimental uncertainty analysis) 1307
5.7.17.1 Absolute and Relative Uncertainties (direct calculation of bias) 1308
5.7.17.2 Statistical Errors 1309
5.7.17.3 Repeatability 1312
5.7.17.4 Error propagation (linearised approximation) 1313
5.7.17.5 Error propagation (statistical approach) 1316
5.7.17.6 Significant Numbers 1318
5.7.18 Spatial statistics 1320
5.7.18.1 2D Poisson's spatial distance model 1320
5.7.19 Surveys 1324
5.7.19.1 Survey designs 1327
5.7.19.1.1 Simple random sampling design without replacement and equal probabilities (RSEPR) 1334
5.7.19.2 Survey Sampling Error and Non-Sampling Error 1335
5.7.20 A World without statistics 1339
5.7.20.1 Data fallacies 1342

6 Algebra **1344**

6.1 Calculus 1350
6.1.1 Equations and Inequations 1351
6.1.1.1 Equations 1352
6.1.1.2 Inequations 1356
6.1.2 Remarkable Identities 1361
6.1.3 Polynomials 1368
6.1.3.1 General relations between roots and coefficients 1373
6.1.3.2 Multiplication of Polynomials 1374
6.1.3.3 Euclidean Division of Polynomials 1374
6.1.3.4 Factorization Theorem of Polynomials 1377
6.1.3.5 Diophantine equation 1378
6.1.3.6 First order univariate Polynomial and Equations 1379
6.1.3.6.1 Solving by circular reference 1380
6.1.3.7 Second order univariate Polynomial and Equations 1382
6.1.3.7.1 Irrational Equations 1387
6.1.3.7.2 Golden Ratio 1389
6.1.3.8 Third order univariate Polynomial and Equations 1390
6.1.3.9 Fourth order univariate Polynomial and Equations 1394
6.1.3.10 Trigonometric Polynomials 1397
6.1.3.11 Cyclotomic Polynomials 1398
6.1.3.12 Legendre Polynomials 1400
6.1.3.13 Laguerre Polynomials 1404

6.1.3.13.1 Associated Laguerre Polynomials . . . 1411
6.1.3.13.2 Laguerre functions . . . 1414
6.2 Set Algebra . . . 1417
6.2.1 Groups Algebra and Geometry . . . 1417
6.2.1.1 Cyclic Groups . . . 1418
6.2.1.2 Transformations Groups . . . 1421
6.2.1.3 Group of Symmetries . . . 1430
6.2.1.3.1 Orbits and Stabilizers . . . 1434
6.2.1.4 Permutations Groups . . . 1435
6.2.2 Galois Theory . . . 1445
6.2.2.1 Elementary symmetric and Invariant Polynomials . . . 1446
6.2.2.2 General Vieta's formulas . . . 1449
6.3 Differential and Integral Calculus . . . 1450
6.3.1 Differential Calculus . . . 1450
6.3.1.1 Differentials . . . 1464
6.3.1.1.1 Implicit functions . . . 1473
6.3.1.2 Usual Derivatives . . . 1475
6.3.1.3 Implicit Differentiation . . . 1492
6.3.1.4 Smoothness . . . 1496
6.3.2 Integral Calculus . . . 1497
6.3.2.1 Definite Integral . . . 1497
6.3.2.2 Indefinite Integral . . . 1508
6.3.2.3 Double Integral . . . 1518
6.3.2.3.1 Fubini's theorem . . . 1521
6.3.2.4 Integration by Substitution . . . 1522
6.3.2.4.1 Jacobian . . . 1524
6.3.2.5 Integration by Parts . . . 1532
6.3.2.6 Usual Primitives . . . 1534
6.3.2.7 Leibniz integral rule . . . 1560
6.3.2.8 Integral representation of first kind Bessel's function . . . 1564
6.3.2.9 Dirac Function . . . 1569
6.3.2.9.1 Heaviside function . . . 1572
6.3.2.10 Gamma Euler Function . . . 1574
6.3.2.10.1 Incomplete regularized Gamma function . . . 1578
6.3.2.10.2 Euler-Mascheroni Constant . . . 1581
6.3.2.11 Curvilinear Integrals . . . 1583
6.3.2.11.1 Curvilinear Integral of a scalar field . . . 1583
6.3.2.11.2 Curvilinear Integral of a vector field . . . 1584
6.3.2.12 Surface Integrals . . . 1588
6.3.2.13 Integrals involving parametric equations . . . 1593
6.3.2.14 Improper Integrals . . . 1595
6.3.2.15 Elliptic Integrals . . . 1599
6.3.2.15.1 Incomplete Elliptic Integrals . . . 1600
6.3.2.15.2 Complete Elliptic Integrals . . . 1601
6.3.3 Differential Equations . . . 1603

6.3.3.1 First order differential equations . . . 1605
6.3.3.2 Linear differential equations . . . 1605
6.3.3.3 Resolution methods of linear differential equations . . . 1607
6.3.3.3.1 Method of characteristic polynomial . . . 1608
6.3.3.3.1.1 Resolution of the H.E. of the first order L.D.E. with constant coefficients . . . 1609
6.3.3.3.1.2 Resolution of the H.E. of the first order L.D.E. with non-constant coefficients . . . 1609
6.3.3.3.1.3 Resolution of the H.E. of the second order L.D.E. with constant coefficients . . . 1610
6.3.3.3.2 Integrating Factor Method (Euler's Method) . . . 1615
6.3.3.3.3 Method of separation of variables . . . 1617
6.3.3.3.4 Method of constant variation . . . 1618
6.3.3.4 Classification of differential equations . . . 1621
6.3.3.5 Systems of Differential Equations . . . 1627
6.3.3.5.1 Matrix exponentiation method . . . 1627
6.3.3.5.2 Phase space . . . 1633
6.3.3.5.2.1 Stability theory . . . 1636
6.3.3.6 Nonlinear dynamical systems . . . 1644
6.3.3.7 Regular Methods of Perturbations . . . 1645
6.3.3.7.1 Perturbation theory for algebraic equations . . . 1645
6.3.3.7.2 Perturbation theory of differential equations . . . 1647
6.4 Sequences and Series . . . 1651
6.4.1 Sequences . . . 1651
6.4.1.1 Arithmetic Sequences . . . 1653
6.4.1.2 Harmonic Sequences . . . 1655
6.4.1.3 Geometric Sequences . . . 1656
6.4.1.4 Cauchy Sequence . . . 1657
6.4.1.5 Fibonacci Sequence . . . 1661
6.4.1.6 Logic Sequences/Psychologist Sequences . . . 1663
6.4.2 Series . . . 1664
6.4.2.1 Gauss Series . . . 1666
6.4.2.1.1 Bernoulli's Numbers and Polynomials . . . 1669
6.4.2.1.2 Euler-Maclaurin formula . . . 1674
6.4.2.2 Arithmetic Series . . . 1679
6.4.2.3 Geometric Series . . . 1679
6.4.2.3.1 Zeta function and Euler's product formula . . . 1682
6.4.2.4 Telescoping Series . . . 1691
6.4.2.5 Grandi's Series . . . 1691
6.4.2.6 Taylor and Maclaurin Series . . . 1696
6.4.2.6.1 Usual Maclaurin developments . . . 1703
6.4.2.6.2 Taylor series of bivariate functions (multivariate Taylor series) . . . 1714
6.4.2.6.3 Quadratic Form . . . 1717
6.4.2.6.4 Lagrange Remainder . . . 1722
6.4.2.6.5 Taylor Series with Integral Remainder . . . 1724
6.4.2.6.6 Lagrange inversion theorem . . . 1725
6.4.2.7 Fourier Series (trigonometric series) . . . 1727
6.4.2.7.1 Fourier series generative art . . . 1749

6.4.2.7.2 Power of a signal 1756

6.4.2.7.3 Fourier Transform 1758

6.4.2.8 Bessel Series 1768

6.4.2.8.1 Zero order Bessel's Functions 1768

6.4.2.8.2 n order Bessel's Functions 1768

6.4.2.8.3 Bessel's Differential Equations of order n 1774

6.4.3 Convergence Criteria 1775

6.4.3.1 Integral Test 1776

6.4.3.2 D'Alembert Rule 1777

6.4.3.3 Alternating Series Test 1780

6.4.3.4 Fixed Point Theorem 1780

6.4.4 Generating Functions 1783

6.4.4.1 Ordinary Generating Functions (transformation of a sequence into a series) 1783

6.4.4.1.1 Composition of Generating functions 1787

6.4.4.2 Multivariate Generating Functions 1789

6.4.4.3 Functional Generating Functions 1789

6.5 Vector Calculus 1794

6.5.1 Concept of Arrow 1795

6.5.2 Set of Vectors 1797

6.5.2.1 Pseudo-Vectors 1799

6.5.2.2 Normal vector 1801

6.5.2.3 Multiplication by a scalar 1802

6.5.2.3.1 Rule of three 1803

6.5.3 Vector Spaces 1805

6.5.3.1 Linear Combinations 1806

6.5.3.2 Sub-vector spaces 1806

6.5.3.3 Generating families 1807

6.5.3.4 Linear Dependence or Independence 1807

6.5.3.5 Base of a vectorial space 1809

6.5.3.6 Direction Angles 1811

6.5.3.7 Dimensions of a vector space 1812

6.5.3.8 Extension of a free family 1813

6.5.3.9 Rank of a finite family 1813

6.5.3.10 Direct Sums 1814

6.5.3.11 Affine spaces 1815

6.5.4 Euclidean Vector Spaces 1818

6.5.4.1 Scalar Product (Dot Product) 1821

6.5.4.1.1 Cauchy–Schwarz inequality 1826

6.5.4.1.2 Triangle Inequalities 1828

6.5.4.1.3 General Scalar/Dot Product 1829

6.5.4.2 Cross Product 1830

6.5.4.3 Mixed Product (triple product) 1838

6.5.5 Vectorial Functional Space 1840

6.5.6 Hermitian Vector Space 1842

6.5.6.1 Hermitian Inner Product 1843

6.5.6.2 Hermitian Cauchy-Schwarz inequality 1844
6.5.6.3 Types of Vectors Spaces 1845
6.5.7 System of Coordinates 1846
6.5.7.1 Cartesian (rectangular) Coordinate System 1846
6.5.7.2 Spherical Coordinate System 1849
6.5.7.2.1 Orthodromic distance 1854
6.5.7.2.2 Haversine distance 1858
6.5.7.3 Cylindrical Coordinate System 1861
6.5.7.4 Polar Coordinate System 1863
6.5.8 Differential Operators 1866
6.5.8.1 Gradients of Scalar Field 1869
6.5.8.2 Gradients of Vector Field 1878
6.5.8.3 Divergences of a Vector Field 1879
6.5.8.4 Rotationals of a Vector Field (Curl) 1888
6.5.8.5 Laplacians of Scalar Fields (Laplace Operator) 1899
6.5.8.6 Laplacians of Vector Fields 1903
6.5.8.7 Remarkable Identities 1911
6.5.8.8 Summary 1918
6.6 Linear Algebra 1924
6.6.1 Linear Systems 1929
6.6.2 Linear Transformations 1934
6.6.3 Matrices 1936
6.6.3.1 Rank of a matrix 1938
6.6.3.2 Matrix Algebra 1945
6.6.3.3 Type of Matrices 1948
6.6.3.4 Determinant 1961
6.6.3.4.1 Derivative of a Determinant relatively to a parameter 1974
6.6.3.4.2 Derivative of logarithm of a determinant 1976
6.6.3.4.3 Cramer’s rule 1977
6.6.3.4.4 Determinant Cofactor and Matrix Inverse 1980
6.6.3.4.5 Determinant of block-diagonal or block-triangular matrices 1986
6.6.3.4.6 Inverse and Determinant of a partitioned symmetric matrix 1989
6.6.4 Change of Basis (frames) 1992
6.6.5 Eigenvalues and Eigenvectors 1997
6.6.5.1 Rotation Matrices and Eigenvalues 2001
6.6.6 Spectral Theorem 2004
6.6.7 Matrix Decompositions 2011
6.6.7.1 Singular Value Decomposition (SVD) 2011
6.6.7.1.1 Singular Vectors and Values 2013
6.6.7.1.2 Special case of SVD for symmetric matrices 2020
6.6.7.1.3 Orthogonal decomposition (eigenvalue spectral decomposition) 2022
6.6.7.2 LU Decomposition 2024
6.6.7.3 Cholesky’s Decomposition 2027
6.6.7.4 QR Decomposition 2030
6.6.8 Woodbury Matrix identity and Sherman-Morrison relation 2035

6.7 Tensor Calculus . . . 2038
6.7.1 Tensor . . . 2039
6.7.2 Indicial Notation . . . 2042
6.7.2.1 Summation on multiple index . . . 2044
6.7.2.2 Kronecker Symbol . . . 2044
6.7.2.3 Antisymmetric Symbol (Levi-Civita symbol) . . . 2045
6.7.3 Metric and Signature . . . 2051
6.7.4 Gram's Determinant . . . 2054
6.7.5 Contravariant and Covariant Components . . . 2059
6.7.6 Operation in Basis . . . 2062
6.7.6.1 Gram-Schmidt Orthogonalization Method . . . 2062
6.7.6.2 Change of Basis . . . 2066
6.7.6.3 Reciprocal Basis (Dual Basis) . . . 2067
6.7.7 Euclidean Tensors (cartesian tensor) . . . 2071
6.7.7.1 Fundamental Tensor . . . 2072
6.7.7.2 Tensor product (dyadic) of two vectors and matrices . . . 2073
6.7.7.3 Tensor Spaces . . . 2076
6.7.7.4 Linear combination of tensors . . . 2080
6.7.7.5 Contraction of indices . . . 2080
6.7.7.5.1 Raising and lowering indices . . . 2081
6.7.8 Special Tensors . . . 2085
6.7.8.1 Symmetric Tensor . . . 2085
6.7.8.2 Antisymmetric Tensor . . . 2088
6.7.8.3 Fundamental Tensor . . . 2091
6.7.9 Curvilinear Coordinates . . . 2091
6.7.9.1 Natural basis in spherical coordinates (curvilinear basis in spherical coordinates) . . . 2097
6.7.9.2 Natural basis in polar coordinates (curvilinear basis in polar coordinates) . . . 2099
6.7.9.3 Natural basis in cylindrical coordinates (cylindrical basis in polar coordinates) . . . 2099
6.7.10 Christoffel symbols . . . 2103
6.7.11 Ricci Theorem . . . 2112
6.7.12 Riemann-Christoffel symbols . . . 2122
6.7.13 Ricci curvature (Ricci tensor) . . . 2127
6.7.14 Einstein Tensor . . . 2141
6.8 Spinor Calculus . . . 2145
6.8.1 Unit Spinor . . . 2146
6.8.2 Geometric Properties . . . 2152
6.8.2.1 Plane Symmetries . . . 2153
6.8.2.2 Rotations . . . 2155
6.8.2.3 Properties of Pauli Matrices . . . 2161
7 Analysis **2168**
7.1 Functional Analysis . . . 2170
7.1.1 Representations . . . 2171
7.1.1.1 Tabular Representations . . . 2171
7.1.1.2 Graphical Representations . . . 2177

7.1.1.2.1 2D representations . . . 2178
7.1.1.2.2 3D representations . . . 2186
7.1.1.2.3 2D Vector representations . . . 2193
7.1.1.2.4 Properties of visual representations . . . 2196
7.1.1.3 Analytical Representations . . . 2206
7.1.2 Functions . . . 2209
7.1.2.1 Limits and Continuity of functions . . . 2226
7.1.2.1.1 Squeeze theorem . . . 2239
7.1.2.1.2 Usual Limits (and Continuities) . . . 2242
7.1.2.1.3 Limit laws . . . 2268
7.1.2.1.4 Asymptotes . . . 2269
7.1.2.1.5 Discontinuities . . . 2276
7.1.2.2 Concavity/Convexity of a function . . . 2280
7.1.2.3 Euler theorem for homogeneous functions . . . 2282
7.1.3 Logarithms . . . 2284
7.1.4 Convolutions . . . 2293
7.1.4.1 Continuous and Discrete Linear Convolution Product . . . 2293
7.1.4.2 Matrix Convolution . . . 2299
7.1.5 Integral Transforms . . . 2303
7.1.5.1 Fourier Transform . . . 2305
7.1.5.1.1 Continuous Time Fourier Transform . . . 2305
7.1.5.1.1.1 Properties of Fourier Transform . . . 2309
7.1.5.1.1.2 Usual Fourier transforms . . . 2319
7.1.5.1.2 Discrete Time Fourier Transform . . . 2327
7.1.5.1.3 Properties of Discrete Fourier Transform . . . 2329
7.1.5.2 Laplace Transform . . . 2336
7.1.5.2.1 Inverse Laplace Transform . . . 2339
7.1.5.2.2 Region of Convergence . . . 2340
7.1.5.2.3 Zeros and Poles of Laplace Transform . . . 2342
7.1.5.2.4 Properties of region of convergence . . . 2343
7.1.5.2.5 Properties of Laplace Transform . . . 2345
7.1.5.2.6 Usual Laplace transforms . . . 2348
7.1.5.2.7 Solving differential systems . . . 2356
7.1.5.2.8 Unilateral Laplace Transform . . . 2358
7.1.5.3 $\mathcal{Z}$-Transform . . . 2360
7.1.5.3.1 Transfer function of LTI system . . . 2363
7.1.5.3.2 Conformal mapping between s-plane to z-plane . . . 2363
7.1.5.3.3 Region of Convergence . . . 2364
7.1.5.3.4 Zeros and Poles of $\mathcal{Z}$-Transform . . . 2366
7.1.5.3.5 Properties of $\mathcal{Z}$-Transform . . . 2367
7.1.5.3.6 Usual $\mathcal{Z}$-Transforms . . . 2371
7.1.5.3.7 Unilateral $\mathcal{Z}$-Transform . . . 2373
7.1.5.4 Hilbert Transform . . . 2376
7.1.6 Functional dot product (inner product) . . . 2379
Orthogonality of trigonometric functions . . . 2380

Orthogonality of complex exponential functions 2382
Orthogonality of Bessel functions 2382
Orthogonality of Hermite polynomial 2385
Orthogonality of Laguerre polynomial 2386
Orthogonality of Legendre polynomial 2387
7.1.6.1 Cauchy-Schwarz inequality for integrals 2388
7.2 Complex Analysis 2390
7.2.1 Linear Applications 2390
7.2.2 Holomorphic Functions 2399
7.2.2.1 Orthogonality of real and imaginary iso-curves 2405
7.2.3 Complex Logarithm 2407
7.2.4 Complex Integral Calculus 2411
7.2.4.1 Convergence of a complex series 2418
7.2.5 Path Decomposition 2427
7.2.5.1 Inverse Path 2428
7.2.6 Laurent Series 2430
7.2.7 Singularities 2437
7.2.8 Residue Theorem 2440
7.2.8.1 Pole at infinity 2444
7.3 Topology 2446
7.3.1 General Topology 2447
7.3.1.1 Topological Spaces 2448
7.3.2 Metric Space and Distance 2450
7.3.2.1 Equivalent Distances 2457
7.3.2.2 Lipschitz Functions 2457
7.3.2.3 Continuity and Uniform Continuity 2460
7.3.3 Opened and Closed Set 2463
7.3.3.1 Balls 2464
7.3.3.2 Partitions 2467
7.3.3.3 Formal Ball 2470
7.3.3.4 Diameter 2470
7.3.4 Varieties 2473
7.3.4.1 Subvariety 2475
7.3.4.2 Surfaces Homeomorphism 2476
7.3.4.3 Differential Varieties 2479
7.4 Measure Theory 2481
7.4.1 Measurable Spaces 2482
7.4.1.1 Monotone Classes 2491
7.4.2 Lebesgue measure in probabilities 2493
7.4.2.1 Lebesgue Measure of the rational and irrational numbers 2496

8 Geometry 2499

8.1 Trigonometry 2503
8.1.1 Radian 2503
8.1.2 Circle Trigonometry 2508

8.1.2.1 Remarkable trigonometric triangle identities . . . 2519
8.1.2.1.1 Laws of Cosines (cosine rule) . . . 2524
8.1.2.1.2 Laws of Sines (sine rule) . . . 2528
8.1.3 Hyperbolic Trigonometry . . . 2532
8.1.3.1 Remarkable hyperbolic identities . . . 2539
8.1.4 Spherical Trigonometry . . . 2542
8.1.4.1 Girard's theorem . . . 2546
8.1.4.2 Four-parts formula . . . 2548
8.1.5 Solid Angle . . . 2549
8.2 Euclidean Geometry . . . 2553
8.2.1 Objects of Euclidean Geometry . . . 2553
8.2.1.1 Dimensions . . . 2555
8.2.2 Euclid's Constructions . . . 2563
8.2.2.1 Segments and Lines . . . 2566
8.2.2.1.1 Quantities of the same type . . . 2568
8.2.3 Plane Geometry . . . 2573
8.2.3.1 Displacements and Turnarounds . . . 2574
8.2.3.2 Plane angles . . . 2575
8.2.3.2.1 Angle Measurements . . . 2583
8.2.3.2.2 Units of Angle Measurements . . . 2586
8.2.3.2.3 Bisector . . . 2591
8.2.3.3 Triangles . . . 2593
8.2.3.3.1 Equal Triangles (congruent triangles) . . . 2594
8.2.3.3.2 Isosceles Triangles . . . 2597
8.2.3.3.3 Equilateral Triangles . . . 2601
8.2.3.3.4 Right Triangle . . . 2601
8.2.3.3.5 Right Isosceles Triangle . . . 2603
8.2.3.3.6 Inequalities in the triangles . . . 2605
8.2.3.3.7 Triangles remarkable interior lines . . . 2607
8.2.3.3.8 Pythagorean theorem . . . 2609
8.2.3.3.9 Thales' Theorem (intercept theorem) . . . 2610
8.2.3.3.9.1 Euclid's (triangle) theorem . . . 2615
8.2.3.4 Parallelism . . . 2618
8.2.3.5 Circle . . . 2620
8.2.3.5.1 Circumscribed circle theorem . . . 2623
8.2.3.5.2 Inscribed circle theorem . . . 2624
8.2.3.5.3 Thales' theorem of the circle . . . 2626
8.2.3.5.4 Central angle theorem . . . 2627
8.2.4 Hilbert's Axioms . . . 2631
8.2.4.1 Incidence Axioms (axioms of association) . . . 2632
8.2.4.2 Order Axioms . . . 2632
8.2.4.3 Congruence Axioms . . . 2633
8.2.4.4 Continuity Axioms . . . 2633
8.2.4.5 Parallels Axioms . . . 2634
8.2.5 Barycenter (centroid) . . . 2635

8.2.6 Geometric Transformations . . . 2640
8.2.6.1 Translation . . . 2642
8.2.6.2 Homothety (scaling) . . . 2643
8.2.6.3 Shear (skew) transformation . . . 2645
8.2.6.4 Rotation . . . 2647
8.2.6.4.1 Gimbal lock . . . 2652
8.2.6.4.2 Euler angles . . . 2656
8.2.6.5 Reflection . . . 2659
8.3 Non-Euclidean Geometry . . . 2663
8.3.1 Curvature(s) . . . 2665
8.3.2 Axioms of non-euclidean geometry . . . 2667
8.3.3 Generalized angle sum theorems . . . 2668
8.3.4 Geodesic and Metric Equation . . . 2669
8.3.5 Riemann Spaces . . . 2673
8.4 Projective Geometry . . . 2679
8.4.1 Conical Perspective (Central Perspective) . . . 2680
8.4.1.1 Images of Points . . . 2684
8.4.1.2 Images of Straight Lines . . . 2701
8.4.2 Affine projections . . . 2712
8.4.2.1 Isometric perspective . . . 2714
8.4.2.2 Oblique perspective . . . 2720
8.4.2.3 Orthogonal projection . . . 2722
8.4.3 Spherical projections . . . 2727
8.4.3.1 Stereographic projection . . . 2729
8.4.3.2 Cylindrical projection . . . 2738
8.4.3.3 Mercator projection . . . 2739
8.4.3.4 Lambert's equivalent projection (Peters projection) . . . 2743
8.4.3.5 Mollweide projection . . . 2746
8.4.4 Other perspectives and Laws of linear perspective . . . 2751
8.4.5 Homogeneous Coordinates (projection coordinates) . . . 2759
8.4.5.1 $\mathcal{P}^2$ Projective Space . . . 2759
8.4.5.2 $\mathcal{P}^3$ Projective Space . . . 2761
8.5 Analytical Geometry . . . 2764
8.5.1 Conics . . . 2764
8.5.1.1 Algebraic approach . . . 2765
8.5.1.2 Geometric Approach . . . 2775
8.5.1.3 Dandelin Theorem (Dandelin spheres) . . . 2793
8.5.1.4 Classification of conical by the determinant . . . 2794
8.5.2 Parametrizations . . . 2800
8.5.2.1 Equation of the Plane . . . 2800
8.5.2.2 Equation of the Straight line . . . 2804
8.5.2.2.1 Distance from a line to a point . . . 2807
8.5.2.2.2 Line defined by the intersection of planes . . . 2808
8.5.2.2.3 Parametric equation of a line in $\mathbb{R}^3$ . . . 2809
8.5.2.3 Equation of a Circle . . . 2810

8.5.2.4 Equation of a Square . . . 2812
8.5.2.5 Equation of an Hexagon . . . 2813
8.5.2.6 Equation of a Cycloid . . . 2814
8.5.2.7 Equation of an Epicycloid . . . 2816
8.5.2.8 Equation of a Spiral . . . 2818
8.5.2.9 Equation of a Hypocycloid . . . 2821
8.5.2.10 Surface of revolution . . . 2823
8.5.2.10.1 Cone . . . 2823
8.5.2.10.2 Sphere . . . 2825
8.5.2.10.2.1 Great circle of sphere . . . 2828
8.5.2.10.2.2 n-Sphere volume . . . 2829
8.5.2.10.3 Ellipsoid (spheroid) . . . 2834
8.5.2.10.4 Cylinder . . . 2837
8.5.2.10.5 Paraboloid . . . 2839
8.5.2.10.6 Hyperboloid . . . 2841
8.5.2.10.7 Torus . . . 2848
8.6 Differential Geometry . . . 2850
8.6.1 Parametric Curves . . . 2850
8.6.2 Isolines . . . 2857
8.6.3 Frenet Frame . . . 2863
8.6.4 Surface Patches . . . 2878
8.6.4.1 Metric of a Surface Patch . . . 2879
8.6.4.1.1 Regularity of a Surface . . . 2882
8.7 Geometric Shapes . . . 2886
8.7.1 Usual Surfaces (areas) . . . 2888
8.7.1.1 Polygons . . . 2888
8.7.1.2 Rectangle . . . 2894
8.7.1.3 Square . . . 2895
8.7.1.4 Unspecified Triangle . . . 2896
8.7.1.5 Isosceles Triangle . . . 2900
8.7.1.6 Equilateral Triangle . . . 2902
8.7.1.7 Right Triangle . . . 2905
8.7.1.8 Trapezoid . . . 2906
8.7.1.9 Parallelogram . . . 2908
8.7.1.10 Hexagon . . . 2909
8.7.1.11 Rhombus . . . 2914
8.7.1.12 Circle . . . 2915
8.7.1.13 Ellipse . . . 2917
8.7.1.13.1 Ellipse section surface . . . 2920
8.7.2 Usual Volumes . . . 2923
8.7.2.1 Polyhedron . . . 2924
8.7.2.1.1 Parallelepiped . . . 2925
8.7.2.1.1.1 Moment of Inertia of a rectangular plate . . . 2926
8.7.2.1.1.2 Moment of Inertia of a triangular plate . . . 2929
8.7.2.1.2 Pyramid . . . 2931

8.7.2.1.2.1 Moment of Inertia of a regular square pyramid . . . 2932
8.7.2.1.3 Right Prism . . . 2933
8.7.2.1.4 Regular Polyhedron . . . 2934
8.7.2.1.5 Regular Tetrahedron . . . 2939
8.7.2.1.6 Regular hexahedron (cube) . . . 2942
8.7.2.1.7 Regular octahedron . . . 2943
8.7.2.1.8 Regular Icosahedron . . . 2947
8.7.2.1.9 Regular Dodecahedron . . . 2951
8.7.2.2 Usual Solids of Revolution . . . 2956
8.7.2.2.1 Cylinder . . . 2959
8.7.2.2.2 Cone . . . 2963
8.7.2.2.3 Sphere . . . 2964
8.7.2.2.3.1 Spherical cap surface . . . 2969
8.7.2.2.3.2 Sphere packing . . . 2970
8.7.2.2.4 Torus . . . 2973
8.7.2.2.5 Ellipsoid (spheroid) . . . 2977
8.7.2.2.6 Paraboloid . . . 2983
8.7.2.2.7 Gabriel's horn . . . 2986
8.7.2.2.8 Wine Barrel with Circular Section . . . 2987
8.7.3 Isoperimetric inequality (2D) . . . 2993
8.8 Graph Theory . . . 2999
8.8.1 Type of Graphs and Structures . . . 3000
8.8.2 Graph Adjacency Matrix . . . 3026
8.8.3 Categories . . . 3031
8.9 Knot Theory . . . 3035
8.9.1 Braids Representation . . . 3035
8.9.1.1 Braids Group . . . 3037
8.9.2 Knot Representation . . . 3040
8.9.2.1 Knots Group . . . 3044
8.9.3 Tait's Knot . . . 3048
8.9.4 Mathematical Formalisation . . . 3052
8.9.4.1 Planar Representation . . . 3061

9 Mechanics 3063

9.1 Principia . . . 3068
9.1.1 System of Units . . . 3072
9.1.1.1 Dimensional Analysis . . . 3081
9.1.1.1.1 Time . . . 3083
9.1.1.1.2 Length . . . 3085
9.1.1.1.3 Mass . . . 3087
9.1.1.1.4 Energy . . . 3092
9.1.1.1.5 Electric Charge . . . 3095
9.1.1.2 Scientific Notation and Metric Prefixes . . . 3098
9.1.1.3 Scales of Measurements . . . 3101
9.1.2 Distributions . . . 3112

9.1.3 Constants 3113
9.1.3.1 Mathematical Constants 3114
9.1.3.2 Universal Constants (fundamental constants) 3114
9.1.3.3 Astronomical/Astrophysical parameters and constants 3119
9.1.3.4 Chemical parameters 3119
9.1.3.5 Material parameters 3121
9.1.3.6 Planck's constants 3122
9.1.4 Principles of Physics 3127
9.1.4.1 Principle of Causality 3128
9.1.4.2 Principle of Conservation of Energy 3129
9.1.4.3 Principle of Least Action 3131
9.1.4.4 Noether's Principle (Noether's theorem) 3132
9.1.4.4.1 Invariance by translation in space 3134
9.1.4.4.2 Invariance by rotation in space 3136
9.1.4.4.3 Invariance by translation in time 3136
9.1.4.4.4 Noether's theorem 3139
9.1.4.5 Curie's Principle 3144
9.1.5 Point Spaces 3145
9.2 Analytical Mechanics 3152
9.2.1 Lagrangian formalism 3155
9.2.1.1 Generalized coordinates and frames 3156
9.2.1.2 Variational Principle 3160
9.2.1.3 Euler-Lagrange Equation 3162
9.2.1.3.1 Beltrami Identity 3177
9.2.1.3.2 Theorem of Variational Calculus 3178
9.2.2 Canonical Formalism 3180
9.2.2.1 Legendre Transform 3180
9.2.2.2 Hamiltonian 3181
9.2.2.3 Poisson bracket 3189
9.2.2.4 Canonical transformations 3192
9.3 Classical Mechanics 3194
9.3.1 Newton's laws 3199
9.3.1.1 Newton's first law (Inertia law) 3199
9.3.1.2 Newton's second law (Fundamental Principle of Dynamics) 3202
9.3.1.3 Newton's third law (Law of Action and Reaction) 3209
9.3.2 Center of Mass and Reduced Mass 3213
9.3.2.1 Center of Mass Theorem 3216
9.3.2.2 Guldin's Theorem 3221
9.3.3 Kinematics of Rectilinear Motion 3224
9.3.3.1 Position 3225
9.3.3.2 Velocity and speed 3225
9.3.3.2.1 Motion parallax 3229
9.3.3.3 Acceleration 3231
9.3.3.3.1 Osculator Plane 3237
9.3.3.4 Galilean Relativity Principle 3239

9.3.4 Angular Momentum 3244
9.3.4.1 Moments 3251
9.3.5 Static Forces 3256
9.3.5.1 Lami's theorem 3261
9.3.6 Ballistics 3263
9.3.6.1 Ballistic equation without friction 3265
9.3.6.2 Ballistic equation with friction 3269
9.3.6.3 Ballistic accounting Earth's curvature 3271
9.3.7 Kinematics of Circular Motion (rotational kinematics) 3280
9.3.8 Energy, Work and Power 3293
9.3.8.1 Conservative vector field 3294
9.3.8.2 Kinetic Energy 3297
9.3.8.2.1 Moment of inertia 3297
9.3.8.2.2 Euler's rotation equation 3317
9.3.8.2.2.1 Derivation in the stationary reference frame 3319
9.3.8.2.2.2 Derivation in the body reference frame 3320
9.3.8.2.3 Gyroscope 3324
9.3.8.2.3.1 Classical Approach with precession only 3326
9.3.8.2.3.2 Lagrangian Approach with precession and nutation 3332
9.3.8.2.4 König's kinetic and angular momentum theorems 3347
9.3.8.2.4.1 First König's Theorem (König's angular momentum theorem) 3348
9.3.8.2.4.2 Second König's Theorem (König's kinetic energy theorem) 3349
9.3.8.3 Gravitational Potential Energy 3351
9.3.8.3.1 Gravitational Potential Energy of a Material Sphere 3356
9.3.8.3.2 Gravitational Effective potential 3358
9.3.8.4 Conservation of Total Mechanical Energy 3360
9.3.8.4.1 Generalized Newton law 3362
9.3.8.5 Conservation of Linear Momentum 3367
9.3.8.5.1 Elastic Collision in 1-dimensions 3369
9.3.8.5.2 Elastic Collision in 2-dimensions 3371
9.3.8.5.3 Inelastic Collision in 2-dimensions 3372
9.3.8.5.4 Rebound (stacked ball drop) 3374
9.3.8.6 Power 3379
9.3.8.6.1 Power of a turning machine 3381
9.3.8.6.1.1 Power yield 3381
9.3.9 Relative Movements and Inertial Forces 3383
9.3.9.1 Coriolis force and deflection magnitude 3391
9.3.9.1.1 Coriolis deviation of free falling objects 3396
9.3.9.2 Eötvös effect 3398
9.3.10 Oscillating Movements 3403
9.3.10.1 Newton's cradle 3403
9.3.10.2 Simple Pendulum (mass-wire pendulum) 3405
9.3.10.3 Physical Pendulum 3413
9.3.10.4 Elastic Pendulum (mass-spring pendulum) 3416
9.3.10.4.1 One degree of freedom elastic pendulum with/without friction 3416

9.3.10.4.2 Harmonic oscillator (Lorentz oscillator) . . . 3421
9.3.10.4.3 Two degrees of freedom elastic pendulum without friction . . . 3421
9.3.10.5 Conical Pendulum . . . 3423
9.3.10.6 Torsion Pendulum . . . 3425
9.3.10.7 Foucault's Pendulum . . . 3426
9.3.10.8 Huygens' Pendulum (and brachistochrone curve) . . . 3433
9.3.10.9 Double Pendulum . . . 3440
9.3.10.10 Triple linear coupled pendulum . . . 3446
9.3.10.11 Inverted Pendulum . . . 3448
9.3.11 Tribology . . . 3452
9.3.11.1 Exponential Friction . . . 3462
9.3.11.2 Horizontal Viscous Friction . . . 3465
9.3.11.3 Vertical Viscous Friction . . . 3466
9.3.11.4 Stokes' Vertical Viscous Friction . . . 3467
9.3.11.5 Stokes' Horizontal Viscous Friction . . . 3470
9.3.11.6 Friction's Heat Factor . . . 3473
9.4 Wave Mechanics . . . 3475
9.4.1 Wave Function . . . 3475
9.4.2 Wave Equation . . . 3476
9.4.3 Type of Waves . . . 3478
9.4.3.1 Periodic Waves . . . 3478
9.4.3.2 Harmonic Waves . . . 3479
9.4.3.2.1 Phase velocity and Group velocity . . . 3481
9.4.3.3 Stationary Waves . . . 3483
9.4.3.4 Vibration Modes in a Stretched String . . . 3486
9.4.3.4.1 Dirichlet Conditions . . . 3488
9.4.3.4.2 Neumann Conditions . . . 3495
9.4.4 Non-relativistic Lagrangian of a String . . . 3496
9.4.5 Vibrational modes of a circular membrane . . . 3500
9.4.6 Phasors . . . 3510
9.4.7 Superposition of periodic waves . . . 3517
9.4.7.1 AM/FM Broadcasting . . . 3523
9.5 Statistical Mechanics . . . 3525
9.5.1 Statistical Information Theory . . . 3525
9.5.1.1 High Entropy Structures . . . 3536
9.5.1.2 Usual Shannon entropies . . . 3538
9.5.1.3 Perplexity . . . 3541
9.5.1.4 Kullback–Leibler divergence . . . 3542
9.5.2 Boltzmann law . . . 3549
9.5.3 Statistical Physics Distributions . . . 3554
9.5.3.1 Maxwell Distribution (velocity distribution) . . . 3554
9.5.3.2 Maxwell-Boltzmann Distribution . . . 3561
9.5.3.3 Boltzmann Distribution . . . 3568
9.5.3.4 Fermi-Dirac Distribution . . . 3571
9.5.3.5 Bose-Einstein Distribution . . . 3576

9.5.4 Brownian Motion (random walk) 3582
9.6 Thermodynamics 3597
9.6.1 Thermodynamic Variables 3599
9.6.2 Thermodynamics Systems 3602
9.6.3 Thermodynamic Transformations 3603
9.6.4 State Variables 3607
9.6.4.1 Phases 3609
9.6.5 Equation of State 3611
9.6.5.1 Ideal Gas law 3612
9.6.5.2 State equation of a Liquid 3613
9.6.5.3 State equation of Solids 3615
9.6.6 Laws of Thermodynamics 3618
9.6.7 Calorific Capacities (heat capacity) 3622
9.6.7.1 Newton's Law of Cooling 3635
9.6.8 Internal Energy 3637
9.6.8.1 Work (energy) of Mechanical Forces 3641
9.6.8.2 Enthalpy 3643
9.6.8.3 Laplace's law 3646
9.6.8.4 Saint-Venant Thermodynamic Equation 3651
9.6.8.5 Thermoelastics coefficients 3653
9.6.9 Heat 3656
9.6.9.1 Entropy 3658
9.6.9.1.1 Life, evolution, and the second law of thermodynamics 3665
9.6.9.1.2 Heat Flow 3666
9.6.9.2 Carnot Cycle 3668
9.6.10 Maxwell relations 3672
9.6.11 Continuity Equation 3679
9.6.11.1 Heat Equation 3684
9.6.11.1.1 Fick's laws of diffusion 3698
9.6.12 Thermal radiation 3703
9.6.12.1 Black Body radiation 3704
9.6.12.1.1 Stefan–Boltzmann law 3706
9.6.12.1.2 Planck's law 3716
9.6.12.2 Summing up all types of heat transfer 3731
9.7 Continuum Mechanics 3734
9.7.1 Rigid Bodies 3735
9.7.1.1 Pressures 3735
9.7.1.2 Elasticity of Solids 3738
9.7.1.2.1 Hooke's law 3744
9.7.1.2.2 Shear Modulus 3751
9.7.1.2.3 Compressibility Modulus (bulk modulus) 3760
9.7.1.2.4 Flexural Modulus (bending modulus) 3762
9.7.1.2.5 Tranverse Wave in Solids 3765
9.7.2 Liquids (fluid mechanics) 3770
9.7.2.1 Pascal's Theorem 3773

9.7.2.2 Viscosity . . . 3775
9.7.2.2.1 Poiseuille's law . . . 3779
9.7.2.2.2 Viscous force on a fluid element . . . 3782
9.7.2.3 Bernoulli's Theorem . . . 3783
9.7.2.3.1 Torricelli's barometer . . . 3791
9.7.2.3.2 Torricelli's law . . . 3793
9.7.2.3.3 Siphon . . . 3798
9.7.2.3.4 Communicating vessels . . . 3800
9.7.2.3.5 Venturi effect . . . 3803
9.7.2.3.6 Pitot Tube . . . 3806
9.7.2.3.7 Pressure drop (pressure loss) . . . 3809
9.7.2.4 Navier-Stokes Equations . . . 3811
9.7.2.4.1 Incompressible flow . . . 3828
9.7.2.4.1.1 Vorticity and vorticity equation . . . 3836
9.7.2.4.1.2 Circulation . . . 3845
9.7.2.4.2 Compressible flow . . . 3846
9.7.2.4.3 Static flow . . . 3846
9.7.2.4.4 Reynolds number . . . 3847
9.7.2.4.5 Boussinesq approximation (buoyancy) . . . 3851
9.7.2.4.6 Stokes' law . . . 3853
9.7.2.4.6.1 Why don't clouds fall? . . . 3861
9.7.2.4.7 Betz's law . . . 3862
9.7.2.4.7.1 Wind pressure . . . 3867
9.7.2.5 Levelling law . . . 3869
9.7.2.6 Hydrostatic Pressure . . . 3872
9.7.2.7 Archimedes' principle . . . 3875
9.7.2.8 Surface tension . . . 3882
9.7.2.9 Speed of sound in a liquid . . . 3885
9.7.2.10 Summary of fluid dynamics laws . . . 3886
9.7.3 Gas . . . 3888
9.7.3.1 Types of Gas . . . 3888
9.7.3.1.1 Perfect Gas . . . 3888
9.7.3.1.2 Real Gas . . . 3897
9.7.3.2 Virial Theorem . . . 3897
9.7.3.2.1 Equipartition theorem of energy derivation based on quantum physics and statistical mechanics . . . 3911
9.7.3.3 Kinetic Pressures (kinetic theory of gases) . . . 3918
9.7.3.4 Kinetic Temperature . . . 3921
9.7.3.5 Amagat and Dalton's law . . . 3924
9.7.3.6 Mean free path (in kinetic theory) . . . 3925
9.7.3.7 Sackur-Tetrode equation . . . 3931
9.7.3.7.1 Case for distinguishable particles . . . 3931
9.7.3.7.2 Case for indistinguishable particles . . . 3933
9.7.3.8 Generalized Bernoulli equation (for gases) . . . 3936
9.7.3.9 Magnus (gravitational) Effect . . . 3938
9.7.3.9.1 Naive approach for the spherical and cylinder rotating case . . . 3939

9.7.3.10 Summary of gas laws . . . 3943
9.7.4 Plasmas . . . 3945
9.7.4.1 Plasma Frequency . . . 3947
9.7.4.2 Lawson Criterion for D-T . . . 3951
9.7.4.3 Saha ionization equation . . . 3954
9.7.4.4 Ionospheric propagation . . . 3958

10 Electromagnetism 3964

10.1 Electrostatics . . . 3968
10.1.1 Electric Force . . . 3969
10.1.2 Electric Potential . . . 3976
10.1.2.1 Path Independence . . . 3979
10.1.3 Equipotential and Field lines . . . 3980
10.1.3.1 Infinite straight wire . . . 3982
10.1.3.2 Electric Rigid Dipole . . . 3984
10.1.4 Electric Flux . . . 3996
10.1.4.1 Capacitor . . . 3997
10.1.4.1.1 Dielectric strength . . . 4003
10.1.5 Electrostatic potential energy . . . 4005
10.2 Magnetostatics . . . 4008
10.2.1 Ampère's theorem . . . 4011
10.2.1.1 Infinitely long solenoid . . . 4014
10.2.1.2 Toroidal coils . . . 4018
10.2.1.3 Electromagnet . . . 4019
10.2.1.3.1 Strength of a magnet or electromagnet . . . 4020
10.2.2 Maxwell-Ampère Relation . . . 4022
10.2.3 Biot-Savart law . . . 4023
10.2.3.1 Magnetic field for a current loop . . . 4025
10.2.3.2 Magnetic field for an infinite wire . . . 4026
10.2.4 Magnetic dipole . . . 4030
10.2.4.1 Magnetic torque . . . 4050
10.2.5 Lorentz law (Lorentz force) . . . 4052
10.2.5.1 Magnetic Vector Potential . . . 4057
10.2.5.2 Work of Magnetic Field . . . 4062
10.2.5.3 Classical Hall effect . . . 4065
10.2.5.4 Larmor radius . . . 4068
10.2.5.5 Energy of a magnetic dipole . . . 4081
10.2.6 Langevin treatment of Diamagnetism and Paramagnetism . . . 4085
10.2.6.1 Langevin model of diamagnetism . . . 4085
10.2.6.2 Langevin model of paramagnetism . . . 4088
10.3 Electrodynamics . . . 4094
10.3.1 Maxwell Equations . . . 4095
10.3.1.1 First Maxwell Equation (constant electric flow) . . . 4095
10.3.1.2 Second Maxwell Equation (non-existence of magnetic monopole) . . . 4099
10.3.1.3 Third Maxwell Equation . . . 4101

10.3.1.3.1 Betatron . . . 4103
10.3.1.4 Fourth Maxwell Equation . . . 4106
10.3.1.5 Magnetic Monopoles . . . 4110
10.3.2 Charge conservation equation . . . 4114
10.3.3 Gauge Theory . . . 4115
10.3.3.1 Electromagnetic field tensor . . . 4121
10.3.4 Electromagnetic wave equation . . . 4138
10.3.4.1 Foucault light speed experiment . . . 4145
10.3.4.2 Helmholtz equation . . . 4152
10.3.4.3 Energy flow transportation (Poynting vector) . . . 4156
10.3.4.4 Emissions . . . 4160
10.3.5 Synchrotron radiation (bremsstrahlung) . . . 4163
10.3.5.1 Liénard-Wiechert potentials . . . 4168
10.3.5.2 Retarded Electric and Magnetic fields . . . 4175
10.3.5.3 Power radiated through an elementary surface . . . 4188
10.3.5.3.1 Larmor's equation (non-relativistic derivation) . . . 4198
10.3.5.3.2 Abraham-Lorentz force . . . 4201
10.3.6 Dipole antenna radiation pattern derivation (Hertz Electric Dipole) . . . 4203
10.3.6.1 The $L = \lambda/2$ (half-wave) far-field resonant antenna . . . 4214
10.4 Electrokinetics . . . 4219
10.4.1 Kirchhoff's laws . . . 4221
10.4.1.1 Mesh law (Kirchhoff's Loop Law) . . . 4221
10.4.1.2 Nodes law (Kirchhoff's Point Law) . . . 4222
10.4.2 Drude model . . . 4226
10.4.3 Ohm's law . . . 4233
10.4.3.1 Complex Conductivity . . . 4238
10.4.3.2 Simple (harmonic oscillator) model of dielectric and conductors (and plasmas) . . . 4240
10.4.3.2.1 Dielectrics . . . 4241
10.4.3.2.2 Conductors . . . 4244
10.4.3.3 Equivalent Resistance . . . 4246
10.4.3.4 Equivalent Capacities . . . 4248
10.4.3.5 Equivalent Inductance . . . 4250
10.4.4 Electromotive Force . . . 4253
10.4.4.1 Faraday's law of induction . . . 4258
10.4.4.1.1 Lenz's law . . . 4260
10.4.4.2 Inductance . . . 4261
10.4.5 Electromagnetic Waves in Conductors . . . 4268
10.4.5.1 Skin effect . . . 4268
10.4.5.2 Phase and group velocity of EM waves inside conductors (modified Maxwell equations) . . . 4272
10.4.6 Semiconductors . . . 4279
10.4.6.1 Non-degenerated statistic density of negative electric charge carriers . . . 4291
10.4.6.2 Non-degenerated statistic density of positive electric charge carriers . . . 4296
10.4.6.3 Energy bands (conduction band, band gap, valence band) . . . 4300
10.4.6.4 Ohm's law of Semiconductors . . . 4304
10.4.7 Superconductivity . . . 4307

10.4.7.1 Meissner effect (second London's equation) 4308
10.5 Optics (ray optics) 4314
10.5.1 Sources and Shadows 4315
10.5.2 Colors 4325
10.5.3 Radiometry/Photometry 4337
10.5.3.1 Energy flow 4337
10.5.3.1.1 Beer–Lambert law 4337
10.5.3.2 Light Intensity (Radiant Intensity) 4339
10.5.3.3 Energy Emittance (Radiant Emittance) 4340
10.5.3.4 Radiance and Luminance 4343
10.5.3.4.1 Lambert's law 4350
10.5.3.5 Kirchhoff's law of Radiation 4351
10.5.3.6 Spectral Decomposition 4352
10.5.4 Law of Refraction and Reflection 4353
10.5.4.1 Refractive index 4358
10.5.4.1.1 Index of refraction rate of change with altitude 4360
10.5.4.2 Snell's law 4364
10.5.4.3 Rainbow 4372
10.5.4.4 The 22° Halo 4385
10.5.4.5 Optical Mirages 4397
10.5.4.5.1 Lagrangian approach to optical mirage 4400
10.5.4.5.2 Snell's law differential approach to air mirroring 4403
10.5.4.5.3 Differential form of long range mirages 4410
10.5.4.6 Cherenkov radiation 4414
10.5.5 Descartes' Formulas 4416
10.5.5.1 Stigmatism 4420
10.5.5.2 Lenses 4426
10.5.5.2.1 Optical Magnification (zoom) 4441
10.5.5.2.2 Angle of View 4451
10.5.5.2.3 Human eyes 4458
10.5.5.2.3.1 Eye magnification 4459
10.5.5.3 Triangular Prism 4462
10.5.5.4 Pentaprism 4466
10.6 Wave Optics 4470
10.6.1 Huygens' principle 4470
10.6.1.1 Normal refraction 4479
10.6.1.2 Thin film interference 4481
10.6.1.3 Fresnel coefficients 4485
10.6.2 Fraunhofer Diffraction 4492
10.6.2.1 Case of a small rectangular aperture (slit) 4492
10.6.2.1.1 Optical resolution in the far field region for small rectangular aperture 4500
10.6.2.2 Case of a network of small rectangular apertures (slits) 4502
10.6.2.3 Young's interference experiment 4509
10.6.3 Fresnel diffraction 4517
10.6.3.1 Case of a rectangular aperture 4517

10.6.3.2 Case of a circular aperture . . . 4521
10.6.3.2.1 Optical resolution in the near field region for circular aperture . . . 4529
10.6.3.3 Case of a circular obstacle . . . 4536
10.6.4 Light polarization . . . 4538
10.6.4.1 Linear polarization . . . 4543
10.6.4.2 Elliptical polarization . . . 4544
10.6.4.3 Circular polarization . . . 4546
10.6.4.4 Natural polarization . . . 4547
10.6.4.5 Malus' law . . . 4550
10.6.5 Coherence and interference (interferometry) . . . 4554
10.6.5.1 Michelson Interferometer . . . 4562
10.6.5.1.1 Case of circular Haidinger fringes . . . 4568
10.6.5.1.2 Case of hyperbolic Haidinger fringes . . . 4569
10.6.5.1.3 Case of straight Haidinger fringes . . . 4573
10.6.5.1.4 Michelson interferometer and Special Relativity . . . 4577
10.6.5.2 Sagnac interferometer (Sagnac effect) . . . 4579
10.6.6 LASER . . . 4587
10.6.6.1 LASER Dispersion/Divergence . . . 4595
10.6.7 Fiber Optics . . . 4604

11 Atomistic **4612**

11.1 Corpuscular Quantum Physics . . . 4616
11.1.1 Dalton's Model . . . 4620
11.1.2 Thomson's Model . . . 4622
11.1.3 Rutherford's Model . . . 4623
11.1.4 Bohr's Model . . . 4626
11.1.4.1 Bohr's Postulates . . . 4626
11.1.4.2 Quantification . . . 4627
11.1.4.3 Hydrogen Type Atoms Model without dragging . . . 4629
11.1.4.4 Hydrogen Type Atoms Model with dragging . . . 4638
11.1.4.5 Neutron Assumption . . . 4641
11.1.5 Wilson and Sommerfeld's Model . . . 4644
11.1.6 Relativistic Sommerfeld Model . . . 4648
11.1.6.1 Magnetic dipole moment . . . 4666
11.1.6.2 Spin . . . 4669
11.1.6.3 Pauli exclusion principle . . . 4672
11.1.7 Electron configuration (atomic orbital) . . . 4675
11.2 Wave Quantum Physics . . . 4688
11.2.1 Postulates . . . 4690
11.2.1.1 1st Postulate: Quantum State . . . 4691
11.2.1.2 2nd postulate: Time evolution of a quantum state . . . 4694
11.2.1.3 3rd postulate: Observables and operators . . . 4695
11.2.1.4 4th postulate: Measure of a property (collapse postulate) . . . 4697
11.2.1.5 5th postulate: Average of a property . . . 4699
11.2.2 Classical principles of uncertainty . . . 4704

11.2.2.1 First classical uncertainty relation 4705
11.2.2.2 Second classical uncertainty relation 4706
11.2.2.3 Third classical uncertainty relation 4708
11.2.3 Quantum algebra 4711
11.2.3.1 Linear functional operators 4711
11.2.3.1.1 Hermitian and Self-adjoint operators 4714
11.2.3.1.2 Commutators and Anticommutators 4718
11.2.3.1.3 Representatives 4723
11.2.3.1.4 Eigenvalues and Eigenfunctions 4725
11.2.3.1.4.1 Orthogonality of eigenfunctions 4727
11.2.3.1.5 Dirac formalism 4729
11.2.3.1.5.1 Kets and Bras 4729
11.2.3.2 Robertson uncertainty relations 4733
11.2.3.2.1 Energy-Time Heisenberg's uncertainty principle 4739
11.2.3.2.2 Zero-Point Energy from the uncertainty principles 4740
11.2.3.3 Fourier uncertainty principle 4744
11.2.4 Schrödinger Model 4749
11.2.4.1 de Broglie associated wave 4749
11.2.4.1.1 Electronic microscopy 4752
11.2.4.2 Classical (time independent) Schrödinger Wave Equation 4758
11.2.4.2.1 Schrödinger Hamiltonian 4760
11.2.4.2.2 De Broglie normalization condition 4764
11.2.4.2.3 Bound and unbound states 4768
11.2.4.3 Classical Shrödinger equation of evolution 4769
11.2.4.3.1 Operator of evolution 4770
11.2.4.3.2 Conserved Quantities and Commutators 4776
11.2.4.3.3 Schrödinger picture by separation of variables 4777
11.2.4.3.4 Linear combination of eigen-states (quantum superposition) 4779
11.2.4.3.5 Schrödinger continuity equation 4783
11.2.4.4 Implications and Applications 4784
11.2.4.4.1 Free particles (zero potential) 4785
11.2.4.4.1.1 Gaussian Wavepacket 4792
11.2.4.4.2 Infinite (rectangular) potential well 4796
11.2.4.4.3 First approach 4797
11.2.4.4.4 Second approach 4802
11.2.4.4.5 Fermi Energy 4805
11.2.4.4.6 1D-Harmonic oscillator 4807
11.2.4.4.7 Planck's Zero-Point Energy 4822
11.2.4.4.8 Electromagnetic Casimir Force 4825
11.2.4.4.8.1 One dimensional case (naive approach) 4826
11.2.4.4.8.2 Three dimensional case 4829
11.2.4.4.9 Tunnel Effect (barrier potential / radioactivity) 4840
11.2.4.4.9.1 Quantum time tunneling 4856
11.2.4.5 Superposition principle 4857
11.2.4.6 Ehrenfest theorems 4870

11.2.4.6.1 Generalized Ehrenfest Theorem . . . 4874
11.2.5 Angular momentum and Spin . . . 4877
11.2.5.1 Spin–orbit interaction (LS coupling) . . . 4899
11.2.5.2 Spin operator for spin $1/2$ particles . . . 4903
11.2.6 Planck’s dimensions . . . 4905
11.2.7 Wave Quantum Physics Interpretation . . . 4911
11.2.7.1 Copenhagen interpretation . . . 4911
11.2.7.1.1 EPR paradox . . . 4913
11.2.7.2 De Broglie-Bohm (pilot-wave) interpretation . . . 4915
11.3 Relativistic Quantum Physics . . . 4919
11.3.1 Relativistic Schrödinger evolution equation . . . 4920
11.3.1.1 Antimatter . . . 4922
11.3.2 Generalized Klein-Gordon Equation . . . 4924
11.3.2.1 Generalized Klein-Gordon Equation continuity equation . . . 4927
11.3.3 Classical free Dirac equation . . . 4931
11.3.3.1 Gamma, alpha, beta matrices . . . 4941
11.3.3.2 Free particle solution . . . 4946
11.3.3.3 Free particle solution at rest . . . 4952
11.3.3.4 Majorana equation . . . 4954
11.3.4 Linearised Dirac Equation . . . 4957
11.3.4.1 Linearised Dirac Equation continuity equation . . . 4970
11.3.4.2 Linearised Dirac Equation Angular Momentum + Helicity Conservation . . . 4973
11.3.5 Generalized Dirac Equation . . . 4978
11.3.6 Pauli Equation . . . 4979
11.3.6.1 Landé g-factor . . . 4989
11.3.6.2 Lamb Shift . . . 4996
11.3.7 CPT Symmetries . . . 5002
11.3.7.1 Charge symmetry . . . 5002
11.3.7.2 Parity symmetry . . . 5003
11.3.7.3 CP symmetry (from Matter to Antimatter) . . . 5005
11.3.7.4 Time Reversal symmetry . . . 5005
11.3.7.5 CP and T symmetries . . . 5006
11.3.7.6 CPT violations . . . 5007
11.3.7.7 CPT classical analogy (with Maxwell equations) . . . 5011
11.3.7.8 CPT with Dirac equation . . . 5013
11.4 Nuclear Physics . . . 5019
11.4.1 Nuclear Weapon . . . 5019
11.4.2 Radioactivity . . . 5023
11.4.2.1 Disintegration . . . 5029
11.4.2.1.1 Half-life and Mean life isotope . . . 5030
11.4.2.2 Activity . . . 5033
11.4.2.2.1 Radiometric dating . . . 5036
11.4.2.2.1.1 Carbon-14 dating (radiocarbon dating) . . . 5036
11.4.2.2.1.2 Rubidium-87 dating (meteorites dating/isochron dating) . . . 5042
11.4.2.2.1.3 Young Earth heat problem . . . 5047

11.4.2.2.2 Radioactive Decay chain 5050
11.4.2.3 Two level radioactive cascade 5051
11.4.2.3.1 Secular equilibrium 5053
11.4.2.3.2 Transient equilibrium 5055
11.4.2.3.3 Non-equilibrium 5055
11.4.2.4 Radioactive phenomena 5056
11.4.2.4.1 Nuclear Fusion (1) 5064
11.4.2.4.2 Nuclear Fission (2) 5068
11.4.2.4.3 Alpha Disintegration (3) 5075
11.4.2.4.4 Beta- Disintegration (4) 5084
11.4.2.4.5 Beta+ Disintegration (5) 5087
11.4.2.4.6 Electronic capture (6) 5088
11.4.2.4.7 Gamma emission (7) 5089
11.4.2.4.8 Internal conversion (8) 5090
11.4.3 Radiation protection 5092
11.4.3.1 Bethe formula 5092
11.4.3.2 Compton scattering 5096
11.4.3.2.1 Thomson scattering 5100
11.4.3.3 Rayleigh scattering 5102
11.4.3.4 Photoelectric effect 5106
11.4.3.5 Rutherford scattering 5111
11.4.3.6 X-rays and Gamma rays 5120
11.4.3.6.1 Electron-Positron pair production 5126
11.4.3.7 Radiation dosimetry 5129
11.4.4 Liquid Drop Model of Nucleus 5141
11.4.4.1 Volumetric binding energy 5143
11.4.4.2 Superficial binding energy 5143
11.4.4.3 Coulomb electrostatic repulsive energy 5144
11.4.4.4 Energy of asymmetry (Pauli energy) 5145
11.4.4.5 Pairing Energy 5149
11.5 Quantum Field Theory 5158
11.5.1 Yukawa potential 5163
11.5.1.1 Mass fields 5166
11.5.1.2 Non-mass fields 5167
11.5.2 Euler-Lagrange equation for Fields 5169
11.5.2.1 Field interactions ($\lambda\varphi^4$ theory) 5182
11.5.3 Gauge Theories 5185
11.5.3.1 Global Gauge invariance 5186
11.5.3.2 Local Gauge invariance 5188
11.6 Elementary Particle Physics 5194
11.6.1 Coupling Constants 5196
11.6.2 Spin magnetic resonance 5206

12 Cosmology 5212

12.1 Astronomy (Celestial Mechanics) 5217

12.1.1 Sun, Moon distances and radii . . . 5218
12.1.2 Drake Equation . . . 5225
12.1.2.1 Habitale Zone (HZ) . . . 5227
12.1.3 Kepler's Laws . . . 5233
12.1.3.1 First Kepler's Law (conicity law) . . . 5234
12.1.3.2 Second Kepler's Law (area law) . . . 5234
12.1.3.2.1 Acceleration variation on an elliptic orbit . . . 5237
12.1.3.2.2 Time of flight . . . 5238
12.1.3.3 Third Kepler's Law (periods' law) . . . 5242
12.1.4 Newton's Gravitational Law . . . 5255
12.1.4.1 Cavendish Experiment . . . 5270
12.1.4.2 Gaussian formulation of Newtonian Gravity . . . 5277
12.1.4.3 Shell Theorem . . . 5281
12.1.4.4 Orbital speed . . . 5282
12.1.4.5 Gravitational field of a homogeneous disc . . . 5284
12.1.4.6 Collision time of two objects at rest . . . 5295
12.1.4.7 Asteroids/Meteors impact velocity . . . 5297
12.1.4.8 Spherisation of Celestial Bodies . . . 5299
12.1.4.8.1 Flattening of Celestial Bodies (rotational flattening) . . . 5300
12.1.4.8.2 Empirical evidence that Earth isn't globally flat . . . 5314
12.1.4.9 Stability of Atmospheres . . . 5331
12.1.4.10 Planetary equilibrium temperature . . . 5335
12.1.5 Roche's Limit . . . 5337
12.1.6 Keplerian Orbitals . . . 5340
12.1.6.1 First Binet Formula (radial velocity) . . . 5340
12.1.6.2 Second Binet Formula (radial acceleration) . . . 5346
12.1.6.3 Keplerian orbital period . . . 5350
12.1.6.4 Average distance between two planets . . . 5356
12.1.6.5 Classical deflection of light (light bending) . . . 5358
12.1.6.6 Classical precession of perihelion (apsidal precession) . . . 5361
12.1.7 Duration of the diurnal arc . . . 5372
12.1.7.1 Evolution of diurnal arc (i.e. Earth-Moon distance) . . . 5379
12.1.8 Trigonometric parallax . . . 5383
12.1.9 Planets' Motion . . . 5388
12.1.9.1 Synodic and Sidereal period . . . 5390
12.1.9.2 Planet's apparent retrograde motion . . . 5395
12.1.10 Equation of Time and Analemma . . . 5404
12.1.10.1 Eccentricity component (variable angular velocity of Earth) . . . 5407
12.1.10.2 Obliquity (inclined plane of the ecliptic) and Sun anomaly . . . 5419
12.1.10.3 Calculation of the Equation of Time . . . 5423
12.1.11 Sundials . . . 5438
12.1.11.1 Equatorial sundials . . . 5441
12.1.11.2 Vertical sundials . . . 5443
12.1.11.3 Horizontal sundials . . . 5446
12.1.11.4 Anammelatic (azimuthal) sundials . . . 5449

12.1.12 Apparent solar time (true solar time) 5461
12.1.13 Celestial sphere and coordinate systems (positional astronomy) 5465
12.1.13.1 Local Sidereal Time 5467
12.1.13.2 Equatorial coordinates to Azimuthal (horizontal) coordinates 5472
12.1.13.3 Azimuthal (horizontal) coordinates to Equatorial coordinates 5474
12.1.13.4 Apparent daily path of the Sun (solargraph) 5476
12.1.14 Eclipses and saros cycles 5486
12.1.15 Lagrange Points 5498
12.1.15.1 Equilibrium points of the first type 5509
12.1.15.1.1 L1 Lagrange point 5509
12.1.15.1.2 L2 Lagrange point 5514
12.1.15.1.3 L3 Lagrange point 5517
12.1.15.2 Equilibrium points of the second type 5519
12.1.15.2.1 L4, L5 Lagrange points 5519
12.1.15.3 Hill sphere (sphere of influence) 5525
12.1.16 Relativistic Doppler-Fizeau Effect 5529
12.1.16.1 Apparent speed 5536
12.2 Astrophysics 5539
12.2.1 Stars 5539
12.2.1.1 Stellar Physics 5549
12.2.1.1.1 Collapse of an Interstellar Cloud 5549
12.2.1.1.1.1 Limit Mass Cloud for Ionization (rogue planets) 5552
12.2.1.1.1.2 Limit Mass Cloud for Fusion (Black Dwarf) 5553
12.2.1.1.1.3 Entropy of a gas collapsing under gravity 5555
12.2.1.1.2 Nuclear Duration Life 5558
12.2.1.1.3 Internal Temperature 5562
12.2.1.1.4 External Temperature 5566
12.2.1.1.5 Equation of Hydrostatic Equilibrium 5573
12.2.1.1.6 Brightness 5576
12.2.1.1.7 Shining (apparent brightness) 5577
12.2.1.1.8 Apparent magnitude 5580
12.2.1.1.9 Absolute magnitude 5586
12.2.1.1.9.1 Why are stars not visible on most space photos and videos 5591
12.2.1.1.10 Eddington limit 5597
12.2.1.2 Pulsative Variable Stars 5601
12.2.1.3 Neutron Stars (magnetars/pulsars) 5606
12.2.1.3.1 Chandrasekhar limit 5607
12.2.1.3.2 Rotation break limit 5612
12.2.1.3.3 Neutron Star Masses and Densities 5614
12.2.1.3.4 Neutron Star magnetic field 5615
12.2.1.3.5 Spin-Down Luminosity 5616
12.2.2 Galaxies 5620
12.2.2.1 Probability of collisions in merging galaxies 5623
12.2.2.2 Radial Speed Anomaly 5625
12.3 Special Relativity 5627

12.3.1 Assumptions and Principles . . . 5628
12.3.1.1 Postulate of Invariance . . . 5629
12.3.1.2 Cosmological Principle . . . 5630
12.3.1.3 Special Relativity Principle . . . 5630
12.3.2 Lorentz Transformations/Boost . . . 5633
12.3.2.1 Displacement four-vector . . . 5637
12.3.2.1.1 Wave Equation Invariance . . . 5640
12.3.2.1.2 Hypergeometric interpretation . . . 5641
12.3.2.2 Velocity four-vector . . . 5642
12.3.2.3 Current four-vector . . . 5646
12.3.2.4 Acceleration four-vector . . . 5647
12.3.2.5 Relativistic sum of velocities . . . 5652
12.3.2.6 Relativistic lengths variation (length contraction) . . . 5654
12.3.2.7 Relativistic apparent angle rotation (Lampa–Terrell–Penrose effect) . . . 5657
12.3.2.8 Relativistic time variation (time dilatation) . . . 5659
12.3.2.8.1 Lewis Carrol Epstein interpretation . . . 5662
12.3.2.8.2 Hafele–Keating experiment (special relativity version) . . . 5664
12.3.2.8.3 Twins paradox . . . 5669
12.3.2.9 Apparent relativistic mass . . . 5670
12.3.2.9.1 Why is Gold yellow? . . . 5675
12.3.2.9.2 Mass–Energy equivalence . . . 5679
12.3.2.9.3 Relativistic Lagrangian . . . 5681
12.3.2.9.4 Relativistic (linear) momentum . . . 5684
12.3.2.9.4.1 Einstein relation . . . 5688
12.3.2.9.4.2 Time of flight . . . 5689
12.3.2.9.4.3 Relativistic force . . . 5691
12.3.2.10 Relativistic electrodynamics . . . 5694
12.3.2.10.1 Tensor field transformation . . . 5702
12.3.3 Minkowski space-time . . . 5703
12.3.3.1 Four-vectors . . . 5706
12.3.3.2 Universe light cone . . . 5707
12.3.4 Empirical evidence of Special Relativity theory . . . 5716
12.4 General Relativity . . . 5720
12.4.1 Assumptions and Principles . . . 5721
12.4.1.1 Equivalence Postulates . . . 5721
12.4.1.2 Mach Principle . . . 5725
12.4.2 Metrics . . . 5728
12.4.2.1 Schild Criteria (Einstein redshift effect Newtonian approach) . . . 5735
12.4.3 Equations of movement . . . 5738
12.4.3.1 Geodesic equations . . . 5747
12.4.3.2 Newtonian Limit . . . 5751
12.4.4 Stress-Energy Tensor . . . 5754
12.4.4.1 Stress-Energy Tensor for a perfect fluid . . . 5762
12.4.4.2 Electromagnetic stress–energy tensor . . . 5763
12.4.5 Einstein’s Field Equations . . . 5766

12.4.5.1 Cosmological Constant (CC) 5774
12.4.5.2 Weak field approximation with cosmological constant 5778
12.4.5.3 Einstein-Maxwell equations 5780
12.4.5.4 Schwarzschild Solution 5781
12.4.6 Empirical evidence of General Relativity theory 5795
12.4.6.1 Gravitational Redshift 5796
12.4.6.2 Precession of Mercury's Perihelion 5800
12.4.6.3 Deflection of Light (light bending) 5812
12.4.6.4 Shapiro Effect (delay) 5821
12.4.6.5 Hafele–Keating experiment (general relativity version) 5827
12.4.6.6 Black Holes 5835
12.4.6.6.1 Schwarzschild Black Holes 5836
12.4.6.6.1.1 Black Hole innermost stable circular orbit 5840
12.4.6.6.1.2 Black Hole photon sphere 5853
12.4.6.6.1.3 Black Hole free fall proper time to singularity 5855
12.4.6.6.1.4 Black Hole Flamm's paraboloid representation 5858
12.4.6.6.1.5 White Hole 5860
12.4.6.7 Gravitational waves 5861
12.4.7 Einstein-Hilbert action 5869
12.4.7.1 Einstein-Hilbert action with cosmological constant 5875
12.4.7.2 Einstein-Hilbert action with matter 5876
12.5 Cosmogony 5878
12.5.1 Newtonian (i.e. non-relativistic) Cosmological Models 5881
12.5.1.1 Hubble's Law 5884
12.5.2 Friedmann Equations (Newtonian derivation) 5892
12.5.2.1 Critical Density and Density parameters 5897
12.5.3 Friedmann-Lemaître Euclidean Cosmological Models 5905
12.5.3.1 Flat spaces ($k = 0$) 5905
12.5.3.1.1 Flat space dominated by matter 5907
12.5.3.1.2 Flat space dominated by radiation 5908
12.5.3.1.2.1 Standard form of the Friedmann equation 5911
12.5.3.1.3 Flat space dominated by vacuum 5918
12.5.3.2 Spherical spaces ($k > 0$) 5920
12.5.3.2.1 Spherical space dominated by matter 5927
12.5.3.2.2 Spherical space dominated by radiation 5930
12.5.3.2.3 Spherical space dominated by vacuum 5932
12.5.3.3 Hyperbolic spaces ($k < 0$) 5934
12.5.3.3.1 Hyperbolic space dominated by matter 5938
12.5.3.3.2 Hyperbolic space dominated by radiation 5941
12.5.3.3.3 Hyperbolic space dominated by vacuum 5944
12.5.4 Observable Universe 5946
12.5.5 Cosmic Microwave Background (CMB) 5966
12.5.5.1 Decoupling time 5968
12.5.5.2 Matter/Photon decoupling (recombination temperature) 5973
12.5.5.3 Background radiation temperature 5984

12.5.6 Friedmann-Lemaître-Robertson-Walker Cosmological Models . . . 5987
  12.5.6.1 Robertson-Walker metric . . . 5987
  12.5.6.2 Cosmological Redshift . . . 5994
  12.5.6.3 Comobile Universe Diameter . . . 5998
  12.5.6.4 Size of the Universe depending on its curvature . . . 6000
  12.5.6.5 General Friedmann equations . . . 6004
  12.5.6.6 Einstein (static) Universe model . . . 6012
  12.5.6.7 Milne linear Universe model . . . 6013
  12.5.6.8 Friedmann Universe models . . . 6014
12.5.7 Friedmann historical derivation . . . 6050
12.5.8 The Black Hole Universe . . . 6058
12.5.9 Borde-Guth-Vilenkin Theorem (BGV Theorem) . . . 6061
12.5.10 Empirical evidence of the Cosmic Inflation theory (i.e. Big Bang) . . . 6069
12.6 Quantum Gravity/Cosmology . . . 6073
12.6.1 Newton Quantum Gravity . . . 6074
12.6.2 Wheeler-DeWitt Quantum Cosmology . . . 6083
  12.6.2.1 Wheeler-DeWitt equation for Newtonian Universe . . . 6085
  12.6.2.2 Wheeler-DeWitt equation for FLRW Universe . . . 6090
12.6.3 Why our Universe is a quite massive unpredictable mess . . . 6097
12.7 String Theory . . . 6099
12.7.1 Wave equation of a transversal string . . . 6101
12.7.2 Non-relativistic Wave equation of a transversal string . . . 6104
  12.7.2.1 Nambu-Goto Action . . . 6109
12.7.3 Lagrangian of a String . . . 6117

13 Chemistry 6122

13.1 Quantum Chemistry . . . 6123
13.1.1 Infinite three-dimensional rectangular potential . . . 6125
13.1.2 Molecular Vibrations . . . 6128
13.1.3 Hydrogenoid Atom (Schrödinger model) . . . 6130
  13.1.3.1 Rigid Rotator . . . 6134
  13.1.3.2 Non-Rigid Rotator . . . 6173
    13.1.3.2.1 Potential Profile . . . 6187
13.2 Molecular Chemistry . . . 6190
13.2.1 Orbital Approximations . . . 6192
13.2.2 LCAO Method . . . 6196
13.2.3 Molecular Rotational Energy Levels . . . 6205
13.2.4 Molecular Vibrational Energy Levels . . . 6211
13.2.5 Morse potential . . . 6216
13.3 Analytical Chemistry . . . 6228
13.3.1 Analytical chemistry process . . . 6234
13.3.2 Simple Mixtures . . . 6235
13.3.3 Chemical formula . . . 6237
13.3.4 Reactions . . . 6241
13.4 Thermochemistry . . . 6249

13.4.1 Chemical transformations . . . 6249
13.4.2 Molar Quantities . . . 6252
13.4.2.1 Standard enthalpy of reaction . . . 6259
13.4.2.1.1 Kirchhoff's Enthalpy Law . . . 6262

14 Theoretical Computing 6267

14.1 Numerical Methods/Analysis . . . 6272
14.1.1 Computer Representation of Numbers . . . 6274
14.1.1.1 Decimal System . . . 6274
14.1.1.2 Binary system . . . 6275
14.1.1.2.1 Floating Point (Real Numbers) . . . 6277
14.1.1.2.2 Binary arithmetic (for fixed point/integers) . . . 6279
14.1.1.3 Hexadecimal System . . . 6280
14.1.1.4 Octal System . . . 6280
14.1.1.5 Conversion of decimal system to non-decimal system: . . . 6282
14.1.2 Algorithm Complexity . . . 6284
14.1.2.1 NP-Completude . . . 6291
14.1.3 Integer Part . . . 6293
14.1.4 Heron's Square Root Algorithm . . . 6295
14.1.5 Archimedes Algorithm . . . 6298
14.1.6 Euler's Number e . . . 6301
14.1.7 Stirling's factorial approximation . . . 6301
14.1.8 Linear Systems of Equations . . . 6304
14.1.8.1 One equation with one unknown . . . 6304
14.1.8.2 Two equations with two unknowns . . . 6304
14.1.8.3 Three equations with three unknowns . . . 6306
14.1.8.4 n equations with n unknowns . . . 6310
14.1.9 Polynomials . . . 6312
14.1.10 Regression Techniques . . . 6316
14.1.10.1 Univariate linear regression model . . . 6322
14.1.10.1.1 Regression line . . . 6323
14.1.10.1.2 Ordinary Least Squares Method (OLSM) . . . 6324
14.1.10.1.3 Univariate Regression Variance Analysis . . . 6329
14.1.10.1.4 F-test for Regression (significance test for linear regression) . . . 6335
14.1.10.1.4.1 Test for Lack of Fit . . . 6345
14.1.10.2 Univariate linear regression Gaussian Model . . . 6348
14.1.10.2.1 Pearson Correlation Coefficient Test . . . 6358
14.1.10.2.2 Confidence interval of predicted values . . . 6361
14.1.10.3 Linear univariate regression forced through the origin (RTO) . . . 6369
14.1.10.4 Deming regression (orthogonal regression) . . . 6372
14.1.10.4.1 Deming orthogonal regression log-likelihood derivation . . . 6372
14.1.11 Multiple linear regression Gaussian Model . . . 6379
14.1.11.1 Fisher Matrix . . . 6389
14.1.11.2 R^2 decomposition in multiple regression . . . 6390
14.1.11.3 Influential points . . . 6391

14.1.11.4 Significance of regression coefficients (variable importance) . . . 6395
14.1.11.5 Variance Inflation Factor (multicollinearity detection) . . . 6398
14.1.11.5.1 Gauss-Markov theorem . . . 6405
14.1.11.6 Weighted Least Squares (WLS) . . . 6407
14.1.11.7 Model Selection Criteria . . . 6412
14.1.11.7.1 Mallows's C_p Linear Regression Prediction . . . 6413
14.1.11.7.2 Akaike Information Criterion (AIC) . . . 6418
14.1.11.7.3 Cross-Validation metrics . . . 6428
14.1.12 Ridge, LASSO and Elastic Net regularization . . . 6435
14.1.12.1 L_2 Regularization or Ridge Regularization . . . 6437
14.1.12.2 L_1 Regularization or LASSO Regularization . . . 6441
14.1.12.3 $L_1 + L_2$ Regularization or Elastic Net Regularization . . . 6443
14.1.12.4 Polynomial regression . . . 6445
14.1.12.5 Kernel regression . . . 6447
14.1.12.6 Logistic regressions (LOGIT) . . . 6449
14.1.12.6.1 Binomial Logistic regression . . . 6450
14.1.12.6.1.1 Binomial Logistic regression relative variable importance . . . 6465
14.1.12.6.1.2 Binomial logistic regression residuals . . . 6466
14.1.12.6.1.3 Hosmer–Lemeshow test . . . 6468
14.1.12.6.1.4 Logistic Kolmogorov-Smirnov statistic . . . 6470
14.1.12.6.1.5 Equivalence between the 2-sample Wald's Z-test and the Rao score test over logistic regression with a single binary predictor . . . 6471
14.1.12.6.2 Multinomial Logistic regression . . . 6477
14.1.12.6.2.1 F_1 score . . . 6482
14.1.12.6.3 Ordinal Logistic regression . . . 6487
14.1.12.6.4 Log-Loss . . . 6494
14.1.12.6.5 Odds Ratio Confidence Interval . . . 6496
14.1.12.6.5.1 Four-fold plot . . . 6501
14.1.12.6.6 Risk Ratio and its Confidence Interval . . . 6503
14.1.12.6.7 ROC and Lift curves . . . 6511
14.1.12.6.7.1 ROC curve . . . 6512
14.1.12.6.7.2 Lift curve . . . 6522
14.1.12.7 General Linear Models (GLM) . . . 6530
14.1.12.7.1 Normal GLM . . . 6532
14.1.12.7.2 Poisson GLM . . . 6534
14.1.12.7.3 Binomial GLM . . . 6538
14.1.12.7.4 Binomial Negative GLM . . . 6541
14.1.12.7.5 Gamma GLM . . . 6549
14.1.12.8 Robust, M-estimators and W-estimators . . . 6557
14.1.13 Interpolation Techniques . . . 6569
14.1.13.1 Bézier Curves (Bézier Splines) . . . 6569
14.1.13.2 Linear ordering isotonic regression . . . 6582
14.1.13.3 Euler Method . . . 6587
14.1.13.4 Polynomial of collocation . . . 6591
14.1.13.5 Lagrange polynomials . . . 6594
14.1.13.6 Newton polynomials (divided differences) . . . 6597

14.1.13.7 Errors in Polynomial Interpolation . . . 6602
14.1.14 Roots search . . . 6603
14.1.14.1 Proportional parts methods . . . 6603
14.1.14.2 Secant method (Regula Falsi or False Position) . . . 6606
14.1.14.3 Bisection method (interval-halving) . . . 6609
14.1.14.4 Newton's method . . . 6612
14.1.15 Numerical Differentiation . . . 6619
14.1.16 Numerical Integration . . . 6621
14.1.16.1 Rectangles method . . . 6623
14.1.16.2 Trapezoidal method . . . 6624
14.1.16.3 Newton–Cotes formulas . . . 6626
14.1.16.3.1 Trapezoidal Rule . . . 6627
14.1.16.3.2 Composite Trapezoidal Rule . . . 6628
14.1.16.3.3 Simpson's Rule . . . 6629
14.1.16.4 Multidimensional integrals . . . 6632
14.1.16.5 Numerical solution of ordinary differential equations . . . 6637
14.1.16.5.1 Runge-Kutta methods . . . 6637
14.1.16.5.1.1 1st order Runge-Kutta method . . . 6637
14.1.16.5.1.2 2nd order Runge-Kutta method . . . 6638
14.1.16.5.1.3 4th order Runge-Kutta method . . . 6642
14.1.17 Optimization . . . 6645
14.1.17.1 Linear programming (Linear Optimization) . . . 6646
14.1.17.1.1 Graphical LP resolution . . . 6649
14.1.17.1.2 Algebraic LP resolution . . . 6651
14.1.17.1.3 Simplex algorithm LP resolution . . . 6657
14.1.17.2 Nonlinear programming (Nonlinear optimization) . . . 6662
14.1.17.2.1 Substitution Method . . . 6666
14.1.17.2.2 Lagrange Multipliers Method . . . 6667
14.1.17.2.3 Newton-Raphson Method (Quadratic Newton) . . . 6673
14.1.17.2.4 Gauss-Newton method's (Tangent Newton) . . . 6678
14.1.17.2.4.1 Gradient Descent/Ascent techniques summary . . . 6691
14.1.17.3 Expectation-Maximization (EM) algorithm . . . 6697
14.1.18 Resampling statistics . . . 6707
14.1.18.1 Monte Carlo Simulations . . . 6708
14.1.18.1.1 Inverse Transform Sampling . . . 6710
14.1.18.1.2 Random number generation . . . 6713
14.1.18.1.3 Monte Carlo integration . . . 6717
14.1.18.1.4 Monte Carlo estimation of π . . . 6721
14.1.18.1.5 Monte Carlo Modelling . . . 6723
14.1.18.1.6 Monte Carlo variance reduction . . . 6727
14.1.18.1.6.1 Method of Antithetic Variables . . . 6729
14.1.18.1.6.2 Method of Control Variables . . . 6730
14.1.18.1.6.3 Method of Conditioning . . . 6731
14.1.18.1.6.4 Method of Stratified Sampling . . . 6733
14.1.18.1.6.5 Importance Sampling (IS) . . . 6737

14.1.18.2 Bootstrapping . . . 6740
14.1.18.3 Jackknife Resampling . . . 6747
14.1.18.4 Permutation Tests . . . 6749
14.1.19 Finite difference method (F.D.M.) . . . 6752
14.1.19.1 One space dimension F.D.M. . . . 6753
14.1.19.1.1 Numerical instability . . . 6756
14.1.19.1.2 von Neumann stability . . . 6757
14.1.19.2 Space-time F.D.M (finite-volume method) . . . 6759
14.1.20 Data Science & Data Mining . . . 6768
14.1.20.1 Training vs Test Data . . . 6795
14.1.20.2 Hyperparameters vs Parameters . . . 6803
14.1.20.3 Association Rules . . . 6808
14.1.20.3.1 ZeroR . . . 6809
14.1.20.3.2 OneR (classification by induction) . . . 6810
14.1.20.3.3 Apriori . . . 6814
14.1.20.4 Equivalence CLAss Transformation (ECLAT) . . . 6820
14.1.20.5 Clustering and Classification . . . 6822
14.1.20.5.1 Naive Bayes classifiers . . . 6826
14.1.20.5.1.1 Binomial Naive Bayes classifier . . . 6826
14.1.20.5.1.2 Gaussian Naive Bayes classifier . . . 6830
14.1.20.5.2 k-nearest neighbours classifier (k-nn) . . . 6832
14.1.20.5.3 Regression and classification trees . . . 6835
14.1.20.5.3.1 Binary classification trees with quantitative variables . . . 6836
14.1.20.5.3.2 Detection of significant splits . . . 6846
14.1.20.5.3.3 Classification (and clustering) validation . . . 6847
14.1.20.5.3.4 Random Forests . . . 6852
14.1.20.5.4 k-means clustering . . . 6854
14.1.20.5.5 Support Vector Machines (SVM) classifier . . . 6867
14.1.20.5.6 Gaussian Mixture Model (GMM) clustering . . . 6874
14.1.20.5.7 Density-Based Spatial Clustering of Applications with Noise (DBSCAN) . . . 6883
14.1.20.5.8 Mean shift clustering . . . 6888
14.1.20.5.9 Hierarchical Ascendant Clustering (HAC) . . . 6896
14.1.20.5.10 Linear Discriminant Analysis . . . 6908
14.1.20.5.10.1 Two classes problem . . . 6909
14.1.20.5.10.2 Multiclass problem . . . 6916
14.1.20.6 Dimensionality Reduction . . . 6918
14.1.20.6.1 Principal Component Analysis . . . 6919
14.1.20.6.2 Linear Discriminant Analysis (LDA) . . . 6919
14.1.20.6.3 MultiDimensional Scaling (MDS) . . . 6920
14.1.20.6.3.1 Classical MDS (cMDS) . . . 6921
14.1.20.6.4 Factor Analysis (Exploratory Factor Analysis (EFA)) . . . 6925
14.1.20.6.4.1 Expectation-Maximization (EM) algorithm solution . . . 6932
14.1.20.6.4.2 Principal Factor Analysis Extraction solution . . . 6935
14.1.20.6.5 T-distributed Stochastic Neighbor Embedding (T-SNE) . . . 6939
14.1.20.6.5.1 KL gradient of Stochastic Neighbour Embedding (SNE) . . . 6942

14.1.20.6.5.2 KL gradient of T-distributed Stochastic Neighbour Embedding (T-SNE) . . . 6943
14.1.20.7 Gradient Boosting . . . 6945
14.1.20.8 Neural networks . . . 6951
14.1.20.8.1 Neuron model . . . 6953
14.1.20.8.2 Transfer functions . . . 6958
14.1.20.8.3 Various cost functions . . . 6961
14.1.20.8.4 Network Architecture . . . 6964
14.1.20.8.5 Backpropagation . . . 6978
14.1.20.8.6 Recursive Neural Networks (RNN) . . . 6986
14.1.20.8.7 Convolutional Neural Networks (CNN) . . . 6996
14.1.20.8.8 Generative Adversarial Network (GAN) . . . 7011
14.1.20.8.9 Objective, cost and loss function . . . 7020
14.1.20.9 Genetic Algorithms . . . 7022
14.1.20.9.1 Encoding and Initial population . . . 7026
14.1.20.9.2 Operators . . . 7028
14.1.20.9.2.1 Operator of selection . . . 7029
14.1.20.9.2.2 Crossover operator . . . 7030
14.1.20.9.2.3 Mutation operator . . . 7031
14.1.20.10Total Unduplicated Reach and Frequency Analysis (TURF) . . . 7036
14.2 Fractals . . . 7038
14.2.1 IFS Fractals . . . 7041
14.2.1.1 Fractals Metric Space . . . 7054
14.2.2 Fractals Visualization . . . 7059
14.2.2.1 Cantor’s Fractal (Cantor Set) . . . 7059
14.2.2.2 Triangle Sierpinski Fractal . . . 7062
14.2.2.3 Sierpinski carpet fractal . . . 7067
14.2.2.4 Fractal spirals . . . 7071
14.2.2.5 Von Koch fractal (Koch snowflake) . . . 7073
14.2.2.5.1 Coastline paradox . . . 7077
14.2.2.6 Natural fractals . . . 7078
14.2.2.6.1 Branch . . . 7079
14.2.2.6.2 Snowflake . . . 7081
14.2.2.6.3 Tree . . . 7085
14.2.2.6.4 Fern . . . 7087
14.2.3 Escape Time Algorithm Fractals . . . 7091
14.2.3.1 Mandelbrot Set . . . 7093
14.2.3.2 Julia Set . . . 7098
14.2.3.3 Newton Set . . . 7102
14.3 Logical Systems . . . 7104
14.3.1 Strict Logic . . . 7104
14.3.1.1 Boolean Algebra . . . 7105
14.3.1.2 Logical Functions (Boolean operators) . . . 7112
14.3.1.3 Karnaugh maps . . . 7116
14.3.1.4 Arithmetic Boolean (binary) operations . . . 7125
14.3.1.4.1 Prefix codes . . . 7133

14.3.2 Fuzzy logic . . . 7136
14.3.2.1 Fuzzy set . . . 7147
14.4 Error-Correcting Codes . . . 7152
14.4.1 CheckSum . . . 7156
14.4.1.1 Luhn algorithm . . . 7157
14.4.2 Check Digit . . . 7159
14.4.2.1 European Article Numbering (EAN-13) . . . 7159
14.4.2.2 Swiss Post Payment slip . . . 7161
14.4.2.3 International Bank Account Number (IBAN) . . . 7163
14.4.2.4 UIC wagon numbers . . . 7164
14.4.3 Permutations . . . 7166
14.4.4 Encoders . . . 7167
14.4.4.1 Block code . . . 7172
14.4.4.2 Systematic codes . . . 7179
14.5 Automata Theory . . . 7181
14.5.1 Von Neumann machine . . . 7182
14.5.2 Turing machine . . . 7184
14.5.3 Chomsky hierarchy . . . 7191
14.5.3.1 Formal language . . . 7191
14.5.3.2 Syntax . . . 7193
14.5.3.3 Grammar . . . 7194
14.5.3.4 Associated automata . . . 7197
14.5.4 Terminology . . . 7198
14.5.4.1 Words . . . 7198
14.5.4.2 Languages . . . 7201
14.5.4.3 Equations . . . 7203
14.5.4.4 Codes . . . 7205
14.5.4.4.1 Prefix codes . . . 7207
14.5.5 Linguistic algorithms . . . 7209
14.5.5.1 Huffman algorithm . . . 7209
14.5.5.2 Sardinas and Patterson algorithm . . . 7214
14.6 Cryptography . . . 7216
14.6.1 Cryptographic systems . . . 7217
14.6.1.1 Kerckhoffs' principle . . . 7221
14.6.2 Traps . . . 7222
14.6.3 Secret-key encryption system . . . 7223
14.6.3.1 Feistel Schemes . . . 7225
14.6.4 Public key encryption . . . 7230
14.6.4.1 Diffie-Hellman protocol . . . 7232
14.6.4.2 R.S.A system . . . 7236
14.6.4.2.1 Euler's theorem . . . 7236
14.6.5 Hash functions . . . 7249
14.6.5.1 MD5 message digest condensation function . . . 7249
14.6.5.2 SHA-1 Secure Hash Algorithm condensation function . . . 7253
14.6.6 Certificate based authentication . . . 7255

14.6.7 Quantum cryptography . . . 7259
14.6.8 Alternative cryptography . . . 7262
14.7 Quantum Computing . . . 7263
14.7.1 Schrödinger’s Cat superposition . . . 7265
14.7.2 Photon polarization . . . 7267
14.7.3 Qubit . . . 7269
14.7.3.1 Bloch sphere . . . 7275
14.7.3.1.1 Qubit of polarization . . . 7280
14.7.3.1.2 1/2 spin Qubit . . . 7283
14.7.4 Entangled qubit system . . . 7288
14.7.5 Quantum logic gates . . . 7288

15 Social Sciences 7292

15.1 Population Dynamics . . . 7297
15.1.1 Birth rate and mortality tables (biometric features) . . . 7298
15.1.1.1 Population Renewal . . . 7308
15.1.2 Population Models . . . 7310
15.1.2.1 Exponential disease model . . . 7310
15.1.2.2 Deterministic Logistic Model (Verlhust) . . . 7313
15.1.2.3 Chaotic Logistic Model . . . 7317
15.1.2.3.1 Feigenbaum’s Bifurcation Diagram . . . 7327
15.1.2.4 Malthusian Growth Law . . . 7332
15.1.2.5 Leslie model . . . 7333
15.1.2.6 Epidemic models for spread of disease . . . 7335
15.1.2.6.1 Exponential Model . . . 7336
15.1.2.6.2 SI Model (Susceptible-Infected) . . . 7336
15.1.2.6.3 SIS Model (Susceptible-Infected-Susceptible) . . . 7338
15.1.2.6.4 SIR Model (Susceptible-Infected-Recovered) . . . 7339
15.1.2.6.5 Incidence rate . . . 7348
15.1.2.7 Lotka–Volterra predator–prey model . . . 7350
15.1.3 Schaefer’s optimal capture model . . . 7359
15.1.4 Hardy-Weinberg model . . . 7361
15.1.5 Mendel’s law . . . 7371
15.1.6 Growth rate with temperature . . . 7375
15.2 Game and Decision Theory . . . 7376
15.2.1 Behavioural decision bias (cognitive bias) . . . 7381
15.2.1.1 Sunk Cost . . . 7386
15.2.1.2 Anchoring Bias . . . 7388
15.2.1.3 Wisdom of Crowds . . . 7389
15.2.1.4 Delphi Method . . . 7392
15.2.2 Utility . . . 7394
15.2.2.1 Pareto Optimum . . . 7394
15.2.2.2 Nash Equilibrium . . . 7395
15.2.3 Games Representations . . . 7398
15.2.3.1 Extensive representation of a Game . . . 7399

15.2.3.2 Extensive representation of a Decision . . . 7400
15.2.3.2.1 Real Options . . . 7408
15.2.3.3 Normal representation of a Game . . . 7410
15.2.3.3.1 Repetitive Games . . . 7417
15.2.3.4 Set representation of a Game . . . 7420
15.2.3.5 Graphical representation of a Game . . . 7424
15.2.4 Expected Utility . . . 7428
15.2.4.1 Hurwitz Criterion . . . 7429
15.2.4.1.1 Pascal's wager . . . 7431
15.2.4.2 Laplace Criterion . . . 7436
15.2.4.3 Kelly criterion . . . 7437
15.2.4.3.1 Discrete time Kelly criterion . . . 7439
15.2.4.3.2 Continuous time Kelly criterion . . . 7442
15.2.5 Evolutionary Games . . . 7444
15.2.5.1 Dove & Hawk game in pure strategy (without probabilities) . . . 7448
15.2.5.2 Dove & Hawk game in stable evolutionary strategy (with probabilities) . . . 7449
15.2.6 Cournot Competition . . . 7452
15.2.7 Markov Decision Processes (MDPs) . . . 7456
15.2.8 Multi-Criteria Decision Making (MCDM) . . . 7462
15.2.8.1 Analytic Hierarchy Process (AHP) . . . 7462
15.2.9 Electoral systems . . . 7473
15.2.10 Ethics and Moral . . . 7478
15.2.10.1 Golden Rule of morality (Tit for Tat) . . . 7479
15.3 Economy . . . 7485
15.3.1 Concepts . . . 7486
15.3.1.1 Microeconomics . . . 7488
15.3.1.1.1 Willingness-To-Pay . . . 7496
15.3.1.1.2 Pricing optimization . . . 7497
15.3.1.1.3 Average & Marginal Cost/Revenue . . . 7502
15.3.1.2 Macroeconomics . . . 7508
15.3.1.2.1 Cobb-Douglas Model . . . 7509
15.3.2 Monetary Model . . . 7515
15.3.2.1 Monetary base conservation . . . 7517
15.3.2.2 Walras' law . . . 7520
15.3.3 Price Index and GDP . . . 7525
15.3.3.1 Paasche and Laspeyres price indices . . . 7526
15.3.3.2 Fisher index and Marshall–Edgeworth index . . . 7527
15.3.3.3 Gross domestic product (GDP) . . . 7528
15.3.4 Supply and Demand Theory . . . 7530
15.3.4.1 Expected utility theory . . . 7530
15.3.5 Net gain/loss opposite feedback model . . . 7538
15.3.6 Capitalization and Actuarial . . . 7547
15.3.6.1 Dates Intervals . . . 7549
15.3.6.2 Rates Equivalence . . . 7552
15.3.6.3 Simple Interest . . . 7553

15.3.6.3.1 Discounts . . . 7555
15.3.6.4 Compound Interest . . . 7557
15.3.6.5 Continuous Interest . . . 7560
15.3.7 Progressive interest (annuities) . . . 7562
15.3.7.1 Postnumerando annuities . . . 7563
15.3.7.2 Praenumerando annuities . . . 7565
15.3.8 Rounding . . . 7567
15.3.9 Loans Amortization/Repayments . . . 7570
15.3.9.1 Fixed-Term Loan . . . 7572
15.3.9.2 Loan with constant amortization . . . 7573
15.3.9.3 Loan with constant annuity . . . 7573
15.3.9.3.1 Equivalent Monthly Installment (EMI) . . . 7575
15.3.10 Modern Portfolio Theory . . . 7581
15.3.10.1 No Arbitrage Opportunity (N.A.O.) . . . 7585
15.3.10.2 Liquidity and Illiquidity . . . 7590
15.3.10.2.1 Bias ratio . . . 7593
15.3.10.2.2 Bid-Ask (spread) Roll Model . . . 7596
15.3.10.3 Portfolios . . . 7599
15.3.10.3.1 Stocks (shares of stocks)/Equities . . . 7607
15.3.10.3.1.1 Dividend Yield . . . 7609
15.3.10.3.1.2 Shares Benchmark Indices . . . 7612
15.3.10.3.1.3 Durand Model . . . 7612
15.3.10.3.2 Obligations (Bonds) . . . 7617
15.3.10.3.2.1 MacCaulay Duration, Modified Duration and Convexity . . . 7633
15.3.10.3.2.2 Equivalence of a Bond rate for a Treasury bill . . . 7640
15.3.10.3.3 Warrants . . . 7643
15.3.10.3.4 Futures & Forwards . . . 7647
15.3.10.3.4.1 Futures & Forwards naive pricing . . . 7652
15.3.10.3.4.2 Cox-Ingersoll-Ross equality of the Future/Forward price . . . 7656
15.3.10.3.4.3 Futures & Forwards commodity hedging . . . 7656
15.3.10.3.4.4 Options . . . 7663
15.3.10.3.5 Returns and Investments rates . . . 7679
15.3.10.3.5.1 Return on Investment (R.O.I.) . . . 7680
15.3.10.3.5.2 Internal Rate of Return (I.R.R.) . . . 7681
15.3.10.3.5.3 Modified Internal Rate of Return (M.I.R.R.) . . . 7682
15.3.10.3.5.4 Money Weighted Rate of Return (M.W.R.R.) . . . 7684
15.3.10.3.5.5 Time Weighted Rate of Return (T.W.R.R.) . . . 7687
15.3.10.3.6 Theory of Speculation . . . 7689
15.3.10.3.7 Portfolio efficient diversification models . . . 7698
15.3.10.3.7.1 Overall minimum variance portfolio (Markowitz portfolio) . . . 7699
15.3.10.3.7.2 Overall maximum Sharpe portfolio . . . 7713
15.3.10.3.7.3 Equal Risk Contribution optimal weights portfolio . . . 7732
15.3.10.3.7.4 Error tracking optimal weights portfolio . . . 7733
15.3.10.3.7.5 Sharpe optimal weights portfolio . . . 7733
15.3.10.3.7.6 Maximum diversification optimal weights portfolio . . . 7734

15.3.10.3.7.7 Treynor-Black optimal weights portfolio 7735

15.3.10.3.7.8 Risk Parity Portfolio 7738

15.3.10.3.8 Capital Asset Pricing Model (CAPM) 7740

15.3.10.3.8.1 Hedging a future and commodity portfolio with the CAPM 7757

15.3.10.3.9 Black & Scholes option pricing models 7759

15.3.10.3.9.1 Put-Call parity equation 7760

15.3.10.3.9.2 Efficient Market Hypothesis 7764

15.3.10.3.9.3 Wiener Process 7765

15.3.10.3.9.4 Generalized Brownian motion 7770

15.3.10.3.9.5 Brownian bridge 7775

15.3.10.3.9.6 Itō's process 7779

15.3.10.3.9.7 Itō's product and quotient rule 7793

15.3.10.3.9.8 Black & Scholes differential equation 7794

15.3.10.3.9.9 Self-financing portfolio on underlying 7798

15.3.10.3.9.10 Greeks and others... 7799

15.3.10.3.9.11 Solving the Black & Scholes differential equation 7804

15.3.10.3.10 Binomial option pricing model (CRR model) 7827

15.3.10.3.11 Bachelier option pricing model 7842

15.3.10.3.12 Black Model (Future options pricing model) 7845

15.3.10.3.13 VIX volatility index 7847

15.3.10.3.14 Value at Risk 7858

15.3.10.3.14.1 Relative Value at Risk 7859

15.3.10.3.14.2 Absolute Value at Risk 7864

15.3.10.3.14.3 Delta-Normal Value at Risk 7865

15.3.10.3.14.4 (Sorted) Historical Value at Risk 7866

15.3.10.3.14.5 Credit Value at Risk (CVaR) 7867

15.3.10.3.14.6 Operational Value at Risk 7869

15.3.10.3.14.7 Surplus Value at Risk (Pension VaR) 7871

15.3.10.3.14.8 Variance-Covariance Value at Risk 7872

15.3.10.3.14.9 Marginal and Component Value at Risk 7874

15.3.10.3.14.10 Conditional Value at Risk (CVaR) 7876

15.3.10.3.14.11 Modified duration Value At Risk (MdVaR) 7878

15.3.10.3.14.12 Backtesting Value at Risk 7879

15.3.11 Exchange rate 7880

15.3.12 Credit default Risk 7884

15.3.12.1 Credit spread risk 7885

15.3.12.2 Value At Risk equity coverage rating 7887

15.3.12.3 Altman Z-score 7889

15.3.12.4 Merton's Distance to Default and KMV Expected Default Frequency 7890

15.3.13 Time Series Analysis 7896

15.3.13.1 Type of Errors 7903

15.3.13.2 Decompositions 7907

15.3.13.3 Types of Forecasting Models 7914

15.3.13.3.1 Simple Moving Average (moving average smoothing) 7919

15.3.13.3.2 Linear Model With Seasonal Coefficients (LMSC) 7922

15.3.13.3.3 Exponential Smoothing models . . . 7927
15.3.13.3.3.1 Simple Exponential Smoothing (EWMA) . . . 7928
15.3.13.3.3.2 Double Exponential Smoothing with One Parameter (Brown Method) . . . 7937
15.3.13.3.3.3 Holt's Double Exponential Smoothing with 2 Parameters (Additive Method) . . . 7945
15.3.13.3.3.4 Holt's and Winter Triple Exponential Smoothing with 3 Parameters (Multiplicative Method) . . . 7951
15.3.13.3.3.5 ETS (Error, Trend, Seasonality) Models . . . 7958
15.3.13.3.4 Logistic Model . . . 7965
15.3.13.4 Autoregressive Models . . . 7972
15.3.13.4.1 AR(p) Autoregressive processes . . . 7983
15.3.13.4.2 MA(q) Moving Average stationary process . . . 7986
15.3.13.4.3 ARMA(p, q) Autoregressive non-seasonal moving average processes . . . 7988
15.3.13.4.4 ARIMA(p, d, q) Autoregressive non-seasonal integrated moving average processes . . . 7989
15.3.13.4.5 ARCH-GARCH processes . . . 7999
15.3.13.4.5.1 ARCH (Autoregressive Conditional Heteroskedasticity) . . . 8000
15.3.13.4.5.2 GARCH (Generalized Autoregressive Conditional Heteroskedasticity) . . . 8004
15.3.13.5 Forecasting Stationary Processes . . . 8005
15.3.13.5.1 Linear Minimal Mean Square Estimation (LMMSE) . . . 8010
15.3.13.6 TBATS . . . 8012
15.3.13.7 Recipes for modelling time series data . . . 8013
15.3.13.7.1 Durbin-Watson autocorrelation test . . . 8015
15.4 Quantitative Management . . . 8021
15.4.1 Corporate and Government Finance Management . . . 8024
15.4.1.1 Basic Accounting Equation . . . 8026
15.4.1.2 Ratio Analysis . . . 8027
15.4.1.2.1 Short term solvency or liquidity measure . . . 8029
15.4.1.2.2 Long term solvency or Liquidity measure . . . 8030
15.4.1.2.3 Profitability Measures . . . 8031
15.4.1.2.4 Growth Rate . . . 8032
15.4.1.2.5 Asset management or turnovers measures . . . 8033
15.4.1.2.6 Market Value Measures . . . 8033
15.4.1.2.7 Data Envelopment Analysis . . . 8034
15.4.1.3 Weighted Average Cost of Capital (WACC) . . . 8044
15.4.1.4 Break-even Point Analysis (BEPA) . . . 8047
15.4.1.5 Investment Strategies . . . 8049
15.4.1.5.1 Net Present Value . . . 8050
15.4.1.5.2 Internal Rate of Return . . . 8057
15.4.1.5.3 Payback . . . 8059
15.4.1.6 Company Valuation Methods . . . 8060
15.4.1.6.1 Balance sheets-based method . . . 8062
15.4.1.6.1.1 Book Value . . . 8062
15.4.1.6.1.2 Adjusted Book Value . . . 8063
15.4.1.6.1.3 Liquidation Value . . . 8064
15.4.1.6.1.4 Substantial Value . . . 8064
15.4.1.6.1.5 Book Value and Market Value . . . 8065
15.4.1.6.2 Income Statemented-Based Methods . . . 8065

15.4.1.6.2.1 Value of Earnings with PER 8066
15.4.1.6.2.2 Value of Dividends 8067
15.4.1.6.2.3 Sales Multiples 8069
15.4.1.6.3 Goodwill-Based Methods 8069
15.4.1.7 Capital Goods 8070
15.4.1.7.1 Linear Amortization 8070
15.4.1.7.2 Arithmetic Declining Amortization 8071
15.4.1.7.3 Geometric Declining Amortization (declining balance) 8072
15.4.2 Project Management 8074
15.4.2.1 Probabilistic PERT (Beta-PERT Distribution) 8074
15.4.2.2 Project planning variance reduction 8083
15.4.2.3 Process Reliability 8086
15.4.2.4 Divide and rule? 8088
15.4.3 Lean Management (Six Sigma Process) 8092
15.4.3.1 Pareto Analysis 8095
15.4.3.1.1 Gini Index 8099
15.4.3.2 Weighted Ishikawa Diagram 8102
15.4.4 Supply Chain Management 8106
15.4.4.1 Supply Chain Management in uncertain future 8108
15.4.4.1.1 Safety Stock formula 8110
15.4.4.2 Optimal initial stock management with zero rotation 8112
15.4.4.3 Wilson's Models 8119
15.4.4.3.1 Wilson's model with resupply 8124
15.4.4.3.2 Wilson's model without resupply 8133
15.4.4.3.3 Wilson's model with resupply and break-up 8136
15.4.5 Queueing Theory 8141
15.4.5.1 $M/M/\ldots$ arrival times modelisation 8145
15.4.5.2 $M/M/\ldots$ service times modelisation 8150
15.4.5.3 Kendall queues notation 8152
15.4.5.4 Modelling of arrivals and departures $M/M/1$ 8154
15.4.5.5 Probability of standby in a $M/M/k/k$ queue (Erlang-B formula) 8162
15.4.5.6 Probability $M/M/K/+\infty$ of standby (Erlang-C formula) 8165
15.4.6 Insurance 8170
15.4.6.1 Premium pricing 8172
15.4.6.2 Taking account of the return on experience 8175
15.4.6.3 Discounting factor of a retirement insurance 8177
15.4.6.4 Pension insurance (life annuities) 8180
15.4.6.4.1 Temporary life annuity 8181
15.4.6.4.2 Deferred life annuity 8183
15.4.7 Sensitivity Analysis 8184
15.4.7.1 Direct Bias Method 8186
15.4.7.2 Correlation Method 8190
15.5 Music Mathematics (physics of hearing) 8195
15.5.1 Longitudinal Sound Waves 8195
15.5.1.1 Power carried by a sound wave 8203

15.5.1.2 Measuring the intensity of a sound . . . 8207
15.5.2 Spherical Sound Waves . . . 8210
15.5.3 Doppler effect . . . 8212
15.5.3.1 Fixed source-Moving observer . . . 8214
15.5.3.2 Moving source-Fixed observer . . . 8215
15.5.3.3 Moving source and observer . . . 8217
15.5.4 Shock waves . . . 8218
15.5.5 Music Scales . . . 8220
15.5.6 Harmonic Oscillator . . . 8224
15.5.6.1 Damping and Resonance . . . 8227

16 Engineering 8234

16.1 Marine & Weather Engineering . . . 8239
16.1.1 Visual horizon . . . 8244
16.1.1.1 Discerning the curvature of the Earth . . . 8257
16.1.2 Wind direction . . . 8261
16.1.3 Atmospheric Profile Models . . . 8264
16.1.3.1 Atmospheric Exponential Profile Model . . . 8264
16.1.3.2 Adiabatic Atmosphere Model . . . 8268
16.1.3.2.1 Hypsometric equation . . . 8271
16.1.4 Earth’s equilibrium temperature . . . 8272
16.1.4.1 Greenhouse effect . . . 8275
16.1.4.2 Milankovitch cycles . . . 8276
16.1.5 Temperature variation range models . . . 8278
16.1.5.1 Diurnal temperature range (DTR) . . . 8278
16.1.5.2 Seasonal temperature range (STR) . . . 8281
16.1.6 Weather (sounding) balloon . . . 8288
16.1.7 Cyclogenesis and Anticyclogenesis . . . 8291
16.1.8 Tides . . . 8301
16.1.8.1 Astronomical tides (influence Moon/Sun) . . . 8303
16.1.8.1.1 First approach (naive center of mass model) . . . 8304
16.1.8.1.2 Second approach (vector field model) . . . 8308
16.1.8.1.2.1 Tides hight . . . 8316
16.1.9 Lorenz equations . . . 8318
16.1.9.1 Rayleigh-Bénard convection cells (Benard-Marangoni instability) . . . 8328
16.1.9.2 Lorenz attractor and chaos . . . 8337
16.1.10 Waves . . . 8344
16.1.10.1 Depth of a wave . . . 8356
16.1.10.2 Wave’s amplitude . . . 8357
16.2 Mechanical Engineering . . . 8359
16.2.1 Mechanical Efficiency . . . 8360
16.2.2 Gears . . . 8362
16.2.2.1 Transmission ratios . . . 8365
16.2.2.2 Gears association . . . 8368
16.2.2.2.1 Odd/Even Gear "problem" . . . 8373

16.2.2.3 Types of Gears 8374
16.2.3 Strength of materials 8382
16.2.3.1 Quadratic moments 8385
16.2.3.2 Equation of the elastic line 8388
16.2.3.2.1 Euler-Bernoulli Beam equation 8394
16.2.3.2.2 Potential elastic energy 8403
16.2.3.3 Torsion 8404
16.2.3.4 Buckling 8408
16.2.3.4.1 Self-buckling 8415
16.2.3.5 Traction 8425
16.2.3.6 Pressure Vessels 8426
16.2.3.6.1 Thin cylinders subjected to internal pressure 8427
16.2.3.6.1.1 Hope stress σ_θ 8428
16.2.3.6.1.2 Longitudinal stress σ_L 8429
16.2.3.6.2 Thin sphere subjected to internal pressure 8431
16.2.3.7 Centrifugal expansion of a rotating ring 8434
16.3 Electrical Engineering 8437
16.3.1 Elementary primitive electrical symbols 8441
16.3.2 Alternative current VS Direct current 8444
16.3.2.1 DC generator (dynamo) 8447
16.3.2.2 Average power 8451
16.3.3 Transformers 8455
16.3.3.1 Transformer universal EMF equation 8461
16.3.4 Steady State Linear Circuits 8463
16.3.4.1 RC series circuit 8467
16.3.4.1.1 RC series step response 8467
16.3.4.1.2 RC series linear response 8471
16.3.4.2 RL series circuit 8472
16.3.4.3 Source-Free serial RLC circuit 8475
16.3.4.3.1 Critically damped response ($R^4 = 4L/C$) 8477
16.3.4.3.2 Overdamped (hypercritic) response ($R^2 > 4L/C$) 8478
16.3.4.3.3 Underdamped (decaying oscillation) response ($R^2 < 4L/C$) 8480
16.3.4.4 Telegrapher's Equations 8486
16.3.4.4.1 Thomson's law of squares of transmission lines (latency metallic cables) 8490
16.3.5 Linear Circuit (RLC) in Forced Regime 8497
16.3.5.1 Low-pass filter 8501
16.3.5.2 High-pass filter 8502
16.3.5.3 Integrator and differentiator 8504
16.3.6 Signal Processing 8506
16.3.6.1 Nyquist-Shannon Theorem (Nyquist sampling theorem) 8506
16.3.7 Semiconductor devices 8519
16.3.7.1 Diodes 8519
16.3.7.1.1 Ideal Diode Equation 8522
16.3.7.1.1.1 Carrier concentrations in pure semiconductors 8522
16.3.7.1.1.2 Carrier concentrations in doped crystals 8524

16.3.7.1.1.3 pn-junctions 8526
16.3.7.1.1.4 Forward-biased pn-junction 8529
16.3.7.2 Amplifiers 8536
16.4 Civil Engineering 8539
16.4.1 Static 8541
16.4.2 Pulleys 8542
16.4.2.1 Windlass 8551
16.4.3 Cornu spiral 8553
16.4.4 Overhead cable 8558
16.4.4.1 Free overhead cable (catenary) 8559
16.4.4.2 Charges overhead cable (suspended bridge) 8567
16.4.4.3 Very tense cable 8571
16.4.5 Falling chimney (naive approach) 8573
16.4.6 Dams 8579
16.4.6.1 Power extracted from a Dam 8582
16.4.7 Tuned Mass Damper 8587
16.5 Aerospace Engineering 8590
16.5.1 Airfoil Lift 8591
16.5.1.1 Newton's lift argument (skipping stone argument) 8593
16.5.1.2 Bernoulli's lift argument (equal time argument) 8594
16.5.1.3 Euler's lift argument 8596
16.5.1.4 Coandă lift argument 8600
16.5.1.5 Kutta-Joukowski lift argument 8601
16.5.2 Cosmic speeds 8603
16.5.3 Fundamental Equation of Propulsion (Tsiolkovsky rocket equation) 8606
16.5.4 Geostationary orbit 8615
16.5.5 Vis-Viva equation 8619
16.5.6 Gravity assist 8622
16.5.6.1 Speed variation (in solar reference frame) 8623
16.5.6.2 Approaching speed from infinity 8626
16.5.6.3 Velocity variation and impact factor 8627
16.5.7 Hohmann Transfer orbit 8633
16.6 Software Engineering 8637
16.6.1 Algorithm 8637
16.6.2 Dichotomic Search algorithm 8639
16.6.2.1 Bisection algorithm 8639
16.6.2.2 Binary search algorithm 8642
16.6.3 Tower of Hanoi algorithm 8644
16.6.4 Sorting algorithms 8650
16.6.4.1 Bubble sort 8650
16.6.4.2 QuickSort algorithm 8652
16.6.5 Dijkstra's algorithm 8654
16.6.6 Google PageRank algorithm 8662
16.6.6.1 Weighted Count 8664
16.6.6.2 Recursive counting 8665

16.6.6.3 Absorbing states . . . 8669
16.7 Industrial Engineering . . . 8673
16.7.1 Six Sigma . . . 8674
16.7.1.1 Quality Control . . . 8677
16.7.1.2 Defaults/Errors . . . 8678
16.7.1.3 Capability Indices . . . 8683
16.7.1.4 Quality Levels . . . 8697
16.7.2 Taguchi Model . . . 8710
16.7.3 Preventive Maintenance . . . 8716
16.7.3.1 Planned Obsolescence . . . 8717
16.7.3.2 Reliability Empirical Estimators . . . 8720
16.7.3.2.1 Average Failure Rate . . . 8734
16.7.3.3 Weibull Distribution . . . 8736
16.7.3.3.1 Two-parameter Weibull distribution linearisation . . . 8741
16.7.3.4 Topology of Systems . . . 8743
16.7.3.4.1 Fault Tree Analysis . . . 8755
16.7.3.4.2 Markov Chain Reliability Model . . . 8757
16.7.3.5 Maximum Likelihood for failure rate determination of samples . . . 8761
16.7.3.6 Kaplan-Meier Survival Rate . . . 8763
16.7.3.7 ABC Method . . . 8772
16.7.3.8 XYZ Method . . . 8777
16.7.4 Design of Experiments (DoE) . . . 8780
16.7.4.1 Two levels factorial Designs . . . 8794
16.7.4.1.1 Replicated Full Factorial Designs . . . 8800
16.7.4.1.2 Plackett-Burman Designs . . . 8807
16.7.4.1.3 Fractional Factorial Designs . . . 8813
16.7.4.2 General factorial Designs . . . 8823
16.7.4.3 Taguchi Designs and Nomenclature (robust designs) . . . 8837
16.7.4.3.1 Signal Factor . . . 8846
16.7.4.4 Response Surface Methodology (Box Domains) . . . 8848
16.7.4.4.1 Pure quadratic curvature test . . . 8850
16.7.4.4.2 Box-Wilson Central Composite Designs . . . 8853
16.7.4.4.2.1 Circumscribed Center Designs . . . 8856
16.7.4.4.2.2 Face Centered Designs . . . 8861
16.7.4.5 Optimal Designs . . . 8865
16.7.4.6 Mixture Design . . . 8874
16.7.4.6.1 Network mixture designs (simplex lattice designs) . . . 8878
16.7.4.6.2 Full Factorial Combined with Mixture Design-Crossed Design . . . 8893
16.7.4.7 General DoE diagnostic tools . . . 8896
16.7.4.7.1 Lenth's PSE Pareto Margin Error for unreplicated factorial designs . . . 8896
16.7.4.7.2 Pareto Margin Error for replicated factorial designs . . . 8899
16.7.4.7.3 Desirability . . . 8900
16.7.5 Quality Control on Reception (Lot Acceptance Sampling Plans) . . . 8904
16.7.5.1 Simple acceptance sampling plan by measurement for a unique tolerance with known standard deviation . . . 8908

16.7.5.1.1 Calculation of the parameters using the norm AF-X06-023 8915
16.7.5.2 Simple acceptance sampling plan by attribute 8918
16.7.5.2.1 Calculation of the parameters using the norm ISO 2859-1 8924
16.7.5.3 Double acceptance sampling plan by attribute 8926
16.7.5.4 Operating characteristic curve (OC) 8927
16.7.5.5 Average outgoing quality (AOQ) 8930
16.7.6 Quality Control Charts (CC) 8934
16.7.6.1 WECO's empirical rules 8938
16.7.6.2 Sample size and Sampling frequency for Control Charts 8940
16.7.6.3 Attributes Control Charts (qualitative CC) 8944
16.7.6.3.1 P Control Charts (binomial proportion CC) 8945
16.7.6.3.2 NP Control Charts (binomial counting CC) 8949
16.7.6.3.3 C Control Charts (Poisson counting CC) 8953
16.7.6.3.4 U Control Charts (normalized Poisson) 8957
16.7.6.3.5 Laney's p' and u' control charts 8961
16.7.6.4 Measurement Control Charts (quantitative CC) 8963
16.7.6.4.1 Individual measurement control chart with required limits 8964
16.7.6.4.2 Individual measurement control chart with moving limits 8966
16.7.6.4.3 Subgroups measurement control chart with standard error 8972
16.7.6.4.4 $\overline{S}-S$ Subgroups measurement control chart for standard deviation 8977
16.7.6.4.5 $\overline{X}-S$ Subgroups measurement control chart 8986
16.7.6.4.6 $\overline{X}-S_P$ Subgroups measurement control chart with pooled variance 8991
16.7.6.4.7 $\overline{R}-R$ Subgroups measurement control chart 8996
16.7.6.4.8 $\overline{X}-R$ Subgroups measurement control chart 9001
16.7.6.5 Autocorrelated Measurement Control Charts (time weighted control charts) 9007
16.7.6.5.1 $\mathrm{I}-\mathrm{MR}/\overline{X}$ Individual moving range measurement control chart 9007
16.7.6.5.2 $\mathrm{I}-\mathrm{MR}/\overline{\mathrm{MR}}$ Individual moving range measurement control chart 9011
16.7.6.5.3 Individual Moving Average control chart 9014
16.7.6.5.4 CUSUM (cumulated sum) control chart with empirical V-mask 9019
16.7.6.5.5 EWMA control charts (exponential weighted moving average) with fixed limits 9026
16.7.6.6 Rare events control charts 9034
16.7.6.6.1 Frequency T control chart with probabilistic limits 9035
16.7.6.6.2 Frequency G control chart of rare events 9040
16.7.6.7 Control Charts Operating Characteristic (OC) Curves 9045
16.7.6.7.1 OC for $\bar{X}$ measurement control charts 9045
16.7.6.7.2 OC for P-type attribute control charts 9047
16.7.7 Design of reliability tests 9049
16.7.7.1 Chi-squared time of test 9050
16.7.7.2 Binomial sampling size 9053
16.7.7.3 Test to demonstrate MTBF (or MTTF) 9057
16.7.7.4 Beta-binomial sampling size 9059

17 Epilogue **9062**

18 Biographies **9064**

A . . . 9065
B . . . 9067
C . . . 9075
D . . . 9084
E . . . 9088
F . . . 9091
G . . . 9096
H . . . 9101
I . . . 9109
J . . . 9109
K . . . 9112
L . . . 9115
M . . . 9124
N . . . 9132
O . . . 9136
P . . . 9138
R . . . 9143
S . . . 9144
T . . . 9152
V . . . 9155
W . . . 9156
Y . . . 9160
Z . . . 9162

19 Chronology 9166

20 Humor 9211

20.1 Situations . . . 9212
20.2 Mathematics . . . 9227
20.3 Physics . . . 9249
20.4 Statistics . . . 9273
20.5 Chemistry . . . 9280
20.6 Engineering . . . 9287
20.7 Computing . . . 9297
20.8 Social Sciences . . . 9308

21 Links 9315

21.1 Exact Sciences . . . 9316
21.2 Publishing/Magazines . . . 9317
21.3 Associations . . . 9319
21.4 Jobs . . . 9320
21.5 Television/Radio . . . 9320
21.6 Other Sciences . . . 9322
21.7 Softwares/Applications . . . 9324

22 Quotes 9330

23 Changelog 9347

24 Nomenclature **9372**

List of Figures **9379**

List of Tables **9430**

List of Algorithms **9440**

Bibliography **9443**

Index **9478**

25 Donate **9556**

Dedicated to all people interested
in enlightened knowledge rather
than blind faith!

2

Warnings

Contents

2.1 Impressum **4**

2.1.1 Use of content 4

2.1.2 How to use this textbook 5

2.1.2.1 Style guide 6

2.1.2.2 Credits 7

2.1.2.3 Videos and Animations 7

2.1.2.4 Ancillaries 8

2.1.3 Data Protection 10

2.1.4 Use of data 10

2.1.5 Data transmission 10

2.1.6 Agreement 10

2.1.7 Errata 10

2.2 License **12**

2.2.1 Preamble 12

2.2.2 Applicability and Definitions 13

2.2.3 Verbatim Copying 14

2.2.4 Copying in Quantity 14

2.2.5 Modifications 15

2.2.6 Combining Documents 16

2.2.7 Collections of Documents 17

2.2.8 Aggregation with independent Works 17

2.2.9 Generating this document 17

2.2.10 Translation 18

2.2.11 Permissions and Copyrights 18

2.2.12 Termination 18

2.2.13 Future revisions of this License 18
2.3 Safety Guides **19**
2.4 Roadmap (To Do) **21**

2.1. Impressum

2.1.1 Use of content

The contents of this textbook are elaborated by a development process by which volunteers reach a consensus. This process that brings together volunteers, research also the point of view of people interested in the topics of this textbook. The person in charge of this textbook administers the process and establishes rules to promote fairness in the consensus approach. It is also responsible for drafting the text, sometime for testing/evaluating or independently verifying the accuracy or completeness of the presented information.

We decline no responsibility for any injury, damage or any other kind, special, incidental, consequential or compensatory, arising from the publication, application or reliance on the content of this textbook. We make no express or implied warranty on the accuracy or completeness of any information published in this textbook, and do not guarantee that the information contained in this textbook meet any specific need or goal of the reader. We do not guarantee the performance of products or services of one manufacturer or vendor solely by virtue of this textbook content.

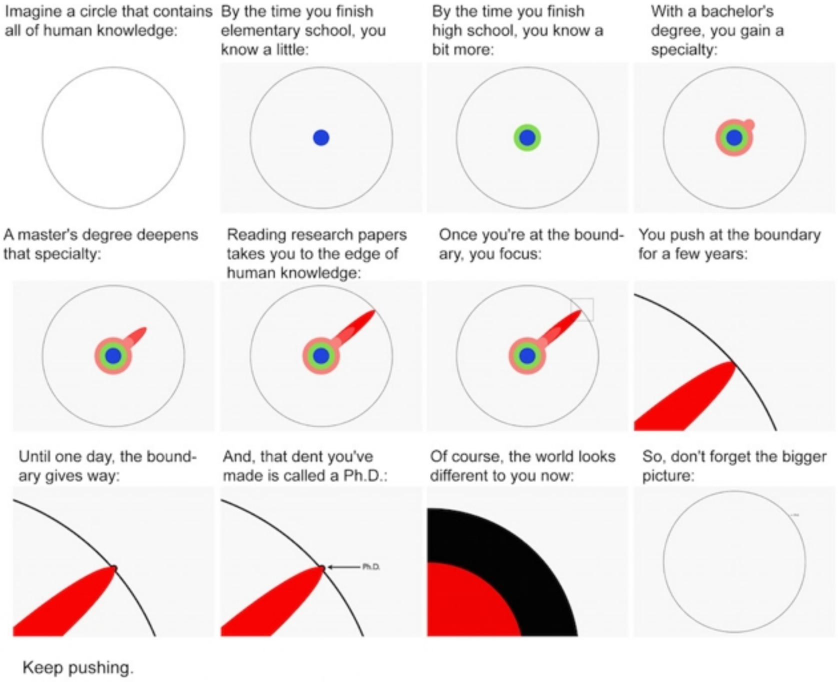
The technical descriptions, procedures, and computer programs in this textbook have been developed without care, therefore they are provide without warranty of any kind. We make also no warranties that the equations, programs, and procedures in this textbook or its associated software are free of error, or are consistent with any particular standard of merchantability, or will meet your requirements for any particular application. They should not be relied upon for solving a problem whose incorrect solution could result in injury to a person or loss of property. Any use of the content of this textbook as at the reader's own risk. The authors, redactors, and publisher disclaim all liability for direct, incidental, or consequent damages resulting form use of the content of this textbook or the associated software.

By publishing texts, it is not the intention of this textbook to provide services on behalf of any person or entity or performing any task to be accomplished by any person or entity for the benefit of a third party. Anyone using this textbook should rely on its own independent judgement or, where that is appropriate, seek the advice of a qualified expert to determine how to exercise reasonable care under all circumstances. The information and standards on the topics covered by this textbook may be available from other sources that the reader may wish to visit in search of points of view or additional information not covered by the contents of this textbook.

We have no power in order to enforce compliance with the contents of this textbook, and we do not undertake to monitor or enforce such compliance. We have no certification, testing or inspection activity of products, designs or installations for safety or health of persons and property. Any certification or other statement of compliance regarding information relating to health or safety of persons and property, mentioned in this textbook, cannot possibly be attributed to the content of this textbook and remains under the responsibility of the certification center or the concerned reporter.

2.1.2 How to use this textbook

At the university level, this textbook can be used for a PhD, graduate level or advanced undergraduate level seminar in various exact and pure fields of science. In reality this textbook also aims to cover the full Kindergarten to PhD curriculum.


Imagine a circle that contains all of human knowledge:
By the time you finish elementary school, you know a little:
By the time you finish high school, you know a bit more:
With a bachelor's degree, you gain a specialty:
A master's degree deepens that specialty:
Reading research papers takes you to the edge of human knowledge:
Once you're at the boundary, you focus:
You push at the boundary for a few years:
Until one day, the boundary gives way:
And, that dent you've made is called a Ph.D.:
Ph.D.
Of course, the world looks different to you now:
So, don't forget the bigger picture:
Keep pushing.

⟨ Berger | Dillon ⟩
@InertialObservr

before getting a PhD i felt dumb.. & at the end of my PhD i still feel dumb, but like in a more advanced way

1:53 AM · Aug 12, 2021

243 Retweets **74** Quotes **3,495** Likes **43** Bookmarks

Figure 2.1: **Education and the bigger picture**

Because the methods of Applied Mathematics are learned by practice and experience, we view a seminar on Applied Mathematics as a learning-by-doing (project oriented) seminar. We structure our mathematical modelling seminars around a set of problems that require the trainee to construct models that help with planning and decision making. The imperative is that the models should be consistent with the theory and back-tested. To fulfil this imperative, it is necessary for the trainee to combine mathematical theory with modelling. The result is that the trainee learns the theory, and more importantly, learns how that theory is applied and combined in the real world. The ability to criticize and identify limitations of dangerous mathematical tools is the most valuable feature of our seminars.

quote 2.1:5.#2

> *Mathematics is not a deductive science - that's a cliché. When you try to prove a theorem, you don't just list the hypotheses, and then start to reason. What you do is trial and error, experimentation, guesswork. You want to find out what the facts are, and what you do is in that respect similar to what a laboratory technician does!*
>
> Paul Halmos

Remark

Mathematicians have always practised experimental mathematics. Existing records of early mathematics, such as Babylonian mathematics, typically consist of lists of numerical examples illustrating algebraic identities. However, modern mathematics, beginning in the 117th century (according to holocene calendar), developed a tradition of publishing results in a final, formal and abstract presentation. The numerical examples that may have led a mathematician to originally formulate a general theorem were not published, and were generally forgotten.

The problems with solutions in this textbook provide the opportunity to apply the text material to a comprehensive set of fairly realistic situations. By the end of the seminars the trainees will have enhanced their skills and knowledge of the most important theoretical and computing tools. These are valuable skills that are in demand by the businesses at the highest levels.

It is very difficult to cover all the material in this textbook in a semester. It takes a lot of time to explain the concepts to the trainees. The reader is encouraged to pick and choose which topics will be covered during the term. It is not necessary strictly necessary to cover them in sequence but it can help in a significant way?

In a nutshell, this textbook offers you a wide variety of topics that are amenable to modelling. All are practical.

2.1.2.1 Style guide

We use a few standard conventions throughout this textbook. They were prepared obviously with LaTeX which automatically numbers sections and the hyperref package provides links within the PDF for the Table of Contents as well as other references made within the body of the text.

We use color and some boxes to set apart some points for convenient reference. In particular:

- Clickable links and cross-references[1] are most of time in magenta (only a few special cases are in blue)
- Texts expliciting the target of cross-references are in light grey

---

[1] According to the Cambridge dictionary and to the scientific consensus a "cross-reference" is a note in a book or document that tells you to look somewhere else in the book or document for more information.

- New terms (most of time associated with definitions) are in red
- Remarks are in greyed out boxes
- Examples are in black bordered boxes
- Proofs start with the word "proof" and are concluded with a ■
- Supplementary examples[2] start with the symbol ✍

Doubtless, we have failed to live up to our legalism somewhere. If the reader keeps a list of these transgressions to give it to us at the end of the reading of this textbook would be very likely worthwhile for the redactors...

Caution!!! Any interpretation of the technical sections (i.e. of the mathematical proofs) in this textbook and related explanations in plain text has to be avoid (if we could we would even say that it is vorbidden!). The context is self-sufficient and therefore no words have other meaning, or implicit meaning, than the academic definition(s) that makes consensus at the time this textbook is published! All numbers are written in the decimal system and no prefix of the latter are implicitly hidden! There are also no hidden message in this textbook! The reader should focus only on the mathematical calculations and predictions rather than worrying about the underlying physical interpretations or philosophical implications!

2.1.2.2 Credits

The redactors have attempted (because the lack of answer is often the rule rather than the exception...) to credit and obtain the necessary authorizations for the use of any material protected by copyright, such as photographs, animations, videos and illustrations from third-party sources.

If, however, copyright infringement is unintentionally committed, we encourage copyright owners to contact us so that we can take appropriate action.

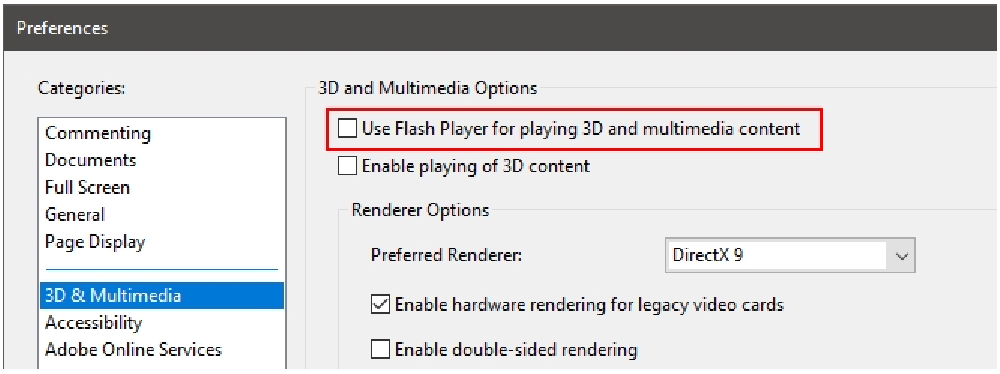
2.1.2.3 Videos and Animations

This textbook contains a dozens of small videos and animations integrated inside the PDF pages themselves. However to be able to play them you will need:

1. Adobe Flash Player for Firefox - NPAPI (the latest version if this technology still exists when you read these lines...)
2. To activate the following option available in most Adobe PDF readers on Microsoft Windows:

---

[2] A companion example to a theorem doesn't have any specific symbol



Preferences

Categories:

- Commenting
- Documents
- Full Screen
- General
- Page Display
- 3D & Multimedia
- Accessibility
- Adobe Online Services

3D and Multimedia Options

☐ Use Flash Player for playing 3D and multimedia content

☐ Enable playing of 3D content

Renderer Options

Preferred Renderer: DirectX 9

☑ Enable hardware rendering for legacy video cards

☐ Enable double-sided rendering

This option needs to be reactivate each time you close and reopen the reader depending on your IT policy.

2.1.2.4 Ancillaries

We offer an array of ancillaries for students, instructors and practitioners.

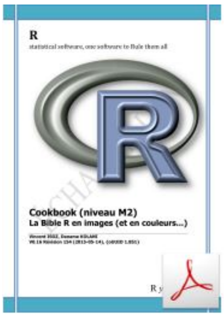
First there are some free companion eBooks and tools in French and English for the people who want to put in practice the theories and models presented in this textbook (see a working and interactive example on a screen may help for the understanding).

Here is the list:

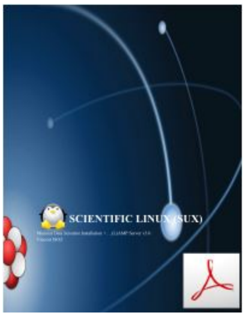
- MATLAB® in English (1,400 pages):
  https://archive.org/details/Matlab
- Maple™ in French (129 pages):
  https://archive.org/details/maple
- R in French (3,134 pages):
  https://archive.org/details/R
- Minitab in French (1,290 pages):
  https://archive.org/details/Minitab
- Scientific Linux installation & Configuration (330 pages):
  https://archive.org/details/scientific-linux-installation

MATLAB
Vincent ISOZ
Cookbook (M2)
Alphorm.com

Waterloo Maple


R
Cookbook (niveau M2)
La Bible R en images (et en couleurs...)

MINITAB (v15-v17)
Exercices de base (niveau PhD)
SPC

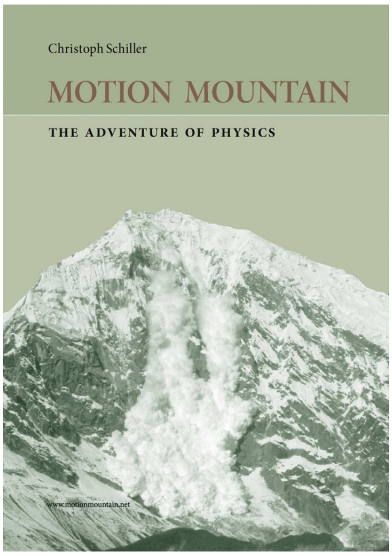

SCIENTIFIC LINUX (SUX)

And as any technical textbook should have a forum, the reader can go through this link for any discussions about the content of this textbook: `https://www.physicsforums.com`.

For people who want to help us to translate this textbook in other languages you can contact us through our social networks.

Because this textbook mainly focus on mathematical aspect of physical phenomena we can only strongly recommend to the reader another free textbook that is in our point of view actually the best one that focus on the popular science aspect of the subjects that we will cover:

Motion Mountain by Dr. Christoph Schiller: `http://www.motionmountain.net`


Christoph Schiller
MOTION MOUNTAIN
THE ADVENTURE OF PHYSICS
www.motionmountain.net

2.1.3 Data Protection

When looking at information on the Internet companion website, some data are automatically saved during a month after which detailed data are destroyed. We try to save as less as possible data and as brief as possible. Wherever we can, we ave only anonymous data. We undertake to process the data you send us personally with the utmost diligence. You can object at any time in the publication of your data by contacting us.

2.1.4 Use of data

Your data are only used for sending the a newsletter. Communication of personal data (except the e-mail address, title and name) is optional. When registering for the newsletter, you can of course specify an alternate address and/or a fictitious name.

2.1.5 Data transmission

We will never sell or commercialize the data of our customers or interested parties and will never affects the rights of the person. In addition, we will not rent mailing lists and will not send you advertising from third parties or on our behalf.

2.1.6 Agreement

When you provide us personal information, you authorize us to save them and use them within the meaning of the Swiss Federal Law on Data Protection. If you ask us not to send you emails, we are obliged, in your interest, save your e-mail in an internal negative list.

2.1.7 Errata

The redactor have made every effort to ensure the accuracy and completeness of the information contained in this textbook. However, we assume no responsibility for any errors, omissions, or differences of opinion that may be present. Readers are strongly advised to consult other scientific peer review sources with high impact factor to supplement and update their knowledge on the topics covered in this textbook.

Illustrations, photographs, and diagrams included in this textbook are for illustrative purposes only and are not intended to be exact representations of the objects or phenomena described. The editors and artists cannot be held responsible for any errors or inaccuracies that these visual elements may contain since they are voluntary for pedagogical reasons!

Although we have taken every care to ensure the accuracy of our content, mistakes do happen. If you find a mistake in this textbook - maybe a mistake in the text, scripts or illustrations - we would be grateful if you would report this to us. By doing so, you can save other readers from frustration and help us improve subsequent versions of this textbook. Our e-mail is given on the footer every page of this textbook. Once your errata are verified, your submissions will be accepted and the error will be visible on the changelog (see page 9347) of updated versions.

2.2. License

The entire contents of this textbook is subject to the GNU Free Documentation License, which means:

- that everyone has the right to freely use the texts for non-commercial usage (Google Ads or any equivalent being considered as a commercial usage!)
- that any person is authorized to broadcast items for non-commercial usage (Google Ads or any equivalent being considered as a commercial usage!)
- that anyone can freely edit the texts for non-commercial usage (Google Ads or any equivalent being considered as a commercial usage!)

and bla bla bla...

in accordance with the license described below:

https://creativecommons.org/licenses/by-nc-sa/4.0/

Version 1.1, March 12000 (holocene calendar)

Copyright (C) 12000 (holocene calendar) Free Software Foundation, Inc. 59 Temple Place, Suite 330, Boston, MA 02111-1307 USA Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

2.2.1 Preamble

The purpose of this License is to make a manual, textbook, or other written document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it only a non-commercial purpose[3]. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

---

[3] The immediate benefit of this is that you can generate the textbook, with brand new problem sets, and distribute it to your students simply as a PDF (in an email, for instance). But more generally, if you are not keen on the way we explained (or failed to explain) something, then you are free to rewrite it. If you would like to cover more (or less) material, then you are free to add (or delete) whatever Chapters/Sections/Paragraphs that you wish. And since we can provide the source code on request, the reader do not need to reinvent the wheel.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed textbook. We recommend this License principally for works whose purpose is instruction or reference.

2.2.2 Applicability and Definitions

This License applies to any manual or other work that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. The "Document", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "you".

A "Modified Version" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "Secondary Section" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject (for example, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics). The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "Invariant Sections" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License.

The "Cover Texts" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License.

A "Transparent" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, whose contents can be viewed and edited directly and straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup has been designed to thwart or discourage subsequent modification by readers is not Transparent. A copy that is not "Transparent" is named "Opaque".

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML designed for human modification. Opaque formats include

PostScript, PDF, proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML produced by some word processors for output purposes only.

The "Title Page" means, for a printed textbook, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

2.2.3 Verbatim Copying

You may copy and distribute the Document in any medium, non-commercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

2.2.4 Copying in Quantity

If you publish printed copies of the Document numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a publicly-accessible computer-network location containing a complete Transparent copy of the Document, free of added material, which the general network-using public has access to download anonymously at no charge using public-standard network protocols. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

2.2.5 Modifications

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

- Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the Changelog section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.
- List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has less than five).
- State on the Title page the name of the publisher of the Modified Version, as the publisher.
- Preserve all the copyright notices of the Document.
- Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.
- Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.
- Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.
- Include an unaltered copy of this License.
- Preserve the section entitled "Changelog", and its title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section entitled "Changelog" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.
- Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "Changelog" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.

- In any section entitled "Acknowledgements" or "Dedications", preserve the section's title, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.
- Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.
- Delete any section entitled "Endorsements". Such a section may not be included in the Modified Version.
- Do not retitle any existing section as "Endorsements" or to conflict in title with any Invariant Section.
- If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.
- You may add a section entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties–for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.
- You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.
- The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

2.2.6 Combining Documents

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections entitled "Changelog" in the various original documents, forming one section entitled "Changelog"; likewise combine any sections entitled "Acknowledgements", and any sections entitled "Dedications". You must delete all sections entitled "Endorsements."

2.2.7 Collections of Documents

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

2.2.8 Aggregation with independent Works

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, does not as a whole count as a Modified Version of the Document, provided no compilation copyright is claimed for the compilation. Such a compilation is named an "aggregate", and this License does not apply to the other self-contained works thus compiled with the Document, on account of their being thus compiled, if they are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one quarter of the entire aggregate, the Document's Cover Texts may be placed on covers that surround only the Document within the aggregate. Otherwise they must appear on covers around the whole aggregate.

2.2.9 Generating this document

You will need four things to generate this document for yourself:

1. Install MiKTeX
2. Install TeXMaker
3. An internet connection
4. Configure MikTeX as indicated in the remarks at the beginning of the *LaTeX_SciencesCh_EN.tex* file

2.2.10 Translation

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of the corresponding section about transformation. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License provided that you also include the original English version of this License. In case of a disagreement between the translation and the original English version of this License, the original English version will prevail.

2.2.11 Permissions and Copyrights

Many figures and tables in this textbook have been reproduced from copyrighted sources. Permission from the publishers and authors for the use of the material is gratefully acknowledged. Some of the sources are directly identified in captions and footnotes, while many others are cited by alphanumeric references or specified as unknown (despite some effort to find them).

2.2.12 Termination

You may not copy, modify, sublicense, or distribute the Document except as expressly provided for under this License. Any other attempt to copy, modify, sublicense or distribute the Document is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

2.2.13 Future revisions of this License

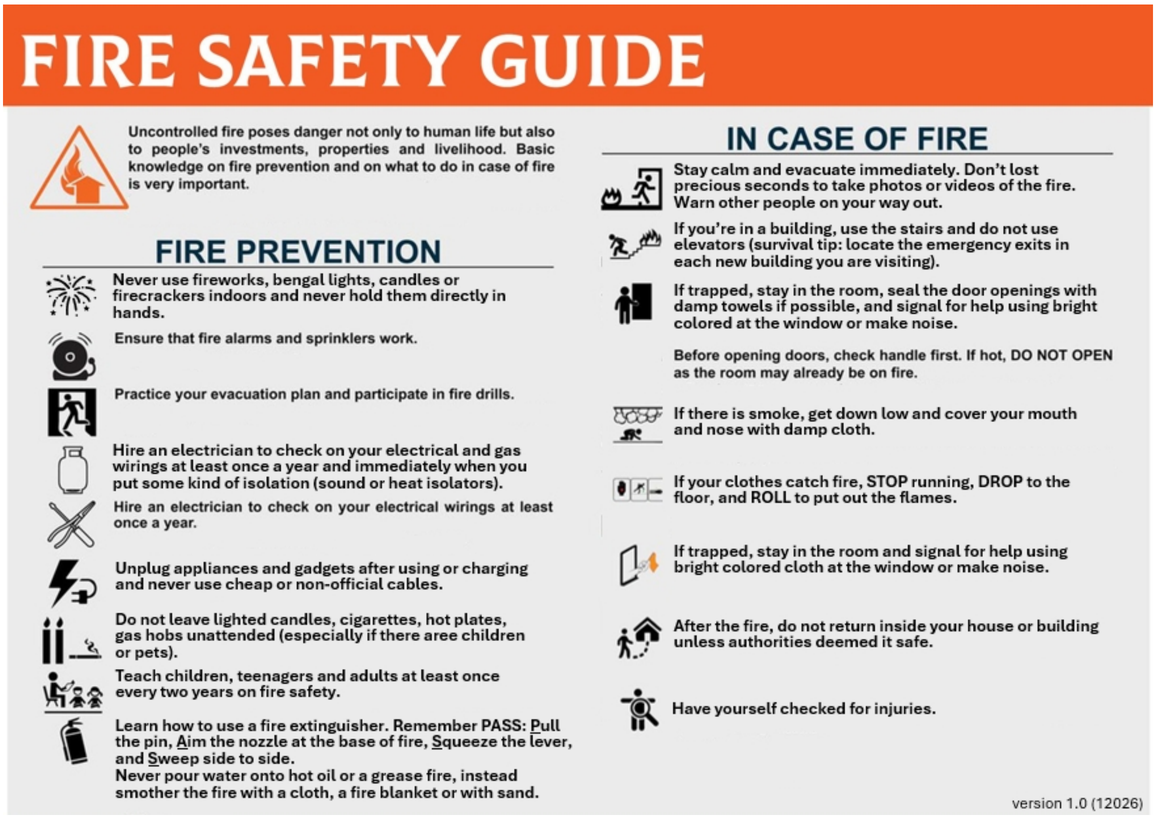
The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See `http://www.gnu.org/copyleft`.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation.

Please consider the environment before printing

2.3. Safety Guides

Revisiting with fire, chemistry and electrical safety rules every year is essential for physics, chemistry, nuclear and engineering teachers because the subject constantly brings students into contact with heat sources, electrical equipment, components, and experiments that can escalate quickly without disciplined habits; from kindergarten to graduate studies, learners evolve in maturity and technical skill, yet the fundamental risks remain the same, and regular reminders cultivate a culture of vigilance, responsible behavior, and scientific respect for hazardous phenomena. This yearly ritual that should be mandatory starting from kindergarten and displayed in every textbook and on the doors of every lab reinforces not only practical safety but also the deeper message that scientific exploration is inseparable from ethical care for oneself and others, a lesson that grows in importance as students advance toward more complex laboratory work and interacts with complex environments while becoming adults.



FIRE SAFETY GUIDE

Uncontrolled fire poses danger not only to human life but also to people's investments, properties and livelihood. Basic knowledge on fire prevention and on what to do in case of fire is very important.

FIRE PREVENTION

Never use fireworks, bengal lights, candles or firecrackers indoors and never hold them directly in hands.

Ensure that fire alarms and sprinklers work.

Practice your evacuation plan and participate in fire drills.

Hire an electrician to check on your electrical and gas wirings at least once a year and immediately when you put some kind of isolation (sound or heat isolators).

Hire an electrician to check on your electrical wirings at least once a year.

Unplug appliances and gadgets after using or charging and never use cheap or non-official cables.

Do not leave lighted candles, cigarettes, hot plates, gas hobs unattended (especially if there aree children or pets).

Teach children, teenagers and adults at least once every two years on fire safety.

Learn how to use a fire extinguisher. Remember PASS: Pull the pin, Aim the nozzle at the base of fire, Squeeze the lever, and Sweep side to side.
Never pour water onto hot oil or a grease fire, instead smother the fire with a cloth, a fire blanket or with sand.

IN CASE OF FIRE

Stay calm and evacuate immediately. Don't lost precious seconds to take photos or videos of the fire. Warn other people on your way out.

If you're in a building, use the stairs and do not use elevators (survival tip: locate the emergency exits in each new building you are visiting).

If trapped, stay in the room, seal the door openings with damp towels if possible, and signal for help using bright colored at the window or make noise.

Before opening doors, check handle first. If hot, DO NOT OPEN as the room may already be on fire.

If there is smoke, get down low and cover your mouth and nose with damp cloth.

If your clothes catch fire, STOP running, DROP to the floor, and ROLL to put out the flames.

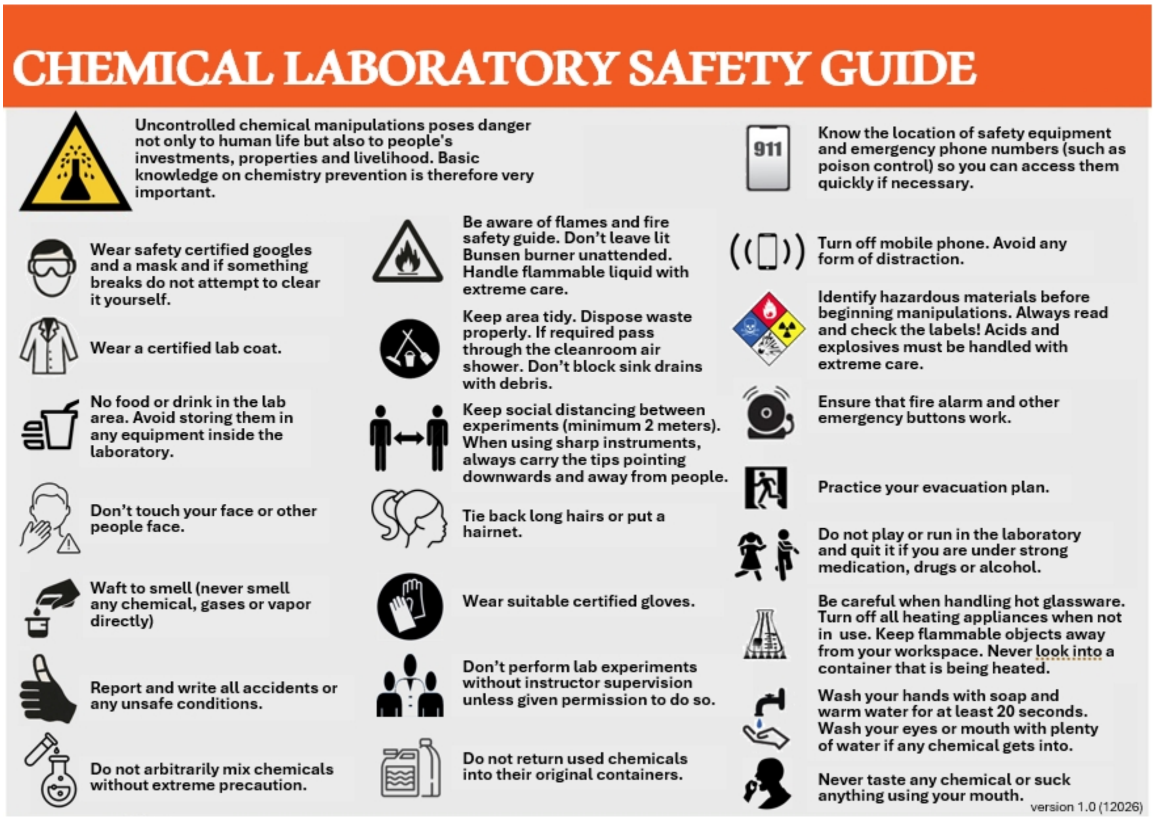
If trapped, stay in the room and signal for help using bright colored cloth at the window or make noise.

After the fire, do not return inside your house or building unless authorities deemed it safe.

Have yourself checked for injuries.

version 1.0 (12026)

Figure 2.2: Fire safety guide



CHEMICAL LABORATORY SAFETY GUIDE

Uncontrolled chemical manipulations poses danger not only to human life but also to people's investments, properties and livelihood. Basic knowledge on chemistry prevention is therefore very important.

Wear safety certified googles and a mask and if something breaks do not attempt to clear it yourself.

Wear a certified lab coat.

No food or drink in the lab area. Avoid storing them in any equipment inside the laboratory.

Don't touch your face or other people face.

Waft to smell (never smell any chemical, gases or vapor directly)

Report and write all accidents or any unsafe conditions.

Do not arbitrarily mix chemicals without extreme precaution.

Be aware of flames and fire safety guide. Don't leave lit Bunsen burner unattended. Handle flammable liquid with extreme care.

Keep area tidy. Dispose waste properly. If required pass through the cleanroom air shower. Don't block sink drains with debris.

Keep social distancing between experiments (minimum 2 meters). When using sharp instruments, always carry the tips pointing downwards and away from people.

Tie back long hairs or put a hairnet.

Wear suitable certified gloves.

Don't perform lab experiments without instructor supervision unless given permission to do so.

Do not return used chemicals into their original containers.

Know the location of safety equipment and emergency phone numbers (such as poison control) so you can access them quickly if necessary.

Turn off mobile phone. Avoid any form of distraction.

Identify hazardous materials before beginning manipulations. Always read and check the labels! Acids and explosives must be handled with extreme care.

Ensure that fire alarm and other emergency buttons work.

Practice your evacuation plan.

Do not play or run in the laboratory and quit it if you are under strong medication, drugs or alcohol.

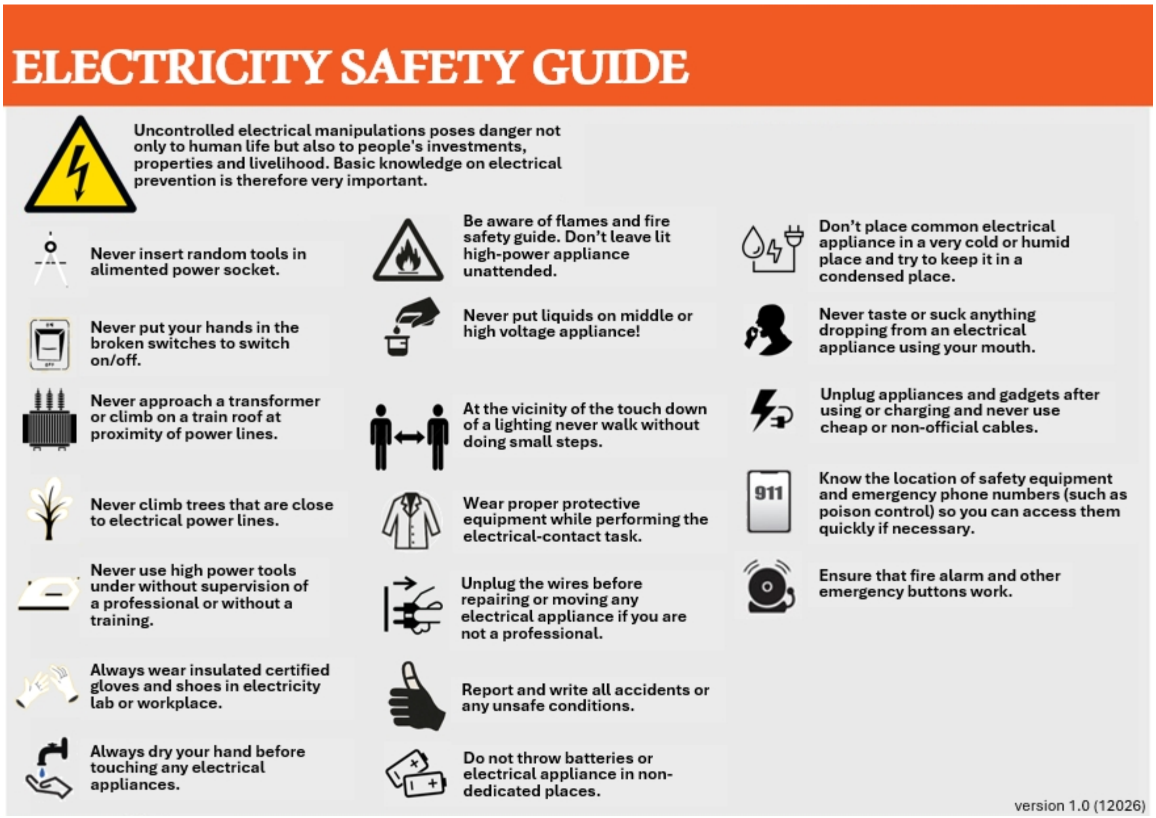
Be careful when handling hot glassware. Turn off all heating appliances when not in use. Keep flammable objects away from your workspace. Never look into a container that is being heated.

Wash your hands with soap and warm water for at least 20 seconds. Wash your eyes or mouth with plenty of water if any chemical gets into.

Never taste any chemical or suck anything using your mouth.

version 1.0 (12026)

Figure 2.3: **Chemical laboratory safety guide**



ELECTRICITY SAFETY GUIDE

Uncontrolled electrical manipulations poses danger not only to human life but also to people's investments, properties and livelihood. Basic knowledge on electrical prevention is therefore very important.

Never insert random tools in alimented power socket.

Never put your hands in the broken switches to switch on/off.

Never approach a transformer or climb on a train roof at proximity of power lines.

Never climb trees that are close to electrical power lines.

Never use high power tools under without supervision of a professional or without a training.

Always wear insulated certified gloves and shoes in electricity lab or workplace.

Always dry your hand before touching any electrical appliances.

Be aware of flames and fire safety guide. Don't leave lit high-power appliance unattended.

Never put liquids on middle or high voltage appliance!

At the vicinity of the touch down of a lighting never walk without doing small steps.

Wear proper protective equipment while performing the electrical-contact task.

Unplug the wires before repairing or moving any electrical appliance if you are not a professional.

Report and write all accidents or any unsafe conditions.

Do not throw batteries or electrical appliance in non-dedicated places.

Don't place common electrical appliance in a very cold or humid place and try to keep it in a condensed place.

Never taste or suck anything dropping from an electrical appliance using your mouth.

Unplug appliances and gadgets after using or charging and never use cheap or non-official cables.

Know the location of safety equipment and emergency phone numbers (such as poison control) so you can access them quickly if necessary.

Ensure that fire alarm and other emergency buttons work.

version 1.0 (12026)

Figure 2.4: **Electrical safety guide**

2.4. Roadmap (To Do)

This textbook has a simple progression rule that is: 1 new A4 page per day since May 12001 (holocene calendar) on subjects that interest the supervisor of the *Original* distribution of the textbook *Opera Magistris*. The following subjects below are already planned - of course any external help is welcome! - for a near or far future still with the same level of details and pedagogical approach in the proofs as the rest of the textbook (all the subjects below should take almost $1,500$ to $2,500$ pages and are already in writing progress):

- Introduction:
    - Add acknowledgements to all people that who have developed MikTeX/LaTeX distro and packages used for this textbook
- Proof Theory:
    - Modal logic
    - Probabilistic logic
    - Statistical abduction
- Probabilities:
    - Explaining conditional plots and spinplots
    - Hidden Markov Chains
    - Fuzzy probability
- Statistics:
    - Mode and Median of statistical laws
    - Bayesian rating average
    - Derivation of standard error of the β of Cox proportional hazard model
    - Likelihood of censored data
    - Normal Law Entropy
    - Propension score
    - Standardized Pearson residuals
    - ANCOVA
    - Wald-Wolfowitz test (binary sequence)
    - Levene-Wolfwitz test[4] (continuous up/down sequence)
    - Ellipse of control

---

[4] also named "turning point test" or "trend test"

  - Ansari-Bradley test
  - Homogeneity of variances tests (Levene, Bartlett, Fligner, Brown-Forsythe)
  - Regular Dickey-Fuller test
  - Augmented Dickey-Fuller test
  - Canonical Correlation
  - G-test of periodicity
  - Gaussian and Student copula
  - Extreme Values Theory
  - Pickands-Balkema-De Haan theorem
  - Schoenfeld's global test for Cox proportional hazards model
  - Survey Theory
  - Canonical Analysis
  - Tucker's Inter-battery factor analysis
  - PLS regression (partial least squares)
  - Two-Stage Least Squares (2SLS) regression
  - Logic regression
  - Adjusted Chi-squared
  - Schoenfeld residuals
  - Jarque-Bera test
  - Dixon, Grubbs, Chi-square outliers tests
  - Somers'd rank coefficient
  - Kaplan-Meier sample size[5]
  - Weighted PCA
  - Bayesian ANOVA
  - Bayesian one sample T-test
  - Bayesian binomial test
  - Bayesian multinomial test
  - Bayesian Mann-Withney test
  - Bayesian correlation
  - Bayesian contingency tables

- Geometry:
  - Gauss remarkable theorem (Theorema Egregium)

- Differential Calculus
  - Lebesgue Integral with numerical application in MATLAB®

---

[5] Sample size calculation: Survival analysis (logrank test)

  - Heun function

- Analysis:
  - Finish the discrete Fourier transformation
  - Finish to introduce the unilateral Laplace transformation
  - Finish to introduce the Hilbert transformation
  - Fourier inequality
  - Add a section with the detailed proof that the euclidean, manhattan, chebychev, minkowski, cosine, pearson, mahalanobis, haversine, jaccard, family-tree, jensen-shannon, chi-square, canberra, spearman, hamming and levenstein distances respect the four axioms of a metric
- Complex Analysis:
  - Residue Theorem for polynomial ratios
- Differential Geometry:
  - Finish to introduce the isometric perspective
  - Normal coordinates
  - Gauss curvature
  - Circumference on a plane, sphere and hyperbolic surface
- Principia:
  - Charge conservation according to Noether theorem (global gauge symmetry)
- Mechanics:
  - Flow through submerged orifice
  - Intermediate axis theorem
  - Flow over notches and weirs
  - Force due to the flow around a pipe bend
  - Force on a pipe nozzle
  - Impact of a jet on plane
  - Force due to a jet hitting an inclined plane
  - Fluid Hammer Effect
  - Kelvin-Helmholtz instability
  - Saint-Venant equations
  - Kutta-Joukowski lift
  - Layer Boundary (Blasius boundary layer)
- Statistical Mechanics:
  - Maxwell–Jüttner distribution

- Electrodynamics:
    - Electromagnetic Fields of a rotating shell of charge
    - Electromagnet rotor torque
    - Corneal Topography (Pentacam system)
    - Sun glitter
- Electrokinetics:
    - Photoelectric converter
    - Norton and Thevenin theorem
- Astronomy:
    - Venus transit method for the distance of the Sun
    - MacCullagh's formula
    - Body flatness indirect calculation
    - Synchronous locking of tidally evolving satellites
    - Flatness spheroid
    - Oberth Effect
    - Clairaut's theorem
    - Planetary departure
    - Nutation-precession calculation method of planets
- Astrophysics:
    - Worm holes
    - Tolman-Oppenheimer-Volkoff limit
    - Hawking's Radiation
    - Black Hole life-time (based on Hawking's radiation)
- General Relativity:
    - Real volume of an object in General Relativity
    - Einstein radius derivation
    - General Birkhoff's theorem
    - Global Positioning System (GPS)
    - Kerr metric and Kerr solution (+ergosphere)
    - Naked singularities
    - Lense–Thirring precession
- Cosmology:
    - Kaluza-Klein theory

  - Inflation theory
  - Borde-Guth-Vilenkin theorem (BGV)
  - Derivation of Wheeler-DeWitt metric
  - Cosmic inflation theory (i.e. Big Bang) Nucleosynthesis
  - Inflationary space-time is past-incomplete
- Nuclear Physics:
  - Critical mass
  - Neutron transport
  - Bateman equation
  - Compton scattering cross section classical derivation
- Wave Quantum Physics:
  - Bell inequalities
  - Lamb shift detailed calculation
  - Davisson-Germer experiment
  - EPR paradox formalism
  - Hydrogen fine structure derivation
  - Multiphoton ionization
  - Entanglement
- Relativistic Quantum Physics:
  - Photon "spin" (helicity) and polarity relation
  - Parity, charge conjugation and time reversal for electroweak theory
  - Free Dirac equation covariance proof
- Quantum Field Theory:
  - Second quantization
  - Brout-Englert-Higgs mechanism
  - Feynman rules and Feynman diagrams
  - Dyson series
  - Light-Light scattering
  - Wigner-Weisskopf theory of spontaneous decay
- Quantum Chemistry:
  - Valence shell electron pair repulsion theory
  - Density Functional Theory (DFT)
  - Hund's rules

- Numerical Methods:
  - Relational algebra
  - Monte Carlo Markov chains
  - Univariate optimization problem with substitution method
  - Acceptation/Rejection Sampling
  - Gibbs Sampling
  - Cronbach coherence indicator
  - Quadratic discriminant Analysis (QDA)
  - Multiple Correspondence Analysis
  - Linear Mixture Model (LMM)
  - Generalized Least Squares (GLS)
  - Correspondence Factor Analysis
  - Generalized Reduced Gradient (GRG) optimization method
  - Normalized mutual information
  - Chi-squared Automatic Interaction Detector (CHAID)
  - Latent semantic analysis (LSA)
  - Correlation based feature selection
  - PRISM, AQ, CN2 and C4.5 algorithms
  - Procrustes analysis
  - Independent component analysis (ICA)
  - Uplift modelling
  - Differences between natural, basis and restricted cubic splines
  - Gauss-Jordan method
  - Gauss-Seidel method
  - Learning Vector Quantization (LVQ)
  - Karush-Kuhn-Tucker conditions
  - Latin Hypercube Sampling (LHS)
  - Universal Approximation Theorem
- Quantum Computing:
  - No-cloning theorem
- Cryptography:
  - Elliptic curves
- Engineering:
  - Box Domains
  - Definitive Screening Design

  - Split-plots designs
  - Central Composite Design
  - Center Face Cube Design
  - Modelization by structural equations
  - Accelerated life testing
  - Microelectronics (npn/pnp junctions, amplifiers)
  - Stress of a rotating disc and rotating ring
  - Shelf life for Stability Study
  - Trochoidal waves (i.e. Gerstner waves)
  - Atmospheric tides
  - Soliton wave (Korteweg de Vries equation)

- Game Theory:
  - Coalition
  - Shapley value
  - Arrow's impossibility theorem

- Economy:
  - Continuous yield rate (Nielson-Siegel)
  - Equivalence of an obligation rate for a treasure bond
  - Options Portfolio hedging
    - Protective Put/Call
    - Bull Spread/Call
    - Bear Spread/Call
    - Butterfly
    - Straddle
    - Strangle
    - Collar
    - Box spreads
    - Calendar spreads
  - Jarrow-Rudd, Tian and Leisen-Reimer binomial models
  - Greeks for binomial trees
  - Swaps
  - Black-Litterman portfolio
  - Margrabe's formula
  - Kirk's (approximation) formula
  - CreditRisk+
  - Fokker-Planck equation

  - Vector autoregressive models for multivariate time series
  - Granger test for times series causation
  - Event Study Analysis
  - Kalman Filter
  - Heston option pricing model
  - Spread option pricing
  - Cornish-Fisher VaR
  - Close to close, Parkinson, Garman-Klass, Rogers-Satcheli, Garman-Klass Yang-Zhang, Yang-Zhang volatility estimates
  - Asian geometric and arithmetic average option pricing
  - Digital options
  - Dynamic stochastic general equilibrium
  - Wold's theorem
  - Changepoint detection analysis
  - SABR model closed form approximate solution
  - RAROC, RORAC, RARORAC
  - Causal Inference (instrumental variables, panel data, difference in difference, regression discontinuity, potential outcome modeling)

- Quantitative Management:
  - Gale-Shapley Algorithm
  - Pareto/NBD, BG/NBD and BG/BB customer lifetime value models
  - Bullwhip Effect
  - Condorcet paradox
  - Computerized Relative Allocation of Facilities Technique (CRAFT)
  - Differed Capital in living case (life assurance)
  - CBOE's Volatility Index (started but not finished)
  - Death differed temporary (life assurance)
  - Ponzi system and pyramid scheme

- Population Dynamics:
  - Neutral theory of molecular evolution

- Biographies:
  - On the request of a reader, add for each biography the religion

Furthermore we want to compile a summarized version of this PDF (without the detailed developments but only with the main results) and a companion presentation including slides (typically made with Beamer) only with the main results, texts and pictures/plots that could be usable by any teacher for free around the World.

Remember that the LaTeX sources of this textbook (with the equations and TikZ/Pgfplots codes!) can be obtained actually depending on your donation on Patreon, Paypal, Tipeee or if you participate to the translation of the textbook in another language.

A copy of the PDF file is available at the Internet Archive (540 Megabytes PDF...):

INTERNET ARCHIVE

https://archive.org/details/@scientific_solutions

To cite this textbook:

```
@book{OperaMagistris12024v4r2,
title = {Opera Magistris - Elements of Applied Mathematics for Engineers},
year = {12024},
keywords = {science, physics, maths, engineering, finance, management},
isbn = {9782399909327},
DOI={10.5281/zenodo.5816819}
}
```

3

Acknowledgements

The ideas and evidence compiled in this textbook have been developed and reinforced by many people during the twenty years of its redaction. I have greatly benefited from my regular interactions with students (university, engineering schools) and executives from all backgrounds, including CEOs, CFOs, PMs of many companies around the world, teaching sessions, developing company-specific programs, consulting, and even informal conversations. I am grateful to them for sharing their wisdom and experience with me and inspiring many of the ideas in the book. I am also grateful to all scientists, economists, engineers, illustrators and science popularizers of the last millennial without who this textbook would not exist in it's actual shape. This includes the following approximately $1,500$ people:

John Cannin, Karl Brandan Mollweide, Howard Haber, Juan F. G. Cascales, Forhad Ahmed, Xavier Hubaut, Sanjay Sekaran, William George Horner, Adrian Olszweski, Anders Jensen, Wojciech Kotłowski, Saed Sayad, Federico Errica, Shi Yan, Karl Pearson, Louis Agassiz, Maria Gaetana Agnesi, Al-Battani, Christrophe Michel, Abu Nasr Al-Farabi, Alhazen, Jim Al-Khalili, Muhammad Ibn Musa al-Khwarizmi, Mihailo Petrovic Alas, Angel Alcala, Vesto Slipher, Salim Ali, Luis Alvarez, Andre Marie Ampère, Anaximander, Carl David Anderson, Mary Anning, Milner Baily Schaefer, Charles Geschke, Virginia Apgar, Archimedes of Syracuse, Agnes Arber, Aristarchus of Samos, Aristotle, Svante Arrhenius, Oswald Avery, Amedeo Avogadro, Waloddi Weibull, David Brewster, Avicenna, Charles Babbage, Francis Thomas Bacon, Alexander Bain, T. Johnson, John Logie Baird, Pierre Barbier, Joseph Banks, Ramon Barba, John Bardeen, Fjodor van Veen, Donald E. Knuth, Paul Gingrich, Leonard Henry Caleb Tippett, Benoît de Maillet, Charles Barkla, Stichting Tegenbeld, Pablo Carlos Budassi, Ibn Battuta, William Bayliss, Colin Maclaurin, Neyman Jerzy, George Beadle, Arnold Orville Beckman, Henri Becquerel, Georges Geiser, Gilles Delpin, Steven Weinberg, R. Ellis, Richard Fitzpatrick, Alain Vienne, Emil Adolf Behring, Jean le Rond Alembert, Alexander Graham Bell, Alonso Marcelo, Larry Sorensen, Emily Lakdawalla, Carl R. Rod Nave, Jérôme Cardan, Jasper Habicht, Ole Witt-Hansen, J. Finn Edward, Lawrence Schwartz, Yves Ayant, Aleksander Gwiazad, Elie Belorizky, Prasanta Chandra Mahalanobis, Aco Z. Muradjan, Jean-Marie Duhamel, Louis Bachelier, Wihelm Weber, Bass Issa, Thomas Pyzdek, Jean-Pierre Durandeau, Giulio Fagnano, Jean-Claude Thiebaud, Emil Gumbel, Martin Knudsen, William de Sitter, Simon H. Hess, Robert Papanicola, Roger Bautrant, Paul Bramand, Régis Martegoutes, Joël Sornette, Luque Manuel, Thomas Söll, Alexandra Elbakyan, Ignacio Icke, Christophe de la Chapelle, Roger Sahun, Olivier Berruyer, Janosh Riebesell, Geroges Thomassier, Daniel Schwen, Chris Fach, Walmes Zeviani, Hubert Reeves, Etienne Klein, Michael S. Fuhrer, Shaffique Adam, Dexter Chua, Emmanuel Rosencher, Gilles Triscone, Paul Keller, Serge Mehl, Christophe Afonso, Riadh Ben Hamouda, William Ogilvy Kermack, Bernard Schaeffer, John J. Uicker, Friedrich Prym, L. Aron Nelson, Ralph Apher, Frank Wang, Anthony Quintano, Joseph Raphson, Charles Pache, G.R. Pennock, Eric Rolph, Maurice Pillet, G. I. Taylor, Joseph Edward Shigley, Lucien Borel, Sebastien Bossu, Felix Klein, Philippe Henrotte, Jean-Louis Boursin, Martin Bircher, Andrew Z. Colvin, Bui Duc Nha, Nicomachus of Gerase, William Forsyth Sharpe, Frank Bowman, Sébastien Tanzilli, Bernard Cagnac, Jean-Clean Pebay-Peyroula, Apollonius of Tyana, Saung Tadashi, Robert P. Kirshner, Vittoria Rezzonico, Max Roser, C. Hebeisen, Martin Kutta, Nikolay Joukowski, Michel Tenenhaus, Gareth Jones, Renzo Cairoli, Hermann Weyl, Robin Plackett, J.P. Burman, Jan Hladik, Pierre-Emmanuel Hladik, Leon Walras, George E. P. Box, K.B. Wilson, Jawed Karim, Chard Hurley, Steve Chen, Johannes Diderik Van Der Waals, Robert Cobbaut, Pascual Jordan, François-Éric Racicot, Raymond Théoret, Nikolay Yakovlevich Sonin, Sam Denby, Sandip Banerjee, Vilfredo Pareto, Genichi Taguchi, David Parker, Christian Schirm, François Viète, Jean-Baptiste Biot, Richard Baraniuk, Thomas Bauer, Thierry Roncalli, Jimmy Roussel, Olivier Cogis, Claudine Robert, Howard Curtis, Zoran Popovic, B. D. Popovic, Matthew Schwartz, Thomas Bayes, Alexander M. Mood, Juergen Giesen, John. D. Cutnell, Felix Bernstein, Oldrich Alfons Vasicek, Stephen Kealhofer, Charles C. Holt, Peter Winters, James Durbin, Corrado Gini, Kaoru Ishikawa, Henry Laurence Gantt, Kenneth W. Johnson, Harry Nyquist, Alastair L. Day, Paul Langevin, Walter Hohmann, Edsger W. Dijkstra, Ferdinand Georg Frobenius, Jean-Marie De Koninck, Armel Mercier, Jean-Loup Delcroix, Gregorio Ricci-Curbastro, Félix Savart, Abraham Bers, Bipin R. Desai, René Didier, Brook Taylor, Pierre Grecias, Bernard Diu, Claudine Guthmann, Teller Edward, Bernard Roulet, Aqeel Anwar, Danielle Lederer, Jacques Douchet, Bruno Zwahlen, Dean G. Duff, Georges Duvaut, Roger Penrose, Edgard Elbaz, Mario Novello, Christian Gentili, Vincent Giard, Carl D. Murray, Myron Scholes, Edward Witten, Stanley F. Dermott, Jimmy Wales, Larry Sanger, Larry Page, Sergey Brin, Jean-Marc Huguenin, Peter Atkins, Claude Elwood Shannon, Ronald S. Friedman, Normand M. Laurendeau, Viktor Witkovsky, Pieter Zeeman, Martin Thoma, Kevin Binz, Dave Jones, Stephan Kulla, Joseph G. Eisenhauer, Stefan Banach, Denis Boigelot, Peter Hoyng, Abu Yusuf Ya'qub Ibn Al-Kindi, Salman Khan, Jeff Atwood, Joel Spolsky, Jason Condon, AMS Lawrence, Frank Morgan, M. Leemis, Jacquelyn T. McQueston, Niels Henrik Abel, Randall Munroe, Scott M. Ransom, Sam Derbyshire, Johann Jakob Balmer, Hugues Vermeiren, Nicholas L. Squeglia, Pierre Alphonse Laurent, Robert Phillips, Gustav Robert Kirchhoff, Knut Hinkelmann, Camille Jordan, Oscar Kempthorne, Ferdinand von Lindemann, Edward Lucas, Eric D. Feigelson, G. Jogesh Babu, Charles-Emile Picard, George Boole, Rob J. Hyndman, Edmond Laguerre, George Athanasopoulos, Ray D'Inverno, Hans Stephani, Lewis Ryder, George B. Arfken, Hans J. Weber, Frank E. Harris, Arthur Asquith Rayner, Kristopher Tapp, Abu Ali al-Husayn ibn Abd Allah Ibn Sina, James P. Howard, E. Joseph Billo, Tobias Holzmann, Josh Neufeld, Abu'l-Hasan al-Uqlidisi, John D. Lawson, Charles E. McCulloch, Shayle R. Searle, Wessel N. van Wieringen, Leucippus of Miletus, Oscar Levin, Robert Vincent Hogg, Tanis Elliot, Bessel Friedrich, Dale Zimmerman, Sepp Hochreiter, Jurgen Schmidhuber, Kurt Gieck, Jack Gieck, K. Kanjii Gopal, Jacques Alexis, Jacques Goupy, Ernst Hairer, Gerhard Wanner, Robert Resnick, David Halliday, Paul Richard, Eugene Hecht, Russell Hibbeler, Jean Hladik, John C. Hull, Jacques Istas, Luan Jaupi, Reza N. Jazar, Richard Arnold Johnson, Dean W. Wichern, Philippe Jorion, Toon Koppelaars, Israel Koren, C. Mani Krishna, Amitabha Lahiri, Palash Baran Pal, Frederic Laroche, J.-Ch. Corbaz, Daniel Goetschi, Malcolm Ludvigsen, Marc Lachieze-Rey, Patrick Lyonnet, Franz Mandl, Henry Mathieu, Autolycus of Pitane, Herve Fanet, Douglas C. Montgomery, Chee-Hock Ng, Soong Boon-Hee, Nikolai Semenovich Piskunov, Richard Wallace Robinett, Francois Rothen, R. Paul Singh, Dennis R. Heldman, James Stewart, Daniel Thiel, Herve Thiriez, William S.C. Williams, Menaech-mus, Paul Wilmott, Reza Zafarani, Mohammad Ali Abbasi , Huan Liu, Wilford Zdunkowski, Andreas Bott, Barton Zwiebach, Christian Gary D., Partha Dasgupta, Kevin Schug, Tilo Arens, Frank Hettlich, Christian Karpfinger, Stephen A. Ross, Randolph Westerfield, Bradford Jordan, Carlos Oliveira, Abu al'Fath Khazini,Abu al'Fath KhaziniAbu al'Fath Khazini Jonathan Cryer, Kung-Sik Chan, Ivo Welch, Mike Zabrocki, William H. Press, Saul. A. Teukolsky, William T. Vetterling, Michel Vermot, Philippe Bovay, Damien Prongue, Sebastien Dorder, Pascal Winkler, J.W. Dettman, D.W. Thormas, C.S. Barrett, C.W. LeSage, William H. Louisell, Ferdinand Filip, Gilbert Strang, Edwin Herman, Lev Davidovich Landau, Evgeny Lifshitz, Emile Berliner, Claude Bernard, Timothy John Berners-Lee, Kolani Daname, Daniel Bernoulli, Jacques Bernoulli, Bernoulli Jean, Jacob Berzelius, Henry Bessemer, Hans Bethe, Homi Jehangir Bhabha, Alfred Binet, Clarence Birdseye, Kristian Birkeland, James Black, Bhagabati Sutra, Elizabeth Blackwell, Alfred Blalock, Katharine Burr Blodgett, Franz Boas, David Bohm, Aage Bohr, Niels Henrik David Bohr, Ludwig Boltzmann, Max Born, Carl Bosch, Robert Bosch, Jagadish Chandra Bose, Satyendra Nath Bose, Walther Wilhelm Georg Bothe, Robert Boyle, William Lawrence Bragg, Tycho Brahe, Brahmagupta, Hennig Brand, Georg Brandt, Karl Schwarzschild, Thucydides, Wernher Von Braun, J Harlen Bretz, Louis de Broglie, Alexander Brongniart, Robert Goodell Brown, Michael E. Brown, Lester R. Brown, Eduard Buchner, Linda Buck, William Buckland, Georges-Louis Leclerc, Mathias Legrand, Comte de Buffon, Sulbar Sutras of Manava, Robert Wilhelm Bunsen, Luther Burbank, Jocelyn Bell Burnell, Macfarlane Burnet, Thomas Burnet, Benjamin Cabrera, Santiago Ramon y Cajal, Rachel Carson, Rudolf Bauer, George Washington Carver, Benoit Mandelbrot, Henry Cavendish, Anders Celsius, James Chadwick, Subrahmanyan Chandrasekhar, Erwin

Chargaff, Noam Chomsky, Âryabhat, Steven Chu, Leland Clark, John Cockcroft, Arthur Compton, Nicolaus Copernicus, Gerty Theresa Cori, Vincent Isoz, Charles-Augustin de Coulomb, Jacques Cousteau, Brian Cox, Francis Crick, James Croll, Emile Borel, Nicholas Culpeper, Marie Curie, Pierre Curie, Georges Cuvier, Adalbert Czerny, Stephen Boyd, Lieven Vandenberghe, Dmitry S. Gorbunov, Valery Rubakov, Rob J. Hyndman, Yanan Fan, Karl Theodor Wilhelm Weierstrass, Mozi, Ruey S. Tsay, Alireza Navabi, Apastamba Dharmasutra, Nicki Hise, Jose D. Perezgonzalez, Yutaka Sasaki, Supriyo Bandyopadhyay, Marc Cahay, Frederick K. Lutgens, Edward J. Tarbuck, Albert E. Beaton, John Wilder Tukey, Michael Kutner, Piotr Korbel, Milan Novak, A.R.Harish, M. Sachidananda, Alain Thys, Yunus A Cengel, John Cimbala, Aapo Hyvarinen, Erkki Oja, Davide Tavella, Florian Gouthiere, Frank Harrell, James F. Crow, Motoo Kimuram, T.H.G. Megson, Sophocles J. Orfanidis, Eric Barends, Denise M. Rousseau, Rafal Jozefowicz, Wojciech Zaremba, Ilya Sutskever, Matthew Hamilton, Philip C. Magnusson, Andreas Weisshaar, Vijai K. Tripathi, Gerald C. Alexander, Paul J. Nahin, Jean-Baptiste Joseph Fourier, Paul J. Nahin, Raymond A. Serway, John W. Jewett, David W. Ball, Daniel Fels, Michal Cifra, Felix Scholkmann, Gottlieb Daimler, John Dalton, James Dwight Dana, Charles Darwin, Humphry Davy, Peter Joseph Wilhelm Debye, Max Delbruck, Jean Andre Deluc, Democritus of Abdera, René Descartes, Rudolf Christian Karl Diesel, Diophantus of Alexandria, Paul Adrien Maurice Dirac, Prokop Divis, Theodosius Dobzhansky, Frank Drake, K. Eric Drexler, John Eccles, Arthur Eddington, Thomas Edison, Harmel Léon, Paul Ehrlich, Albert Einstein, Gertrude Elion, Empedocles, Hidek Yukawa, Eratosthenes of Cyrene, Euclid, Eudoxus of Cnidus, Leonhard Euler, Michael Faraday, Pierre de Fermat, Enrico Fermi, Richard Phillips Feynman, Fibonacci – Leonardo of Pisa, Emil Fischer, Ronald Aymler Fisher, Alexander Fleming, Ambrose Fleming, Howard Florey, Henry Ford, Ruben Ricchiuto, Lee De Forest, Dian Fossey, Thomas Young, Leon Foucault, Benjamin Franklin, Rosalind Franklin, Sigmund Freud, Elizebeth Smith Friedman, Galen, Galileo Galilei, Francis Galton, Luigi Galvani, George Gamow, Martin Gardner, Carl Friedrich Gauss, Murray Gell-Mann, Sophie Germain, Willard Josiah Gibbs, William Gilbert, Sheldon Lee Glashow, Robert Goddard, Maria Goeppert-Mayer, Thomas Gold, Jane Goodall, Stephen Jay Gould, Otto von Guericke, Fritz Haber, Ernst Haeckel, Otto Hahn, Albrecht von Haller, Edmund Halley, Alister Hardy, Thomas Harriot, William Harvey, Stephen Hawking, Felix Bloch, Pierre Francois Verhulst, Patrick H. Leslie, Alfred J. Lotka, Vito Volterra, Godfrey Harold Hardy, Wilhelm Weinberg, Edward John Routh, Adolf Hurwitz, Charles Cobb, Paul Douglas, Hermann Paasche, Etienne Laspeyres, Alfred Marshall, Francis Ysidro Edgeworth, Richard Roll, Robert B. Durand, Frederick Macaulay, John C. Cox, Jonathan E. Ingersoll, Fischer Black, Jack L. Treynor, Fischer Black, Norbert Wiener, Robert C. Merton, Otto Haxel, Karl Werner Heisenberg, Hermann Ludwig Ferdinand von Helmholtz, Empedocles of Acragas, Jan Baptist von Helmont, Joseph Henry, Caroline Herschel, Thomas Robert Malthus, John Herschel, William Herschel, Gustav Ludwig Hertz, Heinrich Hertz, George Gabriel Stokes, Karl F. Herzfeld, George de Hevesy, Jabir Ibn Hayyan, Antony Hewish, David Hilbert, Maurice Hilleman, Hipparchus of Nicaea, Hippocrates, Shintaro Hirase, Dorothy Hodgkin, Robert Hooke, Frederick Gowland Hopkins, William Hopkins, Grace Murray Hopper, Frank Hornby, Jack Horner, Leopold Kronecker, Bernardo Houssay, Abu Sahl al-Qouhi, Joseph Liouville, Fred Hoyle, Edwin Hubble, Alexander von Humboldt, Zora Neale Hurston, Frank Yates, James Hutton, Christiaan Huygens, Hypatia, August Ferdinand Möbius, Ernesto Illy, Ernst Ising, Keisuke Ito, Mae Carol Jemison, Edward Jenner, Johannes Hans Daniel Jensen, Irene Joliot-Curie, Jonathan B. Berk, Peter DeMarzo, Chuong B. Do, Serafim Batzoglou, Augustin-Louis Cauchy, Jan Egbertus Luitzen Brouwer, Benjamin Bray, Pontus Olofsson, Andrei Kolmogorov, Marco Taboga, Ibn Al Haytham, Ken Dill, Sarina Bromberg, Karen K. Yuen, Rand R. Wilcox, Kamil Akhmetovich Valiev, Helmuth Spieler, Simon Min Sze, Myoungkyu Lee, Ulrich Walter, Max Kuhn, Hannu Karttunen, Pekka Kroger, Heikki Oja, Markku Poutanen, Karl Johan Donner, Edward W. Kolb, Lee Tsung-Dao, Horacio Santana Vieira, Anderson Gray McKendrick, Valdir B. Bezerra, Horacio Santana Vieira, Valdir B. Bezerra, David L. Pulfrey, Jean-Philippe Duzan, Victor J. Stenger, Richard Dawkings, Michael Martin, Ricki Monnier, Christopher Hitchens, David L. Streiner, Herbert Benson, Jeffery A. Dusek, Jane B. Sherwood, Peter Lam, Charles F. Bethea, William Carpenter, Georg Cantor, Sidney Levitsky, Peter C. Hill, Donald W. Clem Jr., Manoj K. Jain, Randolph C. Byrd, Sebastien Ruder, Kimball A. Milton, James Prescott Joule, Percy Lavon Julian, Michio Kaku, Heike Kamerlingh Onnes, Pyotr Kapitsa, Friedrich August Kekulé, Frances Kelsey, Pearl Kendrick, Johannes Kepler, Abdul Qadeer Khan, Omar Khayyam, Frank Anscombe, Alfred Kinsey, Gustav Kirchhoff, Martin Klaproth, Robert Koch, Emil Kraepelin, Thomas Kuhn, Stephanie Kwolek, Joseph-Louis Lagrange, Jean-Baptiste Lamarck, Hedy Lamarr, Edwin Herbert Land, Karl Landsteiner, Pierre-Simon Laplace, Max von Laue, Antoine Lavoisier, Ernest Lawrence, Henrietta Leavitt, Antonie van Leeuwenhoek, Inge Lehmann, Gottfried Wihelm Leibniz, Georges Lemaître, Leonardo da Vinci, Niccolo Leoniceno, Henri Bénard, Aldo Leopold, Rita Levi-Montalcini, Claude Levi-Strauss, Willard Frank Libby, Justus von Liebig, Heraclides Ponticus, Carolus Linnaeus, Joseph Lister, John Locke, Hendrik Antoon Lorentz, Konrad Lorenz, Ada Lovelace, Percival Lowell, Lucretius, Charles Lyell, Trofim Lysenko, Jean-Baptiste Say, Pascal Brachet, Ernst Mach, Marcello Malpighi, Jane Marcet, Guglielmo Marconi, Lynn Margulis, Polly Matzinger, Matthew Maury, Muhammad Ibn Ahmad Abul-Rayhan Al-Biruni, James Clerk Maxwell, Ernst Mayr, Barbara McClintock, Lise Meitner, Gregor Johann Mendel, Yajnavalkya, Dmitri Mendeleev, Franz Mesmer, Antonio Meucci, John Michell, Albert Abraham Michelson, Thomas Midgeley Jr., Milutin Milanković, Maria Mitchell, Mario Molina, Thomas Hunt Morgan, Samuel Morse, Henry Moseley, Ukichiro Nakaya, John Napier, Henry Oldenburg, Giulio Natta, John Needham, John von Neumann, Thomas Newcomen, Adam Smith, Isaac Newton, Charles Nicolle, Florence Nightingale, Tim Noakes, Alfred Nobel, Emmy Noether, Christiane Nusslein-Volhard, Bill Nye, Hans Christian Örsted, Georg Simon Ohm, Frank Wilcoxon, J. Robert Oppenheimer, Wilhelm Ostwald, William Oughtred, Josef Stefan, Blaise Pascal, Louis Pasteur, Wolfgang Ernst Pauli, Linus Pauling, Charles François Sturm, Randy Pausch, Ivan Pavlov, Cecilia Payne-Gaposchkin, Wilder Penfield, Marguerite Perey, William Perkin, John Philoponus, Jean Piaget, Philippe Pinel, Max Planck, Pliny the Elder, Henri Poincaré, Hermann Minkowski, Karl Popper, Beatrix Potter, Joseph Priestley, Proclus, Claudius Ptolemy, Pythagoras, Adolphe Quetelet, Harriet Quimby, Thabit Ibn Qurra, Nikolai Ivanovich Lobachevsky, C. V. Raman, Srinivasa Ramanujan, William Ramsay, John Ray, John Henry Poynting, René Thom, Siméon Denis Poissson, Prafulla Chandra Ray, Arthur Cayley, Nicolas Léonard Sadi Carnot, Gaspard Monge, Francesco Redi, Karl Marx, Sally Ride, Georg Friedrich Bernhard Riemann, Wilhelm Röntgen, John Forbes Nash, Élie Cartan, Hermann Rorschach, Ronald Ross, Ibn Rushd, Ernest Rutherford, Carl Sagan, Abdus Salam, Jonas Salk, Robert Cox Merton, Frederick Sanger, John William Strutt Rayleigh, Henri Navier, Alberto Santos-Dumont, André-Marie Legendre, Walter Schottky, Erwin Schrödinger, Theodor Schwann, John Maynard Keynes, Glenn Seaborg, Hans Selye, Carl Jacobi Charles Sherrington, Gene Shoemaker, Ernst Werner von Siemens, George Gaylord Simpson, Edward H. Simpson, Milutin Milankovic, B. F. Skinner, William Smith, Frederick Soddy, Mary Somerville, Henri Léon Lebesgue, Arnold Sommerfeld, Hermann Staudinger, Nicolas Steno, Nettie Stevens, Andrei Andreyevich Markov, William John Swainson, Alexis-Claude Clairaut, Leo Szilard, Niccolo Tartaglia, Edward Teller, Nikola Tesla, Thales of Miletus, Theon of Alexandria, Benjamin Thompson, Joseph John Thomson, William Thomson, Henry David Thoreau, Kip S. Thorne, Clyde Tombaugh, Susumu Tonegawa, Evangelista Torricelli, Charles Hard Townes, Youyou Tu, Alan Turing, Jürgen Wagner, Neil deGrasse Tyson, Harold Urey, Tullio Levi-Civita, Craig Venter, Vladimir Vernadsky, Maric Vladimir, Andreas Vesalius, Rudolf Virchow, Artturi Virtanen, Alessandro Volta, Selman Waksman, George Wald, Alfred Russel Wallace, Peter W. Milonni, Meng-Li Shih, Jeremy N. Munday, Federico Capasso, V. Adrian Parsegian, H. Vic Dannon, Leonard S. Levitt, Marco Scutari, Jean-Baptiste Denis, A. Ronveaux, Felix Medland Arscott, Gerhard Wolf, Plinio Maroni, Anne Duval, Christian P. Robert, George Casella, Gunter Meissner, Ernst Gobel, Uwe Siegner, Russell C. Hibbeler, Tristan Fletcher, Dan Ventura, Michael Ware, Justin Peatross, Michael Edward Peskin, John Wallis, Ernest Walton, James Watson, James Watt, Alfred Wegener, John Archibald Wheeler, Maurice Wilkins, Thomas Willis, E. O. Wilson, William Wilson, Sven Wingqvist, Sergei Winogradsky, Carl Woese, Friedrich Wöhler, Wilbur and Orville Wright, Wilhelm Wundt, Chen-Ning Yang, Ahmed Zewail, V. B. Berestetskii, Lev Petrovich Pitaevskii, John Stewart Bell, M. J. Kearsley, John Bradbury Sykes, Zivorad R. Lazic, Jean-Pierre Demailly, John A. Cornell, Alan Sokal, Jean Bricmont, Mohammed J. Zaki, Wagner Meira, Raghu N. Kacker, Eric S. Lagergren, James J. Filliben, Cleve Moler, Norman H. Nie, C. Hadlai Hull, Dale H. Bent, Keith Geddes, Gaston Gonnet, Bill Gates, Spencer Kimball, Peter Mattis, H.C. Co, Raymond Theoret, Don M. Chance, Mao Xin, Andrew Fraknoi, David Morrison, Sidney C. Wolff, Philippe Mangin, Remi Kahn, Yung Kuo Lim, Robert Eisberg, Thomas Juffman, Adriana Milic, Michael Mullneritsch, Peter Asenbaum, Alexander Tsukernik, Jens Tuxen, Marcel Mayor, Ori Cheshnovsky, Markus Arndt, Christian Wagner, Simon Miller, Jonathan M. Garibaldi, David A. Freedman, Walter Schachermayer, Josef Teichmann, Walter Zucchini, Adrian W. Bowman , Adelchi Azzalini, A. Berzel, O. Nenadic, Ross Ihaka, Robert Gentleman, Leslie Lamport, David A. Grossman, Ophir Frieder, Stan Uryasev, Sergey Saryakin, Gaia Serraino, Konstantin Kalinchenko, Jeffrey Strickland, Giordano Bruno, Douglas Bates, John Chambers, Peter Dalgaard, Kurt Hornik, Tomas Kalibera, Michael Lawrence, Friedrich Leisch, Uwe Ligges, Thomas Lumley, Martin Maechler, Martin Morgan, Paul Murrell, Martyn Plummer, Brian Ripley, Deepayan Sarkar, Duncan Temple Lang, Luke Tierney, Simon Urbanek, Heiner Schwarte, Guido Masarotto, Stefano Iacus, Seth Falcon, Duncan Murdoch, Paul Joseph Cohen, Gaspard Coriolis, Daniel McFadden, Alain Cohen, Antoine Augustin Cournot, Rudolf Clausius, Gabriel Cramer, Sophus Lie, Irène Joliot-Curie, Frédéric Joliot-Curie, Harry Maurice Markowitz, George Bernard Dantzig, Peter Gustav Dirichlet(-Lejeune), Christian Doppler, Paul Karl Ludwig Drude, Paul Erdös, Agner Krarup Erlang, Mikhail Vasilyevich Ostrogradsky, Mitchell Feigenbaum, Joseph von Fraunhofer, Augustin Jean Fresnel, Alexander Friedmann, Kurt Gödel, Evariste Galois, William Sealy Gosset, Frege Friedrich Ludwig Gottlob, Alexander Grothendrieck, Edwin Herbert Hall, William Rowan Hamilton, Felix Hausdorff, Oliver Heaviside, Abu al-Wafa' Buzjani, Charles Hermite, Harold Hotelling, Gregory Benford, R. Length, David B. Laney, William Gemmell Cochran, Sture Holm, Carlo Emilio Bonferroni, Zbynek Sidak, Nikolai Smirnov, R. Length, David B. Laney, William Gemmell Cochran, Sture Holm, Carlo Emilio Bonferroni, Zbynek Sidak, Theodore Wilbur, Donald Allan Darling Anderson, Harald Cramer, Richard von Mises, Frank Wilcoxon, Henry Mann, Donald Ransom Whitney, William Kruskal, Wilson Allen Wallis, Milton Friedman, Charles Spearman, Maurice Kendall, R. H. Wilson, Agner Krarup Erlang, Quinn McNemar, Edward L. Kaplan, Paul Meier, Nathan Mantel, William M. Haenszel, David Cox, Guido Fubini, Lorenzo Mascheroni, Luigi Guido Grandi, Andre-Louis Cholesky, Jorgen Pedersen Gram, Elwin Bruno Christoffel, Germinal Pierre Dandelin, Rudolf Lipschitz, Johann Samuel König, Waclaw Sierpinski, Thomas Simpson, Carl Runge, Barbara Ryan, Brian Joiner, Jonathan Cryer, Carl Neumann, Adhemar Jean Claude Barre de Saint-Venant, Adolf Fick, Warren McCulloch, Walter Pitts, Wesley A. Clark, Frank Rosenblatt, Alexey Ivakhnenko, Henry J. Kelley, Arthur E. Bryson, Ian Goodfellow, Kunihiko Fukushima, David E. Rumelhart, Jean Leonard Marie Poiseuille, Giovanni Battista Venturi, Ian Goodfellow, Henri Pitot, Osborne Reynolds, Joseph Valentin Boussinesq, Hugo Tetrode, Otto Sackur, Meghnad Saha, Joseph Larmor, Alfred-Marie Lienard, Emil Wiechert, Max Abraham, Emil Lenz, August Beer, Johann Heinrich Lambert, Willebrord Snellius, Cherenkov radiation, Etienne-Louis Malus, Howard P. Robertson, Hendrik Casimir, Paul Ehrenfest, Alfred Lande, Willis Eugène Lamb, Jacques Philippe Marie Binet, Hippolyte Fizeau, Joseph Carl Hafele, Richard E. Keating, Alfred Schild, Irwin Shapiro, Howard P. Robertson, Arthur Geoffrey Walker, Bryce DeWitt, John Archibald Wheeler, Yoichiro Nambu, Tetsuo Goto, Hero of Alexandria, James Stirling, Abraham de Moivre, W. Edwards Deming, Colin Lingwood Mallows, Al-Sijistani Zuraqi, Hirotugu Akaike, Roger Cotes, Helge von Koch, Gaston Julia, Shelly Palmer, Maurice Karnaugh, Edward W. Veitch, John Venn, David A. Huffman, August Albert Sardinas, George W. Patterson, Auguste Kerckhoffs, Horst Feistel, Whitfield Diffie, Martin Hellman, Ron Rivest, Adi Shamir, Leonard Adleman, Geoffrey Everest Hinton, Michel Mayor, Didier Queloz, Eric Cornell, Carl Wieman, Andrew Wiles, Bill Smith, Mikel J. Harry, Carlo Rubbia, Simon van der Meer, Alain Aspect, Leonard E. Baum, Peter Mansfield, Andrew Maudsley, John Holl, Irving Wingate Burr, Arno Penzias, Robert Wilson, Robert Dicke, James W. Cooley, Ali Javan, Theodore Maiman, Irving Reed, Gustave Solomon, Yakir Aharonov, Robert G. Chambers, Leon Neil Cooper, John Robert Schrieffer, Clyde Cowan, Fred Reines, Herbert Simon, Owen Chamberlain, Emilio Gino Segrè, Clyde Wiegand, Thomas Ypsilantis, W. K. Hastings, Nicolas Metropolis, Gerhart Lüders, Julian Schwinger, Richard Wesley Hamming, Polykarp Kusch, Cecil Powell, Richard Powell, Eugene Alvin Villar, Cesare Mansueto Giulio Latte, Giuseppe Occhialini, Walter H. Brattain, William Schockley, William Bradford Schockley, Dominique François Jean Arago, Tomonaga Sin-Itiro, Ernst Stueckelberg, William Donald Kerst, Tommy Flowers, Sidney Broadhurst, Abu al-Fath Kbhahazini, William Chandler, Fritz Strassmann, Mark Ott, Otto Robert Frisch, Markus Fierz, Frederik Jozef Belinfante, Isidor Isaac Rabi, Carl von Weizsäcker, Piotr Leonidovich Kapitza, Seth Neddermeyer, Jabez Curry Street, E.C. Stevenson, Abd al-Rahman al-Sufi, Walter Andrew Shewhart, Boris Podolsky, Nathan Rosen, Pavel Cherenkov Alekseyevich, Ernst August Friedrich Ruska, John Cockcroft, Fritz London, Robert Van de Graaff, Friedrich Hund, Robert S. Mulliken, Walter Heitler, Clinton Joseph Davisson, Lester Germer, George Paget Thomson, Ma'ruf al-Karkhi, Oscar Klein, Walter Gordon, Pierre Auger, George Uhlenbeck, Samuel Goudsmit, Pascual Jordan, John Lennard-Jones, Otto Stern, Walter Gerlach, Theodor Franz Eduard Kaluza, Vesto Melvin Slipher, Ernst Julius Öpik, Harlow Shapley, Satyam Bhuyan, Santanu Mukherjee, Gilbert Lewis, Jeff Hein, Irving Langmuir, Marcel Grossmann, William Henry Bragg, Johannes Stark, Walter Friedrich Knipping, Paul Knipping, Jean Perrin, Robert Andrews Millikan, Hans Geiger, Ernest Marsden, Thomas Royds, Walther Nernst, Paul Villard, Alfred-Marie Liénard, George Johnstone Stoney, Wilhelm Wien Carl Werner, Irving Stringham, Giuseppe Peano, Issai Schur, Edward Morley, Karl Hermann Amandus Schwarz, Simon Newcomb, William Kingdon Clifford, William Thomson Kelvin, Gregor Johann Mendel, Urbain Le Verrier, Bernardus Placidus Johann Nepomuk Bolzano, James Joseph Sylvester, Edouard Roche, Julius von Mayer, Émile Clapeyron, Heinrich Lenz, George Green, William Sturgeon, Pierre Frédéric Sarrus, Jean-Victor Poncelet, Joseph Diaz Gergonne, Christian Kramp, Jean-Robert Argand, Joseph Louis Gay-Lussace, William Nicholson, Anthony Carlisle, Caspar Wessel, Jacques Alexandre César Charles, Carl Scheele, Joseph Black, Pierre-Louis Moreau, Jonathan Hulls, Geralamo Saccheri, Stephen Gray, Guillaume de l'Hopital, Denis Papin, Olaus Roemer, Nicolas Lemery, Johan de Witt, James Gregory, John Graunt, Edme Mariotte, Albert Girard, Santorio Santorio, Hans Lippershey, Paolo Sarpi, David Fabricius, Gernian Montanari, Nathaniel Pigott, John Goodricke, Johann Georg Palitzsh, Hans Janssen, Zacheraius Janssen, Simon Stevin, Rafaelle Bombelli, Ludovico Ferrari, Niccolò Fontana, Robert Recorde, Michael Stifel, Christoff Rudolff, Scipione del Ferro, Johannes Gensfleisch zur Laden zum Gutenberg, Albrecht Dürer, Johannes Widmannwe, Nicolas Chuquet, Jamshid al-Kashi, Zhu Sjijie, Raymon Lulle, Pierre de Maricourt, Roger Bacon, Jordan Nemore, Al-Hassãr, Bhaskara, Apollonius of Perga, Philo of Byzantium, Allah Abu'l-Barakat Hibat al-Baghdaadi, Henry Norris Russell, Charles Messier, Jason Cole, Gloria Faccononi, James Childs, Ejnar Hertzsprung, Zhang Heng, Antonio Ciccolella, Tunç Tezel, Karl Garnham, Christoph Schiller, Rhys Taylor, Wang Ch'ung, NASA, Stefano Meschiari, John Schwabish, Severino Ribecca, Afshine Amidi, Shervine Admidi, Rafel Del Rio, Josef Ricafort Jr., ESA, Zhang Long, Arthur Dempster, Nan Laird, Donald Rubin, Jeremy Leconte, Yanghong Huang, Till Tantau, Christian Feuersänger, Vincent Florin, Joram Soch, Alessandro Roussel, Sam Dietterich, Giacomo Tommei, Stefano Marò, Vedat Tanriverdi

Obviously... the views expressed in this textbook are those of the redactors and do not necessarily reflect the views of all the people cited above!

This textbook would not have been possible without the valuable support of the people mentioned below who greatly helped to make original texts more readable for the layman. They find here the expression of my gratitude (and for sure if some errors remains in this textbook this is obviously their fault...):

- **HARMEL Léon** (†12012), graduate electrical engineer with a specialization in electronics and automation, responsible in the physical research laboratory at ACEC in Charleroi (BEL), for the redaction of almost 400 pages used in the sections of Corpuscular Quantum Physics, Wave Quantum Physics, Quantum Field Theory, Spinor Calculus and General Relativity.

- **KOLANI Daname**, PhD student in portfolio management at UM5-Rabat (MAR) for his help on the redaction of almost 700 pages of the `R` and MATLAB® companion books, proofreading and complements of the section in this textbook about data mining and machine learning and translation of a few hundred pages of the original french website into this LaTeX book.

- **LEGRAND Mathias**, PhD of the École Centrale de Nantes (FRA) for his help on the translation of the first 550 pages of the original french website into this LaTeX book.

- **RICCHIUTO Ruben**, engineer degree in physics HES (B.Sc.) from the Engineering School of Geneva (CHE) and mathematician from the University of Geneva for his valuable help (almost 150 pages) in plasma physics, electromagnetism, quantum physics, statistics, topology, quantum chemistry, fractals theory, analysis and many other areas affecting pure mathematics and computing.

- The regular users of Les Mathematiques.net, Futura-Sciences.com, Maths Stackexchange.com, Physics Stackexchange.com and LaTeX Stackexchange.com forums, for their valuable assistance in many areas of mathematics and physics. The debates and discussions that took place on the forums helps to constantly improve the educational aspect of this textbook.

- The Wikipedia, Wikipedia and ProofWiki websites to whom I am indebted to many borrow almost word by word (and this seems sometimes to be mutual...).

- The support of some artificial intelligences like Microsoft Copilot, Goole Bard, NotebookLM to improve some sentences, proofs and create a few illustrations.

- The professors Gabriel Nagy of Michigan State University, Liam Revell of Massachusetts Boston University, Jorge Cham (of PhD Comics), Lili Xiao (Department of Mathematical Sciences, Tsinghua University, Beijing, China), Dennis K. J. Lin (Department of Statistics, Pennsylvania State University), Fengshan Bai (Department of Mathematical Sciences, Tsinghua University, Beijing, China) and the OpenStax team that have received and read my multiple e-mails inquiries about images credits and texts credits but who never answered to me (not quite good examples of what should be the "scientific collaboration and spirit"...)

And thanks to all readers, webmasters and teachers for their websites and quality documents available for free and anonymously on the Internet and also to regular forum stakeholders. I sometimes verbatim recovered their explanations that didn't require additions or corrections. It's probably needless to say that the reader should not assume that these people are in total agreement with the scientific purposes views expressed in this textbook; and are not responsible for any errors or obscurities that the reader might accidentally find in it.

Thanks also to the few colleagues and customers who were willing to give me their comments to improve the content of this textbook. However, it is certain that it can still be improved on many points.

Several scientific figures were kind enough to support us in our approach to providing free, high-quality access to a scientific reference compendium for the general public. Here are their encouragements:

Les concepts et les mots pertinents de la science
sont, tôt ou tard, destinés à être exprimés
en concepts et en mots qui ont un sens pour
le public instruit, et à s'inscrire dans une
image générale du monde. Le site www.sciences.ch
œuvre admirablement pour cette noble entreprise.
En faisant jaillir l'émotion du savoir réputé froid,
le site ouvre les portes à la surprise, à la
curiosité, à l'humour, à l'émerveillement

PhD. Jean-Pierre Luminet (11951-) French astrophysicist specializing in black holes and cosmologyand emeritus research director at the CNRS

Comment se passer de la science
si l'on cherche la lucidité ?

PhD. Albert Jacquard (11925 – 12013) French genetics popularizer of science and essayist

4

Introduction

Contents

4.1 Forewords 35
- 4.1.1 Motivation and goals 41
- 4.1.2 Who this textbook is intended for 47
- 4.1.3 How is this textbook structured 49

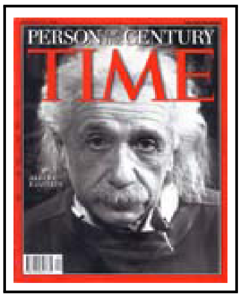
4.2 Methods 50
- 4.2.1 Descartes' Method 56
  - 4.2.1.1 Blind studies 58
- 4.2.2 Research Integrity and Engineering/Scientific Ethics 66
  - 4.2.2.1 Singapore Statement on Research Integrity 66
  - 4.2.2.2 European Code of Conduct for Research Integrity 67
  - 4.2.2.3 Archimedean Oath 68
- 4.2.3 Scientific Publication Rules (SPR) 70
- 4.2.4 Scientific Mainstream Media communication 79
  - 4.2.4.1 Social Networks 81
  - 4.2.4.2 Expert Opinions 85

4.3 Vocabulary 88
- 4.3.1 On Sciences 89
- 4.3.2 Terminology 99

4.4 Science and Faith 103
- 4.4.1 Baloney detection kit 109

4.5 Scientific communication backfire 115

4.6 Is Science dogmatic? 120



This textbook whose first edition has been published in 12001 is designed so that the knowledge required to read it is as basic as possible. It is not necessary to have a PhD to consult it, you just have to know reasoning, to think critically, to observe and have a lot of free time available...

"Simplicity is the seal of truth and it radiates beauty"
Albert Einstein

4.1. Forewords

It is very likely that no human endeavour has had more impact than Science[1] on our lives, our conception of the Universe and ourselves these last centuries. Its theories, models and methodologies conquests and results are all around us in our industrial countries for the best and also sometimes... for the worst.

Omnipresent in the industry (aerospace, imaging, cryptography, transportation, chemistry, algorithmic, etc.) or in the services (banking, fintech, insurance, human resources, projects, logistics, architecture, communications, etc.), Applied Mathematics are a powerful tool that approximated very well our reality and that also appear in many other areas like: surveys, risk modelling, data protection, politics, etc. Applied Mathematics (also sometimes named "Mathematics Machinery") influence our lives (telecommunications, transport, medicine, meteorology, music, project management) and contribute to the resolution of current issues: energy, health, environment, climate, optimization, sustainable development, etc. much more than any soft skill technique or methodology! They great success are their fabulous dispersion in the real world and their increasing integration in all human and artificial intelligence activities. We are going therefore to a situation where mathematicians and engineers will no longer have the monopoly of mathematics, but where almost any graduate job position will have to do advanced mathematics in a way or an other.

As a former student in the field of engineering I have often regretted the absence of a single textbook fairly comprehensive, detailed (without going to the extreme...) and educational if possible free (!) and portable (being personally a fan of eBooks...) containing at least a non exhaustive idea of the overall program of Applied Mathematics in engineering schools with an overview of what is used for real in companies with more intuitive than rigorous proofs but with enough details to avoid unnecessary effort to the reader. Also a textbook that does not require the reader to adopt each time a new notation or terminology specific to the author when it is not outright to change to a foreign language... and where anyone can suggest improvements or additions (through the forum, guest-books or by e-mail).

I was also frustrated during my studies to have quite often have to swallow "formulas" or "laws[2]" supposedly (and wrongly) non-provable or too complicated as my teachers says or even disappointed by renowned authors textbooks (where developments are left to the reader or as exercise and no real applications are even mention...).

---

[1] From Latin *scientia* "knowledge, a knowing, expertness". Itself from *sciens* (genitive scientis) that means "intelligent, skilled", present participle of *scire* that means "to know" probably originally comes from "to separate one thing from another, to distinguish" related to *scindere* "to cut, divide".

[2] "Laws" are descriptive in Science not prescriptive like in religions or in human legal system.

In this textbook predominates the will to avoid to confuse the reader with a majority of empty sentences like "it is evident that...", "it is easy to prove that...", "we leave it to the reader as an exercise...", "it should be known that...", since all developments are presented - we expect - in sufficient detail. But I'm not a purist of maths! I have only one ambition: to explain things the easiest possible possible without loosing too much rigour.

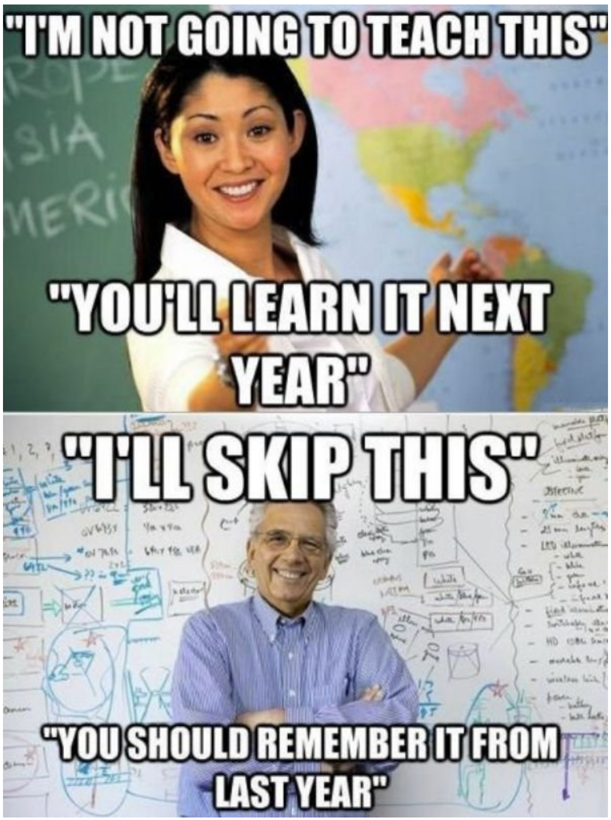
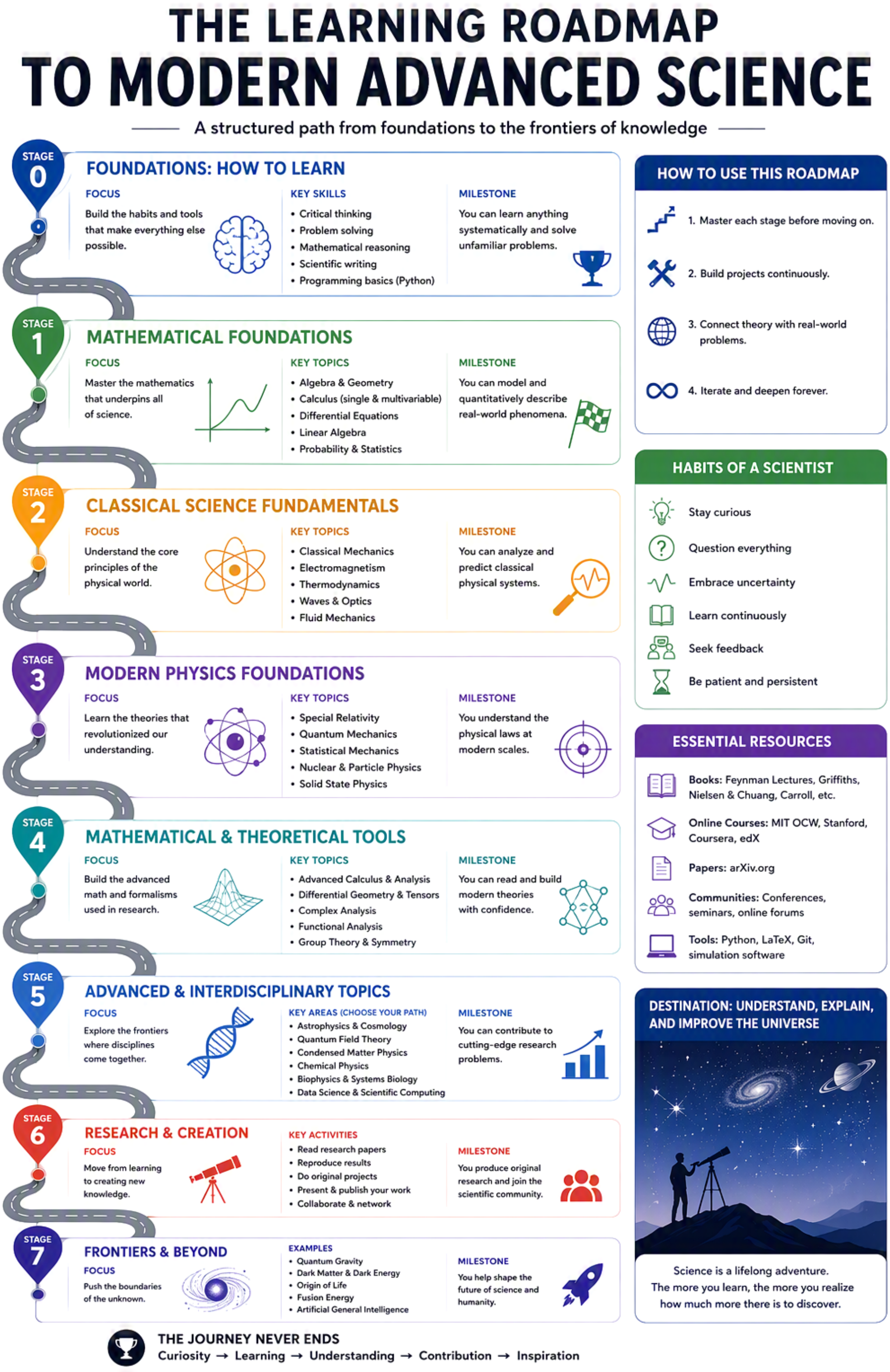

"I'M NOT GOING TO TEACH THIS"
"YOU'LL LEARN IT NEXT YEAR"
"I'LL SKIP THIS"
"YOU SHOULD REMEMBER IT FROM LAST YEAR"

Figure 4.1: **Typical kind of situation we want to avoid in this textbook**

Although I have to admit that prove some mathematical relations presented within the engineering schools curriculum can not be done because of a lack of time in the official program or size limit in a textbook, I can not accept that a teacher or author tells his students (respectively, his readers) that some relations are non-provable (because most of the time this is not true!) or that such or such proof is too complicated without giving a reference (where the student can find the information necessary to satisfy his curiosity) or at least a simplified but satisfactory proof.

Moreover, I think that it is totally archaic today that some teachers continue to ask to their students to take a massive quantity of notes during classes. It would be much more favourable and optimal to distribute a course handout containing all the details in order to be able to concentrate on the essentials points with students, that is to say the oral explanations, interpretations, understanding, reasoning, critical thinking and practice rather than excessive and time-consuming blackboard copy... Obviously by giving a complete course handout some students will be brilliant by their absence but ... it is the better! Thus, those who are passionate can deepen subjects at home or at the university library, the weak do what they have to do and the rest (struggling students but workers) will follow the course given by the teacher to profit to ask questions rather than mindlessly copying a blackboard.

Inspired on a learning model of an American scholar, whose I forgot the name (...), this textbook proposes and imposes the following properties to the reader: discover, memorize, cite, integrate, explain, restate, infer, select, use, decompose, compare, interpret, judge, argue, model, develop, create, search, reasoning, develop in a clear progressive teaching way to develop the analytic skills and openness.



THE LEARNING ROADMAP
TO MODERN ADVANCED SCIENCE

A structured path from foundations to the frontiers of knowledge

STAGE 0 — FOUNDATIONS: HOW TO LEARN

FOCUS
Build the habits and tools that make everything else possible.

KEY SKILLS
- Critical thinking
- Problem solving
- Mathematical reasoning
- Scientific writing
- Programming basics (Python)

MILESTONE
You can learn anything systematically and solve unfamiliar problems.

STAGE 1 — MATHEMATICAL FOUNDATIONS

FOCUS
Master the mathematics that underpins all of science.

KEY TOPICS
- Algebra & Geometry
- Calculus (single & multivariable)
- Differential Equations
- Linear Algebra
- Probability & Statistics

MILESTONE
You can model and quantitatively describe real-world phenomena.

STAGE 2 — CLASSICAL SCIENCE FUNDAMENTALS

FOCUS
Understand the core principles of the physical world.

KEY TOPICS
- Classical Mechanics
- Electromagnetism
- Thermodynamics
- Waves & Optics
- Fluid Mechanics

MILESTONE
You can analyze and predict classical physical systems.

STAGE 3 — MODERN PHYSICS FOUNDATIONS

FOCUS
Learn the theories that revolutionized our understanding.

KEY TOPICS
- Special Relativity
- Quantum Mechanics
- Statistical Mechanics
- Nuclear & Particle Physics
- Solid State Physics

MILESTONE
You understand the physical laws at modern scales.

STAGE 4 — MATHEMATICAL & THEORETICAL TOOLS

FOCUS
Build the advanced math and formalisms used in research.

KEY TOPICS
- Advanced Calculus & Analysis
- Differential Geometry & Tensors
- Complex Analysis
- Functional Analysis
- Group Theory & Symmetry

MILESTONE
You can read and build modern theories with confidence.

STAGE 5 — ADVANCED & INTERDISCIPLINARY TOPICS

FOCUS
Explore the frontiers where disciplines come together.

KEY AREAS (CHOOSE YOUR PATH)
- Astrophysics & Cosmology
- Quantum Field Theory
- Condensed Matter Physics
- Chemical Physics
- Biophysics & Systems Biology
- Data Science & Scientific Computing

MILESTONE
You can contribute to cutting-edge research problems.

STAGE 6 — RESEARCH & CREATION

FOCUS
Move from learning to creating new knowledge.

KEY ACTIVITIES
- Read research papers
- Reproduce results
- Do original projects
- Present & publish your work
- Collaborate & network

MILESTONE
You produce original research and join the scientific community.

STAGE 7 — FRONTIERS & BEYOND

FOCUS
Push the boundaries of the unknown.

EXAMPLES
- Quantum Gravity
- Dark Matter & Dark Energy
- Origin of Life
- Fusion Energy
- Artificial General Intelligence

MILESTONE
You help shape the future of science and humanity.

HOW TO USE THIS ROADMAP

1. Master each stage before moving on.
2. Build projects continuously.
3. Connect theory with real-world problems.
4. Iterate and deepen forever.

HABITS OF A SCIENTIST

- Stay curious
- Question everything
- Embrace uncertainty
- Learn continuously
- Seek feedback
- Be patient and persistent

ESSENTIAL RESOURCES

- **Books:** Feynman Lectures, Griffiths, Nielsen & Chuang, Carroll, etc.
- **Online Courses:** MIT OCW, Stanford, Coursera, edX
- **Papers:** arXiv.org
- **Communities:** Conferences, seminars, online forums
- **Tools:** Python, LaTeX, Git, simulation software

DESTINATION: UNDERSTAND, EXPLAIN, AND IMPROVE THE UNIVERSE

Science is a lifelong adventure.
The more you learn, the more you realize how much more there is to discover.

THE JOURNEY NEVER ENDS
Curiosity → Learning → Understanding → Contribution → Inspiration

So, in my mind, this non-exhaustive textbook (and its associated companion PDFs) must be a substitute, free of charge for all students and employees around the World, to many references and gaps of the scholar system, allowing any curious student not to be frustrated for many years during his academic curriculum. Otherwise, the science of the engineer could have the aspect of a frozen science, apart from the scientific and technical developments, a heterogeneous accumulation of knowledge and especially of formulas which made he considered as a tasteless subproduct of mathematics and that brings companies and governments to many false results and bad decisions...

This textbook has also been designed to meet the needs of executives, both finance as well as non-finance managers. Any executive who wants to probe further and grasp the fundamentals of strategic finance, strategic marketing or project management engineering and supply chain issues will benefit from its lecture.

This textbook has also for purpose to describe and explain how our observable Universe and our World (also other "worlds" in our Universe) works in a much more accurate, more complete and detailed way than any Holy book. This textbook will give you help the reader to understand how science is done and how scientists and engineers work, it will provide the reader models and quantification methods for the origin of species, of galaxies, of planets, of quantum phenomenon, of physics movements, of stellar physics, of extreme observable events and also extreme rare events and explains social strategies and modern technologies in a mathematical and provable way that everyone can check by himself and by exposing every-time the assumptions that any reasonable entity should take care of!

Obviously Applied Mathematics is such an abundant topic that a textbook of this scale can only accommodate the basis (we just scratch the surface of gigantic topics!). Readers are certainly encourage to go beyond this (see the bibliography at the end of the textbook at page 9443).

Now, those who see Applied Mathematics only as a tool (what it also is), or as the enemy of religious beliefs, or as a boring school field school, are legion. However, it is perhaps useful to recall that, as Galileo said, «*the book of nature is written in the language of mathematics*» (without wishing to do scientism!). When you go to China, you learn chinese. When you want go to the Universe you learn maths! Because maths are the language of the Universe. This is why maths are so fundamental and their are amazing as they apply to the whole Universe and across time. It is in this spirit that this textbook discusses Applied Mathematics for students in the Natural, Earth and Life sciences, as well as for all those who have an occupation related to the various subjects including philosophy or for anyone curious to learn about the involvement of science in everyday life.

The choice to study engineering in this textbook as a branch of Applied Mathematics comes from the fact that the differences between all areas of physics (formerly known as "natural philosophy") and mathematics are so hardly notable that Fields medal (the highest award today in the field of mathematics) was awarded in 11990 (holocene calendar) to physicist Edward Witten, who used physical ideas to prove a mathematical theorem. This trend is certainly not fortuitous, because we can observe that all science, since it seeks to achieve a more detailed understanding of the subject it studies, always finish its trials in the pure mathematics (the absolute path by excellence ...). Thus, we can predict in a far future, the convergence of all the sciences (pure, exact or social) to the mathematics for the modelisation techniques (see for example the French PDF *L'explosion des mathématiques* available in the download page of the companion website).

It can sometimes seem to us difficult (due to irrational as obscure and unjustified fear of pure sciences in a large fraction of our contemporaries) to transmit the feeling of the mathematical beauty of nature, its deepest harmony and the well-oiled mechanics of the Universe, to those who know only the basics of algebra. The physicist Richard Feynman spoke a day of "two cultures": people who have and those who do not have sufficient understanding of mathematics to appreciate the scientific structure of nature. It is a pity that mathematics are necessary to deeply understand nature and that they also have a bad reputation. For the record, it is claimed that a King who asked Euclid to teach him geometry complained about its difficulty. Euclid replied, "There is no royal road". Physicists and mathematicians can not convert themselves to a different language. If you want to learn about nature, to appreciate its true value, you must understand its language. The nature is revealed only in this form and we can not be pretentious to the point of asking him to change this fact.

In the same way, no intellectual discussion will allow you to communicate with a deaf person what you really feel while listening music. Similarly, all discussion of the world remain powerless to transmit an intimate understanding of the nature of those of the "other culture". Philosophers and theologians may try to give you qualitative ideas about the Universe. The fact that the scientific method (in the full sense of the term) can not convince the World of its strengths through its iterative process (the fact that it is rewritten and rewritten and rewritten by incremental improvement as in the DMAIC Six Sigma process[3]), is perhaps the fact of the limited horizon of some people who imagine that the human or another intuitive concept, sentimental or arbitrarily is the center of the Universe (anthropocentric principle).

---

[3] DMAIC (an acronym for Define, Measure, Analyze, Improve and Control) refers to a data-driven improvement cycle used for improving, optimizing and stabilizing processes and designs. The DMAIC improvement cycle is the core tool used to drive Six Sigma projects that we will study in-deep in the section of Industrial Engineering. However, DMAIC is not exclusive to Six Sigma and can be used as the framework for other improvement applications.

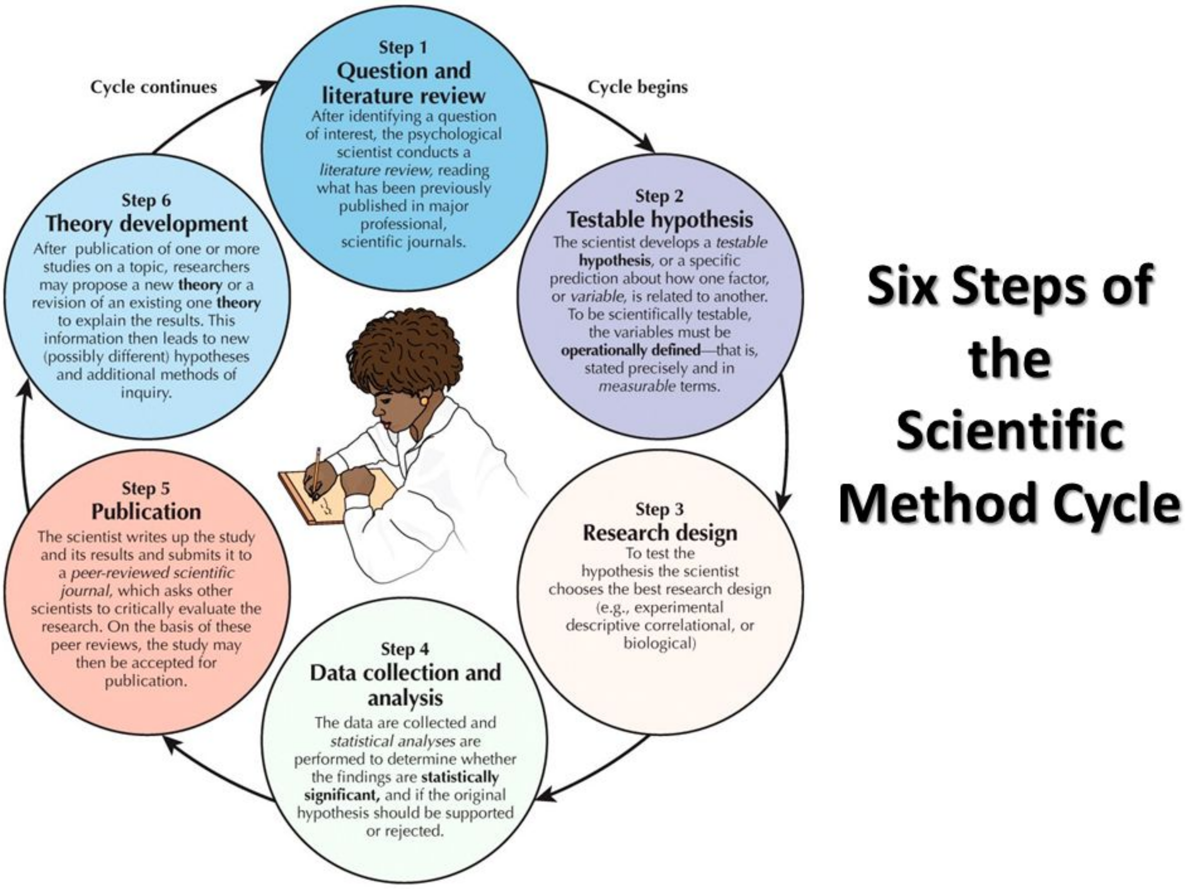

Cycle continues
Step 1
Question and literature review
After identifying a question of interest, the psychological scientist conducts a *literature review,* reading what has been previously published in major professional, scientific journals.
Cycle begins
Step 2
Testable hypothesis
The scientist develops a *testable* **hypothesis**, or a specific prediction about how one factor, or *variable,* is related to another. To be scientifically testable, the variables must be **operationally defined**—that is, stated precisely and in *measurable* terms.
Step 3
Research design
To test the hypothesis the scientist chooses the best research design (e.g., experimental descriptive correlational, or biological)
Step 4
Data collection and analysis
The data are collected and *statistical analyses* are performed to determine whether the findings are **statistically significant,** and if the original hypothesis should be supported or rejected.
Step 5
Publication
The scientist writes up the study and its results and submits it to a *peer-reviewed scientific journal,* which asks other scientists to critically evaluate the research. On the basis of these peer reviews, the study may then be accepted for publication.
Step 6
Theory development
After publication of one or more studies on a topic, researchers may propose a new **theory** or a revision of an existing one **theory** to explain the results. This information then leads to new (possibly different) hypotheses and additional methods of inquiry.
Six Steps of the Scientific Method Cycle

Figure 4.2: **Scientific method cycle process (source: ?)**

Definition (#4.1.1): The "scientific method" is the iterative process that involves making reproducible observations following a robust protocol, formulating a null hypothesis statistical test to explain the observed phenomenon, predicting new phenomenons with numerical outputs based on the hypothesis, conducting experiments to be able to refute the hypothesis, analysing statistically the data collected from experiments from significant power size samples using the triple blinded method and publishing the results by respecting the scientific publication rules.

quote
4.1:40.#3

> *The internet is flooded with people who think common sense talking points are how science is done!*
>
> ?

Remark

If you are an honest scientist (the word "scientist" seems to have been used for the first time in the 118th century according to holocene calendar), the huge majority of your ideas, even good ideas (!), are going to be ruled out, not by new experiments but already by inconsistency with old experiments. This is what really makes Science very different, and what gives us an internal notion of right and wrong before new experiments. So, quite contrary to the sense that some sceptical layperson would get, the idea that you can just make up crap is wrong!

So in science we have theory, experiment, and, in the middle between them, "phenomenology". Phenomenologists are the ones who coax predictions out of theories, usually by simplifying the math and figuring out what can be measured, to which precision, and how (and, not rarely, also by whom).

> ❞ *If you are not a professional active scientist, and you disagree with scientists about Science, it's actually not a disagreement! You're just wrong!! Science is not truth. Science is finding the evidence. When science changes its opinion, it didn't lie to you. It learned more!* ❞
>
> ?

quote 4.1:41.#4

4.1.1 Motivation and goals

The main purpose of this textbook is to make Modern Science known throughout the world so that all human beings can understand and know the origins and mechanisms of Nature around them.

Of course, in order to share this mathematical knowledge, it may seem paradoxical to increase, with our work, the long list of books already available in libraries, in commerce and on the Internet. Nevertheless, I must be able to present arguments that justifies the creation of such a textbook (and its associated website) as compared to textbooks such as that of Feynman ($\sim 1,500$ pages), Landau ($\sim 3,000$ pages) or Bourbaki ($\sim 8,200$ pages) and Wikipedia/Wolfram webpages themselves or Khan Academy videos or OpenStax PDFs ($\sim 5,500$ pages) or even the *Enzyklopädie der mathematischen Wissenschaften* ($\sim 20,000$ pages). So what do I think I can add to such a wealth of material?

1. The great pleasure that we take to write this textbook ("keep the hand" and improve our skills) and have a detailed high quality compendium of tools for our customers and our students (and also all those around the World) for free.
2. The passion for sharing knowledge for free (battle again "copyright madness" (RIP Aaron Swartz!)) and without frontiers with a tool of quality as LaTeX (at the opposite of Wikipedia that mixes LaTeX and normal text and the awful and shameful content of Khan Academy[4]).
3. Support free scientific education, critical thinking and evidence-based understanding of the natural World. Furthermore it is clear that there is an undeserved appetite for people to understand and this textbook has been written for this purpose (we will for example

---

[4] OpenStax has good undergraduate PDF - especially the example in their books - but there are between 40-60% of missing proofs and the table of contents of their PDF and also the Index are not interactive... and major issue...: the content is limited only to undergraduate subjects

introduce the scientific definition of a vaccine, we will deal mathematically with some of the fallacious arguments of flat earthers, we will also prove mathematically why it is normal not to see the stars in many photos or videos taken in space, we we will prove why the Big Bang (i.e. cosmic inflation theory) at the contrary to an urban legend is neither an explosion nor a singularity and was not discovered by G. Lemaître, we will prove mathematically that the Universe can emerge spontaneously by disintegration of the quantum vacuum, that the emergence of life from non-life is only a question of probabilities, etc.).

> ❞ *We don't debunk lies to convince the liars, we debunk lies to educate other people that lies aren't going unchallenged!* ❞

4. Helping defenceless people who are being dragged into false dichotomous beliefs by oversimplified false rhetorical statements or claims to acquire a certain level of critical method and investigation techniques.

5. Write a modern 3rd millennial version of the "Almagest" hence the name "Opera Magistris" that means in English "Major Work".

6. Because we can't wait as there are places in the world where the absence of teaching modern science and its methodology takes peoples to have believes that bring them to some dangerous and obscure paths.

7. We want to offer Applied Mathematics in an enjoyable and easy-to-learn manner ("keep it simple and stupid" at the opposite of the 9 Landau's graduate level books), because we think that Applied Mathematics change the way we understand the Universe.

8. This textbook was first written in French before (in year 12001 according to holocene calendar) that the French version of Wikipedia had good mathematical content and long before Khan Academy or OpenStax did even exist.

9. The quick updates/corrections opportunities to the last scientific discoveries (at the opposite of Khan Academy) and collaborations of a free e-book (with associated effective search tools) without having topics that disappears (at the opposite of Wikipedia).

10. The content depending on readers requests/comments and on our interests (at the opposite of Khan Academy, OpenStax or Landau textbooks)!

11. At the opposite of Scientific publications (PRL or other similar) that sucks because don't give detailed proofs and sometimes turn in an infinite loop of references.

12. The access to LaTeX sources to everybody so nobody need to reinvent the wheel and loose hundred or thousand of hours on redaction instead of innovation (at the opposite of Landau, Feynman and Bourbaki textbooks)!

13. Rigorous presentation with simplified detailed proofs of all presented concepts (at the opposite of Wikipedia, Khan Academy and OpenStax that focus only of the mathematical proofs of undergraduate concepts).

14. The presentation of many advanced and detailed mathematical tools used in business and R&D keeping in mind that the mathematical language seems eternal and to be one of the only common denominator between all countries in the World.

15. The opportunity for students and teachers to reuse content by copy/paste (at the opposite of Landau, Feynman and Bourbaki textbooks)!.
16. Constant and fixed notation (at the opposite of Wikipedia, Khan Academy and OpenStax) throughout the textbook, for mathematical operators, a clear language on all topics (3.C. criterion: clear, complete and concise) and focus on the basics to make an important pedagogical work on the subjects (at the opposite of Landau's textbooks).
17. Gather as much information about pure and exact sciences in one electronic (portable), homogeneous and rigorous textbook (but that don't go as far as Landau's textbooks).
18. Release from all pseudo-truths, only truths[5] that can be proven mathematically with detailed reasoning.
19. Benefit from the development of teaching methods that use the Internet to search for the solution of mathematical problems.
20. The dramatic improvement of automatic translation software and computing power that will make of this textbook, at least we hope, a reference in the fields of sciences.
21. A PDF is better than a website as first all people that use the Internet since 11990 (holocene calendar) know that the huge majority of website disappear after ten years and secondly it is well known that some countries block Wikipedia and other knowledge website to keep their population in the ignorance (and block a PDF that can be shared in a e-mail is much more difficult).
22. and... because Applied Mathematics are beautiful and especially when written in LaTeX and illustrated (at the opposite of Landau or Bourbaki textbooks whose illustrations and equation rendering quality are quite old and poor).

And also ... I think that the results of individual research are the property of humanity and should be available to all those who explore anywhere the phenomena of nature. In this way the work of each benefit to all, and that is for all humanity that our knowledge cumulates and this is the trend that allows Internet.

I do not hide that my contribution is limited largely to this day to that of a collector who gleans his information in the works of masters or publications or from anonymous web pages and who completes and argues developments and improved them when this is possible. Therefore some of the material in this textbook is original, and some comes from primary literature. However the vast majority of what we wrote is a rephrasing of results presented in the existing vast library of some (rare) fantastic textbooks. For those who would accuse me of plagiarism, they should think on the fact that the theorems presented in most non-free textbooks and commercially available have been discovered and written by their predecessors and their own personal contribution was also made, like mine, to put all this information in a clear and modern form a few hundred years later. In addition, it can be seen as doubtful that we ask to pay for access to a culture that is certainly the only truly valid and fair one in this world and where there is no patent or intellectual property rights.

---

[5] Here the word "truth" must not be taken in the sense of "absolute truth" that anyway doesn't exist in experimental science! But only as "something that works in a very reliable way as far as we know"!

quote
4.1:44.#5

> *If you steal from one author, it's plagiarism; if you steal from many, it's research!*
>
> Wilson Mizner

This textbook also reflects my own intellectual limitations. Although I try to study as much science and math fields as possible, it is impossible to master them all. This textbook shows clearly only my own interests and experiences as consultant, but also my strengths and my weaknesses. I am responsible for the selection of inputs and, of course, of possible errors and imperfections.

After attempting a strict (linear) order of presentation of the subject, I decided to arrange this textbook in a more pedagogical (thematic) way and always with practical examples o applications. It is in my opinion very difficult to speak of so vast subject in a purely mathematical order in only one human life, that is to say, when the concepts are introduced one by one, from those already known (where each theory, operator, tools, etc.. would not appear before its definition). Such a plan would require cutting the book, in pieces that are not more thematic. So I decided to present things in a logical order and not in order of need. Thus the reader will encounter, as the editor himself, to the extreme complexity of the subject.

The consequences of this choice are the following:

1. Sometimes it will necessary to admit certain concepts, even to understand later.
2. It will probably be necessary for the reader to go at least twice throughout the book. At the first reading, we apprehend the essential and at the second reading, we understand the details (I congratulate this who understand all the subtleties the first time).
3. You must accept the fact that some topics are repeated and that there are many cross-references and complementary remarks.

Some know that for every theorem and mathematical model, there are almost always several methods of proofs. I've always tried to choose the one that seemed the most simple (e.g. in relativity and quantum physics there is the algebraic and matrix formalism). The objective is to arrive at the same result anyway.

This textbook being in its draft version, it necessarily has lacks on convergence controls, on continuity, grammar and others... (which will horrify some readers and mathematicians ...)! However, I have avoided (or, otherwise, I indicate it) the usual approximations of physics and the use of dimensional analysis, by using it as little as possible. I also try to avoid as much as possible subjects with mathematical tools that have not previously been presented and demonstrated rigorously.

Finally, this presentation, that can still be improved, is not an absolute reference and contains errors. Any comment is welcome. I shall endeavour, as far as possible, to correct the weaknesses and make the necessary changes as soon as possible.

However, while mathematics is accurate and indisputable, theoretical physics (its models), is still interpreted in the common vocabulary (but not in the mathematical vocabulary) and its conclusions all relative. I can only advise, when you read this textbook, to read by for yourself and not to be subjected to outside influences. You must have a very (very) critical mind, take nothing for granted and question everything without hesitation. In addition, the keyword of good scientist should be: "Doubt, doubt, doubt ... doubt still, and always checks.". We also recall that "nothing that we can see, hear, smell, touch or taste, is what it seems to be", therefore do not rely on your daily experience to draw hasty conclusions, be critical, Cartesian, rational and rigorous in your development, reasoning and conclusions!

Remark

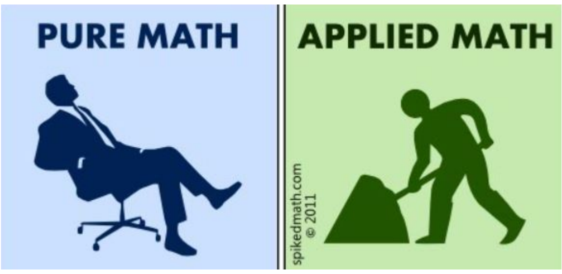
One of the causes of anti-intellectualism probably lies actually in the inability of educational institutions to inculcate critical thinking at high-school. In particular universities, unable to develop critical thinking of their students. The Internet and its amalgamation of true and false also have their part of responsibility. In addition, some experts increase the problem by speaking publicly on issues outside their area of expertise!!!!

I want to say to those who would try to find themselves the results of some developments of this textbook, do not worry if they do not success or if they doubt about their competences because of the time spent solving an equation or problem: some theories that seem obvious or easy today, have sometimes needed several weeks, months, even years, to be developed by mathematicians or leading physicists in the past!

I also tried to ensure that this textbook is pleasing to the eye and to read through.

Finally, I have chosen to write this work in the first person plural form: "we". Indeed, the mathematical physics is not a science that has been made or has evolve through individual work but with intensive collaboration between people connected by the same passion and desire of knowledge. Thus, by making use of "we", I would like to pay tribute to the dead and missing scientists, to contemporary and future researchers for the work they will perform in order to approach the truth[6] and wisdom.

I am happy to make this precious work available to you, and I hope that you will enjoy reading it a little every day!



---

[6] It is clear from the above discussion that scientific research does not and can not lead to a knowledge of nature that is completely free from error. Rather it leads and is able to lead only to an unending process in which the degree of truth in our knowledge is continually increasing. Thus, with the further progress of science into new domains, it becomes possible for us to define the errors in older laws in more and more detail and in more and more respects, and in this way to delimit the domains of validity of these laws more precisely and more nearly completely.

quote
4.1:45.#6

> *Truth is a word that oddly is used most by those who spread lies, that they themselves know to be lies!*
>
> ?

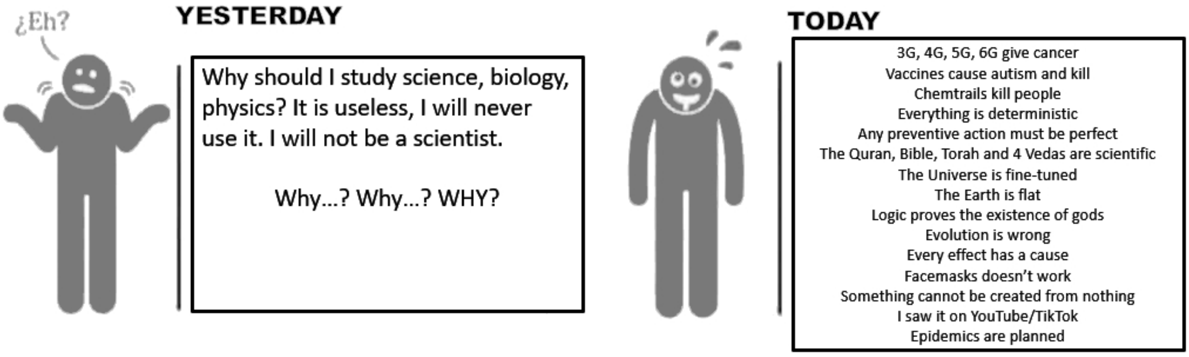
4.1.2 Who this textbook is intended for

This textbook is intended for anyone who is curious about Science and wants to learn more about the Natural world, its origin and mechanisms. It is written in a clear and concise way, making it easy to understand even if you have no prior knowledge of Science. Whether you are a student, a teacher, a hobbyist, or just a casual reader, this textbook will provide you with an engaging and accessible introduction to the fundamental concepts and methods of Science. You will explore detailed proofs from kindergarten level to PhD level on topics such as mathematics, probabilities, physics, engineering, management, finance, chemistry, biology, astronomy, geology, ecology, and more, using real life examples. You will also discover how science relates to everyday life and how it can help you solve problems and make informed decisions and to not be manipulated by the lies of religions/sects and charlatans. this textbook is not a comprehensive textbook, but rather a guide to spark your interest and inspire you to further your scientific knowledge and skills. By reading this textbook, you will develop a scientific mindset and a lifelong passion for learning.

Here are some specific examples of who might find this textbook useful:

- Students in any grades who are learning about science
- Teachers who are looking for a free resource to use in their classrooms
- Consultant who need to provide an unique rigorous source to their customers
- Adults who are interested in learning more about science
- Anyone who is curious about the world around them

This textbook is designed to be flexible and adaptable to your needs and preferences. You can use it as a self-study guide, a reference book, a supplement to your school curriculum, or a source of inspiration for further exploration.


¿Eh?
YESTERDAY
Why should I study science, biology, physics? It is useless, I will never use it. I will not be a scientist.
Why...? Why...? WHY?
TODAY
3G, 4G, 5G, 6G give cancer
Vaccines cause autism and kill
Chemtrails kill people
Everything is deterministic
Any preventive action must be perfect
The Quran, Bible, Torah and 4 Vedas are scientific
The Universe is fine-tuned
The Earth is flat
Logic proves the existence of gods
Evolution is wrong
Every effect has a cause
Facemasks doesn't work
Something cannot be created from nothing
I saw it on YouTube/TikTok
Epidemics are planned

By reading this textbook, on the philosophical point of view, you will typically learn and be exposed to the mathematical proofs and scientific experimental evidence of the following topics:

- Why there is something rather than nothing
- Why the Big Bang isn't an explosion that started at a point
- Why there are a dozens of different types of logics build on different axioms
- Where morality comes from
- That consciousness is only chemical
- Why mathematics are build in such a way to describe to real world so well
- Where does information comes from
- Where particles, atoms, stars, planets and galaxies are formed
- How did life arise from non-life
- Why the our universe isn't fine-tuned for life
- That the purpose of life is to follow the arrow of entropy
- Why time is an illusion in the framework of the Wheeler-DeWitt universe model
- Why everything around you is only a problem of probabilities
- Why scientists don't trust their intuition or reason but instead follow the evidence

4.1.3 How is this textbook structured

This science textbook is structured in a way that gradually increases in complexity. The first few chapters introduce basic concepts and principles, and then the later chapters build on these concepts to explore more advanced topics. The textbook is divided into 12 parts, each covering a major branch of science: mathematics, physics, computing, social science, finance and engineering. Each part consists of several chapters that introduce the key concepts, principles, and phenomena of that branch.

This structure allows readers to learn at their own pace. If you are new to science, you can start with the early chapters and then move on to the later chapters as you become more comfortable with the material. The chapters are organized in a quite logical sequence, building on the previous ones and preparing you for the next ones when possible.

Each chapter begins with an overview of the topic that will be covered. This overview provides a framework for understanding the material that follows. The chapter then proceeds to discuss the topic in more detail, using clear and concise language.

You can follow the textbook from start to finish, or you can choose the topics that interest you the most and skip the ones that you are already familiar with. The textbook also provides cross-references, glossaries, appendices, and indexes that help you navigate the textbook and find the information you need.

To help you understand the material, the textbook includes a variety of features, such as:

- The textbook includes many illustrations to help you visualize the concepts being discussed
- The textbook includes examples to help you understand how the concepts being discussed apply to the real world

I hope this helps you understand how this textbook is structured.

4.2. Methods

SCIENCE is the set of all systematic efforts (scrupulous observations and plausible assumptions until the evidence of the contrary) to acquire knowledge about our environment, to organize and synthesize them into testable laws and theories, whose main purpose is to explain the "how" of things (and NOT the why!) often by a four-step approach:

- What do we have?
- Where will we go?
- What is our goal?
- Does it fit the data?

Scientists have to submit their ideas and results to independent verification and replication of their peers[7] ("peer review"). They must abandon or modify their conclusions when confronted with more complete or different evidences. The credibility of Science is based therefore on this self-correcting mechanism and this is what still makes in the 121st century (holocene calendar) that Science is not the best tool (as we do not know what will exist in the future...) but is has been proven as being the best investigation method for truth (i.e. not the "philosophical truth", but the "truth" as the best explanation at our actual level of knowledge) in comparison for all other actual existing methods or beliefs. The history of science shows that this system works very well compared to all the others. In each area, progress has been spectacular. However, the system sometimes fails and has also to be corrected before small drifts accumulate.

The downside is that scientists are humans. They have the imperfections of all humans, and especially, vanity, pride, anger and conceit. Nowadays, it happens that many people working on the same topic for a given time develop a common faith and believe[8] they hold the truth. The leader of the faith is the Pope and distils his opinion. The Pope that plays the game, takes his miter and his pilgrim's staff to evangelize his fellow heretics. Until then, this makes smile. But, as in real religions, they are sometimes annoying to want to expand their opinion to those who do not believe. Some of these "churches" do not hesitate to behave like the Inquisition. Those who dare to express a different opinion are burned at every opportunity, during conferences, or at their place of work. Some young researchers, uninspired, prefer to convert to the dominant religion, to become clerics faster rather than innovative researchers or even iconoclasts. The great Pope write his Bible to disseminate his ideas, imposes it to read to students and newcomers. He formats then the thought of younger generations and ensures his throne. This is a medieval attitude that can block progress. Some Popes go so far that they believe be the pope in their specialization field automatically gives them the same throne in all other areas...

---

[7] Although peer review is an essential tool for maintaining the quality of scientific research, it is obviously not without flaws since it is performed by humans with all their usual biases. However, it remains the method with the lowest error rate to date.

[8] There are multiple definitions of "faith". Wikipedia gives: « *Confidence or trust in a person, thing, or concept. Belief in God or in the doctrines or teachings of religion.* ». The christian bible (Hebrew 11:1) gives: « *Now faith is confidence in what we hope for and assurance about what do not see.* »

This warning, and the reminders that will follow, must serve the scientist or any reader to ask himself by making good use of what we consider today as the good working/reasoning practices (we will discuss the principles of the Descartes method more below) to solve problems or develop theoretical models. And we do not use the models to build up some metaphysical system. We use the models to predict the outcomes of other measurements and to aid us in putting these to practical use!

For this purpose, here is a summary table that provides the steps that should be followed by a scientist who works in mathematics or theoretical physics (for definitions, see just below):

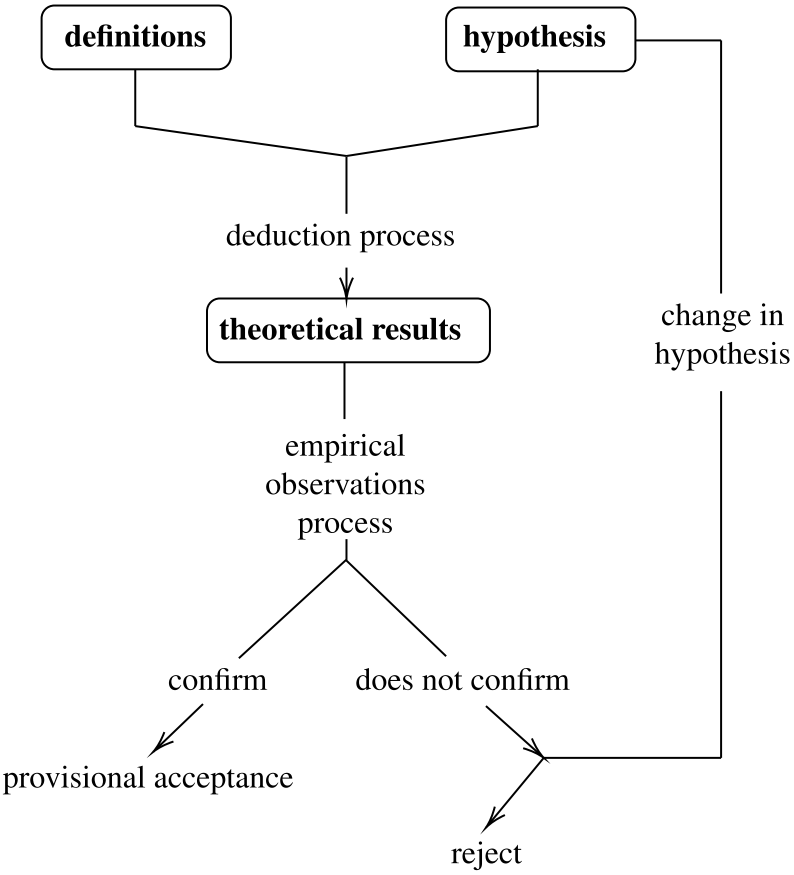
Mathematics	Physics
1. Expose formally or in common language the "hypothesis", the "conjecture" the "property" to prove (hypothesis are denoted H1., H2., etc. the conjectures CJ1., CJ2., etc. and the properties P1., P2., etc.).	**1.** Expose correctly in a formal or common language all the details of the "problems" to solve (problems are denoted P1., P2., etc.) and defines the vocabulary and notations rigorously.
2. Define the "axioms" (non-demonstrable, independent and non-contradictory) that will give the starting points and establish restrictions on the developments (the axioms are denoted A1., A2., etc.)[9]. In the same vein, the mathematicians define the specialized vocabulary related to mathematical operators which will be denoted by D1., D2., etc.	**2.** Define (or state) the "postulates" or "principles" or the "hypothesis" and "assumptions" (supposedly unprovable but supported by datas!) that will give the starting point and establish restrictions on the developments (typically, assumptions and principles are denoted P1., P2., etc. and assumptions H1., H2., etc. trying to avoid the notation confusion between postulates and principles)[10]
3. Once the Axioms laid, pull directly "lemmas" or "properties" whose validity follows directly and prepare the development of theorem supposed to validate departure hypothesis or conjectures (lemmas being denoted L1., L2., etc. and properties P1., P2., etc.).	**3.** Once the "theoretical model" (not to be confuse with the word "theory"...!) developed, check equations units for possible errors in the developments (such checks being denoted VA1., VA2., etc.).
4. Once the "theorems" (noted T1., T2., etc.) proved conclude on "consequences" (denoted C1., C2., etc.) and even properties (denoted P1., P2., etc.).	**4.** Search for borderline cases (including "singularities") of the model to verify its validity/behaviour intuitively (these borderline controls are denoted CL1., CL2., etc.).
5. Test the strength (robustness) or usefulness of the conjectures or hypothesis by proving the reciprocal of the theorem or by comparing them with other examples of mathematical well-know theories to see if form together a coherent structure (examples being denoted E1., E2., etc.).	**5.** Experimentally test the theoretical model obtained and submit work to compare with other independent research teams. The new model should provide experimental results and new never observed one (predictions to falsify). If the model is validated beyond reasonable doubt then it can get the official status of "theory".
6. Possible remarks may be shown in a hierarchically structured order and denoted R1., R2., etc.	**6.** Possible remarks may be shown in a hierarchically structured order and denoted R1., R2., etc.

Table 4.1: **Methodology for Maths & Physics developments**

Remark

The fact that a question can be phrased in a grammatically correct English sentence doesn't make it meaningful, or entitle it to our serious attention. Nor, the fact that a word exist in the dictionary (like the word "soul") make it real...

Proceed as in the above table is a possible workflow basis for people active in the field in mathematics or physics, or for any individual that wants to follow a rigorous critical thinking process[11]. Obviously, proceed cleanly and traditionally as above takes a little or much more time than doing things no matter how (this is why most teachers do not follow these rules, as they would otherwise not have enough time to cover their entire course program) and this is one of the reasons why science takes a majority of people outside of their comfort zone (as most people are looking to fix problems and interrogations in less than a few dozens of minutes).



It must also be known to the reader that we insist on the fact that real scientists should no have emotions behind the subjects they study or speak about. They have to only use evidence (facts based on data, peer review, reproducible experiences, consensus of scientific community) rather

---

[9] Sometimes "properties", "conditions" and "axioms" are confused while the concept of axiom is much more accurate and profound.

[10] The reader should not forget, however, that the validity of a model is not dependent on the realism of its assumptions but on the conformity of its implications with reality.

[11] To educate and exercise yourself or students on the topic of critical thinking we strongly recommend [432] and critical thinking tests like the Ennis-Weir test (free-writing test in which the test-taker evaluates, paragraph by paragraph, a given reasoned case) or the California Critical Thinking Disposition Inventory (CCTDI) that measures dispositions to use critical thinking skills like the Watson-Glaser Critical Thinking Appraisal (WGCTA).

than emotional, biased[12], subjective educational individual analysis that are not data driven. Obviously we may question the assumption that the used sensory probes detect something real or not? But professional scientists don't care about such a philosophical question. The only thing that matters (putting a part the fun of it...) is that these sensory probes give reliable, useful and consistent measures to test models that will help to build new tools for humanity!

Definitions (#4.2.2):

D1. The "reality" is the total set of phenomena that can be observed, measured, and tested, and whose existence is supported by public, reproducible empirical evidence.

D2. An "ontological fact" concerns the fundamental nature of reality and what exists, assumed to be independent of human perception and instruments.

D3. An "epistemological fact" relates to how we come to know or understand something - our knowledge, methods, and justifications.

> *What we observe is not Nature in itself but Nature exposed to our method of investigation.*
>
> Werner Heisenberg

quote 4.2:53.#7

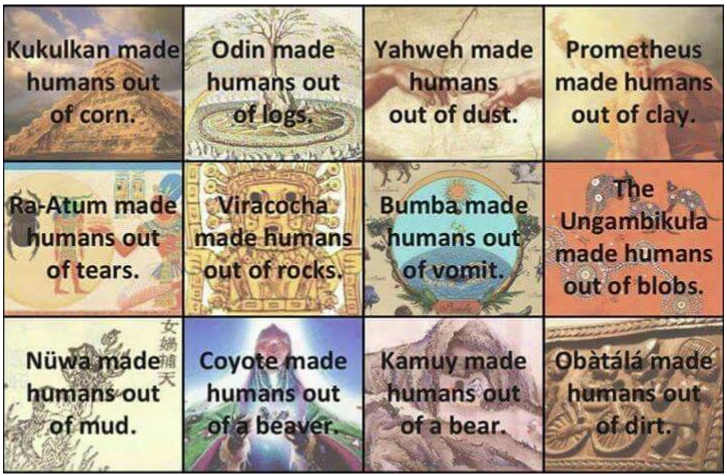

Kukulkan made humans out of corn.
Odin made humans out of logs.
Yahweh made humans out of dust.
Prometheus made humans out of clay.
Ra-Atum made humans out of tears.
Viracocha made humans out of rocks.
Bumba made humans out of vomit.
The Ungambikula made humans out of blobs.
Nüwa made humans out of mud.
Coyote made humans out of a beaver.
Kamuy made humans out of a bear.
Obàtálá made humans out of dirt.

Figure 4.3: **An example of absence of consensus in a small sample of religions**

> *Science doesn't set out to refute religions; science sets out to discover what is true. Religions are just collateral damage!*
>
> ?

quote 4.2:53.#8

---

[12] Among all major biases that we will introduce later, the Confirmation Bias - the tendency to search for, interpret, favour, and recall information that confirms or supports one's prior personal beliefs or values - is very likely the most common by scientifically illiterate people.

Notice also a funny shape of scientific 10 commandments:

1. The phenomena you will observe
   And never measures you will falsify
   (attention to the confirmation bias[13]!)
2. Hypothesis you will proposed
   That with experiment you will test
3. The experiment precisely you will describe and the data and algorithms you will provide
   Because your colleague will reproduce it
   (attention to the narrative discipline trap: the facts will be fitted to the desired results)
4. With your results
   A theory you will build
5. Parsimony you will use
   And the simplest hypothesis you will retain
6. Ultimate truth will never be (epistemic humility)
   And always you will search for the truth
7. From a non-refutable thesis you will refrain
   Because outside of the science it will remain
8. All failures will be like a success
   Because science can confirm but also invalidate
9. My authority I will not use (authority bias)
   To bias people opinions in fields where I have no proven expertise
10. The Archimedean Oath and the Scientific Publication Rules I will respect
    As science must be transparent and responsible

quote 4.2:54.#9

> ❞ *Science knows it doesn't know everything, otherwise it would stop.* ❞
>
> Dara Ó Brian

---

[13] It's very strongly advised to read our introduction to cognitive biases in the section of Decision Theory page 7381

Caution! It is very easy to make new physical theories by just aligning words. This is named "philosophy" or "rhetoric" and the Greeks thought of the atoms in this method. This can lead with a lot of luck to a true theory. Against it is much more difficult to make a "predictive theory", that is to say with equations that predict the outcome of an experiment (all physical theories must posit a correspondence between their mathematical apparatus and the physical world that they are attempting to describe).

Moreover many philosophers reinvent arguments physicists have long known to be wrong. We have heard philosophers worry about paradoxes physicists solved ages ago, and we have heard philosophers deduce how natural laws should be while ignoring how natural laws are. In short, there are unfortunately many philosophers who don't notice when they are out of their depth. The same can be said of physicists, though. Some physicists draw on philosophical arguments more frequently than they like to admit. It's easy enough however for us to discard philosophy as useful - because it is useless.

> ” *Science doesn't know everything, religion doesn't know anything!* ”
>
> L. Aron Nelson

quote
4.2:55.#10

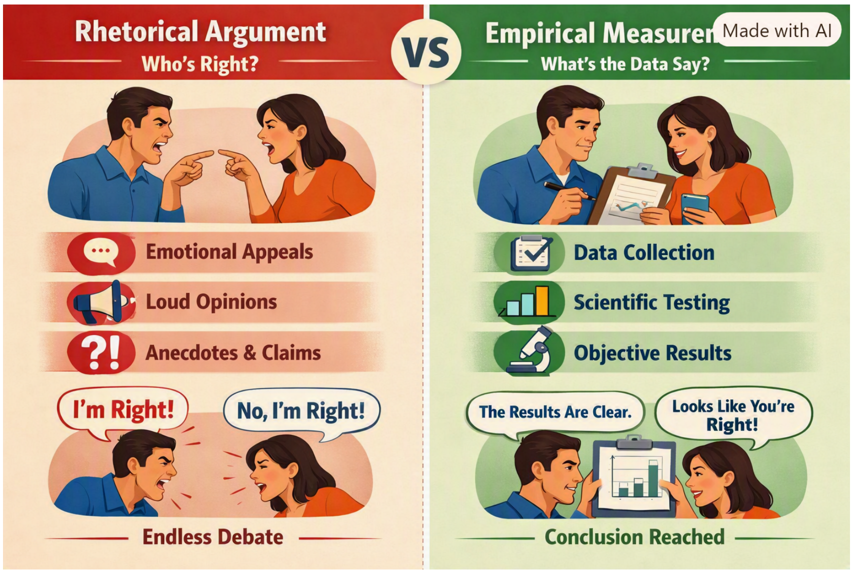
Remark

What separates mathematics and physics is that in mathematics, the hypothesis is always true. Mathematical discourse is not a proof of an external seeking truth, but a target of consistency. What should be correct is just the reasoning.

> ” *When people say "The reason you can't trust science is that scientists used to say <thing that turned out be wrong> but now they say <different thing>", they are actually explaining why science works so well!* ”
>
> Matt Blaze

quote
4.2:55.#11

When these rules are not respected, we speak of "scientific fraud" or of "intellectual fraud" (which often leads to being fired from his job but unfortunately we still not retired the diplomas when it happens). In general, scientific fraud itself comes in four main forms: plagiarism, fabrication of data and alteration of results unfavourable to the hypothesis, the omission of clear working hypotheses and collected datas, fallacious and biased arguments (remember that especially a rhetorical argument that isn't backed up by experimental evidence has no interest for professional scientists!).


Rhetorical Argument
Who's Right?
VS
Empirical Measurement
What's the Data Say?
Made with AI
Emotional Appeals
Loud Opinions
Anecdotes & Claims
I'm Right!
No, I'm Right!
Endless Debate
Data Collection
Scientific Testing
Objective Results
The Results Are Clear.
Looks Like You're Right!
Conclusion Reached

To these frauds we can also add behaviours that pose problems regarding to the quality of work or more specifically to ethics, such as submitting for example several times the same publication with only a few modifications, the omission of conflict of interest (financial, religious, political), the dangerous experiments, the non-conservation of primary data, etc.

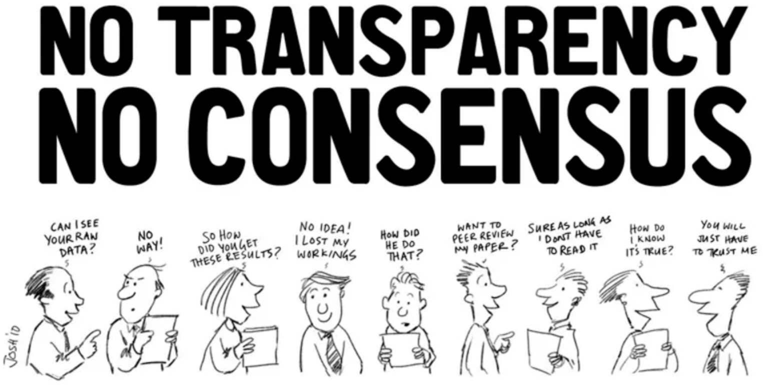

NO TRANSPARENCY
NO CONSENSUS
CAN I SEE YOUR RAW DATA?
NO WAY!
SO HOW DID YOU GET THESE RESULTS?
NO IDEA! I LOST MY WORKINGS
HOW DID HE DO THAT?
WANT TO PEER REVIEW MY PAPER?
SURE AS LONG AS I DON'T HAVE TO READ IT
HOW DO I KNOW IT'S TRUE?
YOU WILL JUST HAVE TO TRUST ME
JOSH 10

Figure 4.4: Source: http://cartoonsbyjosh.co.uk

4.2.1 Descartes' Method

Now we present the four principles of the Descartes' method which, as remind, is considered as the first scientist in history by his method of analysis:

P1. Never accept anything as true that I obviously knew to be such. That is to say, carefully avoid precipitation and to understand nothing more in my judgements than what would appear so clearly and distinctly to my mind, that I had no occasion to doubt.

P2. Divide each of the difficulties I have to examine into as many parts as possible (scrupulous observations and plausible hypothesis until evidence of the opposite), and that would be required to resolve them in the best way.

P3. Driving my thoughts in order, beginning with the simplest objects and easiest to know, to go up gradually by degrees to the knowledge of the most compounds, and even assuming the order between those who not naturally precede each other.

P4. Make everywhere so complete enumerations and so general reviews, that I'm sure not to omit anything.

> ”” *The Art and Science of asking questions is the source of all Knowledge.* ””
>
> Thomas Berger

quote 4.2:57.#12

Nullius in Verba

Nullius in verba (Latin for "on the word of no one" or "take nobody's word for it") is the motto of the Royal Society. It is an expression of the determination of Fellows to withstand the domination of authority and to verify all statements by an appeal to facts determined by reproducible experiments and by cautious scientific peer review[14].

> ”” *You are entitled to your opinion. But you are not entitled to your own facts.* ””
>
> Daniel Patrick Moynihan

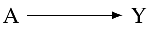
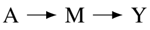
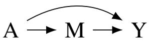
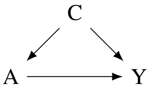
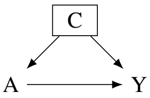
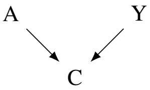
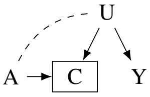
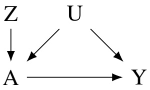
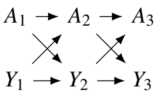
quote 4.2:57.#13

---

[14] The process of peer review generally applies to journal articles, but it is possible for a textbook to be peer reviewed as well. Although many textbooks go through some sort of editorial or review process, there is not an easy method for determining whether a textbook is peer reviewed. One method for locating peer reviewed textbooks is to take a look at textbook publications from university presses. Textbooks published by university presses almost always go through a process of peer review. Textbooks from university presses are typically written by faculty members are who are under immense pressure to produce authoritative scholarly literature. The process of peer review for university presses typically involves two or three independent referees who will initially review the manuscript. If the manuscript receives positive review, the university press will send it to their editorial board, who are all faculty members, for final review.

4.2.1.1 Blind studies

Scientific experiments[15] are designed to determine facts about our world using either "retrospective studies" based on the search of correlations by exploiting existing databases (hence looks backwards) or "prospective studies[16]" based on the search of causalities using controlled/randomized/double-blinded experiments (respectively frequently abbreviated RCT or RCDBT where the "T" stand for "Trial" instead of using the word "Experiment" or "Study"). But in complicated analyses, there's a risk that researchers will unintentionally skew results to match what they were expecting to find. To reduce or eliminate this potential bias, scientists apply a method known as "blind analysis".

Phenomenon	**Causal diagram**	**Structural equations**	**Description**
Causation		$y = f(a)$	A causes Y.
Mediation (full)		$m = f_1(a)$ $y = f_2(m)$	The effect of A on Y is fully mediated by M.
Mediation (partial)		$m = f_1(a)$ $y = f_2(a, m)$	The effect of A on Y is partially mediated by M.
Confounding		$a = f_1(c)$ $y = f_2(a, c)$	C is a common cause (or confounder) of A and Y.
Conditioning[1]		$a = f_1(c)$ $y = f_2(a, c)$	Analysis or study design is conditional on C (i.e. controlling for C, adjusting for C, etc.).
Collider		$c = f(a, y)$	C is a collider of A and Y, i.e. conditioning on C would induce a spurious association between A and Y.
Collider bias[1]		$c = f_1(a, u)$ $y = f_2(u)$	Y is independent of A, but conditioning on the collider C creates a spurious association between A and U (dashed line), opening a back-door path between A and Y.
Instrumental variable		$y = f_1(a, u)$ $a = f_2(z, u)$	Z is an instrumental variable for the exposure A, i.e. the association between Z and Y can be used to derive an estimate of the effect of A on Y.
Feedback loop		$a_2 = f_1(a_1, y_1)$ $a_3 = f_2(a_2, y_2)$ $y_2 = f_3(a_1, y_1)$ $y_3 = f_4(a_2, y_2)$	Example of a cyclical process or feedback loop, where the exposure A is affected by past values of the outcome Y and vice versa.

[1] Note that conditioning on the variable C is only represented in the causal diagram, not in the structural equations.
A: exposure. C: confounder or collider. M: mediator. U: unmeasured confounder. Y: outcome. Z: instrumental variable.

Figure 4.5: **Most common types of causal relationships in Science (source: [275])**

---

[15] This text is a copy/paste of an article written by Manuel Gnida at `http://www.symmetrymagazine.org/article/the-facts-and-nothing-but-the-facts`

[16] Prospective studies have usually fewer potential sources of bias and confounding than retrospective studies.

Definition (#4.2.3): A "randomized controlled trial" (RCT) is a form of scientific experiment used to neutralize factors not under direct experimental control by randomly allocating participants among compared treatments. This random allocation achieve sufficient conversion of confounding factors (confounding factors are those that may compete with the exposure of interest in explaining the effect of a treatment) as white noise to deliver a useful comparison of the treatments study.

Blind studies are probably best known from their use in clinical drug trials[17] (the term "triple-blinding" sometimes refers to this), in which patients are kept in the dark about - or blind to - whether they're receiving an actual drug or a placebo[19] or standard of care (indeed dummy pills, with no pharmacological activity at all, demonstrably improve health at a low statistical level of significance that is why double-blinded drug trials must use placebos or standard of care as controls to neutralize the placebo effect and permit the blinding of the treatment) but also the doctors themselves (to minimize behavioural bias) and the people who collects the data (to minimize the corruption), we speak then of "double-blinded randomised placebo controlled trial". This approach helps researchers judge whether their results stem from the treatment itself or from the patients' belief that they are receiving it. But the method is also use in Gastronomy tasting or in forensic laboratories as well.

Particle physicists and astrophysicists do blind studies, too (see [299])!!! The approach is particularly valuable when scientists search for extremely small effects hidden among background noise that point to the existence of something new, not accounted for in the current model. Examples include the much-publicized discoveries of the Higgs boson by experiments at CERN's Large Hadron Collider[20] (a double-blind data under wraps was done by the DZero and CDF groups of the Higgs experiment) and of gravitational waves by the Advanced LIGO detector.

«*Scientific analyses are iterative processes, in which we make a series of small adjustments to theoretical models until the models accurately describe the experimental data.*» says Elisabeth Krause, a postdoc at the Kavli Institute for Particle Astrophysics and Cosmology, which is jointly operated by Stanford University and the Department of Energy's SLAC National Accelerator Laboratory. «*At each step of an analysis, there is the danger that prior knowledge guides the way we make adjustments. Blind analyses help us make independent and better decisions*».

Return on experience (REX) shows as expected that blind analyses need to be designed individually for each experiment. The way the blinding is done needs to leave researchers with enough information to allow a meaningful analysis, and it depends on the type of data coming out of a specific experiment.

---

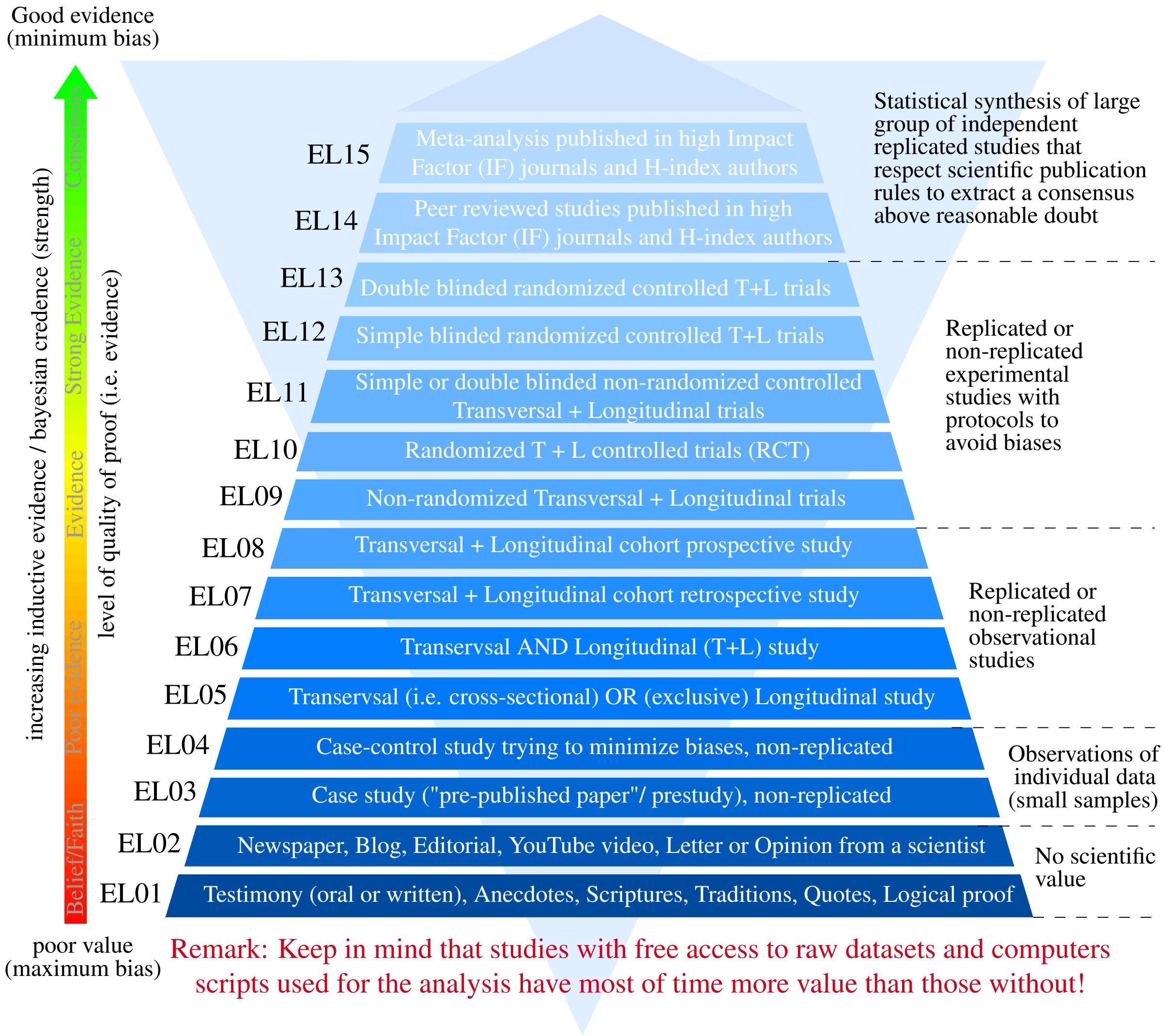
[17] The simplest trial design is a "single-arm trial". In this design, a sample of individuals with the targeted medical condition is given the experimental therapy and then followed over time to observe their response. In a "double-arm trial" half of the people will receive a "placebo" or "standard of care[18]" as reference (exposure group vs non-exposure group). In a "triple-arm trial" a third will receive a competitive treatment. In the general case we speak of "multi-arm trial".

[19] In the science of drugs, a "pure placebo" is a treatment without any active substance; an "impure placebo" is a pharmacologically active product but having no effect on the pathology treated, or whose efficacy has not been sufficiently demonstrated.

[20] If you doubt of the utility in every day life of the CERN, just keep in mind that it is in this fundamental research center that the world wide web has been invented, but also touch-screens, tracker balls, robust chips for nuclear power plants, bid data visualisation softwares, hybrid pixelated detectors, gas electron multiplier, superconductivity technologies for tokamaks and nuclear magnetic resonance scanner, etc.

A common approach is to base the analysis on only some of the data, excluding the part in which an anomaly is thought to be hiding. The excluded data is said to be in a "black box" or "hidden signal box[21]".

Take the search for the Higgs boson. Using data collected with the CERN's Large Hadron Collider (LHC) until the end of 12011 (holocene calendar), researchers saw hints of a bump as a potential sign of a new particle with a mass of about 125 gigaelectronvolts. So when they looked at new data, they deliberately quarantined the mass range around this bump and focused on the remaining data instead.


Good evidence
(minimum bias)
Consensus
Strong Evidence
Evidence
Poor evidence
Belief/Faith
increasing inductive evidence / bayesian credence (strength)
level of quality of proof (i.e. evidence)
EL15
Meta-analysis published in high Impact Factor (IF) journals and H-index authors
EL14
Peer reviewed studies published in high Impact Factor (IF) journals and H-index authors
EL13
Double blinded randomized controlled T+L trials
EL12
Simple blinded randomized controlled T+L trials
EL11
Simple or double blinded non-randomized controlled Transversal + Longitudinal trials
EL10
Randomized T + L controlled trials (RCT)
EL09
Non-randomized Transversal + Longitudinal trials
EL08
Transversal + Longitudinal cohort prospective study
EL07
Transversal + Longitudinal cohort retrospective study
EL06
Transervsal AND Longitudinal (T+L) study
EL05
Transervsal (i.e. cross-sectional) OR (exclusive) Longitudinal study
EL04
Case-control study trying to minimize biases, non-replicated
EL03
Case study ("pre-published paper"/ prestudy), non-replicated
EL02
Newspaper, Blog, Editorial, YouTube video, Letter or Opinion from a scientist
EL01
Testimony (oral or written), Anecdotes, Scriptures, Traditions, Quotes, Logical proof
Statistical synthesis of large group of independent replicated studies that respect scientific publication rules to extract a consensus above reasonable doubt
Replicated or non-replicated experimental studies with protocols to avoid biases
Replicated or non-replicated observational studies
Observations of individual data (small samples)
No scientific value
poor value
(maximum bias)

Remark: Keep in mind that studies with free access to raw datasets and computers scripts used for the analysis have most of time more value than those without!

Figure 4.6: **Scientific evidence hierarchy (assuming each level is made "at its best")**

The figure above[22] that applies to multiple fields of Science (physics, chemistry, biology, genetics, epidemiology, social science, economy, psychology, archaeology, astrophysics, astronomy, cosmology, epidemiology) illustrates in the lowest level of evidence EL01 the obvious fact that we should never trust the opinions and beliefs of scientists (even less "corporate scientists"...), physicians, doctors and engineers (these latter are not scientists by the way, see for example [500] and [190]!) - because science must not be driven by humans beliefs and opinions (!) - but only by the evidence provided by the **DATA** from reproducible independent experiments provided with their statistical analysis at a meta-analysis level inhigh Impact Factor (IF) peer reviewed papers with high H-index (HI) authors !

---

[21] Injecting false signals in scientific detectors is also a periodic common quality routine test in professional research to check the robustness of the scientific processes and the unbiased behavior of researchers.

[22] The huge majority of techniques and methods mentioned in each of the rows of the figure are proven and explained in the section Statistics starting from page 542 and in the section of Numerical methods starting from page 6316.

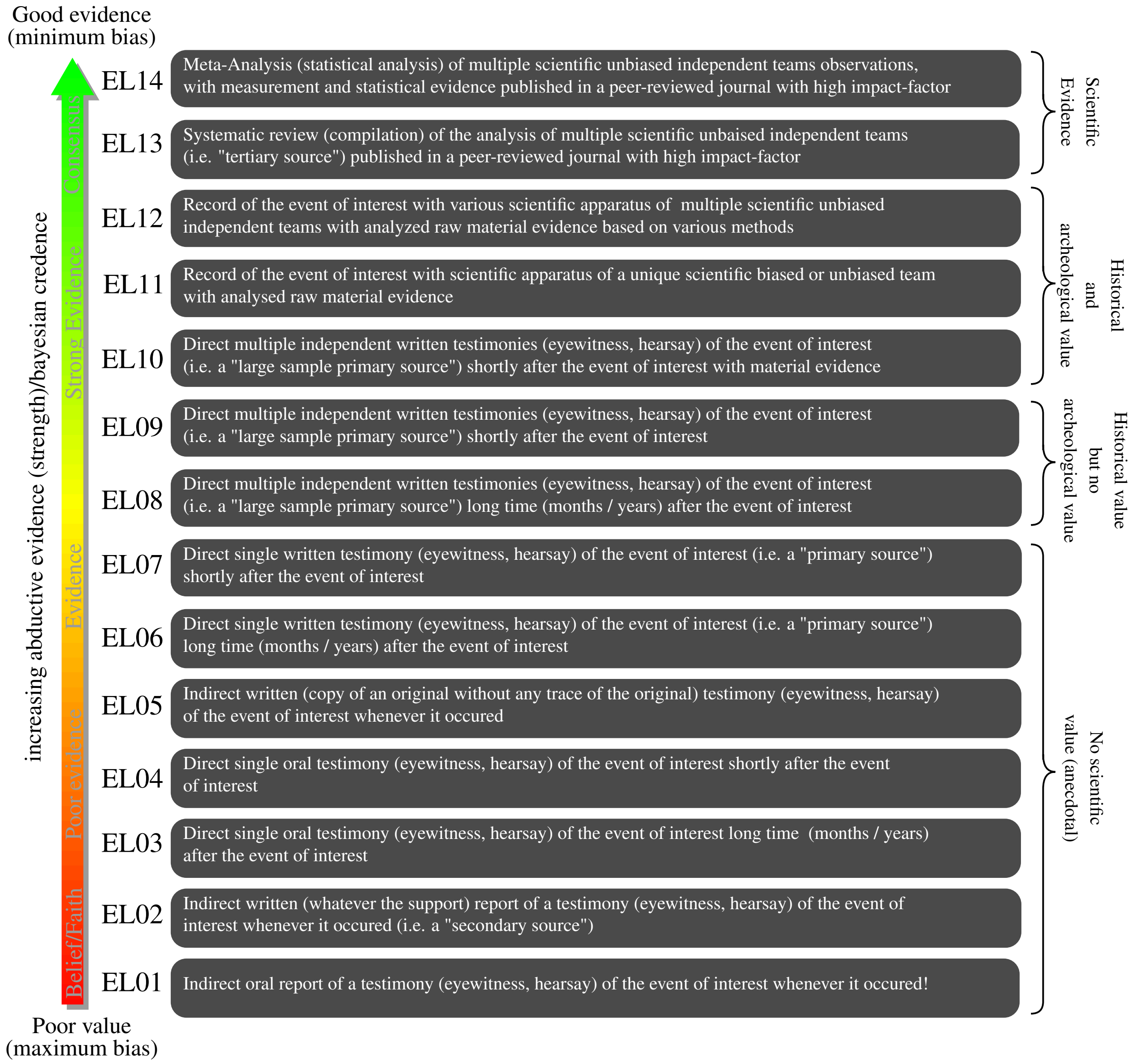
Caution! We can find on Internet plenty of professional looking websites administered by anonymous people (that try hard to hide who they are, in what country they live and what business they run...) compiling meta-analysis! However a website that is anonymous - having by definition a zero H-index - is obviously suspect especially when they cherry-pick only studies that goes in the direction of their political and religious agenda and don't provide the source code to reproduce the results. The same goes for some textbooks that want to pass themselves off as textbooks and published by editor with dubious catalogue and having no affiliation with a major university. In this early 121st century, the content of journal with an H-index below 500 and an Impact Factor below 10 must be taken with extreme precaution and not be considered as written by highly skilled professional scientists.

For example the above level of evidence doesn't applied to History as we cannot repeat or observe the past (at the opposite of astrophysics where the further we look, the further we go back into the life of the Universe we see!). In History the basic rules (undergraduate level) for example are to consider the three different type of sources below:

PRIMARY SOURCE	SECONDARY SOURCE	TERTIARY SOURCE
Definition: raw data; original source of information before it has been analyzed; written by the protagonist of the topic	**Definition:** sources that analyze, interpret or story tell primary data testimonials; they do not offer new evidence and don't constitute evidence; not original	**Definition:** Sources that compile data on a particular topic.
Characteristics: first-hand observation/account;protagonist wrote it;contemporary accounts of events; original dated document	**Characteristics:** interpretation of information; written after the event of interest; offer review or critique;	**Characteristics:** collections or lists of primary an secondary sources; reference works;finding tools for sources
Examples: interviews; speeches; diaries; birth certificates; coins; dated measurement original articles; originals written at the time; experiments account	**Examples:** biographies; journal articles; textbooks; commentaries; editorial; holy books; videos reproducing the events	**Examples:** encyclopedias; bibliographies; abstracts indexes; literature; reviews; library catalogs; databases; this textbook (Opera Magistris)

Table 4.2: **Reminder of the difference scale of Sources in Sciences**

But at a graduate and postgraduate level we have the following actual (nothing is fixed for ever) hierarchy of scientific evidence in the field of History (this hierarchy is rarely taught because they may hurt and offend many believers and even put in cognitive dissonance a significant proportion of professionals - teachers and professors - who have important anchor biases):


Good evidence
(minimum bias)
increasing abductive evidence (strength)/bayesian credence
Consensus
Strong Evidence
Evidence
Poor evidence
Belief/Faith
Poor value
(maximum bias)
Scientific Evidence
Historical and archeological value
Historical value but no archeological value
No scientific value (anecdotal)

EL14 Meta-Analysis (statistical analysis) of multiple scientific unbiased independent teams observations, with measurement and statistical evidence published in a peer-reviewed journal with high impact-factor

EL13 Systematic review (compilation) of the analysis of multiple scientific unbaised independent teams (i.e. "tertiary source") published in a peer-reviewed journal with high impact-factor

EL12 Record of the event of interest with various scientific apparatus of multiple scientific unbiased independent teams with analyzed raw material evidence based on various methods

EL11 Record of the event of interest with scientific apparatus of a unique scientific biased or unbiased team with analysed raw material evidence

EL10 Direct multiple independent written testimonies (eyewitness, hearsay) of the event of interest (i.e. a "large sample primary source") shortly after the event of interest with material evidence

EL09 Direct multiple independent written testimonies (eyewitness, hearsay) of the event of interest (i.e. a "large sample primary source") shortly after the event of interest

EL08 Direct multiple independent written testimonies (eyewitness, hearsay) of the event of interest (i.e. a "large sample primary source") long time (months / years) after the event of interest

EL07 Direct single written testimony (eyewitness, hearsay) of the event of interest (i.e. a "primary source") shortly after the event of interest

EL06 Direct single written testimony (eyewitness, hearsay) of the event of interest (i.e. a "primary source") long time (months / years) after the event of interest

EL05 Indirect written (copy of an original without any trace of the original) testimony (eyewitness, hearsay) of the event of interest whenever it occured

EL04 Direct single oral testimony (eyewitness, hearsay) of the event of interest shortly after the event of interest

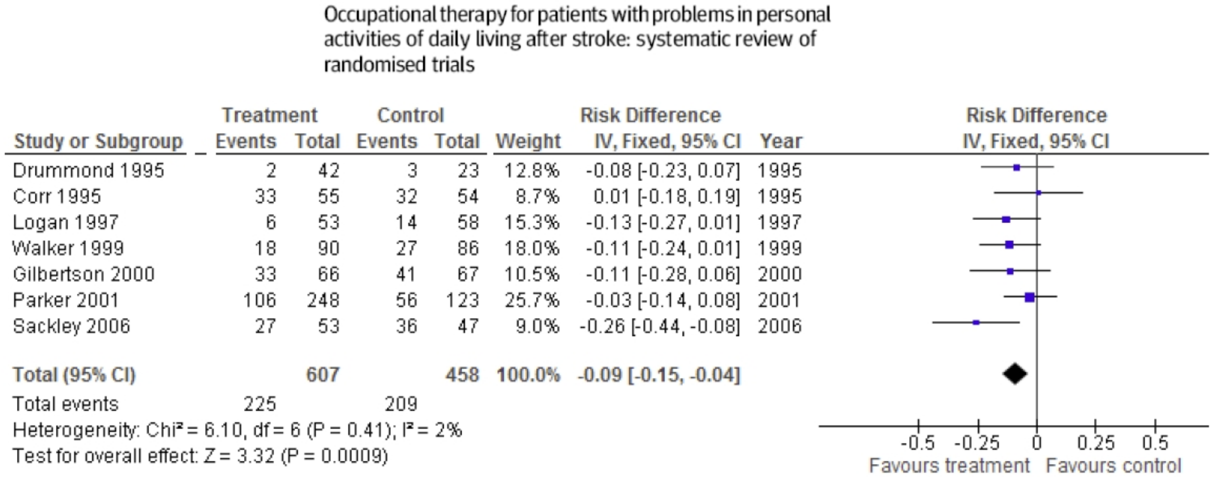
EL03 Direct single oral testimony (eyewitness, hearsay) of the event of interest long time (months / years) after the event of interest

EL02 Indirect written (whatever the support) report of a testimony (eyewitness, hearsay) of the event of interest whenever it occured (i.e. a "secondary source")

EL01 Indirect oral report of a testimony (eyewitness, hearsay) of the event of interest whenever it occured!

Figure 4.7: **Levels of Evidence in a social science like History**

As we can see from the both scales of evidence above, an important safety check to ensure that scientific claims are trustworthy is "replication"! If novel findings from scientific research can be reproduced, it means they are more likely to be valid. On the other hand, if the findings cannot be replicated they are likely to be invalid. Therefore, no single study (not even a randomized controlled trial) can be considered to be strong evidence – it is merely indicative. For this reason, we prefer meta-analyses. A meta-analysis is a summary of studies in which statistical analysis techniques are used to combine the results of individual studies to get a more accurate estimate of the effect. Thus, irrespective of the type of question, meta-analyses are often the most appropriate research designs. In other words, it is the body of evidence that is important – not single studies!



Occupational therapy for patients with problems in personal activities of daily living after stroke: systematic review of randomised trials

Study or Subgroup	Treatment Events	Treatment Total	Control Events	Control Total	Weight	Risk Difference IV, Fixed, 95% CI	Year
Drummond 1995	2	42	3	23	12.8%	-0.08 [-0.23, 0.07]	1995
Corr 1995	33	55	32	54	8.7%	0.01 [-0.18, 0.19]	1995
Logan 1997	6	53	14	58	15.3%	-0.13 [-0.27, 0.01]	1997
Walker 1999	18	90	27	86	18.0%	-0.11 [-0.24, 0.01]	1999
Gilbertson 2000	33	66	41	67	10.5%	-0.11 [-0.28, 0.06]	2000
Parker 2001	106	248	56	123	25.7%	-0.03 [-0.14, 0.08]	2001
Sackley 2006	27	53	36	47	9.0%	-0.26 [-0.44, -0.08]	2006
Total (95% CI)		**607**		**458**	**100.0%**	**-0.09 [-0.15, -0.04]**	
Total events	225		209				

Heterogeneity: Chi^2 = 6.10, df = 6 (P = 0.41); I^2 = 2%

Test for overall effect: Z = 3.32 (P = 0.0009)

Figure 4.8: **Special example of a meta-analysis in the field of clinical trials with concepts that we will detail without exception in this entire textbook (source: [344])**

Remark

Testimonials or personal anecdotes have almost no values if the sample size is small and biased and if there is no direct way to measure the related event directly! This is why Historical Evidence based on only a few testimonial written in a unique thousands of year old book have no scientific value (even if there is a dozens of such books with concordant testimonials).

Those who quote Nietzche:

> *There are no facts, only interpretations!*
>
> Friedrich Nietzsche

quote 4.2:63.#14

also don't understand that they are at the level EL01 because is that claim itself a fact...????

The LHC (CERN) scientists used that data to make sure they were working with a sufficiently accurate model. Then they "opened the box" (unblinding) and applied that same model to the untouched region. The bump turned out to be the long-sought Higgs particle[23].

That worked well for the Higgs researchers. However, as scientists involved with the Large Underground Xenon (LUX) experiment reported at the workshop, the "black box" method of blind analysis can cause problems if the data you're expressly not looking at contains rare events crucial to figuring out your model in the first place.

LUX has recently completed one of the world's most sensitive searches for WIMPs - hypothetical particles of dark matter[24], an invisible form of matter that is five times more prevalent than regular matter. LUX scientists have done a lot of work to guard LUX against background particles-building the detector in a cleanroom, filling it with thoroughly purified liquid, surrounding it with shielding and installing it under a mile of rock. But a few stray particles make it through nonetheless, and the scientists need to look at all of their data to find and eliminate them.

---

[23] Blinding and unblinding of experimental data analysis should be mandatory in all field of Science to avoid methodological biases!

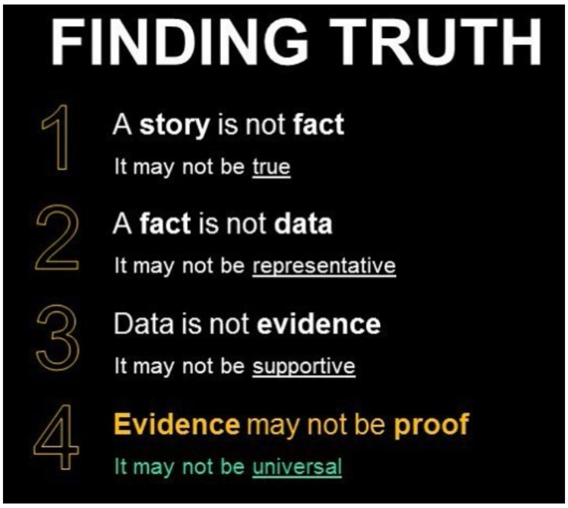
For that reason, LUX researchers chose a different blinding approach for their analyses. Instead of using a "black box", they use a process called "salting".

LUX scientists not involved in the most recent LUX analysis added fake events to the data - simulated signals that just look like real ones. Just like the patients in a blind drug trial, the LUX scientists didn't know whether they were analysing real or placebo data. Once they completed their analysis, the scientists that did the "salting" revealed which events were false.

A similar technique was used by LIGO scientists, who eventually made the first detection of extremely tiny ripples in space-time (for a naive and technical definition of "space-time" see the section of Special Relativity at page 5708) named "gravitational waves".

Not everyone in the scientific community is convinced that blinding is necessary. Blind analyses are more complicated to design than non-blind analyses and take more time to complete. Some scientists participating in blind analyses inevitably spend time looking at fake data, which can feel like a waste.

Typically some quite famous medical doctors and engineers (all are mathematics haters because they were very bad in this field during their studies and don't understand how to apply advanced statistics nor how to read the corresponding analytical results!) castigates the "totally randomized double-blind religion" and a medicine/engineering that has passed from the humanist hands of caregivers/inventors to the cold ones of statisticians and "methodologists". For sure (...) we know since long time that the rule of thumb, feelings and beliefs work much better than statistical method...



FINDING TRUTH

1 A **story** is not **fact**
It may not be true

2 A **fact** is not **data**
It may not be representative

3 Data is not **evidence**
It may not be supportive

4 **Evidence** may not be **proof**
It may not be universal

---

[24] Since June 12025 (holocene calendar) it is known thanks to FRB (Fast Radio Bursts) that approximately 76% of the Universe's baryonic matter lies in the IGM (InterGalactic Medium). About 15% resides in galaxy halos, and a small fraction is burrowed in stars or amid cold galactic galactic gas ([109]).

The reader must also keep in mind that we never claimed in this textbook that meta-analysis or RCT (randomized control trial) are the golden rule of evidence based science. We just claim that they are the tools that seems actually - at the time we write these lines - to give the best results. Obviously they are not perfect (as the humans conducting the experiments are also not perfect beings...) and sometimes they have failed but anyone criticizing meta-analysis and RCT should provide quantitative based evidence that there are other methods (mentioning which one) that performs statistically significantly better!

> ❞ *Most people don't really want the truth; they just want reassurance that what they already believe, is the truth.* ❞
>
> L. Aron Nelson

quote 4.2:65.#15

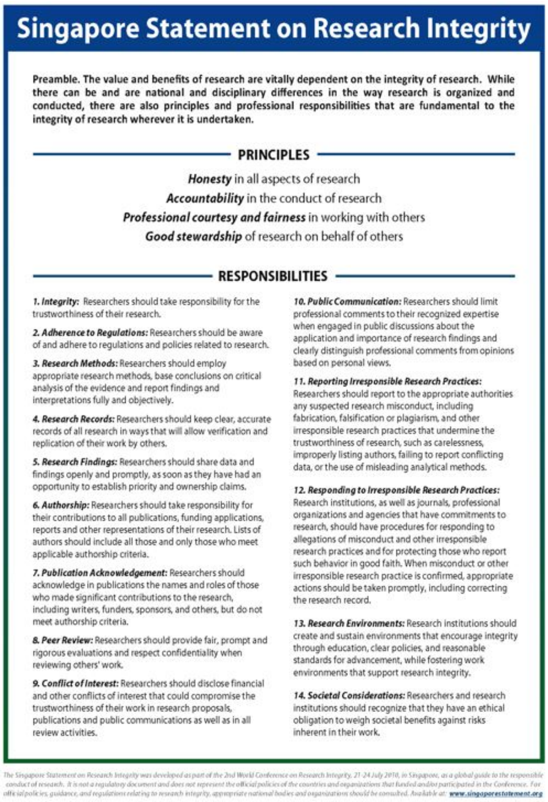
4.2.2 Research Integrity and Engineering/Scientific Ethics

Humans are trying at multiple levels to impose international standards of research integrity and engineering/scientific ethics.

Let us introduce three of the most famous one:

4.2.2.1 Singapore Statement on Research Integrity

The first one is the Singapore Statement on Research Integrity, drafted at the Second World Conference on Research Integrity, which took place in Singapore from July 21 to 24, 12010 (holocene calendar), that is an important step toward promoting ethical conduct among scientists around the world. The 340 conference attendees included scientists, journal editors, academic and industry leaders, and representatives from government funding agencies and publishers from over 51 countries. Nanyang Technological University, the National University of Singapore, the Singapore Management University, and the Agency for Science, Technology, and Research hosted the gathering, with support from Singapore's Ministry of Education and National Research Foundation. The Singapore statement was drafted by conference co-chairs, Nicholas Steneck (University of Michigan) and Tony Mayer (Nanyang Technological University), and the incoming chair for the next World Conference, Melissa Anderson (University of Minnesota). In contrast to the First World Conference, which took place in Lisbon, Portugal in 12007 (holocene calendar) and focused on misconduct issues, the goal of the Second World Conference was to make a concerted effort to promote global research integrity. The Singapore Statement is the fruit of this endeavour.



Singapore Statement on Research Integrity

Preamble. The value and benefits of research are vitally dependent on the integrity of research. While there can be and are national and disciplinary differences in the way research is organized and conducted, there are also principles and professional responsibilities that are fundamental to the integrity of research wherever it is undertaken.

PRINCIPLES

Honesty in all aspects of research
Accountability in the conduct of research
Professional courtesy and fairness in working with others
Good stewardship of research on behalf of others

RESPONSIBILITIES

1. Integrity: Researchers should take responsibility for the trustworthiness of their research.

2. Adherence to Regulations: Researchers should be aware of and adhere to regulations and policies related to research.

3. Research Methods: Researchers should employ appropriate research methods, base conclusions on critical analysis of the evidence and report findings and interpretations fully and objectively.

4. Research Records: Researchers should keep clear, accurate records of all research in ways that will allow verification and replication of their work by others.

5. Research Findings: Researchers should share data and findings openly and promptly, as soon as they have had an opportunity to establish priority and ownership claims.

6. Authorship: Researchers should take responsibility for their contributions to all publications, funding applications, reports and other representations of their research. Lists of authors should include all those and only those who meet applicable authorship criteria.

7. Publication Acknowledgement: Researchers should acknowledge in publications the names and roles of those who made significant contributions to the research, including writers, funders, sponsors, and others, but do not meet authorship criteria.

8. Peer Review: Researchers should provide fair, prompt and rigorous evaluations and respect confidentiality when reviewing others' work.

9. Conflict of Interest: Researchers should disclose financial and other conflicts of interest that could compromise the trustworthiness of their work in research proposals, publications and public communications as well as in all review activities.

10. Public Communication: Researchers should limit professional comments to their recognized expertise when engaged in public discussions about the application and importance of research findings and clearly distinguish professional comments from opinions based on personal views.

11. Reporting Irresponsible Research Practices: Researchers should report to the appropriate authorities any suspected research misconduct, including fabrication, falsification or plagiarism, and other irresponsible research practices that undermine the trustworthiness of research, such as carelessness, improperly listing authors, failing to report conflicting data, or the use of misleading analytical methods.

12. Responding to Irresponsible Research Practices: Research institutions, as well as journals, professional organizations and agencies that have commitments to research, should have procedures for responding to allegations of misconduct and other irresponsible research practices and for protecting those who report such behavior in good faith. When misconduct or other irresponsible research practice is confirmed, appropriate actions should be taken promptly, including correcting the research record.

13. Research Environments: Research institutions should create and sustain environments that encourage integrity through education, clear policies, and reasonable standards for advancement, while fostering work environments that support research integrity.

14. Societal Considerations: Researchers and research institutions should recognize that they have an ethical obligation to weigh societal benefits against risks inherent in their work.

The Singapore Statement on Research Integrity was developed as part of the 2nd World Conference on Research Integrity, 21-24 July 2010, in Singapore, as a global guide to the responsible conduct of research. It is not a regulatory document and does not represent the official policies of the countries and organizations that funded and/or participated in the Conference. For official policies, guidance, and regulations relating to research integrity, appropriate national bodies and organizations should be consulted. Available at: **www.singaporestatement.org**

4.2.2.2 European Code of Conduct for Research Integrity

The second one is the European Code of Conduct for Research Integrity that serves the European research community as a framework for self-regulation across all scientific and scholarly disciplines and for all research settings.

The European Code of Conduct for Research Integrity is a PDF of twenty A4 pages where only seven pages contains the code. The people interested to download it for free can just click on the following link:

https://www.allea.org/wp-content/uploads/2017/05/ALLEA-European-Code-of-Conduct-for-Research-Integrity-2017.pdf

ALLEA
All European Academies

integrity

The European
Code of Conduct for
Research Integrity
REVISED EDITION

4.2.2.3 Archimedean Oath

The last one is inspired by the Hippocratic Oath, a group of students of the École Polytechnique Fédérale de Lausanne in 11990 (holocene calendar) developed an oath of Archimedes expressing the responsibilities and duties of the engineer and technician. It was taken in various versions by other European engineering schools and could serve as basic inspiration oath for scientific researchers (however a few points are missing like as in medicine[25], to be struck off the order of scientists in the event of serious ethical error).

"Considering the life of Archimedes of Syracuse which illustrated as of Antiquity the ambivalent potential of the technique, considering the responsibility increasing for the engineers and scientists with regard to the men and nature, considering the importance of the ethical problems that the technique and its applications raise, today, I pledge following and will endeavour to tend towards the ideal which they represent:

01 I will practice my profession for the good of the people, in the respect of the Human Rights and of the Environment.

02 I will recognize, being as well as possible informed to me, the responsibility for my acts and will not discharge me to in no case on others.

03 I will endeavour to perfect my professional competences.

04 In the choice and the realization of my projects, I will remain attentive with their context and their consequences, in particular from the point of view technical, economic, social, ecological... I will pay a detailed attention to the projects being able to have fine soldiers.

05 I will contribute, in the measurement of my means, to promote equitable relationships between humans and to support the development of the countries lower-income group.

06 I will transmit, with rigour and honesty, with interlocutors chosen with understanding, any information important, if it represents an asset for the company or if its retention constitutes a danger to others. In the latter case, I will take care that information leads to concrete provisions.

07 I will not let myself dominate by the defense of my interests or those of my profession.

08 I will make an effort, in the measurement of my means, to lead my company to take into account the concerns of this Oath.

09 I will practice my profession in all intellectual honesty, with conscience and dignity.

10 I promise it solemnly, freely and on my honour."

---

[25] Let us recall that contrary to a well spread misconception that doctors and surgeons or any medical personnel are obviously not scientists (we leave it to the reader to do an extensive fact checking if that surprises him)!

Remark

Obviously, any engineer that has worked for Fortune 500 corporations knows that such an oath cannot be applied when the board committee and the managers are themselves not engineers and don't care about this oath (at the opposite of medicine where most of time all the hierarchy is made of doctor!). So the only possibility to apply this oath for engineers working is such structures is either to negotiate the respect of this oath during the hiring process (good luck...!) or to resign the contract... (what most engineers don't do and therefore they don't respect this oath...).

Sadly this oath should be completed with the "Münich Declaration of the Duties and Rights of Journalists" (year 11971 according to holocene calendar). That is, the essential duties of the scientist in gathering, reporting on and commenting on data consist in:

- Respecting the truth no matter what consequences it may bring about to him, and this is because the right of the public is to know the truth.
- Defending the freedom of information, of commentaries and of criticism.
- Publishing only such pieces of information the origin of which is known or – in the opposite case – accompanying them with due reservations; not suppressing essential information and not altering texts and documents.
- Not making use of disloyal methods to get information, photographs and documents.
- Feeling obliged to respect the private life of people.
- Correcting any published information which has proved to be inaccurate.
- Observing the professional secrecy and not divulging the source of information obtained confidentially.
- Abstaining from plagiarism, slander, defamation and unfounded accusations as well as from receiving any advantage owing to the publication or suppression of information.
- Never confusing the profession of journalist with that of advertiser or propagandist and not accepting any consideration, direct or not, from advertisers.
- Refusing any pressure and accepting editorial directives only from the leading persons in charge in the editorial office. Every journalist worthy of this name feels honoured to observe the above-mentioned principles; while recognising the law in force in each country, he does accept only the jurisdiction of his colleagues in professional matters, free from governmental or other interventions.

4.2.3 Scientific Publication Rules (SPR)

It is impossible to have a constructive debate or analysis if the basis material is unusable, unavailable or poorly defined. Sadly still in the 121st century (holocene calendar) it is quite easy to find Nobel prize peer-review scientific publications[26] (PRSP) that are scientifically unusable (furthermore there is a huge issue with profit oriented scientific journals that reject replication studies most of the time considering them as "boring", i.e. non-money profitable...). This is why we recall here the basic publication rules to follow to be accepted by a scientific peer review committee:

01 Use of LaTeX for the writing of the publication

02 All redaction files and raw data files must have ISO 9660 compliant names

03 The publication should have a GUID (a unique Digital Object Identifier - DOI - alike code)

04 Put the history of major and minor versions of the publication (eg: v3.6 r58) with dates (ISO 8601 date/time format) and respective status (eg: working paper, preprint, submitted, peer review, accepted, published, retracted) and related taxonomy field keywords at the beginning

05 Put the experiment (development) period date (ISO 8601 date/time format)

06 The typographical rules of the language in which the publication is written must be strictly respected

07 Write an abstract (problem statement, indication of methodology, hypotheses, protocol, main findings and conclusions)

08 Write an introduction (including literature review)

09 All measurement[27] units and mathematical notation must follow ISO 80000 standards

10 Use the "principle of precaution[28]" (use of conditional)

11 Use "reactive responses", that is to say the make the confrontations between hypotheses / data, hypotheses / facts, hypotheses / observations

12 Any borrowing or reproduction of a text from an external source (human or algorithmic) must be placed between quotation marks and referenced

---

[26] Some studies get published with no peer review at all, even studies from Nobel Prizes (...), are named "predatory publishers" flood the scientific literature with journals that are essentially fake, publishing any author who pays (see an exhaustive online list on `https://beallslist.net`).

[27] Measurement by the way must be recordable, reproducible and answer to one or multiple identified causalities.

[28] The of expressions like "...(we) believe...", "....(we have) faith...", "...(we) think..." in the conclusions are forbidden.

⑬ Use, when available, "leverage factors" to give substance and credit to the work by making reference to other corresponding publications on the same subject[29]

⑭ Material and Methods should be described in details and new vocabulary defined with precision. For theoretical papers, they should provide a link (URL) or reference where the full detailed proof can be found (if detailed proof is omitted in the original publication!). For experiments they should provide the detailed randomized double-blinded protocol[30]

⑮ Put high resolution print-screens of charts (including the mandatory measurement errors and confidence/prediction intervals visible on the charts with the source code to generate them for reproducibility!) or photos

⑯ Write the results and for experimental data always provide a statistical analysis to show if the effect seems significant or not (effect sizes[31], fluctuation intervals, confidence interval, averages, medians, standard deviations, standard errors, sample sizes, kurtosis, bayes factor, skewness, Risk Ratio (RR), Odds Ratio (OR), Absolute Risk Reduction (ARR) and if the p-value is communicated then the power of the test must also always be communicated and be at least around 80% to have a scientific value!)

⑰ Calculate the propagation of errors of measurement instruments

⑱ Write the precautional conclusion to avoid HARKing (Hypothesizing After Results Are Known)[32] and always remind that a single study is not conclusive but indicative AND also that the paper may be retracted if it contains methodological errors

⑲ Remove from the publication any explicit personal political position agenda (unscientific statements like "socialism is a mortal enemy"), social beliefs (unscientific statements like "there is only two genders anyway") or religious beliefs (unscientific statements like "peace be upon him" or "may ... have mercy" or "for the sake of..." or "prayers and thoughts")

⑳ Give access to the raw data in a non-proprietary format to the scientific community[33]

㉑ Give access to the scripts/codes (of Open Source softwares!) used for data analysis to the scientific community[34]

㉒ Give access to the LaTeX sources of the publication to the scientific community (the sources can therefore contain in the format of LaTeX comments many additional details on the experiment or on the mathematical proofs)

---

[29] This also the very important step of "personal review", that is to say a personal analysis of several tens / hundreds of scientific publications and that you have made one critical analysis that you use to build your own argument.

[30] Double-blinded to avoid typical scandals like, among many others well-known examples, the Jacques Benveniste affair...

[31] This indicator has for pros of not being dependent of the sample size.

[32] The conclusion for experimental results (reject null hypothesis or not) should be written before (!) the experiment is run and not changed afterwards to avoid human cognitive biases.

[33] There are international standards for this purpose like Study Data Tabulation (SDTM), Analysis Data Model (ADaM) and Standard for Exchange of Nonclinical Data (SEND)

[34] To avoid case like the famous (and shameful) Reinhart-Rogoff error...

㉓ Provide the exact version (with possibly minor version!) of the software used for the experimentation[35] (calculations, writing and others) as well as devices/machines[36]

㉔ Put the bibliography with the references (ISO 690 - Numerical) and give access to the corresponding BibTeX file freely (check at the same time that the articles cited really exist and have not been retracted)

㉕ Cite equivalent studies for meta-analysis[37]

㉖ For meta-analysis weighted confidence intervals, the Harbord's test, the Begg rank correlation test, the Egger test, the heterogeneity test (chi-square), the Macaskill test and the funnel plot must be provided

㉗ Acknowledgements and % financial support of each sponsor must be detailed (competing interests[38], funding sources)

㉘ Submit the paper to the peer review committee[39] (in single or double blind way[40])

㉙ List all actors (with position, grade from which university, e-mail[41] and religious beliefs) and peer reviewers (only name for that latter) of the publication

Any publication that doesn't respect at least one of this rule cannot be considered as a "scientific" (i.e. "serious") publication! Many of the points above also applies to video content (typically TEDx videos or YouTube videos - or any other platform - where the speaker in a comfortable monologue[42]... - doesn't cite the sources and meta-analysis when arguing or exposing its "personal experience", "opinions", or "expertise").

The list of constraints above leads us to the obvious conclusion that just because you're reading an article about science doesn't mean it's a scientific article!

---

[35] Ideally you should keep a virtual machine with the configuration, the software version and the libraries for subsequent reproducibility.

[36] The experiment must be sufficiently detailed to be identically reproducible (or it is also necessary to specify the factors which would make the experiment likely to be non-reproducible).

[37] If there are no equivalent studies, then no meta-analysis are possible, then the results and conclusions don't reach any scientific consensus for recall!

[38] Not only industrial and economic, but also religious like working for a non-secular university!

[39] Notice that if all scientists followed these rule, peer review committees, with their bias and mood influence, would not be necessary anymore as the validation of a scientific paper could be quite easily automated.

[40] "single blind" is that the peer reviews doesn't know the name of the authors, "double blind" is that neither the authors nor the reviewers know each others' identities.

[41] And eventually with gender type, country of origin, religious belief, and birth year for statistical purposes. Like: Albert Einstein (Researcher, PhD Physics ETHZ, a.einstein@ethz.ch, M, CHE, Deist, 11879)

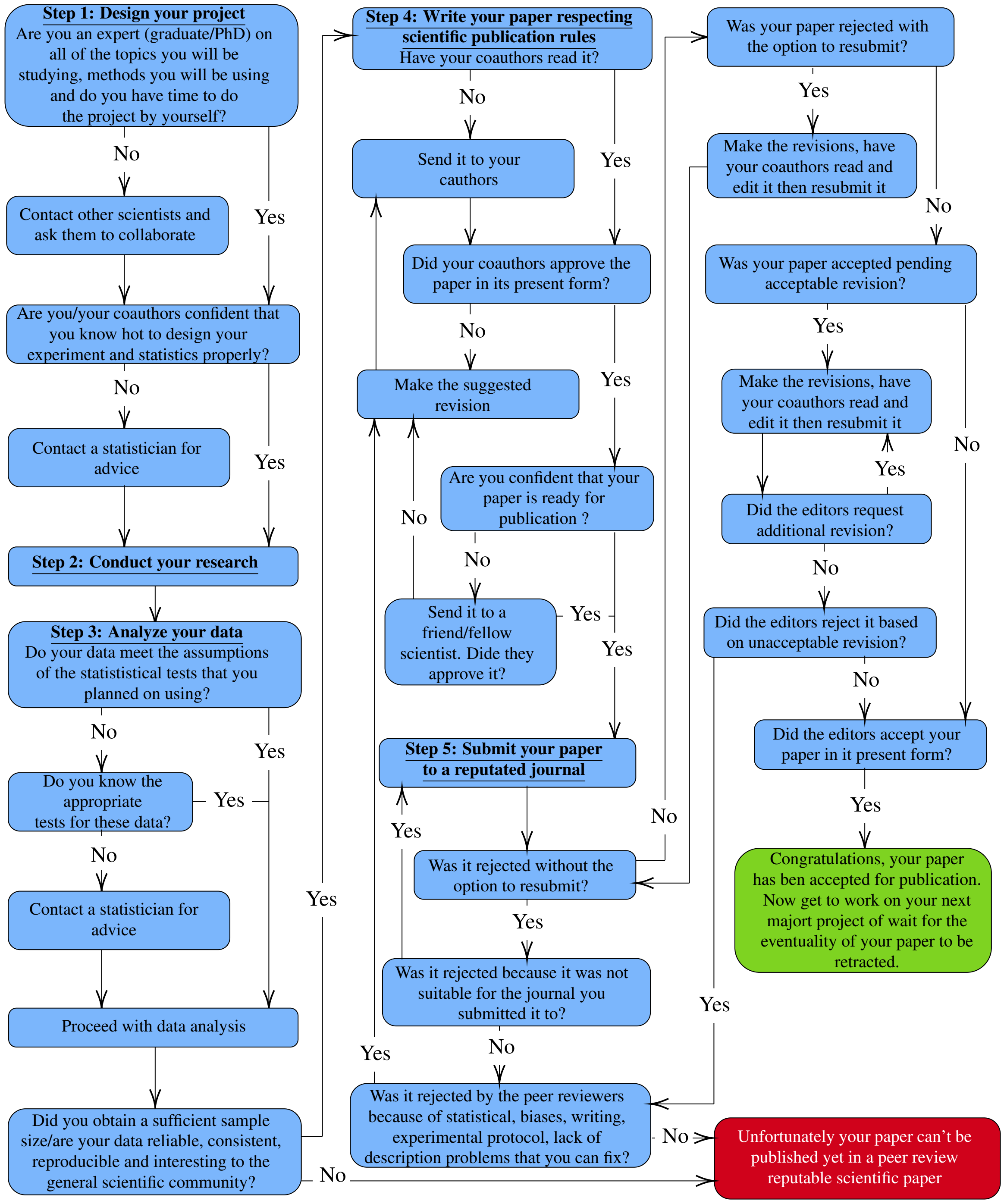
[42] People should not trust monologues - whether on radio, television or any social network - because there are no experts to counter the possible wrong arguments or ill-defined concepts that may lead a part of the audience to wrong speculative interpretations! In addition, humans under the stress of knowing that they are recorded are naturally prone to vocabulary errors and that's without counting the more than 200 cognitive biases in the brain which sometimes lead to the erroneous simplification of complex thoughts... !

Remark

There is an equivalent good list of 29 rules that has been which was proposed in 11996 (holocene calendar) dedicated to scientific systematic review reports! The last version is named "PRISMA 2020" where "PRISMA" stand for: *Preferred Reporting Items for Systematic Reviews and Meta-Analyses*. For more see: `http://prisma-statement.org`.

A photograph or video can be regarded as scientific evidence only if it meets the same methodological standards as empirical measurements: the phenomenon must be independently recorded by multiple qualified scientific teams, modelled in a way that yields predictive outcomes, subjected to rigorous statistical analysis, authenticated through forensic examination of the raw, unaltered files to exclude any computer-generated manipulation, and ultimately published in peer-reviewed journals in accordance with the scientific publication rules given above!

How To Publish Peer Reviewed Literature


Step 1: Design your project
Are you an expert (graduate/PhD) on all of the topics you will be studying, methods you will be using and do you have time to do the project by yourself?
No
Yes
Contact other scientists and ask them to collaborate
Are you/your coauthors confident that you know hot to design your experiment and statistics properly?
No
Yes
Contact a statistician for advice
Step 2: Conduct your research
Step 3: Analyze your data
Do your data meet the assumptions of the statistical tests that you planned on using?
No
Yes
Do you know the appropriate tests for these data?
Yes
No
Contact a statistician for advice
Proceed with data analysis
Did you obtain a sufficient sample size/are your data reliable, consistent, reproducible and interesting to the general scientific community?
Yes
No
Step 4: Write your paper respecting scientific publication rules
Have your coauthors read it?
No
Yes
Send it to your cauthors
Did your coauthors approve the paper in its present form?
No
Yes
Make the suggested revision
Are you confident that your paper is ready for publication ?
No
No
Yes
Send it to a friend/fellow scientist. Dide they approve it?
Yes
Yes
Step 5: Submit your paper to a reputated journal
Yes
Was it rejected without the option to resubmit?
No
Yes
Was it rejected because it was not suitable for the journal you submitted it to?
Yes
No
Was it rejected by the peer reviewers because of statistical, biases, writing, experimental protocol, lack of description problems that you can fix?
No
Unfortunately your paper can't be published yet in a peer review reputable scientific paper
Was your paper rejected with the option to resubmit?
Yes
Make the revisions, have your coauthors read and edit it then resubmit it
No
Was your paper accepted pending acceptable revision?
Yes
Make the revisions, have your coauthors read and edit it then resubmit it
Yes
No
Did the editors request additional revision?
No
Did the editors reject it based on unacceptable revision?
No
Yes
Did the editors accept your paper in it present form?
Yes
Congratulations, your paper has ben accepted for publication. Now get to work on your next majort project of wait for the eventuality of your paper to be retracted.

How To Publish A Blog/Internet/Social Network Post

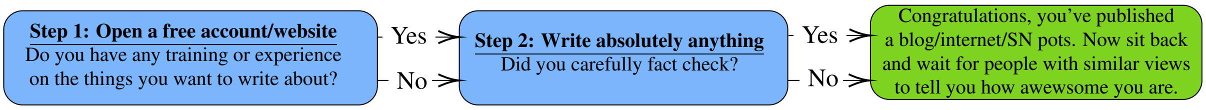

Step 1: Open a free account/website
Do you have any training or experience on the things you want to write about?
Yes
No
Step 2: Write absolutely anything
Did you carefully fact check?
Yes
No
Congratulations, you've published a blog/internet/SN pots. Now sit back and wait for people with similar views to tell you how awesome you are.

Figure 4.9: **How to publish peer review literature (source: https://www.thelogicofscience.com)**

The fact of being published in a peer-reviewed journal, however, does not guarantee the quality of an article! This is indeed an interesting fact that deserves to be underlined again and studied. Is it surprising ? Obviously not! All "publishers" know that the "reviewers" who judge the quality of articles submitted to journals are themselves researchers, caught up in the spiral of their own work and teachings, which have neither the desire nor the material possibility of check step by step all the calculations of an article in theoretical physics or all the references of an article in human sciences. This is not their mission. The vast majority of clearly erroneous articles are rejected by the system (the authors of the failed hoaxes do not suck and no one will know how many were foiled). Some, however, pass through the filters. It is obviously regrettable but perfectly well known to each concerned community. That of the hard sciences is not spared and notorious charlatans have managed to publish in respected and recognized journals. The hard sciences have not naturally been disqualified! These articles have simply been ignored for the vast majority of them (so unfortunately not all and especially those that do not detail all mathematical developments!): Neither read nor cited.

Remark

Even if there is a consensus at a given time between scientists (in 12021 according to holocene calendar, there was an estimate of approximately 9 million scientists around the World according to UNESCO), a unique oriented study (which can be very important) can be used to influence the opinion of mainstream media, governments and people. This is why a study must always be repeated, peer-reviewed and meta-analysed by independent teams and laboratories.

The existing H-index and IF (Impact Factor) index already used in the scientific publications field could be completed by an "SPR index" (Scientific Publication Rules) that gives the percentage of the number of rules listed above respected in a scientific article to give the level of scientific rigour of its content.

Caution! Some people think that a "scientific consensus" refers to a large group of scientists who agree in majority that something is "true" (i.e. have enough evidence to be actually considered as the most accurate model) at a given time. In reality, a scientific consensus is a large body of scientific studies that all agree with and support each other ("consensus of data"). The agreement among the scientists themselves is simply a by-product of the consistent evidence.

A well known example of non-existing consensus are religions. Indeed, if someone argue that as the statistics don't lie the christian god must exist as it is the most followed religion in the world with 2 billion Christians and that 2 billion people can't be wrong, you can recall this same person that as there is 7 billion people in the World, the 5 other billion that not believe in the christian god cannot be wrong as... statistic don't lie... Same if you merge Muslims and Christians together, then only 55% of the people in the World believe in a unique god and 55% is statistically not enough to reach the scientific consensus that is at a threshold level of 95%...

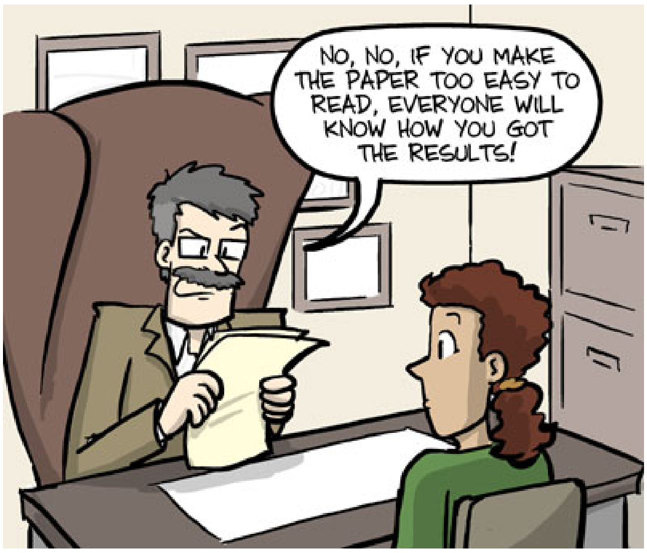
> ❞ *If a religion had even one true evidence beyond reasonable doubt, there would be no other religions!*

quote 4.2:75.#16

”

Remark

It is possible to logically provide evidence of the non-existence of any gods with certain attributes, by showing an inconsistency between those attributes and either the definition of the god or other established facts. For many examples of this, see [372].


NO, NO, IF YOU MAKE THE PAPER TOO EASY TO READ, EVERYONE WILL KNOW HOW YOU GOT THE RESULTS!

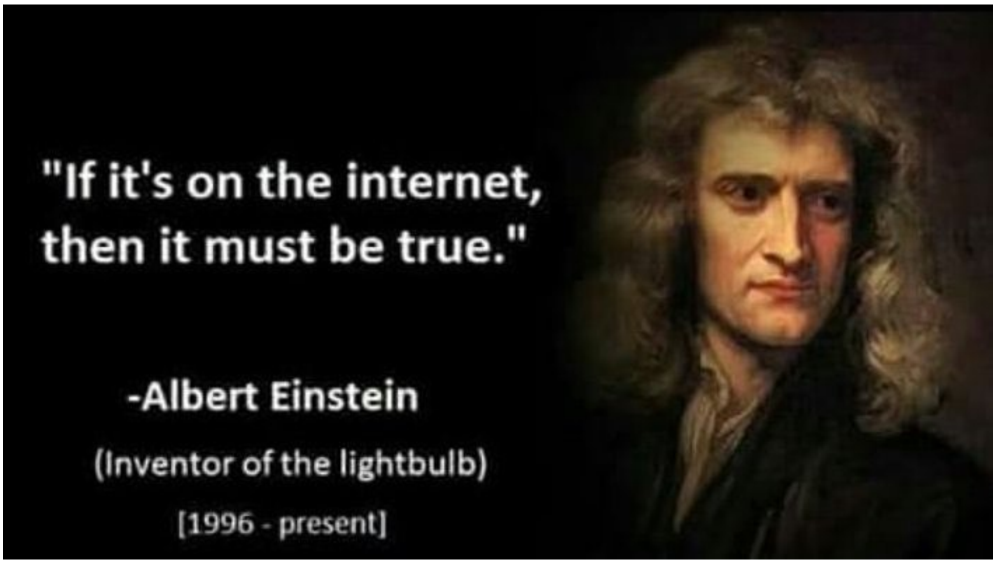
It is then easy to understand why Internet web pages and YouTube videos (or any other similar platform) are not a reliable scientific sources according to the above protocol as:

1. The peer-reviewers names are the huge majority of time not indicated
2. Contributor/Editors are anonymous and can therefore not be identified (typically an issue of Wikipedia)
3. The mathematical details are not provided (or even worst, there is not equations given at all!) so it is hard or even impossible to check by yourself if the reasoning is accurate
4. The experiment exact protocol is not given so it is impossible to know if the results are fake or real
5. No sources or cross-references are given
6. The content is not in a reliable format (a video or a web page are not perennial and protected[43] sources)
7. The new presented theoretical models predict indeed what the previous one do, but doesn't predict anything new and is therefore not falsifiable

---

[43] In the 121st century (holocene calendar) a PDF for example should be protected against edition and should also be electronically signed

8. The speaker on the video makes assumption that are not falsifiable (reference to god or to theories who mathematical details are not provided)
9. etc.


"If it's on the internet, then it must be true."
-Albert Einstein
(Inventor of the lightbulb)
[1996 - present]

The strength of scientific evidence produced by different types of studies (for instance systematic reviews, meta-analysis, randomised control trials, observational research, animal studies, cell studies and expert opinions) can vary.

Remark

The world is not systematically in line with our desires, our hopes, our prejudices, our philosophical aspirations, nor even with our apprehensions or our anxieties. It is not because it would be super cool to be able to move objects by the mere thought that this is necessarily possible. Likewise, it is not because he would be nice that a treatment for a rare disease exists that a therapeutic claim relating to this pathology is necessarily true. Our dreams, our hopes and our imagination are precious attributes of our humanity, which can put us in great danger if we forget that the associated intellectual productions are not due to us by reality. In the same vein, one thing is not necessarily true because it is conceivable that it is.

This infographic will help you understand the advantages and limitations of different types of scientific evidence:

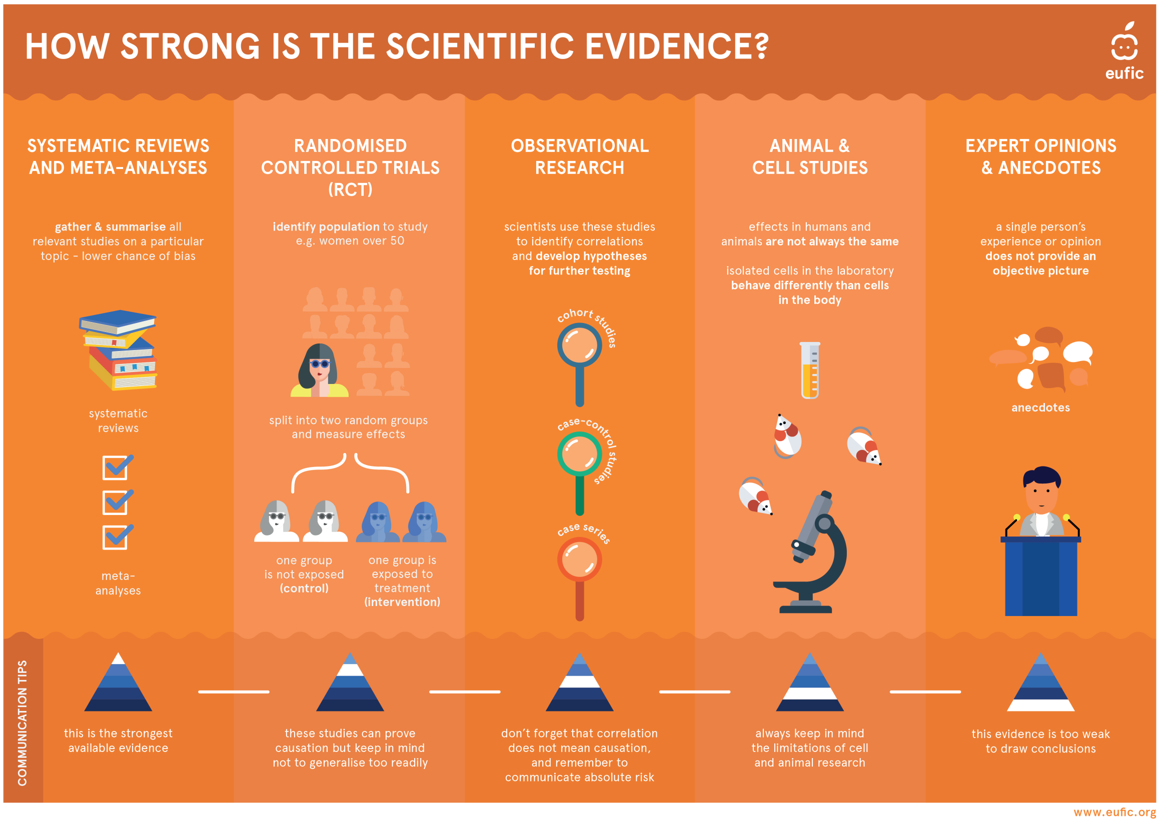

HOW STRONG IS THE SCIENTIFIC EVIDENCE?
eufic
SYSTEMATIC REVIEWS AND META-ANALYSES
gather & summarise all relevant studies on a particular topic - lower chance of bias
systematic reviews
meta-analyses
RANDOMISED CONTROLLED TRIALS (RCT)
identify population to study e.g. women over 50
split into two random groups and measure effects
one group is not exposed (control)
one group is exposed to treatment (intervention)
OBSERVATIONAL RESEARCH
scientists use these studies to identify correlations and develop hypotheses for further testing
cohort studies
case-control studies
case series
ANIMAL & CELL STUDIES
effects in humans and animals are not always the same
isolated cells in the laboratory behave differently than cells in the body
EXPERT OPINIONS & ANECDOTES
a single person's experience or opinion does not provide an objective picture
anecdotes
COMMUNICATION TIPS
this is the strongest available evidence
these studies can prove causation but keep in mind not to generalise too readily
don't forget that correlation does not mean causation, and remember to communicate absolute risk
always keep in mind the limitations of cell and animal research
this evidence is too weak to draw conclusions
www.eufic.org

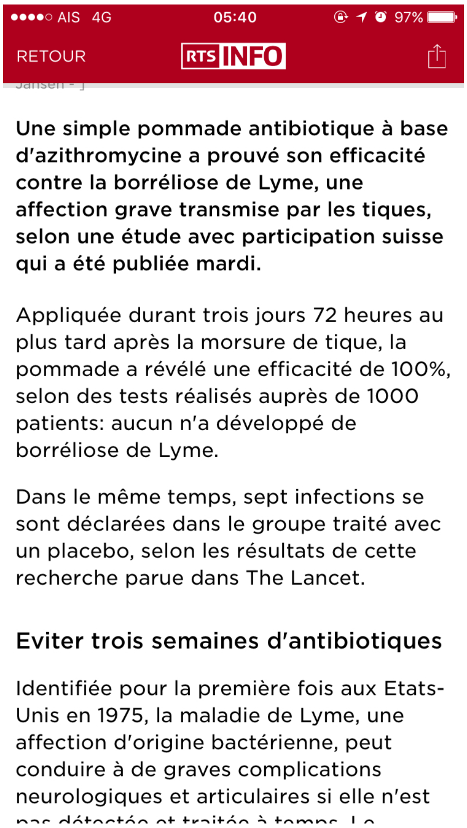
Figure 4.10: **Different types of scientific evidence (source: EUFIC)**

Anyway, we have to keep in mind that the popularity of a idea has no a priori relevance; the arguments must relate to the quality of the evidence, not its notoriety!

4.2.4 Scientific Mainstream Media communication

The reader of mainstream media or also (even worst!) social networks must never trust a scientific study if the reference and peer-reviewed paper is not given as link[44] (and if the paper does not respect the scientific publication rules that we have introduced earlier above page 70!). The study must also not be taken as "absolute truth" by the reader if there is a consensus of the scientific community on only... A UNIQUE ... study/publication[45] at a given time. The only way to be almost sure is to read the study itself and check if that latter respects the previous scientific publication rules, and this even more when you know that most mainstream medias are World champions in fall into the trap of cognitive biases (see section Game and Decision Theory page 7381) and data fallacies (see section Statistics page 1342).

A typical example of such an issue is a news that was taken by many international mainstream media on the Lyme borreliosis disease as following:



Une simple pommade antibiotique à base d'azithromycine a prouvé son efficacité contre la borréliose de Lyme, une affection grave transmise par les tiques, selon une étude avec participation suisse qui a été publiée mardi.

Appliquée durant trois jours 72 heures au plus tard après la morsure de tique, la pommade a révélé une efficacité de 100%, selon des tests réalisés auprès de 1000 patients: aucun n'a développé de borréliose de Lyme.

Dans le même temps, sept infections se sont déclarées dans le groupe traité avec un placebo, selon les résultats de cette recherche parue dans The Lancet.

Eviter trois semaines d'antibiotiques

Identifiée pour la première fois aux Etats-Unis en 1975, la maladie de Lyme, une affection d'origine bactérienne, peut conduire à de graves complications neurologiques et articulaires si elle n'est pas détectée et traitée à temps. Le

Figure 4.11: Swiss TV publication about Lyme borreliosis treatment the 12017-01-08 (source: RTS App)

In summary what the "scientific journalist" (humm humm... I think it must be a new intern in fact...), of one of the main National Swiss Television (so a TV that has enough money to investigate correctly any news... at least in theory... in a country that assess to be number one in almost everything...), has published is a very bad (catastrophic) interpretation of the original scientific publication. The above article report that: «*...a treatment applied during* 3

---

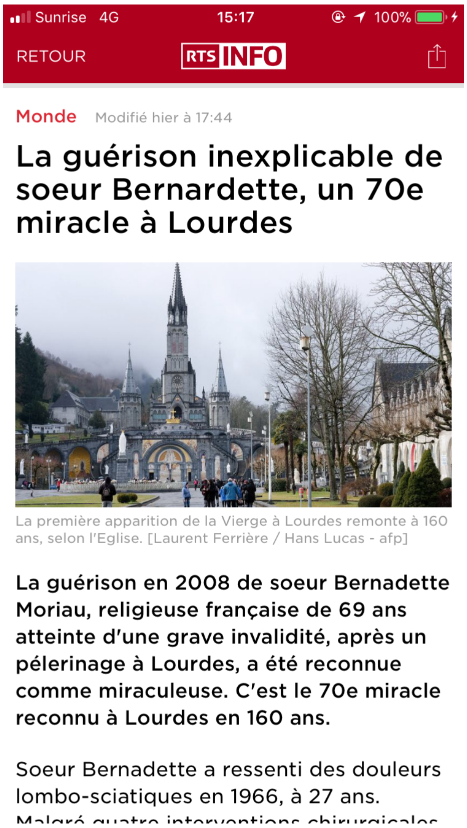
[44] And this is even more true for every information or brain-candy statistics as we will see later during our studies of percentages.

[45] Keep in mind that even a broken clock is right twice a day...

days not later than 72 hour after the bite of the tick has revealed and efficiency of 100%...». The most funny thing is that this article is provided by the Swiss Telegraph Agency (and afterwards relayed by Swiss TV) that says stupidly to be 100% accurate (this is the result of a country that gives a journalist accreditation after only 50 days of training...).

In reality (if medias did have read the publication until the end...) the study was stopped after 8 weeks and it has been shown that the treatment has no better effect than a placebo...

A second typical and recurrent mistake of the mainstream media is the confirmation bias (we will see the study of the biases later) of which here is a boring and ashamed example so much it is repeated (we may really think they do such errors on a given purpose ...):



Monde Modifié hier à 17:44

La guérison inexplicable de soeur Bernardette, un 70e miracle à Lourdes

La première apparition de la Vierge à Lourdes remonte à 160 ans, selon l'Eglise. [Laurent Ferrière / Hans Lucas - afp]

La guérison en 2008 de soeur Bernadette Moriau, religieuse française de 69 ans atteinte d'une grave invalidité, après un pélerinage à Lourdes, a été reconnue comme miraculeuse. C'est le 70e miracle reconnu à Lourdes en 160 ans.

Soeur Bernadette a ressenti des douleurs lombo-sciatiques en 1966, à 27 ans.
Malgré quatre interventions chirurgicales

Figure 4.12: **Swiss TV publication about a miracle in Lourdes the 12018-02-12 (source: App RTS/AFP)**

Of course, any person with a minimum of general knowledge can quite simply check through existing meta-analysis that "miracles[46]" are also occurring in hospitals and that in terms of "rates" of remission, Lourdes does not no better than mere chance in comparison to hospitals scattered all over the planet relative to this type of observation.

---

[46] A "miracle" is the term used by many people when they do not know how to explain an observed phenomenon. Eclipses were an example of a "miracle" not so long ago... If something can happen and does, it's hardly miraculous. So, a miracle must be something that can't happen. But, if a miracle does happen, then clearly, it can happen. And therefore it wasn't a miracle to begin with. So if life is full of miracles then you don't understand Nature and probabilities! Miracles have always been the snare of the ignorant and the refuge of the ambitious...

> *Everything is a sign or miracle when you don't understand Physics.*

quote 4.2:80.#17

It is therefore again shameful that one of the main Swiss national channel (a service that has enough money to properly investigate all information before relaying it ... at least in theory ... in a country that exclaims to be the number one in almost everything ...) has published biased information and this is even more shameful that this type of information comes from the AFP (Agence France Presse)...

We give also another similar problematic example in the section of Population Dynamics that occurred during the COVID-19 pandemic in France in 12021 (holocene calendar) at page 7345.

4.2.4.1 Social Networks

About social networks and scientific communication ... A priori we could see as well on Facebook, YouTube, Twitter, Instagram, TikTok during discussions or debates that:

- If we simplify the scientific discourse (thinking that I may be a good idea...) on some hot topics[47] we can be quickly enough be accused of distorting reality (this is the problem of indeed of simplification...!) or even to be condescending. And if you point out that you have simplified to make scientific illiterate people able to understand your argumentation, you will probably be accused of ad hominem attack. This is a situation afterwards where it is difficult or even impossible to restore confidence.
- If we do not simplify the scientific discourse (by using the vocabulary and quantitative methods of the field), or if we communicate the links to the scientific studies[48] or theories themselves, we are accused of hiding the truth or make statistics lie (or even to be condescending again!) thanks to an abscond vocabulary and the use of too technical terms and tools. The result is even worse when access the scientific papers are not in free access!
- If we are talking about a subject on which we have no expertise or diploma, or which is not our field of activity, we get "repacked" quickly (rightly!)
- On some groups and social networks accounts, some messages from a discussion may be deleted by administrators, this does not permit the possibility to support arguments or counter-arguments or just completely biases the discussion because some answers disappear or just simply never appear (some stakeholders or messages are typically hidden or deleted by the administrator).
- A small percentage of people are very well educated but may be biased because their are indoctrinated since childhood and even worst (because quite difficult to detect)... some are just trolls (with postgraduate curriculum) who enjoy to trigger people on social networks

---

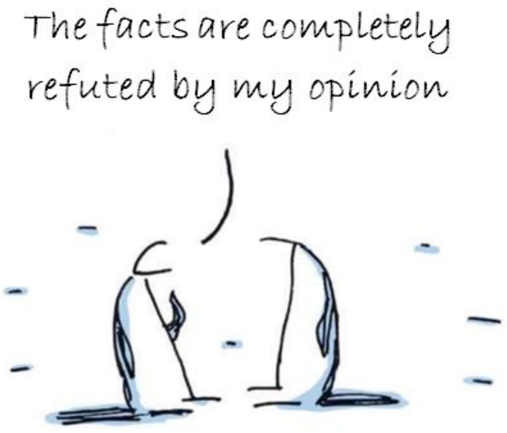
[47] Topics where typically people will tell you that: "the facts are denied by my opinion" ...

[48] Note that it seems to happen regularly that when links to the studies or meta-analysis are provided, there is almost always people to say that they are financed by lobbyists, or the links they were chosen in the sense of the defended arguments, or even that as not all the links towards all the studies of the world are provided then the given one are not representative...

just for the fun of reading the answers or... for the pleasure to analyse by curiosity the answers of people to their trolling.

- Statistically, 2.2% of open social networks users like Facebook, TikTok, Snapchat, Instagram (this excludes automatically social networks like ResearchGate obviously!) have an IQ below 70 points (and even some with higher IQ suffer of a sort of psychological paedomorphosis or of childhood indoctrination). In a country like France, that has in 12021 (holocene calendar) forty millions of daily active users, that makes $880,000$ users... and therefore quite a lot of uneducated, scientifically illiterate and non-rational people (taking into account that these kind of people are most of time jobless and therefore have much more to spend on social networks than PhD holder). That's why the mainstream social networks are frequented by plenty of WDKS representatives (World Dunning-Kruger Society).
- Finally keep in mind that all social networks do not support LaTeX, so it is impossible to have real scientific discussion (that is to say... using mathematical equations or chemical formulas).

The goal here is not to provide a scientific solution to these problems (it would be relevant, however, to conduct some studies on the subject ...!). However ... tricks that seem to work quite well are blocking comments on the content published by scientists or by the scientific community. To intervene in exchanges only if and only if you are invited to communicate as an external expert (and not to intervene on your own initiative).


The facts are completely refuted by my opinion

Finally, let us mention the spurious common argumentation techniques that are particularly flagrant on social networks (and in other media in general too ...) from those who are impervious to the scientific method and statistical analyses and simulations and that are voluntarily (or not?) inspired from the principles of war propaganda that the historian Anne Morelle has stated:

1. We do not want war (i.e. we do not want any "changes")
2. The opponents are the only one responsible for the war (it is he who forces the "unnatural changes")
3. The leader of the opponents has the face of the devil

4. This is a noble cause that we defend (eg a "public service") and not individual interests
5. The opponents knowingly provokes atrocities (eg murders, dismissals, deaths, ...) and if we commit mistakes it is involuntarily
6. The opponents uses unauthorized weapons (i.e. we do not understand the arguments of the opponents)
7. We suffer very little loss, the opponents losses are huge (the current system is good so we must not change it because it will be worse)
8. Artists and intellectuals support our cause
9. Our cause has a sacred character (eg truth seeker, public service, etc.)
10. Those who put in doubt our propaganda are traitors (i.e. they weaken the social cohesion, the national cohesion, try to make profit on the poorest, etc.)

This is typically "post-truth" circumstances as defined by the Oxford dictionary[49].

We can also add to this list the famous: «*I did my own research*». The guys searched two hundred hours on Twitter, YouTube and amateurs Blogs for scientific evidence by cherry picking and ignoring meta-analysis and this without the technical skills to understand scientific protocols and statistical (literally they pick up the first results on Google thinking that any text on the Internet or in a holy fictional book has a scientific value...).

NUCLEAR AND QUANTUM LAB
CHEMISTRY LAB
BIOMOLECULAR LAB
«DID MY OWN RESEARCH» LAB
Google

---

[49] Post-truth: relating to or denoting circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief.

No wonder, if the world is filled with empty and ridiculous opinions, there is nothing more capable of giving them course, than ignorance!

Furthermore, keep in mind the "Dunning-Kruger effect" that is a type of cognitive bias in which people believe that they are smarter and more capable than they really are. Essentially, low ability people do not possess the skills needed to recognize their own incompetence. The first rule of Dunning-Kruger club is you don't know you're in Dunning-Kruger club!

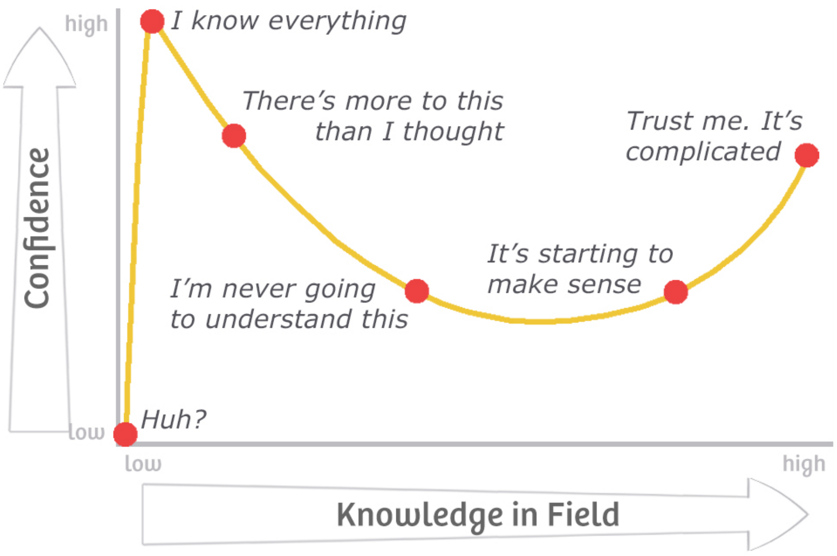


Figure 4.13: **Caricatural Dunning-Kruger effect curve (see the original paper of Dunning-Kruger for the original one)**

Remark

The most important mistake people make about the Dunning-Kruger effect, according to Dr. Dunning, has to do with who falls victim to it. «*The effect is about us, not them*» he wrote. «*The lesson of the effect was always about how we should be humble and cautious about ourselves.*» The Dunning-Kruger effect is not about dumb people. It's mostly about all of us when it comes to things we are not very competent at. Also some sociologist claim that these effect may not exist in reality and that it's a data artefact. As always anyway the conclusion is that educated people know that things are always more complicated than they seem to be...!

quote 4.2:84.#18

> *Ignorance more frequently begets confidence than does knowledge: it is those who know little, not those who know much, who so positively assert that this or that problem will never be solved by science.*
>
> Charles Darwin

4.2.4.2 Expert Opinions

We must also be careful with the method mainstream media choose to interview "experts".

Indeed, in science it doesn't make really sense to invite an expert to speak especially if the latter:

- Use arguments without providing detailed evidence (name of peer-reviewed meta-analysis)
- Don't make difference between "opinions" and "scientific based evidence"
- Speaks about subjects outside of his field of specialization
- Don't work anymore since many years in laboratories
- Use his awards and books to give himself a legitimate expert position
- Use his laboratory team work to promote himself
- Is alone doing a monologue[50] (typical of TEDx as already mentioned)
- Is interviewed by a scientifically illiterate journalist
- Is full of certitudes (the stupid are cocksure while the intelligent[51] are full of doubt)
- Has a mental illness (not to be confused with a physical disease)

There a lot of famous examples (like Nobel Prizes that have turned obsess by irrational subjects outside of their field of competence). However let us give two examples that we will meet later in this textbook again in different sections.

The first example is that Nosek's team that invited researchers to take part in a crowd-sourcing data analysis project. The setup was simple. Participants were all given the same data set and prompt: Do soccer referees give more red cards to dark-skinned players than light-skinned ones? They were then asked to submit their analytical approach for feedback from other teams before diving into the analysis.

Twenty-nine teams with a total of 61 analysts took part. The researchers used a wide variety of methods, ranging - for those of you interested in the methodological gore - from simple linear regression techniques to complex multilevel regressions and Bayesian approaches. They also made different decisions about which secondary variables to use in their analyses.

Despite analysing the same data, the researchers got a variety of results. Twenty teams concluded that soccer referees gave more red cards to dark-skinned players, and nine teams found no significant relationship between skin color and red cards.

---

[50] As already said people should not trust monologues - whether on radio, television or any social network - because there are no experts to counter the possible wrong arguments or ill-defined concepts that may lead a part of the audience to wrong speculative interpretations! In addition, humans under the stress of knowing that they are recorded are naturally prone to vocabulary errors and that's without counting the more than 200 cognitive biases in the brain which sometimes lead to the erroneous simplification of complex thoughts... !

[51] See the definition of "intelligence" at page 7385.

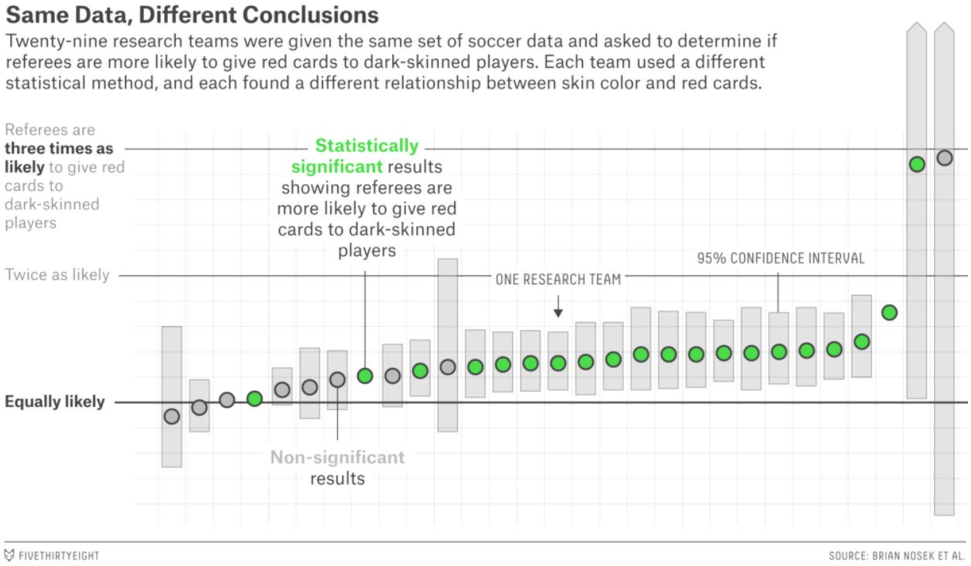


Figure 4.14: **Same data, different conclusions (purpose of meta-analysis)**

The variability in results wasn't due to fraud or sloppy work. These were highly competent analysts who were motivated to find the truth, said Eric Luis Uhlmann, a psychologist at the Insead Business School in Singapore and one of the project leaders. Even the most skilled researchers must make sometimes subjective choices that have a huge impact on the result they find. That's why again it is quite a non-sense to give only one expert to speak in mainstream media in a monologue[52]...

The second recent example (obviously it's special case that must not be generalized!) was the spread of the COVID-19 in the first trimester of year 12020 (holocene calendar) in the USA. It has been asked to a few US experts what is their prognosis for the number of positive COVID-19 cases in March 29 in USA. This was summarised in the following figure:

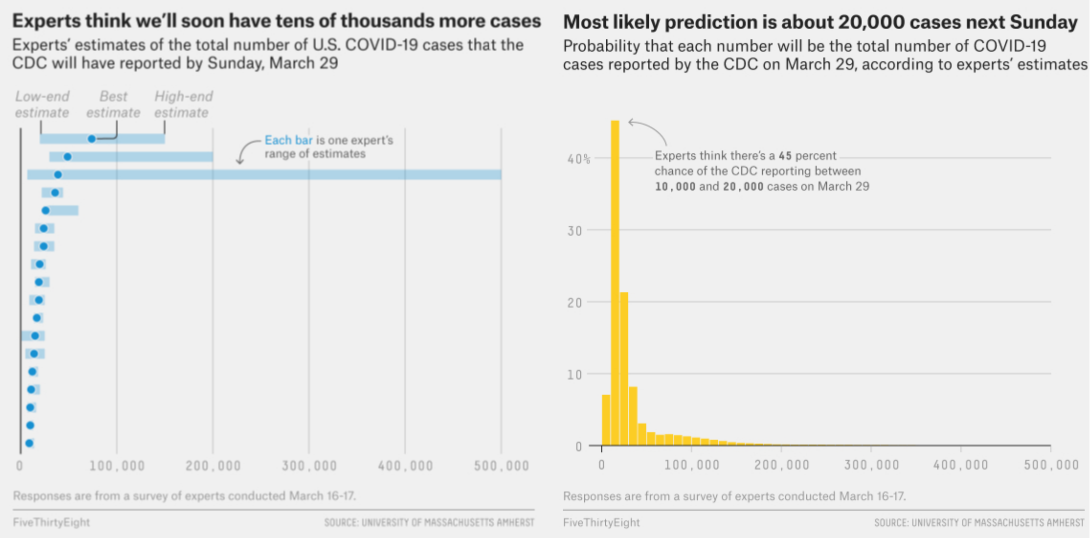


Figure 4.15: **Experts opinions/estimation issue**

---

[52] As already said people should not trust monologues - whether on radio, television or any social network - because there are no experts to counter the possible wrong arguments or ill-defined concepts that may lead a part of the audience to wrong speculative interpretations! In addition, humans under the stress of knowing that they are recorded are naturally prone to vocabulary errors and that's without counting the more than 200 cognitive biases in the brain which sometimes lead to the erroneous simplification of complex thoughts... !

For information, the number of positive cases known in March 29 in USA was in fact a bit above $100,000$. This is again a good example why a unique expert opinion or estimation - whatever the field - may not be always very reliable.

Obviously the reader must keep in mind that the variability exists in the above two examples because the scientists were not authorized to use the "Delphi method". Basically, the Delphi method is a process used to arrive at a group opinion or decision by surveying a panel of experts. Experts respond to several rounds of questionnaires, and the responses are aggregated and shared with the group after each round using statistical techniques (we will come back on that technique in the section Game and Decision Theory page 7392).

Statisticians do not in general exactly agree on how to analyse anything but the simplest of problems. The fact that statistical inference uses mathematics does not imply there is only one reasonable or useful way to conduct an analysis. Engineering uses math as well, but there are many ways to build a bridge. So trying multiple analyses given to multiple independent teams of researchers may be considered scientifically prudent to test the robustness of findings.

4.3. Vocabulary

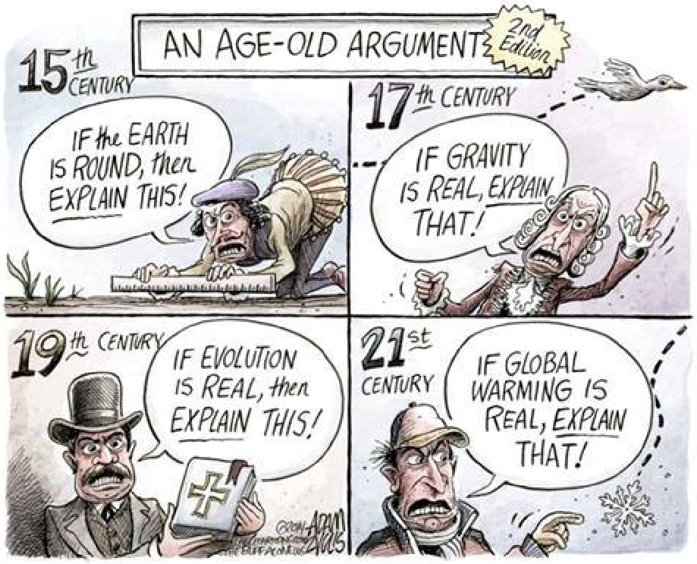
PHYSICS AND MATHEMATICS, like any field of specialization, has its own vocabulary. So that the reader is not lost in the understanding of certain texts he can read in this PDF, we have chosen to present here a few fundamentals words, abbreviations and definitions to know.

Thus, the mathematician like to finish his proofs (when he thinks they are correct) by the abbreviation "Q.E.D." which means "Quod Erat Demonstrandum" (this is Latin).

And during definitions (they are many in math and physics ...) scientist often use the following terminology:

- ... it is sufficient that ...
- ... if and only if ...
- ... necessary and sufficient ...
- ... means ...
- ... prove it ...

These five are not equivalent (identical in the strict sense). Because "it is sufficient that" correspond to a sufficient condition, but not to a necessary condition. Also it must be notice that these five are placed in the context of data analysis, data accuracy, reproduction and peer review and not on any personal or common belief or also emotional aspect of a group of people (even if this group of people is more than a few billion individuals...)!


AN AGE-OLD ARGUMENT
2nd Edition
15th CENTURY
IF the EARTH IS ROUND, then EXPLAIN THIS!
17th CENTURY
IF GRAVITY IS REAL, EXPLAIN THAT!
19th CENTURY
IF EVOLUTION IS REAL, then EXPLAIN THIS!
21st CENTURY
IF GLOBAL WARMING IS REAL, EXPLAIN THAT!

4.3.1 On Sciences

It is important that we define rigorously the different types of sciences to which humans often refers. Indeed, it seems that in the 121st century (holocene calendar) a misnomer is established and that it became impossible for people to distinguish the "intrinsic quality" between a "science" and another one.

Remark

Etymologically, the word "science" comes from the Latin "Scienta" (knowledge) whose root is the verb "scire" which means "to know".

This abuse of language is probably the fact that pure and accurate sciences lose their illusions of universality and objectivity, in the sense that they are self-correcting. This has for effect that some sciences are relegated to the background and try to borrow these methods, principles and origins to generate confusion. We must therefore be very careful about the claims of scientificity in the human sciences, and this is also (or especially) true for the dominant trends in economics, sociology and psychology. Quite simply, the issues addressed by the human sciences are extremely complex, poorly reproducible, and empirical arguments supporting their hypotheses are often quite low.

By itself, however, science does not produce absolute truth as we have already explained it numerous times. By principle, a scientific theory is valid as long as it can predict measurable and reproducible results that can be analysed statistically[53]. But the problems of interpretation of these results are part of natural philosophy.

No scientific theory is proven or provable (i.e. an absolute truth)! It is simply not falsified as long as a set of experiments have not provided contradictory evidence!!

Definitions (#4.3.4):

D1. A "scientific theory" is the best well-substantiated explanation/hypothesis of some aspect of the natural world that is based on a body of empirical[54], repeatable, measurable, controllable, verifiable, calculable and falsifiable evidence (generally condensed into a meta-analysis or systematic review analysis) and that has been tested and confirmed by inferential or abductive statistics at a significant threshold through multiples independent observations and experimentations. Unlike the everyday use of the word "theory", which might imply a guess or speculation, a scientific theory is grounded in empirical evidence and is widely accepted within the scientific community until proven otherwise.

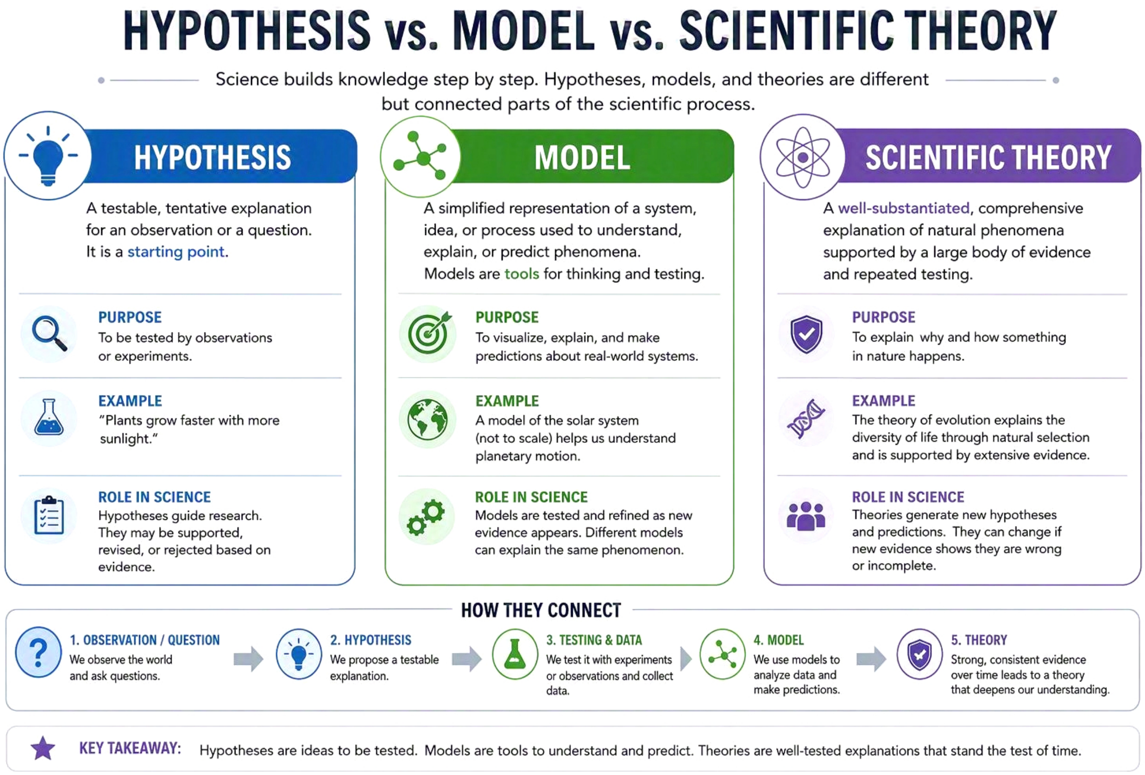
D2. A "scientific proof" is an exhaustive set of testable, measurable, and reproducible independent experiments that converge statistically, based on NHST (Null Hypothesis Statistical Testing) or BHST (Bayesian Statistical Testing(, on the same conclusion beyond reasonable doubt.

---

[53] So what we can read in all the various religious books existing around the world are not "scientific theories" but instead "speculative theories"!

[54] "Empirism" is academically defined by the state of a science when the facts based on the observation of the reality by accurate instruments and methods are not yet linked by any general theoretical model.

D3. A "scientific law", or commonly named "law", is a predictive formula obtained through experiment that hasn't any theoretical explanation yet (like Newton's law that was only later explained by the theoretical framework of General Relativity). Such laws work in the observable universe and are intelligible because they are human construction to approximate the reality.



HYPOTHESIS vs. MODEL vs. SCIENTIFIC THEORY

Science builds knowledge step by step. Hypotheses, models, and theories are different but connected parts of the scientific process.

HYPOTHESIS

A testable, tentative explanation for an observation or a question. It is a starting point.

PURPOSE
To be tested by observations or experiments.

EXAMPLE
"Plants grow faster with more sunlight."

ROLE IN SCIENCE
Hypotheses guide research. They may be supported, revised, or rejected based on evidence.

MODEL

A simplified representation of a system, idea, or process used to understand, explain, or predict phenomena. Models are tools for thinking and testing.

PURPOSE
To visualize, explain, and make predictions about real-world systems.

EXAMPLE
A model of the solar system (not to scale) helps us understand planetary motion.

ROLE IN SCIENCE
Models are tested and refined as new evidence appears. Different models can explain the same phenomenon.

SCIENTIFIC THEORY

A well-substantiated, comprehensive explanation of natural phenomena supported by a large body of evidence and repeated testing.

PURPOSE
To explain why and how something in nature happens.

EXAMPLE
The theory of evolution explains the diversity of life through natural selection and is supported by extensive evidence.

ROLE IN SCIENCE
Theories generate new hypotheses and predictions. They can change if new evidence shows they are wrong or incomplete.

HOW THEY CONNECT

1. OBSERVATION / QUESTION We observe the world and ask questions. →
2. HYPOTHESIS We propose a testable explanation. →
3. TESTING & DATA We test it with experiments or observations and collect data. →
4. MODEL We use models to analyze data and make predictions. →
5. THEORY Strong, consistent evidence over time leads to a theory that deepens our understanding.

★ **KEY TAKEAWAY:** Hypotheses are ideas to be tested. Models are tools to understand and predict. Theories are well-tested explanations that stand the test of time.

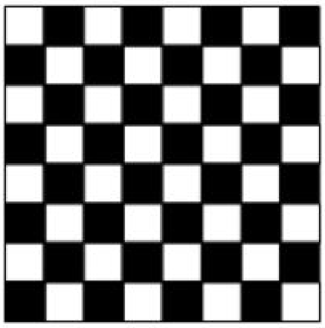
However, the scientific method is reliable enough so that Justice is not legitimate to take position on any scientific knowledge. And also the reader should keep in mind that actually, and as far as we know, it is impossible for the scientific method to prove that anything is empirically true. At best it can improve the likelihood of something to be true. The scientific method builds knowledge on experimental evidence, and provable mathematics models, not axioms. But experimental evidence verifies or falsifies theories, it does not prove them!!!!!!!!

Remark

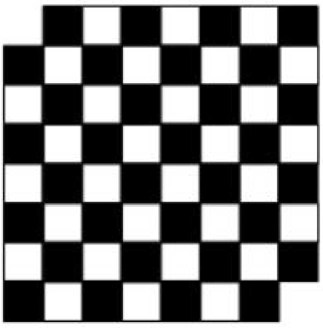
Hence keep in mind that a testimonial, or a simple sentence in a book or in a seminar even if said by a scientist has NO SCIENTIFIC VALUE, or no value for drawing conclusions, if it is not accompanied by data (typically proper, replicable, controlled scientific trials) and a mathematical detailed model! However, testimonials remains only useful by building speculative hypotheses and designing experiments and studies. If a speculative hypothesis is not accompanied by a methodology to test it experimentally, then it's still NOT science, but just free bullshit speculation! In a more elegant way we may say that it is: *pseudoscientific rhetoric aimed at a mathematically unsophisticated audience*!

If you don't get it (or you know people that don't get it), here is an illustrated example that may help to differentiate a "proof" and an "evidence":

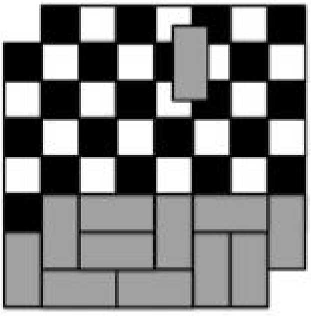
Let's imagine a chess board (a square board with black and white boxes in staggered rows) of 8 boxes per side. The board contains 64 boxes.



If we remove the two diagonally opposed boxes (two white boxes, in our case), there are 62 boxes remaining.



We propose to find a way to cover up all the remaining boxes with dominoes exactly the size of two chess boxes. All the chess boxes must be covered, and no domino must exceed the chess board.



The question is very simple: « is there at least one solution in which all the remaining chess boxes are covered, and no domino exceeds the chess board? »

- **Scientific Method:**

  An exact science is a science for which all theories are based on experiments. The scientific method consists thus to try all the possible ways to lay out the dominoes in order to cover up the chess board. After a certain number of tries, one might conclude that « It's impossible! I tried everything. » Scientifically speaking, the evidence is made that the problem has no solution.

But you will keep a doubt, deep inside you. Because maybe if you systematically eliminate all layouts, you'll finally end up on a layout that works. This would then break down the evidence of impossibility.

- **Mathematical method:**

  Each domino covers two chess boxes. A white one and a black one. On 62 remaining boxes, there are 32 black ones and 30 white ones.

  When 30 dominoes will be placed on the board, all the white boxes will be covered and there will be 2 black boxes not yet covered.

  Those two black boxes can not touch each other by a side. The best would be touching each other by a corner. A domino will thus be unable to cover both remaining boxes.

  The mathematical proof can not be argued and will never be contradicted.

Given the diversity of phenomena to be studied, over the centuries there has been a growing number of disciplines such as chemistry, biology, thermodynamics, etc. All these disciplines that are a priori heterogeneous have common foundation physics, for language mathematics, as grammar the statistics and for elementary principle the scientific method.

> ❞ *The sciences do not try to explain, they hardly even try to interpret, they mainly makes models. By a model is meant a mathematical construct which, with the addition of certain verbal interpretations, describes observed phenomena. The justification of such a mathematical construct is solely and precisely that it is expected to work-that is, correctly to describe phenomena from a reasonable wide area.* ❞
>
> J. von Neuman

quote 4.3:92.#19

Thus, a small memory refresh seems useful:

Definitions (#4.3.5):

D1. We define as "pure science" any set of knowledge based on rigorous reasoning (deductive logic) valid whatever the (arbitrary) elementary factor selected (when we say then "independent of sensible reality") and restricted to the minimum necessary. Only mathematics (often named the "queen of sciences") can be classified in this category.

D2. We define as "exact science" or "hard science" or "empirical science" or "evidence based science", any set of knowledge - appealing to no supernatural cause - based on the advanced statistical study of protocolized observations in the real world, observations that have been transcribed in symbolic form and that can be massively reproduced and refuted (theoretical physics for example... sometimes...). Primarily, the purpose of this kind of science is not to explain the "why" but the "how" and is based on deductive and abductive logic.

And never forget... Science (especially physics) doesn't have to "make sense" it just has to make all the right, testable predictions (instrumentalism)! According to the philosopher Karl Popper, a theory is scientifically acceptable if, as presented, it can be "falsifiable" (synonyms are "refutable" or "testable"), i.e. subjected to experimental tests[55] (or if it is possible to conceive of an observation or an argument which negates the statement in question). The "scientific knowledge" is then by definition the set of theories that have resisted to falsification (no amount of experimentation can ever prove you are right, but a single experiment can prove you are wrong!). Science is by nature subject to continuous questioning.

---

[55] That's why philosophy and naive human logic themselves can't "prove" anything. Because they strongly depend and vary from one socio-cultural context where people are born to another one (people from New-York have a quite different logic and philosophy from the Sentinelese tribe...). Without falsification science would be an anarchy of logically consistent but still useless models that simply suit someone's fancy.

quote
4.3:94.#20

> *A theory which is not refutable by any conceivable event is non-scientific. Irrefutability is not a virtue of a theory (as people often think) but a vice.*
>
> Karl Popper

Logical Structure of Falsificationism

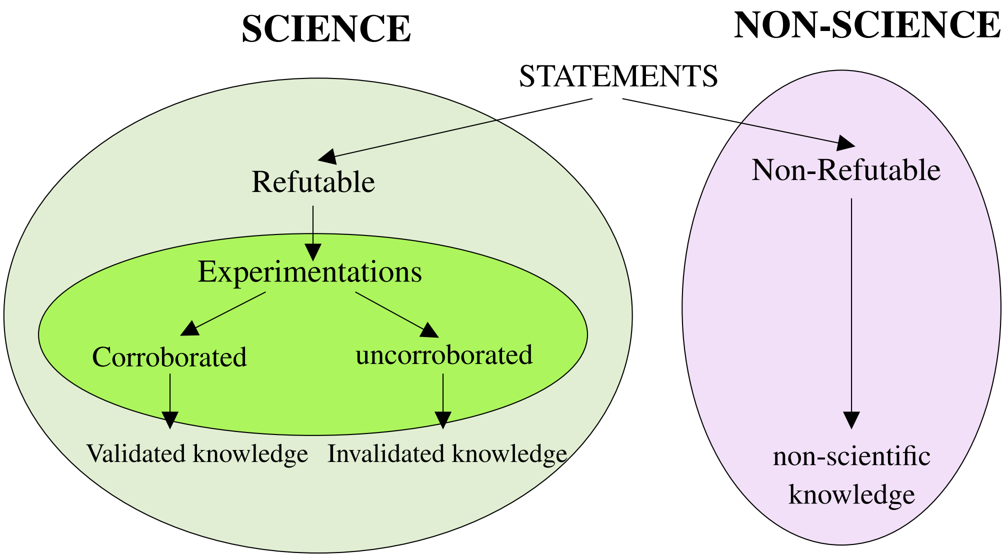
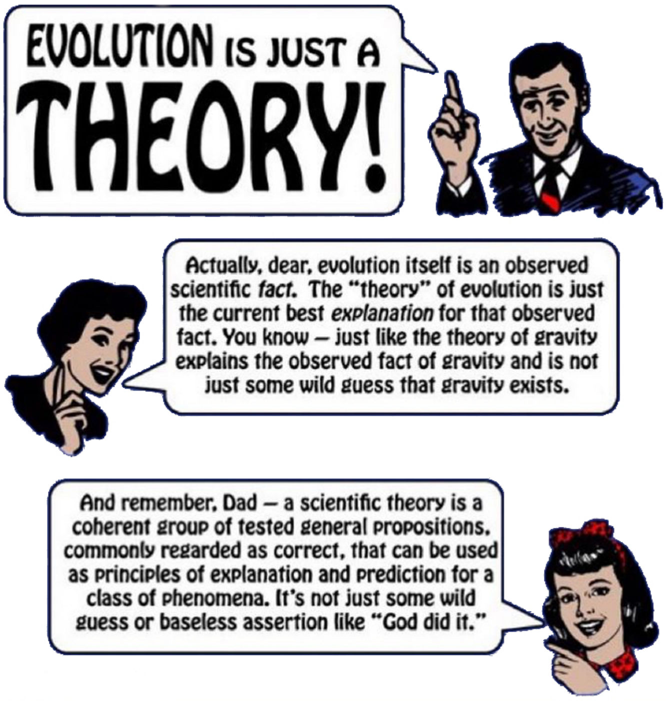

SCIENCE
NON-SCIENCE
STATEMENTS
Refutable
Non-Refutable
Experimentations
Corroborated
uncorroborated
Validated knowledge
Invalidated knowledge
non-scientific knowledge

Figure 4.16: **Logical structure of falsificationism**

Caution! There is no doubt that the exact sciences have yet an enormous prestige, even among their opponents because of their theoretical and practical success. It is certain that some scientists sometimes abuse of this prestige by showing a sense of superiority that is not necessarily justified. Moreover, it often happens that this same scientists expose in the popular literature, very speculative ideas as if they were very approved, and extrapolate their results outside the context in which they were tested (and ... under the hypotheses they were checked once...).


EVOLUTION IS JUST A THEORY!
Actually, dear, evolution itself is an observed scientific *fact*. The "theory" of evolution is just the current best *explanation* for that observed fact. You know – just like the theory of gravity explains the observed fact of gravity and is not just some wild guess that gravity exists.
And remember, Dad – a scientific theory is a coherent group of tested general propositions, commonly regarded as correct, that can be used as principles of explanation and prediction for a class of phenomena. It's not just some wild guess or baseless assertion like "God did it."

> **Remark**
>
> The two previous definitions are often included in the definition of "deductive sciences" or even "phenomenological science".

D3. We define as "rational science" the intellectual activity that starts from self-evident maybe biased claims and moves by rigorous first-order logical (see section Proof Theory page 163) arguments (that may be disconnected from the reality of experimental science) to a conclusion about some new logical truth (not related to experimental truth), this is named a "deduction" and may be or not wrong relatively to the conclusions of empirical science.

D4. We define as "engineering science" or "applied science" any set of knowledge or practices applied to the needs of human society such as electronics, chemistry, computer science, telecommunications, robotics, aerospace, biotechnology...

SCIENCE
ENGINEERING
SCIENCE IS ABOUT KNOWING
ENGINEERING IS ABOUT DOING

Remark

A common confusion by the population is to argue that scientists are the reason why guns and bombs exists. In facts physicists (and other "real" scientists) focus on the theories, the understanding of Nature, not on the business or political applications. The people that converts theories into business applications are "engineers" and engineers are not scientists. Engineers just convert science knowledge into business application in a genius way. That's what engineers are!

D5. We define as "science" any body of knowledge based on studies or observations of events whose interpretation has not yet been transcribed and verified with mathematical rigour, characteristic of previous sciences, but using comparative statistics. We include in this definition: medicine (we should however be careful because some parts of medicine are studying phenomena using mathematical descriptions such as neural networks or other phenomena associated with known physical causes), sociology, psychology, history, biology, etc.

Some teachers like to play with the word "science" as the acronym of (that's not stupid for college students): **S**olve, **C**reate, **I**nvestigate, **E**valuate, **N**otice, **C**lassify, **E**xperiment.

D6. We define a (modern) "scientist" as a professional owning a Doctorate and who systematically and daily gathers and uses reproducible research and experimental evidence, to make statistical hypothesis and test them using cutting edge scientific methods (like PhD level statistics or advanced reproducible numerical simulations), to gain and share peer-reviewed understanding and knowledge in famous scientific journals in a very specialized field of expertise.

The definition therefore exclude the huge majority of engineers (who don't publish in peer-reviewed papers and doesn't know how to analyse data using PhD statistical methods), laboratory assistants who just make experimental manipulations but don't make the hypothesis nor analyse the resulting measurements of their experiments but also medical doctors (see [500] and [190]) who only apply the results of scientific research but don't do any as specified above. Mathematicians are not excluded from this definition as their mathematical theories (proofs) are published in peer-reviewed papers and are reproduced (checked) independently by other mathematicians.

D7. We define as "soft science", "para-science" or "pseudo-science" any set of knowledge or practices that are currently based on non-verifiable and non-refutable facts (not scientifically reproducible) by experience or by mathematics. We include in this definition typically: astrology, theology, paranormal (which was demolished by zetetic science), graphology, justice[56], etc.

As some scientists say: *«It looks like science, it use the vocabulary of science... but that's not science at all.»*

Especially pseudo-sciences are characterized by:

- They start with a conclusion (belief), then work backwards to confirm
- They make sometimes unfalsifiable, vague or unobservable claims

---

[56] Indeed, for example in Switzerland, it is common that the cantonal Judge and the Federal Judge don't give the same judgement as that latter is non-scientific but rather subjectively based on the Judge experience of life

- They relied heavily on anecdotes, personal experiences, and testimonials
- They are hostile to criticism
- They use circular reasoning/arguments and other logical fallacies
- They use vague jargon to confuse and evade
- They use subtle strategies to change people minds (especially children)
- They make people feel special and unique
- They do cherry picking only on favourable evidence[57]
- They use non-reproducible/non-refutable methods with unrepeatable results
- They use bullshit-random language (technobabble) to impress the audience
- They use inconsistent and invalid logic
- They profess certainty, talk of "proof" with great confidence
- They claim there's a conspiracy to suppress or hide their ideas
- They go directly to the public, avoiding scientific scrutiny and consensus
- People working in the field are dogmatic and unyielding
- They highly monetize their books, videos (YouTube), conferences, etc.

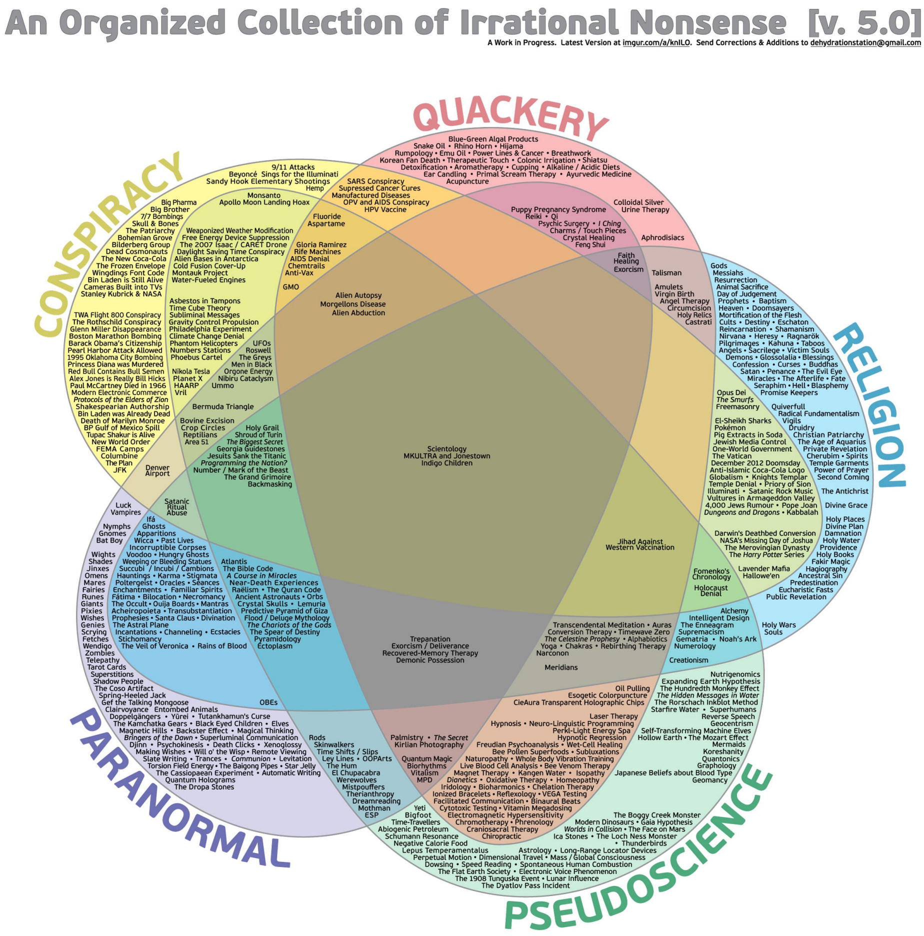

An Organized Collection of Irrational Nonsense [v. 5.0]
A Work in Progress. Latest Version at imgur.com/a/knlLO. Send Corrections & Additions to dehydrationstation@gmail.com
QUACKERY
CONSPIRACY
RELIGION
PARANORMAL
PSEUDOSCIENCE
Scientology
MKULTRA and Jonestown
Indigo Children

Figure 4.17: Organized collection of irrational nonsense

---

[57] Typical especially for all thousand of religions and also pseudo-medicine like acupuncture that can be scientifically test by having a control group with no treatment and a placebo group treated with retractable needles...

And don't forget:

What can be asserted without scientific evidence can be dismissed without scientific evidence!!

D8. We define as a "scientific fact" a fact (i.e. simple eyewitness observation) that has been confirmed repeatedly using the scientific method (with modern apparatus, advanced statistical analysis, bias reduction, independent methods, independent teams and confusion variables control) and is accepted as true (although its truth is never final and must be continuously challenged!) by confirmation of meta-analysis (highest level of scientific evidence).

D9. We define as "phenomenological science" or "natural sciences", any science which is not included in the above definitions (history, sociology, psychology, zoology, biology, ...).

D10. "Scientism" is an ideology that considers experimental science is the only valid mode of knowledge, or, at least, superior to all other forms of interpretation in the world. In this perspective, there is no philosophical, religious or moral truths superior to scientific theories. Only account what is scientifically proven (i.e. evidence beyond reasonable doubt!).

D11. "Positivism" is a set of ideas that considers that only the analysis and understanding of facts verified by experience can explain the phenomena of the sensible world. Certainty is provided solely by the scientific experiment. He rejects introspection, intuition and metaphysical approach to explain any knowledge of the phenomena.

What is interesting about this doctrine is that it is certainly one of the few that requires people to have to think for themselves and to understand the environment around them by continually questioning everything and by never accepting anything as granted (...). In addition, the real sciences have this extraordinary property that they give the possibility to understand things beyond what we can see.

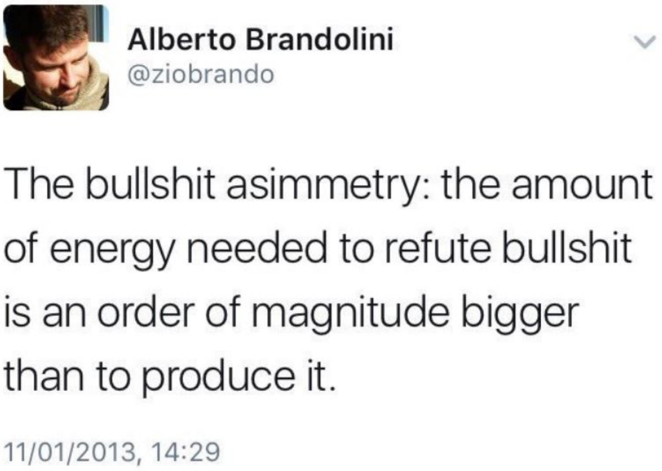
But, science is science, and nothing more: a certain ordering, not too bad success, things that no longer leads to the metaphysics as the time of Aristotle, but that does not pretend to give us the whole story on reality or even the bottom of visible things.

People must remember that Science is a robust cross-validated investigation method and NOT a belief system!!

Remark

A quite common bias is to argue that «*scientists believes in their actual models and the things they based their observations on*». That's totally wrong! Indeed, trained and real scientists are professionals (they are paid every month to do research and development - most of the time to find methods to reject existing theories - at least 30 to 70 hours per week after 5 to 7 years of graduate studies!) and at this level they are supposed to not be attached sentimentally to the models they're working on or things they used to make observations (even if some scientists - and sometimes some famous one - cannot control their cognitive biases!). The "*actual models*" and "*things they based their observations on*" are for professional scientists just "toy models" and "calibrated measurement tools" that seems actually to work well to explain and study our surrounding. When they have to build new models, this is because they are paid for! Not because they believe in them! There are absolutely no beliefs! Do you think that a garage owner really "believe" in it's wrench to work? He doesn't give a f**** about believing in it or not, it's just a well designed tool (maybe there is a better one?) that he uses to do his work conscientiously - sometimes with passion - to get his wage at the end of the month to pay his rent and food for his family!

And by the way... also don't forget that:



Alberto Brandolini
@ziobrando

The bullshit asimmetry: the amount of energy needed to refute bullshit is an order of magnitude bigger than to produce it.

11/01/2013, 14:29

Figure 4.18: **The Alberto Brandolini bullshit asymmetry law (source: Twitter)**

4.3.2 Terminology

The table of methods we presented earlier above (see page 51) contains terms that may perhaps seem unknown or barbarous for you. This is why it seems important to provide definitions of these and some other equally important that can avoid important confusion.

Definitions (#4.3.6):

D1. Beyond its negative sense, the idea of "problem" refers to the first step of the scientific method. Formulate a problem is also essential for its resolution and allows to properly understand what is the problem and see what needs to be resolved.

The concept of "problem" is intimately connected to the concept of "assumption" which will see the definition below.

D2. A "hypothesis" is always, in the context of a theory already established or underlying, a supposition awaiting confirmation or refutation that attempts to explain a group of facts or predict the onset of new facts.

Thus, a hypothesis can be at the origin of a theoretical problem that has to be resolved formally.

D3. The "postulate" or "assumption" in physics corresponds frequently to a principle (see definition below) which admission is required to establish a proof (we mean that this is a non-provable proposition).

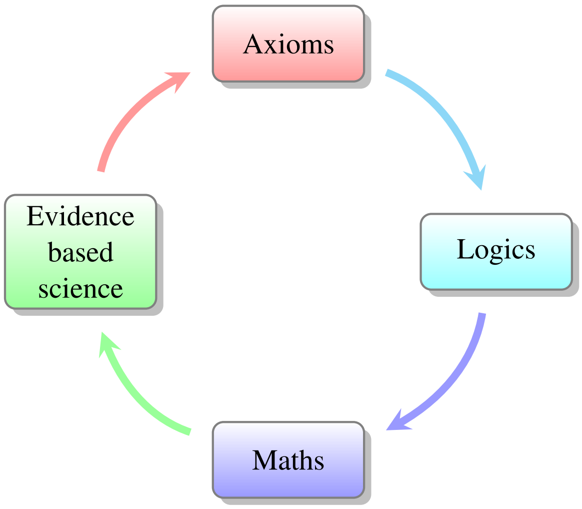
The mathematical equivalent (but in a more rigorous version) of the assumption is the "axiom" for which we will see the definition below.

D4. A "principle" (close parent of "postulate") is a proposal accepted as a basis for reasoning or a general theoretical guide line for reasoning that needs to be performed. In physics, it is also a general law governing a set of phenomena and verified by the accuracy of its consequences.

The word "principle" is used with abuse in small classes or engineering schools by teachers not knowing (which is very rare), or unwilling (rather common), or that can't because lack of time (almost exclusively) prove a relation.

The equivalent of the postulate or principle in mathematics is the "axiom" which we define as follows:

D5. Naively (in Euclid's time...) an "axiom" is a self-evident proposition or truth by itself which admission is necessary to establish a logical proof (this naive definition therefore does not require the resulting logical proof to be connected to reality!). In modern times, however, technically in first order logic (see page 131) an "axiom" is a formula $R(t_1,...,t_n)$ where R is a n-ary predicate symbol and the t_i are terms which admission is necessary and compliant with evidence based science to establish a logical proof.



We can show that an axiom is inconsistent with the remaining axioms of an axiomatic system by proving the negation of the axiom in the system. That is a form of disproof.

We can also prove an axiom by proving that it is implied by the remaining axioms in that same system.

In either case, we would then want to remove the axiom from the system.

If a system made of a single axiom contradicts reality, then it can be disproved by itself by experimental evidence. Like the axiom "every effect has a cause" is wrong in modern physics as we will see it during our study of Quantum Wave Theory (Casimir Effect) or also in our study of Nuclear Physics (radioactivity) or Cosmology (creation ex-nihilo of our universe).

If, however, the axiom is independent of the remaining axioms — very desirable, but we can't always show that — then it can be neither proven nor disproven in the system. Euclid's fifth postulate (that we will introduce during our study of Euclidean geometry at page 2553) is an example of that (as is the axiom of choice and the continuum hypothesis). Indeed, while axioms are accepted as true without proof, mathematicians often explore their implications and study their consequences. For example proofs related to the axiom of choice in Set Theory typically (see page 382) involve demonstrating its equivalence to other statements or exploring its applications. That's why we say that *the axiom of choice can neither be proved nor refuted in ZF* (which is all the ZFC axioms except the axiom of choice) and this result do not require the Gödel incompleteness theorems. And another famous example already mentioned just above is the 5th axiom of Euclidean geometry that has been proven wrong by Nikolai Lobachevsky in non-Euclidean geometries (it needs to be reformulated to hold!).

Remarks

R1. We could say that this is something we define as the truth for the speech that we argue, like a rule of the game, and that it does not necessarily a universal truth value in the sensitive world around us.

R2. Axioms must always be independent (one should not be able to be proved from the other) and non-contradictory (sometimes we also say that they must be "consistent").

D6. The "corollary" is a term unfortunately almost non-existent in physics (wrongly!) and that is in fact a proposal resulting from a truth already demonstrated. We can also say that a corollary is and obvious and necessary consequence of a theorem (or sometimes of a postulate in physics).

D7. A "lemma" is a proposal deduce from one or more assumptions or axioms and that for which the proof prepares this of a theorem.

Remark

The concept of "lemma" is also (and this is unfortunate) almost used only in the field of mathematics.

D8. A "conjecture" is a supposition or opinion based on the likelihood of a mathematical result.

Many conjectures have as little similar to lemmas, as they are checkpoints to obtain significant results.

D9. Beyond its weak conjecture sense, a "theory" or "theorem" (whose etimology means “a spectacle”, “subject for contemplation” or “proposition to be proved”) is a set articulated around a hypothesis and supported by a set of facts or developments that give it a positive content and make the hypothesis well-founded (or at least plausible in the case of theoretical physics).

D10. A "singularity" is an indeterminacy in a calculation That takes the appearance of a division by zero. This term is both used in mathematics and in physics.

D11. A "proof" is a set and sequence of mathematical procedures to use and follow to derive a result already known or not of a theorem, lemma or corollary (as already mentioned, this term should obviously be prohibited in experimental sciences as he doesn’t highlight the underlying level of evidence!).

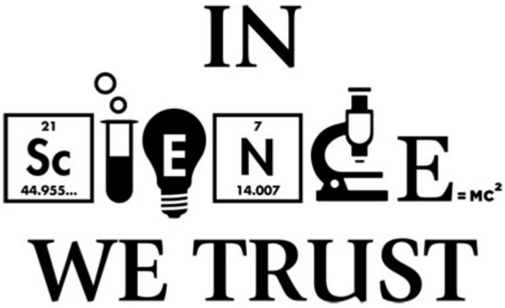
D12. If the word "paradox" etymologically means: contrary to common opinion, it is not by pure taste for provocation, but rather for solid reasons. A "sophism" meanwhile, is a deliberately provocative statement, a false proposition based on an apparently valid reasoning. Thus we speak about the "Zeno’s paradox" when in reality it is only a sophism. The paradox is not limited to falsity, but implies the coexistence of truth and falsity, so that one can no longer distinguish true and the false. The paradox appears as an unsolvable problem an "aporia".

Remark

It should be added that the well-knows paradoxes, by the questions they raised, have permitted significant advances to science and led to major conceptual revolutions in mathematics as in theoretical physics (the paradoxes on sets and on infinity in mathematical, and those at the base of relativity and quantum physics).

4.4. Science and Faith

We will see that in Science, a (scientific) theory is usually incomplete because it can not fully describe the complexity of the real world or because it does not predict what we don't know (excepted for Quantum Physics or General Relativity). It is thus for theories like the Cosmic Inflation Theory (see section Astrophysics page 5539) or the Evolution of Species (see sections Populations Dynamics page 7297 or Decision and Games Theory page 7376) because they are not reproducible in laboratories under identical conditions. But some other (scientific) theories are so accurate to predict physical phenomena that some people believe that mathematics is the nearest language with some kind of gods (at least for those who believe in a divinity[58]...) even if we know, as we have already mention it, that science is (should) be driven only by data and peer review publications.



We should distinguish between different scientific currents:

- "Realism" is a doctrine where physical theories have the aim to describe reality as it is in itself, in its unobservable components.
- "Instrumentalism" is a doctrine where physical theories are only tools to predict observations but do not describe reality itself.
- "Fictionalism" is the doctrine where the content repository (principles and postulates) of physical theories is just an illusion, useful only to ensure the linguistic articulation of the fundamental equations.

Even if today the scientific theories are sponsored by many specialists, alternative models have valid arguments and we can not totally dismiss them. However, the creation of the World in six or seven days as described in some books is difficult to accept, and many believers recognize that a literal reading of religious books is not compatible with the current state of our knowledge and that it is more prudent to interpret it as a parable (even if their own books forbids explicitly such an intellectual exercise...!). If Science never provides definitive answer, it is no longer possible to ignore it!

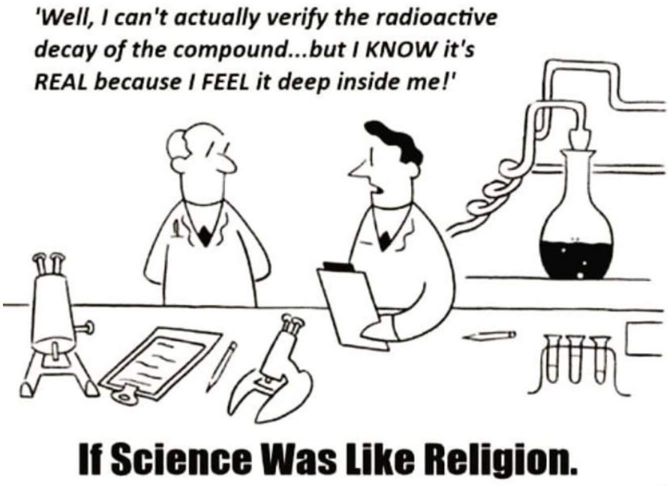
---

[58] Keep in mind that according to what we have seen before, the existence of a Divinity is an "Hypothesis" at the opposite of the Evolution of Species for example that is a "Theory" (see [492])!

Faith (whether religious, superstitious, pseudo-scientific or other not data driven) on the contrary is intended to provide absolute truths of a different nature as it is a personal unverifiable belief (for example, Science requires proofs to be evidence based, religions requires beliefs to be proved...). This is why many people say that «*Science adjusts views based on what's observed when Faith is the denial of observation so that belief can be preserved*».... In fact, one of the functions of religion is to give meaning to the phenomena that can not be explained rationally with actual knowledge[59]. Progress of knowledge trough Science therefore cause sometimes (...) questioning the religious dogma[60].

quote
4.4:104.#21

> *SCIENCE is not defined by what we believe. IT IS NOT a belief system. It has no religious doctrine! It's just an intellectual investigation method that rejects claims that have no supporting experimental reproducible evidence.*


'Well, I can't actually verify the radioactive decay of the compound...but I KNOW it's REAL because I FEEL it deep inside me!'
If Science Was Like Religion.

Conversely, except try to impose his own faith (which is nothing but a subjective and intimate personal conviction) to others, we must defy the natural temptation to characterize scientifically proven fact extrapolations of scientific models beyond their scope.

The word "Science" is, as we have already mentioned above, increasingly used to argue that there is a scientific evidence where there is only a belief (some web pages like this proliferate always more and more and especially to get followers and a lot of clicks on the Internet). According to its detractors it is, for example, the case of the movement of Scientology (but there are many others). According to them, we should rather speak about "occult sciences".

---

[59] This was the case with the rain, the thunder, diseases, stars, comets, earthquakes, volcanic eruptions, birth, evolution, etc. a few hundred years ago and is often designated by scientists under the name of "argument of ignorance"

[60] When theists say that a god must exist because Earth is "perfectly" place for life, they assume that their god is restricted to placing life where it could naturally occur anyway...

The occult sciences and traditional sciences exist since antiquity; they consist on a series of mysterious knowledge and practices designed to penetrate and dominate the secrets of nature. Over the past centuries, they have been progressively excluded from science. The philosopher Karl Popper has for a long time questioned himself about the nature of the demarcation between science and pseudoscience. After noticing that it is possible to find observations to confirm almost any theory, he proposes a methodology based on falsifiability. A theory must according to him, to deserve the adjective "scientific", guarantee the impossibility of some events. It becomes therefore refutable, so (and only then) capable of integrating science. It would suffice to observe any of these events to invalidate the theory, and therefore take the way to improving it.

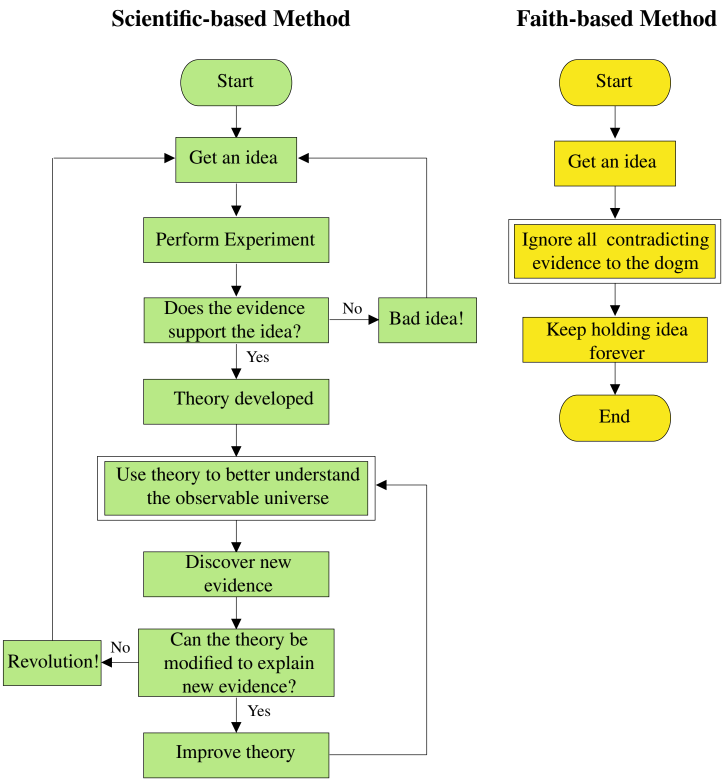
And also let us notice that major difference between scientific textbooks and religious books is that if you destroyed that latter, in a thousand year's time that wouldn't come back just as it was. Whereas if we took every science textbook and every fact and destroyed them all, in a thousand years they'd all be back. Because all the same tests would lead to the same results again!

> ❞ *The reason there is a conflict between science and religions is that science keep disproving the things religions claims to be true or accurate.* ❞

quote 4.4:105.#22

Remark

If science, and over everything MATHEMATICS are the language of a God, then why are all holy books written without maths? Oh, umh... we know why!


Scientific-based Method **Faith-based Method**

Start
Get an idea
Perform Experiment
Does the evidence support the idea?
No
Bad idea!
Yes
Theory developed
Use theory to better understand the observable universe
Discover new evidence
Can the theory be modified to explain new evidence?
No
Revolution!
Yes
Improve theory
Start
Get an idea
Ignore all contradicting evidence to the dogm
Keep holding idea forever
End

This is important because every one of the endless series of pseudo-proofs of the existence of various gods that has been proposed, from antiquity to the present day, is automatically a failure because, as already mentioned, a logical deduction tells us nothing that is not already embedded in its premises. All logic can do for us is test the self-consistency of those premises. There is only one reliable way that humans have discovered so far to obtain knowledge they do not already possess: observation! And science is the methodical collecting of observations and the building and testing of models to describe those observations. Without falsification science would be an anarchy of logically consistent but still useless models that simply suit someone's fancy!

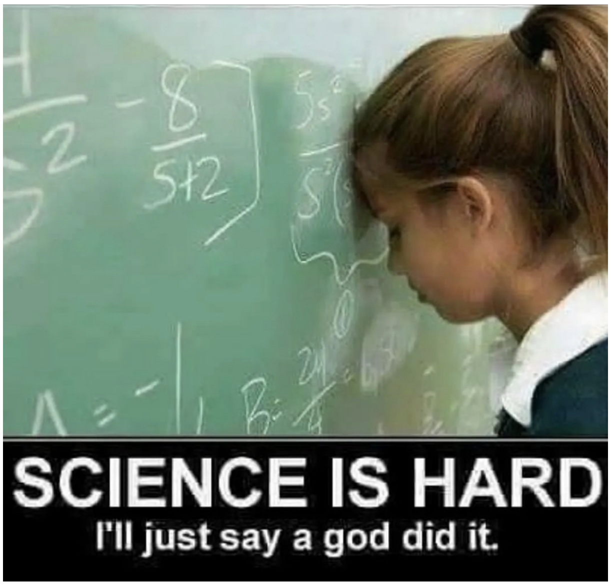
quote
4.4:106.#23

> ❞ *Religion is the practice of using nonsense, in order to explain ignorance.* ❞

Furthermore, religion has had thousands of years to provide evidence that any god exist. Yet can theists manage no better than: «*YOU JUST WAIT UNTIL YOU'RE DEAD*». That tells a lot... all the more because if gods were real they wouldn't need people to argue their existence! And if theists really value faith over science then they can prove it easily: the next time they get sick they can go to a church instead of going to the hospital!

quote
4.4:106.#24

> ❞ *A scientist will read dozens of books in his lifetime, but still known that he has a lot more to learn. A religious fanatic barely reads one book, and think that they know it all.* ❞


SCIENCE IS HARD
I'll just say a god did it.

The reader must also know that at the opposite of a common myth by scientifically illiterate people, any phenomenon or assumption (even supernatural!) if well defined can be tested by the scientific method. That's why we have (among many others) scientific evidence that prayers don't provide any therapeutic effect better than a placebo. See for example about supernatural stuff the following paper *Study of the Therapeutic Effects of Intercessory Prayer (STEP) in cardiac bypass patients: a multicenter randomized trial of uncertainty and certainty of receiving intercessory prayer* [51] for a quite good scientific protocol and analysis or the following paper *Positive therapeutic effects of intercessory prayer in a coronary care unit population* [84] for a typical bad scientific protocol and analysis. There exist also a few meta-analysis on this supernatural topic...

> ❞ *It doesn't matter if a great scientist was religious. What matters is that none of them ever proved their god exists!* ❞

quote 4.4:107.#25

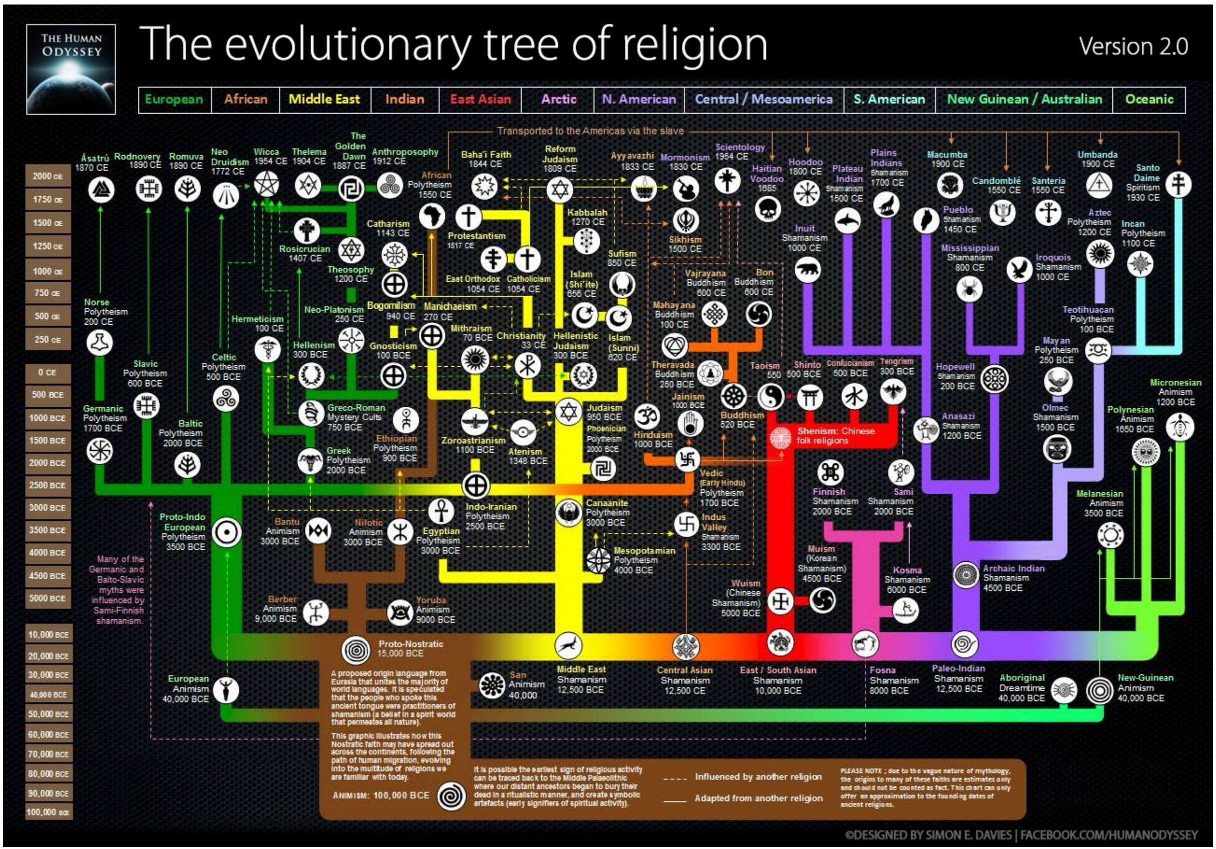
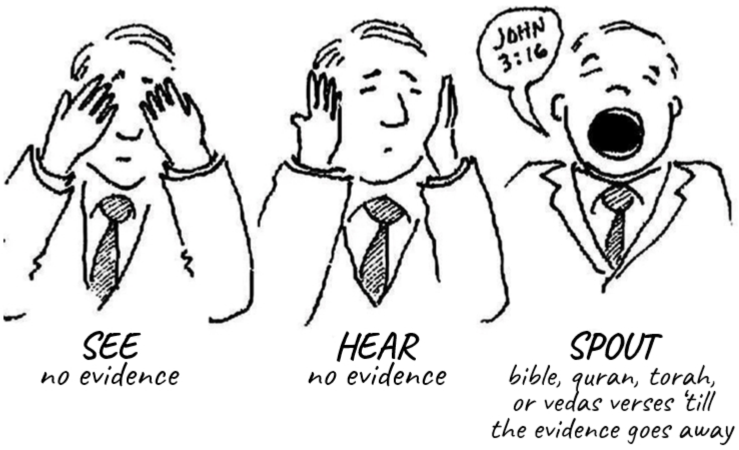

The Human Odyssey
The evolutionary tree of religion
Version 2.0
European
African
Middle East
Indian
East Asian
Arctic
N. American
Central / Mesoamerica
S. American
New Guinean / Australian
Oceanic
Transported to the Americas via the slave
2000 CE
1750 CE
1500 CE
1250 CE
1000 CE
750 CE
500 CE
250 CE
0 CE
500 BCE
1000 BCE
1500 BCE
2000 BCE
2500 BCE
3000 BCE
3500 BCE
4000 BCE
4500 BCE
5000 BCE
10,000 BCE
20,000 BCE
30,000 BCE
40,000 BCE
50,000 BCE
60,000 BCE
70,000 BCE
80,000 BCE
90,000 BCE
100,000 BCE
Ásatrú 1870 CE
Rodnovery 1890 CE
Romuva 1890 CE
Neo Druidism 1772 CE
Wicca 1954 CE
Thelema 1904 CE
The Golden Dawn 1887 CE
Anthroposophy 1912 CE
African Polytheism 1550 CE
Baha'i Faith 1844 CE
Reform Judaism 1809 CE
Ayyavazhi 1833 CE
Mormonism 1830 CE
Scientology 1954 CE
Haitian Voodoo 1685
Hoodoo 1800 CE
Plateau Indian Shamanism 1500 CE
Plains Indians Shamanism 1700 CE
Macumba 1900 CE
Candomblé 1550 CE
Santeria 1550 CE
Umbanda 1900 CE
Santo Daime Spiritism 1930 CE
Catharism 1143 CE
Protestantism 1517 CE
Kabbalah 1270 CE
Rosicrucian 1407 CE
Theosophy 1200 CE
East Orthodox 1054 CE
Catholicism 1054 CE
Islam (Shi'ite) 656 CE
Sufism 850 CE
Sikhism 1500 CE
Inuit Shamanism 1000 CE
Pueblo Shamanism 1450 CE
Aztec Polytheism 1200 CE
Incan Polytheism 1100 CE
Mississippian Shamanism 800 CE
Iroquois Shamanism 1000 CE
Vajrayana Buddhism 600 CE
Bon Buddhism 600 CE
Mahayana Buddhism 100 CE
Norse Polytheism 200 CE
Hermeticism 100 CE
Neo-Platonism 250 CE
Bogomilism 940 CE
Manichaeism 270 CE
Mithraism 70 BCE
Christianity 33 CE
Hellenistic Judaism 300 BCE
Islam (Sunni) 620 CE
Teotihuacan Polytheism 100 BCE
Mayan Polytheism 250 BCE
Hellenism 300 BCE
Gnosticism 100 BCE
Theravada Buddhism 250 BCE
Taoism 550
Shinto 500 BCE
Confucianism 500 BCE
Tengrism 300 BCE
Hopewell Shamanism 200 BCE
Slavic Polytheism 600 BCE
Celtic Polytheism 500 BCE
Jainism 1000 BCE
Micronesian Animism 1200 BCE
Germanic Polytheism 1700 BCE
Greco-Roman Mystery Cults 750 BCE
Judaism 950 BCE
Buddhism 520 BCE
Anasazi Shamanism 1200 BCE
Olmec Shamanism 1500 BCE
Polynesian Animism 1650 BCE
Baltic Polytheism 2000 BCE
Ethiopian Polytheism 900 BCE
Zoroastrianism 1100 BCE
Atenism 1348 BCE
Phoenician Polytheism 2000 BCE
Hinduism 1000 BCE
Shenism: Chinese folk religions
Greek Polytheism 2000 BCE
Vedic (Early Hindu) Polytheism 1700 BCE
Finnish Shamanism 2000 BCE
Sami Shamanism 2000 BCE
Melanesian Animism 3500 BCE
Proto-Indo European Polytheism 3500 BCE
Bantu Animism 3000 BCE
Nilotic Animism 3000 BCE
Egyptian Polytheism 3000 BCE
Indo-Iranian Polytheism 2500 BCE
Canaanite Polytheism 3000 BCE
Indus Valley Shamanism 3300 BCE
Many of the Germanic and Balto-Slavic myths were influenced by Sami-Finnish shamanism.
Mesopotamian Polytheism 4000 BCE
Muism (Korean Shamanism) 4500 BCE
Kosma Shamanism 6000 BCE
Archaic Indian Shamanism 4500 BCE
Wuism (Chinese Shamanism) 5000 BCE
Berber Animism 9,000 BCE
Yoruba Animism 9000 BCE
Proto-Nostratic 15,000 BCE
European Animism 40,000 BCE
San Animism 40,000
Middle East Shamanism 12,500 BCE
Central Asian Shamanism 12,500 CE
East / South Asian Shamanism 10,000 BCE
Fosna Shamanism 8000 BCE
Paleo-Indian Shamanism 12,500 BCE
Aboriginal Dreamtime 40,000 BCE
New-Guinean Animism 40,000 BCE
A proposed origin language from Eurasia that unites the majority of world languages. It is speculated that the people who spoke this ancient tongue were practitioners of shamanism (a belief in a spirit world that permeates all nature).
This graphic illustrates how this Nostratic faith may have spread out across the continents, following the path of human migration, evolving into the multitude of religions we are familiar with today.
Animism: 100,000 BCE
It is possible the earliest sign of religious activity can be traced back to the Middle Palaeolithic where our distant ancestors began to bury their dead in a ritualistic manner, and create symbolic artefacts (early signifiers of spiritual activity).
---- Influenced by another religion
—— Adapted from another religion
PLEASE NOTE ; due to the vague nature of mythology, the origins to many of these faiths are estimates only and should not be counted as fact. This chart can only offer an approximation to the founding dates of ancient religions.
©DESIGNED BY SIMON E. DAVIES | FACEBOOK.COM/HUMANODYSSEY

Figure 4.19: **Evolutionary tree of religions (author: Simon E. Davies)**

Facts enrages believers, who prompts them to solipsism. Their only way to defend beliefs that are not concordant with scientific data is to challenge reality itself (that's why before debating with such people scientists have to agree with them about a definition of "reality" otherwise any debate is pointless)! So every time you argue with believers, because they can't address the facts they will attack the epistemology. As it is as if their position is so weak that their gods couldn't be real unless reality isn't...


JOHN 3:16
SEE
no evidence
HEAR
no evidence
SPOUT
bible, quran, torah, or vedas verses 'till the evidence goes away

quote 4.4:108.#26

> *Faith leads human astray and is a function of his lack of scientific knowledge.*

4.4.1 Baloney detection kit

Through their training, scientists are equipped with what Carl Sagan name the "baloney detection kit" or "bullshit detection kit" that is a set of cognitive tools and techniques that fortify the mind against penetration by falsehoods and to draw boundaries between science and pseudo-science. It isn't merely a tool of science, it contains invaluable tools of healthy scepticism that apply just as elegantly, and just as necessarily, to everyday life. By adopting the kit, we can all shield ourselves against clueless guile and deliberate manipulation.

There are many version of these detection tool but here is a quite complete one (but still incomplete by construction) a proposed by Michael Shermer (founding publisher of Skeptic Magazine and author of *The Borderlands of Science*):

01 ***How reliable is the source of the claim?***

Pseudo-scientists often appear quite reliable, but when examined closely, the facts and figures they cite are distorted, taken out of context or occasionally even fabricated. Of course, everyone makes some mistakes. And as historian of science Daniel Kevles showed so effectively in his book The Baltimore Affair, it can be hard to detect a fraudulent signal within the background noise of sloppiness that is a normal part of the scientific process. The question is, Do the data and interpretations show signs of intentional distortion? When an independent committee established to investigate potential fraud scrutinized a set of research notes in Nobel laureate David Baltimore's laboratory, it revealed a surprising number of mistakes. Baltimore was exonerated because his lab's mistakes were random and nondirectional... So in science, there are no authorities and therefore claims are also not proofs (believing in something does not make it true!!!). At most, there are experts!

02 ***Does this source often make similar claims?***

Pseudo-scientists have a habit of going well beyond the facts. Flood geologists (creationists who believe that Noah's flood can account for many of the Earth's geologic formations) consistently make outrageous claims that bear no relation to geological science. Of course, some great thinkers do frequently go beyond the data in their creative speculations. Thomas Gold of Cornell University is notorious for his radical ideas, but he has been right often enough that other scientists listen to what he has to say. Gold proposes, for example, that oil is not a fossil fuel at all but the by-product of a deep, hot biosphere (micro-organisms living at unexpected depths within the crust). Hardly any Earth scientists with whom I have spoken think Gold is right, yet they do not consider him a crank. Watch out for a pattern of fringe thinking that consistently ignores or distorts data.

03 ***Have the claims been verified by another source?***

Typically pseudo-scientists make statements that are unverified or verified only by a source within their own belief circle. We must ask, Who is checking the claims, and even who is checking the checkers? The biggest problem with the cold fusion debacle, for instance, was not that Stanley Pons and Martin Fleischman were wrong. It was that they announced their spectacular discovery at a press conference before other laboratories verified it. Worse, when cold fusion was not replicated, they continued to cling to their claim. Outside verification is crucial to good science.

04 ***How does the claim fit with what we know about how the world works?***

An extraordinary claim must be placed into a larger context to see how it fits. When people claim that the Egyptian pyramids and the Sphinx were built more than 10,000 years ago by an unknown, advanced race, they are not presenting any context for that earlier civilization. Where are the rest of the artefacts of those people? Where are their works of art, their weapons, their clothing, their tools, their trash? Archaeology simply does not operate this way.

05 ***Has anyone gone out of the way to disprove the claim, or has only supportive evidence been sought?***

This is the "confirmation bias" (we will come back on cognitive biases in the section of Decision Theory page 7381), or the tendency to seek confirmatory evidence and to reject or ignore disconfirmatory evidence. The confirmation bias is powerful, pervasive and almost impossible for any of us to avoid. It is why the methods of science that emphasize checking and rechecking, verification and replication, and especially attempts to falsify a claim, are so critical.

06 ***Does the preponderance of evidence point to the claimant's conclusion or to a different one?***

The theory of evolution, for example, is "proved" through a convergence of evidence from a number of independent lines of inquiry. No one fossil, no one piece of biological or palaeontological evidence has "evolution" written on it; instead tens of thousands of evidentiary bits add up to a story of the evolution of life. Creationists conveniently ignore this confluence, focusing instead on trivial anomalies or currently unexplained phenomena in the history of life.

07 ***Is the claimant employing the accepted rules of reason and tools of research, or have these been abandoned in favour of others that lead to the desired conclusion?***

A clear distinction can be made between SETI (Search for Extraterrestrial Intelligence) scientists and UFOlogists. SETI scientists begin with the null hypothesis that ETIs do not exist and that they must provide concrete evidence before making the extraordinary claim that we are not alone in the universe. UFOlogists begin with the positive hypothesis that ETIs exist and have visited us, then employ questionable research techniques to support that belief, such as hypnotic regression (revelations of abduction experiences), anecdotal reasoning (countless stories of UFO sightings), conspiratorial thinking (governmental cover-ups of alien encounters), low-quality visual evidence (blurry photographs and grainy videos), and anomalistic thinking (atmospheric anomalies and visual misperceptions by eyewitnesses).

08 ***Is the claimant providing an explanation for the observed phenomena or merely denying the existing explanation?***

This is a classic debate strategy-criticize your opponent and never affirm what you believe to avoid criticism. It is next to impossible to get creationists to offer an explanation for life (other than "God did it"). Intelligent Design (ID) creationists have done no better, picking away at weaknesses in scientific explanations for difficult problems and offering in their stead. "ID did it." This stratagem is unacceptable in science.

(09) ***If the claimant proffers a new explanation, does it account for as many phenomena as the old explanation did?***

Many HIV/AIDS sceptics argue that lifestyle causes AIDS. Yet their alternative theory does not explain nearly as much of the data as the HIV theory does. To make their argument, they must ignore the diverse evidence in support of HIV as the causal vector in AIDS while ignoring the significant correlation between the rise in AIDS among haemophiliacs shortly after HIV was inadvertently introduced into the blood supply.

⑩ ***Do the claimant's personal beliefs and biases drive the conclusions, or vice versa?***

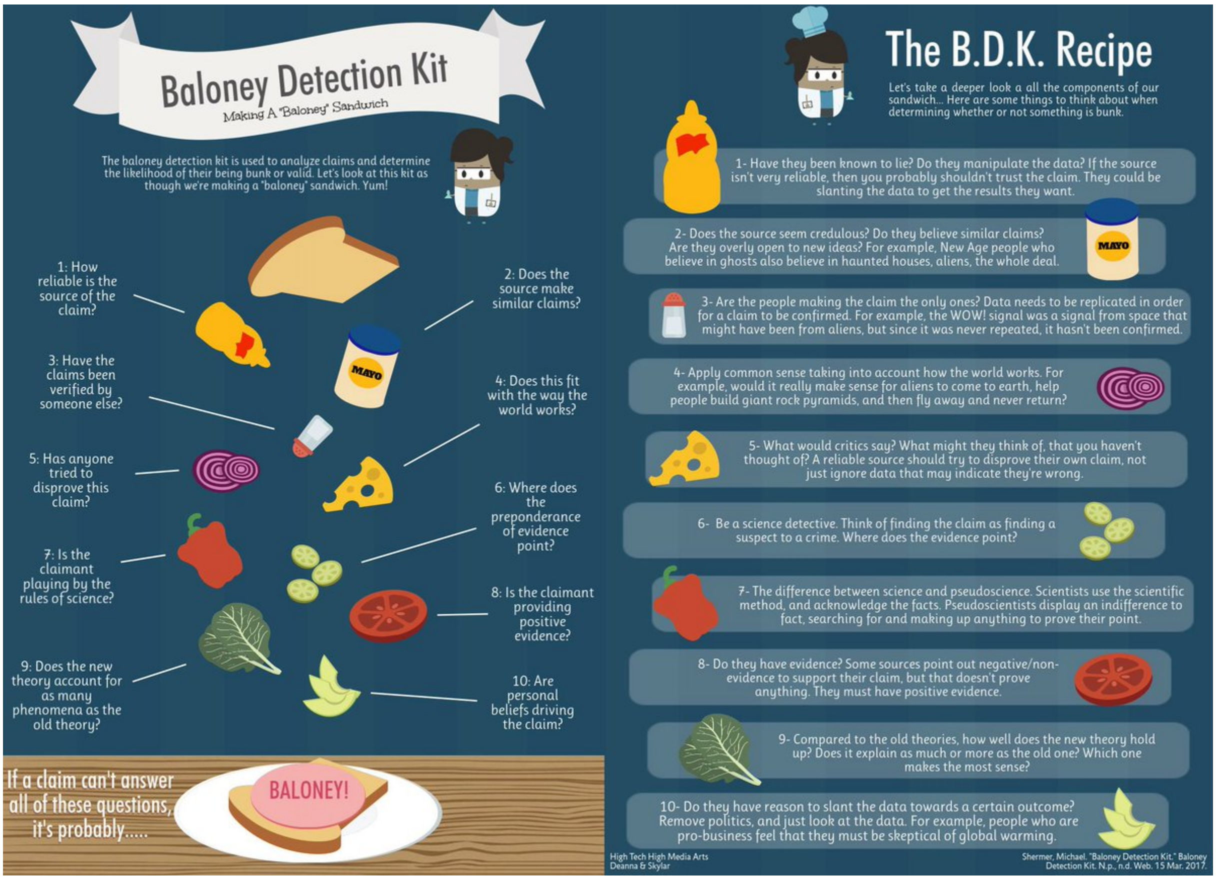
All scientists hold social, political and ideological beliefs that could potentially slant their interpretations of the data (this is a "confirmation bias" also named "cherry picking" that is also by non-scientists the main cause of rejecting science results and tools), but how do those biases and beliefs affect their research in practice? Usually during the peer review system, such biases and beliefs are rooted out, or the paper or book is rejected.

Some humans and especially believers (not just religious believers but on any subject for which there is no experimental evidence beyond reasonable doubt) are good at what is named "concordism". The underlying idea of concordism is to interpret texts or find random analogies in Nature (obviously having zero knowledge of combinatorics and probabilities) to justify that all the inventions or discoveries of modern science have been announced in their "holy" books or invented by people belonging to the same belief group as themselves hundreds or thousands of years before others. Obviously all this masked under a hazy rhetoric, analogies that make no sense (compare statistical studies with simple texts...), no statistical methodology, omitting to quote contradictory facts or conclusions, and with a total absence of cross-referencing sources and peer review! This obscurantism at its finest!!!

It must be noted that this concordist approach is the prerogative of people with little education or indoctrinated since childhood (plus those suffering from mental disorders for various reasons) who are not able to identify their own biases, who are closed to any contradiction, and who are not capable of identifying the real complexity (multifactoriality) and the nuances of even a simple subject (they tend to binarize everything in True/False).

> ❞ *You can beat 100,000 scholars with one fact, but you will very likely never beat an idiot even with 100 facts.* ❞

quote 4.4:111.#27



Baloney Detection Kit

Making A "Baloney" Sandwich

The baloney detection kit is used to analyze claims and determine the likelihood of their being bunk or valid. Let's look at this kit as though we're making a "baloney" sandwich. Yum!

1: How reliable is the source of the claim?

2: Does the source make similar claims?

3: Have the claims been verified by someone else?

4: Does this fit with the way the world works?

5: Has anyone tried to disprove this claim?

6: Where does the preponderance of evidence point?

7: Is the claimant playing by the rules of science?

8: Is the claimant providing positive evidence?

9: Does the new theory account for as many phenomena as the old theory?

10: Are personal beliefs driving the claim?

If a claim can't answer all of these questions, it's probably.....

The B.D.K. Recipe

Let's take a deeper look a all the components of our sandwich... Here are some things to think about when determining whether or not something is bunk.

1- Have they been known to lie? Do they manipulate the data? If the source isn't very reliable, then you probably shouldn't trust the claim. They could be slanting the data to get the results they want.

2- Does the source seem credulous? Do they believe similar claims? Are they overly open to new ideas? For example, New Age people who believe in ghosts also believe in haunted houses, aliens, the whole deal.

3- Are the people making the claim the only ones? Data needs to be replicated in order for a claim to be confirmed. For example, the WOW! signal was a signal from space that might have been from aliens, but since it was never repeated, it hasn't been confirmed.

4- Apply common sense taking into account how the world works. For example, would it really make sense for aliens to come to earth, help people build giant rock pyramids, and then fly away and never return?

5- What would critics say? What might they think of, that you haven't thought of? A reliable source should try to disprove their own claim, not just ignore data that may indicate they're wrong.

6- Be a science detective. Think of finding the claim as finding a suspect to a crime. Where does the evidence point?

7- The difference between science and pseudoscience. Scientists use the scientific method, and acknowledge the facts. Pseudoscientists display an indifference to fact, searching for and making up anything to prove their point.

8- Do they have evidence? Some sources point out negative/non-evidence to support their claim, but that doesn't prove anything. They must have positive evidence.

9- Compared to the old theories, how well does the new theory hold up? Does it explain as much or more as the old one? Which one makes the most sense?

10- Do they have reason to slant the data towards a certain outcome? Remove politics, and just look at the data. For example, people who are pro-business feel that they must be skeptical of global warming.

High Tech High Media Arts
Deanna & Skylar

Shermer, Michael. "Baloney Detection Kit." Baloney Detection Kit. N.p., n.d. Web. 15 Mar. 2017.

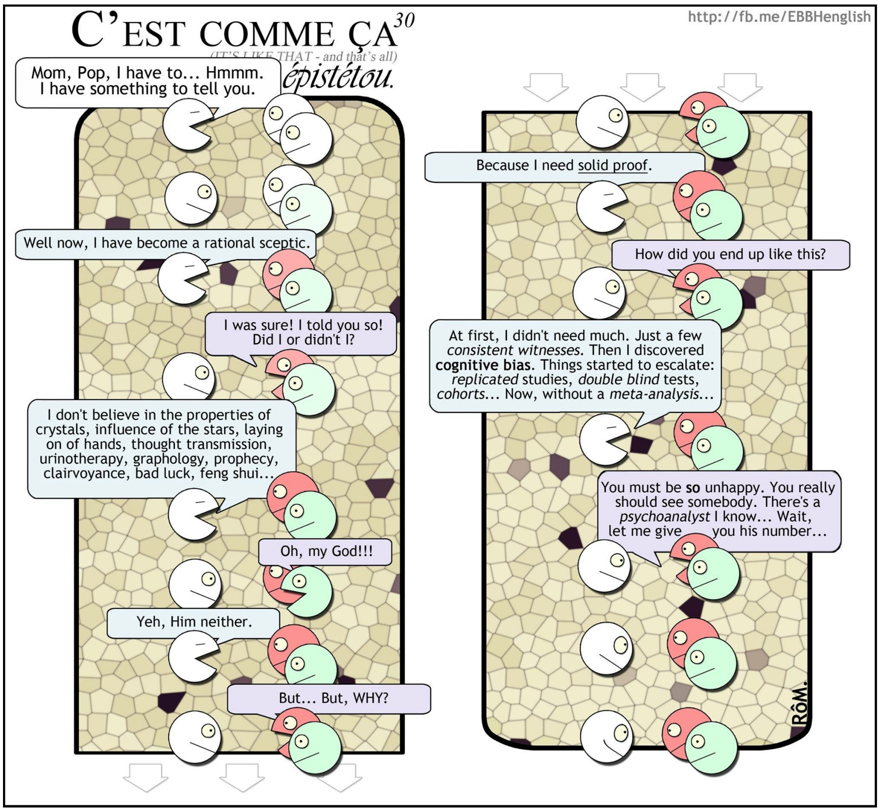
By fine tuning we can go far more about reasoning fallacies. Here is a small non-exhaustive list:

1. Ad hominem: An ad hominem argument attacks the messenger, not the message itself.
2. Argument from authority: Argument that relies on the identity of an authority rather than the components of the argument itself.
3. Argument from adverse consequences: Saying that because the implications of a statement being true would generate negative results, it must not be true.
4. Appeal to ignorance: If something is not known to be false, it must be true.
5. Special pleading: Stating a universal principle, then insisting that it doesn't apply to your assertions for some reason.
6. Begging the question/ assuming the answer: This occurs when a statement has an unproven premise. It is also named "circular reasoning" or "circular logic".
7. Observational selection: Looking at only positive evidence while ignoring the negative and vice versa.
8. Statistics of small numbers: Using small numbers in order to report large percentage increases.

9. Misunderstanding of the nature of statistics: Ignorance about central statistical assumptions and the definition of metrics (the confusion of correlation and causation, the sample size and hate of maths bias are well known example).
10. Post hoc, ergo propter hoc: Basing an effect on a cause only on the basis of chronology.
11. Excluded middle, or false dichotomy: Portraying an issue or argument as having only two options and no spectrum in between.
12. Short-term vs. long-term: Assuming a current trend has remained constant throughout its history and will continue to do so in the future, even though no evidence suggests such an extrapolation is justified.
13. Slippery slope, related to excluded middle: Saying something is wrong because it is next to or loosely related to something wrong.
14. Suppressed evidence and half-truths: Drawing an unwarranted conclusion from premises that are at least in part correct.
15. Weasel words: The usage of vague, non-specific references.

In addition to teaching us what to do when evaluating a claim to knowledge, any good baloney detection kit must also teach us what not to do. It helps us recognize the most common and perilous fallacies of logic and rhetoric. Many good examples can be found in religion and politics, because their practitioners are so often obliged to justify two contradictory propositions.

Finally, we would like to quote Antoine Lavoisier: «*The physicist may also, in the silence of his laboratory and his cabinet, perform patriotic functions; he can thanks to his works reduce the mass of evils which afflict happiness and, had he not, contributed by the new roads that he opened to himself, only to delay of a few years, of a few days, the average life of humans, he could also aspire to the glorious title of benefactor of humanity.*»


C'EST COMME ÇA[30]
(IT'S LIKE THAT - and that's all)
épistétou.
http://fb.me/EBBHenglish
Mom, Pop, I have to... Hmmm. I have something to tell you.
Well now, I have become a rational sceptic.
I was sure! I told you so! Did I or didn't I?
I don't believe in the properties of crystals, influence of the stars, laying on of hands, thought transmission, urinotherapy, graphology, prophecy, clairvoyance, bad luck, feng shui...
Oh, my God!!!
Yeh, Him neither.
But... But, WHY?
Because I need solid proof.
How did you end up like this?
At first, I didn't need much. Just a few consistent witnesses. Then I discovered cognitive bias. Things started to escalate: replicated studies, double blind tests, cohorts... Now, without a meta-analysis...
You must be so unhappy. You really should see somebody. There's a psychoanalyst I know... Wait, let me give you his number...
RôM.

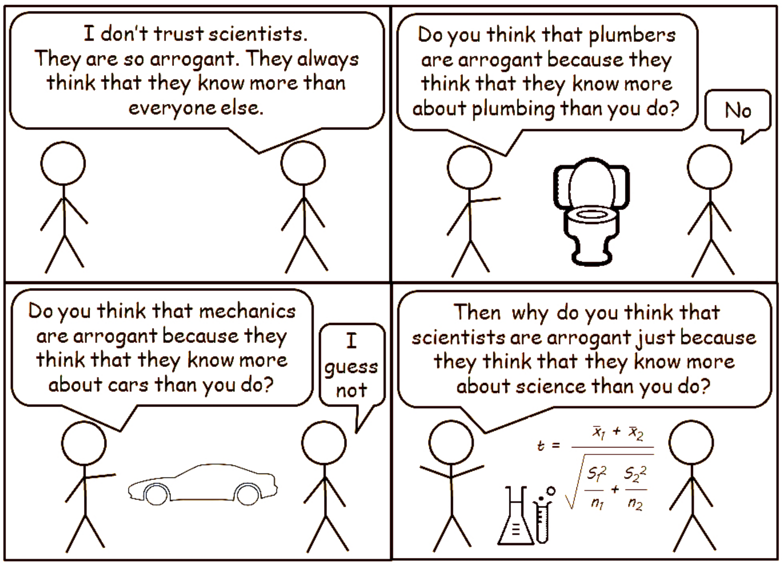
Caution! Be extremely careful in life with non peer review textbooks (or videos/podcasts) distributed by non-academic publisher whose content use pleasant story telling, psychological techniques (beautiful design, well-chosen music, etc.), flattery and extraordinary claims (without referencing to a quite large bibliography of empirical evidence reported in papers that respect the scientifical publication rules as given at page 70) to manipulate you and get your money by selling various services (books, events, trainings, etc.) and unfalsifiable promises! Science at the opposite will mostly provide you highly complex unpleasant knowledge and information and also (unperfect) tools that work and are useful in real life...

4.5. Scientific communication backfire

Another point that may be important to highlight about scientific communication: Scientists need to stop thinking that explaining science will fix all issues from the world and people's biases (at best they can be mitigated a little bit)! Especially nowadays where conspiracies about flat Earth, vaccines, climate change, Moon landing, etc. spread out thanks to social networks and that many mainstream media don't know how to communicate scientific papers correctly (or do it on purpose just for clickbait...). For example, in 12010 (holocene calendar) a landmark study (see [422]) showed that confronting people with hard evidence can backfire, making them more entrenched in their biases and misperceptions.

The main reasons are as follow and apply outside the case where people come listen to scientists in the context of a conference or seminar or participate on scientific debate on social networks live stream:

1. Some people don't really want to listen about the Scientific Method especially when they never asked themselves «what is the level of evidence?», «is it the best method?», «is this not a bias?». Moreover if they use the argument «...their science...» just to point out that scientists may be wrong (without knowing that scientists like to be wrong and that's how science works...!) about a given topic during a debate or discussion, correcting this fallacy may just take such individuals out of their comfort zone and make them even more hate (or reject) science and therefore we get a backfire effect.


I don't trust scientists. They are so arrogant. They always think that they know more than everyone else.
Do you think that plumbers are arrogant because they think that they know more about plumbing than you do?
No
Do you think that mechanics are arrogant because they think that they know more about cars than you do?
I guess not
Then why do you think that scientists are arrogant just because they think that they know more about science than you do?
$t = \frac{\bar{x}_1 + \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

thelogicofscience.com

2. Most humans are full of biases and they don't like to admit that this is a true fact as they assume the human is the top species of evolution and therefore could have such biases.... So when you explain such people that they have biases, you just point out that they are not reliable in some aspect of their observation of Nature. Speak about bias only if people ask you to do so or you will very likely get again a backfire effect!

3. A significant percentage of people believe that their personal experience is more robust than the hundred of years of peer-review, tests, checks of the "scientific method" that has seems so far, if not THE best, at least the best one investigation method of the Nature we know nowadays.

Remark

"Anecdotally" there are also many other reasons (humanly understandable...) why science or members of the scientific community are defamed, such as (non-exhaustive list in random order) the fact that: popular science uses analogies that are distressing false equivalences and annoys some individuals, scientific communication uses sometimes religious vocabulary (such as "God particle" instead of "Higgs boson") or unsuitable vocabulary (such as "Big Bang theory" instead of "cosmic inflation theory") to describe phenomena and this generates a mixture of genres that strongly irritates some individuals, science often imposes itself in small classes only with rhetoric and without mathematical proofs or experimental evidence which can make it confused with a religious dogma and therefore a belief which tends to conflict with people with other beliefs, quantum physics is non-intuitive (non-locality, phenomena without cause, spontaneous appearance of vacuum energy, etc.) and greatly destabilizes the knowledge frame of the macroscopic world, some people have been ejected from the school system because of a bitter failure in the scientific branches and they find themselves marked for life (impossibility of doing the job of their dreams), science contradicts/challenges the dogmas of all religions/beliefs (evolution, genesis, non-existence of the soul, mental illnesses of pseudo-prophets, false miracles, etc.), scientific publications have destroyed (or put into puerile) the business of certain charlatans (clairvoyance, homeopathy, physiotherapy, acupuncture, etc.), some individuals have lost their jobs because of technological evolution falsely attributed to science (while it is the engineers who must find practical applications for science), some individuals have experienced a serious medical error and associate Medicine with Science (while medicine is an art and that doctors are not scientists), the worldview of some individuals is Manichean while science is probabilistic and certainties ("absolute truths") are more the exception than the rule, science is iterative and constantly questions itself and evolves destabilizes some people who need stability/certainties in their lives, some are jealous of the high salaries of scientists or their high level of postgraduate university training (anti-intellectualism), some individuals think that it is science and scientists who develop nuclear weapons and bombs while they are mostly engineers (and again: engineers are not scientists!), the results of science are above human legal laws is poorly accepted by some, the loss of loved ones or serious accidents that people cannot simply associate with the hazard (probabilities) and then need to find a culprit to designate, some individuals are afraid of developments in engineering often wrongly associated with science (artificial intelligence, genetic modifications, injections of treatments), Western science has given Western names to certain theorems which were not discovered by Westerners (this causes jealousy in certain cultures), scientific culture is over-represented in certain countries which have entered into wars or competition with other countries and have totally crushed them with their techno-scientific advances, etc. etc. etc.!

> *No amount of peer-reviewed evidence will ever persuade an idiot, all the more if the latter is scientifically illiterate and doesn't master top level statistical calculations and the subtleties of the scientific method.*

quote
4.5:117.#28

Now let us quote some paragraphs of an excellent article of Tim Requarth as it is almost perfect:

«The theory many scientists seem to swear by is technically known as the deficit model, which states that people's opinions differ from scientific consensus because they lack scientific knowledge. In 12010 (holocene calendar), Dan Kahan, a Yale psychologist, essentially proved this theory wrong. He surveyed over 1,500 Americans, classifying each person's "cultural worldview" on a scale that roughly correlates with politically liberal or conservative. He then assessed each person's scientific literacy with questions such as "True or False: Electrons are smaller than atoms". Finally, he asked them about climate change. If the deficit model were correct, Kahan reasoned, then people with increased scientific literacy, regardless of worldview, should agree with scientists that climate change poses a serious risk to humanity.

That's not what he found. Instead, Kahan found that increased scientific literacy actually had a small negative effect: The conservative-leaning respondents who knew the most about science thought climate change posed the least risk. Scientific literacy, it seemed, increased polarization. In a later study, Kahan added a twist: He asked respondents what climate scientists believed. Respondents who knew more about science generally, regardless of political leaning, were better able to identify the scientific consensus-in other words, the polarization disappeared. Yet, when the same people were asked for their own opinions about climate change, the polarization returned. It showed that even when people understand the scientific consensus, they may not accept it.

> *Having no postgraduate education on this subject, nor having any biases and having done no research in a lab at a professional level whatsoever, I don't understand it. Therefore, it doesn't make sense and is false!*
>
> Average social network user

quote
4.5:117.#29

The takeaway is clear: Increasing science literacy alone won't change minds. In fact, well-meaning attempts by scientists to inform the public might even backfire. Presenting facts that conflict with an individual's worldview, it turns out, can cause people to dig in further. Psychologists, aptly, dubbed this the "backfire effect".

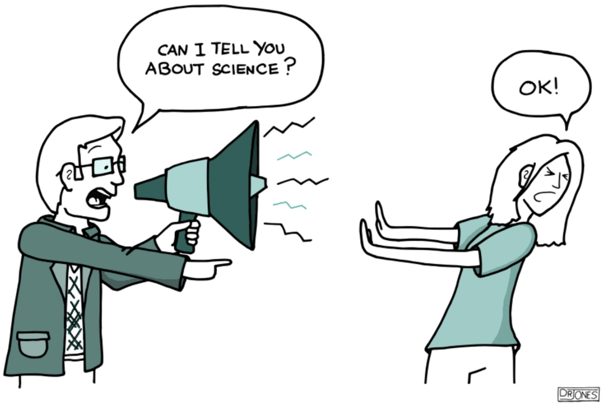

CAN I TELL YOU ABOUT SCIENCE?
OK!
DRJONES

Figure 4.20: **source:** https://www.ratbotcomics.com, **author: Dr. Jones**

If scientists simply want to explain science to a curious audience, disseminate their research more broadly, or write for fun, this doesn't matter much. But if scientists are motivated to change minds-and many enrolled in science communication workshops do seem to have this goal-they will be sorely disappointed.

That's not to say scientists should return to the bench and keep their mouths shut. They should just realize that closing the "information gap" isn't the goal. And instead, they need to learn how to communicate science strategically.

There are obvious reasons why science communication is a necessary and worthwhile endeavour, but a huge one is that there's a politically motivated push to destabilize scientific authority. At a Heartland Institute conference last month, Lamar Smith, the Republican chairman of the House science committee, told attendees he would now refer to "climate science" as "politically correct science", to loud cheers. This lumps scientists in with the nebulous "left" and, as Daniel Engber pointed out here in Slate about the upcoming March for Science, rebrands scientific authority as just another form of elitism.

Is it any surprise, then, that lectures from scientists built on the premise that they simply know more (even if it's true) fail to convince this audience? Rather than fill the information deficit by building an arsenal of facts, scientists should instead consider how they deploy their knowledge. They may have more luck communicating if, in addition to presenting facts and figures, they appeal to emotions. This could mean not simply explaining the science of how something works but spending time on why it matters to the author and why it ought to matter to the reader. Research also shows that science communicators can be more effective after they've gained the audience's trust. With that in mind, it may be more worthwhile to figure out how to talk about science with people they already know, through, say, local and community interactions, than it is to try to publish explainers on national news sites. And they might consider writing op-eds for their local papers, focusing on why science matters to their particular communities.

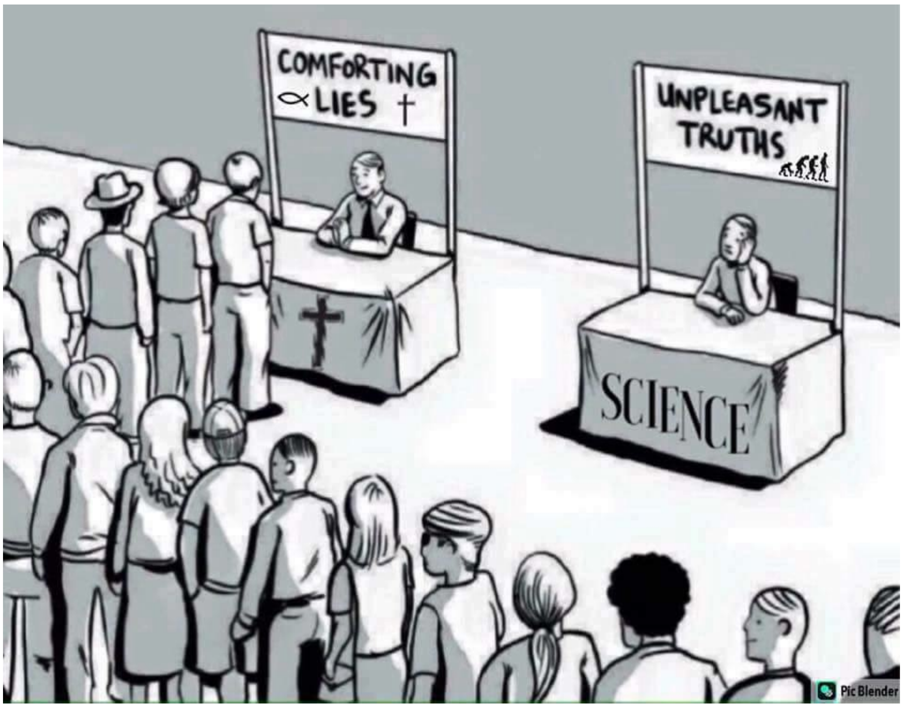
Scientists can also learn to avoid certain pitfalls. I spoke with Gretchen Goldman, research director of the Union of Concerned Scientists' Center for Science and Democracy, which offers communication and advocacy workshops. A counter-intuitive lesson she's learned is that refuting stories that deny climate change by addressing each claim and explaining why it's wrong is not that productive. In fact, it could be counter-productive: "If you repeat the myth, that's the part people remember even if you immediately debunk it", she says. A better approach, she suggests, is to reframe the issue. Don't just keep explaining why climate change is real, explain how climate change will hurt public health or the local economy. Communication that appeals to values, not just intellect, research shows, can be far more effective.

[...] But the obstacles faced by science communicators are not epistemological but cultural. The skills required are not those of a university lecturer but a rhetorician.

So it's an admirable goal to communicate about science, but almost certainly destined to fail. This is because the way most scientists think about science communication - that just explaining the real science better will help - is quite wrong. It's so wrong that it may have the opposite effect of what they're trying to achieve. [...]»

> ❝ *Silence is sometimes the best answer to the fools.* ❞

quote
4.5:119.#30



> ❝ *As a scientist (and atheist) you will often be hated by all those who believe in fairy tales and think that their opinions, anecdotes, circular logic and scientific illiteracy are worth it, but those who stands firm to the Scientific Method and to the precautionary principle will be saved and will help to save the haters. Comforting lies have always and are still taking the elevator and the Scientific Method takes the stairs, the lies will come first but one day or another the Science will always catch up with its unpleasant evidence.* ❞

quote
4.5:119.#31

4.6. Is Science dogmatic?

We will here mainly repeat things that we have already mentioned earlier above. However as some scientifically illiterate people still think in this beginning of the 121st century (holocene calendar) that YouTube videos containing some rhetorical monologue[61] or some books without proofs and without statistical data analysis constitute some sort of "evidence" it is necessary maybe to come back on a few topics but with a different perspective[62]!

For Science, if something exists with evidence beyond reasonable doubt (BRD), at its actual state of knowledge, then it means it can be measured! If it cannot be measured or is ill-defined, well, then science can't provide evidence that it doesn't exist (don't forget that the purpose of Science is to refute models and if it fails to do so a long time enough then a Model may take the status of "Theory"!). In a simplistic and slightly out of context metaphor, this is equivalent to say that for the blind (taking away all their other senses) the world doesn't exist with strong evidence because they cannot see it nor feel it. Some people then say that Science is materialist! However this doesn't mean at all that Science has failed as an investigation method, however it just means that Science knows it doesn't know everything otherwise it would stopped and that the scientific method must be corrected constantly according to the new available evidences.

DATA
over
DOGMA

Typically Science doesn't reject completely parapsychology powers or the existence of one of the thousand of deities created around the World by humans even if actually all experimental tests have rejected them with strong evidence (but not definitively!). The advocates of parapsychology or religions (or of some alternative medicines) may argue that Science can't prove nor disprove what it can't measure directly or indirectly. And they are very likely absolutely right! Science can only fail to reject with a given level of evidence if something exist or not at its actual level of knowledge. So if someone argue that flying unicorns exist without providing a protocol so that thousands of other people can check this fact in a reproducible way... then Science (scientists) humbly say us that: «*we don't have any evidence beyond reasonable doubt at our actual state of knowledge that unicorn exist or not!*»

So it's not Science that is dogmatic, nor the huge majority of the scientific community. But only some bad educated and biased scientists (nobody is perfect...).

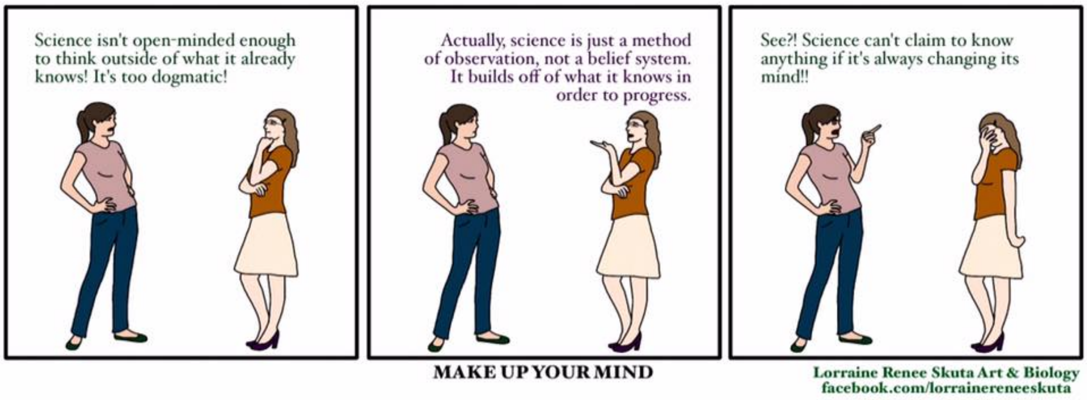
---

[61] As already said people should not trust monologues - whether on radio, television or any social network - because there are no experts to counter the possible wrong arguments or ill-defined concepts that may lead a part of the audience to wrong speculative interpretations! In addition, humans under the stress of knowing that they are recorded are naturally prone to vocabulary errors and that's without counting the more than 200 cognitive biases in the brain which sometimes lead to the erroneous simplification of complex thoughts... !

[62] And please don't forget that the purpose of Science is not the "Truth" but it's only a tool that explains quite well things that we see or feel following the best actual models. And also please don't forget that quoting a book, a famous scientist, a blog or a YouTube video is only at best an evidence of level 2 or 3!

Remark

We will also see later that Gödel's incompleteness theorem proves, from the non-conservativity argument, that there exist in certain mathematical theories (i.e. in certain systems of axioms) truth sentences which cannot be considered true without a notion of a more transcendental truth with respect to the basic underlying theory! Which is further proof that Science itself has demonstrated that it cannot in certain systems be a dogma!


Science isn't open-minded enough to think outside of what it already knows! It's too dogmatic!
Actually, science is just a method of observation, not a belief system. It builds off of what it knows in order to progress.
See?! Science can't claim to know anything if it's always changing its mind!!
MAKE UP YOUR MIND
Lorraine Renee Skuta Art & Biology
facebook.com/lorrainereneeskuta

Some humans don't like that scientists may not know everything about something, or that scientists do errors, and that science takes time and that even worst, these same people will maybe never have an answer to their main questions before their death. But that's how Science works actually!!! If anyone finds a more reliable and fast way to investigate the Universe and its phenomena than the actual state of the Scientific Method, then the majority of the scientific community - especially the academic part - is waiting to check it and if it works indeed better to adopt it!

> ❞ *Extraordinary claims require extraordinary evidence.* ❞
>
> Carl Sagan

quote 4.6:121.#32

So some people may ask Science to be truly scientific, meaning, it should question itself regarding the certainty of their own postulates and instruments. But as we already said many times in the previous sections above, this is what researchers do in their daily work!!! They try to find new evidence to reject actual theories, models, methods, postulates or bad instruments... otherwise Science would stop and scientists would lost their job. However the process is slow and takes days, weeks, months, years, decades, centuries and sometimes even millenniums.

Remark

Some people point out that the version of the Universe scientists currently have is only what their instruments, and especially their imagination allows them to understand. And they are right! We - all professional academic researchers - know that fact since centuries in Science and that's why every day we try to push the boundaries of knowledge (hence also of imagination) and to develop new instruments and methods to measure things that were not even known a few decades before! Keep in mind this is what we are paid for! If there is nothing new to discover we will all lost our jobs!

So yes in its current state, Science cannot account very well for consciousness phenomena, synchronicity, or even near death experiences or spontaneous epiphanies, etc. (even if tools like fMRI and EEG help to observe correlations with electric signals). But well educated professionals scientists won't claim they do or don't exist because we don't have yet any experimental reproducible evidence to support any of these positions! Scientists are just waiting for those who claim that such things exist - and even run quite good business with it - to provide us reproducible experimental evidence. Sadly so far all the people who made such claim and run business with them have failed to provide any evidence or were just simply debunked and that's all.

Anyone should ask himself why evangelist pastors or any kind of miraculous healer (...) don't practice their art in the scientific controlled environment of Hospitals rather that in their church (in front of a public in the huge majority biased, uneducated and scientifically illiterate) or with a small group of unknown and suspicious people behind a camera.......

People like Rupert Sheldrake, a PhD holder (from Cambridge) in biochemistry and retired researcher in the field of parapsychology who proposed the concept of morphic resonance... and also author of numerous rewarding books...... explores ten dogmas of Science that should be reconsidered according to his personal and subjective opinion...

Let us present below these ten dogmas and... we will comment them as Dr. Sheldrake doesn't have a PhD in Physics, nor in Cosmology but has a good sense of rhetoric especially when he does a monologue in front a public of neophytes (the reader and Dr. Sheldrake will be able to find the mathematical proof of our answers in the $9,000$ pages of this textbook if they are curious - as at the opposite of Dr. Sheldrake we like proofs and experimental evidence more than rhetorical monologue...):

1. “Nature is mechanic: all the creatures and systems of nature are robots made to follow a given genetic program.”

   ▷ Our comment: We are not sure of how the word "Mechanics" is defined in this sentence (the basics of a debate is to agree on the meaning of words...) but obviously a biological system cannot be compared to a mechanical one (mechanical system doesn't evolve). However as we will prove it in this textbook Nature is Information based at all levels, Probabilistic based at the microscope level, and behaves Statistically and Mechanically at macroscopic level. So as most of the time in Science..., it's not as easy as it seems to be (it seems that Dr. Sheldrake has a binary vision of the World and of the Universe that is quite surprising given his supposed level of education).

2. “Matter is unconscious: Plants, stars, animals and elements are material things that are and cannot have a consciousness of themselves.”

   ▷ Our comment: First we should ask how is "consciousness" defined by Dr. Sheldrake. Secondly, who asserted that in the scientific community? Is there a written scientific consensus on this subject or it comes out from the imagination of Dr. Sheldrake?

3. “The laws of nature are fixed: At the moment of the Big Bang (i.e. starting point of the expansion of our universe) all the necessary constants until the end of time were established. The habits of nature do not evolve.”

   ▷ Our comment: What scientific consensus stated that? That’s very likely wrong as actually we have a few mathematical-physics models that proves that the constants of the Universe may have changed and also experimental evidence that maybe the laws of Nature may have changed. But one thing is almost sure: there is no scientific consensus yet on that topic!

4. “The amount of nature and energy in the Universe is always the same.”

   ▷ Our comment: The scientific community has strong evidence beyond reasonable doubt of that statement for the observable Universe. However the dynamic of the Universe disprove energy conservation at large scales (this is derived from Noether theorems!). Notice that we have however no evidence yet of this statement of energy conservation for the whole Universe (the observable and not observable one).

5. “Nature has no purpose: there is no design in nature in terms of intention and the process of evolution is mechanic.”

   ▷ Our comment: We have strong evidence beyond reasonable doubt indeed that there is no design as the existing design are flawed in many ways. And the process of evolution as proven mathematically in this textbook and experimentally (beyond any reasonable doubt) in laboratories is not mechanical but stochastic.

6. “Biologic Heritage: The plans to produce a living being are composed within the physical matter lodged in their genes.”

   ▷ Our comment: That’s not completely correct. Experimental observations provide us that some basics of the plans are random and influenced by external and internal modifications. Again Dr. Sheldrake gives a binary view of a phenomenon which is far more complex. But this is something we know as typical of scientific illiterate people: they reduce something complex to something simple because they can’t grasp the complexity of the World and of the Universe.

7. “Memory is kept in the brain as material prints: memory is made of proteins and nerve endings organized as a drawer within itself.”

   ▷ Our comment: If it was the case we wouldn’t forget things... That’s because the brain is far more complex and involves probabilities, Bayesian and stochastic process that we know why the human brain forgets things and has sometimes building issues...

8. “Mind is in the head: The mind has a physical connection with the head and the brain, relegating intellectual subordination on the rest of the body.”

   ▷ Our comment: What is the "Mind" for Dr. Sheldrake? As far as we know there is no scientific consensus about its definition. Furthermore is Mind what we observe in MNR scanners?

9. “Phenomena like telepathy are impossible: thoughts have no effect in the world because of number 8 on the list (the mind is in the head).”

   ▷ Our comment: What scientific consensus or community stated that? Actually Science has no evidence that telepathy works or exist, yes (!) - but no well educated scientist would say it is "impossible" (by the way any well educated scientist know that is better to avoid using the word "impossible" about an unknown future or unmeasurable phenomena).

10. “Only mechanical medicine works: It is merely by chance or placebo effects that traditional healing practices or natural remedies have any effect on people’s health.”

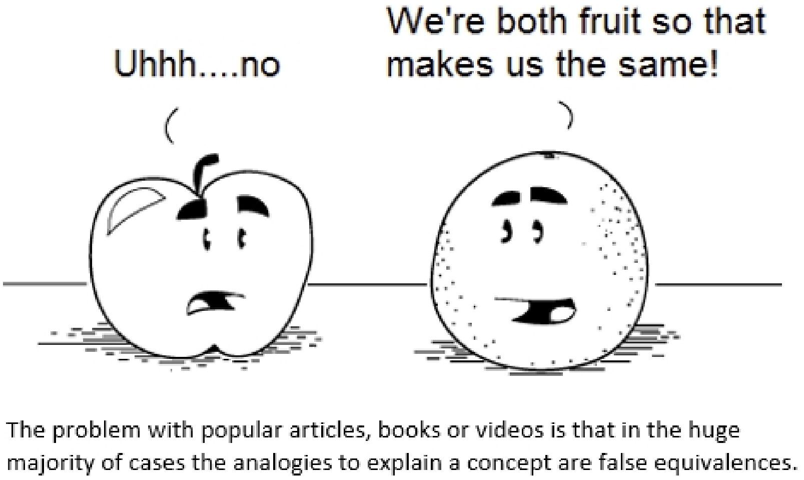
    ▷ Our comment: Again... what scientific consensus or community stated that? That’s not accurate. Maybe scientists agree on the fact that actually not other method than the Scientific Medicine has provided better odds ratio than other medicines. But if one day some people provide robust evidence that traditional healing practices or natural remedies work, then almost surely they will be promoted by the scientific academic community.

If this is the best support Dr. Sheldrake has for his position, then this means that there is no support. It is entirely based on gish gallop and speculation, pointing to common misconceptions and popular opinions rather than evidence and examples of "dogma" in science.

We can understand why someone reading or listening this kind of gish gallop may think that there is some basis of evidence. Gish gallop means making loads of claims without evidence, leaving the opposition the large job of checking each and every claim. Even so, when someone claims the "establishment" is dogmatic, immoral or whatsoever, we may sincerely hope that any reader will critically assess that person claims.

quote 4.6:124.#33

> *A religious belief without dogma is not possible and science is not compatible with any dogma. So science can only be incompatible with religious beliefs.*
>
> ?


Uhhh....no
We're both fruit so that makes us the same!

The problem with popular articles, books or videos is that in the huge majority of cases the analogies to explain a concept are false equivalences.

Figure 4.21: **False equivalence fallacy**

Section quality score: ★★★★☆ 151 votes, 75.23%

5

Arithmetic

Mathematics is the ultimate form of forced art. (unknown)

Contents

5.1 Proof Theory **131**
  5.1.1 Foundations Crisis 134
  5.1.2 Paradoxes 138
    5.1.2.1 Hypothetical-Deductive Reasoning 141
  5.1.3 Propositional Calculus 142
    5.1.3.1 Propositions (premises) 144
    5.1.3.2 Connectors 146
    5.1.3.3 Decision procedures 157
    5.1.3.4 Quantifiers 163
  5.1.4 Predicate Calculus 167
    5.1.4.1 Grammar 168
    5.1.4.2 Languages 169
  5.1.5 Proofs 176
    5.1.5.1 Rules of Proofs 177
  5.1.6 Laws of thought 187
    5.1.6.1 Some famous logical fallacies 199
    5.1.6.2 Is logic a science? 204
5.2 Numbers **209**
  5.2.1 Digital Bases 212
  5.2.2 Type of Numbers 215
    5.2.2.1 Natural Integer Numbers 215
    5.2.2.2 Relative Integer Numbers 222
    5.2.2.3 Rational Numbers 223

5.2.2.4 Irrational Numbers . . . 228
5.2.2.5 Real Numbers . . . 230
5.2.2.6 Transfinite Numbers . . . 239
5.2.2.7 Complex Numbers . . . 243
5.2.2.8 Quaternion Numbers . . . 261
5.2.2.9 Algebraic and Transcendental Numbers . . . 276
5.2.2.10 Universe Numbers (normal numbers) . . . 279
5.2.2.11 Abstract Numbers (variables) . . . 280
5.3 Arithmetic Operators . . . 285
5.3.1 Binary Relations . . . 286
5.3.1.1 Equalities . . . 287
5.3.1.2 Comparators . . . 288
5.3.2 Fundamental Arithmetic Laws . . . 296
5.3.2.1 Addition . . . 296
5.3.2.2 Subtraction . . . 303
5.3.2.3 Multiplication . . . 308
5.3.2.4 Division . . . 313
5.3.2.5 n-root . . . 320
5.3.3 Arithmetic Polynomials . . . 323
5.3.4 Absolute Value . . . 324
5.3.5 Calculation Rules (operators priorities) . . . 328
5.4 Number Theory . . . 333
5.4.1 Principle of good order . . . 333
5.4.2 Induction Principle . . . 334
5.4.3 Divisibility . . . 336
5.4.3.1 Euclidean Division . . . 338
5.4.3.2 Euclidean Algorithm . . . 343
5.4.3.3 Least Common Multiple . . . 347
5.4.3.4 Fundamental Theorem of Arithmetic . . . 352
5.4.3.5 Congruences (modular arithmetic) . . . 356
5.4.3.6 Continued fractions . . . 369
5.5 Set Theory . . . 378
5.5.1 Zermelo-Fraenkel Axiomatic . . . 382
5.5.1.1 Cardinals . . . 388
5.5.1.2 Cartesian Product . . . 391
5.5.1.3 Intervals . . . 392
5.5.2 Set Operations . . . 393

5.5.2.1 Inclusion . . . 394
5.5.2.2 Intersection . . . 394
5.5.2.3 Union . . . 396
5.5.2.4 Difference . . . 397
5.5.2.5 Symmetric Difference . . . 398
5.5.2.6 Product . . . 399
5.5.2.7 Complementarity . . . 399
5.5.3 Functions and Applications . . . 403
5.5.3.1 Cantor-Bernstein Theorem . . . 410
5.5.4 Structures . . . 416
5.5.4.1 Magma . . . 418
5.5.4.2 Monoid . . . 419
5.5.4.3 Groups . . . 421
5.5.4.4 Ring . . . 426
5.5.4.5 Field . . . 429
5.5.4.6 Vector Spaces . . . 431
5.5.4.7 C-algebra A . . . 433
5.5.4.8 Summary . . . 435
5.5.5 Morphisms . . . 439
5.5.5.1 Ideal . . . 444
5.6 Probabilities . . . 450
5.6.1 Event Universe . . . 452
5.6.1.1 Infinite monkey theorem (Borel's law) . . . 455
5.6.1.2 Do probabilities refute complex structures? . . . 459
5.6.2 Kolmogorov's Axioms . . . 476
5.6.3 Conditional Probabilities . . . 485
5.6.3.1 Conditional Expectation and Variance . . . 499
5.6.3.2 Bayesian Networks . . . 504
5.6.4 Geometric Probabilities . . . 516
5.6.5 Martingales . . . 518
5.6.5.1 Gambling . . . 520
5.6.6 Combinatorial Analysis . . . 522
5.6.6.1 Simple Arrangements with Repetitions . . . 523
5.6.6.2 Simple Permutations without Repetitions . . . 525
5.6.6.3 Simple Permutations with Repetitions . . . 526
5.6.6.4 Simple Arrangements without Repetitions . . . 527
5.6.6.5 Simple Combinations without Repetitions . . . 528

5.6.6.6 Simple Combinations with Repetitions 530
5.6.7 Markov Chains 537

5.7 Statistics 542

5.7.1 Samples 545
5.7.2 Averages 546
5.7.2.1 Laplace Smoothing 570
5.7.2.2 Means and Averages properties 571
5.7.3 Type of variables 580
5.7.3.1 Discrete Variables and Moments 582
5.7.3.2 Continuous Variables and Moments 633
5.7.4 Fundamental postulate of statistics 643
5.7.5 Diversity Index 644
5.7.6 Distribution Functions (probabilities laws) 647
5.7.6.1 Discrete Uniform Distribution 649
5.7.6.2 Bernoulli Distribution 652
5.7.6.3 Geometric Distribution 654
5.7.6.4 Binomial Distribution 658
5.7.6.5 Negative Binomial Distribution 667
5.7.6.6 Hypergeometric Distribution 672
5.7.6.7 Multinomial Distribution 679
5.7.6.8 Poisson Distribution 686
5.7.6.9 Normal & Gauss-Laplace Distribution 692
5.7.6.10 Log-Normal Distribution 734
5.7.6.11 Continuous Uniform Distribution 738
5.7.6.12 Triangular Distribution 743
5.7.6.13 Pareto Distribution 746
5.7.6.14 Exponential Distribution 751
5.7.6.15 Cauchy Distribution 755
5.7.6.16 Beta Distribution 760
5.7.6.17 Gamma Distribution 767
5.7.6.18 Chi-Square (Pearson) Distribution 773
5.7.6.19 Student Distribution 783
5.7.6.20 Fisher Distribution 790
5.7.6.21 General Folded Normal Distribution 795
5.7.6.22 Laplace Distribution 800
5.7.6.23 Benford Distribution 803
5.7.6.24 ex-Gaussian distribution 806

5.7.7 Likelihood Estimators 818
5.7.7.1 Normal Distribution MLE 819
5.7.7.2 Poisson Distribution MLE 825
5.7.7.3 Binomial Distribution MLE 826
5.7.7.4 Geometric Distribution MLE 826
5.7.7.5 Weibull Distribution MLE 827
5.7.7.6 Gamma Distribution MLE 830
5.7.7.7 Pareto Distribution MLE 831
5.7.7.8 Censored data MLE 832
5.7.7.9 Fisher Information Matrix 835
5.7.8 Finite Population Correction Factor 843
5.7.9 Confidence Intervals (inference) 848
5.7.9.1 Frequentist Inference 848
5.7.9.2 Bayesian Inference 894
5.7.10 Weak Law of Large Numbers (Khinchin's law) 920
5.7.11 Characteristic Function 926
5.7.11.1 Moment generating function and cumulants 931
5.7.12 Central Limit Theorem 938
5.7.13 Univariate Hypothesis and Adequation Tests (NHST) 944
5.7.13.1 Direction of hypothesis tests and p-values 948
5.7.13.2 Fisher's method for multiple p-values 963
5.7.13.3 Power of a Test 972
5.7.13.4 A/B Tests 991
5.7.13.5 Analysis Of VAriance (ANOVA) 1001
5.7.13.6 Equivalence Tests 1101
5.7.13.7 Cochran C-test 1104
5.7.13.8 Family-wise error rate (adjusted p-values) 1106
5.7.13.9 Adequation Tests (Goodness of Fit tests) 1111
5.7.13.10 Kernel density estimation 1143
5.7.13.11 Likelihood-ratio Tests 1150
5.7.14 Robustness and Nonparametric Statistics 1153
5.7.14.1 Rank Statistics 1156
5.7.14.2 Range Statistics 1203
5.7.14.3 Extreme Value Theory 1211
5.7.15 Multivariate Statistics 1212
5.7.15.1 Principal Component Analysis 1213
5.7.15.2 Chi-square Correspondence Factor Analysis (C2CFA) 1233

5.7.15.3 Chi-squared Test of Independence 1238
5.7.15.4 Pearson's phi coefficient 1250
5.7.15.5 Exact Fisher's Test 1254
5.7.15.6 Cohen's kappa agreement 1260
5.7.15.7 McNemar's test 1264
5.7.16 Survival Statistics 1271
5.7.16.1 Kaplan-Meier Survival Rate 1273
5.7.16.2 Cochran–Mantel–Haenszel Tests 1283
5.7.16.3 Cox (Proportional Hazard) Model 1291
5.7.17 Propagation of Errors (experimental uncertainty analysis) 1307
5.7.17.1 Absolute and Relative Uncertainties (direct calculation of bias) 1308
5.7.17.2 Statistical Errors 1309
5.7.17.3 Repeatability 1312
5.7.17.4 Error propagation (linearised approximation) 1313
5.7.17.5 Error propagation (statistical approach) 1316
5.7.17.6 Significant Numbers 1318
5.7.18 Spatial statistics 1320
5.7.18.1 2D Poisson's spatial distance model 1320
5.7.19 Surveys 1324
5.7.19.1 Survey designs 1327
5.7.19.2 Survey Sampling Error and Non-Sampling Error 1335
5.7.20 A World without statistics 1339
5.7.20.1 Data fallacies 1342

5.1. Proof Theory

We have chosen to begin the study of Applied Mathematics by the theory that seems to us the most fundamental and important in the field of pure and exact sciences: Proof Theory and its underlying formal systems! The proof theory and theory of propositional calculus (logic) have five main objectives through this book:

1. Teach to the reader how to reason and demonstrate (prove), and this independently of the specialization field.
2. Show that the process of a demonstration (proof) is independent of the language used.
3. Prepare the reader to the Logic Theory (see section Logic Systems page 7104).
4. Prepare the path to Gödel's incompleteness theorem (main goal of this section!).
5. Prepare the reader to the Automata Theory (see section Automata Theory page 7181).

The two Gödel theorems are probably the most exciting points of Proof Theory because if we define religion as a system of thought that contains unprovable statements, then it contains elements of faith, and Gödel tells us that mathematics is not only a religion, but that it is the only religion that can prove itself that it is one!

Gödel's incompleteness theorem state that in any axiomatic system there are some statements that cannot be determined to be true or false (but some are!). The incompleteness theorem seems like something horrible, and for mathematics, it is! However, it does not apply to physics (or any other experimental science), because physics does not rely on logical proof in axiomatic systems, it relies on empirical evidence!

Remarks

R1. It is (very) strongly advised to read this section in parallel with those on Automata Theory (page 7181) and Logical Systems (page 7104), including Boolean Algebra (page 7105), available in Theoretical Computing (page 6267) chapter of this book.

R2. We must approach Proof Theory as a sympathetic curiosity but which basically brings nothing much except working/reasoning methods. Moreover, its purpose is not to show that everything is demonstrable but that any proof can be done on a common language starting from a finite number of rules.

As illustrated in the figure below, taken from the book [531], formal system are at the root of the two main targets of the present book that are General Relativity and Quantum Field Theory:

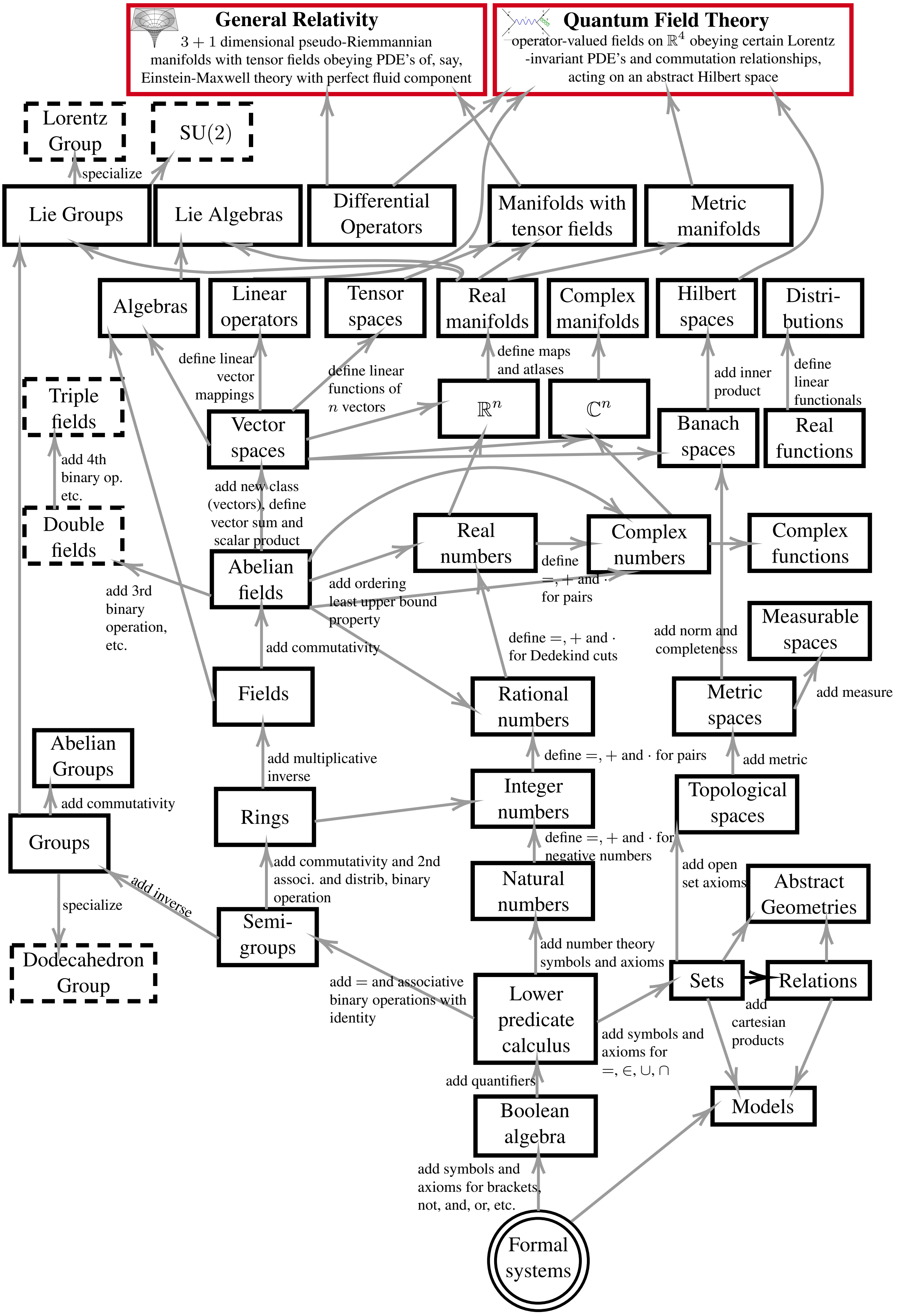

General Relativity
3 + 1 dimensional pseudo-Riemmannian manifolds with tensor fields obeying PDE's of, say, Einstein-Maxwell theory with perfect fluid component
Quantum Field Theory
operator-valued fields on $\mathbb{R}^4$ obeying certain Lorentz -invariant PDE's and commutation relationships, acting on an abstract Hilbert space
Lorentz Group
SU(2)
specialize
Lie Groups
Lie Algebras
Differential Operators
Manifolds with tensor fields
Metric manifolds
Algebras
Linear operators
Tensor spaces
Real manifolds
Complex manifolds
Hilbert spaces
Distri-butions
define linear vector mappings
define linear functions of n vectors
define maps and atlases
add inner product
define linear functionals
Triple fields
Vector spaces
$\mathbb{R}^n$
$\mathbb{C}^n$
Banach spaces
Real functions
add 4th binary op. etc.
add new class (vectors), define vector sum and scalar product
Double fields
Abelian fields
Real numbers
Complex numbers
Complex functions
define =, + and · for pairs
add 3rd binary operation, etc.
add ordering least upper bound property
add commutativity
define =, + and · for Dedekind cuts
add norm and completeness
Measurable spaces
add measure
Fields
Rational numbers
Metric spaces
Abelian Groups
add commutativity
add multiplicative inverse
define =, + and · for pairs
add metric
Groups
Rings
Integer numbers
Topological spaces
define =, + and · for negative numbers
add open set axioms
specialize
add inverse
add commutativity and 2nd associ. and distrib, binary operation
Natural numbers
Abstract Geometries
Semi-groups
Dodecahedron Group
add number theory symbols and axioms
Sets
Relations
add = and associative binary operations with identity
Lower predicate calculus
add symbols and axioms for =, ∈, ∪, ∩
add cartesian products
add quantifiers
Models
Boolean algebra
add symbols and axioms for brackets, not, and, or, etc.
Formal systems

Figure 5.1: **Oversimplified relations between various important mathematical structures for physics (author: Max Tegmark)**

Often when a student arrives in a graduate class he learned how to calculate or use algorithms but almost only a little or even not at all to reason (for the evolutionary origin of reasoning see [387]). For all the reasoning the visual media is a powerful tool (a picture is worth a thousand words) and people who do not see that in tracing a given curve or straight line the solution appears or who do not see in space are really penalized.

During high school we already manipulate unknown objects but especially to make calculations and when we reason about objects represented by letters, we can replace them visually by a real number, a vector, etc. At a given level we ask people to reason on more abstract structures and therefore to work on unknown objects which are elements of a set itself unknown, for example elements of any group (see section Set Theory page 421). This visual support thus doesn't exist anymore.

We ask so often to students to reason, to demonstrate the properties, but almost no one has ever taught them to reason properly, writing proofs, control proofs. If we ask a graduate student what is a proof, it most likely he will have some difficulty to answer. He can say that it is a text in which there are keywords like "therefore", "because", "if", "if and only if", "take a x such that", "assume", "lemma", "theorem", "let us look for a contradiction", etc. But he will probably be unable to provide the grammar of these texts nor their basics, and besides, its teachers, if they have not taken a course in Proof Theory, would probably be unable too.

To understand this situation, remember that to speak a child does not need to know the grammar. He imitates his surroundings and it works very well: most of time a six year old child know to use complicated sentences without ever having done grammar. Most teachers also do not know the grammar of reasoning but, for them, the imitation process has work well and thus they reason correctly. The experience of the majority of university teachers shows that this process of imitation works well for very good students, and then it is enough, but it works much less, if not at all, for many others.

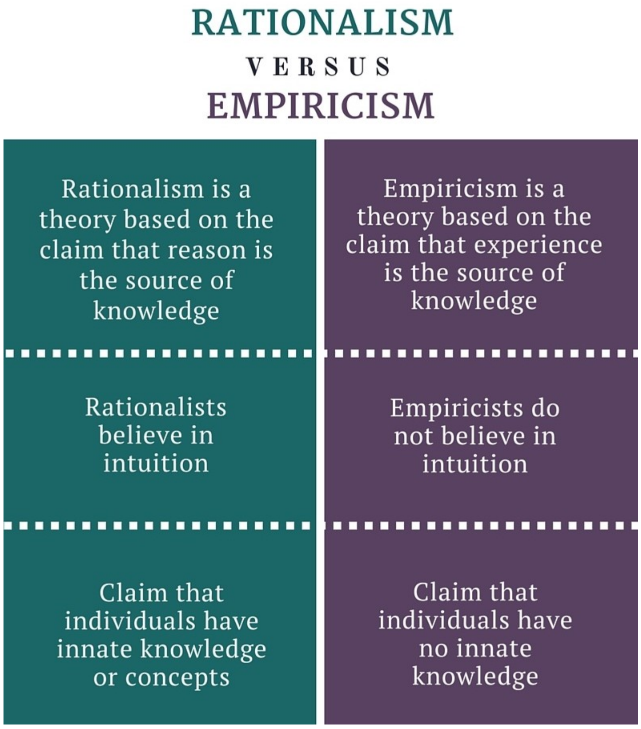
As the complexity level is low (especially during an "equational" type reasoning), grammar is almost useless but when the level increase or when we do not understand why something is wrong, it becomes necessary to do some grammar to progress. Teachers and students are familiar with the following situation: in a school assignment the corrector barred whole page of a large red line and write "false" in the margin. When the student asks what is wrong, the corrector can only say things like "this has no relation with the requested proof", "nothing is right", ..., which help obviously not the student to understand. This is partly because the text written by the student uses the appropriate words but in a more or less random way and can not give meaning to the assembly of these words. In addition, the teacher does not have the tools to explain what is wrong. We must therefore give them to him!

These tools exist but are fairly recent. The proof theory is a branch of mathematical logic whose origin is the crisis of the foundations: there was a doubt about what we had the "right" to do in a mathematical reasoning (see the "foundations crisis" further below). Paradoxes appeared, and it was then necessary to clarify the rules of proof and to verify that these rules are not contradictory. This theory appeared in the early 120th century (holocene calendar), which is very new since most of the mathematics taught in the first half of the university is known since the 116th-117th century (holocene calendar).

5.1.1 Foundations Crisis

For the Greeks philosophers geometry was considered the highest form of knowledge, a powerful key to the metaphysical mysteries of the universe. It was rather a mystical belief and the link between mysticism and religion was made explicit in cults like those of the Pythagoreans.

No culture has since challenged a man for having discovered a geometrical theorem! Later, mathematics was regarded as the model of a priori knowledge in the Aristotelian tradition of rationalism.



RATIONALISM
VERSUS
EMPIRICISM

Rationalism is a theory based on the claim that reason is the source of knowledge	Empiricism is a theory based on the claim that experience is the source of knowledge
Rationalists believe in intuition	Empiricists do not believe in intuition
Claim that individuals have innate knowledge or concepts	Claim that individuals have no innate knowledge

The astonishment of the Greeks philosophers for mathematics has not left us, we find it in the traditional metaphor of mathematics as "Queen of Science". It was strengthened by the spectacular success of mathematical models in science, success that the Greeks (even ignoring the simple algebra) had not anticipated. Since the discovery by Isaac Newton's of integral calculus and the inverse square law of gravity in the late 11600s (holocene calendar) the phenomenal sciences and higher mathematics remained in close symbiosis - to the point that a predictive mathematical formalism was became the hallmark of a "hard science".

After Newton, during the next two centuries, science aspired to that kind of rigour and purity that seemed inherent in mathematics. The metaphysical question seemed simple: mathematics seemed to have a perfect a priori knowledge, and among all sciences, those that were able to mathematize most perfectly were the most effective for predicting phenomena. The perfect knowledge therefore, was in a mathematical formalism that, once reached by science and embracing all aspects of reality, could found a posteriori empirical knowledge on an a priori rational logic. It was in this spirit that Marie Jean-Antoine Nicolas de Caritat, Marquis de Condorcet (French philosopher and mathematician), undertook to imagine describing the entire Universe as a set of partial differential equations being solved one after the other.

The first break in this inspiring picture appeared in the second half of the 119th century (holocene calendar), when Riemann and Lobachevsky separately proved that Euclid's parallel axiom (see the definition page 100) could be replaced by other geometries that produced "consistent" (we will come back more on this word further below). Riemannian geometry was modelled on a sphere, these of Lobatschewsky, on rotation of a hyperboloid.

The impact of this discovery was later obscured by great upheaval, but at the time it made a thunderclap in the intellectual world. The existence of mutually inconsistent axiomatic systems, each of which could be a model for the phenomenal Universe, relied entirely into question the relation between mathematics and theoretical physics.

When we knew only Euclid, there was only one possible geometry. One could believe that the Euclid's postulates (see section Euclidean Geometry page 2563) were a kind of knowledge a priori perfect on the geometry in the phenomenal world. But suddenly we had three geometries, embarrassing for metaphysical subtleties.

Why would we choose between the axioms of plane geometry, spherical and hyperbolic geometry as real descriptions? Because all three are consistent, we can not choose any a priori as a foundation - the choice must be empirical, based on their predictive power in a given situation.

Of course, the theoretical physicists have long been accustomed to choose a formalism to study a scientific problem. But it was already accepted widely, if not unconsciously, that the need to do so was based on human ignorance, and with logic or good enough mathematics, one could infer the right choice from principles first, and produce a priori descriptions of reality that had to be confirmed afterwards by empirical verification.

However, Euclidean geometry, seen for hundreds of years as the model of axiomatic perfection of mathematics, had been dethroned. If we could not know a priori something as basic as the geometry in space, what hope was there for a pure rational theory that would encompass all of nature? Psychologically, Riemann and Lobachevsky had struck at the heart the mathematical enterprise as it had been designed before.

Moreover, Riemann and Lobachevsky have pushed the nature of mathematical intuition into question. It was easy to believe implicitly that mathematical intuition was a form of perception - a way to glimpse the Platonic world behind reality. But with two other geometries pushing the Euclid one in it's limit, no one could never be sure to know what the world really looks like.

Mathematicians responded to this dual problem with excessive rigour, trying to apply the axiomatic method in all mathematics. In the pre-axiomatic period, the proofs were often based on commonly accepted intuitions of the "reality" of mathematics, which could not automatically be regarded as valid.

The new way of thinking about mathematics led to a series of spectacular success. Yet this had also a price. The axiomatic method made the connection between mathematics and the phenomenal reality increasingly close. Meanwhile, discoveries suggested that mathematical axioms that appeared to be consistent with phenomenal experience could lead to dizzying contradictions with this experience.

Most mathematicians quickly became "formalist" arguing that pure mathematics could only be regarded as a kind of elaborate philosophy game that was played with symbols on paper (that's the theory that is behind the mathematical prophetic qualification "zero content system" by Robert Heinlein). The "Platonic" belief in the reality of mathematical objects, in the old-fashioned way, seemed good 20th the trash, despite the fact that mathematicians still feel like platonicians during the process of discovery of mathematics.

Philosophically, then, the axiomatic method led most mathematicians to abandon previous beliefs in the metaphysical specificity of mathematics. It also produced the contemporary rupture between pure and Applied Mathematics. Most of the great mathematicians of the early modern period - Newton, Leibniz, Fourier, Gauss and others - also occupied phenomenal science. The axiomatic method had hatched the modern idea of the pure mathematician as a great aesthete, heedless of physics. Ironically, formalism gave the pure mathematicians a bad addiction to the Platonic attitude. The researchers in Applied Mathematics ceased to meet physicists and learned to put themselves in their behind.

This takes us to the early 120th century (holocene calendar). For the beleaguered minority of Platonists, the worst was yet to come. Cantor, Frege, Russell and Whitehead showed that all pure mathematics could be built on the simple foundation of the Set Theory axiomatic. This suited well the formalists: the mathematics were reunifying, at least in principle, from a small set of rules detached of a big one. Platonicians also were satisfied, if a great structure appeared, consistent keystone for the whole mathematics, the metaphysical specificity of mathematics could still be saved.

In a negative way, though, a Platonist had the last word. Kurt Gödel put his grain of sand in the program of axiomatization formalism when he proved that any sufficiently powerful axiom system to include integers numbers had to be either inconsistent (contain contradictions) or incomplete (too weak to decide the rightness or the falsity of some statements of the system). And that's more or less where things stand today. Mathematicians know that many attempts to advance mathematics as a priori knowledge of the Universe must face numerous paradoxes and unable to decide which axiom system describes the real mathematics. They have been reduced to hope that standards axiomatizations are not inconsistent but just incomplete, and wondering anxiously what contradictions or unprovable theorems are waiting to be discovered elsewhere.

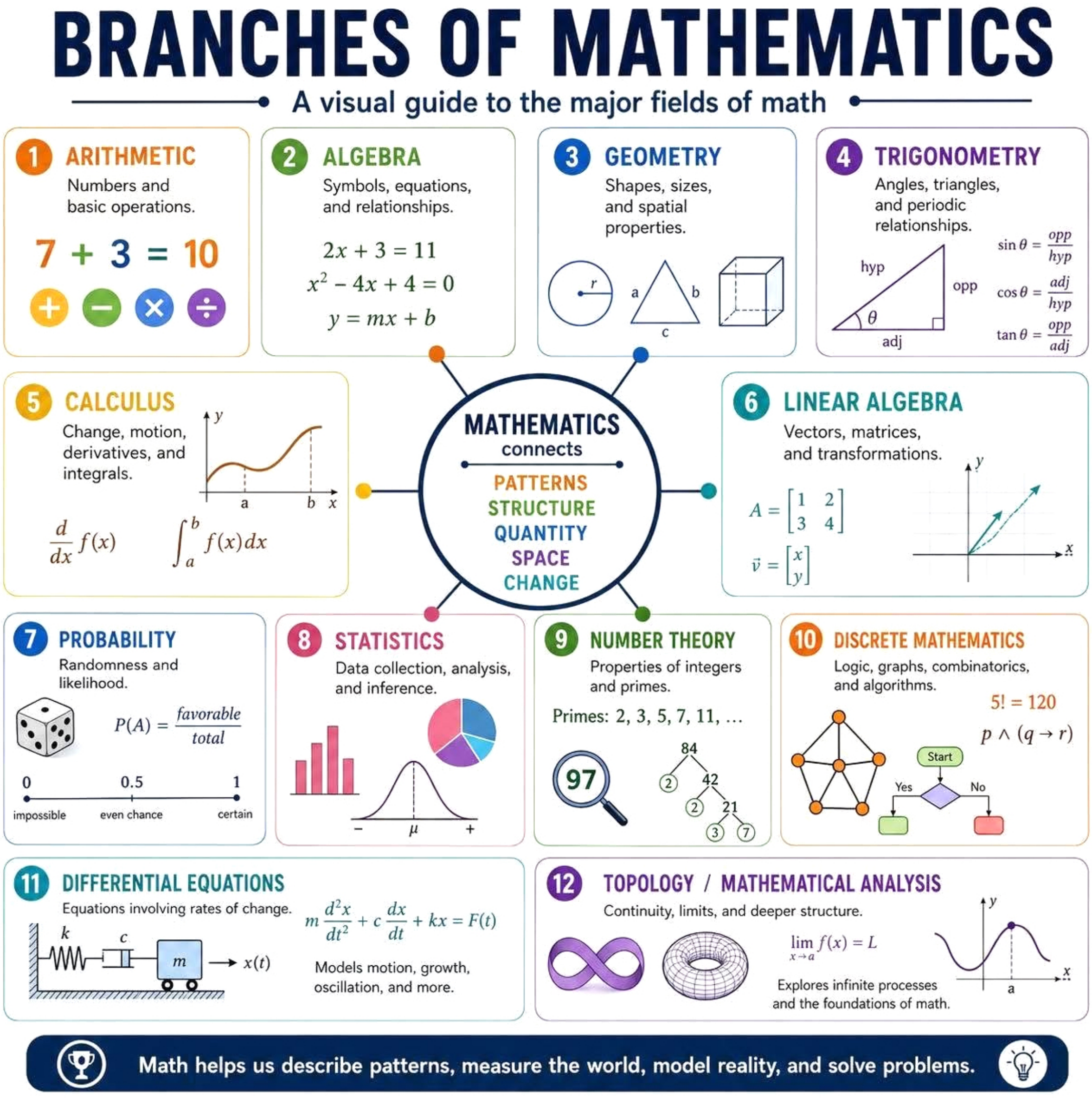
However, on the front of empiricism, mathematics was always a spectacular success as a theoretical construction tool. The great success of physics in the 120th century (General Relativity and Quantum physics) pushed so far out of the realm of physical intuition, they could only be understood by meditating deeply on their mathematical formalism, and extending their logical conclusions, even when those findings seemed wildly bizarre. What irony! Just as the mathematical perception were to appear always less reliable in pure mathematics, it became more and more indispensable in phenomenal science.

In contrast to this background, the applicability of mathematics to phenomenal science poses a more difficult problem than at first appears. The relation between the mathematical models and prediction of phenomena is complex, not only in practice but also in principle. Even more complex, as we now know, there are ways to axiomatize mathematics that mutually exclude themselves!

But why is there only one good choice of mathematical model? That is, why is there a mathematical formalism, for example for quantum physics, so productive that it predicts the discovery of new observable particles?

To answer this question we will can observe that, as well, works as a kind of definition. For many phenomenal systems, such exact predictive formalism has not been found, and none seem plausible. You can easily find such examples: climate or the behaviour of a superior economy to that of a town - systems so chaotically interdependent that exact prediction is actually impossible (not only in practice but in principle).

> *One should study mathematics simply because it helps to arrange one's ideas!*


BRANCHES OF MATHEMATICS
A visual guide to the major fields of math
1 ARITHMETIC
Numbers and basic operations.
7 + 3 = 10
2 ALGEBRA
Symbols, equations, and relationships.
$2x + 3 = 11$
$x^2 - 4x + 4 = 0$
$y = mx + b$
3 GEOMETRY
Shapes, sizes, and spatial properties.
r
a
b
c
4 TRIGONOMETRY
Angles, triangles, and periodic relationships.
hyp
opp
adj
θ
$\sin\theta = \frac{opp}{hyp}$
$\cos\theta = \frac{adj}{hyp}$
$\tan\theta = \frac{opp}{adj}$
5 CALCULUS
Change, motion, derivatives, and integrals.
y
a
b
x
$\frac{d}{dx} f(x)$
$\int_a^b f(x)dx$
MATHEMATICS
connects
PATTERNS
STRUCTURE
QUANTITY
SPACE
CHANGE
6 LINEAR ALGEBRA
Vectors, matrices, and transformations.
$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$
$\vec{v} = \begin{bmatrix} x \\ y \end{bmatrix}$
y
x
7 PROBABILITY
Randomness and likelihood.
$P(A) = \frac{favorable}{total}$
0
0.5
1
impossible
even chance
certain
8 STATISTICS
Data collection, analysis, and inference.
−
μ
+
9 NUMBER THEORY
Properties of integers and primes.
Primes: 2, 3, 5, 7, 11, …
97
84
2
42
2
21
3
7
10 DISCRETE MATHEMATICS
Logic, graphs, combinatorics, and algorithms.
5! = 120
$p \wedge (q \rightarrow r)$
Start
Yes
No
11 DIFFERENTIAL EQUATIONS
Equations involving rates of change.
k
c
m
$x(t)$
$m\frac{d^2x}{dt^2} + c\frac{dx}{dt} + kx = F(t)$
Models motion, growth, oscillation, and more.
12 TOPOLOGY / MATHEMATICAL ANALYSIS
Continuity, limits, and deeper structure.
$\lim_{x \to a} f(x) = L$
Explores infinite processes and the foundations of math.
y
x
a
Math helps us describe patterns, measure the world, model reality, and solve problems.

5.1.2 Paradoxes

Since ancient times, some logicians had noticed the presence of many paradoxes within rationality. In fact, we can say that despite their number, these paradoxes are merely illustrations of a few paradoxical structures. Let us look to for general culture to the most famous which constitute the class of "undecidable propositions".

Let us see one of the most well known example: The paradox of the class of classes (Russell)!

✍**Example:**

There are two types of classes: those that contain themselves (or reflexive classes: the class of non-empty sets, the class of classes, ...) and those who do not contains themselves (or non-reflexives classes: the class of work to be returned, the class of blood oranges, ...). The question is the following: is the class of non-reflexives classes itself reflexive or non-reflexive? If it is reflexive, it contains itself and is thus in the class of non-reflexives classes that it represents, which is contradictory. If it is non-reflexive, it must be included in the class of non-reflexives classes and becomes ipso facto reflexive, we are facing again a contradiction.

This Russell's paradox is often known mainly under the two following variants:

- Does the set of all sets that do not contain themselves contain himself?

  The answer is: If "Yes", then "No" and if "No" then "Yes"...

  Remark

  A religious funny version of this paradox is the following: Can the Almighty (whatever the deity of one of the 45 monotheistic religions) create an entity more powerful than Him? No! Then he is not all-powerful. Yes, then he is not all-powerful.

- Those who do not shave themselves are shaved by the barber but not those who shave themselves. So who shaves the barber?

  The answer is: If the barber shave himself he enters in the category of people that shave themselves so he does not shave himself because he is the barber... But if he does not shave himself he enters in the category of people that are shaved by the barber... The answer is also undecidable...

The Universe is a reflexive class (it contains itself). Therefore the Russell paradox doesn't apply to it by construction!

Russell's paradox challenges the notion of a set as a collection defined by common ownership! In one shot it destroys the logic (undecidable proposition) and set theory... because the overall concept of all sets is an impossibility!!! The self-reference is the center of this logical problem!

Formally, Russell's paradox is stated and proved in just a few lines:

$$\boxed{\begin{gathered}S=\{x:x\notin x\}\\ S\in S\Longleftrightarrow S\notin S\end{gathered}} \tag{5.1.1}$$

This paradox also returns to the question whether a math question correctly formulated (logical) necessarily admits an answer? Said in another way: is any mathematical statement provable... and it is Gödel that many years after the statement of Russell's paradox proved mathematically that the answer is No !!!!!! In other words, there will always be questions unanswered because any system (living language or mathematical tool) based itself is necessarily incomplete! This is the famous impact of Gödel's incompleteness theorem that is technically written as following (and that we will have to prove):

$$\boxed{\exists P:\quad \neg\mathrm{Dem}(P)\wedge\neg\mathrm{Dem}(\neg P)} \tag{5.1.2}$$

that means that given a proposition P, whatever the set of axioms, there exist proposition that we will be never able to prove as true or false.

Let us see another application of the Russell's paradox:

✍**Example:**

In a library, there are two types of catalogues: Those who mention themselves and those who does not mention themselves. A librarian must draw up a catalogue of all catalogues that do not mention themselves. Having completed its work, our librarian asks whether or not to mention the catalogue that is precisely drafting. At this point, he is struck perplexity. If he does not mention this catalogue it will be a catalogue that is not mentioned and which should therefore be included in the list of catalogues that does not mention themselves. On the other hand, if he mentions the catalogue, this catalogue will become a catalogue that is mentioned and must therefore not be included in this catalogue, since it is the catalogue of catalogues which does not mention themselves.

A variations of the previous paradox is the well-known liar paradox:

✍**Example:**

Let us provisionally define lying as the work of making a false proposition. The Cretan poet Epimenides said: "All Cretans are liars", this is the proposal P. How to decide the truthfulness of P? If P is true, as Epimenides is Cretan, P must be false. P must therefore be false to be true, which is contradictory.

As would have made understand the logician Ludwig Wittgenstein, these paradoxes ultimately show that mathematics is a pretty good tool to show the logic but not to talk about it. Give with mathematics an independent existence to this algebraic entities is madness and it is this that produces monsters like the set of all the sets... The logic is empty and can not tell the reality, it restrict to be just a picture of it.

The reader may be always careful to distinguish paradoxes, sophisms and ill-defined (ill-specified) problems or popular myths (provided without evidence for that latter obviously...). That's one of the biggest source of misunderstanding in many debates.

✍**Example:**

If you choose an answer to this question at random, what is the chance you will be correct?

A) 25%

B) 50%

C) 60%

D) 25%

This is an ill-defined question. There are no solution and lead therefore to a paradox.

5.1.2.1 Hypothetical-Deductive Reasoning

The hypothetical-deductive reasoning is, we know (see the Introduction of the book), the ability of the learner to deduce conclusions from pure hypotheses and not only of a real observation. It is a thought process that seeks to identify a causal explanation of any phenomenon (we will come back on this during our first steps in physics). The learner who uses this type of reasoning begins with a hypothesis and then tries to prove or disprove his hypothesis following the block diagram below:

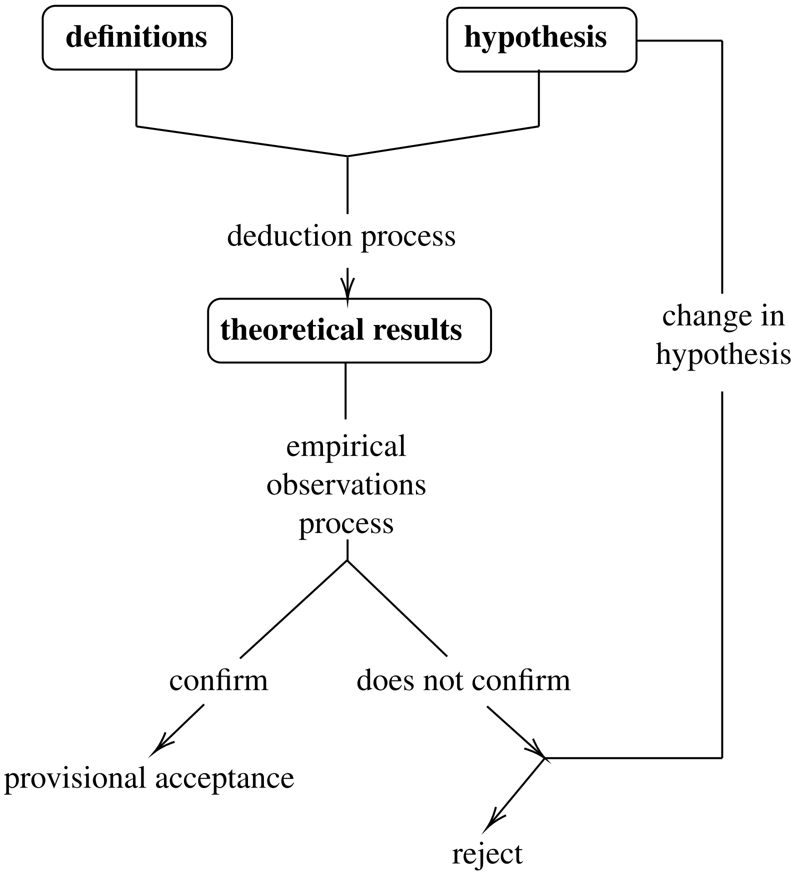

definitions
hypothesis
deduction process
theoretical results
change in hypothesis
empirical observations process
confirm
does not confirm
provisional acceptance
reject

Figure 5.2: **Hypothetical-Deductive Reasoning block diagram**

The deductive procedure is to hold as true, provisionally, this first proposal that we name, in logic a "predicate" (see further below for more details) and to draw all the consequences logically necessary, that is to say to look for its implications.

✍**Example:**

Consider the proposition (antecedent) P: "X is a man", it implies the following proposition (consequent) Q: "X is mortal".

The expression $P \Rightarrow Q$ (if it is a human it is necessarily mortal) is a predicative implication (hence the term "predicate"). There is no case in this example where we can state P without Q. This example is that of strict implication, as we find in the "syllogism" (logical reasoning figure).

In logic, a "syllogism" is a logical reasoning relating at least three propositions: two (named commonly "minor premise", "major premise") or more of them, named "premises", lead to a "conclusion".

A very well-known example of syllogism is: «*All men are mortal, but Socrates is a man; therefore Socrates is mortal*»: the two premises (named "major premise" and "minor premise") are propositions given and assumed to be true, the syllogism makes it possible to establish the formal validity of the conclusion, which is necessarily true if the premises are true.

Remark

Experts have shown that the hypothetical-deductive reasoning develops gradually by children from six to seven years old and that this kind of reasoning is used systematically starting with a strict propositional function until the age of eleven-twelve.

5.1.3 Propositional Calculus

First let us define what is "logic" in its contemporary sense otherwise using old definition we can prove that unicorns exist (as we will prove it further below!).

Definition (#5.1.7): "Logic" is the scientific study, especially formal, of the standards of truth (i.e. of correct reasoning) in a given axiomatic framework. "Formal logic", that is a subfamily, is the science of deductively valid inferences or logical truths. It studies how conclusions follow from premises due to the structure of arguments alone, independent of their topic and content. "Informal logic", that is another subfamily, is associated with informal fallacies, critical thinking, and argumentation theory.

The "propositional calculus" (or "propositional logic") is an absolutely indispensable preliminary to tackle a background in science, philosophy, law, politics, economics, etc. This type of calculation allows for decisions or testing procedures. These help to determine when a logical expression (proposition) is true and especially if it is always true.

Definitions (#5.1.8):

D1. An expression that is always true whatever the content language of the variables that compose it is named a "valid expression", a "tautology" or a "law of propositional logic".

In other words a tautology is a statement that is always true because it includes all logical possibilities.

D2. An expression that is always false is named a "contradiction" or "antilogy".

D3. An expression that is sometimes true, sometimes false is named a "contingent expression" (this is often linked to the analogy that something can exist but does not necessarily have to exist.).

D4. We name "assertion" an expression that we can say unambiguously whether it is true or false.

D5. The "object language" is the language used to write logical expressions.

D6. The "meta-language" is the language used to talk about the object language in everyday language.

Remarks

R1. There are expressions that are actually not assertions. For example, the statement "this statement is false" is a paradox that can be neither true nor false.

R2. Consider a logical expression A. If it is a tautology, we denote it by $\models A$ and by $A \models$ or $\not\models A$ if it is a contradiction.

R3. In mathematics we can try to prove in a general way that an assertion is true, but not that it is false (if this is the case we give just one example).

5.1.3.1 Propositions (premises)

Definition (#5.1.9): In logic, a "proposition" is a statement that has meaning. That means we can say unambiguously whether this statement is true (T) or false (F). This is what we name the "law of excluded middle".

✍**Examples:**

E1. "I lie" is not a proposition (premise). If we assume that this statement is true, it is an affirmation of his own disability, so we should conclude that it is false. But if we assume that it is false, then the author of this statement does not lie, so he told the truth, thus the proposal would be true...

E2. Another funny example is:

- Everything has a creator
- God is that creator
- God does not have creator

It's a solution that fails since it violates its own premise...

Definition (#5.1.10): A proposition in binary logic (where the proposals are either true or false) is therefore never true and false at the same time. This is what we name the "principle of non-contradiction".

Thus, a property on the set of propositions E is an application P from E to the set of "truth values True, False" $\{T,F\}$:

$$P : E \mapsto \{T,F\} \tag{5.1.3}$$

We speak about "associated subset" , when the proposition only generates a portion E' of E and vice versa.

✍**Example:**

In $E = \mathbb{N}$, if $P(x)$ states "x is even", then $P = \{0,2,4,\ldots,2k,\ldots\}$ which is indeed only an associated subset of E but of same cardinal (see section Set Theory page 388).

Definition (#5.1.11): Let P be a property of the set E. A property Q on E is a "negation" of P if and only if, for any $x \in E$:

- $Q(x)$ is F (false) if $P(x)$ is T (true)
- $Q(x)$ is T (true) if $P(x)$ is F (false)

We can gather these conditions in a table named "truth table":

P	Q
T	F
F	T

Table 5.1: **Truth table of values**

Table that we can also find or also write in the most explicit following form:

P	Q
True	False
False	True

Table 5.2: **Truth table of explicit values**

or in binary form:

P	Q
1	0
0	1

Table 5.3: **Truth table of binary values**

In other words, P and Q always have opposite truth values. We denote this kind of statement "Q is a negation of P":

$$Q \Leftrightarrow \neg P \tag{5.1.4}$$

where the symbol $\neg$ is the "negation connector".

Remark

The expressions must be well-formed expressions (often abbreviated "WFE"). By definition, any variable is a well-formed expression, thus $\neg P$ is a well-formed expression. If P, Q are well-formed formulas, then $P \Rightarrow Q$ is a well-formed expression (the expression "I am lying" is not well-formed because it contradicts itself).

5.1.3.2 Connectors

There are other types of logical connectors:

Definition (#5.1.12): Let P and Q two properties set defined on the same set E. $P \vee Q$ (read "P `OR` Q") is a statement (property) on E defined by:

- $P \vee Q$ is true if at least P or Q are true
- $P \vee Q$ is false otherwise

We can write the truth table of the "`OR` connector" or "disjunction connector" $\vee$:

P	Q	$P \vee Q$
T	T	T
T	F	T
F	T	T
F	F	F

Table 5.4: `OR` inclusive truth table

The reader must me careful not to confound the Exclusive `OR` that means "*either this or that, but not both*" with the Inclusive `OR` that means "*either this, or that, or both*". The truth table of that first one, denoted most of times with the symbol $\oplus$ (and very rarely denoted with the symbol $\veebar$), is obviously a bit different:

P	Q	$P \oplus Q$
T	T	F
T	F	T
F	T	T
F	F	F

Table 5.5: `OR` exclusive (XOR) truth table

Every natural number is either even or odd, but not both. So every natural number is even xor odd (that's short for exclusive or).

It should be easy to convince yourself that if the parts P, Q of E are respectively associated with the properties P, Q thus $P \cup Q$ (see section Set Theory page 383) is associated to $P \vee Q$:

$$\begin{aligned} P &\leftrightarrow P \\ Q &\leftrightarrow Q \\ P \vee Q &\leftrightarrow P \cup Q \end{aligned} \tag{5.1.5}$$

The connector $\vee$ is associative (no doubt about the fact that it is commutative!). For proof, just do a truth table where you can check that:

$$[P \vee (Q \vee R)] = [(P \vee Q) \vee R] \tag{5.1.6}$$

Definition (#5.1.13): There is also the "`AND` connector" or also named "conjunction connector" $\wedge$ for whatever are P, Q two properties defined on E, $P \wedge Q$ is a property defined on E by:

- $P \wedge Q$ is true if both properties P, Q are true (the famous syllogism: *All men are mortal, Socrates is a man, therefore Socrates is mortal* is a famous example).
- $P \wedge Q$ if false otherwise

We can write the truth table of the "`AND` connector" or "conjunction connector" $\vee$:

P	Q	$P \wedge Q$
T	T	T
T	F	F
F	T	F
F	F	F

Table 5.6: `AND` **truth table**

It should be also almost easy to convince yourself that if the parts P, Q of E are respectively associated with the properties P, Q thus $P \cap Q$ (see section Set Theory page 394) is associated to $P \wedge Q$:

$$\begin{gathered} P \leftrightarrow P \\ Q \leftrightarrow Q \\ P \wedge Q \leftrightarrow P \cap Q \end{gathered} \tag{5.1.7}$$

The connector $\wedge$ is associative (no doubt about the fact that it is commutative!). For proof, just do a truth table where you can check that:

$$[P \wedge (Q \wedge R)] = [(P \wedge Q) \wedge R] \tag{5.1.8}$$

The connectors $\vee, \wedge$ are distributive one on the other. Using a simple truth table, we can show that (ask me if you want a put the truth table):

$$[P \vee (Q \wedge R)] = [(P \vee Q) \wedge (P \vee R)] \tag{5.1.9}$$

as well as:

$$[P \wedge (Q \vee R)] = [(P \wedge Q) \vee (P \wedge R)] \tag{5.1.10}$$

Definition (#5.1.14): The "negation" operator $\neg$ transforms a True value into a False value such that:

$$\begin{gathered} \neg T = F \\ \neg F = T \end{gathered} \tag{5.1.11}$$

So in logic, negation, also named "logical complement", is an operation that takes a proposition P to another proposition "not P", denoted $\neg P$ or sometimes $\bar{P}$, which is interpreted intuitively as being True when P is false and False when P is True. Negation is thus a unary (single-argument) logical connective.

As we will prove it in detail in the section of Logic System page 7110 (using a simple truth table), the "De Morgan's laws" provide a way of distributing negation over disjunction and conjunction:

$$\begin{gathered} \neg(P \wedge Q) \Leftrightarrow [(\neg P) \vee (\neg Q)] \\ \neg(P \vee Q) \Leftrightarrow [(\neg P) \wedge (\neg Q)] \end{gathered} \tag{5.1.12}$$

Remark

To see the details of all logical operators, the reader should read the section of Logical Systems page 7104 where the identity, the double negative, the idempotence, associativity, the distributive properties, the De Morgan relations are presented more formally and with full details.

Let us now come back on the "logical implication connector" sometimes also named just the "conditional" denoted by the symbol $\Rightarrow$.

Remark

In some books on propositional calculus, this connector is denoted by the symbol $\supset$ and as part of the proof theory we often prefer the symbol $\rightarrow$.

Let P, Q two properties given on E. $P \Rightarrow Q$ is a property on E defined by:

P1. $P \Rightarrow Q$ is False if P is True and Q is False.

P2. $P \Rightarrow Q$ is True otherwise.

In other words, the fact that P logically implies Q means that Q is True for any assessment for which P is True. The implication is therefore the famous "if... then ...".

If we write the truth table of the implication (caution with the before last line!!!):

P	Q	$P \Rightarrow Q$
T	T	T
T	F	F
F	T	T
F	F	T

Table 5.7: **Implication truth table**

In other terms, a False proposition implies that any conclusion will always be True. This is named a "informal fallacy[1]" and it is a very important concept in null hypothesis statistical tests (see section Statistics page 948). If the proposition is True the implication can be True only if the result is True.

---

[1] A "paralogism" is a false argument that appears to be valid, especially to its author, who is honest, unlike a "sophism" which is a fallacious argument (a "fallacy") intended to deceive.

A given set of logical connectives is "truth functionally complete" if and only if all possible connectives can be expressed through suitable combinations of the ones in the given set.

It turns out the set of five logical connectives introduced in this chapter is truth functionally complete. In fact, the subset containing only the three connectives AND, OR, and NOT is also truth functionally complete. This means that you don't need the other two connectives-implication and equivalence-because you can express them in terms of NOT, AND, and OR:

$$\begin{aligned}(P \Rightarrow Q) &\text{ can be expressed as } (\neg P) \vee Q\\ (P \Leftrightarrow Q) &\text{ can be expressed as } (P \wedge Q) \vee (\neg P \wedge \neg Q)\end{aligned} \tag{5.1.13}$$

Although the set of connectives AND, OR, NOT is indeed both sufficient and convenient to express all possible compound predicates, you can achieve the same goal with even a single logical connective: the NAND operator (not and).

Remark

Strictly speaking, the above table should use the symbol $\rightarrow$, and the symbol $\Rightarrow$ should be reserved only for implications that are true.

Example:

Consider the proposition: "If you get your diploma, I buy you a computer".

Of all cases, only one corresponds to a broken promise: the one where the child graduates, and still has no computer (second line in the table above).

What means exactly this promise, that we will write as following:

$$\text{"You have your degree"} \Rightarrow \text{"I buy you a computer"?}$$

Exactly this:

- If you have get graduate, for sure, I will buy you a computer (I can not not buy it).
- If you do not get graduate, I said nothing.

The implication gives us that from any false proposition we can deduce any proposal (last two lines)!!!

Remark

Informal fallacy is very interesting to apply in the case of debates on social networks. Indeed, if someone ask you to prove your premise (null hypothesis) that he assumes to be false, then he didn't understand anything to logic as he should know that from any false premise (null hypothesis) you can always have a true conclusion (you don't reject the null hypothesis) whatever the reasoning! What the guy must do is assume your null hypothesis as true and found at least one example making it false (we reject the null hypothesis in favour of the alternative hypothesis). This is how science works and also the fact that this is not obvious explains why so many people loose their time to argue in social networks using biased arguments.

✍**Example:**

In a course taught by Russell on the subject from a false proposition, any proposal can be inferred, a student asked him the following question (anecdote or legend?):

- "Are you saying that from the proposition $2+2=5$, it follows that you are the Pope?".
- "Yes", answered Russell.
- "And could you prove it!", asked the student sceptical...
- "Certainly", answered Russell, who immediately offered the following proof:
  1. Suppose that $2+2=5$.
  2. Subtract 3 from each member of the equality, we thus get $1=2$.
  3. By symmetry $2=1$.
  4. The Pope and I are 2. Since $2=1$, Pope and I are 1. It follow I'm the Pope.

The implication connector is essential in mathematics, philosophy, etc. It is a backbone of any proof, evidence or deduction. It has the following useful properties (normally easy to check with a small truth table):

$$\begin{aligned} P \Rightarrow Q \Leftrightarrow [(\neg Q) \Rightarrow (\neg P)] \\ P \Rightarrow Q \Leftrightarrow [(\neg P) \vee Q] \end{aligned} \tag{5.1.14}$$

And we have from the last property (again verifiable by a truth table):

$$\neg(P \Rightarrow Q) \Leftrightarrow [P \wedge (\neg Q)] \tag{5.1.15}$$

The "logical equivalence connector" or "biconditional connector" denoted most of times by "$\Leftrightarrow$" or sometimes by "$\leftrightarrow$" meaning by definition:

$$(P \Leftrightarrow Q) \Leftrightarrow [(P \Rightarrow Q) \wedge (Q \Rightarrow P)] \tag{5.1.16}$$

in other words, the first expression has the same value for all evaluation of the second. It is the same with the following relation that is more "atomic" as the logical equivalence is reduced only to the use of $\wedge, \vee$ and negation $\neg$ (combination of what we have seen above):

$$(P \Leftrightarrow Q) \Leftrightarrow [(P \wedge Q) \vee (\neg P \wedge \neg Q)] \tag{5.1.17}$$

When we prove such equivalence of two expressions we can therefore say that: "we proves that the equivalence is a tautology".

A "necessary and sufficient condition" requires that both of the implications $P \Rightarrow Q$ and $Q \Rightarrow P$ (the latter of which can also be written as $P \Leftarrow Q$) hold. The first implication suggests that P is a sufficient condition for Q, while the second implication suggests that Q is a necessary condition for P. This is expressed as "Q is necessary and sufficient for P", "P if and only if Q", or $Q \Leftrightarrow P$.

The truth table of the equivalence is logically given by:

P	Q	$P \Rightarrow Q$	$Q \Rightarrow P$	$Q \Leftrightarrow P$
T	T	T	T	T
T	F	F	T	F
F	T	T	F	F
F	F	T	T	T

Table 5.8: **Truth table of the equivalence**

$P \Leftrightarrow Q$ means (when its true!) that "P and Q always have the same truth value" or "P and Q are equivalent." This is True if P and Q have the same value, False otherwise.

Of course (it is a tautology):

$$\neg(P \Leftrightarrow Q) \Rightarrow (P \Rightarrow \neg Q) \qquad (5.1.18)$$

The relation $P \Leftrightarrow Q$ is equivalent to that P is a necessary and sufficient condition for Q and that Q is a necessary and sufficient condition for P.

Remark

Strictly speaking, the above table should use the symbol $\leftrightarrow$, and the symbol $\Leftrightarrow$ should be reserved only for equivalences that are true.

The conclusion is that the conditions of the types: "necessary", "sufficient", "necessary and sufficient" can be reformulated with the terms "only if", "if", "if and only if".

Therefore:

1. $P \Rightarrow Q$ reflects the fact that Q is a necessary condition for P or in other words, P is True only if Q is True (in the truth table, when $P \Rightarrow Q$ is equal to T we see that P is T only if Q is also T). The assertion that Q is necessary for P is colloquially equivalent to "P cannot be true unless Q is true" or "if Q is false, then P is false".

2. $P \Leftarrow Q$ or what is the same $Q \Rightarrow P$ reflects the fact that Q is a sufficient condition for P or in other words, P is True if Q is True (in the truth table, when $Q \Rightarrow P$ takes the value T we see that P is T if Q is T too). If Q is sufficient for P, then knowing Q to be true is adequate grounds to conclude that P is true; however, knowing Q to be false does not meet a minimal need to conclude that P is false.

3. $P \Leftrightarrow Q$ reflects the fact that Q is a necessary and sufficient condition for P or in other words, P is True if and only if Q is True (in the truth table, when $P \Leftrightarrow Q$ takes the value T we see that P is 1 if Q is T and if and only if Q is equal to T).

Remark

The expression "if and only if" therefore corresponds to a logical equivalence and can only be used to describe a bi-implication !!

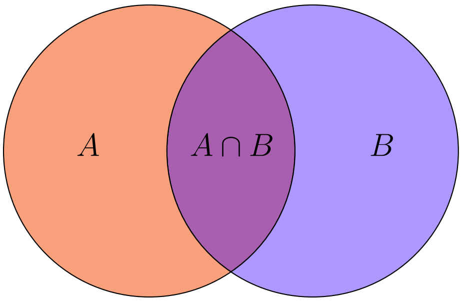
✍**Examples:**

E1. For the whole numbers greater than two, being odd is necessary to being prime, since two is the only whole number that is both even and prime.

E2. «John is a king» implies that John is male. So knowing that John is a king is sufficient to knowing that he is a male.

E3. A number's being divisible by 4 is sufficient (but not necessary) for it to be even, but being divisible by 2 is both sufficient and necessary for it to be even.

E4. Being in the purple region in the figure below is sufficient for being in A, but not necessary. Being in A is necessary for being in the purple region, but not sufficient. Being in A and being in B is necessary and sufficient for being in the purple region.



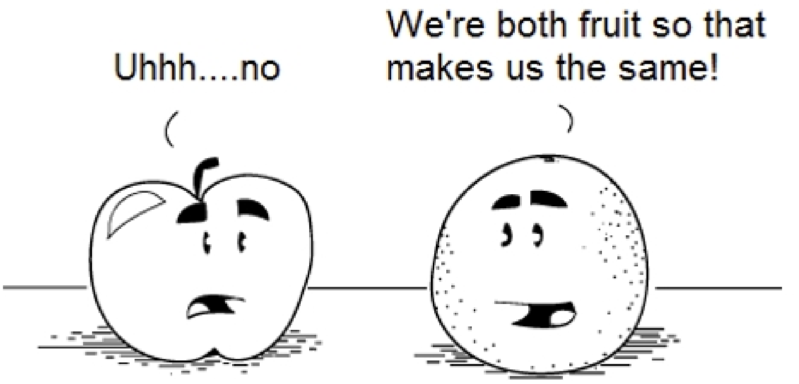
False equivalence is a logical fallacy in which an equivalence is drawn between two subjects based on flawed or false reasoning. This fallacy is categorized as a fallacy of inconsistency. Colloquially, a false equivalence is often named « *comparing apples and oranges* ».

This fallacy is committed when one shared trait between two subjects is assumed to show equivalence, especially in order of magnitude, when equivalence is not necessarily the logical result. False equivalence is a common result when an anecdotal similarity is pointed out as equal, but the claim of equivalence does not bear scrutiny because the similarity is based on oversimplification or ignorance of additional factors. The pattern of the fallacy is often as such:

If $\{x_1,\ldots,x_k,\ldots,x_n\} \in A$ and $\{x_1,\ldots,x_{k-m},\ldots,x_{k-m},\ldots,x_n\} \in B$ with $0 < m < n \in \mathbb{N}$, then since they both contain $\{x_1,\ldots,x_{k-m},x_{k-m},\ldots,x_n\}$ then $A = B$. Since $\{x_{k-m},\ldots,x_{k-m}\}$ is not required to exist in both sets; only a passing similarity is required to cause this fallacy to be used. More formally if $0 < m < n \in \mathbb{N}$:

$$(\{x_1,\ldots,x_k,\ldots,x_n\} \in A) \cap (\{x_1,\ldots,x_{k-m},\ldots,x_{k-m},\ldots,x_n\} \in B) \neq \varnothing \Rightarrow A \neq B \tag{5.1.19}$$

False equivalence arguments are often used on social networks, journalism, and in politics, where one strategical choice may be compared to flaws of a wholly different nature of another.


Uhhh....no
We're both fruit so that makes us the same!

The problem with popular articles, books or videos is that in the huge majority of cases the analogies to explain a concept are false equivalences.

Figure 5.3: **False equivalence fallacy**

Proof 5.0.1. Let two concepts or objects be represented by sets A and B. Each set contains the properties (or elements) that define the object:

$$A = \{a_1, a_2, \ldots, a_n\} \qquad B = \{b_1, b_2, \ldots, b_m\} \tag{5.1.20}$$

In set theory, equivalence (or identity) is defined by the axiom of extensionality. Two sets are equivalent if and only if they have exactly the same elements:

$$A = B \Leftrightarrow (\forall x, x \in A \Leftrightarrow x \in B) \tag{5.1.21}$$

This implies that $A \subseteq B$ and $B \subseteq A$. If even one property differs, the equivalence is broken.

An analogy does not assert total equality, but a non-empty intersection of specific properties in a given context.

If we draw an analogy between A and B based on a property P, this formally means:

$$\exists P : (P \in A \cap B) \tag{5.1.22}$$

However, by the very definition of analogy (which compares two different things), there exists a set of complementary non-empty properties:

- $A \backslash B \neq \varnothing$ (properties specific to A)
- $B \backslash A \neq \varnothing$ (properties specific to B)

To prove that an analogy is a false equivalence, we use the following reasoning:

1. **Analogy Hypothesis:**

   We set $A \approx B$ because they share the subset $S = A \cap B$.

2. **Equivalence Fallacy:**

   We attempt to assert that $A = B$ based solely on S.

3. **Contradiction :**

   Since this is an analogy and not an identity, there exists by definition an element x such that $x \in A$ but $x \notin B$ (or vice versa).

   Now, according to the definition of equivalence seen just above, if $x \in A$ and $x \notin B$, then $A \approx B$.

■ Q.E.D.

The first stage of propositional calculus is the formalization of natural language statements. To make this work, the propositional calculus finally provides three types of tools:

1. The "propositional variables" $(P, Q, R, \ldots)$ symbolize any simple proposals. If the same variable occurs multiple times, each time it symbolizes the same proposal.
2. The five logical operators: $\neg, \wedge, \vee, \Leftrightarrow, \Rightarrow$.
3. Punctuation are reduced to only opening and closing parentheses that organize reading so as to avoid ambiguity.

Here is a summary table:

Description	Symbol	Usage
The "Negation" is an operator that act only on one proposal; it is unary or monadic. "*It is not raining*" will be written $\neg P$. This statement is true if and only if P is false (in this case if it is false that it is raining). The conventional use of negation is characterized by the double negation law: $\neg\neg P$ is equivalent to P.	$\neg$	$\neg P$
The "conjunction" or "logical product" is a binary operator; it connects two proposals. "*Every man is mortal AND my car loses oil*" is written $P \wedge Q$. This latter expression is true if and only if P is true and Q is true.	$\wedge$	$P \wedge Q$
The "disconnection" or "logical sum" is also a binary operator. $P \vee Q$ is true if and only if P is true OR Q is true or both are true. We can understand the OR into two ways: either inclusively or exclusively. In the first case $P \vee Q$ is true if P is true, if Q is true or if P and Q are both true. In the second case, $P \vee Q$ is true if P is true or if Q is true but not if both are. The disjunction of propositional calculus is the inclusive OR and we give to the exclusive OR, that is to say the XOR, the name of "alternative".	$\vee$	$P \vee Q$
The "implication" or "inference" is also a binary operator. It corresponds roughly to the linguistic pattern "*If ... then ...*". "*If I have time, I will go see this movie*" will be written $P \Rightarrow Q$. This latter relation is false if P is true and Q is false. If the result (here Q) is true, the implication $P \Rightarrow Q$ is true. When the antecedent (here P) is false, the implication is always true. This latter remark can be understood if one refers to statements of the type: «*If we could put Paris in a bottle, the Eiffel Tower would be used as a plug*». In summary, an implication is false if and only if its antecedent is true and its consequent is false.	$\Rightarrow$	$P \Rightarrow Q$
The "bi-implication" or "equivalence" $\Rightarrow$ is, too, a binary operator: it symbolizes the terms "... *if and only if* ..." and "... *is equivalent to* ...". The equivalence of two propositions is true if they have the same truth value. The bi-implication therefore expressed as a form of identity, which is why it is often used in definitions.	$\Leftrightarrow$	$P \Leftrightarrow Q$

Table 5.9: Summary of logical core operators

And the corresponding summary truth table:

P	Q	$P \wedge Q$	$P \vee Q$	$P \Rightarrow Q$	$P \Leftrightarrow Q$	$\neg P$
T	T	T	T	T	T	F
T	F	F	T	F	F	F
F	T	F	T	T	F	T
F	F	F	F	T	T	T

Table 5.10: Truth table of main logical operators

The reader may read some authors that like to use at minimum the natural language in their books (...) and that use the symbol "$\therefore$" that is sometimes placed before a logical consequence, such as the conclusion of a syllogism. This symbol consists of three dots placed in an upright triangle and must be read as the word "*therefore*". We can also make use of the symbol "$\because$" that must be read as the word "*because*".

✍**Example:**

∵ All men are mortal.
∵ Socrates is a man.
∴ Socrates is mortal.

In this book we will however avoid using this notation as the engineers don't make use of that latter a lot.

It is possible to establish equivalences between these operators presented earlier. We have already seen how the biconditional could be defined as a product of reciprocal conditional, let us see now other equivalences:

$$\boxed{\begin{aligned}&(P\Rightarrow Q)\Leftrightarrow\neg(P\wedge\neg Q)\\&(P\Rightarrow Q)\Leftrightarrow(\neg P\vee Q)\\&(P\vee Q)\Leftrightarrow(\neg P\Rightarrow Q)\\&(P\wedge Q)\Leftrightarrow\neg(P\Rightarrow\neg Q)\end{aligned}}\tag{5.1.23}$$

Remark

The classical operators $\wedge,\vee,\Leftrightarrow$ can therefore be defined using the canonical operators $\neg,\Rightarrow$ through equivalence laws between them.

Also notice the two relations of De Morgan (see section of Boolean Algebra page 7110 for the proof):

$$\boxed{\begin{aligned}&\neg(P\vee Q)\Leftrightarrow(\neg P\wedge\neg Q)\\&\neg(P\wedge Q)\Leftrightarrow(\neg P\vee\neg Q)\end{aligned}}\tag{5.1.24}$$

that allow us to transform the disjunction into conjunction and vice versa:

$$\boxed{\begin{aligned}&(P\vee Q)\Leftrightarrow\neg(\neg P\wedge\neg Q)\\&(P\wedge Q)\Leftrightarrow\neg(\neg P\vee\neg Q)\end{aligned}}\tag{5.1.25}$$

To summarize the most classical laws of formal deductive logic:

Basics Laws of FOL	Formalism
Detachment (aka Modus Ponens)	$(p\rightarrow q)\wedge p\Rightarrow q$
Indirect reasoning (aka Modus Tollens)	$(p\rightarrow q)\wedge\neg q\Rightarrow\neg p$
Disjunctive addition	$p\Rightarrow(p\vee q)$
Conjunctive simplification	$(p\wedge q)\Rightarrow p$
Disjunctive simplification	$(p\vee q)\wedge\neg p\Rightarrow q$
Chain rule	$(p\rightarrow q)\wedge(q\rightarrow r)\Rightarrow(p\rightarrow r)$
Conditional equivalence	$p\rightarrow q\Leftrightarrow\neg p\vee q$
Biconditional equivalences	$(p\leftrightarrow q)\Leftrightarrow(p\rightarrow q)\wedge(q\rightarrow p)\Leftrightarrow(p\wedge q)\vee(\neg p\wedge\neg q)$
Contrapositive	$(p\rightarrow q)\Leftrightarrow(\neg q\rightarrow\neg p)$

Table 5.11: **Basics Deductive Logical Laws - Common Implications and Equivalences**

A part of the above table is often represented differently:

Rule of inference	**Name**	**Rule of inference**	**Name**
$\begin{array}{r} \neg q \\ p \to q \\ \hline \therefore \neg p \end{array}$	Modus Tollens	$\begin{array}{r} p \\ \hline \therefore p \lor q \end{array}$	Disjunctive amplification
$\begin{array}{r} p \\ p \to q \\ \hline \therefore q \end{array}$	Modus Ponens	$\begin{array}{r} \neg p \to \text{False} \\ \hline \therefore p \end{array}$	Contradiction
$\begin{array}{r} p \land q \\ \hline \therefore p \end{array}$	Simplification	$\begin{array}{r} p \land q \\ p \to (q \to r) \\ \hline \therefore (p \lor q) \to r \end{array}$	Conditional proof
$\begin{array}{r} p \\ q \\ \hline \therefore p \land q \end{array}$	Conjuction	$\begin{array}{r} p \to r \\ q \to r \\ \hline \therefore (p \lor q) \to r \end{array}$	Proof by cases
$\begin{array}{r} p \lor q \\ \neg p \\ \hline \therefore q \end{array}$	Disjunctive syllogism	$\begin{array}{r} p \to q \\ q \to r \\ \hline \therefore p \to r \end{array}$	Law of syllogism

5.1.3.3 Decision procedures

We have previously introduced the basic elements allowing us to operate on expressions from properties (propositional variables) without saying much about the handling of such expressions. So now you need to know that in propositional calculus there are two ways to establish that a proposition is a law of propositional logic. We can either:

1. Use non-automated procedures
2. Use axiomatic and demonstrative procedures

Remark

In many books these procedures are presented before the structure of the propositional language. We chose here to do the opposite thinking that the approach would be easier.

5.1.3.3.1 Non-axiomatic procedural decisions

Several of these methods exist, but we will limit ourselves here to the simplest of them, that of the matrix calculation, often referred to as "methods of truth tables".

The construction procedure is as we have already seen quite simple. Indeed, the truth value of a complex expression is a function of the truth value of the simple statements that compose it, and finally based on the truth value of propositional variables that makes it. Considering all possible combinations of truth values of propositional variables, we can determine the truth values of the complex expression.

Truth tables, as we have seen it, gives us the possibility to decide, about any proposition, if this latter is a tautology (always true), a contradiction (always false) or a contingent expression (sometimes true, sometimes false).

We can distinguish at least four ways to combine propositional variables, brackets and connectors:

	Name	Description	Example
1	Malformed statement	Nonsense. Neither true nor false	$(\vee P)Q$
2	Tautology	Statement always true	$P \vee \neg P$
3	Contradiction	Statement always false	$P \wedge \neg P$
4	Contingent statement	Statement sometimes true, sometimes false	$P \vee Q$

Table 5.12: **Combination of propositional variables**

The method of truth tables helps to determine the type of expression that are well-formed to which we have to face. It requires, in principle, no invention, it is "only" a mechanical procedure. Axiomatized procedures, however, are not entirely mechanical. Inventing a proof as part of an axiomatized system requires sometimes hability, practice or luck. Regarding to truth tables, here is the protocol to follow:

When facing a well-formed expression, or function of truth, we first determine how many distinct propositional variables we are dealing with. We then examine the various arguments that form this expression. We then construct a table with 2^n columns (n being the number of variables and without forgetting that they are binary variables!) and a number of columns equal to the number of arguments plus columns for the expression itself and its other components (see previous examples). Then we assign to the variables the various combinations of True (1) and False (0) values that may be conferred upon them. Each row corresponds to a possible outcome and all of the rows is the set of all possible outcomes. There is, for example, a possible outcome wherein P is a true statement while Q is false.

5.1.3.3.2 Axiomatic procedural decisions

The axiomatization of a theory implies, besides its formalization, that we start form a finite number of axioms and that through the controlled transformation of these, we can get all the theorems of this theory. So we start from a few axioms whose truth is a statement (not proven). We determine afterwards deduction rules for manipulating the axioms or expression obtained from these. The sequence of these deductions is a proof that leads to a theorem, a law or lemma.

We will now briefly present two axiomatic systems, each consisting of axioms using two specific rules named "inference rules" (intuitive rules):

1. The "modus ponens": If we prove A and $A \Rightarrow B$, then we can deduce B. A is named the "minor premise" and $A \Rightarrow B$ the "major premise" of the modus ponens rule.

   ✍**Example:**

   From:

   $$x>y$$

   and:

   $$(x>y)\Rightarrow(y<x)$$

   we can deduce that:

   $$y<x$$

   Remark

   Humans typically communicate in a way that resists shallow logical analysis. In a real conversation, people use words rather than terms, make utterances rather than sentences, and employ a wider variety of inference methods than modus ponens. A great deal of what is communicated and inferred in a conversation depends on context, the speakers and audience, their history, their shared knowledge and confidences, the feelers they lay out to establish mutual trust and rapport.

2. The "substitution": we can in a schema of axioms replace a letter by any formula, at the condition that all identical letters are replaced by identical formulas!

   Let us give as an example, two axiomatic systems: the axiomatic system of Whitehead and Russell, the axiomatic system of Lukasiewicz.

   (a) The axiomatic system of Russell and Whitehead adopts $\neg, \vee$ as primitive symbols and define $\Rightarrow, \wedge, \Leftrightarrow$ from these latter as follows (easily verifiable relations with truth tables):

$$\boxed{\begin{aligned}&(A \Rightarrow B) \Leftrightarrow \neg A \vee B\\&(A \wedge B) \Leftrightarrow \neg(\neg A \vee \neg B)\\&(A \Leftrightarrow B) \Leftrightarrow (A \Rightarrow B) \wedge (B \Rightarrow A)\end{aligned}} \tag{5.1.26}$$

This system includes 5 axioms, somewhat quite obvious plus two rules of inference. Axioms are given here using non-primitive symbols, as did Whitehead and Russell:

A1. $(A \vee A) \Rightarrow A$
A2. $B \Rightarrow (A \vee B)$
A3. $(A \vee B) \Rightarrow (B \vee A)$
A4. $(A \vee (B \vee C)) \Rightarrow (B \vee (A \vee C))$
A5. $(B \Rightarrow C) \Rightarrow ((A \vee B) \Rightarrow (A \vee C))$

we have already presented above some of these.

Remark

These five axioms are not independent of each other. The fourth can be obtained from the other four.

For example, to justify that $\neg A \vee A$ has a sense[2], we can proceed as following (in this system we only prove logical truths, meaning that every line need to be a logical truth by itself):

(1)	$B \Rightarrow (A \vee B)$	Axiom A2
(2)	$A \Rightarrow (A \vee A)$	(1) and substitution
(3)	$(B \Rightarrow C) \Rightarrow ((A \vee B) \Rightarrow (A \vee C))$	Axiom A5 (necessary)
(4)	$(B \Rightarrow C) \Rightarrow ((\neg A \vee B) \Rightarrow (\neg A \vee C))$	(3) and substitution
(5)	$(B \Rightarrow C) \Rightarrow ((A \Rightarrow B) \Rightarrow (A \Rightarrow C))$	(4) and property of $\Rightarrow$
(6)	$((A \vee A) \Rightarrow A) \Rightarrow ((A \Rightarrow (A \vee A)) \Rightarrow (A \Rightarrow A))$	(5) and substitution
(7)	$(A \Rightarrow (A \vee A)) \Rightarrow (A \Rightarrow A)$	(6) (modus ponens)
(8)	$(A \Rightarrow A) \Rightarrow (A \Rightarrow A)$	(7) and axiom A1
(9)	$A \Rightarrow A$	(8) and modus ponens
(10)	$\neg A \vee A$	(9) and property of $\Rightarrow$

(b) The axiomatic system of Lukasiewicz includes three axioms, plus the two rules of inference (modus ponens and substitution):

A1. $(A \Rightarrow B) \Rightarrow ((B \Rightarrow C) \Rightarrow (A \Rightarrow C))$
A2. $A \Rightarrow (\neg A \Rightarrow B)$
A3. $(\neg A \Rightarrow A) \Rightarrow A$

Here is the proof of the first two axioms in the system of Russell and Whitehead. These are the formulas (6) and (16) of the following derivation:

---

[2] It's named the "law of the excluded middle": A proposition is either true or false. This has been widely held to be self-evident for thousands of years

(1)	$(A \vee (B \vee C)) \Rightarrow (B \vee (A \vee C))$	Axiom A4
(2)	$(\neg(B \Rightarrow C) \vee (\neg(A \vee B) \vee (A \vee C))) \Rightarrow (\neg(A \vee B) \vee (\neg(B \Rightarrow C) \vee (A \vee C)))$	(1) and substitution
(3)	$\neg(A \vee B) \vee (\neg(B \Rightarrow C) \vee (A \vee C))$	A4 on (2) and modus ponens
(4)	$(A \vee B) \Rightarrow ((B \Rightarrow C) \Rightarrow (A \vee C))$	(3) by the property of $\Rightarrow$
(5)	$(\neg A \vee B) \Rightarrow ((B \Rightarrow C) \Rightarrow (\neg A \vee C))$	(4) and substitution
(6)	$(A \Rightarrow B) \Rightarrow ((B \Rightarrow C) \Rightarrow (A \Rightarrow C))$	(5) and property of $\Rightarrow$
(7)	$(B \Rightarrow (A \vee B)) \Rightarrow \begin{pmatrix} ((A \vee B) \Rightarrow (B \vee A)) \\ \Rightarrow (B \Rightarrow (B \vee A)) \end{pmatrix}$	(6) and substitution
(8)	$((A \vee B) \Rightarrow (B \vee A)) \Rightarrow (B \Rightarrow (B \vee A))$	(7) modus ponens
(9)	$B \Rightarrow (B \vee A)$	(8) modus ponens
(10)	$\neg B \Rightarrow (\neg B \vee A)$	(9) and substitution
(11)	$\neg\neg B \vee (\neg B \vee A)$	(10) and property of $\Rightarrow$
(12)	$\neg\neg B \vee (\neg B \vee A) \Rightarrow (\neg B \vee (\neg\neg B \vee A))$	A4 + substitution with (11)
(13)	$\neg B \vee (\neg\neg B \vee A)$	(12) and modus ponens
(14)	$B \Rightarrow (\neg\neg B \vee A)$	(19) and property of $\Rightarrow$
(15)	$B \Rightarrow (\neg B \vee A)$	(14) and property of $\Rightarrow$
(16)	$A \Rightarrow (\neg A \Rightarrow B)$	(15) and substitution

These axiomatizations let us found as theorems all tautologies or laws of the propositional logic. From everything that has been said so far, we can try to define what is a proof!!!!

Definition (#5.1.15): A finite sequence of formulas $B_1, B_2, \ldots, B_m$ is named a "proof" from the assumptions/hypothesis $A_1, A_2, \ldots, A_n$ if for each i:

- B_i is one of the assumptions/hypothesis $A_1, A_2, \ldots, A_n$
- or B_i is a variant of an axiom
- or B_i is inferred (by the application of the modus ponens rule) from the major premise B_j and minor premise B_k where $j, k < i$
- or B_i is inferred (by the application of the substitution rule) from an anterior premise B_j, the replaced variable not appearing in $A_1, A_2, \ldots, A_n$

Such a sequence of formulas, B_m being the final formula of the sequence, is more explicitly named "proof of B_m" from the assumptions/hypothesis (or axioms) $A_1, A_2, \ldots, A_n$, what we also write:

$$\boxed{A_1, A_2, \ldots, A_n \vdash B_m} \tag{5.1.27}$$

More explicitly a proof is a deductive argument for a mathematical statement. In the argument, other previously established statements, such as theorems, can be used. In principle, a proof can be traced back to self-evident or assumed statements, known as axioms.

Proofs employ logic but usually include some amount of natural language which usually admits some ambiguity. In fact, the vast majority of proofs in written mathematics can be considered as applications of rigorous informal logic. Purely formal proofs, written in symbolic language instead of natural language, are considered in proof theory.

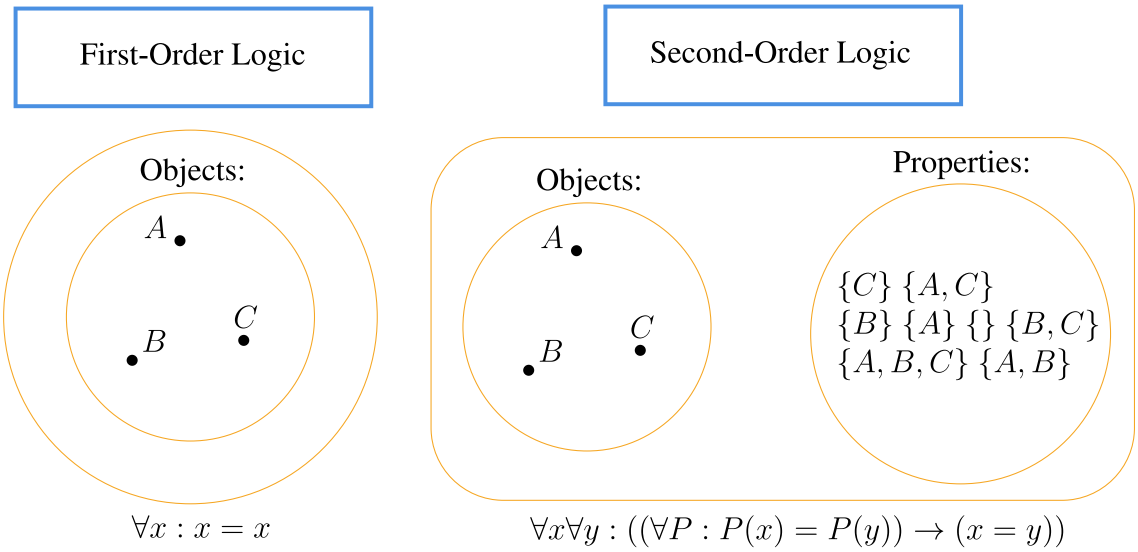
Remark

Note that when we try to prove a result from a number of assumptions, we do not try to prove the assumptions themselves!

5.1.3.4 Quantifiers

We have to complete the use of the connectors of propositional calculus by what we name "quantifiers" if we wish to solve some problems. Indeed, the propositional calculus does not allow us to state general things about the elements of a set, for example. In that sense, propositional logic is only part of the reasoning. The calculus of predicates on the contrary allows to formally handle statements such as "there exists an x such that [x has an American car]" or "for all x [if x is a dachshund, then x is small]". In short, we extend the composed formulas in order to assert existential quantifiers ("there...") and universal quantifiers ("for every..."). The examples we just gave involve a bit special proposals like "x has an American car." This is proposition with a variable. These proposals are in fact the application of a function to x. This function, is this that associates "x has an American car" with x. We will denote this function by "... has an American car" and we say that is a propositional function because it is a function whose value is a proposal. Or a "predicate" as we already know.

The existential and universal quantifiers go hand in hand with the use of propositional functions. The predicate calculus is however limited in the existential and universal formulas. Thus, we prohibit ourselves to use sentences like "there is an affirmation of x such that ...". In fact, we allow ourselves to quantify only "individuals". This is why predicate logic is named "first-order logic" (FOL) or "first-order predicate logic" (FOPL) because it uses variables as basic mathematical objects (while in the second-order logic they can also be sets).



Before going back to the study of the predicate calculus we define:

D1. The "universal quantifier": $\forall$ (for all)

D2. The "existential quantifier": $\exists$ (exists)

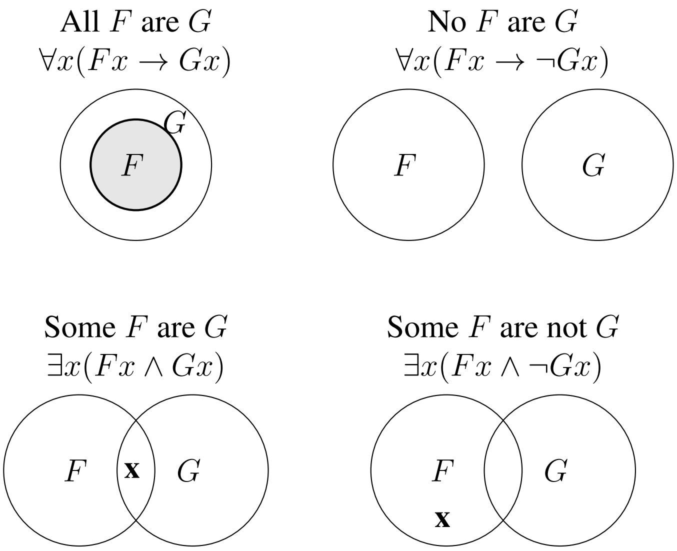

All F are G
$\forall x(Fx \rightarrow Gx)$
G
F
No F are G
$\forall x(Fx \rightarrow \neg Gx)$
F
G
Some F are G
$\exists x(Fx \wedge Gx)$
F
x
G
Some F are not G
$\exists x(Fx \wedge \neg Gx)$
F
G
x

Figure 5.4: **The four standard types of quantified statements**

✍**Example:**

If any complex number is the product of a non-negative number and a number of modulus 1 we will write:

$$\forall z \in \mathbb{C}, \exists p \in \mathbb{Z} \geq 0, \exists u \in \mathbb{C} : \quad (|u| = 1 \wedge z = pu) \tag{5.1.28}$$

The order of quantifiers is critical to meaning, as illustrated by the following two propositions (by the way it's a good test to check if someone know the basics of logic):

For every natural number n, there exists a natural number s such that $s = n^2$.
$\forall n \in \mathbb{N}, \exists s \in \mathbb{N}, s = n^2$

This is clearly true! It just asserts that every natural number has a square. The meaning of the assertion in which the quantifiers are turned around is different:

There exists a natural number s such that for every natural number n, $s = n^2$.
$\exists s \in \mathbb{N}, \forall n \in \mathbb{N}, s = n^2$

This is clearly false! It asserts that there is a single natural number s that is at the same time the square of every natural number.

A frequent question in physics and mathematics is to know if the universal quantifier has to be before of after the predicates they refer to. In fact, strictly in terms of formal logic, quantifiers are always at the beginning of any formula. However, almost no one gives a proof that is written in the formal language. Even simple proofs would be very long and unreadable. But anyone, regardless of what natural language they speak, will interpret a sentence in the formal language in the same way. The price for this clarity of course is readability. Natural languages, because of their inherent ambiguity, are subject to many more limitations.

Obviously the proper usage of a formal notation or of a more informal one depends particularly on the context of presentation. It is essential to whom we communicate an idea and this should guide us to use a suitable level of formal notation.

We use the sometimes the symbol ∃! to say briefly: "there is one and only one".

✍**Example:**

A famous example is the way to explicit that the logarithm is a bijective function:

$$\forall x \in \mathbb{R}^{+}, \exists ! y \in \mathbb{R}_{+}^{*}: \quad x=\ln (y) \tag{5.1.29}$$

Suppose now we claim that there is no smallest number in $\mathbb{R}$. We can translate this into symbols as:

$$\neq \exists x \in \mathbb{R} \forall y \in \mathbb{R}:(x \leq y) \tag{5.1.30}$$

Literally: «it is not true that there is a real number x such tat for all real numbers y, x is less than or equal to y».

However, we can negate quantifiers: we can pass a negation over a quantifier by switching the quantifier type (between universal and existential). So the statement above should be logically equivalent to:

$$\forall x \in \mathbb{R} \exists y \in \mathbb{R}:(y<x) \tag{5.1.31}$$

Notice that $y<x$ is the negation of $x \leq y$. This literally says: «for every real number x there is a real number y which is small than x». We see that this is another way to make our original claim.

Before moving to predicate calculus, let us have a non exhaustive summary that is common in the literature and uses negation:

Expression		Is equivalent to
$\neg\neg A$	$\Leftrightarrow$	A
$\neg A \Rightarrow \neg B$	$\Leftrightarrow$	$B \Rightarrow A$
$\neg A \Rightarrow B$	$\Leftrightarrow$	$\neg B \Rightarrow A$
$A \Rightarrow \neg B$	$\Leftrightarrow$	$(A \wedge B) \Rightarrow F$
$\neg A \wedge \neg B$	$\Leftrightarrow$	$\neg(A \vee B)$
$A \wedge \neg B$	$\Leftrightarrow$	$\neg(\neg A \vee B)$
$\wedge A \neg B$	$\Leftrightarrow$	$\neg(A \vee \neg B)$
$\neg A \vee \neg B$	$\Leftrightarrow$	$\neg(A \wedge B)$
$A \vee \neg B$	$\Leftrightarrow$	$\neg(\neg A \wedge B)$
$\neg A \vee B$	$\Leftrightarrow$	$\neg(A \wedge \neg B)$
$\neg(\exists x \in X : \neg P(x))$	$\Leftrightarrow$	$(\forall x \in X : P(x))$
$(\exists x \in X : \neg P(x))$	$\Leftrightarrow$	$\neg(\forall x \in X : P(x))$
$(\exists x \in X : P(x))$	$\Leftrightarrow$	$(\forall x \in X : \neg P(x))$
$(\exists x \in X : P(x))$	$\Leftrightarrow$	$\neg(\forall x \in X : \neg P(x))$

We will see now that the Proof Theory and Set Theory are the exact transcription of the principles and results of Logic (the one with a capitalized "L"!!!).

5.1.4 Predicate Calculus

In mathematics courses (algebra, analysis, geometry, ...), we prove the properties of different object types (integer, real, matrices, sequences, continuous functions, curves, triangles, ...). To prove these properties, we need of course that the objects on which we work are clearly defined (what is a set, what is a real, what is point, ...?).

In first-order logic and, in particular, in proof theory, the objects we study are the formulas and proofs. We must therefore give a precise definition of what are these objects. The terms and formulas are the grammar of a language, oversimplified and calculated exactly to say what we want without ambiguity and without unnecessary detour.

In the previous sections we saw that syllogistics treats sentences differently than propositional logic. Syllogistics is concerned with the components of sentences, notably the quantifiers «all», «none» and «some», while propositional logic abbreviates any sentence – regardless of its components – with letters like « p », « q », « r », etc. Thus, propositional logic is not capable of explaining the validity of inferences from typical human rhetorical sentences.

Propositional logic does not allow us to explain the validity of rhetorical inferences, because it does not take into account the internal structure of the simple sentences it processes: it treats « p » and « q » as atoms and does not analyze their internal structure.

To formalize rhetorical inferences and explain their validity, it is necessary to use the notion of predicate and convert the concepts into algebraic letters. For example, the following inference:

$$\begin{array}{c}\text{All humans are mortal}\\ \text{All philosophers are humans}\\ \hline \text{All philosophers are mortal}\end{array} \tag{5.1.32}$$

Therefore becomes:

$$\begin{array}{c}\text{All the } M \text{ are } P\\ \text{All the } S \text{ are } M\\ \hline \text{All the } S \text{ are } P\end{array} \tag{5.1.33}$$

If we abbreviate « all the . . . are . . . » by « $\forall x(\ldots x\ldots \rightarrow \cdots x\ldots)$ », we get :

$$\begin{array}{c}\forall x(Mx \rightarrow Px)\\ \forall x(Sx \rightarrow Mx)\\ \hline \forall x(Sx \rightarrow Px)\end{array} \tag{5.1.34}$$

Thus we have converted the syllogism into a valid inference!

Remark

In standard first-order logic, the statement « *All men are mortals* » is written as $\forall x(M(x) \rightarrow \text{Mortal}(x))$. In standard logic, saying « *It is false that all men are mortal* » (written technically $\neg\forall x(M(x) \rightarrow \text{Mortal}(x))$) means exactly the same thing as saying « *There is at least one man who is not mortal* »: $\exists x(M(x) \rightarrow \neg\,\text{Mortal}(x))$.

5.1.4.1 Grammar

Definitions (#5.1.16):

D1. The "terms" designate items for which we want to prove some properties (we will discussed the latter much more in details further below):

- In algebra, the terms refer to the elements of a set (group, ring, field, vector space, etc.). We also manipulate sets of objects (subgroups, subrings, subfields, etc.). The terms which will designate the objects are named "second-order terms".
- In analysis, the terms refer most of time to real numbers (for example, if we put ourselves in functional spaces) or functions.

D2. The "formulas", are the properties of objects we study (we will discussed the latter also much more in details further below):

- In algebra, we can write formulas to express that two elements commute, that a subspace is of dimension 3, etc.
- In analysis, we will write formulas to express the continuity of a function, the convergence of a sequence, etc.
- In set theory, formulas can express inclusion of two sets, membership of an element in a set, etc.

D3. The "proof", enable to check if a formula is true. The precise meaning of this word will also need to be defined. More precisely, they are deductions under assumptions, they allow to "lead from truth to truth", the question of the truth of the conclusion is then returned to that of the hypothesis, which does not look at the logic but is based on the knowledge we have on things we talk about.

5.1.4.2 Languages

In mathematics we use, depending on the area, different languages that are distinguished by the symbols used. The definition below simply expresses that it is sufficient to only have to give the list of symbols to specify the language.

Definition (#5.1.17): A "language" is the content of a family (not necessarily finite) of symbols. We distinguish three kinds of languages: symbols, terms and formulas.

Remarks

R1. We use sometimes the word "vocabulary" or "signature" instead of the word "language".

R2. We already know that the word "predicate" is used instead of the word "relation". We speak then of "predicate calculus" instead of "first-order logic").

5.1.4.2.1 Symbols

There are different types of symbols we will try to define!

Definitions (#5.1.18):

D1. The "constant symbols" (see note below)

✍**Example:**

The neutral element n in Set Theory (see section Set Theory page 417)

D2. The "function symbols" or "functors". To each function symbol is assigned a strictly positive integer that we name the "arity": this is the number of arguments of the function arguments. If the arity is 1 (resp. $2, \ldots, n$), we say then that the function is unary (resp. binary, ..., n-ary).

✍**Example:**

The binary functor of multiplication $\times$ or $\cdot$ in group theory (see section Set Theory page 436) is 2-ary or more commonly a "binary" operator.

D3. The "relation symbol". Similarly to the previous definition, every relation symbol is associated with a positive or null integer (its arity) that corresponds to its number of arguments and we talk then of unary, binary, n-ary relation.

✍**Example:**

The relation $=$ is a binary relation (see section Operators page 287)

D4. The "individual variables". In what will follow we will give us an infinite set $\mathcal{V}$ of variables. The variables will be recorded as it is traditional by the latin lowercase letters: x, y, z (possibly indexed: x_1, x_2, x_3).

D5. To this we should add the connectors $\neg, \Rightarrow, \vee, \wedge$ and quantifiers $\forall, \exists, \exists!$ that we extensively discussed above, on which it is now useless to return.

Remarks

R1. A constant symbol can be seen as a function symbol with 0 argument (arity zero).

R2. We consider (unless otherwise stated) that each language contains the binary relation symbol $=$ (read "equal") and the relation symbol with zero argument denoted $\perp$ (read "bottom" or "absurd/contradiction") represents the value FALSE and $\top$ (read "top or "tautology") represents the value TRUE. In the description of a language, we will often omit to mention them. The symbol $\perp$ is often redundant. We can indeed, without using it, write a formula that is always false. However, it can represent the FALSE in a canonical way and therefore gives the possibility to write general proof rules.

R3. The role of functions and relations is very different. As we will see, the function symbols are used to construct the terms (of language objects) and the relation symbols to build formulas (the properties of these objects).

5.1.4.2.2 Terms

The terms, we also say the "first order terms", are the objects associated with the language.

Definitions (#5.1.19):

D1. Given $\mathcal{L}$ a language, the set $\mathcal{T}$ of terms on $\mathcal{L}$ is the smallest set containing the variables, the constants and stable (we do not go out of the set) by applying function symbols of $\mathcal{L}$ to the terms.

D2. A "closed term" is a term that does not contain variables (by extension, only constants).

D3. For a more formal definition, we can write:

$$\mathcal{T}_0 = \{t\} \tag{5.1.35}$$

where t is a variable or constant symbol and, for any $k \in \mathbb{N}$:

$$\mathcal{T}_{k+1} = \mathcal{T}_k \cup \{f(t_1, \ldots, t_n) | t_i \in \mathcal{T}_k\} \tag{5.1.36}$$

where f is obviously a function of arity n (let us recall that the arity is the number of function arguments). Thus, for each arity, there is a degree of set of terms. We have finally:

$$\mathcal{T}=\bigcup_{k\in\mathbb{N}}\mathcal{T}_k \tag{5.1.37}$$

D4. We name "height" of a term t the smallest k such that $t\in\mathcal{T}_k$.

This definition means that variables and constants are terms and that if f is a n-ary function symbol and $t_1,\ldots,t_n$ are terms then $f(t_1,\ldots,t_n)$ is also a term itself. The set $\mathcal{T}$ of terms is defined by the grammar:

$$\boxed{\mathcal{T}=\{V|S_c|S_f(\mathcal{T},\ldots,\mathcal{T})\}} \tag{5.1.38}$$

This expression must be read as follows: a element of the set $\mathcal{T}$ we are defining is either an element of V (variable) or an element of S_c (the set of symbols of constant), or the application of a n-ary function symbol $f\in S_f$ (constants or variable) of $\mathcal{T}$.

Caution! The fact that f is of the good arity is only implicit in this notation. Moreover, writing $S_f(\mathcal{T},\ldots,\mathcal{T})$ does not mean that all function arguments are identical, but simply that these arguments are elements of $\mathcal{T}$.

Remark

It is often convenient to see a term (expression) as a tree, where each node is labelled with a function symbol (operator or function) and each sheet by a variable or constant.

In what follows, we will almost always define concepts (or prove results) "by recurrence" (see page 218) on the structure or the size of a term.

Definitions (#5.1.20):

D1. To prove a property P on the terms, it suffices to prove P for the variables and constants and to prove $P(f(t_1,\ldots,t_n))$ from $P(t_1),\ldots,P(t_n)$. We do then here a "proof by induction on the height of a term". It is a technique that we will find in the following chapters.

Mathematical induction as an inference rule can be formalized as a second-order axiom. The axiom of induction is, in logical symbols:

$$\forall P.\,[[P(0)\wedge\forall(k\in\mathbb{N}).\,[P(k)\Rightarrow P(k+1)]]\Rightarrow\forall(n\in\mathbb{N}).\,P(n)] \tag{5.1.39}$$

In words, the basis $P(0)$ and the inductive step (namely, that the inductive hypothesis $P(k)$ implies $P(k+1)$) together imply that $P(n)$ for any natural number n. The axiom of induction asserts that the validity of inferring that $P(n)$ holds for any natural number n from the basis and the inductive step.

Induction can be compared to falling dominoes: whenever one domino falls, the next one also falls. The first step, proving that $P(1)$ is true, starts the infinite chain reaction.

D2. To define a function Φ based on the terms, it is enough to define it on the variables and constants and tell how we get $\Phi(f(t_1,\ldots,t_n))$ from $\Phi(t_1),\ldots,\Phi(t_n)$. We do here again a "definition by induction on the height of a term".

✍**Example:**

The size (we also say the "length") of a term t (size denoted $\tau(t)$) is the number of function symbols occurring in t. formally:

- $\tau(x)=\tau(c)=0$ if x is a variable and c is a constant.
- $\tau(f(t_1,\ldots,t_n))=1+\sum\limits_{i\leq i\leq n}\tau(t_i)$

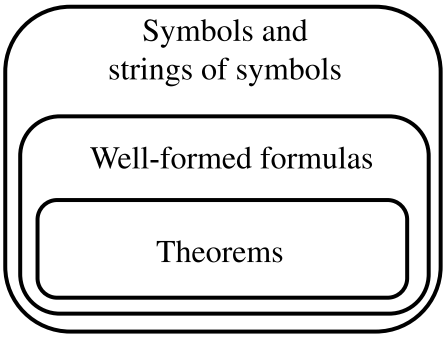
where the 1 in the last relation represents the term f itself.

5.1.4.2.3 Formulas

Definition (#5.1.21): A "well-formed formula" (WFF), often simply "formula" is a word (i.e. a finite sequence of symbols from a given alphabet) that is part of a formal language. A formal language can be considered to be identical to the set containing all and only its formulas.

The formulas of propositional calculus, also named "propositional formulas", are expressions such as $(A\wedge(B\vee C))$.

An atomic formula is a formula that contains no logical connectives nor quantifiers, or equivalently a formula that has no strict sub-formulas. The precise form of atomic formulas depends on the formal system under consideration; for propositional logic, for example, the atomic formulas are the propositional variables. For predicate logic, the atoms are predicate symbols together with their arguments, each argument being a term.



Definition (#5.1.22): Formulas are built from "atomic formulas" using connectors and quantifiers. We will use the following connectors and quantifiers (which we already known):

- Unary negation connector: $\neg$
- Binary connectors of conjunction and disjunction and implication: $\wedge,\vee,\rightarrow$
- Quantifiers: $\exists$ which must be read "it exists" and $\forall$ that must be read "for all"

This notation of the connectors is almost (it should at least). It is used to avoid confusion between the formulas and the current language (metalanguage).

Definitions (#5.1.23):

D1. Given $\mathcal{L}$ a language, the "atomic formulas" of $\mathcal{L}$ are the formulas of the form $R(t_1,\ldots,t_n)$ where R is an n-ary relation symbol of $\mathcal{L}$ of and $t_1,\ldots,t_n$ are terms of $\mathcal{L}$. We denote by "Atom" all atomic formulas. If we denote by S_r the set of relation symbols, we can write all terms related between them by the expression:

$$\boxed{\text{Atom} = S_R(\mathcal{T},\ldots,\mathcal{T})\quad \mathcal{T}\in\mathcal{L},\forall t_i\in\mathcal{T}}\tag{5.1.40}$$

The set $\mathcal{F}$ of formulas of the first order logic of $\mathcal{L}$ is thus defined by the grammar (where x is a variable):

$$\boxed{\mathcal{F} = \text{Atom}|F\wedge F|F\vee F|F\rightarrow F|F\neg F|\exists x\mathcal{F}|\forall x\mathcal{F}}\tag{5.1.41}$$

where should be read: the set of formulas is the smallest set containing formulas and such that if F_1 and F_2 are formulas then $F_1\vee F_2$, etc. are formulas and can be related to each other.

✍**Example:**

The relations symbols of the propositional language are relations of 0-arity (even the symbol "$=$" is absent), the quantifiers are then useless (since a propositional formula can not contain variables). We then obtain the propositional calculus $\mathcal{C}_P$ defined by:

$$\boxed{\mathcal{C}_P = V_P|\bot|\neq\mathcal{C}_P|\mathcal{C}_P\vee\mathcal{C}_P|\mathcal{C}_P\wedge\mathcal{C}_P|\mathcal{C}_P\rightarrow\mathcal{C}_P}\tag{5.1.42}$$

Note the presence of the symbol "bottom" $\bot$ meaning the "false" that we did not mention in our previous study of propositional logic.

The reader must be careful to not to confuse terms and formulas. $\sin(x)$ is a term (function), $x=3$ is a formula. But $\sin(x)\wedge(x=3)$ is nothing: we can not, in fact, put a connector between a term and a formula (meaningless).

Remarks

R1. To define a function Φ based on formulas, we simply need to define Φ on atomic formulas.

R2. To prove a property P on formulas, it suffices to prove P for the atomic formulas.

R3. To prove a property P on the formulas, it is enough assume that the property holds for all formulas of size $p < n$ and to prove the property for formulas of size n.

D2. A "sub-formula" of a formula (or expression) F is one of its components, verbatim a formula from which F is built. Formally, we define the set $\text{SF}(F)$ of the sub-formulas F by:

- If F is atomic:

$$\text{SF}(F)=\{F\} \tag{5.1.43}$$

- If $F=F_1\oplus F_2$ (that is to say a composition) with $\oplus\in\{\vee,\wedge,\rightarrow\}$:

$$\text{SF}=(F)=\{F\}\cup\text{SF}(F_1)\cup\text{SF}(F_2) \tag{5.1.44}$$

- If $F=\neg F$ or $\text{Q}\in F_1$ with $\text{Q}\in\{\forall,\exists\}$:

$$\text{SF}(F)=\{F\}\cup\text{SF}(F_1) \tag{5.1.45}$$

D3. A formula F of $\mathcal{L}$ uses only a finite number of symbols of $\mathcal{L}$. This subset is named the "language of the formula" and is denoted by $\mathcal{L}(F)$.

D4. The "size (or length) of a formula F", denoted by $\tau(F)$ is the number of connectors or quantifiers occurring in F. Formally:

- $\tau(F)=0$ if F is an atomic formula
- $\tau(F_1\oplus F_2)=1+\tau(F_1)+\tau(F_2)$ where once again $\oplus\in\{\vee,\wedge,\rightarrow\}$
- $\tau(\neg F_1)=\tau(\text{Q}xF_1)=1\tau(F_1)$ with once again $\text{Q}\in\{\forall,\exists\}$

D5. The "main operator" (we also say the "main connector") of a formula is defined as:

- If A is atomic, so it has then it has no main operator
- If $A=\neg B$, the $\neg$ is the main operator of A
- If $A=B\oplus C$ where once again $\oplus\in\{\vee,\wedge,\rightarrow\}$ then $\oplus$ is the main operator of A
- If $A=\text{Q}xB$ where once again $\text{Q}\in\{\forall,\exists\}$, then Q is the main operator of A

D6. Given F a formula. The set $\text{FV}(F)$ of free variables of F and the set $\text{DV}(F)$ of dummies variables (or "linked variables") of F are defined by induction on $\tau(F)$.

An occurrence of a given variable is named "linked variable" or "dummy variable" in a formula F in a formula F, if in this formula a quantifier refers to it. Otherwise, we say we have a "free variable".

Remark

An occurrence of a variable x in a formula F is a position of the variable in the formula F. Do not confuse with the object that is the variable itself!

To clarify the possible free variables of a formula F, we write $F[x_1, \ldots, x_n]$. This means that free variables of F are among the variables $x_1, \ldots, x_n$, verbatim if y is free in F, then is one of the x_i but the x_i do not necessarily appear in F.

We can define the dummy or free variables more formally:

(a) If $F = R(t_1, \ldots, t_n)$ is atomic then $\mathrm{FV}(F)$ is the set of free variables appearing in the t_i and we have then for dummy variables $\mathrm{DV}(F) = \varnothing$.

(b) If $F = F_1 \oplus F_2$ where $\oplus \in \{\vee, \wedge, \rightarrow\}$: $\mathrm{FV}(F) = \mathrm{FV}(F_1) \cup \mathrm{FV}(F_2)$ then $\mathrm{DV}(F) = \mathrm{DV}(F_1) \cup \mathrm{DV}(F_2)$.

(c) If $F = \neg F_1$ then $\mathrm{FV}(F) = \mathrm{FV}(F_1)$ and $\mathrm{DV}(F) = \mathrm{DV}(F_1)$.

(d) If $F = \mathrm{Q}xF_1$ with $\mathrm{Q} \in \{\forall, \exists\}$: $\mathrm{FV}(F) = \mathrm{FV}(F_1) - \{x\}$ and $\mathrm{DV}(F) = \mathrm{DV}(F_1) \cup \{x\}$.

✍**Examples:**

E1. Given F: $\forall x\ (x \cdot y = y \cdot x)$ then $\mathrm{FV}(F) = \{y\}$ and $\mathrm{DV}(F) = \{x\}$

E2. Given G: $\{\forall x \exists y (x \cdot z = z \cdot y)\} \wedge \{x = z \cdot z\}$ then $\mathrm{FV}(G) = \{x, z\}$ and $\mathrm{DV}(G) = \{x, y\}$.

D7. We say that formulas F and G are "α-equivalent" if they are (syntactically) identical only after the renaming of their related variables.

D8. A "closed formula" is a formula without free variables.

D9. Given F a formula, x a variable and t a term. $F[x := t]$ is the formula obtained by replacing in F all free occurrences of x by t, after possible renaming of linked occurrences in F which are free in t.

Remarks

R1. We will notice in the examples seen previously that a variable can have both free occurrences and linked occurrences. So we do not always have $\mathrm{FV}(F) \cap \mathrm{DV}(F) = \varnothing$.

R2. We can not rename y into x in the formula $\forall y\ (x \cdot y = y \cdot x)$ of the previous example and get the formula $\forall x\ (x \cdot x = x \cdot x)$: the variable x would be then "captured". We therefore can not rename variables without precautions: we must avoid capture of free occurrences.

5.1.5 Proofs

The proofs that we found in mathematical books or theoretical physics books are assemblies of mathematical symbols and sentences containing keywords such as: "So", "because", "if", "if and only if", "it is necessary", "just", "take an x such that", "therefore", "assume", "seek a contradiction", etc. These words are assumed to be understood by all in the same way, which is in fact, not always the case.

In any work, the purpose of a proof is to convince the reader of the truth of a statement by show him the intellectual path that gives him the possibility to control himself the truth and rigour of the statement. Depending on the level of the reader, this proof will be more or less detailed or formal: something that can be considered obvious in a graduate course may not be in a undergraduate level course.

In a homework, the corrector know that the result given by the student is (normally) true and he knows the proof of it. The student must prove (correctly) the required result. The level of detail that the student must give will depends sometimes on the confidence possessed by the corrector: in a good copy, a "proof by an evident recurrence" will be accepted, while a copy where there previously had an "obvious" which was ... obviously false, will not pass!

To properly manage the level of details, we should know what is a complete proof. This work of formalization has been done at the beginning of the 120th century only (holocene calendar)!!

Several things may seem surprising:

1. There is only a finite number of rules: two for each of the connectors (and the equality) more three general rules. It was not at all clear before that a priori a finite number of rules were enough to prove all that is true. We will show this result (this is essentially the Completeness Theorem). The proof is not trivial at all.

2. These are the same rules for all the mathematics and physics: algebra, analysis, geometry, etc. This means that we have managed to isolate what is general in reasoning!!! We will see later that a proof is an assembly of pairs (Γ, A), where Γ is a set of formulas (the assumptions) and A a formula (the conclusion). When we do the arithmetic, geometry or real analysis, we use, in addition to rules, assumptions that are named as we know "axioms". These express the particular properties of objects that we manipulate (for details on the concept of axioms see the section Introduction section page 100).

Remark

It is important to study the abstract theory of Proofs, even if the "Münchhausen trilemma" (see section Principia page 3128) seems to give evidence in a rhetoric way (with all the weaknesses associated to this art...) the impossibility of proving any truth in the fields of logics and mathematics (keep in mind however that pure sciences are not similar to experimental sciences!).

We prove therefore, in general, formulas using a set of assumptions, and this set can vary during the proof: when we say "suppose F and let us prove G", F is then a new hypothesis that we can use to prove G. to formalize this, we introduce the concept of "sequent".

Definitions (#5.1.24):

D1. A "sequent" is a pair $\Gamma \vdash F$ where:

  (a) Γ is a finite set of formulas that represents the assumptions that we can use. This set is also named the "context of the sequent".

  (b) F is a formula. This is the formula we want to prove. We say that this formula is the "conclusion of the sequent".

  The sign "$\vdash$" must be read "thesis" or "prove that".

  Remarks

  R1. If $\Gamma = \{A_1, \ldots, A_n\}$ we can then write $A_1, \ldots, A_n \vdash F$ instead of $\Gamma \vdash F$.

  R2. We write $\vdash F$ a sequent whose set of assumptions is empty and $\Gamma_1, \ldots, \Gamma_n \vdash F$ a sequent whose set of assumptions is $\bigcup_i \Gamma_i$.

  R3. We write $\Gamma \nvdash F$ to say that "F is not provable".

D2. A sequent $\Gamma \vdash F$ is said to be "provable" (or demonstrable) if it can be obtained by a finite application of rules. A formula F is provable if the sequent $\vdash F$ is provable.

5.1.5.1 Rules of Proofs

Proofs rules are the bricks used to build demonstration steps. A formal demonstration is a finite (and correct!) assembly of rules. This assembly is not linear (not a suite) but a "tree." Indeed, we are often forced to make connections.

We will present a choice of rules. We could have introduced other (instead of or in addition) that would give the same notion of provability. Those that have been chosen are "natural" and correspond to the arguments that we usually made in mathematics. In the common practice we use, in addition to the rules below, many other rules, but these can be deduced from previous ones. We name them "derived rules".

It is traditional to write the root of the tree (the sequent conclusion) at the bottom, the leaves at the top: the nature is done as this... As it is also tradition to write on a sheet of paper, from up to down, it would not be unreasonable to write the root at the top and the leaves at the bottom. We must make a choice!

A rule consists of:

- a set of "premises" each is a sequent. There may be zero, one or more of them.
- the conclusion sequent of the rule
- a horizontal bar separating the premises (top) from the conclusion (bottom). On the right of the bar, we will indicate the name of the rule.

✍**Example:**

Let us consider:

$$\frac{\Gamma \vdash A \rightarrow B \; \Gamma \vdash A}{\Gamma \vdash B} \rightarrow_e \tag{5.1.46}$$

This rule has two premises ($\Gamma \vdash A \rightarrow$ and $\Gamma \vdash A$) and a conclusion ($\Gamma \vdash B$) and is denoted in an abbreviated under the form $\rightarrow_e$. It can be read in two ways:

- from bottom to top: if we want to prove the conclusion, it suffices by using the rule to prove the premises. This is what we do when we seek a proof. This corresponds to the "analysis".
- from top to bottom: if we proved the premises, so we also proved the conclusion. This is what we do when we write a proof. This corresponds to the "synthesis".

For the proofs there is a finite number of 17 rules in number that we will define below:

1. Axiom:

$$\boxed{\frac{}{\Gamma, A \vdash A}\text{ax}} \tag{5.1.47}$$

   From bottom to top: if the conclusion of the sequent is one of the hypothesis, then the sequent is provable.

2. Weakening:

$$\boxed{\frac{\Gamma \vdash A}{\Gamma, B \vdash A}\text{aff}} \tag{5.1.48}$$

   Explanations:

   - From top to bottom: if we prove A under the assumptions Γ, adding other hypotheses can still prove A.
   - From bottom to top: there are assumptions that may not serve

3. Introduction of implication:

$$\boxed{\frac{\Gamma, A \vdash B}{\Gamma \vdash B \rightarrow A} \rightarrow_i} \tag{5.1.49}$$

   From bottom to top: to prove that $A \rightarrow B$ we assume A (that is to say, we add it to the assumptions) and we prove B.

4. Elimination of implication:

$$\boxed{\frac{\Gamma \vdash A \rightarrow B \quad \Gamma \vdash A}{\Gamma \vdash B} \rightarrow_e} \tag{5.1.50}$$

   From bottom to top: to prove B, if we know a theorem of the form $A \rightarrow B$ and if we can prove the lemma $A \rightarrow B$, it suffices to prove A.

5. Introduction to the conjunction:

$$\boxed{\dfrac{\Gamma\vdash A \quad \Gamma\vdash B}{\Gamma\vdash A\wedge B}\wedge_i} \tag{5.1.51}$$

From bottom to top: to prove $A\wedge B$, it suffices to prove A and prove B.

6. Elimination of the conjunction:

$$\boxed{\dfrac{\Gamma\vdash A\wedge B}{\Gamma\vdash A}\wedge_e^g \quad \text{or} \quad \dfrac{\Gamma\vdash A\wedge B}{\Gamma\vdash B}\wedge_e^d} \tag{5.1.52}$$

From top to bottom: from $A\wedge B$, we can deduce A (left elimination) and B (right elimination).

7. Introduction to the disjunction:

$$\boxed{\dfrac{\Gamma\vdash A}{\Gamma\vdash A\vee B}\vee_i^g \quad \text{or} \quad \dfrac{\Gamma\vdash B}{\Gamma\vdash A\vee B}\vee_i^d} \tag{5.1.53}$$

From bottom to top: to prove $A\vee B$, it suffices to prove A (left disjunction) or prove B (right disjunction).

8. Elimination of the disjunction:

$$\boxed{\dfrac{\Gamma\vdash A\vee B \quad \Gamma, A\vdash C \quad \Gamma, B\vdash C}{\Gamma\vdash C}\vee_e} \tag{5.1.54}$$

From bottom to top: if we want to prove C and we know we have $A\vdash B$, it is enough to prove in first time by assuming A, and in a second time by assuming B. This is a case-based reasoning.

9. Introduction of the negation:

$$\boxed{\dfrac{\Gamma, A\vdash \bot}{\Gamma\vdash \neg A}\neg_i} \tag{5.1.55}$$

From bottom to top: to show $\neg A$, we assume A and we prove the absurd ($\bot$).

10. Elimination of the negation:

$$\boxed{\dfrac{\Gamma\vdash \neg A \quad \Gamma\vdash A}{\Gamma\vdash \bot}\neg_e} \tag{5.1.56}$$

From top to bottom: if we proved $\neg A$ and A, then we proved by absurdity ($\bot$).

11. Classic absurdity (reductio ad absurdum):

$$\boxed{\dfrac{\Gamma, \neg A\vdash \bot}{\Gamma\vdash A}\bot_c} \tag{5.1.57}$$

From bottom to top: to prove A, it suffices to prove the absurdity by assuming $\neg A$.

This rule is equivalent to say: A is true if and only if it is false that A is false. This rule is not obvious: it is necessary to prove some results (there are results we can not prove if we do not have this rule). Contrary to many others, this rule may also be applied at any time. We can, in fact, always say: To prove A, I suppose $\neg A$ and I will seek for a contradiction.

12. Introduction of the universal quantifier:

$$\boxed{\dfrac{\Gamma \vdash A \; x \text{ is not free in the formulas of } \Gamma}{\vdash \forall x A}\forall_i} \tag{5.1.58}$$

From bottom to top: to prove $\forall x\ A$, it suffices to show A doing no assumption about x.

Remark

For proofs this check (no assumption on x) is often a source of error.

13. Elimination of the universal quantifier:

$$\boxed{\dfrac{\Gamma \vdash \forall x A}{\Gamma \vdash A[x := t]}\forall_e} \tag{5.1.59}$$

From top to bottom: from $\forall x\ A$, we can deduce $A[x : 0t]$ for any term t. What we can also say under the form: if we proved A for all x, then we can use A with any object t (!!).

14. Introduction of the existential quantifier:

$$\boxed{\dfrac{\Gamma \vdash A[x := t]}{\Gamma \vdash \exists x A}\exists_i} \tag{5.1.60}$$

From bottom to top: to prove $\exists x\ A$, it suffices to find an object (verbatim a term t) for which we know how to prove $A[x := t]$.

15. Elimination of the existential quantifier:

$$\boxed{\dfrac{\Gamma \vdash \exists x A \; \Gamma, A \vdash C \; x \text{ is not free in the formulas of } \Gamma, \text{ neither in } C}{\Gamma \vdash C}\exists_e} \tag{5.1.61}$$

From bottom to top: we prove that there is indeed a set of assumptions such that $\exists x\ A$ and hence this result as new hypothesis, we prove C. This formula C inherits then from the formula $\exists x\ A$ and therefore x is not free in C because it already was not in $\exists x\ A$.

16. Introduction of equality:

$$\boxed{\dfrac{}{\Gamma \vdash t = t} =_i} \tag{5.1.62}$$

From bottom to top: we can always prove $t = t$. This rule means that equality is reflexive (see section Operators page 290).

17. Elimination of equality:

$$\boxed{\dfrac{\Gamma \vdash A[x := t] \; \Gamma \vdash t = u}{\Gamma \vdash A[x := u]} =_e} \tag{5.1.63}$$

From top to bottom: if we prove that $\Gamma \vdash A[x := t]$ and $t = u$, then we have proved $A[x := u]$. This rule expresses that equal objects have the same properties. We notice however that the formulas (or relations) $t = u$ and $u = t$ are not, formally, identical. We will have to prove that equality is symmetric (we will benefit also to prove on the way that equality is transitive).

Let us see now three examples by introducing them in the form of theorems as it should be in proof theory!

Theorem 5.1. The equality is symmetric (a little bit not trivial but quite good to begin with the subject):

Proof 5.1.1.

$$\dfrac{\dfrac{\dfrac{\dfrac{}{x_1=x_2\vdash x_1=x_1}=_i \quad \dfrac{}{x_1=x_2\vdash x_1=x_2}\text{ax}}{x_1=x_2\vdash x_2=x_1}=_e}{\vdash x_1=x_2\rightarrow x_2=x_1}\rightarrow_i}{\vdash \forall x_1,x_2(x_1=x_2\rightarrow x_2=x_1)}\forall_i\times 2 \tag{5.1.64}$$

From top to bottom: we introduce the equality $=_i$ and prove from the assumption $x_1=x_2$ the formula $x_1=x_1$. At the same time, we define the axiom as what $x_1=x_2$. Then from these premises, we eliminate the equality $=_e$ by substituting the terms so that from the assumption $x_1=x_2$ (from the axiom) we get $x_2=x_1$. Then, the elimination of equality automatically implies without assumption that $x_1=x_2\rightarrow x_2=x-1$. Therefore, we simply insert the universal quantifier for each variable (i.e. twice) without any assumption to achieve that equality is symmetric.

■ Q.E.D.

Theorem 5.2. The equality is transitive (that is to say if $x_1=x_2$ and $x_2=x_3$ then $x_1=x_3$). By denoting F the formula $(x_1=x_2)\wedge(x_2=x_3)$:

Proof 5.2.1.

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{}{F\vdash x_1=x_2\wedge x_2=x_3}\text{ax}}{F\vdash x_1=x_2}\wedge_e^g \quad \dfrac{\dfrac{}{F\vdash x_1=x_2\wedge x_2=x_3}\text{ax}}{F\vdash x_2=x_3}\wedge_e^d}{F\vdash x_1=x_3}=_e}{\vdash(x_1=x_2\wedge x_2=x_3)\rightarrow x_1=x_3}\rightarrow_i}{\forall x_1,x_2,x_3((x_1=x_2\wedge x_2=x_3)\rightarrow x_1=x_3)}\forall_i\times 3 \tag{5.1.65}$$

What do we do here? We first introduce the formula F twice as axiom to "dissect" it latter left and right (we do not introduce the equality supposed already introduced as a rule). Once done, we eliminate on the left and on the right the conjunction on the formula to work on left and right terms only and we introduce the equality of the two terms which makes that from the formula we have the transitive equality. It follows without any assumption that automatically implies that equality is transitive and finally we say that this is valid for any value of the different variables (if the formula is true, then equality is transitive).

■ Q.E.D.

And now last big example always in the form of a theorem:

Theorem 5.3. Any involution is a bijection (see section Set Theory page 407).

Proof 5.3.1. Let f be a unary function symbol (with one variable), we write (for the details see the section of Set Theory page 403):

- Inj$[f]$ the formula:

  $$\forall x, y: \quad f(x)=f(y) \rightarrow x=y \tag{5.1.66}$$

  which means that f is injective.

- Surj$[f]$ the formula:

  $$\forall y, \exists x: \quad f(x)=y \tag{5.1.67}$$

- Bij$[f]$ the formula:

  $$\text{Inj}[f] \wedge \text{Surj}[f] \tag{5.1.68}$$

- Inv$[f]$ the formula:

  $$\forall x: \quad f(f(x))=x \tag{5.1.69}$$

which means that f is an involution (we also write this $f \circ f = \text{Id}$ that this is to say that the composition of f is the identity).

We would like to know if:

$$\text{Inv}[f] \vdash \text{Bij}[f] \tag{5.1.70}$$

We will present (trying this to be done as easy as possible) this proof in four (!!!) different ways: traditional (informal), classic (pseudo-formal), formal in tree and formal in-line.

- **Traditional method (informal):**

  We must prove that if f is involutive then it is bijective. So we have two things to prove (and both must be satisfied simultaneously): the function is injective and surjective.

  1. So we prove first that involution is injective.

     We assume for this, because f is an involution it therefore injective, such that:

     $$\forall x, y: \quad f(x)=f(y) \tag{5.1.71}$$

     implies that:

     $$x=y \tag{5.1.72}$$

     However, this assumption automatically comes from the definition of involution that:

     $$\forall x: f(f(x))=x, \forall y: f(f(y))=y \tag{5.1.73}$$

     and to the application of f to the relation:

     $$f(x)=f(y) \tag{5.1.74}$$

     (thus three equalities so far) such that:

     $$f(f(x))=f(f(y)) \tag{5.1.75}$$

     we therefore have:

     $$x=y \tag{5.1.76}$$

2. Let us prove that involution is surjective.
   If it is surjective, then we must have:

   $$\forall y \exists x : \quad f(x)=y \tag{5.1.77}$$

   But, let us define the variable x by definition of the involution itself:

   $$x := f(y) \tag{5.1.78}$$

   as $y=f(x)$..., and a change of variable after we get:

   $$f(f(y))=y \tag{5.1.79}$$

   and therefore surjectivity is ensured.

- **Pseudo-formal method:**

  We take again the same and we inject in it the rules of proof theory:

  We must show that f is involutive and therefore bijective. So we have two things to show $(\wedge_i)$ (and both must be satisfied simultaneously): that the function injective and surjective:

  $$\text{Inj}[f] \wedge \text{Surj}[f] \tag{5.1.80}$$

1. Let us first prove that involution is injective. We assume for this, since f is involutive and therefore injective, that:

   $$(\forall i) \quad \forall x,y : \quad f(x)=f(y) \tag{5.1.81}$$

   implies:

   $$(\rightarrow_i) \quad x=y \tag{5.1.82}$$

   However, this assumption automatically comes from the definition of involution that:

   $$(\forall_e, \text{ax}) \quad \forall x\ f(f(x))=x, \forall y\ f(f(y))=y \tag{5.1.83}$$

   and from the application of f to the relation:

   $$f(x)=f(y) \tag{5.1.84}$$

   (therefore three equalities $=_e\ \times 2$) such that:

   $$f(f(x))=f(f(y)) \tag{5.1.85}$$

   We therefore have:

   $$x=y \tag{5.1.86}$$

2. Let us prove that involution is surjective. If it is surjective, then we must have:

   $$\forall y \exists x : \quad f(x)=y \quad (\forall i) \tag{5.1.87}$$

Now, we define the variable x by definition of the involution itself:

$$(\exists_i)\quad x=f(y)\tag{5.1.88}$$

since $y=f(x)$....., after a change of variables we get:

$$f(f(y))=y\tag{5.1.89}$$

and therefore:

$$(\forall_e,\text{ax})\tag{5.1.90}$$

surjectivity is assured.

- **Formal tree method:**
  Let us do this now with the graphical method that we have presented above.

  1. Let us prove that involution is injective:
     For this we must prove first that:

     $$f(x)=f(y)\vdash f(f(x))=f(f(y))\tag{5.1.91}$$

     Therefore:

     $$\dfrac{\dfrac{\dfrac{}{\text{Inv}[f]\vdash\forall y(f(f(y))=y)}\,\text{ax}}{\text{Inv}[f]\vdash\{f(y)=x\}[y:=f(x)]}\,\forall_e\qquad\dfrac{}{f(x)=f(y)\vdash f(x)=f(y)}\,\text{ax}}{f(x)=f(y)\vdash\{f(x)=y\}[x:=f(y)]}\,=_e\ \textbf{(a)}\tag{5.1.92}$$

     $$\dfrac{\dfrac{\dfrac{}{\text{Inv}[f]\vdash\forall x(f(f(x))=x)}\,\text{ax}}{\text{Inv}[f]\vdash\{f(x)=y\}[x:=f(y)]}\,\forall_e\qquad\dfrac{}{f(x)=f(y)\vdash f(x)=f(y)}\,\text{ax}}{f(x)=f(y)\vdash\{f(y)=x\}[y:=f(x)]}\,=_e\ \textbf{(b)}\tag{5.1.93}$$

     That bring us to write:

     $$\dfrac{\textbf{(a)(b)}}{f(x)=f(y)\vdash f(f(x))=f(f(y))}\,=_e\ \textbf{(c)}\tag{5.1.94}$$

     > **Remark**
     >
     > The latter relation is abbreviated $=_c$ and named (as other existing) "derived rule" because it is an argument that is often made during proofs and a little time consuming to develop each time ...

     Therefore:

     $$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{}{\text{Inv}[f]\rightarrow\text{Inv}[f]}\,\text{ax}}{\text{Inv}[f]\vdash f(f(x))=x}\,\forall_e\qquad\dfrac{\dfrac{}{\text{Inv}[f]\rightarrow\text{Inv}[f]}\,\text{ax}}{\text{Inv}[f]\vdash f(f(y))=y}\,\forall_e}{f(f(x))=f(f(y))\vdash x=y}\,=_e\ \textbf{(c)}}{\text{Inv}[f],f(x)=f(y)\vdash x=y}\,=_e\ \textbf{(d)}}{\text{Inv}[f]\vdash f(x)=f(y)\rightarrow x=y}\,\rightarrow_i}{\text{Inv}[f]\vdash\text{Inv}[f]}\,\forall_i\tag{5.1.95}$$

2. Let us prove now that involution is surjective:

$$\cfrac{\cfrac{\cfrac{\cfrac{}{\text{Inv}[f]\vdash\forall x\{f(f(x))=x\}}\,\text{ax}}{\text{Inv}[f]\vdash\{f(x)=y\}[x:=f(y)]}\,\forall_e}{\text{Inv}[f]\vdash\exists x\{f(x)=y\}}\,\exists_i}{\text{Inv}[f]\vdash\text{Surj}[f]}\,\forall_i\ \textbf{(e)} \tag{5.1.96}$$

It follows:

$$\cfrac{\textbf{(d)(e)}}{\text{Inv}[f]\vdash\text{Inj}[f]\wedge\text{Surj}[f]}\,\wedge_i \tag{5.1.97}$$

- **Formal in-line method:**

  We can do the same thing in a slightly less... wide form ... and more ... tabbed (it is no less indigestible):

$$\begin{array}{ll}
\text{Inv}[f]\vdash\text{Bij}[f] & \vee\, i\\
\quad 1:=\text{Inv}[f]\vdash\text{Inj}[f] & \\
\quad 2:=\text{Inv}[f]\vdash\text{Surj}[f] &
\end{array} \tag{5.1.98a}$$

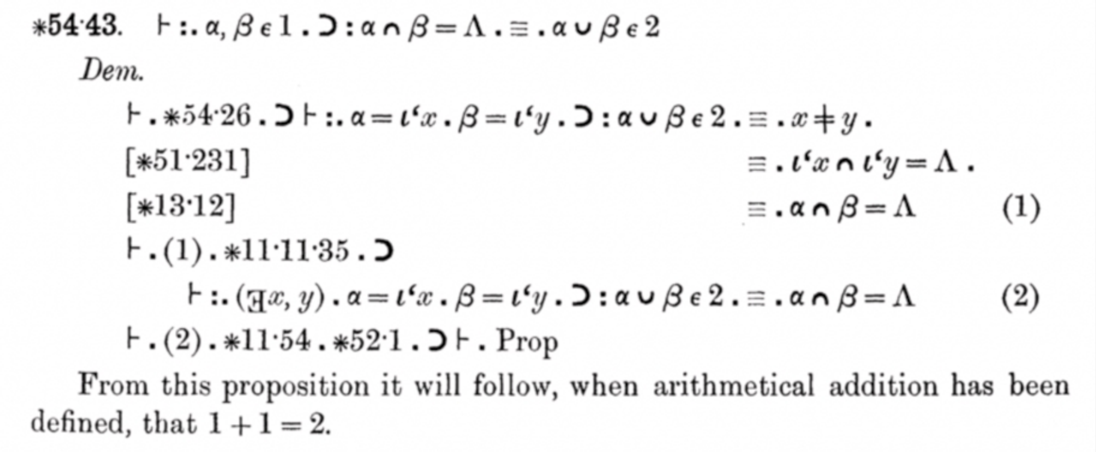
$$\begin{array}{ll}
(1)\text{Inv}[f]\vdash\text{Inj}[f] & \forall i\\
\quad \text{Inv}[f]\vdash f(x)=f(y)\rightarrow x=y & \rightarrow_i\\
\quad \text{Inv}[f],f(x)=f(y)\vdash x=y & =_e\ \times 2(i)(ii)(iii)\\
\quad (i)\quad \text{Inv}[f]\vdash f(f(x))=x & \forall_e\\
\qquad\quad \text{Inv}[f]\vdash\text{Inv}[f] & \text{ax}\\
\quad (ii)\quad \text{Inv}[f]\vdash f(f(y))=y & \forall_e\\
\qquad\quad \text{Inv}[f]\vdash\text{Inv}[f] & \text{ax}\\
\quad (iii)\quad f(x)=f(y)\vdash f(f(x))=f(f(y)) & =_c\\
\qquad (1')f(x)=f(y)\vdash f(x)=f(y) & \text{ax}
\end{array} \tag{5.1.98b}$$

$$\begin{array}{ll}
(2)\text{Inv}[f]\vdash\text{Surj}[f] & \forall i\\
\quad \text{Inv}[f]\vdash\exists x\{f(x)=y\} & \exists_i\\
\quad \text{Inv}[f]\vdash\{f(x)=y\}[x:=f(y)] & \forall_e\\
\quad \text{Inv}[f]\vdash\forall x\{f(f(x))=x\} & \text{ax}
\end{array} \tag{5.1.98c}$$

■ Q.E.D.

Remark

The reader interested in going further in the topic can read the three famous volumes of *Principia Mathematica* of Whitehead and Russell [571] (considered as the "bible" in Proof Theory by many specialists) that have an approximate total of 1900 pages where the first 380 pages of the first volume are dedicated to the proof that $1+1=2$...



✱54·43. $\vdash :. \alpha, \beta \in 1 . \supset : \alpha \cap \beta = \Lambda . \equiv . \alpha \cup \beta \in 2$

Dem.

$\vdash . *54\cdot26 . \supset \vdash :. \alpha = \iota'x . \beta = \iota'y . \supset : \alpha \cup \beta \in 2 . \equiv . x \neq y .$
[✱51·231] $\equiv . \iota'x \cap \iota'y = \Lambda .$
[✱13·12] $\equiv . \alpha \cap \beta = \Lambda$ (1)
$\vdash . (1) . *11\cdot11\cdot35 . \supset$
$\vdash :. (\exists x, y) . \alpha = \iota'x . \beta = \iota'y . \supset : \alpha \cup \beta \in 2 . \equiv . \alpha \cap \beta = \Lambda$ (2)
$\vdash . (2) . *11\cdot54 . *52\cdot1 . \supset \vdash .$ Prop

From this proposition it will follow, when arithmetical addition has been defined, that $1 + 1 = 2$.

Obviously keep in mind that in binary we have...:

$$1 + 1 = 10$$

or in unary (base 1):

$$1 + 1 = 11$$

We can also define a group of two elements $\{0, 1\}$ therefore the addition would give:

$$1 + 1 = 0$$

However... all this highly technical formalism don't make it always obvious to find where is the error in the following "pseudo-proof":

Assume:

$$a = b \tag{5.1.99}$$

Let us multiply both side by a:

$$a^2 = ab \tag{5.1.100}$$

Let us subtract by b^2:

$$a^2 - b^2 = ab - b^2 \tag{5.1.101}$$

We factorize:

$$(a + b)(a - b) = b(a - b) \tag{5.1.102}$$

We divide by $(a - b)$:

$$a + b = b \tag{5.1.103}$$

Since $a = b$, we have:

$$b + b = b \tag{5.1.104}$$

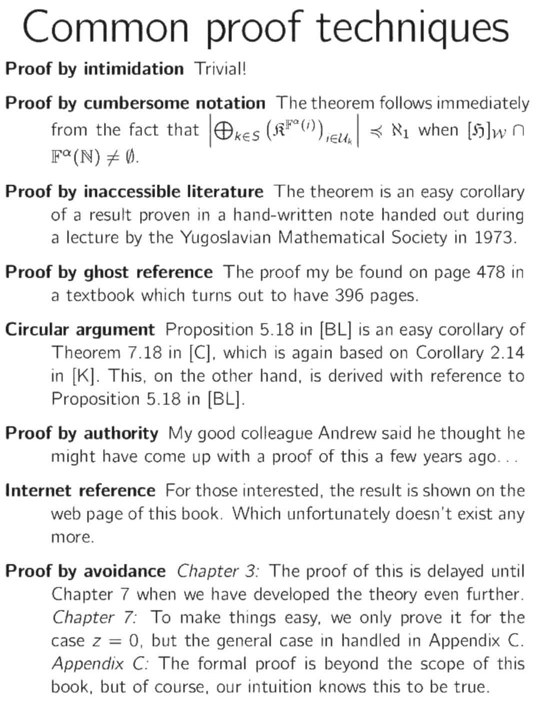
Therefore:

$$2 = 1 \tag{5.1.105}$$

If you don't see the error, here is the analysis:

Written	Reality	Judgement
$a=b$	$a=a$	True
$a^2=ab$	$a^2=aa$	True
$a^2-b^2=ab-b^2$	$a^2-a^2=aa-a^2$	True $(=0)$
$(a+b)(a-b)=b(a-b)$	$(a+a)(a-a)=a(a-a)$	True $(=0)$
$a+b=b$	$a+a=a$	False (unless $a=0$) and forbidden as dividing by $a-b$ is equivalent as dividing by 0
$b+b=b$	$a+a=a$	False
$2b=b$	$2a=a$	False
$2=1$	$2=1$	False

Obviously there are a lot of other kind of proofs as illustrated in the figure below... ☺:



Common proof techniques

Proof by intimidation Trivial!

Proof by cumbersome notation The theorem follows immediately from the fact that $\left|\bigoplus_{k\in S}\left(\mathfrak{K}^{\mathbb{F}^{\alpha}(i)}\right)_{i\in\mathcal{U}_k}\right| \preccurlyeq \aleph_1$ when $[\mathfrak{H}]_{\mathcal{W}} \cap \mathbb{F}^{\alpha}(\mathbb{N}) \neq \emptyset$.

Proof by inaccessible literature The theorem is an easy corollary of a result proven in a hand-written note handed out during a lecture by the Yugoslavian Mathematical Society in 1973.

Proof by ghost reference The proof my be found on page 478 in a textbook which turns out to have 396 pages.

Circular argument Proposition 5.18 in [BL] is an easy corollary of Theorem 7.18 in [C], which is again based on Corollary 2.14 in [K]. This, on the other hand, is derived with reference to Proposition 5.18 in [BL].

Proof by authority My good colleague Andrew said he thought he might have come up with a proof of this a few years ago...

Internet reference For those interested, the result is shown on the web page of this book. Which unfortunately doesn't exist any more.

Proof by avoidance *Chapter 3:* The proof of this is delayed until Chapter 7 when we have developed the theory even further. *Chapter 7:* To make things easy, we only prove it for the case $z=0$, but the general case in handled in Appendix C. *Appendix C:* The formal proof is beyond the scope of this book, but of course, our intuition knows this to be true.

5.1.6 Laws of thought

The laws of thought are fundamental axiomatic rules upon which rational (or philosophical) discourse itself is often considered to be based. The formulation and clarification of such rules have a long tradition in the history of philosophy and logic. Generally they are taken as laws that guide and underlie everyone's thinking, thoughts, expressions, discussions, etc. However, such classical ideas are often questioned or rejected in more recent developments such as intuitionist logic, dialetheism and fuzzy logic.

Definition (#25): The "laws of thought" are axioms by which - or in accordance with which - valid thought proceeds, or that justify valid inference, or to which all valid rhetorical (i.e.

semantic) deductions are reducible and not necessarily connected to evidence based science. Laws of thought are rules that apply without exception to any subject matter of thought, etc. however they can be applied only to sentences, not to objects (!); sometimes they are said to be the "object of logic" where the term "logic" is the scientific evidence based study of the formal truth.

Caution! As already mentioned, as any axiomatic system, the laws of thought are submitted to the both Gödel's incompleteness theorems (proven in year 11931 - holocene calendar) that are for recall among the most important results in modern logic, and have deep implications! They concern the limits of provability in formal axiomatic theories. The first **incompleteness** theorem states that in any consistent formal system F within which a certain amount of arithmetic can be carried out, there are statements of the language of F which can neither be proved nor disproved in F. Formally:

$$F \vdash \neg\text{Provable}_F\left(\ulcorner G_F \urcorner\right)$$

According to the second incompleteness theorem, such a formal system cannot prove that the system itself is **consistent** (assuming it is indeed consistent). Formally:

$$F \nvdash \text{Cons} \equiv \neg\text{Provable}_F\left(0 = 0'\right)$$

Furthermore a proof in a given axiomatic system isn't necessarily true in another one!

In other words Gödel's theorems does not disprove anything in mathematics. Rather, it proves that certain truths in a given axiomatic framework exist which cannot be proved using currently accepted framework or can be disproved in another axiomatic framework (or by experimental evidence).

Or again said otherwise (because some people have difficulties to understand), the theorem doesn't undermine logic and doesn't destroy the fabric of thought. It just says that certain axiomatic systems, satisfying various conditions, aren't complete, which means that certain statements can neither be proved nor disproved by those systems (but some can!). Those very same "undecidable" statements may well be proved or disproved in other axiomatic systems or by evidence based science.

quote
5.1:188.#34

> *Many logicians have been driven to the conclusion that there are unreal objects. [...] we can make true propositions of which these are the subjects; hence they must have some kind of logical being, since otherwise the propositions in which they occur would be meaningless. [...] Logic, I should maintain, must no more admit a unicorn than zoology can; for logic is concerned with the real world just as truly as zoology [...]*
>
> Bertrand Russell

Surprised? People usually come to hold, through their lack of any formal graduate or postgraduate training in logic, that logic is far more than it really is. Therefore, a further understanding of just what logic is, can be enhanced by delineating it from what it is not:

- Logic is not the study of what, if anything, comprises the "groundness of being for the universe" - That's metaphysics!
- Logic is not a set of laws that governs the universe - That's physics!
- Logic is not an immaterial "entity" that transcends reality - Such discussions belong to the realm of theology!
- Logic is not a method for "studying the world" - That's evidence based science!
- Logic is not the method for assessing axioms - That's a matter of pure reason!
- Logic is not a way of evaluating "truth" - That's philosophy!
- Logic is not a set of laws that governs human behavior - That's psychology!
- Logic doesn't applied to all words and sentences (try for example with expressions like "*Close the door!*" or "*What time is it?*") - That's Language theory!
- Logic is not even a study of how people reason - Fortunately there is more to human reason than just logic...

Remark

The reader muser also remember the Münchhausen trilemma (see page 3128) that is a thought experiment to demonstrate the theoretical impossibility of proving any truth, even in the fields of logic and mathematics, without appealing to accepted assumptions like: 1) The circular argument, in which the proof of some proposition presupposes the truth of that very proposition 2) The regressive argument, in which each proof requires a further proof, ad infinitum 3) The dogmatic argument, which rests on accepted precepts which are merely asserted rather than defended.

These warnings (reminders) having been made, let us return to the laws of thought (which fall within the framework of logical "tautologies") in the subjective framework of Aristotelian logic:

A1. **The law of identity (ID):** everything is (i.e., is identical to) itself, or Whatever is, is (*Metaphysics IV (Gamma)*, 3, Aristotle, year 9650 according to holocene calendar):

$$\forall A: \quad A = A \tag{5.1.106}$$

A2. **The law of contradiction (or non-contradiction; NC):** no thing having a given quality also has the negative of that quality (e.g., no even number is non-even), or Nothing can both be and not be (*Metaphysics IV (Gamma)*, 3, Aristotle, year 9650 according to holocene calendar):

$$\neg(A \wedge \neg A) \tag{5.1.107}$$

Remark

Lukasiewicz three valued logic is an example of a logical system in which the classical laws of non-contradiction and the excluded middle (see below) are explicitly rejected as general laws applicable to all propositions.

There are objects (i.e. sentences) that violate the non-contradiction law and named "dialetheias". Like the sentence (object): *This statement is false* known as the Epimenedes paradox. If we assume it is true, then it must be false (since it claims to be false). But if we assume it is false, then it must be true (since it claims to be false). It is then an undecidable statement (undecidable statements generally rely on a hidden self-reference).

A3. **The law of excluded middle (EM):** every thing either has a given quality or has the negative of that quality (e.g., every number is either even or non-even), or Everything must either be or not be (*Metaphysics IV (Gamma)*, 6, Aristotle, year 9650 according to holocene calendar):

$$A \vee \neg A \tag{5.1.108}$$

While the law of non-contradiction tells us that no statement can be both true and false, the law of excluded middle tell us that they must all be one or the other.

Remark

The law of excluded middle is removed in "intuitionistic logic" (also in "fuzzy logic", see page 7136), sometimes more generally named "constructive logic", as at the opposite of "classical logic" it does not focus so much on whether statements are "really" true or false, in that sense - it focuses on whether we know a statement is true, or know it is false! So, before someone working in constructive mathematics can assert a statement, he needs to have a concrete justification for the truth of the statement. Arguments such as proof by contradiction are not permitted!

One kind of example of how excluded middle fails in this framework was named a "weak counterexample" by L. E. J. Brouwer (regarded as one of the greatest mathematicians of the 120th century according to holocene calendar). Here is one: Consider the number π. As a fact, even in classical mathematics we do not currently know whether there are infinitely many occurrences of the digit 5 in the decimal expansion of π. Let A be the statement «*there are infinitely many occurrences of* 5 *in the decimal expansion of* π ». Then someone working in constructive mathematics cannot assert «A *or not* A», because she does not know which of the two options holds. Of course «A *or not* A» is an instance of the law of the excluded middle, and is valid in classical mathematics.

In indeterministic physics the future is open. Consequently, statements and propositions about the future need not be either true or false. For example, the proposition "it will rain in exactly one year time from now at Piccadilly Circus" is neither True (because it is not predetermined that it will rain), nor is it False (because it is not predetermined that it will not rain). Hence, in a world with an open future, the law of the excluded middle doesn't hold!

More technically the three axioms (laws) above can be denoted:

$$\begin{gathered}\forall a(a \equiv \forall P\ P(a)) \\ \forall a \neg \exists P\ (P(a) \wedge \neg P(a)) \\ \forall a \forall P\ (P(a) \vee \neg P(a))\end{gathered} \tag{5.1.109}$$

Some people add a fourth law (axiom):

A4. **Sufficient Reason (SR):** Truth is the reference of a judgement to something outside it as its sufficient reason or ground. Of everything that is, it can be found why it is:

$$\forall a \exists b\ (b \rightarrow a) \tag{5.1.110}$$

Therefore the four axioms of thought (law) can be written as:

$$\boxed{\begin{gathered}\forall a(a \equiv \forall P\ P(a)) \\ \forall a \neg \exists P\ (P(a) \wedge \neg P(a)) \\ \forall a \forall P\ (P(a) \vee \neg P(a)) \\ \forall a \exists b\ (b \rightarrow a)\end{gathered}} \tag{5.1.111}$$

The expression "laws of thought" gained added prominence through its use by George Boole (11815–11864) to denote theorems of his "algebra of logic"; in fact, he named his second logic book *An Investigation of the Laws of Thought on Which are Founded the Mathematical Theories of Logic and Probabilities* (11854). Modern logicians, in almost unanimous disagreement with Boole, take this expression to be a misnomer; none of the above propositions classed under "laws of thought" are explicitly about thought per se, a mental phenomenon studied by psychology, nor do they involve explicit reference to a thinker or knower as would be the case in pragmatics or in epistemology. The distinction between psychology (as a study of mental phenomena) and logic (as a study of valid inference) is widely accepted.

The reader must remember, as we have already mentioned it earlier (!), that logic isn't connected to evidence based science but only to sentences (rhetorical art, i.e. semantic) as already stated earlier in the definition! For example some people claim that the law of non-contradiction is violated by:

- The photon has the property of being at the same time a *particle* and it's own opposite[3] (anti-particle[4])
- There is a non-zero probability of a particle to be at multiple positions at the same time before its measurement (collapse of the wave function)
- The Klein bottle that has it's external surface that is at the same time its own opposite surface (the Möbius strip has the same property as the Klein bottle)

---

[3] Consider an annihilation reaction of some massive particle-antiparticle pair. As we know there is a good probability that this will produce a pair of photons. The inverse reaction will destroy the photons and produce a massive particle antiparticle pair. This proves that a photon with spin $+S$ is the antiparticle of a photon with spin $-S$ and the same frequency.

[4] Half-integer spin particles have well-defined antiparticles, integer spin particles exhibit this fascinating self-duality.

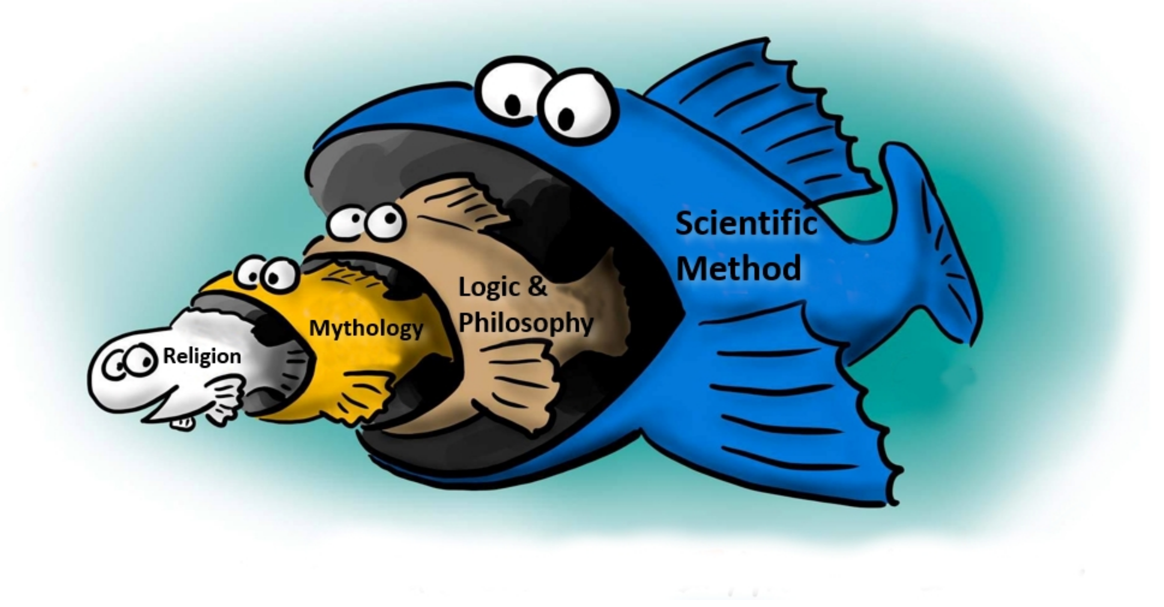
- The sphere that is endless and boundless is at the same time not infinite in size (hence its finite and infinite at the same time)
- The segment that has bounds and finite size contains at the same time infinitely many points (hence its finite and infinite at the same time)
- The Gabriel's horn that has infinite surface area but at the same time a finite volume
- Many mathematical series have infinite terms but at the same time finite value
- The number 0 is an even number or odd number or both at the same?
- The circle is at the same time a square under a given topological transformation in advanced mathematics (for the same reason a cup of coffee is topologically a torus)
- At the quantum level, cause may occur after effect (retrocausality)
- In the Copenhagen interpretation of Quantum Physics, Schrödinger's cat is both dead and alive at the same time (which it is not in the multiverse interpretation of quantum physics).

We see with these toy examples above that the three axioms of classical logic (i.e. laws of thoughts) are not always enough to have a rhetorical debate on the real world as modern science knows it!

Remark

The problem with Classical Logic is that some logicians claim that something that is "opposite" in real life does not make the difference logically necessary........!

However, none of the examples violates the law of non-contradiction! The trivial reason is that the law of non-contradiction refers to the truth value of propositions (sentences) but not to objects or entities !



Remark

Garrett Birkhoff and John Von Neuman proposed in the 11930s (holocene calendar) that the paradoxes of Quantum Mechanics can be explained if we abandoned Classical Logic and used some form of Quantum Logic instead (Birkhoff, Garrett; von Neumann, John. "*The Logic of Quantum Mechanics*". Ann. Math. 37 (4): 823–843). Such a Quantum Logic would change or abandon all together some of the rules of Classical Logic, and would be a perfect case of logical axioms arrived by observation.

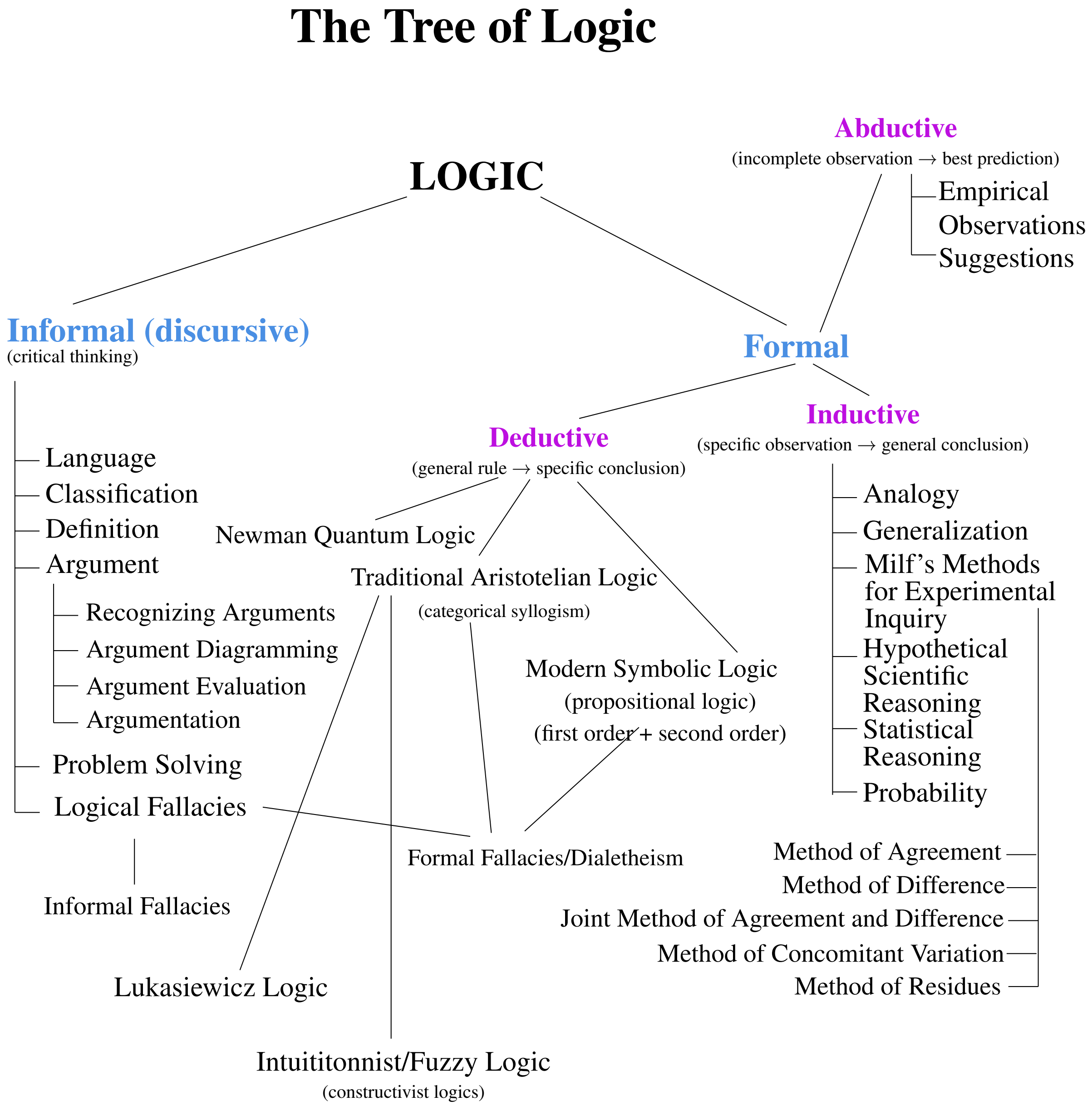

The Tree of Logic
LOGIC
Abductive
(incomplete observation → best prediction)
Empirical
Observations
Suggestions
Informal (discursive)
(critical thinking)
Formal
Language
Classification
Definition
Argument
Recognizing Arguments
Argument Diagramming
Argument Evaluation
Argumentation
Problem Solving
Logical Fallacies
Informal Fallacies
Deductive
(general rule → specific conclusion)
Newman Quantum Logic
Traditional Aristotelian Logic
(categorical syllogism)
Modern Symbolic Logic
(propositional logic)
(first order + second order)
Formal Fallacies/Dialetheism
Lukasiewicz Logic
Intuititonnist/Fuzzy Logic
(constructivist logics)
Inductive
(specific observation → general conclusion)
Analogy
Generalization
Milf's Methods
for Experimental
Inquiry
Hypothetical
Scientific
Reasoning
Statistical
Reasoning
Probability
Method of Agreement
Method of Difference
Joint Method of Agreement and Difference
Method of Concomitant Variation
Method of Residues

Figure 5.5: **The tree of Logic**

Definitions (#5.1.26):

D1. A system is said to be "consistent" if we cannot prove both a statement and its opposite. We cannot prove something like $1 = 2$. In other words, in a consistent system, everything that is provable is true!

D2. A system is said to be "complete" if every true statement can be proven. Every assertion can be either proved or disproved. In other words, in a complete system, everything true is provable!

D3. A system is said to be "decidable" if for every type of problem, there is a procedure that, in theory, will give a solution. A problem is decidable if a “true” or “false” as an answer can be given rather than “don’t know”, crashing or looping forever in an interminable state of indecision.

Here is a table that summarize some well known result of Gödel’s theorems:

Theory	**Is it Consistent?**	**Is it Complete?**	**Is it Decidable?**
Propositional calculus	Yes	Yes	Yes
Euclidean geometry	Yes	Yes	Yes
1^{st} order logic	Yes	Yes	**No**
Arithmetic $(+,-)$ only	Yes	Yes	Yes
Arithmetic in full $(+,-,\times,\div)$	??	No	No

Table 5.13: **Consistency, Completeness, Decidability of common axiomatic systems**

The fact that first-order logic is undecidable has significant implications for computer science, mathematics, and philosophy! In philosophy:

- The undecidability of first-order logic has implications for the nature of truth and knowledge. It suggests that there are certain truths that are beyond the reach of human reason.
- It also has implications for the nature of logic itself. It suggests that logic is not a complete and perfect system for reasoning about the world.

The table above known thanks to Kurt Gödel is famous for the following corollaries:

T1. Any formal system able to do elementary arithmetic is either inconsistent or incomplete

T2. Any formal system able to express its own consistency can prove its own consistency if and only if it is inconsistent

The reader must remember that just because something is logical does not mean that the conclusion is true in the framework of evidenced based science as logic is a rhetorical (i.e. semantic) axiomatic system disconnected from reality! Especially in logic we can use unsupported premises or assumption or even worst, premises or assumption that are evidently shown wrong by evidence based science!

Gödel's ontological proof (logical modal proof that a deity exist written in 11941 according to holocene calendar) is such a famous example:

Ax. 1. $(P(\varphi) \wedge \Box\forall x(\varphi(x) \Rightarrow \psi(x))) \Rightarrow P(\psi)$
Ax. 2. $P(\neg\varphi) \Leftrightarrow \neg P(\varphi)$
Th. 1. $P(\varphi) \Rightarrow \diamond\exists x\varphi(x)$
Df. 1. $G(x) \Leftrightarrow \forall\varphi(P(\varphi) \Rightarrow \varphi(x))$
Ax. 3. $P(G)$
Th. 2. $\diamond\exists xG(x)$
Df. 2. $\varphi \text{ ess. } x \Leftrightarrow \varphi(x) \wedge \forall\psi(\psi(x) \Rightarrow \Box\forall y(\varphi(y) \Rightarrow \psi(y)))$
Ax. 4. $P(\varphi) \Rightarrow \Box P(\varphi)$
Th. 3. $G(x) \Rightarrow G \text{ ess. } x$
Df. 3. $E(x) \Leftrightarrow \forall\varphi(\varphi \text{ ess. } x \Rightarrow \Box\exists y\varphi(y))$
Ax. 5. $P(E)$
Th. 4. $\Box\exists xG(x)$

In normal language, this gives:

- **Axiom 1:** If φ is a positive property, and if it is necessarily true (true in all possible worlds) that every object with property φ also has property ψ, then ψ is also a positive property.
- **Axiom 2:** The negation of a property φ is positive if, and only if, φ is not positive.
- **Theorem 1:** If a property φ is positive, then it is possible that there exists an object x that has this property (in at least one possible world, there exists an object x that has this property).
- **Definition 1:** An object x is God-like if, and only if, x has all positive properties.
- **Axiom 3:** The property of being God-like is itself a positive property.
- **Theorem 2:** It is possible that there exists a God-like object x (in at least one possible world, there exists a God-like object x).
- **Definition 2:** A property φ is an essential property of an object x if, x has property φ, and every property ψ of x necessarily (in all possible worlds) and generally (for all objects) follows from φ.
- **Axiom 4:** If a property φ is positive, then it is necessarily positive (positive in all possible worlds).
- **Theorem 3:** If x is God-like, then being God-like is an essential property of x.
- **Definition 3:** An object x "exists necessarily" if each of its essential properties φ applies, in each possible world, to some object y.
- **Axiom 5:** "Necessary existence" is a positive property.
- **Theorem 4:** It is necessarily true (true in all possible worlds) that a God-like object exists.

Most criticism of Gödel's proof is aimed at its axioms: as with any proof in any logical system, if the axioms of the proof depends on are doubted, then the conclusions can be doubted! It is particularly applicable to Gödel's proof – because it is based on five axioms which are considered questionable. A proof does not necessitate that the conclusion be correct, but rather that by accepting the axioms, the conclusion follows logically.

And once again, if we need to repeat ourselves..., Gödel's ontological proof above is not evidence, since the empirical scientific method that describes reality is based on inductive and abductive logic. The modal logic on which the above proof is based is disconnected from these two and does not allow us to test and verify reality; this is why empirical scientists reject modal logic as a tool for investigating the World.

We can also prove that most gods doesn't exist using logic (i.e. a syllogism):

- Premise (major premise): Everything that exists has a cause[5]
- Property (minor premise): God is a concept without a cause
- Therefore: God has no cause therefore he does not exist

Or formally:

$$\begin{array}{ll} \text{Ax. 1.} & \exists P \forall x P(x) \\ \text{Df. 1.} & \mathcal{G} \rightarrow \neg P(x) \\ \text{Th. 1.} & \Box \neg P(\mathcal{G}) \rightarrow \nexists \mathcal{G} \end{array}$$

Which is a bit similar in idea to this proof:

- Major premise: Major premise: If God can exist without a cause then something can exist without a cause
- Minor premise: If something exists without a cause then why not the Cosmos[6]
- So therefore God solves nothing given that the existence of a God is a problem as big as that of the existence of the Cosmos

Another well known proof of the non-existence of some gods (typically abrahamic one) is (based on the catechism of the Catholic Church, number 385):

- God's omnipotence implies that he can prevent evil
- The goodness of God implies that God must prevent all evil

---

[5] We known since the emergence of the development of quantum physics and quantum cosmology that this premise is obviously false as we will prove during our study of these two fields of physics later!

[6] Indeed we will prove during our study of Cosmogony that contrary to an urban legend peddled by the mass media and bad science popularizers, our observable universe is without cause and extends in a container that is usually named the "cosmos" in which time does not exist.

- However evil exist
- So a good and all-powerful God cannot exist

The philosopher PhD Michael L. Martin has shown that it is possible to prove that an omnipotent god cannot exist. Here is the procedure (syllogism):

1. (major premise) Logic only makes sense if it establishes principles that are necessarily true (principle of non-contradiction). For example, if $1 + 1 = 2$ is not necessarily true, the result becomes variable and can no longer prove anything logically.
2. (minor premise) If Logic is subservient to a god (created by it and contingent upon it), its principles are no longer necessary, but entirely dependent on that god (principle of excluded middle). Therefore, this god can change the status of logical relations at will, which potentially renders Logic invalid (modus ponens).
3. According to the major premise, if a god exists, then Logic is potentially invalid and cannot allow us to reach certain conclusions. Therefore, Logic can no longer be used to definitively prove the existence of a god... if one exists.
4. According to (3), if we want to use Logic to prove the existence of a god with certainty, we must admit that the god cannot change the laws of logic (monotonicity of implication). Therefore, Logic can only lead to the conclusion that an omnipotent god cannot exist!
5. In conclusion, if Logic is a valid means of proving assertions, then it is logically proven that an omnipotent god cannot exist.

> ❞ *The question of the truth of a religion is one thing, but the question of its usefulness is another. I am as firmly convinced that religions do harm as I am that they are untrue.* ❞
>
> Bertrand Russell

quote 5.1:197.#35

Another famous similar example is the proof of the existence of unicorns...! Indeed, consider we want top prove that unicorns, i.e. horses with one horn, exist. The proof consist rhetorically on the following sequence:

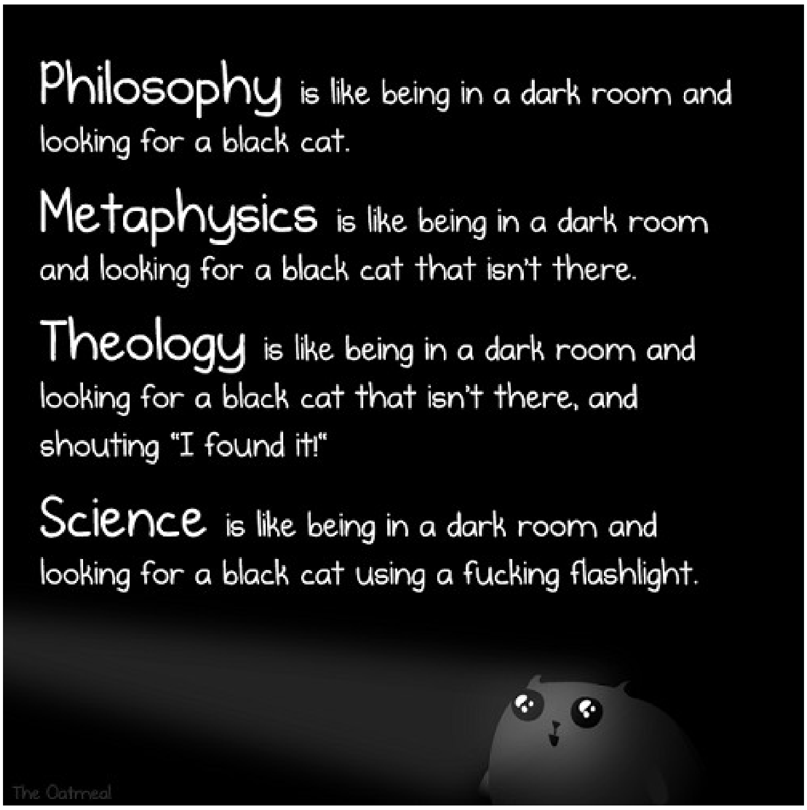
- To prove: Unicorns, i.e. a horses with one horn exist
- (Df. 1.) Unicorns are as we all know horses with one horn!
- (Ax. 1.) Let us suppose the contrary, that unicorns do not exist!
- (Th. 1) Unicorns does not have a horn since, a non-existent entity cannot have anything...
- (Df. 2) But this contradicts the fact that unicorns have one horn by definition

- (Df. 3) This contradiction occurred for wrongly assuming that unicorns do not exist
- (Th. 4) Q.E.D. Hence, unicorns exist!

Formally the above can be rewritten:

Df. 1. $\forall x(\text{Unicorn}(\text{x}) \Rightarrow \text{HasHorn}(\text{x}))$
Ax. 1. $\neg\exists x\, \text{Unicorn}(x)$
Th. 1 $\forall x(\text{Unicorn}(x) \Rightarrow \neg\, \text{HasHorn}(x))$
Df. 2 $(\forall x(\text{Unicorn}(\text{x}) \Rightarrow \text{HasHorn}(\text{x})) \wedge \forall x(\text{Unicorn}(x) \Rightarrow \neg\, \text{HasHorn}(x))) \Rightarrow\perp$
Df. 3 $\neg\exists x\, \text{Unicorn}(x) \Rightarrow\perp$
Th. 4 $\Box\exists x\, \text{Unicorn}(x)$

The error in this example is the claim that a non-existent object cannot have attributes/parts, i.e. $\neq \exists(x) \wedge \text{Predicate}(\text{Q}) \Rightarrow \neg Q(\text{x})$. We cannot assert something about the non-existence of unicorns without evidence based science!



Philosophy is like being in a dark room and looking for a black cat.

Metaphysics is like being in a dark room and looking for a black cat that isn't there.

Theology is like being in a dark room and looking for a black cat that isn't there, and shouting "I found it!"

Science is like being in a dark room and looking for a black cat using a fucking flashlight.

5.1.6.1 Some famous logical fallacies

Definition (#5.1.27): The "principle of explosion", or the "principle of Pseudo-Scotus" (falsely attributed to Duns Scotus), is the law according to which any statement can be proven from a contradiction. That is, from a contradiction, any proposition (including its negation) can be inferred from it; this is known as "deductive explosion". Due to the principle of explosion, the existence of a contradiction (inconsistency) in a formal axiomatic system is disastrous; since any statement can be proven, it trivializes the concepts of truth and falsity...!

In symbolic logic, the principle of explosion can be expressed schematically in the following way:

$$P, \neg P \vdash Q \tag{5.1.112}$$

For any statements P and Q, if P and $\neg P$ are both true, then it logically follows that Q is true.

Proof 5.3.2. Below is a formal proof of the principle using symbolic logic:

Step	Proposition	Derivation
1	P	Premise
2	$\neg P$	Premise
3	$P \vee Q$	Disjunction introduction (1)
4	Q	Disjunctive syllogism $(3, 2)$

■ Q.E.D.

Definition (#5.1.28): The "principle of sufficient reason" (PSR) states that everything must have a reason (explanation) or a cause. The principle was articulated and made prominent by Gottfried Wilhelm Leibniz, with many antecedents, and was further used and developed by Arthur Schopenhauer and Sir William Hamilton, 9th Baronet. It can be stated in various forms that are:

- For every entity X, if X exists, then there is a sufficient explanation for why X exists.
- For every event E, if E occurs, then there is a sufficient explanation for why E occurs.
- For every proposition P, if P is true, then there is a sufficient explanation for why P is true.

In symbolic logic, the principle of sufficient reason can be expressed schematically in the following way:

$$\forall P \exists Q (Q \rightarrow P) \tag{5.1.113}$$

It is an open question whether the principle of sufficient reason can be applied to axioms within a logic construction like a mathematical or a physical theory, because axioms are propositions accepted as having no justification possible within the system. However, the reader must remember that Gödel has shown that for every sufficiently expressive deductive system a proposition exists that can neither be proved nor disproved ! The principle of sufficient reason demands that everything must have a reason. So what is the reason for the principle itself? There isn't one - it's simply stated as an axiom with all what that implies...

The indeterministic interpretations of Quantum Mechanics[7] (Copenhagen and the like) that declare probabilistic predictions complete as to what there is to be known prior to the collapse contradicts the principle of sufficient reason. However other interpretation like the Many World one doesn't.

Remark

The philosopher Robert Nozick also said Quantum Mechanics does violate the principle of sufficient reason. Only a weakened form of the PSR can survive. The mathematician John Conway who wrote on quantum foundations also explicitly stated the original principle of sufficient reason contradicts Quantum Mechanics as such.

Now that we have presented most famous logical fallacies, let us recall all symbols seen so far:

---

[7] For example, the decay of a given radioactive atom has no cause, it is a probabilistic event. According to the Copenhagen interpretation of Quantum Mechanics there are no hidden parameters which give a sufficient cause - not to name a sufficient reason! Furthermore Bell showed that hidden variables weren't possible!

Connective	Symbol	Statement
TRUE value	$\top$	Logical statement is true
FALSE value	$\bot$	Logical statement is false
AND	$\wedge$	Conjunction
OR	$\vee$	Disjunction (inclusive or)
XOR	$\oplus$	Exclusive or
NOT	$\neg, \bar{p}$	Negation
if...then	$\rightarrow, \Rightarrow$	Implication (conditional statement) (necessary condition)
if and only if	$\leftrightarrow, \Leftrightarrow$	Biconditional
equivalent to	$\equiv$	Equivalence
there exists	$\exists$	Existence
there does not exists	$\nexists$	Non-Existence
there is exactly	$\exists!$	Unique Existence
for all	$\forall$	Quantifier
because	$\because$	Because / Since
therefore	$\therefore, \Longrightarrow$	Therefore / Consequence
proves that	$\vdash$	
is not provable	$\nvdash$	
is a tautology	$\models$	Tautology
is a contradiction	$\nvDash$	Contradiction
set of formulas	Γ	
possibly	$\Diamond$	Non-obligatory assumption
end of proof	$\square, \blacksquare$	
code number of the Gödel sentence G	$\ulcorner G \urcorner$	

Table 5.14: **Common symbols in Classical Logic and Proof Theory**

Remark

The logical equivalence $\equiv$ is by tradition something that goes outside of a logical statement (between two of them in particular). If and only if $\leftrightarrow, \Leftrightarrow$ is a symbol that belongs inside a logical statement.

To summarize, below we give a table with most common tautologies, inferences rules and fallacies:

PROPOSITIONAL LOGIC RULES AND FALLACIES

Name	Formula	Description
Tautology	$\varphi = p(x) \therefore A \Rightarrow A$ (i.e. $\models A$)	assertion that A is true in every possible interpretation
Law of Identity	$\forall A : A = A$	for all A, A is what it is
Law of Non-Contradiction	$\neg(A \wedge \neg A)$	A cannot not be and be at the same type its opposite
Law of excluded middle	$A \vee \neg A$	every thing has a given quality or has the negative of that quality
Modus Ponens	$[(p \rightarrow q) \wedge p] \Rightarrow q$	if p then q; p; therefore q
Modus Tollens	$[(p \rightarrow q) \wedge \neg q] \Rightarrow p$	if p then q; not q; therefore not p
Hypothetical Syllogism	$[(p \rightarrow q) \wedge (q \rightarrow r)] \Rightarrow (p \rightarrow r)$	if p then q; if q then r; therefore, if p then r
Disjunctive Syllogism	$(p \vee q; \neg p) \therefore q$	either p or q; not p; therefore, q
Constructive Dilemma	$(p \rightarrow q) \wedge (r \rightarrow s); p \vee r \therefore q \vee s$	if p then q; and if r then s; but either p or r; therefore either q or p
Destructive Dilemma	$(p \Rightarrow q) \wedge (r \Rightarrow s); \neg q \vee \neg s \therefore \neg p \vee \neg r$	if p then q; and if r then s; but either not q or not s; therefore rather not p or not r
Simplification	$[(p = \top) \wedge (q = \top)] \therefore p = \top$	p and q are true; therefore p is true
Conjunction	$(p = \top, q = \top) \therefore p \wedge q$	p and q are true separately; therefore they are true conjointly
Addition	$p = \top \therefore p \vee q$	p is true; therefore the disjunction (p or q) is true
Composition	$[(p \rightarrow q) \wedge (p \rightarrow r)] \Rightarrow [p = \top \rightarrow (q \wedge r) = \top]$	if p then q; and if p then r; therefore if p is true then q and r are true
De Morgan’s Theorems	$\neg(p \wedge q) \equiv (\neg p \vee \neg q)$ $\neg(p \vee q) \equiv (\neg p \wedge \neg q)$	the negation of (p and q) is equiv. to (not p or not q) the negation of (p or q) is equiv. to (not p and not q)
Commutation	$(p \vee q) \equiv (q \vee p)$ $(p \wedge q) \equiv (q \wedge p)$	(p or q) is equiv. to (q or p) (p and q) is equiv. to (q and p)
Association	$[p \vee (q \vee r)] \equiv [(p \vee q) \vee r]$ $[p \wedge (q \wedge r)] \equiv [(p \wedge q) \wedge r]$	p or (q or r) is equiv. to (p or q) or r p and (q and r) is equiv. to (p and q) and r
Distribution	$[p \wedge (q \vee r)] \equiv [(p \wedge q) \vee (p \wedge r)]$ $[p \vee (q \wedge r)] \equiv [(p \vee q) \wedge (p \vee r)]$	p and (q or r) is equiv. to (p and q) or (p and (p or r) p or (q and r) is equiv. to (p or q) and (p or r)
Double Negation	$p \equiv \neg\neg p$	p is equivalent to the negation of not p
Transposition	$(p \rightarrow q) \equiv (\neg q \rightarrow \neg p)$	if p then q is equiv. to if not q then not p
Material Implication	$(p \rightarrow q) \equiv (\neg p \vee q)$	if p then q is equiv. to either not p or q
Material Equivalence	$(p \equiv q) \equiv [(p \rightarrow q) \vee (q \rightarrow p)]$ $(p \equiv q) \equiv [(p \wedge q) \vee (\neg p \wedge \neg q)]$	(p is equiv. to q) means, either (if p is true then q is true) or (if q is true then p is true) (p is equiv. to q) means, either (p and q are true) or (both p and q are false)
Exportation	$[(p \wedge q) \rightarrow r] \equiv [p \rightarrow (q \rightarrow r)]$	(if p and q are true then r is true) is equivalent to (if q is true then r is true, if p is true)
Importation	$[p \rightarrow (q \rightarrow r)] \equiv [(p \wedge q) \rightarrow r]$	if p is true then (if q then r) is true is equivalent to p and q is true then r is true
Principle of explosion (fallacy)	$\forall p, q : (p \wedge \neg p = \top) \vdash q$	for any statements p and q, if p and $\neg p$ are both true, then we can always prove that q is true
Vacuous truth (fallacy)	$\forall x \in P : Q(x)$ where $P = \varnothing$	statement that asserts that all members of the empty set have a certain property
Existential fallacy	$A \Rightarrow B, B \therefore A$	one presupposes that a class B has members A when one is not supposed to do so
Existence property fallacy	$\exists A \neq P(A)$	existence is a quantifier but not a predicate (i.e. existence is not a property!)
Affirming the consequent (fallacy)	$(P = \top \wedge C = \top); C = \top! \therefore P = \top$	if the premise and conclusions are both true, the conclusion is not a necessary consequence of the premise
Denying the antecedent (fallacy)	$(P = \top \rightarrow C \Rightarrow \top); C = \top! \therefore P = \top$	while the conclusion may be true, it does not follow from the premise
Affirming a disjunct (fallacy)	$A \vee B = \top; B = \top! \therefore A = \perp$	the conclusion does not follow from the premise as it could be the case that A and B are both true
Denying a conjunction (fallacy)	$(B \vee A) = \perp! \therefore A = \top$	A and B cannot be both true the conclusion does not follow from the premise as it could be the case that A and B are both false
Undistributed middle (fallacy)	$Z \in B, Y \in B \therefore Y = Z$	it is not necessarily true that "all Bs are Zs," which is ignored in the argument
False dichotomy (fallacy)	$(p \rightarrow q), (r \rightarrow q), (p \vee q) \therefore q$	it is a false dichotomy as whatever the choice it leads to the same conclusion
A priori argument (motivated reasoning) (fallacy)	$\overline{\Gamma, A \vdash A}^{\text{ax}}$	using an axiom to prove the axiom itself doesn’t make sense as the axiom is therefore beyond refutation

Table 5.15: Most well known rules of inference and fallacies in Classical Logic

With all these tools it’s now easy to disprove any ontological argument for any of the gods of the 45 monotheistic revealed religions actually practised around the world! We already have for example explained the fallacies in Gödel’s ontological argument earlier above. Another type of ontologic arguements are those used mainly by the eleven Abrahamic religions and named the "Thomas Aquinas' Five Ways[8]" (Thomas d'Aquinas was a Dominican friar and priest and influential philosopher and theologian from the 113th century according to holocene calendar):

- **The argument of the unmoved mover:** As particles pop-up from vacuum with non-zero speed in the quantum realm, they don't have any mover. They just extract, as we will prove during our study of Quantum Physics, their energy from the vacuum zero-energy level that has a non-zero value (at small scales principle of conservation of energy is still conserved as a photon is it's own anti-particle and other particles appear with opposite counterpart but at cosmological scales the principle of conservation of energy is violated as we will also prove it). Also light emitted from atoms caused by electrons levels changes aren't accelerated by a mover until they reach the speed of light, they emerge from vacuum instantaneously at the speed of light! Thomas d'Aquinas wasn't able to know that Quantum Physics would exist and most philosophers are illiterate in science and doesn't know nothing about vacuum fluctuations (typically they know nothing about the Euler–Heisenberg Lagrangian nor the Uehling potential).

- **Causation of existence:** Theoretically nothing forbids for example our universe to create itself. Indeed according to Roger Penrose speculative Conformal Cyclic Cosmology (CCC) model the end of the Universe generates it's own next beginning. However at the contrary of the science evidence based results that refutes Aquinas first way introduced previously, there are for the moment in the early 121st century (according to holocene calendar) no strong experimental evidence for Roger Penrose model.

- **Contingent and necessary objects:** This way of Aquinas is proven mathematically wrong by formal logic as it is based on three first order logic fallacies: 1) the a priori argument where Aquinas uses an axiom that leads to a conclusion that is the axiom itself (see the table of logical fallacies above) 2) the principle of explosion as Aquinas uses a contradiction to prove a statement (see the table of logical fallacies above) 3) the existence property fallacy as Aquinas uses a quantifier that is not a predicate (see the table of logical fallacies above).

- **Argument from degrees and perfection:** This one is different from the previous ways! It's primarily semantical nonsense based on solely linguistic definitions that doesn't prove a thing! However an nuclear chemist would very likely say that this argument is scientifically speaking pure nonsense and therefore wrong...

- **Argument from intelligent design:** This way of Aquinas uses a ill-defined concept in science. The "mind" is indeed a philosophical/religious construct, with a fluid definition that is difficult to pin down. No one can prove anything without clearly scientifically defining it and there isn't (sadly) any scientific consensus about its definition still in this early 121st century (holocene calendar). Nevertheless, we actually agree that some objects clearly are what the layman would say "mindless" (a stone for example) while others appear to have what we the layman commonly names a "mind" in vernacular usage.

Obviously we strongly advise the reader to check the original text of Thomas d'Aquinas because taking them out of their context can be misleading (some philosophers always argue that

---

[8] The reader must not forget that philosophers aren't skilled scientists and that's a severe handicap when confronting a philosophical argument to evidence based science.

Aquinas five ways are wrongly interpreted when the latter are debunked by modern science) and we must also remember that nowadays someone like Thomas d'Aquinas wouldn't even obtain a high-school grade in modern science...

Therefore, we can conclude that theist attempt to prove the existence of the god of one of the 45 revealed monotheistic religions based on rhetorical formal logic is easily exposed as vacuous!

5.1.6.2 Is logic a science?

quote 5.1:204.#36

> *Logic is the art of going wrong with confidence.*
>
> Joseph Wood Krutch

Let us recall that this question in the broad sense of the word "science" has obviously a positive answer. However as we have seen in the introduction of this book, there are different sciences!

If we designate by "science" any field centered around the "scientific method": making hypotheses, running experiments in the purpose to refute (i.e. falsify) them and so on, then Logic is nothing like this (logic doesn't have for purpose to refute his premises). Experimental science is inherently based on the physical, observable world. While there is deep thinking involved, a posteriori assumptions, the ultimate arbiter of any experimental scientific idea is the sensible world, through experimentation.

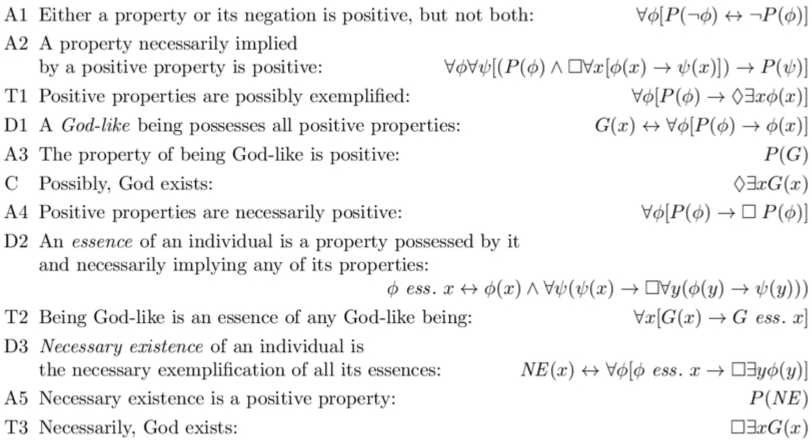
Logic, on the other hand, is entirely self-contained. It's the study of formal systems: sets of axioms and rules on how to derive theorems. Logic is a game where all the rules are made up.

Although logic is not an empirical science or a physical science, it is nonetheless a science in the broad sense of the term, i.e., a systematic body of rational knowledge.

While any use of logic probably may falls back to the real world, logic itself doesn't. It's entirely self-contained. When we prove something logically, we are not making a direct statement about the universe: we are making a claim consistent with the assumptions of our logic system and that may furthermore not be refutable (i.e. not falsifiable) - or even worse - the major well known historical flaw/issue of logic (we have more than 2500 years of evidence of such flaws) is that the conclusion of any logical reasoning is always in the premises. The relationship between logic and the universe is separate: it's what makes logic useful, but it is not what makes logic.

One of the most well known example in "modern" graduate level science of fallacious logic is the *Kalâm cosmological argument* (an argument for the existence of a deity that created our Universe) the second one is very likely the Pascal's wager. The premise of the Kalâm cosmological argument, «*Whatever begins has a cause*», may be kinda logic for an undergraduate or scientifically illiterate individual or a non-professional physicist. But for professionals postgraduate level physicists this is proven mathematically and experimentally wrong at least since the years 11950 (we will see the proofs in the section of Wave Quantum Physics page 4688 and Quantum Gravity page 6074 much later in this book). So anyone using this argument as a logical claim just provide evidence for his ignorance of modern physics[9] and also of logics itself (as the conclusion is in the premise... facepalm). And furthermore as we have also seen during our introduction of this book, a claim or a quote is at maximum an evidence of level 02 (see page 60)...

Very likely the only well known example in "modern" postgraduate level science of fallacious logic (where again the conclusion is in the premises[10]) is the *Gödel's ontological proof* for the existence of (very likely...) the Christian deity... as it suits well Gödel's beliefs... Most theist share the following proof summary on social network this without obviously recognizing all biases as they are for the huge majority of them scientifically illiterate:



A1	Either a property or its negation is positive, but not both:	$\forall\phi[P(\neg\phi)\leftrightarrow\neg P(\phi)]$
A2	A property necessarily implied by a positive property is positive:	$\forall\phi\forall\psi[(P(\phi)\wedge\Box\forall x[\phi(x)\rightarrow\psi(x)])\rightarrow P(\psi)]$
T1	Positive properties are possibly exemplified:	$\forall\phi[P(\phi)\rightarrow\Diamond\exists x\phi(x)]$
D1	A *God-like* being possesses all positive properties:	$G(x)\leftrightarrow\forall\phi[P(\phi)\rightarrow\phi(x)]$
A3	The property of being God-like is positive:	$P(G)$
C	Possibly, God exists:	$\Diamond\exists xG(x)$
A4	Positive properties are necessarily positive:	$\forall\phi[P(\phi)\rightarrow\Box\, P(\phi)]$
D2	An *essence* of an individual is a property possessed by it and necessarily implying any of its properties:	$\phi\ ess.\ x\leftrightarrow\phi(x)\wedge\forall\psi(\psi(x)\rightarrow\Box\forall y(\phi(y)\rightarrow\psi(y)))$
T2	Being God-like is an essence of any God-like being:	$\forall x[G(x)\rightarrow G\ ess.\ x]$
D3	*Necessary existence* of an individual is the necessary exemplification of all its essences:	$NE(x)\leftrightarrow\forall\phi[\phi\ ess.\ x\rightarrow\Box\exists y\phi(y)]$
A5	Necessary existence is a positive property:	$P(NE)$
T3	Necessarily, God exists:	$\Box\exists xG(x)$

Most criticism of Gödel's proof (and other equivalent one) is aimed at its axioms (premises): as with any proof in any logical system, if the axioms of the proof depends on are doubted, then the conclusions can be doubted (that why in logic you can build any proof and that afterwards you need experimental science to sort the garbage from the evidence). It is particularly applicable to Gödel's proof – because it rests on five axioms, some of which are questionable (there are more than $10,000$ gods around the World for example and most of them don't have positive[11] properties... especially in the Abrahamic religions) and some other are refuted theoretically and experimentally (the axiom A1 is wrong as we know from our actual knowledge of Quantum Physics as proven in the chapter on that topic at page 4688) since the 11950 at least... A proof does not necessitate that the conclusion be correct, but rather that by accepting the axioms, the conclusion follows logically. Also Gödel's proof is wrong because again... the conclusion is in the premises. And finally Gödel's ontological proof also suffer from the Gödel's incompleteness theorems!

In a certain set, logic is pure "thought-stuff". It's pure reasoning distilled to a set of simple, verifiable rules. The ultimate arbiter for any logical proof is the logic system, not the universe at large. We can even have perfectly reasonable logic systems that do not map back to our universe at all! It would still be logic, but it probably wouldn't be very interesting or useful! For all these reasons, you can't run an experiment on logic! On the one hand, there is nothing to experiment on, logic is only about validity (when science is about evidence). Again, logic is perfectly self-contained and not dependent on physical reality. On the other hand, logic does not have uncertainty: either you have a proof for something, or you don't. So you don't need to run experiments in the first place, since you have a more powerful notion of proof within the system itself.

---

[9] Another such typical example are that because of existence of truth table, some people again argue that logic can lead to true conclusions. But truth tables depends again on the premise and if the premise is false... guess what!

[10] An argument is "circular" just in case there is a premise, either implicit or explicit, that is logically equivalent to the conclusion. In many cases, circularity is a problem and we name such circularity a "vicious circularity".

[11] And that concept wasn't defined by Gödel...

So logic doesn't involve any experimental configuration and doesn't describe our observable Universe. It is therefore definitely not a scientific method[12]! Let us see a non-exhaustive logical/intuition errors list from some famous philosophers, engineers and scientists that prove the point:

- Aristotle wasn't the first to say that the Earth occupies the center of the universe, but he was the most dogmatic about it, and believed he had established it to be incontrovertibly true by using... logic.
- Euclid was sure by logic that the axiom of parallels (parallel postulate) should always have been true. However modern mathematics and modern experimental science proved he was wrong.
- According to Kepler's logic, planet orbits are ellipses because of his religious faith in perfection. We know now that in fact planets describe, from the point of view of a centered galactic observer, a helicoidal movement and describe a given ellipse whose parameters changes across billions years and may even become chaotic over time.
- According to Boyle's intuition (logic), light had to propagate in something named an "aether". This was later shown experimentally and theoretically wrong.
- According to Darwin's intuition (logic), characteristics of the mother and the father simply get blended in the offspring just as a can of black paint and a can of white paint blend to generate gray when combined. This was proven wrong later by Mendel and modern genetics.
- According to Kelvin's reasoning method, Earth's age would have been 50 times younger than what modern radio-isotopic dating methods provide as evidence.
- According to Kelvin's logic, heavier than air flying machines must have been impossible...
- In the early 120th century (holocene calendar), the vast majority of geologists rejected Wegener's idea of continental drift, arguing that ocean floors were fixed and continents could not logically move.
- According to Einstein's intuition (logic), the Universe must have been static. So he introduced the cosmological constant in it's theory. Evidence proved his logic was flawed.
- According to Einstein's (and others!) intuition and logic, spooky action at a distance (entanglement) was something impossible and wrong.
- Einstein denied several times that Black Holes could form.

---

[12] There is a quite famous 700 pages book [503] written by a philosopher that presents and criticize in details (in a horrible way because obviously most philosophers don't know how to write properly textbooks in LaTeX) most existing ontological proofs and disproofs of gods but however without relying to experimental scientific evidence and focusing only on the main Abrahamic religions. Typical errors of most occidental philosophers by the way...

- In the early days of quantum mechanics, Einstein tried to show that the uncertainty principle was not valid. By 11927 (holocene calendar) he had become convinced of its utility, but he always opposed it
- Eddington's failed to take Chandrasekhar's theory of gravitational collapse seriously because for him it was something illogical.
- etc...

Obviously some people may argue that the errors above were technical errors and not logical errors. That latter point is a matter of semantic !

But how is experimental science a logic?

Well, the core concept in the scientific method is inductive reasoning. Roughly put, it's the idea that if we do X and Y happens $1,000$ times in a row, X will probably lead to Y next time too. If we do this enough times - and are sufficiently careful - we'll just conclude that X causes Y. That's the reason experiments make sense and let us predict the future. As we will see during our study of statistical hypothesis test later, these latter are build on a kind of probabilistic inference and therefore on logic. So the scientific logic is based on logic but... logic is not a scientific method!

Oh! And by the way... Some non-academic people argue that there is only one unique "scientific method". It is quite ambiguous to make such a statement as it is constantly evolving! The documented scientific method by Francis Bacon in 1620 has almost nothing to do anymore with the scientific method of the early 121st century (holocene calendar)!

To finish, let us quote now a part of a text written by Donald E. Simanek on it's website, because it's worth it:

«Scientists do not arrive at models and theories by application of logic. They arrive at them by many processes lumped under the name "induction". Induction cannot be reduced to a set of logical rules (though many have tried). To see patterns (sometimes subtle and hidden ones) in data and observations requires creative ability. This is the ability to think ahead and say: «*What model, set of statements (laws) or theoretical construct could I devise from which these observations and data might be deduced?*»

We can't find, discover, or construct scientific laws and theories by mathematics and logic alone. But we can derive testable and useful results by application of mathematics and logic to laws and theories, and if those deduced results pass experimental tests, our confidence in the validity of the theory from which they were derived is strengthened.

Some people are profoundly disturbed by the fact that reason alone can't generate truths. When the use of mathematics and logic in science is explained to them they respond, «*If mathematics and logic can't produce absolute truths, then they produce only untruths or partial truths, and are therefore worthless.*» This sentence is itself an example of nonsense clothed in the appearance of logic.

It must be admitted at the outset that science is not in the business of finding absolute truths. Science proceeds as if there are no absolute truths, or if there are such truths, we can never know what they are. As the pre-Socratic skeptics observed: If we were to stumble upon an absolute truth, we'd have no way to be certain it is an absolute truth. The models and theories of science are approximations to nature - never perfect. But in most cases we know rather well how good they are. We can state quantitatively the limits of uncertainty of numeric results, and their range of applicability. Yet there's always the possibility that we may find exceptions to one of our accepted laws, or may even find alternative theories that do a better job than older ones.

[...]

Some critics of science attack this process of science, on the grounds that it cannot produce absolute truths. Theirs is a black/white view of the scientific process. Never mind that they have not proposed any other process that is capable of producing anything near the power and comprehensiveness of present science. They say that "Theory X" isn't perfect therefore it is "wrong".

[...]

The fact that science claims no absolute truths is seized upon by people who hold strong religious beliefs and who dislike those conclusions of science that run counter to their emotional convictions. To them, if a thing is not absolutely and finally true, it is false, and therefore the methods used to formulate it must be flawed.»

quote
5.1:208.#37

> ❠ *Philosophy debates God, maths ignores him, climatology doesn't need him, physics can't find him, biology contradicts him, logic kills him!* ❞
>
> Bill Flavell

~80 %

★★★★☆

64 votes, 80.94%

5.2. Numbers

THE basis of mathematics, apart the reasoning (see section Proof Theory page 131), is undoubtedly to ordinary people: arithmetic. It is therefore mandatory that we make a stop on it to study its origin, some of its properties and consequences.

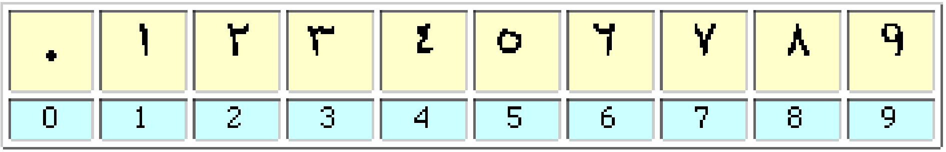
The numbers, like geometric figures are the basis of Geometry, are the basis of Arithmetic. These are also the historical basis because mathematics probably started with the study of these objects, but also the educational foundation, because it is by learning to count that we enter in the world of mathematics.

The history of numbers, also sometimes named "scalar", is far too long to be told here! But we can only advise you one of the best book on the subject: The Universal History of Numbers (approximately 2,000 pages in three volumes) Georges Ifrah, ISBN: 2221057791.

But here's a little flange of this latter which seems fundamental to us:

Our current decimal system, on base 10, uses the digits 0 to 9, named "Arabic numbers", but the fact of Indian origin (Hindus). The first numbers seems to have been invented in the 97th (holocene calendar) in India by Brahmagupta, an Indian mathematician. he invented the figures in Devanagari.

Indeed, the Arabic numbers (of Indian origin...) in the table below are the first line and we see that they are significantly different from the "Indian numbers" of the second line:



٠	١	٢	٣	٤	٥	٦	٧	٨	٩
0	1	2	3	4	5	6	7	8	9

Figure 5.6: Indo-Arabic numbers

You have to read this table as following from left to right: 0 "zero", 1 "one", 2 "two", 3 "three", 4 "four", 5 "five", 6 "six", 7 "seven", 8 "eight, 9 "nine". This system is much more efficient than the Roman numerals (try doing a calculation with Roman notation system you will see...).

It is commonly accepted that these numbers were introduced in Europe only about the year 11000 (holocence calendar). Used in India, they were transmitted by Arabs to the Western world by the Pope Gerbert of Aurillac during his stay in Andalusia at the end of the 109th century (holocene calendar).

Remark

The French word "chiffre" (number) is a corruption of the Arabic word "sifr" meaning "zero". In Italian, "zero" is "zero", and seems to be a contraction of "zefiro", we again see here an Arabic root but the "zero" could also be of Indian origin... So the words "chiffre" and "zero" have the same origin.

The early use of a numerical symbol for the "nothing" in the sense of "no amount", i.e. our "zero" is because the Indians used a system named "positional system". In such a system, the position of a digit in the writing of a number expresses the power of 10 and the number of times it occurs ... and the absence of a position in this system arise from huge proofreading problems and could lead to large errors in calculations. The revolutionary and simple introduction of the concept of "nothing" allowed a proofreading without error of numbers.

The absence of a power is denoted by a small circle...: the zero. Our current system is thus the "decimal and positional system".

✍**Example:**

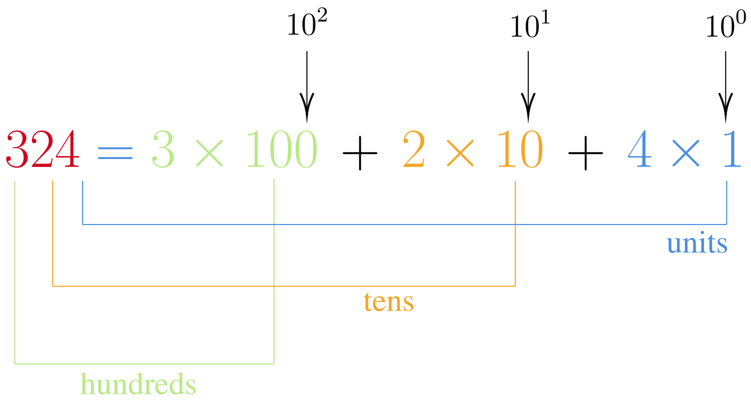


Figure 5.7: **Description of decimal and positional system**

The number 324 is written from left to right as three hundred: 3 times 100, two tens: 2 times 10 and four units: 4 times 1.

Thus a "decimal number" is thus a number that has a finite writing in base 10.

We sometimes see (and this is recommended) a thousands separator represented by a coma in United States (put all three numbers from the first from the right for the whole numbers). Thus, we write $1,034$ instead of 1034 or $1,344,567,569$ instead of 1344567569. Thousand separators permits to quickly quantify the magnitude of the read numbers.

So:

- If we see only one coma we know that the number is about thousands
- If we see two apostrophes we know that the number is about millions

- If we see three apostrophes we know that the number is about billions
- etc.

and so on... also with decimals this gives:

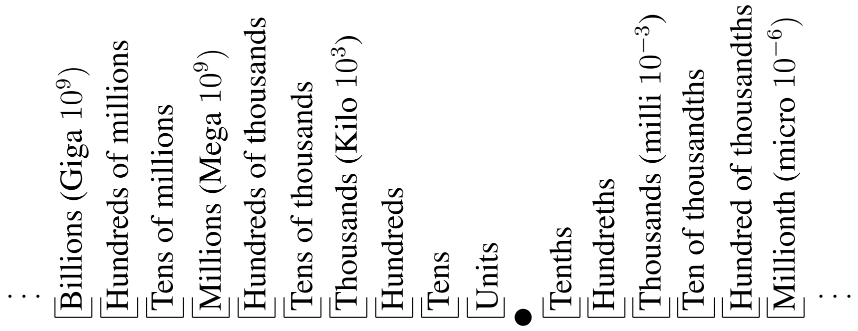

. . .
Billions (Giga 10^9)
Hundreds of millions
Tens of millions
Millions (Mega 10^9)
Hundreds of thousands
Tens of thousands
Thousands (Kilo 10^3)
Hundreds
Tens
Units
●
Tenths
Hundreths
Thousands (milli 10^{-3})
Ten of thousandths
Hundred of thousandths
Millionth (micro 10^{-6})
. . .

Figure 5.8: **Scale representation of the positional system**

In fact, any integer other than the unit can be taken as the basis of a numbering system. We have for example the binary, ternary, quaternary, ..., decimal, duodecimal numbering systems which correspond respectively to the bases two, three, four, ..., ten, twelve.

A generalization of what has been seen above, can be written as follows:

Any positive integer N can be represented in a base b as a sum, where each coefficient a_i are multiplied by their respective weight b^i. Such as:

$$N=a_{n-1}b^{n-1}+a_{n-2}b^{n-2}+\ldots+a_1b^1+a_0b^0 \tag{5.2.1}$$

More elegantly written:

$$\boxed{N=\sum_{i=0}^{n-1}a_ib^i} \tag{5.2.2}$$

with $a_i\in[0,b-1]$ and $b_i\in[1,b^{n-1}]$

Remarks

R1. As frequently in mathematics, we will replace numbers with Latin or Greek letters in order to generalize their representation. So when we speak of a base b, the value of b can take any positive integer value $1,2,3,\ldots$

R2. When we take the value 2 for b, the maximum value of N will be 2^n-1. The numbers that are written in this form are named "Mersenne numbers". These numbers can be prime numbers (see further below what a prime number is) if and only if n is also a prime number.

Indeed, if we take (for example) $b=10$ and $n=3$ the largest value we can get will be:

$$(b-1)10^2+(b-1)10^1+(b-1)10^0=900+90+9=999=10^3-1=b^3-1 \tag{5.2.3}$$

R3. When a number is the same read from left to right or right to left, we name it a "palindrome".

5.2.1 Digital Bases

To write a number in base b system, we must first adopt b characters for representing the b first numbers for example in the decimal system: $\{0,1,2,3,4,5,6,7,8,9\}$. These characters are as we already defined them, the "digits" that we pronounce as usual $\{\text{zero}, \text{one}, \text{two}, \text{three}, \text{four}, \text{five}, \text{six}, \text{seven}, \text{eigth}, \text{nine}\}$.

For the written numbers, we make this convention that a digit, placed to the left of another represents the order units immediately above, or b times larger. To take the place of units that may be lacking in certain orders, we use the zero "0" and consequently, the number of digits may vary.

Definition (#5.2.29): For the spoken numbers, we agree to name "single unit", "ten", "hundred", "thousand", etc., units of the first, second, third, fourth order, etc. Thus the numbers $10,11,\ldots,19$ will be readed in the same way in all numbering systems. The numbers $1a,1b,a0,b0,\ldots$ will be readed ten-a, ten-b, a-ten, b-ten, etc. Thus, the number 5b6a71c will be readed:

five million be-hundred sixty-a thousand seven hundred ten-c

This small example is relevant because it shows the general expression of the spoken language we use daily is intuitively in base ten (fault of our education).

Remarks

R1. The rules of mathematical operations defined for numbers written in the decimal system are the same for numbers written in any numbering system.

R2. To quickly operate in any numbering system, it is useful to know by heart all sums and products of two numbers of a single digit.

R3. The decimal seems has its origin in the fact that humans being have ten fingers.

Let's see how we convert a numbering system in another one:

✍**Examples:**

E1. In base ten we have seen above that $142,713$ will be written as:

$$142,713_{10}=1\cdot 10^5+4\cdot 10^4+2\cdot 10^3+7\cdot 10^2+1\cdot 10^1+3\cdot 10^0 \tag{5.2.4}$$

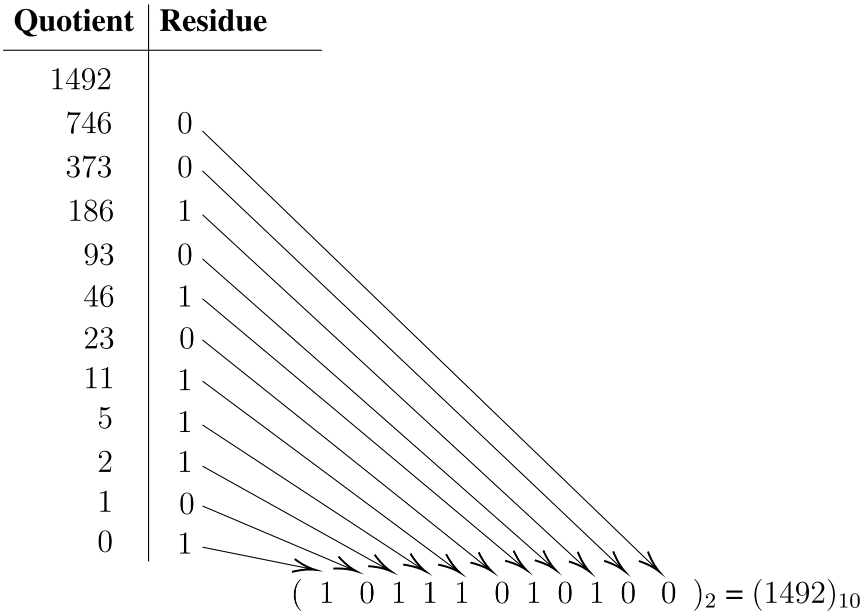
E2. The number 0110 that is in base two (binary base) would be written in base 10:

$$0110_2=0\cdot 2^3+1\cdot 2^2+1\cdot 2^1+0\cdot 2^0=6_{10} \tag{5.2.5}$$

and so on... The reverse operation is often a little trickier (for example the case of the binary base):

✍**Examples:**

E1. Converting the decimal number $1,492$ in binary base is done by successive divisions by 2 of residues and gives (the principle is about the same for all other bases):



Quotient	Residue
1492	
746	0
373	0
186	1
93	0
46	1
23	0
11	1
5	1
2	1
1	0
0	1

$$(1\ 0\ 1\ 1\ 1\ 0\ 1\ 0\ 1\ 0\ 0)_2=(1492)_{10}$$

Figure 5.9: **Decimal to binary conversion**

E2. To convert the number $142,713$ (decimal base) in duodecimal base (base twelve) we have (notation: q is the "quotient", and r is the "residue"):

$$\begin{aligned}\frac{142,713}{12} &\Rightarrow q=11,892 \quad r=9\\ \frac{11,892}{12} &\Rightarrow q=991 \quad r=0\\ \frac{991}{12} &\Rightarrow q=82 \quad r=7\\ \frac{82}{12} &\Rightarrow q=6 \quad r=10\\ \frac{6}{12} &\Rightarrow q=0 \quad r=6\end{aligned} \tag{5.2.6}$$

Thus we have the residues 6, 10, 7, 0, 9 which leads us to write:

$$142,713_{10}=6a709_{12} \tag{5.2.7}$$

where we have chosen for this particular example the symbolism that we have previously defined (a-ten) to avoid any confusion.

5.2.2 Type of Numbers

Now that we know that number is a mathematical object used to count, measure and label it must be know that it exists in mathematics a wide variety of numbers (natural, rational, real, irrational, complex, p-adic quaternions, transfinite, algebraic, constructibles, etc.) since any mathematician may at leisure invent its own numbers just by defining axioms (rules) to manipulate them.

However, there are a few of them that we find much more often than others through this book and some that serve as basic construction for others and which that should be defined sufficiently rigorously (without going to the extremes) in order to know what we will talk about when we will use them.

5.2.2.1 Natural Integer Numbers

The idea of "integer" (the numbers for which there are no decimals) is the fundamental concept of mathematics and comes at the view of a group of objects of the same types (a sheep, another sheep, yet another sheep, etc.).

When the amount of objects in a group is different from that of another group when the speak about a group that is numerically higher or lower regardless of the type of objects in these groups. When the amount of objects of one or multiples groups is equivalent, then we speak about "equality".

To each single object the number "one" or "unit" denoted by "1" in the decimal system will be used.

To form groups of objects, we can operate as follows: to an object, add another object, then another, and so on... each of the clusters, from the point of view of its community, is characterized by a number. It follows from this that a number can be regarded as representing a group of units (single items) such that each unit corresponds to one single object of the collection.

Definition (#5.2.30): Two numbers are said to be "equal" if each of the units of one we can match a unique unit of the other and vice versa (in a bijective way as seen in the section of Set Theory page 407). If this does not hold true when we talk about "inequality".

Below we have a graphical example of a set defined by two inequalities:

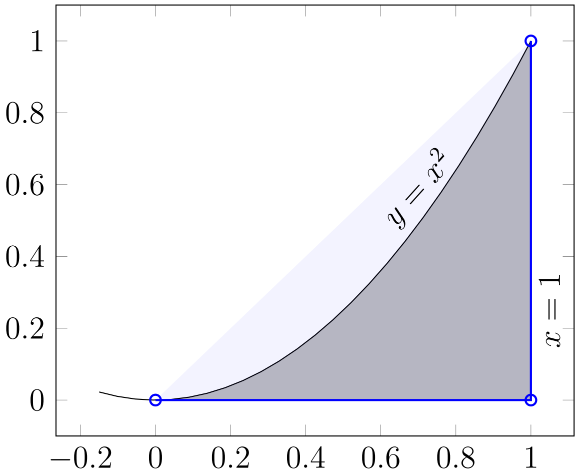


Figure 5.10: $R = \{(x, y) \mid 0 \leq y \leq x^2, x \leq 1\}$

Let us take an object, then another, then to the formed group add again an object and so on. The groups thus formed are characterized by numbers which, taken in the same order as the groups successively obtained, are the "natural sequence $\mathbb{N}$" or "natural integers", also sometimes named "whole numbers", and denoted by:

$$\boxed{\mathbb{N} = \{0, 1, 2, 3, \ldots\}} \tag{5.2.8}$$

To be unambiguous about whether 0 is included or not, sometimes an index (or superscript) is added in the former case:

$$\mathbb{N}^* = \mathbb{N}^1 = \{1, 2, 3, \ldots\} \tag{5.2.9}$$

Remark

The presence of the 0 (zero) in our definition above of $\mathbb{N}$ is debatable since 0 is neither positive nor negative. That is why in many books you will find a definition of $\mathbb{N}$ without the 0 (hence corresponding to $\mathbb{Z}^+$) and therefore the set $\mathbb{N} \cup \{0\}$ (also sometimes denoted $\mathbb{Z}_{\geq 0}$ or $\mathbb{N}_0$) is named the "calculative set" or "whole numbers". But keep in mind that there seems to be no general agreement and it is actually not defined in the ISO 80000 norm series!!!

The components of this natural set can be defined by (we own this definition to the mathematician Frege Gottlob) the following the properties (having read first the section on Set Theory page 378 is strongly recommended...):

P1. 0 (read "zero") is the number of elements (defined as an equivalence relation) of all sets equivalent to (in bijection with) the empty set.

P2. 1 (read "one") is the number of elements of all sets equivalent to the set whose only element is 1.

P3. 2 (read "two") is the number of elements of all sets equivalent to the set whose only element are 1 and 2.

P4. In general, an integer is the number of elements of all sets equivalent to the set of integers preceding it!

The construction of the set of natural numbers is made of the most natural and consistent manner. Natural numbers get their name from what they were, in the beginnings of their existence, to count quantities and things of nature or intervened in human life. The originality of this set lies in the empirical way he has been built since it does not actually the result of a mathematical definition, but more by awareness of the human by the concept of countable quantity, of number and operations that reflect the relations between them.

The question about the origin of $\mathbb{N}$ is therefore the question of the origin of mathematics. And since thousands of years debates confronting the thoughts of the greatest philosophical minds have attempted to elucidate this deep mystery as to whether mathematics is a pure creation of the human mind or whether the man has only rediscovered a science that already existed in nature. Besides the many philosophical questions that the set of Natural numbers can generate, it is nonetheless interesting from an exclusively mathematical point of view. Because of its structure, it has remarkable properties that can be very useful when we practice some given reasoning or calculations.

The sequence of natural numbers $\mathbb{N}$ is unlimited (see section Numbers page 215) but countable (we will this property in details below), because in a group of objects that is represented by a number n, it will be enough to add an object to get another group that will be defined by the integer $n+1$.

Definition (#5.2.31): Two integers that differs from a single positive unit are said to be "consecutive".

5.2.2.1.1 Peano axioms

During the crisis of foundations of mathematics, mathematicians have obviously sought to axiomatize the set $\mathbb{N}$ and we own the actual axiomatisation to Peano and Dedekind.

The axioms of this system include the symbols $<$ and $=$ to represent the relations "smaller than" and "equal to" (see section Operators page 288). They include also the symbols "0" for the number zero and s to represent the "successor" number. In this system, 1 is denoted by:

$$1 = s(0) \tag{5.2.10}$$

named "successor to zero" and 2 is denoted by:

$$2 = s(s(0)) = s(1) \tag{5.2.11}$$

The Peano axioms that builds $\mathbb{N}$ are (see section of Proof Theory page 131 for details on some of the symbols used below):

A1. 0 is a natural number (this permits $\nvdash$ to be not empty).

A2. Every natural number n has a successor, denoted by $s(n)$.

Therefore s is an injective application (see section Set Theory page 406), that is to say:

$$\forall x, y \quad s(x) = s(y) \Leftrightarrow x = y \tag{5.2.12}$$

That is to say that if two successors are equal, they are the successors of the same number.

A3. The successor of a natural number is never zero (therefore $\mathbb{N}$ has a first element):

$$\forall x \quad \neg(s(x) = 0) \tag{5.2.13}$$

A4. If we prove a property φ that is true for x and its successor $s(x)$, then this property is true for any x (this is named the "axiom of recurrence"):

$$(\varphi(x) \Rightarrow \varphi(s(x))) \Rightarrow \forall x\ \varphi(x) \tag{5.2.14}$$

So the set of all the numbers satisfying the four above axioms is denoted by:

So the set of all the numbers satisfying the four above axioms is denoted by:

$$\mathbb{N} = \{0, 1, 2, 3, \ldots, n, \ldots\} \tag{5.2.15}$$

Remark

The Peano axioms allow to build very rigorously the two basic operations of arithmetic in $\mathbb{N}$ that are addition and multiplication (see section Arithmetic Operators page 296 and page 296) and so all the other sets that we will see later (subtraction in $\mathbb{N}$ can not be applied because it can give negative numbers).

5.2.2.1.2 Odd, Even and Perfect Numbers

In arithmetic, study the parity of an integer, its determiner if this integer is or not a multiple of 2. An integer multiple of 2 is an even integer, the others are odd integers.

Definitions (#5.2.32):

D1. The numbers obtained by counting by step of 2 from zero (i.e.. $0, 2, 4, 6, 8, \ldots$) in the set of natural integer numbers $\mathbb{N}$ are named "even numbers".

The n^{th} even number is obviously given by the relation:

$$\mathbb{E}: \quad 2n = n + n \quad \forall \mathbb{N} \tag{5.2.16}$$

Or more aesthetically:

$$\mathbb{E} = \{x \in \mathbb{N} : \exists n \in \mathbb{N}(x = 2n)\} \tag{5.2.17}$$

D2. The numbers we get by counting by step of 2 starting from 1 (i.e. $1, 3, 5, 7, \ldots$) in the set of natural integer numbers $\mathbb{N}$ are named "odd numbers".

The $(n+1)^{\text{th}}$ even number is almost obviously given by the relation:

$$\mathbb{O}: \quad 2n + 1 = \overline{\mathbb{E}} \quad \forall \mathbb{N} \tag{5.2.18}$$

Or more aesthetically:

$$\mathbb{O} = \{x \in \mathbb{N} : \exists n \in \mathbb{N}(x = 2n + 1)\} \tag{5.2.19}$$

Remark

We name "perfect numbers", numbers equal to the sum of their integer divisors strictly smaller than themselves (concept we will see in detail later) such as: $6 = 1 + 2 + 3$ and $28 + 1 = 2 + 4 + 7 + 14$.

What is the sum of the first n odd integers?

Some checks are necessary. The first odd integer is 1. The sum of the first two odd numbers is:

$$1 + 3 = 4 \tag{5.2.20}$$

The sum of the first three odd numbers is:

$$1 + 3 + 5 = 9 \tag{5.2.21}$$

The sum of the first four odd numbers is:

$$1 + 3 + 5 + 7 = 16 \tag{5.2.22}$$

We can think that the sum of the first n odd integers is worth n^2 in other words that:

$$1 + 3 + 5 + \ldots + (2n - 1) = n^2 \tag{5.2.23}$$

Let $\mathrm{P}(n)$ be this relation that we wish to establish for all n.

Let's unfold the three phases of recurrence:

- First phase: the initial verification. It consists of verifying the relation for the first index. Here 1. For $n=1$, in each member of the relation above, we find 1, we can claim that $P(1)$ is true. Note that we have just previously made the verifications at rank $2,3$ and 4.

- Second phase: we establish the hereditary nature of the property (the passage for all n from n to $n+1$). Let the recurrence hypothesis be: «We assume that $P(n)$ is true for a certain rank n». Passage to $n+1$: We demonstrate that the relation is true at rank $n+1$, in other words that:

  $$1+3+5+\ldots+(2n-1)+(2n+1)=(n+1)^2 \qquad (5.2.24)$$

  In the first member we recognize the sum of the first n odd numbers (which is n^2 according to the recurrence hypothesis) to which we have added the $(n+1)$-th odd number $2n+1$. Thus:

  $$1+3+5+\ldots+(2n-1)+(2n+1)=n^2+2n+1 \qquad (5.2.25)$$

  Or:

  $$n^2+2n+1=(n+1)^2 \qquad (5.2.26)$$

- Third phase: The closing clause allows us to take make the summary. By recurrence on n, we have just established that the sum of the first n odd integers is worth n^2.

5.2.2.1.3 Prime Numbers

Definition (#5.2.33): A "prime number" is an integer with exactly two positive divisors (these divisors are both: "1" and the number itself). In the case where there are more than two dividers it is named a "composite number". The property of being prime (or not) is named "primality".

The study of prime numbers is a huge subject in mathematics (see for a small example the section of Number Theory page 352 or of Cryptography page 7216). There are books of thousands of pages on the subject and probable hundreds of research article per month even nowadays. Most theorems are largely out of the study of this book (and out of the interest of its main author...)!

Here is the set of prime numbers less than 1000:

2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 61, 67, 71, 73, 79, 83, 89, 97, 101, 103, 107, 109, 113, 127, 131, 137, 139, 149, 151, 157, 163, 167, 173, 179, 181, 191, 193, 197, 199, 211, 223, 227, 229, 233, 239, 241, 251, 257, 263, 269, 271, 277, 281, 283, 293, 307, 311, 313, 317, 331, 337, 347, 349, 353, 359, 367, 373, 379, 383, 389, 397, 401, 409, 419, 421, 431, 433, 439, 443, 449, 457, 461, 463, 467, 479, 487, 491, 499, 503, 509, 521, 523, 541, 547, 557, 563, 569, 571, 577, 587, 593, 599, 601, 607, 613, 617, 619, 631, 641, 643, 647, 653, 659, 661, 673, 677, 683, 691, 701, 709, 719, 727, 733, 739, 743, 751, 757, 761, 769, 773, 787, 797, 809, 811, 821, 823, 827, 829, 839, 853, 857, 859, 863, 877, 881, 883, 887, 907, 911, 919, 929, 937, 941, 947, 953, 967, 971, 977, 983, 991, 997

The whole set of prime numbers is sometimes denoted by $\mathbb{P}$.

Remark

Note that the primes numbers set does not include the number "1" because it has a only a single divider (himself) and not two as is the definition.

We can ask ourselves if there are infinitely many prime numbers? The answer is YES and here is a proof (among others) by contradiction.

Proof 5.3.3. Suppose that there is a finite number of prime numbers that would be denoted by:

$$p_1, p_2, \ldots, p_n \tag{5.2.27}$$

We create a new number from the product of this prime number to which we add "1":

$$N = (p_1 p_2 \ldots p_n) + 1 \tag{5.2.28}$$

According to our initial hypothesis and the fundamental theorem of arithmetic (see section Number Theory page 352) the new number N should be divisible by one of the existing prime p_i such that we can write:

$$N = q \cdot p_i \tag{5.2.29}$$

where q is an integer. We can make the division:

$$q = \frac{(p_1 p_2 \ldots p_n) + 1}{p_i} = \frac{p_1 p_2 \ldots p_n}{p_i} + \frac{1}{p_i} \tag{5.2.30}$$

The first term is simplified as p_i is in the product. Let us note the resulting integer E:

$$q = E + \frac{1}{p_i} \tag{5.2.31}$$

But, q and E are integers, so $1/p_i$ should be an integer. But p_i is by definition greater than 1. So $1/p_i$ is not an integer and so is also q.

Then there is contradiction, and we can conclude that the prime numbers are not finite but are infinite.

■ Q.E.D.

Remarks

R1. The product $p_n = p_1 p_2 \ldots p_n$ of the indexed prime numbers $\leq n$ is named the "n-th primorial".

R2. We send the reader to the section Cryptography of the chapter on Theoretical Computing page 7216 (or Number Theory section of the chapter Arithmetic page 365) for the study of some remarkable properties of prime numbers including the famous Euler ϕ function (also named "indicator function") and a 120th-121th century (holocene calendar) industrial application of prime numbers.

5.2.2.2 Relative Integer Numbers

The set of natural integers $\mathbb{N}$ has a few issues that we did not set out earlier. For example, subtracting two numbers into $\mathbb{N}$ does not always have a result in $\mathbb{N}$ (negative numbers not existing in this set). Other issue... dividing two numbers in $\mathbb{N}$ also does not always have a result in $\mathbb{N}$ (fractional numbers - rational or irrational - not existing in this set). We then say in the language of set theory that: the subtraction and division is not an internal operation of $\mathbb{N}$.

We can first resolve the problem of subtraction by adding to the set of natural numbers $\mathbb{N}$, negative integers (revolutionary concept for those who where behind this concept at their time!) to get the set of "relative integers" denoted by $\mathbb{Z}$ (for "Zahl" from German, meaning "Number"):

$$\boxed{\mathbb{Z} = \ldots, -3, -2, -1, 0, 1, 2, 3, \ldots = \{p - q \mid p, q \in \mathbb{N}\}} \tag{5.2.32}$$

The set of natural integers is therefore included in the set of relative integers. This is what we denote by (see section Set Theory page 379):

$$\mathbb{N} \subset \mathbb{Z} \tag{5.2.33}$$

and we have by definition (it is a notations to be learned!!!):

$$\begin{aligned} \mathbb{Z}^+ &= \mathbb{Z}_{>0} = \{n \in \mathbb{Z} | n > 0\} = \mathbb{N}^* \\ \mathbb{Z}_0^+ &= \mathbb{Z}_{\geq 0} = \{n \in \mathbb{Z} | n \geq 0\} = \mathbb{N} \\ \mathbb{Z}_0^- &= \mathbb{Z}_{<0} = \{n \in \mathbb{Z} | n < 0\} \\ \mathbb{Z}^* &= \mathbb{Z}_{\neq 0} = \{n \in \mathbb{Z} | n \neq 0\} \end{aligned} \tag{5.2.34}$$

This set was originally created to make the natural numbers an object that we name a "group" (see section Set Theory page 421) relatively to the addition.

Definition (#5.2.34): We say that a set A is a "countable set", if it is equipotent to $\mathbb{N}$. That is to say if there is a bijection (see section Set Theory page 407) of S on $\mathbb{N}$. Thus, roughly said, two equipotent sets have the same number of elements in the meaning of their cardinal (see section Set Theory page 388), or at least the same infinity.

The purpose of this concept is to understand that the sets $\mathbb{N}$ and $\mathbb{Z}$ are countable.

Proof 5.3.4. Let us show that $\mathbb{Z}$ is countable by writing:

$$x_{2k} = k \quad \text{and} \quad x_{2k+1} = -k - 1 \tag{5.2.35}$$

for any integer $k \geq 0$. This gives the following ordered list:

$$0, -1, 1, -2, 2, -3, 3, \ldots \tag{5.2.36}$$

of all relative integers from natural integers only!

■ Q.E.D.

5.2.2.3 Rational Numbers

The set of relative integers $\mathbb{Z}$ also still has an issue. Dividing two numbers in $\mathbb{Z}$ also does not always have a result in $\mathbb{Z}$ (fraction numbers - rational or irrational - not existing in this set). We then say in the language of set theory that: the division is not an internal operation of $\mathbb{Z}$.

We can thus define a new set that contains all the numbers which can be written as a "fraction" that is to say the ratio of a dividend (numerator) and a divider (denominator). When a number can be written in this form, we say that it is a "fractional number":

$$\text{fraction}:\left\{\frac{2}{3}\right.\begin{array}{l}\swarrow\ \text{Numerator}\\ \longleftarrow\ \text{fraction bar}\\ \nwarrow\ \text{Denominator}\end{array}\tag{5.2.37}$$

A fraction can be used to express a part or fraction of something (of an object, of a distance, of a land, of an amount of money, of a cake...).

To better understand rational number (fractions) let us consider two individuals: Andy and Bobby that bot love pizza. On Monday night, they share a pizza equally. How much of the pizza does each one get? Are you thinking that each boy gets half of the pizza? That's right. There is one whole pizza, evenly divided into two parts, so each boy gets one of the two equal parts. In math, we write $\dfrac{1}{2}$ to mean one out of two parts:

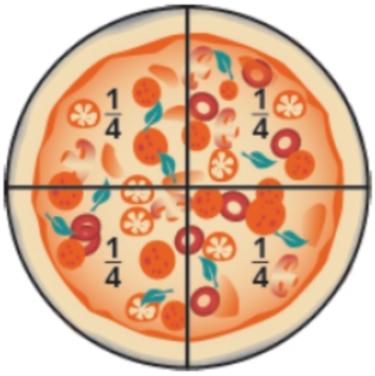


Figure 5.11: **A $1/2$ pizza fraction example (source: OpenStax)**

On Tuesday, Andy and Bobby share a pizza with their parents, Fred and Christy, with each person getting an equal amount of the whole pizza. How much of the pizza does each person get? There is one whole pizza, divided evenly into four equal parts. Each person has one of the four equal parts, so each has $\dfrac{1}{4}$ of the pizza:

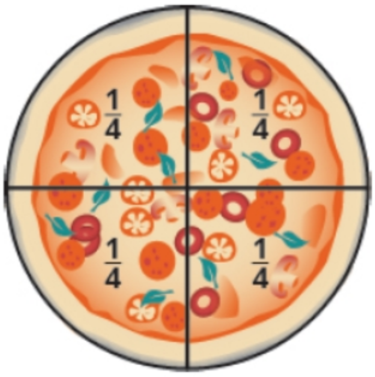


Figure 5.12: **A $1/4$ pizza fraction example (source: OpenStax)**

On Wednesday, the family invites some friends over for a pizza dinner. There are a total of 12 people. If they share the pizza equally, each person would get $\dfrac{1}{12}$ of the pizza:

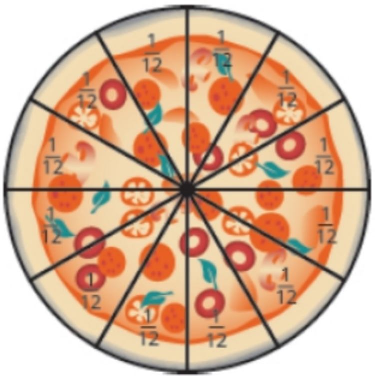


Figure 5.13: **A $1/12$ pizza fraction example (source: OpenStax)**

By definition, the "set of rational numbers" is given by:

$$\boxed{\mathbb{Q}=\left\{\frac{p}{q}|(p,q)\in\mathbb{Z},q\neq 0\right\}} \tag{5.2.38}$$

In other words, any rational number is any number that can be expressed as the quotient or fraction p/q of two integers, p and q, with the denominator q not equal to zero. Since q may be equal to 1, every integer is a rational number.

In the case we put an additional restriction on the denominator such that $q=10^n$ and $n\in\mathbb{N}$ we define therefore the "set of decimal numbers" given by:

$$\boxed{\mathbb{D}=\left\{\frac{p}{10^n}|p\in\mathbb{Z},n\in\mathbb{N}\right\}} \tag{5.2.39}$$

Remark

Pure mathematicians will argue that the definition above is wrong. Yes indeed it is not exact but good enough for engineers and physicists. So here is the right one:

$$\mathbb{Q}=\left\{\frac{p}{q}|(p,q)\in\mathbb{Z},q>0 \text{ and } \gcd(p,q)=1\right\} \tag{5.2.40}$$

Where $q>0$ avoids the division by zero AND the repetition of the sign "$-$" above or below the fraction. We have also added an important property: the greatest common divisor (see section Number Theory page 340) between the both number must be equal to zero. Thus in the purpose that we consider, for example, $3/6$ and $1/2$ are not two distinct numbers. In fact we know that we can always bring a fraction to its irreducible form, where in the latter case we will have $\gcd(p,q)=1$.

We also assume as obvious that:

$$\mathbb{N}\subset\mathbb{Z}\subset\mathbb{D}\subset\mathbb{Q} \tag{5.2.41}$$

The logic of the creation of the set of rational numbers $\mathbb{Q}$ is similar to that of relative integers $\mathbb{Z}$. Indeed, mathematicians wanted to make of the set of relative numbers $\mathbb{Z}$ a "group" with respect to the laws of multiplication and division (see section Set Theory page 308 and page 313).

Moreover, contrary to the intuition of most people, the set of natural integers $\mathbb{N}$ and rational numbers $\mathbb{Q}$ are equipotent (i.e. $\mathbb{Q}\simeq\mathbb{N}$). We can convince ourselves of this equipotence by ranking, as Cantor did, rational numbers in a first time as follows:

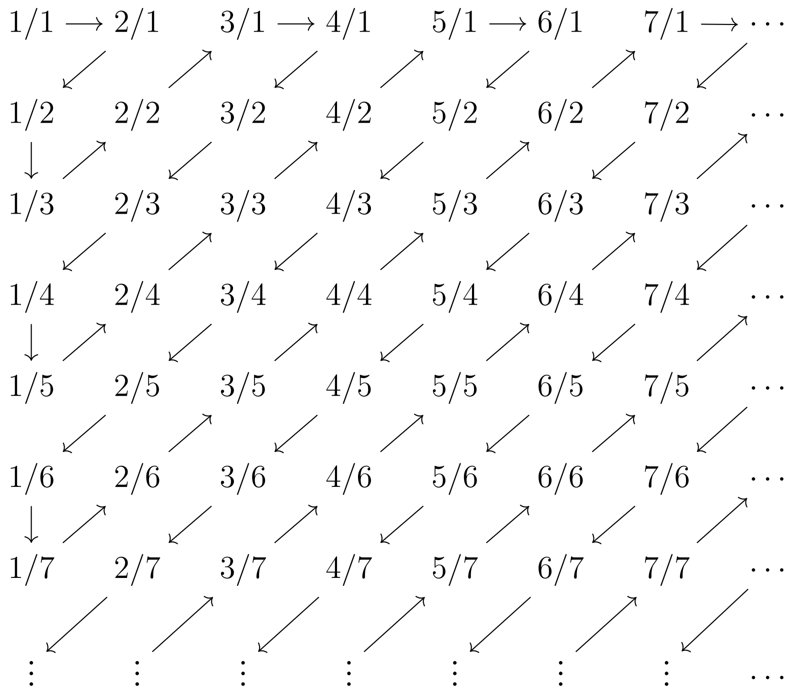


Figure 5.14: **Cantor diagonal method**

This table is constructed so that each rational number appears only once (in the sense of its decimal value) by diagonal hence the name of the method: "Cantor's diagonal".

If we eliminate from each diagonal the rational numbers that appear more than one time (the "equivalent fractions") in order to keep only those who are irreducible (i.e. those with the greatest common divisor of the numerator and denominator is equal to 1), then we can with this distinction define an application $f : \mathbb{N} \Rightarrow \mathbb{Q}$ that is injective (two distinct rational numbers have distinct ranks) and surjective (at any place will be written a rational number).

The application f is therefore bijective: $\mathbb{N}$ and $\mathbb{Q}$ are then effectively equipotent!

The definition a little bit more rigorous (and therefore less funny) of $\mathbb{Q}$ from $\mathbb{Z}$ is as follows (it is interesting to see the notation used):

On the set $\mathbb{Z} \times \mathbb{Z}\ \{0\}$, which should be read as the set constructed from two relative integer whose zero is excluded from the second one, we consider the relation $\mathcal{R}$ between two relative pairs of integers defined by:

$$(a, b)\mathcal{R}(a', b') \Leftrightarrow ab' = a'b \tag{5.2.42}$$

We then easily verify that R is an equivalence relation (see section Operators page 290) on $\mathbb{Z} \times \mathbb{Z} \setminus \{0\}$.

The set of equivalence classes for this relation $\mathcal{R}$ denoted then $(\mathbb{Z} \times \mathbb{Z} \setminus \{0\})\,/\mathcal{R}$ is by definition $\mathbb{Q}$. That is to say that we write therefore more rigorously:

$$\mathbb{Q} = (\mathbb{Z} \times \mathbb{Z} \setminus \{0\})/\mathcal{R} \tag{5.2.43}$$

The equivalence class $(a, b) \in \mathbb{Z} \times \mathbb{Z} \setminus \{0\}$ is explicitly denoted by:

$$\frac{a}{b} \tag{5.2.44}$$

in accordance with the notation that everyone is accustomed to use.

We easily check the addition and multiplication operations that were operations defined on $\mathbb{Z}$ pass without problems to $\mathbb{Q}$ by writing:

$$\frac{a}{b} \cdot \frac{p}{q} = \frac{a \cdot p}{b \cdot q} \quad \text{and} \quad \frac{a}{b} + \frac{p}{q} = \frac{aq + bp}{bq} \tag{5.2.45}$$

Moreover, these operations provide $\mathbb{Q}$ with the structure of a field (see section Set Theory page 429) with $\dfrac{0}{1}$ as neutral element for the addition and $\dfrac{1}{1}$ as neutral element for the multiplication. Thus, any non-zero element of $\mathbb{Q}$ is reversible, in fact:

$$\frac{a}{b} \cdot \frac{b}{a} = \frac{ab}{ab} = \frac{1}{1} \tag{5.2.46}$$

what is written also more technically:

$$(ab, ab)\mathcal{R}(1, 1) \tag{5.2.47}$$

Remark

Even if we want to define $\mathbb{Q}$ as the being the set $\mathbb{Z} \times \mathbb{Z} \setminus 0$ where $\mathbb{Z}$ represents the numerators and $\mathbb{Z} \setminus 0$ the denominators of the rationals, this is not possible because otherwise we would for example $(1,2) \neq (2,4)$ while we expect for an equality.

Hence the need to introduce an equivalence relation which enables us to identify, to return to the previous example, with $(1,2)$ and $(2,4)$. The relation $\mathcal{R}$ that we have defined does not fall from heaven, indeed the reader who handled the rational so far without ever having seen their formal definition knows that:

$$\frac{a}{b}=\frac{a'}{b'} \Leftrightarrow ab'=a'b \tag{5.2.48}$$

It is therefore almost natural to define the relation $\mathcal{R}$ as we have done. In particular, regarding the above example, $\dfrac{1}{2}=\dfrac{2}{4}$ because $(1,2)\mathcal{R}(2,4)$ and the problem is solved.

In addition to the historical circumstances of its establishment, this new entity (set) is distinguished from relative numbers because it induces the original and paradoxical concept of partial quantities. This notion that a priori does not make sense, find its place in the mind of man thanks to the geometry where the idea of fraction of length, of proportion are illustrated more intuitively.

5.2.2.4 Irrational Numbers

It can seem obvious to present irrational numbers before real number (see further below) but this can be explained by the fact that it is the order of the discovering in the human history and therefore it seems more pedagogical to us to present them in this order.

So the set of rational $\mathbb{Q}$ is limited and sadly not sufficient too. Indeed, we may think that all mathematical computation with commonly known operations are reduced to this set but it is not the case!

✍**Examples:**

E1. Let us calculate the square root of two which we denote $\sqrt{2}$ (thing to Pythagorean theorem with a triangle of side 1 and 1 then the third one is of size $\sqrt{2}$). Suppose it is a rational root. So if this is truly a rational, we should be able to express it as a/b, where by the definition of a rational a and b are integers with no common factors. For this reason, a and b can not both be even numbers. There are three remaining possibilities:

1. a is odd (then b is even)
2. a is even (then b is odd)
3. a is odd (then b is odd)

By squaring, we have:

$$\sqrt{2}=\frac{a}{b} \tag{5.2.49}$$

That can be written:

$$2b^2=a^2 \tag{5.2.50}$$

Since the square of an odd number is odd and the square of an even number is even, the case (1) is not possible because a^2 would be odd and $2b^2$ would be even.

If case (2) is also impossible, because then we could write $a=2c$, where c is any integer, and so if we take the square then we have $a^2=4c^2$ that is to say an even number on both sides of equality. Substituting in $2b^2=a^2$ we obtain after simplification that $b^2=2c^2$. Then b^2 would be odd while $2c^2$ would even.

The case (3) is also impossible because a^2 is then odd and $2b^2$ is even (that b is even or odd!).

There is the no solutions! That is to say that the start assumption is false and there does not two integers a and b such that $\sqrt{2}=a/b$.

E2. Let us prove by contradiction, that the famous Euler number e is irrational. To do this, remember that e (see section Functional Analysis page 2289) can also be defined by the Taylor series (see section Sequences and Series page 1703):

$$e=1+\frac{1}{1!}+\frac{1}{2!}+\frac{1}{3!}+\ldots+\frac{1}{n!}+\ldots \tag{5.2.51}$$

Then if e is rational, it could be written in the form p/q (with $q>1$, because we know that is not an integer). Let us multiply both sides of the equality by $q!$:

$$q!e=q!+\frac{q!}{1!}+\frac{q!}{2!}+\frac{q!}{3!}+\ldots+\frac{q!}{q!}+\frac{q!}{(q+1)!}+\frac{q!}{(q+2)!}+\ldots \tag{5.2.52}$$

The first member $q!e$ would then be an integer, because by definition of the factorial:

$$q!=q\cdot(q-1)\cdot(q-2)\ldots2\cdot1 \tag{5.2.53}$$

is an integer.

The first terms of the second member of the previous prior-previous relation, until the term $q!/q!=1$ are also integer because $q!/m!$ is simplified if $q>m$. So by subtraction we find:

$$q!e-\left(q!+\frac{q!}{1!}+\frac{q!}{2!}+\frac{q!}{3!}+\ldots+\frac{q!}{q!}+\ldots\right)=\frac{q!}{(q+1)!}+\frac{q!}{(q+2)!}+\ldots \tag{5.2.54}$$

when the right sequences should be an integer!

After simplification, the second member of the equality becomes:

$$\frac{1}{q+1}+\frac{1}{(q+1)(q+2)}+\ldots \tag{5.2.55}$$

the first term in this sum is strictly less than $1/2$, the second strictly less than $1/4$ second, the third strictly less than $1/8$, etc.

So, since each term is strictly less than the following harmonic series which converges to 1:

$$\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\ldots=1 \tag{5.2.56}$$

then therefore the sequence is not an integer as being strictly less than 1. This is a contradiction!

Thus, the rational numbers do not satisfy the numerical expression of $\sqrt{2}$ and e (to cite only these two particular examples).

They must therefore be complemented by the set of all numbers that can not be written as a fraction (the ratio of an integer dividend and an integer divisor without common factors) and that we name "irrational numbers". Finally we can say that:

Definition (#5.2.35): In mathematics, an "irrational number" is any real number that cannot be expressed as a ratio of integers. Irrational numbers cannot be represented as terminating or repeating decimals.

5.2.2.5 Real Numbers

Definition (#5.2.36): The union of rational and irrational numbers gives the set of "real numbers" that we denote by:

$$\mathbb{Q} \subset \mathbb{R} \tag{5.2.57}$$

Remark

Mathematicians in their usual rigour have different techniques to define real numbers. They use the properties of topology (among others) and especially Cauchy sequences but that's another story that goes beyond the formal scope of this section. For a "set point of view" definition of ℝ (the real numbers are often described as "the complete ordered field") the reader should report to the section on Set Theory page 429.

Below we can see a figure of most common sets of numbers (notice that the decimal set 𝔻 is omitted as it is taught mainly only in France):

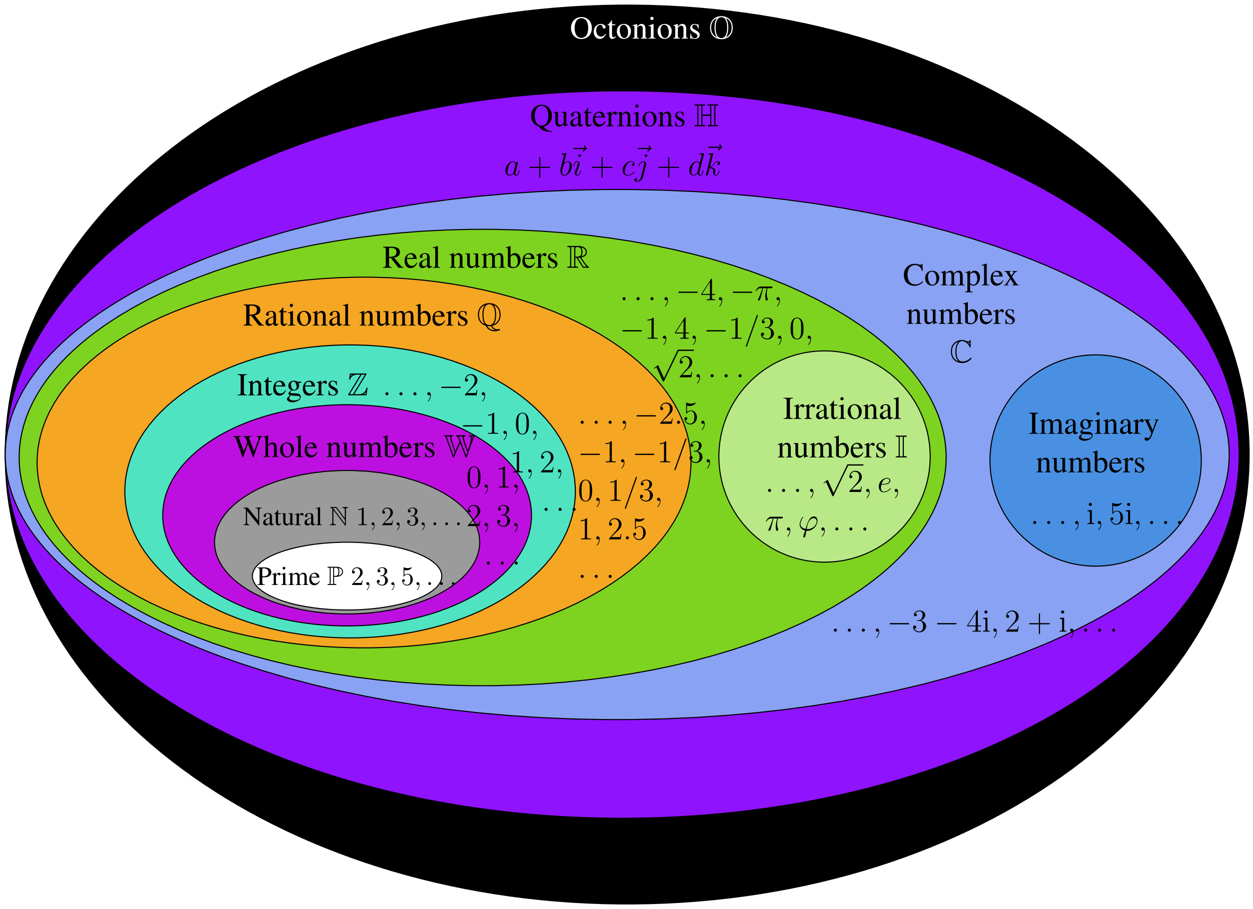

Octonions 𝕆
Quaternions ℍ
$a+b\vec{i}+c\vec{j}+d\vec{k}$
Real numbers ℝ
$\ldots,-4,-\pi,$
$-1,4,-1/3,0,$
$\sqrt{2},\ldots$
Complex numbers ℂ
Rational numbers ℚ
Integers ℤ $\ldots,-2,$
$-1,0,$
$1,2,$
$\ldots$
Whole numbers 𝕎
$0,1,$
$2,3,$
$\ldots$
Natural ℕ $1,2,3,\ldots$
Prime ℙ $2,3,5,\ldots$
$\ldots,-2.5,$
$-1,-1/3,$
$0,1/3,$
$1,2.5$
$\ldots$
Irrational numbers 𝕀
$\ldots,\sqrt{2},e,$
$\pi,\varphi,\ldots$
Imaginary numbers
$\ldots,\mathrm{i},5\mathrm{i},\ldots$
$\ldots,-3-4\mathrm{i},2+\mathrm{i},\ldots$

Figure 5.15: **Simple numbers sets summary (transcendent and algebraic numbers are not represented)**

Obviously we are led to ask ourselves whether ℝ is countable or not. In other words: can pure science - i.e. mathematics - quantify the unquantifiable? The answer and the proof is quite simple.

Proof 5.3.5. By definition, we have seen above that there must be a bijective correspondence between $\mathbb{Q}$ and $\mathbb{R}$ to that $\mathbb{R}$ is countable.

For simplicity, we will show that the interval $[0, 1[$ is then not countable. This will involve of course by extension that $\mathbb{R}$ is not countable!

The elements of this interval are represented by infinite sequences between 0 and 9 (in the decimal system):

- Some of these suites are zero from starting from a given rank, some not.
- So we can identify $[0, 1[$ to the set of all sequences (finite or infinite) of integers between 0 and 9.

$n°1$	x_{11}	x_{12}	x_{13}	x_{14}	$\cdots$	x_{1p}	$\cdots$
$n°2$	x_{21}	x_{22}	x_{23}	x_{24}	$\cdots$	x_{2p}	$\cdots$
$n°3$	x_{31}	x_{32}	x_{33}	x_{34}	$\cdots$	x_{3p}	$\cdots$
$n°4$	x_{41}	x_{42}	x_{43}	x_{44}	$\cdots$	x_{4p}	$\cdots$
$n°5$	x_{51}	x_{52}	x_{53}	x_{54}	$\cdots$	x_{5p}	$\cdots$
$n°6$	x_{61}	x_{62}	x_{63}	x_{64}	$\cdots$	x_{6p}	$\cdots$
	$\cdots$						
		$\cdots$					
			$\cdots$				
$n°k$	x_{k1}	x_{k2}	x_{k3}	x_{k4}	$\cdots$	x_{kp}	$\cdots$
						$\cdots$	

If this set was countable, we could classify these sequences (with a first, second, etc.). Thus, the sequence $x_{11}x_{12}x_{13}x_{14}\ldots x_{1p}\ldots$ would be classified first and so on ... as proposed in the above table.

We could then edit this infinite matrix as follows: to each element of the diagonal, we add 1, according to the rule: $0 + 1 = 1, 1 + 1 = 2, 8 + 1 = 9$ and $9 + 1 = 0$:

$n°1$	$x_{11}+1$	x_{12}	x_{13}	x_{14}	$\cdots$	x_{1p}	$\cdots$
$n°2$	x_{21}	$x_{22}+1$	x_{23}	x_{24}	$\cdots$	x_{2p}	$\cdots$
$n°3$	x_{31}	x_{32}	$x_{33}+1$	x_{34}	$\cdots$	x_{3p}	$\cdots$
$n°4$	x_{41}	x_{42}	x_{43}	$x_{44}+1$	$\cdots$	x_{4p}	$\cdots$
$n°5$	x_{51}	x_{52}	x_{53}	x_{54}	$\cdots$	x_{5p}	$\cdots$
$n°6$	x_{61}	x_{62}	x_{63}	x_{64}	$\cdots$	x_{6p}	$\cdots$
	$\cdots$						
		$\cdots$					
			$\cdots$				
$n°k$	x_{k1}	x_{k2}	x_{k3}	x_{k4}	$\cdots$	x_{kp}	$\cdots$
						$\cdots$	

Then let us consider the sequence on the diagonal:

- It cannot be equal to the first sequence of the first row in the prior-previous table since it is distinguished at least by the first element.

- It cannot be equal to the second sequence of the second row of the prior-previous table since is distinguished at least by the second element.
- It cannot be equal to the third sequence of the second row of the prior-previous table since is distinguished at least by the third element.

and so on ... It the cannot be equal to any of the sequences in this table!

So whatever the chosen classification of infinite sequences of $0 \ldots 9$, there is always one who escapes this classification! So it is that it is impossible to number them ... simply because they do not form a countable set (i.e. $\mathbb{R} \not\approx \mathbb{N}$)!

■ Q.E.D.

The technique that has allowed us to achieve this result is known as the "Cantor diagonal process" (because similar to that used for equipotence between the natural and rational set) and the set of real numbers is said to have the "power of the continuum" by the fact that it is uncountable.

We may assume that it is intuitive for the reader that any real number can be approximated infinitely close by a rational number (for irrational numbers we simply stop at a given number of decimals and find the corresponding rational). Mathematicians say therefore that $\mathbb{Q}$ is "dense" in $\mathbb{R}$ and denote this by:

$$\bar{\mathbb{Q}} = \mathbb{R} \qquad (5.2.58)$$

To see this (in the mathematician way of life...) let us take a real number x. For any $n \in \mathbb{N}$ we denote by:

$$p_n = [10^n x] \qquad (5.2.59)$$

the integer part of x that was previously multiplied by 10^n.

By definition of the integer part of a real number, we have:

$$p_n \leq 10^n x < p_n + 1 \qquad (5.2.60)$$

This inequality is equivalent to:

$$\frac{[10^n x]}{10^n} \leq x < \frac{[10^n x] + 1}{10^n} \qquad (5.2.61)$$

Let us put for all $n \in \mathbb{N}$:

$$a_n = \frac{[10^n x]}{10^n} \qquad \text{and} \qquad b_n \frac{[10^n x] + 1}{10^n} \qquad (5.2.62)$$

where a_n and b_n are respectively the approached decimal values to $x \cdot 10^{-n}$ by default or by excess.

Notice that $\forall n \in \mathbb{N}$ we have:

$$b_n - a_n = 10^{-n} \qquad (5.2.63)$$

and that $\forall n \in \mathbb{N}$ we have $\{a_n, b_n\} \in \mathbb{Q}$.

✍**Example:**

We want to calculate $\sqrt{2}$. Then:

n		a_n	b_n	error
1	$1 < \sqrt{2} < 2$	1	2	1
2	$1.4 < \sqrt{2} < 1.5$	1.4	1.5	0.1
3	$1.41 < \sqrt{2} < 1.42$	1.41	1.42	0.01
4	$1.414 < \sqrt{2} < 1.415$	1.414	1.415	0.001

Theorem 5.4. If the sequences (a_n) and (b_n) are adjacents, that is to say that (a_n) is ascending and (b_n) descending then we have:

$$b_n - a_n \xrightarrow[k\to\infty]{} 0 \tag{5.2.64}$$

Proof 5.4.1. Given $n \in \mathbb{N}$, let us recall that:

$$p_n \leq 10^n x < p_n + 1 \tag{5.2.65}$$

where $p_n = [10^n x]$. Let us multiply both sides by 10 then we get:

$$10p_n \leq 10^{n+1}x < 10(p_n + 1) \tag{5.2.66}$$

Then:

- We have quite obviously:

  $$10p_n \leq p_{n+1} \tag{5.2.67}$$

  hence:

  $$\frac{p_n}{10^n} \leq \frac{p_{n+1}}{10^{n+1}} \tag{5.2.68}$$

  and therefore $a_n \leq a_{n+1}$. The sequence (a_n) is therefore ascending.

- And we also have quite obviously:

  $$1 + p_{n+1} < 10p_{n+1} \tag{5.2.69}$$

  hence:

  $$\frac{1 + p_{n+1}}{10^{n+1}} \leq \frac{p_{n+1}}{10^n} \tag{5.2.70}$$

  and therefore $b_{n+1} \leq b_n$. The sequence (b_n) is therefore descending.

As:

$$b_n - a_n = 10^{-n} \tag{5.2.71}$$

We have indeed:

$$b_n - a_n \underset{k\to\infty}{\longrightarrow} 0 \tag{5.2.72}$$

and therefore we have proved that the both sequences (a_n) and (b_n) are adjacents. They are convergent and converge both to a same limit $L \in \mathbb{R}$.

As $\forall n \in \mathbb{N}$:

$$p_n \leq 10^n x < p_n + 1 \tag{5.2.73}$$

We have necessarily $L = x$.

■ Q.E.D.

Corollary 5.4.1. An real number is a limit of a sequence of rational numbers.

Proof 5.4.2. This is an immediate consequence of the previous theorem. Indeed, the sequences (a_n) and (b_n) are sequences of rationals.

■ Q.E.D.

5.2.2.5.1 Percentages

In business[13] (but not only obviously!) it is of usage with real numbers to communicate in percentages or per-thousand when we have to deal with ratios.

Definitions (#5.2.37):

D1. Given a scalar $x \in \mathbb{R}$ then expressed in "percentage" it will be denoted by:

$$x\% = x \cdot 100 \tag{5.2.74}$$

D2. Given a scalar $x \in \mathbb{R}$ then expressed in "per-thousand" it will be denoted by:

$$x‰ = x \cdot 1,000 \tag{5.2.75}$$

✍**Examples:**

E1. The scalar 0.5 expressed in percentage is given by:

$$0.5 \cdot 100 = 50\%$$

E2. The scalar 1 expressed in percentage is given by:

$$1 \cdot 100 = 100\%$$

E3. The percentage 75% expressed as a scalar is given by:

$$\frac{75}{100} = 0.75$$

---

[13] For interest rates, discounts, bonus, etc.

But we also have another well known percentage...:

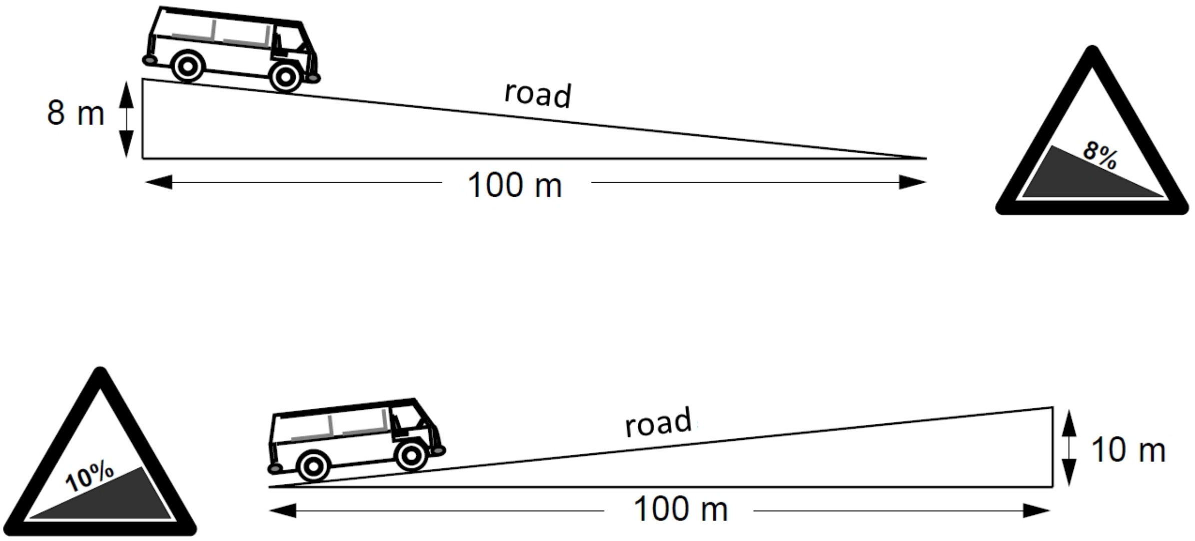


Figure 5.16: **Percentage usage illustration**

Remark

Notice something funny about percentages! We have:

$$x\% \text{ of } y = y\% \text{ of } x$$

That's simply because:

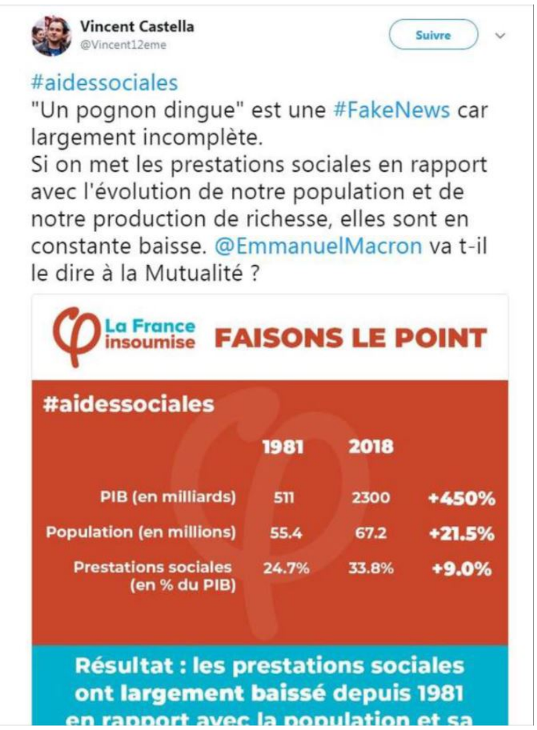
$$y\frac{x}{100} = x\frac{y}{100} \tag{5.2.76}$$

So, in order to calculate a percentage in your head you might want to turn it around: What's 2% of 50? It's 50% of 2.

You must be especially carefully when manipulating percentages (that are relative values for recall!) and even more when you compare that to absolute values. A famous recent "fake news" (in June 12018 according to holocene calendar) made by a socialist activist (supposing to be graduate in engineering...) in France on Twitter (a social media well-known for its fake news) made a lot of people laugh... but at the same time a huge number of people believed that the "fact" was true[14]. Here is the communication (try to guess where is the error, and notice that the author of the post says that France government communicates "fake news" but in reality... he is the guy who communicated the "fake news"... LoL):

---

[14] Serious newspapers, like *Le Monde*, also relayed the information as obviously wrong



Vincent Castella
@Vincent12eme

#aidessociales
"Un pognon dingue" est une #FakeNews car largement incomplète.
Si on met les prestations sociales en rapport avec l'évolution de notre population et de notre production de richesse, elles sont en constante baisse. @EmmanuelMacron va t-il le dire à la Mutualité ?

La France insoumise **FAISONS LE POINT**

#aidessociales

	1981	2018	
PIB (en milliards)	511	2300	+450%
Population (en millions)	55.4	67.2	+21.5%
Prestations sociales (en % du PIB)	24.7%	33.8%	+9.0%

Résultat : les prestations sociales ont largement baissé depuis 1981 en rapport avec la population et sa

Ok let us check if really the government welfare has indeed decrease in France while the population and the Gross Domestic Product (GDP) has at the same time increased as the author of this Tweet argue...

First we have (if we do the calculations correctly):

$$\text{Increase of GDP} = \frac{2300 - 511}{511} \cdot 100 \cong +350.1\% \qquad (5.2.77)$$

Secondly:

$$\text{Increase of Population} = \frac{67.2 - 55.4}{55.4} \cdot 100 \cong +21.3\% \qquad (5.2.78)$$

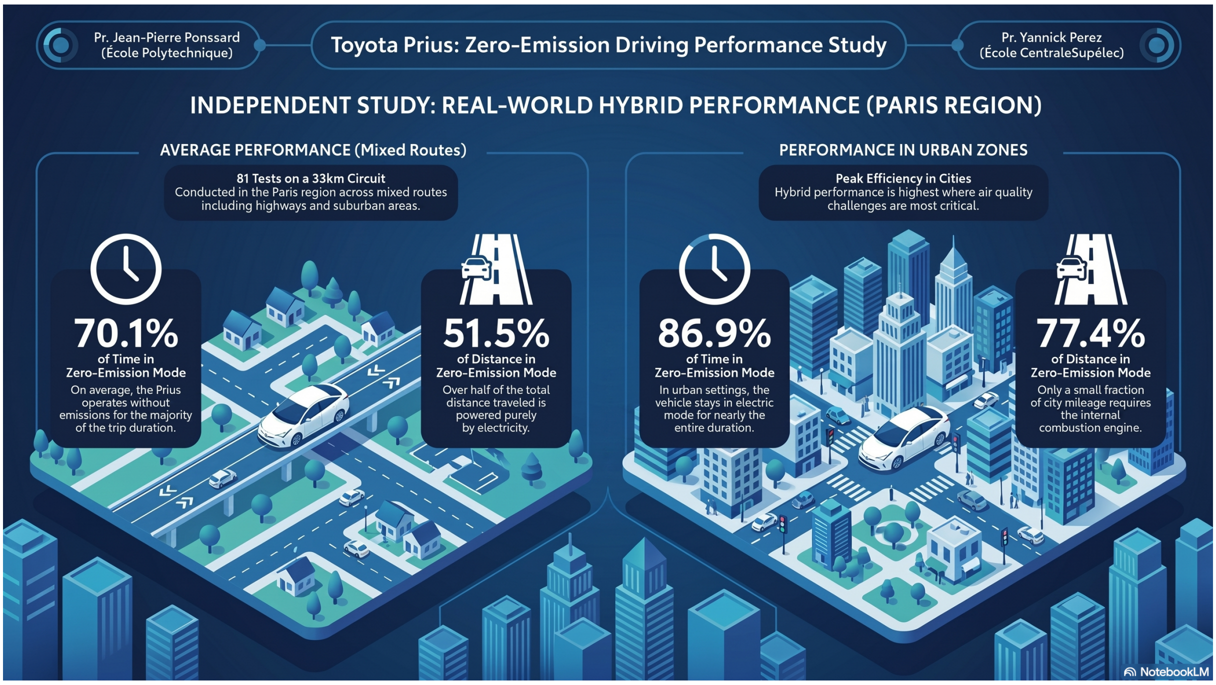
Finally:

$$\text{Increase social welfare} = \frac{33.8 - 24.7}{24.7} \cdot 100 \cong +36.84\% \qquad (5.2.79)$$

You see the error now...? The author of the tweet calculates for the welfare the increase in what we name in finance "basis points", and not in "percentages" as he does for the two firsts. This a very nice example on how to use statistics to lie... (lies works only with scientific illiterate people obviously!).

We don't need to say that this "tweet" was removed a few days after by his author from Twitter...

Another interesting example of pseudo-deception with percentages is the case of Toyota, which communicated the performance of its cars in electric mode in terms of time rather than distance. An independent study has shown afterward that in reality:



Pr. Jean-Pierre Ponssard (École Polytechnique)

Toyota Prius: Zero-Emission Driving Performance Study

Pr. Yannick Perez (École CentraleSupélec)

INDEPENDENT STUDY: REAL-WORLD HYBRID PERFORMANCE (PARIS REGION)

AVERAGE PERFORMANCE (Mixed Routes)

81 Tests on a 33km Circuit
Conducted in the Paris region across mixed routes including highways and suburban areas.

70.1% of Time in Zero-Emission Mode
On average, the Prius operates without emissions for the majority of the trip duration.

51.5% of Distance in Zero-Emission Mode
Over half of the total distance traveled is powered purely by electricity.

PERFORMANCE IN URBAN ZONES

Peak Efficiency in Cities
Hybrid performance is highest where air quality challenges are most critical.

86.9% of Time in Zero-Emission Mode
In urban settings, the vehicle stays in electric mode for nearly the entire duration.

77.4% of Distance in Zero-Emission Mode
Only a small fraction of city mileage requires the internal combustion engine.

You must also be extremely careful relatively to mass-media scientific communications (and even more on social networks...) when you read something like « ... *product X increase deadly syndrome Y of* 400%... ». Because sometimes, the increase could be (random example) in fact of 2 individual on $100,000$ to 8 individuals (still on $100,000$ obviously!). This is a way to bias our brain associating huge percentage to quantities of small magnitudes... So in such situations always look for the original absolute values and to corresponding the risk ratio (see section Numerical Methods page 6503).

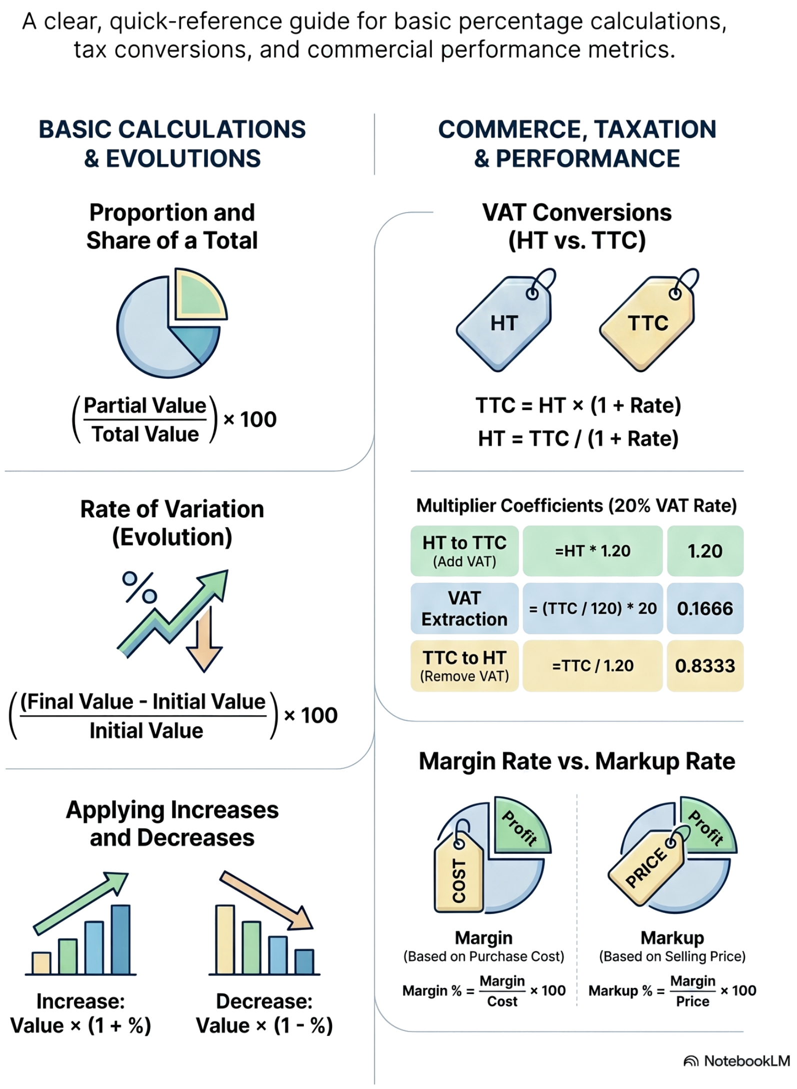
Another famous academic example on how tricky the manipulation of percentages can be is the so named "potato paradox". The paradox is stated as: «*You have* 100 *[kg] of potatoes, which are* 99% *water by weight. You let them dehydrate until they're* 98% *water. How much do they weight now?*» The result using algebra is that after evaporating of the water, the remaining total quantity, x, contains 1 [kg] of pure potatoes and $(98/100)x$ of water. The relation becomes:

$$1+\frac{98}{100} \Rightarrow 1=\frac{1}{50}x \tag{5.2.80}$$

resulting in $x=50$ [kg].

A last one used by some dishonest politicians, managers and marketing specialists is to say that an economic indicator that should be with a high value had increased of $X\%$ a few years before, and that now thanks to them it has decrease also of the same percentage $X\%$ and that we fall back then on the same value as before and like this nothing has changed! But we let the reader check if $C_0(1+X\%)(1-X\%)=C_0$ or not....

Here below is a summary of the most basic well known percentage calculations that any people having a family or any worker will use hundred of even thousands of times in its life:



A clear, quick-reference guide for basic percentage calculations, tax conversions, and commercial performance metrics.

BASIC CALCULATIONS & EVOLUTIONS
Proportion and Share of a Total
(Partial Value / Total Value) × 100
Rate of Variation (Evolution)
%
((Final Value - Initial Value) / Initial Value) × 100
Applying Increases and Decreases
Increase: Value × (1 + %)
Decrease: Value × (1 - %)
COMMERCE, TAXATION & PERFORMANCE
VAT Conversions (HT vs. TTC)
HT
TTC
TTC = HT × (1 + Rate)
HT = TTC / (1 + Rate)
Margin Rate vs. Markup Rate
COST
Profit
PRICE
Profit
Margin
(Based on Purchase Cost)
Margin % = Margin / Cost × 100
Markup
(Based on Selling Price)
Markup % = Margin / Price × 100
NotebookLM

Multiplier Coefficients (20% VAT Rate)

HT to TTC (Add VAT)	=HT * 1.20	1.20
VAT Extraction	= (TTC / 120) * 20	0.1666
TTC to HT (Remove VAT)	=TTC / 1.20	0.8333

Also keep in mind that in practice it is frequent to normalize a series of values $x_1, \ldots, x_n$ with $x_i \in \mathbb{R}$ (considered as a vector $\vec{x}$) into percentages, such that $x_{p,i} \in [0, 1]$ using the simple obvious following transformation[15] named a "Min-Max scaling":

$$\boxed{x_{p,i} = \frac{x_i - \min(\vec{x})}{\max(\vec{x}) - \min(\vec{x})}} \tag{5.2.81}$$

This is especially well known in spreadsheet software like Microsoft® Excel® when using Normalized bar charts, or normalized key performance indicators (conditional formatting), or also in Data Science where plenty of other transformations (see page 6772) are made for the use of some neural networks algorithms for example.

---

[15] The reader must avoid the confusion between "normalizing a variable" and "center reducing a random variable (i.e. standardizing)" to make it a $Z = \mathcal{N}(0, 1)$ statistical variable as we will see further below!

Transformation	Use Case	Equation / Calculation Method
Log Transformation	Stabilizing variance, normalizing data, useful for data with exponential changes (e.g., economic data).	$y = \log(x)$
Square Root Transformation	Stabilizing variance, normalizing data, suitable for data sets with non-negative values (counts, areas).	$y = \sqrt{x}$
Box-Cox Transformation	Generalized form for stabilizing variance, making data more normal-like, applicable in various scenarios.	$y(\lambda) = \frac{x^\lambda - 1}{\lambda}$ for $x > 0$ and $\lambda \neq 0$
Z-Score/Standard Score	Standardizing data to have mean of 0 and standard deviation of 1, used in outlier detection and normalization.	$z = \frac{x-\mu}{\sigma}$
Min-Max Scaling	Scaling data to a fixed range $(0$-$1)$, useful in scale-sensitive algorithms (neural networks, k-NN).	$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$
Normalization (L1, L2 norms)	Scaling individual samples, essential for algorithms needing comparable scales (support vector machines).	$\lVert\vec{x}\rVert_2 = \sqrt{\sum_{i=1}^{n} \lvert x_i \rvert^2}$
Difference Transformation	Making time-series data stationary by subtracting previous observation from the current one.	$\Delta x_t = x_t - x_{t-1}$
Categorical Encoding	Converting categorical data into a numeric format for use in mathematical models (One-Hot, Label Encoding).	One-Hot: Binary vectors, Label: Integer encoding
Binning/Discretization	Transforming continuous variables into discrete bins, simplifying complex relationships in data.	Interval: Equal-Width Quantile: Equal Frequency Count Tree-Based: Decision Tree
...	...	...

Table 5.16: **Some well known basic variables transformations**

5.2.2.6 Transfinite Numbers

We now are with an infinity of real numbers which is different from that of natural numbers. Cantor then dared what no one had dared since Aristotle: the positive integers sequence is also infinite, the set $\mathbb{N}$, is then a set that has a countable infinity of elements, then he said that the cardinal (see section Set Theory page 388) of this set was a number that existed as such, without that we need to use the tote symbol ∞, he denoted it:

$$\aleph_0 = \text{Card}(\mathbb{N}) \tag{5.2.82}$$

This symbol is as we know (see section Set Theory page 390) the first letter of the Hebrew alphabet, pronounced "aleph zero". Cantor was going to name this strange number, a "transfinite number".

The decisive act is to assert that there is, after the finite, a transfinite, that is to say an unlimited scale of determined modes which by nature are infinite, and yet can be specified, as for the finite, by specific numbers, well defined and distinguishable from each other!! This tool was necessary as a set cardinal can be equal to one of its parts as we will see just below!

After this first stroke going against most ideas for over two thousand years, Cantor would continue its path and build the calculation rules, paradoxical at first glance, of the transfinite numbers. These rules were based, as we said earlier, on the fact that two infinite sets are equivalent if there exists a bijection between the two sets.

Thus, we can easily show that the infinity of even numbers is equivalent to the infinity of integers: for this, it suffices to show that for every integer, we can associate an even number, his double, and vice versa. Therefore the cardinal of integers is equal to those of even numbers (the cardinal of a set can be equal to one of its parts!).

Thus, although if even numbers are included in the set of integers, there is an infinity α_0 of them, the two sets are equipotent. By stating that a set can be equal to one of its parts, Cantor goes against what seemed obvious to Aristotle and Euclid: the set of all sets is infinite! This will shake the whole of mathematics and will bring the Zermelo-Fraenkel axiomatic that we will see in the section of Set Theory at page 382.

From the above, Cantor define the following calculations rules on the cardinals:

$$\boxed{\aleph_0+1=\aleph_0 \qquad \aleph_0+\aleph_0=\aleph_0 \qquad \aleph_0^2=\aleph_0} \tag{5.2.83}$$

At first glance these rules seem non-intuitive, but in fact they are! Indeed, Cantor defined the addition of two transfinite numbers as cardinal of the disjoint union of the corresponding sets.

✍**Examples:**

E1. By denoting $\aleph_0$ the cardinal of $\mathbb{N}$ we have $\aleph_0+\aleph_0$ which is equivalent to saying that we sum the cardinal of $\mathbb{N}$ disjoint union $\mathbb{N}$. But as $\mathbb{N}$ disjoint union $\mathbb{N}$ is equipotent to $\mathbb{N}$ then $\aleph_0+\aleph_0=\aleph_0$ (it is enough to be convinced to take the set of odd and even integers which are both countable and which disjoint union is also countable).

E2. Other trivial example: $\aleph_0+1$ corresponds to the cardinality of $\mathbb{N}$ union a point. This set is still equipotent to $\mathbb{N}$ therefore $\aleph_0+1=\aleph_0$.

We will also during our study of the section Set Theory (page 391) that the concept of Cartesian product of two countable sets is such that we have:

$$\text{Card}(\mathbb{N}\times\mathbb{N})=\text{Card}(\mathbb{N}^2)=[\text{Card}(\mathbb{N})]^2 \tag{5.2.84}$$

and therefore:

$$\aleph_0^2=\aleph_0 \tag{5.2.85}$$

Similarly (see section Set Theory page 383) since $\mathbb{Z}=\mathbb{Z}^+\cup\mathbb{Z}^-$ we have:

$$\aleph_0+\aleph_0=\aleph_0 \tag{5.2.86}$$

and identifying $\mathbb{Q}$ to $\mathbb{Z}\times\mathbb{Z}$ (ratio of a numerator over denominator) we have immediately:

$$\aleph_0\times\aleph_0=(\aleph_0)^2=\aleph_0 \tag{5.2.87}$$

We can also prove an interesting statement: if we consider the cardinality of the set of all the cardinals, it is necessarily greater than all the cardinals, including itself (it is better to have read previously the section of Set Theory page 378)! In other words: the cardinality of the set of all sets of A is greater than the cardinal of A itself.

This implies that there is no set containing all sets since there is always a bigger one (it is an equivalent form of the famous old Cantor's paradox)!!!

In technical language it means considering a non-empty set A and then to state that:

$$\text{Card}(A)\leq\text{Card}(\mathcal{P}(A)) \tag{5.2.88}$$

where $\mathcal{P}(A)$ is the set of subsets of A.

That is to say, by definition of the order relation $<$ (strictly less), it suffices to prove that there is no surjective application $f:A\mapsto\mathcal{P}(A)$, in other words that to each element of the set of parts of A it does not match at least one pre-image in A.

Remark

The set $\mathcal{P}(\mathbb{N})$ for example consists of the set of even numbers, odd numbers, natural numbers, as well as the empty set itself, etc. $\mathcal{P}(\mathbb{N})$ is therefore the set of all "potatoes" (to borrow the vocabulary of high school ...) that make $\mathbb{N}$.

Proof 5.4.3. Suppose that we can number each potato of $\mathcal{P}(A)$ with at least one element of A (imagine that with $\mathbb{N}$). In other words it is equivalent to suppose that $f : A \mapsto \mathcal{P}(A)$ is surjective and let us consider a subset E of A such that:

$$E = \{x \in A | x \notin f(x)\} \tag{5.2.89}$$

that is to say the set of elements x of A that do not belong to the set numbered by x (the element x does not belong to the "potato" that it numbers in other terms...).

Or, if f is surjective it must also be a $y \in A$ for this subset E such that:

$$f(y) = E = \{x \in A | x \notin f(x)\} \tag{5.2.90}$$

since E is also a subset of A.

Suppose that y belongs to E. In this case, by definition of E, $y \notin f(y) = E$ (by definition of E that applies for every x and x can also be obviously y or z or doesn't matter what). By consequence, $y \notin E$, but in this second case, always by definition of E, $y \in f(y) = E$ (as y is not in E). We see therefore that the element y cannot exists and therefore f cannot be surjective.

We strongly recommend the reader to review the previous sentence more than on time if necessary.

■ Q.E.D.

Remark

David Hilbert famously argued that infinity cannot exist in physical reality. The consequence of this statement still under debate today. However... infinity has countless practical applications! Here's one way to think about it: do negative numbers have any practical applications? I mean you can't really have a negative amount of anything, can you? You can't have negative five apples! But negative number are useful in accounting! You can also take limits of functions as the variable goes to infinity. In this way you can calculate things like terminal velocity and escape velocity and many others! Infinity can be kinda observed as the more distant an object is the more the light it emits will be red-shifted, i.e. the wavelength tends to infinity as the object approach the Hubble sphere horizon (or maybe one day in the future we will be able to observe an object entering the horizon of a Black Hole where the same kind of effect occurs).

Engineers consider infinity ∞ just a being what is too big for being measurable. As illustrated on the graduation (top left corner) of the multimeter below:

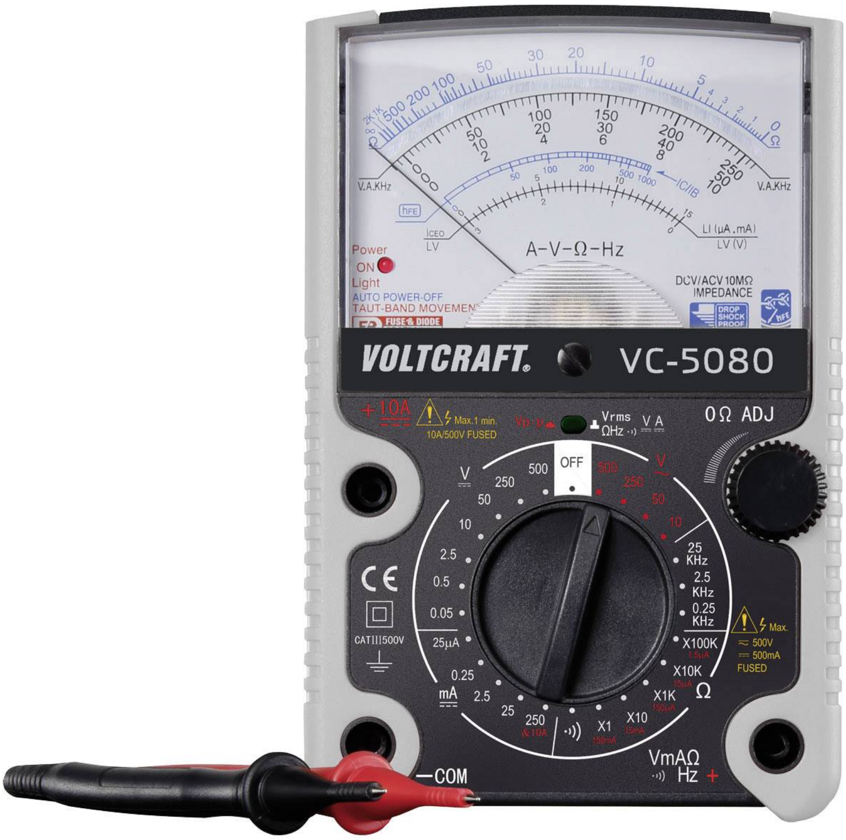


Figure 5.17: **VOLTCRAFT VC-5080 Multimeter with ohmic graduation displaying ∞ symbol**

5.2.2.7 Complex Numbers

Invented in the 116th century (holocene calendar) among others by Girolamo Cardano and Rafaello Bombelli, "complex numbers" (also named "imaginary numbers") are used to solve problems with no solutions in $\mathbb{R}$ and also used to mathematically formalize certain transformations in the plan such as rotation, similarity, translation, etc. and also to generalized some theorem restricted to $\mathbb{R}$ and therefore hiding some interesting results for practical engineering. For physicists, complex numbers above are also a very convenient way to simplify notations. It is thus very difficult to study wave phenomena, General Relativity or quantum mechanics without using complex numbers and expressions.

The importance and utility of complex numbers in biology, cosmology, physics, astrophysics, electronics, electromechanics, astrophysics, microelectronics, electrical engineering, statistics, optics, mechanical engineering, fluid mechanics etc. are almost never questioned by students and professionals as they appear very quickly in the corresponding courses (most of times just after thirty to fifty hours) or textbooks. However a significant percentage of students and professional in Finance and Insurance question the utility of complex numbers as their application cases appear in the fields most of the time only at the postgraduate level (PhD). So here is a small non exhaustive list of topics in Finance and Insurance where complex numbers appear and are really useful (most of the topics below are treated in the present book):

- In Fourier transforms for solving nonlinear partial differential equations (Black-Sholes for example)
- In Laplace transforms for solving nonlinear partial differential equations (Heston model)
- In Fourier transforms for spectral analysis of time series
- In solving polynomials of auto-regressive time series models
- In the use of the characteristic function in the CreditRisk First Boston model
- In PCA (Principal Component Analysis) analysis when IEEE rounding errors with `R` and Python™ produce complex eigenvalues

There are several ways to construct complex numbers. The first is typical of the construction way that mathematicians used as part of Set Theory. They define a couple of real numbers and define the operations between these couples to finally arrive at a meaning of the complex number concept. The second one is less rigorous but its approach is simpler and consist to define the pure unit imaginary number i and then build arithmetic operations from its definition. We will opt for the second method in the texts that will follow!

Definitions (#5.2.38):

D1. We define the "unit pure imaginary number" that we denote by i by the following property:

$$\boxed{\mathrm{i}^2=-1 \Leftrightarrow \mathrm{i}=\sqrt{-1}} \tag{5.2.91}$$

D2. A "complex number" is a pair of a real number a and an imaginary number $\mathrm{i}b$ and generally written in the following form:

$$\boxed{z = a + \mathrm{i}b} \tag{5.2.92}$$

where a and b are numbers belonging to $\mathbb{R}$.

D3. We note the set of complex numbers by $\mathbb{C}$ and therefore we have by construction:

$$\mathbb{R} \subset \mathbb{C} \tag{5.2.93}$$

Remark

The set $\mathbb{C}$ is identified to the oriented Euclidean plane E (see section Vector Calculus page 1830) thanks to the choice of a direct orthonormal basis (we therefore get the "Argand-Cauchy plane", also named "Gauss-Argand plane" or more commonly "Gauss plane" that we will see a little further below and that seems have be defined for the first time in 11806 according to holocene calendar).

The set of complex numbers that constitutes a field (see section Set Theory page 429) and denoted by $\mathbb{C}$, is defined (in a simple way to start) in the notation of set theory by:

$$\boxed{\mathbb{C} = \{z = (x + \mathrm{i}y) | x, y \in \mathbb{R}\}} \tag{5.2.94}$$

In other words we say that the field $\mathbb{C}$ is the field $\mathbb{R}$ to which we have added the imaginary number i. Which is formally denoted by:

$$\mathbb{R}[\mathrm{i}] \tag{5.2.95}$$

The addition and multiplication of complex numbers are internal operations to the set (field) of complex numbers (we will come back much more in detail on certain properties of complex numbers in the section of Set Theory page 421) and defined by:

$$\begin{aligned} z_1 + z_2 &= (x_1 + x_2) + \mathrm{i}(y_1 + y_2) \\ z_1 \cdot z_2 &= (x_1 x_2 - y_1 y_2) + \mathrm{i}(x_1 y_2 + x_2 y_1) \end{aligned} \tag{5.2.96}$$

The "real part" of z is traditionally denoted by:

$$\Re(z) = x \tag{5.2.97}$$

The "imaginary part" of z is traditionally denoted by:

$$\Im(z) = x \tag{5.2.98}$$

The "conjugate" of z is defined by:

$$\bar{z} = \overline{x + \mathrm{i}y} = x - \mathrm{i}y \tag{5.2.99}$$

and is sometimes also denoted z^* (particularly in quantum physics in some books!).

From a complex and its conjugate, it is possible to find its real and imaginary parts. These are the following obvious relations:

$$\Re(z) = \frac{z + \bar{z}}{2} \quad \text{and} \quad \Im(z) = \frac{z - \bar{z}}{2\mathrm{i}} \tag{5.2.100}$$

The "module" of z (or "norm") is the length from the center of the Gaussian plane (see further below a figure of the Gaussian plane) and is simply calculated using the Pythagorean theorem:

$$\boxed{|z|=\sqrt{x^2+y^2}=\sqrt{z\cdot\bar{z}}} \qquad (5.2.101)$$

and is always a positive number or equal to zero.

We will consider as obvious that it satisfies all the properties of a distance (see section of Topology page 2446 and Vector Calculus page 1794).

Remark

The notation $|z|$ for the module is not innocent since $|z|$ coincides with the absolute value of z when z is real.

The division between two complex number is calculated as (the denominator is obviously not zero):

$$\dfrac{z_1}{z_2}=\dfrac{x_1+\mathrm{i}y_1}{x_2+\mathrm{i}y_2}=\dfrac{x_1+\mathrm{i}y_1}{x_2+\mathrm{i}y_2}\cdot\dfrac{x_2-\mathrm{i}y_2}{x_2-\mathrm{i}y_2}=\dfrac{(x_1x_2+y_1y_2)-\mathrm{i}(x_1y_2-x_2y_1)}{x_2^2+y_2^2} \qquad (5.2.102)$$

The opposite of a complex number is calculated similarly:

$$\dfrac{1}{x+\mathrm{i}y}=\dfrac{x-\mathrm{i}y}{(x+\mathrm{i}y)(x-\mathrm{i}y)}=\dfrac{x-\mathrm{i}y}{x^2+y^2}=\dfrac{x}{x^2+y^2}-\mathrm{i}\dfrac{y}{x^2+y^2} \qquad (5.2.103)$$

We can therefore list 8 important properties of the module and the complex conjugate:

P1. We affirm that:

$$|z|=0\Leftrightarrow z=0 \qquad (5.2.104)$$

Proof 5.4.4. By definition of the module $|z|=\sqrt{x^2+y^2}$ so that the sum x^2+y^2 is zero, the necessary condition is that as $(x,y)\in\mathbb{R}$:

$$x=y=0 \qquad (5.2.105)$$

■ Q.E.D.

P2. We affirm that:

$$|z|=|-z|=|\bar{z}| \qquad (5.2.106)$$

Proof 5.4.5. This is immediate by:

$$|z|=\sqrt{x^2+y^2}=\sqrt{(-x)^2+(-y)^2}=|-z|=\sqrt{(x)^2+(-y)^2}=|\bar{z}| \qquad (5.2.107)$$

■ Q.E.D.

P3. We affirm that:

$$\begin{aligned}|\Re(z)|&\leqslant|z| \quad \text{with equality iif } z \text{ is real}\\|\Im(z)|&\leqslant|z| \quad \text{with equality iif } z \text{ is imaginary}\end{aligned} \qquad (5.2.108)$$

Proof 5.4.6. The two above inequalities can be written:

$$\begin{aligned}|x| &\leq \sqrt{x^2+y^2}\\ |y| &\leq \sqrt{x^2+y^2}\end{aligned}\tag{5.2.109}$$

thus equivalent respectively to:

$$\begin{aligned}x^2 &\leq \sqrt{x^2+y^2}\\ y^2 &\leq \sqrt{x^2+y^2}\end{aligned}\tag{5.2.110}$$

which are trivial. The rest of the proof is therefore trivial!

■ Q.E.D.

P4. We have:

$$\forall z_1, z_2 \in \mathbb{C}\times\mathbb{C} \qquad |z_1 z_2| = |z_1||z_2|\tag{5.2.111}$$

and if $z_2 \neq 0$:

$$\left|\frac{z_1}{z_2}\right| = \frac{|z_1|}{|z_2|}\tag{5.2.112}$$

Proof 5.4.7. First:

$$|z_1 z_2|^2 = (z_1 z_2)(\overline{z_1 z_2}) = (z_1 z_2)(\bar{z}_1\bar{z}_2) = (z_1\bar{z}_1)(z_2\bar{z}_2) = |z_1|^2|z_2|^2 \Rightarrow |z_1 z_2| = |z_1||z_2|\tag{5.2.113}$$

(we will prove a little further below that generally $\overline{z_1 z_2} = \bar{z}_1\bar{z}_2$) and for $z_2 \neq 0$:

$$\left|\frac{z_1}{z_2}\right|^2 = \frac{z_1}{z_2}\frac{\overline{z_1}}{z_2} = \frac{z_1}{z_2}\bar{z}_1\frac{\overline{1}}{z_2} = \frac{z_1}{z_2}\bar{z}_1\frac{1}{\bar{z}_2} = \frac{z_1\bar{z}_1}{z_2\bar{z}_2} = \frac{|z_1|^2}{|z_2|^2}\tag{5.2.114}$$

and taking root square this finish the proof.

■ Q.E.D.

P5. We affirm that:

$$|z|^2 = z\bar{z}\tag{5.2.115}$$

Proof 5.4.8. This is immediate:

$$|z|^2 = (\sqrt{x^2+y^2})^2 = (x+\mathrm{i}y)(x-\mathrm{i}y) = x^2 - \mathrm{i}xy + \mathrm{i}yx + y^2 = x^2+y^2 = \bar{z}z\tag{5.2.116}$$

■ Q.E.D.

P6. We affirm that:

$$\forall z, z' \in \mathbb{C} \quad \bar{\bar{z}} = z, \quad \overline{z+z'} = \bar{z}+\bar{z}', \quad \overline{zz'} = \bar{z}\bar{z}'\tag{5.2.117}$$

Proof 5.4.9. The first one is immediate:

$$\bar{\bar{z}}=\overline{\overline{x+\mathrm{i}y}}=\overline{x-\mathrm{i}y}=x+\mathrm{i}y=z \tag{5.2.118}$$

and:

$$\begin{aligned}\overline{z+z'}&=\overline{(x_1+\mathrm{i}y_1)+(x_2+\mathrm{i}y_2)}=\overline{(x_1+x_2)+\mathrm{i}(y_1+y_2)}=(x_1+x_2)-\mathrm{i}(y_1+y_2)\\&=(x_1-\mathrm{i}y_1)+(x_2-\mathrm{i}y_2)=\bar{z}+\bar{z}'\end{aligned} \tag{5.2.119}$$

and:

$$\begin{aligned}\overline{zz'}&=\overline{(x_1+\mathrm{i}y_1)(x_2+\mathrm{i}y_2)}=\overline{(x_1x_2-y_1y_2)+\mathrm{i}(x_1y_2+y_2x_2)}\\&=(x_1x_2-y_1y_2)+\mathrm{i}(x_1y_2+y_2x_2)=(x_1-\mathrm{i}y_1)(x_2-\mathrm{i}y_2)=\bar{z}\bar{z}'\end{aligned} \tag{5.2.120}$$

■ Q.E.D.

Remarks

R1. In mathematical terms, the first proof helps to show that complex conjugation is what is named an "involution" (in the sense that it is changing anything ...).

R2. Also in mathematical terms (it is only the vocabulary!), the second proof shows that the combination of the sum of two complex numbers is what we name a "group automorphism $(\mathbb{C},+)$" (see section Set Theory page 443).

R3. Again, for vocabulary ... the third proof show that the combination of the product of two complex numbers is what we name a "field automorphism $(\mathbb{C},+,\times)$" (see section Set Theory page 443).

P7. We affirm that for z different from zero:

$$\overline{\left(\frac{1}{z}\right)}=\frac{1}{\bar{z}}\qquad\overline{\left(\frac{z}{z'}\right)}=\frac{\bar{z}}{\bar{z}'} \tag{5.2.121}$$

Proof 5.4.10. We will restrict ourselves to the proof of the second relation that is a general case of the first (for $z=1$).

$$\begin{aligned}\overline{\left(\frac{z}{z'}\right)}&=\overline{\left(\frac{x_1+\mathrm{i}y_1}{x_2+\mathrm{i}y_2}\right)}=\overline{\left(\frac{x_1+\mathrm{i}y_1}{x_2+\mathrm{i}y_2}\right)\left(\frac{x_2-\mathrm{i}y_2}{x_2-\mathrm{i}y_2}\right)}=\overline{\frac{(x_1x_2+y_1y_2)+\mathrm{i}(y_1x_2-y_2x_1)}{x_2^2+y_2^2}}\\&=\frac{x_1x_2+y_1y_2}{x_2^2+y_2^2}-\mathrm{i}\frac{y_1x_2-y_2x_1}{x_2^2+y_2^2}=\frac{(x_1x_2+y_1y_2)+\mathrm{i}(y_2x_1-y_1x_2)}{x_2^2+y_2^2}\\&=\frac{(x_1-\mathrm{i}y_1)(x_2+\mathrm{i}y_2)}{x_2^2+y_2^2}=\frac{x_1-\mathrm{i}y_1}{x_2-\mathrm{i}y_2}\frac{x_2+\mathrm{i}y_2}{x_2+\mathrm{i}y_2}=\frac{\bar{z}}{\bar{z}'}\end{aligned} \tag{5.2.122}$$

■ Q.E.D.

P8. We have:

$$|z_1+z_2|\leqslant|z_1|+|z_2| \tag{5.2.123}$$

for any complex number z_1, z_2 (strictly speaking non-zero complex numbers, otherwise the concept of argument of the complex number that we will see further below is undetermined). Furthermore the equality holds if and only if z_1 and z_2 are collinear (the vectors are "on the same straight line") and of the same direction, in other words.... if it exist $\lambda \in \mathbb{R}$ such as $\lambda z_1 = z_2$.

Proof 5.4.11. Directly we have:

$$|z_1+z_2|^2 = |z_1|^2 + 2\Re(z_1\bar{z}_2) + |z_2|^2 \leqslant |z_1|^2 + 2|z_1\bar{z}_2| + |z_2|^2 = (|z_1|+|z_2|)^2 \tag{5.2.124}$$

This inequality may not be obvious to everyone, therefore let us develop it a bit and let us assume it true:

$$\begin{aligned}
&|z_1+z_2|^2 \overset{?}{\leqslant} (|z_1|+|z_2|)^2\\
&\sqrt{(x_1+x_2)^2+(y_1+y_2)^2}^2 \overset{?}{\leqslant} \left(\sqrt{x_1^2+y_1^2}+\sqrt{x_2^2+y_2^2}\right)^2\\
&(x_1+x_2)^2+(y_1+y_2)^2 \overset{?}{\leqslant} (x_1^2+y_1^2)+2\sqrt{x_1^2+y_1^2}\sqrt{x_2^2+y_2^2}+(x_2^2+y_2^2)\\
&x_1^2+2x_1x_2+x_2^2+y_1^2+2y_1y_2+y_2^2 \overset{?}{\leqslant} (x_1^2+y_1^2)+2\sqrt{x_1^2+y_1^2}\sqrt{x_2^2+y_2^2}+(x_2^2+y_2^2)
\end{aligned} \tag{5.2.125}$$

After simplification:

$$\begin{aligned}
&x_1x_2+y_1y_2 \overset{?}{\leqslant} \sqrt{x_1^2+y_1^2}\sqrt{x_2^2+y_2^2}\\
&(x_1x_2+y_1y_2)^2 \overset{?}{\leqslant} (x_1^2+y_1^2)(x_2^2+y_2^2)\\
&x_1^2x_2^2+2x_1x_2y_1y_2+y_1^2y_2^2 \overset{?}{\leqslant} x_1^2x_2^2+x_1^2y_2^2+y_1^2x_2^2+y_1^2y_2^2
\end{aligned} \tag{5.2.126}$$

and again after simplification:

$$\begin{aligned}
&2x_1x_2y_1y_2 \overset{?}{\leqslant} x_1^2y_2^2+y_1^2x_2^2\\
&0 \overset{?}{\leqslant} x_1^2y_2^2-2x_1x_2y_1y_2+y_1^2x_2^2\\
&0 \overset{?}{\leqslant} (x_1y_2-y_1x_2)^2
\end{aligned} \tag{5.2.127}$$

So as the square brackets is necessarily positive or zero it follows:

$$0 \leq (x_1y_2-y_1x_2)^2 \tag{5.2.128}$$

This last relation thus shows that inequality is true.

■ Q.E.D.

Remark

In fact there is a more general form of this inequality named "Minkowski inequality", proved in the section of Vector Calculus page 1828 (complex numbers can indeed be written in the form of vectors as we will see later).

5.2.2.7.1 Geometric Interpretation of Complex Numbers

We can also represent any complex number $a + ib$ or $a - ib$ in a plane defined by two axes (two dimensions) of infinite length and orthogonal between them. The vertical axis represents the imaginary part of a complex number and the horizontal axis the real part (see figure below).

So there is correspondence between the set of complex numbers $\mathbb{C}$ and the set of vectors of the Gaussian plane $\mathbb{R}^2$.

We sometimes name this type of representation "Gauss plane" or "Gauss map" and we name "affix" the point with Cartesian coordinates (a, b) that is identified with the complex number $a + ib$ itself.

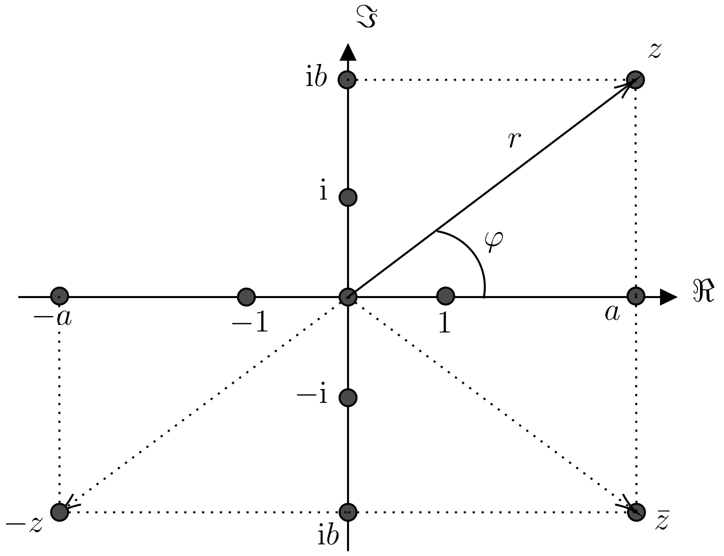


Figure 5.18: Complex Gauss plane

and then we write:

$$\text{Aff}(\vec{r}) = a + \mathrm{i}b \tag{5.2.129}$$

We see on this diagram that a complex number has thus a vector interpretation (see section Vector Calculus page 1797) given by:

$$z = a\begin{pmatrix}1\\0\end{pmatrix} + b\begin{pmatrix}0\\1\end{pmatrix} = a + \mathrm{i}b \tag{5.2.130}$$

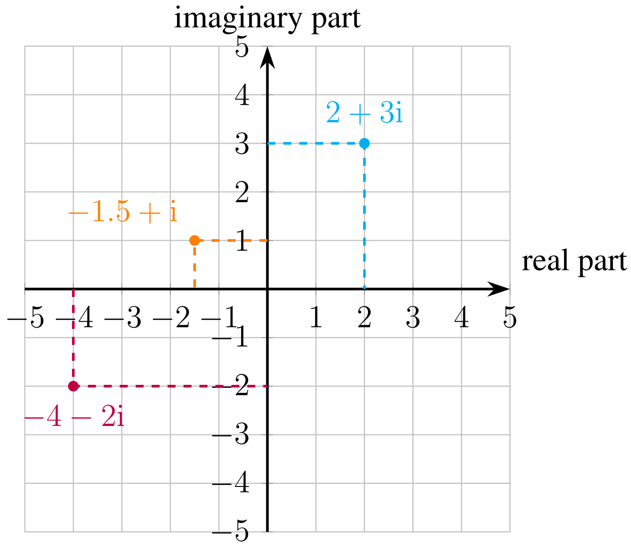
where the canonical basis is defined such as:

$$\left\{1 = \begin{pmatrix}1\\0\end{pmatrix}, \mathrm{i} = \begin{pmatrix}0\\1\end{pmatrix}\right\} \tag{5.2.131}$$

with:

$$r=|z|=\sqrt{a^2+b^2} \tag{5.2.132}$$

Thus, $\begin{pmatrix}1\\0\end{pmatrix}$ is the unitary basis vector of the carried by the horizontal axis $\mathbb{R}$ and $\begin{pmatrix}0\\1\end{pmatrix}$ is the unitary basis vector carried by the vertical imaginary axis $\mathbb{R}_\mathrm{i}$ and r is the module (the norm) that is positive or zero.



This has to be compared with the canonical vectors basis of $\mathbb{R}^2$ (see section Vector Calculus page 1810):

$$\vec{v}=x\vec{e}_1+y\vec{e}_2=x\begin{pmatrix}1\\0\end{pmatrix}+y\begin{pmatrix}0\\1\end{pmatrix}=\begin{pmatrix}x\\y\end{pmatrix} \tag{5.2.133}$$

with:

$$\|\vec{v}\|=\sqrt{x^2+y^2} \tag{5.2.134}$$

so that we can identify the complex plane with the Euclidean plane.

Thanks to the geometric interpretation of the Gaussian plane, the equality below is immediate for example and avoids making some developments:

$$\left|\frac{a+b\mathrm{i}}{b+a}\right|=1 \tag{5.2.135}$$

In addition, the definitions of the cosine and sine (see section Trigonometry page 2508) give us:

$$a=r\cos(\varphi) \qquad b=r\sin(\varphi) \tag{5.2.136}$$

Finally:

$$\boxed{\begin{aligned}&r=\sqrt{a^2+b^2}\\&\varphi^{-1}=\cos^{-1}\left(\frac{a}{r}\right)=\sin^{-1}\left(\frac{b}{r}\right)=\tan^{-1}\left(\frac{b}{a}\right)\end{aligned}} \tag{5.2.137}$$

Therefore:

$$z=a+\mathrm{i}b=r\cos(\varphi)+\mathrm{i}r\sin(\varphi)=r(\cos(\varphi)+\mathrm{i}\sin(\varphi))=r\mathrm{cis}(\varphi) \tag{5.2.138}$$

complex number which is always equal to itself modulo 2π by the properties of trigonometric functions:

$$\boxed{z=r(\cos(\varphi)+\mathrm{i}\sin(\varphi))=r(\cos(\varphi+2k\pi)+\mathrm{i}\sin(\varphi+2k\pi))} \tag{5.2.139}$$

with $k\in\mathbb{N}$ and where φ is named the "argument of z" and is traditionally denoted by:

$$\arg(z) \tag{5.2.140}$$

The properties of the cosine and sine (see section Trigonometry page 2514) lead us directly to write for the argument:

$$\arg(\bar{z})=-\arg(z)\quad\text{and}\quad\arg(-z)=\arg(z+\pi) \tag{5.2.141}$$

We also prove among other things with the Taylor series (see section Sequences and Series page 1704) that:

$$\cos(\varphi)=1-\frac{\varphi^2}{2!}+\frac{\varphi^4}{4!}-\ldots+(-1)^k\frac{\varphi^{2k}}{(2k)!}+\ldots \tag{5.2.142}$$

and:

$$\sin(\varphi)=\varphi-\frac{\varphi^3}{3!}+\frac{\varphi^5}{5!}-\ldots+(-1)^k\frac{\varphi^{2k+1}}{(2k+1)!}+\ldots \tag{5.2.143}$$

which sum is similar to:

$$e^{\varphi}=1+\varphi+\frac{\varphi^2}{2!}+\frac{\varphi^3}{3!}+\ldots+\frac{\varphi^k}{k!}+\ldots \tag{5.2.144}$$

but instead perfectly identical to the Taylor expansion of $e^{\mathrm{i}x}$:

$$e^{\mathrm{i}\varphi}=1+\mathrm{i}\varphi-\frac{\varphi^2}{2!}-\mathrm{i}\frac{\varphi^3}{3!}+\ldots+\mathrm{i}^k\frac{\varphi^k}{k!}+\ldots=\cos(\varphi)+\mathrm{i}\sin(\varphi) \tag{5.2.145}$$

So finally, we can write:

$$\boxed{z=r(\cos(\varphi)+\mathrm{i}\sin(\varphi))=re^{\mathrm{i}\varphi}} \tag{5.2.146}$$

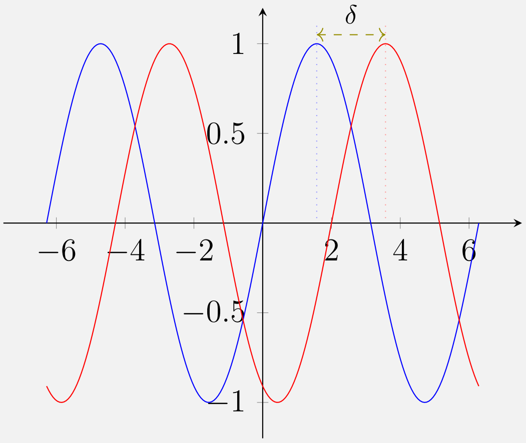
relation named "Euler's formula".

Remark

Notice that:

$$z=r(\cos(\varphi+\delta)+\mathrm{i}\sin(\varphi+\delta))=re^{\mathrm{i}(\varphi+\delta)}=re^{\mathrm{i}\varphi}e^{\mathrm{i}\delta} \tag{5.2.147}$$

where δ is named a "phase-shift" and $e^{\mathrm{i}\delta}$ is named a "phase-shift factor":



Using the properties of trigonometric functions:

$$\begin{aligned}\cos(\varphi)+\mathrm{i}\sin(\varphi)&=e^{\mathrm{i}\varphi}\\ \cos(\varphi)-\mathrm{i}\sin(\varphi)&=e^{-\mathrm{i}\varphi}\end{aligned}\tag{5.2.148}$$

Depending on we sum or subtract this gives us the "Euler formulas" or "de Moivre and Euler formulas":

$$\boxed{\begin{aligned}\cos(\varphi)&=\frac{e^{\mathrm{i}\varphi}+e^{-\mathrm{i}\varphi}}{2}\\ \sin(\varphi)&=\frac{e^{\mathrm{i}\varphi}-e^{-\mathrm{i}\varphi}}{2\mathrm{i}}\end{aligned}}\tag{5.2.149}$$

Notice that the angle can be a purely a complex number! This is to say that in all generality trigonometric functions can be considered as functions that go from $\mathbb{C}$ to $\mathbb{C}$.

Remark

Notice that these relations are very useful to linearise expressions such as $\cos^k(\theta)$ or $\sin^k(\theta)$. Indeed:

$$\begin{aligned}\sin^6(\theta)&=\left(\frac{e^{\mathrm{i}\theta}-e^{-\mathrm{i}\theta}}{2\mathrm{i}}\right)^6\\ &=-\frac{1}{64}\left(e^{\mathrm{i}6\theta}-6e^{\mathrm{i}5\theta}e^{-\mathrm{i}\theta}+15e^{\mathrm{i}4\theta}e^{-\mathrm{i}2\theta}-20e^{\mathrm{i}3\theta}e^{-\mathrm{i}3\theta}+15e^{\mathrm{i}2\theta}e^{-\mathrm{i}4\theta}-6e^{\mathrm{i}\theta}e^{-\mathrm{i}5\theta}+e^{-\mathrm{i}6\theta}\right)\\ &=-\frac{1}{64}\left(e^{\mathrm{i}6\theta}-6e^{\mathrm{i}4\theta}+15e^{\mathrm{i}2\theta}-20+15e^{-\mathrm{i}2\theta}-6e^{-\mathrm{i}4\theta}+e^{-\mathrm{i}6\theta}\right)\\ &=-\frac{1}{64}\left(e^{\mathrm{i}6\theta}+e^{-\mathrm{i}6\theta}-6\left(e^{\mathrm{i}4\theta+e^{-\mathrm{i}4\theta}}\right)+15\left(e^{\mathrm{i}2\theta}+e^{-\mathrm{i}2\theta}\right)\right)\\ &=-\frac{1}{32}\left(\cos(6\theta)-6\cos(4\theta)+15\cos(2\theta)-10\right)\end{aligned}\tag{5.2.150}$$

A well known special case for the formulas above is the case where $\varphi = \pi$ and $r = 1$. Then we have:

$$e^{\mathrm{i}\pi} = \cos(\pi) + \mathrm{i}\sin(\pi) = -1 \tag{5.2.151}$$

after rearranging we get the famous geek relation:

$$e^{\mathrm{i}\pi} + 1 = 0 \tag{5.2.152}$$

Another famous geek case is when we take still $r = 1$ but with $\varphi = \pi/2$ and we raise to the power of i. Then we get:

$$e^{\mathrm{i}\pi/2} = \cos(\pi/2) + \mathrm{i}\sin(\pi/2) = \mathrm{i} \tag{5.2.153}$$

and if we raise that to the power of i:

$$\mathrm{i}^{\mathrm{i}} = \left(e^{\mathrm{i}\pi/2}\right)^{\mathrm{i}} = e^{\mathrm{i}^2\pi/2} = e^{-\pi/2} \cong 0.20788 \tag{5.2.154}$$

Some people consider then as a curious fact that i raised to the i-th power is actually a real number.

Thanks to the exponential form of a complex number, very commonly used in many fields of physics and engineering, we can easily draw relations such that starting from (remember that cis is an old notation that stands for the $\cos(\varphi) + \mathrm{i}\sin(\varphi)$ being in the parenthesis):

$$\begin{aligned} z &= r(\cos(\varphi) + \mathrm{i}\sin(\varphi)) = r\text{cis}(\varphi) = re^{\mathrm{i}\varphi} \\ z_1 &= r_1(\cos(\varphi_1) + \mathrm{i}\sin(\varphi_1)) = r_1\text{cis}(\varphi_1) = r_1e^{\mathrm{i}\varphi_1} \\ z_2 &= r_2(\cos(\varphi_2) + \mathrm{i}\sin(\varphi_2)) = r_1\text{cis}(\varphi_2) = r_2e^{\mathrm{i}\varphi_2} \end{aligned} \tag{5.2.155}$$

and assuming known the basic trigonometric identities (see section Trigonometry page 2519) we have the following relations for the multiplication of two complex numbers:

$$z_1z_2 = r_1r_2\left[\cos(\varphi_1 + \varphi_2) + \mathrm{i}\sin(\varphi_1 + \varphi_2)\right] = r_1e^{\mathrm{i}\varphi_1}r_2e^{\mathrm{i}\varphi_2} = r_1r_2\text{cis}(\varphi_1 + \varphi_2) = r_1r_2e^{\mathrm{i}(\varphi_1+\varphi_2)} \tag{5.2.156}$$

therefore:

$$\arg(z_1z_2) = \arg(z_1) + \arg(z_2) \tag{5.2.157}$$

and therefore if n is a positive integer:

$$\arg(z^n) = n\arg(z) \tag{5.2.158}$$

For the module (norm) of the multiplication:

$$|z_1z_2| = |r_1e^{\mathrm{i}\varphi_1}r_2e^{\mathrm{i}\varphi_2}| = |r_1r_2e^{\mathrm{i}(\varphi_1+\varphi_2)}| = r_1r_2 = |r_1||r_2| \tag{5.2.159}$$

Therefore:

$$|z^m| = |z|^m \tag{5.2.160}$$

For the division of two complex numbers:

$$\frac{z_1}{z_2}=\frac{r_1}{r_2}\left[\cos(\varphi_1-\varphi_2)+\mathrm{i}\sin(\varphi_1-\varphi_2)\right]=\frac{r_1}{r_2}\mathrm{cis}(\varphi_1-\varphi_2)=\frac{r_1}{r_2}\frac{e^{\mathrm{i}\varphi_1}}{e^{\mathrm{i}\varphi_2}}=\frac{r_1}{r_2}e^{\mathrm{i}(\varphi_1-\varphi_2)} \quad (5.2.161)$$

The module of their division then comes immediately:

$$\left|\frac{z_1}{z_2}\right|=\frac{|z_1|}{|z_2|} \quad (5.2.162)$$

therefore we have for the argument:

$$\arg\left(\frac{z_1}{z_2}\right)=\arg\left(\frac{r_1e^{\mathrm{i}\phi_1}}{r_2e^{\mathrm{i}\phi_2}}\right)=\arg\left(\frac{r_1}{r_2}e^{\mathrm{i}(\varphi_1-\varphi_2)}\right)=\varphi_1-\varphi_2=\arg(z_1)-\arg(z_2) \quad (5.2.163)$$

and it comes immediately:

$$\arg(z^{-1})=\arg\left(\left(re^{\mathrm{i}\phi}\right)^{-1}\right)=\arg\left(\frac{1}{r}e^{\mathrm{i}\varphi}\right)=-\varphi=-\arg(z) \quad (5.2.164)$$

For the power of a complex number (or root):

$$z^m=r^me^{\mathrm{i}m\varphi}=r^m\left[\cos(m\varphi)+\mathrm{i}\sin(\varphi)\right]=r^m\mathrm{cis}(m\varphi) \quad (5.2.165)$$

which gives us immediately as already proved previously:

$$|z^m|=|z|^m \quad (5.2.166)$$

and for the argument:

$$\arg(z^m)=\arg\left(\left(re^{\mathrm{i}\varphi}\right)^m\right)=\arg\left(r^me^{m\mathrm{i}\varphi}\right)=m\varphi=m\arg(z) \quad (5.2.167)$$

In case we have a unit module (norm equal to 1) as $z=\cos(\varphi)+\mathrm{i}\sin(\varphi)$ we then have the relation:

$$(\cos(\varphi)+\mathrm{i}\sin(\varphi))^m=\left(e^{\mathrm{i}\varphi}\right)^m=e^{\mathrm{i}m\varphi}=\cos(m\varphi)+\mathrm{i}\sin(m\varphi) \quad (5.2.168)$$

named "de Moivre formula".

For the natural logarithm of a complex number, we trivially have the following relation which is discussed in the section of Functional Analysis (page 2284):

$$\ln(z)=\ln\left(re^{\mathrm{i}\varphi}\right)=\ln(r)+\mathrm{i}\varphi \quad (5.2.169)$$

where $\ln(z)$ is often in the complex case written $\mathrm{Log}(z)$ with an uppercase "L".

All previous relations could of course be obtained with the trigonometric form of complex numbers but then require some additional lines of mathematical developments.

Remark

Notice that the previous developments allow us to calculate things (...) like $\cos(\bar{z})$. Indeed, let $z = x + \mathrm{i}y$ and consider:

$$\overline{e^{\mathrm{i}z}} = \overline{e^{\mathrm{i}(x+\mathrm{i}y)}} = \overline{e^{-y}e^{\mathrm{i}x}} = e^{-y}e^{-\mathrm{i}x} = e^{-\mathrm{i}(x-\mathrm{i}y)} = e^{-\mathrm{i}\bar{z}} \tag{5.2.170}$$

Similarly, we have:

$$\overline{e^{-\mathrm{i}z}} = e^{\mathrm{i}\bar{z}} \tag{5.2.171}$$

Now consider:

$$\overline{\cos(z)} = \overline{\frac{1}{2}(e^{\mathrm{i}z} + e^{-\mathrm{i}z})} = \frac{1}{2}(e^{-\mathrm{i}z} + e^{\mathrm{i}\bar{z}}) = \cos(\bar{z}) \tag{5.2.172}$$

5.2.2.7.1.1 Fresnel Vectors (phasors)

A sinusoidal variation $f(t) = r\sin(\omega t)$ can be represented as the projection (see section Trigonometry page 2503) on the vertical y-axis (imaginary axis the set $\mathbb{C}$) of a rotating vector $\vec{r}$ at angular velocity ω around the origin in the plane xOy:

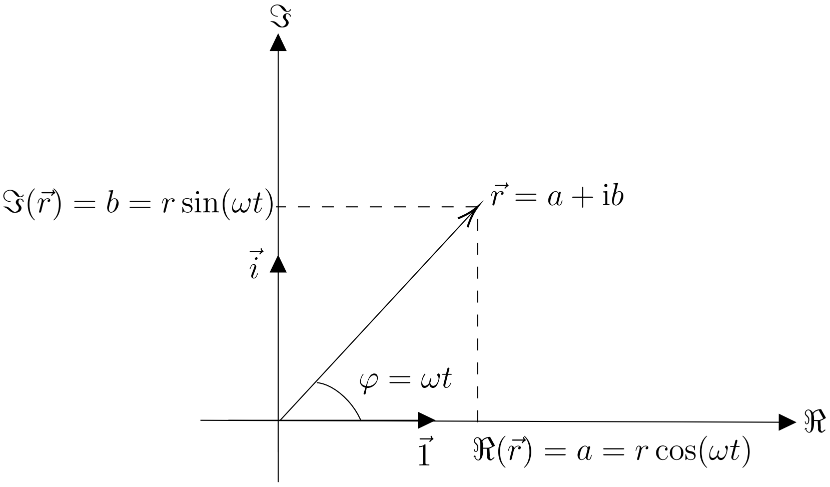
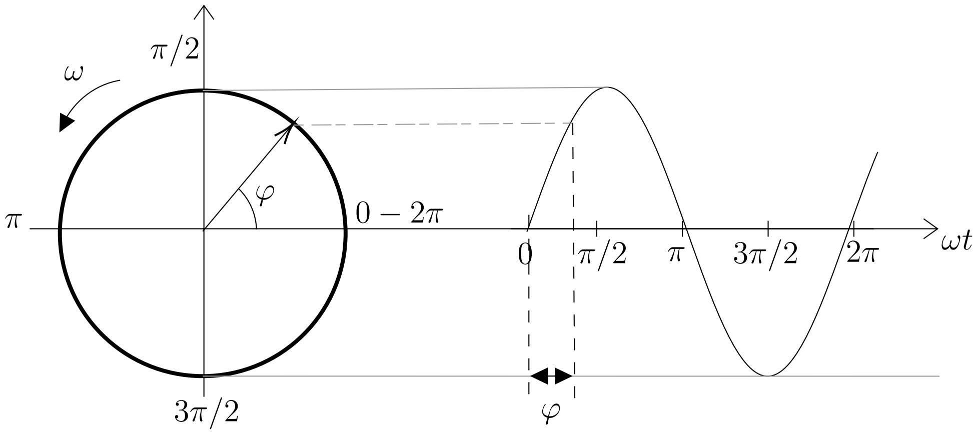

$\Im$
$\Im(\vec{r}) = b = r\sin(\omega t)$
$\vec{r} = a + \mathrm{i}b$
$\vec{i}$
$\varphi = \omega t$
$\vec{1}$
$\Re(\vec{r}) = a = r\cos(\omega t)$
$\Re$

Figure 5.19: **Fresnel representation**

Such a rotating vector is named "Fresnel vector" and can be well interpreted as the imaginary part of a complex number given by:

$$\vec{r} = \Re(\vec{r}) + \mathrm{i}\Im(\vec{r}) = r(\cos(\omega t) + \mathrm{i}\sin(\omega t)) = re^{\mathrm{i}\omega t} = re^{\mathrm{i}\phi} \tag{5.2.173}$$

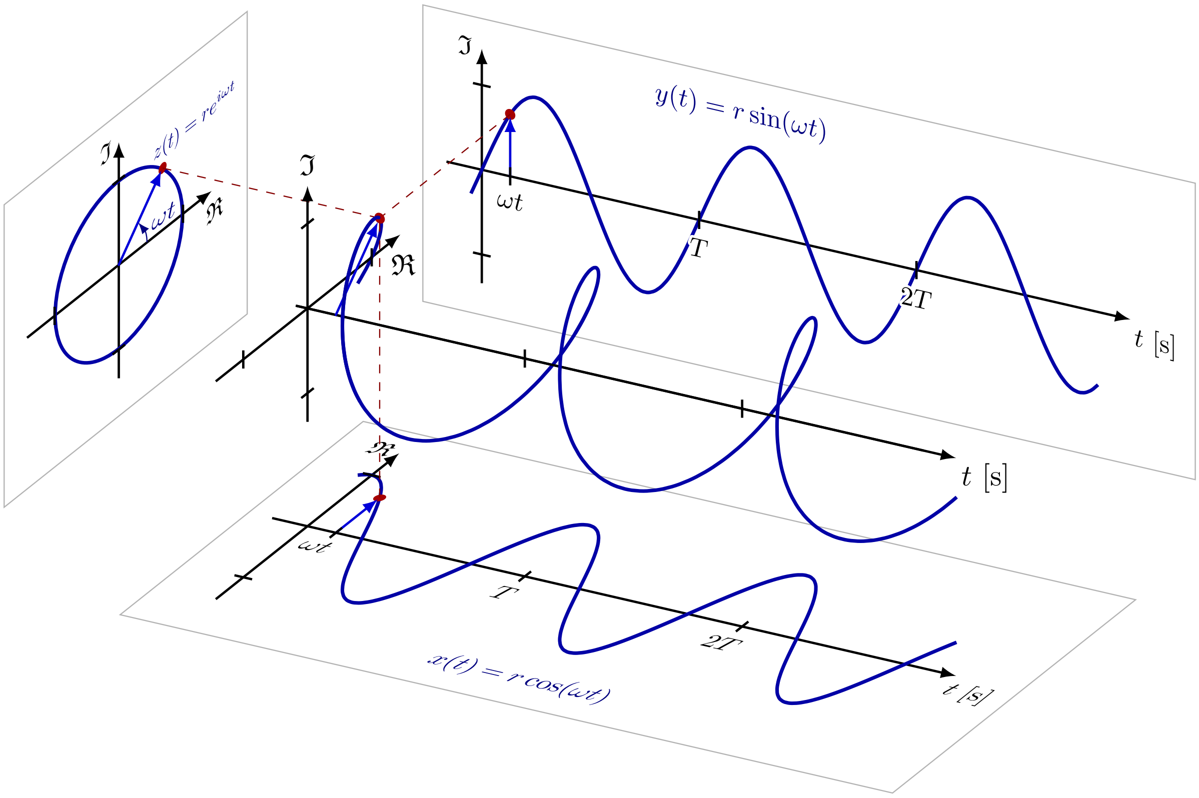
That is to say:



$\pi/2$
ω
φ
π
$0-2\pi$
$3\pi/2$
0
$\pi/2$
π
$3\pi/2$
2π
ωt
φ

Figure 5.20: **Fresnel rotating vector (complex oscillator)**

Or with a supplementary dimension:



$z(t)=re^{i\omega t}$
$\Im$
ωt
$\Re$
$\Im$
$\Re$
$\Im$
$y(t)=r\sin(\omega t)$
ωt
T
$2T$
t [s]
t [s]
$\Re$
ωt
T
$2T$
$x(t)=r\cos(\omega t)$
t [s]

Figure 5.21: **3D complex oscillator (author: Izaak Neutelings)**

We will see the phasor again explicitly in our study of wave mechanics and geometrical optics (as part of diffraction) in the sections with the corresponding names page 3475 and page 4314.

5.2.2.7.2 Transformation in the plane

It is customary to represent real numbers as points on a graduated line. The algebraic operations have their geometric interpretation on it: the addition is a translation, a multiplication a centered scaling.

In particular we can talk about the "square root of a transformation." A translation of amplitude T may be obtained as the iteration of a translation of amplitude $T/2$. Similarly, a scaling of amplitude S can be achieved as iterated scaling of faction $\sqrt{S}$. In particular an homothety (scaling) of a factor 9 can be composed of two homotheties (scaling) of respectively 3 (or -3).

Then we can say that the square root takes on a geometric sense. But what about the square root of negative numbers? In particular of the square root of -1???

A scaling of factor -1 can be seen as a symmetry with respect to the origin. But if we see this transformation in a continuous manner. Therefore a -1 scaling factor also be seen as rotation of π rotation around the origin.

So, the problem of negative square root is simplified. Indeed, it is not difficult to break down a rotation of π radians into two transformations: we can repeat either a rotation of $\pi/2$ or of $-\pi/2$. The image of 1 is the square root of -1 and i is situated on a perpendicular to the origin at a distance 1 either up or down.

Having successfully positioned the number i it not difficult anymore to put other complex numbers in the Gauss plane. We can therefore associate to $2\mathrm{i}$ the product of the scaling of a factor 2 (see section Euclidean Geometry page 2643) by the rotation of center O with angle of $\pi/2$, that is to say a similitude centered at the origin. This is what we will endeavour to prove now.

Given:

$$z_1 = x_1 + \mathrm{i}y_1 = ae^{\mathrm{i}\alpha}, z_2 = x_2 + \mathrm{i}y_2 = be^{\mathrm{i}\beta} \qquad (5.2.174)$$

We have the following geometric transformations properties for complex numbers (see the section Trigonometry at page 2519 for the properties of sine and cosine) that we can happily combine at our discretion:

P1. The multiplication of z_1 by a real number λ in the Gauss plane corresponds trivially to a homothety of center O (the intersection of real and imaginary axis for recall...) and of ratio λ.

Indeed:

$$\lambda z_1 = (\underline{\lambda a})e^{\mathrm{i}\alpha} \qquad (5.2.175)$$

P2. Multiplication of z_1 by a complex number of unit module corresponds a rotation of center O and of angle corresponding to the argument of z_1. Indeed:

$$z_0 z_1 = e^{\mathrm{i}\omega}ae^{\mathrm{i}\alpha} = ae^{\mathrm{i}(\alpha+\omega)} \qquad (5.2.176)$$

Remark

Then we see immediately, for example, that multiplying a complex number by i (that is to say a complex number with $\sin(\omega)=1,\cos(\omega)=0$) corresponds a rotation of $\pi/2$.

Theorem 5.5. It is interesting to notice that in vector form the rotation of center O of z_1 by z_0 can be written using the following matrix:

$$\begin{bmatrix} x_0 & -y_0 \\ y_0 & x_0 \end{bmatrix} \tag{5.2.177}$$

Proof 5.5.1. We have just seen before that z_0z_1 is a rotation of center O of and angle ω. We just need to write it first in the old style:

$$z_0z_1=(x_0+\mathrm{i}y_0)(x_1+\mathrm{i}y_1)=(x_0x_1-y_0y_1)+\mathrm{i}(x_0y_1+y_0x_1) \tag{5.2.178}$$

giving in vector form:

$$z_0z_1=\begin{pmatrix} x_0x_1-y_0y_1 \\ 0 \end{pmatrix}+\mathrm{i}\begin{pmatrix} 0 \\ x_0y_1+y_0x_1 \end{pmatrix} \tag{5.2.179}$$

thus the linear equivalent application is:

$$\begin{bmatrix} x_0 & -y_0 \\ y_0 & x_0 \end{bmatrix}\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}=\begin{pmatrix} x_0x_1-y_0x_1 \\ x_0y_1+y_0y_1 \end{pmatrix} \tag{5.2.180}$$

or as well (we fall back on the rotation matrix in the plane that we will see in the section of Euclidean Geometry page 2648 which is a remarkable result!) using:

$$z=r(\cos(\varphi)+\mathrm{i}\sin(\varphi))=r(\cos(\varphi+k2\pi)+\mathrm{i}\sin(\varphi+k2\pi)) \tag{5.2.181}$$

and in the particular and arbitrary case where r is unitary (in order to have a pure rotation!):

$$z=\cos(\varphi+k2\pi)+\mathrm{i}\sin(\varphi+k2\pi)=x_0+\mathrm{i}y_0 \tag{5.2.182}$$

we have immediately (we took again the same notations for the angle as the one we have in the chapter Geometry):

$$\begin{bmatrix} x_0 & -y_0 \\ y_0 & x_0 \end{bmatrix}\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}=\begin{bmatrix} \cos\omega & -\sin\omega \\ \sin\omega & \cos\omega \end{bmatrix}\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}=\begin{pmatrix} x_0x_1-y_0y_1 \\ x_0y_1+y_0x_1 \end{pmatrix}=\begin{pmatrix} \cos(\omega)x_1-\sin(\omega)y_1 \\ \cos(\omega)y_1+\sin(\omega)x_1 \end{pmatrix} \tag{5.2.183}$$

Notice that the 2D rotation matrix can also be written as:

$$\begin{bmatrix} \cos(\omega) & -\sin(\omega) \\ \sin(\omega) & \cos(\omega) \end{bmatrix}=\cos(\omega)\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}+\sin(\omega)\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}=\cos(\omega)I+\sin(\omega)J \tag{5.2.184}$$

as well:

$$\begin{bmatrix} x_0 & -y_0 \\ y_0 & x_0 \end{bmatrix}=x_0\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}+y_0\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}=x_0\cdot I+y_0\cdot J \tag{5.2.185}$$

■ Q.E.D.

Thus we see that the rotation matrices are not only applications but also are complex numbers (well it was obvious from the start but we had to show it in an aesthetic and simple way).

So, we have for usage to put that:

$$\boxed{1=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\quad\text{and}\quad \mathrm{i}=\begin{bmatrix}0 & -1\\ 1 & 0\end{bmatrix}}\tag{5.2.186}$$

or with another common notation in linear algebra:

$$\mathbb{1}=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\quad\text{and}\quad \mathrm{i}=\begin{bmatrix}0 & -1\\ 1 & 0\end{bmatrix}\tag{5.2.187}$$

The field of complex numbers is isomorphic to the field of real square matrices of dimension 2 of the type:

$$\begin{bmatrix}x_0 & -y_0\\ y_0 & x_0\end{bmatrix}\tag{5.2.188}$$

It is a result that we use many times in various sections of this book for specific studies in algebra, geometry and relativistic quantum physics.

P3. The multiplication of two complex corresponds to a homothety added to a rotation. In other words, a "direct similarity".

Proof 5.5.2.

$$z_1z_2=ae^{\mathrm{i}\alpha}be^{\mathrm{i}\beta}=(\underline{ab})e^{\mathrm{i}\underline{(\alpha+\beta)}}\tag{5.2.189}$$

so this is indeed a similarity of ratio b and angle β.

At the opposite, the following operation:

$$z_1\overline{z_2}=ae^{\mathrm{i}\alpha}be^{-\mathrm{i}\beta}=(\overline{ab})e^{\mathrm{i}\underline{(\alpha-\beta)}}\tag{5.2.190}$$

will be named a "retrograde linear similarity".

Otherwise, it returns trivially an already known following relation:

$$\arg(z_1z_2)=\arg(z_1)+\arg(z_2)\tag{5.2.191}$$

Remarks

R1. As the sum of two complex numbers z_1+z_2 can not have a special simplified mathematical notation in any form whatsoever, then we say that the resulting quantity is equivalent to an "amplitude translation".

R2. The combination of a direct linear similarity (multiplication of two complex numbers) and an amplitude translation (sum by a third complex number) is what we name a "direct linear similarity".

■ Q.E.D.

P4. The conjugate of a complex number is geometrically symmetrical with respect to the real axis such that:

$$\bar{z}_1 = x_1 - \mathrm{i}y_1 = r(\cos(\alpha) - \mathrm{i}\sin(\alpha)) = r(\cos(-\alpha) - \mathrm{i}\sin(-\alpha)) = re^{-\mathrm{i}\alpha} \tag{5.2.192}$$

without forgetting that (basis of trigonometry):

$$\cos(\varphi) = \cos(\varphi + k2\pi) \qquad \sin(\varphi) = \sin(\varphi + k2\pi) \tag{5.2.193}$$

This gives us a known result:

$$\arg(\bar{z}_1) = -\arg(z_1) \tag{5.2.194}$$

From which we get the following property:

$$r(\cos(\varphi + \pi) + \mathrm{i}\sin(\varphi + \pi)) = r\left(-\cos(\varphi) - \mathrm{i}\sin(\varphi)\right) = -r\left(\cos(\varphi) + \mathrm{i}\sin(\varphi)\right) = -z_1 \tag{5.2.195}$$

Hence:

$$\arg(z_1) + \pi = -\arg(-z_1) \tag{5.2.196}$$

P5. The negation of the conjugate of a complex number is geometrically its symmetrical with respect to the imaginary axis such that:

$$-\bar{z}_1 = -x_1 + \mathrm{i}y_1 = r(-\cos(\alpha) + \mathrm{i}\sin(\alpha)) = r(\cos(\pi \pm \alpha) + \mathrm{i}\sin(\alpha)) \tag{5.2.197}$$

Remarks

R1. The combination of the properties P4, P5 is named a "retrograde similarity".

R2. The geometric operation that consist to take the inverse of the conjugate of a complex number (that is to say $\bar{z}^{-1}$) is named a "pole inversion".

P6. The rotation of coordinate center c and angle φ is given and denoted by:

$$\mathcal{R}(z_1) = c + e^{\mathrm{i}\varphi}(z_1 - c) \tag{5.2.198}$$

Some explanations could be useful for some readers:

The complex c gives a point in the Gaussian plane, which will be the center of rotation. The difference $z_1 - c$ gives the chosen radius r. The multiplication by $e^{\mathrm{i}\varphi}$ is the counter-clockwise rotation of the radius from the origin of the Gaussian plane. Finally, the addition by c is the necessary translation to take back the rotated radius r at its original place before the rotation (center c). Which gives schematically:

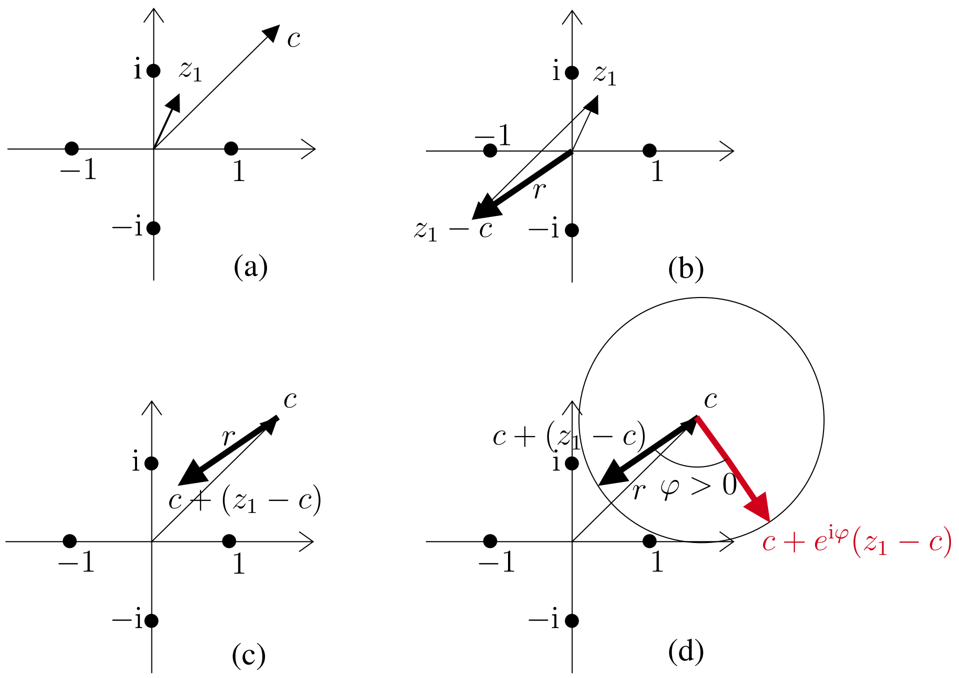


Figure 5.22: **Representation of the complex rotation**

P7. On the same idea, we get and denote an homothety of center c and ratio λ by:

$$\mathcal{H}(z_1) = c + \lambda(z_1 - c) \tag{5.2.199}$$

Some explanations could be useful for some readers:

The difference $z_1 - c$ always gives the radius r and c a central point in the Gauss plane. The expression $\lambda(z_1 - c)$ gives the homothety of the radius from the origin of the Gaussian plane, and finally by adding c gives the necessary translation for the homothety to be seen as being made from center c.

5.2.2.8 Quaternion Numbers

Also named "hypercomplex" quaternions numbers were invented in 11843 (holocene calendar) by William Rowan Hamilton to generalize complex numbers.

Definition (#5.2.39): A "quaternion" is an element $(a, b, c, d) \in \mathbb{R}^4$ and for which we denote by $\mathbb{H}$ the set that contains it and what we name the "set of quaternions".

A quaternion can also be represented in a row or column such as:

$$(a, b, c, d) = \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} \tag{5.2.200}$$

We define the sum of two quaternions (a, b, c, d) and (a', b', c', d') by:

$$(a, b, c, d) + (a', b', c', d') = (a + a', b + b', c + c', d + d') \tag{5.2.201}$$

Remark

It is the natural addition in $\mathbb{R}^4$ seen as a $\mathbb{R}$-vector space (see section Set Theory page 1805).

The associativity is verified by applying the corresponding properties of the operations on $\mathbb{R}$.

We also define the multiplication:

$$(a, b, c, d) \cdot (a', b', c', d') \tag{5.2.202}$$

of two quaternions (a, b, c, d) and (a', b', c', d') by the expression:

$$(a, b, c, d) \cdot (a', b', c', d') = \begin{pmatrix} aa' - bb' - cc' - dd' \\ (ab' + ba') + (cd' - dc') \\ (ac' + ca') - (bd' - db') \\ (da' + ad') + (bc' - cb') \end{pmatrix} \tag{5.2.203}$$

It may be hard to accept but we will be a little further below that there is a family resemblance with the complex numbers.

We can notice that the law of multiplication is not commutative. Indeed, taking the definition of the multiplication above, we have:

$$\begin{aligned} (0, 1, 0, 0) \cdot (0, 0, 1, 0) &= (0, 0, 0, 1) \\ (0, 0, 1, 0) \cdot (0, 1, 0, 0) &= (0, 0, 0, -1) \end{aligned} \tag{5.2.204}$$

But we can also notice that:

$$(0, 1, 0, 0) \cdot (0, 0, 1, 0) = -1(0, 0, 1, 0) \cdot (0, 1, 0, 0) \tag{5.2.205}$$

Remark

It is the natural multiplication in $\mathbb{R}^4$ seen as a $\mathbb{R}$-vector space (see section Set Theory page 1805).

The law of multiplication is distributive with the addition law but it is an excellent example where we must still be careful to prove the left and right distributivity, since the product is not commutative!

The multiplication has for neutral element:

$$(1, 0, 0, 0) \tag{5.2.206}$$

Indeed:

$$(1, 0, 0, 0) \cdot (a, b, c, d) = (a, b, c, d) \cdot (1, 0, 0, 0) = (a, b, d, c) \tag{5.2.207}$$

Any element:

$$(a, b, c, d) \in \mathbb{H}^* = \mathbb{H} - \{(0, 0, 0, 0)\} \tag{5.2.208}$$

is invertible.

Indeed, if (a,b,c,d) is a non-null quaternion, we then have necessarily:

$$a^2+b^2+c^2+d^2\neq 0 \tag{5.2.209}$$

otherwise the four numbers a,b,c,d are of square null, so all zero. Given then the quaternion (a_1,b_1,c_1,d_1) defined by:

$$\begin{aligned} a_1&=\frac{a}{a^2+b^2+c^2+d^2}\\ b_1&=\frac{-b}{a^2+b^2+c^2+d^2}\\ c_1&=\frac{-c}{a^2+b^2+c^2+d^2}\\ d_1&=\frac{-d}{a^2+b^2+c^2+d^2} \end{aligned} \tag{5.2.210}$$

then by applying mechanically the definition of the multiplication of quaternions, we check that:

$$(a,b,c,d)\cdot(a_1,b_1,c_1,d_1)=(a_1,b_1,c_1,d_1)\cdot(a,b,c,d)=(1,0,0,0) \tag{5.2.211}$$

this latter quaternion is therefore the inverse for the multiplication!

Let us prove (for general knowledge) that the field of complex numbers $(\mathbb{C},+,\times)$ is a subfield of $(\mathbb{H},+,\times)$.

Remark

We could also have put this proof in the section of Set Theory because we will make use of a lot of concepts that are have seen there but it seemed to us a little more relevant to put instead the proof here. We expect the reader to tolerate this choice.

Given $\mathbb{H}'$ a set of quaternions of the form $(a,b,0,0)$. If $\mathbb{H}'$ is not empty, and if $(a,b,0,0)$, $(a',b',0.0)$ are elements $\mathbb{H}'$ then $(\mathbb{H}',+\times)$ is a field. Indeed:

P1. For subtraction (and therefore the addition):

$$(a,b,0,0)-(a',b',0,0)=(a-a',b-b',0,0)\in\mathbb{H}' \tag{5.2.212}$$

P2. The multiplication:

$$(a,b,0,0)\cdot(a',b',0,0)=(aa'-bb',ab'+ba',0,0)\in\mathbb{H}' \tag{5.2.213}$$

P3. The neutral element:

$$(1,0,0,0)\in\mathbb{H}' \tag{5.2.214}$$

P4. And finally the inverse:

$$\left(\frac{a}{a^2+b^2},-\frac{b}{a^2+b^2},0,0\right)\in\mathbb{H}' \tag{5.2.215}$$

of $(a,b,0,0)$ is still in $\mathbb{H}'$.

Therefore $(\mathbb{H}', +, \times)$ is a subfield of $\mathbb{H}$. Given then the application:

$$\begin{gathered} f : a + \mathrm{i}b \to (a, b, 0, 0) \\ \mathbb{C} \to \mathbb{H} \end{gathered} \tag{5.2.216}$$

f is bijective, and we easily check that for any complex z_1, z_2, we have:

$$\begin{gathered} f(z_1 + z_2) = f(z_1) + f(z_2) \\ f(z_1 z_2) = f(z_1) f(z_2) \end{gathered} \tag{5.2.217}$$

Therefore f is an isomorphism of $(\mathbb{C}, +, \times)$ on $(\mathbb{H}', +, \times)$.

This isomorphism has for interest (caused) to identify $\mathbb{C}$ to $\mathbb{H}'$ and to write $\mathbb{C} \subset \mathbb{H}$, the laws of addition and subtraction on $\mathbb{H}$ extending the already known operations of $\mathbb{C}$.

Thus, by convention, we will write any element of $(a, b, 0, 0)$ of $\mathbb{H}'$ in the complex form $a + \mathrm{i}b$. Particularly 0 is the element $(0, 0, 0, 0)$, 1 is the element $(1, 0, 0, 0)$ and i and the element $(0, 1, 0, 0)$.

We denote by analogy and by extension j the element $(0, 0, 1, 0)$ and k the element $(0, 0, 0, 1)$. The family $\{1, \mathrm{i}, \mathrm{j}, \mathrm{k}\}$ form a basis of all quaternions seen as a vector space on $\mathbb{R}$, and we will write:

$$\underbrace{a}_{\text{scalar part}} + \underbrace{b\mathrm{i} + c\mathrm{j} + d\mathrm{k}}_{\text{vector part}} \tag{5.2.218}$$

the quaternion (a, b, c, d).

The notation of quaternions as defined above is perfectly suited to the multiplication operation. For the product of two quaternions we get by developing the expression:

$$(a + b\mathrm{i} + c\mathrm{j} + d\mathrm{k}) \cdot (a' + b'\mathrm{i} + c'\mathrm{j} + d'\mathrm{k}) \tag{5.2.219}$$

16 terms that we have to identify to the original definition of multiplication of quaternions to get the following relations:

$$\begin{aligned} &\mathrm{i} \cdot \mathrm{j} = \mathrm{k} = -\mathrm{j} \cdot \mathrm{i} \\ &\mathrm{j} \cdot \mathrm{k} = \mathrm{i} = -\mathrm{k} \cdot \mathrm{j} \\ &\mathrm{k} \cdot \mathrm{i} = \mathrm{j} = -\mathrm{i} \cdot \mathrm{k} \\ &\mathrm{i}^2 = \mathrm{j}^2 = \mathrm{k}^2 = -1 \end{aligned} \tag{5.2.220}$$

Which can be summarized in a table:

$\cdot$	1	i	j	k
1	1	i	j	k
i	i	-1	k	$-\mathrm{j}$
j	j	$-\mathrm{k}$	-1	i
j	j	j	$-\mathrm{i}$	-1

(5.2.221)

Or in a 3D representation of what happens in the corresponding 4D sphere (red, green and blue for i, j and k respectively):

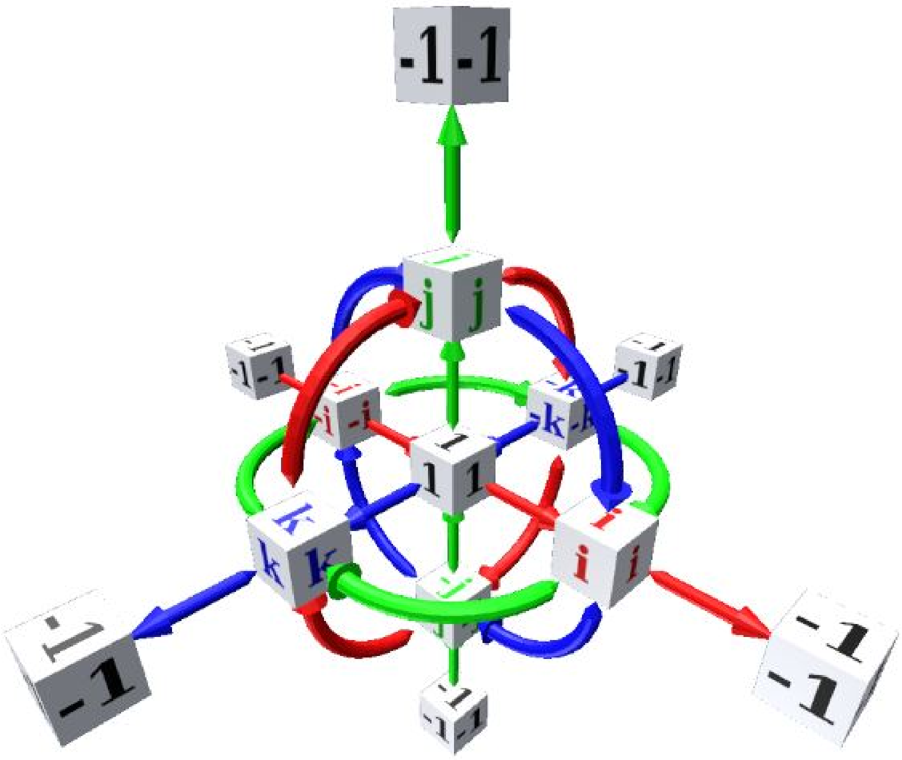


Figure 5.23: **Starting situation for quaternion rotations**

We can see that the expression of the multiplication of two quaternions looks partly much like a vector product (denoted × in this book) and dot product (denoted ∘ in this book):

$$\begin{pmatrix} aa'-bb'-cc'-dd' \\ (ab'+ba')+(cd'-dc') \\ (ac'+ca')-(bd'-db') \\ (da'+ad')+(bc'-cb') \end{pmatrix} \tag{5.2.222}$$

If this is not evident (which would be quite understandable), let us make a concrete example:

✍**Example:**

Given two quaternions without real part:

$$p=x\mathrm{i}+y\mathrm{j}+z\mathrm{k} \qquad q=x'\mathrm{i}+y'\mathrm{j}+z'\mathrm{k} \tag{5.2.223}$$

and $\vec{u},\vec{v}$ the vectors of $\mathbb{R}^3$ of respective components (x,y,z) and (x',y',z'). Then the product:

$$pq=(0,\vec{u})(0,\vec{v}) \tag{5.2.224}$$

is equal to:

$$p\cdot q=(-xx'-yy'-zz',yz'-zy',-xz'+zx',xy'-yx')=(-\vec{u}\circ\vec{v},\vec{u}\times\vec{v}) \tag{5.2.225}$$

We can also for curiosity interest us to the general case ... Given for this two quaternions:

$$p=(a,\vec{u}) \qquad q=(b,\vec{v}) \tag{5.2.226}$$

Then we have:

$$\begin{aligned} p\cdot q &= (a+(0,\vec{u}))\cdot(b+(0,\vec{v}))\\ &= ab+(0,a\vec{v})+(0,b\vec{u})+(-\vec{u}\circ\vec{v},\vec{u}\times\vec{v})\\ &= (ab-\vec{u}\circ\vec{v},a\vec{v}+b\vec{u}+\vec{u}\times\vec{v}) \end{aligned} \tag{5.2.227}$$

Definition (#5.2.40): The center of the non-commutative field $(\mathbb{H},+,\times)$ is the set of elements of $\mathbb{H}$ commuting for the law of multiplication with all the elements of $\mathbb{H}$.

Theorem 5.6. The center of $(\mathbb{H},+,\times)$ is the set of real numbers!

Proof 5.6.1. Give $\mathbb{H}_1$ is the center of $(\mathbb{H},+,\times)$, and (x,y,z,t) a quaternion. We must have the following conditions that are met:

Given $(x,y,z,t)\in\mathbb{H}_1$ then for any $(a,b,c,d)\in\mathbb{H}$ we seek:

$$(x,y,z,t)\cdot(a,b,c,d)=(a,b,c,d)\cdot(x,y,z,t) \tag{5.2.228}$$

which give by developing:

$$\begin{aligned} xa-yb-zc-td&=ax-by-cz-dt\\ xb+ya+zd-tc&=ay+bx+ct-dz\\ xc+za-yd+tb&=az+cx-bt+dy\\ ta+xd+yc-zb&=dx+at+bz-cy \end{aligned} \tag{5.2.229}$$

after simplification (the first line of the previous system is equal to zero on both sides of equality):

$$\begin{aligned} ct-dz&=0\\ bt-dy&=0\\ bz-cy&=0 \end{aligned} \tag{5.2.230}$$

the resolution of this system gives us:

$$y=z=t=0 \tag{5.2.231}$$

So that the quaternion (x,y,z,t) is the center of $\mathbb{H}$ it must be real (not imaginary parts)!

■ Q.E.D.

Just as for complex numbers, we can define a conjugate of quaternions:

Definition (#5.2.41): The conjugate of a quaternion $Z=(a,b,c,d)$ is the quaternion $\bar{Z}=(a,-b,-c,-d)$.

Just as for the complex number, we notice that:

1. First clearly that if $Z=\bar{Z}$ then it means that $Z\in\mathbb{R}$

2. That $Z+\bar{Z}\in\mathbb{R}$
3. That by developing the product $Z\bar{Z}$ we have:

$$\begin{aligned}Z\bar{Z}&=(a,b,c,d)\cdot(a,-b,-c,-d)\\&=(a^2+b^2+c^2+d^2,-ab+ba-cd+dc,-ac+ca+bd-db,da-ad-bc+cb)\\&=(a^2+b^2+c^2+d^2,0,0,0)=a^2+b^2+c^2+d^2\in\mathbb{R}\end{aligned} \tag{5.2.232}$$

that we will adopt, by analogy with complex numbers, as a definition of the norm (or module) of quaternions such as:

$$|Z|=\sqrt{Z\cdot\bar{Z}} \tag{5.2.233}$$

Therefore we also have immediately (relation which will be useful later):

$$|ZZ'|=\sqrt{(ZZ')(\overline{ZZ'}} \tag{5.2.234}$$

As for complex numbers (see below), it is easy to show that the conjugation is an automorphism of the group $(\mathbb{H},+)$.

Indeed, given $Z=(a,b,c,d)$ and $Z'=(a',b',c',d')$ then:

$$\begin{aligned}\overline{Z+Z'}&=(a+a',-b-b',-c-c',-d-d')\\&=(a,-b,-c,-d)+(a',-b',-c',-d')\\&=\bar{Z}+\bar{Z}'\end{aligned} \tag{5.2.235}$$

It is also easy to prove that it is involutive. Indeed:

$$\bar{\bar{Z}}=(a,-(-b),-(-c),-(-d))=(a,b,c,d)=Z \tag{5.2.236}$$

But the conjugation is not a multiplicative automorphism of the field $(\mathbb{H},+,\times)$. Indeed, if we consider the multiplication of Z, Z' and take the conjugate:

$$ZZ'=\begin{pmatrix}aa'-bb'-cc'-dd'\\(ab'+ba')+(cd'-dc')\\(ac'+ca')-(bd'-db')\\(da'+ad')+(bc'-cb')\end{pmatrix}\qquad\overline{ZZ'}=\begin{pmatrix}aa'-bb'-cc'-dd'\\-(ab'+ba')-(cd'-dc')\\-(ac'+ca')+(bd'-db')\\-(da'+ad')-(bc'-cb')\end{pmatrix} \tag{5.2.237}$$

we see immediately (at least for the second row) that we have:

$$ZZ'\neq\overline{ZZ'} \tag{5.2.238}$$

Let us now back to our norm (or module) .... For this, let us calculate the square of the norm $|ZZ'|$:

$$|ZZ'|^2=(ZZ')\cdot(\overline{ZZ'}) \tag{5.2.239}$$

We know (by definition) that:

$$ZZ'=\begin{pmatrix}aa'-bb'-cc'-dd'\\(ab'+ba')+(cd'-dc')\\(ac'+ca')-(bd'-db')\\(da'+ad')+(bc'-cb')\end{pmatrix} \tag{5.2.240}$$

Let us denote this product in such a way that:

$$Z\cdot Z'=(\alpha,\beta,\gamma,\lambda)=Z\text{"}\tag{5.2.241}$$

Then we have:

$$Z\text{"}\overline{Z\text{"}}=\alpha^2+\beta^2+\gamma^2+\lambda^2\tag{5.2.242}$$

substituting it comes:

$$\begin{aligned}(ZZ')\cdot\overline{(ZZ')}=&(aa'-bb'-cc'-dd')^2+(ab'+ba'+cd'-dc')^2\\&+(ac'+ca'-bd'+db')^2+(da'+ad'+bc'-cb')^2\end{aligned}\tag{5.2.243}$$

after an elementary algebraic development (frankly boring) we find:

$$(ZZ')\cdot\overline{(ZZ')}=(a'^2+b'^2+c'^2+d'^2)(a^2+b^2+c^2+d^2)=|Z'|^2|Z|^2\tag{5.2.244}$$

Therefore:

$$(ZZ')\cdot\overline{(ZZ')}=|Z|^2|Z'|^2=|ZZ'|^2\tag{5.2.245}$$

Remark

The norm is therefore a homomorphism (see section Set Theory page 441) of $(\mathbb{H},\times)$ in $(\mathbb{R},\times)$. Subsequently, we will denote by $\mathbb{G}$ all the quaternions of unit norm.

5.2.2.8.1 Matrix Interpretation of Quaternions

Given q and p two quaternions and given the application:

$$p\rightarrow qp\tag{5.2.246}$$

The (left) multiplication can be made with a linear application (see section Linear Algebra page 1934) on $\mathbb{H}$.

If q is written:

$$a+b\mathrm{i}+c\mathrm{j}+d\mathrm{k}\tag{5.2.247}$$

this application has for matrix in the basis $(1,\mathrm{i},\mathrm{j},\mathrm{k})$:

$$\begin{bmatrix}a&-b&-c&-d\\b&a&-d&c\\c&d&a&-b\\d&-c&b&a\end{bmatrix}\tag{5.2.248}$$

What we check well:

$$ZZ'=\begin{bmatrix}a&-b&-c&-d\\b&a&-d&c\\c&d&a&-b\\d&-c&b&a\end{bmatrix}\begin{pmatrix}a'\\b'\\c'\\d'\end{pmatrix}=\begin{pmatrix}aa'-bb'-cc'-dd'\\(ab'+ba')+(cd'-dc')\\(ac'+ca')-(bd'-db')\\(da'+ad')+(bc'-cb')\end{pmatrix}\tag{5.2.249}$$

In fact, we can then define the quaternions as the set of matrices with the visible structure above if we wanted to. This will then reduce them to a sub vector space of $M_4(\mathbb{R})$.

Especially, the matrix of 1 (the real part of the quaternion q) is then nothing other than the identity matrix:

$$M_1=\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix}=\mathbb{1} \tag{5.2.250}$$

as well:

$$M_i=\begin{bmatrix}0&-1&0&0\\1&0&0&0\\0&0&0&-1\\0&0&1&0\end{bmatrix},\quad M_j=\begin{bmatrix}0&0&-1&0\\0&0&0&1\\1&0&0&0\\0&-1&0&0\end{bmatrix},\quad M_k=\begin{bmatrix}0&0&0&-1\\0&0&-1&0\\0&1&0&0\\1&0&0&0\end{bmatrix} \tag{5.2.251}$$

5.2.2.8.2 Rotations with Quaternions

We will see now that conjugation by an element of the group $\mathbb{G}$ of the quaternions of unit norm can be interpreted as a pure rotation in space!

Definition (#5.2.42): The "conjugation" by a non-null quaternion q of unit norm is the application S_q defined on $\mathbb{H}$ by:

$$\boxed{S_q : p \rightarrow q\cdot p\cdot q^{-1}=q\cdot p\cdot\bar{q}} \tag{5.2.252}$$

and we affirm that this application is a rotation.

Remarks

R1. As q is of unit norm 1, we have obviously $|q|=q\bar{q}=1$ therefore $q^{-1}=\bar{q}$. This quaternion can be seen as the proper value (of unit norm) to the application (matrix) p on the vector $\bar{q}$ (we are in a similar situation as the orthogonal rotation matrices seen in the section of Linear Algebra page 1954).

R2. S_q is a linear application (so if it is rotation, the rotation can be decomposed into several rotations). Indeed, let consider two quaternions p_1,p_2 and two real number λ_1,λ_2, then we have:

$$S_q(\lambda_1p_1+\lambda_2p_2)=q(\lambda_1p_1+\lambda_2p_2)\bar{q}=\lambda_1qp_1\bar{q}+\lambda_2qp_2\bar{q}=\lambda_1S_q(p_1)+\lambda_2S_q(p_2) \tag{5.2.253}$$

Let us now check that the application is indeed a pure rotation. As we saw in our study of Linear Algebra and in particular of orthogonal matrices (see section Linear Algebra page 1954), a first obvious condition is that the application conserves the norm.

Let us check this:

$$|S_q(p)|=|qp\bar{q}|=|q||p||\bar{q}|=|p| \tag{5.2.254}$$

Moreover, we can check that a rotation of a purely complex quaternion (such that then we restrict ourselves to $\mathbb{R}^3$) and the same summed reverse rotation is zero (the vector sum up to its opposite cancel):

$$S_q(p)+\overline{S_q(p)}=qp\bar{q}+\overline{qp\bar{q}}=qp\bar{q}+\overline{q(p\bar{q})} \tag{5.2.255}$$

we trivially check that if we have two quaternions q,p then $\overline{p\cdot q}=\bar{q}\bar{p}$ since then:

$$\begin{aligned}S_q(p)+\overline{S_q(p)}&=qp\bar{q}+\overline{q(p\bar{q})}=qp\bar{q}+\overline{(p\bar{q})}\bar{q}\\&=q\cdot p\cdot\bar{q}+\bar{\bar{q}}\cdot\bar{p}\cdot\bar{q}=q\cdot p\cdot\bar{q}+q\cdot\bar{p}\cdot\bar{q}\\&=S_q(p+\bar{p})\end{aligned} \tag{5.2.256}$$

for this operation to be zero, we immediately see that we need to restrict ourselves to the purely complex quaternions p. Since then:

$$S_q(p+\bar{p})=S_q(0)=0 \tag{5.2.257}$$

We conclude then that p must be purely complex so the for the application S_q is a rotation and that $S_q(p)$ is a pure quaternion. In other words, this application is stable (in other words: a pure quaternion by this application remains a pure quaternion).

The application S_q restricted to all purely complex quaternions is thus a vectorial isometry, that is to say a symmetry or a rotation.

We have also seen during our study of the rotation matrices in the section of Linear Algebra page 2001 and Euclidean Geometry page 2648 that such matrices should have a determinant equal to $+1$ so that we have a rotation. Let's see if this is the case of S_q:

For this, we explicitly calculate in function of:

$$q=a+b\mathrm{i}+c\mathrm{j}+d\mathrm{k} \tag{5.2.258}$$

the matrix (in the canonical basis $(\mathrm{i},\mathrm{j},\mathrm{k})$) of S_q and we calculate its determinant. Thus we obtain the coefficients of the columns of this application by remembering that:

$$\begin{aligned}\mathrm{ij}&=\mathrm{k}=-\mathrm{ji}\\\mathrm{jk}&=\mathrm{i}=-\mathrm{kj}\\\mathrm{ki}&=\mathrm{j}=-\mathrm{ik}\\\mathrm{i}^2&=\mathrm{j}^2=\mathrm{k}^2=-1\end{aligned} \tag{5.2.259}$$

and then by calculating:

$$\begin{aligned}S_q(\mathrm{i})&=(a+b\mathrm{i}+c\mathrm{j}+d\mathrm{k})\mathrm{i}(a-b\mathrm{i}-c\mathrm{j}-d\mathrm{k})\\&=(a\mathrm{i}+b(\mathrm{i}^2)+c(\mathrm{ji})+d(\mathrm{ki}))(a-b\mathrm{i}-c\mathrm{j}-d\mathrm{k})\\&=(a\mathrm{i}-b-c\mathrm{k}+d\mathrm{j})(a-b\mathrm{i}-c\mathrm{j}-d\mathrm{k})\\&=(a2\mathrm{i}+ab-ac\mathrm{k}+ad\mathrm{j})-(ba-b^2\mathrm{i}-bc\mathrm{j}-bd\mathrm{k})\\&-(ca\mathrm{k}-cb\mathrm{j}+c^2\mathrm{i}+cd)+(da\mathrm{j}+db\mathrm{k}+cd-d^2\mathrm{i})\\&=(a^2+b^2-c^2-d^2)\mathrm{i}+(\cancel{ab}-ac\mathrm{k}+ad\mathrm{j})-(\cancel{ba}-bc\mathrm{j}-bd\mathrm{k})\\&-(ca\mathrm{k}-cb\mathrm{j}+\cancel{cd})+(da\mathrm{j}+db\mathrm{k}+\cancel{cd})\\&=(a^2+b^2-c^2-d^2)\mathrm{i}+2(ad+bc)\mathrm{j}+2(bd-ac)\mathrm{k}\end{aligned} \tag{5.2.260}$$

$$\begin{aligned}
S_q(\mathrm{j}) &= (a+b\mathrm{i}+c\mathrm{j}+d\mathrm{k})\mathrm{j}(a-b\mathrm{i}-c\mathrm{j}-d\mathrm{k})\\
&= (a\mathrm{j}+b(\mathrm{ij})+c(\mathrm{j}^2)+d(\mathrm{kj}))(a-b\mathrm{i}-c\mathrm{j}-d\mathrm{k})\\
&= (a\mathrm{j}+b\mathrm{k}-c-d\mathrm{i})(a-b\mathrm{i}-c\mathrm{j}-d\mathrm{k})\\
&= (a^2\mathrm{j}+ab\mathrm{k}+ac-ad\mathrm{i})+(ba\mathrm{k}-b^2\mathrm{j}+bc\mathrm{i}+bd)\\
&-(ca-cb\mathrm{i}-c^2\mathrm{j}-cd\mathrm{k})-(da\mathrm{i}+db-dc\mathrm{k}+d^2\mathrm{j})\\
&= (a^2-b^2+c^2-d^2)\mathrm{j}+(ab\mathrm{k}+\cancel{ac}-ad\mathrm{i})+(ba\mathrm{k}+bc\mathrm{i}+\cancel{bd})\\
&-(\cancel{ca}-cb\mathrm{i}-cd\mathrm{k})-(da\mathrm{i}+\cancel{db}-dc\mathrm{k})\\
&= 2(bc-ad)\mathrm{i}+(a^2-b^2+c^2-d^2)\mathrm{j}+2(ab+cd)\mathrm{k}
\end{aligned} \tag{5.2.261}$$

$$\begin{aligned}
S_q(\mathrm{k}) &= (a+b\mathrm{i}+c\mathrm{j}+d\mathrm{k})\mathrm{k}(a-b\mathrm{i}-c\mathrm{j}-d\mathrm{k})\\
&= (a\mathrm{k}+b(\mathrm{ik})+c(\mathrm{jk})+d)(\mathrm{k}^2))(a-b\mathrm{i}-c\mathrm{j}-d\mathrm{k})\\
&= (a\mathrm{k}-b\mathrm{j}+c\mathrm{i}-d)(a-b\mathrm{i}-c\mathrm{j}-d\mathrm{k})\\
&= (a^2\mathrm{k}-ab\mathrm{j}+ac\mathrm{i}+ad)-(ab\mathrm{j}+b^2\mathrm{k}+bc-bd\mathrm{i})\\
&+(ca\mathrm{i}+cb-c^2\mathrm{k}+cd\mathrm{j})-(da-db\mathrm{i}-dc\mathrm{j}-d^2\mathrm{k})\\
&= (a^2-b^2-c^2+d^2)\mathrm{k}+(-ab\mathrm{j}+ac\mathrm{i}+\cancel{ad})-(ab\mathrm{j}+\cancel{bc}-bd\mathrm{i})\\
&+(ca\mathrm{i}+\cancel{cb}+cd\mathrm{j})-(\cancel{da}-db\mathrm{i}-dc\mathrm{j})\\
&= 2(ac+bd)\mathrm{i}+2(cd-ab)\mathrm{j}+(a^2-b^2-c^2+d^2)\mathrm{k}
\end{aligned} \tag{5.2.262}$$

We must then calculate the determinant of the following matrix (pfff ...):

$$\begin{bmatrix}
a^2+b^2-c^2-d^2 & 2(ad+bc) & 2(bd-ac)\\
2(bc-ad) & a^2-b^2+c^2-d^2 & 2(ab+cd)\\
2(ac+bd) & 2(cd-ab) & (a^2-b^2-c^2+d^2)
\end{bmatrix} \tag{5.2.263}$$

remembering that (which also simplifies the expression of the terms of the diagonal as we can see in some books):

$$a^2+b^2+c^2+d^2=1 \tag{5.2.264}$$

we find that the determinant is indeed equal to 1. Otherwise, we can check this with Maple™ 4.00b:

```
>with(linalg):
>A:=linalg[matrix](3,3,[a^2+b^2-c^2-d^2,2*(a*d+b*c),
2*(b*d-a*c),2*(b*c-a*d),a^2-b^2+c^2-d^2,2*(a*b+c*d),
2*(a*c+b*d),2*(c*d-a*b),a^2-b^2-c^2+d^2]);
>factor(det(A));
```

Let us now show that this rotation is a half axis turn (the example that may seem particular is in fact general!):

First, if:

$$q=x\mathrm{i}+y\mathrm{j}+z\mathrm{k} \tag{5.2.265}$$

we have:

$$S_q(q)=qq\bar{q}=q \qquad (5.2.266)$$

which means that the axis of rotation (x,y,z) is fixed by the application S_q itself!

On the other hand, we have seen that if q is a purely complex quaternion of norm 1 then:

$$q^{-1}=\bar{q} \qquad \text{and} \qquad \bar{q}=-q \qquad (5.2.267)$$

Which gives us the relation:

$$q^2=q\cdot(-\bar{q})=q\cdot(-(q^{-1}))=-1 \qquad (5.2.268)$$

This result leads us to calculate the rotation of a rotation:

$$S_q(S_q(p))=q(qp\bar{q})\bar{q}=q^2p\bar{q}^2=S_{q^2}(p)=(-1)p(-q)^2=-pq^2=(-1)p(-1)=p \qquad (5.2.269)$$

Conclusion: Since the rotation of a rotation is a full turn, then S_q is necessarily a half-turn:

$$S_q(p)=-p \qquad (5.2.270)$$

relatively (!) to the axis (x,y,z).

At this stage, we can say that any rotation of the space can be represented by S_q (the conjugation by a quaternion q of norm 1). Indeed, the half turns generates the group of rotations, that is to say that any rotation can be expressed as the product of a finite number of half-turns, and therefore by conjugation of a product of quaternions unitary norm (product which is itself a quaternion of unitary norm...).

We will still give an explicit form connecting a rotation and the quaternion that represents it, just as we did for complex numbers.

Theorem 5.7. Given $\vec{u}(x,y,z)$ a unit vector and $\theta\in[0,2\pi]$ angle. Then we affirm that the rotation of axis $\vec{u}$ and angle θ corresponds to the application S_q, where q is the quaternion:

$$\boxed{q=\cos\left(\frac{\theta}{2}\right)+x\sin\left(\frac{\theta}{2}\right)+y\sin\left(\frac{\theta}{2}\right)\mathrm{j}+z\sin\left(\frac{\theta}{2}\right)\mathrm{k}=\left(\cos\left(\frac{\theta}{2}\right),\sin\left(\frac{\theta}{2}\right)\vec{u}\right)\begin{pmatrix}\cos\left(\frac{\theta}{2}\right)\\ x\sin\left(\frac{\theta}{2}\right)\\ y\sin\left(\frac{\theta}{2}\right)\\ z\sin\left(\frac{\theta}{2}\right)\end{pmatrix}} \qquad (5.2.271)$$

For this assertion is verified, we know we need that:

- The norm of q is equal to 1
- The determinant of the application S_q is equal to 1
- The application S_q conserves the norm
- The application S_q returns all collinear vector to the axis of rotation on the axis of rotation itself.

Proof 5.7.1. Ok let us check every point:

1. The norm of the quaternion previously proposed is indeed equal to 1:

$$\begin{aligned}|q|&=\cos^2\left(\frac{\theta}{2}\right)+x^2\sin^2\left(\frac{\theta}{2}\right)+y^2\sin^2\left(\frac{\theta}{2}\right)+z^2\sin^2\left(\frac{\theta}{2}\right)\\&=\cos^2\left(\frac{\theta}{2}\right)+\sin^2\left(\frac{\theta}{2}\right)(x^2+y^2+z^2)\\&=\cos^2\left(\frac{\theta}{2}\right)+\sin^2\left(\frac{\theta}{2}\right)=1\end{aligned}\tag{5.2.272}$$

   and as $\vec{u}(x,y,z)$ is of unit norm, we have:

$$x^2+y^2+z^2=1\tag{5.2.273}$$

   Therefore:

$$|q|=\cos^2\left(\frac{\theta}{2}\right)+\sin^2\left(\frac{\theta}{2}\right)(x^2+y^2+z^2)=\cos^2\left(\frac{\theta}{2}\right)+\sin^2\left(\frac{\theta}{2}\right)=1\tag{5.2.274}$$

2. The fact that q is a quaternion of unit norm immediately leads to the fact that the determinant of the application S_q is also equal to 1. We have already proved it above in the general case of any quaternion of norm 1 (necessary and sufficient condition).

3. It is the same for the conservation of the norm. We have already proved earlier above that this was the case anyway when the quaternion q of norm 1 (necessary and sufficient condition).

4. Let us now prove that all collinear vector to the axis of rotation is projected onto the axis of rotation itself. Let us denote by q' the purely imaginary unitary quaternion $xi+yj+zk$. Then we have:

$$q=\cos\left(\frac{\theta}{2}\right)+\sin\left(\frac{\theta}{2}\right)q'\tag{5.2.275}$$

   Then:

$$S_q(q')=qq'\bar{q}\tag{5.2.276}$$

   but as q' is the restriction of q to the pure elements that constitute it, this is equivalent as to write:

$$S_q(q')=S_q(q)=qq\bar{q}=q\tag{5.2.277}$$

   Let us now show why we choose the writing $\theta/2$. If $\vec{v}=(x_1,y_1,z_1)$ denotes a unit vector orthogonal to $\vec{u}$ (therefore perpendicular to the axis of rotation), and p the quaternion $x\mathrm{i}+y\mathrm{j}+z\mathrm{k}$ then we have:

$$\begin{aligned}S_q(p)&=\left(\cos\left(\frac{\theta}{2}\right)+\sin\left(\frac{\theta}{2}\right)q'\right)p\left(\cos\left(\frac{\theta}{2}\right)-\sin\left(\frac{\theta}{2}\right)q'\right)\\&=\cos^2\left(\frac{\theta}{2}\right)p+\cos\left(\frac{\theta}{2}\right)\sin\left(\frac{\theta}{2}\right)(q'p-pq')-\sin^2\left(\frac{\theta}{2}\right)q'pq'\end{aligned}\tag{5.2.278}$$

We have shown during the definition of multiplication of two quaternions that:

$$pq' = -q'p \tag{5.2.279}$$

therefore we get:

$$\begin{aligned} S_q(p) &= \cos^2\left(\frac{\theta}{2}\right)p + 2\cos\left(\frac{\theta}{2}\right)\sin\left(\frac{\theta}{2}\right)q'p - \sin^2\left(\frac{\theta}{2}\right)q'pq' \\ &= \cos^2\left(\frac{\theta}{2}\right)p + 2\cos\left(\frac{\theta}{2}\right)\sin\left(\frac{\theta}{2}\right)q'p + \sin^2\left(\frac{\theta}{2}\right)q'p(-q') \\ &= \cos^2\left(\frac{\theta}{2}\right)p + 2\cos\left(\frac{\theta}{2}\right)\sin\left(\frac{\theta}{2}\right)q'p + \sin^2\left(\frac{\theta}{2}\right)q'p\bar{q}' \end{aligned} \tag{5.2.280}$$

We have also prove earlier above that:

$$S_q(p) = -p = qp\bar{q} \tag{5.2.281}$$

Therefore:

$$q'p\bar{q}' = S_{q'}(p) = -p \tag{5.2.282}$$

(the half turn of axis (x, y, z)). So:

$$\begin{aligned} S_{q'}(p') &= \cos^2\left(\frac{\theta}{2}\right)p + 2\cos\left(\frac{\theta}{2}\right)\sin\left(\frac{\theta}{2}\right)q'p - \sin^2\left(\frac{\theta}{2}\right)p \\ &= \left(\cos^2\left(\frac{\theta}{2}\right) - \sin^2\left(\frac{\theta}{2}\right)\right)p + 2\cos\left(\frac{\theta}{2}\right)\sin\left(\frac{\theta}{2}\right)q'p \\ &= \cos(\theta)p + \sin(\theta)q'p \end{aligned} \tag{5.2.283}$$

Remark

We are beginning to see here already the usefulness of having chose from the beginning $\theta/2$ for the angle!

■ Q.E.D.

We know that p is the pure quaternion likened to a unit vector $\vec{v}$ orthogonal to the axis of rotation $\vec{u}$ itself equated with the purely imaginary part of q'. We notice then immediately that the imaginary part of the product (defined!) of the quaternion $q'p$ is equal to the cross product $\vec{u} \times \vec{v} = \vec{w}$. This vector product therefore generates a vector perpendicular to $\vec{u}, \vec{v}$.

The pair $(\vec{v}, \vec{w})$ thus form a plane perpendicular to the axis of rotation $\vec{u}$ (that's as for the simple complex numbers $\mathbb{C}$ in which we have the Gaussian plane and perpendicular to it the axis of rotation!).

Then finally:

$$S_q(p) = \cos(\theta)p + \sin(\theta)q'p = \cos(\theta)\vec{v} + \sin(\theta)\vec{w} \tag{5.2.284}$$

We fall back with on rotation based on a plane (but therefore be in space!) identical to that shown earlier above with the standard complex numbers $\mathbb{C}$ in the Gaussian plane. For more details the reader can refer the section of Spinor Calculus page 2145.

So we know how to do any kind of rotation in space in a single mathematical operation and with a bonus: with the free choice of the axis!

We can now better understand why the algebra of quaternions is not commutative. Indeed, the vector rotations of the plan are commutative but those of space are not like show us the example below:

Given the initial configuration:

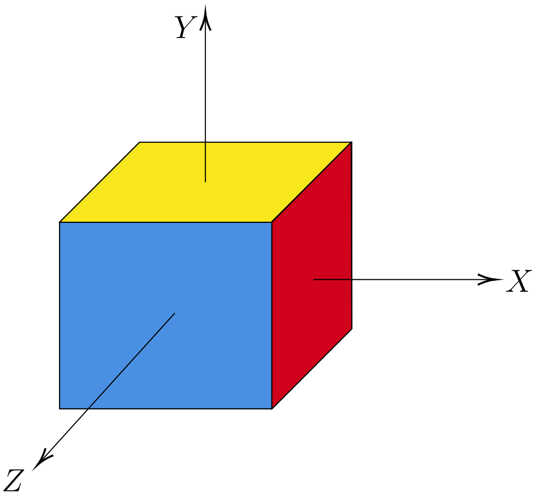


Figure 5.24: **Starting situation for quaternion rotations**

Then a rotation about the X-axis followed by a rotation around the Y axis:

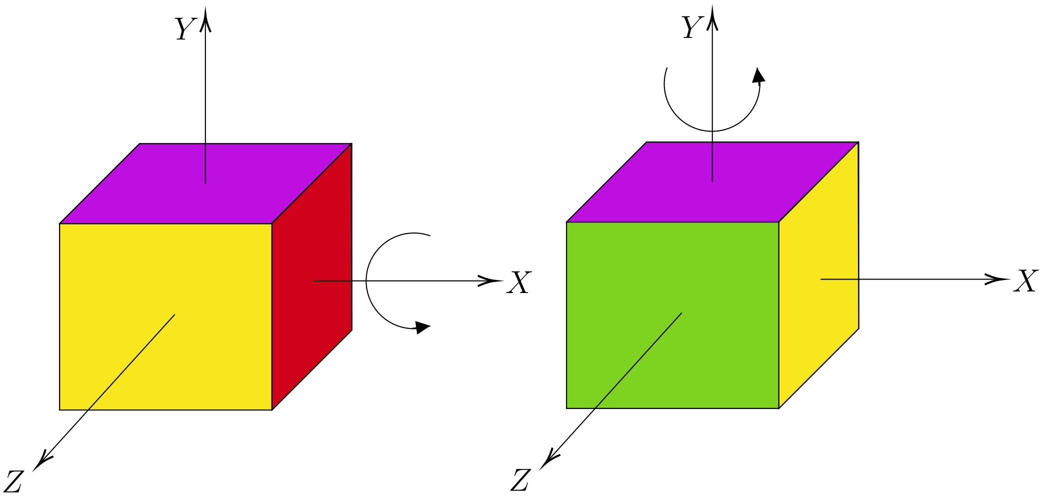


Figure 5.25: **Example quaternion $X - Y$ rotations**

is not equal to a rotation around the Y-axis followed by a rotation about the axis X:

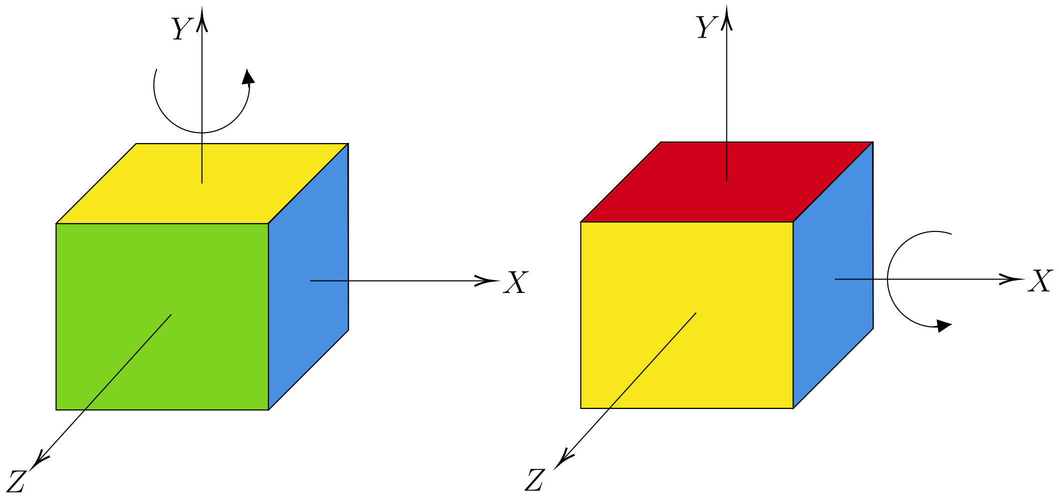


Figure 5.26: **Example of non-equivalence for quaternion rotation**

The results will be fundamental for our understanding of spinors (see section Spinor Calculus page 2145)!

> **Remark**
>
> Notice that we have (where $\mathbb{K}$ is a type of numbers named "Cayley numbers" that we have not planned to present in this book):
>
> $$\mathbb{R} \subset \mathbb{C} \subset \mathbb{H} \subset \mathbb{K} \tag{5.2.285}$$
>
> and that at each stage some element of abstract structure is lost. Ordering is lost on the passage from $\mathbb{R}$ to $\mathbb{C}$. Commutativity disappears on going from $\mathbb{C}$ to $\mathbb{H}$. The Cayley numbers $\mathbb{K}$ lose associativity.

5.2.2.9 Algebraic and Transcendental Numbers

Definitions (#5.2.43):

D1. We name "algebraic integer of degree n", any complex number that is a solution of an univariate algebraic equation of degree n, i.e. a polynomial of degree n (concept that we will discuss in the chapter of Algebra) whose coefficients are integers and whose dominant coefficient is equal to 1.

D2. We name "algebraic number of degree n", any complex number that is a solution of an univariate algebraic equation of degree n, i.e. a polynomial of degree n whose coefficients are rational.

The set of algebraic number is sometimes denoted by $\overline{\mathbb{Q}}$ or $\mathbb{A}$.

Theorem 5.8. A first interesting result and particularly in this area of study (mathematical curiosity...) is that a rational number is an "algebraic integer of degree n" if and only if it's an integer (read several times need...). In scientific terms, we the say that the ring $\mathbb{Z}$ is "fully closed".

Proof 5.8.1. We will assume that the number p/q , where p and q are two prime integers (that is to say that their ratio does not give an integer or more rigorously ... that the greatest common divisor of p,q is equal to 1! , is a root of the following polynomial (see section Calculus page 1368) with relative integer coefficients ($\in \mathbb{Z}$) and whose dominant coefficient is equal to 1:

$$x^n+a_{n-1}x^{n-1}+\ldots+a_1x+a_0 \tag{5.2.286}$$

where the equality with zero of the polynomial is implicit.

In this case:

$$p^n=-(a_{n-1}p^{n-1}+\ldots+a_1pq^{n-2}+a_0q^{n-1})q \tag{5.2.287}$$

Since the coefficients are by definition all integers and their multiple in the parenthesis also, then the parenthesis has necessarily a value in $\mathbb{Z}$.

Therefore, q (at the right of the parenthesis) divides a power of p (at the left of the equality), which is possible, in the set $\mathbb{Z}$ (because our bracket has a value in this same set for recall...), only if q is equal to ± 1 (as they were prime together).

So among all rational numbers the only that are solutions of polynomial equations with relative integer coefficients ($\in \mathbb{Z}$) for which the dominant coefficient is equal to 1 are relative integers!

■ Q.E.D.

To take another interesting and particular case, it is easy to show that any rational number is an algebraic number. Indeed, if we take the simplest following univariate polynomial:

$$qx-p=0 \tag{5.2.288}$$

where p and q are relatively prime and where q is different from 1. So as this is a simple polynomial with rational coefficients ($\in \mathbb{Q}$), after remaniment we have:

$$x=\frac{p}{q} \tag{5.2.289}$$

So since p and q are relatively prime and q is different from 1, we have indeed that every rational number is an "algebraic number of degree 1".

We also have the real (and irrational) number $\sqrt{2}$ which is an "algebraic integer of degree 2" because it is the root of:

$$x^2-2=0 \tag{5.2.290}$$

and the complex number i is also an "algebraic integer of degree 2" because it is the root of the equation:

$$x^2+1=0 \tag{5.2.291}$$

etc.

Definition (#5.2.44): A "transcendental number" is a real or complex number that is not algebraic. That is, it is not a root of a non-zero polynomial equation with rational coefficients.

Theorem 5.9. The set of all transcendental numbers is uncountable. The proof is simple and requires no difficult mathematical development.

Proof 5.9.1. Indeed, since the polynomial with integer coefficients are countable, and since each of these polynomials has a finite number of roots (see the Factorization Theorem in the section Calculus page 1377), the set of algebraic numbers is countable! But the argument of Cantor's diagonal (see section Set Theory page 225) states that real numbers (and therefore also the complex numbers) are uncountable, so the set of all transcendental numbers must be uncountable.

In other words, there is much more transcendental numbers than algebraic numbers...

■ Q.E.D.

The best known transcendent numbers are π and e. We are still looking to provide you a proof more nice and intuitive than that of Hilbert or Lindemann–Weierstrass.

5.2.2.10 Universe Numbers (normal numbers)

Definition (#5.2.45): A "Universe number" also named "normal number" is a real number whose infinite sequence of digits in every base b is distributed uniformly in the sense that each of the b digit values has the same natural density $1/b$. Intuitively this means that no digit, or (finite) combination of digits, occurs more frequently than any other. The set of Universe numbers is sometimes denoted $\mathbb{U}$.

While a general proof can be given that almost all purely real numbers are Universe numbers [181] this proof is not constructive and only very few specific numbers have been shown to be Universe numbers. It is widely believed that the (computable) numbers $\sqrt{2}$, π, and e are Universe numbers, but a proof remains elusive still in this year 12016 (holocene calendar). All of them however are strongly conjectured to be because of some empirical evidence. It is not even known whether all digits occur infinitely often in the decimal expansions of those constants. In particular, the popular claim "every string of numbers eventually occurs in π" or "the whole Holy book is contained in π is not known to be true. It has been conjectured that every irrational algebraic number is a Universe number, while no counterexamples are known, there also exists no algebraic number that has been proven to be a Universe number in any base.

More formally, let $\sum$ be a finite alphabet of b digits, and $\sum^\infty$ the set of all sequences that may be drawn from that alphabet. Let $S \in \sum^\infty$ be such a sequence. For each a in $\sum$ let $N_S(a,n)$ denote the number of times the letter a appears in the first n digits of the sequence S. We say that S is a "simple Universe number" if the limit:

$$\lim_{n\rightarrow+\infty}\frac{N_S(a,n)}{n}=\frac{1}{b} \qquad (5.2.292)$$

for each a is satisfied.

Now let w be any finite string in $\sum^*$ and let $N_S(w,n)$ to be the number of times the string w appears as a substring in the first n digits of the sequence S (for instance, if $S = 01010101\ldots$, then $N_S(010,8) = 3$). Then S is a "Universe number" if, for all finite strings $w \in \sum^*$:

$$\lim_{n\rightarrow+\infty}\frac{N_S(a,n)}{n}=\frac{1}{b^n} \qquad (5.2.293)$$

S is therefore a Universe number if all strings of equal length occur with equal asymptotic frequency. A given infinite sequence is either a Universe number or not, whereas a pure real number, having a different base-b expansion for each integer $b \geq 2$, may be a Universe number in one base but not in another. A "disjunctive sequence" is a sequence in which every finite string appears. A Universe number sequence is a "disjunctive sequence" but a disjunctive sequence need obviously not be a Universe number.

It is possible to prove (yet we don't wish not present this proof in a book on applied mathematics) with the "Universe number theorem" that almost all pure real numbers are Universe number. The set of non-Universe numbers, though "small" in the sense of being a null set, is "large" in the sense of being uncountable (for example no rational number is normal to any base, since the digit sequences of rational numbers are eventually periodic!). For instance, there are uncountable many numbers whose decimal expansion does not contain the digit 5, and none of these are Universe number.

5.2.2.11 Abstract Numbers (variables)

Definition (#5.2.46): A number may be considered as doing abstraction from the nature of the objects that constitute the group that it characterizes as well as how to codify it (Indian notation, Roman notation, etc.). We then say that the number is an "abstract number". In other words, an abstract number, is a number that does not designate the quantity of any particular kind of thing.

Remark

Arbitrarily, the human being has adopted a numerical system mainly used in the World and represented by the symbols $0, 1, 2, 3, 4, 5, 7, 8, 9$ of the decimal system that will be supposedly known both in writing than orally by the reader (language learning).

For mathematicians, it is not advantageous to work with these symbols because they represent only specific cases. What seek theoretical physicists and mathematicians are "literal relations" applicable in a general case and that engineers can according to their needs change these abstract numbers by numeric values that correspond to the problem they need resolve.

These abstract numbers today commonly named "variable" or "unknown", used in the context of "literal calculation" are very often represented since the 116th century (holocene calendar) by:

1. The Latin alphabet:

$$a, b, c, d, e, \ldots, x, y, z; A, B, C, D, E, \ldots, X, Y, Z$$

   where the first lower case letters of the latin alphabet ($a, b, c, d, e \ldots$) are often used to represent an abstract constant, while the lowercase letters of the end of the latin alphabet ($\ldots, x, y, z$) are used to represent entities (variables or unknowns) we seek the value. Commonly, uppercase latin letters are reserved to represent matrices or random variables.

2. The Greek alphabet:

$\mathrm{A}\alpha$	Alpha	$\Lambda\lambda$	Lambda
$\mathrm{B}\beta$	Beta	$\mathrm{M}\mu$	Mu
$\Gamma\gamma$	Gamma	$\mathrm{N}\nu$	Nu
$\Delta\delta$	Delta	$\Xi\xi$	Xi
$\mathrm{E}\epsilon\varepsilon$	Epsilon	$\mathrm{O}o$	Omicron
$\mathrm{Z}\zeta$	Zeta	$\Pi\pi$	Pi
$\mathrm{H}\eta$	Eta	$\mathrm{P}\rho$	Rho
$\Theta\theta\vartheta$	Theta	$\Sigma\sigma$	Sigma
$\mathrm{I}\iota$	Iota	$\mathrm{T}\tau$	Tau
$\mathrm{K}\kappa$	Kappa	$\Upsilon\upsilon$	Upsilon
$\Phi\phi\varphi$	Phi	$\mathrm{X}\chi$	chi
$\Psi\psi$	Psi	$\Omega\omega$	Omega

Table 5.17: **Greek Alphabet**

   which is particularly used to represent more or less complex mathematical operators (such as the index sum Σ, the indexed product Π, the variational δ, the infinitesimal element ε, partial differential ∂, etc.) or variables in the field of physics (as ω for the pulsation, ν for the frequency, ρ for the density, etc.).

3. The modernized Hebrew alphabet (with less intensity...)

   As we have seen, a transfinite cardinal for example is denoted by the letter "aleph": $\aleph_0$.

Although these symbols can represent any number there are some who can represent physical constants also named "Universal constant" as the speed of light c, the gravitational constant G, the Planck constant h, the number π, etc.

We use very often still other symbols that we will introduce and define when reading this book.

Remark

The letters to represent numbers had been used for the first time by Vieta in the 116th century (holocene calendar).

5.2.2.11.1 Domain of a Variable

A variable is therefore likely to take different numerical values. All these values can vary according to the character of the problem considered.

Given two numbers a and b such that $a < b$, then:

Definitions (#5.2.47):

D1. We name "domain of definition" of a variable, all numerical values it is likely to take between two specified limits (endpoints) or on a set (like $\mathbb{N},\mathbb{R},\mathbb{R}^+$, etc.).

D2. We name "closed interval with endpoints a and b", the set of all numbers x between these two values and we denote as example as follows:

$$[a,b]=\{x\in\mathbb{R}\mid a\leq x\leq b\} \tag{5.2.294}$$

The left notation is named obviously "interval notation", the right one is named "set-builder notation".

D3. We name "open interval with endpoints a and b", the set of all numbers x between these two values not included and we denote it as example as follows:

$$]a,b[=\{x\in\mathbb{R}\mid a< x< b\} \tag{5.2.295}$$

D4. We name "interval closed, left open right" or "semi-closed left" the following relation as example:

$$[a,b[=\{x\in\mathbb{R}\mid a\leq x< b\} \tag{5.2.296}$$

D5. We name "interval open left, closed right" or "semi-closed right" the following relation as example:

$$]a,b]=\{x\in\mathbb{R}\mid a< x\leq b\} \tag{5.2.297}$$

Or in a summary and imaged form and as often denoted in Switzerland:

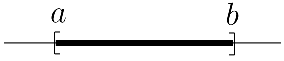
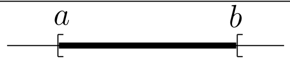
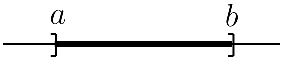
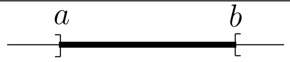
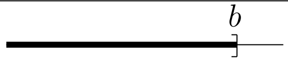
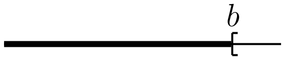
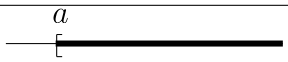
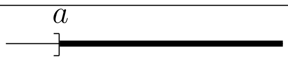
Type	Visual	Math notation	Explicitly
$[a,b]$		$a \leq x \leq b$	Closed bounded interval
$[a,b[$		$a \leq x < b$	Semi-closed and bounded interval on a and semi-open on b (or left semi-closed and right semi-open)
$]a,b]$		$a < x \leq b$	Semi-open bounded interval on a and semi-closed on b (or left semi-open and right semi-closed)
$]a,b[$		$a < x < b$	Bounded open interval
$]-\infty,b]$		$x \leq b$	Unbounded interval closed on b (or closed right)
$]-\infty,b[$		$x \leq b$	Unbounded interval open on b (or open right)
$[a,+\infty[$		$a \leq x$	Unbounded interval closed on a (or closed left)
$]a,+\infty[$		$a < x$	Unbounded interval open on a (or open left)

Table 5.18: **Summary of typical domain of definitions in Switzerland**

and according to the international norm ISO 80000-2:2009 *Mathematical signs and symbols to be used in the natural sciences and technology* (since Switzerland is quite good in the no respect of international norms and standards):

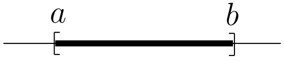
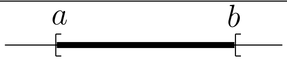
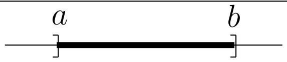
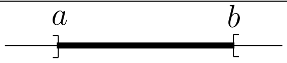
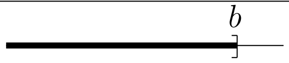
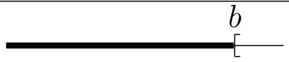
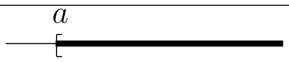
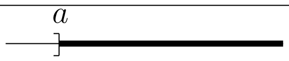
Type	Visual	Math notation	Explicitly
$[a,b]$		$a \leq x \leq b$	Closed bounded interval
$[a,b)$		$a \leq x < b$	Semi-closed and bounded interval on a and semi-open on b (or left semi-closed and right semi-open)
$(a,b]$		$a < x \leq b$	Semi-open bounded interval on a and semi-closed on b (or left semi-open and right semi-closed)
(a,b)		$a < x < b$	Bounded open interval
$(-\infty,b]$		$x \leq b$	Unbounded interval closed on b (or closed right)
$(-\infty,b[$		$x \leq b$	Unbounded interval open on b (or open right)
$[a,+\infty)$		$a \leq x$	Unbounded interval closed on a (or closed left)
$(a,+\infty)$		$a < x$	Unbounded interval open on a (or open left)

Table 5.19: Summary of typical domain of definitions according to ISO norms

Remarks

R1. The notation $\{x \text{ such that } a < x < b\}$ denotes the set of real numbers x strictly greater than x and strictly less than b.

R2. To fact that an interval is for example opened on b means that the real number b is not part thereof. By cons, if it had been closed then b would be part of it.

R3. If the variable x can take all possible negative and positive values we write therefore: $]-\infty,+\infty[$ where the symbol "∞" means "infinite". Obviously there can be combinations of open infinite right intervals with left endpoint and vice versa.

R4. We will recall some of these concepts with a different approach when studying Algebra (literal calculation).

We say that the variable x is an "ordered variable" if by representing its domain of definition by a horizontal axis where each point on the axis represents a value of x, then for each pair of values, we can say that there is an "antecedent" and one that is a "subsequent". Here the notion of antecedent and subsequent is not related to the concept of time it expresses just how the values of the variable are ordered.

Definitions (#5.2.48):

D1. A variable is said to be "increasing" if each subsequent value is greater than each antecedent value.

D2. A variable is said to be "decreasing" if each subsequent value is smaller than each antecedent value.

D3. The increasing and decreasing variables are named "variables with monotonic variations" or simply "monotonic variables".

∼90 %

★★★★☆

31 votes, 69.68%

5.3. Arithmetic Operators

TALKING about numbers like we did in the previous section naturally leads us to consider the operations of calculus. It is therefore logic that we make a non-exhaustive description of the operations that may exist between the numbers. This will be the goal of this section.

We will consider in this book that there are two types of key tools in Arithmetics (we do not speak of Algebra but Arithmetic!):

- Arithmetic operators:

  There are two basic operators (addition "$+$" and subtraction "$-$") from which we can build other operators: the "multiplication" (whose contemporary symbol $\times$ was introduced in 11574 - according to holocene calendar - by William Oughtred) and the "division" (whose old symbol was "$\div$" but since the end of the 120th century (holocene calendar) we use simple the slash $/$ symbol).

  These four operators $(+, -, \times, /)$ are commonly named "rational operators". We will see them more in details after setting the binary relations.

  Remark

  Rigorously addition could be enough if we consider the common set of real number $\mathbb{R}$ because therefore the subtraction is only the addition of a negative number.

- Binary operators (relations):

  There are six basic binary relations (equal $=$, different $\neq$, greater than $>$, less than $<$, greater or equal $\geq$, less than or equal $\leq$) that compare the order of amplitude of elements that are on the left and on the right of these relations (thus at the number of two, hence the name "binary") in order to draw some conclusions. The majority of binary relations symbols were introduced by Vieta and Harriot in the 116th century (holocene calendar).

It is obviously essential to know as best a possible these tools and their properties before going through into more strenuous calculations.

5.3.1 Binary Relations

The concept of "relation" is the basis of all mathematics whose purpose is to study - by observation and deduction (reasoning), calculation and comparison - configurations or abstract/concrete relations of its objects (numbers, forms, structures) by seeking to establish the logical, numerical or conceptual links between these objects.

Definitions (#5.3.49):

D1. Consider two non-empty sets E and F (see section Set Theory page 383) not necessarily identical. If to some given elements x of A we can associate with a precise mathematical rule $\mathcal{R}$ (unambiguous) one element y of F, we define therefore a "functional relation" that maps E to F and that we write:

$$\mathcal{R} : E \mapsto F \tag{5.3.1}$$

Thus, more generally, a functional relation $\mathcal{R}$ can be defined as a mathematical rule that associates to given components x of E, some given elements y of F.

So, in this more general context, if $x\mathcal{R}y$, we say that there y is an "image" of x through $\mathcal{R}$ and that x is a "precedent" or "preimage" of y.

The set of pairs (x, y) such that $x\mathcal{R}y$ is a true statement generates a "graph" or "representation" of the relation $\mathcal{R}$. We can represent these couples in a proper chosen way to make a graphical representation of the relation $\mathcal{R}$.

This is a type of relation on which we will come back in the section Functional Analysis (page 2214) under the form: $\mathcal{R} : f(x) = y \circ f$ and that does not interest us directly in this section.

D2. Consider a non-empty set E, if we associate with this set (and only to this one!) tools to compare its items between them then we talk about a "binary relation" or "comparison relation" and that we write for any element x and y of E:

$$x\mathcal{R}y \tag{5.3.2}$$

These relations can also most of time be presented graphically. In the case of conventional binary operators comparison where E is the set of natural numbers $\mathbb{N}$, relative $\mathbb{Z}$, rationals $\mathbb{Q}$ or real $\mathbb{R}$, that is graphically represented by a horizontal line (typically...); in the case of congruence (see section Number Theory page 356) it is represented by lines in the plane whose points are given by the constraint of congruence.

As we already mentioned, there are 6 fundamental binary relationships (equal $=$, different from $\neq$, greater than $>$, smaller than $<$, bigger or equal $\geq$, smaller or equal $\leq$). But we will see a little further that the rigorous definition of binary relations allows us to build more abstract tools (such as congruence well known by students of small classes and that we will study in the section of Number Theory).

5.3.1.1 Equalities

It is difficult to define the term "equality" in a general case applicable to any situation. For our part, we will allow ourselves for this definition to take the inspiration of the extensionality axiom of Set Theory (discussed later at page 382).

Definitions (#5.3.50):

D1. Two elements are "equal" if and only if they have the same values. The strict equality is denoted by the symbol $=$ that therefore means "equal to" (this symbol was introduced in 11557 - according to holocene calendar - by Robert Rocorde).

If we have $a = b$ and c is any given number (or vector/matrix) and $\star$ any operation (such as addition, subtraction, multiplication or division) then:

$$a \star c = b \star c \tag{5.3.3}$$

This property is used to solve or simplify any type of equations. In practice, the abbreviation "LHS" is informal shorthand for the left-hand side of an equality. Similarly, "RHS" is the right-hand side abbreviation of that latter.

Obviously we have (property of reflexivity):

$$a = b \Leftrightarrow b = a \tag{5.3.4}$$

And also (property of transitivity):

$$\left.\begin{array}{l} a = c \\ b = c \end{array}\right\} a = b$$

We will not enumerate the other properties of the equality in this section (for more details see the section Set Theory page 287).

D2. If two elements are not strictly equal, that is to say "inequal"..., we are connecting them by the symbol $\neq$ and we say they are "not equal".

If we have $a > b$ or $a < b$ then:

$$a \neq b \tag{5.3.5}$$

There are still other equality symbols, which are an extension of two we have defined previously. Unfortunately, they are often misused (we could say rather that they are used in the wrong places) in most of the books available on the market (and this book is not an exception):

1. $\cong$: Should be used for congruence but in fact is mostly used to indicate an approximation.
2. $\approx$: Should be used for approximations but in fact $\cong$ is used instead.
3. $\equiv$: Should be used to say that two elements are equivalent but in practice most people use $=$.

4. $:=$: Is used to say that one element is by definition equal to another one.

5. $\doteq$: Should be used to say "equal by definition to" but in fact most people use instead $:=$.

6. $\sim$: Is used most of time in Statistics to say "follows the law..." but some practitioners use instead $=$ or to say "asymptotically equal".

5.3.1.2 Comparators

The comparators are tools that allow us to compare and order any pair of numbers (and also Sets!).

The possibility of ordering numbers is fundamental in mathematics. Otherwise (if it was not possible to order them), there would be a lot of things that would shock our habits, for example (some of the concepts presented in the following sentence have not yet been presented but we would still make reference to them): no more monotonic functions (especially sequences) and linked to it the derivation would therefore indicate nothing more about the "variation direction", no more approach of roots of polynomial by dichotomy (classical research algorithm in an ordered set that split in two at each iteration), no more segments in geometry, no more than half space, no more convexity, we can not oriented space anymore, etc. It is therefore important to be able to order things as you can see...!

Thus, for any $a, b, c \in \mathbb{R}$ we write when a is greater than or equal to b:

$$a \geqslant b \tag{5.3.6}$$

and when a is less than or equal to b:

$$a \leqslant b \tag{5.3.7}$$

Remark

It is useful to recall that the set of real numbers $\mathbb{R}$ is a totally ordered group (see section Set Theory page 421), otherwise we could not establish order relations among its elements (which is not the case for complex numbers $\mathbb{C}$ that we can not order!).

Definition (#5.3.51): The symbol $\leq$ is an "relation order" (see the rigorous definition further below!) which means "less than or equal to" and conversely the symbol $\geq$ is also an order relation that means "greater than or equal to".

We also have relatively to the strict comparison the following properties that are relatively intuitive:

$$\begin{aligned} a<b \quad &\text{and} \quad b<c \quad \Rightarrow \quad a<c \\ a>b \quad &\text{and} \quad b>c \quad \Rightarrow \quad a>c \end{aligned} \tag{5.3.8}$$

and:

$$\begin{aligned} a>b \quad &\text{and} \quad b=c \quad \Rightarrow \quad a>c \\ a<b \quad &\text{and} \quad b=c \quad \Rightarrow \quad a<c \end{aligned} \tag{5.3.9}$$

if:

$$a < b \quad \text{and} \quad c > 0 \quad \Rightarrow \quad ac < bc \tag{5.3.10}$$

if:

$$a < b \Rightarrow a + c < b + c \qquad \text{and} \qquad a < b \Rightarrow a - c < b - c \tag{5.3.11}$$

and vice versa:

$$a > b \Rightarrow a + c > b + c \qquad \text{and} \qquad a > b \Rightarrow a - c > b - c \tag{5.3.12}$$

We also have:

$$0 < a < b \Rightarrow \frac{1}{a} > \frac{1}{b} \tag{5.3.13}$$

and vice versa:

$$b < a < 0 \Rightarrow \frac{1}{a} < \frac{1}{b} \tag{5.3.14}$$

We can obviously multiply, divide, add or subtract a term from each side of the relation as it is always true. Notice, however, that if you multiply both sides by a negative number it will obviously change as the comparator such that:

$$a > b \quad \text{and} \quad c > 0 \quad \Rightarrow \quad ac < bc \tag{5.3.15}$$

and vice versa:

$$a < b \quad \text{and} \quad c < 0 \quad \Rightarrow \quad ac > bc \tag{5.3.16}$$

We also have:

$$0 < a < b \quad \text{and} \quad p \in \mathbb{R}^{*+} \quad \Rightarrow \quad a^p < b^p \tag{5.3.17}$$

Consider now that $b < a < 0$ and $p \in \mathbb{N}^*$. Then if p is an even integer:

$$0 < a^p < b^p \tag{5.3.18}$$

else if p is odd:

$$a^p > b^p \tag{5.3.19}$$

This result simply comes from the multiplication of signs rule since the power when not fractional is only a multiplication.

Finally:

$$0 < a < b \quad \text{and} \quad n \in \mathbb{N}^* \quad \Rightarrow \quad \sqrt[n]{a} < \sqrt[n]{b} \tag{5.3.20}$$

The relations:

$$\boxed{> \quad < \quad \leqslant \quad \geqslant \quad \gg \quad \ll} \tag{5.3.21}$$

thus correspond respectively to: (strictly) greater than, (strictly) smaller than, smaller or equal, greater or equal, much bigger than, much smaller than.

These relations can be defined in a little more subtle and rigorous way and apply not only to comparators (see for example the congruence relation in the section of Set Theory page 356)!

Let us see this (the vocabulary that follows is also defined in the section of Set Theory page 406):

Definition (#5.3.52): Given a binary relation $\mathcal{R}$ of a set A to itself, a relation R on A is a subset of the Cartesian product $\mathcal{R} \subseteq A \times A$ (that is to say, the binary relation generates a subset by the constraints it imposes on the elements of A satisfying the relation) with the property of being:

P1. A "reflexive relation" if $\forall x \in A$:

$$x\mathcal{R}x \tag{5.3.22}$$

P2. A "symmetrical relation" if $\forall x, y \in A$:

$$x\mathcal{R}y \Rightarrow y\mathcal{R}x \tag{5.3.23}$$

P3. An "anti-symmetrical relation" if $\forall x, y \in A$:

$$(x\mathcal{R}y \quad \text{and} \quad y\mathcal{R}x) \Rightarrow x = y \tag{5.3.24}$$

P4. A "transitive relation" if $\forall x, y, z \in A$:

$$(x\mathcal{R}y \quad \text{and} \quad y\mathcal{R}z) \Rightarrow x\mathcal{R}z \tag{5.3.25}$$

P5. An "connex relation" if $\forall x, y \in A$:

$$\forall x, y \in A \Rightarrow x\mathcal{R}y \quad \text{or} \quad y\mathcal{R}x \tag{5.3.26}$$

Mathematicians have given special names to the families of relations satisfying some of these properties.

Definitions (#5.3.53):

D1. A relation is named "strict order relation" if and only if it is only transitive (some specify then that it is necessarily antireflexive but this last fact is then obvious...).

D2. A relation is named a "preorder" if and only if it is reflexive and transitive.

D3. A relation is named an "equivalence relation" if and only if it is reflexive, symmetric, and transitive.

D4. A relation is named "order relation" if and only if it is reflexive, transitive and antisymmetric (thus the relations $>, <$ are not order relations because obviously not reflexive relations).

D5. A relation is named "total order relation" if and only if it is reflexive, transitive, connex and antisymmetric.

For the other combinations it seems (as far as we know) that there are no special name among the mathematicians...

Remark

The binary relations have all similar properties in natural sets $\mathbb{N}$, relative $\mathbb{Z}$, rational $\mathbb{Q}$ and real $\mathbb{R}$ (there is no natural order relation on the set of complex numbers $\mathbb{C}$).

✍**Example:**

Of the below set of eight copies of books, the relation "... has the same ISBN as ..." is an equivalence relation:

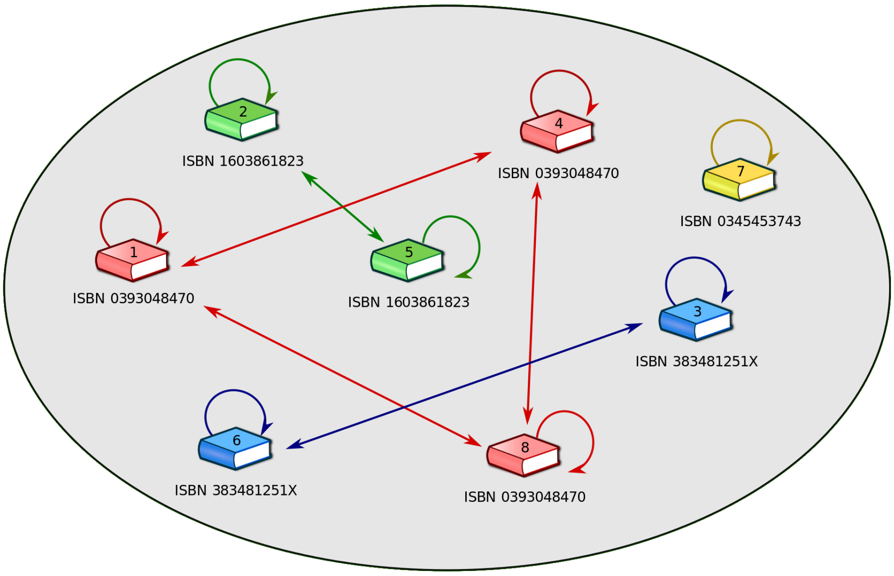

2
ISBN 1603861823
4
ISBN 0393048470
7
ISBN 0345453743
1
ISBN 0393048470
5
ISBN 1603861823
3
ISBN 383481251X
6
ISBN 383481251X
8
ISBN 0393048470

Figure 5.27: **Illustration of an equivalence relation (source: Wikipedia, author: Stephan Kulla)**

If we summarize:

BINARY RELATION	$=$	$\neq$	$>$	$<$	$\leqslant$	$\geqslant$
reflexive	yes	no	no	no	yes	yes
symmetric	yes	yes	no	no	no	no
transitive	yes	no	yes	yes	yes	yes
connex	no	no	no	no	yes	yes
antisymmetric	yes	no	no	no	yes	yes

Table 5.20: **Binary Relations**

Thus we see that the binary relations $\leq, \geq$ form with the previously mentioned sets, total order relations and it is very easy to see which binary relations are partial, total or equivalence order relations.

Definition (#5.3.54): If $\mathcal{R}$ is an equivalence relation on A. For $\forall x \in A$, the "equivalence class" of x is by definition the set:

$$\boxed{[x] = \{y \in A : x\mathcal{R}y\}} \tag{5.3.27}$$

$[x]$ is therefore a subset of A ($x \subseteq A$) which we denote also thereafter ... R.

We thus have a new set that is named the "set of equivalence classes" or "quotient set" denoted in this book by A/R. So:

$$\boxed{A/\mathcal{R} = \{[x] | x \in A\}} \tag{5.3.28}$$

You should know that in $A/\mathcal{R}$ we do not look anymore at $[x]$ as a subset of A, but as an element!

A relation of equivalence, presented in a popularized manner... thus serves to stick one unique label to items that satisfy the same property, and to confuse them with the said label (knowing what we do with this label).

✍**Examples:**

E1. Set of the classes of equivalence of the previously illustration about the ISBN:

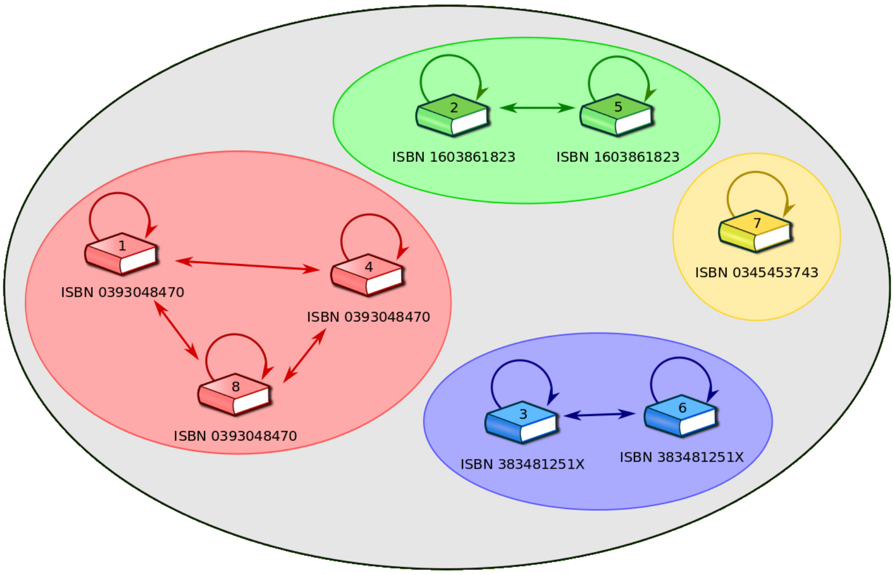


Figure 5.28: **Illustration of set of equivalence classes (source: Wikipedia, author: Stephan Kulla)**

E2. In the set of integers $\mathbb{Z}$, if we study the remains of the division of number by 2, we have that the result is always 0 or 1.

The zero equivalence class is then named the "set of even integers numbers", the one equivalence class is therefore named the "set of odd integers". So we have two classes of equivalence for two partitions of $\mathbb{Z}$ (always keep in mind this simple example for theoretical elements that follow it helps a lot!).

If we name the first 0 and the second 1, we fall back on the operation rules between odd and even numbers:

$$0+0=0 \qquad 0+1=0 \qquad 1+1=0$$

which respectively means that the sum of two even integers is even, that the sum of an even and an odd integer is odd and that the sum of two odd integer is even.

And for the multiplication:

$$0\times 0=0 \qquad 0\times 1=0 \qquad 1\times 1=1$$

which respectively means that the two product of two even integer is even, the product of an even and an odd integer is even and that the product of two odd integer is odd.

Now, to verify that we are dealing with an equivalence relation, we should still check that it is reflexive ($x\mathcal{R}x$), symmetrical (if $x\mathcal{R}y$ then $y\mathcal{R}x$) and transitive (if $x\mathcal{R}y$ and $y\mathcal{R}z$ then $x\mathcal{R}z$). We will see how to check it a few paragraphs further below because this example is a very special case of congruence relation.

Definition (#5.3.55): The application $f: A \mapsto A/\mathcal{R}$ defined by $x \mapsto [x]$ is named "canonical projection". Any element $z \in [x]$ is therefore named "class representative" of $[x]$.

Theorem 5.10. Now consider a set E. Then we propose to proved that there is correspondence between the set of equivalence relations on E and all partitions of E. In other words, this theorem says that an equivalence relation on E is nothing more but a partition on E (this is intuitive).

Proof 5.10.1. Let $\mathcal{R}$ be an equivalence relation on E. We choose $I = E/\mathcal{R}$ as set partition indexing and all we ask for any $[x] \in E/\mathcal{R}$, $E_{[x]} = [x]$.

We just have to check the following two properties of the definition of partitions to show that the family $\left(E_{[x]}\right)_I$ is a partition of E:

P1. Given $[x],[y] \in E/\mathcal{R}$ such that $[x] \neq [y]$ then (obvious) $E_{[x]} \cap E_{[y]} = \varnothing$.

P2. $E = \bigcup\limits_{[x]\in E/\mathcal{R}}$ is obvious because if $x \in E$ then $x \in [x] = E_{[x]}$.

■ Q.E.D.

Again, it should by easy to check with the practical example of the division by 2 given previously that the partition of even and odd numbers satisfies these two properties (if not reader can contact us we will add this as an example).

We have therefore associated to the equivalence relation $\mathcal{R}$ a partition E. Conversely, if $(E_i)_I$ is a partition of E then we almost easily verify that the relation $\mathcal{R}$ is defined by $x\mathcal{R}y$ if and only if there exists $i \in I$ such as $x, y \in E_i$ is an equivalence relation! Both applications are thus bijective and the inverses of each other.

✍**Example:**

We will now apply an example a little less trivial than the last we have seen to the construction of rings $\mathbb{Z}/d\mathbb{Z}$ after a few reminders equation (for the concept of ring see the section Set Theory page 426).

Reminders:

1. Given two numbers $n, m \in \mathbb{Z}$. We say that "n divides m" and we write $n|m$ if and only if there exists an integer $k \in \mathbb{Z}$ such as $m = kn$ (see section Numbers Theory page 313).
2. Given $d \geq 1$ is an integer. We define the relation $\mathcal{R}$ by $n\mathcal{R}m$ if and only if $d|(n-m)$ or in other words $n\mathcal{R}m$ if and only if there exists $d \in \mathbb{Z}$ such that $n = m + kd$. Usually we write this $n \equiv m \pmod{d}$ instead of $n\mathcal{R}m$ and we say that "n is congruent to m modulo d". Remember also that $n \equiv 0 \pmod{d}$ if and only if d divides n (see section Numbers Theory page 356).

We will now introduce an equivalence relation on $\mathbb{Z}$. Let us prove that for any integer $d \geq 1$, the congruence modulo d is an equivalence relation on $\mathbb{Z}$ (we have already proved this in the section of Number Theory in our study of congruence on page 356, but let us redo this work for the fun...).

To prove this we simply have to control the three properties of the equivalence relation:

P1. Reflexivity: $n \equiv n$ since $n = n + 0d$.

P2. Symmetry: If $n \equiv m$ then $n = m + kd$ and therefore $m = n + (-k)d$ that is to say $m \equiv n$.

P3. Transitivity: If $n \equiv m$ and then $m \equiv j$ then $n = m + kd$ and $m = j + k'd$ therefore $n = j + (k + k')d$ that is to say $n \equiv j$.

In the above situation, we denote by $\mathbb{Z}/d\mathbb{Z}$ the set of equivalence classes and we will denote by $[n]_d$ the equivalence class of congruence of a given integer n given by:

$$[n]_d = \{\ldots, n-2d, n-d, n, n+d, n+2d, n+3d, \ldots\} \tag{5.3.29}$$

(each difference of two values in the braces is divisible by d and this is therefore an equivalence class), thus:

$$\mathbb{Z}/d\mathbb{Z} = \{[0]_d, [1]_d, [2]_d, \ldots, [d-1]_d\} \tag{5.3.30}$$

In particular (trivial since we obtain thus the all $\mathbb{Z}$):

$$\mathbb{Z}/2\mathbb{Z} = \{[0]_2, [1]_2\} \tag{5.3.31}$$

Thus, we see that the first example we gave with the even and odd numbers is a particularly simple case of congruence equivalence classes modulo 2 because they all reduce to only two classes.

Remark

The operations of addition and multiplication on $\mathbb{Z}$ define also the operation of addition and multiplication on $\mathbb{Z}/d\mathbb{Z}$. Then we say that these operations are compatible with the equivalence relation and then form a ring (see section Set Theory page 426).

5.3.2 Fundamental Arithmetic Laws

As we have said before, there is a fundamental operator (addition) from which we can define multiplication, subtraction (provided that the chosen Numbers Set is adapted to it....) and division (provided that the chosen Numbers Set is also adapted to it....) and around which we can build the entire Analytical Mathematics.

Obviously there are some subtleties to be considered when the level of rigour increase. The reader can then refer to the whole section of Set Theory (page 378) where fundamental laws are redefined more accurately than what will follow.

5.3.2.1 Addition

Definition (#5.3.56): The "addition" of integers is an operation denoted "$+$" which has for only purpose to bring together in one number all the units contained in several others. The result of the operation is named the "sum", the "total" or "cumul". The numbers to be added are named therefore "terms of the addition".

Remark

The signs of addition "$+$" and subtraction "$-$" are due to the German mathematician Johannes Widmann (11489 according to holocene calendar).

Thus, $A + B + C\ldots$ are the terms of the addition and the result is the sum of the terms of the addition.

Or in schematic form of a special case:

$$0+4+3=4+3=7$$

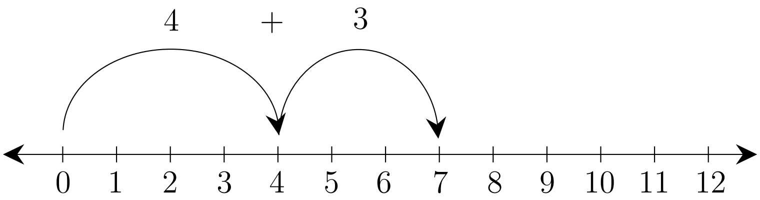


Figure 5.29: **One possible schema for addition**

Here is a list of some intuitive properties that we assume without proofs (as in fact they are axioms) of the operation of addition:

P1. The sum of several numbers do not depend on the order of terms. Then we say that the addition is a "commutative operation". This means concretely for any two numbers:

$$A+B=B+A \tag{5.3.32}$$

P2. The sum of several numbers does not change if we replace two or more of them by their intermediate result. Then we say that the addition is an "associative operation":

$$(A+B)+C=A+(B+C) \tag{5.3.33}$$

P3. The Zero is the neutral element of addition because any number added to zero gives that number:

$$A+0=0 \tag{5.3.34}$$

P4. Depending on the set in which we work ($\mathbb{Z},\mathbb{Q},\mathbb{R},\ldots$), the addition may include a term in such a way that a sum is zero. Then we say that there exists an "opposite" to the sum such as:

$$A+\bar{A}=0 \tag{5.3.35}$$

We have define more rigorously the addition using the Peano axioms in the particular case of all natural numbers $\mathbb{N}$ as we have already see in the section Numbers (page 215). So, with these axioms it is possible to prove that there exists one and only one application (uniqueness), denoted "+" of $\mathbb{N}\times\mathbb{N}$ in $\mathbb{N}$ satisfying:

$$\begin{aligned}
\forall n\in\mathbb{N}: &\quad n+0=n\\
\forall p\in\mathbb{N},\forall q\in\mathbb{N}: &\quad p+s(q)=s(p+q)\\
\forall n\in\mathbb{N}: &\quad s(n)=n+1
\end{aligned} \tag{5.3.36}$$

where s means "successor".

Remark

As this textbook has not be written for mathematicians, we will pass the proof (relatively long and of little interest in the case of business) and we will assume that the application "+" exists and is unique ... and that it follows from the above properties.

Let $x_1, x_2, \ldots, x_n$ be any numbers then we can write the sum as following:

$$\boxed{x_1 + x_2 + \ldots + x_n = \sum_{i=1}^{n} x_i} \tag{5.3.37}$$

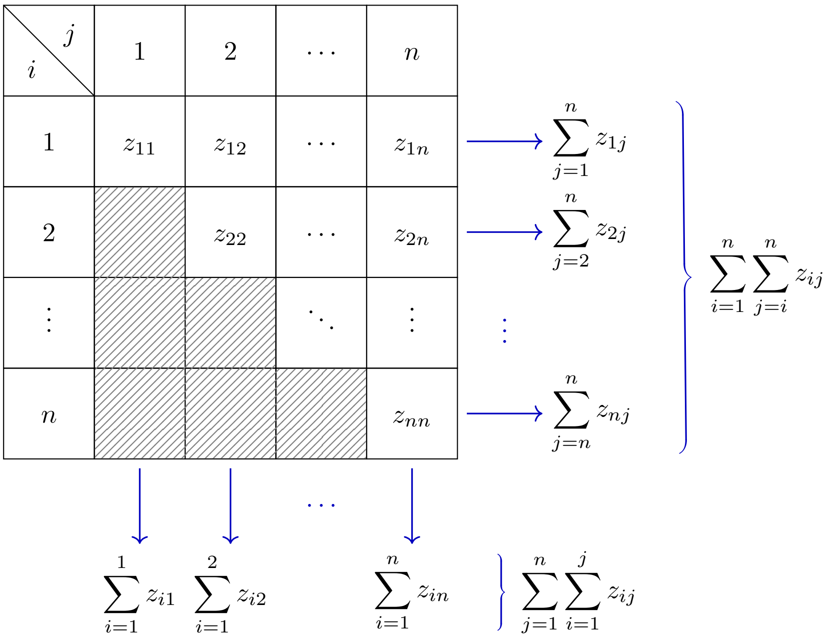
by defining upper and lower bound to the indexed sum (below and above the uppercase greek symbol Sigma).

Here are some properties relatively to this condenses notation that should be obvious (if not the reader can send us a request we will add the details):

$$\sum_{i=1}^{n} kx_i = k\sum_{i=1}^{n} x_i \qquad \sum_{i=1}^{n} k = nk \qquad \sum_{i=1}^{n}(x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i \tag{5.3.38}$$

where k is a constant.

Such a notation is especially useful when we have to sum for example double entry tables like the special case illustrated below (symmetric square matrix):



j
i
1
2
$\cdots$
n
1
z_{11}
z_{12}
$\cdots$
z_{1n}
2
z_{22}
$\cdots$
z_{2n}
$\vdots$
$\ddots$
$\vdots$
n
z_{nn}
$\sum_{j=1}^{n} z_{1j}$
$\sum_{j=2}^{n} z_{2j}$
$\vdots$
$\sum_{j=n}^{n} z_{nj}$
$\sum_{i=1}^{n}\sum_{j=i}^{n} z_{ij}$
$\sum_{i=1}^{1} z_{i1}$
$\sum_{i=1}^{2} z_{i2}$
$\cdots$
$\sum_{i=1}^{n} z_{in}$
$\sum_{j=1}^{n}\sum_{i=1}^{j} z_{ij}$

Let us see now some concrete examples of additions of various simple number in the purpose to practice the basis:

✍**Examples:**

The addition of two numbers relatively small is quite easy since we have learn by heart to count to a number resulting of the operation. Therefore (examples taken on decimal basis):

$$\begin{array}{rr} & 5 \\ + & 2 \\ \hline & 7 \end{array} \tag{5.3.39}$$

and:

$$\begin{array}{rcc} & 1 & 0 \\ + & & 3 \\ \hline & 1 & 3 \end{array} \tag{5.3.40}$$

and:

$$\begin{array}{rcccc} & 1 & 0 & 1 & 4 \\ + & & & & 3 \\ \hline & 1 & 0 & 1 & 7 \end{array} \tag{5.3.41}$$

For more bigger number we can adopt another method that human must also learn by heart. For example:

$$\begin{array}{rcccc} & 9 & 2 & 4 & 4 \\ + & 3 & 4 & 7 & 5 \\ \hline & & & & ? \end{array} \tag{5.3.42}$$

The algorithm (process) is therefore the following: We add the columns (4 columns in this example) from right to left. For the first column we have therefore $4+5=9$ this gives:

$$\begin{array}{rcccc} & 9 & 2 & 4 & 4 \\ + & 3 & 4 & 7 & 5 \\ \hline & & & & 9 \end{array} \tag{5.3.43}$$

and we continue like this for the second column where we have $4+7=11$ at the difference that now we have a number >10, then we report the first left digit on the next (left) column for the addition. Therefore:

$$\begin{array}{rcccc} & 9 & 2^{+1} & 4 & 4 \\ + & 3 & 4 & 7 & 5 \\ \hline & & & 1 & 9 \end{array} \tag{5.3.44}$$

The third column we be calculated therefore as $1+2+4=7$ which give us:

$$\begin{array}{rcccc} & 9 & 2^{+1} & 4 & 4 \\ + & 3 & 4 & 7 & 5 \\ \hline & & 7 & 1 & 9 \end{array} \tag{5.3.45}$$

For the last column we have $9+3=12$ and once again we report the first digit from the left on the next column of the addition. Therefore:

$$\begin{array}{rccccc} & {}^{+1}9 & 2^{+1} & 4 & 4 \\ + & 3 & 4 & 7 & 5 \\ \hline & 2 & 7 & 1 & 9 \end{array} \tag{5.3.46}$$

Finally:

$$\begin{array}{rcccccc} & & {}^{+1}9 & 2^{+1} & 4 & 4 \\ + & & 3 & 4 & 7 & 5 \\ \hline & 1 & 2 & 7 & 1 & 9 \end{array} \tag{5.3.47}$$

This example shows how we can proceed for the addition of any real numbers: we do an addition column by column from the right to the left and if the result of one addition is greater than 10, we report the left digit on the next (left) column.

Following a reader request, here is also an illustrated example on how we teach to add simple fractions together at the middle school level, first in the case where the denominators are the same:

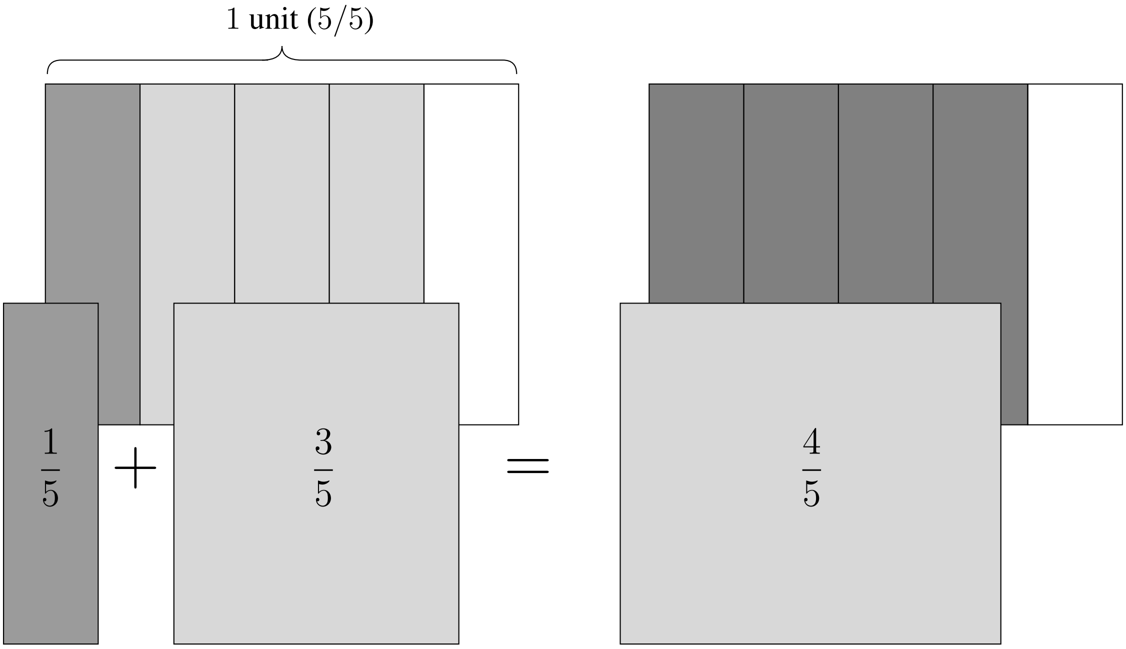

1 unit (5/5)
$\frac{1}{5}$
+
$\frac{3}{5}$
=
$\frac{4}{5}$

Figure 5.30: **Addition of fractions with common denominator**

Or if the denominators are not common, we seek common fractions equivalent to each term of the addition, in such a way that all the fractions have a common denominator and afterwards we add as we did before:

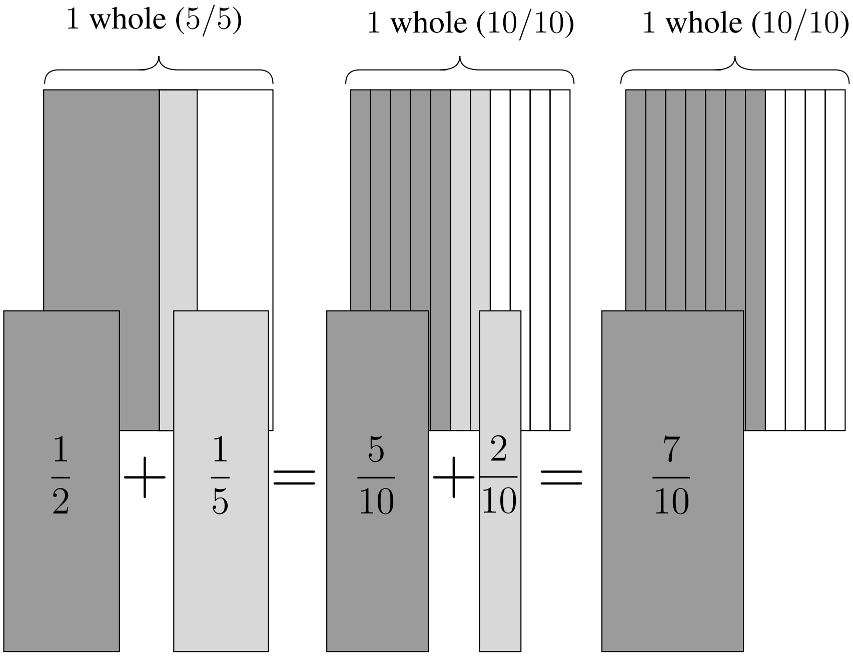

1 whole (5/5)
1 whole (10/10)
1 whole (10/10)
$\frac{1}{2}$
+
$\frac{1}{5}$
=
$\frac{5}{10}$
+
$\frac{2}{10}$
=
$\frac{7}{10}$

Figure 5.31: **Addition of fractions without common denominators**

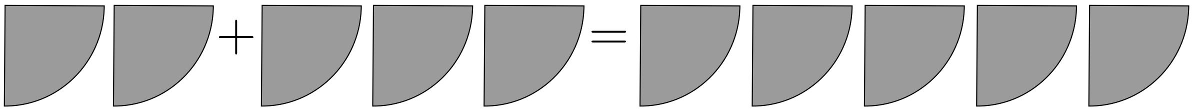
We then see above that to add a fraction from another we first put the fraction at the same denominator, we then add the numerators and we keep the denominator if there is no possible simplification:

$$\frac{1}{2}+\frac{1}{5}=\frac{5}{5}\cdot\frac{1}{2}-\frac{2}{2}\cdot\frac{1}{4}=\frac{5}{10}+\frac{2}{10}=\frac{7}{10} \tag{5.3.48}$$

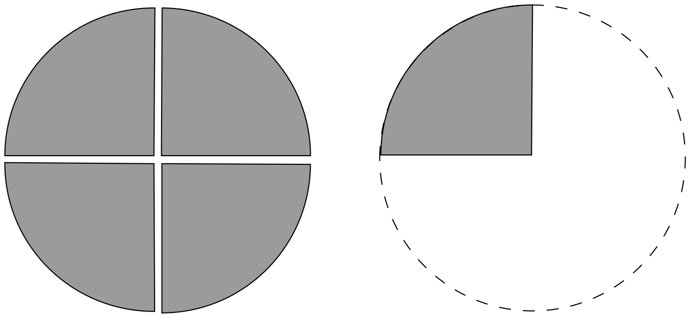
Obviously when we add fractions together we could also get a number that represents another number that is greater than 1 as for example:

$$\frac{2}{4}+\frac{3}{4}=\frac{5}{4}=1.25 \tag{5.3.49}$$

This can be illustrated as following:



If you want to assemble these quarter discs, here is what we get:



You understand then why we can say that we have $5/4$ of a disc!

This algorithm (process or methodology) of addition is quite simple to understand and to execute. We will not go further on this subject at this day!

Below a typical addition table for the first twelve natural numbers:

+	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	2	3	4	5	6	7	8	9	10	11	12	13
2	3	4	5	6	7	8	9	10	11	12	13	14
3	4	5	6	7	8	9	10	11	12	13	14	15
4	5	6	7	8	9	10	11	12	13	14	15	16
5	6	7	8	9	10	11	12	13	14	15	16	17
6	7	8	9	10	11	12	13	14	15	16	17	18
7	8	9	10	11	12	13	14	15	16	17	18	19
8	9	10	11	12	13	14	15	16	17	18	19	20
9	10	11	12	13	14	15	16	17	18	19	20	21
10	11	12	13	14	15	16	17	18	19	20	21	22
11	12	13	14	15	16	17	18	19	20	21	22	23
12	13	14	15	16	17	18	19	20	21	22	23	24

Figure 5.32: **Addition table of the first 12 natural numbers**

5.3.2.2 Subtraction

Definition (#5.3.57): The "subtraction" is a mathematical operation that represents the operation of removing objects from a collection. More formally the subtraction of the number A by the number B denoted by the symbol "$-$" consist in founding the number C such that added to B gives A.

Remark

As we saw it in the section of Set Theory (page 404) the subtraction in the set $\mathbb{N}$ could be possible only if $A > B$.

Formally we write an inline literal subtraction in the form:

$$A - B = C \tag{5.3.50}$$

That must satisfies:

$$A = B + C \tag{5.3.51}$$

Or in schematic form of a special case:

$$10 - 3 - 4 = 7 - 4 = 3$$

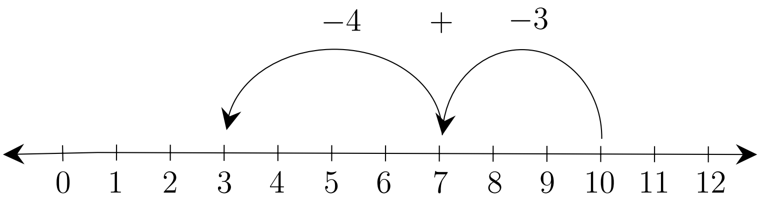


Figure 5.33: One possible schema for subtraction

Here are some intuitive properties that we assume without proof for the subtraction operation (as it can be deduce from the addition...):

P1. The subtraction of several numbers depends on the order of the terms. We say when than subtraction is a "non-commutative operation". Indeed:

$$5 - 2 \neq 2 - 5 \tag{5.3.52}$$

P2. The subtraction of several numbers change if we replace two or more of them by their intermediate result. We say when the subtraction is a "non-associative operation". Indeed:

$$5 - (3 - 2) \neq (5 - 3) - 2 \tag{5.3.53}$$

P3. The zero is not the neutral element of subtraction. Indeed, any number to which we subtract zero gives the same number, so zero is neutral on the right ... but not left because any number we subtract to zero does not give zero! We then say that the zero is only "neutral on the right" in the case of subtraction. Indeed:

$$0 - 5 \neq 0 \tag{5.3.54}$$

P4. Depending on the set in which we work, the subtraction can include a term in such a way that the total is zero. We then say that there is a "opposite" for the subtraction.

In most complicated cases we have a special vocabulary:

$$\begin{array}{rrrl} -1 & & & \longleftarrow \text{carry} \\ 7 & 0 & 4 & \longleftarrow \text{Minuend} \\ 5 & 1 & 2 & \longleftarrow \text{Subtrahend} \\ \cline{1-3} 1 & 9 & 2 & \longleftarrow \text{Rest or Difference} \end{array} \tag{5.3.55}$$

The "minuend" is 704, the "subtrahend" is 512. The minuend digits are $m_3 = 7$, $m_2 = 0$ and $m_1 = 4$. The subtrahend digits are $s_3 = 5$, $s_2 = 1$ and $s_1 = 2$. Beginning at the one's place, 4 is not less than 2 so the difference 2 is written down in the result's one place. In the ten's place, 0 is less than 1, so the 0 is increased by 10, and the difference with 1, which is 9, is written down in the ten's place. The American method corrects for the increase of ten by reducing the digit in the minuend's hundreds place by one. That is, the 7 is struck through and replaced by a 6. The subtraction then proceeds in the hundreds place, where 6 is not less than 5, so the difference is written down in the result's hundred's place. We are now done, the result is 192.

Let us see now some concrete examples of subtractions of various simple number in the purpose to practice the basis:

✍**Example:**

The subtraction of two relatively small numbers is pretty easy once we memorized to count to at least the number resulting from this operation. So:

$$\begin{array}{rr} & 5 \\ - & 2 \\ \hline & 3 \end{array} \tag{5.3.56}$$

and:

$$\begin{array}{rr} & 10 \\ - & 3 \\ \hline & 7 \end{array} \tag{5.3.57}$$

and:

$$\begin{array}{rr} & 1014 \\ - & 3 \\ \hline & 1011 \end{array} \tag{5.3.58}$$

For larger numbers another possible method must be learned by heart (as well as for the addition). For example:

$$\begin{array}{rr} & 4574 \\ - & 3785 \\ \hline & ? \end{array} \tag{5.3.59}$$

we subtract the columns (4 columns in this example) from right to left. In the first column we have $4-5=-1<0$ so we report -1 to the next column (second one) and we write $10-1=9$ below the horizontal line of the first column:

$$\begin{array}{rrrrr} & 4 & 5 & 7^{-1} & 4 \\ - & 3 & 7 & 8 & 5 \\ \hline & & & & 9 \end{array} \tag{5.3.60}$$

and we continue as well for the second column $7-8=-1<0$ so that we report -1 on the next column (third one) and as $-1-1=-2$ we report $10-2=8$ below the horizontal bar of the second column:

$$\begin{array}{rrrrr} & 4 & 5^{-1} & 7^{-1} & 4 \\ - & 3 & 7 & 8 & 5 \\ \hline & & & 8 & 9 \end{array} \tag{5.3.61}$$

The third column is calculated as $5-7=-2<0$ and we report -1 on the next column (fourth one) and as $-1-2=-3$ we report $10-3=7$ below the line of the third column bar:

$$\begin{array}{rrrrr} & 4^{-1} & 5^{-1} & 7^{-1} & 4 \\ - & 3 & 7 & 8 & 5 \\ \hline & & 7 & 8 & 9 \end{array} \tag{5.3.62}$$

In the last column we have $4-3=1>0$ therefore we report the nothing on the next column and as $1-1=0$ we report 0 below the line of the fourth column bar:

$$\begin{array}{rrrrr} & 4^{-1} & 5^{-1} & 7^{-1} & 4 \\ - & 3 & 7 & 8 & 5 \\ \hline & 0 & 7 & 8 & 9 \end{array} \tag{5.3.63}$$

That's how we therefore we proceed to subtracting any numbers. We make a subtraction by column from the right to the left and if the result is a subtraction is less than zero we report -1 to the next column and the addition of the latest report on the subtraction obtained below the line.

Following a reader request, here is also an illustrated example on how we teach to subtract simple fractions together at the middle school level:

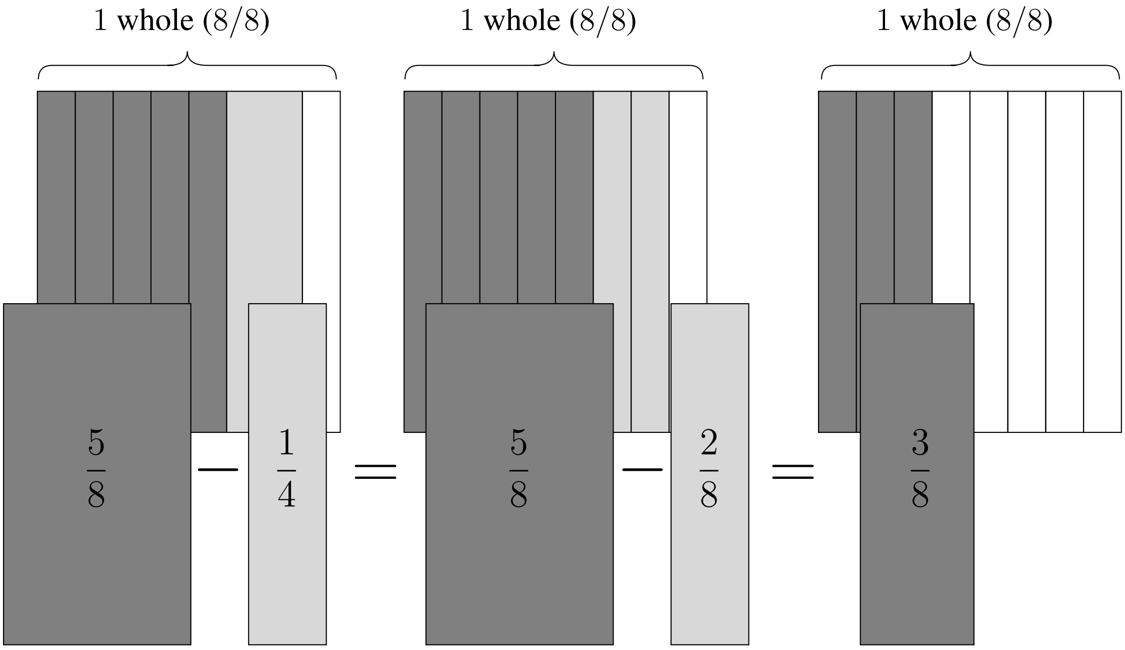

1 whole (8/8)
1 whole (8/8)
1 whole (8/8)
$\frac{5}{8}-\frac{1}{4}=\frac{5}{8}-\frac{2}{8}=\frac{3}{8}$

Figure 5.34: **Subtraction of fractions without common denominators**

We then see above that to subtract a fraction from another we first put the fraction at the same denominator, we then subtract the numerators and we keep the denominator if there is no possible simplification:

$$\frac{5}{8}-\frac{1}{4}=\frac{5}{8}-\frac{2}{2}\cdot\frac{1}{4}=\frac{5}{8}-\frac{2}{8}=\frac{3}{8} \tag{5.3.64}$$

We have when we mix the addition and subtraction the following resulting relation that should be obvious for most readers:

$$\begin{aligned} a+(b-c)&=(a+b)-c\\ a-(b+c)&=(a-b)-c\\ a-(b-c)&=(a-b)+c\\ a-b&=(a-c)-(b-c) \end{aligned} \tag{5.3.65}$$

The methodology used for subtraction being based on exactly the same rules that for addition we will expand the subject more as this seems actually useless in our point of view. This method is very simple and of course requires some habits to work with numbers to be fully understood and mastered.

Below a typical subtraction table for the first twelve natural numbers:

−	1	2	3	4	5	6	7	8	9	10	11	12
1	0	1	2	3	4	5	6	7	8	9	10	11
2	−1	0	1	2	3	4	5	6	7	8	9	10
3	−2	−1	0	1	2	3	4	5	6	7	8	9
4	−3	−2	−1	0	1	2	3	4	5	6	7	8
5	−4	−3	−2	−1	0	1	2	3	4	5	6	7
6	−5	−4	−3	−2	−1	0	1	2	3	4	5	6
7	−6	−5	−4	−3	−2	−1	0	1	2	3	4	5
8	−7	−6	−5	−4	−3	−2	−1	0	1	2	3	4
9	−8	−7	−6	−5	−4	−3	−2	−1	0	1	2	3
10	−9	−8	−7	−6	−5	−4	−3	−2	−1	0	1	2
11	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	0	1
12	−11	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	0

Figure 5.35: **Subtraction table of the first 12 natural numbers**

Remark

Addition seems to be a simple operation however in some well known book simple kindergarten level sums are wrong. Like for example in the King James Version of Bible (one of the many different existing translated version of the Bible existing around the World), Ezra 1:9-10 (old testament). The inventory of the sum leads to a wrong total ($5,469$ instead of $2,499$):

Verse Number	Description	Quantity
`I-9`	gold dishes	30
	silver dishes	1,000
	knives	29
`I-10`	basons of gold	30
	silver basons of a second sort	410
	other vessels	1,000
`I-11`	Total	5,469

Some high level reputed professional theologians tried to find a workaround to explain this error in a "godly" book but failed (see [487]). And such an error occurs multiple times in the Bible not only once!

5.3.2.3 Multiplication

Definition (#5.3.58): The "multiplication" of numbers is an operation that has for purpose, given two numbers, one named "multiplier" m, and the other "multiplicand" M, to find a third number named "product" P that is the sum (multiplication is only a successive number of sums!) as many equal numbers to the multiplicand as there are units multiplier:

$$m \times M = \underset{(1)}{M} + \underset{(2)}{M} + \underset{(3)}{M} + \ldots + \underset{(m)}{M} = \sum_{i=1}^{m} M = P \qquad (5.3.66)$$

The multiplicand and multiplier are named "product factors".

The multiplication is indicated in kindergarten by the symbol "$\times$"of the elevated dot symbol in higher classes "$\cdot$" or even when there is no possible confusion... without anything:

$$\boxed{a \times b = a \cdot b = ab} \qquad (5.3.67)$$

We can define the multiplication using the Peano axioms in the special case of natural numbers $\mathbb{N}$ as already mentioned in the section Numbers (page 218). Thus, with these axioms it is possible to prove that there is (exists) one and only one (unique) application, denoted "$\times$" or more often "$\cdot$" of $\mathbb{N}^2$ to $\mathbb{N}$ satisfying:

$$\begin{aligned} &\forall n \in \mathbb{N}: & n \cdot 0 &= 0 \\ &\forall p \in \mathbb{N}, \forall q \in \mathbb{N}: & p(q+1) &= pq + q \end{aligned} \qquad (5.3.68)$$

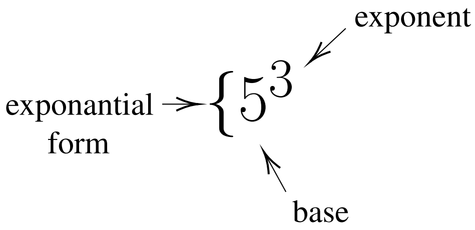
Remark

As this textbook has not be written for mathematicians, we will pass the proof (relatively long and of little interest in the case of business) and we will assume that the application "$\times$" exists and is unique ... and that it follows from the above properties.

The "power" is a specific notation of a special case of the multiplication. When to multiplicand(s) and the multiplier(s) are typically identical in numerical values, we denote therefore the multiplication by (for example):

$$n \cdot n \cdot n \cdot n \cdot n \cdot n \cdot n \cdot n = n^8 \qquad (5.3.69)$$

This is what we name the "power notation" or "exponentiations". The number in superscript is what we the name the "power" or the "exponent" of the number. The notation with exponents is said to be seen for the first time in a book of Chuquet in 11484 (holocene calendar).



In general:

$$\underbrace{n \cdot n \cdot n \cdot \ldots \cdot n \cdot n \cdot n \cdot n}_{m \text{ times}} = n^m \qquad (5.3.70)$$

You can check by yourself that its properties are the following (for example):

$$n^x n^y = n^{x+y} \tag{5.3.71}$$

and also:

$$a^x b^x = (ab)^x \tag{5.3.72}$$

Here are some obvious properties about the multiplication that we will admit without proof (this is a Set properties point of view listing):

P1. The multiplication of several numbers does not depend on the order of terms. Then we say that multiplication is a "commutative operation".

P2. The multiplication of several numbers does not change if we replace two or more of them by their intermediate result. We then say that the multiplication is an "associative operation".

P3. The unit is the neutral element of the multiplication as any multiplicand multiplied by the multiplier 1 is equal to the multiplicand itself.

P4. The multiplication may have a term such that the product is equal to unity (the neutral element). Then we say that there exists a "multiplicative inverse" (but this depends strictly speaking in what set of numbers we work as in some the concept of decimal number does not exist!).

P5. Multiplication is a "distributive operation", that is to say:

$$a \cdot (b + c) = ab + ac \tag{5.3.73}$$

the reverse being named a "factorization operation".

✍**Example:**

Probably the best-known case in the world of factorization (by users of spreadsheets softwares like Microsoft® Excel®) is the calculation of the addition or removal of a tax on a given amount! For example, consider adding 8% VAT to an amount of 100.-. This will typically be written in one of the following ways:

$$100+100\cdot8\% = 100\cdot1+100\cdot8\% = 100\cdot(1+8\%) = 100\cdot(1+8/100) = 100\cdot1.08$$

P6. Any number at the power of zero is equal to 1 as:

$$n^0 = n^m n^{-m} = \frac{n^m}{n^m} = 1 \tag{5.3.74}$$

From this result some deduce too quickly (and even some professors teach it!) that $0^0 = 1$. But in reality 0^0 is just undefined as it the ratio of two zeros.

Remark

There are two main reasons why some professors teach $0^0 = 1$. The first one is an application of the binomial theorem (see section Calculus page 1364):

$$(1+x)^n = \sum_{k=0}^{n} \begin{pmatrix} n \\ k \end{pmatrix} 1^{n-k} x^k$$

When $x = 0$, the expression becomes:

$$\begin{aligned}(1+0)^n &= \sum_{k=0}^{n} \begin{pmatrix} n \\ k \end{pmatrix} 1^{n-k} 0^k \\ 1^n &= 1 \cdot 0^0 \\ 1 &= 1 \cdot 0^0\end{aligned}$$

In this context, 0^0 has to be equal to 1.

The second go with the reasoning that 1^1 is 1, $0.5^{0.5}$ is 0.70 etc. x^x goes down from one until it hits 0.4 and then it goes back up and approaches the limit of one but never hits. I.e. $0.000000001^{0.000000001}$ is almost 1, so some people round 0^0 to 1.

To summarize it seems that there are no international convention as far as we know.

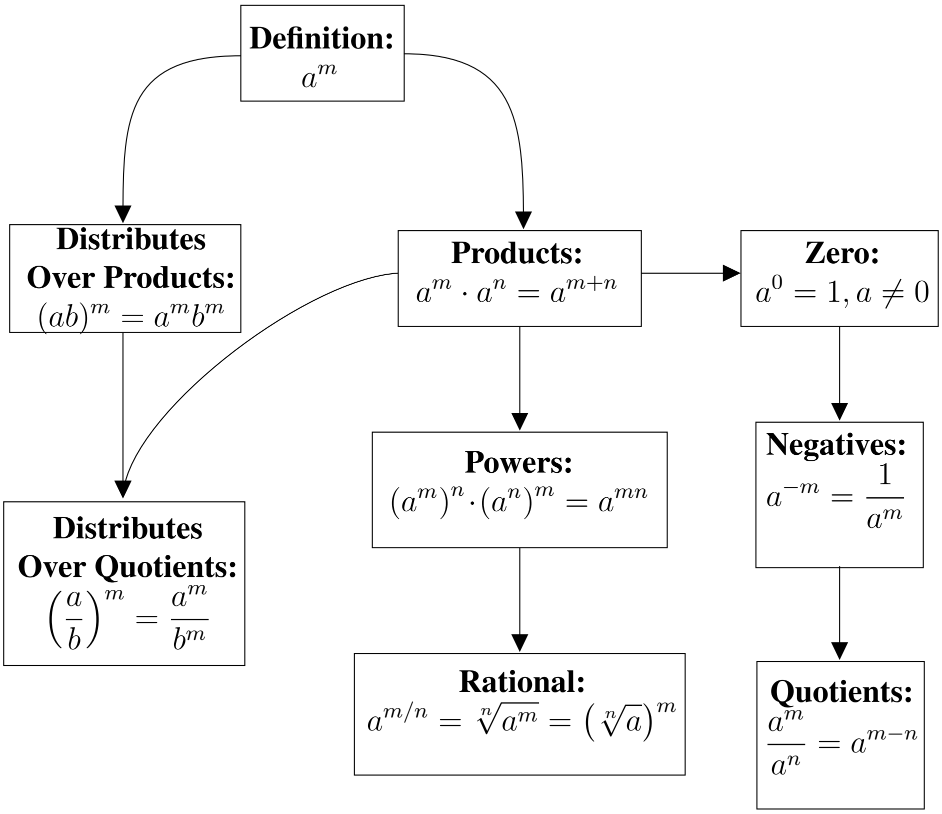

Definition: a^m
Distributes Over Products: $(ab)^m = a^m b^m$
Products: $a^m \cdot a^n = a^{m+n}$
Zero: $a^0 = 1, a \neq 0$
Distributes Over Quotients: $\left(\frac{a}{b}\right)^m = \frac{a^m}{b^m}$
Powers: $(a^m)^n \cdot (a^n)^m = a^{mn}$
Negatives: $a^{-m} = \frac{1}{a^m}$
Rational: $a^{m/n} = \sqrt[n]{a^m} = \left(\sqrt[n]{a}\right)^m$
Quotients: $\frac{a^m}{a^n} = a^{m-n}$

Figure 5.36: **Summary of exponents rules**

Let us also introduce some special notations for the multiplication:

1. Given any numbers $x_1, x_2, \ldots, x_n$ (not necessarily equal) then we can write the product as following:

$$\boxed{x_1 \cdot x_2 \cdot \ldots \cdot x_n = \prod_{i=1}^{n} x_i} \tag{5.3.75}$$

   by defining upper and lower bounds to the indexed product (above and below the uppercase Greek letter "Pi").

   We trivially have respectively to the latter notation (on request we can detail more...):

$$\prod_{i=1}^{n} kx_i = k^n \prod_{i=1}^{n} x_i \tag{5.3.76}$$

   for any number k such that:

$$\prod_{i=1}^{n} k = k^n \tag{5.3.77}$$

   We also have for example:

$$\prod_{i=1}^{n} (x+y) = (x+y)^n \tag{5.3.78}$$

2. We define the "factorial" simply ("simply"... because it exists also a more complex way of defining it through the Euler Gamma function as it is done in the section of Integral and Differential Calculus page 1574) by:

$$\boxed{1 \times 2 \times 3 \times 4 \times \cdots \times n = n!} \tag{5.3.79}$$

   with the special fact that (only the complex definition mentioned before can make this fact obvious... so for the proof see again page 1577):

$$0! = 1 \tag{5.3.80}$$

Let us see some simple examples of basic multiplications:

✍**Examples:**

E1. The multiplication of two relatively small numbers is fairly easy once we have memorized count to at least the number resulting from this operation. So:

$$\begin{array}{rr} & 5 \\ \times & 2 \\ \hline & 10 \end{array} \qquad \begin{array}{rr} & 10 \\ \times & 3 \\ \hline & 30 \end{array} \qquad \begin{array}{rr} & 1014 \\ \times & 3 \\ \hline & 3042 \end{array} \tag{5.3.81}$$

E2. For much larger numbers we must adopt another method that has to be memorized.

For example:

$$\begin{array}{r} 4\,5\,7\,4 \\ \times \qquad 8 \\ \hline 3\,2 \\ 5\,6 \\ 4\,0 \\ 3\,2 \end{array} \tag{5.3.82}$$

This methodology is very logical if you understand how we build a number in base ten. Thus we have (we'll assume that the distributive property is mastered):

$$\begin{aligned} 8\times 4574 &= 8\times(4\cdot 10^3+5\cdot 10^2+7\cdot 10^1+4\cdot 10^0)\\ &= 8\times 4000+8\times 500+8\times 70+8\times 4\\ &= 32\cdot 10^3+40\cdot 10^2+56\cdot 10^1+32\cdot 10^0\\ &= 36592 \end{aligned} \tag{5.3.83}$$

To avoid overloading the notations in the multiplication by the "vertical" method, we do not represent the zeros that would overload unnecessarily the calculations (and even more if the multiplier and / or the multiplicand are very large numbers).

Following a reader request, here is also an illustrated example on how we teach (on possible way among others) to multiply simple fractions together at the middle school level:

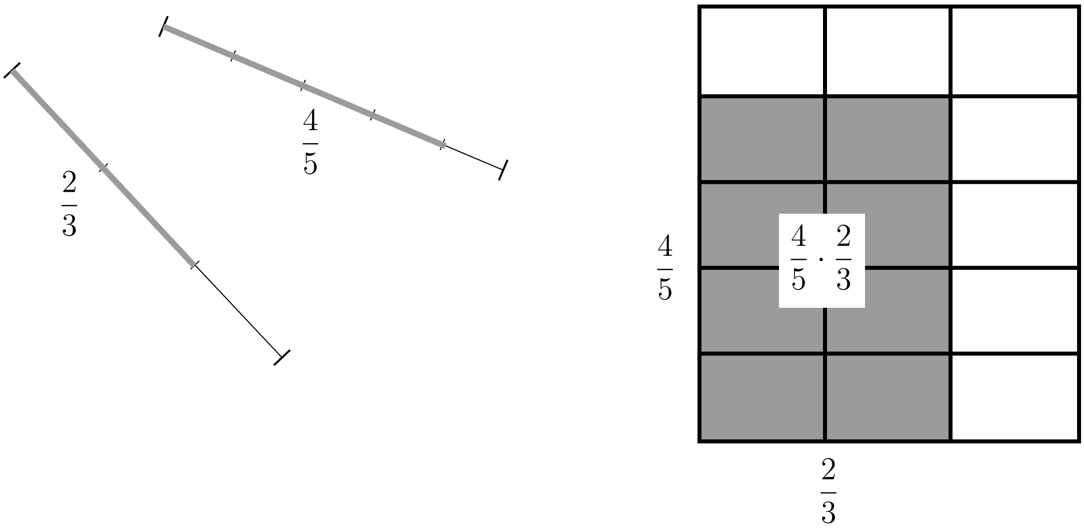


Figure 5.37: **Multiplication of fractions without common denominators**

We have therefore:

$$\frac{2}{3}\cdot\frac{4}{5}=\frac{8}{15} \tag{5.3.84}$$

This example illustrates that to multiply two fractions we simply multiply the numerators together and the denominators together.

Below a typical multiplication table for the first twelve natural numbers:

×	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	1	2	3	4	5	6	7	8	9	10	11	12
2	2	4	6	8	10	12	14	16	18	20	22	24
3	3	6	9	12	15	18	21	24	27	30	33	36
4	4	8	12	16	20	24	28	32	36	40	44	48
5	5	10	15	20	25	30	35	40	45	50	55	60
6	6	12	18	24	30	36	42	48	54	60	66	72
7	7	14	21	28	35	42	49	56	63	70	77	84
8	8	16	24	32	40	48	56	64	72	80	88	96
9	9	18	27	36	45	54	63	72	81	90	99	108
10	10	20	30	40	50	60	70	80	90	100	110	120
11	11	22	33	44	55	66	77	88	99	110	121	132
12	12	24	36	48	60	72	84	96	108	120	132	144

Figure 5.38: **Multiplication table of the first 12 natural numbers**

Caution!! Multiplication can't be applied universally in life or science, especially when dimensions are involved and that is related to what we name "dimensional analysis" that has for purpose to check the consistency of mathematical operations applied to real-life cases. Multiplication (and derived operations) is valid only when the resulting quantity makes physical sense therefore multiplying 0 cow by 5 cows doesn't make sense and is a calculation that must not be done as the dimensional result are cows squared, i.e. $[\text{cow}^2]$. Dimensional analysis is like a grammar for the language of physics. Just as you wouldn't mix verbs and nouns randomly in a sentence, you can't mix units and dimensions arbitrarily in equations. Multiplication is powerful—but only when it respects the structure of the physical world.

5.3.2.4 Division

Definition (#5.3.59): The "division" of integers (to start with the simplest case ...) is an operation, which aims, given two integers, one named "dividend" D, the other named "divider" d , to find a third number named "quotient" Q which is the largest number whose product by the divisor can be subtracted (so the division result of the subtraction!) the dividend (the difference being named the "rest" R or sometimes the "congruence").

Remark

In the case of real numbers there are never any rest at the end of the division operation (because the quotient multiplied by the divisor gives always exactly the dividend)!

Generally in the context of integers (or algebraic equation division), if we denote by D the dividend and by d the divisor, the quotient Q and the remainder R we have the relation:

$$\boxed{D = Q \cdot d + R} \tag{5.3.85}$$

knowing that the division was initially written as follows:

$$D : d = \frac{D}{d} \tag{5.3.86}$$

We indicate the operation of division by placing between the two numbers, the dividend and the divider, a symbol ":" or a slash "/" or even in kindergarten with the symbol $\div$.

We refer also often by the term "fraction" (instead of "quotient"), the ratio of two numbers or in other words, the division of the first by the second.

Remark

The sign of division ":" is said to be due to Gottfried Wilhelm Leibniz. The slash symbol could have been see for the first time in the works of Leonardo Fibonacci (11202) and is probably due to the Hindus.

If we divide two numbers and we want an integer as quotient and as remainder (if there is one...), then we speak of "Euclidiean division".

For example, dividing a cake, is not a Euclidean division because the quotient is not an integer, except if one takes the four quarters...:

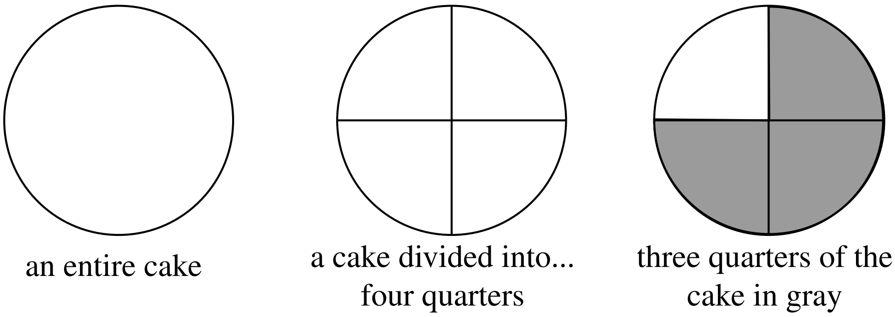
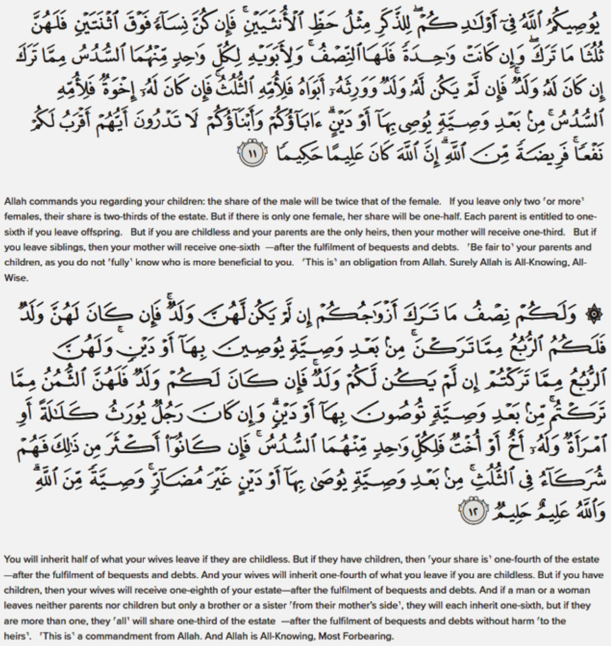

an entire cake
a cake divided into... four quarters
three quarters of the cake in gray

Figure 5.39: Schematic example of a division (fractions)

Remark

Fractions can seem easy to handle. However this is not the easy for everyone. Like in the Quran for example they tried to do a general rule of inheritance that fail in some cases and that's quite a shame for a "godly" book with rules provided by an omniscient entity... Let's see that with the Surah An-Nisa 4:11-12 from the Warsh Quran (one of the many arabic different existing Qurans around the World):



يُوصِيكُمُ ٱللَّهُ فِيٓ أَوۡلَٰدِكُمۡۖ لِلذَّكَرِ مِثۡلُ حَظِّ ٱلۡأُنثَيَيۡنِۚ فَإِن كُنَّ نِسَآءٗ فَوۡقَ ٱثۡنَتَيۡنِ فَلَهُنَّ
ثُلُثَا مَا تَرَكَۖ وَإِن كَانَتۡ وَٰحِدَةٗ فَلَهَا ٱلنِّصۡفُۚ وَلِأَبَوَيۡهِ لِكُلِّ وَٰحِدٖ مِّنۡهُمَا ٱلسُّدُسُ مِمَّا تَرَكَ
إِن كَانَ لَهُۥ وَلَدٞۚ فَإِن لَّمۡ يَكُن لَّهُۥ وَلَدٞ وَوَرِثَهُۥٓ أَبَوَاهُ فَلِأُمِّهِ ٱلثُّلُثُۚ فَإِن كَانَ لَهُۥٓ إِخۡوَةٞ فَلِأُمِّهِ
ٱلسُّدُسُۚ مِنۢ بَعۡدِ وَصِيَّةٖ يُوصِي بِهَآ أَوۡ دَيۡنٍۗ ءَابَآؤُكُمۡ وَأَبۡنَآؤُكُمۡ لَا تَدۡرُونَ أَيُّهُمۡ أَقۡرَبُ لَكُمۡ
نَفۡعٗاۚ فَرِيضَةٗ مِّنَ ٱللَّهِۗ إِنَّ ٱللَّهَ كَانَ عَلِيمًا حَكِيمٗا ١١

Allah commands you regarding your children: the share of the male will be twice that of the female. If you leave only two 'or more' females, their share is two-thirds of the estate. But if there is only one female, her share will be one-half. Each parent is entitled to one-sixth if you leave offspring. But if you are childless and your parents are the only heirs, then your mother will receive one-third. But if you leave siblings, then your mother will receive one-sixth —after the fulfilment of bequests and debts. 'Be fair to' your parents and children, as you do not 'fully' know who is more beneficial to you. 'This is' an obligation from Allah. Surely Allah is All-Knowing, All-Wise.

۞ وَلَكُمۡ نِصۡفُ مَا تَرَكَ أَزۡوَٰجُكُمۡ إِن لَّمۡ يَكُن لَّهُنَّ وَلَدٞۚ فَإِن كَانَ لَهُنَّ وَلَدٞ
فَلَكُمُ ٱلرُّبُعُ مِمَّا تَرَكۡنَۚ مِنۢ بَعۡدِ وَصِيَّةٖ يُوصِينَ بِهَآ أَوۡ دَيۡنٖۚ وَلَهُنَّ
ٱلرُّبُعُ مِمَّا تَرَكۡتُمۡ إِن لَّمۡ يَكُن لَّكُمۡ وَلَدٞۚ فَإِن كَانَ لَكُمۡ وَلَدٞ فَلَهُنَّ ٱلثُّمُنُ مِمَّا
تَرَكۡتُمۚ مِّنۢ بَعۡدِ وَصِيَّةٖ تُوصُونَ بِهَآ أَوۡ دَيۡنٖۗ وَإِن كَانَ رَجُلٞ يُورَثُ كَلَٰلَةً أَوِ
ٱمۡرَأَةٞ وَلَهُۥٓ أَخٌ أَوۡ أُخۡتٞ فَلِكُلِّ وَٰحِدٖ مِّنۡهُمَا ٱلسُّدُسُۚ فَإِن كَانُوٓاْ أَكۡثَرَ مِن ذَٰلِكَ فَهُمۡ
شُرَكَآءُ فِي ٱلثُّلُثِۚ مِنۢ بَعۡدِ وَصِيَّةٖ يُوصَىٰ بِهَآ أَوۡ دَيۡنٍ غَيۡرَ مُضَآرّٖۚ وَصِيَّةٗ مِّنَ ٱللَّهِۗ
وَٱللَّهُ عَلِيمٌ حَلِيمٞ ١٢

You will inherit half of what your wives leave if they are childless. But if they have children, then 'your share is' one-fourth of the estate —after the fulfilment of bequests and debts. And your wives will inherit one-fourth of what you leave if you are childless. But if you have children, then your wives will receive one-eighth of your estate—after the fulfilment of bequests and debts. And if a man or a woman leaves neither parents nor children but only a brother or a sister 'from their mother's side', they will each inherit one-sixth, but if they are more than one, they 'all' will share one-third of the estate —after the fulfilment of bequests and debts without harm 'to the heirs'. 'This is' a commandment from Allah. And Allah is All-Knowing, Most Forbearing.

Let's imagine the particular but simple case that we are a man, who has 1 wife, 2 daughters, a father and a mother. We have no sisters or brothers. Unfortunately consider that we died in a car accident and left $1,000$ USD as an inheritance. So let's apply the above rules:

1. According to verse `4:12`, our wife will receive $1/8$ of our $1,000$ USD. Hence: $(1/8)\cdot 1,000 = 125$ USD.
2. According to verse `4:11`, our 2 daughters will receive $2/3$. Hence: $(2/3)\cdot 1000 = 666.\bar{6}$ USD.
3. According to verse `4:11`, our father will receive $1/6$. Hence: $(1/6)\cdot 1000 = 166.\bar{6}$ USD.
4. According to verse `4:11`, our mother will receive $1/6$. Hence: $(1/6)\cdot 1000 = 166.\bar{6}$ USD.

Hence a total of:

$$\left(\frac{1}{8}+\frac{2}{3}+\frac{1}{6}+\frac{1}{6}\right)\cdot 1000 = 125 + 666.\bar{6} + 166.\bar{6} + 166.\bar{6} = \frac{27}{24}\cdot 1000 = 1125$$

So a fraction of $27/24 > 1$ of the initial amount (money was created from nothing...). Huge failure... !

If we have:

$$D : d = D\cdot\frac{1}{d} = D\cdot i_D \qquad (5.3.87)$$

we denote by i_D the inverse of the dividend. At any number is associated an inverse that satisfies this condition. From this definition it comes the notation (with x being any number other than zero)

$$\frac{x}{x}=x^1\cdot x^{-1}=x^{1-1}=x^0=1 \tag{5.3.88}$$

In the case of two fractional numbers, we say they are "inverse" or "reciprocal", when their product is equal to unity (as the previous relations).

Remarks

R1. A division by zero is what we name a "singularity". That is to say the result of the division is: undetermined!!

R2. When we multiply the dividend and the divisor of a division (fraction) by a same number, the quotient does not change (this is an: "equivalent fraction"), but the remainder is multiplied by that number.

R3. Divide a number by a product made of several factors is equivalent to divide this number successively by each of the factors of the product and vice versa.

R4. Fractions that are greater than 0 but less than 1 are named "proper fractions". In proper fractions, the numerator is less than the denominator. When a fraction has a numerator that is greater than or equal to the denominator, the fraction is an "improper fraction". An improper fraction is always 1 or greater than 1. And, finally, a mixed number is a combination of a whole number and a proper fraction.

The properties of the divisions with the condensed power notations (exponentiation) are typically as example (we will leave to the reader the fact to check this up to with numerical values):

$$\frac{x\cdot x\cdot x}{y\cdot y}=x^3\cdot y^{-2} \tag{5.3.89}$$

or obviously another example:

$$\frac{x\cdot x\cdot x}{x\cdot x}=\frac{x}{x}\cdot\frac{x}{x}\cdot x=1\cdot 1\cdot x=x^3\cdot x^{-2}=x^1=x \tag{5.3.90}$$

We therefore deduce that:

$$\frac{x^p}{x^q}=x^{p-q} \tag{5.3.91}$$

Let us recall that a prime number (relative integer $\mathbb{Z}$) is a number greater than 1 that has for divisors only itself and unity (remember that 2 is prime for example). Therefore any number that is not prime has at least one prime number as a divisor (except 1 by definition!). The smallest divisors of an integer is a prime number (we will detail the properties of prime numbers relatively to the operation of division in the section Numbers Theory page 220).

Let us see some properties of the division (some of us are already known because they arise from logical reasoning of the multiplication properties):

$$\begin{aligned}
&\frac{a}{b}=\frac{c}{d}\Leftrightarrow a\cdot d=b\cdot c && \frac{a\cdot d}{b\cdot d}=\frac{a}{b}\cdot\frac{d}{d}=\frac{a}{b}\cdot 1=\frac{a}{b}\\
&\frac{a}{-b}=\frac{-a}{b}=-\frac{a}{b} && \frac{a}{b}+\frac{c}{b}=\frac{a+c}{b}\\
&\frac{a}{b}+\frac{c}{d}=\frac{a\cdot d+b\cdot c}{b\cdot d} && \frac{a}{b}\cdot\frac{c}{d}=\frac{a\cdot c}{b\cdot d}\\
&\frac{a}{b}\div\frac{c}{d}=\frac{a}{b}\cdot\frac{d}{c}=\frac{a\cdot d}{b\cdot c} && \frac{a}{b}b=a
\end{aligned}\tag{5.3.92}$$

where:

$$\frac{a}{b}=\frac{c}{d}\Leftrightarrow ad=bc\tag{5.3.93}$$

is what we name a "terms amplification" and:

$$\frac{a}{b}+\frac{c}{b}=\frac{a+c}{b}\tag{5.3.94}$$

is an operation consisting by putting everything with a "common denominator".

We also have the following properties:

P1. The division of several numbers depend on the order of terms. We then say that the division is a "non-commutative operation". This means we have when a that is different from b and that both are different from zero:

$$\frac{a}{b}\neq\frac{b}{a}\tag{5.3.95}$$

P2. The result of the division of several numbers change if we replace two or more of them by their intermediate result. We then say that the division is a "non-associative operation":

$$\frac{\left(\dfrac{a}{b}\right)}{c}\neq\frac{a}{\left(\dfrac{b}{c}\right)}\tag{5.3.96}$$

P3. The unit is the neutral element has that we multiply the dividend or the divider by 1 the result of the division remains the same:

$$\frac{a}{b}=\frac{1\cdot a}{b}=\frac{a}{1\cdot b}=1\cdot\frac{a}{b}\tag{5.3.97}$$

P4. The division may include a divider in such a way that the division is equal to unity (neutral element 1). We then say that there exist a "symmetrical to the division" that is obviously equal to the numerator (dividend) itself.

P5. The incrementation of numerator and denominator by a constant value is not equal to the initial ratio in the general case where $a\neq b$:

$$\frac{a}{b}\neq\frac{a+c^{te}}{b+c^{te}}\tag{5.3.98}$$

Now that we know the multiplication (and therefore power notation) and division, if we consider a and b are two positive real numbers, different from zero we have:

$$a^pa^q=\underbrace{a\cdot a\ldots\cdot a}_{p\text{-times}}\cdot\underbrace{a\cdot a\ldots\cdot a}_{q\text{-times}}=a^{p+q} \tag{5.3.99}$$

and (named sometimes the "zero exponent rule of exponents"):

$$\frac{a^p}{a^q}=\frac{\underbrace{a\cdot a\ldots\cdot a}_{p\text{-times}}}{\underbrace{a\cdot a\ldots\cdot a}_{q\text{-times}}}=\underbrace{a\cdot a\ldots\cdot a}_{p-q\text{-times}}=a^{p-q}\Rightarrow a^{-q}=\frac{1}{a^q} \tag{5.3.100}$$

and:

$$a^p\cdot b^p=(a\cdot b)^p\Leftrightarrow\frac{a^p}{b^p}=\left(\frac{a}{b}\right)^p \tag{5.3.101}$$

We have also obviously:

$$\frac{a^n}{a^n}=\frac{\underbrace{a\cdot a\ldots\cdot a}_{n\text{-times}}}{\underbrace{a\cdot a\ldots\cdot a}_{n\text{-times}}}=\underbrace{a\cdot a\ldots\cdot a}_{n-n\text{-times}}=a^0=1 \tag{5.3.102}$$

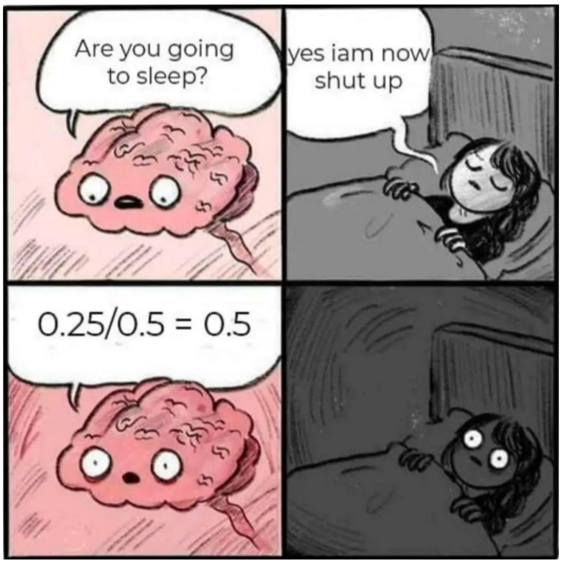
Therefore (not directly related to the subject of division... but shouldn't we introduce this now...?):

$$(a^n)^m=\underbrace{\underbrace{a\cdot a\ldots\cdot a}_{n\text{-times}}\cdot\ldots\underbrace{a\cdot a\ldots\cdot a}_{n\text{-times}}}_{m\text{ -times}}=\underbrace{a\cdot a\ldots\cdot a}_{mn\text{-times}}=a^{mn} \tag{5.3.103}$$

Remark

When dealing with division, be extremely careful abound dividing stuff by zero. This may lead to some absurd results, like the non-less famous one shared a lot on social networks:

$$\begin{aligned}a&=b\\a^2&=ab\\a^2-b^2&=ab-b^2\\(a+b)(a-b)&=b(a-b)\\a+b&=b\\2b&=b\\2&=1\end{aligned}$$


Are you going to sleep?
yes iam now shut up
0.25/0.5 = 0.5

5.3.2.5 n-root

Now that we have introduce in a simple and not too much formal way the operations of multiplication (and power notation) and division we can introduction the concept of n-root.

As we know for example that:

$$2^3 2^2 = 2^{3+2}$$

we can by reverse inference for example also write:

$$2^1 = 2^{0.5+0.5} = 2^{0.5} 2^{0.5} = 2^{1/2} 2^{1/2} \tag{5.3.104}$$

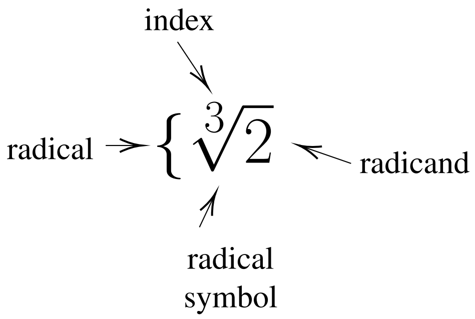
and therefore it means that fractional power exist! This is what we name n-root (in the above example we speak of 2-root).

We can now define the principal n-root of any number!

Definition (#5.3.60): In mathematics, the n-th root of a number a, where n is a positive integer, is a number r which, when raised to the power n yields x. That is to say such that: $r^n = x$, where n is the degree of the root. By convention we write:

$$\boxed{r = x^{1/n} := \sqrt[n]{x}} \tag{5.3.105}$$

Roots are usually written using the "radical" symbol $\sqrt[n]{\ldots}$ or also named the "radix". The number $n \in \mathbb{N}^*$ is named the "radicand" and sometimes the "index".



From what has been said for the powers, we can easily conclude that the n-th root of a product of several factors is the product of n-th roots of each factor:

$$\sqrt[n]{a} \cdot \sqrt[n]{b} = \sqrt[n]{a \cdot b} \Leftrightarrow \frac{\sqrt[n]{a}}{\sqrt[n]{b}} = \sqrt[n]{\frac{a}{b}} \tag{5.3.106}$$

as (seen previously):

$$(a^p)^q = a^{p \cdot q} = (a^q)^p \tag{5.3.107}$$

And therefore:

$$a^{\frac{p}{q}} = \sqrt[q]{a^p} = (\sqrt[q]{a})^p \quad \text{and} \quad \sqrt[p]{\sqrt[q]{a}} = \sqrt[pq]{a} \tag{5.3.108}$$

Obviously it comes:

$$\underbrace{\sqrt[n]{a} \cdot \sqrt[n]{a} \cdot \ldots \sqrt[n]{a}}_{n\text{-times}} = n \tag{5.3.109}$$

We also have if $a < 0$:

$$\sqrt[n]{a^n} = a \tag{5.3.110}$$

if $n \in \mathbb{N}^*$ is odd and:

$$\sqrt[n]{a^n} = |a| \tag{5.3.111}$$

if $n \in \mathbb{N}^*$ is even.

If $x < 0$ and $n \in \mathbb{N}^*$ is odd then:

$$y = a^{1/n} = \sqrt[n]{x} \tag{5.3.112}$$

is the number y such that:

$$b^n = x \tag{5.3.113}$$

If $n \in \mathbb{N}^*$ is even then obviously, as we already have seen it earlier, the root belong to $\mathbb{C}$ (see section Numbers page 243).

If the denominator of a fraction contains a factor of the form $\sqrt[n]{a^k}$ with $a \neq 0$, by multiplying the numerator and denominator by $\sqrt[n]{a^{n-}}$, we will remove the root of the denominator, since:

$$\frac{x}{\sqrt[n]{a^k}} \frac{\sqrt[n]{a^{n-k}}}{\sqrt[n]{a^{n-k}}} = \frac{x\sqrt[n]{a^{n-k}}}{\sqrt[n]{a^k}\sqrt[n]{a^{n-k}}} = \frac{x\sqrt[n]{a^{n-k}}}{\sqrt[n]{a^k a^{n-k}}} = \frac{x\sqrt[n]{a^{n-k}}}{\sqrt[n]{a^{k+n-k}}} = \frac{x\sqrt[n]{a^{n-k}}}{\sqrt[n]{a^n}} = \frac{x\sqrt[n]{a^{n-k}}}{|a|} \tag{5.3.114}$$

✍**Example:**

Let us see a world famous example of the application of the root about the origin of the ISO paper formats: A6, A5, A4, A3, A2, A1, A0, etc.

These format of papers have in fact the property (there is a goal at the origin!) to keep the proportions when we bend or cut the sheet in half in its largest dimension. Thus, if we denote by L the length and W the width of the sheet, we have:

$$\frac{L}{W}=\frac{W}{\frac{L}{2}}\Rightarrow L^2=2l^2 \tag{5.3.115}$$

Hence we have:

$$L=\sqrt{2}W \tag{5.3.116}$$

As the A0 format by definition has an area of 1 $[\mathrm{m}^2]$. For this format we have then:

$$LW=1\;[\mathrm{m}^2] \tag{5.3.117}$$

Therefore we deduce that:

$$LW=\sqrt{2}W^2=\frac{L^2}{\sqrt{2}}=1\;[\mathrm{m}^2] \tag{5.3.118}$$

and therefore:

$$W=2^{-1/4}\cdot 1\;[m]\cong 84.1\;[\mathrm{cm}] \tag{5.3.119}$$

from whence we derive:

$$L=2^{-1/4}\cdot 1\;[m]\cong 118.9\;[\mathrm{cm}] \tag{5.3.120}$$

Obviously the reader at this point may ask himself: *What types of algorithms do computers/calculators typically use to compute fractional exponents?*

The basic idea is to use logarithms (see section Analysis page 2284). If $y=x^p$, then:

$$\log(y)=p\log(x) \tag{5.3.121}$$

so:

$$y=e^{p\log(x)} \tag{5.3.122}$$

The actual value of $\log(x)$ and $e^{p\log(x)}$ are then calculated using like a Taylor polynomial (see section Sequences and Series page 1723).

5.3.3 Arithmetic Polynomials

Definition (#5.3.61): An "arithmetic polynomial" (not to be confused with "algebraic polynomial" that will be studied later in the section Calculus page 1368) is a set of numbers separated from each other by the operators of addition or subtraction ($+$ or $-$) including therefore the multiplication...

The components enclosed in the polynomial are known as "terms" of the polynomial. When the polynomial contains a single term, then we speak of "monomial", if there are two terms we speak of "binomial", and so on...

Theorem 5.11. The value of an arithmetic polynomial is equal to the excess of the sum of the terms preceded by the $+$ sign on the sum of the terms preceded by the sign $-$.

Proof 5.11.1.

$$\begin{aligned} n_1 - n_2 + n_3 - n_4 + n_5 - n_6 + \ldots - n_{i-1} + n_i = \\ (n_1 + n_3 + n_5 + \ldots + n_i) \\ + (-1)(n_2 + n_4 + n_6 + \ldots + n_{i-1}) \end{aligned} \tag{5.3.123}$$

whatever the values of the terms.

■ Q.E.D.

Highlight the negative unit -1 is what we name, as we already know, a "factorization". The reverse operation is named as we also already know a "distribution" or "development".

The product of several polynomials can always be replaced by a single polynomial that we name the... "resulting product". We usually operate as follows: we multiply successively all the terms of the first polynomial, starting from the left, with the first, the second, ..., the last by the second polynomial. We obtain a first partial product. We do, if necessary, a reduction (simplification) of similar terms. We then multiply each of the terms of the partial product successively by the first, the second, ..., the last term of the third polynomial starting from the left and so on.

✍**Example:**

$$\begin{aligned}
P_1 \cdot P_2 \cdot P_3 &= (a+b+c)(d+e+f)(g+h+i)\\
&= a(d+e+f)(g+h+f) + b(d+e+f)(g+h+i)\\
&+ c(d+e+f)(g+h+i)\\
&= ad(g+h+i) + ae(g+h+i) + af(g+h+i) + bd(g+h+i)\\
&+ be(g+h+i) + bf(g+h+i) + cd(g+h+i) + ce(g+h+i)\\
&+ cf(g+h+i)\\
&= adg + adh + adi + aeg + aeh + aei + afg + afh + afi + bdg + bdh\\
&+ bdi + beg + beh + bei + bfg + bfh + bfi + cdg + cdh + cdi + ceg + ceh\\
&+ cei + cfg + cfh + cfi
\end{aligned} \tag{5.3.124}$$

The product of the polynomials $P_1, P_2, P_3, \ldots, P_k, \ldots$ is the sum of all products of n_i factors formed with a term of P_i, of a term of P_2, ..., and a term of P_k and so. if there is no reduction, the number of terms is equal to the product of the numbers of terms of each polynomial such that the final number of therms is equals to:

$$n = \prod_{i=1}^{n} n_i \tag{5.3.125}$$

5.3.4 Absolute Value

Definition (#5.3.62): In mathematics, the "absolute value" $|x|$ of a real number x is the non-negative value of x without regard to its sign. Namely, $|x| = x$ for a positive x, $|x| = -x$ for a negative x (in which case $-x$ is positive), and $|0| = 0$. For example, the absolute value of 3 is 3, and the absolute value of -3 is also 3. The absolute value of a number may be thought of as its distance from zero.

Remarks

R1. The term absolute value has been used in this sense from at least 11806 (holocene calendar). The notation $|x|$, with a vertical bar on each side, was introduced by Karl Weierstrass in 11841 (holocene calendar).

R2. For plots about the absolute value, the reader is referred to the Functional Analysis section of this textbook (page 2216).

For any real number x, the "absolute value" x, is formally given by:

$$|x| = \begin{cases} +x & \text{if } x > 0 \\ -x & \text{if } x < 0 \\ 0 & \text{if } x = 0 \end{cases} \tag{5.3.126}$$

At the origin the absolute value was defined as:

$$|x|=\sqrt{x^2} \tag{5.3.127}$$

We notice that also the following possible notation:

$$|x|=\max(-x,x) \tag{5.3.128}$$

And the equivalent expressions:

$$\begin{aligned} x &\leqslant |x| \\ |-x| &= |x| \end{aligned} \tag{5.3.129}$$

and also:

$$\begin{aligned} |x|\leqslant y &\Leftrightarrow -y\leqslant x\leqslant y \\ |x|\geqslant y &\Leftrightarrow x\leqslant -y \vee x\geqslant y \end{aligned} \tag{5.3.130}$$

the latter being often used in the context of solving inequalities.

✍**Example:**

Solving an inequality such that:

$$|x-3|\leq 9 \tag{5.3.131}$$

is then solved simply by using the intuitive concept of distance. The solution is the set of real numbers whose distance from the real number 3 is less than or equal to 9. This is the range of center 3 and radius 9 or formally:

$$[3-9,3+9]=[-6,12] \tag{5.3.132}$$

Let us indicate that it is also useful to interpret the term:

$$\boxed{|x-y|=\sqrt{(x-y)^2}} \tag{5.3.133}$$

as the (euclidean!) distance between the two numbers x and y on the real line. Thus, by providing the set of real numbers of the absolute value distance , it becomes a metric space (see the section of Topology page 2450 to have a robust introduction to what is a distance)!!!

The absolute value has some trivial properties that we will give without proof (excepted on reader request) as they seem to us quite intuitive:

P1. Non-negativity:

$$|x|\geq 0 \tag{5.3.134}$$

P2. Positive-definiteness:

$$|x|=0 \iff x=0 \tag{5.3.135}$$

P3. Multiplicativeness:

$$|xy| = |x||y| \tag{5.3.136}$$

The proof is straightforward:

$$|ab| = \sqrt{(ab)^2} = \sqrt{a^2b^2} = \sqrt{a^2}\sqrt{b^2} = |a||b| \tag{5.3.137}$$

P4. Subadditivity ("first" triangle inequality):

$$|x+y| \leq |x| + |y| \tag{5.3.138}$$

This can be proven in one shot:

$$\begin{gathered}
|a|^2 + |b|^2 + 2|a||b| \geq a^2 + b^2 + 2ab \\
(|a| + |b|)^2 \geq (a+b)^2 \quad \left(\because \forall x \in \mathbb{R}; x^2 = |x|^2\right) \\
||a| + |b|| \geq |a+b| \\
\therefore |a| + |b| \geq |a+b|
\end{gathered} \tag{5.3.139}$$

Other important properties of the absolute value include:

P5. Idempotence (the absolute value of the absolute value is the absolute value):

$$|(|x|)| = |x| \tag{5.3.140}$$

P6. Evenness (reflection symmetry of the graph):

$$|-x| = |x| \tag{5.3.141}$$

The proof is straightforward:

$$|-a| = \sqrt{(-a)^2} = \sqrt{a^2} = |a| \tag{5.3.142}$$

P7. Preservation of division (equivalent to multiplicativeness) if $y \neq 0$:

$$\left|\frac{x}{y}\right| = \frac{|x|}{|y|} \tag{5.3.143}$$

The proof is straightforward:

$$|a \cdot \frac{1}{b}| = \sqrt{\left(a\frac{1}{b}\right)^2} = \sqrt{a^2\left(\frac{1}{b}\right)^2} = \sqrt{a^2}\sqrt{\left(\frac{1}{b}\right)^2} = \frac{|a|}{|b|} \tag{5.3.144}$$

P8. Reverse ("second") triangle inequality (equivalent to subadditivity):

$$|x-y| \geq ||x| - |y|| \tag{5.3.145}$$

For all $x, y \in \mathbb{R}$, the triangle inequality gives:

$$\begin{gathered}
|x| = |x - y + y| \leq |x - y| + |y|, \\
|x| - |y| \leq |x - y|
\end{gathered} \tag{5.3.146}$$

Interchaning $x \leftrightarrow y$ gives:

$$|y|-|x| \leq |y-x| \tag{5.3.147}$$

which when rearranged gives:

$$-(|x|-|y|) \leq |x-y| \tag{5.3.148}$$

Now combining:

$$|x|-|y| \leq |x-y| \quad \text{with} \quad -(|x|-|y|) \leq |x-y| \tag{5.3.149}$$

gives:

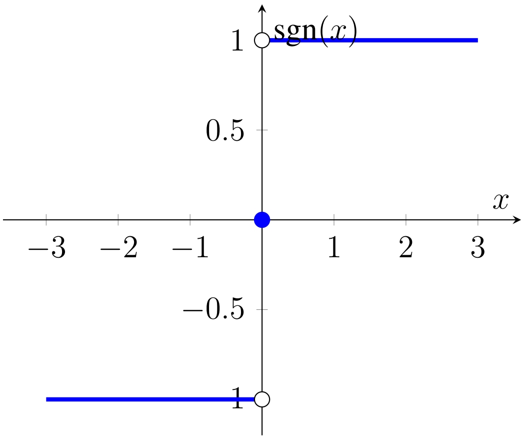
$$-|x-y| \leq |x|-|y| \leq |x-y| \tag{5.3.150}$$

Hence the desired result:

$$||x|-|y|| \leq |x-y| \tag{5.3.151}$$

The absolute function is strongly related to some other functions that we will use quite some times in different section of this textbook. A well known an important example is the "signum function" or "sign function" that is an odd mathematical function that extracts the sign of a real number and defined by:

$$\operatorname{sgn}(x) := \begin{cases} -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0, \\ +1 & \text{if } x > 0. \end{cases} \tag{5.3.152}$$



Alternatively (the signum function is the derivative of the absolute value function, up to the indeterminacy at zero):

$$\operatorname{sgn}(x) = \frac{\mathrm{d}}{\mathrm{d}x}|x|, \quad x \neq 0 \tag{5.3.153}$$

Any real number can be expressed as the product of its absolute value and its sign function:

$$x = \operatorname{sgn}(x) \cdot |x| \tag{5.3.154}$$

5.3.5 Calculation Rules (operators priorities)

Frequently in computing (in development in particular), we speak of "operators precedence". In mathematics we speak of "priority of the sets of operations and rules of signs". What is this exactly?

We have already seen what are the properties of addition, subtraction, multiplication, division and power. We therefore insist that the reader distinguishes the concept of "property" of this of "priority" (that we will immediately see) which are (obviously) two completely different things!

In mathematics, in particular, we first define the priorities of the symbols $\{[()]\}$:

1. Operations that are in parenthesis $()$ should be performed first in the polynomial.
2. Operations that are in brackets $[\,]$ should be made afterwards from the results of operations that were in parenthesis $()$.
3. Finally, from the intermediate results of operations that were in parenthesis $()$ and brackets $[\,]$, we calculate the operations that are between the braces $\{\}$.

Let us do an example, this will be more telling.

✍**Example:**

Consider the calculation of the following arithmetic polynomial:

$$\{[5\cdot(8+2)+3\cdot[4+(8+6)\cdot 2]]\cdot(1+9)\}\cdot 7+1 \tag{5.3.155}$$

According to the rules we defined earlier, we first calculate all the elements that are in parenthesis $()$, that is to say:

$$8+2=10 \quad (8+6)=14 \quad (1+9)=10 \tag{5.3.156}$$

Which give us:

$$\{[5\cdot 10+3\cdot[4+14\cdot 2]]\cdot 10\}\cdot 7+1 \tag{5.3.157}$$

Always according to the rules we defined earlier, now we calculate all the elements between brackets by always starting to calculate the terms that are in brackets $[\,]$ at the lowest level of the other brackets $[\,]$. Thus, we first calculate the expression $[4+14\cdot 2]$ that is in the top-level bracket: $[5\cdot 10+3\cdot\ldots]$.

This gives us $[4+14\cdot 2]=32$ and therefore:

$$\{[5\cdot 10+3\cdot 32]\cdot 10\}\cdot 7+1 \tag{5.3.158}$$

It remains to us to calculate now $[5\cdot 10+3\cdot 32]=146$ and therefore:

$$\{146 \cdot 10\} \cdot 7 + 1 \tag{5.3.159}$$

We now calculate the single term in braces, which gives us:

$$\{146 \cdot 10\} = 1460 \tag{5.3.160}$$

Finally it remains:

$$1460 \cdot 7 + 1 = 10221 \tag{5.3.161}$$

Obviously this is a special case ... But the idea remains the same in general.

The priority of arithmetic operators is a problem mainly related to computer languages (as we have already mentioned) because we can only write mathematical relation on a single line and this is many times as source of confusion for people not having technical skill.

Therefore in computing science, the following expression:

$$\frac{-a \cdot (b+c)^d}{e^f} - g \tag{5.3.162}$$

will be written (pretty much on most computer languages):

$$-a * (b+c)\hat{\ }d/e\hat{\ }f - g \tag{5.3.163}$$

A non initiated could read this in many ways:

$$\frac{-a \cdot (b+c)^d}{e^f} - g; \quad \frac{[-a \cdot (b+c)]^d}{e^f} - g; \quad \left\{\frac{-a \cdot (b+c)^d}{e}\right\}^f - g; \quad [-a \cdot (b+c)]^{\frac{d}{ef}} - g \tag{5.3.164}$$

what you will agree, is very dangerous because we will arrive at different results each time (special cases aside...)!

Thus it has logically be defined an order of prioritization of operators such that the operations are carried out in the following order:

1. $-$ Negation
2. ˆ Power (exponentiation)
3. $*$ Multiplication and $/$ division[16]
4. $\backslash$ Integer division (specific to computer science)
5. mod Modulus (see section Number Theory page 356)
6. $+, -$ Addition and subtraction

---

[16] The *Physical Review* journals, for example, state that in the submission instruction that multiplication is of higher precedence than division!

Obviously the rules of parentheses (), brackets [], and braces {} that were defined in mathematics apply also to computing.

Thus we get in the order (we replace every transaction made with a symbol):

First the terms in parentheses:

$$-a * (b + c)\hat{\ } d/e\hat{\ } f - g = -a * \alpha\hat{\ } d/e\hat{\ } f - g \tag{5.3.165}$$

Afterwards we can apply the priority rules of the operators in the order defined previously:

1. First the negation (rule 1):

$$-a * \alpha\hat{\ } d/e\hat{\ } f - g = \beta * \alpha\hat{\ } d/e\hat{\ } f - g \tag{5.3.166}$$

2. The power (rule 2):

$$\beta * \alpha\hat{\ } d/e\hat{\ } f - g = \beta * \chi/\delta - g \tag{5.3.167}$$

3. We apply the multiplication (rule 3):

$$\beta * \chi/\delta - g = \varepsilon/\delta - g \tag{5.3.168}$$

4. And we apply division (rule 3 again):

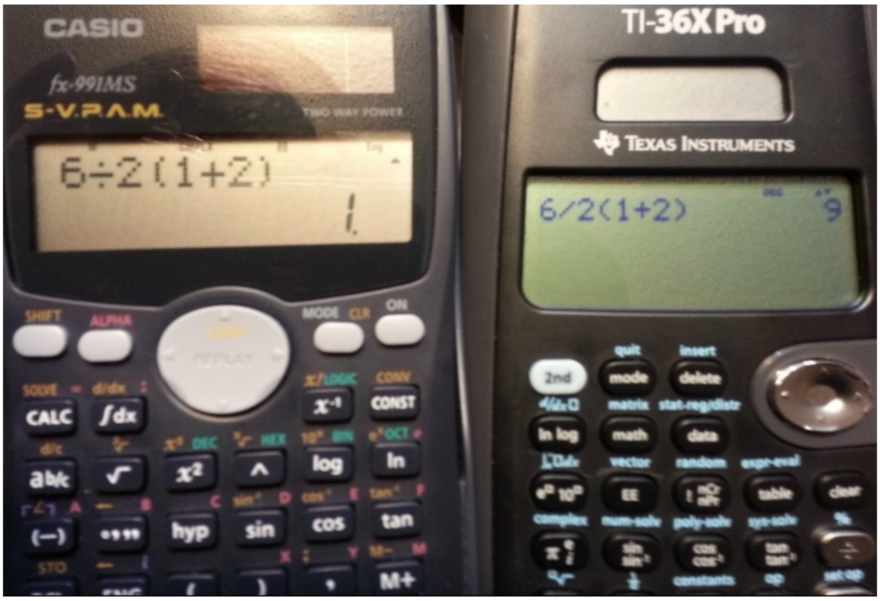
$$\varepsilon/\delta - g = \phi - g \tag{5.3.169}$$

   The rules (4) and (5) does not apply to this particular example.

5. And Finally (rule 6):

$$\phi - g = \varphi \tag{5.3.170}$$

Thus, following these rules, neither a computer nor a human can (should) be wrong in interpreting an equation written on a single line to avoid such issues (hence the important to have ISO standards in all field of the industry and administration):



Remark

Mnemonics are often used to help students in highs-school to remember the most basic rules, but the rules taught by the use of acronyms can be misleading. In the United States, the acronym "PEMDAS" is common. It stands for Parentheses, Exponents, Multiplication, Division, Addition, Subtraction. PEMDAS is often expanded to the mnemonic "Please Excuse My Dear Aunt Sally". Canada and New Zealand use BEDMAS, standing for Brackets, Exponents, Division, Multiplication, Addition, Subtraction. Most common in the UK, India and Australia are BODMAS meaning "B"rackets, "O"f or "O"rder, "D"ivision, "M"ultiplication, "A"ddition and "S"ubtraction. Nigeria and some other West African countries use BIDMAS.

In computer code, however, there are several operators that we do not always find in pure mathematics and which order property frequently change depending from a computer language to another. We will not dwell too much on that stuff as it is almost without end, however, we have below a small description:

- The concatenation operator "&" is evaluated before comparisons operators.
- Comparison operators ($=, <, >, \ldots$) all have equal priority.

However, the leftmost operator in an expression, hold a higher priority.

The logical operators are evaluated in the following order of priority in most computing languages:

1. Not ($\neg$)
2. And ($\wedge$)
3. Or ($\vee$)
4. Xor ($\oplus$)
5. Eqv ($\Leftrightarrow$)
6. Imp ($\Rightarrow$)

Now that we have seen the operator priorities, what are the rules about signs applicable in mathematics and computing science?

First, you must know that these latter rules only apply in the case of multiplication and division. Given two positive numbers $(+x),(+y)$. We have:

$$(+x)\cdot(+y)=(+x)\cdot(+y)=+(x\cdot y) \tag{5.3.171}$$

In other words, the multiplication of two positive numbers is a positive number and this can be generalized to the multiplication of n positive numbers.

We have:

$$(-x)\cdot(+y)=(+x)\cdot(-y)=-(x\cdot y) \tag{5.3.172}$$

In other words, the multiplication of a positive number to a negative number is negative. Which can be generalized: to a positive result of a multiplication if there is an even number of negative numbers, and a negative result if there is an odd number of negative numbers on all n numbers included in the multiplication.

We have:

$$(-x)\cdot(-y)=(-x)\cdot(-y)=+(x\cdot y) \tag{5.3.173}$$

In other words, multiplying two negative numbers is positive. What can be generalized: to a positive result of the multiplication if there is an even number of negative numbers and a negative result if there is an odd number of negative numbers.

About divisions, the reasoning is the same:

$$\frac{(+x)}{(+y)}=+\left(\frac{x}{y}\right) \quad \text{and} \quad \frac{(+y)}{(+x)}=+\left(\frac{y}{x}\right) \tag{5.3.174}$$

In other words, if the numerator and denominator are positive, then the result of the division will be positive.

We have:

$$\frac{(+x)}{(-y)}=\frac{(-x)}{(+y)}=-\left(\frac{x}{y}\right) \quad \text{and} \quad \frac{(-y)}{(+x)}=\frac{(+y)}{(-x)}=-\left(\frac{y}{x}\right) \tag{5.3.175}$$

In other words, if either the numerator or denominator is negative, then the result of the division will be necessarily negative.

We have:

$$\frac{(-x)}{(-y)}=+\left(\frac{x}{y}\right) \quad \text{and} \quad \frac{(-y)}{(-x)}=+\left(\frac{y}{x}\right) \tag{5.3.176}$$

In other words, if the numerator and denominator are positive, then the result of the division, will necessarily be positive.

Obviously if we have a subtraction of terms, it is possible to rewrite it in the form:

$$x-y=x+(-1)y=-1\cdot(-x+y) \tag{5.3.177}$$

~80 %

★★★★☆

11 votes, 76.36%

5.4. Number Theory

TRADITIONALLY, number theory is a branch of mathematics that deals with properties of integers, whether natural or whole integers. More generally, the field of study of this theory concerns a broad class of problems that naturally come from the study of integers. Number theory can be divided into several branches of study (algebraic number theory, computational number theory, etc.) depending on the methods used and the issues addressed.

Remark

The term "arithmetic" was also used to refer to number theory but it is a fairly old term, which is not as popular as in the past.

We chose to introduce in this section only the subjects that are essential to the study of mathematics and theoretical physics of this textbook as well as those to be absolutely part of the general culture of the engineer (some results have application in Biostatistics!).

5.4.1 Principle of good order

We will take for granted the principle that says that every non-empty set $S \subset \mathbb{N}$ contains a smaller element.

We can use this theorem to prove an important property of numbers named " Archimedean property" or "Archimedes' axiom" which states:

For $\forall a, b \in \mathbb{N}$ where a is non-zero, there is at least one positive integer n such that:

$$n \cdot a \geqslant b \tag{5.4.1}$$

In other words, for two unequal values, there is always an integer multiple of the smallest, bigger than the larger one. We name "Archimedean", structures whose elements satisfy a comparison property (see section Set Theory page 380).

While this is trivial to understand in the case of integers let us prove it because it allows us to see the type of approaches used by mathematicians when they must prove trivial items like this...

Proof 5.11.2. Let us suppose the opposite by saying that for $\forall n \in \mathbb{N}$ we have:

$$n \cdot a < b \tag{5.4.2}$$

If we can prove that it is absurd for any n then we will have prove the Archimedean property (and also if a, b are real).

Let us consider then the set:

$$S = \{b - na | n \in \mathbb{N}\} \tag{5.4.3}$$

Using the principle of good order, we deduce that there exist $s_0 \in S$ such as $s_0 \leq s$ for all $s \in S$. Let us write that this smaller element is:

$$s_0 = b - n_0 a \tag{5.4.4}$$

and therefore we also have:

$$b - (n_0 + 1)a \in S \tag{5.4.5}$$

As by hypothesis $na \leq b$ then we must have:

$$b - (n_0 + 1)a \geq b - n_0 a \tag{5.4.6}$$

and if we reorganize and simplify:

$$-(n_0 + 1) \geq -n_0 \tag{5.4.7}$$

and that we simplify the negative sign we had to get...:

$$n_0 + 1 \geq n_0 \tag{5.4.8}$$

an obvious contradiction!

This contradiction leads that the initial assumption as $na < b$ for all n then is false and therefore the Archimedean property is proved by the absurd.

■ Q.E.D.

5.4.2 Induction Principle

Let S be a set of natural numbers that has the following two properties:

P1. $1 \in S$

P2. If $k \in S$, then $k + 1 \in S$

then:

$$S = \mathbb{N} \setminus \{0\} = \mathbb{N}^* \tag{5.4.9}$$

We are build like this the set of natural numbers (refer to the section Set Theory page 382 to see the rigorous construction of the set of natural number $\mathbb{N}$ with the Zermelo-Fraenkel axioms).

Theorem 5.12. Given now:

$$B = \mathbb{N}^* \setminus S \tag{5.4.10}$$

the symbol "\" meaning for recall "excluding". We want to prove that:

$$B = \varnothing \tag{5.4.11}$$

Again, even if it is trivial to understand, let us do the proof because it allows us to see the type of approaches used by mathematicians when they must prove trivial stuff like this...

Proof 5.12.1. Let us suppose the opposite, that is to say:

$$B \neq \varnothing \tag{5.4.12}$$

By the principle of good order, since $B \subset \mathbb{N}$, B must have a smallest element which we will denote by b_0.

But since $1 \in S$ by the property (P1), we have that $b_0 > 1$ and of course also that $1 \in B$, that is to say also $b_0 - 1 \in S$. By using the property (P2), we finally have $b_0 \in S$, that is to say that $b_0 \notin B$, therefore we get a contradiction.

■ Q.E.D.

✍**Example:**

We want to show thanks to the induction principle, that the sum of the first n square equals $n(n+1)(2n+1)/6$, that is to say for $n \geq 1$, we would have to (see section Sequences and Series page 1667):

$$1+2^2+\ldots+n^2=\sum_{i=1}^{n} i^2=\frac{n(n+1)(2n+1)}{6} \tag{5.4.13}$$

First the above relation is easily verified for $n = 1$ we will show that $n = k + 1$ also verifies that relation. Under the induction hypothesis:

$$\begin{aligned}1^2+2^2+\ldots+k^2+(k+1)^2&=\sum_{i=1}^{k+1} i^2=\frac{k(k+1)(2k+1)}{6}+(k+1)^2\\&=\frac{(k+1)(k+2)(2k+3)}{6}\end{aligned} \tag{5.4.14}$$

although we fall back on the assumption of the validity of the first relation but with $n = k + 1$, hence the result.

This proved process is therefore of great importance in the study of arithmetic. Often observation and induction have led to a suspicion of laws it would have been more difficult to find by a priori. We realize the accuracy of formulas by the previous method that gave birth to modern algebra by Fermat and Pascal studies on the Pascal's triangle (see section Calculus page 1363).

5.4.3 Divisibility

Definition (#5.4.63): Given $A, B \in \mathbb{Z}$ with $A \neq 0$. We say that "A divides B (without rest)" if there is an integer q (the quotient) such that:

$$B = Aq \tag{5.4.15}$$

in which case we write to differentiate of the classic division:

$$A|B \tag{5.4.16}$$

Otherwise, we write (if A doesn't divide B without rest):

$$A \nmid B \tag{5.4.17}$$

and we say that "A does not divide B".

Remarks

R1. Remember that $|$ is a relation when the symbol $/$ is an operation!

R2. Do not confuse the expression "A divides B" which means that rest is necessarily zero and the expression "A is the divisor of the division by B" indicating that the rest is not necessarily zero!

Moreover, if $A|B$, we also say that "B can be divided by A" or "B is a multiple of A".

In case where $A|B$ and that $1 \geq A < B$, we will say that A is a "proper divisor" of B.

Moreover, it is clear that $A|0$ regardless of $A \in \mathbb{Z} \setminus \{0\}$ otherwise what we have a singularity.

Here are some basic theorems relating to the division:

Theorem 5.13. If $A|B$, then $A|BC$ whatever $C \in \mathbb{Z}$. Or more formally:

$$\forall C \in \mathbb{Z} : \quad A|B \Rightarrow A|BC \tag{5.4.18}$$

Proof 5.13.1. If $A|B$, the it exists an integer q such that:

$$B = Aq \tag{5.4.19}$$

Then:

$$BC = (Aq)C = A(qC) \tag{5.4.20}$$

and therefore:

$$A|BC \tag{5.4.21}$$

■ Q.E.D.

Theorem 5.14. If $A|B$ and $B|C$, then $A|C$ or more formally:

$$A|B \wedge A|B \Rightarrow A|C \tag{5.4.22}$$

Proof 5.14.1. If $A|B$ and $B|C$ then, there exists two integers q and r such that $B = Aq$ and $C = Br$. More formally:

$$A|B \wedge B|C \rightarrow \exists(q, r) \in \mathbb{N} : \quad B = Aq \wedge C = Br \tag{5.4.23}$$

Therefore:

$$C = A(qr) \tag{5.4.24}$$

and hence:

$$C = A|C \tag{5.4.25}$$

■ Q.E.D.

Theorem 5.15. If $A|B$ and $A|C$ then:

$$A|(Bx + Cy) \quad \forall x, y \in \mathbb{Z} \tag{5.4.26}$$

Proof 5.15.1. If $A|B$ and $A|C$ then, there exists two integers q and r such that $B = Aq$ and $C = Ar$. It follows:

$$Bx + Cy = (Aq)x + (Ar)y = A(qx + ry) \tag{5.4.27}$$

and therefore:

$$A|(Bx + Cy) \quad \forall x, y \in \mathbb{Z} \tag{5.4.28}$$

■ Q.E.D.

Theorem 5.16. If $A|B$ and $B|A$ then:

$$A = \pm B \tag{5.4.29}$$

Proof 5.16.1. If $A|B$ and $B|A$ then, there exists two integers q and r such that $B = Aq$ and $A = Br$.

We then have:

$$B = B(qr) \tag{5.4.30}$$

and thus $qr = 1$. This is why we can have $q = \pm 1$ if $r = \pm 1$ and thus:

$$A = \pm B \tag{5.4.31}$$

■ Q.E.D.

Theorem 5.17. If $A|B$ and $B \neq 0$ then:

$$|A| \leq |B| \tag{5.4.32}$$

Proof 5.17.1. If $A|B$ then there exist an integer $q \neq 0$ such that $B = Aq$. But then:

$$|B| = |A||q| \geq |A| \tag{5.4.33}$$

as $|q| \geq 1$.

■ Q.E.D.

5.4.3.1 Euclidean Division

The Euclidean division is an operation that, to two integers named respectively the "dividend" and "divisor" combines two other integers named the "quotient" and "remainder". Initially define only for non-zero integers, it can be generalized to relative integers and polynomials, for example.

Definition (#5.4.64): We name "euclidean division" or "integer division" of two numbers A and B the operation of dividing B by A, stopping when the rest is strictly less than A.

Let us recall (see section Numbers page 220) that any number which admits exactly two euclidean divisors (such that division gives no remainder) that are the 1 and itself is named a "prime number" (which excludes the number 1 of the list of primes) and that any pair of numbers which have only 1 as common Euclidean divider are said to be "relatively prime", "mutually prime", or "coprime".

Theorem 5.18. Given $A, B \in \mathbb{Z}$ with $A > 0$. The "theorem of the Euclidean division" state that there are unique integers q (quotient) and r (remainder) such as:

$$B = Aq + r \Leftrightarrow r = B - Aq \tag{5.4.34}$$

where $0 \leq r < A$. Furthermore, if $A \nmid B$, then $0 < r < A$.

✍**Example:**

We have one cake with 9 parts (B), we then have to divide it between 4 people (A) with one part remaining ($r = 1$) such that $q = 2$:

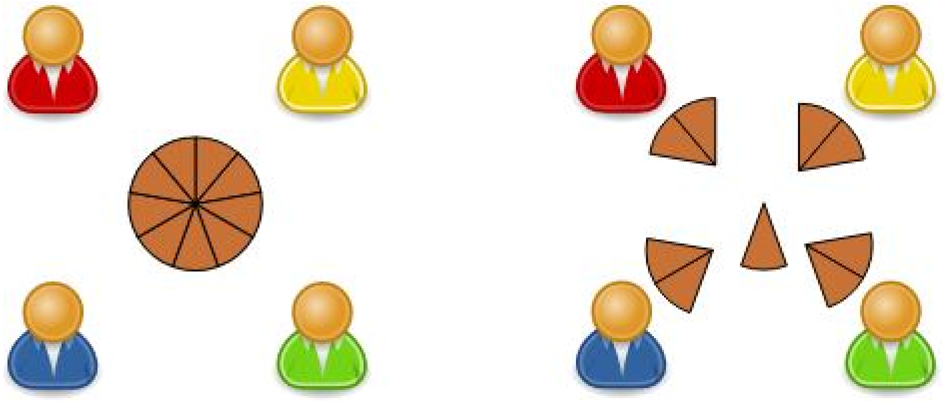


Figure 5.40: **The pie has 9 slices, so each of the 4 people receive 2 slices and 1 is left over**

and therefore:

$$9 = 2 \cdot 4 + 1 \tag{5.4.35}$$

Proof 5.18.1. Let us consider the set:

$$S = \{r = B - qA | q, B \in \mathbb{Z}, A \in \mathbb{Z}^*, B - qA \geq 0\} \tag{5.4.36}$$

It is relatively easy to see that $S \subset \mathbb{N}^* \cup \{0\}$ and that $S \neq \varnothing$, hence, according to the principle of good order, we conclude that S contains a smaller element $r \geq 0$.

Given q the integer satisfying thus:

$$r = B - Aq \tag{5.4.37}$$

We want to first show that $r < A$ assuming the opposite (proof ad absurdum), that is to say that $r \geq A$. So, in this case, we have:

$$B - qA = r \geq A \tag{5.4.38}$$

which is equivalent to:

$$B - (q+1)A = r - A \geq 0 \tag{5.4.39}$$

but $B - (q+1)A \in S$ and:

$$B - (q+1)A < B - qA \tag{5.4.40}$$

This contradicts the fact that:

$$r = B - qA \tag{5.4.41}$$

is the smallest element of S. So $r < A$. Finally, it is clear that if $r = 0$, we have $A|B$, hence the second statement of the theorem.

■ Q.E.D.

Remark

In the statement of the Euclidean division, we assumed that $A > 0$. What do we get when $A < 0$? In this situation, $-A$ is obviously positive, and then we can apply the Euclidean division to B and $-A$. Therefore, there are integers q and r integers such that:

$$B = q(-A) + r \tag{5.4.42}$$

where $0 \geq r < |A|$. But this relation can be written:

$$B = -q(A) + R \tag{5.4.43}$$

where obviously, $-q$ is an integer. The conclusion is that the Euclidean division can be stated in a more general form.

Given $A, B \in \mathbb{Z}$, there exist two integers q and r such that:

$$B = Aq + r \tag{5.4.44}$$

where $0 \geq r < |A|$. Furthermore, if $A \nmid B$, then $0 < r < |A|$

The integers q and r are unique in the Euclidean division. Indeed, if there are two other integers q' and r' such as:

$$B = Aq' + r' \tag{5.4.45}$$

always with $0 \leq r' < A$, then:

$$A(q'-q) = r - r' \tag{5.4.46}$$

and therefore:

$$A|(r-r') \tag{5.4.47}$$

Following the previous theorem we have if $r - r' \neq 0$ that $|r - r'| \geq A$.

But, this last inequality is impossible as by construction $-A < r - r'$. Therefore $r = r'$ and, as $A \neq 0$, then $q' = q$ hence the unicity.

In mathematics and more precisely in Set Algebra, in the context of ring theory, a ring is named an "Euclidean ring" if it is possible to define a Euclidean division in it. The relative integers form the most simplest case of Euclidean ring. This set verifies the following property:

$$\forall a \in \mathbb{Z} \quad \forall b \in \mathbb{Z}\backslash\{0\} \quad \exists q, r \in \mathbb{Z} \quad \text{such that} \quad a = bq + r \quad \text{and} \quad |r| < |b| \tag{5.4.48}$$

The reader must remember that natural numbers with addition and multiplication as the operations do not form a ring or a field. They don't form a ring because addition does not have inverses in the natural numbers which is a property required for a ring. Hence the fact that the smallest Euclidean ring is the made of relative integers!

5.4.3.1.1 Greatest common divisor

The "greatest common divisor" (gcd) (also known as "greatest common factor" (gcf), "highest common factor" (hcf), "greatest common measure" (gcm), or "highest common divisor") of two or more integers, when at least one of them is not zero, is the largest positive integer that divides the numbers without a remainder.

Definition (#5.4.65): Given $a, b \in \mathbb{Z}$ such as $ab \neq 0$. The "greatest common divisor" (gc) of a and b, denoted:

$$(a, b) \tag{5.4.49}$$

is the positive integer n that satisfies the following two properties:

P1. $d|a$ and $d|b$ (so without remainder r in the division!)

P2. If $c|a$ and $c|b$ then $c \leq d$ and $c|d$ (by definition!)

Note that 1 is always a common divisor of two arbitrary integers.

✍**Example:**

Let us consider the positive integers 36 and 54. A common divisor of 36 and 54 is a positive integer that divides 36, and also 54. For example, 1 and 2 are common divisors 36 and 54.

$$\begin{aligned}\text{Div}_{36}&=\{1,2,3,4,6,9,12,18,36\}\\ \text{Div}_{54}&=\{1,2,3,6,9,18,27,54\}\end{aligned}\qquad(5.4.50)$$

We have the intersection represented by the following Venn diagram (see section Set Theory page 378):

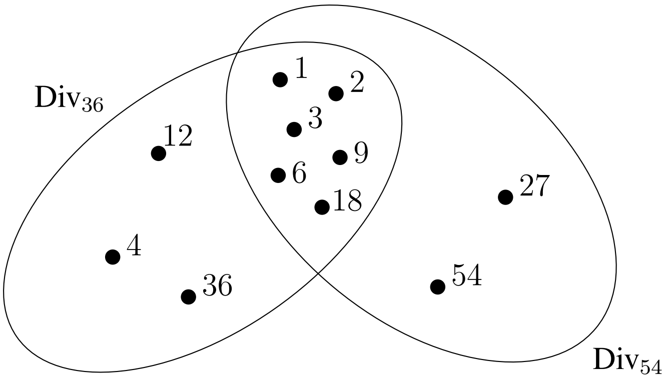


Figure 5.41: **Venn diagram of common divisors**

with the following set of common divisors:

$$\text{Div}_{36}\cap\text{Div}_{54}=\{1,2,3,6,9,18\}\qquad(5.4.51)$$

and hence, the highest common factor is:

$$\max\{1,2,3,6,9,18\}=18\qquad(5.4.52)$$

and we find that the set of common divisors of 36 and 54 is also the set of dividers of 18.

However it is not necessarily obvious that the greatest common divisor other than 1 of two integers a and b that are not relatively prime always exists. This is proved by the following theorem (however, if the gcd exists, it is by definition unique!) named "Bézout's theorem" that can also gives the opportunity to prove other interesting properties of two numbers as we shall see later.

Theorem 5.19. Given $a, b \in \mathbb{Z}$ such that $ab \neq 0$. If d divides a and d divides b (for both without remainder r !) then there must two integers x and y such that:

$$\boxed{d=(a,b)=ax+by}\qquad(5.4.53)$$

This relation is named the "Bézout's identity" and it is a linear Diophantine equation (see section Calculus page 1378).

Proof 5.19.1. Obviously, if a and b are relatively prime we know that d is then 1.

To prove the Bézout identity let first consider the set:

$$S=\{d=ax+by|x,y\in\mathbb{Z},ax+by>0\}\qquad(5.4.54)$$

As $S \subset \mathbb{N}$ and $S \neq \varnothing$, we can use the principle of good order and conclude that S has a smaller element d. We can then write:

$$d = ax_0 + by_0 \tag{5.4.55}$$

for some given choice $x_0, y_0 \in \mathbb{Z}$. So it is sufficient to prove that $d = (a, b)$ to prove the Bézout identity!

Let us proceed with a proof by contradiction by assuming $d \nmid a$. Then if this is the case, following the Euclidean division, there exist $q, r \in \mathbb{Z}$ such as $a = qd + r$, where $0 < r < d$. But then:

$$r = a - qd = a - q(ax_0 + by_0) = a(1 - qx_0) + b(-qy_0) \tag{5.4.56}$$

Thus we have that $r \in S$ and $r < d$, which contradicts the fact that d is the smallest possible element of S. Thus we have proven not only that $d|a$, but also that d always exists and, in the same way we prove that $d|b$.

■ Q.E.D.

Corollary 5.19.1. As important corollary let us now prove that if $a, b \in \mathbb{Z}$ such that $ab \neq 0$, then:

$$S = \{ax + by | x, y \in \mathbb{Z}\} \tag{5.4.57}$$

is the set of all multiples of $d(a, b)$.

Proof 5.19.2. As $d|a$ and $d|b$, then we have necessarily $dax + by|$ for any $x, y \in \mathbb{Z}$. Either $M = \{nd|n \in \mathbb{Z}\}$. Our problem is then reduced to prove the fact that $S = M$.

Given first $s \in S$ which means that $d|s$ and involves $s \in M$.

Given a $m \in M$, this would mean that $m = nd$ for a certain $n \in \mathbb{Z}$.

As $d = ax_0 + by_0$ for any choice of integers $x_0, y_0 \in \mathbb{Z}$, then:

$$m = nd = n(ax_0 + by_0) = a(nx_0) + b(ny_0) \in S \tag{5.4.58}$$

■ Q.E.D.

The assumptions may seem complicated but put your attention a given time on the last equality. You will quickly understand!

Remark

If instead of defining the greatest common divisor of two non-zero integers, we allow one of them to be equal to 0, say: $a \neq b$, $b = 0$. In this case, we have $a|b$ and, according to our definition of the GCD, it is clear that $(a, 0) = |a|$.

Given $d = (a, b)$ and $m \in \mathbb{Z}$, then we have the following properties of the GCD (without proof but if a reader request them we will give the details):

P1. $(a,b+ma)=(a,b)=(a,-b)$

P2. $(am,bm)=|m|(a,b)$ where $m\neq 0$

P3. $\left(\dfrac{a}{d},\dfrac{b}{d}\right)=1$

P4. If $g\in\mathbb{Z}\setminus\{0\}$ such that $g|a$ and $g|b$ then $\left(\dfrac{a}{g},\dfrac{b}{g}\right)=\dfrac{1}{|g|}(a,b)$

In some books, these four properties are proved using intrinsically the property itself. Personally we abstain make usage of this approach because doing this is more ridiculous than anything else as the statement of the property is a proof in itself.

✍**Example:**

In a bathroom, a high-end client wants to cover the wall above the bathtub with a whole number of (very expensive) square-shaped earthenware tiles whose side is a whole number of centimetres as large as possible. Knowing that the wall is 210 [cm] high and 135 [cm] wide, what is the length, in centimetres, of the side of a tile. How many tiles will then be needed?

For the number of tiles to be a whole number, the size of a tile must be a common divisor of both 210 and 135. The largest possible size corresponds to the GCD of 210 and 135, we need tiles with a side of 15 [cm]. So we will need $210/15=14$ tiles in height and $135/15=9$ tiles in width. So a total of $14\cdot 9=126$ tiles.

Let us now develop a method (algorithm) that will be very useful to us to calculate (determine) the greatest common divisor of two integers (sometimes useful computing science).

5.4.3.2 Euclidean Algorithm

The Euclidean algorithm is an algorithm for determining the greatest common divisor of two integers (we have hesitate to put this subject in the section of Theoretical Computing...).

To address this method intuitively, you must know that you need to see that an integer as a length, a pair of integers as a rectangle (sides) and their GCD is the size of the largest square for tile (paving) their rectangle by definition (yes if you think for a moment it's quite logical!).

The algorithm decomposes the original rectangle into squares, always smaller and smaller, by successive Euclidean division of the length by the width, then the width by the remainder until a zero remainder. We must understand this geometric approach to then understand the algorithm.

✍**Example:**

Let us consider that we seek the GCD of (a,b) where b is equal 21 and a is equal 15 and keep in mind that the GCD, besides the fact that it divides a and b, must leave a zero remainder! In other words it must divide the remainder of the division of b by a also!

So we have the following rectangle of 21 by 15:

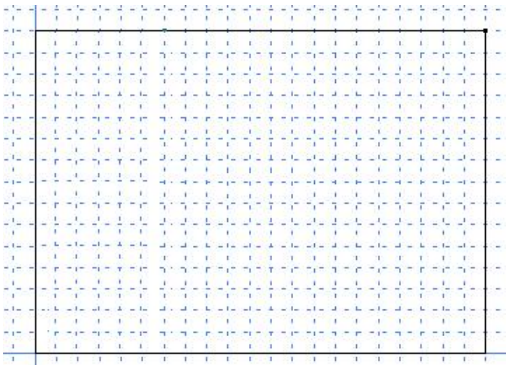


Figure 5.42: **First step of the GCD algorithm**

First we see if 15 is the GCD (it always starts with the smallest). We then divide 21 by 15, which is equivalent geometrically to:

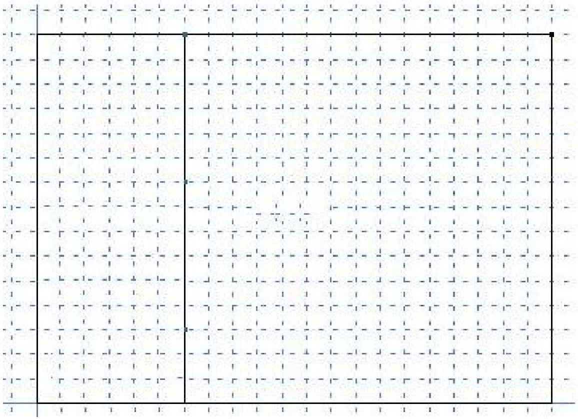


Figure 5.43: **Second step of the GCD algorithm**

15 is therefore not the GCD (we suspected it...). We immediately see that we can not pave the rectangle with a square of 15 by 15.

So we have a remainder of 6 (left rectangle). The GCD as we know must, if it exists, by definition divide that remains and leave a zero remainder.

So we have a rectangle of 15 by 6. So we are looking now to pave this new rectangle because we know that the greatest common divisor is by construction less than or equal to 6. Then we have:

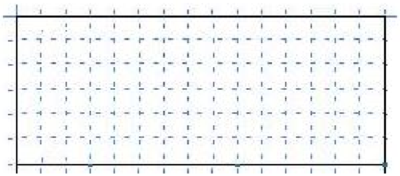


Figure 5.44: **Third step of the GCD algorithm**

So we divide 15 by remainder 6 (this result will be less than 6 and immediately permits to tests whether the remainder will be the GCD). We are getting:

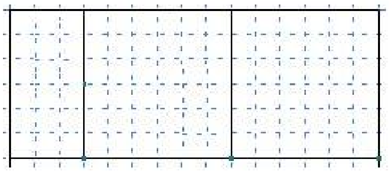


Figure 5.45: **Fourth step of the GCD algorithm**

Again, we can not pave the rectangle only with squares. In other words, we have a non-zero remainder which is 3. Given now a rectangle of 6 by 3. So we are looking now to pave the new rectangle because we know that the greatest common divisor is by construction less than or equal to 3 and that it will leave a remainder equal to zero, if it exists. We then have geometrically:

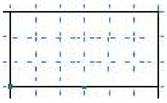


Figure 5.46: **Fifth step of the GCD algorithm**

We divide 6 by 3 (which will be less than 3 and permits us to test immediately whether the rest will be the GCD):

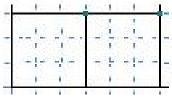


Figure 5.47: **Sixth step of the GCD algorithm**

and it's all good! We then have 3 that leaves us with a remainder equal to zero and divides the remainder 6 so this is the GCD. So we have in the end:

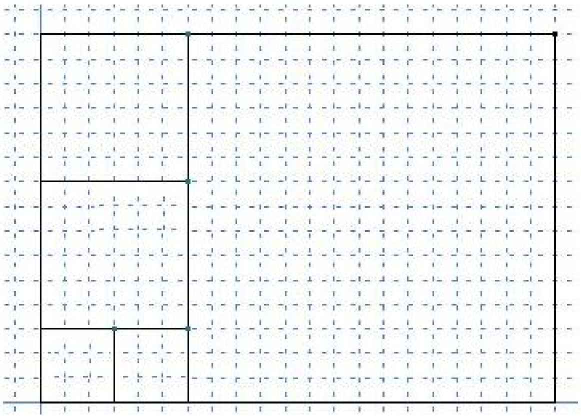


Figure 5.48: **Summary of the GCD algorithm**

Now let us see the equivalent formal approach.

Given $a,b\in\mathbb{Z}$, where $a>0$. Applying successively the Euclidean division (with $b>a$), we get the following sequence of equations:

$$\begin{aligned} b &= qa_1 + r_1 & 0 < r_1 < a \\ a &= r_1 q_2 + r_2 & 0 < r_2 < r_1 \\ r_1 &= r_1 2 q_3 + r_3 & 0 < r_3 < r_2 \\ &\ldots \\ r_{j-2} &= r_{j-1} q_j + r_j & 0 < r_j < r_{j-1} \\ r_{j-1} &= r_j q_{j+1} \end{aligned} \tag{5.4.59}$$

if $d = (a, b)$, then $d = r_j$. With the corresponding pseudocode algorithm:

Algorithm 1: GCD pseudocode algorithm

```
Data: $a$,$b$
Result: $b$
initialization;
$r = a \mod b$;
while $r \neq 0$ do
    $a := b$;
    $b := r$;
    $f(b) := f(a)$;
    $a := x_1$;
    $f(a) := f(x_1)$;
Display $x_1$;
```

Otherwise even more formally:

Proof 5.19.3. We want first prove that $r_j = (a, b)$. But, following the property P1:

$$(a, b + ma) = (a, b) = (a, -b) \tag{5.4.60}$$

we have:

$$(a, b) = (a, r_1) = (r_1, r_2) = \ldots = (r_{j-1}, r_j) \tag{5.4.61}$$

To prove the second property of the Euclide's algorithm, we write the prior-previous equation of the system under the form:

$$r_j = r_{j-2} - q_j r_{j-1} \tag{5.4.62}$$

Now, using the previous equation this prior-previous equation of the system, we have:

$$r_j = r_{j-2} - q_j(r_{j-3} - q_{j-1} r_{j-2}) = (1 + q_j q_{j-1}) r_{j-2} + (-q_j) r_{j-3} \tag{5.4.63}$$

Continuing this process, we can express r_j as a linear combination of a and b.

■ Q.E.D.

✍**Example:**

Let us calculate the greatest common divisor of $(429, 966)$ and express this number as a linear combination of 429 and 966.

We apply obviously Euclide's algorithm:

$$\begin{aligned}966 &= 429 \cdot 2 + 108\\ 429 &= 108 \cdot 3 + 105\\ 108 &= 105 \cdot 1 + 3\\ 105 &= 35 \cdot 3\end{aligned} \tag{5.4.64}$$

We therefore conclude that:

$$r_j = d = (966, 429) = 3 \tag{5.4.65}$$

and, in addition, that:

$$\begin{aligned}3 &= 108 - 105 \cdot 1 = 108 - (429 - 108 \cdot 3) = 108 \cdot 4 - 429\\ &= (966 - 429 \cdot 2) \cdot 4 - 429\\ &= 966 \cdot 4 - 429 \cdot 9 = 966 \cdot 4 + 429 \cdot (-9)\end{aligned} \tag{5.4.66}$$

Thus the GCD is indeed expressed as a linear combination of a and b and constitutes as such the GCD.

Definition (#5.4.66): We say that the integers $a_1, a_2, \ldots, a_n$ are for recall "relatively prime" if:

$$(a_1, a_2, \ldots, a_n) = 1 \tag{5.4.67}$$

5.4.3.3 Least Common Multiple

The least common multiple (also named the "lowest common multiple" or "smallest common multiple") of two integers a and b, usually denoted by $\text{LCM}(a, b)$, is the smallest positive integer that is divisible by both a and b. Since division of integers by zero is undefined, this definition has meaning only if a and b are both different from zero.

The LCM is familiar from grade-school arithmetic as the "lowest common denominator (LCD)" (also named "smallest common denominator") that must be determined before fractions can be added, subtracted or compared. The LCM of more than two integers is also well-defined: it is the smallest positive integer that is divisible by each of them.

For example, to perform the calculation:

$$\frac{65}{18372} - \frac{21}{15310} \tag{5.4.68}$$

We can reduce the fractions to a common denominator. In the absence of another method, we write:

$$\frac{65}{18372}-\frac{21}{15310}=\frac{65\cdot 15310}{18372\cdot 15310}-\frac{21\cdot 18372}{15310\cdot 18372}=\frac{995150}{281275320}-\frac{385812}{281275320}=\frac{609338}{281275320} \tag{5.4.69}$$

We then simplify:

$$\frac{609338}{281275320} \tag{5.4.70}$$

by finding a common divisor between 609338 and 281275320 (...) and we get:

$$\frac{199}{91860} \tag{5.4.71}$$

If we had been able to anticipate that 91860 was a common multiple of 18372 and 15310 (and even that it was the smallest common multiple!), the calculations would have been much easier.

Indeed, since 91860 is equal to $18372\cdot 5$ and to $15310\cdot 6$, we could have written:

$$\frac{65}{18372}-\frac{21}{15310}=\frac{65\cdot 5}{18372\cdot 5}-\frac{21\cdot 6}{15310\cdot 6}=\frac{325}{91860}-\frac{126}{91860}=\frac{199}{91860} \tag{5.4.72}$$

Therefore, finding the common multiples of two numbers is of definite interest for calculation when we don't have any computer assistance!

Definitions (#5.4.67):

D1. Given $a_1,a_2,\ldots,a_n\in\mathbb{Z}\setminus\{0\}$, we say that m is a "common multiple" of $a_1,a_2,\ldots,a_n$ if $a_i|m$ for $i=1,2,\ldots,n$

D2. Given $a_1,a_2,\ldots,a_n\in\mathbb{Z}\setminus\{0\}$, we name "lowest common multiple" (LCM) of $a_1,a_2,\ldots,a_n$ if $a_i|m$ for $i=1,2,\ldots,n$ denoted traditionally by:

$$[a_1,a_2,\ldots,a_n] \tag{5.4.73}$$

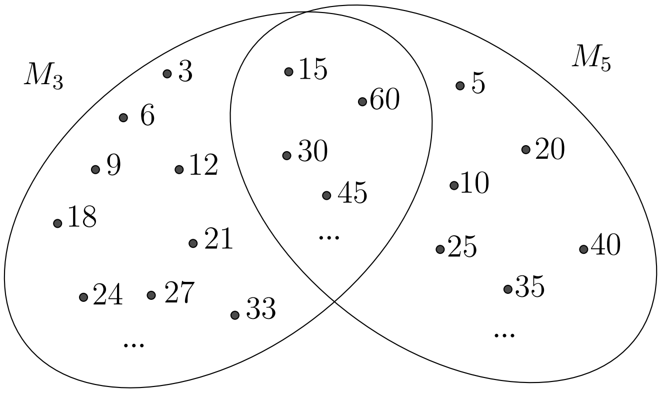
the lowest integer positive common multiple to all common multiples of $a_1,a_2,\ldots,a_n$.

✍**Examples:**

E1. Let us consider the positive integers 3 and 5. A common multiple of 3 and 5 is a positive integer which is both a multiple of 3, and a multiple of 5. In other words, which is divisible by 3 and 5. We have therefore:

$$\begin{aligned}M_3&=\{3,6,9,12,15,18,21,24,27,30,\ldots\}\\M_5&=\{5,10,15,20,25,30,35,40,45,\ldots\}\end{aligned} \tag{5.4.74}$$

We then have the intersection represented by the following Venn diagram (see section Set Theory 378):


M_3
3
6
9
12
18
21
24
27
33
...
15
60
30
45
...
M_5
5
20
10
25
40
35
...

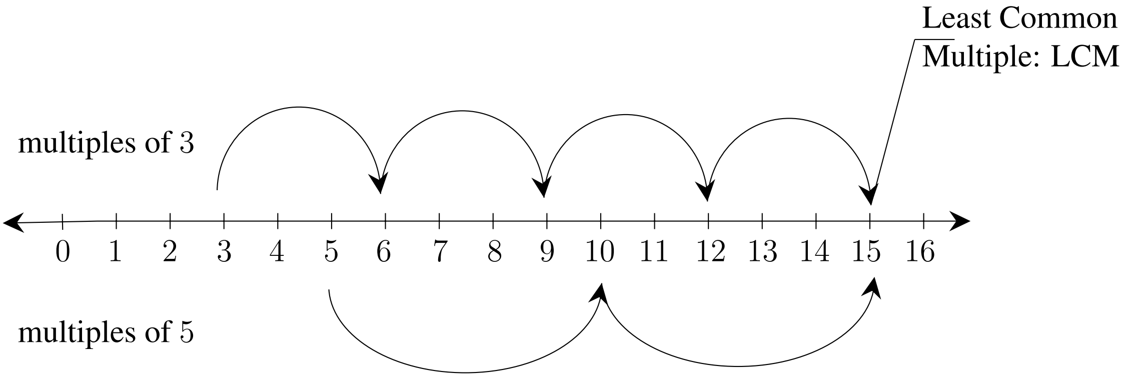
we then have the following set of common multiples:

$$M_3 \cap M_5 = \{15, 30, 45, 60, \ldots\} \tag{5.4.75}$$

and therefore the LCM is given by:

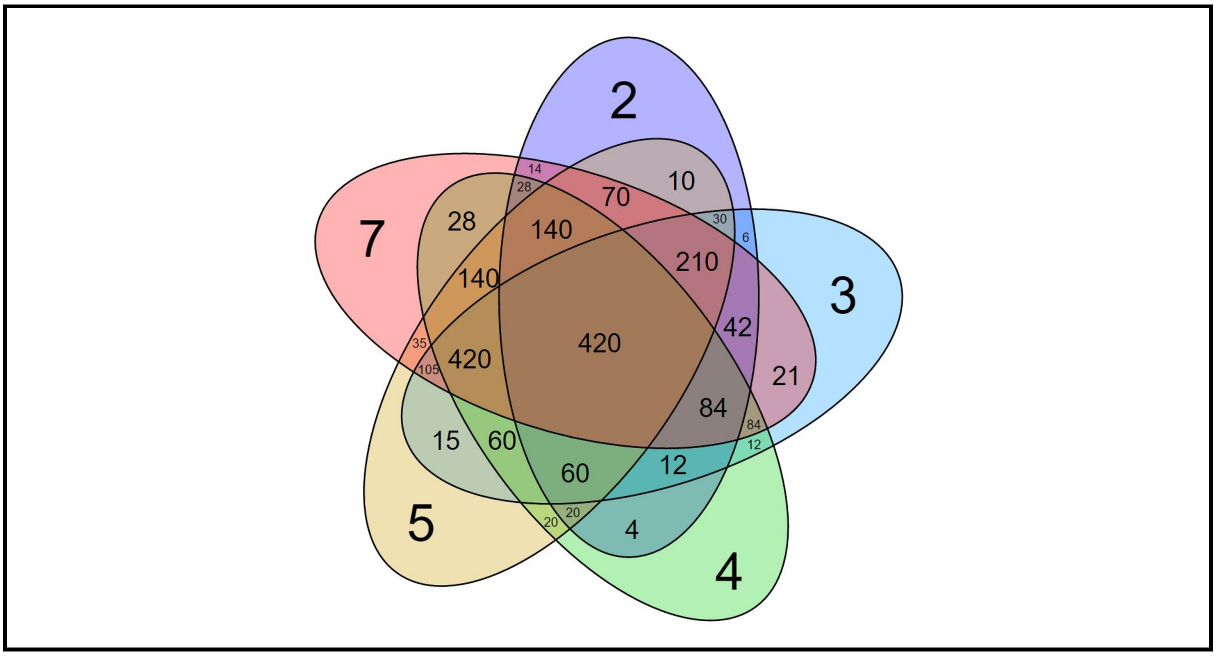
$$\text{LCM} = \min\{15, 30, 45, 60, \ldots\} = 15 \tag{5.4.76}$$

Or if it can help here is another possible visualization of the concept:


Least Common
Multiple: LCM
multiples of 3
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
multiples of 5

We see obviously that all the common multiples of 3 and 5 is the set of multiples of 15.

E2. Wikipedia gives also a nice example of a Venn Diagram (see section Set Theory page 378) when we seek for a LCM of 5 multiples using a visual tool:



Remark

Given $a_1, a_2, \ldots, a_n \in \mathbb{Z} \setminus \{0\}$. Then the least common multiple exists. Indeed, consider the set E of natural integers m that for all i divide a_i. What we will write:

$$E = \{m|a_i|m \in \mathbb{N}, i = 1, 2, \ldots, n\} \tag{5.4.77}$$

Since we have necessarily $|a_1 a_2 \ldots a_n| \in E$, then the set is not empty and, according to the axiom of good order, the set E contains a smaller positive element.

Let us now see some theorems related to the LCM:

Theorem 5.20. If m is any common multiple of $a_1, a_2, \ldots, a_n$ then $[a_1, a_2, \ldots, a_n]|m$ that is to say that m divides each of the a_i.

Proof 5.20.1. Given $M = [a_1, a_2, \ldots, a_n]$. Then, by the Euclidean division, there are integers q and r such that:

$$m = qM + r \quad 0 \leq r < M \tag{5.4.78}$$

It suffices to show that $r = 0$. Let us suppose that $r \neq 0$ (reductio ad absurdum). Since $a_i|m$ and $a_i|M$, then we have $a_i|r$ and this for $i = 1, 2, \ldots, n$. So r is common multiple of $a_1, a_2, \ldots, a_n$ of the smallest than the LCM. We just obtained a contradiction, which proves the theorem.

■ Q.E.D.

Theorem 5.21. If $k > 0$, then $[ka_1, ka_2, \ldots, ka_n] = k[a_1, a_2, \ldots, a_n]$

The proof will be assumed obvious (if not as always contact us and will add the details!)

Theorem 5.22. $[a, b] \cdot (a, b) = |ab|$. Or more explicitly: LCM$(a; b) \cdot$ GCD$(a; b) = |a \cdot b|$.

Proof 5.22.1.

Lemma 5.22.1. For this proof, we will use the "Euclid's lemma" that says that if $a|bc$ and $(a, b) = 1$ then $a|c$.

In other words Euclid's lemma captures a fundamental property of prime numbers, namely: If a prime divides the product of two numbers, it must divide at least one of those numbers. It is also named "Euclid's first theorem". This lemma is the key of the proof of the fundamental theorem of arithmetic that we will see just further below!

Indeed, this can be easily verified because we have seen that there exists $x, y \in \mathbb{Z}$ such as $1 = ax + by$ and then $c = acx + bcy$. But $a|ac$ and $a|bc$ imply that $a|(acx + bcy)$, that is to say also that $a|c$.

Ok let us now return to our theorem:

Since $(a, b) = (a, -b)$ and $[a, b] = [a, -b]$, it suffices to prove the result for positive integers a and b.

First of all, let consider the case where $(a, b) = 1$. The integer $[a, b]$ being a multiple of a, we can write $[a, b] = ma$. Thus, we have $b|ma$ and since $(a, b) = 01$, it follows, by Euclid's lemma, that $b|m$. Therefore, $b \leq m$ and then $ab \leq am$. But ab is a common multiple of a and b that can not be smaller than the LCM. therefore $ab = ma = [a, b]$.

For the general case, that is to say $(a, b) = d > 1$, we have, according to the property:

$$\left(\frac{a}{d}, \frac{b}{d}\right) = 1 \tag{5.4.79}$$

and with the result obtained previously that:

$$\left[\frac{a}{d}, \frac{b}{d}\right]\left(\frac{a}{d}, \frac{b}{d}\right) = \frac{a}{d}\frac{b}{d} \tag{5.4.80}$$

When we multiply both sides of the equation by d^2, the result follows and the proof is done.

■ Q.E.D.

Here are some trivial properties that we give without proof:

P1. $\text{LCM}(a; b) = \text{LCM}(b; a)$

P2. If b is a multiple of a, then $\text{LCM}(a; b) = b$

P3. If a and b are mutually prime, then we have $\text{LCM}(a; b) = a \cdot b$

P4. If k is a natural integer, then we have $\text{LCM}(ka; kb) = k\text{LCM}(a; b)$ (homogeneity)

5.4.3.4 Fundamental Theorem of Arithmetic

The fundamental theorem of arithmetic says that every natural number $n > 1$ can be written as a product of primes, and this representation is unique, except for the order in which the prime factors are arranged.

The theorem establishes the importance of prime numbers. Essentially, they are the building blocks of building positive integers, each positive integer containing primes in a unique way.

Remark

This theorem is sometimes named "factorization theorem" (wrongly ... because some other theorems have the same name ...).

So let's go:

Theorem 5.23. Every integer greater than 1 either is prime itself or is the product of prime numbers, and that this product is unique, up to the order of the factors.

Remark

This theorem is one of the main reasons why 1 is not considered a prime number: if 1 were prime, the factorization would not be unique.

Proof 5.23.1. The proof uses Euclid's lemma: if a prime p divides the product of two natural numbers a and b, then either p divides a or p divides b (or both).

If n is prime, and therefore product of a unique prime integer, namely itself, the result is true and the proof is complete (say that a prime number is product of itself is obviously a misnomer!). Suppose that n is not prime and therefore strictly greater than 1 and consider the set:

$$D = \{d|n \text{ and } 1 < d < n\} \tag{5.4.81}$$

So, $D \subset \mathbb{N}$ and since n is composite, we have that $D \neq \varnothing$. According to the principle of good order, D has a smaller element p_1 that is prime, otherwise the minimum choice of p_1 is contradicted. We can the write $n = p_1 n_1$. If n_1 is prime, then the proof is complete. If n_1 is also composite, then we repeat the same argument as before and we deduce the existence of a prime number p_2 and of an integer $n_2 < n_1$, such as $n = p_1 p_2 n_2$. By continuing we come inevitably to the conclusion that n_k will be prime.

So finally we have well shown that any number can be decomposed into prime numbers factors with the principle of good order.

■ Q.E.D.

We do not know to this day a simple law that allows to calculate the n-th prime factor p_n. Thus, to know if an integer m is a prime, it is almost easier at this date to verify its presence in a table of prime numbers.

In fact, we use nowadays the following method:

Given an integer m, if we want to determine whether it is prime or not, we calculate if it is divisible by the primes number p_n belonging to the set:

$$\{p_n \in \mathbb{N} | p_n < \sqrt{M}\} \tag{5.4.82}$$

✍**Example:**

The integer 223 is neither divisible by 2 or by 3 or by 5 or by 7, or by 11, or by 13. It is useless to continue with the next prime number, because $17^2 = 289 > 223$. We conclude therefore that the number 223 is prime.

We can now address the concept of "set of divisors ". In $\mathbb{N}$ the set of division of 200 is $\{1; 2; 4; 5; 8; 10; 20; 25; 40; 50; 100; 200\}$.

We can derive this result from the prime factorization of 200. Indeed, this factorization is:

$$200 = 2^3 \cdot 5^2 \tag{5.4.83}$$

We can then verify that the divisors of 200 are the numbers that can be written in the form $2^{\beta_1}5^{\beta_2}$ with $0 \leq \beta_1 \leq 3$ and $0 \leq \beta_2 \leq 2$.

Theorem 5.24. Given n an integer greater than or equal to 2, whose decomposition in prime factors is $n = {p_1}^{\alpha_1}{p_2}^{\alpha_2}\ldots p_k^{\alpha_k}$. Set set of natural divisors of n is the set of integers d that can be writtent as $d = p_1^{\beta_1}p_2^{\beta_2}\ldots p_k^{\beta_k}$ where $\beta_1, \beta_2, \ldots, \beta_k$ are natural integers such that $0 \leq \beta_1 \leq \alpha_1, 0 \leq \beta_2 \leq \alpha_2, \ldots, 0 \leq \beta_k \leq \alpha_k$.

Proof 5.24.1. n is an integer greater than or equal to 2, whose prime factorization is:

$$n = p_1^{\alpha_1}p_2^{\alpha_2}\ldots p_k^{\alpha_k} \tag{5.4.84}$$

Let us consider an integer d written in the form:

$$d = p_1^{\beta_1}p_2^{\beta_2}\ldots p_k^{\beta k} \tag{5.4.85}$$

where $\beta_1, \beta_2, \ldots, \beta_k$ are natural integers such that $0 \leq \beta_1 \leq \alpha_1, 0 \leq \beta_2 \leq \alpha_2, \ldots, 0 \leq \beta_k \leq \alpha_k$.

We can then write:

$$n = p_1^{\alpha_1}p_2^{\alpha_2}\ldots p_k^{\alpha_k} = \left(p_1^{\beta_1}p_2^{\beta_2}\ldots p_k^{\beta_k}\right) \cdot \left(p_1^{\alpha_1-\beta_1}p_2^{\alpha_2-\beta_2}\ldots p_k^{\alpha_k-\beta_k}\right) \tag{5.4.86}$$

Hence:

$$n = d \cdot q \tag{5.4.87}$$

with:

$$q = p_1^{\alpha_1-\beta_1}p_2^{\alpha_2-\beta_2}\ldots p_k^{\alpha_k-\beta_k} \tag{5.4.88}$$

We know that $\beta_1, \beta_2, \ldots, \beta_k$ are natural integers such that $0 \leq \beta_1 \leq \alpha_1, 0 \leq \beta_2 \leq \alpha_2, \ldots, 0 \leq \beta_k \leq \alpha_k$.

Hence $\alpha_1 - \beta_1; \alpha_2 - \beta_2; \ldots; \alpha_k - \beta_k$ are natural integers.

Therefore, q is a natural number and d is therefore a divisor of n.

Consider an integer d that is a divisor of n:

- If $d=1$, d can be written as $d=p_1^{\beta_1}p_2^{\beta_2}\ldots p_k^{\beta_k}$ with $\beta_1=\beta_2=\ldots=\beta_k=0$.
- If $d>1$, given $d=q_1^{\gamma_1}q_2^{\gamma_2}\ldots q_r^{\gamma_r}$ the decomposition of d in prime factorization.

  Then $q_1,q_2,\ldots,q_r$ are prime divisors of d, therefore they are prime divisors of n, so these are some of the numbers $p_1,p_2,\ldots,p_k$. We can then write $d=p_1^{\beta_1}p_2^{\beta_2}\ldots p_k^{\beta_k}$ where $\beta_1,\beta_2,\ldots,\beta_k$ are natural integers.

  If the number p_i does not appear in the decomposition of d, then we have $\beta_i=0$.

Let us prove that we have then that $\beta_1\leq\alpha_1$:

We know that d divides n, and $p_1^{\beta_1}$ divides d, therefore $p_1^{\beta_1}$ divides n, hence $p_1^{\beta_1}$ divides $p_1^{\alpha_1}p_2^{\alpha_2}\ldots p_k^{\alpha_k}$.

By contradiction (reductio ad absurdum), if β_1 was striclty greater than α_1, then $p_1^{\beta_1-\alpha_1}$ would divide $p_2^{\alpha_2}\ldots p_k^{\alpha_k}$ threfore p_1 would divide $p_2^{\alpha_2}\ldots p_k^{\alpha_k}$, which is not possible since p_1 is not one of the numbers $p_2,\ldots,p_k$.

We have therefore $\beta_1\leq\alpha_1$. Similarly, we can justify that $\beta_2\leq\alpha_2,\ldots,\beta_k\leq\alpha_k$ following the same procedure.

Therefore d can be written as $d=p_1^{\beta_1}p_2^{\beta_2}\ldots p_k^{\beta_k}$ where $\beta_1,\beta_2,\ldots,\beta_k$ are natural integers such as $0\leq\beta_1\leq\alpha_1,0\leq\beta_2\leq\alpha_2,\ldots,0\leq\beta_k\leq\alpha_k$...

■ Q.E.D.

Remark

If n has a prime factorization $n=p_1^{\alpha_1}p_2^{\alpha_2}\ldots p_k^{\alpha_k}$, the number of natural divisors of n is then $(\alpha_1+1)\cdot(\alpha_2+1)\cdot\ldots\cdot(\alpha_k+1)$.

Indeed, every natural divisor can be written $p_1^{\beta_1}p_2^{\beta_2}\ldots p_k^{\beta_k}$ and every number β_i can take the (α_i+1) values from 0 to α_i.

The prime factorization of 200 is: $200=2^3\cdot 5^2$.

This allows us to say that the number of natural divisors of 200 is: $(3+1)\cdot(2+1)=4\cdot 3=12$.

Thanks to this fundamental theorem we can prove powerful results on the GCD and LCM, the proof of which is also very instructive from a methodological point of view.

Let us begin by recalling that Euclid's algorithm, seen above, allows us to justify that:

$$\text{GCD}(1'500;4'725)=75 \tag{5.4.89}$$

We have also proved earlier above that:

$$\text{GCD}(a;b)\cdot\text{LCM}(a;b)=a\cdot b \tag{5.4.90}$$

Hence:

$$\text{GCD}(1,500;4,725)\cdot\text{LCM}(1,500;4,725)=1,500\cdot 4,725 \tag{5.4.91}$$

From this we deduce that:

$$\text{LCM}(1,500;4,725)=\frac{1,500\cdot 4,725}{\text{GCD}(1,500;4,725)}=\frac{7,087,500}{75}=94,500 \tag{5.4.92}$$

We have:

$$1,500=2^2\cdot 3\cdot 5^3 \quad\text{and}\quad 4'725=3^3\cdot 5^2\cdot 7 \tag{5.4.93}$$

If we write the decomposition using the same prime factors for both numbers and allowing the use of zero exponents, we can write:

$$1,500=2^2\cdot 3^1\cdot 5^3\cdot 7^0 \quad\text{and}\quad 4,725=2^0\cdot 3^3\cdot 5^2\cdot 7^1 \tag{5.4.94}$$

We can then observe that the number obtained by taking the smallest exponents is:

$$2^0\cdot 3^1\cdot 5^2\cdot 7^0=75=\text{GCD}(1,500;4,725) \tag{5.4.95}$$

and the number obtained by taking the largest exponents is:

$$2^2\cdot 3^3\cdot 5^3\cdot 7^1=94,500=\text{LCM}(1,500;4,725) \tag{5.4.96}$$

Theorem 5.25. Let a and b be two natural numbers greater than or equal to 2, decomposed in the form:

$$a=p_1^{\alpha_1}p_2^{\alpha_2}\ldots p_k^{\alpha_k} \quad\text{and}\quad b=p_1^{\beta_1}p_2^{\beta_2}\ldots p_k^{\beta_k} \tag{5.4.97}$$

where $p_1,p_2,\ldots,p_k$ are prime numbers, $\alpha_1,\alpha_2,\ldots,\alpha_k$ and $\beta_1,\beta_2,\ldots,\beta_k$ natural numbers, possibly equal to zero.

For each value of i between 1 and k, we define:

$$\delta_i:=\min(\alpha_i,\beta_i) \quad\text{and}\quad \gamma_i:=\max(\alpha_i,\beta_i) \tag{5.4.98}$$

Therefore:

$$\text{GCD}(a;b)=p_1^{\delta_1}p_2^{\delta_2}\ldots p_k^{\delta_k} \quad\text{and}\quad \text{LCM}(a;b)=p_1^{\gamma_1}p_2^{\gamma_2\cdots}p_k^{\gamma k} \tag{5.4.99}$$

Proof 5.25.1. All divisors of a are of the form:

$$d=p_1^{\theta_1}p_2^{\theta_2}\ldots p_k^{\theta_k} \tag{5.4.100}$$

where $\theta_1,\theta_2,\ldots,\theta_k$ are natural numbers such that:

$$0\leq\theta_1\leq\alpha_1,0\leq\theta_2\leq\alpha_2,\ldots,0\leq\theta_k\leq\alpha_k \tag{5.4.101}$$

All divisors of b are of the form:

$$D=p_1^{\sigma_1}p_2^{\sigma_2}\ldots p_k^{\sigma_k} \tag{5.4.102}$$

where $\sigma_1, \sigma_2, \ldots, \sigma_k$ are natural numbers such that:

$$0 \leq \sigma_1 \leq \beta_1, 0 \leq \sigma_2 \leq \beta_2, \ldots, 0 \leq \sigma_k \leq \beta_k \tag{5.4.103}$$

The common divisors of a and b are then of the form:

$$p_1^{\pi_1} p_2^{\pi_2} \ldots p_k^{\pi_k} \tag{5.4.104}$$

where $\pi_1, \pi_2, \ldots, \pi_k$ are natural numbers such that:

$$0 \leq \pi_1 \leq \alpha_1, \quad 0 \leq \pi_2 \leq \alpha_2, \ldots, 0 \leq \pi_k \leq \alpha_k \quad \text{et} \quad 0 \leq \pi_1 \leq \beta_1, 0 \leq \pi_2 \leq \beta_2, \ldots, 0 \leq \pi_k \leq \beta_k \tag{5.4.105}$$

We therefore conclude that:

$$0 \leq \pi_1 \leq \delta_1, \quad 0 \leq \pi_2 \leq \delta_2, \ldots, 0 \leq \pi_k \leq \delta_k \tag{5.4.106}$$

The greatest common divisor (GCD) of a and b is obtained by definition for exponents equal to the largest possible values of π_i, that is, for exponents equal to δ_i. Therefore, we have:

$$\text{GCD}(a;b) = p_1^{\delta_1} p_2^{\delta_2} \ldots p_k^{\delta_k} \tag{5.4.107}$$

Now, the reader may remember that we demonstrated above that:

$$\text{GCD}(a;b) \cdot \text{LCM}(a;b) = a \cdot b \tag{5.4.108}$$

We then have:

$$\text{LCM}(a;b) = \frac{a \cdot b}{\text{GCD}(a;b)} = \frac{p_1^{\alpha_1} p_2^{\alpha_2} \ldots p_k^{\alpha_k} \cdot p_1^{\beta_1} p_2^{\beta_2} \ldots p_k^{\beta_k}}{p_1^{\delta_1} p_2^{\delta_2} \ldots p_k^{\delta_k}} = p_1^{\alpha_1+\beta_1-\delta_1} p_2^{\alpha_2+\beta_2-\delta_2} \ldots p_k^{\alpha_k+\beta_k-\delta_k} \tag{5.4.109}$$

On the other hand, we have:

$$\alpha_i + \beta_i = \min(\alpha_i, \beta_i) + \max(\alpha_i, \beta_i) = \delta_i + \gamma_i \tag{5.4.110}$$

Therefore:

$$\alpha_i + \beta_i - \delta_i = \gamma_i \tag{5.4.111}$$

We therefore conclude:

$$\text{LCM}(a;b) = p_1^{\gamma_1} p_2^{\gamma_2} \ldots p_k^{\gamma_k} \tag{5.4.112}$$

■ Q.E.D.

5.4.3.5 Congruences (modular arithmetic)

Modular arithmetic is a system of arithmetic for integers, where numbers "wrap around" upon reaching a certain value, named the "modulus" (plural moduli).

A familiar use of modular arithmetic is in the 12-hour clock (and also the calendar), in which the day is divided into two 12-hour periods. If the time is 7:00 now, then 8 hours later it will be 3:00. Usual addition would suggest that the later time should be $7+8=15$, but this is not the answer because clock time "wraps around" every 12 hours; in 12-hour time, there is no "15 o'clock". Likewise, if the clock starts at 12:00 (noon) and 21 hours elapse, then the time will be $9:00$ the next day, rather than 33:00. Because the hour number starts over after it reaches 12, this is arithmetic modulo 12. According to the definition below, 12 is congruent not only to 12 itself, but also to 0, so the time "12:00" could also be written "0:00", since 12 is congruent to 0 modulo 12.

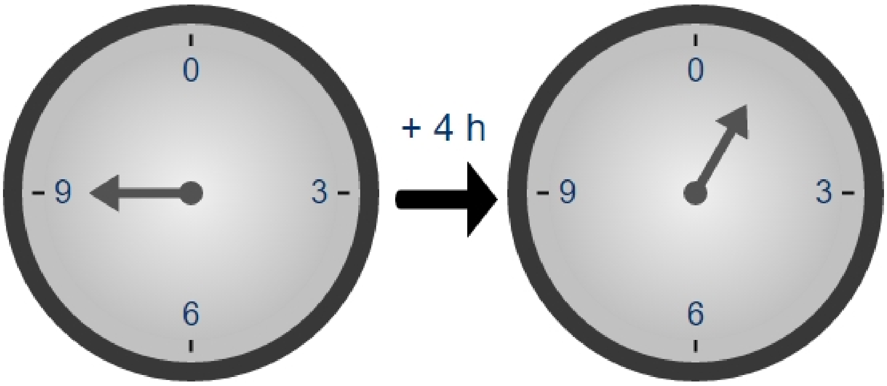


Figure 5.49: Time-keeping on a clock using modulo 12 arithmetics (source: Wikipedia)

Definition (#5.4.68): Let $m \in \mathbb{Z} \setminus 0$. If a and b have the same remainder when divided by m in the Euclidean division then we say "a is congruent to b modulo m", and we write:

$$\boxed{a \equiv b \mod (m)} \tag{5.4.113}$$

or equivalently there is (at least) one relative integer k such that:

$$\boxed{a = b + km} \tag{5.4.114}$$

Often denoted:

$$x = qm + r \tag{5.4.115}$$

where q is the "quotient" and r the "remainder" of the euclidean division.

✍**Example:**

We have:

$$51 \equiv 1 \mod (10)$$

as (we apply $a=b+km$):

$$51 = 1 + 5 \cdot 10$$

And we have:

$$-51 \equiv 9 \mod (10)$$

as (we apply again $a = b + km$):

$$-51 = 9 + (-6) \cdot 10$$

We also name the number b "residue". Thus, a residue is an integer congruent to another, modulo a given integer m. The reader can verify that this requires that:

$$m|(a-b) \tag{5.4.116}$$

The case with negative number is many times difficult to grasp. So let us do a step by step in-deep companion example!

Suppose we try to calculate $-3 \div 12$. A possible answer is:

$$-3 = -1 \cdot 12 + 9$$

In other words, that $-3 \mod 12 = 9$. But can't we represent this as:

$$-3 = -2 \cdot 12 + 21$$

It's true that $-3 \equiv 21 \mod 12$. And also that $-3 \equiv 117 \mod 12$, and of course $-3 \equiv 9 \mod 12$ (or alternatively $-3 \mod 12 = 9$). All of these numbers are in the same "congruence or residue class". Typically the smallest non-negative value is used to represent the class by tradition such that $0 \geq b < m$.

So we could write this congruence class as $\{\ldots, -15, -3, 9, 21, 33, \ldots\}$ but that would become tedious so by convention we represent the congruence class by just one of its members, often the smallest non-negative member. So we represent $\{\ldots, -15, -3, 9, 21, 33, \ldots\}$ by $9 \mod 12$.

✍**Example:**

All spreadsheets and business intelligence softwares have a modulo function often denoted `MOD(value,congruence)` (that returns the b of $a = b + km$) and used for various reasons. But the most used and well known practical case is the calculation of the fiscal month, given the civil month (for example the civil month starts obviously in January (1) but the fiscal month may start for a corporation in April (4). So how can we implement a formula that give the right column in the table below (in the practice you can't generate the column by writing the fiscal months manually because most corporations work on a daily basis and have thousands of days for which they want to know what is the fiscal month!):

	Civil Month (CM)	**Fiscal Month** (FM)
JAN	1	$(-3 \mod 12)+1=9+1=10$
FEB	2	$(-2 \mod 12)+1=10+1=11$
MAR	3	$(-1 \mod 12)+1=11+1=12$
APR	4	$(0 \mod 12)+1=0+1=1$
MAY	5	$(1 \mod 12)+1=1+1=2$
JUN	6	$(2 \mod 12)+1=2+1=3$
JUL	7	$(3 \mod 12)+1=3+1=4$
AUG	8	$(4 \mod 12)+1=4+1=5$
SEP	9	$(5 \mod 12)+1=5+1=6$
OCT	10	$(6 \mod 12)+1=6+1=7$
NOV	11	$(7 \mod 12)+1=7+1=8$
DEC	12	$(8 \mod 12)+1=8+1=9$

In the special case above where the fiscal month is April (4), in all spreadsheet softwares, business intelligence softwares, engineering softwares and programming languages, the formula would be:

```
Fiscal Month = MOD(Civil Month-4,12)+1
```

Or written in more general way:

```
MOD(Civil Month-Fiscal Month Start,12)+1
```

This is equivalent to:

$$a=b+km \Leftrightarrow \texttt{b = MOD(a,m)} \Rightarrow \texttt{FM=MOD(CM-FMS,12)+1}$$

Remarks

R1. The reader must well understand that congruence implies a null remainder for the division!

R2. We exclude in addition to the 0 also the 1 and $-1$for the possible values of m in the definition of congruence in some books.

R3. Behind the term congruence are hidden similar concepts of different levels of abstraction:

- In modular arithmetic, so we say that "two integers a and b are congruent modulo m if they have the same remaining in the Euclidean division by m". We can also say that they are congruent modulo m if their difference is a multiple of m.

- In the study of oriented angles, we say that "two measurements are congruent modulo 2π [rad] if and only if their difference is a multiple of 2π [rad]". This characterize two measures of the same angle (see section Trigonometry page 2513).

- In algebra, we speak of congruence modulo I in a commutative ring (see section Set Theory page 426) for which I is an ideal: "x is congruent to y modulo I

if and only if their difference belongs to I". This congruence is an equivalence relation compatible with the operations of addition and multiplication and gives the possibility to define a quotient ring of the parent set with its ideal I.

- We sometimes see in the study of geometry (see section Euclidean Geometry page 2633) the term "congruence" used in place of "similar". It is then a simple equivalence relation on the set of plane figures.

The relation of congruence $\equiv$ is an equivalence relation (see section Operators page 290), in other words, given $a, b, c, m \in \mathbb{Z}, m > 1$ then the congruence relation is:

P1. Reflexive:

$$a \equiv a \mod (m) \tag{5.4.117}$$

P2. Symmetric:

$$a \equiv b \mod (m) \Leftrightarrow b \equiv b \mod (m) \tag{5.4.118}$$

P3. Transitive:

$$a \equiv b \mod (m), b \equiv c \mod (m) \Rightarrow a \equiv b \mod (m) \tag{5.4.119}$$

The properties P1 and P2 are obvious (if this is not the case please let us know we will develop!). We will prove only P3.

Proof 5.25.2. The assumptions imply that:

$$b = a + km \qquad c = b + lm \tag{5.4.120}$$

But then:

$$c = b + lm = (a + km) + lm = a + (k + l)m \tag{5.4.121}$$

This proves that a and c are congruent modulo m.

■ Q.E.D.

The relation of congruence $\equiv$ is compatible with the sum and the product (remember that power is ultimately an extension of the product!).

Indeed, given $(a, b, a', b', m) \in \mathbb{Z}, m > 1$ such that $a \equiv \mod (m)$ and $a' \equiv b'$ then:

P1. $a + a' \equiv b + b' \mod (m)$

P2. $aa' \equiv bb' \mod (m)$

Proof 5.25.3. We have:

$$a = b + km \qquad a' = b' + lm \tag{5.4.122}$$

by hypothesis. But then:

$$a + a' = b + b' + (l + k)m \tag{5.4.123}$$

which proves P1. We also have:

$$aa' = bb' + blm + b'km + klm'2 = bb' + sm \tag{5.4.124}$$

which proves P2.

■ Q.E.D.

Remark

The congruence relation behaves in many point like the relation of equality. However a property of the relation of equality $=$ is not true for that of congruence $\equiv$, namely the simplification: If $ab \equiv \mod (m)$, we do not have necessarily $b \equiv c \mod (m)$.

✍**Examples:**

E1. Consider the following simple case:

$$2 \cdot 1 \equiv 2 \cdot 3 \mod (4) \quad \text{but} \quad 1 \not\equiv 3 \mod (4) \tag{5.4.125}$$

E2. Gangsters have just stolen a million dollars, in denominations of $1, 10, 100$ and $1,000$ dollars. One of them claims that there are exactly $92,000$ in bills. Is this statement correct?

According to the gangster's statement, the stolen amount is written as:

$$1'000'000 = a + 10b + 100c + 1'000d$$

where:

$$a + b + c + d = 92,000$$

Modulo 9 (this is a good choice!), this gives:

$$1 \equiv a + b + c + d$$

hence:

$$1 \equiv 2$$

Therefore we have the proof of the gangster's dishonesty (but that was something we suspected).

So far we have seen the properties of congruences involving a single modulus. We will now study the behaviour of the congruence relation on a change of modulus:

P1. If $a \equiv b \mod (m)$ and $d|m$, then $a \equiv b \mod (d)$

P2. If $a \equiv b$ and $a \equiv b \mod (s)$ then a and b are congruate modulus $[r, s]$

We think this two properties are obvious. We do not need to go into details for P1. For P2, since $b - a$ is a multiple of r and s since by hypothesis:

$$\frac{b-a}{r} = k, \frac{b-a}{s} = l \Rightarrow b - a = rk = sl \qquad (5.4.126)$$

$b - a$ is then a multiple of the LCM of r and s, which proves P2.

From these properties it follows that if we denote by $f(x)$ a polynomial with integer coefficients (positive or negative):

$$f(x) = Ax^n + Bx^{n-1} + Cx^{n-2} + \ldots + Kx + L \qquad (5.4.127)$$

The congruence $a \equiv b \mod (m)$ will also give $f(a) \equiv f(b) \mod (m)$.

If we replace x successively by all integers in a polynomial $f(x)$ with integer coefficients, and if we take the remaining modulus m, these remaining are reproduced from m to m (in the sense where the congruence is satisfied), since we have, regardless of the number m and x:

$$f(x) \equiv f(x+m) \mod (m) \qquad (5.4.128)$$

We then deduce then the impossibility of solving the following congruence:

$$f(x) \equiv r \mod (m) \qquad (5.4.129)$$

with integer numbers, if r is anyone of the "non-remaining" (a residue that does not satisfy the congruence).

5.4.3.5.1 Congruence Class

Definition (#5.4.69): We name "modulo m congruence class", the subset of the set $\mathbb{Z}$ defined by the property that two elements a and b of $\mathbb{Z}$ are in the same class if and only if $a \equiv b \mod (m)$ or that a set of elements are congruent by this same modulo.

Remark

We saw in the section Operators that this is in fact an equivalence class (page 292) as the congruence modulo m is, as we have proved above, a relation of equivalence!!!

✍**Example:**

Given $m = 3$. We divide the set of integers into congruence classes of modulo 3. Here are for example three sets whose elements are congruent with one another without rest (see well what gives the set of all these classes together!):

$$\begin{gathered}\{\ldots,-9,-6-3,0,3,6,9,12,\ldots\}\\ \{\ldots,-8,-5,-2,1,4,7,10,13,\ldots\}\\ \{\ldots,-7,-4,-1,2,5,8,11,\ldots\}\end{gathered} \tag{5.4.130}$$

Thus we see that for each pair of elements of a congruence class, the congruence modulo 3 exists. However, we see that we can not take that $-9 \equiv -8 \mod (3)$ where -9 is in the first class and -8 in the second.

The smallest non-negative number of the first class is 0, this of the second is 1 and the last is 2. Thus, we will denote these three classes respectively $[0]_3$, $[1]_3$, $[2]_3$, the number 3 in the index indicating the modulus.

It is interesting to notice that if we take any number of the first class and any number of the second class, then their sum is always in the second class. This can be generalized and allows to define a sum of classes modulo 3 by writing:

$$\begin{gathered}[0]_3+[0]_3=[0]_3;\quad [0]_3+[1]_3=[1]_3\\ [0]_3+[2]_3=[2]_3;\quad [1]_3+[1]_3=[2]_3\\ [1]_3+[2]_3=[0]_3;\quad [2]_3+[2]_3=[1]_3\end{gathered} \tag{5.4.131}$$

and also:

$$\begin{gathered}[0]_3\times[0]_3=[0]_3;\quad [0]_3\times[1]_3=[0]_3\\ [0]_3\times[2]_3=[0]_3;\quad [1]_3\times[1]_3=[1]_3\\ [1]_3\times[2]_3=[2]_3;\quad [2]_3\times[2]_3=[1]_3\end{gathered} \tag{5.4.132}$$

Thus, for any $m > 1$, the congruence class:

$$a \mod (m) \tag{5.4.133}$$

is the set of integers congruent to a modulo m (and congruent modulo m between them)!!! This class is denoted by:

$$\boxed{[a]_m := a \mod (m)} \tag{5.4.134}$$

Remark

Having bracketed the expression "and congruent modulo m between them" is due to the fact that the congruence, being an equivalence relation we have as we have proved above that $b \equiv a \mod (m)$, $c \equiv \mod (m)$, then $b \equiv a \mod (m)$.

Definition (#5.4.70): The set of congruence classes $[a]_m$ (that forms by the fact that congruence is an equivalence relation: "equivalence classes"), for a fixed m gives what we name a "quotient set" (see section Operators page 292). More rigorously, we speak of the "quotient set of $\mathbb{Z}$ by the congruence relation" whose elements are the congruence classes (or: equivalence classes) and then form the ring $\mathbb{Z}/m\mathbb{Z}$.

We deduce from the definition the following two trivial properties:

1. The number b is in the class $[a]_m$ if and only if $a \equiv b \mod (m)$
2. The classes $[a]_m$ and $[b]_m$ are equal if and only if $a \equiv b \mod (m)$

Theorem 5.26. There are exactly m different congruence classes of modulo m, i.e. $[0]_m, [1]_m, \ldots, [m-1]_m$.

Proof 5.26.1. Given $m > 1$, than any integer a is congruent modulo m to one and only one integer r of the set $\{0, 1, 2, \ldots, m-1\}$ (notice well, it is important, that we restrict ourselves to the positive integers without taking into account the negative one!). In addition, this integer r is exactly the remaining of the division of a by m. In other words, if $0 \leq r < m$, then:

$$a \equiv r \mod (m) \tag{5.4.135}$$

if and only if $a = qm + r$ where q is the quotient of a by m and r is the remainder. The proof is an immediate consequence of the definition of the congruence and of the Euclidean division.

■ Q.E.D.

Definition (#5.4.71): An integer b in a congruence class modulo m is named a "representative of this class" (it is clear that by the equivalence relation that two representative of the same class are congruent modulo m to each other).

We can now be able to build an addition and a multiplication on the congruence classes. To define the sum of two classes $[a]_m, [b]_m$, it suffices to take one representative from each class, to their sum and take the congruence class of the result. Thus (see examples above):

$$[a]_m + [b]_m = [a + b]_m \tag{5.4.136}$$

And same for the multiplication:

$$[a]_m \cdot [b]_m = [a \cdot b]_m \tag{5.4.137}$$

By construction of the addition and multiplication, we see that 0 (zero) is the neutral element for addition:

$$[a]_m + [0]_m = [a]_m, \quad \forall a \in \mathbb{Z}, \forall m \in \mathbb{N} \tag{5.4.138}$$

and the class of the integer 1 is the neutral element for multiplication:

$$[a]_m \cdot [1]_m = [a]_m, \quad \forall a \in \mathbb{Z}, \forall m \in \mathbb{N} \tag{5.4.139}$$

Definition (#5.4.72): An element $[a]_m$ of $\mathbb{Z}/m\mathbb{Z}$ is "one unit" if there is an element $[b]_m \in \mathbb{Z}/m\mathbb{Z}$ such that $[a]_m \cdot [b]_m$. The following theorem helps to characterize classes modulo m which are units in $\mathbb{Z}/m/\mathbb{Z}$:

Theorem 5.27. Given $[a]$ an element of $\mathbb{Z}/m/\mathbb{Z}$. Then $[a]$ is a unit if and only if $(a, m) = 1$.

Proof 5.27.1. Suppose first that $(a, m) = 1$. Then by Bézout's theorem, we have its identity:

$$as + mr = 1 \tag{5.4.140}$$

In other words, as is congruent to 1 modulo m. But this is equivalent to write by definition that $[a][s] = 1$ showing that $[a]$ is a unit. Conversely, if $[a]$ is a unit, this implies that there exists a class $[s]$ such that $[a][s] = 1$.

Thus, we have just proved that $\mathbb{Z}/\mathbb{Z}$ is indeed a ring since it has an addition, a multiplication, a neutral element and an inverse!!

■ Q.E.D.

5.4.3.5.2 Complete set of residues

Definition (#5.4.73): A set of numbers $a_0, a_1, \ldots, a_{(m-1)} \mod (m)$ form a "complete set of residues", also named a "covering system", if they satisfy $a_i = i \mod (m)$ for $i = 0, 1, \ldots, m-1$.

This type of systems will help us to introduce in the section of Cryptography (page 7216) an important function used to secured communication devices at the end of the 120th century (holocene calendar) and at beginning of the 121st century.

To introduce this concept, consider the following finite system of congruences modulo 6:

$$\begin{aligned} 6 &\equiv 0 \mod (6)\\ 13 &\equiv 1 \mod (6)\\ 2 &\equiv 2 \mod (6)\\ -3 &\equiv 3 \mod (6)\\ 22 &\equiv 4 \mod (6)\\ 11 &\equiv 5 \mod (6) \end{aligned} \tag{5.4.141}$$

where as the reader will have probably noticed it: no residue is repeated in the list and no residue taken in pairs are congruent between them modulo m (is this last point that oblige to stop at 5 in our example). We then say that the residues are "mutually incongruent".

If these conditions are met, then we say that the ordered set $\{6, 13, 2, -3, 22, 11\}$ is a "complete system of residues modulo m" as already defined. Such a set is not unique for a given module. Thus, the set $\{0, 1, 2, 3, 4, 5\}$ is also a complete (trivial) system of residues modulo 6.

If we eliminate from this entire system all numbers that are not prime to m, then we have a "system of reduced residue modulo m". So in the above example, the reduced residue system modulo 6 will be $\{13, 11\}$.

Reduces systems will be useful to us in the section Cryptography (page 7216) to prove an important result in the asymmetric public key systems.

We will see also in the section Cryptography (page 7236), the "Euler indicator function" when m is prime (which is not the case in the previous example) gives the cardinal of the reduced system modulus m as being equal to:

$$\boxed{\phi(m) = m - 1 = \text{Card}\left(\{n \leq m, (n, m) = 1\}\right)} \tag{5.4.142}$$

So under the assumption condition that m is prime, the reduced system of residue is obviously written:

$$\{r_1, r_2, \ldots, r_\phi(m)\} \tag{5.4.143}$$

5.4.3.5.3 Chinese remainder theorem

In its basic form, the Chinese remainder theorem will determine a number n that, when divided by some given divisors, leaves given remainders. In Sun Tzu's example (stated in modern terminology), what is the smallest number n that when divided by 3 leaves a remainder of 2, when divided by 5 leaves a remainder of 3, and when divided by 7 leaves a remainder of 2?

The Chinese remainder theorem can therefore be seen as solving a linear system but in a modular system. For many students and future engineers, this theorem will never be used in practice, but some will see it again it in the field of cryptography (in the context of decryption especially).

There are several possible proofs as always but we opted for the one that, as always for us, seemed the most educational.

Given m and n both prime integers between them. Then special case of a system of two congruences (see further below for an example of resolution of a system of three congruences):

$$\begin{aligned} x &\equiv a \mod (m)\\ x &\equiv b \mod (m) \end{aligned} \tag{5.4.144}$$

has a unique solution.

Proof 5.27.2. As m and n are assumed as prime between them, there exists then u and v two integers such as (application of the Bézout identity proved earlier above):

$$um+vn=1 \tag{5.4.145}$$

Therefore we have:

$$aum+avn=a \tag{5.4.146}$$

That is to say:

$$avn\equiv a \mod (m) \tag{5.4.147}$$

Then we have also by extension:

$$bum+bvn=b \tag{5.4.148}$$

That is to say:

$$bum\equiv b \mod (n) \tag{5.4.149}$$

So to be clear, we have so far:

$$um+vn=1 \Rightarrow \begin{cases} avn\equiv a \mod (m)\\ bum\equiv b \mod (n) \end{cases} \tag{5.4.150}$$

We then have for recall:

$$bum\equiv b \mod (n) \Rightarrow bum+bvn=b \tag{5.4.151}$$

But we can also write with $k \in \mathbb{Z}$:

$$bum + \underbrace{(k + av)}_{=bv} n = b \tag{5.4.152}$$

Therefore:

$$bum + avn \equiv b \mod (n) \tag{5.4.153}$$

Then we also have:

$$avn \equiv a \mod (m) \Rightarrow\Rightarrow aum + avn = a \tag{5.4.154}$$

But we can also write with $k \in \mathbb{Z}$:

$$\underbrace{(K + bu)}_{=au} m + avn = a \tag{5.4.155}$$

Therefore:

$$bum + avn \equiv a \mod (m) \tag{5.4.156}$$

So to be always clear, we have so far:

$$\begin{aligned} um + vn = 1 = (m, n) &\Rightarrow \begin{cases} avn \equiv a \mod (m) \\ bum \equiv b \mod (n) \end{cases} \\ &\Rightarrow \begin{cases} bum + avn \equiv a \mod (m) \\ bum + avn \equiv b \mod (n) \end{cases} \end{aligned} \tag{5.4.157}$$

So finally we get that:

$$x = bum + avn \tag{5.4.158}$$

is a particular solution of the system. But we also have $\forall i, j \in \mathbb{Z}$ by the definition of the congruence:

$$\begin{aligned} bum + avn + im \equiv a \mod (m) \\ bum + avn + jn \equiv b \mod (n) \end{aligned} \tag{5.4.159}$$

So that x is always solution of the system, we must have:

$$\begin{aligned} i = n \\ j = m \end{aligned} \tag{5.4.160}$$

and therefore:

$$\begin{aligned} bum + avn + nm \equiv a \mod (m) \\ bum + avn + mn \equiv b \mod (n) \end{aligned} \tag{5.4.161}$$

Therefore a little bit more general solution is:

$$x = bum + avn + nm \tag{5.4.162}$$

But by extension, we have the general solution:

$$x = bum + avn + znm \tag{5.4.163}$$

with $z \in \mathbb{Z}$. We then say sometimes that the solution is "x modulo nm".

■ Q.E.D.

✍**Example:**

As an example, consider the problem of finding an integer x such that:

$$\begin{aligned} x &\equiv 2 \mod (3) \\ x &\equiv 3 \mod (4) \\ x &\equiv 1 \mod (3) \end{aligned} \tag{5.4.164}$$

A brute-force approach converts these congruences into sets and writes the elements out to the product of $3 \cdot 4 \cdot 5 = 60$ (the solutions modulo 60 for each congruence):

$$\begin{aligned} &x \in \{2, 5, 8, \mathbf{11}, 14, 17, 20, 23, 26, 29, 32, 35, 38, 41, 44, 47, 50, 53, 56, 59\} \\ &x \in \{3, 7, \mathbf{11}, 15, 19, 23, 27, 31, 35, 39, 43, 47, 51, 55, 59\} \\ &x \in \{1, 6, \mathbf{11}, 16, 21, 26, 31, 36, 41, 46, 51, 56\} \end{aligned} \tag{5.4.165}$$

To find an x that satisfies all three congruences, intersect the three sets to get:

$$x = 11 \tag{5.4.166}$$

This solution is modulo 60, hence all solutions are expressed as:

$$x \equiv 11 \mod (60) \tag{5.4.167}$$

Another way to find a solution is with basic algebra, modular arithmetic, and stepwise substitution.

We start by translating these congruences into equations for some t, s, and u:

$$\begin{aligned} x &= 2 + 3t \\ x &= 3 + 4s \\ x &= 1 + 5u \end{aligned} \tag{5.4.168}$$

After we substitute the x from the first equation into the second congruence:

$$2 + 3t \equiv 3 \mod (4) \tag{5.4.169}$$

That is to say:

$$3t \equiv 1 \mod (4) \tag{5.4.170}$$

Hence:

$$t \equiv 3 \mod (4) \tag{5.4.171}$$

meaning that $t = 3 + 4s$ for some integer s. Substitute now t into the first equation:

$$x = 2 + 3t = 2 + 3(3 + 4s) = 11 + 12s \tag{5.4.172}$$

Substitute this x into the third congruence:

$$11+12s \equiv 1 \mod (5) \tag{5.4.173}$$

That is to say:

$$1+1s \equiv 1 \mod (5) \tag{5.4.174}$$

Hence:

$$2s \equiv 0 \mod (5) \tag{5.4.175}$$

meaning that $s=0+5u$ for some integer u. Finally:

$$x=11+12s=11+12(5u)=11+60u \tag{5.4.176}$$

So, we have solution $\{11,71,131,191,\ldots\}$.

5.4.3.6 Continued fractions

Definition (#5.4.74): A "continued fraction" is an expression obtained through an iterative process of representing a number as the sum of its integer part and the reciprocal of another number, then writing this other number as the sum of its integer part and another reciprocal, and so on. In a finite continued fraction (or terminated continued fraction), the iteration/recursion is terminated after finitely many steps by using an integer in lieu of another continued fraction. In contrast, an infinite continued fraction is an infinite expression. In either case, all integers in the sequence, other than the first, must be positive. The integers a_i are named the "coefficients" or "terms" of the continued fraction.

The notion of continued fraction come back from the time of Fermat and culminated with the work of Lagrange and Legendre in the late 118th century (holocene calendar). These fractions are important in physics because we find them back in our study of acoustic and also in the thought process that led Galois to develop his group theory and also in the studies of gear ratios (for watch complications as discussed in the section of Mechanical Engineering page 8368). Continued fractions will also be very useful to us in the section of Astronomy to determine the saros cycles of lunar and solar eclipses (see page 5486).

To understand the motivation of continued fraction let us introduce a basic example.

Consider a typical rational number:

$$\frac{415}{93} \tag{5.4.177}$$

which is around 4.4624.

As first approximation, start with 4, which is the integer part:

$$\frac{415}{93}=4+\frac{43}{93} \tag{5.4.178}$$

Note that the fractional part is the reciprocal of $93/43$ which is about 2.1628. Use the integer part, 2, as an approximation for the reciprocal, to get a second approximation of:

$$\frac{93}{43}=2+\frac{7}{43} \tag{5.4.179}$$

So we have so far:

$$\frac{415}{93}=4+\cfrac{1}{2+\cfrac{7}{43}} \tag{5.4.180}$$

Note that the fractional part is the reciprocal of $43/7$ which is about 6.1429. Use the integer part, 6, as an approximation for the reciprocal, to get a second approximation of:

$$\frac{43}{7}=6+\frac{1}{7} \tag{5.4.181}$$

Therefore:

$$\frac{415}{93}=4+\cfrac{1}{2+\cfrac{1}{6+\cfrac{1}{7}}} \tag{5.4.182}$$

Note that the fractional part $1/7$ is the reciprocal of 7 which is about... 7 Use the integer part, 7, as an approximation for the reciprocal, to get a second approximation of:

$$\frac{7}{1}=7+\frac{0}{1} \tag{5.4.183}$$

Therefore we get:

$$\frac{415}{93}=4+\cfrac{1}{2+\cfrac{1}{6+\cfrac{1}{7+\cfrac{0}{1}}}}=4+\cfrac{1}{2+\cfrac{1}{6+\cfrac{1}{7}}} \tag{5.4.184}$$

This expression is named as we know the "continued fraction representation of the number".

Dropping some of the less essential parts of the expression:

$$\frac{415}{93}=4+\cfrac{1}{2+\cfrac{1}{6+\cfrac{1}{7}}} \tag{5.4.185}$$

gives the abbreviated notation:

$$\frac{415}{93}=[4;2,6,7] \tag{5.4.186}$$

Note that it is customary to replace only the first comma by a semicolon.

As generalization of the previous example let us consider in a first time the rational number a/b with $(a,b)=1$ with $b>0$ and $a>b$. We know that all the quotients q_i and the remaining r_i are within the scope of the Euclidean division positive integers.

Let us recall that the Euclidean algorithm already seen earlier (but written in a slightly different way):

$$\begin{aligned}\frac{a}{b}&=q_1+\frac{r_1}{b}\\\frac{b}{r_1}&=q_2+\frac{r_2}{r_1}\\\frac{r_1}{r_2}&=q_3+\frac{r_3}{r_2}\\&\ldots\\\frac{r_{n-2}}{r_{n-1}}&=q_n\end{aligned}\tag{5.4.187}$$

By successive substitutions, we get:

$$\begin{aligned}\frac{a}{b}&=q_1+\frac{1}{b/r_1}=q_1+\cfrac{1}{q_2+\cfrac{1}{r_1/r_2}}\\&=\ldots\\&=q_1+\cfrac{1}{q_2+\cfrac{1}{q_3+\cfrac{1}{\ddots\cfrac{}{\cfrac{1}{q_{n_1}+\cfrac{1}{q_n}}}}}}\end{aligned}\tag{5.4.188}$$

What is also sometimes written:

$$\frac{a}{b}=q_1+\frac{1}{1/\varepsilon_1}=q_1+\cfrac{1}{q_2+\cfrac{1}{\varepsilon_2}=\ldots}\tag{5.4.189}$$

So any positive rational number can be expressed as a finite continued fraction where $q_n\in\mathbb{N}$.

Taking our introducing example:

$$\frac{415}{93}=4+\cfrac{1}{2+\cfrac{1}{6+\cfrac{1}{7+\cfrac{0}{1}}}}=4+\cfrac{1}{2+\cfrac{1}{6+\cfrac{1}{7}}}\tag{5.4.190}$$

we notice indeed that $q_n\in\mathbb{N}$ and that we have by construction:

$$q_n=\left[\frac{1}{\varepsilon_{n-1}}\right]\tag{5.4.191}$$

where the brackets represent the integer part and that we also have:

$$\frac{1}{\varepsilon_{n_1}}=q_n+\varepsilon\tag{5.4.192}$$

The development of the number a/b is named the "development of the number a/b in finite continued fraction" and is condensed in the following notation:

$$\boxed{[q_1; q_2, q_3, \ldots, q_n]} \tag{5.4.193}$$

Let us see now another example:

✍**Example:**

Let us see how the extract the square root of a number A (for example $A = 2$ such that we want to extract $\sqrt{2}$) by the continued fraction method.

Given a the largest integer whose square a^2 is smaller than A. We subtract it to A. So there is a remaining of (for $A = 2$, we have $a = 1$):

$$r = A - a^2 = (\sqrt{A} - a)(\sqrt{A} + a) \tag{5.4.194}$$

where we have used a remarkable identities that we will prove in the section Calculus later (page 1361). Hence dividing both members by the second parenthesis, we have:

$$\sqrt{A} - a = \frac{r}{\sqrt{A} + a} \tag{5.4.195}$$

Therefore:

$$\sqrt{A} = a + \frac{r}{\sqrt{A} + a} \tag{5.4.196}$$

In the denominator, we replace $\sqrt{A}$ by:

$$a + \frac{r}{\sqrt{A} + a} \tag{5.4.197}$$

That gives:

$$\sqrt{A} = a + \cfrac{r}{2a + \cfrac{r}{\sqrt{A} + a}} \tag{5.4.198}$$

etc... we thus see that the system is simple to determine the expression of a root square in terms of continued fraction.

We consider now as intuitive that every rational number can be expressed as finite continued fraction and conversely that any finite continued fraction represents a rational number. By extension, an irrational number is represented by an infinite continued fraction!

Now consider $[q_1; q_2, q_3, \ldots, q_n]$ a finite continued fraction. The continued fraction:

$$C_k = [q_1; q_2, \ldots, q_n] \tag{5.4.199}$$

where $k = 1, 2, \ldots, n$ is named the "k-th reduced" or "k-th convergent" or the "k-th partial quotient".

With this notation, we have:

$$\begin{aligned} C_1 &= [q_1] = q_1 \\ C_2 &= [q_1; q_2] = q_1 + \frac{1}{q_2} = \frac{q_1 q_2 + 1}{q_2} \\ C_3 &= [q_1; q_2, q_3] = q_1 + \cfrac{1}{q_2 + \cfrac{1}{q_3}} = \frac{(q_1 q_2 + 1) \cdot q_3 + q_1}{q_2 q_3 + 1} \\ C_4 &= [q_1; q_2, q_3, q_4] = \frac{((q_1 q_2 + 1) \cdot q_3 + q_1) \cdot q_4 + (q_1 q_2 + 1)}{(q_2 q_3 + 1) \cdot q_4 + q_2} \end{aligned} \tag{5.4.200}$$

To simplify the expressions above, we introduce the sequences $\{n_i\}, \{d_i\}$ (n is for **n**umerator and d for **d**enominator) defined by:

$$\begin{aligned} n_0 &= 1 \qquad n_1 = q_1, \ldots \qquad n_i = q_i n_{i-1} + n_{i-2} \\ d_0 &= 0 \qquad d_1 = 1, \ldots \qquad d_i = q_i d_{i-1} + d_{i-2} \end{aligned} \tag{5.4.201}$$

Thanks to this construction, we have an interesting immediate little inequality that will be useful to us further below:

$$0 = d_0 < d_1 < d_2 < d_3 < \ldots \tag{5.4.202}$$

With the above definition, we find that:

$$C_1 = \frac{n_1}{d_1}, \quad C_2 = \frac{n_2}{d_2}, \quad C_3 = \frac{n_3}{d_3}, \quad \ldots, \quad C_i = \frac{n_i}{d_i} \tag{5.4.203}$$

Either by generalizing:

$$C_k = [q_1, q_2, q_3, \ldots, q_k] = \frac{n_k}{d_k} \tag{5.4.204}$$

Now let us show for later use that for $i \geq 1$, we have:

$$n_i d_{i-1} - d_i n_{i-1} = (-1)^i \tag{5.4.205}$$

The result is immediate for $i = 1$. Assuming that the result is true for i let us show that it is true for $i + 1$. Since:

$$n_{i+1} d_i - d_{i+1} n_i = (q_{i+1} n_i + n_{i-1}) d_i - (q_{i+1} d_i + d_{i-1}) n_i = -(d_{i-1} n_i - n_{i-1} d_i) = (-1)^{i+1} \tag{5.4.206}$$

then using the induction hypothesis, we get the result!

We can now establish a vital relation for what will follow.

Theorem 5.28. Let us prove that if C_k is the k-th reduced to of the simple finite continued fraction $[q_1; q_2, \ldots, q_n]$ then:

$$C_1 < C_3 < C_5 < \ldots < C_6 < C_4 < C_2 \tag{5.4.207}$$

Proof 5.28.1.

$$\begin{aligned}C_{k+2}-C_k&=(C_{k+2}-C_{k+1})+(C_{k+1}-C_k)=\left(\frac{n_{k+2}}{d_{k+2}}-\frac{n_{k+1}}{d_{k+1}}\right)+\left(\frac{n_{k+1}}{d_{k+1}}-\frac{n_k}{d_k}\right)\\&=\left(\frac{n_{k+2}d_{k+1}-n_{k+1}d_{k+2}}{d_{k+2}d_{k+1}}\right)+\left(\frac{n_{k+1}d_k-n_kd_{k+1}}{d_{k+1}d_k}\right)=\frac{(-1)^{k+2}}{d_{k+2}d_{k+1}}+\frac{(-1)^{k+1}}{d_{k+1}d_k}\\&=\frac{(-1)^{k+2}d_k+(-1)^{k+1}d_{k+2}}{d_{k+2}d_{k+1}d_k}=\frac{(-1)^{k+1}(-1)d_k+(-1)^{k+1}d_{k+2}}{d_{k+2}d_{k+1}d_k}\\&=\frac{-(-1)^{k+1}d_k+(-1)^{k+1}d_{k+2}}{d_{k+2}d_{k+1}d_k}=\frac{(-1)^{k+1}(d_{k+2}-d_k)}{d_{k+2}d_{k+1}d_k}\end{aligned} \tag{5.4.208}$$

as:

$$0=d_0<d_1<d_2<d_3<\ldots \tag{5.4.209}$$

therefore:

$$d_{k+2}-d_k>0 \tag{5.4.210}$$

indicating to us that the sign of $C_{k+1}-C_k$ is the same as $(-1)^{k+1}$.

It follows that $C_{k+2}>C_k$ for an odd k, and $C_{k+2}<C_k$ for k even. Then:

$$C_1<C_3<C_5<\ldots \quad \text{and} \quad C_2>C_4>C_6>\ldots \tag{5.4.211}$$

and after as:

$$C_k-C_{k-1}=\frac{n_k}{d_k}-\frac{n_{k-1}}{d_{k-1}}=\frac{n_kd_{k-1}-n_{k-1}d_k}{d_kd_{k-1}}=\frac{(-1)^k}{d_kd_{k-1}} \tag{5.4.212}$$

So for k even, we have $C_k>C_{k-1}$, we therefore deduce that:

$$C_1<C_3<C_5<\ldots<C_6<C_4<C_2 \tag{5.4.213}$$

■ Q.E.D.

Let us show now that every infinite continued fraction can be any irrational number.

In formal terms, if $\{q_n\}$ is a sequence of positive integers and that we consider $C_n=[q_1;q_2,\ldots,q_n]$ then it necessarily converges to a real number if $n\rightarrow+\infty$.

Actually it is not difficult to observe (it's quite intuitive) with a practical example that we have:

$$C_k-C_{k-1}\rightarrow 0 \tag{5.4.214}$$

when $k\rightarrow+\infty$.

Now, let us denote by x any real number and $q_1=[x]$ the integer part of this real number. Then we saw at the beginning of our study of continued fractions that:

$$x=q_1+\varepsilon_1 \tag{5.4.215}$$

Therefore it comes:

$$\varepsilon_1 = x - q_1 \tag{5.4.216}$$

Let's look for the needs of the section on Acoustics (page 8195) on the calculation of a continued fraction of a logarithm using the previous relation!

First let us recall that:

$$\frac{a}{b} = q_1 + \frac{1}{1/\varepsilon_1} = q_1 + \frac{1}{q_2 + 1/\varepsilon_2} = \ldots \tag{5.4.217}$$

That is (relation proved in the section of Functional Analysis page 2284):

$$x = \log_a(u) = \frac{\ln(u)}{\ln(a)} \tag{5.4.218}$$

with $1 < a < u$ and $(a, u) = 1$.

Given y_n defined by:

$$y_{-1} = u \quad y_0 = a \quad y_{n+1} = \frac{y_{n-1}}{y_n^{q_1}} \tag{5.4.219}$$

Therefore let us prove that:

$$\varepsilon_n = \frac{\ln(y_{n-2})}{\ln(y_{n-1})} - q_n \tag{5.4.220}$$

Indeed for $n = 1$ we have:

$$\varepsilon_1 = x - q_1 = \frac{\ln(u)}{\ln(a)} - q_1 = \frac{\ln(y_{-1})}{\ln(y_0)} - q_1 \tag{5.4.221}$$

for $n = 2$ we use first the fact that:

$$\begin{aligned}\varepsilon_1 = x - q_1 = \frac{\ln(u)}{\ln(a)} - q_1 &= \frac{\ln(u) - q_1\ln(a)}{\ln(a)} = \frac{\ln(u) - \ln(a^{q_1})}{\ln(a)} \\ = \frac{\ln\left(\dfrac{u}{a^{q_1}}\right)}{\ln(a)} &= \frac{\ln\left(\dfrac{y_{-1}}{y_0^{q_1}}\right)}{\ln(a)} = \frac{\ln(y_1)}{\ln(a)}\end{aligned} \tag{5.4.222}$$

Therefore:

$$\frac{1}{\varepsilon_1} = \frac{\ln(a)}{\ln(y_1)} \tag{5.4.223}$$

and as we had proved that:

$$\frac{1}{\varepsilon_{n-1}} = q_n + \varepsilon_n \Rightarrow \varepsilon_2 = \frac{1}{\varepsilon_1} - q_2 = \frac{\ln(a)}{\ln(y_1)} - q_2 = \frac{\ln(y_0)}{\ln(y_1)} - q_2 \tag{5.4.224}$$

etc... by induction demonstrating our right to use this notation changes.

✍**Example:**

Let us look for the expression of the continuous fraction of:

$$x=\log_a(u)=\log_2(3) \tag{5.4.225}$$

We know by playing with the definition of the logarithm that:

$$2^1<3<2^2 \tag{5.4.226}$$

therefore:

$$\begin{aligned}&\log_2(2^1)<\log_2(3)<\log_2(2^2)\Rightarrow\log_2(2)<\log_2(3)<2\log_2(2)\\&\Rightarrow 1<\dfrac{\log_2(3)}{\log_2(2)}<2\Rightarrow 1<\log_2(3)<2\end{aligned} \tag{5.4.227}$$

therefore $q_1=1$. Then we have:

$$\varepsilon_1=\dfrac{\ln(y_{1-2})}{\ln(y_{1_1})}-q_1=\dfrac{\ln(y_{-1})}{\ln(y_0)}-q_1 \tag{5.4.228}$$

and as:

$$y_{-1}=u=3\qquad y_0=a=2 \tag{5.4.229}$$

it comes:

$$\varepsilon_1=\dfrac{\ln(3)}{\ln(2)}-1 \tag{5.4.230}$$

So we have the first partial quotient:

$$\dfrac{a}{b}=\log_2(3)=q_1+\dfrac{1}{1/\varepsilon_1}=1+\dfrac{1}{\dfrac{1}{\frac{\ln(3)}{\ln(2)}-1}}=\ldots \tag{5.4.231}$$

Verbatim we have already:

$$y_{0+1}=y_1=\dfrac{y_{-1}}{y_0^{q_1}}=\dfrac{3}{2^1}=\dfrac{3}{2} \tag{5.4.232}$$

Let us simplify:

$$\left(\dfrac{\ln(3)}{\ln(2)}-1\right)^{-1}=\dfrac{\ln(2)}{\ln\left(\frac{3}{2}\right)}=\dfrac{\ln(2)}{\ln(y_1)} \tag{5.4.233}$$

So the first partial quotient can be written:

$$\log_2(3)=1+\dfrac{1}{\dfrac{\ln(2)}{\ln(y_1)}}=\ldots \tag{5.4.234}$$

and let us go to the second partial quotient. We already know for this that:

$$1<\frac{\ln(2)}{\ln\left(\frac{3}{2}\right)}<2 \tag{5.4.235}$$

So it is immediate that $q_2=1$ and then as:

$$\varepsilon=\frac{\ln(y_{n-2})}{\ln(y_{n-1})}-q_n \tag{5.4.236}$$

We have:

$$\varepsilon_2=\frac{\ln(y_0)}{\ln(y_1)}-q_2=\frac{\ln(2)}{\ln\left(\dfrac{3}{2}\right)}-1=\frac{\ln(2)-\ln\left(\dfrac{3}{2}\right)}{\ln\left(\dfrac{3}{2}\right)}=\frac{\ln\left(\dfrac{4}{3}\right)}{\ln\left(\dfrac{3}{2}\right)} \tag{5.4.237}$$

Finally we get:

$$\log_2(3)=q_1+\frac{1}{q_2+1/\varepsilon_2}=1+\cfrac{1}{1+\cfrac{1}{\cfrac{\ln\left(\dfrac{4}{3}\right)}{\ln\left(\dfrac{3}{2}\right)}}}=\cfrac{1}{1+\cfrac{\ln\left(\dfrac{3}{2}\right)}{\ln\left(\dfrac{4}{3}\right)}} \tag{5.4.238}$$

etc... etc.

~90 %

★★★★☆

21 votes, 68.57%

5.5. Set Theory

During our study of numbers, operators, and number theory (in the chapters of the same name), we often used the terms "groups", "rings", "fields", "homomorphism", etc. and thereafter we will continue to do it again many times. Besides the fact that these concepts are of utmost importance, to give demonstrations or build mathematical concepts essential to the study of contemporary theoretical physics (quantum field physics, string theory, standard model, ...), they allow us to understand the components and the basic properties of mathematics and its operators by storing them in separate categories. So, choose to present the Set Theory as the 5th chapter of this textbook is a very questionable choice as rigorously that it is where almost everything begins... However, we still needed to expose the Proof Theory for the notations and methods that will be used here.

Moreover, when teaching modern mathematics in the secondary or primary (in the years 11970 according to holocene calendar), the language of sets and the preliminary study of binary relations to a more rigorous approach to the notion of functions and applications of mathematics in general was introduced (see definition below).

Definition (#5.5.75): We talk about "arrow diagram" (or "sagittal diagram" from latin "sagitta" = arrow) to all diagram showing a correspondence between the two sets of components connected wholly or partially by a set of arrows. When there is no arrows, we then just speak of "Venn diagrams".

For example, the graphical representation of a defined function of the set $E = \{-3,-2,-1,0,1,2,3\}$ to the set $F = \{0,1,2,4,9\}$ (usually denoted $E \mapsto F$) leads to the sagittal diagram below:

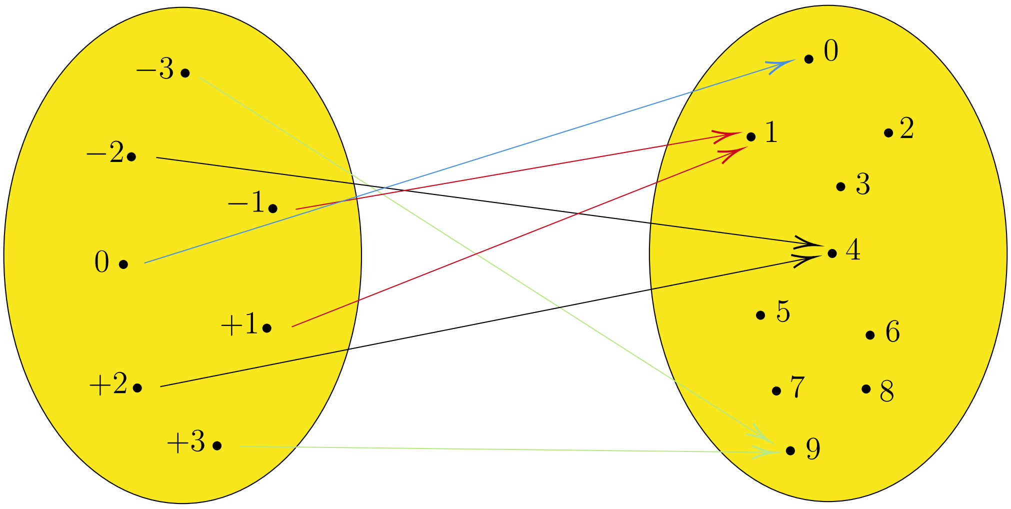


Figure 5.50: **Sagittal diagram example from a definition set to a destination set**

A relation from E to E (usually denoted $E \mapsto E$) provides an arrow diagram of the type:

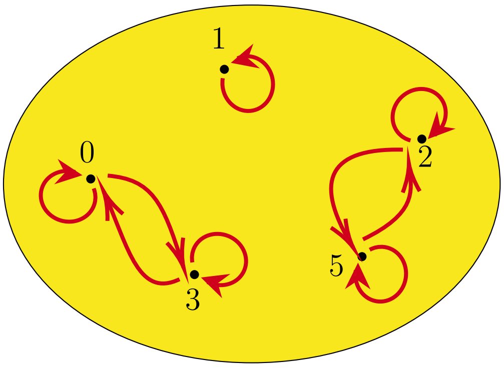


Figure 5.51: **Function returning in its own set of definitions**

The closure of each element showing a "reflexive relation" and the systematic presence of a back arrow indicating a "symmetrical relation".

Definition (#5.5.76): If the target set is identical to the original set, we say that we have a "binary relation".

However choosing to introduce the Set Theory in school classrooms has also some other reason. In fact, for the sake of internal rigour (i.e. not related to reality), a very large part of mathematics was rebuilt within a single axiomatic framework, so named "Set Theory", in the sense that each mathematical concept (previously independent of the other) is returned to a definition where all the logical components come from this same framework: it is regarded as fundamental! Thus, the rigour of reasoning carried out within Set Theory is guaranteed by the fact that the frame is "non-contradictory" or "consistent". Let us see now the definitions that build this framework.

Definitions (#5.5.77):

D1. We name "set" any list, collection or gathering of well-defined objects, explicitly or implicitly.

D2. A "Universe" U is an object whose constituents are sets.

Note that what mathematicians name "Universe" is not a set! In fact it is a model that satisfies to the axioms of sets.

Indeed, we will see that we can not talk about the set of all sets (because this is not a set) to designate the object that consists of all the sets and that's why we talk about "Universe".

D3. We name "elements" or "members of the set" objects belonging to the set and we write:

$$p \in A \tag{5.5.1}$$

if p is an element of the set A and in the contrary case:

$$p \notin A \tag{5.5.2}$$

If B is a "part" of A, or "subset" of A, we write this:

$$B \subset A \quad \text{or} \quad A \supset B \tag{5.5.3}$$

Thus:

$$\forall x, x \in B \Rightarrow x \in A \tag{5.5.4}$$

✍**Examples:**

E1. $A = \{1, 2, 3\}$

E2. $X = \{X \mid x \text{ is a positive integer}\}$

D4. We can provide sets with a number of relations that compare (useful sometimes...) their elements or to compare some of their properties. These relations are named "comparison relations" or "order relations" like $\leq, \geq$ (see section Operators page 290).

D5. Let E be a set, and let $\mathcal{P}$ be a statement depending on a parameter $x \in E$. Let $\mathcal{P}(x)$ be this statement evaluated at x. We will say that $\mathcal{P}$ is a "property on the elements of E" if, for each $x \in E$, the statement $\mathcal{P}(x)$ is a proposition in the mathematical sense, and is therefore either true or false (excluded middle). In this case, we will denote by:

$$\{x \in E | \mathcal{P}(x)\} \tag{5.5.5}$$

the subset of E formed by the elements $x \in E$ for which $\mathcal{P}(x)$ is true.

We note that the excluded middle condition on $\mathcal{P}$ guarantees that the subset:

$$A = \{x \in E | \mathcal{P}(x)\} \tag{5.5.6}$$

is well-defined, meaning that its elements have been well-characterized: if $x \in E$, then $x \in A$ if $\mathcal{P}(x)$ is true, and $x \notin A$ if $\mathcal{P}(x)$ is false. We don't give the list of elements of A, but a property that characterizes its elements: we say that A is given "by comprehension" (as opposed to a set "given by extension"). For example:

$$\{1, 2, 3, 4\} \quad \text{and} \quad \{n \in \mathbb{N}; 1 \leqslant n \leqslant 4\} \tag{5.5.7}$$

are two notations for the same subset of $\mathbb{N}$: the first is in extension, the second in comprehension.

Remarks

R1. The structure of ordered set has original been set up in the framework of the Numbers Theory by Cantor and Dedekind.

R2. As we have proved in the section on Operators, $\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}$ are totally ordered by the usual relations $\leq, \geq$. The relation $<$, often named "strict order" is not an order relation because not reflexive and not antisymmetric (see section Operators page 290). For example, in $\mathbb{N}$ the relation "a divides b" , often denoted by the symbol "|" is a partial order.

R3. If $\mathcal{R}$ is an ordering on E and F is a subset of E, the restriction to F of the relation $\mathcal{R}$ is an order on F, named "order induced by $\mathcal{R}$ in F".

R4. If $\mathcal{R}$ is an order on E, the relation $\mathcal{R}'$ defined by:

$$x\mathcal{R}'y \Leftrightarrow y\mathcal{R}x$$

is an order on E, named "reciprocal order" of $\mathcal{R}$. The reciprocal order $\leq$ of the usual order is the order noted $\geq$ and reciprocal order to the order "a divides b" in $\mathbb{N}$ is the order "b is a multiple of a".

The set is the basic mathematical entity whose existence is defined: it is not defined as itself but by its properties, given by the axioms. It uses a human process: a kind of categorization feature, which allows thought to distinguish several independent qualified elements.

Theorem 5.29. We can demonstrate from these concepts, that the number of subsets of a set of cardinal n is 2^n.

Proof 5.29.1. First there is the empty set $\varnothing$, that is 0 items chosen from n, i.e. C_0^n (notation of binomial coefficient non-conform with ISO 31-11!) as we have seen in the section of Probabilities:

$$C_k^n=\vec{v}=\binom{n}{k}=\frac{A_n^k}{k!}=\frac{n!}{k!\,(n-k)} \qquad (5.5.8)$$

and so on...

The number of subsets (cardinal) of E corresponds to the summation of all binomial coefficients:

$$\boxed{\operatorname{Card}\left(P\left(E\right)\right)=\sum_{k=0}^{n}C_k^n} \qquad (5.5.9)$$

But, we have (see section Calculus page 1364):

$$(x+y)^n=\sum_{k=0}^{n}C_k^n x^k y^{n-k} \qquad (5.5.10)$$

therefore:

$$\boxed{\operatorname{Card}\left(P\left(E\right)\right)=\sum_{k=0}^{n}C_n^k 1^k 1^{n-k}=(1+1)^n=2^n} \qquad (5.5.11)$$

■ Q.E.D.

✍**Example:**

Consider the set $S = \{x_1, x_2, x_3\}$, we have the set of all parts of $P(S)$ consisting of:

- The empty set: $\{\} = \varnothing$
- The singletons: x_1, x_2, x_3
- The duets: $x_1, x_2, x_1, x_3, x_2, x_3$
- Itself: $\{x_1, x_2, x_3\}$

Such that:

$$P(S) = \{\varnothing, x_1, x_2, x_3, x_1, x_2, x_1, x_3, x_2, x_3, x_1, x_2, x_3\} \tag{5.5.12}$$

What makes indeed 8 elements!

Remark

The order in which the elements are differentiated does not come into account when counting parts of the original set.

In Applied Mathematics, we work almost exclusively with sets of numbers. Therefore, we will limit our study of definitions and properties of these.

Now let us formalize the basic concepts for working with the most common sets we encounter in the basic school curriculum.

5.5.1 Zermelo-Fraenkel Axiomatic

The Zermelo-Fraenkel axiomatic, abbreviated sometimes "ZF-C axioms" shown below was formulated by Ernst Zermelo and Abraham Adolf Fraenkel specified by the early 120th century (holocene calendar) and completed by the axiom of choice (hence the capital C in ZF-C). It is considered as the most natural axiomatic structure in the context of set theory.

Remark

There are many other axiomatic structures, based on the more general concept of "class", as developed by von Neumann, Bernays and Gödel (for the notations, see section Proof Theory page 131).

Strictly technically speaking..., the ZF axioms are statements of calculation for first order predicate (see section Proof Theory page 163) egalitarian in a language with only one primitive symbol for membership (binary relation). The following should therefore only be seen as an attempt (...) to express in English the expected significance of these axioms.

A1. Axiom of extensionality:

Two sets are equal if and only if they have the same elements. This is what we denote:

$$A = B \Leftrightarrow (\forall x \in A, x \in B) \wedge (\forall x \in B, x \in A) \tag{5.5.13}$$

So A and B are equal if every element x of A is also in B and every element x of B also belongs to A.

A2. Axiom of empty set:

The empty set exists, we denote it:

$$\varnothing \tag{5.5.14}$$

and it has no element, its cardinality is therefore 0.

In fact this axiom can be deduced from another axiom that we will see a little further but it is convenient to introduce it by convenience for teaching in high-school classes.

A3. Axiom of pairing:

If A and B are two sets, then, there exist a set C containing A and B alone and as components. This set C is then denoted $\{A, B\}$.

From the perspective of the sets considered elements that gives:

$$\forall A \forall B \exists C : \quad A \in C \wedge B \in C \tag{5.5.15}$$

This axiom also shows the existence of the "singleton" a set noted:

$$\{X\} \tag{5.5.16}$$

which is a set whose only element is X (and therefore with unitary cardinal). We simply need to apply the axiom asking equality between A and B.

A4. Axiom of the sum (also named "axiom of union"):

This axiom allows us to build the union (merge) of sets. Said in a most common way: the union (merge) of any family of a set, is... a set.

The union of any family of sets is often denoted:

$$\bigcup_i A_i \tag{5.5.17}$$

or if we take some of its elements:

$$\bigcup_{x \in A} x_i \tag{5.5.18}$$

A5. Axiom of subsets:

This axiom defines that for any set A, the set of all its parts $P(A)$ exists (do not confuse with the "P" of probability!).

So for any set A, we can associate a set B which contains exactly the parts C (verbatim the subsets) of the first:

$$\forall A \exists B \forall C : \quad (C \in B \Leftrightarrow C \subset A) \tag{5.5.19}$$

A6. Axiom of infinity:

This axiom express the fact that there exist an infinite set. To formalize it, we say that there exist a set, named "autosuccessor set" A containing $\varnothing$ (the empty set) such that if x belongs to A, then also $x \cup \{x\}$ belongs to A:

$$A \text{ is autosuccessor} \Leftrightarrow (\varnothing \in A) \wedge (x \in A \Rightarrow (x \cup \{x\}) \in A) \tag{5.5.20}$$

This axiom expresses for example that the set of integers exists. Indeed, $\mathbb{N}$ is so the smallest autosuccessor set in the sense of inclusion $\mathbb{N} = \{\varnothing, \{\varnothing, \{\varnothing, \ldots\}\}\}$ and by convention we note (where we build the Natural Set):

$$\begin{aligned} 0 &= \varnothing \\ 1 &= \{\varnothing\} \\ 2 &= \{\varnothing, \{\varnothing\}\} \\ \ldots \end{aligned} \tag{5.5.21}$$

A7. Axiom of regularity (also named "foundation axiom"):

The main purpose of this axiom is just to eliminate the possibility of having A as part of itself.

Thus, for any non-empty set A, there exists a set B which is an element of A such that no element of A is an element of B (you must distinguish the level of the language used, a set and its elements have not the same status!) that we note:

$$\forall A \neq \varnothing : \quad \exists B \in A, A \cap B = \varnothing \tag{5.5.22}$$

and thus result we expected to have:

$$\forall A, A \notin A \tag{5.5.23}$$

Proof 5.29.2. Indeed, let A be a set such that $A \in A$. Consider the singleton $\{A\}$, set whose only element is A. According to the axiom of foundation, we must have an element of this singleton that has no element in common with him. But the only possible element is A itself , that is to say that we must have:

$$A \cap \{A\} = \varnothing \tag{5.5.24}$$

But by hypothesis $A \in A$ and by construction $A \in \{A\}$. So:

$$A \in (A \cap \{A\}) \tag{5.5.25}$$

which contradicts the previous assertion. Therefore:

$$A \notin A \tag{5.5.26}$$

■ Q.E.D.

Remark

Every set contains as a subset the empty set $\emptyset = \{\}$, and it also equally trivially contains itself as a subset. For the empty set these two cases are in fact the same, and indeed the empty set is the unique subset of the empty set. So we can form the set $\{\emptyset\}$ whose only member is $\emptyset$ but note that $\{\emptyset\} \neq \emptyset$, because $\emptyset \in \{\emptyset\}$ however $\emptyset \notin \emptyset$. In other words: nothing can be part of something but cannot contain itself! That's why there is always something instead of nothing!

A8. Axiom of replacement (also named "Axiom schema of replacement"):

This axiom expresses the fact that if a formula f is a functional then for any set A, there is a set B consisting precisely of the images of A by this function.

So, in a little more formally way, the set A of elements a and a binary relation f (which is quite generally a functional), there exist a set B consisting of elements b such that $f(a, b)$ is true. If f is a function where b is not free then it means that:

$$b = f(A) \text{ and } B = f(A) \tag{5.5.27}$$

In a technical way we write this axiom as following:

$$\forall A\ \exists a \in A\ \exists! f : \quad f(a,b) \Rightarrow \exists B \forall a \in A \exists b \in B f(a,b) \tag{5.5.28}$$

So for every set A and any item it contains, there is one and only one b defined by the functional f such that there exists a set B for which any element a belonging to the set A there is a b belonging to set B defined by the functional f.

Let's see an example with the following binary predicate that for the value of any a from A determines the value of any b of B:

$$P(a,b) = (a = 1 \wedge b = 2) \vee (a = 3 \wedge b = 4) \tag{5.5.29}$$

Therefore from the knowledge that a is equal 1 we derive that b is equal 2 and similarly (i.e. by replacement) when a is equal 3, we derive that b is equal 4.

We see well through this small example the strong relation that exists considering the predicate P as a naive function! Moreover, as there is an infinity of possible functions f, the replacement scheme is considered as an infinite number of axioms.

A9. Axiom of selection (also named "Axiom comprehension schema"):

This axiom simply expresses that for any set A and any property P expressible in the language of set theory, the set of all elements of A satisfying the property P exist.

So more formally, to any set A and any condition or proposition $P(x)$, there is a set B whose elements are exactly the elements x of A for which $P(x)$ is true. This is what we write:

$$B = \{x \in A : P(x)\} \tag{5.5.30}$$

In a more comprehensive and rigorous way we have in fat for any functional f that does not include a as free variable:

$$\forall A \exists B \forall a : \quad a \in B \Leftrightarrow a \in A \wedge f \tag{5.5.31}$$

It is typically the axiom that we use to construct the set of even numbers:

$$\{a \in \mathbb{N} \mid \exists b \in \mathbb{N}, a = 2b\} \tag{5.5.32}$$

or to prove the existence of the empty set (which invalidates the axiom of the empty set) because you just have to ask that there exist a set that satisfies the property:

$$A \neq A \tag{5.5.33}$$

and regardless of the set A. And only the empty set satisfies this property by the selection axiom.

The compliance with the strict conditions of this axiom eliminates the paradoxes of the "naive set theory", as Russell's paradox or Cantor's paradox who invalidated the naive set theory.

For example, consider the Russell set R of all sets that do not contain themselves (note that we give a property of R without specifying what is this set):

$$R = \{E : E \notin E\} \tag{5.5.34}$$

The problem is to know whether or not R contains itself or not. If $R \in R$, then, R is self-contained, and, by definition $R \notin R$, and vice versa. Each possibility is contradictory.

If we now denote by C the set of all sets (Cantor Universe) we have in particular:

$$P(C) \in C \tag{5.5.35}$$

which is impossible (i.e. with the power of the continuum of real numbers), according to Cantor's theorem (see section Numbers page 225).

These "paradoxes" (or "syntactic antinomies") come from a non-compliance with the conditions of application of the selection axiom: to define E (in the example of Russell), there must be a proposition P which bears on the set R, which should be explicated. The proposal defining the set of Russell or that of Cantor does not indicate what is the set E. It is therefore invalid!

A very nice and well known example (this is why we present it) helps to better understand (this is the "Russell paradox" which we have already spoken about in length in the section on Proof Theory page 138):

A young student went one day to his barber. He entered into conversation and asked him if he had many competitors in his pretty city. Seemingly innocent way, the barber replied, *« I have no concurrence. Because of all the men of the city, I obviously do not shave those who shave themselves, but I am fortunate to shave. all those who do not shave themselves»*.

What then in such a so simple statement could take to the fault the logic of our young smart student?

The answer is in fact innocent, until we decide to apply to the case of the barber: Does he shaves himself, Yes or No?

Suppose he shaves himself: he then belongs to the category of those who shave themselves, those who the barber said he did of course not shave.... So he does not shave himself..........

Finally, this unfortunate barber is in a strange position: if he shaves himself, he does not shave himself, and if he does not shave himself, he shaves himself. This logic is self-destructive, contradictory stupidly, rationally irrational.

Then comes the selection axiom: We exclude the barber of all persons to which the declaration applies. Because in reality, the problem is that the barber is a member of the set of all the men of the city. So what applies to all men does not apply to the individual case of the barber.

A10. Axiom of choice:

Given a set A of non-empty mutually disjoint sets, there exist a set B (the set of choices for A) containing exactly one element for each member A.

However let us indicate that the issue of the axiomatization and therefore of the foundations found himself still shaken by two questions at the time of their construction: what valid axioms must be chosen and in a system of axioms are the mathematics coherent (do we not have a risk of seeing a contradiction)?

The first issue was first raised by the continuum hypothesis: if we can put two sets of numbers in correspondence term to term, they have the same number of elements (cardinal). We can thus map all integer numbers $\mathbb{N}$ with rational numbers $\mathbb{Q}$ as we have shown in the section on Numbers (page 225), so they have the same cardinality, be we can not map integer numbers with all the real numbers. The question then is whether there is a set whose number of elements would be located between the two or not? This question is important to build the classical theory of analysis and mathematicians usually choose to say there is none, but we can also say the opposite.

In fact the continuum hypothesis is linked in a more profound way we could thing to the axiom of choice which can also be formulated as follows: if C is a collection of non-empty sets then we can select any element of each set of the collection. If C has a finite number of elements or a countable number of elements, the axiom seems pretty trivial: we can sort and number the sets of C and the selection of an element in each set is simple. Where it begin to get complicated is when the set C has the power of the continuum: how to choose the elements if it is not possible to number them?

Finally in 11938 (holocene calendar) Kurt Gödel shows that set theory is consistent without the axiom of choice and without the continuum hypothesis as well as with! And to end it all Paul Cohen in 11963 (holocene calendar) shows that the axiom of choice and the continuum hypothesis are not related.

Ok to make a pedagogical summary of all this stuff consider the following figure (excluding the axiom of choice):

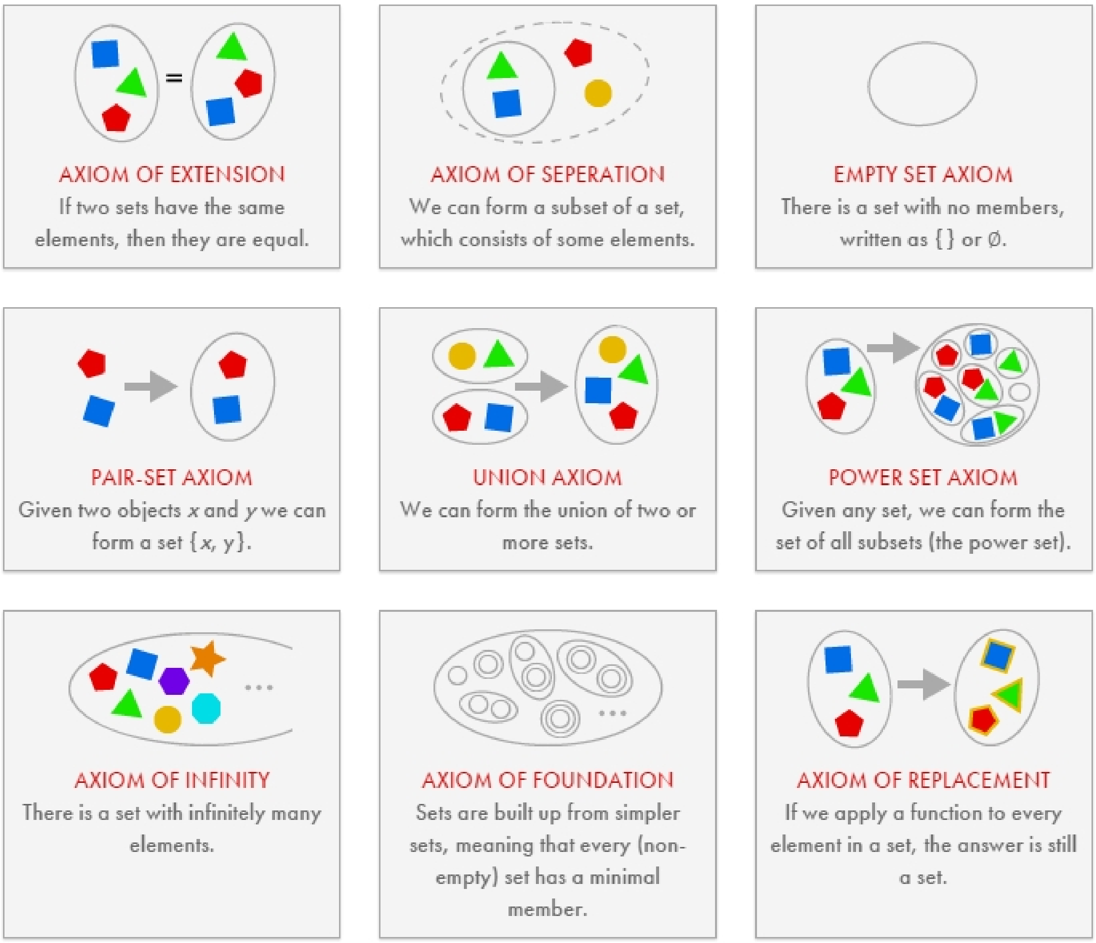

AXIOM OF EXTENSION
If two sets have the same elements, then they are equal.
AXIOM OF SEPERATION
We can form a subset of a set, which consists of some elements.
EMPTY SET AXIOM
There is a set with no members, written as {} or Ø.
PAIR-SET AXIOM
Given two objects x and y we can form a set {x, y}.
UNION AXIOM
We can form the union of two or more sets.
POWER SET AXIOM
Given any set, we can form the set of all subsets (the power set).
AXIOM OF INFINITY
There is a set with infinitely many elements.
AXIOM OF FOUNDATION
Sets are built up from simpler sets, meaning that every (non-empty) set has a minimal member.
AXIOM OF REPLACEMENT
If we apply a function to every element in a set, the answer is still a set.

Figure 5.52: **Zermelo-Frankel axioms visual summary (source:?)**

5.5.1.1 Cardinals

Definition (#5.5.78): Sets are said to be "equipotent" if there exists a bijection (one-one correspondence) between these sets. We thus say they have same "cardinal" that the norm ISO 3111 advocated to write $card(S)$ but in this textbook we will also use the notation $\text{Card}(S)$ (many U.S. books use non-official notation that looks exactly like the absolute value $\mid S\mid$ or $\#S$).

Thus, more rigorously, a cardinal (which quantifies the number of items in the set) is an equivalence class (see section Operators page 292) for the relation of equipotence.

Remark

Cantor is the main creator of set theory, in a form that we name today "naive set theory". But, apart to elementary considerations, his theory was also consisting of higher abstraction levels. The real novelty of the Cantor theory is that it lets talk about infinity. For example, an important idea Cantor was precisely to define the "equipotence".

If we write $c_1 = c_2$ as equality of cardinals, we mean by that there are two equipotent sets A and B such that:

$$c_1 = \text{Card}(A) \quad \text{and} \quad c_2 = \text{Card}(B) \tag{5.5.36}$$

Cardinals can be compared. The order thus defined is a total ordering (see section Operators page 291) between the Cardinals (the proof that the order relation is complete uses the axiom of choice and the proof that it is antisymmetric is known under the name of Cantor-Bernstein's theorem that we will demonstrate further below).

Say that $c_1 < c_2$ means in simple language that A is equipotent to a proper part of B, but B is not equipotent to any own part of A. Mathematicians would say that $\text{Card}(A)$ is smaller or equal to the $\text{Card}(B)$ if there is an injection of A into B.

We saw during our study of numbers (see section Numbers page 389), especially of transfinite numbers, that an equipotent set (or bijection) to $\mathbb{N}$ was told to "countable set".

Let us now see this notion a little more in detail:

Let A be a set, if there is an integer n such that there is at least for each element of A a corresponding item int the set $\{1, 2, \ldots, n\}$ (in fact this is rigorously a bijection... concept that we will define later) then we say that the cardinal of A, denoted $\text{Card}(A)$ or $\text{Card}(A)$ is a "finite cardinal" and its value is n.

Otherwise, we say that the set A has an "infinite cardinal" and we write:

$$\text{Card}(A) = +\infty \tag{5.5.37}$$

A set A is "countable" if there is a bijection between A and (N). A set of numbers A is "countable" if there is a bijection between A and part of (N). A set at maximum countable is thus of finite cardinal, or countable.

We can therefore check the following proposals:

P1. A part of a countable set is at most countable.

P2. A set containing a non-countable set is also not countable.

P3. The product of two countable sets is countable.

Remark

We can restrict a set of numbers relatively with the null element and the negative or positive elements in it and therefore we write (example for the real set):

$$\begin{aligned}
\mathbb{R}^* &= \mathbb{R} \setminus \{\varnothing\} \\
\mathbb{R}_+ &= \{x \mid x \in \mathbb{R} \text{ and } x \geq 0\} \\
\mathbb{R}_- &= \{x \mid x \in \mathbb{R} \text{ and } x \leq 0\} \\
\mathbb{R}_+^* &= \mathbb{R}_+ \setminus \{\varnothing\} \\
\mathbb{R}_-^* &= \mathbb{R}_- \setminus \{\varnothing\}
\end{aligned} \tag{5.5.38}$$

These concepts are similar for $\mathbb{N}, \mathbb{Z}, \mathbb{Q}$ (the set of complex numbers $\mathbb{C}$ being not ordered, the second and third line does not apply to).

So any infinite subset of $\mathbb{N}$ is equipotent to $\mathbb{N}$ itself, what may seem counter-intuitive at first...!

In particular, there are as many even integers as any natural integer numbers (use the bijection $f(n) = 2n$) from $\mathbb{N}$ to P, where P is the set of even natural numbers. As many relative numbers as integers, as many integers as rational numbers (see the section Numbers page 225 for the proofs).

Thus we can write:

$$\text{Card}(\mathbb{N}) = \text{Card}(\mathbb{Z}) = \text{Card}(\mathbb{Q}) = \aleph_0 \tag{5.5.39}$$

and more generally, any infinite part of $\mathbb{Q}$ is countable.

Thus we have an important result: any infinite set therefore has an infinite countable part.

Since we have proved in the section on Numbers (page 232) that the set of real numbers $\mathbb{R}$ has the "power of the continuum" and that the set of natural numbers $\mathbb{N}$ has transfinite cardinal $\aleph_0$, Cantor raised the question whether there was a cardinal between the transfinite cardinal $\aleph_0$ and the cardinal of $\mathbb{R}$? In other words, we have an infinite amount of integers, and an even greater amount of real numbers. So does it exist an infinite greater than the infinite of integers and smaller than that of the real numbers?

The problem arose by writing $\aleph_0$ the cardinal of $\mathbb{N}$ and $\aleph_1$ (new) the cardinal of $\mathbb{R}$ and offering to demonstrate or contradict that:

$$\aleph_1 = 2^{\aleph_0} \tag{5.5.40}$$

according to the combinatorial law that gives the number of elements that we can get from all subsets of a set (as we have proved it before).

The rest of his life, Cantor tried, in vain, to prove this result that we name the "continuum hypothesis". He did not succeed and descended into madness. In 19000 (holocene calendar), the International Congress of Mathematicians, Hilbert considered that this was one of the 23 major issues that should be resolved in the 120th century (holocene calendar).

This problem is solved in a rather surprising way. First, in 11938, one of the greatest logicians of the 120th century (holocene calendar), Kurt Gödel showed that the hypothesis of Cantor was not refutable, that is to say, we could never prove that it was false. Then in 11963 (holocene calendar), the mathematician Paul Cohen closed the debate. He demonstrated that we could never prove that it was true!!! We can conclude rightly that Cantor had become mad to try to demonstrate a problem that could not be proved.

5.5.1.2 Cartesian Product

Definition (#5.5.79): If E and F are two sets, we name "Cartesian product of E by F" the set noted $E \times F$ (not to be confused with the vector product notation) consisting of all possible pairs (e, f) where e is an element of E and f an element of F.

More formally:

$$\boxed{E \times F = \{(e, f) \mid e \in E \wedge f \in F\}} \tag{5.5.41}$$

We note the Cartesian product of E by itself:

$$E \times E = E^2 \tag{5.5.42}$$

and then we say that E^2 is the "set of pairs of elements of E".

We can perform the Cartesian product of a sequence $E_1 \times E_2 \times \ldots \times E_n$ of sets and get all n-tuples $(e_1, e_2, \ldots, e_n)$ where $e_1 \in E_1, e_2 \in E_2, \ldots, e_n \in E_n$.

In the case where all sets E_i are identical to E, the Cartesian product $E_1 \times E_2 \times \ldots \times E_n$ is obviously noted E^n. We then say that e^n is the "set of all n-tuples of elements of E".

If E and F are finite then the Cartesian product $E \times F$ is finished. Moreover:

$$\text{Card}(E \times F) = \text{Card}(E) \cdot \text{Card}(F) \tag{5.5.43}$$

From here we see that if the sets $E_1, E_2, \ldots, E_n$ are finished then the Cartesian product $E_1 \times E_2 \times \ldots \times E_n$ is finished and we have:

$$\text{Card}(E_1 \times E_2 \times \ldots \times E_n) = \prod_{i=1}^{n} \text{Card}(E_i) \tag{5.5.44}$$

In particular:

$$\text{Card}(E^n) = [\text{Card}(E)]^n \tag{5.5.45}$$

if E is a finite set.

✍**Examples:**

E1. If $\mathbb{R}$ is the set of real numbers, then $\mathbb{R}^2$ is the set of all couples of real numbers. In the plane reported to a referential, any point M has the coordinates that are an element of $\mathbb{R}^2$.

E2. When we run two dice whose faces are numbered 1 through 6, each die can be symbolized by the set $E = \{1, 2, 3, 4, 5, 6\}$. The outcome of a roll of dies is then an element of $E^2 = E \times E$. The Cardinal of $E \times E$ is then 36. There are therefore 36 possible results when we launch two dices whose faces are numbered 1 to 6.

Remark

Set theory and the concept of cardinal is the theoretical basis of relational database softwares.

5.5.1.3 Intervals

Let M be a set of any numbers so that $M \subset \mathbb{R}$ (particular but frequent example). We have for definitions: **Definitions (#5.5.80):**

D1. $x \in \mathbb{R}$ is named "upper bound" of the set M, if $x \geq m$ for $\forall m \in M$. Conversely, we speak about "lower bound" (so do not confuse the concept of terminal with the concept of interval!).

D2. Either $M \subset \mathbb{R}, M \neq \varnothing$. $x \in \mathbb{R}$ is named the "smallest upper bound" noted:

$$x = \text{sup} M \tag{5.5.46}$$

of M if x is an upper bound of M and if for any upper bound $y \in \mathbb{R}$ we have $x \leq y$. Conversely, we speak about the "smaller lower bound" that we note:

$$x = \text{inf} M \tag{5.5.47}$$

The definitions are equivalent in the context of functional analysis (see section of the same name page 2170) as the functions are defined on sets.

Indeed, let f be a function whose domain of definition I swept all $\mathbb{R}$. We note that:

$$f : E \mapsto \mathbb{R} \tag{5.5.48}$$

and let $x_0 \in \mathbb{R}$.

Definitions (#5.5.81):

D1. We say that f has a "global maximum" on x_0 if:

$$\forall x \in E : \quad f(x) \leq f(x_0) \tag{5.5.49}$$

D2. We say that f has a "global minimum" on x_0 if:

$$\forall x \in E : \quad f(x) \geq f(x_0) \tag{5.5.50}$$

In each of these cases, we say that f has an "global extremum" on x_0 (it is a concept that we often use in the sections of Analytical Mechanics page 3152 and Numerical Methods page 6272!).

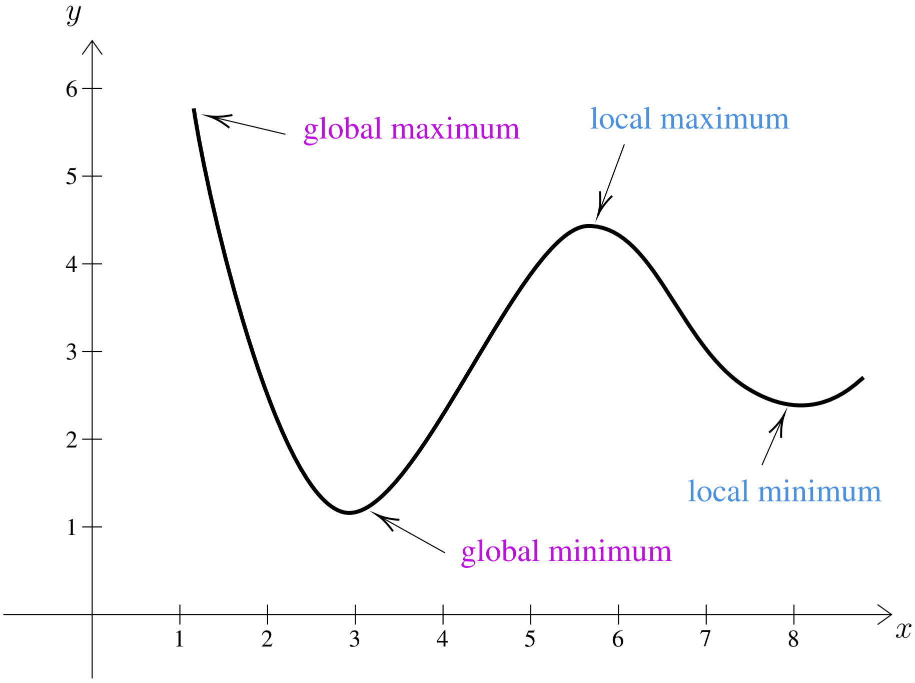


Figure 5.53: Global/Local naximum and minimum example (source: Wikipedia)

D3. f is "upper bounded" if there is a real number M such as $\forall x \in I, f(x) \leq M$. In this case, the function has an upper bound of f on its domain of definition I traditionally denoted:

$$\sup_I f \tag{5.5.51}$$

D4. f is "lower bounded" if there is a real M such that $\forall x \in I, f(x) \geq M$. In this case, the function has a lower bound of f on its domain of definition I traditionally denoted:

$$\inf_I f \tag{5.5.52}$$

D5. We say that f is "bounded" if it is both lower bounded and upper bounded (typically the case of trigonometric functions).

5.5.2 Set Operations

We can build from at least three sets A, B, C all sets operations (which notations are due to Richard Dedekind) existing in set theory (very useful in the study of probability and statistics).

Remark

Some of the notations below will be frequently use later in relatively complex theorems, so it is necessary to understand them deeply!

Thus, we can construct the following set operations:

5.5.2.1 Inclusion

In the simplest case, we define the "inclusion" as:

$$\boxed{A \subset B \Leftrightarrow \forall x \left[x \in A \Rightarrow x \in B \right]} \tag{5.5.53}$$

In a non-specialized language here's C: A is "included" (is a "part", or is a "subset") in B then for all x belonging to each of these x also belongs to B:

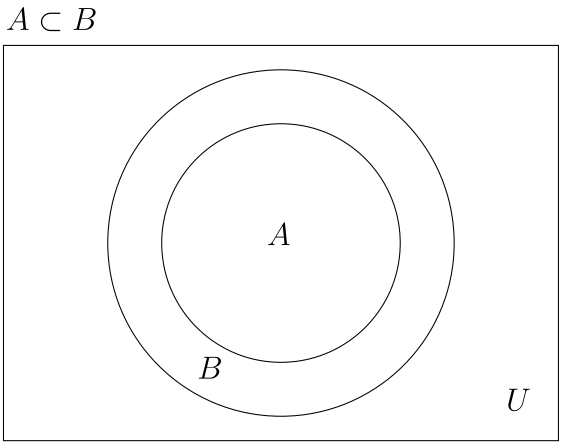


Figure 5.54: **Visual example (Euler diagram) of the inclusion**

where the U in the lower right corner of the figure represents the Cantor Universe.

From this it follows the following properties:

P1. If $A \in B$ and $B \in A$ then it implies $A = B$ and vice versa.

P2. If $A \in B$ and $B \in C$ then implies $A \in C$.

5.5.2.2 Intersection

In the simplest case, we define the "intersection" as:

$$\boxed{A \cap B = \{x \mid x \in A \wedge x \in B\}} \tag{5.5.54}$$

In a non-specialized language here's what you have to read: the "intersection" of sets A and B consists of all the elements that are both in A and in B:

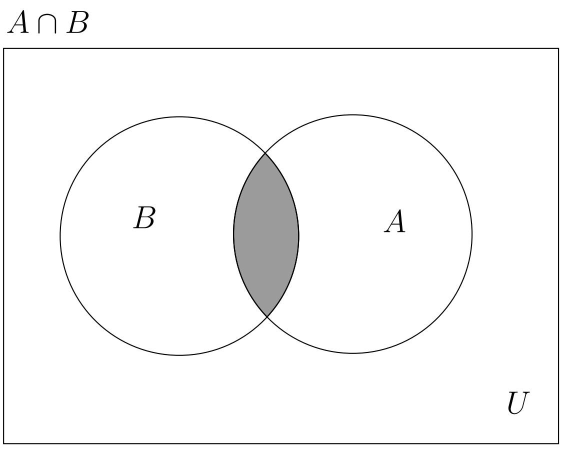


Figure 5.55: **Visual example (Euler diagram) of the intersection**

More generally, if (A_i) is a family of sets indexed by $i \in I$, the intersection of the $(A_i), i \in I$ is denoted:

$$\bigcap_{i \in I} A_i \qquad (5.5.55)$$

This intersection is explicitly defined by:

$$\bigcap_{i \in I} A_i = \{x \mid \forall i \in I x \in A_i\} \qquad (5.5.56)$$

That is to say the intersection of the family of indexed sets includes all x that are located in each set of all sets of the family.

Given two sets A and B, we say they are "disjoint" if and only if:

$$A \cap B = \varnothing \qquad (5.5.57)$$

Furthermore, if:

$$A \cap B = \varnothing \Leftrightarrow \text{Card}(A \cup B) = \text{Card}(A) + \text{Card}(B) \qquad (5.5.58)$$

Mathematicians note that:

$$A \sqcup B \qquad (5.5.59)$$

and name it "disjoint union".

We sometimes joke that knowledge is built on the disjunction... (those who understand will appreciate...).

Definition (#5.5.82): A collection $S = \{S_i\}$ of non-empty sets form a "partition" of a set A if the following properties hold:

P1. $\forall S_i, S_j \in S$ and $i \neq j \Rightarrow S_i \cap S_j = \varnothing$

P2. $A=\bigcup_{S_i\in S} S_i$

✍**Example:**

The set of even numbers and the set of odd numbers are a partitions of $\mathbb{Z}$.

The intersection law is trivially a commutative law (see further below the definition of the concept of "law") as:

$$A\cap B=B\cap A \tag{5.5.60}$$

5.5.2.3 Union

In the simplest case, we define the "union" (also sometimes named "merge") as:

$$\boxed{A\cup B=\{x\mid x\in A\vee x\in B\}} \tag{5.5.61}$$

In a non-specialized language here's what you have to read: the "union" (or "merge") of the sets A and B is the set of elements that are in A plus those that are in B.

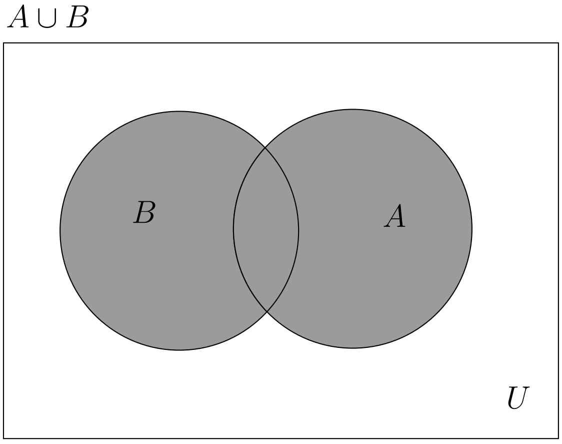


Figure 5.56: Visual example (Euler diagram) of the union

More generally, if (A_i) is a family of sets indexed by $i\in I$, the union of the $(A_i),i\in I$ is denoted:

$$\bigcup_{i\in I} A_i \tag{5.5.62}$$

This union is explicitly defined by:

$$\bigcup_{i\in I} A_i=\{x\mid \exists i\in I x\in A_i\} \tag{5.5.63}$$

That is to say that the union of the family of indexed sets includes all x for which there is a set indexed by i such that x is included in one of the set A_i.

We have the following distributive properties:

$$\begin{aligned}\left(\bigcup_{i\in I} A_i\right)\cap B=\bigcup_{i\in I}(A_i\cap B)\\ \left(\bigcap_{i\in I} A_i\right)\cup B=\bigcap_{i\in I}(A_i\cup B)\end{aligned} \tag{5.5.64}$$

The law of union $\cup$ is a commutative law (see further below the definition of the concept of "law") as:

$$A\cup B=B\cup A \tag{5.5.65}$$

We also name "idempotences laws" the relations (note that for the general culture):

$$\begin{aligned}A\cap A=A\\ A\cup A=A\end{aligned} \tag{5.5.66}$$

and "absorptions laws" the relations:

$$\begin{aligned}A\cap(A\cup B)=A\\ A\cup(A\cap B)=A\end{aligned} \tag{5.5.67}$$

The laws of intersection and union are associative, such that:

$$\begin{aligned}A\cap(B\cap C)=(A\cap B)\cap C\\ A\cup(B\cup C)=(A\cup B)\cup C\end{aligned} \tag{5.5.68}$$

and distributive such that:

$$\begin{aligned}A\cap(B\cup C)=(A\cap B)\cup(A\cap C)\\ A\cup(B\cap C)=(A\cup B)\cap(A\cup C)\end{aligned} \tag{5.5.69}$$

If we recall the concept of "cardinal" (see above) we have with the previously defined operations, the following relation:

$$\text{Card}(A\cup B)=\text{Card}(A)+\text{Card}(B)-\text{Card}(A\cap B) \tag{5.5.70}$$

Hence if $A\cap B=\varnothing$:

$$\text{Card}(A\cup B)=\text{Card}(A)+\text{Card}(B) \tag{5.5.71}$$

5.5.2.4 Difference

In the simplest case, we define the "difference" as:

$$\boxed{A\setminus B=\{x\mid x\in A\wedge x\notin B\}} \tag{5.5.72}$$

In a non-specialized language here's what you have to read: The "difference" of the sets A and B consists of all the elements found only in A (and thus excluding those of B):

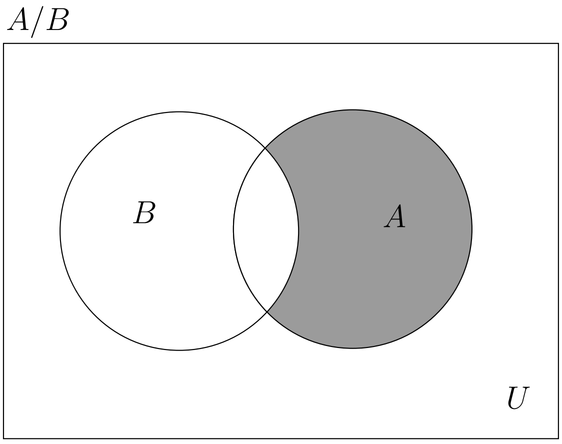


Figure 5.57: Visual example (Euler diagram) of the difference

5.5.2.5 Symmetric Difference

Let U be a set. For any $A, B \subseteq U$ we define the "symmetric difference" $A\delta B$ between A and B by:

$$\boxed{A \triangle B = (A \setminus B) \cup (B \setminus A)} \tag{5.5.73}$$

Or more explicitly using the "exclusive or" in (boolean) logic (see page 146):

$$\begin{aligned} A \oplus B = A\triangle B &= \{x : (x \in A) \oplus (x \in B)\} \\ &= \left(A \cap B^{\complement}\right) \cup \left(A^{\complement} \cap B\right) = \{x : ((x \in A) \wedge (x \notin B)) \vee ((x \notin A) \wedge (x \in B))\} \\ &= (A \cup B) \cap (A \cap B)^{\complement} = \{x : ((x \in A) \vee (x \in B)) \wedge \neg((x \in A) \wedge (x \in B))\} \end{aligned} \tag{5.5.74}$$

In a non-specialized language here's what you have to read: The "symmetric difference" of the sets A and B consists of all items that are only in A and those found only in B (we pass aside elements that are common):

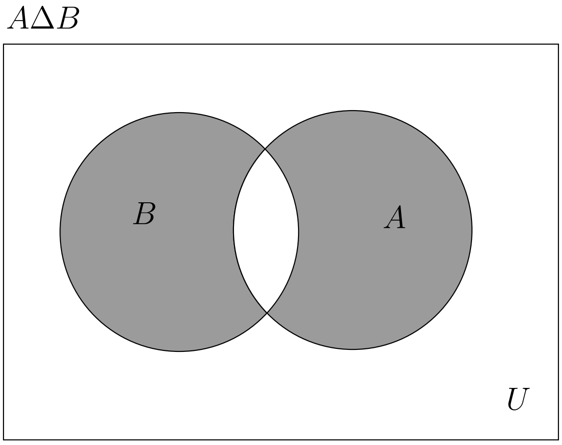


Figure 5.58: Visual example (Euler diagram) of the symmetric difference

So as we can see we have:

$$A \triangle B = (A \cup B) \setminus (A \cap B) \tag{5.5.75}$$

Some trivial properties are given below:

- P1. Commutativity: $A \bigtriangleup B = B \bigtriangleup A$
- P2. Complementarity (see definition below): $A^c \bigtriangleup B^c = A \bigtriangleup B$
- P3. $A \bigtriangleup B = (A \cup B) \setminus (A \cap B)$

5.5.2.6 Product

In the simplest case, we define the "set product" or "cartesian product" as:

$$\boxed{A \times B = \{(x,y) \mid x \in A \wedge y \in B\}} \tag{5.5.76}$$

In a non-specialized language here's what to you have to read: "product" (not to be confused with the multiplication or cross product of vectors) of two sets A and B is the set of pairs such as each element of each set is combined with all elements of the other set.

The product set of real numbers for example generates the plane where each element is defined by X and Y axis.

We often find products sets in mathematics and physics when we work with functions. For example, a function of two real variables which gives real output will be written:

$$\underset{\mathbb{R}\times\mathbb{R}\longrightarrow\mathbb{R}}{f(x,y)} \rightarrow z \tag{5.5.77}$$

or more simply:

$$f(x,y) = z \tag{5.5.78}$$

5.5.2.7 Complementarity

In the simplest case, we define the "complementarity" as:

$$\boxed{\forall A \subset U \quad \bar{A} = \{x \mid x \in U, x \notin A\}} \tag{5.5.79}$$

In a non-specialized language here's what you have to read: The "complementary" is defined as taking a set U and a subset A of U then the complement of A in U is the set of elements that are in U but not in A:

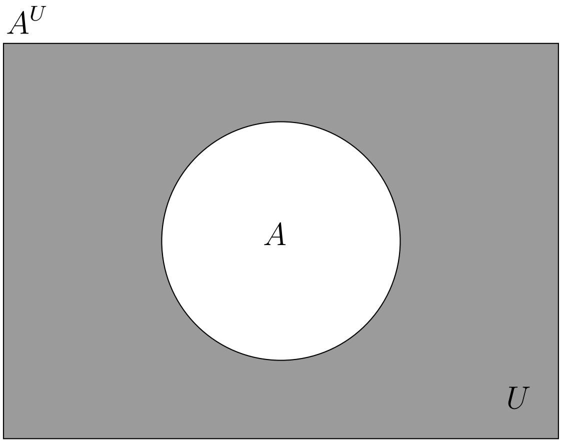


Figure 5.59: **Visual example (Euler diagram) of the complementarity**

Other notations of complementarity that is sometimes found in the literature and the following book are (depending on the context to avoid confusion with other stuff):

$$_U A^c \text{ or } A^c \tag{5.5.80}$$

or in the particular example above, we could also just write $U \setminus A$.

We have for properties for all A_i included in any B:

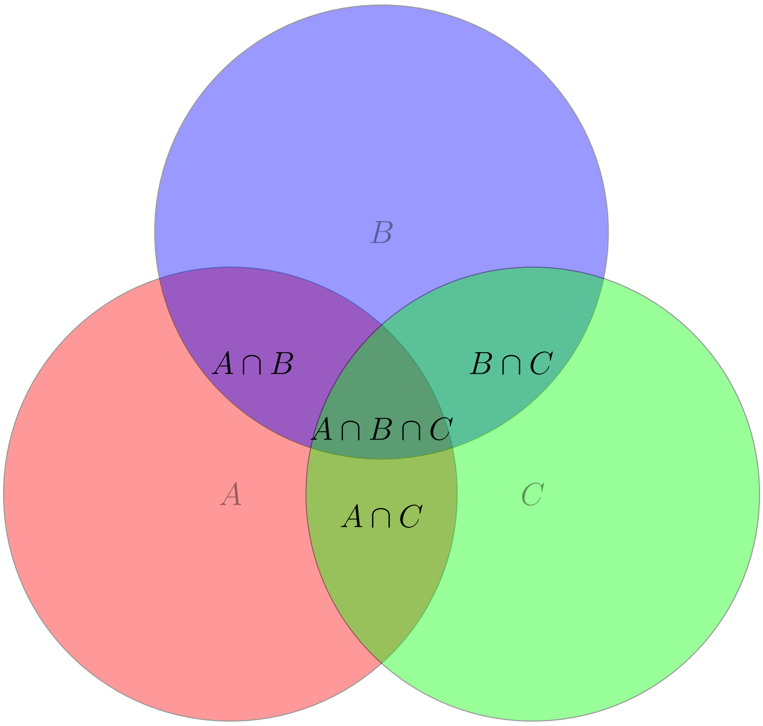
$$\left(\bigcup_{i\in I} A_i\right)^c = \bigcap_{i\in I} A_i^c \tag{5.5.81}$$

$$\left(\bigcap_{i\in I} A_i\right)^c = \bigcup_{i\in I} A_i^c \tag{5.5.82}$$

Here are some trivial properties regarding to complementarity:

$$\begin{aligned}\bar{\bar{A}} &= A\\ A\cap\bar{A} &= \varnothing\\ A\cup\bar{A} &= U\end{aligned} \tag{5.5.83}$$

There are other very important relations that also applied to Boolean logic (see section Logic Systems page 7104). If we consider three sets A, B, C as shown below:



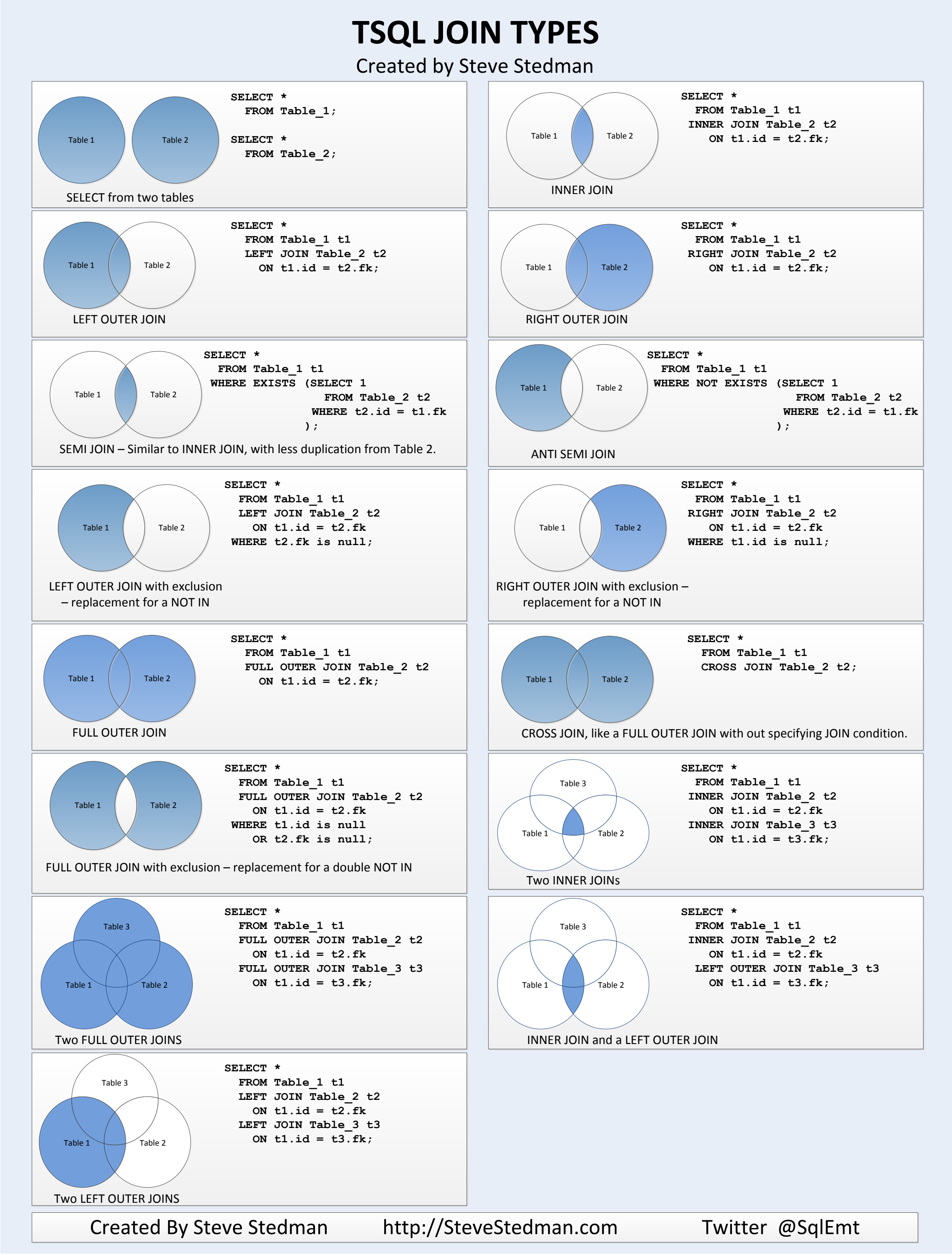
then we have:

$$\begin{aligned} A \setminus (B \cap C) &= (A \setminus B) \cup (A \setminus C) \\ A \setminus B \setminus C &= (A \setminus B) \cup (A \cap B) \end{aligned} \quad (5.5.84)$$

and the famous "De Morgan's laws" in set form (see section Logic Systems page 7110), which are given by the relations:

$$\boxed{\begin{aligned} \overline{A \cap B} &= \bar{A} \cup \bar{B} \\ \overline{A \cup B} &= \bar{A} \cap \bar{B} \end{aligned}} \quad (5.5.85)$$

We would like indicate before moving on to another topic, that a significant number of adults in employment (mostly managers) having forgotten the previous defined concepts after leaving high school must study them again when they learn the SQL language (Structured Query Language) which is the most common worldwide language to query corporate databases servers in the 120th and 121st century (holocene calendar). Most of them learn in training centers the following scheme to build queries with joins:



TSQL JOIN TYPES

Created by Steve Stedman

```
SELECT *
  FROM Table_1;

SELECT *
  FROM Table_2;
```

SELECT from two tables

```
SELECT *
  FROM Table_1 t1
 INNER JOIN Table_2 t2
    ON t1.id = t2.fk;
```

INNER JOIN

```
SELECT *
  FROM Table_1 t1
  LEFT JOIN Table_2 t2
    ON t1.id = t2.fk;
```

LEFT OUTER JOIN

```
SELECT *
  FROM Table_1 t1
 RIGHT JOIN Table_2 t2
    ON t1.id = t2.fk;
```

RIGHT OUTER JOIN

```
SELECT *
  FROM Table_1 t1
 WHERE EXISTS (SELECT 1
                 FROM Table_2 t2
                WHERE t2.id = t1.fk
              );
```

SEMI JOIN – Similar to INNER JOIN, with less duplication from Table 2.

```
SELECT *
  FROM Table_1 t1
 WHERE NOT EXISTS (SELECT 1
                     FROM Table_2 t2
                    WHERE t2.id = t1.fk
                  );
```

ANTI SEMI JOIN

```
SELECT *
  FROM Table_1 t1
  LEFT JOIN Table_2 t2
    ON t1.id = t2.fk
 WHERE t2.fk is null;
```

LEFT OUTER JOIN with exclusion
– replacement for a NOT IN

```
SELECT *
  FROM Table_1 t1
 RIGHT JOIN Table_2 t2
    ON t1.id = t2.fk
 WHERE t1.id is null;
```

RIGHT OUTER JOIN with exclusion –
replacement for a NOT IN

```
SELECT *
  FROM Table_1 t1
  FULL OUTER JOIN Table_2 t2
    ON t1.id = t2.fk;
```

FULL OUTER JOIN

```
SELECT *
  FROM Table_1 t1
 CROSS JOIN Table_2 t2;
```

CROSS JOIN, like a FULL OUTER JOIN with out specifying JOIN condition.

```
SELECT *
  FROM Table_1 t1
  FULL OUTER JOIN Table_2 t2
    ON t1.id = t2.fk
 WHERE t1.id is null
    OR t2.fk is null;
```

FULL OUTER JOIN with exclusion – replacement for a double NOT IN

```
SELECT *
  FROM Table_1 t1
 INNER JOIN Table_2 t2
    ON t1.id = t2.fk
 INNER JOIN Table_3 t3
    ON t1.id = t3.fk;
```

Two INNER JOINs

```
SELECT *
  FROM Table_1 t1
  FULL OUTER JOIN Table_2 t2
    ON t1.id = t2.fk
  FULL OUTER JOIN Table_3 t3
    ON t1.id = t3.fk;
```

Two FULL OUTER JOINS

```
SELECT *
  FROM Table_1 t1
 INNER JOIN Table_2 t2
    ON t1.id = t2.fk
  LEFT OUTER JOIN Table_3 t3
    ON t1.id = t3.fk;
```

INNER JOIN and a LEFT OUTER JOIN

```
SELECT *
  FROM Table_1 t1
  LEFT JOIN Table_2 t2
    ON t1.id = t2.fk
  LEFT JOIN Table_3 t3
    ON t1.id = t3.fk;
```

Two LEFT OUTER JOINS

Created By Steve Stedman http://SteveStedman.com Twitter @SqlEmt

Figure 5.60: Common SQL query expressions with joins

5.5.3 Functions and Applications

Definition (#5.5.83): In mathematics, an "application" (or "function") denoted typically f - in analysis - or A - in linear algebra - is the information of two sets, the departure set E and arrival set F (or "image of E"), and a relation associating each element x of the departure set one and only one element of the arrival set, which we name "image of x by f" in the analysis field we note that $f(x)$ or $f(E)$ to explicit the departure set. We name "images" the elements of $f(E)$ and the elements of E are named the "antecedents".

Then we say that f is an application from E to F denoted:

$$f : E \mapsto F \tag{5.5.86}$$

(remember the first arrow/sagittal diagram presented at the beginning of this section), or we also say that this is an application of arguments in E and values in F.

Remark

The term "function" is often used for applications with scalar numeric values, real or complex, that is to say when the arrival set is $\mathbb{R}$ or $\mathbb{C}$. We speak then of "real function" or "complex function". In the case of vector we prefer to use the word "application" as we already mention it in the definition.

✍**Example:**

Consider the following arbitrary function:

$$f(\vec{x},\vec{p},\vec{d}) = \sum_{i=1}^{n} \frac{p_i x_i}{(1+d_i)^i} \tag{5.5.87}$$

with $\vec{x} \in \mathbb{R}^n$, $\vec{p} \in [0,1]^n$ and $d \in \mathbb{R}_+^n$ then the application will be written:

$$f(\vec{x},\vec{p},\vec{d}) : \mathbb{R}^n \times [0,1]^n \times \mathbb{R}_+^n \mapsto \mathbb{R} \tag{5.5.88}$$

Definitions (#5.5.84):

D1. The "graph" or "plot" (or also named "graphic" or "representative") of an application or function $f : E \mapsto F$ is the subset of the cartesian product $E \times F$ consisting of pairs $(x, f(x))$ for x varying in E. The data of the graph f determines its starting set (by projection on the first argument often denoted x) and image (projection on the second argument often denoted y).

D2. If the triplet $f(E, F, \Gamma)$ is a function where E and F are two sets and $\Gamma \subset (E \times F)$ is a graph, E and F are the source and purpose of f respectively. The "definition domain" or "departure set" of f is:

$$D_f = I = \{x \in E \mid \exists y \in F, (x, y) \in \Gamma\} \tag{5.5.89}$$

D3. Given three non empty sets E, F, G, any function of $E \times F$ to G is named a "composition law" of $E \times F$ with values in G.

D4. An "internal composition law" (or simply "internal law") in E is a composition law of $E \times E$ with values in E (that is to say this is the case $E = F = G$).

Remark

The subtraction in $\mathbb{N}$ is not an internal composition law although it is part of the four basic high-school arithmetic operators. But the addition in $\mathbb{N}$ is such an internal law.

D5. An "external composition law" (or simply "external law") in E is a composition law of $F \times E$ with values in E, where F is a separate set of E. In general, F is a set, named "scalar set".

✍**Example:**

In the case of a vector space (see definition much lower) the multiplication of a vector (whose components are based on a given set) by a real scalar is an example of external composition law.

Remark

An external composition law with values in E is also named "action of F on E". The set F is then the field operators. They also say that F operates on E (keep in mind the example of the vectors mentioned above).

D5. We name "image of f", and note $\text{Im}(f)$, the subset defined by:

$$\boxed{\text{Im}(f) = f(E) = \{y \in E \mid \exists x \in E, y = f(x)\}} \tag{5.5.90}$$

Thus, "the image" of a function $f : E \mapsto F$ is the collection of $f(x)$ for x browsing E. It is a subset of F.

And we name "ker of f", and we note $\ker(f)$, the very important subset in mathematics defined by:

$$\boxed{\ker(f) = f(\{0\}) = \{x \in E \mid f(x) = 0\}} \tag{5.5.91}$$

According to the figure (you must deeply understand this concept because we will reuse the ker many times to prove theorems that have important practical applications later in various chapters):

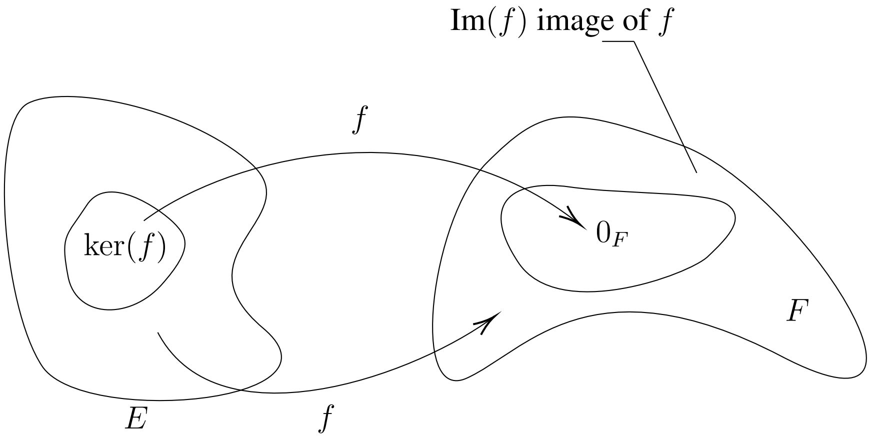


Figure 5.61: **ker concept of a function**

Remarks

R1. $\ker(f)$ is derived from the German "Kern", simply meaning "kernel".

R2. Note that $\ker(f) \subset \ker\left(f^2\right)$ because $f(x) = 0$ implies $f(f(x)) = 0$ (and same for any power n of f^n). It follows that $\ker(f) = \ker\left(f^n\right)$ (i.e. $\ker(f^m) = \ker(f^{m+1})$).

R3. Normally the notations Im and ker are reserved for group homomorphisms, rings, fields and to linear applications between vector spaces and modules, etc. (see further below). We do not usually use them for any applications between any sets. But ... it does not really matter for the moment at this level of the book.

✍**Examples:**

E1. The sine function has from its argument a kernel that is $\ker(\sin(x)) = 2\pi\mathbb{Z}$.

E2. We want to determine the kernel of the matrix:

$$A = \begin{bmatrix} 1 & 2 & 4 \\ 2 & 3 & 3 \end{bmatrix} \tag{5.5.92}$$

For this purpose we transform the matrix into the following system:

$$\begin{cases} x + 2y + 4z = 0 \\ 2x + 3y + 3z = 0 \end{cases} \tag{5.5.93}$$

Simplifying, we get:

$$\begin{cases} x + 2y + 4z = 0 \\ \quad - y - 5z = 0 \end{cases} \tag{5.5.94}$$

Hence:

$$\begin{cases} x - 6z = 0 \\ y + 5z = 0 \end{cases} \tag{5.5.95}$$

Finally:

$$\ker(A)=\{(6z,-5z,z)\mid z\in\mathbb{R}\}=\begin{pmatrix}6\\-5\\1\end{pmatrix} \tag{5.5.96}$$

Applications and functions can have a phenomenal amount of properties. Below you can find some easy one that are part of the general knowledge of the physicist (for more information about what a function is, see the section on Functional Analysis page 2170).

Let f be an application or function of a set E to a set F then we have the following properties:

P1. An application or function is said to be "surjective" if:

Any element y of F is the image by f of at least (we emphasize on the "at least") an element of E. We thus say that it is a "surjection" from E to F. It follows from this definition, that an application or function $f:E\mapsto F$ is surjective and denoted:

$$\boxed{f:E\twoheadrightarrow F} \tag{5.5.97}$$

if and only if $F=\text{Im}f$. In other words, we also write this definition as following:

$$\forall y\in F,\exists x\in E:\quad y=f(x) \tag{5.5.98}$$

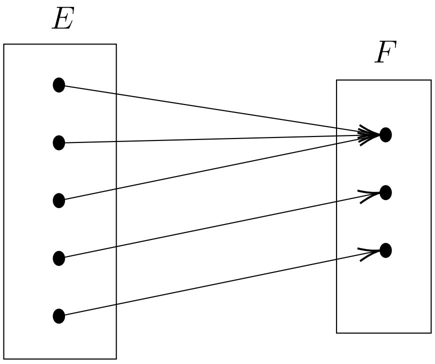


Figure 5.62: **Schematic representation of a surjective application or function**

P2. An application or function is said to be "injective" if:

Any element y of F is the image by f of at most (we emphasize the "at most") a single element of E. We thus say that f is an injection of E to F. It follows from this definition, that an application or function $f:E\mapsto F$ is injective and denoted:

$$\boxed{f:E\rightarrowtail F} \tag{5.5.99}$$

if and only if the relations $x_1,x_2\in E$ and $f(x_1)=f(x_2)$ involve. In other words: an application or function for which two separate elements have distinct images is say to be "injective". Or an application or function is injective at least if one of the following equivalent properties holds:

P2.1 $\forall x,y \in E^2: \; f(x)=f(y) \Rightarrow x=y$

P2.2 $\forall x,y: \; x \neq y \Rightarrow f(x) \neq f(y)$

P2.3 $\forall y \in F$ the equation in x, $y=f(x)$ has at least one solution in E

All this can be resumed by:

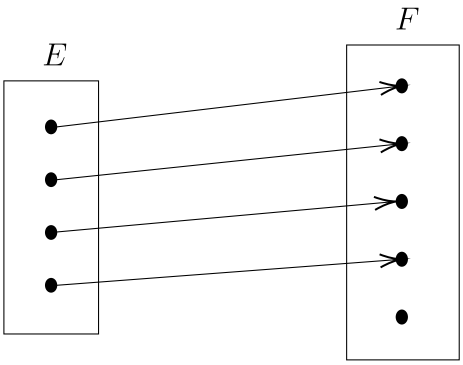


Figure 5.63: **Schematic representation of an injective application or function**

P3. An application or function is said to be "bijective" or "total application/function" if:

An application or function f from E to F is both injective and surjective. In this case, we have that for any element y of F, the equation $y=f(x)$ admits in E a single (not "at least" or not "at most") pre-image x and denoted:

$$\boxed{f: E \leftrightarrow F} \tag{5.5.100}$$

In other words, we also write this definition as following:

$$\forall y \in F, \exists ! x \in E: \quad y=f(x) \tag{5.5.101}$$

This is illustrated by:

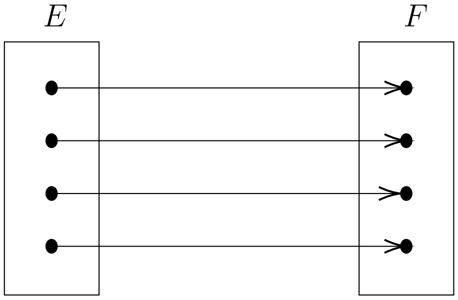


Figure 5.64: **Schematic representation of a bijective application or function**

We are thus naturally led to define a new application from F to E, named "inverse function" or "reciprocal function" of f and denoted f^{-1} (some authors and teachers use the notation ${}^r f$ instead...) that to every element of F matches the unique pre-image element x of E (also named sometimes "solution") of the equation $y=f(x)$. In other words:

$$x=f^{-1}(y) \tag{5.5.102}$$

The existence of an inverse (reciprocal) function or application implies that the graph of a bijective function or application (in the set of real numbers...) and that of its inverse (reciprocal) are symmetric with respect to the right of equation $y = x$.

Indeed, we notice that if $y = f(x)$ is equivalent to $x = f^{-1}(y)$, then these equations imply that the point (x, y) is on the graph of f if and only if the point (y, x) is the graph of equation f^{-1}.

As we can see for example in the figure below with the sinus function (see section Trigonometry page 2503):

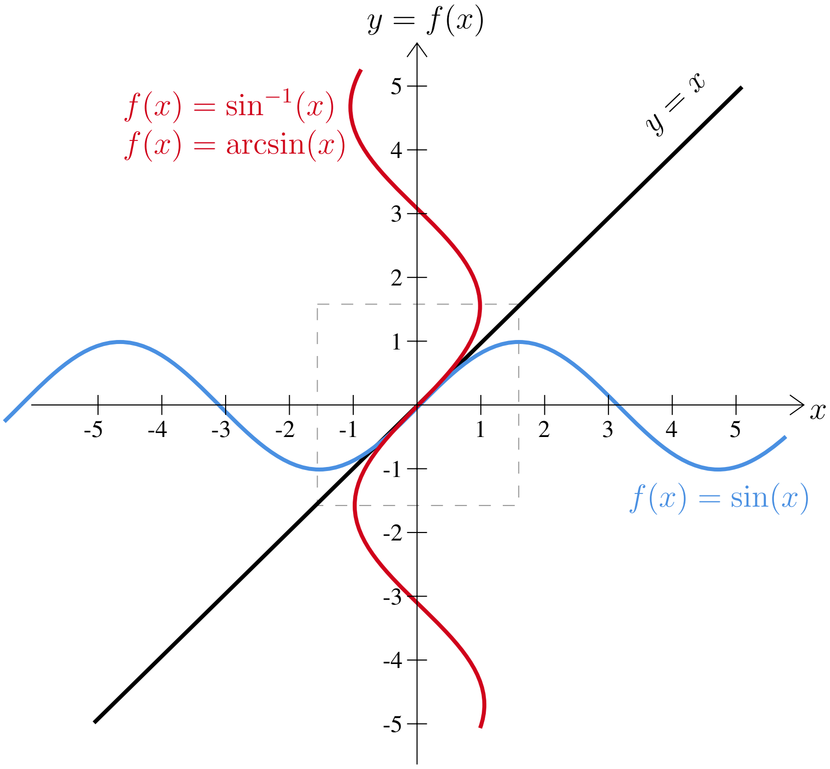


Figure 5.65: **Bijective function example**

✍**Example:**

Take the case of a holiday station where a group of tourists must be housed in a hotel. Each way to allocate these tourists in hotel rooms may be represented by an application of all tourists to all the rooms (to each tourist is assigned a room).

- Tourists want the application to be injective, that is to say, each of them has a single room. This is only possible if the number of tourists does not exceed the number of rooms.
- The hotel manager hopes that the application is surjective, that is to say, each room is occupied. This is only possible if there are at least as many tourists than rooms.
- If it is possible to spread the tourists so that there is only one per room, and all the rooms are occupied: the application will then be both injective and surjective that is to say bijective.

Remarks

R1. It comes from the definitions above that f is bijective in the set of real numbers if and only if any horizontal line intersects the graph of the function at a single point. This leads us to the second following remark:

R2. An application that satisfies the test of the horizontal line is continuously increasing or decreasing at any point in its domain.

P4. An application or function is named "composite application" or "composite function" if:

Let φ be an application or function from E to F and ψ an application or function of F in G. The application or function that associates to each element x of the set E an element $\psi(\varphi(x))$ of G is named "composed application" of φ and ψ and is denoted by:

$$\psi \circ \varphi \tag{5.5.103}$$

where the symbol "$\circ$" is named "round" (do not confused with the scalar product we will see later in the section of Vector Calculus at page 1821!). Thus, the above relation is written "psi round phi" but has to be read "phi round psi" (...). So:

$$\boxed{(\psi \circ \varphi)(x) = \psi(\varphi(x))} \tag{5.5.104}$$

Let, moreover, χ be an application (not a function!) of G in H. We check immediately that the composition operation is associative for applications (for more details see the section of Linear Algebra page 1947):

$$\chi \circ (\psi \circ \varphi) = (\chi \circ \psi) \circ \varphi \tag{5.5.105}$$

This allows us to omit parentheses and write more simply:

$$\chi \circ \psi \circ \varphi \tag{5.5.106}$$

In the particular case where φ would be an application or function from E to E, we note φ^k the composed application $\varphi \circ \varphi \circ \ldots \circ \varphi$ (k times).

What's important in what we have seen until now in this section is that all defined properties listed above are applicable to numbers' sets.

Let us see a concrete and very powerful example:

5.5.3.1 Cantor-Bernstein Theorem

Caution!! This theorem, for which the result may seem trivial, is not necessarily easy to approach (its mathematical formalism is not very aesthetic...). We advise you to read the proof slowly and imagine the sagittal diagrams in your head during the developments.

Here is the hypothesis to prove:

Theorem 5.30. Let X and Y be two sets. If there is an injection (remember the definition of an injective function or application above) from X to Y and another from Y to X, then both sets are in bijection (remember the definition of a bijective function or application above). It is therefore an antisymmetric relation.

This is illustrated by:

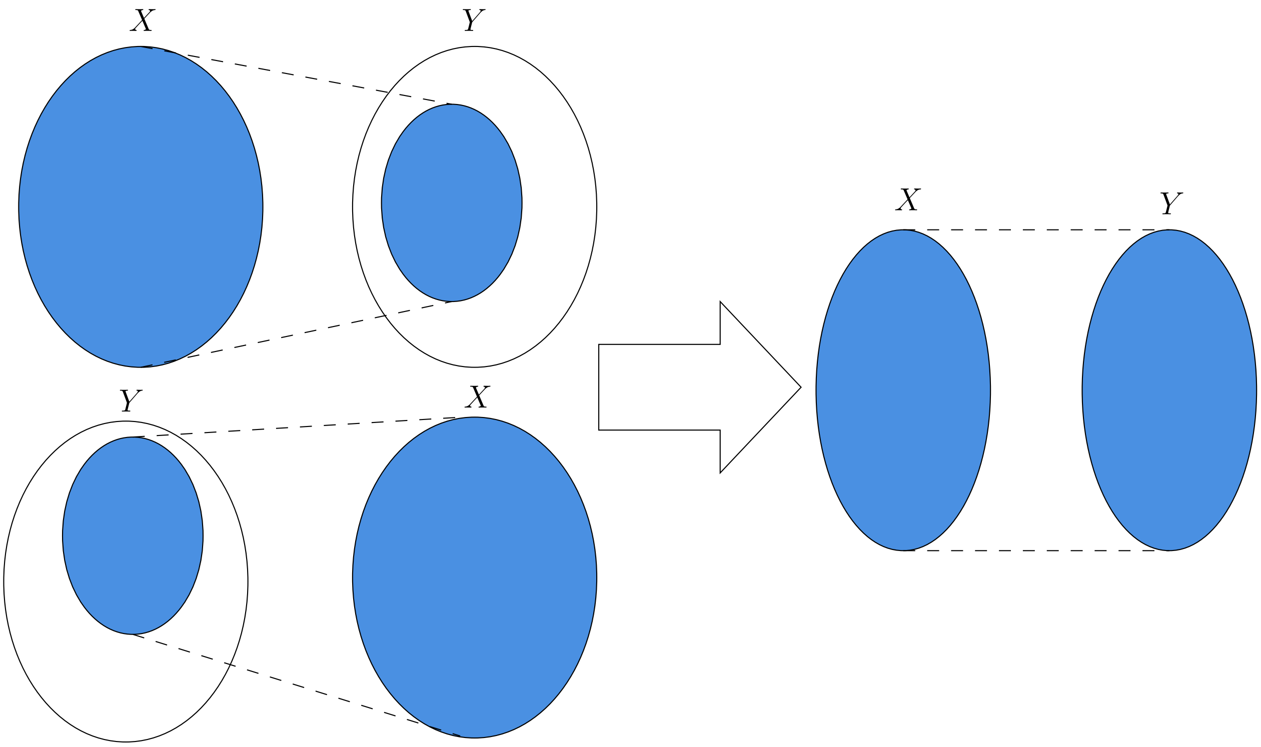


Figure 5.66: **Representation of an antisymmetric relation**

Formally this theorem is sometimes written:

$$\text{Card}(Y) \leq \text{Card}(X) \quad \text{and} \quad \text{Card}(X) \leq \text{Card}(Y) \Leftrightarrow \text{Card}(X) = \text{Card}(Y) \tag{5.5.107}$$

or more technically:

$$(X \rightarrowtail Y) \wedge (Y \rightarrowtail X) \Leftrightarrow X \leftrightarrow Y \tag{5.5.108}$$

For the proof we need rigorously to demonstrate beforehand a lemma (intuitively obvious... but not formally) who's statement is as follows:

Lemma 5.30.1. Let X, Y, Z three sets such that $X \subseteq Z \subseteq Y$. If X and Y are in bijection through a function f, then X and Z are in bijection through a function g.

Technically, we write this:

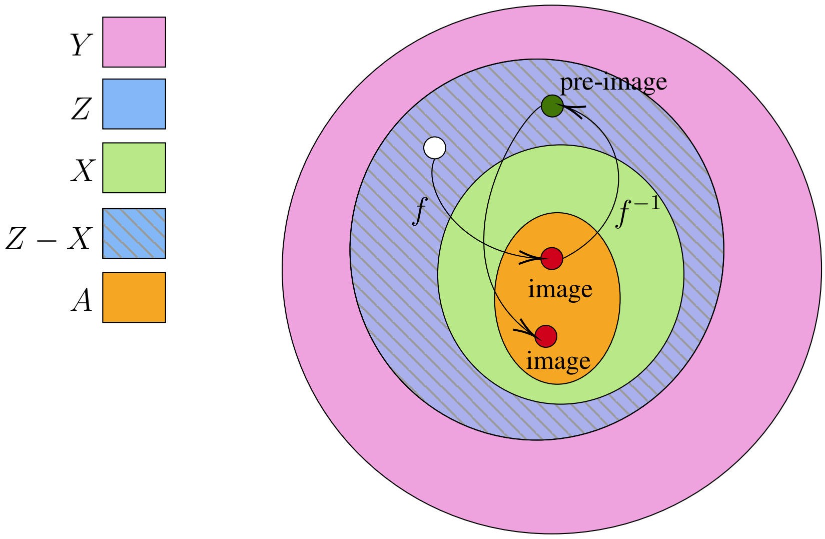
$$\left.\begin{array}{l} X \subseteq Z \subseteq Y \\ f : X \leftrightarrow Y \end{array}\right\} g : X \leftrightarrow Z \tag{5.5.109}$$

An example of application of this lemma is the set of natural numbers $\mathbb{N}$ and rational numbers $\mathbb{Q}$ which are in bijection (see the section of Number Theory page 225 for the proof). Therefore, all the rational numbers are in bijection with the set of natural numbers since $\mathbb{N} \subseteq \mathbb{Z} \subseteq \mathbb{Q}$.

Proof 5.30.1. First, formally, we define a function f from Y to X such that it is bijective:

$$f : Y \leftrightarrow X \tag{5.5.110}$$

To continue we need a set A that will be defined by the union of the images of the functions of the functions f (of the kind $f(f(f\ldots)))$... and build such a tool is the trick of the proof!) of the pre-images of the set Z (remember that $Z \subseteq Y$) from which we exclude the elements of X (that we will be noted for this proof: $Z - X$):



In other words (if the first form is not clear...) we define the set A as the union of images of $(Z - X)$ by the applications $f \circ f \circ \ldots \circ f$. What we write in a condensed a pretty way as following:

$$A = \bigcup_{n=1}^{+\infty} f^n(Z - X) \tag{5.5.111}$$

Because $f: Y \mapsto X$ and that $(Z - X) \subseteq Y$ we have by construction $A \subseteq X$ and thus:

$$((Z-X) \cup A) \subseteq Z \tag{5.5.112}$$

Notice that we also have:

$$A = \bigcup_{n=1}^{+\infty} f^n(Z-X) \Rightarrow f(A) = f\left(\bigcup_{n=1}^{+\infty} f^n(Z-X)\right) = \bigcup_{n=1}^{+\infty} f\left(f^n(Z-X)\right) = \bigcup_{n=1}^{+\infty} f^{n+1}(Z-X) \tag{5.5.113}$$

and by reindexing:

$$f(A) = \bigcup_{n=2}^{+\infty} f^n(Z-X) \tag{5.5.114}$$

We then have (make a pattern in your head of the arrow diagrams can help at that level of the proof...) whatever A:

$$f((Z-X) \cup A) = A \tag{5.5.115}$$

We can elegantly prove this last equality as this is one of the main result (!):

$$f((Z-X) \cup A) = f(Z-X) \cup f(A) = f(Z-X) \cup \bigcup_{n=2}^{+\infty} f^n(Z-X) = \bigcup_{n=1}^{+\infty} f^n(Z-X) \tag{5.5.116}$$

Since Z can be partitioned (nothing prevents us to do this!) in two disjoint subsets (we can draw a figure on request):

$$(Z-X) \cup A \qquad \text{and} \qquad (X-A) \tag{5.5.117}$$

and without forgetting that $X \subseteq Z \subseteq Y$ and $A \subseteq X$, now we introduce as definition the function g (we don't give more information about it yet) such that:

$$g: Z \mapsto X \tag{5.5.118}$$

and for every pre-image a of g of the partition $((Z-X) \cup A) \subseteq Z$ we have:

$$\forall a \in ((Z-X) \cup A) \to f(a) \tag{5.5.119}$$

This means that because $((Z-X) \cup A) \subseteq Z$ and $Z \subseteq Y$ we can thus apply (associate) the bijective function f (remember that $f: Y \to X$) as equivalent of the function g to any element of $((Z-X) \cup A)$.

We also have also for every pre-image a of g of the partition $(X-A)$ (remember that $A \subseteq X$):

$$\forall a \in (X-A) \to a \tag{5.5.120}$$

this means that we just apply the identity function we could also associate it to g.

So to sum up, we have build a function g that is then bijective because its restrictions to the $((Z-X) \cup A)$ is f and that to $(X-A)$ is the identity which are both bijective, the first by definition, the second by construction!

Finally there exists, by construction, a bijection between X and Z and we have proved indeed, the lemma that:

$$\boxed{\left.\begin{array}{l} X \subseteq Z \subseteq Y \\ f : X \leftrightarrow Y \end{array}\right\} g : X \leftrightarrow Z} \tag{5.5.121}$$

■ Q.E.D.

Now that we have proved the Lemma let us recall the assumptions of the Cantor-Bernstein theorem using the result of the Lemma:

Consider φ an injection from X to Y and ψ an injection for Y to X with $X \subseteq Y$.

We thus have:

$$\varphi(X) \subseteq Y \qquad \text{and} \qquad \psi(Y) \subseteq X \tag{5.5.122}$$

Hence:

$$\psi(\varphi(X)) \subset \psi(Y) \subset X \tag{5.5.123}$$

Then so far we have:

- As φ is injective, then X and $\varphi(X)$ are by definition bijective (yes! indeed, an injective function is by definition bijective when we reduce its image set to the pre-image set[17])
- As ψ is injective, $\psi(\varphi(X))$ and $\varphi(X)$ are bijective.

Therefore from the previous proved lemma, X and $\psi(\varphi(X))$ are also in bijection!

By using also the lemma on $\psi(\varphi(X))$, $\psi(Y)$ and X, it follows that $\psi(\varphi(X))$ is the also in bijection with $\psi(Y)$ which gives us with what have just seen, that since $\psi(\varphi(X))$ and $\varphi(X)$ are in bijection, that $\psi(Y)$ is in bijection with $\varphi(X)$, then X and Y are indeed related by a bijective application (phew!!! it is a beautiful reasoning but it is as vicious as simple...) and we have:

$$\boxed{(X \rightarrowtail Y) \wedge (Y \rightarrowtail X) \Leftrightarrow X \leftrightarrow Y} \tag{5.5.124}$$

This theorem can obviously be interpreted in the following way: If we can count a part of a set with all the elements of another set, and vice versa, then they have the same number of elements. This is quite obvious for finished sets. The theorem above then generalizes this notion for infinite sets and this is its strength!

From there, this theorem represents one of the basic bricks to generalize the notion of set sizes to infinite sets.

---

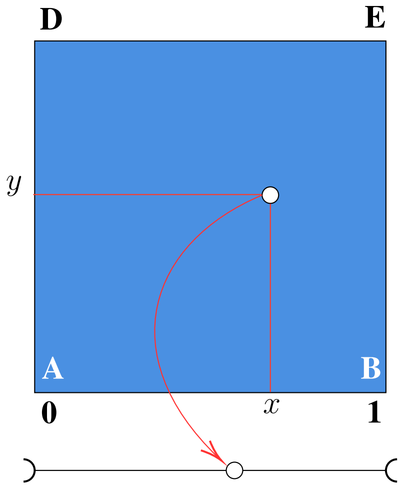
[17] but it must be notice that this work only for infinite sets!

✍**Example:**

The line is a geometric object of dimension 1, like its intervals. The plane has a dimension equal to 2. Its nature is therefore very different from that of the line. However, these two sets of points are still equipotent! In other words the open interval $]0,1[$ has the same power as the open square $]0,1[\times]0,1[$.

The only technical point to take into consideration is the following: Any real number is perfectly determined by the data of its proper decimal development.

Let a point of the open square with coordinates (x,y) with $x=0.x_0x_1x_2\ldots$ and $y=0.y_0y_1y_2\ldots$. We associate with it the point $z=0.x_0y_0x_1y_1x_2y_2\ldots$ of the open interval $]0,1[$.



The application f thus defined is clearly injective!

Let us now consider a real number z of $]0,1[$ with $z=0.z_0z_1z_2z_3z_4z_5\ldots$ This point is the image of the point of the square whose coordinates are:

$$0.z_0z_2z_4\ldots \text{ and } y=0.z_1z_3z_5\cdots$$

Surjectivity seems to follow from this.

The meticulous reader will probably show the real $0.909090\ldots$ image of the point $(0.999\ldots,0.000\ldots)$, in other words $(1,0)$ which has the notable defect of not belonging to the open square!

We have, in fact, the same problem with all numbers of the type $0.9y_09y_1\ldots$ "images" of the "segment" $\{1\}\times]0,1[$ (the "open" segment $]BC[$ in the figure above).

Our mapping f from the interval $]0,1[$ to the open square $]0,1[\times]0,1[$ is therefore injective but not surjective! It is therefore not bijective, contrary to our first expectation. In a way, this result is even more surprising than the one announced. That one can "inject" a two-dimensional object "into" and not "onto" an absolutely flat object defies all intuition! «*I see it, but I don't believe it*» in Cantor's own words.

So there is an injection of the open square into $]0,1[$.

The Cantor-Berstein theorem saves the day! As we proved just before, this theorem states that if there exists an injection of X onto Y and an injection of Y onto X, then there exists a bijection from X onto Y:

$$(X \mapsto Y) \wedge (Y \mapsto X) \Leftrightarrow X \leftrightarrow Y$$

We can conclude: the open interval $]0,1[$ and the open square $]0,1[\times]0,1[$ are equipotent!

5.5.4 Structures

The so-named "modern algebra" or "abstract algebra" begins with the theory of algebraic structures due in part to Carl F. Gauss and especially to Évariste Galois. These structures exist in a very large number but only the fundamentals will interest us here (this textbook being mainly dedicated, for recall, to engineers). Before describing them, here is a synoptic diagram of these main structures and their hierarchy:

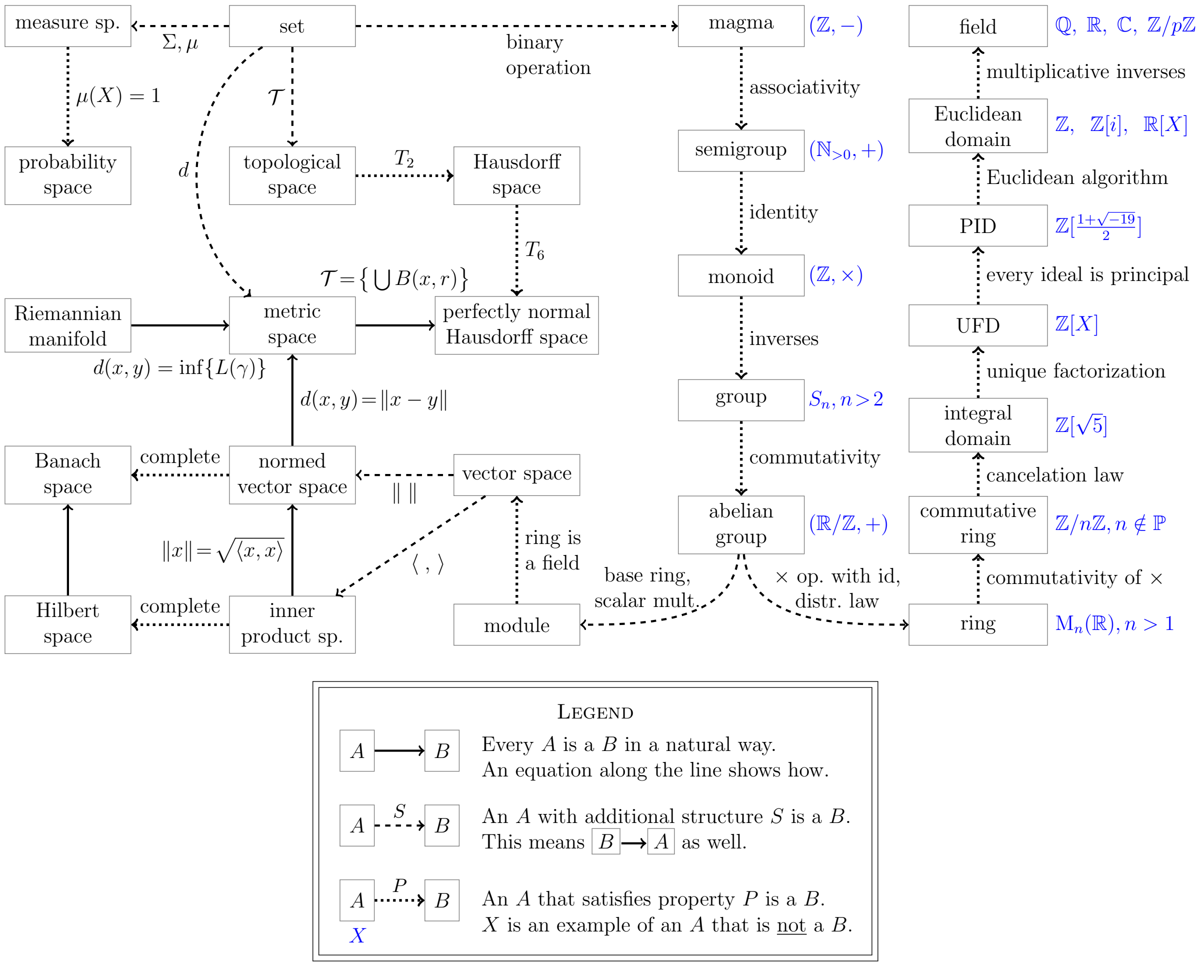

measure sp.
set
magma
$(\mathbb{Z}, -)$
field
$\mathbb{Q},\ \mathbb{R},\ \mathbb{C},\ \mathbb{Z}/p\mathbb{Z}$
Σ, μ
binary operation
$\mu(X)=1$
$\mathcal{T}$
associativity
multiplicative inverses
Euclidean domain
$\mathbb{Z},\ \mathbb{Z}[i],\ \mathbb{R}[X]$
probability space
d
topological space
T_2
Hausdorff space
semigroup
$(\mathbb{N}_{>0}, +)$
Euclidean algorithm
identity
PID
$\mathbb{Z}[\frac{1+\sqrt{-19}}{2}]$
T_6
$\mathcal{T}=\{\bigcup B(x,r)\}$
monoid
$(\mathbb{Z}, \times)$
every ideal is principal
Riemannian manifold
metric space
perfectly normal Hausdorff space
inverses
UFD
$\mathbb{Z}[X]$
$d(x,y)=\inf\{L(\gamma)\}$
unique factorization
$d(x,y)=\|x-y\|$
group
$S_n, n>2$
integral domain
$\mathbb{Z}[\sqrt{5}]$
Banach space
complete
normed vector space
vector space
commutativity
cancelation law
$\|\ \|$
abelian group
$(\mathbb{R}/\mathbb{Z}, +)$
commutative ring
$\mathbb{Z}/n\mathbb{Z}, n \notin \mathbb{P}$
$\|x\|=\sqrt{\langle x,x\rangle}$
ring is a field
$\langle\ ,\ \rangle$
base ring, scalar mult.
× op. with id, distr. law
commutativity of ×
Hilbert space
complete
inner product sp.
module
ring
$\mathrm{M}_n(\mathbb{R}), n>1$
LEGEND
A B
Every A is a B in a natural way.
An equation along the line shows how.
A S B
An A with additional structure S is a B.
This means $B \longrightarrow A$ as well.
A P B
X
An A that satisfies property P is a B.
X is an example of an A that is not a B.

Figure 5.67: **Synoptic diagram of common algebraic structures (author: ?)**

Remark

At the top of the diagram, the structure has the minimum number of constraints (exactly one), at the bottom, a maximum (to seven). The more we descend, the more specialized the structure is.

To simplify the notations, let us assume that $\star$ and $\circ$ represents a composition[18] like (like the addition, subtraction, multiplication or division, ...), then:

Definitions (#5.5.85): Given $\star$ and $\circ$ the symbols of internal laws (this could be addition and multiplication to take the best known case) to a given set E then:

---

[18] This generalized notation is sometimes named "stellar notation".

D1. $\star$ is a "commutative law" if:

$$a \star b = b \star a \tag{5.5.125}$$

D2. $\star$ is an "associative law" if:

$$a \star (b \star c) = (a \star b) \star c \tag{5.5.126}$$

D3. n is the "neutral element" for $\star$ if:

$$a \star n = n \star a = a \tag{5.5.127}$$

We will also admit without proof (it is quite intuitive) that if there is a neutral element, then it is unique.

D4. a' is the "symmetrical element" (in the general sense of the "opposite", for example for addition and "inverse" for multiplication) of a for $\star$ if:

$$a' \star a = a \star a' = n \tag{5.5.128}$$

We shall also admit, and without proof, that the symmetric of any element is unique.

D5. $\circ$ is a "distributive law" with respect to $\star$ if:

$$\begin{gathered} a \circ (b \star c) = (a \circ b) \star (a \circ c) \\ (a \star b) \circ c = (a \circ b) \star (b \circ c) \end{gathered} \tag{5.5.129}$$

D6. b is the "absorbing element" if for all a and a law $\star$ we have:

$$a \star b = 0 \tag{5.5.130}$$

Remarks

R1. If a is its own symmetric with respect to the law $\star$, mathematicians say that a is "involutive".

R2. If an element b of E satisfies:

$$a \star b = b \star a = b \tag{5.5.131}$$

then b is say to be an "absorbing element" for the law $\star$.

R3. It must always be checked that the neutral and the symmetrical elements are such on the "left" and on the "right". Thus, for example, in $(\mathbb{Z}, -)$, the element 0 is neutral to the right because $x - 0 = x$ but $0 - x = -x$.

5.5.4.1 Magma

Definition (#5.5.86): We denote a set by the term "magma M", if the components constituting it are operable with respect to an internal law $\star$:

$$(M,\star) \text{ is a magma} \Longleftrightarrow \begin{cases} \star \text{ is an operation} \\ \star \text{ is an internal law} \end{cases} \tag{5.5.132}$$

It is therefore important to remember that if we designate an algebraic structure by the term "magma", it does not mean in any case that the internal law is commutative, associative or even that it possesses a neutral element!

Remarks

R1. If, moreover, the internal law $\star$ is commutative, we speak of "commutative magma".

R2. If, moreover, the internal law $\star$ is commutative, we speak of "associative magma".

R3. If, moreover, the internal law $\star$ possesses a neutral element n, we speak sometimes of "unitary associative magma" or respectively "unitary commutative magma" but we will see just below that the both have in practice other official names.

Definition (#5.5.87): In a magma $(M,\star)$, an element x is named "regular element" (or "simplifiable element") to the left if for any pair $(a,b)\in M$ we have:

$$x\star a = x\star b \Longrightarrow a=b \tag{5.5.133}$$

Remark

We also define following the same logic a regular element to the right.

Thus, an element is say to be "regular" if it is regular to the right and to the left. If $\star$ is commutative (which is the case for a commutative magma), the notions of regular element to the left or to the right coincide.

A magma $(M,\star)$ is therefore an elementary algebraic structure. There are more subtle structures (monoids, groups, rings, fields, vector space, etc.) in which a set is provided with several laws and different properties. We will see them right now and use them throughout this textbook explicitly or implicitly.

5.5.4.2 Monoid

Definition (#5.5.88): If the law $\star$ is associative and has a neutral element n, then we say that the "unitary associative magma" is a "monoid":

$$(\ddot{M},\star)\text{ is a monoid} \quad\Longleftrightarrow\quad \begin{cases}\star\text{ is associative}\\ \exists\text{ a neutral element } n\in\ddot{M}\text{ for the law }\star\end{cases} \tag{5.5.134}$$

Remarks

R1. If the internal law $\star$ is commutative then we say that the structure forms an "abelian monoid" (or simply "commutative monoid").

R2. In some textbooks we also find that the monoid is a "semi-group" (with an associative $\star$ law!) that has a neutral element n.

Let us show immediately that the set of natural integers $\mathbb{N}$ is a totally ordered abelian monoid (as we have partially seen it in the section Operators page 288) with respect to the laws of addition and multiplication:

The addition law "+" is an internal operation such that $\forall a,b\in\mathbb{N}$ we have:

$$a+b=c \tag{5.5.135}$$

We can prove that this is indeed the case by knowing that 1 belongs to $\mathbb{N}$ such that:

$$\sum_{i=1}^{a}1+\sum_{i=1}^{b}1=\sum_{i=1}^{a+b}1 \tag{5.5.136}$$

Therefore $c\in\mathbb{N}$ and the addition is indeed an internal law (we also say that the set $\mathbb{N}$ is "stable" with respect to addition) and at the same time associative since 1 can be added to itself by definition in any order without the result being altered. If you remember that multiplication is a law that is built on addition (see section Operators page 308), then the multiplication law $\times$ is also an internal and associative law!

We will assume starting from here that it is trivial that the addition law $+$ is also commutative and that the zero "0" is the neutral element n. Thus, the multiplication law $\times$ is also commutative and it is trivial that "1" is the neutral element n.

On the other hand, to speak about something that is not directly related to the monoids... but that will be useful to us a little further below, does it exist in line with the previous example for the addition law $+$ a symmetric $\exists c$ such that $\forall a,b\in\mathbb{N}$ we have:

$$\left(\sum_{i=1}^{a}1+\sum_{i=1}^{b}1\right)+c=\left(\sum_{i=1}^{a+b}1\right)+c=0 \tag{5.5.137}$$

with $c\in\mathbb{N}$?

It is quite trivial that for this equality to be satisfied we must have:

$$\left(\sum_{i=1}^{a+b}1\right)=-c \tag{5.5.138}$$

thus:

$$a+b=-c \tag{5.5.139}$$

but negative numbers do not exist in $\mathbb{N}$. This also leads us to the conclusion that the addition law $+$ has no symmetric element in $\mathbb{N}$ and that the subtraction law $-$ does not exist in $\mathbb{N}$ (the subtraction being rigorously the addition of a number negative!).

Similarly, since it will also be useful to us a little further below, is there a symmetric a' for the multiplication law $\times$ such that $\forall a \in \mathbb{N}$ we have:

$$a' \cdot a = n = 1 \tag{5.5.140}$$

with $a' \in \mathbb{N}$?

First it is obvious that:

$$a' = \frac{1}{a} \tag{5.5.141}$$

But excepted for $a = 1$, the quotient $1/a$ does not exist in $\mathbb{N}$. Therefore we must conclude that there do not exist for any element of $\mathbb{N}$ for the multiplication law $\times$ and therefore that the division law $/$ does not exist in $\mathbb{N}$ and that the multiplication law does not form a monoid in this set .

Synthesis:

$\mathbb{N}$	$(+)$	$(-)$	$\times$	$/$
Internal operation (closure)	yes	no	yes	no
Commutative	yes		yes	
Neutral Element	yes (0)		yes (1)	
Absorbing element	no		yes (0)	
Symmetrical	no		no	

Table 5.21: **Laws and their properties in $\mathbb{N}$**

We have, for example, the following properties with respect to the set of natural integers $\mathbb{N}$ and the concept of monoid:

P1. $(\mathbb{N}, \leq, \geq)$ is completely ordered (noticed that this notation is somewhat abusive, it suffices that there is just one of the two order relations $\mathcal{R}$ so that the set is totally ordered).

P2. $(\mathbb{N}, +)$ and $(\mathbb{N}, \times)$ are abelian monoids.

P3. The element zero "0" is the absorbing element for the monoid $(\mathbb{N}, \times)$.

P4. The laws of subtraction $-$ and division $/$ do not exist in the set $\mathbb{N}$.

P5. $\mathbb{N}$ is an abelian monoid totally ordered with respect to the laws of addition $+$ and multiplication $\times$ (caution! the following notation is abusive because the monoid is composed of only one internal law and a unique order relation $\mathcal{R}$ which would give in total 4 unique monoids):

$$(\mathbb{N}, +, \leq, \geq) \tag{5.5.142}$$

Remarks

R1. It is rare to use monoids in daily work of the engineer or physicist. Indeed, this is because often when we find ourselves faced with a structure too poor to really discuss about something or develop a physical model. Then we extend it to something richer, like a group, or a field (see below) such as the set of relative integers $\mathbb{Q}$ or of real number $\mathbb{R}$ (at least...).

R2. Say that an algebraic structure is totally ordered with respect to some laws means that given a law $\star$, and $\mathcal{R}$ an order relation and a, b, c, d four elements of the structure concerned, then if $a\ \mathcal{R}\ b$ and $c\ \mathcal{R}\ d$ imply $(a\star c)\mathcal{R}(b\star d)$, we denote this structure $(S,\star,\mathcal{R})$ or simply $(S,\mathcal{R})$ and indicating the concerned law later in the text.

5.5.4.3 Groups

Definition (#5.5.89): We designate a set by the term "group", if the constituent components satisfy the three conditions of what we name the "internal group law", defined below:

$$\boxed{(G,\star)\text{ is a group}\quad\Longleftrightarrow\quad\begin{cases}\star\text{ is associative}\\ \exists\in G\text{ a neutral element for }\star\\ \forall x\in G\text{ has a symmetrical (inverse) for }\star\end{cases}}\tag{5.5.143}$$

In this case, the law of internal composition $\star$ will often (but not exclusively!) be denoted "$+$" and named "addition", the neutral element is denoted e and its value is equal to "0" and the symmetric of x is denoted "$-x$".

Let us emphasize that group structure is probably one of the most important in the practice of modern engineering and physics in general. This is why it is necessary to pay special attention to it (see section Set Algebra page 421)!

If, moreover, the internal law $\star$ is also commutative, then we say that the group is an "abelian group" or simply a "commutative group".

If there exists in G at least one element a such that every element of G is a power of$a$ or of the symmetric a' of a, we say that $(G,\star)$ is a "cyclic group of generator a" if it is finite, otherwise say that it is "monogenic" (we will come back in details on cyclic groups in the section of Set Algebra page 1418).

More generally, a group $(G,\star)$ of neutral element e, not reduced only to $\{e\}$ will be monogeneous, if there exists an element a of G distinct from e such that $G=\{e,a^1,a^2,a^3,\ldots,a^n,\ldots\}$. Such a group will be cyclic, if there exists a non-zero integer n for which $a^n=e$. The smallest non-zero integer satisfying this equality is then the "order of the group".

Let us show immediately that the set of relative integers $\mathbb{Z}$ is a totally ordered abelian group (as we have seen partially in the section Operators page 288) with respect to the laws of addition $+$ and multiplication $\times$.

First, to shorten the developments, it is useful to recall that the set $\mathbb{Z}$ is an "extension" of $\mathbb{N}$ by the fact that we have added to it all the symmetric negative sign numbers ($\mathbb{N} \subset \mathbb{Z}$).

Thus, always with the abuse of notations (because normally a group has only one law $\star$ and one order relation $\mathcal{R}$ that is sufficient to order it):

$$(\mathbb{Z}, +, -, \leq, \geq) \tag{5.5.144}$$

forms a completely ordered abelian group (4 groups by the way!) and:

$$(\mathbb{Z}, \times, \leq, \geq) \tag{5.5.145}$$

a completely ordered abelian monoid (two monoids in facts!).

Let us also notice that the law of division does not exist for any element of the set $\mathbb{Z}$! So in general we say that the division does not exist in it!

Synthesis:

$\mathbb{Z}$	$(+)$	$(-)$	$(\times)$	$(/)$
Internal operation	yes	yes	yes	no
Associativity	yes	no	yes	
Commutativity	yes	no	yes	
Neutral element	yes (0)	no (0 not neutral to the left)	yes (1)	
Absorbing element	no	no	yes (0)	
Symmetrical	yes (opposed sign)	yes	no	

Table 5.22: **Laws and their properties in $\mathbb{Z}$**

We therefore have the following properties:

P1. $(\mathbb{Z}, \leq, \geq)$ is completely ordered (caution! again this notation is a little abusive! It is enough that there is just one of the two order relation $\mathcal{R}$ so that the set is totally ordered).

P2. $(\mathbb{Z}, +)$ is a commutative group whose zero "0" is the neutral element.

P3. The division law does not exist in the set $\mathbb{Z}$.

P4. The set $\mathbb{Z}$ is an abelian group totally ordered with respect to the law of addition $+$ (caution! the following notation is again abusive because the group is composed only of one order relation $\mathcal{R}$ which would give a total of two groups):

$$(\mathbb{Z}, +, \leq, \geq) \tag{5.5.146}$$

The set $\mathbb{Z}$ set is not a commutative group totally ordered with respect to the law of multiplication:

$$(\mathbb{Z}, \times, \leq, \geq) \tag{5.5.147}$$

We see then immediately that $\mathbb{Z}$ has too restricted properties, that is why it is interesting to extend it by the set of rational $\mathbb{Q}$ defined in a very simplistic way ... by (see section Numbers page 224):

$$\mathbb{Q}=\left\{\frac{p}{q}\middle|(p,q)\in\mathbb{Z}/q=0\right\} \qquad (5.5.148)$$

This means for recall that the set of rationals $\mathbb{Q}$ is defined by the set of quotients p and q belonging each to $\mathbb{Z}$ of which we exclude to q to take the value zero (the notation $/q$ signifying for recall the "exclusion").

And we obviously have:

$$(\mathbb{N},\mathbb{Z})\subset\mathbb{Q} \qquad (5.5.149)$$

It is therefore obvious (without proof and always using the abusive notation already commented many times earlier above...) that $(\mathbb{Q},\leq,\geq)$ is also totally ordered and also that $\mathbb{Q}$ is an abelian group totally ordered with respect to the law of addition $+$ only:

$$(\mathbb{Q},+,\leq,\geq) \qquad (5.5.150)$$

What becomes interesting with $\mathbb{Q}$ is that the law of multiplication becomes an internal law and forms a commutative abelian group named "multiplicative group" with respect to $\mathbb{Q}^*$.

Proof 5.30.2. Let us prove then that the symmetric exists for the law of multiplication $\times$ such that:

$$a\times a=n=1 \qquad (5.5.151)$$

Since in $\mathbb{Q}^*$ any number can be put in the form:

$$a=\frac{p}{q} \qquad (5.5.152)$$

with $p\in\mathbb{Z}$, $q\in\mathbb{Z}^*$.

So since:

$$a'=\frac{1}{a}=\frac{1}{\left(\dfrac{p}{q}\right)}=\frac{q}{p} \qquad (5.5.153)$$

There is therefore a symmetric to every rational in $\mathbb{Q}^*$ for the law of multiplication.

■ Q.E.D.

By definition, or by construction, the division exists in $\mathbb{Q}^*$ and is an internal operation. But is it associative such that $\forall(p,q,r)\in\mathbb{Q}^*$ we have:

$$\frac{\dfrac{p}{q}}{r}=\frac{p}{\dfrac{q}{r}} \qquad (5.5.154)$$

In fact, the verification is fairly trivial if we remember that division is defined from the law of multiplication of the inverse and that the latter law is... associative! So it comes:

$$\frac{\frac{p}{q}}{r}=\left(p\frac{1}{q}\right)\frac{1}{r}=p\left(\frac{1}{q}\frac{1}{r}\right)=p\left(\frac{1}{qr}\right)\neq\frac{p}{\frac{q}{r}}=\frac{p}{q\frac{1}{r}}=p\frac{r}{q} \tag{5.5.155}$$

We can also ask ourselves the division law $(/)$ is commutative such that the relation:

$$\frac{a}{b}=\frac{b}{a} \tag{5.5.156}$$

with $\forall(a,b)\in\mathbb{Q}^*$.

We see very well that this is not the case since we can write this last relation in the form:

$$\frac{a}{b}=\frac{b}{a}=\pm 1\Rightarrow a=\pm b \tag{5.5.157}$$

To sum up:

$\mathbb{Q}$	$(+)$	$(-)$	$(\times)$	$(/)$
Internal law	yes	yes	yes	yes
Associativity	yes	no	yes	no
Commutativity	yes	no	yes	no
Neutral element	yes (0)	no (0 not neutral to the left)	yes (1)	yes (1 neutral to the right)
Absorbing element	no	no	yes (0)	yes (0 at the numerator)
Symmetrical	yes (opposed sign)	yes (opposed sign)	no (excepted in $\mathbb{Q}^*$)	no

Table 5.23: **Laws and their properties in $\mathbb{Q}$**

We therefore have the following properties:

P1. $(\mathbb{Q},\leq,\geq)$ is totally ordered

P2. $(\mathbb{Q},+),(\mathbb{Q}^*,\times)$ are independently totally ordered abelian groups

P3. Zero "0" is the absorbent element with respect to the group $(\mathbb{Q}^*,\times)$

P4. The set $\mathbb{Q}$ is an abelian group totally ordered with respect to the laws of addition and multiplication that we denote:

$$(\mathbb{Q},+,\leq,\geq)\qquad\text{and}\qquad(\mathbb{Q}^*,\times,\leq,\geq) \tag{5.5.158}$$

The same properties are applicable to $\mathbb{R}$ and $\mathbb{C}$ but with the difference that the latter are not orderable.

However, it may be understandable that for $\mathbb{C}$ the reader is sceptical. Let us develop all this:

We must make sure that the sum "+", the difference "$-$", the product "$\times$" and the quotient "$/$" of two numbers of the type $x+\mathrm{i}y$ gives something of the same type again (that is to say an application $\mathbb{C}\mapsto\mathbb{C}$).

Let us add the numbers $a+\mathrm{i}b$ and $c+\mathrm{i}d$ where a, b, c and d are real numbers:

$$\begin{aligned}&(a+\mathrm{i}b)+(c+\mathrm{i}d)\\&\quad=a+\mathrm{i}b+c+\mathrm{i}d\\&\quad=a+c+\mathrm{i}b+\mathrm{i}d\\&\quad=(a+c)+\mathrm{i}(b+d)\end{aligned}\tag{5.5.159}$$

Therefore the addition is indeed a commutative and associative internal law for which there exists a neutral and symmetric element in the set of complex numbers $\mathbb{C}$.

Let us subtracts the numbers $a+\mathrm{i}b$ and $c+\mathrm{i}d$ where a, b, c and d are here again, real numbers:

$$\begin{aligned}&(a+\mathrm{i}b)-(c+\mathrm{i}d)\\&\quad=a+\mathrm{i}b-c-\mathrm{i}d\\&\quad=a-c+\mathrm{i}b-\mathrm{i}d\\&\quad=(a-c)+\mathrm{i}(b-d)\end{aligned}\tag{5.5.160}$$

Therefore the subtraction is an internal law, but it is not commutative, neither associative it has no neutral element to the left and neither symmetric elements.

Let us now multiply the numbers $a+\mathrm{i}b$ and $c+\mathrm{i}b$ where a, b, c and d are still real numbers. To achieve our work, we use the distributivity of multiplication (see section Calculus page 1361) with respect to addition:

$$\begin{aligned}&(a+\mathrm{i}b)(c+\mathrm{i}d)\\&\quad=ac+a\mathrm{i}d+\mathrm{i}bc+\mathrm{i}b\mathrm{i}d\\&\quad=ac+\mathrm{i}ad+\mathrm{i}bc+\mathrm{i}^2bd\\&\quad=ac+\mathrm{i}ad+\mathrm{i}bc+(-1)bd\\&\quad=(ac-bd)+\mathrm{i}(ad+bc)\end{aligned}\tag{5.5.161}$$

Thus the law of multiplication is indeed a commutative, associative and distributive (!) internal operation for which there exists a neutral and symmetric element in $\mathbb{C}^*$ (see below).

A division is before all a multiplication by the inverse. Prove that there exists an inverse proves that there exists a symmetric for multiplication. Let us therefore inverse the number $x+\mathrm{i}y$ where x and y are still real numbers (different from zero!):

$$\begin{aligned}\frac{1}{x+\mathrm{i}y}&=\frac{(x-\mathrm{i}y)}{(x+\mathrm{i}y)(x-\mathrm{i}y)}\\&=\frac{x-\mathrm{i}y}{x^2-(\mathrm{i}y)^2}\\&=\frac{x-\mathrm{i}y}{x^2+y^2}\\&=\frac{x}{x^2+y^2}-\mathrm{i}\frac{y}{x^2+y^2}\end{aligned}\tag{5.5.162}$$

So the inverse of a complex number is indeed an internal non-associative and non-commutative operation for which there exists a neutral element, and it is symmetric. The same is true for division, which corresponds to the product by the inverse of a complex number.

Let us now consider quickly an example of a cyclic group (we will in-deep the subject in the section of Set Algebra page 1418): In $\mathbb{C}$, let us consider $G=\{1,\mathrm{i},-1,-\mathrm{i}\}$ equipped with the usual multiplication of complex numbers. Then $(G,\times)$ is obviously an abelian group. Such a group is also monogenic because it is generated by the powers of one of its elements: i (or $-\mathrm{i}$). This monogenic group being finite, it is then a named a "cyclic group".

5.5.4.4 Ring

The ring is the heart of the commutative algebra which is the algebraic structure corresponding to the high-school concepts of addition, subtraction, and multiplication.

Definition (#5.5.90): A commutative group (or "abelian group") A is a "ring" if it is provided with a second internal law of composition satisfying the following properties:

$$(A,+,\times) \text{ is a ring} \quad \Longleftrightarrow \quad \begin{cases} (A,+) \text{ is an abelian group} \\ \text{The law } \times \text{ is associative} \\ \text{The law } \times \text{ is distributive relatively to the law } + \end{cases} \tag{5.5.163}$$

As we already know, the neutral element of the first internal composition law "$+$" is denoted "0" and named "zero" of the ring. The second internal law $\times$ is often denoted by a mid-height point "$\cdot$" (or by a cross $\times$ in high-school) and named "multiplication".

Remarks

R1. If, moreover, the second internal law of composition "$\times$" is also commutative, the ring is say to be a "commutative ring". We also encounter non-commutative rings in which the commutativity relation is not imposed or just not necessary and then we sometimes have to impose it, then we must reinforce the property of the neutral element of this second law by imposing "1" to be a neutral element both to right and to left such that: $1a=a1=a$ (an example of a non-commutative ring is provided by the set of square $n\times n$ matrices with coefficients in a ring A, for example $M_n(\mathbb{R})$ as seen in the section of Linear Algebra page 1947).

R2. If, in addition, there exists in A a neutral element for the second internal composition law "$\times$", and that this neutral element is the unit "1" then we say that the ring is a "unitary ring" and "1" is named "unit of the ring". If the ring is commutative and has a neutral element for the second internal composition law then we speak of "commutative unitary ring".

R3. If $a\times b=0\Rightarrow(a=0 \text{ or } b=0)$, regardless of the elements a, b of A, the ring is named an "integral ring" or "ring without zero divisors" (if not, it is obviously named "non-integral ring").

R4. A "factorial ring" is a unitary and integral commutative ring in which the fundamental theorem of arithmetic (see section Number Theory page 352) is verified.

Definitions (#5.5.91):

D1. An element a of a ring A is a "unitary element" if there is a $b \in A$ such that $ab = ba = 1$. If such a b exists it is unique (we saw such an example in our study of congruence classes in the section of Number Theory page 356).

D2. An element a of a ring A is named a "left zero divisor" if there exists $x \neq 0$ such that $ax = 0$. Similarly, an element a of a ring is named a "right zero divisor" if there exists $y \neq 0$ such that $ya = 0$. An element a that is both a left AND a right zero divisor is named a "two-sided zero divisor[19]".

✍**Examples:**

E1. The only zero divisor of the ring $\mathbb{Z}$ of integers is 0.

E2. In the ring of 2×2 matrices (over any non-zero ring) we can find and a a and a non zero x as following:

$$\begin{pmatrix}1 & 1\\ 2 & 2\end{pmatrix}\begin{pmatrix}1 & 1\\ -1 & -1\end{pmatrix} = \begin{pmatrix}-2 & 1\\ -2 & 1\end{pmatrix}\begin{pmatrix}1 & 1\\ 2 & 2\end{pmatrix} = \begin{pmatrix}0 & 0\\ 0 & 0\end{pmatrix} \tag{5.5.164}$$

In the latter example it is difficult to choose which of the matrix is the "zero divisor" this is why we can find in some textbooks that a AND x are considered as "zero divisors".

Remarks

R1. It should be obvious from the preceding relations that clear that a ring is integral if and only if it has no zero divisors (excepted the 0 as already mentioned earlier!).

R2. The concepts of unit and zero divisor are incompatible but one element of a ring can be neither. This is the case, for example, of all integers different from $\{0, -1, 1\}$ in $\mathbb{Z}$. These are neither units nor divisors from zero! We name them "regular elements".

We will see an important example of ring in the context of our study of polynomials (see section Calculus page 1368), but we have already seen some very important ones in our study of congruence classes in the section of Number Theory page 356.

Let's see some examples of rings! During our study of groups earlier above, we have seen that the structures:

$$(\mathbb{Z},+) \qquad (\mathbb{Q},+) \qquad (\mathbb{R},+) \qquad (\mathbb{C},+) \tag{5.5.165}$$

are all four abelian groups and the first three are in addition totally ordered.

---

[19] If the ring is commutative, then the left and right zero divisors are obviously the same.

Since the law of division is in no way associative, we can restrict ourselves to studying for each of the above groups the pair of laws: "$+$" and "$\times$".

So it comes very quickly that:

$$(\mathbb{Z},+,\times) \qquad (\mathbb{Q},+,\times) \qquad (\mathbb{R},+,\times) \qquad (\mathbb{C},+,\times) \tag{5.5.166}$$

constitute unitary and integral commutative rings.

Remark

We will consider as obvious, at this level of our discourse, that the reader will have noticed that $\mathbb{Z}$ is a "sub-ring" of $\mathbb{Q}$ in the sense that the operations defined are internal to each set and that the neutral and identity elements are identical and that there exists for each element of these sets an opposite which is in the same set. We will deepen the concept of "sub-ring" a little further.

Let A be a ring. We have the following properties:

P1. $a+b=a+c \Rightarrow (b=c) \qquad \forall a,b,c \in A$

Proof 5.30.3. This derives from the definition D4 seen at the beginning of the part concerning the algebraic structures (every element has an opposite / symmetric). Indeed, we can add to:

$$a+b=a+c \tag{5.5.167}$$

the element $-a$. We then get:

$$(-a)+a+b=(-a)+a+c \tag{5.5.168}$$

by the existence of the opposite this gives:

$$0+b=0+c \tag{5.5.169}$$

hence:

$$b=c \tag{5.5.170}$$

■ Q.E.D.

P2. $0\cdot a=0 \qquad \forall a\in A$

Proof 5.30.4. This property derives from the definitions D3 (existence of the neutral element), and D4 (existence of the opposite / symmetric), and D5 (distributivity with respect to the other law) and finally of property P1 above. Indeed, we have:

$$0\cdot a+a=0\cdot a+1\cdot a=(0+1)\cdot a=1\cdot a=0a \tag{5.5.171}$$

We therefore have:

$$0\cdot a+a=a \tag{5.5.172}$$

The property P1 above allows us to conclude that:

$$0\cdot a=0 \tag{5.5.173}$$

(we could discuss the relevance of this kind of proof...).

■ Q.E.D.

P3. $(-1)\cdot a=-a \qquad \forall a\in A$

Proof 5.30.5. This property is proved using P2. We have:

$$0=0\cdot a=(1+(-1))\cdot a=1\cdot a+(-1)\cdot a \tag{5.5.174}$$

by adding $-a$ to this last equality, we get:

$$(-a)=(-1)\cdot a \tag{5.5.175}$$

■ Q.E.D.

5.5.4.4.1 Sub-ring

Definitions (#5.5.92): Given A a ring and $S\subset A$ a subset of A. We say that S is a "sub-ring" of A if:

P1. $n\in S$ (the neutral element of A is also that of S)

P2. $a\in S\Rightarrow -a\in S$

P3. $(a,b)\in S\Rightarrow a+b\in S$

P4. $(a,b)\in S\Rightarrow a\cdot b\in S$

✍**Example:**

The ring $\mathbb{Z}$ is a sub-ring of $\mathbb{Q}$ and $\mathbb{R}$.

5.5.4.5 Field

Definition (#5.5.93): We denote a set of numbers by the term "field" if:

$$\boxed{(F,+,\times)\text{ is a field if }\quad\Longleftrightarrow\quad\begin{cases}(F,+,\times)\text{ is a unitary ring}\\(F-\{0\},\times)\text{ is a group}\end{cases}} \tag{5.5.176}$$

Hence a field is a non-zero ring in which any non-zero element is invertible or in other words: a ring of which all non-zero elements are unit elements is a field.

Remarks

R1. If the internal law × is also commutative, the field is obviously named a "commutative field".

R2. The quaternions (see section Numbers page 261) form a "non-commutative field" for addition and multiplication.

Let us see examples of field among the following unitary rings:

$$(\mathbb{Z},+,\times)\qquad(\mathbb{Q},+,\times)\qquad(\mathbb{R},+,\times)\qquad(\mathbb{C},+,\times)\tag{5.5.177}$$

We must first determine which ones do not constitute groups with respect to the internal law of multiplication "$\times$".

As we have already seen in our study of the groups earlier above, it is evident that we must eliminate $(\mathbb{Z},+,\times)$ because of the existence of inverses which is not assured in this set.

Thus, the fundamental fields of arithmetic are:

$$(\mathbb{Q},+,\times)\qquad(\mathbb{R},+,\times)\qquad(\mathbb{C},+,\times)\tag{5.5.178}$$

and since the law of multiplication "$\times$" is commutative in these sets, we can affirm that these fields are also commutative fields.

We often have in the small classes the following diagram for the most important field used by students, engineers and managers:

$$\underset{\text{commutative field totally ordered}}{(\mathbb{R},+,\times)}\left\{\begin{array}{l}\underset{\text{commutative field}}{(\mathbb{R},+)}\left\{\begin{array}{l}\underset{\text{commutative group}}{(\mathbb{R},+)}\left\{\begin{array}{l}\text{The addition is an internal law}\\\text{The addition is an internal associative law}\\\text{The addition is a commutative law}\\\text{There exist a neutral element for the additon: }0\\\text{Any real number has an opposite}\end{array}\right.\\\underset{\text{commutative group}}{(\mathbb{R},\times)}\left\{\begin{array}{l}\text{The multiplication is an internal law}\\\text{The multiplication is an internal associative law}\\\text{The multiplication is an internal commutative law}\\\text{There exist a neutral element for the multiplication: }1\\\text{Every real number has an inverse}\end{array}\right.\\\text{The multiplication is distributive relatively to the addition}\end{array}\right.\\\underset{\text{totally ordered set}}{(\mathbb{R},\leq)}\left\{\begin{array}{l}\text{The relation }\leq\text{ is reflexive}\\\text{The relation }\leq\text{ is antisymmetric}\\\text{The relation }\leq\text{ is transitive}\end{array}\right.\end{array}\right.\tag{5.5.179}$$

Thus, we named "Field" a set F of real or complex numbers a such that the sum, difference, product and quotient of any two of these numbers a belong to the same system F.

We also state this property in the following way: the numbers of a field reproduce by the rational operations (addition, subtraction, multiplication, division). Thus it is evident that the number zero can never form the denominator of a quotient and the set of integers can not form a field because the division in the set of integers does not necessarily give a numerical value that exist in this same set.

5.5.4.6 Vector Spaces

When we define a "vector" (see section Vector Calculus page 1797), we usually refer in high-school to an "Euclidean space" (see also the section Vector Calculus) of n dimensions denoted $\mathbb{R}^n$. However, the notion of vector space is much more larger as we will see by reading the other sections and chapters of this textbook than the latter which represents only one particular case.

Definition (#5.5.94): A "vector space (EV)" or "K-vector space" (abbreviated sometimes: K-ev) on the field K (we will frequently take and use for this field $\mathbb{R}$ or $\mathbb{C}$) is a set $(E,+,\cdot)$ have the following properties:

$$\boxed{(E,+,\cdot)\quad\Longleftrightarrow\quad\begin{cases}(E,+)\text{ is an abelian group}\\ \cdot\text{ is an external law defined by }\begin{cases}\cdot:E\times K\mapsto E\\(x,a)\to a\cdot x\end{cases}\end{cases}}\tag{5.5.180}$$

We thus have two composition laws (taking the traditional notations of the vectors which will perhaps be more intuitive and useful for the reader and for what will also follow...):

1. An internal law of composition: the addition denoted "+" which satisfies:
   1.1. Associativity:
   $$\forall\vec{x},\vec{y}\vec{z}\in E:\quad(\vec{x}+\vec{y})+\vec{z}=\vec{x}+(\vec{y}+\vec{z})\tag{5.5.181}$$
   1.2. Commutativity:
   $$\forall\vec{x},\vec{y}\in E:\quad\vec{x}+\vec{y}=\vec{y}+\vec{x}\tag{5.5.182}$$
   1.3. Neutral element:
   $$\exists\vec{0}\in E:\quad\forall\vec{x}\in E,\vec{x}+\vec{0}+\vec{x}\tag{5.5.183}$$
   1.4. Opposite element:
   $$\forall\vec{x}\in E,\exists\vec{x}'\in E:\quad\vec{x}+\vec{x}'=\vec{0}\tag{5.5.184}$$
2. An external composition law: the multiplication by a scalar, denoted "$\cdot$" (to avoid the confusion with the cross product "×" that we will introduce in the section of Vector Calculus page 1830), which satisfies:
   2.1. Associativity:
   $$\forall\lambda,\mu\in K,\forall\vec{x}\in E:\quad\lambda\cdot(\mu\cdot\vec{x})=(\lambda\cdot\mu)\cdot\vec{x}\tag{5.5.185}$$
   2.2. Distributivity on the right with respect to the field K:
   $$\forall\lambda,\mu\in K,\forall\vec{x}\in E:\quad(\lambda+\mu)\cdot\vec{x}=\lambda\cdot\vec{x}+\mu\cdot\vec{x}\tag{5.5.186}$$
   2.3. Left distributivity with respect to E:
   $$\forall\lambda\in K,\forall\vec{x},\vec{y}\in E:\quad\lambda\cdot(\vec{x}+\vec{y})=\lambda\cdot\vec{x}+\lambda\cdot\vec{y}\tag{5.5.187}$$

2.4. Neutral element (from K to E):

$$\forall \vec{x} \in E: \quad 1 \cdot \vec{x} = \vec{x} \tag{5.5.188}$$

Hence 10 properties at the total (yes indeed... the fact of having an internal or external law is a property on itself!).

We then say that the vector space has a "vectorial algebraic structure" and that these elements are "vectors" (at least in the most common case...), the elements of K are the "scalars" (see section Numbers page 209).

Remarks

R1. The respective internal laws that are frequently used as addition and multiplication that we already know very well on $\mathbb{R}$, which is very convenient for our habits...

R2. From now on, to distinguish the elements of the field K from the set E, we shall denote those of K by Greek letters and those of E by Latin capital letters.

It is should not be necessary to prove that these properties are satisfied for $\mathbb{R}^n$ and consequently for $\mathbb{R}^2$. We can, however, ask ourselves some subsets of $\mathbb{R}^n$?

✍**Examples:**

E1. Let us consider the rectangular region of $\mathbb{R}^3$ shown in figure (a) and in perspective in the (c) below:

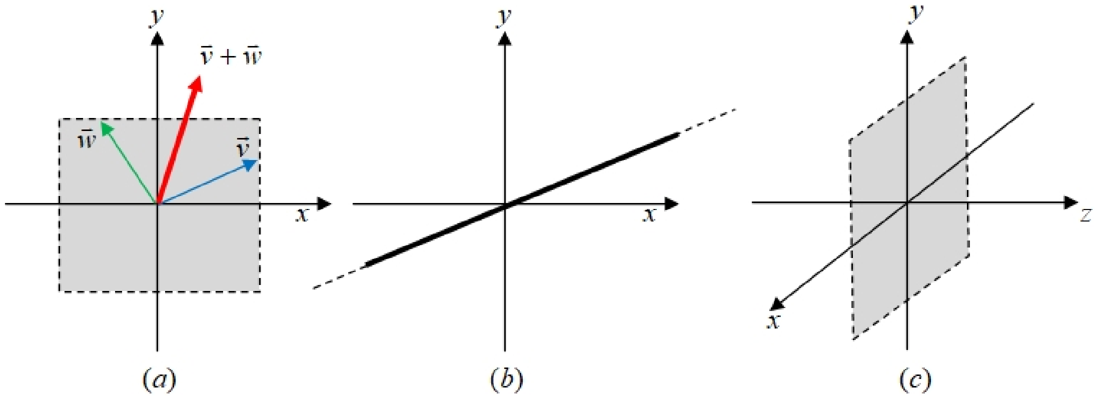


Figure 5.68: **Example of a concept of vector space**

This subset (the rectangle) of $\mathbb{R}^2$ is not a vector space because, among other things, the internal operation property (closure) of the abelian group is not satisfied. Indeed, if we take two vectors $\vec{v}$,$\vec{w}$ inside the rectangle and add them, the result $\vec{v}+\vec{w}$ may come out of the rectangle. On the other hand, it is easy to see that the (infinite) line $\mathbb{R}^1$ illustrated in figure (b) follows all the properties enumerated above and, consequently, defines a vector space. Let us notice, however, that this line must pass through the origin, otherwise the property of the neutral element of the abelian group would not be respected (the neutral element then would no longer exists).

E2. Another example of a vector space is the "polynomial vector space" with real coefficients of degree two or less denoted $\mathcal{P}^2$ (see section Calculus page 1368). For example, two randomly chosen elements of this space are:

$$\begin{aligned} x &= t^2+2t+3 \\ y &= -t+5 \end{aligned} \tag{5.5.189}$$

This polynomial set respects the properties of a vector space. Indeed, if we add two polynomials of degree two or less, we get another polynomial of degree two or less. We can also multiply a polynomial by a scalar without changing the order (or degree) of that latter, and so on.... We can therefore represent a polynomial by vectors whose terms are the coefficients of the polynomial.

Let us notice that we can also form vector spaces with sets of functions more general than polynomials. It is only important to respect the 10 fundamental properties of a vector space!

So defined, a vector space E on K is an action of $(K,\times)$ on $(E,+)$ which is compatible with the group law (by extension an "automorphism" - see definition below - on $(E,+)$).

Definition (#5.5.95): Let E be a vector space, we name "vectorial subspace" F of E a subset of E if and only if (as written by nice mathematicians...):

$$\begin{aligned} &F \neq \varnothing \\ &\forall (X,Y): \quad \in F^2, X+Y \in F \\ &\forall \lambda \in K, \forall X \in F: \quad \lambda X \in F \end{aligned} \tag{5.5.190}$$

or by using another notation (the one used more by physicists and engineers):

$$\begin{aligned} &F \neq \varnothing \\ &\forall (\vec{x},\vec{y}) \in F^2: \quad \vec{x}+\vec{y} \in F \\ &\forall \lambda \in K, \forall \vec{x} \in F: \quad \lambda \vec{x} \in F \end{aligned} \tag{5.5.191}$$

5.5.4.7 C-algebra A

Definition (#5.5.96): A "C-algebra A" where C is a commutative field (also named sometimes "K-algebra A" where the K state for "Körper" in German) is a set A with two internal composition laws: the addition "$+$" and the cross product "$\times$" and one external law (multiplication) "$\cdot$" with operator domain C (multiplication by a scalar) if and only if:

$$(A,C,+,\cdot,\times) \text{ is a } C\text{-algebra} \quad \Longleftrightarrow \quad \begin{cases} (A,+,\cdot) \text{ is a } C\text{-vector space} \\ (A,+,\times) \text{ is a unitary ring} \\ \forall \lambda \in C, \forall a,b \in A: \quad (\lambda a)\cdot b = a \cdot (\lambda b) = \lambda (a \cdot b) \end{cases} \tag{5.5.192}$$

✍**Examples:**

E1. To take an example in the line of the one on vector space examples, the Euclidean space $\mathbb{R}^3$ with the addition "$+$", the multiplication "$\cdot$" and the cross product "$\times$" is a non-associative and non-commutative $\mathbb{R}$-algebra denoted $(\mathbb{R}^3, \mathbb{R}, +, \cdot, \times)$.

E2. The set $\mathbb{C}$ is an $\mathbb{R}$-algebra (as complex numbers can be seen as a vector with two components according to what we saw in the section Numbers page 244).

5.5.4.8 Summary

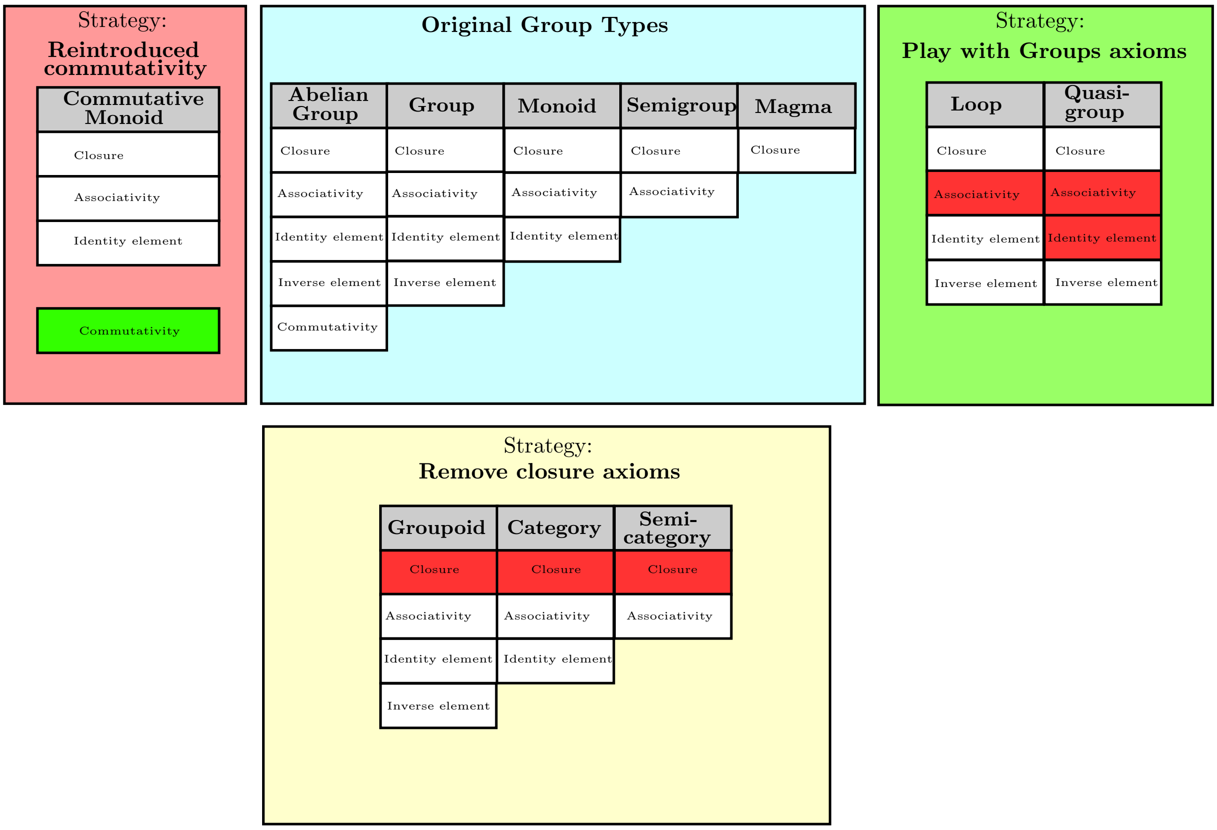
Ok... so far a lot of definitions and concepts. To sum up the most important algebraic structures seen so far we get the authorization of Kevin Binz to reproduce very nice visual figures that could help the reader brain to have an overview of most important concepts seen so far.

So first let us introduce the different set structures in the point of view where we build them by adding each time a new property:

Magma	**Semigroup**	**Monoid**	**Group**	**Abelian Group**
Closure	Closure	Closure	Closure	Closure
	Associativity	Associativity	Associativity	Associativity
		Identity element	Identity element	Identity element
			Inverse element	Inverse element
				Commutativity

Table 5.24: Summary of most common algebraic structures
(source: `https://kevinbinz.com/tag/identity-element/`, author: Kevin Binz)

Of course, there is no particularly strong reason to "knife" axioms in this order. More esoteric options are available (in the below, red axioms are removed, green are re-introduced) introducing at the same time some set structure that we did not introduce before:



Strategy:
Reintroduced commutativity

Commutative Monoid
Closure
Associativity
Identity element
Commutativity

Original Group Types

Abelian Group	**Group**	**Monoid**	**Semigroup**	**Magma**
Closure	Closure	Closure	Closure	Closure
Associativity	Associativity	Associativity	Associativity	
Identity element	Identity element	Identity element		
Inverse element	Inverse element			
Commutativity				

Strategy:
Play with Groups axioms

Loop	**Quasi-group**
Closure	Closure
Associativity	Associativity
Identity element	Identity element
Inverse element	Inverse element

Strategy:
Remove closure axioms

Groupoid	**Category**	**Semi-category**
Closure	Closure	Closure
Associativity	Associativity	Associativity
Identity element	Identity element	
Inverse element		

Figure 5.69: Summary of most common algebraic structures
(source: `https://kevinbinz.com/tag/identity-element/`, author: Kevin Binz)

Closure: For all a, b in S, $a + b$ is also in S

Associativity: For all a, b, c in S, the relation $(a + b) + c = a + (b + c)$ holds

Identity element: There exists the element 0 such that for each a in S, $0 + a = a + 0 = a$ holds.

Inverse element: For each a in S, there exists an element $-a$ such that $a + (-a) = -a + a = 0$

Commutativity: For all a, b in S, the relation $a + b = b + a$ holds.

Figure 5.70: Group Theory addition summary
(source: https://kevinbinz.com/tag/identity-element/, author: Kevin Binz)

If we think of the above as a function, three inputs are relevant to us: the target set, the operator, and the identity element. Thus, we can condense the above into $(S, +, 0)$.

Closure: For all a, b in S, $a \times b$ is also in S

Associativity: For all a, b, c in S, the relation $(a \times b) \times c = a \times (b \times c)$ holds

Identity element: There exists the element 1 such that for each a in S, $1 \times a = a \times 1 = a$ holds.

Inverse element: For each a in S, there exists an element $1/a$ such that $a \times (1/a) = 1/a \times a = 1$

Commutativity: For all a, b in S, the relation $a \times b = b \times a$ holds.

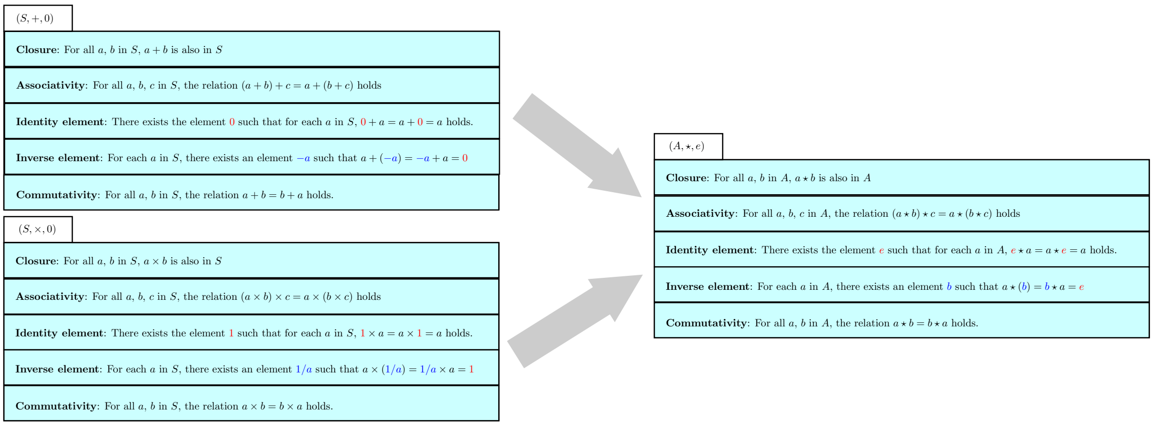
Figure 5.71: Group Theory multiplication summary
(source: https://kevinbinz.com/tag/identity-element/, author: Kevin Binz)

If we think of the above as a function, three inputs are relevant to us: the target set, the operator, and the identity element. Thus, we can condense the above into $(S, \times, 0)$.

Did the above two sections feel painfully similar? Yes it seems!

One lesson they teach you in computer science is: if you notice yourself copy-pasting code, you should try to consolidate your software into one function.

Analogously, we can generalize addition and multiplication, like this:



$(S, +, 0)$

Closure: For all a, b in S, $a + b$ is also in S

Associativity: For all a, b, c in S, the relation $(a + b) + c = a + (b + c)$ holds

Identity element: There exists the element 0 such that for each a in S, $0 + a = a + 0 = a$ holds.

Inverse element: For each a in S, there exists an element $-a$ such that $a + (-a) = -a + a = 0$

Commutativity: For all a, b in S, the relation $a + b = b + a$ holds.

$(S, \times, 0)$

Closure: For all a, b in S, $a \times b$ is also in S

Associativity: For all a, b, c in S, the relation $(a \times b) \times c = a \times (b \times c)$ holds

Identity element: There exists the element 1 such that for each a in S, $1 \times a = a \times 1 = a$ holds.

Inverse element: For each a in S, there exists an element $1/a$ such that $a \times (1/a) = 1/a \times a = 1$

Commutativity: For all a, b in S, the relation $a \times b = b \times a$ holds.

$(A, \star, e)$

Closure: For all a, b in A, $a \star b$ is also in A

Associativity: For all a, b, c in A, the relation $(a \star b) \star c = a \star (b \star c)$ holds

Identity element: There exists the element e such that for each a in A, $e \star a = a \star e = a$ holds.

Inverse element: For each a in A, there exists an element b such that $a \star (b) = b \star a = e$

Commutativity: For all a, b in A, the relation $a \star b = b \star a$ holds.

Figure 5.72: Merging of addition and multiplication
(source: https://kevinbinz.com/tag/identity-element/, author: Kevin Binz)

Here, we generalize our three inputs defined above:

- S becomes a specific instance of an input set
- $+$ and $\times$ become specific instances of a general class of operator.
- 0 and 1 become specific instances of a general class of identity elements.

We have seen earlier that this particular set of five axioms is an "Abelian group" (also known as a "commutative group") that can be summarized as:

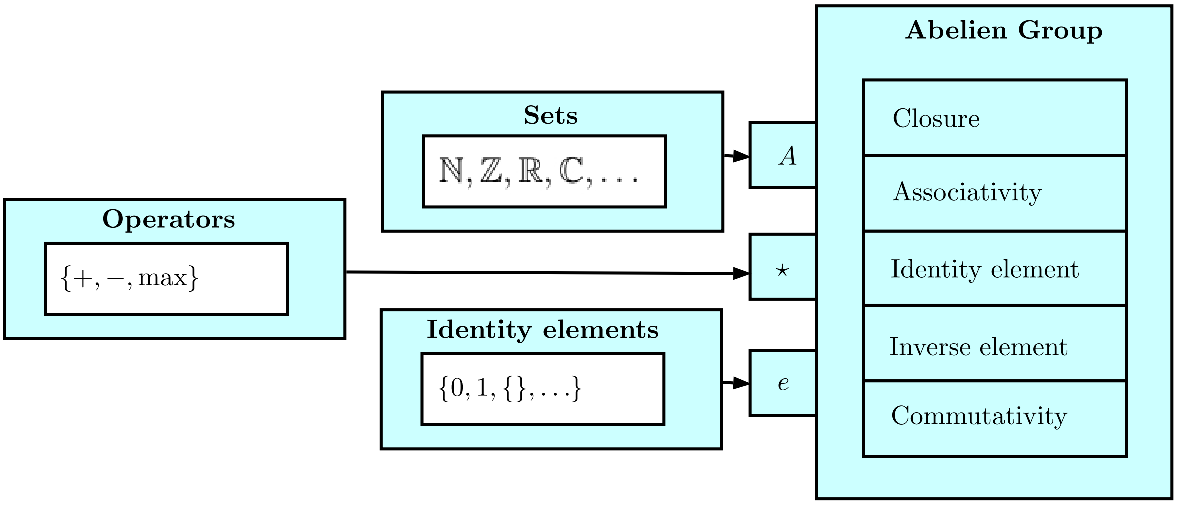


Figure 5.73: Abelian group structure (source: https://kevinbinz.com/tag/identity-element/, author: Kevin Binz)

Let us marry addition-groups and multiplication-groups together, into a field:

Field $(F,+,0,\times,1)$	
Closure: For all a, b in F, $a+b$ is also in F	Abelian Group $F(+,0)$
Associativity: For all a, b, c in F, the relation $(a+b)+c=a+(b+c)$ holds	
Identity element: There exists the element 0 such that for each a in F, $0+a=a+0=a$ holds.	
Inverse element: For each a in F, there exists an element $-a$ such that $a+(-a)=-a+a=0$	
Commutativity: For all a, b in F, the relation $a+b=b+a$ holds.	
Closure: For all a, b in F, $a\times b$ is also in F	Abelian Group $(F-\{0\},\times,1)$
Associativity: For all a, b, c in F, the relation $(a\times b)\times c=a\times(b\times c)$ holds	
Identity element: There exists the element 1 such that for each a in F, $1\times a=a\times 1=a$ holds.	
Inverse element: For each a in F, there exists an element $1/a$ such that $a\times(1/a)=1/a\times a=1$	
Commutativity: For all a, b in F, the relation $a\times b=b\times a$ holds.	
Left distributivity: For all a, b, c in F, the equation $a\times(b+c)=(a\times b)+(a\times c)$ holds	Relations between "$+$" and "$\times$"
Right distributivity: For all a, b, c in F, the equation $(a+b)\times c=(a\times c)+(b\times c)$ holds	

Figure 5.74: Field set structure (source: `https://kevinbinz.com/tag/identity-element/`, author: Kevin Binz)

5.5.5 Morphisms

In many fields of mathematics, morphism refers to a structure-preserving map from one mathematical structure to another. The notion of morphism recurs in much of contemporary mathematics. In set theory, morphisms are functions; in linear algebra, linear transformations; in group theory, group homomorphisms; in topology, continuous functions, and so on.

The concept of homomorphisms (from the Greek homoios = similar and morphê = form) was defined by mathematicians because it makes it possible to highlight remarkable properties of functions in particular with their structures, their kernel, and what we name the "ideals" (see further below). They will allow us to identify one algebraic structure of another.

Definitions (#5.5.97):

D1. If $(A,\star)$ and $(B,\circ)$ are two magmas (regardless of the notation used for internal laws), an application f of A into B is a named "homomorphism of magma" or "morphism of magma" (by abuse of language we sometimes write just "homomorphism") if:

$$\boxed{f(a\star b)=f(a)\circ f(b)\quad \forall a,b\in A} \tag{5.5.193}$$

In other words: if the image of a composition in A is the composition of the images in B.

D2. If $(A,\star)$ and $(B,\circ)$ are two monoids, an application f of A into B is a "homomorphism of monoid" if:

$$\boxed{\begin{aligned} f(a\star b)&=f(a)\circ f(b)\quad \forall a,b\in A\\ f(1_A)&=1_B \end{aligned}} \tag{5.5.194}$$

where 1_A, 1_B are the respective neutral elements of the monoids A, B.

D3. If A, B are two rings, a "homomorphism of rings" (very important for the Cryptography section of this textbook page 7216!) of A in B is a function $f:A\mapsto B$ such that for any $a,a'\in A$ we have:

$$\boxed{\begin{aligned} f(1_A)&=1_B\\ f(a+a')&=f(a)+f(a')\\ f(a\cdot a')&=f(a)f(a') \end{aligned}} \tag{5.5.195}$$

where 1_A, 1_B are the neutral elements of the rings A, B with respect to the multiplication "$\cdot$".

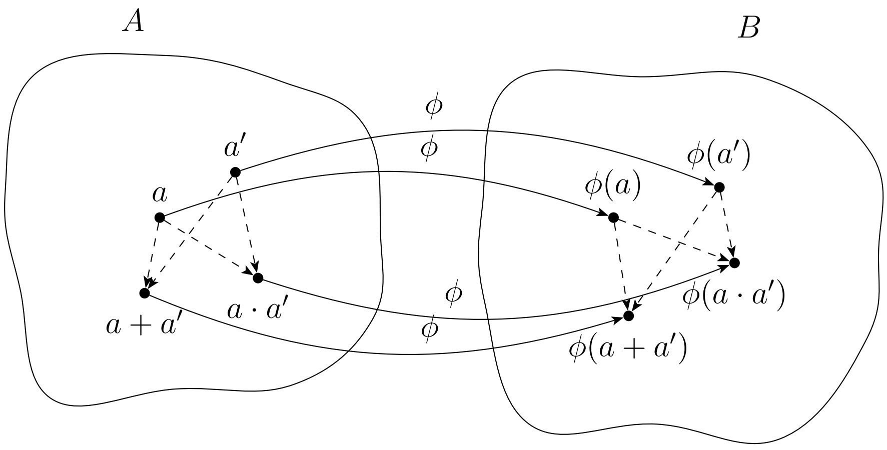


Figure 5.75: **Ring homomorphism**

Given $f : A \mapsto B$ a homomorphism of rings. Then:

P1. $f(0) = 0$

Proof 5.30.6. By:

$$f(a + a') = f(a) + f(a') \tag{5.5.196}$$

we have

$$f(a + 0) = f(a) + f(0) \tag{5.5.197}$$

By adding $-f(a)$ on both sides of the equality, we get:

$$0 = f(0) \tag{5.5.198}$$

■ Q.E.D.

P2. $f(-a) = -f(a)$

Proof 5.30.7. This property also follows from:

$$f(a + a') = f(a) + f(a') \tag{5.5.199}$$

and of the property P1. Indeed, we have:

$$f(0) = f(a + (-a)) = f(a) + f(-a) = 0 \tag{5.5.200}$$

By adding $-f(a)$ to the two sides of the last equality, we get:

$$f(-a) = -f(a) \tag{5.5.201}$$

■ Q.E.D.

P3. If a is a unit of A, then $f(a)$ is a unit of B and $f(a)^{-1} = f(a^{-1})$.

Proof 5.30.8. Given $a, b \in A$ such that:

$$ab = ba = 1 \tag{5.5.202}$$

Then by $f(a \cdot b) = f(a) \cdot f(b)$ and $f(1_A) = 1_B$, we have:

$$1 = f(1) = f(ab) = f(a)f(b) \tag{5.5.203}$$

and also:

$$1 = f(b)f(a) \tag{5.5.204}$$

which shows that $f(b)$ is the inverse of $f(a)$ if b is the inverse of a.

■ Q.E.D.

Let us now introduce a quite powerful theorem:

Theorem 5.31. Let us now show that a homomorphism of rings $f : A \mapsto B$ is injective if and only if the element 0 is the only pre-image of 0 (and therefore reciprocally), which is technically written:

$$\ker(f) = \{0\} \tag{5.5.205}$$

That is to say: the kernel is "trivial".

Proof 5.31.1. The condition is clearly necessary. Let us show that it is sufficient:

We therefore assume that $\ker(f) = 0$. Given $a, a' \in A$ such that $f(a) = f(a')$. Then as we have a homomorphism of ring we can write:

$$0 = f(a) - f(a') = f(a - a') \tag{5.5.206}$$

Which implies that $a - a' = 0$ and therefore that $a = a'$.

This shows that f is injective if it is a homomorphism and that $\ker(f) = 0$ is indeed a sufficient condition.

■ Q.E.D.

D4. Given $(A, +)$ and $(B, \star)$, two groups and f an application $f : A \mapsto B$. We say that f is a "homomorphism of group" if (we could just as well put the "$\star$" instead of the "$+$" in the first group and the "$+$" instead of the "$\star$" in the second group, the definition would remain the same by simply replacing the respective operators !):

$$\boxed{\begin{gathered} f(x + y) = f(x) \star f(y) \quad \forall x, y \in A \\ f(1_A) = 1_B \\ f(x^{-1}) = (f(x))^{-1} \quad \forall x \in A \end{gathered}} \tag{5.5.207}$$

where 1_A, 1_B are the respective neutral elements of the groups A, B. We notice that the only difference between a homomorphism of rings and a homomorphism of groups is that it has two laws instead of one and that we add the concept of inverse.

That said, the third proposition above is in fact a consequence of the definition composed only of the first two lines. Indeed, consider a homomorphism f between the groups $(A, +)$ and $(B, \star)$ with 1_A and 1_B respectively the neutral elements of A and B, then we have:

$$\forall x \in A : \quad 1_B = f(1_A) = f(x + x^{-1}) = f(x) \star f(x^{-1}) \tag{5.5.208}$$

Hence:

$$\forall x \in A: \quad 1_B = f(x) \star f(x^{-1}) \tag{5.5.209}$$

Therefore:

$$\forall x \in A: \quad f(x) = \frac{1_B}{f(x^{-1})} = (f(x^{-1}))^{-1} \tag{5.5.210}$$

✍**Example:**

The exponential function e (see section Functional Analysis page 2289) and the logarithm function $\ln$ are a group homorphism of the group $(\mathbb{R}^+, +)$ on the group $(\mathbb{R}^+, \cdot)$.

D5. Let f be an application $f: A \mapsto B$ from one field to another. We say that f is a "homomorphism of field" if f is a homomorphism of ring...

Indeed, the fact that the homomorphism of field is the same as that of a ring is due to the fact that the difference between the two structures is that the elements of the field are all invertible (no law or law property differs between the both according to their definition of the homomorphism).

Let us show now that every homomorphism of field is injective ("injective homomorphism") by remembering that earlier above we have proved that every homomorphism of rings was injective!

Proof 5.31.2. If a is different from 0 and $b = a^{-1}$ (we use here the property that the elements of a field are invertible!) then:

$$1 = f(1) = f(a \cdot b) = f(a) \cdot f(b) \tag{5.5.211}$$

So when a is different from zero, $f(a)$ is also different from 0 which proves that $\ker(f) = \{0\}$ and therefore that f is injective.

■ Q.E.D.

D6. Given A and B be two K-vector spaces and $f: A \mapsto B$ an application of A into B. We say that f is a "linear mapping" or "linear application" or "homomorphism of vector spaces" (it is implicit that this is relative to the indicated laws and for the chosen application) if:

$$\begin{aligned} &f(x+y) = f(x) + f(y) \quad \forall (x,y) \in A \\ &f(\lambda x) = \lambda f(x) \quad \forall x \in A, \forall \lambda \in K \end{aligned} \tag{5.5.212}$$

and we denote by $L(A, B)$ the set of linear applications.

Remarks

R1. We have already defined the concept of linear application but did not specify that the two sets A and B were K-vector spaces.

R2. A linear application is named a "linear form" if and only if $B = K$.

D7. If the homomorphism is bijective we will say that f is an "isomorphism". If there is an isomorphism between A and B, we say that A and B are "isomorphic" and we will denote this $A \simeq B$.

Remark

The isomorphism allows the identification of two sets with an identical algebraic structure (group, ring, etc.) but whose elements are named in a different way.

D8. If the homomorphism f is a purely internal mapping, then we will say that f is an "endomorphism" (in other words, we have an endomorphism if in the definition of the homomorphism we have $A = B$).

Remark

If we have an endomorphism f of E, then f is then restricted to $\Im(f)$. So the term "endomorphism" just means that the application f arrives in E and not that it touches all the elements of E itself. We have then $f(E) \subset E$ and not necessarily $f(E) = E$ because in the latter case we say that f is surjective as we have already seen it.

D9. If the endomorphism f is furthermore bijective (hence in other words if the homomorphism is an endomorphism and an isomorphism), then we will say that f is an "automorphism".

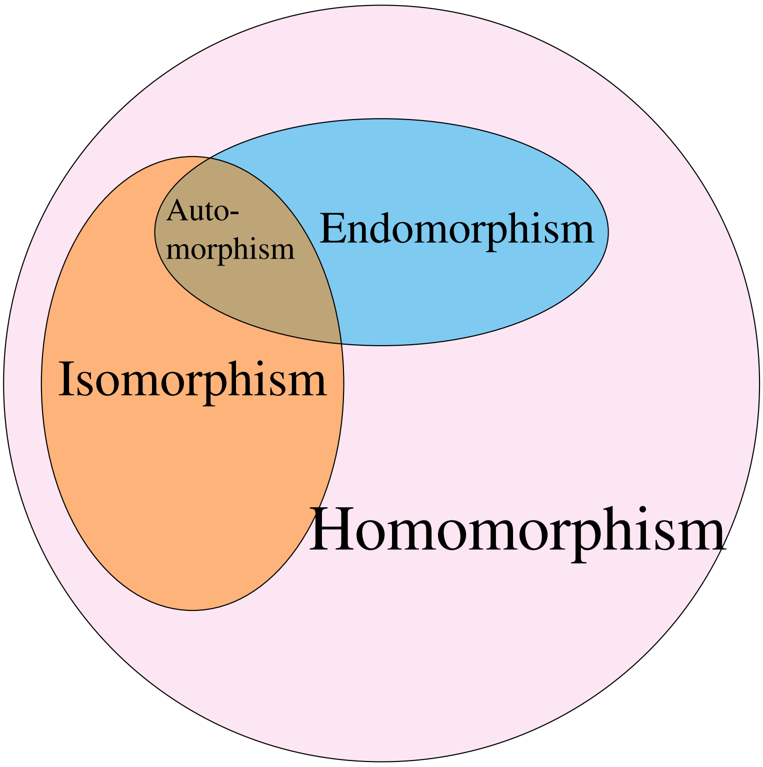


Figure 5.76: **Venn diagram of different types of homomorphisms (source: Wikipedia, author: Martin Thoma)**

5.5.5.1 Ideal

Definition (#5.5.98): Let A be a commutative ring (like $(\mathbb{R},+,\cdot)$ for example). A subset $S \subset A$ is an "ideal" if:

P1. For all $a, a' \in S$:

$$a + a' \in S \qquad (5.5.213)$$

P2. For all $a \in S$ and all $r \in A$:

$$r \cdot a \in S \qquad (5.5.214)$$

In other words, an ideal is a closed subset for addition and stable for the multiplication by any element of A.

✍**Example:**

The set of even numbers $\mathbb{Z}_{2k}$ is an example of ideal of the set of relative integers $\mathbb{Z}$.

Remark

The ideals $S = \{0\}$ and $S = A$ are named the "trivial ideals".

To know if an ideal is equal to the whole ring, it is useful to use the following property which specifies that if A is a ring and I an ideal of A, then if $1 \in A$ we have $I = A$.

Proof 5.31.3. This results from the property P2 of the definition of an ideal:

For any $r \in A$, we have $r = r \cdot 1 \in 1$ because $1 \in I$.

■ Q.E.D.

Another quite general example of ideal is given by the kernel of a homomorphism of rings. Indeed, let us prove that the kernel of a homomorphism $F : R \mapsto S$ is an ideal of R.

Proof 5.31.4. Given $a, a' \in \ker(f)$. Then:

$$f(a + a') = f(a) + f(a') = 0 + 0 = 0 \qquad (5.5.215)$$

which shows that $a + a' \in \ker(f)$. Given $r \in R$, then:

$$f(ra) = f(r)f(a) = f(r) \cdot 0 = 0 \qquad (5.5.216)$$

Which shows that $r \cdot a \in \ker(f)$.

■ Q.E.D.

Proposition: Let A be a ring and let $a \in A$. The subset:

$$\{ax|x \in A\} \tag{5.5.217}$$

Denoted (a) or aA, is an ideal (we will see a concrete example after the next small set of definitions).

Definitions (#5.5.99):

D1. An ideal $I \neq A$ of a ring A is named a "principal ideal" if there exists $a \in A$ such as $I = (a)$.

D2. A ring of which all the ideals are principal is named "principal ring".

✍**Example:**

Let us now show that the ring $\mathbb{Z}$ is principal (because all its ideals are principal).

Let I be an ideal of $\mathbb{Z}$ (it is easy to choose one: for example, all multiples of 2 or 3, etc.). Let $r \in I$ be the smallest non-zero positive integer of I. We will show that $I = r\mathbb{Z} = (r)$.

Let a be any element of I. The Euclidean division allows us to write:

$$a = qr + r' \tag{5.5.218}$$

with $0 \leq r' < r$ (as we have already proveed it).

But as $r' = a - qr$ and that $a, r \in I$, by the definition of an ideal, we have $r' \in I$ (the sum or difference of the elements of an ideal belonging to the ideal). By the choice of r (r' being less than r) this implies that $r' = 0$ and therefore that $a = rq$.

Thus every element of I is a multiple r of q:

$$I = r\mathbb{Z} \tag{5.5.219}$$

So for the set of even number (denoted as we know $2\mathbb{Z}$ or $\mathbb{Z}_{2k}$) we have:

$$I_{2\mathbb{Z}} = 2 \cdot \mathbb{Z} \tag{5.5.220}$$

The example above uses only the Euclidean division on $\mathbb{Z}$. We can then generalize this result to the rings which possess a Euclidean division. Thus, for example, the ring $K[X]$ of the polynomials (see section Calculus page 1368) with coefficients in a field K is a principal ring because it has an Euclidean division.

Theorem 5.32. The ring $K[X]$ of the polynomials with coefficients in a field K is a principal ring (all its ideal are principal).

Proof 5.32.1. Let I be an ideal of $K[X]$. Let us denote by d the smallest degree that can have a non-zero polynomial of I. If $d = 0$ then $1 \in I$ and therefore $I = 1 \cdot K[X] = K[X]$. Otherwise, given $a(X)$ a polynomial of degree d. If $u(X) \in I$ then one can divide $u(X)$ by $a(X)$. There exists therefore $q(X), r(X) \in K[X]$ such as $\deg(r) < \deg(a) = d$ and:

$$u(X) = q(X) \cdot a(X) + r(X) \tag{5.5.221}$$

Therefore $r(X) \in I$ which lead to $r = 0$ (otherwise contradiction with the minimality of d). Therefore:

$$u(X) = q(X) \cdot a(X) \tag{5.5.222}$$

We have just shown that:

$$I = a(X) \cdot K[X] \tag{5.5.223}$$

■ Q.E.D.

To come back on $\mathbb{Z}$... we have then proved that the only ideals are those of the form $r\mathbb{Z}$. Moreover if we have d and r which are integers > 1. Then $r\mathbb{Z} \subset d\mathbb{Z}$ if and only if $d|r$.

Proof 5.32.2. If $d|r$ then there exists n with $r = d \cdot n$. Let $m \cdot a$ be an element of $r\mathbb{Z}$. Therefore:

$$r \cdot a = d \cdot n \cdot a \in d\mathbb{Z} \tag{5.5.224}$$

which shows that $r\mathbb{Z} \subset d\mathbb{Z}$.

Conversely, if $r \in d\mathbb{Z}$ this implies that r is of the form $d \cdot n$ and this proves that d divides r.

■ Q.E.D.

Theorem 5.33. Let us also prove that a ring R is a field if and only if it possesses only the trivial ideals:

$$I = \{\{0\}, R\} \tag{5.5.225}$$

Proof 5.33.1. Let us show that the condition is necessary: Let I be a non-zero ideal of R (that is so say different of $\{0\}$) and $r \in I$ a non-zero element. By hypothesis (that it is a field), it is invertible, that is to say that there exists $t \in R$ such that:

$$r \cdot t = t \cdot r = 1 \tag{5.5.226}$$

This implies that $1 \in I$ and therefore, by a result obtained earlier above $I = R$.

Conversely, suppose that every ideal $I \neq R$ is the null ideal (that is to say $\{0\}$). Then if $r \in R$ is a non-zero element of R, the principal ideal (r) must be equal to R. But this implies that $1 \in (r)$ and therefore that there exists $x \in R$ with $r \cdot x = 1$ which shows that r is invertible. The ring R is therefore a field.

■ Q.E.D.

This characterization will enable us to quite easily prove that:

Theorem 5.34. Any homomorphism starting from a field is injective. That is, if $f : R \mapsto S$ is a homomorphism where R is a field, then f is injective.

Proof 5.34.1. We put together what has been seen so far:

- We have proved earlier above that the kernel $\ker(f)$ of a homomorphism of ring is an ideal.
- We have also proved just earlier above that set was a field if $\{0\}$ and R were the only trivial ideals.

Therefore for the both point above we have either $\ker(f) = 0$ or $\ker(f) = R$ for the field (as it encompass the concept of ring!).

But since $f(1) = 1 \neq 0$ (by definition of a homomorphism!) it follows that it remains to us only the choice $\ker(f) = \{0\}$. This implies by a previous theorem (where we have proved long time before that if $\ker(f) = \{0\}$ the homomorphism is injective) that ... f is injective.

■ Q.E.D.

Let us now study the homomorphisms whose starting ring is $\mathbb{Z}$. Let A be a ring and $f : \mathbb{Z} \mapsto A$ a homomorphism. By definition of a homomorphism and by its properties, it is necessary that as we have defined long time before that $f(0) = 0$ and $f(1) = 1$ (among others). But we also still need that:

$$f(n) = \underbrace{f(1 + \ldots + 1)}_{\text{k}} = \underbrace{(1_A + \ldots + 1_A)}_{\text{k}} = n \cdot 1_A \quad (5.5.227)$$

for any $k \in \mathbb{Z}$. Thus f is completely determined by the information of $f(1)$ and is therefore unique.

Conversely, we show that the application $f : \mathbb{Z} \mapsto A$ defined by:

$$f(k) = k \cdot 1_A \quad (5.5.228)$$

is a homomorphism of rings. In summary, there exists one and only one homomorphism of $\mathbb{Z}$ in any ring A.

Definition (#5.5.100): Given R a ring and $f : \mathbb{Z} \mapsto R$ the unique homomorphism defined just previously. If f is injective, we say that R is of "zero characteristic" and we denote it:

$$\text{char}(R) = k = 0 \quad (5.5.229)$$

Otherwise, $\ker(f)$ is a non trivial ideal of $\mathbb{Z}$ and as $\mathbb{Z}$ is therefore principal (as we have demonstrated above) it is of the form $k\mathbb{Z}$ with $k > 0$. The integer k is named the "characteristic of R" and we have therefore:

$$\text{char}(R) = k > 0 \quad (5.5.230)$$

Remark

Less formally, the characteristic of a A ring is the smallest positive integer k such that $k \cdot 1_A = 0$. If there is none, then the characteristic is zero.

The definition above may not be very clear (at least for me it is not!). So let us see another approach to introduce the characteristic much more detailed...:

Given R a ring and any of its elements r, let us denote integers with boldface type. So **1** is the integer number one, for instance, **0** is the integer zero, whereas 0_R and 1_R are the neutral elements in R.

Since R under addition is an abelian group, we can as usual define:

$$\mathbf{0}r = 0_R \qquad \text{and} \qquad (\boldsymbol{n}+\mathbf{1})r = \boldsymbol{n}r + r \tag{5.5.231}$$

By definition of a homomorphism, we have if $f : \mathbb{Z} \mapsto R$ that:

$$f(\boldsymbol{n}) = \boldsymbol{n}1_R \tag{5.5.232}$$

then for recall:

- $f(\boldsymbol{m}+\boldsymbol{n}) = f(\boldsymbol{m}) + f(\boldsymbol{n})$
- $f(\boldsymbol{mn}) = f(\boldsymbol{m})f(\boldsymbol{n})$
- $f(\mathbf{1}) = 1_R$

so f is a homomorphism of rings. Like all ring homomorphisms, $f : \mathbb{Z} \mapsto R$ has a kernel which is an ideal as we have proved just earlier, so we have $\ker(f) = \boldsymbol{k}\mathbb{Z}$ for a unique $\boldsymbol{k} \geq \mathbf{0}$.

If $k = 0$, then as we have proved much earlier above, then f is injective. Otherwise, we have:

$$\boldsymbol{k}1_R = 0_R \tag{5.5.233}$$

by definition of the kernel, which means that:

$$\boldsymbol{k}r = \underbrace{r + r + \ldots + r}_{\boldsymbol{k}} = \underbrace{(1_R + 1_R + \ldots + 1_R)}_{\boldsymbol{k}} r = (\boldsymbol{k}1)r = 0_R r = 0_R \tag{5.5.234}$$

for every $r \in R$. This integer $\boldsymbol{k}$ is the "characteristic" of S.

That is, char(R) is the smallest positive number $\boldsymbol{k}$ such that:

$$\underbrace{1_R + 1_R + \ldots + 1_R}_{\boldsymbol{k}\text{ times}} = 0 \tag{5.5.235}$$

if such a number $\boldsymbol{k} \in \mathbb{N}$ exists, and 0 otherwise.

✍**Examples:**

E1. The only ring that has $\text{char}(R) = 1$ such that:

$$\boldsymbol{k} \cdot 1_R = 1 \cdot 1_R = 1_R = 0_R \tag{5.5.236}$$

is the trivial ring $R = \{0_R\}$.

E2. The ring $\mathbb{Z}$ is of zero characteristic because the unique homomorphism $f : \mathbb{Z} \mapsto \mathbb{Z}$ is the identity. It is therefore injective.

E3. The injections $\mathbb{Z} \mapsto \mathbb{Q}$ and $\mathbb{Z} \mapsto \mathbb{R}$ show that $\mathbb{Q}$ and $\mathbb{R}$ (and also $\mathbb{C}$) are fields of zero characteristic.

Theorem 5.35. We now propose to prove that the characteristic of an integral ring (and in particular of a field) is equal to 0 or to a prime number p.

Proof 5.35.1. We show the contraposition. Let R be a ring of characteristic $m \neq 0$ with m not prime.

There are then natural $n, r \in \mathbb{N}$ constraint by $n, r < m$ such that $m = n \cdot r$.

Given $f : \mathbb{Z} \mapsto R$ the unique homomorphism (defined earlier above). By definition of m we have then $f(m) = 0_R$ but... (!) $f(r) \neq 0 \neq f(n)$. But then (and this it the trick of the proof!):

$$f(r)f(n) = f(r \cdot n) = f(m) = 0 \tag{5.5.237}$$

which shows that R is not integral.

■ Q.E.D.

Remark

The reciprocal of the theorem is not true as shown in the example of the ring $\mathbb{R} \times \mathbb{R}$ where addition and multiplication are made component by component. It is a ring of zero characteristic but with divisors of zero:

$$(1, 0) \cdot (0, 1) = 0 \tag{5.5.238}$$

~95 %

★★★★☆

16 votes, 62.5%

5.6. Probabilities

PROBABILITY is the measure of the likelihood that an event will occur and therefore the calculation of probabilities handles random phenomena (known more aesthetically as "stochastic processes" when their are time-dependent), that is to say, phenomena that do not always lead to the same outcome and that can be studied using numbers their implications and occurrences. However, even if these phenomena have variable outcomes, depending on chance, we observe a certain statistical regularity.

Probability is quantified as a number between 0 and 1 (where 0 indicates impossibility and $1$indicates certainty). The higher the probability of an event, the more certain we are that the event will occur.

Probabilities are very important in practice, especially at very high management or political level. Indeed, in many cases, it is more practical to use a simple but uncertain rule rather than a complex but certain one, even if the true rule is deterministic and our modelling system has the fidelity to accommodate a complex rule. For example, the simple rule "the sales will increase" is cheap to develop and is broadly useful, while a rule of the form, "the sales will increase if the inflation index also increase, and there is no epidemia, no earthquakes, no conflict, no asteroids, no major flood, no political change, no strikes..." is expensive to develop, maintain and communicate, and after all of this effort is still very brittle and prone to failure.

The concepts related to probabilities have been given an axiomatic mathematical formalization in probability theory (see further below), which is used widely in such areas of study as mathematics, statistics, finance, gambling, science (in particular physics), artificial intelligence/Machine Learning, computer science, game theory, and philosophy to, for example, draw inferences about the expected frequency of events. Probability theory is also used to describe the underlying mechanics and regularities of complex systems.

quote
5.6:450.#38

> ❞ *We measure the degree of intelligence of an individual by the quantity of incertitudes he can handle.* ❞
>
> Emmanuel Kant

Definitions (#5.6.101): There are several ways to define a "probability". Mainly we are talking about (if we put apart the formalism of Measure Theory that we will study much later at page 2493):

D1. "Experimental or inductive probability" which is the probability derived from the whole population or a sample of the latter trough the study of the underlying phenomena with experimentation.

D2. "Theoretical or deductive probability" which is the known probability through the study of the underlying phenomenon without experimentation. It is therefore an "a priori" knowledge as opposed to the previous definition that was rather referring to a notion of "a posteriori" probability.

As it is not always possible to determine a priori probabilities, we are often asked to perform experiments. We must then be able to pass from the first to the second solution. This passage is supposed to be possible in terms of limit (with a population sample whose size approaches the size of the whole population).

The formal modelling of the probability calculus was invented by A.N Kolmogorov in a book published in 11933 (holocene calendar). This model is based on the probability space (U, A, P) that we will define a little further and that we can relate to the theory of measurement (see section Measure Theory page 2481). However, the probabilities were studied in the scientific point of view by Pierre de Fermat and Blaise Pascal in the mid 117th century (holocene calendar).

Remark

If you have a teacher or trainer who dare to teach statistics and probabilities only with examples based on gambling (cards, dice, match, toss, etc.) dispose it to whom it may concern because it would mean that he has no experience in the field and he will teach you anything and no matter how (examples could normally be based on industry, pharmaceutics, economy or R&D, in short: areas daily used in companies but especially not on gambling...!).

It is common to distinguish three different main types of probabilities in practice:

D1. "Frequentist probability", also known as "frequentism", is an interpretation of probability that defines an event's probability as the limit of its relative frequency in many trials. In other words, it considers probabilities based on how often an event occurs in repeated experiments or observations.

D2. "Bayesian probability" is interpreted as reasonable expectation or quantification of personal belief rather than just frequency or propensity. A Bayesian probabilist starts by specifying a prior probability for a hypothesis based on available information. As new relevant data (evidence) becomes available, the prior probability is updated to a posterior probability using Bayes theorem (see page 485).

D3. "Fuzzy probability" is an extension of traditional probability theory that deals with mixed probabilistic and non-probabilistic uncertainty. Fuzzy probabilities reflect both aleatory (random) uncertainty and epistemic (knowledge-based) uncertainty separately, providing extended information and decision aids.

5.6.1 Event Universe

quote
5.6:452.#39

> *In physics anything that can happen (meaning has a non-zero probability of occurring) eventually will happen!*

Definitions (#5.6.102):

D1. The "universe of events", or "universe of observables", U is the set of all possible outcomes (results), named "elementary events" that occur during a random determined test. The universe can be finite (countable) if the elementary events are finite or continuous (uncountable) if they are infinite.

D2. Any "event" A is a set of elementary events and is part of the universe of possible U. It is possible that an event is composed of only a single elementary event.

✍**Examples:**

E1. Consider the universe of all possible blood groups, then event A "the individual is Rh positive" is represented by:

$$A=\{A+,B+,AB+,O+\}\subset U \tag{5.6.1}$$

while the event B "the individual is a universal donor" is represented by:

$$B=\{O-\}\subset U \tag{5.6.2}$$

which is therefore an elementary event.

E2. In $(\mathbb{N},+)$ every element is regular and $(\mathbb{N},\times)$ any non-zero element is regular.

D3. Let U be a universe and A an event, we say that the event A "occurs" (or "is realized") if during the run of the trial the issue i $(i\in U)$ occurs and that $i\in A$. Otherwise, we say that A "was not realized".

D4. The empty subset $\varnothing$ of U is named "impossible event". Indeed, if during a trial where the event i occurs, we always have $i\in\varnothing$ and the event $\varnothing$ then never occurred.

If U is finite, or countably infinite, any subset of U is an event, that is no longer true if U is uncountable (we will see in the chapter Statistics why).

D5. The set U is also named "certain event". Indeed, if at the end of the trial the event i occurs, we have always $i\in U$ (since U is the universe of events). The event U then always occurred!

D6. Let A and B be two subsets of U. We know that the events $A \cup B$ and $A \cap B$ are both subsets of U then, events that are respectively "joint events" and "disjoint events".

D7. If two events A and B are such that:

$$A \cap B = \varnothing \tag{5.6.3}$$

the two events may not be feasible during the same trial, then we say that they are "mutually exclusive events".

D8. If two events A and B are such that:

$$A \cap B \neq \varnothing \tag{5.6.4}$$

the two events may be feasible during the same trial (the possibility to see a black cat when we pass under a ladder, for example...), we say conversely that they are "independent events".

D9. "Randomness" (i.e. "random"), at the opposite of "probabilistic" (i.e. "probability"), is the lack of pattern or predictability in events. A random sequence of events, symbols or steps has no order and does not follow an intelligible pattern or combination. Individual random events are by definition unpredictable, but in many cases the frequency of different outcomes over a large number of events (or "trials") is predictable.

Remark

If we randomly (i.e. uniformly) choose a real number in the interval $[0,1]$ then for every number there is a zero probability that we will pick this number. This does not mean that we did not pick any number at all!

Similarly with the rationals, while infinite, and dense and all that, they are very very sparse in the aspect of measure and probability. It is perfectly possible that if we throw countably many darts at the real line we will hit exactly all the rationals and every rational exactly once. This scenario is highly unlikely, because the rational numbers is a measure zero set.

Probability deals with *what are the odds of that happening*? a priori, not a posteriori!! So we are interested in measuring a certain structure a set has, in modern aspects of probability and measure, the rationals have size zero and this means zero probability. More formally, as we will see in the section Statistics page 647, the probabilities are obtained by integrating a probability density function $f(x)$ over an interval. The function is non negative and it has the property:

$$\int\limits_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1 \tag{5.6.5}$$

The probability of selecting a real in an interval $[a,b] \subset \mathbb{R}$ is then given by:

$$P(X \in [a,b]) = \int\limits_{a}^{b} f(x)\mathrm{d}x = 1 \tag{5.6.6}$$

The reader can see now that, given a real number $x_0 \in \mathbb{R}$, we have:

$$P(X = x_0) = P(X \in [x_0, x_0]) = \int_{x_0}^{x_0} f(x)\mathrm{d}x = 0 \qquad (5.6.7)$$

quote
5.6:454.#40

> *Probability theory is nothing but common sense reduced to calculation.*
>
> P. S. de Laplace

5.6.1.1 Infinite monkey theorem (Borel's law)

The "infinite monkey theorem ", also named sometimes "Borel's law", states that a monkey hitting keys at random on a typewriter keyboard for an infinite amount of time will almost surely type a given text, such as the complete works of William Shakespeare. In fact the monkey would almost surely type every possible finite text an infinite number of times. However, the probability of a universe full of monkeys typing a complete work such as Shakespeare's Hamlet is so tiny that the chance of it occurring during a period of time hundreds of thousands of orders of magnitude longer than the age of the universe is extremely low (but technically not zero).

In this context, "almost surely" is a mathematical term with a precise meaning, and the "monkey" is not an actual monkey, but a metaphor for an abstract device that produces an endless random sequence of letters and symbols.

Theorem 5.36. If we have an infinite number of monkeys each hitting keys at random on typewriter keyboards then, with probability 1, one of them will type the complete works of William Shakespeare (something like a randomness machine, which is typing keys randomly at a typewriter, would eventually write any book that's ever existed, given enough time).

Proof 5.36.1. Let A_n be the event that the nth monkey types the complete works of Shakespeare. Then if there are m characters on the keyboard and N characters in the complete works of Shakespeare, then the probability to get the complete works of Shakespeare should be (see further below page 523):

$$P(A_n)=\frac{1}{m^N} \tag{5.6.8}$$

for each n. Furthermore the A_n are mutually independent (disjoint event). Hence:

$$\sum_{i=1}^{+\infty}P(A_n)=\sum_{n=1}^{+\infty}m^{-N}=+\infty \tag{5.6.9}$$

Infinitely many of the events A_n occur i.e. infinitely many monkeys will type the complete works of Shakespeare.

■ Q.E.D.

Therefore if we accept infinity in an argument, then we can end up also accepting that "given infinity, anything can happen" (the next example about complex structures will also lead us to the same conclusion!).

Remark

This result is a special case of a more general one named "second Borel–Cantelli lemma". This lemma states that if events E_n are independent and the sum of the probabilities of the E_n diverges to infinity, then the probability that infinitely many of them occur is 1. Ie, if $\sum_{n=1}^{+\infty} P\left(E_n\right) = +\infty$ and the events $\left(E_n\right)_{n=1}^{+\infty}$ are independent, then $P\left(\limsup_{n\to+\infty} E_n\right) = 1$. Formerly:

$$\because \sum_{n=1}^{+\infty} \mathbb{P}\left(E_n\right) = +\infty \wedge \left(\perp\!\!\!\perp E_n\right)_{n=1}^{+\infty} \Rightarrow \mathbb{P}\left(\limsup_{n\to+\infty} E_n\right) = 1 \tag{5.6.10}$$

See the proof just below for more.

quote 5.6:456.#41

> *The true logic of this world is in the calculus of probabilities.*
>
> James Clerk Maxwell

5.6.1.1.1 Second Borel-Cantelli lemma

Theorem 5.37. Let the events E_n be independent and let the sum of the probabilities of the E_n diverges to infinity.

Then the probability that infinitely many of them occur is 1.

That is, if:

$$\sum_{n=1}^{+\infty} P\left(E_n\right) = +\infty \tag{5.6.11}$$

and the events $\langle E_n\rangle_{n=1}^{+\infty}$ are independent, then:

$$P\left(\limsup_{n\to\infty} E_n\right) = 1 \tag{5.6.12}$$

This means that under the right conditions, certain events will almost surely happen infinitely often. This lemma is particularly useful in ergodic theory, number theory, and stochastic processes, where it helps determine whether certain behaviors persist indefinitely.

Proof 5.37.1. Let:

$$\sum_{n=1}^{+\infty} P\left(E_n\right) = +\infty \tag{5.6.13}$$

Let $\langle E_n\rangle_{n=1}^{+\infty}$ be independent. It is sufficient to show the event that the E_n s did not occur for infinitely many values of n has probability 0. This is just to say that it is sufficient to show that:

$$1 - P\left(\limsup_{n\to+\infty} E_n\right) = 0 \tag{5.6.14}$$

Noting that[20]:

$$\begin{aligned}1-P\left(\limsup_{n\to+\infty} E_n\right)&=1-P\left(\{E_n\text{ i.o. }\}\right)=P\left(\{E_n\text{ i.o. }\}^c\right)\\&=P\left(\left(\bigcap_{N=1}^{+\infty}\bigcup_{n=N}^{+\infty}E_n\right)^c\right)=P\left(\bigcup_{N=1}^{+\infty}\bigcap_{n=N}^{+\infty}E_n^c\right)\\&=P\left(\liminf_{n\to+\infty}E_n^c\right)=\lim_{N\to+\infty}P\left(\bigcap_{n=N}^{+\infty}E_n^c\right)\end{aligned}\tag{5.6.15}$$

It is enough to show:

$$P\left(\bigcap_{n=N}^{+\infty}E_n^c\right)=0\tag{5.6.16}$$

Since the $\langle E_n\rangle_{n=1}^{+\infty}$ are independent:

$$P\left(\bigcap_{n=N}^{+\infty}E_n^c\right)=\prod_{n=N}^{+\infty}P\left(E_n^c\right)=\prod_{n=N}^{+\infty}\left(1-P\left(E_n\right)\right)\tag{5.6.17}$$

Let us now introduce the "exponential function inequality", i.e., by the definition of the exponential function (see section Functional Analysis page 2289):

$$e^x=\lim_{n\to+\infty}\left(1+\frac{x}{n}\right)^n\tag{5.6.18}$$

We need now and intermediate result that states that given $x\in\mathbb{R}$ a real number such that $x>-1$ and $n\in\mathbb{Z}_{\geq 0}$ a positive integer, then we get the "Bernoulli's inequality":

$$(1+x)^n\geq 1+nx\tag{5.6.19}$$

- Base Case: For $n=1$, $1+x=1+x$ so the inequality holds.
- Induction Assumption: Assume that for some integer $k\geq 1$, $(1+x)^k\geq 1+kx$.
- Inductive Step: We must show that $(1+x)^{k+1}\geq 1+(k+1)x$

So let's do the proof of the inductive step that confirms Bernoulli's inequality:

$$\begin{aligned}(1+x)^{k+1}&=(1+x)(1+x)^k\\&\geq(1+x)(1+kx)\\&=1+(k+1)x+kx^2\\&\geq 1+(k+1)x,\end{aligned}\tag{5.6.20}$$

since $kx^2\geq 0$. This completes the induction step!

---

[20] In mathematical proofs, the abbreviation "i.o." stands for "infinitely often".

Hence for $x > -1$, we can write using Bernoulli's inequality:

$$\begin{aligned} e^x &= \lim_{n\to+\infty}\left(1+\frac{x}{n}\right)^n \\ &\geq \lim_{n\to+\infty}(1+x) \\ &= 1+x \end{aligned} \tag{5.6.21}$$

Or (we just replace x by $-x$):

$$e^{-x} \geq 1-x \tag{5.6.22}$$

Therefore (the condition $P(E_n) > -1$ is satisfied here to apply Bernoulli's inequality!):

$$\begin{aligned} P\left(\bigcap_{n=N}^{+\infty} E_n^c\right) &= \prod_{n=N}^{+\infty}\left(1-P\left(E_n\right)\right) \\ &\leq \prod_{n=N}^{+\infty}\exp\left(-P\left(E_n\right)\right) \\ &= \exp\left(-\sum_{n=N}^{+\infty}P\left(E_n\right)\right) \\ &= 0 \end{aligned} \tag{5.6.23}$$

This completes the proof.

■ Q.E.D.

5.6.1.2 Do probabilities refute complex structures?

> *The most important questions in life are indeed, for the most part, really only problems of probability.*
>
> P. S. de Laplace

quote 5.6:459.#42

Both traditional creationists and intelligent design writers have invoked probability arguments in criticisms of biological evolution. They argue that some features of biology (for an introduction to biology see [393] and for an excellent introduction to origin of life - abiogenesis - see [453]) are so fantastically improbable that they could never have been produced by a purely natural, "random" process, even assuming the billions of years of history asserted by geologists and astronomers. They often equate the hypothesis of evolution to the absurd suggestion that monkeys randomly typing at a typewriter could compose a selection from the works of Shakespeare, or that an explosion in an aerospace equipment yard could produce a working 747 airliner (we won't give the name of the people that wrote that because the quality of their book and their religious agenda don't deserve to be quoted here!). More recent studies of this type, in an attempt to promote an "intelligent design" world-view, argue that functional biology operates on an exceedingly small subset of the space of all possible DNA sequences, and that any changes to the "computer program" of biology are, like changes to human computer programs, almost certain to make the organism non-functional (again we won't give the name of people who claim that for the same reasons as before!).

Definition (#5.6.103): "Biological evolution" is the cumulative change in the heritable characteristics properties of groups of structures over the course of generations through seconds to millions years depending on the type of structures. It embraces everything from slight changes in the proportions of different forms of a coding structure due to natural environment constraints, random sequential genetic drift and genetic variations within a population to the alterations that led from the earliest self-replicating molecules structures to organisms, dinosaurs, bees, oaks, humans, etc. (notice the subtlety of the definition thanks to the underlined word that avoid to have the "fire" to be an entity having a biological evolution...).

While not generally appreciated by the public at large, it is a well-known fact in the world of scientific research that arguments based on probability and statistics are fraught with potential fallacies and errors, and even "expert" researchers can fool themselves with invalid reasoning. For these reasons, rigorous courses in probability and statistics are now required of students in virtually all fields of science (keep in mind that engineering is not science, this is why actually the huge majority of engineers have a very low understanding of probabilities and statistics), and in numerous other disciplines as well. Attorneys need to be at least moderately well-versed in probability and statistics arguments and how they can go awry in the courtroom arguments. In the finance world, statistical over-fitting and other errors of probability and statistics are now thought to be a leading reason behind the fact that many strategies and investment funds which look great on paper often fail miserably in real-world usage.

To illustrate the difficulties with probability arguments, mathematics teachers often ask their class (let's say it has 23 students) if they think it is likely that two or more persons in the class have exactly the same birthday. Most students say that it is highly unlikely, thinking that the chances that two people have the same particular birthday is $1/365$, and so 23 times this amount is only $23/365$. But this argument is fallacious, indeed, if one numbers the 23 people from 1 to 23, the event that all 23 people have different birthdays is the same as the event that person 2 does not have the same birthday as person 1, and that person 3 does not have the same birthday as either person 1 or person 2, and so on, and finally that person 23 does not have the same birthday as any of persons 1 through 22.

Let these events respectively be named "Event 2", "Event 3", and so on. One may also add an "Event 1", corresponding to the event of person 1 having a birthday, which occurs with probability 1. This conjunction of events may be computed using conditional probability: the probability of Event 2 is $364/365$, as person 2 may have any birthday other than the birthday of person 1. Similarly, the probability of Event 3 given that Event 2 occurred is $363/365$, as person 3 may have any of the birthdays not already taken by persons 1 and 2. This continues until finally the probability of Event 23 given that all preceding events occurred is $343/365$. Finally, the principle of conditional probability implies that the probability $P(A')$ that no two people in the room have the same birthday is equal to the product of these individual probabilities:

$$P(A')=\frac{365}{365}\cdot\frac{364}{365}\cdot\frac{363}{365}\cdot\frac{362}{365}\cdot\ldots\cdot\frac{343}{365}=\left(\frac{1}{365}\right)^{23}\cdot(365\cdot 364\cdot 363\cdot\ldots\cdot 343)\cong 0.492703 \tag{5.6.24}$$

Therefore, the probability that at least two people in the class have the same birthday is given by (as both events are mutually exclusive):

$$P(A)=1-P(A')\cong 1-0.492703\cong 0.507297 \tag{5.6.25}$$

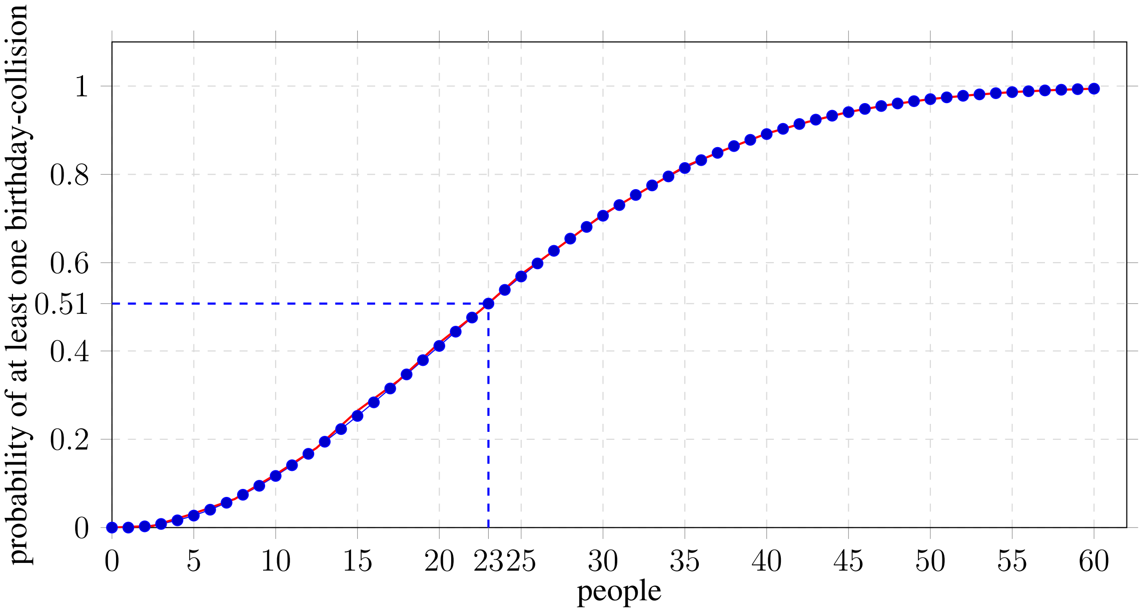


Figure 5.77: **Birthday paradox probability distribution**

One creationist-intelligent design argument (high-school level argument...) goes like this: the human alpha-globin molecule (abbreviated HBA1), a component of haemoglobin that performs a key oxygen transfer function, is a protein chain based on a sequence of 141 amino acids[21] (average mass of 10^{-25} [kg]), keeping in mind that the smallest known protein at the day we write these lines is made of only 10 amino acids (see [260]). There are 20 common different amino acids in living systems[22]:

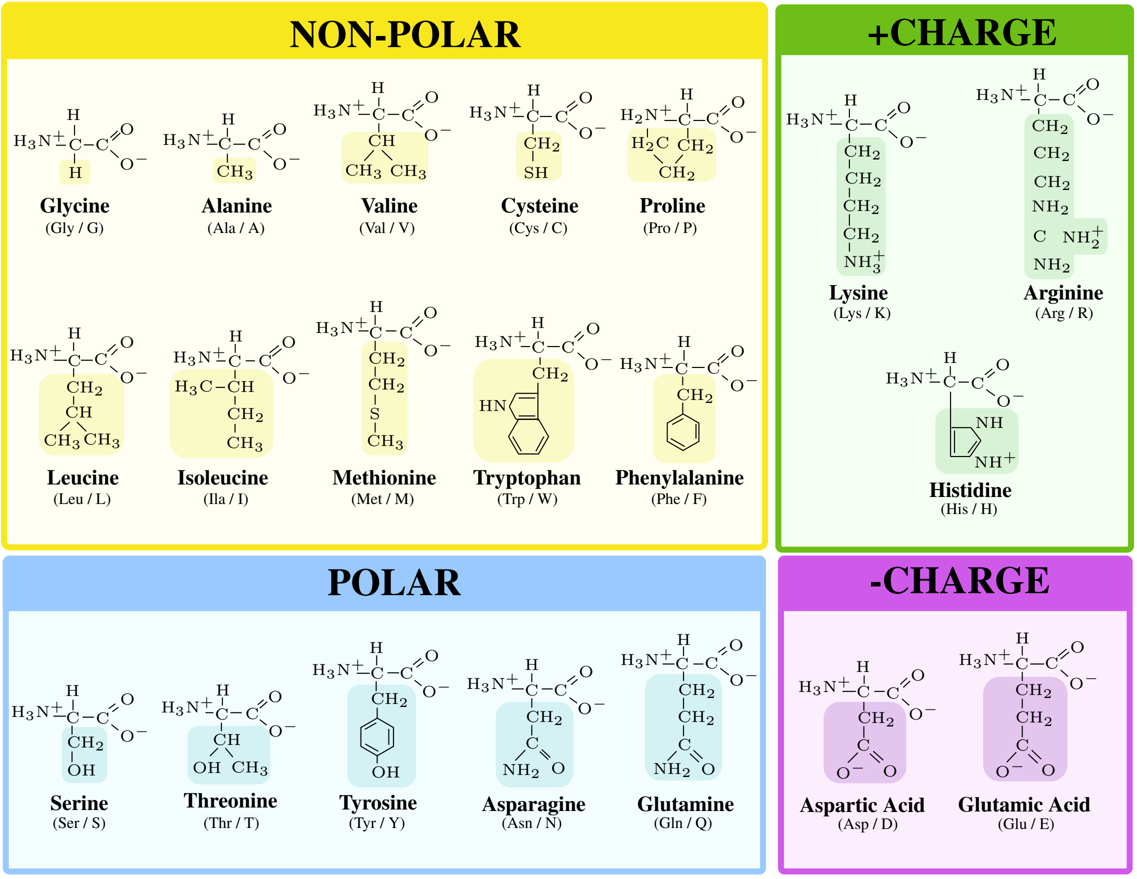

NON-POLAR
Glycine
(Gly / G)
Alanine
(Ala / A)
Valine
(Val / V)
Cysteine
(Cys / C)
Proline
(Pro / P)
Leucine
(Leu / L)
Isoleucine
(Ila / I)
Methionine
(Met / M)
Tryptophan
(Trp / W)
Phenylalanine
(Phe / F)
+CHARGE
Lysine
(Lys / K)
Arginine
(Arg / R)
Histidine
(His / H)
POLAR
Serine
(Ser / S)
Threonine
(Thr / T)
Tyrosine
(Tyr / Y)
Asparagine
(Asn / N)
Glutamine
(Gln / Q)
-CHARGE
Aspartic Acid
(Asp / D)
Glutamic Acid
(Glu / E)

Figure 5.78: Common amino-acids

Amino-acids have been reproduced with natural phenomenon in the Miller-Urey experiment (see [179]) and have also been extracted in meteorites (see [117]):

---

[21] Haemoglobin molecule is made up of four polypeptide chains: two alpha chains of 141 amino acids residues each and two beta chains of 146 amino acids residues each.

[22] Notice that a total of 23 types of amino acids were found in Ryugu asteroid samples taken in November 12019 (holocene calendar) and brought back on Earth in December 12020 (holocene calendar) by Japan's Hayabusa 2 space probe. This gives a strong indication that amino acides also form in outer space.

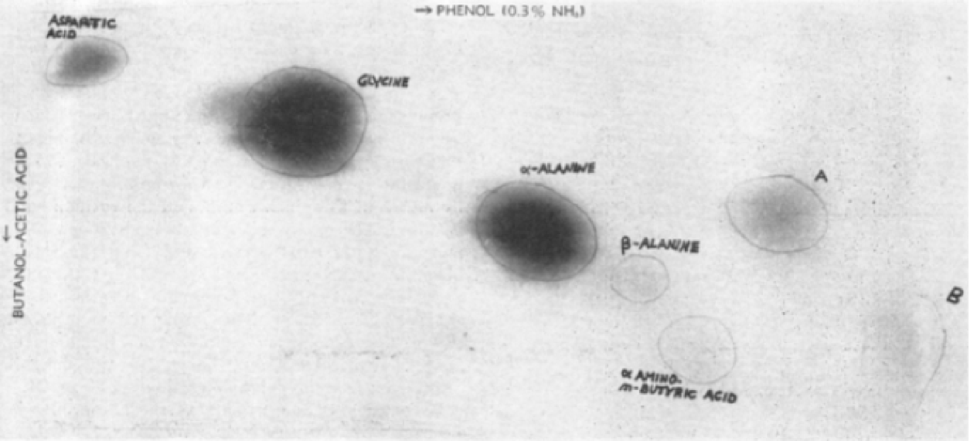

ASPARTIC ACID
→ PHENOL (0.3% NH₃)
GLYCINE
BUTANOL-ACETIC ACID ←
α-ALANINE
A
β-ALANINE
B
α AMINO-n-BUTYRIC ACID

Figure 5.79: **Paper chromatography performed by Miller, published on the 11953 (holocene calendar) paper, in which the amino-acids found in his experiment can be observed**

More recent reproduction of the Miller-Urey experiment like the one in 12025 (holocene calendar) has not only once again confirmed the production of amino-acids but also of concomitant formation of protocells and prebiotic compounds (see [291]):

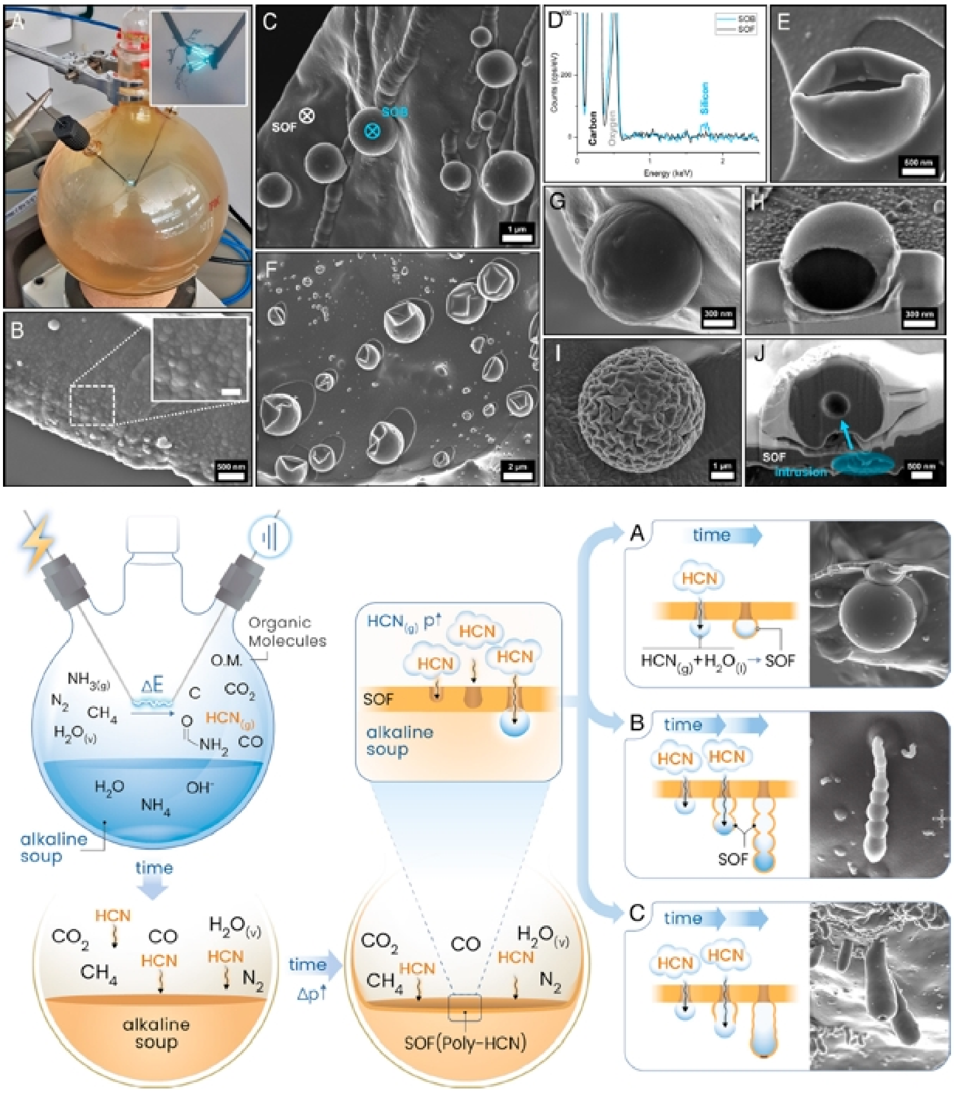

A
B
C
SOF
SOB
1 µm
D
Counts (cps/eV)
Carbon
Oxygen
Silicon
Energy (keV)
SOB
SOF
E
500 nm
F
2 µm
G
300 nm
H
300 nm
I
1 µm
J
SOF
Intrusion
500 nm
500 nm
Organic Molecules
O.M.
$NH_{3(g)}$
ΔE
C
CO_2
N_2
CH_4
O
$HCN_{(g)}$
$H_2O_{(v)}$
NH_2
CO
H_2O
OH^-
NH_4
alkaline soup
time
HCN
CO_2
CO
$H_2O_{(v)}$
CH_4
HCN
HCN
N_2
alkaline soup
time
Δp↑
$HCN_{(g)}$ p↑
HCN
HCN
HCN
SOF
alkaline soup
CO_2
CO
$H_2O_{(v)}$
HCN
HCN
CH_4
N_2
SOF(Poly-HCN)
A
time
HCN
$HCN_{(g)}+H_2O_{(l)}$ → SOF
B
time
HCN
HCN
SOF
C
time
HCN
HCN

Figure 5.80: **Concomitant formation of protocells and prebiotic compounds in the Miller-Urey experiment**

With the sequencing of the alpha-globin that is:

```
Val-Leu-Ser-Pro-Ala-Asp-Lys-Thr-Asn-Val-Lys-Ala-Ala-Trp-Gly-Lys-Val-Gly
-Ala-His-Ala-Gly-Glu-Tyr-Gly-Ala-Glu-Ala-Leu-Glu-Arg-Met-Phe-Ser-Phe-Pro
-Thr-Thr-Lys-Thr-Tyr-Phe-Pro-His-Phe-Leu-Ser-His-Gly-Ser-Ala-GIn-Val-Lys
-Gly-His-Gly-Lys-Lys-Lys-Val-Ala-Asp-Ala-Leu-Thr-Ala-Val-His-Val-Hal-Hal
-Hsp-Met-Pro-Asn-Ala-Leu-Ser-Ala-Leu-Ser-Asp-Leu-His-Ala-Heu-Arg-Val-Asp
-Pro-Val-Asn-Phe-Lys-Leu-Leu-Ser-His-Cys-Leu-Leu-Leu-Ala-His-Leu-Pro-Ala
-Glu-Phe-Thr-Pro-Ala-Val-His-Ala-Ser-Leu-Asp-Lys-Phe-Leu-Ala-Ser-Val-Ser
-Thr-Val-Leu-Thr-Ser-Lys-Tyr-Arg
```

Or for an abbreviated version of the four polypeptide chains:

- Chain A:
  ```
  VLSPADKTNV KAAUGKVGAH AGEYGAEALE RMFLSFPTTK TYFPHFDLSH GSAOVKGHGK
  KVADALTHAV AHVDDMPMAL SALSDLHAHK LRVDPVNFKL. LSHCLLVTLA AHLPAEFTPA
  VHASLDKFLA SVSTVLTSKY R
  ```
- Chain B:
  ```
  VHLTPEEKSA VTALUGKVNV DEVGGEALGR LLVVYPUTOR FFESFGDLST pDAVHGWRV
  KAHGKKWLGA FSDGLAHLDN LKGTFATLSE LHCDKLHWDP ENFRLLGNVL VCVLAHHFGK
  EFTPPVOAAY OKVVAGVANA LAHKYH
  ```
- Chain C:
  ```
  WLSPADKTNV KAAUGKVGAH AGEYGAEALE RMFLSFPTTK TYFPHFDLSH GSAOVKGHGK
  KVADALTNAV AHVDDHPNAL SALSDLHAHK LRVDPVNFKL LSHCLLVTLA AHLPAEFTPA
  VHASLDKFLA SVSTVLTSKY R
  ```
- Chain D:
  ```
  WHLTPEEKSA VTALUGKVNV DEVGEEALGR LLWYPUTOR FFESFGDLST PDAVMGNPKV
  KAHGKKVLGA FSDGLAHLDN LKGTFATLSE LHCDKLHVDP ENFRLLGNVL VCVLAHHFGK
  EFTPPVQ
  ```

So the number of potential chains of length 141 is[23] (see further below page 523 for the proof):

$$\bar{A}_m^n = n^m = 20^{141} \cong 10^{183} \tag{5.6.26}$$

The creationists and also many theists argue then that this figure is so enormous that even after billions of years of random molecular trials, no human alpha-globin protein molecule would ever appear "at random," and thus the hypothesis that human alpha-globin arose by an evolutionary process is decisively refuted.

---

[23] This calculation is anyway quite wrong as symmetrical permutations (mirror results) are counted - but it won't however change the order of magnitude of the result - and experimental evidence have shown that there are forbidden/unstable sequences of amino acids!

Definition (#5.6.104): "Life" (i.e. "autocatalytic set" in chemistry) is a chemical system capable of undergoing biological evolution (see definition above at page 459) through self-replication[24] using the resources of its environment (notice again the subtlety of this definition referring to that of "biological evolution" to avoid to have the "fire" classified as a living entity...!).

Remark

In 12023 (holocene calendar), Ryo Mizuuchi and Norikazu Ichihashi (see [396]), identified a 20-nucleotide RNA variant that self-reproduces! As a nucleotide is made of 5 carbon atoms, 8 hydrogen atoms, 1 nitrogen atom, 1 phosphorus atom, 4 oxygen atoms therefore a single nucleotide contains 19 atoms (the number of possible combinations based on the highest valence case of each atom is therefore $4^5 1^8 5^1 5^1 2^4 = 409,600 \cong 10^{5.612}$ hence much much smaller than the 10^{183} argument used by creationists...). That means the smallest known RNA that can self-reproduce has only $19 \cdot 20 = 380$ atoms! Furthermore, another scientific team (see [354]) was able to develop spontaneous emergence of self-replicating molecules containing nucleobases and amino acids (what some people qualify of "creating life from non-life" or "non-life producing life") and another team was able to develop synthetic self-replicating cells (see [308])! Another interesting paper is the one relating to an experiment observing a small non-synthetic polymerase of only 45 nucleotides from a pool of random sequences that reproduces itself (see [201]).

One major fallacy in the alpha-globin argument mentioned above, common to many others of this type, is that it ignores the fact that a large class of alpha-globin molecules can perform the essential oxygen transfer function, so that the computation of the probability of a single instance is misleadingly remote. Indeed, most of the 141 amino acids in alpha-globin can be changed without altering the key oxygen transfer function, as can be seen by noting the great variety in alpha-globin molecules across the animal kingdom. When one revises the calculation above, based on only 25 locations essential for the oxygen transport function (which is a generous over-estimate), one obtains 10^{33} fundamentally different chains, an enormous figure but incomparably smaller than 10^{183}. And keep in mind that it's the number if we assume a total independence! The slightest constraint can greatly reduce this value. Keep in mind, for example, the problem of the 100 prisoners whose probability goes from $8 \cdot 10^{-31}$ to about $\sim 31\%$ just by adding a stupid constraint (see Wikipedia for the explanation of this problem and the detailed proof)...

A calculation such as this can be refined further, taking into account other features of alpha-globin and its related biochemistry. Some of these calculations produce probability values even more extreme than the above. But do any of these calculations really matter[25]? The main problem is that all such calculations, whether done accurately or not, suffer from the fatal fallacy of presuming that a structure such as human alpha-globin arose by a single all-at-once random trial event. But generating a molecule "at random" in a single shot is decidedly not the scientific hypothesis in question – this is a creationist theory, not a scientific theory. Instead, available evidence from hundreds of published studies on the topic has shown above any reasonable doubt that alpha-globin arose as the end product of a long sequence of intermediate steps, each of which was biologically useful in an earlier context.

---

[24] Autocatalysis, the ability of systems (molecules, metabolic networks or compartments) to make copies of themselves, is central to all evolutionary scenarios. Systems where autocatalysis is accompanied by information transfer and heredity are said to be self-replicating!

[25] For example a physicist applying the scientific method with the frequentist approach will calculate the odds of the appearance of life with the available experimental evidence! Thus the current data available to us gives us 1 observable universe and 1 life, meaning 100% probability (i.e. the probability of observing life statistically according to the available sample).

In short, the creationist-intelligent design argument claiming that scientists assert an all-at-once "at random" creation of various biomolecules, and then asserting that this is probabilistically impossible, is a classic "straw man" fallacy. Scientists do not believe this, so this line of argumentation is completely invalid. In other words, it does not matter how good or how bad the mathematics used in the analysis is, if the underlying model is a fundamentally invalid description of the phenomenon in question. Any simplistic probability calculation of evolution that does not take into account the step-by-step process by which the structure came to be is almost certainly fallacious and can easily mislead!

Some of the difficulties with creationist probability arguments can be illustrated by considering snowflakes (even if we must keep in mind that *comparison is not reason*!). Bentley and Humphrey's book *Snow Crystals* includes over $2,000$ high-resolution black-and-white photos of real snowflakes, each with intricate yet highly regular patterns that are almost perfectly six-way symmetric (see [350] for more details).

By employing a reasoning based on six-way symmetry, one can calculate the chances that one of these structures can form "at random" as roughly one part in 10^{2500}. This probability figure is even more extreme than some that have appeared in the creationist-intelligent design literature. So is this proof that each individual snowflake has been designed by a supernatural intelligent entity? Obviously not.

One of the fallacy here, once again, is presuming an all-at-once random assembly of molecules. Instead, snowflakes, like biological organisms, are formed as the product of a long series of steps acting under well-known physical laws, and the outcomes of such processes very sensitively depend on the starting conditions and numerous environmental parameters. It is thus folly to presume that one can correctly reckon the chances of a given outcome by means of superficial probability calculations that ignore the processes by which they formed.

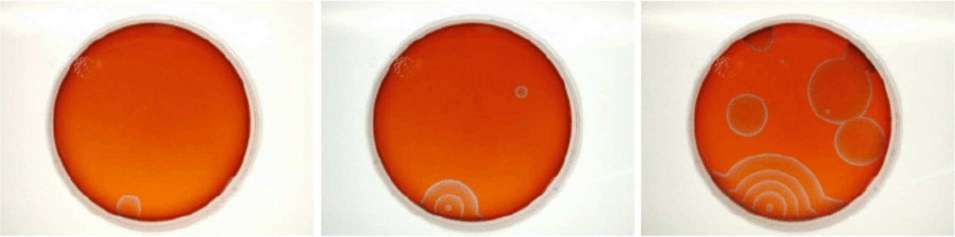
Remark

Notice also the art of cherry picking from theists and creationists. A deity created something "perfect" as DNA (or trees whatsoever!). However he did it with almost 4 or 6 thousands genetic disease (coding errors), with a light source (Sun) that gives cancer, on a planet (Earth) where the majority of water is undrinkable[a], near a star (Sun) that will explode, in a orbit that will make the planet unlivable in a few billion years, in a finite resource ecosystem with deadly rampant diseases, in a universe in majority uninhabitable... Yes logic makes sense!

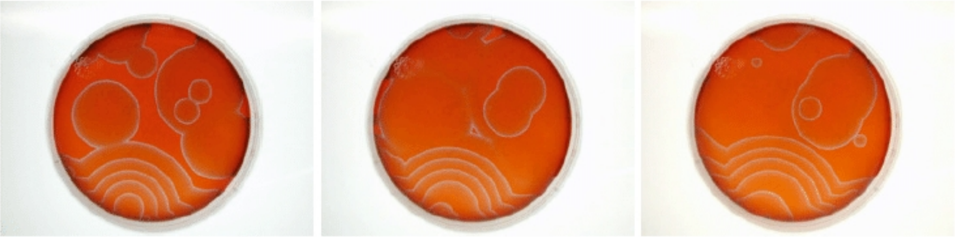
---

[a] Water quantities types varies hugely across time on Earth (for example new research indicates that our planet was an ocean world some 3 billion years ago). But actually in the 21st century water from oceans and marginal seas, saline groundwater and water from saline closed lakes amount to over 97% of the water on Earth (corresponding to approximately $70.9 \pm 3.5\%$ of Earth's area, seas represents approximetely 11.3%). Only 2.5% is made of freshwater (itself made of 1.2% of surface water, 30.1% of ground-water and 88.7% of glaciers and ice caps).

Another famous example of randomness creating structures are the Belousov–Zhabotinsky reaction, that is a class of reactions that serve as a classical example of non-equilibrium thermodynamics, resulting in the establishment of a non-linear chemical oscillator (it is "random" in the way that the final patterns we get are very unlikely to be identical between two same experimental configuration!):



(a) Initial (b) $t = 3sec$ (c) $t = 9sec$



(d) $t = 12sec$ (e) $t = 15sec$ (f) $t = 18sec$

It's time to put some numbers on the formulas of statistical mechanics to bring home just how impressive the things going on inside our cells actually are.

To start, we live (actually...) at temperatures of approximately 300 [K]. If you have studied statistical mechanics you know that the kinetic energy of a molecule is given by (see section Continuum Mechanics page 3897):

$$E_c=\frac{N}{2}mv^2=\frac{3}{2}kNT=\frac{3}{2}\frac{R}{N_A}NT \tag{5.6.27}$$

Therefore:

$$\bar{v}=\sqrt{\frac{3RT}{mN_A}} \tag{5.6.28}$$

Or in terms of molar mass:

$$\bar{v}=\sqrt{\frac{3RT}{M_m}} \tag{5.6.29}$$

Therefore for water with $18.01528\cdot 10^{-3}$ $[\text{kg}\cdot\text{mol}^{-1}]$ this gives at room temperature:

$$\bar{v}=\sqrt{\frac{3\cdot 8.314\cdot 300}{18.01528\cdot 10^{-3}}}\cong 644\;[\text{m}\cdot\text{s}^{-1}] \tag{5.6.30}$$

This is not a too bad result as many sources give 590 $[\text{m}\cdot\text{s}^{-1}]$. We will take that latter value.

The thermal velocity relation works even for something as large as RNA polymerase II ($\sim 830\cdot 10^{-24}$ [kg]).

To make things really easy lets work with a molecular complex of mass even bigger (i.e. more complex). Something as: $1\cdot 10^{-21}$ [kg]. Such a mass would have an average thermal velocity of:

$$\bar{v}=\sqrt{\frac{3RT}{mN_A}}\cong\sqrt{\frac{3\cdot 8.314\cdot 300}{1\cdot 10^{-21}\cdot 6.0221\cdot 10^{23}}}\cong 3.52\;[\text{m}\cdot\text{s}^{-1}] \tag{5.6.31}$$

Let us take a lower bound of 2 $[\text{m}\cdot\text{s}^{-1}]$ according to some non peer-reviewed sources.

Cells are small. The three polymerases transcribed DNA into RNA have masses in the order of $1\cdot 10^{-21}$ [kg] range. So how long should it take them to go through a quite big cell nucleus of 10 microns (10^{-5} [m]) in diameter? If it's going at 2 $[\text{m}\cdot\text{s}^{-1}]$ so it will traverse it statistically in average $250,000$ times in a second[26] or one time every 4 $[\mu\text{s}]$.

We clearly left something out — nothing in the cell moves in a straight line. It is very crowded, so that even though things are moving very quickly their trajectory isn't straight obviously!

We have proved in the section of Continuum Mechanics (page 3925) that the average distance travelled by a molecule (assumed as spherical) between two collisions is given by:

$$\lambda=\frac{1}{n\pi r^2\sqrt{2}} \tag{5.6.32}$$

---

[26] For people interested to see a numerical simulation video of a Brownian motion of proteins inside a bacterial cytoplasm like that of E. Coli, see [379].

And we proved at the same place that the number of collision per second were given by:

$$C=n\pi r^2\sqrt{2}\bar{v}=n\pi r^2\sqrt{2}\sqrt{\frac{3RT}{mN_A}} \tag{5.6.33}$$

To calculate how much water can fit into a cell nucleus, we need to know how big is that latter. A source says that water can be considered a squashed sphere of maximum radius of 1.41 Angstroms (10^{-10} [m]).

So what is the volume of a water molecule? It's:

$$V_{\text{H}_2\text{O}}=\frac{4}{3}\pi r^3=\frac{4}{3}\pi(1.41\cdot 10^{-10})^3\cong 1.174\cdot 10^{-29}\;[\text{m}^3] \tag{5.6.34}$$

So the density of molecules is given by:

$$n=\frac{1}{1.174\cdot 10^{-29}}\cong 8.5\cdot 10^{28}\;[\text{molecules}\cdot\text{m}^{-3}] \tag{5.6.35}$$

Therefore:

$$\lambda=\frac{1}{n\pi r^2\sqrt{2}}=\frac{1}{3.5\cdot 10^{14}\cdot\pi\cdot(1.41\cdot 10^{-10})^2\sqrt{2}}\cong 1.4\cdot 10^{-10}\;[\text{m}] \tag{5.6.36}$$

and (still for water only and in one cube meter!):

$$C=n\pi r^2\sqrt{2}\bar{v}=8.5\cdot 10^{28}\cdot\pi\cdot(1.41\cdot 10^{-10})^2\cdot\sqrt{2}\cdot 644\cong 4.8\cdot 10^{12}\;[\text{collisions}\cdot\text{s}^{-1}] \tag{5.6.37}$$

so it's like to meet another molecule of water every (roughly) 10^{-12} seconds (1 pico-second).

Notice a billion years has:

$$60\cdot 60\cdot 24\cdot 365\cdot 1\cdot 9\cong 3.15\cdot 10^{16}\;[\text{s}] \tag{5.6.38}$$

So we have roughly by billion years a total of collision for one cubic meter equal to:

$$C_{\text{bill. years}}=4.8\cdot 10^{12}\cdot 3.15\cdot 10^{16}\cong 1.5\cdot 10^{29}\;[\text{collisions}\cdot\text{billion years}^{-1}] \tag{5.6.39}$$

And as the amount of water on Earth is estimated roughly at (assuming that all this water was involved equally on the whole planet to the same chemical processes and conditions...):

$$V_{\text{Earth H}_2\text{O}}\cong 1\cdot 10^{18}\;[\text{m}^3] \tag{5.6.40}$$

Therefore on the whole Earth we have:

$$C_{\text{Tot}}\cong 1.5\cdot 10^{29}\cdot 1\cdot 10^{18}\cong 1\cdot 10^{36}\;[\text{collisions}\cdot\text{billion years}^{-1}] \tag{5.6.41}$$

So per billion years we have $1,000$ times more collisions than the number 10^{33} of combinations of different chains of the alpha-globin. Even if we consider a factor 100 of errors in the number of collisions per second (it's not impossible that it is our magnitude of error in our estimations!), we would need only 1 billion years to try all 10^{33} combination just by random collisions.

Keep in mind that this is the result for considering only one Earth planet like (ours). Whatever the result, we should have multiplied it by the lower bound approximation of Earth like planets in our Galaxy (estimated in the year 12020 according to holocene calendar between 1 to 6 billion, i.e. $\sim 10^9$, see [36]) and the number of galaxies in the observable Universe (estimated between 100 billion and 2 trillion, i.e. between 10^{11} and 10^{12}, see [111]). And this is only for the observable Universe... (keeping in mind that the Universe is to our actual knowledge very likely flat and therefore infinite with $\Omega_0 = 1$ or $k = 0$ as seen in the section of Cosmogony page 5897 or at least 250 bigger as detailed in [555]).

At the time we write these lines, early 121st century according to Holocene calendar, it is pretty much sure that thanks to the James Webb Telescope biosignatures (i.e. life) have been found on other planets, it's just that astrophysicists are not yet allowed to say it. In the professional circle of astrophysicists in spectroscopy it is known that there are so many plausible exoplanetary atmospheric biosignatures that have been detected that there's no longer any doubt, but the standard of proof for official acceptance is just incredibly high. The most statistical significant case - at a 3σ level actually - being the planet K2-18b (this planet is a little larger than Earth in a solar system about 124 light years away in the habitable zone of a red dwarf) with the detection of dimethyl sulfide at concentrations that seem to be thousands of higher than on Earth. On Earth this gas is produced almost exclusively by marine microbes.

The results of these calculations demonstrate that the assertions about the "impossible" probabilities of life - made without mathematical proof by Nobel laureate George Wald and by astronomer Fred Hoyle - are entirely unfounded. This illustrates why even Nobel Prize winners, like all scientists, should refrain from making authoritative claims on topics that fall outside their own field of expertise

Let us quote one of the 12019 (holocene calendar) Nobel Prize in Physics:

> *The emergence of life is ultimately a chemical process. If the conditions are right, life will emerge. I see no need for an initial push from a creator god.*
>
> Didier Queloz

quote 5.6:469.#43

Anyway the reader must remember that the probability of something happening as a determinate to it happening is not relevant if it has actually happened!

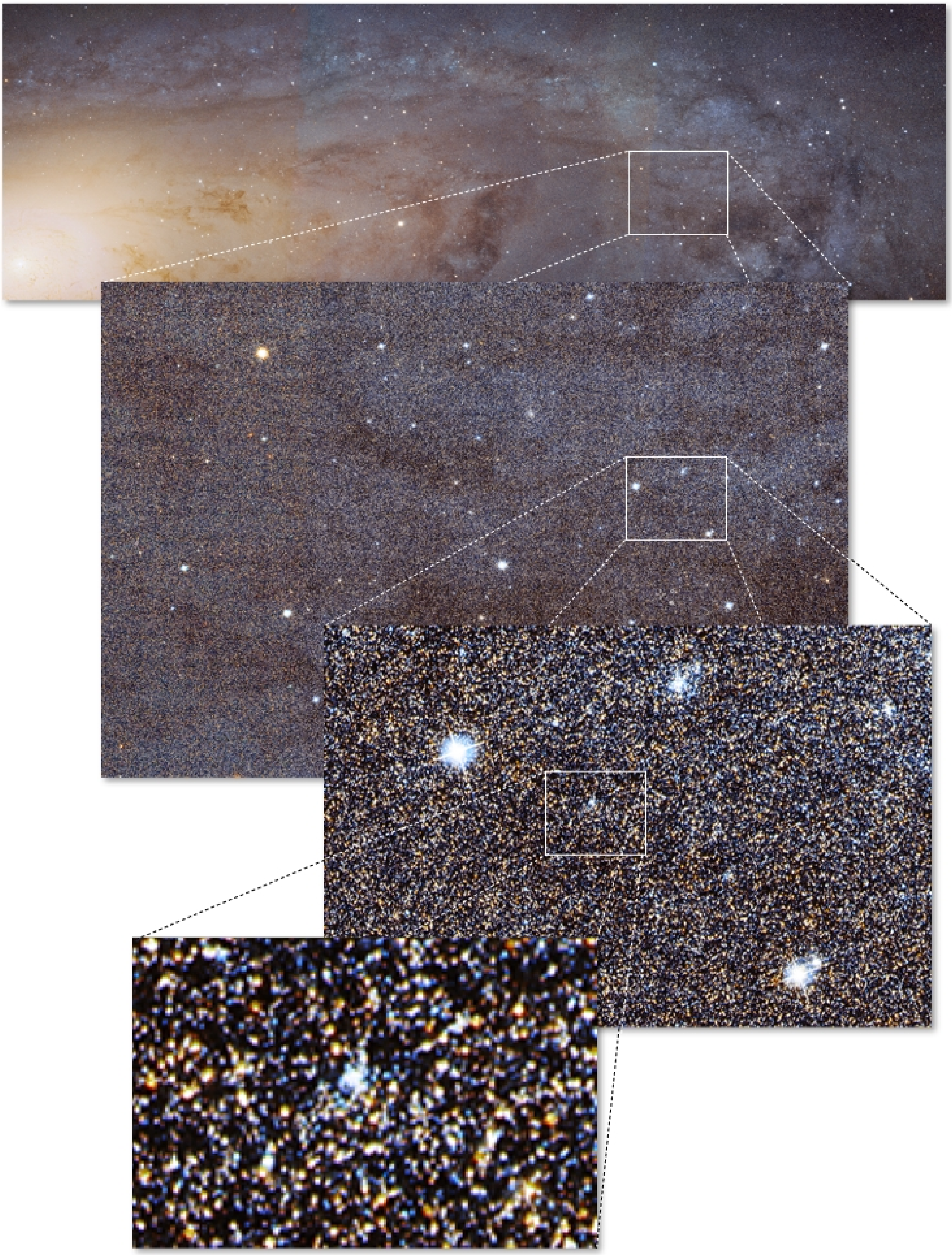


Figure 5.81: Hubble space telescope sharpest and largest M31 (Andromeda galaxy) image ever with 1.5 billion pixels with approximately 100 million stars across over 40,000 light-years released the in 12015, 5 January (credit: NASA, ESA, J. Dalcanton (University of Washington, USA), B. F. Williams (University of Washington, USA), L. C. Johnson (University of Washington, USA), the PHAT team, and R. Gendler)

Remark

So far we see that the main factor here for the number of collisions is not the radius of the molecules involved, nor their speed or their mass but their quantity (i.e. density). Obviously the time is important but there is now too much incertitude about the range of that latter.

But let us be more realistic! We should take into account that:

1. Water is not amino acids (the biggest amino acid has a weight of 204.2 $[\text{g}\cdot\text{mole}^{-1}]$) and there are elastic collisions between water molecules and such organic molecules.
2. The water molecules density is not equal to that of amino acids density per litre of water. Sadly this is a value it seems we don't have access to and is highly speculative[27].

Let us consider the relation for an elastic collision in one dimension (see section Classical Mechanics page 3369):

$$v_{2f}=\frac{2m_1}{m_2+m_1}v_{1i}+\frac{m_2-m_1}{m_2+m_1}v_{2i} \tag{5.6.42}$$

And let us consider the worst case where $v_{2i}=0$. Then:

$$v_{2f}=\frac{2m_1}{m_2+m_1}v_{1i}=\frac{2\cdot 18}{18+204.2}\cdot 644\cong 104\;[\text{m}\cdot\text{s}^{-1}] \tag{5.6.43}$$

So the speed of amino acids would only decrease our number of collisions by maximum a factor 10. Not enough to make "life" something rare on the scale of a billion years (to make it rare we should have at least a decrease in an order of $10,000$).

Let us do now some reverse engineering. To make emerging life rare on a billion years scale, as already mentioned and thanks to previous calculations, we should have a least a decrease of $10,000$. Thanks the to the elastic collision, we can take down this value to $1,000$.

So we should found somewhere, something in our calculations, to make the collisions $1,000$ times more rare. The idea would take this into the parameter n (molecules density) in the above relations. If we assume that the ratio of amino acid to water molecules is at 1 for $1,000$ then it starts to make things difficult on the scale of a billion years! That means in 1 mole of water (i.e. 18 [g] or 18 [ml]), we would have 0.018 [g] of amino acid.

So as other extra-solar planets were something unlikely a century ago, it is very likely that the future provides evidence that life is something very common on planets that have water or atmospheres with the right components and temperature on a time range of the billion years.

Remark

Some scientifically illiterate people may argue that Roger Penrose calculated that the probability for our Universe (on the assumption that the latter is in a Black Hole...) to be in the low entropy he is now is about 10^{123} to 1. Hence it is impossible for our Universe to not have been created by a deity (passing under silence the question of who created that complex deity...). But it is irrelevant! Even if what Penrose said is 100% accurate or off by a factor of a googol, it's irrelevant!! Indeed, if we run a computer program that emulates a coin flipped one billion times and that we write the sequence of heads and tails to a file, then we just produced a sequence more than a trillion trillion trillion times rarer!! In fact, we could run that same computer program until the end of human existence and it would likely never produce that same string of outcomes. And yet it occurred nonetheless...

---

[27] It is related to the problem of proportion compounds of the Miller-Urey experiment.

If one feels inclined, one can try to estimate the probability of one's own existence. A naïve calculation would multiply the chance that a specific sperm met a specific egg by the chance that one's parents happened to meet, that one's grandparents crossed paths, and so on through the entire ancestral line. Unsurprisingly, the resulting probability is vanishingly small. Yet this is precisely the point: events of extremely low probability occur constantly in our universe. And once such an event has happened, its probability becomes 100%.

Remark

The reader will also note the art of cherry picking among believers and creationists. A deity would have created something "perfect" like DNA (or trees, whatever!). However it would have done it with nearly 4 to 6 thousand genetic diseases (coding errors), with a source of light (Sun) which gives cancer, on a planet (Earth) where the majority of water is undrinkable near a star (Sun) which will explode and which is in an orbit which will make it unlivable in a few billion years, with an ecosystem of limited resources with endemic deadly diseases, in a universe mostly uninhabitable... Yes that makes sense hmmm...

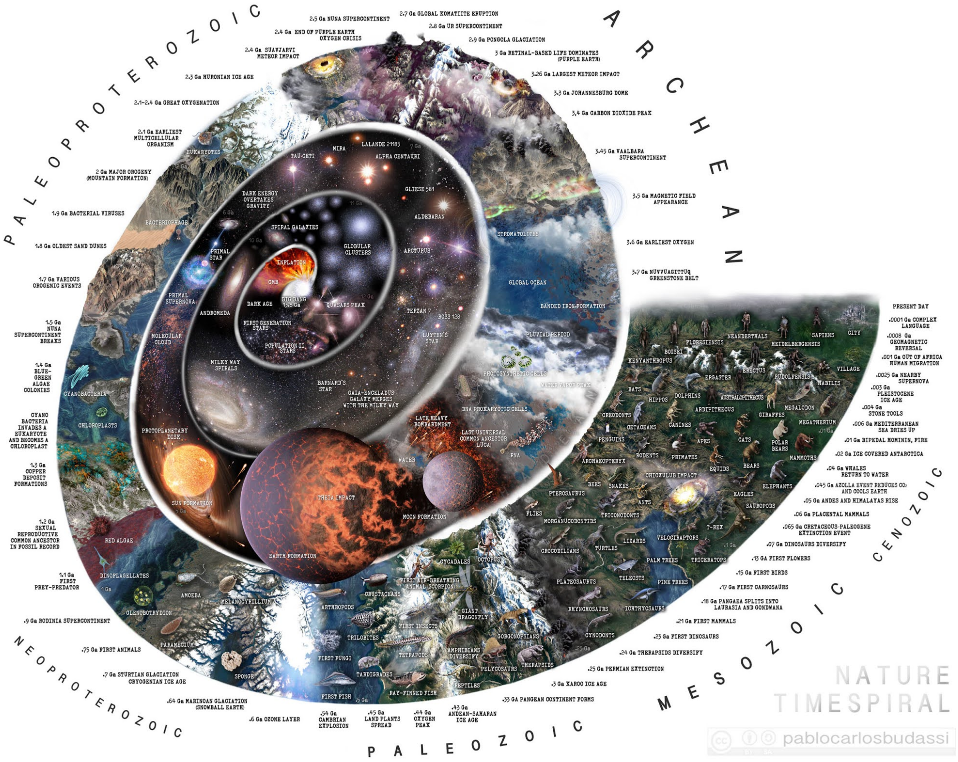


Figure 5.82: Nature Timespiral (author: Pablo Carlos Budassi)

By the way in 12025 (holocene calendar) scientists have found abundant ammonia and nitrogen-rich soluble organic matter in samples from asteroid (101955) Bennu (see [205]). This includes amino acids (including 14 of the 20 used in terrestrial biology), amines, formaldehyde, carboxylic acids, polycyclic aromatic hydrocarbons and N-heterocycles (including all five nucleobases found in DNA and RNA).

Remark

Another fallacy of bigots is that the probability of cellular life's arising from non-living matter (abiogenesis) as calculated by Fred Hoyle is about one-in-$10^{40,000}$. But there is no sources from Fred Hoyle with the detailed mathematical derivation that lead to this value. Another lie from religious people...! There is also the Junkyard tornado analogy used by bigots but as we have already proved it: all analogies are wrong equivalences!

> ❞ *I have your letter and the best thing I can do is refer to my published works, both scientific and popular. The creationist movement is lead by a dishonest bunch of operators and misquotation is the hall mark of their work. Responding to them is time wasting and a letter would not be adequate to put your questions to rest. There are some things best ignored and the stupidity of these so called religious fanatics continues to astonish me.* ❞
>
> Richard Leakey

quote 5.6:473.#44

Description	**Probability** (or estimated)
Probability of dying one day or another	100%
Probability of >7 Richter scale earthquake somewhere on Earth during your lifetime	100%
Probability of erupting volcano somewhere on Earth during your lifetime	100%
Six Sigma management probability to produce a defect-free product	99.99966%
Banking regulatory sector (Basel Committee 12022) Value at Risk (VaR) threshold	99%
Banking regulatory sector (Basel Committee 12022) Expected Shortfall (ES) threshold	97.5%
Prior statistical test typical power (type II error)	80%
Probability of flipping a coin and it landing on heads	50%
Probability of experiencing a pandemic in one's lifetime	$\sim$ 38%
The 100 prisoners problem optimal solution	$\sim$ 31.18%
Probability of massive star forming a black hole (unovae)	$\lesssim$ 10-30%
Probability of identical birthday in a group of 23 people	50%
Probability of dying from any given cancer during lifetime	$\sim$ 18%
Probability of rolling a six on a six-sided dice	$16.\bar{6}$%
Probability of miscarriage	$\sim$ 15.3%
Typical p-value threshold in social and medical science (type I error)	5%
Probability of a baby to born with a birth defect	$\sim$ 2.63%
Probability of winning roulette game in a casino	$\sim$ 2.6%
Probability of drawing an Ace of Spades from a deck of cards	$\sim$ 1.92%
Probability of having a stroke for people aged between 35-40	1.4-2.3%
Typical p-value threshold in engineering science (type I error)	1%
Probability of dying from a heart attack during lifetime	$\sim$ 0.99%
Probability of a newborn to have a genetic anomaly	$\sim$ 0.2%
Probability 10 km size asteroid colliding Earth during your lifetime	0.0000830%
Probability that you will die tomorrow as "average" human	0.000001%
Typical p-value threshold in particle physics (type I error)	$\sim$ 0.000001%
Probability of getting a royal flush in poker	0.00000154%
Probability of winning the Powerball jackpot	0.00000021%
Fatality probability for an average person using an average commercial flight	$\sim 3\cdot 10^{-8}$
Probability of two identical fingerprints according to Jackson Gold and Maria Cuellar	$\sim 7\cdot 10^{-8}$
Fatality probability for an average person crossing an average road	$\sim 4\cdot 10^{-9}$
Fatality probability for an average person using an average lift	$\sim 1\cdot 10^{-10}$
Probability of tunnel effect in polonium decay process	10^{-15}
Probability of alpha-globin formed completely by chance based on amino-acides on Earth like planets in the observable Universe	$\sim 10^{-4}$-10^{-10}
Probability of alpha-globin formed completely by chance based on amino-acides only on Earth	$\sim 10^{-25}$-10^{-31}
Probability of two identical fingerprints according to Francis Galton	2^{-36}
Probability of the actual entropy of our Universe according to Roger Penrose calculation	$\sim 10^{-123}$%
Probability of rolling fifty six-sided dices and get a given result	$\sim 10^{-300}$%
Probability a monkey rewrites *Hamlet* by chance	$\sim 10^{-182,856}$
Probability of being born	$\sim 10^{-2,685,000}$
Probability of human to cross a wall by tunnelling effect	$\sim 10^{-4.10^{30}}$
Probability of our universe to exist as it is	$\sim 10^{-272,000}$ -$10^{-10^{500}}$

Table 5.25: **Some famous probabilities (and estimations) values as known in the early 121st century**

> *Nature wants 5 of your 7 children dead. It wants you dead by 50. Everything better than that is brought to you by science and technology.*
>
> David Frum

quote 5.6:474.#45

5.6.2 Kolmogorov's Axioms

In Chapter 2 of *A First Course in Probability*, Sheldon Ross cautions us against "defining" probability as the limiting proportion of events over repeated samples/ trials/experiments: «*...it does not at all seem to be a priori evident that this [limiting frequency] need be the case. Also,..., how do we know that if the experiment is repeated..., we shall again obtain the same limiting proportion...? ...would it not be more reasonable to assume a set of simpler... axioms... and then attempt to prove [mathematically] that such a constant limiting frequency does in some sense exist... [using] the law of large numbers... This latter approach is the modern axiomatic approach to probability theory...*»

On the surface, this seems entirely appropriate; however, under this formalism, it appears we must treat probability, P, as an ethereal and ephemeral characteristic of an event A pertaining to any particular trial - i.e., an unfalsifiable causative propensity. Then, the strong law of large numbers (SLLN) is a theorem to be rigorously proven[28] mathematically stating that: the limiting proportion of events over repeated trials converges to this very same propensity. Ie:

$$\lim_{n\to+\infty}\sum_{i=1}^{n}\frac{X_i}{n}=\bar{X} \qquad (5.6.44)$$

From this perspective, the existence of a limiting proportion is a consequence resting on unfalsifiable axioms!

Remark

The weak law of large numbers states that for a specified large n, the average $\bar{X}_n$ is likely to be near μ (see page 920 for the detailed proof). Thus, it leaves open the possibility that $\left|\bar{X}_n-\mu\right|>\varepsilon$ happens an infinite number of times, although at infrequent intervals.

The strong law of large numbers shows that this almost surely will not occur. Note that it does not imply that with probability 1, we have that for any $\varepsilon>0$ the inequality $\left|\bar{X}_n-\mu\right|<\varepsilon$ holds for all large enough n, since the convergence is not necessarily uniform on the set where it holds.

The probability of an event is somehow responding to the notion of frequency of a random phenomena, in other words, at each event we will attach a real number in the interval $[0,1]$, which measure the probability (chance) of realization. The properties of frequencies we can highlight during various trials allow us to determine the properties of probabilities.

Let U be a universe.

Definition (#5.6.105): We say that we define a "probability" on the events of U if to any event A of U we associate a number or measure $P(A)$, named "a priori probability of event A" or "marginal probability of A" and also shortly named "credence of A" (strength of conviction, expressed as a percentage, of a rational agent, about a hypothesis as defined in [284], it is therefore not a belief as the latter is associated to irrational agents that are unable to quantify their hypothesis).

---

[28] The detailed proof of the strong law of large numbers is actually, due to its difficulty, out of the scope of this textbook. That's also the reason why the huge majority of engineers and postgraduate physicists, astrophysicts, cosmologists, biologists, chemists, etc. don't study it during their curriculum.

Here are the "Kolmogorov's axioms":

A1. For any event A:

$$\boxed{1 \geq P(A) \geq 0} \tag{5.6.45}$$

Thus, the probability of any event is a real number between 0 and 1 inclusive (this is common human sense...).

A2. The probability of the certain event or of the set (sum) of possible events is equal to 1:

$$\boxed{P(U) = 1} \tag{5.6.46}$$

A3. If $A \cap B = \varnothing$, i.e. the two events are incompatible (disjoint), then:

$$\boxed{P(A \cup B) = P(A) + P(B)} \tag{5.6.47}$$

the probability of the merge ("or") of two mutually incompatibles events (or mutually exclusive) is therefore equal to the sum of their probabilities (law of addition). We then speak of "disjoint probability".

We understand better that the third axiom requires $A \cap B = \varnothing$ otherwise the sum of all probabilities could be greater than 1 (imagine again the set diagram of the two events in your head!).

Remarks

R1. Probabilities are often communicated as percentages. So be very careful about typical biases of mass media (and not only!) when dealing with them. We have already warn you, with example, about this trap on page 234.

R2. You must also be very careful with random match probabilities! The most well known case is named the "prosecutor's fallacy". The underlying trap can be understand by considering a crime-scene DNA sample compared against a database of $20,000$ men. A match is found, an individual is then accused and at his trial, it is testified that the probability that two DNA profiles match by chance is only 1 in $10,000$ (prosecutor's argument!). This does not mean the probability that the suspect is innocent is 1 in $10,000$ (what jurors may wrongly understand...)! Since $20,000$ individuals were tested, there were $20,000$ opportunities to find a match by chance (to see the detailed calculation on how to get the probability of getting at least one match based on frequentist approach, see page 659 and page 496 for the bayesian approach).

Let A be an event, P be the probability measure. A has zero probability if $P(A) = 0$. A is impossible if $A = \varnothing$.

Impossibility implies zero probability, but the reverse is false! Consider the real line $\mathbb{R}$; if you randomly select a number x, the probability that $x = 0$ is 0, but this is not impossible. In fact, the probability that x belongs to some countable set, e.g $\mathbb{Q}$, is also 0.

What we mean is that $P(A) = 0$ does not imply $A = \varnothing$, i.e. knowing probability measure $= 0$ does not help you figure out if a set is empty or not!

✍**Example:**

Suppose the typewriter has 50 keys, and the word to be typed is *banana*. If the keys are pressed randomly and independently, it means that each key has an equal chance of being pressed. Then, the chance that the first letter typed is *b* is $1/50$, and the chance that the second letter typed is *a* is also $1/50$, and so on. Therefore, the chance of the first six letters spelling *banana* is:

$$\frac{1}{50}\cdot\frac{1}{50}\cdot\frac{1}{50}\cdot\frac{1}{50}\cdot\frac{1}{50}\cdot\frac{1}{50}=\left(\frac{1}{50}\right)^6=\frac{1}{15625000000} \tag{5.6.48}$$

less than one in 15 billion, but not zero, hence a possible outcome.

From the above, the chance of not typing banana in a given block of 6 letters is $1-(1/50)^6$. Because each block is typed independently, the chance X_n of not typing banana in any of the first n blocks of 6 letters is:

$$X_n=\left(1-\frac{1}{50^6}\right)^n \tag{5.6.49}$$

As n grows, X_n gets smaller. For an n of a million, X_n is roughly 0.9999, but for an n of 10 billion X_n is roughly 0.53 and for an n of 100 billion it is roughly 0.0017. As n approaches infinity, the probability X_n approaches zero; that is, by making n large enough, X_n can be made as small as is desired, and the chance of typing banana approaches 100%.

We will find an example of this kind of disjoint probability in the chapter of Industrial Engineering when studying F.M.E.A. (Failure Modes and Effects Analysis) for fault analysis systems with a complex structure.

In other words in a more general form if $(A_i)_{i\in\mathbb{N}}$ is a sequence of pairwise disjoint events (A_i and A_j can not occur at the same time though $i\neq j$) then:

$$\boxed{P\left(\bigcup_{i=1}^{n}A_i\right)=\sum_{i=1}^{n}P(A_i)} \tag{5.6.50}$$

We then speak of "σ-additivity" because if we look more closely at the three axioms above the measure P forms a σ-algebra (see section Measure Theory page 2483).

At the opposite, if the events are not incompatibles (they can overlap or in other words: they have a joint probability), we then have for probability that at most one of the two takes place:

$$P(A\cup B)=P(A)+P(B)-P(A\cap B) \tag{5.6.51}$$

This means that the probability that at most one of the events A or B occurs is equal to the sum of the probabilities for the realization of A or B occurred, minus the probability that A and B

occurred simultaneously (we will show later that this is simply equal to the probability that the two do not occur at the same time!).

✍**Examples:**

E1. Consider that in a given area, over 50 years, the probability of a major earthquake is 5% and on the same period the probability a major flood is 10%... We would like to know what is the probability that a nuclear plant meets at most one of two events during the same period if the are not incompatibles. We then calculate the probability from the above relation that gives 14.5%...

E2. If we throw a handful of two dices, the probability of getting a double is:

$$P=\frac{6}{36}=\frac{1}{6} \tag{5.6.52}$$

Why ? Because all the possible outcomes at launch give $6^2=36$ issues and the number of favourable cases of double 1, double 2, ... double 6 is 6 and so $P=\frac{6}{6^2}=\frac{6}{36}=\frac{1}{6}$.

And thus if they were incompatibles we would have then $A\cap B=\varnothing$ we find again the disjoint probability:

$$P(A\cup B)=P(A)+P(B) \tag{5.6.53}$$

If we generalize to $n=3$ events we get:

$$\begin{aligned}P(A\cup B\cup C)=&P(A)+P(B)+P(C)\\&-P(A\cap B)-P(A\cap C)-P(B\cap C)\\&+P(A\cap B\cap C)\end{aligned} \tag{5.6.54}$$

The general relation for n events, named the "inclusion-exclusion formula" is:

$$\boxed{\begin{aligned}P\left(\cup_{i=1}^n A_i\right)=&\sum_{i=1}^n P\left(A_i\right)-\sum_{i<j}P\left(A_i\cap A_j\right)+\sum_{i<j<k}P\left(A_i\cap A_j\cap A_k\right)\\&\ldots+(-1)^{n+1}P\left(A_1\cap\ldots\cap A_n\right)\end{aligned}} \tag{5.6.55}$$

In words, we take all possible intersections of one, two, ... n sets and the signs of the sums alternate.

Proof 5.37.2. A point that is in k sets is counted k times by the first sum, C_2^k by the second, C_3^k by the third and so on until it is counted 1 time by the kth term.The net result is:

$$C_1^k-C_2^k+C_3^k+\ldots+(-1)^{k+1}1 \tag{5.6.56}$$

To show that this adds up to 1, we recall the Binomial theorem (see section Calculus page 1364):

$$(a+b)^k=a^k+C_1^k a^{k-1}b+C_2^k a^{k-2}b^2+\ldots+b^k \tag{5.6.57}$$

Setting $a=1$ and $b=-1$ we have:

$$0=1-C_2^k+C_3^k\ldots-(-1)^{k+1}1 \tag{5.6.58}$$

which proves the desired result.

■ Q.E.D.

Remark

This lead us also naturally to "Boole's inequality", that the probability of one or more of events occurring is no greater than the sum of the probabilities of the individual events:

$$P\left(\bigcup_i A_i\right) \leq \sum_i P(A_i) \tag{5.6.59}$$

An immediate consequence of the axioms (A2) and (A3) is the relations between the probability of an event A and its complement, noted $\bar{A}$ (or more rarely in accordance with the notation used in the chapter of Proof Theory the complementary may be noted $\neg A$):

$$P(\bar{A}) = 1 - P(A) \tag{5.6.60}$$

Let U be a universe with a finite number of n possible outcomes:

$$U = \{i_1, i_2, \ldots, i_n\} \tag{5.6.61}$$

where the events:

$$I_1 = \{i_1\}, I_2 = \{i_2\}, I_3 = \{i_3\}, \ldots, I_n = \{i_n\} \tag{5.6.62}$$

are named "elementary events". When these events have the same probability, we say they are "equiprobables". In this case, it is very easy to calculate the probability. Indeed, these events being by definition incompatible with each other at this level of our discussion, we have under the third axiom (A3) of probabilities:

$$P(I_1 \cup I_2 \cup \ldots \cup I_n) = P(I_1) + P(I_2) + \ldots + P(I_n) \tag{5.6.63}$$

but since:

$$P(I_1 \cup I_2 \cup \ldots \cup I_n) = P(U) = 1 \tag{5.6.64}$$

and that the probability of the right hand are by hypothesis equiprobable, we have:

$$P(I_1) = P(I_2) = \ldots = P(I_n) = \frac{1}{n} \tag{5.6.65}$$

Definition (#5.6.106): If A and B are not mutually exclusive but independent, we know that by their (in)compatibility $A \cap B = \varnothing$, that (very important in statistics!):

$$P(A \cap B) = P(A) \cdot P(B) \tag{5.6.66}$$

the probability of the intersection ("and" operator) of two independent events is equal to the product of their probabilities (law of multiplication). We name it "joint probability" (this is the most common case).

✍**Example:**

Consider that in a given area, over 50 years, the probability of a major earthquake is 5% and on the same period the probability a major flood is 10%... Assume that these two events are not mutually exclusive. In other words that they are compatible. We will be interested to their independence. Thus, we would like to know what is the probability that a nuclear power plants meets the two events at the same time, at any time, during this same period. We then calculate the probability from the above equation that gives 0.05%...

Under a more general form, the events $A_1, A_2, \ldots, A_n$ are independent if the probability of the intersection is the product of the probabilities:

$$\boxed{P\left(\bigcap_{i=1}^{n} A_i\right)=\prod_{i=1}^{n} P(A_i)} \tag{5.6.67}$$

Remark

Be careful to not confuse "independent" and "incompatible" (i.e. mutually exclusive)!

So far to summarize a bit we have:

Type	Expression
2 incompatibles events (disjoints)	$P(A\cup B)=P(A)+P(B)$
2 not incompatibles events (joints) non-mutually exclusive events	$P(A\cup B)=P(A)+P(B)-P(A\cap B)$
2 not incompatibles but independents events	$P(A\cap B)=P(A)\cdot P(B)$

Table 5.26: **Classical cases of probabilities**

Thanks the above definition, we can show that the probability that either A or B is to take place (e.g. at least one of the two but not both at the same time), is simply equal to... the probability that the two do not does not occur at the same time:

$$\begin{aligned}P(A\cup B)&=P(A)+P(B)-P\left(A\cap B)\right)\\&=P(A)+P(B)-P(A)P(B)=1-P(\bar{A})P(\bar{B})\\&=1-P(1-A)P(1-B)\end{aligned} \tag{5.6.68}$$

We can also use this definition to determine the probability that only one of two events occurs:

$$\begin{aligned}P(A\oplus B)&=P(A)P(\bar{B})+P(B)P(\bar{A})\\&=P(A)(1-P(B))+P(B)(1-P(A))=P(A)+P(B)-2P(A)P(B)\\&=P(A)+P(B)-2P(A\cap B)\end{aligned} \tag{5.6.69}$$

✍**Example:**

Consider that in a given area, over 50 years, the probability of a major earthquake is 5% and on the same period the probability a major flood is 10%.... We would like to know what is the probability that a nuclear power plant exactly meets one of the both events during the same period, assuming they can not occur at the same time. We then calculate the probability from the above equation and that gives 14%...

There is a common (among others) and important area in the industry where the four following relations are frequently used:

$$\begin{aligned}\text{AND} &= P(A)\cdot P(B)\\ \text{OR COMPATIBLE} &= P(A)+P(B)-P(A\cap B)\\ \text{OR INCOMPATIBLE} &= P(A)+P(B)\\ \text{XOR} &= P(A)+P(B)-2P(A\cap B)\end{aligned} \tag{5.6.70}$$

Remark

Data scientists also use the some of the results above especially when working with proportions in surveys or classification problems (see section Numerical Methods page 6818).

This is the "tree analysis error" or "probabilistic tree analysis" which is used to analyze the possible reasons for failure of a system of any kind (industrial, administrative or other).

To close this part of the chapter consider the following figure displaying Venn diagrams (see section Set Theory page 378) for all 16 events (including the impossible event) that can be described in terms of two given events A and B. In each case, the event is represented by the red area:

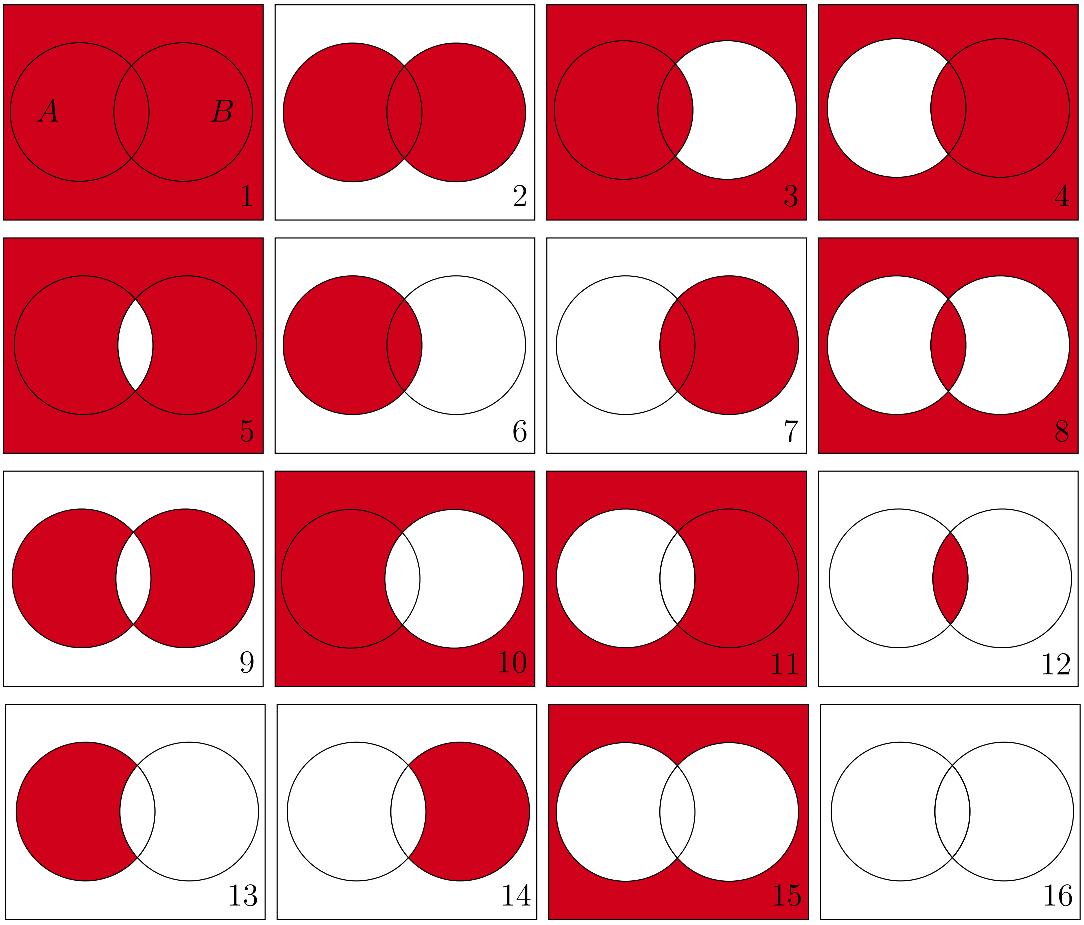


Figure 5.83: Possible Venn diagrams for two events

Consider the situation where A represents and earthquake and B represents a major flood and U the universe of all dramatics events for a nuclear power plant. We consider that the two events are independents. Then each of the 16 events can be described as follows, either mathematically or verbally:

1. An earthquake can occur or a flood or nothing or the both together or any other event (to summarize: any event can occur).
$$P(U)=1=100\% \tag{5.6.71}$$
2. $A\cup B$: Any event with an earthquake a flood or the both event together can occur.
$$P(A\cup B)=P(A)+P(B)-P(A\cap B)=P(A)+P(B)-P(A)P(B) \tag{5.6.72}$$
3. $A\cup B^c$: Any event with earthquake can occur with or without a flood excepted events with a flood not associated to an earthquake.
$$P(A\cup B^c)=P(U)-P(B)+P(A\cap B)=1-P(B)+P(A)P(B) \tag{5.6.73}$$
4. $A^c\cup B$: Any event with earthquake can occur with or without a flood excepted events with a flood not associated to an earthquake.
$$P(A^c\cup B)=P(U)-P(A)+P(A\cap B)=1-P(A)+P(A)P(B) \tag{5.6.74}$$

5. $A^c \cup B^c$: Any event can occur excepted those associated with an earthquake together with a flood.

$$P(A^c \cup B^c) = P(U) - P(A \cap B) = 1 - P(A)P(B) \tag{5.6.75}$$

6. A: Any event with an earthquake can occur (this include the events associating an earthquake and a flood).

$$P(A) = P(A) \tag{5.6.76}$$

7. B: Any event with a flood can occur (this include the events associating a flood and an earthquake).

$$P(B) = P(B) \tag{5.6.77}$$

8. $(A \cap B) \cup (A^c \cap B^c)$: Any event can occur excepted those including an earthquake without a flood or those including a flood without an earthquake.

$$\begin{aligned} P((A \cap B) \cup (A^c \cap B^c)) &= P(U) - P(A) - P(B) + 2P(A \cap B) \\ &= 1 - P(A) - P(B) + 2P(A)P(B) \end{aligned} \tag{5.6.78}$$

9. $(A \cap B^c) \cup (A^c \cup B)$: Any event including an earthquake without a flood or a flood without an earthquake can occur.

$$P((A \cap B^c) \cup (A^c \cup B)) = P(A) + P(B) + 2P(A \cap B) \tag{5.6.79}$$

10. B^c: Any event excepted those including a flood can occur.

$$P(B^c) = P(U) - P(B) = 1 - P(B) \tag{5.6.80}$$

11. A^c: Any event excepted those including an earthquake can occur.

$$P(A^c) = P(U) - P(A) = 1 - P(A) \tag{5.6.81}$$

12. $A \cap B$: Any events associating an earthquake and a flood together can occur.

$$P(A \cap B) = P(A)P(B) \tag{5.6.82}$$

13. $A \cap B^c$: Any event with an earthquake and without a flood can occur.

$$P(A \cap B^c) = P(A) - P(A)P(B) \tag{5.6.83}$$

14. $A^c \cap B$: Any event with a flood and without an earthquake can occur.

$$P(A^c \cap B) = P(B) - P(A)P(B) \tag{5.6.84}$$

15. $A^c \cap B^c$: Any event can occur excepted those including an earthquake and/or a flood.

$$P(A^c \cap B^c) = P(U) - P(A \cup B) = 1 - P(A) - P(B) - P(A)P(B) \tag{5.6.85}$$

16. $A \cap A^c$ or $B \cap B^c$: Impossible Event.

$$P(A \cap A^c) = P(B \cap B^c) = P(\varnothing) = 0 \tag{5.6.86}$$

5.6.3 Conditional Probabilities

What can we infer about the probability of an event B knowing that an event A has occurred, aware that there is a link between A and B? In other words, if there is a link between A and B, the completion of A has to change our understanding of B and we want to know if it is possible to define the conditional probability of an event (relatively) to another event.

This type of probability is named "conditional probability" or "a posterior probability" of B knowing A, and is denoted in the context of the study of conditional probabilities:

$$P(B/A) \tag{5.6.87}$$

and often in practice to avoid confusion with a possible division:

$$P(B \mid A) \tag{5.6.88}$$

and we sometimes find in U.S. books the notation:

$$P(B \wedge A) \tag{5.6.89}$$

or also:

$$P_B(A) \tag{5.6.90}$$

We also have the case:

$$P(A/B) \tag{5.6.91}$$

which is named "likelihood function of A" or "a priori probability of A given B".

Historically, the first mathematician to have used the correct notion of conditional probability was Thomas Bayes (1702-1761). This is why we often say "Bayes" or "Bayesian" probabilities as soon as conditional probabilities are involved: "Bayes formula", "Bayesian statistics", etc.

The notion of conditional probability that we will introduce is much less simple than it first appears and the conditionals problems are an inexhaustible source of errors of any kind (there are famous paradoxes on the subject and even expert requires peer review to minimize mistakes).

Let's start with a simple example: Suppose we have two dice. Now imagine that we only launched the first die. We want to know what is the probability that by throwing the second dice, the sum of the two numbers is equal to a given minimum value. Thus, the probability of obtaining the minimum value given the value of the first die is totally different from the probability of obtaining the same minimum value in throwing two dice at the same time. How to calculate this new probability?

Let us now formalize the process! After the launch of the first dice, we have:

$$A = \{\text{the result of the first throw is...}\} \tag{5.6.92}$$

Under the hypothesis that $B \subset A$, we feel that $P(B/A)$ must be proportional to $P(B)$, the proportionality constant being determined by the normalization:

$$P(A/A) = 1 \tag{5.6.93}$$

Now let $B\subset A^c$ (B is included in the complement of A so that the events are mutually exclusive). It is then relatively intuitive .... that under the previous hypothesis of incompatibility we have the conditional probability:

$$P(B/A)=0 \tag{5.6.94}$$

This leads us to the following definitions of respectively a posteriori and a priori probabilities:

$$\boxed{P(B/A)=\dfrac{P(A\cap B)}{P(A)}}\quad \text{and}\quad \boxed{P(A/B)=\dfrac{P(A\cap B)}{P(B)}} \tag{5.6.95}$$

Thus, the fact to know that A has occurred reduces all possible outcomes of the universe U of B. From there, only the events of type $A\cap B$ are important. The probability of A given B or vice versa (by symmetry) must be proportional to $P(A\cap B)$!

The coefficient of proportionality is the denominator and it ensures the certain event. Indeed, if two events A and B are independent (think the black cat and the scale for example), then we have:

$$P(A\cap B)=P(A)P(B) \tag{5.6.96}$$

and then we see that $P(B/A)$ is equal to $P(B)$ and therefore the event A adds no information to B and vice versa! So in other words, if A and B are independent, we have:

$$P(B/A)=P(B)\quad \text{and}\quad P(A/B)=P(A) \tag{5.6.97}$$

Another fairly intuitive way to see things is to represent the probability measure P as a measure of subsets areas (surface) of $\mathbb{R}^2$.

Indeed, if A and B are two subsets of respective areas $P(A)$ and $P(B)$ then the question of what is the probability that a point in the plane belongs to B knowing that it belongs to A it is quite obvious that the probability is given by answer:

$$\frac{\text{Surface}(A)\cap\text{Surface}(B)}{\text{Surface}(A)} \tag{5.6.98}$$

We would like to indicate that the definition of conditional probability is often used in the following way:

$$P(A\cap B)=P(A/B)P(B)=P(B/A)P(A) \tag{5.6.99}$$

named "formula of compound probabilities" or "product rule". It is written in many different ways (just a matter of notation convention):

$$P(A\cap B)=P(A/B)P(B)\qquad \text{or}\qquad P(A,B)=P(A|B)P(B) \tag{5.6.100}$$

Thus, the a posteriori probability of B knowing A can also be written as:

$$\boxed{P(B/A)=\frac{P(A/B)P(B)}{P(A)}} \tag{5.6.101}$$

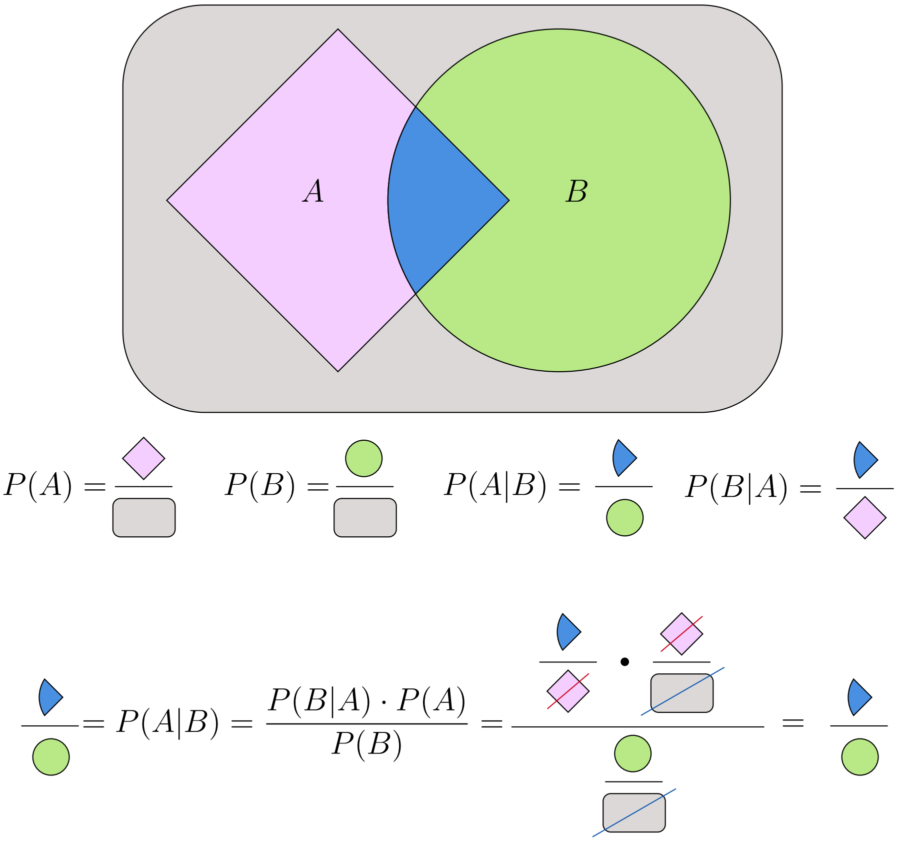


Figure 5.84: Bayes' theorem illustrated

Or in human words:

$$\text{Posterior probability} = \frac{\text{Likelihood} \cdot \text{prior probability}}{\text{expectedness}} \tag{5.6.102}$$

but also sometimes:

$$\text{Plausability} = \frac{\text{Likelihood} \cdot \text{Credence}}{\text{Partition function}} \tag{5.6.103}$$

The way that this formula gives an update of the probability hypothesis, B, in light of some body of data, A, is named the "diachronic interpretation". "Diachronic" meaning that something is happening over time, in this case the probability of the hypothesis changes, over time, as we see new data.

In this interpretation the different terms have a name:

- $P(B)$ is the probability of the hypothesis before we see the data, name as we already know, the "prior probability[29]", or just the "prior".
- $P(B/A)$ is what we want to compute, the probability of the hypothesis after we see the data, named as we already know, the "posterior".

---

[29] Where does the prior come from is the million dollar question... In principle the bayesian is supposed to choose a prior that represents their prior information. This will be challenging in high dimensional cases to say the least. Also, critics will say that someone's prior opinions should not be included in a data analysis because this is not scientific.

- $P(A/B)$ is the probability of the data under the hypothesis, named the "likelihood".
- $P(A)$ is the probability of the data under the hypothesis, named the "normalizing constant" or "expectedness".

✍Examples:

E1. Let us consider a disease like meningitis. The priori probability of having the meningitis will be denoted by $P(M) = 0.001$ (arbitrary value for the example) and the expectedness of having headache will be denoted $P(S) = 0.1$. We assume known that the likelihood of having a headache if we have meningitis is:

$$P(S/M) = 0.9 \tag{5.6.104}$$

So we will use the Bayes' theorem:

$$\underbrace{P(A/B)}_{\text{posterior probability}} = \frac{\overbrace{P(B/A)}^{\text{Likelihood}} \cdot \overbrace{P(A)}^{\text{prior probability}}}{\underbrace{P(B)}_{\text{expectedness}}} \tag{5.6.105}$$

The Bayes' theorem then gives the a posteriori probability of having meningitis if we have a headache:

$$P(M/S) = \frac{P(S/M)P(M)}{P(S)} = 0.09 \tag{5.6.106}$$

E2. In a quiz-show (this example is named the Monty Hall paradox), the candidate has to select one of three doors. Behind two doors there is a goat (an undesirable prize). Behind the one door there is a car. At the beginning the candidate tries to guess the door with the car. Afterwards the moderator Monty Hall opens one of the doors with a goat which is not the candidate's door. So two doors are left where one of them is the one, the candidate has chosen. Monty asks the candidate whether he would like to change his mind; he is free to select the other door.

In general, is it better to change the mind, stick to the first selection or doesn't it matter?

Assumption (one of three cases): The candidate selects door #1.

D_1	Car behind door #1
D_2	Car behind door #2
D_3	Car behind door #3

We define the corresponding probabilities:

$$P(D_1) = \frac{1}{3} \qquad P(D_2) = \frac{1}{3} \qquad P(D_3) = \frac{1}{3} \tag{5.6.107}$$

We somehow have to model the behaviour of the moderator. Therefore we introduce a hypothesis H that he opens door #3. The probability that he selects door #3 in favour of door #2 is $\frac{1}{2}$, because #1 is already taken by the candidate:

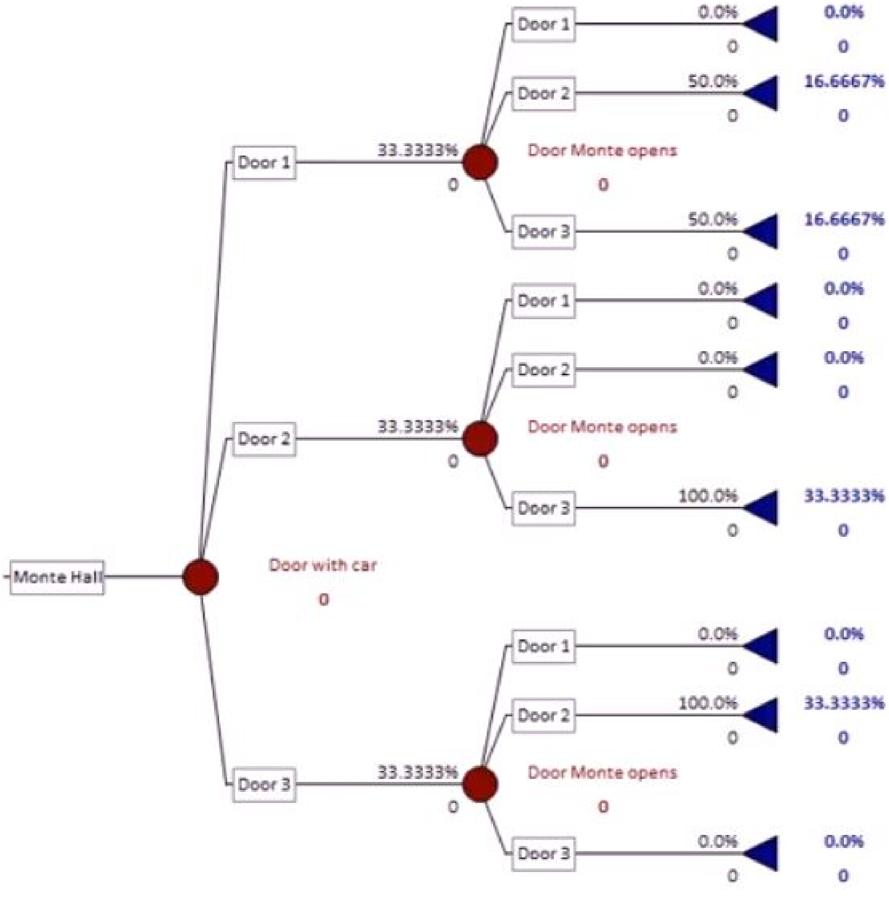
$$P(H) = \frac{1}{2} \tag{5.6.108}$$

Furthermore:

$$P(H/D_1)=\frac{1}{2} \qquad P(H/D_2)=1 \qquad P(H/D_3)=0 \tag{5.6.109}$$

So the probability that he opens door #3 under the condition that door #1 contains the car is $\frac{1}{2}$. The probability is 1 if the car is in door #2. And if the car is in door #3, the moderator will never show door #3.

The above scenario can be build easily with a software like @Risk of Palisade (the PrecisionTree module):


Monte Hall
Door with car
0
Door 1
33.3333%
0
Door Monte opens
0
Door 1
0.0%
0
0.0%
0
Door 2
50.0%
0
16.6667%
0
Door 3
50.0%
0
16.6667%
0
Door 2
33.3333%
0
Door Monte opens
0
Door 1
0.0%
0
0.0%
0
Door 2
0.0%
0
0.0%
0
Door 3
100.0%
0
33.3333%
0
Door 3
33.3333%
0
Door Monte opens
0
Door 1
0.0%
0
0.0%
0
Door 2
100.0%
0
33.3333%
0
Door 3
0.0%
0
0.0%
0

In the above figure we assume that the contester has already chooses door #1. If the car is behind door #1 Monty well flip a coin to open door #2 or #3 (hence the 50% and 50%). However if the car is behind door #2 Monty will surely open door #3 (hence the 100%). And finally if the car is behind door #3 Monty will surely open door #2 (hence the second 100%).

We can spot some asymmetry here. This point is evidently responsible why an intuition like *it does not matter* might be wrong.

So we will use the Bayes' Theorem:

$$\underbrace{P(A/B)}_{\text{posterior probability}} = \frac{\overbrace{P(B/A)}^{\text{Likelihood}} \cdot \overbrace{P(A)}^{\text{prior probability}}}{\underbrace{P(B)}_{\text{expectedness}}} \tag{5.6.110}$$

The important question is: Under the assumption that we have selected door #1 (our first assumption) and Monty opens door #3 (our hypothesis or expectedness), what is the probability that the car is behind door #1 compared to door #2?

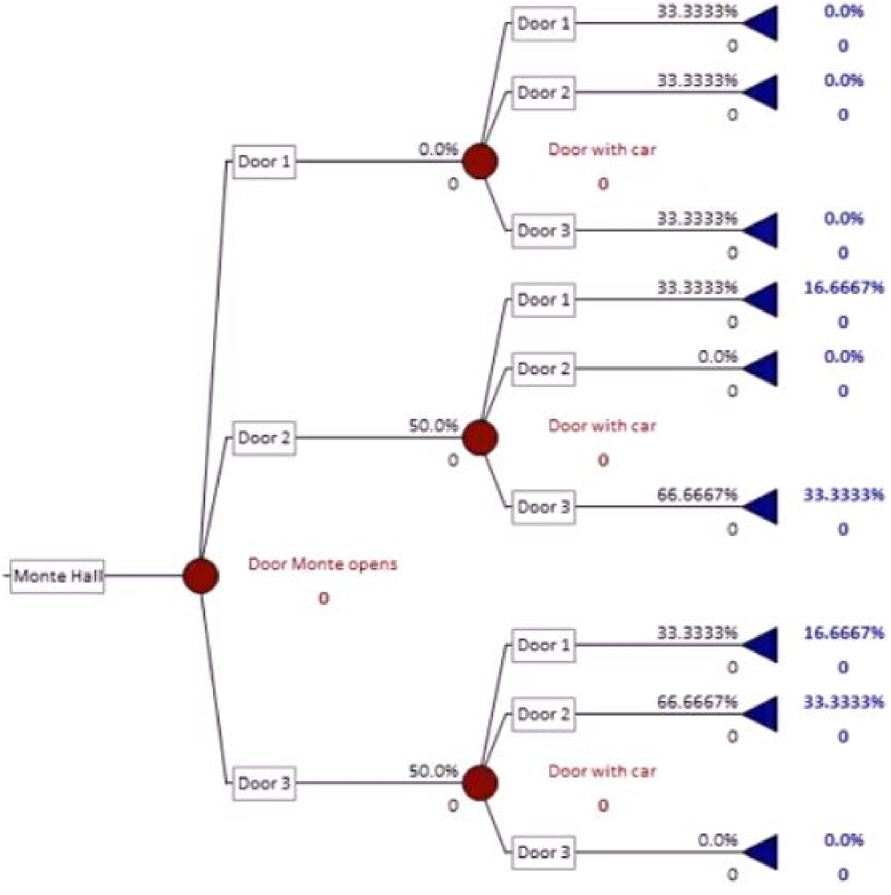
$$P(D_1|H) = \frac{P(H|D_1) \cdot P(D_1)}{P(H)} = \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2}} = \frac{1}{3} \tag{5.6.111}$$

$$P(D_2|H) = \frac{1 \cdot \frac{1}{3}}{\frac{1}{2}} = \frac{2}{3} \tag{5.6.112}$$

So apparently, the probability that door #2 contains the car is higher than for door #1. Or in other words: in this case the candidate should prefer to change his mind.

Now we have to evaluate this for other hypotheses (moderator opens door #2) and other candidate selections (door #2 and door #3). The approach is the same and not carried out here. But the conclusion keeps the same in all cases: Being open to change your mind will make you probabilistically win the car.

So here Bayes rule is necessary because the contester needs to flip the tree above. To get the nodes in the opposite order. The reason is that the contester will first see the door that Monty opens. At that point the contester need to know the posterior probability. In @Risk Palisade has a Bayes revision option that flips the tree and gives:


Monte Hall
Door Monte opens
0
Door 1
0.0%
0
Door 2
50.0%
0
Door 3
50.0%
0
Door with car
0
Door 1
33.3333%
0
0.0%
0
Door 2
33.3333%
0
0.0%
0
Door 3
33.3333%
0
0.0%
0
Door 1
33.3333%
0
16.6667%
0
Door 2
0.0%
0
0.0%
0
Door 3
66.6667%
0
33.3333%
0
Door 1
33.3333%
0
16.6667%
0
Door 2
66.6667%
0
33.3333%
0
Door 3
0.0%
0
0.0%
0

Now Monty action is on the left and the location is on the right. Can you see that the contester should switch no matter what Monty does? Monty won't certainly open door 1 because the contester has already chosen it. You can see on the left that the chance are 50/50 that Monty opens door #2 or door #3. If he opens door #2 the posterior probability that the car is behind door #3 is $2/3$ that is twice the probability which is twice the posterior probability that it is behind door #1. In the same way if he opens door #3, the posterior probability that the car is behind door #2 is $2/3$ that is twice the probability which is twice the posterior probability that it is behind door #1.

Some people think that this analysis is all smoke and mirrors. In fact some people continue to believe that Bayes rule is quite controversial. However everything here is based on mathematical ground and there is nothing controversial about it. The key about understanding the apparent paradox, that the contester can do better by switching, is that Monty has supplied some useful information by showing a door where the car is not located. This information together with Bayes rule allow the contester to improve his chances of winning the car by switching doors.

We also notice that:

$$\begin{aligned}P(A)&=P\left((B_1\cup B_2\cup\ldots\cup B_i)\cap A\right)\\&=P\left((B_1\cap A)\cup(B_2\cap A)\cup\ldots\cup(B_i\cap A)\right)\\&=P\left((B_1\cap A)+(B_2\cap A)+\ldots+(B_i\cap A)\right)\\&=P(A/B_1)P(B_1)+P(A/B_2)P(B_2)+\ldots+P(A/B_n)P(B_n)\\&=\sum_{i=1}^{n}P(A/B_i)P(B_i)\end{aligned}\tag{5.6.113}$$

So we can know the probability of the event A knowing the elementary probabilities $P(B_i)$ of its causes and the conditional probabilities of A for each B_i:

$$\boxed{P(A)=\sum_{i=1}^{n}P(A/B_i)P(B_i)}\tag{5.6.114}$$

which is named the "formula of total probabilities" or "total probabilities theorem" (relation that is also very useful when we don't know the value of $P(A)$ or can't estimate it directly easily). But also, for any j, we have the following corollary using the previous results that gives us following an event A, the probability that it is the cause B_i that produced it:

$$\boxed{P(B_j/A)=\frac{P(B_j\cap A)}{P(A)}=\frac{P(A/B_j)P(B_j)}{\displaystyle\sum_i P(A/B_i)P(B_i)}}\tag{5.6.115}$$

which is the general form of the "Bayes formula" or "Bayes' theorem" that we will use a little in the Statistical Mechanics chapter and through the study of the queueing theory (see section Quantitative Management page 8141). You should know that the implications of this theorem are, however, considerable in daily life, in medicine, in industry and in the field of Data Mining.

In fact, every valid form of hypothetico-deductive method (HDM) is described by Bayes' theorem. The hypothetico-deductive method consists of formulating a hypothesis H, deducing the observations that would be made if H were true and those that would be made if H were false, and making numerous observations until the probability of H far exceeds all known alternatives, or becomes implausible.

We often find in the literature many examples of applications of the previous relation with only two possible outcomes B with respect to the event A. Therefore we find the Bayes formula written in the following form for each issue:

$$\begin{aligned}P(B_1/A)&=\frac{P(A/B_1)P(B_1)}{P(A/B_1)P(B_1)+P(A/B_2)P(B_2)}=\frac{P(A/B_1)P(B_1)}{P(A/B_1)P(B_1)+P(A/\bar{B}_1)P(\bar{B}_1)}\\P(B_2/A)&=\frac{P(A/B_2)P(B_2)}{P(A/B_1)P(B_1)+P(A/B_2)P(B_2)}=\frac{P(A/B_2)P(B_2)}{P(A/\bar{B}_2)P(\bar{B}_2)+P(A/B_2)P(B_2)}\end{aligned}\tag{5.6.116}$$

and note that in this particular case (binary outcomes):

$$\begin{aligned}
P(B_1/A)+P(B_2/A) &= P(B_1/A)+P(\bar{B}_1/A)\\
&= \frac{P(A/B_1)P(B_1)}{P(A/B_1)P(B_1)+P(A/\bar{B})P(\bar{B})}+\frac{P(A/B_2)P(B_2)}{P(A/\bar{B}_2)P(\bar{B}_2)+P(A/B_2)P(B_2)}\\
&= \frac{P(A/B_1)P(B_1)}{P(A/B_1)P(B_1)+P(A/\bar{B})P(\bar{B})}+\frac{P(A/B_2)P(B_2)}{P(A/B_1)P(B_1)+P(A/\bar{B}_1)P(\bar{B}_1)}\\
&= \frac{P(A/B_1)P(B_1)+P(A/B_2)P(B_2)}{P(A/B_1)P(B_1)+P(A/\bar{B}_1)P(\bar{B}_1)}\\
&= \frac{P(A/B_1)P(B_1)+P(A/\bar{B}_1)P(\bar{B}_1)}{P(A/B_1)P(B_1)+P(A/\bar{B}_1)P(\bar{B}_1)}=1
\end{aligned} \tag{5.6.117}$$

is an intuitive result.

For binary events, we also have (returning to the theorem of total probabilities seen above):

$$\begin{aligned}
P(A) &= \sum_{i=1}^{n} P(A/B_i)P(B_i)\\
&= P(A/B_1)P(B_1)+P(A/B_2)P(B_2)\\
&= P(A/B_1)P(B_1)+P(A/\bar{B}_1)P(\bar{B}_1)
\end{aligned} \tag{5.6.118}$$

Or in the case of odds ratio:

$$\frac{P(A/B)}{P(\bar{A}/B)}=\frac{\dfrac{P(B/A)P(A)}{P(B)}}{\dfrac{P(B/\bar{A})P(\bar{A})}{P(B)}}=\frac{P(B/A)P(A)}{P(B/\bar{A})}\frac{P(A)}{P(\bar{A})} \tag{5.6.119}$$

✍**Examples:**

E1. A disease affects 10 people on $10,000$ ($0.1\% = 0.001$). A test has been developed which has a 5% false positives (people not having the disease but for which the test says they are affected) but still always detects the disease if a person has it. What is the probability that a random person for which the test gives a positive result really has this disease?

There is therefore $10,000$ people, 500 of which are false positives, and we know a posteriori that 10 people have really the disease. Then the probability that somebody who has a positive test result is really sick is:

$$\begin{aligned}P(M)&=0.001\\P(\bar{M})&=0.999\\P(T/M)&=1\\P(T/\bar{M})&=0.05\\P(T)&=P(T/M)P(M)+P(T/\bar{M})P(\bar{M})=0.05095\\P(M/T)&=\frac{P(T/M)P(M)}{P(T)}=\frac{1\cdot 0.001}{0.05095}=0.19627\end{aligned}\tag{5.6.120}$$

This is often a shocking and counter-intuitive result. It also highlights why diagnostic tests must be extremely reliable!

E2. Two machines M_1 and M_2 produce respectively 100 and 200 pieces. M_1 produces 5% defective pieces and M_2 produces 6% (posterior probabilities). What is the a priori probability that a defective piece was manufactured by the machine M_1?

We then have:

$$\begin{aligned}P(M_j/A)&=\frac{P(A/M_j)P(M_j)}{\sum\limits_i P(A/M_i)P(M_i)}=\frac{P(A/M_1)P(M_1)}{P(A/M_1)P(M_1)+P(A/M_2)P(M_2)}\\&=\frac{\dfrac{5}{100}\dfrac{300}{100}}{\dfrac{5}{100}\dfrac{100}{300}+\dfrac{6}{100}\dfrac{200}{300}}=\frac{5\%\dfrac{1}{3}}{5\%\dfrac{1}{3}+6\%\dfrac{2}{3}}\cong 39\%\end{aligned}\tag{5.6.121}$$

E3. From a batch of 10 pieces with 30% defective, we take a sample of size 3 without replacement. What is the probability that the second piece is correct (whatever the first is)?

We have (as the problem is binary, i.e. correct/not correct):

$$\begin{aligned}P(A)&=\sum_{i=1}^{n}P(A/B_i)P(B_i)=\sum_{i=1}^{2}P(A/B_i)P(B_i)\\&=P(A/B_1)P(B_1)+P(A/B_2)P(B_2)=\frac{7}{9}\frac{3}{10}+\frac{6}{9}\frac{7}{10}=70\%\end{aligned}\tag{5.6.122}$$

E4. Consider that two duelists (A and B) have a revolver with a rotatable magazine and they want to play the russian roulette game. The magazine contains 6 slots, and the duelists load it with one shot (S):

1. Duelist A rotates the magazine at random, points the revolver at his/her head and presses the trigger (we consider that the rotation of the magazine is completely fair).
2. If duelist A survives, he/she hands the revolver to duelist B who repeats step 1 above.

3. The duelists continue in this manner until one of them are dead.

What is the probability that duelist A dies as he/her starts the game ?

As it is a binary case so too as the previous example, we can write ($P\left(A/S\right)$ is the conditional probability of A dies given the first slot contains the shot and $P\left(A/\bar{S}\right)$ is interpreted that duelist B begins the game again as duelist A didn't get the kill shot):

$$\begin{aligned} P(A) &= \sum_{i=1}^{n} P(A/S_i)P(S_i) = P\left(A/S\right)P\left(S\right) + P\left(A/\bar{S}\right)P\left(\bar{S}\right) \\ &= 1\frac{1}{6} + P\left(A/\bar{S}\right)\frac{5}{6} \\ &= \frac{1}{6} + P(B)\cdot\frac{5}{6} = \frac{1}{6} + (1-P(A))\frac{5}{6} \end{aligned} \tag{5.6.123}$$

Therefore solving this last line (equation with one unknown):

$$P(A) = \frac{1}{6} + \frac{5}{6} - \frac{5}{6}P(A) \Rightarrow \frac{11}{6}P(A) = 1 \Rightarrow P(A) = \frac{6}{11} = 54.\overline{54}\% \tag{5.6.124}$$

Hence the duelist that starts the game is more likely to die than the first (that makes quite sense as this duelist will always have one shot more in advance).

E5. Consider that corporations assess employees knowledge with exams or assessments in the form of multiple choice questions (MCQs). If an employee responds to a question there are two issues: either he knows the answer or he try to guess it. Let p be the probability that the employee know the answer and therefore $1-p$ that he guess it. We admit that the employee who guesses will correctly answer with a probability of $1/m$ where m is the number of proposed answers. What is the a priori probability that an employee (really) knows the answer to a question with 5 choices if he answered correctly?

Let B and A be respectively the events "the employee knows the answer" and "the employee correctly answers the question". Then the a priori probability that an employee knows (really) the answer to a question that he answered correctly is:

$$P(B/A) = \frac{P(A/B)P(B)}{P(A/B)P(B) + P(A/\bar{B})P(\bar{B})} = \frac{1\cdot p}{1\cdot p + \dfrac{1}{m}(1-p)} = \frac{5}{6} \cong 83\% \tag{5.6.125}$$

E6. Suppose a doctor thinks the odds are $2/1$ that a patient has a certain disease. She orders two independents tests. Let H be the event the patient has the disease and E_1 and E_2 be the event the tests are positive. Suppose the first test has probability 0.1 of a false positive and probability 0.05 of a false negative. The second test has probabilities 0.05 and 0.08 of a false positive and false negative, respectively. If both tests are positive, what is the posterior probability the patient has the disease?

Assuming that E_1 and E_2 are conditionally independent of H and also of $\bar{H}$, we work first in terms of odds, then convert to probability:

$$\frac{P(H/E_1E_2)}{P(\bar{H}/E_1E_2)}=\frac{P(H)}{P(\bar{H})}\frac{P(E_1E_2/H)}{P(E_1E_2/\bar{H})}=\frac{P(H)}{P(\bar{H})}\frac{P(E_1/H)P(E_2/H)}{P(E_1/\bar{H})P(E_2/\bar{H})} \tag{5.6.126}$$

The data are:

$$\frac{P(H)}{P(\bar{H})}=2, P(E_1/H)=0.95, P(E_1/\bar{H})=0.1, P(E_2/H)=0.92, P(E_2/\bar{H})=0.05 \tag{5.6.127}$$

Substituting value, we get:

$$\frac{P(H/E_1E_2)}{P(\bar{H}/E_1E_2)}=2\cdot\frac{0.95\cdot 0.92}{0.10\cdot 0.05}=349.6 \tag{5.6.128}$$

Similarly we get:

$$P(H/E_1E_2)=99.71\% \tag{5.6.129}$$

E7. Let us come back again on prosecutor's fallacy with the DNA case (see earlier above page 477). Let us consider that DNA profiling of a suspect gives rise to a rare trait, X, such that:

$$P(X\text{occurs in general population})=\frac{1}{10000} \tag{5.6.130}$$

which is equivalent to:

$$P(\text{Match}/\text{Innocent})=P(\text{M}/\text{I})=\frac{1}{10000} \tag{5.6.131}$$

This is of no interest! We want (G is for "potential guilty"):

$$P(\text{I}/\text{M})=\frac{P(\text{M}/\text{I})P(\text{I})}{P(\text{M}/\text{I})P(\text{I})+P(\text{M}/\text{G})P(\text{G})}=\frac{\dfrac{1}{10000}\dfrac{19999}{20000}}{\dfrac{1}{10000}\dfrac{19999}{20000}+1\cdot\dfrac{1}{20000}}\cong 66.67\% \tag{5.6.132}$$

Caution! Be careful when reading books containing probabilities calculations especially when the authors are theists and even more when the probabilities involve Bayes' theorem. The believing physicist Stephen Unwin, as presented in his 12003 (holocene calendar) book, calculates the probability of his christian god to be 67%. He begins his calculation by assuming the prior probability of christian god's existence, before any evidence is considered, to be 50%... How does he know that when we know that there are thousands of very different gods existing on Earth? If we compare Unwin's calculation with the unpublished one of the non-believing physicist Larry Ford. Ford used all the same equations as Unwin, differing only in the numbers inserted in the calculation. He argued the prior probability of Unwin's christian god must be compared with other unobserved entities such as Big Foot or the Loch Ness Monster, which he took to be one in a million (10^{-6}). With his set of numbers, Ford's conclusion was that the probability christian god exists is one in a hundred thousand trillion (10^{-17}).

Bayesian analysis provides also a powerful tool to formalize reasoning under uncertainty and the examples we have shown above illustrate how this tool can be difficult to use.

Finally, for a future use in the section of Theoretical Computing, for the Data Mining technique named "binomial naive Bayes classification" we will need a last result!

Let us recall the product rule seen just earlier above:

$$P(A\cap B)=P(A/B)P(B) \tag{5.6.133}$$

We can extend this for three variables. Indeed, let us consider:

$$P(A\cap D)=P(A/D)P(D) \tag{5.6.134}$$

and let us put $D=B\cap C$ then:

$$P(A\cap (B\cap C))=P(A/(B\cap C))P(B\cap C) \tag{5.6.135}$$

But as the product rule gives us also:

$$P(B\cap C)=P(B/C)P(C) \tag{5.6.136}$$

Then:

$$P(A\cap (B\cap C))=P(A/(B\cap C))P(B/C)P(C) \tag{5.6.137}$$

also sometimes denoted:

$$P(A,B,C)=P(A|B,C)P(B|C)P(C) \tag{5.6.138}$$

and in general for n variables:

$$\boxed{P(A_1,A_2,\ldots,A_n)=P(A_1|A_2,\ldots,A_n)P(A_2|A_3,\ldots,A_n)P(A_{n-1}|A_n)P(A_n)} \tag{5.6.139}$$

In general we refer to this as the "chain rule" or to avoid any confusion with the same term in the field of Differential and Integral Calculus we refer to this as the "Bayes' chain rule".

This formula is especially significant for Bayesian Belief Nets. It provides a means of calculating the full joint probability distribution ; in BBNs many of the variables A_i will be conditionally independent.

Now the reader must keep something in mind: Any probability result that is true for unconditional probability remains true if everything is conditioned on some event!

We know that by definition:

$$P(A/B)=\frac{P(A\cap B)}{P(B)} \tag{5.6.140}$$

and so if we condition everything on C having occurred, we get that:

$$P(A/(B\cap C))=\frac{P((A\cap B)/C)}{P(B/C)} \tag{5.6.141}$$

Hence:

$$P((A\cap B)/C)=P(A/(B\cap C))P(B\cap C) \tag{5.6.142}$$

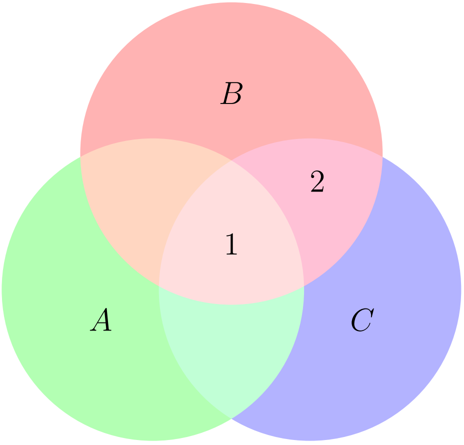
also sometimes (badly) denoted:

$$P(A,B/C)=P(A/B,C)P(B/C) \tag{5.6.143}$$

or:

$$P(A,B|C)=P(A|B,C)P(B|C) \tag{5.6.144}$$

Intuitively to get this result it can help to draw the corresponding Venn diagram (almost as always):



We then have:

$$P(A\cap B/C)=\frac{\text{"1"}}{\text{"C"}},\quad P(B/C)=\frac{\text{"1"}+\text{"2"}}{\text{"C"}},\quad P(A/B\cap C)=\frac{\text{"1"}}{\text{"1"}+\text{"2"}} \tag{5.6.145}$$

and then just do the multiplications to see that everything is fine.

5.6.3.1 Conditional Expectation and Variance

Now, we will see the continuous version of the conditional probability by introducing the subject directly with a particular example (the general theory being indigestible) infinitely important in the field of social statistics and quantitative finance. However, this choice (the study of a particular case) implies that the reader has read the first chapter of Statistics to study the functions of continuous distributions and especially that of the Pareto law.

So here's the scenario: Often, in social sciences or economics, we find in the literature dealing with the laws of Pareto statements like the following (but almost never with a detailed proof): whatever your income, the average income of those who have an income above yours is in a constant ratio, greater than 1, to your income if it follows a Pareto random variable. Then we say that the law is isomorphic to any truncated part itself.

Let's see what it is exactly!

Let X be a random variable equal to the income and following a Pareto with the density (see section Statistics page 746):

$$f(x)=k\frac{x_m^k}{x^{k+1}} \tag{5.6.146}$$

with $k>1,x_m>0,x\geq x_m$ and that has for distribution function (see also the Statistics chapter for the detailed proof):

$$P(X\leq x)=1-\left(\frac{x_m}{x}\right)^k \tag{5.6.147}$$

The sentence begins with «*whatever your income* ...», then select any income $x_0\geq x_m$.

Now we need to compute «*the average income of those with income higher than* x_0». It is therefore asked to calculate the expected (average income) of a new random variable Y that is equal to X, but restricted to the population of people with an income above x_0:

$$Y=X|_{\{X\geq x_0\}} \tag{5.6.148}$$

The distribution function of Y is given by:

$$P(Y\leq x)=P\left(X\leq x \mid X\geq x_0\right) \tag{5.6.149}$$

This expression is of course equal to zero if $x\leq x_0$. Well, so far we have only do vocabulary. First recall the following conditional probability relation already seen before:

$$P(B/A)=\frac{P(A\cap B)}{P(A)} \tag{5.6.150}$$

for $x\geq x_0$ we have for the conditional law:

$$\boxed{P\left(X\leq x \mid X\geq x_0\right)=\frac{P(x_0\leq X\leq x)}{P(X\geq x_0)}} \tag{5.6.151}$$

Before going further, you should be aware that the numerator and denominator are independent but that the whole must be considered, however, as the realization of a single random variable which we denote Y. Furthermore, only the numerator is a dependent variable. The denominator can it be considered as a normalization constant.

So we see that the density of Y is given by the function:

$$f_Y(y)=\begin{cases} 0 & y<x_0 \\ \dfrac{f(y)}{P(X\geq x_0)} & y\geq x_0 \end{cases}$$

Now we can calculate the expected mean (see section Statistics page 634) of Y (assuming $X\geq x_0$):

$$\begin{aligned}\mathrm{E}(Y)=\mathrm{E}(Y|X\geq x_0)&=\int yf_Y(y)\mathrm{d}y=\int\limits_{x_0}^{+\infty}\frac{f(y)}{P(X\geq x_0)}\mathrm{d}y=\frac{1}{P(X\geq x_0)}\int\limits_{x_0}^{+\infty}yf(y)\mathrm{d}y\\ &=\frac{1}{P(X\geq x_0)}\int\limits_{x_0}^{+\infty}y\frac{x_m^k}{y^{k+1}}\mathrm{d}y=\frac{kx_m^k}{P(X\geq x_0)}\int\limits_{x_0}^{+\infty}\frac{1}{y^k}\mathrm{d}y\\ &=\frac{kx_m^k}{P(X\geq x_0)}\frac{1}{(k-1)y^{k-1}}\Bigg|_{x_0}^{+\infty}=\frac{1}{P(X\geq x_0)}\frac{kx_m^k}{(k-1)x_0^{k-1}}\end{aligned}\tag{5.6.152}$$

Knowing that (see section Statistics page 750):

$$P(X\geq x_0)=\int\limits_{x_0}^{+\infty}k\frac{x_m^k}{x^{k+1}}\mathrm{d}x=\frac{x_m^k}{x_0^k}=\left(\frac{x_m}{x_0}\right)^k\tag{5.6.153}$$

We finally have:

$$\mathrm{E}(Y)=\frac{k}{k-1}x_0\tag{5.6.154}$$

$\mathrm{E}(Y)$ represents also the average income of those with a income above x_0 and as can be seen from the above equality it is in a constant ratio, greater than 1, to your income x_0.

We can check this result by doing a Monte Carlo simulation in a spreadsheet software (it is interesting to mention it to generalize to situations not computable by hand). You just need to simulate the inverse of the distribution function:

$$X=\left(\frac{x_m^k}{1-P(X\leq x)}\right)^{1/k}\tag{5.6.155}$$

in Microsoft® Excel® 11.8346:

```
=($B$7^$B$6/(1-RANDBETWEEN(1,10000)/10000))^(1/$B$6)
```

and then take the average of the values obtained above or equal to a given X (which corresponds to x_0) and ensure that we get the good results as proved above!

Therefore, generally for a discrete variable, the "conditional mean" or "conditional expectation" is commonly defined by:

$$\mathrm{E}(X|Y)(y) \equiv \mathrm{E}(X|Y=y) \equiv \sum_x x \cdot P(X=x|Y=y) \tag{5.6.156}$$

But this is only the case when we take an equality with the prior variable! If we generalize it to any other order relation, then a better definition would be:

$$\boxed{\mathrm{E}(X|Y\mathcal{R}y) \equiv \sum_x x \cdot P(X=x|Y\mathcal{R}y)} \tag{5.6.157}$$

Obviously, we could also calculate the conditional variance (verbatim the conditional standard deviation) as we will see it further below!

Theorem 5.38. Let us prove now a property that will be useful to us especially during our study of financial time series. It is based on the fact that sometimes we have to calculate such relations:

$$\boxed{\mathrm{E}\left(\mathrm{E}(X|Y)(y)\right) = \mathrm{E}(X)} \tag{5.6.158}$$

that are named "iterated conditional mean" or "law of total expectation" (and even sometimes named the "tower rule"...) also sometimes denoted $\mathrm{E}\left(\mathrm{E}(X|Y=y)\right)$ or $\mathrm{E}\left(\mathrm{E}(X|\pi)\right)$.

Proof 5.38.1.

$$\begin{aligned}\mathrm{E}\left(\mathrm{E}(X|Y)\right) &= \sum_y \mathrm{E}(X|Y=y)\cdot P(Y=y)\\ &= \sum_y\left(\sum_x x\cdot P(X=x|Y=y)\right)\cdot P(Y=y)\\ &= \sum_y\sum_x x\cdot P(X=x|Y=y)\cdot P(Y=y)\\ &= \sum_y\sum_x x\cdot P(Y=y|X=x)\cdot P(X=x)\\ &= \sum_x x\cdot P(X=x)\cdot\left(\sum_y P(Y=y|X=x)\right)\\ &= \sum_x x\cdot P(X=x)\\ &= \mathrm{E}(X)\end{aligned} \tag{5.6.159}$$

■ Q.E.D.

Let us do the same for the variance (see section Statistics page 634)! First let us recall the definition of the variance (or "unconditional variance"). If Y is a random variable, and we denote:

$$Z = (Y - \mathrm{E}(Y))^2 \tag{5.6.160}$$

then the unconditional variance of Y is obviously defined as:

$$\mathrm{V}(Y)=\mathrm{E}(Z) \tag{5.6.161}$$

Now let us now consider two random variables X, Y and we want to calculate the conditional variance of Y given X that we will denote $\mathrm{V}(Y|X)$. Similarly as the unconditional variance we can put:

$$Z=(Y-\mathrm{E}(Y|X))^2 \tag{5.6.162}$$

The conditional variance will be defined as:

$$\mathrm{V}(Y|X)=\mathrm{E}(Z|X) \tag{5.6.163}$$

Keep in mind that the unconditional variance is a scalar, when the conditional variance remains a random variable itself! So to get a scalar for the conditional variance, we have to estimate that latter for a given value of x, what we will denote by $\mathrm{V}(Y|X=x)$.

Thus:

$$\begin{aligned}\mathrm{V}(Y|X=x)&=\mathrm{E}\left((Y-\mathrm{E}(Y|X=x))^2|X\right)=\mathrm{E}\left(Y^2-2Y\mathrm{E}(Y|X=x)+(\mathrm{E}(Y|X=x)^2)|X\right)\\&=\mathrm{E}\left(Y^2|X=x\right)-2\mathrm{E}\left(Y\mathrm{E}(Y|X=x)|X=x\right)+\mathrm{E}\left(\mathrm{E}(Y|X=x)^2|X=x\right)\end{aligned} \tag{5.6.164}$$

As the terms $\mathrm{E}(Y|X=x)$ are scalars, we can take them out:

$$\begin{aligned}\mathrm{V}(Y|X=x)&=\mathrm{E}\left(Y^2|X=x\right)-2\mathrm{E}(Y|X=x)\mathrm{E}\left(Y|X=x\right)+\mathrm{E}(Y|X=x)^2\mathrm{E}\left(1|X=x\right)\\&=\mathrm{E}\left(Y^2|X=x\right)-2\left(\mathrm{E}(Y|X=x)\right)^2+\mathrm{E}(Y|X=x)^2\mathrm{E}\left(1|X=x\right)\end{aligned} \tag{5.6.165}$$

The expected conditional value of a constant is always equal to that constant, i.e.:

$$\mathrm{E}(1|X=x)=1 \tag{5.6.166}$$

Therefore:

$$\begin{aligned}\mathrm{V}(Y|X=x)&=\mathrm{E}\left(Y^2|X=x\right)-2\left(\mathrm{E}(Y|X=x)\right)^2+\mathrm{E}(Y|X=x)^2\\&=\mathrm{E}\left(Y^2|X=x\right)-\left(\mathrm{E}(Y|X=x)\right)^2\end{aligned} \tag{5.6.167}$$

We recognize in the last equality above, the Huygens relation (see page 593) of the unconditional mean. Therefore the conditional variance is defined as:

$$\boxed{\mathrm{V}(Y|X=x)=\mathrm{E}\left(Y^2|X=x\right)-\left(\mathrm{E}(Y|X=x)\right)^2} \tag{5.6.168}$$

Explicitly this will be written for continuous variables:

$$\mathrm{V}(Y|X=x)=\int\limits_y y^2f_{y|x}(y|x)\mathrm{d}y-\left(\int\limits_y yf_{y|x}(y|x)\mathrm{d}y\right)^2 \tag{5.6.169}$$

Theorem 5.39. Now let us prove the law of "iterated conditional variance", also named the "law of total variance":

$$\boxed{\mathrm{V}(Y)=\mathrm{E}(\mathrm{V}(Y|X=x))+\mathrm{V}(\mathrm{E}(Y|X=x))} \tag{5.6.170}$$

Proof 5.39.1. First we start form Huygens relation (see page 593):

$$\mathrm{V}(Y)=\mathrm{E}(Y^2)-\mathrm{E}(Y)^2 \tag{5.6.171}$$

The law of iterated means give us:

$$\mathrm{E}(Y^2)=\mathrm{E}\left(\mathrm{E}(Y^2|X=x)\right) \tag{5.6.172}$$

Injecting it in the previous relation, we get:

$$\mathrm{V}(Y)=\mathrm{E}\left(\mathrm{E}(Y^2|X=x)\right)-\mathrm{E}(Y)^2 \tag{5.6.173}$$

Using the result we have get earlier above:

$$\mathrm{V}(Y|X=x)=\mathrm{E}\left(Y^2|X=x\right)-(\mathrm{E}(Y|X=x))^2 \Leftrightarrow \mathrm{E}\left(Y^2|X=x\right)=\mathrm{V}(Y|X=x)+(\mathrm{E}(Y|X=x))^2 \tag{5.6.174}$$

and injecting that in the previous relation, we get:

$$\begin{aligned}\mathrm{V}(Y)&=\mathrm{E}\left[\mathrm{V}(Y|X=x)+(\mathrm{E}(Y|X=x))^2\right]-\mathrm{E}(Y)^2\\&=\mathrm{E}\left(\mathrm{V}(Y|X=x)\right)+\mathrm{E}\left((\mathrm{E}(Y|X=x))^2\right)-\mathrm{E}(Y)^2\end{aligned} \tag{5.6.175}$$

Now we use again the law of iterated means that we square:

$$(\mathrm{E}(Y))^2=(\mathrm{E}(\mathrm{E}(Y|X=x)))^2 \tag{5.6.176}$$

And we inject it in the previous relation to get:

$$\mathrm{V}(Y)=\mathrm{E}\left(\mathrm{V}(Y|X=x)\right)+\mathrm{E}\left((\mathrm{E}(Y|X=x))^2\right)-(\mathrm{E}(\mathrm{E}(Y|X=x)))^2 \tag{5.6.177}$$

We recognize in that latter equality the Huygens relation (see page 593)! Indeed:

$$\mathrm{V}(Y)=\mathrm{E}\left(\mathrm{V}(Y|X=x)\right)+\underbrace{\mathrm{E}\left((\mathrm{E}(Y|X=x))^2\right)-(\mathrm{E}(\mathrm{E}(Y|X=x)))^2}_{\mathrm{V}(\mathrm{E}(Y|X=x))} \tag{5.6.178}$$

So finally:

$$\mathrm{V}(Y)=\mathrm{E}(\mathrm{V}(Y|X=x))+\mathrm{V}(\mathrm{E}(Y|X=x)) \tag{5.6.179}$$

■ Q.E.D.

5.6.3.2 Bayesian Networks

"Bayesian networks", also named "Bayesian Belief Nets" abbreviated BNNs, or "Causal Bayesian networks, are simply a graphical representation of a problem of conditional probabilities to better visualize the interaction between the different variables when they begin to be in large numbers.

This is a technique increasingly used in decision aided software (Data Mining), artificial intelligence (A.I.) and also in the analysis and risk management (ISO 31010 norm).

Bayesian networks are by definition directed acyclic graphs (see section Graphs Theory page 3017), so that an event can not (even indirectly) influence its own probability, with quantitative description of dependencies between events.

These graphs are used for both knowledge representation models and calculating conditional probabilities machines. They are mainly used for diagnosis (medical and industrial), risk analysis (diagnostics failures, faults or accidents), spam detection (Bayesian filter), voice text and image opinions analysis, fraud detection or bad payers as well as data mining (M.K.M.: Mining and Knowledge Management) in general.

Remark

Many systems and software based on drawings or on information in existing databases exists to build and analyse Bayesian networks. Paid solutions: SQL Server, Oracle, Hugin. Free solutions (at this date): Tanagra, Microsoft Belief Network MSBNX 1.4.2, RapidMiner, `R` (`bnlearn` package). Personally for pedagogical reasons, I prefer the simplicity of the small software MSBNX from Microsoft. For information, in 15 years of professional experience as a consultant I have met so far only one company on more than 800 multinationals which used Bayesian networks... (in the field of transportation).

Use a Bayesian network is assimilated to do "Bayesian inference" or "Causal inference". Based on observed information (causal probabilities), we calculate the probability of possible known data but not observed.

For a given domain (e.g. medical), we describe the causal relations between variables of interest by a graph (we do not need again to specify that it is acyclic). In this graph, the causal relations between variables are not deterministic, but probabilistic. Thus, the observation of a cause or multiple causes does not always implies the effect or effects that depend on it, but only changes the probability of observing them.

The particular interest of Bayesian networks is to consider simultaneously a priori knowledge of experts (in the graph) and experience contained in the data.

Example of five variables with their relations (directed acyclic graph) and numbering of states/variables:

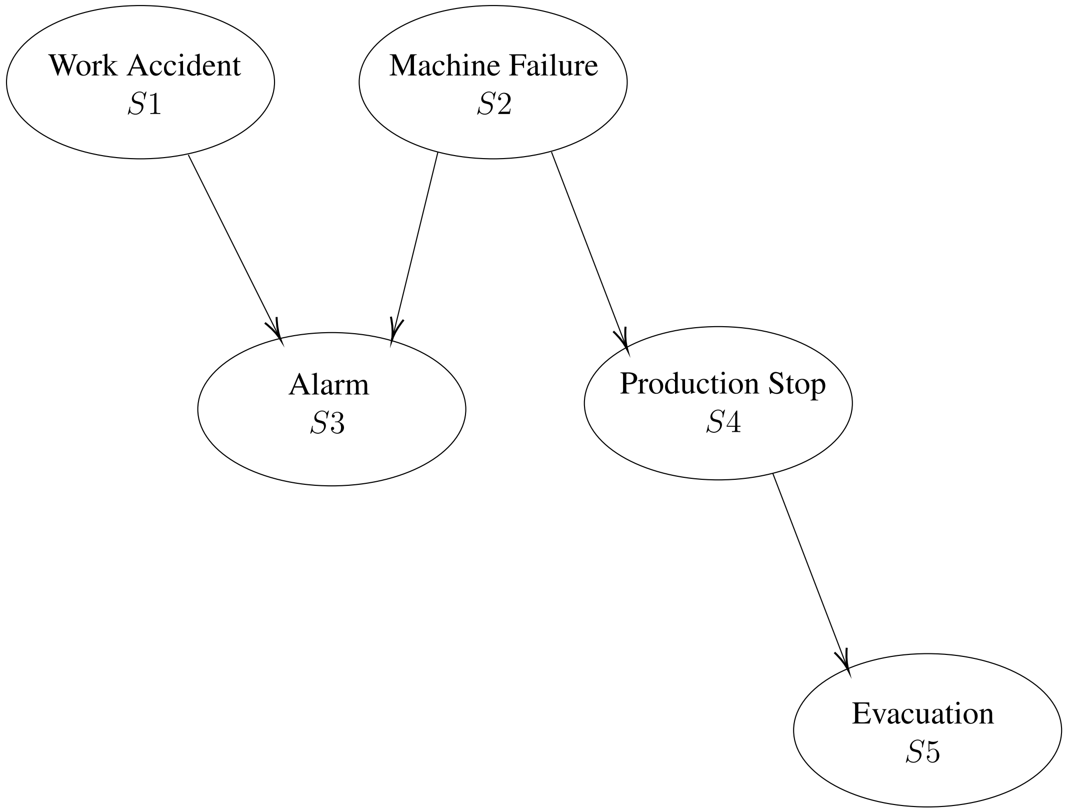


Figure 5.85: **Example of Bayesian network (acyclic oriented) with 5 states**

Obviously, the construction of the causal graph is based primarily on return of experience (REX) and sometimes results on standards or reports of expert committees. In computing science, the causal graph automatically change depending on the content of databases (think at the Amazon book store in real time target advertisements based on your past purchases or at the Genius Apple service). But we can rarely think to all possibilities and there will sometimes hidden states between two known states that have been forgotten and that would have allowed to better modelize the situations.

Suppose in the example above that with the help of a corporate database, we know that in about $100,000$ man-days, we had in the company $1,000$ accidents (i.e.. 1% of total) and 100 machines failures (i.e.. $0.01\%$of total). Then we represent it in the traditional form as follows:

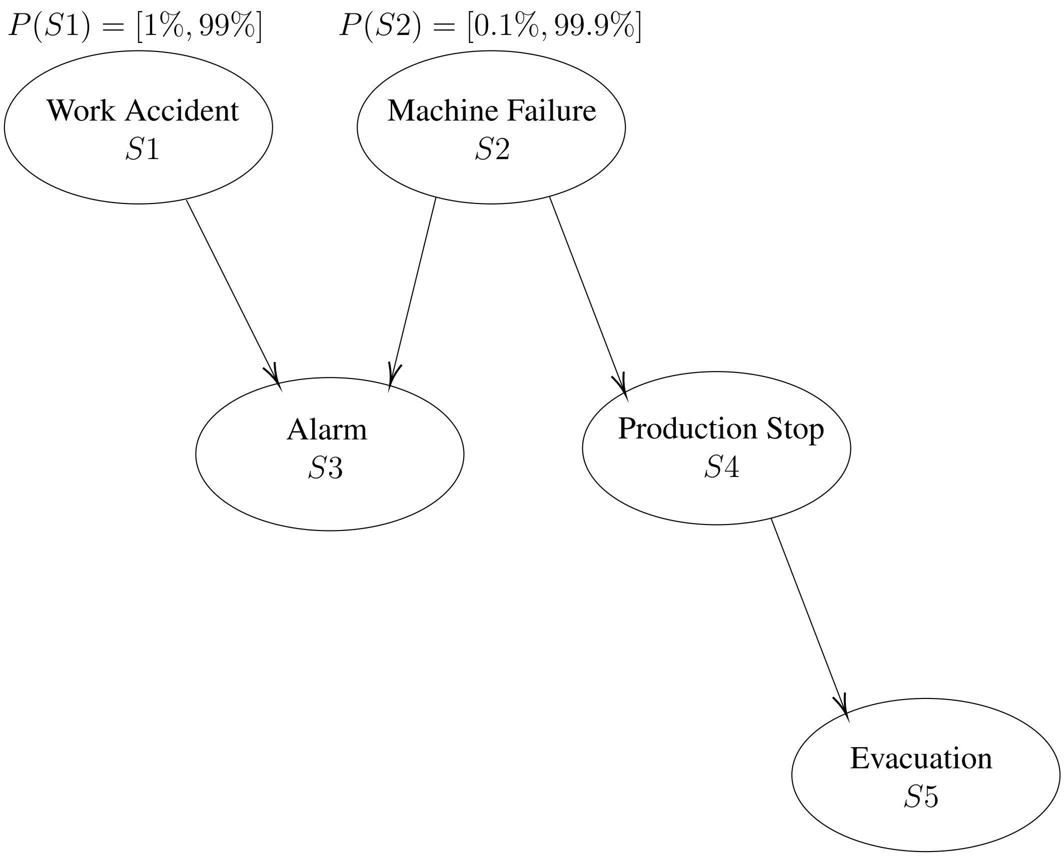


Figure 5.86: **Directed acyclic Bayesian network with probability of departure**

where we have the subset $S2, S4, S5$ which is what experts name a "serial or linear connection", the triplet $S3, S2, S4$ is a named a "divergent relation" (if the arrows were reversed for the triplet, we would have a "converging relation").

Before going further with our example we will make some observations in relation to these three types of relations:

For clarity, we distinguish first "conditional independence" and "conditional dependence".

We say that events A and C are "conditionally independent" if given an event B the following equality holds:

$$P(A/B) = P(A/B, C) \tag{5.6.180}$$

So the term "conditional" implies the presence of B and the fact that C does not influence the probability of the event A.

About "conditional dependence", this time we can distinguish three types of relations.

1. The conditional dependence of the following type is named a "serial or linear connection" (already mentioned above):

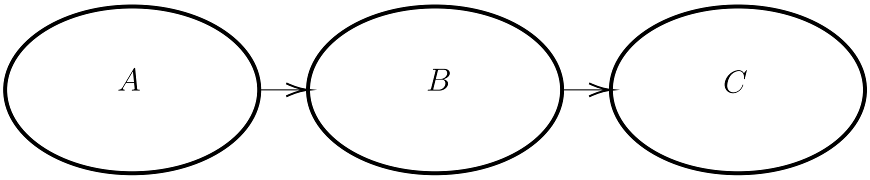


Figure 5.87: **Serial/linear Bayesian network**

where A, B and C are dependent (in this particular example there are 3 dependent nodes A, B and C, but in general this dependence relates to all nodes if there were more than 3)

In addition, A and C are conditionally dependent to B. But if the variable B is known, A no longer provides any useful information about C (the path of uncertainty is somehow broken) and therefore A and C become conditionally independent. We then have the conditional probability that simplify as follows:

$$P(C/B,A)=P(C/B) \tag{5.6.181}$$

2. The conditional dependence of the following type is named a "divergent connection" (as already mentioned above):

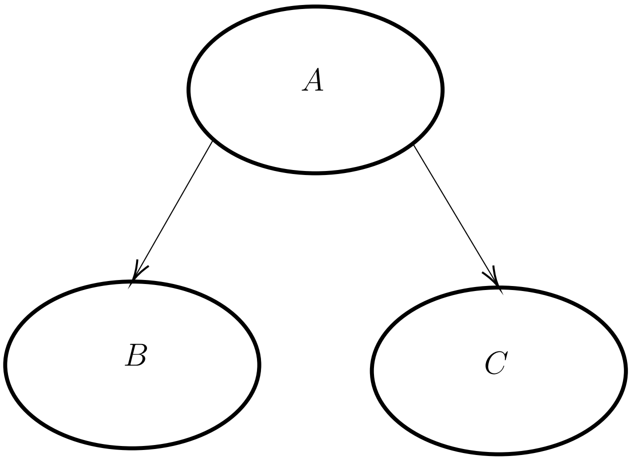


Figure 5.88: **Divergent Bayesian network**

In addition, B and C are conditionally dependent on A. But if A is known, B does not provide any more information on C (again the path of uncertainty is somehow broken) and therefore B and C become independent. We then have for example if A is known:

$$P(C/A,B)=P(C/A) \tag{5.6.182}$$

3. The conditional dependence of the following type is named a "convergence connection" or "V-Structure" (as already mentioned above):

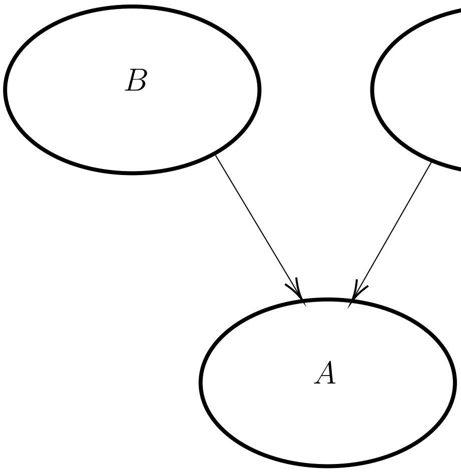


Figure 5.89: **Convergent Bayesian network**

where this time the parents are independent.

So B and C are independent but become conditionally dependent on A. If A is known, then we have:

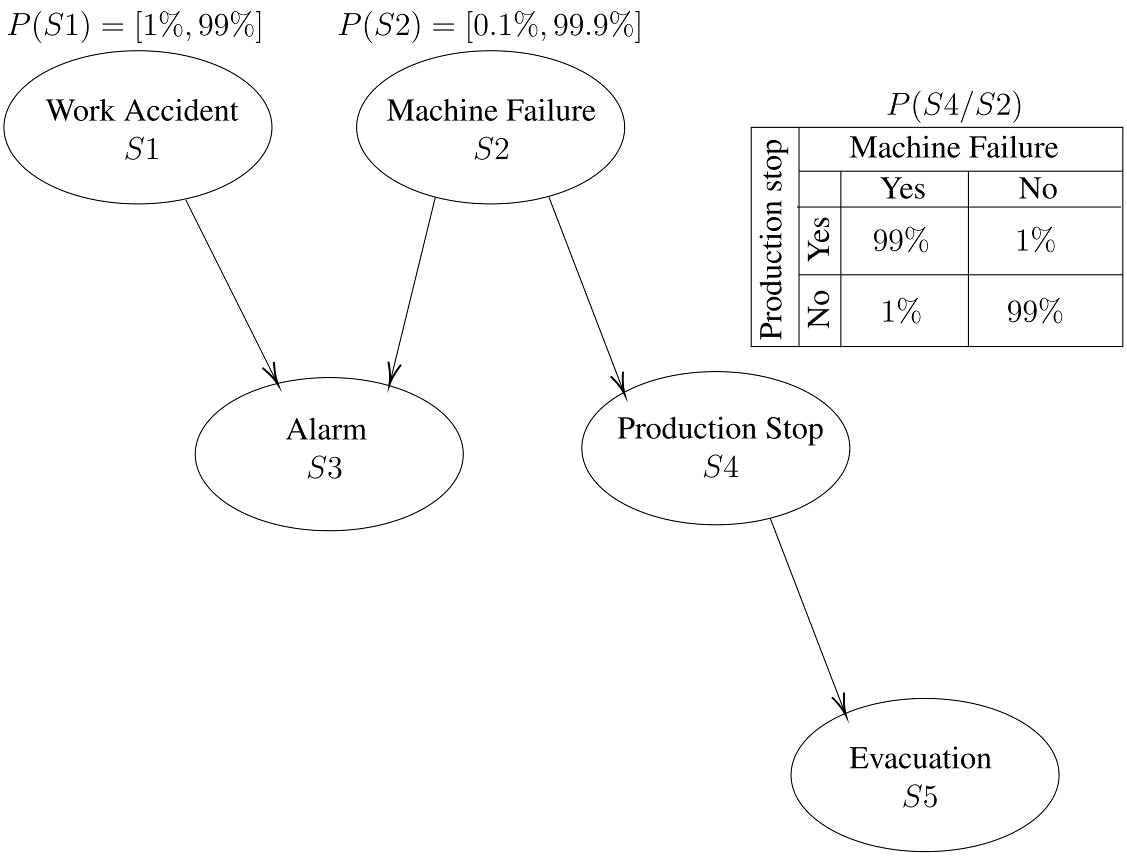
$$P(A/B,C)=P(A/B) \tag{5.6.183}$$

The dependence between parents therefore requires the observation of their common child.

Now to make a concrete example, let us first recall the Bayes' theorem:

$$\underbrace{P(A/B)}_{\text{posterior probability}}=\frac{\overbrace{P(B/A)}^{\text{Likelihood}}\cdot\overbrace{P(A)}^{\text{priori probability}}}{\underbrace{P(B)}_{\text{expectedness}}} \tag{5.6.184}$$

and now suppose that we have a knowledge database that gives us (thanks to quality managers who always inputs the quality issues) that when a machine failure occurred, 99 times out of 100 (99%) there has been a total production stop (i.e. 1% of time there was no production stop) and on all stop production 1% was not due to a machine failure. What we traditionally represent as follows:


$P(S1) = [1\%, 99\%]$
$P(S2) = [0.1\%, 99.9\%]$
Work Accident $S1$
Machine Failure $S2$
Alarm $S3$
Production Stop $S4$
Evacuation $S5$

$P(S4/S2)$

		Machine Failure	
		Yes	No
Production stop	Yes	99%	1%
	No	1%	99%

Figure 5.90: Directed acyclic Bayesian network with probability of departure

So the "implicit probability" (expectedness) that there is a production stop is given by:

$$P(S4) = P(S4/S2)P(S2) + P(S4/\overline{S2})P(\overline{S2}) = 99\% \cdot 0.1\% + 1\% \cdot 99.9\% \cong 1.098\% \quad (5.6.185)$$

This value represents the implicit proportion of productions stop from the $100,000$ man-days (so we can give a proportion of rows in the database that represents a production stop regardless of the cause and even without knowing the details of the database).

It then follows immediately that the implicit probability (expectedness) that there is no production stop is given by:

$$P(S4) = 1 - P(\overline{S4}) \cong 98.902\% \quad (5.6.186)$$

This is consistent with what gives us the freeware MSBNX 1.4.2:

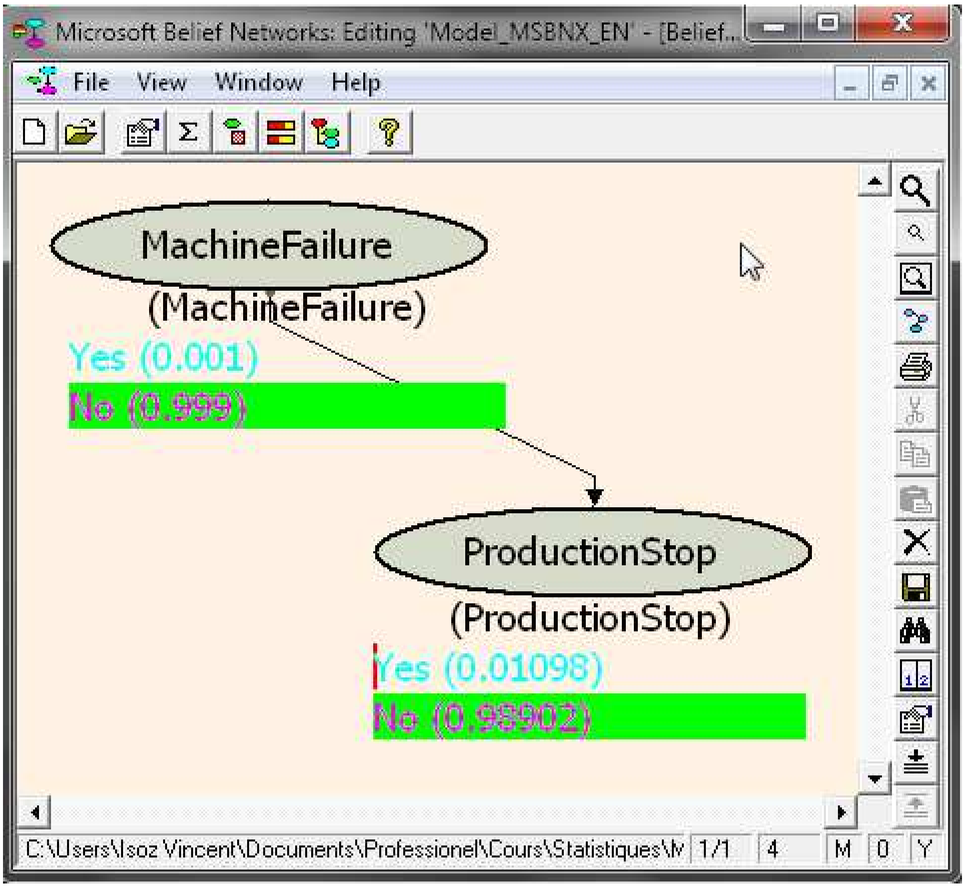
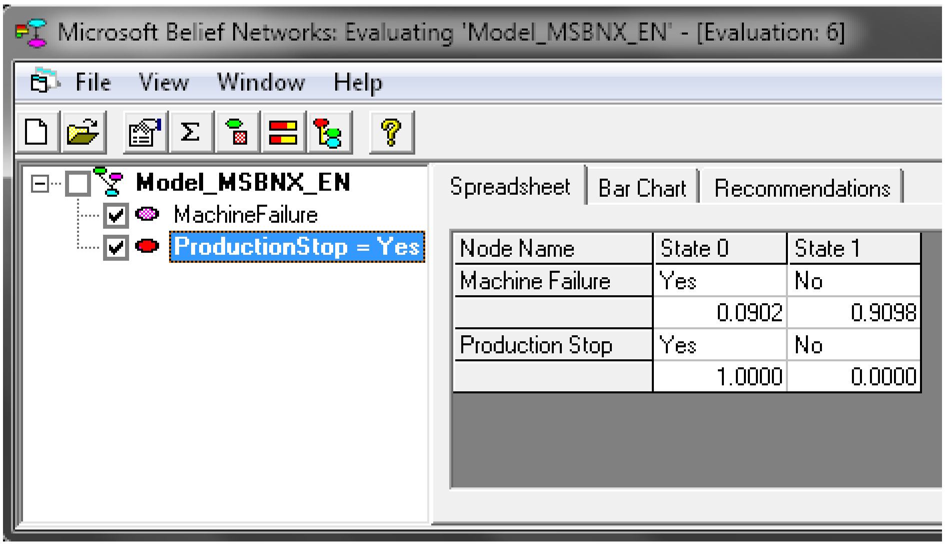

Microsoft Belief Networks: Editing 'Model_MSBNX_EN' - [Belief...
File View Window Help
MachineFailure
(MachineFailure)
Yes (0.001)
No (0.999)
ProductionStop
(ProductionStop)
Yes (0.01098)
No (0.98902)
C:\Users\Isoz Vincent\Documents\Professionel\Cours\Statistiques\M 1/1 4 M 0 Y

Figure 5.91: **Beginning of the Bayesian network with MSBNX 1.4.2**

Now suppose we observed a production stop. What is the a posteriori probability that it is due to a machine failure? We then have:

$$P(S2/S4)=\frac{P(S4/S2)P(S2)}{P(S4)}=\frac{99\%\cdot 0.1\%}{1.098\%}\cong 9.02\% \tag{5.6.187}$$

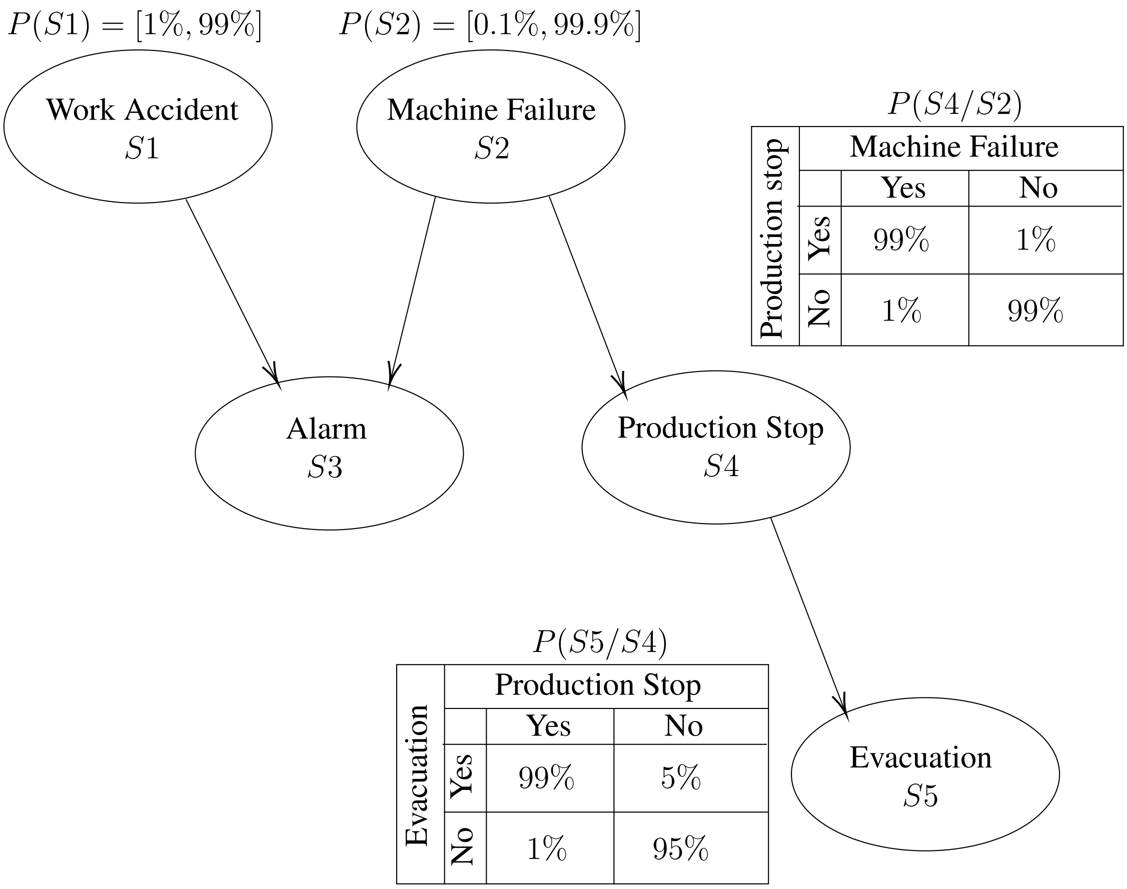
We can also check this with the software MSBNX 1.4.2:


Microsoft Belief Networks: Evaluating 'Model_MSBNX_EN' - [Evaluation: 6]
File View Window Help
Model_MSBNX_EN
MachineFailure
ProductionStop = Yes
Spreadsheet
Bar Chart
Recommendations

Node Name	State 0	State 1
Machine Failure	Yes	No
	0.0902	0.9098
Production Stop	Yes	No
	1.0000	0.0000

Figure 5.92: **A Posteriori probability of a production stop due to a machine failure in MSBNX 1.4.2**

Now, imagine that our database gives us (always thanks to quality managers who ensured to input quality issues) that 99 times out of 100 (99%) when there was a production stop, there was an evacuation. However 5% of evacuations were identified as having nothing to do with a production stop (i.e. 95% of evacuations are due to fire exercises OR other events):



$P(S4/S2)$

Production stop	Machine Failure: Yes	Machine Failure: No
Yes	99%	1%
No	1%	99%

$P(S5/S4)$

Evacuation	Production Stop: Yes	Production Stop: No
Yes	99%	5%
No	1%	95%

Figure 5.93: **2nd level Bayesian network**

Now to calculate the implicit probability (expectedness) of evacuations compared to machines failures, we saw that when we had a conditional dependence series, the conditional probability depends only on the direct parent. Thus, we get the following expectedness:

$$P(S5)=P(S5/S4)P(S4)+P(S5/\overline{S4})P(\overline{S4})=99\%\cdot 1.089\%+5\%\cdot 98.092\%\cong 6.03\% \tag{5.6.188}$$

We can also check that this result is compliant with the on given by the software MSBNX 1.4.2:

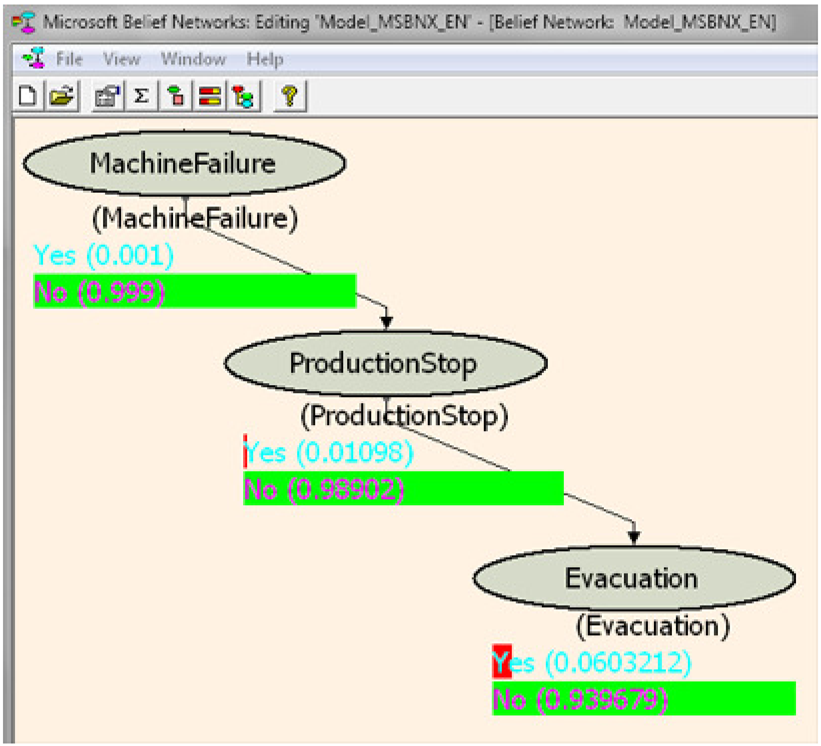

Microsoft Belief Networks: Editing 'Model_MSBNX_EN' - [Belief Network: Model_MSBNX_EN]
File View Window Help
MachineFailure
(MachineFailure)
Yes (0.001)
No (0.999)
ProductionStop
(ProductionStop)
Yes (0.01098)
No (0.98902)
Evacuation
(Evacuation)
Yes (0.0603212)
No (0.939679)

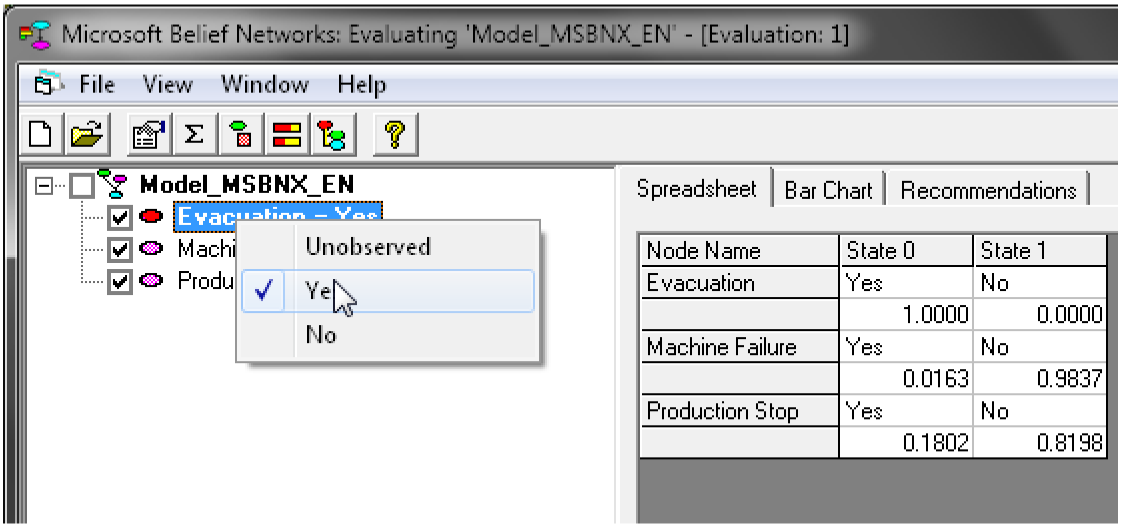
Figure 5.94: Implicit probability of an evacuation in MSBNX 1.4.2

So the implicit probability (expectedness) of evacuation does not actually depend on machine failures.

Now suppose we have observed an evacuation. We want to know what is the a posteriori probability that it is due to a machine failure! Then we have:

$$P(S4/S5)=\frac{P(S5/S4)P(S4)}{P(S5)}=\frac{99\%\cdot 1.098\%}{6.03\%}\cong 18.02\% \qquad (5.6.189)$$

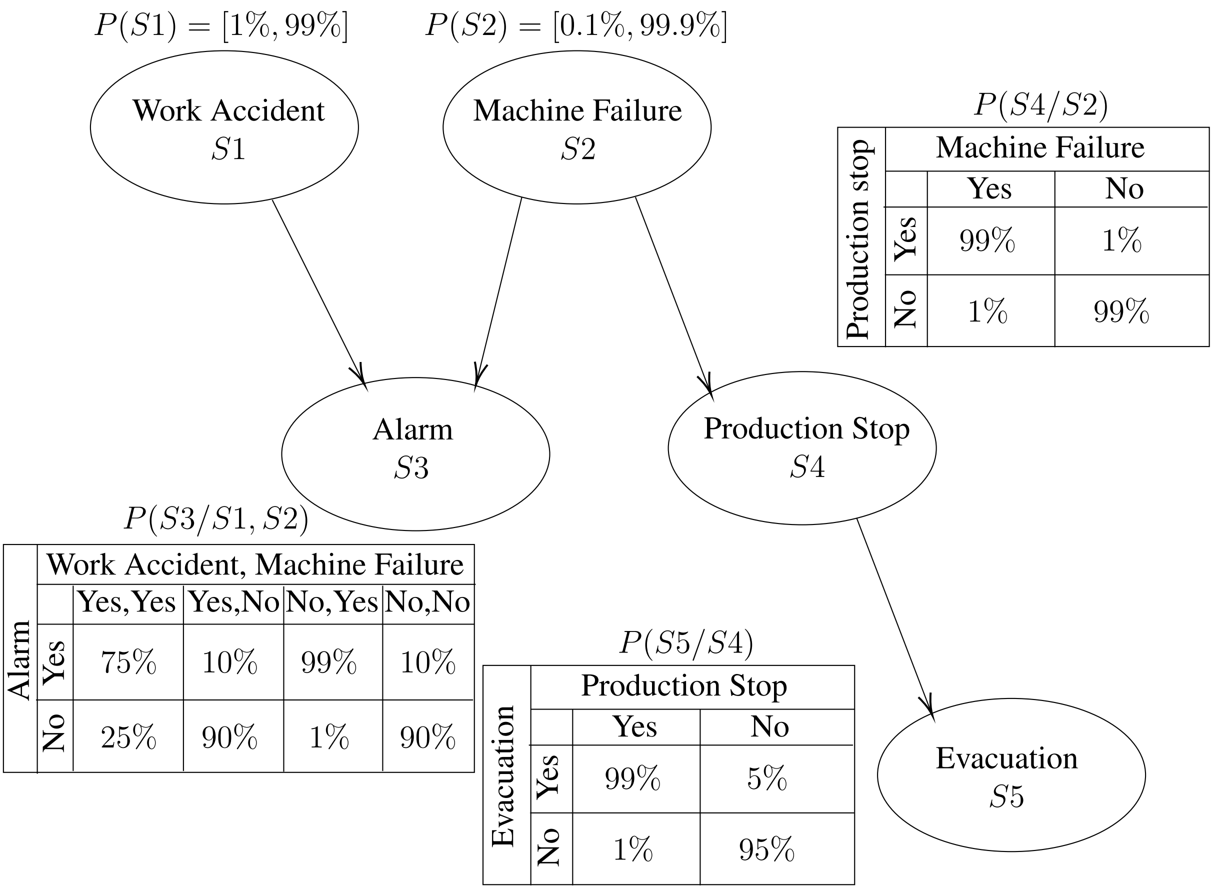
We can also check this with the software MSBNX 1.4.2:


Microsoft Belief Networks: Evaluating 'Model_MSBNX_EN' - [Evaluation: 1]
File View Window Help
Model_MSBNX_EN
Evacuation = Yes
Machi
Produ
Unobserved
Yes
No
Spreadsheet
Bar Chart
Recommendations

Node Name	State 0	State 1
Evacuation	Yes	No
	1.0000	0.0000
Machine Failure	Yes	No
	0.0163	0.9837
Production Stop	Yes	No
	0.1802	0.8198

Figure 5.95: A Posteriori probability of an evacuation due to a machine failure in MSBNX 1.4.2

Now we study the case with the alarm and again a database allows us to build a table with different probabilities:



$P(S4/S2)$

		Machine Failure	
		Yes	No
Production stop	Yes	99%	1%
	No	1%	99%

$P(S3/S1,S2)$

		Work Accident, Machine Failure			
		Yes,Yes	Yes,No	No,Yes	No,No
Alarm	Yes	75%	10%	99%	10%
	No	25%	90%	1%	90%

$P(S5/S4)$

		Production Stop	
		Yes	No
Evacuation	Yes	99%	5%
	No	1%	95%

Figure 5.96: **2nd level Bayesian network with second branch**

Now to calculate the implicit probability (expectedness) that there is an alarm, we will have to consider the four possible situations. We then use again the theorem of total probability:

$$\begin{aligned}P(S3)=&P(S3/S1,S2)P(S1)P(S2)+P(S3/\overline{S1},S2)P(\overline{S1})P(S2)\\&+P(S3/S1,\overline{S2})P(S1)P(\overline{S2})+P(S3/\overline{S1},\overline{S2})P(\overline{S1})P(\overline{S2})\end{aligned}\tag{5.6.190}$$

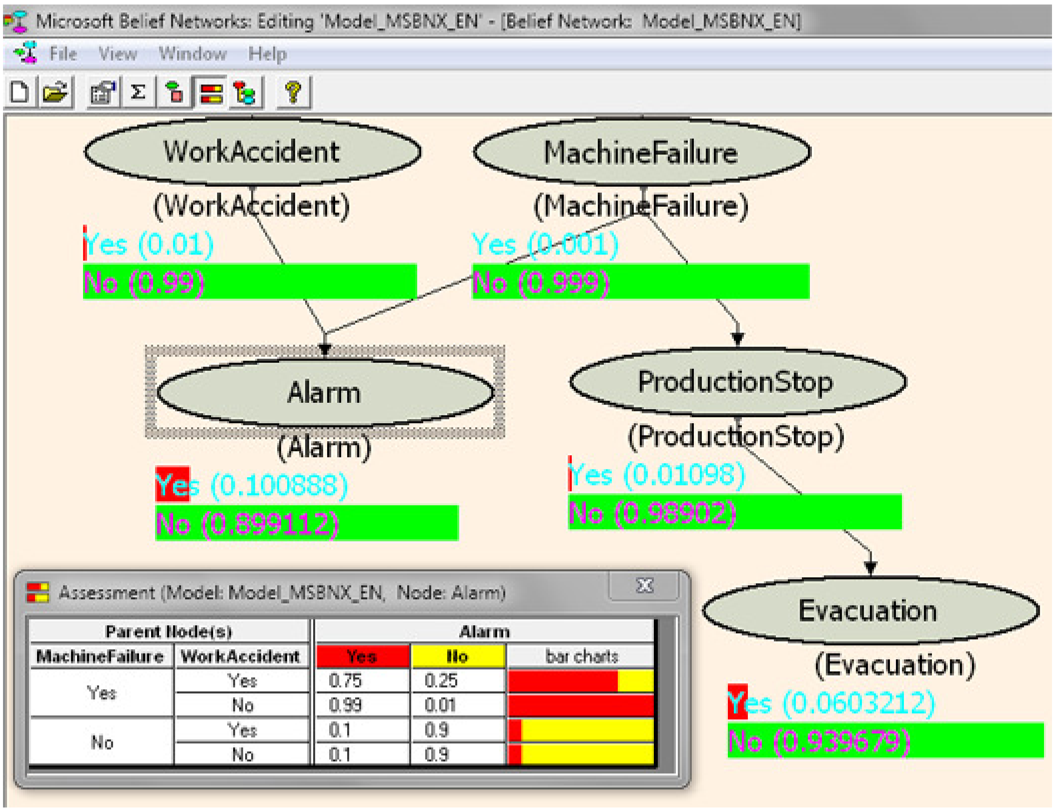
What a little more rigorously should be written:

$$\begin{aligned}P(S3)=&P(S3/S1\wedge S2)P(S1)P(S2)+P(S3/\overline{S1}\wedge S2)P(\overline{S1})P(S2)\\&+P(S3/S1\wedge\overline{S2})P(S1)P(\overline{S2})+P(S3/\overline{S1}\wedge\overline{S2})P(\overline{S1})P(\overline{S2})\end{aligned}\tag{5.6.191}$$

The numerical application therefore provides for the implied probability (expectedness) of an alarm:

$$\begin{aligned}P(S3)=&75\%\cdot 1\%\cdot 0.1\%+99\%\cdot 99\%\cdot 0.1\%+10\%\cdot 1\%\cdot 99.9\%\\&+10\%\cdot 99\%\cdot 99.9\%\cong 10.089\%\end{aligned}\tag{5.6.192}$$

What can be built and check as follows with MSBNX 1.4.2:


Microsoft Belief Networks: Editing 'Model_MSBNX_EN' - [Belief Network: Model_MSBNX_EN]
File View Window Help
WorkAccident
(WorkAccident)
Yes (0.01)
No (0.99)
MachineFailure
(MachineFailure)
Yes (0.001)
No (0.999)
Alarm
(Alarm)
Yes (0.100888)
No (0.899112)
ProductionStop
(ProductionStop)
Yes (0.01098)
No (0.98902)
Evacuation
(Evacuation)
Yes (0.0603212)
No (0.939679)
Assessment (Model: Model_MSBNX_EN, Node: Alarm)

Parent Node(s)		Alarm		
MachineFailure	WorkAccident	Yes	No	bar charts
Yes	Yes	0.75	0.25	
	No	0.99	0.01	
No	Yes	0.1	0.9	
	No	0.1	0.9	

Figure 5.97: **Implicit probability in MSBNX 1.4.2**

It may be useful to the reader to know that he can sometimes found in the literature the following notations:

$$P(S3/S1=\text{Yes},S2=\text{Yes})=P(S3/S1,S2)=P(S3/S1\wedge S2) \tag{5.6.193}$$

or:

$$P(S3|S1=\text{Yes},S2=\text{Yes})=P(S3|S1,S2)=P(S3|S1\wedge S2) \tag{5.6.194}$$

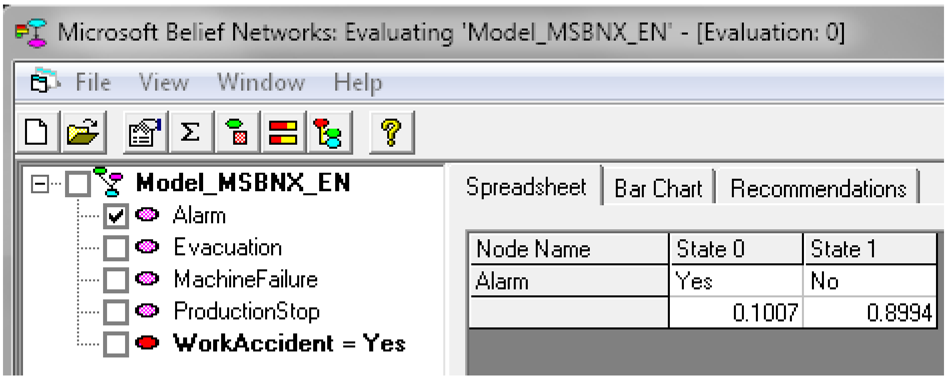
Remark

In the particular example studied above all event have only two states. But in practice they can have 3, 4 and more states. Therefore probabilities cross-tables quickly become enormous.

As in the previous case, suppose we know that there was a working accident. We wish then calculate the a posteriori probability of an alarm. We then have (observe that the probability depends actually only to the state $S2$ state since the state $S1$ is completely known!):

$$\begin{aligned}P(S3/S1)&=P(S3/S1,S2)P(S2)+P(S3/S1,\overline{S2})P(\overline{S2})\\&=75\%\cdot 0.1\%+10\%\cdot 99.9\%=10.065\%\end{aligned} \tag{5.6.195}$$

We can also check this with the software MSBNX 1.4.2:


Microsoft Belief Networks: Evaluating 'Model_MSBNX_EN' - [Evaluation: 0]
File
View
Window
Help
Spreadsheet
Bar Chart
Recommendations

☐ **Model_MSBNX_EN**
- ☑ Alarm
- ☐ Evacuation
- ☐ MachineFailure
- ☐ ProductionStop
- ☐ **WorkAccident = Yes**

Node Name	State 0	State 1
Alarm	Yes	No
	0.1007	0.8994

Figure 5.98: **Implicit probability of an alarm in MSBNX 1.4.2**

To complete this example, we would calculate the a posteriori probabilities $P(S2/S3)$ and $P(S1/S3)$. To do this, we must first calculate the a priori probabilities $P(S3/S2)$ and $P(S3/S1)$ (this last one has been calculated just before).

We have for the missing value (which can be easily checked as before with MSBNX 1.4.2 software):

$$\begin{aligned} P(S3/S2) &= P(S3/S2,S1)P(S1) + P(S3/S2,\overline{S1})P(\overline{S1}) \\ &= 75\% \cdot 1\% + 99\% \cdot 99\% \cong 98.76\% \end{aligned} \tag{5.6.196}$$

We then have:

$$\begin{aligned} P(S3/S1) &= 10.065\% \\ P(S3/S2) &= 98.76\% \end{aligned} \tag{5.6.197}$$

We now have everything we need to calculate the a posteriori probability of $P(S2/S3)$ and $P(S1/S3)$:

$$\begin{aligned} P(S2/S3) &= \frac{P(S3/S2)P(S2)}{P(S3)} = \frac{98.76\% \cdot 0.1\%}{10.089\%} = 0.9789\% \\ P(S1/S3) &= \frac{P(S3/S1)P(S1)}{P(S3)} = \frac{10.065\% \cdot 1\%}{10.089\%} = 0.9976\% \end{aligned} \tag{5.6.198}$$

So the a posteriori probability that there is a machine breakdown when we know that there is an alarm is 0.979% (i.e. 0.021% that the trigger of the alarm is not a priori due to a machine failure). Respectively there is, a posteriori, 0.998% probability that there is a work accident when we know there is an alarm (and then 0.002% that it is not a priori due to a work accident).

From the critical point of view, when there is finally an alarm we can not say a lot of things.... This is because, in this case, to the fact that the events of significant interest both have low probability to occur (work accident and machine failure) and that the employees respond quite well at the start of the alarm (otherwise if the a priori probabilities were high it would mean that the behaviour of the employees is not good because we can guess - with exasperation - in advance which problem occurs with a good confidence).

Remark

We did not find how to check these last calculations with MSBNX 1.4.2. If someone finds how to do, it would be great to give us the details of the process.

To conclude, the reader will have noticed that the calculations can quickly become annoying when the graph becomes complex and this explains the use of computer software. Furthermore, in the banking sector that uses for example Bayesian networks for credit risk, the a priori probability can be more complex. For example we might want to know the a priori probability that there is a machine failure knowing that we have an alarm and an accident:

$$P(S2/S3,S1)=P(S2/S3=\text{Yes},S1=\text{Yes}) \tag{5.6.199}$$

For people wishing to go deeper into the subject, with practical cases on `R` software and the use of Bayesian Gaussian networks, we strongly recommend reading [486]!

Remark

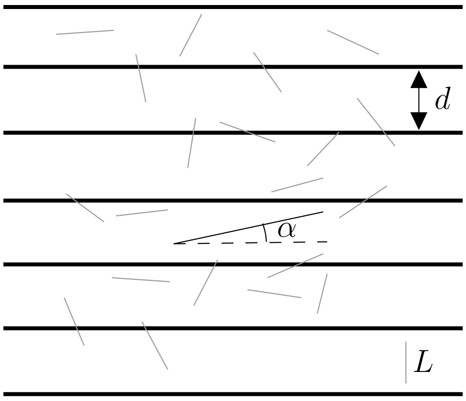
The Bayesian causal networks we have looked at so far do not show the passage of time. Yet time is essential to causality, insofar as effects must not precede their causes. Fortunately, it is easy to incorporate time into BNs by using successive time periods, named "time slices", to distinguish between the values of the same variables at different times or in different time periods. This creates a "dynamic Bayesian network" (DBN) that shows dependencies among the values of different random variables over time.

5.6.4 Geometric Probabilities

In mathematics, Buffon's needle problem is a question first posed in the 118th century (holocene calendar) by Georges-Louis Leclerc, Comte de Buffon. Buffon's needle is very interesting because it is known as the earliest problem in "geometric probability" to be solved using integral geometry. Furthermore there are many similar cases of calculations in particle physics and astrophysics.

The problem is stated as following: Suppose that you drop a short needle on ruled paper what is then the probability that the needle comes to lie in a position where it crosses one of the lines?

The probability depends on the distance d between the lines of the ruled paper, and it depends on the length L of the needle that we drop - or rather it depends only on the ratio L/d. A short needle for our purpose is one of length $L\leq d$. In other words, a short needle is one that cannot cross two lines at the same time (and will come to touch two lines only with probability zero).



The trick to get a solution is to consider the slope of the needle; let's say it drops to lie with an angle of α away from horizontal, where α will be in the range $0 \leq \alpha \leq \frac{\pi}{2}$ (we will ignore the case where the needle comes to lie with negative slope, since that case is symmetric to the case of positive slope, and produces the same probability). A needle that lies with angle α has height $L\sin(\alpha)$, and the probability that such a needle crosses one of the horizontal lines of distance d is:

$$\frac{L\sin(\alpha)}{d} \tag{5.6.200}$$

Thus we get the probability by averaging over the possible angles α, as:

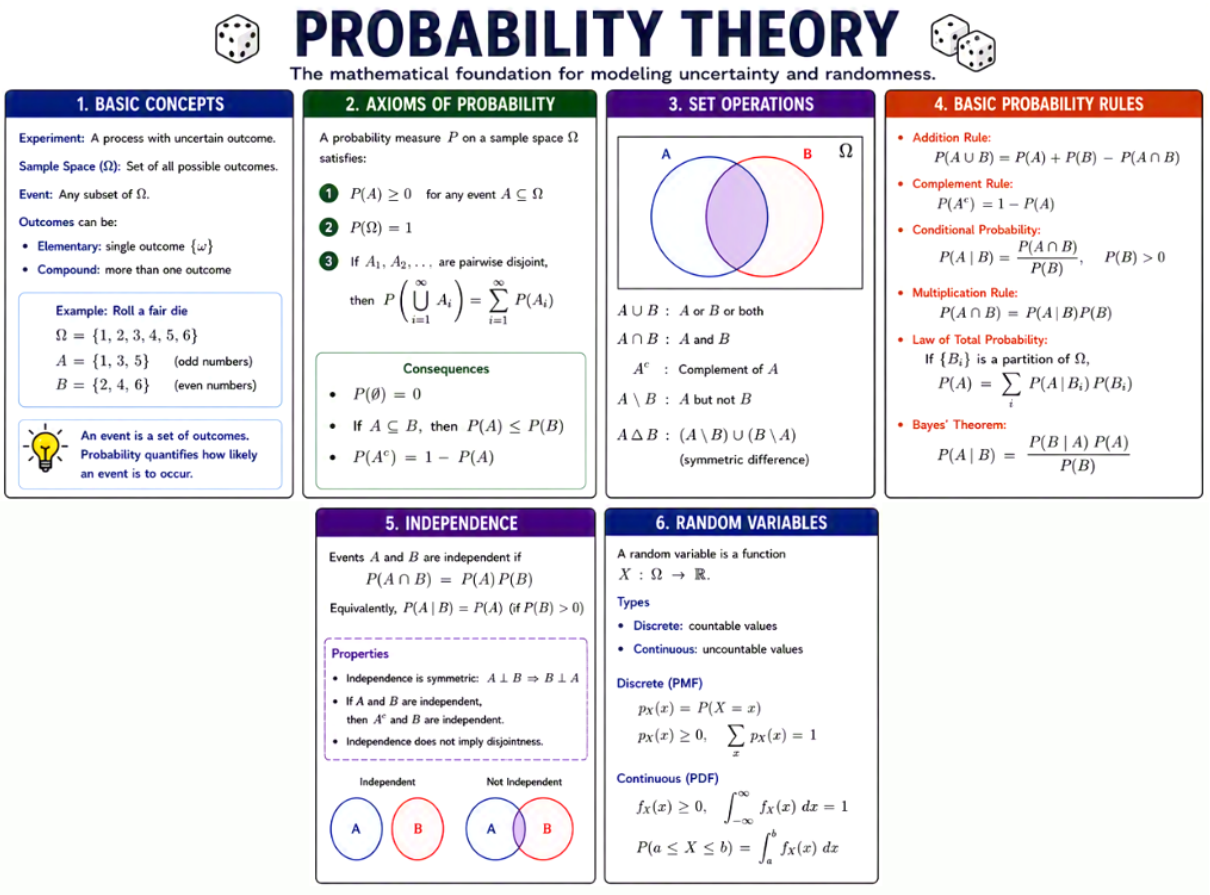
$$P = \frac{1}{\pi/2}\int\limits_0^{\pi/2}\frac{L\sin(\alpha)}{d}\mathrm{d}\alpha = \frac{2}{\pi}\frac{L}{d}\int\limits_0^{\pi/2}\sin(\alpha)\mathrm{d}\alpha = \frac{2}{\pi}\frac{L}{d}\left[-\cos(x)\right]_0^{\pi/2} = \frac{2}{\pi}\frac{L}{d} \tag{5.6.201}$$

For a long needle, we get the same probability $\frac{L\sin(\alpha)}{d}$ as long as $L\sin(\alpha) \leq d$ that is, in the range $0 \leq \alpha \leq \arcsin(d/L)$. However, for larger angles c the needle must cross a line, so the probability is 1. Hence we compute:

$$\begin{aligned} P &= \frac{2}{\pi}\left(\int\limits_0^{\arcsin(d/L)}\frac{L\sin(\alpha)}{d}\mathrm{d}\alpha + \int\limits_{\arcsin(d/L)}^{\pi/2} 1 d\mathrm{d}\alpha\right) \\ &= \frac{2}{\pi}\left[\frac{L}{d}[-\cos\alpha]_0^{\arcsin(d/L)} + \left(\frac{\pi}{2} - \arcsin\left(\frac{d}{L}\right)\right)\right] \\ &= 1 + \frac{2}{\pi}\left[\frac{L}{d}\left(1 - \sqrt{1 - \frac{d^2}{L^2}}\right) - \arcsin\left(\frac{d}{L}\right)\right] \end{aligned} \tag{5.6.202}$$

for $L \leq d$.

This last relation yields obviously $2/\pi$ for $L = d$. And it tends to 1 for $L \to +\infty$.


A
B
Ω
Independent
Not Independent
A
B
A
B

PROBABILITY THEORY

The mathematical foundation for modeling uncertainty and randomness.

1. BASIC CONCEPTS

Experiment: A process with uncertain outcome.

Sample Space (Ω): Set of all possible outcomes.

Event: Any subset of Ω.

Outcomes can be:

- **Elementary:** single outcome $\{\omega\}$
- **Compound:** more than one outcome

Example: Roll a fair die

$\Omega = \{1, 2, 3, 4, 5, 6\}$

$A = \{1, 3, 5\}$ (odd numbers)

$B = \{2, 4, 6\}$ (even numbers)

An event is a set of outcomes. Probability quantifies how likely an event is to occur.

2. AXIOMS OF PROBABILITY

A probability measure P on a sample space Ω satisfies:

① $P(A) \geq 0$ for any event $A \subseteq \Omega$

② $P(\Omega) = 1$

③ If $A_1, A_2, \ldots$ are pairwise disjoint, then $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$

Consequences

- $P(\emptyset) = 0$
- If $A \subseteq B$, then $P(A) \leq P(B)$
- $P(A^c) = 1 - P(A)$

3. SET OPERATIONS

$A \cup B$: A or B or both

$A \cap B$: A and B

A^c : Complement of A

$A \setminus B$: A but not B

$A \Delta B$: $(A \setminus B) \cup (B \setminus A)$ (symmetric difference)

4. BASIC PROBABILITY RULES

- **Addition Rule:** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- **Complement Rule:** $P(A^c) = 1 - P(A)$
- **Conditional Probability:** $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}, \quad P(B) > 0$
- **Multiplication Rule:** $P(A \cap B) = P(A \mid B)P(B)$
- **Law of Total Probability:** If $\{B_i\}$ is a partition of Ω, $P(A) = \sum_i P(A \mid B_i)\, P(B_i)$
- **Bayes' Theorem:** $P(A \mid B) = \dfrac{P(B \mid A)\, P(A)}{P(B)}$

5. INDEPENDENCE

Events A and B are independent if

$$P(A \cap B) = P(A)\, P(B)$$

Equivalently, $P(A \mid B) = P(A)$ (if $P(B) > 0$)

Properties

- Independence is symmetric: $A \perp B \Rightarrow B \perp A$
- If A and B are independent, then A^c and B are independent.
- Independence does not imply disjointness.

6. RANDOM VARIABLES

A random variable is a function $X : \Omega \rightarrow \mathbb{R}$.

Types

- **Discrete:** countable values
- **Continuous:** uncountable values

Discrete (PMF)

$$p_X(x) = P(X = x)$$

$$p_X(x) \geq 0, \quad \sum_x p_X(x) = 1$$

Continuous (PDF)

$$f_X(x) \geq 0, \quad \int_{-\infty}^{\infty} f_X(x)\, dx = 1$$

$$P(a \leq X \leq b) = \int_a^b f_X(x)\, dx$$

5.6.5 Martingales

A martingale in probabilities (there is another one in stochastic processes) is a technique to increase the chances of winning in gambling while respecting the rules of the game. The principle is completely dependent on the type of game we are focusing, but this term is accompanied by an aura of mystery that some players would know efficient secret techniques to cheat with chance. For example, many players (or candidates to play) search THE martingale that will beat the bank in the most common games in casinos (institutions whose profitability relies almost entirely on the difference - however small - between the chances of winning and losing as we will see in details further below).

Many martingales are the dream of their author, some are actually inapplicable, some could actually give the possibility to cheat a little. Gambling in general are unfair: whatever the shot played, the probability of winning of the casino (or of the State in the case of a lottery) is greater than this of the player. In this type of game, it is not possible to inverse the chances, just to minimize the probability of gambler's ruin.

The most common example is the roulette wheels martingale. It consists to play a single chance to the roulette wheels (red or black, odd or even) to win, for example, a unit in a series of moves by doubling his bet if we lose, and that until we earn. Example: the player bets 1 unit on red, if red comes out, it stops playing and won 1 unit (2 units less gain setting unit), if black comes out, he doubles his betting by 2 units on red and so on until he wins.

Figure 5.99: **Casino roulette wheel**

Having a chance on two to win, he may think he will eventually win, and when he wins, he is necessarily paid for everything he has played more one unit of his initial bet.

This martingale appears to be safe in practice. Note that in theory, to be sure of winning, we should have the opportunity to play an unlimited number of times.... This has major drawbacks:

This martingale is in fact limited by the bets that the player can do because you have to double the bet every time you lose: 2 times the initial bet, then 4, 8, 16 .... if he loses 10 times, he must be able to bet 1024 times its initial investment for the 11th party! Therefore a lot of money for little gain!

The roulette wheels also have a "0" which is neither red nor black. The risk of losing at every shot is then larger than $1/2$...

In addition, to paralyse this strategy, casinos offer table games per set: from 1 to 100.-, from 2 to 200.-, from 5 to 500.-, ... Therefore it is impossible to use this method on a large number of shots, which increases the risk of losing it all.

Blackjack is a game that has winning strategies: several playing techniques, which usually require to memorize the cards, can overturn the chances in favour of the player. The mathematician Edward Thorp has published in 11962 (holocene calendar) a book that was at the time a real best-seller. But all these methods require long training weeks and are easily detectable by the croupier (sudden changes in the amounts of bets are typical). The casino has then the opportunity to banish from its establishment the players using this playing martingale.

It should be noted that there are enough advanced methods. One of them is based on the less played combinations. In games where the gain depends on the number of winning players (Lotto...), playing the least played combinations maximize gains. This is how some people sell combinations that would be statistically very rarely used by other players.

Based on this reasoning, we can still conclude that a player who would have been able to determine statistically the least played combinations, to maximize its expected payoff, will in fact certainly not be the only player to have achieved this by the analysis of these famous combinations! This means that in theory the numbers the least played are actually overplayed combinations, the best might be to achieve a mix of played numbers and overplayed numbers to play for the ideal combinations. Another conclusion to all this is maybe that the best is to play random combinations which ultimately are less likely to be chosen by the players who incorporate a human and harmonious factor in the choice of their numbers.

5.6.5.1 Gambling

At its core the business of casino gaming is pretty simple. Casinos make money on their games because of the mathematics behind the games. As Nico Zographos, dealer-extraordinaire for the 'Greek Syndicate' in Deauville, Cannes, and Monte Carlo in the 11920s (holocene calendar) observed about casino gaming: «*There is no such thing as luck. It is all mathematics.*»

With a few notable exceptions, the house always wins in the long run[30] because of the mathematical advantage the casino enjoys over the player.

A common question about gampling is: How does house always win in gambling, even if the chances are $50/50$?

The answer relies in the fact that when a player lose the hand, they take his bet. If you win the hand, they only give you back your bet plus 95% of his bet. In order to calculate winnings we need to understand the concept of expected value.

In short, expected value is the expected payout from a particular probabilistic event. In simple case, we can calculate it by summing all the possible probabilities p_i multiplied by their payouts Pay-out$_i$. In this example if the player bets \$$1$ every game, he will gain \$$0.95$ when he wins and lost \$$1$ when he lose. Ignoring the possibility of a tie this comes out to!

The expected payout value is therefore:

$$\text{E(Pay-out)} = \sum_i p_i \text{Pay-out}_i = 0.5 \cdot (-1) + 0.5 \cdot (0.95) = -0.025 \qquad (5.6.203)$$

This means that on any given hand, the player will expect to lose about 2 or 3 cents. Really, he is actually gaining \$$0.95$ or losing \$$1$ but on average if he plays forever, it will come out to \$$-0.025$ per hand and he will slowly lose money. Even though the chances of winning any given hand are $50/50$.

In other words, by the law of large numbers (see further below page 920) says that as the number of bets placed goes to infinity, the average profit or loss realized by a player per bet will converge to the expected value. Since the expected value is negative, that means that the infinitely busy gambler will lose infinitely much money (or more realistically, go bankrupt). Or from the another point of view, the house always wins because the casino makes a LARGE number of bets.

---

[30] Without this underlined term, the sentence makes no sense. The devil is in the details!

The house advantage - the all-important percentage that explains how casinos make money - is also named the "house edge", the "theoretical win percentage", and "expected win percentage". In double-zero roulette, this figure is 5.3%. In the long run the house will retain 5.3% of the money wagered. In the short term, of course, the actual win percentage will differ from the theoretical win percentage (the magnitude of this deviation can be predicted from statistical theory). The actual win percentage is just the (actual) win divided by the handle. Because of the law of large numbers - or as some prefer to name it, the "law of averages" - as the number of trials gets larger, the actual win percentage should get closer to the theoretical win percentage.

Remark

In an effort to entice players and increase business, casinos occasionally offer novel wagers, side bets, increased payoffs, or rule variations. These promotions have the effect of lowering the house advantage and the effective price of the game for the player. This is sound reasoning from a marketing standpoint, but can be disastrous for the casino if care is not taken to ensure the math behind the promotion is sound (the are some famous examples of such disasters with a Illinois riverboat Casino, also an Indian Casino in California and some Casinos in Las Vegas, etc.).

The two tables below show the house advantages for many of the popular casino games:

Popular Casino Games (under optimal strategies)	**House Advantage**
Roulette (double-zero)	5.3%
Craps (pass/come)	1.4%
Craps (pass/come with double odds)	0.6%
Blackjack - average player	2.0%
Blackjack - 6 decks, basic strategy	0.5%
Blackjack - single deck, basic strategy	0.0%
Baccarat (no tie bets)	1.2%
Caribbean Stud	5.2%
Let It Ride	3.5%
Three Card Poker	3.4%
Pai Gow Poker (ante/play)	2.5%
Slots	$5-10\%$
Video Poker	$0.5-3\%$
Keno (average)	27%

Let us prove for example where the value 5.3% for the Roulette above comes from!

The house edge is the percentage of each bet that the casino expects to win over the long run. The easiest way to calculate this number is to assume that someone is playing for \$100 a bet, and determining how much he'll lose on average over the long run, then divide that into the number of bets.

I'll use roulette as an example again, because it's illustrative.

You place 38 bets on a single number at the roulette table. If you had a mathematically perfect set of spins (which is what you want to assume when you're calculating the house edge), you'll win once and lose 37 times.

The winning bet pays off at 35 to 1, so you'll win \$$3500$ on it. But you'll lose \$$3700$ on the 37 losing spins. Your net loss is \$$3700-$ \$$3500$, or \$$200$.

If you divide \$$200$ by 38, you get \$$5.26$ per spin, which is your average loss per spin.

That is the house edge for the game -5.26%.

Caution!!!!!!!!! It's important to remember that the house edge is a long term mathematical expectation. In the short term, anything can (and often will) happen.

We will stop here about the subject of gambling because it's not the most ethical business and there are already plenty of graduate level textbooks dedicated to that topic.

5.6.6 Combinatorial Analysis

"Combinatorial analysis" (counting techniques) is the field of mathematics that deals with the study of all the issues, events or facts (distinguishable or indistinguishable) with their arrangements (combinations) ordered or not according to some given constraints.

Definitions (#5.6.107):

D1. A sequence of objects (events, issues, objects, ...) is said "ordered" if each suite with a particular order of objects is recognized as a particular configuration.

D2. A sequel is "unordered" if and only if we are interested in the frequency of appearance of objects regardless of their order.

D3. The objects (of a sequence) are said "distincts" if their characteristics can not to be confused with the other objects.

Remark

We chose to put combinatorial analysis in this chapter because when we calculate probabilities, we also often need to know what is the probability of finding a combination or arrangement of given events under given constraints. Also we have decided (as always!) to focus on combinatorial mathematics that have important practical applications in others sections of this textbook!

Students often have difficulty remembering the difference between a permutation, an arrangement and a combination. Here is a little summary of what we'll see:

- Permutation: We take all the objects.
- Arrangement: We choose objects from the original set and the order intervenes.
- Combination: Same as for the arrangement, but the order does not intervenes.

You must not forget that for each result, the reverse will give the probability of falling respectively on a given permutation/arrangement/combination!

We will present and demonstrate below the 6 most common cases from which we can find (usually) all others:

5.6.6.1 Simple Arrangements with Repetitions

Definition (#5.6.108): A "simple arrangement with repetition" or equivalently a "random sampling with replacement"" (RSWR) is an ordered sequence of length m of n distinct objects not necessarily all different in the sequence (either: with possible repetitions!).

Let A and B be two finite sets of respective cardinal m, n such that there is trivially m ways to choose an object in A (of type a) and n ways to choose an object in B (of type b).

We saw in the section Set Theory (page 388) that if A and B are disjoint, that:

$$\text{Card}(A \cup B) = \text{Card}(A) + \text{Card}(B) = m + n \tag{5.6.204}$$

We therefore deduce the following properties:

P1. If an object can not be at the same time of type a and type b and if there is m ways to select an object of type a and n ways to choose an object of type b, then the union of objects gives $m + n$ selections (this is typically the result of the SQL UNION queries without filters in corporate Relational Databases Management System).

P2. If we can choose an object type of type a in m ways then an object of type b in n ways, then there is according to the Cartesian product of two sets (see section Set Theory page 391):

$$\text{Card}(A \times B) = \text{Card}(A) \cdot \text{Card}(B) = m \cdot n \tag{5.6.205}$$

ways to choose a single object of type a then an object of type b.

With the same notation for m and n, we can choose for each element of A, its single image among the n elements of B. So there are n ways to choose the image of the first element of A, then also n ways to choose the image of the second element of A, ..., and n ways to choose the image of the m-th element of A. The total number of consecutive possible applications from A to B is thus equal to the m product of n (thus m times the cartesian product of the cardinality of the set B with itself!). It is usual to write it under the following way (we have indicated the different ways to write his result as it can be found in various textbooks):

$$\boxed{\text{Card}(B^A)=\text{Card}\left(\underbrace{B\times B\times\ldots\times B}_{m\text{ times}}\right)=\text{Card}(B)^m=\bar{A}_n^m=n^m} \quad (5.6.206)$$

where B^A is the set of applications from A to B. The increase in the number of possibilities is geometric (not "exponential" as it is often wrongly said!).

This result is mathematically similar to the ordered result (an arrangement where the order of elements in the sequence is taken into account) of m trials in a bag containing n different balls with replacement after each trial. In France this result is traditionally named a "p-list".

✍**Examples:**

E1. How many (ordered) "words" of 7 letters can we form from a separate alphabet of 24 letters (very useful to know the number of trials to find a password for example)? The solution is:

$$\bar{A}_{24}^7=24^7=4,586,471,424 \quad (5.6.207)$$

Such a huge number is not very convenient to indicate the strength of a password. That's why IT professional prefer to use a measure based on entropy as we will see in the section of Statistical Mechanics (see page 3530).

E2. How many groups of people will we have in a referendum on 5 subjects and where each can be either accepted or rejected? The solution is (widely used in some Swiss companies):

$$\bar{A}_2^5=2^5=32 \quad (5.6.208)$$

A simple generalization of this result can consist of the following problem statement:

If we have m such objects $k_1,k_2,\ldots,k_m$ as k_i may take n_i different values then the number of possible combinations is:

$$\boxed{A_{mn}=n_1n_2\ldots n_i\ldots n_m} \quad (5.6.209)$$

And if $n_1=n_2=\ldots=n_m$ then we fall back on:

$$A_{mn}=A_n^m=n_1n_2\ldots n_i\ldots n_m=n^m \quad (5.6.210)$$

5.6.6.2 Simple Permutations without Repetitions

Definition (#5.6.109): A "simple permutation without repetitions" (formerly named "substitution") of n distinct objects is an ordered (different) sequence of these n objects all different by definition in the sequence (without repetition).

Remark

Be careful not to confuse the concept of permutation (n elements between them) and this of arrangement (of n elements among m)!

The number of permutations of n items can be calculated by induction: there are n places for a first element, $n-1$ for a second element, ..., and there will be only one place for the last remaining element.

It is therefore trivial that we the number of permutations is given by:

$$n(n-1)(n-2)(n-3)\ldots(n-(n-1)) \tag{5.6.211}$$

Recall that the product:

$$n(n-1)(n-2)(n-3)\ldots(n-(n-1))=\prod_{i=1}^{n} i \tag{5.6.212}$$

is named the "factorial of n" and we note it $n!$ for $n \in \mathbb{N}$.

There is therefore for n distinguishable elements:

$$\boxed{A_n=\prod_{i=1}^{n} i=n!} \tag{5.6.213}$$

as possible permutations. This type of calculation can be useful for example in project management (calculation of the number of different ways to get in a production line n different parts ordered from external suppliers).

✍**Example:**

How many (ordered) "words" of 7 different letters without repetition can we create?

$$A_n=7!=5040 \tag{5.6.214}$$

This result leads us to assimilate it to the ordered results (an arrangement A_n in which the order of elements in the sequence is taken into account) of the trial of balls that are all different from a bag containing n distinguishable balls without replacement.

The huge values that can bring the factorial may be quite surprising or even sometimes counter-intuitive (consider for example the number of ways you can create groups of just say...7 friends!). By the way it also question if the Universe is computable or not as such calculations leads to huge numbers!

5.6.6.3 Simple Permutations with Repetitions

Definition (#5.6.110): A "simple permutation with repetition" is when we consider the number of ordered permutations (different) of a sequence of n distinct objects not necessarily all different in a given quantity.

Remark

Do not confuse this definition with the "simple arrangement with repetition" seen before!

When some elements are not all distinguishable in a sequence of objects (they are repeated in the sequence), then the number of permutations that we can be do are then trivially reduced to a smaller number then if all the elements were all distinguishable.

Consider n_i as the number of objects of the type i, with:

$$n_1 + n_2 + \ldots + n_k = n \tag{5.6.215}$$

then we write:

$$\bar{A}_n(n_1, \ldots, n_i, \ldots, n_k) \tag{5.6.216}$$

the number of possible permutations (yet unknown) with repetition (one or more elements in a sequence of repetitive elements are not distinguishable by permutation).

If each of the n_i positions occupied by identical elements were occupied by different elements, the number of permutations could then have to be multiplied by each of the $n_i!$ (previous case).

It comes that we fall back on the factorial such as:

$$\bar{A}_n(n_1, n_2, \ldots, n_k)n_1!n_2!\ldots n_k! = n! \tag{5.6.217}$$

from what we deduce immediately:

$$\boxed{\bar{A}_n(n_1, n_2, \ldots, n_k) = \frac{n!}{n_1!n_2!\ldots n_k!}} \tag{5.6.218}$$

If the n objects are all different in the sequence, we then have:

$$n_1! = n_2! = \ldots = n_k! = 1! = 1 \tag{5.6.219}$$

and we fall back again on a simple permutation (without repetition) as:

$$\bar{A}_n(n_1, n_2, \ldots, n_k) = \frac{n!}{n_1!n_2!\ldots n_k!} = \frac{n!}{1!1!\ldots 1!} = n! \tag{5.6.220}$$

It should therefore be remembered that the permutations with repetition are in smaller quantity than those without repetition (obvious since we do not take into account the permutations of identical elements between them!).

✍**Example:**

How many (ordered) "words" can we generate with the letters of the word "Mississippi":

$$\bar{A}_4(\underbrace{1}_{\text{M}},\underbrace{2}_{\text{PP}},\underbrace{4}_{\text{IIII}},\underbrace{4}_{\text{SSSS}})=\frac{11!}{1!2!4!4!}=34,650 \qquad (5.6.221)$$

This result leads us to assimilate it to an ordered result (a permutation $\bar{A}_n$ where the order of elements in the sequence is not taken into account) of the trial of n balls that are not all different from a bag containing $k \geq n$ balls with limited replacement for each ball.

5.6.6.4 Simple Arrangements without Repetitions

Definition (#5.6.111): A "simple arrangement without repetition" or equivalently a "random sampling without replacement" (RSWOR) is an ordered sequence of p objects all distinct taken from n distinct objects with $n \geq p$.

We now propose to enumerate the possible arrangements of n objects among p without repetition. We denote A_n^p the number of these arrangements.

It is easy to calculate that $A_n^1=n$ and to check that $A_n^2=n(n-1)$. Indeed, there are n ways to choose the first object and $(n-1)$ ways to choose the second when we already have the first.

To determine a nice expression for A_n^p , we reason by induction. We assume A_n^{p-1} known and we deduce that:

$$A_n^p=A_n^{p-1}\left[n-(p-1)\right]=A_n^{p-1}(n-p+1) \qquad (5.6.222)$$

It comes:

$$A_n^p=n(n-1)(n-2)(n-3)\ldots(n-(p-1)) \qquad (5.6.223)$$

then:

$$n!=A_n^p(n-p)!=\left[n(n-1)(n-2)\ldots(n-p+2)(n-p+1)\right](n-p)(n-p-1)\ldots \qquad (5.6.224)$$

whence:

$$\boxed{A_n^p=\frac{n!}{(n-p)!}} \qquad (5.6.225)$$

This result leads us to assimilate it to the ordered results (an arrangement A_n^p in which the order of elements in the sequence is taken into account) of the trial of p distinct balls from a bag containing n different balls without replacement.

✍**Example:**

Consider the 24 letters of the alphabet, how many (ordered) "words" of 7 distinct letters can we generate?

$$A_{24}^{7}=\frac{24!}{(24-7)!}=1,744,364,160 \tag{5.6.226}$$

The reader may have noticed that if $p=n$ we end up with:

$$A_{n}^{p}=\frac{n!}{0!}=n! \tag{5.6.227}$$

So we can say that a simple permutation of n elements without repetition is like a simple arrangement without repetition when $n=p$.

5.6.6.5 Simple Combinations without Repetitions

Definition (#5.6.112): A "simple combination without repetitions" or "choice function" is a non-ordered sequence (where the order doesn't interest us!) of p elements all different (not necessarily in the visual sense of the word!) selected from n distinct objects and is by definition denoted C_p^n in this textbook and named the "binomial" or "binomial coefficient".

If we permute the elements of each simple arrangement of elements p of n, we get all simple permutations and we know that there are in a number of $p!$, using the notation convention of this textbook we then have (contrary to that recommended one by ISO 31-11!):

$$\boxed{C_p^n=\binom{n}{p}=\frac{A_n^p}{p!}=\frac{n!}{p!(n-p)!}} \tag{5.6.228}$$

It is a relation often used in gambling but also in the industry trough the hypergeometric distribution (see section Industrial Engineering page 8677).

A simple way to remember this function is the following trick: Consider we must select p (3 individuals) among n (6 individuals) independently of the order what are the number of possibilities?

We know that we have $6\cdot 5\cdot 4=120$ possibilities to select them all taking into account the order! The calculation we just made is obviously equal to $n!/p!=6!/3!=6\cdot 5\cdot 4$. But as the order must not be taken into account we must divide the 120 by the number of ways we can arrange the 3 people in the group. So we divide 120 by 3! or more generally and logically by $(n-p)!$. Hence the relation above!

Remarks

R1. We have necessarily by construction $C_p^n\leq A_n^p$.

R2. Depending on the authors we inverse the index and suffix of C then you must be careful!

This result leads us to assimilate it to the unordered result (an arrangement C_n^p in which the order of elements of the sequence is not taken into account) of the trial of p balls of a bag containing n different balls without replacement.

✍**Examples:**

E1. Consider the 24 letters of the alphabet, how many choices do we have to take 7 letters in the 24 without taking into account the trial order?

$$C_7^{24} = \frac{24!}{7!(24-7)!} = 346,104 \tag{5.6.229}$$

The same value can be obtained with the function `COMBIN( )` of Microsoft® Excel® 11.8346 (English version).

E2. In a Design of Experiment (see section Industrial Engineering page 8780) we have 2 factors of $L = 3$ levels each and therefore we need $N = 9$ runs to completely determine all the interactions. If we consider that we can take a subset of $S = 3$ runs, how many combinations of 3 among the 9 can we choose if repetitions are forbidden?

$$C_3^9 = \frac{9!}{3!(9-3)!} = 84 \tag{5.6.230}$$

We understand therefore why in Design of Experiments it is important to find a trick to choose the best subset (D-optimum designs).

There is, in relation to the binomial coefficient, another relation very often used in many case studies and also more globally in physics or functional analysis. This is the "Pascal's Formula":

Proof 5.39.2.

$$\begin{aligned} C_{p-1}^{n-1} + C_p^{n-1} &= \frac{(n-1)!}{((n-1)-(p-1))!(p-1)!} + \frac{(n-1)!}{(n-1-p)!p!} \\ &= \frac{(n-1)!}{(n-p)!(p-1)!} + \frac{(n-1)!}{(n-1-p)!p!} \end{aligned} \tag{5.6.231}$$

We also have $p! = p(p-1)!$, then:

$$(p-1)! = \frac{p!}{p} \tag{5.6.232}$$

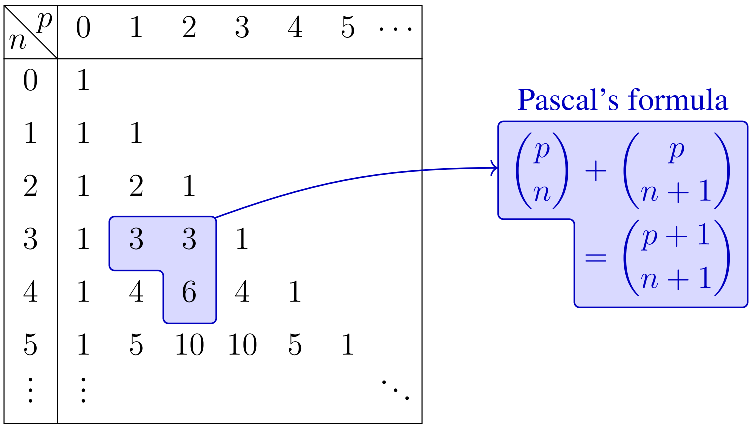
and because $(n-p)(n-p-1)! = (n-p)!$:

$$(n-p-1)! = \frac{(n-p)!}{n-p} \tag{5.6.233}$$

Then:

$$\begin{aligned} C_{p-1}^{n-1} + C_p^{n-1} &= \frac{(n-1)!}{(n-p)!(p-1)!} + \frac{(n-1)!}{(n-1-p)!p!} = \frac{(n-1)!p}{(n-p)!p!} + \frac{(n-1)!(n-p)}{(n-p)!p!} \\ &= \frac{(n-1)!}{(n-p)!p!}[p+(n-p)] = \frac{(n-1)!n}{(n-p)!p!} = \frac{n!}{(n-p)!p!} = C_p^n \end{aligned} \tag{5.6.234}$$

■ Q.E.D.



$n \backslash p$	0	1	2	3	4	5	$\cdots$
0	1						
1	1	1					
2	1	2	1				
3	1	3	3	1			
4	1	4	6	4	1		
5	1	5	10	10	5	1	
$\vdots$	$\vdots$						$\ddots$

Pascal's formula

$$\binom{p}{n}+\binom{p}{n+1}=\binom{p+1}{n+1}$$

Figure 5.100: **Pascal's formula illustrated**

Another very common relation that will be useful to us is the "binomial coefficient identity", given by:

$$\begin{aligned}\binom{n}{a}\binom{n-a}{b-a}&=\frac{n!}{(n-a)!a!}\frac{(n-a)!}{(n-a-(b-a))!(b-a)!}\\&=\frac{n!}{a!}\frac{1}{(n-b)!(b-a)!}\\&=\frac{n!}{(n-b)!b!}\frac{b!}{(b-a)!a!}\\&=\binom{n}{b}\binom{b}{a}\end{aligned}\qquad(5.6.235)$$

where the third equality follows by multiplying and dividing by $b!$.

5.6.6.6 Simple Combinations with Repetitions

Definition (#5.6.113): A "simple combination with repetition" of p elements of n is a collection of p non-ordered elements, not necessarily distinct.

Simple combinations with repetition are very important for the Wald-Wolfowitz statistical test used in economics and biology and that we will study in the Statistics section.

We will introduce this kind of combination directly with an example an ingenious approach that we have thanks to the physicist and 11938 (holocene calendar) Nobel Prize in Physics: Enrico Fermi.

Consider $\{a,b,c,d,e,f\}$ a set having a number n of elements equal to 6 and where we draw a number of elements p equal to 8. We would like to calculate the number of combinations with repetition of elements in a starting set of cardinal 6 in a destination set of cardinal 8.

Consider, for example, the following three combinations:

$$\begin{gathered} a\ a\ b\ b\ b\ e\ e\ f \\ b\ b\ d\ d\ d\ e\ e\ e \\ b\ b\ b\ d\ d\ d\ d\ d \end{gathered} \tag{5.6.236}$$

where as the order of elements does not occur, we have grouped the elements to facilitate the reading. Now represent all the above elements by the same symbol "0" and separate the groups consisting of a single element by bars (this is the trick Enrico Fermi). Thus, when one or more elements are not included in a combination, we still denote the separation bars (corresponding to the number of missing elements + the separation of group). Thus, the three combinations above can be written as:

$$\begin{gathered} 0\,0\ |\ 0\,0\,0\ |||\ 0\,0\ |\ 0 \\ |\,0\,0\ ||\ 0\,0\,0\ |\ 0\,0\,0\ | \\ |\,0\,0\,0\,0\ ||\ 0\,0\,0\,0\ || \end{gathered} \tag{5.6.237}$$

We see above that in each case, there are eight "0" (logic. ..) but also that there are also always five "|". The number of combinations with repetitions of six elements of a starting set to the final one of 8 elements is equal to the number of permutations with repetitions of $8+5=13$ elements, so:

$$\frac{13!}{5!8!}=1287 \tag{5.6.238}$$

We also see that in the general case the number of combinations without consideration of repetitions order can also be written:

$$\frac{(n+p-1)!}{(n-1)!p!} \tag{5.6.239}$$

It is traditional to write this:

$$\Gamma_p^n=\frac{(n+p-1)!}{(n-1)!p!} \tag{5.6.240}$$

We also see that:

$$\Gamma_p^n=\frac{(n+p-1)!}{(n-1)!p!}=\frac{(n+p-1)!}{((n+p-1)-p)!p!}=C_p^{n+p-1} \tag{5.6.241}$$

Then:

$$\boxed{\Gamma_p^n=C_p^{n+p-1}=\frac{(n+p-1)!}{(n-1)!p!}} \tag{5.6.242}$$

That we also sometimes write:

$$\Gamma_p^n=C_p^{n+p-1}=C_{n-1}^{n+p-1}=\frac{(n+p-1)!}{(n-1)!p!} \tag{5.6.243}$$

✍**Example:**

How many different combinations of outcomes can we make by rolling three standard (6-sided) dice if the order of the dice does not matter?

The answer is:

$$\Gamma_p^n=C_p^{n+p-1}=\frac{(n+p-1)!}{(n-1)!p!}=\frac{(6+3-1)!}{(6-1)!3!}=56 \tag{5.6.244}$$

But if the order matter, then it was simply $6\cdot 6\cdot 6=216$.

To summarize:

Type	Expression
Simple arrangement with repetitions (denoted ${}^R V_n^m$ according to ISO 80000-2:2009)	$\bar{A}_n^m=n^m$
Simple arrangement without repetitions (denoted V_n^m according to ISO 80000-2:2009)	$A_m^n=\dfrac{n!}{(n-m)!}$
Simple permutation without repetitions	$A_n=n!$
Simple permutation with repetitions	$\bar{A}_n(n_1,n_2,\ldots,n_k)=\dfrac{n!}{n_1!n_2!\ldots n_k!}$
Simple combination without repetitions: case of the simple arrangement without repetitions where the order is not taken into account (denoted ${}^R C_n^p$ according to ISO 80000-2:2009)	$C_m^n=\dbinom{n}{p}=\dfrac{A_n^m}{m!}=\dfrac{n!}{m!(n-m)!}$
Simple combination with repetitions (case of the simple permutation with repetition where the order is not taken into account)	$\Gamma_p^n=C_p^{n+p-1}=\dfrac{(n+p-1)!}{(n-1)!p!}$

Table 5.27: **Resume of main Combinatorial Analysis cases**

Four of the relations above can also be presented as following:

	Order Matters	**Order doesn't Matters**
With Replacement	$\bar{A}_n^m=n^m$	$\Gamma_p^n=C_p^{n+p-1}=\dfrac{(n+p-1)!}{(n-1)!p!}$
Without Replacement	$A_m^n=\dfrac{n!}{(n-m)!}$	$C_m^n=\dbinom{n}{p}=\dfrac{A_n^m}{m!}=\dfrac{n!}{m!(n-m)!}$

Remark

If we consider three elements A, B and C the total cumulated number of combinations where the order matter is quite obviously (this is straightforward from the Binomial theorem proved at page 1364, where we put $a=1$ and $b=1$):

$$\begin{aligned}\sum_{p=0}^{n}\frac{n!}{p!}-1&=\frac{3!}{0!}+\frac{3!}{1!}+\frac{3!}{2!}+\frac{3!}{3!}-1\\&=\{A,B,C,AB,AC,BC,BA,CA,CB,ABC,BCA,CBA,BAC,ACB,CAB\}\\&=(6+6+3+1)-1=15\end{aligned}\tag{5.6.245}$$

But if the order doesn't matter, we have:

$$\begin{aligned}\sum_{p=1}^{n}\binom{n}{p}=\sum_{p=0}^{n}\binom{n}{p}-1=\binom{n}{0}+\binom{n}{1}+\ldots+\binom{n}{n}-1\\=\{A,B,C,AB,BC,CA,ABC\}=(1+3+3+1)-1=7=2^n-1\end{aligned}\tag{5.6.246}$$

We also have the number of ways n different items can be arranged into n bags with exactly n objects each (we can reuse any objective, however no repetition is allowed inside the bags), that is given by:

$$(n!)^n\tag{5.6.247}$$

and the number of ways in which $m\cdot n$ different items can be divided equally into m groups each containing n objects and where order of the groups matters that is given by:

$$\frac{(mn)!}{(n!)^m}\tag{5.6.248}$$

and the number of ways in which $m\cdot n$ different items can be divided equally into m groups, each containing n objects and where the order of the groups is doesn't matter is:

$$\frac{(mn)!}{(n!)^m m!}\tag{5.6.249}$$

The both tables above can also be summarized and illustrated as following:

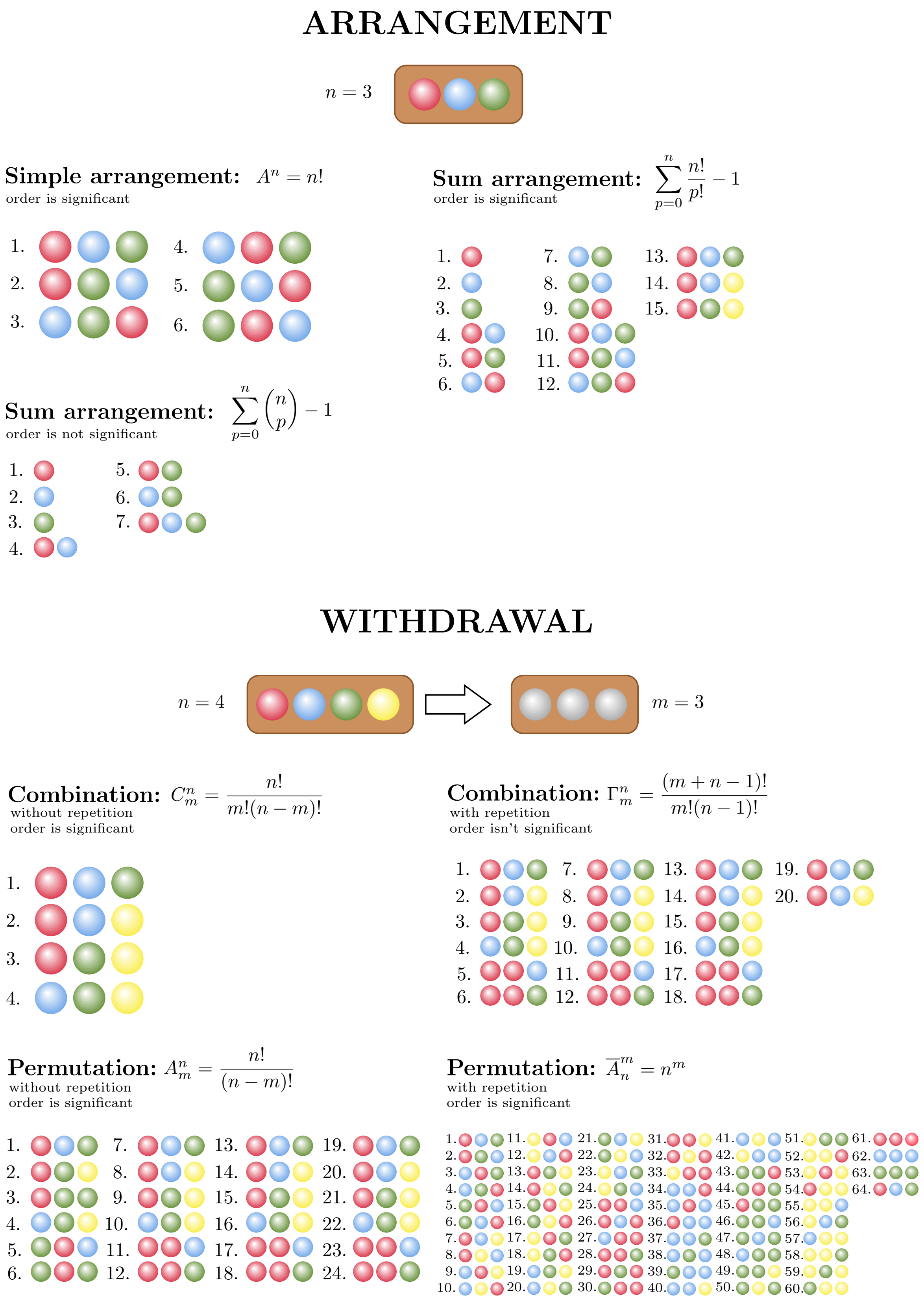
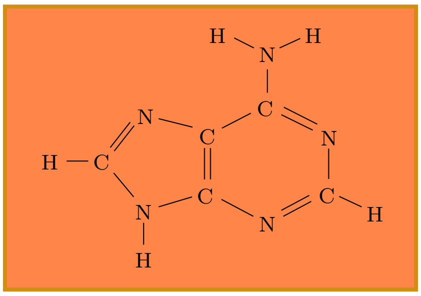

ARRANGEMENT
$n=3$
Simple arrangement: $A^n = n!$
order is significant
Sum arrangement: $\sum_{p=0}^{n} \frac{n!}{p!} - 1$
order is significant
Sum arrangement: $\sum_{p=0}^{n} \binom{n}{p} - 1$
order is not significant
WITHDRAWAL
$n=4$
$m=3$
Combination: $C_m^n = \frac{n!}{m!(n-m)!}$
without repetition
order is significant
Combination: $\Gamma_m^n = \frac{(m+n-1)!}{m!(n-1)!}$
with repetition
order isn't significant
Permutation: $A_m^n = \frac{n!}{(n-m)!}$
without repetition
order is significant
Permutation: $\overline{A}_n^m = n^m$
with repetition
order is significant

Figure 5.101: **Combinations, permutations and arrangements simple summary**

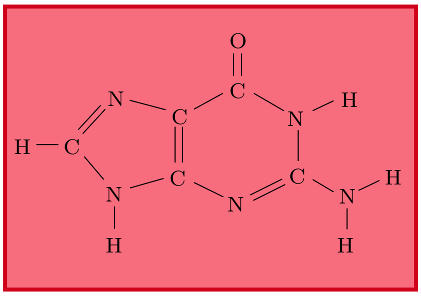
You're better in theory than in practice

To close this study on the basics of combinatorial analysis, there would be another important case but that seems unsolved still in this beginning of the 121st century (holocene calendar) and related to chemistry (some experts speak about "combinatorial chemistry"). Indeed, given a number of pure chemical elements with their respective valence in some thermodynamics conditions, how many molecules involving one of all of them can be formed?

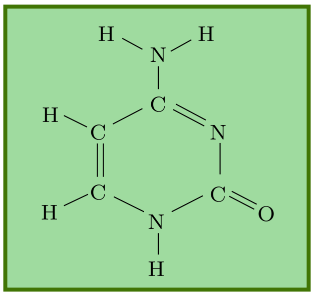
For example, with the Adenine (one of the four fundamentals components of DNA) given by $C_5N_5H_5$ and visible in the figure below:



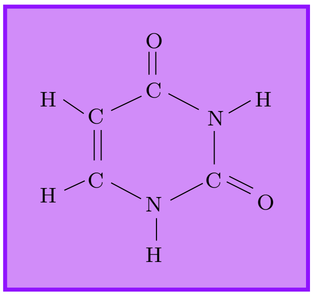
Adenine



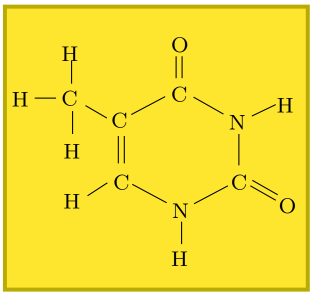
Guanine



Cytosine



Uracil



Thymine

We then could ask ourselves what is total number of combinations involving $5\mathrm{C}$, $5\mathrm{N}$ and $5\mathrm{H}$?

An upper bound to this question would be given (based on their respective valence by):

$$4^5 \cdot 3^5 \cdot 2^5 = 7,962,624 \qquad (5.6.250)$$

But if we draw all physical possible combination in standard temperature and pressure conditions, it seems that an upper limit is rather between 200 and $2,000$ different molecules.

So this is still and important "open problem" in chemistry with huge application possibilities (some industries actually use computers to find feasible combinations of pure chemical elements that may have applications in medicine!).

5.6.7 Markov Chains

Markov chains are simple but powerful probabilistic and statistical tools but for which the choice of the mathematical presentation can sometimes be a nightmare... We will try here to simplify a maximum the writings to introduce this great tool widely used in businesses to manage supply chain, in queuing theory for call centers or cash desk, in failure theory for preventive maintenance, statistical physics and biological engineering and also in time series analysis and forecasting (and the list goes on and for more details the reader should refer to the relevant chapters available in this textbook...).

Definitions (#5.6.114):

D1. We note by $\{X(t)\}_{t\in T}$ a probabilistic process function of time whose value at any time depends on the outcome of a random experiment. Thus, at each time t, $X(t)$ is a random variable that we name "stochastic process" (for more details on financial applications see the chapter Economy).

D2. If we consider a discrete time, we then note "discrete time stochastic process" as $\{X_n\}_{n\in\mathbb{N}}$.

D3. If we further assume that the random variables X_n can take only a discrete set of values we then speak of "process in discrete time and discrete space".

Remark

It is quite possible as in the queuing theory (see section Management Techniques page 8141) to have a continuous time process with discrete state space.

Definition (#5.6.115): $\{X_n\}_{n\in\mathbb{N}}$ is a "Markov chain" if and only if:

$$P(X_n=j\mid X_{n-1}=i_{n-2},X_{n-2}=i_{n-2},\ldots,X_0=i_0)=P(X_n=j\mid X_{n-1}=i_{n-1})\qquad(5.6.251)$$

in other words (it is very easy!) the probability that the chain is in a certain state on the n-th step of the process depends only on the state of the process at step $n-1$ and not on any previous steps!

Remark

Also in probabilities a stochastic process verifies the Markov property if and only if the conditional probability distribution of future states, given the present moment, depends only on the present state and not even past states as the relation above. A process with this property is also named a "Markov process".

Definition (#5.6.116): A "homogeneous Markov chain" is a chain such that the probability that it has to go in a certain state at the n-th stage is independent of time. In other words, the probability distribution characterizing the next step does not depend on time (of the previous step), at all times the probability distribution of the chain is always the same for characterizing the transition to the current step.

We can then define (reduce) the law of "probability transition" of a state i to state j by:

$$p_{ij}=P(X_n=j \mid X_{n-1}=i) \tag{5.6.252}$$

It is then natural to define the "transition matrix" or "stochastic matrix":

$$M=\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mn} \end{bmatrix} \tag{5.6.253}$$

as the matrix that contains all possible transition probabilities between states in an oriented graph.

Markov chains can be represented graphically as an oriented graph G (see section Graph Theory page 3012) sometimes named "automata" having for the top points (states) i and for the edges the oriented couples (i,j). We then associate to each component an oriented arc and a transition probability.

✍**Example:**

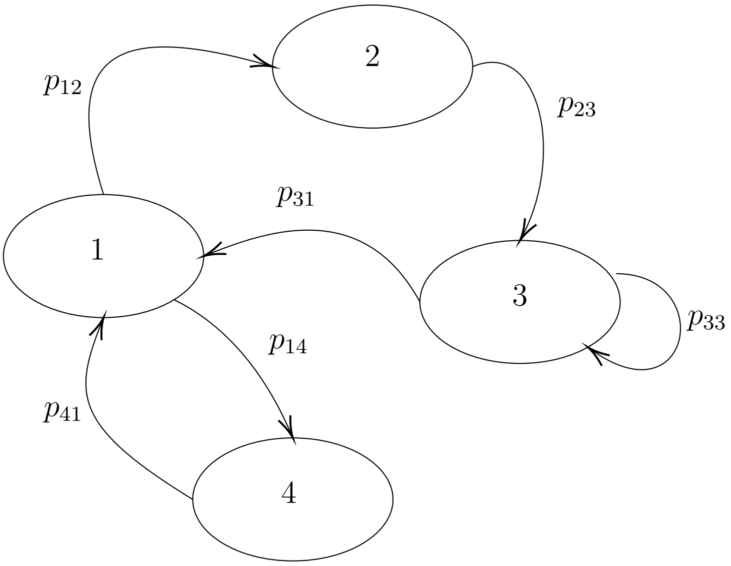


Figure 5.102: Generic example of a Markov chain

Thus, in the example of the oriented graph above, the only allowed transitions from the above 4 states (4×4 matrix) are those indicated by the arrows. So that the transition matrix is then simplified to:

$$P=\begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{bmatrix}=\begin{bmatrix} 0 & p_{12} & 0 & p_{14} \\ 0 & 0 & p_{23} & 0 \\ p_{31} & 0 & p_{33} & 0 \\ p_{41} & 0 & 0 & 0 \end{bmatrix} \tag{5.6.254}$$

The reader has seen in the previous example that we have the trivial property (by construction!) that the sum of the terms (probabilities) of a row of the matrix P is always unitary (and therefore the sum of the terms of a column of the transpose of the matrix unit is still equal to the unit too):

$$\sum_j p_{ij} = 1 \tag{5.6.255}$$

and that the matrix is positive (meaning that all its terms are non-negative).

Remark

Remember that the sum of the probabilities of the columns is always equal to 1 for the transpose of the stochastic matrix!

The analysis of transient state (or: random walk) of a Markov chain consist to determine (or to impose!) to the column-matrix (vector) $p(n)$ to be in a given state at n-th step of the walk:

$$p(n) = \begin{pmatrix} p_1(n) \\ p_2(n) \\ \vdots \\ p_{\text{Card}(E)}(n) \end{pmatrix} \tag{5.6.256}$$

with the sum of the components that is always equal 1 (since the sum of the probabilities of being in any of the vertices of the graph at given a time/step must be equal to 100%).

We frequently name this column matrix "stochastic vector" or "probability measure on the vertex i".

Theorem 5.40. We want to prove that the total probability of this stochastic vector is always equal to 1.

Proof 5.40.1. If $p(n)$ is a stochastic vector, then its image:

$$p(n+1) = P^T p(n) \tag{5.6.257}$$

is also a stochastic vector. Effectively, $p_i(n+1) \geq 0$ because:

$$p_i(n+1) = \sum_i p_{ij} p_j(n) \tag{5.6.258}$$

is a sum of positive or zero values. Furthermore, we find:

$$\begin{aligned} \sum p_i(n+1) &= \sum_i \sum_j p_{ij} p_j(n) = \sum_j \sum_i p_{ij} p_j(n) = \sum_j \left(\sum_i p_{ij} \right) p_j(n) \\ &= \sum_j 1 \cdot p_j(n) = \sum_j p_j(n) = 1 \end{aligned} \tag{5.6.259}$$

■ Q.E.D.

This probability vector whose components are positive or zero, depends (it's pretty intuitive) on the transition matrix P and the vector of initial probabilities $p(0)$.

Although even if it is provable (Perron-Frobenius theorem), the reader may verify by a practical case (computerized or not!) that if we choose any vector state $p(n)$ then there exists for any stochastic matrix P a unique probability vector traditionally noted π as:

$$P^T\pi = \pi \tag{5.6.260}$$

Such a probability measure π satisfying the above equation is named an "invariant measure" or "stationary measure" or "balanced measure" which represents the equilibrium state of the system. In terms of linear algebra (see section of the same name page 1924) for the eigenvalue 1, π is an eigenvector of P (see section Linear Algebra page 1997).

Iterations obtained before a Markov chain reaches the stationary measure are named "burn-in" or "warmup".

We will see a trivial example in the Graph Theory section (page 3026) which will be redeveloped in detailed as in the section of Game and Decision Theory in the context of pharmaco-economics (page 7456) and in the section of Software Engineering (page 8662) when we will study the fundamentals of the Google PageRank algorithm. But also note that the Markov chains are used for example in meteorology (or in the case of computer passwords hacks):

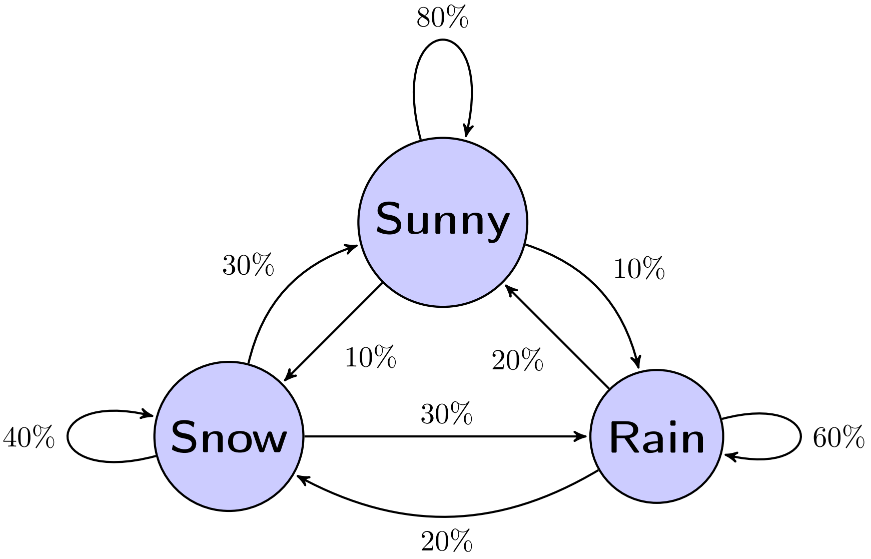


Figure 5.103: **Concrete example of a very simple Markov chain**

or in medicine, finance, transportation, marketing, etc.

Let us also point out a quite trivial equality to which the mathematician would have given the name of "Chapman-Kolmogorov equation". It is simply the following equality obtained by recurrence (which saves a lot of time in terms of applied calculations):

$$\boxed{p(n) = P^T p(n-1) = P^T\left(P^T p(n-2)\right) = \ldots = \left(P^T\right)^n p(0)} \tag{5.6.261}$$

In the field of language analysis, from the frequency analysis of a sequence of words, computers are able to also build Markov chains and therefore propose a more correct semantic during grammatical computerized corrections or in written transcription of oral presentations.

Definitions (#5.6.117):

D1. A Markov chain is said to be an "irreducible Markov chain" if all states are bound to others (it's the case of the example in the figure above).

D2. A Markov chain is said to be an "absorbing Markov chain" if one of the states of the chain absorbs the transitions (so nothing comes out just to say things in a more simple way!).

~90 %

★★★★☆

27 votes, 51.11%

5.7. Statistics

STATISTICS is a science that concerns the systematic grouping of facts or recurring events that lend themselves to a numerical or qualitative assessment over time according to a given law. In the industry and the economy in general, statistics is a science that helps in an uncertain environment to make valid inferences. Fundamentally, statistics rarely concern themselves with proving anything true or false in experimental science, but rather with assessing how convincing competing models/hypotheses/effects are!

You should know that among all areas of mathematics, the one that is used the most widely in business and research centers is Statistics and especially since softwares greatly facilitates the calculations! This is why this section is one of the biggest in this textbook even if only the basic concepts are presented!

Note also that Statistics have a very bad reputation at the university because the notations are often confusing and vary greatly from one teacher to another, from one book to another, from one practitioner to another. Strictly speaking, it should comply with the vocabulary and notation of the ISO 3534-1:2006 *Statistics - Vocabulary and symbols* standard and unfortunately this section was written before the publication of this standard... a certain period of adaptation will be necessary to obtain the full compliance.

It is perhaps useless to precise that Statistic is widely used in engineering, robotics, theoretical physics, fundamental physics, econometrics, project management and in the industry of process, in the fields of life and non-life insurance, in the actuarial science or in the database analysis (with Microsoft® Excel® very often ... unfortunately ....) and the list goes on. We will also meet quite often the tools presented here in the sections of Fluid Mechanics, Thermodynamics, Quantum Field Theory, Technical Management, Industrial Engineering and Economy (especially in the last two). The reader can then refer to them to have some concrete practical applications of the most important theoretical elements that will be seen here.

Note also that in addition to a few simple examples on these pages, many other application examples are given on the exercise server of the companion website in the categories Probability and Statistics, Industrial Engineering, Econometrics and Management Techniques.

Definition (#5.7.118): The main purpose of Statistics is to determine the characteristics of a given population from the study of a part of the population, named "sample" or "representative sample". The determination of these characteristics should enable statistics to be a tool for the decision help!

Remark

The data processing concerns the "descriptive statistics" or also named the "descriptive analytics". The interpretation of data from estimators is named "statistical inference" (or "inferential statistics"), and mass data analysis "statistical frequency" as opposed to

Bayesian inference (see section Probabilities page 485).

When we observe an event taking into account some given factors, there can happen that a second observation takes place in conditions that seem identical. By repeating these steps several times on different supposedly similar objects, we find that the observed results are statistically distributed around a mean value that is ultimately the most likely possible outcome. In practice, however, we sometimes perform a single measurement and then the goal is determine the value of the error we make by adopting it as average measure. This determination requires knowledge of the type of statistical distribution we are dealing with and that is on what we will focus (among others) to study here (at least the basics!). However, there are several common methodological approaches when we face the probabilistic events (less common are not mentioned yet):

1. A first is to simply ignore the random elements, for the simple reason that we do not know how to integrate them. We then use the "scenarios method" also named "deterministic simulation". This is typically the tool used by financial managers and non-graduates managers with tools like Microsoft® Excel® (which includes a scenarios management tool) or Microsoft Project (which includes a tool to manage the deterministic optimistic, pessimistic and expected scenarios).

2. A second possible approach, when we do not know how to associate probabilities to specific future random events, is game theory (see section Game and Decision Theory page 7376) where semi-empirical criteria of selection are used as the criterion of maximax, minimax, Laplace, Savage, etc.

3. Finally, when we can link probabilities to random events, whether these probabilities derived from calculations or measurements, whether they are based on experience from previous similar situations as the current situation, we can use descriptive and inferential statistics (contents of this chapter) to obtain usable and relevant information from this mass of acquired datas.

4. A last approach when we know the relative probabilities from intervening events in response to strategic choices is the use of decision theory (see section Game and Decision Theory pages 7400 or 7426, 7429, 7449 and 7456).

STANDARD DEVIATION
SAMPLE
HYPOTHESIS
INFERENTIAL
PREDICTION
EXPERIMENT
CONFIDENCE INTERVAL
CORRELATION
DESCRIPTIVE
EMPIRICAL
P-VALUE
PERCENTAGE
CONTINUOUS
FREQUENCY
Z-SCORE
QUALITATIVE
ESTIMATE
ANALYSIS
DISCRETE
MODE
BIAS
TYPE I ERROR
RESIDUAL
RANGE
STATISTICS
DOUBLE-BLIND
MEAN
PROPORTION
SYMMETRICAL
TEST STATISTIC
RESEARCH
SIGNIFICANCE
ANOVA
CENTRAL LIMIT THEOREM
NORMAL
GRAPH
DATA
VARIABLE
OUTLIER
VARIANCE
STANDARDIZING
REGRESSION
PARAMETER
TYPE II ERROR
QUANTITATIVE
PROBABILITY
ALPHA
DISTRIBUTION
POPULATION
MARGIN OF ERROR
PERCENTILE
SAMPLE SPACE
MEDIAN

Remarks

R1. Without mathematical statistics, a calculation on datas (e.g. an average), is a "punctual indicator". This is mathematical statistics which gives it the status of estimator whose bias, uncertainty and other statistical characteristics are controlled. We generally seek to ensure that the estimator is unbiased, convergent and efficient (we will see during our study of estimators further what is exactly all that stuff).

R2. When we communicate a statistic it should be an obligation to specify the confidence interval, the p-value (see page 948) and the size of the studied sample (absolute statistics) and its detailed characteristics and make available the sources and data protocol otherwise it has almost no scientific value (we will see all these concepts in detail further below). A common mistake is to communicate in relative values. For example, on a test group of $1,000$ women, 5 women will die from breast cancer without screening check, with screening check 4 women will die. Some will say a little to quickly (typically physicians....) that screening checks saves 20% of women (relative value as one the of five could have been saved...). In fact this is wrong since the absolute benefit of screening is insignificant!

R3. If you have a teacher or trainer who dares to teach you statistics and probabilities only with examples based on gambling (cards, dice, matches, coins, etc.) dispose or denounce him. Normally examples should be based on the industry, economy or R&D, i.e. in areas used in daily by businesses!).

Let us introduce before continuing some definitions that will be useful for the future on the concept of samples and averages (we will come back on this basic definitions later in a more technical way!):

5.7.1 Samples

During the statistical analysis of sets of information, the way to select the sample is as important as how to analyse it.The sample must be representative of the population (we do not necessarily make reference to human populations!). For this, the random sampling is the best way to achieve it.

Definitions (#5.7.119):

D1. The statistician always starts from the observation of a finite number of elements, which we name the "population". The observed elements, in quantity n, are all of the same nature, but this nature can be very different from one population to another.

D2. We are in the presence of a "quantitative character" when each observed element is explicitly subject to the same measure. To a given quantitative character, we associate a "quantitative variable" continuous or discrete, which summarizes all the possible values that the measure can take (such information being represented by functions like the Gauss-Laplace distribution, the beta distribution, the Poisson distribution, etc.).

> **Remark**
>
> We will come back on the concept of "variable" and "distribution" a little further...

D3. We are in the presence of a "qualitative character" when each observed element is explicitly subject to a single connection to a "modality" from a set of exclusive modalities (e.g.: man | woman) that permits to classify all studied elements in a given certain point of view (such information being represented by bar charts, sector charts, bubble charts, etc.). All modalities of a character can be established a priori before the survey (a list, a nomenclature, a code) or after the survey. A study population can be represented by a mixed character, or set of modalities such as gender, wage range, age, number of children, marital status for example for an individual.

D4. A "random sample" is by default (without more precision) a sample in which all individuals in a population have the same chance, or "equally likely probability" (and we emphasize that this probability must be equal), to end up in the sample.

D5. In the opposite in a sample whose elements were not chosen randomly, then we talk about a "biased sample" (in the opposite case we talk about a "non-biased sample").

> **Remark**
>
> A small representative sample is by far preferable to a large biased sample. But when the sample sizes are small, the probabilities are so that sometimes we can have a result worst than the biased one...

5.7.2 Averages

The concept of "average" or "central tendency" (financial analysts name it a "measure of location"...) is with the notion of "variable" at the basis of statistics.

This notion seems very familiar to us and we talk a lot about it without asking too many questions. But there are various qualifiers (we emphasize that this are only qualifiers!) to distinguish the way of the resolution of a problem of calculating the average.

Thus, you must be very very careful about the calculations of averages because there is a tendency in business to rush and to systematically use the arithmetic mean without thinking, which can lead to serious errors! A nice example (for an analogy) is that a considerable number of laws require only moderate levels of pollution per year, while for example, smoking one cigarette per day during 365 days does not have the same impact as smoking 365 cigarettes in one day during one year when both have the same average taken over a year ... This is a clear evidence of statistical incompetence of the legislature.

Here is a small sample of common mistakes:

- Consider that the arithmetic mean is the value that divides the population into two equal parts (although it is the median that does this).
- Consider that the average of the ratios of the type goals/realisations is equal to the ratio of the average of the goals and of the average of the realizations (although it is not the same thing!).
- Consider that the average salaries of different subsidiaries, is equal to the global average (while this is true if and only if there is the same number of employees in each subsidiary of the company).
- Consider that the average of the average of the rows in a table is always equal to the average of the columns of the same table (although this is true if and only if the cell contents are not empty).
- Calculate the arithmetic average growth of the revenue in $\%$ (as the geometric mean must be used).
- etc.

We will see below different average with examples relative to arithmetic, to the enumeration, to physics, to econometrics, to geometry and sociology. The reader will find other practical examples by going through the entire book.

Definitions (#5.7.120): As given x_i real numbers, then we have:

D1. The "arithmetic average" or "sample average" (the most commonly known) is defined as the quotient of the sum of n observed x_i values by the total size n of the sample:

$$\boxed{\mu_a = \frac{1}{n}\sum_{i=1}^{n} x_i} \tag{5.7.1}$$

and is very often written $\bar{x}$ or $\widehat{\mu}$ and is for any discrete or continuous statistical distribution an unbiased estimator of the mean.

The arithmetic average represents a statistical measure expressing the magnitude that would have each member of a set of measures if the total must be equal to the product of the arithmetic average by the number of members.

✍**Example:**

Consider three products of respective prices $\{10.\text{-}, 30.\text{-}, 60.\text{-}\}$ whose price has increase of respectively $\{1.\text{-}, 6.\text{-}, 18.\text{-}\}$, hence respectively a total increase of $25.\text{-}$. The arithmetic average increase is therefore:

$$\mu_a = \frac{1+6+18}{3} = 8.\bar{3} \Rightarrow 3 \cdot 8.\bar{3} = 25.\text{-}$$

Remark

The statistical fact that when the characteristics of individual objects within a population of similar objects are distributed according to a given law (curve) lead us to the observation that when we sample such a population we are more likely to find a property near the average than one out on either tail of the curve. This is named the "principle of mediocrity".

If some values repeats more than once in the measurements, the arithmetic mean is then often formally denoted as following:

$$\mu_a = \frac{1}{n}\sum_{i=1}^{n} n_i x_i = \sum_{i=1}^{n} \frac{n_i}{n} x_i = \sum_{i=1}^{n} w_i x_i \tag{5.7.2}$$

and the last equality is named "weighted average" and often denoted:

$$\boxed{\mu_w = \sum_{i=1}^{n} w_i x_i} \tag{5.7.3}$$

Notice that the arithmetic average is a special case of the weighted average! Because we may have $w_i \neq n_i/n$ in some very common practical applications!

✍**Example:**

Consider three products of respective prices $\{10.\text{-}, 30.\text{-}, 60.\text{-}\}$ (hence a total of $100.\text{-}$) whose price has increase of respectively $\{1.\text{-}, 6.\text{-}, 18.\text{-}\}$, i.e. $\{10\%, 20\%, 30\%\}$, hence respectivelly a total incrase of 25.- (from $100.\text{-}$ to $125.\text{-}$ hence it is immediate that the global increase is of 25%). The arithmetic average increase in percentage is therefore:

$$\mu_a = \frac{10\% + 20\% + 30\%}{3} = 20\%$$

and the weighted average increase is of:

$$\mu_w = \frac{10.\text{-}}{100.\text{-}}10\% + \frac{30.\text{-}}{100.\text{-}}20\% + \frac{60.\text{-}}{100.\text{-}}30\% = 25\%$$

It is then easy to identify that the result that makes sense in this case is the weighted average and not the arithmetic average!

Finally, we could indicate that under this approach, the actual weighted average will be named "mathematical mean" or just "mean" in the field of study of probabilities.

We may as well use the frequencies of occurrence of the observed values named "classes frequencies":

$$f_i = \frac{n_i}{n} \tag{5.7.4}$$

So that we get another equivalent definition named the "weighted average by the classes frequencies":

$$\mu_a = \frac{1}{n}\sum_{i=1}^{n} n_i x_i = \sum_{i=1}^{n} \frac{n_i}{n} x_i = \sum_{i=1}^{n} f_i x_i \tag{5.7.5}$$

Remark

The reader may found interesting to know that in many business applications of statistics (especially in "Machine Learning"), when the statistics must be updated "on the fly" each time new data appear we speak then of "data flow statistics" or "data stream mining" or even "online learning". So an efficient algorithmic way must be found to update the calculations without recalculating everything from scratch! The simplest typical example is the mean! To calculate the new mean for each new incoming data given a series of $x_1, x_2, \ldots, x_n$ data, we simply use:

$$\mu_{k+1} = \frac{\mu_k \cdot k + x_{k+1}}{k+1} = \mu_k + \frac{x_{k+1} - \mu_k}{k+1} \tag{5.7.6}$$

Trying to express in a streaming way a maximum of machine learning/statistical algorithms or methods is a full time job (and plenty are already done)!

Before continuing, it's important to know that in the field of statistics it is useful and often necessary to combine measurements/data in class intervals of a given width (see examples below). We often have to make several tries to choose the intervals even if there are semi-empirical formulas for choosing the number of classes when we have n available values. One of these semi-empirical rules used by many practitioners is to retain the smallest integer k of classes such as:

$$\boxed{2^k \geq n} \tag{5.7.7}$$

the width of the class interval is then obtained by dividing the range (difference between the maximum and minimum measured value) by k. Thus:

$$k \geq \log_2(n) \tag{5.7.8}$$

By convention and rigorously... (so rarely respected in the notations), a class interval is closed on the left and open on the right (see section Numbers page 2211):

$$[\ldots,\ldots[\tag{5.7.9}$$

This empirical rule is named the "Sturges rule" and is based on the following reasoning:

We assume that the values of the binomial coefficient C_k^i gives the number of individuals in an ideal histogram (we let the reader check this simply with a spreadsheet software like Microsoft® Excel® 11.8346 and the `COMBIN(k,i)` function) of k intervals for the i-th interval. As k becomes large the histogram looks more and more like a continuous curve named the "Normal curve" or "bell curve" as we will see later.

Therefore, based on the binomial theorem (see section Calculus page 1364), we have:

$$n=\sum_{i=0}^{k} C_k^i=\sum_{i=0}^{k} C_k^i a^{i-k} b^k=\sum_{i=0}^{k} C_k^i 1^{i-k} 1^k=\sum_{i=0}^{k} C_k^i=(a+b)^k=(1+1)^k=2^k \tag{5.7.10}$$

Then, for each interval i the practitioner will traditionally take the average between the lower and upper limit for the calculation and multiply it by the corresponding class frequency f_i. Therefore, the grouping of class frequencies implies that:

(a) The weighted average by the frequencies differs from the arithmetic average.

(b) As the approximation seen above it will be a worst indicator compare to the arithmetic average...

(c) It is very sensitive to the choice of the number of classes, than very bad at this level.

There are many other empirical rules for the discretization of random variables. For example, the software XLStat offers not less than 10 rules (constant amplitude, Fisher algorithm, k-means, 20/80, etc.).

In practice we prefer to use ECDF (Empirical Cumulated Distribution Function) or QQ-plots because those latter do not rely on bin-size!

Later, we will see two very important properties of the arithmetic average and of the mean that you will have to understand absolutely (the weighted average of deviations from the average and the average deviations from the average).

Remark

The "mode", denoted Mod or simply M_0, is defined as the value that appears most often in a set of data. In Microsoft® Excel® 11.8346, it is important to know that the `MODE( )` function returns the first value in the order of values having the largest number of occurrences therefore assuming a unimodal distribution.

D2. The "median", also named the "middle value", denoted M_e (or more simply denoted M when any confusion with something else is avoided...) of a set of numbers is the value separating the higher half from the lower half of a data sample, a population, or a probability distribution into two equal parts. In the case of a continuous statistical distribution $f(x)$ of a random variable X, it is the value that represents the value that has 50% of cumulative probability to occur (we will see further in details the concept of statistical distribution):

$$P(X \leq M_e) = P(X \geq M_e) = \int\limits_{-\infty}^{M_e} f(x)\mathrm{d}x = \int\limits_{M_e}^{+\infty} f(x)\mathrm{d}x = 0.5 = F(X) \tag{5.7.11}$$

In the case of a series of ordered values $x_1, x_2, \ldots, x_i, \ldots, x_n$, the median is therefore by definition the value such that we have the same number of values that are greater than or equal to it than the number of values that are less or equal to it.

Remarks

R1. The median is mainly used for skewed distributions, because it represent them better than the arithmetic average.

R2. The median is in practice often not a single value (at least in the case where n is even). Indeed, between the values corresponding to ranges $\dfrac{n}{2}$ and $\dfrac{n}{2}+1$ there is an infinite number of values to choose which cut the population in half. This is why for some people the median is not intuitive and especially when there are a lot of duplicates values. That's why in the case of duplicates, using in a spreadsheet software a formula like `=COUNTIF(range of cells,"<=Median)/COUNT(range of cells)` to calculate the proportion of cells having a value less than the median (or more than the median) and expecting a result of 50% is obviously wrong!

R3. Some corporations remove the duplicate values before calculating the Median. That's scientifically wrong but may be arguable in context of internal governance of the board committee (typically if evidence has shown that this lead to better decisions!).

More rigorously:

- If the number of terms is odd, i.e. of the form $2n+1$, the median of the series is the term of order $n+1$ (that the terms are all distinct or not!).
- If the number of terms is even, i.e. of the form $2n$, the median of the series is half-sum (arithmetic average) of the values of the terms of rank n and $n+1$ (that the terms are all distinct or not!).

$$\text{median}(x) = \begin{cases} x_{(n+1)/2}, & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{n/2} + x_{n/2+1}\right) & \text{if } n \text{ is even} \end{cases} \tag{5.7.12}$$

In any case, by this definition, it follows that there are at least 50% of the terms of the series that are smaller than or equal to the median, and at least 50% of the terms of the series that are greater than or equal to the median.

For example, consider the table of wages below:

Employee N°	Wage	Cumulated Employees	%Cumulated Employees
1	1,200	1	6%
2	1,220	2	12%
3	1,250	3	18%
4	1,300	4	24%
5	1,350	5	29%
6	1,450	6	35%
7	1,450	7	41%
8	1,560	8	47%
9	1,600	9	53%
10	1,800	10	59%
11	1,900	11	65%
12	2,150	12	71%
13	2,310	13	76%
14	2,610	14	82%
15	3,000	15	88%
16	3,400	16	94%
17	4,800	17	100%

Table 5.28: **Identification of the median**

There is in the table an odd number $2n + 1$ of values. So the median of the series is the term of rank $n + 1$. This is $1,600$.- (result that give any spreadsheet software). The arithmetic average is in this case about $2,020$.-.

As in so many cases, the simple explanation for the median from school is simple, but not actually true. The defitinion of a median is clear for odd vector lengths, but there are different options for even vector length. That's why the `R` software implements 9 of them and the default one, named "sample median" isn't at the days we write these lines compliant with results given by SAS, SPSS not even with its own `quantile()` function.

Indeed, the usual mathematical definition given first above is based on the inverse of the CDF, which gives the "left median", the lower end of the median interval. This also has the valuable property that the median is one of the observed data values - in the rare cases where you aren't interested in the median as an estimator, but just as a summary of all the data, it should certainly be a data value.

The problem with the left median is symmetry. Why not the right median? The inverse CDF gives you the left median because CDFs are right-continuous, but they could have been left-continuous; mathematicians just picked one possibility. So, the right median is another possibility. Or, you might want some compromise between the two, and that's one way to get the school median. But we've now got at least two answers and possibly three (if you count the right median).

In direct relation with the median it is important to define the following concept to understand the underlying mechanism:

D3. Be given a statistical series $x_1, x_2, \ldots, x_i, \ldots, x_n$, we name "dispersion of absolute differences" around x the number $\varepsilon'(x)$ defined by:

$$\varepsilon'(x) = \sum_i \mid x_i - x \mid \qquad (5.7.13)$$

$\varepsilon'(x)$ is minimum for a value of x closest to a given value x_i in the sense of the absolute error value. The median is the value that achieves this minimum (extremum)! The idea will then be to study the variations of the function to find the position of this extremum.

Indeed, we can write:

$$\forall x \in [x_r, x_{r+1}], r \in \{1, 2, 3, \ldots, n-1\} \quad \varepsilon'(x) = \sum_{i=1}^{r} \mid x_i - x \mid + \sum_{i=r+1}^{n} \mid x_i - x \mid \qquad (5.7.14)$$

Then by definition of the x value:

$$\begin{aligned}\varepsilon'(x) &= \sum_{i=1}^{r} \mid x_i - x \mid + \sum_{i=r+1}^{n} \mid x_i - x \mid \\ &= [rx - (x_1 + x_2 + \ldots + x_r) - [(x_{r+1} + \ldots + x_n - (n-r)x]] \\ &= (2r - n)x + (x_{r+1} + \ldots + x_n) - (x_1 + x_2 + \ldots + x_r)\end{aligned} \qquad (5.7.15)$$

What allows us to skip the absolute values is simply the choice of the index r that is taken so that the series of values in practice can always be split into two parts: all that is less than the element indexed by r and all that is superior to it (i.e.. the median by anticipation...).

$\varepsilon'(x)$ is also a piecewise (discrete) affine function (similar to the equation of a line for fixed values of r and n) where we see that by analogy the factor:

$$2r - n \qquad (5.7.16)$$

is the slope of the function and:

$$(x_{r+1} + \ldots + x_n) - (x_1 + x_2 + \ldots + x_r) \qquad (5.7.17)$$

the Y-intercept (ordinates at the origin).

The function is decreasing (negative slope) until r is less than $\dfrac{n}{2}$ and increasing when r is greater than $\dfrac{n}{2}$ (it passes trough an extremum!). Specifically, we distinguish two particularly cases of interest since n is an integer:

- If n is even, we can say that $n = 2n'$, then the slope can be written $2(r - n')$ and it is equal to zero if $r = n$ and then, as the result is valid by construction only for $\forall x \in [x_r, x_{r+1}]$ then $\varepsilon'(x)$ is constant on $[x_{n'}, x_{n'+1}]$ and we have an extremum necessarily in the middle of this range (arithmetic average of the two terms).

- If n is odd, we can say that $n = 2n' + 1$ (we cut the series into two equal parts), then the slope can be written $(2r - 2n' - 1)$ and it is zero if $r = n' + \dfrac{1}{2}$, as the result is only valid for $\forall x \in [x_r, x_{r+1}]$ then it is immediate that the middle value is the median $x_{n'+1}$.

We find out the median in both cases. We will also see later how the median is defined for a continuous random variable (the underlying idea is the same).

There is another practical case where the statistician has at its disposal only the values grouped in intervals of statistical classes. The procedure for determining the median is then different:

When we have at our disposal only a values grouped in intervals of statistical classes, the abscissa of the point of the median is generally within a class. To then get a more accurate value of the median, we perform a linear interpolation. This is what we name the "linear interpolation method of the median".

The median value can be read from the graph or calculated analytically. Indeed, consider the graph of the cumulative probability $F(x)$ in class intervals as below where the bounds of the intervals were connected by straight lines:

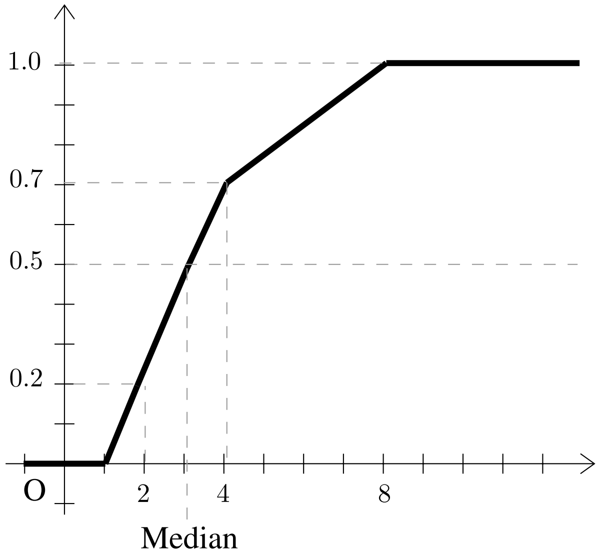


Figure 5.104: Graphical representation of the estimation of the median by linear interpolation

The value of the median M_e is obviously located at the crossroads between the cumulated probability of 50% (0.5) and the abscissa. Thus, by applying the basics of functional analysis, we have (just by observing that the slope in the interval containing the median is equal in the half-interval to the left and to right adjacent to the median):

$$\frac{\Delta x}{\Delta y} = \frac{M_e - 2}{0.5 - 0.2} = \frac{4 - 2}{0.7 - 0.2} \tag{5.7.18}$$

What we frequently write:

$$\frac{M_e - a}{0.5 - F(a)} = \frac{b - a}{F(b) - F(a)} \tag{5.7.19}$$

Thus the value of the median:

$$M_e = a + (b-a)\frac{0.5 - F(a)}{F(b) - F(a)} \tag{5.7.20}$$

Consider the following table that we will see again much later in this chapter:

Price of tickets	**Number of tickets**	**Cumulated number of tickets**	**Relative frequencies of tickets**
[0,50[	668	668	0.068
[50,100[	919	1,587	0.1587
[100,150[	1,498	3,085	0.3085
[150,200[	1,915	5,000	0.5000
[200,250[	1,915	6,915	0.6915
[250,300[	1,498	8,413	0.8413
[300,350[	919	9,332	0.9332
[350,400[	440	9,772	0.9772
[400 and +	228	10,000	1

Table 5.29: **Identification of the median and the mode**

We see that the "median class" is in the range $[150, 200]$ because the cumulative value of 0.5 is there (column at the right of the table) but the median has, using the previously established relation, the precise value of (it is trivial in the particular example of the table above, but we still do the calculation...):

$$M_e = 150 + (200-150)\frac{0.5 - 0.3805}{0.5 - 0.3805} = 200 \tag{5.7.21}$$

and of course we can do the same with any other percentile!

Definition (#5.7.121): We can also give on the way a definition to determine the "modal value" (the value that occurs the most frequently) often denoted:

$$M_0 = \text{mode}(X) = \arg\max_x f_X(x) \tag{5.7.22}$$

if we are only in possession of the frequencies of class intervals. To see that we start with diagram below named "grouped distribution" in frequencies bar:

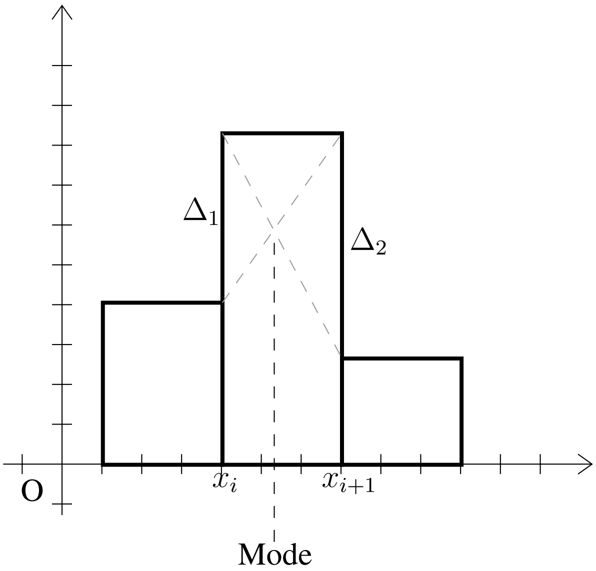


Figure 5.105: **Graphical representation of the estimation of the modal value with classes intervals**

Using the Thales relations (see section Euclidean Geometry 2610), it comes immediately, noting M_0 the modal value:

$$\frac{M_0 - x_i}{\Delta_1} = \frac{x_{i+1} - M_0}{\Delta_2} \tag{5.7.23}$$

As in a proportion, we do not change the value of the ratio by adding the numerators and adding the denominators, we get:

$$\frac{M_0 - x_i}{\Delta_1} = \frac{x_{i+1} - M_0}{\Delta_2} = \frac{x_{i+1} - x_i}{\Delta_1 + \Delta_2} \tag{5.7.24}$$

We then have:

$$M_0 = x_i + \frac{\Delta_1}{\Delta_1 + \Delta_2}(x_{i+1} - x_i) \tag{5.7.25}$$

With the previous example this gives then:

$$\begin{aligned} M_0 &= 150 + \frac{(1,915 - 1,498)}{(1,915 - 1,498) + (1,915 - 1,498)}(250 - 150) \\ &= 150 + \frac{1}{2}(250 - 150) = 150 + \frac{1}{2}100 = 200 \end{aligned} \tag{5.7.26}$$

The question that then arises is to the appropriateness of the choice of the mean, mode or median in terms of communication ... (normally we communicate them all three in corporate reports!).

A good example is that of the labour market where in general, while the average wage and the median wage are quite different, the institutions of state statistics calculate the median than many traditional media then explicitly equate to he concept of "arithmetic average" in their news...

Remark

To avoid getting an arithmetic average having little sense, we often calculate a "trimmed average", i.e. an arithmetic average calculated after removing outliers in the series (using Grubbs or Dixon tests or other some naive methods). See the section of Numerical Methods page 6560 for more details on this naive method.

The "quantile" generalize the concept of "median" by cutting the distribution in sets of equal parts (of the same cardinality we might say ...) or in other words in regular intervals. We define the "quartiles", the "decile" and "percentile" on the population, ordered in ascending order, that we divide by 4, 10 or 100 parts of the same size. So we talk about the 90th percentile to indicate the value separating the first 90% of the population and the 10% remaining.

Let X be a random variable with cumulated distribution function $F(x)$. The τth quantile of X is defined by:

$$Q_Y(\tau)\equiv\inf_x\{x:F(x)\geq\tau\} \tag{5.7.27}$$

or also sometimes denoted:

$$Q_X(p)=\min\{x\in\mathbb{R}\mid F_X(x)=p\} \tag{5.7.28}$$

The "midhinge" is the average of the first and third quartiles and is thus a measure of location. Equivalently, it is the 25% trimmed mid-range or 25% midsummary; it is an L-estimator (see page 1156) defined explicitly by:

$$\text{MH}(X)=\overline{Q_{1,3}(X)}=\dfrac{Q_1(X)+Q_3(X)}{2}=\dfrac{P_{25}(X)+P_{75}(X)}{2}=M_{25}(X) \tag{5.7.29}$$

The midhinge is related to the interquartile range (IQR), the difference of the third and first quartiles (i.e. $IQR=Q_3-Q_1$), which is a measure of statistical dispersion. The two are complementary in sense that if one knows the midhinge and the IQR, one can find the first and third quartiles.

The use of the term "hinge" for the lower or upper quartiles derives from John Tukey's work on exploratory data analysis in the late 1970 s, and "midhinge" is a fairly modern term dating from around that time. The midhinge is slightly simpler to calculate than the "trimean" (TM), which originated in the same context and equals the average of the median $\left(\tilde{X}=Q_2=P_{50}\right)$ and the midhinge:

$$\text{MH}(X)=2\,\text{TM}(X)-\text{med}(X)=2\dfrac{Q_1+2Q_2+Q_3}{4}-Q_2 \tag{5.7.30}$$

Definition (#5.7.122): The "P-th percentile" ($0<P\leq 100$) of a list of N ordered values (sorted from least to greatest) is the smallest value in the list such that no more than P percent of the data is strictly less than the value and at least P percent of the data is less than or equal to that value. This is obtained by first calculating the ordinal rank and then taking the value from the ordered list that corresponds to that rank. The ordinal rank n is calculated using the relation (using the round to nearest lower integer notation):

$$n=\left\lfloor\dfrac{P}{100}\cdot N\right\rfloor \tag{5.7.31}$$

Note that in Microsoft® Excel® 11.8346 the functions `QUARTILE( )`, `PERCENTILE( )`, `MEDIAN( )`, `PERCENTRANK( )` are available and it can be useful that we specify that there are several variants (at least nine as detailed in the excellent paper [271] and that are all available with the function `quantile( )` of the `R` statistical software!) of calculating these percentiles that explains possible variation between the results of different spreadsheet and statistical softwares[31]. When publishing scientific papers, scientists and engineers should always specify which on the nine methods of computing the quantiles was used!

✍**Example:**

The most famous example is to take the median for a better understanding of the vocabulary related to this topic! The median is indeed also the second quartile ($2\cdot 25\%$), but it's also at the same time the 0.5 quantile and the 50% percentile (i.e. centile) and the 5th decile.

More technically in the `R` software the 9 different types of quantile are defined as weighted averages of consecutive order statistics:

$$Q_i(p)=(1-\gamma)x_j+\gamma x_{j+1} \tag{5.7.32}$$

where $1\leq i\leq 9$, $\frac{j-m}{n}\leq p<\frac{j-m+1}{n}$, x_j is the jth order statistic, n is the sample size, the value of γ Is a function of $j=\lfloor np+m\rfloor$ and $g=np+m-j$, and m is a constant determined by the sample quantile type[32].

Remark

About Microsoft® Excel®... and it's conditional formatting tool... his user must not confuse the option "percentile" that is based on row numbers on the sorted values of interest on that of "percentage" defined for a vector $\vec{X}$ of values by:

$$\text{Percentage}(x_i)=\frac{x_i-\min\vec{X}}{R_{\vec{X}}}=\frac{x_i-\min\vec{X}}{\max\vec{X}-\min\vec{X}} \tag{5.7.33}$$

This concept is very important in the context of confidence intervals that we will see much further in this section and very useful in the field of quality with the use of "box plots" (also named "Box & Whiskers plots") to compare ("discriminate" as experts say) quickly two populations of data or more and especially to eliminate outliers (taking as reference the median will just make more sense!):

---

[31] The `R` Statistical software for example include all nine methods through a parameter of it's function `quantile()`.

[32] Based on the definitions, type = 1,3,4 should always give the same result for probs = 0.5 (i.e. median), and type = 2,5,6,7,8,9 should always give the same result for probs = 0.5 (i.e. median).

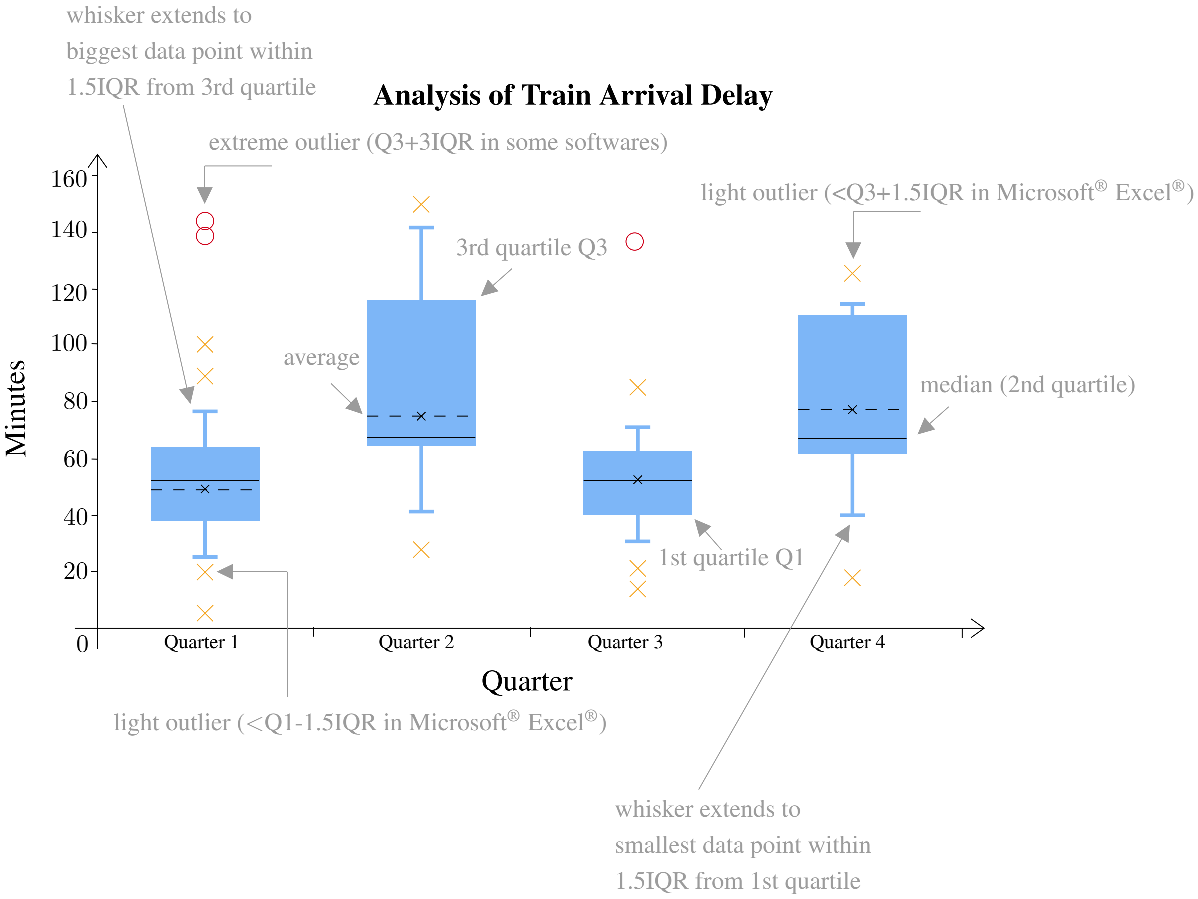
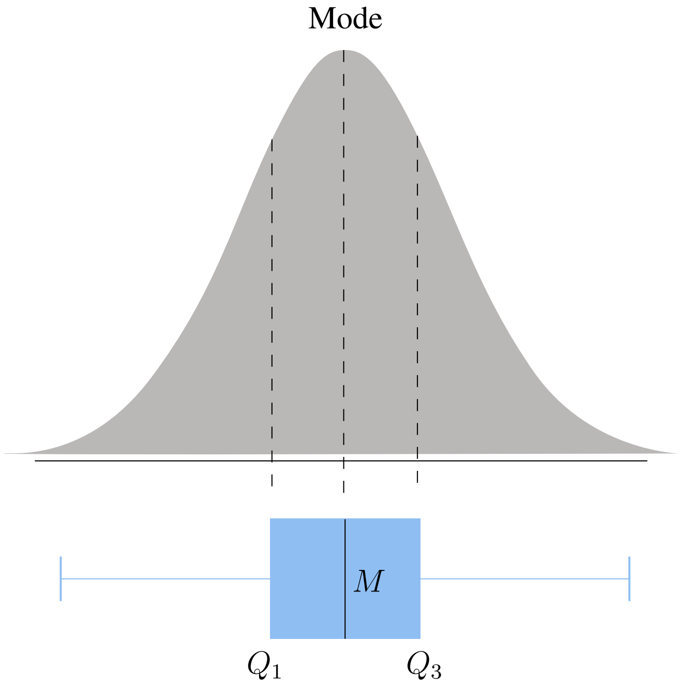


Figure 5.106: Box & Whiskers plot ideal chart type

Another very important mental representation of box plots is the following (it can get an idea of the asymmetry of the distribution as is able to do the R software):



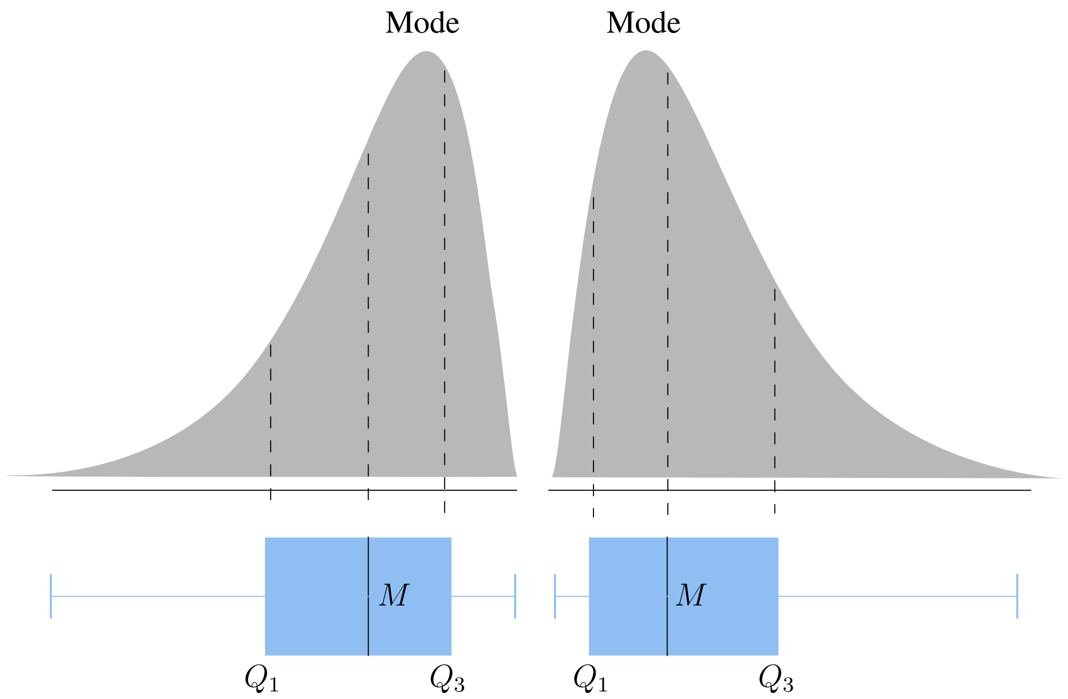
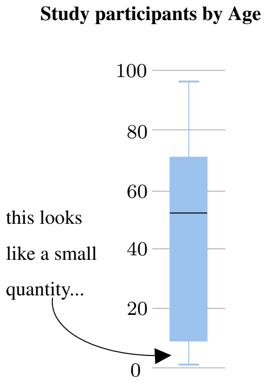
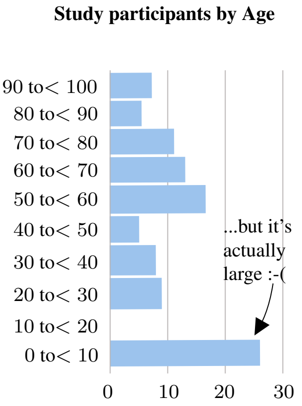
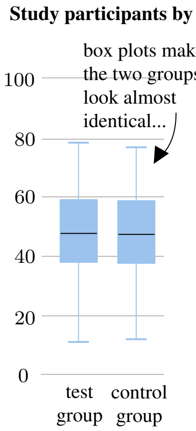


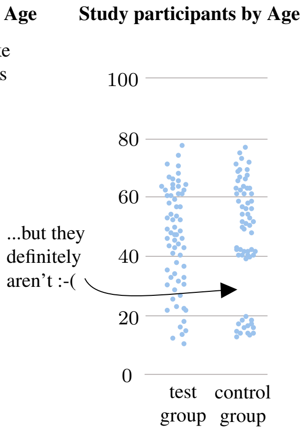
Figure 5.107: **Graphical representation of the mode, median and 1st + 3rd quartile compared to a distribution**

However notice that Box & Whiskers plots have important issues as depicted below:

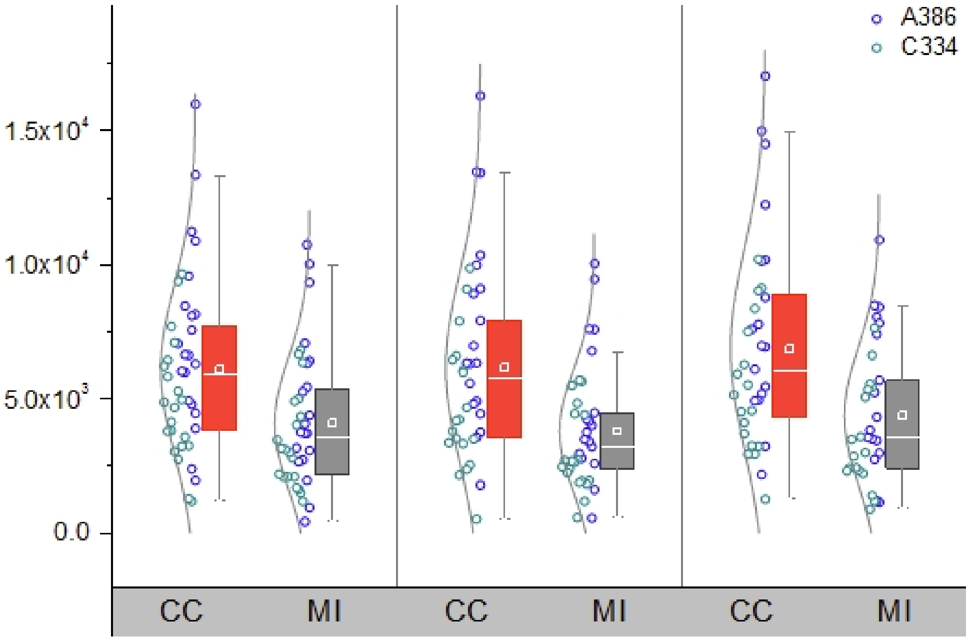








That's why we prefer in practice to use pirate plots (that are a mixed of violin plots and Box & Whiskers plots):



Remark

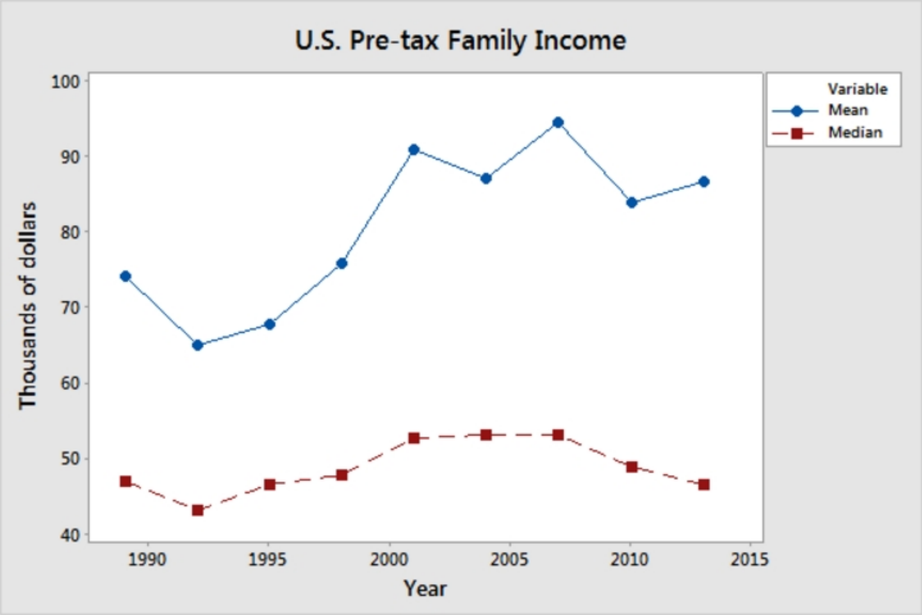
As we will prove it during our study of Inverse Transform Sampling in the section of Numerical Methods (see page 6710), the percentiles of any distribution are always uniformly distributed.

The concepts of median, outliers and confidence intervals that have yet been proved and/or just defined are so significant that there exists international standards for their proper use. First let us cite the norm ISO 16269-7:2001 *Median - Estimation and confidence intervals* and also the norm ISO 16269-4:2010 *Detection and treatment of outliers*. For more details and more outliers detection tests the reader can take a look to an entire textbook dedicated to outliers identification as [237] or look at the Data Science mind map (see page 6770) where more than twenty different tests are classified.

A beautiful demonstration on how governments can make statistics lie (hopefully now most of them has to give the data to the population for free) is for example the USA Federal Reserve that pointed out the fact a growth in average income since the 12010s (holocene calendar). But when we have access to the raw data:

`https://www.federalreserve.gov/econres/scfindex.htm`

and plot the median versus the average with for example Minitab 17 we get (the years in the plot are based on the Gregorian calendar):



and after some analysis you can point out that the increase in the average is largely due to increases in the highest 10% of incomes. The decline in the median at the same time suggests that typical Americans are in reality not doing as well since 12010 (holocene calendar)... Q.E.D!

Remark

Be careful with obvious conclusions when analyzing trends! As we will see during our study of Simpson's paradox (see further below page 967), when a body of data displays one trend, yet when broken (i.e. stratified) into subgroups, the opposite trend may appear into view for each of those subgroups!

D4. By analogy with the median, we define the "medial" as the value (in ascending order of values) that shares the (cumulative) sum values into two equal masses (i.e. the total sum divided by two).

In the case of our wages example, while the median gives the 50% of the salaries being below and above the medial gives how many employees share (and therefore the sharing wage) the first half and how many employees share the second half of the total of the wages costs.

For example, to return to our table on wages:

Employee N°	Wage	Cumulated Wages	% Cumulated Wages
1	1,200	1,200	3.5%
2	1,220	2,420	7%
3	1,250	3,670	10.7%
4	1,300	4,970	14.5%
5	1,350	6,320	18.4%
6	1,450	7,770	22.6%
7	1,450	9,220	26.8%
8	1,560	10,780	31.4%
9	1,600	12,380	36.1%
10	1,800	14,180	41.3%
11	1,900	16,080	46.8%
12	2,150	18,230	53.1%
13	2,310	20,540	59.8%
14	2,610	23,140	67.4%
15	3,000	26,140	76.1%
16	3,400	29,540	86%
17	4,800	34,340	100%

Table 5.30: **Identification of the mediale**

The sum of all wages is equal to $34,340$ and therefore the medial is $17,170$ then the medial is between the employees No. 11 and 12, while the median was $1,600$. We see then that the medial corresponds to 50% of the aggregate. This is a very useful indicator in Pareto or Lorenz analysis (see section Quantitative Management page 8095).

D5. Like the median, the "pseudomedian", used in some nonparametric statistical tests, is a measure of centrality for a dataset. The pseudomedian is derived by taking the median of the set of the means of pairs of data points, where the first is less than or equal to the second.

So, if the data were $\{1,3,7\}$, the pseudomedian would be the median of $\{$ mean $(1,1)$, mean $(1,3)$, mean $(1,7)$, mean $(3,3)$, mean $(3,7)$, mean $(7,7)\}$, which is 3.5. More technically, if the data are $x_1,x_2,\ldots,x_n$, then their pseudomedian is given by taking the median of the collection of $\frac{n}{2}(n+1)$ values of the form $w=\frac{x_i+x_j}{2}$ for $1\leq i\leq j\leq n$.

D6. The "root mean square" sometimes denoted simply Q which comes from the general relation:

$$\boxed{\mu_q=\sqrt[m]{\sum_{i=1}^{n}\frac{x_i^m}{n}}} \tag{5.7.34}$$

but where we take $m=2$.

Remark

It is one of the best known averages in statistics because the standard deviation is a root mean square (see below)!

✍**Example:**

Consider a square of side a, and another square of side b. The average area of two squares equals one square of side:

$$\mu_q^2=\frac{a^2+b^2}{2}\Rightarrow\mu_q=\sqrt{\frac{a^2+b^2}{2}} \tag{5.7.35}$$

In Microsoft® Excel® 11.8346 you can combine the functions `SUMSQ( )`, `COUNT( )` and to quickly calculate the root mean square as following:

```
=(SUMSQ(...)/COUNT(...))^(1/COUNT(...))
```

D7. The "harmonic mean" sometimes simply denoted H is defined by:

$$\boxed{\frac{1}{\mu_h}=\frac{\sum\limits_{i=1}^{n}x_i^{-1}}{n}} \tag{5.7.36}$$

It is little known but is often the result of simple and relevant arguments (typically the equivalent resistance of an electrical circuit with several resistors in parallel). There is a `HARMEAN( )` function in Microsoft® Excel® 11.8346 to calculate it.

✍**Example:**

Consider a distance d travelled in one direction at the speed v_1 and in the other direction (or not) at the speed v_2. The arithmetic average speed will be obtained by dividing the total distance $2d$ by the time of the travel:

$$v=\frac{2d}{t} \tag{5.7.37}$$

If we calculate the time it takes when travel d with a speed v_i that is simply the quotient:

$$t_i=\frac{d}{v_i} \tag{5.7.38}$$

The total time is therefore:

$$t=\frac{2d}{v}=t_1+t_2=\frac{d}{v_1}+\frac{d}{v_2} \tag{5.7.39}$$

If the distance is not the same for the both velocities anyway each velocity remains the same this is why d disappears!

In other words: We use the harmonic mean when are given to us ratios. Let's see other important examples to illustrate this:

✍**Examples:**

E1. Three investments have a Price / Return ratio of respectively 104%, 106% and 109% (so we lost money in all three investments in this particular case!). Knowing that the price of the three investments was initially the same, we use the harmonic average:

$$\mu_{\mathrm{h}}=\frac{3}{\frac{1}{104}+\frac{1}{106}+\frac{1}{109}}\cong 106.29\% \tag{5.7.40}$$

While an arithmetic average would give about 106.33%. Which makes a big difference when we manage millions in cash!

Note on the way that the following development shows that the harmonic mean is a special case of the "weighted arithmetic mean":

$$\begin{aligned}\mu_{a,y}=\sum_i w_i x_i=&\frac{\frac{1}{104}}{\frac{1}{104}+\frac{1}{106}+\frac{1}{109}}104+\frac{\frac{1}{106}}{\frac{1}{104}+\frac{1}{106}+\frac{1}{109}}106+\frac{\frac{1}{109}}{\frac{1}{104}+\frac{1}{106}+\frac{1}{109}}\\=&\frac{1}{\frac{1}{104}+\frac{1}{106}+\frac{1}{109}}+\frac{1}{\frac{1}{104}+\frac{1}{106}+\frac{1}{109}+\frac{1}{109}+\frac{1}{106}+\frac{1}{109}}=\frac{3}{\frac{1}{104}+\frac{1}{106}+\frac{1}{109}}\end{aligned} \tag{5.7.41}$$

If all weights are equal-weighted then we fall back on the standard arithmetic average!

E2. Consider that an investor puts us a disposition 300.- each month to buy a given asset. The first month this asset is worth 9.- so we can buy 33.333 units of it, the second month the asset is worth 11.- so we can buy 27.27227 units of it, finally the last and the third month the asset is worth 4.- so we can buy 75 units of it. The question is, what is the average price of this asset in our portfolio? We then have:

$$\mu_{\mathrm{k}}=\frac{900}{33.333+27.2727+75}=\frac{900}{\frac{300}{9}+\frac{300}{11}+\frac{300}{4}}=\frac{3}{\frac{1}{9}+\frac{1}{11}+\frac{1}{4}}\cong 6.6369.\text{-} \tag{5.7.42}$$

E3. Alice can finish a task in 2 hours and Bob takes 3 hours to finish the same task, how long would it take them if they worked together?

Alice works therefore a rate of $1/2$ (in on hour she can do 50% of the task) and Bob therefore at a rate of $1/3$ (in one hour he can do $33.\bar{3}\%$ of the taks). Together they work therefore at a rate of:

$$\frac{1}{2}+\frac{1}{3}=\frac{5}{6}$$

So it would take them $6/5=1.2$ of an hour (72 minutes) to finish it if they worked together.

D8. The "geometric mean" sometimes simply denoted G is defined in the general case by:

$$\boxed{\mu_g = \sqrt[n]{\prod_{i=1}^{n} x_i}} \tag{5.7.43}$$

This average is often forgotten by undergraduate employees but famous it is famous in the field of finance (see section Economy page 7548) this is also why there is an `GEOMEAN()` function in Microsoft® Excel® 11.8346 to calculate it.

A geometric mean is often used when comparing different items – finding a single "figure of merit" for these items – when each item has multiple properties that have different numeric ranges. For example, the geometric mean can give a meaningful "average" to compare two companies which are each rated at 0 to 5 for their environmental sustainability, and are rated at 0 to 100 for their financial viability. If an arithmetic mean were used instead of a geometric mean, the financial viability is given more weight because its numeric range is larger so a small percentage change in the financial rating (e.g. going from 80 to 90) makes a much larger difference in the arithmetic mean than a large percentage change in environmental sustainability (e.g. going from 2 to 5). The use of a geometric mean "normalizes" the ranges being averaged, so that no range dominates the weighting, and a given percentage change in any of the properties has the same effect on the geometric mean. So, a 20% change in environmental sustainability from 4 to 4.8 has the same effect on the geometric mean as a 20% change in financial viability from 60 to 72.

Remark

In 12010 (holocene calendar), the geometric mean was introduced to compute the Human Development Index by United Nations Development Programme. Poor performance in any dimension is directly reflected in the geometric mean. That is to say, a low achievement in one dimension is not anymore linearly compensated for by high achievement in another dimension. The geometric mean reduces the level of substitutability between dimensions and at the same time ensures that a 1% decline in index of, say, life expectancy has the same impact on the HDI as a 1% decline in education or income index. Thus, as a basis for comparisons of achievements, this method is also more respectful of the intrinsic differences across the dimensions than a simple average.

As for zero values, it is impossible to calculate the geometric mean with negative numbers! However, there are several workarounds for this problem, all of which require that the negative values be converted or transformed to a meaningful positive equivalent value. Most often this problem arises when it is desired to calculate the geometric mean of a percent change in a population or a financial return, which includes negative numbers.

For example, to get the geometric mean of the values $+12\%, -8\%, 0\%$ and $+2\%$, instead of calculating the geometric mean of their decimal multiplier equivalents of $1.12, 0.92, 1$ and 1.02, to compute a geometric mean equal to 1.0125. Subtracting 1 from this value gives the geometric mean of $+1.25\%$ as a net rate of growth or in financial circles is named the "Compound Annual Growth Rate C.A.G.R."), often denoted:

$$\text{CAGR} = \left(\left(\frac{\text{EV}}{\text{BV}}\right)^{\frac{1}{n}} - 1\right) \cdot 100 \tag{5.7.44}$$

where EV is the Ending Value of the investment, BV the Beginning value of the investment and n the number of years involved.

✍**Example:**

Suppose that a bank offers an investment opportunity and plans for the first year an interest (this is absurd, but this is an example) with a rate $(X-Y)\%$ but for the second year with an interest rate $(X+Y)\%$. At the same time another bank provides a constant interest rate for two years: $X\%$. We will say a little bit to fast that this is the same... In fact the two investments do not have the same profitability!

In the first bank, a capital C_0 will give after the first year of interest:

$$(X-Y)\%\cdot C_0 \tag{5.7.45}$$

and the second year:

$$(X+Y)\%\left[(X-Y)\%\cdot C_0\right] \tag{5.7.46}$$

In the other bank we will have after one year:

$$X\%\cdot C_0 \tag{5.7.47}$$

and after the second year:

$$X\%(X\%\cdot C_0) \tag{5.7.48}$$

and so on...

As you can probably see it the placement will not be identical if $Y\neq 0$! $X\%$ is the not the arithmetic average of $(X-Y)\%$ and $(X+Y)\%$.

Now if we write:

$$r_1=(X+Y)\% \quad \text{and} \quad r_2=(X-Y)\% \tag{5.7.49}$$

What is in reality the average value of the global interest rate r?

After 2 years (for example), the capital is multiplied by $r_1\cdot r_2$. If an average exists it will be denoted by r and the capital will thus be multiplied by r^2. Then we have the relation:

$$r^2=r_1r_2\Leftrightarrow r=\sqrt{r_1r_2} \tag{5.7.50}$$

This is an example where we therefore see the geometric mean. Forgetting to use of the geometric mean a common mistake in companies where some employees calculate the arithmetic average rate of increase of a reference value.

The reader can notice for the fun that the discrete arithmetic mean, geometric mean and harmonic mean that we have introduce earlier above can be encompassed in the following general expression:

$$f^{-1}\left(\frac{f(x_1)+\ldots+f(x_2)}{n}\right) \tag{5.7.51}$$

D9. The "moving average" of order k, also named sometimes "rolling mean", is defined on a sequence $x_1,\ldots,x_i,\ldots,x_n$ of realizations by:

$$\boxed{\forall i\in[1\ldots n-k]:\ \text{MA}_{i+k}(k)=\frac{\sum\limits_{j=i}^{i+k-1}x_j}{k}} \tag{5.7.52}$$

The moving average is used particularly in economics, where it represents a trend of a series of values, where the number of points is equal to the total number of points of the series of values less the number that you specify for the period (for a detailed example see the section of Economy page 7919).

Remark

On some Internet pages and statistical softwares, the "rolling mean" of order k (must be an odd positive number greater than 2!) is defined for a sequence $x_1,\ldots,x_i,\ldots,x_n$ (with $n\geq k$) by:

$$\forall i\in[(k+1)/2\ldots n-(k+1)/2]:\ \text{RM}_i=\frac{\sum\limits_{j=i+(k+1)/2-1}^{i+(k+1)/2-1}x_j}{k} \tag{5.7.53}$$

A moving average in finance is calculated from the average of a stock price over a given period: each point of a moving average of 100 sessions is the average of 100 last current values. This curve, displayed simultaneously with the evolution of the curve of the values, smooths the daily changes in the value and gives the possibility to better see the trends.

The moving averages can be calculated for different time periods, which can generate short-term trends MMC (20 sessions according to the habits of the domain), medium (50-100 sessions) or long-term MML (over 100 sessions):

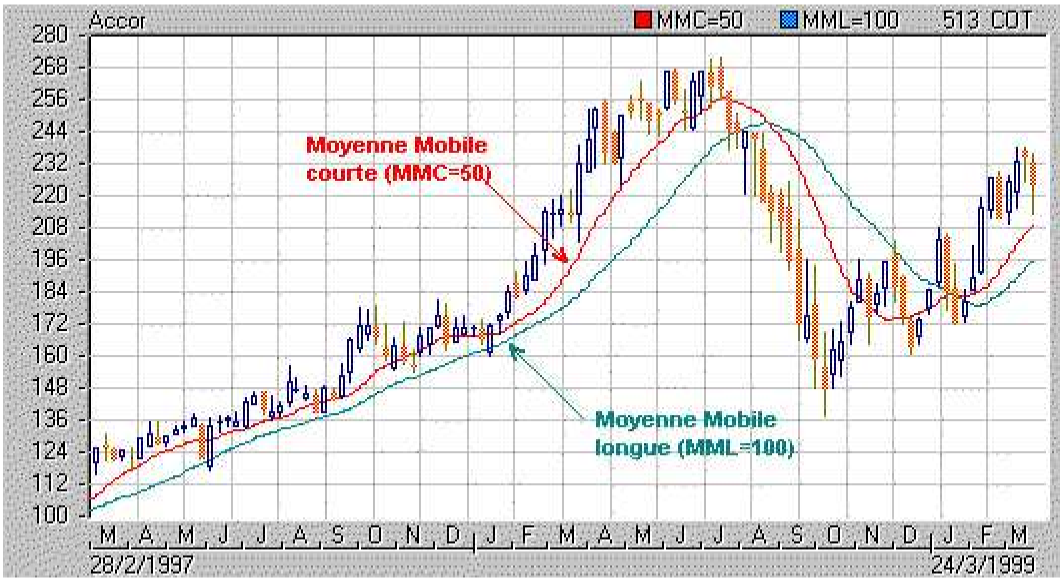


Figure 5.108: **Graphical representation of a few moving averages**

The crosses of the moving averages with the price curve (cutted with a certain granularity) of the value generate purchase or sale (basic) signals depending on the case:

- Buy signal: when the price curve crosses the MM up.
- Sell signal: when the price curve crosses the MM down.

In addition to the moving average, note that there are a lot of other artificial indicators often used in finance (the `R` software has a package dedicated only to such indicators). As for example the "upside/downside ratio".

The idea is the following: If you have a financial product (see section Economy page 7604) whose current price is P_c for which you have a goal of high gain corresponding to a high price, which we will denote by P_h (high price) and conversely, the potential loss that you feel is at a price P_l (low price).

Then the ratio:

$$\frac{P_h - P_c}{P_c - P_l} = \text{UD}_R \qquad (5.7.54)$$

gives the upside/down ratio!

✍**Examples:**

E1. For example, a financial product of 10.- with a low price of 5.- and a high price of 15.- has a ratio of $\text{UD}_R = 1$ and therefore an identical speculative factor to allow a gain or loss of 5.-.

E2. A financial product of 10.- with a low price of 5.- and a high price of 20.- has a ratio of $\text{UD}_R = 2$ and therefore twice the speculative potential gain compared to the loss.

Remark

Some financial institutions recommend to reject the UD_R below 3. Investors also tend to reject too high UD_R that can be a sign of artificial inflation.

D10. The "weighted average" (the moving average and arithmetic average are just a special cases of the weighted average with $w_i = 1$) is defined by:

$$\boxed{\mu_p = \frac{\sum\limits_{i=1}^{n} w_i x_i}{\sum\limits_{i=1}^{n} w_i}} \tag{5.7.55}$$

Is used for example in geometry to locate the centroid of a polygon, in physics to determine the center of gravity or in statistics to calculate the mean and other advanced regression techniques and in project management for estimating task durations forecast.

In the general case the weights w_i represents the weighted influence or arbitrary/empirical influence of the element x_i relatively to the others one.

D11. The "functional mean" or "integral average" is defined as:

$$\mu_f = \frac{1}{b-a}\int\limits_a^b f(x)\mathrm{d}x \tag{5.7.56}$$

where μ_f depends of a function f of a real integrated variable (see section Differential and Integral Calculus page 1497) on a range $[a, b]$. It is often used in signal theory (electronics, electrotechnichs).

5.7.2.1 Laplace Smoothing

To come back to our class frequencies seen above and before proceeding with the study of some mathematical properties of the averages... you should know that when we work with discrete probabilities distributions it happens very (very) often that we meet a typical problem whose source is the size of the population.

Consider as an example the case where we have 12 documents and that we would like estimate the probability of occurrence of the word "Viagra". We have on a sample the following values:

Document ID	Word occurrences
1	1
2	0
3	2
4	0
5	4
6	6
7	3
8	0
9	6
10	2
11	0
12	1

Table 5.31: **Class frequencies of the word**

Table that we can represent in another way:

Word occurrences	Documents	Probability
0	4	0.33
1	2	0.17
2	2	0.17
3	1	0.083
4	1	0.083
5	0	0
6	2	0.17

Table 5.32: **Respective frequencies classes of documents**

And here we have a common phenomenon. There is no record with 5 occurrences of the word of interest. The idea (very common in the field of Data Mining) is then to add artificially and empirically using a count using a technique named "Laplace smoothing" which involves adding k units at each occurrence. Therefore the table becomes:

Word occurrences	Documents	Probability
0	5	0.26
1	3	0.16
2	3	0.16
3	2	0.11
4	2	0.11
5	1	0.05
6	3	0.16

Table 5.33: **Frequencies classes of documents with smoothing**

Obviously this type of technique is debatable and beyond the scientific framework ... We even hesitated to introduce this technique in the section of Numerical Methods (with the rest of all the empirical numerical techniques) rather than here...

5.7.2.2 Means and Averages properties

Now we will see some relevant properties that connect some of these means and averages or are specific to a particular mean/average.

The first properties are important so beware to understand them:

P1. The calculation of the arithmetic, root mean square and harmonic average/mean can be generalized using the following expression:

$$\boxed{\mu=\sqrt[m]{\frac{1}{n}\sum_{i=1}^{n}x_i^m}} \tag{5.7.57}$$

where we see:

(a) For $m=1$, we get the arithmetic average

(b) For $m=2$, we get the root mean square

(c) For $m=-1$ we get the harmonic mean

P2. The arithmetic average has the property of linearity, that is to say (without proof because it is simple to check):

$$\boxed{\overline{\lambda x+\mu}=\lambda\bar{x}+\mu} \tag{5.7.58}$$

This is the statistical version of the property of the mean in the field of probabilities that we will see further.

P3. The weighted sum of the deviations from the arithmetic average is zero.

Proof 5.40.2. First, by definition, we know that:

$$n=\sum_{i=1}^{r}n_i \quad \text{and} \quad \mu=\frac{1}{n}\sum_{i=1}^{r}n_ix_i \tag{5.7.59}$$

then we have:

$$\boxed{\sum_{i=1}^{r} n_i(x_i-\mu)=\sum_{i=1}^{r} n_ix_i-\mu\sum_{i=1}^{r} n_i=\sum_{i=1}^{r} n_ix_i-\left(\frac{1}{n}\sum_{i=1}^{r} n_ix_i\right)n=\sum_{i=1}^{r} n_ix_i-\sum_{i=1}^{r} n_ix_i=0} \tag{5.7.60}$$

Thus, this tool can not be used as a measure of dispersion!

By extension, the arithmetic average of the weighted deviations from the average is also equal to zero:

$$\frac{\sum\limits_{i=1}^{r} n_i(x_i-\mu)}{n}=0 \tag{5.7.61}$$

■ Q.E.D.

This result is quite important because it will further be useful for a better understanding of the concept of standard deviation and variance.

P4. Now we would like to prove that:

$$\mu_h < \mu_g < \mu_a < \mu_q \tag{5.7.62}$$

> **Remark**
>
> The comparisons between the above means/averages and the median or the weighted or moving averages does not make sense this is why we won't compare them.

Proof 5.40.3. First, we consider two non-zero real numbers x_1 and x_2 as $x_2 > x_1 > 0$ and then we write:

(a) The arithmetic average:

$$\mu_a=\frac{x_1+x_2}{2} \tag{5.7.63}$$

(b) The geometric mean:

$$\mu_g=\sqrt{x_1x_2} \tag{5.7.64}$$

(c) The harmonic mean:

$$\frac{1}{\mu_h}=\frac{\dfrac{1}{x_1}+\dfrac{1}{x_2}}{2}=\frac{\dfrac{x_2}{x_1x_2}+\dfrac{x_1}{x_2x_1}}{2}=\frac{x_1+x_2}{2x_1x_2}\Leftrightarrow \mu_h=\frac{2x_1x_2}{x_1+x_2} \tag{5.7.65}$$

(d) The root mean square:

$$\mu_q^2=\frac{x_1^2+x_2^2}{2} \tag{5.7.66}$$

We will start to prove that $\mu_g > \mu_h$ by contradiction by putting $\mu_g - \mu_h < 0$:

$$\begin{aligned}\mu_g - \mu_h = \sqrt{x_1x_2} - \frac{2x_1x_2}{x_1+x_2} &= \frac{\sqrt{x_1x_2}x_1 + \sqrt{x_1x_2}x_2 - 2x_1x_2}{x_1+x_2} < 0\\ &\Rightarrow \sqrt{x_1x_2}x_1 + \sqrt{x_1x_2}x_2 - 2x_1x_2 < 0\\ &\Rightarrow \sqrt{x_1x_2}x_1 + \sqrt{x_1x_2}x_2 < 2x_1x_2\\ &\Rightarrow \frac{\sqrt{x_1x_2}}{x_2} + \frac{\sqrt{x_1x_2}}{x_1} < 2\\ &\Rightarrow \sqrt{\frac{x_1}{x_2}} + \sqrt{\frac{x_2}{x_1}} < 2\end{aligned} \tag{5.7.67}$$

By convenience we will now put:

$$y = \sqrt{\frac{x_2}{x_1}} \tag{5.7.68}$$

and we know that $y > 1$. We therefore have:

$$\sqrt{\frac{x_1}{x_2}} + \sqrt{\frac{x_2}{x_1}} = \frac{1}{y} + y \tag{5.7.69}$$

and remember we search if it is possible that:

$$\sqrt{\frac{x_1}{x_2}} + \sqrt{\frac{x_2}{x_1}} < 2 \tag{5.7.70}$$

We can now easily check this statement from the following equivalences:

$$\sqrt{\frac{x_1}{x_2}} + \sqrt{\frac{x_2}{x_1}} = \frac{1}{y} + y = \frac{y^2+1}{y} < 2 \Leftrightarrow y^2 + 1 - 2y < 0 \Leftrightarrow (y-1)^2 < 0 \tag{5.7.71}$$

There is also a contradiction, and this validates our initial hypothesis:

$$\mu_g - \mu_h > 0 \Leftrightarrow \mu_g > \mu_h \tag{5.7.72}$$

Let see if $\mu_g > \mu_a$.

Under the hypothesis $x_2 > x_1 > 0$. We search now to prove that:

$$\frac{x_1+x_2}{2} > \sqrt{x_1x_2} \tag{5.7.73}$$

Now we have the following equivalences:

$$\frac{x_1+x_2}{2} > \sqrt{x_1x_2} \Leftrightarrow (x_1+x_2)^2 > 4x_1x_2 \Leftrightarrow x_1^2 + x_2^2 - 2x_1x_2 > 0 \Leftrightarrow (x_1-x_2)^2 > 0 \tag{5.7.74}$$

and the last expression is obviously correct because the square of a (real) number is always positive which verifies our initial hypothesis:

$$\mu_a - \mu_g > 0 \Leftrightarrow \mu_a > \mu_g \tag{5.7.75}$$

We will prove now that $\mu_q > \mu_a$ by contradiction by putting $\mu_q - \mu_a < 0$:

$$\begin{aligned}\mu_q-\mu_a&=\sqrt{\frac{x_1^2+x_2^2}{2}}-\frac{x_1+x_2}{2}\\&\Rightarrow\sqrt{\frac{x_1^2+x_2^2}{2}}-\frac{x_1+x_2}{2}<0\\&\Rightarrow\sqrt{\frac{x_1^2+x_2^2}{2}}<\frac{x_1+x_2}{2}\\&\Rightarrow\frac{x_1^2+x_2^2}{2}<\frac{x_1^2+2x_1x_2+x_2^2}{4}\\&\Rightarrow x_1^2+x_2^2<\frac{x_1^2+2x_1x_2+x_2^2}{2}\\&\Rightarrow x_1^2-2x_1x_2+x_2^2<0\\&\Rightarrow(x_1-x_2)^2<0\end{aligned} \tag{5.7.76}$$

But the square of a (real) number is always positive which verifies our initial hypothesis:

$$\mu_q-\mu_a>0\Leftrightarrow\mu_q>\mu_a \tag{5.7.77}$$

We then have:

$$\boxed{\mu_h<\mu_g<\mu_a<\mu_q} \tag{5.7.78}$$

■ Q.E.D.

Once these inequalities proved, we can then move on to a figure that we attribute to Archimedes to place three of these averages. The interest of this example is to show that there are some remarkable relations between statistics and geometry (coincidence??).

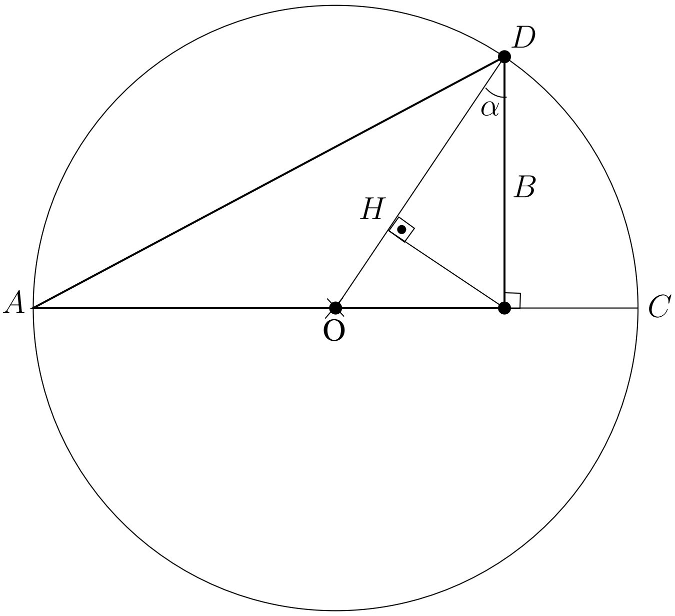


Figure 5.109: **Starting point for the geometric representation for the various averages**

We will first write $a = \overline{AB}, b = \overline{BC}$ and O is the midpoint of $\overline{AC}$. Thus, the circle is drawn with center O and radius $\overline{\text{O}A}$. D is the intersection of the perpendicular segment

$\overline{AC}$ through B and of the circle Ω (we can choose the intersection we want). H is itself the orthogonal projection of B on $\overline{\mathrm{O}D}$.

Archimedes says that $\overline{\mathrm{O}A}$ is the arithmetic average of a and b and that $\overline{BD}$ is the geometric mean of a and b, and $\overline{DH}$ is the harmonic mean of a and b.

We then prove that (could be trivial):

$$\overline{\mathrm{O}A}=\frac{\overline{AC}}{2}=\frac{a+b}{2} \tag{5.7.79}$$

Therefore $\overline{\mathrm{O}A}$ is the arithmetic average μ_a of a and b.

We have in the right-angled triangle ADB:

$$\overline{AD}^2=\overline{DB}^2+\overline{BA}^2 \tag{5.7.80}$$

Then we have in the right-angled triangle BDC:

$$\overline{DC}^2=\overline{BC}^2+\overline{DB}^2 \tag{5.7.81}$$

We then add these two relations, and we get:

$$2\overline{DB}^2+\overline{BA}^2+\overline{BC}^2=\overline{AD}^2+\overline{DC}^2 \tag{5.7.82}$$

We know that D is on a circle of diameter $\overline{AC}$, so ADC is rectangle on D. Therefore:

$$\overline{AD}^2+\overline{DC}^2=\overline{AC}^2 \tag{5.7.83}$$

And then we replace $\overline{BA}$ and $\overline{BC}$ by a and b:

$$2\overline{DB}^2+a^2+b^2=\overline{AC}^2=(a+b)^2\Leftrightarrow 2\overline{DB}^2=2ab \tag{5.7.84}$$

So finally:

$$DB=\sqrt{ab} \tag{5.7.85}$$

And therefore, $\overline{DB}$ is the geometric mean μ_g of a and b.

We have now to prove that $\overline{DH}$ is the harmonic mean of a and b. We have in a first time using the orthogonal projection as studied in the section of Vector Calculus (page 1821):

$$\overrightarrow{D\mathrm{O}}\circ\overrightarrow{DB}=\|\overrightarrow{D\mathrm{O}}\|\|\overrightarrow{DB}\|\cos(\alpha)=\overline{D\mathrm{O}}\cdot\overline{DH}=\frac{a+b}{2}\overline{DH} \tag{5.7.86}$$

Then we also have (again orthogonal projection):

$$\overrightarrow{D\mathrm{O}}\circ\overrightarrow{DB}=\left(\|\overrightarrow{D\mathrm{O}}\|\cos(\alpha)\right)\|\overrightarrow{DB}\|=\overline{DB}\cdot\overline{DB}=\overline{DB}^2 \tag{5.7.87}$$

Therefore we have:

$$\overline{DB}^2=\frac{a+b}{2}\overline{DH} \tag{5.7.88}$$

and since $DB=\sqrt{ab}$, we have then:

$$\overline{DH}=\frac{2ab}{a+b} \tag{5.7.89}$$

$\overline{DH}$ is therefore the harmonic mean of a and b. Archimedes was not wrong!

Definition (#5.7.123): In a contingency table (a table showing the distribution of one variable in rows and another in columns like the one visible below), the "marginal means" of one variable are the means for that variable averaged across every level of the other variable.

As the name implies, these means appear in the margins of the contingency table like the one below:

	Technique 1	**Technique 2**	**Technique 3**	*Marginal Mean*
Male	79.5	88.7	89.2	$(79.5+88.7+89.2)/3=85.8$
Female	88.3	87.7	90.6	$(88.3+87.7+90.6)/3=88.87$
Marginal Mean	$(79.5+88.3)/2=83.9$	$(88.7+87.7)/2=88.2$	$(89.2+90.6)/2=89.9$	

Table 5.34: **Marginal Mean example**

Marginal means are useful because they tell us the overall average value for a specific level of some variable. In essence, marginal means offer a simple way for us to understand the means for specific levels of variables.

Marginal means from descriptives like above are said to be "completely unconstrained", where as "estimated marginal means' are based on a model (like ANOVA, ANCOVA, etc.), and the assumptions of the model impose constraints on them. this is why they're named *estimated* marginal means, because they're estimated from a model!

Caution! Estimated marginal means are popular for summarizing linear models that include factors. For balanced experimental designs (see page 1007), they are just equal to the marginal means. For unbalanced data, they in essence estimate the marginal means you would have observed that the data arisen from a balanced experiment! Earlier developments regarding estimated marginal means were developed in a least-squares context and are sometimes referred to as "least-squares means". Since its early development, the concept has expanded far beyond least-squares settings!

5.7.2.2.1 Jensen Inequality

The "Jensen's inequality" is a particularly important relation (or property) in finance and insurance because it shows by the proof why an options seller will take the expected mean of payoff rather than payoff of the expect mean or why ultimately the insured will always pay more or equal than the differential of the expected mean of accident costs, and also in time series analysis why GARCH processes are naturally leptokurtic and also for the proof of the Kullback-Leibler divergence why that latter is strictly positive or null (see section Statistical Mechanics page 3542).

Theorem 5.41. Jensen's inequality will allow us to prove that for a convex function (see section Functional Analysis page 2280):

$$\boxed{f\left(\sum_{i=1}^{m} p_i x_i\right) \leq \sum_{i=1}^{m} p_i f(x_i)} \tag{5.7.90}$$

with $p_i > 0$, $\sum_i p_i = 1$. Either by using notations specific to the domain of statistics:

$$f(\mathrm{E}(X)) \leq \mathrm{E}(f(X)) \tag{5.7.91}$$

and to the valuation of options (see section Economy page 7788):

$$P(\mathrm{E}(S_T)) \leq \mathrm{E}(P(S_T)) \tag{5.7.92}$$

Before proceeding further, let us recall that a function f is concave if $-f$ is convex and vice versa. It is then immediate that:

$$f\left(\sum_{i=1}^{m} p_i x_i\right) \geq \sum_{i=1}^{m} p_i f(x_i) \tag{5.7.93}$$

Either by using notations specific to the domain of statistics:

$$f(\mathrm{E}(X)) \geq \mathrm{E}(f(X)) \tag{5.7.94}$$

Let's go to the proof and then we will make a small practical simplified example with derivatives in finance.

Proof 5.41.1. Let us see if:

$$f\left(\sum_{i=1}^{m} p_i x_i\right) \leq \sum_{i=1}^{m} p_i f(x_i) \tag{5.7.95}$$

is true by proceeding by induction. First for $m = 2$, we fall back on the classical definition of a convex function:

$$f(p_1x_1 + p_2x_2) \leq p_1 f(x_1) + p_2 f(x_2) \tag{5.7.96}$$

Or written differently as $p_1 + p_2 = 1$:

$$f(p_1x_1 + (1-p_1)x_2) \leq p_1 f(x_1) + (1-p_1) f(x_2) \tag{5.7.97}$$

Now, we assume the previous relation true for $m = n$, and we show that it is also true for $m = n + 1$:

$$f\left(\sum_{i=1}^{n+1} p_i x_i\right) = f\left(p_{n+1}x_{n+1} + \sum_{i=1}^{n} p_i x_i\right) = f\left(p_{n+1}x_{n+1} + (1-p_{n+1})\sum_{i=1}^{n} \frac{p_i}{1-p_{n+1}} x_i\right) \tag{5.7.98}$$

Since is f is convex, we can write:

$$f\left(p_{n+1}x_{n+1} + (1-p_{n+1})\sum_{i=1}^{n} \frac{p_i}{1-p_{n+1}} x_i\right) \leq p_{n+1} f(x_{n+1}) + (1-p_{n+1}) f\left(\sum_{i=1}^{n} \frac{p_i}{1-p_{n+1}} x_i\right) \tag{5.7.99}$$

Let us put:

$$q_i = \frac{p_i}{1-p_{n+1}} \tag{5.7.100}$$

for $i = 1 \ldots n$. We have $q_i > 0$ for all $i = 1 \ldots n$ and:

$$\sum_{i=1}^{n} q_i = \sum_{i=1}^{n} \frac{p_i}{1-p_{n+1}} = \frac{1}{1-p_{n+1}} \sum_{i=1}^{n} p_i = \frac{1-p_{n+1}}{1-p_{n+1}} = 1 \tag{5.7.101}$$

because for recall:

$$\sum_{i=1}^{n+1} p_i = 1 \tag{5.7.102}$$

and consequently:

$$\sum_{i=1}^{n} p_i = 1 - p_{n+1} \tag{5.7.103}$$

By hypothesis of recurrence we then have:

$$f\left(\sum_{i=1}^{n} q_i x_i\right) \leq \sum_{i=1}^{n} q_i f(x_i) \tag{5.7.104}$$

And therefore:

$$f\left(\sum_{i=1}^{n+1} p_i x_i\right) \leq p_{n+1} f(x_{n+1}) + (1-p_{n+1}) \sum_{i=1}^{n} \frac{p_i}{1-p_{n+1}} f(x_i) \tag{5.7.105}$$

by simplifying the expression on the right we get indeed:

$$f\left(\sum_{i=1}^{n+1} p_i x_i\right) \leq \sum_{i=1}^{n+1} p_i f(x_i) \tag{5.7.106}$$

■ Q.E.D.

✍**Example:**

We know that the payoffs of a Call and a Put are respectively at maturity (from the point of view of a buyer and therefore by symmetry from the point of view of a seller also in absolute value):

$$\begin{gathered} \max(S_T - K, 0) \\ \max(K - S_T, 0) \end{gathered} \tag{5.7.107}$$

Let us consider a price at maturity of 102.- and that the possible strikes at maturity are $\{100, 110, 150\}$, we have then have:

$$\begin{aligned} \mathrm{E}(P(S_T)) &= \frac{\max(x_{T,1} - K, 0) + \max(x_{T,2} - K, 0) + \max(x_{T,3} - K, 0)}{3} \\ &= \frac{\max(100 - 102, 0) + \max(110 - 102, 0) + \max(150 - 102, 0)}{3} = 18.\bar{6}.\text{-} \end{aligned} \tag{5.7.108}$$

and:

$$\begin{aligned}P(\mathrm{E}(S_T)) &= \max\left(\frac{x_{T,1}-K+x_{T,2}-K+x_{T,3}-K}{3},0\right)=\max\left(\frac{x_{T,1}+x_{T,2}+x_{T,3}-3K}{3},0\right)\\ &= \max\left(\frac{100+110+150-3\cdot 102}{3}\right)=\max(18,0)=18.\text{-}\end{aligned} \tag{5.7.109}$$

and so we have well for a Call from the point of view of a buyer or seller in absolute values:

$$P(\mathrm{E}(S_T))\leq \mathrm{E}(P(S_T)) \tag{5.7.110}$$

All what we have seen so far (and that we will see further!) is related to random values localized on a straight line. For statistics on a circle or a sphere, named "directional statistics", we strongly recommend the reader to look at [370].

5.7.3 Type of variables

In talking about variables quantitative or qualitative variables, sometimes you hear variables being described as categorical (or sometimes nominal), or ordinal, or interval. Below we will define these terms and explain why they are important.

Definitions (#5.7.124):

D1. The "discrete variables" (by counting) that belongs to $\mathbb{Z}$: Are analysed with statistical laws based on a countable definition domain always strictly positive (the Poisson or Hypergeoemetric distribution are such typical case in the industry). Are almost always represented graphically by histograms.

D2. The "continuous variables" (by measure) that belong to $\mathbb{R}$: Are analysed with statistical laws based on an uncountable domain of definition strictly positive or may take any positive or negative value (typically the Normal distribution in the industry). Are almost always represented graphically by histograms with class intervals.

D3. The "attribute variables" (by classification): They are not digital data (only when they are coded with digits!) but qualitative data type {Yes, No}, {Passed, Failed}, {On time, Late}, {red, green blue, black}, etc. The binary data type attribute follow typically a Bernoulli or Binomial distribution while higher order qualitative variables have no average or standard deviation (indeed... try to calculate the mean and standard deviation between the qualitative variables Red, Green and Pink...).

In attribute variable we mainly distinct two subtypes of variables:

(a) A "categorical variable" (sometimes named a "nominal variable", or "multicategorical variable" or "polytomous variable") is one that has two or more categories, but there is no intrinsic ordering to the categories. For example, gender is a categorical variable having two categories (male and female) and there is no intrinsic ordering to the categories. Hair color is also a categorical variable having a number of categories (blonde, brown, brunette, red, etc.) and again, there is no agreed way to order these from highest to lowest. A purely categorical variable is one that simply allows you to assign categories but you cannot clearly order the variables. If the variable has a clear ordering, then that variable would be an ordinal variable, as described below.

(b) An "ordinal variable" is similar to a categorical variable. The difference between the two is that there is a clear ordering of the variables. For example, suppose you have a variable, named *economic status*, with three categories: {low, high, medium}. In addition to being able to classify people into these three categories, you can order the categories as low (1) $<$ medium (2) $<$ high (3). Now consider a variable like *educational experience* (with values such as: {high school graduate, college graduate, some college, elementary school graduate}). These also can be ordered as: elementary school graduate (1) $<$ high school graduate (2) $<$ some college (3) $<$ college graduate (4).

A pictorial and more complete summary of variable types in statistics is given in the figure below:

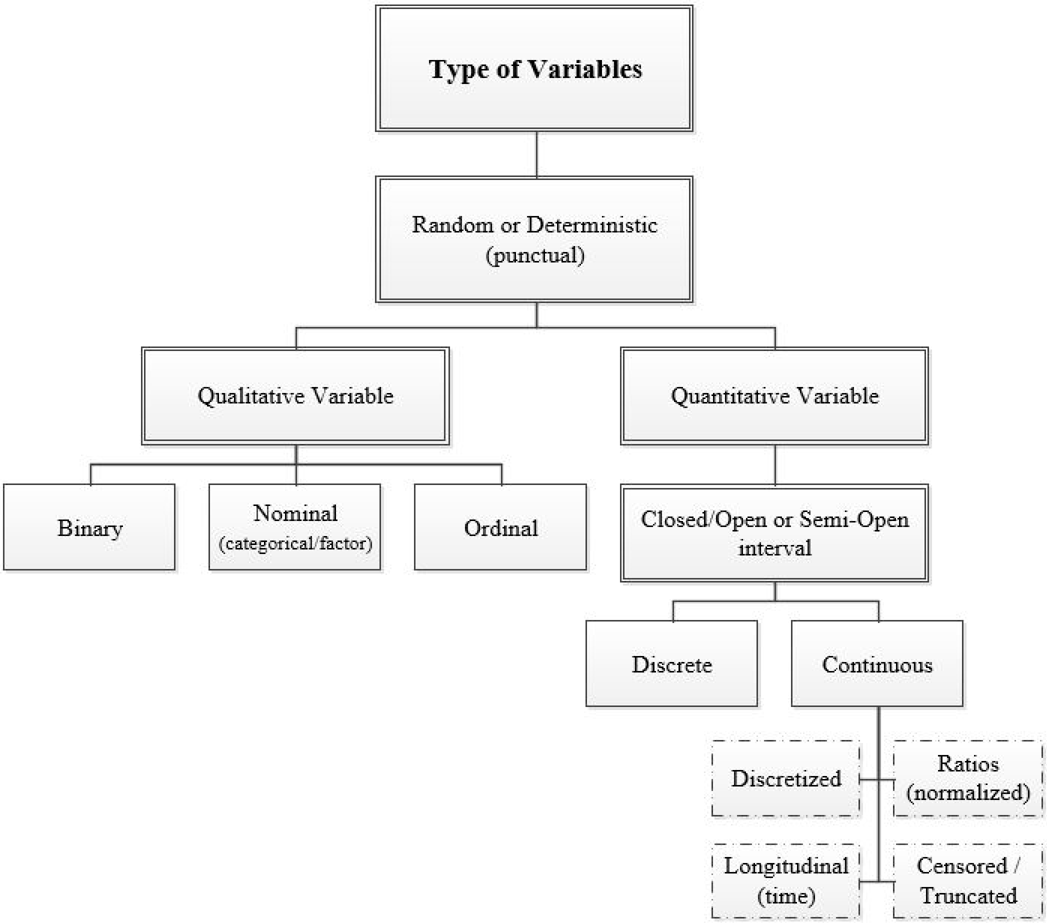

Type of Variables
Random or Deterministic (punctual)
Qualitative Variable
Quantitative Variable
Binary
Nominal (categorical/factor)
Ordinal
Closed/Open or Semi-Open interval
Discrete
Continuous
Discretized
Ratios (normalized)
Longitudinal (time)
Censored / Truncated

Figure 5.110: **Type of variables in statistics**

Remark

In principle, any continuous variable or any interval or ratio variable with more than two observed values is a polytomous variable, but the term is usually reserved for nominal and ordinal variables, named "unordered and ordered polytomous variables".

Understanding the different types of data is an important discipline for the engineers because it has important implications for the type of analysis tools and techniques that will be used.

A common question regarding the collection of data is what is the amount that should be collected. In fact it depends on the desired level of accuracy. We will see much further in this section (with proof!) how to mathematically determine the amount of data to collect.

Now that we are relatively familiar with the concept of average (mean), we can discuss on more formal calculations and that will make sense.

5.7.3.1 Discrete Variables and Moments

Consider X is an independent variable (an individual of a sample, whose property is independent of other individuals) that can take discrete random values (realizations of the vector $(X_1, X_2, \ldots, X_n)$) with respective probabilities $(p_1, p_2, \ldots, p_n)$ where, by the axioms of probabilities (see section Probabilities page 476):

$$p_i \in [0,1] \qquad \sum_i p_i = 1 \tag{5.7.111}$$

Let X be a numeric (quantitative) random variable (r.v.). It will be fully described in practice most of time by the value of the probability (for discrete variables) for a realization of this variable or by:

Definitions (#5.7.125):

D1. The probability corresponding to its value, named "probability density function[33]" (P.D.F.) and denoted by:

$$P(X) = f(X) \tag{5.7.112}$$

also sometimes (and more properly) denoted:

$$P(X = x) = f_X(x) \tag{5.7.113}$$

keeping in mind that $\sum f_X(x)$ must be equal to 1.

D2. The cumulative probability (for discrete AND continuous variables) to be typically less than or equal X for all realizations x. This "cumulative distribution function" (C.D.F.) is denoted by:

$$F(x) = P(x \leq X) \qquad \forall x \in \mathbb{R} \tag{5.7.114}$$

also sometimes (and more properly) denoted (for recall D is the domain of definition as defined on page 2211 in the section of Functional Analysis):

$$F_X(t) = P(X \leq t) = P(X \in]\min(D), t]) \qquad \forall t \in \mathbb{R} \tag{5.7.115}$$

with:

$$0 \leq P(X) \leq 1 \quad \text{and} \quad 0 \leq F(X) \leq 1 \tag{5.7.116}$$

where $F(x)$ is also named the "repartition function" of the random variable X. It is the theoretical proportion of the population whose value is less than or equal to x. It follows for example:

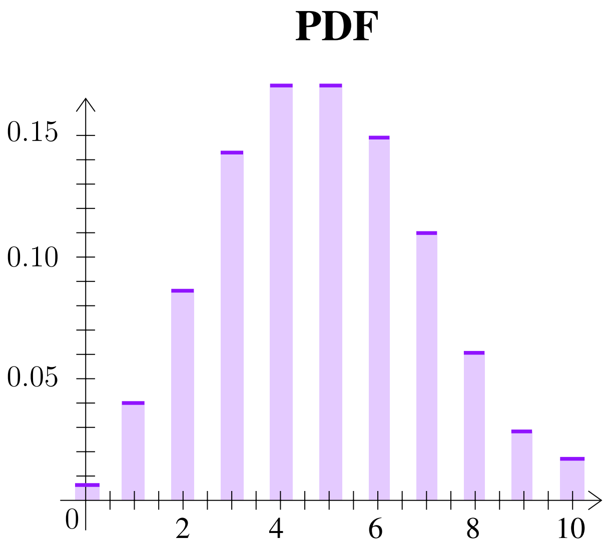
$$P(X > x) = 1 - F(x) \Leftrightarrow P(X \leq x) + P(X > x) = 1 \tag{5.7.117}$$

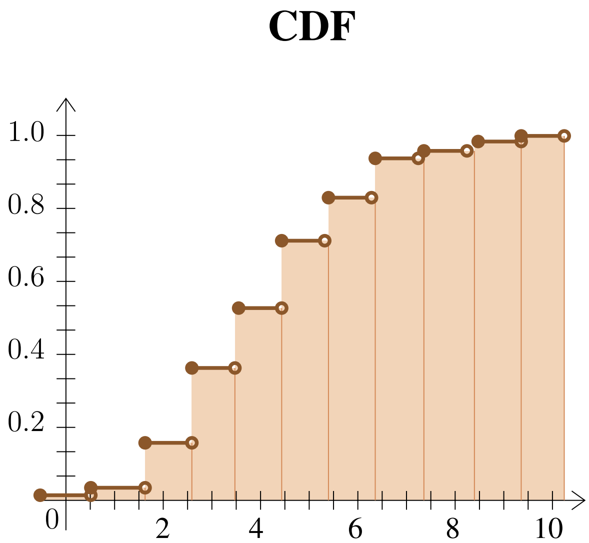
---

[33] For discrete variable some use instead the terminology: "probability mass function" (PMF).

More generally, for any two numbers a and b with $a < b$, we have:

$$P(a \leq x \leq b) = F(b) - F(a) \tag{5.7.118}$$


PDF


CDF

This definition obviously also applies to continuous variables!

D3. The "empirical cumulated distribution function" (E.C.D.F.) is naturally defined by (we have indicated the different notations that you can find in the literature):

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{x_i \leq x} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i \leq x} = \frac{\#(i : x_i \leq x)}{n} \tag{5.7.119}$$

with $\hat{F}_n(x) \in [0,1]$ and associated with the sample of independent and identically distributed variables which as we know is named a "random vector" denoted by $\vec{X} = (x_1, x_2, \ldots, x_n)$.

It is simply the cumulative frequencies of appearance normalized to unity below a certain fixed value (approach that the majority of human beings are naturally using when seeking the repartition function).

So if we take again the example of wages already used above, then we have for example for x fixed to 1,800:

Ordered Wages $x_i \leq x$	**Frequencies** $1_{x_i \leq x}$
$1,200$	1
$1,220$	1
$1,250$	1
$1,300$	1
$1,350$	1
$1,450$	1
$1,450$	1
$1,560$	1
$1,600$	1
$1,800$	1
$1,900$	0
$2,150$	0
$2,310$	0
$2,600$	0
$3,000$	0
$3,400$	0
$4,800$	0

Table 5.35: **Example of the empirical repartition function**

And then:

$$\hat{F}_{17}(x \leq 1,800) = \frac{1}{17}\sum_{i=1}^{17} 1_{x_i \leq x} = \frac{1}{17} \cdot 10 \cong 59\% \tag{5.7.120}$$

The repartition function is clearly a monotonically increasing function (or more precisely "non-decreasing") whose values range from 0 to 1.

Let's look at a property that we will use several times and which is important in practice (property valid for both discrete and continuous random variables):

Let X be a random variable of repartition function F_X and density function f_X (let us recall that: $f_X = F'_X$) and $c^{te} \in \mathbb{R}^*$ a constant such that $y = c^{te}X$. We want to determine the repartition function and the density function of the random variable Y from F_X and f_X respectively.

By definition for $y \in \mathbb{R}$:

$$F_Y(y) = P(Y \leq y) = P\left(c^{te}X \leq y\right) = P\left(X \leq \frac{y}{c^{te}}\right) = F_X\left(\frac{y}{c^{te}}\right) \tag{5.7.121}$$

Therefore:

$$\forall y \in \mathbb{R} \quad F_Y(y) = F_X\left(\frac{y}{c}\right) \tag{5.7.122}$$

With regard to the density function we have:

$$f_Y(y) = F'_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}F_X\left(\frac{y}{c^{te}}\right) = \frac{1}{c^{te}}F'_X\left(\frac{y}{c^{te}}\right) = \frac{1}{c^{te}}f_X\left(\frac{y}{c^{te}}\right) \tag{5.7.123}$$

Therefore:

$$\forall y \in \mathbb{R} \quad f_Y(y) = \frac{1}{c^{te}}f_X\left(\frac{y}{c^{te}}\right) \tag{5.7.124}$$

5.7.3.1.1 Mean and Variance of Discrete Random Variables

Definition (#5.7.126): We define the "expectation" or "mean", also named "moment of order 1", of the random variable X by the relation (with various notations):

$$\boxed{\mu_X=\mathbb{E}(X)=\text{E}(X)=\sum_i P(X=x_i)x_i=\sum_i p_i x_i} \tag{5.7.125}$$

also sometimes named "parts rule".

In other words, we know that for every event in the sample space is associated with a probability that we also associate with a value (given by the random variable). The question then is to know what value we can get on the long term? The expected value (the mean...) is then the weighted average, by the probability, of all values of the events of sample space.

If the probability is given by a discrete distribution function $f(x_i)$ (see the definitions of distribution functions further below in the text) of the random variable, we then have:

$$p_i=f(x_i)\Rightarrow \mu_X=\text{E}(X)=\sum_i x_i f(x_i) \tag{5.7.126}$$

Remarks

R1. The mean μ_X can also be written just simply μ if there are no possible confusion on the random variable.

R2. If we consider each realization of the random variables $(x_1,x_2,\ldots,x_n)$ as the components of a vector $\vec{x}$ and each associated probability (or ponderation) $(p_1,p_2,\ldots,p_n)$ as the components of a vector $\vec{p}$ we can write the mean in a technical way using the scalar product (see section Vector Calculus page 1821) often written:

$$\vec{p}\circ\vec{x}=\langle\vec{p},\vec{x}\rangle=\sum_{i=1}^{n}p_i x_i=\text{E}(X) \tag{5.7.127}$$

Here are the most important mathematical properties of the mean for any random variable (whatever the distribution law!) and that we will often use throughout this section (and many other involving statistics):

P1. Multiplication by a constant (homogeneous):

$$\text{E}(\alpha X)=\sum_i \alpha x_i P(X=x_i)=\alpha\sum_i x_i P(X=x_i)=\alpha\text{E}(\alpha X) \tag{5.7.128}$$

P2. Sum of two random variables (independent or not!):

$$\begin{aligned}\mathrm{E}(X+Y) &= \sum_{i,j}\left[(x_i+y_j)P\left((X=x_i)\cap(Y=y_i)\right)\right]\\ &= \sum_{i,j}\left[x_iP\left((X=x_i)\cap(Y=y_j)\right)\right]+\sum_{i,j}\left[y_iP\left((X=x_i)\cap(Y=y_j)\right)\right]\\ &= \sum_i x_i\left[\sum_j P\left((X=x_i)\cap(Y=y_i)\right)\right]+\sum_j y_j\left[\sum_i P\left((X=x_i)\cap(Y=y_j)\right)\right]\\ &= \sum_i x_iP\left[(X=x_i)\cap\bigcup_j(Y=y_i)\right]+\sum_j y_jP\left[(Y=y_j)\cap\bigcup_i(X=x_i)\right]\\ &= \sum_i x_iP(X=x_i)+\sum_j y_jP(Y=y_j)=\mathrm{E}(X)+\mathrm{E}(Y)\end{aligned} \tag{5.7.129}$$

Where we used in the 4th line, the property view in the section of Probabilities (page 477):

$$P\left(\bigcup_{i\in\mathbb{N}}A_i\right)=\sum_{i\in\mathbb{N}}P(A_i) \tag{5.7.130}$$

We deduce that for n random variables X_i following any probability distribution:

$$\mathrm{E}\left(X_i\right)=\sum_i \mathrm{E}(X_i) \tag{5.7.131}$$

P3. Then mean of a constant a is equal to the constant itself:

$$\mathrm{E}(a)=\sum_i ap_i=a\sum_i p_i=a\cdot 1=a \tag{5.7.132}$$

P4. Mean of a product of two random variables:

$$\mathrm{E}(X\cdot Y)=\sum_{i,j}x_iy_iP(X=x_i,Y=y_j) \tag{5.7.133}$$

And if the two random variables are independent, then the probability is equal to the joint probability (see section Probabilities page 480). Therefore we have:

$$\begin{aligned}\mathrm{E}(X\cdot Y) &= \sum_{i,j}x_iy_jP(X=x_i,Y=y_j)=\sum_{i,j}x_iy_jP(x_i)P(y_j) \quad =\sum_{i,j}x_iP(x_i)y_iP(y_i)\\ &= \sum_{i,j}x_iP(x_i)\sum_{i,j}y_iP(y_i)=\mathrm{E}(X)\mathrm{E}(Y)\end{aligned} \tag{5.7.134}$$

✍**Example:**

For a unique dice the expected value is:

$$\text{E}(X)=\frac{1}{6}\cdot 1+\frac{1}{6}\cdot 2+\frac{1}{6}\cdot 3+\frac{1}{6}\cdot 4+\frac{1}{6}\cdot 5+\frac{1}{6}\cdot 6=\frac{1}{6}\cdot 21=3.5$$

Using the property of linearity of the mean, we have then for the sum of n dices:

$$\text{E}_n(X)=n\text{E}(X)=n\cdot 3.5$$

For the special case of two dices, there the most likely sum will be:

$$\text{E}_2(X)=2\text{E}(X)=2\cdot 3.5=7$$

So the mean of the product of independent random variables is always equal to the product of their means.

We will assume as obvious that these four properties extend to the continuous case!

Definition (#5.7.127): After having translated the trend by the mean it is interesting to have and indicator that reflects the dispersion or "dispersion" around the mean by a value named "variance of X" or "second-order centered moment" or "mean square error (MSE), written $\text{V}(X)$ or σ_X^2 (read "sigma square") and given in its discrete form by:

$$\boxed{\sigma_X^2=\text{V}(X)=\text{MSE}=\text{E}\left[(X-\mu_X)^2\right]=\sum_i (x_i-\mu_X)^2 f(x_i)=\sum_i (x_i-\mu_X)^2 p_i} \tag{5.7.135}$$

In other words the variance measure of how much a distribution of values is spread from the mean.

The variance is however not directly comparable to the mean because of the fact that the units of the variance are the square of the unit of the random variable, which follows directly from its definition. To have an indicator of dispersion that can be compared to the parameters of central tendency (mean, median and ... mode), it then suffices to take the square root of the variance.

For convenience, we define the "standard deviation" of X by:

$$\boxed{\sigma_x=\sigma(X)=\sqrt{\text{V}(X)}} \tag{5.7.136}$$

The standard deviation is therefore the root mean square of the deviations (or "mean squared difference") between the observations and their mean.

Remarks

R1. The standard deviation σ_X of the random variable X can be simply denoted σ if there is no possible confusion.

R2. The standard deviation and variance are, in the literature, often named "dispersion parameters" or "scale parameters" as opposed to the mean, mode and median that are named "positional parameters".

Notice that when two different types of measurement methods give different statistical distributions of the variable of interest, then in physics and statistics, the distance between the two peaks of some unimodal distributions is named the "tension". For example in the year 12021 (holocene calendar), the tension between the cosmological radio-wave background measurement method and the direct measurement method of the Hubble constant has a tension of 5σ as illustrated below:

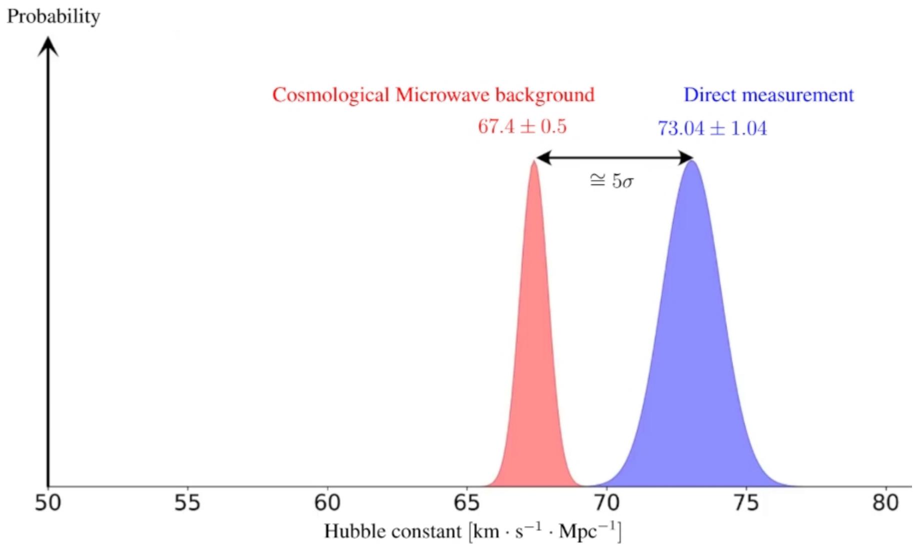


Figure 5.111: **Illustration of "tension" in statistics**

Definition (#5.7.128): The ratio (expressed in %):

$$\boxed{\mathrm{CV}=\frac{\sigma}{\mu}} \tag{5.7.137}$$

is often used in business to compare the mean and the standard deviation and is named the "coefficient of variation C.V." because it has no units (which is it's main advantage!) and because many industrial statistical methods consider that a good C.V should ideally be just about a few % only.

More generally for any statistics estimator $\hat{\theta}$ (sum, average, median, etc.) we can build a coefficient of variation such that:

$$\mathrm{CV}=\frac{\sigma_{\hat{\theta}}}{\hat{\theta}} \tag{5.7.138}$$

Thus, in practice we consider that:

Coefficient of variation	**Quality**
20%	Poor
10%	Acceptable
5%	Controlled
2.5%	Excellent
1.25%	World Class
0.0625%	Rarely achieved

Table 5.36: **Qualitative judgements of C.Vs commonly accepted**

Why do we find a square (respectively a square root) in the definition of the variance? The intuitive reason is simple (the rigorous much less ...). Remember, that we have shown above that the sum of the deviations from the actual weighted average is always zero:

$$\sum_{i=1}^{r} n_i(x_i - \mu) = 0 \tag{5.7.139}$$

If we assimilate the size of each sample by the probability by normalizing the sample size with respect to n, we come upon a relation that is the same as the variance with the difference that the term in brackets is not squared. And then we immediately see the problem... the dispersion measure is always zero, hence the need to bring this to the square.

We could, however, imagine to use the absolute value of deviations from the mean, but for a number of reasons that we will see later during our study of estimators, the choice of squaring is quite natural.

Note, however, the common use in the industry of two common other indicators of dispersion:

1. The "mean absolute deviation" (mean of the absolute values of deviations from the mean):

   $$\text{AVEDEV} = \frac{\sum\limits_{i=1}^{n} \mid x_i - \mu \mid}{n} \tag{5.7.140}$$

   Which is a very elementary indicator used when we do not want to make statistical inference on a series of measures. This deviation can be easily calculated in the English version of Microsoft® Excel® 11.8346 using the `AVEDEV( )` function.

2. The "median absolute deviation" denoted MAD (median of absolute values of deviations from the median):

   $$\text{MAD} = M_e\left(\mid X - M_e(X) \mid\right) \tag{5.7.141}$$

   which is considered as a more robust measure of dispersion than those given by the mean absolute deviation or the standard deviation (unfortunately this indicator is not natively integrated in spreadsheets softwares). That latter must not be confused with the "mean median absolute deviation" defined by:

   $$\text{MMAD} = \frac{\mid X - M_e(X) \mid}{n} \tag{5.7.142}$$

Remark

In order to use the MAD as a consistent estimator for the estimation of the standard deviation σ (see further below for the definition of the standard deviation) as we will need it for our study of Control Charts (see section of the corresponding name page 8934), one takes:

$$\hat{\sigma} = k \cdot \text{MAD} \tag{5.7.143}$$

where k is a constant scale factor, which depends on the distribution.

For normally distributed data as we will see during our study of the half-Normal distribution (see page 796) we take:

$$\text{MAD} \cong 2\sqrt{2}\text{erf}^{-1}(0.50) \cong 0.67499\sigma \tag{5.7.144}$$

Hence:

$$\sigma \cong 1.4826\text{MAD} \tag{5.7.145}$$

Having seen the definition of the both concepts let us see now through an example how to calculate them and their difference on a same set of data!

✍**Example:**

Consider the following measure of a random variable X:

$$(1, 1, 2, 2, 4, 6, 9) \tag{5.7.146}$$

and where the median value is given as we know by:

$$M_e(X) = M_e(1, 1, 2, 2, 4, 6, 9) = 2 \tag{5.7.147}$$

The absolute deviations from the median are then:

$$\mid X - M_e(x) \mid = (1, 1, 0, 0, 2, 4, 7) \tag{5.7.148}$$

Placed in ascending order, we then have:

$$(0, 0, 1, 1, 2, 4, 7) \tag{5.7.149}$$

where we easily identify the absolute deviation from the median, which is:

$$\text{MAD} = M_e(0, 0, 1, 1, 2, 4, 7) = 1 \tag{5.7.150}$$

In the case where we have at disposition a series of measures, we can estimate the experimental value of the mean (expectation) and of the variance with the following estimators (it is simply the of average and standard deviation of a sample when the events are equally likely) with the specific notation:

$$\boxed{\hat{\mu}=\sum_{i=1}^{n}x_i \quad \text{and} \quad \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2} \tag{5.7.151}$$

Proof 5.41.2. First for the mean:

$$\mu=\mathrm{E}(X)=\sum_{i=1}^{n}p_ix_i=\sum_{i=1}^{n}\frac{1}{n_i}x_i=\frac{1}{n}\sum_{i=1}^{n}x_i=\hat{\mu} \tag{5.7.152}$$

And for the variance:

$$\begin{aligned}\sigma^2&=\mathrm{E}((X-\mathrm{E}(X))^2)=\mathrm{E}((X-\hat{\mu})^2)=\sum_{i=1}^{n}p_i(x_i-\mu)^2=\sum_{i=1}^{n}\frac{1}{n_i}(x_i-\mu)^2\\&=\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2=\hat{\sigma}^2\end{aligned} \tag{5.7.153}$$

■ Q.E.D.

Theorem 5.42. Let us prove now a very nice little property as the arithmetic average is an optimum for the sum of squared errors.

Proof 5.42.1.

$$\sum_{i=1}^{n}(x_i-\alpha)^2=\sum_{i=1}^{n}x_i^2-2\alpha\sum_{i=1}^{n}x_i+n\alpha^2 \tag{5.7.154}$$

And if we search for α as the derivative of the above expression is equal to zero:

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\left(\sum_{i=1}^{n}x_i^2-2\alpha\sum_{i=1}^{n}x_i+n\alpha^2\right)=0 \tag{5.7.155}$$

then α is an optimum. We have therefore:

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\left(\sum_{i=1}^{n}x_i^2-2\alpha\sum_{i=1}^{n}x_i+n\alpha^2\right)=-2\sum_{i=1}^{n}x_i+2n\alpha=0 \tag{5.7.156}$$

or after rearrangement and an elementary simplification we get:

$$\alpha=\frac{1}{n}\sum_{i=1}^{n}x_i \tag{5.7.157}$$

■ Q.E.D.

It is effectively the arithmetic average! Now to see if it is a maximum extrema or minimum extrema we just need calculate the second derivative (see section Differential and Integral Calculus page 1454) and see if it gives a positive constant (i.e. the first derivative increases when α increase). Therefore we immediately see that it is effectively a minimum extrema!!!

Remark

The term of the sum that we see in the expression of the variance (standard deviation) is named the "sum of squared deviations from the mean" or "sum of squared errors from the mean". We also name it the "total sum of squares", or "total variation", or "sum of square errors" in the context of the study of the ANOVA (see the further below)

Before that we continue, let us recall the concept of geometric mean seen above (widely used for returns in finance or growth analyses in % of sales):

$$\mu_g=\sqrt[n]{\prod_{i=1}^{n} x_i}=\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} \tag{5.7.158}$$

It's fine but employees in financial departments also need to calculate the standard deviation of this average. The idea is then to take the logarithm to reduce it to a simple arithmetic mean (it is still obviously an estimator!):

$$\ln(\mu_g)=\ln\left(\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}\right)=\frac{1}{n}\ln\left(\left(\prod_{i=1}^{n} x_i\right)\right)=\frac{1}{n}\sum_{i=1}^{n}\ln(x_i) \tag{5.7.159}$$

Remark

For the section of Finance, if we denote the decimal multiplier equivalent $R_i=1+r_i$ or a random variable return r_i then:

$$\begin{aligned}\mu_a(\ln(R))&=\frac{1}{n}[\ln(R_1)+\ln(R_2)+\ldots+\ln(R_n)]=\frac{1}{n}\ln(R_1R_2\ldots R_n)=\ln\left[(R_1R_2\ldots R_n)^{1/n}\right]\\&=\ln\left(\sqrt[n]{R_1R_2\ldots R_n}\right)=\ln(\mu_g(R))\end{aligned} \tag{5.7.160}$$

However if $\mu_g(R)\cong 1$ then using first order Taylor expansion $\ln(x)\cong x-1$ (around $x\cong 1$ obviously!) we get:

$$\mu_a(\ln(R))=\ln(\mu_g(R))=\mu_g(R)-1 \tag{5.7.161}$$

So in other words: the arithmetic average of the log-returns (their decimal multiplier equivalent) is approximately equal to geometric mean of the returns (their decimal multiplier equivalent). It is an approximation you can often found in many financial computer codes.

Therefore, since taking the logarithm of the values we have the arithmetic mean of the log values, then the logarithm of the geometric standard deviation (with physicist reasoning like...) will be:

$$\ln\left(\hat{\sigma}_g\right)=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\ln(x_i)-\ln(\mu_g))}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\ln\left(\frac{x_i}{\mu_g}\right)} \tag{5.7.162}$$

Then we just take the exponential of the standard deviation of the logarithms of the values to have the "geometric standard deviation":

$$\hat{\sigma}_g=e^{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\ln\left(\frac{x_i}{\mu_g}\right)}} \tag{5.7.163}$$

It can be easily proved, using the same method as earlier, that the geometric standard deviation of the returns (the last relation above) is approximately equal to the Pearson standard deviation of the log-returns.

By the way, remember that we have proved earlier that $\mu_g < \mu_a$.

The variance can also be written in the very important way named the "Huygens relation" or "König-Huygens theorem" or "Steiner translation theorem" that we will reuse several times thereafter. Let's see what it is:

$$\begin{aligned}\mathrm{V}(X)&=\mathrm{E}[(X-\mu_X)^2]=\sum_i(x_i-\mu_X)^2f(x_i)=\sum_i\left(x_i^2-2\mu_Xx_i+\mu_X^2\right)f(x_i)\\&=\sum_i x_i^2f(x_i)-2\mu_X\sum_i x_if(x_i)+\mu_X^2\sum_i f(x_i)=\sum_i x_i^2f(x_i)-2\mu_X^2+\mu_X^2\\&=\sum_i x_i^2f(x_i)-\mu_X^2=\mathrm{E}(X^2)-\mu_X^2=\mathrm{E}(X^2)-\mathrm{E}(X)^2\end{aligned} \tag{5.7.164}$$

Remark

The reader may found interesting to know that in many business applications of statistics (especially in "Machine Learning"), when the statistics must be updated "on the fly" each time new data appear we speak then of "data flow statistics" or "data stream mining" or even "online learning". So an efficient algorithmic way must be found to update the calculations without recalculating everything from scratch! The simplest typical example is the variance! Updating the sample variance calculation looks obviously quite consuming if we just add new data according to:

$$\hat{\sigma}=\sqrt{\sum_i\frac{1}{n}(x_i-\hat{\mu})^2} \tag{5.7.165}$$

But using the Huygens relation, we see that this sample variance becomes much less consuming to calculate as it becomes only a problem of updating two sums:

$$\hat{\sigma}=\sqrt{\mathrm{E}(X^2)-\mathrm{E}(X)^2}=\sqrt{\frac{\sum_i x_i^2}{n}-\frac{\left(\sum_i x_i\right)^2}{n}} \tag{5.7.166}$$

Let us now do a relatively small hook to a common scenario generator of errors in business when several statistical series are handled (very common case in the industry as well as in insurance or finance)!

Consider two data series on the same character:

- $(x_1,n_1),(x_2,n_2),\ldots,(x_p,n_p)$ sample of total size n, arithmetic average $\bar{x}$, standard deviation σ_x.
- $(y_1,m_1),(y_2,m_2),\ldots,(y_p,m_q)$ sample of total size m, arithmetic average $\bar{y}$, standard deviation σ_y.

with $n+m=N$.

We then have:

$$\bar{z}=\frac{\sum_{i=1}^{p} x_i+\sum_{i=1}^{q} y_i}{n+m}=\frac{n\frac{1}{n}\sum_{i=1}^{p} x_i+m\frac{1}{m}\sum_{i=1}^{q} y_i}{n+m}=\frac{n\bar{x}+m\bar{y}}{n+m} \tag{5.7.167}$$

So the average of the averages is not equal to the overall average (first common mistake in business) except if the two data series have the same sample size ($n=m={}^1\!/\!_2N$)!!!

Let have a look at the standard deviation always with the same situation. First remember that we have:

$$\begin{aligned}\sigma_x^2&=\sum_{i=1}^{p} p_i(x_i-\bar{x})^2=\frac{1}{n}\sum_{i=1}^{p} n_i(x_i-\bar{x})^2=\frac{1}{\sum_{i=1}^{p} n_i}\sum_{i=1}^{p} n_i(x_i-\bar{x})^2\\ \sigma_y^2&=\sum_{i=1}^{q} p_i(y_i-\bar{y})^2=\frac{1}{m}\sum_{i=1}^{q} m_i(y_i-\bar{y})^2=\frac{1}{\sum_{i=1}^{q} m_i}\sum_{i=1}^{q} m_i(y_i-\bar{y})^2\end{aligned} \tag{5.7.168}$$

To continue, recall that we have previously proved the Huygens theorem and therefore:

$$\mathrm{V}(Z)=\sigma_z=\mathrm{E}(Z)-\mathrm{E}(Z)^2=\frac{1}{k}\sum_{i=1}^{k} k_i z_i^2-\bar{z}^2 \tag{5.7.169}$$

Therefore we have:

$$\begin{aligned}
\sigma_z^2 &= \frac{1}{k}\sum_{i=1}^{k} k_i z_i^2 - \bar{z}^2 = \frac{1}{k}\sum_{i=1}^{k} k_i z_i^2 - \left(\frac{n\bar{x}+m\bar{y}}{n+m}\right)^2 \\
&= \frac{1}{n+m}\left(\sum_{i=1}^{p} n_i x_i^2 + \sum_{i=1}^{q} m_i y_i^2\right) - \left(\frac{n\bar{x}+m\bar{y}}{n+m}\right)^2 \\
&= \frac{n\dfrac{1}{n}\displaystyle\sum_{i=1}^{p} n_i x_i^2 + m\dfrac{1}{m}\displaystyle\sum_{i=1}^{q} m_i y_i^2}{n+m} - \left(\frac{n\bar{x}+m\bar{y}}{n+m}\right)^2 \\
&= \frac{n\dfrac{1}{n}\displaystyle\sum_{i=1}^{p} n_i x_i^2 - n\bar{x}^2 + m\dfrac{1}{m}\displaystyle\sum_{i=1}^{q} m_i y_i^2 - m\bar{y}^2}{n+m} + \frac{n\bar{x}+m\bar{y}^2}{n+m} - \left(\frac{n\bar{x}+m\bar{y}}{n+m}\right)^2 \\
&= \frac{n\left(\dfrac{1}{n}\displaystyle\sum_{i=1}^{p} n_i x_i^2 - \bar{x}^2\right) + m\left(\dfrac{1}{m}\displaystyle\sum_{i=1}^{q} m_i y_i^2 - \bar{y}^2\right)}{n+m} + \frac{n\bar{x}+m\bar{y}^2}{n+m} - \left(\frac{n\bar{x}+m\bar{y}}{n+m}\right)^2 \\
&= \frac{n\sigma_x^2+m\sigma_y^2}{n+m} + \frac{n\bar{x}+m\bar{y}^2}{n+m} - \left(\frac{n\bar{x}+m\bar{y}}{n+m}\right)^2 \\
&= \frac{n\sigma_x^2+m\sigma_y^2}{n+m} + \frac{(n\bar{x}^2+m\bar{y}^2)(n+m)-(n\bar{x}+m\bar{y}^2)}{(n+m)^2} \\
&= \frac{n\sigma_x^2+m\sigma_y^2}{n+m} + \frac{(n^2\bar{x}^2+mn\bar{y}^2+mn\bar{x}^2+m^2\bar{y}^2)-(n^2\bar{x}^2+2n\bar{x}m\bar{y}+m^2\bar{y}^2)}{(n+m)^2} \\
&= \frac{n\sigma_x^2+m\sigma_y^2}{n+m} + \frac{mn\bar{y}^2+mn\bar{x}^2-2n\bar{x}m\bar{y}}{(n+m)^2} = \frac{n\sigma_x^2+m\sigma_y^2}{n+m} + nm\frac{\bar{y}^2+\bar{x}^2-2\bar{x}\bar{y}}{(n+m)^2} \\
&= \frac{n\sigma_x^2+m\sigma_y^2}{n+m} + nm\frac{(\bar{y}-\bar{x})^2}{(n+m)^2}
\end{aligned} \tag{5.7.170}$$

Redoing exactly the same developments with the unbiased variance we would get:

$$\sigma_z^2 = \frac{(n-1)\hat{\sigma}_x^2+(m-1)\hat{\sigma}_y^2}{n+m-1} + nm\frac{(\bar{y}-\bar{x})^2}{(n+m)(n+m-1)} \tag{5.7.171}$$

So we see that the overall variance is not equal to the sum of the variance (second common mistake in business)!!!

Consider now X being a random variable of mean μ (constant and determined value) and variance σ^2 (constant and determined value), we define the "reduced centered variable" by the relation:

$$\boxed{Y = \frac{X-\mu}{\sigma}} \tag{5.7.172}$$

Remark

When the real mean is unknowns and same for the variance, we denoted the above relation as following (various notation are represented):

$$Y=\frac{X-\hat{\mu}}{\hat{\sigma}}=\frac{X-\bar{X}}{S} \tag{5.7.173}$$

and this is named a "studentized variable".

Theorem 5.43. We prove in a very simple way by using the property of linearity of the mean and property of scalar multiplication of the variance that:

$$\mathrm{E}(Y)=0,\mathrm{V}(Y)=1 \tag{5.7.174}$$

Proof 5.43.1. For the proof we just use the definitions of the expected mean and variance (using Huygens theorem for this latter). So let us begin with the mean:

$$\mathrm{E}(Y)=\mathrm{E}\left(\frac{X-\mu}{\sigma}\right)=\frac{1}{\sigma}\mathrm{E}\left(X-\mu\right)=\frac{1}{\sigma}(\mathrm{E}(X)-\mathrm{E}(\mu))=\frac{1}{\sigma}(\mu-\mu)=0=\mu_Y \tag{5.7.175}$$

And now with the variance using the Huygens theorem:

$$\begin{aligned}\mathrm{V}(Y)&=\mathrm{E}(Y^2)-\mathrm{E}(Y)^2=\mathrm{E}(Y^2)-\mu_Y^2=\mathrm{E}(Y^2)-0^2=\mathrm{E}(Y^2)\\&=\mathrm{E}\left(\left(\frac{X-\mu}{\sigma}\right)^2\right)=\mathrm{E}\left(\frac{1}{\sigma^2}\left(X^2-2X\mu+\mu^2\right)\right)=\frac{1}{\sigma^2}\left(\mathrm{E}(X^2)-2\mu\mathrm{E}(X)+\mu^2\right)\\&=\frac{1}{\sigma^2}\left(\mathrm{E}(X^2)-2\mu\mu+\mu^2\right)=\frac{1}{\sigma^2}\left(\mathrm{E}(X^2)-\mu^2\right)=\frac{1}{\sigma^2}\left(\left(\mathrm{V}(X)+\mathrm{E}(X)^2\right)-\mu^2\right)\\&=\frac{1}{\sigma^2}\left(\mathrm{V}(X)+\mu^2-\mu^2\right)=\frac{1}{\sigma^2}\mathrm{V}(X)=\frac{1}{\sigma^2}\sigma^2=1\end{aligned} \tag{5.7.176}$$

■ Q.E.D.

Thus, any statistical distribution defined by a mean and standard deviation can be transformed into another distribution often easier to analyse statistical. Therefore making this transformation, we obtain a random variable for which the parameters of the distribution low are now useless to know. When we do that with other laws, and in the general case, when we speak of "pivotal variables".

Here are some important mathematical properties of the variance:

P1. Multiplication by a constant (scaling upon multiplication):

$$\mathrm{V}(\alpha X)=\sum_i f(x_i)(\alpha x_i-\alpha\mu)^2=\alpha^2\sum_i f(x_i)(x_i-\mu)^2=\alpha^2\mathrm{V}(X) \tag{5.7.177}$$

P2. Sum of two random variables (additivity under dependence):

$$\begin{aligned}
\mathrm{V}(X+Y) &= \sum_i \sum_j \left((x_i+y_i)-(\mu_X+\mu_Y)\right)^2 f(x_i,y_i)\\
&= \sum_i \sum_j \left((x_i-\mu_X)+(y_i-\mu_Y)\right)^2 f(x_i,y_i)\\
&= \sum_i \sum_j \left((x_i-\mu_X)^2+2(x_i-\mu_X)(y_i-\mu_Y)+(y_i-\mu_Y)^2\right)^2 f(x_i,y_i)\\
&= \sum_{i,j}(x_i-\mu_X)^2 f(x_i,y_i)+2\sum_{i,j}(x_i-\mu_X)(y_i-\mu_Y)f(x_i,y_i)\\
&+\sum_{i,j}(y_i-\mu_Y)^2 f(x_i,y_i)\\
&= \mathrm{V}(X)+2\sum_{i,j}(x_i-\mu_X)(y_i-\mu_Y)f(x_i,y_i)+\mathrm{V}(Y)\\
&= \mathrm{V}(X)+2\mathrm{E}\left[(X-\mu_X)(Y-\mu_Y)\right]+\mathrm{V}(Y)\\
&:= \mathrm{V}(X)+2\mathrm{cov}(X,Y)+\mathrm{V}(Y)
\end{aligned} \tag{5.7.178}$$

Where we meet for the first time the concept of "covariance" denoted by cov(). We will see further below at page 600 the general case to n random variables!

Obviously under independence we have $\mathrm{V}(X+Y)=\mathrm{V}(X)+\mathrm{V}(Y)$.

P3. Affine transformation:

$$\begin{aligned}
\mathrm{V}(\alpha+\beta X) &= \mathrm{E}\left[(\alpha+\beta X-\mathrm{E}(\alpha+\beta X))^2\right]\\
&= \mathrm{E}\left[(\alpha+\beta X-\alpha+\beta\mathrm{E}(X))^2\right]\\
&= \mathrm{E}\left[\beta^2(X-\mathrm{E}(X))^2\right]\\
&= \beta^2\mathrm{E}\left[(X-\mathrm{E}(X))^2\right]\\
&= \beta^2\mathrm{V}(X)
\end{aligned} \tag{5.7.179}$$

without too much surprises... The fact that $\mathrm{V}(\alpha+X)=\mathrm{V}(X)$ is named "invariance under addition" and the one that $\mathrm{V}(\beta X)=\beta^2\mathrm{V}(X)$ is named "scaling upon multiplication".

P4. Product of two random variables (using the Huygens theorem):

$$\mathrm{V}(X\cdot Y)=\mathrm{E}\left((XY)^2\right)-\mathrm{E}(XY)^2=\mathrm{E}(X^2Y^2)-\mathrm{E}(XY)^2 \tag{5.7.180}$$

And if the two random variables are independent, we get:

$$\mathrm{V}(X\cdot Y)=\mathrm{E}\left(X^2\right)\mathrm{E}\left(Y^2\right)-\mathrm{E}\left(X\right)^2\mathrm{E}\left(Y\right)^2 \tag{5.7.181}$$

What we can rewrite using once again the Huygens theorem:

$$\mathrm{V}(X\cdot Y)=(\mathrm{V}(X)-\mathrm{E}(X)^2)(\mathrm{V}(Y)-\mathrm{E}(Y)^2)-\mathrm{E}\left(X\right)^2\mathrm{E}\left(Y\right)^2 \tag{5.7.182}$$

We will assume as obvious that these four properties extend to the continuous case!

5.7.3.1.1.1 Semivariance

The "semi-variance" is an indicator that may have been proposed by Harry Markowitz instead of the variance for portfolio optimization especially.

Analyses in the labs may have shown that portfolios optimized for semivariance minimization (see section Economy page 7698) have better performance than those based on variance. Indeed, the variance considers extremes of positive and negative returns as having equal undesirability, whereas as we shall see, semi-variance considers only the negative part. On the other hand, if the distribution of returns tends to a perfect symmetry, the two methods tend (obviously) to give exactly the same result.

Definition (#5.7.129): The "semi-variance", also named "Lower Partial Standard Deviation" (LSPD), is a measure of the dispersion only of the realization of a random variables X which are less than the expected mean of this same random variable according to:

$$\boxed{\mathrm{V}_{\text{semi}}=\hat{\sigma}^2_{\text{semi}}=\mathrm{E}\left(\min(X-\mathrm{E}(X),0)^2\right)} \tag{5.7.183}$$

or denoted in the common discrete biased case, as standard deviation:

$$\sigma_{\text{semi}}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\min(x_i-\hat{\mu}_X,0)^2\right)} \tag{5.7.184}$$

Remark

By the way the following trap must be avoid:

$$\mathrm{E}\left(\min(X-\mathrm{E}(X),0)^2\right)\neq\mathrm{V}\left(\min(X-\mathrm{E}(X),0)^2\right) \tag{5.7.185}$$

There trough this definition we see that the semivariance is similar to variance. However, it only considers observations below the mean. A useful tool in portfolio or asset analysis, semivariance provides a measure for "downside risk". While standard deviation and variance provide measures of volatility, semivariance only looks at the negative fluctuations of an asset. By neutralizing all values above the mean, or an investor's target return, semivariance estimates the average loss that a portfolio could incur.

For risk averse investors, solving for optimal portfolio allocations by minimizing semivariance would limit the likelihood of a large loss.

✍**Example:**

Let us consider the small following table:

Date	**Price**	**Return (X)**	$(\min(X-\mathbf{E}(X),0))^2$
March	20		
April	19	-5%	$(-4.6\%)^2$
May	17.10	-10%	$(-9.6\%)^2$
June	18.81	$+10\%$	0^2
July	19.25	$+5\%$	0^2
August	19.36	-1.99%	$(-1.59\%)^2$
Arithmetic mean:		$\mathrm{E}(X)\cong 0.4\%$	

The semi-variance value is then:

$$\mathrm{V}_{\text{semi}}=\mathrm{E}\left(\min(X-\mathrm{E}(X),0)^2\right)=\frac{1}{5}\left[(-4.6\%)^2+(-9.6\%)^2+(-1.59\%)^2\right]\cong 0.232\% \tag{5.7.186}$$

5.7.3.1.2 Discrete Covariance

We have seen in one of the last equations earlier above the concept of "covariance" for which we will determine a more convenient expression later:

$$\boxed{\mathrm{cov}(X,Y)=\mathrm{c}_{X,Y}=\mathrm{E}\left[(X-\mu_X)(Y-\mu_Y)\right]} \tag{5.7.187}$$

The concept of covariance can be understood geometrically as illustrated below (we will define further below the concept of correlation abbreviated Corr in the figure):

- The rectangle is green when either:
  - Both x and y are greater than their respective means
  - Both x and y are less than their respective means
- The rectangle is red when either:
  - x is greater than μ_X but y is less than μ_Y
  - x is less than μ_X but y is greater than μ_Y

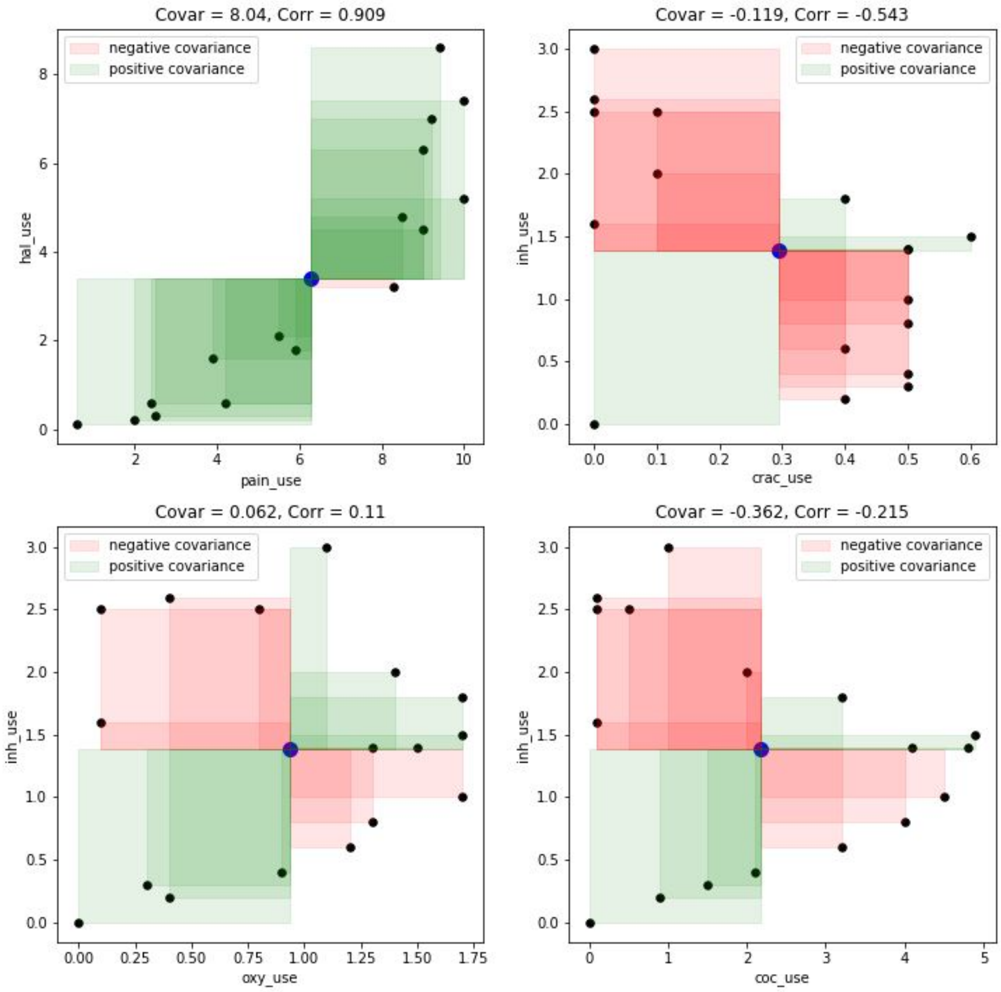


Figure 5.112: **Geometric interpretation of covariance**

We introduce now a more general and very important expression of the covariance in many application fields:

$$\begin{aligned}
\mathrm{V}(X+Y+Z) &= \sum_i\sum_j\sum_k ((x_i+y_j+z_k)-(\mu_X+\mu_Y+\mu_Z))^2 f(x_i,y_j,z_k)\\
&= \sum_i\sum_j\sum_k ((x_i-\mu_X)+(y_j-\mu_Y)+(z_k-\mu_Z))^2 f(x_i,y_j,z_k)\\
&= \sum_i\sum_j\sum_k (A+B+C)^2 f(x_i,y_j,z_k)\\
&= \sum_i\sum_j\sum_k (A^2+B^2+C^2+2AB+2BC+2AC) f(x_i,y_j,z_k)\\
&= \mathrm{V}(X)+\mathrm{V}(Y)+\mathrm{V}(Z)+2\sum_i\sum_j\sum_k (AB+BC+AC) f(x_i,y_j,z_k)\\
&= \mathrm{V}(X)+\mathrm{V}(Y)+\mathrm{V}(Z)+2\mathrm{cov}(X,Y)+2\mathrm{cov}(Y,Z)+2\mathrm{cov}(X,Z)
\end{aligned} \tag{5.7.188}$$

Remark

Some readers may have noticed that with 1 variable, we have 0 covariance terms, with 2 variables we have 1 covariance term, with 3 variables we have $2+2+2=6$ covariance terms, and continuing we would see that we have in the general case $n(n-1)$ covariance terms for n variables.

Now we change the notation to simplify even more:

$$\begin{aligned}\mathrm{V}(X_1+X_2+X_3)&=\mathrm{V}(X_1)+\mathrm{V}(X_2)+\mathrm{V}(X_3)+2\mathrm{cov}(X_1,X_2)+2\mathrm{cov}(X_2,X_3)+2\mathrm{cov}(X_1,X_3)\\&=\sum_{i=1}^{3}\mathrm{V}(X_i)+2\sum_{i<j}^{3}\mathrm{cov}(X_i,X_j)\end{aligned} \tag{5.7.189}$$

Therefore in the general case:

$$\boxed{\mathrm{V}\left(\sum_{i=1}^{n}X_i\right)=\sum_{i=1}^{n}\mathrm{V}(X_i)+2\sum_{i<j}^{n}\mathrm{cov}(X_i,X_j)} \tag{5.7.190}$$

Or using standard deviation:

$$\boxed{\sigma_{\sum_{i=1}^{n}X_i}=\sqrt{\sum_{i=1}^{n}\sigma_{X_i}^2+2\sum_{i<j}^{n}\mathrm{cov}(X_i,X_j)}} \tag{5.7.191}$$

Using the properties of the mean (especially $\mathrm{E}(X)=c^{te}$ and $\mathrm{E}(c^{te})=c^{te}$) we can write the covariance in a much simpler way for calculation purposes:

$$\begin{aligned}\mathrm{cov}(X,Y)&=\mathrm{E}\left[(X-\mathrm{E}(X))(Y-\mathrm{E}(Y))\right]=\mathrm{E}\left[XY-\mathrm{E}(X)Y-X\mathrm{E}(Y)+\mathrm{E}(X)\mathrm{E}(Y)\right]\\&=\mathrm{E}(XY)-\mathrm{E}(\mathrm{E}(X)Y)-\mathrm{E}(X\mathrm{E}(Y))+\mathrm{E}(\mathrm{E}(X)\mathrm{E}(Y))\\&=\mathrm{E}(XY)-\mathrm{E}(X)\mathrm{E}(Y)-\mathrm{E}(X)\mathrm{E}(Y)+\mathrm{E}(X)\mathrm{E}(Y)\\&=\mathrm{E}(XY)-\mathrm{E}(X)\mathrm{E}(Y)\end{aligned} \tag{5.7.192}$$

and we obtain the relation widely used in statistics and finance in practice named the "covariance formula"...:

$$\boxed{\mathrm{c}_{X,Y}=\mathrm{cov}(X,Y)=\mathrm{E}(XY)-\mathrm{E}(X)\mathrm{E}(Y)} \tag{5.7.193}$$

which is however best known when written as:

$$\mathrm{c}_{X,Y}=\mathrm{cov}(X,Y)=\frac{1}{n}\sum_{i}x_iy_i-\bar{x}\bar{y} \tag{5.7.194}$$

If $X=Y$ (equivalent to a univariate covariance) we fall back again on the Huygens theorem:

$$\boxed{\mathrm{c}_{X,X}=\mathrm{cov}(X,X)=\mathrm{E}(XX)-\mathrm{E}(X)\mathrm{E}(X)=\mathrm{E}(X^2)-\mathrm{E}(X)^2} \tag{5.7.195}$$

Remark

Statistics can be partitioned according to the number of random variables we study. Thus, when a single random variable is studied, we speak of "univariate statistics", for two random variables of "bivariate statistics" and in general, of "multivariate statistics".

If and only if the variables are equally likely, we find the covariance in the literature in the following form, sometimes named "Pearson covariance", which derives from calculations that we have done previously with the mean:

$$\mathrm{c}_{X,Y}=\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu_Y)(x_i-\mu_X) \tag{5.7.196}$$

Covariance is a measure of the simultaneous variation of X and Y. Indeed, if X and Y generally grow simultaneously, the products $(y_i-\mu_Y)(x_i-\mu_X)$ will be positive (positively correlated), whereas if Y decreases as X increases, these same products will be negative (negative correlation).

Note that if we distribute the terms of the last equation, we have:

$$\begin{aligned}\mathrm{c}_{X,Y}=\mathrm{cov}(X,Y)&=\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu_Y)(x_i-\mu_X)=\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})\\&=\frac{1}{n}\sum_{i=1}^{n}x_i(y_i-\bar{y})-\bar{x}(y_i-\bar{y})=\frac{1}{n}\left(\sum_{i=1}^{n}x_i(y_i-\bar{y})-\bar{x}\sum_{i=1}^{n}(y_i-\bar{Y})\right)\end{aligned} \tag{5.7.197}$$

and we have already shown that the sum of the deviations from the mean is zero. Hence we get another common way to write the covariance:

$$\mathrm{c}_{X,Y}=\mathrm{cov}(X,Y)=\frac{1}{n}\sum_{i=1}^{n}x_i(y_i-\bar{y}) \tag{5.7.198}$$

and by symmetry:

$$\mathrm{c}_{X,Y}=\mathrm{cov}(X,Y)=\frac{1}{n}\sum_{i=1}^{n}y_i(x_i-\bar{x}) \tag{5.7.199}$$

So in the end, in the equiprobable case, we finally have the equivalent three important relations used in various sections of this textbook:

$$\boxed{\begin{aligned}\mathrm{c}_{X,Y}&=\mathrm{cov}(X,Y)=\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})\\\mathrm{c}_{X,Y}&=\mathrm{cov}(X,Y)=\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\bar{x}\bar{y}\\\mathrm{c}_{X,Y}&=\mathrm{cov}(X,Y)=\frac{1}{n}\sum_{i=1}^{n}x_i(y_i-\bar{y})=\frac{1}{n}\sum_{i=1}^{n}y_i(x_i-\bar{x})\end{aligned}} \tag{5.7.200}$$

In the section Theoretical Computing for the study of linear regression (page 6324) and factor analysis we will need the explicit expression of the bilinearity property of the variance. To see what it is exactly, consider three random variables X, Y and Z and two constants a and b. Then using the third relation given above, we have:

$$\begin{aligned}\text{cov}\,(Y, aX+bZ) &= \frac{1}{n}\sum_i (y_i-\bar{y})(ax_i+bz_i) = \frac{1}{n}\sum_i \left[(y_i-\bar{y})ax_i+(y_i-\bar{y})bz_i\right]\\ &= \frac{1}{n}\left[\sum_i (y_i-\bar{y})ax_i+\sum_i (y_i-\bar{y})bz_i\right]\\ &= a\frac{1}{n}\sum_i (y_i-\bar{y})x_i+b\frac{1}{n}\sum_i (y_i-\bar{y})z_i\\ &= a\text{cov}(X,Y)+b\text{cov}(Y,Z)\end{aligned} \tag{5.7.201}$$

The last relation is also important and will be used in several sections of this textbook (Economy page 7485, Numerical Methods page 6272). It also allows us to directly obtain the covariance for the sums of various random variables.

✍**Example:**

If X, Y, Z, T are four random variables defined on the same population, we want to compute the following covariance:

$$\text{cov}(3X+5Y, 4Z-2T) \tag{5.7.202}$$

We will develop that in two phases (this is also why we name that "bilinearity property"). First with respect to the second argument (random choice!):

$$\text{cov}(3X+5Y, 4Z-2T) = 4\text{cov}(3X+5Y, Z) - 2\text{cov}(3X+5Y, T) \tag{5.7.203}$$

And then with respect to the first:

$$\text{cov}(3X+5Y, 4Z-2T) = 4\left[3\text{cov}(X,Z)+5\text{cov}(Y,Z)\right]-2\left[3\text{cov}(X,T)+5\text{cov}(Y,T)\right] \tag{5.7.204}$$

So in the end:

$$\text{cov}(3X+5Y, 4Z-2T) = 12\text{cov}(X,Z)+20\text{cov}(Y,Z)-6\text{cov}(X,T)-10\text{cov}(Y,T) \tag{5.7.205}$$

There is an elegant and simple method to remember how to calculate the covariance of two random variables. Indeed, remembering that we have proved just earlier that:

$$\begin{aligned}\text{V}(X+Y) &= \text{V}(X)+\text{V}(Y)+2\text{cov}(X,Y)\\ \text{V}(X-Y) &= \text{V}(X)+\text{V}(Y)-2\text{cov}(X,Y)\end{aligned} \tag{5.7.206}$$

It comes by subtracting these two relations and simplifying / reordering:

$$\boxed{\text{cov}(X,Y) = \frac{\text{V}(X+Y)-\text{V}(X-Y)}{4}} \tag{5.7.207}$$

and that's it! A simple way to remember the compute the covariance.

Now, consider a set of random vectors $\vec{X}_i := X_i$ of components $(x_1, x_2, \ldots, x_n)_i$. The calculation of the covariance of the components by pairs gives what is named the "covariance matrix" (a tool widely used in finance, management and statistical numerical methods!).

Indeed, we define the component (m, n) of the covariance matrix by:

$$\text{cov}_{X_m,X_n} = \text{E}\left[(X_m - \mu_{X_m})(X_n - \mu_{X_n})\right] := \text{c}_{m,n} \tag{5.7.208}$$

We can therefore write a symmetric matrix (usually in practice it must be a square matrix...) in the form:

$$\Sigma = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix} \tag{5.7.209}$$

where Σ is the usual tradition letter to denote the "covariance matrix".

By symmetry and because it is a square n by n matrix only the number $\dfrac{n(n+1)}{2}$ of components is useful for us to determine the whole matrix (trivial but important information for when we will study the structural equation modelling in the Numerical Methods section).

This matrix has the remarkable property that if we take the set of all random vectors and we calculate the covariance matrix, then the diagonal will give us obviously the variances of each pair of vectors (see examples in the chapters Economics, Numerical Methods or Industrial Engineering) because we have for recall:

$$\text{cov}_{X_m,X_m} = \text{E}\left[(X_m - \mu_{X_m})(X_n - \mu_{X_m})\right] = \text{E}\left[(X_m - \mu_{X_m})^2\right] = \text{V}(X_m) = \sigma^2_{X_m} \tag{5.7.210}$$

Hence the name "variance-covariance matrix":

$$\Sigma = \begin{bmatrix} \text{V}_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & \text{V}_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & \text{V}_{nn} \end{bmatrix} = \begin{bmatrix} \sigma^2_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & \sigma^2_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & \sigma^2_{nn} \end{bmatrix} \tag{5.7.211}$$

And this is a little bit abusively sometimes written as:

$$\Sigma = \begin{bmatrix} \sigma^2_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma^2_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma^2_{nn} \end{bmatrix} \tag{5.7.212}$$

This matrix has the advantage of quickly showing what pairs of random variables have a negative covariance and there... for which random variable the variance of the sum is smaller than the sum of the variances!

Remark

As we already mention it, this matrix is very important and we will often see it again in the section Economy during our study of modern portfolio theory (page 7698) and also for data mining techniques in the section of Theoretical Computing (page 6768) and also for principal components analysis (page 1213) and also in Industrial Engineering during our study of bivariate control charts (page 8934).

Recall now that we have an axiom in probability (see section Probabilities page 476) which lead us to the fact that two events A and B are independent if and only if:

$$P(A\cap B)=P(A)P(B) \tag{5.7.213}$$

Similarly, by extension, we define the independence of discrete random variables.

Definition (#5.7.130): Let X,Y be two discrete random variables. We say that X,Y are independent if and only if:

$$\forall x,y\in\mathbb{R}\quad P(X=x,Y=y)=P(X=x)P(Y=y) \tag{5.7.214}$$

More generally, the discrete variables $X_1,X_2,\ldots,X_n$ are independent (in block) if:

$$\forall x_1,\ldots,x_n\in\mathbb{R}\quad P(X_1=x_1,X_2=x_2,\ldots,X_3=x_3)=\prod_{i=1}^{n}P(X_i=x_i) \tag{5.7.215}$$

Theorem 5.44. The independence of two random variables implies that their covariance is zero (the opposite is false!).

Proof 5.44.1. We will prove this in the case where the random variables take only a finite number of values $\{x_i\}_I$ and $\{y_j\}_J$, respectively, with I,J finite sets.

For the proof let us recall that:

$$\begin{aligned}\mathrm{E}(XY)&=\sum_{i,j}P(X=x_i,Y=y_j)x_iy_j=\sum_{i,j}P(X=x_i)P(Y=y_j)x_iy_j\\&=\sum_{i}P(X=x_i)x_i\sum_{j}P(Y=y_j)y_j=\mathrm{E}(X)\mathrm{E}(Y)\end{aligned} \tag{5.7.216}$$

and therefore:

$$\mathrm{c}_{X,Y}=\mathrm{E}(XY)-\mathrm{E}(X)\mathrm{E}(Y)=\mathrm{E}(X)\mathrm{E}(Y)-\mathrm{E}(X)\mathrm{E}(Y)=0 \tag{5.7.217}$$

Remark

So smaller is the covariance (near to zero), more the series are independent. Conversely, the greater the covariance (in absolute value) higher the series are dependent.

Given that:

$$\mathrm{V}(X+Y)=\mathrm{V}(X)+\mathrm{V}(Y)+2c_{X,Y} \tag{5.7.218}$$

and the fact that if X and Y are independent we have $c_{X,Y}=0$. Then:

$$\mathrm{V}(X+Y)=\mathrm{V}(X)+\mathrm{V}(Y) \tag{5.7.219}$$

More generally if $X_1, \ldots, X_n$ are independent (in block) then for any discrete or continuous statistical distribution law (!) we have using the two most common notations:

$$\mathrm{V}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{V}(X_i) \tag{5.7.220}$$

Or using the standard deviation:

$$\sigma_{\sum_{i=1}^{n} X_i} = \sqrt{\sum_{i=1}^{n} \sigma_{X_i}^2} \tag{5.7.221}$$

■ Q.E.D.

There is another important relation that we will need for our study of Data Mining/Machine Learning and especially to compare different models and involving the covariance matrix.

Let $X = (X_i)$ be an $n \times 1$ random vector and let A be an $n \times n$ symmetric matrix. If $\mathrm{E}(X) = \mu$ and $\mathrm{V}(X) = \Sigma = \sigma_{ij}$ then $\mathrm{E}[X^T AX] = \mathrm{tr}(A\sigma) + \mu^T A\mu$ (keeping in mind that $X^T AX$ is a 1×1 matrix, or in other words: a scalar!).

To prove this first notice that:

$$X^T AX = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} X_i X_j. \tag{5.7.222}$$

So what is its expectation? Clearly we have first by using the linearity proper of the expected mean:

$$\mathrm{E}[X^T AX] = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} \mathrm{E}[X_i X_j] \tag{5.7.223}$$

Applying the covariance relation proved just earlier above:

$$\mathrm{cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) \Leftrightarrow \mathrm{E}(XY) = \mathrm{cov}(X,Y) + \mathrm{E}(X)\mathrm{E}(Y) \tag{5.7.224}$$

then we get:

$$\mathrm{E}[X^T AX] = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\mathrm{E}[X_i X_j] = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}(\sigma_{ij} + \mu_i\mu_j) \tag{5.7.225}$$

As Σ is symmetric:

$$\begin{aligned}\mathrm{E}[X^T AX] &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}(\sigma_{ij} + \mu_i\mu_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\sigma_{ji} + \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\mu_i\mu_j \\ &= \sum_{i=1}^{n} [A\Sigma]_{ii} + \mu^T A\mu = \mathrm{tr}(A\Sigma) + \mu^T A\mu\end{aligned} \tag{5.7.226}$$

The following relation will be important to us during our study of the Akaike information criterion (see section Numerical Methods page 6420):

$$\boxed{\mathrm{E}[X^T AX] = \mathrm{tr}(A\Sigma) + \mu^T A\mu} \tag{5.7.227}$$

Notice that the notation above is common in the field of statistics but more rigorously we should write:

$$\boxed{\mathrm{E}[\vec{x}^T A\vec{x}] = \mathrm{tr}(A\Sigma) + \vec{\mu}^T A\vec{\mu}} \tag{5.7.228}$$

We need also another property of the covariance matrix for our study of Akaike information criterion (see section Numerical Methods page 6420). Let us consider $\vec{x} \in \mathbb{R}^p$ be a random (column) vector with covariance matrix $\sigma \in \mathbb{R}^{p\times p}$. Let $A \in \mathbb{R}^{p\times p}$ be a non-random matrix. Then:

$$\begin{aligned}
\mathrm{V}(A\vec{x}) &= \mathrm{E}[(A\vec{x} - \mathrm{E}(A\vec{x}))(A\vec{x} - \mathrm{E}(A\vec{x}))^T]\\
&= \mathrm{E}[(A\vec{x} - A\mathrm{E}(\vec{x}))(A\vec{x} - A\mathrm{E}(\vec{x}))^T]\\
&= \mathrm{E}[(A\vec{x} - A\mathrm{E}(\vec{x}))(\vec{x}^T A^T - \mathrm{E}(\vec{x})^T A^T)]\\
&= \mathrm{E}[A(\vec{x} - \mathrm{E}(\vec{x}))(\vec{x} - \mathrm{E}(\vec{x}))^T A^T]\\
&= A\,\underbrace{\mathrm{E}[(\vec{x} - \mathrm{E}(\vec{x}))(\vec{x} - \mathrm{E}(\vec{x}))^T]}_{=\Sigma}\,A^T\\
&= A\Sigma A^T
\end{aligned} \tag{5.7.229}$$

We get the same result if we replace the vector $\vec{x}$ by a matrix X such that:

$$\mathrm{V}(AX) = A\Sigma A^T \tag{5.7.230}$$

5.7.3.1.2.1 Quasi-correlation matrix

In practice, some financial analysts estimate qualitatively the matrix of correlations based on their experience, we then talk about the "quasi-correlation matrix", and remember that the variance-covariance matrix and the matrix of correlations are the same thing but with two different perspectives, however it is not possible to make an estimate randomly (without rules).

Indeed, let us consider n real random variables $X_1, \ldots, X_n$ of respective means $\mu_1, \ldots, \mu_n$. Let us denote exceptionally by Σ the symmetric matrix of variances-covariances defined by:

$$\Sigma = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n}\\ c_{21} & c_{22} & \cdots & c_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ c_{n1} & c_{n2} & \cdots & c_{nn}\end{bmatrix} = \begin{bmatrix} \mathrm{V}(X_1) & \mathrm{cov}(X_1,X_2) & \cdots & \mathrm{cov}(X_1,X_n)\\ \mathrm{cov}(X_2,X_1) & \mathrm{V}(X_2) & \cdots & \mathrm{cov}(X_2,X_n)\\ \vdots & \vdots & \ddots & \vdots\\ \mathrm{cov}(X_n,X_1) & \mathrm{cov}(X_n,X_2) & \cdots & \mathrm{V}(X_n)\end{bmatrix} \tag{5.7.231}$$

Let us prove that Σ is positive semi-definite (see section Linear Algebra page 1957):

Proof 5.44.2. Given $\vec{\alpha} = (\alpha_1, \ldots, \alpha_n)^T$ a vector of $\mathbb{R}^n$. We have:

$$\begin{aligned}
\mathrm{V}\left(\sum_{i=1}^n \alpha_i X_i\right) &= \mathrm{E}\left(\left(\sum_{i=1}^n \alpha_i(X_i - \mu_i)\right)^2\right) = \mathrm{E}\left(\sum_{i,j} \alpha_i\alpha_j(X_i - \mu_i)(X_j - \mu_j)\right)\\
&= \sum_{,j} \alpha_i\alpha_j \mathrm{E}\left((X_i - \mu_i)(X_j - \mu_j)\right) = \sum_{i,j} \alpha_i\alpha_j \mathrm{cov}(X_i, X_j) = \vec{\alpha}^T\Sigma\vec{\alpha}
\end{aligned} \tag{5.7.232}$$

Therefore:

$$\vec{\alpha}^T\Sigma\vec{\alpha} \geq 0 \tag{5.7.233}$$

Since the variance is positive or zero by construction! Which means that Σ is semi-positive. This is a very important property for many statistical applications (Mahalonobis distance, Cholesky decomposition, etc.).

■ Q.E.D.

Now, by the generality of $\vec{\alpha}$, let us admit (nothing prevents us to do so!) that the latter is an eigenvector (see section Linear Algebra page 1997). Then, since:

$$\vec{\alpha}^T\Sigma\vec{\alpha} \geq 0 \tag{5.7.234}$$

and since by definition of their eigenvalues and eigenvectors:

$$\Sigma\vec{\alpha} = \lambda_i\vec{\alpha} \tag{5.7.235}$$

Then we have:

$$\vec{\alpha}^T\Sigma\vec{\alpha} = \vec{\alpha}^T\lambda_i\vec{\alpha} \geq 0 \tag{5.7.236}$$

And let us concentrate on the right term:

$$\vec{\alpha}^T\lambda_i\vec{\alpha} \geq 0 \tag{5.7.237}$$

Since the eigenvalue is a scalar we can rearrange this for example in the following form (commutativity):

$$\left(\vec{\alpha}^T\vec{\alpha}\right) \geq 0 \tag{5.7.238}$$

and as $\vec{\alpha}^T\vec{\alpha}$ is necessarily positive then the λ_i can only be positive or null.

So we got a way to check that our matrix of variances-covariances (or correlation) is indeed what it is supposed to be !!!

Moreover, as we have demonstrated in our study of the spectral theorem in the Linear Algebra section page 2004 that for a matrix such that is Σ (symmetric, real, positive semi-definite):

$$\det(\Sigma) = \prod \lambda_i \tag{5.7.239}$$

Then another way of controlling is Σ is positive semi-definite is that the determinant of the variance-covariance matrix is positive or zero. And keep in mind that if the determinant is zero, then the matrix is not invertible! Hence the variance-covariance matrix is not always invertible!

Remark

However, the reciprocal is false (obviously!), we can have a positive determinant without the matrix being positive. While for eigenvalues there is equivalence: positive matrix is equivalent to all eigenvalues are positive.

5.7.3.1.2.2 Anscombe's famous quartet

Anscombe's quartet comprises four datasets that have nearly identical elementary statistical properties, yet appear very different when graphed or analysed with undergraduate statistics rather than high-school one. Each dataset consists of eleven (x, y) points. They were constructed in 11973 (holocene calendar) by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties. This quartet is also used to test if an analytical tool can be accepted a "statistics-compliant" (as the six corresponding used statistics should be the minimum provided by any high-school level analytical tool!).

The datasets are as follows. The x values are the same for the first three datasets:

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Table 5.37: **Anscombe's quartet**

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relation, and the inadequacy of basic statistic properties for describing realistic datasets.

With Microsoft® Excel® 14.0.7166 we get:

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10.00	8.04	10.00	9.14	10.00	7.46	8.00	6.58
	8.00	6.95	8.00	8.14	8.00	6.77	8.00	5.76
	13.00	7.58	13.00	8.74	13.00	12.74	8.00	7.71
	9.00	8.81	9.00	8.77	9.00	7.11	8.00	8.84
	11.00	8.33	11.00	9.26	11.00	7.81	8.00	8.47
	14.00	9.96	14.00	8.10	14.00	8.84	8.00	7.04
	6.00	7.24	6.00	6.13	6.00	6.08	8.00	5.25
	4.00	4.26	4.00	3.10	4.00	5.39	19.00	12.50
	12.00	10.24	12.00	9.13	12.00	8.15	8.00	5.56
	7.00	4.82	7.00	7.26	7.00	6.42	8.00	7.91
	5.00	5.68	5.00	4.74	5.00	5.73	8.00	6.89
Mean	9.0000	7.5009	9.0000	7.5009	9.0000	7.5000	9.0000	7.5009
Median	9.0000	7.5800	9.0000	8.1400	9.0000	7.1100	8.0000	7.0400
Variance	10.0000	3.7521	10.0000	3.7524	10.0000	3.7478	10.0000	3.7484
Skewness	0.0000	−0.0650	0.0000	−1.3158	0.0000	1.8555	3.3166	1.5068
Kurtosis	1.2000	−0.5349	−1.2000	0.8461	1.2000	4.3841	11.0000	3.1513
Correlation	0.8164		0.8162		0.8164		0.8165	

Table 5.38: **Anscombe's quartet statistics summary**

As we can see with elementary statistical indicators it is almost impossible to guess a difference between the four data sets. But if we use the skewness or the kurtosis (see page 636) this change everything! The mutual information criterion (see section Statistical Mechanics page 3529) can also discriminate between the four plots.

Looking to the corresponding charts we get the same conclusion:

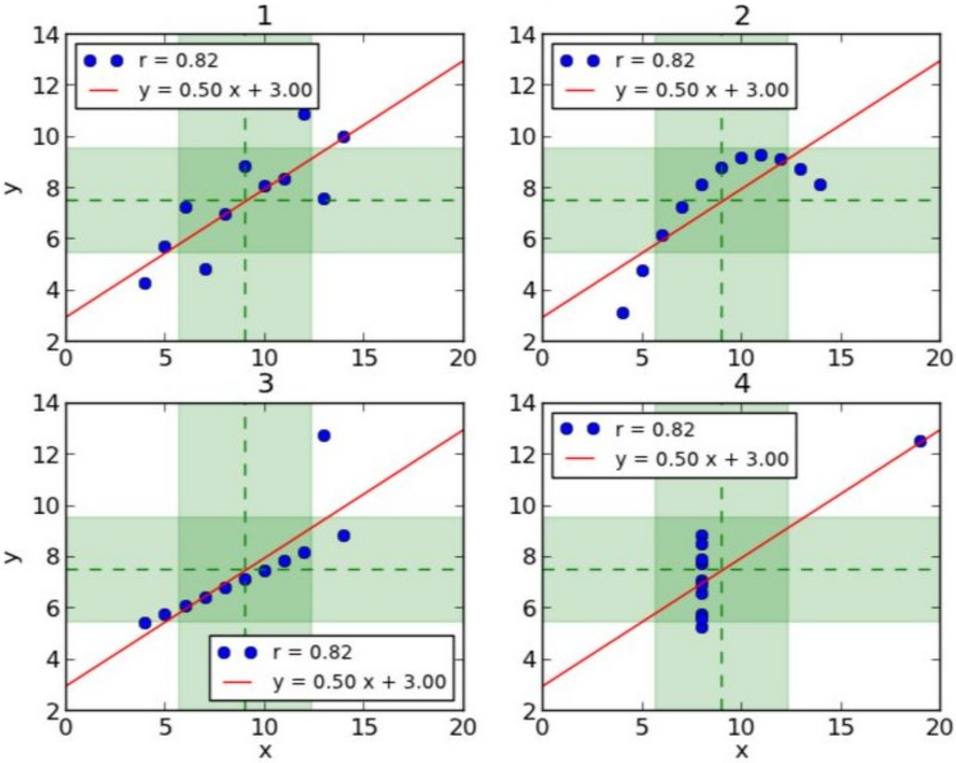


Figure 5.113: **Anscombe's quartet graphs summary**

5.7.3.1.3 Mean and Variance of the Average

Often in statistics, it is (verrrrry!) useful to determine the standard deviation of the sample mean and to work with it to get important analytical results in management and manufacturing. Let's see what it is!

Given the average of a series of terms, each determined by the measurement of several values (it is in fact its estimator in a particular case as we will see later):

$$\bar{X}=\frac{1}{n}\left(X_1+X_2+\ldots+X_n\right) \tag{5.7.240}$$

then using the properties of the mean:

$$\mathrm{E}\left(\bar{X}\right)=\frac{1}{n}\left(\mathrm{E}(X_1)+\mathrm{E}(X_2)+\ldots+\mathrm{E}(X_n)\right) \tag{5.7.241}$$

and if all the random variables are independent and identically distributed then we have:

$$\mathrm{E}\left(\bar{X}\right)=\frac{1}{n}\left(\mu+\mu+\ldots+\mu\right)=\frac{1}{n}n\mu=\mu \tag{5.7.242}$$

Remark

We will prove much further below that if all the random variables are independent and identically distributed with finite variance, then the mean follows asymptotically what we name a "Normal distribution".

For the variance, the same reasoning applies:

$$\mathrm{V}(\bar{X})=\sigma_{\bar{X}}^2=\frac{1}{n^2}\left(\mathrm{V}(X_1)+\mathrm{V}(X_2)+\ldots+\mathrm{V}(X_n)\right)=\frac{1}{n^2}\left(\sigma_1^2+\sigma_2^2+\ldots+\sigma_2^n\right) \tag{5.7.243}$$

And if the random variables are independent and identically distributed (we will study further the very important case current in practice where the last condition is not satisfied):

$$\boxed{\mathrm{V}(\bar{X})=\sigma_{\bar{X}}^2=\frac{\sigma^2}{n}} \tag{5.7.244}$$

Then we get the "standard deviation of the mean" also named "standard error" or "non-systematic variation":

$$\boxed{\mathrm{SE}=\sigma_{\bar{X}}=\frac{\sigma}{\sqrt{n}}} \tag{5.7.245}$$

and this is strictly the standard deviation of the estimator of mean!

The more intuitive form to express the Standard Error in terms of percent for non-analytical workers, managers and chief executives is named "Relative Standard Error (RSE)" which is the expression of the Standard Error as percent, that is:

$$\boxed{\mathrm{RSE}=\frac{\sigma_{\bar{X}}}{\bar{X}}} \tag{5.7.246}$$

The latter is quite useful when we have to deal with many variables with different units!!

The value of $\sigma_{\bar{X}}$ is available in many softwares including Microsoft® Excel® charts (but there is no built-in function in Microsoft® Excel®) and is written with the standard deviation (as above) or with the notation of the variance (then we only have to take the square root...).

Note that the last relation can be used even if the average of n random variables is not the same! The main condition is just that the standard deviations are all equal and this is the case in the industry (production).

We then have:

$$\boxed{\begin{aligned} &\text{E}(S_n)=n\mu && \text{E}(M_n)=\mu \\ &\text{V}(S_n)=\sigma_{S_n}^2=n\sigma^2 && \text{V}(M_n)=\sigma_{M_n}^2=\frac{\sigma^2}{n} \end{aligned}} \tag{5.7.247}$$

where S_n is the sum of n independent identically distributed random variables and M_n their estimated average.

The reduced centered variable that we introduced earlier:

$$Y=\frac{X-\mu}{\sigma} \tag{5.7.248}$$

can then be written in several very useful ways:

$$\boxed{Y=\frac{S_n-n\mu}{\sigma_{S_n}\sqrt{n}}=\frac{M_n-\mu}{\sigma_{M_n}/\sqrt{n}}} \tag{5.7.249}$$

Furthermore, assuming that the reader already knows what is a Normal distribution $\mathcal{N}(\mu,\sigma)$, we will show later in detail because it is extremely important (!) that the probability of the random variable $\bar{X}$, average of n identically distributed and linearly independent random variables, has for law (obviously):

$$\mathcal{N}\left(\mu,\frac{\sigma}{\sqrt{n}}\right) \tag{5.7.250}$$

5.7.3.1.4 Coefficient of Correlation

Now consider X and Y two random variables having for covariance:

$$c_{X,Y}=\mathrm{E}\left[(X-\mu_X)(Y-\mu_Y)\right] \tag{5.7.251}$$

Theorem 5.45. We have:

$$\left(c_{X,Y}\right)^2\leq \mathrm{V}(X)\mathrm{V}(Y) \tag{5.7.252}$$

We will prove this relation immediately because the use of the covariance alone for data analysis is not always great because it is not strictly limited and easy to use (at interpretation). We will construct an indicator easier to use in business.

Proof 5.45.1. We choose any constant a and we calculate the variance of:

$$aX+Y \tag{5.7.253}$$

We can then immediately write using the properties of the variance and the of the mean:

$$\mathrm{V}(aX+Y)=a^2\mathrm{V}(X)+\mathrm{V}(Y)+2ac_{X,Y} \tag{5.7.254}$$

The right quantity is positive or null for any a by construction of the variance (left). So the discriminant of the expression (see section Calculus page 1383), seen as a polynomial in a is of the type:

$$\begin{gathered}P(x)=ax^2+bx+c=a\left[\left(x+\frac{b}{2a}\right)^2-\frac{b^2-4ac}{4a^2}\right]\\ \Rightarrow\\ P(a)=\mathrm{V}(X)a^2+2c_{X,Y}a+\mathrm{V}(Y)=\mathrm{V}(X)\left[\left(a+\frac{2c_{X,Y}}{2\mathrm{V}(X)}\right)^2-\frac{(2c_{X,Y})^2-4\mathrm{V}(X)\mathrm{V}(Y)}{4\mathrm{V}(X)^2}\right]\end{gathered} \tag{5.7.255}$$

Because $P(a)$ is positive for any a we have as only possibility that:

$$(2c_{X,Y})^2-4\mathrm{V}(X)\mathrm{V}(Y)\leq 0 \tag{5.7.256}$$

Therefore after simplification:

$$\left(c_{X,Y}\right)^2\leq \mathrm{V}(X)\mathrm{V}(Y) \tag{5.7.257}$$

■ Q.E.D.

This gives us also:

$$\frac{(c_{X,Y})^2}{\mathrm{V}(X)\mathrm{V}(Y)}\leq 1\Leftrightarrow\frac{|c_{X,Y}|}{\sqrt{\mathrm{V}(X)\mathrm{V}(Y)}}\leq 1 \tag{5.7.258}$$

Finally we get some statistical inequality (it's a special form of the Cauchy-Schwarz inequality that we will prove later in the section of Vector Calculus page 1826):

$$-1 \leq \frac{c_{X,Y}}{\sqrt{\mathrm{V}(X)\mathrm{V}(Y)}} \leq 1 \tag{5.7.259}$$

If the variances of X and Y are non-zero, the correlation between X and Y is defined by the "linear correlation coefficient" (it is a standardized covariance so that its amplitude does not depend on the chosen unit measure) and written:

$$R_{X,Y} = \frac{c_{X,Y}}{\sqrt{\mathrm{V}(X)\mathrm{V}(Y)}} \tag{5.7.260}$$

and also named "product moment correlation coefficient" or simply "correlation coefficient".

Which can also be written in an expanded form (using Huygens theorem):

$$R_{X,Y} = \frac{\mathrm{E}(XY)-\mathrm{E}(X)\mathrm{E}(Y)}{\sqrt{\mathrm{V}(X)\mathrm{V}(Y)}} = \frac{\mathrm{E}(XY)-\mathrm{E}(X)\mathrm{E}(Y)}{\sqrt{\left(\mathrm{E}(X^2)-\mathrm{E}(X)^2\right)\left(\mathrm{E}(Y^2)-\mathrm{E}(Y)^2\right)}} \tag{5.7.261}$$

or more condensed:

$$\boxed{R_{X,Y} = \frac{c_{X,Y}}{\sigma_X\sigma_Y}} \tag{5.7.262}$$

Remark

Note that normally, the letter R is reserved to say that this is an estimator of the correlation coefficient but the definition above is not an estimator (the variances doesn't have the small hat...) and that, strictly speaking, we should then write $\rho_{X,Y}$ according to the traditions of use.

Whatever the units and the orders of magnitude, the correlation coefficient is a number between -1 and $+1$ without units (so its value does not depend on the unit of measure, which is by far not the case for all statistical indicators!). It reflects more or less the linear dependence of X and Y or geometrically more or less the flatness magnitude. We can therefore say that a coefficient of correlation of zero or close to 0 correlation means that there is no linear relation between the characters. But it does not involve any notion of more general independence!

So when the correlation is zero:

$$\rho_{X,Y} = \frac{c_{X,Y}}{\sigma_X\sigma_Y} = 0 \tag{5.7.263}$$

Then:

$$c_{X,Y} = \mathrm{E}(XY)-\mathrm{E}(X)\mathrm{E}(Y) = 0 \tag{5.7.264}$$

Hence:

$$\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y) \tag{5.7.265}$$

And we know that this relations holds when random variables are independent. But think that in the general case "*uncorrelated*" means "*independence*" is totally wrong! Only independence implies uncorrelation! The opposite is not true (if X and Y are uncorrelated, then they can still be dependent)!

✍**Example:**

Consider as example of uncorrelated variables that are not independent, the following example given (useful example in finance!) by:

$$X = \mathcal{N}(0,1) \qquad Y = X^2 \tag{5.7.266}$$

Y is clearly dependent upon X. As we will prove later, during our study of the Normal centered reduced distribution, we have:

$$\mathrm{E}(X) = 0 \qquad \mathrm{E}(X^3) = 0 \tag{5.7.267}$$

And as:

$$c_{X,Y} = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) = \mathrm{E}(X \cdot X^2) - \mathrm{E}(X)\mathrm{E}(X) = 0 \tag{5.7.268}$$

Then we have indeed dependent variables that are uncorrelated!

The only general case when lack of correlation implies independence is when the joint distribution of X and Y is Gaussian!

The figure below illustrates some common cases and the last row highlights well situations with dependent but uncorrelated variables:

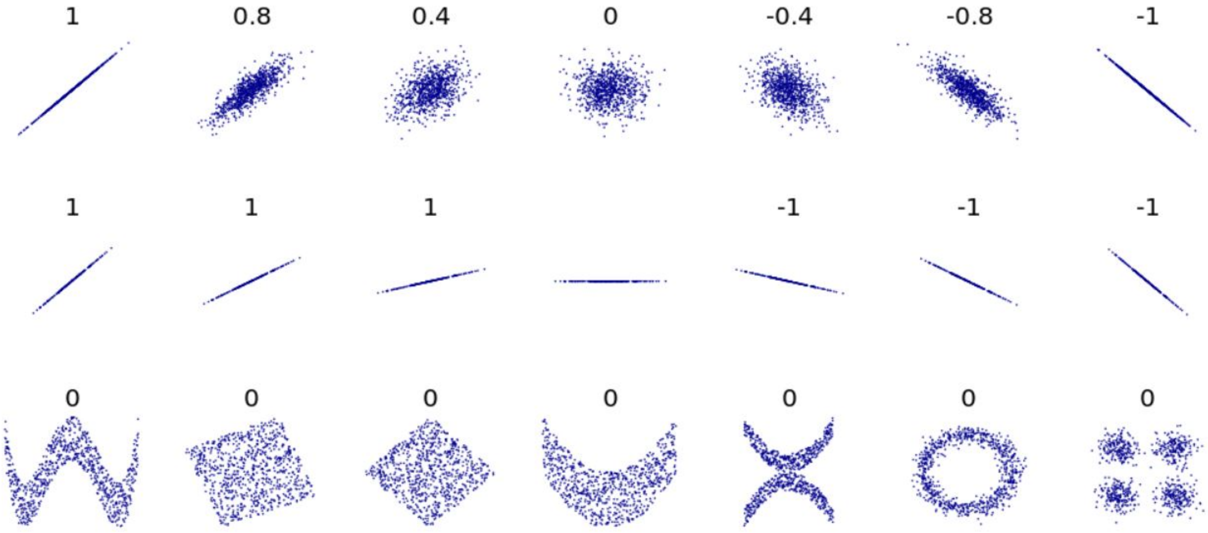


Figure 5.114: **Several sets of (x, y) points, with the corresponding Pearson correlation (source: Wikipedia, author: Denis Boigelot)**

When the correlation coefficient is near $+1$ or -1, the characters are said to be strongly correlated. We must be careful with the frequent confusion between correlation and causality. Thus, two phenomena that are correlated does not imply in any way that one is the cause of the other (this fallacy is also known as "spurious correlation" or "cum hoc ergo propter hoc[34]", latin for "with this, therefore because of this," and "false cause")!!!!

Indeed, for any two correlated events, A and B, the different possible relationships include:

- A causes B (direct causation);
- B causes A (reverse causation);
- A and B are consequences of a common cause, but do not cause each other;
- A causes B and B causes A (bidirectional or cyclic causation);
- A causes C which causes B (indirect causation);
- There is no connection between A and B;
- The correlation is a coincidence.

Remark

A situation in which a measure of association or relationship between exposure and outcome is distorted by the presence of another variable is named a "confounding". An extraneous variable that wholly or partially accounts for the observed effect is named a "confounder variable".

So we know that correlation doesn't imply necessarily causation and that in practice times we can't always have mathematical (i.e. statistical) evidence that the correlation implies a causation! However there is a consensus in the scientific field that a correlation implies a causation when the following so named "Braford Hill Criterion" are fulfilled[35]:

---

[34] *Post hoc ergo propter hoc* (latin of "after this, therefore because of this:) is an informal fallacy that states: "Since event Y followed event X, event Y must have been caused by event X." It is often shortened simply to *post hoc fallacy*. A logical fallacy of the questionable cause variety, it is subtly different from the fallacy *cum hoc ergo propter hoc* ("with this, therefore because of this"), in which two events occur simultaneously or the chronological ordering is insignificant or unknown. Post hoc is a logical fallacy in which an event says to be the cause of a later event because it occurred earlier!

[35] Keep in mind that a "systematic review" answers a defined research question by collecting and summarising all empirical evidence that fits pre-specified eligibility criteria. A "meta-analysis" is the use of statistical methods to summarise the results of these studies.

Criteria	Meaning
Strength of Association	A strong association is more likely to have a causal component than is a modest association. Strength of the association is determined by the types of existing studies. The highest-level studies from the evidence pyramid would represent the strongest associations, (i.e., Systematic Reviews with meta-analysis and RCTs). Results from these studies must demonstrate an Odds Ratio (OR) or Relative Risk (RR) of at least 2 or above in order to be meaningful. Anything between 1 and 2 is weak, while > 2 is moderate and > 4 is considered strong.
Consistency	A relationship is repeatedly observed in all available studies.
Specificity	A factor influences specifically a particular outcome or population. The more specific an association between a factor and an effect, the greater the probability that it is causal.
Temporality	The cause must precede the outcome it is assumed to affect (e.g. smoking before the appearance of lung cancer) when the outcome is measured over time (longitudinal study).
Biological Gradient (dose-response)	The outcome increases monotonically with increasing dose of exposure or according to a function predicted by a substantive theory (i.e. the more cigarettes one smokes, the greater the chance of the cancer occurring).
Plausability	The observed association can be plausibly explained by substantive matter, i.e., biologically possible.
Coherence	A causal conclusion should not fundamentally contradict present substantive knowledge (studies must not contradict each other).
Experiment	Causation is more likely if evidence is based on randomized experiments or a systematic review with a meta-analysis of randomized experiments. However, if RCT's are not ethically possible then prospective cohort studies may provide the highest level of evidence.
Analogy	For analogous exposures and outcomes, an effect has already been shown, e.g., effects first demonstrated on animals or an effect previously occurring on humans such as the effects of thalidomide on a foetus during pregnancy.

Table 5.39: **Bradford Hill correlation-causation criterion**

Coming back to the mathematical aspect of the correlation:

- If $R_{X,Y}=-1$ we are dealing with a "pure negative correlation" (in the case of a linear relation all measurement points are located on a straight line with a negative slope).
- If $-1<R_{X,Y}<+1$ we are dealing with a negative or positive correlation named "imperfect correlation" (in the case of a linear relation all measurement points are located on a straight line negative or positive slope respectively).

- If $R_{X,Y}=0$ the correlation is zero... (in the case of a linear relation all the measurement points are located on a straight line of slope zero).

- If $R_{X,Y}=+1$ we are dealing with a "pure positive correlation" (in the case of a linear relation all measurement points are located on a straight positive slope).

The analysis of the correlation coefficient has the objective of determining the degree of association between variables: it is often expressed as the coefficient of determination, which is the square of the correlation coefficient. The coefficient of determination thus measures the contribution of a variable to the explanation of the second. However, asking «*how high should R (or R^2) be?* doesn't make sense. A low R doesn't negate a significant predictor or change the meaning of its coefficient. R is simply whatever value it is, and it doesn't need to be any particular value to allow for a valid interpretation».

Using the expressions of mean and standard deviation of equiprobable variables as demonstrated above (thus the idea of computing the correlation of two random variables is a good idea if they are jointly gaussian), we start:

$$R_{X,Y}=\frac{\mathrm{E}(XY)-\mathrm{E}(X)\mathrm{E}(Y)}{\sqrt{(\mathrm{E}(X^2)-\mathrm{E}(X)^2)\,(\mathrm{E}(Y^2)-\mathrm{E}(Y)^2)}} \tag{5.7.269}$$

To obtain the estimator of the coefficient of correlation:

$$R_{X,Y}=\frac{\dfrac{1}{n}\sum_{i=1}^{n}x_iy_i-\left(\dfrac{1}{n}\sum_{i=1}^{n}x_i\right)\left(\dfrac{1}{n}\sum_{i=1}^{n}y_i\right)}{\sqrt{\left(\dfrac{1}{n}\sum_{i=1}^{n}x_i^2-\dfrac{1}{n^2}\left(\sum_{i=1}^{n}x_i\right)^2\right)\left(\dfrac{1}{n}\sum_{i=1}^{n}y_i^2-\dfrac{1}{n^2}\left(\sum_{i=1}^{n}y_i\right)^2\right)}} \tag{5.7.270}$$

where we see that the covariance becomes the average of the products minus the product of averages.

Thus after simplification we get a famous expression:

$$\boxed{R_{X,Y}=\frac{\sum_{i=1}^{n}x_iy_i-\dfrac{1}{n}\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{\sqrt{\left(\sum_{i=1}^{n}x_i^2-\dfrac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right)\left(\sum_{i=1}^{n}y_i^2-\dfrac{1}{n}\left(\sum_{i=1}^{n}y_i\right)^2\right)}}} \tag{5.7.271}$$

The correlation coefficient can be calculated in the English version of Microsoft® Excel® 11.8346 and others with the integrated `CORREL( )` function.

We will see in the section of Numerical Methods a more general expression of the correlation coefficient (page 6332) and in the section of Statistical Mechanics an alternative indicator (page 3531).

Remarks

R1. In the literature, the experimental correlation coefficient is often named "sampling Pearson coefficient" (in the equiprobable case), or "Pearson product-moment", and when we carry it to the square, then we name it the "coefficient of determination".

R2. Often the square of the coefficient is somewhat improperly interpreted as the % of variation explained in the response variable Y by the explanatory variable X.

Finally, note that we have the following relation which is used a lot in practice (see the section Economics page 7698 for famous detailed examples!):

$$\text{V}(X+Y)=\text{V}(X)+\text{V}(Y)+2c_{X,Y}=\text{V}(X)+\text{V}(Y)+2R_{X,Y}\sqrt{\text{V}(X)\text{V}(Y)} \tag{5.7.272}$$

or the version with the standard deviation even more famous:

$$\boxed{\sigma_{X+Y}=\sqrt{\sigma_X^2+\sigma_Y^2+2R_{X,Y}\sigma_X\sigma_Y}} \tag{5.7.273}$$

It is a relation that we can often see in finance in the calculation of the VaR (Value at Risk) according to RiskMetrics methodology proposed by JP Morgan (see section Economy page 7859).

Let us see a small application example of the correlation but that has nothing to do with VaR (at least for the moment...).

✍**Example:**

An airline company has 120 seats available that she reserves for connecting passengers from two flights arrived earlier in the journey and that have to go to Frankfurt. The first flight arrived from Manila and the number of passengers on board follows a Normal distribution with mean 50 and variance 169. The second flight arrives in Taipei and the number of passengers on board follows a Normal distribution with mean 45 and variance 196.

The linear correlation coefficient between the number of passengers of both flights was measured as:

$$R_{X,Y}=0.5 \tag{5.7.274}$$

The law that follows the number of passengers for Frankfurt if we assume that the law of the couple also follows a Normal distribution (according to statement!) is:

$$X+Y=\mathcal{N}\left(\mu_Y+\mu_X,\sigma_{X+Y}\right) \tag{5.7.275}$$

with:

$$\begin{aligned}&\mu_Y+\mu_X=50+45=95\\&\sigma_{X+Y}=\sqrt{\sigma_X^2+\sigma_Y^2+2R_{X,Y}\sigma_X\sigma_Y}=\sqrt{169+196+2\cdot 0.5\sqrt{169\cdot 196}}\cong 23.38\end{aligned} \tag{5.7.276}$$

This is a bad start for customer satisfaction in the long term...

We have proved earlier that for a sum of independent and identically distributed random variables, the standard-deviation of their mean was given by the standard-error:

$$\sigma_{\bar{X}}=\frac{\sigma}{\sqrt{n}} \tag{5.7.277}$$

And what if their are not independent but still identically distributed of variance σ^2 (important case in Machine Learning but not only!)? First let us recall that:

$$\mathrm{V}\left(\sum_{i=1}^{n}X_i\right)=\sum_{i=1}^{n}\mathrm{V}(X_i)+2\sum_{i<j}^{n}\mathrm{cov}(X_i,X_j) \tag{5.7.278}$$

Therefore:

$$\mathrm{V}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\mathrm{V}(\bar{X})=\frac{n\sigma^2}{n^2}+\frac{2\displaystyle\sum_{i<j}^{n}\mathrm{cov}(X_i,X_j)}{n^2} \tag{5.7.279}$$

Using the definition of correlation seen just above:

$$R_{X_i,Y_i}=\frac{\mathrm{cov}_{X_i,Y_i}}{\sigma_{X_i}\sigma_{Y_i}} \tag{5.7.280}$$

We get (keeping in mind that all variances are assumed to be the same such that $\sigma_{X_i}\sigma_{Y_i}=\sigma$):

$$\mathrm{V}(\bar{X})=\frac{n\sigma^2}{n^2}+\frac{2\displaystyle\sum_{i<j}^{n}R_{X_i,Y_i}\sigma^2}{n^2}=\frac{\sigma^2}{n}+\frac{2\sigma^2\sum_{i<j}^{n}R_{X_i,Y_i}}{n^2} \tag{5.7.281}$$

Let us now assume that all pairwise correlations are equals (such that $R_{X_i,Y_i}=R$), then we get:

$$\begin{aligned}\mathrm{V}(\bar{X})&=\frac{\sigma^2}{n}+\frac{2\sigma^2\displaystyle\sum_{i<j}^{n}R}{n^2}=\frac{\sigma^2}{n}+\frac{\sigma^2 n(n-1)R}{n^2}=\frac{\sigma^2}{n}+\frac{\sigma^2(n-1)R}{n}\\&=R\sigma^2+\frac{1-R}{n}\sigma^2\end{aligned} \tag{5.7.282}$$

That latter relation is often denoted by data scientists as following:

$$\mathrm{V}(\bar{X})=\rho\sigma^2+\frac{1-\rho}{B}\sigma^2 \tag{5.7.283}$$

Finally, let us recall that the variance-covariance matrix is given by:

$$\Sigma=\begin{bmatrix}\mathrm{V}_{11} & c_{12} & \cdots & c_{1n}\\ c_{21} & \mathrm{V}_{22} & \cdots & c_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ c_{n1} & c_{n2} & \cdots & \mathrm{V}_{nn}\end{bmatrix}=\begin{bmatrix}\sigma_{11}^2 & c_{12} & \cdots & c_{1n}\\ c_{21} & \sigma_{22}^2 & \cdots & c_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ c_{n1} & c_{n2} & \cdots & \sigma_{nn}^2\end{bmatrix} \tag{5.7.284}$$

Dividing each term by the corresponding terms $\sqrt{\text{V}_i}\sqrt{\text{V}_j}$ gives us the "correlation matrix":

$$R=\begin{bmatrix} \dfrac{\text{V}_{11}}{\sqrt{\text{V}_1}\sqrt{\text{V}_1}} & \dfrac{c_{12}}{\sqrt{\text{V}_1}\sqrt{\text{V}_2}} & \cdots & \dfrac{c_{1n}}{\sqrt{\text{V}_1}\sqrt{\text{V}_n}} \\ \dfrac{c_{21}}{\sqrt{\text{V}_2}\sqrt{\text{V}_1}} & \dfrac{\text{V}_{22}}{\sqrt{\text{V}_2}\sqrt{\text{V}_2}} & \cdots & \dfrac{c_{2n}}{\sqrt{\text{V}_2}\sqrt{\text{V}_n}} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{c_{n1}}{\sqrt{\text{V}_n}\sqrt{\text{V}_1}} & \dfrac{c_{n2}}{\sqrt{\text{V}_n}\sqrt{\text{V}_2}} & \cdots & \dfrac{\text{V}_{nn}}{\sqrt{\text{V}_n}\sqrt{\text{V}_n}} \end{bmatrix}=\begin{bmatrix} 1 & R_{12} & \cdots & R_{1n} \\ R_{21} & 1 & \cdots & R_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ R_{n1} & R_{n2} & \cdots & 1 \end{bmatrix} \tag{5.7.285}$$

Notice that if the variances are all equal to the unit ($\text{V}_i=1$) then the covariance matrix is equal to the correlation matrix, i.e. $R=\Sigma$ (this is a useful property that we will use much later during our study of the variance inflation factor for multicollinearity detection)!

Let us see what we just said in another way! As we know the equation for the standardization of a variable is written as:

$$z=\frac{x_i-\bar{x}}{\sigma_x} \tag{5.7.286}$$

We also know that the original covariance matrix is defined as (using unbiased estimator):

$$\sigma_{xy}=\frac{1}{n-1}\sum_i^n(x_i-\bar{x})(y_i-\bar{y}) \tag{5.7.287}$$

And after standardizing both variables:

$$x'=\frac{x-\bar{x}}{\sigma_x} \qquad \text{and} \qquad y'=\frac{y-\bar{y}}{\sigma_y} \tag{5.7.288}$$

we get:

$$\begin{aligned}\sigma'_{xy}&=\frac{1}{n-1}\sum_i^n(x'_i-0)(y'_i-0)=\frac{1}{n-1}\sum_i^n\left(\frac{x-\bar{x}}{\sigma_x}\right)\left(\frac{y-\bar{y}}{\sigma_y}\right)\\&=\frac{1}{(n-1)\cdot\sigma_x\sigma_y}\sum_i^n(x_i-\bar{x})(y_i-\bar{y})=\frac{\sigma_{xy}}{\sigma_x\sigma_y}\end{aligned} \tag{5.7.289}$$

So finally:

$$\sigma'_{xy}=\frac{\sigma_{xy}}{\sigma_x\sigma_y} \tag{5.7.290}$$

So the variance-covariance matrix becomes indeed the correlation matrix with standardized variables!

5.7.3.1.4.1 Geometric interpretation

A measure of the relationship between two random variables is the covariance as we know given by:

$$c_{X,Y} = \mathrm{E}\left[(X-\mu_X)(Y-\mu_Y)\right] \tag{5.7.291}$$

Covariance normalized to unity is the correlation coefficient as we have just seen it:

$$\rho_{X,Y} = \frac{c_{X,Y}}{\sigma_X \sigma_Y} = \frac{\mathrm{E}\left[(X-\mu_X)(Y-\mu_Y)\right]}{\sqrt{\mathrm{E}[(X-\mu_X)^2]}\sqrt{\mathrm{E}[(Y-\mu_Y)^2]}} \tag{5.7.292}$$

Hence:

$$\rho_{X,Y} = \frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}[(X_i-\bar{x})(Y_i-\bar{y})]}{\sqrt{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(X_i-\bar{x})^2}\sqrt{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(Y_i-\bar{y})^2}} = \frac{\displaystyle\sum_{i=1}^{n}[(X_i-\bar{x})(Y_i-\bar{y})]}{\sqrt{\displaystyle\sum_{i=1}^{n}(X_i-\bar{x})^2}\sqrt{\displaystyle\sum_{i=1}^{n}(Y_i-\bar{y})^2}} \tag{5.7.293}$$

The resulting expression:

$$\rho_{X,Y} = \frac{\displaystyle\sum_{i=1}^{n}[(X_i-\bar{x})(Y_i-\bar{y})]}{\sqrt{\displaystyle\sum_{i=1}^{n}(X_i-\bar{x})^2}\sqrt{\displaystyle\sum_{i=1}^{n}(Y_i-\bar{y})^2}} \tag{5.7.294}$$

is the ratio of two elements. The numerator is the scalar product of two vectors, while the denominator is the product of their lengths:

$$\rho_{X,Y} = \frac{\displaystyle\sum_{i=1}^{n}[(X_i-\bar{x})(Y_i-\bar{y})]}{\sqrt{\displaystyle\sum_{i=1}^{n}(X_i-\bar{x})^2}\sqrt{\displaystyle\sum_{i=1}^{n}(Y_i-\bar{y})^2}} = \frac{\vec{x}\circ\vec{y}}{\|\vec{x}\|\cdot\|\vec{y}\|} = \cos(\vec{x},\vec{y})) \tag{5.7.295}$$

This expression shows the formal identity between the correlation coefficient, and the cosine of the angle between two random vectors.

5.7.3.1.4.2 Point biserial correlation coefficient

A "point-biserial correlation", denoted r_{pb}, is used to measure the strength and direction of the association that exists between one continuous variable and one dichotomous variable. It is a special case of the Pearson's product-moment correlation, which is applied when you have two continuous variables, whereas in this case one of the variables is measured on a dichotomous scale.

For example, we could use a point-biserial correlation to determine whether there is an association between salaries, measured in US dollars, and gender (i.e., your continuous variable would be "salary" and your dichotomous variable would be "gender", which has two categories: "males" and "females").

To calculate r_{pb} we assume that the dichotomous variable Y has the two values 0 and 1. If we divide the data set into two groups, group 1 which received the value "1" on Y and group 2 which received the value "0" on Y, then the point-biserial correlation coefficient is calculated as follows:

$$\boxed{r_{\text{pb}} = \frac{\mu_1 - \mu_0}{\sigma_n}\sqrt{\frac{n_1 n_0}{n^2}}} \tag{5.7.296}$$

where σ_n is the standard deviation used when data are available for every member of the population (obviously there are version that use the estimator of the standard deviation instead of the biased standard deviation!):

$$\sigma_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{5.7.297}$$

μ_1 being the mean value on the continuous variable X for all data points in group 1, and μ_0 the mean value on the continuous variable X for all data points in group 2. Further, n_1 is the number of data points in group 1, n_0 is the number of data points in group 2 and n is the total sample size.

Proof 5.45.2. Let the n data consist of $n_0 > 0 (x, 0)$ pairs and $n_1 > 0 (x, 1)$ pairs. Their Pearson correlation coefficient will be the same as the reversed data consisting of corresponding $(0, x)$ and $(1, x)$ pairs. Because there are exactly two distinct values of the first coordinates, the regression line of the reversed data must pass through the mean points $(0, \mu_0)$ and $(1, \mu_1)$, whence it has slope:

$$a = \frac{\mu_1 - \mu_0}{1 - 0} = \mu_1 - \mu_0 \tag{5.7.298}$$

We have proved (see section Numerical Methods page 6332) that the correlation coefficient for a univariate regression is also given by:

$$R_{X,Y}^2 = \frac{a^2 \text{V}(X)}{\text{V}(Y)} \Leftrightarrow R_{X,Y} = \frac{a}{\sigma_Y}\sigma_X \tag{5.7.299}$$

Therefore the slope must be multiplied by the standard deviation of the first coordinates and divided by the standard deviation of the second coordinates (the original x values), denoted traditionally (sadly) by $\hat{\sigma}_n$. The standard deviation of the first coordinates is readily computed from the fact that they consist of n_0 zeros (0) and n_1 ones (1). Considering this set of values as a Bernoulli random variable, we have proved earlier above during our study of the Bernoulli distribution that:

$$\sigma_X = \sqrt{pq} = \sqrt{p(1-p)} \tag{5.7.300}$$

Therefore:

$$\sigma_X = \sqrt{pq} = \sqrt{\frac{n_0}{n}\frac{n_1}{n}} = \sqrt{\frac{n_0 n_1}{n^2}} \tag{5.7.301}$$

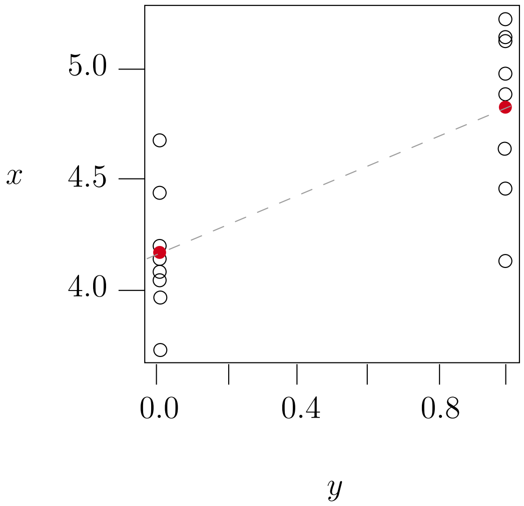
Consequently the point biserial Pearson correlation coefficient is:

$$r_{\text{pb}}=\frac{\mu_1-\mu_0}{\sigma_n}\sqrt{p(1-p)}=\frac{\mu_1-\mu_0}{\sigma_n}\sqrt{\frac{n_0n_1}{n^2}} \tag{5.7.302}$$

So the point biserial correlation coefficient is just a special case of the Pearson correlation coefficient. However remember that in most statistical computer softwares the Pearson correlation coefficient is calculated using the unbiased variance and the Point biserial correlation coefficient used the biased one, this is why we don't get the same numerical results when using both corresponding functions in softwares!

■ Q.E.D.

The idea is depicted here:



The heights of the red dots depict the mean values μ_0 and μ_1 of each vertical strip of points. The dashed gray line is the regression line.

Remark

R software native function computes the point biserial correlation coefficient with the corr and corr.test function because the point-biserial correlation is simply a special case of Pearson's product-moment correlation, which is applied when you have two continuous variables, whereas in this case one of the variables is measured on a dichotomous scale. But its also available as biserial.cor in various packages.

✍**Example:**

Let us imagine that the level of operational risk in an institution's business lines is graded as high or low based on the opinion of an operational risk analyst. Suppose we are interested in assessing the degree of correlation between operational risk and gross income of the business lines. Since in this case operational risk is a binary variable and gross income a continuous variable, we can use the biserial coefficient to estimate correlation!

5.7.3.1.4.3 Partial correlation and semipartial correlation

Partial correlation measures the degree of association between two random variables, with the effect of a set of controlling random variables removed. If we are interested in finding whether or to what extent there is a numerical relationship between two variables of interest, using their correlation coefficient will give misleading results if there is another, "confounding", variable that is numerically related to both variables of interest. This misleading information can be avoided by controlling (i.e. cancelling) for the confounding variable, which is done by computing the partial correlation coefficient.

When in the case of three variables we want to control (cancel) one of the variable and that this one is the unique that is qualitative, the strategy is then simple as it consist to calculate the correlation between the two remaining variable for each level of the qualitative variable. But things start to be more complicated when the variable to control (cancel) is continuous. This is why we need to build a more sophisticated approach!

The "partial correlation coefficient", denoted here $R_{AB.C}$, makes it possible to know the value of the correlation between two variables A and B, if the variable C had remained constant for the series of observations considered.

In other words, the partial correlation coefficient $R_{AB.C}$ is the total correlation coefficient between variables A and B when their best linear explanation has been removed in terms of C. It is given by the relation (in the obvious special case of three variables!):

$$\boxed{R_{AB.C}=\frac{R_{AB}-R_{AC}R_{BC}}{\sqrt{1-R_{AC}^2}\sqrt{1-R_{BC}^2}}} \tag{5.7.303}$$

The quickest proof of this relation is to rely on the geometric interpretation of the correlation (cosine).

But notice before that like the correlation coefficient, the partial correlation coefficient takes on a value in the range from -1 to $+1$. The value -1 conveys a perfect negative correlation controlling for some variables (that is, an exact linear relationship in which higher values of one variable are associated with lower values of the other); the value $+1$ conveys a perfect positive linear relationship, and the value 0 conveys that there is no linear relationship.

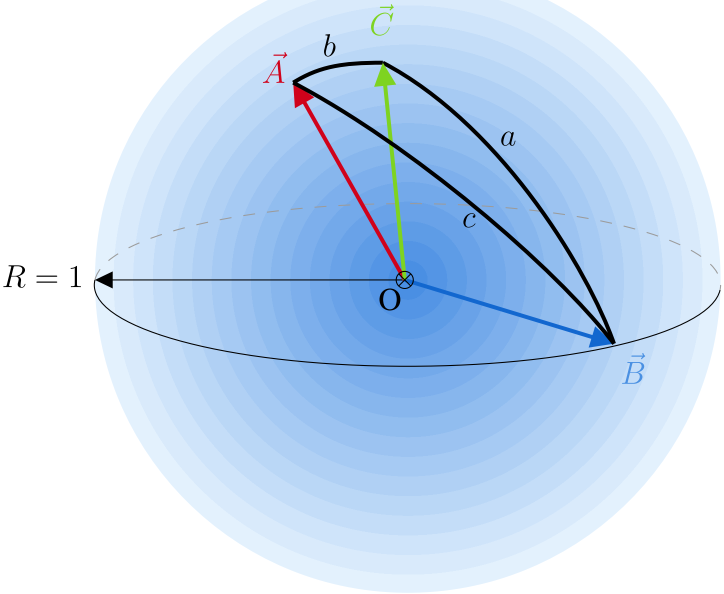
Also notice that if R_{AC} and R_{BC} are both equal to zero (i.e. there is no confounding variable) then we have the equality:

$$R_{AB.C}=\frac{R_{AB}-0\cdot 0}{\sqrt{1-0}\sqrt{1-0}}=R_{AB} \tag{5.7.304}$$

So it is easy to understand that in general when there is somewhere a correlation between some of the variables, then $R_{AB.C}>R_{AB}$. For other configuration values, the reader must be very precautions on the interpretation as the result depends obviously on the value of R_{AB}!

Ok now let's go for the proof in the special case of three variables!

The series of observations A, B and C, once centered reduced (thus of radius 1), are centered vectors $\vec{A}$, $\vec{B}$, $\vec{C}$ of unit length on the same sphere:



Their extremities (vertices) determine a spherical triangle ABC, whose sides a, b and c are the arcs of large circles $\overset{\frown}{BC}$, $\overset{\frown}{AC}$ and $\overset{\frown}{AB}$. The correlation coefficients between these vectors are (let us recall that for a unitary circle/sphere we have $\cos(\alpha) = \cos(a)$ as seen in the section of Trigonometry):

$$R_{BC} = \cos(a) \qquad R_{AC} = \cos(b) \qquad R_{AB} = \cos(c) \tag{5.7.305}$$

We have then by the cosine formula of the spherical triangles (see section Trigonometry page 2544) the following value for the opening angle of C:

$$\cos(\hat{C}) = \frac{\cos(c) - \cos(a)\cos(b)}{\sin(a)\sin(b)} = \frac{\cos(c) - \cos(a)\cos(b)}{\sqrt{1-\cos^2(a)}\sqrt{1-\cos^2(b)}} \tag{5.7.306}$$

Just as c is the angle between the points A and B (and on a unite sphere also the arclength $\overset{\frown}{AB}$), seen from the center of the sphere, $\hat{C}$ is the spherical angle between points A and B, seen from point C to the surface of the sphere, and:

$$R_{AB.C} = \cos(\hat{C}) \tag{5.7.307}$$

is the "partial correlation" between A and B when C is fixed.

Remarks

R1. If we partial one variable out of a correlation, that partial correlation is named a "first order partial correlation". If we partial out 2 variables from that correlation (e.g., $R_{12.34}$), we have a "second order partial correlation", and so forth. It is customary to refer to "unpartialed (raw, as it were) correlations" as "zero order correlations".

R2. Keep in mind that as for the correlation, the partial correlation when high, doesn't necessarily imply causation!

With partial correlation, we find the correlation between A and B holding C constant for both A and B. Sometimes, however, we want to hold C constant for just A or just B. In that case, we compute a "semipartial correlation coefficient".

Let's compare the correlational formulas for the partial and semipartial:

- Partial:

$$R_{AB.C}=\frac{R_{AB}-R_{AC}R_{BC}}{\sqrt{1-R_{AC}^2}\sqrt{1-R_{BC}^2}} \tag{5.7.308}$$

- Semipartial:

$$R_{A(B.C)}=\frac{R_{AB}-R_{AC}R_{BC}}{\sqrt{1-R_{BC}^2}} \quad \text{and} \quad R_{B(A.C)}=\frac{R_{AB}-R_{AC}R_{BC}}{\sqrt{1-R_{AC}^2}} \tag{5.7.309}$$

Note that the partial and semipartial correlation formulas are the same in the numerator and almost the same in the denominator. The partial contains something extra, that is, something missing from the semipartial correlation in the denominator. This means that the partial correlation is going to be larger in absolute value than the semipartial. This will be true except when the controlling or partialling variable is uncorrelated with the variable to be controlled or residualized; this is a trivial case.

5.7.3.1.4.4 Intraclass coefficient of correlation

The "intraclass correlation coefficient", denoted ICC, is a metric that can be used when quantitative measurements are made on measures that are organized into groups. It describes how the measurements of the same group are similar (thus it is still one of the many indicators of the family of repeatability and reproducibility studies like Cohen's kappa of approval that we will introduce later). This statistic is considered as a type of correlation and, unlike most other correlation measures, it therefore works on data structured in groups.

An important application is the evaluation of the consistency or the reproducibility of quantitative measurements made by different observers measuring the same quantity (R&R Gage studies - GRR). For example, if several doctors are asked to mark the results of a scan for signs of cancer progression, we may wonder how the scores are consistent from one physician to another (i.e. ANOVA with random factors).

An important aspect of this statistic (and also a weakness) is that it quantifies both inter-observer variability and intra-observer variability. Inter-observer variability refers to systematic differences between observers - for example, a physician may always mark patients at a higher risk level than other physicians. Intra-observer variability refers to measurement deviations of a single observer.

> **Remark**
>
> Since the intraclass correlation coefficient gives a composition of inter- and intra-observer variability, its results are sometimes considered difficult to interpret when observers are not exchangeable (random). Alternative measures such as the Cohen approval kappa statistic (see below), the Fleiss kappa and the correlation coefficient of concordance were proposed as more appropriate measures of agreement between non-exchangeable observers.

There are several kinds of intraclass correlations. They relate to different ways of conceptualizing the variables that are often referred to generically as "subjects" and "judges." Either of these variables may be a random variable (their levels were chosen at random from a population of possible levels) or a fixed variable, the different levels of which were specifically selected for this design, and the same levels would be selected again in a replication. In addition, we could use individual ratings, or we could take the mean of a set of ratings by different raters. Doing so will alter the intraclass correlation coefficient by providing a more reliable measure.

Here is a figurative example of a ICC (for numerical application see the Minitab or `R` companion books):

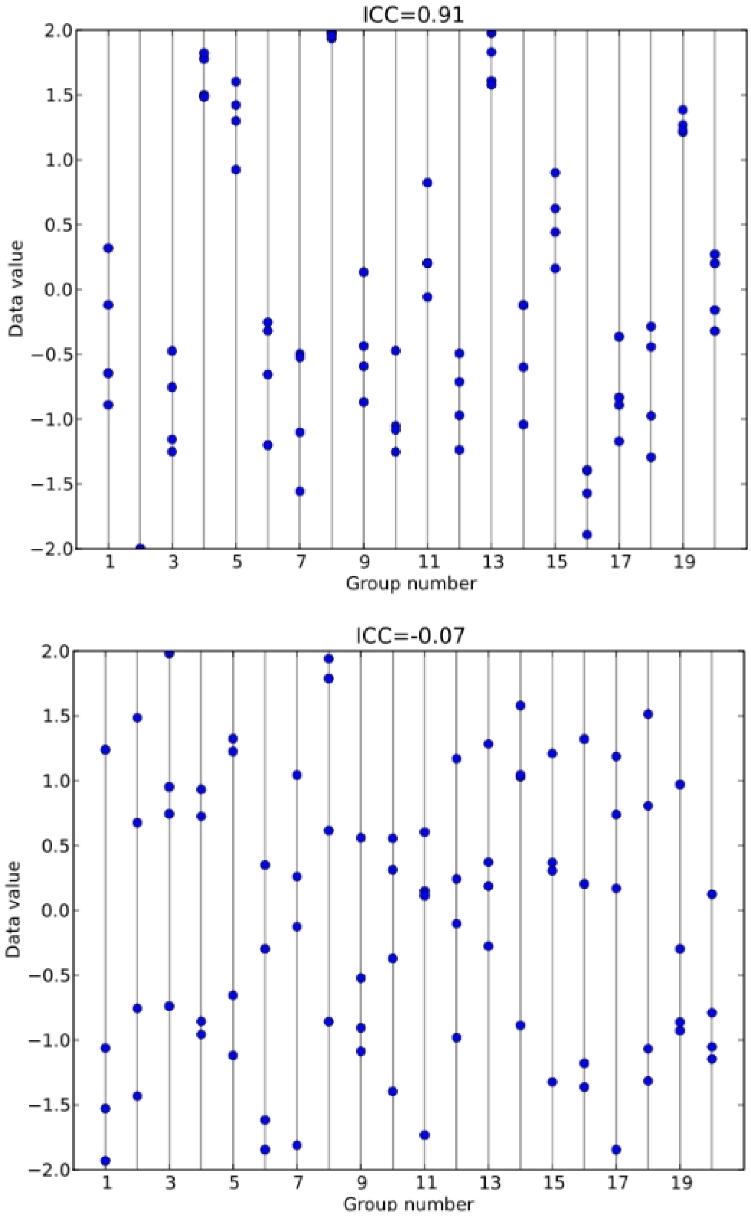


Figure 5.115: **Intra-class correlation coefficient illustration (source: Wikipedia)**

The intra-class correlation coefficient that will interest us here, denoted ICC1, that is most often used in practice is to consider a fixed factor (non-random) and the other random (there are therefore other variants that must be applied with caution!).

Indeed, imagine a study concerning the productivity of employees of a large company. The management committee wants to have an idea of the daily production and how it depends on the production machines used by the employees. If we take for the study a random sample of n employees among N, with all the machines at a quantity of k then we have a model with one fixed factor (machines) and the other one that is random (employees).

The empirical idea is then to write the model in the following very specific form (for the other more general models it will be necessary to wait until we write the developments!):

$$y_{ij}=\mu+\alpha_j+\varepsilon_{ij} \tag{5.7.310}$$

where i is the measure of the fixed factor (often as rows in ICC analysis tables) and j is the random factor represented by the judges (often as columns in ICC analysis tables) and thus ε_{ij} describes the variability of measurement of the elements in a given level of random factor and the error of the judges is inseparable from the element measured. The previous relation is also sometimes denoted:

$$y_{ij} = \mu_j + \varepsilon_{ij} \tag{5.7.311}$$

where obviously $\mu_j = \mu + \tau_j$.

As usual (...) we assume that:

$$\begin{aligned} \alpha_j &= \mathcal{N}(0, \sigma_\alpha^2) \\ \varepsilon_{ij} &= \mathcal{N}(0, \sigma_\varepsilon^2) \end{aligned} \tag{5.7.312}$$

So the variances of the α_j are assumed to be identical (which is nevertheless quite a strong hypothesis in practice!).

We have, therefore, from the linearity of the mean:

$$\text{E}(y_{ij}) = \text{E}(\mu + \alpha_j + \varepsilon_{ij}) = \mu \tag{5.7.313}$$

The decomposition of the variance is immediate if the errors are independent of the random factors:

$$\text{V}(y_{ij}) = \text{V}(\mu + \alpha_j + \varepsilon_{ij}) = \text{V}(\alpha_j) + \text{V}(\varepsilon_{ij}) = \sigma_\alpha^2 + \sigma_\varepsilon^2 \tag{5.7.314}$$

We have therefore:

$$y_{ij} = \mathcal{N}(\mu, \sigma_\alpha^2 + \sigma_\varepsilon^2) \tag{5.7.315}$$

For two elements (observations) taken under the same level of factor i, we have:

$$\begin{aligned} \text{cov}(y_{ij}, y_{ik}) &= \text{cov}(\mu + \alpha_j + \varepsilon_i j, \mu + \alpha_k + \varepsilon_{ik}) \\ &= \text{cov}(\alpha_j + \varepsilon_{ij}, \alpha_k + \varepsilon_{ik}) = \text{cov}(\alpha_j + \varepsilon_{ij}, \alpha_k) + \text{cov}(\alpha_j + \varepsilon_{ij}, \varepsilon_{ik}) \\ &= \text{cov}(\alpha_j, \alpha_k) + \text{cov}(\varepsilon_{ij}, \alpha_k) + \text{cov}(\alpha_j, \varepsilon_{ik}) + \text{cov}(\varepsilon_{ij}, \varepsilon_{ik}) \end{aligned} \tag{5.7.316}$$

But as the variances of all α_j are assumed to be identical and the errors independent of the fixed effects, the latter reduce to:

$$\text{cov}(y_{ij}, y_{ik}) = \text{cov}(\alpha_j, \alpha_k) = \text{cov}(\alpha_j, \alpha_j) = \text{V}(\alpha_j) = \sigma_\alpha^2 \tag{5.7.317}$$

And their correlation is then:

$$\boxed{\rho(y_{ij}, y_{ik}) = \frac{\text{cov}(y_{ij}, y_{ik})}{\text{V}(y_{ij})\text{V}(y_{ik})} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2}} \tag{5.7.318}$$

This is often denoted in textbooks as:

$$\text{ICC} = \frac{S_b^2}{S_b^2 + S_w^2} \tag{5.7.319}$$

Where S_w^2 is the pooled variance (see page 1011) within subjects, and S_b^2 is the variance of the measurements between subjects. $S_b^2+S_w^2$ is the total variance.

It is therefore the ratio of the variance of the random factors to the total variance (the denominator including the variance of the error) which is often denoted ICC1 or ICC(1,1) to say that it is the Intraclass correlation coefficient for 1 random factor and 1 fixed factor.

So if $\sigma_\varepsilon\rightarrow 0$, the intraclass correlation is unitary and this is then and indicator of good reproducibility!

Let us now recall from our study of one fixed factor ANOVA (page 1004) that we had the following decomposition in sum of the squares:

$$\begin{array}{ccccc} Q_T & = & Q_A & + & Q_R \\ \sum_{ij}(y_{ij}-\bar{\bar{y}})^2 & = & \sum_i n_i(\bar{y}_i-\bar{\bar{y}})^2 & + & \sum_{ij}(y_{ij}-\bar{y}_i)^2 \end{array} \tag{5.7.320}$$

If $n_i=n$ (balanced experiment), then we have obviously:

$$\begin{array}{ccccc} Q_T & = & Q_A & + & Q_R \\ \sum_{ij}(y_{ij}-\bar{\bar{y}})^2 & = & n\sum_i(\bar{y}_i-\bar{\bar{y}})^2 & + & \sum_{ij}(y_{ij}-\bar{y}_i)^2 \end{array} \tag{5.7.321}$$

This is often written in the context of the use of the intra-class correlation coefficient:

$$\frac{\sum_{ij}(y_{ij}-\bar{\bar{y}})^2}{N-1}=n\frac{\sum_i(\bar{y}_i-\bar{\bar{y}})^2}{k-1}+\frac{\sum_{ij}(y_{ij}-\bar{y}_i)^2}{k(n-1)}=n\sigma_\alpha^2+\sigma_\varepsilon^2 \tag{5.7.322}$$

where for recall $N=kn$. Here we assume therefore that the judges are indexed by i, the observations by j and that we have n judges and k observations!

And we also regularly have the following notations in some textbooks:

$$\text{MS}_{\text{w/intra}}=\sigma_\varepsilon^2 \qquad \text{and} \qquad \text{MS}_{\text{b/inter}}=n\sigma_\alpha^2+\sigma_\varepsilon^2 \tag{5.7.323}$$

or:

$$\text{SSW}=\sigma_\varepsilon^2 \qquad \text{and} \qquad \text{SSB}=n\sigma_\alpha^2+\sigma_\varepsilon^2 \tag{5.7.324}$$

where SSW stands for "Sum of Squares Within" and SSB for "Sum of Squares Between" or (...):

$$\text{MSE}=\sigma_\varepsilon^2 \qquad \text{and} \qquad \text{MSk}=n\sigma_\alpha^2+\sigma_\varepsilon^2 \tag{5.7.325}$$

where MSE stands for "Mean Square for Error" and MSk for "Mean Square for treatments". Therefore:

$$n\sigma_\alpha^2=\text{MS}_{\text{b/inter}}-\sigma_\varepsilon^2=\text{MS}_{\text{b/inter}}-\text{MS}_{\text{w/intra}} \tag{5.7.326}$$

Hence:

$$\sigma_\alpha^2=\frac{\text{MS}_{\text{b/inter}}-\text{MS}_{\text{w/intra}}}{n} \tag{5.7.327}$$

Remark

As the reader have maybe notice it, the numerator may be negative and therefore so may be σ_α^2... If we get negative variance components, the following are possible ways to deal with the negative estimates:

- Accept the estimate as evidence of a true value of zero and use zero as the estimate, recognizing that the estimator will no longer be unbiased.
- Retain the negative estimate, recognizing that subsequent calculations using the results might not make much sense.
- Interpret that the negative component estimate indicates an incorrect statistical model.
- Collect more data and analyse them separately or in conjunction with the existing data and hope that increased information will yield positive estimates.

For example, when a variance component estimate is less than zero, the software Minitab displays the negative estimate, but sets the estimate to zero in order to calculate the percent of total variability.

This is the reasons why we often find the intraclass correlation coefficient in the following common form in the textbooks because it allows a sometimes a more intuitive approach to the concept:

$$\begin{aligned}\text{ICC} &= \frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\varepsilon^2} = \frac{\dfrac{\text{MS}_{\text{b/inter}}-\text{MS}_{\text{w/intra}}}{n}}{\dfrac{\text{MS}_{\text{b/inter}}-\text{MS}_{\text{w/intra}}}{n}-\text{MS}_{\text{w/intra}}} = \frac{\text{MS}_{\text{b/inter}}-\text{MS}_{\text{w/intra}}}{\text{MS}_{\text{b/inter}}-\text{MS}_{\text{w/intra}}+n\text{MS}_{\text{w/intra}}}\\ &= \frac{\text{MS}_{\text{b/inter}}-\text{MS}_{\text{w/intra}}}{\text{MS}_{\text{b/inter}}-(n-1)\text{MS}_{\text{w/intra}}}\end{aligned} \tag{5.7.328}$$

That is also often written is some textbooks as:

$$\text{ICC} = \frac{\text{MS}_{\text{bet S}}-\text{MS}_{\text{residual}}}{\text{MS}_{\text{bet S}}+(n-1)\text{MS}_{\text{residual}}} \tag{5.7.329}$$

Dominic V. Cicchetti (11994 according to holocene calendar) gives the following often quoted guidelines for interpretation for kappa or ICC inter-rater agreement measures:

ICC value	Criteria
$0.4 <$	Poor
$[0.4, 0.59]$	Fair
$[0.60, 0.74]$	Good
> 0.74	Excellent

Table 5.40: Interpretation for kappa or ICC inter-rater agreement

5.7.3.2 Continuous Variables and Moments

Definitions (#5.7.131):

D1. We say that X is a continuous variable if its "cumulative distribution function" (C.D.F.) is continuous (already defined above). The distribution function of X is defined by for $x \in \mathbb{R}$ or a truncated subset of $\mathbb{R}$:

$$F(x) = P(X \leq x) \tag{5.7.330}$$

that is the cumulative probability that the random variable X is smaller than or equal to the set value x. More properly, as for discrete variables, this is denoted by:

$$F_X(t) = P(X \leq t) = P(X \in]\min(D), t]) \qquad \forall t \in \mathbb{R} \tag{5.7.331}$$

with:

$$0 \leq P(X) \leq 1 \quad \text{and} \quad 0 \leq F(X) \leq 1 \tag{5.7.332}$$

D2. We denote by:

$$G(x) = 1 - F(x) = P(X > x) \tag{5.7.333}$$

the "survival function" or "tail function".

D3. If furthermore the distribution function F of X is continuously differentiable of derivative f (or sometimes denoted by ρ) named "probability density function[36]" or just simply "distribution function" then we say that X is absolutely continuous and in this case we have:

$$P(x_1 \leq X \leq x_2) = \int\limits_{x_1}^{x_2} f(x)\mathrm{d}x = F(x_2) - F(x_1) \tag{5.7.334}$$

with the normalization condition:

$$P(X < -\infty) = \int\limits_{-\infty}^{+\infty} f(x)\mathrm{d}x = \int\limits_{-\infty}^{+\infty} \mathrm{d}F(x) = 1 \tag{5.7.335}$$

Any probability distribution function must satisfy the integral of normalization in its domain of definition!

Let us consider a typical transformation of random variables that is quite common to generalize some special cases to more generic cases (typically from the standard Normal law to the general Normal law or from the standard Cauchy law to the generalized Cauchy law)! Let X be a continuous random variable be then set $Y = a + bX$ where a is named the "location parameter" and with $b > 0$ named the "scale parameter" and let us denote F the distribution function of X and G the distribution function of Y. Then:

---

[36] Keep in mind that the expression "probability mass function" is reserved to discrete variables.

$$G(y) = P(Y \leq y) = P(a + bX \leq y) = P\left(X \leq \frac{y-a}{b}\right) = F\left(\frac{y-a}{b}\right) \tag{5.7.336}$$

Now we differentiate and we get the important and trivial result:

$$g(y) = f\left(\frac{y-a}{b}\right)\frac{1}{b} \tag{5.7.337}$$

where g is the density function for Y and f is the density function for X.

We can also consider the case where $b < 0$. Then:

$$G(y) = P(Y \leq y) = P(a - bX \leq y) = P\left(X \geq \frac{y-a}{b}\right) = 1 - F\left(\frac{y-a}{b}\right) \tag{5.7.338}$$

differentiating:

$$g(y) = -f\left(\frac{y-a}{b}\right)\frac{1}{b} = f\left(\frac{y-a}{b}\right)\frac{1}{|b|} \tag{5.7.339}$$

They reader must try to keep in mind these relations if possible.

Remark

It is interesting to note that the definition implies that the probability that a completely continuous random variable takes a given value tends to zero! So it is not because an event has almost a zero probability that it can not happen!!!

The average being defined by a sum weighted by probabilities for a discrete variable, it becomes an integral for a continuous variable (the upper and lower integration limits are not always from $-\infty$ to $+\infty$ obviously! It depends of the definition domain of the random variable of interest):

$$\boxed{\mathrm{E}(X) = \int_{-\infty}^{+\infty} x f(x)\mathrm{d}x} \tag{5.7.340}$$

and therefore the variance for a continuous variable: is written as:

$$\boxed{\mathrm{V}(X) = \int_{-\infty}^{+\infty} [x - \mathrm{E}(X)]^2 f(x)\mathrm{d}x} \tag{5.7.341}$$

Then we have also the median that is logically redefined in the case of a continuous random variable by:

$$\boxed{P(X \leq M_e) = P(X \geq M_e) = \int_{-\infty}^{M_e} f(x)\mathrm{d}x = \int_{M_e}^{+\infty} f(x)\mathrm{d}x = \frac{1}{2}} \tag{5.7.342}$$

and it rarely coincides with the average!

And the modal value is given by the value of x where:

$$\frac{\mathrm{d}f(x)}{\mathrm{d}x}=0 \tag{5.7.343}$$

Statisticians often use the following notations for the expected mean of a continuous variable:

$$\mathrm{E}(X), M(X), \mu_X, \mu \tag{5.7.344}$$

and for the variance:

$$\mathrm{V}(X), S(X), \sigma_X^2, \sigma^2 \tag{5.7.345}$$

That is the same as for the moment of discrete variable.

Thereafter, we will calculate these different moments indicators with detailed proofs only for the most used cases.

But before... let us discuss about some other well-knows moments and highly discussed...!

5.7.3.2.1 Skewness and Kurtosis

Caution! I have hesitated during many years (more than 10 years!) to write something about the skewness and kurtosis in this textbook as they are most of time misleading concepts and that they are difficult (or even impossible...) to introduce with mathematical rigour. I changed my mind when I saw that my students were expecting something more complete and robust than Wikipedia. The reader must keep in mind that these both indicators must be use with caution and are good tools if we know that we are working with Normally distributed data.

As already mentioned in the previous warning, skewness and kurtosis are somewhat vague notions. The reader should therefore not expect any proof of their origin here... Furthermore, the concept of "symmetry", is mathematically precise, but skewness by contrast is surprisingly slippery. Kurtosis is perhaps even more so.

The underlying ideas corresponding definitions are the following (the definitions are given for discrete variable but can easily be extend to continuous variable by converting the sums into integrals!):

Definitions (#5.7.132):

D1. If the "skewness", denoted in general γ, and associated most of time with the moment of order 3 also named the "Fisher-Pearson coefficient of skewness":

$$\boxed{\gamma_1=\frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\sqrt{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^3}} \qquad (5.7.346)$$

is zero, the distribution is symmetric (this does not mean that the symmetry takes place on a peak of the distribution because in the bimodal case the axis of symmetry may be between the two symmetric modal values). If the skewness is positive (right-modal), the distribution (modal / median value) is lean to the right (or there are extreme values to the right). If the skewness is negative (left-modal), the distribution (the modal / median value) is leaned to the left (or there are extreme values to the left).

Remark

A common question is why the skewness is defined as third central moment, and not fifth or any other number? What's the logic behind it? In fact nothing special! We could choose any odd power but the thing is that the moment of order 1 is the expected mean (already taken...) and the next smallest order easy to manage for further calculations is the power of 3 (we let the reader imagine the integration of distribution function with variables of order like 5 or bigger...).

It's possible for a distribution to be symmetric but not have zero third moment. A simple counterexample is any T-distribution with 3 or fewer degrees of freedom (see proof during our study of the corresponding distribution). In samples, however, symmetry implies zero third moment - but samples are almost never perfectly symmetric, so it's not much use there either. It's also possible to build asymmetric distributions that have non-zero third moment (see our `R` companion book!)!!! So symmetry doesn't necessarily imply zero third moment and zero third moment doesn't necessarily imply symmetry.

D2. If the "kurtosis", denoted in general κ (see the maths just after the definitions), and associated most of time with the moment of order 4 (hence always positive or equal to zero):

$$\boxed{\kappa_1=\frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left(\sqrt{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^4}} \tag{5.7.347}$$

and if we define the "excess kurtosis" as κ_1-3. There are three distinct regimes as described below:

- The excess kurtosis is zero, i.e. "mesokurtic", then the flattening is similar to that of a Normal distribution[37].
- If the value is greater than zero, i.e. "leptokurtic", then the distribution of interest is higher than that of a Normal distribution with equal mean.
- The excess kurtosis is smaller than zero, i.e. "platikurtic", then the distribution of interest is lower than that of a Normal distribution with equal mean.

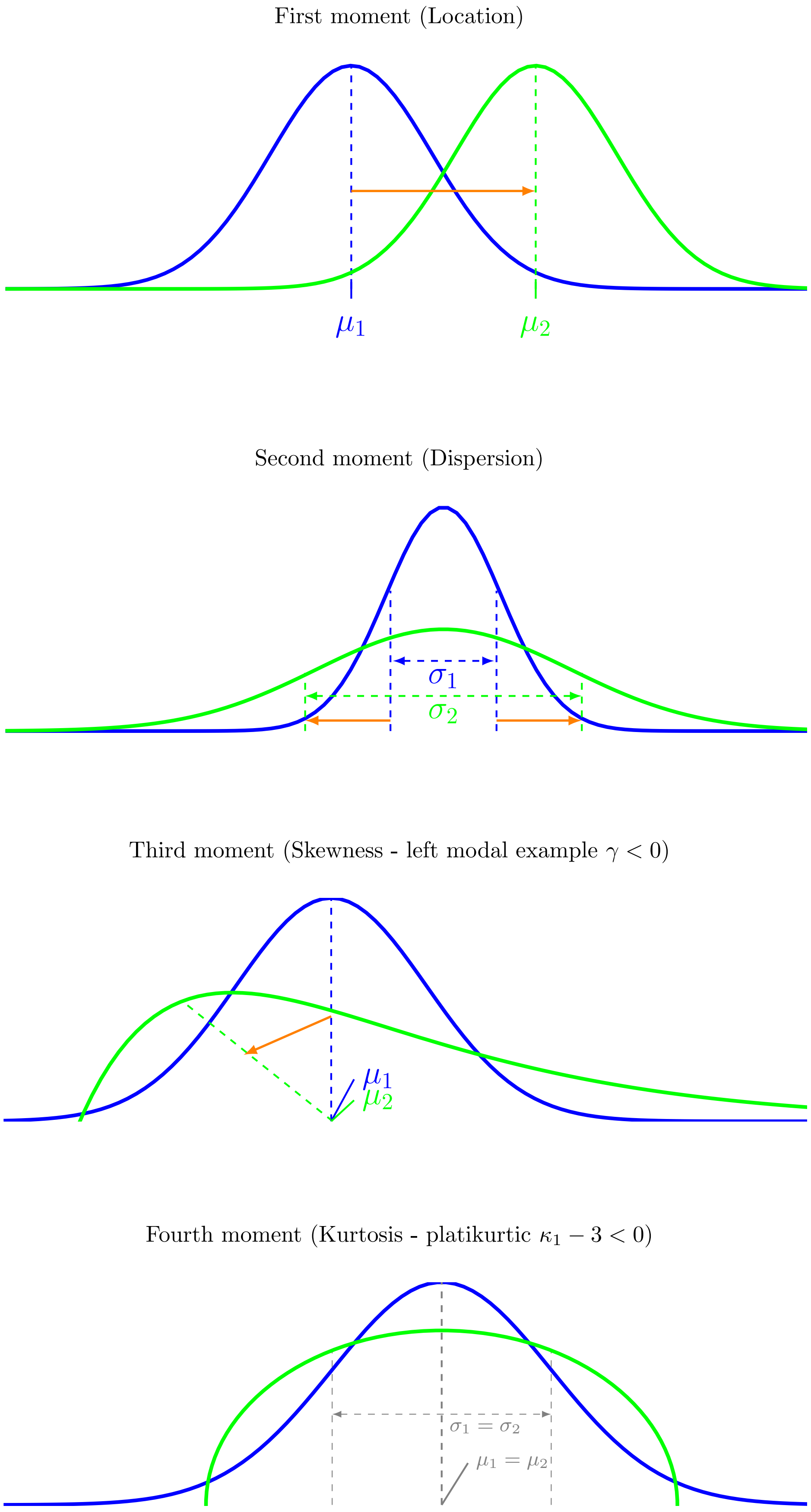
Rigorously kurtosis measure how far the underlying distribution is from being symmetric and bimodal (algebraically, a perfectly symmetric and bimodal distribution will have a kurtosis of 1, which is the smallest possible value the kurtosis can have). Most elementary texts describe kurtosis as a measure of the "peakedness" of a distribution. This term is misleading, and that a far better term for describing kurtosis is "bimodality," where the lower the kurtosis, the greater the bimodality.

Notice that the kurtosis may seems similar to the idea of measuring the dispersion (i.e. variance). However, one way to look at kurtosis is that it measures the width of a random distribution compared to that of a Normal distribution with the same mean and variance as the distribution in question. If the distribution is symmetric about the mean, a distribution with a positive kurtosis will have a sharper peak and much longer tails than does the Normal distribution. The opposite is the case for distributions with a negative kurtosis.

Here is an illustration:

---

[37] Rigorously a Normal distribution has a zero kurtosis. But the opposite is not true! Distributions that don't look Normal at all can have zero kurtosis!


First moment (Location)

Second moment (Dispersion)

Third moment (Skewness - left modal example $\gamma < 0$)

Fourth moment (Kurtosis - platikurtic $\kappa_1 - 3 < 0$)

Figure 5.116: **Kurtosis and Skewness range map of some distributions (author: Simon H. Hess)**

Remark

If one does image searches for the "kurtosis" on Google, quite a few images show the kurtosis as being only a Gaussian distributions with a greater or smaller second order moment (i.e. the Gaussian is just larger or smaller). Such images are misleading representations of kurtosis.

There have been a fairly large number of attempts to give measures corresponding to these notions, but these often-useful measures can be surprisingly counter-intuitive at times. For example, the moment-based skewness can be zero when the distribution is asymmetric (contradicting an assertion one can surprisingly often find when reading elementary texts which discuss skewness).

In facts moments of a continuous distribution, and functions of them like the kurtosis, tell you extremely little about the graph of its density function!!!! It should be abundantly clear that kurtosis and skewness is not an easily interpretable or intuitive measure of symmetry, unimodality, bimodality, convexity, or any other familiar geometric characterization of a curve. Functions of moments, therefore (and kurtosis as a special case) do not describe geometric properties of the graph of the probability density function. This intuitively makes sense: because a probability density function represents probability by means of area, we can almost freely shift probability density around from one location to another, radically changing the appearance of the probability density function, while fixing any finite number of pre-specified moments.

Here are the skewness used by `R` (implements γ_1, γ_2, γ_3), MATLAB® (implements γ_1, γ_2), Mathematica (implements γ_1), and MINITAB/IBM SPSS and Microsoft® Excel® (that implement γ_3):

$$\begin{aligned}
\gamma_1 &= \frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\sqrt{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^3}=\frac{\mu_3}{\sigma^3}\\
\gamma_2 &= \gamma_1\frac{n}{(n-1)(n-2)}\\
\gamma_3 &= \gamma_1\left(\frac{n-1}{n}\right)^{3/2}\\
\gamma_{\text{Pearson mode}} &= \frac{\mu-M_0}{\sigma}\\
\gamma_{\text{Pearson median}} &= \frac{3(\mu-M_e)}{\sigma}\\
\gamma_{\text{Bowley}} &= B_1=\frac{Q_3+Q_1-2Q_2}{Q_3-Q_1}=\frac{\dfrac{Q_3+Q_1}{2}-Q_2}{\dfrac{Q_3-Q_1}{2}}\\
&\ldots
\end{aligned}\tag{5.7.348}$$

and the kurtosis used by `R` (implements κ_3, κ_4, κ_5), MATLAB® (implements κ_1, κ_2), Mathematica (implements κ_1), and MINITAB/IBM SPSS and Microsoft® Excel® (that implement

κ_3):

$$\begin{aligned}
\kappa_1 &= \frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left(\sqrt{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^4} = \frac{\mu_4}{\sigma^4}\\
\kappa_2 &= \frac{(N-1)((N+1)\kappa_1-3(N+1))}{(N-2)(N-3)}+3\\
\kappa_3 &= \frac{N(N+1)}{(N-1)(N-2)(N-3)}\kappa_1-\frac{3(N-1)^2}{(N-2)(N-3)}\\
\kappa_4 &= \kappa_1-3\\
\kappa_5 &= (\kappa_4+3)\left(1-\frac{1}{N}\right)^2-3\\
&\ldots
\end{aligned} \tag{5.7.349}$$

From the relations above, only γ_1, γ_2, κ_1 and κ_3 are actually in the ISO 3534-1 standard.

Remark

It exist at least 11 methods to calculate the Skewness. See the study by Tabor, J. (12010 according to holocene calendar), *Investigating the Investigative Task: Testing for Skewness - An Investigation of Different Test Statistics and their Power to Detect Skewness*, Journal of Statistics Education, 18, 1-13. www.amstat.org/publications/jse/v18n2/tabor.pdf

Skewness γ_1 can be expressed in terms of the non-central moment $\mathrm{E}[X^3]$ by expanding its general definition (this form is very useful in practice to calculate the explicit skewness of a given distribution!):

$$\begin{aligned}
\gamma_1 &= \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]\\
&= \frac{\mathrm{E}[X^3]-3\mu\mathrm{E}[X^2]+3\mu^2\mathrm{E}[X]-\mu^3}{\sigma^3}\\
&= \frac{\mathrm{E}[X^3]-3\mu(\mathrm{E}[X^2]-\mu\mathrm{E}[X])-\mu^3}{\sigma^3}\\
&= \frac{\mathrm{E}[X^3]-3\mu\sigma^2-\mu^3}{\sigma^3}
\end{aligned} \tag{5.7.350}$$

For a sample of n values, a natural method of moments estimator of the population skewness γ_1 is obviously as already defined earlier above:

$$\gamma_1 = \frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\sqrt{\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(x_i-\bar{x})^3}\right)^3} \tag{5.7.351}$$

Same for the Kurtosis:

$$\begin{aligned}\kappa_1 &= \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]\\ &= \frac{\mathrm{E}[X^4]-4\mu\mathrm{E}[X^3]+6\mu^2\mathrm{E}[X^2]-4\mu^3\mathrm{E}[X]-\mu^4}{\sigma^4}\end{aligned} \tag{5.7.352}$$

Here is a figure giving an idea of the range values of skewness squared and kurtosis for some distributions:

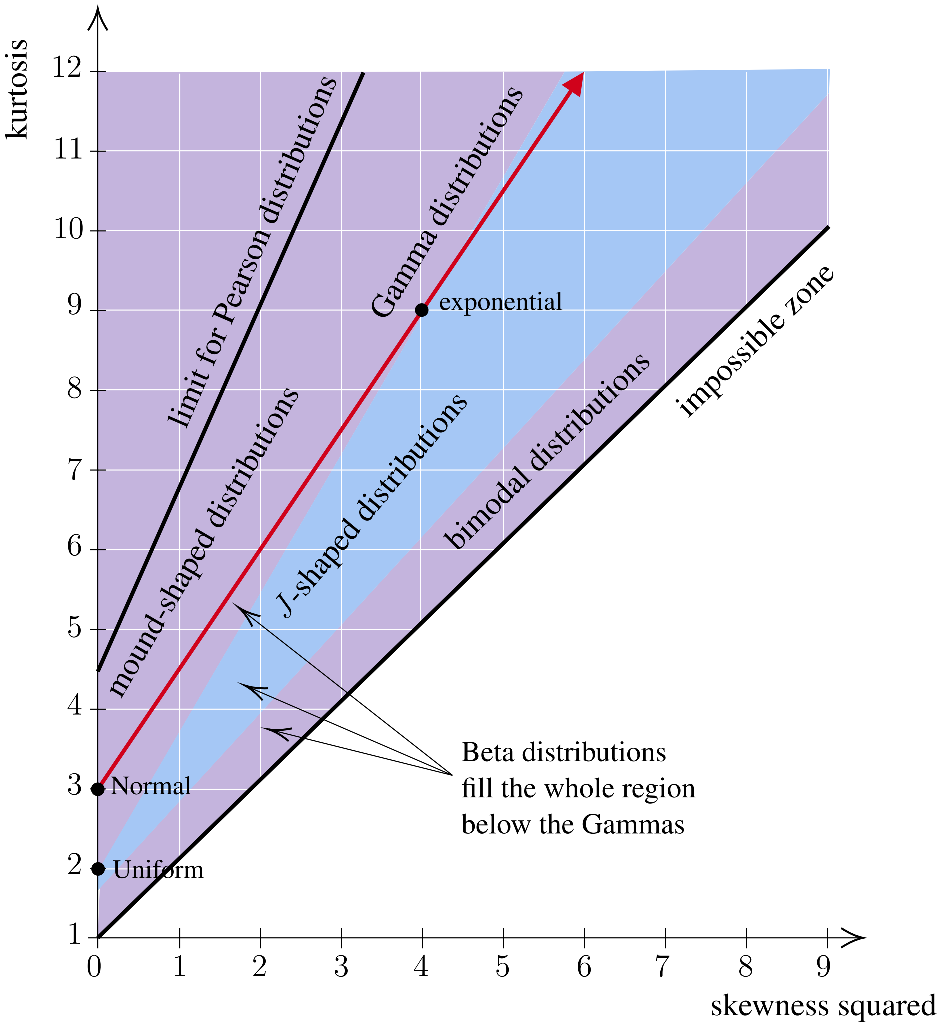


Figure 5.117: The Beta and Gamma families of distributions on the shape characterization plane

Later during the study of some important distributions, we will prove in details the values of the skewness and kurtosis for: the Uniform, Triangular, Normal, Student and the Beta distribution.

But before doing this let us prove an interesting theorem!

Theorem 5.46. Given a random variable X with symmetric probability density function $f(x)$, we have $\mathrm{E}(X)=a\mathrm{E}(X)=a$ where a is the point of symmetry assuming that $\mathrm{E}(X)$ exists.

Proof 5.46.1. First, it is obvious that symmetry about a means according to the assumptions of the theorem statement that:

$$f(a+z)=f(a-z) \tag{5.7.353}$$

for all z.

Suppose that the expectation $\mathrm{E}(X)$ exists. Then we know that:

$$\mathrm{E}(X)=\int\limits_{-\infty}^{+\infty} xf(x)\mathrm{d}x \tag{5.7.354}$$

Let us rewrite this as:

$$\mathrm{E}(X)=\int\limits_{-\infty}^{+\infty} (a+(x-a))f(x)\mathrm{d}x \tag{5.7.355}$$

which is equal to:

$$\int\limits_{-\infty}^{+\infty} af(x)\mathrm{d}x+\int\limits_{-\infty}^{+\infty} (x-a)f(x)\mathrm{d}x \tag{5.7.356}$$

The first integral is obviously equal to a as $f(x)$ is a density function.

In the second integral we make the change of variable $y=x-a$. Then this second integral becomes:

$$\int\limits_{-\infty}^{+\infty} yf(a+y)\mathrm{d}y \tag{5.7.357}$$

Break up at $y=0$. We get:

$$\int\limits_{-\infty}^{0} yf(a+y)\mathrm{d}y+\int\limits_{0}^{+\infty} yf(a+y)\mathrm{d}y \tag{5.7.358}$$

For the first integral, make the change of variable $z=-y$. We get:

$$\int\limits_{+\infty}^{0} (-1)(-z)f(a-z)\mathrm{d}z \tag{5.7.359}$$

Using $f(a-z)=f(a+z)$, and some minor fooling with minus signs, we end up with:

$$-\int\limits_{0}^{+\infty} zf(a+z)\mathrm{d}z \tag{5.7.360}$$

Therefore:

$$\int\limits_{-\infty}^{0} yf(a+y)\mathrm{d}y+\int\limits_{0}^{+\infty} yf(a+y)\mathrm{d}y \tag{5.7.361}$$

becomes:

$$-\int_{0}^{+\infty} zf(a+z)\mathrm{d}z+\int_{0}^{+\infty} yf(a+y)\mathrm{d}y \tag{5.7.362}$$

And this is equal to zero! Then we have:

$$\int_{-\infty}^{+\infty} af(x)\mathrm{d}x+\int_{-\infty}^{+\infty}(x-a)f(x)\mathrm{d}x=a+\left(-\int_{0}^{+\infty} zf(a+z)\mathrm{d}z+\int_{0}^{+\infty} yf(a+y)\mathrm{d}y\right)=a+0=a \tag{5.7.363}$$

■ Q.E.D.

Remark

The skewness and kurtosis are fundamental moments to develop the famous Jarque-Bera test of normality or the famous Cornish-Fisher Value at Risk (we will see all that later in this textbook).

5.7.4 Fundamental postulate of statistics

One of the ultimate goals of statistics is, starting from a sample, to find the analytical distribution function that gave birth to the sample. This goal will be presented in this textbook as a postulate (even if this assumption is quite difficult to apply in practice).

Postulate: For any empirical distribution function $\hat{F}_n(x)$ of the n-th measurement of the x random variable we can associate a theoretical distribution function $F(x)$ to which it converges when the sample size is large enough if:

$$X_n=\sup|\hat{F}_n(x)-F(x)| \tag{5.7.364}$$

is the random variable defined as the largest difference (in absolute value) between $\hat{F}_n(x)$ and $F(x)$ (observed for all values of x for a given sample), then X_n converges to 0 almost surely.

Remark

Mathematicians of Statistics prove this postulate rigorously as a theorem named the "fundamental theorem of statistics" or the "Glivenko-Cantelli theorem" regarding continuous functions. Personally, even if we offends the experts, we think that this proof is not one because it is very far away from the practical reality (yes this is our physicist side that emerges...) and this theoretical result leads many practitioners do their utmost (excluding data, transformations and other abominations) to find a known distribution law that they can adjust to their measured data.

5.7.5 Diversity Index

It happens in the field of biology or business that you it is asked to a statistician or analyst to measure the diversity of a number of predefined elements. For example, imagine a multinational with a range of well-defined products and some of the stores (customers) in the world can choose a subset of this range for their business sales. The request is then to make a ranking of stores that sell the widest range of branded products and that by taking also into account the quantity.

For example, we have a list of a total 4 products in our catalogue. By hazard, three of our customers sell our 4 products but we would like to know which customers sells the greatest diversity and this by taking into account the quantities.

We have the following sales data by product for the customer 1:

Customer 1	
Product 1	5
Product 2	5
Product 3	5
Product 4	5

For the customer 2:

Customer 2	
Product 1	1
Product 2	1
Product 3	1
Product 4	17

and for the customer 3:

Customer 3	
Product 1	2
Product 2	2
Product 3	2
Product 4	34

A measure of information (diversity of states) that is well suited to this purpose is the Shannon formula introduced in the section of Statistical Mechanics (page 3530) whose mean:

$$S(x)=\mathrm{E}(h(x))=-\lambda\sum_{i=1}^{n}p_i\log(p_i) \tag{5.7.365}$$

Arbitrarily, we will take $\lambda=1$ and the logarithm in base 10 (so, if we have 10 equiprobable variables, entropy is unitary for example...).

Therefore we have:

$$S(x)=-\sum_{i=1}^{n}p_i\log_{10}(p_i) \tag{5.7.366}$$

We will rewrite this more adequately for the application in business. Thus, if n is the number of products and p_i the proportion (or "relative frequency") of sales of product i from all sales N then:

$$p_i=\frac{f_i}{N} \tag{5.7.367}$$

Then we have:

$$\begin{aligned}S(x)&=-\sum_{i=1}^{n}\frac{f_i}{N}\log_{10}(\frac{f_i}{N})=-\frac{1}{N}\sum_{i=1}^{n}f_i\,[\log_{10}(f_i)-\log_{10}(N)]\\&=\frac{\displaystyle\sum_{i=1}^{n}f_i\,[\log_{10}(N)-\log_{10}(f_i)]}{N}=\frac{\log_{10}(N)\displaystyle\sum_{i=1}^{n}f_i-\sum_{i=1}^{n}f_i\log_{10}(f_i)}{N}\\&=\frac{N\log_{10}(N)-\displaystyle\sum_{i=1}^{n}f_i\log_{10}(f_i)}{N}\end{aligned} \tag{5.7.368}$$

This gives for the customer 1 (we stay in base 10 for the logarithm):

$$\begin{aligned}S_1&=\frac{N\log(N)-\displaystyle\sum_{i=1}^{n}f_i\log(f_i)}{N}\\&=\frac{20\log(20)-(5\log(5)+5\log(5)+5\log(5)+5\log(5))}{20}\\&=\frac{20\log(20)-20\log(5)}{20}=\log(20)-\log(5)=\log(4)=0.602\end{aligned} \tag{5.7.369}$$

which is the maximum possible value (each state is equally likely). And for customer 2 we have:

$$\begin{aligned}S_2&=\frac{N\log(N)-\displaystyle\sum_{i=1}^{n}f_i\log(f_i)}{N}\\&=\frac{20\log(20)-(1\log(1)+1\log(1)+1\log(1)+17\log(17))}{20}=0.255\end{aligned} \tag{5.7.370}$$

And finally for customer 3:

$$\begin{aligned}S_3&=\frac{N\log(N)-\displaystyle\sum_{i=1}^{n}f_i\log(f_i)}{N}\\&=\frac{40\log(40)-(2\log(2)+2\log(2)+2\log(2)+34\log(34))}{40}=0.255\end{aligned} \tag{5.7.371}$$

Thus, the customer that has the greatest diversity is the first one. We also see an interesting property of the Shannon formula with customer 2 and 3 and this is that the quantity does not affect diversity (since the only difference between the two customers is that the quantity is multiplied by a factor of 2 and not diversity)!

Notice that if we put all the products in only one customer (set), we get:

$$\begin{aligned}S_{\text{tot}} &= \frac{N\log(N)-\displaystyle\sum_{i=1}^{n}f_i\log(f_i)}{N} \\ &= \frac{80\log(80)-(8\log(8)+8\log(8)+8\log(8)+56\log(56))}{80}=0.408\end{aligned} \tag{5.7.372}$$

Therefore we guess that (without general proof):

$$\boxed{S_{\text{tot}}\leq\sum_i S_i} \tag{5.7.373}$$

Hence the fact that it is better to have various product in one place than in multiple place (result used for example in the field of document management where the purpose is to avoid to have the same type of documents saved in different share drive or libraries).

5.7.6 Distribution Functions (probabilities laws)

When we observe probabilistic phenomena, and we take note of the values taken by them and that we report them graphically, we can observe that the individual measurements follow a typical characteristic which is sometimes adjustable theoretically with a good level of quality.

In the field of probabilities and statistics, we name these characteristics "distribution functions" because they indicate the frequency with which the random variable appears for given values.

Remark

We sometimes simply use the term "function" or "law" to describe these characteristics.

These functions are in practice bounded by what we name the "range of the distribution" which is the difference between the maximum value (on the right) and the minimum value (on the left) of the observed values:

$$R = \max - \min \tag{5.7.374}$$

In theory they are not necessarily bounded and then we talk (see section Functional Analysis page 2207) about a "domain of definition" or more simply about the "support" of the function.

If the observed values are distributed in a certain way then there is a probability (or "cumulative probability" in the case of continuous distribution functions) to have a certain value of the distribution function.

In industrial practice (see section Industrial Engineering page 8684), the range of statistical values is important (as well as the standard deviation) because it gives an indication of the variation of a process (variability).

If L denote any possible univariate distribution function the range of the function is simply denoted by L if its domain of definition is $\mathbb{R}$ otherwise if it is bounded you will typically see something like $L_{]a,b]}$.

Definitions (#5.7.133):

D1. The mathematical relation that gives the probability of a given value of the distribution function a random variable is named the "density function" (or "probability density function"), "mass function" or "marginal function".

D2. The mathematical relation that gives the cumulative probability that a random variable to be lower than or equal to a certain value of the distribution function is referred to as the "repartition function" or "cumulative function" or "cumulative distribution function".

D3. Random variables are "independent and identically distributed (i.i.d.)" if they all follow the same distribution function, with the same parameters values and that they are independent.

Such functions are very numerous, we offer then here to the reader a detailed study of the most known only.

Before going any further it could be useful to know that if X is a continuous or discrete random variable, then are several tradition of notation in the literature to indicate that it follows a given probability distribution L. Here are the most common:

$$\begin{aligned} X &\sim L\\ X &\overset{\mathcal{L}}{=} L\\ X &\mapsto L\\ X &= L \end{aligned}\tag{5.7.375}$$

In this section and throughout the book in general, we will use the last notation!

Here is the list of the univariate and multivariate distribution functions that we will see here as well as distribution functions commonly used in the industry and located in other chapters/sections and those whose proof has yet still to be written:

- Discrete Uniform distribution $\mathcal{U}(a,b)$ (see below)
- Bernoulli distribution $\mathrm{B}(1,p)$ (see below)
- Geometric distribution $\mathcal{G}(N)$ (see below)
- Binomial distribution $\mathcal{B}(N,k)$ (see below)
- Binomial Negative distribution $\mathrm{NB}(N,k,p)$ (see below)
- Hypergeometric distribution $\mathcal{H}(n,p,m,k)$ (see below)
- Multinomial distribution $\mathcal{H}(\vec{k},\vec{p},m)$ (see below)
- Poisson distribution $\mathcal{P}(\mu,k)$ (see below)
- Gauss-Laplace/Normal distribution $\mathcal{N}(\mu,\sigma)$ (see below)
- General Folded Normal distribution (see below)
- Half-Normal distribution (see below)
- Log-Normal distribution $\mathcal{LN}(\mu,\sigma)$ (see below)
- Continuous Uniform distribution (see below)
- Triangular distribution (see below)
- Pareto distribution (see below)
- Exponential distribution (see below)
- Weibull distribution (see section Industrial Engineering page 8736)
- Generalized Exponential distribution (still must be written)
- Erlang/Erlang-B/Erlang-C distributions (see section Quantitative Management page 8147)
- Cauchy distribution (see below)
- Beta distribution (below and section Quantitative Management page 8076)
- Gamma distribution (see below)
- Chi-squared distribution (see below)
- Student distribution (see below)
- Fisher-Snedecor distribution (see below)
- Benford distribution (see below)
- Logistic distribution (see section Numerical Methods page 6453)
- Laplace distribution $\mathcal{L}(\mu,b)$ (see below)
- Square Gauss Distribution (see below)
- Exponentially modified Gaussian distribution (see below)
- Extreme value Distribution (still must be written)
- Vasicek distribution (already written but need to be reviewed)

All these distributions are mathematical evidence that randomness can produce precision!

By the way, for a more exhaustive list (almost a hundred!) and classification of distribution functions, the reader can see refer to the Data Science mind map at page 6770.

Remark

The reader will find the mathematical developments of the Weibull distribution function in the section on Industrial Engineering (page 8736), and the logistic distribution function in the section of Numerical Methods (page 6453).

> *All epistemologic value of the theory of probability is based on this: that large-scale random phenomena in their collective action create strict, non-random regularity.*
>
> Vladimirovitch Gnedenko and Andrey Kolmogorov

quote 5.7:649.#46

5.7.6.1 Discrete Uniform Distribution

If we accept that it is possible to associate a probability to an event, we can conceive of situations where we can assume a priori that all elementary events are equally likely (that is to say, they have the same probability to occur). We then use the ratio between the number of favourable cases and the number of possible cases to calculate the probability of all events in the Universe of events U. More generally, if U is a finite set of equally likely events and A is part of U, then we have using set theory notation (see section Set Theory page 388):

$$P(A)=\frac{\text{Card}(A)}{\text{Card}(U)}=\frac{\#A}{\#U} \tag{5.7.376}$$

More commonly, if e is an event that may have N equally likely possible outcomes. Then the probability of observing the outcome of this given event follows a "discrete uniform function" (or "discrete uniform law") given by the relation:

$$\boxed{P_e=\frac{1}{N}} \tag{5.7.377}$$

And obviously this is also equal to the modal (mode) value of the distribution.

Whose mean (or average) is given by:

$$\text{E}(X)=\sum_i p_i x_i=\sum_i P_e x_i=P_e\sum_i x_i=\frac{1}{N}\sum_i x_i \tag{5.7.378}$$

If we put ourselves in the particular case where $x_i=1$ with $i=1\ldots N$. We then have (see section Sequences and Series page 1666):

$$\text{E}(X)=\frac{1}{N}\sum_i x_i=\frac{1}{N}\sum_{i=1}^{N} i=\frac{1}{N}\frac{N(N+1)}{2}=\frac{N+1}{2} \tag{5.7.379}$$

Obviously the as this distribution is symmetric to the mean, the median is equal to mean.

If the random variable e take all values between $[a,b]$ (another special case) such the distribution will be now denoted by $\mathcal{U}(a,b)$ and the probability mass function of e is:

$$f_e(x)=\frac{1}{b-a+1}\qquad\text{where}\qquad x\in\{a,a+1,\ldots,b-1,b\}\tag{5.7.380}$$

as the number of integers between and including a and b is $n=b-a+1$. It is immediate that the cumulative distribution function is given by:

$$F_e(x)=\begin{cases}0\text{ if }x<a\\ \dfrac{\lfloor x\rfloor-a+1}{b-a+1}\text{ if }a\leq x\leq b\\ 1\text{ if }x>b\end{cases}\tag{5.7.381}$$

and it should be obvious that we have for the expected mean:

$$\begin{aligned}\mathrm{E}(X)&=\sum_{i=a}^{b}P_e i=P_e\sum_{i=a}^{b}i=\frac{1}{b-a+1}\sum_{i=a}^{b}i=\frac{1}{b-a+1}\left(\sum_{i=1}^{b}i-\sum_{i=1}^{a-1}i\right)\\&=\frac{1}{b-a+1}\left(\frac{b(b+1)}{2}+\frac{(a-1)((a-1)+1)}{2}\right)=\frac{1}{b-a+1}\frac{b(b+1)-a(a-1)}{2}\\&=\frac{1}{b-a+1}\frac{(b-a+1)(b+a)}{2}=\frac{a+b}{2}\end{aligned}\tag{5.7.382}$$

For the variance we have (always using the results of the section on Sequences and Series page 1667):

$$\begin{aligned}\mathrm{V}(X)&=\sum_{i=1}^{N}p_i(i-\mu)^2=\sum_{i=1}^{N}P_e(i-\mu)^2=\sum_{i=1}^{N}\frac{1}{N}(i-\mu)^2=\frac{1}{N}\sum_{i=1}^{N}(i-\mu)^2\\&=\frac{1}{N}\left(\sum_{i=1}^{N}i^2-2\mu\sum_{i=1}^{N}i+\sum_{i=1}^{N}\mu^2\right)=\frac{1}{N}\left(\sum_{i=1}^{N}i^2-2\mu\sum_{i=1}^{N}i+N\mu^2\right)\\&=\frac{1}{N}\sum_{i=1}^{N}i^2-\frac{2}{N}\mu\sum_{i=1}^{N}i+\mu^2=\frac{1}{N}\frac{N(N+1)(2N+1)}{6}-\frac{2}{N}\frac{N+1}{2}\sum_{i=1}^{N}i+\left(\frac{N+1}{2}\right)^2\\&=\frac{(N+1)(2N+1)}{6}-\frac{N+1}{N}\sum_{i=1}^{N}i+\frac{(N+1)^2}{4}\\&=\frac{(N+1)(2N+1)}{6}-\frac{N+1}{N}\frac{N(N+1)}{2}+\frac{(N+1)^2}{4}\\&=\frac{(N+1)(2N+1)}{6}-\frac{(N+1)^2}{2}+\frac{(N+1)^2}{4}=\frac{(N+1)(2N+1)}{6}-\frac{(N+1)^2}{4}\\&=\frac{2(N+1)(2N+1)-3(N+1)^2}{12}=\frac{2(2N^2+N+2N+1)-3N^2-6N-3}{12}\\&=\frac{4N^2+6N+2-3N^2-6N-3}{12}=\frac{N^2-1}{12}\end{aligned}\tag{5.7.383}$$

If the random variable e take all values between $[a,b]$ (another special case) such the distribution will be now denoted by $\mathcal{U}(a,b)$ then it should be obvious that we have for the variance:

$$\text{V}(X)=\frac{N^2-1}{12}=\frac{(b-a+1)^2+1}{12} \tag{5.7.384}$$

By symmetry of the distribution if all values of the domain of definition $[a,b]$ are taken by the random variable we have for the median:

$$M_e=\text{E}(X)=\frac{a+b}{2} \tag{5.7.385}$$

Here is a plot example of the mass distribution function and cumulative distribution function respectively for discrete uniform law of parameters $\{1,5,8,11,12\}$ (we see that each value is equally likely):

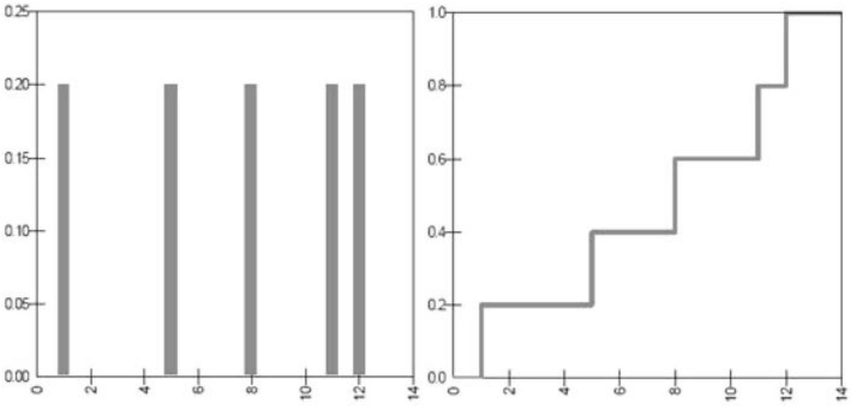


Figure 5.118: **Uniform law $\mathcal{U}$ (density and cumulative distribution function)**

As we can see in the above diagram the cumulative distribution function can be written:

$$F(x)=P(X\leq x)=\frac{\#(i:x_i\leq x)}{N} \tag{5.7.386}$$

Remark

For sure the discrete uniform distribution has no specific modal value M_0!

Let e be a random variable following a discrete uniform distribution $e\sim\mathcal{U}(a,b)$. Then, the quantile function of e is:

$$Q_e(p)=a(1-p)+(b+1)p-1 \quad \text{when } p\in\left\{\frac{1}{n},\frac{2}{n},\ldots,\frac{b-a}{n},1\right\} \tag{5.7.387}$$

with $n=b-a+1$. Remember that the quantile function $Q_X(p)$ is in general defined as the smallest x, such that $F_X(x)=p$:

$$Q_X(p)=\min\{x\in\mathbb{R}\mid F_X(x)=p\} \tag{5.7.388}$$

Because the cumulative distribution function only returns multiples of $1/n$ with $n = b - a + 1$, the quantile function is only defined for such values. First, we have $Q_X(p) = -\infty$, if $p = 0$. Second, since the cumulative probability increases step-wise by $1/n$ at each integer between and including a and b, the minimum x at which:

$$F_e(x) = \frac{c}{n} \quad \text{where} \quad c \in \{1, \ldots, n\} \tag{5.7.389}$$

is given by:

$$Q_e\left(\frac{c}{n}\right) = a + \frac{c}{n} \cdot n - 1 = a + c - 1 \tag{5.7.390}$$

Substituting $p = c/n$ and $n = b - a + 1$, we can finally show:

$$\begin{aligned} Q_e(p) &= a + p \cdot (b - a + 1) - 1 \\ &= a + pb - pa + p - 1 \\ &= a(1-p) + (b+1)p - 1 \end{aligned} \tag{5.7.391}$$

We will prove in the section of Statistical Mechanics (see page 3542) that the discrete uniform distribution maximizes (Shannon) entropy for a random variable with finite support.

5.7.6.2 Bernoulli Distribution

If we are dealing with a binary observation then the probability of an event is constant from one observation to the other if there is no memory effect (in other words: a sum of Bernoulli variables, two by two independent).

We name this kind of observations where the random variables takes the values 0 (false) or 1 (true), with probability $q = 1 - p$ respectively p, "Bernoulli trials" with "contrary events with contrary probabilities".

Thus, a random variable X follows a "Bernoulli function" $\text{B}(1, p)$ (or "Bernoulli law") if it can take only the values 0 or 1 , associated with probabilities p and q and so that $q + p = 1$ and:

$$\boxed{P(X = 0) = q \qquad P(X = 1) = p = 1 - q} \tag{5.7.392}$$

The classic example of such a process is the game of piece face or sampling with replacement or be considered as such (this last case is very important in industrial practice). There certainly is no need for the reader to formally verify that the cumulative probability is unitary...

Remark

The introduction above is perhaps not relevant for business, but we will see in the section of Quantitative Techniques (page 8141) that the Bernoulli function naturally appears at the beginning of our study of queuing theory.

Note that, by extension, if we consider N events where we get in a particular order k times one possible outcomes (success) and the other $N - k$ (fail) times, then the probability of such a series (k successes and $N - k$ failures ordered in any particular way) is given intuitively by (applying independent events rule):

$$P(N, k) = p^k (1-p)^{N-k} = p^k q^{N-k} \tag{5.7.393}$$

with $N \in \mathbb{N}^*$.

Here is an example plot of the cumulative distribution function for $q = 0.3$:

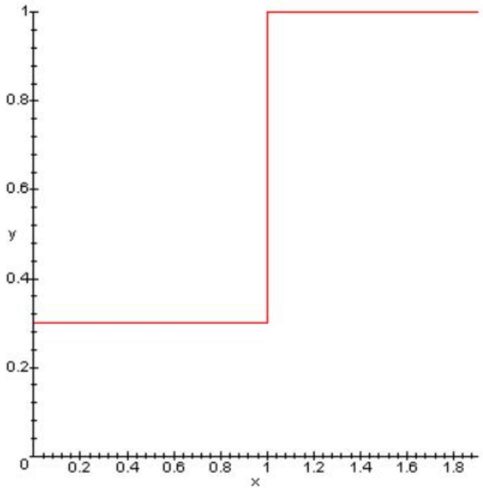


Figure 5.119: **Bernoulli law B (cumulative distribution function)**

The Bernoulli function has therefore for expected mean (average) choosing p as the probability of the event of interest:

$$\mu = \mathrm{E}(X) = \sum_i p_i x_i = p \cdot 1 + (1-p) \cdot 0 = p \tag{5.7.394}$$

and for variance (we use the Huygens theorem proved above):

$$\mathrm{V}(X) = \sigma^2 = \mathrm{E}(X^2) - \mathrm{E}(X)^2 = p - p^2 = p(1-p) = pq \tag{5.7.395}$$

The modal value M_0 of the Bernoulli law depends on the values of p or q. So we have (it could be obvious for the reader):

$$\begin{aligned} M_0 &= 0 \Leftrightarrow p < q \\ M_0 &= \{0,1\} \Leftrightarrow p = q \\ M_0 &= 1 \Leftrightarrow q > 0 \end{aligned} \tag{5.7.396}$$

Remark

For sure the Bernoulli distribution has no specific median value M_e!

5.7.6.3 Geometric Distribution

The geometric law $\mathcal{G}(N)$ or "Pascal's law" consist in a Bernoulli trial, where the probability of success is p and that of failure $q = 1 - p$ are constant, that we renew independently until the first success.

Remember that during our presentation of the Bernoulli law we have deduce an extension to N such that:

$$P(N,k) = p^k(1-p)^{N-k} = p^k q^{N-k} \tag{5.7.397}$$

Therefore the probability to get the first success $k = 1$ after N trials is:

$$\boxed{\mathcal{G}(N) = p(1-p)^{N-1} = pq^{N-1}} \tag{5.7.398}$$

with $N \in \mathbb{N}^*$.

As you can see, greater is N, smaller is the probability $\mathcal{G}(N)$. This can be seem non-logic but in fact it is! Indeed in the sentence «*the probability to get the first success after N trials*», you must not forget that it is written after and not during.

Therefore for sure... the probability to have $N - 1$ failures followed by 1 success will be always be smaller when N increase (have a look to the figure a little bit further below for $p = 0.5$, it may help to understand).

This law has for expected mean:

$$\mu = \mathrm{E}(X) = \sum_i p_i x_i = \sum_{N=1}^{+\infty} pq^{N-1}N = \sum_{N=1}^{+\infty} p(1-p)^{N-1}N = p\sum_{N=1}^{+\infty}(1-p)^{N-1}N \tag{5.7.399}$$

However, the last relation can also be written:

$$\sum_{N=1}^{+\infty} Nq^{N-1} = \frac{1}{(1-q)^2} \tag{5.7.400}$$

Indeed, we proved in the section of Sequences and Series (page 1679) during our study of geometric series that:

$$\sum_{k=0}^{n} q^k = \frac{1-q^{n+1}}{1-q} \tag{5.7.401}$$

Taking the limit $n \to +\infty$ when we get:

$$\sum_{k=0}^{+\infty} q^k = \frac{1}{1-q} \tag{5.7.402}$$

because $0 \leq q < 1$. Then we just derive both members of equality with respect to q and we get:

$$\sum_{k=1}^{+\infty} kq^{k-1} = \frac{1}{(1-q)^2} \tag{5.7.403}$$

This done let us continue...

We have then the average number of trials X it takes to get the first success (or in other words, the expected rank - number of expected trials - to see the first success):

$$\text{E}(X)=\sum_{N=0}^{+\infty}NP(X=N)=\sum_{N=0}^{+\infty}Npq^{N-1}=\frac{p}{(1-q)^2}=\frac{p}{p^2}=\frac{1}{p} \tag{5.7.404}$$

Now we calculate the variance and reminding once again (Huygens theorem):

$$\text{V}(X)=\text{E}(X^2)-\text{E}(X)^2 \tag{5.7.405}$$

So let's start by calculating $\text{E}(X^2)$:

$$\begin{aligned}\text{E}(X^2)&=\sum_{N=0}^{+\infty}N^2P(X=N)=p\sum_{N=0}^{+\infty}N^2q^{N-1}=p\sum_{N=0}^{+\infty}N(N-1+1)q^{N-1}\\&=p\sum_{N=1}^{+\infty}N(N-1)q^{N-1}+p\sum_{N=0}^{+\infty}Nq^{N-1}\end{aligned} \tag{5.7.406}$$

The last term of this expression is equivalent to the expected mean calculated previously. Thus:

$$p\sum_{N=0}^{+\infty}Nq^{N-1}=\frac{1}{p} \tag{5.7.407}$$

It remains to calculate:

$$p\sum_{N=1}^{+\infty}N(N-1)q^{N-1} \tag{5.7.408}$$

We have:

$$p\sum_{N=1}^{+\infty}N(N-1)q^{N-1}=pq\sum_{N=2}^{+\infty}N(N-1)q^{N-2} \tag{5.7.409}$$

But deriving the following equality:

$$\sum_{k=1}^{+\infty}kq^{k-1}=\frac{1}{(1-q)^2} \tag{5.7.410}$$

We get:

$$\sum_{k=1}^{+\infty}k(k-1)q^{k-2}=\frac{2}{(1-q)^3} \tag{5.7.411}$$

Therefore:

$$pq\sum_{N=2}^{+\infty}N(N-1)q^{N-2}=pq\frac{2}{(1-q)^3}=pq\frac{2}{p^3}=\frac{2q}{p^2} \tag{5.7.412}$$

Thus:

$$\mathrm{E}(X^2)=\frac{2q}{p^2}+\frac{1}{p^2} \tag{5.7.413}$$

Finally when it comes to ranking the expected variance of the first success (i.e.: the variance expected number before the first successful trials):

$$\mathrm{V}(X)=\sigma^2=\mathrm{E}(X^2)-\mathrm{E}(X)^2=\frac{2q}{p^2}+\frac{1}{p}-\frac{1}{p^2}=\frac{1-p}{p^2} \tag{5.7.414}$$

The modal value is easy to get because we need to find the value of N that maximize the definition of the geometric law:

$$pq^{N-1} \tag{5.7.415}$$

and we hope that it is immediate to the reader that this is satisfy when $N=1$ therefore:

$$M_0=pq^{1-1}=pq^0=p \tag{5.7.416}$$

Now let us determine the median M_e. For this, by definition we know we must have:

$$\sum_{N=1}^{M_e}pq^{N-1}=\sum_{N=M_e}^{+\infty}pq^{N-1}=0.5 \tag{5.7.417}$$

But we can rewrite:

$$\sum_{N=1}^{M_e}pq^{N-1}=\sum_{N=M_e}^{+\infty}pq^{N-1}=pq^{M_e-1}\sum_{N=0}^{+\infty}q^N=pq^{M_e-1}\frac{1}{1-q}=pq^{M_e-1}\frac{1}{p}=q^{M_e-1}=0.5 \tag{5.7.418}$$

Therefore (in base 10):

$$\log\left(q^{M_e-1}\right)=(M_e-1)\log(q)=\log(0.5) \tag{5.7.419}$$

Finally, based on our definition of the median, we get:

$$M_e=\frac{\log(0.5)}{\log(q)}+1 \tag{5.7.420}$$

Now we determine the cumulative function of the geometrical law. We start from:

$$\mathcal{G}(N)=pq^{N-1} \tag{5.7.421}$$

Then we have, by definition of the cumulative probability, that the experience is successful in the first N trials:

$$P(X\leq N)=1-\sum_{j=N+1}^{+\infty}pq^{j-1}=1-p\sum_{j=N+1}^{+\infty}q^{j-1} \tag{5.7.422}$$

with N being for sure an integer of values $0,1,2,\ldots$.

We write:

$$j-1=n+k \Rightarrow k=n-j+1 \tag{5.7.423}$$

We then have for the cumulated distribution function:

$$\begin{aligned}P(X \leq N) &= 1-p\sum_{k=0}^{+\infty} q^{N+k} = 1-p\sum_{k=0}^{+\infty} q^N q^k = 1-pq^N\sum_{k=0}^{+\infty} q^k = 1-pq^N\sum_{k=0}^{+\infty} q^k \\ &= 1-pq^N\frac{1}{1-q} = 1-(1-q)q^N\frac{1}{1-q} = 1-q^N\end{aligned} \tag{5.7.424}$$

✍**Example:**

You try late at night and in the dark, to open a lock with a bunch of five keys, without attention, because you are a little tired (or a little tipsy ...) you will try each key. Knowing that only one key will work, what is the probability of using the right key at the N-th test?

The solution is:

$$\mathcal{G}(N) = pq^{N-1} = \frac{1}{5}\left(\frac{4}{5}\right)^{N-1} \tag{5.7.425}$$

Plot of the mass function and cumulative distribution function for the Geometric law with parameter $p=0.5$:

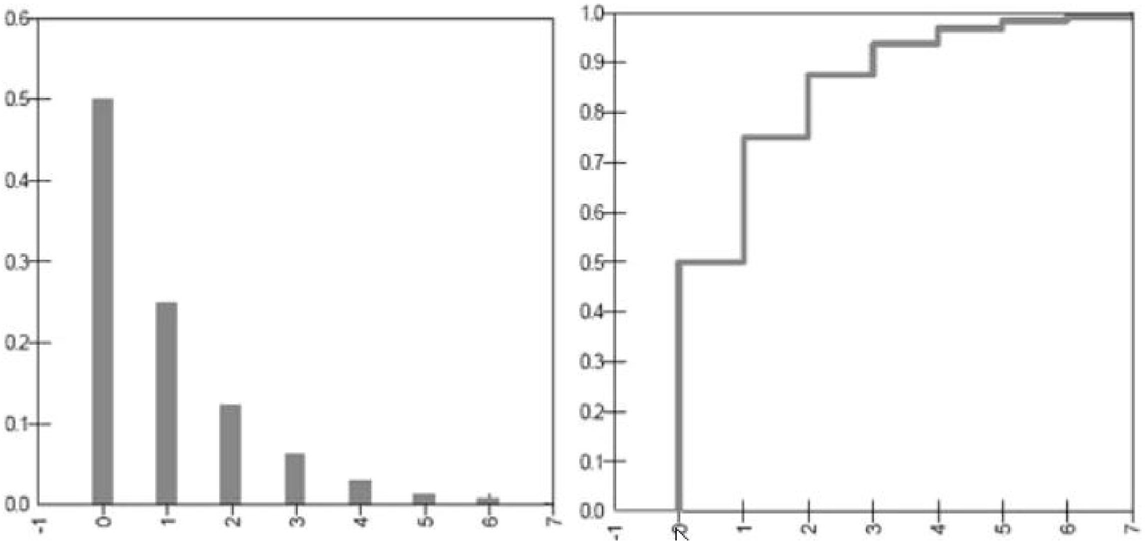


Figure 5.120: **Geometric law $\mathcal{G}$ (mass and cumulative distribution function)**

5.7.6.4 Binomial Distribution

We come back now to our Bernoulli experiment. More generally, any particular N-tuple consisting of k successes and of $N-k$ failures will have for probability (within a sampling with replacement or without replacement if the population is large ... in a first approximation):

$$P(N,k)=p^k(1-p)^{N-k}=p^kq^{N-k} \tag{5.7.426}$$

to be drawn (or appear) whatever the order of appearance of successes and failures (the reader will have perhaps notice that this is a generalization of the geometric distribution, just write $k=1$ to find the geometric distribution back).

But we know that the combinatorial determines the number of N-tuples of this type (the number of ways to order the appearance of failures and successes). The number of possible arrangements is, as we proved it (see section Probabilities page 528), given by the binomial coefficient (we recall that the notation in this textbook does not comply with ISO standard 31-11):

$$C_k^N=\frac{N!}{k!(N-k)!} \tag{5.7.427}$$

So as the probability of obtaining a given series of k successes and $N-k$ failures is always the same (regardless of the order) then we have just to multiply the probability of a particular series by the binomial coefficient (this is equivalent to a sum) such that:

$$P(N,k)=C_k^NP=\frac{N!}{k!(N-k)!}p^k(1-p)^{N-k}=\frac{N!}{k!(N-k)!}p^kq^{N-k} \tag{5.7.428}$$

to get the total probability to obtain any of these possible series (since each is possible).

Remark

This is equivalent to the study of a sampling with simple replacement (see section Probabilities page 528) with constraint on the order or to the study of a series of successes and failures. We will use this relation in the context of the queuing theory or reliability (see section Industrial Engineering page 8673). Note that in the case of large populations, even if the sampling is not with replacement it can be considered as with...

Written in another way this gives the "binomial function" (or "binomial law"), also known as the following distribution function:

$$\boxed{\mathcal{B}(N,k)=C_k^Np^k(1-p)^{N-k}=C_k^Np^kq^{N-k}=\binom{N}{k}p^kq^{N-k}} \tag{5.7.429}$$

and sometimes also denoted by $\beta(n,p)$ with a lowercase n or uppercase N (it does not really matter...) and can be calculated in the English version of Microsoft® Excel® 11.8346 using the `BINOMDIST( )` function.

We sometimes say that the binomial law is not exhaustive as the size of the initial population is not apparent in the expression of the law.

The binomial distribution appears across a wide span of scientific and engineering fields wherever researchers model processes made of repeated, independent trials with two possible outcomes. It is central in biology and medicine for analyzing genetic inheritance patterns, mutation rates, and clinical trial success counts; in physics and chemistry for modeling particle detection events, reaction successes, or defect formation; in computer science and information theory for packet-loss modeling, randomized algorithms, and error-correcting codes; in electrical and mechanical engineering for reliability analysis, component failure probabilities, and quality-control sampling; in environmental science for species-occurrence surveys and detection probabilities; and in social sciences for survey response modeling and behavioral event counts, and it also plays a clear role in quantitative finance, where it underpins discrete-time models of asset prices, option valuation, and risk assessment. Across these domains, the binomial distribution provides a mathematically clean way to quantify how often a “success” should occur when conditions are stable and repeatable.

Remark

The Binomial distribution is named "Symmetric Binomial Distribution" when $p = 0.5$.

✍**Examples:**

E1. We want to test the alternator of a generator. The probability of failure at solicitation of this material is estimated to be 1 failure per $1,000$ starts.

We decided to test 100 starts. The probability of observing one failure in this test is:

$$\begin{aligned}\mathcal{B}(N=100,k=1)=C_k^N p^k(1-p)^{N-k}&=\frac{N!}{k!(N-k)!}p^k(1-p)^{N-k}\\&=\frac{100!}{1!(100-1)!}\left(\frac{1}{1000}\right)^1\left(1-\frac{1}{1000}\right)^{99}\cong 9\%\end{aligned}\tag{5.7.430}$$

E2. Maybe you remember the prosecutor’s fallacy we have introduced at page 477. Let us come back on the example related to the crime-scene DNA sample. Now that we have introduced the binomial distribution, we can calculate the probability of getting at least one match among the record:

$$\mathcal{B}(N=20000,k=1)=C_k^N p^k(1-p)^{N-k}=\frac{N!}{k!(N-k)!}p^k(1-p)^{N-k}\cong 27.07\%\tag{5.7.431}$$

So, this evidence alone is an uncompelling data hacking result. If the culprit was in the database then he and one or more other men would probably be matched; in either case, it would be a fallacy to ignore the number of records searched when weighing the evidence. "Cold hits" like this on DNA databanks are now understood to require careful presentation as trial evidence.

We obviously have for the cumulative distribution function (very useful in practice for suppliers batch control or reliability as we will see in the section of Industrial Engineering page 8904!):

$$\sum_{k=0}^{N}C_k^N p^k(1-p)^{N-k}=1\tag{5.7.432}$$

Indeed, we have proved in the section Calculus (page 1364) the "binomial theorem":

$$(x+y)^n=\sum_{k=0}^{n} C_k^n x^k y^{n-k}=1 \tag{5.7.433}$$

Therefore:

$$\sum_{k=0}^{N} C_k^N p^k (1-p)^{N-k}=(p+(1-p))^N=1^N=1 \tag{5.7.434}$$

Instead of calculating such cumulated probability rather than hand it is better to use Microsoft® Excel® 11.8346 (or any other widely known software) with the function `CRITBINOM()` to not bother to calculate these type of values.

The expected mean (average) of $\mathcal{B}(N,k)$ is given by:

$$\mathrm{E}(X)=\sum_{k=0}^{N} p_i k=\sum_{k=0}^{N} P(X=k)k=\sum_{k=0}^{N} C_k^N p^k (1-p)^{N-k}k=\sum_{k=1}^{N} C_k^N p^k (1-p)^{N-k}k \tag{5.7.435}$$

But having:

$$C_k^N=\frac{N}{k}C_{k-1}^{N-1}\Leftrightarrow kC_k^N=NC_{k-1}^{N-1} \tag{5.7.436}$$

We finally get:

$$\begin{aligned}\mathrm{E}(X)&=\sum_{k=1}^{N} C_k^N p^k (1-p)^{N-k}=N\sum_{k=1}^{N} C_{k-1}^{N-1} p^k (1-p)^{N-k}\\&=Np\sum_{k=1}^{N} C_{k-1}^{N-1} p^{k-1} (1-p)^{(N-1)-(k-1)}=Np\sum_{k=0}^{N-1} C_k^{N-1} p^k (1-p)^{(N-1)-k}\\&=Np(p+(1-p))^{N-1}=Np\end{aligned} \tag{5.7.437}$$

that gives the average number of times that we will get the desired outcome of probability p after N trials.

The mean of the binomial distribution is sometimes noted in the specialized literature with the following notation if r is the potential number of possible expected outcomes in a population of size n:

$$\mathrm{E}(X)=Np=N\frac{r}{n} \tag{5.7.438}$$

Before calculating the variance, we need to introduce the following equality:

$$\sum_{s=0}^{N-1} s\frac{(N-1)!}{s!(N-1-s)!}p^s q^{N-1-s}=(N-1)p \tag{5.7.439}$$

Indeed, let us proof this relation using the previous developments:

$$\begin{aligned}&\sum_{s=0}^{N-1} s\frac{(N-1)!}{s!(N-1-s)!}p^sq^{N-1-s}=\sum_{s=0}^{N-1} s\frac{(N-1)!}{s(s-1)!(N-1-s)!}p^sq^{N-1-s}\\&=\sum_{s=1}^{N-1}\frac{(N-1)!}{(s-1)!(N-1-s)!}p^sq^{N-1-s}=(N-1)\sum_{s=1}^{N-1}\frac{(N-2)!}{(s-1)!(N-1-s)!}p^sq^{N-1-s}\\&=(N-1)\sum_{j=0}^{N-2}\frac{(N-2)!}{j!(N-2-j)!}p^{j+1}q^{N-2-j}=(N-1)p\sum_{j=0}^{N-2}C_j^{N-2}p^jq^{N-2-j}\\&=(N-1)p\sum_{j=0}^{N-2}C_j^{N-2}p^jq^{N-2-j}\end{aligned} \tag{5.7.440}$$

We recognize in the last equality the cumulative distribution function that is equal to 1. Therefore:

$$(N-1)p\sum_{j=0}^{N-2}C_j^{N-2}p^jq^{N-2-j}=(N-1)p\cdot 1=(N-1)p \tag{5.7.441}$$

We start now the (long) calculation of the variance of the binomial distribution by using the previous results:

$$\begin{aligned}
\text{V}(k) &= \sum_{k=0}^{N} p_i(k-\mu)^2 = \sum_{k=0}^{N} \frac{N!}{k!(N-k)!}p^kq^{N-k}(k-Np)^2 = \sum_{k=0}^{N} \frac{N!}{k!(N-k)!}p^kq^{N-k}(k-\mu)^2 \\
&= \sum_{k=0}^{N} k^2\frac{N!}{k!(N-k)!}p^kq^{N-k} + \sum_{k=0}^{N} \mu^2\frac{N!}{k!(N-k)!}p^kq^{N-k} - \sum_{k=0}^{N} 2k\mu\frac{N!}{k!(N-k)!}p^kq^{N-k} \\
&= \sum_{k=0}^{N} k^2\frac{N!}{k!(N-k)!}p^kq^{N-k} + \mu^2\sum_{k=0}^{N} \frac{N!}{k!(N-k)!}p^kq^{N-k} - 2\mu\sum_{k=0}^{N} k\frac{N!}{k!(N-k)!}p^kq^{N-k} \\
&= \sum_{k=0}^{N} k^2\frac{N!}{k!(N-k)!}p^kq^{N-k} + \mu^2\underbrace{\sum_{k=0}^{N} C_k^N p^kq^{N-k}}_{=1} -2\mu\underbrace{\sum_{k=0}^{N} k\frac{N!}{k!(N-k)!}p^kq^{N-k}}_{=\mu=Np} \\
&= \sum_{k=0}^{N} k^2\frac{N!}{k!(N-k)!}p^kq^{N-k} + \mu^2 - 2\mu^2 = \sum_{k=0}^{N} k^2\frac{N!}{k!(N-k)!}p^kq^{N-k} - \mu^2 \\
&\underset{s=k-1}{=} -\mu^2 + N\sum_{s=0}^{N-1}(s+1)\frac{(N-1)!}{s!(N-1-s)!}p^sq^{N-(s+1)} \\
&= -\mu^2 + Np\sum_{s=0}^{N-1}(s+1)\frac{(N-1)!}{s!(N-1-s)!}p^sq^{N-s-1} \\
&= -\mu^2 + Np\sum_{s=0}^{N-1} s\frac{(N-1)!}{s!(N-1-s)!}p^sq^{(N-1)-s} + Np\sum_{s=0}^{N-1} s\frac{(N-1)!}{s!((N-1-s)!}p^sq^{(N-1)-s)} \\
&= -\mu^2 + Np\sum_{s=0}^{N-1} sC_s^{N-1}p^sq^{(N-1)-s} + Np\sum_{s=0}^{N-1} sC_s^{N-1}p^sq^{(N-1)-s)} \\
&= -\mu^2 + Np(N-1)p + Np\sum_{s=0}^{N-1} sC_s^{N-1}p^sq^{(N-1)-s)} \\
&= -\mu^2 + Np^2(N-1) + Np\cdot 1 = -\mu^2 + Np^2(N-1) + Np \\
&= -\mu^2 + N^2p^2 - Np^2 + Np = -\mu^2 + \mu^2 - Np^2 + Np = Np(1-p) \\
&= Npq
\end{aligned} \tag{5.7.442}$$

Finally:

$$\text{V}(k) = \sigma^2 = Np(1-p) = Npq \tag{5.7.443}$$

The standard deviation of the binomial distribution is sometimes noted in the specialized literature in the following way if r is the potential number of expected outcomes in a population of size n and s the not expected one:

$$\sigma = \sqrt{Npq} = \sqrt{N\frac{r}{n}\frac{s}{n}} = \frac{1}{n}\sqrt{Nrs} \tag{5.7.444}$$

Here is a plot example of the binomial $\mathcal{B}(10, 0.5)$ distribution and cumulative distribution function:

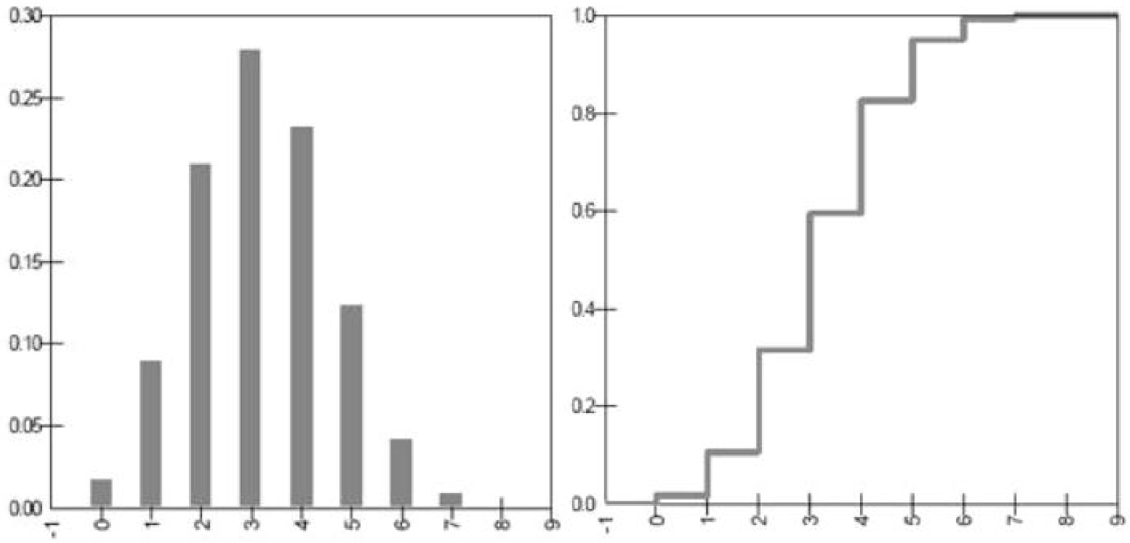


Figure 5.121: Binomial law $\mathcal{B}$ (mass and cumulative distribution function)

It could be useful to note that some employees in companies (especially physicians in the field of survival analysis - clinical studies - as we will see later!) normalize the calculation of the mean and standard deviation to the unit of N. Then we have:

$$\begin{gathered}\mathrm{E}\left(\frac{k}{N}\right)=\frac{Np}{N}=p\\ \mathrm{V}\left(\frac{k}{N}\right)=\frac{1}{N^2}\mathrm{V}(X)=\frac{1}{N^2}Npq=\frac{1}{N}p(1-p)\Leftrightarrow\sigma_{k/N}=\sqrt{\frac{p(1-p)}{N}}\end{gathered}\tag{5.7.445}$$

✍**Example:**

In a sample of 100 workers, 25% are late at least once a week. The mean number and variance of late people is then:

$$\begin{gathered}\mathrm{E}(k)=Np=100\cdot 0.25=25\\ \sigma_k=\sqrt{Np(1-p)}=\sqrt{100\cdot 0.25(1-0.25)}\cong 4.33\end{gathered}\tag{5.7.446}$$

Normalized to the unit of N this give us:

$$\begin{gathered}\mathrm{E}\left(\frac{k}{N}\right)=p=0.25\\ \sigma_{k/N}=\sqrt{\frac{p(1-p)}{N}}=0.0433\end{gathered}\tag{5.7.447}$$

Let us now calculate the mode. Because the function is discrete we can not use derivative. Then we will use a hint. We compute the ratio:

$$\frac{\mathcal{B}(N,k)}{\mathcal{B}(N,k+1)}\tag{5.7.448}$$

and we check that this ratio is > 1 for every $k < k^*$ and ≤ 1 for every $k \geq k^*$, for some integer k^* that is the k value corresponding to the modal value.

Let $a_k = P(X = k)$. We have:

$$a_k = C_k^N p^k q^{N-k} \qquad a_{k+1} = C_{k+1}^N p^{k+1} q^{N-k-1} \tag{5.7.449}$$

We calculate the ratio $\dfrac{a_{k+1}}{a_k}$. Note that:

$$\frac{a_{k+1}}{a_k} = \frac{C_{k+1}^N p^{k+1} q^{N-k-1}}{C_k^N p^k q^{N-k}} = \frac{\dfrac{N!}{(k+1)!(N-(k+1))!}}{\dfrac{N!}{k!(N-k)!}} \frac{p^{k+1} q^{N-k-1}}{p^k q^{N-k}} = \frac{n-k}{k+1}\frac{p}{q} = \frac{n-k}{k+1}\frac{p}{1-p} \tag{5.7.450}$$

What is important now is to analyse:

$$\frac{n-k}{k+1}\frac{p}{1-p} = \frac{np-kp}{k-kp+1-p} \tag{5.7.451}$$

depending on the value of k. First we can see that this ratio is equal to 1 and therefore we have two modes if:

$$\frac{np-kp}{k-kp+1-p} = 1 \Leftrightarrow np-kp = k-kp+1-p \tag{5.7.452}$$

That is to say if $k = np+p-1 = p(n+1)-1$. This can be seen as the limit point of interest. But don't forget we are looking for the k such that the ratio is less than 1. So we try two values:

$$\begin{aligned} k &= [p(n+1)-1]+1 \\ k &= [p(n+1)-1]-1 \end{aligned} \tag{5.7.453}$$

Injecting this in our ratio we see that:

$$k = [p(n+1)-1]+1 = k^* = M_0 \tag{5.7.454}$$

Is the value we were looking for! Finally there are two possible values for the modes. A unique modal value and a double modal value.

As we know, the median value is the value of X such that we have:

$$\sum_{k=0}^{X} C_k^N p^k (1-p)^{N-k} = 0.5 \tag{5.7.455}$$

But we did not yet found an easy proof to determine M_e in the general case for the Binomial law.

To conclude on the binomial law, we will develop now a result that we will need to build the McNemar paired test for a square contingency table (and as it is squared it is also dichotomous) that we will study further below (page 1264).

We need for this test to calculate the covariance of two-paired binomial random variables (this is why the covariance is non-zero):

$$\text{cov}(n_i,n_j)=\text{E}(n_in_j)-\text{E}(n_i)(n_j) \tag{5.7.456}$$

As they are paired, this means that:

$$\begin{gathered}n_i+n_j=n\\ p_i=1-p_j\\ 1-p_i=p_j\end{gathered} \tag{5.7.457}$$

And therefore:

$$\text{cov}(n_i,n_j)=\text{E}(n_in_j)-\text{E}(n_i)(n_j)=\text{E}(n_in_j)-(np_i)(np_j)=\text{E}(n_in_j)-n^2p_j(1-p_j) \tag{5.7.458}$$

Now comes the difficulty that is to calculate $\text{E}(n_in_j)$. To calculate this term it does not exist to our knowledge other methods than looking for the law of the pair (sometimes we can get around such approach). In this case it is a multinomial distribution (more precisely: trinomial) that it is customary to write in the following way by construction:

$$\mathcal{M}(n_i,n_j,p_i,p_j,1-p_i-p_j)=\frac{n!}{n_i!n_j!(n-n_i-n_j)!}p_i^{n_i}p_j^{n_j}(1-p_i-p_j)^{n-n_i-n_j} \tag{5.7.459}$$

that we will write now temporarily as following to condense the expression:

$$\mathcal{M}(n,k,l,p,q,r)=\frac{n!}{k!l!(n-k-l)!}p^kq^lr^{n-k-l}=\frac{n!}{k!l!(n-k-l)!}p^kq^l(1-p-q)^{n-k-l} \tag{5.7.460}$$

So we have a trinomial law as we are looking for the number of times we have the event k, the event l and neither one nor the other (so the rest of the time).

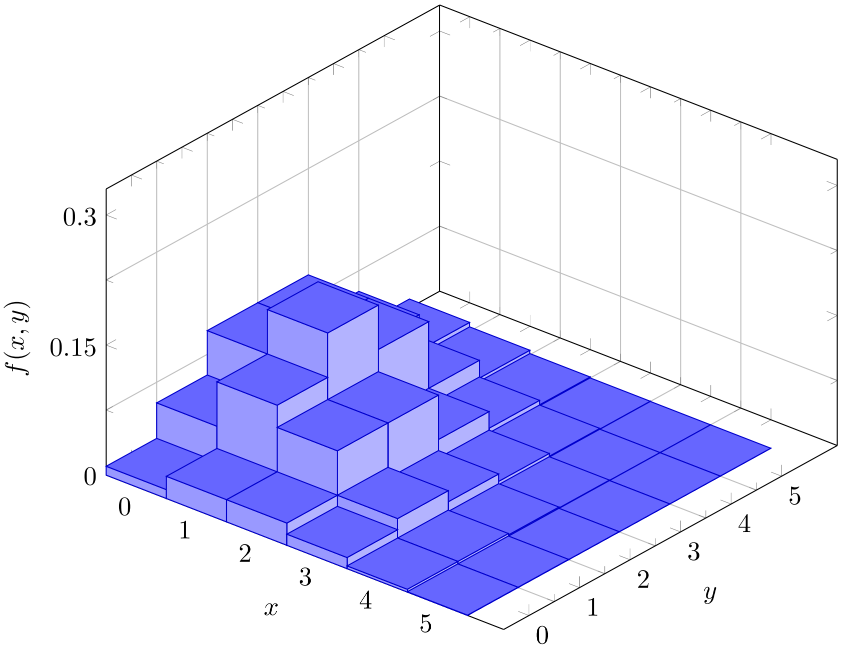


Figure 5.122: Trinomoial distribution with $p=1/5,q=2/5$ and $n=5$

We then get:

$$\begin{aligned}\mathrm{E}(kl)&=\sum_{k,l}klP(kl)=\sum_{k,l}\frac{n!}{k!l!(n-k-l)!}p^kq^lr^{n-k-l}\\&=\underbrace{0}_{k=0,l\neq 0}+\underbrace{0}_{l=0,k\neq 0}+\underbrace{0}_{k=0,l=0}+\sum_{k\geq 1,l\geq 1}kl\frac{n!}{k!l!(n-k-l)!}p^kq^lr^{n-k-l}\end{aligned}\tag{5.7.461}$$

If $k\geq 1$ and $l\geq 1$, we obtain:

$$\begin{aligned}kl\frac{n!}{k!l!(n-k-l)!}&=\frac{n!}{k(k-1)!l(l-1)!(n-k-l)!}=\frac{n(n-1)(n-2)!}{(k-1)!(l-1)!(n-k-l)!}\\&=n(n-1)\frac{(n-2)!}{(k-1)!(l-1)!(n-k-l)!}\end{aligned}\tag{5.7.462}$$

Now we use this relation in the joint mean:

$$\begin{aligned}\mathrm{E}(kl)&=\sum_{k\geq 1,l\geq 1}kl\frac{n!}{k!l!(n-k-l)!}p^kq^lr^{n-k-l}\\&=n(n-1)\sum_{1\leq k\leq n,1\leq l\leq n}kl\frac{(n-2)!}{k(k-1)!l(l-1)!(n-k-l)!}p^kq^lr^{n-k-l}\\&=n(n-1)pq\sum_{1\leq k\leq n,1\leq l\leq n}\frac{(n-2)!}{(k-1)!(l-1)!(n-k-l)!}p^{k-1}q^{l-1}r^{n-k-l}\end{aligned}\tag{5.7.463}$$

Consider now the special case where n is equal 2. We then have:

$$\begin{aligned}&\sum_{1\leq k\leq n,1\leq l\leq n}\frac{(n-2)!}{(k-1)!(l-1)!(n-k-l)!}p^{k-1}q^{l-1}r^{n-k-l}\\&=\sum_{1\leq k\leq 2,1\leq l\leq 2}\frac{0!}{(k-1)!(l-1)!(2-k-l)!}p^{k-1}q^{l-1}r^{2-k-l}\\&=\frac{0!}{(1-1)!(1-1)!(2-1-1)!}p^{1-1}q^{1-1}r^{2-1-1}=1\end{aligned}\tag{5.7.464}$$

where the sum is reduced to only one term because if we take for example $k=2,l=1$ we get a negative factorial at the denominator.

For n equal 3, the result will be also 1, and so on (we will assume to simplify... that some numerical examples will suffice to convince the reader of the generality of this property because it is very boring to write with LaTeX).

Then we have:

$$\mathrm{E}(kl)=n(n-1)pq\underbrace{\sum_{1\leq k\leq n,1\leq l\leq n}\frac{(n-2)!}{(k-1)!(l-1)!(n-k-l)!}p^{k-1}q^{l-1}r^{n-k-l}}_{=1}=n(n-1)pq\tag{5.7.465}$$

So in the end:

$$\text{cov}(n_i,n_j)=\text{E}(n_in_j)-n^2p_j(1-p_j)=n(n-1)p_ip_j-n^2p_ip_j=-np_ip_j \tag{5.7.466}$$

It does make sense for the covariance to be negative, since the more there are in category 1, the fewer are available for category 2. And this is the major result we will need for the study of the McNemar test.

Let X and Y be independent binomial random variables with parameters (n,p) and (m,p), respectively. Let $Z=X+Y$. Then the distribution of Z is given by:

$$\begin{aligned}
P(Z=k)&=\sum_{i=0}^{k}P(X=i)P(Y=k-i)\\
&=\sum_{i=0}^{k}\binom{n}{i}p^i(1-p)^{n-i}\binom{m}{k-i}p^{k-i}(1-p)^{m-(k-i)}\\
&=\sum_{i=0}^{k}\binom{n}{i}p^k(1-p)^{n-i}\binom{m}{k-i}(1-p)^{m-k+i}\\
&=\sum_{i=0}^{k}\binom{n}{i}p^k(1-p)^{n+m-k}\binom{m}{k-i}\\
&=p^k(1-p)^{n+m-k}\sum_{i=0}^{k}\binom{n}{i}\binom{m}{k-i}=p^k(1-p)^{n+m-k}\binom{n+m}{k}
\end{aligned} \tag{5.7.467}$$

Hence the binomial distribution is indeed stable by addition:

$$\mathcal{B}(n,p)+\mathcal{B}(m,p)=\mathcal{B}(n+m,p) \tag{5.7.468}$$

Doing the same development with different probabilities and without forgetting that $p_i+p_j=1$ then we have:

$$\mathcal{B}(n,p_i)+\mathcal{B}(m,p_j)=\mathcal{B}(n+m,p_i+p_J) \tag{5.7.469}$$

5.7.6.5 Negative Binomial Distribution

The negative binomial distribution is applied in the same situation as for binomial distribution, but it gives the probability to have E failures before the R-th success when the probability of success is p (or, at contrary, the probability to have R success before E-th failure when the failure probability is p).

We will introduce this important distribution with an example. Consider this for this purpose the following probabilities:

$$P(\text{success})=0.2=p \qquad P(\text{failure})=0.8=1-p=q \tag{5.7.470}$$

Imagine that we have done 10 trials and we wanted to stop at the third success and that the 10th trial is the third successful one! We will write this:

$$[1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9]\ 10 \tag{5.7.471}$$

Now we highlight what we will consider as the successes (R) and failures (E):

$$\begin{matrix}[1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9] & 10\\ [\texttt{E} & \texttt{E} & \texttt{R} & \texttt{E} & \texttt{E} & \texttt{E} & \texttt{R} & \texttt{E} & \texttt{E}] & \texttt{R}\end{matrix} \qquad (5.7.472)$$

We have also 7 failures and 3 successes. In an experiment where the draws are independent (or can be considered as independent...), the probability that we get this particular result is:

$$(0.8)^7(0.2)^3 \qquad (5.7.473)$$

But the order of successes and failures in the bracketed part is irrelevant. So as we have 2 success among the 9 trial in brackets it follows that the probability of obtaining the same result regardless of the order is then using combinatorics:

$$C_2^9(0.8)^7(0.2)^3=\binom{9}{2}(0.8)^7(0.2)^3\cong 0.0603 \qquad (5.7.474)$$

Which corresponds to the probability of having 7 failures before the 3rd success (or otherwise seen: 3 successes after 10 trials). This can be written with Microsoft® Excel® 14.0.6123 or later ($7+3=10$ trials, 7 failures including 3 successes):

```
=NEGBINOMDIST(7,3,0.2,0)=0.0604
```

We now generalize the prior-previous notation by writing the number of failures k, N the total number of trials and p the probability of success:

$$\text{NB}(N,k,p)=C_{N-k-1}^{N-1}q^kp^{N-k}=\binom{N-k-1}{N-1}q^kp^{N-k} \qquad (5.7.475)$$

However, there are several possible notations because the previous relation is not very intuitive to practice as may have perhaps noticed the reader. Thus, if we denote k as the number of successes and not the number of failures, then we have (the most common writing way from my point of view among a lot of others notations) the following probability of having $N-k$ success before having a number k of failures with a probability p (or of failures before having k successes ... it's symmetrical!):

$$\text{NB}(N,k,p)=C_{N-1}^{k-1}q^{N-k}p^k=\binom{N-1}{k-1}p^kq^{N-k} \qquad (5.7.476)$$

therefore the comparison with the formulation of the binomial distribution proved above is then perhaps more obvious!

However, it is more common to write the previous relation by removing N because for the moment the notation is still perhaps not very clear. For this, we note R the number of successes, E the number of failures, p the probability of success and then comes the probability of having R success after E failures (this is perhaps much more clear...):

$$\text{NB}(R,E,p)=C_{E+R-1}^{E-1}q^{(E+R)-E}p^E=\binom{E+R-1}{E-1}p^Eq^R \qquad (5.7.477)$$

We sometimes find this last relation with another relation using explicitly the binomial coefficient:

$$\begin{aligned}\text{NB}(R,E,p)&=\binom{E+R-1}{E-1}q^Rp^E=\frac{(E+R-1)!}{(E-1)!R!}q^Rp^E=\frac{(E+R-1)!}{R!(E-1)!}q^Rp^E\\&=\binom{E+R-1}{R}q^Rp^E\end{aligned} \tag{5.7.478}$$

The cumulative probability that we have at least R successes before the E-th failure is obviously given by:

$$\sum_{R=0}^{N}\binom{E+R-1}{E-1}q^Rp^E \tag{5.7.479}$$

Remark

The name of this law comes from the fact that some statisticians use a definition of the binomial coefficient with a negative value for the expression of the function. Since this is a rather a rare notation, we do not want lose time to prove the origin of the name. You should also know that this law is also known as the "Pascal's law" (as well as the geometric distribution ...) in honour of Blaise Pascal and also as "Polya's law" in honour of George Pólya.

✍**Examples:**

E1. A long-term quality control has enabled us to compute the estimator of the proportion p of nonconforming pieces as equal to 2% at the output of a production line. We would like to know the cumulative probability to have 200 pieces before the 3rd defective piece appears. With Microsoft® Excel® 14.0.6123 or later it comes using the negative binomial distribution:

`=NEGBINOM.DIST(200,3,0.02,1)=77.35%`

E2. To compare with the binomial distribution, we can ask ourselves what is the cumulative probability of drawing 198 non-defective parts from 201 using Microsoft® Excel® 14.0.6123 or later:

`=BINOM.DIST(198,201,0.98,1)=76.77%`

Therefore we see that the difference is small. In fact the difference between the two laws is in practice so small that we then use almost always the binomial law (but you should still be careful with this choice!).

As usual, we will now determine the variance and mean of this law. Let's start with the mean of having R successes when the E-th failure appears knowing that the probability of a failure is p. For this we will use a very simple and ingenious trick (all art was thinking about it...). If we return to our initial example:

$$\begin{matrix}[1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9] & 10\\ [\texttt{R} & \texttt{R} & \texttt{E} & \texttt{R} & \texttt{R} & \texttt{R} & \texttt{E} & \texttt{R} & \texttt{R}] & \texttt{E}\end{matrix} \tag{5.7.480}$$

and we rewrite this example as follows:

$$\begin{matrix}[1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10] \\ [\underbrace{\texttt{R} \;\; \texttt{R} \;\; \texttt{E}}_{=X_1} & \underbrace{\texttt{R} \;\; \texttt{R} \;\; \texttt{R} \;\; \texttt{E}}_{=X_2} & \underbrace{\texttt{R} \;\; \texttt{R} \;\; \texttt{E}}_{=X_3}]\end{matrix} \quad (5.7.481)$$

We then notice that the third success R of the first notation can be decomposed into the sum of three geometric random variables such that:

$$R = X_1 + X_2 + \ldots + X_n \quad (5.7.482)$$

With in the case of this particular example $n = 3$ corresponding in fact to $E = 3$. So quite generally the sum of n random geometric variables always gives a negative binomial distribution if the probability p is equal for each geometric variable!!! Anyway... as we have proved the expression of the mean and variance of the Geometric law as (thus giving us the mean rank of the first failure):

$$\mathrm{E}(X) = \frac{1}{p} \qquad \mathrm{V}(X) = \frac{1-p}{p^2} = \frac{q}{p^2} \quad (5.7.483)$$

since the random variables are independent and of same parameters then it comes for the negative binomial the mean of the rank of E-th failure using the property of the mean:

$$\mathrm{E}(X) = \mathrm{E}(X_1 + X_2 + \ldots + X_n) = \mathrm{E}(X_1) + \mathrm{E}(X_2) + \ldots \mathrm{E}(X_n) = n\mathrm{E}(X) = n\frac{1}{p} = E\frac{1}{p} \quad (5.7.484)$$

And therefore for the variance of the negative binomial distribution:

$$\mathrm{V}(X) = \mathrm{V}(X_1 + X_2 + \ldots + X_n) = \mathrm{V}(X_1) + \mathrm{V}(X_2) + \ldots \mathrm{V}(X_n) = n\mathrm{V}(X) = n\frac{1-p}{p^2} = E\frac{q}{p^2} \quad (5.7.485)$$

So the mean and variance of the rank (corresponding to the number of trials N or from another point of view: the mean number of successes by the simple subtraction $X - E$) to have the E-th failure is then to summarize:

$$\begin{aligned}\mathrm{E}(X) &= \frac{E}{p} \\ \mathrm{V}(X) &= E\frac{1-p}{p^2}\end{aligned} \quad (5.7.486)$$

Thus, putting $E = 1$, we fall back on the mean and variance of the geometric distribution!

Now, let Y be the random variable representing the number of trials before the R-th success. We then have the following expressions for the variance and the mean that are very common in the literature (these expressions of mean and variance corresponds to what we can find for the negative binomial law in Wikipedia for example):

$$\begin{aligned}\mathrm{E}(Y) &= \mathrm{E}(X-1) = \mathrm{E}(X) - \mathrm{E}(1) = \mathrm{E}(X) - 1 = \frac{E}{p} - 1 = \frac{E-p}{p} = \frac{\mathrm{E}(1-p)}{p} \\ \mathrm{V}(Y) &= \mathrm{V}(X-1) = \mathrm{V}(X) + \mathrm{V}(1) = \mathrm{V}(X) = \frac{\mathrm{E}(1-p)}{p^2}\end{aligned} \quad (5.7.487)$$

✍**Example:**

What is the expected number (mean) of trials we can expect before we fall on the third non-conforming part, knowing that the probability of a non-conforming part is 2%?

$$\text{E}(X)=\dfrac{E}{p}=\dfrac{3}{2\%}=150 \tag{5.7.488}$$

and for the standard deviation:

$$\sigma=\sqrt{E\dfrac{q}{p^2}}=\sqrt{3\dfrac{(1-2\%)}{(2\%)^2}}=85.732 \tag{5.7.489}$$

Like always the reader will find below a plot example of the distribution and cumulative distribution function for the negative binomial law of parameters $\text{NB}(N,k,p)=P(N,3,0.6)$ based on the example of the begging, but where the only difference is the probability of success where we he have taken 60% instead of 20%.

Thus, there is 21.6% of probability of having the third success after the third successive trial (i.e. 0 trials more than the number of successes), 25.92% of probability of having the third success after the fourth successive trial (i.e. one trial more than the number of successes), 20.7% of probability of having the third success after the fifth successive trial (i.e. two trial more than the number of successes) and so on...:

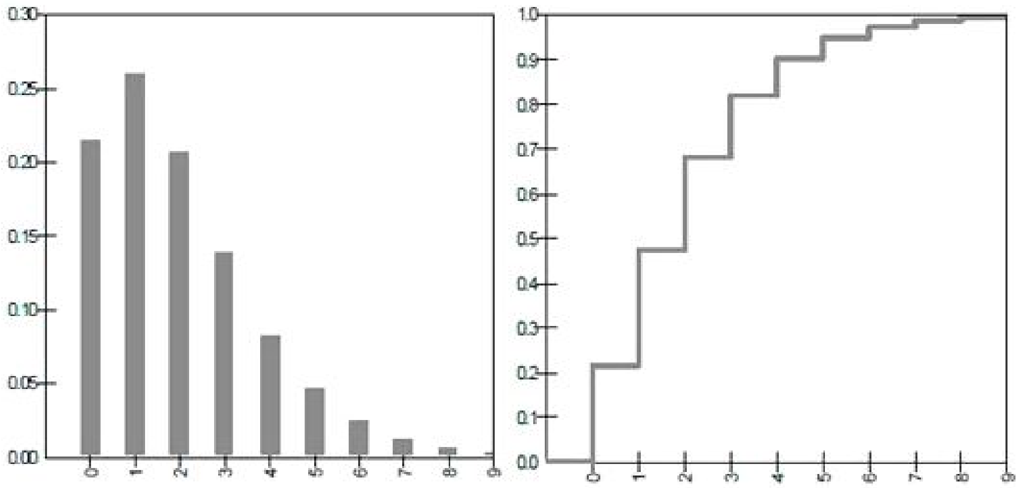


Figure 5.123: **Negative Binomial law `NB` (mass and cumulative distribution function)**

The above distributions are truncated to 9 (corresponding to 12 trials) but they theoretically continue indefinitely. What particularly distinguishes the binomial and geometric distributions from the negative binomial are the tails of the distribution.

The binomial negative distribution has an important place in a special regression technique that we will see in the section of Theoretical Computing at page 6316.

5.7.6.6 Hypergeometric Distribution

We consider to approach this function a simple example (but not very interesting in practice) that is this of an urn containing n balls where m are black and the other m' white (for several important examples used in the industry refer to the sections of Industrial Engineering page 8677). We take successively, and without replacement in the urn, p balls. The question is to find the probability that among the p balls, there is k that are black (in this statement the order of the drawing does not interest us).

We often talk about "exhaustive sampling" with the hypergeometric distribution because at the opposite of to the binomial distribution, the size of the lot which is the basis for the sampling will appear in the law.

> **Remark**
>
> This is also equivalent to a non-ordered sampling without replacement/repetition (see section Probabilities page 528) with constraint on the occurrences sometimes named "simultaneous sampling". We will often use the hypergeometric distribution in the field of quality and reliability where the black balls are associated to items with defects and the white one to items without defects.

The p balls can be chosen among n balls in C_p^n ways (thus representing the number of possible different outcomes) with as reminder (see section Probabilities page 528):

$$C_p^n = \binom{n}{p} = \frac{A_n^p}{p!} = \frac{n!}{p!(n-p)!} \tag{5.7.490}$$

The k black balls can be chosen among the m black in C_k^m ways. The $p-k$ white balls can be chosen in C_{p-k}^{n-m} ways. There is therefore $C_k^m C_{p-k}^{n-m}$ ways to have k black balls and $p-k$ white balls.

The searched probability is therefore given by (we will see an alternative notation in the section of Industrial Engineering page 8677):

$$\boxed{\mathcal{H}(n,p,m,k) = \frac{C_k^m C_{p-k}^{n-m}}{C_p^n} = \frac{\dfrac{m!}{k!(m-k)!}\dfrac{(n-m)!}{(p-k)!((n-m)-(p-k))!}}{\dfrac{n!}{p!(n-p)!}}} \tag{5.7.491}$$

and is said to follow a "Hypergeometric distribution" (or "Hypergeometric law") and can be obtained fortunately directly in Microsoft® Excel® 14.0.7153 with the function `HYPGEOM.DIST()`.

✍**Examples:**

E1. We want to develop a small computer program of $10,000$ lines of code (n). The return on experience shows that the probability of failure is one bug per $1,000$ lines of code (or 0.1% of $10,000$ lines) that corresponds to the value m.

We test about 50% of the functionality of the software randomly before sending it to the customer (corresponding to the equivalent of $5,000$ lines that is p). The probability of observing 5 bugs (k) is then given with Microsoft® Excel® 14.0.715:

`HYPGEOMDIST(k,p,m,n)=HYPGEOMDIST(5,5000,1%*10000,10000)=24.62%`

E2. In a small single production of a batch of $1,000$ pieces we know that 30% on average are bad because of the complexity of the pieces and by return on experience from a previous similar manufacturing. We know that a customer will randomly draw 20 pieces to decide whether to accept or reject the lot. He will not reject the lot if he finds zero defective pieces on the 20. What is the probability of having exactly 0 defective?

`=HYPGEOMDIST(0,20,300,1000)=0.073%`

and as we require a null draw drawing result, the calculation of the hypergeometric distribution simplifies manually to:

$$\frac{700}{1000}\frac{699}{999}\frac{698}{998}\frac{697}{997}\ldots\frac{681}{981}=0.073\% \tag{5.7.492}$$

It is not forbidden to make direct calculation of the mean and variance of the hypergeometric distribution, but the reader will without much trouble guess that this calculation will be ... relatively indigestible. Then we can use an indirect method that is much more interesting!

First, the reader will perhaps, even certainly, have noticed that experienced of the hypergeometric distribution is a series of Bernoulli trials (without replacement of course!).

So we will cheat by using initially the property of linearity of the mean. We define for this purpose a new variable corresponding implicitly in fact to the experience of the hypergeometric distribution (a sequence of k Bernoulli trials!):

$$X=\sum_{i=1}^{k}X_i \tag{5.7.493}$$

where X_i is the success of obtaining at the i-th drawing a black ball (either 0 or 1). But, we know that for all i the random variable X_i follows a Bernoulli function for which we have proved in our study of the Bernoulli distribution that $\text{E}(X_i)=p$.

Therefore, by the property of linearity of the mean we have (caution! here p is not the number of balls, but the probability associated with an expected event!):

$$\text{E}(X)=\text{E}\left(\sum_{i=1}^{k}X_i\right)=kp \tag{5.7.494}$$

In the Bernoulli trial, p is the probability of obtaining the desired item or event (for reminder...). In the hypergeometric distribution what interests us is the probability of a black ball (which are in quantity m, therefore with m' white balls) compared to the total amount of n balls. And the ratio obviously gives us this probability. Thus, we have:

$$\mu=\text{E}(X)=kp=k\frac{m}{m+m'}=k\frac{m}{n} \tag{5.7.495}$$

where k is the number of trials (do not confuse with the notation of the initial statement where it was by the variable p!). The mean gives then the average number of black balls in a drawing of k balls among n, where m are known as being black. The reader will have noticed that the mean of the hypergeometric distribution is the same as the binomial distribution!

To determine the variance, we use the variance of the Bernoulli distribution and the following relation proved during the introduction of the mean and covariance at the beginning of this chapter:

$$\text{V}(X+Y)=\text{V}(X)+\text{V}(Y)+2\text{cov}(X,Y)+\text{V}(X)+\text{V}(Y)+2(\text{E}(XY)-\text{E}(X)\text{E}(Y)) \tag{5.7.496}$$

Recalling that we have $X=\sum_{i=1}^{k} X_i$, we get:

$$\text{V}(X)=\text{V}\left(\sum_{i=1}^{k} X_i\right)=\sum_{i=1}^{k}\text{V}(X_i)+2\sum_{i=1\leq 1<j\leq n}^{k}(\text{E}(X_iX_j)-\text{E}(X_i)\text{E}(X_j)) \tag{5.7.497}$$

However, for the Bernoulli law, we have:

$$\text{V}(X)=pq=\frac{m}{m+m'}\frac{m'}{m+m'}=\frac{mm'}{(m+m')^2}=\frac{mm'}{n^2} \tag{5.7.498}$$

Then we first already get:

$$\text{V}\left(\sum_{i=1}^{k} X_i\right)=k\frac{mm'}{(m+m')^2}=k\frac{mm'}{n^2} \tag{5.7.499}$$

Afterwards, we have easily:

$$\text{E}(X_i)\text{E}(X_j)=\frac{m}{m'+m}\frac{m}{m'+m}=\left(\frac{m}{m'+m}\right)^2 \tag{5.7.500}$$

The calculation of $\text{E}(X_iX_j)$ requires a good understanding of probabilities (this will be a good refresh!).

The mean $\text{E}(X_iX_j)$ is given (implicitly), as we know, by the weighted sum of the probabilities that two events occur at the same time. However, our events are binary: either it is a black ball (1) or it is a white ball (0). So all terms of the sum without two consecutive black balls consecutively will be null!

The problem is then to calculate the probability of having two consecutive black balls and it is thus written:

$$\begin{aligned}P(X_iX_j=1)&=P((X_i=1)\cap(X_j=1))=P(X_i=1)P_{X_i=1}P(X_j=1)\\&=\frac{m}{m'+m}\frac{m-1}{(m'+m)-1}\end{aligned} \tag{5.7.501}$$

So we finally have:

$$\text{E}(X_iX_j)=\frac{m}{m'+m}\frac{m-1}{(m'+m)-1} \tag{5.7.502}$$

Therefore:

$$\begin{aligned}\text{cov}(X_i,X_j)&=\text{E}(X_iX_j)-\text{E}(X_i)(X_j)=\frac{m}{m'+m}\frac{m-1}{(m'+m)-1}-\left(\frac{m}{m'+m}\right)^2\\&=\frac{-mm'}{(m'+m)^2(m'+m-1)}\end{aligned}\tag{5.7.503}$$

Finally (using the result of Gauss series seen in the section of Sequences and Series page 1666):

$$\begin{aligned}\text{V}(X)&=\sum_{i=1}^{k}\text{V}(X_i)+2\sum_{i=1\leq 1<j\leq n}^{k}(\text{E}(X_iX_j)-\text{E}(X_i)\text{E}(X_j))\\&=k\frac{mm'}{(m'+m)^2}+2\frac{-mm'}{(m'+m)^2(m'+m-1)}\sum_{i=1\leq 1<j\leq n}^{k}1\\&=k\frac{mm'}{(m'+m)^2}+2\frac{-mm'}{(m'+m)^2(m'+m-1)}\frac{(k-1)k}{2}\\&=\frac{kmm'(m'+m-1)-mm'k(k-1)}{(m'+m)^2(m'+m-1)}=\frac{kmm'(m'+m-1-k+1)}{(m'+m)^2(m'+m-1)}\\&=\frac{kmm'(m'+m-k)}{(m'+m)^2(m'+m-1)}=k\frac{m}{m'+m}\frac{m'}{m'+m}\frac{m'+m-k}{m'+m-1}\\&=kpq\frac{m'+m-k}{m'+m-1}=kpq\frac{n-k}{n-1}\end{aligned}\tag{5.7.504}$$

where we have used the fact that:

$$\sum_{i<j}\text{cov}(X_i,X_j)\tag{5.7.505}$$

is composed of:

$$C_2^k=\binom{k}{2}=\frac{k(k-1)}{2}\tag{5.7.506}$$

terms that correspond to the number of ways there are to choose a pair (i,j) with $i<j$.

Because:

$$\text{V}(X)=kpq\frac{n-k}{n-1}=k\frac{m}{n}\frac{n-m}{n}\frac{n-k}{n-1}\tag{5.7.507}$$

We can write:

$$\sigma=\sqrt{kpq}\sqrt{\frac{n-k}{n-1}}=\sqrt{k\frac{m}{n}\frac{n-m}{n}}\sqrt{\frac{n-k}{n-1}}\tag{5.7.508}$$

In the specialized literature, we often find the variance written in the following way by noting the expected event r and the non-expected event s:

$$\text{V}(X)=kpq\frac{n-k}{n-1}=k\frac{r}{n}\frac{s}{n}\frac{n-k}{n-1}=\frac{krs(n-k)}{n^2(n-1)}=\frac{klrs}{n^2(n-1)}\tag{5.7.509}$$

so with $l=n-k$. This last notation will be very useful further below during our study of the Mantel-Haenszel test (page 1283). Furthermore, we see that in:

$$\sigma=\sqrt{kpq}\sqrt{\frac{n-k}{n-1}} \tag{5.7.510}$$

there is the standard deviation as the binomial distribution, at the difference of a factor that is noted:

$$\text{fpc}=\sqrt{\frac{n-k}{n-1}} \tag{5.7.511}$$

that we found often in statistics and is named "finite population correction factor".

Here is an example plot example of the distribution function and cumulative distribution for the Hypergeometric function of parameters $(n,p,m,k)=(10,6,5,k)$:

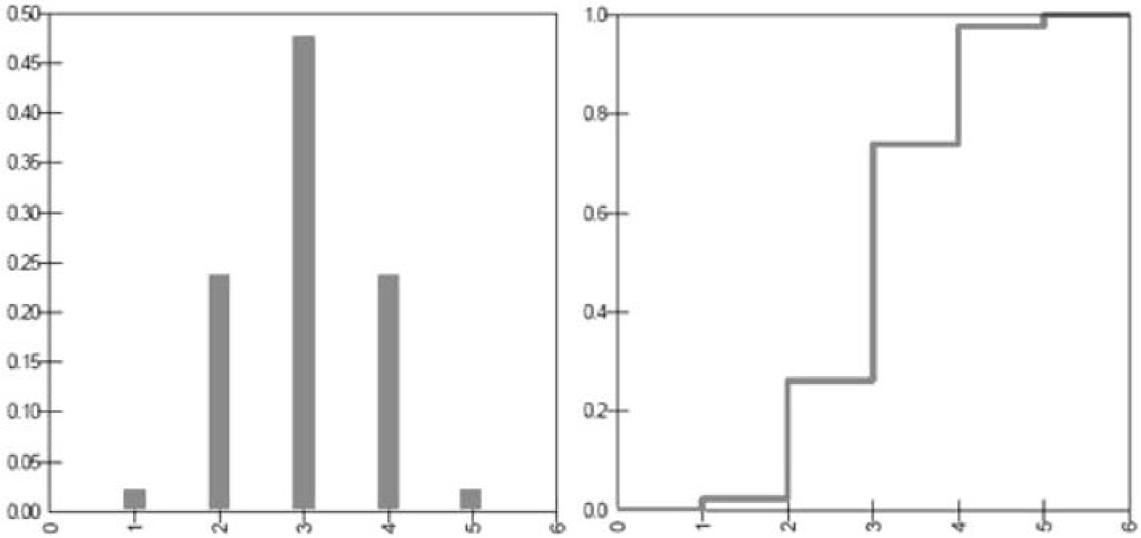


Figure 5.124: **Hypergeometric law $\mathcal{H}$ (mass and cumulative distribution function)**

We will prove now that the Hypergeometric distribution tends to a binomial distribution since this property is used many times in different sections of this textbook (especially the section of Industrial Engineering page 8677).

To do this, we decompose:

$$\mathcal{H}(n,p,m,k)=\frac{C_k^m C_{p-k}^{n-m}}{C_p^n} \tag{5.7.512}$$

We then get:

$$\begin{aligned}\frac{C_k^m C_{p-k}^{n-m}}{C_p^n}&=\frac{\dfrac{m!}{k!(m-k)!}\dfrac{(n-m)!}{(p-k)!((n-m)-(p-k))!}}{\dfrac{n!}{p!(n-p)!}}\\&=\frac{m!}{k!(m-k)!}\frac{(n-m)!}{(p-k)!((n-m)-(p-k))!}\frac{p!(n-p)!}{n!}\\&=\frac{p!}{k!(p-k)!}\frac{m}{(m-k)!}\frac{(n-m)!}{(p-k)!((n-m)-(p-k))!}\frac{(n-p)!}{n!}\\&=C_p^k\frac{m}{(m-k)!}\frac{(n-m)!}{(p-k)!((n-m)-(p-k))!}\frac{(n-p)!}{n!}\end{aligned}\tag{5.7.513}$$

For the second term:

$$\frac{m}{(m-k)!}=\frac{1\cdot 2\cdot 3\cdot\ldots\cdot m}{1\cdot 2\cdot 3\cdot\ldots\cdot(m-k)}=m(m-1)\ldots(m-k+1)\tag{5.7.514}$$

For $m\rightarrow+\infty$ (...) all the terms are of then of the order of m. Then we have:

$$m(m-1)\ldots(m-k+1)\cong m^k\tag{5.7.515}$$

For the third term, a identical development to the previous one provides (for sure we need that also $n\rightarrow+\infty$ (...)):

$$\frac{(n-m)!}{(p-k)!((n-m)-(p-k))!}\cong(n-m)^{p-k}\tag{5.7.516}$$

And for sure... we can debate therefore about $n-m$ when both terms tends to infinity...

Ditto for the fourth term:

$$\frac{(n-p)!}{n!}\cong n^{-p}\tag{5.7.517}$$

In conclusion we have:

$$\frac{C_k^m C_{p-k}^{n-m}}{C_p^n}\underset{n,m\rightarrow+\infty}{=}C_p^k\frac{m^k(n-m)^{p-k}}{n^p}\tag{5.7.518}$$

We change the notation by writing p (the number of individuals drawn) as being N. We get then:

$$\frac{C_k^m C_{N-k}^{n-m}}{C_N^n}\underset{n,m\rightarrow+\infty}{=}C_k^N\frac{m^k(n-m)^{N-k}}{n^N}\tag{5.7.519}$$

We make another change of notation by writing b the black balls and w the white balls. We get then:

$$\frac{C_k^b C_{N-k}^{n-b}}{C_N^n}\underset{n,b\rightarrow+\infty}{=}C_k^N\frac{b^k(n-b)^{N-k}}{n^N}=C_k^N\frac{b^k w^{N-k}}{n^N}\tag{5.7.520}$$

Finally, we note p the proportion of black balls and q that of white balls in the lot. We then get:

$$\frac{C_k^{np}C_{N-k}^{n-np}}{C_N^n} \underset{np\to+\infty}{=} C_k^N\frac{(np)^k(nq)^{N-k}}{n^N}=C_k^N(np)^k(nq)^{N-k}n^{-N}=C_k^Np^kq^{N-k} \tag{5.7.521}$$

We find out the binomial distribution! In practice, it is common to approximate the hypergeometric distribution with a binomial distribution when the ratio of the number of individuals from the total number of individuals is less than 10% (that is to say when the sample is 10 times smaller than the whole population). It follows that the hypergeometric distribution tends also (as we will show later) to a Normal distribution when the population tends to infinity and that the sample is small.

In practice, Monte Carlo simulations with testing adjustments (see later in this chapter) have shown that the hypergeometric distribution could be approximated by a Normal distribution (very important case in contingency statistical tests that we will study further below in this section) if the following three conditions are met simultaneously:

$$k>9\frac{n-m}{m}\qquad k>9\frac{m}{n-m}\qquad k<\frac{n}{10} \tag{5.7.522}$$

Thus graphically and approximately...:

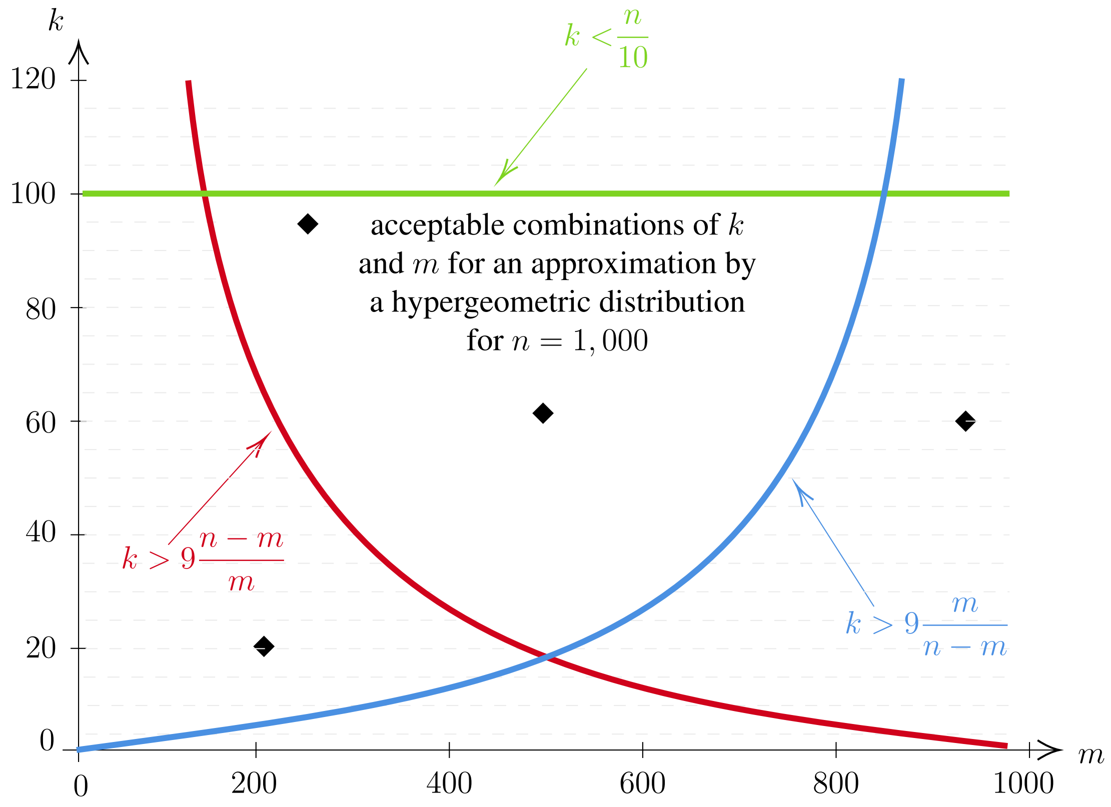


Figure 5.125: Conditions of application of the approximation by a Normal distribution

Thus:

$$\mathcal{H}(k,m,n)\cong\mathcal{N}(\mu,\sigma)=\mathcal{N}\left(k\frac{m}{n},\sqrt{\frac{krs(n-k)}{n^2(n-1)}}\right)=\mathcal{N}\left(k\frac{m}{n},\sqrt{\frac{kmm'(n-k)}{n^2(n-1)}}\right) \tag{5.7.523}$$

5.7.6.7 Multinomial Distribution

The multinomial distribution (so named because it involves several times the binomial coefficient) is a law applicable to n distinguishable events, each with a given probability, which occur one or more times and it is not necessarily ordered. This is a frequent case in marketing research and that will be useful to build the statistical McNemar test that we will study much later (see further below page 1264) and also the Multinomial logistic regression (see section Numerical Methods page 6477).

The multinomial law is a generalization of the binomial law applicable for example to n trials of a six-sided die but with the difference that probabilities are not equal!

More technically, consider the space of events $\Omega = \{1, 2, \ldots, m\}$ with the probabilities $P(\{i\}) = p_i, i = 1, 2, \ldots, n$. We take n times a given element of Ω with replacement with the probability $p_i, i = 1, 2, \ldots, n$. We will search what is the probability of getting in a non-necessarily ordered way the event 1, k_1 times, event 2, k_2 times and this on a sequence of n drawings.

Remark

This is equivalent to the study of a sampling with replacement AND constraints on the occurrences! So without constraints we will see with an example that we fall back on a sampling with simple replacement.

We know that (see section Probabilities page 528) if we take a set of events with multiple outcomes, then the different combinations of sequences we can get taking p selected elements among n is given by:

$$C_p^n = \frac{n!}{p!(n-p)!} \tag{5.7.524}$$

We have therefore:

$$C_{k_1}^n = \frac{n!}{k_1!(n-k_1)!} \tag{5.7.525}$$

different ways to get k_1 times a given event. Thus an associated probability of (it's the binomial distribution):

$$P(n, k_1) = C_{k_1}^n p_1^{k_1} q_1^{n-k_1} = C_{k_1}^n p_1^{k_1} (1-p_1)^{n-k_1} \tag{5.7.526}$$

Now comes the particularity of the multinomial distribution!: there are no failures in contrast to the binomial distribution. Each "pseudo-failure" can be considered as a subset draw of k_2 items from the $n - k_1$ remaining elements.

Thus the term:

$$C_{k_1}^n p_1^{k_1} q_1^{n-k_1} \tag{5.7.527}$$

will be written on the whole experience if we consider a particular case limited to two types of events:

$$C_{k_1}^n p_1^{k_1} q_1^{n-k_1} := C_{k_1}^n p_1^{k_1} (C_{k_2}^{n-k_1} p_2^{k_2}) \tag{5.7.528}$$

so with:

$$C_{k_2}^{n-k_1}=\frac{(n-k_1)!}{k_2!((n-k_1)-k_2)!} \tag{5.7.529}$$

which gives us the number of different times to get k_2 times a second event because in the whole sequence of n elements k_1 of them have already been taken so have now only $n-k_1$ remaining on which we can get the k_2 desired.

These relations then show us that this is a situation where each event probability is considered as a binomial (hence its name ...).

So we have in the case of two sets of t-uples:

$$\begin{aligned}C_{k_1}^{n}p_1^{k_1}C_{k_2}^{n-k_1}p_2^{k_2}&=C_{k_1}^{n}C_{k_2}^{n-k_1}p_1^{k_1}p_2^{k_2}=\frac{n!}{k_1!(n-k_1)!}\frac{(n-k_1)!}{k_2!((n-k_1)-k_2)!}p_1^{k_1}p_2^{k_2}\\&=\frac{n!}{k_1!k_2!}\frac{1}{n-k_1-k_2}p_1^{k_1}p_2^{k_2}\end{aligned} \tag{5.7.530}$$

and because:

$$-k_1-k_2=-n \tag{5.7.531}$$

we get:

$$P=\frac{n!}{k_1!k_2!}\frac{1}{0!}p_1^{k_1}p_2^{k_2}=\frac{n!}{k_1!k_2!}p_1^{k_1}p_2^{k_2} \tag{5.7.532}$$

and we see that the construction of this distribution therefore requires that:

$$\sum_i p_i=1 \qquad \sum_i k_i=n \tag{5.7.533}$$

Thus, by induction we have the probability $\mathcal{M}$ we were looking for and named the "Multinomial function" and using the previous relation given by:

$$\boxed{\mathcal{M}(\vec{k},\vec{p},m)=\frac{n!}{\prod\limits_{i=1}^{m}k_i!}\prod_{i=1}^{m}p_i^{k_i}=\frac{\left(\sum\limits_{i=1}^{m}k_i\right)!}{\prod\limits_{i=1}^{m}k_i!}\prod_{i=1}^{m}p_i^{k_i}} \tag{5.7.534}$$

In other words if there are m random variables, i.e. X_i, $i\in[1,n]$, where X_i represents the number of occurrences of item i in a choice of n items, with entry i in the vector of probabilities $\vec{p}$, p_i giving the probability of drawing item i. The probability of selecting k_i of item $k_1\ldots k_m$ of item m is then given by $\mathcal{M}(\vec{k},\vec{p},m)$ above (if this explanation is not clear, see the example further below!).

Remark

Note that if $m=2$ then the above relation becomes:

$$\mathcal{M}=\frac{(k_1+k_2)!}{k_1!k_2!}p_1^{k_1}p_2^{k_2} \tag{5.7.535}$$

But as $k_1+k_2=n$ then:

$$\mathcal{M}=\frac{n!}{k_1!(n-k_1)!}p_1^{k_1}p_2^{n-k_1} \tag{5.7.536}$$

and as $p_1+p_2=1$ we finally have:

$$\mathcal{M}=\frac{n!}{k_1!(n-k_1)!}p_1^{k_1}(1-p_1)^{n-k_1} \tag{5.7.537}$$

We recognize here the binomial distribution!

In a spreadsheet software Microsoft® Excel® 11.8346, the term:

$$\frac{\left(\sum\limits_{i=1}^{m} k_i\right)!}{\prod\limits_{i=1}^{m} k_i!} \tag{5.7.538}$$

is named "multinomial coefficient" and is available under the name of the function `MULTINOMIAL( )`. In the literature we also find this term sometimes under the following respective notations:

$$\binom{k_1+k_2+\ldots+k_m}{k_1,k_2,\ldots,k_m} \qquad \binom{n}{k_1,k_2,\ldots,k_m} \tag{5.7.539}$$

Which gives then for the expression of the multinomial distribution:

$$\mathcal{M}(\vec{k},\vec{p},m)=\binom{n}{k_1,k_2,\ldots,k_m}\prod_{i=1}^{m}p_i^{k_i} \tag{5.7.540}$$

Theorem 5.47. We will show now that the multinomial distribution is effectively a probability distribution (because we could doubt ...). If this is the case, as we know it, the sum of the probabilities must be equal to 1.

Proof 5.47.1. Let us recall that we proved the binomial theorem (see section Calculus page 1364):

$$(x+y)^n=\sum_{k=0}^{n}C_k^n x^k y^{n-k} \tag{5.7.541}$$

Now do a little bit of notation:

$$(x_1+x_2)^n=\sum_{k_1=0}^{n}C_k^n x_1^{k_1}x_2^{n-k_1}=\sum_{k_1=0}^{n}\frac{n!}{k_1!(n-k_1)!}x_1^{k_1}x_2^{n-k_1} \tag{5.7.542}$$

and this time a change of variables:

$$(x_1+x_2)^n=\sum_{k_1=0}^{n}C_k^n x_1^{k_1}x_2^{n-k_1}=\sum_{k_1=0}^{n}\frac{n!}{k_1!k_2!}x_1^{k_1}x_2^{k_2} \tag{5.7.543}$$

This last relation (which is a special case with two terms of the "multinomial theorem") will be useful to us to show that the multinomial distribution is effectively a probability distribution. We also take the special case with two groups of drawing:

$$\mathcal{M}(\vec{k},\vec{k},m)=\frac{n!}{\prod\limits_{i=1}^{m}k_i!}\prod_{i=1}^{m}p_i^{k_i}=\frac{n!}{k_1!k_2!}p_1^{k_1}p_2^{k_2} \tag{5.7.544}$$

which can is also written by the construction of the multinomial distribution:

$$\mathcal{M}=\frac{n!}{k_1!(n-k_1)!}p_1^{k_1}p_2^{n-k_1} \tag{5.7.545}$$

and therefore, the sum must be equal to the unit such that:

$$\sum_{k_1=0}^{n}\frac{n!}{k_1!(n-k_1)!}p_1^{k_1}p_2^{n-k_1}=1 \tag{5.7.546}$$

To check this we use the multinomial theorem shown above:

$$(p_1+p_2)^n=\sum_{k_1=0}^{n}\frac{n!}{k_1!k_2!}p_1^{k_1}p_2^{k_2} \tag{5.7.547}$$

However, by construction of the multinomial sum of probabilities is unitary, we have effectively:

$$(p_1+p_2)^n=(1)^n=1=\sum_{k_1=0}^{n}\frac{n!}{k_1!(n-k_1)!}p_1^{k_1}p_2^{n-k_1} \tag{5.7.548}$$

■ Q.E.D.

✍**Examples:**

E1. We launch an unbiased die 12 times. The probability that all 6 faces appear the same number of times (not necessarily consecutively!) is given by:

$$\mathcal{M}=\frac{n!}{\prod\limits_{i=1}^{m}k_i!}\prod_{i=1}^{m}p_i^{k_i}=\frac{12!}{\prod\limits_{i=1}^{6}2!}\prod_{i=1}^{6}\left(\frac{1}{6}\right)^2=\frac{12!}{2^6 6^{12}}=0.34\% \tag{5.7.549}$$

where we see well that m is the number of successes.

E2. We launch an unbiased die 12 times. The probability that a single unique face appears

12 times (hence the "1" appears 12 times, or the "2" or the "3", etc.) is given by:

$$\mathcal{M}=\frac{n!}{\prod\limits_{i=1}^{m}k_i!}\prod_{i=1}^{m}p_i^{k_i}=\frac{12!}{\prod\limits_{i=1}^{1}12!}\prod_{i=1}^{1}\left(\frac{1}{6}\right)^{12}=\frac{12!}{6^{12}} \tag{5.7.550}$$

So we end up with this last example known a being a binomial distribution result.

Let us recall that we have proved just earlier above (we just changes the notation a bit):

$$\mathcal{M}(\vec{x},\vec{p},r)=\frac{n!}{\prod\limits_{i=1}^{r}x_i!}\prod_{i=1}^{r}p_i^{x_i}=\frac{n!}{x_1!x_2!\ldots x_r!}p_1^{x_1}p_2^{x_2}\ldots p_r^{x_r} \tag{5.7.551}$$

And then we have obviously:

$$\sum_{\substack{x_1,x_2,\ldots,x_r\\ \sum_{i=1}^{r}x_i=n}}\frac{n!}{x_1!x_2!\ldots x_r!}p_1^{x_1}p_2^{x_2}\ldots p_r^{x_r}=1 \tag{5.7.552}$$

The mean of the multinomial distribution is in fact a mean of vector. But as the mean of a vector is equal to the mean of its components we want to calculate $\text{E}(x_i)$. For this, without loss of generality, let us calculate $\text{E}(x_1)$. To do this let us recall that by definition of the mean:

$$\text{E}(x_1)=\sum_{x_1=0}^{n}x_1P(X_1=x_1) \tag{5.7.553}$$

But for $x_1=0$ the corresponding terms is equal to zero. Then we can directly write:

$$\text{E}(x_1)=\sum_{x_1=1}^{n}x_1P(X_1=x_1) \tag{5.7.554}$$

Now as the sum already acts on the term $P(X_1=x_1)$, for that latter we have only to sum on all terms excepted x_1 such that:

$$\begin{aligned}\text{E}(x_1)&=\sum_{x_1=1}^{n}x_1\sum_{\substack{x_2,\ldots,x_r\\ \sum_{i=2}^{r}x_i=n-x_1}}\frac{n!}{x_1!x_2!\ldots x_r!}p_1^{x_1}p_2^{x_2}\ldots p_r^{x_r}\\&=\sum_{x_1=1}^{n}\sum_{\substack{x_2,\ldots,x_r\\ \sum_{i=2}^{r}x_i=n-x_1}}\frac{n!}{(x_1-1)!x_2!\ldots x_r!}p_1p_1^{x_1-1}p_2^{x_2}\ldots p_r^{x_r}\end{aligned} \tag{5.7.555}$$

Now we make the following change of variables $x_1'=x_1-1$. Then:

$$\begin{aligned}\mathrm{E}(x_1)&=\sum_{x_1'=0}^{n-1}\sum_{\substack{x_2,\ldots,x_r\\ \sum_{i=2}^{r}x_i=n-x_1'+1}}\frac{n!}{x_1'!x_2!\ldots x_r!}p_1p_1^{x_1'}p_2^{x_2}\ldots p_r^{x_r}\\&=\sum_{x_1'=0}^{n-1}\sum_{\substack{x_2,\ldots,x_r\\ x_1'+x_2+\ldots+x_r=n-1}}\frac{n!}{x_1'!x_2!\ldots x_r!}p_1p_1^{x_1'}p_2^{x_2}\ldots p_r^{x_r}\\&=np_1\sum_{x_1'=0}^{n-1}\sum_{\substack{x_2,\ldots,x_r\\ x_1'+x_2+\ldots+x_r=n-1}}\frac{(n-1)!}{x_1'!x_2!\ldots x_r!}p_1^{x_1'}p_2^{x_2}\ldots p_r^{x_r}\end{aligned}\tag{5.7.556}$$

But the second summation is knows to us! Indeed, it's the cumulated probability on all $n-1$ (instead of the classical n). And that one is equal to 1. Therefore we have:

$$\begin{aligned}\mathrm{E}(x_1)&=np_1\sum_{x_1'=0}^{n-1}\underbrace{\sum_{\substack{x_2,\ldots,x_r\\ x_1'+x_2+\ldots+x_r=n-1}}\frac{(n-1)!}{x_1'!x_2!\ldots x_r!}p_1^{x_1'}p_2^{x_2}\ldots p_r^{x_r}}_{=1}\\&=np_1\sum_{x_1'=0}^{n-1}1=np_i\end{aligned}\tag{5.7.557}$$

Therefore:

$$\mathrm{E}(\vec{x})=n\vec{p}\tag{5.7.558}$$

Now let us see the variance (it's also a vector of variance in fact!). The idea is quite the same as for the mean. First we will use the Huygens relation:

$$\mathrm{V}(x_i)=\mathrm{E}(x_i^2)-\mathrm{E}(x_i)^2=\mathrm{E}(x_i^2)-(np_i)^2\tag{5.7.559}$$

So only the first term $\mathrm{E}(x_i^2)$ is unknown to us. Without loss of generality, and again by taking the definition of the expected mean, let us consider:

$$\begin{aligned}\mathrm{E}(x_1^2)&=\sum_{x_1=0}^{n}x_1^2P(X_1=x_1)=\sum_{x_1=1}^{n}x_1^2P(X_1=x_1)\\&=\sum_{x_1=1}^{n}x_1^2\sum_{\substack{x_2,\ldots,x_r\\ \sum_{i=2}^{r}x_i=n-x_1}}\frac{n!}{x_1!x_2!\ldots x_r!}p_1^{x_1}p_2^{x_2}\ldots p_r^{x_r}\\&=\sum_{x_1=1}^{n}x_1\sum_{\substack{x_2,\ldots,x_r\\ \sum_{i=2}^{r}x_i=n-x_1}}\frac{n!}{(x_1-1)!x_2!\ldots x_r!}p_1^{x_1}p_2^{x_2}\ldots p_r^{x_r}\end{aligned}\tag{5.7.560}$$

Let us do now the change of variable:

$$x_1'=x_1-1\Leftrightarrow x_1=x_1'+1\tag{5.7.561}$$

Therefore:

$$
\begin{aligned}
\mathrm{E}(x_1^2) &= \sum_{x_1'=0}^{n-1}(x_1'+1)\sum_{\substack{x_2,\ldots,x_r\\ \sum_{i=2}^r x_i=n-x_1'-1}}\frac{n!}{x_1'!x_2!\ldots x_r!}p_1^{x_1'+1}p_2^{x_2}\ldots p_r^{x_r}\\
&=\sum_{x_1'=0}^{n-1}x_1'\sum_{\substack{x_2,\ldots,x_r\\ \sum_{i=2}^r x_i=n-x_1'-1}}\frac{n!}{x_1'!x_2!\ldots x_r!}p_1^{x_1'+1}p_2^{x_2}\ldots p_r^{x_r}+\sum_{x_1'=0}^{n-1}\sum_{\substack{x_2,\ldots,x_r\\ \sum_{i=2}^r x_i=n-x_1'-1}}\frac{n!}{x_1'!x_2!\ldots x_r!}p_1^{x_1'+1}p_2^{x_2}\ldots p_r^{x_r}\\
&=np_1\sum_{\substack{x_1',\ldots,x_r\\ x_1'+\ldots+x_r=n-1}}\frac{(n-1)!}{x_1'!x_2!\ldots x_r!}p_1^{x_1'}p_2^{x_2}\ldots p_r^{x_r}+np_1\sum_{\substack{x_1',\ldots,x_r\\ x_1'+\ldots+x_r=n-1=n-1}}\frac{(n-1)!}{x_1'!x_2!\ldots x_r!}p_1^{x_1'}p_2^{x_2}\ldots p_r^{x_r}\\
&=np_1\begin{pmatrix}\text{expected valued of}\\ \text{multinomial with } n-1\end{pmatrix}+np_1\underbrace{\begin{pmatrix}\text{cumulated probability}\\ \text{of a multinomial } n-1\end{pmatrix}}_{=1}\\
&=np_1\left((n-1)p_1\right)+np_1
\end{aligned}
\tag{5.7.562}
$$

So finally:

$$
\mathrm{V}(x_i)=\mathrm{E}(x_i^2)-\mathrm{E}(x_i)^2=\left[np_1\left((n-1)p_1\right)+np_1\right]-(np_i)^2=np_1(1-p_1)
\tag{5.7.563}
$$

5.7.6.8 Poisson Distribution

For some rare events, the probability p is very small and tends to zero. However, the average value np tends to a fixed value as n tends to infinity.

We start from a binomial distribution with mean $\mu = np$ that we will assume finite when n tends to infinity.

The probability of k successes after n trials is (Binomial distribution):

$$\mathcal{B}(n,k) = C_k^n p^k (1-p)^{n-k} = C_k^n p^k q^{n-k} = \binom{n}{k} p^k q^{n-k} \tag{5.7.564}$$

By writing $p = \dfrac{m}{n}$ (where m will be temporarily the new notation for the mean according to $\mu = np$), this expression can be rewritten as:

$$\mathcal{B}(n,k) = \binom{n}{k} p^k q^{n-k} = \frac{n!}{k!(n-k)!}\left(\frac{n}{m}\right)^k \left(1-\frac{m}{n}\right)^{n-k} \tag{5.7.565}$$

By grouping the terms, we can put the value $\mathcal{B}(n,k)$ under the form:

$$\mathcal{B}(n,k) = \frac{m^k}{k!}\left(1-\frac{m}{n}\right)^n \frac{n!}{(n-k)!n^k\left(1-\dfrac{m}{n}\right)^k} \tag{5.7.566}$$

We recognize that when n tends to infinity, the second factor of the product has for limit $e^{-\mu}$ (see Functional Analysis page 2289). The third factor, since we are interested to the small values of k (the probability of success is very small), its limit for n tending to infinity is equal to 1.

This technique of passing to the limit is sometimes named in this context: the "Poisson limit theorem".

So we get the "Poisson distribution" (or "Poisson law"), also sometimes named the "law of rare events" therefore given by:

$$\boxed{\mathcal{P}(\mu,k) = \frac{\mu^k}{k!}e^{-\mu}} \tag{5.7.567}$$

which can be obtained in Microsoft® Excel® 11.8346 with the function `POISSON( )` and in practice and the specialized literature is often indicated by the letter u.

It is indeed a probability distribution since using the Taylor series (see section Sequences and Series page 1703), we show that the sum of the cumulative probabilities is:

$$\sum_{k=0}^{+\infty}\frac{\mu^k}{k!}e^{-\mu} = e^{-\mu}\sum_{k=0}^{+\infty}\frac{\mu^k}{k!} = e^{-\mu}e^{\mu} = 1 \tag{5.7.568}$$

Remark

We will frequently encounter this distribution in different sections of the book such as in the study of preventive maintenance in the section of Industrial Engineering page 8716 or in the section of Quantitative Management for the study of queuing theory page 8141 (the reader can refer to them for interesting and pragmatic examples) and finally in the field of life and non-life insurance.

Here is a plot example of the Poisson distribution and cumulative distribution function with parameter $\mu = 3$:

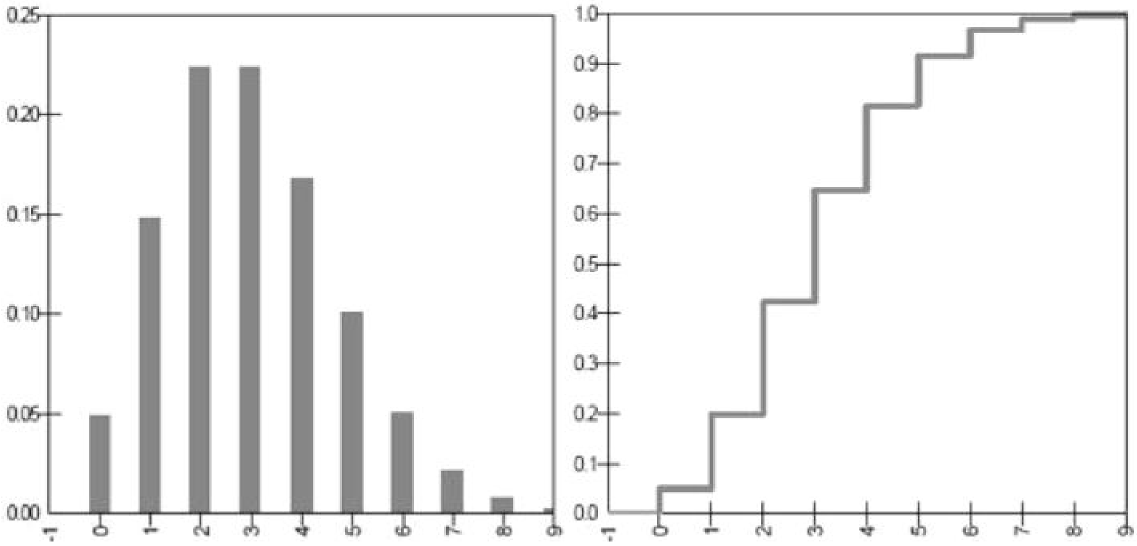


Figure 5.126: Poisson law $\mathcal{P}$ (mass and cumulative distribution function)

This distribution is important because it describes many processes whose probability is small and constant. It is often used in the "queuing theory" (waiting time), acceptability and reliability test and statistical quality control. Among other things, it applies to processes such as the emission of light quanta by excited atoms, the number of red blood cells seen under the microscope, the number of incoming calls to a call center. The Poisson distribution is valid for many observations in nuclear and particle physics.

The mean (average) of the Poisson distributions is (we use the Taylor series of the exponential):

$$\mu = \mathrm{E}(k) = \sum_{k=0}^{+\infty} k\mathcal{P}(\mu, k) = \sum_{k=0}^{+\infty} k\frac{\mu^k}{k!}e^{-\mu} = e^{-\mu}\mu\sum_{k=1}^{+\infty}\frac{\mu^{k-1}}{(k-1)!} = e^{-\mu}\mu e^{\mu} = \mu \tag{5.7.569}$$

and gives the average number of times that you get the desired outcome.

This result may seem confusing .... the mean is expressed by the mean?? Yes must simply not forget that it is given since the beginning by:

$$\mu = np \tag{5.7.570}$$

Remark

For more details the reader may also refer to the subsection on "estimators" page 818.

The variance of the Poisson distribution function is itself given by (again we use the Taylor series):

$$\begin{aligned}
\sigma^2=\mathrm{V}(k)&=\sum_{k=0}^{+\infty}[k-\mu^2]P_k=\sum_{k=0}^{+\infty}[k-\mu^2]\frac{\mu^k}{k!}e^{-\mu}\\
&=\sum_{k=0}^{+\infty}k^2\frac{\mu^k}{k!}e^{-\mu}-\sum_{k=0}^{+\infty}2k\mu\frac{\mu^k}{k!}e^{-\mu}+\sum_{k=0}^{+\infty}\mu^2\frac{\mu^k}{k!}e^{-\mu}\\
&=\sum_{k=0}^{+\infty}k^2\frac{\mu^k}{k!}e^{-\mu}-2\mu\mu\sum_{k=1}^{+\infty}\frac{k\mu^{k-1}}{k(k-1)!}e^{-\mu}+\mu^2\sum_{k=0}^{+\infty}\frac{\mu^k}{k!}e^{-\mu}\\
&=\sum_{k=0}^{+\infty}k^2\frac{\mu^k}{k!}e^{-\mu}-2\mu^2\sum_{k=1}^{+\infty}\frac{\mu^{k-1}}{(k-1)!}e^{-\mu}+\mu^2\cdot 1=\sum_{k=0}^{+\infty}k^2\frac{\mu^k}{k!}e^{-\mu}-2\mu^2+\mu^2\\
&=e^{-\mu}\sum_{k=0}^{+\infty}k^2\frac{\mu^k}{k!}-\mu^2=\mu e^{-\mu}\sum_{k=1}^{+\infty}k\frac{\mu^{k-1}}{(k-1)!}-\mu^2\\
&\underset{s=k-1}{=}\mu e^{-\mu}\sum_{s=0}^{+\infty}(s+1)\frac{\mu^s}{s!}-\mu^2=\mu e^{-\mu}\left(\sum_{s=0}^{+\infty}s\frac{\mu}{s!}+\sum_{s=0}^{+\infty}\frac{\mu^s}{s!}\right)-\mu^2\\
&=\mu e^{-\mu}\left(\mu\sum_{s=1}^{+\infty}\frac{\mu^{s-1}}{(s-1)!}+e^{\mu}\right)-\mu^2=\mu e^{-\mu}(\mu e^{\mu}+e^{\mu})-\mu^2\\
&=\mu e^{-\mu}\mu e^{\mu}+\mu e^{-\mu}e^{\mu}-\mu^2=\mu^2-\mu-\mu^2=\mu
\end{aligned}\tag{5.7.571}$$

always with:

$$\mu=np\tag{5.7.572}$$

The important fact for the Poisson distribution is that the variance that is equal to the mean is named the "equidispersion property of the Poisson distribution". This is a property often used in practice as an indicator to identify whether the data (with discrete support) are distributed according to a Poisson distribution.

The theoretical laws of statistical distributions are determined assuming completion of an infinite number of measurements. It is obvious that we can only perform in practice a finite number N. Hence the utility to establish correspondence between the theoretical and experimental values. For the experimental values we obviously obtain only an approximation whose validity is, however, often accepted as sufficient.

Now we will prove an important property of the Poisson distribution in the field of engineering and statistics that we name "stability by addition". The idea is as follows:

Let X and Y be two independent random variables with Poisson distribution of respective parameters λ and μ. We want to ensure that their sum is also a Poisson distribution:

$$X+Y=P_{\lambda+\mu}\tag{5.7.573}$$

See this:

$$P(X+Y=k)=\sum_{i=0}^{k}P\left[(X=i)\cap(Y=k-i)\right]=\sum_{i=0}^{k}P(X=i)P(Y=k-i) \quad (5.7.574)$$

because the events are independent. Then we have:

$$\begin{aligned}P(X+Y=k)&=\sum_{i=0}^{k}P(X=i)P(Y=k-i)=\sum_{i=0}^{k}\frac{\lambda^i e^{-\lambda}}{i!}\frac{\mu^{k-i}e^{\mu}}{(k-i)!}\\&=\frac{e^{-(\lambda+\mu)}}{k!}\sum_{i=0}^{k}\frac{k!}{i!(k-i)!}\lambda^i\mu^{k-i}\end{aligned} \quad (5.7.575)$$

However, by applying the binomial theorem (see section Calculus page 1364):

$$\sum_{i=0}^{k}\frac{k!}{i!(k-i)!}\lambda^i\mu^{k-i}=\sum_{i=0}^{k}C_i^k\lambda^i\mu^{k-i}=(\lambda+\mu)^k \quad (5.7.576)$$

So in the end:

$$P(X+Y=k)=(\lambda+\mu)^k\frac{e^{-(\lambda+\mu)}}{k!} \quad (5.7.577)$$

and therefore the Poisson distribution is stable by addition. So any Poisson distribution where the parameter is verbatim indefinitely dividable into a finite or infinite sum of independent Poisson distributions.

For the people who may not have access to a spreadsheet or statistical software here are some useful tables:

POISSON CUMULATIVE DISTRIBUTION FUNCTION

k/μ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	0.905	0.819	0.741	0.670	0.607	0.549	0.497	0.449	0.407	0.368
1	0.995	0.982	0.963	0.938	0.910	0.878	0.844	0.809	0.772	0.736
2	1.000	0.999	0.996	0.992	0.986	0.977	0.966	0.953	0.937	0.920
3	1.000	1.000	1.000	0.999	0.998	0.997	0.994	0.991	0.987	0.981
4	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.998	0.996
5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999
6	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

k/μ	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
0	0.333	0.301	0.273	0.247	0.223	0.202	0.183	0.165	0.150	0.135
1	0.699	0.663	0.627	0.592	0.558	0.525	0.493	0.463	0.434	0.406
2	0.900	0.879	0.857	0.833	0.809	0.783	0.757	0.731	0.704	0.677
3	0.974	0.966	0.957	0.946	0.934	0.921	0.907	0.891	0.875	0.857
4	0.995	0.992	0.989	0.986	0.981	0.976	0.970	0.964	0.956	0.947
5	0.999	0.998	0.998	0.997	0.996	0.994	0.992	0.990	0.987	0.983
6	1.000	1.000	1.000	0.999	0.999	0.999	0.998	0.997	0.997	0.995
7	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.999
8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

k/μ	2.2	2.4	2.6	2.8	3.0	3.2	3.4	3.6	3.8	4.0
0	0.111	0.091	0.074	0.061	0.050	0.041	0.033	0.027	0.022	0.018
1	0.355	0.308	0.267	0.231	0.199	0.171	0.147	0.126	0.107	0.092
2	0.623	0.570	0.518	0.469	0.423	0.380	0.340	0.303	0.269	0.238
3	0.819	0.779	0.736	0.692	0.647	0.603	0.558	0.515	0.473	0.433
4	0.928	0.904	0.877	0.848	0.815	0.781	0.744	0.706	0.668	0.629
5	0.975	0.964	0.951	0.935	0.916	0.895	0.871	0.844	0.816	0.785
6	0.993	0.988	0.983	0.976	0.966	0.955	0.942	0.927	0.909	0.889
7	0.998	0.997	0.995	0.992	0.988	0.983	0.977	0.969	0.960	0.949
8	1.000	0.999	0.999	0.998	0.996	0.994	0.992	0.988	0.984	0.979
9	1.000	1.000	1.000	0.999	0.999	0.998	0.997	0.996	0.994	0.992
10	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.998	0.997
11	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999
12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

k/μ	4.2	4.4	4.6	4.8	5.0	5.2	5.4	5.6	5.8	6.0
0	0.015	0.012	0.010	0.008	0.007	0.006	0.005	0.004	0.003	0.002
1	0.078	0.066	0.056	0.048	0.040	0.034	0.029	0.024	0.021	0.017
2	0.210	0.185	0.163	0.143	0.125	0.109	0.095	0.082	0.072	0.062
3	0.395	0.359	0.326	0.294	0.265	0.238	0.213	0.191	0.170	0.151
4	0.590	0.551	0.513	0.476	0.440	0.406	0.373	0.342	0.313	0.285
5	0.753	0.720	0.686	0.651	0.616	0.581	0.546	0.512	0.478	0.446
6	0.867	0.844	0.818	0.791	0.762	0.732	0.702	0.670	0.638	0.606
7	0.936	0.921	0.905	0.887	0.867	0.845	0.822	0.797	0.771	0.744
8	0.972	0.964	0.955	0.944	0.932	0.918	0.903	0.886	0.867	0.847
9	0.989	0.985	0.980	0.975	0.968	0.960	0.951	0.941	0.929	0.916
10	0.996	0.994	0.992	0.990	0.986	0.982	0.977	0.972	0.965	0.957
11	0.999	0.998	0.997	0.996	0.995	0.993	0.990	0.988	0.984	0.980
12	1.000	0.999	0.999	0.999	0.998	0.997	0.996	0.995	0.993	0.991
13	1.000	1.000	1.000	1.000	0.999	0.999	0.999	0.998	0.997	0.996
14	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.999
15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999
16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

k/μ	6.5	7.0	7.5	8.0	8.5	9.0	9.5	10.0	10.5	11.0
0	0.002	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.011	0.007	0.005	0.003	0.002	0.001	0.001	0.000	0.000	0.000
2	0.043	0.030	0.020	0.014	0.009	0.006	0.004	0.003	0.002	0.001
3	0.112	0.082	0.059	0.042	0.030	0.021	0.015	0.010	0.007	0.005
4	0.224	0.173	0.132	0.100	0.074	0.055	0.040	0.029	0.021	0.015
5	0.369	0.301	0.241	0.191	0.150	0.116	0.089	0.067	0.050	0.038
6	0.527	0.450	0.378	0.313	0.256	0.207	0.165	0.130	0.102	0.079
7	0.673	0.599	0.525	0.453	0.386	0.324	0.269	0.220	0.179	0.143
8	0.792	0.729	0.662	0.593	0.523	0.456	0.392	0.333	0.279	0.232
9	0.877	0.830	0.776	0.717	0.653	0.587	0.522	0.458	0.397	0.341
10	0.933	0.901	0.862	0.816	0.763	0.706	0.645	0.583	0.521	0.460
11	0.966	0.947	0.921	0.888	0.849	0.803	0.752	0.697	0.639	0.579
12	0.984	0.973	0.957	0.936	0.909	0.876	0.836	0.792	0.742	0.689
13	0.993	0.987	0.978	0.966	0.949	0.926	0.898	0.864	0.825	0.781
14	0.997	0.994	0.990	0.983	0.973	0.959	0.940	0.917	0.888	0.854
15	0.999	0.998	0.995	0.992	0.986	0.978	0.967	0.951	0.932	0.907
16	1.000	0.999	0.998	0.996	0.993	0.989	0.982	0.973	0.960	0.944
17	1.000	1.000	0.999	0.998	0.997	0.995	0.991	0.986	0.978	0.968
18	1.000	1.000	1.000	0.999	0.999	0.998	0.996	0.993	0.988	0.982
19	1.000	1.000	1.000	1.000	0.999	0.999	0.998	0.997	0.994	0.991
20	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.998	0.997	0.995
21	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.998
22	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999
23	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

5.7.6.9 Normal & Gauss-Laplace Distribution

This is the most important function of distribution in the field of statistics following a famous theorem named the "central limit theorem", which as we will see later, permits to prove (among other things) that any sum of independent identically distributed random variables with a finite mean and variance converges to a Laplace-Gaussian function (Normal distribution).

It is therefore very important to focus your attention on the developments that will be presented right now!

Let start from a binomial function and make tender the number of trials n to infinity. If p is set from the beginning , the mean $\mu = np$ also tends to infinity, furthermore the standard deviation $\sigma = npq$ also tends to infinity.

Remark

The case where p varies and tends to 0 while keeping fixed the mean has already been studied during the study of the Poisson function.

If we want to calculate the limit of the binomial function, it will then be necessary to make a change of origin, which stabilizes the mean, to 0 for example, and a change of unit change that stabilizes the standard deviation to 1 for example.

Let us now denote by $P_n(k)$ the binomial probability of k success and let's see first how $P_n(k)$ vary with k and calculate the difference:

$$\begin{aligned} P_n(k+1)-P_n(k) &= \binom{n}{k+1}p^{k+1}q^{n-k-1}-\binom{n}{k}p^kq^{n-k} \\ &= \binom{n}{k}p^kq^{n-k}\left(\frac{(n-k)p}{(k+1)q}-1\right)=P_n(k)\frac{np-k-q}{(k+1)q} \end{aligned} \tag{5.7.578}$$

We conclude that $P_n(k)$ is an increasing function of k, as $np-k-q$ is positive (for n, p and q fixed). Too see it, just take a few values (of the right term of the equality) or to observe the graph of the binomial distribution function, remembering that:

$$\mu = np \tag{5.7.579}$$

As $q<1$ it is therefore evident that the value of k close to the mean $\mu = np$ of the binomial distribution is the maxima of $P_n(k)$.

On the other hand the difference $P_n(k+1)-P_n(k)$ is the increase rate of the function $P_n(k)$. We can then write:

$$\frac{\Delta P_n(k)}{\Delta k}=\frac{P_n(k+1)-P_n(k)}{(k+1)-k} \tag{5.7.580}$$

as being the slope of the function.

We now define a new random variable such that its average is zero (negligible variations) and its standard deviation equal to the unit (this will be a centered-reduced variable in other words). Then we have:

$$x=\frac{k-np}{\sqrt{npq}} \tag{5.7.581}$$

Then we also have with this new random variable:

$$\Delta x=\frac{(k+1)-np}{\sqrt{npq}}-\frac{k-np}{\sqrt{npq}}=\frac{[(k+1)-np]-(k-np)}{\sqrt{npq}} \tag{5.7.582}$$

Let us denote $F(x)$ as being $P_n(k)$ calculated using the new random variable with zero mean and unit standard deviation which we seek the expression when n tends to infinity.

Let's go back to:

$$\frac{\Delta P_n(k)}{\Delta k}=\frac{np-k-q}{(k+1)q}P_n(k)=\frac{-(k-np)-q}{(k+1)q}P_n(k) \tag{5.7.583}$$

To simplify the study of this relation when n tends to infinity and k to mean $\mu=np$, multiply both sides by $npq/\sqrt{npq}$:

$$\frac{\Delta P_n(k)}{\Delta k}\frac{npq}{\sqrt{npq}}=\frac{np-k-q}{(k+1)q}\frac{npq}{\sqrt{npq}}P_n(k)=\frac{-(k-np)-q}{(k+1)q}\frac{npq}{\sqrt{npq}}P_n(k) \tag{5.7.584}$$

We rewrite now the right-hand side of this equality. Then we get:

$$\frac{-(k-np)-q}{(k+1)}\frac{np}{\sqrt{npq}}P_n(k)=\frac{[-(k-np)-q]np}{(k+1)\sqrt{npq}}P_n(k) \tag{5.7.585}$$

And now let us rewrite the left term of the prior-previous relation. We then get:

$$\begin{aligned}\frac{\Delta P_n(k)}{\Delta k}\frac{npq}{\sqrt{npq}}&=\frac{\Delta P_n(k)}{(k+1)-1}\frac{npq}{\sqrt{npq}}=\frac{\Delta P_n(k)}{[(k+1)-np]-(k-np)}\frac{npq}{\sqrt{npq}}\\&=\frac{\Delta P_n(k)}{\dfrac{[(k+1)-np]-(k-np)}{\sqrt{npq}}}\frac{npq}{\sqrt{npq}\sqrt{npq}}\\&=\frac{\Delta P_n(k)}{\dfrac{[(k+1)-np]-(k-np)}{\sqrt{npq}}}=\frac{\Delta P_n(k)}{\Delta x}\end{aligned} \tag{5.7.586}$$

After passing to the limit for n tending to infinity we have in a first time for the denominator of the second term of the prior-previous relation:

$$\frac{[-(k-np)-q]np}{(k+1)\sqrt{npq}}P_n(k) \tag{5.7.587}$$

the following simplification:

$$(k+1)\sqrt{npq}\underset{n\rightarrow+\infty}{=}k\sqrt{npq} \tag{5.7.588}$$

Thus:

$$\frac{-[(k-np)+q]np}{k\sqrt{npq}}P_n(k)=-\frac{k-np}{k\sqrt{npq}}npP_n(k)-\frac{qnp}{k\sqrt{npq}}P_n(k) \tag{5.7.589}$$

and in a second time, taking into account that the considered values of k are then in the neighborhood of the mean np, we get:

$$\frac{k-np}{k\sqrt{npq}}np=\frac{np}{k}\frac{k-np}{\sqrt{npq}}\underset{n\rightarrow+\infty}{\cong}\frac{k-np}{\sqrt{npq}}=x \tag{5.7.590}$$

and:

$$\frac{qnp}{k\sqrt{npq}}\underset{n\rightarrow+\infty}{\cong}\frac{qnp}{np\sqrt{npq}}=\frac{q}{\sqrt{npq}}\underset{n\rightarrow+\infty}{\cong}0 \tag{5.7.591}$$

Thus:

$$\frac{-[(k-np)+q]np}{k\sqrt{npq}}\underset{n\rightarrow+\infty}{\cong}-x \tag{5.7.592}$$

and as:

$$P_n(k)\underset{n\rightarrow+\infty}{:=}F(x) \tag{5.7.593}$$

where $F(x)$ represents (awkwardly) for the few lines that follow, the density function as n tends to infinity.

Finally we have:

$$\frac{\mathrm{d}F(x)}{\mathrm{d}x}=-xF(x) \tag{5.7.594}$$

This relation can also be rewritten rearranging the terms:

$$\frac{1}{F(x)}\mathrm{d}F(x)=-x\mathrm{d}x \tag{5.7.595}$$

and by integrating both sides of this equality we get (see section Differential and Integral Calculus page 1497):

$$\ln(F(x))=-\frac{x^2}{2}+c^{te} \tag{5.7.596}$$

The following function is a solution of the above equation:

$$F(x)=\lambda e^{-\frac{x^2}{2}} \tag{5.7.597}$$

Effectively:

$$\ln\left(\lambda e^{-\frac{x^2}{2}}\right)=\ln(\lambda)+\ln\left(e^{-\frac{x^2}{2}}\right)=c^{te}+\left(-\frac{x^2}{2}\right)=-\frac{x^2}{2}+c^{te} \tag{5.7.598}$$

The constant is determined by the condition that:

$$\int\limits_{-\infty}^{+\infty}F(x)\mathrm{d}x \tag{5.7.599}$$

which represents the sum of all probabilities, that mus be equal to 1. We prove for this that we need to have:

$$\lambda=\frac{1}{\sqrt{2\pi}} \tag{5.7.600}$$

Proof 5.47.2. We have:

$$\int\limits_{-\infty}^{+\infty} e^{\frac{-x^2}{2}}\mathrm{d}x=\int\limits_{-\infty}^{+\infty} e^{\left(-\frac{x}{\sqrt{2}}\right)^2}\mathrm{d}x=\sqrt{2}\int\limits_{-\infty}^{+\infty} e^{-z^2}\mathrm{d}z \tag{5.7.601}$$

So let us focus on the last term of this equality named the "Gauss integral" or "Gaussian integral" (in fact the case below is a special case of that integral) or also "probability integral". Thus:

$$I=\int\limits_{-\infty}^{+\infty} e^{-x^2}\mathrm{d}x=2\int\limits_{0}^{+\infty} e^{-x^2}\mathrm{d}x \tag{5.7.602}$$

since e^{-x^2} is an even function (see section Functional Analysis page 2212). Now we write the square of the integral as follows:

$$I^2=4\lim_{R\rightarrow+\infty}\left[\left(\int\limits_{0}^{R} e^{-x^2}\mathrm{d}x\right)\left(\int\limits_{0}^{R} e^{-y^2}\mathrm{d}y\right)\right]=4\lim_{R\rightarrow+\infty}\left(\int\limits_{0}^{R}\int\limits_{0}^{R} e^{-(x^2+y^2)}\mathrm{d}x\mathrm{d}y\right) \tag{5.7.603}$$

and make a change of variable passing in polar coordinates, therefore we also use the Jacobian in these same coordinates (see section Differential and Integral Calculus page 1524):

$$\begin{aligned}I^2&=4\lim_{R\rightarrow+\infty}\left(\int\int e^{-r^2}r\mathrm{d}r\mathrm{d}\phi\right)=4\lim_{R\rightarrow+\infty}\left(\int\limits_{0}^{\frac{\pi}{2}}\int\limits_{0}^{R} e^{-r^2}r\mathrm{d}r\mathrm{d}\phi\right)\\&=4\frac{\pi}{2}\lim_{R\rightarrow+\infty}\left(\int\limits_{0}^{R} e^{-r^2}r\mathrm{d}r\right)=2\pi\left(-\frac{1}{2}e^{-r^2}\Big|_0^{+\infty}\right)=2\pi\left(0+\frac{1}{2}\right)=\pi\end{aligned} \tag{5.7.604}$$

Therefore:

$$I=\sqrt{\pi} \tag{5.7.605}$$

By extension for $e^{-\frac{x^2}{2}}$ we have:

$$I=\lambda^{-1}=\sqrt{2\pi} \tag{5.7.606}$$

Hence, the reader must keep in mind for further developments that:

$$\boxed{I=\int\limits_{-\infty}^{+\infty} e^{-x^2}\mathrm{d}x=\sqrt{\pi}} \tag{5.7.607}$$

and that:

$$\boxed{I=\int\limits_{-\infty}^{+\infty} e^{-\frac{x^2}{2}}\mathrm{d}x=\sqrt{2\pi}} \tag{5.7.608}$$

■ Q.E.D.

We thus obtain the "standard Normal distribution" denoted as probability density function (with the capital letter F that may unfortunately lead to confusion in the present development with the notation of the cumulative distribution function... we apologize...):

$$F(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \tag{5.7.609}$$

which can be calculated in Microsoft® Excel® 11.8346 with the function `NORMSDIST( )`. Notice also that is some textbooks that latter relation is denoted by $\phi(x)$.

Remark

Originally Gauss used the term "normal" (in the sense of "orthogonal") referring to the geometric interpretation of a system of linear (normal) equations from which the distribution bearing his name was derived. Pearson in 11920 (holocene calendar) credited himself for first using the term "normal curve" noting that: « *while it avoids an international question of priority [of Gauss vs Laplace], has the disadvantage of leading people to believe that all other distributions are 'abnormal'.*»

For information, a random variable following a Normal centered reduced distribution is by tradition often denoted by the letter Z ("Zentriert" in German meaning "centered" in English) and its realization by a lowercase z.

Returning to non-normalized variables (i.e. non-standardized) by remembering that:

$$x=\frac{k-np}{\sqrt{npq}}=\frac{k-\mu}{\sigma} \tag{5.7.610}$$

We get the "Gauss-Laplace function" (or "Gauss-Laplace law") or also named "Normal distribution" given in the form of probability density in this textbook by:

$$\boxed{P(k,\mu,\sigma)=\mathcal{N}(\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(k-\mu)^2}{2\sigma^2}}} \tag{5.7.611}$$

Remark

So in general the reader must keep in mind that when a random variable x follows a $\mathcal{N}(\mu,\sigma)$, then standardizing it by subtracting it's mean (i.e. a "mean normalization") and dividing it by it's own standard deviation, make him follow a Normal centered reduced distribution Z so that its realization will be denoted:

$$z=\frac{x-\mu_x}{\sigma_x} \tag{5.7.612}$$

and named "Z-factor". Using this factor for non-Normal distributions can be a bit tricky. Without checking the Normality of the underlying data with a Kolmogorov-Smirnov, Anderson-Darling, Ryan-Joiner or one of the other many Normality tests (see page 1111) it should be avoided! Also doing a data adjustment like the Johnson transformation can completely change the interpretation of the results and in most cases should also be avoided. So as always in science: don't use any metric blindly without checking how it really works!

The cumulative probability (distribution cumulative function) to have a certain value k is obviously given in this special case by:

$$P(k \leq x) = \Phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-\frac{(k-\mu)^2}{2\sigma^2}} \mathrm{d}k \tag{5.7.613}$$

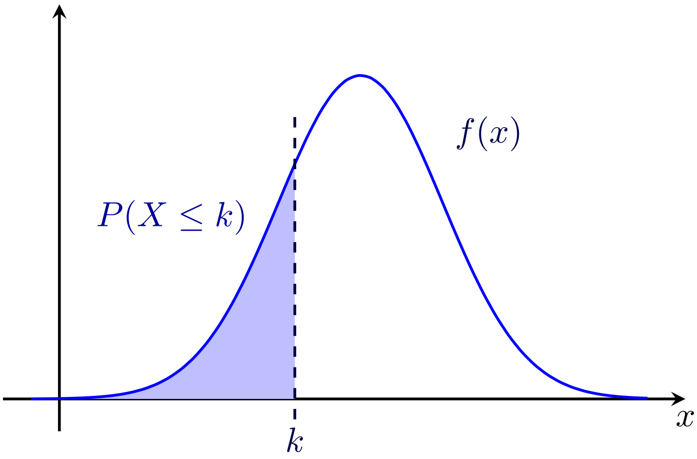


Figure 5.127: **Gaussian distribution cumulated area (author: Izaak Neutelings)**

On the request of a reader let us see how the cumulative function is often presented in textbooks... First remember that the density function of the Normal distribution is given by:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{5.7.614}$$

Thus, the cumulative distribution function is:

$$\begin{aligned} F_X(x) &= \int\limits_{-\infty}^{x} \mathcal{N}\left(z;\mu,\sigma^2\right) \mathrm{d}z \\ &= \int\limits_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2\right] \mathrm{d}z \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int\limits_{-\infty}^{x} \exp\left[-\left(\frac{z-\mu}{\sqrt{2}\sigma}\right)^2\right] \mathrm{d}z \end{aligned} \tag{5.7.615}$$

Substituting $t=(z-\mu)/(\sqrt{2}\sigma)$, i.e. $z=\sqrt{2}\sigma t+\mu$, this becomes:

$$\begin{aligned}F_X(x)&=\frac{1}{\sqrt{2\pi}\sigma}\int\limits_{(-\infty-\mu)/(\sqrt{2}\sigma)}^{(x-\mu)/(\sqrt{2}\sigma)}\exp\left(-t^2\right)\mathrm{d}(\sqrt{2}\sigma t+\mu)\\&=\frac{\sqrt{2}\sigma}{\sqrt{2\pi}\sigma}\int\limits_{-\infty}^{\frac{x-\mu}{\sqrt{2}\sigma}}\exp\left(-t^2\right)\mathrm{d}t\\&=\frac{1}{\sqrt{\pi}}\int\limits_{-\infty}^{\frac{x-\mu}{\sqrt{2}\sigma}}\exp\left(-t^2\right)\mathrm{d}t\\&=\frac{1}{\sqrt{\pi}}\int\limits_{-\infty}^{0}\exp\left(-t^2\right)\mathrm{d}t+\frac{1}{\sqrt{\pi}}\int_0^{\frac{x-\mu}{\sqrt{2}\sigma}}\exp\left(-t^2\right)\mathrm{d}t\\&=\frac{1}{\sqrt{\pi}}\int\limits_{0}^{\infty}\exp\left(-t^2\right)\mathrm{d}t+\frac{1}{\sqrt{\pi}}\int\limits_{0}^{\frac{x-\mu}{\sqrt{2}\sigma}}\exp\left(-t^2\right)\mathrm{d}t\end{aligned}\tag{5.7.616}$$

We recognize here the Error function (see section Thermodynamics page 3697) defined for reminder by:

$$\mathrm{erf}(x)=\frac{2}{\sqrt{\pi}}\int\limits_0^x\exp\left(-t^2\right)\mathrm{d}t\tag{5.7.617}$$

Hence the last equality can be rewritten:

$$\begin{aligned}F_X(x)&=\frac{1}{2}\lim_{x\to\infty}\mathrm{erf}(x)+\frac{1}{2}\mathrm{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\\&=\frac{1}{2}+\frac{1}{2}\mathrm{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\\&=\frac{1}{2}\left[1+\mathrm{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right]\end{aligned}\tag{5.7.618}$$

Hence the famous relation:

$$F_X(x)=\frac{1}{2}\left[1+\mathrm{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right]\tag{5.7.619}$$

Caution! Because there is no closed formula for the cumulative distribution of the Normal law (and the same occurs for others statistical laws), there are plenty of different numerical implementation of the latter. For example eighteen (!!!) different numerical implementation exist for the Normal law as detailed in the paper [587] (the Zelen and Severo, Hart, Marsaglia, Cody, Dia, Shore methods etc.). That's why you will get the same result in SAS and Microsoft® Excel® at the eighth decimal but the result will be slightly different from Python™ and the GNU Scientific Library that doesn't use the same numerical implementation. None of the returned value is more exact than another one! It would be a nonsense ton claim the opposite! That's why professional scientists and statisticians communicate and validate their results for very important and critical statistics using three different tools/softwares at least!

Here is a plot example of the distribution and cumulative distribution function for the Normal law with the parameters example $(\mu, \sigma) = (0, 1)$ that is therefore the standard centered reduced Normal distribution:

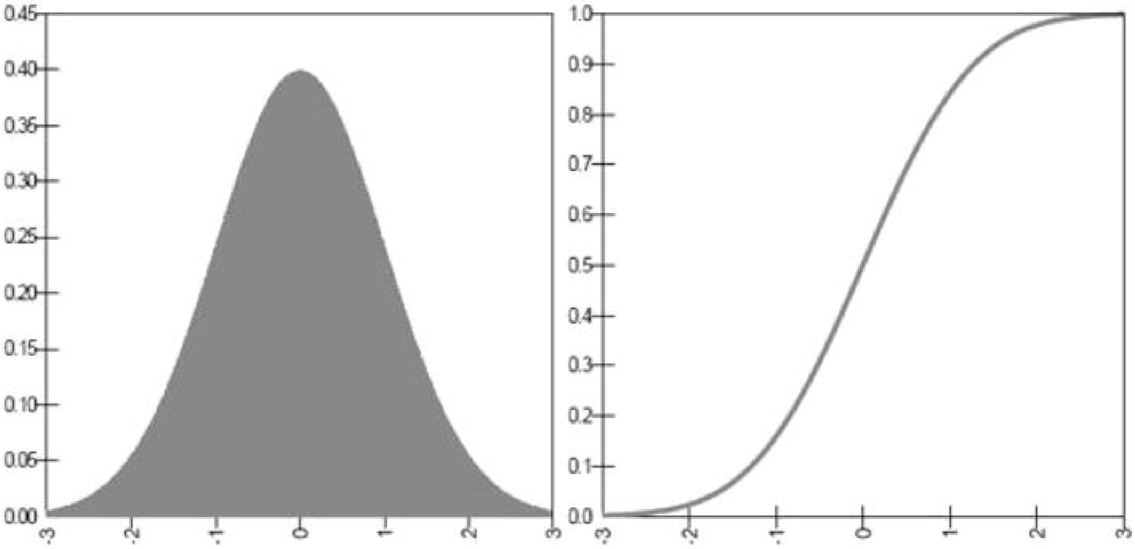


Figure 5.128: **Normal law $\mathcal{N}$ (mass and cumulative distribution function)**

This law governs under very general conditions, and often encountered, in many random phenomena. It is also symmetrical with respect to the mean (this is important to remember!).

We will now show that μ well represents the mean (or average) of x (this is a bit silly but we can still check ...):

$$\text{E}(X)=\int\limits_{-\infty}^{+\infty} xf(x)\mathrm{d}x=\frac{1}{\sigma\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\mathrm{d}x \tag{5.7.620}$$

We put:

$$u=\frac{x-\mu}{\sigma} \tag{5.7.621}$$

We then have:

$$\begin{aligned}\mathrm{E}(X)&=\int\limits_{-\infty}^{+\infty}xf(x)\mathrm{d}x=\frac{1}{\sigma\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}(\sigma u+\mu)e^{-\frac{u^2}{2}}\sigma\mathrm{d}u\\&=\sigma\frac{1}{\sigma\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}\sigma ue^{-\frac{u^2}{2}}\mathrm{d}u+\sigma\frac{1}{\sigma\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}\mu e^{-\frac{u^2}{2}}\mathrm{d}u\\&=\sigma\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}ue^{-\frac{u^2}{2}}\mathrm{d}u+\frac{1}{\sqrt{2\pi}}\mu\int\limits_{-\infty}^{+\infty}e^{-\frac{u^2}{2}}\mathrm{d}u\end{aligned}\tag{5.7.622}$$

Let us now calculate the first integral:

$$J=\int\limits_{-\infty}^{+\infty}ue^{-\frac{u^2}{2}}\mathrm{d}u=-e^{-\frac{u^2}{2}}\Big|_{-\infty}^{+\infty}=e^{-\frac{u^2}{2}}\Big|_{+\infty}^{-\infty}=0-0=0\tag{5.7.623}$$

So we finally get:

$$\mathrm{E}(X)=\frac{1}{\sqrt{2\pi}}\mu\int\limits_{-\infty}^{+\infty}e^{-\frac{u^2}{2}}\mathrm{d}u=\frac{1}{\sqrt{2\pi}}\mu\sqrt{2\pi}=\mu\tag{5.7.624}$$

Remarks

R1. The reader might find confusing at first that the parameter of a function is one of the results that we seek of this same function (as for the Poisson distribution). What bothers is to put in practice such a thing. In fact, everything will be more clear when we will discuss later in this chapter the concepts of "likelihood estimators".

R2. It could be interesting to know for the reader that in practice (finance, quality assurance, etc.) it is common to have to calculate only mean only positive values of the random variable which is then naturally defined as "positive mean" and given by:

$$\mathrm{E}^{+}(X)=\frac{1}{\sigma\sqrt{2\pi}}\int\limits_{0}^{+\infty}xe^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\mathrm{d}x\tag{5.7.625}$$

We will see a practical example of this last relation in the section Economy during our study of the theoretical model of speculation of Louis Bachelier (page 7689).

Also we will prove now (...) that σ is the standard deviation of X (in other words to prove that $\mathrm{V}(X)=\sigma^2$) and for this we recall that we had prove that (Huygens relation):

$$\mathrm{V}(X)=\mathrm{E}(X^2)-\mathrm{E}(X)^2\tag{5.7.626}$$

We already know that at the level of the notations we have:

$$\mathrm{E}(X)=\mu\Rightarrow\mathrm{E}(X)^2=\mu^2\tag{5.7.627}$$

then we first calculate $\mathrm{E}(X^2)$:

$$\mathrm{E}(X^2)=\int x^2 f(x)\mathrm{d}x=\frac{1}{\sigma\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} x^2 e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\mathrm{d}x \qquad (5.7.628)$$

Let $y=(x-\mu)/\sqrt{2}\sigma$ that therefore leads us to:

$$\begin{aligned}\mathrm{E}(X^2)&=\int x^2 f(x)\mathrm{d}x=\frac{1}{\sigma\sqrt{\pi}}\int\limits_{-\infty}^{+\infty}(y\sqrt{2}\sigma+\mu)^2 e^{-y^2}\mathrm{d}x\\&=\frac{2}{\sigma\sqrt{\pi}}\int\limits_{-\infty}^{+\infty} y^2 e^{-y^2}\mathrm{d}y+2\mu\sigma\sqrt{\frac{2}{\pi}}\int\limits_{-\infty}^{+\infty} y e^{-y^2}\mathrm{d}y+\frac{\mu^2}{\sqrt{\pi}}\int\limits_{-\infty}^{+\infty} e^{-y^2}\mathrm{d}y\end{aligned} \qquad (5.7.629)$$

And we know that (already proved above):

$$\int\limits_{-\infty}^{+\infty} y e^{-y^2}\mathrm{d}y=0 \qquad \int\limits_{-\infty}^{+\infty} e^{-y^2}\mathrm{d}y=\sqrt{\pi} \qquad (5.7.630)$$

It remains to calculate therefore only the first integral. To do this, we proceed by integration by parts (see section Differential and Integral Calculus page 1532):

$$\int\limits_a^b f(t)g'(t)\mathrm{d}t=f(t)g(t)|_a^b-\int\limits_a^b f'(t)g(t)\mathrm{d}t \qquad (5.7.631)$$

leads us to:

$$\int\limits_{-\infty}^{+\infty} y^2 e^{-y^2}\mathrm{d}y=\int\limits_{-\infty}^{+\infty} y\left(ye^{-y^2}\right)\mathrm{d}y=-y\frac{e^{-\frac{y^2}{2}}}{2}\Bigg|_a^b+\int\limits_{-\infty}^{+\infty}\frac{e^{-y^2}}{2}\mathrm{d}y=\frac{\sqrt{\pi}}{2} \qquad (5.7.632)$$

Then we get:

$$\mathrm{E}(X^2)=\frac{2}{\sqrt{\pi}}\sigma^2\frac{\sqrt{\pi}}{2}+2\mu\sigma\sqrt{\frac{2}{\pi}}0+\frac{\mu^2}{\sqrt{\pi}}\sqrt{\pi}=\mu^2+\sigma^2 \qquad (5.7.633)$$

And finally:

$$\mathrm{V}(X)=\mathrm{E}(X^2)-\mathrm{E}(X)^2=(\mu^2+\sigma^2)-\mu^2=\sigma^2 \qquad (5.7.634)$$

An additional signification of the standard deviation of Gauss-Laplace distribution is a measure of the width of the distribution as (this can be checked only with the aid of integration by using numerical methods) for any non-zero mean and standard deviation we have (thanks to John Cannin for the LaTeX figure):

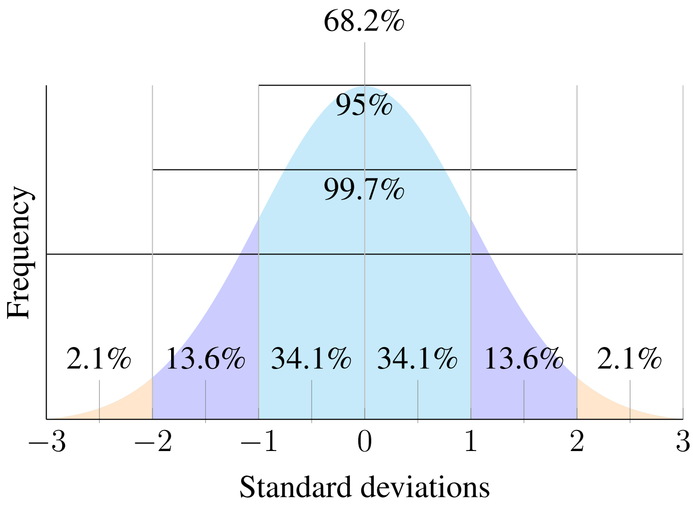


Figure 5.129: **Sigma intervals for the Normal distribution**

The width of the interval has a great importance in the interpretation of uncertainties measurement. The presentation of a result like $\bar{N} \pm \sigma$ has for signification that the average value has about 68.3% chance (probability) to lie between the limits of $\bar{N}-\sigma$ and $\bar{N}+\sigma$ or has approximately 95.4% to lie between the limits of $\bar{N}-2\sigma$ and $\bar{N}+2\sigma$ etc.

Let X be a random variable following a Normal distribution with mean μ and variance σ^2. Then, about 68%, 95% and 99.7% of the values of X will fall within 1, 2 and 3 standard deviations from the mean, respectively:

$$\begin{aligned} P(\mu-1\sigma\leq X\leq \mu+1\sigma)&\cong 68\%\\ P(\mu-2\sigma\leq X\leq \mu+2\sigma)&\cong 95\%\\ P(\mu-3\sigma\leq X\leq \mu+3\sigma)&\cong 99.7\% \end{aligned} \tag{5.7.635}$$

this is often named the "68-95-99.7 rule".

Indeed, the as we have seen just earlier, the cumulative distribution function of a Normally distributed random variable X is:

$$F_X(x)=\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right] \tag{5.7.636}$$

where $\operatorname{erf}(x)$ is for recall the error function defined as:

$$\operatorname{erf}(x)=\frac{2}{\sqrt{\pi}}\int\limits_0^x \exp\left(-t^2\right)\mathrm{d}t \tag{5.7.637}$$

which exhibits a point-symmetry property:

$$\operatorname{erf}(-x)=-\operatorname{erf}(x) \tag{5.7.638}$$

Thus, the probability that X falls between $\mu - a\cdot\sigma$ and $\mu + a\cdot\sigma$ is equal to:

$$\begin{aligned}
p(a) &= P(\mu - a\sigma \leq X \leq \mu + a\sigma)\\
&= F_X(\mu + a\sigma) - F_X(\mu - a\sigma)\\
&= \frac{1}{2}\left[1 + \text{erf}\left(\frac{\mu + a\sigma - \mu}{\sqrt{2}\sigma}\right)\right] - \frac{1}{2}\left[1 + \text{erf}\left(\frac{\mu - a\sigma - \mu}{\sqrt{2}\sigma}\right)\right]\\
&= \frac{1}{2}\left[\text{erf}\left(\frac{\mu + a\sigma - \mu}{\sqrt{2}\sigma}\right) - \text{erf}\left(\frac{\mu - a\sigma - \mu}{\sqrt{2}\sigma}\right)\right]\\
&= \frac{1}{2}\left[\text{erf}\left(\frac{a}{\sqrt{2}}\right) - \text{erf}\left(-\frac{a}{\sqrt{2}}\right)\right]\\
&= \frac{1}{2}\left[\text{erf}\left(\frac{a}{\sqrt{2}}\right) + \text{erf}\left(\frac{a}{\sqrt{2}}\right)\right]\\
&= \text{erf}\left(\frac{a}{\sqrt{2}}\right)
\end{aligned} \tag{5.7.639}$$

With that, we can use numerical implementations of the error function to calculate:

$$\begin{aligned}
P(\mu - 1\sigma \leq X \leq \mu + 1\sigma) = p(1) = 68.27\%\\
P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = p(2) = 95.45\%\\
P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = p(3) = 99.73\%
\end{aligned} \tag{5.7.640}$$

Remark

This concept is widely used in quality management in industrial business especially with the Six Sigma methodology (see Industrial Engineering page 8674) which requires a mastery of 6 around each side of the mean (!) of the manufacturing (or anything else whose deviation is measured).

Sigma Quality Level	Insured non-defection rate in %	Defection rate in parts per million
1σ	68.26894	317,311
2σ	95.4499	45,500
3σ	99.73002	2,700
4σ	99.99366	63.4
5σ	99.999943	0.57
6σ	99.9999998	0.002

Table 5.41: **Sigma quality level with defection / non-defection rate**

The second column of the table can easily be obtained with Maple™ 4.00b (or also with the spreadsheet software from Microsoft). For example for the first line:

```
>S:=evalf(int(1/sqrt(2*Pi)*exp(-x^2/2),x=-1..1));
```

and the first row of the third column:

```
>(1-S)*1E6;
```

If the Normal distribution was not centered, then we just would write for the second column:

```
>S:=evalf(int(1/sqrt(2*Pi)*exp(-(x-mu)^2/2),x=-1..1));
```

and so on for any deviation and mean we will then obtain exactly the same intervals!!!

Let X be a random variable following a Normal distributions :

$$X \sim \mathcal{N}\left(\mu, \sigma^2\right) \tag{5.7.641}$$

Then, the quantile function of X is:

$$Q_X(p) = \sqrt{2}\sigma \cdot \text{erf}^{-1}(2p-1) + \mu \tag{5.7.642}$$

where $\text{erf}^{-1}(x)$ is the inverse error function.

Indeed, we have proven earlier that the cumulative function of the Normal distribution was:

$$F_X(x) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right] \tag{5.7.643}$$

Because the cumulative distribution function (CDF) is strictly monotonically increasing, the quantile function is equal to the inverse of the CDF :

$$Q_X(p) = F_X^{-1}(x) \tag{5.7.644}$$

This can be derived by rearranging the prior-previous relation:

$$\begin{aligned} p &= \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right] \\ 2p-1 &= \text{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right) \\ \text{erf}^{-1}(2p-1) &= \frac{x-\mu}{\sqrt{2}\sigma} \\ x &= \sqrt{2}\sigma \cdot \text{erf}^{-1}(2p-1) + \mu \end{aligned} \tag{5.7.645}$$

Let us now seek the median of the Normal law (obviously by its symmetry that's quite obvious...)! Remember that the cumulative distribution function of the Normal distribution is given by:

$$F_X(x) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right] \tag{5.7.646}$$

where $\text{erf}(x)$ is the error function. Thus, the inverse cumulative distribution function is:

$$x = \sqrt{2}\sigma \cdot \text{erf}^{-1}(2p-1) + \mu \tag{5.7.647}$$

where $\text{erf}^{-1}(x)$ is the inverse error function. Setting $p = 1/2$, we get:

$$\text{median}(X) = M_e(X) = \sqrt{2}\sigma \cdot \text{erf}^{-1}(0) + \mu = \mu \tag{5.7.648}$$

Following a reader request, let's prove how to obtain the modal value of the Normal law (even if it is trivial given its symmetry and shape...). So Let X be a random variable following a Normal distribution:

$$X \sim \mathcal{N}\left(\mu, \sigma^2\right) \tag{5.7.649}$$

The probability density function of the Normal distribution is for reminder given by:

$$\text{mode}(X) = \underset{x}{\arg\max} f_X(x) \tag{5.7.650}$$

The first derivative of this function is:

$$f'_X(x) = \frac{\mathrm{d}f_X(x)}{\mathrm{d}x} = \frac{1}{\sqrt{2\pi\sigma^3}} \cdot (-x+\mu) \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{5.7.651}$$

We can therefore calculate the root of the first derivative to get the result:

$$\begin{aligned} f'_X(x) = 0 &= \frac{1}{\sqrt{2\pi\sigma^3}} \cdot (-x+\mu) \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \\ 0 &= -x+\mu \\ x &= \mu \end{aligned} \tag{5.7.652}$$

Let us now derive the skewness of the centered reduced Normal distribution define for recall as:

$$\gamma_1 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mathrm{E}[X^3] - 3\mu\sigma^2 - \mu^3}{\sigma^3} \tag{5.7.653}$$

And hence for the Normal centered reduced distribution $\mathcal{N}(0,1)$:

$$\gamma_1 = \frac{\mathrm{E}[X^3] - 3\cdot 0 \cdot 1 - 0^3}{1^3} = \mathrm{E}[X^3] \tag{5.7.654}$$

Let now $Y = X - a$ be a random variable. Now note that due to symmetry of the Normal law Y and $-Y$ have the same distribution. That implies:

$$\mathrm{E}[Y^3] = \mathrm{E}[(-Y)^3] \tag{5.7.655}$$

This implies that the only possible solution is $\mathrm{E}[Y^3] = 0$. Hence for the Normal Law:

$$\gamma_1 = \mathrm{E}[X^3] = 0 \tag{5.7.656}$$

Let us now derive the kurtosis κ_1 and κ_4 for the centered reduced Normal distribution:

$$\mathcal{N}(0,1) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} e^{\frac{x^2}{2}} \mathrm{d}x \tag{5.7.657}$$

For this purpose we use the following usual primitive (see section Differential and Integral Calculus page 1534):

$$\int x^n e^{ax} \mathrm{d}x = \frac{x^n e^{ax}}{a} - \frac{n}{a} \int x^{n-1} a^{ax} \mathrm{d}x \tag{5.7.658}$$

Therefore using also the Gauss integral:

$$\begin{aligned}
\mu_4 &= \int\limits_{-\infty}^{+\infty} x^4 f(x)\mathrm{d}x = \frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} x^4 e^{-\frac{x^2}{2}}\mathrm{d}x\\
&= \frac{2}{\sqrt{2\pi}}\int\limits_{0}^{+\infty} x^4 e^{-\frac{x^2}{2}}\mathrm{d}x = \frac{1}{\sqrt{2\pi}}\int\limits_{0}^{+\infty} y^{\frac{3}{2}} e^{-\frac{y}{2}}\mathrm{d}y\\
&= \frac{1}{\sqrt{2\pi}}\left(-2y^{\frac{3}{2}}e^{-\frac{y}{2}}\Bigg|_0^{+\infty} + 3\int\limits_{0}^{+\infty} y^{\frac{1}{2}} e^{-\frac{y}{2}}\mathrm{d}y\right)\\
&= \frac{1}{\sqrt{2\pi}}\left(-2y^{\frac{3}{2}}e^{-\frac{y}{2}}\Bigg|_0^{+\infty} + 3\left[-2y^{\frac{1}{2}}e^{-\frac{y}{2}}\Bigg|_0^{+\infty} + \int\limits_{0}^{+\infty} y^{-\frac{1}{2}} e^{-\frac{y}{2}}\mathrm{d}y\right]\right)\\
&= \frac{1}{\sqrt{2\pi}}\left(-2y^{\frac{3}{2}}e^{-\frac{y}{2}}\Bigg|_0^{+\infty} + 3\left[-2y^{\frac{1}{2}}e^{-\frac{y}{2}}\Bigg|_0^{+\infty} + 2\int\limits_{0}^{+\infty} e^{-\frac{x^2}{2}}\mathrm{d}x\right]\right)\\
&= \frac{1}{\sqrt{2\pi}}\left(-2y^{\frac{3}{2}}e^{-\frac{y}{2}}\Bigg|_0^{+\infty} + 3\left[-2y^{\frac{1}{2}}e^{-\frac{y}{2}}\Bigg|_0^{+\infty} + 2\frac{\sqrt{2\pi}}{2}\right]\right)\\
&= \frac{1}{\sqrt{2\pi}}\left(-2y^{\frac{3}{2}}e^{-\frac{y}{2}} - 6y^{\frac{1}{2}}e^{-\frac{y}{2}}\right)\Bigg|_0^{+\infty} + 3\\
&= 3
\end{aligned} \tag{5.7.659}$$

Therefore, the kurtosis is:

$$\kappa_1 = \frac{\mu_4}{\sigma^4} = \frac{3}{1^4} = 3 \tag{5.7.660}$$

Hence the normalized kurtosis:

$$\kappa_4 = \kappa_1 - 3 = 0 \tag{5.7.661}$$

Now for our study later of diffraction, let us calculate the "full width at half maximum" (FWHM) which is the difference between the two values of the independent variable at which the dependent variable is equal to half of its maximum value. In other words, it is the width of a spectrum curve measured between those points on the y-axis which are half the maximum amplitude.

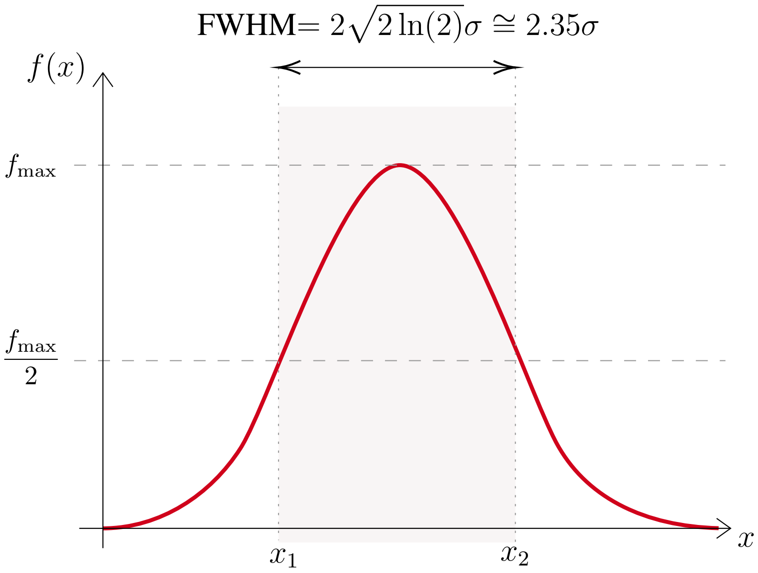


Figure 5.130: **Full width at half maximum (FWHM) illustrated**

FWHM is applied to such phenomena as the duration of pulse waveforms and the spectral width of sources used for optical communications and the resolution capabilities of spectrometers (to calculate crystallite site). The convention of "width" meaning "half maximum" is also widely used in signal processing to define bandwidth as "width of frequency range where less than half the signal's power is attenuated", i.e., the power is at least half the maximum. In signal processing terms, this is at most -3 [dB] of attenuation, named "half-power point" or, more specifically, "half-power bandwidth". When half-power point is applied to antenna beam width, it is named "half-power beam width".

So let X be a random variable following a Normal distribution:

$$X \sim \mathcal{N}\left(\mu,\sigma^2\right) \tag{5.7.662}$$

Then, the full width at half maximum (FWHM) of X is:

$$\text{FWHM}(X) = 2\sqrt{2\ln(2)}\sigma \tag{5.7.663}$$

Indeed, remember that the probability density function of the Normal distribution is as we have seen earlier above:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma}\cdot\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{5.7.664}$$

and we have proven just earlier that mode of the Normal distribution is:

$$\text{mode}(X) = \mu \tag{5.7.665}$$

such that:

$$f_{\text{max}} = f_X(\text{mode}(X)) = f_X(\mu) = \frac{1}{\sqrt{2\pi}\sigma} \tag{5.7.666}$$

The FWHM bounds satisfy the equation:

$$f_X\left(x_{\text{FWHM}}\right) = \frac{1}{2}f_{\text{max}} = \frac{1}{2\sqrt{2\pi}\sigma} \tag{5.7.667}$$

Using this cumulative distribution function we can develop this equation as follows:

$$\begin{aligned}
\frac{1}{\sqrt{2\pi}\sigma}\cdot\exp\left[-\frac{1}{2}\left(\frac{x_{\text{FWHM}}-\mu}{\sigma}\right)^2\right] &= \frac{1}{2\sqrt{2\pi}\sigma}\\
\exp\left[-\frac{1}{2}\left(\frac{x_{\text{FWHM}}-\mu}{\sigma}\right)^2\right] &= \frac{1}{2}\\
-\frac{1}{2}\left(\frac{x_{\text{FWHM}}-\mu}{\sigma}\right)^2 &= \ln\frac{1}{2}\\
\left(\frac{x_{\text{FWHM}}-\mu}{\sigma}\right)^2 &= -2\ln\left(\frac{1}{2}\right)\\
\frac{x_{\text{FWHM}}-\mu}{\sigma} &= \pm\sqrt{2\ln(2)}\\
x_{\text{FWHM}}-\mu &= \pm\sqrt{2\ln(2)}\sigma\\
x_{\text{FWHM}} &= \pm\sqrt{2\ln(2)}\sigma+\mu
\end{aligned}\tag{5.7.668}$$

This implies the following two solutions for x_{FWHM}:

$$\begin{aligned}
x_1 &= \mu-\sqrt{2\ln(2)}\sigma\\
x_2 &= \mu+\sqrt{2\ln(2)}\sigma
\end{aligned}\tag{5.7.669}$$

such that the full width at half maximum of X is:

$$\begin{aligned}
\text{FWHM}(X) &= \Delta x = x_2-x_1\\
&= (\mu+\sqrt{2\ln(2)}\sigma)-(\mu-\sqrt{2\ln(2)}\sigma)\\
&= 2\sqrt{2\ln(2)}\sigma
\end{aligned}\tag{5.7.670}$$

The probability density function of the Normal distribution with mean μ and variance σ^2 has two inflection points at $x=\mu-\sigma$ and $x=\mu+\sigma$, i.e. exactly one standard deviation away from the expected value!

To prove this let us remind that the probability density function of the Normal distribution is given by:

$$f_X(x)=\frac{1}{\sqrt{2\pi}\sigma}\cdot\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]\tag{5.7.671}$$

The first derivative of this function is:

$$f'_X(x)=\frac{\mathrm{d}f_X(x)}{\mathrm{d}x}=\frac{1}{\sqrt{2\pi}\sigma}\cdot\left(-\frac{x-\mu}{\sigma^2}\right)\cdot\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]\tag{5.7.672}$$

The second derivative is therefore:

$$\begin{aligned}
f''_X(x)=\frac{\mathrm{d}^2f_X(x)}{\mathrm{d}x^2} &= \frac{1}{\sqrt{2\pi}\sigma}\cdot\left(-\frac{1}{\sigma^2}\right)\cdot\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]+\frac{1}{\sqrt{2\pi}\sigma}\cdot\left(\frac{x-\mu}{\sigma^2}\right)^2\cdot\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]\\
&= \frac{1}{\sqrt{2\pi}\sigma}\cdot\left[\left(\frac{x-\mu}{\sigma^2}\right)^2-\frac{1}{\sigma^2}\right]\cdot\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]
\end{aligned}\tag{5.7.673}$$

Therefore second derivative is zero, if and only if:

$$\begin{aligned}
0 &= \left[\left(\frac{x-\mu}{\sigma^2}\right)^2-\frac{1}{\sigma^2}\right]\\
0 &= \frac{x^2}{\sigma^4}-\frac{2\mu x}{\sigma^4}+\frac{\mu^2}{\sigma^4}-\frac{1}{\sigma^2}\\
0 &= x^2-2\mu x+\left(\mu^2-\sigma^2\right)\\
x_{1,2} &= -\frac{-2\mu}{2}\pm\sqrt{\left(\frac{-2\mu}{2}\right)^2-\left(\mu^2-\sigma^2\right)}\\
x_{1,2} &= \mu\pm\sqrt{\mu^2-\mu^2+\sigma^2}\\
x_{1,2} &= \mu\pm\sigma
\end{aligned} \tag{5.7.674}$$

hence there are inflection points at $x_{1/2}=\mu\pm\sigma$.

The Gauss-Laplace distribution is also not only a tool for data analysis but also for data generation. Indeed, this distribution is one of the largest used in the world of multinationals that use statistical tools for risk management, project management and simulation where a large number of random variables are to be controlled. The best examples of applications use the softwares Crystal Ball or Palisade @Risk (this last one being my favorite...).

In this context of application (project management), it is also very common to use the sum (task duration) or the product of random variables (customer uncertainty) following Gauss-Laplace distributions. We will see now how to calculate this:

5.7.6.9.1 Sum of two random Normal variables

Let X, Y be two independent random variables. Suppose that X follows the distribution $\mathcal{N}(\mu_1,\sigma_1)$ and that Y follows the distribution $\mathcal{N}(\mu_2,\sigma_2)$. Then the random variable $Z=X+Y$ has a density equal to the convolution product of f_x and f_y (see section Functional Analysis page 2293). That is to say:

$$f_z(s)=\int\limits_{-\infty}^{+\infty} f_X(x)f_Y(y)(s-x)\mathrm{d}x=\frac{1}{2\pi\sigma_1\sigma_2}\int\limits_{-\infty}^{+\infty} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}e^{-\frac{(s-x-\mu_2)^2}{2\sigma_2^2}}\mathrm{d}x \tag{5.7.675}$$

which is equivalent to the joint product (see section Probabilities page 480) of the probabilities of occurrence of the two continuous variables (remember the same kind of calculation in discrete form!).

To simplify the expression, make the change of variable $t=x-\mu_1$ and let us write:

$$a=\mu_1+\mu_2-s \qquad \sigma=\sqrt{\sigma_1^2+\sigma_2^2} \tag{5.7.676}$$

As:

$$\left(s-y-\mu_2^2\right)^2=-\left(-s+y+\mu_2^2\right)^2=\left(-s+y+\mu_2^2\right)^2=(t+a)^2 \tag{5.7.677}$$

we get after a hard to guess rearrangement trick:

$$\begin{aligned}f_z(s)&=\dfrac{1}{2\pi\sigma_1\sigma_2}\int\limits_{-\infty}^{+\infty}e^{-\frac{t^2}{2\sigma_1^2}}e^{-\frac{(t+a)^2}{2\sigma_2^2}}\mathrm{d}t=\dfrac{1}{2\pi\sigma_1\sigma_2}\int\limits_{-\infty}^{+\infty}e^{-\frac{\left(\sigma t+\frac{a\sigma_1^2}{\sigma}\right)^2+\frac{\sigma_1^2a^2\sigma_2^2}{\sigma^2}}{2\sigma_1^2\sigma_2^2}}\mathrm{d}t\\&=\dfrac{1}{2\pi\sigma_1\sigma_2}e^{-\frac{a^2}{2\sigma^2}}\int\limits_{-\infty}^{+\infty}e^{-\frac{\left(\sigma t+\frac{a\sigma_1^2}{\sigma}\right)^2}{2\sigma_1^2\sigma_2^2}}\mathrm{d}t\end{aligned}\tag{5.7.678}$$

We write:

$$u=\dfrac{\sigma t+\dfrac{a\sigma_1^2}{\sigma}}{\sqrt{2}\sigma_1\sigma_2}\Rightarrow\dfrac{\mathrm{d}u}{\mathrm{d}t}=\dfrac{\sigma}{\sqrt{2}\sigma_1\sigma_2}\Rightarrow\mathrm{d}t=\sqrt{2}\sigma_1\sigma_2\dfrac{\mathrm{d}u}{\sigma}\tag{5.7.679}$$

Then:

$$f_z(s)=\dfrac{1}{2\pi\sigma_1\sigma_2}e^{-\frac{a^2}{2\sigma^2}}\int\limits_{-\infty}^{+\infty}e^{-\frac{\left(\sigma t+\frac{a\sigma_1^2}{\sigma}\right)^2}{2\sigma_1^2\sigma_2^2}}\mathrm{d}t=\dfrac{1}{\sqrt{2}\pi\sigma}e^{-\frac{a^2}{2\sigma^2}}\int\limits_{-\infty}^{+\infty}e^{-u^2}\mathrm{d}u\tag{5.7.680}$$

Knowing that (already proved earlier above!):

$$\int\limits_{-\infty}^{+\infty}e^{-u^2}\mathrm{d}u=\sqrt{\pi}\tag{5.7.681}$$

and:

$$a^2=(-s+\mu_1+\mu_2)^2=(s-\mu_1-\mu_2)^2\tag{5.7.682}$$

our relation becomes:

$$\boxed{f_z(s)=\dfrac{1}{\sigma\sqrt{2}\sqrt{\pi}}e^{-\frac{(s-\mu_1-\mu_2)^2}{2\sigma^2}}}\tag{5.7.683}$$

We recognize the expression of the Gauss-Laplace distribution (Normal law) of mean $\mu_1+\mu_2$ and standard deviation $\sigma=\sqrt{\sigma_1^2+\sigma_2^2}$.

Therefore, $X+Y$ follows the distribution as written by physicists (both argument have same units):

$$\mathcal{N}\left(\mu_1+\mu_2,\sqrt{\sigma_1^2+\sigma_2^2}\right)\tag{5.7.684}$$

and as noted by most mathematicians and statisticians:

$$\mathcal{N}\left(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2\right)\tag{5.7.685}$$

The fact that the sum of two Normal random variables always give also a Normal distribution is what we name in statistics "stability of the sum" of the Gauss-Laplace distribution (Normal law). We will find such properties for other distribution that will be discussed later.

So as well as for Poisson random variables, any Normal random variable whose parameters are known is verbatim indefinitely divisible into a finite or infinite number of independent Normal variables that are summed as:

$$\mathcal{N}(\mu,\sigma)=\sum_{i=1}^{n}\mathcal{N}\left(\frac{\mu}{n},\frac{\sigma^2}{n}\right) \tag{5.7.686}$$

Remark

The families of stable distribution by the sum is an important field of study in physics, finance and statistics named "Levy alpha-stable distribution". If time permits, I will present the details of this extremely important study in this chapter.

Notice that from what we have derive above, we can conclude that if we have a random variable X that is Normally distributed and we want to keep it's mean but we want to increase it's standard deviation by a constant c, then we just have to add to itself the quantity $c\cdot(X-\mu)/\sigma$.

5.7.6.9.2 Product of two random Normal variables

Let X,Y be two real independent random variables. We denote by f_X and f_Y the corresponding densities and we seek to determine the density of the variable $Z=XY$ (very important case, particularly in engineering).

Let f denote the density function of the pair (X,Y). Since X,Y are independent (see section Probabilities page 480):

$$f(x,y)=f_X(x)f_Y(y) \tag{5.7.687}$$

The distribution function of Z is:

$$F(Z)=P(Z\leq z)=P(XY\leq z)=\iint_D f(x,y)\mathrm{d}x\mathrm{d}y=\iint_D f_X(x)f_Y(y)\mathrm{d}x\mathrm{d}y \tag{5.7.688}$$

where $D=\{(x,y)|xy<z\}$.

D can be rewritten as a disjoint union (we do this for anticipating in the future change of variables a division by zero):

$$D=D_1\sqcup D_2\sqcup D_3 \tag{5.7.689}$$

with:

$$\begin{aligned}D_1&=\left\{(x,y)\in\mathbb{R}^2|xy\leq z\wedge x>0\right\}\\D_2&=\left\{(x,y)\in\mathbb{R}^2|xy\leq z\wedge x<0\right\}\\D_3&=\left\{(x,y)\in\mathbb{R}^2|xy\leq z\wedge x=0\right\}\end{aligned} \tag{5.7.690}$$

We have:

$$F(Z)=\iint_{D_1}f_X(x)f_Y(y)\mathrm{d}x\mathrm{d}y+\iint_{D_2}f_X(x)f_Y(y)\mathrm{d}x\mathrm{d}y+\underbrace{\iint_{D_3}f_X(x)f_Y(y)\mathrm{d}x\mathrm{d}y}_{=0} \tag{5.7.691}$$

The last integral is equal to zero because D_3 is of measurement (thickness) zero for the integral along x.

We then perform the following change of variable:

$$x = x \qquad t = xy \tag{5.7.692}$$

The Jacobian of the transformation (see section Differential and Differential Calculus page 1524) is:

$$\left| \begin{bmatrix} 1 & 0 \\ -t/x^2 & 1/x \end{bmatrix} \right| = \frac{1}{|x|} \tag{5.7.693}$$

Thus:

$$F(z) = \int_{-\infty}^{z} \int_{0}^{+\infty} \frac{f_X(x) f_Y(t/x)}{|x|} \mathrm{d}x\mathrm{d}t + \int_{-\infty}^{z} \int_{-\infty}^{0} \frac{f_X(x) f_Y(t/x)}{|x|} \mathrm{d}x\mathrm{d}t = \int_{-\infty}^{z} \int_{-\infty}^{+\infty} \frac{f_X(x) f_Y(t/x)}{|x|} \mathrm{d}x\mathrm{d}t \tag{5.7.694}$$

Let f_Z be the density of the variable Z. By definition:

$$F(Z \leq z) = F(z) = \int_{-\infty}^{z} f_Z(t) \mathrm{d}t \tag{5.7.695}$$

On the other hand:

$$F(z) = \int_{-\infty}^{z} \int_{-\infty}^{+\infty} \frac{f_X(x) f_Y(t/x)}{|x|} \mathrm{d}x\mathrm{d}t \tag{5.7.696}$$

as we have seen. Therefore:

$$\boxed{f_Z(t) = \int_{-\infty}^{+\infty} \frac{f_X(x) f_Y(t/x)}{|x|} \mathrm{d}x} \tag{5.7.697}$$

What is a bit sad is that in the case of a Gauss-Laplace distribution (Normal distribution), this integral can only be easily calculated numerically ... it is then necessary to use Monte Carlo integration type methods (see section Numerical Methods page 6717).

However according to some research done on the Internet, but without certainty, this integral may be calculated and give a new distribution named "Bessel distribution".

5.7.6.9.3 Bivariate Normal Distribution

If two Normal distributed random variables are independent, we know that their joint probability is equal to the product of their probabilities. So we have:

$$P=P_1P_2=\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}\frac{1}{\sqrt{2\pi}\sigma_2}e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}=\frac{1}{2\pi\sigma_1\sigma_2}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}-\frac{(x-\mu_2)^2}{2\sigma_2^2}} \tag{5.7.698}$$

Now comes an approach that we will often find in the follow developments: to generalize simple algebra models, you have to think in a Linear Algebra way! Therefore we are left with two vectors involving a scalar product:

$$P=P_1P_2=\frac{1}{2\pi\sigma_1\sigma_2}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}-\frac{(x-\mu_2)^2}{2\sigma_2^2}}=\frac{1}{2\pi\sigma_1\sigma_2}e^{-\frac{1}{2}\begin{pmatrix}x_1-\mu_1\\x_2-\mu_2\end{pmatrix}^T\begin{pmatrix}\frac{x_1-\mu_1}{2\sigma_1^2}\\\frac{x_2-\mu_2}{2\sigma_1 2^2}\end{pmatrix}} \tag{5.7.699}$$

But we can do even better because for the moment there is no added value to this notation! Effectively a subtle idea is to involve the determinant of a matrix (see section Linear Algebra page 1961) and the inverse of this same matrix in the previous relation:

$$\begin{aligned}P=P_1P_2&=\frac{1}{2\pi\sigma_1\sigma_2}e^{-\frac{1}{2}\begin{pmatrix}x_1-\mu_1\\x_2-\mu_2\end{pmatrix}^T\begin{pmatrix}\frac{x_1-\mu_1}{2\sigma_1^2}\\\dfrac{x_2-\mu_2}{2\sigma_1 2^2}\end{pmatrix}}\\&=\frac{1}{2\pi(\sigma_1^2\sigma_2^2-0\cdot 0)^{1/2}}e^{-\frac{1}{2}\begin{pmatrix}x_1-\mu_1\\x_2-\mu_2\end{pmatrix}^T\begin{bmatrix}\frac{1}{\sigma_1^2}&0\\0&\frac{1}{\sigma_2^2}\end{bmatrix}\begin{pmatrix}x_1-\mu_1\\x_2-\mu_2\end{pmatrix}}\\&=\frac{1}{2\pi\left|\begin{bmatrix}\sigma_1^2&0\\0&\sigma_2^2\end{bmatrix}\right|^{1/2}}e^{-\frac{1}{2}\begin{pmatrix}x_1-\mu_1\\x_2-\mu_2\end{pmatrix}^T\begin{bmatrix}\sigma_1^2&0\\0&\sigma_2^2\end{bmatrix}^{-1}\begin{pmatrix}x_1-\mu_1\\x_2-\mu_2\end{pmatrix}}\end{aligned} \tag{5.7.700}$$

We thus find a particular case of the variance-covariance matrix. In the field of the bivariate Normal distribution is it is customary to write this last relation in the following form:

$$\boxed{f(X_1,X_2)=\mathcal{N}(\vec{\mu},\Sigma)=\frac{1}{2\pi|\Sigma|^{1/2}}e^{-\frac{1}{2}\overrightarrow{(x-\mu)}^T\Sigma^{-1}\overrightarrow{(x-\mu)}}} \tag{5.7.701}$$

If we make a plot of this function we get:

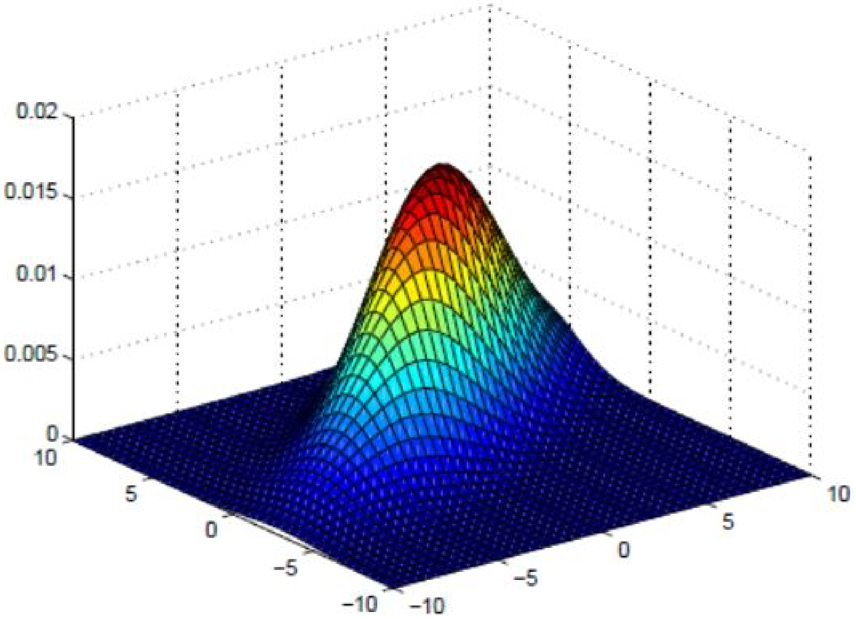
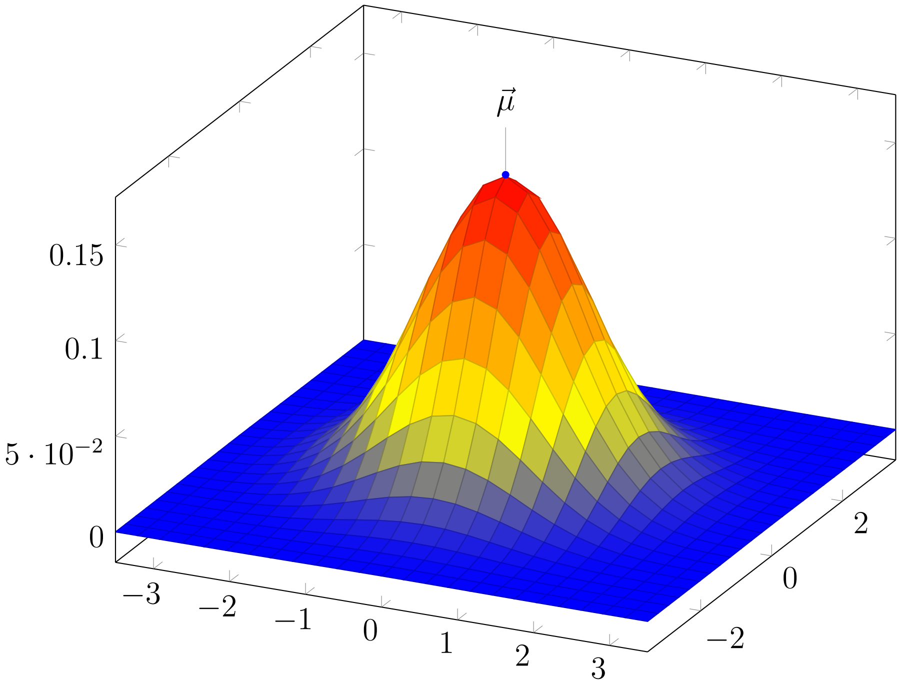


Figure 5.131: **Plot of the bivariate Normal function in MATLAB®**

or in high definition with TikZ:



or another one (not with the same values) with corresponding projections:

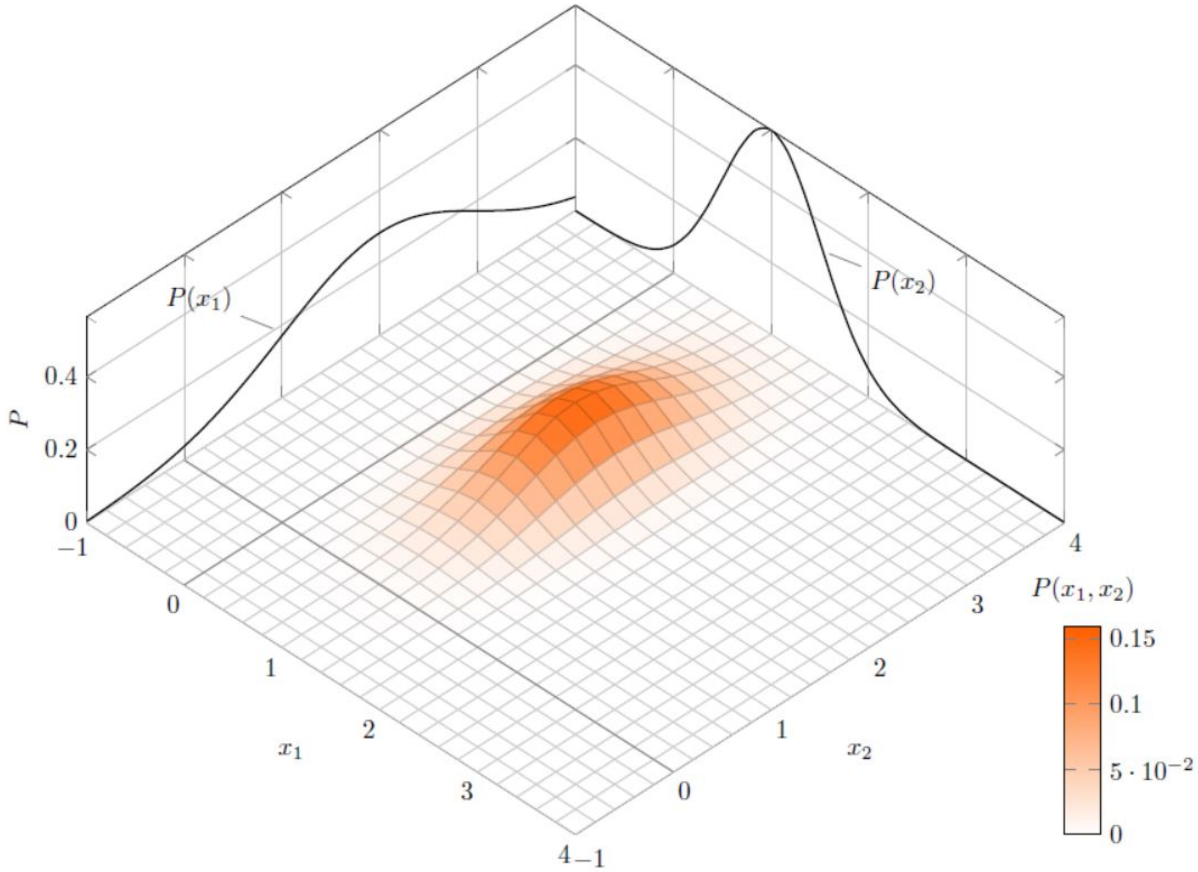


Figure 5.132: Plot of the bivariate Normal function with pgfplots

Now consider the important case in engineering, astronomy and quantum physics by returning to the following notation:

$$P = f(X_1, X_2) = \frac{1}{2\pi\sigma_1\sigma_2}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}-\frac{(x-\mu_2)^2}{2\sigma_2^2}} \tag{5.7.702}$$

and by focusing on to the iso-lines such that for any pair of values of the two random variables, we have:

$$\frac{1}{2\pi\sigma_1\sigma_2}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}-\frac{(x-\mu_2)^2}{2\sigma_2^2}} = c^{te} \tag{5.7.703}$$

By doing some very basic algebraic manipulations, we get:

$$-\frac{(x-\mu_1)^2}{2\sigma_1^2}-\frac{(x-\mu_2)^2}{2\sigma_2^2} = \ln\left(2\pi\sigma_1\sigma_2 c^{te}\right) \tag{5.7.704}$$

Thus:

$$\frac{(x-\mu_1)^2}{2\sigma_1^2}+\frac{(x-\mu_2)^2}{2\sigma_2^2} = \ln\left(\frac{1}{2\pi\sigma_1\sigma_2 c^{te}}\right) \tag{5.7.705}$$

and we get:

$$\frac{(x-\mu_1)^2}{2\sigma_1^2\ln\left(\frac{1}{2\pi\sigma_1\sigma_2 c^{te}}\right)}+\frac{(x-\mu_2)^2}{2\sigma_2^2\ln\left(\frac{1}{2\pi\sigma_1\sigma_2 c^{te}}\right)} = 1 \tag{5.7.706}$$

We recognize here the analytical equation of an ellipse (see section Analytical Geometric page 2768)!

A plot of iso-lines with $\mu=\begin{pmatrix}3\\2\end{pmatrix}, \Sigma=\begin{bmatrix}25 & 0\\0 & 9\end{bmatrix}$ give us:

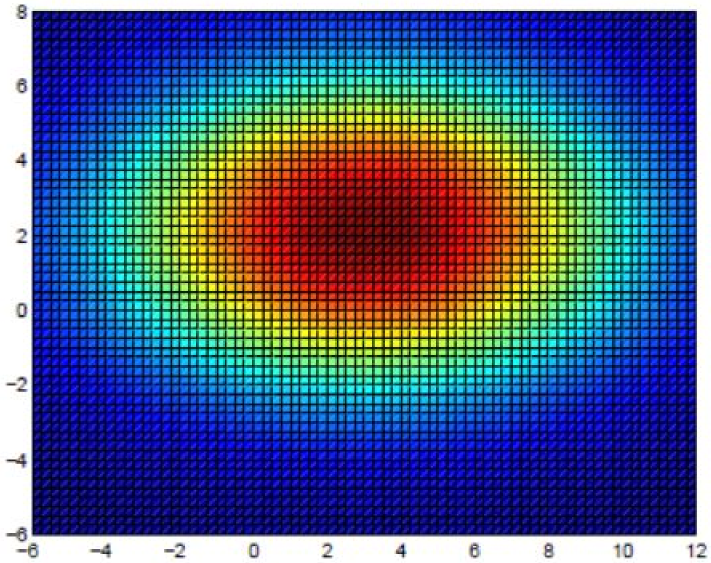


Figure 5.133: Plot of the iso-lines of the bivariate Normal function (non-correlated case)

But now recall that when we got:

$$f(X_1,X_2)=\frac{1}{2\pi|\Sigma|^{1/2}}e^{-\frac{1}{2}\overrightarrow{(x-\mu)}^T\Sigma^{-1}\overrightarrow{(x-\mu)}} \tag{5.7.707}$$

the variance-covariance matrix was zero everywhere except on the diagonal, implying verbatim the independence of the two random variables. We can obviously guess that the generalization is that the variance-covariance matrix is non-zero in the diagonal and then the two random variables are correlated. Consequently, the iso-lines become with values such as $\mu=\begin{pmatrix}3\\2\end{pmatrix}, \Sigma=\begin{bmatrix}10 & 5\\5 & 5\end{bmatrix}$:

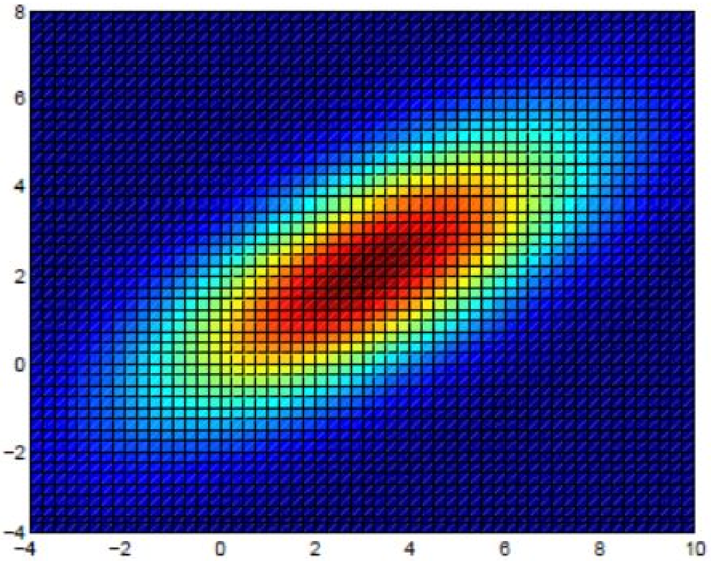


Figure 5.134: Plot of the iso-lines of the bivariate Normal function (correlated case)

So the correlation rotates the axis of the ellipses! Note that we have therefore:

$$\Sigma=\begin{bmatrix}\sigma_{11}^2 & \sigma_{12}\\ \sigma_{21} & \sigma_{22}^2\end{bmatrix}=\begin{bmatrix}\sigma_{11}^2 & \sigma_{12}\\ \sigma_{12} & \sigma_{22}^2\end{bmatrix}\Rightarrow\Sigma^{-1}=\frac{1}{\sigma_{11}^2\sigma_{22}^2-\sigma_{12}^2}\begin{bmatrix}\sigma_{11}^2 & -\sigma_{12}\\ \sigma_{21} & \sigma_{22}^2\end{bmatrix}\tag{5.7.708}$$

and thus verbatim:

$$|\Sigma|=\sigma_{11}^2\sigma_{22}^2-\sigma_{12}^2\tag{5.7.709}$$

Recall that we saw during our study of the correlation coefficient that (well... normally... the R notation for the correlation is used only if the variances are estimated but as it is the most common notation in practice we will still us it...):

$$\sigma_{12}=\text{cov}(X_1,X_2)=\text{c}(X_1,X_2)=R_{X_1,X_2}\sigma_{11}\sigma_{22}\tag{5.7.710}$$

Thus:

$$\Sigma^{-1}=\frac{1}{\sigma_{11}^2\sigma_{22}^2-\sigma_{12}^2}\begin{bmatrix}\sigma_{11}^2 & -\sigma_{12}\\ \sigma_{21} & \sigma_{22}^2\end{bmatrix}=\frac{1}{\sigma_{11}^2\sigma_{22}^2(1-R_{12}^2)}\begin{bmatrix}\sigma_{11}^2 & -\sigma_{12}\\ \sigma_{21} & \sigma_{22}^2\end{bmatrix}\tag{5.7.711}$$

and the exponent of the exponential of the bivariate Normal takes a form that we can find very often in the literature:

$$\begin{aligned}&\begin{pmatrix}x_1-\mu_1\\ x_2-\mu_2\end{pmatrix}^T\left(\frac{1}{\sigma_{11}^2\sigma_{22}^2(1-R_{12}^2)}\begin{bmatrix}\sigma_{11}^2 & -\sigma_{12}\\ \sigma_{21} & \sigma_{22}^2\end{bmatrix}\right)\begin{pmatrix}x_1-\mu_1\\ x_2-\mu_2\end{pmatrix}\\ &=\begin{pmatrix}x_1-\mu_1\\ x_2-\mu_2\end{pmatrix}^T\left(\frac{1}{\sigma_{11}^2\sigma_{22}^2(1-R_{12}^2)}\begin{bmatrix}\sigma_{11}^2 & -R_{12}\sigma_{11}\sigma_{22}\\ -R_{12}\sigma_{11}\sigma_{22} & \sigma_{22}^2\end{bmatrix}\right)\begin{pmatrix}x_1-\mu_1\\ x_2-\mu_2\end{pmatrix}\\ &=\frac{\sigma_{22}^2(x_1-\mu_1)^2+\sigma_{11}^2(x_2-\mu_2)^2-2R_{12}\sigma_{11}\sigma_{22}(x_1-\mu_1)(x_2-\mu_2)}{\sigma_{11}^2\sigma_{22}^2(1-R_{12}^2)}\\ &=\frac{1}{1-R_{12}^2}\left(\left(\frac{x_1-\mu_1}{\sigma_1 1}\right)^2+\left(\frac{x_2-\mu_2}{\sigma_2 2}\right)^2-2R_{12}\left(\frac{x_1-\mu_1}{\sigma_1 1}\right)\left(\frac{x_2-\mu_2}{\sigma_2 2}\right)\right)\end{aligned}\tag{5.7.712}$$

Note that if the random variables are centered reduced, then we have:

$$\Sigma^{-1}=\frac{1}{1-R_{12}^2} \tag{5.7.713}$$

and thus the exponent of the exponential of the bivariate Normal distribution becomes:

$$\Sigma^{-1}=\frac{1}{1-R_{12}^2}\left[x_1^2+x_2^2-2R_{12}x_1x_2\right] \tag{5.7.714}$$

Thus, the density function of the bivariate Normal centered reduced distribution will be written:

$$\begin{aligned}
f(x_1,x_2,R)&=\frac{1}{2\pi(1-R_{12}^2)}e^{-\frac{x_1^2+x_2^2-2R_{12}x_1x_2}{2(1-R_{12}^2)}}\\
&=\frac{1}{\sqrt{2\pi(1-R_{12}^2)}}e^{-x_1^2}\frac{1}{\sqrt{2\pi(1-R_{12}^2)}}e^{-x_2^2}e^{\frac{R_{11}x_1x_2}{1-R_{12}^2}}\\
&=\mathcal{N}(0,1-R_{12}^2)\mathcal{N}(0,1-R_{12}^2)e^{\frac{R_{11}x_1x_2}{1-R_{12}^2}}\\
&=\mathcal{N}(0,1)\mathcal{N}(0,1)\left(1-R_{12}^2\right)^2e^{\frac{R_{11}x_1x_2}{1-R_{12}^2}}
\end{aligned} \tag{5.7.715}$$

Thus, we can see that a bivariate Normal reduced centered distribution function normal can be constructed by the multiplication of two Normal centered and reduced distributions themselves multiplied by a term that depends mainly on the correlation parameter. The latter term includes the nature of the dependence of the two random variables and provides the link between the marginal distributions (both Normal centered and reduced) to obtain the joint bivariate Normal distribution.

If necessary (this can be very useful in practice), here is the Maple™ 4.00b code to plot a bivariate Normal function (taking the last example) even if it is also quite simple to do with a spreadsheet software like Microsoft® Excel®:

```
>f:=(x,y,rho,mu1,mu2,sigma1,sigma2)->(1/(2*Pi*sqrt(sigma1*sigma2*(1-rho^2))))
*exp((-1/(2*(1-rho^2)))*(((x-mu1)/sqrt(sigma1))^2+((y-mu2)/sqrt(sigma2))^2
-2*rho*((x-mu1)/sqrt(sigma1))*((y-mu2)/sqrt(sigma2))));


>plot3d(f(x,y,5/sqrt(10*5),3,2,10,5),x=-4..10,y=-4..9,grid=[40,40]);
```

and for the plot with the iso-lines:

```
>with(plots):
>contourplot(f(x,y,5/sqrt(10*5),3,2,10,5),x=-4..10,y=-4..9,grid=[40,40]);
```

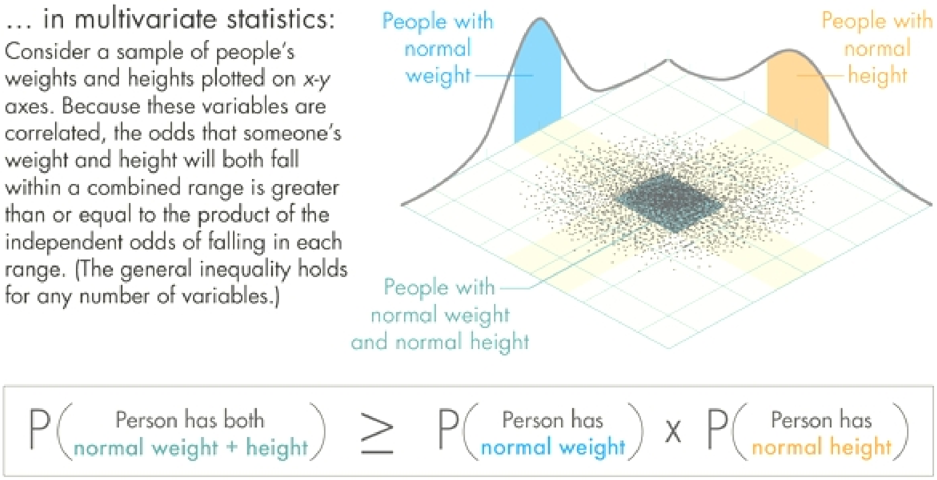
and we can check that it is a probability density function by writing:

```
>int(int(f(x,y,5/sqrt(10*5),3,2,10,5),x=-infinity...+infinity)
,y=-infinity...+infinity);
```

or calculate the cumulative probability between two intervals:

```
>evalf(int(int(f(x,y,5/sqrt(10*5),3,2,10,5),x=-3...+4),y=-5...+2));
```

Notice that it is only in 12014 (holocene calendar) that the "Gaussian correlation conjecture" (also named "Gaussian correlation inequality") was proven by Thomas Royen, a retired German professor of statistics. This conjecture can be summarized with the following figure:



... in multivariate statistics:

Consider a sample of people's weights and heights plotted on x-y axes. Because these variables are correlated, the odds that someone's weight and height will both fall within a combined range is greater than or equal to the product of the independent odds of falling in each range. (The general inequality holds for any number of variables.)

People with normal weight
People with normal height
People with normal weight and normal height
P(Person has both normal weight + height) ≥ P(Person has normal weight) x P(Person has normal height)

We will not provide the proof here as I never needed it for my business or also my customers business but it is quite funny that to know that an inequality that seems quite obvious needed so many years before someone found the proof.

Remark

On July 17, 12014 (holocene calendar), a few years after his retirement, when brushing his teeth, Royen had a flash of insight: how to use the Laplace transform of the multivariate gamma distribution to achieve a relatively simple proof for the Gaussian correlation inequality, a conjecture on the intersection of geometry, probability theory and statistics, formulated after work by Dunnett and Sobel (11955 according to holocene calendar) and the American statistician Olive Jean Dunn (11958 according to holocene calendar) that had remained unsolved since then. He sent a copy of his proof to Donald Richards, an acquainted American mathematician, who worked on a proof of the GCI for 30 years. Richards immediately saw the validity of Royen's proof and subsequently helped him to transform the mathematical formulas into LaTeX (hence the importance to lean LaTeX when we are engineers or scientists!). When Royen contacted other reputed mathematicians, though, they didn't bother to investigate his proof, because Royen was relatively unknown, and these mathematicians therefore estimated the chance that Royen's proof would be false as very high (especially also by sending a Microsoft Word file for peer review your chances to be read are quite low).

5.7.6.9.4 Multivariate Normal distribution

The bivariate Normal distribution is obviously a special case of the multivariate one. We will see now some important properties of the multivariate Normal distribution that we will need later for various statistical tests and methods.

Let us consider the following multivariate Normal distribution $P(\vec{X}) = \mathcal{N}(\vec{\mu}_X, \Sigma_X)$. We want to prove the following multivariate affine transformation that will be useful to us for the study of Factor Analysis later:

$$P(\vec{Y}) = P(A\vec{X} + \vec{b}) = \mathcal{N}(\vec{\mu}_Y, \Sigma_Y) \tag{5.7.716}$$

where:

$$\vec{\mu}_Y = A\vec{\mu}_X + \vec{b} \qquad \Sigma_y = A\Sigma_Y A^T \tag{5.7.717}$$

Proof 5.47.3. Let us see first for the mean:

$$\vec{\mu}_X = \mathrm{E}(A\vec{X} + \vec{b}) = A\mathrm{E}(\vec{X}) + \vec{b} = A\vec{\mu}_X + \vec{b} \tag{5.7.718}$$

and now for the variance-covariance matrix:

$$\begin{aligned}
\Sigma_Y &= \mathrm{E}\left[(\vec{Y} - \vec{\mu}_Y)(\vec{Y} - \vec{\mu}_Y)^T\right] \\
&= \mathrm{E}\left\{\left[(A\vec{X} + \vec{b}) - (A\vec{\mu}_X + \vec{b})\right]\left[(A\vec{X} + \vec{b}) - (A\vec{\mu}_X + \vec{b})\right]^T\right\} \\
&= \mathrm{E}\left[(A\vec{X} - A\vec{\mu}_X)(A\vec{X} - A\vec{\mu}_X)^T\right] \\
&= A\mathrm{E}\left[(\vec{X} - \vec{\mu}_X)(\vec{X} - \vec{\mu}_X)^T\right] \\
&= A\Sigma_X A^T
\end{aligned} \tag{5.7.719}$$

■ Q.E.D.

Notice some special cases:

- If $A = \mathbb{1}$, then:

$$P(\vec{X} + \vec{b}) = \mathcal{N}(\vec{\mu}_X + \vec{b}, \Sigma_X) \tag{5.7.720}$$

- If $\vec{b} = \vec{0}$, then:

$$P(A\vec{X}) = \mathcal{N}(\vec{\mu}_X, A\Sigma_X A^T) \tag{5.7.721}$$

- If $P(\vec{X}) = \mathcal{N}\left([\mu_1, \ldots, \mu_n]^T, \mathrm{diag}[\sigma_1^2, \ldots, \sigma_2^2]\right)$, then:

$$P\left(\sum_{i=1}^{n} x_i\right) = \mathcal{N}\left(\sum_{i}^{n} \mu_i, \sum_{i}^{n} \sigma_i^2\right) \tag{5.7.722}$$

Now let us move to another property of multivariate Normal distribution.

For this, let us recall first that the product of two univariate gaussian probability density functions is given by:

$$\begin{aligned}P(x)P(y)&=\mathcal{N}(\mu_x,\sigma_x)\mathcal{N}(\mu_y,\sigma_y)\\&=\frac{1}{\sqrt{2\pi\sigma_x}}e^{-\frac{(x-\mu_x)^2}{\sigma_x^2}}\frac{1}{\sqrt{2\pi\sigma_y}}e^{-\frac{(y-\mu_y)^2}{\sigma_y^2}}=\frac{1}{2\pi\sigma_x\sigma_y}e^{-\left(\frac{(x-\mu_x)^2}{2}+\frac{(y-\mu_y)^2}{2}\right)}\end{aligned} \tag{5.7.723}$$

Remark

Do not confuse the product of the Normal distributed variables, and the product of their respective probability density functions!!!!

Indeed, the product of the probability density functions of two random variables X and Y will give the joint distribution of the vector-valued random variable (X, Y) in the case that X and Y are independent. Therefore, if X and Y are normally distributed independent random variables, the product of their respective probability density function is bivariate normal with zero correlation (but the result isn't technically speaking a probability density function!).

On the other hand, even in the case that X and Y are independent and identically standard Normal random variables, their product is not itself Normal (it's easy to check in the case where we put $Y = X$ as X^2 cannot be negative, hence it's not Normally distributed!).

Now let us consider for further needs the following multivariate gaussian variable notation:

$$\begin{aligned}P_i(\vec{x})&=\mathcal{N}\left(\vec{x},\vec{\mu}_i,\Sigma_i\right)=\frac{1}{(2\pi)^{n/2}\left|\Sigma_i\right|^{1/2}}\exp\left[-\frac{1}{2}\left(\vec{x}-\vec{\mu}_i\right)^T\Sigma_i^{-1}\left(\vec{x}-\vec{\mu}_i\right)\right]\\&=\frac{1}{(2\pi)^{n/2}\left|\Sigma_i\right|^{1/2}}\exp\left[-\frac{1}{2}\left(\vec{x}^T\Sigma_i^{-1}\vec{x}-2\vec{x}^T\Sigma_i^{-1}\vec{\mu}_i+\vec{\mu}_i\Sigma_i^{-1}\vec{\mu}_i\right)\right]\\&=\exp\left[-\frac{1}{2}\left[n\ln(2\pi)+\ln\left|\Sigma_i\right|+\vec{\mu}_i\Sigma_i^{-1}\vec{\mu}_i\right]\right]\exp\left[-\frac{1}{2}\left(\vec{x}^T\Sigma_i^{-1}\vec{x}-2\vec{x}^T\Sigma_i^{-1}\vec{\mu}_i\right)\right]\\&=\exp\left[-\frac{1}{2}\left[\vec{x}^T\Sigma_i^{-1}\vec{x}-2\vec{x}^T\Sigma_i^{-1}\vec{\mu}_i+c_i\right]\right]\end{aligned} \tag{5.7.724}$$

where:

$$c_i=n\ln(2\pi)+\ln\left|\Sigma_i\right|+\vec{\mu}_i\Sigma_i^{-1}\vec{\mu}_i \tag{5.7.725}$$

Now for the need of our study of Factor Analysis, let us consider a same random variable $\vec{x}$ inject into two multivariate Normal distribution with different mean and variance-covariance matrix:

$$\begin{aligned}P_1(\vec{x})P_2(\vec{x}) &= \mathcal{N}\left(\vec{x},\vec{\mu}_1,\Sigma_1\right)\mathcal{N}\left(\vec{x},\vec{\mu}_2,\Sigma_2\right)\\ &= \exp\left[-\frac{1}{2}\left[\vec{x}^T\Sigma_1^{-1}\vec{x}-2\vec{x}^T\Sigma_1^{-1}\vec{\mu}_1+c_1+\vec{x}^T\Sigma_2^{-1}\vec{x}-2\vec{x}^T\Sigma_2^{-1}\vec{\mu}_2+c_2\right]\right]\\ &= \exp\left[-\frac{1}{2}\left[\vec{x}^T\left(\Sigma_1^{-1}+\Sigma_2^{-1}\right)\vec{x}-2\vec{x}^T\left(\Sigma_1^{-1}\vec{\mu}_1+\Sigma_2^{-1}\vec{\mu}_2\right)+c_1+c_2\right]\right]\\ &= \exp\left[-\frac{1}{2}\left[\vec{x}^T\Sigma^{-1}\vec{x}-2\vec{x}^T\Sigma^{-1}\vec{\mu}+c+(c_1+c_2-c)\right]\right]\\ &= \mathcal{N}(\vec{x},\vec{\mu},\Sigma)\exp\left(c_1+c_2-c\right)\propto\mathcal{N}(\vec{x},\vec{\mu},\Sigma)\end{aligned} \tag{5.7.726}$$

Here we have defined:

$$\begin{aligned}c &= n\ln(2\pi)+|\Sigma|+\vec{\mu}^T\Sigma^{-1}\vec{\mu}\\ \Sigma &= \left(\Sigma_1^{-1}+\Sigma_2^{-1}\right)^{-1}\end{aligned} \tag{5.7.727}$$

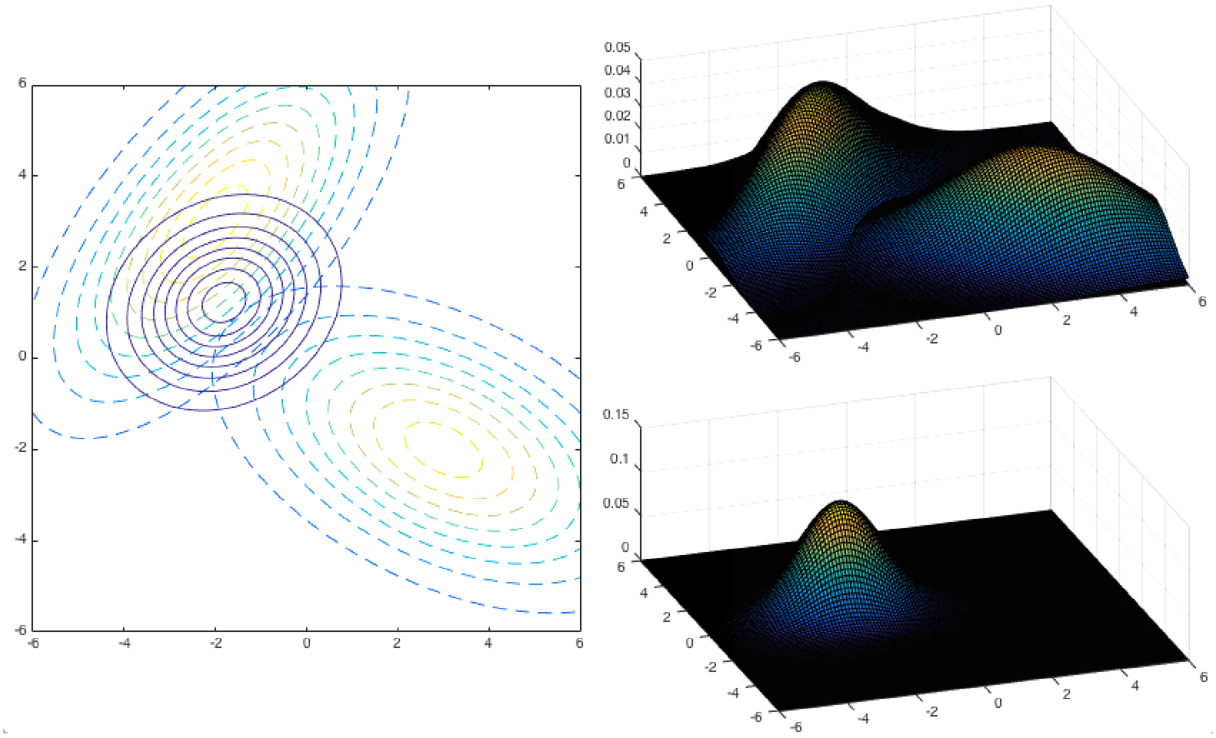
and:

$$\Sigma^{-1}\vec{\mu}=\Sigma_1^{-1}\vec{\mu}_1+\Sigma_2^{-1}\vec{\mu}_2 \tag{5.7.728}$$

i.e.:

$$\vec{\mu}=\Sigma\left(\Sigma_1^{-1}\vec{\mu}_1+\Sigma_2^{-1}\vec{\mu}_2\right)=\left(\Sigma_1^{-1}+\Sigma_2^{-1}\right)^{-1}\left(\Sigma_1^{-1}\vec{\mu}_1+\Sigma_2^{-1}\vec{\mu}_2\right) \tag{5.7.729}$$

We see then that the product of two multivariate gaussians with same random vector is also a scaled Gaussian:



5.7.6.9.4.1 Marginal and conditional distributions of multivariate Normal distributions

For our study later of Factor Analysis we will also need what follows!

A random vector $\vec{x}=[\vec{x}_1^T,\vec{x}_2^T]^T$ with a Normal distribution $\mathcal{N}(\vec{x},\vec{\mu},\Sigma)$ can be partitioned into two subvectors $\vec{x}_1$ and $\vec{x}_2$:

$$P\left(\left[\begin{array}{c}\vec{x}_1\\ \vec{x}_2\end{array}\right]\right)=P(\vec{x})=\mathcal{N}(\vec{x},\vec{\mu},\Sigma)=\mathcal{N}\left(\left[\begin{array}{c}\vec{x}_1\\ \vec{x}_2\end{array}\right],\left[\begin{array}{c}\vec{\mu}_1\\ \vec{\mu}_2\end{array}\right],\left[\begin{array}{cc}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{array}\right]\right) \tag{5.7.730}$$

where:

$$\Sigma_{ij}=\Sigma_{ji}^T,\Sigma_{ii}=\Sigma_{ii}^T \qquad i,j=1,2,i\neq j \tag{5.7.731}$$

Given this Normal joint distribution of $\vec{x}_1$ and $\vec{x}_2$, we can find the Normal conditional distribution of $\vec{x}_i$ given $\vec{x}_j$:

$$P\left(\vec{x}_i|\vec{x}_j\right)=\mathcal{N}\left(\vec{x}_i,\vec{\mu}_{i|j},\Sigma_{i|j}\right) \tag{5.7.732}$$

in terms of the mean vector and the covariance matrix:

$$\begin{cases}\vec{\mu}_{i|j}=\vec{\mu}_i+\Sigma_{ij}\Sigma_{jj}^{-1}\left(\vec{x}_j-\vec{m}_j\right)\\ \Sigma_{i|j}=\Sigma_{ii}-\Sigma_{ij}\Sigma_{jj}^{-1}\Sigma_{ji}\end{cases} \tag{5.7.733}$$

Proof 5.47.4. The joint distribution of $\vec{x}$ is defined as:

$$\begin{aligned}P(\vec{x})=P\left(\vec{x}_1,\vec{x}_2\right)&=\dfrac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp\left[-\dfrac{1}{2}(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})\right]\\ &=\dfrac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp\left[-\dfrac{1}{2}Q\left(\vec{x}_1,\vec{x}_2\right)\right]\end{aligned} \tag{5.7.734}$$

Here we have defined:

$$\begin{aligned}Q\left(\vec{x}_1,\vec{x}_2\right)&=(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})\\ &=\left[\left(\vec{x}_1-\vec{\mu}_1\right)^T,\left(\vec{x}-\vec{\mu}_2\right)^T\right]\left[\begin{array}{cc}\Sigma^{11} & \Sigma^{12}\\ \Sigma^{21} & \Sigma^{12}\end{array}\right]\left[\begin{array}{c}\vec{x}_1-\vec{\mu}_1\\ \vec{x}_2-\vec{\mu}_2\end{array}\right]\\ &=\left(\vec{x}_1-\vec{\mu}_1\right)^T\Sigma^{11}\left(\vec{x}_1-\vec{m}_1\right)+2\left(\vec{x}_1-\vec{m}_1\right)^T\Sigma^{12}\left(\vec{x}_2-\vec{m}_2\right)+\left(\vec{x}_2-\vec{m}_2\right)^T\Sigma^{22}\left(\vec{x}_2-\vec{m}_2\right)\end{aligned} \tag{5.7.735}$$

where we put:

$$\Sigma^{-1}=\left[\begin{array}{cc}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{array}\right]^{-1}=\left[\begin{array}{cc}\Sigma^{11} & \Sigma^{12}\\ \Sigma^{21} & \Sigma^{22}\end{array}\right] \tag{5.7.736}$$

According to the property of the inverse of a partitioned symmetric matrix (see section Linear Algebra page 1989), we have:

$$\begin{aligned}\Sigma^{11}&=\left(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)^{-1}=\Sigma_{11}^{-1}+\Sigma_{11}^{-1}\Sigma_{12}\left(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)^{-1}\Sigma_{21}\Sigma_{11}^{-1}\\ \Sigma^{22}&=\left(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)^{-1}=\Sigma_{22}^{-1}+\Sigma_{22}^{-1}\Sigma_{21}\left(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)^{-1}\Sigma_{12}\Sigma_{22}^{-1}\\ \Sigma^{12}&=-\Sigma_{11}^{-1}\Sigma_{12}\left(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)^{-1}=\left(\Sigma^{21}\right)^T\end{aligned} \tag{5.7.737}$$

Substituting the second expression for Σ^{11}, the first expression for Σ^{22}, and Σ^{12} into $Q(\vec{x}_1,\vec{x}_2)$ we get:

$$\begin{aligned}
Q\left(\vec{x}_1,\vec{x}_2\right) = {} & \left(\vec{x}_1-\vec{\mu}_1\right)^T\left[\Sigma_{11}^{-1}+\Sigma_{11}^{-1}\Sigma_{12}\left(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)^{-1}\Sigma_{21}\Sigma_{11}^{-1}\right]\left(\vec{x}_1-\vec{\mu}_1\right)\\
& -2\left(\vec{x}_1-\vec{\mu}_1\right)^T\left[\Sigma_{11}^{-1}\Sigma_{12}\left(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)^{-1}\right]\left(\vec{x}_2-\vec{\mu}_2\right)\\
& +\left(\vec{x}_2-\vec{\mu}_2\right)^T\left[\left(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)^{-1}\right]\left(\vec{x}_2-\vec{\mu}_2\right)\\
& =\left(\vec{x}_1-\vec{\mu}_1\right)^T\Sigma_{11}^{-1}\left(\vec{x}_1-\vec{\mu}_1\right)\\
& +\left(\vec{x}_1-\vec{\mu}_1\right)^T\Sigma_{11}^{-1}\Sigma_{12}\left(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)^{-1}\Sigma_{21}\Sigma_{11}^{-1}\left(\vec{x}_1-\vec{\mu}_1\right)\\
& -2\left(\vec{x}_1-\vec{\mu}_1\right)^T\Sigma_{11}^{-1}\Sigma_{12}\left(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)^{-1}\left(\vec{x}_2-\vec{\mu}_2\right)\\
& +\left(\vec{x}_2-\vec{\mu}_2\right)^T\left(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)^{-1}\left(\vec{x}_2-\vec{\mu}_2\right)\\
& =\left(\vec{x}_1-\vec{\mu}_1\right)^T\Sigma_{11}^{-1}\left(\vec{x}_1-\vec{\mu}_1\right)\\
& +\left[\left(\vec{x}_2-\vec{\mu}_2\right)-\Sigma_{21}\Sigma_{11}^{-1}\left(\vec{x}_1-\vec{\mu}_1\right)\right]^T\left(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)^{-1}\left[\left(\vec{x}_2-\vec{\mu}_2\right)-\Sigma_{21}\Sigma_{11}^{-1}\left(\vec{x}_1-\vec{\mu}_1\right)\right]
\end{aligned}\tag{5.7.738}$$

The last equal sign is due to the following equations for any vectors $\vec{u}$ and $\vec{v}$ and a symmetric matrix $A=A^T$:

$$\begin{aligned}
\vec{u}^TA\vec{u}-2\vec{u}^TA\vec{v}+\vec{v}^TA\vec{v} & =\vec{u}^TA\vec{u}-\vec{u}^TA\vec{v}-\vec{u}^TA\vec{v}+\vec{v}^TA\vec{v}\\
& =\vec{u}^TA(\vec{u}-\vec{v})-(\vec{u}-\vec{v})^TA\vec{v}=\vec{u}^TA(\vec{u}-\vec{v})-\vec{v}^TA(\vec{u}-\vec{v})\\
& =(\vec{u}-\vec{v})^TA(\vec{u}-\vec{v})=(\vec{v}-\vec{u})^TA(\vec{v}-\vec{u})
\end{aligned}\tag{5.7.739}$$

To simplify the notations, let us now put:

$$\left\{\begin{array}{l}
\vec{b}=\vec{\mu}_2+\Sigma_{21}\Sigma_{11}^{-1}\left(\vec{x}_1-\vec{\mu}_1\right)\\
A=\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}
\end{array}\right.\tag{5.7.740}$$

and also:

$$\left\{\begin{aligned}
Q_1\left(\vec{x}_1\right) & =\left(\vec{x}_1-\vec{\mu}_1\right)^T\Sigma_{11}^{-1}\left(\vec{x}_1-\vec{\mu}_1\right)\\
Q_2\left(\vec{x}_1,\vec{x}_2\right) & =\left[\left(\vec{x}_2-\vec{\mu}_2\right)-\Sigma_{21}\Sigma_{11}^{-1}\left(\vec{x}_1-\vec{\mu}_1\right)\right]^T\left(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)^{-1}\left[\left(\vec{x}_2-\vec{\mu}_2\right)-\Sigma_{21}\Sigma_{11}^{-1}\left(\vec{x}_1-\vec{\mu}_1\right)\right]\\
& =\left(\vec{x}_2-\vec{b}\right)^TA^{-1}\left(\vec{x}_2-\vec{b}\right)
\end{aligned}\right.\tag{5.7.741}$$

Then the expression above for $Q(\vec{x}_1,\vec{x}_2)$ can be written as:

$$Q\left(\vec{x}_1,\vec{x}_2\right)=Q_1\left(\vec{x}_1\right)+Q_2\left(\vec{x}_1,\vec{x}_2\right)\tag{5.7.742}$$

and the joint distribution can be written as:

$$\begin{aligned}
P(\vec{x}) & =P\left(\vec{x}_1,\vec{x}_2\right)=\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}Q\left(\vec{x}_1,\vec{x}_2\right)\right]\\
& =\frac{1}{(2\pi)^{n/2}\left|\Sigma_{11}\right|^{1/2}\left|\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right|^{1/2}}\exp\left[-\frac{1}{2}Q_1\left(\vec{x}_1\right)\right]\exp\left[-\frac{1}{2}Q_2\left(\vec{x}_1,\vec{x}_2\right)\right]\\
& =\frac{1}{(2\pi)^{p/2}\left|\Sigma_{11}\right|^{1/2}}\exp\left[-\frac{1}{2}\left(\vec{x}_1-\vec{\mu}_1\right)^T\Sigma_{11}^{-1}\left(\vec{x}_1-\vec{\mu}_1\right)\right]\frac{1}{(2\pi)^{q/2}|A|^{1/2}}\exp\left[-\frac{1}{2}\left(\vec{x}_2-\vec{b}\right)^TA^{-1}\left(\vec{x}_2-\vec{b}\right)\right]\\
& =\mathcal{N}\left(\vec{x}_1,\vec{\mu}_1,\Sigma_{11}\right)\mathcal{N}\left(\vec{x}_2,\vec{b},A\right)
\end{aligned}\tag{5.7.743}$$

where we have used the expression of the determinant of a partitioned matrix (see section Linear Algebra page 1990):

$$|\Sigma|=|\Sigma_{11}||\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}| \tag{5.7.744}$$

The marginal distribution of $\vec{x}_1$ is then:

$$\begin{aligned}P_1(\vec{x}_1)&=\int P(\vec{x}_1,\vec{x}_2)\mathrm{d}x_2=\mathcal{N}(\vec{x}_1,\vec{\mu}_1,\Sigma_{11})\int\mathcal{N}(\vec{x}_2,\vec{b},A)\mathrm{d}x_2=\mathcal{N}(\vec{x}_1,\vec{\mu}_1,\Sigma_{11})\cdot 1\\&=\mathcal{N}(\vec{x}_1,\vec{\mu}_1,\Sigma_{11})=\frac{1}{(2\pi)^{p/2}|\Sigma_{11}|^{1/2}}\exp\left[-\frac{1}{2}(\vec{x}_1-\vec{\mu}_1)^T\Sigma_{11}^{-1}(\vec{x}_1-\vec{\mu}_1)\right]\end{aligned} \tag{5.7.745}$$

and the conditional distribution of $\vec{x}_2$ given $\vec{x}_1$ is then:

$$P_{2|1}=\frac{P(\vec{x}_1,\vec{x}_2)}{P_1(\vec{x}_1)}=\mathcal{N}(\vec{x}_2,\vec{b},A)=\frac{1}{(2\pi)^{q/2}|A|^{1/2}}\exp\left[-\frac{1}{2}\left(\vec{x}_2-\vec{b}\right)^T A^{-1}\left(\vec{x}_2-\vec{b}\right)\right] \tag{5.7.746}$$

with for recall:

$$\begin{cases}\vec{b}=\vec{\mu}_{2|1}=\vec{\mu}_2+\Sigma_{21}\Sigma_{11}^{-1}(\vec{x}_1-\vec{\mu}_1)\\ A=\Sigma_{2|1}=\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\end{cases} \tag{5.7.747}$$

■ Q.E.D.

Then we can deduce two famous results for the special bivariate case (using $\sigma_{12}=\text{cov}(X_1,X_2)=R_{X_1,X_2}\sigma_{11}\sigma_{22}$), i.e. (with the traditional notation in the field of statistics):

$$\begin{cases}\mathrm{E}(Y|X)=\vec{\mu}_{2|1}=\vec{\mu}_2+\Sigma_{21}\Sigma_{11}^{-1}(\vec{x}_1-\vec{\mu}_1)=\mu_Y+\rho\dfrac{\sigma_Y}{\sigma_X}(x-\mu_X)\\ \mathrm{V}(Y|X)=\Sigma_{2|1}=\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}=\sigma_Y^2-\sigma_Y^2\rho^2=\sigma_Y^2(1-\rho^2)\end{cases} \tag{5.7.748}$$

5.7.6.9.5 Normal Reduced Centered Distribution

The Gauss-Laplace distribution is not tabulated as we must then have so many numerical tables as possible values for the mean μ and standard deviation σ (which are the parameters of the function as we have seen it).

Therefore, by a change of variable, the Normal distribution becomes the Normal reduced centered distribution more often named the "standard Normal distribution" where:

1. "Centered" refers to subtracting the mean μ to the measures (thus the distribution function is symmetric to the vertical axis).
2. "Reduced" refers to the division by the standard deviation σ (thus the distribution function has a unit variance).

By this change of variable, the variable k is replaced by the reduced centered random variable:

$$k^{*}=\frac{k-\mu}{\sigma} \tag{5.7.749}$$

If the variable k has for mean μ and standard deviation σ then the variable k^{*} has a mean of 0 and standard deviation of 1 (this last variable is usually denoted by the letter Z).

Thus the relation:

$$P(k,\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(k-\mu)^2}{2\sigma^2}} \tag{5.7.750}$$

is therefore written (trivially) more simply:

$$\boxed{P(k^{*},0,1)=\mathcal{N}(0,1)=\frac{1}{\sqrt{2\pi}}e^{-\frac{k^{*2}}{2}}} \tag{5.7.751}$$

which is just the explicit expression of the reduced centered Normal distribution ("standard Normal") often denoted $\mathcal{N}(0,1)$ which we will find very often in the sections of physics, finance, quantitative management and engineering!

The derivation of the inverse of the standard normal PDF is simple (it is useful to know it as not implemented by default in all statistical softwares):

$$\begin{aligned}
y&=\frac{1}{\sqrt{2\pi}}e^{\frac{-1}{2}x^2}\\
\ln(y)&=\log\left(\frac{1}{\sqrt{2\pi}}\right)-\frac{1}{2}x^2\\
x^2&=-2\left(\ln(y)-\log\left(\frac{1}{\sqrt{2\pi}}\right)\right)\\
x&=\sqrt{-2\ln(y\sqrt{2\pi})}
\end{aligned} \tag{5.7.752}$$

Remark

Calculate the integral of the previous relation for an interval can not be done accurately formally speaking. One possible and simple idea is then to express the exponential in a Taylor series and then be integrated term by term of the series (making sure to take enough terms for convergence!).

5.7.6.9.6 Henry's Line

Often in business it is the Gauss-Laplace (Normal) distribution that is analysed but common and easily accessible software like Microsoft® Excel® are unable to verify that the measured data follow a Normal distribution when we do the frequency analysis (there are no default integrated tool allowing users to check this assumption) and we do not have the original ungrouped data.

The trick then is then to use the reduced centered variable that is build as we have seen above with the following relation:

$$k^* = \frac{k-\mu}{\sigma} \tag{5.7.753}$$

The idea of the Henry's Line is then to use the linear relation between k and k^* given by the equation of the line:

$$k^* = f(k) = \frac{k}{\sigma} - \frac{\mu}{\sigma} \tag{5.7.754}$$

and which can be plotted to determine the mean and the standard deviation of the Normal distribution.

✍**Example:**

Suppose we have the following frequencies analysis of $10,000$ receipts in a supermarket:

Price of receipts	**Number of receipts**	**Cumulated number of receipts**	**Relative frequencies of receipts**
$[0,50[$	668	668	0.068
$[50,100[$	919	$1,587$	0.1587
$[100,150[$	$1,498$	$3,085$	0.3085
$[150,200[$	$1,915$	$5,000$	0.5000
$[200,250[$	$1,915$	$6,915$	0.6915
$[250,300[$	$1,498$	$8,413$	0.8413
$[300,350[$	919	$9,332$	0.9332
$[350,400[$	440	$9,772$	0.9772
$[400$ and $+$	228	$10,000$	1

Table 5.42: **Supermarket receipt amount distribution**

If we now plot this in Microsoft® Excel® 11.8346 we get:

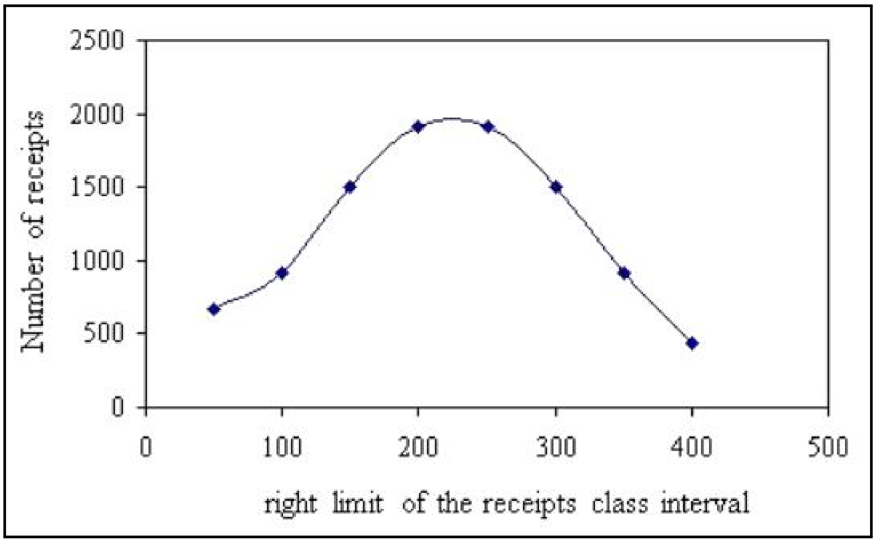


Figure 5.135: **Distribution of receipts amount**

What looks terribly like a Normal distribution, thus the authorization, without too much risk to use in this example the technique of Henry's line.

But what can we do now? Well... now that we know the cumulative frequency, it remains for us to calculate each k^* using numerical tables or the `NORMSINV( )` function of Microsoft® Excel® 11.8346 (remember that formal integration of the Gaussian function is not easy...).

This will give us the values of the standard Normal distribution $\mathcal{N}(0,1)$ of these respective cumulative frequencies (cumulative distribution function). So we get (we leave to the reader to take its statistic table or open its favorite software...):

Upper limit of the interval	**Cumulated relative frequencies**	**Correspondence for k^* of $\mathcal{N}(0,1)$**
50	0.068	-1.5
100	0.1587	-1
150	0.3085	-0.5
200	0.5000	0
250	0.6915	0.5
300	0.8413	1
350	0.9332	1.5
400	0.9772	2
-	1	-

Table 5.43: **Cumulative relative frequencies to the Henry's line**

Note that in the type of table above, in Microsoft® Excel®, the null and unit cumulative frequencies will generated some errors. You should then play a little bit...
As we specified earlier, we have under discrete form:

$$k^*=f(k_i)=\frac{k_i}{\sigma}-\frac{\mu}{\sigma} \quad (5.7.755)$$

So graphically in Microsoft® Excel® 11.8346 we can thanks to our table plot the following chart (obviously we could do strictly a linear regression in the rules of art as seen in the section of Numerical Methods with confidence, prediction intervals and other stuffs...):

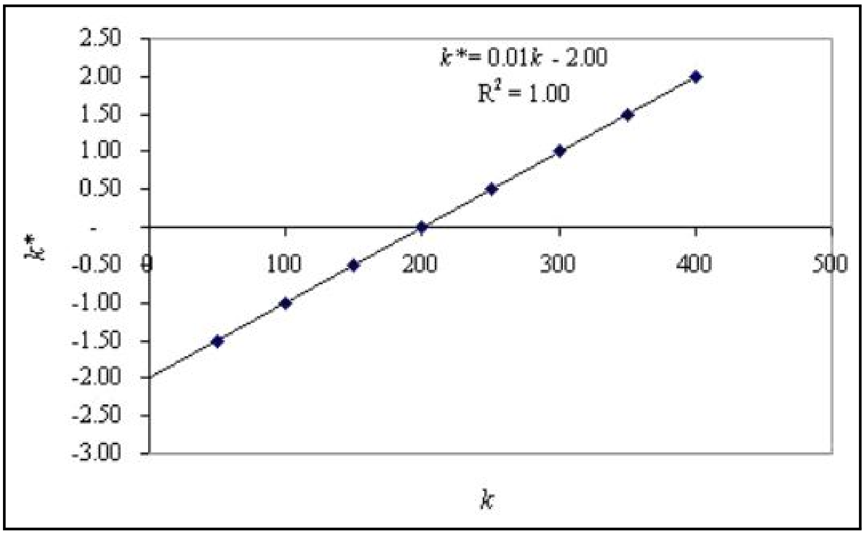


Figure 5.136: **Linearized form of the distribution**

So thanks to the linear regression given by Microsoft® Excel® 11.8346 (or calculated by you using the techniques of linear regressions seen in the section of Numerical Methods page 6316). It comes:

$$k^* = f(k) = \frac{k}{\sigma} - \frac{\mu}{\sigma} = 0.01k - 2 \qquad (5.7.756)$$

we immediately deduce that:

$$\sigma = 100 \quad \mu = 200 \qquad (5.7.757)$$

This is thus a particular technique for a particular distribution! Similar techniques more or less simple (or complicated depending on the case...) exist for others distributions.

See now another approximate approach to solve this problem. Let's take again our table for this example:

Price of receipts	Upper limit of the interval	Center	Relative cumulative frequencies in %
$[0, 50[$	50	25	6.8
$[50, 100[$	100	75	15.87
$[100, 150[$	150	125	30.85
$[150, 200[$	200	175	50.00
$[200, 250[$	250	225	69.15
$[250, 300[$	300	275	84.13
$[300, 350[$	350	325	93.32
$[350, 400[$	400	375	97.72
$[400$ and $+$	$-$	$-$	100

The average is now calculated using the central value of the intervals and sample sizes according to the relation we have seen at the beginning of this section (page 546):

$$\hat{\mu}=\frac{\sum\limits_{i=1}^{N} n_i x_i}{\sum\limits_{i=1}^{N} n_i}=\frac{\sum\limits_{i=1}^{N} n_i x_i}{n} \tag{5.7.758}$$

Price of receipts	Center	Relative cumulative frequencies	Calculation
$[0,50[$	25	668	$16,700$
$[50,100[$	75	919	$68,925$
$[100,150[$	125	$1,498$	$187,250$
$[150,200[$	175	$1,915$	$335,125$
$[200,250[$	225	$1,915$	$430,875$
$[250,300[$	275	$1,498$	$411,950$
$[300,350[$	325	919	$411,950$
$[350,400[$	375	440	$165,000$
[400 and $+$	$-$	$-$	$-$
	Sum:	$9,772$	$1,914,500$
		Average:	$\dfrac{1,914,500}{9,772}=195.92$

The average that we have calculated yet is also quite close to the average obtained previously with the Henry's line.

The standard deviation will now be calculated using also the central value of the intervals and sample sizes according to the relation seen at the beginning of this chapter:

$$\hat{\sigma}=\sqrt{\frac{\sum\limits_{i=1}^{N} n_i(x_i-\hat{\mu})}{\sum\limits_{i=1}^{N} n_i}} \tag{5.7.759}$$

Price of receipts	Center	Relative cumulative frequencies	Calculation %
$[0, 50[$	25	668	16, 700
$[50, 100[$	75	919	68, 925
$[100, 150[$	125	1, 498	187, 250
$[150, 200[$	175	1, 915	335, 125
$[200, 250[$	225	1, 915	430, 875
$[250, 300[$	275	1, 498	411, 950
$[300, 350[$	325	919	411, 950
$[350, 400[$	375	440	165, 000
[400 and +	–	228	–
		Variance:	8364.16
		Standard Deviation:	91.45

The standard deviation that we have calculated yet is also quite close to the standard deviation obtained with the method of the Henry's line.

5.7.6.9.7 Q-Q plot

Another way to judge of the quality of fit of experimental data with a theoretical distribution (whatever that is!) is the use of a "quantile-quantile plot" or simply named "q-q plot".

The idea is pretty simple, it based on the comparison the experimental data relatively to the theoretical data that are supposed to follow a particular distribution. Thus, in the case of our example, if we take the values of the mean ($\sim$ 200) and standard deviation ($\sim$ 100) obtained with the Henry's line as theoretical parameters for the Normal distribution, we get:

Price of receipts	Upper experimental limit (imposed)	Relative cumulative frequencies in %	Upper theoretical limit (calculated)
$[0, 50[$	50	6.80%	50.91
$[50, 100[$	100	15.87%	100.02
$[100, 150[$	150	30.85%	149.99
$[150, 200[$	200	50.00%	200
$[200, 250[$	250	69.15%	250.01
$[250, 300[$	300	84.13%	299.98
$[300, 350[$	350	93.32%	350.00
$[350, 400[$	400	97.72%	399.90
[400 and +	–	100%	–

Plotted, this gives us the famous Q-Q plot:

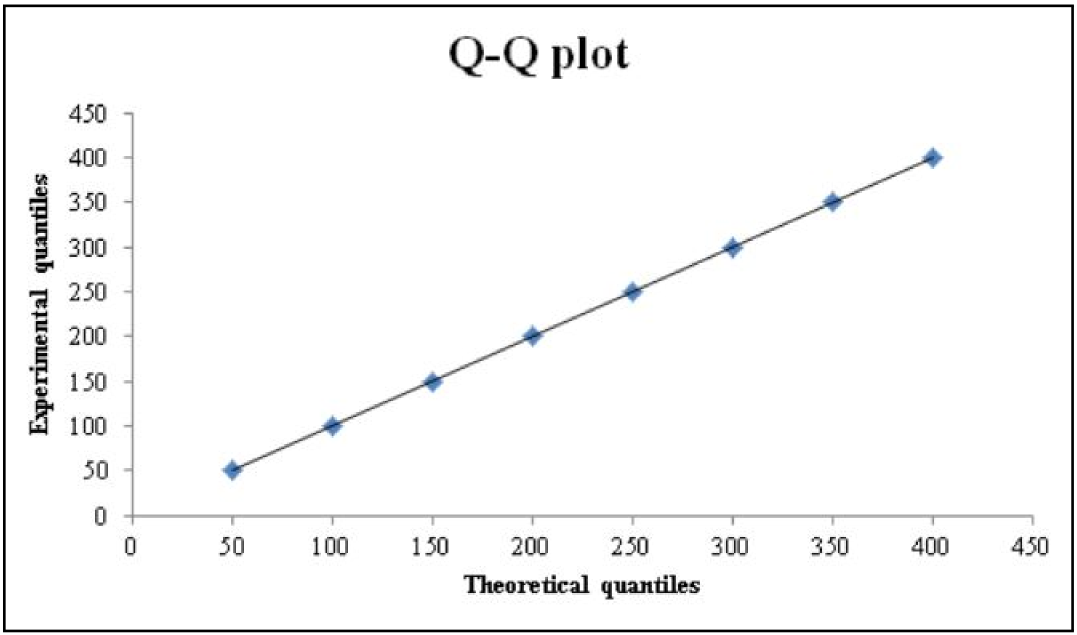


Figure 5.137: **Q-Q plot of the distribution**

And of course we can compare the observed quantiles with the supposed theoretical distribution. More the points will be aligned on the line of unit slope and zero intercept origin, the better will be the fit! It's very visual, very simple and widely used by non-specialists in business statistics.

For the people who may not have access to a spreadsheet or statistical software here is a table that may be useful relatively to the Normal distribution:

NORMAL CUMULATIVE DISTRIBUTION FUNCTION

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

5.7.6.10 Log-Normal Distribution

We say that a positive random variable X follows a "log-normal function" (or "log-Normal distribution") if by writing:

$$y=\ln(x) \tag{5.7.760}$$

we see that y follows a Normal distribution of mean μ and variance σ^2 (moments of the Normal distribution).

Verbatim by the properties of logarithms, a variable can be modelled by a log-Normal distribution if it results of the multiplication of many small independent factors (property of the product in sum of the logarithms and stability of the Normal distribution by the addition).

The density function of X for $x \geq 0$ is then (this will be justified just further below):

$$\boxed{f(x)=\mathcal{LN}(\mu,\sigma)=\frac{1}{\sigma x\sqrt{2\pi}}e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}} \tag{5.7.761}$$

that can be calculated in Microsoft® Excel® 11.8346 with the `LOGNORMDIST( )` function or its inverse by `LOGINV( )`.

This type of scenario is frequently encountered in physics, maintenance techniques, and financial markets in option pricing models (see the respective chapters of this textbook for practical examples), but also sometimes in genetics for modeling mutation rates. There is also an important point to note regarding the log-normal distribution in the later discussion of the central limit theorem!

Let us show that the cumulative probability function corresponds to a Normal distribution if we make the change of variables mentioned above:

$$\int\limits_0^{+\infty} f(x)\mathrm{d}x=\int\limits_0^{+\infty}\frac{1}{\sigma x\sqrt{2\pi}}e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}\mathrm{d}x=\frac{1}{\sigma\sqrt{2\pi}}\int\limits_0^{+\infty}\frac{1}{x}e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}\mathrm{d}x \tag{5.7.762}$$

by writing:

$$y=\ln(x)\Rightarrow\frac{\mathrm{d}y}{\mathrm{d}x}=\frac{1}{x}\Leftrightarrow \mathrm{d}x=x\mathrm{d}y \tag{5.7.763}$$

and (by definition):

$$x=e^y \tag{5.7.764}$$

we then get:

$$\begin{aligned}\int\limits_0^{+\infty} f(x)\mathrm{d}x&=\frac{1}{\sigma\sqrt{2\pi}}\int\limits_0^{+\infty}\frac{1}{x}e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}\mathrm{d}x=\frac{1}{\sigma\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}\frac{1}{x}e^{-\frac{(y-\mu)^2}{2\sigma^2}}x\mathrm{d}y\\&=\frac{1}{\sigma\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}e^{-\frac{(y-\mu)^2}{2\sigma^2}}\mathrm{d}y\end{aligned} \tag{5.7.765}$$

So we indeed fall back on the Normal distribution! Hence the cumulative function is similarly given by:

$$F_X(x)=\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{\ln(x)-\mu}{\sqrt{2}\sigma}\right)\right] \tag{5.7.766}$$

with for reminder:

$$\operatorname{erf}(x)=\frac{2}{\sqrt{\pi}}\int\limits_0^x \exp\left(-t^2\right)\mathrm{d}t \tag{5.7.767}$$

Identically as for the Normal distribution, we get the quantile function immediately:

$$Q_X(p)=\exp\left(\mu+\sqrt{2}\sigma\cdot\operatorname{erf}^{-1}(2p-1)\right) \tag{5.7.768}$$

And again identically as for the Normal distribution, we get then the median that is given by:

$$\operatorname{median}(X)=M_e=e^{\mu} \tag{5.7.769}$$

The mean (average) of X is then given by (the natural logarithm being not defined for $x<0$ we start the integral from zero):

$$\begin{aligned}\mathrm{E}(X)&=\int\limits_0^{+\infty}xf(x)\mathrm{d}x=\frac{1}{\sigma\sqrt{2\pi}}\int\limits_0^{+\infty}x\frac{1}{x}e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}\mathrm{d}x=\frac{1}{\sigma\sqrt{2\pi}}\int\limits_0^{+\infty}e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}\mathrm{d}x\\&=\frac{1}{\sigma\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}e^{-\frac{(u-\mu)^2}{2\sigma^2}+u}\mathrm{d}u\end{aligned} \tag{5.7.770}$$

where we performed the change of variable:

$$u=\ln(x)\Leftrightarrow x=e^u\qquad x\mathrm{d}u=\mathrm{d}x \tag{5.7.771}$$

The expression:

$$\frac{(u-\mu)^2}{2\sigma^2}+u \tag{5.7.772}$$

moreover being equal to:

$$-\frac{1}{2\sigma^2}\left(\left(u-(\mu+\sigma^2)\right)^2-\left(\mu+\sigma^2\right)^2+\mu^2\right) \tag{5.7.773}$$

the last integral also becomes:

$$\begin{aligned}\mathrm{E}(X)&=\frac{e^{\frac{\left(\mu+\sigma^2\right)^2-\mu^2}{2\sigma\pi}}}{\sigma\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}e^{-\frac{1}{2\sigma\pi}\left(u-\left(\mu+\sigma^2\right)^2\right)}\mathrm{d}u\\&=e^{\frac{\left(\mu+\sigma^2\right)^2-\mu^2}{2\sigma\pi}}=e^{\mu+\frac{\sigma^2}{2}}\end{aligned} \tag{5.7.774}$$

and where we used the property that emerged during our study of the Normal distribution, that is to say that any integral of the form:

$$\int\limits_{-\infty}^{+\infty} e^{-\frac{(x-c^{\text{te}})^2}{2\sigma^2}}\mathrm{d}x=\sigma\sqrt{2\pi} \tag{5.7.775}$$

always has the same value!

To calculate the variance, recall that for a random variable X, we have the Huygens theorem:

$$\mathrm{V}(X)=\mathrm{E}(X^2)-\mathrm{E}(X)^2 \tag{5.7.776}$$

Let us calculate $\mathrm{E}(X^2)$ by performing similarly to previous developments:

$$\begin{aligned}
\mathrm{E}(X^2)&=\int\limits_0^{+\infty} x^2 f(x)\mathrm{d}x=\frac{1}{\sigma\sqrt{2\pi}}\int\limits_0^{+\infty} xe^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}\mathrm{d}x\\
&=\frac{1}{\sigma\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} e^u e^{-\frac{(u-\mu)^2}{2\sigma^2}}e^u\mathrm{d}u=\frac{1}{\sigma\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} e^{-\frac{(u-\mu)^2}{2\sigma^2}+2u}\mathrm{d}u\\
&=\frac{1}{\sigma\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} e^{-\frac{\left(u-\left(\mu+2\sigma^2\right)\right)^2}{2\sigma^2}}e^{\frac{\left(\mu+2\sigma^2\right)^2-\mu^2}{2\sigma^2}}\mathrm{d}u\\
&=\frac{e^{\frac{\left(\mu+2\sigma^2\right)-\mu^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} e^{-\frac{\left(u-\left(\mu+2\sigma^2\right)\right)^2}{2\sigma^2}}\mathrm{d}u\\
&=e^{\frac{\left(\mu+2\sigma^2\right)-\mu^2}{2\sigma^2}}=e^{\frac{4\mu\sigma^2+4\sigma^4}{2\sigma^2}}=e^{2\mu+2\sigma^2}=e^{2(\mu+\sigma^2)}
\end{aligned} \tag{5.7.777}$$

where once again we have the change of variable:

$$u=\ln(x)\Leftrightarrow x=e^u\Rightarrow\mathrm{d}x=e^u\mathrm{d}u \tag{5.7.778}$$

and where we transformed the expression:

$$-\frac{(u-\mu)^2}{2\sigma^2}+2u \tag{5.7.779}$$

as:

$$-\frac{1}{2\sigma^2}\left(\left(u-(\mu+2\sigma^2)\right)^2-\left(\mu+2\sigma^2\right)^2+\mu^2\right) \tag{5.7.780}$$

Then:

$$\begin{aligned}
\mathrm{V}(X)&=\mathrm{E}(X^2)-\mathrm{E}(X)^2=e^{2\mu+2\sigma^2}-\left(e^{\mu+\frac{\sigma^2}{2}}\right)^2\\
&=e^{2\mu+2\sigma^2}-e^{2\mu+\sigma^2}=e^{2\mu+\sigma^2}\left(e^{\sigma^2}-1\right)
\end{aligned} \tag{5.7.781}$$

Here is a plot example of the distribution and cumulative distribution of the Log-Normal function of parameters $(\mu,\sigma)=(0,1)$:

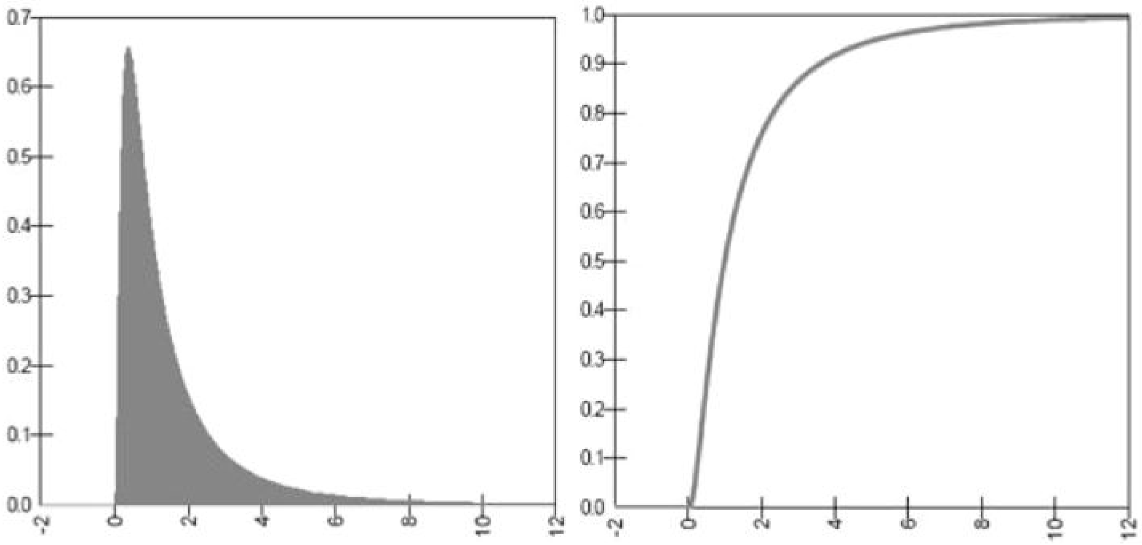


Figure 5.138: **Log-Normal law (mass and cumulative distribution function)**

Remark

Notice that if $y=\log(x)$ follows a $\mathcal{N}(\mu,\sigma)=\mathcal{N}(0,1)$ distribution. Then we have for the mean and variance of the log-Normal distribution:

$$\text{E}(X)=e^{\mu+\frac{\sigma^2}{2}}=e^{0.5}\cong 1.648 \quad\text{and}\quad \text{V}(X)=e^{2\mu+\sigma^2}\left(e^{\sigma^2}-1\right)=e^1\left(e^1-1\right)\cong 4.67 \tag{5.7.782}$$

The median of the log-Normal distribution can be obtained quite straightforward. Let us recall that M_e is given by the value of x for which the first derivative cancels, i.e.:

$$\frac{\partial}{\partial x}\frac{1}{\sigma x\sqrt{2\pi}}e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}=0 \tag{5.7.783}$$

We get then immediately:

$$M_e=e^{\mu} \tag{5.7.784}$$

The modal value of the log-Normal distribution is kinda a trap. Some people guess sometimes that the modal value is $M_0=e^{\mu}$ too. But that's wrong! Indeed, let us remind first that the modal value is defined by:

$$\text{mode}(X)=\underset{x}{\arg\max} f_X(x) \tag{5.7.785}$$

And remember that the density function of the log-Normal distribution is:

$$f_X(x)=\frac{1}{x\sigma\sqrt{2\pi}}\cdot\exp\left[-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right] \tag{5.7.786}$$

Therefore if we calculate the root of the first derivative:

$$\begin{aligned}f'_X(x)=0&=-\frac{1}{x^2\sigma\sqrt{2\pi}}\cdot\exp\left[-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right]\cdot\left(1+\frac{\ln(x)-\mu}{\sigma^2}\right)\\-1&=\frac{\ln(x)-\mu}{\sigma^2}\\x&=e^{\left(\mu-\sigma^2\right)}\end{aligned} \tag{5.7.787}$$

Hence the modal value of the log-Normal distribution given by:

$$M_0 = e^{\mu-\sigma^2} \tag{5.7.788}$$

5.7.6.11 Continuous Uniform Distribution

Let us choose $a < b$. We define the continuous uniform distribution function or "uniform function" by the relation:

$$\boxed{\mathcal{U}_{a,b}(x) = \frac{1}{b-a}1_{[a,b]}} \tag{5.7.789}$$

where $1_{[a,b]}$ means that outside the domain of definition $[a,b]$ the distribution function is zero. We will find this type of notation later in some other distribution functions. The case $\mathcal{U}_{0,1}(x)$ is commonly named the "standard uniform function".

Obviously the value $1/(b-a)$ is also the value of the modal (mode) value as the distribution is flat!

So we have for the cumulative distribution function:

$$P(X \leq x) = \int_a^x \frac{1}{b-a}1_{[a,b]}\mathrm{d}x = \frac{1}{b-a}1_{[a,b]}\int_a^x \mathrm{d}x = \frac{x_{[a,b]}-a}{b-a} \tag{5.7.790}$$

It is indeed a distribution function because it satisfies (simple integral):

$$\int_{-\infty}^{+\infty} P_{a,b}\mathrm{d}x = \int_{-\infty}^{+\infty}\frac{1}{b-a}1_{[a,b]}\mathrm{d}x = \frac{1}{b-a}1_{[a,b]}\int_a^b \mathrm{d}x = \frac{1}{b-a}1_{[a,b]}x\Big|_a^b = 1_{[a,b]}\frac{b-a}{b-a} = 1 \tag{5.7.791}$$

The continuous uniform function has for expected mean (and as the distribution is symmetric to the mean that's obviously also the value of the median):

$$\begin{aligned}\mu = \mathrm{E}(X) &= \int_a^b xf(x)\mathrm{d}x = \frac{1}{b-a}\int_a^b x\mathrm{d}x = \frac{1}{b-a}\frac{x^2}{2}\Bigg|_a^b = \frac{1}{b-a}\frac{b^2-a^2}{2}\\ &= \frac{1}{b-a}\frac{(b+a)(b-a)}{2} = \frac{a+b}{2}\end{aligned} \tag{5.7.792}$$

and for the variance using the Huygens theorem:

$$\begin{aligned}\mathrm{V}(X)=\mathrm{E}(X^2)-\mathrm{E}(X)^2&=\int\limits_a^b x^2 f(x)\mathrm{d}x-\left(\frac{a+b}{2}\right)^2=\frac{1}{b-a}\int\limits_a^b x^2\mathrm{d}x-\left(\frac{a+b}{2}\right)^2\\&=\frac{1}{b-a}\left.\frac{x^3}{3}\right|_a^b-\left(\frac{a+b}{2}\right)^2=\frac{1}{b-a}\frac{b^3-a^3}{3}-\left(\frac{a+b}{2}\right)^2\\&=\frac{1}{b-a}\frac{(b-a)(b^2+ab+a^2)}{3}-\left(\frac{a+b}{2}\right)^2\\&=\frac{1}{3}(b^2+ab+a^2)-\frac{1}{4}(b^2+2ab+a^2)=\frac{4(b^2+ab+a^2)-3(b^2+2ab+a^2)}{12}\\&=\frac{4b^2+4ab+4a^2-3b^2-6ab-3a^2)}{12}=\frac{b^2-2ab-a^2}{12}=\frac{(b-a)^2}{12}\end{aligned}\tag{5.7.793}$$

Here is a plot example of the distribution and cumulative distribution of the continuous uniform function of parameters $(a,b)=(0,1)$:

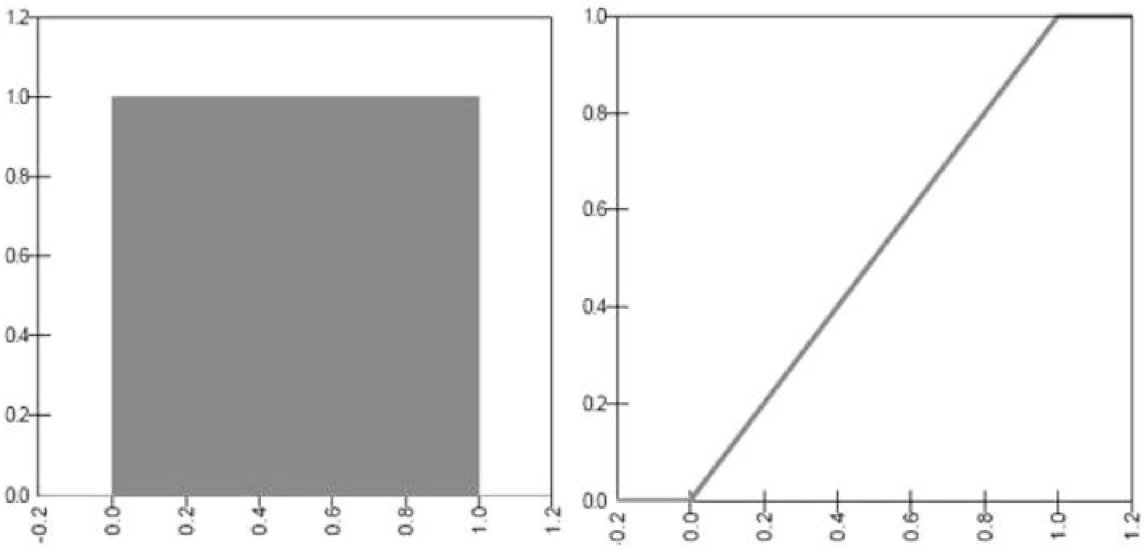


Figure 5.139: **Uniform continuous law (mass and cumulative distribution function)**

Using the same method above we can derive a quite useful special case (that we will use in the chapter of Theoretical Computing). That is the excepted mean of $\mathcal{U}_{0,1}(x)$:

$$\mathrm{E}(\mathcal{U}_{0,1}^2)=\int_0^1 x^2 f(x)\mathrm{d}x=\int_0^1 u^2\cdot 1\mathrm{d}x=\frac{1}{3}\tag{5.7.794}$$

Remark

The uniform function $\mathcal{U}_{a,b}(x)$ is often used in business simulation to indicate that the random variable has equal probabilities to have a value within a certain interval (typically in portfolio returns or in the estimation of project durations). The best example of application is again Crystal Ball or @Risk software that integrate with Microsoft Project.

Let X be a random variable following a continuous uniform distribution:

$$X=\mathcal{U}_{a,b}(x)=\frac{1}{b-a}1_{[a,b]} \tag{5.7.795}$$

Then, the quantile function of X is:

$$Q_X(p)=\begin{cases}-\infty \text{ if } p=0\\ bp+a(1-p) \text{ if } p>0\end{cases} \tag{5.7.796}$$

The cumulative distribution function of the continuous uniform distribution is:

$$F_X(x)=\begin{cases}0 & \text{if } x<a\\ \dfrac{x-a}{b-a} & \text{if } a\leq x\leq b\\ 1 & \text{if } x>b\end{cases} \tag{5.7.797}$$

The quantile function is defined for recall as the smallest x, such that $F_X(x)=p$:

$$Q_X(p)=\min\{x\in\mathbb{R}\mid F_X(x)=p\} \tag{5.7.798}$$

Thus, we have $Q_X(p)=-\infty$, if $p=0$. When $p>0$, it holds that:

$$Q_X(p)=F_X^{-1}(x) \tag{5.7.799}$$

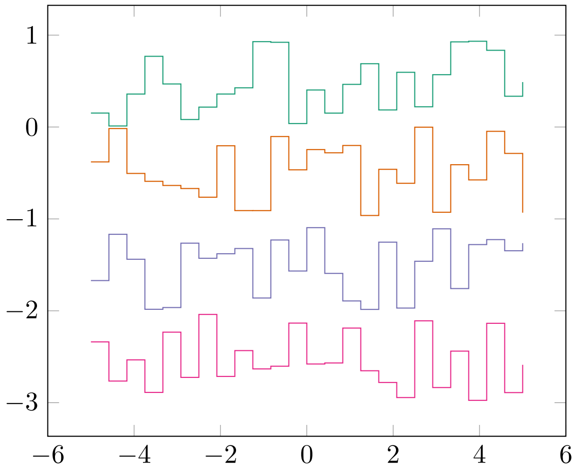
This can be derived by rearranging the cumulative function:

$$p=\frac{x-a}{b-a} \tag{5.7.800}$$

As:

$$x=p(b-a)+a=bp+a(1-p) \tag{5.7.801}$$

The figure below shows the generation (25 trials) of four variables uniformly distributed with different shifts:



We can typically recognize above patterns that are very similar to those observed in nuclear and material laboratories that study the quantum vacuum fluctuations!

Let us see an interesting result of the continuous uniform distribution (and that applies also to the discrete one as well...).

I often hear managers (who consider themselves at high level) that if we have a measure with an equal probability to occur in a closed given interval, then the sum of two such independent random variables have also the same equal probability in the same interval!

Now we will prove here that this is not the case (if someone has a more elegant proof I'm interested)!

Proof 5.47.5. Consider two independent random variables X and Y that follow a uniform distribution in a closed interval $[0, a]$. We are searching the density of their sum will be written:

$$Z = X + Y \tag{5.7.802}$$

Then we have:

$$f_X(x) = f_Y(y) = \begin{cases} +1 & \text{if } 0 \leq x, y \leq a \\ 0 & \text{otherwise} \end{cases} \tag{5.7.803}$$

with the variable:

$$0 \leq z \leq 2a \tag{5.7.804}$$

To calculate the distribution of the sum, remember that we know that in discrete terms this is equivalent to the joint product of probabilities (see section Probabilities page 480) of the occurrence of two continuous variables (remember the same kind of calculation in the discreet form!).

That is to say:

$$f_Z(z) = \int\limits_{-\infty}^{+\infty} f_X(z-y) f_Y(y) \mathrm{d}y \tag{5.7.805}$$

As $f_Y(y) = 1$ if $0 \leq y \leq a$ and 0 otherwise then the product of the previous convolution reduces to:

$$f_Z(z) = \int\limits_0^a f_X(z-y) \mathrm{d}y \tag{5.7.806}$$

The integrand is by definition 0 except by construction in the interval $0 \leq z - y \leq a$ it is then 1.

Let us focus on the limits of the integral that is in this case the only one that is interesting ....

First we make a change of variables by writing:

$$u = z - y \tag{5.7.807}$$

thus:

$$\mathrm{d}u = -\mathrm{d}y \tag{5.7.808}$$

The integral can be then written in this interval after the change of variable:

$$f_Z(z)=\int\limits_0^a f_X(z-y)\mathrm{d}y=-\int\limits_z^{z-a} f_X(u)\mathrm{d}u=\int\limits_z^{z-a}\mathrm{d}u \tag{5.7.809}$$

Remembering that we have seen at the beginning that $0\leq z\leq 2a$, then we have immediately if $z<0$ and $z>2a$ that the integral is zero.

We will consider two cases for the interval because the convolution of these two rectangular functions can be distinguished according to the situation where at first they cross (nest), that is to say where $0\leq z\leq a$, and then recede from each other, that is to say $a<z\leq 2a$.

- In the first case (nest) where $0\leq z\leq a$:

  $$f_Z(u)=\int\limits_{z-a}^{z}\mathrm{d}u=\int\limits_0^z\mathrm{d}u=u|_0^z=z \tag{5.7.810}$$

  where we changed the lower bound to 0 because anyway $f_X(u)$ is zero for any negative value (and when $0\leq z\leq a$,$z-a$ is precisely zero or negative!).

- In the second case (dislocation) where $a<z\leq 2a$:

  $$f_Z(u)=\int\limits_{z-a}^{z}\mathrm{d}u=\int\limits_{z-a}^{a}\mathrm{d}u=a-(z-a)=2a-z \tag{5.7.811}$$

  where we changed the upper terminal a because anyway $f_X(u)$ is zero for any higher value (and when $a\leq z\leq 2a$, z is just larger than a).

  So in the end, we have:

  $$f_Z(z)=\begin{cases} z & \text{if } 0\leq z\leq a\\ 2a-z & \text{if } a\leq z\leq 2a\\ 0 & \text{otherwise}\end{cases} \tag{5.7.812}$$

■ Q.E.D.

This is a particular case, deliberately simplified, of the triangular distribution that we will discover just after...

This result (which may seems perhaps not intuitive) can be checked in a few seconds with a spreadsheet software like Microsoft® Excel® 11.8346 using the `RANDBETWEEN()` and the `FREQUENCY( )` functions.

5.7.6.12 Triangular Distribution

Let $a < c < b$. We define the "triangular distribution" (or "triangular function") by construction based on the following two distribution functions:

$$\boxed{P_{a,c}(x)=\frac{2(x-a)}{(b-a)(c-a)}\cdot 1_{[a,c]} \quad \text{and} \quad P_{c,b}(x)=\frac{2(b-x)}{(b-a)(b-c)}\cdot 1_{]c,b]}} \tag{5.7.813}$$

where a is often assimilated with the optimistic value, c to the modal value and b the pessimistic value.

It is also the only way to write this distribution function if the reader keeps in mind that the base of a triangle of length $c-a$ must have a height h equal to $2/(c-a)$ as its total area is equal to unity (we will soon prove it).

Here is a plot example of the triangular distribution and cumulative distribution for the parameters $(a,c,b)=(0,3,5)$:

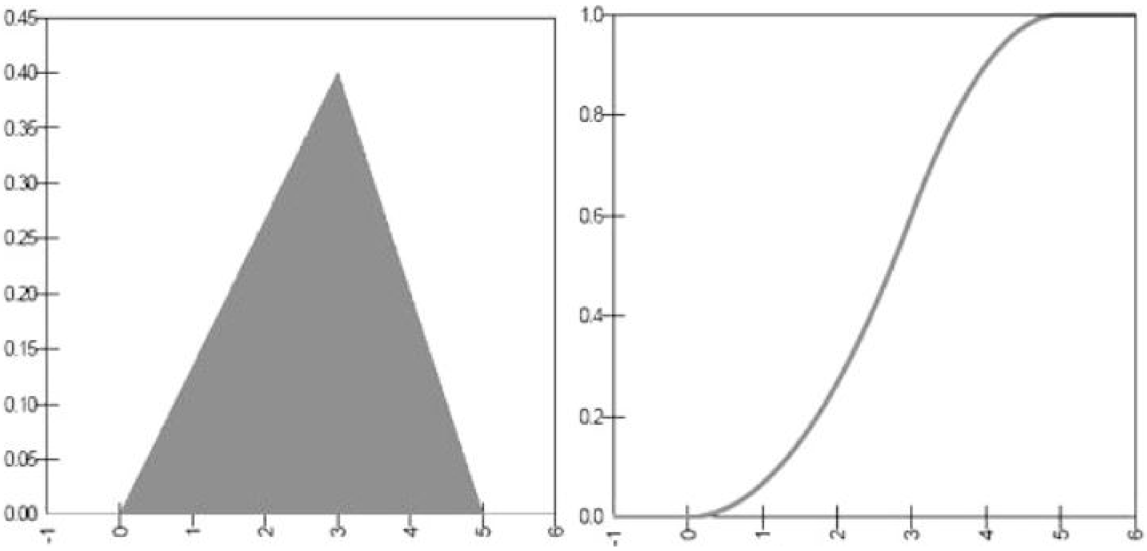


Figure 5.140: **Triangular law (mass and cumulative distribution function)**

The slope of the first straight line (increasing from left) is obviously:

$$\frac{2a}{(b-a)(c-a)} \tag{5.7.814}$$

and the slope of the second straight line (decreasing to the right):

$$\frac{-2}{(b-a)(b-c)} \tag{5.7.815}$$

This function is a distribution function if it satisfies:

$$\rho=\int\limits_{-\infty}^{+\infty}(P_{a,c}(x)+P_{c,b}(x))\mathrm{d}x=1 \tag{5.7.816}$$

It is in this case, simply the area of the triangle which we recall is simply the base multiplied by the height divided by 2 (see section Geometric Shapes page 2896):

$$\rho=\frac{1}{2}\left((b-a)\cdot\frac{2}{b-a}\right)=1 \tag{5.7.817}$$

Remark

This function is widely used in project management in the context of task duration estimations or in industrial simulations. Where a corresponds to the optimistic value, c to the expected value (mode) and the value b to the pessimistic value. The best example of application is again the softwares Crystal Ball or @Risk that are add-ins for Microsoft Project.

The triangular function has also for mean (average):

$$\begin{aligned}
\mu&=\int\limits_{-\infty}^{+\infty}xf(x)\mathrm{d}x=\int\limits_{a}^{c}x\frac{2(x-a)}{(b-a)(c-a)}\mathrm{d}x+\int\limits_{c}^{b}x\frac{2(b-x)}{(b-a)(b-c)}\mathrm{d}x\\
&=\frac{2}{(b-a)(c-a)}\left(\frac{1}{3}x^3-\frac{1}{2}ax^2\right)\Bigg|_a^c+\frac{2}{(b-a)(c-a)}\left(\frac{1}{2}bx^2-\frac{1}{3}x^3\right)\Bigg|_c^b\\
&=\frac{2}{(b-a)(c-a)}\left(\left(\frac{1}{3}c^3-\frac{1}{2}ac^2\right)-\left(\frac{1}{3}a^3-\frac{1}{2}a^3\right)\right)\\
&-\frac{2}{(b-a)(c-a)}\left(\left(\frac{1}{2}b^3-\frac{1}{3}b^3\right)-\left(\frac{1}{2}bc^2-\frac{1}{3}c^3\right)\right)\\
&=2\frac{(b-c)\left(\left(\frac{1}{3}c^3-\frac{1}{2}ac^2\right)+\frac{a^3}{6}\right)+(c-a)\left(\frac{1}{6}b^3-\left(\frac{1}{2}bc^2-\frac{1}{3}c^3\right)\right)}{(b-a)(c-a)(b-c)}\\
&=2\frac{\frac{1}{3}bc^3-\cancel{\frac{1}{2}abc^2}+\frac{ba^3}{6}-\cancel{\frac{1}{3}c^4}+\frac{1}{2}ac^3-\cancel{\frac{a^3c}{6}}}{(b-a)(c-a)(b-c)}\\
&+2\frac{\frac{1}{6}cb^3-\frac{1}{2}bc^3+\cancel{\frac{1}{3}c^4}-\frac{ab^3}{6}+\cancel{\frac{1}{2}abc^2}-\frac{1}{3}ac^3}{(b-a)(c-a)(b-c)}\\
&=\frac{\frac{2}{3}bc^3+ac^3-\frac{1}{3}a^3c+\frac{1}{3}cb^3+\frac{ba^3}{3}-\frac{ab^3}{3}-bc^3-\frac{2}{3}ac^3}{(b-a)(c-a)(b-c)}\\
&=\frac{-\frac{1}{3}bc^3-\frac{1}{3}a^3c+\frac{1}{3}ba^3-\frac{1}{3}\frac{1}{3}ab^3+\frac{1}{3}cb^3+\frac{1}{3}ac^3}{(b-a)(c-a)(b-c)}\\
&=\frac{1}{3}\frac{-bc^3-a^3c+ba^3-ab^3+cb^3+ac^3}{(b-a)(c-a)(b-c)}=\frac{1}{3}\frac{(a+b+c)(b-a)(c-a)(b-c)}{(b-a)(c-a)(b-c)}\\
&=\frac{a+b+c}{3}
\end{aligned} \tag{5.7.818}$$

and for variance:

$$\begin{aligned}\sigma^2&=\int\limits_{-\infty}^{+\infty}(x-\mu)^2f(x)\mathrm{d}x=\int\limits_{a}^{c}(x-\mu)^2\frac{2(x-a)}{(b-a)(c-a)}\mathrm{d}x+\int\limits_{c}^{b}(x-\mu)^2\frac{2(b-x)}{(b-a)(b-c)}\mathrm{d}x\\&=-\frac{1}{6}\frac{c\left(-3c^3+8c^3\mu+4c^2a-6c\mu^2-12c\mu a+12\mu^2a\right)}{(b-a)(c-a)}+\frac{1}{6}\frac{a^2(a^2-4\mu a+6\mu^2}{(b-a)(c-a)}\\&=-\frac{1}{6}\frac{-3c^3+8c^2\mu+4c^2b-6c\mu^2-12c\mu b+12\mu^2b}{(b-a)(c-a)}+\frac{1}{6}\frac{b^2(b^2-4\mu b+6\mu^2)}{(b-a)(c-a)}\\&=\frac{1}{6}b^2+\frac{1}{6}ba-\frac{2}{3}\mu b+\frac{1}{6}a^2-\frac{2}{3}\mu a+\mu^2+\frac{1}{6}ca-\frac{2}{3}c\mu+\frac{1}{6}c^2\end{aligned}\tag{5.7.819}$$

We can replace μ by the result obtained before and we get after simplification (it is boring algebra...):

$$\sigma^2=\frac{a^2+b^2+c^2-ab-ac-bc}{18}\tag{5.7.820}$$

We can show that the sum of two independent random variables, each uniformly distributed on $[a,b]$ (i.e. independent and identically distributed) follows a triangular distribution on $[2a,2b]$ but if they do not have the same limits, then their sum gives something that has no name to my knowledge...

As already mentioned earlier in the previous remark box, the triangular distribution is used a lot in Risk Management by beginners (the level above consisting in using the beta distribution that we will see further below). A quite common figure you can find in project management textbooks for beginners is the impact/probability matrix give below ("impact" is modelled according to the distance between optimistic and pessimistic points):

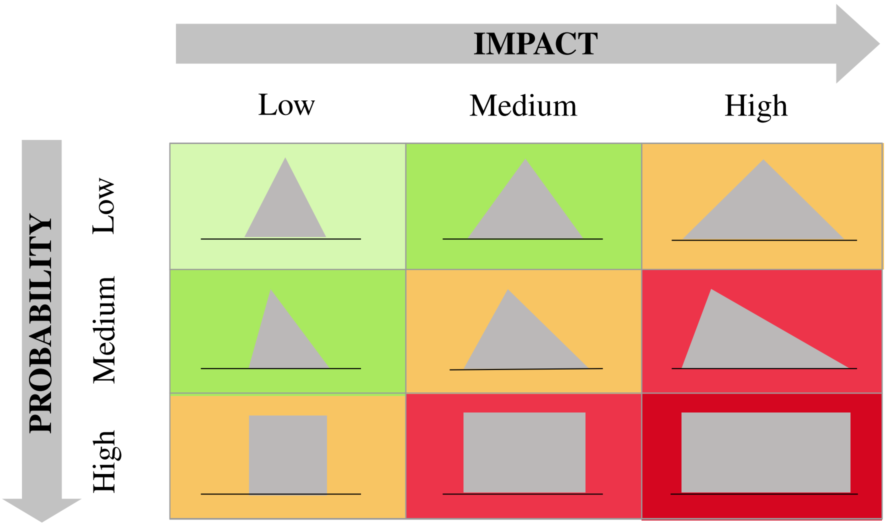


Figure 5.141: **Naive risk matrix in project management (author: Jose Barato)**

The idea of a risk matrix was first developed in the Electronic System Centre, US Air force (see [197]).

5.7.6.13 Pareto Distribution

The "Pareto distribution" (or "Pareto law"), also named "power law" or "scale law" is the formalization of the $80-20$ principle. This decision tool helps determine the critical factors (about 20%) influencing the majority (80%) of the goal.

Remark

This distribution is a fundamental and basic tool in quality management (see Industrial Engineering page 8772 and Quantitative Management page 8095). It is also used in reinsurance. The theory of queues had also some interest in this distribution when some research in the 11990s (holocene calendar) showed that this distribution also seems ton explain well a number of variables observed in the Internet traffic (and more generally on all high speed data networks).

A random variable is said by definition follow a Pareto distribution if its cumulative distribution function is given by:

$$P(X\leqslant x)=1-\left(\frac{x_m}{x}\right)^k \tag{5.7.821}$$

with x that must be greater than or equal to x_m.

The Pareto density function (distribution function) is then given by:

$$f(x)=\frac{\mathrm{d}}{\mathrm{d}x}\left(1-\left(\frac{x_m}{x}\right)^k\right)=-\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{x_m}{x}\right)^k=-x_m^k\frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{x^k}=k\frac{x_m^k}{x^{k+1}} \tag{5.7.822}$$

with $k\in\mathbb{R}_+$ and $x\geq x_m\geq 0$ (then $x>0$).

The Pareto distribution is defined by two parameters, x_m and k (named "Pareto index"). This distribution is also said to be "scale invariant" or "fractal distribution", because of the following property:

$$f\left(c^{te}\cdot x\right)=kx_m^k\left(c^{te}\cdot x\right)^{-k-1}=\left(c^{te}\right)^{-k-1}kx_m^kx^{-k-1}=\left(c^{te}\right)^{-k-1}f(x)\propto f(x) \tag{5.7.823}$$

The Pareto function is also well a distribution function as the cumulative distribution function known we have:

$$\begin{aligned}\int\limits_0^{+\infty}f(x)\mathrm{d}x&=\left.\left(1-\left(\frac{x_m}{x}\right)^k\right)\right|_{x_m}^{+\infty}\\&=\left(1-\left(\frac{x_m}{+\infty}\right)^k\right)-\left(1-\left(\frac{x_m}{x_m}\right)^k\right)=(1-0)-(1-1^k)=1\end{aligned} \tag{5.7.824}$$

The expected mean is given by:

$$\mu=\mathrm{E}(X)=\int\limits_{x_m}^{+\infty}xf(x)\mathrm{d}x=\int\limits_{x_m}^{+\infty}xk\frac{x_m^k}{x^{k+1}}\mathrm{d}x=kx_m^k\int\limits_{x_m}^{+\infty}x\frac{1}{x^k}\mathrm{d}x=-\left.\frac{kx_m^k}{k-1}\frac{1}{x^{k-1}}\right|_{x_m}^{+\infty}=\frac{kx_m}{k-1} \tag{5.7.825}$$

if $k > 1$. If $k \leq 1$, the mean does not exist.

To calculate the variance, using the Huygens theorem:

$$\mathrm{V}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 \tag{5.7.826}$$

we get:

$$\mathrm{E}(X^2) = \int\limits_{x_m}^{+\infty} x^2 f(x)\mathrm{d}x = kx_m^k \int\limits_{x_m}^{+\infty} \frac{1}{x^{k-1}}\mathrm{d}x = -\frac{kx_m^k}{k-2}\frac{1}{x^{k-2}}\Bigg|_{x_m}^{+\infty} = \frac{kx_m^2}{k-2} \tag{5.7.827}$$

if $k > 2$. If $k \leq 2$, $\mathrm{E}(X^2)$ doesn't exist!

So if $k > 2$:

$$\sigma^2 = \mathrm{V}(X) = \frac{kx_m^2}{k-2} - \left(\frac{kx_m}{k-1}\right)^2 = \frac{kx_m^2}{(k-1)^2(k-2)} \tag{5.7.828}$$

If $k \leq 2$, the variance doesn't exists.

Here is a plot example of the Pareto distribution and cumulative distribution for the parameters $(x, x_m, k) = (x, 1, 2)$:

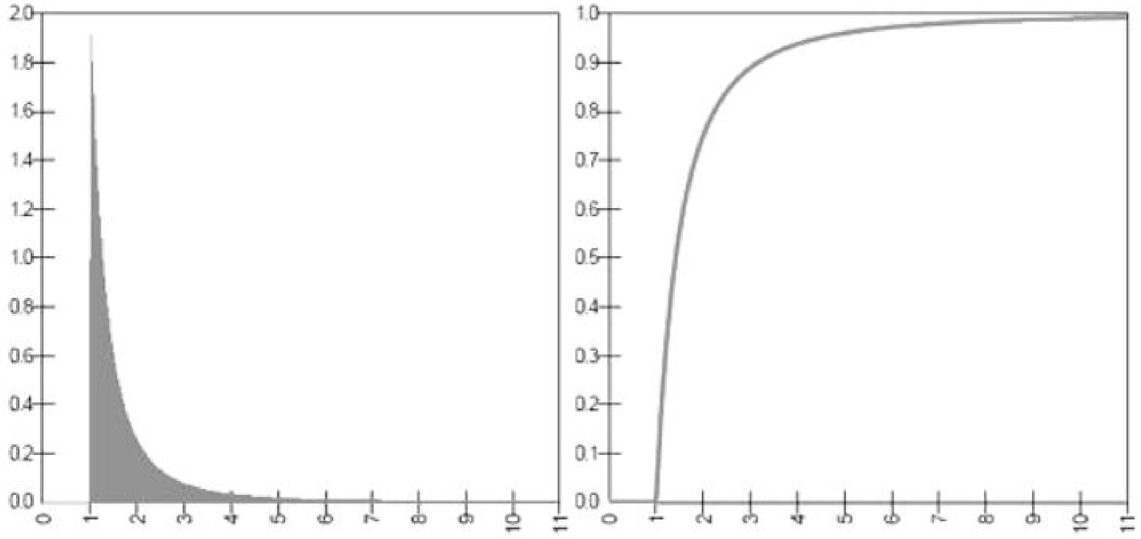


Figure 5.142: **Pareto law (mass and cumulative distribution function)**

Remark

See that when $k \rightarrow +\infty$ the distribution approaches $\delta(x - x_m)$ where δ is the Dirac delta function.

There is another important way to deduce the family of Pareto distributions that allows us to understand many things about other distributions and that is often presented as follows:

Let us write x_0 the threshold beyond which we calculate the mean of the considered quantity, and $\mathrm{E}(Y)$ the mean beyond this threshold x_0 as it is proportional (linearly dependent) to the chosen threshold:

$$\mathrm{E}(Y) = ax_0 + b \tag{5.7.829}$$

This functional relation expresses the idea that the conditional mean beyond the threshold x_0 is a multiple of this threshold plus a constant, that is to say a linear function of the threshold.

Thus, in project management, for example, we could say that once a certain threshold of time is exceeded, the expected duration is a multiple of this threshold plus a constant.

If a linear relation of this type exists and is satisfied, then we talk about a probability distribution in the form of a generalized Pareto distribution.

Consider the mean of the Bayesian conditional function given by (see section Probabilities page 499):

$$\mathrm{E}(Y)=\frac{1}{P(X\geq x_0)}\int\limits_{x_0}^{+\infty} yf(y)\mathrm{d}y \tag{5.7.830}$$

where the notation on the left hand side is somewhat shortened but the reader will implicitly understand that it is a conditional expectation.

If we write $F(y)$ the cumulative distribution function $f(y)$, then we have by definition:

$$\mathrm{d}F(y)=f(y)\mathrm{d}y \tag{5.7.831}$$

Thus:

$$\mathrm{E}(Y)=\frac{1}{P(X\geq x_0)}\int\limits_{x_0}^{+\infty} y\mathrm{d}F(y) \tag{5.7.832}$$

and if we define:

$$\bar{F}_X(x)=P(X\geq x_0)=1-F_X(x) \tag{5.7.833}$$

what we can assimilate to the "tail of the distribution".

We get:

$$\mathrm{E}(Y)=\frac{1}{\bar{F}(x_0)}\int\limits_{x_0}^{+\infty} y\mathrm{d}F(y) \tag{5.7.834}$$

and therefore we seek the very special case where:

$$\mathrm{E}(Y)=\frac{1}{\bar{F}(x_0)}\int\limits_{x_0}^{+\infty} y\mathrm{d}F(y)=ax+b \tag{5.7.835}$$

this is to say:

$$\int\limits_{x_0}^{+\infty} y\mathrm{d}F(y)=(ax+b)\bar{F}(x_0) \tag{5.7.836}$$

Differentiating with respect to x, we find:

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\int_{x_0}^{+\infty} y\mathrm{d}F(y)\right)=\frac{\mathrm{d}}{\mathrm{d}x}\left((ax+b)\bar{F}(x_0)\right) \tag{5.7.837}$$

The derivative of the integral defined above will be the derivative of a constant (valorisation of the integral in $+\infty$) minus the derivative of the analytical expression of the integral for x_0. So we have:

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\int_{x_0}^{+\infty} y\mathrm{d}F(y)\right)=\frac{\mathrm{d}}{\mathrm{d}x}\left(\int_{x_0}^{+\infty} yf(y)\mathrm{d}y\right)=-xf(x)=-x\frac{\mathrm{d}F(x)}{\mathrm{d}x}=a\bar{F}(x)+(ax+b)\frac{\mathrm{d}\bar{F}(x)}{\mathrm{d}x} \tag{5.7.838}$$

Thus:

$$-x\frac{\mathrm{d}F(x)}{\mathrm{d}x}=a\bar{F}(x)+(ax+b)\frac{\mathrm{d}\bar{F}(x)}{\mathrm{d}x} \tag{5.7.839}$$

and as:

$$\mathrm{d}F=\mathrm{d}(1-\bar{F})=-\mathrm{d}\bar{F} \tag{5.7.840}$$

it comes:

$$x\frac{\mathrm{d}\bar{F}}{\mathrm{d}x}=a\bar{F}(x)+(ax+b)\frac{\mathrm{d}\bar{F}}{\mathrm{d}x} \tag{5.7.841}$$

After simplification and rearrangement we obtain:

$$a\bar{F}(x)\mathrm{d}x=-\left(x(a-1)+b\right)\mathrm{d}\bar{F}(x) \tag{5.7.842}$$

which is a differential equation in $\bar{F}(x)$. Its Resolution provides all forms of seek Pareto distributions, according to the values taken by the parameters a and b.

To solve this differential equation, consider the special case where $a>1, b=0$. Then we have:

$$a\bar{F}(x)\mathrm{d}x=-x(a-1)\mathrm{d}\bar{F}(x) \tag{5.7.843}$$

By writing:

$$k=\frac{a}{a-1} \tag{5.7.844}$$

We then get:

$$-\frac{1}{x}\mathrm{d}x=\frac{1}{k}\frac{1}{\bar{F}(x)}\mathrm{d}\bar{F}(x) \tag{5.7.845}$$

and therefore:

$$-\ln(x)=\frac{1}{k}\ln(\bar{F}(x))+c^{te} \tag{5.7.846}$$

It comes:

$$e^{\ln\left(\frac{1}{x}\right)}=e^{\frac{1}{k}\ln(\bar{x})+c^{te}}=e^{\frac{1}{k}\ln\left(\bar{F}(x)\right)}e^{c^{te}} \tag{5.7.847}$$

and therefore:

$$\ln\left(\frac{1}{x}\right)=\bar{F}(x)^{\frac{1}{k}}c^{te} \tag{5.7.848}$$

So we have:

$$\bar{F}(x)=\left(\frac{c^{te}}{x}\right)^k=\left(\frac{x_m}{x}\right)^k \tag{5.7.849}$$

Then it comes form the cumulative distribution function:

$$F(x)=1-\bar{F}(x)=1-\left(\frac{x_m}{x}\right)^k \tag{5.7.850}$$

If we seek for the distribution function, we derive by x to get:

$$f(x)=\frac{\mathrm{d}F(x)}{\mathrm{d}x}=k\frac{x_m^k}{x^{k+1}} \tag{5.7.851}$$

This is the Pareto distribution we have used since the beginning and named "Pareto distribution of type I" (we won't see in this textbook those of type II).

An interesting thing to observes is the case of the resolution of the following differential equation:

$$a\bar{F}(x)\mathrm{d}x=-(x(a-1)+b)\mathrm{d}\bar{F}(x) \tag{5.7.852}$$

when $a=1,b>0$. The differential equation is then reduced to:

$$\bar{F}(x)\mathrm{d}x=-b\mathrm{d}\bar{F}(x) \tag{5.7.853}$$

Thus:

$$-\frac{1}{b}\mathrm{d}x=\frac{1}{\bar{F}(x)}\mathrm{d}\bar{F}(x) \tag{5.7.854}$$

After integration:

$$-\frac{1}{b}x=\ln(\bar{F}(x)) \tag{5.7.855}$$

and therefore:

$$\bar{F}(x)=e^{-\frac{1}{b}x} \tag{5.7.856}$$

If we make a small change in notation:

$$\bar{F}(x)=e^{-\lambda x} \tag{5.7.857}$$

and that we write the distribution function:

$$F(x)=1-\bar{F}(x)=1-e^{-\lambda x} \tag{5.7.858}$$

and by derivating we get the distribution function of the exponential distribution:

$$F(x)=\lambda e^{-\lambda x} \tag{5.7.859}$$

So the exponential distribution has a conditional mean threshold that is equal to:

$$\mathrm{E}(Y)=ax_0+b\underset{a=1}{=}x_0+b\underset{b=\frac{1}{\lambda}}{=}x_0+\frac{1}{b}=x_0+\sigma \tag{5.7.860}$$

So the conditional mean threshold is equal to itself plus the standard deviation of the distribution.

5.7.6.14 Exponential Distribution

We define the "exponential distribution" (or "exponential law") by the following distribution function:

$$\boxed{P(x)=\lambda e^{-\lambda x}1_{[0,+\infty]}} \tag{5.7.861}$$

with $\lambda>0$ that as we will immediately see is in the fact that the inverse of the mean and where x is a random variable without memory. This law is also sometimes denoted $\mathcal{E}(\lambda)$.

In fact the exponential distribution naturally appears from simple developments (see the Nuclear Physics chapter for example) under assumptions that impose a constance in the aging of phenomenon. In the section of Quantitative Management, we have also proved in detail in the section on the theory of queues (page 8149), that this law was without memory. That is to say, that the cumulative probability of a phenomenon occurs between the time t and $t+s$, if it is not realized before, is the same as that the cumulative probability of occurring between the time 0 and s.

Remarks

R1. This function is occurring frequently in nuclear physics (see chapter of the same name) or quantum physics (see also chapter of the same name) as well in reliability (see section Quantitative Management page 8673) or in the theory of queues (see section Quantitative Management page 8141).

R2. We can get this distribution in Microsoft® Excel® 11.8346 with the `EXPONDIST( )` function.

It is also really a distribution function because it verifies:

$$\int\limits_{-\infty}^{+\infty} P_\lambda(x)\mathrm{d}x=\lambda\int\limits_{0}^{+\infty} e^{-\lambda x}\mathrm{d}x=\lambda\left(-\frac{1}{\lambda}\right)e^{-\lambda x}\Big|_0^{+\infty}=-\left(e^{-\infty x}-e^{0x}\right)=-(0-1)=1 \tag{5.7.862}$$

The exponential distribution has for expected mean using integration by parts:

$$\begin{aligned}\mu&=\int\limits_{-\infty}^{+\infty} xP_\lambda(x)\mathrm{d}x=\lambda\int\limits_{0}^{+\infty} xe^{-\lambda x}\mathrm{d}x=-\left.xe^{-\lambda x}\right|_0^{+\infty}-\int\limits_{0}^{+\infty}-e^{-\lambda x}\mathrm{d}x=\int\limits_{0}^{+\infty}e^{-\lambda x}\mathrm{d}x\\&=-\frac{1}{\lambda}\left.e^{-\lambda x}\right|_0^{+\infty}=\frac{1}{\lambda}\end{aligned} \tag{5.7.863}$$

and for variance using once again the Huygens relation:

$$\mathrm{V}(X)=\mathrm{E}(X^2)-\mathrm{E}(X)^2 \tag{5.7.864}$$

it remains for us to only the to calculate:

$$\mathrm{E}(X^2)=\int\limits_{0}^{+\infty}\lambda x^2e^{-\lambda x}x \tag{5.7.865}$$

A variable change $y = \lambda$ leads us to:

$$\mathrm{E}(X^2) = \frac{1}{\lambda^2}\int_0^{+\infty} y^2 e^{-y}\mathrm{d}y \tag{5.7.866}$$

A double integration by parts gives us:

$$\begin{gathered}\int_a^b f(t)g'(t)\mathrm{d}t = f(t)g(t)|_a^b - \int_a^b f'(t)g(t)\mathrm{d}t \\ \int_0^{+\infty} y^2e^{-y}\mathrm{d}y = -y^2e^{-y}|_0^{-\infty} + 2\int_0^{+\infty} ye^{-y}\mathrm{d}y = 2\left[-ye^{-y}|_0^{+\infty}\right] + 2\int_0^{+\infty} e^{-y}\mathrm{d}y = 2\end{gathered} \tag{5.7.867}$$

Hence:

$$\mathrm{E}(X^2) = \frac{2}{\lambda^2} \tag{5.7.868}$$

we have therefore:

$$\mathrm{V}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2} \tag{5.7.869}$$

So the standard deviation (square root of the variance for recall) and mean have exactly the same expression!

Here is a plot example of the exponential distribution and cumulative distribution for the parameter $\lambda = 1$:

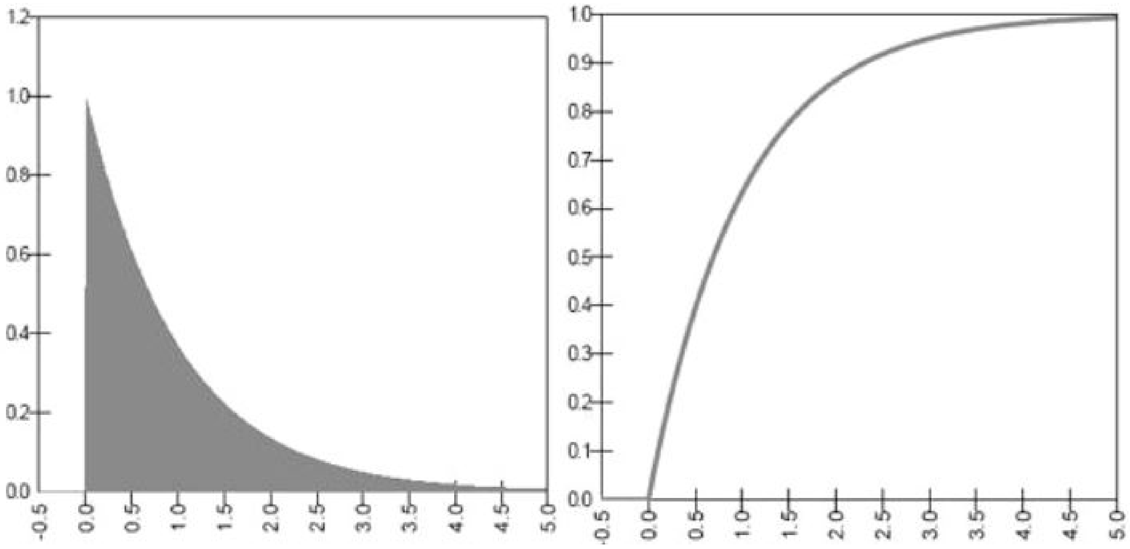


Figure 5.143: Exponential law (mass and cumulative distribution function)

Now let us determine the cumulative distribution function of the exponential law:

$$P(X \leqslant x) = \int_0^x \lambda e^{-\lambda t}\mathrm{d}t = \lambda\int_0^x e^{-\lambda t}\mathrm{d}t = \lambda\left(-\frac{1}{\lambda}e^{-\lambda t}\right)\Big|_0^x = 1 - e^{-\lambda x} \tag{5.7.870}$$

Let X be a random variable following an exponential distribution:

$$X \sim \text{Exp}(\lambda) \tag{5.7.871}$$

Then, the quantile function of X is:

$$Q_X(p) = \begin{cases} -\infty, \text{ if } p = 0 \\ -\dfrac{\ln(1-p)}{\lambda}, \text{ if } p > 0 \end{cases} \tag{5.7.872}$$

Indeed, the cumulative distribution function of the exponential distribution is as we have just shown:

$$F_X(x) = \begin{cases} 0, \text{ if } x < 0 \\ 1 - \exp[-\lambda x], \text{ if } x \geq 0 \end{cases} \tag{5.7.873}$$

Remember that the quantile function $Q_X(p)$ is defined as the smallest x, such that $F_X(x) = p$:

$$Q_X(p) = \min\{x \in \mathbb{R} \mid F_X(x) = p\} \tag{5.7.874}$$

Thus, we have $Q_X(p) = -\infty$, if $p = 0$. When $p > 0$, it holds that:

$$Q_X(p) = F_X^{-1}(x) \tag{5.7.875}$$

This can be derived by rearranging the equation :

$$\begin{aligned} p &= 1 - \exp[-\lambda x] \\ \exp[-\lambda x] &= 1 - p \\ -\lambda x &= \ln(1-p) \\ x &= -\frac{\ln(1-p)}{\lambda} \end{aligned} \tag{5.7.876}$$

Let X be a random variable following an exponential distribution:

$$X \sim e^{\lambda} \tag{5.7.877}$$

Then, the median of X is:

$$\text{median}(X) = M_e(X) = \frac{\ln(2)}{\lambda} \tag{5.7.878}$$

Indeed, remember that the median is the value at which the cumulative distribution function is $1/2$:

$$F_X(\text{median}(X)) = \frac{1}{2} \tag{5.7.879}$$

The cumulative distribution function of the exponential distribution is as we have just shown earlier above:

$$F_X(x) = 1 - e^{-\lambda x}, \quad x \geq 0 \tag{5.7.880}$$

Thus, the inverse cumulative distribution function is:

$$x = -\frac{\ln(1-p)}{\lambda} \tag{5.7.881}$$

and setting $p = 1/2$, we get:

$$\text{median}(X) = -\frac{\ln\left(1-\dfrac{1}{2}\right)}{\lambda} = \frac{\ln(2)}{\lambda} \tag{5.7.882}$$

Let X be a random variable still following an exponential distribution:

$$X \sim e^{\lambda} \tag{5.7.883}$$

Then, the mode of X is:

$$\text{mode}(X) = M_0(X) = 0 \tag{5.7.884}$$

Indeed, remember that the mode is the value which maximizes the probability density function:

$$\text{mode}(X) = \underset{x}{\arg\max} f_X(x) \tag{5.7.885}$$

The probability density function of the exponential distribution is for reminder:

$$f_X(x) = \begin{cases} 0, \text{ if } x < 0 \\ \lambda e^{-\lambda x}, \text{ if } x \geq 0 \end{cases} \tag{5.7.886}$$

Since:

$$f_X(0) = \lambda \tag{5.7.887}$$

and:

$$0 < e^{-\lambda x} < 1 \quad \text{for any} \quad x > 0 \tag{5.7.888}$$

it follows that:

$$\text{mode}(X) = 0 \tag{5.7.889}$$

Let us now compute the skewness of the exponential distribution as defined by (see page 636):

$$\gamma_1 = \frac{\text{E}\left(X^3\right) - 3\mu\sigma^2 - \mu^3}{\sigma^3} \tag{5.7.890}$$

where μ and σ are respectively the mean and standard deviation of X. Since X follows an exponential distribution, the mean of X is given by:

$$\mu = \text{E}(X) = \frac{1}{\lambda} \tag{5.7.891}$$

and the standard deviation of X is given by:

$$\sigma = \sqrt{\text{V}(X)} = \sqrt{\frac{1}{\lambda^2}} = \frac{1}{\lambda} \tag{5.7.892}$$

Therefore:

$$\begin{aligned}\gamma_1 &= \frac{\text{E}\left(X^3\right) - 3\mu\sigma^2 - \mu^3}{\sigma^3} = \frac{\text{E}\left(X^3\right)}{\sigma^3} - \frac{3\mu\sigma^2 + \mu^3}{\sigma^3} \\ &= \frac{\text{E}\left(X^3\right)}{\left(\frac{1}{\lambda}\right)^3} - \frac{3\left(\frac{1}{\lambda}\right)\left(\frac{1}{\lambda}\right)^2 + \left(\frac{1}{\lambda}\right)^3}{\left(\frac{1}{\lambda}\right)^3} = \lambda^3 \cdot \text{E}\left(X^3\right) - \frac{\frac{3}{\lambda^3} + \frac{1}{\lambda^3}}{\frac{1}{\lambda^3}} \\ &= \lambda^3 \cdot \text{E}\left(X^3\right) - 4\end{aligned} \tag{5.7.893}$$

Thus, the remaining work is to compute $\mathrm{E}\left(X^3\right)$. To do this, we use the moment-generating function of the exponential distribution to calculate:

$$\mathrm{E}\left(X^3\right) = M'_X(0) \tag{5.7.894}$$

based on the relationship between raw moment and moment-generating function. First, we differentiate the moment-generating function of the exponential distribution (see page 933 where we have derived the moment generating function of the Gamma distribution and if we put $\alpha = 1$ we get that of the exponential distribution!):

$$M_X(t) = \frac{\lambda}{\lambda - t} = \lambda(\lambda - t)^{-1} \tag{5.7.895}$$

with respect to t. Using the chain rule gives:

$$M'_X(t) = -1 \cdot \lambda(\lambda - t)^{-2} \cdot (-1) = \lambda(\lambda - t)^{-2} \tag{5.7.896}$$

We continue using the chain rule to obtain the second derivative:

$$M''_X(t) = -2 \cdot \lambda(\lambda - t)^{-3} \cdot (-1) = 2\lambda(\lambda - t)^{-3} \tag{5.7.897}$$

Finally, one more application of the chain rule gives us the third derivative:

$$M'''_X(t) = -3 \cdot 2\lambda(\lambda - t)^{-4} \cdot (-1) = 6\lambda(\lambda - t)^{-4} = \frac{6\lambda}{(\lambda - t)^4} \tag{5.7.898}$$

Therefore:

$$\mathrm{E}\left(X^3\right) = M'''_X(0) = \frac{6\lambda}{(\lambda - 0)^4} = \frac{6\lambda}{\lambda^4} = \frac{6}{\lambda^3} \tag{5.7.899}$$

We now substitute this relation into the skewness, giving:

$$\gamma_1 = \lambda^3 \cdot \mathrm{E}\left(X^3\right) - 4 = \lambda^3 \cdot \left(\frac{6}{\lambda^3}\right) - 4 = 6 - 4 = 2 \tag{5.7.900}$$

Remark

We will see later that the exponential distribution is a special case of a more general distribution which is the chi-square distribution, the chi-square is also a special case of a more general distribution that is the Gamma distribution. This is a very important property used in the "Poisson test" for rare events (see also below).

5.7.6.15 Cauchy Distribution

Let X, Y be two independent random variables following a Normal reduced centered distribution (with zero mean and unit variance). Thus the density function is given for each variable by:

$$f_X(x) = f_Y(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{5.7.901}$$

The random variable:

$$T=\frac{X}{|Y|} \tag{5.7.902}$$

(the absolute value will be useful in an integral during a change of variable) follows a characteristic appearance named the "standard Cauchy distribution" (or "standard Cauchy law") or even "standard Lorentz law" (as we took as starting point the special case of two centred reduced Normal laws!).

Let us now determine its density function f. To do this, recall that f is determined by the (general) relation:

$$\forall t \in \mathbb{R}, P(T \leqslant t)=\int\limits_{-\infty}^{t} f(x)\mathrm{d}x \tag{5.7.903}$$

So (application of elementary differential calculus):

$$f(t)=\frac{\mathrm{d}}{\mathrm{d}t}P(T \leq t) \tag{5.7.904}$$

in the case where f is continuous.

Since X and Y are independent, the density function of the random vector is given by the third axioms of probabilities (see section Probabilities page 480):

$$(x,y)\mapsto f_X(x)\cdot f_Y(x) \tag{5.7.905}$$

therefore:

$$P(T\leqslant t)=P\left(\frac{X}{|Y|}<t\right)=P(X\leqslant t|Y|)=\int\limits_D f_X(x)\cdot f_Y(x)\mathrm{d}x\mathrm{d}y \tag{5.7.906}$$

where $D=\{(x,y)|x<t|y|\}$. This last integral becomes:

$$\int\limits_D f_X(x)\cdot f_Y(x)\mathrm{d}x\mathrm{d}y=\int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{t\cdot|y|} f_X(x)\cdot f_Y(x)\mathrm{d}x\mathrm{d}y \tag{5.7.907}$$

Let us make the following change of variables $x=u|y|$ in the inner integral. We obtain:

$$P(T\leqslant t)=\int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{t} f_X(u\cdot|y|)\cdot f_Y(y)|y|\mathrm{d}u\mathrm{d}y=\int\limits_{-\infty}^{t}\int\limits_{-\infty}^{+\infty} f_X(u\cdot|y|)\cdot f_Y(y)|y|\mathrm{d}y\mathrm{d}u \tag{5.7.908}$$

Therefore:

$$f(t)=\frac{\mathrm{d}}{\mathrm{d}t}P(T\leqslant t)=\int\limits_{-\infty}^{+\infty} f_X(t\cdot|y|)\cdot f_Y(y)|y|\mathrm{d}y=\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} e^{-\frac{y^2(t^2+1)}{2}}|y|\mathrm{d}y \tag{5.7.909}$$

Now the absolute value will be useful to write:

$$f(t)=\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} e^{-\frac{y^2(t^2+1)}{2}}|y|\mathrm{d}y=-\frac{1}{2\pi}\int\limits_{-\infty}^{0} e^{-\frac{y^2(t^2+1)}{2}}y\mathrm{d}y+\frac{1}{2\pi}\int\limits_{0}^{+\infty} e^{-\frac{y^2(t^2+1)}{2}}y\mathrm{d}y \qquad (5.7.910)$$

For the first integral we have:

$$\frac{1}{2\pi}\int\limits_{-\inf}^{0} e^{-\frac{y^2(t^2+1)}{2}}y\mathrm{d}y=\left.\frac{e^{-\frac{y^2(t^2+1)}{2}}}{t^2+1}\right|_{-\infty}^{0}=-\frac{1}{t^2+1}+\frac{1}{t^2+1}=0 \qquad (5.7.911)$$

It remains therefore only the second integral and making the change of variable $v=y^2$, we get:

$$f(t)=\frac{1}{2\pi}\int\limits_{0}^{+\infty} e^{-\frac{\nu(t^2+1)}{2}}\mathrm{d}\nu=-\left.\frac{e^{-\frac{\nu(t^2+1)}{2}}}{\pi(t^2+1)}\right|_{0}^{+\infty}=\frac{1}{\pi(t^2+1)} \qquad (5.7.912)$$

What we will denote thereafter (to respect the notations adopted so far):

$$\boxed{P(x)=\frac{1}{\pi(x^2+1)}} \qquad (5.7.913)$$

and that is simply the so named "standard Cauchy distribution" (as we took as starting point two centred reduced Normal laws!).

It is also effectively a distribution function because it verifies (see section Differential and Integral Calculus page 1534):

$$\int\limits_{-\infty}^{+\infty} P(x)\mathrm{d}x=\frac{1}{\pi}\int\limits_{-\infty}^{+\infty}\frac{1}{x^2+1}\mathrm{d}x=\frac{1}{\pi}(\arctan(+\infty)-\arctan(-\infty))=\frac{1}{\pi}\left(\frac{\pi}{2}-\left(-\frac{\pi}{2}\right)\right)=1 \qquad (5.7.914)$$

It is obvious that we get therefore for the cumulative distribution function:

$$\begin{aligned}P(X\leq x)&=\int\limits_{-\infty}^{x} P(x)\mathrm{d}x=\frac{1}{\pi}\int\limits_{-\infty}^{x}\frac{1}{x^2+1}\mathrm{d}x=\frac{1}{\pi}(\arctan(x)-\arctan(-\infty))=\\&=\frac{1}{\pi}\left(\arctan(x)-\left(-\frac{\pi}{2}\right)\right)=\frac{1}{\pi}\arctan(x)+\frac{1}{2}\end{aligned} \qquad (5.7.915)$$

Here is plot example of the Cauchy distribution:

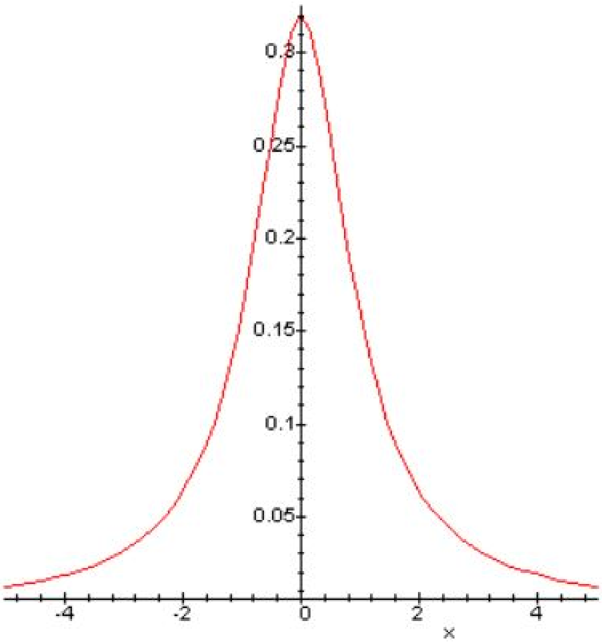


Figure 5.144: **Cauchy law (mass function)**

The Cauchy distribution has for expected mean:

$$\begin{aligned}\mu &= \int_{-\infty}^{+\infty} xP(x)\mathrm{d}x = \frac{1}{\pi}\int_{-\infty}^{+\infty}\frac{x}{1+x^2}\mathrm{d}x = \frac{1}{\pi}\left(\int_{-\infty}^{0}\frac{x}{1+x^2}\mathrm{d}x + \int_{0}^{+\infty}\frac{x}{1+x^2}\mathrm{d}x\right)\\ &= \frac{1}{2\pi}\left(\ln(1+x^2)\big|_{-\infty}^{0} + \ln(1+x^2)\big|_{0}^{+\infty}\right) = \frac{1}{2\pi}(-\ln(\infty)+\ln(\infty)) = 0\end{aligned} \tag{5.7.916}$$

Caution!!!! The above calculations do not give zero in facts because the subtraction of infinite is not zero but indeterminate! The Cauchy distribution therefore and strictly speaking does not admits an expected mean!

Thus, even if we can build a variance:

$$\begin{aligned}\sigma^2 &= \int_{-\infty}^{+\infty}(x-\mu)^2 f(x)\mathrm{d}x = \int_{-\infty}^{+\infty} x^2P(x)\mathrm{d}x = \frac{1}{\pi}\int_{-\infty}^{+\infty}\frac{x^2}{1+x^2}\mathrm{d}x = \frac{1}{\pi}\int_{-\infty}^{+\infty}\left(1-\frac{1}{1+x^2}\right)\mathrm{d}x\\ &= \frac{2}{\pi}\lim_{t\rightarrow+\infty}\int_{0}^{t}\left(1-\frac{1}{1+x^2}\right)\mathrm{d}x = 2\pi\lim_{t\rightarrow+\infty}(t-\arctan(t)) = +\infty\end{aligned} \tag{5.7.917}$$

this is absurd and does not exist strictly speaking as the mean doesn't exists...!

The Cauchy distribution is used a lot in financial engineering as it is heavy tailed and therefore a very good candidate to be more accurate in predicting extreme values at the opposite to the Normal distribution that has the tails decreasing to quick. Further the Cauchy distribution is a heavy tailed law with a support on $\mathbb{R}$ when the Pareto distribution (also heavy tailed) is defined only on $\mathbb{R}^+$.

Software such as JASP and the R package BayesFactor (Morey et al., 12018) rely on the Cauchy distribution as a prior for many of the Bayesian analyses.

The Cauchy distribution if one of the most famous distribution function that... we cannot found in the spreadsheet softwares like Microsoft® Excel®. To be able to get the closed form of the inverse Cauchy CDF we start from the CDF proven previously:

$$P(X \leq x) = \frac{1}{\pi}\arctan(x) + \frac{1}{2} \tag{5.7.918}$$

and therefore if we let:

$$y = \frac{1}{\pi}\arctan(x) + \frac{1}{2} \tag{5.7.919}$$

We immediately get the inverse CDF:

$$x = \tan\left(\pi\left(y - \frac{1}{2}\right)\right) \tag{5.7.920}$$

That is useful in finance as we know (see section Numerical Methods page 6710) to simulate a Cauchy variable when we use the inverse transform sampling:

$$X = \tan\left(\pi\left(\mathcal{U}_{[0,1]} - \frac{1}{2}\right)\right) \tag{5.7.921}$$

Using the random variable transformation (see page 633) we immediately derive the "general Cauchy distribution" or simply named the "Cauchy distribution" and given by:

$$\boxed{f(x) = \frac{1}{b\pi\left[1 + \left(\dfrac{x-a}{b}\right)^2\right]} = \frac{1}{\pi b}\frac{b^2}{(x-a)^2 + b^2}} \tag{5.7.922}$$

In Bayesian inference the scale parameter b is often denoted ω and named the "width parameter".

By default, the Cauchy distribution in JASP and in the BayesFactor of the R software package is centred around zero. At first, this may seem counter-intuitive, because this is the prior parameter for the alternative hypothesis. The reason for the centring is that it allows us to test a non-directional hypothesis. Given that the prior distribution is symmetrical and centred around zero, effect sizes of, say, greater than $+0.5$ or smaller than -0.5 are equally likely. The bell-shape of the Cauchy distribution further indicates a belief that smaller effect sizes are more likely than larger effect sizes. When we increase the width parameter ω, we increase the a priori probability of large effect sizes (though small effect sizes will continue to be more likely).

5.7.6.16 Beta Distribution

Let us first recall that the Euler Gamma function is defined by the relation (see section Differential and Integral Calculus page 1574):

$$\Gamma(z)=\int\limits_0^{+\infty} e^{-x}x^{z-1}\mathrm{d}x \tag{5.7.923}$$

We proved (see section Differential and Integral Calculus page 1574) that a non-trivial property of this function is:

$$\Gamma(z+1)=z\Gamma(z) \tag{5.7.924}$$

Let us now write:

$$\Gamma(a)\Gamma(b)=\lim_{R\rightarrow+\infty}\iint\limits_{A_R} e^{-x-y}x^{a-1}y^{b-1}\mathrm{d}x\mathrm{d}y \tag{5.7.925}$$

where:

$$A_R=\{(x,y)|x\geq 0,y\geq 0,x+y\leq R\} \tag{5.7.926}$$

By the change of variables:

$$\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}u-v\\v\end{pmatrix} \tag{5.7.927}$$

we get:

$$\Gamma(a)\Gamma(b)=\lim_{R\rightarrow+\infty}\iint\limits_{A_R} e^{-x-y}x^{a-1}y^{b-1}\mathrm{d}x\mathrm{d}y=\lim_{R\rightarrow+\infty}\int\limits_0^R e^{-u}\left(\int\limits_0^u (u-v)^{a-1}v^{b-1}\mathrm{d}v\right)\mathrm{d}u \tag{5.7.928}$$

For the internal integral we now use the substitution $v=ut, 0\leq t\leq 1$ and therefore we find:

$$\begin{aligned}\Gamma(a)\Gamma(b)&=\lim_{R\rightarrow+\infty}\int\limits_0^R e^{-u}\left(\int\limits_0^u (u-v)^{a-1}v^{b-1}\mathrm{d}v\right)\mathrm{d}u\\&=\lim_{R\rightarrow+\infty}\int\limits_0^R e^{-u}u^{a+b-1}\mathrm{d}u\int\limits_0^1(1-t)^{a-1}t^{b-1}\mathrm{d}t\overset{\text{def}}{=}B(a,b)\int\limits_0^{+\infty}e^{-u}u^{a+b-1}\mathrm{d}u\\&=B(a,b)\Gamma(a+b)\end{aligned} \tag{5.7.929}$$

Remark

For information, the beta function appears in Quantum Electrodynamics (QED) when calculating the one-loop Feynamn diagram for the regulation of the gauge boson self energy.

The function B that appears in the expression above is named "beta function[38]" and therefore we have:

$$\boxed{B(a,b)=B(b,a)=\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}} \tag{5.7.930}$$

Now that we have defined what we name the "beta function", consider the two parameters $a>0,b>0$ and consider also the special relation below as the "beta distribution" or "beta law" (there are several formulations of the beta distribution and a very important one is studied in detail in the section of Quantitative Management page 8074):

$$\boxed{P_{a,b}(x)=\frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}1_{]0,1[}} \tag{5.7.931}$$

where:

$$\boxed{B(a,b)=\int\limits_0^1 x^{a-1}(1-x)^{b-1}\mathrm{d}x} \tag{5.7.932}$$

We will also encounter this latter function in the context of Quantum Field Theory!

Historically, this function comes from Bayesian statistics (notably a special case treated by Laplace) of the ratio of the Binomial distribution to the sum of all possible outcomes of this same Binomial distribution such that:

$$P(x\mid p,q)=\frac{C_{p+q}^{p}x^p(1-x)^q}{\int C_{p+q}^{p}y^p(1-y)^q\mathrm{d}y}=\frac{x^p(1-x)^q}{\int y^p(1-y)^q\mathrm{d}y} \tag{5.7.933}$$

Remark

The integral rewritten as following:

$$B(a+1,b+1)=\int\limits_0^1 x^a(1-x)^b\mathrm{d}x \tag{5.7.934}$$

is named the "Chebyshev integral".

The integral generalized as following:

$$B(x,a,b)=\int\limits_0^x t^{a-1}(1-t)^{b-1}\mathrm{d}t \tag{5.7.935}$$

is named the "incomplete beta function". While best known for its applications in Statistics, it is also widely used in many other fields such as actuarial science, economics, finance, survival analysis, life testing and telecommunications.

---

[38] Historically this function comes from the calculation of $m!n!=\lim_{a^2\rightarrow+\infty}\int\limits_0^{a^2} e^{-u}u^m\mathrm{d}u\int\limits_0^{a^2} e^{-v}v^n\mathrm{d}v$, using the Euler-Gamma function and that leads to $B(m+1,n+1)\equiv 2\int_0^{\pi/2}\cos^{2m+1}(\theta)\sin^{2n+1}(\theta)\mathrm{d}\theta=\frac{m!n!}{(m+n+1)!}$.

We first check that $P_{a,b}(x)$ that is effectively a distribution function (without getting into too much details ...):

$$\begin{aligned}\int\limits_{-\infty}^{+\infty} P_{a,b}(x)x &= \int\limits_{-\infty}^{+\infty} \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}\cdot 1_{]0,1[}\mathrm{d}x = \int\limits_{-\infty}^{+\infty} \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}\cdot 1_{]0,1[}\mathrm{d}x \\ &= \frac{1}{B(a,b)}\int\limits_{0}^{1} x^{a-1}(1-x)^{b-1}\mathrm{d}x = \frac{B(a,b)}{B(a,b)} = 1\end{aligned} \tag{5.7.936}$$

Let us now calculate the expected mean:

$$\begin{aligned}\mu &= \int\limits_{-\infty}^{+\infty} xP_{a,b}(x)\mathrm{d}x = \frac{1}{B(a,b)}\int\limits_{0}^{1} x^{a}(1-x)^{b-1}\mathrm{d}x \\ &= \frac{B(a+1,b)}{B(a,b)} = \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)}\cdot\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} = \frac{a}{a+b}\end{aligned} \tag{5.7.937}$$

by using the relation:

$$\Gamma(z+1) = z\Gamma(z) \tag{5.7.938}$$

and its variance:

$$\begin{aligned}\sigma^2 &= \int\limits_{-\infty}^{+\infty} (x-\mu)^2 f(x)\mathrm{d}x = \frac{1}{B(a,b)}\int\limits_{0}^{1} (x-\mu)^2 x^{a-1}(1-x)^{b-1}\mathrm{d}x \\ &= \frac{1}{B(a,b)}\left(\int\limits_{0}^{1} x^{a+1}(1-x)^{b-1}\mathrm{d}x - 2\mu\int\limits_{0}^{1} x^{a}(1-x)^{b-1}\mathrm{d}x + \mu^2\int\limits_{0}^{1} x^{a-1}(1-x)^{b-1}\mathrm{d}x\right) \\ &= \frac{B(a+2,b)-2\mu^2 B(a,b)+\mu^2 B(a+1,b)}{B(a,b)} = \frac{B(a+2,b)-\mu^2 B(a,b)}{B(a,b)}\end{aligned} \tag{5.7.939}$$

As we know that:

$$\Gamma(z+1) = z\Gamma(z) \quad \text{and} \quad B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{5.7.940}$$

we find:

$$B(a+2,b) = \frac{\mu\cdot(a+1)}{a+b+1} \tag{5.7.941}$$

and therefore:

$$\sigma^2 = \frac{\mu\cdot(a+1)}{a+b+1} - \mu^2 = \frac{ab}{(a+b)^2(a+b+1)} \tag{5.7.942}$$

Examples of plots of the beta distribution function for $(a,b) = (0.1,0.5)$ in red, $(a,b) = (0.3,0.5)$ in green, $(a,b) = (0.5,0.5)$ in black, $(a,b) = (0.8,0.8)$ in blue, $(a,b) = (1,1)$

in magenta, $(a,b)=(1,1.5)$ in cyan, $(a,b)=(1,2)$ in gray, $(a,b)=(1.5,2)$ in turquoise, $(a,b)=(2,2)$ in yellow,$(a,b)=(3,3)$ in gold color:

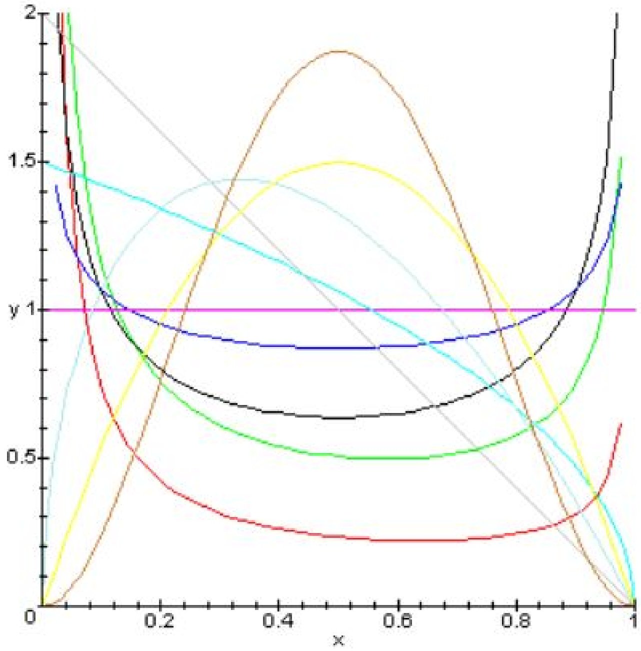


Figure 5.145: **Some Beta law mass functions**

Here is a plot example of the beta distribution and cumulative distribution for the parameters $(a,b)=(2,3)$:

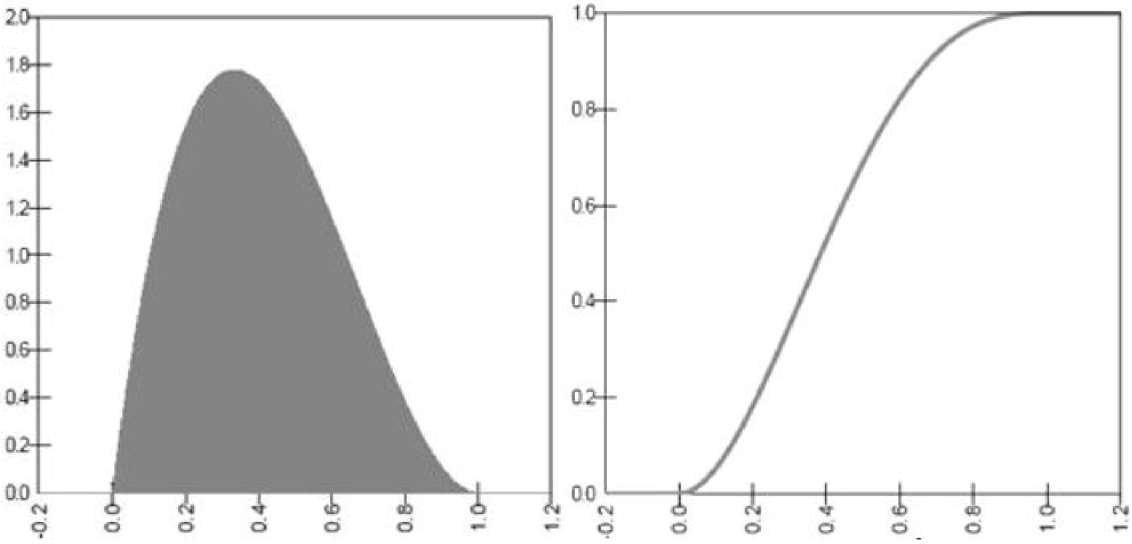


Figure 5.146: **Beta law (mass and cumulative distribution function)**

The fact that the beta law is one of the rare distribution functions whose support is between $]0,1[$ explains its frequent use in Bayesian statistics as an a priori law of the distribution of a proportion!

Remark

Notice the important special case that when $a, b = 1$, the beta distribution, reduce to a uniform continuous distribution! Indeed:

$$P_{a,b} = P_{1,1} = \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)} p^{1-1}(1-p)^{1-1} 1_{]0,1[} = 1 \cdot 1_{]0,1[} \tag{5.7.943}$$

In general, if you want a beta prior distribution with parameters a and b but you only know μ and σ^2, they can be found via the both relations:

$$\begin{aligned} \mu &= \frac{a}{a+b} \\ \sigma^2 &= \frac{ab}{(a+b)^2(a+b+1)} \end{aligned} \tag{5.7.944}$$

We can solve the first one as follows for b:

$$b = \frac{a(1-\mu)}{\mu} \tag{5.7.945}$$

Substituting that latter into the above σ^2 yields:

$$\begin{aligned} \sigma^2 = \frac{a^2(1-\mu)}{\mu(a+b)^2(a+b+1)} &= \frac{a^2(1-\mu)}{\mu(a[1+(1-\mu)/\mu])^2(a+a(1-\mu)/\mu+1)} \\ &= \frac{(1-\mu)}{\mu(1+(1-\mu)/\mu)^2(a/\mu+1)} \end{aligned} \tag{5.7.946}$$

Rearranging that latter yields:

$$a = \frac{1-\mu-\sigma^2\mu^3}{\sigma^2\mu^2} \tag{5.7.947}$$

and therefore:

$$b = \frac{(1-\mu)\left[1-\mu-\sigma^2\mu^3\right]}{\sigma^2\mu^3} \tag{5.7.948}$$

5.7.6.16.1 Regularized Incomplete beta function integral

What follows should better be placed in the section of Functional Analysis however it has seem more pedagogical to us to place it just here for obvious reasons.

Consider the "regularized incomplete Beta function $I_x(a,b)$" (or "regularized beta function" for short) defined in terms of the incomplete beta function and the complete beta function as (for $\alpha > 0$):

$$\boxed{I_x(a,b) = \frac{B(x,a,b)}{B(a,b)} = \frac{\displaystyle\int_0^x t^{a-1}(1-t)^{b-1}\,\mathrm{d}t}{\displaystyle\int_0^1 t^{a-1}(1-t)^{b-1}\,\mathrm{d}t}} \tag{5.7.949}$$

where for recall $a > 0$ and $b > 0$. This function will be useful to us in some Bayesian A/B statistics (see section Statistics page 996) but also in String Theory (see section String Theory page 6099)!

Notice the canonical values that are immediate:

$$I_0(a,b)=0 \qquad I_1(a,b)=0 \tag{5.7.950}$$

We want to prove the recursive property:

$$I_x(a+1,b)=I_x(a,b)-\frac{x^a\,(1-x)^b}{a\,B\,(a,b)} \tag{5.7.951}$$

Remark

This also means that:

$$I_x(a-1,b)=I_x(a,b)-\frac{x^{a-1}\,(1-x)^b}{(a-1)B\,(a-1,b)} \tag{5.7.952}$$

but for $a>1$ (and still $b>0$).

Proof 5.47.6. Using the following Beta function (see page 761) property:

$$B(a+1,b)=\frac{a!(b-1)!}{(a+b)!}=\frac{(a-1)!(b-1)!}{(a+b-1)!}\frac{a}{a+b}=\frac{aB(a,b)}{a+b} \tag{5.7.953}$$

Remark

From the above relation and because of symmetry of the beta function, the following three results follow almost immediately:

$$\begin{aligned}B(a,b+1)&=\frac{bB(a,b)}{a+b}\\ B(a+1,b)+B(a,b+1)&=B(a,b)\\ B(a+1,b-1)&=\frac{aB(a,b)}{b-1}\end{aligned} \tag{5.7.954}$$

We compute the derivative:

$$I'_x(a+1,b)=\frac{\mathrm{d}I_x(a+1,b)}{\mathrm{d}t}=\frac{x^a(1-x)^{b-1}}{B(a+1,b)}=\frac{x^a(1-x)^{b-1}}{B(a,b)a/(b+a)}=(a+b)\frac{x^a(1-x)^{b-1}}{B(a,b)a} \tag{5.7.955}$$

Now let:

$$f(a,b,x)=I_x(a+1,b)-I_x(a,b)+\frac{x^a\,(1-x)^b}{a\,B\,(a,b)} \tag{5.7.956}$$

Then $f(a,b,0)=0$ and:

$$\begin{aligned}
f'(a,b,x) &= I'_x(a+1,b)-I'_x(a,b)+\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{x^a\,(1-x)^b}{a\,B\,(a,b)}\right)\\
&=(a+b)\frac{x^a(1-x)^{b-1}}{B(a,b)a}-\frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}+\frac{x^{a-1}(1-x)^b}{B(a,b)}-\frac{bx^a(1-x)^{b-1}}{aB(a,b)}\\
&=\frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}\left(\frac{(a+b)x}{a}-1+1-x-\frac{bx}{a}\right)=0
\end{aligned}\tag{5.7.957}$$

therefore $f\equiv 0$ and the recursive property is proven.

■ Q.E.D.

So we have proved that:

$$I_x(a-1,b)=I_x(a,b)-\frac{x^{a-1}\,(1-x)^b}{(a-1)B\,(a-1,b)}\tag{5.7.958}$$

And notice that we have the limit case:

$$I_x(1,b)=\frac{\displaystyle\int\limits_0^x t^0\,(1-t)^{b-1}\,\mathrm{d}t}{\displaystyle\int\limits_0^1 t^0\,(1-t)^{b-1}\,\mathrm{d}t}=\frac{-\left.\dfrac{(1-t)^b}{b}\right|_0^x}{-\left.\dfrac{(1-t)^b}{b}\right|_0^1}=\frac{-\dfrac{(1-x)^b}{b}+\dfrac{1}{b}}{\dfrac{1}{b}}=1-(1-x)^b\tag{5.7.959}$$

So now recursively iterating:

$$I_x(a-1,b)=I_x(a,b)-\frac{x^{a-1}\,(1-x)^b}{(a-1)B\,(a-1,b)}\tag{5.7.960}$$

Until the case:

$$I_x(1,b)=1-(1-x)^b\tag{5.7.961}$$

$$I_x(a,b)=1-(1-x)^b-\sum_{j=1}^{a-1}\frac{x^{a-j}(1-x)^b}{(a-j)B(a-j,b)}\tag{5.7.962}$$

Subsuming the zeroth term into the sum:

$$I_x(a,b)=1-\sum_{i=0}^{a-1}\frac{x^i(1-x)^b}{(b+i)B(1+i,b)}\tag{5.7.963}$$

And we will stop here for the study of this function as this is all what we need in the actual state of this textbook for practical applications!

5.7.6.17 Gamma Distribution

The Euler Gamma function being known, consider two parameters $a > 0, \lambda > 0$ and let us define the "Gamma distribution" (or "Gamma law") as given by the relation (density function):

$$\boxed{P_{a,\lambda}(x) = \frac{x^{a-1}e^{-\lambda x}}{\int\limits_{0}^{+\infty} x^{a-1}e^{-\lambda x}\mathrm{d}x} \cdot 1_{]0,+\infty[}} \tag{5.7.964}$$

By the change of variables $t = \lambda x$ we obtain:

$$\int\limits_{0}^{+\infty} x^{a-1}e^{-\lambda x}\mathrm{d}x = \frac{\Gamma(a)}{\lambda^a} \tag{5.7.965}$$

and we can then write the relation in a more conventional form that we find frequently in the literature:

$$\boxed{P_{a,\lambda}(x) = \frac{x^{a-1}e^{-\lambda x}}{\Gamma(a)}\lambda^a \cdot 1_{]0,+\infty[}} \tag{5.7.966}$$

and it is under this notation that we find this distribution function in Microsoft® Excel® 11.8346 under the name `GAMMADIST( )` and its inverse by `GAMMAINV( )`. Notice that the previous relation is also often denoted:

$$\Gamma(a,\lambda) = \frac{x^{a-1}e^{-\lambda x}}{\Gamma(a)}\lambda^a \cdot 1_{]0,+\infty[} \tag{5.7.967}$$

Let us now see a simple property of the Gamma distribution that will be partially useful for the study of the Welch statistical test (see page 1020). First recall that we have shown above that:

$$\forall y \in \mathbb{R} \quad f_Y(y) = \frac{1}{c^{te}} f_X\left(\frac{y}{c^{te}}\right) \tag{5.7.968}$$

Let us write $Y = c^{te}X$, then we have immediately:

$$f_Y(y) = \frac{y^{a-1}e^{\lambda \frac{y}{c^{te}}}}{\Gamma(a)}\left(\frac{\lambda}{c^{te}}\right) 1_{]0,+\infty[} \tag{5.7.969}$$

So the multiplication by a constant of random variable that follows a Gamma distribution has only for effect of dividing the parameter λ by the same constant. This is the reason why λ is named "scale parameter".

If $a \in \mathbb{N}$, the Gamma distribution at the denominator becomes (see section Differential and Integral Calculus page 1574) the factorial $(a-1)!$. The Gamma function can then be written:

$$P_{a,\lambda} = \frac{x^{a-1}\lambda^a e^{-\lambda x}}{(a-1)!} = \frac{(x\lambda)^{a-1}}{(a-1)!}\lambda e^{-\lambda x} \tag{5.7.970}$$

This particular notation of the Gamma distribution is named the "Erlang distribution" that we find naturally in the theory of queues and that is very important in practice!

Remark

If $a = 1$ then $\Gamma(a) = 1$ and $x^{a-1} = 1$ and we fall back on the exponential distribution.

Then we check with a similar reasoning to this of the beta distribution that $P_{a,\lambda}(x)$ is a distribution function:

$$\int\limits_{-\infty}^{+\infty} P_{a,\lambda} \mathrm{d}x = 1 \tag{5.7.971}$$

Examples of plots of the beta distribution function for $(a, \lambda) = (0.5, 1)$ in red, $(a, \lambda) = (1, 1)$ in green, $(a, \lambda) = (2, 1)$ in black, $(a, \lambda) = (4, 2)$ in blue, $(a, \lambda) = (16, 8)$ in magenta:

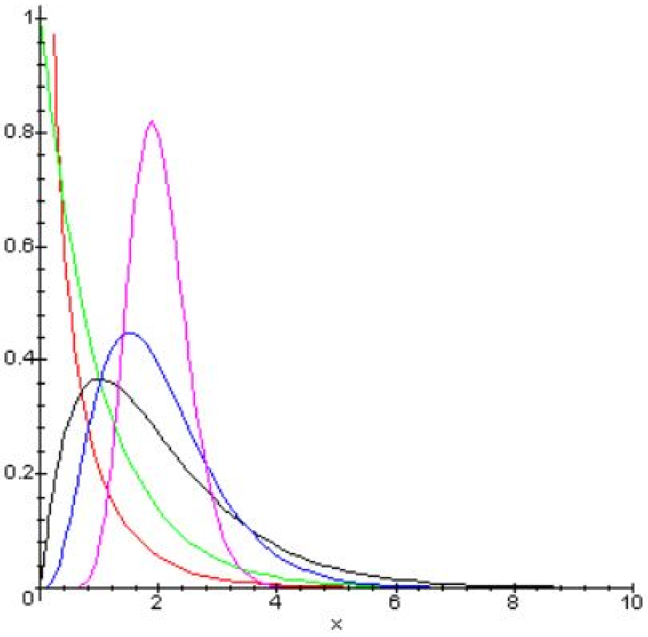


Figure 5.147: **Some Gamma law mass functions**

and a plot example of the Gamma distribution and cumulative distribution for the parameters $(a, \gamma) = (4, 1)$:

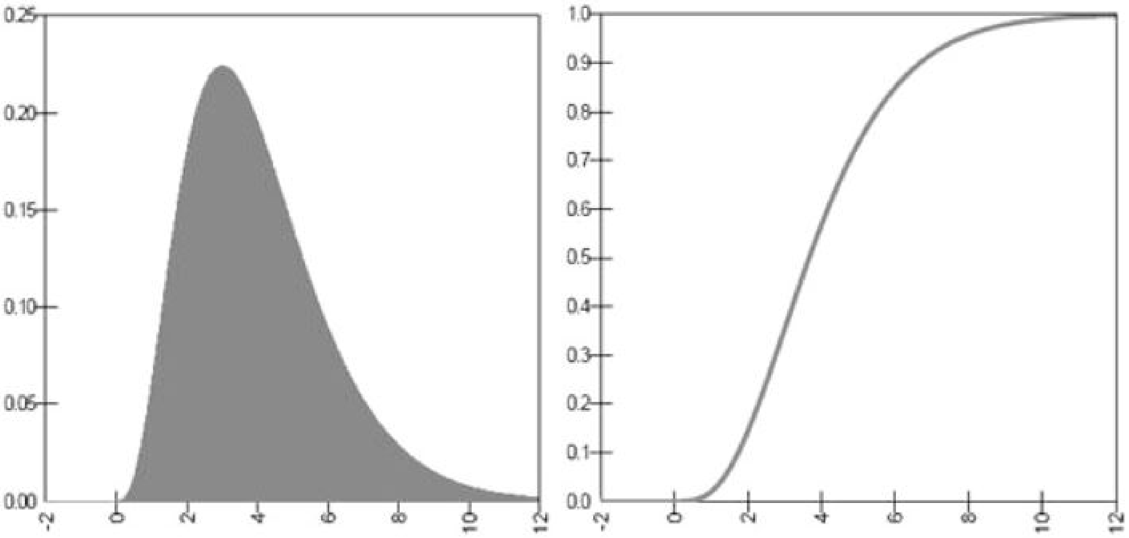


Figure 5.148: Gamma law (mass and cumulative distribution function)

Let X be a positive random variable following a Gamma distribution:

$$X \sim \text{Gam}(a,b) \tag{5.7.972}$$

Then, the cumulative distribution function of X is:

$$F_X(x)=\frac{\gamma(a,bx)}{\Gamma(a)} \tag{5.7.973}$$

Indeed, the probability density function of the Gamma distribution is for reminder:

$$f_X(x)=\frac{b^a}{\Gamma(a)}x^{a-1}\exp[-bx] \tag{5.7.974}$$

Thus, the cumulative distribution function is:

$$\begin{aligned}F_X(x)&=\int\limits_0^x \text{Gam}(z;a,b)\mathrm{d}z\\&=\int\limits_0^x \frac{b^a}{\Gamma(a)}z^{a-1}\exp[-bz]\mathrm{d}z\\&=\frac{b^a}{\Gamma(a)}\int\limits_0^x z^{a-1}\exp[-bz]\mathrm{d}z\end{aligned} \tag{5.7.975}$$

Substituting $t=bz$, i.e. $z=t/b$, this becomes:

$$\begin{aligned}F_X(x)&=\frac{b^a}{\Gamma(a)}\int\limits_{b\cdot 0}^{b\cdot x}\left(\frac{t}{b}\right)^{a-1}\exp\left[-b\left(\frac{t}{b}\right)\right]\mathrm{d}\left(\frac{t}{b}\right)\\&=\frac{b^a}{\Gamma(a)}\cdot\frac{1}{b^{a-1}}\cdot\frac{1}{b}\int\limits_0^{bx} t^{a-1}\exp[-t]\mathrm{d}t\\&=\frac{1}{\Gamma(a)}\int\limits_0^{b\cdot x} t^{a-1}\exp[-t]\mathrm{d}t\end{aligned} \tag{5.7.976}$$

With the definition of the "lower incomplete gamma function":

$$\gamma(s,x)=\int_0^x t^{s-1}\exp[-t]\mathrm{d}t \tag{5.7.977}$$

we arrive at the final result given earlier above:

$$F_X(x)=\frac{\gamma(a,bx)}{\Gamma(a)} \tag{5.7.978}$$

The Gamma function has also for expected mean:

$$\mu=\int\limits_{-\infty}^{+\infty} xf(x)\mathrm{d}x=\frac{\lambda^a}{\Gamma(a)}\int\limits_{-\infty}^{+\infty} x^a e^{-\lambda x}\mathrm{d}x=\frac{\lambda^a}{\Gamma(a)}\frac{\Gamma(a+1)}{\lambda^{a+1}}\frac{\lambda^a}{\Gamma(a)}\frac{a\Gamma(a)}{\lambda^{a+1}}=\frac{a}{\lambda} \tag{5.7.979}$$

and for variance:

$$\begin{aligned}\sigma^2&=\int\limits_{-\infty}^{+\infty}(x-\mu)^2 f(x)\mathrm{d}x=\frac{\lambda^a}{\Gamma(a)}\int\limits_{-\infty}^{+\infty}(x-\mu)^2x^{a-1}e^{-\lambda x}\mathrm{d}x\\&=\frac{\lambda^a}{\Gamma(a)}\left(\int\limits_{-\infty}^{+\infty}x^{a+1}e^{-\lambda x}x-2\mu\int\limits_{-\infty}^{+\infty}x^a e^{-\lambda x}x+\mu^2\int\limits_{-\infty}^{+\infty}x^{a-1}e^{-\lambda x}\mathrm{d}x\right)\\&=\frac{\lambda^a}{\Gamma(a)}\left(\frac{\Gamma(a+2)}{\lambda^{a+2}}+\mu^2\frac{\Gamma(a)}{\lambda^a}-2\mu\frac{\Gamma(a+1)}{\lambda^{a+1}}\right)\\&=\frac{1}{\lambda^2\Gamma(a)}\left(\Gamma(a+2)+a^2\Gamma(a)-2a\Gamma(a+1)\right)\\&=\frac{1}{\lambda^2\Gamma(a)}\left((a+1)a\Gamma(a)+a^2\Gamma(a)-2a^2\Gamma(a)\right)=\frac{a}{\lambda^2}\end{aligned} \tag{5.7.980}$$

Let us now prove a property of the Gamma distribution that will permit us later in this chapter, during our study of the analysis of variance and confidence intervals based on small samples, another extremely important property of the Chi-square distribution.

As we know, the distribution function of a random variable following a Gamma function of parameters $a,\lambda>0$ is:

$$P_{a,\gamma}(x)=f(x)=\frac{e^{-\lambda x}x^{a-1}}{\Gamma(a)}\lambda^a 1_{[0,+\infty]} \tag{5.7.981}$$

with (see section Differential and Integral Calculus page 1574) the Euler Gamma function:

$$\Gamma(a)=\int\limits_0^{+\infty} e^{-x}x^{a-1}\mathrm{d}x \tag{5.7.982}$$

Moreover, when a random variable follows a Gamma function we often notice it in the following way:

$$X=\gamma(a,\lambda) \tag{5.7.983}$$

Let X, Y be two independent variables. We will prove that if $X = \gamma(p, \lambda)$ and $Y = \gamma(q, \lambda)$, hence with the same scale parameter, then:

$$X + Y = \gamma(p + q, \lambda) \tag{5.7.984}$$

We write f the density function of the pair $X, Y, f(x)$ the density function of X and f_Y the density function of Y. Because X and Y are independent, we have:

$$f(x, y) = f_X(x) \cdot f_Y(y) \tag{5.7.985}$$

for all $x, y > 0$.

Let $Z = X + Y$. The distribution function of Z is therefore:

$$F(z) = P(Z \leqslant z) = P(X + Y \leqslant z) = \iint\limits_D f(x, y)\mathrm{d}x\mathrm{d}y \tag{5.7.986}$$

where $D = \{(x, y)|x + y \leq z\}$.

Remark

As we already know we name such a calculation a "convolution" and statisticians often have to handle such entities because they work on many random variables that they have to sum or even to multiply.

Simplifying:

$$F(z) = \int\limits_{-\infty}^{+\infty} \int\limits_{-\infty}^{z-x} f_X(x)f_Y(y)\mathrm{d}x\mathrm{d}y \tag{5.7.987}$$

We perform the following change of variable $x = x, y = s - x$. The Jacobian is therefore (see section Differential and Integral Calculus page 1524):

$$J = \begin{vmatrix} \dfrac{\partial x}{\partial x} & \dfrac{\partial x}{\partial s} \\ \dfrac{\partial y}{\partial x} & \dfrac{\partial y}{\partial s} \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ -1 & 0 \end{vmatrix} = 1 \tag{5.7.988}$$

Therefore with the new integration limits $s = x + y = x + (z - x) = z$ we have:

$$F(z) = \int\limits_{-\infty}^{+\infty} \int\limits_{-\infty}^{z} f_X(x)f_Y(s - x)\mathrm{d}s\mathrm{d}x = \int\limits_{-\infty}^{z} \int\limits_{-\infty}^{+\infty} f_X(x)f_Y(s - x)\mathrm{d}x\mathrm{d}s \tag{5.7.989}$$

If we denote by g the density function Z we have:

$$F(z) = \int\limits_{-\infty}^{z} \int\limits_{-\infty}^{+\infty} f_X(x)f_Y(s - x)\mathrm{d}x\mathrm{d}s = \int\limits_{-\infty}^{z} g(s)\mathrm{d}s \tag{5.7.990}$$

Then it follows:

$$g(s) = \int\limits_{-\infty}^{+\infty} f_X(x)f_Y(s - x)\mathrm{d}x \tag{5.7.991}$$

f_X and f_Y being null when the argument is negative, we can change the limits of integration:

$$g(s)=\int_0^s f_X(x)f_Y(s-x)\mathrm{d}x \quad \text{for} \quad s\geqslant 0 \tag{5.7.992}$$

Let us calculate g:

$$g(s)=\dfrac{\lambda^{p+q}e^{-\lambda s}}{\Gamma(p)\Gamma(q)}\int_0^s x^{p-1}(s-x)^{q-1}\mathrm{d}x \tag{5.7.993}$$

After the change of variable $x=st$ we obtain:

$$g(s)=\dfrac{\lambda^{p+q}e^{-\lambda s}}{\Gamma(p)\Gamma(q)}s^{p+q-1}\int_0^1 t^{p-1}(1-t)^{q-1}\mathrm{d}t=\dfrac{\lambda^{p+q}e^{-\lambda s}}{\Gamma(p)\Gamma(q)}s^{p+q-1}B(p,q) \tag{5.7.994}$$

where B is the beta function we saw earlier in our study of the beta distribution. But we have also proved the relation:

$$B(p,q)=\dfrac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} \tag{5.7.995}$$

Therefore:

$$B(p,q)=\dfrac{\lambda^{p+q}e^{-\lambda s}}{\Gamma(p+q)}s^{p+q-1} \tag{5.7.996}$$

More explicitly:

$$g(x,y)=\dfrac{\lambda^{p+q}e^{-\lambda(x+y)}}{\Gamma(p+q)}(x+y)^{p+q-1} \tag{5.7.997}$$

Which finally gives us:

$$P(Z\leqslant z)=P(X+Y\leqslant z)=\int_0^s g(s)\mathrm{d}s=\int_0^s\dfrac{\lambda^{p+q}e^{-\lambda s}}{\Gamma(p+q)}s^{p+q-1}\mathrm{d}s \tag{5.7.998}$$

This shows that if two random variables follow a Gamma distribution then their sum will also follow a Gamma distribution with parameters:

$$X+Y=\gamma(p+q,\lambda) \tag{5.7.999}$$

So the Gamma distribution is stable by addition as are all distribution arising from the Gamma distribution that we will see below.

5.7.6.17.1 Generalized Gamma Distribution

The generalized gamma distribution is a continuous probability distribution with three parameters. It is a generalization of the two-parameter gamma distribution. Since many distributions commonly used for parametric models in survival analysis (such as the exponential distribution, the Weibull distribution and the Gamma distribution, and log-normal) are special cases of the generalized gamma, it is sometimes used to determine which parametric model is appropriate for a given set of data.

Therefore let us notice that if we write after trials and errors the following density function named "generalized Gamma law":

$$\boxed{f(x)=\frac{\eta}{\alpha\Gamma(\kappa)}\left(\frac{x}{\alpha}\right)^{\kappa\eta-1}e^{-\left(\frac{x}{\alpha}\right)^{\eta}}} \tag{5.7.1000}$$

with $x>0$, $\alpha>0$, $\eta>0$, $\kappa>0$.

Then, for $\kappa=1$ we fall back on the density function of the Weibull distribution (see section Industrial Engineering page 8736) that is with our own notations of the corresponding section is given by:

$$f(x)=\frac{\alpha}{\beta}\frac{x^{\alpha-1}}{\beta^{\alpha-1}}e^{-\frac{x^\alpha}{\beta^\alpha}}=\frac{\alpha}{\beta^\alpha}x^{\alpha-1}e^{-\left(\frac{\alpha}{\beta}\right)^\alpha} \tag{5.7.1001}$$

For $\eta=1$, we fall back the Gamma density function just introduced before:

$$f(x)=\frac{x^{a-1}e^{-\lambda x}}{\Gamma(a)}\lambda^a\cdot 1_{]0,+\infty[} \tag{5.7.1002}$$

For $\kappa=1$ and $\eta=1$, we fall back on the exponential distribution also seen previously:

$$f(x)=\lambda e^{\lambda x}\cdot 1_{[0,+\infty[} \tag{5.7.1003}$$

and finally for $\eta\rightarrow 0$, $\kappa\rightarrow+\infty$ we fall back on a log-Normal distribution after developing the limits using the Stirling, Hospital and Taylor techniques (see section Numerical Methods page 6301, Differential and Integral Calculus page 1461, Sequences and Series page 1696):

$$f(x)=\frac{1}{\sigma x\sqrt{2\pi}}e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}} \tag{5.7.1004}$$

As always, on request we can detail the developments!

5.7.6.18 Chi-Square (Pearson) Distribution

The "chi-square distribution" (also named "chi-square law" or "Pearson law") has a very important place in the industrial practice for some common hypothesis tests (see far below...) and is by definition only a particular case of the Gamma distribution in the case where $a=k/2$ and $\lambda=1/2$, when k is a positive integer:

$$\boxed{P_k(x)=\frac{1}{2^{\frac{k}{2}}\Gamma(k/2)}x^{\frac{k}{2}-1}e^{-\frac{x}{2}}\cdot 1_{[0,+\infty[}} \tag{5.7.1005}$$

Remark

We will also prove later during our study of the confidence interval on the variance with known mean (see page 857), the very important property that if X follows a Normal standard distribution $\mathcal{N}(0,1)$ then $X^2=\chi_1^2$.

This relation that connects the chi-square distribution with the Gamma distribution is important in the in Microsoft® Excel® 11.8346 as the function `CHIDIST( )` returns the confidence level and not the distribution function. Then you must use the function `GAMMADIST()` with the parameters given above (except that you must take the inverse of $1/2$: also 2 as parameter) to get the distribution and cumulative functions.

The reader who wishes to check that the Chi-square distribution is only a special case of the Gamma distribution can write in Microsoft® Excel® 14.0.6123:

```
=CHISQ.DIST(2*x,2*k,TRUE)
=GAMMA.DIST(x,k,1,TRUE)
```

All calculations made previously still apply and we get immediately:

$$\mu=k,\sigma^2=2k \tag{5.7.1006}$$

Examples of plots of the chi-squared distribution function for $k=1$ in red, $k=3$ in green, in black, $k=4$ in blue:

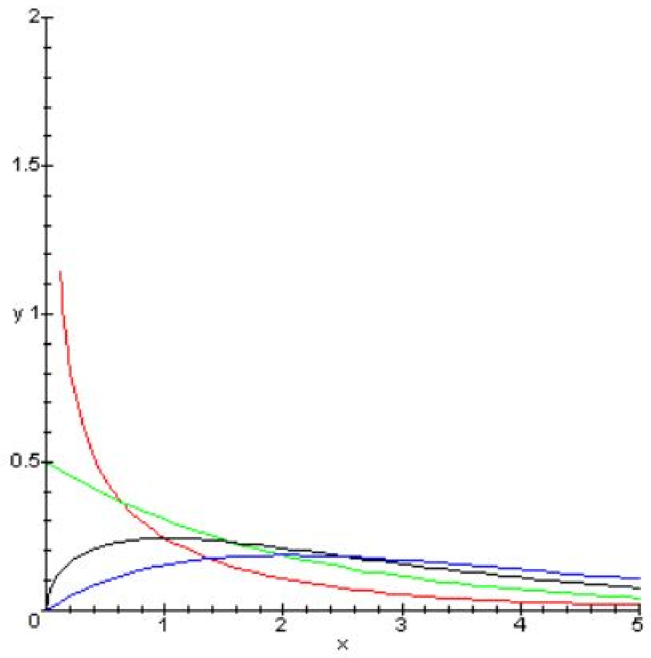


Figure 5.149: **Some χ^2 law mass functions**

and a plot example of the chi-squared distribution and cumulative distribution for the parameter $k=2$ (for more plots about this distribution see further below on page 388):

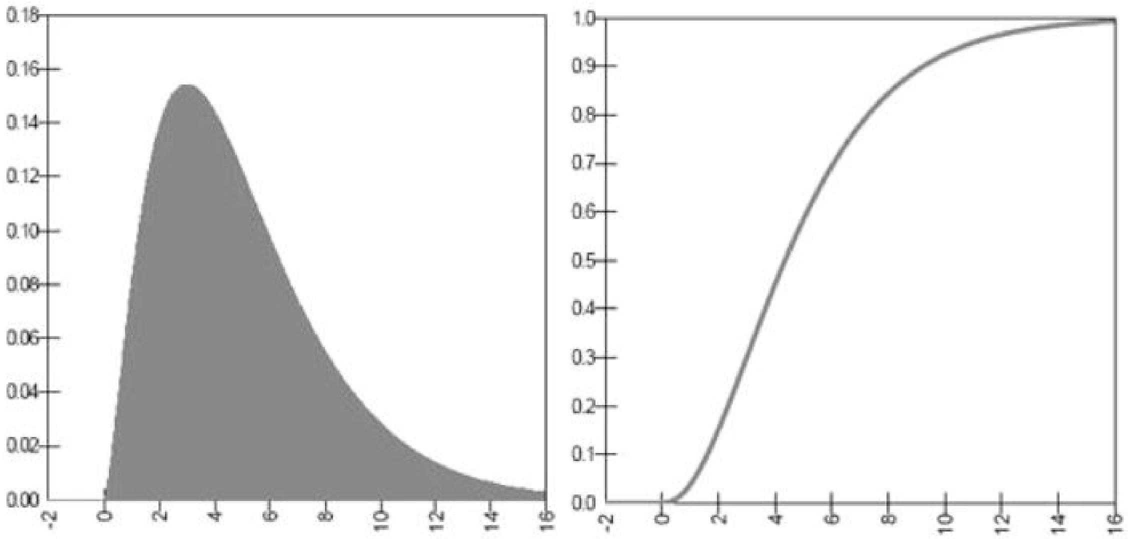


Figure 5.150: χ^2 **law (mass and cumulative distribution function)**

In the literature, it is traditional to write:

$$X=\chi_k^2 \quad \text{or} \quad X=\chi^2(k) \tag{5.7.1007}$$

to indicate that the distribution of the random variable X is a chi-square distribution. Furthermore it is common to name the parameter k "degrees of freedom" and abbreviate it "df".

The χ^2 distribution is therefore a special case of the Gamma distribution and by taking $k=2$ we also find the exponential distribution (see above) for $\lambda=1/2$:

$$P_2(x)=\frac{1}{2^1\Gamma(1)}x^{2/2-1}e^{-x/2}\cdot 1_{[0,+\infty[}=\frac{1}{2}e^{-\frac{x}{2}}1_{[0,+\infty[} \tag{5.7.1008}$$

Moreover, since (see section Differential and Integral Calculus page 1574):

$$\Gamma\left(\frac{1}{2}\right)=\sqrt{\pi} \tag{5.7.1009}$$

the χ^2 distribution with k equal to unity can be written as:

$$P_1(x)=\frac{1}{2^{1/2}\Gamma(1/2)}x^{1/2-1}e^{-x/2}\cdot 1_{[0,+\infty[}=\frac{1}{\sqrt{2\pi x}}e^{-x/2}\cdot 1_{[0,+\infty[} \tag{5.7.1010}$$

Finally, let us finish with a fairly large property in the field of statistical tests that we will investigate a little further and particularly for confidence intervals of rare events and the famous Fisher method for multiple p-value null hypothesis tests (see page 948). Indeed, the reader can check in a spreadsheet software like Microsoft® Excel® 14.0.6123 that we have:

=POISSON.DIST($x\in\mathbb{N},\mu$,TRUE)
=1-CHISQ.DIST($2\mu,2(x+1)$,TRUE)
=1-GAMMA.DIST($2\mu,x+1$,TRUE)
=1-EXPON.DIST($x,0.5$,TRUE)

So we need to prove this relation between law χ^2 and Poisson distributions. See it starting from the Gamma distribution:

$$P_{a,\lambda}(x)=\frac{x^{a-1}e^{-\lambda x}}{\Gamma(a)}\lambda^a 1_{]0,+\infty[} \tag{5.7.1011}$$

If we write $\lambda=1/2$ and $a=k/2$ then we have the χ^2 distribution with k degrees of freedom:

$$P_k(x)=\frac{1}{2^{k/2}\Gamma(k/2)}x^{k/2-1}e^{-x/2}1_{]0,+\infty[} \tag{5.7.1012}$$

Now remember that we have seen in the section Sequences and Series (page 1703), the following Taylor (Maclaurin) series from order $n-1$ around 0 to λ with integral rest:

$$e^{\lambda}=\sum_{k=0}^{n-1}\frac{\lambda^k}{k!}e^x(0)+\int_0^{\lambda}e^n(t)\frac{(\lambda-1)^{n-1}}{(n-1)!}\mathrm{d}t=\sum_{k=0}^{n-1}\frac{\lambda^k}{k!}+\int_0^{\lambda}e^t\frac{(\lambda-1)^{n-1}}{(n-1)!}\mathrm{d}t \tag{5.7.1013}$$

$$\underset{u=\lambda-t}{=}\sum_{k=0}^{n-1}\frac{\lambda^k}{k!}-\int_{\lambda}^{0}e^{\lambda-u}\frac{u^{n-1}}{(n-1)!}\mathrm{d}u=\sum_{k=0}^{n-1}\frac{\lambda^k}{k!}+\int_0^{\lambda}e^{\lambda-u}\frac{u^{n-1}}{(n-1)!}\mathrm{d}u \tag{5.7.1014}$$

We multiply by $e^{-\lambda}$:

$$e^{\lambda}e^{-\lambda}=e^{-\lambda}\left(\sum_{k=0}^{n-1}\frac{\lambda^k}{k!}+\int_0^{\lambda}e^{\lambda-u}\frac{u^{n-1}}{(n-1)!}\mathrm{d}u\right)\Rightarrow 1=\sum_{k=0}^{n-1}\frac{\lambda^k}{k!}e^{-\lambda}+\int_0^{\lambda}e^{-u}\frac{u^{n-1}}{(n-1)!}\mathrm{d}u \tag{5.7.1015}$$

And therefore:

$$\sum_{k=0}^{n-1}\frac{\lambda^k}{k!}e^{-\lambda}=1-\int_0^{\lambda}e^{-u}\frac{u^{n-1}}{(n-1)!}\mathrm{d}u \tag{5.7.1016}$$

Now, let us focus on the term:

$$\int_0^{\lambda}e^{-u}\frac{u^{n-1}}{(n-1)!}\mathrm{d}u \tag{5.7.1017}$$

and make a first change of variable:

$$\int_0^{\lambda}e^{-u}\frac{u^{n-1}}{(n-1)!}\mathrm{d}u\underset{u=x/2}{=}\int_0^{\lambda}\frac{1}{2^{n-1}(n-1)!}x^{n-1}e^{-x/2}\frac{1}{2}\mathrm{d}x=\int_0^{\lambda}\frac{1}{2^n(n-1)!}x^{n-1}e^{-x/2}\mathrm{d}x \tag{5.7.1018}$$

and a second change of variable (caution! the k in the change of variable is not the same as this in the Poisson sum...):

$$\int_0^{\lambda}\frac{1}{2^n(n-1)!}x^{n-1}e^{-x/2}\mathrm{d}x\underset{n=k/2}{=}\int_0^{\lambda}\frac{1}{2^{k/2}\left(\dfrac{k}{2}-1\right)!}x^{k/2-1}e^{-x/2}\mathrm{d}x \tag{5.7.1019}$$

However, we have shown in the section of Differential and Integral Calculus (page 1574) that if x is a positive integer:

$$x! = \Gamma(x+1)! \tag{5.7.1020}$$

Then it comes:

$$\left(\frac{k}{2}-1\right)! = \Gamma\left(\left(\frac{k}{2}-1\right)+1\right) = \Gamma\left(\frac{k}{2}\right) \tag{5.7.1021}$$

Finally we have:

$$\int\limits_0^\lambda \frac{1}{2^{k/2}\left(\dfrac{k}{2}\right)!} x^{k/2-1} e^{-x/2} \mathrm{d}x \tag{5.7.1022}$$

where we find out the chi-squared distribution under the integral! So at the end:

$$\sum_{k=0}^{n-1} \frac{\lambda^k}{k!} e^{-\lambda} = 1 - \int\limits_0^\lambda \frac{1}{2^{k/2}\left(\dfrac{k}{2}\right)!} x^{k/2-1} e^{-x/2} \mathrm{d}x \tag{5.7.1023}$$

This explains the formulas given above for the spreadsheet software, remembering that we choose to write:

$$\begin{aligned} n &= \frac{k}{2} \Leftrightarrow k = 2n \\ u &= \frac{x}{x} \Leftrightarrow x = 2u = 2(\lambda - t)_{t=0} = 2\lambda \end{aligned} \tag{5.7.1024}$$

Finally for people that don't have an access to spreadsheet or statistical software:

CHI-SQUARED PERCENTAGE POINTS

k	0.1%	0.5%	1.0%	2.5%	5.0%	10.0%	12.5%	20.0%	25.0%	33.3%	50.0%
1	0.000	0.000	0.000	0.001	0.004	0.016	0.025	0.064	0.102	0.186	0.455
2	0.002	0.010	0.020	0.051	0.103	0.211	0.267	0.446	0.575	0.811	1.386
3	0.024	0.072	0.115	0.216	0.352	0.584	0.692	1.005	1.213	1.568	2.366
4	0.091	0.207	0.297	0.484	0.711	1.064	1.219	1.649	1.923	2.378	3.357
5	0.210	0.412	0.554	0.831	1.145	1.610	1.808	2.343	2.675	3.216	4.351
6	0.381	0.676	0.872	1.237	1.635	2.204	2.441	3.070	3.455	4.074	5.348
7	0.598	0.989	1.239	1.690	2.167	2.833	3.106	3.822	4.255	4.945	6.346
8	0.857	1.344	1.646	2.180	2.733	3.490	3.797	4.594	5.071	5.826	7.344
9	1.152	1.735	2.088	2.700	3.325	4.168	4.507	5.380	5.899	6.716	8.343
10	1.479	2.156	2.558	3.247	3.940	4.865	5.234	6.179	6.737	7.612	9.342
11	1.834	2.603	3.053	3.816	4.575	5.578	5.975	6.989	7.584	8.514	10.341
12	2.214	3.074	3.571	4.404	5.226	6.304	6.729	7.807	8.438	9.420	11.340
13	2.617	3.565	4.107	5.009	5.892	7.042	7.493	8.634	9.299	10.331	12.340
14	3.041	4.075	4.660	5.629	6.571	7.790	8.266	9.467	10.165	11.245	13.339
15	3.483	4.601	5.229	6.262	7.261	8.547	9.048	10.307	11.037	12.163	14.339
16	3.942	5.142	5.812	6.908	7.962	9.312	9.837	11.152	11.912	13.083	15.338
17	4.416	5.697	6.408	7.564	8.672	10.085	10.633	12.002	12.792	14.006	16.338
18	4.905	6.265	7.015	8.231	9.390	10.865	11.435	12.857	13.675	14.931	17.338
19	5.407	6.844	7.633	8.907	10.117	11.651	12.242	13.716	14.562	15.859	18.338
20	5.921	7.434	8.260	9.591	10.851	12.443	13.055	14.578	15.452	16.788	19.337
21	6.447	8.034	8.897	10.283	11.591	13.240	13.873	15.445	16.344	17.720	20.337
22	6.983	8.643	9.542	10.982	12.338	14.041	14.695	16.314	17.240	18.653	21.337
23	7.529	9.260	10.196	11.689	13.091	14.848	15.521	17.187	18.137	19.587	22.337
24	8.085	9.886	10.856	12.401	13.848	15.659	16.351	18.062	19.037	20.523	23.337
25	8.649	10.520	11.524	13.120	14.611	16.473	17.184	18.940	19.939	21.461	24.337
26	9.222	11.160	12.198	13.844	15.379	17.292	18.021	19.820	20.843	22.399	25.336
27	9.803	11.808	12.879	14.573	16.151	18.114	18.861	20.703	21.749	23.339	26.336
28	10.391	12.461	13.565	15.308	16.928	18.939	19.704	21.588	22.657	24.280	27.336
29	10.986	13.121	14.256	16.047	17.708	19.768	20.550	22.475	23.567	25.222	28.336
30	11.588	13.787	14.953	16.791	18.493	20.599	21.399	23.364	24.478	26.165	29.336
35	14.688	17.192	18.509	20.569	22.465	24.797	25.678	27.836	29.054	30.894	34.336
40	17.916	20.707	22.164	24.433	26.509	29.051	30.008	32.345	33.660	35.643	39.335
45	21.251	24.311	25.901	28.366	30.612	33.350	34.379	36.884	38.291	40.407	44.335
50	24.674	27.991	29.707	32.357	34.764	37.689	38.785	41.449	42.942	45.184	49.335
55	28.173	31.735	33.570	36.398	38.958	42.060	43.220	46.036	47.610	49.972	54.335
60	31.738	35.534	37.485	40.482	43.188	46.459	47.680	50.641	52.294	54.770	59.335

CHI-SQUARED PERCENTAGE POINTS

k	60.0%	66.7%	75.0%	80.0%	87.5%	90.0%	95.0%	97.5%	99.0%	99.5%	99.9%
1	0.708	0.936	1.323	1.642	2.354	2.706	3.841	5.024	6.635	7.879	10.828
2	1.833	2.197	2.773	3.219	4.159	4.605	5.991	7.378	9.210	10.597	13.816
3	2.946	3.405	4.108	4.642	5.739	6.251	7.815	9.348	11.345	12.838	16.266
4	4.045	4.579	5.385	5.989	7.214	7.779	9.488	11.143	13.277	14.860	18.467
5	5.132	5.730	6.626	7.289	8.625	9.236	11.070	12.833	15.086	16.750	20.515
6	6.211	6.867	7.841	8.558	9.992	10.645	12.592	14.449	16.812	18.548	22.458
7	7.283	7.992	9.037	9.803	11.326	12.017	14.067	16.013	18.475	20.278	24.322
8	8.351	9.107	10.219	11.030	12.636	13.362	15.507	17.535	20.090	21.955	26.125
9	9.414	10.215	11.389	12.242	13.926	14.684	16.919	19.023	21.666	23.589	27.877
10	10.473	11.317	12.549	13.442	15.198	15.987	18.307	20.483	23.209	25.188	29.588
11	11.530	12.414	13.701	14.631	16.457	17.275	19.675	21.920	24.725	26.757	31.264
12	12.584	13.506	14.845	15.812	17.703	18.549	21.026	23.337	26.217	28.300	32.910
13	13.636	14.595	15.984	16.985	18.939	19.812	22.362	24.736	27.688	29.819	34.528
14	14.685	15.680	17.117	18.151	20.166	21.064	23.685	26.119	29.141	31.319	36.123
15	15.733	16.761	18.245	19.311	21.384	22.307	24.996	27.488	30.578	32.801	37.697
16	16.780	17.840	19.369	20.465	22.595	23.542	26.296	28.845	32.000	34.267	39.252
17	17.824	18.917	20.489	21.615	23.799	24.769	27.587	30.191	33.409	35.718	40.790
18	18.868	19.991	21.605	22.760	24.997	25.989	28.869	31.526	34.805	37.156	42.312
19	19.910	21.063	22.718	23.900	26.189	27.204	30.144	32.852	36.191	38.582	43.820
20	20.951	22.133	23.828	25.038	27.376	28.412	31.410	34.170	37.566	39.997	45.315
21	21.991	23.201	24.935	26.171	28.559	29.615	32.671	35.479	38.932	41.401	46.797
22	23.031	24.268	26.039	27.301	29.737	30.813	33.924	36.781	40.289	42.796	48.268
23	24.069	25.333	27.141	28.429	30.911	32.007	35.172	38.076	41.638	44.181	49.728
24	25.106	26.397	28.241	29.553	32.081	33.196	36.415	39.364	42.980	45.559	51.179
25	26.143	27.459	29.339	30.675	33.247	34.382	37.652	40.646	44.314	46.928	52.620
26	27.179	28.520	30.435	31.795	34.410	35.563	38.885	41.923	45.642	48.290	54.052
27	28.214	29.580	31.528	32.912	35.570	36.741	40.113	43.195	46.963	49.645	55.476
28	29.249	30.639	32.620	34.027	36.727	37.916	41.337	44.461	48.278	50.993	56.892
29	30.283	31.697	33.711	35.139	37.881	39.087	42.557	45.722	49.588	52.336	58.301
30	31.316	32.754	34.800	36.250	39.033	40.256	43.773	46.979	50.892	53.672	59.703
35	36.475	38.024	40.223	41.778	44.753	46.059	49.802	53.203	57.342	60.275	66.619
40	41.622	43.275	45.616	47.269	50.424	51.805	55.758	59.342	63.691	66.766	73.402
45	46.761	48.510	50.985	52.729	56.052	57.505	61.656	65.410	69.957	73.166	80.077
50	51.892	53.733	56.334	58.164	61.647	63.167	67.505	71.420	76.154	79.490	86.661
55	57.016	58.945	61.665	63.577	67.211	68.796	73.311	77.380	82.292	85.749	93.168
60	62.135	64.147	66.981	68.972	72.751	74.397	79.082	83.298	88.379	91.952	99.607

5.7.6.18.1 Noncentral chi-square distribution

The "noncentral chi-squared distribution" can be found in two major situation in industrial statistics (and hence in business):

- In the lower and upper bound estimations of tolerance interval for an univariate Normal distribution
- As the distribution of the alternative hypothesis (H_1) of many NHST tests and hence used to calculate the power of the corresponding tests.

Let us now determine the density probability function of the noncentered chi-square distribution.

We know that the density probability function of $X^2 = \mathcal{N}(0,1)^2$ follows by definition a χ^2 distribution of one degree of freedom. The question now is what distribution follows $\mathcal{N}(\lambda,1)^2$??

First we know that we can rewrite this as:

$$T = X - \lambda = \mathcal{N}(0,1) \tag{5.7.1025}$$

Hence T has distribution $\Phi(t)$ and density function (see earlier our study of the Normal distribution):

$$\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \tag{5.7.1026}$$

The distribution of X^2 is obviously:

$$\begin{aligned} P(X^2 \leq x) &= P\left(-\sqrt{x} \leq X \leq \sqrt{x}\right) \\ &= P\left(-\sqrt{x} - \lambda \leq X_1 - \lambda \leq \sqrt{x} - \lambda\right) \\ &= \Phi(\sqrt{x} - \lambda) - \Phi(-\sqrt{x} - \lambda) \end{aligned} \tag{5.7.1027}$$

Hence the density is:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x}\left[\Phi(\sqrt{x} - \lambda) - \Phi(-\sqrt{x} - \lambda)\right] &= \frac{1}{2\sqrt{x}}\left[\phi(\sqrt{x} - \lambda) + \phi(\sqrt{x} + \lambda)\right] \\ &= \frac{1}{2\sqrt{x}}\left[\frac{1}{\sqrt{2\pi}} e^{\frac{(\sqrt{x}-\lambda)^2}{2}} + \frac{1}{\sqrt{2\pi}} e^{\frac{(\sqrt{x}+\lambda)^2}{2}}\right] \\ &= \frac{1}{\sqrt{2\pi x}} e^{-\frac{1}{2}(x+\lambda^2)}\left[\frac{e^{\lambda\sqrt{x}} + e^{-\lambda\sqrt{x}}}{2}\right] \end{aligned} \tag{5.7.1028}$$

And we know (see section Sequences and Series page 1703) the following Maclaurin series:

$$e^x = \sum_{k=0}^{+\infty} \frac{1}{k!} x^k \tag{5.7.1029}$$

Hence:

$$\begin{aligned}\frac{1}{\sqrt{2\pi x}}e^{-\frac{1}{2}(x+\lambda^2)}\left[\frac{e^{\lambda\sqrt{x}}+e^{-\lambda\sqrt{x}}}{2}\right]&=\frac{1}{\sqrt{2\pi x}}e^{-\frac{1}{2}(x+\lambda^2)}\left[\sum_{k=0}^{+\infty}\frac{(\lambda\sqrt{x})^{2k}}{(2k)!}\right]\\&=\sum_{k=0}^{+\infty}\left\{e^{-\frac{\lambda^2}{2}}\frac{(\lambda^2/2)^k)}{k!}\right\}\left\{e^{-\frac{x}{2}}\left(\frac{x}{2}\right)^k\frac{2^{2k}k!}{(2k)!}\frac{1}{\sqrt{2\pi x}}\right\}\end{aligned}\tag{5.7.1030}$$

Now let us recall that if K is a random variable with a Poisson distribution and mean $\lambda^2/2$, then its distribution is equal to:

$$P_\lambda(K=k)=e^{-\lambda^2/2}\frac{(\lambda^2/2)^k}{k!}\tag{5.7.1031}$$

with $k=0,1,2,\ldots$.

Therefore:

$$\sum_{k=0}^{+\infty}\left\{e^{-\frac{\lambda^2}{2}}\frac{(\lambda^2/2)^k)}{k!}\right\}\left\{e^{-\frac{x}{2}}\left(\frac{x}{2}\right)^k\frac{2^{2k}k!}{(2k)!}\frac{1}{\sqrt{2\pi x}}\right\}=\sum_{k=0}^{+\infty}P_\lambda(K=k)\left\{e^{-\frac{x}{2}}\left(\frac{x}{2}\right)^k\frac{2^{2k}k!}{(2k)!}\frac{1}{\sqrt{2\pi x}}\right\}\tag{5.7.1032}$$

Now let us recall that the chi-square distribution is given by:

$$P_k(x)=\frac{1}{2^{\frac{k}{2}}\Gamma(k/2)}x^{\frac{k}{2}-1}e^{-\frac{x}{2}}\cdot 1_{[0,+\infty[}\tag{5.7.1033}$$

Hence:

$$\begin{aligned}P_{2k+1}(x)&=\frac{1}{2^{\frac{2k+1}{2}}\Gamma((2k+1)/2)}x^{\frac{(2k+1)}{2}-1}e^{-\frac{x}{2}}\cdot 1_{[0,+\infty[}=\frac{1}{2^{k+\frac{1}{2}}\Gamma(k+\frac{1}{2})}x^{k-\frac{1}{2}}e^{-\frac{x}{2}}\cdot 1_{[0,+\infty[}\\&=\frac{1}{2\Gamma(k+\frac{1}{2})}\left(\frac{x}{2}\right)^{k-\frac{1}{2}}e^{-\frac{x}{2}}\cdot 1_{[0,+\infty[}\end{aligned}\tag{5.7.1034}$$

So:

$$P_{2k+1}(x)2\Gamma\left(k+\frac{1}{2}\right)\left(\frac{x}{2}\right)^{\frac{1}{2}}\cdot 1_{[0,+\infty[}=\left(\frac{x}{2}\right)^k e^{-\frac{x}{2}}\tag{5.7.1035}$$

Therefore:

$$\begin{aligned}\sum_{k=0}^{+\infty}P_\lambda(K=k)\left\{e^{-\frac{x}{2}}\left(\frac{x}{2}\right)^k\frac{2^{2k}k!}{(2k)!}\frac{1}{\sqrt{2\pi x}}\right\}&=\sum_{k=0}^{+\infty}P_\lambda(K=k)\left\{P_{2k+1}(x)2\Gamma\left(k+\frac{1}{2}\right)\left(\frac{x}{2}\right)^{\frac{1}{2}}\cdot 1_{[0,+\infty[}\frac{2^{2k}k!}{(2k)!}\frac{1}{\sqrt{2\pi x}}\right\}\\&=\sum_{k=0}^{+\infty}P_\lambda(K=k)\left\{P_{2k+1}(x)\Gamma\left(k+\frac{1}{2}\right)\frac{2^{2k}k!}{(2k)!}\frac{1}{\sqrt{\pi}}\right\}1_{[0,+\infty[}\\&=\sum_{k=0}^{+\infty}P_\lambda(K=k)P_{2k+1}(x)\left\{\Gamma\left(k+\frac{1}{2}\right)\frac{2^{2k}k!}{(2k)!}\frac{1}{\sqrt{\pi}}\right\}1_{[0,+\infty[}\end{aligned}\tag{5.7.1036}$$

Now let us recall that we have proved in the section Differential and Integral Calculus (see page 1574) that:

$$\Gamma\left(k+\frac{1}{2}\right)=\frac{(2k-1)!}{2^{2k-1}(k-1)!}\sqrt{\pi} \tag{5.7.1037}$$

Therefore:

$$\begin{aligned}
&\sum_{k=0}^{+\infty} P_\lambda(K=k)P_{2k+1}(x)\left\{\Gamma\left(k+\frac{1}{2}\right)\frac{2^{2k}k!}{(2k)!}\frac{1}{\sqrt{\pi}}\right\}1_{[0,+\infty[}\\
&=\sum_{k=0}^{+\infty} P_\lambda(K=k)P_{2k+1}(x)\left\{\frac{(2k-1)!}{2^{2k-1}(k-1)!}\sqrt{\pi}\frac{2^{2k}k!}{(2k)!}\frac{1}{\sqrt{\pi}}\right\}1_{[0,+\infty[}\\
&=\sum_{k=0}^{+\infty} P_\lambda(K=k)P_{2k+1}(x)\left\{\frac{(2k-1)!}{2^{2k-1}(k-1)!}\frac{2^{2k}k!}{(2k)!}\right\}1_{[0,+\infty[}\\
&=\sum_{k=0}^{+\infty} P_\lambda(K=k)P_{2k+1}(x)\frac{1}{2^{-1}}\frac{k}{2k}1_{[0,+\infty[}\\
&=\sum_{k=0}^{+\infty} P_\lambda(K=k)P_{2k+1}(x)1_{[0,+\infty[}
\end{aligned} \tag{5.7.1038}$$

So finally the noncentral chi-square distribution function with one degree of freedom is given by:

$$\boxed{\chi_1(\lambda)=\sum_{k=0}^{+\infty} P_\lambda(K=k)P_{2k+1}(x)1_{[0,+\infty[}} \tag{5.7.1039}$$

Or explicitly:

$$\chi_1(\lambda)=\sum_{k=0}^{+\infty}\left\{e^{-\frac{\lambda^2}{2}}\frac{(\lambda^2/2)^k)}{k!}\right\}\left\{\frac{1}{2^{k+\frac{1}{2}}\Gamma(k+\frac{1}{2})}x^{k-\frac{1}{2}}e^{-\frac{x}{2}}\cdot 1_{[0,+\infty[}\right\} \tag{5.7.1040}$$

Remark

It can be proved (but so far we don't need this result actually anywhere in this textbook), that the general case with n degrees of freedom is given by:

$$\chi_n(\lambda)=\sum_{k=0}^{+\infty} P_\lambda(K=k)P_{2k+n}(x)1_{[0,+\infty[} \tag{5.7.1041}$$

That's actually all we need for our study later of tolerance intervals! So we will not derive the mean and variance of this distribution as is it would be useless in the actual state of this textbook!

5.7.6.19 Student Distribution

The "Student distribution" (or "Student's law") of parameter k is defined by the relation (we will prove however it's historical origin at the next page):

$$\boxed{P(x)=\frac{\Gamma\left(\dfrac{k+1}{2}\right)}{\Gamma\left(\dfrac{k}{2}\right)\sqrt{k\pi}}\left(1+\frac{x^2}{k}\right)^{-\frac{k+1}{2}}} \tag{5.7.1042}$$

with k being the degree of freedom of the χ^2 distribution underlying the construction of the Student function as we will see.

This is in reality the "centered reduced Student's distribution" because as we will see it further below, it is derived by a relation involving the centered reduced Normal law.

Let us indicate that this distribution can also be obtained in Microsoft® Excel® 11.8346 using the `TDIST( )` function and its inverse by `TINV()`.

It is indeed a distribution function because it also satisfies (remains to be proved directly, but as we will see it is the product of two distribution functions thus indirectly...):

$$\int\limits_{-\infty}^{+\infty} P_k(x)\mathrm{d}x=1 \tag{5.7.1043}$$

Let us see the easiest proof to justify the provenance of the Student distribution and that will also be very useful further in statistical inference and analysis of variance.

For this proof, remember first that:

1. If X and Y are two independent random variables with respective densities f_X, f_Y, the distribution of the pair (X, Y) has a density f satisfying (axiom of probabilities!):

$$f(x,y)=f_X(x)f_Y(y) \tag{5.7.1044}$$

2. The distribution $\mathcal{N}(0,1)$ is given by (see above):

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2} \tag{5.7.1045}$$

3. The distribution χ_n^2 is given by (see above):

$$f(y)=\frac{1}{2^{n/2}\Gamma(n/2)}e^{-\frac{y}{2}}y^{\frac{n}{2}-1} \tag{5.7.1046}$$

   for $y\geq 0$ and $n\geq 1$.

4. The Euler function Γ is defined for all $\alpha > 0$ by (see section Differential and Integral Calculus page 1574):

$$\Gamma(\alpha)=\int\limits_{0}^{+\infty} e^{-x}x^{\alpha-1}\mathrm{d}x \tag{5.7.1047}$$

and satisfies (see section Differential and Integral Calculus page 1574):

$$\Gamma(\alpha-1)=\frac{\Gamma(\alpha)}{\alpha-1} \tag{5.7.1048}$$

for $\alpha \geq 2$.

These reminders made, now consider a random variable X that follows the distribution $\mathcal{N}(0,1)$ and Y a random variable following the distribution χ_n^2.

We assume X and Y being independent and we consider the random variable (this is at the origin the historical study of the Student distribution in the framework of statistical inference which led to define this variable for which we will deepen the origin later):

$$T=\sqrt{n}\frac{X}{\sqrt{Y}}=\frac{X}{\sqrt{Y/n}}=\frac{\mathcal{N}(0,1)}{\sqrt{\chi_n^2/n}} \tag{5.7.1049}$$

We will prove that T follows a Student distribution of parameter n.

Proof 5.47.7. Let F and f be respectively the repartition and density functions and T, f_X, f_Y the density functions of X, Y and (X,Y) respectively. Then we have for all $t\in\mathbb{R}$:

$$F(t)=P(T\leqslant t)=P\left(\sqrt{n}\frac{X}{\sqrt{Y}}\leqslant t\right)=\iint\limits_{D} f(x,y)\mathrm{d}x\mathrm{d}y=\iint\limits_{D} f_X(x)f_Y(y)\mathrm{d}x\mathrm{d}y \tag{5.7.1050}$$

where:

$$D=\left\{(x,y)\in\mathbb{R}\times\mathbb{R}^*|x\leq\frac{t\sqrt{y}}{\sqrt{n}}\right\} \tag{5.7.1051}$$

the imposed positive and non-zero value and y being due to the fact that it is under a root and furthermore at the denominator.

Thus:

$$F(t)=\iint\limits_{D} f_X(x)f_Y(y)\mathrm{d}x\mathrm{d}y=\int\limits_{0}^{+\infty} f_Y(y)\mathrm{d}y\int\limits_{-\infty}^{t\sqrt{y}/\sqrt{n}} f_X(x)\mathrm{d}x=\int\limits_{0}^{+\infty} f_Y(y)\Phi(t\sqrt{y}/\sqrt{n})\mathrm{d}y \tag{5.7.1052}$$

where because X follows a Normal $\mathcal{N}(0,1)$ distribution:

$$\Phi(u)=\int\limits_{-\infty}^{u}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\mathrm{d}x \tag{5.7.1053}$$

is the Normal centered reduced cumulative distribution.

Thus, we obtain the density distribution function of T by deriving F:

$$f(t)=F'(t)=\frac{1}{\sqrt{n}}\int\limits_0^{+\infty} f_Y(y)f_X(t\sqrt{y}/\sqrt{n})\sqrt{y}\mathrm{d}y \tag{5.7.1054}$$

because (the derivative of a function is equal to its derivatives multiplied by its inner derivative):

$$\frac{\mathrm{d}\Phi(t\sqrt{n}/\sqrt{t})}{\mathrm{d}t}=f_X(t\sqrt{n}/\sqrt{t})\frac{\sqrt{y}}{\sqrt{n}} \tag{5.7.1055}$$

Therefore:

$$\begin{aligned}
f(t)&=\frac{1}{\sqrt{n}}\int\limits_0^{+\infty} f_Y(y)f_X(t\sqrt{y}/\sqrt{n})\sqrt{y}\mathrm{d}y\\
&=\frac{1}{\sqrt{n}}\int\limits_0^{+\infty}\frac{1}{2^{n/2}\Gamma(n/2)}e^{-y/2}y^{n/2-1}\frac{1}{\sqrt{2\pi}}e^{-(t\sqrt{y}/\sqrt{n})^2/2}\sqrt{y}\mathrm{d}y\\
&=\frac{1}{\sqrt{n}}\frac{1}{2^{n/2}\Gamma(n/2)}\frac{1}{\sqrt{2\pi}}\int\limits_0^{+\infty}e^{-y/2}y^{n/2-1}e^{-(t\sqrt{y}/\sqrt{n})^2/2}\sqrt{y}\mathrm{d}y\\
\\
&=\frac{1}{2^{n/2}\Gamma(n/2)\sqrt{2\pi n}}\int\limits_0^{+\infty}e^{-\frac{y(1+t^2/n)}{2}}y^{(n-1)/2}\mathrm{d}y
\end{aligned} \tag{5.7.1056}$$

By making the change of variable:

$$u=\frac{y}{2}\left(1+\frac{t^2}{n}\right)\Rightarrow\begin{cases}\mathrm{d}y=\left(\dfrac{2}{1+\dfrac{t^2}{n}}\right)\mathrm{d}u\\ y=\frac{2}{1+\frac{t^2}{n}}u\end{cases} \tag{5.7.1057}$$

we get:

$$\begin{aligned}
f(t)&=\frac{1}{2^{n/2}\Gamma(n/2)\sqrt{2\pi n}}\left(\frac{2}{1+t^2/n}\right)^{(n+1)/2}\underbrace{\int\limits_0^{+\infty}e^{-u}u^{(n-1)/2}\mathrm{d}u}_{=\Gamma\left(\frac{n+1}{2}\right)}\\
&=\frac{\Gamma\left(\dfrac{n+1}{2}\right)}{2^{n/2}\Gamma(n/2)\sqrt{2\pi}\sqrt{n}}\left(\frac{2}{1+t^2/n}\right)^{(n+1)/2}\\
&=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma(n/2)\sqrt{\pi n}}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}
\end{aligned} \tag{5.7.1058}$$

what is the Student distribution of parameter n.

■ Q.E.D.

Let us now prove what is the mean of the Student distribution:

$$T=\sqrt{n}\frac{X}{\sqrt{Y}} \tag{5.7.1059}$$

We have:

$$\mathrm{E}(T)=\mathrm{E}(\sqrt{n}X)\mathrm{E}\left(\frac{1}{\sqrt{Y}}\right) \tag{5.7.1060}$$

But $\mathrm{E}\left(\dfrac{1}{\sqrt{Y}}\right)$ exists if and only if $n\geq 2$. Effectively for $n=1$:

$$\mathrm{E}\left(\frac{1}{\sqrt{Y}}\right)=\int\limits_0^{+\infty}\frac{f_Y(y)}{\sqrt{y}}\mathrm{d}y=\frac{1}{2^{\frac{n}{2}}\Gamma(n/2)}\int\limits_0^{+\infty}e^{-\frac{y}{2}}y^{\frac{n-3}{2}}\mathrm{d}y \tag{5.7.1061}$$

and:

$$\int\limits_0^{+\infty}e^{-\frac{y}{2}}y^{\frac{1-3}{2}}\mathrm{d}y=\int\limits_0^{+\infty}\frac{e^{-\frac{y}{2}}}{y}\mathrm{d}y\geqslant\int\limits_0^{1}\frac{e^{-\frac{y}{2}}}{y}\mathrm{d}y\geqslant e^{-\frac{1}{2}}\int\limits_0^{+\infty}\frac{1}{y}\mathrm{d}y\rightarrow+\infty \tag{5.7.1062}$$

Whereas for $n\geq 2$ we have:

$$\int\limits_0^{+\infty}e^{-\frac{y}{2}}y^{\frac{n-3}{2}}\mathrm{d}y=2^{\frac{n-1}{2}}\int\limits_0^{+\infty}e^{-u}u^{\frac{n-3}{2}}\mathrm{d}u=2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)<+\infty \tag{5.7.1063}$$

Thus, for $n=1$ the mean does not exist.

So for $n\geq 2$:

$$\mathrm{E}(T)=\sqrt{n}\underbrace{\mathrm{E}(X)}_{=0}\mathrm{E}\left(\frac{1}{\sqrt{Y}}\right)=0 \tag{5.7.1064}$$

Now let us see the value of the variance. So we have:

$$\mathrm{V}(T)=\mathrm{E}(T^2)-\mathrm{E}(T)^2 \tag{5.7.1065}$$

First we will discuss the existence of $\mathrm{E}(T^2)$. We have trivially:

$$\mathrm{E}(T^2)=nE\left(\frac{X^2}{Y}\right)=n\mathrm{E}(X^2)\mathrm{E}\left(\frac{1}{Y}\right) \tag{5.7.1066}$$

X follows a Normal centered reduced distribution thus:

$$\mathrm{V}(X)=1=\mathrm{E}(X^2)-\underbrace{\mathrm{E}(X)^2}_{=0}=\mathrm{E}(X^2)\quad\Rightarrow\quad\mathrm{E}(X^2)=1 \tag{5.7.1067}$$

With regard to $\mathrm{E}\left(\dfrac{1}{Y}\right)$ we have:

$$\mathrm{E}\left(\frac{1}{Y}\right)=\int\limits_0^{+\infty}\frac{f_Y(y)}{y}\mathrm{d}y=\frac{1}{2^{\frac{n}{2}}\Gamma(n/2)}\int\limits_0^{+\infty}e^{-\frac{y}{2}}y^{\frac{n}{2}-2}\mathrm{d}y=\frac{1}{2^{\frac{n}{2}}\Gamma(n/2)}\int\limits_0^{+\infty}e^{-u}u^{\frac{n}{2}-2}\mathrm{d}u=\frac{\Gamma\left(\frac{n}{2}-1\right)}{2\Gamma(n/2)} \tag{5.7.1068}$$

where we made the change of variable $u=y/2$.

But the integral defining $\Gamma\left(\dfrac{n}{2}-1\right)$ converges only if $n\geq 3$.

Therefore $\mathrm{E}(T^2)$ exists if and only if $n\geq 3$ so it's value is according to the properties of the Euler Gamma function demonstrated in the chapter of Differential and Integral Calculus:

$$n\mathrm{E}\left(T\right)=n\frac{\Gamma\left(\frac{n}{2}-1\right)}{2\Gamma(n/2)}=\frac{n}{n-2} \tag{5.7.1069}$$

Therefore for $n\geq 3$:

$$\mathrm{V}(T)=\frac{n}{n-2} \tag{5.7.1070}$$

It is also important to note that this law is symmetrical about 0!

Here is a plot example of the Student distribution and cumulative distribution for the parameter $k=3$ (for more plots about this distribution see further below on page 388):

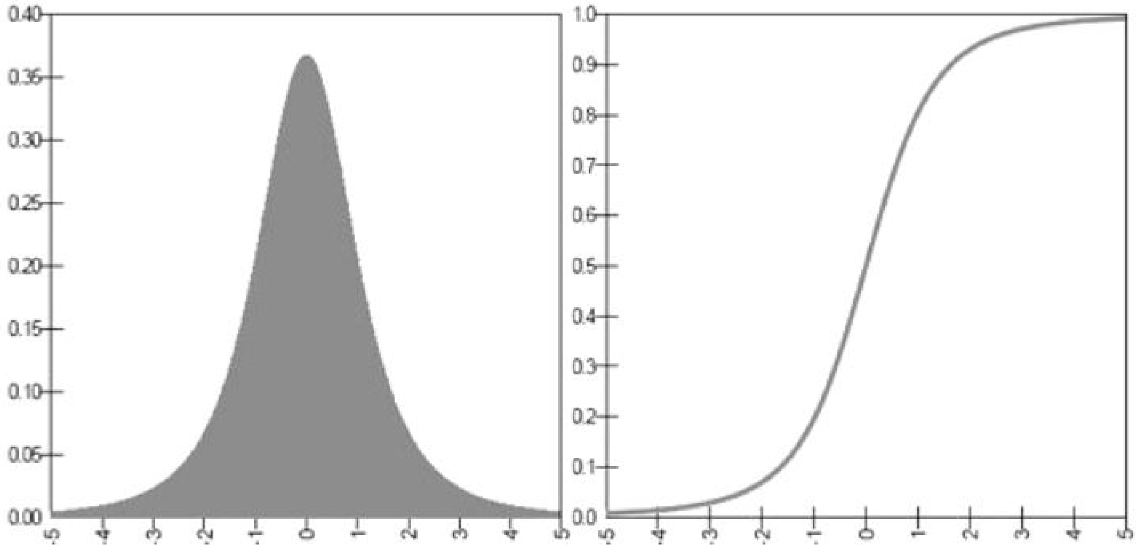


Figure 5.151: **Student T law (mass and cumulative distribution function)**

Finally for people that don't have an access to spreadsheet or statistical software:

STUDENT'S T PERCENTAGE POINTS

k	60.0%	66.7%	75.0%	80.0%	87.5%	90.0%	95.0%	97.5%	99.0%	99.5%	99.9%
1	0.325	0.577	1.000	1.376	2.414	3.078	6.314	12.706	31.821	63.657	318.31
2	0.289	0.500	0.816	1.061	1.604	1.886	2.920	4.303	6.965	9.925	22.327
3	0.277	0.476	0.765	0.978	1.423	1.638	2.353	3.182	4.541	5.841	10.215
4	0.271	0.464	0.741	0.941	1.344	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.457	0.727	0.920	1.301	1.476	2.015	2.571	3.365	4.032	5.893
6	0.265	0.453	0.718	0.906	1.273	1.440	1.943	2.447	3.143	3.707	5.208
7	0.263	0.449	0.711	0.896	1.254	1.415	1.895	2.365	2.998	3.499	4.785
8	0.262	0.447	0.706	0.889	1.240	1.397	1.860	2.306	2.896	3.355	4.501
9	0.261	0.445	0.703	0.883	1.230	1.383	1.833	2.262	2.821	3.250	4.297
10	0.260	0.444	0.700	0.879	1.221	1.372	1.812	2.228	2.764	3.169	4.144
11	0.260	0.443	0.697	0.876	1.214	1.363	1.796	2.201	2.718	3.106	4.025
12	0.259	0.442	0.695	0.873	1.209	1.356	1.782	2.179	2.681	3.055	3.930
13	0.259	0.441	0.694	0.870	1.204	1.350	1.771	2.160	2.650	3.012	3.852
14	0.258	0.440	0.692	0.868	1.200	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.439	0.691	0.866	1.197	1.341	1.753	2.131	2.602	2.947	3.733
16	0.258	0.439	0.690	0.865	1.194	1.337	1.746	2.120	2.583	2.921	3.686
17	0.257	0.438	0.689	0.863	1.191	1.333	1.740	2.110	2.567	2.898	3.646
18	0.257	0.438	0.688	0.862	1.189	1.330	1.734	2.101	2.552	2.878	3.610
19	0.257	0.438	0.688	0.861	1.187	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.437	0.687	0.860	1.185	1.325	1.725	2.086	2.528	2.845	3.552
21	0.257	0.437	0.686	0.859	1.183	1.323	1.721	2.080	2.518	2.831	3.527
22	0.256	0.437	0.686	0.858	1.182	1.321	1.717	2.074	2.508	2.819	3.505
23	0.256	0.436	0.685	0.858	1.180	1.319	1.714	2.069	2.500	2.807	3.485
24	0.256	0.436	0.685	0.857	1.179	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.436	0.684	0.856	1.178	1.316	1.708	2.060	2.485	2.787	3.450
26	0.256	0.436	0.684	0.856	1.177	1.315	1.706	2.056	2.479	2.779	3.435
27	0.256	0.435	0.684	0.855	1.176	1.314	1.703	2.052	2.473	2.771	3.421
28	0.256	0.435	0.683	0.855	1.175	1.313	1.701	2.048	2.467	2.763	3.408
29	0.256	0.435	0.683	0.854	1.174	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.435	0.683	0.854	1.173	1.310	1.697	2.042	2.457	2.750	3.385
35	0.255	0.434	0.682	0.852	1.170	1.306	1.690	2.030	2.438	2.724	3.340
40	0.255	0.434	0.681	0.851	1.167	1.303	1.684	2.021	2.423	2.704	3.307
45	0.255	0.434	0.680	0.850	1.165	1.301	1.679	2.014	2.412	2.690	3.281
50	0.255	0.433	0.679	0.849	1.164	1.299	1.676	2.009	2.403	2.678	3.261
55	0.255	0.433	0.679	0.848	1.163	1.297	1.673	2.004	2.396	2.668	3.245
60	0.254	0.433	0.679	0.848	1.162	1.296	1.671	2.000	2.390	2.660	3.232
∞	0.253	0.431	0.674	0.842	1.150	1.282	1.645	1.960	2.326	2.576	3.090

5.7.6.19.1 Multivariate Student T-distribution

There is sadly many various different definition of the "multivariate T-distribution" (we speak then of multivariate T-distribution "candidates"). We will focus here on the one often named "canonical multivariate T-distribution".

First if the reader redo the proof of the derivation of the T distribution but use $\mathcal{N}(0,\sigma^2)$ instead of $\mathcal{N}(0,1)$, the reader should easily see that the result is:

$$f(t)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma(n/2)\sqrt{\pi\sigma n}}\left(1+\frac{t^2}{n\sigma^2}\right)^{-\frac{n+1}{2}} \tag{5.7.1071}$$

that is named the "centered Student distribution" (as it is not reduced!).

And using the rule of thumb it's also quite easy to guess that if the variable was not centered then we would have:

$$f(t)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma(n/2)\sqrt{\pi\sigma n}}\left(1+\frac{(t-\mu)^2}{n\sigma^2}\right)^{-\frac{n+1}{2}} \tag{5.7.1072}$$

that is the (real) "Student's law" (i.e. the non-centered and non-reduced one...)!

The T-distribution above is obviously a special case of the multivariate T-distribution with number of variables $n=1$, i.e. random vector $x\in\mathbb{R}$, mean $\mu=0$ and covariance matrix $\Sigma=1$.

We can guess by the rule of thumb that the probability density function of the "canonical multivariate T-distribution" knowing how the multivariate Normal distribution looks like (see page 720) is given by:

$$\boxed{T(\vec{x};\vec{\mu},\Sigma,\nu)=\frac{\Gamma([\nu+n]/2)}{\Gamma(\nu/2)\sqrt{(\nu\pi)^n|\Sigma|}}\left[1+\frac{1}{\nu}(\vec{x}-\vec{\mu})^{\mathrm{T}}\Sigma^{-1}(\vec{x}-\vec{\mu})\right]^{-(\nu+n)/2}} \tag{5.7.1073}$$

with $\vec{x}\in\mathbb{R}^n$ and Σ being the $n\times n$ variance-covariance matrix.

Setting $n=1$, such that $x\in\mathbb{R}$, as well as $\mu=0$ and $\Sigma=1$, we get:

$$\begin{aligned}T(\vec{x};0,1,\nu)&=\frac{\Gamma([\nu+1]/2)}{\Gamma(\nu/2)\sqrt{(\nu\pi)^1|1|}}\left[1+\frac{1}{\nu}(x-0)^{\mathrm{T}}1^{-1}(x-0)\right]^{-(\nu+1)/2}\\&=\frac{\Gamma([\nu+1]/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}}\left[1+\frac{x^2}{\nu}\right]^{-(\nu+1)/2}\\&=\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}}\left[1+\frac{x^2}{\nu}\right]^{-\frac{\nu+1}{2}}\end{aligned} \tag{5.7.1074}$$

which is equivalent to the probability density function of the (centered-reduced) T-distribution.

5.7.6.20 Fisher Distribution

The "Fisher distribution" (or "Fisher-Snedecor distribution") of parameters k and l is defined by the relation:

$$\boxed{P(x)=\frac{\Gamma\left(\dfrac{k+1}{2}\right)\left(\dfrac{k}{l}\right)^{\frac{k}{2}}}{\Gamma\left(\dfrac{k}{2}\right)\Gamma\left(\dfrac{l}{2}\right)}x^{\frac{k}{2}-1}\left(1+\frac{kx}{l}\right)^{-\frac{k+l}{2}}} \tag{5.7.1075}$$

if $x \geq 0$ (see the plot further below on page 388). The parameters k and l are positive integers and correspond to the two degrees of freedom of the underlying chi-square distributions. This distribution is often denoted by $F_{k,l}$ or by $F(k,l)$ and can be obtained in Microsoft® Excel® 11.8346 with the `FDIST( )` function.

It is indeed a distribution function because it satisfies the property (remains to us to prove that directly but as we will see, it is the product of two distribution functions therefore indirectly...):

$$\int\limits_0^{+\infty} F_{k,l}(x)\mathrm{d}x=1 \tag{5.7.1076}$$

Let us see the easiest proof to justify the provenance of the Fisher distribution and that we will be us also very useful further in statistical inference and analysis of variance.

For this proof, recall that:

1. The distribution χ^2_n is given by (see above):

$$f(y)=\frac{1}{2^{n/2}\Gamma(n/2)}e^{-y/2}y^{n/2-1} \tag{5.7.1077}$$

for $y \geq 0$ and $n \geq 1$.

2. The Euler Gamma function Γ is defined for all $\alpha > 0$ by (see section Differential and Integral Calculus page 1574):

$$\Gamma(\alpha)=\int\limits_0^{+\infty} e^{-x}x^{\alpha-1}\mathrm{d}x \tag{5.7.1078}$$

Let X, Y be two independent random variables following respectively the distributions χ^2_n and χ^2_m.

We consider the random variable:

$$T=\frac{X/n}{Y/m}=\frac{\chi^2_n/n}{\chi^2_m/m} \tag{5.7.1079}$$

We will prove that the distribution of T is the Fisher-Snedecor distribution of parameters n, m.

Let us note for this purpose F and f the cumulative and density function of T and f_X, f_Y, f density functions of X, Y and respectively (X, Y). We have for all $t \in \mathbb{R}$:

$$F(t)=P(T\leqslant t)=P\left(\frac{X/n}{Y/m}\leqslant t\right)=\iint\limits_{D} f(x,y)\mathrm{d}x\mathrm{d}y=\iint\limits_{D} f_X(x)f_Y(y)\mathrm{d}x\mathrm{d}y \tag{5.7.1080}$$

where:

$$D=\left\{(x,y)\in\mathbb{R}^*\times\mathbb{R}^* | x\leq\frac{nt}{m}y\right\} \tag{5.7.1081}$$

where the imposed positive values comes in fact that behind them there is a chi-square for x and y. Therefore:

$$F(t)=\int\limits_{0}^{+\infty} f_Y(y)\mathrm{d}y\int\limits_{0}^{\frac{nty}{m}} f_X(x)\mathrm{d}y \tag{5.7.1082}$$

We obtain the density function of T by deriving F. First the inner derivative:

$$f(t)=F'(t)=\frac{n}{m}\int\limits_{0}^{+\infty} f_Y(y)f_X\left(\frac{nty}{m}y\right)y\mathrm{d}y \tag{5.7.1083}$$

Then explicitly because:

$$f(y)=\frac{1}{2^{m/2}\Gamma(m/2)}e^{-y/2}y^{m/2-1} \quad \text{and} \quad f\left(\frac{nt}{m}y\right)=\frac{1}{2^{n/2}\Gamma(n/2)}e^{-\frac{nt}{2m}y}\left(\frac{nt}{m}y\right)^{n/2-1} \tag{5.7.1084}$$

we then have:

$$\begin{aligned}
f(t)&=\frac{n}{m}\int\limits_{0}^{+\infty}\frac{1}{2^{m/2}\Gamma(m/2)}e^{-y/2}y^{\frac{m}{2}-1}\frac{1}{2^{\frac{n}{2}}\Gamma(n/2)}e^{-\frac{nt}{2m}y}\left(\frac{nt}{m}y\right)^{\frac{n}{2}-1}y\mathrm{d}y\\
&=\frac{1}{2^{n/2}\Gamma(n/2)2^{m/2}\Gamma(m/2)}\frac{n}{m}\int\limits_{0}^{+\infty}e^{-y/2}y^{m/2-1}e^{-\frac{nt}{2m}y}\left(\frac{nt}{m}y\right)^{\frac{n}{2}-1}y\mathrm{d}y\\
&=\frac{1}{2^{n/2}\Gamma(n/2)2^{m/2}\Gamma(m/2)}\frac{n}{m}\left(\frac{nt}{m}\right)^{n/2-1}\int\limits_{0}^{+\infty}e^{-y/2}y^{m/2-1}e^{-\frac{nt}{2m}y}y^{n/2-1}y\mathrm{d}y\\
&=\frac{1}{2^{\frac{n+m}{2}}\Gamma(n/2)\Gamma(m/2)}\frac{n}{m}\left(\frac{nt}{m}\right)^{\frac{n}{2}-1}\int\limits_{0}^{+\infty}y^{\frac{n+m}{2}-1}e^{-\frac{y}{2}\left(1+\frac{nt}{m}\right)}\mathrm{d}y
\end{aligned} \tag{5.7.1085}$$

By making the change of variable:

$$u=\frac{y}{2}\left(1+\frac{nt}{m}\right)\Rightarrow y=\frac{2u}{1+\frac{nt}{m}}\Rightarrow \mathrm{d}y=\mathrm{d}u\frac{2}{1+\frac{nt}{m}} \tag{5.7.1086}$$

we get:

$$\begin{aligned}
f(t) &= \frac{n}{m}\frac{1}{2^{\frac{n+m}{2}}\Gamma(n/2)\Gamma(m/2)}\left(\frac{nt}{m}\right)^{\frac{n}{2}-1}\left(\frac{2}{1+\frac{nt}{m}}\right)^{\frac{n+m}{2}}\underbrace{\int\limits_0^{+\infty} u^{\frac{n+m}{2}-1}e^{-u}\mathrm{d}u}_{\Gamma\left(\frac{n+m}{2}\right)} \\
&= \frac{\frac{n}{m}\Gamma\left(\frac{n+m}{2}\right)}{\Gamma(n/2)\Gamma(m/2)}\left(\frac{nt}{m}\right)^{\frac{n}{2}-1}\left(1+\frac{nt}{m}\right)^{-\frac{n+m}{2}} \\
&= \frac{\left(\frac{n}{m}\right)^{\frac{n}{2}}\Gamma\left(\frac{n+m}{2}\right)}{\Gamma(n/2)\Gamma(m/2)}t^{\frac{n}{2}-1}\left(1+\frac{nt}{m}\right)^{-\frac{n+m}{2}}
\end{aligned} \tag{5.7.1087}$$

For the people that may not have access to a spreadsheet or statistical software here are some useful tables:

PERCENTAGE POINTS OF THE F DISTRIBUTION

$\nu_2\backslash\nu_1$	q	2	3	4	5	6	7	8	10	12	15	20	30	50	∞
1	0.900	49.5	53.6	55.8	57.2	58.2	59.1	59.7	60.5	61.0	61.5	62.0	62.6	63.0	63.3
	0.950	199.	216.	225.	230.	234.	237.	239.	242.	244.	246.	248.	250.	252.	254.
	0.975	800.	864.	900.	922.	937.	948.	957.	969.	977.	985.	993.			
	0.990														
	0.999														
2	0.900	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.39	9.41	9.43	9.44	9.46	9.47	9.49
	0.950	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5
	0.975	39.0	39.2	39.2	39.3	39.3	39.4	39.4	39.4	39.4	39.4	39.4	39.5	39.5	39.5
	0.990	99.0	99.2	99.2	99.3	99.3	99.4	100.	100.	100.	100.	100.	100.	100.	99.5
	0.999	999.	999.												
3	0.900	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.23	5.22	5.20	5.18	5.17	5.15	5.13
	0.950	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.79	8.74	8.70	8.66	8.62	8.58	8.53
	0.975	16.0	15.4	15.1	14.9	14.7	14.6	14.5	14.4	14.3	14.3	14.2	14.1	14.0	13.9
	0.990	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.2	27.1	26.9	26.7	26.5	26.4	26.1
	0.999	149.	141.	137.	135.	133.	132.	131.	129.	128.	127.	126.	125.	125.	123.
4	0.900	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.92	3.90	3.87	3.84	3.82	3.79	3.76
	0.950	6.94	6.59	6.39	6.26	6.16	6.09	6.04	5.96	5.91	5.86	5.80	5.75	5.70	5.63
	0.975	10.6	9.98	9.60	9.36	9.20	9.07	8.98	8.84	8.75	8.66	8.56	8.46	8.38	8.26
	0.990	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.5	14.4	14.2	14.0	13.8	13.7	13.5
	0.999	61.2	56.2	53.4	51.7	50.5	49.7	49.0	48.0	47.4	46.8	46.1	45.4	44.9	44.1
5	0.900	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.30	3.27	3.24	3.21	3.17	3.15	3.10
	0.950	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.74	4.68	4.62	4.56	4.50	4.44	4.36
	0.975	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.62	6.52	6.43	6.33	6.23	6.14	6.02
	0.990	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.1	9.89	9.72	9.55	9.38	9.24	9.02
	0.999	37.1	33.2	31.1	29.8	28.8	28.2	27.6	26.9	26.4	25.9	25.4	24.9	24.4	23.8
6	0.900	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.94	2.90	2.87	2.84	2.80	2.77	2.72
	0.950	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.06	4.00	3.94	3.87	3.81	3.75	3.67
	0.975	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.46	5.37	5.27	5.17	5.07	4.98	4.85
	0.990	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.87	7.72	7.56	7.40	7.23	7.09	6.88
	0.999	27.0	23.7	21.9	20.8	20.0	19.5	19.0	18.4	18.0	17.6	17.1	16.7	16.3	15.7
7	0.900	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.70	2.67	2.63	2.59	2.56	2.52	2.47
	0.950	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.64	3.57	3.51	3.44	3.38	3.32	3.23
	0.975	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.76	4.67	4.57	4.47	4.36	4.28	4.14
	0.990	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.62	6.47	6.31	6.16	5.99	5.86	5.65
	0.999	21.7	18.8	17.2	16.2	15.5	15.0	14.6	14.1	13.7	13.3	12.9	12.5	12.2	11.7
8	0.900	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.54	2.50	2.46	2.42	2.38	2.35	2.29
	0.950	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.35	3.28	3.22	3.15	3.08	3.02	2.93
	0.975	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.29	4.20	4.10	4.00	3.89	3.81	3.67
	0.990	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.81	5.67	5.52	5.36	5.20	5.07	4.86
	0.999	18.5	15.8	14.4	13.5	12.9	12.4	12.0	11.5	11.2	10.8	10.5	10.1	9.80	9.33

PERCENTAGE POINTS OF THE F DISTRIBUTION

$\nu_2 \backslash \nu_1$	q	2	3	4	5	6	7	8	10	12	15	20	30	50	∞
9	0.900	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.42	2.38	2.34	2.30	2.25	2.22	2.16
	0.950	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.14	3.07	3.01	2.94	2.86	2.80	2.71
	0.975	5.71	5.08	4.72	4.48	4.32	4.20	4.10	3.96	3.87	3.77	3.67	3.56	3.47	3.33
	0.990	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.26	5.11	4.96	4.81	4.65	4.52	4.31
	0.999	16.4	13.9	12.6	11.7	11.1	10.7	10.4	9.89	9.57	9.24	8.90	8.55	8.26	7.81
10	0.900	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.32	2.28	2.24	2.20	2.16	2.12	2.06
	0.950	4.10	3.71	3.48	3.33	3.22	3.14	3.07	2.98	2.91	2.84	2.77	2.70	2.64	2.54
	0.975	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.72	3.62	3.52	3.42	3.31	3.22	3.08
	0.990	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.85	4.71	4.56	4.41	4.25	4.11	3.91
	0.999	14.9	12.6	11.3	10.5	9.93	9.52	9.20	8.75	8.45	8.13	7.80	7.47	7.19	6.76
11	0.900	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.25	2.21	2.17	2.12	2.08	2.04	1.97
	0.950	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.85	2.79	2.72	2.65	2.57	2.51	2.40
	0.975	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.53	3.43	3.33	3.23	3.12	3.03	2.88
	0.990	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.54	4.40	4.25	4.10	3.94	3.81	3.60
	0.999	13.8	11.6	10.3	9.58	9.05	8.66	8.35	7.92	7.63	7.32	7.01	6.68	6.42	6.00
12	0.900	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.19	2.15	2.10	2.06	2.01	1.97	1.90
	0.950	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.75	2.69	2.62	2.54	2.47	2.40	2.30
	0.975	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.37	3.28	3.18	3.07	2.96	2.87	2.72
	0.990	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.30	4.16	4.01	3.86	3.70	3.57	3.36
	0.999	13.0	10.8	9.63	8.89	8.38	8.00	7.71	7.29	7.00	6.71	6.40	6.09	5.83	5.42
13	0.900	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.14	2.10	2.05	2.01	1.96	1.92	1.85
	0.950	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.67	2.60	2.53	2.46	2.38	2.31	2.21
	0.975	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.25	3.15	3.05	2.95	2.84	2.74	2.60
	0.990	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.10	3.96	3.82	3.66	3.51	3.37	3.17
	0.999	12.3	10.2	9.07	8.35	7.86	7.49	7.21	6.80	6.52	6.23	5.93	5.63	5.37	4.97
14	0.900	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.10	2.05	2.01	1.96	1.91	1.87	1.80
	0.950	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.60	2.53	2.46	2.39	2.31	2.24	2.13
	0.975	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.15	3.05	2.95	2.84	2.73	2.64	2.49
	0.990	6.51	5.56	5.04	4.69	4.46	4.28	4.14	3.94	3.80	3.66	3.51	3.35	3.22	3.00
	0.999	11.8	9.73	8.62	7.92	7.44	7.08	6.80	6.40	6.13	5.85	5.56	5.25	5.00	4.60
15	0.900	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.06	2.02	1.97	1.92	1.87	1.83	1.76
	0.950	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.54	2.48	2.40	2.33	2.25	2.18	2.07
	0.975	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.06	2.96	2.86	2.76	2.64	2.55	2.40
	0.990	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.80	3.67	3.52	3.37	3.21	3.08	2.87
	0.999	11.3	9.34	8.25	7.57	7.09	6.74	6.47	6.08	5.81	5.53	5.25	4.95	4.70	4.31
16	0.900	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.03	1.99	1.94	1.89	1.84	1.79	1.72
	0.950	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.49	2.42	2.35	2.28	2.19	2.12	2.01
	0.975	4.69	4.08	3.73	3.50	3.34	3.22	3.12	2.99	2.89	2.79	2.68	2.57	2.47	2.32
	0.990	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.69	3.55	3.41	3.26	3.10	2.97	2.75
	0.999	11.0	9.01	7.94	7.27	6.80	6.46	6.19	5.81	5.55	5.27	4.99	4.70	4.45	4.06
17	0.900	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.00	1.96	1.91	1.86	1.81	1.76	1.69
	0.950	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.45	2.38	2.31	2.23	2.15	2.08	1.96
	0.975	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.92	2.82	2.72	2.62	2.50	2.41	2.25
	0.990	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.59	3.46	3.31	3.16	3.00	2.87	2.65
	0.999	10.7	8.73	7.68	7.02	6.56	6.22	5.96	5.58	5.32	5.05	4.77	4.48	4.24	3.85

5.7.6.21 General Folded Normal Distribution

The "folded Normal distribution" is the distribution of the absolute value of a random variable with a Normal distribution[39]. As we have mentioned before, the Normal distribution is perhaps the most important in probability and is used to model an incredible variety of random phenomena. Since one may only be interested in the magnitude of a Normally distributed variable, the folded Normal arises in a very natural way especially in Finance and Industrial Engineering for Design of Experiments (see section Industrial Engineering page 8780). The name stems from the fact that the probability measure of the Normal distribution on $]-\infty,0]$ is folded over to $[0,+\infty[$.

Definition (#5.7.134): Suppose that X has a Normal distribution with mean $\mu\in\mathbb{R}$ and standard deviation $\sigma\in[0,+\infty[$. Then $Y=|X|$ has the fold Normal distribution with parameters μ and σ.

Suppose that Z follows the standard Normal distribution. Let us recall that then Z has probability density function ϕ and distribution function Φ given by:

$$\begin{aligned}\phi(z)&=\frac{1}{\sqrt{2\pi}}e^{-z^2/2}\\ \Phi(z)&=\int\limits_{-\infty}^{z}\phi(x)\mathrm{d}x=\int\limits_{-\infty}^{z}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\mathrm{d}x\end{aligned}\tag{5.7.1088}$$

with $z\in\mathbb{R}$.

If $\mu\in\mathbb{R}$ and $\sigma\in[0,+\infty[$, then $X=\mu+\sigma Z$ has the Normal distribution with mean μ and standard deviation σ, and therefore it is obvious at this level that:

$$Y=|X|=|\mu+\sigma Z|\tag{5.7.1089}$$

has the folded Normal distribution with parameters μ and σ.

Now let us determine the cumulative probability CDF function of such a variable! For $y\in[0,+\infty[$:

$$\begin{aligned}F(y)=P(Y\leq y)&=P(|X|\leq y)=P(|\mu+\sigma Z|\leq y)=P(-y\leq\mu+\sigma Z\leq y)\\ &=P\left(\frac{-y-\mu}{\sigma}\leq Z\leq\frac{y-\mu}{\sigma}\right)=\Phi\left(\frac{y-\mu}{\sigma}\right)-\Phi\left(\frac{-y-\mu}{\sigma}\right)\end{aligned}\tag{5.7.1090}$$

Since $\Phi(-Z)=1-\Phi(Z)$ we have:

$$\begin{aligned}F(y)&=\Phi\left(\frac{y-\mu}{\sigma}\right)-\Phi\left(\frac{-y-\mu}{\sigma}\right)=\Phi\left(\frac{y-\mu}{\sigma}\right)+\Phi\left(\frac{y+\mu}{\sigma}\right)-1\\ &=\int\limits_{0}^{y}\frac{1}{\sigma\sqrt{2\pi}}\left\{e^{-\frac{1}{2}\left(\frac{x+\mu}{\sigma}\right)^2}+e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\right\}\mathrm{d}x\end{aligned}\tag{5.7.1091}$$

We cannot compute the quantile function F^{-1} in closed form, but values of this function can be approximated.

---

[39] The majority of the text below comes from `http://www.math.uah.edu/stat/`

It comes therefore immediately that Y has probability density function f given by:

$$\boxed{f(|Y|)=\frac{1}{\sigma\sqrt{2\pi}}\left\{e^{-\frac{1}{2}\left(\frac{x+\mu}{\sigma}\right)^2}+e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\right\}} \tag{5.7.1092}$$

This follow from differentiating the CDF with respect to y as we know!

Now as always in this textbook we will focus only in what we need for the applications in the other chapters! So as we don't need the moments of the folder Normal distribution we will not calculate them. The only purpose of the above development were to build the tools to be able to introduce a special case of the folder Normal distribution.

5.7.6.21.1 Half-Normal distribution

In probability theory and statistics, the "half-Normal distribution" is a special case of the folded Normal distribution.

Let X follow an ordinary Normal distribution, $\mathcal{N}(0,\sigma^2)$, then $Y=|X|$ follows a half-Normal distribution. Thus, the half-Normal distribution is a fold at the mean of an ordinary Normal distribution with mean $\mu=0$.

Thus, let $Y=|\sigma Z|=\sigma|Z|$ where Z has a standard Normal distribution and $\sigma\in[0,+\infty[$. Clearly σ is a scale parameter, unlike the case for the general folded Normal distribution. The distribution of Y when $\sigma=1$, $Y=|Z|$ has the "standard half-Normal distribution".

As the half-Normal distribution is just a special case of the folded Normal distribution with $\mu=0$ it comes immediately:

$$\boxed{f(y)=\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}e^{-\frac{y^2}{2\sigma^2}}} \tag{5.7.1093}$$

with $y\in\mathbb{R}^+$ and sometimes denoted $\mathcal{HN}(0,\sigma^2)$.

Now what interest us for the others chapters of this textbook (especially Design of Experiments) are the moments of that latter!

To calculate them, first remember that:

$$\phi(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2} \tag{5.7.1094}$$

with $z\in\mathbb{R}$. Therefore it is immediate that:

$$\phi'(z)=-z\phi(z) \tag{5.7.1095}$$

Also remember that we have already proved that:

$$\text{E}(Z)=0 \qquad \text{E}(Z^2)=\text{V}(Z)=1 \tag{5.7.1096}$$

We first need to determiner the moments for the Normal distribution! So for $n \in \mathbb{N}^+$:

$$\mathrm{E}(Z^{n+1}) = \int\limits_{-\infty}^{+\infty} z^{n+1}\phi(z)\mathrm{d}z = \int\limits_{-\infty}^{+\infty} z^n z\phi(z)\mathrm{d}z = -\int\limits_{-\infty}^{+\infty} z^n\phi'(z)\mathrm{d}z \tag{5.7.1097}$$

Now we integrate by parts (see section Differential and Integral Calculus page 1532), with $u = z^n$ and $\mathrm{d}v = \phi'(z)\mathrm{d}z$ to get:

$$\mathrm{E}(Z^{n+1}) = -z^n\phi(z)\Big|_{-\infty}^{+\infty} + \int\limits_{-\infty}^{+\infty} nz^{n-1}\phi(z)\mathrm{d}z = 0 + n\mathrm{E}(Z^{n-1}) \tag{5.7.1098}$$

Therefore for $n \in \mathbb{N}$, with $n > 1$:

$$\mathrm{E}(Z^{n+1}) = n\mathrm{E}(Z^{n-1}) \tag{5.7.1099}$$

The moments of the standard Normal distribution are now easy to compute. First we know that:

- $\mathrm{E}(Z) = 0$
- $\mathrm{E}(Z^2) = 1$

Therefore:

- $\mathrm{E}(Z) = 0$
- $\mathrm{E}(Z^2) = 1$
- $\mathrm{E}(Z^3) = \mathrm{E}(Z^{2+1}) = 2 \cdot \mathrm{E}(Z^{2-1}) = 2 \cdot \mathrm{E}(Z) = 0$
- $\mathrm{E}(Z^4) = \mathrm{E}(Z^{3+1}) = 3 \cdot \mathrm{E}(Z^{3-1}) = 3 \cdot \mathrm{E}(Z^2) = 3 \cdot 1 = 1 \cdot 3 = 3$
- $\mathrm{E}(Z^5) = \mathrm{E}(Z^{4+1}) = 4 \cdot \mathrm{E}(Z^{4-1}) = 4 \cdot \mathrm{E}(Z^3) = 4 \cdot 0 = 0$
- $\mathrm{E}(Z^6) = \mathrm{E}(Z^{5+1}) = 5 \cdot \mathrm{E}(Z^{5-1}) = 5 \cdot \mathrm{E}(Z^4) = 5 \cdot 3 = 1 \cdot 3 \cdot 5 = 15$
- $\mathrm{E}(Z^7) = \mathrm{E}(Z^{6+1}) = 6 \cdot \mathrm{E}(Z^{6-1}) = 6 \cdot \mathrm{E}(Z^5) = 6 \cdot 0 = 0$
- $\mathrm{E}(Z^8) = \mathrm{E}(Z^{7+1}) = 7 \cdot \mathrm{E}(Z^{7-1}) = 7 \cdot \mathrm{E}(Z^6) = 7 \cdot 15 = 1 \cdot 3 \cdot 5 \cdot 7$
- $\ldots$

Therefore we see that for the odd powers, that is to say Z^{2n+1} with $n \in \mathbb{N}$ then:

$$\mathrm{E}(Z^{2n+1}) = 0 \tag{5.7.1100}$$

and for even powers:

$$\mathrm{E}(Z^{2n}) = 1 \cdot 3 \cdot \ldots \cdot (2n-1) = \frac{(2n)!}{n!2^n} \tag{5.7.1101}$$

It follows for $X=\mathcal{N}(0,\sigma)$ (just check with the special case of $n=0$ and $n=1$):

$$\boxed{\mathrm{E}(X^{2n+1})=0 \qquad \mathrm{E}(X^{2n})=\sigma^{2n}\frac{(2n)!}{n!2^n}} \tag{5.7.1102}$$

The moments of the half-Normal distribution can now be computed explicitly.

First it should be quite obvious by construction that the even order moments of Y are the same as the even order moments of Z (in both case the values are all positive and therefore equal). Hence:

$$\mathrm{E}(Y^{2n})=\sigma^{2n}\frac{(2n)!}{n!2^n} \tag{5.7.1103}$$

For the odd order moments we must use (see above):

$$f(Y)=\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}e^{-\frac{y^2}{2\sigma^2}} \tag{5.7.1104}$$

with $x\in\mathbb{R}^+$. Therefore by definition:

$$\mathrm{E}(Y^{2n+1})=\int\limits_0^{+\infty} y^{2n+1}f(y)\mathrm{d}y=\int\limits_0^{+\infty} y^{2n+1}\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}e^{-\frac{y^2}{2\sigma^2}}\mathrm{d}y \tag{5.7.1105}$$

hence:

$$\mathrm{E}(Y^{2n+1})=\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\int\limits_0^{+\infty} y^{2n+1}e^{-\frac{y^2}{2\sigma^2}}\mathrm{d}y \tag{5.7.1106}$$

Now we make the change of variable $u=y^2/(2\sigma^2)$, therefore first we have (this is obvious):

$$e^{-\frac{y^2}{2\sigma^2}}=e^{-u} \tag{5.7.1107}$$

and:

$$\mathrm{d}u=\frac{2y\mathrm{d}y}{2\sigma^2}\Leftrightarrow \mathrm{d}y=\frac{\sigma^2}{y}\mathrm{d}u \tag{5.7.1108}$$

Therefore we have so far:

$$\sigma\sqrt{\frac{2}{\pi}}\int\limits_0^{+\infty} y^{2n+1}\frac{1}{y}e^{-u}\mathrm{d}u=\sigma\sqrt{\frac{2}{\pi}}\int\limits_0^{+\infty}\left(y^2\right)^n e^{-u}\mathrm{d}u \tag{5.7.1109}$$

So finally:

$$\mathrm{E}(Y^{2n+1})=\sigma\sqrt{\frac{2}{\pi}}\int\limits_0^{+\infty}\left(2\sigma^2u\right)^n e^{-u}\mathrm{d}u=\sigma^{2n+1}2^n\sqrt{\frac{2}{\pi}}\int\limits_0^{+\infty}u^ne^{-u}\mathrm{d}u \tag{5.7.1110}$$

We recognize in this expression the Gamma Euler function integral (see section Differential and Integral Calculus page 1574)! Therefore it is immediate that:

$$\mathrm{E}(Y^{2n+1})=\sigma\sqrt{\frac{2}{\pi}}\Gamma_0(u)=\sigma^{2n+1}2^n\sqrt{\frac{2}{\pi}}n! \tag{5.7.1111}$$

So as summary:

$$\mathrm{E}(Y^{2n})=\sigma^{2n}\frac{(2n)!}{n!2^n} \qquad \mathrm{E}(Y^{2n+1})=\sigma^{2n+1}2^n\sqrt{\frac{2}{\pi}}n! \tag{5.7.1112}$$

So finally we get the result we need for some properties of the Brownian motion in finance:

$$\begin{aligned}&\mathrm{E}(Y)=\mathrm{E}(Y^{2\cdot 0+1})=\sigma\sqrt{\frac{2}{\pi}}\\&\mathrm{V}(Y)=\mathrm{E}(Y^2)-(\mathrm{E}(Y))^2=\mathrm{E}(Y^{2\cdot 1})-(\mathrm{E}(Y))^2=\sigma^2-\left(\sigma\sqrt{\frac{2}{\pi}}\right)^2=\sigma^2\left(1-\frac{2}{\pi}\right)\end{aligned} \tag{5.7.1113}$$

But still one property is missing and now for our needs in the section of Industrial Engineering (page 8780): the value of the Median!

So let M_e denote the median of the half-Normal distribution. Then by definition it follows that:

$$0.50=F(M_e)=\int\limits_0^{M_e} f(y)\mathrm{d}y=\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\int\limits_0^{M_e} e^{-\frac{1}{2}\left(\frac{y}{\sigma}\right)^2}\mathrm{d}y \tag{5.7.1114}$$

Substituting $y/(\sqrt{2}\sigma)=u$ we have:

$$0.50=F(M_e)=\frac{2}{\sqrt{\pi}}\int\limits_0^{\frac{M_e}{\sqrt{2}\sigma}} e^{-u^2}\mathrm{d}u \tag{5.7.1115}$$

We recognize here the Error function (see section Thermodynamics page 3697). Therefore:

$$0.50=F(M_e)=\mathrm{erf}\left(\frac{M_e}{\sqrt{2}\sigma}\right) \tag{5.7.1116}$$

Therefore:

$$\mathrm{erf}^{-1}(0.50)=\frac{M_e}{\sqrt{2}\sigma} \tag{5.7.1117}$$

A spreadsheet software like Microsoft® Excel® gives us for the complementary Error function:

```
=ERFC(0.5)=0.479500122
```

Therefore:

$$\boxed{M_e\cong 0.479500122\sqrt{2}\sigma\cong 0.67\sigma} \tag{5.7.1118}$$

The technique that we will see in the section Industrial Engineering (page 8896) makes the approximation that therefore:

$$1.5\cdot M_e\cong\sigma \tag{5.7.1119}$$

indeed... it's engineering...

5.7.6.22 Laplace Distribution

In probability theory and statistics, the "Laplace distribution" is a continuous probability distribution named also sometimes named the "double exponential distribution", because it can be thought of as two exponential distributions (with an additional location parameter) spliced together back-to-back (although the term is also sometimes used to refer to the Gumbel distribution). This distribution has many applications but the one that interest us in this textbook is the LASSO regression (see section Numerical Methods page 6441).

The difference between two independent identically distributed exponential random variables is governed by a Laplace distribution.

Proof 5.47.8. Let X_1 and X_2 be independent exponential random variable with population means α_1 and α_2 respectively. Define $Y = X_1 - X_2$. The goal is to find the distribution of Y by the cumulative distribution function technique. First we know that:

$$f_{X_1}(x_1) = \frac{1}{\alpha_1}e^{-x_1/\alpha_1} \qquad x_1 > 0 \tag{5.7.1120}$$

and:

$$f_{X_2}(x_x) = \frac{1}{\alpha_2}e^{-x_2/\alpha_2} \qquad x_2 > 0 \tag{5.7.1121}$$

By independence, it follows that the joint probability density function of X_1 and X_2 is:

$$f_{X_1,X_2}(x_1,x_2) = \frac{1}{\alpha_1\alpha_2}e^{-x_1/\alpha_1}e^{-x_2/\alpha_2} \qquad x_1 > 0, x_2 > 0 \tag{5.7.1122}$$

We want $F_Y(y) = P(Y \leq y)$ where $-\infty < y < +\infty$. Looking at the support of $f_{X_1,X_2}(x_1,x_2)$ given above, we see that the resulting cumulative distribution function for Y will (must) be piecewise, where the pieces are separated at $y = 0$. So we first consider the case where $y \leq 0$. The double integral giving us the cumulative distribution function $F_Y(y)$ of $y \leq 0$ is:

$$\begin{aligned}
F_Y(y) &= P(Y \leq y) = P(X_1 - X_2 \leq y) = P(X_2 \geq X_1 - y)\\
&= \int_0^{+\infty}\int_{x_1-y}^{+\infty} f_{X_1,X_2}(x_1,x_2)\mathrm{d}x_2\mathrm{d}x_1 = \int_0^{+\infty}\int_{x_1-y}^{+\infty}\frac{1}{\alpha_1\alpha_2}e^{-x_1/\alpha_1}e^{-x_2/\alpha_2}\mathrm{d}x_2\mathrm{d}x_1\\
&= \frac{\alpha_2}{\alpha_1+\alpha_2}e^{y/\alpha_2} \qquad y \leq 0
\end{aligned} \tag{5.7.1123}$$

For $y > 0$, the cumulative distribution function of Y is:

$$\begin{aligned}
F_Y(y) &= P(Y \leq y) = 1 - P(Y > y) = 1 - P(X_1 - X_2 > y) = 1 - P(X_2 \geq X_1 - y)\\
&= 1 - \int_y^{+\infty}\int_0^{x_1-y} f_{X_1,X_2}(x_1,x_2)\mathrm{d}x_2\mathrm{d}x_1 = 1 - \int_y^{+\infty}\int_0^{x_1-y}\frac{1}{\alpha_1\alpha_2}e^{-x_1/\alpha_1}e^{-x_2/\alpha_2}\mathrm{d}x_2\mathrm{d}x_1\\
&= 1 - \frac{\alpha_1}{\alpha_1+\alpha_2}e^{-y/\alpha_1} \qquad y > 0
\end{aligned} \tag{5.7.1124}$$

Differentiating with respect to y gives:

$$f_Y(y)=\begin{cases}\dfrac{e^{y/\alpha_2}}{\alpha_1+\alpha_2} & \text{if } y\leq 0\\ \dfrac{e^{-y/\alpha_1}}{\alpha_1+\alpha_2} & \text{if } y>0\end{cases} \tag{5.7.1125}$$

which is the probability density function of a Laplace random variable with parameters α_1 and α_2.

■ Q.E.D.

A random variable has a $\mathcal{L}(\mu,b)$ distribution if its probability density function is:

$$\boxed{\mathcal{L}(x\mid\mu,b)=\frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)=\frac{1}{2b}\begin{cases}\exp\left(-\frac{\mu-x}{b}\right) & \text{if } x<\mu\\ \exp\left(-\frac{x-\mu}{b}\right) & \text{if } x\geq\mu\end{cases}} \tag{5.7.1126}$$

Here, μ is a location parameter and $b>0$, which is sometimes referred to as the diversity, is a scale parameter. If $\mu=0$ and $b=1$, the positive half-line is exactly an exponential distribution scaled by $1/2$.

Notice that the Laplace distribution is often denoted:

$$\mathcal{L}(x\mid\mu,\sigma)=\frac{1}{2\sigma}\exp\left(-\frac{|x-\mu|}{\sigma}\right) \tag{5.7.1127}$$

The probability density function of the Laplace distribution is also reminiscent of the Normal distribution; however, whereas the Normal distribution is expressed in terms of the squared difference from the mean μ, the Laplace density is expressed in terms of the absolute difference from the mean. Consequently, the Laplace distribution has fatter tails than the Normal distribution.

The Laplace distribution is easy to integrate (if one distinguishes two symmetric cases) due to the use of the absolute value function. Its cumulative distribution function is as follows:

$$\begin{aligned}F(x)&=\int\limits_{-\infty}^{x} f(u)\,\mathrm{d}u=\begin{cases}\frac{1}{2}\exp\left(\frac{x-\mu}{b}\right) & \text{if } x<\mu\\ 1-\frac{1}{2}\exp\left(-\frac{x-\mu}{b}\right) & \text{if } x\geq\mu\end{cases}\\ &=\frac{1}{2}+\frac{1}{2}\text{sng}(x-\mu)\left(1-\exp\left(-\frac{|x-\mu|}{b}\right)\right)\end{aligned} \tag{5.7.1128}$$

Let us now derive the mean and variance of the Laplace distribution adopting the following notation:

$$\mathcal{L}(x\mid\mu,\sigma)=\frac{1}{2\sigma}\exp\left(-\frac{|x-\mu|}{\sigma}\right) \tag{5.7.1129}$$

Let $u = x - \mu$, then we have $x = u + \mu$ and $\mathrm{d}x = \mathrm{d}u$. By linearity we have $\mathrm{E}(X) = \mathrm{E}(U+\mu) = \mathrm{E}(U) + \mu$, hence:

$$\begin{aligned}
\mathrm{E}[X] &= \frac{1}{2\sigma}\int\limits_{-\infty}^{+\infty} u\, e^{-\frac{|u|}{\sigma}}\,\mathrm{d}u + \mu \\
&= \underbrace{\frac{1}{2\sigma}\int\limits_{-\infty}^{0} u\, e^{\frac{u}{\sigma}}\,\mathrm{d}u}_{\text{set } u=-u} + \frac{1}{2\sigma}\int\limits_{0}^{+\infty} u\, e^{-\frac{u}{\sigma}}\,\mathrm{d}u + \mu \\
&= -\frac{1}{2\sigma}\int\limits_{0}^{+\infty} u\, e^{-\frac{u}{\sigma}}\,\mathrm{d}u + \frac{1}{2\sigma}\int\limits_{0}^{+\infty} u\, e^{-\frac{u}{\sigma}}\,\mathrm{d}u + \mu \\
&= \mu
\end{aligned} \tag{5.7.1130}$$

and:

$$\begin{aligned}
\mathrm{E}\left[X^2\right] &= \mathrm{E}\left[(U+\mu)^2\right] = \mathrm{E}\left[U^2\right] + 2\mu\mathrm{E}\left[U\right] + \mu^2 = \mathrm{E}\left[U^2\right] + 2\mu\,\mathrm{E}\left[X-\mu\right] + \mu^2 \\
&= \frac{1}{2\sigma}\int\limits_{-\infty}^{+\infty} u^2 e^{-\frac{|u|}{\sigma}}\,\mathrm{d}u + \mu^2 = \underbrace{\frac{1}{2\sigma}\int\limits_{-\infty}^{0} u^2 e^{\frac{u}{\sigma}}\,\mathrm{d}u}_{\text{set } u=-u} + \frac{1}{2\sigma}\int\limits_{0}^{+\infty} u^2 e^{-\frac{u}{\sigma}}\,\mathrm{d}u + \mu^2 \\
&= \frac{1}{2\sigma}\int\limits_{0}^{+\infty} u^2 e^{-\frac{u}{\sigma}}\,\mathrm{d}u + \frac{1}{2\sigma}\int\limits_{0}^{+\infty} u^2 e^{-\frac{u}{\sigma}}\,\mathrm{d}u + \mu^2 \\
&= \underbrace{\frac{1}{\sigma}\int\limits_{0}^{+\infty} u^2 e^{-\frac{u}{\sigma}}\,\mathrm{d}u}_{\text{set } v=\frac{u}{\sigma}} + \mu^2 = \sigma^2\int\limits_{0}^{+\infty} v^2 e^{-v}\,dv + \mu^2 \\
&= 2\sigma^2 + \mu^2
\end{aligned} \tag{5.7.1131}$$

where:

$$\Gamma(n+1) = \int\limits_{0}^{+\infty} v^n e^{-v}\,\mathrm{d}v = n! \qquad \text{for } n \text{ natural number.} \tag{5.7.1132}$$

Thus:

$$\mathrm{V}[X] = \mathrm{E}\left[X^2\right] - \left(\mathrm{E}[X]\right)^2 = 2\sigma^2 \tag{5.7.1133}$$

5.7.6.23 Benford Distribution

This distribution was discovered first in 11881 (holocene calendar) by Simon Newcomb, an American astronomer, after he saw that the wear (and so the use) of the preferred first pages of logarithms tables (at this time there we compiled into books). Frank Benford, around 11938 (holocene calendar) remarked at his turn this unequal wear, believing he was the first to formulate this law that unduly bears his name today and arrived at the same results after having listed tens of thousands of data (lengths of rivers , stock quotes, etc.).

There is also one possible explanation: we need more often to extract the logarithm of numbers starting with 1 that numbers starting with 9, implying that the first are in "bigger quantity" than the second one.

Although this idea may seem to him quite implausible, Benford began to test his hypothesis. Nothing more simple: he studies tables of numerical values and calculates the percentage of occurrence of the left-most digit (first decimal). The results obtained confirm his intuition:

First position number	1	2	3	4	5	6	7	8	9
Apparition probability (%)	30.1	17.6	12.5	9.7	7.9	6.7	5.8	5.1	4.6

Table 5.44: **Occurrence of a digit following the Benford distribution**

From these data, Benford found experimentally that the cumulative probability of a number beginning with the digit n (except 0) is (we will prove this later) is given by the relation:

$$P(n)=\log_{10}\left(1+\frac{1}{n}\right) \tag{5.7.1134}$$

named "Benford distribution" (or "Benford law").

Here is a Maple™ plot of the previous function:

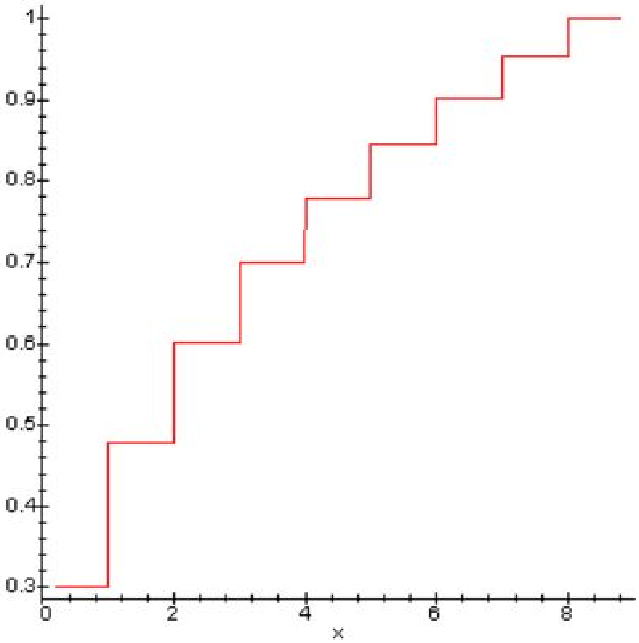


Figure 5.152: **Plot of the Benford function (cumulative distribution function)**

It should be noted that this distribution applies only to lists of values that are "natural", that is to say numbers with physical meaning. It obviously does not work on a list of numbers randomly drawn.

The Benford distribution has been tested on all kinds of tables: length of the rivers of the world, country area, election results, price list of grocery store ... It is true almost every time.

The distribution is said to be independent of the selected unit. If we take for example a supermarket price list , it also works well with the costs expressed in dollars as with the same costs converted into Euros.

This strange phenomenon remained unexplained and little studied until quite recently. Then a general proof was given in 11996 (holocene calendar), which uses the central limit theorem.

As surprising as it may seem, this distribution has found application: it is said that the IRS use it to detect false statements. The principle is based on the restriction seen above: Benford's distributions applies only to values with physical meaning.

Thus, if there is a universal probability distribution $P(n)$ on such numbers, they should be invariant under scaling such that:

$$P(kn)=f(k)P(n) \tag{5.7.1135}$$

If:

$$\int P(n)\mathrm{d}n=1 \tag{5.7.1136}$$

Then:

$$\int P(kn)\mathrm{d}n=\frac{1}{k} \tag{5.7.1137}$$

and the normalization of the distribution gives:

$$f(k)=\frac{1}{k} \tag{5.7.1138}$$

If we derive $P(kn)=f(k)P(n)$ with respect to k we obtain:

$$\frac{\mathrm{d}}{\mathrm{d}k}P(kn)=\frac{\mathrm{d}}{\mathrm{d}k}\frac{1}{k}P(n)\Rightarrow nP'(kn)=P(n)\frac{\mathrm{d}}{\mathrm{d}k}\frac{1}{k}\Rightarrow nP'(kn)=-P(n)\frac{1}{k^2} \tag{5.7.1139}$$

choosing $k=1$ we have:

$$nP'(n)=-P(n) \tag{5.7.1140}$$

This differential equation has for solution:

$$P(n)=\frac{1}{n} \tag{5.7.1141}$$

This function is not strictly speaking a distribution function (it diverges) and secondly, the physics and human laws impose limits.

So we have to compare this distribution with respect to an arbitrary reference. Thus, if the decimal number studied contains power of 10 (10 in total: $0,1,2,3,4,5,6,7,9$) the probability that the first non-zero digit (decimal) is D is also given by:

$$P_D=\frac{\int\limits_D^{D+1}P(n)\mathrm{d}n}{\int\limits_1^{10}P(n)\mathrm{d}n} \tag{5.7.1142}$$

The limits of the integral are from 1 to 10 because the null value is prohibited.

The integral in the denominator gives:

$$\int\limits_0^{10}P(n)\mathrm{d}n=\int\limits_0^{10}\frac{1}{n}\mathrm{d}n=\ln(10)-\ln(1)=\ln(10) \tag{5.7.1143}$$

The integral in the numerator gives:

$$\int\limits_D^{D+1}P(n)\mathrm{d}n=\int\limits_D^{D+1}\frac{1}{n}\mathrm{d}n=\ln(D+1)-\ln(D)=\ln\left(\frac{D+1}{D}\right) \tag{5.7.1144}$$

Finally:

$$P_D=\frac{\ln\left(\dfrac{D+1}{D}\right)}{\ln(10)}=\frac{\ln\left(1+\dfrac{1}{D}\right)}{\ln(10)} \tag{5.7.1145}$$

By the properties of logarithms (see section Functional Analysis page 2284) we have:

$$\boxed{P_D=\log_{10}\left(1+\frac{1}{D}\right)} \tag{5.7.1146}$$

However, the Benford's distribution applies not only to non-scaling data but also to numbers from any sources. Explain this case involves a more rigorous investigation using the central limit theorem. This demonstration was conducted only in 11996 (holocene calendar) by T. Hill by an approach using the distribution of distributions.

5.7.6.24 ex-Gaussian distribution

In probability theory, an "exponentially modified Gaussian distribution" (EMG, also known as "exGaussian distribution") describes the sum of independent Normal and exponential random variables. An exGaussian random variable Z may be expressed as $Z = X + Y$, where X and Y are independent, X is Gaussian with mean μ and variance σ^2 and Y is exponential of rate λ. It has a characteristic positive skew from the exponential component.

It may also be regarded as a weighted function of a shifted exponential with the weight being a function of the Normal distribution.

Remark

The distribution is used as a theoretical model for the shape of chromatographic peaks. It has been proposed as a statistical model of intermitotic time in dividing cells. It is also used in modelling cluster ion beams. It is commonly used in psychology and other brain sciences in the study of response times In a slight variant where the mean of the Normal component is set to zero, it is also used in Stochastic Frontier Analysis, as one of the distributional specifications for the composed error term that models inefficiency. In signal processing, EMGs have been extended to the multimodal case with an optional oscillation term to represent digitized sound signals.

More technically, let A be a random variable that is Normally distributed with mean μ and variance σ^2, and let B be a random variable that is exponentially distributed with rate λ. Suppose further that A and B are independent. Then the sum $X = A + B$ is said to have an exponentially-modified Gaussian (i.e., ex-Gaussian) distribution, with parameters μ, σ, and λ; that is:

$$X \sim \text{ex-Gaussian}(\mu, \sigma, \lambda) \tag{5.7.1147}$$

where $\mu \in \mathbb{R}, \sigma > 0$, and $\lambda > 0$.

The probability density function of X is:

$$\boxed{\text{ex-Gaussian}(\mu, \sigma, \lambda) = \frac{\lambda}{\sqrt{2\pi}} \exp\left[\frac{\lambda^2\sigma^2}{2} - \lambda(t-\mu)\right] \cdot \int\limits_{-\infty}^{\frac{t-\mu}{\sigma}-\lambda\sigma} \exp\left[-\frac{1}{2}y^2\right] \mathrm{d}y} \tag{5.7.1148}$$

Indeed, suppose X follows an ex-Gaussian distribution. Then $X = A + B$, where A and B are independent A is normally distributed with mean μ and variance σ^2, and B is exponentially distributed with rate λ. Then, the probability density function for A is given by:

$$f_A(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right] \tag{5.7.1149}$$

and the probability density function for B is given by:

$$f_B(t) = \begin{cases} \lambda \exp[-\lambda t], \text{ if } t \geq 0 \\ 0, \text{ if } t < 0 \end{cases} \tag{5.7.1150}$$

Thus, the probability density function for the sum $X = A+B$ is given by taking the convolution of f_A and f_B :

$$\begin{aligned} f_X(t) &= \int_{-\infty}^{+\infty} f_A(x) f_B(t-x)\mathrm{d}x \\ &= \int_{-\infty}^{t} f_A(x) f_B(t-x)dx + \int_{t}^{\infty} f_A(x) f_B(t-x)\mathrm{d}x \\ &= \int_{-\infty}^{t} f_A(x) f_B(t-x)\mathrm{d}x \end{aligned} \tag{5.7.1151}$$

which follows from the fact that $f_B(t-x) = 0$ for $x > t$. From here, we substitute these expressions for the probability density functions f_A and f_B:

$$\begin{aligned} f_X(t) &= \int_{-\infty}^{t} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \cdot \lambda \exp[-\lambda(t-x)]\mathrm{d}x \\ &= \frac{\lambda}{\sigma\sqrt{2\pi}} \int_{-\infty}^{t} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \cdot \exp[-\lambda t + \lambda x]\mathrm{d}x \\ &= \frac{\lambda}{\sigma\sqrt{2\pi}} \int_{-\infty}^{t} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \cdot \exp[-\lambda t] \cdot \exp[\lambda x]\mathrm{d}x \\ &= \frac{\lambda \exp[-\lambda t]}{\sigma\sqrt{2\pi}} \int_{-\infty}^{t} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 + \lambda x\right] \mathrm{d}x \end{aligned} \tag{5.7.1152}$$

We can further simplify the integrand with a substitution; to this end, let:

$$y = \frac{x-\mu}{\sigma} - \lambda\sigma \tag{5.7.1153}$$

This gives the following three identities:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{1}{\sigma}, \qquad \text{or equivalently,} \qquad \mathrm{d}x = \sigma\mathrm{d}y \tag{5.7.1154}$$

and:

$$\frac{x-\mu}{\sigma} = y + \lambda\sigma \tag{5.7.1155}$$

and:

$$x = y\sigma + \lambda\sigma^2 + \mu \tag{5.7.1156}$$

Substituting these identities into:

$$f_X(t) = \frac{\lambda \exp[-\lambda t]}{\sigma\sqrt{2\pi}} \int_{-\infty}^{t} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 + \lambda x\right] \mathrm{d}x \tag{5.7.1157}$$

gives:

$$\begin{aligned}
f_X(t) &= \frac{\lambda\exp[-\lambda t]}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{\frac{x-\mu}{\sigma}+\lambda\sigma} \exp\left[-\frac{1}{2}(y+\lambda\sigma)^2+\lambda\left(y\sigma+\lambda\sigma^2+\mu\right)\right]\sigma\mathrm{d}y\\
&= \frac{\lambda\sigma\exp[-\lambda t]}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{\frac{x-\mu}{\sigma}+\lambda\sigma} \exp\left[-\frac{1}{2}\left(y^2+2y\lambda\sigma+\lambda^2\sigma^2\right)+\lambda y\sigma+\lambda^2\sigma^2+\lambda\mu\right]\mathrm{d}y\\
&= \frac{\lambda\exp[-\lambda t]}{\sqrt{2\pi}} \int\limits_{-\infty}^{\frac{x-\mu}{\sigma}+\lambda\sigma} \exp\left[-\frac{1}{2}y^2-y\lambda\sigma-\frac{\lambda^2\sigma^2}{2}+\lambda y\sigma+\lambda^2\sigma^2+\lambda\mu\right]dy\\
&= \frac{\lambda\exp[-\lambda t]}{\sqrt{2\pi}} \int\limits_{-\infty}^{\frac{x-\mu}{\sigma}+\lambda\sigma} \exp\left[-\frac{1}{2}y^2\right]\cdot\exp\left[\frac{\lambda^2\sigma^2}{2}+\lambda\mu\right]\mathrm{d}y\\
&= \frac{\lambda\exp[-\lambda t]}{\sqrt{2\pi}}\cdot\exp\left[\frac{\lambda^2\sigma^2}{2}+\lambda\mu\right] \int\limits_{-\infty}^{\frac{x-\mu}{\sigma}+\lambda\sigma} \exp\left[-\frac{1}{2}y^2\right]\cdot\mathrm{d}y\\
&= \frac{\lambda}{\sqrt{2\pi}}\cdot\exp\left[-\lambda t+\frac{\lambda^2\sigma^2}{2}+\lambda\mu\right] \int\limits_{-\infty}^{\frac{x-\mu}{\sigma}+\lambda\sigma} \exp\left[-\frac{1}{2}y^2\right]\mathrm{d}y\\
&= \frac{\lambda}{\sqrt{2\pi}}\cdot\exp\left[\frac{\lambda^2\sigma^2}{2}-\lambda(t-\mu)\right] \int\limits_{-\infty}^{\frac{x-\mu}{\sigma}+\lambda\sigma} \exp\left[-\frac{1}{2}y^2\right]\mathrm{d}y
\end{aligned} \tag{5.7.1158}$$

and this finishes the proof!

Let X still be a random variable following an ex-Gaussian distribution:

$$X \sim \text{ex-Gaussian}(\mu,\sigma,\lambda) \tag{5.7.1159}$$

Then, the mean or expected value of X is:

$$\mathrm{E}(X) = \mu + \frac{1}{\lambda} \tag{5.7.1160}$$

The mean or expected value $\mathrm{E}(X)$ is the first raw moment of X, so we can use the moment-generating function of the ex-Gaussian distribution to calculate (see page 933):

$$\mathrm{E}(X) = M_X'(0) \tag{5.7.1161}$$

First, we differentiate:

$$M_X(t) = \left(\frac{\lambda}{\lambda-t}\right)\exp\left[\mu t+\frac{1}{2}\sigma^2t^2\right] \tag{5.7.1162}$$

with respect to t. Using the product rule and chain rule gives:

$$\begin{aligned}M_X'(t) &= \frac{\lambda}{(\lambda-t)^2}\exp\left[\mu t+\frac{1}{2}\sigma^2t^2\right]+\left(\frac{\lambda}{\lambda-t}\right)\exp\left[\mu t+\frac{1}{2}\sigma^2t^2\right](\mu+\sigma^2t)\\ &=\left(\frac{\lambda}{\lambda-t}\right)\cdot\exp\left[\mu t+\frac{1}{2}\sigma^2t^2\right]\cdot\left[\frac{1}{\lambda-t}+\mu+\sigma^2t\right]\end{aligned}\tag{5.7.1163}$$

Evaluating the latter at $t=0$ gives the desired result:

$$\begin{aligned}M_X'(0) &= \left(\frac{\lambda}{\lambda-0}\right)\cdot\exp\left[\mu\cdot 0+\frac{1}{2}\sigma^2\cdot 0^2\right]\cdot\left[\frac{1}{\lambda-0}+\mu+\sigma^2\cdot 0\right]\\ &=1\cdot 1\cdot\left[\frac{1}{\lambda}+\mu\right]\\ &=\mu+\frac{1}{\lambda}\end{aligned}\tag{5.7.1164}$$

Let X be a random variable following an ex-Gaussian distribution:

$$X\sim\text{ex-Gaussian}(\mu,\sigma,\lambda)\tag{5.7.1165}$$

Then the variance of X is:

$$\mathrm{V}(X)=\sigma^2+\frac{1}{\lambda^2}\tag{5.7.1166}$$

To compute the variance of X, we partition the variance into expected values using Huygens relation (see page 593):

$$\mathrm{V}(X)=\mathrm{E}\left(X^2\right)-\mathrm{E}(X)^2\tag{5.7.1167}$$

We then use the moment-generating function of the ex-Gaussian distribution to calculate:

$$M_X(t)=\left(\frac{\lambda}{\lambda-t}\right)\exp\left[\mu t+\frac{1}{2}\sigma^2t^2\right]\tag{5.7.1168}$$

with respect to t. Using the product rule and chain rule gives:

$$\begin{aligned}M_X'(t) &= \frac{\lambda}{(\lambda-t)^2}\exp\left[\mu t+\frac{1}{2}\sigma^2t^2\right]+\left(\frac{\lambda}{\lambda-t}\right)\exp\left[\mu t+\frac{1}{2}\sigma^2t^2\right](\mu+\sigma^2t)\\ &=\left(\frac{\lambda}{\lambda-t}\right)\cdot\exp\left[\mu t+\frac{1}{2}\sigma^2t^2\right]\cdot\left[\frac{1}{\lambda-t}+\mu+\sigma^2t\right]\\ &=M_X(t)\cdot\left[\frac{1}{\lambda-t}+\mu+\sigma^2t\right]\end{aligned}\tag{5.7.1169}$$

We now use the product rule to obtain the second derivative:

$$\begin{aligned}M_X''(t) &= M_X'(t)\cdot\left[\frac{1}{\lambda-t}+\mu+\sigma^2t\right]+M_X(t)\cdot\left[\frac{1}{(\lambda-t)^2}+\sigma^2\right]\\ &=M_X(t)\cdot\left[\frac{1}{\lambda-t}+\mu+\sigma^2t\right]^2+M_X(t)\cdot\left[\frac{1}{(\lambda-t)^2}+\sigma^2\right]\\ &=M_X(t)\cdot\left[\left(\frac{1}{\lambda-t}+\mu+\sigma^2t\right)^2+\frac{1}{(\lambda-t)^2}+\sigma^2\right]\\ &=\left(\frac{\lambda}{\lambda-t}\right)\cdot\exp\left[\mu t+\frac{1}{2}\sigma^2t^2\right]\cdot\left[\left(\frac{1}{\lambda-t}+\mu+\sigma^2t\right)^2+\frac{1}{(\lambda-t)^2}+\sigma^2\right]\end{aligned}\tag{5.7.1170}$$

Applying:

$$\mathrm{E}\left(X^2\right)=M''_X(0) \qquad (5.7.1171)$$

gives:

$$\begin{aligned}\mathrm{E}\left(X^2\right)&=M''_X(0)\\&=\left(\dfrac{\lambda}{\lambda-0}\right)\cdot\exp\left[\mu\cdot 0+\dfrac{1}{2}\sigma^2\cdot 0^2\right]\cdot\left[\left(\dfrac{1}{\lambda-0}+\mu+\sigma^2\cdot 0\right)^2+\dfrac{1}{(\lambda-0)^2}+\sigma^2\right]\\&=1\cdot 1\cdot\left[\left(\dfrac{1}{\lambda}+\mu\right)^2+\dfrac{1}{\lambda^2}+\sigma^2\right]\\&=\dfrac{1}{\lambda^2}+\dfrac{2\mu}{\lambda}+\mu^2+\dfrac{1}{\lambda^2}+\sigma^2=\dfrac{2}{\lambda^2}+\dfrac{2\mu}{\lambda}+\mu^2+\sigma^2\end{aligned} \qquad (5.7.1172)$$

Since the mean of an ex-Gaussian distribution is given by:

$$\mathrm{E}(X)=\mu+\frac{1}{\lambda} \qquad (5.7.1173)$$

we can apply the Huygens relation to show:

$$\begin{aligned}\mathrm{V}(X)&=\mathrm{E}\left(X^2\right)-\mathrm{E}(X)^2\\&=\left[\dfrac{2}{\lambda^2}+\dfrac{2\mu}{\lambda}+\mu^2+\sigma^2\right]-\left(\mu+\dfrac{1}{\lambda}\right)^2\\&=\dfrac{2}{\lambda^2}+\dfrac{2\mu}{\lambda}+\mu^2+\sigma^2-\mu^2-\dfrac{2\mu}{\lambda}-\dfrac{1}{\lambda^2}\\&=\sigma^2+\dfrac{1}{\lambda^2}\end{aligned} \qquad (5.7.1174)$$

and this completes the proof.

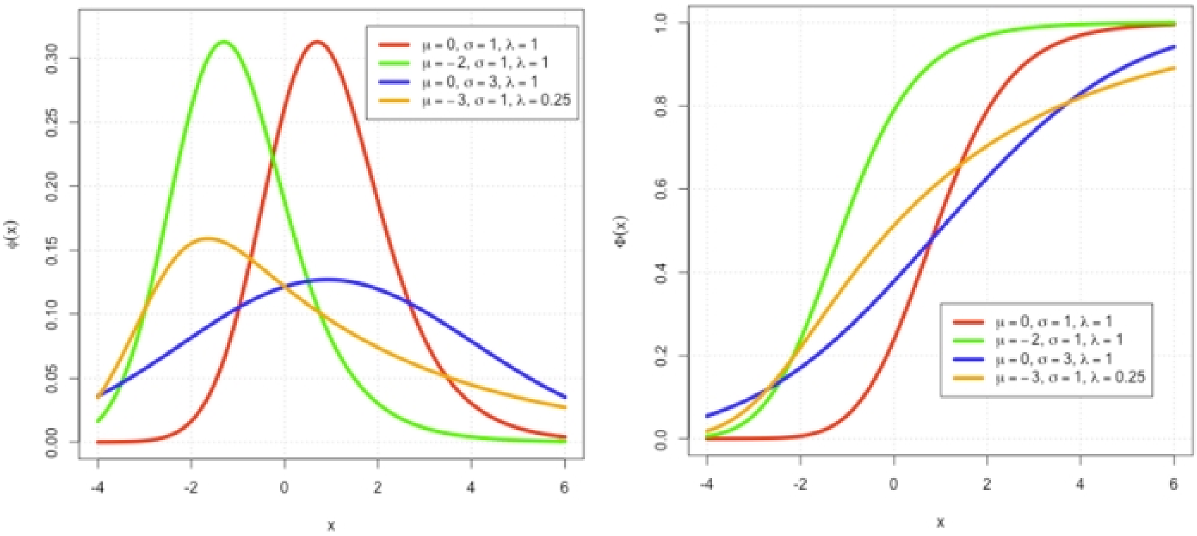


Figure 5.153: ex-Gaussian law (mass and cumulative distribution function)

Let X still be a random variable following an ex-Gaussian distribution:

$$X \sim \text{ ex-Gaussian } (\mu, \sigma, \lambda) \tag{5.7.1175}$$

Then the skewness of X is:

$$\text{Skew}(X) = \gamma_1 = \frac{2}{\lambda^3 \left(\sigma^2 + \frac{1}{\lambda^2}\right)^{\frac{3}{2}}} \tag{5.7.1176}$$

Indeed, remember first that the Fisher-Pearson coefficient of skewness is defined by (see page **??**):

$$\gamma_1 = \frac{\text{E}\left(X^3\right) - 3\mu\sigma^2 - \mu^3}{\sigma^3} \tag{5.7.1177}$$

where μ and σ are the mean and standard deviation of X, respectively. To prevent confusion between the labels used for the ex-Gaussian parameters and the mean and standard deviation of X, we rewrite the above relation as:

$$\gamma_1 = \frac{\text{E}\left(X^3\right) - 3 \cdot \text{E}(X) \cdot \text{V}(X) - \text{E}(X)^3}{\text{V}(X)^{\frac{3}{2}}} \tag{5.7.1178}$$

Since X follows an ex-Gaussian distribution, the mean of X is given by:

$$\text{E}(X) = \mu + \frac{1}{\lambda} \tag{5.7.1179}$$

and the variance by:

$$\text{V}(X) = \sigma^2 + \frac{1}{\lambda^2} \tag{5.7.1180}$$

Thus, the primary work is to compute $\text{E}\left(X^3\right)$. To do this, we use again the moment-generating function of the ex-Gaussian distribution to calculate:

$$\text{E}\left(X^3\right) = M_X'''(0) \tag{5.7.1181}$$

based on the relationship between raw moment and moment-generating function. First, we differentiate the moment-generating function of the ex-Gaussian distribution:

$$\begin{aligned} M_X'(t) &= \frac{\lambda}{(\lambda - t)^2} \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right] + \left(\frac{\lambda}{\lambda - t}\right) \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right] \left(\mu + \sigma^2 t\right) \\ &= \left(\frac{\lambda}{\lambda - t}\right) \cdot \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right] \cdot \left[\frac{1}{\lambda - t} + \mu + \sigma^2 t\right] \\ &= M_X(t) \cdot \left[\frac{1}{\lambda - t} + \mu + \sigma^2 t\right] \end{aligned} \tag{5.7.1182}$$

We then use the product rule to obtain the second derivative:

$$\begin{aligned} M_X''(t) &= M_X'(t) \cdot \left[\frac{1}{\lambda - t} + \mu + \sigma^2 t\right] + M_X(t) \cdot \left[\frac{1}{(\lambda - t)^2} + \sigma^2\right] \\ &= M_X(t) \cdot \left[\frac{1}{\lambda - t} + \mu + \sigma^2 t\right]^2 + M_X(t) \cdot \left[\frac{1}{(\lambda - t)^2} + \sigma^2\right] \\ &= M_X(t) \cdot \left[\left(\frac{1}{\lambda - t} + \mu + \sigma^2 t\right)^2 + \frac{1}{(\lambda - t)^2} + \sigma^2\right] \end{aligned} \tag{5.7.1183}$$

Finally, we use the product rule and chain rule to obtain the third derivative:

$$\begin{aligned}M_X'''(t)&=M_X'(t)\left[\left(\frac{1}{\lambda-t}+\mu+\sigma^2t\right)^2+\frac{1}{(\lambda-t)^2}+\sigma^2\right]\\&\quad+M_X(t)\left[2\left(\frac{1}{\lambda-t}+\mu+\sigma^2t\right)\left(\frac{1}{(\lambda-t)^2}+\sigma^2\right)+\frac{2}{(\lambda-t)^3}\right]\end{aligned} \tag{5.7.1184}$$

Applying:

$$\mathrm{E}\left(X^3\right)=M_X'''(0) \tag{5.7.1185}$$

Into the previous relation, we get:

$$\begin{aligned}\mathrm{E}\left(X^3\right)&=M_X'''(0)\\&=M_X'(0)\left[\left(\frac{1}{\lambda}+\mu\right)^2+\frac{1}{\lambda^2}+\sigma^2\right]+M_X(0)\left[2\left(\frac{1}{\lambda}+\mu\right)\left(\frac{1}{\lambda^2}+\sigma^2\right)+\frac{2}{\lambda^3}\right]\\&=\left(\mu+\frac{1}{\lambda}\right)\left(\frac{1}{\lambda^2}+\frac{2\mu}{\lambda}+\mu^2+\frac{1}{\lambda^2}+\sigma^2\right)+\left(\frac{2}{\lambda^3}+\frac{2\sigma^2}{\lambda}+\frac{2\mu}{\lambda^2}+2\mu\sigma^2+\frac{2}{\lambda^3}\right)\\&=\left(\mu+\frac{1}{\lambda}\right)\left(\frac{2}{\lambda^2}+\frac{2\mu}{\lambda}+\mu^2+\sigma^2\right)+\left(\frac{4}{\lambda^3}+\frac{2\sigma^2}{\lambda}+\frac{2\mu}{\lambda^2}+2\mu\sigma^2\right)\\&=\frac{2\mu}{\lambda^2}+\frac{2\mu^2}{\lambda}+\mu^3+\mu\sigma^2+\frac{2}{\lambda^3}+\frac{2\mu}{\lambda^2}+\frac{\mu^2}{\lambda}+\frac{\sigma^2}{\lambda}+\frac{4}{\lambda^3}+\frac{2\sigma^2}{\lambda}+\frac{2\mu}{\lambda^2}+2\mu\sigma^2\\&=\frac{6\mu}{\lambda^2}+\frac{6}{\lambda^3}+\frac{3\mu^2+3\sigma^2}{\lambda}+3\mu\sigma^2+\mu^3\end{aligned} \tag{5.7.1186}$$

We now substitute the previous results:

$$\begin{aligned}\mathrm{E}(X)&=\mu+\frac{1}{\lambda}\\\mathrm{V}(X)&=\sigma^2+\frac{1}{\lambda^2}\\\mathrm{E}\left(X^3\right)&=M_X'''(0)=\frac{6\mu}{\lambda^2}+\frac{6}{\lambda^3}+\frac{3\mu^2+3\sigma^2}{\lambda}+3\mu\sigma^2+\mu^3\end{aligned} \tag{5.7.1187}$$

Into the numerator of:

$$\text{Skew}(X)=\gamma_1=\frac{\mathrm{E}\left(X^3\right)-3\cdot\mathrm{E}(X)\cdot\mathrm{V}(X)-\mathrm{E}(X)^3}{\mathrm{V}(X)^{\frac{3}{2}}} \tag{5.7.1188}$$

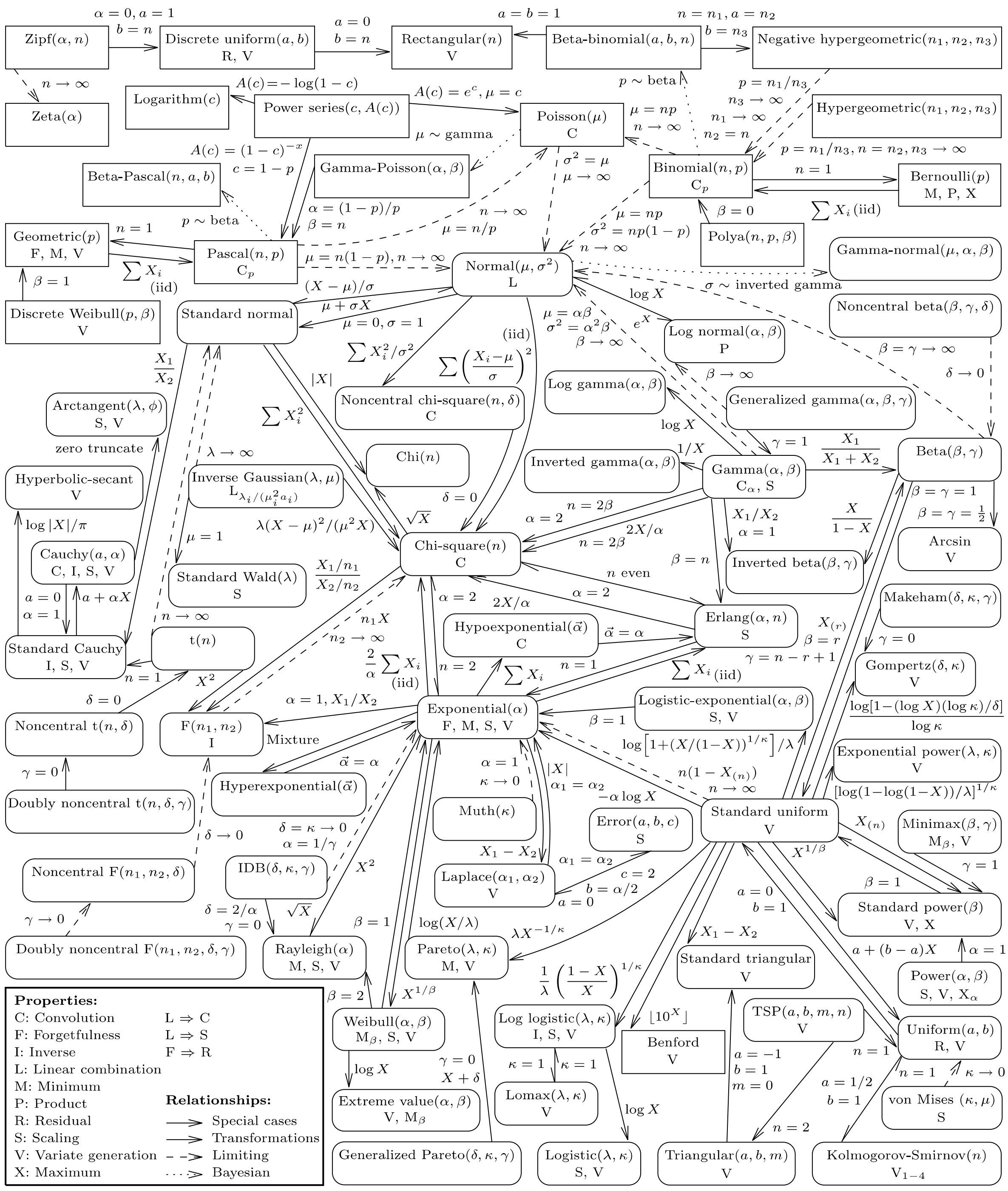
Gives:

$$\begin{aligned}\mathrm{E}\left(X^3\right)-3\cdot\mathrm{E}(X)\cdot\mathrm{V}(X)-\mathrm{E}(X)^3&=\left(\frac{6\mu}{\lambda^2}+\frac{6}{\lambda^3}+\frac{3\mu^2+3\sigma^2}{\lambda}+3\mu\sigma^2+\mu^3\right)-3\left(\mu+\frac{1}{\lambda}\right)\left(\sigma^2+\frac{1}{\lambda^2}\right)-\left(\mu+\frac{1}{\lambda}\right)^3\\&=\frac{6\mu}{\lambda^2}+\frac{6}{\lambda^3}+\frac{3\mu^2+3\sigma^2}{\lambda}+3\mu\sigma^2+\mu^3-3\mu\sigma^2-\frac{3\mu}{\lambda^2}-\frac{3\sigma^2}{\lambda}-\frac{3}{\lambda^3}-\mu^3-\frac{3\mu^2}{\lambda}-\frac{3\mu}{\lambda^2}-\frac{1}{\lambda^3}\\&=\frac{2}{\lambda^3}\end{aligned} \tag{5.7.1189}$$

Thus, we have:

$$\begin{aligned}
\text{Skew}(X)=\gamma_1&=\frac{\mathrm{E}\left(X^3\right)-3\cdot\mathrm{E}(X)\cdot\mathrm{V}(X)-\mathrm{E}(X)^3}{\mathrm{V}(X)^{\frac{3}{2}}}\\
&=\frac{\frac{2}{\lambda^3}}{\left(\sigma^2+\frac{1}{\lambda^2}\right)^{\frac{3}{2}}}\\
&=\frac{2}{\lambda^3\left(\sigma^2+\frac{1}{\lambda^2}\right)^{\frac{3}{2}}}.
\end{aligned} \tag{5.7.1190}$$

To summarize an important part of everything we've seen so far, the picture below is very useful because it summarizes the relation between 76 of the most common univariate distributions (57 continuous and 19 discrete):



Zipf(α, n)
Discrete uniform(a, b) R, V
Rectangular(n) V
Beta-binomial(a, b, n)
Negative hypergeometric(n₁, n₂, n₃)
Zeta(α)
Logarithm(c)
Power series(c, A(c))
Poisson(μ) C
Hypergeometric(n₁, n₂, n₃)
Beta-Pascal(n, a, b)
Gamma-Poisson(α, β)
Binomial(n, p) Cₚ
Bernoulli(p) M, P, X
Geometric(p) F, M, V
Pascal(n, p) Cₚ
Normal(μ, σ²) L
Polya(n, p, β)
Gamma-normal(μ, α, β)
Discrete Weibull(p, β) V
Standard normal
Noncentral beta(β, γ, δ)
Log normal(α, β) P
Noncentral chi-square(n, δ) C
Log gamma(α, β)
Generalized gamma(α, β, γ)
Arctangent(λ, φ) S, V
Chi(n)
Inverted gamma(α, β)
Gamma(α, β) Cα, S
Beta(β, γ)
Hyperbolic-secant V
Inverse Gaussian(λ, μ)
Chi-square(n) C
Arcsin V
Cauchy(a, α) C, I, S, V
Standard Wald(λ) S
Inverted beta(β, γ)
Makeham(δ, κ, γ)
Standard Cauchy I, S, V
t(n)
Hypoexponential(α) C
Erlang(α, n) S
Gompertz(δ, κ) V
Noncentral t(n, δ)
F(n₁, n₂) I
Exponential(α) F, M, S, V
Logistic-exponential(α, β) S, V
Exponential power(λ, κ) V
Doubly noncentral t(n, δ, γ)
Hyperexponential(α)
Muth(κ)
Error(a, b, c) S
Standard uniform V
Minimax(β, γ) M_β, V
Noncentral F(n₁, n₂, δ)
IDB(δ, κ, γ)
Laplace(α₁, α₂) V
Standard power(β) V, X
Doubly noncentral F(n₁, n₂, δ, γ)
Rayleigh(α) M, S, V
Pareto(λ, κ) M, V
Standard triangular V
Power(α, β) S, V, X_α
Weibull(α, β) M_β, S, V
Log logistic(λ, κ) I, S, V
Benford V
TSP(a, b, m, n) V
Uniform(a, b) R, V
Extreme value(α, β) V, M_β
Lomax(λ, κ) V
von Mises (κ, μ) S
Generalized Pareto(δ, κ, γ)
Logistic(λ, κ) S, V
Triangular(a, b, m) V
Kolmogorov-Smirnov(n) V₁₋₄
Properties:
C: Convolution
F: Forgetfulness
I: Inverse
L: Linear combination
M: Minimum
P: Product
R: Residual
S: Scaling
V: Variate generation
X: Maximum
L ⇒ C
L ⇒ S
F ⇒ R
Relationships:
Special cases
Transformations
Limiting
Bayesian

Figure 5.154: Relations between distributions (authors: AMS Lawrence M. Leemis and Jacquelyn T. McQueston)

And we shouldn't forget the traditional summary tables that every good textbook provides (they should obviously add the median, the skewness and kurtosis but there is sadhl most of time no room for it):

	Notation	$F_X(x)$	$f_X(x)$	$\mathrm{E}(X)$	$\mathrm{V}(X)$	$M_X(s)$
Uniform	Unif $\{a,\ldots,b\}$	$\begin{cases}0 & x<a\\ \frac{\lfloor x\rfloor-a+1}{b-a} & a\leq x\leq b\\ 1 & x>b\end{cases}$	$\dfrac{I(a\leq x\leq b)}{b-a+1}$	$\dfrac{a+b}{2}$	$\dfrac{(b-a+1)^2-1}{12}$	$\dfrac{e^{as}-e^{-(b+1)s}}{s(b-a)}$
Bernoulli	$\mathrm{Bern}(p)$	$(1-p)^{1-x}$	$p^x(1-p)^{1-x}$	p	$p(1-p)$	$1-p+pe^s$
Binomial	$\mathrm{Bin}(n,p)$	$I_{1-p}(n-x,x+1)$	$\left(\begin{array}{c}n\\x\end{array}\right)p^x(1-p)^{n-x}$	np	$np(1-p)$	$(1-p+pe^s)^n$
Multinomial	Mult (n,p)		$\dfrac{n!}{x_1!\ldots x_k!}p_1^{x_1}\ldots p_k^{x_k} \quad \sum_{i=1}^{k}x_i=n$	$\begin{pmatrix}np_1\\ \vdots\\ np_k\end{pmatrix}$	$\begin{pmatrix}np_1(1-p_1) & -np_1p_2\\ -np_2p_1 & \ddots\end{pmatrix}$	$\left(\sum_{i=0}^{k}p_ie^{s_i}\right)^n$
Hypergeometric	$\mathrm{Hyp}(N,m,n)$	$\approx\Phi\left(\dfrac{x-np}{\sqrt{np(1-p)}}\right)$	$\dfrac{\left(\begin{array}{c}m\\x\end{array}\right)\left(\begin{array}{c}N-m\\n-x\end{array}\right)}{\left(\begin{array}{c}N\\n\end{array}\right)}$	$\dfrac{nm}{N}$	$\dfrac{nm(N-n)(N-m)}{N^2(N-1)}$	
Negative Binomial	$\mathrm{NBin}(r,p)$	$I_p(r,x+1)$	$\left(\begin{array}{c}x+r-1\\r-1\end{array}\right)p^r(1-p)^x$	$r\dfrac{1-p}{p}$	$r\dfrac{1-p}{p^2}$	$\left(\dfrac{pe^s}{1-(1-p)e^s}\right)^r$
Geometric	$\mathrm{Geo}(p)$	$1-(1-p)^x \quad x\in\mathbb{N}^+$	$p(1-p)^{x-1} \quad x\in\mathbb{N}^+$	$\dfrac{1}{p}$	$\dfrac{1-p}{p^2}$	$\dfrac{pe^s}{1-(1-p)e^s}$
Poisson	$\mathrm{Po}(\lambda)$	$e^{-\lambda}\sum_{i=0}^{x}\dfrac{\lambda^i}{i!}$	$\dfrac{\lambda^xe^{-\lambda}}{x!}$	λ	λ	$e^{\lambda(e^s-1)}$

Table 5.45: **Summary of most common discrete distributions**

	Notation	$F_X(x)$	$f_X(x)$	$\mathrm{E}(X)$	$\mathrm{V}(X)$	$M_X(s)$
Uniform	$\mathrm{Unif}(a,b)$	$\begin{cases}0 & x<a\\ \frac{x-a}{b-a} & a<x<b\\ 1 & x>b\end{cases}$	$\dfrac{I(a<x<b)}{b-a}$	$\dfrac{a+b}{2}$	$\dfrac{(b-a)^2}{12}$	$\dfrac{e^{sb}-e^{sa}}{s(b-a)}$
Normal	$\mathcal{N}\left(\mu,\sigma^2\right)$	$\Phi(x)=\int_{-\infty}^{x}\phi(t)\,dt$	$\phi(x)=\dfrac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\dfrac{(x-\mu)^2}{2\sigma^2}\right\}$	μ	σ^2	$e^{\mu s+\frac{\sigma^2s^2}{2}}$
Log-Normal	$\ln\mathcal{N}\left(\mu,\sigma^2\right)$	$\dfrac{1}{2}+\dfrac{1}{2}\mathrm{erf}\left[\dfrac{\ln x-\mu}{\sqrt{2\sigma^2}}\right]$	$\dfrac{1}{x\sqrt{2\pi\sigma^2}}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$	$e^{\mu+\sigma^2/2}$	$(e^{\sigma^2}-1)e^{2\mu+\sigma^2}$	
Multivariate Normal	$\mathrm{MVN}(\mu,\Sigma)$		$(2\pi)^{-k/2}\lvert\Sigma\rvert^{-1/2}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$	μ	Σ	$e^{\mu^Ts+\frac{1}{2}s^T\Sigma s}$
Student's t	Student(ν)	$I_x\left(\dfrac{\nu}{2},\dfrac{\nu}{2}\right)$	$\dfrac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\dfrac{x^2}{\nu}\right)^{-(\nu+1)/2}$	$0 \quad \nu>1$	$\begin{cases}\dfrac{\nu}{\nu-2} & \nu>2\\ \infty & 1<\nu\leq 2\end{cases}$	
Chi-square	χ_k^2	$\dfrac{1}{\Gamma(k/2)}\gamma\left(\dfrac{k}{2},\dfrac{x}{2}\right)$	$\dfrac{1}{2^{k/2}\Gamma(k/2)}x^{k/2-1}e^{-x/2}$	k	$2k$	$(1-2s)^{-k/2}\ s<1/2$
F	$\mathrm{F}(d_1,d_2)$	$I_{\frac{d_1x}{d_1x+d_2}}\left(\dfrac{d_1}{2},\dfrac{d_2}{2}\right)$	$\dfrac{\sqrt{\frac{(d_1x)^{d_1}d_2^{d_2}}{(d_1x+d_2)^{d_1+d_2}}}}{x\mathrm{B}\left(\frac{d_1}{2},\frac{d_1}{2}\right)}$	$\dfrac{d_2}{d_2-2}$	$\dfrac{2d_2^2(d_1+d_2-2)}{d_1(d_2-2)^2(d_2-4)}$	
Exponential	$\mathrm{Exp}(\beta)$	$1-e^{-x/\beta}$	$\dfrac{1}{\beta}e^{-x/\beta}$	β	β^2	$\dfrac{1}{1-\frac{s}{\beta}}\,(s<\beta)$
Gamma	$\mathrm{Gamma}(\alpha,\beta)$	$\dfrac{\gamma(\alpha,\beta x)}{\Gamma(\alpha)}$	$\dfrac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$	$\dfrac{\alpha}{\beta}$	$\dfrac{\alpha}{\beta^2}$	$\left(\dfrac{1}{1-\frac{s}{\beta}}\right)^\alpha(s<\beta)$
Inverse Gamma	$\mathrm{InvGamma}(\alpha,\beta)$	$\dfrac{\Gamma\left(\alpha,\frac{\beta}{x}\right)}{\Gamma(\alpha)}$	$\dfrac{\beta^\alpha}{\Gamma(\alpha)}x^{-\alpha-1}e^{-\beta/x}$	$\dfrac{\beta}{\alpha-1}\ \alpha>1$	$\dfrac{\beta^2}{(\alpha-1)^2(\alpha-2)}\ \alpha>2$	$\dfrac{2(-\beta s)^{\alpha/2}}{\Gamma(\alpha)}K_\alpha\left(\sqrt{-4\beta s}\right)$
Dirichlet	$\mathrm{Dir}(\alpha)$		$\dfrac{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}{\prod_{i=1}^{k}\Gamma\left(\alpha_i\right)}\prod_{i=1}^{k}x_i^{\alpha_i-1}$	$\dfrac{\alpha_i}{\sum_{i=1}^{k}\alpha_i}$	$\dfrac{\mathrm{E}X_i(1-\mathrm{E}X_i)}{\sum_{i=1}^{k}\alpha_i+1}$	
Beta	$\mathrm{Beta}(\alpha,\beta)$	$I_x(\alpha,\beta)$	$\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$	$\dfrac{\alpha}{\alpha+\beta}$	$\dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	$1+\sum_{k=1}^{\infty}\left(\prod_{r=0}^{k-1}\dfrac{\alpha+r}{\alpha+\beta+r}\right)\dfrac{s^k}{k!}$
Weibull	Weibull(λ,k)	$1-e^{-(x/\lambda)^k}$	$\dfrac{k}{\lambda}\left(\dfrac{x}{\lambda}\right)^{k-1}e^{-(x/\lambda)^k}$	$\lambda\Gamma\left(1+\dfrac{1}{k}\right)$	$\lambda^2\Gamma\left(1+\dfrac{2}{k}\right)-\mu^2$	$\sum_{n=0}^{\infty}\dfrac{s^n\lambda^n}{n!}\Gamma\left(1+\dfrac{n}{k}\right)$
Pareto	$\mathrm{Pareto}(x_m,\alpha)$	$1-\left(\dfrac{x_m}{x}\right)^\alpha\ x\geq x_m$	$\alpha\dfrac{x_m^\alpha}{x^{\alpha+1}}\quad x\geq x_m$	$\dfrac{\alpha x_m}{\alpha-1}\ \alpha>1$	$\dfrac{x_m^2\alpha}{(\alpha-1)^2(\alpha-2)}\ \alpha>2$	$\alpha(-x_ms)^\alpha\Gamma(-\alpha,-x_ms)\ s<0$
Laplace	$\mathrm{Laplace}(\mu,b)$	$\dfrac{1}{2}+\dfrac{1}{2}\mathrm{sgn}(x-\mu)\left(1-\exp\left(-\dfrac{\lvert x-\mu\rvert}{b}\right)\right)$	$\dfrac{1}{2b}\exp\left(-\dfrac{\lvert x-\mu\rvert}{b}\right)$	μ	$2b^2$	$\dfrac{\exp(\mu t)}{1-b^2t^2}\ \lvert t\rvert<1/b$
exGaussian	$\mathrm{ex}-\mathrm{Gaussian}(\mu,\sigma,\lambda)$	$\Phi(x,\mu,\sigma)-\dfrac{1}{2}e^{\frac{\lambda}{2}(2\mu+\lambda\sigma^2-2x)}\,\mathrm{erfc}\left(\dfrac{\mu+\lambda\sigma^2-x}{\sqrt{2}\sigma}\right)$	$\dfrac{\lambda}{2}e^{\frac{\lambda}{2}(2\mu+\lambda\sigma^2-2x)}\,\mathrm{erfc}\left(\dfrac{\mu+\lambda\sigma^2-x}{\sqrt{2}\sigma}\right)$	$\mu+1/\lambda$	σ^2+1/λ^2	$\left(1-\dfrac{t}{\lambda}\right)^{-1}\exp\left(\mu t+\dfrac{1}{2}\sigma^2t^2\right)$

Table 5.46: **Summary of most common continuous distributions**

And here a first high quality vectorial visual summary of some discrete and continuous probability density functions :

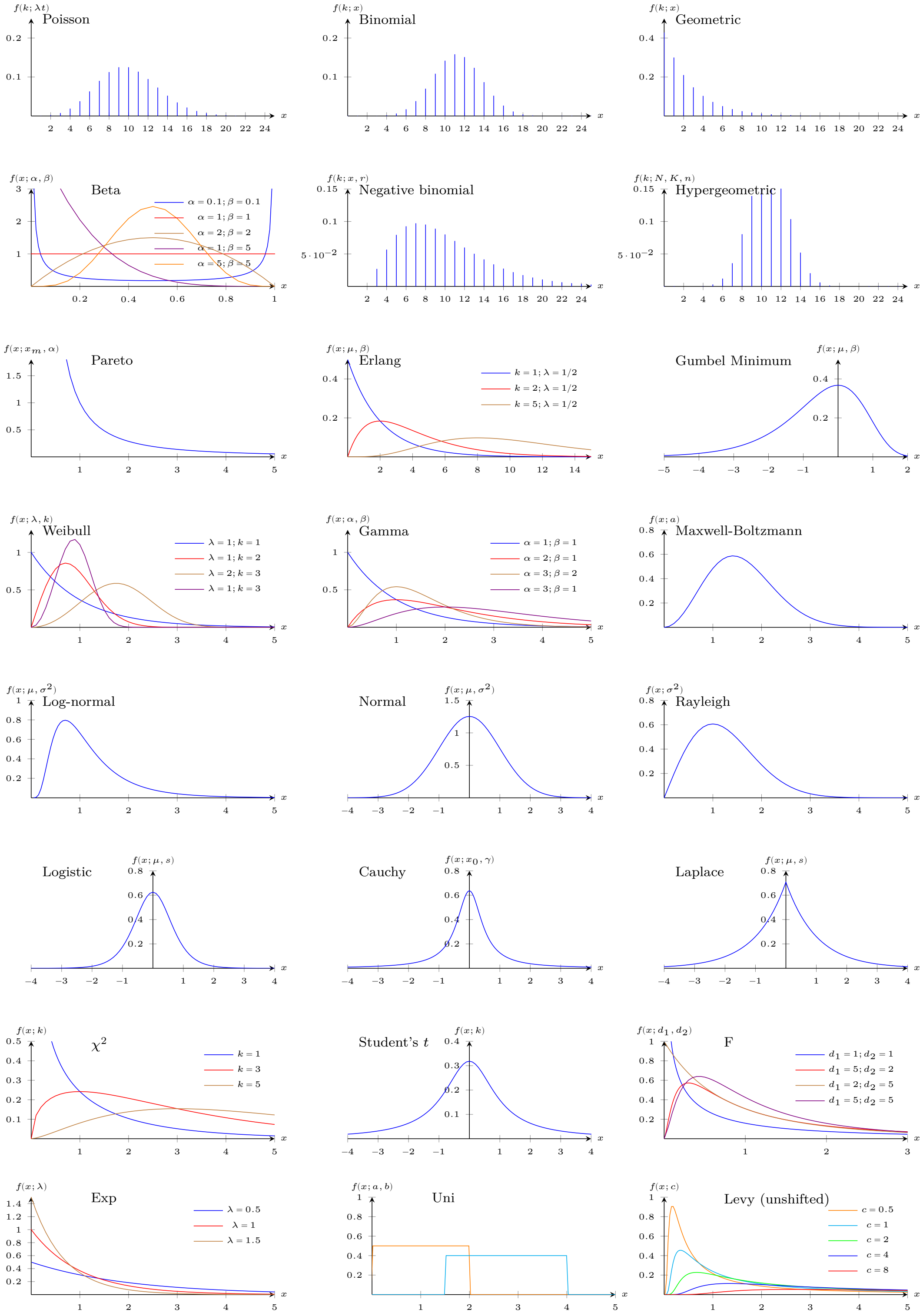


Figure 5.155: **Visual summary of some important continuous and discrete statistical distributions (authors: Adrian Tam, V. Isoz)**

and another visual summary of some important continuous distribution with their PDF and CDF:

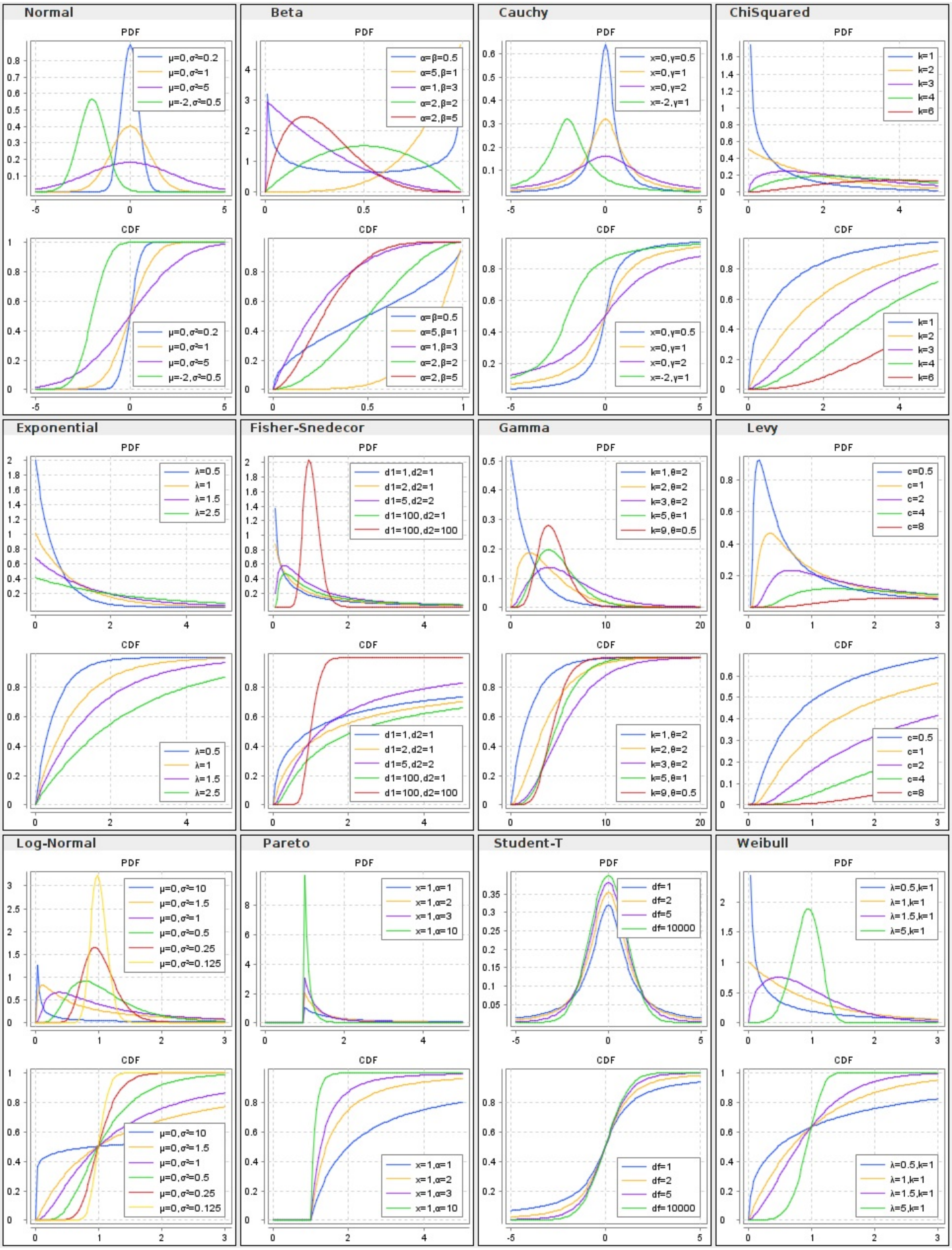


Figure 5.156: Visual summary of some important continuous statistical distributions (source: ?)

5.7.7 Likelihood Estimators

What follows is of extreme importance in the field of statistics and is used widely in practice. It is important therefore to pay attention! Besides the fact that we will use massively likelihood estimators in this chapter, we shall find it in the chapter of Numerical Methods for advanced and generalized linear regression (including their generalization that are M-estimators or even W-estimators!) and also in the chapter of Industrial Engineering in the context of parametric estimation of reliability.

We assume that we have observations $x_1, x_2, x_3, \ldots, x_n$ which are realizations of unbiased independent random variables (in the sense that they are randomly selected from a batch) $X_1, X_2, X_3, \ldots, X_n$ of a unknown probability distribution but having the same one.

Suppose we proceed by trial and error to estimate the unknown probability distribution P. One way to proceed is to ask if the observations $x_1, x_2, x_3, \ldots, x_n$ had a high probability to get out or not with this arbitrary probability distribution P.

We need for this to calculate the joint probability that the observations $x_1, x_2, x_3, \ldots, x_n$ had to get out with the probabilities $p_1, p_2, p_3, \ldots, p_n$. This joint probability is equal to (see section Probabilities page 480):

$$\prod_{i=1}^{n} P(X_i = x_i) \tag{5.7.1191}$$

noting by the letter P the assumed probability distribution associated to $p_1, p_2, p_3, \ldots, p_n$. You must admit that it would be particularly awkward, at the intuition level of risk, to choose a probability distribution (with its parameters!) that minimizes this quantity...

Instead, we will seek the probabilities $p_1, p_2, p_3, \ldots, p_n$ (or the associated parameters of the probability distribution) that maximizes $\prod_{i=1}^{n} P(X_i = x_i)$, that is to say, that makes the observations $x_1, x_2, x_3, \ldots, x_n$ the most likely possible.

This leads us to seek the parameter(s) θ that maximizes the quantity:

$$L_n(\theta) = \prod_{i=1}^{n} P_\theta(X_i = x_i) \tag{5.7.1192}$$

and where the parameter θ is often in undergraduate school level problems a first order moment (mean) or second order moment (variance).

The quantity L is named "likelihood". It is a function of the parameter(s) θ and observations θ.

The value(s) of the parameter(s) θ that maximize the likelihood $L_n(\theta)$ are named "maximum likelihood estimators" (MLE estimators).

In the very special case but useful of the Normal distribution, one of the parameters θ will be the variance (see a little further concrete example) and can be considered intuitive to the physicist that to maximize the probability, the standard deviation should be as small as possible (so that the maximum numbers of events are in the same interval). Thus, when we calculate a MLE which is the smallest among several possible, then we are talking about a UMV estimator for "Uniform Minimum Variance Unbiased" because their own variance should be as small as possible. This can be demonstrated (but the proof is not very elegant) using the definition of the Fisher Information and the Fréchet theorem (or Rao-Cramér) that makes use of the Cauchy-Schwarz inequality (see section Vector Calculus page 1826) and the analogy between mean and scalar product ... This demonstration will not be in this textbook.

Let us still do seven small examples (very classic, useful and important in the industry) with in order of importance (i.e. not necessarily in order of ease...) the distribution function of Gauss-Laplace (Normal distribution), the Poisson distribution, the binomial distribution, the Geometric distribution, the Weibull distribution, the Gamma distribution and finally the Pareto distribution.

Remarks

R1. These seven examples are important as used in SPC (statistical process control) in various international companies around the world (see section Industrial Engineering page 8673).

R2. A frequentist maximum likelihood estimate is like to the "mode" of the likelihood function, so in a Bayesian context with uniform prior, the MLE is like a mode of the posterior.

R3. In Bayesian statistics (see further below) we have the "maximum a posteriori" (MAP) which finds a maximum of the posterior distribution for a uniform prior probability. The MAP estimator coincides with the maximum-likelihood estimator.

5.7.7.1 Normal Distribution MLE

Let be $x_1, x_2, \ldots, x_n$ an n-sample of identically distributed random variables assumed to follow a Gaussian-Laplace (Normal) distribution of parameters μ and σ^2.

We are looking what are the values of the maximum likelihood estimators $\vec{\theta}$ that maximize the likelihood $L_n(\theta)$ of the Normal distribution?

Remark

It is trivial that the maximum likelihood estimators vector is here:

$$\vec{\theta} = (\mu, \sigma) \tag{5.7.1193}$$

We have proved earlier above that the density of a Gaussian random variable was given by:

$$P(x, \mu, \mu) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{5.7.1194}$$

The likelihood is then given by:

$$L(\mu, \sigma) = \prod_{i=1}^{n} P(x_i, \mu, \mu) = \frac{1}{\sigma^n \sqrt{2\pi}^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \tag{5.7.1195}$$

Maximize a function or maximize its logarithm is equivalent, therefore the "log-likelihood" will be:

$$\ln(L(\mu,\sigma)) = -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \tag{5.7.1196}$$

To determine the two estimators of the Normal distribution, first let us fix the standard deviation. To do this, we derive $\ln(L(\mu,\sigma))$ over μ and look for what the average value of the function is equal to zero.

Remark

The partial derivative that we seek to cancel is often named the "score function" (we will often find this concept during our study of optimization techniques and machine learning in the chapter of Theoretical Computing):

$$\frac{\partial}{\partial\theta}\ln\left[L(\theta)\right] \tag{5.7.1197}$$

It remains after simplification the following term that is equal to zero:

$$\sum_{i=1}^{n}(x_i-\mu)^2 \tag{5.7.1198}$$

Thus, the maximum likelihood estimator of the expected mean of the Normal distribution is after rearrangement:

$$\boxed{\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}x_i} \tag{5.7.1199}$$

and we see that it is simply the arithmetic mean (or also named "sample mean").

Let us now fix the mean. The cancellation of the derivative of $\ln(L(\mu,\sigma))$ over σ leads us to:

$$\frac{\partial}{\partial\sigma}\ln(L(\mu,\sigma)) = -\frac{n}{\sigma^2} - \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 = 0 \tag{5.7.1200}$$

This allows us to write the maximum likelihood estimator for the standard deviation (the variance when the mean is known under the an assumed distribution also supposed known!):

$$\boxed{\hat{\sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2} \tag{5.7.1201}$$

that some people also name "Pearson standard deviation"...

Even if it is a little bit redundant some people asked us to show the proof of the estimator of the covariance matrix (and therefore the correlation matrix).

Remember that we have proved earlier that for the bivariate case we have:

$$f(X_1,X_2) = \frac{1}{2\pi|\Sigma|^{1/2}}e^{-\frac{1}{2}\overrightarrow{(x-\mu)}^T\Sigma^{-1}\overrightarrow{(x-\mu)}} \tag{5.7.1202}$$

In fact the relation is the same for the multivariate case with T!

The log-likelihood is therefore immediate by analogy with the univariate case:

$$\ln\left(L(\vec{\mu},\Sigma)\right)=-\frac{nT}{2}\ln(2\pi)-\frac{T}{2}\ln(\Sigma)-\frac{1}{2}\sum_{i=1}^{T}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) \tag{5.7.1203}$$

That we can also write (as Σ is diagonal and Σ^{-1} also):

$$\begin{aligned}\ln\left(L(\vec{\mu},\Sigma)\right)&=-\frac{nT}{2}\ln(2\pi)-\frac{T}{2}\ln(\Sigma)-\frac{1}{2}\sum_{i=1}^{T}\text{tr}\left((x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right)\\&=-\frac{nT}{2}\ln(2\pi)-\frac{T}{2}\ln(\Sigma)-\frac{1}{2}\sum_{i=1}^{T}\text{tr}\left(\Sigma^{-1}(x_i-\mu)^T(x_i-\mu)\right)\\&:=-\frac{nT}{2}\ln(2\pi)-\frac{T}{2}\ln(\Sigma)-\frac{1}{2}\text{tr}\left(\Sigma^{-1}S\right)\end{aligned} \tag{5.7.1204}$$

Where by definition and using the estimator of the mean:

$$S=\sum_{i=1}^{T}(x_i-\hat{\mu})^T(x_i-\hat{\mu}) \tag{5.7.1205}$$

Then we deduce that:

$$\frac{\partial}{\partial\Sigma^{-1}}\ln\left(L(\vec{\hat{\mu}},\Sigma)\right)=\frac{T}{2}\Sigma-\frac{1}{2}\Sigma=0 \tag{5.7.1206}$$

and we get finally:

$$\hat{\Sigma}=\frac{1}{T}S \tag{5.7.1207}$$

However, we have not yet defined what is a good estimator! What we mean here is:

- If the mean of an estimator is equal to itself, we say that this estimator is "unbiased" and that's obviously what we want!
- If the mean of an estimator is not equal to itself, then we say that this estimator is "biased" and is necessarily less good...

In the previous example, the average is unbiased (this is trivial as the average of the arithmetic mean is equal to itself). But what about the variance (verbatim the standard deviation)?

A simple little calculation by linearity of the mean (since the random variables are identically distributed) will give us the answer in the case where the theoretical average (mean) is approximated as in practice (industry) by the estimator of the mean (most common case).

So we have for the calculation of the mean of the "sample variance":

$$\begin{aligned}\mathrm{E}(\hat{\sigma}^2)&=\mathrm{E}(\mathrm{V}(X))=\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\right)=\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i^2-2x_i\hat{\mu}+\hat{\mu}^2)\right)\\&=\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}x_i^2-2\frac{1}{n}\sum_{i=1}^{n}x_i\hat{\mu}+\frac{1}{n}\sum_{i=1}^{n}\hat{\mu}^2)\right)=\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}x_i^2-2\hat{\mu}\frac{1}{n}\sum_{i=1}^{n}x_i+\frac{1}{n}n\hat{\mu}^2)\right)\\&=\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}x_i^2-2\hat{\mu}^2+\hat{\mu}^2)\right)=\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}x_i^2-2\hat{\mu})\right)=\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(x_i^2)-\mathrm{E}(\hat{\mu}^2)\end{aligned} \tag{5.7.1208}$$

However, as the variables are supposed to be identically distributed:

$$\mathrm{E}(\hat{\sigma}^2)=\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(x_i^2)-\mathrm{E}(\hat{\mu}^2)=\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(x^2)-\mathrm{E}(\hat{\mu}^2)=\mathrm{E}(x^2)-\mathrm{E}(\hat{\mu}^2) \tag{5.7.1209}$$

And as we have (Huygens theorem):

$$\begin{aligned}\mathrm{V}(X)&=\mathrm{E}(X^2)-\mathrm{E}(X^2)\\\mathrm{V}(\hat{\mu})&=\mathrm{E}(\hat{\mu}^2)-\mathrm{E}(\hat{\mu})^2=\mathrm{E}(\hat{\mu}^2)-\mathrm{E}(X)^2\end{aligned} \tag{5.7.1210}$$

wherein the second relation can be written only because we use the maximum likelihood estimator of the average (empirical average). Therefore combining the two above relations with the prior-previous one we get:

$$\mathrm{E}(\hat{\sigma}^2)=\mathrm{E}(x^2)-\mathrm{E}(\hat{\mu}^2)=\left(\mathrm{V}(X)+\mathrm{E}(X)^2\right)-\left(\mathrm{V}(\hat{\mu})+\mathrm{E}(X)^2\right)=\mathrm{V}(X)-\mathrm{V}(\hat{\mu}) \tag{5.7.1211}$$

and as:

$$\mathrm{V}(X)=\sigma^2\quad\text{and}\quad\mathrm{V}(\hat{\mu})=\frac{\sigma^2}{n} \tag{5.7.1212}$$

Finally we have:

$$\boxed{\mathrm{E}(\hat{\sigma}^2)=\sigma^2-\frac{\sigma^2}{n}=\left(1-\frac{1}{n}\right)\sigma^2=\frac{n-1}{n}\sigma^2} \tag{5.7.1213}$$

so we have a bias of at least one standard error:

$$-\frac{\sigma^2}{n} \tag{5.7.1214}$$

then we say that this estimator has a negative bias (it underestimates the true value!).

We also note that the estimator tends towards to an unbiased estimator of the variance (USV) when the number of items tends to infinity $n\rightarrow+\infty$. We say that we have a "asymptotically unbiased" or "asymptotically unbiased estimator".

Then, the "bias of an estimator" is defined, according to the special case above, in general as:

$$\boxed{\text{bias}(\hat{\theta})=\mathrm{E}(\hat{\theta})-\theta} \tag{5.7.1215}$$

It is important to note that we have yet proved that the empirical variance tends towards the theoretical variance when n tends to infinity and ... that the data follows or not a Normal distribution!

By the properties of the mean, we get now:

$$\mathrm{E}(\mathrm{E}(\hat{\sigma}^2)) = \mathrm{E}(\hat{\sigma}^2) = \mathrm{E}\left(\frac{n-1}{n}\sigma^2\right) \tag{5.7.1216}$$

We have then:

$$\sigma = \sqrt{\frac{n}{n-1}\hat{\sigma}^2} = \sqrt{\frac{n}{n-1}\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\hat{\mu})^2} \tag{5.7.1217}$$

simply named the "standard deviation" ... (that must not be confused with the "standard error" as we shall see later).

An estimator is named "consistent estimator" if it converges in probability (as it can be considered itself as random variable following it own distribution), when $n \rightarrow +\infty$, towards the true parameter value. Technically, this condition will be written:

$$\boxed{\forall \varepsilon > 0 : \lim_{n\rightarrow+\infty} P\left(\left|\hat{\theta}-\theta\right| > \varepsilon\right) = 0} \tag{5.7.1218}$$

Maybe, the following example may help. Let's suppose the samples are taking from a Normal distribution. Then using the fact that $\frac{(n-1)S^2}{\sigma^2}$ (see page 860) is a chi-squared random variable with $n-1$ degrees of freedom, we get:

$$\begin{aligned} \mathrm{V}\left(\frac{(n-1)S^2}{\sigma^2}\right) &= \mathrm{Var}\ \chi^2_{n-1} \\ \frac{(n-1)^2}{\sigma^4}\mathrm{V}(S^2) &= 2(n-1) \\ \mathrm{V}(S^2) &= \frac{2(n-1)\sigma^4}{(n-1)^2} = \frac{2\sigma^4}{(n-1)}, \end{aligned} \tag{5.7.1219}$$

where we have used that fact that (see page 773):

$$\mathrm{V}\left(\chi^2_{n-1}\right) = 2(n-1) \tag{5.7.1220}$$

So written differently (according to above notation):

$$\boxed{\mathrm{V}\left(\hat{\sigma}^2\right) = \frac{2\sigma^2}{n-1}} \tag{5.7.1221}$$

From these facts we can informally see that the distribution of $\mathrm{V}\left(\hat{\sigma}^2\right)$ is becoming more and more concentrated at σ^2 as the sample size increases since the mean $\mathrm{E}\left(\hat{\sigma}^2\right)$ is converging to σ^2 (i.e. $\hat{\sigma}$ is unbiased) and the variance $\mathrm{V}\left(\hat{\sigma}^2\right)$ is converging to 0 (i.e. $\hat{\sigma}$ is consistent)!

So we finally summarize as following the two important previous results:

1. The "biased maximum likelihood estimator" or also named "empirical standard deviation" or "sample standard deviation" or "Pearson standard deviation" ... is therefore given by:

   $$\boxed{\hat{\sigma}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2}} \tag{5.7.1222}$$

   when $n\rightarrow+\infty$. We find this standard deviation depending on the context (by tradition) noted in five other ways that are:

   $$\sigma^*,S^*,\sigma_*,S_*,S_n \tag{5.7.1223}$$

   and sometimes (but this is very awkward because it often generates confusion with the unbiased estimator) σ or S.

2. The "unbiased maximum likelihood estimator" or simply named "standard deviation" with the "Bessel correction" (the -1 at the denominator is the correction[40]!):

   $$\boxed{\sigma=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\hat{\mu})^2}} \tag{5.7.1224}$$

   which as we can see is a consistent estimator (when n tends to infinity it tends to the biased maximum likelihood estimator).

   We find this standard deviation depending on the context (by tradition) noted in three other ways that are:

   $$\sigma,S,S_{n-1} \tag{5.7.1225}$$

We find these last two notations often in tables and in many softwares and we will use them later in the development of confidence intervals and hypothesis testing!

For example, in the Microsoft® Excel® 11.8346 the unbiased estimator is given by the `STDEV( )` function and the non-biased by `STDEVP()`.

In total, this makes us is three estimators for the same indicator! As in the overwhelming majority of cases of the industry the mean is not known, we usually use the last two relations bordered above. Now this is where comes the vicious part: when we calculate the bias of this two estimators, the first is biased, the second is not. So we tend to use only the latter. Nay! Because we could also talk about the variance and precision of an estimator, which are also important criteria for judging the quality of an estimator relative to another. If we were to calculate the variance of the two estimators, then the first, which is biased, is smaller than the second which is unbiased variance! All that to say that the criteria of bias is not (by far) the only one to be study to judge the quality of an estimator.

Finally, it is important to remember that the factor -1 in the denominator of the unbiased maximum likelihood estimator stems from the need to correct the mean of the biased estimator initially subtracted by one time the standard error!

---

[40] Some textbooks also define the "Bessel factor" as $n/(n-1)$ like if we multiply the uncorrected sample variance by it we get the unbiased estimator of the population variance.

5.7.7.2 Poisson Distribution MLE

Using the same method as for the Normal (Gauss-Laplace) distribution, we will seek the maximum likelihood estimators of the Poisson distribution which for recall is given by:

$$P_k=\frac{\mu^k}{k!}e^{-\mu} \tag{5.7.1226}$$

Thus, the likelihood is given by:

$$L(\mu)=\prod_{i=1}^{n}P(x_i,\mu)=\frac{\mu^{\sum_{i=1}^{n}x_i}}{\prod_{i=1}^{n}x_i!}e^{-\mu n} \tag{5.7.1227}$$

Maximize a function or maximize its logarithm is equivalent therefore:

$$\ln[L(\mu)]=\ln(\mu)\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}\ln(x_i!)-\mu n \tag{5.7.1228}$$

We are now looking to maximize it:

$$\frac{\partial\ln[L(\mu)]}{\partial\mu}=\frac{1}{\mu}\sum_{i=1}^{n}x_i-n=0 \tag{5.7.1229}$$

and thus we obtain the only maximum likelihood estimator that will be:

$$\boxed{\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i} \tag{5.7.1230}$$

It is quite normal to find in this example the sample mean because it is the best possible estimator for the parameter of the Poisson distribution (which also represents the mean of a Poisson distribution).

Knowing that the standard deviation of this particular distribution (see above during the development of the Poisson distribution) is the square root of the mean, then we have for the standard deviation maximum likelihood:

$$\boxed{\hat{\sigma}=\sqrt{\frac{1}{\displaystyle\sum_{i=1}^{n}x_i}}} \tag{5.7.1231}$$

Remark

We show in the same way identical results for the exponential distribution that is widely used in preventive maintenance and reliability!

5.7.7.3 Binomial Distribution MLE

Suppose we have a Bernoulli model in which each observation has a constant and equal chance of success denoted p. A Bernoulli variable takes one of two values, conventionally 1 or 0, which indicate a 'success' or 'failure'. The probability distribution for this variable is:

$$\begin{aligned} f(1) &= p \\ f(0) &= 1-p \end{aligned} \tag{5.7.1232}$$

If S stands for the number of successes and F for the number of failures, then the likelihood is:

$$L(p)=\prod_{i=1}^{S} p \prod_{i=1}^{F}(1-p)=p^S(1-p)^F \tag{5.7.1233}$$

and the log-likelihood would then be:

$$\ln(L(p))=S\ln(p)+F\ln(1-p) \tag{5.7.1234}$$

We then take the derivative of this with respect of p:

$$\dfrac{\partial \ln(L(p))}{\partial p}=\dfrac{S}{p}-\dfrac{F}{1-p} \tag{5.7.1235}$$

We now set this equal to zero and solve for p and then we get immediately:

$$\boxed{\hat{p}=\dfrac{S}{S+F}} \tag{5.7.1236}$$

5.7.7.4 Geometric Distribution MLE

Using the same method as for the Normal distribution (Gauss-Laplace) and the Poisson distribution, we will seek the maximum likelihood estimator of the Binomial which we recall, is given by:

$$P(N,k)=C_k^N\cdot P=\dfrac{N!}{k!(N-k!)}p^k q^{N-k}=\dfrac{N!}{k!(N-k!)}p^k(1-p)^{N-k} \tag{5.7.1237}$$

Accordingly, the likelihood is given by:

$$L(p)=\prod_{i=1}^{N}P(x_i,p)=C_k^N p^k(1-p)^{N-k} \tag{5.7.1238}$$

It should be remembered that the factor following the combinatorial term already expressed the successive variables according to what we saw during our study of the Bernoulli and Binomial distribution functions. Hence the disappearance of the product in the preceding equality.

Maximize a function or maximize its logarithm is equivalent therefore:

$$\ln[L(p)]=\ln(C_k^N)+k\ln(p)+(N-k)\ln(1-p) \tag{5.7.1239}$$

We are now looking to maximize it:

$$\frac{\partial \ln[L(p)]}{\partial p}=\frac{k}{p}-\frac{N-k}{1-p}=0 \tag{5.7.1240}$$

The reader may have perhaps noticed that the binomial coefficient has disappeared. Therefore, we immediately deduce that the estimator of the binomial distribution is the same as the geometric distribution.

Which gives:

$$k(1-p)-p(N-k)=k-kp-pN+pk=k-pN=0 \tag{5.7.1241}$$

from which we derive the maximum likelihood estimator:

$$\boxed{\hat{p}=\frac{k}{N}} \tag{5.7.1242}$$

This result is quite intuitive if we consider the classic example of a coin that has a chance on two of dropping on one of its faces. The probability p being the number of times k a given face where was observed in the total number of tests (all sides combined).

Remark

In practice, it is not as easy to apply these estimators! We must carefully consider which are most suitable for a given experiment and ideally also calculate the mean squared error (standard error) of each of the estimators of the mean (as we have already done for the empirical mean earlier). In short... it is a long process of reflection.

5.7.7.5 Weibull Distribution MLE

We saw in the section of Industrial Engineering a very detailed study of the three-parameter Weibull distribution with its standard deviation and mean because as we mentioned it is quite used in the field of reliability engineering (see section Industrial Engineering page 8736).

Unfortunately, the three parameters of this distribution are unknown in practice. Using estimators however we can determine the expression of two of the three assuming γ as zero. This gives us the following Weibull distribution named "Weibull distribution with two parameters":

$$P(x,\beta,\eta)=\frac{\beta}{\eta}\left(\frac{x}{\eta}\right)^{\beta-1}e^{-\left(\frac{x}{n}\right)^{\beta}} \tag{5.7.1243}$$

and for recall with $\beta>0$ and $\eta>0$.

Therefore the likelihood is given by:

$$L(\beta,\eta)=\prod_{i=1}^{n}P(x_i,\beta,\eta)=\prod_{i=1}^{n}\frac{\beta}{\eta}\left(\frac{x_i}{\eta}\right)^{\beta-1}e^{-\left(\frac{x_i}{n}\right)^{\beta}}=\left(\frac{\beta}{\eta}\right)^{n}\prod_{i=1}^{n}\left(\frac{x_i}{\eta}\right)^{\beta-1}e^{-\sum_{i=1}^{n}\left(\frac{x_i}{n}\right)^{\beta}} \tag{5.7.1244}$$

Maximize a function or maximize its logarithm is equivalent therefore:

$$\begin{aligned}\ln(L(\beta,\eta)) &= \ln\left(\left(\frac{\beta}{\eta}\right)^n\prod_{i=1}^n\left(\frac{x_i}{\eta}\right)^{\beta-1}e^{\sum_{i=1}^n\left(\frac{x_i}{\eta}\right)^\beta}\right)\\ &= n\ln\left(\frac{\beta}{\eta}\right)-\frac{1}{\eta^\beta}\sum_{i=1}^n x_i^\beta+\sum_{i=1}^n\ln\left[\left(\frac{x_i}{\eta}\right)^{\beta-1}\right]\\ &= n\ln\left(\frac{\beta}{\eta}\right)-\eta^{-\beta}\sum_{i=1}^n x_i^\beta+(\beta-1)\sum_{i=1}^n\ln\left(\frac{x_i}{\eta}\right)\end{aligned}\tag{5.7.1245}$$

Now we seek to maximize this by remembering that (see section Differential and Integral Calculus page 1475):

$$\frac{\mathrm{d}}{\mathrm{d}x}a^x=a^x\ln(a)\quad\text{and}\quad\frac{\mathrm{d}}{\mathrm{d}x}a^{-x}=-a^x\ln(a)\tag{5.7.1246}$$

then:

$$\begin{aligned}\frac{\partial\ln(\beta,\eta)}{\partial\beta} &= n\frac{1}{\beta}+\frac{1}{\eta^\beta}\ln(\eta)\sum_{i=1}^n x_i^\beta-\frac{1}{\eta^\beta}\sum_{i=1}^n x_i^\beta\ln(x_i)+\sum_{i=1}^n\ln\left(\frac{x_i}{\eta}\right)\\ &= n\frac{1}{\beta}+\frac{1}{\eta^\beta}\left[\sum_{i=1}^n x_i^\beta\left(\ln(\eta)-\ln(x_i)\right)\right]+\sum_{i=1}^n\ln\left(\frac{x_i}{\eta}\right)\\ &= n\frac{1}{\beta}-\frac{1}{\eta^\beta}\left[\sum_{i=1}^n x_i^\beta\left(\ln(x_i)-\ln(\eta)\right)\right]+\sum_{i=1}^n\ln\left(\frac{x_i}{\eta}\right)\\ &= n\frac{1}{\beta}-\frac{1}{\eta^\beta}\sum_{i=1}^n x_i^\beta\ln\left(\frac{x_i}{\eta}\right)+\sum_{i=1}^n\ln\left(\frac{x_i}{\eta}\right)=0\end{aligned}\tag{5.7.1247}$$

And we get for the second parameter:

$$\frac{\partial\ln(\beta,\eta)}{\partial\eta}=-n\frac{1}{\nu}-\frac{\beta}{\eta^\beta\eta}\sum_{i=1}^n x_i^\beta+(1-\beta)n\frac{1}{\eta}=0\tag{5.7.1248}$$

then:

$$-\frac{1}{\eta^\beta}\sum_{i=1}^n x_i^\beta-n=0\tag{5.7.1249}$$

Finally to summarize with the correct notations (and in the resolution order in practice):

$$n\frac{1}{\hat{\beta}}-\frac{1}{\hat{\eta}^{\hat{\beta}}}\sum_{i=1}^n x_i^{\hat{\beta}}\ln\left(\frac{x_i}{\eta}\right)+\sum_{i=1}^n\ln\left(\frac{x_i}{\eta}\right)=0\quad\text{and}\quad-\frac{1}{\hat{\eta}^{\hat{\beta}}}\sum_{i=1}^n x_i^{\hat{\beta}}-n=0\tag{5.7.1250}$$

Solving these equations involves heavy computations and we can a priori do nothing with that in conventional spreadsheets softwares such as Microsoft® Excel® or Open Office Calc without programming (at least as far as we know...).

We then take a different approach by writing our Weibull distribution with two parameters as follows:

$$P(x,\beta,\theta)=\frac{\beta}{\theta}x^{\beta-1}e^{\frac{x^\beta}{\theta}} \tag{5.7.1251}$$

with for recall $\beta>0$ and $\theta>0$.

Therefore the likelihood is given by:

$$L(\beta,\eta)=\prod_{i=1}^{n}P(x_i,\beta,\theta)=\prod_{i=1}^{n}\frac{\beta}{\theta}x_i^{\beta-1}e^{-\frac{x_i^\beta}{\theta}}=\left(\frac{\beta}{\theta}\right)^n e^{-\sum\limits_{i=1}^{n}\dfrac{x_i^\beta}{\theta}}\prod_{i=1}^{n}x_i^{\beta-1} \tag{5.7.1252}$$

Maximize a function or maximize its logarithm is equivalent therefore:

$$\begin{aligned}\ln(L(\beta,\eta))&=\ln\left[\left(\frac{\beta}{\theta}\right)^n e^{-\sum\limits_{i=1}^{n}\dfrac{x_i^\beta}{\theta}}\prod_{i=1}^{n}x_i^{\beta-1}\right]\\&=n\ln\frac{\beta}{\theta}-\frac{1}{\theta}\sum_{i=1}^{n}x_i^\beta+\sum_{i=1}^{n}\ln(x_i^{\beta-1})=n\ln\frac{\beta}{\theta}-\frac{1}{\theta}\sum_{i=1}^{n}x_i^\beta+(\beta-1)\sum_{i=1}^{n}\ln(x_i)\end{aligned} \tag{5.7.1253}$$

Now we seek to maximize this by remembering that (see section Differential and Integral Calculus page 1475):

$$\frac{\mathrm{d}}{\mathrm{d}x}a^x=a^x\ln(a)\quad\text{and}\quad\frac{\mathrm{d}}{\mathrm{d}x}a^{-x}=-a^x\ln(a) \tag{5.7.1254}$$

then:

$$\frac{\partial\ln(\beta,\theta)}{\partial\beta}=\frac{n}{\beta}-\frac{1}{\theta}\sum_{i=1}^{n}x_i^\beta\ln(x_i)+\sum_{i=1}^{n}\ln(x_i)=0 \tag{5.7.1255}$$

And we have for the second parameter:

$$\frac{\partial\ln(\beta,\theta)}{\partial\theta}=-\frac{n}{\theta}+\frac{1}{\theta^2}\sum_{i=1}^{n}x_i^\beta=0 \tag{5.7.1256}$$

It is then immediate that:

$$\bar{\theta}=\frac{1}{n}\sum_{i=1}^{n}x_i^\beta \tag{5.7.1257}$$

injected into the equation:

$$\frac{n}{\beta}-\frac{1}{\theta}\sum_{i=1}^{n}x_i^\beta\ln(x_i)+\sum_{i=1}^{n}\ln(x_i)=0 \tag{5.7.1258}$$

We get:

$$\frac{n}{\beta}-\frac{\displaystyle\sum_{i=1}^{n}x_i^{\beta}\ln(x_i)}{\displaystyle\frac{1}{n}\sum_{i=1}^{n}x_i^{\beta}}+\sum_{i=1}^{n}\ln(x_i)=0 \tag{5.7.1259}$$

simplifying:

$$\frac{\displaystyle\sum_{i=1}^{n}x_i^{\beta}\ln(x_i)}{\displaystyle\frac{1}{n}\sum_{i=1}^{n}x_i^{\beta}}-\frac{1}{\beta}=\frac{1}{n}\sum_{i=1}^{n}\ln(x_i) \tag{5.7.1260}$$

The resolution of the two equations (in order from top to bottom):

$$\boxed{\begin{aligned}&\frac{\displaystyle\sum_{i=1}^{n}x_i^{\beta}\ln(x_i)}{\displaystyle\frac{1}{n}\sum_{i=1}^{n}x_i^{\beta}}-\frac{1}{\beta}-\frac{1}{n}\sum_{i=1}^{n}\ln(x_i)=0\\&\frac{1}{n}\sum_{i=1}^{n}x_i^{\beta}-\bar{\theta}\end{aligned}} \tag{5.7.1261}$$

can easily be calculated with the Target Tool of Microsoft® Excel® or Open Office Calc.

5.7.7.6 Gamma Distribution MLE

Here we will use a technique named "method of moments" to determine the estimators of the parameters of the Gamma distribution.

Suppose that $X_1,\ldots,X_n$ are independent and identically distributed random variables according to the Gamma distribution with density:

$$P_{a,\lambda}(x)=\frac{x^{a-1}e^{-\lambda x}}{\Gamma(a)}\lambda^{a}1_{[0,+\infty[} \tag{5.7.1262}$$

We seek to estimate a,λ. For this, we first determine some theoretical moments. The first moment is the expected mean that as we have proved before is given by:

$$\mathrm{E}(X)=m_1=\frac{a}{\lambda} \tag{5.7.1263}$$

and the second moment, the mean of the square of the random variable, is as we have implicitly proved in the proof of the variance of the Gamma distribution given by:

$$\mathrm{E}(X^2)=m_2=\frac{a(a+1)}{\lambda^2} \tag{5.7.1264}$$

We then express the relation between the parameters and the theoretical moments:

$$\begin{aligned} m_1 &= \frac{a}{\lambda} \\ m_2 &= \frac{a(a+1)}{\lambda^2} \end{aligned} \tag{5.7.1265}$$

The resolution of this simple system gives:

$$\begin{aligned} a &= \frac{m_1^2}{m_2 - m_1^2} \\ \lambda &= \frac{m_2 - m_1^2}{m_1^2} \end{aligned} \tag{5.7.1266}$$

Once this system established, the method of moments consist to use the empirical moments, i.e. for our example the first two, $\hat{m}_1, \hat{m}_2$:

$$\begin{aligned} \hat{m}_1 &= \frac{X_1 + \ldots + X_n}{n} \\ \hat{m}_2 &= \frac{X_1^2 + \ldots + X_n^2}{n} \end{aligned} \tag{5.7.1267}$$

that we define as equal to the true theoretical moments ... Therefore, it comes:

$$\boxed{\begin{aligned} a &= \frac{\hat{m}_1^2}{\hat{m}_2 - \hat{m}_1^2} \\ \lambda &= \frac{\hat{m}_2 - \hat{m}_1^2}{\hat{m}_1^2} \end{aligned}} \tag{5.7.1268}$$

5.7.7.7 Pareto Distribution MLE

Let us recall the Pareto density given by:

$$f(x) = k\frac{x_m^k}{x^{k+1}} \tag{5.7.1269}$$

Let us choose a more convenient and traditional notation:

$$f(x, \alpha, \beta) = \alpha\frac{\beta^\alpha}{x^\alpha} \qquad x \geq \beta \qquad \alpha, \beta > 0 \tag{5.7.1270}$$

The likelihood function of the Pareto distribution given a sample $x = (x_1, \ldots, x_n)$ is given by:

$$L(\alpha, \beta, x) = \prod_{i=1}^{n} \alpha\frac{\beta^\alpha}{x_i^{\alpha+1}} = \alpha^n \beta^{n\alpha} \prod_{i=1}^{n} \frac{1}{x_i^{\alpha+1}} \tag{5.7.1271}$$

Taking the log of the likelihood, we obtain:

$$\ln[L(\alpha, \beta; x)] \equiv \ell(\alpha, \beta; x) = n\ln(\alpha) + n\alpha\ln(\beta) - (\alpha+1)\sum_{i=1}^{n} \ln(x_i). \tag{5.7.1272}$$

Since a higher β will always result in a higher likelihood - $\ln(\beta)$ is monotonically increasing - we maximize the likelihood by setting $\hat{\beta}$ as high as possible (we can't use differentiation to maximize the likelihood in this case). Since $\beta \leq x_i$ for all i, we maximize the likelihood by setting $\hat{\beta} = \min\limits_i x_i$, the smallest x_i in the sample.

For α, we set the partial derivative of ℓ with respect to α equal to 0:

$$\frac{\partial \ell(\alpha,\beta)}{\partial \alpha} = \frac{n}{\alpha} + n\ln(\beta) - \sum_{i=1}^{n} \ln\left(x_i\right) = 0 \qquad (5.7.1273)$$

Therefore:

$$\hat{\alpha} = \frac{n}{\displaystyle\sum_{i=1}^{n} \ln\left(x_i\right) - n\ln(\hat{\beta})} = \frac{1}{\displaystyle\frac{1}{n}\sum_{i=1}^{n} \ln\left(x_i\right) - n\ln(\hat{\beta})} \qquad (5.7.1274)$$

5.7.7.8 Censored data MLE

The "likelihood of censored data" is of crucial importance in the field of industrial reliability and also in pharmacies and clinical studies, knowing that for reasons of cost and unforeseen circumstances, there is often a limit to the time of observation or events which make that the subjects of study leave the experiment. Knowing that there are four types of situations (see figure below if needed):

1. "Uncensored data" (un), the event of interest is observed between the start time of measurement and end of measurement.
2. "Right-censored" (rc), the event did not occur during the limited observation time, that is, between the measurement start and end time.
3. "Left-censored" (lc), the object of study left the experiment for whatever reasons observed between the time of beginning of measurement and end of measurement, or we do not know with precision at which what should have been measured started.
4. "Censorship interval" (ci), an event of interest took place between two instants, but we do not know exactly when in this interval (before, between or after the measurement periods).

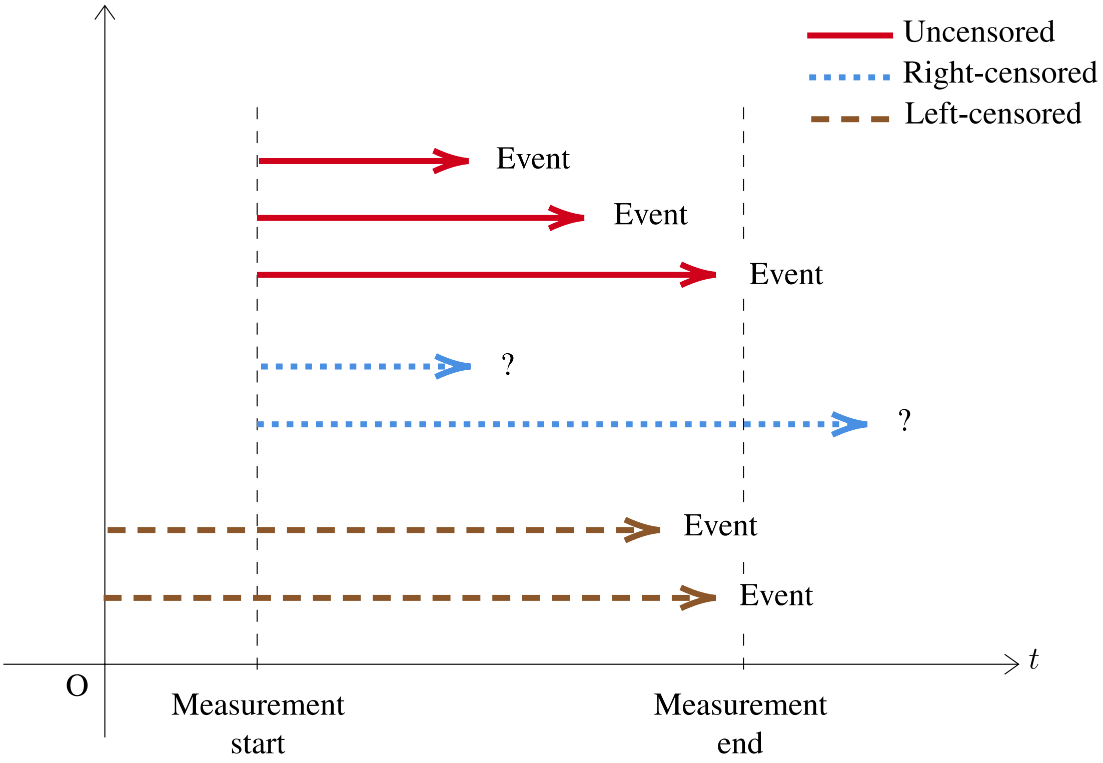

Uncensored
Right-censored
Left-censored
Event
Event
Event
?
?
Event
Event
t
O
Measurement start
Measurement end

Figure 5.157: **Type of censored data**

It is then relatively intuitive that if all events are independent then the likelihood function will be given by:

$$L(\theta)=\prod_{T_i\in\text{uc}}P(T=T_i,\theta)\prod_{T_i\in\text{lc}}P(T<T_i,\theta)\prod_{T_i\in\text{rc}}P(T>T_i,\theta)\prod_{T_i\in\text{ci}}P(T_{i,l}<T<T_{i,r},\theta) \tag{5.7.1275}$$

where the index l is for "left" and the index r is for "right".

We have obviously (think of the discrete case before thinking of the continuous case):

$$\begin{aligned}P(T=T_i,\theta)&=f(T_i,\theta)\\P(T<T_i,\theta)&=F(T_i,\theta)=1-R(T_i,\theta)\\P(T>T_i,\theta)&=1-F(T_i,\theta)=R(T_i,\theta)\\P(T_{i,l}<T<T_{i,r},\theta)&=F(T_{t,r},\theta)-F(T_{t,l},\theta)=R(T_{i,l},\theta)-R(T_{i,r},\theta)\end{aligned} \tag{5.7.1276}$$

In the field of reliability engineering that interest us here, we often reduce ourselves to the following cases:

$$L(\theta)=\prod_{T_i\in\text{uc}}P(T=T_i,\theta)\prod_{T_i\in\text{rc}}P(T>T_i,\theta) \tag{5.7.1277}$$

but with the following notation (...):

$$\boxed{L(\theta)=\prod_{T_i\in\text{uc}}P(T=T_i,\theta)\prod_{T_i\in\text{rc}}R(T_i,\theta)} \tag{5.7.1278}$$

In many experiments, the number of faulty elements f to be observed on n is fixed in advance (we often speak of "type II censorship") or simply known afterwards. Therefore, we have:

$$L(\theta)=\prod_{i=1}^{f} P(T=T_i,\theta)\prod_{i=f+1}^{n} R(T_i,\theta) \tag{5.7.1279}$$

Let us explicit this probability with Weibull's law. Let us recall that a possible form of writing of the density function is written for the latter (see section Industrial Engineering page 8736):

$$f(x)=\frac{\alpha}{\beta}\frac{x^{\alpha-1}}{\beta^{\alpha-1}}e^{-\frac{x^{\alpha}}{\beta^{\alpha}}}=\frac{\alpha}{\beta}x^{\alpha-1}e^{-\left(\frac{x}{\beta}^{\alpha}\right)} \tag{5.7.1280}$$

and that we have proved that its survival function was given by:

$$R(x)=e^{-\left(\frac{x}{\beta}^{\alpha}\right)} \tag{5.7.1281}$$

Therefore:

$$L(\alpha,\beta)=\prod_{i=1}^{f}\frac{\alpha}{\beta}\left(\frac{t_i}{\beta}\right)^{\alpha-1}e^{-\left(\frac{t_i}{\beta}\right)^{\alpha}}\prod_{i=f+1}^{n}e^{-\left(\frac{t_i}{\beta}\right)^{\alpha}}=\left(\frac{\alpha}{\beta^{\alpha}}\right)^{f}e^{-\frac{1}{\beta^{\alpha}}\sum_{i=1}^{n}t_i^{\alpha}}\prod_{i=1}^{f}t_i^{\alpha-1} \tag{5.7.1282}$$

Maximizing the likelihood is equivalent to maximizing the logarithm. Since then:

$$\ln[L(\alpha,\beta)]=f\ln(\alpha)-f\alpha\ln(\beta)-\sum_{i=1}^{n}\left(\frac{t_i}{\beta}\right)^{\alpha}+(\alpha-1)\sum_{i=1}^{f}\ln(t_i) \tag{5.7.1283}$$

We calculate the partial derivative with respect to α, β (see section of Differential and Integral Calculus page 1475):

$$\begin{aligned}\frac{\partial\ln[L(\alpha,\beta)]}{\partial\alpha}&=\frac{f}{\alpha}-f\ln(\beta)-\sum_{i=1}^{n}\left(\frac{t_i}{\beta}\right)^{\alpha}\ln\left(\frac{t_i}{\beta}\right)+\sum_{i=1}^{f}\ln(t_i)=0\\ \frac{\partial\ln[L(\alpha,\beta)]}{\partial\alpha}&=-\frac{f\alpha}{\beta}+\frac{\alpha}{\beta}\sum_{i=1}^{n}\left(\frac{t_i}{\beta}\right)^{\alpha}=0\end{aligned} \tag{5.7.1284}$$

The second relation can easily be explicited with respect to β such that:

$$\hat{\beta}=\left(\frac{1}{f}\sum_{i=1}^{n}t_i^{\alpha}\right)^{1/\hat{\alpha}} \tag{5.7.1285}$$

In any case, it seems that it will be necessary to use numerical optimization techniques to determine all parameters.

5.7.7.9 Fisher Information Matrix

We will introduce now a matrix that will be useful to us much more later for some advanced statistical tests but also to determine (among others!) non-obvious statistical law parameters and also an important information criterion named the "Akaike Information Criterion" (see section Numerical Methods page 6420).

We have seen earlier that the partial derivative of the log-likelihood (that we will denote $\mathcal{L}$ for the following developments) relatively to a given parameter θ_i of a probability function (depending of a set of parameters $\vec{\theta}$) will give us an estimator of this parameter if put it as equal to zero.

Also respectively we then have that the partial derivative valuated at this correspond estimator then gives:

$$\frac{\partial\mathcal{L}(\vec{\theta})}{\partial\theta_i}(\hat{\theta})=0 \tag{5.7.1286}$$

Remark

The expression:

$$\mathcal{S}_i(\vec{\theta})=\frac{\partial\mathcal{L}(\vec{\theta})}{\partial\theta_i}(\hat{\theta})=\frac{\partial\ln(L(\vec{\theta}))}{\partial\theta_i}(\hat{\theta}) \tag{5.7.1287}$$

is known as the "score function".

Let us recall that during our study of bivariate Taylor series (see section Sequences and Series page 1714), we have proved that:

$$f(x_0+h,y_0+k)\cong f(x_0,y_0)+\frac{\partial f}{\partial x}(x_0,y_0)h+\frac{\partial f}{\partial y}(x_0,y_0)k+\frac{1}{2!}(h\ k)H\begin{pmatrix}h\\k\end{pmatrix} \tag{5.7.1288}$$

where:

$$H=\begin{bmatrix}\dfrac{\partial^2 f}{\partial x^2} & \dfrac{\partial^2 f}{\partial x\partial y}\\ \dfrac{\partial^2 f}{\partial y\partial x} & \dfrac{\partial^2 f}{\partial y^2}\end{bmatrix} \tag{5.7.1289}$$

Let us do the following replacement:

$$h=x-x_0\qquad k=y-y_0 \tag{5.7.1290}$$

We then have:

$$f(x,y)\cong f(x_0,y_0)+\frac{\partial f}{\partial x}(x_0,y_0)(x-x_0)+\frac{\partial f}{\partial y}(x_0,y_0)(y-y_0)+\frac{1}{2!}(x-x_0\ y-y_0)H\begin{pmatrix}x-x_0\\y-y_0\end{pmatrix} \tag{5.7.1291}$$

Remark

In the univariate case this is also written:

$$f(x) \cong f(x_0) + \frac{\partial f}{\partial x}(x_0)(x-x_0) + \frac{1}{2!}(x-x_0)H(x-x_0) \tag{5.7.1292}$$

or in some textbooks:

$$f(x) \cong f(x_0) + \frac{\partial f}{\partial x}(x_0)(x-x_0) + \frac{1}{2!}(x-x_0)H(x_0)(x-x_0) \tag{5.7.1293}$$

In vector form this is often written:

$$f(\vec{v}) \cong f(\vec{v}_0) + \left(\vec{\nabla} f(\vec{v}_0)\right) \circ (\vec{v}) + \frac{1}{2!}\vec{v}^T H \vec{v} \tag{5.7.1294}$$

Now remember that $f(\vec{v}_0)$ and $\vec{\nabla} f(\vec{v}_0)$ does not contain anymore any term in x or y as the function is 'valuated'. Hence if we take the gradient of the whole previous expression, we get:

$$\vec{\nabla} f(\vec{v}) \cong \underbrace{\vec{\nabla} f(\vec{v}_0)}_{=0} + \vec{\nabla}\left(\vec{\nabla} f(\vec{v}_0) \circ (\vec{v})\right) + \frac{1}{2!}\vec{\nabla}\left(\vec{v}^T H \vec{v}\right) = \vec{\nabla} f(\vec{v}_0) + H\vec{v} \tag{5.7.1295}$$

If you don't understand that result, do it again with the univariate case may help. Indeed, starting from:

$$f(x) \cong f(x_0) + \frac{\partial f}{\partial x}(x_0)(x-x_0) + \frac{1}{2!}(x-x_0)H(x_0)(x-x_0) \tag{5.7.1296}$$

We then have in a very very detailed way:

$$\begin{aligned}
\frac{\partial f(x)}{\partial x} &\cong \underbrace{\frac{\partial f(x_0)}{\partial x}}_{=0} + \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}(x_0)(x-x_0)\right) + \frac{1}{2!}\frac{\partial}{\partial x}\left((x-x_0)H(x_0)(x-x_0)\right)\\
&= \underbrace{\frac{\partial^2 f}{\partial x^2}(x_0)(x-x_0)}_{=0} + \frac{\partial f}{\partial x}(x_0)\underbrace{\frac{\partial f}{\partial x}(x-x_0)}_{=1}\\
&+ \frac{1}{2!}\left(\underbrace{\frac{\partial f}{\partial x}(x-x_0)}_{=1}H(x_0)(x-x_0) + \underbrace{(x-x_0)\frac{\partial f}{\partial x}H(x_0)(x-x_0)}_{=0} + (x-x_0)H(x_0)\underbrace{\frac{\partial f}{\partial x}(x-x_0)}_{=1}\right)\\
&= \frac{\partial f}{\partial x}(x_0) + \frac{1}{2!}\left(H(x_0)(x-x_0) + (x-x_0)H(x_0)\right) = \frac{\partial f}{\partial x}(x_0) + \frac{1}{2!}\left(2H(x_0)(x-x_0)\right)\\
&= \frac{\partial f}{\partial x}(x_0) + H(x_0)(x-x_0)
\end{aligned} \tag{5.7.1297}$$

That latter equality can indeed be written:

$$\vec{\nabla} f(x) \cong \vec{\nabla} f(x_0) + H(x_0)(x-x_0) \tag{5.7.1298}$$

Now let us come back to the more general equality:

$$\boxed{\vec{\nabla} f(\vec{v}) \cong \vec{\nabla} f(\vec{v}_0) + H\vec{v}} \tag{5.7.1299}$$

We can solve this expression for the optimal point $\vec{\nabla} f(\vec{v}) = \vec{0}$, giving:

$$\vec{0} \cong \vec{\nabla} f(\vec{v}_0) + H\vec{v} = \vec{\nabla} f \begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix} + H \begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix} \tag{5.7.1300}$$

Hence:

$$\boxed{\begin{pmatrix} x \\ y \end{pmatrix} \cong \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - H^{-1}\vec{\nabla} f \begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix}} \tag{5.7.1301}$$

This is the Newton-Raphson method for multidimensional optimization by the way...

Now let us focus on the following equality from the previous one:

$$\begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix} \cong -H^{-1}\vec{\nabla} f \begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix} \tag{5.7.1302}$$

And let us change the notation and replace f by $\mathcal{L}$:

$$\begin{pmatrix} \theta_1 - \theta_{1,0} \\ \theta_2 - \theta_{2,0} \end{pmatrix} \cong -H^{-1}\vec{\nabla}\mathcal{L} \begin{pmatrix} \theta_1 - \theta_{1,0} \\ \theta_2 - \theta_{2,0} \end{pmatrix} = H^{-1}\vec{\nabla}\mathcal{L}(\vec{\theta}) \tag{5.7.1303}$$

where $\vec{\nabla}\mathcal{L}(\vec{\theta})$ (the gradient of the likelihood at the point $\vec{\theta}$) is named the "score vector".

Remark

In the univariate case this is will also be reduced to:

$$\theta - \theta_0 \cong -H^{-1}\frac{\partial \mathcal{L}(\theta_0)}{\partial \theta} = -\left[\frac{\partial^2 \mathcal{L}(\theta_0)}{\partial \theta^2}\right]^{-1} \frac{\partial \mathcal{L}(\theta_0)}{\partial \theta} \tag{5.7.1304}$$

Now let us have a bit 'fun'... Let us take the outer product (see section Linear Algebra page 2014) of $\vec{\theta}$, i.e.:

$$\vec{\theta}\vec{\theta}^T = \begin{pmatrix} \theta_1 - \theta_{1,0} \\ \theta_2 - \theta_{2,0} \end{pmatrix} \begin{pmatrix} \theta_1 - \theta_{1,0} & \theta_2 - \theta_{2,0} \end{pmatrix} = \begin{bmatrix} (\theta_1 - \theta_{1,0})^2 & (\theta_1 - \theta_{1,0})(\theta_2 - \theta_{2,0}) \\ \theta_2 - \theta_{2,0} & (\theta_2 - \theta_{2,0})^2 \end{bmatrix} \tag{5.7.1305}$$

Ok so far nothing interesting. But let us take the expected mean:

$$\mathrm{E}(\vec{\theta}\vec{\theta}^T) = \begin{bmatrix} \mathrm{E}(\theta_1 - \theta_{1,0})^2 & \mathrm{E}((\theta_1 - \theta_{1,0})(\theta_2 - \theta_{2,0})) \\ \mathrm{E}((\theta_2 - \theta_{2,0})(\theta_1 - \theta_{1,0})) & \mathrm{E}(\theta_2 - \theta_{2,0})^2 \end{bmatrix} \tag{5.7.1306}$$

Here we recognize something! The elements of the diagonal look similar to a variance... The variance of θ_1 relatively to $\theta_{1,0}$. It is common to assume that θ_1 is the estimator and $\theta_{1,0}$ is the true value, then we write (keep in mind that this can be generalized to more than 2 parameters!):

$$\mathrm{E}(\vec{\theta}\vec{\theta}^T) = \begin{bmatrix} \mathrm{E}(\hat{\theta}_1 - \theta_1)^2 & \mathrm{E}((\hat{\theta}_1 - \theta_1)(\hat{\theta}_2 - \theta_2)) \\ \mathrm{E}((\hat{\theta}_2 - \theta_2)(\hat{\theta}_1 - \theta_1)) & \mathrm{E}(\hat{\theta}_2 - \theta_2)^2 \end{bmatrix} = \begin{bmatrix} \mathrm{V}(\hat{\theta}_1) & \mathrm{cov}(\hat{\theta}_1, \hat{\theta}_2) \\ \mathrm{cov}(\hat{\theta}_1, \hat{\theta}_2) & \mathrm{V}(\hat{\theta}_2 - \theta_2)) \end{bmatrix} = \Sigma(\vec{\theta}) \tag{5.7.1307}$$

We recognize here the variance-covariance matrix[41] (see above page 604). Ok that's nice but so what??? What does this bring us? In fact using previous equalities we have:

$$\mathrm{E}(\vec{\theta}\vec{\theta}^T) = \Sigma(\vec{\theta}) \cong \mathrm{E}\left[\left(H^{-1}\vec{\nabla}\mathcal{L}(\vec{\theta})\right)\left(H^{-1}\vec{\nabla}\mathcal{L}(\vec{\theta})\right)^T\right] \tag{5.7.1308}$$

---

[41] Sadly denoted in some textbooks $\mathrm{V}(\vec{\theta})$...

Using the properties of transposition as proved in the section of Linear Algebra and the property of matrix multiplication association property, that latter relation can be written:

$$\Sigma(\vec{\theta}) \cong \mathrm{E}\left[\left(H^{-1}\vec{\nabla}\mathcal{L}(\vec{\theta})\right)\left(\vec{\nabla}\mathcal{L}(\vec{\theta})^T(H^{-1})^T\right)\right] = \mathrm{E}\left[H^{-1}\left(\vec{\nabla}\mathcal{L}(\vec{\theta})\vec{\nabla}\mathcal{L}(\vec{\theta})^T\right)(H^{-1})^T\right] \tag{5.7.1309}$$

Using another property of the transposed, this can be also written:

$$\Sigma(\vec{\theta}) \cong \mathrm{E}\left[H^{-1}\left(\vec{\nabla}\mathcal{L}(\vec{\theta})\vec{\nabla}\mathcal{L}(\vec{\theta})^T\right)(H^T)^{-1}\right] \tag{5.7.1310}$$

Assuming the Hessian is symmetric such that $H^T = H$ (if Schwarz's theorem[42] applies on the function considered inside the Hessian!), we then have:

$$\Sigma(\vec{\theta}) \cong \mathrm{E}\left[H^{-1}\left(\vec{\nabla}\mathcal{L}(\vec{\theta})\vec{\nabla}\mathcal{L}(\vec{\theta})^T\right)H^{-1}\right] \tag{5.7.1311}$$

Using the linearity property of the mean:

$$\Sigma(\vec{\theta}) \cong \left[\mathrm{E}(H^{-1})\mathrm{E}\left(\vec{\nabla}\mathcal{L}(\vec{\theta})\vec{\nabla}\mathcal{L}(\vec{\theta})^T\right)\mathrm{E}(H^{-1})\right] \tag{5.7.1312}$$

Now let us focus on:

$$\mathrm{E}\left(\vec{\nabla}\mathcal{L}(\vec{\theta})\vec{\nabla}\mathcal{L}(\vec{\theta})^T\right) \tag{5.7.1313}$$

But we need an important result before! As the likelihood $L(\vec{x}),\vec{\theta})$ is a density, we know that:

$$\int L(\vec{x},\vec{\theta})\mathrm{d}x = 1 \tag{5.7.1314}$$

Now by differentiating both sides with respect to one of the parameter θ_i we get:

$$\frac{\partial}{\partial\theta_i}\int L(\vec{x},\vec{\theta})\mathrm{d}x = 0 \tag{5.7.1315}$$

By assuming (this assumption must be checked on every case!) that we have:

$$\frac{\partial}{\partial\theta_i}\int L(\vec{x},\vec{\theta})\mathrm{d}x = \int\frac{\partial L(\vec{x},\vec{\theta})}{\partial\theta_i}\mathrm{d}x \tag{5.7.1316}$$

Then:

$$\int\frac{\partial L(\vec{x},\vec{\theta})}{\partial\theta_i}\mathrm{d}x = 0 \tag{5.7.1317}$$

But this is also equal to:

$$\int\frac{\partial\ln(L(\vec{x},\vec{\theta}))}{\partial\theta_i}L(\vec{x},\vec{\theta})\mathrm{d}x = 0 \tag{5.7.1318}$$

---

[42] See section Differential and Integral Calculus page 1469

Differentiating again with respect to θ_i and taking the derivative inside gives:

$$\begin{aligned}\frac{\partial}{\partial\theta_i}\int\frac{\partial\ln(L(\vec{x},\vec{\theta}))}{\partial\theta_i}L(\vec{x},\vec{\theta})\mathrm{d}x&=\int\frac{\partial^2\ln(L(\vec{x},\vec{\theta}))}{\partial\theta_i^2}L(\vec{x},\vec{\theta})\mathrm{d}x+\int\frac{\partial\ln(L(\vec{x},\vec{\theta}))}{\partial\theta_i}\frac{\partial L(\vec{x},\vec{\theta})}{\partial\theta_i}\mathrm{d}x\\&=\int\frac{\partial^2\ln(L(\vec{x},\vec{\theta}))}{\partial\theta_i^2}L(\vec{x},\vec{\theta})\mathrm{d}x+\int\frac{\partial\ln(L(\vec{x},\vec{\theta}))}{\partial\theta_i}\frac{1}{L(\vec{x},\vec{\theta})}\frac{\partial L(\vec{x},\vec{\theta})}{\partial\theta_i}L(\vec{x},\vec{\theta})\mathrm{d}x\\&=\int\frac{\partial^2\ln(L(\vec{x},\vec{\theta}))}{\partial\theta_i^2}L(\vec{x},\vec{\theta})\mathrm{d}x+\int\frac{\partial\ln(L(\vec{x},\vec{\theta}))}{\partial\theta_i}\frac{\partial\ln(L(\vec{x},\vec{\theta}))}{\partial\theta_i}L(\vec{x},\vec{\theta})\mathrm{d}x\\&=0\end{aligned}\tag{5.7.1319}$$

Hence:

$$\int\frac{\partial\ln(L(\vec{x},\vec{\theta}))}{\partial\theta_i}\frac{\partial\ln(L(\vec{x},\vec{\theta}))}{\partial\theta_i}L(\vec{x},\vec{\theta})\mathrm{d}x=-\int\frac{\partial^2\ln(L(\vec{x},\vec{\theta}))}{\partial\theta_i^2}L(\vec{x},\vec{\theta})\mathrm{d}x\tag{5.7.1320}$$

Written in a more common way by statisticians as following:

$$\mathrm{E}\left(\frac{\partial\ln(L(\vec{x},\vec{\theta}))}{\partial\theta_i}\frac{\partial\ln(L(\vec{x},\vec{\theta}))}{\partial\theta_i}\right)=-\mathrm{E}\left(\frac{\partial^2\ln(L(\vec{x},\vec{\theta}))}{\partial\theta_i^2}\right)\tag{5.7.1321}$$

And using the notation defined earlier:

$$\mathrm{E}\left(\frac{\partial\mathcal{L}(\vec{x},\vec{\theta})}{\partial\theta_i}\frac{\partial\mathcal{L}(\vec{x},\vec{\theta})}{\partial\theta_i}\right)=-\mathrm{E}\left(\frac{\partial^2\mathcal{L}(\vec{x},\vec{\theta})}{\partial\theta_i^2}\right)\tag{5.7.1322}$$

If we do the same development again but when taking the partial derivative a second time but relatively to θ_j instead of θ_i we get immediately:

$$\mathrm{E}\left(\frac{\partial\mathcal{L}(\vec{x},\vec{\theta})}{\partial\theta_i}\frac{\partial\mathcal{L}(\vec{x},\vec{\theta})}{\partial\theta_j}\right)=-\mathrm{E}\left(\frac{\partial^2\mathcal{L}(\vec{x},\vec{\theta})}{\partial\theta_i\partial\theta_j}\right)=-\mathrm{E}\left(\frac{\partial^2\mathcal{L}(\vec{x},\vec{\theta})}{\partial\vec{\theta}}\right)\tag{5.7.1323}$$

This can be put obviously in matrix form for each component i,j. Then it is usual to define the mean of matrix made of the components $\frac{\partial\mathcal{L}(\vec{x},\vec{\theta})}{\partial\theta_i}\frac{\partial\mathcal{L}(\vec{x},\vec{\theta})}{\partial\theta_j}$ as the "Fisher information matrix" denoted $\mathcal{I}(\vec{\theta})$ and we recognize that a matrix made of the components $\frac{\partial^2\mathcal{L}(\vec{x},\vec{\theta})}{\partial\theta_i\partial\theta_j}$ is the "Hessian" (see section Sequences and Series page 1715) denoted for recall $H(\vec{\theta})$, then we have:

$$\boxed{\mathcal{I}(\vec{\theta})=-\mathrm{E}\left(H(\vec{\theta})\right)}\tag{5.7.1324}$$

And this is the relation we were expecting to find!

So now let us come back to:

$$\Sigma(\vec{\theta})=\left[\mathrm{E}(H^{-1})\mathrm{E}\left(\vec{\nabla}\mathcal{L}(\vec{\theta})\vec{\nabla}\mathcal{L}(\vec{\theta})^T\right)\mathrm{E}(H^{-1})\right]\tag{5.7.1325}$$

That we can now rewrite as:

$$\Sigma(\vec{\theta})=\left[\mathrm{E}(H^{-1})\mathcal{I}(\vec{\theta})\mathrm{E}(H^{-1})\right]=\left[\mathrm{E}(H^{-1})(-\mathrm{E}(H(\vec{\theta}))\mathrm{E}(H^{-1})\right]=-\left[\mathrm{E}(H^{-1})\mathrm{E}(H)\mathrm{E}(H^{-1})\right]\tag{5.7.1326}$$

So finally we get the important final result we were seeking from the start named the "information matrix equality":

$$\boxed{\Sigma(\vec{\theta}) \cong -\mathrm{E}(H^{-1})} \tag{5.7.1327}$$

That is that the covariance-variance matrix (itself equal to the outer product of $\vec{\theta}$ with itself!), is equal to the negative on the inverse Hessian! Or in other words: the hessian of a likelihood function equals the information matrix, or the variance-covariance matrix of the score functions!

✍**Example:**

Our sample is made up of the first n terms of an independent and identically distributed sequence $\{X_n\}$ of normal random variables having mean μ and variance σ^2. The probability density function of a generic term of the sequence is:

$$f_X(x_j) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x_j-\mu)^2}{\sigma^2}} \tag{5.7.1328}$$

The mean μ and the variance σ^2 are the two parameters that need to be estimated.

We can read off the standard error of $\vec{\theta}$ from the square roots of the diagonal elements of the matrix Σ. Note that these are only correct asymptotically and are hard to calculate in finite samples.

We know that the likelihood function is:

$$L(\mu, \sigma^2, x_1, \ldots, y_n) = L(\vec{\theta}, \vec{x}) = \frac{1}{(\sqrt{2\pi\sigma^2})^2} e^{-\frac{1}{2}\sum_{j=1}^{n}\frac{(x_j-\mu)^2}{\sigma^2}} \tag{5.7.1329}$$

and hence (for recall):

$$\ln(L(\vec{\theta}, \vec{x})) = \mathcal{L}(\vec{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\mu)^2 \tag{5.7.1330}$$

The likelihood and their solution are for recall:

$$\begin{aligned}
\frac{\partial \ln(L(\vec{\theta})}{\partial \mu} &= \frac{\mathcal{L}(\vec{\theta})}{\partial \mu} = \frac{1}{\sigma^2}\sum_{j=1}^{n}(x_t-\mu) = 0 \Rightarrow \hat{\mu} = \frac{1}{n}\sum_{j=1}^{n} x_j \\
\frac{\partial \ln(L(\vec{\theta})}{\partial \sigma^2} &= \frac{\mathcal{L}(\vec{\theta})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\mu)^2 = 0 \Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_j-\hat{\mu})^2
\end{aligned} \tag{5.7.1331}$$

Now let us calculate the Hessian of $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2$, that is the Hessian of $\mathcal{L}(\vec{\theta})$. We have to compute 4 components of the matrix as the likelihood has 2 parameters ($2 \times 2 = 4$) and the out-diagonal components are equal assuming the Schwarz theorem applies. Then:

$$\begin{aligned}
\frac{\partial^2\mathcal{L}(\vec{\theta})}{\partial\mu^2} &= -\frac{n}{\sigma^2}\\
\frac{\partial^2\mathcal{L}(\vec{\theta})}{\partial\sigma^2} &= \frac{n}{2\sigma^4}-\frac{1}{\sigma^6}\sum_{j=1}^{n}(x_j-\mu)^2\\
\frac{\partial^2\mathcal{L}(\vec{\theta})}{\partial\sigma^2\partial\mu} &= -\frac{1}{\sigma^4}\sum_{j=1}^{n}(x_j-\mu)
\end{aligned}\tag{5.7.1332}$$

When evaluated at $\hat{\vec{\theta}}$:

$$\begin{aligned}
\left.\frac{\partial^2\mathcal{L}(\vec{\theta})}{\partial\mu^2}\right|_{\vec{\theta}=\hat{\vec{\theta}}} &= -\frac{n}{\hat{\sigma}^2}\\
\left.\frac{\partial^2\mathcal{L}(\vec{\theta})}{\partial\sigma^2}\right|_{\vec{\theta}=\hat{\vec{\theta}}} &= \frac{n}{2\hat{\sigma}^4}-\frac{1}{\hat{\sigma}^6}\sum_{j=1}^{n}(x_j-\hat{\mu})^2=\frac{n}{2\hat{\sigma}^4}-\frac{n\hat{\sigma}^2}{\hat{\sigma}^6}=-\frac{n}{2\hat{\sigma}^2}\\
\left.\frac{\partial^2\mathcal{L}(\vec{\theta})}{\partial\sigma^2\partial\mu}\right|_{\vec{\theta}} &= -\frac{1}{\hat{\sigma}^4}\sum_{j=1}^{n}(x_j-\mu)=0
\end{aligned}\tag{5.7.1333}$$

Hence:

$$H(\hat{\vec{\theta}})=\begin{bmatrix}-\dfrac{n}{\hat{\sigma}^2} & 0\\ 0 & -\dfrac{n}{2\hat{\sigma}^4}\end{bmatrix}\tag{5.7.1334}$$

We have also:

$$H(\vec{\theta})=\begin{bmatrix}-\dfrac{n}{\sigma^2} & -\dfrac{1}{\sigma^4}\displaystyle\sum_{j=1}^{n}(x_j-\mu)\\ -\dfrac{1}{\sigma^4}\displaystyle\sum_{j=1}^{n}(x_j-\mu) & \dfrac{n}{2\sigma^4}-\dfrac{1}{\sigma^6}\displaystyle\sum_{j=1}^{n}(x_j-\hat{\mu})^2\end{bmatrix}\tag{5.7.1335}$$

We have:

$$\begin{aligned}
\mathrm{E}\left[-\frac{n}{\sigma^2}\right] &= -\frac{n}{\sigma^2}\\
\mathrm{E}\left[-\frac{1}{\sigma^4}\sum_{j=1}^{n}(x_j-\mu)\right] &= -\frac{1}{\sigma^4}\sum_{j=1}^{n}\mathrm{E}(x_j-\mu)=0\\
\mathrm{E}\left[\frac{n}{2\sigma^4}-\frac{1}{\sigma^6}\sum_{j=1}^{n}(x_j-\hat{\mu})^2\right] &= \frac{n}{2\sigma^4}-\frac{1}{\sigma^6}\sum_{j=1}^{n}\mathrm{E}(x_j-\hat{\mu})^2=\frac{n}{2\sigma^4}-\frac{n\sigma^2}{\sigma^6}=-\frac{n}{2\sigma^4}
\end{aligned}\tag{5.7.1336}$$

Hence, the Fisher information matrix is given by:

$$\mathcal{I}(\vec{\theta})=-\mathrm{E}(H(\vec{\theta}))=\begin{bmatrix}\dfrac{n}{\sigma^2} & 0\\ 0 & \dfrac{n}{2\sigma^4}\end{bmatrix} \tag{5.7.1337}$$

Matrix inversion gives us also (see section Linear Algebra page 1980):

$$\Sigma(\vec{\theta})\cong-\mathrm{E}(H^{-1})=\dfrac{1}{\dfrac{n}{\sigma^2}\dfrac{n}{2\sigma^4}-0\cdot 0}\begin{bmatrix}\dfrac{n}{2\sigma^4} & 0\\ 0 & \dfrac{n}{\sigma^2}\end{bmatrix}=\begin{bmatrix}\dfrac{\sigma^2}{n} & 0\\ 0 & \dfrac{2\sigma^4}{n}\end{bmatrix} \tag{5.7.1338}$$

Hence:

$$\mathrm{V}(\vec{\theta})=\begin{pmatrix}\dfrac{\sigma^2}{n}\\ \dfrac{2\sigma^4}{n}\end{pmatrix} \tag{5.7.1339}$$

Now let us recall that during our proof of the central limit theorem, we have proved that (page 938):

$$Z_n=\dfrac{\hat{\mu}-\mu}{\sigma/\sqrt{n}}\cong\mathcal{N}(0,1) \tag{5.7.1340}$$

Therefore:

$$\sqrt{n}\,(\hat{\mu}-\mu)=\mathcal{N}(0,\sigma) \tag{5.7.1341}$$

Now instead of the estimator $\hat{\mu}$, let us consider the estimate of the score function (the σ is obviously the same as the previous relation):

$$\sqrt{n}\left(\hat{\mathcal{S}}_i(\theta)-\mathcal{S}_i(\theta)\right)=\mathcal{N}(0,\sigma) \tag{5.7.1342}$$

Now we consider that $\hat{\mathcal{S}}$ at the true value of θ (i.e. θ_0) is not zero as it is an estimator (but it's however probable near zero)! But $\mathcal{S}$ is however zero at θ_0. Then we have:

$$\sqrt{n}\left(\hat{\mathcal{S}}_i(\theta_0)-\mathcal{S}_i(\theta_0)\right)=\sqrt{n}\left(\hat{\mathcal{S}}_i(\theta_0)-0\right)=\mathcal{N}(0,\sigma) \tag{5.7.1343}$$

Hence:

$$\sqrt{n}\hat{\mathcal{S}}_i(\theta_0)=\mathcal{N}(0,\sigma) \tag{5.7.1344}$$

Or in vector form:

$$\sqrt{n}\vec{\hat{\mathcal{S}}}(\theta_0)=\mathcal{N}\left(0,\Sigma(\vec{\hat{\mathcal{S}}})\right) \tag{5.7.1345}$$

With explicitly, by construction of the variance-covariance matrix (for the special case of 2×2 to simplify the example):

$$\Sigma(\vec{\hat{\mathcal{S}}})=\begin{bmatrix}\sigma^2_{\hat{\mathcal{S}}_1} & \sigma_{\hat{\mathcal{S}}_1}\sigma^2_{\hat{\mathcal{S}}_2}\\ \sigma_{\hat{\mathcal{S}}_1}\sigma_{\hat{\mathcal{S}}_2} & \sigma^2_{\hat{\mathcal{S}}_2}\end{bmatrix}=\begin{bmatrix}\mathrm{E}((\hat{\mathcal{S}}_1-\mathcal{S}_1)^2) & \mathrm{E}((\hat{\mathcal{S}}_1-\mathcal{S}_1)(\hat{\mathcal{S}}_2-\mathcal{S}_2))\\ \mathrm{E}((\hat{\mathcal{S}}_1-\mathcal{S}_1)(\hat{\mathcal{S}}_2-\mathcal{S}_2)) & \mathrm{E}((\hat{\mathcal{S}}_2-\mathcal{S}_2)^2)\end{bmatrix} \tag{5.7.1346}$$

However, estimated at $\vec{\theta}_0$ the true values $\mathcal{S}_1$ and $\mathcal{S}_2$ are equal to zero. Then the variance-covariance matrix reduces to:

$$\Sigma(\vec{\hat{\mathcal{S}}})=\begin{bmatrix}\mathrm{E}(\hat{\mathcal{S}}_1^2) & \mathrm{E}(\hat{\mathcal{S}}_1\hat{\mathcal{S}}_2)\\ \mathrm{E}(\hat{\mathcal{S}}_1\hat{\mathcal{S}}_2) & \mathrm{E}(\hat{\mathcal{S}}_2^2)\end{bmatrix}=\mathrm{E}\left(\begin{bmatrix}\hat{\mathcal{S}}_1^2 & \hat{\mathcal{S}}_1\hat{\mathcal{S}}_2\\ \hat{\mathcal{S}}_1\hat{\mathcal{S}}_2 & \hat{\mathcal{S}}_2^2\end{bmatrix}\right) \tag{5.7.1347}$$

Or by expliciting the score functions:

$$\Sigma(\vec{\hat{\mathcal{S}}})=\mathrm{E}\left(\begin{bmatrix}\hat{\mathcal{S}}_1^2 & \hat{\mathcal{S}}_1\hat{\mathcal{S}}_2\\ \hat{\mathcal{S}}_1\hat{\mathcal{S}}_2 & \hat{\mathcal{S}}_2^2\end{bmatrix}\right)=\mathrm{E}\left(\begin{bmatrix}\left(\dfrac{\partial\mathcal{L}(\vec{x},\vec{\theta})}{\partial\theta_1}\right)^2 & \dfrac{\partial\mathcal{L}(\vec{x},\vec{\theta})}{\partial\theta_1}\dfrac{\partial\mathcal{L}(\vec{x},\vec{\theta})}{\partial\theta_2}\\ \dfrac{\partial\mathcal{L}(\vec{x},\vec{\theta})}{\partial\theta_1}\dfrac{\partial\mathcal{L}(\vec{x},\vec{\theta})}{\partial\theta_2} & \left(\dfrac{\partial\mathcal{L}(\vec{x},\vec{\theta})}{\partial\theta_2}\right)^2\end{bmatrix}\right) \tag{5.7.1348}$$

We recognize here the Fisher information matrix! Therefore:

$$\Sigma(\vec{\hat{\mathcal{S}}})=\begin{bmatrix}\sigma_{\hat{\mathcal{S}}_1}^2 & \sigma_{\hat{\mathcal{S}}_1}\sigma_{\hat{\mathcal{S}}_2}^2\\ \sigma_{\hat{\mathcal{S}}_1}\sigma_{\hat{\mathcal{S}}_2} & \sigma_{\hat{\mathcal{S}}_2}^2\end{bmatrix}=\mathcal{I}(\vec{\theta}) \tag{5.7.1349}$$

And so finally:

$$\sqrt{n}\vec{\hat{\mathcal{S}}}(\theta_0)=\mathcal{N}(0,\mathcal{I}(\vec{\theta})) \tag{5.7.1350}$$

5.7.8 Finite Population Correction Factor

Now we prove another result which we will be required in some statistical tests that we will see later.

Suppose we have a population of N individuals that we represent by the set $\{1,2,\ldots,N\}$ and a random variable X which is an application of $\{1,2,\ldots,N\}$ in $\mathbb{R}$. We denote by $x_i=X(i)$. The mean of X is thus given by:

$$\mu=\frac{1}{N}\sum_{i=1}^{N}x_i \tag{5.7.1351}$$

Remember the variance of X is by definition:

$$\sigma=\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 \tag{5.7.1352}$$

Now we consider the set E of samples of size n taken in $\{1,2,\ldots,N\}$ with $0<n<N$. Each individual has a probability of being drawn equal to:

$$\frac{1}{N}\frac{1}{N-1}\cdot\ldots\cdot\frac{1}{N-(n-1)}=\frac{(N-n)!}{N!} \tag{5.7.1353}$$

We are interested in the random variable $\bar{X}$ defined on E and that equal to the sample mean. More specifically:

$$\bar{X}(i_1,\ldots,i_n)=\frac{1}{n}\sum_{k=1}^{n}x_{i_k} \qquad (5.7.1354)$$

To calculate the variance $\text{V}(\bar{X})$, we will $\bar{X}$ express as a sum of random variables. Indeed, if we define the variables X_k with $k=1\ldots N$ by:

$$X_k(i_1,\ldots,i_n)=\begin{cases}x_k \text{ if } k\in\{i_1,\ldots,i_n\}\\ 0 \text{ otherwise}\end{cases} \qquad (5.7.1355)$$

We have naturally by the previous definition (see with caution the sum limits!):

$$\bar{X}=\frac{1}{n}\sum_{k=1}^{N}X_k \qquad (5.7.1356)$$

and thus we get:

$$\text{V}(\bar{X})=\frac{\sigma}{n}=\frac{1}{n}\left(\sum_{k=1}^{N}\text{V}(X_k)+\sum_{i\neq j}^{N}\text{cov}(X_i,X_j)\right) \qquad (5.7.1357)$$

The random variables X_k are not independent in pairs, in fact as we shall see, their covariances are not zero if N is finite. Otherwise (zero covariance), we find a result already proved earlier:

$$\text{V}(\bar{X})=\sigma_{\bar{X}}=\frac{\sigma^2}{n} \qquad (5.7.1358)$$

So we need to calculate the variances $\text{V}(X)$ and covariances $\text{cov}(X_i,X_j)$.

For this purpose we will use the Huygens relation and we will start by calculating the mean $\text{E}(X_k)$:

$$\text{E}(X_k)=P(X_k=x_k)x_k \qquad (5.7.1359)$$

But $P(X_k=x_k)$ is the probability that a sample contains k. This probability is obviously equal to n/N and therefore:

$$\text{E}(X_k)=P(X_k=x_k)x_k=\frac{n}{N}x_k \qquad (5.7.1360)$$

Similarly we obtain:

$$\text{E}(X_k^2)=P(X_k=x_k)x_k=\frac{n}{N}x_k^2 \qquad (5.7.1361)$$

We can therefore calculate the variance $\text{V}(X_k)$:

$$\text{V}(X_k)=\text{E}(X_k^2)-\text{E}(X_k)^2=\frac{n}{N}x_k^2-\left(\frac{n}{N}x_k\right)^2=\frac{n(N-n)}{N^2}x_k^2 \qquad (5.7.1362)$$

To calculate the covariances we need now to calculate the means $\text{E}(X_iX_j)$:

$$\text{E}(X_iX_j)=P(X_i=x_i,X_j=x_j)x_ix_j \qquad (5.7.1363)$$

But $P(X_i = x_i, X_j = x_j)$ is the probability that a sample contains i and j. This probability is obviously given by:

$$\frac{n}{N}\frac{n-1}{N-1} \tag{5.7.1364}$$

and therefore:

$$\mathrm{E}(X_iX_j) = \frac{n(n-1)}{N(N-1)}x_ix_j \tag{5.7.1365}$$

We can now compute the covariance:

$$\mathrm{cov}(X_i,X_j) = \mathrm{E}(X_iX_j) - \mathrm{E}(X_i)\mathrm{E}(X_j) = \frac{n(n-1)}{N(N-1)}x_ix_j - \frac{n^2}{N^2}x_ix_j = -\frac{n(N-n)}{N^2(N-1)}x_ix_j \tag{5.7.1366}$$

We are now able to calculate $\mathrm{V}(\bar{X})$:

$$\mathrm{V}(\bar{X}) = \frac{1}{n}\left(\sum_{k=1}^{N}\mathrm{V}(X_k) + \sum_{i\neq j}^{N}\mathrm{cov}(X_i,X_j)\right) = \frac{1}{n}\left(\frac{n(N-n)}{N^2}\sum_{k=1}^{N}x_k^2 - \frac{n(N-n)}{N^2(N-1)}\sum_{i\neq j}^{N}x_ix_j\right) \tag{5.7.1367}$$

Using Huygens theorem we get:

$$\mathrm{V}(X_k) = \sigma^2 = \mathrm{E}(X_k^2) - \mathrm{E}(X_k)^2 = \mathrm{E}(X_k^2) - \mu^2 = \mathrm{E}(X_k^2) - \mu^2 \Leftrightarrow \sigma^2 + \mu^2 = \mathrm{E}(X_k^2) \tag{5.7.1368}$$

Using the result proved above and previous relation:

$$\mathrm{E}(X_k^2) = \frac{n}{N}x_k^2 = \sigma^2 + \mu^2 \tag{5.7.1369}$$

Therefore:

$$x_k^2 = \frac{N}{n}\sigma^2 + \frac{N}{n}\mu^2 \Leftrightarrow \sum_{k=1}^{N}x_k^2 = \frac{N^2}{n}\left(\sigma^2+\mu^2\right) \tag{5.7.1370}$$

For the double sum $\sum_{i\neq j}^{N}x_ix_j$, we have:

$$\begin{aligned}\sum_{i\neq j}^{N}x_ix_j &= \sum_{i=1}^{N}x_i(x_1+\ldots+x_{i-1}+x_{i+1}+\ldots+x_N) = \sum_{i=1}^{N}x_i(N\mu - x_i)\\ &= N\mu\sum_{i=1}^{N}x_i - \sum_{i=1}^{N}x_i^2 = N\mu N\mu - \sum_{i=1}^{N}x_i^2 = N^2\mu^2 - \sum_{i=1}^{N}x_i^2\\ &= N^2\mu^2 - \frac{N^2}{n}\left(\sigma^2+\mu^2\right)\end{aligned} \tag{5.7.1371}$$

Therefore:

$$\begin{aligned}
\mathrm{V}(\bar{X}) &= \frac{1}{n}\left[\frac{n(N-n)}{N^2}\frac{N^2}{n}\left(\sigma^2+\mu^2\right)-\frac{n(N-n)}{N^2(N-1)}\left(N^2\mu^2-\frac{N^2}{n}\left(\sigma^2+\mu^2\right)\right)\right]\\
&= \frac{1}{n}\left[(N-n)\left(\sigma^2+\mu^2\right)-\frac{n(N-n)}{N-1}\left(\mu^2-\frac{1}{n}\left(\sigma^2+\mu^2\right)\right)\right]\\
&= \frac{1}{n}\left((N-n)\left(\sigma^2+\mu^2\right)-\frac{n(N-n)}{N-1}\mu^2+\frac{N-n}{N-1}\sigma^2+\frac{N-n}{N-1}\mu^2\right)\\
&= \frac{1}{n}\left((N-n)\sigma^2+\frac{N-n}{N-1}\sigma^2+\mu^2\left(\frac{N-n}{N-1}-\frac{n(N-n)}{N-1}+(N-n)\right)\right)\\
&= \frac{1}{n}\left((N-n)\sigma^2+\frac{N-n}{N-1}\sigma^2+\mu^2\left(\frac{-nN+n^2+N^2-nN}{N-1}\right)\right)\\
&= \frac{1}{n}\left((N-n)\sigma^2+\frac{N-n}{N-1}\sigma^2+\mu^2\frac{(N-n)^2}{N-1}\right)\\
&= \frac{1}{n}\left(\frac{N(N-n)}{N-1}\sigma^2+\mu^2\frac{(N-n)^2}{N-1}\right)\\
&= ..???.. = \sigma^2\frac{N-n}{n(N-1)}
\end{aligned}\tag{5.7.1372}$$

Thus:

$$\boxed{\sigma_{\bar{X}}=\sqrt{\mathrm{V}(\bar{X})}=\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}\tag{5.7.1373}$$

The famous factor:

$$\text{fpc}=\sqrt{\frac{N-n}{N-1}}\leq 1\tag{5.7.1374}$$

that we have already encountered during our study the hypergeometric distribution is named "finite correction factor (on finite population)" and has the effect of reducing the standard error especially as n is large. This factor is less than 1, so the variance of the hypergeometric is always less than that of the corresponding binomial. This makes sense because the draws are made without replacement in the hypergeometric distribution.

If the sample is relatively large (more than 5% of the population), it is sometimes customary to use this correction factor.

In the science of survey, we can conclude that for the "random sample without replacement (RSWOR)" and for "random sample with replacement (RSWR)" the average is most of times given by:

$$\hat{\mu}_X=\frac{1}{n}\sum_{i=1}^{n}x_i\tag{5.7.1375}$$

but for the standard deviation:

$$\hat{\sigma}_{\text{RSWR}}=\sqrt{\mathrm{V}(\bar{X}_n)}=\frac{\sigma_N^2}{\sqrt{n}}\qquad\hat{\sigma}_{\text{RSWOR}}=\sqrt{\mathrm{V}(\bar{X}_n)}=\frac{\sigma_N^2}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\tag{5.7.1376}$$

where the correction factor is sometimes approximated by:

$$\text{fpc} \cong \sqrt{\frac{N-n}{N}} \tag{5.7.1377}$$

5.7.9 Confidence Intervals (inference)

Until now we have always determined the likelihood estimators or simple estimators (variance, standard deviation) from theoretical statistics distributions or measured on an entire population of data. But now that we have covered descriptive statistics, probability and sampling, we are ready to make "frequentist inference" and "bayesian inference" from our sample data to our population of interest.

In statistical frequentist inference, we take what we know from the sample, apply the underlying theory of sampling (central limit theorem) to make statements about our population of interest. We make estimates about the population through the use of the sample data. Estimates can either be point estimates or interval estimates. In statistical Bayesian inference, we take the likelihood knowledge we have on the data and the prior on some parameters to infer a posterior distribution having new "hyperparameters".

quote 5.7:848.#47

> *Being a statistician means never having to say you are certain!*
>
> ?

5.7.9.1 Frequentist Inference

Definition (#5.7.135): A "confidence interval" (C.I.) in frequentist inference is a pair of numbers (in the univariate case) that defines (a priori) the range of possible values with a certain cumulative probability of the distribution of a given estimator from a sample of an experience (the range of the statistical indicator being usually calculated using real measured parameters). In other words, the C.I. is a collection of intervals with a given percentage (probability) of them containing the true parameter!

Formally, let y be a random sample from a probability distributions governed by a parameter of interest θ and quantities not of interest φ. A confidence interval for θ is defined as an interval $[u(y), v(y)]$ determined by the random variables $u(y)$ and $v(y)$ with the property:

$$P(u(y) < \theta < v(y) \mid \theta, \varphi) = \gamma \qquad \forall(\theta, \varphi) \tag{5.7.1378}$$

where $\gamma = 1 - \alpha$ is the confidence level.

So we must say: The a priori probability that our method produces confidence intervals that will contain θ is $1 - \alpha$.

But we should never say: The probability that this confidence interval we produced does contain θ is $1 - \alpha$.

Remark

The confidence interval is often confused with the "Bayesian credibility interval" that is an interval within which an unobserved parameter value falls with a particular probability. It is an interval in the domain of a posterior probability distribution or a predictive distribution (the generalisation to multivariate problems is the "credible region").

Textbook authors and proponents of confidence intervals bridge the gap seamlessly by claiming that confidence intervals have three desirable properties: first, that the confidence coefficient can be read as a measure of the uncertainty one should have that the interval contains the parameter; second, that the CI width is a measure of estimation uncertainty; and third, that the interval contains the "likely" or "reasonable" values for the parameter. These all involve reasoning about the parameter from the observed data: that is, they are "post-data" inferences.

These interpretations of confidence intervals are not correct. We name the mistake these authors have made the "Fundamental Confidence Fallacy" (FCF) because it seems to flow naturally from the definition of the confidence interval (see [436]).

The reasoning behind the Fundamental Confidence Fallacy seems plausible: on a given sample, we could get any one of the possible confidence intervals. If 95% of the possible confidence intervals contain the true value, without any other information it seems reasonable to say that we have 95% certainty that we obtained one of the confidence intervals that contain the true value. This interpretation is suggested by the name "confidence interval" itself: the word "confident", in lay use, is closely related to concepts of plausibility and belief. The name "confidence interval" — rather than, for instance, the more accurate "coverage procedure" — encourages the Fundamental Confidence Fallacy.

So if the confidence intervals of two parameters overlap you can't conclude that these parameters are not significantly different from each other! And that's because if you construct a hundred 95% confidence intervals, then about 95 of these intervals will contain the population value (you must always keep in mind that you work with statistical objects!!!!). So the specific values of a calculated confidence intervals should rigorously not be directly interpreted!

In short, this is what you need to remember when comparing two confidence intervals:

1. When two confidence intervals of a given parameter do not overlap, the difference between the two parameters will be significant.
2. When two confidence intervals of a given parameter do overlap, the difference between the two parameters can be significant or non-significant.

Keep in mind for example when the parameter is the mean, that standard errors of means are not the same as the standard errors of differences of means!

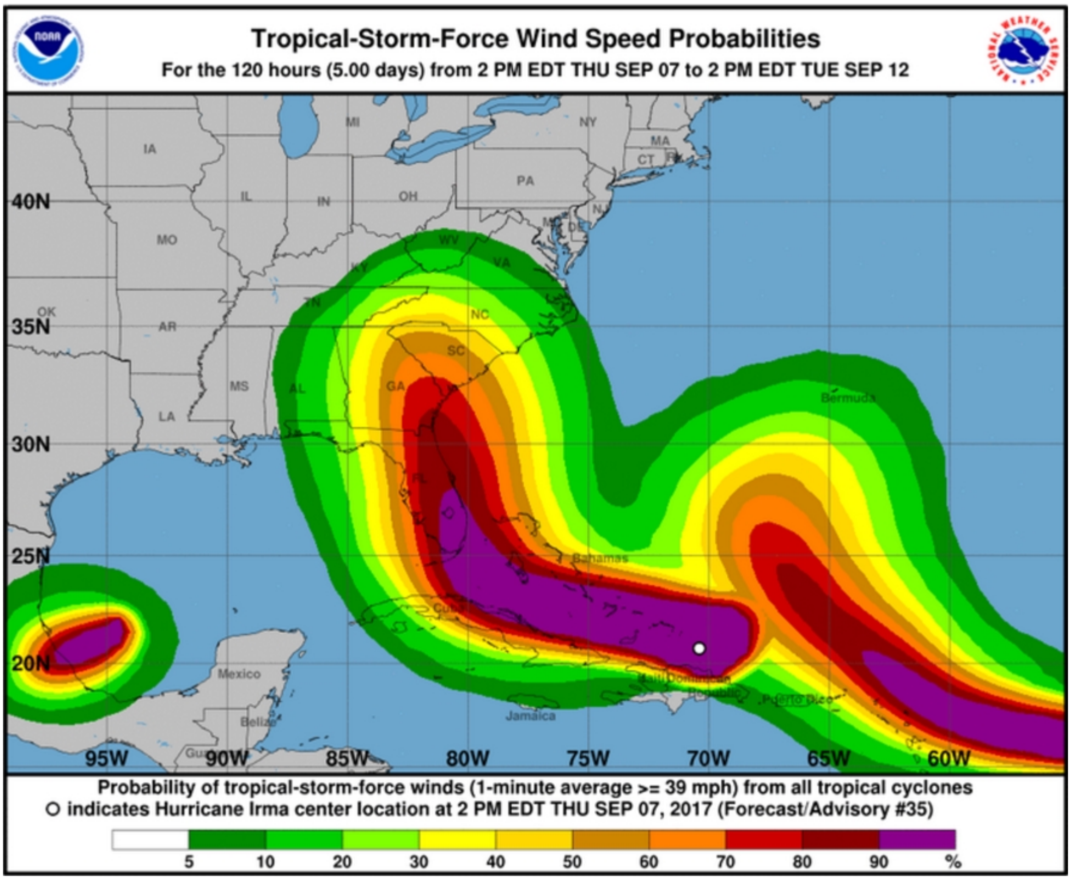

Tropical-Storm-Force Wind Speed Probabilities
For the 120 hours (5.00 days) from 2 PM EDT THU SEP 07 to 2 PM EDT TUE SEP 12
40N
35N
30N
25N
20N
95W
90W
85W
80W
75W
70W
65W
60W
IA
MI
NY
MA
CT
PA
IL
IN
OH
MO
WV
VA
KY
OK
AR
TN
NC
SC
MS
AL
GA
LA
FL
Bermuda
Bahamas
Cuba
Haiti Dominican
Puerto Rico
Jamaica
Mexico
Belize
Probability of tropical-storm-force winds (1-minute average >= 39 mph) from all tropical cyclones
O indicates Hurricane Irma center location at 2 PM EDT THU SEP 07, 2017 (Forecast/Advisory #35)
5
10
20
30
40
50
60
70
80
90
%

Figure 5.158: Example of confidence intervals that have interested million of peoples not so far ago in the past...

We now turn to the task that consists naturally to ask ourselves what must be the sample sizes of our measured data to have some validity for our estimators or even to which confidence interval correspond a given standard deviation or quantile in a Normal centered reduced distribution (for large samples), in a chi-square distribution, Student distribution or Fisher distribution (we will see the last two cases of small sample sizes further below during our study of ANalysis Of VAriance, i.e. ANOVA) when the mean or variance are known or unknown respectively on all or part of the given population.

It is important to know that these confidence intervals often use the central limit theorem that will be proved further below (to avoid any possible frustration) and the developments that we will do now are also useful in the field of (a posteriori) Hypotheses Tests that have a major role in statistics and therefore indirectly in all fields of science!!!

Finally, it could be useful to indicate that a large numbers of organizations (private or institutional) make false statistics because the assumptions and conditions of use of these confidence intervals (verbatim hypotheses tests) are not rigorously verified or simply omitted or worse, the whole base (measurements) is not collected in the rules of art (reliability of the data collection and reproducibility protocols not validated by scientific peer).

The reader must also know that we have put many other confidence interval techniques detailed proofs related for example for regression techniques in the section of Theoretical Computing page 6316.

Remark

The practitioner should be very careful about the calculation of confidence intervals and the use of hypothesis testing in practice. This is why, to avoid trivial usage error or interpretation error, it is important to refer to the following international standards eg: ISO 2602:1980 (*Statistical interpretation of test results - Estimation of the mean - Confidence interval*), ISO 2854:1976 (*Statistical interpretation of data - Techniques of estimation and tests relating to means and variances*), ISO 3301:1975 (*Statistical interpretation of data - Comparison of two means in the case of paired observations*), ISO 3494:1976 (*Statistical interpretation of data - Effectiveness of tests relating to means and variances*), ISO 5479:1997 (*Statistical interpretation of data - Tests for departure from the Normal distribution*), ISO 10725:2000 + ISO 11648-1:2003 + ISO 11648-2:2001 (*Sampling plans and procedures for acceptance for control of bulk materials*), ISO 11453:1996 (*Statistical interpretation of data - Tests and confidence intervals relating to proportions*), ISO 16269-4:2010 (*Statistical interpretation of data - Detection and treatment of outliers*), ISO 16269-6:2005 (*Statistical interpretation of data - Determination of statistical tolerance intervals*), ISO 16269-8:2004 (*Statistical Interpretation of data - Determination of prediction intervals*), ISO / TR 18532:2009 (*Guidelines for the application of statistical quality and industrial standards*).

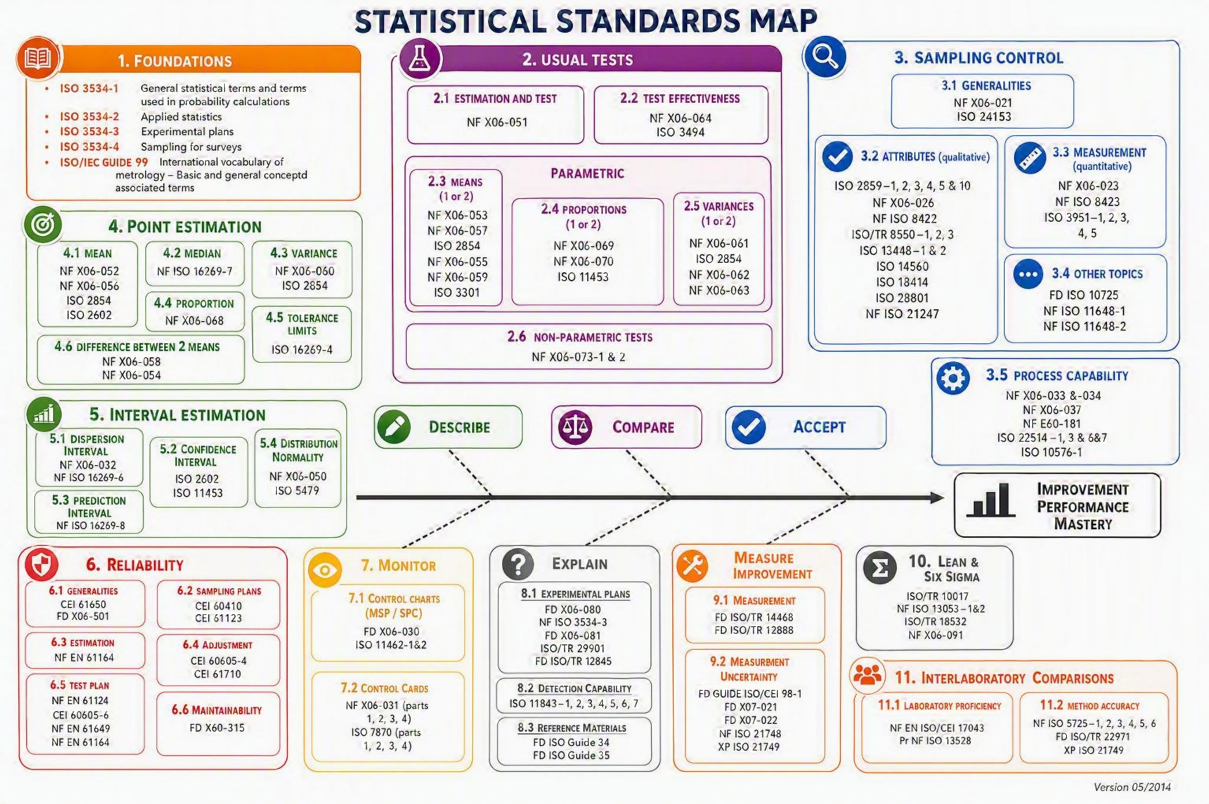

STATISTICAL STANDARDS MAP
1. FOUNDATIONS
ISO 3534-1 General statistical terms and terms used in probability calculations
ISO 3534-2 Applied statistics
ISO 3534-3 Experimental plans
ISO 3534-4 Sampling for surveys
ISO/IEC GUIDE 99 International vocabulary of metrology – Basic and general conceptd associated terms
2. USUAL TESTS
2.1 ESTIMATION AND TEST
NF X06-051
2.2 TEST EFFECTIVENESS
NF X06-064
ISO 3494
PARAMETRIC
2.3 MEANS (1 or 2)
NF X06-053
NF X06-057
ISO 2854
NF X06-055
NF X06-059
ISO 3301
2.4 PROPORTIONS (1 or 2)
NF X06-069
NF X06-070
ISO 11453
2.5 VARIANCES (1 or 2)
NF X06-061
ISO 2854
NF X06-062
NF X06-063
2.6 NON-PARAMETRIC TESTS
NF X06-073-1 & 2
3. SAMPLING CONTROL
3.1 GENERALITIES
NF X06-021
ISO 24153
3.2 ATTRIBUTES (qualitative)
ISO 2859–1, 2, 3, 4, 5 & 10
NF X06-026
NF ISO 8422
ISO/TR 8550–1, 2, 3
ISO 13448–1 & 2
ISO 14560
ISO 18414
ISO 28801
NF ISO 21247
3.3 MEASUREMENT (quantitative)
NF X06-023
NF ISO 8423
ISO 3951–1, 2, 3, 4, 5
3.4 OTHER TOPICS
FD ISO 10725
NF ISO 11648-1
NF ISO 11648-2
3.5 PROCESS CAPABILITY
NF X06-033 &-034
NF X06-037
NF E60-181
ISO 22514 –1, 3 & 6&7
ISO 10576-1
4. POINT ESTIMATION
4.1 MEAN
NF X06-052
NF X06-056
ISO 2854
ISO 2602
4.2 MEDIAN
NF ISO 16269-7
4.3 VARIANCE
NF X06-060
ISO 2854
4.4 PROPORTION
NF X06-068
4.5 TOLERANCE LIMITS
ISO 16269-4
4.6 DIFFERENCE BETWEEN 2 MEANS
NF X06-058
NF X06-054
5. INTERVAL ESTIMATION
5.1 DISPERSION INTERVAL
NF X06-032
NF ISO 16269-6
5.2 CONFIDENCE INTERVAL
ISO 2602
ISO 11453
5.3 PREDICTION INTERVAL
NF ISO 16269-8
5.4 DISTRIBUTION NORMALITY
NF X06-050
ISO 5479
DESCRIBE
COMPARE
ACCEPT
IMPROVEMENT PERFORMANCE MASTERY
6. RELIABILITY
6.1 GENERALITIES
CEI 61650
FD X06-501
6.2 SAMPLING PLANS
CEI 60410
CEI 61123
6.3 ESTIMATION
NF EN 61164
6.4 ADJUSTMENT
CEI 60605-4
CEI 61710
6.5 TEST PLAN
NF EN 61124
CEI 60605-6
NF EN 61649
NF EN 61164
6.6 MAINTAINABILITY
FD X60-315
7. MONITOR
7.1 CONTROL CHARTS (MSP / SPC)
FD X06-030
ISO 11462-1&2
7.2 CONTROL CARDS
NF X06-031 (parts 1, 2, 3, 4)
ISO 7870 (parts 1, 2, 3, 4)
EXPLAIN
8.1 EXPERIMENTAL PLANS
FD X06-080
NF ISO 3534-3
FD X06-081
ISO/TR 29901
FD ISO/TR 12845
8.2 DETECTION CAPABILITY
ISO 11843–1, 2, 3, 4, 5, 6, 7
8.3 REFERENCE MATERIALS
FD ISO Guide 34
FD ISO Guide 35
MEASURE IMPROVEMENT
9.1 MEASUREMENT
FD ISO/TR 14468
FD ISO/TR 12888
9.2 MEASURBMENT UNCERTAINTY
FD GUIDE ISO/CEI 98-1
FD X07-021
FD X07-022
NF ISO 21748
XP ISO 21749
10. LEAN & SIX SIGMA
ISO/TR 10017
NF ISO 13053–1&2
ISO/TR 18532
NF X06-091
11. INTERLABORATORY COMPARISONS
11.1 LABORATORY PROFICIENCY
NF EN ISO/CEI 17043
Pr NF ISO 13528
11.2 METHOD ACCURACY
NF ISO 5725–1, 2, 3, 4, 5, 6
FD ISO/TR 22971
XP ISO 21749
Version 05/2014

Figure 5.159: **Cartography of statistics standards**

Caution! The reader must not expect to find in this textbook a detailed derivation of the almost 500 parametric, non-parametric and bayesian hypothetical tests and their respective confidence intervals! It would take tenth of thousand of pages to do such a work. However we provide in the section of Numerical Methods at page 6775, an exhaustive list of all these kind of tests and even an illustration (see page 6770) categorizing them by family (there may be some categorization errors!).

5.7.9.1.1 C.I. on the Mean with known theoretical Variance

Let's start with the simplest and most common case that is the determination of the number of individuals to have some confidence in the average of the measurements of a random variable assumed to follow a Normal distribution.

First let us recall that we showed at the beginning of this chapter that the standard error (standard deviation of the mean) was under the assumptions of independent and identically distributed variables (i.i.d.):

$$\sigma_{\bar{X}}=\frac{\sigma}{\sqrt{n}} \tag{5.7.1379}$$

Now, before we go any further, consider X as a random variable following a Normal distribution with mean μ and standard deviation σ. We would like that the random variable has for example 95% cumulative probability of being in a given bounded symmetric interval. Which is therefore expressed as follows:

$$P(\mu-\delta\leqslant X\leqslant\mu+\delta)=0.95 \tag{5.7.1380}$$

Remark

Therefore with a confidence interval of 95% you will be right a posteriori 19 times out of 20, or any other level of confidence or risk level α (1-confidence level, 5%) that you will be set up in advance (many practitioners recommend in this beginning of the 121st century (holocene calendar) to take a $\alpha=1\%$). On average, your conclusions will therefore be good, but we can never know whether a particular decision is good! If the risk level is very low but the event still occurs, specialists then speak about a "large deviation" or a "black swan". Management of outliers is addressed in ISO 16269-4:2010 *Detection and treatment of outliers* that any engineer doing business statistics has to follow.

By centering and reducing the random variable:

$$P\left(-\frac{\delta}{\sigma}\leqslant\frac{X-\mu}{\sigma}\leqslant\frac{\delta}{\sigma}\right)=0.95 \tag{5.7.1381}$$

Let us now write Y the reduced centered variable:

$$P\left(-\frac{\delta}{\sigma}\leqslant Y\right)-P\left(Y\geqslant\frac{\delta}{\sigma}\right)=0.95 \tag{5.7.1382}$$

Since the Normal centered reduced distribution is symmetric:

$$1-2P\left(Y\geqslant\frac{\delta}{\sigma}\right)=0.95 \tag{5.7.1383}$$

Therefore:

$$P\left(Y\geqslant\frac{\delta}{\sigma}\right)=0.025 \tag{5.7.1384}$$

From there reading statistical tables of the standard Normal distribution (or by using a simple spreadsheet software), we have to satisfy the equality that:

$$\frac{\delta}{\sigma}\cong 1.96 \quad\Leftrightarrow\quad \delta\cong 1.96\sigma \tag{5.7.1385}$$

Which can easily be obtained with Microsoft® Excel® 11.8346 by using the function:

```
=-NORMSINV((1-0.95)/2)
```

As noted in the traditional way in the general case other than the 95% one (Z is the random variable corresponding to the half quantile of the chosen threshold of the standard Normal distribution):

$$\delta\cong Z\sigma \tag{5.7.1386}$$

Now, consider that the variable X on which we wish to make statistical inference is the average (and we show later that it follows a Normal distribution centered reduced distribution). Therefore:

$$\delta\cong Z\sigma=Z\frac{\sigma}{\sqrt{n}} \tag{5.7.1387}$$

Then we get:

$$n\cong\frac{Z^2\sigma^2}{\delta^2} \tag{5.7.1388}$$

from which we obviously take (normally...) the upper integer value...

The latter notation is usually written in the following way highlighting better the width of the confidence interval of an underlying threshold level[43] α:

$$\boxed{n\cong\left\lceil\left(\frac{Z_{\alpha/2}\sigma}{\delta}\right)^2\right\rceil} \tag{5.7.1389}$$

Relation named "sample size estimation by Normal distribution".

---

[43] Don't forget that in this textbook we write $Z_{\alpha/2}$ where most authors write $Z_{1-\alpha/2}$

Remark

Calculating an appropriate sample size for typically a clinical trial is like walking on a knife edge for the statistician: too low a sample size and the study risks being unable to detect a difference between treatment and the control group, too large a sample size and we will have a "overpowering" of the study that can lead to several issues theoretical issues (p-hacking) that might cause the study to not be as effective or interpretable, and to potentially become unethical.

Thus, we now know the number of individuals we must have to ensure to get a given precision interval δ (margin of error) around the mean and that for a given percentage measures are in this range and assuming the theoretical standard deviation σ is known (or imposed) in advance (typically used in quality engineering or surveys).

In the case of polls / surveys where the sample population is large enough, then we can introduce the population correction factor fpc that we proved earlier above (we speak then of SPWR for "survey plan without replacement"). Then it comes by taking the approximation of the fpc:

$$\begin{aligned} n &\cong \left(\frac{Z_{\alpha/2}\sqrt{1-\dfrac{n}{N}}\sigma}{\delta}\right)^2=\frac{Z_{\alpha/2}^2\left(1-\dfrac{n}{N}\right)\sigma^2}{\delta^2}\\ \Rightarrow n&\left(1+\frac{Z_{\alpha/2}^2\sigma^2}{N}\right)\cong\frac{Z_{\alpha/2}^2\sigma^2}{\delta^2}\\ \Rightarrow n&\cong\frac{Z_{\alpha/2}^2\sigma^2}{\delta^2\left(1+\dfrac{Z_{\alpha/2}^2\sigma^2}{N}\right)}=\frac{NZ_{\alpha/2}^2\sigma^2}{\delta^2 N+Z_{\alpha/2}^2\sigma^2}\end{aligned} \quad (5.7.1390)$$

In other words, we can calculate the number n of individuals to measure to ensure a given confidence interval (associated to the quantile Z) of the measured average assuming the theoretical standard deviation known (or imposed) and wishing a precision δ in absolute value of the mean.

However ... in reality, the variable Z comes from the central limit theorem (see below) that gives for a large sample size (approximately):

$$\boxed{Z=Z_n=\frac{M_n-\mu}{\sigma/\sqrt{n}}=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}} \quad (5.7.1391)$$

Rearranging we get then:

$$\mu=\bar{X}-Z\frac{\sigma}{\sqrt{n}} \quad (5.7.1392)$$

and as Z can be negative or positive then it is more logic to write this as:

$$\mu=\bar{X}\pm Z\frac{\sigma}{\sqrt{n}} \quad (5.7.1393)$$

Thus:

$$\bar{X}-Z\frac{\sigma}{\sqrt{n}}\leqslant\mu\leqslant\bar{X}+Z\frac{\sigma}{\sqrt{n}} \quad (5.7.1394)$$

That engineers sometimes write:

$$\text{LCL} \leqslant \mu \leqslant \text{UCL} \tag{5.7.1395}$$

where LCL is the lower confidence limit and UCL the upper confidence limit. This is the Six Sigma terminology (see section Industrial Engineering 8674).

And we have seen earlier that for a confidence interval of 95% we have $Z = 1.96$. And since the Normal distribution is symmetric:

$$95\% = 1 - 5\% = 1 - \alpha = 1 - 2 \cdot 2.5\% = 1 - 2 \cdot 0.025 \tag{5.7.1396}$$

Thus we finally write[44] the "one sample Z test":

$$\boxed{\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}} \tag{5.7.1397}$$

where we define for all tests having the same structure, the "margin error" by:

$$\text{ME} = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{5.7.1398}$$

If the elements sampled are not independent then we have to write obviously:

$$\boxed{\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\text{fpc} \leqslant \mu \leqslant \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\text{fpc}} \tag{5.7.1399}$$

where for recall as proven earlier:

$$\text{fpc} = \sqrt{\frac{N-n}{N-1}} \tag{5.7.1400}$$

is the finite population correction factor.

As we have already mentioned, and we will prove a little further, the arithmetic reduced centered mean of a series of independent and identically distributed random variables with finite variance asymptotically follows a standard Normal distribution, this is why the confidence interval above is very general! This is why we sometimes speak of "asymptotic confidence interval of the mean".

These intervals obviously have for origin the fact that we work very often in statistics with samples and not the entire available population. The selected sampling thus affects the value of the punctual estimator. We then speak of "sampling fluctuation".

In the particular case of an IC (confidence interval) at 95%, the last relation will be written:

$$\bar{X} - Z_{0.025}\frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \bar{X} + Z_{0.025}\frac{\sigma}{\sqrt{n}} \tag{5.7.1401}$$

Sometimes we find the prior-previous inequality in the following equivalent notation:

$$\bar{X} - Z_{\alpha/2}\sigma_{\bar{X}} \leqslant \mu \leqslant \bar{X} + Z_{\alpha/2}\sigma_{\bar{X}} \tag{5.7.1402}$$

---

[44] Don't forget that in this textbook we write $Z_{\alpha/2}$ where most authors write $Z_{1-\alpha/2}$

or more rarely with the following general notation (for all intervals):

$$\bar{X}-\text{ME}\leqslant\mu\leqslant\bar{X}+\text{ME}\qquad(5.7.1403)$$

where ME stands for "margin of error".

We are thus now able to estimate population sizes needed to obtain a certain level of confidence α in an outcome or to estimate the confidence interval in which is the theoretical mean knowing the experimental (empirical) average and the estimator maximum likelihood of the standard deviation. We can of course therefore also determine the a posteriori probability that the mean is outside a given range ... (one as the other being widely used in the industry).

Finally, note that from the previous result, we deduce immediately the stability property of the Normal distribution (shown above) the following test that we find in many statistical softwares:

$$Z^{\text{calc.}}=\frac{(\bar{X}_2-\bar{X}_1)-(\mu_2-\mu_1)}{\sqrt{\dfrac{\sigma_1^2}{n}+\dfrac{\sigma_2^2}{n}}}\qquad(5.7.1404)$$

named "bilateral Z test on the difference of two means" or also sometimes named "two sample Z test" and with the corresponding confidence interval:

$$\boxed{(\bar{X}_2-\bar{X}_1)-Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{n}}\leq\mu_2-\mu_1\leq(\bar{X}_2-\bar{X}_1)+Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{n}}}\qquad(5.7.1405)$$

And this is not because two means are significantly different that their confidence intervals do not overlap!!!! As shows the graph below obtained with Minitab 16 software where the test-Z of the difference is significant at 95%:

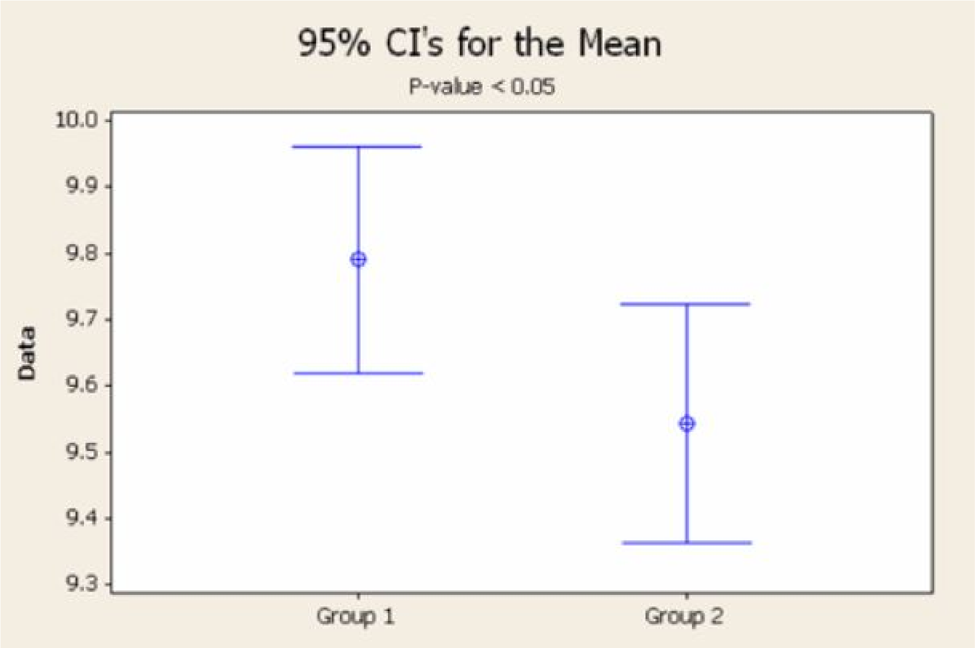


Figure 5.160: **Line plot illustration of the overlay of two 95% confidence intervals (source: Minitab 15)**

while their mean is significantly different to a confidence level of 95%.

Also notice that many people (especially non-scientists) don't understand why a "practical mean difference" is not the same as "statistical mean difference". But the reason is obvious! The first doesn't involve the variance and the second one do! Hence the fact that the second one is more rigorous!!!

Remark

The size of the parent population for the relations developed above does not come into consideration in the calculations of confidence intervals or even not in the sample size, and because it is considered as infinite. So be careful not sometimes not to have sample sizes that are larger than the actual parent population...

5.7.9.1.2 C.I. on the Variance with known theoretical Mean

Let's start by demonstrating a fundamental property of the Chi-square distribution:

If a random variable X follows a Normal centered reduced distribution $X = \mathcal{N}(0,1)$ then its square follows a chi-square distribution of 1 degree of freedom:

$$X^2 = \chi^2(1) \tag{5.7.1406}$$

This result is sometimes named a "Wald statistics" and any statistical test using it directly (we should better speak about a "test family") can be designated under the name "Wald's test" (for a concrete example see further below the Cochran-Mantel-Haenszel test page 1283).

Proof 5.47.9. To prove this property, it suffices to calculate the density of the random X^2 variable with $X = \mathcal{N}(0,1)$. However, if $X = \mathcal{N}(0,1)$ and if we set $Y = X^2$, then for all $y \geq 0$ we get:

$$P(Y \leqslant y) = P(X^2 \leqslant y) = P(-\sqrt{y} \leqslant Y \leqslant \sqrt{y}) \tag{5.7.1407}$$

Since the standard Normal distribution is symmetric about 0 for the random variable X, we can write:

$$P(Y \leqslant y) = 2P(0 \leqslant X \leqslant \sqrt{y}) \tag{5.7.1408}$$

Denoting by Φ the cumulative distribution function of the standard Normal distribution, we have:

$$P(Y \leqslant y) = 2\Phi(\sqrt{y}) - 2 \cdot 0.5 = 2\Phi(\sqrt{y}) - 1 \tag{5.7.1409}$$

and as:

$$\Phi(0) = P(\mathcal{N}(0,1) \leq 0) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{0} e^{-k/2}\mathrm{d}k = 0.5 \tag{5.7.1410}$$

therefore:

$$P(Y \leq y) = 2\Phi(\sqrt{y}) - 2 \cdot 0.5 = 2\Phi(\sqrt{y}) - 1 \tag{5.7.1411}$$

The cumulative distribution function of the random variable $Y = X^2$ is thus given by:

$$P(Y \leq y) = 2\Phi(\sqrt{y}) - 1 \tag{5.7.1412}$$

if y is greater than or equal to zero, null if y is less than zero. We will denote this cumulative distribution $f_Y(y)$ for the further calculations.

Since the density distribution function is the derivative of the cumulative distribution function and X follows a Normal centered reduced distribution so we reduced for the random variable X:

$$P(X = x) = \Phi'(x) = \frac{\mathrm{d}}{\mathrm{d}x}\Phi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2} \tag{5.7.1413}$$

and then it follows for the probability distribution of Y (which is the square of X for reminder!):

$$\frac{\mathrm{d}}{\mathrm{d}y}f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}(2\Phi(\sqrt{y}) - 1) = 2\Phi'(\sqrt{y})\frac{\mathrm{d}}{\mathrm{d}y}\sqrt{y} = \frac{1}{\sqrt{y}}\Phi'(\sqrt{y}) = \frac{1}{\sqrt{y}}\frac{1}{\sqrt{2\pi}}e^{-y/2}\frac{1}{\sqrt{2\pi y}}e^{-y/2} \tag{5.7.1414}$$

this last expression corresponds is exactly the relation we obtained during our study of the chi-square distribution imposing a degree of freedom equal to the unit.

The theorem is therefore proved, that is if X follows a Normal centered reduced distribution while its square follows a Chi-square distribution of 1 degree of freedom as:

$$\boxed{\left(X_{\mathcal{N}(0,1)}\right)^2 = \chi^2(1)} \tag{5.7.1415}$$

■ Q.E.D.

This type of relation is used mainly in industrial processes and their control (see section Industrial Engineering page 8934).

Now let us open a parenthesis that is quite important in some linear regression software reports and especially in the curvature test for design of experiment (see section Industrial Engineering page 8780). Let us recall that we have:

$$T_n = \frac{\mathcal{N}(0,1)}{\sqrt{\chi_n^2/n}} \tag{5.7.1416}$$

And we have just prove above that:

$$Z^2 = \frac{\chi_1^2}{1} \tag{5.7.1417}$$

Therefore:

$$T_n = \frac{\sqrt{\chi_1^2/1}}{\sqrt{\chi_n^2/n}} = \sqrt{\frac{\chi_2^1/1}{\chi_n^2/n}} \tag{5.7.1418}$$

And as we have also seen that:

$$F_{n,m}=\frac{\sqrt{\chi_n^2/n}}{\sqrt{\chi_m^2/m}} \tag{5.7.1419}$$

it follows that:

$$T_n=\sqrt{F_{1,n}} \tag{5.7.1420}$$

or more commonly in practice:

$$\boxed{T_n^2=F_{1,n}} \tag{5.7.1421}$$

We will now use a result proved during our study of the Gamma distribution. We have effectively seen that the sum of two random variables following a Gamma distribution also follows a Gamma distribution where the two parameters are added:

$$X+Y=\gamma(p+q,\lambda) \tag{5.7.1422}$$

As the Chi-square distribution is a special case of the Gamma distribution, the same result applies.

To be more precise, this is equivalent to say: If $X_1,\ldots,X_k$ are random independent and identically distributed (i.i.d.) variables $\mathcal{N}(0,1)$ then by extension of the above proof where we have shown that:

$$\left(X_{\mathcal{N}(0,1)}\right)^2=\chi^2(1) \tag{5.7.1423}$$

and by the property of linearity of the Gamma distribution, then sum of their squares follows a chi-square distribution of degree k such that:

$$\boxed{\chi^2(k)=X_1^2+X_2^2+\ldots+X_k^2} \tag{5.7.1424}$$

Thus, the distribution of χ^2 of k degrees freedom is the probability distribution of the sum of squares of k Normal centered reduced variables linearly independent of each other. It is in fact the linearity property of the chi-square distribution (implicitly the linearity of the Gamma distribution)!

Now see another significant property of the chi-square distribution: If $X_1,\ldots,X_n$ are independent and identically distributed $\mathcal{N}(\mu,\sigma)$ (thus the same mean and the same standard and following a Normal distribution) random variables and if we write the maximum likelihood estimator variance by:

$$\boxed{S_*^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2} \tag{5.7.1425}$$

then, the ratio of the random variable S_*^2 on the standard deviation assumed to be known for the entire population ("the true standard deviation" or "theoretical standard deviation"!) multiplied by the number of individuals n population follows a chi-square distribution of degree n such that:

$$\chi^2(n)=\chi^2_n=\frac{nS^2_*}{\sigma^2} \tag{5.7.1426}$$

This result is named the "Cochran theorem" or "Fisher-Cochran theorem" (in the particular case of Gaussian samples) and thus gives us a distribution for the empirical standard deviations (whose parent law is a Normal distribution!).

Using the value of the standard deviation proved during our study of chi-square distribution we have:

$$\mathrm{V}\left(\frac{nS^2_*}{\sigma^2}\right)=\mathrm{V}(\chi^2(n))=2n \tag{5.7.1427}$$

But n and σ are imposed and are therefore considered as constants. We have therefore:

$$\frac{n^2}{\sigma^4}\mathrm{V}(S^2_*)=\mathrm{V}(\chi^2(n))=2n\Rightarrow \mathrm{V}(S^2_*)=\sigma_{S^2_*}=\sigma^4\frac{2}{n} \tag{5.7.1428}$$

And therefore we have an expression of the standard deviation of the empirical standard deviation if we know the standard deviation of the population:

$$\sigma_{S_*}=\sigma^2\sqrt{\frac{2}{n}} \tag{5.7.1429}$$

But we have proved during our study of estimators that:

$$\hat{\sigma}=\sqrt{\frac{n-1}{n}}\sigma \tag{5.7.1430}$$

It follows:

$$\hat{\sigma}_{S_*}=\sqrt{\frac{n-1}{n}}\sigma_{S^2_*}=\sqrt{\frac{n-1}{n}}\sigma^2\sqrt{\frac{2}{n}}=\sqrt{\frac{2(n-1)}{n^2}}\sigma^2 \tag{5.7.1431}$$

It follows therefore the sometimes important relation in the practice of the estimator of the standard deviation of ... the standard deviation:

$$\boxed{\hat{\sigma}_{S_*}=\sqrt{\frac{2(n-1)}{n^2}}\sigma^2} \tag{5.7.1432}$$

Recall that the parent population is said to be "infinite" if the sample selection with replacement or if the size N of the parent population is much higher than this of the sample of size n.

Remarks

R1. In laboratories the $X_1,\ldots,X_n$ can be seen as a class of individuals of the same product identically studied by different research teams with instruments of the same precision (standard deviation of the measure identically equal).

R2. S^2_* is the "inter-class variance" also named "explained variance". So it gives a measure of the variance occurring in different laboratories.

What is interesting here is that from the calculation of the chi-square distribution and by knowing n and the standard deviation σ^2 it is possible to estimate the interclass variance (and also interclass standard deviation).

To see that this latter property is a generalization of the basic relation:

$$\chi^2(n)=X_1^1+X_2^2+\ldots+X_n^2 \tag{5.7.1433}$$

it suffices to see that the random variable nS_*^2/σ^2 is a sum of n squares of $\mathcal{N}(0,1)$ independent of each other. Indeed, recall that a centered reduced random variable (see our study of the Normal distribution) is given by:

$$Y=\frac{X-\mu}{\sigma} \tag{5.7.1434}$$

Therefore:

$$\frac{nS_*^2}{\sigma^2}=\frac{n}{\sigma^2}\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2=\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \tag{5.7.1435}$$

However, since the random variables $X_1,\ldots,X_n$ are independent and identically distributed according to a Normal distribution, then the random variables:

$$\frac{X_1-\mu}{\sigma},\ldots,\frac{X_n-\mu}{\sigma} \tag{5.7.1436}$$

are also independent and identically distributed according to a Normal distribution but a centered reduced one.

Since:

$$\frac{nS_*^2}{\sigma^2}=\chi^2(n) \tag{5.7.1437}$$

rearranging we get:

$$\sigma^2=\frac{nS_*^2}{\chi^2(n)} \tag{5.7.1438}$$

So on the population of measurements, the true standard deviation follows the relation given above. It is therefore feasible to make statistical inference on the standard deviation When the theoretical mean is known (...).

Since the chi-square distributions is not symmetric, the only way to make this inference is to use numerical calculations and then we denote the confidence interval at the level of 95% (for example ...) as follows:

$$\frac{nS_*^2}{\chi^2_{2.5\%}(n)}\leqslant\sigma^2\leqslant\frac{nS_*^2}{\chi^2_{97.5\%}(n)} \tag{5.7.1439}$$

Either by writing $95\% = 1-\alpha$:

$$\boxed{\frac{nS_*^2}{\chi^2_{\alpha/2}(n)} \leqslant \sigma^2 \leqslant \frac{nS_*^2}{\chi^2_{1-\alpha/2}(n)}} \tag{5.7.1440}$$

the denominator being obviously the quantile of the chi-squared distribution. This relation is rarely used in practice as the theoretical average (mean) is not known. In order to avoid confusion, the latter relation is often denoted as follows:

$$\frac{nS_*^2}{\chi^2_{\alpha/2,n}} \leqslant \sigma^2 \leqslant \frac{nS_*^2}{\chi^2_{1-\alpha/2,n}} \tag{5.7.1441}$$

Let's see the most common case:

5.7.9.1.3 C.I. on the Variance with known empirical Mean

Let us now make statistical inference when the theoretical average μ of the population (i.e. the mean) is not known. To do this, consider now the sum of:

$$\chi^2(n)=\sum_{i=1}^n\left(\frac{X_i-\mu}{\sigma}\right)^2=\frac{1}{\sigma^2}\sum_{i=1}^n(X_i-\mu)^2=\frac{1}{\sigma^2}\sum_{i=1}^n\left((X_i-\bar{X})+(\bar{X}-\mu)\right)^2 \tag{5.7.1442}$$

where for recall is $\bar{X}$ the empirical average (arithmetic mean) of the sample:

$$\bar{X}=\frac{1}{n}\sum_{i=1}^n X_i \tag{5.7.1443}$$

Continuing the development we have:

$$\chi^2(n)=\frac{1}{\sigma^2}\left[\sum_{i=1}^n(X_i-\bar{X})^2+2(\bar{X}-\mu)\sum_{i=1}^n(X_i-\bar{X})+n(\bar{X}-\mu)^2\right] \tag{5.7.1444}$$

However, we have proved earlier in this chapter that the sum of the deviations from the mean was zero. So:

$$\begin{aligned}\chi^2(n)&=\frac{1}{\sigma^2}\left[\sum_{i=1}^n(X_i-\bar{X})^2+2(\bar{X}-\mu)\cdot 0+n(\bar{X}-\mu)^2\right]\\&=\frac{1}{\sigma^2}\left[\sum_{i=1}^n(X_i-\bar{X})^2+n(\bar{X}-\mu)^2\right]\\&=\sum_{i=1}^n\frac{(X_i-\bar{X})^2}{\sigma^2}+\frac{n(\bar{X}-\mu)^2}{\sigma^2}=\sum_{i=1}^n\frac{(X_i-\bar{X})^2}{\sigma^2}+\left(\frac{\bar{X}-\mu}{\sigma^2/\sqrt{n}}\right)^2\end{aligned} \tag{5.7.1445}$$

and by taking back the unbiased estimator of the Normal distribution (we change notation to respect the traditions and differentiate the empirical average of the theoretical mean):

$$S^2=\frac{1}{n-1}\sum_{i=1}^n(X_i-\bar{X})^2 \tag{5.7.1446}$$

Thus:

$$\chi^2(n)=\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{\sigma^2}+\left(\frac{\bar{X}-\mu}{\sigma^2/\sqrt{n}}\right)^2=\frac{(n-1)S^2}{\sigma^2}+\left(\frac{\bar{X}-\mu}{\sigma^2/\sqrt{n}}\right)^2 \tag{5.7.1447}$$

or with another common notation:

$$\chi^2(n)=\frac{(n-1)S^2}{\sigma^2}+\left(\frac{M_n-\mu}{\sigma^2/\sqrt{n}}\right)^2 \tag{5.7.1448}$$

Since the second term (squared) follows a Normal centered reduced distribution too, then if we remove it we get by the proof made earlier above about the chi-square distribution property, the following result:

$$\chi^2(n-1)=\frac{(n-1)S^2}{\sigma^2} \tag{5.7.1449}$$

These developments allow us this time to also make inferences about the variance of a $\mathcal{N}(0,1)$ distribution when the parameters μ and σ of the parent population are both unknown. It is this result that gives us, for example, the confidence interval:

$$\boxed{\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}\leqslant\sigma^2\leqslant\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}} \tag{5.7.1450}$$

when the theoretical average (mean) μ is unknown. And also to avoid any confusion, it is more usual to write:

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}\leqslant\sigma^2\leqslant\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}} \tag{5.7.1451}$$

In the same way as above, we can calculate the standard deviation of the standard deviation that has a great importance in the practice of finance:

$$\begin{aligned}&\text{V}\left(\frac{(n-1)S^2}{\sigma^2}\right)=\text{V}(\chi^2_{n-1}\\&\frac{(n-1)^2}{\sigma^4}\text{V}(S^2)=2(n-1)\\&\text{V}(S^2)=\frac{2\sigma^4}{n-1}\\&\sigma_{S^2}=\sqrt{\frac{2}{n-1}}\sigma^2\end{aligned} \tag{5.7.1452}$$

5.7.9.1.4 C.I. on the Mean with known empirical Variance

We have proved much higher that the Student distribution came from the following relation:

$$T(k)=\frac{Z}{\sqrt{U/k}} \tag{5.7.1453}$$

if Z and U are independent random variables and if Z follows a Normal centered reduced distribution $\mathcal{N}(0,1)$ and U a chi-square distribution $\chi^2(k)$ as:

$$f_T(t)=\frac{\Gamma\left(\dfrac{n+1}{2}\right)}{\Gamma(n/2)\sqrt{k\pi}}\left(1+\frac{t^2}{n}\right)^{-\frac{k+1}{2}} \tag{5.7.1454}$$

and remember that its density function is symmetrical!

Here is a very important application of the above result:

Suppose that $X_1,\ldots,X_n$ is a random sample of size n from a distribution $\mathcal{N}(\mu,\sigma)$. So we can already write that following developments made above:

$$Z=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \tag{5.7.1455}$$

And for U that follows a $\chi^2(k)$ distribution, then if we ask that $k=n-1$ then according to the results above:

$$U=\frac{(n-1)S^2}{\sigma^2}=\chi^2(n-1) \tag{5.7.1456}$$

We then get after some trivial simplifications:

$$\frac{Z}{\sqrt{U/k}}=\frac{\dfrac{X-\mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)S^2}{\sigma^2}/(n-1)}}=\frac{\dfrac{X-\mu}{1/\sqrt{n}}}{\sqrt{\dfrac{(n-1)S^2}{(n-1)}}}=\frac{\dfrac{X-\mu}{1/\sqrt{n}}}{\sqrt{S^2}}=\frac{\bar{X}-\mu}{S/\sqrt{n}} \tag{5.7.1457}$$

So since:

$$T(k)=\frac{Z}{\sqrt{U/k}} \tag{5.7.1458}$$

follows a Student distribution with parameter k then we get the "independent one-sample T-test" or simply named "one-sample T-test":

$$\boxed{\frac{\bar{X}-\mu}{S/\sqrt{n}}=T(n-1)} \tag{5.7.1459}$$

which also follows a Student distribution of parameter $n-1$ and is widely used in laboratories for calibration testing[45].

---

[45] The engineer must not forget that a calibrated instrument, does not mean that is measures correctly! However it means that it has known values of incertitude on its measurements relatively to a reference or a specification!

This gives us also after rearrangement:

$$\mu = \bar{X} - \frac{S}{\sqrt{n}} T(n-1) \tag{5.7.1460}$$

This allows us to make inference about the mean μ of a Normal distribution with the theoretical standard deviation being unknown (meaning that there is not enough experimental values) but where the unbiased estimator of the standard deviation is known. It is this result that gives us the confidence interval[46]:

$$\boxed{\bar{X} - \frac{S}{\sqrt{n}} T_{\alpha/2}(n-1) \leqslant \mu \leqslant \bar{X} + \frac{S}{\sqrt{n}} T_{\alpha/2}(n-1)} \tag{5.7.1461}$$

where we see the same factors as for the statistical inference on the average (mean) of a (theoretical) random with know standard deviation as the Student distribution is asymptotically equal to the Normal distribution for large values of n. Thus, the previous interval and the following interval:

$$\bar{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \bar{X} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \tag{5.7.1462}$$

gives very similar values (to three decimal places) for values of n at around $10,000$ (in practice we consider that for 100 this is the same...).

We immediately deduce by the stability property of the chi-square distribution (proved above in that this property arises from the Gamma distribution) the following test that we find in many statistical software:

$$(\bar{X}_2 - \bar{X}_1) - T_{\alpha/2}(n-2)\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \leq \mu_2 - \mu_1 \leq (\bar{X}_2 - \bar{X}_1) + T_{\alpha/2}(n-2)\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \tag{5.7.1463}$$

named "bilateral t (Student) test on the difference of two means" or more simply "two sample T-test".

We can of course therefore also determine the probability that the mean is inside or outside a certain range ... (the both case being widely used in industry).

The reader can for fun control with Microsoft® Excel® 11.8346 that for a large number of measurements n, the Student distribution tends to the Normal centered reduced distribution by comparing the values of the two functions below:

```
=T.INV(5%/2,n-1)
=NORM.S.INV(5%/2)
```

---

[46] Don't forget that in this textbook we write $T_{\alpha/2}$ where most authors write $T_{1-\alpha/2}$

Remark

The previous result was obtained by William S. Gosset around 11910 (holocene calendar). Gosset who had studied mathematics and chemistry, worked as a statistician for the Guinness brewery in England. At that time, we knew that if $X_1, \ldots, X_n$ are independent and identically distributed random variables then:

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}=\mathcal{N}(0,1) \quad (5.7.1464)$$

However, in statistical applications we were rather obviously interested in the following quantity:

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \quad (5.7.1465)$$

We then merely assume that this amount followed almost a Normal centered reduced distribution, which was not a bad approximation as can show the image below ($\mathrm{d}f = n-1$):

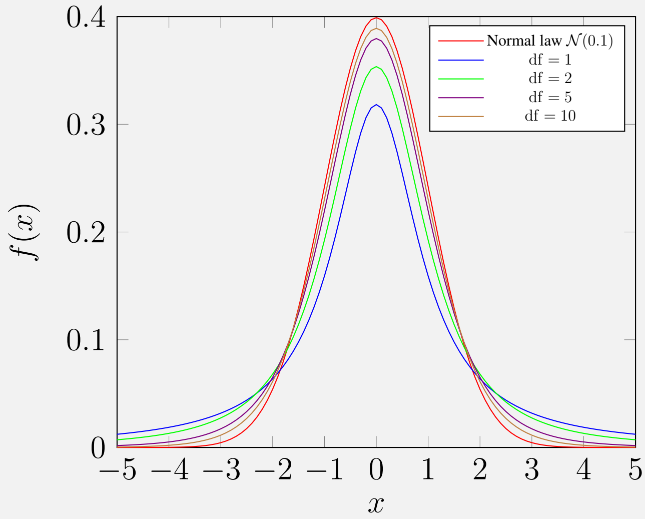


Figure 5.161: **Comparison between the Normal and Student distribution functions**

After numerous simulations, Gosset came to the conclusion that this approximation was valid only when n is large enough (so that gave him the indication that there must be somewhere behind the central limit theorem). He decided to determine the origin of the distribution and after completing a course in statistics with Karl Pearson he obtained his famous result that he published under the pseudonym Student. Thus, is why we name "Student distribution" that law that should have been named the "Gosset distribution".

Finally, note that the Student's T-test is also used to identify whether changes (increasing or vice versa) in the average of two identical populations are statistically significant. That is to say, if the size of two dependent samples is the same then we can develop the following test (we included all different types of writing that can be found in the literature and in many software implementing this test):

$$\begin{aligned}T(n-1)&=\frac{\bar{X}-\mu}{S/\sqrt{n}}=\frac{X_{\bar{D}-\mu_D}}{S_D/\sqrt{n}}=\frac{\bar{D}-\delta_0}{S_D/\sqrt{n}}=\frac{\dfrac{1}{n}\displaystyle\sum_i d_i-\delta_0}{S_D/\sqrt{n}}\\&=\frac{\dfrac{1}{n}\displaystyle\sum_i(X_{i2}-X_{i1})-\delta_0}{S_D/\sqrt{n}}=\frac{\left(\dfrac{1}{n}\displaystyle\sum_i X_{i2}-\dfrac{1}{n}\displaystyle\sum_i X_{i1}\right)-\delta_0}{S_D/\sqrt{n}}=\frac{(\bar{X}_2-\bar{X}_1)-\delta_0}{S_D/\sqrt{n}}\end{aligned}\tag{5.7.1466}$$

With:

$$S_D=\sqrt{\frac{1}{n-1}\sum_i(d_i-\bar{D})}=\sqrt{\mathrm{V}(X_1-X_2)}\tag{5.7.1467}$$

And in the homoscedastic case where $\mathrm{V}(X_1)=\mathrm{V}(X_2)$, remembering that:

$$\begin{aligned}\mathrm{V}(X_1-X_2)&=\mathrm{V}(X_1)+\mathrm{V}(X_2)-2\mathrm{cov}(X_1,X2)=\mathrm{V}(X_1)+\mathrm{V}(X_2)-2\rho\sqrt{\mathrm{V}(X_1)}\sqrt{\mathrm{V}(X_2)}\\&=\sigma^2+\sigma^2-2\rho\sigma\sigma=2\sigma^2(1-\rho)\end{aligned}\tag{5.7.1468}$$

Hence:

$$S_D=\sqrt{\mathrm{V}(X_1-X_2)}=\sqrt{2\sigma^2(1-\rho)}\tag{5.7.1469}$$

The relation:

$$T(n-1)=\frac{(\bar{X}_2-\bar{X}_1)-\delta_0}{S_D/\sqrt{n}}\tag{5.7.1470}$$

is very useful for comparing the same sample twice in different measurement situations (sales before or after a discount on an article for example). This relation is named "T-test (Student) averages two paired samples (or dependent samples)" or more simply "paired sample T-test".

Definition (#5.7.136): We speak of "paired samples" if the sample values are taken 2 times on the same individuals (i.e. the values of the pairs are not independent, unlike two samples taken independently).

Remark

For paired tests the distribution of each sample is completely unimportant! What matters is the distribution of the "pair differences"!!

5.7.9.1.5 C.I. on the Geometric Mean (for known or unknown Variance)

Let's just see how to compute the confidence interval for the geometric mean that is of main importance in finance! Consider we have $X_1, \ldots X_n$ from some distribution with mean μ and variance σ^2.

Consider $Y_i = \ln(X_i)$ (in practice some people take the $\log$ rather than the $\ln$), where the arithmetic mean of Y is μ_y and it corresponding variance is σ_y^2. Consider the average of the Ys: $\bar{Y}_n = \sum_{i=1}^n Y_i/n$. Then due to the central limit theorem:

$$\sqrt{n}\left(\bar{Y}_n - \mu_y\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_y^2\right) \tag{5.7.1471}$$

Now consider $e^{\bar{Y}_n}$:

$$\begin{aligned} e^{\bar{Y}_n} &= \exp\left\{\sum_{i=1}^n \frac{1}{n} Y_i\right\} = \exp\left\{\sum_{i=1}^n \frac{1}{n}\ln(X_i)\right\} = \exp\left\{\sum_{i=1}^n \ln(X_i^{1/n})\right\} \\ &= \prod_{i=1}^n \exp\left\{\ln(X_i^{1/n})\right\} = \prod_{i=1}^n X_i^{1/n} = \sqrt[n]{\prod_{i=1}^n X_i} \end{aligned} \tag{5.7.1472}$$

Thus, $e^{\bar{Y}}$ is the geometric mean of the X_i! So next, we can apply the Delta method to the central limit theorem. Define $g(x) = e^x$, then $g'(x) = e^x$. By the Delta method (see section Theoretical Computing page 6504):

$$\sqrt{n}\left(e^{\bar{Y}_n} - e^{\mu_y}\right) \xrightarrow{d} \mathcal{N}\left(0, e^{2\mu_y}\sigma_y^2\right) \tag{5.7.1473}$$

Notice well the $e^{2\mu_y}$ in front of the σ_y^2! So we you have a tool to make our confidence intervals from!

e^{μ_y} is our true geometric mean of the Ys, and we need to make a confidence interval for this (this is not a confidence interval for the expected value μ). The first step is to estimate σ_y^2. Since σ_y^2 is the variance of the Ys:

$$\sigma_y^2 := \frac{1}{n}\sum_{i=1}^n \left(Y_i - \bar{Y}_n\right)^2 = \frac{1}{n}\sum_{i=1}^n \left(\ln(X_i) - \ln(e^{\bar{Y}_n})\right)^2 = \frac{1}{n}\sum_{i=1}^n \ln\left(\frac{X_i}{e^{\bar{Y}_n}}\right) \tag{5.7.1474}$$

To make our $100(1-\alpha)\%$ confidence interval for the true geometric mean of the Xs we then have:

$$e^{\bar{Y}_n} - Z_{\alpha/2}\frac{e^{\bar{Y}_n}\sigma_y}{\sqrt{n}} \leq G_X \leq e^{\bar{Y}_n} + Z_{\alpha/2}\frac{e^{\bar{Y}_n}\sigma_y}{\sqrt{n}} \tag{5.7.1475}$$

That can also be written:

$$\boxed{\hat{G}_X - Z_{\alpha/2}\frac{e^{\bar{Y}_n}\sigma_y}{\sqrt{n}} \leq G_X \leq \hat{G}_X + Z_{\alpha/2}\frac{e^{\bar{Y}_n}\sigma_y}{\sqrt{n}}} \tag{5.7.1476}$$

This is the "Z-test of the geometric mean" (when the variance is known).

Keep in mind that the Z-test of the geometric mean must be used only if the σ_y is known. Otherwise we should use the Student version of that test! That is:

$$\boxed{\hat{G}_X - T_{\alpha/2}(n-1)\frac{e^{\bar{Y}_n}\hat{\sigma}_y}{\sqrt{n}} \leq G_X \leq \hat{G}_X + T_{\alpha/2}(n-1)\frac{e^{\bar{Y}_n}\hat{\sigma}_y}{\sqrt{n}}} \tag{5.7.1477}$$

This is the "T-test of the geometric mean".

> Caution! Many people and statistical softwares calculate this interval the wrong way. For example in the Stata documentation in 12021 (holocene calendar), the method of calculation is explicitly described as: "*To compute the geometric mean, the function* `ameans` *creates* $u_j = \ln(x_j)$ *for all positive* x_j. *The arithmetic mean of the* u_j *and its confidence interval are then computed as in the function* `ci`. *Let* $\bar{u}$ *be the resulting mean, and let* $[L, U]$ *be the corresponding confidence interval. The geometric mean is then* $\exp(\bar{u})$, *and its confidence interval is* $[\exp(L), \exp(U)]$.". And sadly the package DescTools of the `R` statistical software and its function Gmeans() is based on Stata method...

5.7.9.1.6 Binomial exact Test

Often when measuring we want to compare two small samples taken randomly (without replacement!) from also a small population ... to know if they are statistically significantly different or not as when we were expecting a perfect equality!

We are looking for a suitable test for the following cases:

- To know if the sample of a population prefers to use a given technical method of work rather than another when we expect that the population does not prefer one of the other
- To know if the sample of a population has a predominant characteristic among two possibilities when we expect that the population is well balanced

Before going further into details, let us remind that we must be extremely cautious about how to get the two samples. The experience must be unbiased, this is to say for reminder, that the sampling protocol must not favour one of the both characteristics of the population (if you study the balance between man/woman in a population by attracting people for the survey with a gift in the form of jewellery or just by calling during the workdays you will have a biased sample ... because you'll probably naturally have more women than men...).

This said, this situation match with a binomial distribution for which we proved earlier in this chapter that the probability of k successes in a population of size N with a probability of success is p (probability of failure q being therefore $1-p$) was given by the relation:

$$P(N,k) = C_k^N p^k q^{N-k} = C_k^N p^k (1-p)^{N-k} \tag{5.7.1478}$$

In the case before we are interesting we have $p = q = 0.5$:

$$P(N,k) = C_k^N 0.5^k 0.5^k = 0.5^N C_k^N \tag{5.7.1479}$$

while remembering that the distribution will not be symmetrical and especially if the population size N is small.

If we now denote by x the number of successes (considered as the size of the first sample) and y is the number of failures (considered as the size of the second sample), then we have:

$$P(N,k)=0.5^N C_k^N = P(x+y,k)=0.5^{x+y}C_k^{x+y} \tag{5.7.1480}$$

This being done, to build the test and by the asymmetry of the distribution, we will calculate the cumulative probability that k is smaller than the x obtained by the experience and sum it to the cumulative probability that k is greater than the y obtained by the experiment (which corresponds to a cumulative probability of respectively left and right tails of the distribution). So this sum corresponds to the probability:

$$P=0.5^N\sum_{k=0}^{x} C_k^N + 0.5^N \sum_{k=y}^{N} C_k^N \tag{5.7.1481}$$

and this last relation is named "binomial exact test (two-tailed)".

If the probability P obtained by the sum is above a certain cumulative probability fixed in advance, then we say that the difference with a random sample in a perfectly balanced population is not statistically significant (bilaterally ...) and respectively if it is below, the difference will be statistically significant and therefore we reject the assumed equilibrium.

Therefore, if:

$$0.5^N\sum_{k=0}^{x'} C_k^N + 0.5^N \sum_{k=y'}^{N} C_k^N = 0.5^N\left(\sum_{k=0}^{x'} C_k^N + \sum_{k=y'}^{N} C_k^N\right) \geq \alpha \tag{5.7.1482}$$

the difference with a balanced population will be considered not statistically significant. Often we will α to be at the maximum equal to 5% (but rarely below) which corresponds to a confidence interval of 95%.

Unfortunately from a statistical software to the other the required parameters or results will not necessarily be the same (spreadsheets softwares for example do not include a specific function for the binomial test, will often have to build a table or develop yourself a function). For example, some softwares automatically calculate and impose (which is quite logical in a sense...)

$$0.5^N\left(\sum_{k=0}^{x-1} C_k^N + \sum_{k=y+1}^{N} C_k^N\right) = P \tag{5.7.1483}$$

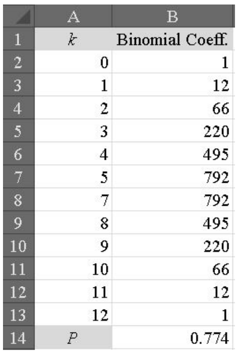
✍**Example:**

From a small population having two particular characteristics x and y that interest us and which we expect to have a perfect balance but as $x=y$ we actually got $x=5$ and $y=7$. We would like do the calculation with Microsoft® Excel® 11.8346 to know whether this difference is statistically significant or not at a level of 5%?

So, to answer this question, we will calculate the cumulative probability:

$$0.5^N\left(\sum_{k=0} C_k^N+\sum_{k=y}^{N} C_k^N\right)=0.5^{12}\left(\sum_{k=0}^{5} C_k^{12}+\sum_{k=7}^{12} C_k^{12}\right) \tag{5.7.1484}$$

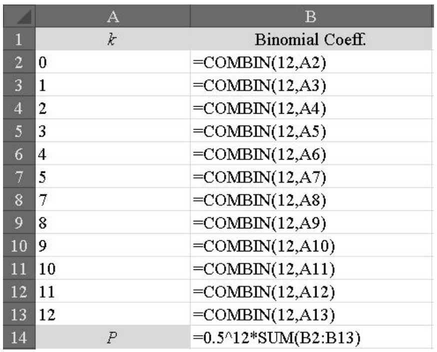
which gives us:



	A	B
1	*k*	Binomial Coeff.
2	0	1
3	1	12
4	2	66
5	3	220
6	4	495
7	5	792
8	7	792
9	8	495
10	9	220
11	10	66
12	11	12
13	12	1
14	*P*	0.774

Figure 5.162: **Calculated values of the binomial coefficients in Microsoft® Excel® 11.8346**

thus explicitly:



	A	B
1	*k*	Binomial Coeff.
2	0	=COMBIN(12,A2)
3	1	=COMBIN(12,A3)
4	2	=COMBIN(12,A4)
5	3	=COMBIN(12,A5)
6	4	=COMBIN(12,A6)
7	5	=COMBIN(12,A7)
8	7	=COMBIN(12,A8)
9	8	=COMBIN(12,A9)
10	9	=COMBIN(12,A10)
11	10	=COMBIN(12,A11)
12	11	=COMBIN(12,A12)
13	12	=COMBIN(12,A13)
14	*P*	=0.5^12*SUM(B2:B13)

Figure 5.163: **Formulas for calculating binomial coefficients in Microsoft® Excel® 11.8346**

thus the cumulative probability being 0.774 (i.e. 77.4%) the difference compared with balanced population will be considered as not statistically significant.

Remark

This test is also used by most statistical software (such as Minitab) to give a confidence interval of the conformity of opinions in relation to that of an expert. This is what we call a R&R study (reproducibility & repeatability) by attributes (see the companion Minitab for an example).

5.7.9.1.7 C.I. for a Proportion

For information some statisticians use the fact that the Normal distribution arises from the Poisson distribution which itself derives from the binomial distribution (we have proved it when n tends to infinity and p and q are of the same order) to build a confidence interval in the context of the analysis of proportions (widely used in the analysis of the quality in the industry).

To see this, we note X_i the random variable defined by:

$$X_i = \begin{cases} 1 & \text{if the } i\text{th element of the sample has the attribute } A \\ 0 & \text{otherwise} \end{cases} \tag{5.7.1485}$$

where the attribute A can be the property "defective" or "non-defective" for example, in an analysis of pieces. We note by k the number of successes of the attribute A.

The random variable $X = X_1 + X_2 + \ldots + X_n$ we have proved it earlier in this chapter, follows a binomial distribution with parameters n and p with the moments:

$$\begin{aligned} \mu &= \mathrm{E}(X) = np \\ \sigma &= \sqrt{\mathrm{V}(X)} = \sqrt{npq} = \sqrt{np(1-p)} \end{aligned} \tag{5.7.1486}$$

That said, we do not know the true value of p. We will use the estimator of the binomial distribution proved above:

$$\hat{p} = \frac{k}{n} = \frac{X}{n} \tag{5.7.1487}$$

Based on the properties of the mean we have then:

$$\mathrm{E}(\hat{p}) = \mu_{\hat{p}} = \mathrm{E}\left(\frac{X}{n}\right) = \frac{np}{n} = p \tag{5.7.1488}$$

And by using the properties of the variance, we have the following relation for the variance of the sample mean of the proportion:

$$\mathrm{V}(\hat{p}) = \sigma_{\hat{p}}^2 = \mathrm{V}\left(\frac{X}{n}\right) = \frac{\mathrm{V}(X)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n} \tag{5.7.1489}$$

This then brings us to:

$$\mu_{\hat{p}} \cong \hat{p} \quad \text{and} \quad \sigma_{\hat{p}} \cong \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{5.7.1490}$$

Finally, remember that we have proved that the Normal distribution resulted from the binomial distribution under certain conditions (practitioners admit that it is applicable as $n > 50$ and $np \geq 5$). In other words, the random variable X following a binomial distribution follows a Normal distribution under certain conditions. Obviously, if X follows a Normal distribution then X/n also (and so do $\hat{p}$...). Therefore we can center and reduce $\hat{p}$ so that it behaves as the reduced Normal centered random variable denoted by Z:

$$Z=\frac{\hat{p}-p}{\sqrt{\dfrac{\bar{p}(1-\bar{p})}{n}}} \tag{5.7.1491}$$

✍**Examples:**

E1. If 5% of the annual production of a business fails, what is the probability that by taking a sample of 75 pieces of the production line only 2% or less will be defective?

We therefore have:

$$Z=\frac{0.02-0.05}{\sqrt{\dfrac{0.05\cdot 0.95}{75}}} \tag{5.7.1492}$$

The corresponding cumulative probability to that value can be easily obtained with Microsoft® Excel® 11.8346:

`=NORMSDIST(-1.19)=11.66%`

But note that we do not have $np\geq 5$ that is satisfied therefore we could exclude to use this result.

E2. In its report from 11998 (holocene calendar), JP Morgan explained that during the year 11998 (holocene calendar) its losses went beyond the Value at Risk (see section Economy page 7858) 20 days on 252 working days of the year based on a 95% temporal VaR (thus 5% of working days considered as loss). At the threshold of 95% it is just bad luck or is that the VaR model used was bad?

$$Z=\frac{\hat{p}-p}{\sqrt{\dfrac{\bar{p}(1-\bar{p})}{n}}}=\frac{n\hat{p}-np}{\sqrt{\hat{p}(1-\hat{p})n}}=\frac{-0.05\cdot 252-20}{\sqrt{0.05(1-0.05)252}}\cong -2.14<-1.96 \tag{5.7.1493}$$

So it was just bad luck.

E3. Some individuals expect all people declared clinically death to experience NDE (Near Death Experience). The study *Near-death experience in survivors of cardiac arrest: a prospective study in the Netherlands* on a sample of 344 patients has 62 patients (18%) that reported a NDE (i.e. 82% did not experience NDE). We want to know if 82% is statistically significantly different of the expected 100% at a threshold of 5%:

$$Z=\frac{n\hat{p}-np}{\sqrt{\hat{p}(1-\hat{p})n}}=\frac{244-344}{\sqrt{0.05(1-0.05)344}}\cong -24.73<-1.96 \tag{5.7.1494}$$

The corresponding cumulative probability to that value can be easily obtained with Microsoft® Excel® 11.8346:

`=NORMSDIST(-24.73)=100%`

Therefore NDE is not a global phenomenon and we can not assign to it, beyond all reasonable doubt, an absolute existence! However other factors may explain the fact that only a few people experience NDE.

We can now approximate the confidence interval for the proportion by using the fact that the binomial distribution has a Normal asymptotic behaviour under the conditions demonstrated during our introduction of the Normal distribution such as we get the "one-proportion Z test" or also more commonly named "one-proportion p test":

$$\boxed{\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \tag{5.7.1495}$$

That (non-parametric) test is also commonly named "binomial test" and must not be confused with the binomial exact test derived at page 869.

Before proceeding to an example, it may be useful to clarify to the reader that this approximation by a Normal distribution is very common and that we'll meet it again numerous times in proofs that will follow. It is even so common that this approximation method has a name..: the "Wald's method" (well actually there are several Wald methods but we will only use the most known one).

Example:

We take $\alpha=5\%$, then we have:

$$\boxed{\hat{p}-Z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\leq p\leq\hat{p}+Z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \tag{5.7.1496}$$

That is to say:

$$\hat{p}-1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\leq p\leq\hat{p}+1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{5.7.1497}$$

On a production of 300 elements we found that 8 of them were defect. What is the confidence interval?

We check first with:

$$\hat{p}=\frac{8}{300}\cong 2.66\% \tag{5.7.1498}$$

that:

$$n\hat{p}=300\cdot\frac{8}{300}=8\geq 5 \tag{5.7.1499}$$

So it is acceptable to use the confidence interval by the Normal distribution. We therefore have:

$$\frac{8}{300}-1.96\sqrt{\frac{\frac{8}{300}\left(1-\frac{8}{300}\right)}{300}}\leq p\leq\frac{8}{300}+1.96\sqrt{\frac{\frac{8}{300}\left(1-\frac{8}{300}\right)}{300}} \tag{5.7.1500}$$

That is to say:

$$0.84\%\leq p\leq 4.49\% \tag{5.7.1501}$$

Caution! Because the comparison of two proportions is of main interest for the pharmaceutical industry, it is therefore a quite intensive mathematical research topic. At the point that the Wald confidence interval for the comparison of two proportions is not the only existing one. Some statistical softwares for (like R for example) give the options to use complementary (more robust or exact) confidence intervals like: the Wilson confidence interval, the Agresti-Coull confidence interval, the Agresti-Caffo confidence interval, the Newcombe confidence interval, the Miettinen-Nurminen confidence interval, the Skewness-corrected asymptotic score (SCAS) confidence interval, the Blaker confidence interval, the Clopper-Pearson confidence interval, the Hepworth confidence interval. As always... never forget the Dunning-Kruger effect... especially in Science!

To conclude this subject, we can obviously be interested to the number of individuals (sample size) necessary to satisfy a certain (imposed) confidence interval accuracy when having an imposed standard deviation.

We therefore have according to the above assumptions and in the acceptance of the approximation by a Normal distribution:

$$Z=\frac{\hat{p}-p}{\sqrt{\dfrac{pq}{n}}}=\frac{\delta}{\sqrt{\dfrac{p(1-p)}{n}}} \tag{5.7.1502}$$

And by proceeding in an identical manner to the developments made above with the Normal distribution, we get the "sample size for proportion estimation by binomial distribution":

$$\boxed{n_p\cong\frac{Z_{\alpha/2}^2p(1-p)}{\delta^2}=\frac{Z_{\alpha/2}^2pq}{\delta^2}} \tag{5.7.1503}$$

obviously we normally take the integer value in practice...

Applying the same reasoning for two groups of proportions we get the "sample size for two proportions estimation by binomial distribution":

$$n_{2p}\cong\frac{Z_{\alpha/2}^2[p_1(1-p_1)+p_2(1-p_2)]}{\delta^2} \tag{5.7.1504}$$

A question that often comes up in practice is the fact to know whether you have to take a unilateral or bilateral test. In fact there is no precise answer, it depends on what we want to highlight.

Remark

The size of the parent population for the relations developed above does not come into consideration in the calculation of confidence intervals or in one of the sample size, and because it is considered infinite. So be careful to not have sometimes sample sizes that are larger than the possible real parent population... Otherwise you should insert the fpc (finite population factor).

The reader can find in most marketing textbooks tabulated values (including the fpc!) in the result above that lead to the construction of the following table:

	Confidence level $=95\%$			**Confidence level** $=99\%$		
	Margin of error			**Margin of error**		
Population size	5%	2.5%	1%	5%	2.5%	1%
100	80	94	99	87	96	99
500	217	377	475	285	421	485
1,000	278	606	906	399	727	943
10,000	370	1,332	4,899	622	2,098	6,239
100,000	383	1,513	8,762	659	2,585	14,227
500,000	384	1,532	9,423	663	2,640	16,055
1,000,000	384	1,534	9,512	663	2,647	16,317

Table 5.47: **One question survey sample size with fpc factor**

✍**Example:**

We would like to know the number of individuals (sample size) to take in a production lot knowing that the proportion of defective units is imposed at 30% with a tolerated error of about 5% between the actual and empirical proportion and to obtain a confidence interval at a level of 95% of the result:

$$n \cong \frac{Z_{5\%/2}^2 \cdot 30\% \cdot (1-30\%)}{5\%^2} = \frac{(-1.96)^2 \cdot 30\% \cdot 70\%}{5\%^2} \cong 322 \tag{5.7.1505}$$

Remark

The last relation is very often used in sampling theory (analysis for referendum with responses of type: Yes/No) where sometimes the sample size n is imposed for costs reasons of the survey and for which we seek to calculate the uncertainty δ and sometimes the reciprocal (the uncertainty is imposed and therefore we seek to know the sample size).

5.7.9.1.7.1 Variance stabilizing transformation

Often, the variance of an estimator depends on the value of the parameter being estimated. For example, if $X_n \sim$ Binomial (n, p), then with $\hat{p}_n = X_n/n$:

$$\mathrm{V}\left[\widehat{p}_n\right] = \frac{p(1-p)}{n} \tag{5.7.1506}$$

In regression situations, for instance, one usually desires the dependent Y_i' s to have the same variance for each i, so that if these Y_i' s are binomial, or Poisson, the variance will not be constant. However, taking a function of the variables may achieve approximately equal variances. Such a function is named a "variance stabilizing transformation". Formally, if $\widehat{\theta}_n$ is an estimator of θ, then we wish to find a g such that:

$$\sqrt{n}\left(g\left(\widehat{\theta}_n\right) - g(\theta)\right) \xrightarrow{d} \mathcal{N}(0,1) \tag{5.7.1507}$$

The "1" for the variance is arbitrary. The important thing is that it does not depend on θ. In the binomial example, the variance stabilizing g would satisfy:

$$\sqrt{n}\left(g\left(\hat{p}_n\right) - g(p)\right) \xrightarrow{d} \mathcal{N}(0,1) \tag{5.7.1508}$$

We know that:

$$\sqrt{n}\left(\widehat{p}_n - p\right) \xrightarrow{d} \mathcal{N}(0, p(1-p)) \tag{5.7.1509}$$

and by the univariate Δ-method (see section Theoretical Computing page 6504):

$$\sqrt{n}\left(g\left(\hat{p}_n\right) - g(p)\right) \xrightarrow{d} \mathcal{N}\left(0, g'(p)^2 p(1-p)\right) \tag{5.7.1510}$$

What should g be so that the variance is equal to 1? We need to solve to answer that question:

$$g'(p) = \frac{1}{\sqrt{p(1-p)}} \tag{5.7.1511}$$

so that:

$$g(p) = \int_0^p \frac{1}{\sqrt{y(1-y)}}\mathrm{d}y \tag{5.7.1512}$$

First, let $u = \sqrt{y}$, so that $y = u^2$ and $\mathrm{d}y = 2u\mathrm{d}u$, and:

$$g(p) = \int_0^{\sqrt{p}} \frac{1}{u\sqrt{1-u^2}} 2u\mathrm{d}u = 2\int_0^{\sqrt{p}} \frac{1}{\sqrt{1-u^2}}\mathrm{d}u \tag{5.7.1513}$$

The integral is $\arcsin(u)$ (see section Differential and Integral Calculus page 1534), which means the variance stabilizing transformation is:

$$g(p) = 2\arcsin(\sqrt{p}) \tag{5.7.1514}$$

Note that adding a constant to g won't change the derivative. The approximation suggested by:

$$\sqrt{n}\left(g\left(\hat{p}_n\right) - g(p)\right) \longrightarrow^{\mathcal{D}} \mathcal{N}(0,1) \tag{5.7.1515}$$

is then after rearranging the previous relation:

$$2\arcsin\left(\sqrt{\hat{p}_n}\right) \cong \mathcal{N}\left(2\arcsin(\sqrt{p}), \frac{1}{n}\right) \qquad (5.7.1516)$$

An approximate 95% confidence interval for $2\arcsin(\sqrt{p})$ is:

$$2\arcsin\left(\sqrt{\widehat{p}_n}\right) \pm \frac{1}{\sqrt{n}} \qquad (5.7.1517)$$

That interval can be inverted to obtain the interval for p, that is, apply $g^{-1}(u)=\sin(u/2)^2$ to both ends:

$$p \in \left(\sin\left(\arcsin\left(\sqrt{\hat{p}_n}\right)-\frac{1}{\sqrt{n}}\right)^2, \sin\left(\arcsin\left(\sqrt{\hat{p}_n}\right)+\frac{1}{\sqrt{n}}\right)^2\right) \qquad (5.7.1518)$$

This interval may be slightly better than the usual approximate interval:

$$\widehat{p}_n \pm 2\sqrt{\frac{\widehat{p}_n\left(1-\hat{p}_n\right)}{n}} \qquad (5.7.1519)$$

5.7.9.1.7.2 Test of equality (or difference) of two Proportions

Always in the same context as the previous approximation of the binomial distribution by a Normal distribution, the industry (especially biostatistics) likes to compare two proportions of two different populations to see if they are statistically equal or not (i.e. said statistically significantly different or not).

Therefore, let us recall that we have proved the stability of the Normal distribution if two random variables are independent and identically distributed (according to a Normal distribution!):

$$X-Y=\mathcal{N}\left(\mu=\mu_X-\mu_Y,\sigma=\sqrt{\sigma_X^2+\sigma_Y^2}\right) \tag{5.7.1520}$$

Under the above assumptions it is then approximately the same for the difference of two proportions:

$$\begin{aligned}\frac{X}{n_1}-\frac{Y}{n_2}&\cong\mathcal{N}\left(\mu_{\bar{p}_X}-\mu_{\bar{p}_Y},\sqrt{\sigma_{\bar{p}_X}^2+\sigma_{\bar{p}_Y}^2}\right)\cong\mathcal{N}\left(\hat{p}_X-\hat{p}_Y,\sqrt{\sigma_{\hat{p}_X}^2+\sigma_{\hat{p}_Y}^2}\right)\\&=\mathcal{N}\left(\hat{p}_X-\hat{p}_Y,\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_1}+\frac{\hat{p}_Y(1-\hat{p}_Y)}{n_2}}\right)\end{aligned} \tag{5.7.1521}$$

Therefore we know that this new reduced centered variable follows a Normal distribution as:

$$Z=\frac{(\hat{p}_X-\hat{p}_Y)-\mu}{\sqrt{\dfrac{\hat{p}_X(1-\hat{p}_X)}{n_1}+\dfrac{\hat{p}_Y(1-\hat{p}_Y)}{n_2}}} \tag{5.7.1522}$$

and as we seek to know the cumulative probability that the mean of the difference is zero, the latter relation is reduced in this case to "Z-test for the equality of two proportions unpooled":

$$Z=\frac{\hat{p}_X-\hat{p}_Y}{\sqrt{\dfrac{\hat{p}_X(1-\hat{p}_X)}{n_1}+\dfrac{\hat{p}_Y(1-\hat{p}_Y)}{n_2}}} \tag{5.7.1523}$$

Obviously we can build (as always...) a confidence interval:

$$\boxed{(\hat{p}_X-\hat{p}_Y)-Z_{\alpha/2}\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_1}+\frac{\hat{p}_Y(1-\hat{p}_Y)}{n_2}}\leq p_X-p_Y\leq(\hat{p}_X-\hat{p}_Y)+Z_{\alpha/2}\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_1}+\frac{\hat{p}_Y(1-\hat{p}_Y)}{n_2}}} \tag{5.7.1524}$$

We refer to this interval as the "Wald interval". Wald intervals for a single proportion or for the difference of two proportions are known to perform poorly, and are usually well below the nominal coverage), although they are still widely used in practice because of their simplicity. The Wald interval also suffers from overshoot,that is, bounds of the calculated interval can be outside the range of possible values!

The above test in the field of sensitivity analysis (especially when applied to the design of websites), is sadly named because of a lack of precision the "proportions A/B test" (not to be confuse with the other kind of A/B tests that we will see further below). Indeed we will see later that a confusion can arise because a pooled version of that test also exist, but also an unpooled χ^2 version and pooled χ^2 version...

Warning! The test of the difference of two proportions of two different samples is obviously not the same as the test of the difference of two proportions in the same sample (covariance oblige since the two proportions are therefore no longer independent)! In the latter case we use McNemar's test (see section Numerical Methods page 1264).

However, it seems that the following approximate relation, named "Z-test for the equality of two proportions pooled", following the return on experience is more correct when taking a modified denominator:

$$Z=\dfrac{\hat{p}_X-\hat{p}_Y}{\sqrt{\dfrac{\hat{p}_X(1-\hat{p}_X)}{n_1}+\dfrac{\hat{p}_Y(1-\hat{p}_Y)}{n_2}}}=\dfrac{\hat{p}_X-\hat{p}_Y}{\sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}} \qquad (5.7.1525)$$

where $\hat{p}$ will be taken as the mixture of the two populations. That is to say:

$$\hat{p}=\dfrac{\text{sum of events in each of the two samples}}{\text{sum of samples sizes}} \qquad (5.7.1526)$$

thus (by changing the notations of the indices of the experimental proportions):

$$\hat{p}=\dfrac{n_1\hat{p}_1+n_2\hat{p}_2}{n_1+n_2} \qquad (5.7.1527)$$

This test is named "two proportions Z test" or more simply "two proportions p test". In medicine, it is named the "test of differences in risk" (meaning implicitly that each proportion is a segment of the studied population in relation to an undesired event).

Remark

Numerical simulations seems to show that the unpooled variance is slightly more biased than the pooled one. We will also see during our study of the χ^2 test of independance (see page 1238) why softwares like `R` use the χ^2 test instead of the Z-test to compare two proportions.

✍**Example:**

In the context of a sampling plans (see section Industrial Engineering page 8904) we have taken on a first batch of 50 individuals, 48 that are in perfect conditions. In a second batch of 30 individuals, 26 were in perfect condition.

Thus we have:

$$\hat{p}_X=\frac{48}{50}=96\%,\hat{p}_Y=\frac{26}{30}=86.66\% \tag{5.7.1528}$$

We would like to know if the difference is statistically significant with a 95% confidence interval or simply due to chance. We then use:

$$\hat{p}=\frac{n_1\hat{p}_1+n_2\hat{p}_2}{n_1+n_2}=\frac{48+26}{50+30}=92.50\% \tag{5.7.1529}$$

and:

$$Z=\frac{\hat{p}_X-\hat{p}_Y}{\sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}}=\frac{0.96-0.8666}{\sqrt{0.925(1-0.925)\left(\dfrac{1}{50}+\dfrac{1}{30}\right)}}\cong 1.535 \tag{5.7.1530}$$

This corresponds to a cumulative probability using Microsoft® Excel® 11.8346:

```
=NORMSDIST(1.535)=93.77%
```

Therefore the difference is due to chance (that said it is almost in extremis...). In other words, it is not statistically significant under the set constraints.

5.7.9.1.8 Sign Test

We measure something on a sample and later, we measure the same thing on this same sample but with a different method (so it is therefore paired samples!). Both ordered rankings of measures are compared and too each case is assigned a sign ("$+$" for an increase in the rankings, "$-$" in case of descent). Those who remain at the same level are eliminated.

According to the hypothesis to be tested, there are so many "$+$" as "$-$", that is to say, the median of the distribution has not changed (this statement may not seem obvious at first reading so be aware to take time to think about it).

The idea is that for each pair of values, there are only two possible signs of change, we have a chance on two (50% probability) that the difference is positive or negative. This test is based only on the study of signs of the differences between the pairs of individuals, regardless of the values o these differences.

We can wish to control two assumptions:

- The inequality of proportions of signs must be statistically significant. So one of the two signs must be in a small number compared to the other, which corresponds to a left-sided test (the cumulative probability of the small number of characters must be below a certain level α).
- The proportion of the two signs must be low unbalanced $(P(+) = P(-) = 0.5)$. It is therefore in this case a bilateral test (the most common case) with a given level α.

To create such a test, we consider the appearance of the "$+$" and "$-$" as a binary random sampling system where the order of success is not taken into account (it is therefore based on a binomial or hypergeometric distribution) and with replacement (which immediately eliminates the hypergeometric distribution that is not symmetric and problematic to use in practice...). To consider a random sampling with replacement (with the fact that we do not reinject each individual in reality), it is necessary that the population N is large. This is why the sign test considers that the paired values should be continuous (which allows verbatim to approach the hypergeometric distribution by the binomial distribution). However, some statistical software use the hypergeometric distribution for the sake of precision.

Remark

You should know that most statistical software, do implicitly the assumption in this test that the data are continuous and use therefore the binomial distribution.

✍**Example:**

Consider two sets of measurements with two different methods. We would like to test the hypothesis with a confidence level α of 5% if the difference between the two methods is statistically significant (thus we expect a balance of signs). This is therefore two samples sign test (knowing that it is also possible to do the same by comparing the values of a single sample to its median):

$$\begin{gathered}20.4, 25.4, 25.6, 25.6, 26.6, 28.6, 28.7, 29, 29.8, 30.5, 30.9, 31.1\\20.7, 26.3, 26.8, 28.1, 26.2, 27.3, 29.5, 32, 30.9, 32.3, 32.3, 31.7\end{gathered} \tag{5.7.1531}$$

Therefore we have the differences:

$$-0.3, -0.9, -1.2, -2.5, 0.4, 1.3, -0.8, -3.0, -1.1, -1.8, -1.4, -0.6 \tag{5.7.1532}$$

With the signs:

$$-, -, -, -, +, +, -, -, -, -, -, - \tag{5.7.1533}$$

Well it is already clear that the result will be the rejection of the hypothesis as there is no difference. But we will still do the calculations. As the test is a two-sided at the level of 5% , the cumulative probability of obtaining at least two signs "+" must not be less than 2.5% and not more than 97.5% if we want to accept (not reject) the assumption as that the difference is not statistically significant.

We then have:

$$P(X \leq 2) = \sum_{k=0}^{2} C_k^{12} 0.5^k (1-0.5)^{12-k} = \sum_{k=0}^{2} C_k^{12} 0.5^k 0.5^{12-k} = 0.5^{12} \sum_{k=0}^{2} C_k^{12} \cong 1.93\% \tag{5.7.1534}$$

Either with Microsoft® Excel® 14.0.6123:

`=BINOMDIST(2,12,0.5,1)=1.928%`

or if we don't do the approximation by being more accurate with the hypergeometric distribution:

`=HYPGEOM.DIST(2,24/2,12,24,TRUE)=0.17%`

which is not really better...!

So the cumulative probability is less than 2.5% and is by far not more than 97.5%, therefore we reject the hypothesis as that the difference is not statistically significant.

We could accept the hypothesis if we take for α the value:

$$p - \text{value} \cong 1.93\% \cdot 2 = 3.86\% \tag{5.7.1535}$$

but this is not the case!

To conclude on this sign test (median test), we have for information some statistical softwares that propose a confidence interval of the median based on the calculation method described previously (confidence interval of the binomial distribution). However, we think that it would better to use bootstrapping techniques as we have seen in the section on Numerical Methods (page 6740), so won't introduce this technique here. In addition it may be useful to know that some make an approximation using the Normal distribution (as with most tests but we won't study this approximation in this context).

5.7.9.1.9 Mood's Median Test

Here we will introduce a test that has many names: "median test", "Mood's median test" or "Westenberg-Mood's median test" or "Brown-Mood's median test" ...

We consider two independent samples $(X_1,\ldots,X_{n_1})$ and $(Y_1,\ldots,Y_{n_2})$. We assume that $(X_1,\ldots,X_{n1})$ is an independent and identically distributed sample from a continuous distribution F and $(Y_1,\ldots,Y_{n_2})$ is an independent and identically distributed sample from a continuous distribution G.

After the grouping of the n_1+n_2 values of the two samples, $k=n_1M_n$ is the number of observations X_i of the first sample that are greater than the median $N=n_1+n_2$ of the observations (the notation is not great because it can give the impression that it is a multiplication...).

Under the null hypothesis that X and Y variables follow the same continuous distribution (that is to say, $G=F$) hypothesis, the variable $k=n_1M_n$ can take the values $0,1,\ldots,n_1$ according to the hypergeometric distribution:

$$P(N,N/2,n_1,k)=\frac{C_k^{n_1}C_{N/2-k}^{n_2}}{C_{N/2}^{N}}=\frac{\dfrac{n_1!}{k!(n_1-k)!}\dfrac{n_2!}{(N/2-k)!(n_2-(N/2-k))!}}{\dfrac{N!}{N/2!(N-N/2)!}} \tag{5.7.1536}$$

Therefore, we can calculate the unilateral cumulative probability of having k. Mood's test is also a purely unilateral test.

✍**Example:**

Consider the two samples:

$$\begin{aligned}&23.4,24.4,24.6,24.9,25.0,26.2,26.3,26.8,26.8,26.9,27.0,27.6,27.7\\&22.5,22.9,23.7,24,24.4,24.5,25.3,26,26.2,26.4,26.7,26.9,27.4\end{aligned} \tag{5.7.1537}$$

The overall median calculated with Microsoft® Excel® 14.0.6123 is 26.10. We have a total of:

$$k=n_1M_n=8 \tag{5.7.1538}$$

Then it comes with Microsoft® Excel® 14.0.6123:

`=HYPGEOM.DIST(8,26/2,13,26,TRUE)=94.24%`

So at a threshold of 5%, we do not reject the null hypothesis (but... being close to the limit this is a bit tedious to conclude that...). If we do the same calculation using the binomial distribution we obtain:

`=BINOM.DIST(8,26/2,0.5,1)=86.65%`

But obviously here the approximation does not apply since a binomial approximation is acceptable in practice when the sample is about 10 times smaller than the population.

Remark

Unfortunately, there are several versions of the Mood test. For example, a software such as Minitab compares thanks to a contingency table... the values above or below the median and made a simple chi-square test of independence (Pearson's test) as seen in the section of Numerical Methods.

5.7.9.1.10 Poisson Test (1 sample)

We know that a number of rare events follow a Poisson distribution. We can then allow us as for any other distribution to calculate the cumulative probability in a given interval (bilateral or unilateral).

So if we have a discrete random variable following a Poisson distribution:

$$P_k = \frac{\mu^k}{k!}e^{\mu} \qquad (5.7.1539)$$

We then have to a certain right sided level of confidence α, the closest value n of k satisfying the condition:

$$\sum_{k=0}^{n} \frac{\mu^k}{k!}e^{\mu} \rightarrow 1-\alpha \qquad (5.7.1540)$$

So to find the value n (strictly positive integer or null value) we should reverse the sum, which is not... something funny to do (this is why most spreadsheet softwares do not offer at this day the inverse of the Poisson distribution).

Now recall that we have seen in the section on Sequences and Series, the following Taylor (Maclaurin) series with full integral rest to order $n-1$ around 0 to λ:

$$\sum_{k=0}^{n-1} \frac{\lambda^k}{k!}e^{\lambda} = 1 - \int\limits_0^{\lambda} \frac{1}{2^{k/2-1}\Gamma(k/2)} x^{k/2-1}e^{-x/2}\mathrm{d}x \qquad (5.7.1541)$$

Result we had also given in the form of functions for Microsoft® Excel® 14.0.6123 so that the reader can verify this equivalence:

`=POISSON(`$x \in \mathbb{N}, \mu$`,TRUE)`
`=1-CHIDIST(`$2\mu, 2(x+1)$`,TRUE)`

Then it follows that in spreadsheets softwares, we can use the inverse chi-square distribution to calculate the inverse of the Poisson distribution with this time however a small nuance: the result will not necessarily be an integer.

If for example we take (always with Microsoft® Excel® 14.0.6123)

```
=1-CHI.DIST(2*20,2*(15+1),TRUE)=15.6513135%
```

The question is then to find the notation for the opposite ... This is then given by (we divide by 2 to fall back on the mean that is the value of interest):

`=CHIINV(1-15.6513135%,2*(15+1))/2=15.53194258`

Finally, the notation of the inverse is relatively natural. Thus, the "1 sample Poissons' test" at a given right sided α level can be written:

$k \leq$`=CHIINV(1-alpha,2*(number of measures+1))/2`

Formally:

$$\boxed{\lambda \leq \frac{\chi^2_{1-\alpha}2(n_{\text{choosed}}+1)}{2}} \tag{5.7.1542}$$

Note however one thing! It seems that some statistical software approximate sometimes with abuse the Poisson distribution by a Normal distribution. Therefore, the unilateral interval is calculated with:

$$\mu \leq \bar{X} + Z_{1_\alpha}\frac{\sigma}{\sqrt{n}} \tag{5.7.1543}$$

But with the Poisson distribution, remember that we have:

$$\begin{aligned} \mu = \sigma^2 &= \lambda \\ \bar{X} &= \hat{\lambda} \end{aligned} \tag{5.7.1544}$$

Therefore:

$$\lambda = \hat{\lambda} + Z_{1-\alpha}\sqrt{\frac{\lambda_{\text{choosed}}}{n}} \tag{5.7.1545}$$

✍**Example:**

A company manufactures televisions in a constant quantity and has measured the number of defective product produced each quarter for the past ten years (so 4 times 10 measures). The stakeholders determines the maximum acceptable number of defective units is 20 per quarter and wants to determine if the production satisfies these requirements (under the assumption that the distribution of defective follow a Poisson distribution) at a confidence level of 5%.

The 40 measures give us an average of:

$$\bar{X} = \hat{\lambda} = 17.825 \tag{5.7.1546}$$

Then we have with the rough approximation:

$$\lambda = 17.825 + Z_{1-5\%}\sqrt{\frac{20}{40}} \tag{5.7.1547}$$

Either in a spreadsheet software like Microsoft® Excel® 14.0.6123:

```
=NORM.S.INV(1-5%)*SQRT(20/40)+17.825=18.988
```

or:

$$\lambda \leq \frac{\chi^2_{1-5\%}2(20+1)}{2} \tag{5.7.1548}$$

Either in a spreadsheet software like Microsoft® Excel® 11.8346:

```
=CHIINV(1-5%,2*(20+1))/2=14.072
```

In both cases, we are below the imposed average of 2 (so we reject the null hypothesis that the number of defects is greater than or equal to 20). Obviously, it is possible for each of the methods to determine what should be the cumulative probability (confidence level) which brings us to the 20 limit (therefore the $p$value in other words on which we will come back further below). With the first method, aka Normal approximation, the p-value (see page 948) is 0.104%.

Obviously, in the bilateral case, we have:

$$\boxed{\frac{\chi^2_{\alpha/2}2(n+1)}{2} \leq \lambda \leq \frac{\chi^2_{1-\alpha/2}2(n+1)}{2}} \tag{5.7.1549}$$

Example:

An airline company had 2 crash on $1,000,000$ flights (very rare event). What is two-sided the confidence at the level of 95% knowing that globally the number of accidents per million is 0.4.

We have therefore:

$$\frac{\chi^2_{5\%/2}2(2+1)}{2} \leq \lambda \leq \frac{\chi^2_{1-5\%/2}2(2+1)}{2} \tag{5.7.1550}$$

Either the upper bound with a spreadsheet software like Microsoft® Excel® 11.8346 is given by:

```
=CHIINV(1-5%/2,2*(2+1))/2=7.224
```

and for the lower bound:

```
=CHIINV(1-5%/2,2*(2+1))/2=0.618
```

So statistically, the company is less secure than all companies.

5.7.9.1.11 Poisson Test (2 samples)

We have just seen that:

$$\frac{\chi^2_{\alpha/2}2(n+1)}{2} \leq \lambda \leq \frac{\chi^2_{1-\alpha/2}2(n+1)}{2} \tag{5.7.1551}$$

However, following the same reasoning that led us to construct the following test of average comparison:

$$(\bar{X}_2-\bar{X}_1)-Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \leq \mu_2-\mu_1 \leq (\bar{X}_2-\bar{X}_1)+Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \tag{5.7.1552}$$

or its equivalent with the Student distribution when the true standard deviation is not known and using the fact that we have shown that the Poisson distribution is stable by the addition (and hence by subtraction), that Gamma distribution was also stable by the addition (and therefore by subtraction) and also the chi-square distribution since it is only a special case of the gamma distribution, we tend to write perhaps a little bit to fast the extension of what we have seen before:

$$\frac{\chi^2_{\alpha/2}(2(n_2+1)-2(n_1+1))}{2} \leq \lambda_2-\lambda_1 \leq \frac{\chi^2_{1-\alpha/2}(2(n_2+1)-2(n_1+1))}{2} \tag{5.7.1553}$$

And in the facts this is a trap as say some practitioners ... because the chi-square distribution has a support which is defined as being strictly positive and the confidence interval could naturally have a negative left terminal (... O_o). One solution could be to use the test of the difference of two proportions that we have already discussed earlier:

$$Z \cong \frac{(\hat{p}_X-\hat{p}_Y)-\mu}{\sqrt{\dfrac{\hat{p}_X(1-\hat{p}_X)}{n_1}+\dfrac{\hat{p}_Y(1-\hat{p}_Y)}{n_2}}} \tag{5.7.1554}$$

Of course, only in the case that the conditions for approaching the test with a Normal distribution are met (proportions have to be typically less than 0.1 and no greater than 50).

Most softwares seem to have implemented this latter method (with which I do not necessarily agree).

✍**Example:**

An airline company had 2 airplanes crash was in $1,000,000$ flights (very rare event). Another company had 3 crashes in $1,200,000$ flights. What is the two sided confidence interval at the level of 95% assuming that the difference is zero.

Therefore, the proportions are:

$$\hat{p}_X=\frac{2}{n_1}=\frac{2}{1,000,000}=0.000002 \quad \text{and} \quad \hat{p}_Y=\frac{3}{n_2}=\frac{3}{1,200,000}=0.0000025 \tag{5.7.1555}$$

We write:

$$\begin{aligned}\Delta&=\hat{p}_X-\hat{p}_Y=-0.0000005\\ \hat{q}_X&=1-\hat{p}_X\\ \hat{q}_Y&=1-\hat{p}_Y\end{aligned} \tag{5.7.1556}$$

Then we have:

$$\Delta-Z_{5\%/2}\sqrt{\frac{\hat{p}_X\hat{q}_X}{n_1}+\frac{\hat{p}_Y\hat{q}_Y}{n_2}}\leq\mu\leq\Delta+Z_{5\%/2}\sqrt{\frac{\hat{p}_X\hat{q}_X}{n_1}+\frac{\hat{p}_Y\hat{q}_Y}{n_2}} \tag{5.7.1557}$$

which gives a confidence interval for the expected theoretical proportion difference:

$$-0.000004461\leq\mu\leq 0.000003461 \tag{5.7.1558}$$

and therefore as -0.0000005 is in this interval, we accept the hypothesis as the difference of proportions are not statistically significant at the threshold of 5%.

Or taking the non approximated expression, we have (with the same conclusion):

$$\begin{gathered}\frac{\chi^2_{5\%/2}(2(3+1)-2(2+1))}{2}\leq\lambda_2-\lambda_1\leq\frac{\chi^2_{1-5\%/2}(2(3+1)-2(2+1))}{2}\\ \Downarrow\\ 0.0253\leq\mu\leq 3.6889\end{gathered}$$

So to summarize some convergence of distributions in all these different tests and intervals that we have seen so far, we offer the reader the following diagram that we hope... will clarify perhaps more or less things:

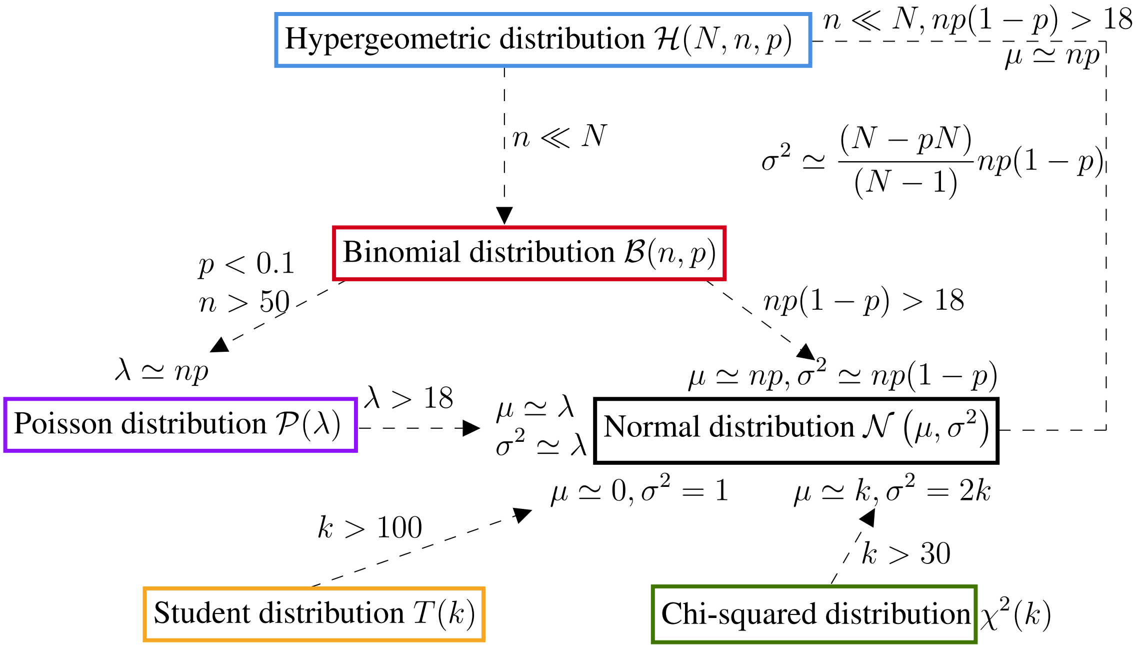

Hypergeometric distribution $\mathcal{H}(N,n,p)$
$n \ll N, np(1-p) > 18$
$\mu \simeq np$
$n \ll N$
$\sigma^2 \simeq \dfrac{(N-pN)}{(N-1)} np(1-p)$
Binomial distribution $\mathcal{B}(n,p)$
$p < 0.1$
$n > 50$
$np(1-p) > 18$
$\lambda \simeq np$
$\mu \simeq np, \sigma^2 \simeq np(1-p)$
Poisson distribution $\mathcal{P}(\lambda)$
$\lambda > 18$
$\mu \simeq \lambda$
$\sigma^2 \simeq \lambda$
Normal distribution $\mathcal{N}(\mu, \sigma^2)$
$\mu \simeq 0, \sigma^2 = 1$
$\mu \simeq k, \sigma^2 = 2k$
$k > 100$
$k > 30$
Student distribution $T(k)$
Chi-squared distribution $\chi^2(k)$

Figure 5.164: Convergence of different customary distributions in elementary statistical inference

And also this table where all relations have been demonstrated in detail above, and some already used (others will be used later):

Statistical sampling	Statistic Average	Statistic Standard-Deviation
Mean (infinite population)	μ	$\dfrac{\sigma}{\sqrt{n}}$
Mean (finite population)	μ	$\dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}}$
Proportion (finite population)	p	$\sqrt{\dfrac{p(1-p)}{n}}$
Proportion (infinite population)	p	$\sqrt{\dfrac{p(1-p)}{n}}\sqrt{\dfrac{N-n}{N-1}}$
$\hat{\sigma}^2$ (infinite population)	$\dfrac{n-1}{n}\sigma^2$	$\sqrt{\dfrac{2(n-1)}{n^2}}\sigma^2$

Table 5.48: Table of statistical sampling proved earlier and used in part until now

Definition (#5.7.137): In Bayesian terms "abduction" is the process of finding the shortest prior hypothesis, technically named the "Minimum Description Length" (MDL) (but also known under the name of "Occam's razor") that explains the best the data. With induction, we are often trying to figure out what will be most likely in the future. With abduction, we are often trying to figure out what best explains something that already occurred.

5.7.9.1.12 Confidence/Tolerance/Prediction/Credibility Intervals

Here we go and in order to avoid frequent confusion and before moving on to more complex subjects, we will compare the confidence interval, the credibility interval, the tolerance interval (often named "fluctuation interval" in some school programs) and finally the prediction interval.

Definitions (#5.7.138):

D1. The "tolerance interval" (or "fluctuation interval"), from the statistician point of view[47], is an interval containing a certain percentage (usually $68.26, 95.44$ or 99.73% in the case of a Normal distribution) of individuals in a population of measures.

D2. The "confidence interval" for a sample mean (or a sample proportion) must be read as the percentage of similarly constructed confidence intervals (usually $90, 95$ or 99% in the two sided case) that will contain the true expected mean (or true expected proportion) knowing the measurement data a posteriori.

D3. The "credibility interval" for a sample mean (or proportion p) is the interval value to a given probability level (usually $90, 95$ or 99% in the two sided case) for the expected mean (true average) or the proportion of the population where we have special prior for the mean/standard deviation or for the proportion (see further below page 910).

D4. The "prediction interval" is used to determine an interval for a single value based on the knowledge of the sample mean and the standard deviation of the population.

An example being more often better than a thousand words, consider the case where the mean and the standard deviation of prices are 49 DVD are given by:

$$\bar{X}=\hat{\mu}=31.5531 \quad S=\hat{\sigma}=0.7992 \tag{5.7.1559}$$

Therefore we have:

$$\begin{aligned}[\bar{X}\pm S]&=[30.8,32.4]\\ [\bar{X}\pm 2S]&=[30.0,33.2]\\ [\bar{X}\pm 3S]&=[29.2,34.2]\end{aligned} \tag{5.7.1560}$$

corresponding to the tolerance intervals according to a Normal distribution of $68.26, 95.44$ and 99.73%.

But a confidence interval of 95% based on the relation proved above:

$$\bar{X}-\frac{S}{\sqrt{n}}T_{\alpha/2}(n-1)\leq\mu\leq\bar{X}-\frac{S}{\sqrt{n}}T_{\alpha/2}(n-1) \tag{5.7.1561}$$

gives:

$$31.32\leq\mu\leq 31.78 \tag{5.7.1562}$$

---

[47] At the opposite of that of the mechanical engineer where the tolerance interval is obviously something the measurement interval in which a piece will not be rejected from the production.

So 95% cumulative probability that the true mean is between 31.32 and 31.78.

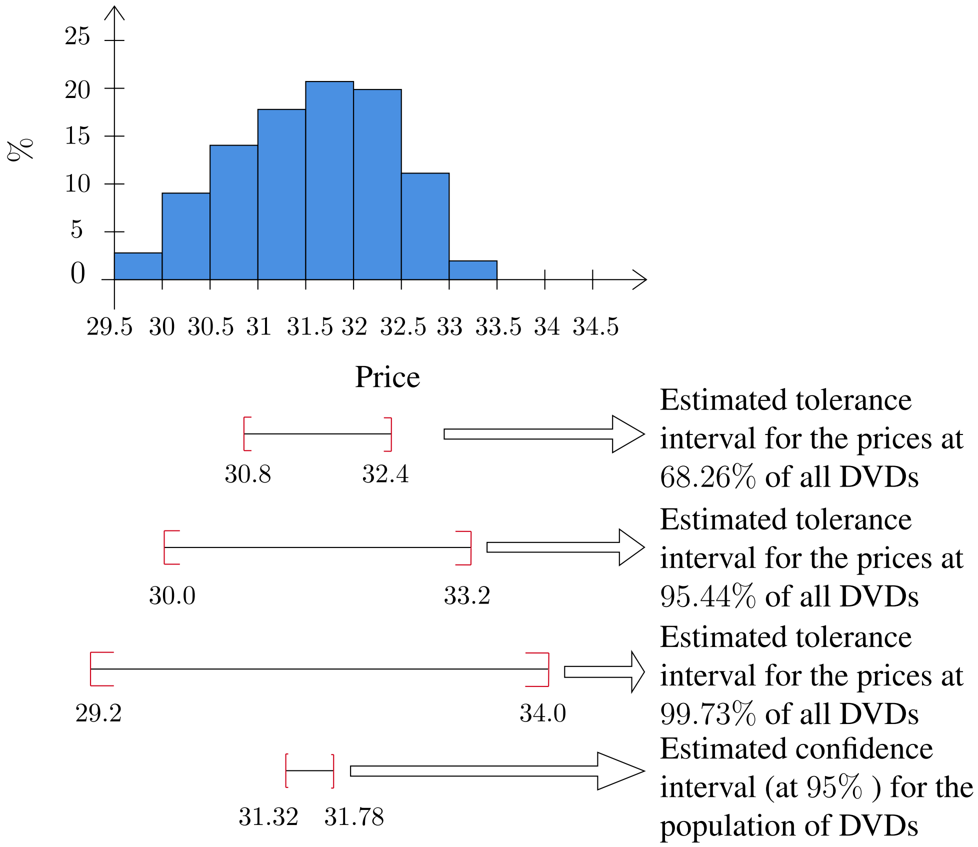


Figure 5.165: **Histogram of the prices of a sample of** 49 **DVDs**

Now lets us introduce a new concept that is rarely addressed in the statistics literature. The idea of the prediction interval is rather than look at the confidence interval of the mean based on an experimental average, to use this experimental average (sample mean) as a basis for predicting the interval of a single value (and not of the average!).

We'll look at the difference between the mean and a punctual value:

$$\bar{x}-x \qquad (5.7.1563)$$

that we will assume close to zero (it is better to have a reliable product and pass the tests to obtain the authorization of sales...). About the variance, what interests us is not just the standard deviation of the mean anymore, but the standard deviation of the difference... and as the sample is independent of the unique value we have:

$$\sigma_{\bar{x}-x}^2=\sigma_{\bar{x}}^2+\sigma_x^2=\frac{\sigma^2}{n}+\sigma^2=\sigma^2\left(1+\frac{1}{n}\right) \qquad (5.7.1564)$$

So we can write as a first approximation:

$$Z\cong\frac{\bar{x}-x}{\sigma_{\bar{x}-x}^2}=\frac{\bar{x}-x}{\sqrt{\sigma^2\left(1+\dfrac{1}{n}\right)}}=\frac{\bar{x}-x}{\sigma\sqrt{1+\dfrac{1}{n}}} \qquad (5.7.1565)$$

And of course after what we saw:

$$T(n-1)\cong\frac{\bar{x}-x}{S\sqrt{1+\dfrac{1}{n}}} \tag{5.7.1566}$$

So we can build verbatim the prediction interval:

$$\boxed{\bar{x}-T_{\alpha/2}(n-1)S\sqrt{1+\frac{1}{n}}\leq x\leq\bar{x}+T_{\alpha/2}(n-1)S\sqrt{1+\frac{1}{n}}} \tag{5.7.1567}$$

5.7.9.2 Bayesian Inference

To understand the idea of "Bayesian inference" (to not be confuse with "Bayesian Hypothesis Testing") remember that in the non-Bayesian (frequentist) approach we measure for example a parameter θ. Afterwards, we calculate the standard error or a confidence interval on which we perform a hypothesis test.

In the Bayesian approach we assume that the parameter θ follows a non-trivial prior distribution. We then calculate its posterior probability distribution knowing the measurements made and we calculate afterwards a confidence interval on that latter posterior.

According to some specialists there is now a plenty of theorem and masses of worked-out numerical examples that shows evidence of the superiority of the Bayesian method in a hundred of different areas! The traditional frequentist methods which use only sampling distributions are usable and useful in many particularly simple, idealized problems; however, they represent the most proscribed special cases of probability theory, because they presuppose conditions (independent repetitions of a random experiment but no relevant information!) that are hardly ever met in real problems.

Remark

The frequentist approach is the case of what we also name "uninformative prior". But to reader must not think that a uniform law prior is an example of uninformative prior (there is a detailed mathematical example further below just after our study of beta-binomial posterior)!!! Indeed, the pure frequentists doesn't make ANY prior (i.e. the probability is completely indeterminate $0/0$)! The pure frequentist always assumes that the sample data is exactly representative of the population because that's all it can do.

Let us assume that we have an univariate bayesian model with a likelihood $P(y|\vec{\theta})$ (also sometimes denoted $f(y|\vec{\theta})$) and an univariate prior distribution $P(\vec{\theta})$ (also sometimes denoted $f(\vec{\theta})$).

Let us recall that (according to what we have seen during our study of Bayesian probabilities page 486) if we multiply the univariate likelihood $P(y|\vec{\theta})$ and the univariate prior probability $P(\vec{\theta})$ we get the univariate posterior distribution $P(\vec{\theta}|y)$ (also denoted $\pi(\vec{\theta}|y)$). Indeed, using what we have seen earlier above we get (the denominator is for recall the "marginal probability" or also sometimes named the "evidence"):

$$\boxed{\text{Posterior}=P(\vec{\theta}|y)=\frac{\text{likelihood}\cdot\text{prior}}{\text{marginal}}=\frac{P(y|\vec{\theta})P(\vec{\theta})}{P(y)}\propto P(y|\vec{\theta})P(\vec{\theta})} \tag{5.7.1568}$$

where $P(y|\vec{\theta})P(\vec{\theta})$ is commonly named the "unnormalized posterior density".

Also often denoted (still in the univariate case and using the definition of marginal distribution):

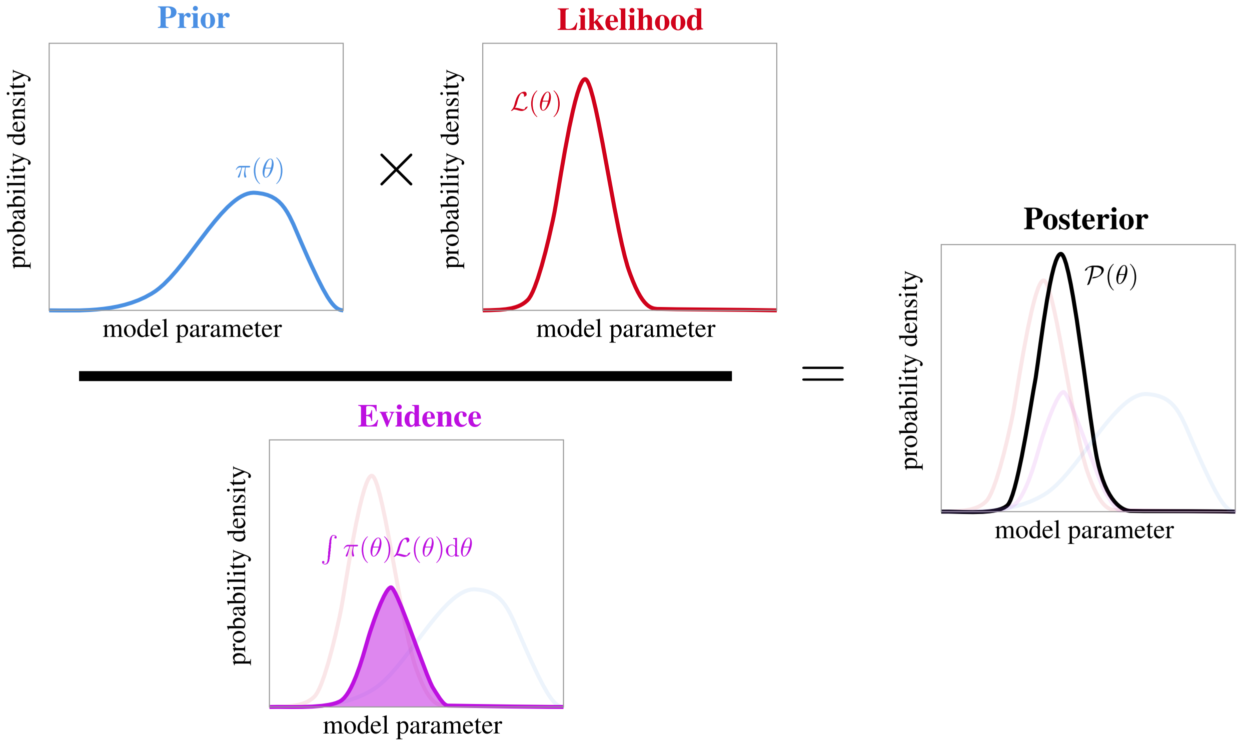
$$\boxed{\pi(\vec{\theta}|y)=\frac{\ell(y|\vec{\theta})f(\vec{\theta})}{f(y)}=\frac{\ell(y|\vec{\theta})f(\vec{\theta})}{\int f(y,\vec{\theta})\mathrm{d}p}\propto f(y|\vec{\theta})f(\vec{\theta})} \tag{5.7.1569}$$

Where the ratio:

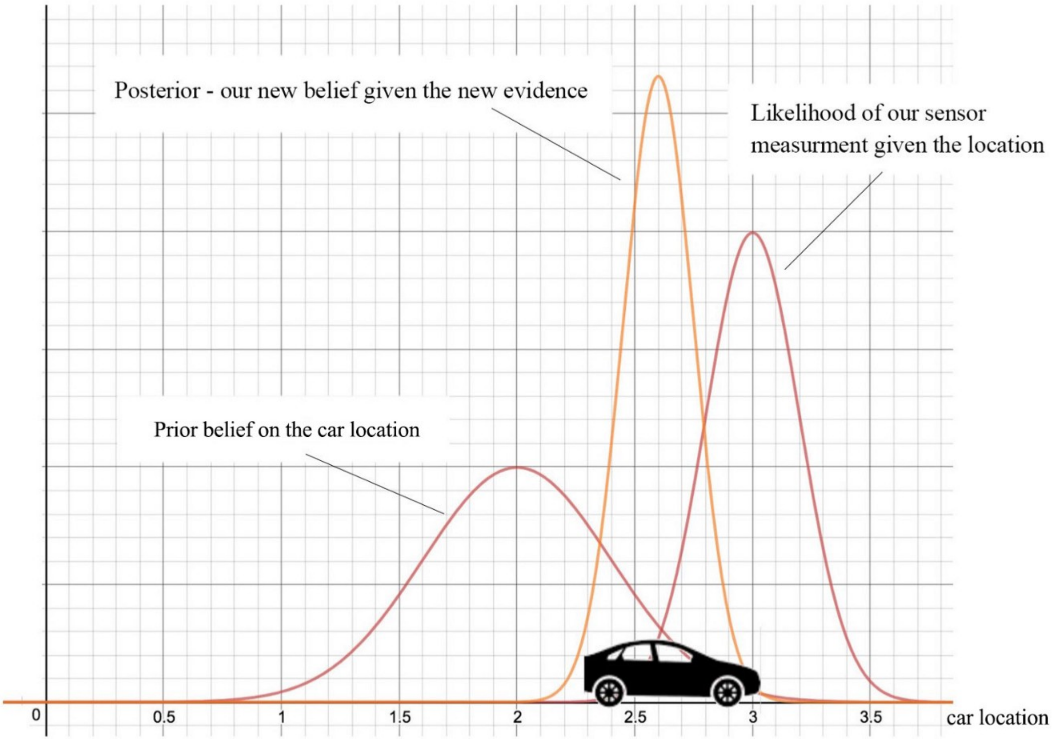
$$\frac{\ell(y|\vec{\theta})}{f(y)} \tag{5.7.1570}$$

is sometimes named the "standardized likelihood".

With the exception of the marginal probability, posterior, prior and likelihood are not a point estimate. They are curves as illustrated below:


Prior
probability density
$\pi(\theta)$
model parameter
×
Likelihood
probability density
$\mathcal{L}(\theta)$
model parameter
=
Posterior
probability density
$\mathcal{P}(\theta)$
model parameter
Evidence
probability density
$\int \pi(\theta)\mathcal{L}(\theta)\mathrm{d}\theta$
model parameter

Or, in the example below, we show the probability density function (PDF) on different car locations:



We strongly recommend the reader to see the examples below for a better understanding of the denominator.

Definition (#5.7.139): A prior $P(\vec{\theta})$ for a sampling model is named a "conjugate prior" if the resulting posterior denoted traditionally $\pi(\vec{\theta}|\vec{y})$ (for multivariate case) is in the same distributional family as the prior.

In other words: if our posterior distribution is of the same family as the prior distribution, we speak of "conjugation of distributions". Then we also say that the prior function is the "conjugation of the likelihood" (in practice it is however the exception rather than the rule ...).

Here is for example a table summarizing the most classic cases:

Likelihood	Prior	Posterior
Binomial	Beta	(Beta-)Binomial distribution
Binomial	Binomial	Binomial distribution
Negative Binomial	Beta	(Beta-)Negative Binomial distribution
Poisson	Gamma	(Poisson-)Gamma distribution
Geometric	Beta	(Beta-)Geometric distribution
Normal (unknown mean)	Normal	Normal distribution
Normal (unknown variance)	Inverse Gamma	(Normal-)Gamma inverse distribution
Normal (all moments known)	Normal/Gamma	Normal/Gamma distribution
Multinomial	Dirichlet	(Multinomial-)Dirichlet distribution

However (and as always!) in this textbook we will detail only the conjugations that have practical applications in other chapters!

A last reminder before continuing. If we know for a given likelihood distribution the expression $P(y_i/\vec{\theta})$, then for a vector of independent realizations of a random variable, the total likelihood is equal to the product of the probabilities:

$$P(\vec{y}|\vec{\theta}) = P(y_1, \ldots, y_i, \ldots, y_n|\vec{\theta}) = \prod_{i=1}^{n} P(y_i|\vec{\theta}) \tag{5.7.1571}$$

Be careful with Bayesian inference! Don't trust absolutely intuition, for even simple prior plus likelihood scenarios defy it! The quartet below shows that the choice of prior and likelihood distribution matters a lot. More specifically the "tailed-ness" of the distribution matters more!!

The author of this quartet depicts four cases:

- Normal prior, Normal likelihood (all is fine here)
- T-distribution prior, T-distribution likelihood (things start to go awry, we now have a bimodal posterior)
- T-distribution prior, Normal likelihood (the posterior mimics the likelihood and in a way, prior had absolutely no sway)
- Normal prior, T-distribution likelihood (the posterior mimics the prior and in a way likelihood had absolutely no sway)

Therefore the Bayesian inference requires a lot of hardwork and knowledge to get it right.

Richard McElreath's Quartet

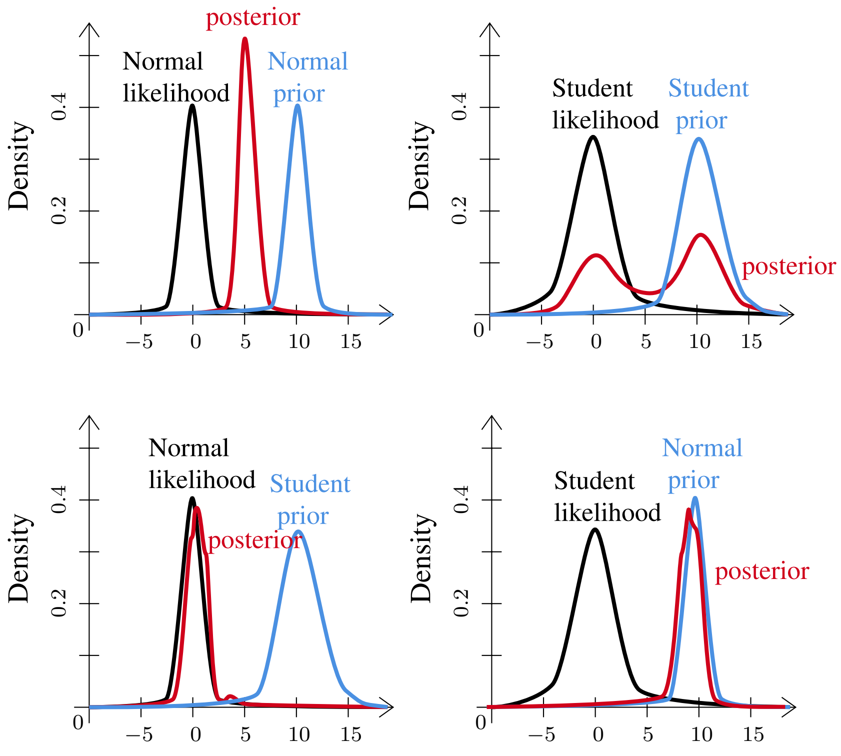


Figure 5.166: Richard McElreath's Quartet

5.7.9.2.1 Beta prior conjugate to Binomial likelihood (beta-binomial distribution)

Let us now introduce the beta-binomial distribution that is used a lot in Engineering but also in Bayesian statistics (especially known for the famous A/B test for example!).

We decide to start with this distribution have a binomial likelihood because it is a law with only one parameter and pedagogically it seems more judicious to start with that latter rather that with a Normal likelihood which has two parameters (furthermore there are three possible conjugates for the Normal likelihood as shows in the above table).

Definition (#5.7.140): The "beta-binomial distribution" is the binomial distribution in which the probability p of success at each trial is fixed but randomly drawn from a beta distribution prior to n Bernoulli trials.

Let us now prove that the Beta distribution is a conjugate distribution of the binomial distribution, but before a definition!

Suppose we observe n independent Bernoulli random variable $x_1,\ldots,x_n$. We wish to estimate the success probability p with the Bayesian approach. We will use a beta $P_{a,b}$ priori for p (as $P_{a,b}\in]0,1[$ it may be a good choice...) and prove this is a conjugate prior.

First using the relationship between the Euler Gamma function and the factorial (see section Differential and Integral Calculus page 1574) we have using the conventional notation (since the probability of the prior is assumed independent of the trial itself we then multiply both probabilities):

$$\begin{aligned}\pi(\vec{\theta}|y)&\propto P(y|\vec{\theta})P(\vec{\theta})=P(y|p,n)P(p,a,b)=\mathcal{B}(n,y)P_{a,b}=\mathcal{B}(n,y)\frac{p^{a-1}(1-p)^{b-1}}{B(a,b)}\\&=\left(\binom{n}{y}p^y(1-p)^{n-y}\right)\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}p^{a-1}(1-p)^{b-1}\right)\\&=\frac{n!}{y!(n-y)}p^y(1-p)^{n-y}\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}p^{a-1}(1-p)^{b-1}\right)\\&=\frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}p^{y+a-1}(1-p)^{n-y+b-1}\end{aligned} \tag{5.7.1572}$$

Notice that we also have the following equivalent result (just the notation differs!):

$$\begin{aligned}\pi(p(a,b)|y,n)&\propto P(y,n|p(a,b))P(p(a,b))=\mathcal{B}(n,y)P_{a,b}\\&=\left(\binom{n}{y}p^y(1-p)^{n-y}\right)\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}p^{a-1}(1-p)^{b-1}\right)\\&=\binom{n}{y}\frac{p^{y+a-1}(1-p)^{n-y+b-1}}{B(a,b)}\end{aligned} \tag{5.7.1573}$$

Although it is not really necessary, let us derive the marginal density of y (in purpose to normalize the posterior density further below!):

$$\begin{aligned}P(y)&=\int\limits_0^1 f(y,n,p(a,b))\mathrm{d}p\\&=\frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int\limits_0^1 p^{y+a-1}(1-p)^{n-y+b-1}\mathrm{d}p\\&=\frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(y+a)\Gamma(n-y+b)}{\Gamma(n+a+b)}\int\limits_0^1\frac{\Gamma(n+a+b)}{\Gamma(y+a)\Gamma(n-y+b)}p^{y+a-1}(1-p)^{n-y+b-1}\mathrm{d}p\\&=\frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(y+a)\Gamma(n-y+b)}{\Gamma(n+a+b)}\end{aligned} \tag{5.7.1574}$$

That latter marginal density is also sometimes written:

$$\begin{aligned}P(y)&=\int\limits_0^1 f(y,p(a,b))\mathrm{d}p=\int\limits_0^1 \mathcal{B}(n,y)P_{a,b}\mathrm{d}p\\&=\int\limits_0^1\binom{n}{y}\frac{p^{y+a-1}(1-p)^{n-y+b-1}}{B(a,b)}\mathrm{d}p=\binom{n}{y}\frac{\int\limits_0^1 p^{y+a-1}(1-p)^{n-y+b-1}\mathrm{d}p}{B(a,b)}\end{aligned} \tag{5.7.1575}$$

So a common (wrong) definition of the beta-binomial distribution is given by these two equivalent relations of the marginal density:

$$\boxed{P(y)=\frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(y+a)\Gamma(n-y+b)}{\Gamma(n+a+b)}=\binom{n}{y}\frac{\int\limits_0^1 p^{y+a-1}(1-p)^{n-y+b-1}\mathrm{d}p}{B(a,b)}} \tag{5.7.1576}$$

Now let us focus on the posterior $\pi(\vec{\theta},y)$ (i.e. the "real" expression of the beta-binomial distribution), then given by:

$$\begin{aligned}\pi(p(a,b)|y,n)&=\frac{P(y|p(a,b))P(p(a,b))}{P(y)}=\frac{\dfrac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)}\dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}p^{y+a-1}(1-p)^{n-y+b-1}}{\dfrac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)}\dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\dfrac{\Gamma(y+a)\Gamma(n-y+b)}{\Gamma(n+a+b)}}\\&=\frac{\Gamma(n+a+b)}{\Gamma(y+a)\Gamma(n-y+b)}p^{y+a-1}(1-p)^{n-y+b-1}=\frac{p^{y+a-1}(1-p)^{n-y+b-1}}{B(y+a,n-y+b)}\end{aligned} \tag{5.7.1577}$$

That latter equality is sometimes denoted:

$$\pi(p(a,b)|y,n)=\frac{p^{s+\alpha-1}(1-p)^{f+\beta-1}}{B(s+\alpha,f+\beta)} \tag{5.7.1578}$$

Clearly this posterior is a beta distribution $P_{a,b}(y+a, n-y+b)$. So we see that only the parameters of the beta prior change into now "hyper-parameters" of a beta posterior! In other words, this means that if the likelihood function is binomial, then a beta prior gives a beta posterior!

Notice that we then quickly get by extension (when drawing N items among the n):

$$\begin{aligned}\pi(p(a,b)|\vec{y},n) &= \prod_{i=1}^{N} \frac{\Gamma(n+a+b)}{\Gamma(y_i+a)\Gamma(n-y_i+b)} p^{y_i+a-1}(1-p)^{n-y_i+b-1} \\ &= P_{a,b}\left(\sum_i^N y_i + a, nN - \sum_i^N y_i + b\right)\end{aligned} \tag{5.7.1579}$$

That latter relation is sometimes denoted $\text{Bbin}(a,b,n)$.

Remark

The worst is that some authors in the field of reliability engineering denote the cumulated probability that $y < n - F$ where F is the number of failures as following (see section Industrial Engineering page 9059):

$$P_{a,b}\left((n-F)+1, n-(n-F)+b\right) = B(n-F+1, F+b) \tag{5.7.1580}$$

Then it does not become obvious at all out of context to know if we are dealing with a mass density or distribution function.

An interesting case to consider is when we have no particular premonition about prior distribution. Therefore, we can take a function that has the same probability for all values in the $[0,1]$ interval of the p parameter of the binomial distribution.

Therefore, the beta function can always be kept as a distribution a priori since it is reduced to the uniform case:

$$P_{a,b} = P_{1,1} = \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)} p^{1-1}(1-p)^{1-1} 1_{]0,1[} = 1 \cdot 1_{]0,1[} \tag{5.7.1581}$$

which is therefore a continuous uniform distribution in the interval $[0,1]$. Hence the posterior function of the parameter p is given by:

$$\pi(p(1,1)|\vec{y},n) = P_{1,1}\left(\sum_i^N y_i + 1, nN - \sum_i^N y_i + 1\right) \tag{5.7.1582}$$

Let us calculate now the expected mean and variance of the beta binomial distribution using[48] the law of total expectation (see page 501) and the law of total variance (page 502):

$$\text{E}(Y) = \text{E}(\text{E}(Y|\pi)) \qquad \text{V}(Y) = \text{E}(\text{V}(Y|\pi)) + \text{V}(\text{E}(Y|\pi)) \tag{5.7.1583}$$

Expectation and variance for the binomial distribution are assumed to be so well known that we will not mention them here (see page 658), and for the beta distribution we have (see page 760):

$$\text{E}(\pi) = \frac{a}{a+b} \qquad \text{V}(\pi) = \frac{ab}{(a+b)^2(a+b+1)} \tag{5.7.1584}$$

---

[48] There are other (complicated) ways to derive expectation and variance, but if you remember these relations, then you have as far as we know the simplest derivation!

Using the law of total expectation , we have:

$$\mathrm{E}(Y)=\mathrm{E}(\mathrm{E}(Y|\pi))=\mathrm{E}(np)=n\mathrm{E}(p)=n\mathrm{E}(\pi)=\frac{na}{a+b} \tag{5.7.1585}$$

and using the law of total variance, we have (in the second line we used the Huygens relation):

$$\begin{aligned}\mathrm{V}(Y)&=\mathrm{E}(\mathrm{V}(Y|\pi))+\mathrm{V}(\mathrm{E}(Y|\pi))=\mathrm{E}(np(1-p))+\mathrm{V}(np)\\&=n\mathrm{E}(p)-n\mathrm{E}(p^2)+n^2\mathrm{V}(p)=n\left(\mathrm{E}(p)-\mathrm{V}(p)-\mathrm{E}(p)^2+n\mathrm{V}(p)\right)\\&=n\left(\frac{a}{a+b}-\frac{ab}{(a+b)^2(a+b+1)}-\left(\frac{a}{a+b}\right)^2+n\frac{ab}{(a+b)^2(a+b+1)}\right)\\&=n\left(\frac{a(a+b)(a+b+1)}{(a+b)^2(a+b+1)}-\frac{ab}{(a+b)^2(a+b+1)}-\frac{a^2(a+b+1)}{(a+b)^2(a+b+1)}+n\frac{ab}{(a+b)^2(a+b+1)}\right)\\&=n\frac{a(a+b)(a+b+1)-ab-a^2(a+b+1)+nab}{(a+b)^2(a+b+1)}\\&=n\frac{(a^2+ab)(a+b+1)-ab-a^3-a^2b-a^2+nab}{(a+b)^2(a+b+1)}\\&=n\frac{a^3+a^2b+a^2+a^2b+ab^2+ab-ab-a^3-a^2b-a^2+nab}{(a+b)^2(a+b+1)}\\&=n\frac{a^2b+ab^2+nab}{(a+b)^2(a+b+1)}=\frac{nab(a+b+n)}{(a+b)^2(a+b+1)}\end{aligned} \tag{5.7.1586}$$

So to summarize we have for posterior expectation and posterior variance:

$$\mathrm{E}(Y)=\frac{na}{a+b}\qquad\mathrm{V}(Y)=\frac{nab(a+b+n)}{(a+b)^2(a+b+1)} \tag{5.7.1587}$$

If we reintroduce p, this can we be written:

$$\mathrm{E}(Y)=np\qquad\mathrm{V}(Y)=np(1-p)\frac{a+b+n}{a+b+1} \tag{5.7.1588}$$

We notice that the variance is higher than for the corresponding binomial distribution. We say then that there is "over-dispersion".

Remark

In case of $a=b=0.5$ for the beta distribution, it is quite common to speak about the "Jeffreys prior".

5.7.9.2.1.1 Uninformative prior

When prior distributions have no population basis, they can be difficult to construct, and there has long been a desire for prior distributions that can be guaranteed to play a minimal role in the posterior distribution. Such distributions are sometimes named "reference prior distributions", and the prior density is described as "uninformative prior" or "noninformative prior". The rationale for using noninformative prior distributions is often said *to be to let the data speak for themselves*, so that inferences are unaffected by information external to the current data.

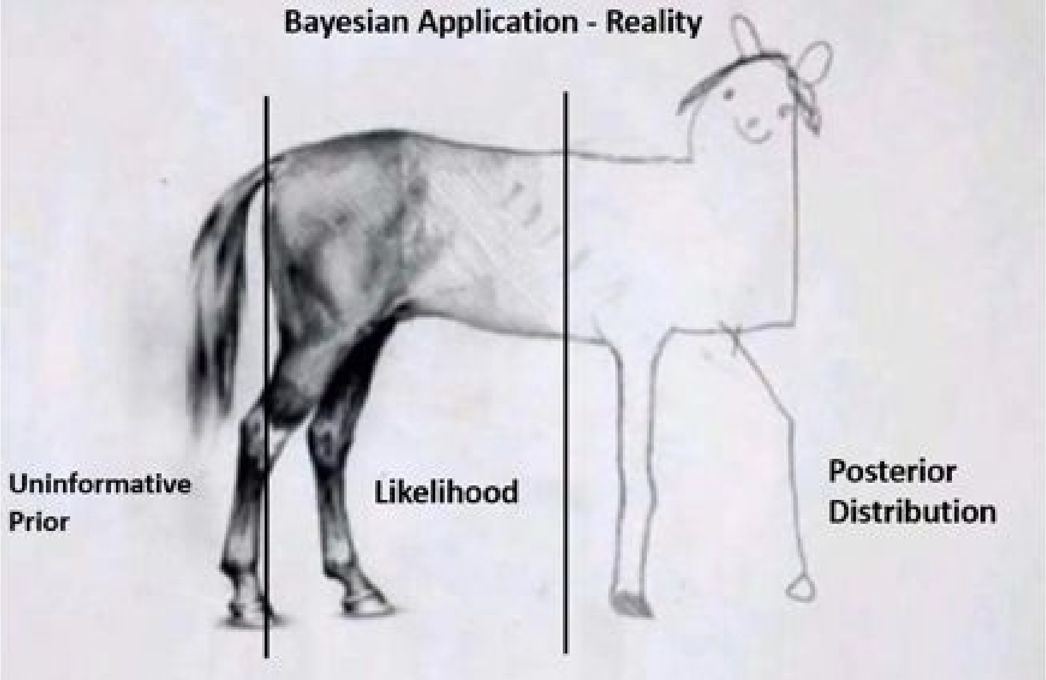


Figure 5.167: **Depending on the choice of the uninformative prior and likelihood sometimes things can go very wrong**

Some people would expect a uniform prior to be a good example of an uninformative prior and get the same result as the frequentist approach. However, this is not the case[49]!

As an example, let's look the classical Bernoulli problem - you flip coins and s are heads, f are tails.

We want to predict p, the probability of heads on any given coin flip.

The frequentist approach is to say that p is simply $s/(s+f)$. In this approach, we assume nothing about the distribution of p and we're simply finding the maximum likelihood estimate. This makes sense!

Now, let's adopt the Bayesian approach and assume we say:

$$P(p|\text{data}) = \frac{P(\text{data}|p)\cdot P(p)}{P(\text{data})} \qquad (5.7.1589)$$

As we have proved it earlier above at page 898 if we make our prior the beta distribution, we get:

$$P(p|\text{data}) = \frac{p^{s+\alpha-1}(1-p)^{f+\beta-1}}{B(s+\alpha,f+\beta)} \qquad (5.7.1590)$$

---

[49] For reasons unknown to us yet, even String theory physicists seems to use the uniform distribution as "natural choice" (naturalness criteria) for uninformative prior. A first attempt to justify the uniform probability distribution for the naturalness criterion might be to say that it doesn't introduce additional parameters into the String theory. But of course it does: it introduces the number 1 as a typical width! Some physicists, however argue without evidence yet, that 1 satisfy naturalness criterion...

Now let's say we want to find the expected value of p:

$$\begin{aligned}\mathrm{E}(p|\text{data})&=\int\limits_0^1 P(p|\text{data})p\mathrm{d}p=\int\limits_0^1 \frac{p^{s+\alpha-1}(1-p)^{f+\beta-1}}{B(s+\alpha,f+\beta)}p\mathrm{d}p\\&=\int\limits_0^1 \frac{p^{s+\alpha}(1-p)^{f+\beta-1}}{B(s+\alpha,f+\beta)}\mathrm{d}p=\frac{B(s+\alpha+1,f+\beta)}{B(s+\alpha,f+\beta)}\end{aligned} \tag{5.7.1591}$$

We fall back obviously on the same expression of the expected mean as the beta distribution (as it is a conjugate prior!). Hence we can do the same simplification:

$$\mathrm{E}(p|\text{data})=\frac{B(s+\alpha+1,f+\beta)}{B(s+\alpha,f+\beta)}=\frac{s+\alpha}{s+\alpha+f+\beta} \tag{5.7.1592}$$

Makes sense so far! Now, let's say we want to choose a really "objective" prior to emulate the frequentist approach, like a uniform distribution. This is equivalent to $B(1,1)$ as we have proved it during our study of the Beta distribution. In this case, we would get:

$$\mathrm{E}(p|\text{data};\alpha=1,\beta=1)=\frac{s+1}{s+f+2} \tag{5.7.1593}$$

This may hurts the brain of some people! Even though we're being as objective as we can ("p is equally likely to be anything"), we're getting a different result than the frequentist approach.

On the flip side, if we choose what we may consider as an improper prior like $B(0,0)$, we get the same result as the frequentist approach:

$$\mathrm{E}(p|\text{data};\alpha=0,\beta=0)=\frac{s}{s+f} \tag{5.7.1594}$$

What makes $B(0,0)$ more similar to the frequentist approach than $B(1,1)$? What assumption are we making in $B(1,1)$? How is $B(0,0)$ "less informed"? In fact you can't ask this question because it is simply inconsistent by definition!

There is no objective, unique prior that represents ignorance. Instead, reference priors are chosen by public agreement, much like units of length and weight. In this interpretation, reference priors are akin to a default option in a computer package. We fall back to the default when there is insufficient information to otherwise define the prior.

Let us briefly describe below the following classes of non-informative priors:

- **Bayes-Laplace uniform prior**: It's the prior we have introduce just earlier above and that are from Bayes and Laplace (118 and 119th century according to holocene calendar).
- **Jeffreys' prior**: Jeffreys (11946 according to holocene calendar) proposed a method to assign priors that is consistent under re-parametrization (we will give more details about this prior the day we will need it for a practical application in this textbook).
- **Jaynes' maximum entropy prior**: Jaynes (11968 according to holocene calendar) proposed another method to elicit uninformative priors. He argued that the probability distribution which best represents a given state of knowledge can be found by maximizing the entropy of the distribution (see section Statistical Mechanics page 3530), subject to some constraints representing the available information.

- **Bernardo's reference prior**: Bernardo (11979 according to holocene calendar) proposed that as the prior should influence the posterior as little as possible for this to happen, the posterior should be on average very different from the prior. Therefore, a reference prior is derived by maximizing the expected value of the Kullback-Leibler divergence (see section Statistical Mechanics page 3542) of the prior from the posterior.

Definitions (#5.7.141):

D1. A distribution is named a "flat prior", if its precision is zero or variance is infinite (in bayesian inference the "precision" is by definition the inverse of the variance).

D2. A distribution is named a "hard prior", if its precision is infinite or variance is zero (in bayesian inference the "precision" is by definition the inverse of the variance).

D3. A distribution is named a "soft prior", if its precision or variance are non-zero and finite (in bayesian inference the "precision" is by definition the inverse of the variance).

D4. A distribution is named an "empirical prior", if it has been derived from empirical data.

D5. A distribution is named an "theoretical prior", if it was specified without regard from empirical data.

5.7.9.2.2 Normal prior conjugate to Normal likelihood

We will focus here only on the case of the unknown mean because in practice it is the most used one (as far as we know) and it is also according to our personal point of view the only case that is intuitive to explain pedagogically speaking...

Let us consider that we want to make inference on the mean. We have the presumption that the measured random variable y probably follows a Normal distribution of mean and standard deviation $\mathcal{N}(y|\mu,\sigma)$ (i.e. the likelihood of y) and with the prior that μ follows a Normal distribution $\mathcal{N}(\mu_0,\sigma_0)$, which can be considered as a good choice as long as the possible values of μ correspond to the definition domain of the Normal distribution.

We have then (remember that in the prior distribution it is the parameter of interest which is the random variable!):

$$
\begin{aligned}
\pi(\mu|\vec{y},\sigma) &\propto P(\vec{y}|\mu,\sigma)P(\mu_0,\sigma_0) = \mathcal{N}(\vec{y}|\mu,\sigma)\mathcal{N}(\mu_0,\sigma_0)\\
&= \prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2}\right)\right]\cdot\frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)\\
&\propto \exp\left(\sum_{i=1}^{n}-\frac{(y_i-\mu)^2}{2\sigma^2}-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right) = \exp\left(-\frac{1}{2}\left[\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma^2}+\frac{(\mu-\mu_0)^2}{\sigma_0^2}\right]\right)\\
&= \exp\left(-\frac{1}{2\sigma^2\sigma_0^2}\left[\sigma_0^2\sum_{i=1}^{n}(y_i-\mu)^2+\sigma^2(\mu-\mu_0)^2\right]\right)\\
&= \exp\left(-\frac{1}{2\sigma^2\sigma_0^2}\left[\sigma_0^2\sum_{i=1}^{n}(y_i^2-2\mu y_i+\mu^2)+\sigma^2(\mu^2-2\mu\mu_0+\mu_0^2)\right]\right)\\
&= \exp\left(-\frac{1}{2\sigma^2\sigma_0^2}\left[\sigma_0^2\sum_{i=1}^{n}y_i^2-2\sigma_0^2\mu\sigma_{i=1}^{n}y_i+\sigma_0^2 n\mu^2+\sigma^2\mu^2-2\sigma^2\mu\mu_0+\sigma^2\mu_0^2\right]\right)\\
&= \exp\left(-\frac{1}{2\sigma^2\sigma_0^2}\left[\sigma_0^2\sum_{i=1}^{n}y_i^2-2\sigma_0^2\mu\frac{n}{n}\sum_{i=1}^{n}y_i+\sigma_0^2 n\mu^2+\sigma^2\mu^2-2\sigma^2\mu\mu_0+\sigma^2\mu_0^2\right]\right)\\
&= \exp\left(-\frac{1}{2\sigma^2\sigma_0^2}\left[\sigma_0^2\sum_{i=1}^{n}y_i^2-2\sigma_0^2\mu n\bar{y}+\sigma_0^2 n\mu^2+\sigma^2\mu^2-2\sigma^2\mu\mu_0+\sigma^2\mu_0^2\right]\right)
\end{aligned}
\tag{5.7.1595}
$$

We group and rearrange some terms:

$$
\begin{aligned}
\pi(\mu|\vec{y},\sigma) &\propto \exp\left(-\frac{1}{2\sigma^2\sigma_0^2}\left[\mu^2(\sigma^2+\sigma_0^2 n)-2\mu(\sigma^2\mu_0+\sigma_0^2 n\bar{y})+\sigma^2\mu_0^2+\sigma_0^2\sum_{i=1}^{n}y_i^2\right]\right)\\
&= \exp\left(-\frac{1}{2}\left[\mu^2\left(\frac{\sigma^2+\sigma_0^2 n}{\sigma^2\sigma_0^2}\right)-2\mu\left(\frac{\sigma^2\mu_0+\sigma_0^2 n\bar{y}}{\sigma^2\sigma_0^2}\right)+\frac{\sigma^2\mu_0^2+\sigma_0^2\sum_{i=1}^{n}y_i^2}{\sigma^2\sigma_0^2}\right]\right)\\
&= \exp\left(-\frac{1}{2}\left[\mu^2\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)-2\mu\left(\frac{\mu_0}{\sigma_0^2}+\frac{n\bar{y}}{\sigma^2}\right)+\frac{\mu_0^2}{\sigma_0^2}+\frac{\sum_{i=1}^{n}y_i^2}{\sigma^2}\right]\right)
\end{aligned}
\tag{5.7.1596}
$$

We now multiply the inside part of the square brackets by:

$$
\frac{\dfrac{1}{\sigma_0^2}+\dfrac{n}{\sigma^2}}{\dfrac{1}{\sigma_0^2}+\dfrac{n}{\sigma^2}}
\tag{5.7.1597}
$$

in the purpose to eliminate the μ^2:

$$\begin{aligned}\pi(\mu|\vec{y},\sigma) &\propto \exp\left(-\frac{1}{2}\frac{\dfrac{1}{\sigma_0^2}+\dfrac{n}{\sigma^2}}{\dfrac{1}{\sigma_0^2}+\dfrac{n}{\sigma^2}}\left[\mu^2\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)-2\mu\left(\frac{\mu_0}{\sigma_0^2}+\frac{n\bar{y}}{\sigma^2}\right)+k\right]\right)\\ &=\exp\left(-\frac{1}{2}\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)\left[\mu^2\left(\frac{\dfrac{1}{\sigma_0^2}+\dfrac{n}{\sigma^2}}{\dfrac{1}{\sigma_0^2}+\dfrac{n}{\sigma^2}}\right)-2\mu\left(\frac{\dfrac{\mu_0}{\sigma_0^2}+\dfrac{n\bar{y}}{\sigma^2}}{\dfrac{1}{\sigma_0^2}+\dfrac{n}{\sigma^2}}\right)+\underbrace{\frac{\dfrac{\mu_0^2}{\sigma_0^2}+\dfrac{\sum_{i=1}^n y_i^2}{\sigma^2}}{\dfrac{1}{\sigma_0^2}+\dfrac{n}{\sigma^2}}}_{:=k}\right]\right)\end{aligned} \tag{5.7.1598}$$

where k does not depend on μ. Let us continue...:

$$\begin{aligned}\pi(\mu|\vec{y},\sigma) &\propto \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)\left[\mu^2-2\mu\left(\frac{\dfrac{\mu_0}{\sigma_0^2}+\dfrac{n\bar{y}}{\sigma^2}}{\dfrac{1}{\sigma_0^2}+\dfrac{n}{\sigma^2}}\right)+k\right]\right)\\ &=\exp\left(-\frac{1}{2}\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)\left[\mu-\left(\frac{\dfrac{\mu_0}{\sigma_0^2}+\dfrac{n\bar{y}}{\sigma^2}}{\dfrac{1}{\sigma_0^2}+\dfrac{n}{\sigma^2}}\right)\right]^2\right)\end{aligned} \tag{5.7.1599}$$

Assuming that it exist a k such that (...):

$$k=\left(\frac{\dfrac{\mu_0}{\sigma_0^2}+\dfrac{n\bar{y}}{\sigma^2}}{\dfrac{1}{\sigma_0^2}+\dfrac{n}{\sigma^2}}\right)^2 \tag{5.7.1600}$$

Remark

In other words we have to check that when using this tool that:

$$\left(\frac{\dfrac{\mu_0}{\sigma_0^2}+\dfrac{n\bar{y}}{\sigma^2}}{\dfrac{1}{\sigma_0^2}+\dfrac{n}{\sigma^2}}\right)^2=\frac{\dfrac{\mu_0^2}{\sigma_0^2}+\dfrac{\sum_{i=1}^n y_i^2}{\sigma^2}}{\dfrac{1}{\sigma_0^2}+\dfrac{n}{\sigma^2}} \tag{5.7.1601}$$

The posterior distribution thus looks indeed like the appearance of a Normal law:

$$\pi(\mu|\vec{y},\sigma)\propto\exp\left(-\frac{1}{2}\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)\left[\mu-\left(\frac{\dfrac{\mu_0}{\sigma_0^2}+\dfrac{n\bar{y}}{\sigma^2}}{\dfrac{1}{\sigma_0^2}+\dfrac{n}{\sigma^2}}\right)\right]^2\right) \tag{5.7.1602}$$

Remark

At the opposite of our study of the beta-binomial posterior distribution, we will not here divide by the marginal distribution that is anyway just a normalization constant and doesn't really interest us for now for the practical application we will have much later in this textbook.

The mean and the variance of the posterior distribution are then, by analogy, equal to:

$$\begin{aligned}\mu_p&=\frac{\dfrac{\mu_0}{\sigma_0^2}+\dfrac{n\bar{y}}{\sigma^2}}{\dfrac{1}{\sigma_0^2}+\dfrac{n}{\sigma^2}}=\frac{\dfrac{\mu_0\sigma^2\sigma_0^2}{\sigma_0^2}+\dfrac{n\bar{y}\sigma^2\sigma_0^2}{\sigma^2}}{\sigma^2+\sigma_0^2n}=\frac{\mu_0\sigma^2+n\bar{y}\sigma_0^2}{\sigma^2+\sigma_0^2n}=\frac{\sigma^2}{\sigma^2+\sigma_0^2n}\mu_0+\frac{\sigma_0^2}{\sigma^2+\sigma_0^2n}n\bar{y}\\ \sigma_p^2&=\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}=\frac{\sigma^2\sigma_0^2}{\sigma_0^2n+\sigma^2}\end{aligned} \tag{5.7.1603}$$

Let us analyse a little more closely how the average prior mean μ_0 and the posterior mean μ_p are related to each other:

$$\mu_p=\frac{\sigma^2}{\sigma^2+\sigma_0^2n}\mu_0+\frac{\sigma_0^2}{\sigma^2+\sigma_0^2n}n\bar{y} \tag{5.7.1604}$$

We can observe that when n increases, the empirical mean $\bar{y}$ (second term) dominates the prior mean μ_0. When the prior variance σ_0^2 decreases, the average a priori (first term) takes a predominant importance.

Recall that in practice the specialist will, based on his return of experience (REX), choose the prior $\{\mu_0,\sigma_0\}$ about μ_p (since that latter depends on them) and that the standard deviation σ is assumed to be known (it is the set of all these parameters that make a lot of practitioners prefer the frequentist approach...).

5.7.9.2.3 Binomial prior conjugate to Binomial likelihood

This case has been sent to us by a reader. We did agree do add it in the present book for two reasons! The first one is because the "conditional binomial" is an interesting case where the prior and likelihood are both discrete (and this is not always intuitive to deal with that), the second one is because it use a kind of notation that should be known to a maximum a people and we can encounter in other textbooks. Let's see that that!

Let X and Y be two random variables where Y is binomially distributed conditional on X:

$$Y\mid X\sim\text{Bin}(X,q) \tag{5.7.1605}$$

and X also follows a binomial distribution but with different success frequency:

$$X\sim\text{Bin}(n,p) \tag{5.7.1606}$$

Then, as we will prove it, the marginal distribution of Y unconditional on X is again a binomial distribution:

$$Y\sim\text{Bin}(n,pq) \tag{5.7.1607}$$

Indeed, to prove that remember we are interested in the probability that Y equals a number m. According to the law of marginal probability or the law of total probability, this probability can be expressed as:

$$P(Y=m)=\sum_{k=0}^{+\infty} P(Y=m \mid X=k)\cdot P(X=k) \tag{5.7.1608}$$

Since, by definitions of the prior and the likelihood, $P(X=k)=0$ when $k>n$ and $P(Y=m \mid X=k)=0$ when $k<m$, we have:

$$P(Y=m)=\sum_{k=m}^{n} P(Y=m \mid X=k)\cdot P(X=k) \tag{5.7.1609}$$

Now we can take the probability mass function of the binomial distribution (see section Statistics page 658) and plug it in for the terms in the sum above to get:

$$P(Y=m)=\sum_{k=m}^{n}\binom{k}{m}q^m(1-q)^{k-m}\cdot\binom{n}{k}p^k(1-p)^{n-k} \tag{5.7.1610}$$

Applying the binomial coefficient identity (see section Probabilities page 530):

$$\binom{n}{k}\binom{k}{m}=\binom{n}{m}\binom{n-m}{k-m} \tag{5.7.1611}$$

and rearranging the terms, we have:

$$P(Y=m)=\sum_{k=m}^{n}\binom{n}{m}\binom{n-m}{k-m}p^k q^m(1-p)^{n-k}(1-q)^{k-m} \tag{5.7.1612}$$

Now we partition $p^k=p^m\cdot p^{k-m}$ and pull all terms dependent on k out of the sum:

$$\begin{aligned}P(Y=m)&=\binom{n}{m}p^m q^m\sum_{k=m}^{n}\binom{n-m}{k-m}p^{k-m}(1-p)^{n-k}(1-q)^{k-m}\\&=\binom{n}{m}(pq)^m\sum_{k=m}^{n}\binom{n-m}{k-m}(p(1-q))^{k-m}(1-p)^{n-k}\end{aligned} \tag{5.7.1613}$$

Then we substitute $i=k-m$, such that $k=i+m$:

$$P(Y=m)=\binom{n}{m}(pq)^m\sum_{i=0}^{n-m}\binom{n-m}{i}(p-pq)^i(1-p)^{n-m-i} \tag{5.7.1614}$$

According to the binomial theorem (see section Calculus page 1364):

$$(x+y)^n=\sum_{k=0}^{n}\binom{n}{k}x^{n-k}y^k \tag{5.7.1615}$$

the sum in the prior-previous relation is therefore equal to:

$$\sum_{i=0}^{n-m}\binom{n-m}{i}(p-pq)^i(1-p)^{n-m-i}=((p-pq)+(1-p))^{n-m} \tag{5.7.1616}$$

Therefore:

$$\begin{aligned}P(Y=m)&=\binom{n}{m}(pq)^m\sum_{i=0}^{n-m}\binom{n-m}{i}(p-pq)^i(1-p)^{n-m-i}\\&=\binom{n}{m}(pq)^m(p-pq+1-p)^{n-m}\\&=\binom{n}{m}(pq)^m(1-pq)^{n-m}\end{aligned}\tag{5.7.1617}$$

which is the probability mass function of the binomial distribution with parameters n and pq, such that:

$$Y\sim\mathrm{Bin}(n,pq)\tag{5.7.1618}$$

5.7.9.2.4 Credibility Intervals

In Bayesian inference, a "credibility interval" also named "credible interval" is an interval within which an unobserved parameter value falls with a particular probability. It is an interval in the domain of a posterior probability distribution or a predictive distribution. The generalisation to multivariate problems is the credible region. Credible intervals are analogous to confidence intervals in frequentist statistics, although they differ on a philosophical basis: Bayesian intervals treat their bounds as fixed and the estimated parameter as a random variable, whereas frequentist confidence intervals treat their bounds as random variables and the parameter as a fixed value. Also, Bayesian credible intervals use (and indeed, require) knowledge of the situation-specific prior distribution, while the frequentist confidence intervals do not.

A Bayesian statistician would say that given his observed data, there is a 95% probability that the true value of θ falls within the credible region while the frequentist statistician would say there is a 95% cumulated probability that when he computes a confidence interval from data of this sort, the true value of θ will fall within it.

Technically the bayesian "credibility interval" of size $1-\alpha$ is an interval $[x_1, x_2]$, where:

$$\boxed{P(x_1 \leq \theta \leq x_2|D) = \int\limits_{x_1}^{x_2} \pi(p(\theta|D))\mathrm{d}\theta = 1-\alpha} \qquad (5.7.1619)$$

Note that it is the parameter and not the interval limits, conditioned on the data D, that is random in this framework. This makes a credibility interval more intuitive because now we can interpret $1-\alpha$ as a (posterior) probability given an observed data set. In the frequentist setting, a given confidence interval is either right or wrong since θ is fixed. The confidence level only tells us what proportion of confidence intervals would be correct if we would repeat our calculations using new independent samples from the same underlying distribution, which many non-statisticians find confusing.

Let us see a companion example well illustrated (by the way...) in our `R` companion book to get more intuition about different approach.

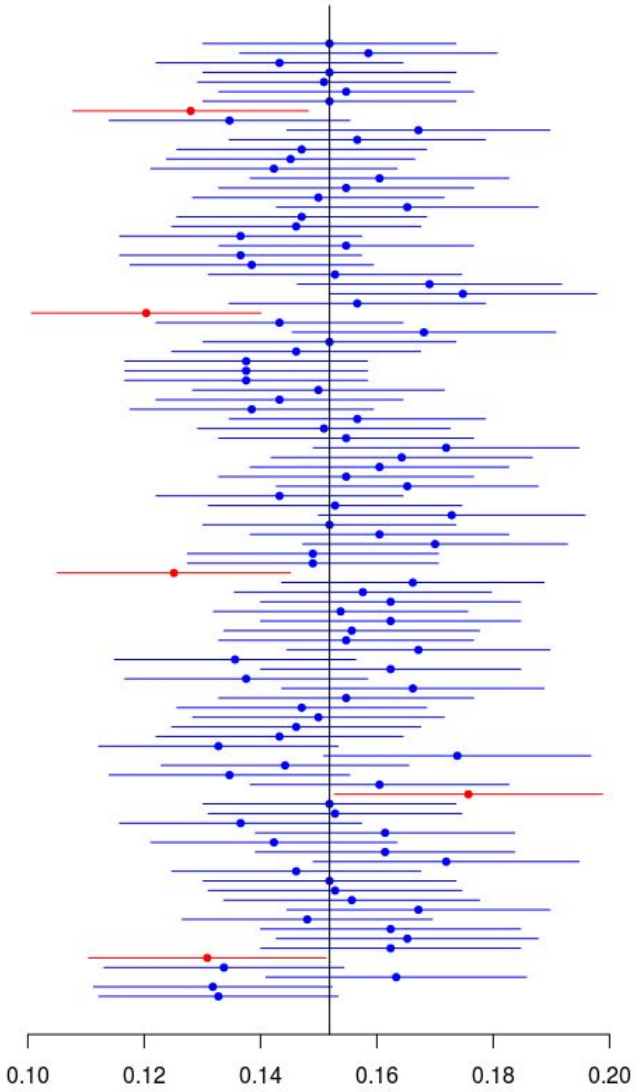
Consider a simple problem with Bernoulli trials in the field of insurance claims. We want to derive some confidence interval for the probability to claim a loss. There were $n=1047$ policies and in $y=159$ claims.

Consider the standard (frequentist) confidence interval of the proportion. What does that mean that:

$$\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad (5.7.1620)$$

is the (asymptotic) 95% confidence interval? The way we may see it is very simple. Let us generate some samples, of size n, with the same probability as the empirical one, i.e. $\hat{p}$ (which is the meaning of *from data of this sort*). For each sample, we compute the confidence interval with the relationship above. It is a 95% confidence interval because in 95% of the scenarios, the empirical value lies in the confidence interval.

If we write a R code that computes 100 of such intervals we get typically:



Now, what about the Bayesian credibility interval ? Assume that the prior distribution for the probability (denoted $\hat{p}$ in the frequentist framework) to claim a loss has a $\mathcal{B}(a,b)$ distribution. We have seen in the earlier above that, since the Beta distribution is the conjugate of the Bernoulli one, the posterior distribution will also be Beta. More precisely:

$$\pi(p(a,b)|\vec{y},n)=P_{a,b}\left(\sum_{i}^{N}y_i+a,nN-\sum_{i}^{N}y_i+b\right) \tag{5.7.1621}$$

Let us choose a uniform prior:

$$\pi(p(1,1)|\vec{y},n)=P_{1,1}\left(\sum_{i}^{N}y_i+1,nN-\sum_{i}^{N}y_i+1\right) \tag{5.7.1622}$$

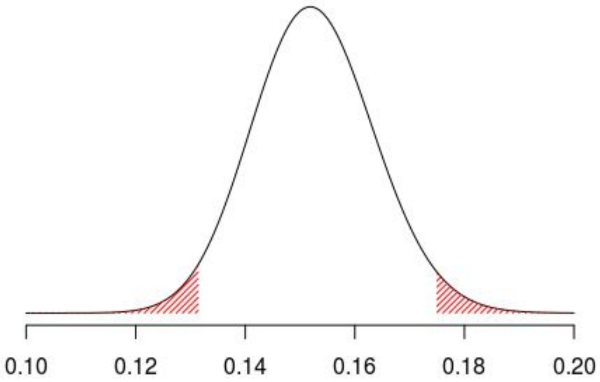
and let us consider we do the experiment only once ($N=1$):

$$\pi(p(1,1)|\vec{y},n)=P_{1,1}\left(y+1,n-y+1\right)=P_{1,1}\left(159+1,1047-159+1\right)=P_{1,1}(160,989) \tag{5.7.1623}$$

Based on that property, the credibility interval at 95% is based on quantiles of that (posterior) distribution such as:

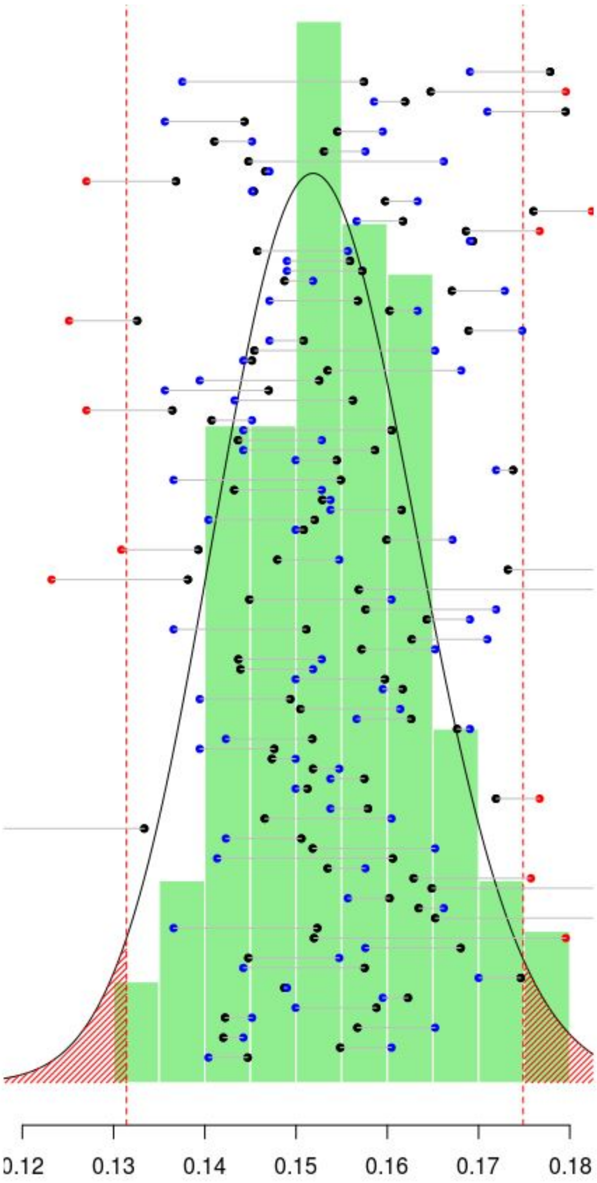
$$P(x_1 \leq \theta \leq x_2|D) = \int\limits_{x_1}^{x_2} \pi(p(\theta|D))\mathrm{d}\theta = \int\limits_{x_1}^{x_2} P_{1,1}(160, 989)\mathrm{d}\theta = 95\% \tag{5.7.1624}$$

and has the following shape if plotted with `R`:



It is straightforward with a statistical software to get the lower bound and upper bound quantile of the credibility interval.

We can now plot again 100 simulations but not generated using all the same estimated proportion as before, but using some possible probabilities (in green, below, we can visualize the histogram of those randomly chosen values), based on that posterior distribution (given the observations):



Here above there is 95% of probability that those empirical proportions lie in the credibility interval, defined using quantiles of the posterior distribution. We can actually visualize all those proportions: in black the proportion used to generate the sample, and then, in blue or red, the averages obtained on those simulated samples.

So what the reader has to remember is that in the Frequentist formulism (paradigm), the confidence intervals are random but the mean is not considered random but fixed. In Bayesian formulism (paradigm), the mean is considered a random variable but credible intervals are not random variables but are fixed!!!

5.7.9.2.5 Bayes Factor (BF)

In statistics, the use of "Bayes Factor" (i.e. "evidence" in experimental science) is a Bayesian alternative to classical p-value hypothesis testing (see page 948). Bayesian model comparison is a method of model selection based on Bayes factors. The models under consideration are statistical models. The aim of the Bayes factor is to quantify the support for a model over another, regardless of whether these models are correct. The technical definition of "support" in the context of Bayesian inference is described below.

Let us recall that according to Bayes' theorem, we can update prior probabilities of some model $\mathcal{M}$ ($P(\mathcal{M})$) to posterior probabilities ($P(\mathcal{M}|D)$) after observing some datum D by accounting for the probability of observing that datum given the model ($P(D|\mathcal{M})$, also known as the "likelihood"):

$$P(\mathcal{M}|D)=\frac{P(D|\mathcal{M})P(\mathcal{M})}{P(D)} \tag{5.7.1625}$$

Using this equation, We can compare the probability-odds of two models:

$$\frac{P(\mathcal{M}_1|D)}{P(\mathcal{M}_2|D)}=\frac{P(D|\mathcal{M}_1)}{P(D|\mathcal{M}_2)}\cdot\frac{P(\mathcal{M}_1)}{P(\mathcal{M}_2)} \tag{5.7.1626}$$

That latter is also often denoted:

$$\frac{P(H_0|D)}{P(H_1|D)}=\frac{P(D|H_0)}{P(D|H_1)}\cdot\frac{P(H_0)}{P(H_1)} \tag{5.7.1627}$$

Where the left-most term are the posterior odds, the right-most term are the prior odds, and the middle term is the Bayes factor:

$$\text{BF}_{12}=\frac{P(D|\mathcal{M}_1)}{P(D|\mathcal{M}_2)} \tag{5.7.1628}$$

That latter is also often denoted:

$$\text{BF}_{10}=\mathfrak{B}_{12}=\frac{P(D|H_0)}{P(D|H_1)} \tag{5.7.1629}$$

Thus, Bayes factors can be seen either as a ratio quantifying the relative probability of some observed data by two models as they can be computed by comparing the marginal likelihoods of the two models, or as the degree by which some prior beliefs about the relative credibility of two models are to be updated as they can be computed by dividing posterior odds by prior odds, as we will soon show it.

Obviously when the two models are equally probable a priori, so that $P(\mathcal{M}_1)=P(\mathcal{M}_2)$, the Bayes factor is equal to the ratio of the posterior probabilities of $\mathcal{M}_1$ and $\mathcal{M}_2$.

Bayes factor BF_{10}	**Label**
> 100	Extreme evidence for H_1
$30 - 100$	Very strong evidence for H_1
$10 - 30$	Strong evidence for H_1
$3 - 10$	Moderate evidence for H_1
$1 - 3$	Anecdotal evidence for H_1
1	No evidence
$1/3 - 1$	Anecdotal evidence for H_0
$1/3 - 1/10$	Moderate evidence for H_0
$1/10 - 1/30$	Strong evidence for H_0
$1/30 - 1/100$	Very strong evidence for H_0
$< 1/100$	Extreme evidence for H_0

Table 5.49: **Interpretations of the Bayes Factor**

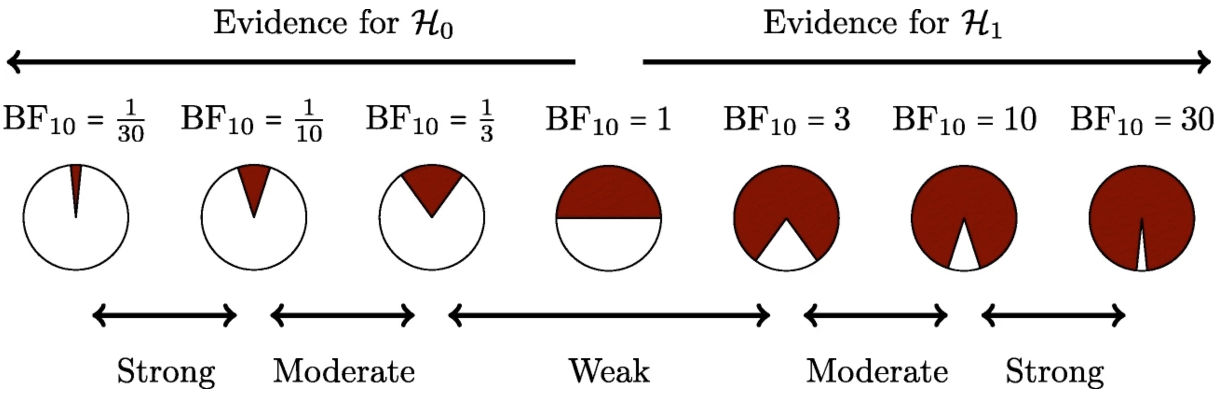

Evidence for $\mathcal{H}_0$
Evidence for $\mathcal{H}_1$
$\mathrm{BF}_{10} = \frac{1}{30}$
$\mathrm{BF}_{10} = \frac{1}{10}$
$\mathrm{BF}_{10} = \frac{1}{3}$
$\mathrm{BF}_{10} = 1$
$\mathrm{BF}_{10} = 3$
$\mathrm{BF}_{10} = 10$
$\mathrm{BF}_{10} = 30$
Strong
Moderate
Weak
Moderate
Strong

Figure 5.168: **A graphical representation of a Bayes factor classification table from the JASP guidelines for conducting and reporting a Bayesian analysis**

✍**Example:**

Suppose we have a random variable that produces either a success or a failure. We want to compare a model $\mathcal{M}_1$ where the probability of success is $q = 1/2$, and another model $\mathcal{M}_2$ where θ is unknown and we take a prior distribution for θ that is uniform on $[0, 1]$. We take a sample of 200, and find 115 successes and 85 failures. The likelihood can be calculated according to the binomial distribution:

$$\binom{200}{115}\theta^{115}(1-\theta)^{85} \tag{5.7.1630}$$

Thus we have for $\mathcal{M}_1$:

$$P\left(X = 115 \mid \mathcal{M}_1\right) = \binom{200}{115}\left(\frac{1}{2}\right)^{200} = 0.005956\ldots \tag{5.7.1631}$$

whereas for $\mathcal{M}_2$ we have (the development involves the beta distribution):

$$\begin{aligned}P\left(X=115 \mid \mathcal{M}_2\right) &= \int\limits_0^1 \binom{200}{115}\theta^{115}(1-\theta)^{85}\mathrm{d}\theta = \binom{200}{115}\cdot\int\limits_0^1 \theta^{115}(1-\theta)^{85}\mathrm{d}\theta \\ &= \binom{200}{115}\cdot B(116,86) = \binom{200}{115}\cdot\frac{\Gamma(116)\cdot\Gamma(86)}{\Gamma(116+86)} \\ &= \frac{200!}{115!\cdot 85!}\cdot\frac{115!\cdot 85!}{201!} = \frac{1}{201} = 0.004975\ldots\end{aligned} \tag{5.7.1632}$$

The ratio is then:

$$\begin{aligned}\mathrm{BF}_{12} &= \frac{P\left(X=115 \mid \mathcal{M}_1\right)}{P\left(X=115 \mid \mathcal{M}_2\right)} = \frac{\mathcal{B}(k,N,0.5)}{\int B(\theta,1,1)\mathcal{B}(k,N,\theta)} \\ &= \frac{\binom{N}{k}0.5^k\left(1-0.5\right)^{N-k}}{\int\limits_0^1\binom{N}{k}\theta^k\left(1-\theta\right)^{N-k}\mathrm{d}\theta} = \frac{\binom{N}{k}0.5^N}{\binom{N}{k}\int\limits_0^1\theta^k\left(1-\theta\right)^{N-k}\mathrm{d}\theta} \\ &= \frac{0.5^N}{B(k+1,N-k+1)} \cong 1.197\ldots\end{aligned} \tag{5.7.1633}$$

which is "barely worth mentioning" even if it points very slightly towards $\mathcal{M}_1$.

A frequentist hypothesis test of $\mathcal{M}_1$ (here considered as a null hypothesis) would have produced a very different result. Such a test says that $\mathcal{M}_1$ should be rejected at the 5% significance level, since the probability of getting 115 or more successes from a sample of 200 if $\theta = 1/2$ is 0.0200, and as a two-tailed test of getting a figure as extreme as or more extreme than 115 is 0.0400. Note that 115 is more than two standard deviations away from 100. Thus, whereas a frequentist hypothesis test would yield significant results at the 5% significance level, the Bayes factor hardly considers this to be an extreme result. Note, however, that a non-uniform prior (for example one that reflects the fact that we expect the number of successes and failures to be of the same order of magnitude) could result in a Bayes factor that is more in agreement with the frequentist hypothesis test.

We can see in this example that the Bayes factor (evidence) of H_0 to H_1 is then the ratio of the likelihood under H_0 to the average likelihood under H_1 (or "odds" of H_0 to H_1)!

A classical likelihood-ratio test would have found the maximum likelihood estimate for θ, namely $115/200 = 0.575$, whence:

$$P\left(X=115 \mid \mathcal{M}_2\right) = \binom{200}{115}\theta^{115}(1-\theta)^{85} = 0.056991 \tag{5.7.1634}$$

(rather than averaging over all possible θ). That gives a likelihood ratio of 0.1045 and points towards $\mathcal{M}_2$.

$\mathcal{M}_2$ is a more complex model than $\mathcal{M}_1$ because it has a free parameter which allows it to model the data more closely. The ability of Bayes factors to take this into account is a reason why Bayesian inference has been put forward as a theoretical justification for and generalisation of Occam's razor, reducing Type I errors.

On the other hand, the modern method of relative likelihood takes into account the number of free parameters in the models, unlike the classical likelihood ratio. The relative likelihood method could be applied as follows. Model $\mathcal{M}_1$ has 0 parameters, and so its AIC (see section Numerical Methods page 6420) value is $2\cdot 0-2\cdot\ln(0.005956)=10.2467$. The model $\mathcal{M}_2$ has 1 parameter, and so its AIC value is $2\cdot 1-2\cdot\ln(0.056991)=7.7297$. Hence $\mathcal{M}_1$ is about $\exp\left((7.7297-10.2467)/2\right)=0.284$ times as probable as $\mathcal{M}_2$ to minimize the information loss. Thus $\mathcal{M}_2$ is slightly preferred, but $\mathcal{M}_1$ cannot be excluded!

Notice from the above example we can deduce even a more general definition of the Bayes factor:

$$\boxed{\text{BF}_{12}=\dfrac{P(D|\mathcal{M}_1)}{P(D|\mathcal{M}_2)}=\dfrac{\int P(D|\theta_1,\mathcal{M}_1)P(\theta_1|\mathcal{M}_1)\mathrm{d}\theta_1}{\int P(D|\theta_2,\mathcal{M}_2)P(\theta_2|\mathcal{M}_2)\mathrm{d}\theta_2}} \tag{5.7.1635}$$

Or written in the form of the null and alternative hypothesis:

$$\text{BF}_{10}=\dfrac{P(D|H_0)}{P(D|H_1)}=\dfrac{\int P(D|\theta_1,H_0)P(\theta_1|H_0)\mathrm{d}\theta_1}{\int P(D|\theta_2,H_1)P(\theta_2|H_1)\mathrm{d}\theta_2} \tag{5.7.1636}$$

The Bayes factor (evidence) of H_0 to H_1 is then ratio of the average likelihood under H_0 to the average likelihood under H_1 (or "odds" of H_0 to H_1)!

In most cases this factor has however to be calculated using Monte Carlo Markov Chains (see section Numerical Methods page **??**).

As already mentioned during our study of the Cauchy distribution, when submitting a paper with Bayes Factors for publication, it is likely that the reviewers will ask the authors to justify their priors and to explicit it in their paper (otherwise a Bayes Factor may have been hacked!). Software such as JASP and the `R` package BayesFactor rely on the Cauchy distribution (with two parameters that are most of time very difficult to justify in practice) as a prior for many of the analyses.

Remark

To avoid hidden Bayes Factor hacking, softwares like JASP automatically provides a plot of the evidence for H_1 or H_0 for different width parameters of the Cauchy prior (named sometimes "sensitivity analysis"). In addition to the plot, JASP also provides a value denoted $\max$ BF10: this gives the Bayes Factor (for the alternative hypothesis) and the width parameter value of the Cauchy distribution (denoted r in JASP) where the largest Bayes Factor is found. Thus, conducting numerous analyses and reporting only the "best" one is a problem both for Frequentist and Bayesian inference. Bayes Factors do not provide a solution to this problem and also doesn't provide a solution to the false positive rate!!! So p-values (see page 948) and Bayes Factors share some of their pitfalls. Neither should be applied without some understanding of the underlying principles! Therefore a shift from Frequentist to Bayesian statistics is unlikely to be sufficient to alleviate the symptoms of the replication crisis...

Then in practice we just replace both numerator and denominator with Monte Carlo estimates. For the numerator we get:

$$\int P(D|\theta_1,\mathcal{M}_1)P(\theta_1|\mathcal{M}_1)\mathrm{d}\theta_1 = \mathrm{E}_{\mathcal{M}_1}\big(P(D|\theta,\mathcal{M}_1)\big) \cong \frac{1}{N}\sum_{j=1}^{N} P(D|\theta_1^{(j)},\mathcal{M}_1) \qquad (5.7.1637)$$

where the subscript $\mathcal{M}_1$ at the expectation indicates that θ in the expectation is distributed according to (the prior of) model 1 and, correspondingly, the $\theta_1^{(j)}$ on the right-hand side are independent and identically distributed samples from the prior of model 1. Using the same approximation for the denominator, we get:

$$\widehat{\mathrm{BF}}_{12} = \frac{\frac{1}{N}\sum_{j=1}^{N} P(D|\theta_1^{(j)},\mathcal{M}_1)}{\frac{1}{N}\sum_{j=1}^{N} P(D|\theta_2^{(j)},\mathcal{M}_2)} \qquad (5.7.1638)$$

The law of large numbers guarantees that this estimate converges to the correct value $\widehat{\mathrm{BF}}_{12}$ as the sample size N increases.

Caution! Their is some confusion between the frequentist and Bayesian paradigms because they use many of the same words to mean very different things. Name, "probability". Beyond Kolmogorov's axioms i would be a good idea to eliminate the word "probability" from the field of statistics. Under the frequentist paradigm, « *the probability...* » should only ever be referred to as, « *the long-run sampling proportion...* ». Likewise under the Bayesian paradigm, « *the probability...* » should only ever be referred as to as « *the unfalsifiable belief of the experimenter...* ». If the field of statistics makes this one simple change it can avoid a lot of nonsensical claims.

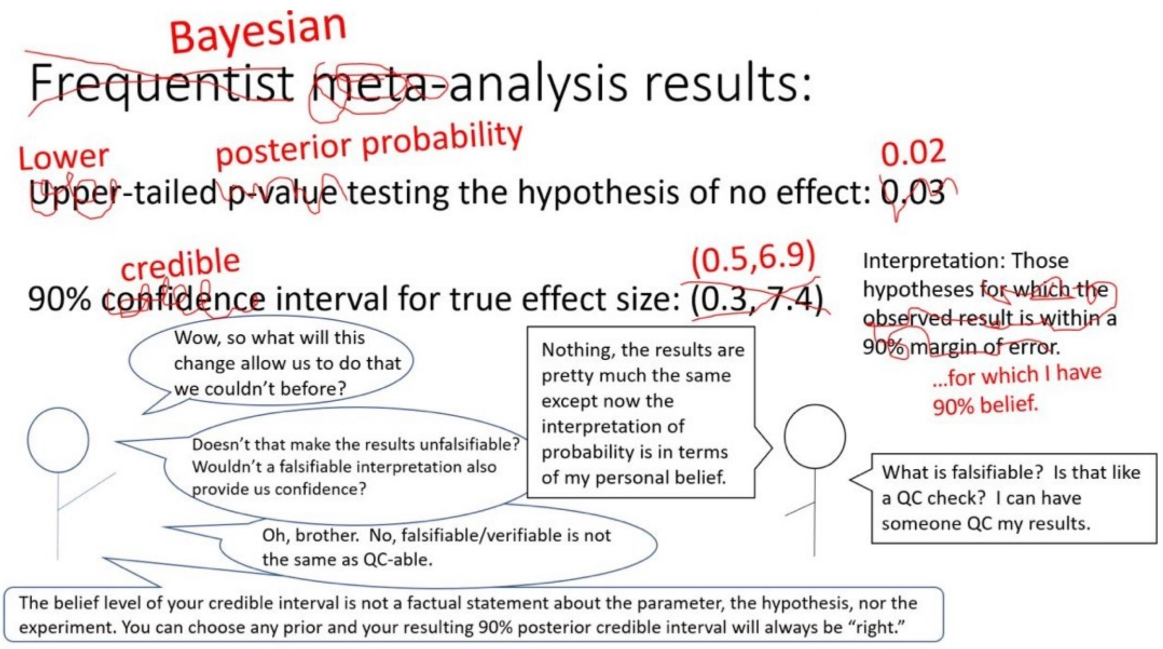
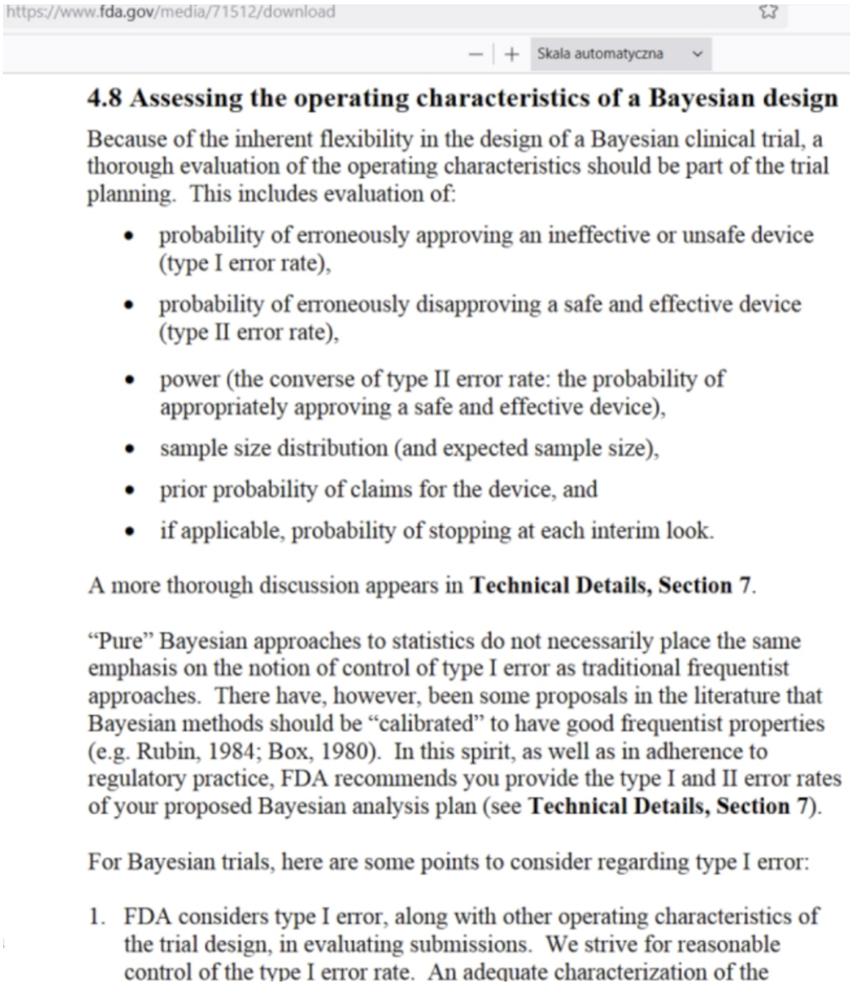

Bayesian
~~Frequentist~~ ~~meta-~~analysis results:
Lower
posterior probability
0.02
~~Upper~~-tailed ~~p-value~~ testing the hypothesis of no effect: ~~0.03~~
credible
(0.5,6.9)
90% ~~confidence~~ interval for true effect size: ~~(0.3, 7.4)~~
Interpretation: Those hypotheses ~~for which the observed result is within a 90% margin of error.~~
...for which I have 90% belief.
Wow, so what will this change allow us to do that we couldn't before?
Nothing, the results are pretty much the same except now the interpretation of probability is in terms of my personal belief.
Doesn't that make the results unfalsifiable? Wouldn't a falsifiable interpretation also provide us confidence?
What is falsifiable? Is that like a QC check? I can have someone QC my results.
Oh, brother. No, falsifiable/verifiable is not the same as QC-able.
The belief level of your credible interval is not a factual statement about the parameter, the hypothesis, nor the experiment. You can choose any prior and your resulting 90% posterior credible interval will always be "right."

Figure 5.169: **What it means to be "revolutionary" in statistics (author: Geoffrey Johnson)**

The Bayes Factor is used quite a lot in cutting edge fields of various sciences like: Social Science (to compare different models of behaviours), History (to quantify the existence of an individual), Cosmology (to compare various cosmological models), Particles Physics (to compare various models of fundamental particles). For more, see the excellent paper [307].

To close this topic let us provide a sample of the regulation of the FDA (Federal Drug Administration) relatively to clinical trials based on bayesian method because it's quite interesting (as a reader make your own conclusions...) :



4.8 Assessing the operating characteristics of a Bayesian design

Because of the inherent flexibility in the design of a Bayesian clinical trial, a thorough evaluation of the operating characteristics should be part of the trial planning. This includes evaluation of:

- probability of erroneously approving an ineffective or unsafe device (type I error rate),
- probability of erroneously disapproving a safe and effective device (type II error rate),
- power (the converse of type II error rate: the probability of appropriately approving a safe and effective device),
- sample size distribution (and expected sample size),
- prior probability of claims for the device, and
- if applicable, probability of stopping at each interim look.

A more thorough discussion appears in **Technical Details, Section 7**.

"Pure" Bayesian approaches to statistics do not necessarily place the same emphasis on the notion of control of type I error as traditional frequentist approaches. There have, however, been some proposals in the literature that Bayesian methods should be "calibrated" to have good frequentist properties (e.g. Rubin, 1984; Box, 1980). In this spirit, as well as in adherence to regulatory practice, FDA recommends you provide the type I and II error rates of your proposed Bayesian analysis plan (see **Technical Details, Section 7**).

For Bayesian trials, here are some points to consider regarding type I error:

1. FDA considers type I error, along with other operating characteristics of the trial design, in evaluating submissions. We strive for reasonable control of the type I error rate. An adequate characterization of the

5.7.10 Weak Law of Large Numbers (Khinchin's law)

We will now focus on a very interesting relation in statistics that can tell a lot of things while having little information and whatever the distribution (which is not bad!). This is a widely used property for example in statistical simulation in the context of the use of Monte Carlo techniques.

Given a random variable with values in $\mathbb{R}^+$. Then we will show the following relation named "Markov's inequality":

$$\boxed{\forall\lambda > 0 : P(X \geqslant \lambda) \leqslant \frac{\mathrm{E}(X)}{\lambda}} \tag{5.7.1639}$$

with in the particular context of probabilities $\mathrm{E}(X) \leq \lambda$.

In other words, we propose to prove that the probability that a random variable is greater than or equal to λ is less then or equal to its mean divided by the value λ considerated and regardless of the probability distribution of the random variable X!

Proof 5.47.10. Let us write the values of X by $(x_1, \ldots, x_n)$, where $0 \leq x_1 < x_2 < \ldots < x_n$ (that is to say sorted in ascending order) and let us also write $x_0 = 0$. We note first that the inequality is trivial in case where $\lambda \geq x_n \geq 0$. Indeed, as X can be included only between 0 and x_n by definition then the probability that it is greater to x_n is equal to zero. In other words:

$$P(X \geqslant \lambda) = 0 \tag{5.7.1640}$$

and X being positive, then $\mathrm{E}(X)$ is also positive, thus the inequality in this special case in a first time.

Otherwise, we have $0 < \lambda \leq x_n$ and then there exist one $k \in (1, \ldots, n)$ such that $x_{k-1} < \lambda \leq x_k$. Thus:

$$\mathrm{E}(X) = \sum_{i=1}^{n} x_i p_i \geqslant \sum_{i=k}^{n} x_i p_i \geqslant \sum_{i=k}^{n} p_i = \lambda P(X \geqslant \lambda) \tag{5.7.1641}$$

■ Q.E.D.

✍**Example:**

We assume that the number of outgoing parts of a given factory during one week is a random variable of mean 50. If we wish to estimate the cumulative probability that production exceeds 75 parts we will simply apply:

$$P(X \geqslant 75) \leqslant \frac{\mathrm{E}(X)}{\lambda} = \frac{50}{75} = 0.6666 = 66.66\% \tag{5.7.1642}$$

Consider now a kind of generalization of this inequality named "Bienayme-Chebyshev inequality" (abbreviated "BC inequality") that will allow us to get a very very interesting and important result a little bit later.

Consider a real random variable X (so we do not limit ourselves to the only cases where it is in $\mathbb{R}^+$). Then we will prove the following Bienayme-Chebyshev inequality:

$$\boxed{\forall\varepsilon>0:P(|X-\mathrm{E}(X)|\geqslant\varepsilon)\leqslant\frac{\mathrm{V}(X)}{\varepsilon^2}=\frac{\sigma^2}{\varepsilon^2}} \tag{5.7.1643}$$

which expresses the fact that the smaller the standard deviation is, more the probability that the random variable X moves away from its expectation is low.

Proof 5.47.11. We obtain this inequality by first writing:

$$\forall\varepsilon>0:P(|X-\mathrm{E}(X)|\geqslant\varepsilon)=P\left([X-\mathrm{E}(X)]^2\geqslant\varepsilon^2\right) \tag{5.7.1644}$$

where the choice of the square will serve us for a future simplification.

Then by applying Markov's inequality (as you see is something that can be useful ...) to the random variable $Y=[X-\mathrm{E}(X)]^2$ with $\lambda=\varepsilon^2$ it comes automatically:

$$P(Y\geqslant\varepsilon^2)=P\left([X-\mathrm{E}(X)]^2\geqslant\varepsilon^2\right)\leqslant\frac{\mathrm{E}(Y)}{\varepsilon^2}=\frac{\mathrm{E}\left([X-\mathrm{E}(X)]^2\right)}{\varepsilon^2} \tag{5.7.1645}$$

Then, using the definition of the variance:

$$\mathrm{E}\left([X-\mathrm{E}(X)]^2\right)=\mathrm{E}\left([X-\mu_X]^2\right)=\mathrm{V}(X)=\sigma^2 \tag{5.7.1646}$$

We get:

$$P(Y\geqslant\varepsilon^2)\leqslant\frac{\mathrm{E}\left([X-\mathrm{E}(X)]^2\right)}{\varepsilon^2}=\frac{\mathrm{V}(X)}{\varepsilon^2}=\frac{\sigma^2}{\varepsilon^2} \tag{5.7.1647}$$

■ Q.E.D.

If we put:

$$\varepsilon=t\sigma \tag{5.7.1648}$$

The equality will be written:

$$\boxed{P(|X-\mathrm{E}(X)|\geqslant t\sigma)\leqslant\frac{1}{t^2}} \tag{5.7.1649}$$

and expresses the cumulative probability in order that X moves away from the mean of more than t times its standard deviation, is below $1/t^2$. But for $t=2$, 3 and 4, we can deduce that there is, respectively, 75%, 89% and 94% of the data between the interval defined by 2, 3 and 4 standard deviations around the mean. There is, in particular, less than 1 chance on 9 that X moves away from its mean by more than three times the standard deviation. This is also this theorem that is used by the Basel Committee to define the Value at Risk correction factor used in finance (see section Economy page 7858).

✍**Example:**

We take the example where the number of outgoing parts of a given factory during one week is a random variable of mean 50. We assume in addition that the variance of the weekly production is 25. We seek to calculate the probability that the production of next week is between 40 and 60 pieces.

To calculate this we must first remember that the BT inequality is based partially on the term $|X - \text{E}(X)|$ so we have:

$$|40 - 50| = |60 - 50| = 10 \tag{5.7.1650}$$

Therefore the BT inequality allows us to work on equal intervals in absolute value which is also written:

$$-10 \leq X - \text{E}(X) \leq 10 \tag{5.7.1651}$$

Then we just have to apply numerical values to the inequality:

$$P(|X - \text{E}(X)| \geqslant \varepsilon) = P(|X - \text{E}(X)| \geqslant 10) = \frac{\text{V}(x)}{\varepsilon^2} = \frac{25}{10^2} = 0.25 = 25\% \tag{5.7.1652}$$

Because it can be applied to completely arbitrary distributions provided they have a known finite mean and variance, the inequality generally gives a poor bound compared to what might be deduced if more aspects are known about the distribution involved:

t	**Min. % within t standard deviations of mean**	**Max. % beyond t standard deviations from mean**
1	0%	100%
$\sqrt{2}$	50%	50%
1.5	55.56%	44.44%
2	75%	25%
$2\sqrt{2}$	87.5%	12.5%
3	88.8889%	11.1111%
4	93.75%	6.25%
5	96%	4%
6	97.2222%	2.7778%
7	97.9592%	2.0408%
8	98.4375%	1.5625%
9	98.7654%	1.2346%
10	99%	1%

Table 5.50: **Chebychev typical intervals to the mean**

There is another inequality sometimes useful in practice. Let us quickly prove it!

Defining $Y = X - \text{E}[X]$, it follows that $\text{E}[Y] = 0$ and $\text{V}(Y) = \text{V}(X) =: \sigma^2 = \text{E}\left[Y^2\right]$.

For $t, u > 0$, using Markov's inequality given for reminder by (see page 920):

$$\forall \lambda > 0 : P(X \geqslant \lambda) \leqslant \frac{\mathrm{E}(X)}{\lambda} \tag{5.7.1653}$$

we have:

$$\begin{aligned} P(Y \geq t) = P(Y + u \geq t + u) \leq P\left((Y+u)^2 \geq (t+u)^2\right) \\ \leq \frac{\mathrm{E}\left[(Y+u)^2\right]}{(t+u)^2} = \frac{\sigma^2 + u^2}{(t+u)^2} =: \varphi(u) \end{aligned} \tag{5.7.1654}$$

the last inequality being a consequence of Markov's inequality! Before we continue with the next steps a reader has asked us to explain how $P(Y + u \geq t + u) \leq P\left((Y+u)^2 \geq (t+u)^2\right)$? This is indeed not necessary trivial at first sight!

First, remember that we have $t, u > 0$. The first thing trivial is that $(Y+u)^2 \geq (Y+u)$. Now what you have to observe numerically is that for some values of $1 > t, u > 0$ we have $t + u < t^2 + 2tu + u^2$. Hence the width of the interval of the probability is bigger and that explains the inequality!

As the above holds for any choice of $u \in \mathbb{R}$, we can choose to apply it with the value that minimize: $\varphi'(u) = 0$, i.e.:

$$\varphi'(u) = 2\frac{u}{(t+u)^2} - 2\frac{\sigma^2 + u^2}{(t+u)^3} = 0 \tag{5.7.1655}$$

gives $u = \sigma^2/t$, and the result follows:

$$\boxed{P(X - \mathrm{E}(X) \geq t) \leq \frac{\sigma^2}{\sigma^2 + t^2}} \tag{5.7.1656}$$

and this is the "Cantelli's inequality" (also named the "Chebyshev-Cantelli inequality" or the "one-sided Chebyshev inequality").

The last two inequalities obtained before the example will allow us to obtain a very important and powerful relation that we call "weak law of large numbers" (WLLN) or "Khinchin theorem".

Consider a random variable X having a variance and $(X_n)_{n \in \mathbb{N}^*}$ a sequence of independent random variables (i.e. uncorrelated by pairs) of the same distribution as this of X and all having the same mean μ and the same standard deviations σ.

What we will show is that if we measure the same random quantity X_n of the same distribution in the process of a series of independent experiments (so in this case, we say technically that the sequence $(X_n)_{n \in \mathbb{N}^*}$ of random variables is defined on the same probability space) , then the arithmetic average of the observed values will stabilize on the mean of X when the number of measurements tends to infinity[50]:

$$\boxed{\forall \varepsilon > 0 : \quad \left(\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| \geqslant \varepsilon \right) \leqslant \frac{\sigma^2}{n\varepsilon^2} \underset{n \to +\infty}{\to} 0} \tag{5.7.1657}$$

---

[50] This explains for example why births, weddings, deaths display a fascinating regularity from one year to the other on huge independent - uncorrelated - population of individuals

when $n \rightarrow +\infty$ this is the very important result which we did mention above! The empirical estimator of the mean tends for any distribution to the true expected mean if n is large! So by this rule we assure that the sample average is a consistent estimator of mean! This result (quite intuitive) is sometimes named the "fundamental theorem of Monte Carlo" because it is central to the simulations principle of the same name (see section Numerical Methods page 6708), which are crucial in the study of advanced statistics.

Remark

This convergence property is also named "regression toward the mean" or simply "reversion to the mean". Implicitly meaning that if a variable is extreme on its first measurement, it will tend to be closer to the average on its second measurement - and if it is extreme on its second measurement, it will tend to have been closer to the average on its first. That's why measuring and individual performance (whatever what kind of performance), only once doesn't make sense scientifically speaking as that performance fluctuates more or less around its mean!!!

One of Galton's first examples was the average height of parents and their children. He found that tall parents had (on average) children who were smaller than them, and that short parents had (on average) children who were taller than them. In both cases the children with parents at the extreme ends of the distribution had heights closer to the population mean height.

So in other words, the cumulative probability that the difference between the arithmetic average and the expected of the observed random variables to be in a given range around the average tends to zero as the number of measured random variables tends to infinity (that which is ultimately intuitive).

This result allows us to estimate the expected mean value using the empirical mean (arithmetic average) calculated on a very large number of experiments.

Proof 5.47.12. We use the Chebyshev-Bienaymé inequality for the random variable (this relation is difficult to interpret but allows us to get the desired result):

$$Y=\frac{1}{n}\sum_{i=1}^{n}X_i \tag{5.7.1658}$$

And we first calculate using the mathematical properties of the mean that we proved earlier:

$$\mathrm{E}(Y)=\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n}X_i\right)=\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i)=\frac{1}{n}\sum_{i=1}^{n}\mu=\mu \tag{5.7.1659}$$

and in a second time using the mathematical properties of the variance as already proved above:

$$\mathrm{V}(Y)=\mathrm{V}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n^2}\mathrm{V}\left(\sum_{i=1}^{n}X_i\right)=\frac{1}{n^2}\left(\sum_{i=1}^{n}\mathrm{V}(X_i)+2\sum_{1\leqslant i<j\leqslant n}\mathrm{cov}(X_i,X_j)\right) \tag{5.7.1660}$$

and since we assumed uncorrelated variables then the covariance between them is zero therefore:

$$\mathrm{V}(Y)=\frac{1}{n^2}\left(\sum_{i=1}^{n}\mathrm{V}(X_i)\right)=\frac{1}{n^2}\left(\sum_{i=1}^{n}\sigma^2\right)=\frac{1}{n^2}n\sigma^2=\frac{\sigma^2}{n^2} \tag{5.7.1661}$$

So by injecting it into the BT inequality:

$$P\left([Y-\mathrm{E}(Y)]^2 \geqslant \varepsilon^2\right) \leqslant \frac{\mathrm{V}(Y)}{\varepsilon^2} \tag{5.7.1662}$$

that becomes:

$$P\left(\left[\frac{1}{n}\sum_{i=1}^{n} X_i-\mu\right]^2 \geqslant \varepsilon^2\right) \leqslant \frac{\sigma^2}{n\varepsilon^2} \tag{5.7.1663}$$

and the inequality tends effectively to zero as well n at the denominator tends to infinity.

■ Q.E.D.

Note that the latter relation is often denoted in some works, and according to what we saw at the beginning of this section:

$$P\left(|M_n-m_X|^2 \geqslant \varepsilon\right) \leqslant \frac{\sigma_X^2}{n\varepsilon^2} \tag{5.7.1664}$$

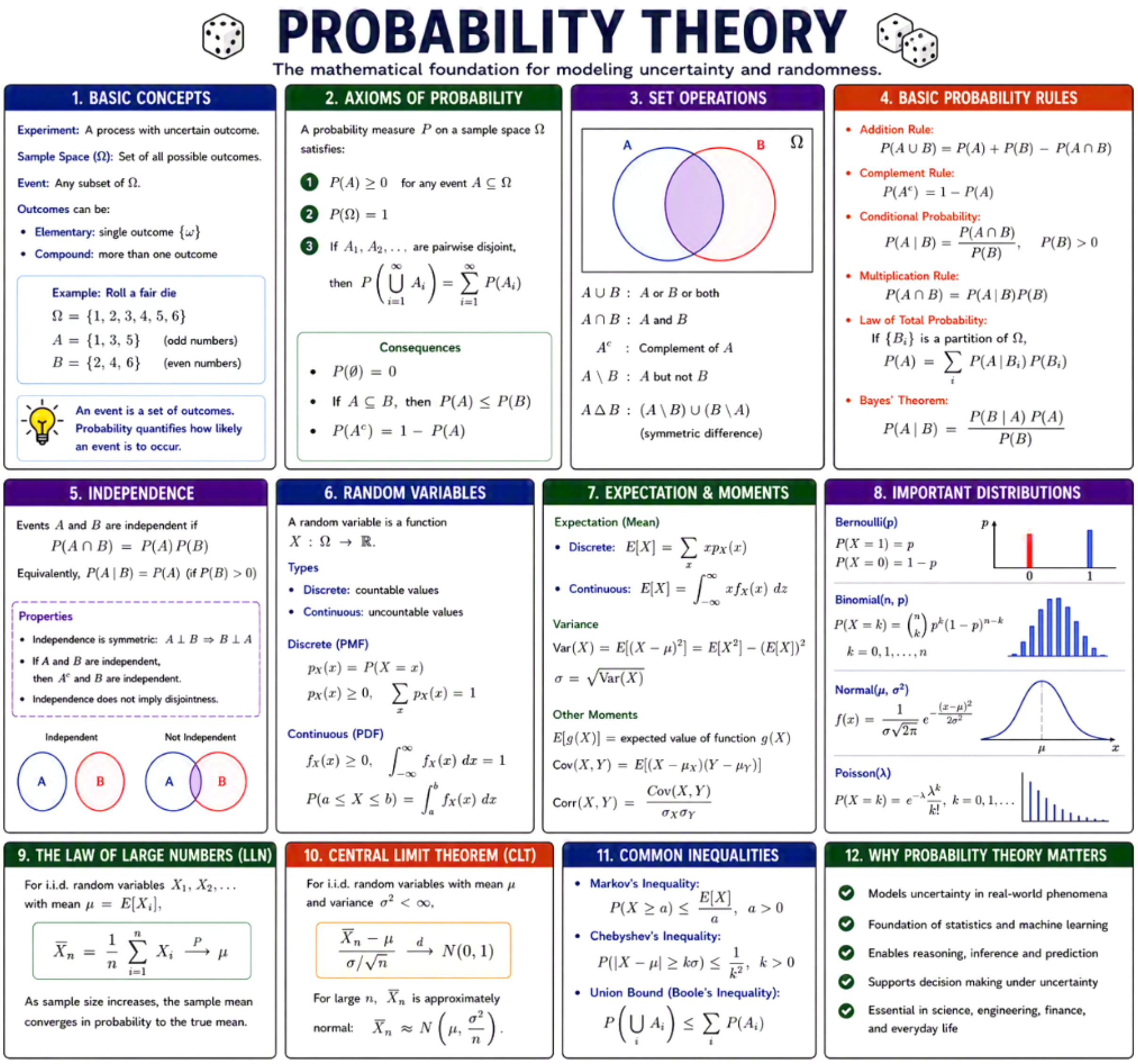
or:

$$P\left(|M_n-m_X| < \varepsilon\right) \geqslant 1-\frac{\sigma_X^2}{n\varepsilon^2} \tag{5.7.1665}$$

Therefore for $\forall\varepsilon > 0$:

$$\lim_{n\rightarrow+\infty} P\left(|M_n-m_X| < \varepsilon\right) = 1 \tag{5.7.1666}$$

A brief reminder of some of the most important elements seen so far about probabilities and statistics :



PROBABILITY THEORY

The mathematical foundation for modeling uncertainty and randomness.

1. BASIC CONCEPTS

Experiment: A process with uncertain outcome.

Sample Space (Ω): Set of all possible outcomes.

Event: Any subset of Ω.

Outcomes can be:

- **Elementary:** single outcome $\{\omega\}$
- **Compound:** more than one outcome

Example: Roll a fair die

$\Omega = \{1, 2, 3, 4, 5, 6\}$

$A = \{1, 3, 5\}$ (odd numbers)

$B = \{2, 4, 6\}$ (even numbers)

An event is a set of outcomes. Probability quantifies how likely an event is to occur.

2. AXIOMS OF PROBABILITY

A probability measure P on a sample space Ω satisfies:

① $P(A) \geq 0$ for any event $A \subseteq \Omega$

② $P(\Omega) = 1$

③ If $A_1, A_2, \ldots$ are pairwise disjoint, then $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$

Consequences

- $P(\emptyset) = 0$
- If $A \subseteq B$, then $P(A) \leq P(B)$
- $P(A^c) = 1 - P(A)$

3. SET OPERATIONS

$A \cup B$: A or B or both

$A \cap B$: A and B

A^c : Complement of A

$A \setminus B$: A but not B

$A \,\Delta\, B$: $(A \setminus B) \cup (B \setminus A)$ (symmetric difference)

4. BASIC PROBABILITY RULES

- **Addition Rule:**
  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- **Complement Rule:**
  $P(A^c) = 1 - P(A)$
- **Conditional Probability:**
  $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}, \quad P(B) > 0$
- **Multiplication Rule:**
  $P(A \cap B) = P(A \mid B)P(B)$
- **Law of Total Probability:**
  If $\{B_i\}$ is a partition of Ω,
  $P(A) = \sum_i P(A \mid B_i)\, P(B_i)$
- **Bayes' Theorem:**
  $P(A \mid B) = \dfrac{P(B \mid A)\, P(A)}{P(B)}$

5. INDEPENDENCE

Events A and B are independent if

$P(A \cap B) = P(A)\, P(B)$

Equivalently, $P(A \mid B) = P(A)$ (if $P(B) > 0$)

Properties

- Independence is symmetric: $A \perp B \Rightarrow B \perp A$
- If A and B are independent, then A^c and B are independent.
- Independence does not imply disjointness.

6. RANDOM VARIABLES

A random variable is a function $X : \Omega \to \mathbb{R}$.

Types

- **Discrete:** countable values
- **Continuous:** uncountable values

Discrete (PMF)

$p_X(x) = P(X = x)$

$p_X(x) \geq 0, \quad \sum_x p_X(x) = 1$

Continuous (PDF)

$f_X(x) \geq 0, \quad \int_{-\infty}^{\infty} f_X(x)\, dx = 1$

$P(a \leq X \leq b) = \int_a^b f_X(x)\, dx$

7. EXPECTATION & MOMENTS

Expectation (Mean)

- **Discrete:** $E[X] = \sum_x x p_X(x)$
- **Continuous:** $E[X] = \int_{-\infty}^{\infty} x f_X(x)\, dz$

Variance

$\mathrm{Var}(X) = E[(X-\mu)^2] = E[X^2] - (E[X])^2$

$\sigma = \sqrt{\mathrm{Var}(X)}$

Other Moments

$E[g(X)]$ = expected value of function $g(X)$

$\mathrm{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$

$\mathrm{Corr}(X,Y) = \dfrac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$

8. IMPORTANT DISTRIBUTIONS

Bernoulli(p)

$P(X=1) = p$

$P(X=0) = 1-p$

Binomial(n, p)

$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$

$k = 0, 1, \ldots, n$

Normal(μ, σ^2)

$f(x) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Poisson(λ)

$P(X=k) = e^{-\lambda}\dfrac{\lambda^k}{k!}, \ k = 0, 1, \ldots$

9. THE LAW OF LARGE NUMBERS (LLN)

For i.i.d. random variables $X_1, X_2, \ldots$ with mean $\mu = E[X_i]$,

$$\boxed{\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu}$$

As sample size increases, the sample mean converges in probability to the true mean.

10. CENTRAL LIMIT THEOREM (CLT)

For i.i.d. random variables with mean μ and variance $\sigma^2 < \infty$,

$$\boxed{\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)}$$

For large n, $\overline{X}_n$ is approximately normal: $\overline{X}_n \approx N\left(\mu, \dfrac{\sigma^2}{n}\right)$.

11. COMMON INEQUALITIES

- **Markov's Inequality:**
  $P(X \geq a) \leq \dfrac{E[X]}{a}, \quad a > 0$
- **Chebyshev's Inequality:**
  $P(|X-\mu| \geq k\sigma) \leq \dfrac{1}{k^2}, \quad k > 0$
- **Union Bound (Boole's Inequality):**
  $P\left(\bigcup_i A_i\right) \leq \sum_i P(A_i)$

12. WHY PROBABILITY THEORY MATTERS

- ✓ Models uncertainty in real-world phenomena
- ✓ Foundation of statistics and machine learning
- ✓ Enables reasoning, inference and prediction
- ✓ Supports decision making under uncertainty
- ✓ Essential in science, engineering, finance, and everyday life

5.7.11 Characteristic Function

In probability theory and statistics, the "characteristic function" of any real-valued random variable completely defines its probability distribution. If a random variable admits a probability density function, we will prove further below that then the characteristic function is the inverse Fourier transform of the probability density function. Thus it provides the basis of an alternative route to analytical results compared with working directly with probability density functions or cumulative distribution functions.

Before giving an engineer proof of the famous central limit theorem, first let us introduce the concept of "characteristic function", which is central in statistics.

First, remember that the Fourier transform is given in its physicist version (see section Sequences and Series page 1758) by the relation:

$$\mathcal{F}(k)=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}f(x)e^{-\mathrm{i}xk}\mathrm{d}x \tag{5.7.1667}$$

Let us recall that the Fourier transform is an analogue of the Fourier series theory for non-periodic functions, and allows to associate them a frequencies spectrum. To a given factor, it is a "bilateral Laplace transform" given by (see section Analysis page 2337):

$$\mathcal{L}[f(x)]=\mathcal{F}(p)=\int\limits_{-\infty}^{+\infty}f(x)e^{-px}\mathrm{d}x \tag{5.7.1668}$$

where p is the complex variable given in the present case by (the real part is zero, because the Fourier transform is the particular case of a Laplace transform whose real part of the variable is zero: then make Fourier transform is like make a Laplace transform on the axis of imaginary numbers only):

$$p=\mathrm{i}k \tag{5.7.1669}$$

Now we want to prove that if:

$$g(x)=f'(x)\quad\Rightarrow\quad G(k)=\mathrm{i}k\mathcal{F}(k) \tag{5.7.1670}$$

In other words, we are looking for a simplified expression of the Fourier transform of the derivative of $f(x)$.

Proof 5.47.13. We start from:

$$G(k)=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}f'(x)e^{-\mathrm{i}xk}\mathrm{d}x \tag{5.7.1671}$$

An integration by parts gives (see section Differential and Integral Calculus page 1532):

$$\int\limits_{-\infty}^{+\infty}f'(x)e^{-\mathrm{i}xk}\mathrm{d}x=f(x)e^{-\mathrm{i}xk}\Big|_{x=-\infty}^{x=+\infty}+\mathrm{i}k\int\limits_{-\infty}^{+\infty}f(x)e^{-\mathrm{i}xk}\mathrm{d}x \tag{5.7.1672}$$

By imposing that f tends to zero at infinity, then we have:

$$f(x)e^{-\mathrm{i}xk}\Big|_{x=-\infty}^{x=+\infty}=0 \tag{5.7.1673}$$

and:

$$G(k)=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}f'(x)e^{-\mathrm{i}xk}\mathrm{d}x=\frac{\mathrm{i}k}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}f(x)e^{-\mathrm{i}xk}\mathrm{d}x=\mathrm{i}k\mathcal{F}(k) \tag{5.7.1674}$$

■ Q.E.D.

This is the first result we needed.

Now let us prove that if:

$$g(x)=xf(x)\quad\Rightarrow\quad G(k)=\mathrm{i}\mathcal{F}'(k) \tag{5.7.1675}$$

Proof 5.47.14. So we start from:

$$\begin{aligned}G(k)&=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}xf(x)e^{-\mathrm{i}xk}\mathrm{d}x=\mathrm{i}\frac{\mathrm{d}}{\mathrm{d}k}\left(\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}f(x)e^{-\mathrm{i}xk}\mathrm{d}x\right)=\mathrm{i}\mathcal{F}'(k)\\&=\mathrm{i}\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}f(x)\frac{\mathrm{d}}{\mathrm{d}k}e^{-\mathrm{i}xk}\mathrm{d}x=\mathrm{i}\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}f(x)(-\mathrm{i}x)e^{-\mathrm{i}xk}\mathrm{d}x=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}xf(x)e^{-\mathrm{i}xk}\mathrm{d}x\end{aligned} \tag{5.7.1676}$$

■ Q.E.D.

This is the second result we needed.

Now let us perform the calculation of the Fourier transform of the Normal centered-reduced distribution $\mathcal{N}(0,1)$ (this choice is not innocent...):

$$y=f_X(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2} \tag{5.7.1677}$$

We know that this latter relation is trivially solution of the following differential equation (or it satisfies it in other words...):

$$y'=-xy \tag{5.7.1678}$$

taking the Fourier transform of both sides of the equality, we use the previous two results:

$$\begin{gathered}g(x)=f'(x)\text{ then }G(k)=\mathrm{i}k\mathcal{F}(k)\\g(x)=xf(x)\text{ then }G(k)=\mathrm{i}\mathcal{F}'(k)\end{gathered} \tag{5.7.1679}$$

We have:

$$\mathrm{i}k\mathcal{F}(k)=-\mathrm{i}\mathcal{F}'(k) \tag{5.7.1680}$$

Or:

$$-k\mathrm{d}k=\frac{1}{\mathcal{F}(k)}\mathrm{d}\mathcal{F}(k) \tag{5.7.1681}$$

Then after integration:

$$\begin{aligned}&\int\limits_0^k-k\mathrm{d}k=\int\limits_0^k\frac{1}{\mathcal{F}(k)}\mathrm{d}\mathcal{F}(k)\Rightarrow-\frac{1}{2}k^2\Big|_0^k=\ln(\mathcal{F}(k))\Big|_0^k\Rightarrow-\frac{1}{2}k^2=\ln(\mathcal{F}(k))-\ln(\mathcal{F}(0))\\&\qquad\Rightarrow-\frac{1}{2}k^2=\ln\left(\frac{\mathcal{F}(k)}{\mathcal{F}(0)}\right)\Rightarrow e^{-\frac{1}{2}k^2}=\frac{\mathcal{F}(k)}{\mathcal{F}(0)}\Rightarrow\mathcal{F}(k)=\mathcal{F}(0)e^{-\frac{1}{2}k^2}\end{aligned} \tag{5.7.1682}$$

As:

$$\mathcal{F}(k)=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} f(x)e^{-\mathrm{i}xk}\mathrm{d}k \tag{5.7.1683}$$

We therefore have:

$$\begin{aligned}\mathcal{F}(0)&=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} f(x)e^{-\mathrm{i}kx}\mathrm{d}x=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} f(x)\mathrm{d}x=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\mathrm{d}x\\&=\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} e^{-x^2/2}\mathrm{d}x\end{aligned} \tag{5.7.1684}$$

We proved during our study of the Normal distribution that:

$$\int\limits_{-\infty}^{+\infty} e^{-x^2/2}\mathrm{d}x=\sqrt{2\pi} \tag{5.7.1685}$$

Therefore:

$$\mathcal{F}(0)=\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} e^{-x^2/2}\mathrm{d}x=\frac{1}{2\pi}\sqrt{2\pi}=\frac{1}{\sqrt{2\pi}} \tag{5.7.1686}$$

We then have (very important result!):

$$\mathcal{F}(k)=\frac{1}{\sqrt{2\pi}}e^{-k^2/2} \tag{5.7.1687}$$

Let us now introduce the characteristic function as defined by statisticians:

$$\boxed{\phi_X(\omega)=\mathrm{E}\left[e^{\mathrm{i}kX}\right]=\int\limits_{-\infty}^{+\infty} e^{\mathrm{i}kx}f_X(x)\mathrm{d}x} \tag{5.7.1688}$$

which is an important and powerful analytical tool for analysing a sum of independent random variables. In addition, this function contains all the information characteristic of the random variable X (all moments) as its Taylor development will show it to us just further below!

Remark

The notation is not innocent since the $\mathrm{E}[\ldots]$ represents an expected mean of the density function with respect to the complex exponential.

Therefore the characteristic function of the Normal reduced centered random variable of distribution law:

$$f_X(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2} \tag{5.7.1689}$$

becomes easy to determine because:

$$\boxed{\phi_X(k)=\int\limits_{-\infty}^{+\infty} e^{\mathrm{i}kx} f_X(x)\mathrm{d}x=\sqrt{2\pi}\mathcal{F}(-k)} \tag{5.7.1690}$$

This is why the characteristic function of the reduced centered Normal distribution is often assimilated to a simple Fourier transform (see section Sequences and Series page 1758).

And thanks to the previous result:

$$\mathcal{F}(k)=\frac{1}{\sqrt{2\pi}}e^{-k^2/2}\Rightarrow F(-k)=\frac{1}{\sqrt{2\pi}}e^{-(-k)^2/2}=\frac{1}{\sqrt{2\pi}}e^{-k^2/2} \tag{5.7.1691}$$

Therefore:

$$\phi_X(k)=\frac{\sqrt{2\pi}}{\sqrt{2\pi}}e^{-k^2/2}=e^{-k^2/2} \tag{5.7.1692}$$

which is the result we will need for the central limit theorem that we will study just after. This characteristic function is equal, to a given constant, to the probability density of the law. Then we say that the characteristic function of a Gaussian function is Gaussian function....

But before that, let us look a little closer this characteristic function:

$$\phi_X(\omega)=\mathrm{E}\left[e^{\mathrm{i}kx}\right]=\int\limits_{-\infty}^{+\infty} f_X(x)\mathrm{d}x \tag{5.7.1693}$$

Using a Maclaurin development (see section Sequences and Series page 1696) and changing some notations we get:

$$\phi_X(\omega)=\int\limits_{-\infty}^{+\infty}\left[\sum_{k=0}^{+\infty}\frac{(\mathrm{i}\omega)^k}{k!}x^k f_X(x)\mathrm{d}x\right] \tag{5.7.1694}$$

and by inverting the sum and the integral, we have:

$$\phi_X(\omega)=\sum_{k=0}^{+\infty}\frac{(\mathrm{i}\omega)^k}{k!}\left[\int\limits_{-\infty}^{+\infty} x^k f_X(x)\mathrm{d}x\right]=\sum_{k=0}^{+\infty}\frac{(\mathrm{i}\omega)^k}{k!}M_X^k \tag{5.7.1695}$$

This characteristic function contains therefore all the moments (general term used for the mean, variance, skewness and kurtosis) of X.

✍**Examples:**

Let us see another famous example with the Poisson distribution (interesting case as it is a discrete distribution function!)!

Proof. Using the definition of characteristic function:

$$\begin{aligned}\varphi_X(t) &= \mathrm{E}\left[\exp(\mathrm{i}tX)\right]\\ &= \sum_{x\in\mathbb{R}_X}\exp(\mathrm{i}tx)P_X(x)\\ &= \sum_{x\in\mathbb{R}_X}[\exp(\mathrm{i}t)]^x\exp(-\lambda)\frac{1}{x!}\lambda^x\\ &= \exp(-\lambda)\sum_{x=0}^{+\infty}\frac{(\lambda\exp(\mathrm{i}t))^x}{x!}\\ &= \exp(-\lambda)\exp(\lambda\exp(\mathrm{i}t))\\ &= \exp(\lambda[\exp(\mathrm{i}t)-1])\end{aligned}$$

where:

$$\exp(\lambda\exp(\mathrm{i}t))=\sum_{x=0}^{+\infty}\frac{(\lambda\exp(\mathrm{i}t))^x}{x!}$$

is the usual Taylor series expansion of the exponential function as proved in the section Sequences and Series page 1703 (note that the series converges for any value of t).

5.7.11.1 Moment generating function and cumulants

In probability theory and statistics, the moment-generating function of a real-valued random variable is an alternative specification of its probability distribution. Thus, it provides the basis of an alternative route to analytical results compared with working directly with probability density functions or cumulative distribution functions. There are particularly simple results for the moment-generating functions of distributions defined by the weighted sums of random variables. However, not all random variables have moment-generating functions.

As its name implies, the moment generating function can be used to compute a distribution's moments: the nth moment about 0 is the nth derivative of the moment-generating function, evaluated at 0.

In addition to real-valued distributions (univariate distributions), moment-generating functions can be defined for vector- or matrix-valued random variables, and can even be extended to more general cases.

The moment-generating function of a real-valued distribution does not always exist, unlike the characteristic function. There are relations between the behaviour of the moment-generating function of a distribution and properties of the distribution, such as the existence of moments.

The moment-generating function of a random variable X is:

$$M_X(t):=\mathrm{E}\left(e^{tX}\right),\quad t\in\mathbb{R} \tag{5.7.1696}$$

wherever this expectation exists. In other words, the moment-generating function of X is the expectation of the random variable e^{tX}.

$M_X(0)$ always exists and is equal to 1 (see proof just below). However, a key problem with moment-generating functions is that moments and the moment generating function may not exist, as the integrals need not converge absolutely. By contrast, the characteristic function or Fourier transform always exists (because it is the integral of a bounded function on a space of finite measure), and for some purposes may be used instead.

The moment-generating function is so named because it can be used to find the moments of distributions. The series expansion of e^{tX} is:

$$e^{tX}=1+tX+\frac{t^2X^2}{2!}+\frac{t^3X^3}{3!}+\ldots+\frac{t^nX^n}{n!}+\ldots \tag{5.7.1697}$$

Hence:

$$\begin{aligned}M_X(t)=\mathrm{E}\left(e^{tX}\right)&=1+t\mathrm{E}(X)+\frac{t^2\mathrm{E}\left(X^2\right)}{2!}+\frac{t^3\mathrm{E}\left(X^3\right)}{3!}+\ldots+\frac{t^n\mathrm{E}\left(X^n\right)}{n!}+\ldots\\&=1+tm_1+\frac{t^2m_2}{2!}+\frac{t^3m_3}{3!}+\ldots+\frac{t^nm_n}{n!}+\ldots\end{aligned} \tag{5.7.1698}$$

where m_n is the n th moment. Differentiating $M_X(t)$ a number i times with respect to t and setting $t=0$, we obtain the ith moment about the origin, m_i.

If it exists for t in a neighbourhood of 0, then all "moments" of X exist, and:

$$\boxed{\left.\frac{\partial^k}{\partial t^k}M_X(t)\right|_{t=0}=\mathrm{E}\left(X^k\right)} \tag{5.7.1699}$$

The "cumulant generating function" is the log of the moment generating function:

$$c_X(t)=\log\left(M_X(t)\right) \tag{5.7.1700}$$

It generates the "cumulants", which are defined by what the cumulant generating function generates, i.e., for a random variable, the k^{th} cumulant is:

$$\boxed{\gamma_k=\left.\frac{\partial^k}{\partial t^k}c_X(t)\right|_{t=0}=\left.\frac{\partial^k}{\partial t^k}\log\left(M_X(t)\right)\right|_{t=0}=\left.\frac{\partial^k}{\partial t^k}\log\left(\mathrm{E}(e^{tX})\right)\right|_{t=0}} \tag{5.7.1701}$$

Cumulants are slightly easier to work with than moments. The first four are:

$$\begin{aligned}\gamma_1&=\mathrm{E}(X)=\mu_1=\mu\\\gamma_2&=\mathrm{V}(X)=\mu_2=\sigma^2\\\gamma_3&=\mathrm{E}\left[(X-\mathrm{E}[X])^3\right]=\mu_3\\\gamma_4&=\mathrm{E}\left[(X-\mathrm{E}[X])^4\right]-3\mathrm{V}[X]^2=\mu_4-3\mu_2^2=\mu_4-3\sigma^4\end{aligned} \tag{5.7.1702}$$

The skewness and kurtosis (see page 636) are then simple functions of the cumulants:

$$\text{Skewness}\,[X]=\frac{\gamma_3}{\sigma^3}\qquad\text{and}\qquad\text{Kurtosis}\,[X]=\frac{\gamma_4}{\sigma^4} \tag{5.7.1703}$$

If X is a continuous random variable, the following relation between its moment-generating function $M_X(t)$ and the two-sided Laplace transform of its probability density function $f_X(x)$ holds:

$$M_X(t)=\mathcal{L}\left\{f_X\right\}(-t) \tag{5.7.1704}$$

since the PDF's two-sided Laplace transform (see section Analysis page 2337) is given as:

$$\mathcal{L}(f(X)) = \int\limits_{-\infty}^{+\infty} e^{-sx} f_X(x)\mathrm{d}x \tag{5.7.1705}$$

and the "moment-generating function's" (MGF) definition expands (by the law of the unconscious statistician) to:

$$\boxed{M_X(t) = \mathrm{E}(e^{tX}) = \int\limits_{-\infty}^{+\infty} e^{tx} f_X(x)\mathrm{d}x} \tag{5.7.1706}$$

The moment-generating function is the expectation of a function of the random variable, it can be written as:

- For a discrete probability mass function, $M_X(t) = \sum_{i=1}^{+\infty} e^{tx_i} p_i$
- For a continuous probability density function, $M_X(t) = \int_{-\infty}^{+\infty} e^{tx} f(x)\mathrm{d}x$

✍**Examples:**

E1. Let us take the case of the Normal distribution.

Suppose $X = \mathcal{N}(\mu, \sigma^2)$ (with $\sigma^2 > 0$). Then its MGF is:

$$M_X(t) = \mathrm{E}\left(e^{tX}\right) = \frac{1}{\sqrt{2\pi}\sigma} \int\limits_{-\infty}^{+\infty} e^{tx} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \mathrm{d}x = \frac{1}{\sqrt{2\pi}\sigma} \int\limits_{-\infty}^{+\infty} e^{tx-\frac{1}{2}\frac{1}{\sigma^2}\left(x^2-2\mu x+\mu^2\right)} \mathrm{d}x$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \int\limits_{-\infty}^{+\infty} e^{-\frac{1}{2}\frac{1}{\sigma^2}\left(x^2-2\left(\mu+t\sigma^2\right)x+\mu^2\right)} \mathrm{d}x$$

In the exponent, complete the square with respect to the x:

$$x^2 - 2\left(\mu + t\sigma^2\right)x = \left(x - \left(\mu + t\sigma^2\right)\right)^2 - \left(\mu + t\sigma^2\right)^2$$

Then:

$$M_X(t) = e^{\frac{1}{2}\frac{1}{\sigma^2}\left(\left(\mu+t\sigma^2\right)^2-\mu^2\right)} \int\limits_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{1}{\sigma^2}\left(x-\left(\mu+t\sigma^2\right)\right)^2} \mathrm{d}x$$

Notice that the integrand is the probability density function of a $\mathcal{N}(\mu + t\sigma^2, \sigma^2)$, which means the integral is 1. Expanding the $\left(\mu + t\sigma^2\right)^2$ and simplifying a bit yields:

$$M_X(t) = e^{t\mu + \frac{1}{2}\sigma^2 t^2}, t \in \mathbb{R}$$

The cumulant generating function is then a simple quadratic:

$$c_X(t) = t\mu + \frac{1}{2}\sigma^2 t^2$$

and it is easy to see that:

$$c'_X(0) = \mu, \quad c''(0) = \sigma^2, \quad c'''(t) = 0$$

Thus the mean is μ and variance is σ^2 (not surprisingly), and all other cumulants are 0. In particular, the skewness and kurtosis are both 0 .

E2. Let us see now an example with the Gamma distribution.

The Gamma distribution has as we know two parameters: $\alpha > 0$ is the shape parameter, and $\lambda > 0$ is the scale parameter. Its space is $\mathcal{X} = [0, +\infty[$, and its probability density function is $cx^{\alpha-1}\exp(-\lambda x)$, where:

$$\frac{1}{c} = \int\limits_0^{+\infty} x^{\alpha-1}e^{-\lambda x}\mathrm{d}x = \int\limits_0^{+\infty} (u/\lambda)^{\alpha-1}e^{-u}\mathrm{d}u/\lambda = \frac{1}{\lambda^\alpha}\int\limits_0^{+\infty} u^{\alpha-1}e^{-u}\mathrm{d}u = \frac{1}{\lambda^\alpha}\Gamma(\alpha)$$

The gamma function is defined to be that very integral, i.e., for $\alpha > 0$ (see section Differential and Integral Calculus page 1574):

$$\Gamma(\alpha) = \int\limits_0^{+\infty} u^{\alpha-1}e^{-u}\mathrm{d}u$$

It generally has no closed-form value, but there are some convenient facts that we have already proven during it's study:

$$\begin{aligned}\Gamma(\alpha+1) &= \alpha\Gamma(\alpha) \text{ for } \alpha > 0\\ \Gamma(n) &= (n-1)! \text{ for } n = 1, 2, \ldots;\\ \Gamma\left(\frac{1}{2}\right) &= \sqrt{\pi}\end{aligned}$$

The probability density function for a $\Gamma(\alpha, \lambda)$ is then:

$$f(x \mid \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}, x \in (0, \infty)$$

If $\alpha = 1$, then we fall back on the "exponential distribution" (λ), with probability density function:

$$g(x \mid \lambda) = \lambda e^{-\lambda x}, x \in \mathcal{X} = [0, +\infty[$$

The MGF is:

$$M_X(t) = \mathrm{E}\left(e^{tX}\right) = \frac{\lambda^\alpha}{\Gamma(\alpha)}\int\limits_0^{+\infty} e^{tx}x^{\alpha-1}e^{-\lambda x}\mathrm{d}x = \frac{\lambda^\alpha}{\Gamma(\alpha)}\int\limits_0^{+\infty} x^{\alpha-1}e^{-(\lambda-t)x}\mathrm{d}x$$

That integral needs $(\lambda - t) > 0$ to be finite, so we need $t < \lambda$, which means the MGF is finite for a neighbourhood of zero, since $\lambda > 0$. Now the integral at the end of the previous relation is the same as that the one just earlier above:

$$\int\limits_0^{+\infty} x^{\alpha-1}\exp(-\lambda x)\mathrm{d}x$$

except that $\lambda - t$ is in place of λ. Thus we have:

$$\mathrm{E}\left(e^{tX}\right)=\frac{\lambda^\alpha}{\Gamma(\alpha)}\frac{\Gamma(\alpha)}{(\lambda-t)^\alpha}=\left(\frac{\lambda}{\lambda-t}\right)^\alpha, t<\lambda$$

We will use the cumulant generating function $c_X(t)=\log\left(M_X(t)\right)$ to obtain the mean and variance, because it is slightly easier. Thus:

$$c'_X(t)=\frac{\partial}{\partial t}\alpha(\log(\lambda)-\log(\lambda-t))=\frac{\alpha}{\lambda-t}\Longrightarrow \mathrm{E}(X)=c'_X(0)=\frac{\alpha}{\lambda}$$

and:

$$c''_X(t)=\frac{\partial^2}{\partial t^2}\alpha(\log(\lambda)-\log(\lambda-t))=\frac{\alpha}{(\lambda-t)^2}\Longrightarrow \mathrm{V}(X)=c''_X(0)=\frac{\alpha}{\lambda^2}$$

In general, the k^{th} cumulant is:

$$\gamma_k=\left.\frac{\partial^k}{\partial t^k}c_X(t)\right|_{t=0}=(k-1)!\frac{\alpha}{\lambda^k}$$

and in particular:

$$\begin{aligned}\text{Skewness }(X)&=\frac{2\alpha/\lambda^3}{\alpha^{3/2}/\lambda^3}=\frac{2}{\sqrt{\alpha}}\\ \text{Kurtosis }(X)&=\frac{6\alpha/\lambda^4}{\alpha^2/\lambda^4}=\frac{6}{\alpha}\end{aligned}$$

Thus the skewness and kurtosis depends on just the shape parameter α. Also, they are positive, but tend to 0 as α increases.

E3. Let us see now an example with the binomial distribution.

The Binomial is as we already know a model for counting the number of successes in n trials, e.g., the number of heads in ten flips of a coin, where the trials are independent and have the same probability p of success:

$$X=\mathcal{B}(n,p)\Longrightarrow f_X(x)=\begin{pmatrix}n\\x\end{pmatrix}p^x(1-p)^{n-x},\quad x\in\mathcal{X}=\{0,1,\ldots,n\}$$

The fact that the probability density function sums to 1 relies on the binomial theorem (see section Calculus page 1364):

$$(a+b)^n=\sum_{x=0}^{n}a^xb^{n-x}$$

with $a=p$ and $b=1-p$. This theorem also helps in finding the MGF:

$$\begin{aligned}M_X(t)=\mathrm{E}\left(e^{tX}\right)&=\sum_{x=0}^{n}e^{tx}f_X(x)=\sum_{x=0}^{n}e^{tx}\begin{pmatrix}n\\x\end{pmatrix}p^x(1-p)^{n-x}\\&=\sum_{x=0}^{n}\begin{pmatrix}n\\x\end{pmatrix}\left(pe^t\right)^x(1-p)^{n-x}=\left(pe^t+1-p\right)^n\end{aligned}$$

It is finite for all $t \in \mathbb{R}$, as is the case for any bounded random variable. The first two moments are:

$$\mathrm{E}(X)=M_X'(0)=n\left(pe^t+1-p\right)^{n-1}pe^t\Big|_{t=0}=np$$

and:

$$\begin{aligned}\mathrm{E}\left(X^2\right)=M_X''(0)&=\left[n(n-1)\left(pe^t+1-p\right)^{n-2}\left(pe^t\right)^2+n\left(pe^t+1-p\right)^{n-1}pe^t\right]\Big|_{t=0}\\&=n(n-1)p^2+np\end{aligned}$$

Thus:

$$\mathrm{V}(X)=\mathrm{E}\left(X^2\right)-\mathrm{E}(X)^2=n(n-1)p^2+np-(np)^2=np-np^2=np(1-p)$$

E4. Let us see now an example with the Poisson distribution.

$$\begin{aligned}M_X(t)=\mathrm{E}\left(e^{tX}\right)=\sum_{x=0}^{n}e^{tx}f_X(x)&=\sum_{x=0}^{+\infty}e^{tx}e^{-\lambda}\frac{1}{x!}\lambda^x\\&=e^{-\lambda}\sum_{x=0}^{+\infty}\frac{\left(\lambda e^t\right)^x}{x!}\\&=e^{-\lambda}e^{\lambda e^t}=e^{\lambda\left(e^t-1\right)}\end{aligned}$$

where:

$$e^{\lambda e^t}=\sum_{x=0}^{+\infty}\frac{\left(\lambda e^t\right)^x}{x!}$$

is the usual Taylor series expansion of the exponential function (see section Sequences and Series page 1703). Furthermore, since the series converges for any value of t, the moment generating function of a Poisson random variable exists for any $t \in \mathbb{R}$.

E5. Let us see now an example with the ex-Gaussian distribution (see section Statistics page 806).

Then, the moment generating function of X is:

$$M_X(t)=\left(\frac{\lambda}{\lambda-t}\right)\exp\left[\mu t+\frac{1}{2}\sigma^2t^2\right] \tag{5.7.1707}$$

Indeed, suppose X follows an ex-Gaussian distribution. Then, $X=A+B$ where A and B are independent A is normally distributed with mean μ and variance σ^2, and B is exponentially distributed with rate λ. We have already proven above that the moment generating function for A (Normal distribution) is given by:

$$M_A(t)=\exp\left[\mu t+\frac{1}{2}\sigma^2t^2\right] \tag{5.7.1708}$$

and we also have already proven the moment generating function for B (Poisson distribution) is given by:

$$M_B(t)=\frac{\lambda}{\lambda-t} \tag{5.7.1709}$$

By definition, X is a linear combination of independent random variables A and B, so the moment generating function of X is the product of $M_A(t)$ and $M_B(t)$. That is:

$$\begin{aligned}M_X(t) &= M_A(t)\cdot M_B(t)\\ &= \exp\left[\mu t+\frac{1}{2}\sigma^2t^2\right]\cdot\left(\frac{\lambda}{\lambda-t}\right)\\ &= \left(\frac{\lambda}{\lambda-t}\right)\exp\left[\mu t+\frac{1}{2}\sigma^2t^2\right]\end{aligned}\tag{5.7.1710}$$

Theorem 5.48. The limiting distribution of a Poisson (λ) distribution as $\lambda\rightarrow+\infty$ is Normal (this will be useful to us in the section of Quality Engineering).

Proof 5.48.1. Let $X_n\sim\text{Poisson}(n)$, for $n=1,2,\ldots$ The probability mass function of X_n is:

$$f_{X_n}(x)=\frac{n^xe^{-n}}{x!}\quad x=0,1,2,\ldots\tag{5.7.1711}$$

The moment generating function of X_n is as proved in the example just above:

$$M_{X_n}(t)=E\left[e^{tX_n}\right]=e^{n\left(e^t-1\right)}\tag{5.7.1712}$$

for $-\infty<t<+\infty$. Taking the limit gets us nowhere because:

$$\lim_{n\rightarrow+\infty}M_{X_n}(t)=\lim_{n\rightarrow\infty}e^{n\left(e^t-1\right)}=\infty\tag{5.7.1713}$$

So now consider a "standardized" Poisson random variable:

$$\frac{X_n-n}{\sqrt{n}}\tag{5.7.1714}$$

which has limiting moment generating function:

$$\begin{aligned}\lim_{n\rightarrow+\infty}M_{(X_n-n)/\sqrt{n}}(t) &= \lim_{n\rightarrow+\infty}\text{E}\left[\exp\left(t\cdot\frac{X_n-n}{\sqrt{n}}\right)\right]\\ &= \lim_{n\rightarrow+\infty}\exp(-t\sqrt{n})\text{E}\left[\exp\left(\frac{tX_n}{\sqrt{n}}\right)\right]\\ &= \lim_{n\rightarrow+\infty}\exp(-t\sqrt{n})\exp\left(n\left(e^{t/\sqrt{n}}-1\right)\right)\\ &= \lim_{n\rightarrow+\infty}\exp\left(-t\sqrt{n}+n\left(tn^{-1/2}+t^2n^{-1}/2+t^3n^{-3/2}/6+\cdots\right)\right)\\ &= \lim_{n\rightarrow+\infty}\exp\left(t^2/2+t^3n^{-1/2}/6+\cdots\right)\\ &= \exp\left(t^2/2\right)\qquad-\infty<t<+\infty\end{aligned}\tag{5.7.1715}$$

by using the moment generating function of a Poisson random variable and expanding the exponential function as a series. This can be recognized as the moment generating function of a standard Normal random variable! This implies that the associated unstandardized random variable X_n has a limiting distribution that is Normal with mean n and variance n. This result is the basis for the "normal approximation to the Poisson distribution"! In practice most people consider the approximation valid enough if $n>20$ (i.e. $\lambda>20$).

5.7.12 Central Limit Theorem

The central limit theorem is a set of results from the early 120th century (holocene calendar) on the weak convergence in probability of a sequence of random variables. Intuitively, from these results, any sum (implicitly: the average of these variable) of independent random variables identically distributed tends to some given random variable. The best known and most important result is simply named "central limit theorem" concerning a sum of independent random variables with existing variance whose number tends to infinity and it is this that we will prove heuristically here.

In the simplest case, considered below for the proof of the theorem, these variables are continuous, independent and have the same mean and the same variance. To try to get a finite result, we should center this sum by subtracting its average and reduce it by dividing it by its standard deviation. Under fairly broad conditions, the probability distribution (of the average) converges to a Normal centered reduced distribution. The ubiquity of the Normal distribution is explained by the fact that many considered random phenomena are due to the superposition of an infinity of many small causes.

This probability theorem has an interpretation in mathematical statistics. The latter associates a probability distribution to a population. Each element extracted from the population is considered as a random variable and by bringing together a number n of these supposed independent variables, we get a sample. The sum of these random variables divided by n gives a new variable named the "empirical mean". This, once reduced, tends to a Normal reduced variable as n approaches infinity as we know.

The central limit theorem tells us what we should expect in terms of sums of independent random variables. But what about the products? Well, the logarithm of a product (strictly positive factors) is the sum of logarithms of factors, so that the logarithm of a product of random variables (with strictly positive values) tends to a Normal distribution, resulting a log-Normal distribution for the product itself (see our study of the Log-Normal distribution earlier above).

In itself, the convergence to the Normal distribution ("asymptotic normality") of many random variables when their number tends to infinity only interests the mathematician. For the practitioner, it is worth stopping shortly before the limit: the sum of a large number of these variables is nearly Gaussian, which often provides a more usable approximation than the exact law.

Moving further away from theory, we can say that a good number of natural phenomena are due to the superposition of numerous causes, more or less independent. As a result, the Normal law represents them reasonably effectively.

Conversely, we can say that no concrete phenomenon is really Gaussian because it can not exceed certain limits, especially if its values are all positive (or all negative).

Proof 5.48.2. Given $\{X_i\}_{i=1...+\infty}$ a suite (sample) of continuous random variables (in our simplified proof...), independent (independent measures of physical or mechanical phenomena for example) and identically distributed, for which the average μ_X and standard deviation σ_X exist (this means that the central limit theorem works for finite variance phenomena only as far as we know!!!).

We saw earlier in this section (see page 940) that:

$$\boxed{Z_n=\frac{S_n-n\mu}{\sigma\sqrt{n}}=\frac{M_n-\mu}{\sigma/\sqrt{n}}=\frac{\textbf{signal}}{\text{estimated noise}}} \tag{5.7.1716}$$

are the same expressions of centered reduced variable generated using a sequence of n identically distributed random variables that by construction therefore has a zero mean and unit variance:

$$\mathrm{E}(Z_n)=0 \quad \text{and} \quad \mathrm{V}(Z_n)=1 \tag{5.7.1717}$$

Let us develop the first form of the prior-previous equality (the both are anyway equal!):

$$Z_n=\frac{1}{\sqrt{n}\sigma}\left[\sum_{i=1}^{n}(X_i-\mu)\right]=\frac{1}{\sqrt{n}\sigma}\left[\sum_{i=1}^{n}X_i-n\mu\right] \tag{5.7.1718}$$

Now using the characteristic function of the Normal centered-reduced distribution (we simplify at the same time the notations of the estimators of the average and standard deviation):

$$\phi_{Z_n}(\omega)=\mathrm{E}[e^{\mathrm{i}\omega Z_n}]=\mathrm{E}\left[\exp\left(\mathrm{i}\omega\frac{\sum_{i=1}^{n}(X_i-\mu)}{\sqrt{n}\sigma}\right)\right] \tag{5.7.1719}$$

As the random variables X_i are independent and identically distributed, we get:

$$\phi_{Z_n}(\omega)=\mathrm{E}\left[\exp\left(\mathrm{i}\omega\frac{X_1-\mu}{\sqrt{n}\sigma}\right)\right]\cdot\ldots\cdot\mathrm{E}\left[\exp\left(\mathrm{i}\omega\frac{X_n-\mu}{\sqrt{n}\sigma}\right)\right]=\left(\mathrm{E}\left[\exp\left(\mathrm{i}\omega\frac{X-\mu}{\sqrt{n}\sigma}\right)\right]\right)^n \tag{5.7.1720}$$

A Taylor expansion (see section Sequences and Series page 1696) of the term between braces gives at the third order (Maclaurin series expansion of the exponential):

$$\mathrm{E}\left[\exp\left(\mathrm{i}\omega\frac{X-\mu}{\sqrt{n}\sigma}\right)\right]\cong\mathrm{E}\left[1+\mathrm{i}\frac{\omega}{\sqrt{n}\sigma_{\bar{X}}}(X-\mu)-\frac{\omega^2}{n\sigma^2}(X-\mu)^2\right] \tag{5.7.1721}$$

Finally:

$$\begin{aligned}
\phi_{Z_n}(\omega)&\cong\left(\mathrm{E}\left[1+\mathrm{i}\frac{\omega}{\sqrt{n}\sigma}(X-\mu)-\frac{\omega^2}{n\sigma^2}(X-\mu)^2\right]\right)^n\\
&=\left[\int_{-\infty}^{+\infty}\left(1+\mathrm{i}\frac{\omega}{\sqrt{n}\sigma}(X-\mu)-\frac{\omega^2}{n\sigma_{\bar{X}}^2}(X-\mu)^2\right)f_X(x)\mathrm{d}x\right]^n\\
&=\left[\int_{-\infty}^{+\infty}f_X(x)\mathrm{d}x+\int_{-\infty}^{+\infty}\frac{\mathrm{i}X\omega}{\sqrt{n}\sigma}f_X(x)\mathrm{d}x-\int_{-\infty}^{+\infty}\frac{\mathrm{i}\mu\omega}{\sqrt{n}\sigma}f_X(x)\mathrm{d}x-\frac{\omega^2}{n\sigma^2}\int_{-\infty}^{+\infty}(X-\mu)^2f_X(x)\mathrm{d}x\right]^n\\
&=\left[1+\frac{\mathrm{i}\omega}{\sqrt{n}\sigma_{\bar{X}}}\int_{-\infty}^{+\infty}Xf_X(x)\mathrm{d}x-\frac{\mathrm{i}\mu\omega}{\sqrt{n}\sigma_{\bar{X}}}\int_{-\infty}^{+\infty}f_X(x)\mathrm{d}x-\frac{\omega^2}{n\sigma_{\bar{X}}^2}\int_{-\infty}^{+\infty}(X-\mu)^2f_X(x)\mathrm{d}x\right]^n\\
&=\left[1+\frac{\mathrm{i}\omega}{\sqrt{n}\sigma_{\bar{X}}}\mu-\frac{\mathrm{i}\mu\omega}{\sqrt{n}}-\frac{\omega^2}{n\sigma^2}\int_{-\infty}^{+\infty}(X-\mu)^2f_X(x)\mathrm{d}x\right]^n\\
&=\left[1-\frac{\omega^2}{n\sigma^2}\int_{-\infty}^{+\infty}(X-\mu)^2f_X(x)\mathrm{d}x\right]^n=\left[1-\frac{\omega^2}{n\sigma_{\bar{X}}^2}\mathrm{V}(X)\right]^n=\left[1-\frac{\omega^2}{2n}\right]^n
\end{aligned} \tag{5.7.1722}$$

Let us put:

$$-\frac{\omega^2}{2n}=\frac{1}{x} \tag{5.7.1723}$$

Then we have:

$$\phi_{Z_n}(\omega)\cong\left[1+\frac{1}{x}\right]^{-\frac{\omega^2 x}{2}}=\left[\left(1+\frac{1}{x}\right)^x\right]^{-\frac{\omega^2}{2}} \tag{5.7.1724}$$

So as x approaches infinity (see section Functional Analysis page 2226):

$$\boxed{\phi_{Z_n}(\omega)\cong\lim_{n\rightarrow+\infty}\left[\left(1+\frac{1}{x}\right)^x\right]^{-\frac{\omega^2}{2}}=\left[\lim_{n\rightarrow+\infty}\left(1+\frac{1}{x}\right)^x\right]^{-\frac{\omega^2}{2}}=e^{-\frac{\omega^2}{2}}} \tag{5.7.1725}$$

We find back the characteristic function of the reduced centered Normal distribution!

In two words, the Central Limit Theorem (CLT) said that for large samples, the centered and reduced sum of n random variables independent ad identically distributed follow a Normal centered and reduced distribution. And so we have verbatim to the empirical mean:

$$\frac{M_n-\mu}{\sigma/\sqrt{n}}=\mathcal{N}(0,1)\Leftrightarrow M_n=\mathcal{N}\left(\mu,\frac{\sigma}{\sqrt{n}}\right) \tag{5.7.1726}$$

Or more elegantly the CLT will be written:

$$\boxed{\sqrt{n}\left(\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)-\mu\right)\xrightarrow{d}\mathcal{N}\left(0,\sigma^2\right)} \tag{5.7.1727}$$

■ Q.E.D.

Caution! In facts there are three central limits theorem (CLT) nowadays... The "Lindeberg-Levy CLT" proven just above that isn't very useful because of the need for identically distributed variables. The "Lyapunov's and Lindeberg CLT" which is much more flexible relaxing the requirement of identically distributed variables at the cost of the necessity of higher moments like skewness and kurtosis to disappear sufficiently quickly (being just finite and quickly go towards zero). The "Gnedenko-Kolmogorov CLT" that is even better as it handles also heavy-tails (but there are also traps as many things in life are not as easy as they seem to...!).

Now we will illustrate the central limit theorem in the case of a sequence $\{X_i\}$ of discrete random variables following a Bernoulli distribution with parameter equal to $1/2$.

We can imagine that XN represents the result obtained in the n-th launch of a coin (assigning the number 1 for head and 0 for tail). Let us write:

$$\bar{X}_n=\frac{1}{n}\sum_{k=1}^{n}X_k \tag{5.7.1728}$$

the average. We obviously have for all n in this special case:

$$\mathrm{E}(X_n)=\frac{1}{2} \qquad \mathrm{V}(X_n)=\frac{1}{4} \tag{5.7.1729}$$

and therefore:

$$\mathrm{E}(\bar{X}_n)=\sum_{k=1}^{n}\frac{1}{2}=\frac{1}{2} \qquad \mathrm{V}(\bar{X}_n)=\frac{1}{n^2}\sum_{k=1}^{n}\frac{1}{4}=\frac{1}{4n} \tag{5.7.1730}$$

After having centered and reduced $\bar{X}_n$ we get:

$$\frac{\bar{X}_n-\mathrm{E}(\bar{X}_n)}{\sqrt{\mathrm{V}(\bar{X}_n)}}=2\sqrt{n}\left(\bar{X}_n-\frac{1}{2}\right) \tag{5.7.1731}$$

Let us note Φ as always the cumulative Normal reduced centered distribution.

The central limit theorem says that for any $t\in\mathbb{R}$:

$$\lim_{n\rightarrow+\infty}P\left(2\sqrt{n}\left(\bar{X}_n-\frac{1}{2}\right)\leq t\right)=\Phi(t) \tag{5.7.1732}$$

Using Maple™ 4.00b we have plot in blue some graphs of the function:

$$t\mapsto F_n(t)=P\left(2\sqrt{n}\left(\bar{X}_n-\frac{1}{2}\right)\leq t\right) \tag{5.7.1733}$$

for different values of n. We have Represented in red the function Φ.

For $n=1$:

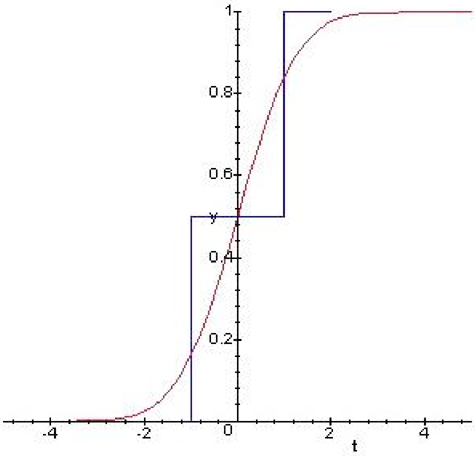


Figure 5.170: First approach of the Bernoulli distribution by the Normal distribution according to the CLT

For $n=2$:

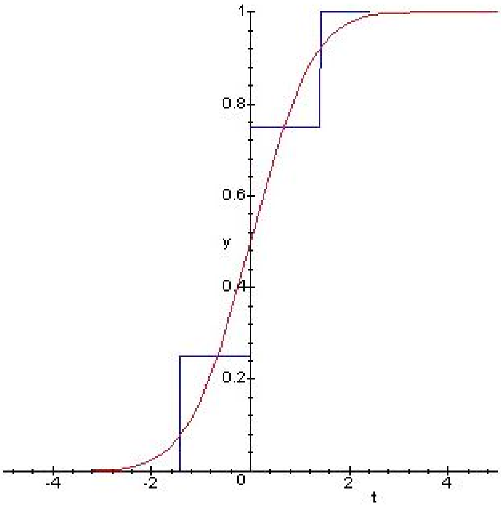


Figure 5.171: Second approach of the Bernoulli distribution by the Normal distribution according to the CLT

For $n = 5$:

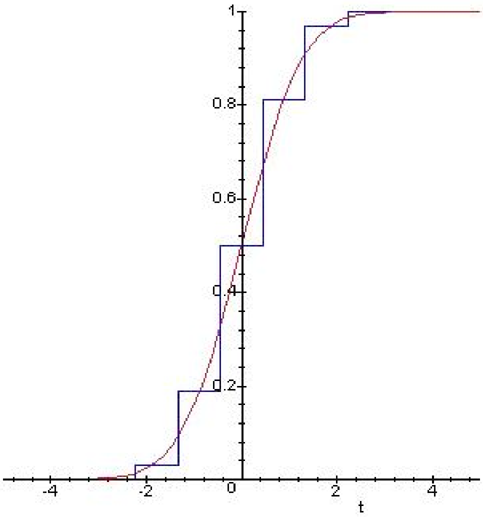


Figure 5.172: Fifth approach of the Bernoulli distribution by the Normal distribution according to the CLT

For $n = 30$:

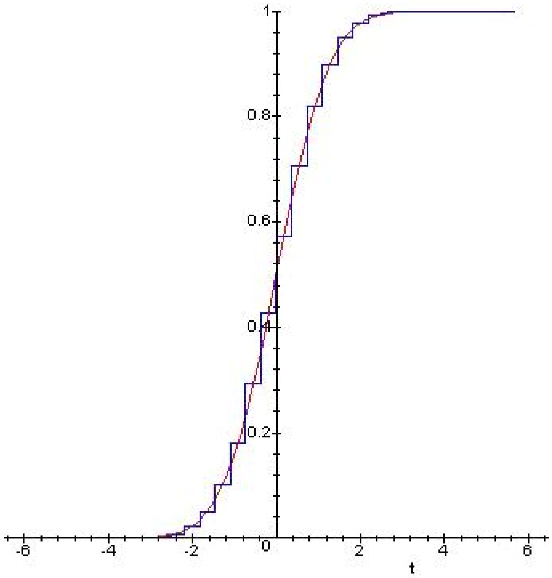


Figure 5.173: **Thirteenth approach of the Bernoulli distribution by the Normal distribution according to the CLT**

These graphs were obtained with Maple™ 4.00b using the following commands:

```
>with(stats):
>with(plots):
>e1:=plot(Heaviside(t+1)*statevalf[dcdf,binomiald[1,0.5]](trunc((t+1)/2))
,t=-2..2,y=0..1,color=blue):
>e2:=plot(Heaviside(t+sqrt(2))*statevalf[dcdf,binomiald[2,0.5]]
(trunc((t*sqrt(2)+2)/2)),t=-sqrt(2)-1..sqrt(2)+1,y=0..1,color=blue):
>e3:=plot(Heaviside(t+sqrt(5))*statevalf[dcdf,binomiald[5,0.5]]
(trunc((t*sqrt(5)+5)/2)),t=-sqrt(5)-1..sqrt(5)+1,y=0..1,color=blue):
>e4:=plot(statevalf[cdf,normald](t),t=-5..5):
>e5:=plot(Heaviside(t+sqrt(30))*statevalf[dcdf,binomiald[30,0.5]]
(trunc((t*sqrt(30)+30)/2)),t=-sqrt(30)-1..sqrt(30)+1,y=0..1,color=blue):
>display(e1,e4);
>display(e2,e4);
>display(e4,e3);
>display(e5,e4);
```

clearly show the convergence of F_n to Φ.

In fact we see that convergence is downright uniform which is confirmed by the "Moivre-Laplace central limit theorem":

Given X_n a sequence of independent random variables with the same Bernoulli parameter p, $0 < p < 1$. Therefore:

$$F_n(t) = P\left(2\sqrt{n}\left(\bar{X}_n - \frac{1}{2}\right) \leq t\right) = \Phi(t) \tag{5.7.1734}$$

tends uniformly to $\Phi(t)$ on $\mathbb{R}$ when $n \rightarrow +\infty$.

5.7.13 Univariate Hypothesis and Adequation Tests (NHST)

Definition (#5.7.142): A "statistical model" is a mathematical tool that embodies a set of statistical assumptions concerning the generation of sample data (and similar data from a larger population), often in considerably idealized form. It is usually specified as a mathematical relationship between one or more random variables and other non-random variables named "parameters". All statistical hypothesis tests and all statistical estimators are derived via statistical models!

During our study of confidence intervals, remember that we get the below few relations (it is only a sample of the most important one proved above!):

$$\bar{X}-Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\leq\mu\leq\bar{X}+Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{5.7.1735}$$

and:

$$\frac{nS_*^2}{\chi^2_{\alpha/2}(n)}\leq\sigma^2\leq\frac{nS_*^2}{\chi^2_{1-\alpha/2}(n)} \tag{5.7.1736}$$

and:

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}\leq\sigma^2\leq\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)} \tag{5.7.1737}$$

and finally:

$$\bar{X}-\frac{S}{\sqrt{n}}T_{\alpha/2}(n-1)\leq\mu\leq\bar{X}+\frac{S}{\sqrt{n}}T_{\alpha/2}(n-1) \tag{5.7.1738}$$

that allowed to do statistical inference based on the knowledge or not of the true mean or variance of the whole or on a sample of the population. In other words, under in what range stood a given moment (mean or variance) as a function of the chosen confidence level α.

We will also prove in details later the following two intervals:

$$\frac{1}{F_{\alpha/2}(n,m)}\frac{S^2_{*X}}{S^2_{*Y}}\leq\frac{\sigma^2_X}{\sigma^2_Y}\leq\frac{1}{F_{1-\alpha/2}(n,m)}\frac{S^2_{*X}}{S^2_{*Y}} \tag{5.7.1739}$$

and:

$$\frac{1}{F_{\alpha/2}(n-1,m-1)}\frac{S^2_{X}}{S^2_{Y}}\leq\frac{\sigma^2_X}{\sigma^2_Y}\leq\frac{1}{F_{1-\alpha/2}(n-1,m-1)}\frac{S^2_{*X}}{S^2_{*Y}} \tag{5.7.1740}$$

These are only a few of the more than 500 parametric and non-parametric confidence intervals - and their respective tests - that exist at the beginning of the 121st century (for more see the Data Science mind map at page 6770)!!!

Definition (#5.7.143): When we want to know if we can trust the value of a statistic (mean, median, variance, correlation coefficient, etc.) with some certainty, we speak of "hypothesis testing" and especially of "conformity test" (we speak of "adequation test" when the purpose is to check that measures will follow a particular distribution law).

Formally, a statistical hypothesis is a statement about the parameters of a distribution describing a population from which observations can be sampled as measured data. More precisely, let m be a generative model describing measured data y in terms of a distribution $\mathcal{D}(\theta)$ with model parameters $\theta \in \Theta$. Then, a statistical hypothesis is formally specified as:

$$H : \theta \in \Theta^* \quad \text{where: } \Theta^* \subset \Theta \tag{5.7.1741}$$

Let H be a statistical hypothesis. Then:

- H is named a "point hypothesis" or "exact hypothesis", if it specifies an exact parameter value:

$$H : \theta = \theta^* \tag{5.7.1742}$$

- H is named a "set hypothesis" or "inexact hypothesis", if it specifies a set of possible values with more than one element for the parameter value (e.g. a range or an interval):

$$H : \theta \in \Theta^* \tag{5.7.1743}$$

The statement which is tested in a statistical hypothesis test is named the "null hypothesis", denoted as H_0. The test is designed to assess the strength of evidence against H_0 and possibly reject it. The opposite of H_0 is named the "alternative hypothesis". Usually, H_0 is a statement that a particular parameter is zero, that there is no effect of a particular treatment or that there is no difference between particular conditions. More precisely, let m be a generative model describing measured data y using model parameters $\theta \in \Theta$. Then, a null hypothesis is formally specified as:

$$H_0 : \theta \in \Theta_0 \quad \text{where: } \Theta_0 \subset \Theta \tag{5.7.1744}$$

Let H_0 be a null hypothesis of a statistical hypothesis test. Then, the corresponding alternative hypothesis, denoted as H_1, is either the negation of H_0 or an interesting sub-case in the negation of H_0, depending on context. The test is designed to assess the strength of evidence against H_0 and possibly reject it in favor of H_1. Usually, H_1 is a statement that a particular parameter is non-zero, that there is an effect of a particular treatment or that there is a difference between particular conditions.

More precisely, let m be a generative model describing measured data y using model parameters $\theta \in \Theta$. Then, null and alternative hypothesis are formally specified as:

$$\begin{aligned} H_0 &: \theta \in \Theta_0 \quad \text{where: } && \Theta_0 \subset \Theta \\ H_1 &: \theta \in \Theta_1 \quad \text{where: } && \Theta_1 = \Theta \setminus \Theta_0 \end{aligned} \tag{5.7.1745}$$

Remark

The reader must also remember, as already said earlier in this section, that we have put many other confidence interval techniques detailed proofs related for example for regression techniques in the section of Theoretical Computing.

Hypothesis testing are intended to check whether the sample can be considered as extracted from a given population or be representative of this population vis-a-vis of parameter such as a mean, variance or the observed frequency. This implies that the theoretical distribution of the parameter is known at the population level. The hypothesis tests are not made to prove the null hypothesis (usually expressing equality or uniformity between different populations), but to eventually reject it (to be exact the rejection of the null hypothesis is more robust). In terms of the communication of statistical tests a number of experts recommend:

1. To always communicate the p-value (see page 948) with four decimal places (we will come back later on this concept).
2. Never say that a low p-value shows a significant magnitude of the effect studied because it is not necessary true (to check you just take a phenomenon of very small amplitude on a large sample and then the p-value will become very small by construction). Once again will discuss this more deeply later.
3. To always give the confidence interval of the test whether it is unilateral or bilateral.
4. To be careful to not to set a rejection threshold to the test except if a standard (norm) or legislation requires it (in which case we will specify which one).
5. To never say that the test is "proved" or "significant" or even "statistically significant". Just say that the result is "statistical" or that we have the "likelihood of the data knowing the null hypothesis" and that's it (in bayesian terms: «*the likelihood of the data is not the credence of the model*»)!
6. If the interest is to show the null hypothesis and that it is not rejected, since often its statistical power is low, you will need to repeat the experience to reinforce the conclusion.
7. If the interest is to reject the null hypothesis and that this is true, a good scientific practice is to look for additional studies that would faulty conclusion.
8. If there is for example no statistical difference between two values, therefore this does not mean that there is presence of statistical equivalence. It is then necessary to proceed to "equivalence tests".
9. The rejection of the null hypothesis does not mean that the mechanism of the phenomenon has been highlighted but just indicates, for refresh, an information about the size of the dataset a posteriori.
10. We communicate the posteriori power of the test.

To summarize, the studies must be communicated respecting the principle of truthfulness, after having been the subject of appropriate controls, and must be exposed, described and presented with impartially (some people say "a-theoric"). We must not confuse between objectives and speculatives results! The conclusions should be the most faithful expression of the content of facts and datas.

The reader has to know that Null Hypothesis Significance Testing (NHST) is the result of the evolution of two hypothesis testing theories, i.e. the Fisher's approach to data testing and the Neyman-Pearson's approach tests of acceptance. The final procedure is the incongruent combination of the previous two theories into the current approach: NSHT. We may also consider that Bayesian Hypothesis Testing (BHT) is itself a kind (...) of evolution of NHST. All these four hypothesis testing theories (Fisher, Neyman-Pearson, NHST, BHT) are complementary and the scientist should avoid to use only one of them when doing analyzing an experiment outcome. As already mentioned at the beginning of the book, different analytical approach are or first importance to compare the results by minimizing the practitioner bias. For more details see [435].

For example, if we want to know with some confidence whether a given mean of a population sample is realistic about the true unknown theoretical mean we will use the "Z-test" which is simply:

$$\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{5.7.1746}$$

Now remember that we have proved that if we have two random variables of law:

$$\mathcal{N}(\mu_1,\sigma_1) \qquad \mathcal{N}(\mu_2,\sigma_2) \tag{5.7.1747}$$

then subtracting (differentiate) the averages gives:

$$\mathcal{N}\left(\mu_2-\mu_1,\sqrt{\sigma_1^2+\sigma_2^2}\right) \tag{5.7.1748}$$

So for the difference of two average of random variables from two independent population samples we obtain directly:

$$\mathcal{N}\left(\bar{X}_2-\bar{X}_1,\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right) \tag{5.7.1749}$$

We can then adapt the Z-test as:

$$Z^{\text{calc}} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{\bar{X}-\mu}{\sqrt{\dfrac{\sigma^2}{n}}} = \frac{(\bar{X}_2-\bar{X}_1)-(\mu_2-\mu_1)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \tag{5.7.1750}$$

The relation that is very useful when for two samples of two populations, we want check if there is a statistically significant difference of the mean to a given fixed level of confidence level α and the associated probability for this difference:

$$\bar{X}_2-\bar{X}_1-Z\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \leqslant \mu_2-\mu_1 \leqslant \bar{X}_2-\bar{X}_1+Z\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \tag{5.7.1751}$$

Therefore:

$$|Z^{\text{calc}}| \leqslant |Z_{\alpha/2}| \tag{5.7.1752}$$

We then speak of "two-sample Z-test" and it is used lost in the industry to ensure the equality of two averages of two measurements of populations when standard deviation is known.

And if the theoretical standard deviation is not known, we use the "Student T-test" (used a lot in pharmacoeconomics) proved above:

$$\bar{X}-\frac{S}{\sqrt{n}}T_{\alpha/2}(n-1)\leq\mu\leq\bar{X}+\frac{S}{\sqrt{n}}T_{\alpha/2}(n-1) \tag{5.7.1753}$$

In the same idea for standard deviation, we use the "Chi-square variance test" as already proved above:

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}\leq\sigma^2\leq\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)} \tag{5.7.1754}$$

And when we want to test for the equality of variance of two populations we use "Fisher F-test" (that will be proved below during our study of the analysis of variance):

$$\frac{1}{F_{\alpha/2}(n-1,m-1)}\frac{S_X^2}{S_Y^2}\leq\frac{\sigma_X^2}{\sigma_Y^2}\leq\frac{1}{F_{1-\alpha/2}(n-1,m-1)}\frac{S_{*X}^2}{S_{*Y}^2} \tag{5.7.1755}$$

In practice we must be aware that the purpose of a test is very often to show that the effect is significant. It is then customary to say that the test is successful if the null hypothesis is rejected in favour of the alternative hypothesis! When the practitioner knows that the effect is significant and his test fails to reject the null hypothesis H_0 this is sometimes named the "dilemma of not rejecting the null hypothesis". As we will see a little further, the idea is then to calculate an "a posteriori test power" (it is then named by some softwares like SPSS, "observed power") and adapt the sample size accordingly to have an acceptable power according to tradition.

5.7.13.1 Direction of hypothesis tests and p-values

The fact that we get all values satisfying a right bounded test AND (!) left bounded is what we name in the general case a "two-tailed test" as it includes the left sided and right sided unilateral tests. Thus, all the above tests are in a bilateral form, but we could make a unilateral use too! We use one-tailed test when the expected difference (or difference to highlight) can only go in one direction (typically in the case of clinical trials or during a corrective quality action in the industry for which we expect a improvement going in one single direction). Unilateral tests are sometimes named "non-inferiority tests" (left-sided) or "non-superiority test" (right-sided).

Below we presented for example a right unilateral test (since the rejection region is on the right and therefore the cumulative probability is left-sided) and a two-sided test of 5% (since the beginning of the 121st century (holocene calendar), more and more practitioners recommend to take a value of 1%):

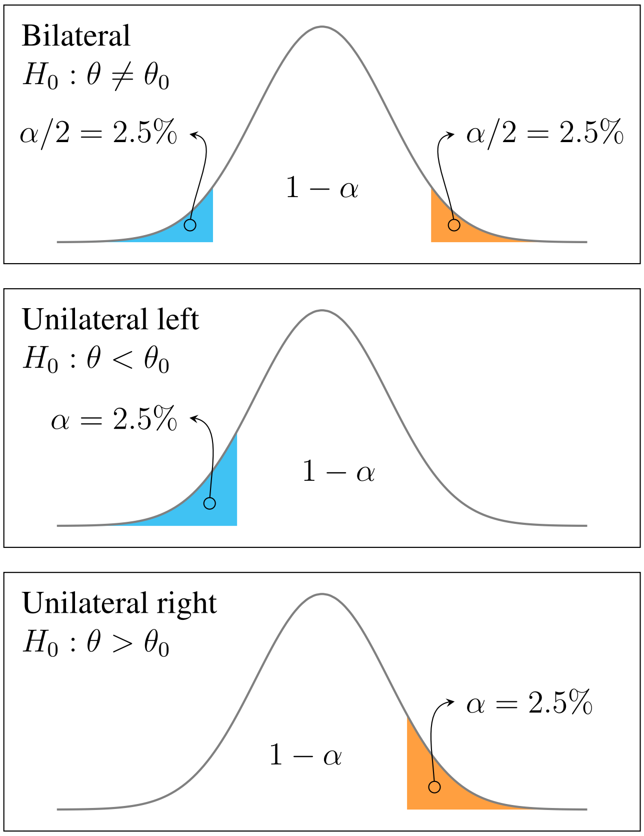

Bilateral
$H_0 : \theta \neq \theta_0$
$\alpha/2 = 2.5\%$
$\alpha/2 = 2.5\%$
$1-\alpha$
Unilateral left
$H_0 : \theta < \theta_0$
$\alpha = 2.5\%$
$1-\alpha$
Unilateral right
$H_0 : \theta > \theta_0$
$\alpha = 2.5\%$
$1-\alpha$

Figure 5.174: **Illustration of a test (or confidence level) unilateral and bilateral (author: Walmes Zeviani)**

We can also summarize how to determine the p-value (which will be discussed more in detail further below) with the following diagram:

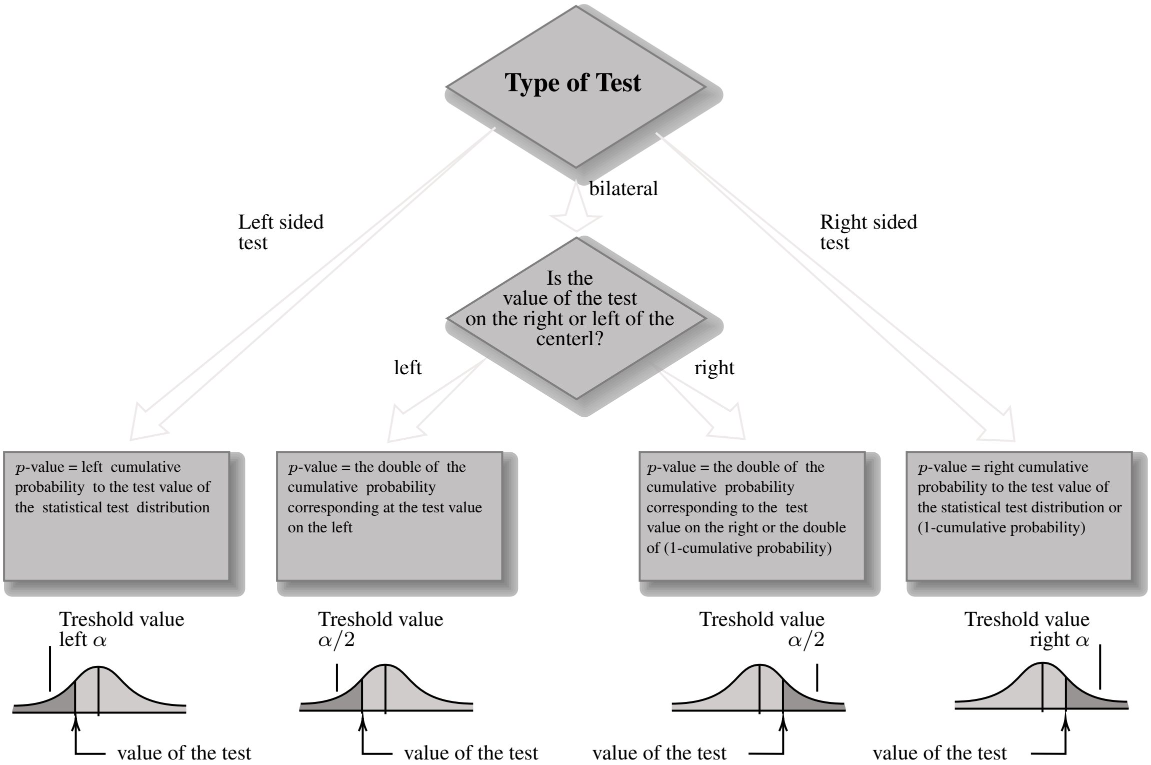

Type of Test
bilateral
Left sided test
Right sided test
Is the value of the test on the right or left of the centerl?
left
right
p-value = left cumulative probability to the test value of the statistical test distribution
p-value = the double of the cumulative probability corresponding at the test value on the left
p-value = the double of the cumulative probability corresponding to the test value on the right or the double of (1-cumulative probability)
p-value = right cumulative probability to the test value of the statistical test distribution or (1-cumulative probability)
Treshold value left α
Treshold value $\alpha/2$
Treshold value $\alpha/2$
Treshold value right α
value of the test
value of the test
value of the test
value of the test

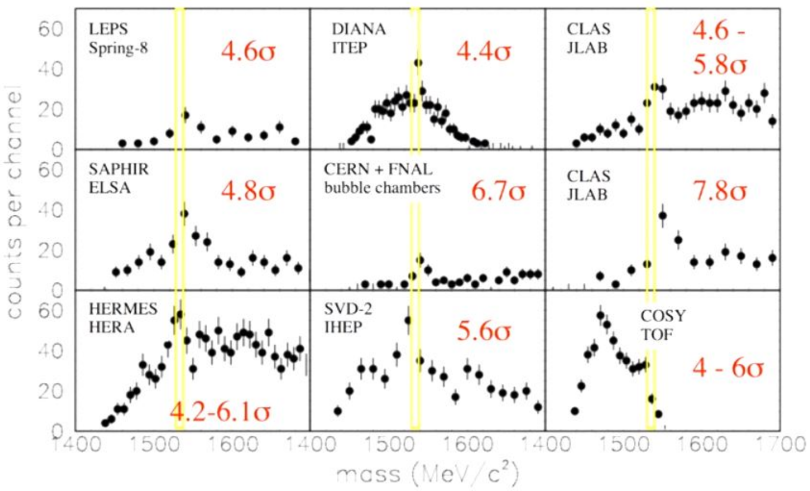
Figure 5.175: Summary figure to determine the p-value of parametrical test with symmetrical distributions

Definition (#5.7.144): The "p-value" is the likelihood of obtaining an effect at least as extreme as the one in your sample data, assuming the truth of the null hypothesis. In other words, if the null hypothesis is true, the p-value is the probability of obtaining your sample data. It answers the question, are your sample data unusual if the null hypothesis is true? If you are thinking that the p-value is the probability that the null hypothesis is true, the probability that you are making a mistake if you reject the null, or anything else along these lines, this is the most common misunderstanding of practitioners.

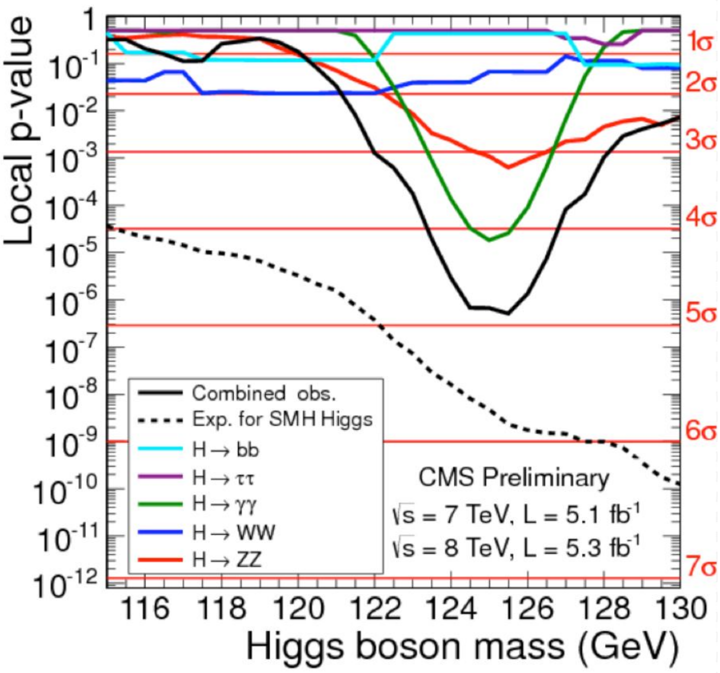
Remark

As the American Statistical Association - ASA wrote in its 12016 (holocene calendar) statement (see `https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1154108`): «*p-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.*» Instead if we are looking for a non-technical description, the ASA offers this statement: «*p-values can indicate how incompatible the data are with a specified statistical model.*»

As already mentioned, the p-value is often considered as significant if it is less than 5% or even better when... below 1%. In particle physics, according to the Gauss distribution (for which it is easy to convert percentages into standard deviation as we already know!), the p-value is often communicated in σ as shown in the random example below of pentaquark discovery studies in 12003 (holocene calendar):



or another random (well-known random...) example is that of the Higgs Boson confirmation discovery at the LHC/CERN in 12012[51] (holocene calendar):



The common misconception is what we'd really like to know. We would loooove to know the probability that a hypothesis is correct, or the probability that we're making a mistake. What we get instead with these frequentist methods is the probability of our observation given the null hypothesis, which just isn't as useful.

It would be great if we could take evidence solely from a sample and determine the probability that the sample is wrong. Unfortunately, that seems not possible with the frequentist methods and this for logical reasons when you think about it. Without outside information, a sample can't tell you whether it's representative of the population.

---

[51] The standard of certainty that particle physicists require at the beginning of the 121st century (holocene calendar) for a discovery is roughly 1 in 3.5 million, i.e. 5σ.

Sigma (unilateral Normal Law)/ Z	p-value	Field
1	84.1344746%	Not used in practice
1.64485363	95.0000000%	Typical in undergraduate curriculum
2	97.7249868%	Not used in practice
2.32634789	99.0000000%	Typical in 21 st century social science
3	99.8650102%	Not used in practice
4	99.9968329%	Not used in practice
5	99.9999713%	Used in particle physics

Table 5.51: **Conversion sigma to p-value in unilateral**

Sigma (bilateral Normal Law)	p-value	Field
1	68.2689492%	Not used in practiced
1.959963988	95.0000000%	Typical in undergraduate curriculum
2	95.4499736%	Not used in practiced
2.5758293	99.0000000%	Typical in 21st century social science
3	99.7300204%	Not used in practiced
4	99.9936658%	Not used in practiced
5	99.9999427%	Used in particle physics

Table 5.52: **Conversion sigma to p-value in bilateral**

The p-values are based exclusively on information contained within a sample. Consequently, p-values can't answer the question that we most want answered, but there seems to be an irresistible temptation towards interpreting it that way.

Let us also note that the hypothesis tests on the standard deviation (variance), the mean or correlation are named "parametric tests" in reverse of the nonparametric tests that we will see much further below.

Remarks

R1. There is also another definition of the concept of parametric and nonparametric tests that we will see further below (a bit different because more precise).

R2. Warning! Some authors or teachers sometimes talk about "left-sided" for a "right-sided test"... In fact it is simply a choice of vocabulary. If the reference of teaching is not the rejection area but the acceptance area, then it is clear that the right and left are concepts that are reversed...

Finally, many software calculate what we name therefore the "p-value" which is the limit calculated risk (probability threshold) α that might have set the statistician to be at the border between acceptance of the null hypothesis and its rejection (remember that a successful test does not prove anything and a rejection is better but still do not prove anything!). So the p-value is a fundamental value in statistics because it gives the possibility to quantify the likelihood of the null hypothesis H_0 (acceptance or rejection).

Definition (#5.7.145): But strictly speaking the "p-value" is the conditional (Bayesian) probability, that our data satisfy the null hypothesis (that H_0 is true[52]), i.e. the likelihood of results:

$$\boxed{p\text{-value}=P(\text{data}|H_0)} \tag{5.7.1756}$$

and (sadly...) not the probability of the null hypothesis knowing the data (plausibility of H_0, i.e. $P(H_0|\text{data})$)!:

$$p\text{-value}\neq P(H_0|\text{data}) \tag{5.7.1757}$$

While the difference may be small as we saw in the section Probabilities, it is not zero! So in reality the p-value says nothing about the hypothesis itself, but it provides information on the experimental data if the experiment is conducted infinite number of times and all relevant assumptions hold! So as a p-value simply gives information only in the case we would repeat the experiment a large number of times we understand better why p-values are often decried in single experiment studies...

As Ronald Fisher famously said: «*A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance*». What this means is p-value less than 0.05 (or as applicable) only tells you that one should repeat the experiment!

It is true that the smaller the p-value, the more unusual the data would be if every single assumption were correct; but a very small p-value does not tell us which assumption is incorrect. For example, the p-value may be very small because the targeted hypothesis is false; but it may instead (or in addition) be very small because the study protocols were violated, or because it was selected for presentation based on its small size. Conversely, a large p-value indicates only that the data are not unusual under the model, but does not imply that the model or any aspect of it (such as the targeted hypothesis) is correct; it may instead (or in addition) be large because (again) the study protocols were violated, or because it was selected for presentation based on its large size!

The general definition of a p-value may help one to understand why statistical tests tell us much less than what many think they do: Not only does a p-value not tell us whether the hypothesis targeted for testing is true or not; it says nothing specifically related to that hypothesis unless we can be completely assured that every other assumption used for its computation is correct, an assurance that is lacking in far too many studies.

Especially keep in mind (because we have a natural cognitive bias about this...) that in general:

$$P(\text{data}|H_0)\neq P(H_0|\text{data}) \tag{5.7.1758}$$

But these two probabilities (likelihood on the left and plausibility on the right) are however linked by the Bayes theorem (see section Probabilities page 492):

$$P(H_0|\text{data})=\frac{P(\text{data}|H_0)P(H_0)}{P(\text{data})} \tag{5.7.1759}$$

---

[52] Equivalently we can say that it is the conditional probability of obtaining data as extreme as the one that was actually observed, assuming that the null hypothesis is true! Another nice way to describe, it is to say that the p-value is the proportion of times that you would see evidence stronger than what was observed, against the null hypothesis, if the null hypothesis were true and you hypothetically repeated the experiment (sampling of individuals from a population) a large number of times!

But the both terms $P(H_0)$ (prior probability of the observed phenomenon) and $P(\text{data})$ are in practice difficult or even impossible to know.

For hypothesis testing, for example, the 5% risk threshold α is the risk to reject the null hypothesis H_0 even though it is true. If the risk imposed/chosen is 5% and the calculated p-value is less (in most tests but be careful because this is not a generality!!!), the test fails (rejection of the null hypotheses) in favour of an alternative hypothesis denoted by H_1 or H_a. Never forget that reject the test is always better in term of power than accepting it. In other words we can say that the p-value is the smallest α at which we would reject H_0.

In a paper published the March 7, 12016 (holocene calendar) by the American Statistical Association we have a list of principles to improve the conduct and interpretation of quantitative science relatively to the usage of the p-value! The statement's six principles, many of which address misconceptions and misuse of the p- value, are the following:

1. p-values can indicate how incompatible the data are with a specified statistical model.
2. p-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

The alternative hypothesis H_1 has of course itself its own risk that we denote by β and its own p-value. So when the null hypothesis is not rejected, the risk associated with this decision is a "risk of the second kind". To assess this, we should calculate the power of the test considered (see proofs further below).

Perhaps, to better understand, here is an illustration of a particular case of an bilateral hypothesis test of the average for a typical random variable following a Normal distribution (basically it's almost the same principle for all tests...):

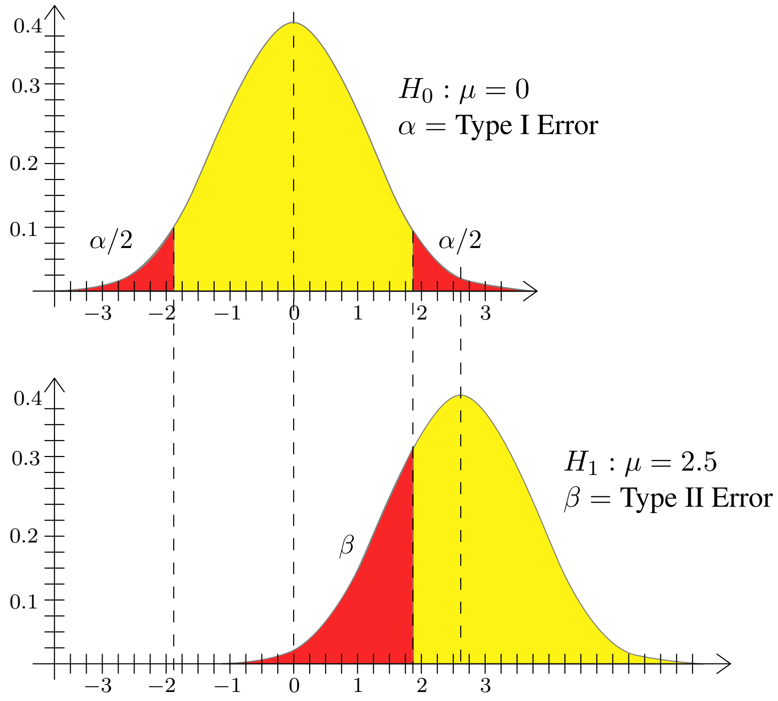


Figure 5.176: Null and alternative hypothesis of a special case of two-sided test

Thus, in the case presented above, we see better why the null hypothesis can be accepted or rejected in favour of the alternative hypothesis (which is of the same law that the null hypothesis but just shifted) depending on the reference value measured that will be used for the test (in the special case of the above figure it is the arithmetic mean of measurements).

We also note that the red area of the alternative hypothesis, corresponding to the cumulated probability β, is partially merged with the yellow part of the null hypothesis. This is why we can sometimes accept the null hypothesis wrongly. However, we see that smaller is β, more the alternative hypothesis would be far from the red boundary zone of the null hypothesis (this would correspond to a translation to the right in this case) and less the likelihood of a false conclusion is big. This is why we talk about "risk β" because smaller it is, the better. Verbatim more $1-\beta$ is big, then smaller is the risk of confusing the null and alternative hypothesis. This is why $1-\beta$ is named "power of the test" (see below the subsection that is devoted to this concept).

We (sadly) do not reject the null hypothesis if the p-value is by tradition greater than 5% (0.05). In fact, bigger is the p-value, the better it is because for the people interested in the null hypothesis because the confidence interval is becoming smaller. If the confidence interval is becoming huge (very close to 100%) because the p-value is very small then the analysis does not really make sense physically speaking anymore in the point of view of the null hypothesis!

Thus, if the p-value is small, that mean we should take a low risk α of error, therefore accept (i.e. do not reject) in almost all cases the tested hypothesis (H_0)...

Remarks

R1. We should never say that we "accept" a hypothesis or that it is "true" or "false" as these terms are too strong and might suggest a scientific proof. We should say if we "reject" or "does not reject" the null hypothesis and it may be possibly "correct" or "incorrect".

R2. For bilateral test hypotheses, we can for example say that we have (or not have) a significant difference between the measured reference value and the expected value. For one-sided tailed, we can say that the measured reference value is significantly bigger or smaller than the expected value.

R3. Moreover if the reader has well understand the construction of hypothesis testing, the fact of wrongly reject a hypothesis ("Type I error" or "Error of the first kind") is more robust than accept it wrongly ("Type II error" or "Error of the second kind").

R4. The reader will also have noticed with the help of the previous figure that a one sided test has a higher power than a one-sided test (for a same risk threshold of course!). Thus, a statistically insignificant difference in bilateral test, can be statistically significant in a one sided test.

R5. If the p-value is close to the limit threshold (rejection) value we say that the "effect is marginally significant".

A big issue with hypothesis testing[53] is what we name "p-hacking": the underlying idea consists of replicating a test dozens of times until it provides the conclusion that the experimenter wants... or try all possible combinations of variables until we found something significant or even just take big samples (as we know that NHST tests fail systemically when samples are big).

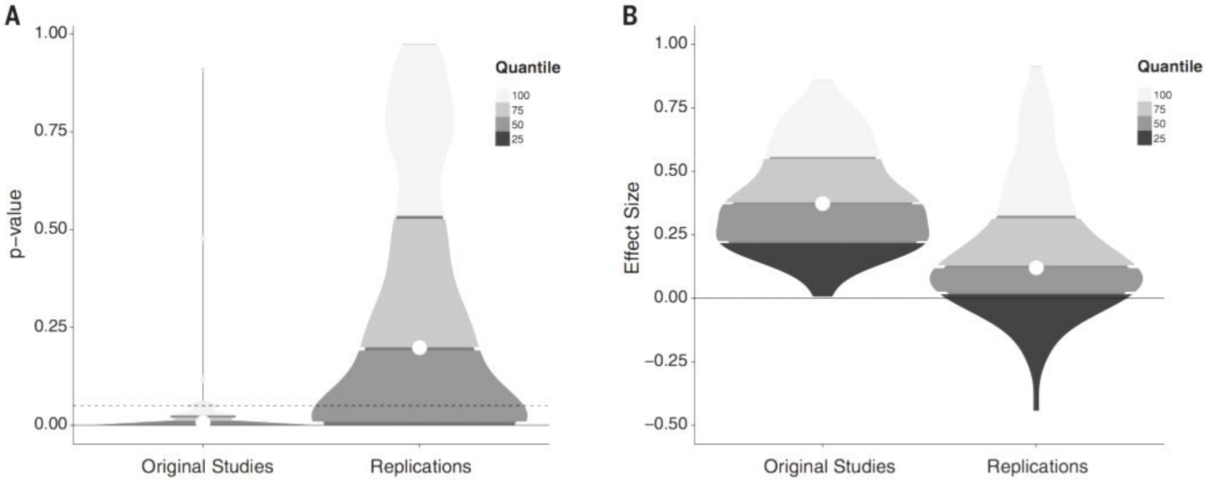


Figure 5.177: p-hacking Open Science Collaboration, *Estimating the reproducibility of physiological science*, Science, 349 (12015)

p-hacking is generally thought of as cheating, but what if we made it compulsory instead? If the purpose of studies is to push the frontiers of knowledge, then perhaps playing around with different methods shouldn't be thought of as a dirty trick, but encouraged as a way of exploring boundaries. A recent project spearheaded by Brian Nosek, a founder of the non-profit Center for Open Science, offered a clever way to do this.

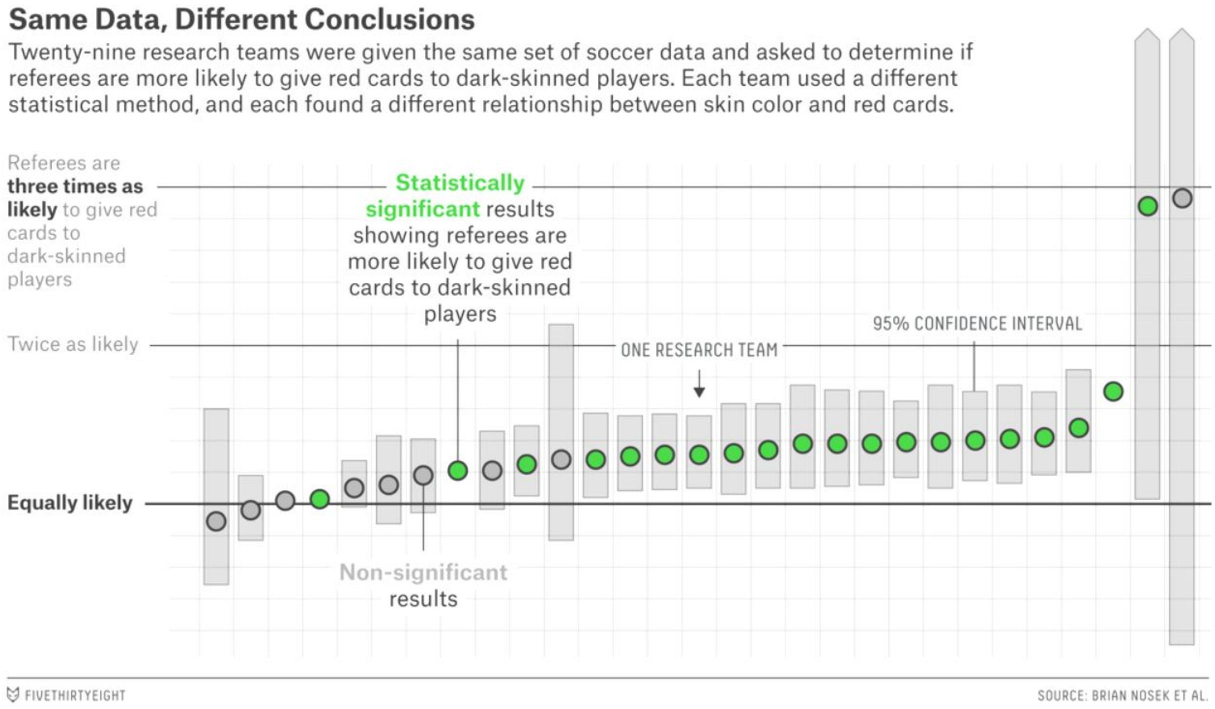
---

[53] A second one is the "significance filter curse" quite well documented in [602].

Nosek's team invited researchers to take part in a crowd-sourcing data analysis project. The setup was simple. Participants were all given the same data set and prompt: Do soccer referees give more red cards to dark-skinned players than light-skinned ones? They were then asked to submit their analytical approach for feedback from other teams before diving into the analysis.

Twenty-nine teams with a total of 61 analysts took part. The researchers used a wide variety of methods, ranging - for those of you interested in the methodological gore - from simple linear regression techniques to complex multilevel regressions and Bayesian approaches. They also made different decisions about which secondary variables to use in their analyses.

Despite analysing the same data, the researchers got a variety of results. Twenty teams concluded that soccer referees gave more red cards to dark-skinned players, and nine teams found no significant relationship between skin color and red cards.



Same Data, Different Conclusions

Twenty-nine research teams were given the same set of soccer data and asked to determine if referees are more likely to give red cards to dark-skinned players. Each team used a different statistical method, and each found a different relationship between skin color and red cards.

SOURCE: BRIAN NOSEK ET AL.

The variability in results wasn't due to fraud or sloppy work. These people are assumed to be highly competent analysts who were motivated to find the evidence, said Eric Luis Uhlmann, a psychologist at the Insead business school in Singapore and one of the project leaders. Even the most skilled researchers must make subjective choices that have a huge impact on the result they find.

But these disparate results don't mean that studies can't inch us toward strong evidence. The important lesson here is that a single analysis is not sufficient to find a definitive answer.

However generally the reader must keep in mind that if there are multiple statistical tests that apply for exactly the same thing and work differently (quite common in statistics like the multiple homogeneity of variance tests, the multiple normality tests, the multiple unit root tests, the multiple stationarity tests, etc.), then a single significant result leads to rejection!

Caution! Furthermore statistical significance has literally zero connection with "practical significance" also named "scientific importance" or "relevance". It will never tell you if something is important or not! Indeed, it's not an indicator knowing the context! That a null hypothesis is rejected or not only means that the hypothesis test has not met the pre-set significance level α and not that the results are significant or not significant for practical applications (they very likely are to the scientist spending nights in the lab collecting the data)! The test statistics typically just looks at the effect (difference, ratio, covariance, etc) and the variance. In other words, it just looks if the signal is visible enough on the background of the noise. It takes the current condition - sample size - into consideration. More data means that smaller effects will be indicated as statistically significant.

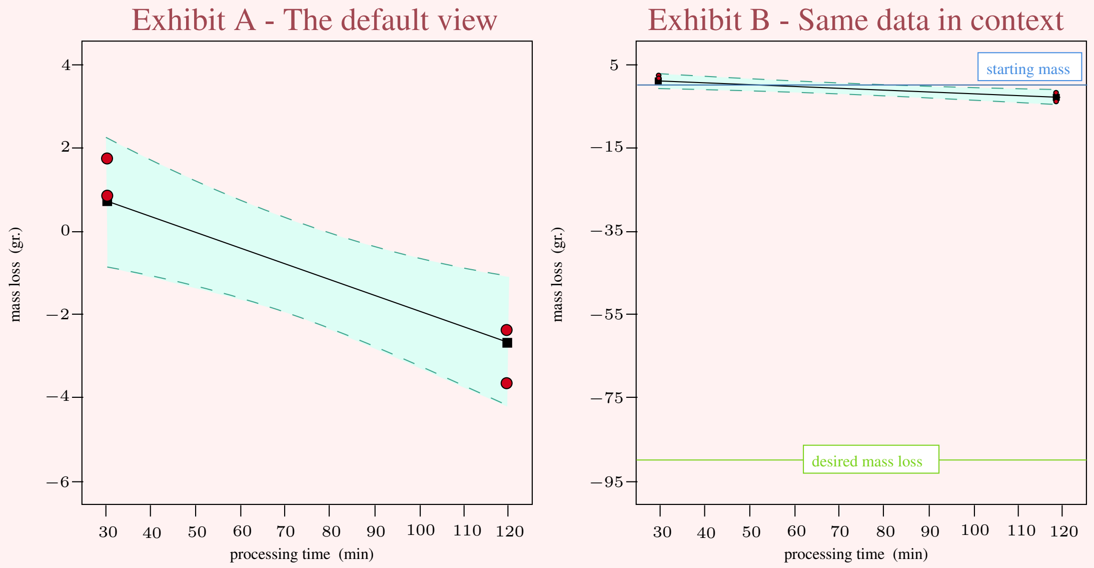


Exhibit A - The default view
Exhibit B - Same data in context

Figure 5.178: High statistical significance which contextualized has low practical significance (source: Danish Technological Institute)

To decide whether something is practically significant, we need the minimal meaningful difference threshold! It doesn't come from some "blind" mathematical procedure, ignoring the context, but rather from the domain knowledge: experts, literature review, your experience, pilot studies, etc. In medicine we name it MCID for "minimum clinically important difference". And we can embed it into statistical hypothesis testing via interval hypothesis, like in non-inferiority, superiority or equivalence clinical studies.

> *Evidence based medicine has taken centuries to advance humans to the point where people are so healthy and living so long that they can deny that science, statistics and medicine is what got us here.*

quote 5.7:958.#48

Keep in mind that as we said in the introduction: What makes science so powerful is that it's self-correcting. Sure, false findings get published, but eventually new studies come along to overturn them, and new evidence is revealed. At least, that's how it's supposed to work. But scientific publishing doesn't have a great track record when it comes to self-correction, but this is changing with new technologies that facilitates such tracking with computer databases.

So the scientific method is still a posteriori far away the most reliable method (less worst) that any other way of knowing that we have TODAY (but it's not an argument however to not try to find a better one)!

Remark

By the way... the reader can try to remember an adage named after economist Charles Goodhart, which has been phrased by Marilyn Strathern as: «*When a measure becomes a target, it ceases to be a good measure.*» But the original formulation was: «*Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes.*»

Definitions (#5.7.146):

D1. The "probability α of Type I error" (error of the first kind / false negative[54]) is the probability of rejecting the null hypothesis when in reality it is likely to be true[55]. In statistics, the "size of a test" is the probability of falsely rejecting the null hypothesis. That is, it is the probability of making a type I error. It is denoted by the Greek letter α (alpha). For a simple null hypothesis, the size is determined by the following conditional probability:

$$P\left(\text{ test rejects } H_0 \mid H_0\right) \tag{5.7.1760}$$

D2. The "probability β of Type II error" (error of the second kind / false positive) is the probability of maintaining the null hypothesis when it is likely to be false. Since, by definition, power is equal to $1-\beta$, the power of a test will get smaller as beta gets bigger. The "power of a test" is the probability of rejecting H_0, given that H_1 is actually true. That is, for given null and alternative hypothesis, the power is determined by the following conditional probability:

$$P\left(\text{ test doesn't reject } H_0 \mid H_1\right) \tag{5.7.1761}$$

	State of Nature	
Test outcome	H_0 **false**	H_0 **true**
Reject H_0	Correctly rejected null hypothesis ($1-\beta$: power of the test)	Type I Error (false negative) (Risk α)
Fail to reject H_0	Type II Error (false positive) (Risk β)	Correctly accepted null hypothesis ($1-\alpha$: confidence level)

---

[54] Notice that depending on the point of view of the scientist, a "false negative" can be named "false positive", this is why many textbooks avoid using this specific vocabulary!

[55] Please keep in mind that as already mentioned in the introduction of this textbook, we are speaking here not about the "philosophical truth", but the "truth" as the best explanation at our actual level of knowledge!!!

Or in a more mathematical form:

	State of Nature	
Test outcome	H_0 **false**	H_0 **true**
Reject H_0	Correctly rejected null hypothesis $P(\text{reject } H_0 \vert H_0 \text{ is false}) = 1-\beta$	Type I Error (false negative) $P(\text{reject } H_0 \vert H_0 \text{ is true}) = \alpha$
Fail to reject H_0	Type II Error (false positive) $P(\text{fail reject } H_0 \vert H_0 \text{ is false}) = \beta$	Correctly accepted null hypothesis $P(\text{fail reject } H_0 \vert H_0 \text{ is true}) = 1-\alpha$

Thus a traditional criteria selection is to use the following principle: among all the tests that have the same size of Type I error, choose one that has the smallest size of the value of the Type II error.

Thus, it can be said that a hypothesis test simply has for purpose to minimize α the error of Type I and, on the contrary, to maximize $1-\beta$, the power of the test.

In general, the magnitude of the Type II error increases when that of the Type I error decreases. We can not, as far as we know, minimize the two errors at the same time. For this reason, we often take a fixed value of α, size of the type I error, and we minimize β, magnitude of the Type II error (i.e. we increase the power $1-\beta$).

Remark

In 12017 (holocene calendar) the false negative rate in the general U.S.A. population for HIV test is around 0.003%, or three times out of every $100,000$ tests.

False positive rates are far lower, between 0.0004% and 0.0007% due to the practice by which the initial positive result is confirmed with a secondary test.

Other example, false negatives occur in 1 of every $2,500$ women screened (NHS Breast Screening Programme – *Helping you decide* leaflet of July 12013 according to holocene calendar) and occur more often in younger women than in older women because the breast tissue of younger women is denser.

To close this subject, here are the three types of situations testing hypotheses on the statistics that is the average in the framework of an underlying Normal distribution and whose mean is in this particular case zero and of unit variance (because we can often come back to this particular case by centering and reducing the underlying random variable or by doing special transformations):

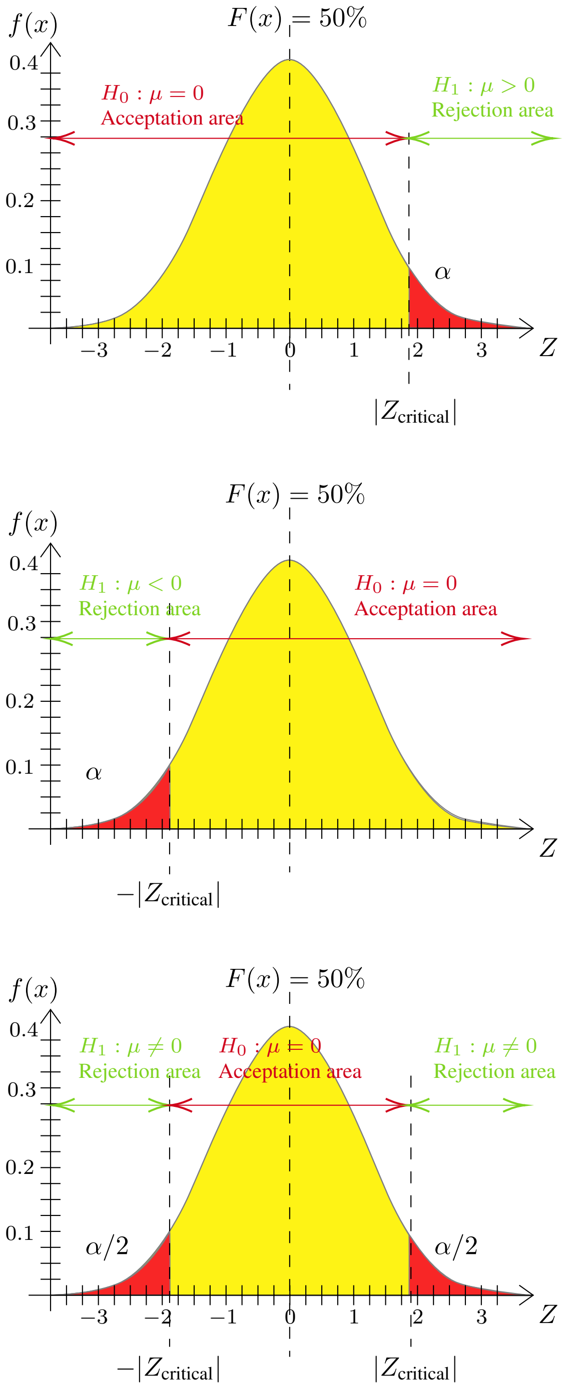


Figure 5.179: **The three possible scenarios for a hypothesis test on the average**

Let us indicate that it makes no sense (as opposed to what we can sometimes read in some paper or electronic media) to have the following null hypotheses in the special case shown above:

$$\begin{gathered} H_0 : \mu \leq 0 \\ H_0 : \mu \geq 0 \\ H_0 : \mu \neq 0 \end{gathered} \tag{5.7.1762}$$

with the alternative hypothesis that automatically follows (we did not write it because it is useless and obvious). The reason is simple: how could you position your reduced centered

Normal distribution if the mean is not fixed ... ??? This is why the null hypothesis in the context of the tests on the mean (and some other tests) are always written with an equality!

Obviously the reader has very likely noticed that:

- If the confidence interval contains the parameter of interest relatively to H_0 the test is not significant
- If the confidence interval excluded the parameter of interest relatively to H_0 the test is significant

To summarize, we can say that if we make a decision, we can make an error and it is best not to make mistakes often. Clearly, the probability of saying something stupid must be known and preferably small.

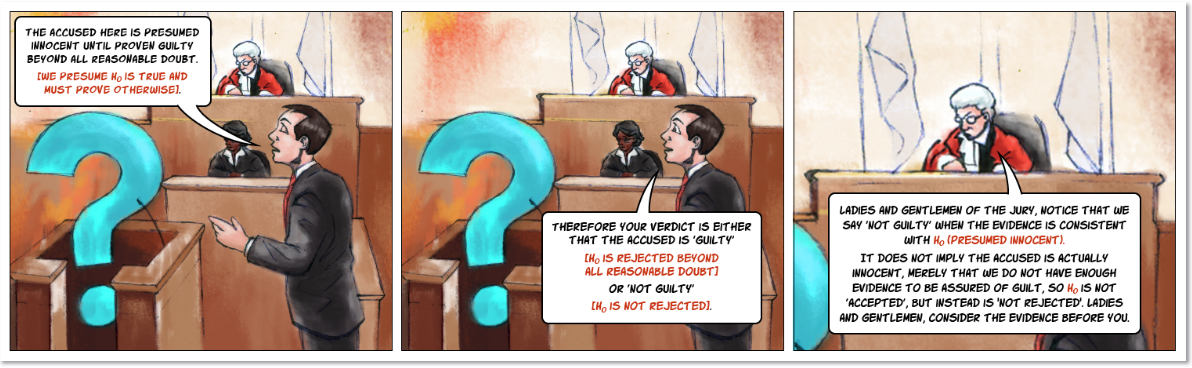
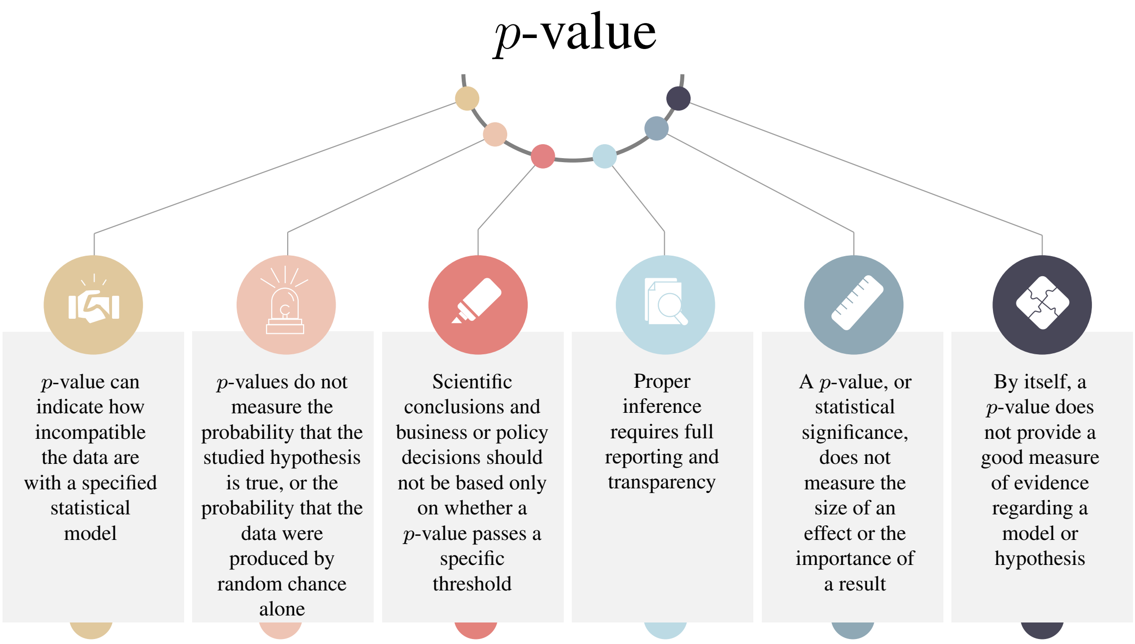

THE ACCUSED HERE IS PRESUMED INNOCENT UNTIL PROVEN GUILTY BEYOND ALL REASONABLE DOUBT. [WE PRESUME H_0 IS TRUE AND MUST PROVE OTHERWISE].
THEREFORE YOUR VERDICT IS EITHER THAT THE ACCUSED IS 'GUILTY' [H_0 IS REJECTED BEYOND ALL REASONABLE DOUBT] OR 'NOT GUILTY' [H_0 IS NOT REJECTED].
LADIES AND GENTLEMEN OF THE JURY, NOTICE THAT WE SAY 'NOT GUILTY' WHEN THE EVIDENCE IS CONSISTENT WITH H_0 (PRESUMED INNOCENT). IT DOES NOT IMPLY THE ACCUSED IS ACTUALLY INNOCENT, MERELY THAT WE DO NOT HAVE ENOUGH EVIDENCE TO BE ASSURED OF GUILT, SO H_0 IS NOT 'ACCEPTED', BUT INSTEAD IS 'NOT REJECTED'. LADIES AND GENTLEMEN, CONSIDER THE EVIDENCE BEFORE YOU.

Figure 5.180: **Comics illustrating Null and Alternative hypothesis (author: ?)**

But keep in mind that a better understanding of the p-value will not take away the human impulse (cognitive bias) to use statistics to create an impossible level of confidence. Most people want something that they can't really get, and that they don't really understand because of a lack of education and critical thinking. They want certainty! And even if you try to explain them you will have a backfire effect as they suffer from the anti-intellectualism syndrome.

Caution! In some high level statistical softwares the p-value may not match the confidence interval. For example, in R with the proportion test (see page 872), the p-value is computed using the Rao score test (based on pooled variance), the one-sided interval is computed with the Wilscon score and the two sided confidence interval is computer with the Wald (normal) non-pooled variance. The same kind of scenario in R (and other statistical softwares) happens with the Fisher's exact test (see page 1254).



p-value can indicate how incompatible the data are with a specified statistical model

p-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone

Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold

Proper inference requires full reporting and transparency

A p-value, or statistical significance, does not measure the size of an effect or the importance of a result

By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis

5.7.13.2 Fisher's method for multiple p-values

In statistics, "Fisher's method, also known as "Fisher's combined probability test", is a technique for data fusion or "meta-analysis" (analysis of analyses). It was developed by and named for Ronald Fisher. In its basic form, it is used to combine the results from several independent tests bearing upon the same overall hypothesis (H_0).

Consider a set of k independent tests, each of these to test a certain null hypothesis $H_{0|i}, i = \{1, 2, \ldots, k\}$. For each test, a significance level p_i, i.e., a p-value, is obtained. All these p-values can be combined into a joint test whether there is a global effect, i.e., if a global null hypothesis H_0 can be rejected.

There are a number of ways to combine these independent, partial tests. The Fisher method is one of these, and is perhaps the most famous and most widely used. The test was presented in Fisher's now classical book, *Statistical Methods for Research Workers* (11932 according to holocene calendar).

The test is based on the fact that the probability of rejecting the global null hypothesis is related to intersection of the probabilities of each individual test, $\prod_i p_i$. However, $\prod_i p_i$ is not uniformly distributed (but each p_i is assumed to be uniformly distributed!), even if the null is true for all partial tests, and cannot be used itself as the joint significance level for the global test. To remediate this fact, some interesting properties and relationships among distributions of random variables were exploited by Fisher and embodied in the succinct excerpt above. The proof is given below:

Proof 5.48.3. So we are looking for the distribution law of:

$$X = \prod_i p_i \tag{5.7.1763}$$

but this seems quite difficult to do as we are dealing with a product. Then we take first logarithm (the natural one as we know in statistics that most of time in is the $\ln$ that appears:

$$\ln(X)=\ln\left(\prod_i p_i\right)=\sum_i \ln(p_i) \tag{5.7.1764}$$

Therefore we have now a sum of the natural logarithm of uniform distribution $\mathcal{U}_{0,1}$. It's a bit better but now we have the problem of the $\ln\left(\mathcal{U}_{0,1}\right)$.

But now let us see something nice! Remember that the cumulative distribution function of an exponential distribution is:

$$F(x)=1-e^{-\lambda x} \tag{5.7.1765}$$

The inverse is then given by:

$$x=-\frac{1}{\lambda}\ln(1-F(x)) \tag{5.7.1766}$$

If P is a random variable uniformly distributed in the interval $[0,1]$, so is $1-P$. As consequence the previous relation can be equivalently written as:

$$x=-\frac{1}{\lambda}\ln(P) \tag{5.7.1767}$$

where once again: $P=\mathcal{U}_{0,1}$. Therefore we have just found that the natural logarithm of a uniform random variable in the interval $[0,1]$ follows an exponential distribution with parameter $\lambda=1$!

But now we have a small a problem. As we have a sum of exponential laws we have never proved if the exponential law is stable by addition...

Let us now recall that the chi-squared distribution with k degrees of freedom is given by:

$$\chi^2_k(x)=\frac{1}{2^{\frac{k}{2}}\Gamma(k/2)}x^{\frac{k}{2}-1}e^{-\frac{x}{2}}\cdot 1_{[0,+\infty[} \tag{5.7.1768}$$

If $k=2$ we get:

$$\chi^2_2(x)=\frac{1}{2^{2/2}\Gamma(2/2)}x^{\frac{2}{2}-1}e^{-\frac{x}{2}}\cdot 1_{[0,+\infty[}=\frac{1}{2\cdot 1}x^0e^{-x/2}=\frac{1}{2}e^{-x/2} \tag{5.7.1769}$$

In other words, chi-squared distribution with two degrees of freedom is equal to a exponential distribution with $\lambda=1/2$!

And remember now during our study of the Gamma distribution we have proved that it was stable by addition. And as we have proved that the chi-square distribution follows from a special case of the Gamma distribution, and just now that the exponential distribution follows from a special case of the chi-squared distribution, then that latter is also stable by addition with degrees of freedom!!!!

Therefore multiplying the sum of logarithms by -2 to get rid of the $-$ coming from the exponential law and the $1/2$ coming from what bounds the chi-squared law to the exponential law, we get:

$$\chi^2_{2k}=-2\sum_{i=1}^{k}p_i \tag{5.7.1770}$$

from which a p-value for the global hypothesis can be easily obtained.

■ Q.E.D.

Under Fisher's method, two small p-values p_1 and p_2 combine to form a smaller p-value. The yellow-green boundary in the figure below defines the region where the meta-analysis p-value is below 0.05. For example, if both p-values are around 0.10, or if one is around 0.04 and one is around 0.25, the meta-analysis p-value is around 0.05:

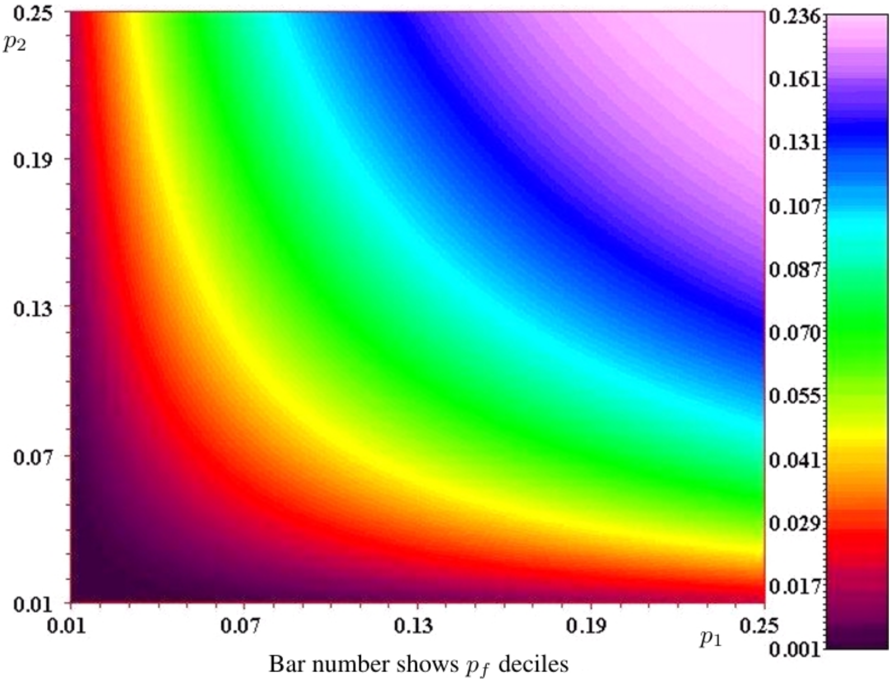


Figure 5.181: Fisher's method for multiple p-values (source: Wikipedia)

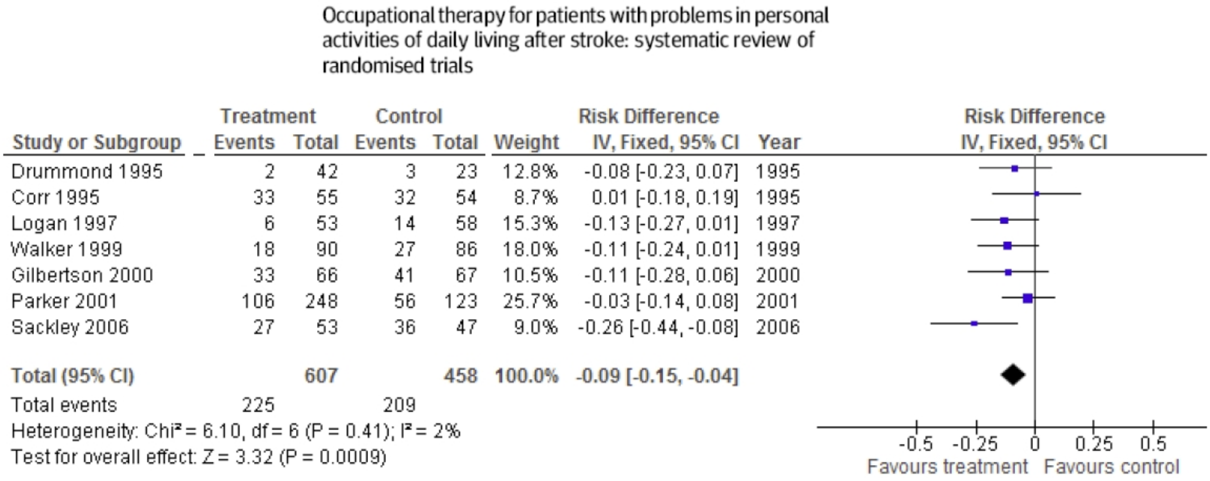
Remark

In the famous judgement case of Lucia de Berk the court made heavy use of statistical calculations to achieve its conviction. In a 12003 (holocene calendar) TV special of NOVA, Dutch professor of Criminal Law Theo de Roos stated: «*In the Lucia de Berk case statistical evidence has been of enormous importance. I do not see how one could have come to a conviction without it*». The law psychologist Henk Elffers, who was used by the courts as expert witness on statistics both in the original case and on appeal, was also interviewed on the programme and stated that the chance of a nurse working at the three hospitals being present at the scene of so many unexplained deaths and resuscitation is one in 342 million. But sadly, this value was wrongly calculated (this is why any important statistical calculations should be peer-reviewed always at least by three other independent people) by making a simple calculations of the multiplication of the p-values... Furthermore, always in the context of Lucai de Berk the statisticians Richard D. Gill and Piet Groeneboom calculated a chance of one in twenty-five that a nurse could experience a sequence of events of the same type as Lucia de Berk.

The use of probability arguments in the De Berk case was discussed in a 12007 (holocene calendar) *Nature* article by Mark Buchanan. He wrote: « *The court needs to weigh up two different explanations: murder or coincidence. The argument that the deaths were unlikely to have occurred by chance (whether* 1 *in* 48 *or* 1 *in* 342 *million) is not that* *meaningful on its own - for instance, the probability that ten murders would occur in the same hospital might be even more unlikely. What matters is the relative likelihood of the two explanations! However, the court was given an estimate for only the first scenario. »*

So we see here that using simple multiplication of Fisher multiplication inference is not the point in such a situation!

Notice that the use of Fisher's method works only if the studies included in a meta-analysis are comparable (sample size). If they aren't we have to use other techniques permitting the usage of weights like illustrated in the figure below that we have already encountered in the Introduction of this textbook:



Occupational therapy for patients with problems in personal activities of daily living after stroke: systematic review of randomised trials

	Treatment		Control			Risk Difference	
Study or Subgroup	Events	Total	Events	Total	Weight	IV, Fixed, 95% CI	Year
Drummond 1995	2	42	3	23	12.8%	-0.08 [-0.23, 0.07]	1995
Corr 1995	33	55	32	54	8.7%	0.01 [-0.18, 0.19]	1995
Logan 1997	6	53	14	58	15.3%	-0.13 [-0.27, 0.01]	1997
Walker 1999	18	90	27	86	18.0%	-0.11 [-0.24, 0.01]	1999
Gilbertson 2000	33	66	41	67	10.5%	-0.11 [-0.28, 0.06]	2000
Parker 2001	106	248	56	123	25.7%	-0.03 [-0.14, 0.08]	2001
Sackley 2006	27	53	36	47	9.0%	-0.26 [-0.44, -0.08]	2006
Total (95% CI)		**607**		**458**	**100.0%**	**-0.09 [-0.15, -0.04]**	
Total events	225		209				

Heterogeneity: Chi² = 6.10, df = 6 (P = 0.41); I² = 2%
Test for overall effect: Z = 3.32 (P = 0.0009)

Figure 5.182: Special example of a meta-analysis in the field of clinical trials with concepts that we will detail without exception in this entire book (source: [344])

The risk difference (R.D.) being simply the difference of two proportions, we use the famous following Wald interval (see page 880) for the calculation of the confidence interval:

$$(\hat{p}_X-\hat{p}_Y)-Z_{\alpha/2}\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_1}+\frac{\hat{p}_Y(1-\hat{p}_Y)}{n_2}}\leq p_X-p_Y\leq(\hat{p}_X-\hat{p}_Y)+Z_{\alpha/2}\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_1}+\frac{\hat{p}_Y(1-\hat{p}_Y)}{n_2}} \tag{5.7.1771}$$

where by definition R.D. $:=p_X-p_Y$.

We see above that the width of the 95% confidence interval of the common risk difference is smaller than the width of the 95% confidence interval of the estimated effect in each of the trials. That's obviously because the pooled effect is more precise than the effect estimated in the trials.

It is also very important to notice that in a meta-analysis (like the example above), even if no trial detects an effect of the experimental intervention (as you can see above in every trial zero is in the confidence interval), the synthesis of the results of the trials (Risk Difference as above) can make it possible to detect an overall effect (the reader can see that in the total confidence interval, the zero is outside the confidence interval). We gain in power of the test!

The reader may ask where does the -0.09 comes from in the "Total (95% CI)" line? It comes from the simple following weighted average of all the risk differences:

$$\widehat{\text{R.D.}}_{\text{weighted}} = 0.128 \cdot (-0.08) + 0.087 \cdot (0.01) + 0.153 \cdot (-0.13) + \ldots = -0.09$$

5.7.13.2.1 Simpson's Paradox (sophism)

The "Simpson's paradox" (in fact it's a sophism and not a paradox because it is just a special case of omitted variable bias), also named "Yule-Simpson effect", is a paradox in probability and statistics, in which a trend appears in different groups of data but disappears or reverses when these groups are combined[56].

This result is often encountered especially in social-science (Human Resources, Marketing, Psychology), and medical-science statistics and is particularly confounding when frequency data are unduly given causal interpretations. Simpson's paradox disappears when causal relations are brought into consideration.

Let us see a famous real example for a real medical study (there is also another famous case with the *Nation at Risk* report, if the reader is interested, just search for it on Internet!):

The table below shows the success rates and numbers of treatments for treatments involving both small and large kidney stones, where Treatment A includes all open surgical procedures and Treatment B is percutaneous nephrolithotomy (which involves only a small puncture). The numbers in parentheses indicate the number of success cases over the total size of the group (for example, 93% equals 81 divided by 87):

	Treatment A	**Treatment B**
Small stones	Group 1 93% $(81/87)$	Group 2 87% $(234/270)$
Large stones	Group 3 73% $(192/263)$	Group 4 69% $(55/88)$
Both	78% $(273/350)$	83% $(289/350)$

The paradoxical conclusion is that treatment A is more effective when used on small stones, and also when used on large stones, yet treatment B is more effective when considering both sizes at the same time. In this example the "lurking" variable (or "confounding variable") of the stone size was not previously known to be important until its effects were included.

Formally, if we denote (we change the notations!!!!) A the result, B the treatment, and C the stones, in the general case, with the Simpson paradox it is possible to have using Bayesian probabilities:

$$P(A|B,C) < P(A|B^c,C) \text{ and } P(A|B,C^c) < P(A|B^cC^c) \text{ yet also: } P(A|B) > P(A|B^c) \tag{5.7.1772}$$

---

[56] Edward H. Simpson first described this phenomenon in a technical paper in 11951 (holocene calendar), but the statisticians Karl Pearson et al., in 11899 (holocene calendar), and Udny Yule, in 11903 (holocene calendar), had mentioned similar effects earlier. The name Simpson's paradox was introduced by Colin R. Blyth in 11972 (holocene calendar).

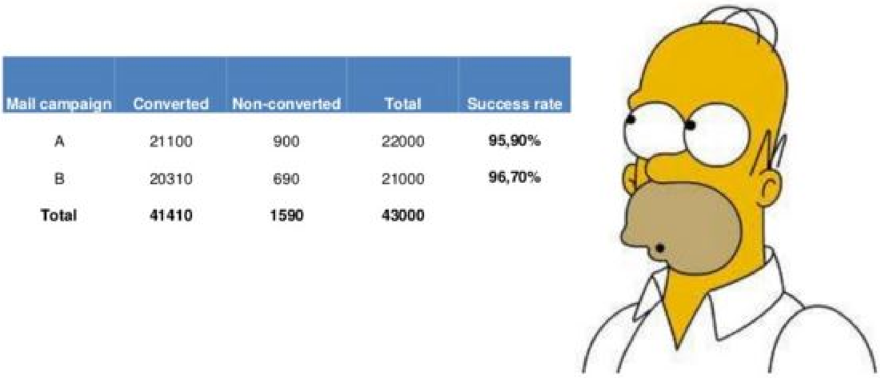
Which treatment is considered better is determined by an inequality between two ratios (successes/total). The reversal of the inequality between the ratios, which creates Simpson's paradox, happens because two effects occur together:

1. The sizes of the groups, which are combined when the lurking variable is ignored, are very different. Doctors tend to give the severe cases (large stones) the better treatment A, and the milder cases (small stones) the inferior treatment B. Therefore, the totals are dominated by groups 3 and 2, and not by the two much smaller groups 1 and 4.
2. The lurking variable has a large effect on the ratios, i.e. the success rate is more strongly influenced by the severity of the case than by the choice of treatment. Therefore, the group of patients with large stones using treatment A (group 3) does worse than the group with small stones, even if the latter used the inferior treatment B (group 2).

Based on these effects, the paradoxical result is seen to arise by suppression of the causal effect of stone size on successful treatment. The paradoxical result can be rephrased more accurately as follows: When the less effective treatment (B) is applied more frequently to easier cases, it can appear to be a more effective treatment.

Simpson's paradox usually fools us on tests of performance. In a famous example, researchers concluded that a newer treatment for kidney stones was more effective than traditional surgery, but it was later revealed that the newer treatment was more often being used on small kidney stones. More recently, on elementary school tests, minority students in Texas outperform their peers in Wisconsin, but Texas has so many minority students that Wisconsin beats it in state rankings. It would be a shame if Simpson's paradox led doctors to prescribe ineffective treatments or Texas schools to waste money copying Wisconsin.

Consider another funny illustrated example in Marketing! Consider we have the following information:



Mail campaign	Converted	Non-converted	Total	Success rate
A	21100	900	22000	95,90%
B	20310	690	21000	96,70%
Total	41410	1590	43000	

Figure 5.183: **Marketing Campaign performance**

Obviously the campaign B perform the best into converting an e-mail to click ("conversion rate")!

But now splitting in clicking high valuable links (costly products: high ticket conversion) or low valuable links (low ticket conversion) we get:

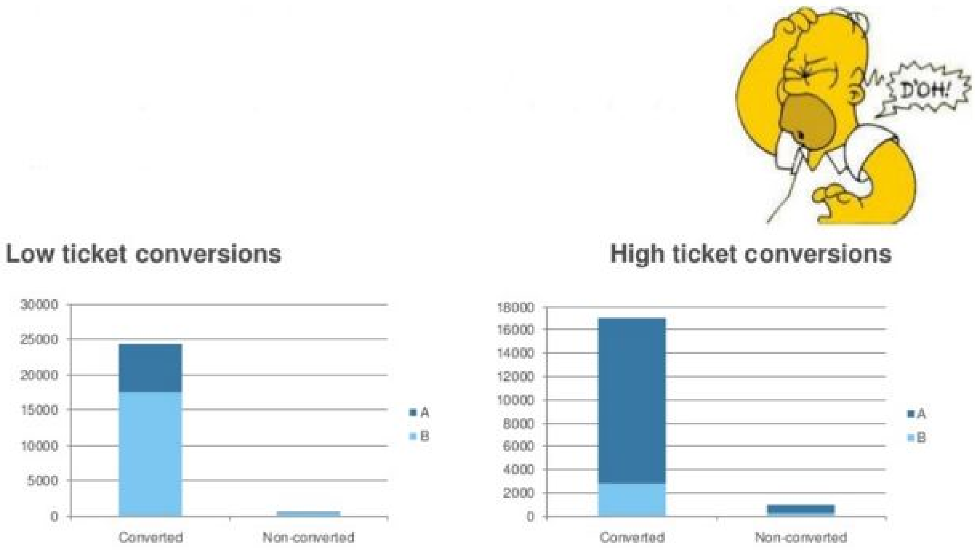


Figure 5.184: **Marketing Campaign performance in ticket conversion covariate**

Or for people that are more visual (with plots) here is an example from our R companion book. This is a simple regression of the IRIS dataset. The linear regression trend seem negative:

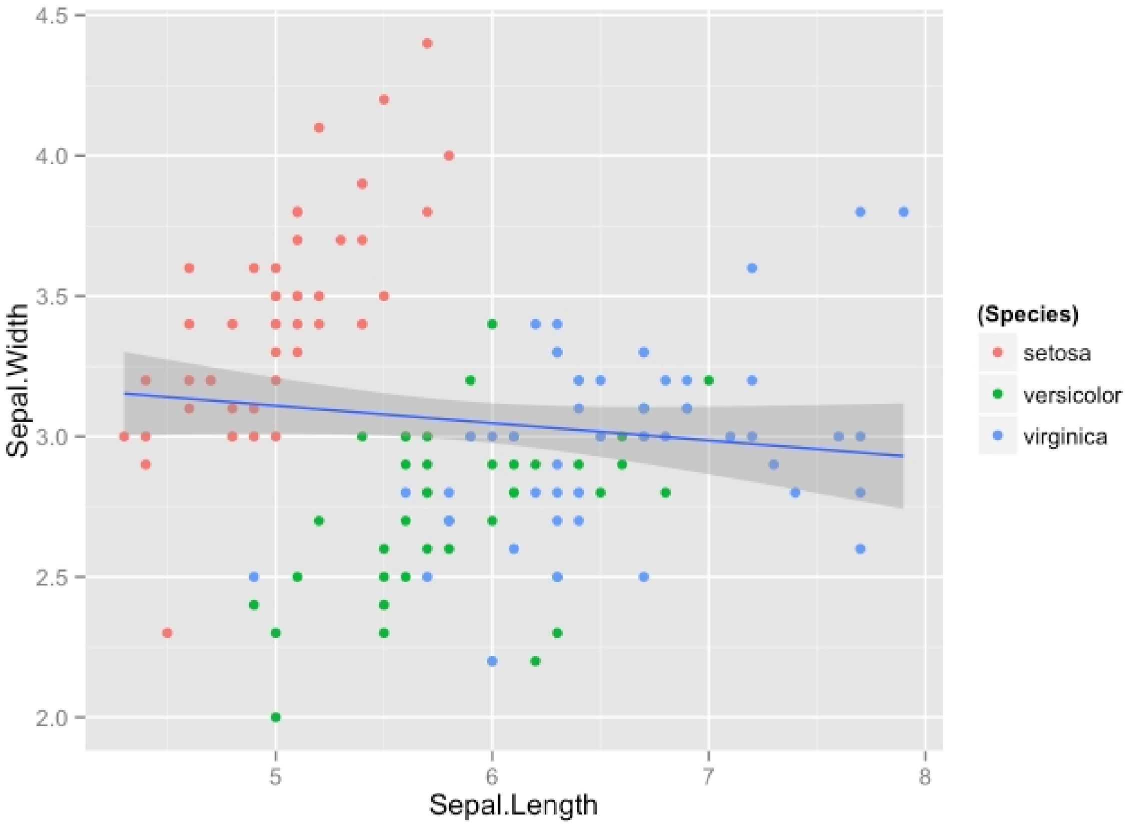


Figure 5.185: **IRIS dataset non-clustered regression in R to illustrate Simpson paradox**

Now if we cluster the regression.... all the trends are positive...:

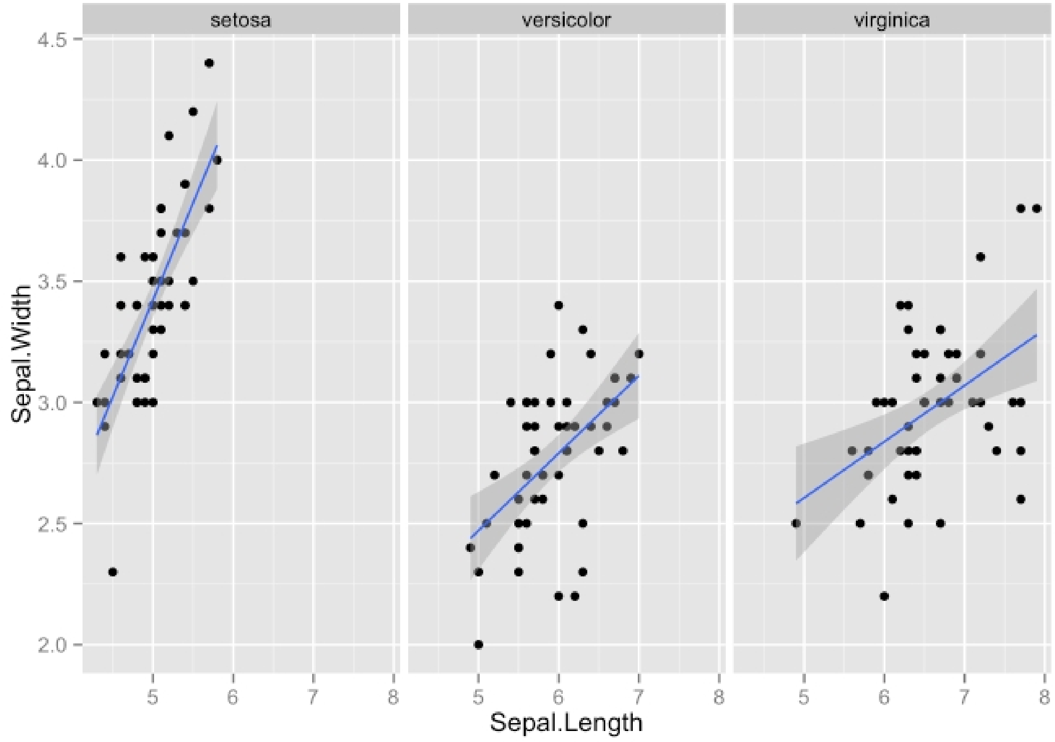
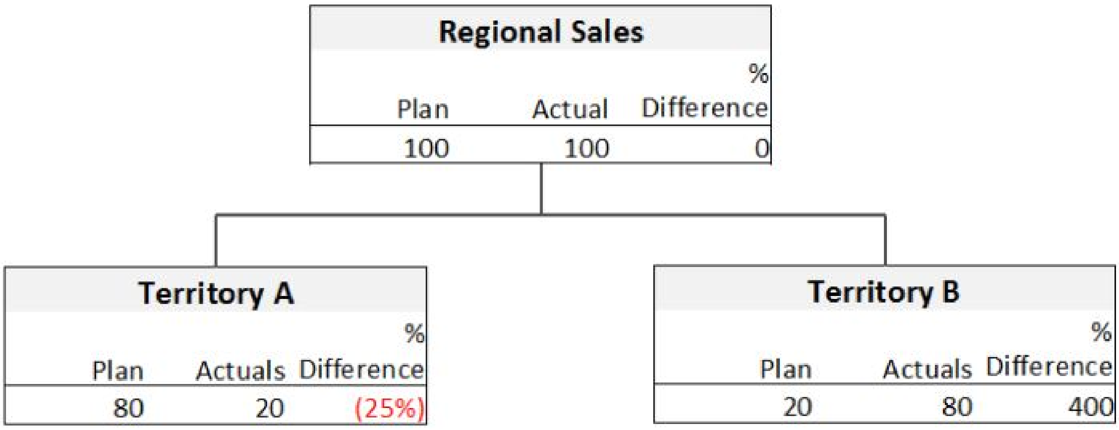


Figure 5.186: **IRIS dataset clustered regression in R to illustrate Simpson paradox**

Some people may argue that this is not business oriented... So here is an example of Simpson's paradox typical in business intelligence reports or dashboards:



Regional Sales

Plan	Actual	% Difference
100	100	0

Territory A

Plan	Actuals	% Difference
80	20	(25%)

Territory B

Plan	Actuals	% Difference
20	80	400

We see above that aggregated regional Sales look like they are on plan. But at the next level of detail (at the Territory level), we can see that Territory A is doing very poorly (25% of plan) and Territory B is doing exceptionally well (400% of plan). This is a major challenge in the world of BI. It's easy to feel content with what the BI reports and dashboards are telling us without having an easy, automated way to delve into the minutiae to find those areas of the business that are significantly over- or under-performing.

Definition (#5.7.147): In the academic literature on information systems and business communication, a "report" is generally defined as a structured layout that systematically presents information, data, analyses, findings, and conclusions regarding a specific subject for the purpose of informing, documenting, evaluating, or supporting decision-making for a specific target audience eventually in an interactive way.

Definition (#5.7.148): In the academic literature on information systems and business communication, a "dashboard" is generally defined as a visual information system that in almost real-time consolidates, organizes, and presents the most important performance indicators, metrics, graphs or data from various reports needed to achieve one or more objectives, displayed on a single screen so that information can be monitored and understood at a glance for a specific target audience in a non-interactive way.

Remark

Another kind of quite famous similar case in statistics is named the "Lord's paradox" (that is also only a sophism in reality!). It just highlights the obvious fact that under some conditions an ANOVA (ANalysis Of VARiance) won't give the same result as an ANCOVA (ANalysis of CoVAriance). These are two different kind of tools that we will see further below.

5.7.13.3 Power of a Test

When the effect is actually large, we can expect that we need less observations to demonstrate it that when the effect is small ... but how much exactly? Did we have the possibility to do this, in terms of number of measures, to "prove" what we seek? Should we go about it differently and change the device to do the observation/experiment?

We have seen earlier that we can never "prove" the null hypothesis[57]; all we can do is reject or fail to reject it. However, there are times when it is necessary to try to "prove" (i.e. get evidence) the non-existence of a difference between groups. This most often happens within the context of comparing a new treatment against an established one and showing that the new intervention is not inferior to the standard (in some circumstances, we need many more subjects to get evidence of non-inferiority than we need to show a difference). Can we do something like this?

To answer all questions above, we need now to study in more detail the concept of "power of a test" that we have so far only mentioned, remember the following figure that we have already seen just a little bit earlier:

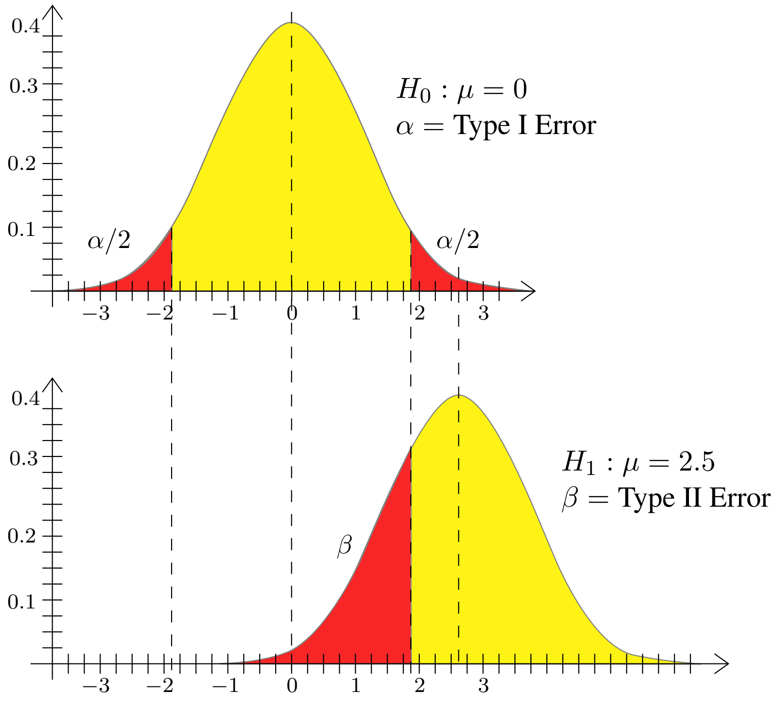


Figure 5.187: Null and alternative hypothesis of a special case of two-sided test

In the particular example above, we will obviously reject the null hypothesis H_0 if $\bar{X} > 1.96$ or $\bar{X} < -1.96$. Imagine that under the alternative hypothesis H_1, if we measured 2.5 for $\bar{X}$, we will have for power of the test:

$$1 - \beta = 1 - P\left(Y_{\mathcal{N}(2.5,1)} \leq 1.96\right) = 1 - P\left(Y_{\mathcal{N}(\bar{X},1)} \leq Z_{1-\alpha/2}\right) = 1 - 0.2946 \cong 0.705 \tag{5.7.1773}$$

---

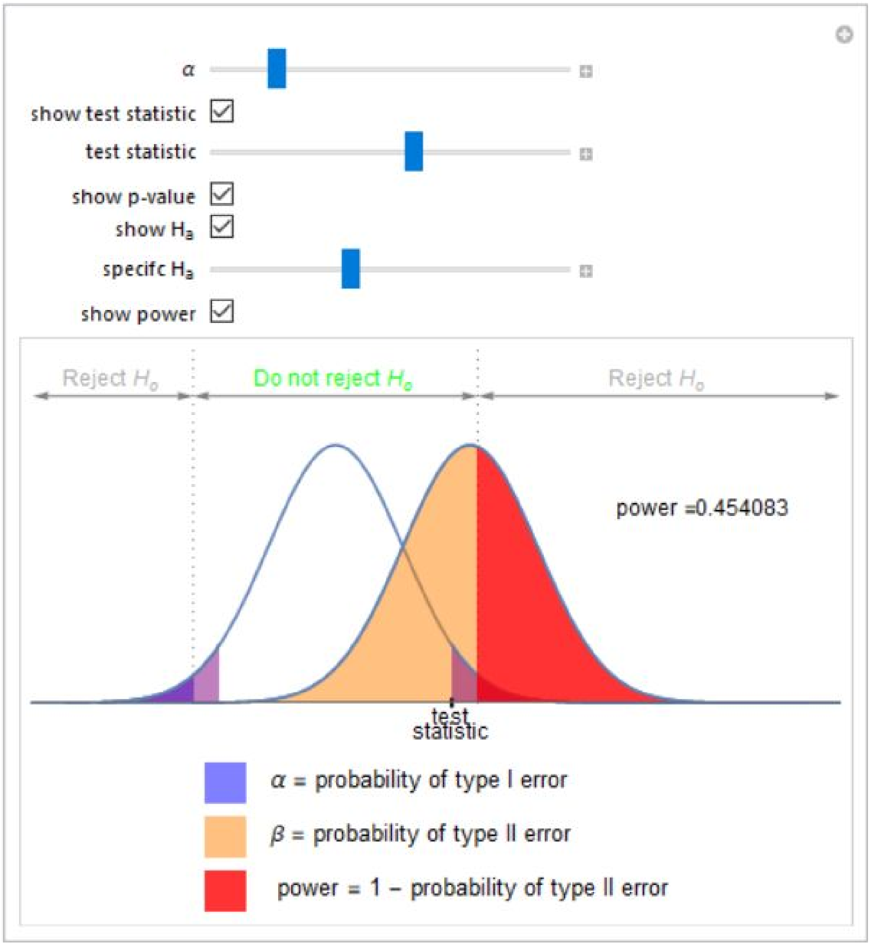
[57] Remember that with NHST, if the p-value is greater than a given threshold, we do should not say that we can accept (or "prove") the null hypothesis; rather, we use the convoluted locution that « *We have failed to disprove the null*».

So the test is relatively strong/powerful (in practice, we consider a test to be powerful if its value is above of 80% - that is to say the $\beta \leq 20\%$). Thus, we see that the (a posteriori!) power $1-\beta$ is even larger than the p-value is small (and respectively the power will be a posteriori even smaller than the p-value is great). Therefore the a posteriori power is in decreasing correspondence with the p-value (in practice it is however a bit absurd to make these calculations a posteriori but some scientific paper require this information).

So in general:

$$\boxed{\text{power} = 1 - \beta = 1 - P\,(\text{not rejecting a false } H_0)} \qquad (5.7.1774)$$

A great way to play with these concepts and visualize them is to take a look at the Wolfram widget available on http://demonstrations.wolfram.com/StatisticalPower/:



α

show test statistic ☑

test statistic

show p-value ☑

show H$_a$ ☑

specifc H$_a$

show power ☑

Reject H_0
Do not reject H_0
Reject H_0
power =0.454083
test statistic
🟦 α = probability of type I error
🟧 β = probability of type II error
🟥 power = 1 – probability of type II error

Before continuing, let us inform you that many don't understand why the probability of getting a type I error when performing a hypothesis test, isn't affected by the sample size?!

If we were using standard hypothesis testing, we may set the confidence level α and the compare the test p-value to it. In this case the sample size will not impact the probability of Type I error because your confidence level α is the probability of Type I error, pretty much by definition. In other words, you set the probability of Type I error by choosing the confidence level! The probability of Type I error is only impacted by your choice of the confidence level and nothing else!

But yes, we could also fix Type II error rate (but this is not the purpose of science as we know to "prove" the null hypothesis as it is practically quite hard most of times!), the Type I error rate usually becomes the unknown parameter and it is dependent on the sample size, the variance and the effect size (delta).

Remark

Many people say that if a Person A posits the existence of some phenomenon, it is incumbent on Person A to prove its existence rather than it being Person B's job to disprove it. For example, if we assert that there are pink unicorns with purple polka dots hiding deep in the forest, I cannot simply say, «*Well, prove that they don't exist*». For the scientific world to accept my claim, I have to bring one back, dead or alive. Is it ever legitimate to violate this injunction and try to prove the non-existence of something? In the text above we then have taken a look at trying to "prove" (i.e. get "enough evidence") that something does not exist[a].

---

[a] Obviously if we are sure that we have investigated ALL possible combinations, places and that the huge majority properties of the unicorn are proven wrong or impossible, then it's very highly likely that the unicorn itself doesn't exist

5.7.13.3.1 Evidence ("proof") of a negative

Georgia minister Nelson L. Price asserts on his website (and many other science popularizer on YouTube and scientifically illiterate social networks users) that « *one of the laws of logic is that you can't prove a negative* » (also named "argumentum ad ignorantium"). The fact is, however, that this supposed "law of logic" is no such thing!

Notice, for a start, that « *You cannot prove a negative* » is itself a negative. So, if it were true, it would itself be unprovable. Notice that any claim can be transformed into a negative by a little rephrasing - most obviously, by negating the claim and then negating it again. « *I exist* is logically equivalent to « *I do not not exist* », which is a negative. Yet here is a negative it seems we might perhaps be able to prove (in the style of Descartes: « *I think, therefore I do not not exist!* »).

Of course, those who say « *You cannot prove a negative* », or who say we do an « *argumentum ad ignorantium* », will insist that we have misunderstood their point. As Steven D. Hales notes, when people say, « *You can't prove a negative* », what they really mean is that you cannot prove that something does not exist. If this point were correct, it would apply not just to supernatural beings lying beyond the cosmic veil but also to things that might be supposed to exist on this side of the veil, such as unicorns, Martians, rabbits with 20 heads, and so on. We would not be able to prove the non-existence of any of these things either.

But is the point correct? Is it true that we can never prove that something does not exist? Again, it depends **IF** you have sufficiently defined the scope of what you are looking at and the level of evidence[58]. If someone claims there's a unicorn in the tool shed (see [519] for a deep treatment of this subject), we can quickly establish this person is mistaken by going and taking a look. We could similarly establish there's no Loch Ness monster by draining the loch. But what of the claim that unicorns once existed? We can't travel back in time and directly observe all of the past as we can every corner of the tool shed or Loch Ness. Does it follow that we can't prove unicorns never existed?

It depends in part, as we already know it, on what we mean by "prove". The word has a variety of meanings. By saying something is "prove", we might mean that it is established beyond all possible doubt. Or we might mean it has been established beyond reasonable doubt (this is the kind of proof in experimental science). Can we establish beyond reasonable doubt that unicorns have never inhabited the Earth? True, the history of our planet has been and gone, so we can no longer directly inspect it. But surely, if unicorns did roam the Earth, we would expect to find some evidence of their presence, such as fossils of unicorns or at least of closely related animals from which unicorns might plausibly have evolved. There is none. We also have plenty of evidence that unicorns are a fictional creation, in which case, it's surely reasonable for us to conclude that there never were any unicorns. Indeed, we would suggest we can give evidence of this beyond reasonable doubt (especially if all inherent properties of unicorns have already been broken down).

In response, it might be said, « *But you can't prove conclusively, beyond all possible doubt, that unicorns never roamed the Earth*[59] ». This is undeniably true! However, this point is not peculiar to negatives. It can be made about any claim about the unobserved, and thus any scientific theory at all, including scientific theories about what does exist. We can prove beyond reasonable doubt that dinosaurs existed, but not beyond all possible doubt!

Despite the mountain of evidence that dinosaurs roamed the Earth, it's still possible that, say, all those dinosaur fossils are fakes placed there by alien pranksters long ago.

Let's sum up. If « *you can't prove a negative* » means you can't prove beyond reasonable doubt that certain things don't exist, then the claim is just false. We prove the non-existence of things on a regular basis. If, on the other hand, "you can't prove a negative" means you cannot prove beyond all possible doubt that something does not exist, well, that may, arguably, be true. But so what? That point is irrelevant so far as defending beliefs in supernatural entities against the charge that science and/or reason have established beyond reasonable doubt that they don't exist.

Let us now consider something has a finite number of observable (i.e. refutable) properties/characteristics in a finite volume. So say we have a bowl. We can prove (i.e. get evidence) the non-existence of milk in our bowl with a photo (or more complex apparatus) of an empty bowl. But this photo does not depict negative milk. It depicts a positive bottom of the bowl, and from that positive, we can deduce the impossibility of milk being there, knowing, from other evidence, what milk looks like. So we are using positive evidence to prove impossibility.

---

[58] For example the huge majority of properties and acts that define some divinities have been showed with strong evidence as wrong. This lead us to the very likely conclusion that there are no divinities - especially divinities that have become mythologies give us a strong evidence for this.

[59] A "conclusive evidence" is evidence that cannot be contradicted by any other evidence. It is so strong as to overbear any other evidence to the contrary. The evidence is of such a nature that it compels a fact-finder to come to a certain conclusion.

Impossibility is the exhaustion of exclusive possibilities. Possibilities are zero or positive. In this case, milk and empty are mutually exclusive, where the likelihood of milk is 0, and the likeliness of empty is 1. This is different from proving a negative. Now as most religions have holy books that depicts quite precisely some properties and actions of their gods on Earth (finite volume), then if we found evidence in that closed volume (Earth) that disprove at least one of the properties (i.e. characteristic) of their god, then that latter very likely doesn't exist beyond any reasonable doubt (B.A.R.D.)... Technically this is denoted:

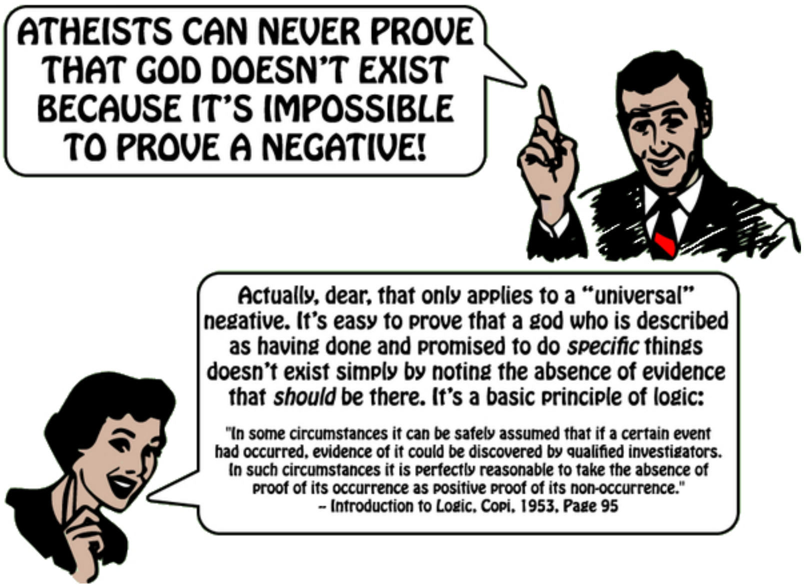
$$\begin{gathered}\forall\ H_{0,i} \in \mathcal{H}_0 = \{H_{0,1}, H_{0,2} \ldots, H_{0,n} | H_{0,j} \cap H_{0,i} = \varnothing \wedge \cap_{i=1}^{n} H_{0,i} \neq \varnothing\ \forall i, j\} \\ \text{if } \exists H_{0,i} = \texttt{FALSE B.A.R.D.} \Rightarrow \mathcal{H}_0 = \texttt{FALSE B.A.R.D.}\end{gathered} \tag{5.7.1775}$$

and named the "Stenger-Neyman-Pearson supernatural falsification method". So applying this relation to religious books (vedas, bible, torah, quran, etc.) knowing that the bible has (without counting the contradictions and historical errors) at least 21 scientific errors, the quran has (without counting the contradictions and historical errors) at least 73 scientific errors, the torah has (without counting the contradictions and historical errors) at least 20 scientific errors, the vedas have (without counting the contradictions and historical errors) at least 26 scientific errors bring us to quite an obvious conclusion about the perfection - and hence existence - of their supposed respective deities...

quote
5.7:976.#49

> *Science doesn't tell us what to think but sometimes it tells us what we can no longer believe.*
>
> Etienne Klein

The old adage «You cannot prove a negative» (the correct formulation should be «You cannot get evidence of a negative») is a typical evidence level 02 (see page 60) argument from scientifically illiterate people.


ATHEISTS CAN NEVER PROVE THAT GOD DOESN'T EXIST BECAUSE IT'S IMPOSSIBLE TO PROVE A NEGATIVE!
Actually, dear, that only applies to a "universal" negative. It's easy to prove that a god who is described as having done and promised to do *specific* things doesn't exist simply by noting the absence of evidence that *should* be there. It's a basic principle of logic:
"In some circumstances it can be safely assumed that if a certain event had occurred, evidence of it could be discovered by qualified investigators. In such circumstances it is perfectly reasonable to take the absence of proof of its occurrence as positive proof of its non-occurrence." -- Introduction to Logic, Copi, 1953, Page 95

So proving (i.e. get evidence) the non-existence depends how tightly defined the thing are checking for does not exist. If we can define some attribute the thing absolutely must have in a finite volume or time range, then we may be able to have strong evidence that the thing with that attribute does not exist. But as soon as there is any vagueness in the description, we cannot do so. So, again in order to prove (i.e. get evidence) something does not exist we must establish the properties of that thing in a finite volume to the point of being able to test its negation.

5.7.13.3.1.1 Two-One-Sided T-test procedure (TOST)

If we want to test the equality of two values $\hat{\theta}_1 = \hat{\theta}_2$ (mean, proportion or other), meaning the non-existence of an effect (or a "negative"), as alternative hypothesis H_A we can't do that with classical hypothesis and alternative test (as it is quite hard to build the statistics of the null hypothesis H_0, i.e. the statistical distribution of $\hat{\theta}_1 \neq \hat{\theta}_2$)!

A well known approach we then use in practice (considering the case with means!) is the "Two One-Sided Tests (TOST) procedure". This involves looking at the difference between the mean values for the two populations then testing a joint null hypothesis: that this difference is less than or equal to $+\delta$ AND that this difference is greater than or equal to $-\delta$ (for some sufficiently small and positive δ). If we can reject this joint null hypothesis, then you can conclude that the absolute value of the difference between the means is less than δ (see [300] for an analysis on how to choose δ).

In other words, the Two-One-Sided T-test procedure (TOST) is used to show that two samples are similar, in contrast to classical statistical tests which check for dissimilarity. The TOST relies on a parameter called delta, which has to be set by the researcher using their intuition. Doing so can be difficult, because of complex interactions of relevant parameters.

In some fields there may be industry standards to use as a guide for δ. If not, the equivalence acceptance criterion is best developed from a consensus among stakeholders such as decision makers, system engineers, operators, and subject matter experts (SMEs). Deciding what difference in SUT performance is "good enough" and determining δ for the equivalence test is not a trivial task and adequate time should be allotted for this in the planning phase. The value of δ should be based on how big a difference is practically important to consider the means not equivalent. Δ is similar to the difference to detect when we size a traditional designed experiment. Once δ is derived, the hypothesis can be stated as:

$$\begin{aligned} &H_0 : |\mu_1 - \mu_2| > \delta\text{: The difference in means is greater than the equivalence acceptance criterion}\\ &H_A : |\mu_1 - \mu_2| \leq \delta\text{: The difference in means is less than or equal to } \delta \text{ (the means are equivalent)} \end{aligned} \tag{5.7.1776}$$

A one-sided test similarity T-test (OST) is appropriate if the test team only cares if the response differs in one direction. For example, if a new rubber compound is used to make a tire with a mileage guarantee the producers may only care if the tire gets fewer miles of use. This is often named a "test for non-inferiority". In these situations, the null hypothesis is often termed the "inferiority" hypothesis and the alternate is the "non inferiority" hypothesis. The following equation is used to compare the t-ratio to the critical value.

As always we no longer use μ, because it represents the true population parameter, which we will likely never know. Instead we use $\hat{\mu}$, which denotes the sample mean. The critical value

$T_{\alpha,\text{df}}$ can be calculated with software or looked up in statistics tables.

We reject H_0 if:

$$\frac{|\hat{\mu}_1 - \hat{\mu}_2| + \delta}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}} > T_{\alpha,n_1+n_2-2} \tag{5.7.1777}$$

In similarity testing, there is no two-sided test. Instead, we conduct two non inferiority tests from both sides:

$$\begin{array}{ll} H_0 : |\mu_1 - \mu_2| > \delta \longrightarrow \begin{cases} H_0 : \mu_1 - \mu_2 > \delta & \text{Null hypothesis expanded for the upper side test} \\ H_0 : \mu_1 - \mu_2 < -\delta & \text{Null hypothesis expanded for the lower side test} \end{cases} \\ H_A : |\mu_1 - \mu_2| \leq \delta \longrightarrow H_A : -\delta \leq \mu_1 - \mu_2 \leq \delta \; \text{The alternate hypothesis is expanded} \end{array} \tag{5.7.1778}$$

When the hypotheses are expanded to create the test statistics we get the following criterion. We reject H_0 if:

$$\frac{|\hat{\mu}_1 - \hat{\mu}_2| + \delta}{\text{SE}} > T_{\alpha,n_1+n_2-2} \quad \text{AND} \quad \frac{|\hat{\mu}_1 - \hat{\mu}_2| - \delta}{\text{SE}} < -T_{\alpha,n_1+n_2-2} \tag{5.7.1779}$$

Note that the two means are considered equivalent if, and only if, both null hypotheses are rejected! We must show that difference in the means is less than $+\delta$ and greater than $-\delta$ to show that the means are equivalent; it cannot be true in just one direction. It is also prudent to point out that this is different than a traditional two-sided hypothesis test. When conducting a two-sided test with a Type I probability of α, we use $\alpha/2$ to compare to the p-value on each side. In this situation we use the full value of α on both tests because they are both one-sided.

✍**Example:**

Suppose we are analysing the effectiveness of the door gunner in a helicopter. It is normal for the door gunner to use the M-249 but a unit wants to instead arm them with the M-4. The M-249 is a machine gun capable of a higher volume of fire, but it is heavier, harder to move, and more prone to jam. Can the M-4 provide equivalent combat power? We have the following data:

Weapon Type	**Target Damage Score**
M-249	180
M-249	143
M-4	65
M-4	112
M-4	139
M-4	112
M-4	125
M-4	78
M-4	138
M-4	84
M-249	117
M-249	169
M-249	111
M-249	114
M-249	166
M-4	134
M-249	131
M-4	69
M-249	93
M-249	177

The reader can see the weapon used and the damage score for each of the 20 trials. This example is small enough that we can go through the calculations manually. Three basic statistics are calculated and shown in the table below:

Weapon	**Sample size**	**Sample mean**	**Sample variance**
M-249	$n_1 = 10$	$\hat{\mu}_1 = 140.1$	$\hat{\sigma}_1^2 = 981.21$
M-4	$n_2 = 10$	$\hat{\mu}_2 = 105.6$	$\hat{\sigma}_2^2 = 849.6$

Using the values calculated in the table above we can calculate the test statistic and critical value. First for the upper end of the curve:

$$\text{SE} = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} = \sqrt{\frac{981.21}{10} + \frac{849.6}{10}} = \sqrt{183.0811} \cong 13.53$$

We reject H_0 if:

$$\frac{|\hat{\mu}_1 - \hat{\mu}_2| + \delta}{\text{SE}} > T_{\alpha, n_1+n_2-2}$$

Hence:

$$\frac{|140.1 - 105.6| + 20}{13.53} > T_{0.05,18}$$

That is (right TOST):

$$4.03 > 1.734 \xrightarrow{\text{yields}} \text{True}$$

The statement is true so we reject H_0 in the first one-sided test. Next, we conduct the second test (left TOST):

$$\frac{|\hat{\mu}_1-\hat{\mu}_2|-\delta}{\text{SE}}<-T_{\alpha,n_1+n_2-2}$$

Hence:

$$\frac{|140.1-105.6|-20}{13.53}<-T_{0.05,18}$$

That is:

$$1.07<-1.734\xrightarrow{\text{yields}}\text{False}$$

The statement is false and we fail to reject H_0 in the second one-side test. Therefore, we also fail to reject the overall null hypothesis that the M-4 is not equivalent to the M-249. This is not a surprising outcome because our initial difference in means was 34.5.

The same reasoning can be applied for other tests. This is why in the `R` software you can find also a function to compute the TOST for proportions!

Remark

Depending on the statistical software packages, the degrees of freedom may be calculated using the Welch correction. Some software packages return only one p-value and only one confidence interval (most of the time for reasons we can't explain they return that of the left TOST only). Some even return a completely wrong calculation of the p-value. In other words it is better to run a TOST on more than three statistical packages to be almost sure to get a converging conclusion...

5.7.13.3.2 Power of the one sample Z-test and T-test

Very generally, in the case of a bilateral test, the above relation will be written:

$$1-\beta=1-P\left(Y_{\mathcal{N}(\bar{X},1)}\leq\mu+Z_{1-\alpha/2}\right)=1-P\left(Y_{\mathcal{N}(0,1)}\leq\mu-\bar{X}+Z_{1-\alpha/2}\right)\quad(5.7.1780)$$

If the standard deviation of the mean (e.g. variance) is not unitary, we have:

$$1-\beta=1-P\left(Y_{\mathcal{N}(0,\sigma_{\bar{X}})}\leq\mu-\bar{X}+Z_{1-\alpha/2}\sigma_{\bar{X}}\right)\quad(5.7.1781)$$

Therefore we get:

$$1-\beta=1-P\left(Y_{\mathcal{N}(0,1)}\leq\frac{\mu-\bar{X}}{\sigma_{\bar{X}}}+Z_{1-\alpha/2}\right)=1-P\left(Y_{\mathcal{N}(0,1)}\leq\frac{\mu-\bar{X}}{\sigma/\sqrt{n}}+Z_{1-\alpha/2}\right)\quad(5.7.1782)$$

written in another way:

$$\begin{aligned}1-\beta&=1-P\left(Y_{\mathcal{N}(0,1)}\leq\frac{\mu-\bar{X}}{\sigma}\sqrt{n}+Z_{1-\alpha/2}\right)=1-P\left(Y_{\mathcal{N}(0,1)}\leq d_\sigma\sqrt{n}+Z_{1-\alpha/2}\right)\\&=P\left(Y_{\mathcal{N}(0,1)}\leq-Z_{1-\alpha/2}-d_\sigma\sqrt{n}\right)=P\left(Y_{\mathcal{N}(0,1)}\leq Z_{\alpha/2}-d_\sigma\sqrt{n}\right)\end{aligned}\quad(5.7.1783)$$

It is in this form that we find the power of a bilateral test of the average (power of a 1 sample Z-test):

$$1-\beta=P\left(Y_{\mathcal{N}(0,1)}\leq Z_{\alpha/2}-d_\sigma\sqrt{n}\right) \tag{5.7.1784}$$

where d_σ is sometimes named the "size effect" and therefore defined by:

$$d_\sigma=\frac{\mu-\bar{X}}{\sigma}=\frac{\delta}{\sigma} \tag{5.7.1785}$$

and δ is named the "difference"!

It goes without saying that if the true variance is not known, we must be replace the Normal law by the Student law as:

$$1-\beta=P\left(Y_{\mathcal{N}(0,1)}\leq T_{\alpha/2}-d_S\sqrt{n}\right) \tag{5.7.1786}$$

with:

$$d_S=\frac{\mu-\bar{X}}{S}=\frac{\delta}{S} \tag{5.7.1787}$$

Caution with a small recurring trap! The development above corresponds to a δ which is negative with respect to the first example! The relation is a little bit different in the case where δ is positive but it does not matter because the power of the test is the same in absolute value!

For the sample size it quite simple. We have by equating the quantiles:

$$Z_{1-\beta}=Z_{\alpha/2}-d_\sigma\sqrt{n} \tag{5.7.1788}$$

And therefore in bilateral:

$$\boxed{n=\left(\frac{Z_{\alpha/2}-Z_{1-\beta}}{d_\sigma}\right)^2} \tag{5.7.1789}$$

where we see that if the power of the test is imposed as being equal to 50%, with having therefore $Z_{1-\beta}$ that is 0 then we fall back (!) on the relation of the sample size for the Normal distribution proved earlier above:

$$n=\left(\frac{Z_{\alpha/2}}{d_\sigma\sqrt{n}}\right)^2 \tag{5.7.1790}$$

Note also that we sometimes find in the literature the prior-previous relation (the sample size n such that power of the test is at least $1-\beta$) as follows:

$$n=\frac{\sigma^2\left(Z_{1-\alpha/2}+Z_{1-\beta}\right)^2}{(\mu_0-\mu_A)^2} \tag{5.7.1791}$$

Obviously we can set other parameters to determine the value of the remaining variable. We could also, for example, look for the value of the power of the test by fixing the standard deviation, the sample size and the risk threshold, etc.

One reader offered us a very elegant way of finding the same result with much less developments... Indeed, you need only to see in the previous figure that we have:

$$Z_{\beta}\frac{\sigma}{\sqrt{n}}+\bar{X}=Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}+\mu \tag{5.7.1792}$$

equality from which we draw immediately an equivalent relation to the two previous one (which obviously gives the same numerical result):

$$n=\left(\frac{\sigma\left(Z_{\beta}-Z_{1-\alpha/2}\right)}{(\mu-\bar{X})}\right)^2 \tag{5.7.1793}$$

In practice, the sample standard deviation will usually be used in the analysis (thus a T-test or interval will be calculated instead of a Z-test or interval). In that case, there is greater uncertainty in the estimate, which must be incorporated into the sample size estimate. For the case of testing the mean of a normal population, the relation will be:

$$n=\left(\frac{\hat{\sigma}\left(T_{\beta,n-1}-T_{1-\alpha/2,n-1}\right)}{(\mu-\bar{X})}\right)^2 \tag{5.7.1794}$$

Remark

The attentive reader will have perhaps noticed that we assumed in the previous developments that the standard deviation of the mean of the null hypothesis and that of the alternative one is implicitly assumed to be the same... In practice this is almost all the time so, this is why almost all statistical software require only one standard deviation to calculate the power of the 1-Sample Z Test. However, for some rare academic softwares, they ask the standard deviation of the two means. But then the above mathematical developments are obviously different.

A test power analysis may have several facets:

1. We know the level of the test, the sample size and effect size (implicitly the difference) and we seek to calculate its power. This allows us to see if our experimental device is properly calibrated.
2. We know the desired test power, the risk threshold level and effect size to detect. We seek then to calculate the sample size needed to mount an effective experimental design.
3. We know the desired test power, the risk threshold level and sample size and we seek to know the effect size we can hope to highlight.

Without exception, we consider as unnecessary to show a test if the expected power is less than 80%. This power corresponds to a 80% probability of not rejecting the null hypothesis wrongly or, what remains to the same: a 20% of Type II error.

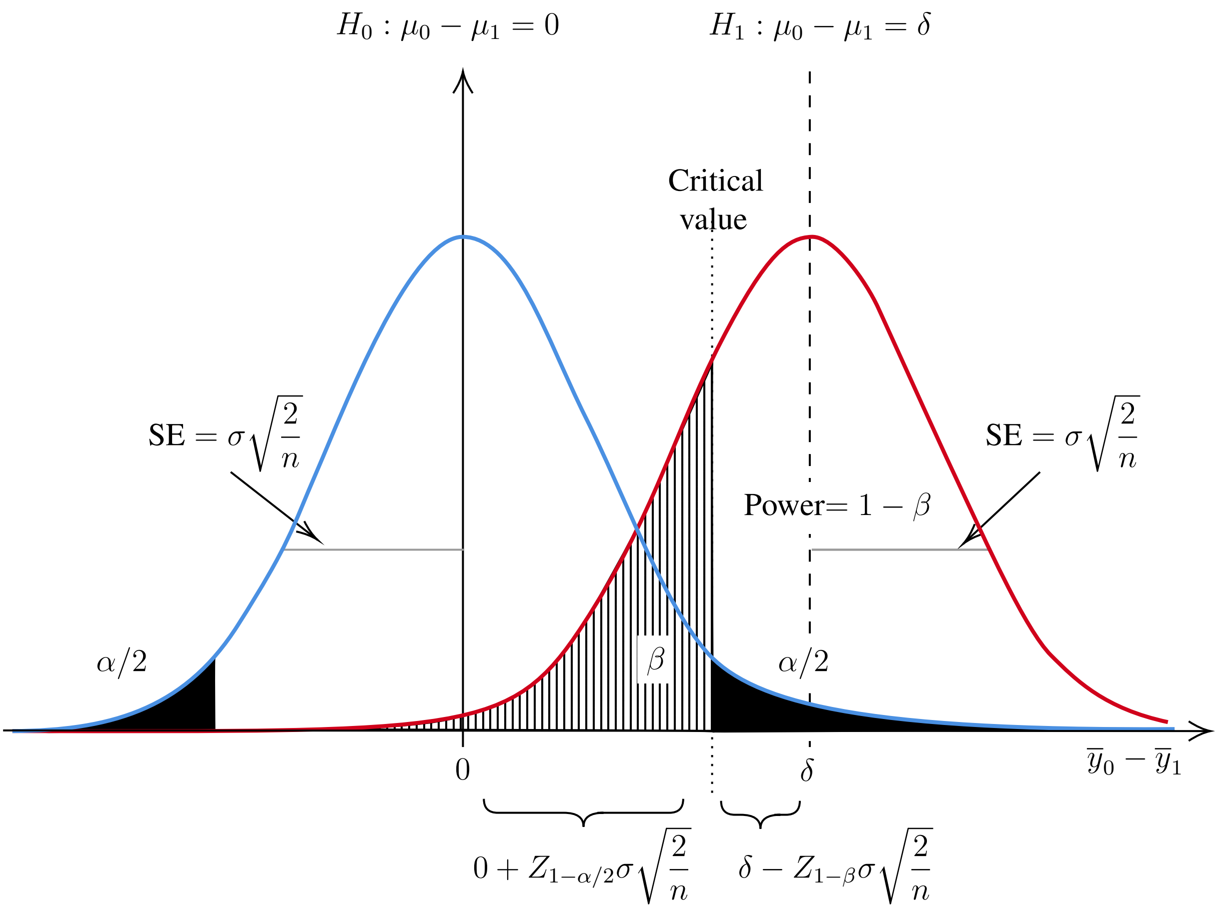
Obviously, it is possible to make the same reasoning (analytically when possible, otherwise numerically) with absolutely ALL the hypothesis tests that we have seen until now. So as there is a little more than a hundred hypothesis testing in the field of statistics as we have already mentioned... it is obvious that we're not going to have.... fun ... to make the same developments to determine the sample size, effect size and power for all these tests but only for the classics one. As long as we have computers at our disposal with the algorithms integrated by computer scientists, we do not need to redo any developments that would not bring much. Moreover, the majority of statistical softwares can calculate usually the power of only 5 to 10 common tests.

Remark

We will not deal with parametric statistical tests for the detection of outliers in this textbook as the Q-Dixon's test or Grubb's test and this just because they have too empirical origin and that they have no interest analytically speaking. By cons, if some readers insist, and that we have the time, we can put the details on these tests with detailed algorithms for calculating critical values using a simple spreadsheet and Monte Carlo technique with any distribution of their choice (but not only for data following a Normal law contrary to what is written in most books).

5.7.13.3.3 Power of the two sample Z-test and T-test

This figure below (taken from van Belle) illustrates some important basic concepts that will follow:


$H_0 : \mu_0 - \mu_1 = 0$
$H_1 : \mu_0 - \mu_1 = \delta$
Critical value
$\text{SE} = \sigma\sqrt{\frac{2}{n}}$
$\text{SE} = \sigma\sqrt{\frac{2}{n}}$
Power$= 1 - \beta$
$\alpha/2$
β
$\alpha/2$
0
δ
$\overline{y}_0 - \overline{y}_1$
$0 + Z_{1-\alpha/2}\sigma\sqrt{\frac{2}{n}}$
$\delta - Z_{1-\beta}\sigma\sqrt{\frac{2}{n}}$

Each curve represents the difference between two groups on some continuous measurement ($\Delta = y_0 - y_1$). The null hypothesis, that there is no difference between the two groups, is represented by the curve on the left ($\mu_\Delta = 0$). The alternative hypothesis, that there is some difference between the two groups, is represented by the curve on the right ($\mu_\Delta = \delta$). Each distribution is also characterized by its variance (σ^2) which is usually assumed to be the same (or similar enough...) for both distributions. The relationship between the standard error of the differences in means and the standard deviation[60]:

$$\text{SE}_\Delta = \sigma\sqrt{\frac{2}{n}} \tag{5.7.1795}$$

allows us to set up calculations for the sample size, n. The term $\sigma\sqrt{\frac{2}{n}}$ derives from the standard error for the difference between two means with equal variances.

The reader can also see the (conditional) probabilities about test results and true underlying differences that make up the classical components of sample size calculations:

- type I error or $\alpha = P(+$ result | - difference), i.e.. false positive
  - the traditional black-shaded $0.05/2$ tails on the null hypothesis
- type II error or $\beta = P($ result | + difference), i.e. false negative
  - the traditional hatched 0.2 left-handed tail on the alternative hypothesis
- power or $1 - \beta = P(+$ result $|$ + difference), i.e. true positive
  - the area to the right of β on the alternative hypothesis

Finally, you can also see that you can increase the power of a study without increasing the sample size by:

1. Measuring your outcome more precisely, which decreases the variance or "shrinks" the two curves towards their respective means and increases the area under the curve to the right of the critical value, or
2. Studying a greater effect size, in this case the difference between $\mu_\Delta = 0$ and $\mu_\Delta = \delta$, which separates the curves from each other and similarly increases the area under the curve to the right of the critical value

Returning to the figure above, the critical value is the point where the upper value of α on the null curve and the value for β on the alternative curve meet. This defines the point at which you accept or reject your null hypothesis, and sets up the sample size equation as:

$$0 + Z_{1-\alpha/2}\sigma\sqrt{\frac{2}{n}} = \delta - Z_{1-\beta}\sigma\sqrt{\frac{2}{n}} \tag{5.7.1796}$$

---

[60] This is based on: $\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \cong \sigma\sqrt{\frac{2}{n}}$

Solving for n allows us to calculate sample sizes "sample size for difference of homoscedastic means":

$$\boxed{n=\frac{2\left(Z_{1-\alpha/2}+Z_{1-\beta}\right)^2}{\left(\frac{\mu_1-\mu_2}{\sigma}\right)^2}=\frac{2\left(Z_{1-\alpha/2}+Z_{1-\beta}\right)^2}{\left(\frac{\delta}{\sigma}\right)^2}}\tag{5.7.1797}$$

Substituting the usual values of 1.96 for α and 0.84 for β, and defining the denominator as the square of the standardized difference (Δ^2), yields the rule of thumb called "Lehr's equation" or the "Rule of 16" :

$$n=\frac{2(1.96+0.84)^2}{\Delta^2}=\frac{15.68}{\left(\frac{\delta}{\sigma}\right)^2}\cong\frac{16\sigma^2}{\delta^2}\tag{5.7.1798}$$

✍**Example:**

Say, for example, you want to demonstrate a 10 point difference in IQ between two groups, one of which is exposed to a potential toxin, the other of which is not. Using a mean population IQ of 100, and a standard deviation of 20 :

$$n=\frac{16}{(100-90/20)^2}=\frac{16}{(0.5)^2}=64$$

or a total sample size of 128.

We can re-arrange Lehr's equation to solve for the detectable standardized difference between two populations:

$$n=\frac{16}{(\Delta)^2}\Leftrightarrow\Delta=\frac{4}{\sqrt{n}}$$

For the unstandardised difference:

$$\delta=\frac{4\sigma}{\sqrt{n}}$$

So, for example, we were interested in knowing the difference in mean IQ we could detect with a sample size of 50. Applying the above relation we get about 12 IQ points.

In practice, the sample standard deviation will usually be used in the analysis (thus a T-test or interval will be calculated instead of a Z-test or interval). In that case, there is greater uncertainty in the estimate, which must be incorporated into the sample size estimate. For the case of testing the mean of a normal population, the formula is now:

$$\boxed{n=\frac{2\left(T_{1-\alpha/2,n-1}+Z_{1-\beta,n-1}\right)^2}{\left(\frac{\mu_1-\mu_2}{\hat{\sigma}}\right)^2}=\frac{2\left(T_{1-\alpha/2,n-1}+T_{1-\beta,n-1}\right)^2}{\left(\frac{\delta}{\hat{\sigma}}\right)^2}}\tag{5.7.1799}$$

5.7.13.3.4 Power of the one and two samples P-test

Same as with the confidence interval of the Normal distribution with known theoretical standard deviation (that is to say the entire population), we can determine the number of individuals (sample size) for a fixed test power for the 1 sample proportion test studied in detail earlier. For this, we use the same technique as for the power of the one sample Z-test. We write so at first:

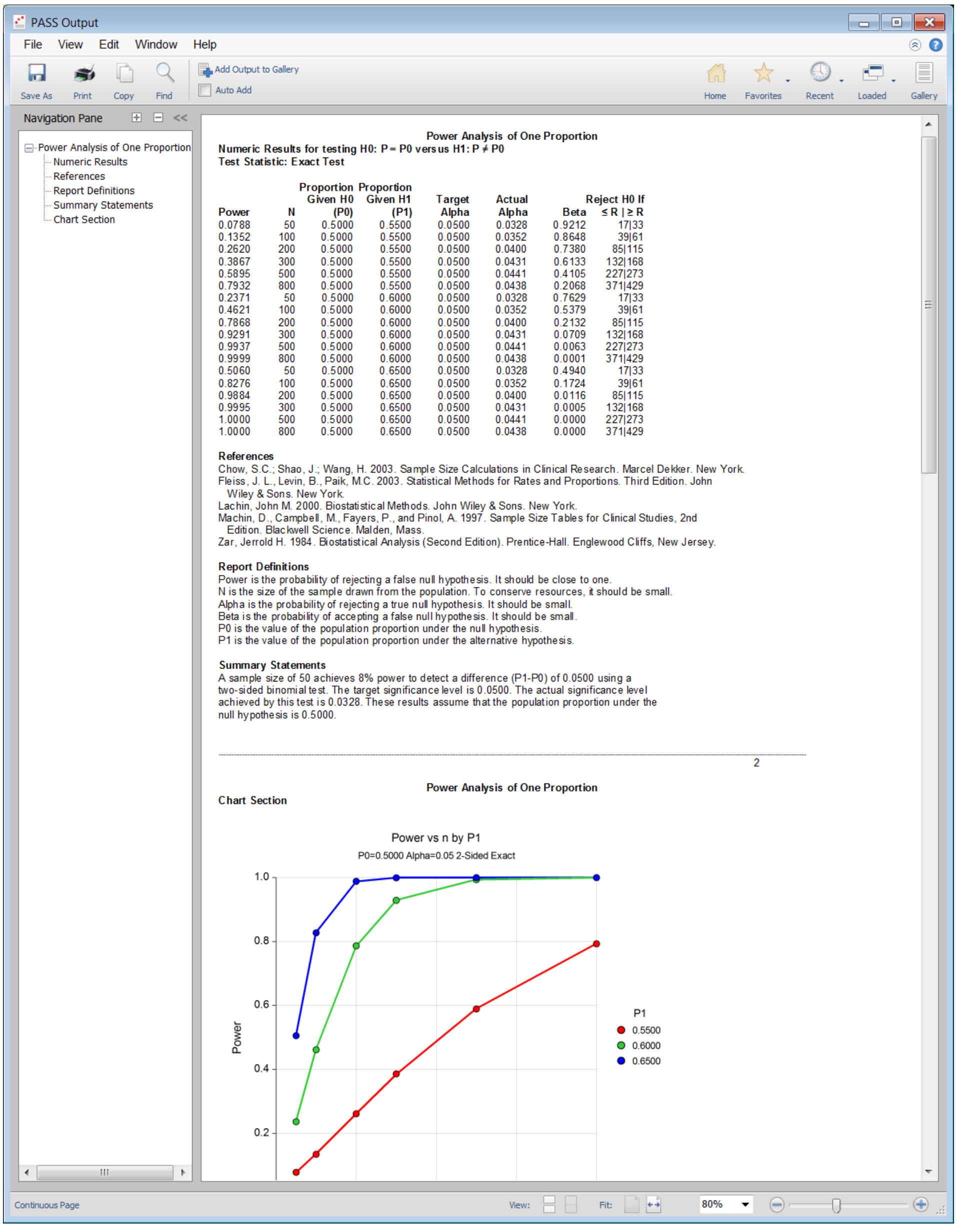
$$Z_\beta\sqrt{\frac{\hat{p}\hat{q}}{n}}+\hat{p}=Z_{1-\alpha/2}\sqrt{\frac{pq}{n}}+p \tag{5.7.1800}$$

Hence we deduce:

$$\boxed{n=\left(\frac{Z_\beta\sqrt{\hat{p}\hat{q}}-Z_{1-\alpha/2}\sqrt{pq}}{p-\hat{p}}\right)^2} \tag{5.7.1801}$$

So if the power is 50% we fall back on:

$$n=\frac{Z_{1-\alpha/2}^2pq}{\delta^2} \tag{5.7.1802}$$


PASS Output
File View Edit Window Help
Save As Print Copy Find
Add Output to Gallery
☐ Auto Add
Home Favorites Recent Loaded Gallery
Navigation Pane
Power Analysis of One Proportion
Numeric Results
References
Report Definitions
Summary Statements
Chart Section
Continuous Page
View: Fit: 80%
Power vs n by P1
P0=0.5000 Alpha=0.05 2-Sided Exact
1.0
0.8
0.6
0.4
0.2
Power
P1
● 0.5500
● 0.6000
● 0.6500

Power Analysis of One Proportion

```
Numeric Results for testing H0: P = P0 versus H1: P ≠ P0
Test Statistic: Exact Test

                 Proportion Proportion
                   Given H0   Given H1     Target     Actual              Reject H0 If
Power        N        (P0)       (P1)      Alpha      Alpha        Beta    ≤ R | ≥ R
0.0788      50      0.5000     0.5500     0.0500     0.0328      0.9212        17|33
0.1352     100      0.5000     0.5500     0.0500     0.0352      0.8648        39|61
0.2620     200      0.5000     0.5500     0.0500     0.0400      0.7380       85|115
0.3867     300      0.5000     0.5500     0.0500     0.0431      0.6133      132|168
0.5895     500      0.5000     0.5500     0.0500     0.0441      0.4105      227|273
0.7932     800      0.5000     0.5500     0.0500     0.0438      0.2068      371|429
0.2371      50      0.5000     0.6000     0.0500     0.0328      0.7629        17|33
0.4621     100      0.5000     0.6000     0.0500     0.0352      0.5379        39|61
0.7868     200      0.5000     0.6000     0.0500     0.0400      0.2132       85|115
0.9291     300      0.5000     0.6000     0.0500     0.0431      0.0709      132|168
0.9937     500      0.5000     0.6000     0.0500     0.0441      0.0063      227|273
0.9999     800      0.5000     0.6000     0.0500     0.0438      0.0001      371|429
0.5060      50      0.5000     0.6500     0.0500     0.0328      0.4940        17|33
0.8276     100      0.5000     0.6500     0.0500     0.0352      0.1724        39|61
0.9884     200      0.5000     0.6500     0.0500     0.0400      0.0116       85|115
0.9995     300      0.5000     0.6500     0.0500     0.0431      0.0005      132|168
1.0000     500      0.5000     0.6500     0.0500     0.0441      0.0000      227|273
1.0000     800      0.5000     0.6500     0.0500     0.0438      0.0000      371|429
```

References

Chow, S.C.; Shao, J.; Wang, H. 2003. Sample Size Calculations in Clinical Research. Marcel Dekker. New York.
Fleiss, J. L., Levin, B., Paik, M.C. 2003. Statistical Methods for Rates and Proportions. Third Edition. John Wiley & Sons. New York.
Lachin, John M. 2000. Biostatistical Methods. John Wiley & Sons. New York.
Machin, D., Campbell, M., Fayers, P., and Pinol, A. 1997. Sample Size Tables for Clinical Studies, 2nd Edition. Blackwell Science. Malden, Mass.
Zar, Jerrold H. 1984. Biostatistical Analysis (Second Edition). Prentice-Hall. Englewood Cliffs, New Jersey.

Report Definitions

Power is the probability of rejecting a false null hypothesis. It should be close to one.
N is the size of the sample drawn from the population. To conserve resources, it should be small.
Alpha is the probability of rejecting a true null hypothesis. It should be small.
Beta is the probability of accepting a false null hypothesis. It should be small.
P0 is the value of the population proportion under the null hypothesis.
P1 is the value of the population proportion under the alternative hypothesis.

Summary Statements

A sample size of 50 achieves 8% power to detect a difference (P1-P0) of 0.0500 using a two-sided binomial test. The target significance level is 0.0500. The actual significance level achieved by this test is 0.0328. These results assume that the population proportion under the null hypothesis is 0.5000.

2

Power Analysis of One Proportion

Chart Section

Figure 5.188: Power analysis table of one proportion test from PSS software

For the power of the test of the difference of two proportions (two sample proportion test), with the objective of determining the sample size, we have to put $n=n_1=n_2$. Therefore the developments we get during our study of the test of the difference of two proportions can be written:

$$Z=\dfrac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_1}-\dfrac{1}{n_2}\right)}}=\dfrac{\hat{p}_1-\hat{p}_2}{\sqrt{\dfrac{2\hat{p}(1-\hat{p})}{n}}}=\dfrac{\hat{p}_1-\hat{p}_2}{\sqrt{\dfrac{2\hat{p}\hat{q}}{n}}} \qquad (5.7.1803)$$

with therefore:

$$\hat{p}=\dfrac{n_1\hat{p}_1+n_2\hat{p}_2}{n_1+n_2}=\dfrac{n(\hat{p}_1+\hat{p}_2)}{2n}=\dfrac{\hat{p}_1+\hat{p}_2}{2} \qquad (5.7.1804)$$

In the same way that we did for the Z-test and the one sample p-test, we have:

$$Z_\beta\sqrt{\mathrm{V}(\hat{p}_2-\hat{p}_1)}+(\hat{p}_2-\hat{p}_1)=Z_{1-\alpha/2}\sqrt{\dfrac{2\hat{p}\hat{q}}{n}}+\dfrac{\hat{p}_1+\hat{p}_2}{2} \qquad (5.7.1805)$$

Therefore:

$$Z_\beta\sqrt{\mathrm{V}(\hat{p}_2-\hat{p}_1)}+(\hat{p}_2-\hat{p}_1)=Z_{1-\alpha/2}\sqrt{\frac{2\hat{p}\hat{q}}{n}}+\hat{p} \qquad (5.7.1806)$$

So what it takes us to assume that the real difference between the two proportions is the average (which is debatable...).

But we also have (as the samples are independent and using the property of the variance):

$$Z_\beta\sqrt{\mathrm{V}(\hat{p}_2)+\mathrm{V}(\hat{p}_1)}+(\hat{p}_2-\hat{p}_1)=Z_{1-\alpha/2}\sqrt{\frac{2\hat{p}\hat{q}}{n}}+\hat{p} \qquad (5.7.1807)$$

Therefore:

$$Z_\beta\sqrt{\frac{\hat{p}_1\hat{q}_1}{n}+\frac{\hat{p}_2\hat{q}_2}{n}}+(\hat{p}_2-\hat{p}_1)=Z_{1-\alpha/2}\sqrt{\frac{2\hat{p}\hat{q}}{n}}+\hat{p} \qquad (5.7.1808)$$

That gives obviously:

$$Z_\beta\sqrt{\frac{\hat{p}_1\hat{q}_1+\hat{p}_2\hat{q}_2}{n}}+(\hat{p}_2-\hat{p}_1)=Z_{1-\alpha/2}\sqrt{\frac{2\hat{p}\hat{q}}{n}}+\hat{p} \qquad (5.7.1809)$$

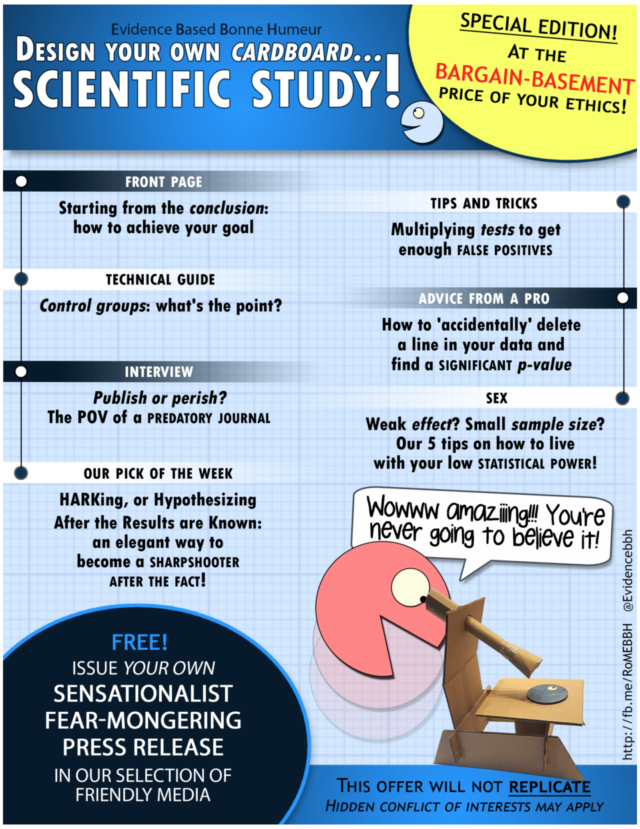
We have then after rearrangement:

$$\boxed{n=\left(\frac{Z_\beta\sqrt{\hat{p}_1\hat{q}_1+\hat{p}_2\hat{q}_2}-Z_{1-\alpha/2}\sqrt{2\hat{p}\hat{q}}}{\hat{p}-(\hat{p}_2-\hat{p}_1)}\right)^2} \qquad (5.7.1810)$$

The `R` software has a function named `power.prop.test( )` that helps us play with the above relation.

Remark

You might find other relations in the literature for the sample size of the one sample or two sample proportions tests. But however be careful...!!! Sadly some textbooks calculates the n based on how to get a given confidence interval, and not on how to get a given power (hence the absence of Z_β the in the corresponding relations). So this has nothing to do with the sample size to get a given power but instead what is the sample size to do p-hacking.......



Evidence Based Bonne Humeur

DESIGN YOUR OWN *CARDBOARD*...
SCIENTIFIC STUDY!

SPECIAL EDITION!
AT THE
BARGAIN-BASEMENT
PRICE OF YOUR ETHICS!

FRONT PAGE
Starting from the *conclusion*:
how to achieve your goal

TIPS AND TRICKS
Multiplying *tests* to get
enough FALSE POSITIVES

TECHNICAL GUIDE
Control groups: what's the point?

ADVICE FROM A PRO
How to 'accidentally' delete
a line in your data and
find a SIGNIFICANT *p-value*

INTERVIEW
Publish or perish?
The POV of a PREDATORY JOURNAL

SEX
Weak *effect*? Small *sample size*?
Our 5 tips on how to live
with your low STATISTICAL POWER!

OUR PICK OF THE WEEK
HARKing, or Hypothesizing
After the Results are Known:
an elegant way to
become a SHARPSHOOTER
AFTER THE FACT!

FREE!
ISSUE *YOUR OWN*
SENSATIONALIST
FEAR-MONGERING
PRESS RELEASE
IN OUR SELECTION OF
FRIENDLY MEDIA

THIS OFFER WILL NOT REPLICATE
HIDDEN CONFLICT OF INTERESTS MAY APPLY

Wowww amaziiing!!! Youre never going to believe it!
http://fb.me/RoMEBBH @Evidencebbh

Now consider that the power is not given to find the sample size, but that we have the sample size and we want to computer the power!

As companion example consider that we are trying to test whether the sex ratio of some sampled individuals significantly differs from the expected sex ratio of 1 assuming we have $n = 64$, of which female=34 and male=30.

In order to find the power, we need to have a specific alternative in mind. Suppose that null hypothesis is $H_0 : p = 0.5$ vs. $H_1 : p > 0.5$, where $p = P(\text{Female})$. Also suppose we have $n = 64$ and you want the power of a test at level $\alpha = 0.05$ against the specific alternative $p = 0.6$.

For an exact binomial test, we need to first to find the critical value c such that $P(X \leq c \,|\, n = 64, p = 0.5)$ is maximized, but still below 0.95. With any statistical software or even spreadsheet software we can find that this value is $c = 40$ (cautious! many statistical softwares do a Normal approximation of the binomial law in such a power calculation...). Then the power of this test against the alternative value $p = 0.6$ is given by $P(X \leq 40 \,|\, n = 64, p = 0.6) = 0.3927$.

5.7.13.3.5 Power calculation based on simulations

In more complicated models it is not possible to determine an explicit expression for the power of a test and calculation of power based on simulations can be an attractive approach. Sometimes it may also just be easier to let the computer do the job by running simulations rather than to get bugged down in derivations of explicit expressions for power even though they may in fact be possible to derive.

Recall that power is the probability of getting a significant result when there is in fact a difference, thus in the long run it is the proportion of significant results to the total number of tests:

$$\text{power} = \frac{\text{no. } p\text{-values } < \alpha}{\text{no. tests}} \tag{5.7.1811}$$

We can let the computer generate random data from the model under the alternative hypothesis and then perform the significance test. We can even do that many many times and record the p-values allowing us to calculate the power via the relation above.

If the number of cases is already available, power calculations are not helpful. It is an attempt to interpret a result retrospectively via the power, which is not useful because no direct connection can be shown. Post-hoc power calculations should therefore be avoided. This has also been discussed in the literature for several years and a well-known article on the subject is [256].

Now that we have seen what are p-values, confidence intervals and power in the framework of null hypothesis statistical testing, the reader interested to take a look at a paper dedicated only to the topic of the most common misinterpretations of these three concepts can refer to: *Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations* (see [216]).

5.7.13.4 A/B Tests

"A/B testing" (also named "bucket tests" or "split-run testing") is a randomized experiment with two variants, A and B. It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics. A/B testing is a way to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective.

Booming in business and a staple analysis in medical trials, the A/B tests assesses the effect of an intervention or treatment by comparing its success rate with that of a control condition. As the name implies, two versions (A and B) are compared, which are identical except for one variation that might affect a user's behaviour. Version A might be the currently used version (control), while version B is modified in some respect (treatment). For instance, on an e-commerce website the purchase funnel is typically a good candidate for A/B testing, as even marginal improvements in drop-off rates can represent a significant gain in sales. Significant improvements can sometimes be seen through testing elements like copy text, layouts, images and colors, but not always.

Remark

When running an A/B tests on something and regularly check ongoing experiments for significant results, you must be careful to not fall into the trap that statisticians call "repeated significance testing errors". Indeed, the significance calculation makes always critical assumption that we have that the sample size fixed in advance! If we wait and the sample size is too big, we fall again in the p-hacking case where the sample size is too big and then creates a false significant test.

Let us see three typical A/B tests by ascending order of mathematical complexity (the first one may be considered as a "frequentist test" and the two others as "bayesian tests").

5.7.13.4.1 Naive proportion approach

The most naive A/B test approach is completely based on the one-proportion p test" that is given for reminder (see page 872) by:

$$\hat{p}-Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\leq p\leq \hat{p}+Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{5.7.1812}$$

but where the notation is traditionally changed as following:

$$\widehat{\text{CR}}-Z_{\alpha/2}\sqrt{\frac{\widehat{\text{CR}}(1-\widehat{\text{CR}})}{n}}\leq \text{CR}\leq \widehat{\text{CR}}+Z_{\alpha/2}\sqrt{\frac{\widehat{\text{CR}}(1-\widehat{\text{CR}})}{n}} \tag{5.7.1813}$$

where CR is the abbreviation of "conversion rate".

Practitioners of that naive test also define the "uplift" as being simply the relative increase in conversion rate between Group A and Group B. It is possible to have negative uplift if our Control group is more effective than the new Test group. Here is the expression:

$$\text{Uplift}=\frac{\text{Conversion rate Test group}}{\text{Conversion rate Control group}}-1 \tag{5.7.1814}$$

Obviously using the standard error measure of the one proportion p test and Z-value we can also compute the minimum and maximum expected uplift within a specific level of confidence if needed!

5.7.13.4.2 Fieller's test (ratio of two means)

The purpose of the "Fieller's test" is the calculation of a confidence interval for the ratio of two means. Its is an especially trendy test in the field of webdesign when two design or user interface of a same website (or the number of clicks) have to be compared. The Fieller's test is also commonly known under the name "A/B test for means".

Let us consider X and Y two random variables Normally distributed representing averages of samples of size n and m respectively and having same variance, such that:

$$\begin{aligned}X &= \mathcal{N}\left(\mu,\frac{\sigma^2}{n}\right)=\mathcal{N}(\mu,a\sigma^2)\\ Y &= \mathcal{N}\left(\nu,\frac{\sigma^2}{m}\right)=\mathcal{N}(\nu,b\sigma^2)\end{aligned} \tag{5.7.1815}$$

So this is a homoscedastic case and then in practice we take the global standard deviation.

In reality, however, we will have access only to the estimators:

$$X=\mathcal{N}\left(\hat{\mu},\frac{\hat{\sigma}^2}{n}\right)=\mathcal{N}(\hat{\mu},aS^2)Y \quad =\mathcal{N}\left(\hat{\nu},\frac{\hat{\sigma}^2}{m}\right)=\mathcal{N}(\hat{\nu},bS^2) \tag{5.7.1816}$$

What interests us is to build a confidence interval for the ratio:

$$\theta=\frac{\mu}{\nu} \tag{5.7.1817}$$

And this is the purpose of Fieller's test obviously!

The idea is then to compute and to group the two random variables into one:

$$X-\theta Y \tag{5.7.1818}$$

which by the null hypothesis will therefore follow a random variable of the type:

$$\mathcal{N}(0,(a+\theta^2 b)\sigma^2) \tag{5.7.1819}$$

We then have in extenso:

$$\frac{X-\theta Y}{\sigma\sqrt{a+\theta^2 b}}=\mathcal{N}(0,1)=Z \tag{5.7.1820}$$

If the standard deviation can only be estimated, we know that we then have:

$$\frac{X-\theta Y}{S\sqrt{a+\theta^2 b}}=T(m+n-2)=T(k) \tag{5.7.1821}$$

It follows, of course, that:

$$P\left(-T_{\alpha/2}(k)\leq\frac{X-\theta Y}{S\sqrt{a+\theta^2 b}}\leq T_{\alpha/2}(k)\right)=1-\alpha \tag{5.7.1822}$$

By symmetry of the Student's law we can therefore write that:

$$P\left(\frac{(X-\theta Y)^2}{S^2(a+\theta^2 b)}\right)=1-\alpha \tag{5.7.1823}$$

After rearrangement, we get:

$$P\left((Y^2-bS^2T^2_{\alpha/2}(k))\theta^2-2XY\theta+(X^2-aS^2T^2_{\alpha/2}(k))\leq 0\right)=1-\alpha \tag{5.7.1824}$$

There are now two interesting situations to consider depending on the sign of the coefficient θ^2 but let's put it aside for now concentrating only on the roots of the polynomial in the parenthesis. We then have (see section Calculus page 1383):

$$\theta_{1,2}=\frac{2XY\pm\sqrt{(-2XY)^2-4(Y^2-bS^2T^2_{\alpha/2}(k))(X^2-aS^2T^2_{\alpha/2}(k))}}{2(Y^2-bS^2S^2_{\alpha/2}(k))} \tag{5.7.1825}$$

This is simplified immediately into:

$$\theta_{1,2}=\frac{XY\pm\sqrt{(XY)^2-(Y^2-bS^2T^2_{\alpha/2}(k))(X^2-aS^2T^2_{\alpha/2}(k))}}{Y^2-bS^2S^2_{\alpha/2}(k)} \tag{5.7.1826}$$

The development in the root gives:

$$\theta_{1,2}=\frac{XY\pm\sqrt{aY^2S^2T_{\alpha/2}(k)+bX^2S^2T^2_{\alpha/2}(k)-abS^4T^4_{\alpha/2}(k)}}{Y^2-bS^2T^2_{\alpha/2}(k)} \tag{5.7.1827}$$

What can be rewritten in a form that will be useful to us:

$$\theta_{1,2}=\frac{\dfrac{X}{Y}\pm\dfrac{1}{Y}\sqrt{aS^2T_{\alpha/2}(k)+b\left(\dfrac{X}{Y}\right)^2S^2T^2_{\alpha/2}(k)-\dfrac{1}{Y^2}abS^4T^4_{\alpha/2}(k)}}{1-\dfrac{bS^2T^2_{\alpha/2}(k)}{Y^2}} \tag{5.7.1828}$$

Let us put:

$$g=\frac{bS^2T^2_{\alpha/2}(k)}{Y^2} \tag{5.7.1829}$$

Our roots can then be written:

$$\theta_{1,2}=\frac{\dfrac{X}{Y}\pm\sqrt{aS^2T_{\alpha/2}(k)+X^2g-gaS^2T^2_{\alpha/2}(k)}}{1-g} \tag{5.7.1830}$$

or:

$$\theta_{1,2}=\frac{\dfrac{X}{Y}\pm\dfrac{1}{Y}\sqrt{X^2g+aS^2T^2_{\alpha/2}(k)(1-g)}}{1-g} \tag{5.7.1831}$$

or otherwise:

$$\theta_{1,2}=\frac{\dfrac{X}{Y}\pm\dfrac{1}{Y}\sqrt{X^2g+a\dfrac{gY^2}{b}(1-g)}}{1-g} \tag{5.7.1832}$$

We do a last simplification:

$$\theta_{1,2}=\frac{\dfrac{X}{Y}\pm\sqrt{\left(\dfrac{X}{Y}\right)^2g+\dfrac{a}{b}g(1-g)}}{1-g} \tag{5.7.1833}$$

Who are therefore for recall the roots of:

$$P\left((Y^2-bS^2T^2_{\alpha/2}(k))\theta^2-2XY\theta+(X^2-aS^2T^2_{\alpha/2}(k))\leq 0\right)=1-\alpha \tag{5.7.1834}$$

The coefficient of θ^2 in the polynomial is $Y^2-bS^2T^2_{\alpha/2}(k)$. Therefore:

- If the latter is greater than zero, then the parabola (see section Analytical Geometric page 2776) is concave (seen from above) and the confidence interval is then between the two roots of the parabola (because the values there are negative according to the imposed inequality).
- If the coefficient of θ^2 in the polynomial is negative then the parabola is convex (seen from above) and the confidence interval is then outside the two roots of the parabola because the values there are negative according to the imposed inequality).
- If the coefficient of θ^2 has a near zero probability of being equal to zero we can ignore this scenario.

Obviously only the first case interests us because the second is difficult to interpret (the intervals would have the infinite bounds which is unrealistic and especially useless). So we must have:

$$Y^2-bS^2T^2_{\alpha/2}(k)>0 \tag{5.7.1835}$$

What can be written:

$$\frac{bS^2T^2_{\alpha/2}(k)}{Y^2}<1 \tag{5.7.1836}$$

Therefore:

$$g<1 \tag{5.7.1837}$$

✍**Example:**

Let us consider the case where we have:

$$\begin{aligned} X &= \mu_X = 160.3684 \\ m &= 19 \\ Y &= \mu_Y = 155 \\ n &= 19 \\ S &= 13.44278 \end{aligned} \tag{5.7.1838}$$

Therefore:

$$\begin{aligned} a &= \frac{1}{m} = \frac{1}{19} \\ b &= \frac{1}{n} = \frac{1}{19} \\ k &= m+n-2 = 36 \\ T^2_{\alpha/2}(k) &= T^2_{5\%/2}(36) = 1.688298 \end{aligned} \tag{5.7.1839}$$

We have:

$$\frac{X}{Y} = \frac{160.3684}{155} \cong 1.034635 \tag{5.7.1840}$$

and:

$$g = \frac{bS^2T^2_{\alpha/2}(k)}{Y^2} = 0.00112839 \tag{5.7.1841}$$

As well as (we give this calculation because some statistical software - `R` in particular - given them in the output while normally it is only useful for Student's t test of two independent homoscedastic samples):

$$\begin{aligned} \mathrm{V}_R &= \frac{(m-1)S_1^2+(n-1)S_2^2}{m+n-2} = 178.228 \\ T(36) &= \frac{(X-Y)-(\mu-\nu)}{\sqrt{\mathrm{V}_r\left(\dfrac{1}{m}+\dfrac{1}{n}\right)}} = \frac{X-Y}{\sqrt{\mathrm{V}_r\left(\dfrac{1}{m}+\dfrac{1}{n}\right)}} = 1.2397 \end{aligned} \tag{5.7.1842}$$

We then have the following interval between the two roots of the parabola:

$$\theta_{1,2} = \frac{\dfrac{X}{Y} \pm \sqrt{\left(\dfrac{X}{Y}\right)^2 g + \dfrac{a}{b}g(1-g)}}{1-g} = \{0.987, 1.084\} \tag{5.7.1843}$$

We can go a little further by considering the case where:

$$\frac{\text{cov}(X,Y)}{\sigma^2} = \rho \Leftrightarrow (X,Y) = \sigma^2\rho \tag{5.7.1844}$$

Either within the estimated framework:

$$\frac{\text{cov}(X,Y)}{\sigma^2} = S^2R \tag{5.7.1845}$$

This gives us the following distribution (assuming Normality) for the random variable grouping the first two according to $X - \theta Y$:

$$\mathcal{N}\left(0, aS^2 - 2\theta RS^2 + b\theta^2 S^2\right) \tag{5.7.1846}$$

We then have:

$$\frac{X - \theta Y}{S\sqrt{a - 2\theta R + b\theta^2}} = T(m + n - 2) = T(k) \tag{5.7.1847}$$

And therefore, as before, we can write:

$$P\left(\frac{(X - \theta Y)^2}{S^2(a - 2\theta R + b\theta^2)} \leq T^2_{\alpha/2}\right) = 1 - \alpha \tag{5.7.1848}$$

What we can rewrite in the following form:

$$P\left((X - \theta Y)^2 \leq T^2_{\alpha/2}(k)S^2(a - 2\theta R + b\theta^2)\right) = 1 - \alpha \tag{5.7.1849}$$

What developed gives:

$$P\left(X^2 - 2\theta XY + \theta^2 Y^2 \leq T_{\alpha/2}(k)S^2 a - 2\theta RT^2_{\alpha/2}(k)S^2 + b\theta^2 T_{\alpha/2}(k)S^2\right) = 1 - \alpha \tag{5.7.1850}$$

We group together the terms as earlier above:

$$P\left(\theta^2(Y^2 - bT^2_{\alpha/2}(k)S^2) - 2\theta(XY - RT^2_{\alpha/2}S^2) + (X^2 - T^2_{\alpha/2}S^2 a) \leq 0\right) \tag{5.7.1851}$$

We then have for roots after simplification of the numerical factors:

$$\theta_{1,2} = \frac{XY \pm \sqrt{(XY - RT^2_{\alpha/2}(k)S^2)^2 - (Y^2 - bT^2_{\alpha/2}(k)S^2)(X^2 - T^2_{\alpha/2}S^2 a)}}{Y^2 - bT^2_{\alpha/2}(k)S^2} \tag{5.7.1852}$$

After the simplification method seen above (pfff...) we finally get:

$$\theta_{1,2} = \frac{\dfrac{X}{Y} \pm \dfrac{1}{Y^2}ST^2_{\alpha/2}(k)\sqrt{-2RXY + RS^2T^2_{\alpha/2}(k) + bX^2 + a(Y^2 - bS^2T^2_{\alpha/2}(k))}}{1 - g} \tag{5.7.1853}$$

5.7.13.4.3 Bayesian A/B testing for binary outcomes

For a binary-outcome test (e.g. a test of conversion rates), the probability that B will beat A in the long run is given by[61]:

$$P(p_B > p_A) = \sum_{i=0}^{\alpha_B - 1} \frac{B(\alpha_A + i, \beta_B + \beta_A)}{(\beta_B + i)B(1 + i, \beta_B)B(\alpha_A, \beta_A)} \tag{5.7.1854}$$

where:

---

[61] The derivation and text mainly comes from https://www.evanmiller.org/bayesian-ab-testing.html.

- α_A is one plus the number of successes for A
- β_A is one plus the number of failures for A
- α_B is one plus the number of successes for B
- β_B is one plus the number of failures for B
- B is the beta function as seen on page 761 (keep in mind we must have $a>0,b>0$):

$$P_{a,b}(x)=\frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}1_{]0,1[}=\frac{x^{a-1}(1-x)^{b-1}}{\int\limits_0^1 x^{a-1}(1-x)^{b-1}\mathrm{d}x}1_{]0,1[} \tag{5.7.1855}$$

Proof 5.48.4. The beta distribution is a convenient prior distribution for modelling a binomial parameter p (as we know that this is one of the rare distribution that has a domain of definition in $[0,1]$). Starting from a non-informative prior, the distribution of p after S successes and F failures is given by $B(S+1,F+1)$. For the remainder of the discussion let $\alpha=S+1$ and $\beta=F+1$, which are just the two parameters of the beta distribution for the belief.

Now suppose we have two experimental branches (A and B) and have a Bayesian belief for each one:

$$p_A=B(\alpha_A,\beta_A)\qquad p_B=B(\alpha_B,\beta_B) \tag{5.7.1856}$$

Using the probability density function of the beta distribution as prior for p, we can get the total probability that p_B is greater than p_A by integrating the joint distribution over all values for which p_B>p_A:

$$P(p_B>p_A)=\int\limits_0^1\int\limits_{p_A}^1\frac{p_A^{\alpha_A-1}(1-p_A)^{\beta_A-1}}{B(\alpha_A,\beta_A)}\frac{{p_B}^{\alpha_B-1}(1-p_B)^{\beta_B-1}}{B(\alpha_B,\beta_B)}\mathrm{d}p_B\mathrm{d}p_A \tag{5.7.1857}$$

sometimes named the "Euler's hypergeometric integral".

Evaluating the inner integral:

$$\begin{aligned}\int\limits_{p_A}^1\frac{{p_B}^{\alpha_B-1}(1-p_B)^{\beta_B-1}}{B(\alpha_B,\beta_B)}\mathrm{d}p_B&=\int\limits_{p_A}^1\frac{{p_B}^{\alpha_B-1}(1-p_B)^{\beta_B-1}}{\int\limits_0^1 x^{a-1}(1-x)^{b-1}\mathrm{d}x}\mathrm{d}p_B\\&=\frac{1}{\int\limits_0^1 x^{a-1}(1-x)^{b-1}\mathrm{d}x}\int\limits_{p_A}^1{p_B}^{\alpha_B-1}(1-p_B)^{\beta_B-1}\mathrm{d}p_B\\&=\frac{1}{B(\alpha_B,\beta_B)}\int\limits_{p_A}^1{p_B}^{\alpha_B-1}(1-p_B)^{\beta_B-1}\mathrm{d}p_B\end{aligned} \tag{5.7.1858}$$

the relation becomes:

$$P(p_B > p_A) = 1 - \int\limits_0^1 \frac{p_A^{\alpha_A - 1}(1-p_A)^{\beta_A - 1}}{B(\alpha_A, \beta_A)} I_{p_A}(\alpha_B, \beta_B) \mathrm{d}p_A \tag{5.7.1859}$$

Where I_x is the incomplete beta function (see section Statistics page 764). Using the identity $I_x(1,b) = 1-(1-x)b$, the recursive relationship:

$$I_x(a,b) = I_x(a-1,b) - \frac{x^{a-1}(1-x)^b}{(a-1)B(a-1,b)} \tag{5.7.1860}$$

and the fact that α and β are integers, we can express I_x as:

$$I_x(a,b) = 1-(1-x)^b - \sum_{j=1}^{a-1} \frac{x^{a-j}(1-x)^b}{(a-j)B(a-j,b)} \tag{5.7.1861}$$

Or equivalently:

$$I_x(a,b) = 1 - \sum_{i=0}^{a-1} \frac{x^i(1-x)^b}{(b+i)B(1+i,b)} \tag{5.7.1862}$$

Our probability integral can therefore be written:

$$\begin{aligned}
P(p_B > p_A) &= 1 - \int\limits_0^1 \frac{p_A^{\alpha_A - 1}(1-p_A)^{\beta_A - 1}}{B(\alpha_A, \beta_A)} I_{p_A}(\alpha_B, \beta_B) \mathrm{d}p_A \\
&= 1 - \int\limits_0^1 \frac{p_A^{\alpha_A - 1}(1-p_A)^{\beta_A - 1}}{B(\alpha_A, \beta_A)} \left(1 - \sum_{i=0}^{\alpha_B - 1} \frac{p_A^i (1-p_A)^{\beta_B}}{(\beta_B + i)B(1+i,\beta_B)}\right) \mathrm{d}p_A \\
&= 1 - 1 + \int\limits_0^1 \frac{p_A^{\alpha_A - 1}(1-p_A)^{\beta_A - 1}}{B(\alpha_A, \beta_A)} \sum_{i=0}^{\alpha_B - 1} \frac{p_A^i (1-p_A)^{\beta_B}}{(\beta_B + i)B(1+i,\beta_B)} \mathrm{d}p_A \\
&= \int_0^1 \sum_{i=0}^{\alpha_B - 1} \frac{p_A^{\alpha_A - 1 + i}(1-p_A)^{\beta_A + \beta_B - 1}}{(\beta_B + i)B(\alpha_A, \beta_A)B(1+i,\beta_B)} \mathrm{d}p_A \\
&= \sum_{i=0}^{\alpha_B - 1} \int\limits_0^1 \frac{p_A^{\alpha_A - 1 + i}(1-p_A)^{\beta_A + \beta_B - 1}}{(\beta_B + i)B(\alpha_A, \beta_A)B(1+i,\beta_B)} \mathrm{d}p_A \\
&= \sum_{i=0}^{\alpha_B - 1} \frac{B(\alpha_A + i, \beta_A + \beta_B)}{(\beta_B + i)B(\alpha_A, \beta_A)B(1+i,\beta_B)} \int\limits_0^1 \frac{p_A^{\alpha_A - 1 + i}(1-p_A)^{\beta_A + \beta_B - 1}}{B(\alpha_A + i, \beta_A + \beta_B)} \mathrm{d}p_A
\end{aligned} \tag{5.7.1863}$$

Finally:

$$\boxed{P(p_B > p_A) = \sum_{i=0}^{\alpha_B - 1} \frac{B(\alpha_A + i, \beta_A + \beta_B)}{(\beta_B + i)B(1+i,\beta_B)B(\alpha_A, \beta_A)}} \tag{5.7.1864}$$

■ Q.E.D.

The reader must know that even if there exist a closed formula, a the day we write these lines, most statistical software implement numerical methods (simulations) to calculate this probability!

Remarks

R1. If we want to assume that each value of p has the same probability, i.e. we have no out-of-data (prior) information that we would like to include in our model, then we can set $\alpha_A = \beta_A = \alpha_B = \beta_B$, that means a uniform distribution for all the values of p in $[0, 1]$ interval.

R2. The reader can refer to our `R` companion free e-book to see how to apply these equations easily on practical cases with a free open-source software and code!

5.7.13.4.4 Bayesian A/B testing for count datas

Analysing count data (for example, if you're comparing the number of sales per salesman, or number of sales week over week) requires a different relation. The probability that group 1 has a higher arrival rate than group 2 is given by:

$$P(\lambda_1 > \lambda_2) = \sum_{k=0}^{\alpha_1 - 1} \frac{(\beta_1+\beta_2)^{-(k+\alpha_2)}\beta_1^k \beta_2^{\alpha_2}}{(k+\alpha_2)B(k+1,\alpha_2)} \tag{5.7.1865}$$

A derivation and implementation follow; they both closely mirror the binary-outcome case.

Proof 5.48.5. Let λ_1 and λ_2 be the Poisson parameter for each group. With a gamma-distributed prior belief (see section Statistics page 767), the posterior beliefs are given by:

$$\lambda_1 = \Gamma(\alpha_1, \beta_1) \qquad \lambda_2 = \Gamma(\alpha_2, \beta_2) \tag{5.7.1866}$$

Using the probability density function of the gamma distribution, we can get the total probability that λ_1 is greater than λ_2 by integrating the joint distribution over all values for which $\lambda_1 > \lambda_2$:

$$P(\lambda_1 > \lambda_2) = \int\limits_0^{+\infty}\int\limits_{\lambda_2}^{+\infty} \Gamma(\alpha_1,\beta_1)\Gamma(\alpha_2,\beta_2)\mathrm{d}\lambda_1\mathrm{d}\lambda_2 = \int\limits_0^{+\infty}\int\limits_{\lambda_2}^{+\infty} \frac{\beta_1^{\alpha_1}\lambda_1^{\alpha_1-1}e^{-\beta_1\lambda_1}}{\Gamma(\alpha_1)}\frac{\beta_2^{\alpha_2}\lambda_2^{\alpha_2-1}e^{-\beta_2\lambda_2}}{\Gamma(\alpha_2)}\mathrm{d}\lambda_1\mathrm{d}\lambda_2 \tag{5.7.1867}$$

Evaluating the inner integral, the equation becomes:

$$\int\limits_0^{+\infty} Q(\alpha_1, \beta_1\lambda_2)\frac{\beta_2^{\alpha_2}\lambda_2^{\alpha_2-1}e^{-\beta_2\lambda_2}}{\Gamma(\alpha_2)}\mathrm{d}\lambda_2 \tag{5.7.1868}$$

Where Q is the upper incomplete regularized gamma function (see section Differential and Integral Calculus page 1578). Using the identity $Q(1, z) = e^{-z}$ and the recursive relationship:

$$Q(a+n, z) = Q(a, z) + z^a e^{-z}\sum_{k=0}^{n-1}\frac{z^k}{\Gamma(a+k+1)} \tag{5.7.1869}$$

we can express Q as (see section Differential and Integral Calculus page 1578):

$$Q(n,z)=e^{-z}\sum_{k=0}^{n-1}\frac{z^k}{\Gamma(k+1)} \tag{5.7.1870}$$

The probability integral can therefore be written (we used the integral already seen during our study of the Gamma function at page 1574):

$$\begin{aligned}P(\lambda_1>\lambda_2)&=\int\limits_0^{+\infty}e^{-\beta_1\lambda_2}\left(\sum_{k=0}^{\alpha_1-1}\frac{(\beta_1\lambda_2)^k}{\Gamma(k+1)}\right)\frac{\beta_2^{\alpha_2}\lambda_2^{\alpha_2-1}e^{-\beta_2\lambda_2}}{\Gamma(\alpha_2)}\mathrm{d}\lambda_2\\&=\sum_{k=0}^{\alpha_1-1}\int\limits_0^{+\infty}\frac{e^{-\beta_1\lambda_2}(\beta_1\lambda_2)^k}{\Gamma(k+1)}\frac{\beta_2^{\alpha_2}\lambda_2^{\alpha_2-1}e^{-\beta_2\lambda_2}}{\Gamma(\alpha_2)}\mathrm{d}\lambda_2\\&=\sum_{k=0}^{\alpha_1-1}\frac{\beta_1^k\beta_2^{\alpha_2}}{\Gamma(k+1)\Gamma(\alpha_2)}\int_0^{+\infty}e^{-(\beta_1+\beta_2)\lambda_2}\lambda_2^{k+\alpha_2-1}\mathrm{d}\lambda_2\\&=\sum_{k=0}^{\alpha_1-1}\frac{\beta_1^k\beta_2^{\alpha_2}}{\Gamma(k+1)\Gamma(\alpha_2)}\frac{\Gamma(k+\alpha_2)}{(\beta_1+\beta_2)^{(k+\alpha_2)}}\\&=\sum_{k=0}^{\alpha_1-1}\beta_1^k\beta_2^{\alpha_2}(\beta_1+\beta_2)^{-(k+\alpha_2)}\frac{\Gamma(k+\alpha_2)}{\Gamma(k+1)\Gamma(\alpha_2)}\frac{k+\alpha_2}{k+\alpha_2}\end{aligned} \tag{5.7.1871}$$

Using $\Gamma(z+1)=z\Gamma(z)$, we have:

$$P(\lambda_1>\lambda_2)=\sum_{k=0}^{\alpha_1-1}\beta_1^k\beta_2^{\alpha_2}(\beta_1+\beta_2)^{-(k+\alpha_2)}\frac{\Gamma(k+\alpha_2+1)}{\Gamma(k+1)\Gamma(\alpha_2)}\frac{1}{k+\alpha_2} \tag{5.7.1872}$$

Replacing the gamma functions with a beta function (see page 761), we have:

$$\boxed{P(\lambda_1>\lambda_2)=\sum_{k=0}^{\alpha_1-1}\frac{(\beta_1+\beta_2)^{-(k+\alpha_2)}\beta_1^k\beta_2^{\alpha_2}}{(k+\alpha_2)B(k+1,\alpha_2)}} \tag{5.7.1873}$$

■ Q.E.D.

Remark

Again, the reader can refer to our R companion free e-book to see how to apply these equations easily on practical cases with a free open-source software and code!

5.7.13.5 Analysis Of VAriance (ANOVA)

"Analysis Of VAriance" is a collection of statistical models used to analyse the differences among group means and their associated procedures (such as "variation" among and between groups), developed by statistician and evolutionary biologist Ronald Fisher. In the ANOVA setting, the observed variance in a particular variable is partitioned into components attributable to different sources of variation.

The main techniques of analysis of variances are the following one (some of them are detailed further below) in the order of most to less used one (and in increasing order of complexity):

- Two variable one-way fixed factor ANOVA (Student T-test)[62]
- One-way fixed factor ANOVA (n independent fixed variables)[63]
- Two-way fixed factor ANOVA with/without repetitions[64]
- Multifactor ANOVA with/without repetition
- Repeated Measures ANOVA (RM-ANOVA)[65]
- Block[66] (randomized) design (ANOVA in blocks)
- Latin square ANOVA
- Greaco-Latin square ANOVA
- Youden square ANOVA[67]
- Split-Plot ANOVA
- Split-Block ANOVA
- Strip-Plot ANOVA
- Split-Split-Plot ANOVA
- Analysis of Covariance (ANCOVA)
- Nested (Hierarchical) analysis of variance (HANOVA)
- Partially Nested ANOVA (PANOVA)
- Multivariate analysis of variance (MANOVA)
- Random effects ANOVA
- Mixed effects (with fixed and random factors) ANOVA
- Cross-Over ANOVA
- Robusts ANOVA (by permutations, M-estimators based)
- ...

and some of the ANOVA above are said "balanced" or "unbalanced", depending if the number of measurement are equal or not equal for each variable and therefore can be calculated with different types of sum of square errors in the unbalance case, named "Type I sum of squares",

---

[62] There is a nonparametric rank-transformed equivalent named Friedman test that we will see further below.

[63] There is a nonparametric rank-transformed equivalent named Kruskal-Wallis where the post-hoc will be typically the Mann-Whitney U test that we will all see further below.

[64] In facts for all ANOVA there exist a mathematical variant for repeated measurements.

[65] Repeated measures ANOVA is the equivalent of the one-way ANOVA, but for related, not independent groups, and is the extension of the dependent T-test. A repeated measures ANOVA is also referred to as a within-subjects ANOVA or ANOVA for correlated samples.

[66] There is a variant where blocked are considered as fixed or random... In facts the block is considered as an another factor. Then it's just a two-way ANOVA

[67] These are incomplete Latin squares, in which the number of columns is not equal to the number of rows.

"Type II sum of squares" or "Type III sum of squares". So we can say that there is at least between 20 and 60 types of different ANOVA[68] (including the special variations but without including the randomized/bootstrapped versions...)!

It must also be known that the terminology of ANOVA is largely from the statistical design of experiments (see section Industrial Engineering page 8780). The experimenter adjusts factors and measures responses in an attempt to determine an effect. Factors are assigned to experimental units by a combination of randomization[69] and blocking to ensure the validity of the results.

Remark

For the reader interested to have a deeper overview, however without too much mathematical details, at the opposite of the present book, we recommend the following actual worldwide reference for experts in ANOVA and Design of Experiments: *Design and Analysis of Experiments*[a] (authors: Hinkelmann, K. and Kempthorne, O.), approximately 730.- USD and 1,800 pages (see [249],[250] and [251]) and *Applied Linear Statistical Models* (see [324]).

[a] Three volume set!!!

The purpose of this section will mainly have for purpose to avoid the following type of telephone conversation between an ANOVA specialist (S) and a research scientist or engineer (R). R: « Hello, Mr. Stat, I wonder whether you have just a minute for a quick statistical question. » S: « Usually I do not do statistical consulting over the phone, but let me see if I can help you. What is the problem? » R: « We are developing new growth media for industrial producers for growing flower plants. We have three such media and we use them with four flower varieties. We have five replications for each combination of medium and flower. We have analysed the data as a 3×4 two-way classification with five observations per cell. But my graduate assistant has talked to one of your students and he is now confused about the validity of this analysis. I just want you to confirm that we have done the right thing. » S: « Well, I do not know.» R: «What do you mean, you do not know? You are the expert!? » S: « I really need to know more about how you performed the experiment. For example, how did you prepare the media that you used in the individual pots? I assume that you grow the flowers in pots in the greenhouse. » R: « Yes, that is right. My graduate assistant simply mixed each medium in a big container, which we then put in the individual pots. » S: « That may be a problem, because now you may not have any replication. » R: « What do you mean, we have no replication? I just told you that we have five replications. » S: « Yes, but ... I think it would be best if you would come to my office for me to explain this to you and to take a closer look at your experiment. » R: « But we have already submitted the paper for publication. » S: « Then why don't you come when you get the reviews back.» Silence. R: « Yes, thank you. I'll do that... »

This is, of course, just a fictitious conversation[70]. But many consulting statisticians have had similar conversations (most of the time the paper was already rejected). The aim of this section is to help engineers not only to better understand each other but also to obtain a better understanding of the intricacies of designing and analysing experiments.

---

[68] For more details see the Data Science mind map at page 6770.

[69] In the pharmaceutical field (mainly!), we use randomization to make the treated groups (if they are sufficiently large and comparable) and untreated groups comparable in terms of potential severity factors, which makes it possible to neutralize the differences due to the factors of severity, in order to isolate the effect of the treatment. It should be noted that randomization does not only neutralize known factors but also unknown severity factors!

[70] Another typical real example can be found on the following YouTube video: `https://vimeo.com/815286997`

5.7.13.5.1 ANOVA with one fixed factor

The objective of the analysis of variance (at the contrary to what its name might suggest...) is a statistical technique for comparing the means of two or more populations and that is very widely used in the pharmaceutical field or in the R&D labs or benches test[71]. This method, however, takes its name from that it use measures of variance to determine the statistical significance, or not, of the differences of measured averages on populations or samples.

More precisely, the real meaning is to know if the fact that the sample averages are (slightly) different can be assigned at random sampling or due to the fact that a variability factor actually generates significantly different samples (if we have the population size, we have obviously not such question!).

Remark

For more information about the vocabulary and application, the engineer and the researcher should refer to standard ISO 3534-3:1999.

For the analysis of variance named also "one factor ANOVA" (Analysis Of VAriance) or "one factor ANAVAR" (ANAlysis of VARiance), or "one-way ANOVA" or more rigorously "one fixed factor ANOVA with repetitions" or "ANOVA with one fixed categorical variable with repetition" or "complete random design" (CRD), we first recall, as we proved, that the Fisher-Snedecor law is given by the ratio of two independent random variables that follow a Chi-square law and divided by the degree of freedom such as:

$$\frac{U/k}{V/l}=F_{k,l}(x)=\frac{\Gamma\left(\dfrac{k+l}{2}\right)\left(\dfrac{k}{l}\right)^{\frac{k}{2}}}{\Gamma\left(\dfrac{k}{2}\right)\Gamma\left(\dfrac{l}{2}\right)}\frac{x^{\frac{k}{2}-1}}{\left(1+\dfrac{k}{l}x\right)^{\frac{k+l}{2}}} \tag{5.7.1874}$$

and we will now see its importance.

Remark

When a factor can have a very large number of levels we consider having chosen some given levels of the factor among a multitude of possible as a random selection. This is why we speak then in this latter case of "random factor" which is the subject of specific ANOVA models that we will study once those on fixed factors mastered (e.g. ANOVA mixing fixed and random factors are named "mixed ANOVA").

Let us consider a random sample of size n, say $X_1,X_2,\ldots,X_n$ sampled from the law $\mathcal{N}(\mu_X,\sigma_X)$ and a random sample of size m, say $Y_1,Y_2,\ldots,Y_m$ sampled from the law $\mathcal{N}(\mu_Y,\sigma_Y)$.

Consider now the maximum likelihood estimators of the standard deviation of the Normal law traditionally noted in the field of analysis of variance:

$$S_{*X}^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_X)^2 \quad\text{and}\quad S_{*Y}^2=\frac{1}{n}\sum_{i=1}^{n}(Y_i-\mu_Y)^2 \tag{5.7.1875}$$

---

[71] In the case of "random factor" (concept that we will see later!) however we instead compare variances....!

The above statistics are those that we could use to estimate the variances if the average theoretical μ_X, μ_Y were known. So we can use a result proved above in our study of confidence intervals:

$$\chi^2(n) = \frac{nS^2_{*X}}{\sigma^2_X} \quad \text{and} \quad \chi^2(m) = \frac{mS^2_{*Y}}{\sigma^2_Y} \tag{5.7.1876}$$

As the X_i are independent of the Y_j (hypothesis that suppose the covariance to be zero, the converse being for reminder not always!), the random variables:

$$\frac{nS^2_{*X}}{\sigma^2_X} \quad \text{and} \quad \frac{mS^2_{*Y}}{\sigma^2_Y} \tag{5.7.1877}$$

are independent from each other.

Remark

There is an existing type of ANOVA for the case where the variables are not independent (we then speak about "covariate"). This is the "ANCOVA" which means "ANalysis of COvariance and VAriance" that uses a mix of linear regression (see section Numerical Methods page 6329) and ANOVA. The purpose of the ANCOVA is to statistically eliminate the indirect effect of the covariate. We will see much later how it works in details.

We can therefore apply the Fisher-Snedecor law with:

$$U = \frac{nS^2_{*X}}{\sigma^2_X};\ k = n \quad \text{and} \quad V = \frac{mS^2_{*Y}}{\sigma^2_Y};\ l = m \tag{5.7.1878}$$

Therefore we get:

$$\frac{U/k}{V/l} = \frac{\frac{nS^2_{*X}}{\sigma^2_X}/n}{\frac{mS^2_{*Y}}{\sigma^2_Y}/m} = \frac{\frac{S^2_{*X}}{\sigma^2_X}}{\frac{S^2_{*Y}}{\sigma^2_Y}} = \frac{S^2_{*X}\sigma^2_Y}{S^2_{*Y}\sigma^2_X} = F_{n,m}(x) \tag{5.7.1879}$$

Finally:

$$\frac{\sigma^2_X}{\sigma^2_Y} = \frac{1}{F_{n,m}(x)}\frac{S^2_{*X}}{S^2_{*Y}} \tag{5.7.1880}$$

This theorem allows us to deduce the confidence interval of the ratio of two variances when the theoretical mean is known. Since Fisher function is not symmetrical, the only possibility to make the inference is to use numerical calculation and then we will denote for a given confidence interval the test as follows:

$$\boxed{\frac{1}{F_{\alpha/2}(n,m)}\frac{S^2_{*X}}{S^2_{*Y}} \leqslant \frac{\sigma^2_X}{\sigma^2_Y} \leqslant \frac{1}{F_{1-\alpha/2}(n,m)}\frac{S^2_{*X}}{S^2_{*Y}}} \tag{5.7.1881}$$

In the case where the averages μ_X, μ_Y are unknown, we use the unbiased estimators of the variances noted traditionally in the field of analysis of variance:

$$S^2_X = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \text{and} \quad S^2_Y = \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \bar{Y})^2 \tag{5.7.1882}$$

To estimate the theoretical variances, we use the result proved above:

$$\frac{(n-1)S_X^2}{\sigma_X^2}=\chi^2(n-1) \quad \text{and} \quad \frac{(m-1)S_Y^2}{\sigma_Y^2}=\chi^2(m-1) \tag{5.7.1883}$$

As the X_i are independent of the Y_j (assumption!), the variables:

$$\frac{(n-1)S_X^2}{\sigma_X^2} \quad \text{and} \quad \frac{(m-1)S_Y^2}{\sigma_Y^2} \tag{5.7.1884}$$

are independent from each other. We can therefore apply the Fisher-Snedecor distribution:

$$U=\frac{(n-1)S_X^2}{\sigma_X^2};\; k=n-1 \quad \text{and} \quad V=\frac{(m-1)S_Y^2}{\sigma_Y^2};\; l=m-1 \tag{5.7.1885}$$

Therefore we get:

$$\frac{U/k}{V/l}=\frac{\frac{(n-1)S_X^2}{\sigma_X^2}/(n-1)}{\frac{(m-1)S_Y^2}{\sigma_Y^2}/(m-1)}=\frac{\frac{S_X^2}{\sigma_X^2}}{\frac{S_Y^2}{\sigma_Y^2}}=\frac{S_X^2\sigma_Y^2}{S_Y^2\sigma_X^2}=F_{n-1,m-1}(x) \tag{5.7.1886}$$

Finally:

$$\frac{\sigma_X^2}{\sigma_Y^2}=\frac{1}{F_{n-1,m-1}(x)}\frac{S_X^2}{S_Y^2} \tag{5.7.1887}$$

This theorem allows us to deduce the confidence interval of the ratio of two variances when the sample mean is known. Since Fisher function is not symmetrical, the only possibility to make the inference is to use numerical calculation and then we will denote for a given confidence interval the "Fisher F-test" as follows:

$$\boxed{\frac{1}{F_{\alpha/2}(n-1,m-1)}\frac{S_X^2}{S_Y^2}\leqslant\frac{\sigma_X^2}{\sigma_Y^2}\leqslant\frac{1}{F_{1-\alpha/2}(n-1,m-1)}\frac{S_X^2}{S_Y^2}} \tag{5.7.1888}$$

keeping in mind that its use implicitly requires constraints of Normality of the studied variables.

Rondal A. Fisher (11890-11962 according to holocene calendar) is, as Karl Pearson (11857-11936 according to holocene calendar), one of the main founders of the modern theory of statistics. Fisher studied at Cambridge, where he obtained in 11912 (holocene calendar) a degree in astronomy. It is by studying the theory of error in astronomical observations that Fisher became interested in statistics. Fisher is the inventor of the branch of statistics named "analysis of variance".

In the early 120th century (holocene calendar), Rondald A. Fisher therefore develops the very important methodology of experimental design (see section Industrial Engineering page 8780). To confirm the usefulness of a factor, he developed a test to ensure that the different samples are of different natures. This test is based on the analysis of variance (of the samples), and named also ANOVA but for "normalized analysis of variance".

Let us take k samples of n random values each. Each of the values being considered as an observation or measurement of something or on the basis of something (a different place or a different object... in short: a single factor of variability between samples!). We will have a total of N observations (measurements) given by:

$$N = kn \qquad (5.7.1889)$$

If each sample has an equal number of values n (sample size) such that $n_1 = n_2 = \ldots = n_k$ we then speak of "balanced plan" or "balanced design" with k levels (or k modalities).

Remark

If we have more factors of variability (e.g. each place compares himself to different laboratories), then we will talk about "multifactorial ANOVA". Therefore, if there are only two sources of variability, we talking about "two factor ANOVA" (see below for details on the various of two-factor ANOVA).

We will consider that each of the k samples is sampled (follows) a random variable following Normally distributed:

Factor 1			
Sample 1	Sample 2	Sample ...i	Sample k
x_{11} x_{21} $\ldots$ x_{n1}	x_{12} x_{22} $\ldots$ x_{n2}	$\ldots$ $\ldots$ x_{ij} $\ldots$	x_{1k} x_{2k} $\ldots$ x_{nk}
Average: $\bar{x}_1$	Average: $\bar{x}_2$	Average: $\bar{x}_i$	Average: $\bar{x}_k$

Table 5.53: **Typical "cross" table of a one factor ANOVA**

Therefore in a one-way analysis of variance (one-way ANOVA), a single factor (treatment) with several levels is considered.

In terms of testing, we test whether the mean of the k samples of size n are equal under the assumption that their variances are equal. What we write as under hypothesis notation as follows:

$$\begin{aligned} H_0 &: \quad \mu_1 = \mu_2 = \ldots = \mu_k = \mu \\ H_1 &: \quad \exists i, j \quad \mu_i \neq \mu_j \end{aligned} \qquad (5.7.1890)$$

In other words, the samples are representative of the same population (i.e. of a same statistical law). That is to say, the variations between the values of different samples are essentially due to probabilities. For this we study the variability of results in the samples and between samples.

Remark

The statistical model of the observations can also be written in the following very important form, named a "factor effects model" or even sometimes a "classificatory linear model":

$$\boxed{x_{ij} = \mu + \tau_i + \varepsilon_{ij}} \qquad (5.7.1891)$$

The parameter μ is the "overall mean", τ_i is the ith "treatment effect", and ε_{ij} is a "random error". It is assumed that ε_{ij}'s are i.i.d $\mathcal{N}(0, \sigma^2)$. This is a completely additive model (there is no interaction!). Under this form the hypotheses are then:

$$\begin{aligned} H_0 &: \quad \sigma_{\mu_i} = 0 \\ H_1 &: \quad \sigma_{\mu_i} > 0 \end{aligned} \tag{5.7.1892}$$

Notice that we have therefore:

$$\mu = \frac{1}{k}\sum_{i=1}^{k}\mu_i \qquad \text{and} \qquad \sum_{i=1}^{k}\tau_i = 0 \tag{5.7.1893}$$

It is exactly the same as we will see later as writing (formulation found in some publications or books):

$$\begin{aligned} H_0 &: \quad \sigma_{\mu} = 0 \\ H_1 &: \quad \sigma_{\mu} > 0 \end{aligned} \tag{5.7.1894}$$

So we will for the rest denote by i the index number of the sample (1 to k) and j the index of the observation (from 1 to n). Therefore x_{ij} will be the value of the j-th observation of the data sample number i (we chose to reverse the usual notation so be careful not to mislead afterwards ... we're sorry ... it was bad choice me made when we started to write this textbook!).

According to the above hypothesis, we have:

$$x_{ij} = \mathcal{N}(\mu_i, \sigma) \tag{5.7.1895}$$

We will denote by $\bar{x}_i$ the empirical/estimated (arithmetic) average of the sample i (often named "marginal average" and sometimes denoted also $\bar{x}_{i.}$):

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} \quad \Rightarrow \quad \sum_{j=1}^{n_i} x_{ij} = n_i\bar{x}_i \tag{5.7.1896}$$

and $\bar{\bar{x}}$ (also sometimes denoted $\bar{x}_{..}$) the empirical/estimated (arithmetic) average of the N values (therefore the average of the $\bar{x}_i$) is then given by:

$$\bar{\bar{x}} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij} = \frac{1}{N}\sum_{i=1}^{k}(n_i\bar{x}_i) \tag{5.7.1897}$$

Using the properties of the mean and variance already proved above we know that:

$$\bar{x}_i = \mathcal{N}\left(\mu_i, \frac{\sigma_i}{\sqrt{n_i}}\right) \quad \text{and} \quad \bar{\bar{x}} = \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \tag{5.7.1898}$$

with μ which is the average of the true averages (expected means) μ_i:

$$\mu = \frac{1}{N}\sum_{i=1}^{k}\mu_i \tag{5.7.1899}$$

Remark

Writing the model as factor effect model:

$$x_{ij}=\mu+\tau_i+\varepsilon_{ij} \tag{5.7.1900}$$

we see that the maximum likelihood estimators of the model parameters are $\hat{\mu}=\bar{\bar{x}}$ and $\hat{\tau}_i=\bar{x}_i-\bar{\bar{x}}$. Hence:

$$\hat{x}_{ij}=\bar{\bar{x}}+(\bar{x}_i-\bar{\bar{x}})=\bar{x}_i \tag{5.7.1901}$$

Let us now introduce three important variances:

1. The "total variance" as being intuitively the estimated unbiased variance unbiased considering the set of N observations as a single sample:

   $$\mathrm{V}_T=\frac{\displaystyle\sum_{ij}(x_{ij}-\bar{\bar{x}})^2}{N-1} \tag{5.7.1902}$$

   where the numerator is named "sum of the squares of total differences".

2. The "variance between samples" (that is to say between the averages of the samples) is also intuitively the variance estimator of the averages of samples:

   $$\mathrm{V}_E=\frac{\displaystyle\sum_{i}(\bar{x}_i-\bar{\bar{x}})^2}{k-1} \tag{5.7.1903}$$

   where the numerator term is named "sum of the squared differences between samples".

   As we have proved that if all variables are identically distributed (same variance and same mean) and independent the variance of individuals is n times that of the average:

   $$\sigma_{\bar{X}}=\frac{\sigma}{\sqrt{n}} \quad\Leftrightarrow\quad \mathrm{V}(\bar{X})n=\mathrm{V}(X) \tag{5.7.1904}$$

   then the "variance of observations" (random variables in a sample) is given by:

   $$\mathrm{V}_A=n\mathrm{V}_E=\frac{n\displaystyle\sum_{i}(\bar{x}_i-\bar{\bar{x}})^2}{k-1} \tag{5.7.1905}$$

   We then have therefore above the hypothesis of equality of variances that is expressed in mathematical form for the developments that will follow.

3. The "residual variance" is the effect of the "uncontrolled factors". It is by definition the average of the sample variances (its like: standard error):

   $$\mathrm{V}_R=\frac{1}{k}\sum_i\frac{\displaystyle\sum_{j}(x_{ij}-\bar{x}_i)^2}{n-1}=\frac{\displaystyle\sum_{ij}(x_{ij}-\bar{x}_i)^2}{k(n-1)} \tag{5.7.1906}$$

   where the numerator term is named "sum of squared residuals errors" or even more often "residual errors".

Finally, these indicators are sometimes summarized as follows:

$$\begin{array}{lll} Q_T=\sum_{ij}(x_{ij}-\bar{\bar{x}})^2 & \mathrm{df}_T=N-1=kn-1 & \mathrm{V}_T=\dfrac{Q_T}{\mathrm{df}_T} \\ Q_A=n\sum_{i}(\bar{x}_i-\bar{\bar{x}})^2 & \mathrm{df}_A=k-1 & \mathrm{V}_A=\dfrac{Q_A}{\mathrm{df}_A} \\ Q_R=\sum_{ij}(x_{ij}-\bar{x}_i)^2 & \mathrm{df}_R=k(n-1) & \mathrm{V}_R=\dfrac{Q_R}{\mathrm{df}_R} \end{array} \tag{5.7.1907}$$

Note that if the samples do not have the same size (which is rare in practice), then we have:

$$\boxed{\begin{array}{lll} Q_T=\sum_{ij}(x_{ij}-\bar{\bar{x}})^2 & \mathrm{df}_T=N-1 & \mathrm{V}_T=\dfrac{Q_T}{\mathrm{df}_T} \\ Q_A=n\sum_{i}(\bar{x}_i-\bar{\bar{x}})^2 & \mathrm{df}_A=k-1 & \mathrm{V}_A=\dfrac{Q_A}{\mathrm{df}_A} \\ Q_R=\sum_{ij}(x_{ij}-\bar{x}_i)^2 & \mathrm{df}_R=\sum_{i}(n_i-1) & \mathrm{V}_R=\dfrac{Q_R}{\mathrm{df}_R} \end{array}} \tag{5.7.1908}$$

Remarks

R1. The term Q_T is often indicated in the industry by the acronym SST meaning in "Sum of Squares Total" or more rarely TSS for "Total Sum of Squares".

R2. The term Q_A is often indicated in the industry by the acronym SSB meaning in "Sum of Squares Between (samples)" or more rarely SSk for "Sum of Squares Between treatments" (hence sometimes denoted also: $\mathrm{SS}_{\mathrm{tr}}$).

R3. The term Q_R is often indicated in the industry by the acronym SSW meaning in "Sum of Squares Within (samples)" or more rarely SSE for "Sum of Squares due to Errors" (hence sometimes denoted also: SS_E).

Let us indicate that we often see in the literature (we will use a little further below this notation):

$$\begin{aligned} Q_R&=\sum_{ij}(x_{ij}-\bar{x}_i)^2=\frac{n_i-1}{n_i-1}\sum_{ij}(x_{ij}-\bar{x}_i)^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2\\ &=\sum_{i=1}^{k}(n_i-1)\sum_{j=1}^{n_i}\frac{(x_{ij}-\bar{x}_i)^2}{n_i-1}=\sum_{i=1}^{k}(n_i-1)S_i^2 \end{aligned} \tag{5.7.1909}$$

so with the unbiased estimator of the variance of the observations:

$$S_i^2=\frac{1}{n_i-1}\sum_{j}^{n_i}(x_{ij}-\bar{x}_i)^2 \tag{5.7.1910}$$

Before going further, let us stop on the residual variance. We then have for samples that are not of the same size:

$$\boxed{\mathrm{V}_R=\frac{\displaystyle\sum_{ij}(x_{ij}-\bar{x}_i)^2}{\displaystyle\sum_i(n_i-1)}}\tag{5.7.1911}$$

This writing is often named "pooled variance". The square root of a pooled variance estimator is knows obviously as a "pooled standard deviation". We also see that if we assume all the mean $\bar{x}_i$ to be equal and also all samples sizes n_i we fall back on the standard error.

Let us now open a small important parenthesis... Let's take the case of two samples only:

$$\mathrm{V}_R=\frac{\displaystyle\sum_j(x_{1j}-\bar{x}_1)^2+\sum_j(x_{2j}-\bar{x}_2)^2}{(n_1-1)+(n_2-1)}=\frac{\displaystyle\sum_j(x_{1j}-\bar{x}_1)^2+\sum_j(x_{2j}-\bar{x}_2)^2}{n_1+n_2-2}\tag{5.7.1912}$$

Therefore by introducing the maximum likelihood estimator of the variance:

$$\mathrm{V}_R=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{(n_1-1)+(n_2-1)}=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}\tag{5.7.1913}$$

Remark

It is then quite easy to understand where the following relation comes from:

$$V_R=S_R^2=\frac{\displaystyle\sum_{i=1}^{p}\left((n_i-1)S_i^2\right)}{\left(\displaystyle\sum_{i=1}^{p}n_i\right)-p}\tag{5.7.1914}$$

also named the "variance of repeatability of p data series".

In laboratories experiments this also leads to the definition of "inter-series variance" (difference between mean values variance of the data series and the variance of repeatability divided by the number of measurements per series), where given p data series (i varies from 1 to p), each data series containing j values (j varies from 1 to n_i):

$$S_g^2=S_{\mu_i}^2-\frac{S_R^2}{n}=\frac{\dfrac{\displaystyle\sum_{i=1}^{p}\left(n_i(\mu_i-\mu_p)^2\right)}{p-1}-S_R^2}{n}\tag{5.7.1915}$$

In practice, if the variance of the means is smaller than the contribution of the variance of repeatability, S_R^2, the inter-series variance is declared equal to zero.

We can also observe that in the specific case where $n_1=n_2=n$:

$$\mathrm{V}_R=\frac{nS_1^2-S_1^2+nS_2^2-S_2^2}{2n-2}=\frac{n(S_1^2+S_2^2)-(S_1^2+S_2^2)}{2(n-1)}=\frac{(n-1)(S_1^2+S_2^2)}{2(n-1)}=\frac{S_1^2+S_2^2}{2}\tag{5.7.1916}$$

So:

$$S_1^2 \leqslant \mathrm{V}_R \leqslant S_2^2 \qquad (5.7.1917)$$

Now suppose we want to compare with a confidence interval the mean of two populations with different variance to know whether they are different or not.

We currently know two tests to check the averages. The Z-test and T-test. As in the industry (practice) it is rare that we have time to take large samples, let's focus on the second we had proved above:

$$\bar{X} - \frac{S}{\sqrt{n}} T_{\alpha/2}(n-1) \leq \mu \leq \bar{X} + \frac{S}{\sqrt{n}} T_{\alpha/2}(n-1) \qquad (5.7.1918)$$

And also recall that:

$$S_{\bar{X}} = \frac{S}{\sqrt{n}} \qquad (5.7.1919)$$

Now let us recall that we have proved that if we have two random variables of distribution:

$$\mathcal{N}(\mu_1, \sigma_1) \qquad \mathcal{N}(\mu_2, \sigma_2) \qquad (5.7.1920)$$

then the subtraction of the averages gives (property of stability of the average):

$$\mathcal{N}\left(\mu_1 - \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right) \qquad (5.7.1921)$$

So for the difference average of two random variables coming from two population samples we obtain directly:

$$\mathcal{N}\left(\bar{X}_2 - \bar{X}_1, \sqrt{S_{\bar{X}_1}^2 + S_{\bar{X}_2}^2}\right) = \mathcal{N}\left(\bar{X}_2 - \bar{X}_1, \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right) \qquad (5.7.1922)$$

And now the idea is to take the approximation (under the assumption that the variances are equal):

$$\mathcal{N}\left(\bar{X}_2 - \bar{X}_1, \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right) \cong \mathcal{N}\left(\bar{X}_2 - \bar{X}_1, \sqrt{\frac{\mathrm{V}_R}{n_1} + \frac{\mathrm{V}_R}{n_2}}\right) = \mathcal{N}\left(\bar{X}_2 - \bar{X}_1, \sqrt{\mathrm{V}_R\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right) \qquad (5.7.1923)$$

This approximation is named "homoscedastic hypothesis".

We then have the following confidence interval (assuming that we know only an estimation of the variance) remembering that the subtraction or the sum of two independent random variables implies that their variances always add (so it is the same for degrees of freedom of the Student law as we have proved above due to the direct connection with the law of the chi-squared):

$$(\bar{X}_2 - \bar{X}_1) - \sqrt{\mathrm{V}_R\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} T_{\alpha/2}(n-1) \leqslant \mu_2 - \mu_1 \leqslant (\bar{X}_2 - \bar{X}_1) + \sqrt{\mathrm{V}_R\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} T_{\alpha/2}(n-1) \qquad (5.7.1924)$$

with:

$$n=n_1+n_2 \tag{5.7.1925}$$

As the idea in practice is often to test the equality of expected means (and therefore that their difference is zero) from the known estimators then:

$$(\bar{X}_2-\bar{X}_1)-\sqrt{\text{V}_R\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}T_{\alpha/2}(n-2)\leqslant 0\leqslant(\bar{X}_2-\bar{X}_1)+\sqrt{\text{V}_R\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}T_{\alpha/2}(n-2) \tag{5.7.1926}$$

In most softwares on the market, the result is given only from the fact that quantile T that we get is included in the $T_{\alpha/2}$ corresponding to the given confidence interval given by (for reminder):

$$T^{\text{calc}}(n-2)=\frac{(\bar{X}_2-\bar{X}_1)-(\mu_2-\mu_1)}{\sqrt{\text{V}_R\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \tag{5.7.1927}$$

in the case of the homoscedastic hypothesis (equality of variances homogeneity of variances).

Remark

The latter relation is named "independent two-sample T-test", or "homoscedastic T-test" or "T-test of equality of 2 observations expectations with equal variances" or more simply but somewhat abusively "2-sample t-Test" with samples of different sizes and equal variances. Often in the literature, both theoretical means are equal when comparing. It follows that we then have:

$$T^{\text{calc}}(n-2)=\frac{\bar{X}_2-\bar{X}_1}{\sqrt{\text{V}_R\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}=\frac{\bar{X}_2-\bar{X}_1}{S_R\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \tag{5.7.1928}$$

Otherwise, in the most general case of the assumption of heteroscedasticity (not equality of variances), we explicitly write (we'll come back on this later during our study of the Welch test...):

$$\boxed{T^{\text{calc}}(n-2)=\frac{(\bar{X}_2-\bar{X}_1)-(\mu_2-\mu_1)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}} \tag{5.7.1929}$$

Therefore:

$$|T^{\text{calc}}(n-1)|\leqslant|T_{\alpha/2}(n-1)| \tag{5.7.1930}$$

Remark

The ante-previous relation is named "independent two-sample T-test", or "heteroscedasticity T-test" or even "test of equality of expected means: two observations with different variances". If the sizes of the samples are equal and that the variances are also equal and that we assume the two expected mean equal during the comparison, it follows that we then have:

$$T^{\text{calc}}(n-2)=\frac{\bar{X}_2-\bar{X}_1}{S_{1,2}\sqrt{\dfrac{2}{n}}} \tag{5.7.1931}$$

If we need to compare two means differentiated by the index i,j among a number k of multiple means of an ANOVA (as post-hoc test to identify which means are different as the ANOVA omnibus[72] test doesn't tell us anything about this!), the previous relation is then generalized to:

$$T^{\text{calc}}(n-k)=\frac{\bar{X}_i-\bar{X}_j}{\sqrt{\text{V}_R\left(\dfrac{1}{n_i}+\dfrac{1}{n_j}\right)}} \tag{5.7.1932}$$

and rewritten traditionally as following:

$$\text{LSD}=\bar{X}_i-\bar{X}_j=T^{\text{calc}}(n-k)\sqrt{\text{V}_R\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} \tag{5.7.1933}$$

and is named the "least significant difference Fisher test" or simply abbreviated LSD[73]. Obviously the test consist to verify that $\bar{X}_i-\bar{X}_j>\text{LSD}$ at a given α threshold and repeat this procedure for all $k(k-1)/2$ comparisons.

Remark

Remember that we have derived the following two-sample Student T-test:

$$T^{\text{calc}}(n_1+n_2-2)=\frac{\bar{X}_{T1}-\bar{X}_2}{\sqrt{\text{V}_R\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}} \tag{5.7.1934}$$

with the following pooled variance:

$$\text{V}_R=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2} \tag{5.7.1935}$$

In practice it may happens that practitioners use the trimmed mean μ_{T_α} of each sample as the means and the Winsorized sample (unbiased) variances $S_{n[\alpha]}$ as the variances (see section of Numerical Methods page 6560). The test is then named the "trimmed means T-test" (yes the name doesn't highlight the fact that the variances are Winsorized... who the hell knows why...????) and written as:

$$T^{\text{calc}}_{\text{T\&W}}(n_1+n_2-2)=\frac{\bar{X}_{T_\alpha 1}-\bar{X}_{T_\alpha 2}}{\sqrt{\text{V}_{R,W}\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}} \tag{5.7.1936}$$

---

[72] "omnibus" derives from the Latin word "for all". In plain language, you can interpret an omnibus test or omnibus statistic like an "overall test" - it is testing a number of things at once.

[73] Post-hoc tests don't need ANOVA to be run first, except the LSD test (ANOVA is so-named "protector test" to it). Post-hoc by no means mean « post-ANOVA » (those who claim it, even authors of textbooks, have clearly no idea what the tests actually do!). Post hoc (or post factum) means « after the data is collected ».

with the following pooled variance:

$$\mathrm{V}_{R,W}=\frac{(n_1-1)S^2_{n[\alpha],1}+(n_2-1)S^2_{n[\alpha],2}}{n_1+n_2-2} \tag{5.7.1937}$$

This is typically a test from the family of "Robust Null Hypothesis Significance Testing" (RNHST).

This finished, we close now this parenthesis and return to our main topic... We were therefore at the following table if you remember:

$$\begin{aligned}
&Q_T=\sum_{ij}(x_{ij}-\bar{\bar{x}})^2 && \mathrm{df}_T=N-1=kn-1 && \mathrm{V}_T=\frac{Q_T}{\mathrm{df}_T}\\
&Q_A=n\sum_{i}(\bar{x}_i-\bar{\bar{x}})^2 && \mathrm{df}_A=k-1 && \mathrm{V}_A=\frac{Q_A}{\mathrm{df}_A}\\
&Q_R=\sum_{ij}(x_{ij}-\bar{x}_i)^2 && \mathrm{df}_R=k(n-1) && \mathrm{V}_R=\frac{Q_R}{\mathrm{df}_R}
\end{aligned} \tag{5.7.1938}$$

Where we have in the case of samples of the same size:

$$\mathrm{df}_T=\mathrm{df}_A+\mathrm{df}_R=k-1+k(n-1)=k-1+kn-k=kn-1 \tag{5.7.1939}$$

and also the total error that is the sum of the errors of the averages (between classes) and of the residual error (intra-class) and this that the samples are of the same size or not:

$$\boxed{Q_T=Q_A+Q_R} \tag{5.7.1940}$$

Remark

Remember that this sum of square total decomposition is also sometime written:

$$\text{SSTO}=\text{SSTR}+\text{SSE} \tag{5.7.1941}$$

where SSTR stands for "Sum of Square Treatment Residuals" (the equivalent of Q_A) and SSE for "Sum of Square Errors" (the equivalent of Q_R).

Indeed:

$$\begin{aligned}
Q_T&=\sum_{ij}(x_{ij}-\bar{\bar{x}})^2=\sum_i\sum_j\left((x_{ij}-\bar{x}_i)+(\bar{x}_i-\bar{\bar{x}})\right)^2\\
&=\sum_i\sum_j(x_{ij}-\bar{x}_i)^2+\sum_i\sum_j(\bar{x}_i-\bar{\bar{x}})^2+2\sum_i\sum_j(x_{ij}-\bar{x}_i)(\bar{x}_i-\bar{\bar{x}})
\end{aligned} \tag{5.7.1942}$$

But we have:

$$\sum_i\sum_j(x_{ij}-\bar{x}_i)(\bar{x}_i-\bar{\bar{x}})=\sum_i\left[(\bar{x}_i-\bar{\bar{x}})\sum_j(x_{ij}-\bar{x}_i)\right]=0 \tag{5.7.1943}$$

Because:

$$\sum_j(x_{ij}-\bar{x}_i)=0\quad\forall i \tag{5.7.1944}$$

Therefore:

$$Q_T=\sum_i\sum_j\left(x_{ij}-\bar{x}_i\right)^2+\sum_i\sum_j\left(\bar{x}_i-\bar{\bar{x}}\right)^2=Q_R+Q_A \tag{5.7.1945}$$

Now, under the strong assumption (which we will be absolutely required a little further below) that the true variances are linked such as:

$$\sigma_R^2=\sigma_A^2=\sigma_T^2=\sigma^2 \tag{5.7.1946}$$

and therefore that respective estimators are asymptotically equal ... which in practice is approximately true only when certain conditions are met (which is why it is absolutely necessary before making an ANOVA to run an a priori calculation of the power and the sample size of the ANOVA test) we have:

$$\frac{Q_R}{\sigma^2}=\frac{\displaystyle\sum_{ij}\left(x_{ij}-\bar{x}_i\right)^2}{\sigma^2}=\chi^2_{N-k} \tag{5.7.1947}$$

that follows immediately from the proof that we made during our study of statistical inference with the Chi-square law where we got (for refresh!):

$$\chi^2_{n-1}=\frac{(n-1)S^2}{\sigma^2}=\frac{(n-1)\left(\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(X_i-\bar{X})^2\right)}{\sigma^2}=\frac{\displaystyle\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2} \tag{5.7.1948}$$

To determine the number of degrees of freedom of the Chi-square law of:

$$\frac{Q_A}{\sigma^2}=\frac{\displaystyle\sum_i n_i\left(\bar{x}_i-\bar{\bar{x}}\right)^2}{\sigma^2} \tag{5.7.1949}$$

we will use the fact that (by the same reasoning as for the ante-previous relation):

$$\frac{Q_T}{\sigma^2}=\frac{\displaystyle\sum_{ij}\left(x_{ij}-\bar{\bar{x}}\right)^2}{N-1}=\chi^2_{N-1} \tag{5.7.1950}$$

and since that $Q_T=Q_A+Q_R$, then we must have:

$$\frac{Q_T}{\sigma^2}=\frac{Q_A+Q_R}{\sigma^2}=\frac{Q_A}{\sigma^2}+\frac{Q_R}{\sigma^2}=\chi^2_?+\chi^2_{N-k}=\chi^2_{N-1} \tag{5.7.1951}$$

It follows by the linearity property of the Chi-Square law:

$$\frac{Q_A}{\sigma^2}=\frac{\displaystyle\sum_i n_i\left(\bar{x}_i-\bar{\bar{x}}\right)^2}{\sigma^2}=\chi^2_{k-1} \tag{5.7.1952}$$

So to summarize we have:

$$\boxed{\frac{Q_A}{\sigma^2}=\chi^2_{k-1}\quad\text{and}\quad\frac{Q_R}{\sigma^2}=\chi^2_{N-k}} \tag{5.7.1953}$$

Now comes the Fisher law in the assumption that the variances are equal (and measurement Normally distributed)! Because:

$$\frac{(k-1)S_A^2}{\sigma^2}=\chi^2_{k-1} \quad \text{and} \quad \frac{(N-k)S_R^2}{\sigma^2}=\chi^2_{N-k} \tag{5.7.1954}$$

What we want to do is to see if there is a difference between the variance of the averages (between classes) and the residual variance (intra-class). To compare two variances when the true averages are unknown we saw that the best was to use the Fisher test. But, we proved in our study of the Fisher law a little above that:

$$\frac{\sigma_X^2}{\sigma_Y^2}=\frac{1}{F_{n-1,m-1(x)}}\frac{S_X^2}{S_Y^2} \tag{5.7.1955}$$

where in our case study:

$$\frac{\sigma_A^2}{\sigma_R^2}=1 \Rightarrow F_{k-1,N-k}=\frac{S_A^2}{S_R^2}=\frac{\mathrm{V}_A}{\mathrm{V}_R}=\frac{\dfrac{\displaystyle\sum_i n_i(\bar{x}_i-\bar{\bar{x}})^2}{k-1}}{\dfrac{\displaystyle\sum_{ij}(x_{ij}-\bar{x}_i)^2}{N-k}}=\frac{\dfrac{Q_A}{k-1}}{\dfrac{Q_R}{N-k}} \tag{5.7.1956}$$

Since there are dozens of different types of ANOVA we have to understand very good this choice of the simplest ANOVA we are studying right now! Thus, if the averages are the same, the null hypothesis is then that variance ratio is equal to unity (under the above assumptions already mentioned far above). If F is too large at a given threshold, then we reject the null hypothesis of equality of means (because verbatim variances will be strongly different). So it seems here logic to compare the variances between groups (numerator) with that in the groups (denominator) but as we shall see this is not always the choice to be made (especially in hierarchical ANOVA)!

Given the assumption of the first equality of the above relation (the one that precedes the implication), we understand at the same time much better the great sensitivity of the ANOVA results to the non equality of true variances!

Let us also indicate that the previous relation:

$$\boxed{F_{k-1,N-k}=\frac{\dfrac{Q_A}{k-1}}{\dfrac{Q_R}{N-k}}=\frac{\dfrac{\text{SSB}}{k-1}}{\dfrac{\text{SSW}}{N-k}}} \tag{5.7.1957}$$

is often given in the literature as follows:

$$F_{k-1,N-k}=\frac{\text{MSk}}{\text{MSE}}=\frac{\text{variance between treatments}}{\text{variance within treatments}}=\frac{\text{signal}}{\text{noise}} \tag{5.7.1958}$$

where MSk is named "Mean Square for treatments" and MSE "Mean Square for Error". This ratio will therefore give us the value of the random variable F (whose support is for reminder bounded at zero). Regarding the choice of the test (right/left one-sided or bilateral), note that if the averages are really equal, then for all i:

$$(\bar{x}_i-\bar{\bar{x}})=0 \tag{5.7.1959}$$

So in this case:

$$F_{k-1,N-k}=\frac{\dfrac{\sum\limits_{i}n_i(\bar{x}_i-\bar{\bar{x}})^2}{k-1}}{\dfrac{\sum\limits_{ij}(x_{ij}-\bar{x}_i)^2}{N-k}}=\frac{\dfrac{\sum\limits_{i}n_i\cdot 0}{k-1}}{\dfrac{\sum\limits_{ij}(x_{ij}-\bar{x}_i)^2}{N-k}}=0 \tag{5.7.1960}$$

Which brings us of course to immediately adopt a right-sided test!

Otherwise, in general, the interpretation of this fraction is basically as following: This is the ratio (normalized to the number of degrees of freedom) of the sum of the error of the averages (between classes) and that of the residual error (intra-class) or in other words the ratio of the interclass variance by the residual variance. The ratio thus follows a Fisher distribution with two parameters given by the degrees of freedom of the respective classes.

Remark

If there are only two populations (samples), we must understand that then that use of the Student T-test is more than enough and is perfectly equivalent! In fact, the ANOVA is an indirect comparison of means, the Student T-test a direct comparison... so it is obvious to guess which one is better in this particular situation!

All calculations we have made until now are very often represented in softwares in a standard table form as represented below (this is for example how Microsoft® Excel® 11.8346 or Minitab 15.1.1 give the results):

Source	**Sum of squares (SSE)**	χ^2 **df**	**Mean squares**	F	**Critical** F
Inter-Class	$Q_A=\sum\limits_i n_i\left(\bar{x}_i-\bar{\bar{x}}\right)^2$	$k-1$	$\text{MSk}=\dfrac{Q_A}{k-1}$	$\dfrac{\text{MSk}}{\text{MSE}}$	$P(F>F_{k-1,N-k})$
Intra-Class	$Q_R=\sum\limits_{ij}(x_{ij}-\bar{x}_i)^2$	$N-k$	$\text{MSE}=\dfrac{Q_R}{N-k}$		
Total	$Q_T=\sum\limits_{ij}(x_{ij}-\bar{\bar{x}})^2$	$N-1$			

Table 5.54: **One-way fixed factor ANOVA table**

The null hypothesis will not be rejected if the value of:

$$F_{k-1,N-k}=\frac{\text{MSk}}{\text{MSE}} \tag{5.7.1961}$$

is smaller or equal to the quantile of the distribution F that corresponds to the cumulative probability to 1 subtracted from confidence level α and denoted by F_c.

The p-value in the ANOVA table and the multiple comparison results like Dunnett's test (see further below) and Tukey's test (see also further below page 1207) are based on different methodologies and can occasionally produce contradictory results. For example, it is possible that the ANOVA p-value can indicate that there are no differences between the means while the multiple comparisons output indicates that some means that are different. In this case, you can generally trust the multiple comparisons output.

It must be remembered that to use ANOVA, we must therefore assume that the samples come from different populations (unpaired data) and follow a Normal law (even if the ANOVA is relatively robust to non-normality). It is therefore necessary to check the normality of the distributions and the homoscedasticity (Levene test for example). Otherwise, nonparametric variants of the analysis of variance must be used (Kruskal-Wallis ANOVA or Friedman's ANOVA, for example among others that we will see later).

Remarks

R1. Notice that in practice, inter-class variance is often named "inter-laboratory variance" and the intra-class variance is verbatim often named "intra-laboratory variance".

R2. There are, in this early 121st century (holocene calendar) more than 50 existing tests or procedures comparison for the variance. The opinion varies among practitioners relatively to their relevance and the effectiveness of homogeneity of variance tests (HVT). Some argue that they are absolutely necessary to be applied before doing any ANOVA, others say that these tests are anyway of poor performance, the ANOVA being anyway more robust to homoscedasticity than the differences than can be detected by HVT, particularly in case of non-Normality. In fact, all these issues are related to the "Behrens-Fisher problem", which is that the comparison of means without assuming the equivariance. However among the fifty existing tests, several comparative studies have highlighted the following efficient tests that we will present further below: Bartlett's test, Levene and Brown-Forsythe.

R3. When certain levels of a factor are combined into one to be compared to a reference level statisticians then talk about creating "contrasts" (see just below).

The distribution of $\bar{x}_i$ is $\mathcal{N}(\mu_i,\sigma_\varepsilon^2/n)$. Therefore as we have already seen before we can write:

$$\bar{x}_i - T_{\alpha/2}(N-k)\frac{\sigma_\varepsilon}{\sqrt{n}} \leq \mu_i \leq \bar{x}_i + T_{\alpha/2}(N-k)\frac{\sigma_\varepsilon}{\sqrt{n}} \tag{5.7.1962}$$

Hence:

$$\bar{x}_i - T_{\alpha/2}(N-k)\frac{\text{MSE}}{\sqrt{n}} \leq \mu_i \leq \bar{x}_i + T_{\alpha/2}(N-k)\frac{\text{MSE}}{\sqrt{n}} \tag{5.7.1963}$$

A confidence interval for the difference of two means is then as usual:

$$(\bar{x}_i-\bar{x}_j) - T_{\alpha/2}(N-k)\sqrt{\frac{2\text{MSE}}{n}} \leq \mu_i-\mu_j \leq (\bar{x}_i-\bar{x}_j) + T_{\alpha/2}(N-k)\sqrt{\frac{2\text{MSE}}{n}} \tag{5.7.1964}$$

5.7.13.5.1.1 Welch Test

During our one-way fixed factor ANOVA study above, we had introduced the fact that it was possible to compare the average of two populations whose data followed a Normal distribution but of unequal variances using the heteroscedastic T-test Student:

$$T^{\text{calc}}(n-2)=\frac{(\bar{X}_2-\bar{X}_1)-(\mu_2-\mu_1)}{\sqrt{\dfrac{S_1^2}{n_1}+\dfrac{S_2^2}{n_2}}} \tag{5.7.1965}$$

with:

$$n=n_1+n_2 \tag{5.7.1966}$$

But in reality we had a little passed into silence ... (some will have perhaps noticed it) that we are not authorized to make this last addition of the degrees of freedom as soon as the variances are not equal! Indeed, let us go back to the source of the Student distribution:

$$T(n-1)=\frac{Z}{\dfrac{U}{n-1}} \tag{5.7.1967}$$

where for recall, the variable Z follows a reduced Normal centered law and the random variable U follows a χ^2 law with $n-1$ degrees of freedom. And remember that we have proved that:

$$\chi^2(n-1)=\frac{(n-1)S^2}{\sigma^2} \tag{5.7.1968}$$

But now, let us consider that this same random variable U is given by (we work as a reminder with the standard deviation of the mean!):

$$\begin{aligned}U&=\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}=\frac{\dfrac{\sigma_1^2}{n_1-1}\chi^2(n_1-1)}{n_1}+\frac{\dfrac{\sigma_2^2}{n_2-1}\chi^2(n_2-1)}{n_2}=\frac{\dfrac{\sigma_1^2}{f_1}\chi^2_{n_1-1}}{n_1}+\frac{\dfrac{\sigma_2^2}{f_2}\chi^2_{n_2-1}}{n_2}\\&=\frac{\sigma_1^2}{n_1f_1}\chi^2_{n_1-1}+\frac{\sigma_2^2}{n_2f_2}\chi^2_{n_2-1}=aX+bY\end{aligned} \tag{5.7.1969}$$

where X and Y are two independent random variables following χ^2 laws with respectively f_1 and f_2 degrees of freedom such that:

$$X=\chi^2(f_1)\qquad\text{and}\qquad Y=\chi^2(f_2) \tag{5.7.1970}$$

where we impose that a and b are positive because otherwise, as proved earlier above, aX and bY would not follow gamma laws.

Let us recall then that (see page 773):

$$\text{E}(X)=f_1\qquad\text{and}\qquad\text{V}(X)=2f_1 \tag{5.7.1971}$$

We then have:

$$aX=\gamma\left(\frac{f_1}{2},\frac{1}{2a}\right)\qquad\text{and}\qquad bY=\gamma\left(\frac{f_2}{2},\frac{1}{2b}\right) \tag{5.7.1972}$$

However $U=aX+bY$ does not follow a gamma law if $a\neq b$ (as we demonstrated during our study of the Gamma law at page 767). Assuming, however, that they are close enough, numerical simulations have shown that we can approximate the law of U by the following gamma law:

$$\gamma\left(\frac{f}{2},\frac{1}{2g}\right) \tag{5.7.1973}$$

which has therefore for expected mean and variance (see page 767):

$$\mu=\frac{\dfrac{f}{2}}{\dfrac{1}{2g}}=fg \qquad \sigma^2=\frac{\dfrac{f}{2}}{\left(\dfrac{1}{2g}\right)^2}=2fg^2 \tag{5.7.1974}$$

and we will determine the positive numbers f and g so that the expected means and the variances coincide with those of U.

We therefore have by the independence of the two random variables:

$$\mathrm{E}(U)=af_1+bf_2 \qquad \text{and} \qquad \mathrm{V}(U)=2(a^2f_1+b^2f_2) \tag{5.7.1975}$$

So that the expectation and the variance of our approximation correspond to that of U, we must have:

$$fg=af_1+bf_2 \qquad \text{and} \qquad 2fg^2=2(a^2f_1+b^2f_2) \tag{5.7.1976}$$

By dividing the second equation by the first the f disappears and we find:

$$g=\frac{a^2f_1+b^2f_2}{af_1+bf_2} \tag{5.7.1977}$$

Replacing in the first we find what it is customary to call the "Welch-Satterthwaite equation":

$$f=\frac{(af_1+bf_2)^2}{af_1+bf_2} \tag{5.7.1978}$$

Thus explicitly:

$$f=\frac{\left(\left(\dfrac{\sigma_1^2}{n_1f_1}\right)f_1+\left(\dfrac{\sigma_2^2}{n_2f_2}\right)f_2\right)^2}{\left(\dfrac{\sigma_1^2}{n_1f_1}\right)^2f_1+\left(\dfrac{\sigma_2^2}{n_2f_2}\right)^2f_2}=\frac{\left(\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}\right)^2}{\dfrac{(\sigma_1^2/n_1)^2}{f_1}+\dfrac{(\sigma_2^2/n_2)^2}{f_2}}=\frac{\left(\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}\right)^2}{\dfrac{(\sigma_1^2/n_1)^2}{n_1-1}+\dfrac{(\sigma_2^2/n_2)^2}{n_2-1}} \tag{5.7.1979}$$

As we do not know in practice the variances (and therefore the true standard deviations), we actually take:

$$\boxed{f=\frac{\left(\dfrac{S_1^2}{n_1}+\dfrac{S_2^2}{n_2}\right)^2}{\dfrac{(S_1^2/n_1)^2}{n_1-1}+\dfrac{(S_2^2/n_2)^2}{n_2-1}}} \tag{5.7.1980}$$

So the expression above rounded to the nearest integer is the number of degrees of freedom that must actually be taken for the Student heteroscedastic test:

$$T^{\text{calc}}([f])=\frac{(\bar{X}_2-\bar{X}_1)-(\mu_2-\mu_1)}{\sqrt{\dfrac{S_1^2}{n_1}+\dfrac{S_2^2}{n_2}}} \qquad (5.7.1981)$$

Finally, let us indicate that in the general case with n terms, the Welch-Satterthwaite equation is written (without proof):

$$f_n=\frac{\left(\displaystyle\sum_{i=1}^{n}\frac{S_i}{n_i}\right)^2}{\displaystyle\sum_{i=1}^{n}\frac{(S_i^2/n_i)^2}{n_i-1}} \qquad (5.7.1982)$$

Remark

For the special case $n=2$ the latter relation can often be found in the literature in the following form (...):

$$\nu=\frac{\left(\frac{s_1^2}{N_1}+\frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2\cdot\nu_1}+\frac{s_2^4}{N_2^2\cdot\nu_2}}=\frac{\left(\frac{s_1^2}{N_1}+\frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2\cdot(N_1-1)}+\frac{s_2^4}{N_2^2\cdot(N_2-1)}} \qquad (5.7.1983)$$

In the field of finance when using the CAPM model (see section Economy page 7740) or when using the ANCOVA (see section Statistics page ??) and not only that ... it is useful and important to compare the regression line slopes to find out if they are significantly different or not!

We have proved in the section of Numerical Methods (page 6348) that for a linear function whose parameters were estimated and which was expressed in the form:

$$\hat{Y}=Ax+B \qquad (5.7.1984)$$

and which exact unknown expression is:

$$Y=ax+b \qquad (5.7.1985)$$

we had:

$$A-T_{\alpha/2}(n-2)\sqrt{\frac{\frac{1}{n-2}\sum e_k^2}{n\mathbb{V}(x)}}\leq a\leq A+T_{\alpha/2}(n-2)\sqrt{\frac{\frac{1}{n-2}\sum e_k^2}{n\mathbb{V}(x)}} \qquad (5.7.1986)$$

and that if the real variance is unknown:

$$A-T_{\alpha/2}(n-2)\sqrt{\frac{\frac{1}{n-2}\sum e_k^2}{(n-1)\hat{\mathbb{V}}(x)}}\leq a\leq A+T_{\alpha/2}(n-2)\sqrt{\frac{\frac{1}{n-2}\sum e_k^2}{(n-1)\hat{\mathbb{V}}}(x)} \qquad (5.7.1987)$$

or sometimes also written:

$$A-T_{\alpha/2}(n-2)\sqrt{\frac{\text{SSE}}{n\text{V}(x)}}\leq a\leq A+T_{\alpha/2}(n-2)\sqrt{\frac{\text{SSE}}{n\text{V}(x)}} \tag{5.7.1988}$$

That has the same shape as the Student confidence interval test for the expected mean!

This being recalled, the objective now is to test if two slopes of two different regressions are equal (very important test for the study of the ANCOVA!), and then we write that the null hypothesis in the context of a hypothesis test obviously as following:

$$H_0:A_1=A_2\Leftrightarrow A_2-A_2=\Delta A=0 \tag{5.7.1989}$$

Since the confidence interval of a single slope of a regression is similar in all aspects to a confidence interval of the Student's mean, the difference of two slopes is then similar in all aspects to the Welch hypothesis test. We then have since we make the null hypothesis that the two true slopes have a null difference (i.e. $a_1=a_2$):

$$T^{\text{calc}}([f])=\frac{(A_2-A_1)-(a_2-a_1)}{\sqrt{\dfrac{\text{SSE}_1}{n_1\text{V}(x_1)}+\dfrac{\text{SSE}_2}{n_2\text{V}(x_2)}}}=\frac{A_2-A_1}{\sqrt{\dfrac{\text{SSE}_1}{n_1\text{V}(x_1)}+\dfrac{\text{SSE}_2}{n_2\text{V}(x_2)}}} \tag{5.7.1990}$$

with:

$$f\cong\frac{\left(\displaystyle\sum_{i=1}^{2}\frac{\dfrac{1}{n_i-2}\sum e_{i,k}^2}{n_i\text{V}(x_i)}\right)^2}{\displaystyle\sum_{i=1}^{2}\frac{\left(\dfrac{\dfrac{1}{n_i-2}\sum e_{i,k}^2}{n_i\text{V}(x_i)}\right)}{n_i-1}} \tag{5.7.1991}$$

Which is often written by tradition:

$$T^{\text{calc}}(f\cong n_1+n_2-4)\cong\frac{\hat{\beta}_2-\hat{\beta}_1}{S_{\hat{\beta}_1-\hat{\beta}_2}} \tag{5.7.1992}$$

Do not forget (very important case in the application of the ANCOVA) that if we work with the estimators, then $n_i\text{V}(x_i)$ become $(n_i-1)\hat{\text{V}}(x_i)$.

Remark

The variants of the T-test are: the Welch test that accounts for unequal group sizes and variance, the Yuen-Welch test that accounts for unequal variances and outliers, the permutation Welch/Yuen-Welch tests that accounts for non-normality and at the same time for unequal variances and outliers!

5.7.13.5.1.2 Contrasts

The F-test is rather unspecific. It basically gives us a "Yes/No" answer to the question: Is there any treatment effect at all? It does not tell us what specific treatment or treatment combination is special. Quite often, we have a more specific question than the aforementioned global null hypothesis. For example, we might want to compare a set of new treatments vs. a control treatment or we want to do pairwise comparisons between many (or all) treatments.

Many multiple comparison methods use the idea of what we name a "contrast". What is this? To introduce this concept let us consider a one-way fixed factor ANOVA where the null hypothesis H_0 was rejected but where actually we don't know what level of the variable cause the difference that brig us to reject the null hypothesis.

Consider for example that among the levels i, j, k, ... we suspect the levels i and k to differ. Therefore we would like to test the hypothesis:

$$\begin{gathered} H_0 : \mu_i = \mu_k \\ H_1 : \mu_i \neq \mu_k \end{gathered} \tag{5.7.1993}$$

or equivalently:

$$\begin{gathered} H_0 : \mu_i - \mu_k = 0 \\ H_1 : \mu_i - \mu_k \neq 0 \end{gathered} \tag{5.7.1994}$$

If we had suspected at the start of the experiment that the average of the two first level (considering an imaginary experiment with a factor having 5 fixed levels) did not differ from the average of the two highest levels, then the hypothesis would have been:

$$\begin{gathered} H_0 : \mu_1 + \mu_2 = \mu_3 + \mu_4 \\ H_1 : \mu_1 + \mu_2 \neq \mu_3 + \mu_4 \end{gathered} \tag{5.7.1995}$$

or:

$$\begin{gathered} H_0 : \mu_1 + \mu_2 - \mu_3 - \mu_4 = 0 \\ H_1 : \mu_1 + \mu_2 - \mu_3 - \mu_4 \neq 0 \end{gathered} \tag{5.7.1996}$$

In general a "contrast" is a linear combination of parameters of the form:

$$\boxed{C = \sum_{i=1}^{k} c_i \mu_i \quad \text{where} \quad \sum_i c_i = 0, c_i = \{\pm 1, 0\}} \tag{5.7.1997}$$

Remark

Two contrasts C_1, C_2 with coefficients $\{c_i\}$ and $\{d_i\}$, i.e.:

$$C_1 = \sum_{i=1}^{k} c_i \mu_i \qquad C_2 = \sum_{i=1}^{n} d_i \mu_i \tag{5.7.1998}$$

are said to be "orthogonal contrasts" if and only if $\sum c_i d_i = 0$.

For example, $C_1 = \mu - 2\mu_2 + \mu_3$ and $C_2 = \mu_1 - \mu_2$ are orthogonal contrast since $(1)(1) + (-2)(0) + (1)(-1) = 0$.

Both of the hypothesis above can then be expressed in terms of contrasts:

$$\boxed{\begin{aligned} H_0 &: \sum_i^k c_i\mu_i = 0 \\ H_1 &: \sum_i^k c_i\mu_i \neq 0 \end{aligned}} \tag{5.7.1999}$$

Therefore in the case of our previous example:

$$\begin{aligned} H_0 &: \mu_1 + \mu_2 - \mu_3 - \mu_4 = 0 \\ H_1 &: \mu_1 + \mu_2 - \mu_3 - \mu_4 \neq 0 \end{aligned} \tag{5.7.2000}$$

we have $c_1 = +1$, $c_2 = +1$, $c_3 = -1$, $c_4 = -1$, $c_5 = 0$.

Testing hypotheses involving contrasts can be done in three ways. We will introduce here only the one that gives the practitioner the possibility to build a confidence interval. This method uses a Student T-test. Indeed, let us write the contrast in terms of estimates:

$$\hat{C} = \sum_{i=1}^k c_i\bar{y}_{i.} \tag{5.7.2001}$$

as under the null hypothesis we should have (don't forget it!):

$$\hat{C} = \sum_{i=1}^k c_i\bar{y}_{i.} = 0 \tag{5.7.2002}$$

The variance of the sum can be written:

$$\text{V}(\hat{C}) = \text{V}\left(\sum_{i=1}^k c_i\bar{y}_i\right) = \sum_{i=1}^k c_i^2\text{V}(\bar{y}_i) \tag{5.7.2003}$$

If the equality of variance assumption is satisfied and the design is equilibrated we have:

$$\text{V}(\mu_i) = \frac{\sigma^2}{n} \tag{5.7.2004}$$

But with estimated we know that this is written:

$$\text{V}(\bar{y}_i) = \frac{S^2}{n} \tag{5.7.2005}$$

and therefore:

$$\text{V}(\hat{C}) = \sum_{i=1}^k c_i^2\text{V}(\bar{y}_i) = \sum_{i=1}^k c_i^2\frac{S^2}{n} = \frac{S^2}{n}\sum_{i=1}^k c_i \tag{5.7.2006}$$

Hence:

$$\frac{C-\mu}{\sigma_C}=\frac{C}{\sqrt{\text{V}(C)}} \tag{5.7.2007}$$

should follow a centered Normal law under the above assumptions:

$$\frac{C}{\sqrt{\text{V}(C)}}=\mathcal{N}(0,1) \tag{5.7.2008}$$

But as we work with estimators we know that we have instead:

$$T(N-k)=\frac{\hat{C}}{\sqrt{\text{V}(\hat{C})}}=\frac{\displaystyle\sum_{i=1}^{k}c_i\bar{y}_i}{\sqrt{\dfrac{S^2}{n}\displaystyle\sum_{i=1}^{k}c_i}} \tag{5.7.2009}$$

That is sometimes denoted in the field of analysis of variance:

$$T(N-k)=\frac{\displaystyle\sum_{i=1}^{k}c_i\bar{y}_i}{\sqrt{\dfrac{\text{MSE}}{n}\displaystyle\sum_{i=1}^{k}c_i}} \tag{5.7.2010}$$

Therefore the $100\cdot(1-\alpha)$ percent confidence interval on the contrast $\sum_{i=1}^{k}c_i\mu_i$ is:

$$\boxed{\sum_{i=1}^{k}c_i\bar{y}_i-T_{\alpha/2}(N-k)\sqrt{\frac{S^2}{n}\sum_{i=1}^{k}c_i}\leq\sum_{i=1}^{k}c_i\mu_i\leq\sum_{i=1}^{k}c_i\bar{y}_i+T_{\alpha/2}(N-k)\sqrt{\frac{S^2}{n}\sum_{i=1}^{k}c_i}} \tag{5.7.2011}$$

Clearly if this confidence interval includes zero, we would be unable to reject the null hypothesis.

Remark

Keep in mind that if you want to test for example in the framework of an ANOVA that $H_0:\ \mu_2=(\mu_1+\mu_3)/2$, it is therefore equivalent to write and test the following contrast $\sum c_i\mu_i=\mu_1-2\mu_2+\mu_3=\bar{y}_1-2\bar{y}_2+\bar{y}_3$.

5.7.13.5.2 ANOVA with two fixed factors without repetitions

We will now see the concept of interaction that is essential to understand what is behind the fixed two-factor ANOVA, also named "ANOVA with two fixed categorical variables" or "two-way ANOVA" or "randomized complete block design[74]", without and especially later with (!) repetitions. Indeed, it is only with two-factor ANOVA with repetitions - by mathematical construction - that can be statistically (under given assumptions) consider objectively if two or more factors interact significantly together.

We must therefore, before moving to the pure mathematical part, introduce some concepts!

Definitions (#5.7.149):

D1. We say that there is "no interaction" when the average of responses by a factor based on its level varies by the same amount and with the same sign as the average of the responses of another factor depending on its levels. Then we say that: the response interaction curves in the diagram are parallel.

> Remark
>
> The parallelism of response curves is normal in the situation of no interaction because it means that whatever the level of one or the other factors, the variation (if there is any) of the response will always be the same amplitude. What is characteristic of independence (at least locally).

D2. We say that two factors are "in interaction" when the average responses by a factor depending on its levels do not change by the same magnitude and / or not with the same sign as the average of the responses of another factor according to their levels. Then we say that: the response curves in the interactions diagram are not parallel.

> Remark
>
> The absence of interaction is a very strong assumption and a rarely observed. Often we have interactions or strong interactions.

To understand the concept, we will use small examples have no repeated measures to such we will have a simple qualitative idea of the phenomenon, but by no means a scientific approach of the concept of interaction (that will be study later in details).

For every small example below we visualize the situations by means of two types of representations: a graph illustrating the main effects on the one hand and a pattern of the interactions on the other hand.

---

[74] As already mentioned, a "block" can be considered as a factor and hence the mathematical treatment is exactly the same!

Consider the following small table with two factors at two levels ("explanatory variables") including four cells ("variable of interest"):

	Factor 2	
Factor 1	Level 1	Level 2
Level 1	3	3
Level 2	3	3

Table 5.55: **First example of 2-factor ANOVA without repetition**

We will get as result with software such as Minitab:

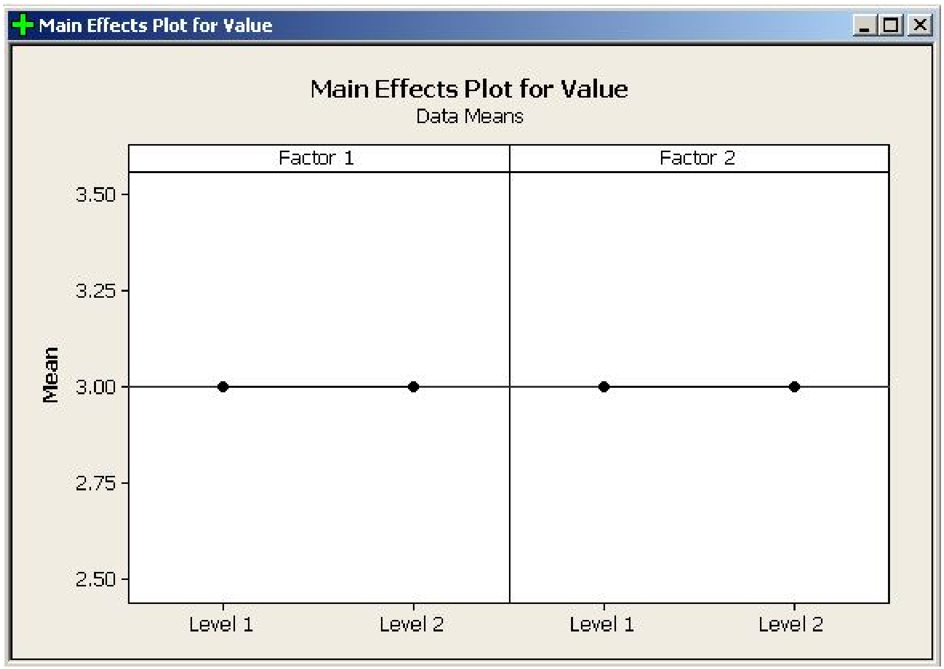

Main Effects Plot for Value
Main Effects Plot for Value
Data Means
Factor 1
Factor 2
3.50
3.25
3.00
2.75
2.50
Mean
Level 1
Level 2
Level 1
Level 2

Figure 5.189: **Main effects plot without interactions in Minitab 15**

We see well that no factor has a major effect on anything. What is relatively obvious given the content of the previous table...

The diagram of interactions (often named "profiler" in the industry) gives:

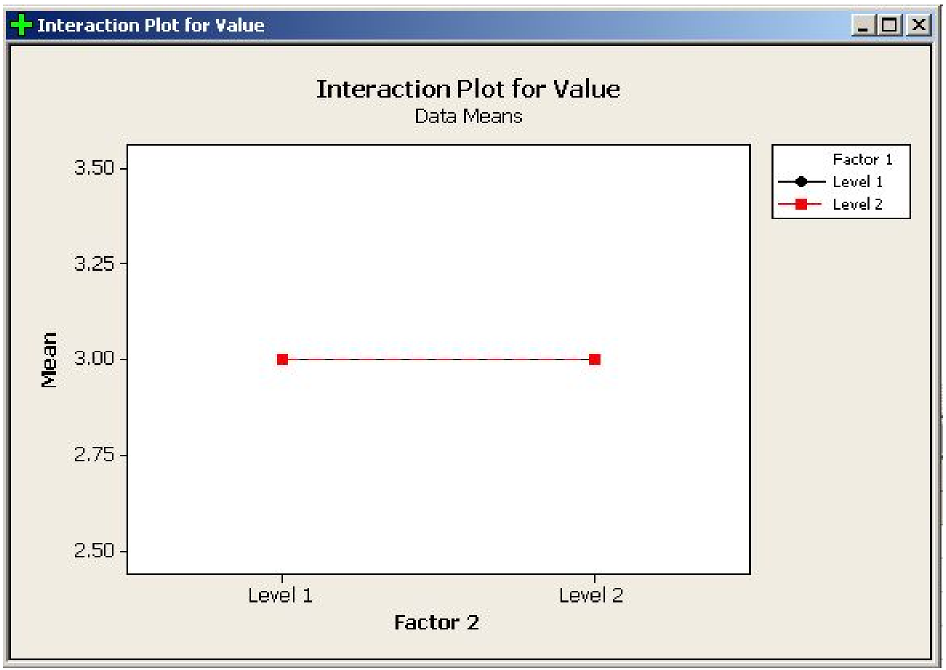


Figure 5.190: **Interactions plot without interactions in Minitab 15**

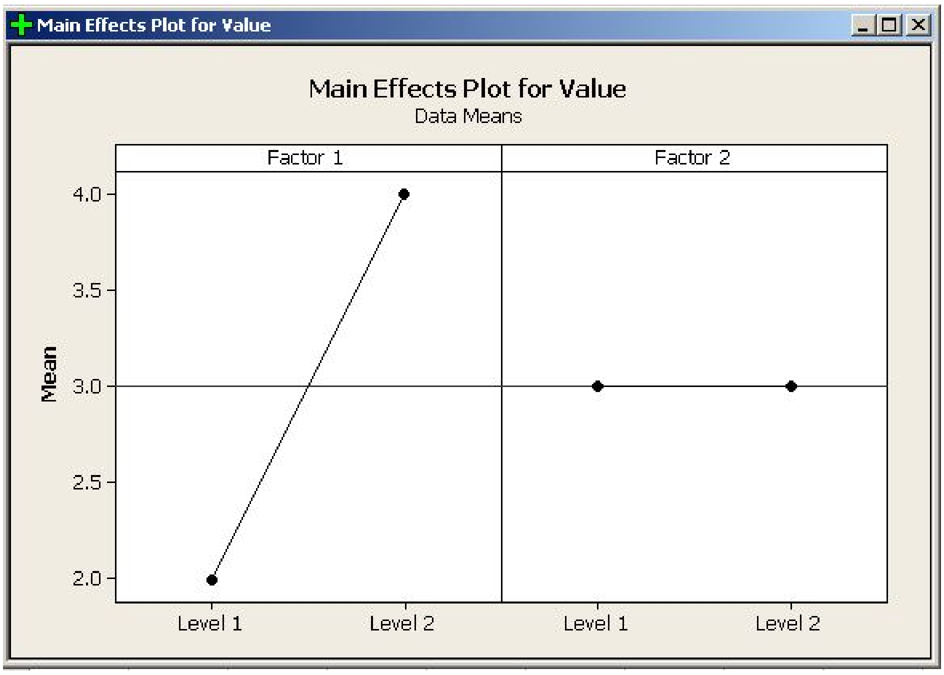
where we can see that the factors do not interact between them (or neutralize themselves it depends...). Then we say that there is "(a priori) no effect or no interaction (locally)". In fact in some experiments the absence of interaction is a very strong assumption and therefore very rare. This is why we must pay attention to the chosen words when interpreting interaction plots (because do not to go through pure calculations is dangerous for this step or simply not scientific!).

Now consider the following table:

	Factor 2	
Factor 1	Level 1	Level 2
Level 1	2	2
Level 2	4	4

Table 5.56: **Second example of 2-factor ANOVA without repetition**

It seems obvious that the factor one through considered with its levels seems to have an influence on the answer. But let us see the different representations:



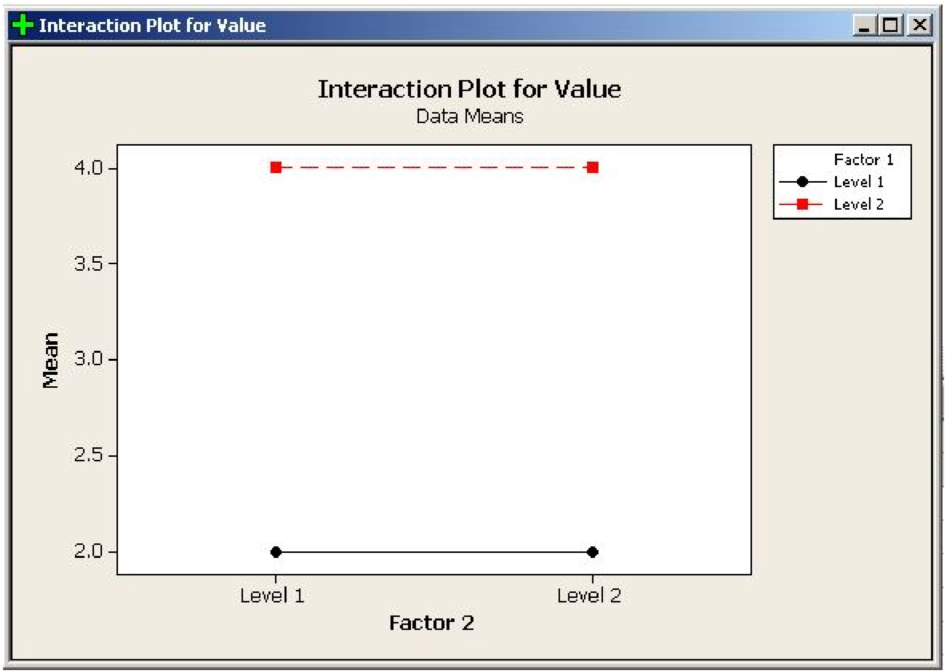


Figure 5.191: **Main effects and interactions plot with interactions in Minitab 15**

Let us comment the first graph as proposed by a reader:

This includes two parts: the left one analyses the effects of Factor 1 through its two levels; that of the right do the same for the Factor 2.

Have a closer look to the left part:

We see two points connected by a line segment. Here the first point, the one for the level 11, is located in the ordinate 2 while the second point, that for level 2, is located in the ordinate 4. Now remember that each point represents an average. And the ordinate of the first point is located at the average of $(2+2)/2=2$.

Having said that and hoping this helped to a better understanding, let us continue...

It is quite clear in the chart above that only the level of Factor 1 influences the answer, while Factor 2 has no effect on the answer. Then we say that: there is a main effect (locally) of the Factor 1.

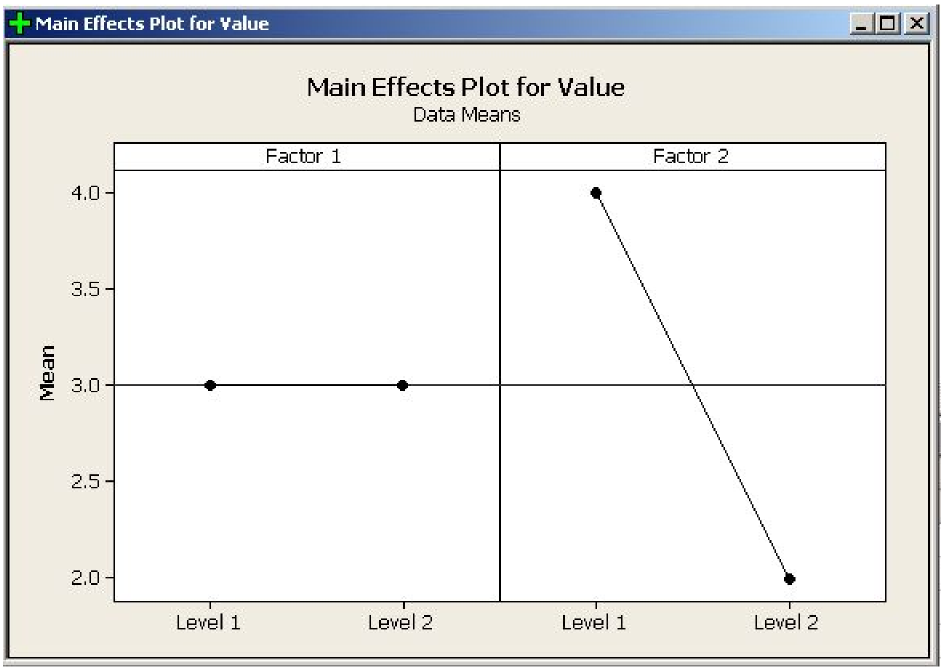
On the diagram of interactions, we have the same information but in a different form. We see that regardless of the level of Factor 2, the answers are horizontal and thus that it does not influence the results. We are then in a situation where "(a priori) the main effect is (locally) Factor 1 and there is an absence of interactions between factors".

Now consider the following table:

	Factor 2	
Factor 1	Level 1	Level 2
Level 1	4	2
Level 2	4	2

Table 5.57: **Third example of 2-factor ANOVA without repetition**

This time we can observe that the Factor 2 has an influence but not the Factor 1. But let us also see this with our two types of representations:


Main Effects Plot for Value
Main Effects Plot for Value
Data Means
Factor 1
Factor 2
Mean
4.0
3.5
3.0
2.5
2.0
Level 1
Level 2
Level 1
Level 2

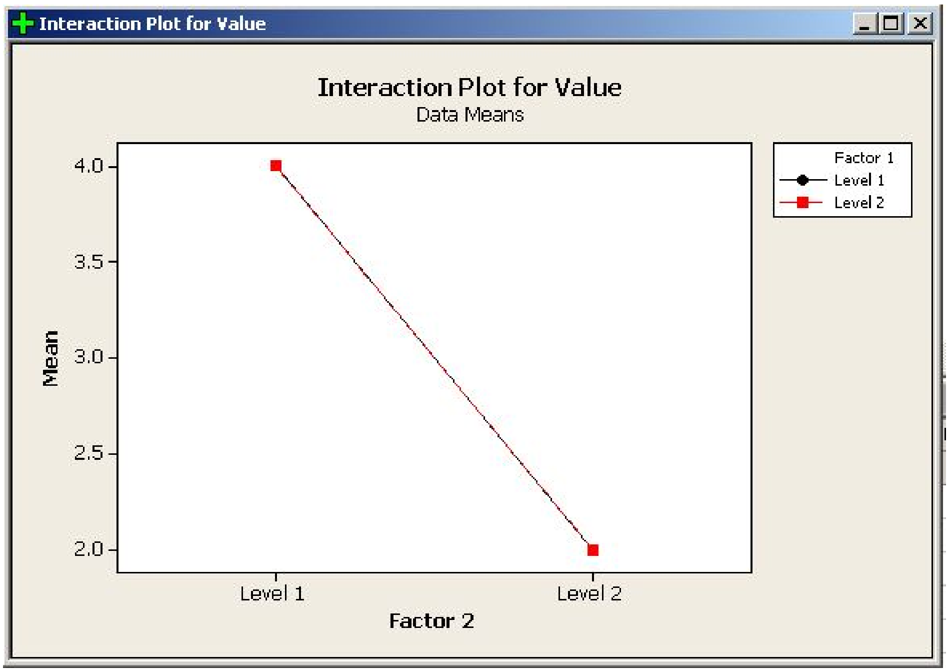


Figure 5.192: **Main effects and interactions plot with interactions in Minitab 15**

We see well on the top diagram that the Factor 1 has no influence. On the bottom diagram it is less obvious but the superposition of the two lines indicates that the Factor 1 has no influence. Then we say that there is: "(a priori) main effect (locally) of Factor 2 and absence of interactions between factors".

Now consider the following table:

	Factor 2	
Factor 1	Level 1	Level 2
Level 1	3	1
Level 2	5	3

Table 5.58: **Fourth example of 2-factor ANOVA without repetition**

We see that both factors have an influence on the answer. We can clearly see it on the both diagrams below:

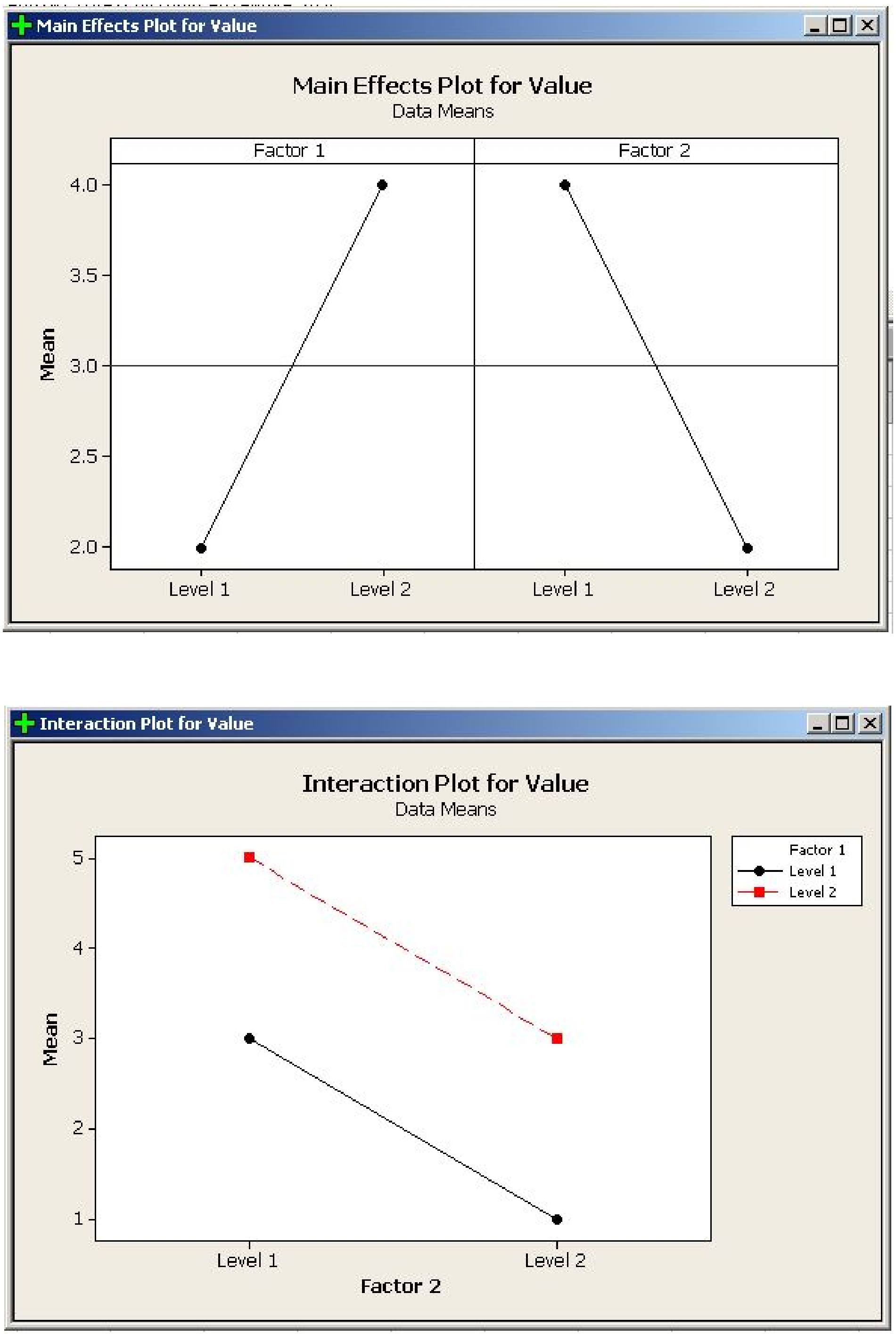


Figure 5.193: **Main effects and interactions plot with interactions in Minitab 15**

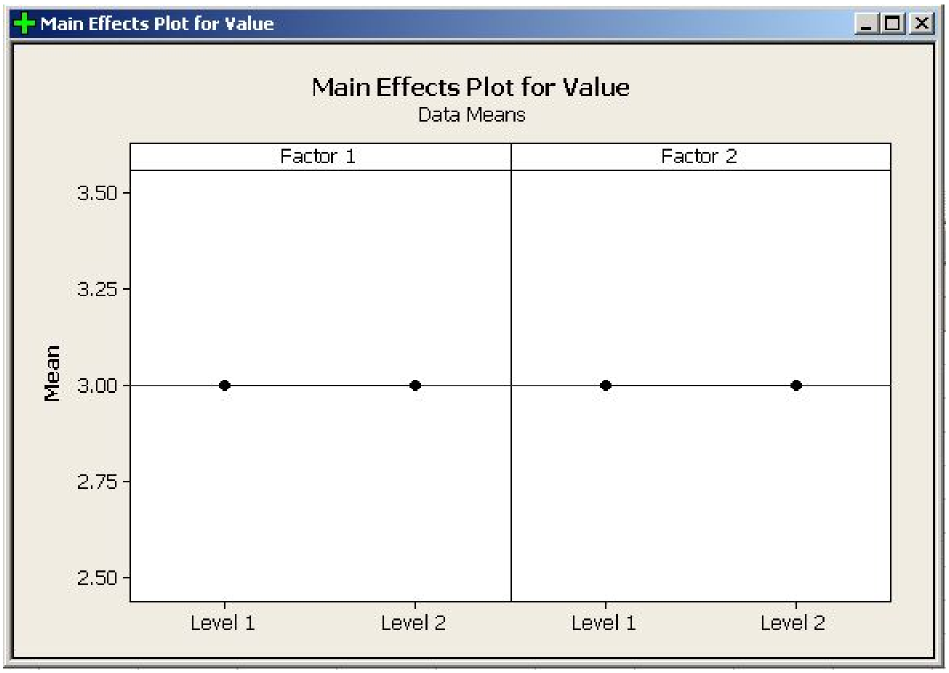
We can see very well on the top diagram that the Factor 1 has an influence on the answer and that we have the same conclusion for the Factor 2 (and furthermore to in the same magnitude regardless of the direction!). On the bottom diagram it is less obvious, but the same conclusion is valid. We then say that: "(a priori) both factors are (locally) significant and without interactions".

Now consider the following table:

	Factor 2	
Factor 1	Level 1	Level 2
Level 1	2	4
Level 2	4	2

Table 5.59: **Fifth example of 2-factor ANOVA without repetition**

which in this form is not always obvious to interpret. But with the diagrams we immediately have a more relevant information:


Main Effects Plot for Value
Main Effects Plot for Value
Data Means
Factor 1
Factor 2
Mean
3.50
3.25
3.00
2.75
2.50
Level 1
Level 2
Level 1
Level 2

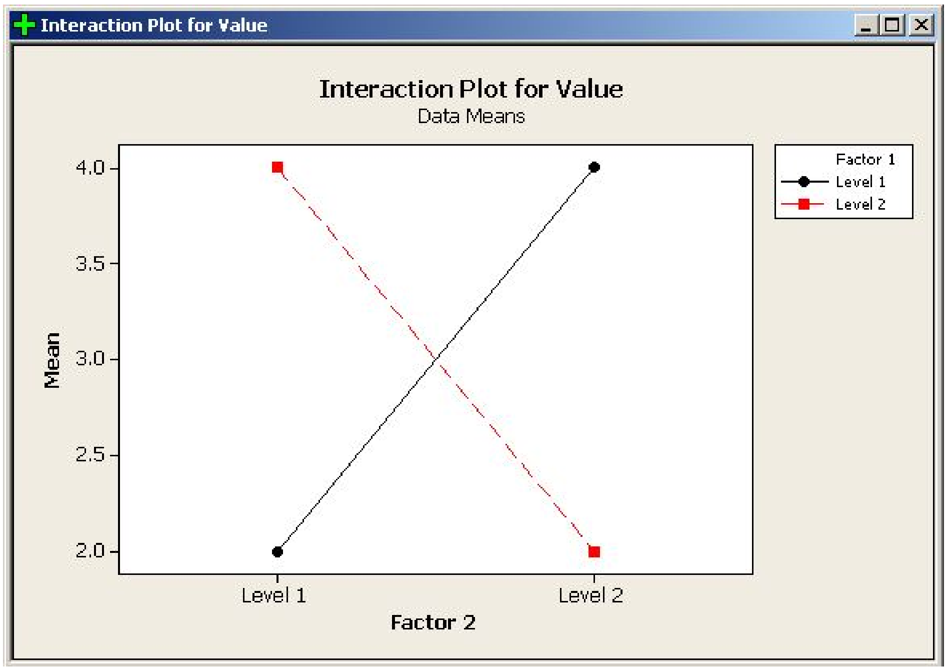

Interaction Plot for Value
Interaction Plot for Value
Data Means
Mean
4.0
3.5
3.0
2.5
2.0
Level 1
Level 2
Factor 2
Factor 1
Level 1
Level 2

Figure 5.194: **Main effects and interactions plot ("disordonal interaction") with interactions in Minitab 15**

We see well on the top diagram that none of the factors has an effect in average on the answer a priori (same diagram as at the very beginning of this series of examples with the same mean!). The bottom diagram gives us additional information by cons (!!!): The factors have a cross-influence and as this cross-influence is of the same magnitude, the effects cancel out. Then we say that: "(a priori) the two factors are (locally) in interaction in F1 * F2".

Caution! It may happen that a "statistical interaction" like the one above may not mean what you thing! In regression, when you predict Y from two variables (say, X and Z), they are therefore said to "interact" when the predictiveness of one variable depends on the other. For example, perhaps X predicts Y strongly when Z is high but terribly otherwise. In this case, you might say that X and Z interact to predict Y. We could think of one predictor (e.g., Z) like a light switch that turns the other variable on or off. For X to predict Y, you also need the right combination of X and Z. Very important! And for this reason, analysts and data scientists often explore this by including product terms in their regressions:

$$Y = X + Z + X * Z$$

If $X * Z$ explains unique variance in Y, this is a sign that Z and X interact. We can re-arrange terms to tell the story of this interaction, but this is how we commonly test it. So.... $X * Z$ is strong and significant. However... (!!!) you can also get the same effect in other cases. Eek! Why? Imagine that X has a quadratic relationship to Y, say:

$$Y = X + X^2$$

This is just a special case of interaction. So, in this case, X "moderates itself" and that makes sense: if the relationship between X and Y is non-linear, it's stronger at some levels of X but weaker at others. The relationship between X and Y depends.... on X! Well, if other variables correlate with X (e.g., Z), guess what? They may act the same way because they are picking up on that quadratic relationship. So, if you run the usual "test for interaction":

$$Y = X + Z + X * Z$$

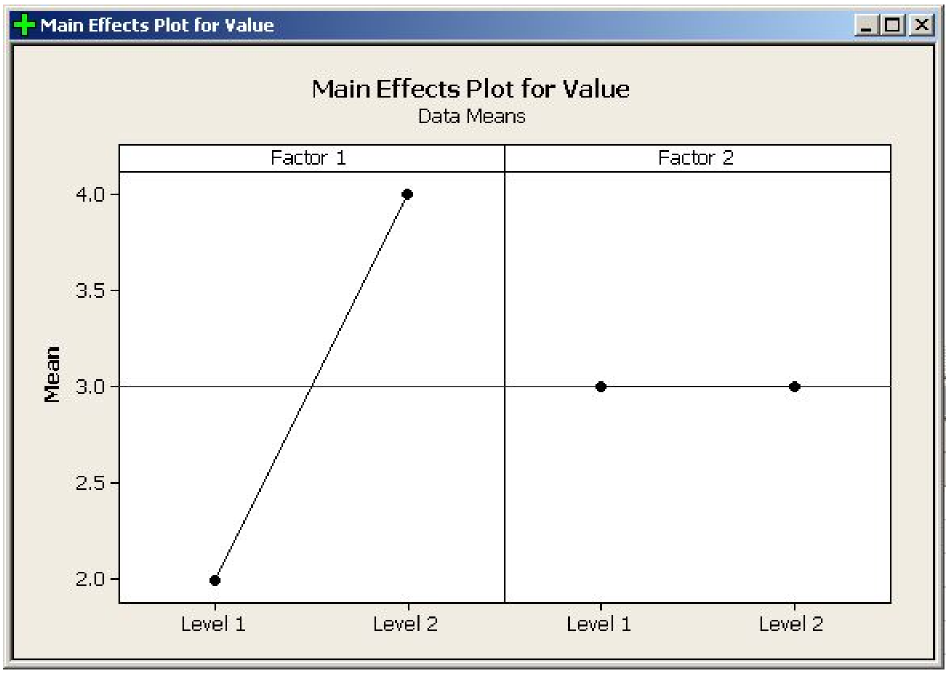
Your term may come up significant because it is tapping into the quadratic relationship X has with itself, which you haven't controlled for! So be sceptical of interactions, particularly when there is non-linearity present! Examine the linearity of relationships between your predictors (single one at a time) and your outcome if you do plan to test for interactions!

Now consider the following table:

	Factor 2	
Factor 1	Level 1	Level 2
Level 1	1	3
Level 2	5	3

Table 5.60: **Sixth example of 2-factor ANOVA without repetition**

which in this form is not also always obvious to interpret. But with the diagrams we immediately have a more relevant information:



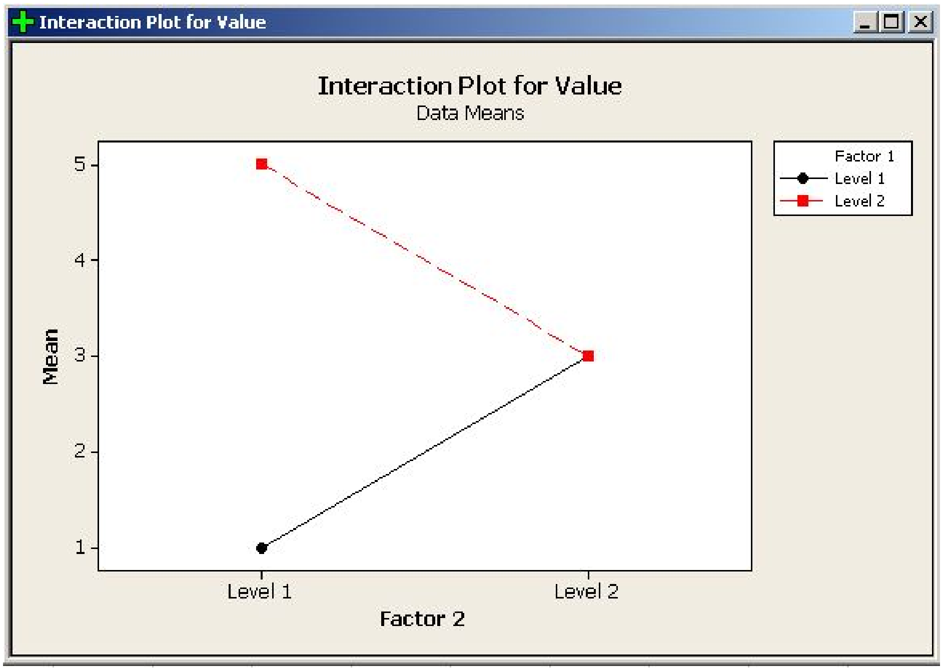


Figure 5.195: **Main effects and interactions plot with interactions in Minitab 15**

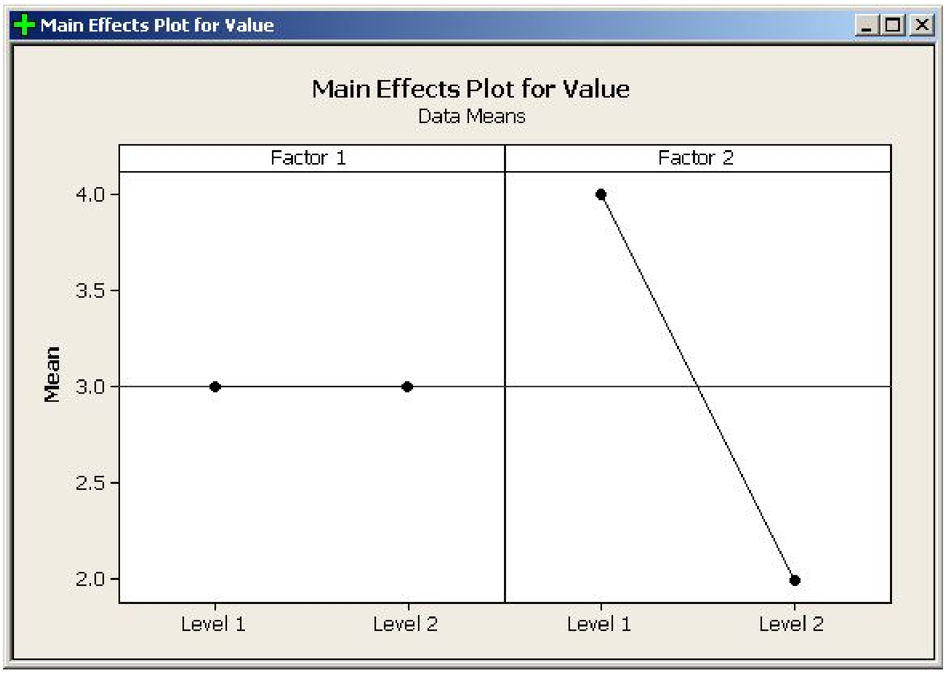
We observe well on the top diagram that the Factor 1 appears to have an influence and the factor 2 no (in average!). The bottom interactions diagram gives us, too, once again, additional information (!!!): It is that the factors are interacting. We say then that we have "(a priori) two factors (locally) in interaction F1 * F2 where the influence of Factor 1 is significant".

Now consider the following table:

	Factor 2	
Factor 1	Level 1	Level 2
Level 1	3	3
Level 2	5	1

Table 5.61: **Seventh example of 2-factor ANOVA without repetition**

We see that both factors influence the answer. We see that well on the both diagrams below:


Main Effects Plot for Value
Main Effects Plot for Value
Data Means
Factor 1
Factor 2
4.0
3.5
3.0
2.5
2.0
Mean
Level 1
Level 2
Level 1
Level 2

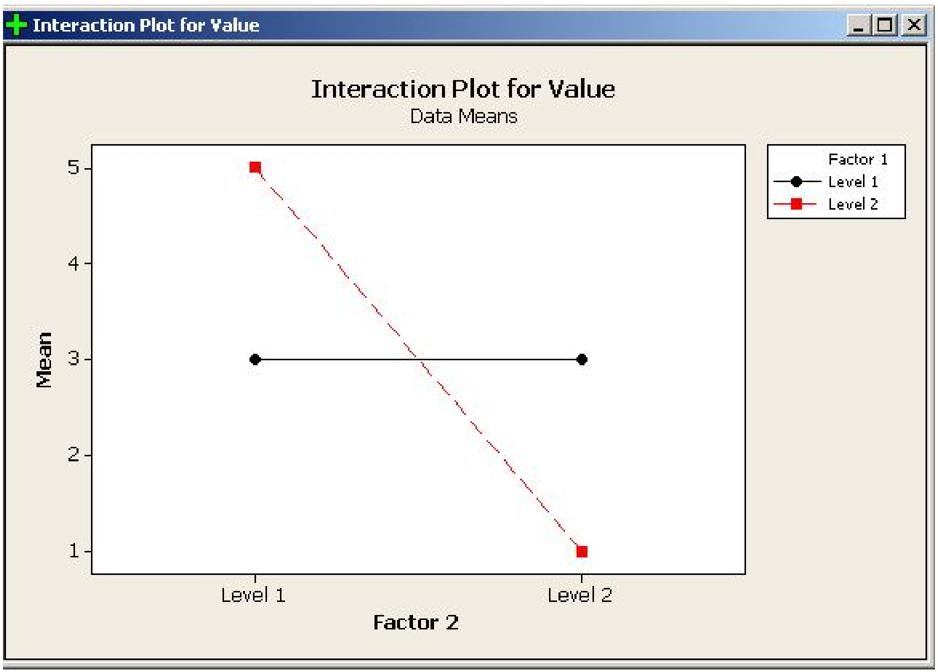

Interaction Plot for Value
Interaction Plot for Value
Data Means
5
4
3
2
1
Mean
Level 1
Level 2
Factor 2
Factor 1
Level 1
Level 2

Figure 5.196: **Main effects and interactions plot with interactions in Minitab 15**

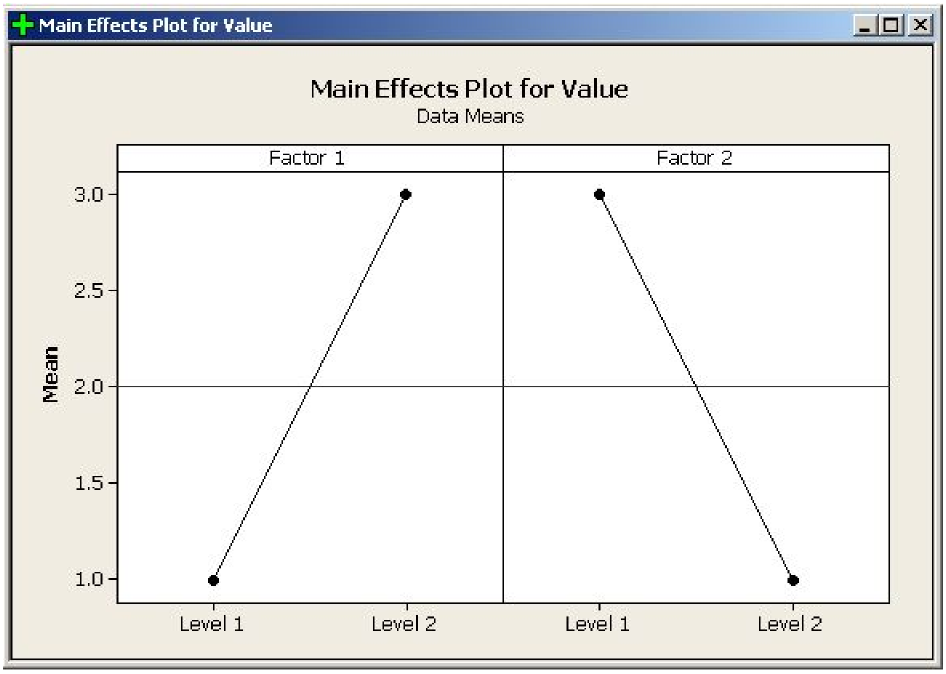
Here we say we have: "(a priori) the two factors (locally) interacting F1 * F2 where the influence of Factor 2 is significant".

And finally a last example:

	Factor 2	
Factor 1	Level 1	Level 2
Level 1	1	5
Level 2	5	1

Table 5.62: **Eighth and last example of 2-factor ANOVA without repetition**

Which gives us the two diagrams:


Main Effects Plot for Value
Main Effects Plot for Value
Data Means
Factor 1
Factor 2
Mean
3.0
2.5
2.0
1.5
1.0
Level 1
Level 2
Level 1
Level 2

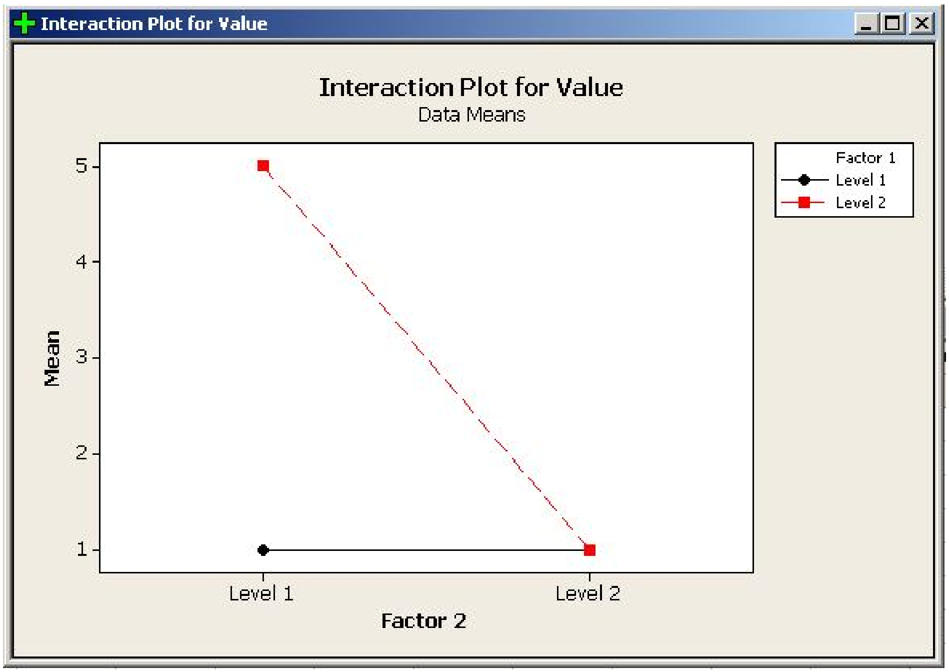


Figure 5.197: **Main effects and interactions plot with interactions in Minitab 15**

Here we say we have: "(a priori) the two factors (locally) interacting F1 * F2 where the influence of both factors is significant".

Remarks

R1. A belief (commonly spread) of people who have a lack experience in laboratories is to think that for an interaction to be significant it is necessary that the both factors that compose it are also significant.

R2. If an interaction is significant, the main effects are objectively rarely interpretable!

After all these diagrams and tables, let us turn back to the mathematical part:

We saw earlier how to perform a one-way fixed factor analysis of variance. To recap, this consists to a test for equality of means for k independent samples of n random variables each (in the case where all samples have the same number of measures: a balanced ANOVA). Each sample is regarded as an experiment on a different subject then considered therefore as an independent variable factor!!!

However it happens in the reality that we do for each sample vary a second parameter, then considered a second variable factor but independent of the first one. We speak then of course analysis of variance with two factors. In addition, we will consider in the first instance to simplify the calculations that the random variables are independent! Therefore a factor has no influence over the other has we already mention it !!! In other words there is no interaction between factors. We speak then of a "two-factor ANOVA without interactions".

To determine the formulation of the test to be performed, remember that for the analysis of variance with factor, we decomposed the total variance in the sum of the variance of average (inter-class) and of the residual variance (intraclass) such that:

$$Q_T = Q_A + Q_R \tag{5.7.2012}$$

explicitly we compare the samples $i=1\ldots k$:

$$\begin{aligned}Q_T&=\sum_{ij}(x_{ij}-\bar{\bar{x}})^2=\sum_i\sum_j(x_{ij}-\bar{\bar{x}})^2=\sum_i\sum_j((x_{ij}-\bar{x}_i)+(\bar{x}_i-\bar{\bar{x}}))^2\\&=\sum_i\sum_j(x_{ij}-\bar{x}_i)^2+\sum_i\sum_j(\bar{x}_i-\bar{\bar{x}})^2+2\sum_i\sum_j(x_{ij}-\bar{x}_i)(\bar{x}_i-\bar{\bar{x}})\end{aligned}\tag{5.7.2013}$$

which had given us in the end:

$$Q_T=\sum_i\sum_j(x_{ij}-\bar{x}_i)^2+\sum_i\sum_j(\bar{x}_i-\bar{\bar{x}})^2=Q_R+Q_A\tag{5.7.2014}$$

For the ANOVA with two fixed factors we will start form the table below:

	Factor A				
Factor B	Sample 1	Sample 2	Sample ...j	Sample r	
Sample 1	x_{11}	x_{12}	$\ldots$	x_{1r}	Average $\bar{x}_1$
Sample 2	x_{21}	x_{22}	$\ldots$	x_{2r}	Average $\bar{x}_2$
Sample i	$\ldots$	$\ldots$	x_{ij}	$\ldots$	Average $\bar{x}_i$
Sample k	x_{n1}	x_{n2}	$\ldots$	x_{kr}	Average $\bar{x}_k$
	Average: $\bar{x}_1$	Average: $\bar{x}_2$	Average: $\bar{x}_j$	Average: $\bar{x}_r$	$\bar{\bar{x}}$

Table 5.63: **Typical cross structure for the analysis of variance of 2 fixed factors without repetitions**

for which in a laboratory the factor kept fixed while the other is varied will be named the "block factor" and the other will be named the "treatment factor" and in practice it will be ensured that the latter does not always change out in the same order in the purpose to eliminate possible inertia effects when switching from one treatment to another (the Americans refer to two-factor controlled ANOVAs without interactions under the terms: "randomized block design" (GRBD)).

To continue the whole trick is to break down the total variance by comparing the average of the rows (observations) indexed this time with $i=1\ldots k$ and of columns (samples) indexed with $j=1\ldots r$ relatively to the total average such as:

$$\begin{aligned}Q_T&=\sum_{ij}(x_{ij}-\bar{\bar{x}})^2=\sum_{ij}(x_{ij}-\bar{\bar{x}})^2=\sum_i\sum_j\left[(\bar{x}_i-\bar{\bar{x}})+(\bar{x}_j-\bar{\bar{x}})+(x_{ij}-\bar{x}_i-\bar{x}_j+\bar{\bar{x}})\right]^2\\&=\sum_{ij}[((\bar{x}_i-\bar{\bar{x}})+(\bar{x}_j-\bar{\bar{x}}))^2+2\left((\bar{x}_i-\bar{\bar{x}})+(\bar{x}_j-\bar{\bar{x}})\right)(x_{ij}-\bar{x}_i-\bar{x}_j+\bar{\bar{x}})\\&+(x_{ij}-\bar{x}_i-\bar{x}_j+\bar{\bar{x}})^2]\\&=\sum_{ij}[(\bar{x}_i-\bar{\bar{x}})^2+2(\bar{x}_i-\bar{\bar{x}})(\bar{x}_j-\bar{\bar{x}})+(\bar{x}_j-\bar{\bar{x}})^2+2((\bar{x}_i-\bar{\bar{x}})+(\bar{x}_j-\bar{\bar{x}}))(x_{ij}-\bar{x}_i-\bar{x}_j+\bar{\bar{x}})\\&+(x_{ij}-\bar{x}_i-\bar{x}_j+\bar{\bar{x}})^2]\end{aligned}\tag{5.7.2015}$$

But we have in a first time:

$$\sum_{ij}2(\bar{x}_i-\bar{\bar{x}})(\bar{x}_j-\bar{\bar{x}})=2\underbrace{\sum_i(\bar{x}_i-\bar{\bar{x}})}_{=0}\underbrace{\sum_j(\bar{x}_j-\bar{\bar{x}})}_{=0}=0\tag{5.7.2016}$$

So it remains:

$$\begin{aligned}Q_T = &\sum_{ij}[(\bar{x}_i-\bar{\bar{x}})^2+(\bar{x}_j-\bar{\bar{x}})^2+2((\bar{x}_i-\bar{\bar{x}})+(\bar{x}_j-\bar{\bar{x}}))(x_{ij}-\bar{x}_i-\bar{x}_j+\bar{\bar{x}})\\ &+(x_{ij}-\bar{x}_i-\bar{x}_j+\bar{\bar{x}})^2]\end{aligned}\tag{5.7.2017}$$

But we also have:

$$\begin{aligned}&\sum_{ij}2((\bar{x}_i-\bar{\bar{x}})+(\bar{x}_j-\bar{\bar{x}}))(x_{ij}-\bar{x}_i-\bar{x}_j+\bar{\bar{x}})\\ &=2\sum_{ij}((\bar{x}_i-\bar{\bar{x}})+(\bar{x}_j-\bar{\bar{x}}))(x_{ij}-\bar{x}_i-\bar{x}_j+\bar{\bar{x}})\\ &=2\sum_{ij}(\bar{x}_i-\bar{\bar{x}})(x_{ij}-\bar{x}_i-\bar{x}_j+\bar{\bar{x}})+2\sum_{ij}(\bar{x}_j-\bar{\bar{x}})(x_{ij}-\bar{x}_i-\bar{x}_j+\bar{\bar{x}})\end{aligned}\tag{5.7.2018}$$

For the rest, let us indicate first that relatively to our above table, we have:

$$\begin{aligned}\bar{\bar{x}}&=\frac{1}{rk}\sum_{ij}x_{ij}\\ \bar{x}_i&=\frac{1}{r}\sum_i x_i\\ \bar{x}_j&=\frac{1}{k}\sum_j x_j\end{aligned}\tag{5.7.2019}$$

It follows then that:

$$\begin{aligned}&\sum_{ij}(\bar{x}_i-\bar{\bar{x}})(x_{ij}-\bar{x}_i-\bar{x}_j+\bar{\bar{x}})=\sum_i(\bar{x}_i-\bar{\bar{x}})\left(\sum_j x_{ij}-\sum_j\bar{x}_i-\sum_j\bar{x}_j+\sum_j\bar{\bar{x}}\right)\\ &=\sum_i(\bar{x}_i-\bar{\bar{x}})\left(r\bar{x}_i-\bar{x}_i\sum_j 1-\sum_j\frac{1}{k}\sum_i x_{ij}+\bar{\bar{x}}\sum_j 1\right)\\ &=\sum_i(\bar{x}_i-\bar{\bar{x}})\left(r\bar{x}_i-\bar{x}_i r-\frac{1}{k}\sum_{ij}x_{ij}+\bar{\bar{x}}r\right)=\sum_i(\bar{x}_i-\bar{\bar{x}})(-r\bar{\bar{x}}+\bar{\bar{x}}r)\\ &=\sum_i(\bar{x}_i-\bar{\bar{x}})\cdot 0=0\end{aligned}\tag{5.7.2020}$$

and then it comes immediately that we have likewise:

$$\sum_{ij}(\bar{x}_j-\bar{\bar{x}})(x_{ij}-\bar{x}_i-\bar{x}_j+\bar{\bar{x}})=0\tag{5.7.2021}$$

So it remains in the end:

$$\begin{aligned}Q_T&=\sum_{ij}[(\bar{x}_i-\bar{\bar{x}})^2+(\bar{x}_j-\bar{\bar{x}})^2+(x_{ij}-\bar{x}_i-\bar{x}_j+\bar{\bar{x}})^2]\\ &=\sum_{ij}(\bar{x}_i-\bar{\bar{x}})^2+\sum_{ij}(\bar{x}_j-\bar{\bar{x}})^2+\sum_{ij}(x_{ij}-\bar{x}_i-\bar{x}_j+\bar{\bar{x}})^2\\ &=k\sum_j(\bar{x}_k-\bar{\bar{x}})^2+r\sum_i(\bar{x}_i-\bar{\bar{x}})^2+\sum_{ij}(x_{ij}-\bar{x}_i-\bar{x}_j+\bar{\bar{x}})^2\end{aligned}\tag{5.7.2022}$$

what we will denote in this textbook in the following condensed form:

$$\boxed{Q_T = Q_A + Q_B + Q_R} \tag{5.7.2023}$$

where Q_A, Q_B are obviously associated with the main effects (comparison of marginal means with the total average).

Therefore compared to the one-way fixed factor ANOVA we have an additional term for the total variance!

In order it is clear that the first sum of the differences from the first column factor:

$$\sum_{ij} (\bar{x}_i - \bar{\bar{x}}) \tag{5.7.2024}$$

will have just as for the one-way fixed factor ANOVA $k-1$ degrees of freedom. That is to say, under the same assumptions as the one-way fixed factor ANOVA:

$$\frac{\sum\limits_{ij} (\bar{x}_i - \bar{\bar{x}})}{\sigma^2} = \chi^2_{k-1} \tag{5.7.2025}$$

The second sum of the differences from the second factor (line factor):

$$\sum_{ij} (\bar{x}_j - \bar{\bar{x}}) \tag{5.7.2026}$$

is new but we prove in a perfectly identical way to the first one it will have $k-1$ degrees of freedom. That is to say, under the same assumptions as the one-way fixed factor ANOVA:

$$\frac{\sum\limits_{ij} (\bar{x}_j - \bar{\bar{x}})}{\sigma^2} = \chi^2_{r-1} \tag{5.7.2027}$$

For the third sum that follows mandatory also a chi-square law (since the total variance follows a chi-square law and that the first two terms of the sum also!):

$$\sum_{ij} (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}})^2 \tag{5.7.2028}$$

it's a little bit trickier ... but there's a trick in the physicist way of life...! We know from our study of the one-way fixed factor ANOVA that the sum of the degrees of freedom of each term must equal the total number of degrees of freedom. In other words, we must have for the two-way fixed factor ANOVA:

$$\text{df}_T = N - 1 = kr - 1 = (k-1) + (r-1) + ??? \tag{5.7.2029}$$

So obviously what is missing is equal to:

$$??? = kr - 1 - [(k-1) + (r-1)] = kr - 1 - k + 1 - r + 1 = kr - k - r + 1 = (k-1)(r-1) \tag{5.7.2030}$$

Therefore:

$$\frac{\sum\limits_{ij}(x_{ij}-\bar{x}_i-\bar{x}_j+\bar{\bar{x}})^2}{\sigma^2}=\chi^2_{(k-1)(r-1)} \tag{5.7.2031}$$

We then have the following table:

$$\boxed{\begin{array}{lll}
Q_T=\sum\limits_{ij}(x_{ij}-\bar{\bar{x}})^2 & \mathrm{df}_T=N-1=kr-1 & \mathrm{V}_T=\dfrac{Q_T}{\mathrm{df}_T}\\
Q_A=k\sum\limits_{j}(\bar{x}_j-\bar{\bar{x}})^2 & \mathrm{df}_A=r-1 & \mathrm{V}_A=\dfrac{Q_A}{\mathrm{df}_A}\\
Q_B=r\sum\limits_{i}(\bar{x}_i-\bar{\bar{x}})^2 & \mathrm{df}_B=k-1 & \mathrm{V}_B=\dfrac{Q_B}{\mathrm{df}_B}\\
Q_R=\sum\limits_{ij}(x_{ij}-\bar{x}_i-\bar{x}_j+\bar{\bar{x}})^2 & \mathrm{df}_R=(k-1)(r-1) & \mathrm{V}_R=\dfrac{Q_R}{\mathrm{df}_R}
\end{array}} \tag{5.7.2032}$$

Finally, the rest is exactly the same as for the one-way fixed factor ANOVA just that we have two tests to do this time that are:

$$F_{k-1,(k-1)(r-1)}=\frac{\dfrac{Q_A}{k-1}}{\dfrac{Q_R}{(k-1)(r-1)}}\quad\text{and}\quad F_{r-1,(k-1)(r-1)}=\frac{\dfrac{Q_B}{r-1}}{\dfrac{Q_R}{(k-1)(r-1)}} \tag{5.7.2033}$$

The choice above seems intuitively practical...

All calculations we have made above are often represented in software in a standard form of a table whose form and content is presented below (this is how Minitab 15.1.1 and Microsoft® Excel® 11.8346 present it for example):

Sum of squares (SSE)	χ^2 **df**	**Mean squares**	F	**Critical** F
$Q_A=k\sum\limits_j(\bar{x}_j-\bar{\bar{x}})^2$	$k-1$	$\text{MSk}A=\dfrac{Q_A}{k-1}$	$\dfrac{\text{MSk}A}{\text{MSE}}$	$P(F>F_{k-1,(k-1)(r-1)})$
$Q_B=r\sum\limits_i(\bar{x}_i-\bar{\bar{x}})^2$	$r-1$	$\text{MSk}B=\dfrac{Q_B}{r-1}$	$\dfrac{\text{MSk}B}{\text{MSE}}$	$P(F>F_{r-1,(k-1)(r-1)})$
$Q_R=\sum_{ij}(x_{ij}-\bar{x}_i-\bar{x}_j+\bar{\bar{x}})^2$	$(k-1)(r-1)$	$\text{MSE}=\dfrac{Q_R}{(k-1)(n-1)}$		
$Q_T=\sum\limits_{ij}(x_{ij}-\bar{\bar{x}})^2$	$N-1$	$\text{MST}=\dfrac{Q_T}{N-1}$		

Table 5.64: Two-way fixed factor ANOVA table without repetitions

and the condition of acceptance of the null hypothesis of equality of means for each factor is the same as for the one-way fixed factor ANOVA (see the server of exercises for a detailed practical example with Microsoft® Excel® 11.8346).

So we have two Fisher tests to know whether if factor A (respectively B) have a significant influence on the measures or not.

Obviously, in the above developments, factors A and B are interchangeable in the developments by symmetry!

Remark

The corresponding factor effect model is (notice that the model is assumed to be purely additive, that is without interactions!):

$$\boxed{x_{ijk} = \mu + \tau_i + \beta_j + \varepsilon_{ij}} \tag{5.7.2034}$$

with $i = 1 \ldots a$, $j = 1 \ldots b$ where μ is the overall mean, τ_i is the effect of the ith of treatment A, β_j is the effect of the jth level of treatment B (main effects) and ε_{ij} is the habitual random error following $\mathcal{N}(0, \sigma_\varepsilon^2)$.

The assumptions on the effect terms are $\sum_{i=1}^{a} \tau_i = 0$, $\sum_{j=1}^{b} \beta_j = 0$.

The hypothesis of interest are $H_0: \; \tau_i = 0$ and $H_0: \; \beta_j = 0$.

We see that the maximum likelihood estimators of the model parameters are $\hat{\mu} = \bar{\bar{x}}$ and $\hat{\tau}_i = \bar{x}_i - \bar{\bar{x}}$ and $\hat{\beta}_j = \bar{x}_j - \bar{\bar{x}}$. Hence:

$$\hat{x_{ij}} = \bar{\bar{x}} + (\bar{x}_i - \bar{\bar{x}}) + (\bar{x}_j - \bar{\bar{x}}) = \bar{x}_i + \bar{x}_j - \bar{\bar{x}} \tag{5.7.2035}$$

5.7.13.5.3 ANOVA with two fixed factors with repetitions

So far we have examined ANOVA on experiments with one or two fixed factors (one or two categorical variables not sampled from a populations). In the case of two factors, we considered that for each combination of factors we only had one measure (cell). But it can happen (and this is better!) that have several measures for one combination of factors! We name this type of study of "experimental design with repeated measures" and the results will be processed with an analysis of variance for repeated measures (RM-ANOVA) with two fixed factors and interactions! This is an extremely important tool as it allows validation studies by several independent laboratories (or employees) and is also associated with many other statistical tools such as the study of reproducibility and repeatability (R&R Study) to name only the best known cases in the industrial field.

You must understand that it is mandatory in the field of statistics to involve interactions between factors systematically when we are dealing with an experience having repeated measures. This for the simple reason that the mathematical interaction term only appears in this situation!

Thus, it may be intuitive (even before the proof) that a two fixed factor ANOVA with repeated measures contains one double interaction, and two main effects. A three fixed factor ANOVA with repeated measures will have one triple interaction, three double interactions and 3 main effects. And so on...

Before we start, we will consider the following measurement table (this kind of table typically belongs to the family of "multi-dimensional crossed-table with subtotals"):

	Factor A				
Factor B	Sample 1	Sample 2	Sample ...j	Sample r	AVERAGE
Sample 1	x_{111}	x_{112}	...	x_{11r}	
Replication 2	x_{211}	x_{212}	...	x_{21r}	
Replication m	...	...	...	...	
Replication n	x_{n11}	x_{n12}	...	x_{n1r}	
Average Sample 1	$\bar{x}_{11}$	$\bar{x}_{12}$	...	$\bar{x}_{13}$	$\bar{x}_{1.}$
Sample 2	x_{121}	x_{122}	...	x_{12r}	
Replication 2	x_{221}	x_{222}	...	x_{22r}	
Replication m	...	...	...	...	
Replication n	x_{n21}	x_{n22}	...	x_{n2r}	
Average Sample 2	$\bar{x}_{21}$	$\bar{x}_{22}$	...	$\bar{x}_{2r}$	$\bar{x}_{2.}$
Sample i	...	...	x_{1ij}	...	
Replication 2	...	...	...	...	
Replication m	...	...	...	...	
Replication n	...	...	x_{nij}	x_{nir}	
Average Sample i	...	...	$\bar{x}_{ij}$	...	$\bar{x}_{i.}$
Sample k	x_{2k1}	x_{2k2}	...	x_{2kr}	
Replication 2	x_{21}	x_{22}	...	x_{2r}	
Replication m	...	...	...	...	
Replication n	x_{nk1}	x_{nk2}	...	x_{nkr}	
Average Sample k	$\bar{x}_{k1}$	$\bar{x}_{k2}$	...	$\bar{x}_{kr}$	$\bar{x}_{k.}$
AVERAGE	$\bar{x}_{.1}$	$\bar{x}_{.2}$	$\bar{x}_{.j}$	$\bar{x}_{.r}$	$\bar{\bar{x}}_{..}$

Table 5.65: **Typical cross structure for the analysis of variance of 2 fixed factors with repetitions**

with the usual properties of the mean (for reminder):

$$\begin{aligned}\bar{x}_{i.} &= \frac{1}{r}\sum_{j=1}^{r}\bar{x}_{ij} = \frac{1}{rn}\sum_{m=1}^{n}\sum_{j=1}^{r}x_{mij}\\ \bar{x}_{.j} &= \frac{1}{k}\sum_{i=1}^{k}\bar{x}_{ij} = \frac{1}{kn}\sum_{m=1}^{n}\sum_{k=1}^{k}x_{mij}\\ \bar{x}_{..} &= \frac{1}{r}\sum_{j=1}^{r}\bar{x}_{.j} = \frac{1}{k}\sum_{i=1}^{k}\bar{x}_{ij} = \frac{1}{krn}\sum_{m=1}^{n}\sum_{k=1}\sum_{j=1}^{r}x_{mij}\end{aligned} \tag{5.7.2036}$$

And remember that the two fixed factor ANOVA without replications (and therefore without possibility to analyse mathematically the interactions), all the trick was to break down the total variance by comparing the average of the lines indexed with $i = 1 \ldots k$ and of columns indexed with $j = 1 \ldots r$ compared with the total average.

The idea will now be just about the same except that we will compare the mean of the lines indexed with $i = 1 \ldots k$ and columns indexed with $j = 1 \ldots r$ not only in comparison to the total average but also that of each row and of each column.

For this purpose with start from what we got for the two fixed factor ANOVA without replication:

$$Q_T = \sum_{ij}(x_{ij}-\bar{\bar{x}})^2 = \sum_{ij}(x_i-\bar{\bar{x}})^2 + \sum_{ij}(x_j-\bar{\bar{x}})^2 + \sum_{ij}(x_{ij}-\bar{x}_i-\bar{x}_j+\bar{\bar{x}})^2 \tag{5.7.2037}$$

but whose notation will just be adapted to the context:

$$Q_T = \sum_{ij}(x_{ij}-\bar{\bar{x}}_{..})^2 = r\sum_{i}(x_{i.}-\bar{\bar{x}}_{..})^2 + k\sum_{j}(x_{.j}-\bar{\bar{x}}_{..})^2 + \sum_{ij}(x_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{\bar{x}}_{..})^2 \tag{5.7.2038}$$

Obviously, with this notation the two fixed factor ANOVA without replication becomes:

$$\begin{aligned}Q_T &= \sum_{ij}(x_{ij}-\bar{\bar{x}})^2 &&\rightarrow\quad Q_T = \sum_{ij}(x_{ij}-\bar{\bar{x}}_{..})^2\\ Q_A &= r\sum_{i}(x_i-\bar{\bar{x}})^2 &&\rightarrow\quad Q_A = r\sum_{i}(x_{i.}-\bar{\bar{x}}_{..})^2\\ Q_B &= k\sum_{i}(x_j-\bar{\bar{x}})^2 &&\rightarrow\quad Q_B = k\sum_{j}(x_{.j}-\bar{\bar{x}}_{..})^2\end{aligned} \tag{5.7.2039}$$

But in the present case, we must add a term for the replication and adapt the notation for the measurements. So, without repeating all the developments (it is a bit cheeky but...) we get directly:

$$Q_T = \sum_{mij}(x_{mij}-\bar{\bar{x}}_{..})^2 = nr\sum_{i}(x_{i.}-\bar{\bar{x}}_{..})^2 + nk\sum_{j}(x_{.j}-\bar{\bar{x}}_{..})^2 + \sum_{ijm}(x_{mij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{\bar{x}}_{..})^2 \tag{5.7.2040}$$

where in order, m is the replication of the sample i of the factor A and of sample j of factor B.

It comes then obviously the interclass variances for factors A and B that are immediate:

$$\begin{aligned}Q_A &= nr\sum_i \left(x_{i.}-\bar{\bar{x}}_{..}\right)^2\\ Q_B &= nk\sum_j \left(x_{.j}-\bar{\bar{x}}_{..}\right)^2\end{aligned}\tag{5.7.2041}$$

where Q_A, Q_B are obviously again associated with main effects (comparisons of marginal averages with the total average).

Now let's play a little by introducing in the sum, in positive and negative in the last term:

$$\sum_{ijm}\left(x_{mij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{\bar{x}}_{..}\right)^2\tag{5.7.2042}$$

the average of replications:

$$\bar{x}_{ij}\tag{5.7.2043}$$

that we will find in fine in the total sum of squares:

$$\begin{aligned}&\sum_{mij}\left(x_{mij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{\bar{x}}_{..}\right)^2=\sum_{mij}\left(x_{mij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{\bar{x}}_{..}+\bar{x}_{ij}-\bar{x}_{ij}\right)^2\\ &=\sum_{mij}\left(\left(\bar{x}_{ij}-x_{i.}-x_{.j}+\bar{\bar{x}}\right)+\left(x_{mij}-\bar{x}_{ij}\right)\right)^2=\sum_{mij}\left(\bar{x}_{ij}-x_{i.}-x_{.j}+\bar{\bar{x}}\right)^2\\ &+2\sum_{mij}\left(\bar{x}_{ij}-x_{i.}-x_{.j}+\bar{\bar{x}}\right)\left(x_{mij}-\bar{x}_{ij}\right)+\sum_{mij}\left(x_{mij}-\bar{x}_{ij}\right)^2\\ &=n\sum_{ij}\left(\bar{x}_{ij}-x_{i.}-x_{.j}+\bar{\bar{x}}\right)^2+2\sum_{mij}\left(\bar{x}_{ij}-x_{i.}-x_{.j}+\bar{\bar{x}}\right)\left(x_{mij}-\bar{x}_{ij}\right)+\sum_{mij}\left(x_{mij}-\bar{x}_{ij}\right)^2\end{aligned}\tag{5.7.2044}$$

Of course, we recognize rather quickly the intra-classes variance (often also named "residual error" or simply in the particular case of the two fixed factor ANOVA with repetitions "repeatability error"):

$$Q_R=\sum_{mij}\left(x_{mij}-\bar{x}_{ij}\right)^2\tag{5.7.2045}$$

and the term we can interpret (by comparison with the two factor ANOVA without repetitions) as the variance of interaction:

$$Q_{A\times B}=Q_{A,B}=n\sum_{ij}\left(\bar{x}_{ij}-x_{i.}-x_{.j}+\bar{\bar{x}}\right)^2=n\sum_{ij}\left(\bar{x}_{ij}-\bar{\bar{x}}\right)\tag{5.7.2046}$$

If the ANOVA is balanced the term:

$$2\sum_{mij}\left(\bar{x}_{ij}-x_{i.}-x_{.j}+\bar{\bar{x}}_{..}\right)\left(x_{mij}-\bar{x}_{ij}\right)\tag{5.7.2047}$$

must vanish. Let us check this:

$$\begin{aligned}\sum_{mij}\left(\bar{x}_{ij}-x_{i.}-x_{.j}+\bar{\bar{x}}\right)\left(x_{mij}-\bar{x}_{ij}\right)&=\sum_{mij}\left(x_{mij}-\bar{x}_{ij}\right)\left(\bar{x}_{ij}-x_{i.}-x_{.j}+\bar{\bar{x}}\right)\\ &=\sum_i\sum_j\sum_m\left(x_{mij}-\bar{x}_{ij}\right)\left(\bar{x}_{ij}-x_{i.}-x_{.j}+\bar{\bar{x}}\right)\end{aligned}\tag{5.7.2048}$$

and therefore for i and j fixed it comes:

$$(\bar{x}_{ij} - x_{i.} - x_{.j} + \bar{\bar{x}}) \underbrace{\sum_m (x_{mij} - \bar{x}_{ij})}_{=0} = 0 \tag{5.7.2049}$$

And therefore the summation over all i and j will be zero also by extension. Those who have a doubt about the cancellation of the two terms of the development above, may be able to reassure themselves by making a numerical application (enjoy!...).

Then finally:

$$\begin{aligned} Q_T &= \sum_{mij} (x_{mij} - \bar{\bar{x}}_{..})^2 \\ &= \underbrace{nk\sum_j (x_{.j} - \bar{\bar{x}}_{..})^2}_{Q_A} + \underbrace{nr\sum_i (x_{i.} - \bar{\bar{x}}_{..})^2}_{Q_B} + \underbrace{n\sum_{ij} (x_{mij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}}_{..})^2}_{Q_{A\times B}} + \underbrace{\sum_{mij} (x_{mij} - \bar{\bar{x}}_{ij})^2}_{Q_R} \end{aligned} \tag{5.7.2050}$$

where for reminder, n is therefore the number of replications, r the number of samples of factor A and k and the number of samples of factor B (the latter two parameters are often mixed by those who make calculations by hand). Result is sometimes denoted as following in the literature:

$$\begin{aligned} Q_T &= \sum_{mij} (x_{mij} - \bar{x}_{...})^2 \\ &= nk\sum_j (x_{.j} - \bar{x}_{...})^2 + nr\sum_i (x_{i.} - \bar{x}_{...})^2 + n\sum_{ij} (x_{mij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{...})^2 + \sum_{mij} (x_{mij} - \bar{\bar{x}}_{ij})^2 \end{aligned} \tag{5.7.2051}$$

So in comparison to the two fixed factor ANOVA without replication, we have an additional term for the total variance.

In the order it is clear that the first sum of the differences from the first factor column:

$$\sum_j (x_{.j} - \bar{\bar{x}}_{..})^2 \tag{5.7.2052}$$

will have just like for one-way fixed factor ANOVA and two-way fixed factor ANOVA without repetition $k-1$ degrees of freedom. That is to say, under the same assumptions of these latter ANOVA, we have:

$$\frac{\displaystyle\sum_j (x_{.j} - \bar{\bar{x}}_{..})^2}{\sigma^2} = \chi^2_{k-1} \tag{5.7.2053}$$

The second sum of the differences relatively to the second factor line:

$$\sum_i (x_{i.} - \bar{\bar{x}}_{..})^2 \tag{5.7.2054}$$

will have under the same assumptions the property:

$$\frac{\sum\limits_{i}(x_{i.}-\bar{\bar{x}}_{..})^2}{\sigma^2}=\chi^2_{r-1} \tag{5.7.2055}$$

Thanks to the reasoning performed with two fixed actor ANOVA without repetition, we know that for the interaction term:

$$\sum_{ij}(x_{mij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{\bar{x}}_{..})^2 \tag{5.7.2056}$$

We have:

$$\frac{\sum\limits_{ij}(x_{mij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{\bar{x}}_{..})^2}{\sigma^2}=\chi^2_{(k-1)(r-1)} \tag{5.7.2057}$$

It remains to determine the number of degrees of freedom of the last term:

$$\sum_{mij}(x_{mij}-\bar{\bar{x}}_{ij})^2 \tag{5.7.2058}$$

To do this, we proceed in the same way as for the two fixed factor ANOVA without repetitions. We know from our study of the one-way fixed factor ANOVA that the sum of the degrees of freedom of each term must be equal tot the total number of degrees of freedom. In other words, we must have for the two fixed factors ANOVA:

$$\mathrm{df}_T=N-1=nkr-1=(k-1)+(r-1)+(k-1)(r-1)+??? \tag{5.7.2059}$$

So what is missing is obvious:

$$\begin{aligned}???&=nkr-1-[(k-1)+(r-1)+(k-1)(r-1)]\\&=nkr-1-k+1-r+1-(kr-k-r+1)\\&=nkr-1-k+1-r+1-kr+k+r-1\\&=nkr-kr=N-kr=kr(n-1)\end{aligned} \tag{5.7.2060}$$

Therefore:

$$\frac{\sum\limits_{mij}(x_{mij}-\bar{x}_{ij})^2}{\sigma^2}=\chi^2_{N-kr} \tag{5.7.2061}$$

We have then the following table:

$$\boxed{\begin{array}{lll}
Q_T=\sum\limits_{mij}(x_{mij}-\bar{\bar{x}}_{..})^2 & \mathrm{df}_T=N-1=nkr-1 & \mathrm{V}_T=\dfrac{Q_T}{\mathrm{df}_T}\\
Q_A=nk\sum\limits_{j}(\bar{x}_{.j}-\bar{\bar{x}}_{..})^2 & \mathrm{df}_A=k-1 & \mathrm{V}_A=\dfrac{Q_A}{\mathrm{df}_A}\\
Q_B=nr\sum\limits_{i}(\bar{x}_{i.}-\bar{\bar{x}}_{..})^2 & \mathrm{df}_B=r-1 & \mathrm{V}_B=\dfrac{Q_B}{\mathrm{df}_B}\\
Q_{A\times B}=n\sum\limits_{ij}(\bar{x}_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{\bar{x}}_{..})^2 & \mathrm{df}_{A\times B}=(k-1)(r-1) & \mathrm{V}_{A\times B}=\dfrac{Q_{A\times B}}{\mathrm{df}_{A\times B}}\\
Q_R=\sum\limits_{mij}(x_{mij}-\bar{\bar{x}}_{..})^2 & \mathrm{df}_R=N-kr & \mathrm{V}_R=\dfrac{Q_R}{\mathrm{df}_R}
\end{array}} \tag{5.7.2062}$$

Finally, the rest is exactly the same as for the two-way fixed factor ANOVA without replication with the difference that now we just have three tests this time:

$$F_{k-1,N-kr}=\dfrac{\dfrac{Q_A}{k-1}}{\dfrac{Q_R}{N-kr}}\qquad F_{r-1,N-kr}=\dfrac{\dfrac{Q_B}{r-1}}{\dfrac{Q_R}{N-kr}}\qquad F_{(k-1)(r-1),N-kr}=\dfrac{\dfrac{Q_{A\times B}}{(k-1)(r-1)}}{\dfrac{Q_R}{N-kr}} \tag{5.7.2063}$$

Again the choice of the ratios is relatively intuitive!

All calculations we have made above are often represented in software in a standard form of a table whose form and content is presented below (this is how Minitab 15.1.1 and Microsoft® Excel® 11.8346 present it for example):

Sum of squares (SSE)	χ^2 **df**	**Mean squares**	F	**Critical** F
$Q_A=nk\sum_j(\bar{x}_{.j}-\bar{\bar{x}}_{..})^2$	$k-1$	$\text{MSk}A=\dfrac{Q_A}{k-1}$	$\dfrac{\text{MSk}A}{\text{MSE}}$	$P(F>F_{k-1,Nk-r})$
$Q_B=nr\sum_i(\bar{x}_{i.}-\bar{\bar{x}}_{..})^2$	$r-1$	$\text{MSk}B=\dfrac{Q_B}{r-1}$	$\dfrac{\text{MSk}B}{\text{MSE}}$	$P(F>F_{r-1,N-kr})$
$Q_{A\times B}=\sum_{mij}(\bar{x}_{mij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{\bar{x}}_{..})^2$	$(k-1)(r-1)$	$\text{MSk}AB=\dfrac{Q_{A\times B}}{(k-1)(n-1)}$	$\dfrac{\text{Msk}AB}{\text{MSE}}$	$P(F>F_{(k-1)(n-1),N-kr})$
$Q_R=\sum_{mij}(\bar{x}_{mij}-\bar{x}_{ij})^2$	$N-kr$	$\text{MSE}=\dfrac{Q_R}{N-kr}$		
$Q_T=\sum_{mij}(x_{mij}-\bar{\bar{x}}_{..})^2$	$N-1$	$\text{MST}=\dfrac{Q_T}{N-1}$		

Table 5.66: **Two-way fixed factor ANOVA table with repetitions**

and the null hypothesis of equality of means for each factor is the same as for one-way fixed factor ANOVA (see the exercise server for a detailed practical example with Microsoft® Excel® 11.8346).

So we have three Fisher tests to know whether for each if factor A, respectively B or interaction AB have a significant influence on the measures or not.

Obviously, in the above developments, factors A and B are interchangeable by symmetry!

Remark

The corresponding factor effect model is (notice that the model is not purely additive!):

$$\boxed{x_{ijk}=\mu+\tau_i+\beta_j+(\tau\beta)_{ij}+\varepsilon_{ijk}} \tag{5.7.2064}$$

with $i=1\ldots a$, $j=1\ldots b$, $k=1\ldots n$ where μ is the overall mean, τ_i is the effect of the ith of treatment A, β_j is the effect of the jth level of treatment B (main effects), $(\tau\beta)_{ij}$ is the effect of the interaction between τ_i and β_j, and ε_{ijk} is the habitual random error following $\mathcal{N}(0,\sigma_\varepsilon^2)$.

The assumptions on the effect terms are $\sum_{i=1}^a\tau_i=0$, $\sum_{j=1}^b\beta_j=0$, $\sum_{i=1}^a(\tau\beta)_{ij}=0$ for $j=1\ldots b$ and $\sum_{j=1}^b(\tau\beta)_{ij}=0$ for $i=1\ldots a$.

The hypothesis of interest are $H_0:\ \tau_i=0$, $H_0:\ \beta_j=0$ or $H_0:\ (\tau\beta)_{ij}=0$.

5.7.13.5.4 Multifactor ANOVA with Repeated Measures

The "multifactor categorical variables ANOVA with repeated measures" is simply the name by which specialists refer to the following fixed-factors ANOVA:

- three factors (fixed) ANOVA with or without repetition
- four factors (fixed) ANOVA with or without repetition
- five factors (fixed) ANOVA with or without repetition
- etc.

Obviously, the ANOVA with one and two fixed factors are also part of the family of multifactor ANOVA. Also be aware that most statistical software manages up to ANOVA with 15 fixed factors fixed factors (categorical variables) on the assumption condition that the plan is balanced (i.e. for each level of each factor, there is the same number of measurements). A spreadsheet software (like Microsoft® Excel® for example) by default is able to deal with ANOVA having a maximum of two fixed factors.

Okay now the reader may be disappointed (well I am also disappointed to have only one life...) because honestly I do not want to redo all the developments seen above for the one-way and two-way fixed factors ANOVA for 3, 4 and up to 15 factors because it would take more than 100 A4 pages to do in a pedagogical and clear form plus it is always based on the same mechanical development. Sadly the generalized theory of ANOVA well being much shorter, it is indigestible for my taste.

Remark

The practitioner (and the reader) must avoid at any cost to do the confusion between "Multifactor" / factorial ANOVA that refer to cases where we have more than one factor (i.e., categorical explanatory variable) and "Multivariate" ANOVA that is also named MANOVA (note the initial M) that refers to cases where we have multiple dependent / response variables (we will study that latter further below page 1091).

5.7.13.5.4.1 Interlude on CRD, RBD, RCBD, IBD, GRBD...

So far we have seen some very common ANOVA. But it is time to discuss the way to do the experiments before going further. But before it would be useful to first make a recall between two main type of experiments in science!

quote 5.7:1052.#50

> *To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of...*
>
> Ronald Fisher

Definition (#5.7.150): The "observational study" is defined as follows:

- Collect data in a way that does not directly interfere with how the data arise, i.e. merely "observe"
- Based on an observational study, we can only establish an **association**, in other words "**correlation**", between the explanatory and response variables
- If an observational study uses data from the past, it is named "retrospective study", whereas if data are collected throughout the study, it is named a "prospective study".

Definition (#5.7.151): The "experiment study" is defined as follows:

- Randomly assign subjects to various treatments
- Can establish **causal** connections between the explanatory and response variables!

As companion example, consider we want to evaluate the relationship between regularly working out and energy level. We can design the study as an observational study or an experiment.

- In an observational study, we sample two types of people from the population, those who choose to work out regularly and those who don't. Then, we find the average energy level for the two groups of people and compare.
- In an experiment, we sample a group of people from the population, and we randomly assign these people into two groups: those who will regularly work out throughout the course of study, and those who will not. The difference is that the decision of whether to work out or not is not left up to the subjects as in the observational study, but is instead imposed by researcher. At the end, we compare the energy level of the two groups.

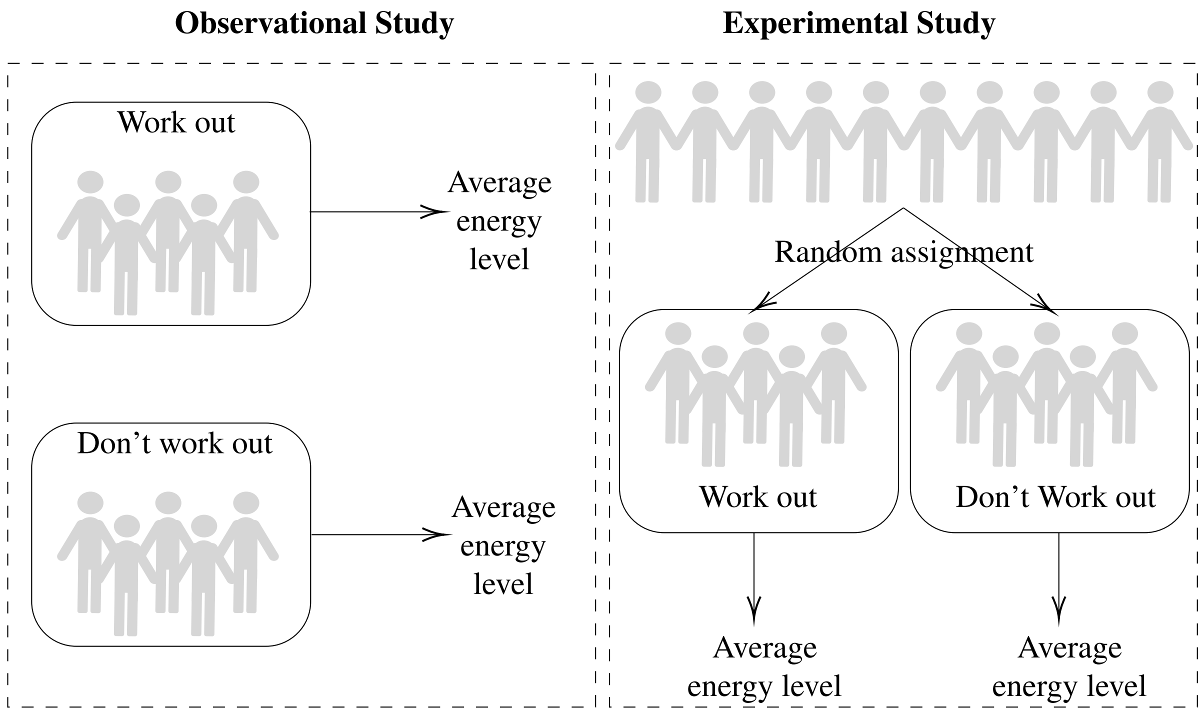
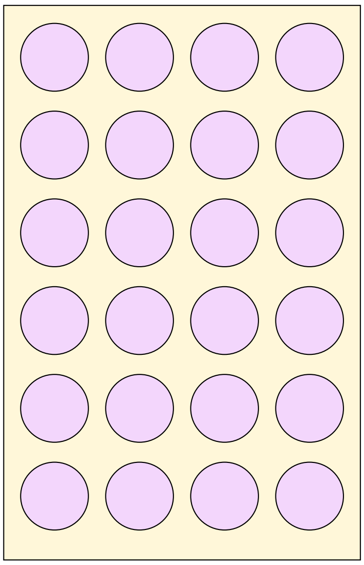

Observational Study
Experimental Study
Work out
Average energy level
Don't work out
Average energy level
Random assignment
Work out
Don't Work out
Average energy level
Average energy level

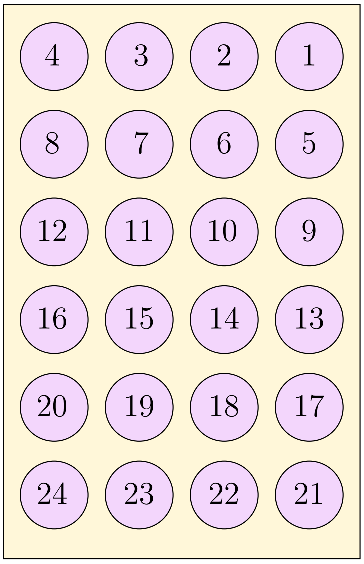
Figure 5.198: **Experimental vs Observational studies**

Based on the observational study, even if we find a difference between the energy levels of these two groups of people, we really can't attribute this difference solely to working out because there may be other variables that we didn't control for in this study that contribute to the observed difference. For example, people who are in better shape might be more likely to regularly work out and also have higher energy levels. However, in the experiment such variables that might also contribute to the outcome are likely equally represented in the two groups due to random assignment. Therefore, if we find a difference between the two averages, we can indeed make (likely) a causal statement attributing this difference to working out.

In practice, starting from the following experimental setup (one experimental unit, 24 observational units):



that we number as following:

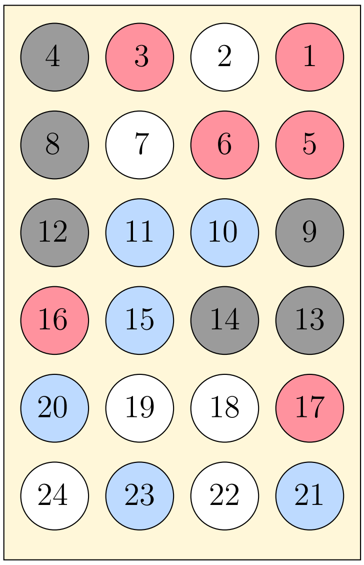


we differentiate the following configurations[75] from the more common to the most rare or advanced type of design (also named "random control trials" (RCT) or "random control designs" (RCD)):

- "Complete randomized design" (CRD):

  One-Way ANOVA (but not only) are typically associated to CRD as the idea behind is to make measurement on experimental units that are randomized for each level of the factor, to avoid a possible bias (confounding). A typical corresponding configuration will be:

  Fertilizer 1 : Blue
  Fertilizer 2 : Red
  Fertilizer 3 : Black
  No Fertilizer: White (control)



  The corresponding linear model is simple, beginning with a grand mean μ, and an "offset" term α_i for each of the a treatment groups:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad i = 1 \ldots a \quad j = 1..n_i \tag{5.7.2065}$$

---

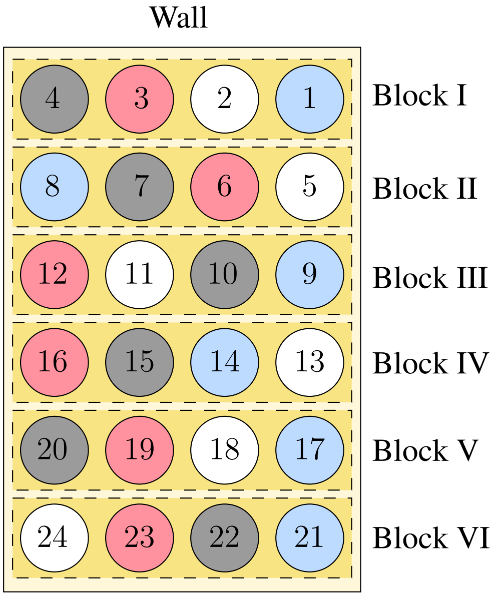
[75] Without taking into account that for each, there is a version for balanced and unbalanced designs and also a version for missing values

- "Randomized block design[76]" (RBD):

  When a factor is kept fixed, while we will vary the others, it will be named the "blocking factor[77]" and the others will be named "treatment factors" (some authors designate the two fixed factor ANOVA without interactions by the name "randomized block design").

  In the figure below, we see some blocks (corresponding typically to hard-to-change factors as days, observers, lands, etc.):

  Fertilizer 1 : Blue
  Fertilizer 2 : Red
  Fertilizer 3 : Black
  No Fertilizer: White (control)



  The corresponding linear model is simple as with just have a CRD applied to each block b:

$$y_{ijk}=\mu+\alpha_i+\beta_j+\varepsilon_{ijk} \qquad i=1\ldots a \quad j=1..b \quad k=1\ldots n_{ij} \tag{5.7.2066}$$

  In practice we will ensure that the randomization in the blocks is not always performed in the same order to eliminate potential inertia when changing from one treatment to the other effects (the reader can see in the previous figure that two blocks `I` and `V` have the same randomization). This lead us to the next type of design.

- "Randomized complete block design" (RCBD):

  The randomized complete block design (RCBD) also uses a restricted randomization scheme in the purpose to reduce the errors which account for differences among observation within each treatment. Within every block (e.g., location), the treatments are randomized to the experimental units (e.g., plots of land). This creates difference among the blocks and makes observation within a block more similar.

  The design is named "complete" because we see the complete set of treatments within every block and there is no identically randomized blocks (we will later also learn about incomplete block designs where this is not the case anymore). Note that blocking already exists at the time of randomization (and not only at the time of the analysis).

---

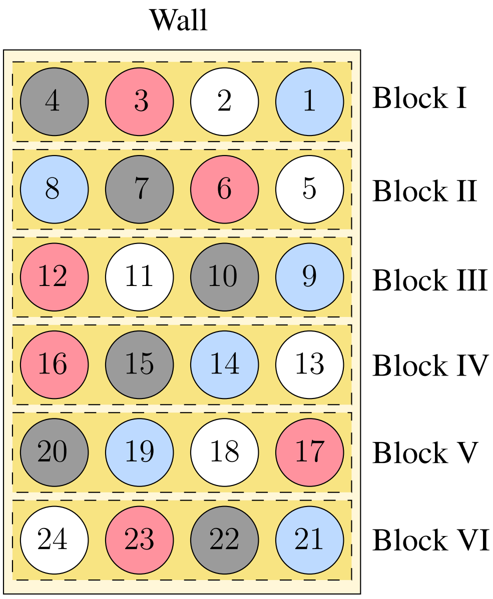
[76] That design and all others that follow belong to a family of designs named sometimes "crossover designs" or "crossover control trials"

[77] More formally a block is a matched group of individual who are similar or identical on a nuisance variable.

The analysis of a randomized complete block design is straightforward. We treat the block factor as "another" factor in our model[78] and then it is just a two-way ANOVA.

Because in the previous RBD the blocks I and V had identical randomization, we have now in RCBD configuration:

Fertilizer 1 : Blue
Fertilizer 2 : Red
Fertilizer 3 : Black
No Fertilizer: White (control)



So if you are able to maintain the condition homogeneous/constant throughout your experiment then CRD is just fine for lab experiment (lab experiments are usually conducted under the controlled conditions therefore the CRD test is usually used). But if you have resources like time, space and funds and you are confident enough to neutralize the variation (error) using suitable bollocking technique then go for the RCBD to reduce the experimental error.

- "Incomplete block design" (IBD):

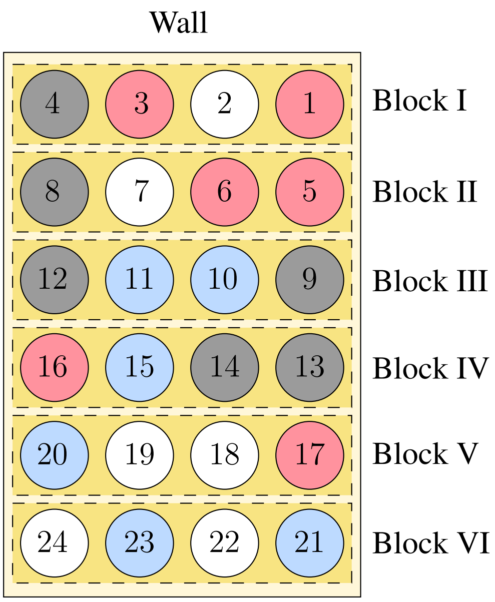
  Some experiments may consist of a large number of treatments and it may not be feasible to run all the treatments in all the blocks. Designs where only some of the treatments appear in every block are known as incomplete block designs.

  Here is an example where we can see that in some blocks, some treatments never appear:

---

[78] We can also block on more than one factor. A special case is the so-named Latin Square Design

Fertilizer 1 : Blue
Fertilizer 2 : Red
Fertilizer 3 : Black
No Fertilizer: White (control)

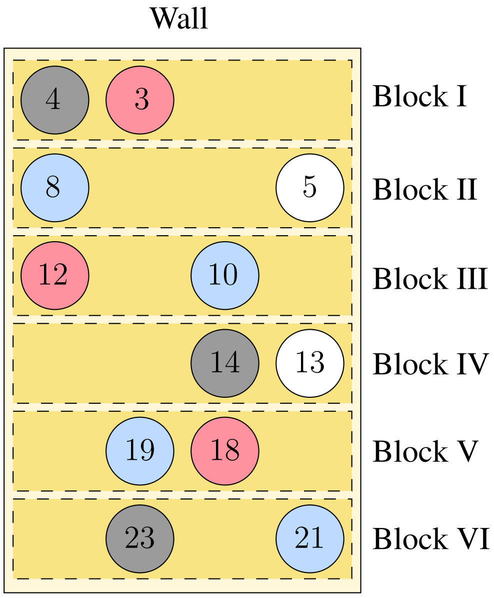

Wall
4 3 2 1
Block I
8 7 6 5
Block II
12 11 10 9
Block III
16 15 14 13
Block IV
20 19 18 17
Block V
24 23 22 21
Block VI

- "Balanced incomplete block design" (BIBD):

  In such a configuration, each block is selected in a balanced manner so that any pair of treatments occur together the same number of times as any other pairs.

  Considering the four treatments of our previous example, there are $\binom{4}{2} = 6$ ways of choosing three out of four:

Fertilizer 1 : Blue
Fertilizer 2 : Red
Fertilizer 3 : Black
No Fertilizer: White (control)


Wall
4 3
Block I
8 5
Block II
12 10
Block III
14 13
Block IV
19 18
Block V
23 21
Block VI

- "General randomized block design" (GRBD):

  Like a randomized complete block design (RCBD), a GRBD is randomized. Within each block, treatments are randomly assigned to experimental units: this randomization is also independent between blocks. In a (classic) RCBD, however, there is no replication of treatments within blocks. This replication allows the estimation and testing of an interaction term in the associated linear model.

- And others like PBIB (partial balanced incomplete design), CPBIB (cyclic balanced incomplete design), GCPIB (general cyclic balanced incomplete design), etc.

Pure random assignment guarantees quite well that the treatment and control groups will have identical characteristics on average, but in any particular random allocation, the two groups will differ along some dimensions! Methods used to pursue greater balance include stratification, pair-wise matching, and re-randomization. Sadly some researchers are not controlling for the method of randomization in their analysis and this leads to tests with incorrect sample size, and can result in lower power than if a pure random draw was used.

One way to control balance is just to do what we name a "randomization test". A randomization test is in reality a family of tests that we have already study before like:

- A series of Student T-test (see page 865) of two groups to compare the means of the treatment and control group variable by variable (at least the known one...) and check that the conclusion of the study remains the same or not.
- A series of Student paired T-test (see page 867) of two groups to compare the means of the treatment and control group knowing that they are correlated and do this variable by variable (at least the known one...) and check that the conclusion of the study remains the same or not.
- A series of Median test (see page 885) of two groups to compare the medians of the treatment and control group variable by variable (at least the known one...) and check that the conclusion of the study remains the same or not.
- A series of factorial ANOVA with or without repeated measures (see page 1001) to compare the averages of the treatment and control group by set of variables (at least the known one...) and check that the conclusion of the study remains the same or not (i.e. remember that's equivalent do to a regression of the type $y = a + b \cdot \text{Treat.} + \varepsilon$ and test whether $b = 0$).
- ...

Then we do this for 20 variables (still the known one...), find one or two are significant at a given threshold and then say «*This is about what we may expected by chance, so it seems randomization has succeeded quite well in generating balance*». But what if we find 3 or 4 differences out of 20 to be significant? Or what if none are individually significant, but the differences are all in the same direction?

An alternative, or complementary approach is to test for joint orthogonality. To do this, take a set of X variables $(x_1, x_2, \ldots, x_{20})$ and we run the following:

$$\text{Treat.} = a + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_{20}x_{20} + \varepsilon$$

And then we test the joint hypothesis $b_1 = b_2 = b_3 = \ldots = b_{20} = 0$. And we will see later in the section of Numerical Methods that this can be run as a linear regression, with an F-test (see page 6335); or as a probit (see page 6453), with a χ^2-test.

If we have multiple treatment groups we can then run a multinomial logit (see page 6477) or any other preferred specification and test for joint orthogonality within this framework!

5.7.13.5.4.2 Interlude on Type I, Type II, Type III ANOVAs

The purpose of this interlude is to clarify the meaning of the different types of sums of squares (and the underlying "philosophies") found in most statistical software in the context of variance (ANOVA). For this, we will place ourselves in ANOVA with two fixed factors.

Indeed, when data is unbalanced, there are different ways to calculate the sums of squares for ANOVA. There are at least 3 approaches, commonly named Type I sums of squares, Type II sums of squares and type III sums of squares (also named "Yates's weighted squares of means"). These names seem to have been introduced into the statistics world from the SAS package but are since decades pretty widespread. Which type to use has led to an ongoing controversy. However, it essentially comes down to testing different hypotheses about the data.

To introduce these three types, let us consider a fixed two-factor analysis model of variance in which:

- The first factor that we will denote by F_1, has J levels ($J \geq 2$) that will be indexed with the letter j
- The second factor that we will denote by F_2, has K levels ($K \geq 2$) that will be indexed with the letter k
- At the intersection of the level j of F_1 and the level k of F_2, we realize n_{jk} observations ($n_{jk} \geq 1$) of a random variable y (and we assume this time that the experimental design is not necessarily balanced)
- Each observation is denoted y_{ijk} (with $i = 1 \ldots n_{jk}$ and $j = 1 \ldots J$ and $k = 1 \ldots K$)
- We put:
  - $n_{j+} = \sum_{k=1}^{K} n_{jk}$ as the marginal sample size of the level j of F_1
  - $n_{+k} = \sum_{j=1}^{J} n_{jk}$ as the marginal sample size of the level k of F_2
  - $n = \sum_{j=1}^{J}\sum_{k=1}^{K} n_{jk}$ as the total number of measurements

Let us introduce the different empirical partial averages of observations y_{ijk}:

$$\begin{aligned}
\bar{y}_{.jk} &= \frac{1}{n_{jk}}\sum_{i=1}^{n_{jk}} y_{ijk}\\
\bar{y}_{.j.} &= \frac{1}{n_{j+}}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}} y_{ijk}\\
\bar{y}_{..k} &= \frac{1}{n_{+k}}\sum_{j=1}^{J}\sum_{i=1}^{n_{jk}} y_{ijk}\\
\bar{y}_{...} &= \frac{1}{n}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}} y_{ijk}
\end{aligned} \tag{5.7.2067}$$

Now let us consider the total sum of squares of the model, quantity with as we know $n-1$ degrees of freedom:

$$\text{TSS}=\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}(y_{ijk}-\bar{y}_{...})^2 \tag{5.7.2068}$$

We will first explicit again the decomposition of the quantity TSS. For this purpose, let us first notice the following equality:

$$y_{ijk}-\bar{y}_{...}=(y_{ijk}-\bar{y}_{.jk})+(\bar{y}_{.j.}-\bar{y}_{...})+(\bar{y}_{..k}-\bar{y}_{...})+(\bar{y}_{.jk}-\bar{y}_{.j.}-\bar{y}_{..k}+\bar{y}_{...}) \tag{5.7.2069}$$

As:

$$\begin{aligned}(a+b+c+d)^2&=a^2+2ab+2ac+2ad+b^2+2bc+2bd+c^2+2cd+d^2\\&=a^2+b^2+c^2+d^2+\sum_{l}^{6}\text{DP}_l(abc)\end{aligned} \tag{5.7.2070}$$

where DP means "double product (terms)". So with our problem we have:

$$(y_{ijk}-\bar{y}_{...})^2=(y_{ijk}-\bar{y}_{.jk})^2+(\bar{y}_{.j.}-\bar{y}_{...})^2+(\bar{y}_{..k}-\bar{y}_{...})^2+(\bar{y}_{.jk}-\bar{y}_{.j.}-\bar{y}_{..k}+\bar{y}_{...})^2+\sum_{l}^{6}\text{DP}_l(ijk) \tag{5.7.2071}$$

By triple summation, we get:

$$\begin{aligned}\text{SST}=\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}(y_{ijk}-\bar{y}_{...})^2&=\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}(y_{ijk}-\bar{y}_{.jk})^2+\sum_{j=1}^{J}n_{j+}(\bar{y}_{.j.}-\bar{y}_{...})^2+\sum_{k=1}^{K}n_{+k}(\bar{y}_{..k}-\bar{y}_{...})^2\\&+\sum_{j=1}^{J}\sum_{k=1}^{K}n_{jk}(\bar{y}_{.jk}-\bar{y}_{.j.}-\bar{y}_{..k}+\bar{y}_{...})^2+\sum_{l}^{6}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}\text{DP}_l(ijk)\end{aligned} \tag{5.7.2072}$$

To detail the double products DP_l, let us first recall that if the quantities x_{jk} are reals independent of i, we can write:

$$\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}x_{jk}(y_{ijk}-\bar{y}_{.jk})=\sum_{j=1}^{J}\sum_{k=1}^{K}x_{jk}\left[\sum_{i=1}^{n_{jk}}(y_{ijk}-\bar{y}_{.jk})\right] \tag{5.7.2073}$$

But we also know that:

$$\sum_{i=1}^{n_{jk}}(y_{ijk}-\bar{y}_{.jk})=0 \tag{5.7.2074}$$

Therefore:

$$\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}x_{jk}(y_{ijk}-\bar{y}_{.jk})=\sum_{j=1}^{J}\sum_{k=1}^{K}x_{jk}\left[\underbrace{\sum_{i=1}^{n_{jk}}(y_{ijk}-\bar{y}_{.jk})}_{=0}\right]=0 \tag{5.7.2075}$$

It follows that the sums of the first three double products (those in which the quantity $(y_{ijk}-y_{jk})$ is in factor) are zero (i.e. $2ab$, $2ac$, $2ad$). By cons, the other three (i.e. $2bc$, $2bd$, $2cd$) are, in general, not zero. The fourth ($2bc$) is written:

$$\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}\text{DP}_4(ijk)=2\sum_{j=1}^{J}\sum_{k=1}^{K}n_{jk}(\bar{y}_{.j.}-\bar{y}_{...})(\bar{y}_{..k}-\bar{y}_{...}) \qquad (5.7.2076)$$

double product that we will denote SDP_1. The sums of the last two double products ($2bd$ and $2cd$ hence $2d(b+c)$) will be grouped in the following expression:

$$\begin{aligned}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}(\text{DP}_5(ijk)+\text{DP}_6(ijk))&=2\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}(\bar{y}_{.jk}-\bar{y}_{.j.}-\bar{y}_{..k}+\bar{y}_{...})\left((\bar{y}_{.j.}-\bar{y}_{...})+(\bar{y}_{..k}-\bar{y}_{...})\right)\\&=2\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}(\bar{y}_{.jk}-\bar{y}_{.j.}-\bar{y}_{..k}+\bar{y}_{...})(\bar{y}_{.j.}+\bar{y}_{..k}-2\bar{y}_{...})\end{aligned} \qquad (5.7.2077)$$

sum of double product that we will denote SDP_2.

Let's now explicit the square sums by introducing the following quantities (we know the degrees of freedom thanks to our study of the two way fixed factor ANOVA for recall!):

- Sum of squares due to the factor F_1 (quantity at $J-1$ degrees of freedom):

$$\text{SS}F_1=\sum_{j=1}^{J}n_{j+}(\bar{y}_{.j.}-\bar{y}_{...})^2 \qquad (5.7.2078)$$

- Sum of squares due to the factor F_2 (quantity at $K-1$ degrees of freedom):

$$\text{SS}F_2=\sum_{k=1}^{K}n_{+k}(\bar{y}_{..k}-\bar{y}_{...})^2 \qquad (5.7.2079)$$

- Sum of squares due to the interactions of the factors (quantity at $(J-1)(K-1)$ degrees of freedom):

$$\text{SS}F_{1*2}=\sum_{j=1}^{J}\sum_{k=1}^{K}n_{jk}(\bar{y}_{.jk}-\bar{y}_{.j.}-\bar{y}_{..k}+\bar{y}_{...})^2 \qquad (5.7.2080)$$

- Sum of squares due to errors (or residual, quantity at $n-JK$ degrees of freedom):

$$\text{SSE}=\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{ijk}}(y_{ijk}-\bar{y}_{.jk})^2 \qquad (5.7.2081)$$

We can finally rewrite the total sum of squares in the form:

$$\text{SST}=\text{SS}F_1+\text{SS}F_2+\text{SS}F_{1*2}+\text{SSE}+\text{SDP}_1+\text{SDP}_2 \qquad (5.7.2082)$$

But most often written:

$$\text{SST}=\text{SS}F_A+\text{SS}F_B+\text{SS}F_{A*B}+\text{SSE}+\text{SDP}_A+\text{SDP}_B \tag{5.7.2083}$$

Or explicitly:

$$\boxed{\begin{aligned}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}(y_{ijk}-\bar{y}_{...})^2=&\sum_{j=1}^{J}n_{j+}(\bar{y}_{.j.}-\bar{y}_{...})^2+\sum_{k=1}^{K}n_{+k}(\bar{y}_{..k}-\bar{y}_{...})^2\\&+\sum_{j=1}^{J}\sum_{k=1}^{K}n_{jk}(\bar{y}_{.jk}-\bar{y}_{.j.}-\bar{y}_{..k}+\bar{y}_{...})^2\\&+\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{ijk}}(y_{ijk}-\bar{y}_{.jk})^2+2\sum_{j=1}^{J}\sum_{k=1}^{K}n_{jk}(\bar{y}_{.j.}-\bar{y}_{...})(\bar{y}_{..k}-\bar{y}_{...})\\&+2\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}(\bar{y}_{.jk}-\bar{y}_{.j.}-\bar{y}_{..k}+\bar{y}_{...})(\bar{y}_{.j.}+\bar{y}_{..k}-2\bar{y}_{...})\end{aligned}} \tag{5.7.2084}$$

We already know, from our derivation of the sum of squares of the balanced two way fixed factor ANOVA that in the balanced case:

$$\text{SDP}_A=\text{SDP}_B=0 \tag{5.7.2085}$$

The decomposition of SST is then of obvious interpretation. On the other hand, this is not the case with unbalanced plans ANOVA for which these two quantities are in general non-zero.

When the quantities SDP_A and SDP_B are non-zero, it is not possible to assign them to a single source of variation (F_A, F_B or $F_A\times F_B$). This explains the difficulties encountered in specifying the sources of variation in a model relating to an unbalanced plan. For this reason, other reasoning is used to specify these sources, which explains the existence of several types of sums of squares, according to the chosen philosophy.

Remark

Remember that obviously the three type of sum of squares that we will see now are always equal, and especially equal to zero as we already know!

Let us now represent the full model by $\text{SS}(A,B,AB)$, i.e.:

$$\begin{aligned}\text{SS}(A,B,AB)=&\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}(y_{ijk}-\bar{y}_{...})^2=\sum_{j=1}^{J}n_{j+}(\bar{y}_{.j.}-\bar{y}_{...})^2+\sum_{k=1}^{K}n_{+k}(\bar{y}_{..k}-\bar{y}_{...})^2\\&+\sum_{j=1}^{J}\sum_{k=1}^{K}n_{jk}(\bar{y}_{.jk}-\bar{y}_{.j.}-\bar{y}_{..k}+\bar{y}_{...})^2\\&+\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{ijk}}(y_{ijk}-\bar{y}_{.jk})^2+2\sum_{j=1}^{J}\sum_{k=1}^{K}n_{jk}(\bar{y}_{.j.}-\bar{y}_{...})(\bar{y}_{..k}-\bar{y}_{...})\\&+2\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}(\bar{y}_{.jk}-\bar{y}_{.j.}-\bar{y}_{..k}+\bar{y}_{...})(\bar{y}_{.j.}+\bar{y}_{..k}-2\bar{y}_{...})\end{aligned} \tag{5.7.2086}$$

Other models are represented similarly:

$$\begin{aligned}\text{SS}(A,B)&=\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}(y_{ijk}-\bar{y}_{...})^2=\sum_{j=1}^{J}n_{j+}(\bar{y}_{.j.}-\bar{y}_{...})^2+\sum_{k=1}^{K}n_{+k}(\bar{y}_{..k}-\bar{y}_{...})^2\\&+\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{ijk}}(y_{ijk}-\bar{y}_{.jk})^2+2\sum_{j=1}^{J}\sum_{k=1}^{K}n_{jk}(\bar{y}_{.j.}-\bar{y}_{...})(\bar{y}_{..k}-\bar{y}_{...})\\&+2\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}(\bar{y}_{.jk}-\bar{y}_{.j.}-\bar{y}_{..k}+\bar{y}_{...})(\bar{y}_{.j.}+\bar{y}_{..k}-2\bar{y}_{...})\end{aligned}\tag{5.7.2087}$$

indicates the model with no interaction. The next one:

$$\begin{aligned}\text{SS}(B,AB)&=\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}(y_{ijk}-\bar{y}_{...})^2=\sum_{k=1}^{K}n_{+k}(\bar{y}_{..k}-\bar{y}_{...})^2\\&+\sum_{j=1}^{J}\sum_{k=1}^{K}n_{jk}(\bar{y}_{.jk}-\bar{y}_{.j.}-\bar{y}_{..k}+\bar{y}_{...})^2\\&+\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{ijk}}(y_{ijk}-\bar{y}_{.jk})^2+2\sum_{j=1}^{J}\sum_{k=1}^{K}n_{jk}(\bar{y}_{.j.}-\bar{y}_{...})(\bar{y}_{..k}-\bar{y}_{...})\\&+2\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}(\bar{y}_{.jk}-\bar{y}_{.j.}-\bar{y}_{..k}+\bar{y}_{...})(\bar{y}_{.j.}+\bar{y}_{..k}-2\bar{y}_{...})\end{aligned}\tag{5.7.2088}$$

indicates the model that does not account for main effects from factor A. We also have:

$$\begin{aligned}\text{SS}(A)&=\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}(y_{ijk}-\bar{y}_{...})^2=\sum_{j=1}^{J}n_{j+}(\bar{y}_{.j.}-\bar{y}_{...})^2\\&+\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{ijk}}(y_{ijk}-\bar{y}_{.jk})^2+2\sum_{j=1}^{J}\sum_{k=1}^{K}n_{jk}(\bar{y}_{.j.}-\bar{y}_{...})(\bar{y}_{..k}-\bar{y}_{...})\\&+2\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}(\bar{y}_{.jk}-\bar{y}_{.j.}-\bar{y}_{..k}+\bar{y}_{...})(\bar{y}_{.j.}+\bar{y}_{..k}-2\bar{y}_{...})\end{aligned}\tag{5.7.2089}$$

and so on.

The influence of particular factors (including interactions) can be tested by examining the differences between models. For example, to determine the presence of an interaction effect, an F-test of the models $\text{SS}(A,B,AB)$ and the no-interaction model $\text{SS}(A,B)$ would be carried out. It is handy to define incremental sums of squares to represent these differences. Let:

$$\begin{aligned}\text{SS}(AB|A,B)&=\text{SS}(A,B,AB)-\text{SS}(A,B)\\\text{SS}(A|B,AB)&=\text{SS}(A,B,AB)-\text{SS}(B,AB)\\\text{SS}(A|B)&=\text{SS}(A,B)-\text{SS}(B)\\\text{SS}(B|A)&=\text{SS}(A,B)-\text{SS}(A)\end{aligned}\tag{5.7.2090}$$

The notation shows the incremental differences in sums of squares, for example $\text{SS}(AB|A,B)$ represents the sum of squares for interaction after the main effects has been removed, and $\text{SS}(A|B)$ is the sum of squares for the A main effect after the B main effect have been removed and ignoring interactions. The different types of sums of squares then arise depending on the stage of model reduction at which they are carried out. In particular:

- The "Type I ANOVA" also named "sequential sum of squares" is computed using $\text{SS}(A)$ for factor A. $\text{SS}(B|A)$ for factor B. $\text{SS}(AB|B,A)$ for interaction AB. This tests the main effect of factor A, followed by the main effect of factor B after the main effect of A, followed by the interaction effect AB after the main effects. Because of the sequential nature and the fact that the two main factors are tested in a particular order, this type of sums of squares will give different results for unbalanced data depending on which main effect is considered first[79]. This is why type I sums of squares, in unbalanced designs, is reserved to models in which there is a natural order between the factors. Notice that for type I, the individual effect SS sum to the total effect SS as:

$$\begin{aligned}\text{SS}(A)+\text{SS}(B|A)+\text{SS}(AB|B,A)&=\text{SS}(A)+[\text{SS}(A,B)-\text{SS}(A)]+[\text{SS}(A,B,AB)-\text{SS}(A,B)]\\&=\text{SS}(A,B,AB)=\text{SST}\end{aligned}\tag{5.7.2091}$$

  Hence the type I ANOVA is term-order-sensitive. Forget about it and your study is doomed! But it handles interactions, follows the principle of marginality and quite simple. This is the default in the `R` statistical software.

- The "Type II ANOVA"is calculated using $\text{SS}(A|B)$ for factor A and $\text{SS}(B|A)$ for factor B. This type tests for each main effect after the other main effect. Note that no significant interaction is assumed (in other words, we should test for interaction first ($\text{SS}(AB|A,B)$) and only if AB is not significant, we continue with the analysis for main effects). If there is indeed no interaction, then type II is statistically more powerful than type III (see below). Notice the important property that type II is order independent at the opposite of the other types. Notice that for type II, it is obvious that the individual effect SS does not sum to the total effect SS at the opposite of Type I. Indeed:

$$\begin{aligned}\text{SS}(A|B)+\text{SS}(B|A)&=[\text{SS}(A,B)-\text{SS}(B)]+[\text{SS}(A,B)-\text{SS}(A)]\\&=2\text{SS}(A,B)-\text{SS}(A)-\text{SS}(B)\\&=\text{SS}(A)+\text{SS}(B)\neq\text{SST}\end{aligned}\tag{5.7.2092}$$

  Hence the type II ANOVA is order-insensitive but doesn't support interactions.

- The "Type III ANOVA" also named "marginal sum of squares" is calculated using $\text{SS}(A|B,AB)$ for factor A. $\text{SS}(B|A,AB)$ for factor B. This type tests for the presence of a main effect after the other main effect and interaction. This approach is therefore valid in the presence of significant interactions. However, it is often not interesting to interpret a main effect if interactions are present (generally speaking, if a significant interaction is present, the main effects should not be further analysed). If the interactions are not significant, type II gives a more powerful test. Notice also that as for type II, it is obvious that the individual effect SS does not sum to the total effect SS:

---

[79] However in the case of balanced designs, the result will be the same and we fall back on the classical two-way fixed factor ANOVA

$$\begin{aligned}\text{SS}(A|B,AB)+\text{SS}(B|A,AB) &= [\text{SS}(A,B,AB)-\text{SS}(B,AB)]+[\text{SS}(A,B,AB)-\text{SS}(A,AB)]\\ &= 2\text{SS}(A,B,AB)-\text{SS}(B,AB)-\text{SS}(A,AB)\\ &= 2\text{SS}(A,B,AB)-[\text{SS}(B)+\text{SS}(AB)]-[\text{SS}(A)+\text{SS}(AB)]\\ &= \text{SS}(A)+\text{SS}(B)\neq \text{SST}\end{aligned} \tag{5.7.2093}$$

Hence the type III ANOVA is order-insensitive and most flexible, but can mislead you no less in presence of disordinal interactions (and sometimes the ordinal too). This is the default in the SAS™ statistical software.

Remarks

R1. When data is balanced, the factors are orthogonal, and types I, II and III all give the same results!

R2. In the beginning of the 121st century (holocene calendar), the softwares SAS and SPSS use SS type III as their default, while functions that ship with base `R` use type I by default. This can lead to different results when analysing the same data with different statistics packages.

In summary, the hypothesis of interest is usually about the significance of one factor while controlling for the level of the other factors. This equates to using type II or III SS. In general, if there is no significant interaction effect, then type II is more powerful, and follows the principle of marginality. If interaction is present, then type II is inappropriate while type III can still be used, but results need to be interpreted with caution (in the presence of interactions, main effects are rarely interpretable).

The subject is quite controversial since decades. This is why, when possible, it would be better to avoid unbalanced experiments!

5.7.13.5.5 Hierarchical (Nested) ANOVA (HANOVA)

The "hierarchical fixed factor ANOVA" (HANOVA) or "nested ANOVA with fixed factors", more often referred to as "two stage nested design", consists in analysing samples where one of the control factors is not independent but dependent on another control factor (this is why often beginners sometimes hesitate somewhat between a two-factor ANOVA and a two-factor hierarchical ANOVA).

Consider to introduce the concept the historical case of hierarchical ANOVA (agriculture!), two completely independent surfaces: Study Surface A and Study Surface B. Each of these surfaces (named a "whole plots" or "block effect") is separated into 4 distinct study squares (named "split-plots" or "subplots" or "sub-factors") with 5 replicas for each square. So we have a total of 2 whole plot, with 8 split-plots (test surfaces) for a total of 40 individuals[80]:

---

[80] In order to better understand the structure of the simple split-plot design it is advantageous to view it as superimposing one RCBD (for the split-plot treatments) on top of another RCBD (for the whole treatments). We shall refer to this as SPD (RCBD,RCBD). Variations of this form of split-plot designs are possible by using other than both RCBD.

Study Surface A

$S_1(A)$	$S_2(A)$	$S_3(A)$	$S_4(A)$
X	X	X	X
X	X	X	X
X	X	X	X
X	X	X	X
X	X	X	X

Study Surface B

$S_1(B)$	$S_2(B)$	$S_3(B)$	$S_4(B)$
X	X	X	X
X	X	X	X
X	X	X	X
X	X	X	X
X	X	X	X

Table 5.67: **Hierarchical ANOVA illustration**

We are therefore in the particular situation where the effects of the levels of the factor of the second level (squares) have no concrete meaning. For example, these squares depend on the level of the *Surface* factor considered and a study of the main effects of the *Square* factor is here irrelevant!

We can consider, for example, that the wholeplot factor represents suppliers, the subplot factor object lot types, and therefore for each supplier a given level of the subplot will physically represent a specific lot for each supplier and therefore not interchangeable.

Or a more common example are the "nested R&R (Repeatability and Reproducibility) Gage studies". Indeed, "measurement systems analysis" (MSA) are essential to the success of any data analysis. If you cannot rely on the tool you're using to take measurements, then why bother collecting data to begin with? It would be like trying to lose weight while relying on a scale that doesn't work. What's the point in weighing yourself?

In MSA studies for continuous measurements (e.g. weight, length, volume) using non-destructive testing, each part can be measured repeatedly. In this case, we can use "crossed Gage studies". However, sometimes we must conduct a MSA where the test required to take the measurement destroys the object or physically changes the characteristic that is being measured.

Suppose again that we are conducting a Gage R&R study for a destructive test with 3 operators and 2 replicates per part. But suppose it is not feasible to obtain 6 specimens that are similar enough to be considered the same part, rather only 2 specimens. Then in this case, you must use a "nested Gage study".

When a nested Gage R&R study is not enough, try an "expanded Gage R&R". Most Gage R&R studies assess the effects of two factors on variation in your measurement system - typically operator and part. However, the effects of operator and part frequently are not enough to provide a complete understanding of the measurement system. Adding a third variable (typically gage) to the standard study is often required. When three or more factors are included in the analysis, we call the study an "expanded Gage R&R". In the following situations, a third factor is crucial to understanding the system.

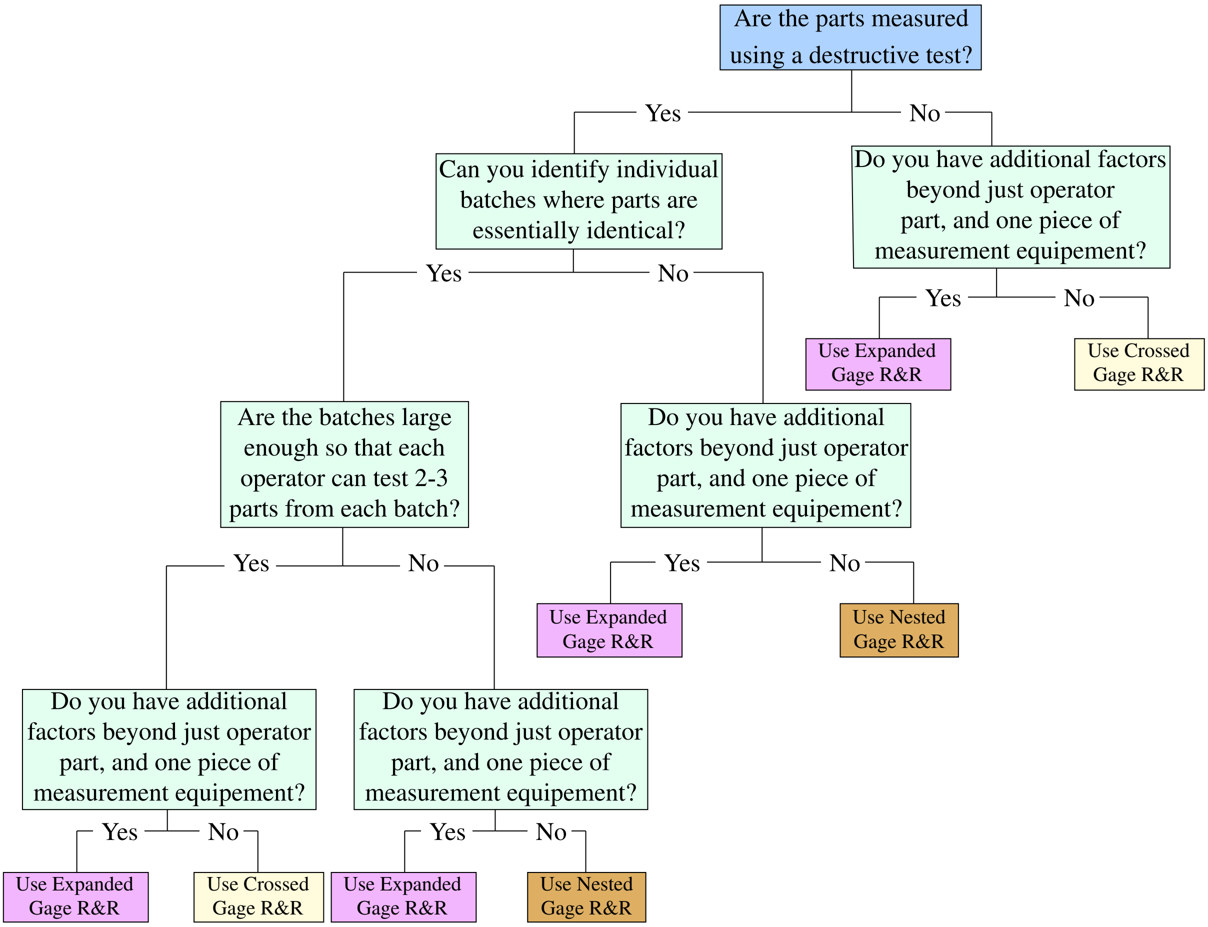

Are the parts measured using a destructive test?
Yes
No
Can you identify individual batches where parts are essentially identical?
Do you have additional factors beyond just operator part, and one piece of measurement equipement?
Yes
No
Use Expanded Gage R&R
Use Crossed Gage R&R
Yes
No
Are the batches large enough so that each operator can test 2-3 parts from each batch?
Do you have additional factors beyond just operator part, and one piece of measurement equipement?
Yes
No
Use Expanded Gage R&R
Use Nested Gage R&R
Yes
No
Do you have additional factors beyond just operator part, and one piece of measurement equipement?
Do you have additional factors beyond just operator part, and one piece of measurement equipement?
Yes
No
Yes
No
Use Expanded Gage R&R
Use Crossed Gage R&R
Use Expanded Gage R&R
Use Nested Gage R&R

Figure 5.199: **How to select a Gage R&R design**

So keep in mind:

- Crossed R&R studies are mostly related to non-destructive measurements (as explained above)
- Nested R&R studies are mostly related to destructive measurements (as explained above)

Each operator measures a different set of parts. Therefore, each part is said to be "nested" within operator rather than "crossed" since each part is unique to an operator. Let's look at a diagram for another view of crossed versus nested studies:

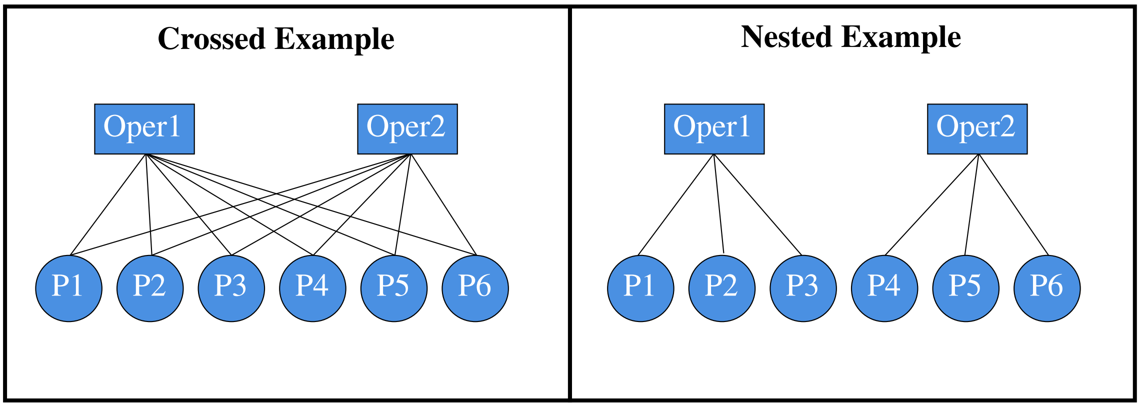


Figure 5.200: **Cross vs Nested studies (source: Minitab)**

The crossed Gage study is associated with a normal fixed factor ANOVA, or a Latin square ANOVA if the measurements are not repeated (see further below page 1074) where the nested Gage study is associated with a nested ANOVA!

It is customary to say that we have a "two-factor split-plot" or a "two-way hierarchical ANOVA" experiment because the behaviour of the squares is not independent of that of the surfaces.

Remarks

R1. To easily differentiate a fixed two-factor ANOVA with a nested fixed two-factor ANOVA with, it is sufficient to note that physically a given level of the first fixed factor represents the same object for each of the levels of the second factor while for the hierarchical ANOVA it is each time a different object specific to the factor under which it is.

R2. We have given above two examples with only one nested level, but it is possible to develop the theory for multiple nesting levels (a level, in a level that is itself in a level, and so on). We are talk then about "multiple hierarchical ANOVA".

A hierarchical experience and the corresponding analysis must first be able to tell us if there is a significant difference between the "whole plots" (controlled factors) and also if there is a significant difference between the "subplots".

It is important to remember that in this type of structure as each level of a factor is present only once in the level of another factor we can not estimate the interactions between the two. Indeed, if we consider that the factor whole-plot is a country and the subplot a lake ... it is difficult then to move the lake in another country ... In taking our example above with the surfaces it is also not very trivial to move a square to another surface...

Remark

We can not use a model where factors are nested unless we have repetitions. In the opposite case where the tests would not be repeated, the effect due to the nested factor can not be studied and the model that we will have to use to analyse the data will be one of those exposed in our study of the analysis of the one-way variance.

In the special case where the statistical study of the hierarchical ANOVA has failed us to highlight a significant difference between the "subplots" of the two "whole plots", but on the other hand has successfully highlight a significant difference between the two "whole plots" this should lead us to conclude to a particular environmental influence.

Denoting by $i=1\ldots M$ the number of levels of the main factor A, $j=i\ldots m$ the number of levels of the nested factor B and $k=1\ldots n$ the number of repeated measurements (thus a total of measurements $M\cdot m\cdot n$) and the averages:

$$\begin{aligned}\bar{x}_{...}&=\frac{1}{Mmn}\sum_{ijk}x_{ijk}=\frac{1}{N}\sum_{ijk}x_{ijk}\\ \bar{x}_{i..}&=\frac{1}{mn}\sum_{jk}x_{ijk}\\ \bar{x}_{ij.}&=\frac{1}{Mmn}\sum_{k}x_{ijk}\end{aligned}\tag{5.7.2094}$$

and under the same assumptions of use as the one-way or two-way ANOVA seen earlier above (independence of the errors ε_{ijk}, the errors have the same variance, the errors follow a Normal law), we write the decomposition of the variance in the form:

$$\begin{aligned}Q_T&=\sum_{ijk}\left(x_{ijk}-\bar{x}_{...}\right)^2=\sum_{ijk}\left(x_{ijk}-\bar{x}_{...}+\bar{x}_{i..}-\bar{x}_{i..}+\bar{x}_{ij.}-\bar{x}_{ij.}\right)^2\\ &=\sum_{ijk}\left((x_{i..}-\bar{x}_{...})+(\bar{x}_{ij.}-\bar{x}_{i..})+(x_{ijk}-\bar{x}_{ij.})\right)^2\\ &=\sum_{ijk}\left((\bar{x}_{i..}-\bar{x}_{...})^2+2(\bar{x}_{i..}-\bar{x}_{...})\left((\bar{x}_{ij.}-\bar{x}_{i..})+(x_{ijk}-\bar{x}_{ij.})\right)+\left((\bar{x}_{ij.}-\bar{x}_{i..})+(x_{ijk}-\bar{x}_{ij.})\right)^2\right)\\ &=\sum_{ijk}\left(\begin{array}{c}(\bar{x}_{i..}-\bar{x}_{...})^2+2(x_{i..}-\bar{x}_{...})\left((\bar{x}_{ij.}-\bar{x}_{i..})+(x_{ijk}-\bar{x}_{ij.})\right)\\ +\left((\bar{x}_{ij.}-\bar{x}_{i..})^2+2(\bar{x}_{ij.}-\bar{x}_{i..})(x_{ijk}-\bar{x}_{ij.})+(x_{ijk}-\bar{x}_{ij.})^2\right)\end{array}\right)\\ &=\sum_{ijk}(\bar{x}_{i..}-\bar{x}_{...})^2+\sum_{ijk}(\bar{x}_{ij.}-\bar{x}_{i..})^2+\sum_{ijk}(x_{ijk}-\bar{x}_{ij.})^2\\ &\quad+2\left[\sum_{ijk}(\bar{x}_{i..}-\bar{x}_{...})\left((\bar{x}_{ij.}-\bar{x}_{i..})+(x_{ijk}-\bar{x}_{ij.})\right)+\sum_{ijk}(\bar{x}_{ij.}-\bar{x}_{i..})(x_{ijk}-\bar{x}_{ij.})\right]\end{aligned}\tag{5.7.2095}$$

But, we have:

$$\sum_{ijk}(\bar{x}_{ij.}-\bar{x}_{i..})(x_{ijk}-\bar{x}_{ij.})=\sum_{ij}\left[(\bar{x}_{ij.}-\bar{x}_{i..})\underbrace{\sum_{k}(x_{ijk}-\bar{x}_{ij.})}_{=0}\right]=0\tag{5.7.2096}$$

and:

$$
\begin{aligned}
\sum_{ijk}(\bar{x}_{i..}-\bar{x}_{...})\left((\bar{x}_{ij.}-\bar{x}_{i..})+(x_{ijk}-\bar{x}_{ij.})\right) &= \sum_{ijk}(\bar{x}_{i..}-\bar{x}_{...})(\bar{x}_{ij.}-\bar{x}_{i..})+\sum_{ijk}(\bar{x}_{i..}-\bar{x}_{...})(x_{ijk}-\bar{x}_{ij.})\\
&= \sum_{ik}(\bar{x}_{i..}-\bar{x}_{...})\underbrace{\sum_{j}(\bar{x}_{ij.}-\bar{x}_{i..})}_{=0}+\sum_{i}\left[(\bar{x}_{i..}-\bar{x}_{...})\sum_{jk}(x_{ijk}-\bar{x}_{ij.})\right]\\
&= \sum_{i}\left[(\bar{x}_{i..}-\bar{x}_{...})\sum_{j}\underbrace{\sum_{k}(x_{ijk}-\bar{x}_{ij.})}_{=0}\right]
\end{aligned}
\tag{5.7.2097}
$$

So it remains in the end:

$$
\begin{aligned}
Q_T=\sum_{ijk}(x_{ijk}-\bar{x}_{...})^2 &= \sum_{ijk}(\bar{x}_{i..}-\bar{x}_{...})^2+\sum_{ijk}(\bar{x}_{ij.}-\bar{x}_{i..})^2+\sum_{ijk}(\bar{x}_{ijk}-\bar{x}_{ij.})^2\\
&= \underbrace{mn\sum_{i=1}^{M}(\bar{x}_{i..}-\bar{x}_{...})^2}_{Q_A}+\underbrace{n\sum_{i=1}^{M}\sum_{j=1}^{m}(\bar{x}_{ij.}-\bar{x}_{i..})^2}_{Q_{(A)B}}+\underbrace{\sum_{i=1}^{M}\sum_{j=1}^{m}\sum_{k=1}^{n}(\bar{x}_{ijk}-\bar{x}_{ij.})^2}_{Q_R}
\end{aligned}
\tag{5.7.2098}
$$

where the summation indices are not traditionally the same as the two-factor ANOVA with repetition (but this is a detail). If we compare with the variance decomposition of the two-factor ANOVA with repetition that was for recall:

$$
\begin{aligned}
Q_T &= \sum_{mij}(x_{mij}-\overline{\overline{x}}_{..})^2\\
&= \underbrace{nk\sum_{j}(\bar{x}_{.j}-\overline{\overline{x}}_{..})^2}_{Q_A}+\underbrace{nr\sum_{i}(\bar{x}_{i.}-\overline{\overline{x}}_{..})^2}_{Q_B}+\underbrace{n\sum_{ij}(\bar{x}_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\overline{\overline{x}}_{..})^2}_{Q_{A\times B}}+\underbrace{\sum_{mij}(x_{mij}-\bar{x}_{ij})^2}_{Q_R}
\end{aligned}
\tag{5.7.2099}
$$

We notice that the main difference lies in the fact we are not interested in the variance of the interaction of the two factors (which was the goal or say rather a... constraint!) and that we then did not have to make it appear.

Now come the quest of determining the degrees of freedom. For Q_T it's $\text{df}_T=Mmn-1$. For the term Q_A (principal factor), the number of degrees of freedom is the same as for the one-way ANOVA or the two-factor ANOVA with/without repetitions. That is $\text{df}_A=M-1$ degrees of freedom. For the residual term it also comes immediately that it is the same thing as the two-factor ANOVA with repetition and therefore that $\text{df}_R=Mm(n-1)=N-Mm$. Finally, the remaining term $Q_{(A)B}$ is new, but we can easily determine its number of degrees of freedom since we must have:

$$
\text{df}_T=\text{df}_A+\text{df}_{(A)B}+\text{df}_R
\tag{5.7.2100}
$$

Then it comes:

$$
Mmn-1=(M-1)+\text{df}_{(A)B}+Mm(n-1)
\tag{5.7.2101}
$$

Hence:

$$\begin{aligned}\mathrm{df}_{(A)B}&=(Mmn-1)-(M-1)-Mm(n-1)\\&=Mmn-1-M+1-Mmm+Mm=-M+Mm=M(m-1)\end{aligned}\tag{5.7.2102}$$

We then have the following table for the hierarchical ANOVA (according to the most common representation in statistical softwares) that we will comment after:

Sum of squares	χ^2 **df**	**Mean squares**	F	**Critical** F
$Q_A=mn\sum_{i=1}^{M}\sum_{ijk}(\bar{x}_{i..}-\bar{x}_{...})^2$ (due to the surface/bloc)	$M-1$	MSk$A=\dfrac{Q_A}{M-1}$	$\dfrac{\text{MSk}A}{\text{MSk}(A)B}$	$P(F>F_{M-1,Mm(n-1)})$
$Q_{(A)B}=n\sum_{i=1}^{M}\sum_{j=1}^{m}(\bar{x}_{ij.}-\bar{x}_{i..})^2$ (due to the plot in the whole plot)	$M(m-1)$	MSk$(A)B=\dfrac{Q_{(A)B}}{M(m-1)}$	$\dfrac{\text{MSk}(A)B}{\text{MSE}}$	$P(F>F_{M(m-1),Mm(n-1)})$
$Q_R=\sum_{i=1}^{M}\sum_{j=1}^{m}\sum_{k=1}^{n}(x_{ijk}-\bar{x}_{ij.})^2$ (residuals)	$Mm(n-1)$	MSE $=\dfrac{Q_R}{Mm(n-1)}$		
$Q_T=\sum_{ijk}(x_{ijk}-\bar{x}_{...})^2$ (total)	$Mmn-1$	MST $=\dfrac{Q_T}{Mmn-1}$		

Table 5.68: **Hierarchical (random) factor ANOVA table**

So we have two Fisher tests just like the two-factor canonical ANOVA which is relatively intuitive. On the other hand what often surprises the practitioners is the first Fisher test where we have:

$$F_{M-1,Mm(n-1)}=\frac{\text{MSk}A}{\text{MSk}(A)B}\tag{5.7.2103}$$

Instead of what we could have written a little bit too quick:

$$F_{M-1,Mm(n-1)}=\frac{\text{MSk}A}{\text{MSE}}\tag{5.7.2104}$$

So what must be understood here is that there is not one that is right and the other wrong!!!!!!!!! It is necessary to know what we are interested in to make the best choice. Thus, comparing the variance of the blocks with the residual variance of the treatments is not necessarily the most interesting for judging the equality of the average treatments between them. This is why it is better to compare the variance of the blocks with the variance of the treatments.

Therefore in most statistical softwares the table above is say to be the "ANOVA table for random factors" and the table below is the version for fixed factors (in case you really need to compare the variance of the blocks with the residuals variance of the treatments!):

Sum of squares	χ^2 df	Mean squares	F	Critical F
$Q_A=mn\sum_{i=1}^{M}\sum_{ijk}(\bar{x}_{i..}-\bar{x}_{...})^2$ (due to the surface/block)	$M-1$	$\text{MSk}A=\dfrac{Q_A}{M-1}$	$\dfrac{\text{MSk}A}{\text{MSE}}$	$P(F>F_{M-1,Mm(n-1)})$
$Q_{(A)B}=n\sum_{i=1}^{M}\sum_{j=1}^{m}(\bar{x}_{ij.}-\bar{x}_{i..})^2$ (due to the plot in the whole plot)	$M(m-1)$	$\text{MSk}(A)B=\dfrac{Q_{(A)B}}{M(m-1)}$	$\dfrac{\text{MSk}(A)B}{\text{MSE}}$	$P(F>F_{M(m-1),Mm(n-1)})$
$Q_R=\sum_{i=1}^{M}\sum_{j=1}^{m}\sum_{k=1}^{n}(\bar{x}_{ijk}-\bar{x}_{ij.})^2$ (residuals)	$Mm(n-1)$	$\text{MSE}=\dfrac{Q_R}{Mm(n-1)}$		
$Q_T=\sum_{ijk}(x_{ijk}-\bar{x}_{...})^2$ (total)	$Mmn-1$	$\text{MST}=\dfrac{Q_T}{Mmn-1}$		

Table 5.69: **Hierarchical (fixed) factor ANOVA table**

Remark

Let us recall that for an ANOVA with fixed factors without replications we had:

$$x_{ijk}=\mu+\tau_i+\beta_j+\varepsilon_{ij} \tag{5.7.2105}$$

with $i=1\ldots a$, $j=1\ldots b$ where μ is the overall mean, τ_i is the effect of the ith of treatment A, β_j is the effect of the jth level of treatment B (main effects) and ε_{ij} is the habitual random error following $\mathcal{N}(0,\sigma_\varepsilon^2)$.

For nested ANOVA, the β_j will be expanded $\alpha_j+\gamma_{ij}$ and the new model may contain an interaction term representing possible interaction between blocking factor and the treatments. Thus the factor effect model may be written:

$$\boxed{x_{ijk}=\mu+(\alpha_j+\gamma_{ij})+\tau_k+(\alpha\tau)_{ik}+\varepsilon_{ijl}} \tag{5.7.2106}$$

Now before we continue with some other major ANOVA type (especially "ANOVA with blocks" and "split-plot ANOVA") let us give in the form of a figure what we have seen so far and what we will see just after:

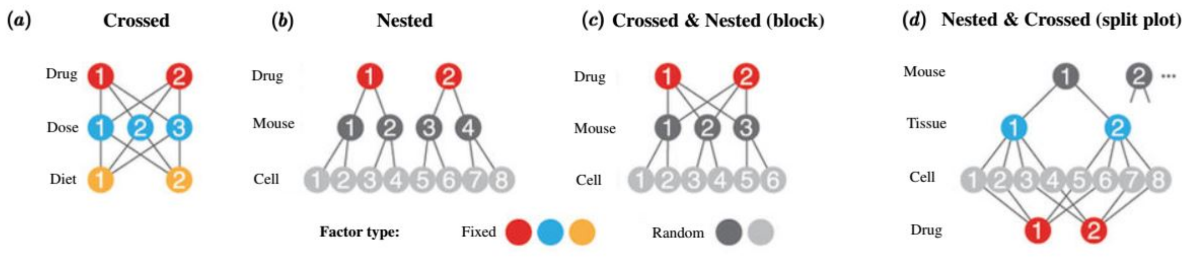


Figure 5.201: **Comparing different some major ANOVA structures**

- *(a)* A crossed design examines every combination of levels for each fixed factor
- *(b)* A nested design can progressively subreplicate a fixed factor with nested levels of a random factor that are unique to the level within which they are nested

(*c*) If a random factor can be reused for different levels of the treatment, it can be crossed with the treatment and modelled as a block

(*d*) A split plot design (also named "mixed-design") is a structure where the fixed effects (tissue, drug) are crossed (each combination of tissue and drug are tested) but themselves nested within replicates.

Thus, in a mixed-design ANOVA model, one factor (a fixed effects factor) is a between-subjects variable and the other (a random effects factor) is a within-subjects variable. Thus, overall, the model is a type of mixed effect model.

Remark

Let us recall that for an ANOVA with fixed factors without replications we had:

$$x_{ijk}=\mu+\tau_i+\beta_j+\varepsilon_{ij} \tag{5.7.2107}$$

with $i=1\ldots a$, $j=1\ldots b$ where μ is the overall mean, τ_i is the effect of the ith of treatment A, β_j is the effect of the jth level of treatment B (main effects) and ε_{ij} is the habitual random error following $\mathcal{N}(0,\sigma_\varepsilon^2)$.

For nested and crossed ANOVA, the β_j will be expanded $\alpha_j+\gamma_{ij}+(\alpha\gamma)_{ij}$ and to inclusion of certain block-treatment interaction terms. Thus the factor effect model may be written:

$$\boxed{x_{ijk}=\mu+(\alpha_j+\gamma_{ij}+(\alpha\gamma)_{ij})+\tau_k+(\alpha\tau)_{ik}+(\gamma\tau)_{jk}+\varepsilon_{ijl}} \tag{5.7.2108}$$

5.7.13.5.6 ANOVA with blocks

With a "randomized block design" (RBD), we have a characteristic of the units-of-analysis that we stratify (block) and then randomize into our treatment conditions within each block. For example, we could block on sex (male and female) and then randomly assign to a treatment and control condition separately for males and females, ensuring balance across the blocks in the number assigned to each group. In this design, we have one factor (treatment/control) and one block (male/female). This design also controls for any variance associated with the block (you would only want to use a block that you have good reason to believe is associated with the dependent variable).

The purpose of the blocking factor is to account for a nuisance factor and/or to reduce the error term used in performing the test for the significance of the treatment effect. For this reason, the significance of the block effect itself is not tested, nor are multiple comparisons done between fixed blocks. Otherwise, a one-way blocked ANOVA is analysed as a two-way ANOVA with no interactions and no replications. That means that the following ANOVA table:

Sum of squares (SSE)	χ^2 df	Mean squares	F	Critical F
$Q_A = k\sum_j (\bar{x}_j - \bar{\bar{x}})^2$	$k-1$	$\text{MSk}A = \dfrac{Q_A}{k-1}$	$\dfrac{\text{MSk}A}{\text{MSE}}$	$P(F > F_{k-1,(k-1)(r-1)})$
$Q_B = r\sum_i (\bar{x}_i - \bar{\bar{x}})^2$	$r-1$	$\text{MSk}B = \dfrac{Q_B}{r-1}$	$\dfrac{\text{MSk}B}{\text{MSE}}$	$P(F > F_{r-1,(k-1)(r-1)})$
$Q_R = \sum_{ij} (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}})^2$	$(k-1)(r-1)$	$\text{MSE} = \dfrac{Q_R}{(k-1)(n-1)}$		
$Q_T = \sum_{ij} (x_{ij} - \bar{\bar{x}})^2$	$N-1$	$\text{MST} = \dfrac{Q_T}{N-1}$		

Table 5.70: **Two-way fixed factor ANOVA table without repetitions**

becomes:

Sum of squares (SSE)	χ^2 df	Mean squares	F	Critical F
$Q_t = k\sum_j (\bar{x}_j - \bar{\bar{x}})^2$ treatments	$k-1$	$\text{MSk}t = \dfrac{Q_t}{k-1}$	$\dfrac{\text{MSk}t}{\text{MSE}}$	$P(F > F_{k-1,(k-1)(r-1)})$
$Q_b = r\sum_i (\bar{x}_i - \bar{\bar{x}})^2$ blocks	$r-1$	$\text{MSk}b = \dfrac{Q_b}{r-1}$	$\dfrac{\text{MSk}b}{\text{MSE}}$	$P(F > F_{r-1,(k-1)(r-1)})$
$Q_R = \sum_{ij} (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}})^2$	$(k-1)(r-1)$	$\text{MSE} = \dfrac{Q_R}{(k-1)(n-1)}$		
$Q_T = \sum_{ij} (x_{ij} - \bar{\bar{x}})^2$	$N-1$	$\text{MST} = \dfrac{Q_T}{N-1}$		

Table 5.71: **One-way blocked ANOVA table**

In practical situations it is sometimes not easy to decide whether block effects should be treated as random effects. For example, are the blocks in a field experiment randomly selected form a larger population of blocks? Most likely they were the only blocks available for the experiment. Of, if the experiment is replicated over two years (setting up a GRBD with nested blocking factors),are those years randomly selected? Certainly not, but still the researcher may want to consider the as "random" years. But if one years turns out to be a dry year and the other to be a wet year, then clearly we have an intrinsic blocking factor with fixed effects. There are obviously many variations of this discussion and thus this question becomes rather philosophical and often controversial. In general, we prefer to make the analysis once by considering them as fixed and once as random (as it is really quick done with computer software), and see if the conclusions are the same.

For two samples, one-way blocked ANOVA is equivalent to the two-sample paired T-test.

5.7.13.5.7 Latin Square ANOVA without repetitions

A "Latin Square ANOVA" is a three factor nested ANOVA but with the subtility that the third nested one has not each of its levels repeated in the levels of the two first but only one time and in such a way that each of its levels appears only one time at each row and each column of the parent level:

1	2	3
3	1	2
2	3	1

4	3	1	2
3	4	2	1
1	2	4	3
2	1	3	4

1	2	4	3	5
4	5	2	1	3
3	4	1	5	2
2	3	5	4	1
5	1	3	2	4

Figure 5.202: Latin squares of orders 3, 4 and 5

The number of level of each factor is imposed as being in equal quantity that is why there is the term "square" in "Latin Square" design! The term "Latin" comes from the fact that no term is repeated on the same line or on the same column! We then logically speak of "balanced Latin square design" or just simply of "Latin square design" (LSD).

So this design (attributed to Leonard Euler in 11783 according to holocene calendar), represents, in some sense, the simplest form of a row-column design as it is used for comparing K treatments in K rows and K columns, where rows and columns represent two blocking factors. They were proposed as experimental design by Ronald Fisher in 11925 (holocene calendar), although François Cretté De Palluel (11788 according to holocene calendar) already utilized the idea of a 4×4 Latin square design for an agricultural experiment.

A complete illustration of such as design can be given by the following example (typical in Gage R&R crossed studies without replications!):

	Operators				
Batch	1	2	3	4	5
1	A=24	B=20	C=19	D=24	E=24
2	B=17	C=24	D=30	E=27	A=36
3	C=18	D=38	E=26	A=27	B=21
4	D=26	E=31	A=26	B=23	C=22
5	E=22	A=30	B=20	C=29	D=31

Table 5.72: Famous Latin square design for the Propellant model of Douglas C. Montgomery

We see why the Latin square ANOVA are relatively rare in practice because they assume that for each object we have the same number of combinations of tests when in reality we could have less or more. Moreover, the Latin squares ANOVA are often designated as incomplete plans, or "incomplete block designs[81]", and thus in extenso economic because if we really had to make all the combinations for a Latin square of dimension n, the number of total combinations would be n^3 whereas in reality the Latin square represents only n^2.

Therefore we can already observe that:

1. The order of performing the trials is random
2. We have many trials (as the example above should have $5 \cdot 5 \cdot 5 = 125$ trials if it were complete)

---

[81] So keep in mind that it is a randomized block design in which every treatment is not present in every block

Note also that for a given dimension a Latin square is not unique. Indeed the following Latin square:

	Cars			
Drivers	1	2	3	4
1	A	B	D	C
2	D	C	A	B
3	B	D	C	A
4	C	A	B	D

is equivalent to that one:

	Cars			
Drivers	1	2	3	4
1	A	B	D	C
2	C	D	B	A
3	B	A	C	D
4	D	C	A	B

The real interest of the Latin square ANOVA is that they are certainly economic since they are incomplete but mainly because they are orthogonal in the sense that they contain the best way of doing the tests by minimizing the quantity of these last ones. Indeed, we see in a Latin square that each letter appears only once in each line and each column (hence the origin of the name!) in order to minimize the error (mathematically there is therefore "orthogonality").

It is customary to represent a Latin square ANOVA in the following more general way:

	1st Factor				
2nd Factor	1	2	3	4	Average of rows
I	y_{111}	y_{122}	y_{134}	y_{143}	$\bar{y}_{1..}$
II	y_{214}	y_{223}	y_{231}	y_{341}	$\bar{y}_{2..}$
III	y_{312}	y_{324}	y_{333}	y_{341}	$\bar{y}_{3..}$
IV	y_{413}	y_{421}	y_{432}	y_{444}	$\bar{y}_{4..}$
Average of columns	$\bar{y}_{.1.}$	$\bar{y}_{.2.}$	$\bar{y}_{.3.}$	$\bar{y}_{.4.}$	$\bar{y}_{...}$

And obviously notice that the number of rows is (obviously!) always equal to the number of columns and for such designs, we also speak of "symmetric designs".

Let us prove that we have:

$$\begin{aligned}\mathrm{Q}_T=&\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{k=1}^{K}\left(y_{ijk}-\bar{y}_{...}\right)^2\\ =&\underbrace{K\sum_{i=1}^{K}\left(\bar{y}_{i..}-\bar{y}_{...}\right)^2}_{\text{SSE}_{\text{rows}}}+\underbrace{K\sum_{j=1}^{K}\left(\bar{y}_{.j.}-\bar{y}_{...}\right)^2}_{\text{SSE}_{\text{columns}}}+\underbrace{K\sum_{k=1}^{K}\left(\bar{y}_{..k}-\bar{y}_{...}\right)^2}_{\text{SSE}_{\text{treatements}}}\\ &+\underbrace{\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{k=1}^{K}\left(y_{ijk}-\bar{y}_{i..}-\bar{y}_{.j.}-\bar{y}_{..k}+2\bar{y}_{...}\right)^2}_{\text{SSE}_{\text{residual}}}\\ =&\,\mathrm{Q}_R+\mathrm{Q}_C+\mathrm{Q}_{RC}+\mathrm{Q}_r\end{aligned}\tag{5.7.2109}$$

where R stands for "rows", C for "columns" and r for "residuals".

Well let's go in joy and good mood ...! We first write:

$$\begin{aligned}\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{k=1}^{K}\left(y_{ijk}-\bar{y}_{...}\right)^2&=\sum_{i,j,k}\left(y_{ijk}-\bar{y}_{...}\right)^2=\sum_{i,j,k}\left(y_{ijk}+\bar{y}_{i..}-\bar{y}_{i..}+\bar{y}_{.j.}-\bar{y}_{.j.}+\bar{y}_{..k}-\bar{y}_{..k}-\bar{y}_{...}\right)^2\\ &=\sum_{i,j,k}\left(y_{ijk}-\bar{y}_{i..}-\bar{y}_{.j.}-\bar{y}_{..k}+2\bar{y}_{...}+\bar{y}_{i..}-\bar{y}_{...}+\bar{y}_{.j.}-\bar{y}_{...}+\bar{y}_{..k}-\bar{y}_{...}\right)^2\\ &=\sum_{i,j,k}\left(y_{ijk}-\bar{y}_{i..}-\bar{y}_{.j.}-\bar{y}_{..k}+2\bar{y}_{...}\right)^2+\sum_{i,j,k}\left(\bar{y}_{i..}-\bar{y}_{...}+\bar{y}_{.j.}-\bar{y}_{...}+\bar{y}_{..k}-\bar{y}_{...}\right)^2\\ &+2\sum_{i,j,k}\left(y_{ijk}-\bar{y}_{i..}-\bar{y}_{.j.}-\bar{y}_{..k}+2\bar{y}_{...}\right)\left(\bar{y}_{i..}+\bar{y}_{.j.}+\bar{y}_{..k}-3\bar{y}_{...}\right)\end{aligned}\tag{5.7.2110}$$

But:

$$\sum_{i,j,k}\left(y_{ijk}-\bar{y}_{i..}-\bar{y}_{.j.}-\bar{y}_{..k}+2\bar{y}_{...}\right)\left(\bar{y}_{i..}+\bar{y}_{.j.}+\bar{y}_{..k}-3\bar{y}_{...}\right)=0\tag{5.7.2111}$$

Indeed, let's start by noticing that:

$$\sum_{i,j,k}y_{ijk}\cdot\bar{y}_{i..}=\sum_{i}\bar{y}_{i..}\sum_{j,k}y_{ijk}=\sum_{i}\bar{y}_{i..}\cdot K\cdot\bar{y}_{i..}=K\sum_{i}(\bar{y}_{i..})^2\tag{5.7.2112}$$

The same calculations show that:

$$\sum_{i,j,k}y_{ijk}\cdot\bar{y}_{.j.}=K\sum_{j}(\bar{y}_{.j.})^2\qquad\text{and}\qquad\sum_{i,j,k}y_{ijk}\cdot\bar{y}_{..k}=K\sum_{k}(\bar{y}_{..k})^2\tag{5.7.2113}$$

Furthermore:

$$\sum_{i,j,k}(-\bar{y}_{i..})\bar{y}_{i..}=-\sum_{i}\sum_{j,k}(\bar{y}_{i..})^2=-K\sum_{i}(\bar{y}_{i..})^2\tag{5.7.2114}$$

and again we have the same result for the indices j and k. That is to say:

$$\sum_{i,j,k}(-\bar{y}_{.j.})\bar{y}_{.j.}=-K\sum_{j}(\bar{y}_{.j.})^2\quad\text{and}\quad\sum_{i,j,k}(-\bar{y}_{..k})\bar{y}_{..k}=-K\sum_{k}(\bar{y}_{..k})^2\tag{5.7.2115}$$

Therefore:

$$\begin{gathered}\sum_{i,j,k}(y_{ijk}-\bar{y}_{i..}-\bar{y}_{.j.}-\bar{y}_{..k}+2\bar{y}_{...})(\bar{y}_{i..}+\bar{y}_{.j.}+\bar{y}_{..k}-3\bar{y}_{...})=\\ \sum_{i,j,k}\left[-3\bar{y}_{...}\cdot y_{ijk}-2\bar{y}_{i..}\bar{y}_{.j.}-2\bar{y}_{i..}\bar{y}_{..k}-2\bar{y}_{.j.}\bar{y}_{..k}+5\bar{y}_{...}(\bar{y}_{i..}+\bar{y}_{.j.}+\bar{y}_{..k})-6(\bar{y}_{...})^2\right]\end{gathered} \tag{5.7.2116}$$

Now, note that:

$$\sum_{i,j,k}\bar{y}_{i..}\cdot\bar{y}_{.j.}=\sum_{i}\bar{y}_{i..}\sum_{j,k}\bar{y}_{.j.}=K\bar{y}\sum_{i}\bar{y}_{i..}=K^2(\bar{y})^2 \tag{5.7.2117}$$

as:

$$\sum_{j,k}\bar{y}_{.j.}=K\bar{y} \tag{5.7.2118}$$

Here again we also have with the same calculations:

$$\sum_{i,j,k}\bar{y}_{i..}\cdot\bar{y}_{..k}=K^2(\bar{y})^2 \quad \text{and} \quad \sum_{i,j,k}\bar{y}_{.j.}\cdot\bar{y}_{..k}=K^2(\bar{y})^2 \tag{5.7.2119}$$

Furthermore:

$$\sum_{i,j,k}(\bar{y}_{...})^2=K^2(\bar{y}_{...})^2 \tag{5.7.2120}$$

Therefore:

$$\begin{gathered}\sum_{i,j,k}(y_{ijk}-\bar{y}_{i..}-\bar{y}_{.j.}-\bar{y}_{..k}+2\bar{y}_{...})(\bar{y}_{i..}+\bar{y}_{.j.}+\bar{y}_{..k}-3\bar{y}_{...})=\\ -12K^2(\bar{y}_{...})^2+\sum_{i,j,k}\left[-3\bar{y}_{...}\cdot y_{ijk}+5\bar{y}_{...}(\bar{y}_{i..}+\bar{y}_{.j.}+\bar{y}_{..k})\right]\end{gathered} \tag{5.7.2121}$$

We have:

$$\sum_{i,j,k}\bar{y}_{...}\cdot y_{ijk}=\bar{y}_{...}\sum_{i,j,k}y_{ijk}=K^2(\bar{y}_{...})^2 \tag{5.7.2122}$$

as:

$$\sum_{i,j,k}y_{ijk}=K\bar{y}_{...} \tag{5.7.2123}$$

Furthermore:

$$\sum_{i,j,k}\bar{y}_{...}\bar{y}_{i..}=\sum_{i}\bar{y}_{...}\sum_{j,k}\bar{y}_{i..}=\sum_{i}\bar{y}_{...}K\cdot\bar{y}_{i..}=K^2(\bar{y}_{...})^2 \tag{5.7.2124}$$

as:

$$\sum_{i}\bar{y}_{i..}=K\bar{y} \tag{5.7.2125}$$

The same result holds for the indexes j and k, i.e.:

$$\sum_{i,j,k} \bar{y}_{...}\bar{y}_{.j.} = K^2(\bar{y}_{...})^2 \qquad \text{and} \qquad \sum_{i,j,k} \bar{y}_{...}\bar{y}_{..k} = K^2(\bar{y}_{...})^2 \tag{5.7.2126}$$

Hence:

$$\sum_{i,j,k} (y_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{y}_{..k} + 2\bar{y}_{...})(\bar{y}_{i..} + \bar{y}_{.j.} + \bar{y}_{..k} - 3\bar{y}_{...}) = 0 \tag{5.7.2127}$$

Back to the beginning...

Then it remains to us:

$$\sum_{i,j,k} (y_{ijk} - \bar{y}_{...})^2 = \sum_{i,j,k} (y_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{y}_{..k} + 2\bar{y}_{...})^2 + \sum_{i,j,k} (\bar{y}_{i..} - \bar{y}_{...} + \bar{y}_{.j.} - \bar{y}_{...} + \bar{y}_{..k} - \bar{y}_{...})^2 \tag{5.7.2128}$$

Let's evaluate the double products in the last sum:

$$\sum_{i,j,k} (\bar{y}_{i..} - \bar{y}_{...})(\bar{y}_{.j.} - \bar{y}_{...}) = \sum_{i,j,k} \bar{y}_{i..}\bar{y}_{.j.} - \sum_{i,j,k} \bar{y}_{...}\bar{y}_{i..} - \sum_{i,j,k} \bar{y}_{...}\bar{y}_{.j.} + \sum_{i,j,k} (\bar{y}_{...})^2 = 0 \tag{5.7.2129}$$

because as seen previously, all these are worth $K^2(\bar{y})^2$.

The other double products are also worth 0 by the same arguments. So:

$$\begin{aligned}\sum_{i,j,k} (y_{ijk} - \bar{y}_{...})^2 =& \sum_{i,j,k} (y_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{y}_{..k} + 2\bar{y}_{...})^2 \\ &+ \sum_{i,j,k} (y_{i..} - \bar{y}_{...})^2 + \sum_{i,j,k} (y_{.j.} - \bar{y}_{...})^2 + \sum_{i,j,k} (y_{..k} - \bar{y}_{...})^2\end{aligned} \tag{5.7.2130}$$

To finish, let us note that:

$$\sum_{i,j,k} (\bar{y}_{i..} - \bar{y}_{...})^2 = \sum_{i} \sum_{j,k} (\bar{y}_{i..} - \bar{y}_{...})^2 = K \sum_{i} (\bar{y}_{i..} - \bar{y}_{...})^2 \tag{5.7.2131}$$

and the same for j and k. Therefore:

$$\begin{aligned}\sum_{i,j,k} (y_{ijk} - \bar{y}_{...})^2 =& K \sum_{i} (\bar{y}_{i..} - \bar{y}_{...})^2 + K \sum_{j} (\bar{y}_{.j.} - \bar{y}_{...})^2 + K \sum_{k} (\bar{y}_{..k} - \bar{y}_{...})^2 \\ &+ \sum_{i,j,k} (y_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{y}_{..k} + 2\bar{y}_{...})^2\end{aligned} \tag{5.7.2132}$$

If we denote by K the dimension of a Latin square, for each of the above terms we have the following degrees of freedom:

$$K^2 - 1 = (K - 1) + (K - 1) + (K - 1) + (K - 1)(K - 2) \tag{5.7.2133}$$

We then have the following ANOVA table:

Sum of squares	χ^2 **df**	**Mean squares**	F	**Critical** F
$Q_R = K\sum_i (\bar{y}_{i..} - \bar{y}_{...})^2$ (rows)	$K-1$	$\text{MSk}R = \dfrac{Q_R}{K-1}$	$\dfrac{\text{MSk}R}{\text{MSE}}$	$P(F > F_{K-1,(K-1)(K-2)})$
$Q_C = K\sum_j (\bar{y}_{.j.} - \bar{y}_{...})^2$ (columns)	$K-1$	$\text{MSk}C = \dfrac{Q_C}{K-1}$	$\dfrac{\text{MSk}C}{\text{MSE}}$	$P(F > F_{K-1,(K-1)(K-2)})$
$Q_{RC} = K\sum_k (\bar{y}_{..k} - \bar{y}_{...})^2$ (treatments)	$K-1$	$\text{MSk}RC = \dfrac{Q_{RC}}{K-1}$	$\dfrac{\text{MSk}RC}{\text{MSE}}$	$P(F > F_{K-1,(K-1)(K-2)})$
$Q_r = \sum_{i,j,k} (y_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{y}_{..k} + 2\bar{y}_{...})^2$ (residuals)	$(K-1)(K-2)$	$\text{MSE} = \dfrac{Q_r}{(K-1)(K-2)}$		
$Q_T = \sum_{i,j,k} (y_{ijk} - \bar{y}_{...})^2$ (total)	K^2-1	$\text{MST} = \dfrac{\text{Q}_\text{T}}{K^2-1}$		

Table 5.73: **Latin square ANOVA table**

Remark

The factor effect model for this design is:

$$\boxed{y_{ijk} = \mu + \alpha_i + \tau_j + \beta_k + \varepsilon_{ijk}} \tag{5.7.2134}$$

where $i, j, k = 1 \ldots N$ and y_{ijk} is the observation in the ith row and kth column for the jth treatment, μ is the overall mean, α_i is the ith row effect, τ_j is the jth treatment effect, β_k is the kth column effect, and ε_{ijk} is a random error.

Notice that only two out of the three subscripts are necessary to identify an observation in the latin square design and that this model is (obviously) purely additive. The assumptions on the effect terms are $\sum_{i=1}^N \alpha_i = 0$, $\sum_{j=1}^N \tau_j = 0$ and $\sum_{k=1}^N \beta_k = 0$.

We are interested (mainly!) for zero treatment effect, that is, $H_0 : \tau_1 = \tau_2 = \ldots = 0$ versus $H_1 : \tau_j \neq 0$.

Keep in mind that the advantage of Latin square designs are:

- They allow experiments with a small number of runs (more economical than crossing all factors)
- Easy to analyse

The disadvantages are:

- The number of levels of each blocking variable must equal the number of levels of the treatment factor
- They assume that there are no interactions between the blocking variables or between the treatment variable and the blocking variable (may be unrealistic)
- It is say that squares smaller than 4×4 or 5×5 generally have too few replications for a desirable level of precision (low power)

- It can be proved that 6×6 Latin square designs are impossible (the proof is actually not in this textbook!)

Remarks

R1. The Latin square designs are part of what we name also sometimes "space filling design" (family that also contains spherical designs, uniform designs with minimum potential designs, or maximum entropy designs or IMSE (integrated mean-square error) Gaussian processes).

R2. There exist also rectangular Latin design for information...

5.7.13.5.8 Graeco-Latin Square ANOVA

A "Graeco-Latin Square ANOVA" is an extension of a Latin square used when the number of block is greater than 2. Concretely, a Graeco-Latin square ANOVA can control up to two nuisance factors and one treatment factor (three sources of extraneous variability...) as represented below:

	Column			
Row	1	2	3	4
1	Aα	Bβ	Cγ	Dδ
2	Bδ	Aγ	Dβ	Cα
3	Cβ	Dα	Aδ	Bγ
4	Dγ	Cδ	Bα	Aβ

Table 5.74: **Graeco-Latin square of order** 4

Use of these designs results in exceptional saving of time and means, as the total number of design points has been drastically reduced as if we consider n level for each of the three factor we have only n^2 experimental points, when compared to n^4 (since there are four factors!!!!) design points in full factorial design. However similar to Latin squares, these designs may be used only when interactions are statistically insignificant.

To figure out a Graeco-Latin square, we consider initially a Latin square $K \otimes K$ on which we superimposed another Latin square $K \otimes K$ which this time is denoted by Greek letters. Afterwards in a second time during the superposition, we must ensure that the following property is respected, "each row and each column can only contain couple of letters (Latin, Greek) that are distincts". Thus, if this property is verified (as shown in the diagram below) we say that the two orthogonal Latin squares are orthogonals:

4×4 **Latin square**
Latin letters

A	B	C	D
B	A	D	C
C	D	A	B
D	C	B	A

4×4 **Latin square**
Greek letters

α	β	γ	δ
δ	γ	β	α
β	α	δ	γ
γ	δ	α	β

4×4 **Graeco-Latin square**

	Columns			
Rows	1	2	3	4
1	Aα	Bβ	Cγ	Dδ
2	Bδ	Aγ	Dβ	Cα
3	Cβ	Dα	Aδ	Bγ
4	Dγ	Cδ	Bα	Aβ

Figure 5.203: **Merge of two Latin square designs**

Remark

The Graeco-Latin square designs are also part of what we name also sometimes "space filling design" as for the Latin square designs.

As you can notice, the Graeco-Latin square design allows the analysis of 4 factors (row, column, Latin letter, Greek letter), each factor having K levels for only $K \otimes K$ or K^2 possible combinations.

For the need of the analysis of variance, we have to define certain variables:

- K the number of levels for each factor
- $N = K^2$, the total number of observations $y_{i,j,k,l}$ in the Graeco-Latin square, where i is the i-th level of the factor represented in row, j the j-th level of the factor represented in column, k the k-th level of the factor represented by Latin letters and the l the l-th level of the factor represented by Greek letters.

Thus, one can formulate as follows:

- The sum and the average of observations:

$$y_{....} = \sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{l=1}^{K} y_{ijkl} = \sum_{i=1}^{K}\sum_{j=1}^{K} y_{ij..} \qquad \bar{y}_{....} = \frac{y_{....}}{K^2} \tag{5.7.2135}$$

- The sum and average for each row:

$$y_{i...} = \sum_{j=1}^{K} y_{ij..} \qquad \bar{y}_{i...} = \frac{y_{i...}}{K} \tag{5.7.2136}$$

- The sum and average for each columns:

$$y_{.j..} = \sum_{i=1}^{K} y_{ij..} \qquad \bar{y}_{.j..} = \frac{y_{.j..}}{K} \tag{5.7.2137}$$

- The sum and average for each Latin letter:

$$y_{..k.} = \sum_{i=1}^{K}\sum_{j=1}^{K} y_{ijk.} \qquad \bar{y}_{..k.} = \frac{y_{..k.}}{K} \tag{5.7.2138}$$

- The sum and average for each Greek letter:

$$y_{...l} = \sum_{i=1}^{K}\sum_{j=1}^{K} y_{ij.l} \qquad \bar{y}_{...l} = \frac{y_{...l}}{K} \tag{5.7.2139}$$

For the development, we will prefer this notation $\sum\limits_{i,j,k,l}^{K}$ to that one $\sum\limits_{i=1}^{K}\sum\limits_{j=1}^{K}\sum\limits_{k=1}^{K}\sum\limits_{l=1}^{K}$ in order to simplify the notations.

Let us deal now with the decomposition of the total variance! Let us prove that the latter:

$$\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{l=1}^{K}(y_{ijkl}-\bar{y}_{....})^2=\sum_{i,j,k,l}^{K}(y_{ijkl}-\bar{y}_{....})^2 \qquad (5.7.2140)$$

can be written:

$$\begin{aligned}\sum_{i,j,k,l}^{K}(y_{ijkl}-\bar{y}_{....})^2 &= K\sum_{i}^{K}(\bar{y}_{i...}-\bar{y}_{....})^2+K\sum_{j}^{K}(\bar{y}_{.j..}-\bar{y}_{....})^2+K\sum_{k}^{K}(\bar{y}_{..k.}-\bar{y}_{....})^2+K\sum_{l}^{K}(\bar{y}_{...l}-\bar{y}_{....})^2\\ &+\sum_{i,j,k,l}^{K}(y_{ijkl}-\bar{y}_{i...}-\bar{y}_{.j..}-\bar{y}_{..k.}-\bar{y}_{...l}+3\bar{y}_{....})^2\end{aligned}$$

For that, let us first put:

$$\sum_{i,j,k,l}^{K}(y_{ijkl}-\bar{y}_{....})^2=\sum_{i,j,k,l}^{K}(y_{ijkl}-\bar{y}_{i...}+\bar{y}_{i...}-\bar{y}_{.j..}+\bar{y}_{.j..}-\bar{y}_{..k.}+\bar{y}_{..k.}-\bar{y}_{...l}+\bar{y}_{...l}-\bar{y}_{....})^2 \qquad (5.7.2141)$$

Let's put some $\bar{y}_{....}$ in it like this:

$$\begin{aligned}&=\sum_{i,j,k,l}^{K}(y_{ijkl}-\bar{y}_{i...}-\bar{y}_{.j..}-\bar{y}_{..k.}-\bar{y}_{...l}+3\bar{y}_{....}+\bar{y}_{i...}-\bar{y}_{....}+\bar{y}_{.j..}-\bar{y}_{....}+\bar{y}_{..k.}-\bar{y}_{....}+\bar{y}_{...l}-\bar{y}_{....})^2\\ &=\sum_{i,j,k,l}^{K}[(y_{ijkl}-\bar{y}_{i...}-\bar{y}_{.j..}-\bar{y}_{..k.}-\bar{y}_{...l}+3\bar{y}_{....})+(\bar{y}_{i...}-\bar{y}_{....}+\bar{y}_{.j..}-\bar{y}_{....}+\bar{y}_{..k.}-\bar{y}_{....}+\bar{y}_{...l}-\bar{y}_{....})]^2\end{aligned} \qquad (5.7.2142)$$

At this level, we fall back on our famous remarkable identity $(a+b)^2=a^2+b^2+2ab$. Thus in the framework our development, we get:

$$\begin{aligned}&=\sum_{i,j,k,l}^{K}\Big[(y_{ijkl}-\bar{y}_{i...}-\bar{y}_{.j..}-\bar{y}_{..k.}-\bar{y}_{...l}+3\bar{y}_{....})^2+(\bar{y}_{i...}-\bar{y}_{....}+\bar{y}_{.j..}-\bar{y}_{....}+\bar{y}_{..k.}-\bar{y}_{....}+\bar{y}_{...l}-\bar{y}_{....})\\ &\quad+2\left(\bar{y}_{i...}+\bar{y}_{.j..}+\bar{y}_{..k}+\bar{y}_{...l}-4\bar{y}_{....}\right)\left(y_{ijkl}-\bar{y}_{i...}-\bar{y}_{.j..}-\bar{y}_{..k.}-\bar{y}_{...l}+3\bar{y}_{....}\right)\Big]\\ &=\sum_{i,j,k,l}^{K}(y_{ijkl}-\bar{y}_{i...}-\bar{y}_{.j..}-\bar{y}_{..k.}-\bar{y}_{...l}+3\bar{y}_{....})^2+\sum_{i,j,k,l}^{K}(\bar{y}_{i...}-\bar{y}_{....}+\bar{y}_{.j..}-\bar{y}_{....}+\bar{y}_{..k.}-\bar{y}_{....}+\bar{y}_{...l}-\bar{y}_{....})^2\\ &\quad+2\sum_{i,j,k,l}^{K}\left(\bar{y}_{i...}+\bar{y}_{.j..}+\bar{y}_{..k.}+\bar{y}_{...l}-4\bar{y}_{....}\right)\left(y_{ijkl}-\bar{y}_{i...}-\bar{y}_{.j..}-\bar{y}_{..k.}-\bar{y}_{...l}+3\bar{y}_{....}\right)\end{aligned}$$

Let us consider now the 3th member of this expression $2\sum ab$:

$$\sum_{i,j,k,l}^{K}\left(\bar{y}_{i...}+\bar{y}_{.j..}+\bar{y}_{..k}+\bar{y}_{...l}-4\bar{y}_{....}\right)\left(y_{ijkl}-\bar{y}_{i...}-\bar{y}_{.j..}-\bar{y}_{..k.}-\bar{y}_{...l}+3\bar{y}_{....}\right) \qquad (5.7.2143)$$

And let us show that the result of that latter is equal to 0. Let's start by developing it as below:

$$\begin{aligned}
&= \sum_{i,j,k,l}^{K} \bar{y}_{i...}\left(y_{ijkl} - \bar{y}_{i...} - \bar{y}_{.j..} - \bar{y}_{..k.} - \bar{y}_{...l} + 3\bar{y}_{....}\right)^{(a)}\\
&\quad + \sum_{i,j,k,l}^{K} \bar{y}_{.j..}\left(y_{ijkl} - \bar{y}_{i...} - \bar{y}_{.j..} - \bar{y}_{..k.} - \bar{y}_{...l} + 3\bar{y}_{....}\right)^{(b)}\\
&\quad + \sum_{i,j,k,l}^{K} \bar{y}_{..k.}\left(y_{ijkl} - \bar{y}_{i...} - \bar{y}_{.j..} - \bar{y}_{..k.} - \bar{y}_{...l} + 3\bar{y}_{....}\right)^{(c)}\\
&\quad + \sum_{i,j,k,l}^{K} \bar{y}_{...l}\left(y_{ijkl} - \bar{y}_{i...} - \bar{y}_{.j..} - \bar{y}_{..k.} - \bar{y}_{...l} + 3\bar{y}_{....}\right)^{(d)}\\
&\quad - 4\sum_{i,j,k,l}^{K} \bar{y}_{....}\left(y_{ijkl} - \bar{y}_{i...} - \bar{y}_{.j..} - \bar{y}_{..k.} - \bar{y}_{...l} + 3\bar{y}_{....}\right)^{(e)}
\end{aligned}$$

Let us develop the first term above denoted by the exponent ${}^{(a)}$:

$$\begin{aligned}
\sum_{i,j,k,l}^{K} \bar{y}_{i...}\left(y_{ijkl} - \bar{y}_{i...} - \bar{y}_{.j..} - \bar{y}_{..k.} - \bar{y}_{...l} + 3\bar{y}_{....}\right) &= \sum_{i,j,k,l}^{K} \bar{y}_{i...}\cdot y_{ijkl} - \sum_{i,j,k,l}^{K} \bar{y}_{i...}\cdot\bar{y}_{i...} - \sum_{i,j,k,l}^{K} \bar{y}_{i...}\cdot\bar{y}_{.j..}\\
&\quad - \sum_{i,j,k,l}^{K} \bar{y}_{i...}\cdot\bar{y}_{..k.} - \sum_{i,j,k,l}^{K} \bar{y}_{i...}\cdot\bar{y}_{...l} + 3\sum_{i,j,k,l}^{K} \bar{y}_{i...}\cdot\bar{y}_{....}
\end{aligned}$$

Let us evaluate each of the six terms above separately:

$$\begin{aligned}
&\sum_{i,j,k,l}^{K} \bar{y}_{i...}\cdot y_{ijkl} = \sum_{i}^{K} \bar{y}_{i...}\cdot\sum_{j,k,l}^{K} y_{ijkl} = K\cdot\bar{y}_{i...}\cdot\bar{y}_{i...} = K\sum_{i}^{K}\left(\bar{y}_{i...}\right)^2\\
&-\sum_{i,j,k,l}^{K} \bar{y}_{i...}\cdot\bar{y}_{i...} = -K\sum_{i}^{K}\left(\bar{y}_{i...}\right)^2\\
&-\sum_{i,j,k,l}^{K} \bar{y}_{i...}\cdot\bar{y}_{.j..} = -\sum_{i}^{K} \bar{y}_{i...}\cdot\sum_{j,k,l}^{K} y_{.j..} = -K\bar{y}_{....}\cdot K\bar{y}_{....} = -K^2\left(\bar{y}_{....}\right)^2\\
&-\sum_{i,j,k,l}^{K} \bar{y}_{i...}\cdot\bar{y}_{..k.} = \ldots = -K^2\left(\bar{y}_{....}\right)^2\\
&-\sum_{i,j,k,l}^{K} \bar{y}_{i...}\cdot\bar{y}_{...l} = \ldots = -K^2\left(\bar{y}_{....}\right)^2\\
&+3\sum_{i,j,k,l}^{K} \bar{y}_{i...}\cdot\bar{y}_{....} = \ldots = +3K^2\left(\bar{y}_{....}\right)^2
\end{aligned}\tag{5.7.2144}$$

So finally we get that the original term denoted by the exponent ${}^{(a)}$:

$$\sum_{i,j,k,l}^{K} \bar{y}_{i...}\left(y_{ijkl} - \bar{y}_{i...} - \bar{y}_{.j..} - \bar{y}_{..k.} - \bar{y}_{...l} + 3\bar{y}_{....}\right)\tag{5.7.2145}$$

is equal to:

$$\sum_{i,j,k,l}^{K} \bar{y}_{i...} \cdot y_{ijkl} - \sum_{i,j,k,l}^{K} \bar{y}_{i...} \cdot \bar{y}_{i...} - \sum_{i,j,k,l}^{K} \bar{y}_{i...} \cdot \bar{y}_{.j..} - \sum_{i,j,k,l}^{K} \bar{y}_{i...} \cdot \bar{y}_{..k.} - \sum_{i,j,k,l}^{K} \bar{y}_{i...} \cdot \bar{y}_{...l} + 3\sum_{i,j,k,l}^{K} \bar{y}_{i...} \cdot \bar{y}_{....} \tag{5.7.2146}$$

That gives:

$$K\sum_{i}^{K} \left(\bar{y}_{i...}\right)^2 - K\sum_{i}^{K} \left(\bar{y}_{i...}\right)^2 - K^2\left(\bar{y}_{....}\right)^2 - K^2\left(\bar{y}_{....}\right)^2 - K^2\left(\bar{y}_{....}\right)^2 + 3K^2\left(\bar{y}_{....}\right)^2 = 0 \tag{5.7.2147}$$

Thus, by the same boring mathematical process, we can prove that each of the other four remaining terms are also equal to zero. Thus:

$$(b)\ \sum_{i,j,k,l}^{K} \bar{y}_{.j..}\left(y_{ijkl} - \bar{y}_{i...} - \bar{y}_{.j..} - \bar{y}_{..k.} - \bar{y}_{...l} + 3\bar{y}_{....}\right) =$$

$$K\sum_{j}^{K} \left(\bar{y}_{.j..}\right)^2 - K^2\left(\bar{y}_{....}\right)^2 - K\sum_{j}^{K} \left(\bar{y}_{.j..}\right)^2 - K^2\left(\bar{y}_{....}\right)^2 - K^2\left(\bar{y}_{....}\right)^2 + 3K^2\left(\bar{y}_{....}\right)^2 = 0$$

$$(c)\ \sum_{i,j,k,l}^{K} \bar{y}_{..k.}\left(y_{ijkl} - \bar{y}_{i...} - \bar{y}_{.j..} - \bar{y}_{..k.} - \bar{y}_{...l} + 3\bar{y}_{....}\right) =$$

$$K\sum_{k}^{K} \left(\bar{y}_{..k.}\right)^2 - K^2\left(\bar{y}_{....}\right)^2 - K\sum_{k}^{K} \left(\bar{y}_{..k.}\right)^2 - K^2\left(\bar{y}_{....}\right)^2 - K^2\left(\bar{y}_{....}\right)^2 + 3K^2\left(\bar{y}_{....}\right)^2 = 0$$

$$(d)\ \sum_{i,j,k,l}^{K} \bar{y}_{...l}\left(y_{ijkl} - \bar{y}_{i...} - \bar{y}_{.j..} - \bar{y}_{..k.} - \bar{y}_{...l} + 3\bar{y}_{....}\right) =$$

$$K\sum_{l}^{K} \left(\bar{y}_{...l}\right)^2 - K^2\left(\bar{y}_{....}\right)^2 - K\sum_{l}^{K} \left(\bar{y}_{...l}\right)^2 - K^2\left(\bar{y}_{....}\right)^2 - K^2\left(\bar{y}_{....}\right)^2 + 3K^2\left(\bar{y}_{....}\right)^2 = 0$$

$$(e)\ -4\sum_{i,j,k,l}^{K} \bar{y}_{....}\left(y_{ijkl} - \bar{y}_{i...} - \bar{y}_{.j..} - \bar{y}_{..k.} - \bar{y}_{...l} + 3\bar{y}_{....}\right) =$$

$$-4K^2\left(\bar{y}_{....}\right)^2 + 4K^2\left(\bar{y}_{....}\right)^2 + 4K^2\left(\bar{y}_{....}\right)^2 + 4K^2\left(\bar{y}_{....}\right)^2 + 4K^2\left(\bar{y}_{...l}\right)^2 - 12K^2\left(\bar{y}_{....}\right)^2 = 0$$

Now that we have proved that $2\sum ab = 0$, that is to say:

$$2\sum_{i,j,k,l}^{K} \left(\bar{y}_{i...} + \bar{y}_{.j..} + \bar{y}_{..k} + \bar{y}_{...l} - 4\bar{y}_{....}\right)\left(y_{ijkl} - \bar{y}_{i...} - \bar{y}_{.j..} - \bar{y}_{..k.} - \bar{y}_{...l} + 3\bar{y}_{....}\right) = 0 \tag{5.7.2148}$$

Let us come back to the beginning (...):

$$\begin{aligned}
&\sum_{i,j,k,l}^{K}\left(y_{ijkl}-\bar{y}_{....}\right)^2\\
&=\sum_{i,j,k,l}^{K}\left(y_{ijkl}-\bar{y}_{i...}-\bar{y}_{.j..}-\bar{y}_{..k.}-\bar{y}_{...l}+3\bar{y}_{....}\right)^2+\sum_{i,j,k,l}^{K}\left(\bar{y}_{i...}+\bar{y}_{.j..}+\bar{y}_{..k}+\bar{y}_{...l}-4\bar{y}_{....}\right)^2\\
&+2\sum_{i,j,k,l}^{K}\left(\bar{y}_{i...}+\bar{y}_{.j..}+\bar{y}_{..k.}+\bar{y}_{...l}-4\bar{y}_{....}\right)\left(y_{ijkl}-\bar{y}_{i...}-\bar{y}_{.j..}-\bar{y}_{..k.}-\bar{y}_{...l}+3\bar{y}_{....}\right)
\end{aligned}$$

Since the 3rd term is equal to it remains obviously only:

$$\sum_{i,j,k,l}^{K}\left(y_{ijkl}-\bar{y}_{i...}-\bar{y}_{.j..}-\bar{y}_{..k.}-\bar{y}_{...l}+3\bar{y}_{....}\right)^2+\sum_{i,j,k,l}^{K}\left(\bar{y}_{i...}+\bar{y}_{.j..}+\bar{y}_{..k.}+\bar{y}_{...l}-4\bar{y}_{....}\right)^2 \tag{5.7.2149}$$

Let us take the 2nd term of the above expression:

$$\begin{aligned}
&\sum_{i,j,k,l}^{K}\left(\bar{y}_{i...}+\bar{y}_{.j..}+\bar{y}_{..k.}+\bar{y}_{...l}-4\bar{y}_{....}\right)^2\\
&=\sum_{i,j,k,l}^{K}\left(\bar{y}_{i...}-\bar{y}_{....}+\bar{y}_{.j..}-\bar{y}_{....}+\bar{y}_{..k.}-\bar{y}_{....}+\bar{y}_{...l}-\bar{y}_{....}\right)^2\\
&=\sum_{i,j,k,l}^{K}\left[\left(\bar{y}_{i...}-\bar{y}_{....}\right)+\left(\bar{y}_{.j..}-\bar{y}_{....}\right)+\left(\bar{y}_{..k.}-\bar{y}_{....}\right)+\left(\bar{y}_{...l}-\bar{y}_{....}\right)\right]^2\\
&=\sum_{i,j,k,l}^{K}\left(\bar{y}_{i...}-\bar{y}_{....}\right)^2+\sum_{i,j,k,l}^{K}\left(\bar{y}_{.j..}-\bar{y}_{....}\right)^2+\sum_{i,j,k,l}^{K}\left(\bar{y}_{..k.}-\bar{y}_{....}\right)^2+\sum_{i,j,k,l}^{K}\left(\bar{y}_{...l}-\bar{y}_{....}\right)^2\\
&+\sum_{i,j,k,l}^{K}\left(\bar{y}_{i...}-\bar{y}_{....}\right)\cdot\left(\bar{y}_{i...}+\bar{y}_{.j..}+\bar{y}_{..k.}+\bar{y}_{...l}-4\bar{y}_{....}\right)\\
&+\sum_{i,j,k,l}^{K}\left(\bar{y}_{.j..}-\bar{y}_{....}\right)\cdot\left(\bar{y}_{i...}+\bar{y}_{.j..}+\bar{y}_{..k.}+\bar{y}_{...l}-4\bar{y}_{....}\right)\\
&+\sum_{i,j,k,l}^{K}\left(\bar{y}_{..k.}-\bar{y}_{....}\right)\cdot\left(\bar{y}_{i...}+\bar{y}_{.j..}+\bar{y}_{..k.}+\bar{y}_{...l}-4\bar{y}_{....}\right)\\
&+\sum_{i,j,k,l}^{K}\left(\bar{y}_{...l}-\bar{y}_{....}\right)\cdot\left(\bar{y}_{i...}+\bar{y}_{.j..}+\bar{y}_{..k.}+\bar{y}_{...l}-4\bar{y}_{....}\right)
\end{aligned}$$

In the above development let us consider the following expression:

$$\sum_{i,j,k,l}^{K}\left(\bar{y}_{i...}-\bar{y}_{....}\right)\left(\bar{y}_{i...}+\bar{y}_{.j..}+\bar{y}_{..k.}+\bar{y}_{...l}-4\bar{y}_{....}\right) \tag{5.7.2150}$$

And let us prove that it is equal to zero:

$$\begin{aligned}
&\sum_{i,j,k,l}^{K} \left(\bar{y}_{i...} - \bar{y}_{....}\right) \cdot \left(\bar{y}_{i...} + \bar{y}_{.j..} + \bar{y}_{..k.} + \bar{y}_{...l} - 4\bar{y}_{....}\right) \\
&= \sum_{i,j,k,l}^{K} \bar{y}_{i...} \cdot \bar{y}_{i...} + \sum_{i,j,k,l}^{K} \bar{y}_{i...} \cdot \bar{y}_{.j..} + \sum_{i,j,k,l}^{K} \bar{y}_{i...} \cdot \bar{y}_{..k.} \\
&+ \sum_{i,j,k,l}^{K} \bar{y}_{i...} \cdot \bar{y}_{...l} - \sum_{i,j,k,l}^{K} \bar{y}_{i...} \cdot 4\bar{y}_{....} - \sum_{i,j,k,l}^{K} \bar{y}_{....} \cdot \bar{y}_{i...} \\
&- \sum_{i,j,k,l}^{K} \bar{y}_{....} \cdot \bar{y}_{.j..} - \sum_{i,j,k,l}^{K} \bar{y}_{....} \cdot \bar{y}_{..k.} - \sum_{i,j,k,l}^{K} \bar{y}_{....} \cdot \bar{y}_{...l} + \sum_{i,j,k,l}^{K} \bar{y}_{....} \cdot 4\bar{y}_{....} \\
&= K^2 \left(\bar{y}_{....}\right)^2 + K^2 \left(\bar{y}_{....}\right)^2 + K^2 \left(\bar{y}_{....}\right)^2 + K^2 \left(\bar{y}_{....}\right)^2 \\
&- 4K^2 \left(\bar{y}_{....}\right)^2 - K^2 \left(\bar{y}_{....}\right)^2 - K^2 \left(\bar{y}_{....}\right)^2 - K^2 \left(\bar{y}_{....}\right)^2 \\
&- K^2 \left(\bar{y}_{....}\right)^2 + 4K^2 \left(\bar{y}_{....}\right)^2 \\
&= 0
\end{aligned}$$

Consequently, each of the following expressions:

$$\sum_{i,j,k,l}^{K} \left(\bar{y}_{.j..} - \bar{y}_{....}\right) \cdot \left(\bar{y}_{i...} + \bar{y}_{.j..} + \bar{y}_{..k.} + \bar{y}_{...l} - 4\bar{y}_{....}\right) \tag{5.7.2151}$$

$$\sum_{i,j,k,l}^{K} \left(\bar{y}_{..k.} - \bar{y}_{....}\right) \cdot \left(\bar{y}_{i...} + \bar{y}_{.j..} + \bar{y}_{..k.} + \bar{y}_{...l} - 4\bar{y}_{....}\right) \tag{5.7.2152}$$

$$\sum_{i,j,k,l}^{K} \left(\bar{y}_{...l} - \bar{y}_{....}\right) \cdot \left(\bar{y}_{i...} + \bar{y}_{.j..} + \bar{y}_{..k.} + \bar{y}_{...l} - 4\bar{y}_{....}\right) \tag{5.7.2153}$$

is equal to 0.

Finally it remains:

$$\begin{aligned}
\sum_{i,j,k,l}^{K} \left(y_{ijkl} - \bar{y}_{....}\right)^2 = &\sum_{i,j,k,l}^{K} \left(\bar{y}_{i...} - \bar{y}_{....}\right)^2 + \sum_{i,j,k,l}^{K} \left(\bar{y}_{.j..} - \bar{y}_{....}\right)^2 + \sum_{i,j,k,l}^{K} \left(\bar{y}_{..k.} - \bar{y}_{....}\right)^2 + \sum_{i,j,k,l}^{K} \left(\bar{y}_{...l} - \bar{y}_{....}\right)^2 \\
&+ \sum_{i,j,k,l}^{K} \left(y_{ijkl} - \bar{y}_{i...} - \bar{y}_{.j..} - \bar{y}_{..k.} - \bar{y}_{...l} + 3\bar{y}_{....}\right)^2 \\
= &\sum_{i}^{K} \sum_{j,k,l}^{K} \left(\bar{y}_{i...} - \bar{y}_{....}\right)^2 + \sum_{j}^{K} \sum_{i,k,l}^{K} \left(\bar{y}_{.j..} - \bar{y}_{....}\right)^2 + \sum_{k}^{K} \sum_{i,j,l}^{K} \left(\bar{y}_{..k.} - \bar{y}_{....}\right)^2 + \sum_{l}^{K} \sum_{i,j,k}^{K} \left(\bar{y}_{...l} - \bar{y}_{....}\right)^2 \\
&+ \sum_{i,j,k,l}^{K} \left(y_{ijkl} - \bar{y}_{i...} - \bar{y}_{.j..} - \bar{y}_{..k.} - \bar{y}_{...l} + 3\bar{y}_{....}\right)^2 \\
= &K \sum_{i}^{K} \left(\bar{y}_{i...} - \bar{y}_{....}\right)^2 + K \sum_{j}^{K} \left(\bar{y}_{.j..} - \bar{y}_{....}\right)^2 + K \sum_{k}^{K} \left(\bar{y}_{..k.} - \bar{y}_{....}\right)^2 + K \sum_{l}^{K} \left(\bar{y}_{...l} - \bar{y}_{....}\right)^2 \\
&+ \sum_{i,j,k,l}^{K} \left(y_{ijkl} - \bar{y}_{i...} - \bar{y}_{.j..} - \bar{y}_{..k.} - \bar{y}_{...l} + 3\bar{y}_{....}\right)^2
\end{aligned}$$

If we denote as usual SSE as the sum of the squares of the deviations then we can rewrite our result in this form:

$$\text{SSE}_T=\text{SSE}_R+\text{SSE}_C+\text{SSE}_l+\text{SSE}_g+\text{SSE}_r \tag{5.7.2154}$$

where:

$$\begin{aligned}\text{SSE}_T&=\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{l=1}^{K}\left(y_{ijkl}-\bar{y}_{....}\right)^2\\ \text{SSE}_R&=K\sum_{i}^{K}\left(\bar{y}_{i...}-\bar{y}_{....}\right)^2\\ \text{SSE}_C&=K\sum_{j}^{K}\left(\bar{y}_{.j..}-\bar{y}_{....}\right)^2\\ \text{SSE}_l&=K\sum_{k}^{K}\left(\bar{y}_{..k.}-\bar{y}_{....}\right)^2\\ \text{SSE}_g&=K\sum_{l}^{K}\left(\bar{y}_{...l}-\bar{y}_{....}\right)^2\end{aligned} \tag{5.7.2155}$$

and:

$$\text{SSE}_r=\sum_{i,j,k,l}^{K}\left(y_{ijkl}-\bar{y}_{i...}-\bar{y}_{.j..}-\bar{y}_{..k.}-\bar{y}_{...l}+3\bar{y}_{....}\right)^2=\text{SSE}_T-\text{SSE}_L-\text{SSE}_C-\text{SSE}_l-\text{SSE}_g \tag{5.7.2156}$$

With for degrees of freedom:

$$\text{df}_T=\text{df}_R+\text{df}_C+\text{df}_l+\text{df}_g+\text{df}_r$$

$$K^2-1=(K-1)+(K-1)+(K-1)+(K-1)+(K-1)\,(K-3)$$

The degree of freedom for the squares of the residual differences is obtained by posing:

$$K^2-1-[(K-1)+(K-1)+(K-1)+(K-1)]=(K-1)\,(K-3)$$

We then have the following ANOVA table:

Sum of squares (SSE)	χ^2 **df**	**Mean squares**	F	**Critical** F
$Q_R = K\sum_i^K (\bar{y}_{i...} - \bar{y}_{....})^2$ (rows)	$K-1$	$\text{MSk}R = \dfrac{Q_R}{K-1}$	$\dfrac{\text{MSk}R}{\text{MSE}}$	$P(F > F_{K-1,(K-1)(K-3)})$
$Q_C = K\sum_j (\bar{y}_{.j.} - \bar{y}_{...})^2$ (columns)	$K-1$	$\text{MSk}C = \dfrac{Q_C}{K-1}$	$\dfrac{\text{MSk}C}{\text{MSE}}$	$P(F > F_{K-1,(K-1)(K-3)})$
$Q_l = K\sum_k^K (\bar{y}_{..k.} - \bar{y}_{....})^2$ (latin letter)	$K-1$	$\text{MSk}l = \dfrac{\text{Q}_\text{l}}{K-1}$	$\dfrac{\text{MSk}l}{\text{MSE}}$	$P(F > F_{K-1,(K-1)(K-3)})$
$Q_g = K\sum_l^K (\bar{y}_{...l} - \bar{y}_{....})^2$ (greek letter)	$K-1$	$\text{MSk}g = \dfrac{Q_l}{K-1}$	$\dfrac{\text{MSk}g}{\text{MSE}}$	$P(F > F_{K-1,(K-1)(K-3)})$
$Q_r = \sum_{i,j,k,l}^K (y_{ijkl} - \bar{y}_{i...} - \bar{y}_{.j..} - \bar{y}_{..k.} - \bar{y}_{...l} + 3\bar{y}_{....})^2$ (residuals)	$(K-1)(K-3)$	$\text{MSE} = \dfrac{Q_r}{(K-1)(K-3)}$		
$Q_T = \sum_{i=1}^K\sum_{j=1}^K\sum_{k=1}^K\sum_{l=1}^K (y_{ijkl} - \bar{y}_{....})^2$ (total)	K^2-1	$\text{MSk}T = \dfrac{Q_T}{K^2-1}$		

Table 5.75: **Greaco-Latin square ANOVA table**

Notice that in the Graeco-Latin Square Design as in the Latin Square Design it is assumed that no interaction among the factors exist. This is an even stronger assumption in the Graeco-Latin Square Design because we have an additional factor and, hence, more possible interactions.

Remark

The factor effect model for this design is (purely additive!):

$$\boxed{y_{ijkl} = \mu + \theta_i + \tau_j + \omega_k + \phi_l + \varepsilon_{ijkl}} \tag{5.7.2157}$$

where $i, j, k, l = 1 \ldots N$ and where y_{ijkl} is the observation in the ith row, Latin letter j, Greek letter k, and lth column, μ is the overall mean, θ_i is the ith row effect, τ_j is the effect of the Latin letter j treatment, ω_k is the effect of Greek letter k, ϕ_l is the lth column effect, and ε_{ijkl} are i.i.d. $\mathcal{N}(0, \sigma_\varepsilon^2)$ errors. Notice that as for the Latin square design two subscripts suffice to identify an observation and 6×6 Graeco-Latin designs are impossible.

We are (mainly!) interested in testing for zero treatment effect, that is, $H_0 : \tau_1 = \tau_2 = \ldots = \tau_N = 0$ versus $H_1 : \tau_j \neq 0$.

Many softwares return only the first Fisher test (that for Q_R, i.e. the treatments) since the blocks of the Greaco-Latin design represent restrictions on randomization and that therefore the other tests may not be appropriate.

5.7.13.5.9 Multivariate ANOVA (MANOVA)

The "Multivariate Analysis of Variance" is an extension of ANOVA where we assume that the categorical variables are linearly dependent and we test whether there is a significant difference in the averages (variables supposed to be of the continuous type) through the different groups under the interdependence of the categorical explanatory variables.

Again, as for the ANOVA, the MANOVA and Multivariate Linear Models are the two facets of a same coin are the ground foundations of a more field of statistics as depicted in the flowchart below (actually in the book, only Multivariate Generalized Linear Models are treated in details in the section of Numerical Methods!):

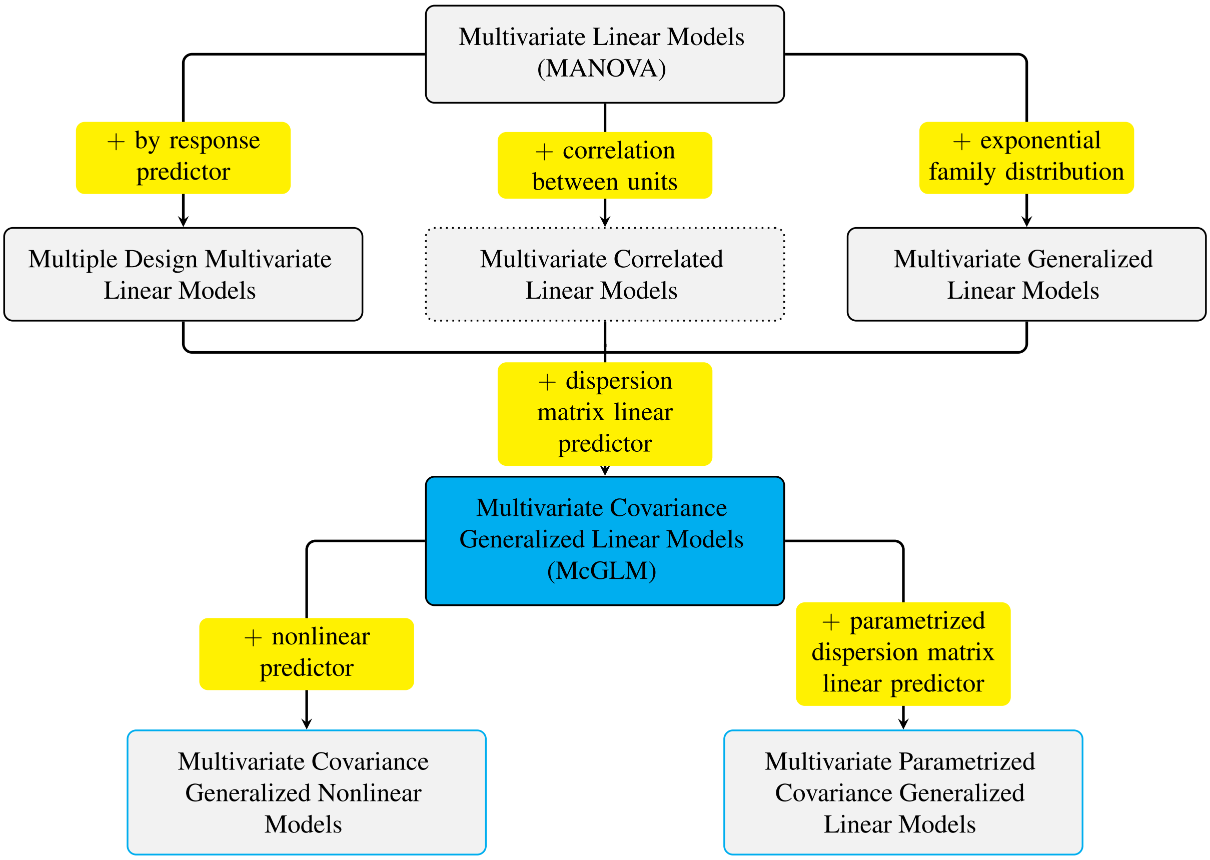

Multivariate Linear Models (MANOVA)
+ by response predictor
+ correlation between units
+ exponential family distribution
Multiple Design Multivariate Linear Models
Multivariate Correlated Linear Models
Multivariate Generalized Linear Models
+ dispersion matrix linear predictor
Multivariate Covariance Generalized Linear Models (McGLM)
+ nonlinear predictor
+ parametrized dispersion matrix linear predictor
Multivariate Covariance Generalized Nonlinear Models
Multivariate Parametrized Covariance Generalized Linear Models

Figure 5.204: **Families of Multivariate Linear Models**

If it is not clear (...), the mathematical notation may be more explicit to the reader. Let us recall first that in the case of the canonical ANOVA with 1 fixed factor at k levels, the null hypothesis was of the form:

$$H_0:\ \bar{x}_1=\bar{x}_2=\ldots=\bar{x}_k \tag{5.7.2158}$$

For the one-way MANOVA, the null hypothesis is:

$$H_0:\ \vec{\mu}_1=\vec{\mu}_2=\vec{\mu}_3=\ldots=\vec{\mu}_k \tag{5.7.2159}$$

where we have therefore "mean vectors" for a given number of dependent group variables in a quantity k.

Thus explicitly for k groups and p categorical variables:

$$H_0 : \begin{pmatrix} \bar{x}_{11} \\ \bar{x}_{21} \\ \vdots \\ \bar{x}_{p1} \end{pmatrix} = \begin{pmatrix} \bar{x}_{12} \\ \bar{x}_{22} \\ \vdots \\ \bar{x}_{p2} \end{pmatrix} = \begin{pmatrix} \bar{x}_{13} \\ \bar{x}_{23} \\ \vdots \\ \bar{x}_{p3} \end{pmatrix} = \ldots = \begin{pmatrix} \bar{x}_{1k} \\ \bar{x}_{2k} \\ \vdots \\ \bar{x}_{pk} \end{pmatrix} \qquad (5.7.2160)$$

As the reader will see in the developments below, the one-way MANOVA has nine assumptions (hypothesis):

H1. Our two or more dependent variables should be measured at a continuous level (i.e., they are interval or ratio variables)

H2. Our independent variable should consist of two or more categorical, independent (unrelated) groups

H3. We should have independence of observations, which means that there is no relationship between the observations in each group or between the groups themselves

H4. We should have an adequate sample size. At least need to have more cases (e.g., participants) in each group of the independent variable than the number of dependent variables we are analysing

H5. There are no univariate or multivariate outliers. First, there can be no (univariate) outliers in each group of the independent variable for any of the dependent variables

H6. There is multivariate Normality

H7. There is a linear relationship between each pair of dependent variables for each group of the independent variable

H8. There is a homogeneity of variance-covariance matrices

H9. There is no multicollinearity

Let us now recall that in the one-factor ANOVA we compute the p-value of the ratio of the estimated variances assuming the true variances as equals:

$$F_{k-1,N-k} = \dfrac{\dfrac{Q_A}{k-1}}{\dfrac{Q_R}{N-k}} = \dfrac{\dfrac{\text{SSB}}{k-1}}{\dfrac{\text{SSA}}{N-k}} \qquad (5.7.2161)$$

The next step in understanding MANOVA is the recognition that mathematicians are lazy and sometimes do engineering statistics... So first we simplify this ratio by multiplying left and right by $(k-1)/(N-1)$. This gives what we name an "A-statistics":

$$A = \dfrac{k-1}{N-1} F_{k-1,N-k} = \dfrac{k-1}{N-1} \dfrac{\dfrac{\text{SSB}}{k-1}}{\dfrac{\text{SSA}}{N-k}} = \dfrac{\text{SSB}}{\text{SSW}} \qquad (5.7.2162)$$

But this is not enough as we are still with independent variables! So the idea is to used variance-covariance like matrix and to get a similar expression of as a ratio we take the determinant... This brings us to write a multivariate like Fisher test for variances (......):

$$\Lambda=\frac{\det(W)}{\det(T)}=\frac{\det(W)}{\det(B+W)} \tag{5.7.2163}$$

better known in the following form:

$$\boxed{\Lambda=\frac{|W|}{|B+W|}=\frac{|W|}{|T|}} \tag{5.7.2164}$$

where W and T are respectively the determinants of the square matrices of the sum of the squares of the within (W) and global (Total: T) deviations. Which means that if the part "within" (Between: B) is really large, then Λ tends to zero. In contrast, if B is very small, tends Λ tends to 1.

To prove that this ratio does not depend on the covariance matrix of the dependent variables, let us recall that considering the structure of W and $B+W$ we can make a spectral decomposition (see further below our study of principal component analysis) in the form:

$$\Lambda=\frac{|W|}{|B+W|}=\frac{|S||W||S^T|}{|S||B+W||S^T|}=\frac{|W|}{|B+W|} \tag{5.7.2165}$$

It can be demonstrated with some approximations that Λ follows a distribution named "Wilk's distribution" (of "Samuel Stanley Wilks").

It should be noted that Wilk's Λ can be expressed as a function of the eigenvalues of $W^{-1}B$. From the definition of Λ, it follows using the properties of the determinant (see section Linear Algebra page 1961) that:

$$\frac{1}{\Lambda}=\frac{|T|}{|W|}=\left|W^{-1}(W+B)\right|=\left|\mathbb{1}_p+W^{-1}B\right| \tag{5.7.2166}$$

We recognize here the determinant of the "eigenvalue equations" (see section Linear Algebra page 1998) on $W^{-1}B$ but with $\lambda=1$. Indeed:

$$\det(A-\lambda I_n)=\prod_{i=1}^{n}(\lambda_i-\lambda) \tag{5.7.2167}$$

So in our case here we have $\lambda=-1$ therefore:

$$\det(A-\lambda I_n)=\prod_{i=1}^{n}(\lambda_i+1) \tag{5.7.2168}$$

Therefore:

$$\frac{1}{\Lambda}=\frac{|T|}{|W|}=\prod_{i=1}^{k}(1+\lambda_i) \tag{5.7.2169}$$

and consequently:

$$\Lambda=\frac{1}{\prod\limits_{i=1}^{k}(1+\lambda_i)}=\prod_{i=1}^{k}\frac{1}{1+\lambda_i} \qquad (5.7.2170)$$

Also, it follows that:

$$\ln(\Lambda)=\ln\left(\frac{1}{\prod\limits_{i=1}^{k}(1+\lambda_i)}\right)=-\sum_{i=1}^{k}(1+\lambda_i) \qquad (5.7.2171)$$

When Wilk's Λ approaches 1, we showed that it means that the difference in means is negligible. This is the case when $\ln(\Lambda)$ approaches 0. However, when Λ approaches 0 or $\ln(\Lambda)$ approaches 1, it means that the difference is large. Therefore, a large value of $\ln(\Lambda)$ (i.e., close to 0) is an indication of the significance of the difference between the means.

Remark

The bad news is that there are in fact four different multivariate tests that are made from the above relation. The reason for four different statistics and four approximations is that the mathematics of MANOVA get so complicated in some cases that no one has ever been able to solve them as far as we know...

Let us see now a companion detailed example:

✍**Example:**

We consider the following set of two dependent variables on three groups of independent variables:

K_1		K_2		K_3	
x_1	x_2	x_1	x_2	x_1	x_2
2	3	4	8	7	6
3	4	5	6	8	7
5	4	6	7	10	8
2	5			9	5
				7	6
$\bar{x}_{11}=3$	$\bar{x}_{21}=4$	$\bar{x}_{12}=5$	$\bar{x}_{22}=7$	$\bar{x}_{13}=8.2$	$\bar{x}_{23}=6.4$

with:

$$\bar{x}_1=\frac{3+5+8.2}{3}=5.4 \qquad \bar{x}_2=\frac{4+7+6.4}{3}=5.8 \qquad (5.7.2172)$$

Now we calculate:

$$W=W_1+W_2+W_3=\begin{bmatrix}\text{SSE}_1 & \text{SSE}_{12}\\ \text{SSE}_{12} & \text{SSE}_2\end{bmatrix}+\begin{bmatrix}\text{SSE}_1 & \text{SSE}_1\\ \text{SSE}_{22} & \text{SSE}_2\end{bmatrix}+\begin{bmatrix}\text{SSE}_1 & \text{SSE}_{32}\\ \text{SSE}_{32} & \text{SSE}_2\end{bmatrix} \qquad (5.7.2173)$$

We then have for example for the first group:

$$\text{SSE}_i = \sum_{j=1}^{n}(x_{ij} - \bar{x}_{i1})^2 : \begin{cases} \text{SSE}_1 = \sum_{j=1}^{n}(x_{1j} - \bar{x}_{11})^2 \\ \text{SSE}_2 = \sum_{j=1}^{n}(x_{2j} - \bar{x}_{21})^2 \end{cases}$$
$$\text{SSE}_{ij} = \text{SSE}_{ji} = \sum_{j=1}^{n}(x_{ij} - \bar{x}_{11})(x_{2j} - \bar{x}_{21}) \qquad (5.7.2174)$$

Which gives us in our case (very easy to check with any spreadsheet software):

$$W = \begin{bmatrix} 6 & 0 \\ 0 & 2 \end{bmatrix} + \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} + \begin{bmatrix} 6.8 & 2.6 \\ 2.6 & 5.2 \end{bmatrix} = \begin{bmatrix} 14.8 & 1.6 \\ 1.6 & 9.2 \end{bmatrix} \qquad (5.7.2175)$$

And we have (we notice that we assume that over all the x_1 and x_2 the number of measurement are $N_1 = N_2 = N$):

$$T = B + W = \begin{bmatrix} \sum_{i=1}^{N}(x_{i1} - \bar{x}_1)^2 & \sum_{i=1}^{N}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) \\ \sum_{i=1}^{N}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) & \sum_{i=1}^{N}(x_{i2} - \bar{x}_2)^2 \end{bmatrix} = \begin{bmatrix} 76.72 & 26.00 \\ 26.00 & 28.25 \end{bmatrix} \qquad (5.7.2176)$$

and therefore:

$$\Lambda = \frac{|W|}{|T|} = \frac{\begin{vmatrix} 14.8 & 1.6 \\ 1.6 & 9.2 \end{vmatrix}}{\begin{vmatrix} 76.72 & 26.00 \\ 26.00 & 28.25 \end{vmatrix}} = 0.0897 \qquad (5.7.2177)$$

After the calculation of the p-value is more delicate and we will ask the reader to refer to the Minitab or `R` companion books for the computational details and the final conclusion.

So we could use a one-way MANOVA to determine whether students' short-term and long-term recall of facts differed based on three different lengths of lecture (i.e., the two dependent variables are "short-term memory recall" and "long-term memory recall", whilst the independent variable is "lecture duration", which has four independent groups: "30 minutes", "60 minutes", "90 minutes" and "120 minutes"). Alternately, a one-way MANOVA could be used to determine whether there is a difference in salary and bonuses based on degree type (i.e., the two dependent variables are "salary" and "bonuses", whilst the independent variable is "degree type", which has five groups: "business studies", "psychology", "biological sciences", "engineering" and "law").

When there is a statistically significant difference between the groups of the independent variable, it is possible to determine which specific groups were significantly different from each other using post hoc tests. You need to conduct these post hoc tests because the one-way MANOVA is an omnibus[82] test statistic and cannot tell you which specific groups were significantly different from each other; it only tells you that at least two groups were different.

---

[82] "omnibus" derives from the Latin word "for all". In plain language, you can interpret an omnibus test or omnibus statistic like an "overall test" - it is testing a number of things at once.

5.7.13.5.9.1 Hotelling's T-squared Test

The "Hotelling's T-squared test" (Hotelling, 11931 according to holocene calendar) is the multivariate counterpart of the T-test. For example, let's say we wanted to compare how well two different sets of students performed in school. We could compare univariate data (e.g. mean test scores) with a T-test. Or, we could use Hotelling's T-squared to compare multivariate data (e.g. the multivariate mean of test scores, GPA, and class grades).

Hotelling's T-squared test is based on Hotelling's T^2 distribution and forms the basis for various multivariate control charts in Industrial Engineering. As the reader will understand it at then end of the development presented below, Hotelling's T-squared test is simply a special case of MANOVA, in which only two batches (groups) are compared!

More technically speaking Hotelling's T^2 test makes it possible to analyse whether the measures of random variables supposed to jointly follow a multivariate Normal law and therefore having a possible correlation (variance-covariance matrix) deviate considerably from Normality or not.

To introduce this test, let us recall that we have demonstrated in the section of Statistics that the Multivariate Normal law could be written:

$$f\left(X_1,X_2\right)=\frac{1}{2\pi|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}\overrightarrow{(x-\mu)}^T\Sigma^{-1}\overrightarrow{(x-\mu)}\right) \tag{5.7.2178}$$

We see that the following term in the parenthesis:

$$D=\overrightarrow{(x-\mu)}^T\Sigma^{-1}\overrightarrow{(x-\mu)} \tag{5.7.2179}$$

giving a scalar can be seen as the unique random variable of the bivariate distribution (as we can do with the square of the random variable in the reduced centered univariate Normal law). So for a given value of D, we are "at a certain height" of the multidimensional Gaussian distribution.

What will interest us here is that when we work with a multivariate distribution some variables have less weight than others because of their large standard deviation and the correlated variables tend to reduce the amplitude Y.

Therefore, a very interesting subject in industrial practice (Hotelling control card and regressions) is then to know that it is the distribution law which describes:

$$D=\overrightarrow{(x-\mu)}^T\Sigma^{-1}\overrightarrow{(x-\mu)} \tag{5.7.2180}$$

So let us deal now with this topic!

The "Mahalanobis distance" between two points $\vec{x}$ and $\vec{y}$ is defined as

$$d(\vec{x},\vec{y})=\sqrt{(\vec{x}-\vec{y})^\top\Sigma^{-1}(\vec{x}-\vec{y})} \tag{5.7.2181}$$

Thus, the squared Mahalanobis distance of a random vector $\vec{x}=\mathcal{N}(\vec{\mu},\Sigma)$ and the center $\vec{\mu}$ of a multivariate Gaussian distribution is defined as:

$$D=d(\vec{x},\vec{\mu})^2=(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu}) \tag{5.7.2182}$$

where Σ is a $n \times n$ covariance matrix and $\vec{\mu}$ is the mean vector. In order to achieve a different representation of D one can first perform an eigenvalue decomposition on Σ^{-1} which is with (see section Linear Algebra page 2023):

$$\Sigma^{-1} = U\Lambda^{-1}U^{-1} = U\Lambda^{-1}U^{T} = \sum_{k=1}^{n} \lambda_k^{-1}\vec{u}_k\vec{u}_k^{T} \tag{5.7.2183}$$

where for recall $\vec{u}_k$ is the kth eigenvector of the corresponding eigenvalue λ_k. Plugging the above relation back into:

$$D = d(\vec{x},\vec{\mu})^2 = (\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu}) \tag{5.7.2184}$$

results in:

$$\begin{aligned} D &= (\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu}) \\ &= (\vec{x}-\vec{\mu})^T\left(\sum_{k=1}^{n}\lambda_k^{-1}\vec{u}_k\vec{u}_k^{T}\right)(\vec{x}-\vec{\mu}) \\ &= \sum_{k=1}^{n}\lambda_k^{-1}(\vec{x}-\vec{\mu})^T\vec{u}_k\vec{u}_k^{T}(\vec{x}-\vec{\mu}) \end{aligned} \tag{5.7.2185}$$

Using the properties of transposed (see section Linear Algebra page 1951) matrices we get:

$$D = \sum_{k=1}^{n}\lambda_k^{-1}\left[\vec{u}_k^{T}(\vec{x}-\vec{\mu})\right]^2 = \sum_{k=1}^{n}\left[\lambda_k^{-\frac{1}{2}}\vec{u}_k^{T}(\vec{x}-\vec{\mu})\right]^2 = \sum_{k=1}^{n}Y_k^2 \tag{5.7.2186}$$

where Y_k is a new random variable based on an affine linear transform of the random vector $\vec{x}$. Indeed, we have:

$$Z = (\vec{x}-\vec{\mu}) = \mathcal{N}(\vec{0},\Sigma) \tag{5.7.2187}$$

If we set:

$$\vec{a}_k^{T} := \lambda_k^{-1/2}\vec{u}_k^{T} \tag{5.7.2188}$$

then we get:

$$Y_k = \lambda_k^{-\frac{1}{2}}\vec{u}_k^{T}(\vec{x}-\vec{\mu}) = \lambda_k^{-1/2}\vec{u}_k^{T}Z = \vec{a}_k^{T}Z \tag{5.7.2189}$$

We can the say that Y_k is the result of an affine transformation of Z.

Note that Y_k is now a random variable drawn from a univariate Normal distribution $Y_k = \mathcal{N}\left(0,\sigma_k^2\right)$, where, according to:

$$\begin{aligned} \Sigma_y &= \mathrm{E}\left\{(\vec{y}-\vec{\mu}_y)(\vec{y}-\vec{\mu}_y)^T\right\} \\ &= \mathrm{E}\left\{\left[(A\vec{x}+\vec{b})-\left(A\vec{\mu}_x+\vec{b}\right)\right]\left[(A\vec{x}+\vec{b})-\left(A\vec{\mu}_x+\vec{b}\right)\right]^T\right\} \\ &= \mathrm{E}\left\{\left[A\left(\vec{x}-\vec{\mu}_x\right)\right]\left[A\left(\vec{x}-\vec{\mu}_x\right)\right]^T\right\} = \mathrm{E}\left\{A\left(\vec{x}-\vec{\mu}_x\right)\left(\vec{x}-\vec{\mu}_x\right)^T A^T\right\} \\ &= A\mathrm{E}\left\{\left(\vec{x}-\vec{\mu}_x\right)\left(\vec{x}-\vec{\mu}_x\right)^T\right\}A^T = A\mathbf{\Sigma}_x A^T \end{aligned} \tag{5.7.2190}$$

Therefore we get:

$$\sigma_k^2=\vec{a}_k^T\Sigma\vec{a}_k=\lambda_k^{-1/2}\vec{u}_k^T\Sigma\lambda_k^{-1/2}\vec{u}_k=\lambda_k^{-1}\vec{u}_k^T\Sigma\vec{u}_k \tag{5.7.2191}$$

If we insert into that last relation the following one:

$$\Sigma=\sum_{j=1}^{n}\lambda_j\vec{u}_j\vec{u}_j^T \tag{5.7.2192}$$

we get:

$$\begin{aligned}\sigma_k^2&=\lambda_k^{-1}\vec{u}_k^T\Sigma\vec{u}_k=\lambda_k^{-1}\vec{u}_k^T\left(\sum_{j=1}^{n}\lambda_j\vec{u}_j\vec{u}_j^T\right)\vec{u}_k\\&=\sum_{j=1}^{n}\lambda_k^{-1}\vec{u}_k^T\lambda_j\vec{u}_j\vec{u}_j^T\vec{u}_k=\sum_{j=1}^{n}\lambda_k^{-1}\lambda_j\vec{u}_k^T\vec{u}_j\vec{u}_j^T\vec{u}_k\end{aligned} \tag{5.7.2193}$$

since all eigenvectors $\vec{u}_i$ are pairwise orthonormal the dotted products $\vec{u}_k^T\vec{u}_j$ and $\vec{u}_j^T\vec{u}_k$ will be zero for $j\neq k$. Only for the case $j=k$ we get:

$$\begin{aligned}\sigma_k^2&=\lambda_k^{-1}\lambda_k\vec{u}_k^T\vec{u}_k\vec{u}_k^T\vec{u}_k\\&=\lambda_k^{-1}\lambda_k\|\vec{u}_k\|^2\|\vec{u}_k\|^2\\&=\lambda_k^{-1}\lambda_k\|\vec{u}_k\|^2\|\vec{u}_k\|^2\\&=1\end{aligned} \tag{5.7.2194}$$

since the norm $\|\vec{u}_k\|$ of a orthonormal eigenvector is equal to 1. The squared Mahalanobis distance can then be expressed as:

$$D=d(\vec{x},\vec{\mu})^2=(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})=\sum_{k=1}^{n}Y_k^2=\sum_{k=1}^{n}(\mathcal{N}(0,1))^2 \tag{5.7.2195}$$

Now the Chi-square distribution with n degrees of freedom is exactly defined as being the distribution of a variable which is the sum of the squares of n random variables being standard normally distributed. Hence, D is Chi-square distributed with n degrees of freedom.

So if the random variables are really jointly Normally distributed according to a correlation matrix Σ we then have:

$$\boxed{T^2=(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})=\sum_{k=1}^{n}(\mathcal{N}(0,1))^2=\chi_n^2} \tag{5.7.2196}$$

which is named the "Hotelling T^2 statistics".

Then to a given confidence threshold we then have:

$$P\left((\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})\leq\chi_{n-1}^2\right) \tag{5.7.2197}$$

that is the probability associated with a certain height (isocline) in the special case of the Normal bivariate distribution. So each quantile of the χ^2 law corresponds for example in the bivariate special case to an elliptical isocline in the general case (when seen from above) or to a circle when the correlation is zero.

So here what we are testing is:

$$H_0 : \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \vdots \\ \mu_{1p} \end{pmatrix} = \begin{pmatrix} \mu_{21} \\ \mu_{22} \\ \vdots \\ \mu_{2p} \end{pmatrix} \text{ against } H_1 : \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \vdots \\ \mu_{1p} \end{pmatrix} \neq \begin{pmatrix} \mu_{21} \\ \mu_{22} \\ \vdots \\ \mu_{2p} \end{pmatrix}$$

Or, in other words:

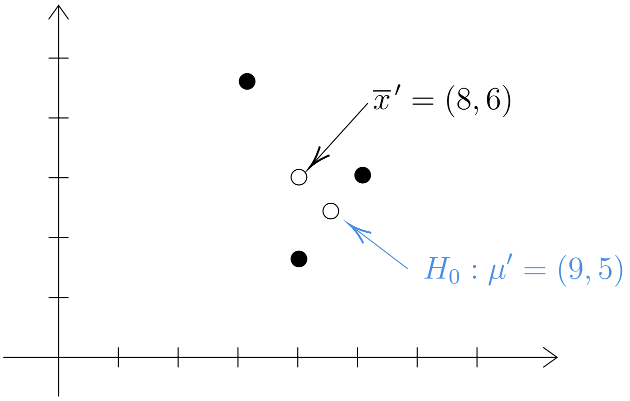
$$H_0 : \mu_{11} = \mu_{21} \text{ and } \mu_{12} = \mu_{22} \text{ and } \ldots \text{ and } \mu_{1p} = \mu_{2p}$$

For the null hypothesis, that is fine only if the population means are identical for all of the variables. The alternative is that at least one pair of these means is different. This is expressed below:

$$H_1 : \mu_{1k} \neq \mu_{2k} \text{ for at least one } k \in \{1, 2, \cdots, p\}$$

✍**Example:**

Consider:



The data are given by:

$$X = \begin{pmatrix} 6 & 9 \\ 10 & 6 \\ 8 & 3 \end{pmatrix} \tag{5.7.2198}$$

and the null hypothesis will be:

$$H_0 : \quad \vec{\mu}_0 = \begin{pmatrix} 9 \\ 5 \end{pmatrix} \tag{5.7.2199}$$

Assuming data come from a multivariate Normal distribution and independent observations (don't forget that S is the traditional notation for the estimate of Σ):

$$\vec{\bar{x}}=\begin{pmatrix}8\\6\end{pmatrix}\quad S=\begin{pmatrix}4&-3\\-3&9\end{pmatrix}\tag{5.7.2200}$$

We have:

$$S^{-1}=\frac{1}{4(9)-(-3)(-3)}\begin{pmatrix}9&3\\3&4\end{pmatrix}=\begin{pmatrix}1/3&1/9\\1/9&4/27\end{pmatrix}\tag{5.7.2201}$$

Hence:

$$\begin{aligned}T^2&=\left(\vec{\bar{x}}-\vec{\mu}_0\right)'S^{-1}\left(\vec{\bar{x}}-\vec{\mu}_0\right)\\&=((8-9),(6-5))\begin{pmatrix}1/3&1/9\\1/9&4/27\end{pmatrix}\begin{pmatrix}(8-9)\\(6-5)\end{pmatrix}\\&=(-1,1)\begin{pmatrix}1/3&1/9\\1/9&4/27\end{pmatrix}\begin{pmatrix}-1\\1\end{pmatrix}=7/27\cong 0.259\end{aligned}\tag{5.7.2202}$$

The value we need for a threshold of 5% is:

$$\chi^2_{n,1-\alpha}=\chi^2_{2,1-5\%}=\chi^2_{2,95\%}\cong 5.991\tag{5.7.2203}$$

We can compare our T^2 to 5.991. We see already that we can't reject H_0 (see how close $\vec{\bar{x}}$ and $\vec{\mu}$ and in the figure).

Alternatively we could compute the p-value easily with any spreadsheet software. This lead us to a p-value of 0.878.

5.7.13.6 Equivalence Tests

As the reader will probably have understood it from the preceding paragraphs, hypotheses tests (NHST) applied to the search for differences between groups do not technically allow us to conclude an equivalence simply because we do not reject the null hypothesis H_0. Yet we also showed that if the power of the test was typically above 80% it did not make a problem to consider the null hypothesis H_0 as true but the problem is that for this we need most of times to consider large samples and in practice this is not always feasible and it can also lead to absurdities since the maths show that with large samples we almost systematically reject the null hypothesis H_0.

The p-value can be used only to statistically reveal a posteriori the rejection of the null hypothesis H_0 in the vast majority of cases. To summarize this problem: «*The absence of evidence, does not imply evidence of absence*». Expressed in other words in a very common case this gives (particular example!) that the non-rejection of the null hypothesis H_0 of equality of two means (for example) does not imply the equality of means!!!

Remark

Some people quite often state wrongly that « *Absence of evidence is NOT evidence of absence.*» Why do they do that? Perhaps because they do not know better? And with the proof below they get the possibility of knowing for sure that "Absence of evidence IS evidence of absence!" and nothing else....

Definitions:

D1. A is evidence of B: $P(B\mid A)>P(B\mid\neg A)$

D2. Absence of evidence: $a=\neg A$

D3. Absence $b=\neg B$

$$\begin{gathered}P(B\mid A)>P(B\mid\neg A)\quad\Leftrightarrow\quad 1-P(\neg B\mid A)>1-P(\neg B\mid\neg A)\\ \Downarrow\\ P(\neg B\mid\neg A)>P(\neg B\mid\neg\neg A)\quad\Leftrightarrow\quad P(\neg B\mid A)<P(\neg B\mid\neg A)\\ \Downarrow\\ P(b\mid a)>P(b\mid\neg a)\end{gathered}\tag{5.7.2204}$$

■ Q.E.D.

An absence of evidence can be evidence of absence when we can reasonably expect evidence to exist (see further below the bayesian approach that confirms this conclusion)!

With the advent of generic medicines from the 11960s (holocene calendar) in the pharmaceutical field, the importance of highlighting equivalences (we are talking instead of "bioequivalences") has taken a growing place and organizations like the F.D.A. (Federal Drug Administration in U.S.A.), EMA (European Medicines Agency) or WHO (World Health Organization) have common guidelines that lead to the approval of generic only if the equivalence is shown under certain empirical conditions but at least standardized according to a consensus of experts.

Thus, most often the following three conditions are required[83]:

1. A comparison criterion is required
2. A confidence interval is required
3. An a priori limit of bioequivalence is required

Let us see a companion example to grasp the concept:

✍**Example:**

Let us consider the Student's T test of the difference of the averages of two unpaired samples $\{$Test, Reference$\}$ which we had demonstrated above:

$$T_{\alpha/2,n_1+n_2-2}=\frac{(\bar{X}_T-\bar{X}_R)-(\mu_T-\mu_R)}{\hat{\sigma}_D\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \tag{5.7.2205}$$

We then know the following two bounds of the confidence interval:

$$\begin{aligned}\text{LCL}&=(\bar{X}_T-\bar{X}_T)-T_{\alpha/2,n_1+n_2-2}\hat{\sigma}_D\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\\ \text{UCL}&=(\bar{X}_T-\bar{X}_T)+T_{\alpha/2,n_1+n_2-2}\hat{\sigma}_D\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\end{aligned} \tag{5.7.2206}$$

Remark

We recognize here a special case of "TOST" test for "Two-One Sided Test" seen earlier.

We will assume that the F.D.A. request:

$$\begin{aligned}\text{LSL}&=-0.2\mu_R\\ \text{USL}&=+0.2\mu_R\\ \delta_L&=80\%\\ \delta_U&=120\%\end{aligned} \tag{5.7.2207}$$

Consider that the measurement have given us:

$$\begin{aligned}\bar{X}_T&=80.272\\ \bar{X}_R&=82.559\\ n_1&=n_2=12\\ T_{\alpha/2,n_1+n_2-2}&=T_{5\%,22}=1.717\\ \hat{\sigma}_D&=9.145\end{aligned} \tag{5.7.2208}$$

Therefore, we have the equivalence limits that will be:

$$\left.\begin{aligned}\text{LSL}&=-0.2\mu_R\\ \text{USL}&=+0.2\mu_R\end{aligned}\right\}=0.2\cdot 82.559\cong\pm 16.61 \tag{5.7.2209}$$

---

[83] Still by respecting the scientific publication rules introduced at the beginning of this textbook

The confidence intervals are given by:

$$(80.272 - 82.559) \pm 1.717 \cdot 9.145\sqrt{0.167} \qquad (5.7.2210)$$

Therefore:

$$(\text{LCL}, \text{UCL}) = (-8.698, 4.123) \qquad (5.7.2211)$$

As we have:

$$(\text{LCL}, \text{UCL}) \subset (\text{LSL}, \text{USL}) \qquad (5.7.2212)$$

The equivalence is then considered by experts consensus as "proven".

Now let us come back to the « *Absence of evidence is NOT evidence of absence.*» based on a bayesian approach.

There's a difference between not looking and therefore not seeing any X, and looking and not seeing any X. The latter is 'evidence', the former is not.

So the hypothesis under test is «*There is a unicorn in that field behind the hill.*» Alice stays where she is and doesn't look. If there is a unicorn in the field, Alice sees no unicorns. If there is no unicorn in the field, Alice sees no unicorns. $P(\text{sees no unicorn} \mid \text{is unicorn}) = P(\text{sees no unicorn} \mid \text{no unicorn}) = 1$. When the hypothesis makes no difference to the observation, the 'evidence' contributed by the observation to belief in the hypothesis is zero.

Bob climbs to the top of the hill and looks down on the field, and sees no unicorn. If there is a unicorn in the field, Bob would see it. If there is no unicorn in the field, Bob would see no unicorn. $P(\text{sees no unicorn} \mid \text{is unicorn}) \neq P(\text{sees no unicorn} \mid \text{no unicorn})$. When the hypothesis being true or false changes the probability of the observation, evidence is contributed. Looking and seeing no unicorns in the field is positive evidence that there are no unicorns in the field.

We can quantify evidence using Bayesian probability:

$$\begin{aligned} P(H_1 \mid O) &= \frac{P(O \mid H_1)\,P(H_1)}{P(O)} \\ P(H_2 \mid O) &= \frac{P(O \mid H_2)\,P(H_2)}{P(O)} \end{aligned} \qquad (5.7.2213)$$

where H_1 is «*there is no unicorn in that field*». H_2 is «*there is a unicorn in that field*», and O is «*I see no unicorn*». Divide one by the other:

$$\frac{P(H_1 \mid O)}{P(H_2 \mid O)} = \frac{P(O \mid H_1)}{P(O \mid H_2)}\frac{P(H_1)}{P(H_2)} \qquad (5.7.2214)$$

Take logarithms to make the multiplication additive:

$$\log\left(\frac{P(H_1 \mid O)}{P(H_2 \mid O)}\right) = \log\left(\frac{P(O \mid H_1)}{P(O \mid H_2)}\right) + \log\left(\frac{P(H_1)}{P(H_2)}\right) \qquad (5.7.2215)$$

We interpret this as saying that the Bayesian belief in favour of H_1 over H_2 after the observation is equal to the evidence in favour of H_1 over H_2 arising from the observation plus the Bayesian belief in favour of H_1 over H_2 before the observation. The additive evidence arising from the experiment is quantified as:

$$\log\left(\frac{P\left(O \mid H_1\right)}{P\left(O \mid H_2\right)}\right) \tag{5.7.2216}$$

The dictum is based on people claiming things like that there are no unicorns in the universe based on having looked at only a tiny portion of it and having seen none. Strictly speaking, there is non-zero evidence arising from this, but it's near zero, being related to the log of the volume of space and time searched divided by the volume of the universe.

Distilled to a sentence: absence of evidence is evidence of absence, but the strength of that evidence depends on how strong your test is (how likely it would be to succeed if the thing did exist).

5.7.13.7 Cochran C-test

The Cochran C-test has for purpose to verify the homogeneity of variances for several populations. This is a preliminary or subsequent test (post hoc) helpful test to do before making a balanced ANOVA (balanced) analysis and is recommended by the standard ISO 5725 (as the Tukey's test that we will see much further at page 1207) !

Although the idea of the Cochran C-test is empirical, it is nevertheless intuitive as are the definitions of Dixon and Grubbs tests. Why then do we present in this textbook in details the proof of the Cochran C-test when we have mentioned that we would not do this for the test of Grubbs and Dixon because the approach of these latter was also empirical? The reason is simple in fact: the test of Grubbs and Dixon require, at least as far as we know, Monte Carlo simulations to determine the critical values of rejection or acceptance of the null hypothesis, while the critical value of Cochran C-test can be obtained relatively easily analytically.

That said ... we define the Cochran C-test by the ratio:

$$C_j=\frac{S_j^2}{\displaystyle\sum_{i=1}^{N} S_i^2} \tag{5.7.2217}$$

where the S_i are the unbiased variances of the N different sources of data, each composed of n samples and the null hypothesis is intuitively the equality of variances against the alternative hypothesis that one of the variance is too big (that is to say: bad) and dismissed because considered as an outlier variance.

The ISO 5725 recommends to repeat this test until there is no longer any aberrant variance (therefore too large AND far from other variances).

To determine the critical value, let us invert the definition of Cochran C-test and do some basic algebraic manipulations:

$$C_j^{-1}=\frac{\displaystyle\sum_{i=1}^{N} S_i^2}{S_j^2}=\frac{S_j^2+\left(\displaystyle\sum_{i=1}^{N} S_i^2-S_j^2\right)}{S_j^2}=1+\frac{\displaystyle\sum_{i=1,i\neq j}^{N} S_i^2}{S_j^2} \tag{5.7.2218}$$

We note that pretty much the second term of the last equality looks like a Fisher law. As the Fisher law is not stable by the addition, we should find a way to turn the term:

$$\sum_{i=1,i\neq j}^{N} S_i^2 \tag{5.7.2219}$$

into a single variance. The idea is then relatively simple but still had someone to think about it... We know the S_i are non-biased variances equation with a factor $1/(n-1)$. Therefore if the N samples (levels) are independent, the overall variance is then by stability of the Normal distribution and by taking the traditional notations of the ANOVA:

$$\begin{aligned}S^2_{\text{global}-j} &= \frac{1}{N-1}\sum_{i=1,i\neq j}^{N} S_i^2 = \frac{1}{N-1}\sum_{i=1,i\neq j}^{N}\frac{Q_i}{n-1} = \frac{1}{(N-1)(n-1)}\sum_{i=1,i\neq j}^{N} Q_i\\ &= \frac{1}{(N-1)(n-1)}Q_{\text{global}-j} = \frac{Q_{\text{global}-j}}{(N-1)(n-1)}\end{aligned} \tag{5.7.2220}$$

Therefore:

$$C_j^{-1} = 1+\frac{\frac{N-1}{N-1}\displaystyle\sum_{i=1}^{N} S_i^2}{S_j^2} = 1+(N-1)\frac{\frac{1}{N-1}\displaystyle\sum_{i=1}^{N} S_i^2}{S_j^2} = 1+(N-1)\frac{\frac{Q_{\text{global}-j}}{(N-1)(n-1)}}{\frac{Q_j}{n-1}} \tag{5.7.2221}$$

We recognize in the last equality the ratio of two squared variances. We then have identically to what we've proved in our study of the one-way fixed factor ANOVA without replication:

$$F_{n-1,(N-1)(n-1)} = \frac{S_j^2}{S^2_{\text{global}-j}} = \frac{\frac{Q_j}{n-1}}{Q_{\text{global}-j}}(N-1)(n-1) \tag{5.7.2222}$$

and therefore it comes:

$$C_j^{-1} = 1+\frac{N-1}{F_{n-1,(N-1)(n-1)}} \tag{5.7.2223}$$

which is therefore independent of j and therefore the left-tail Cochran' C-test (since by definition the Cochran's ratio should be as small as possible) will have for critical value:

$$C(\alpha,N,n) = \left(1+\frac{N-1}{F_{n-1,(N-1)(n-1)}}\right)^{-1} \tag{5.7.2224}$$

If we consider a test with a significance level of $1-\alpha$ (thus corresponding to the cumulative probability of not making a Type I error) and we reiterate it independently again as second time. Then, if the tests are independent, following the axiom of probabilities, the probability of not making a Type I error will be given by the product of probabilities:

$$(1-\alpha)(1-\alpha) = (1-\alpha)^2 \tag{5.7.2225}$$

and so on for n tests. We notice so quickly that the cumulative probability of not making a Type I error decreases very quickly. For example, for 10 independent repeated tests with a 5% level, then we have:

$$(1-\alpha)^n = (1-5\%)^{10} \cong 59\% \tag{5.7.2226}$$

which is catastrophic! So if we want a level of confidence on repeated tests of a certain value which we will denote α_r, it is clear that we must solve the following equation:

$$(1-\alpha)^n=(1-\alpha_r) \quad (5.7.2227)$$

Therefore (relation sometimes named "Šidàk equation"):

$$\alpha=1-(1-\alpha_r)^{1/n}=1-\frac{1}{(1-\alpha_r)^n} \quad (5.7.2228)$$

and with a Taylor expansion to the second order it comes (see section Sequences and Series page 1696):

$$\alpha=1-(1-\alpha_r)^{1/n}\cong 1-(1-n\alpha_r)=n\alpha_r \quad (5.7.2229)$$

that we name "Bonferroni approximation" or sometimes also "Boole approximation" or "Dunn approximation".

Remark

In statistics, "family-wise error rate" (FWER) is the probability of making one or more false discoveries, or type I errors when performing multiple hypotheses tests (typically during meta-analysis).

So in the end, we have for the Cochran C-test:

$$C(\alpha,N,n)=\left(1+\frac{N-1}{F_{\alpha/N,n-1,(N-1)(n-1)}}\right)^{-1} \quad (5.7.2230)$$

that we can compute with the English version of Microsoft® Excel® 14.0.6123 using the following formula:

```
=1/(1+(N-1)/FINV(ALPHA/N,n-1,(N-1)*(n-1)))
```

Remark

We benefit here from assuming that all tests are independent of each other. In practical applications, that is often not the case. Depending on the correlation structure of the tests, the Bonferroni correction could be extremely conservative, leading to a high rate of false negatives!!!

5.7.13.8 Family-wise error rate (adjusted p-values)

Definitions (#5.7.152):

D1. The "family-wise error rate" (FWER) (or "per-experiment error rate") is the probability of making any Type I errors at all (typically during meta-analysis).

D2. The "false discovery rate" (FDR) is the expected proportion of false rejections out of all rejections (typically also during meta-analysis).

Before doing the maths, let us recall the famous case of Nosek's team that invited researchers to take part in a crowd-sourcing data analysis project. The setup was simple. Participants were all given the same data set and prompt: Do soccer referees give more red cards to dark-skinned players than light-skinned ones? They were then asked to submit their analytical approach for feedback from other teams before diving into the analysis.

Twenty-nine teams with a total of 61 analysts took part. The researchers used a wide variety of methods, ranging - for those of you interested in the methodological gore - from simple linear regression techniques to complex multilevel regressions and Bayesian approaches. They also made different decisions about which secondary variables to use in their analyses.

Despite analysing the same data, the researchers got a variety of results. Twenty teams concluded that soccer referees gave more red cards to dark-skinned players, and nine teams found no significant relationship between skin color and red cards.

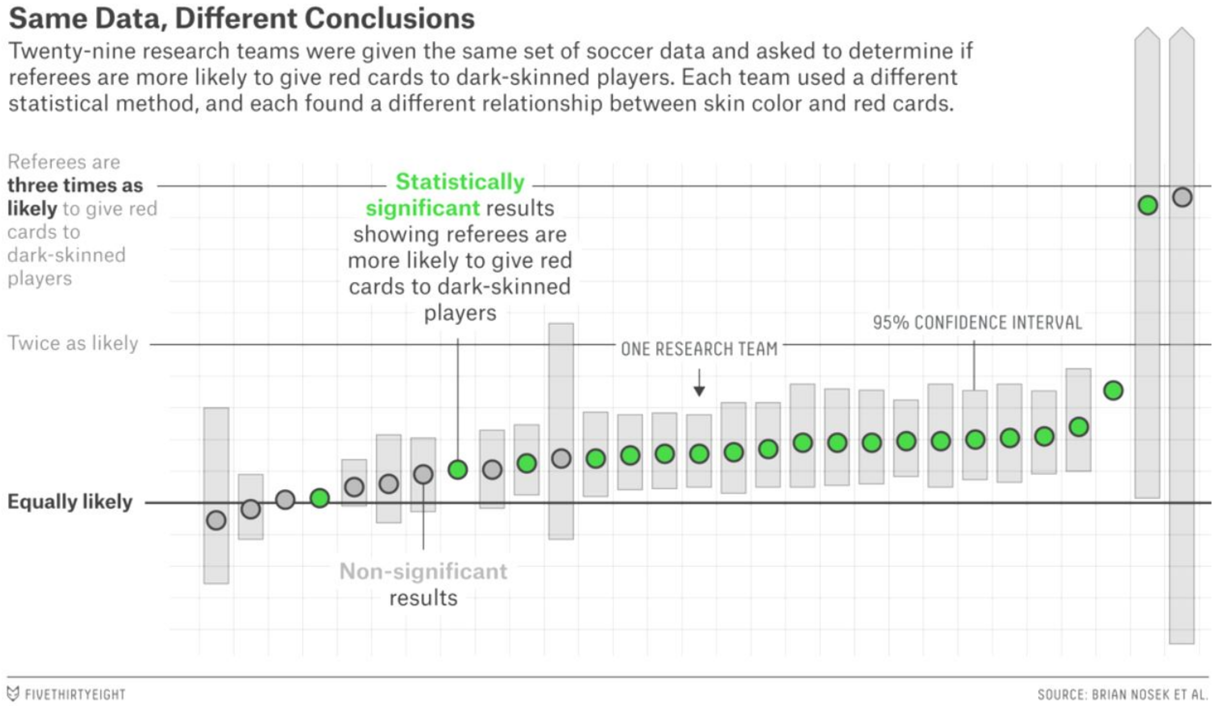


Figure 5.205: **Same data, different conclusions (purpose of meta-analysis)**

The variability in results wasn't due to fraud or sloppy work. These were highly competent analysts who were motivated to find the truth, said Eric Luis Uhlmann, a psychologist at the Insead business school in Singapore and one of the project leaders. Even the most skilled researchers must make subjective choices that have a huge impact on the result they find.

That's why knowing how to deal with multiple p-value is critical and especially when lives are concerned by the result (remember the meta-analysis for the influence of Hydroxychloroquine during the COVID-19 period in year 12020 according to holocene calendar).

The most commonly used method which controls FWER at level α is the Bonferroni's method seen just earlier above. Let us make some recall about that latter!

- The probability of making a Type I error when dealing only with a specific test is denoted $\alpha[\text{PT}]$ (pronounced "alpha per test"). It is also named the "testwise alpha".
- The probability of making at least one Type I error for the whole family of tests denoted $\alpha[\text{PF}]$ (pronounced "alpha per family of tests"). It is also named the "familywise alpha" or the "experimentwise alpha".

During our study of the Cochran C-test earlier above (see page 1105) we have seen that the Bonferroni correction for multiple tests was then written by (assuming independence!):

$$\boxed{\alpha[\text{PF}] = 1 - (1 - \alpha[\text{PT}])^n} \tag{5.7.2231}$$

where n is the number of tests.

Also during our study of the Cochran C-test above we have derived the following first order Taylor approximation:

$$\alpha[\text{PF}] \cong n\alpha[\text{PT}] \tag{5.7.2232}$$

The Šidàk-Bonferonni equations can be used to find the value of $\alpha[\text{PT}]$ when $\alpha[\text{PF}]$ is fixed. For example, suppose that we want to perform four independent tests, and we want to limit the risk of making at least one Type I error to an overall value of $\alpha[\text{PF}] = 0.05$, we will consider a test significant if its associated probability is smaller than (in the case of four associated tests):

$$\alpha[\text{PT}] = 1 - (1 - \alpha[\text{PF}])^{1/n} = 1 - (1 - 0.05)^{1/4} = 0.0127 \tag{5.7.2233}$$

With the Bonferonni approximation, a test reaches significance if its associated probability is smaller than:

$$\alpha[\text{PT}] = \frac{\alpha[\text{PF}]}{n} = \frac{0.05}{4} = 0.0125 \tag{5.7.2234}$$

Remark

Sometimes we may have:

$$\alpha[\text{PF}] \cong n\alpha[\text{PT}] > 1 \tag{5.7.2235}$$

This is why it is not uncommon in the literature to find that latter relation under the form:

$$\alpha[\text{PF}] \cong \{n\alpha[\text{PT}], 1\}_1 \tag{5.7.2236}$$

or:

$$\alpha[\text{PF}] \cong \min\{n\alpha[\text{PT}], 1\} \tag{5.7.2237}$$

that simply means that if $n\alpha[\text{PT}] > 1$, then we replace it by 1.

✍**Example:**

Consider four null hypotheses $H_1, \ldots, H_4$ with p-values:

$$p_1 = 0.01, p_2 = 0.04, p_3 = 0.03, p_4 = 0.005$$

to be tested at significance level $\alpha = 0.05$. First we test $0.01 \cdot 4 = 0.04 < 0.05$: reject $H_{0(1)}$. After we test $0.04 \cdot 4 = 0.16 > 0.05$: Don't reject $H_{0(2)}$. After we test $0.03 \cdot 4 = 0.12 > 0.05$: Don't reject $H_{0(3)}$. Finally we test $0.005 \cdot 4 = 0.02 < 0.05$: Reject $H_{0(4)}$.

The Bonferroni method is guaranteed to control FWER, but it has a big problem. It greatly reduces our power to detect real differences. For example, suppose the effect size is 2 and we are doing a Student T-test, rejecting the null hypothesis for $p < 0.05$. With 10 observations per group, the power is 99%. Now suppose we have $1,000$ tests, and use the Bonferroni method. That means that to reject the null hypothesis, we need $p < 0.00005$. The power is now only 29%. If we have $10,000$ tests (which is small for genomics studies) the power is only 10%.

5.7.13.8.1 Holm-Bonferroni Correction

The "Holm-Bonferroni correction" method is as follows:

- Let $H_1, \ldots, H_m$ be a family of m null hypotheses and $p_1, \ldots, p_m$ the corresponding p-values.
- Start by ordering the p-values (from lowest to highest) $p_{(1)} \ldots p_{(m)}$ and let the associated hypotheses be $H_{(1)}, \ldots, H_{(m)}$.
- For a given significance level α, let k be the minimal index such that $p_{(k)} > \frac{\alpha}{m+1-k}$.
- Reject the null hypotheses $H_{(1)}, \ldots, H_{(k-1)}$ and do not reject $H_{(k)}, \ldots, H_{(m)}$.
- If $k = 1$ then do not reject any of the null hypotheses and if no such k exist then reject all of the null hypotheses.

This method ensures that FWER $\leq \alpha$, where FWER is the family-wise error rate.

Proof 5.48.6. Holm-Bonferroni controls the FWER as follows: Let $H_{(1)}, \ldots, H_{(m)}$ be a family of hypotheses, and $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$ be the sorted p-values. Let I_0 be the set of indices corresponding to the (unknown) true null hypotheses, having m_0 members.

Let us assume that we wrongly reject a true hypothesis. We have to prove that the probability of this event is at most α.

Let h be the first rejected true hypothesis (first in the ordering given by the Holm-Bonferroni test). Then $H_{(1)}, \ldots, H_{(h-1)}$ are all rejected false hypotheses and $h - 1 \leq m - m_0$. From there, we get:

$$\frac{1}{m-h+1} \leq \frac{1}{m_0} \tag{5.7.2238}$$

as h is rejected, we have:

$$p_{(h)} \leq \frac{\alpha}{m-h+1} \tag{5.7.2239}$$

by definition of the test. Using the previous relation the right hand side is at most α/m_0. Thus, if we wrongly reject a true hypothesis, there has to be a true hypothesis with p-value at most α/m_0.

So let us define the random variable $A = \left\{P_i \leq \frac{\alpha}{m_0} \text{ for } i \in I_0\right\}$. Whatever the (unknown) set of true hypotheses I_0 is, we have $P(A) \leq \alpha$ (by the Bonferroni inequalities). Therefore, the probability to reject a true hypothesis is at most α.

■ Q.E.D.

✍**Example:**

Consider four null hypotheses $H_1, \ldots, H_4$ with p-values:

$$p_1 = 0.01, p_2 = 0.04, p_3 = 0.03, p_4 = 0.005$$

to be tested at significance level $\alpha = 0.05$. Since the procedure is step-down, we first test $H_4 = H_{(1)}$, which has the smallest p-value $p_4 = p_{(1)} = 0.005$. The p-value is compared to:

$$\frac{\alpha}{m+1-k} = \frac{\alpha}{4+1-1} = \frac{0.05}{4} = 0.0125$$

the null hypothesis is rejected and we continue to the next one. Since $p_1 = p_{(2)} = 0.01$ and:

$$\frac{\alpha}{m+1-k} = \frac{\alpha}{4+1-2} = \frac{0.05}{3} = 0.0167$$

we reject $H_1 = H_{(2)}$ as well as $0.01 < 0.167$ and continue. The next hypothesis H_3 is not rejected since $p_3 = p_{(3)} = 0.03$ and:

$$\frac{\alpha}{m+1-k} = \frac{\alpha}{4+1-3} = \frac{0.05}{2} = 0.025$$

then $0.03 > 0.025$. We have our k that is equal to $k = 3$. We stop testing and conclude that H_1 and H_4 are rejected and H_2 and H_3 are not rejected while controlling the family-wise error rate at level $\alpha = 0.05$. Note that even though $p_2 = p_{(4)} = 0.04 < 0.05 = \alpha$ applies, H_2 is not rejected. This is because the testing procedure stops once a failure to reject occurs.

5.7.13.8.2 Holm–Šidák Correction

As we know the Bonferroni correction seen earlier above is only an approximation (first order Taylor series) of the Šidák correction.

Therefore it is possible to replace:

$$\frac{\alpha}{m},\frac{\alpha}{m-1},\ldots,\frac{\alpha}{1} \tag{5.7.2240}$$

with:

$$1-(1-\alpha)^{1/m},1-(1-\alpha)^{1/(m-1)},\ldots,1-(1-\alpha)^{1} \tag{5.7.2241}$$

resulting in a slightly more powerful test and named the "Holm–Šidák Correction".

5.7.13.9 Adequation Tests (Goodness of Fit tests)

The goodness of fit (GoF) of a statistical model describes how well it fits a set of observations. Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected under the model in question. Such measures can be used in statistical hypothesis testing, e.g. to test for normality of residuals, to test whether two samples are drawn from identical distributions (see Kolmogorov–Smirnov test further below), or whether outcome frequencies follow a specified distribution (see Pearson's chi-squared test below).

5.7.13.9.1 Pearson's Chi-squared GoF Test

We will study here our first GoF nonparametric test, certainly one of the most known and most simple one (which applies only to non-censored data as far as we know).

To introduce this test, assume that a random variable follow a probability distribution P. If we draw a sample from the population corresponding to this law, the observed distribution, named "sampling distribution", always deviate more or less of the theoretical distribution, taking into account the sampling fluctuations.

Generally, we do not know the shape of the law P, nor the value of its parameters. It is the nature of the studied phenomenon and the analysis of the observed distribution that allow to choose a law likely to be adequate and afterwards to estimate the parameters.

The differences between the theoretical law and the observed distribution can be attributed either to sampling fluctuations, or to the fact that the phenomenon does not follow, in reality, the supposed law.

Basically, if the gaps are small enough, we will assume they are due to random fluctuations and we will accept (not reject in fact!) the supposed law. On the contrary, if the gaps are too high, we conclude that they can not be explained solely by the fluctuations and the phenomenon does not follow the supposed law (reject the null hypothesis).

To assess these gaps and to make a decision, we need:

1. Define the measure of the distance between the empirical distribution and the theoretical resulting from the retention law.

2. Determine the probability law followed by the random variable giving the distance (in fact sadly not reject the null hypothesis).

3. State a decision rule to tell from the observed distribution, if the law chosen is acceptable or not.

First, we will need for this purpose the central limit theorem and secondly recall that during the construction of the Normal distribution, we have proved that the variable:

$$x=\frac{k-np}{\sqrt{npq}}=\frac{k-np}{\sqrt{np(1-p)}} \tag{5.7.2242}$$

follow a Normal distribution centered reduced when n approaches infinity (Laplace condition) and the probability p was very small.

In practice, the approximation is quite acceptable... in some companies... and when $np>5$ and $p\leq 0.5$ therefore (it was one of the terms that needed to tend to zero when we made the proof):

$$\frac{q}{\sqrt{npq}}\cong 0.3 \tag{5.7.2243}$$

For example in the two figures below where we represented the binomial laws approached by the Normal associated laws, we have on the left $n=60,p=1/6,np=10$ and on the right $n=40,p=0.05,np=2$:

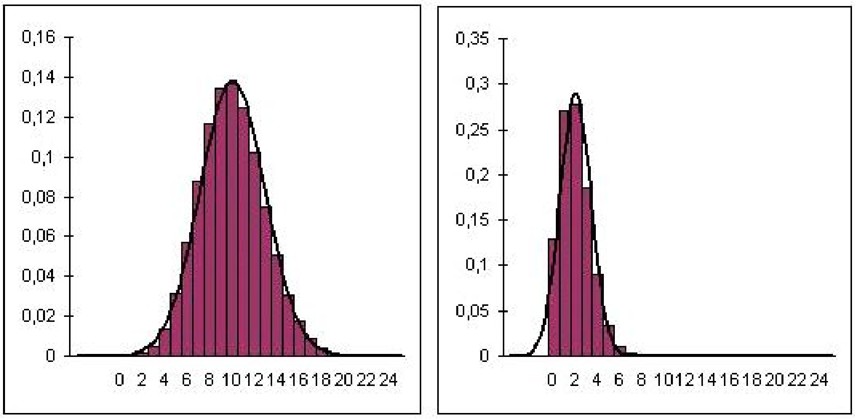


Figure 5.206: **Approach of binomial functions by associated Normal functions**

Finally, remember that we have proved that the sum of squares of n linearly independent reduced centered Normal random variable follows a chi-square with n degrees of freedom denoted by $\chi^2(n)$.

Now consider a random variable X that follows a theoretical distribution function (continuous or discrete) P and let us draw a sample of size n in the population corresponding to this law P.

The n observations are distributed along k terms (class values) $C_1,C_2,\ldots,C_k$, whose probabilities $p_1,p_2,\ldots,p_k$ are determined by the distribution function P (refer to the example with the Henry straight line at page 726).

For each modality C_i, the empirical sample size is a binomial random variable k_i of parameters:

$$\beta(n, p_i) \tag{5.7.2244}$$

This number k_i corresponds indeed to the number of successful events: "result equal to the modality C_i" of probability p_i, obtained during the sampling on n items of the experimental batch (and not in the population of the theoretical law as before!).

We have proved earlier above during our the study of the binomial law that his expected mean was:

$$\text{E}(k_i) = np_i \tag{5.7.2245}$$

represents the expected theoretical sample size of the modality C_i and its variance is given by (when n is very big and p_i very small):

$$\text{V}(k_i) = np_iq_i = np_i(1 - p_i) \cong np_i \tag{5.7.2246}$$

Its standard deviation is therefore:

$$\sigma_{k_i} = \sqrt{np_i} \tag{5.7.2247}$$

Under these conditions, provided that the modality C_i has a size np_i of at least equal to 5, the reduced centered variable:

$$E_i = \frac{k_i - np_i}{\sqrt{np_i}} = \frac{k_i - \mu_i}{\sigma_i} \tag{5.7.2248}$$

between empirical and theoretical sample size can be approximately regarded as a reduced centered Normal variable as we have proved it earlier above. It is also named the "Pearson's residual" for value i, and it compares the observed with the expected counts. The sign (positive or negative) indicates whether the observed frequency k_i is higher or lower than the value fitted under the model, and the magnitude indicates the degree of departure. When data do not fit a model, examination of the Pearson residuals often helps to diagnose where the model has failed.

We now define the variable:

$$D = \sum_{i=1}^{N} E_i^2 = \sum_{i=1}^{N} \frac{(k_i - np_i)^2}{np_i} \tag{5.7.2249}$$

where k_i is often named "experimental frequency" and np_i "theoretical frequency".

If we take the square it is because that in a simple sum certain terms would cancel by opposing effects and thus would mask the differences, if we take the sum of the absolute values the statistical table of D would be difficult to build and the test would be not very robust because of the small gap of distances. The square allows not only to have an easy statistical table of D which is simple since it is based on a law with a single parameter, as we shall see, and the square also sufficiently increases the test's robustness.

Note that this variable is also sometimes (somewhat unfortunately) denoted by:

$$D = \sum_{i=1}^{N} \frac{(n_i - n_i')^2}{n_i'} \tag{5.7.2250}$$

or more often (the E_i below is not the same as the previous E_i just above!!!):

$$\boxed{D=\sum_{i=1}^{N}\dfrac{(O_i-E_i)^2}{E_i}} \tag{5.7.2251}$$

where:

$$\dfrac{(O_i-E_i)^2}{E_i} \tag{5.7.2252}$$

is commonly named, as already mentioned earlier above, the "Pearson's residual".

This variable D, sum of squared variables E_i, provides a measure of a "distance" between empirical and theoretical distribution and empirical distribution. Let us note however that this is not a distance in the usual mathematical sense (topological).

Recall that D can therefore also be written:

$$D=\sum_{i=1}^{N}\dfrac{(k_i-np_i)^2}{np_i}=\sum_{i=1}^{N}\dfrac{(k_i-\mu_i)^2}{\sigma_i^2} \tag{5.7.2253}$$

D is therefore the sum of squares of N reduced centered Normal random variables linked by the single linear relation:

$$k_1+k_2+\ldots+k_N=n \tag{5.7.2254}$$

where n is the sample size. So D follows a chi-square distribution, but with $N-1$ degrees of freedom, so a degree less because of the unique linear relation between them! Indeed, recall that the degree of freedom is the number of independent variables in the sum and not just the number of summed terms.

Therefore:

$$\boxed{D=\chi^2(N-1)=\chi^2_{N-1}=\sum_{i=1}^{N}\dfrac{(k_i-np_i)^2}{np_i}} \tag{5.7.2255}$$

We name this test the "nonparametric Chi-square test" or "Pearson's Chi-square test" or "Chi-square test of adjustment" or "Karl Pearson's test" or "goodness of fit Chi-square test"...

Remark

There are plenty of statistical tests with that contains "chi-square" in it. For example:

- Chi-square Goodness-of-fit test (allow us to determine whether a specified population distribution is valid)
- Chi-square test of association/independence (allow us to determine whether the distribution of one variable has been influenced by another variable)
- Chi-square test of homogeneity (allow us to compare two or more population proportions)

- Chi-square test for outliers (allow us to detect if there is at least one outlier in comparison of a specific distribution)
- Chi-square test for difference of two count data (special case of homogeneity test with only one category!)

Even if all those tests have different purposes, their mathematical framework and procedure are EXACTLY the same and therefore any of these four tests can be run to get at the same time the four conclusions with the same p-values and the same critical chi-square values!!!

We also see by the way that the squared Pearson residual is the individual contribution to the Pearson χ^2 statistic.

Then, the usage is to determine the value of the Chi-square distribution with $N-1$ degrees of freedom with a 5% probability of being exceeded. Thus, in the hypothesis that the studied phenomenon, follows the theoretical distribution P, so there is a 95% cumulative probability that the variable D takes a value less than the one given by the Chi-Square distribution.

If the value of the law of Chi-square obtained from the sample is smaller than that corresponding to 95% of cumulative probability, we do not reject the null hypothesis that the phenomenon follows the law P.

Remarks

R1. The fact that the assumption of the law P is accepted does not mean that this hypothesis is true, merely that the information given by the sample does not allow to reject it. Similarly, the fact that the assumption of the law P is rejected does not necessarily mean that this assumption is false but that the information provided by the sample rather lead to the conclusion that the inadequacy of such a law.

R2. For the variable D to follow a chi-square law, it is necessary that the theoretical values np_i of the different modalities C_i are at least equal to 5, that the sample was drawn randomly (no correlation) and that the probabilities p_i are not close to zero.

This goodness of fit test however suffers from a major issue: it requires to group the measurements in classes C_i and in practice there is no absolute theorem (at least as far we know) to choose the number of classes (and in full width) excepted the Sturge Rule proved earlier. It is this reason that make the chi-squared goodness of fit test is reserved for discrete distributions where the problem of the choice of classes not arise.

However, we will need to create goodness of fit tests that do not require the use of classes and we will see just after ad hoc tools for this purpose (Kolmogorov-Smirnov or Anderson-Darling to name only the most important one).

✍**Example:**

Suppose that the births in a hospital for a period of time, are as follows:

Day	**M**	**T**	**W**	**T**	**F**	**S**	**S**	**Total**
Observations	120	130	125	128	80	70	75	728

We note that there were a total of 728 births. We ask then the following question: How many should there be births, in theory, every day if there is no difference between days? This represents the null hypothesis H_0. In fact the null hypothesis states that the differences between the observed frequencies and conceptual frequencies are relatively small. We take for granted that if there is no difference there should be the same number of births each day. Since there are a total of 728 births for 7 days in theory there should be $728/7 = 104$ births every day. So now we have the following table:

Day	**M**	**T**	**W**	**T**	**F**	**S**	**S**	**Total**
Observations	120	130	125	128	80	70	75	728
Expected	104	104	104	104	104	104	104	104

The total of observed frequencies equals the total expected frequencies. The purpose is therefore to examine the difference between the observed and expected frequencies (assumed to follow a uniform law in this special case) using the Chi-square relation. In other words, we do a fit test between an empirical distribution function (observed) and the uniform distribution function. Then we have:

$$\begin{aligned}
\chi^2(N-1) &= \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i} \\
&= \frac{(120-104)^2}{104} + \frac{(130-104)^2}{104} + \frac{(125-104)}{104^2} + \frac{(128-104)^2}{104} \\
&+ \frac{(80-104)^2}{104} + \frac{(70-104)^2}{104} + \frac{(75-104)^2}{104} \\
&= \frac{16^2}{104} + \frac{26^2}{104} + \frac{21^2}{104} + \frac{24^2}{104} + \frac{(-24)^2}{104} + \frac{(-34)^2}{104} + \frac{(-29)^2}{104} \\
&= \frac{256}{104} + \frac{676}{104} + \frac{441}{104} + \frac{576}{104} + \frac{576}{104} + \frac{1156}{104} + \frac{841}{104} \\
&= 2.46 + 6.5 + 4.24 + 5.54 + 5.54 + 11.12 + 8.09 = 43.49
\end{aligned} \tag{5.7.2256}$$

The χ^2 is therefore 43.49. As such this number means little. This result should be interpreted with the help of the table of critical values of the χ^2_6. Without using the table we understand that it is very unlikely that the observed frequency and theoretical frequency are identical. We accept that there may be some difference (we therefore reject the null hypothesis H_0 in favour of the alternative hypotheses H_A).

So do not forget that this test only applies to uncensored data, that is to say for which the intervals are closed!

Remark

The above test is also used a feature selection technique in Machine Learning and is then named "χ^2 feature selection". The idea is that if we have a dataset with too many columns (categorical or continuous), we which to eliminated some of them by investigating if they are statistically dependent (then we can eliminate one of them) or independent (then we can't eliminate any of them). Notice that to run this technique on continuous variables (columns), we can just group them into bins!

5.7.13.9.2 G-test GoF

We use the "G–test of goodness-of-fit" or also named "G^2-test" (that is a special case of the "likelihood ratio test" or "log-likelihood ratio test" as we will see further below at page 1150) when we have one nominal variable and that we want to see whether the number of observations in each category fits a theoretical expectation, and the sample size is large and the observed total of counts and the expected total of counts is equal!

The statistical null hypothesis is that the number of observations in each category is equal to that predicted by a theory, and the alternative hypothesis is that the observed numbers are different from the expected. The null hypothesis is usually an extrinsic hypothesis, where we know the expected proportions before doing the experiment.

The G-test statistic is proportional to the Kullback–Leibler divergence (see section Statistical Mechanics page 3542) of the theoretical distribution from the empirical distribution:

$$\boxed{G=2\sum_i O_i\cdot\ln\left(\frac{O_i}{E_i}\right)=2\sum_i No_i\cdot\ln\left(\frac{No_i}{Ne_i}\right)=2N\sum_i o_i\cdot\ln\left(\frac{o_i}{e_i}\right)=2N\,D_{\text{KL}}(o\|e)} \tag{5.7.2257}$$

Let us put now:

$$O_i=E_i+\delta_i \tag{5.7.2258}$$

with the constraint that:

$$\sum_i \delta_i \tag{5.7.2259}$$

so that the total number of counts stays the same (i.e. $\sum_i O_i=\sum_i E_i=N$). The G-test is then:

$$G=2\sum_i(E_i+\delta_i)\log\left(\frac{E_i+\delta_i}{E_i}\right)=2\sum_i(E_i+\delta_i)\log\left(1+\frac{\delta_i}{E_i}\right) \tag{5.7.2260}$$

If we Taylor expand this around:

$$\frac{\delta_i}{E_i}\cong 0 \tag{5.7.2261}$$

noticing that this assume $E_i \gg \delta_i$ and hence $E_i \gg 0$ (as for the chi-square test of adequation seen just earlier), then we know that for the logarithm (see section Sequences and Series page 1703):

$$\log(1+x) \cong x - \frac{x^2}{2} \qquad (5.7.2262)$$

Then we get:

$$\begin{aligned}
G &= 2\sum_i (E_i+\delta_i)\log\left(1+\frac{\delta_i}{E_i}\right) \cong 2\sum_i (E_i+\delta_i)\left(\frac{\delta_i}{E_i}-\frac{1}{2}\frac{\delta_i^2}{E_i^2}\right)\\
&= 2\sum_i\left(\delta_i-\frac{1}{2}\frac{\delta_i^2}{E_i}+\frac{\delta_i^2}{E_i}-\frac{1}{2}\frac{\delta_i^3}{E_i^2}\right) = 2\sum_i\left(\delta_i+\frac{1}{2}\frac{\delta_i^2}{E_i}+\frac{1}{2}\frac{\delta_i^3}{E_i^2}\right)\\
&\cong 2\sum_i\left(\delta_i+\frac{1}{2}\frac{\delta_i^2}{E_i}\right) = 2\sum_i\left(\frac{2E_i\delta_i+\delta_i}{2E_i}\right) = \sum_i\left(\frac{2E_i\delta_i+\delta_i^2}{E_i}\right) = \sum_i\left(\frac{(E_i+\delta_i)^2-E_i^2}{E_i}\right)\\
&= \sum_i\left(\frac{O_i^2-E_i^2}{E_i}\right)
\end{aligned} \qquad (5.7.2263)$$

and so, we see that $G=\chi^2$ when:

- O_i is close to E_i
- $\sum_i E_i = \sum_i O_i = N$
- $E_i \gg \delta_i$ i.e. $E_i \gg 0$

However, the more O_i and E_i are different, the less well this approximation will work, and the χ^2 will tend to compute erroneous answers. The effects of a single outlier in a small sample set will be more pronounced, which explains why the χ^2 often fails in situations with little data.

Since the χ^2 value is just an approximation to the G-value, the G-value can also be used in the chi-squared probability test. It seems that actually the G-test p-value need to be computed using Monte Carlo techniques.

Remark

If the expected number of observations in any category is too small, the G–test may give inaccurate results, and we should use an exact test instead (Fisher exact test in case of independence as seen at page 1254, McNemar test in case of dependence as seen at page 1264). In practice, it's a good idea to compute both χ^2 and G to see if they lead to similar results. If the resulting p-values are close, then we can be fairly confident that the large-sample approximation is working well. Obviously the G–test of goodness-of-fit assumes independence!

5.7.13.9.3 Kolmogorov-Smirnov GoF Test

In statistics, the Kolmogorov-Smirnov test is also a goodness of fit test (especially well known as normality test) based on an empirical distance used to determine whether the distribution of a sample follows a well known law given by a continuous distribution function (or for comparing two samples and check if they are dependent or not as similar or not). This test, as well as that of chi-squared GoF test is only valid for non-censored data (at least not without correction obtained by numerical simulations).

To introduce this test, we chose the Lilliefors approach that give the possibility to avoid complex calculations. Furthermore, softwares that provide the "Lilliefors GoF test" do not offer the Kolmogorov-Smirnov test since the latter is correct only asymptotically (which is the case of the software Tanagra 4.14).

Imagine we want to build a nonparametric GoF test who works for both discrete and continuous laws without suffering the same problem as the Chi-square GoF test (clustering in classes).

To build this test, we start from the empirical distribution function already defined at the beginning of this section and given for reminder by:

$$\hat{F}_n(x)=\frac{1}{n}\sum_{i=1}^{n}1_{x_i\leq x}=\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i\leq x}=\frac{\#(i:x_i\leq x)}{n} \tag{5.7.2264}$$

that obviously belong to the interval $[0,1]$.

Let us now denote by $F(x)$, the true supposed law which analytical expression is known and with which we would like to compare $\hat{F}(x)$ and build the distance:

$$D_n(x)=\hat{F}_n(x)-F(x) \tag{5.7.2265}$$

The reference distribution may, however, also originate from another measurement sample. The idea is then simply to compare two empirical distributions. We speak then of "Kolmogorov-Smirnov test for 2 independent samples". Some softwares also manage empirically the case where the two samples do not have the same size.

The problem with this choice of distance is ... what x should we then choose to make a test? Well to answer it is simple to see that it would be foolish to take a x for which this distance is minimal, because have a D_n that can be almost zero does not add much information... Therefore, we rather postponed towards the greatest absolute deviation. Which brings us to redefine the distance D_n as follows:

$$\boxed{D_n(x)=\sup\left|\hat{F}_n(x)-F(x)\right|=\sup\left|\frac{1}{n}\sum_{i=1}^{n}1_{x_i\leq x}-F(x)\right|} \tag{5.7.2266}$$

where D_n is named "Kolmogorov-Smirnov empirical distribution" (for sure we should prove rigorously that it is really a distribution ... but for now it is too complex in terms of the content of this textbook however this can be verified by numerical simulations!). Before going further with respect to the theory, let's look at a practical example (because the example is long we will not put it into the conventional box).

Suppose we measured the following five values:

$$-1.2, 0.2, -0.6, 0.8, -1.0 \tag{5.7.2267}$$

thus ordered:

$$x_1 = -1.2, x_2 = -1.0, x_3 = -0.6, x_4 = 0.2, x_5 = 0.8 \tag{5.7.2268}$$

We want to test the following null hypothesis:

$$H_0 : \hat{F}(x) = \Phi(x) \tag{5.7.2269}$$

where $\Phi(x)$ is as usual the distribution function of the Normal centered reduced distribution.

The empirical distribution function will be given by:

$$\hat{F}(x) = \begin{cases} 0 = \dfrac{0}{n} = \dfrac{0}{5} & \text{for} \quad x < -1.2 \\ 0.2 = \dfrac{1}{n} = \dfrac{1}{5} & \text{for} \quad -1.2 < x < -1 \\ 0.4 = \dfrac{2}{n} = \dfrac{2}{5} & \text{for} \quad -1.0 < x < -0.6 \\ 0.6 = \dfrac{3}{n} = \dfrac{3}{5} & \text{for} \quad -0.6 < x < +0.2 \\ 0.8 = \dfrac{4}{n} = \dfrac{4}{5} & \text{for} \quad +0.2 < x < +0.8 \\ 1 = \dfrac{5}{n} = \dfrac{5}{5} & \text{for} \quad x > +0.8 \end{cases} \tag{5.7.2270}$$

Then we traditionally build the following table:

x	$\hat{F}(x)$	$\Phi(x)$	$\lvert\hat{F}(x) - \Phi(x)\rvert$
-1.2^{-}	0	0.115	0.115
-1.2	0.2	0.115	0.085
-1.0^{-}	0.2	0.159	0.041
-1.0	0.4	0.159	0.241
-0.6^{-}	0.4	0.274	0.126
-0.6	0.6	0.274	**0.326**
$+0.2^{-}$	0.6	0.580	0.020
$+0.2$	0.8	0.580	0.220
$+0.8^{-}$	0.8	0.788	0.012
$+0.8$	1	0.788	0.212

Often associated with the following graph comparing empirical (in red) and theoretical (in blue) distributions:

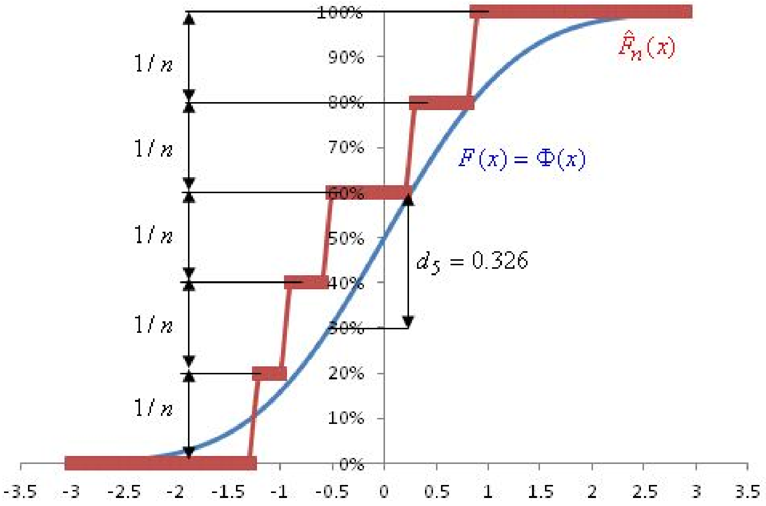


Figure 5.207: **Representation of the approach of the Kolmogorov-Smirnov GoF**

We then observe that the maximum deviation is 0.326. We will denote that for after:

$$d_5 = 0.326 \tag{5.7.2271}$$

that some softwares such as Minitab or `R` denote with the abbreviation: KS.

The reader will have noticed that the biggest deviation above the curve is measured by:

$$D_n^+ = \sup_{i=1}^{n} \left\{ \frac{i}{n} - F(x_i) \right\} = \sup_{i=1}^{n} \left\{ \frac{i}{n} - y_i \right\} \tag{5.7.2272}$$

The largest deviation below the curve is measured by:

$$D_n^+ = \sup_{i=1}^{n} \left\{ F(x_i) - \frac{i-1}{n} \right\} = \sup_{i=1}^{n} \left\{ y_i - \frac{i-1}{n} \right\} \tag{5.7.2273}$$

The biggest deviation is then:

$$D_n = \sup_{i=1}^{n} \left\{ D_n^-, D_n^+ \right\} = \sup_{i=1}^{n} \left\{ \frac{i}{n} - y_i, y_i - \frac{i-1}{n} \right\} \tag{5.7.2274}$$

But what can we do with this value? To what can we compare it? Well the idea is relatively simple and involves generating n values (thus 5 in our example) from the distribution law $F(x)$ of the null hypothesis H_0 and compare them to themselves. In other words, to make a Monte Carlo simulation (see section Numerical Methods page 6708).

Thus, in our example, we generate 5 random value values of $\mathcal{N}(0,1)$ which gives us example with Microsoft® Excel® 11.8346:

```
=NORM.S.INV(RANDBETWEEN(0,1000000)/1000000)
```

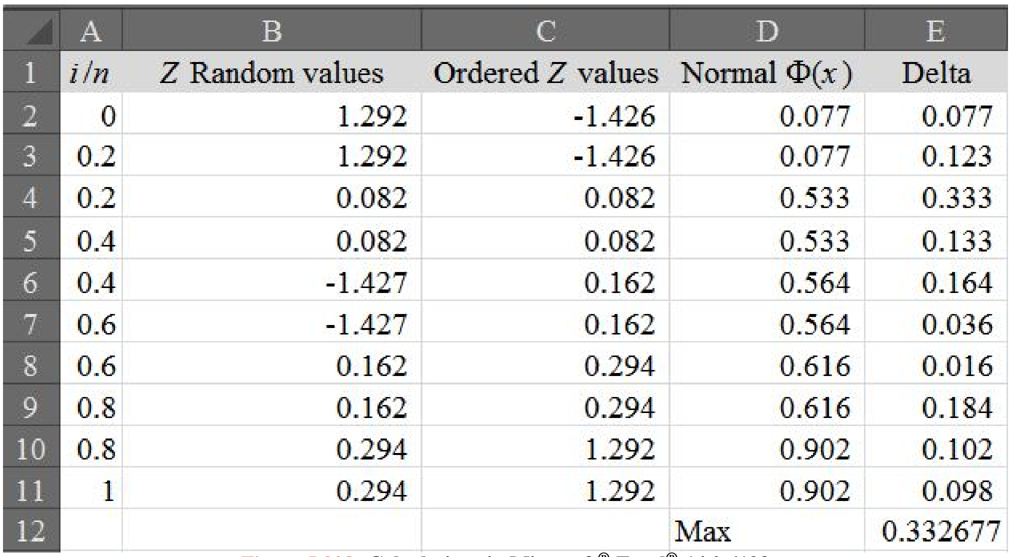
We obtain then 5 values of Z (remember it is the usual notation of a random variable of a Normal centered reduced distribution) which ordered will be for example:

$$-1.427, 0.082, 0.162, 0.294, 1.292 \tag{5.7.2275}$$

and we repeat the same table as before:

x	$\hat{F}(x)$	$\Phi(x)$	$\lvert\hat{F}(x)-\Phi(x)\rvert$
-1.427^-	0	0.077	0.115
-1.427	0.2	0.077	0.085
$+0.082^-$	0.2	0.533	0.041
$+0.082$	0.4	0.533	0.241
$+0.162^-$	0.4	0.564	0.126
$+0.162$	0.6	0.564	**0.326**
$+0.294^-$	0.6	0.616	0.020
$+0.294$	0.8	0.616	0.220
$+1.292^-$	0.8	0.902	0.012
$+1.292$	1	0.902	0.212

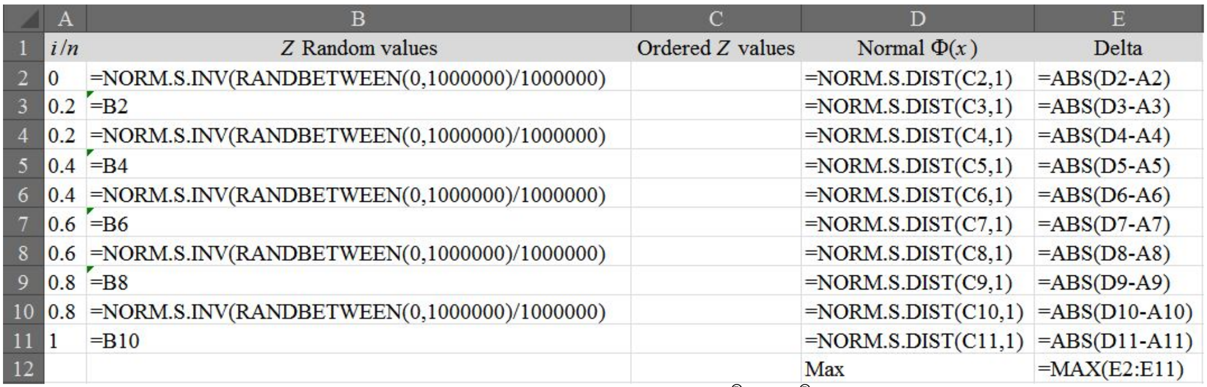
And so we have the maximum deviation that is 0.333. Either with Microsoft® Excel® 14.0.6123:



	A	B	C	D	E
1	i/n	Z Random values	Ordered Z values	Normal $\Phi(x)$	Delta
2	0	1.292	-1.426	0.077	0.077
3	0.2	1.292	-1.426	0.077	0.123
4	0.2	0.082	0.082	0.533	0.333
5	0.4	0.082	0.082	0.533	0.133
6	0.4	-1.427	0.162	0.564	0.164
7	0.6	-1.427	0.162	0.564	0.036
8	0.6	0.162	0.294	0.616	0.016
9	0.8	0.162	0.294	0.616	0.184
10	0.8	0.294	1.292	0.902	0.102
11	1	0.294	1.292	0.902	0.098
12				Max	0.332677

Figure 5.208: Calculations in Microsoft® Excel® 14.0.6123

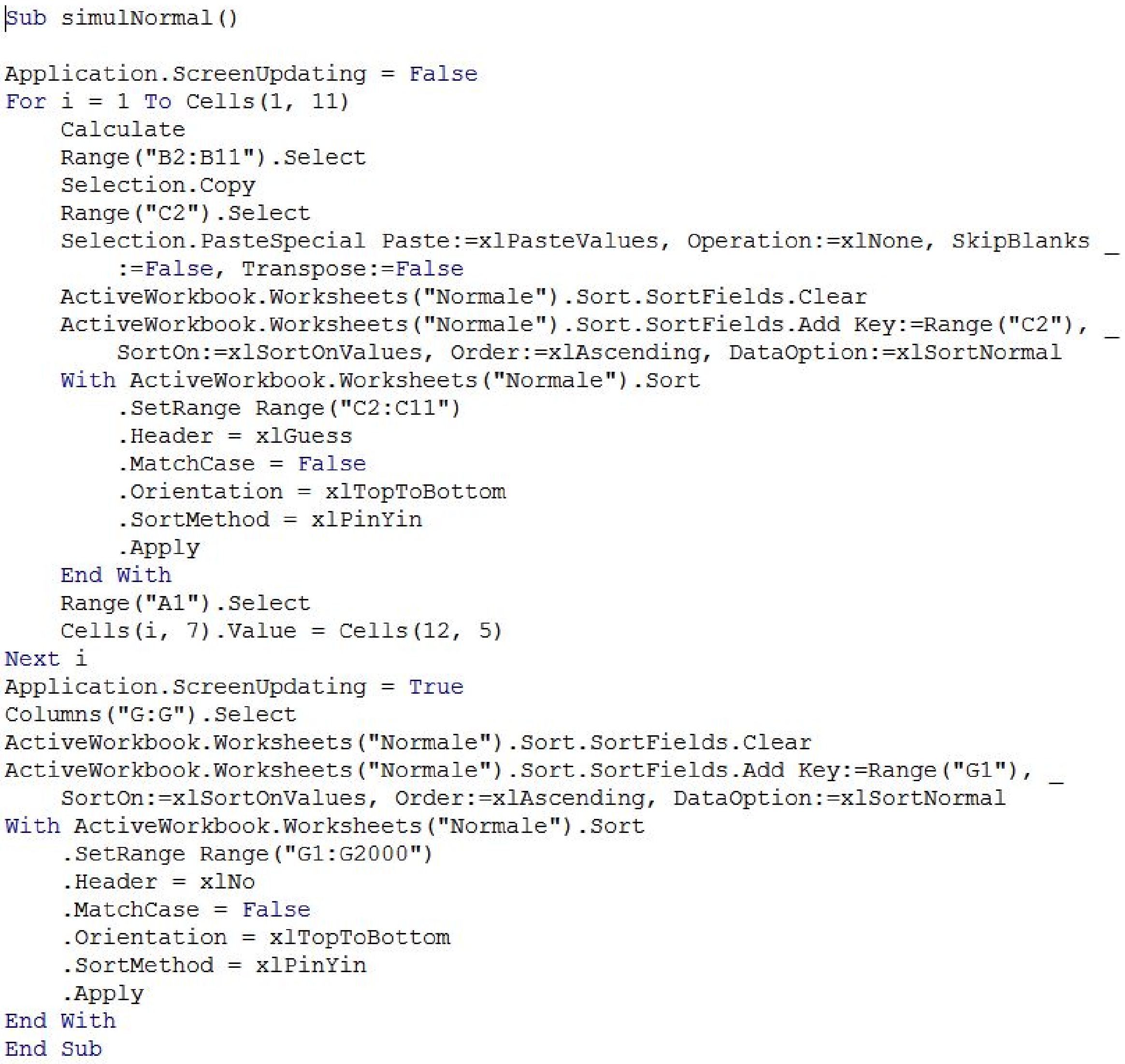
with the explicit formulas:



	A	B	C	D	E
1	i/n	Z Random values	Ordered Z values	Normal $\Phi(x)$	Delta
2	0	=NORM.S.INV(RANDBETWEEN(0,1000000)/1000000)		=NORM.S.DIST(C2,1)	=ABS(D2-A2)
3	0.2	=B2		=NORM.S.DIST(C3,1)	=ABS(D3-A3)
4	0.2	=NORM.S.INV(RANDBETWEEN(0,1000000)/1000000)		=NORM.S.DIST(C4,1)	=ABS(D4-A4)
5	0.4	=B4		=NORM.S.DIST(C5,1)	=ABS(D5-A5)
6	0.4	=NORM.S.INV(RANDBETWEEN(0,1000000)/1000000)		=NORM.S.DIST(C6,1)	=ABS(D6-A6)
7	0.6	=B6		=NORM.S.DIST(C7,1)	=ABS(D7-A7)
8	0.6	=NORM.S.INV(RANDBETWEEN(0,1000000)/1000000)		=NORM.S.DIST(C8,1)	=ABS(D8-A8)
9	0.8	=B8		=NORM.S.DIST(C9,1)	=ABS(D9-A9)
10	0.8	=NORM.S.INV(RANDBETWEEN(0,1000000)/1000000)		=NORM.S.DIST(C10,1)	=ABS(D10-A10)
11	1	=B10		=NORM.S.DIST(C11,1)	=ABS(D11-A11)
12				Max	=MAX(E2:E11)

Figure 5.209: Explicit calculations in Microsoft® Excel® 14.0.6123

with the small corresponding VBA routine quickly and poorly made that will take the number of iterations required in the cell `K1` and will put the empirical distribution of Kolmogorov-Smirnov in column `G` of the active sheet:



```
Sub simulNormal()

Application.ScreenUpdating = False
For i = 1 To Cells(1, 11)
    Calculate
    Range("B2:B11").Select
    Selection.Copy
    Range("C2").Select
    Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
        :=False, Transpose:=False
    ActiveWorkbook.Worksheets("Normale").Sort.SortFields.Clear
    ActiveWorkbook.Worksheets("Normale").Sort.SortFields.Add Key:=Range("C2"), _
        SortOn:=xlSortOnValues, Order:=xlAscending, DataOption:=xlSortNormal
    With ActiveWorkbook.Worksheets("Normale").Sort
        .SetRange Range("C2:C11")
        .Header = xlGuess
        .MatchCase = False
        .Orientation = xlTopToBottom
        .SortMethod = xlPinYin
        .Apply
    End With
    Range("A1").Select
    Cells(i, 7).Value = Cells(12, 5)
Next i
Application.ScreenUpdating = True
Columns("G:G").Select
ActiveWorkbook.Worksheets("Normale").Sort.SortFields.Clear
ActiveWorkbook.Worksheets("Normale").Sort.SortFields.Add Key:=Range("G1"), _
    SortOn:=xlSortOnValues, Order:=xlAscending, DataOption:=xlSortNormal
With ActiveWorkbook.Worksheets("Normale").Sort
    .SetRange Range("G1:G2000")
    .Header = xlNo
    .MatchCase = False
    .Orientation = xlTopToBottom
    .SortMethod = xlPinYin
    .Apply
End With
End Sub
```

Figure 5.210: **Small VBA script for K-S GoF test**

We therefore reiterate the procedure a thousand times and we get the following distribution function (obtained simply by making a scatter chart type in Microsoft® Excel® 14.0.6123 with $2,000$ simulations):

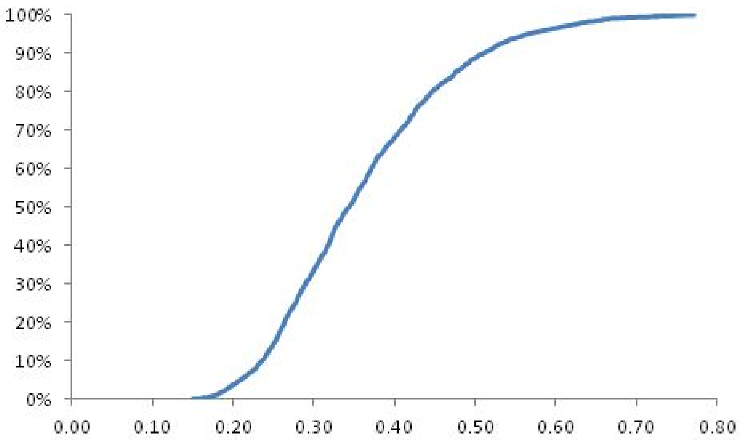


Figure 5.211: **K-S GoF distribution simulation**

and applying a one-sided test with a threshold of $\alpha = 5\%$ we get for the 95th percentile:

$$D_{5,\alpha} \cong 0.564 \tag{5.7.2276}$$

The reader will find the same value in the Kolmogorov-Smirnov tables available in many books. A few thousand simulations are therefore sufficient to restore the values of the tables!

And now, we compare:

$$(D_{5,\alpha} \cong 0.564) > (d_0 \cong 0.326) \tag{5.7.2277}$$

and therefore we do not reject the null hypothesis.

But ... we must still take care with only five values, it is quite likely that the null hypothesis H_0 is not rejected for other distribution laws that the Normal distribution.

Thus, as the reader will have noticed, for each null hypothesis H_0 associated with a particular distribution law, we must tabulate the empirical distribution of Kolmogorov-Smirnov for different values of n and of α using numerical methods. In the majority of books there is only one table with a powerful theorem that shows that in reality, the critical values will be the same.

Remark

Kolmogorov and Smirnov have proved that when n goes very large and that the law of the null hypothesis is continuous, it is no longer necessary to tabulate a Kolmogorov-Smirnov table for each law because we have then:

$$\lim_{n\rightarrow+\infty} P\left(\sqrt{n}D_n \leq x\right) = 1 - 2\sum_{i=1}^{+\infty}(-1)^{k-1}e^{-2\text{i}^2x^2} \tag{5.7.2278}$$

therefore D_n is independent of the distribution law of the null hypothesis H_0. By simulating with the Monte Carlo method, we observe an effective convergence when n exceeds a hundred. But in practice, the vast majority of the time, it is unthinkable to have such a number of measures. Hence the fact that this theoretical result is little used in practice and justifies the absence of proof in this textbook.

To conclude on the K-S GoF test, let us notice to the reader that he will find the mathematical proof of the Anderson-Darling GoF further below.

KOLMOGOROV-SMIRNOV ONE-SIDED TEST

n	0.2	0.1	0.05	0.02	0.01
1	0.9000	0.9500	0.9750	0.9900	0.9950
2	0.6838	0.7764	0.8419	0.9000	0.9293
3	0.5648	0.6360	0.7076	0.7846	0.8290
4	0.4927	0.5652	0.6239	0.6889	0.7342
5	0.4470	0.5094	0.5633	0.6272	0.6685
6	0.4104	0.4680	0.5193	0.5774	0.6166
7	0.3815	0.4361	0.4834	0.5384	0.5758
8	0.3583	0.4096	0.4543	0.5065	0.5418
9	0.3391	0.3875	0.4300	0.4796	0.5133
10	0.3226	0.3687	0.4092	0.4566	0.4889
11	0.3083	0.3524	0.3912	0.4367	0.4677
12	0.2958	0.3382	0.3754	0.4192	0.4490
13	0.2847	0.3255	0.3614	0.4036	0.4325
14	0.2748	0.3142	0.3489	0.3897	0.4176
15	0.2659	0.3040	0.3376	0.3771	0.4042
16	0.2578	0.2947	0.3273	0.3657	0.3920
17	0.2504	0.2863	0.3180	0.3553	0.3809
18	0.2436	0.2785	0.3094	0.3457	0.3706
19	0.2373	0.2714	0.3014	0.3369	0.3612
20	0.2316	0.2647	0.2941	0.3287	0.3524
21	0.2262	0.2586	0.2872	0.3210	0.3443
22	0.2212	0.2528	0.2809	0.3139	0.3367
23	0.2165	0.2475	0.2749	0.3073	0.3295
24	0.2120	0.2424	0.2693	0.3010	0.3229
25	0.2079	0.2377	0.2640	0.2952	0.3166
26	0.2040	0.2332	0.2591	0.2896	0.3106
27	0.2003	0.2290	0.2544	0.2844	0.3050
28	0.1968	0.2250	0.2499	0.2794	0.2997
29	0.1935	0.2212	0.2457	0.2747	0.2947
30	0.1903	0.2176	0.2417	0.2702	0.2899
31	0.1873	0.2141	0.2379	0.2660	0.2853
32	0.1844	0.2108	0.2342	0.2619	0.2809
33	0.1817	0.2077	0.2308	0.2580	0.2768
34	0.1791	0.2047	0.2274	0.2543	0.2728
35	0.1766	0.2018	0.2242	0.2507	0.2690
36	0.1742	0.1991	0.2212	0.2473	0.2653
37	0.1719	0.1965	0.2183	0.2440	0.2618
38	0.1697	0.1939	0.2154	0.2409	0.2584
39	0.1675	0.1915	0.2127	0.2379	0.2552
40	0.1655	0.1891	0.2101	0.2349	0.2521
> 40	$1.07/\sqrt{n}$	$1.22/\sqrt{n}$	$1.36/\sqrt{n}$	$1.52/\sqrt{n}$	$1.63/\sqrt{n}$

Caution! Let us remind what the normality tests answer: *Is there convincing evidence of any deviation from the Gaussian ideal?* With moderately large real data sets, the answer is almost always yes.

The question engineers and scientists often expect the normality test to answer is: *Do the data deviate enough from the Gaussian ideal to "forbid" use of a test that assumes a Gaussian distribution?* Scientists and engineers often want the normality test to be the referee that decides when to abandon conventional (ANOVA, etc.) tests and instead analyze transformed data or use a rank-based nonparametric test or a resampling or bootstrap approach. For this purpose, normality tests are not very useful as many of these tests are robust to some kind of non-normality.

5.7.13.9.4 Ryan-Joiner GoF Test

Consider a random variable X which we know the sampling distribution and for which we would like to check the Normality or not and consider an ordered random variable Y generated by a Normal centered reduced distribution (that's why Ryan-Joiner test is especially well known as normality test). To compare X and Y, we will center X and order its values in ascending order.

For a same given sample size, if the ordered values of X and Y taken in pairs follow the same law, a linear regression based on the other should give a fairly close correlation coefficient of 1. Taking the definition of the squared correlation coefficient, we have then:

$$\rho_{X,Y}^2=\left(\frac{\text{cov}(X,Y)}{\sqrt{\text{V}(X)\text{V}(Y)}}\right)^2 \tag{5.7.2279}$$

Y is assumed to follow a centered reduced Normal distribution. It comes then:

$$\rho_{X,Y}^2=\left(\frac{\text{cov}(X,Y)}{\sqrt{\text{V}(X)\text{V}(Y)}}\right)^2=\left(\frac{\text{E}(XY)-\text{E}(X)\text{E}(Y)}{\sqrt{\text{V}(X)\text{V}(Y)}}\right)^2=\left(\frac{\text{E}(XY)}{\sqrt{\text{V}(X)}}\right)^2=\frac{(\text{E}(XY))^2}{\text{V}(X)} \tag{5.7.2280}$$

and if we take the estimator of the correlation coefficient:

$$R_{X,Y}^2=\frac{\left(\dfrac{1}{n}\displaystyle\sum_{i=1}^n x_iy_i\right)^2}{\dfrac{1}{n}\displaystyle\sum_{n=1}^n x_i^2} \tag{5.7.2281}$$

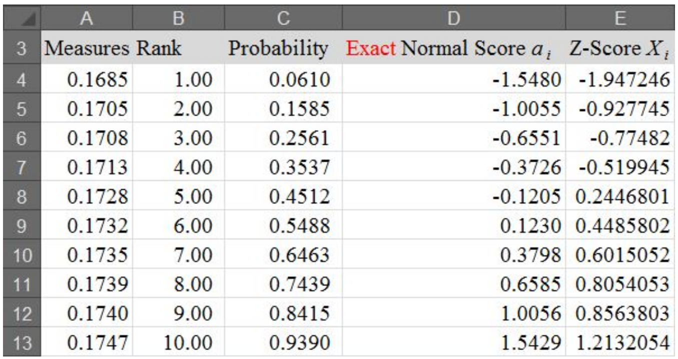
Therefore after simplification:

$$R_{X,Y}^2=\frac{\left(\dfrac{1}{n}\displaystyle\sum_{i=1}^n x_iy_i\right)^2}{\dfrac{1}{n}\displaystyle\sum_{n=1}^n x_i^2}=\frac{\left(\dfrac{\sqrt{n}}{n}\displaystyle\sum_{i=1}^n x_iy_i\right)^2}{\displaystyle\sum_{n=1}^n x_i^2}=\frac{\left(\displaystyle\sum_{i=1}^n \left(\frac{y_i}{\sqrt{n}}\right)x_i\right)^2}{\displaystyle\sum_{n=1}^n x_i^2}=\frac{\left(\displaystyle\sum_{i=1}^n a_ix_i\right)^2}{\displaystyle\sum_{n=1}^n x_i^2} \tag{5.7.2282}$$

This is the Ryan-Joiner approach (implemented in Minitab for example) of the Shapiro-Wilk test. The results of both tests are very similar. The coefficients a_i can be easily obtained using any spreadsheet software by using a Monte Carlo simulation (see section Numerical Methods page 6708). If our reader wish it we will detail how to get the coefficients a_i with Microsoft® Excel® for a given n.

✍**Example:**

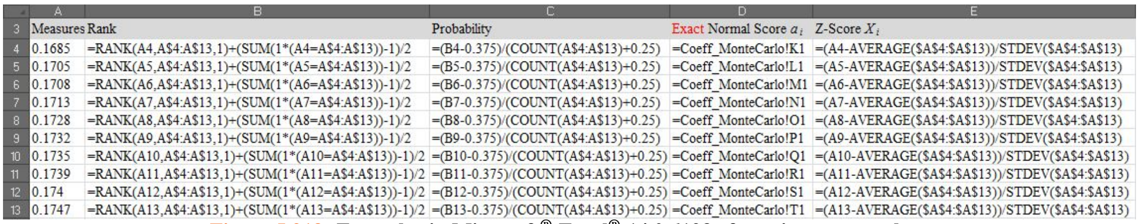
Consider the 10 measures of column `A` already sorted in ascending order:



	A	B	C	D	E
3	Measures	Rank	Probability	Exact Normal Score a_i	Z-Score X_i
4	0.1685	1.00	0.0610	-1.5480	-1.947246
5	0.1705	2.00	0.1585	-1.0055	-0.927745
6	0.1708	3.00	0.2561	-0.6551	-0.77482
7	0.1713	4.00	0.3537	-0.3726	-0.519945
8	0.1728	5.00	0.4512	-0.1205	0.2446801
9	0.1732	6.00	0.5488	0.1230	0.4485802
10	0.1735	7.00	0.6463	0.3798	0.6015052
11	0.1739	8.00	0.7439	0.6585	0.8054053
12	0.1740	9.00	0.8415	1.0056	0.8563803
13	0.1747	10.00	0.9390	1.5429	1.2132054

Figure 5.212: **Example of ordered measures, ranks, RJ coefficient and Z-score**

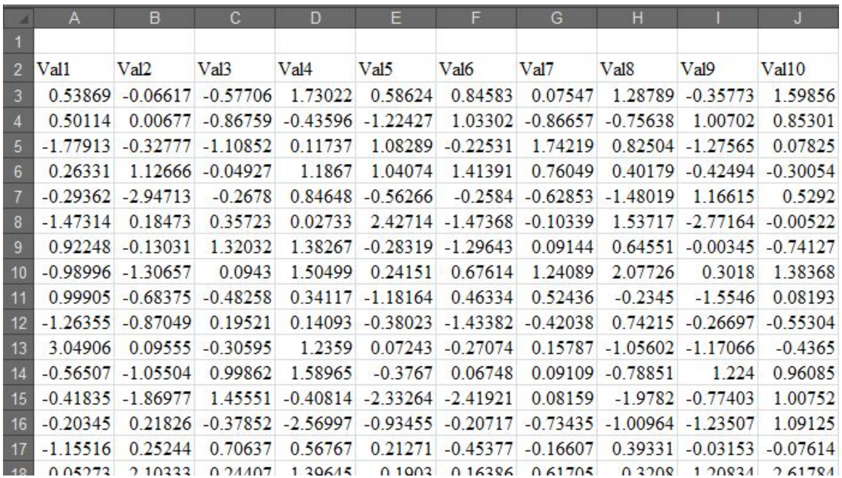
The corresponding formulas are:



	A	B	C	D	E
3	Measures	Rank	Probability	Exact Normal Score a_i	Z-Score X_i
4	0.1685	=RANK(A4,A$4:A$13,1)+(SUM(1*(A4=A$4:A$13))-1)/2	=(B4-0.375)/(COUNT(A$4:A$13)+0.25)	=Coeff_MonteCarlo!K1	=(A4-AVERAGE($A$4:$A$13))/STDEV($A$4:$A$13)
5	0.1705	=RANK(A5,A$4:A$13,1)+(SUM(1*(A5=A$4:A$13))-1)/2	=(B5-0.375)/(COUNT(A$4:A$13)+0.25)	=Coeff_MonteCarlo!L1	=(A5-AVERAGE($A$4:$A$13))/STDEV($A$4:$A$13)
6	0.1708	=RANK(A6,A$4:A$13,1)+(SUM(1*(A6=A$4:A$13))-1)/2	=(B6-0.375)/(COUNT(A$4:A$13)+0.25)	=Coeff_MonteCarlo!M1	=(A6-AVERAGE($A$4:$A$13))/STDEV($A$4:$A$13)
7	0.1713	=RANK(A7,A$4:A$13,1)+(SUM(1*(A7=A$4:A$13))-1)/2	=(B7-0.375)/(COUNT(A$4:A$13)+0.25)	=Coeff_MonteCarlo!N1	=(A7-AVERAGE($A$4:$A$13))/STDEV($A$4:$A$13)
8	0.1728	=RANK(A8,A$4:A$13,1)+(SUM(1*(A8=A$4:A$13))-1)/2	=(B8-0.375)/(COUNT(A$4:A$13)+0.25)	=Coeff_MonteCarlo!O1	=(A8-AVERAGE($A$4:$A$13))/STDEV($A$4:$A$13)
9	0.1732	=RANK(A9,A$4:A$13,1)+(SUM(1*(A9=A$4:A$13))-1)/2	=(B9-0.375)/(COUNT(A$4:A$13)+0.25)	=Coeff_MonteCarlo!P1	=(A9-AVERAGE($A$4:$A$13))/STDEV($A$4:$A$13)
10	0.1735	=RANK(A10,A$4:A$13,1)+(SUM(1*(A10=A$4:A$13))-1)/2	=(B10-0.375)/(COUNT(A$4:A$13)+0.25)	=Coeff_MonteCarlo!Q1	=(A10-AVERAGE($A$4:$A$13))/STDEV($A$4:$A$13)
11	0.1739	=RANK(A11,A$4:A$13,1)+(SUM(1*(A11=A$4:A$13))-1)/2	=(B11-0.375)/(COUNT(A$4:A$13)+0.25)	=Coeff_MonteCarlo!R1	=(A11-AVERAGE($A$4:$A$13))/STDEV($A$4:$A$13)
12	0.174	=RANK(A12,A$4:A$13,1)+(SUM(1*(A12=A$4:A$13))-1)/2	=(B12-0.375)/(COUNT(A$4:A$13)+0.25)	=Coeff_MonteCarlo!S1	=(A12-AVERAGE($A$4:$A$13))/STDEV($A$4:$A$13)
13	0.1747	=RANK(A13,A$4:A$13,1)+(SUM(1*(A13=A$4:A$13))-1)/2	=(B13-0.375)/(COUNT(A$4:A$13)+0.25)	=Coeff_MonteCarlo!T1	=(A13-AVERAGE($A$4:$A$13))/STDEV($A$4:$A$13)

Figure 5.213: **Formulas in Microsoft® Excel® 14.0.6123 of previous screen shot**

and therefore we have in a sheet named *CoeffMonteCarlo* Monte Carlo simulations to determine the 10 coefficients a_i traditionally denoted in the case of 10 measurements in the tables as following: $\left\{a_{1/10}, a_{2/10}, \ldots, a_{10/10}\right\}$. First we must generate 10 columns with Normal centered reduced variables on almost $10,000$ rows with the following Microsoft® Excel® formula:



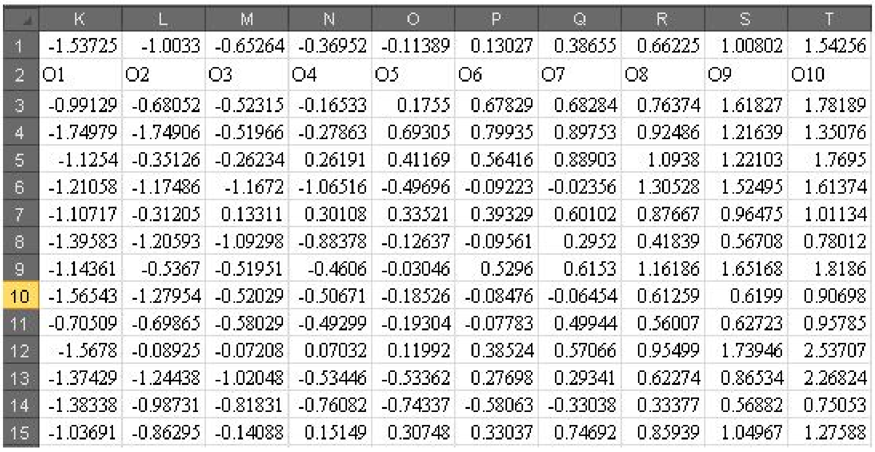
	A	B	C	D	E	F	G	H	I	J
1										
2	Val1	Val2	Val3	Val4	Val5	Val6	Val7	Val8	Val9	Val10
3	0.53869	-0.06617	-0.57706	1.73022	0.58624	0.84583	0.07547	1.28789	-0.35773	1.59856
4	0.50114	0.00677	-0.86759	-0.43596	-1.22427	1.03302	-0.86657	-0.75638	1.00702	0.85301
5	-1.77913	-0.32777	-1.10852	0.11737	1.08289	-0.22531	1.74219	0.82504	-1.27565	0.07825
6	0.26331	1.12666	-0.04927	1.1867	1.04074	1.41391	0.76049	0.40179	-0.42494	-0.30054
7	-0.29362	-2.94713	-0.2678	0.84648	-0.56266	-0.2584	-0.62853	-1.48019	1.16615	0.5292
8	-1.47314	0.18473	0.35723	0.02733	2.42714	-1.47368	-0.10339	1.53717	-2.77164	-0.00522
9	0.92248	-0.13031	1.32032	1.38267	-0.28319	-1.29643	0.09144	0.64551	-0.00345	-0.74127
10	-0.98996	-1.30657	0.0943	1.50499	0.24151	0.67614	1.24089	2.07726	0.3018	1.38368
11	0.99905	-0.68375	-0.48258	0.34117	-1.18164	0.46334	0.52436	-0.2345	-1.5546	0.08193
12	-1.26355	-0.87049	0.19521	0.14093	-0.38023	-1.43382	-0.42038	0.74215	-0.26697	-0.55304
13	3.04906	0.09555	-0.30595	1.2359	0.07243	-0.27074	0.15787	-1.05602	-1.17066	-0.4365
14	-0.56507	-1.05504	0.99862	1.58965	-0.3767	0.06748	0.09109	-0.78851	1.224	0.96085
15	-0.41835	-1.86977	1.45551	-0.40814	-2.33264	-2.41921	0.08159	-1.9782	-0.77403	1.00752
16	-0.20345	0.21826	-0.37852	-2.56997	-0.93455	-0.20717	-0.73435	-1.00964	-1.23507	1.09125
17	-1.15516	0.25244	0.70637	0.56767	0.21271	-0.45377	-0.16607	0.39331	-0.03153	-0.07614

Figure 5.214: Centered reduced normal variables for RJ coefficients

and then we have to build the ranks of all these values row by row such as:

```
=NORM.S.INV(RANDBETWEEN(1;99999999)/100000000)
```

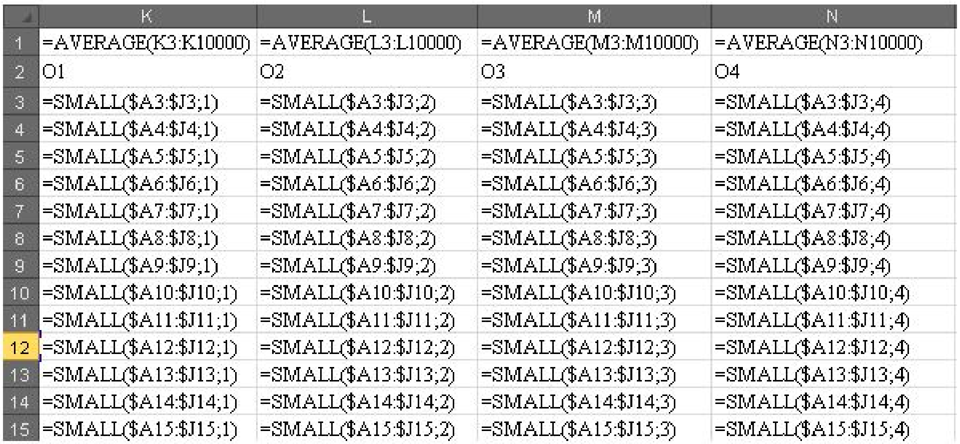
and then we have to build the ranks (ascending order) of all these values row by row such as:



	K	L	M	N	O	P	Q	R	S	T
1	-1.53725	-1.0033	-0.65264	-0.36952	-0.11389	0.13027	0.38655	0.66225	1.00802	1.54256
2	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
3	-0.99129	-0.68052	-0.52315	-0.16533	0.1755	0.67829	0.68284	0.76374	1.61827	1.78189
4	-1.74979	-1.74906	-0.51966	-0.27863	0.69305	0.79935	0.89753	0.92486	1.21639	1.35076
5	-1.1254	-0.35126	-0.26234	0.26191	0.41169	0.56416	0.88903	1.0938	1.22103	1.7695
6	-1.21058	-1.17486	-1.1672	-1.06516	-0.49696	-0.09223	-0.02356	1.30528	1.52495	1.61374
7	-1.10717	-0.31205	0.13311	0.30108	0.33521	0.39329	0.60102	0.87667	0.96475	1.01134
8	-1.39583	-1.20593	-1.09298	-0.88378	-0.12637	-0.09561	0.2952	0.41839	0.56708	0.78012
9	-1.14361	-0.5367	-0.51951	-0.4606	-0.03046	0.5296	0.6153	1.16186	1.65168	1.8186
10	-1.56543	-1.27954	-0.52029	-0.50671	-0.18526	-0.08476	-0.06454	0.61259	0.6199	0.90698
11	-0.70509	-0.69865	-0.58029	-0.49299	-0.19304	-0.07783	0.49944	0.56007	0.62723	0.95785
12	-1.5678	-0.08925	-0.07208	0.07032	0.11992	0.38524	0.57066	0.95499	1.73946	2.53707
13	-1.37429	-1.24438	-1.02048	-0.53446	-0.53362	0.27698	0.29341	0.62274	0.86534	2.26824
14	-1.38338	-0.98731	-0.81831	-0.76082	-0.74337	-0.58063	-0.33038	0.33377	0.56882	0.75053
15	-1.03691	-0.86295	-0.14088	0.15149	0.30748	0.33037	0.74692	0.85939	1.04967	1.27588

Figure 5.215: Ascending row by row sorting of simulations to determine the RJ coefficients

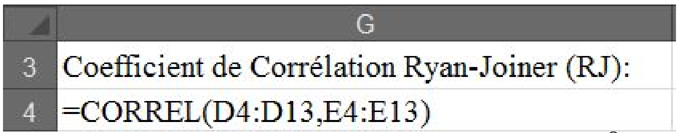
with the following formulas (given only for the first four i because of lack of space on the screenshot):



	K	L	M	N
1	=AVERAGE(K3:K10000)	=AVERAGE(L3:L10000)	=AVERAGE(M3:M10000)	=AVERAGE(N3:N10000)
2	O1	O2	O3	O4
3	=SMALL($A3:$J3;1)	=SMALL($A3:$J3;2)	=SMALL($A3:$J3;3)	=SMALL($A3:$J3;4)
4	=SMALL($A4:$J4;1)	=SMALL($A4:$J4;2)	=SMALL($A4:$J4;3)	=SMALL($A4:$J4;4)
5	=SMALL($A5:$J5;1)	=SMALL($A5:$J5;2)	=SMALL($A5:$J5;3)	=SMALL($A5:$J5;4)
6	=SMALL($A6:$J6;1)	=SMALL($A6:$J6;2)	=SMALL($A6:$J6;3)	=SMALL($A6:$J6;4)
7	=SMALL($A7:$J7;1)	=SMALL($A7:$J7;2)	=SMALL($A7:$J7;3)	=SMALL($A7:$J7;4)
8	=SMALL($A8:$J8;1)	=SMALL($A8:$J8;2)	=SMALL($A8:$J8;3)	=SMALL($A8:$J8;4)
9	=SMALL($A9:$J9;1)	=SMALL($A9:$J9;2)	=SMALL($A9:$J9;3)	=SMALL($A9:$J9;4)
10	=SMALL($A10:$J10;1)	=SMALL($A10:$J10;2)	=SMALL($A10:$J10;3)	=SMALL($A10:$J10;4)
11	=SMALL($A11:$J11;1)	=SMALL($A11:$J11;2)	=SMALL($A11:$J11;3)	=SMALL($A11:$J11;4)
12	=SMALL($A12:$J12;1)	=SMALL($A12:$J12;2)	=SMALL($A12:$J12;3)	=SMALL($A12:$J12;4)
13	=SMALL($A13:$J13;1)	=SMALL($A13:$J13;2)	=SMALL($A13:$J13;3)	=SMALL($A13:$J13;4)
14	=SMALL($A14:$J14;1)	=SMALL($A14:$J14;2)	=SMALL($A14:$J14;3)	=SMALL($A14:$J14;4)
15	=SMALL($A15:$J15;1)	=SMALL($A15:$J15;2)	=SMALL($A15:$J15;3)	=SMALL($A15:$J15;4)

Figure 5.216: **Microsoft® Excel® 14.0.6123 formulas of previous screenshot**

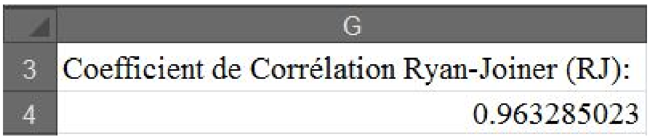
Finally, it only remains to calculate the correlation coefficient between the columns `D` and `E` of the first screenshot:



	G
3	Coefficient de Corrélation Ryan-Joiner (RJ):
4	=CORREL(D4:D13,E4:E13)

Figure 5.217: **Final calculation of the RJ correlation coefficient with Microsoft® Excel® 14.0.6123**

That gives:



	G
3	Coefficient de Corrélation Ryan-Joiner (RJ):
4	0.963285023

where the square of this value is very very close to the Shapiro-Wilk test.

Then, to know if we can reject or not the null hypothesis H_0 of Normality assumption, we should repeat the same procedure with in place of the measurements use random values generated from a Normal distribution and then determine the critical value of acceptance/rejection (it is very simple to make but we can detail on readers request).

After calculations in this special case we sadly do not reject the null hypothesis.

5.7.13.9.5 Anderson-Darling GoF Test

It is surprising that a test reasonably strong (robust) as the Kolmogorov-Smirnov test can be designed based on only a single observation and a single point of the function of candidate distribution!!! It would seem, with more hindsight, more efficient to measure the difference between the two distribution functions by comparing these functions on their entire domain, that is to say from $-\infty$ to $+\infty$.

It exist a family of tests which the statistics are based on the integral of the square of the difference (these tests are often considered as nonparametric but in my opinion wrongly and same for the Kolmogorov-Smirnov GoF test that is also considered as nonparametric):

$$\left(\hat{F}_n(x)-F_n(x)\right)^2 \tag{5.7.2283}$$

between the empirical distribution function and the reference distribution function. The simplest of these statistics is:

$$S=\int\limits_{-\infty}^{+\infty}\left(\hat{F}_n(x)-F_n(x)\right)^2\mathrm{d}x \tag{5.7.2284}$$

which is simply the area between the empirical distribution function and the reference distribution function. Either, by taking the above graph used during our study of the Kolmogorov-Smirnov GoF:

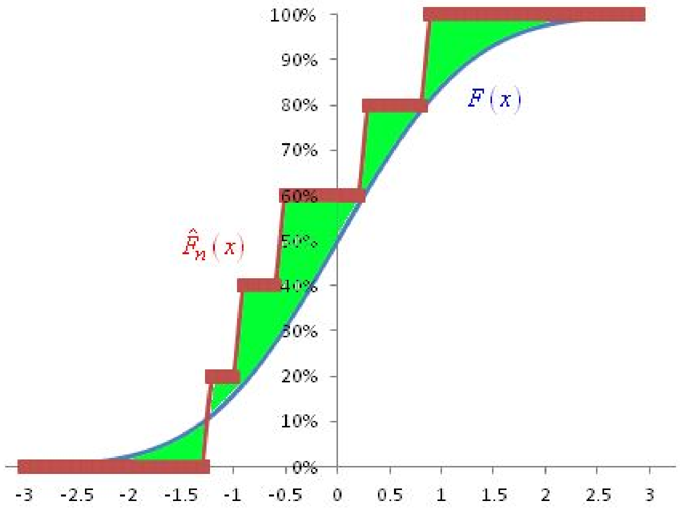


Figure 5.218: Idea behind the Anderson-Darling GoF

However, arbitrarily, we can choose something other than the measurement x for the integral. Thus, a conventional choice is to take the theoretical distribution function itself as a basic measure of the integral. It comes as follows:

$$W^2=n\int\limits_{-\infty}^{+\infty}\left(\hat{F}_n(x)-F_n(x)\right)^2\mathrm{d}F(x) \tag{5.7.2285}$$

Statistic resulting therefrom is named "Cramér-von Mises statistic". However it suffers from a major flaw of robustness when measuring points are on the tails of the distribution.

It then was proposed the following measure that is somewhat less sensitive to measurement points on the tails:

$$\frac{1}{n}W^2=\int\limits_{-\infty}^{+\infty}\frac{\left(\hat{F}_n(x)-F_n(x)\right)^2}{F(x)(1-F(x))}\mathrm{d}F(x)\tag{5.7.2286}$$

named "Anderson-Darling statistic" which was the most used in the late 120th century (holocene calendar) and remains dominant in the early 121st century also (at least while the sample is a fair size!). It is more robust by construction that the Cramér-von Mises statistics , and Kolmogorov-Smirnov simulations, but numerical studies have shown that it is less robust than the Shapiro-Wilk test or Ryan-Joiner test.

Remark

We can put the three statistics next to each other for comparisons purposes (with their traditional notation):

$$D_n(x)=\sup\left|\hat{F}_n(x)-F(x)\right| \qquad W^2=n\int\limits_{-\infty}^{+\infty}\left(\hat{F}_n(x)-F_n(x)\right)^2\mathrm{d}F(x) \qquad A^2=\int\limits_{-\infty}^{+\infty}\frac{\left(\hat{F}_n(x)-F_n(x)\right)^2}{F(x)(1-F(x))}\mathrm{d}F(x)\tag{5.7.2287}$$

Remembering that the definition of the empirical distribution $\hat{F}_n(x)$ in our study of the Kolmogorov-Smirnov GoF test implies that:

$$\hat{F}_n(x)=0\tag{5.7.2288}$$

if $x\in[-\infty,x_1]$ and:

$$\hat{F}_n(x)=\frac{n}{k}\tag{5.7.2289}$$

if $x\in[x_k,x_{k+1}]$ and:

$$\hat{F}_n(x)=1\tag{5.7.2290}$$

if $x\in[x_n,+\infty]$. We then have assuming in addition that F is continuous:

$$\begin{aligned}\frac{1}{n}W^2&=n\int\limits_{-\infty}^{+\infty}\frac{\left(\hat{F}_n(x)-F_n(x)\right)^2}{F(x)(1-F(x))}\mathrm{d}F(x)\\&=\int\limits_{-\infty}^{x_1}\frac{F(x)^2}{F(x)(1-F(x))}\mathrm{d}F(x)+\int\limits_{x_1}^{x_2}\frac{\left(\dfrac{1}{n}-F(x)\right)^2}{F(x)(1-F(x))}\mathrm{d}F(x)+\int\limits_{x_2}^{x_3}\frac{\left(\dfrac{2}{n}-F(x)\right)^2}{F(x)(1-F(x))}\mathrm{d}F(x)+\\&\quad\ldots+\int\limits_{x_n}^{+\infty}\frac{(1-F(x))^2}{F(x)(1-F(x))}\mathrm{d}F(x)\end{aligned}\tag{5.7.2291}$$

Then we make the change of variable $u=F(x)$ and therefore:

$$\mathrm{d}u=\mathrm{d}F(x) \tag{5.7.2292}$$

and without forgetting that the integral bound change as:

$$F(-\infty)=0 \quad \text{and} \quad F(+\infty)=1 \tag{5.7.2293}$$

Therefore we get:

$$\begin{aligned}\frac{1}{n}W^2=&\int\limits_0^{u_1}\frac{u^2}{u(1-u)}\mathrm{d}u+\int\limits_{u_1}^{u_2}\frac{\left(\dfrac{1}{n}-u\right)^2}{u(1-u)}\mathrm{d}u+\int\limits_{u_2}^{u_3}\frac{\left(\dfrac{2}{n}-u\right)^2}{u(1-u)}\mathrm{d}u+\\&\ldots+\int\limits_u^1\frac{(1-u)^2}{u(1-u)}\mathrm{d}u\\=&\int\limits_0^{u_1}\frac{}{u(1-u)}\mathrm{d}u+\int\limits_{u_1}^{u_2}\frac{\left(\dfrac{1}{n}-u\right)^2}{u(1-u)}\mathrm{d}u+\int\limits_{u_2}^{u_3}\frac{\left(\dfrac{2}{n}-u\right)^2}{u(1-u)}\mathrm{d}u+\ldots+\int\limits_u^1\frac{1-u}{u}\mathrm{d}u\end{aligned} \tag{5.7.2294}$$

where we of course have put:

$$u_j=F(x_j) \tag{5.7.2295}$$

We must now calculate these integrals. So we look for the primitive of a function of the type:

$$\frac{(a-u)^2}{u(1-u)}=\frac{a^2-2au+u^2}{u(1-u)}=\frac{a^2}{u(1-u)}-\frac{2a}{1-u}+\frac{u}{1-u} \tag{5.7.2296}$$

The primitive of the following terms:

$$\frac{1}{u(1-u)} \qquad \frac{1}{1-u} \tag{5.7.2297}$$

have been proved in their general form in the section of Differential and Integral Calculus and primitives are respectively:

$$\begin{aligned}&a^2\int\frac{1}{u(1-u)}\mathrm{d}u=a^2\ln\left|\frac{u}{1-u}\right|=a^2\ln|u|-a^2\ln|1-u|=a^2\ln(u)-a^2\ln(1-u)\\&2a\int\frac{1}{1-u}\mathrm{d}u=-2a\ln(1-u)\end{aligned} \tag{5.7.2298}$$

because in view of the possible values of u, then it is unnecessary to indicate the absolute values.

So we have only to calculate the primitive:

$$\frac{u}{1-u} \tag{5.7.2299}$$

where an obvious change of variables (if you want details feel free to ask us) gives us the primitive without the constant:

$$\int \frac{u}{1-u}\mathrm{d}u = -u - \ln(1-u) \tag{5.7.2300}$$

Then we have finally:

$$\begin{aligned}\int \frac{(a-u)^2}{u(1-u)^2}\mathrm{d}u &= \frac{a^2}{u(1-u)} - \frac{2a}{1-u} + \frac{u}{1-u}\\ &= a^2\ln(u) - a^2\ln(1-u) - (-2a\ln(1-u)) - u - \ln(1-u)\\ &= -u + a^2\ln(u) - (a-1)^2\ln(1-u)\end{aligned} \tag{5.7.2301}$$

Therefore we have:

$$\begin{aligned}&\frac{1}{n}W_n^2 = \int\limits_0^{u_1}\frac{u}{1-u}\mathrm{d}u + \sum_{j=1}^{n-1}\left(-u + \left(\frac{j}{n}\right)^2\ln(u) - \left(\frac{j}{n}-1\right)^2\ln(1-u)\right)\Bigg|_{u_j}^{u_{j+1}} + \int\limits_{u_n}^{1}\frac{1-u}{u}\mathrm{d}u\\ &= -u_1 - \ln(1-u_1) + \sum_{j=1}^{n-1}\left(-u + \left(\frac{j}{n}\right)^2\ln(u) - \left(\frac{j}{n}-1\right)^2\ln(1-u)\right)\Bigg|_{u_j}^{u_{j+1}} - 1 + u_n - \ln(u_n)\\ &= -u_1 - \ln(1-u_1) - 1 + u_n - \ln(u_n)\\ &+ \sum_{j=1}^{n-1}\left(-u_{j+1} + u_j + \left(\frac{j}{n}\right)^2\ln(u_{j+1}) - \left(\frac{j}{n}\right)^2\ln(u_j) - \left(\frac{j}{n}-1\right)^2\ln(1-u_{j+1}) + \left(\frac{j}{n}-1\right)^2\ln(1-u_j)\right)\end{aligned} \tag{5.7.2302}$$

We can already notice that in the last equality:

$$-u_1 + u_n + \sum_{j=1}^{n-1}(-u_{j+1} + u_j) = 0 \tag{5.7.2303}$$

Therefore it remains:

$$\begin{aligned}\frac{1}{n}W_n^2 &= -\ln(1-u_1) - 1 - \ln(u_n)\\ &+ \sum_{j=1}^{n-1}\left(\left(\frac{j}{n}\right)^2\ln(u_{j+1}) - \left(\frac{j}{n}\right)^2\ln(u_j) - \left(\frac{j}{n}-1\right)^2\ln(1-u_{j+1}) + \left(\frac{j}{n}-1\right)^2\ln(1-u_j)\right)\\ &= -\ln(1-u_1) - 1 - \ln(u_n) + \frac{1}{n^2}\sum_{j=1}^{n-1}j^2\ln(u_{j+1}) - \frac{1}{n^2}\sum_{j=1}^{n-1}j^2\ln(u_j)\\ &- \frac{1}{n^2}\sum_{j=1}^{n-1}(j-n)^2\ln(1-u_{j+1})\\ &+ \frac{1}{n^2}\sum_{j=1}^{n-1}(j-n)^2\ln(1-u_j)\end{aligned} \tag{5.7.2304}$$

We now make some tricky algebraic manipulations (but simple) to condense the writing of the latter equality.

First, notice that we can write the first sum as (the reader can check the two sums are equal by developing for a small value of n to make an example):

$$\sum_{j=1}^{n-1} j^2\ln(u_{j+1})=\sum_{j=1}^{n}(j-1)^2\left(\ln(u_j)\right) \tag{5.7.2305}$$

which is equivalent to put $j=j-1$.

We transform also the second sum:

$$\sum_{j=1}^{n-1} j^2\ln(u_j)=\left(\sum_{j=1}^{n} j^2\ln(u_j)\right)-n^2\ln(u_n) \tag{5.7.2306}$$

and the reader can check that the equality below for the third sum is verified (if you problems for this don't hesitate to contact us):

$$\sum_{j=1}^{n-1}(j-n)^2\ln(1-u_{j+1})=-n^2\ln(1-u_1)-\left(\sum_{j=1}^{n}(j-1-n)^2\ln(1-u_j)\right) \tag{5.7.2307}$$

which also is equivalent to put $j=j-1$.

Finally, we transform the fourth sum (since in any case when j is equal to n then the term of the sum is zero ...):

$$\sum_{j=1}^{n-1}(j-n)^2\ln(1-u_j)=\sum_{j=1}^{n}(j-n)^2\ln(1-u_j) \tag{5.7.2308}$$

Therefore we have:

$$\begin{aligned}\frac{1}{n}W_n^2=&-\ln(1-u_1)-1-\ln(u_n)+\frac{1}{n^2}\sum_{j=1}^{n}(j-1)^2\left(\ln(u_j)\right)-\frac{1}{n^2}\left[\left(\sum_{j=1}^{n} j^2\ln(u_j)\right)-n^2\ln(u_n)\right]\\&-\frac{1}{n^2}\left[-n^2\ln(1-u_1)-\left(\sum_{j=1}^{n}(j-1-n)^2\ln(1-u_j)\right)\right]+\frac{1}{n^2}\sum_{j=1}^{n}(j-n)^2\ln(1-u_j)\end{aligned} \tag{5.7.2309}$$

Either by eliminating the terms that vanish:

$$\begin{aligned}\frac{1}{n}W_n^2=&-1+\frac{1}{n^2}\sum_{j=1}^{n}(j-1)^2\left(\ln(u_j)\right)-\frac{1}{n^2}\sum_{j=1}^{n} j^2\ln(u_j)+\frac{1}{n^2}\sum_{j=1}^{n}(j-1-n)^2\ln(1-u_j)\\&+\frac{1}{n^2}(j-n)^2\ln(1-u_j)\end{aligned} \tag{5.7.2310}$$

And by grouping the terms with the same form of logarithm:

$$\begin{aligned}\frac{1}{n}W_n^2&=-1-\frac{1}{n^2}\sum_{j=1}^{n}(2j-1)\ln(u_j)+\frac{1}{n^2}\sum_{j=1}^{n}(2j-2n-1)\ln(1-u_j)\\&=-1-\frac{1}{n^2}\sum_{j=1}^{n}\left[(2j-1)\ln(u_j)+(2(j-n)-1)\ln(1-u_j)\right]\end{aligned} \tag{5.7.2311}$$

Therefore:

$$W_n^2 = -n - \frac{1}{n}\sum_{j=1}^{n}\left[(2j-1)\ln(u_j) + (2(j-n)-1)\ln(1-u_j)\right] \tag{5.7.2312}$$

This is one of the closed form of the Anderson-Darling GoF test (especially well known as normality test) and that in the context of a Normal distribution is expressed by tradition in the following form:

$$\boxed{A_n^2 = -n - \frac{1}{n}\sum_{j=1}^{n}\left[(2j-1)\ln(\Phi(Y_j)) + (2(j-n)-1)\ln(1-\Phi(Y_j))\right]} \tag{5.7.2313}$$

But there is another very common simplified expression. To establish it, we start again from the prior-previous expression:

$$W_n^2 = -n - \frac{1}{n}\sum_{j=1}^{n}\left[(2j-1)\ln(u_j) + (2(j-n)-1)\ln(1-u_j)\right] \tag{5.7.2314}$$

By doing the change of variable $j = 1+n-i$ in the last sum, the expression:

$$(2(j-n)-1)\ln(1-u_j) \tag{5.7.2315}$$

becomes:

$$(1-2i)\ln(1-u_{1+n-i}) \tag{5.7.2316}$$

and the bounds of the sum becomes:

$$\begin{aligned} i &= 1+n-j \underset{j=1}{\Rightarrow} i = n \\ i &= 1+n-j \Rightarrow i = 1 \end{aligned} \tag{5.7.2317}$$

Therefore:

$$\sum_{j=1}^{n} = \sum_{i=n}^{1} = \sum_{i=1}^{n} \tag{5.7.2318}$$

Hence:

$$\begin{aligned} \frac{1}{n}W_n^2 &= -1 - \frac{1}{n^2}\sum_{j=1}^{n}(2j-1)\ln(u_j) + \frac{1}{n^2}\sum_{i=1}^{n}(1-2i)\ln(u_{1+n-i}) \\ &= -1 - \frac{1}{n^2}\sum_{j=1}^{n}(2j-1)\ln(u_j) + \frac{1}{n^2}\sum_{j=1}^{n}(1-2j)\ln(u_{1+n-j}) \\ &= -1 - \frac{1}{n^2}\sum_{j=1}^{n}(2j-1)\left[\ln(u_j) + \ln(u_{1+n-j})\right] \end{aligned} \tag{5.7.2319}$$

Finally:

$$\boxed{W_n^2 = \text{AD} = -n - \frac{1}{n}\sum_{j=1}^{n}(2j-1)\left[\ln(u_j) + \ln(u_{1+n-j})\right] = -n - \frac{1}{n}\sum_{j=1}^{n}S_j} \tag{5.7.2320}$$

✍**Example:**

Suppose we have measured the following five values:

$$-1.2, 0.2, -0.6, 0.8, -1.0$$

thus ordered:

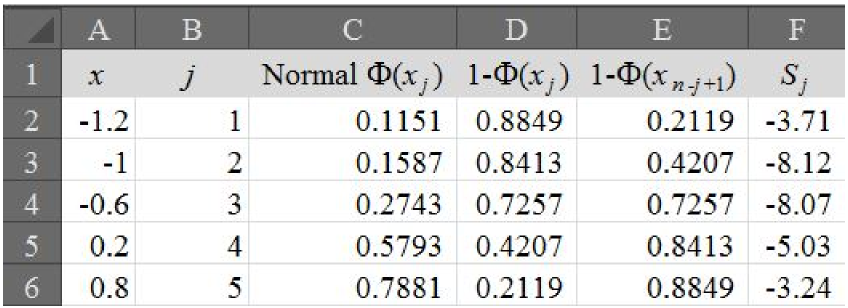
$$x_1 = -1.2, x_2 = -1.0, x_3 = -0.6, x_4 = 0.2, x_5 = 0.8$$

We want to test the following null hypothesis:

$$H_0 : \text{CDF}(x_i) = \Phi(x_i) \tag{5.7.2321}$$

where $\Phi(x)$ is for recall the distribution function of the Normal centered reduced distribution $\mathcal{N}(0,1)$.

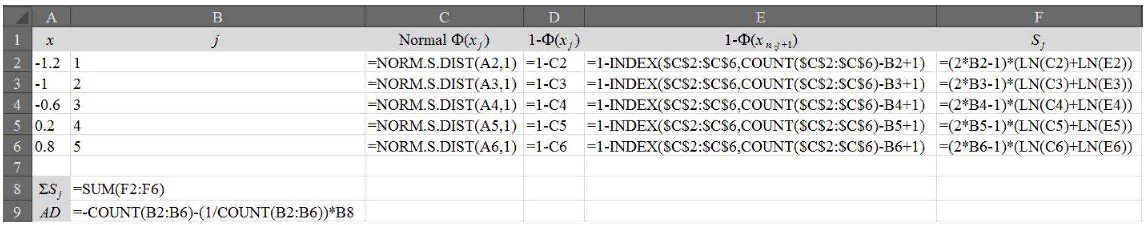
To implement the calculation of the AD in a spreadsheet software like Microsoft® Excel® 14.0.6123 we first do:



	A	B	C	D	E	F
1	x	j	Normal $\Phi(x_j)$	1-$\Phi(x_j)$	1-$\Phi(x_{n-j+1})$	S_j
2	-1.2	1	0.1151	0.8849	0.2119	-3.71
3	-1	2	0.1587	0.8413	0.4207	-8.12
4	-0.6	3	0.2743	0.7257	0.7257	-8.07
5	0.2	4	0.5793	0.4207	0.8413	-5.03
6	0.8	5	0.7881	0.2119	0.8849	-3.24

Figure 5.219: **Initial values for AD GoF test**

Explicitly:



	A	B	C	D	E	F
1	x	j	Normal $\Phi(x_j)$	1-$\Phi(x_j)$	1-$\Phi(x_{n-j+1})$	S_j
2	-1.2	1	=NORM.S.DIST(A2,1)	=1-C2	=1-INDEX($C$2:$C$6,COUNT($C$2:$C$6)-B2+1)	=(2*B2-1)*(LN(C2)+LN(E2))
3	-1	2	=NORM.S.DIST(A3,1)	=1-C3	=1-INDEX($C$2:$C$6,COUNT($C$2:$C$6)-B3+1)	=(2*B3-1)*(LN(C3)+LN(E3))
4	-0.6	3	=NORM.S.DIST(A4,1)	=1-C4	=1-INDEX($C$2:$C$6,COUNT($C$2:$C$6)-B4+1)	=(2*B4-1)*(LN(C4)+LN(E4))
5	0.2	4	=NORM.S.DIST(A5,1)	=1-C5	=1-INDEX($C$2:$C$6,COUNT($C$2:$C$6)-B5+1)	=(2*B5-1)*(LN(C5)+LN(E5))
6	0.8	5	=NORM.S.DIST(A6,1)	=1-C6	=1-INDEX($C$2:$C$6,COUNT($C$2:$C$6)-B6+1)	=(2*B6-1)*(LN(C6)+LN(E6))
7						
8	ΣS_j	=SUM(F2:F6)				
9	AD	=-COUNT(B2:B6)-(1/COUNT(B2:B6))*B8				

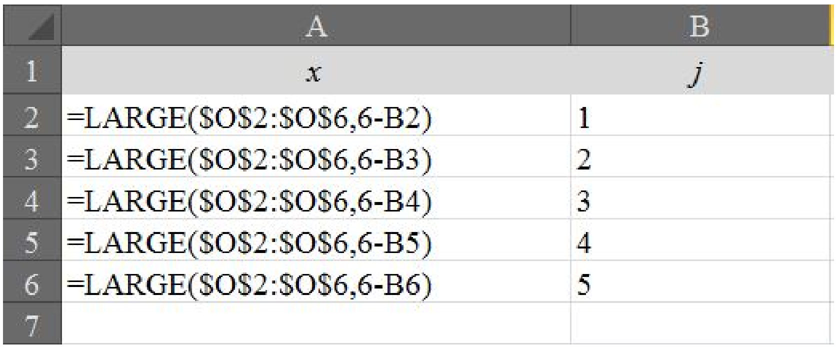
Figure 5.220: **Formulas in previous figure values for AD GoF test**

So we get the same value of the indicator AD that as statistical softwares that allows to choose the law to be compared (and therefore the parameters relating). However, for very small samples statistics software uses the following correction (which we were not able to re-obtain by simulation...):

$$\text{AD}^* = \text{AD}\left(1 + \frac{0.75}{n} + \frac{2.25}{n^2}\right) \tag{5.7.2322}$$

thus in our case AD* is about 0.789 as $\text{AD} = 0.636$.

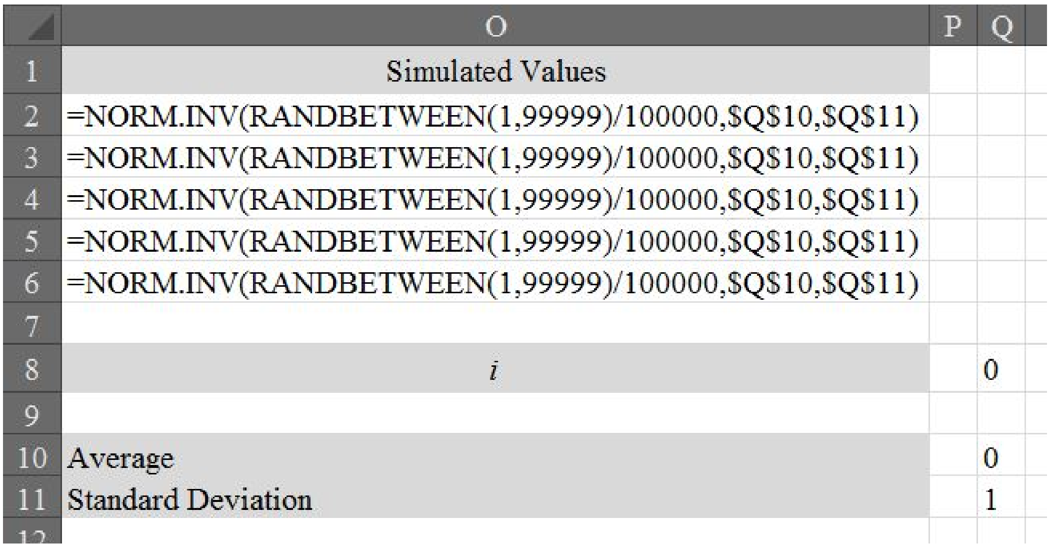
Then to calculate the p-value we need to investigate a curiosity... Indeed if we determine it by a Monte Carlo simulation as we did during our practical application of the Kolmogorov-Smirnov test by first writing in the column `A` from a new sheet the following formulas:



	A	B
1	*x*	*j*
2	=LARGE($O$2:$O$6,6-B2)	1
3	=LARGE($O$2:$O$6,6-B3)	2
4	=LARGE($O$2:$O$6,6-B4)	3
5	=LARGE($O$2:$O$6,6-B5)	4
6	=LARGE($O$2:$O$6,6-B6)	5
7		

Figure 5.221: **Formulas to sort Monte Carlo simulated values for AD GoF test**

values coming from the column `O` where we have:

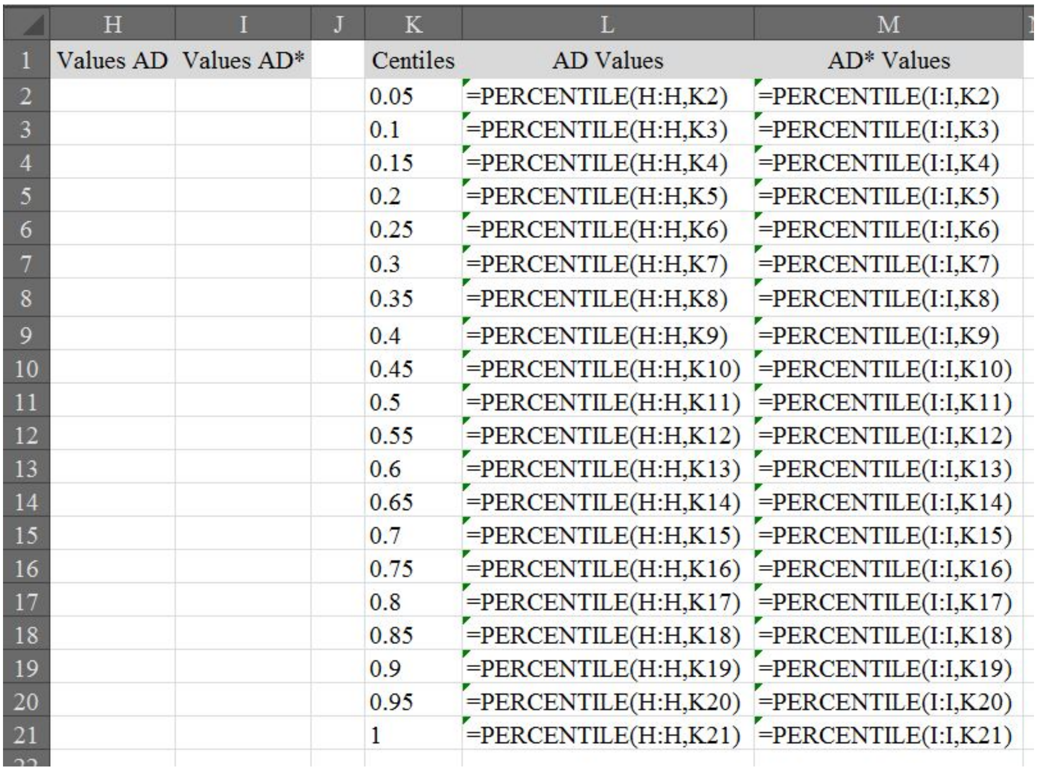


	O	P	Q
1	Simulated Values		
2	=NORM.INV(RANDBETWEEN(1,99999)/100000,$Q$10,$Q$11)		
3	=NORM.INV(RANDBETWEEN(1,99999)/100000,$Q$10,$Q$11)		
4	=NORM.INV(RANDBETWEEN(1,99999)/100000,$Q$10,$Q$11)		
5	=NORM.INV(RANDBETWEEN(1,99999)/100000,$Q$10,$Q$11)		
6	=NORM.INV(RANDBETWEEN(1,99999)/100000,$Q$10,$Q$11)		
7			
8	*i*		0
9			
10	Average		0
11	Standard Deviation		1

Figure 5.222: **Monte Carlo simulated formulas values for AD GoF test**

The reader will notice that this is same as comparing the sample with a uniform distribution!!!

Then having prepare the following columns `H`, `I` (see the figure below) that will contain the simulated values reported by the VBA code given a little further and columns `L`, `M` that allow us after to have the distribution of AD and AD values thanks to the calculation of the percentile:



	H	I	J	K	L	M
1	Values AD	Values AD*		Centiles	AD Values	AD* Values
2				0.05	=PERCENTILE(H:H,K2)	=PERCENTILE(I:I,K2)
3				0.1	=PERCENTILE(H:H,K3)	=PERCENTILE(I:I,K3)
4				0.15	=PERCENTILE(H:H,K4)	=PERCENTILE(I:I,K4)
5				0.2	=PERCENTILE(H:H,K5)	=PERCENTILE(I:I,K5)
6				0.25	=PERCENTILE(H:H,K6)	=PERCENTILE(I:I,K6)
7				0.3	=PERCENTILE(H:H,K7)	=PERCENTILE(I:I,K7)
8				0.35	=PERCENTILE(H:H,K8)	=PERCENTILE(I:I,K8)
9				0.4	=PERCENTILE(H:H,K9)	=PERCENTILE(I:I,K9)
10				0.45	=PERCENTILE(H:H,K10)	=PERCENTILE(I:I,K10)
11				0.5	=PERCENTILE(H:H,K11)	=PERCENTILE(I:I,K11)
12				0.55	=PERCENTILE(H:H,K12)	=PERCENTILE(I:I,K12)
13				0.6	=PERCENTILE(H:H,K13)	=PERCENTILE(I:I,K13)
14				0.65	=PERCENTILE(H:H,K14)	=PERCENTILE(I:I,K14)
15				0.7	=PERCENTILE(H:H,K15)	=PERCENTILE(I:I,K15)
16				0.75	=PERCENTILE(H:H,K16)	=PERCENTILE(I:I,K16)
17				0.8	=PERCENTILE(H:H,K17)	=PERCENTILE(I:I,K17)
18				0.85	=PERCENTILE(H:H,K18)	=PERCENTILE(I:I,K18)
19				0.9	=PERCENTILE(H:H,K19)	=PERCENTILE(I:I,K19)
20				0.95	=PERCENTILE(H:H,K20)	=PERCENTILE(I:I,K20)
21				1	=PERCENTILE(H:H,K21)	=PERCENTILE(I:I,K21)

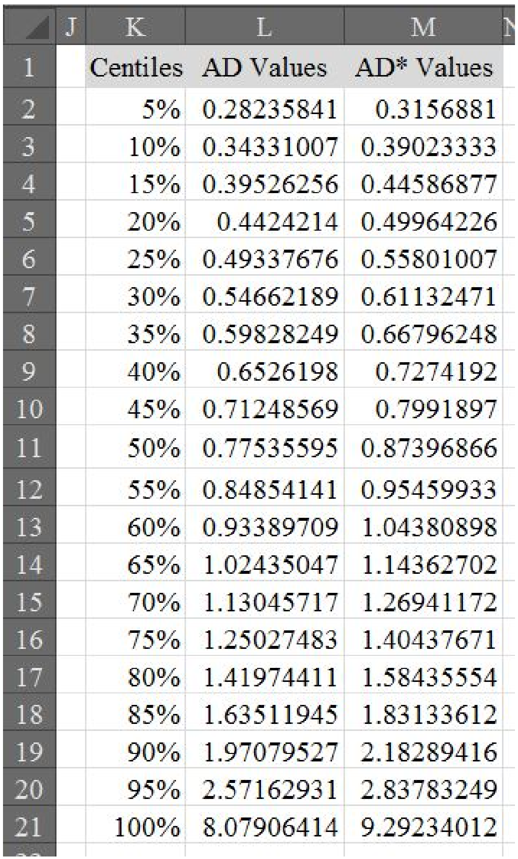
Figure 5.223: Columns for writing the AD/AD* values computed by VBA and the percentiles calculations

with the little VBA code below quickly and dirty made:

Listing 5.1: VBA code of the AD GoF test

```
Sub SimulAndersonDarling()

        Const intSimulations As Integer = 1000
        Dim i As Integer

        For i=1 To intSimulations
                Calculate
                Cells(i+1, 8).Value = Cells(9, 2).Value
                Cells(i+1, 9).Value = Cells(10, 2).Value
                Cells(8, 17).Value = i
        Next i
End Sub
```

we then have with $10,000$ simulations the following distribution of AD and AD* values:



	K	L	M
1	Centiles	AD Values	AD* Values
2	5%	0.28235841	0.3156881
3	10%	0.34331007	0.39023333
4	15%	0.39526256	0.44586877
5	20%	0.4424214	0.49964226
6	25%	0.49337676	0.55801007
7	30%	0.54662189	0.61132471
8	35%	0.59828249	0.66796248
9	40%	0.6526198	0.7274192
10	45%	0.71248569	0.7991897
11	50%	0.77535595	0.87396866
12	55%	0.84854141	0.95459933
13	60%	0.93389709	1.04380898
14	65%	1.02435047	1.14362702
15	70%	1.13045717	1.26941172
16	75%	1.25027483	1.40437671
17	80%	1.41974411	1.58435554
18	85%	1.63511945	1.83133612
19	90%	1.97079527	2.18289416
20	95%	2.57162931	2.83783249
21	100%	8.07906414	9.29234012

Figure 5.224: AD GoF test percentiles

So whether it is for AD or AD* the p-value is in our particular case almost equal to $1-40\%=60\%$ which corresponds to the tabulated values by Peter A.W. Lewis at IBM (11961 according to holocene calendar).

What is curious and that we need to investigate is that the majority of software use the following formulas (R.B. D'Agostino and M.A. Stephens, Eds., 11986 (holocene calendar), Goodness-of-Fit Technology, Marcel Dekker) to avoid Monte Carlo simulations:

$$\begin{aligned}
&\text{AD}^*\geq 0.6\Rightarrow p\text{-values}\cong e^{1.2937-5.709\text{AD}^*+0.0186\text{AD}^{*2}}\\
&0.34\leq\text{AD}^*<0.6\Rightarrow p\text{-values}\cong e^{0.9177-4.279\text{AD}^*-1.38\text{AD}^{*2}}\\
&0.2\leq\text{AD}^*<0.34\Rightarrow p\text{-values}\cong 1-e^{-8.318+42.796\text{AD}^*-59.938\text{AD}^{*2}}\\
&\text{AD}^*\leq 0.2\Rightarrow p\text{-values}\cong 1-e^{-13.436+101.14\text{AD}^*-223.73\text{AD}^{*2}}
\end{aligned}\tag{5.7.2323}$$

and in our case, the application of these formulas give a p-value of about 9% !!! Value actually given by statistical softwares! We will try to find where this difference comes from as our calculation seem subjectively more accurate than this empirical formulas... We asked the support of an American statistical package editor to explain the reason for the difference between the values tabulated by Peter A.W. Lewis and those of R.B D'Augostino and M.A. Stephens but they were not able to answer. We also contacted M.A. Stephens himself so that he can communicate us how he had obtained these formulas but we never get any answers ...

5.7.13.9.6 Cramér–von Mises GoF Test

In statistics the Cramér–von Mises GoF test (considered by many as a nonparametric test) is used for judging the goodness of fit of a cumulative distribution function F^* compared to a given empirical distribution function F_n, or for comparing two empirical distributions (the criterion is named after Harald Cramér and Richard Edler von Mises who first proposed it in 11928–11930 according to holocene calendar). It is also used as a part of other algorithms, such as minimum distance estimation. It is commonly defined as following (as we have already introduced it just earlier above):

$$W^2 = n \int_{-\infty}^{+\infty} (F_n(x) - F^*(x))^2 \, \mathrm{d}F^*(x) \tag{5.7.2324}$$

but should better be written as:

$$W^2 = \int \left(\hat{F}_n(t) - F_0(t)\right)^2 \mathrm{d}F_0(t) \tag{5.7.2325}$$

In one-sample applications F^* is the theoretical distribution and F_n is the empirically observed distribution. Alternatively the two distributions can both be empirically estimated ones; this is named the "two-sample case".

We want to derive the Cramér-von Mises test statistic (said to be with n-degrees of freedom):

$$\boxed{W^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left(U_i - \frac{2i-1}{2n}\right)^2} \tag{5.7.2326}$$

where $U_i = F_0(X_{i:n})$ is the theoretical order statistics from the original (that's why it's a parametric test! We need to estimate the parameters of the original distribution):

$$\frac{W^2}{n} := C_n = \int \left(\hat{F}_n(t) - F_0(t)\right)^2 \mathrm{d}F_0(t) \tag{5.7.2327}$$

Notice that this statistic is tabulated[84] against the null hypothesis H_0 that $\hat{F}_n(t) = F_0(t)$.

Let $X_1, \ldots, X_n$ be a random sample of size n from $f(x)$. The empirical CDF is:

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x) = \begin{cases} 0 & \text{if } x < X_{1:n} \\ i/n & \text{if } X_{i:n} \leq x < X_{i+1:n} \\ 1 & \text{if} X_{n:n} \leq x \end{cases} \tag{5.7.2328}$$

---

[84] Mainly and sadly only for the Normal distribution also in some of the most important statistical softwares.

Let $U_i=F_0(X_{i:n})$, then (the reader will see that there are some quite clever simplifications):

$$\begin{aligned}
C_n &\triangleq \int \left[\hat{F}_n(x)-F_0(x)\right]^2 \mathrm{d}F_0(x)\\
&=\int\limits_{-\infty}^{X_{1:n}}\left[-F_0(x)\right]^2 \mathrm{d}F_0(x)+\sum_{i=1}^{n-1}\int\limits_{X_{i:n}}^{X_{i+1:n}}\left[\frac{i}{n}-F_0(x)\right]^2 \mathrm{d}F_0(x)+\int\limits_{X_{n:n}}^{+\infty}\left[1-F_0(x)\right]^2 \mathrm{d}F_0(x)\\
&=\frac{1}{3}U_1^3+\frac{1}{3}\sum_{i=1}^{n-1}\left[\left(U_{i+1}-\frac{i}{n}\right)^3-\left(U_i-\frac{i}{n}\right)^3\right]-\frac{1}{3}(U_n-1)^3\\
&=\frac{1}{3}U_1^3+\frac{1}{3}\sum_{i=1}^{n-1}\left[U_{i+1}^3-U_i^3+\frac{3i^2}{n^2}\left(U_{i+1}-U_i\right)-\frac{3i}{n}\left(U_{i+1}^2-U_i^2\right)\right]-\frac{1}{3}(U_n-1)^3
\end{aligned}\tag{5.7.2329}$$

Now we use:

$$\begin{aligned}
\sum_{i=1}^{n-1}i^2(U_{i+1}-U_i)&=\sum_{i=2}^{n}(i-1)^2U_i-\sum_{i=1}^{n-1}i^2U_i=(n-1)^2U_n-U_1-\sum_{i=2}^{n-1}(2i-1)U_i\\
&=n^2U_n-\sum_{i=1}^{n}(2i-1)U_i
\end{aligned}\tag{5.7.2330}$$

Therefore:

$$\begin{aligned}
C_n &\triangleq \frac{1}{3}U_1^3+\frac{1}{3}U_n^3-\frac{1}{3}U_1^3+\left(U_n-\sum_{i=1}^{n}\frac{2i-1}{n^2}U_i\right)-\left(U_n^2-\sum_{i=1}^{n}\frac{1}{n}U_i^2\right)-\frac{1}{3}(U_n-1)^3\\
&=\frac{1}{3}+\frac{1}{n}\sum_{i=1}^{n}\left(U_i^2-\frac{2i-1}{n}U_i\right)=\frac{1}{3}+\frac{1}{n}\sum_{i=1}^{n}\left(U_i-\frac{2i-1}{2n}\right)^2-\frac{1}{n}\sum_{i=1}^{n}\left(\frac{2i-1}{2n}\right)^2\\
&=\frac{1}{3}+\frac{1}{n}\sum_{i=1}^{n}\left(U_i-\frac{2i-1}{2n}\right)^2-\frac{1}{4n^2}\sum_{i=1}^{n}(2i-1)^2\\
&=\frac{1}{3}+\frac{1}{n}\sum_{i=1}^{n}\left(U_i-\frac{2i-1}{2n}\right)^2-\frac{1}{4n^2}\frac{1}{3}n(4n^2-1)\\
&=\frac{1}{12n^2}+\frac{1}{n}\sum_{i=1}^{n}\left(U_i-\frac{2i-1}{2n}\right)^2
\end{aligned}\tag{5.7.2331}$$

Therefore:

$$W^2=nC_n=\frac{1}{12n}+\sum_{i=1}^{n}\left(U_i-\frac{2i-1}{2n}\right)^2\tag{5.7.2332}$$

✍Example:

Consider we have the following parameters from a sample:

$$\bar{X} = 6.4034 \qquad \text{and} \qquad \sigma_X = 3.299319$$

Assuming a Normal distribution $\mathcal{N}(\bar{X}, \sigma_X)$, we get for each of the values:

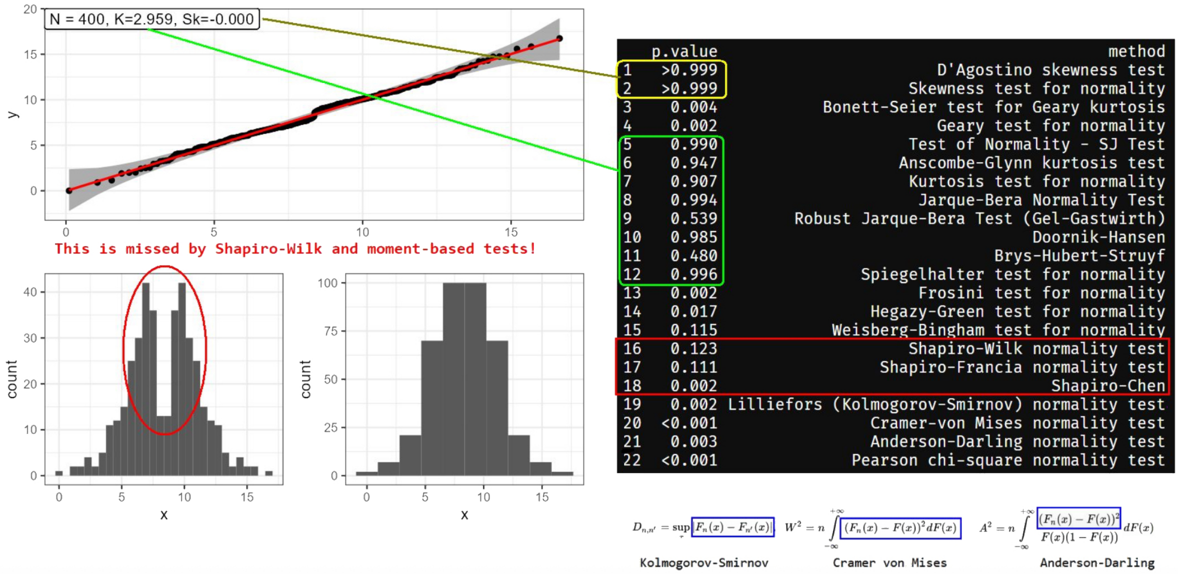
$$F_0(X) = (0.04943849, 0.05925460, 0.06477221, 0.07101753, 0.13526439, 0.25218337, 0.31746223, 0.38651837, 0.48830421, 0.67550752, 0.68425558, 0.69965168, 0.78797818, 0.79458401, 0.79661282, 0.83025028, 0.83051069, 0.83202204, 0.83609969, 0.84033359)$$

Let us calculate now the Cramér-von Mises statistic using:

$$\begin{aligned} W^2 &= \frac{1}{12 \cdot 20} + \sum_{i=1}^{20} \left(\frac{F(x_i) - (2 \cdot i - 1)}{20 \cdot 20} \right)^2 \\ &= \frac{1}{240} + \left(\frac{0.04943849 - (2-1)}{40} \right)^2 + \cdots \left(\frac{0.84033359 - (40-1)}{40} \right)^2 \\ &= \frac{1}{240} + (0.005972399 + 0.0002479177 + 0.003627386 + \cdots + 0.01813504) \\ &= 0.04166667 + 0.1861175 = 0.1902842 \end{aligned} \tag{5.7.2333}$$

We get here $W^2 = 0.1902842$ that we report in the Cramér-von Mises table for 20 degrees of freedom, which gives a p-value of $0.006 < 5\%$. Hence, we reject H_0 and we conclude that the distribution function of X is very likely not drawn from a Normal distribution.

There are almost half a hundred of GoF tests and because all of them have different assumptions they don't lead the same results as illustrated below with a bi-modal quasi-Gaussian distribution:



```
   p.value                                    method
1   >0.999                  D'Agostino skewness test
2   >0.999               Skewness test for normality
3    0.004        Bonett-Seier test for Geary kurtosis
4    0.002                  Geary test for normality
5    0.990               Test of Normality - SJ Test
6    0.947                Anscombe-Glynn kurtosis test
7    0.907               Kurtosis test for normality
8    0.994                Jarque-Bera Normality Test
9    0.539   Robust Jarque-Bera Test (Gel-Gastwirth)
10   0.985                            Doornik-Hansen
11   0.480                        Brys-Hubert-Struyf
12   0.996          Spiegelhalter test for normality
13   0.002                Frosini test for normality
14   0.017           Hegazy-Green test for normality
15   0.115        Weisberg-Bingham test for normality
16   0.123               Shapiro-Wilk normality test
17   0.111           Shapiro-Francia normality test
18   0.002                             Shapiro-Chen
19   0.002 Lilliefors (Kolmogorov-Smirnov) normality test
20  <0.001           Cramer-von Mises normality test
21   0.003           Anderson-Darling normality test
22  <0.001          Pearson chi-square normality test
```

$$D_{n,n'} = \sup_x \boxed{|F_n(x) - F_{n'}(x)|} \qquad W^2 = n \int_{-\infty}^{+\infty} \boxed{(F_n(x) - F(x))^2 dF(x)} \qquad A^2 = n \int_{-\infty}^{+\infty} \frac{\boxed{(F_n(x) - F(x))^2}}{F(x)(1 - F(x))} dF(x)$$

Kolmogorov-Smirnov Cramer von Mises Anderson-Darling

5.7.13.10 Kernel density estimation

In statistics, "kernel density estimation" (KDE) or "kernel smoothing" is a nonparametric way to estimate the probability density function of a random variable, but also to do some special regressions (see section Numerical Methods page 6447). Kernel density estimation is a fundamental data smoothing problem in business and science where inferences about the population are made, based on a finite data sample. In some fields such as signal processing and econometrics it is also termed the "Parzen–Rosenblatt window method", after Emanuel Parzen and Murray Rosenblatt, who are usually credited with independently creating it in its current form.

The nonparametric estimation of a probability density function is very important in all domains using statistics! We already know that the technique making use of histograms is very delicate and questionable because it is necessary to choose the right interval width for each of the bars, which is very subjective and the Sturges rule that we have seen earlier can be discussed!

To introduce the subject we first recall that we have in the univariate case:

$$F(x-h<X<x+h)=P(x-h<X<x+h)=\int\limits_{x-h}^{x+h} f(t)\mathrm{d}t \qquad (5.7.2334)$$

Or what is sometimes written more elegantly:

$$F(x\in I_j)=P(x\in I_j)=\int\limits_{I_j} f(t)\mathrm{d}t \qquad (5.7.2335)$$

Based on the method of numerical integration with rectangles (see section Numerical Methods page 6623) we can write:

$$P(x-h<X<x+h)=\int\limits_{x-h}^{x+h} f(t)\mathrm{d}t\cong 2hf(x) \qquad (5.7.2336)$$

Therefore, we have:

$$f(x)\cong\frac{1}{2h}P(x-h<X<x+h) \qquad (5.7.2337)$$

The above probability can be estimated by the relative frequency of observations:

$$\hat{f}(x)=\frac{1}{2h}\frac{\#\text{ observations in the interval }]x-h,x+h[}{n} \qquad (5.7.2338)$$

What is sometimes written:

$$\hat{f}(x)=\frac{1}{2h}\frac{\#\{x_i\in I_j\}}{n} \qquad (5.7.2339)$$

An alternative non-trivial writing (and equal in numerical value!) is:

$$\hat{f}(x)=\frac{1}{n}\sum_{i=1}^{n}w(x-x_i)=\frac{1}{n}\sum_{i=1}^{n}w(t,h) \qquad (5.7.2340)$$

where $x_1, x_2, \ldots, x_n$ are the observed values and:

$$w(x-x_i,h)=w(t,h)=\begin{cases}\dfrac{1}{2h} & |t|<h \\ 0 & \text{otherwise}\end{cases} \qquad (5.7.2341)$$

is named the "weight of the kernel" function or "univariate Rosenblatt kernel function". The later relation is also often denoted:

$$\hat{f}(x)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right) \qquad (5.7.2342)$$

With (notation a little unfortunate in my point of view...):

$$K(x)=K\left(\frac{x-x_i}{h}\right)=\begin{cases}\dfrac{1}{2} & |x-x_i|<h \\ 0 & \text{otherwise}\end{cases} \qquad (5.7.2343)$$

and where h is named the "bandwidth".

In general, a "kernel function" $K(x)$ is a function that must satisfy the following properties:

P1. The kernel function is greater than or equal to zero $K(x)\geq 0$

P2. The kernel function is maximum in $x=0$

P3. The kernel function is symmetric with respect to the vertical axis passing through $x=0$

P4. The kernel function tends to zero away from $x=0$ such that $\lim_{|x|\to+\infty}K(x)=0$

P5. The kernel function is a probability density function

The property P1 is obvious to control and it does not require any proof even in more complicated cases than the kernel function taken as an example previously.

The property P2 is more an implementation constraint rather than a property. To be satisfied, we must simply center the values (measurements) with the which we work and assume that the distribution is unimodal.

The property P3 is also more an implementation constraint than a property. It will be useful to us to simplify certain calculations such as the bias of the expected mean or of the standard deviation (see below).

The property P4 applies by common sense and derives from the property P2. In the case of unimodal probability density functions (reason why this property is often omitted from the statements).

To check property P5 it is sufficient to show that the cumulative probability is unitary.

Proof 5.48.7. We know by construction of the previous kernel function that on each point is placed a rectangle of height $1/2h$ and of width $2h$. Then, the idea is simply to write:

$$\begin{aligned}\sum_i \hat{f}(x_i)2h &= \sum_i \left[\frac{1}{nh}K\left(\frac{x-x_i}{h}\right)\right]2h = \sum_i \frac{1}{2h}\frac{\#\{x_i \in [x_i-h, x_i+h]\}}{n}2h\\ &= \sum_i \frac{\#\{x_i \in [x_i-h, x_i+h]\}}{n} = \sum_{i=1}^{n}\frac{1}{n} = n\frac{1}{n} = 1\end{aligned} \quad (5.7.2344)$$

■ Q.E.D.

Some examples of kernel functions that satisfy these properties are (the most important ones have already been dissected in the previous subsections) to a given normalization constant:

- The Gaussian kernel:

$$K(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2} \quad (5.7.2345)$$

- The Cauchy kernel:

$$K(x) = \frac{1}{\pi(1+x^2)} \quad (5.7.2346)$$

- The Picard kernel:

$$K(x) = \frac{1}{2}e^{-|x|} \quad (5.7.2347)$$

- The Epanechnikov kernel:

$$K(x) = \begin{cases}\frac{3}{4}(1-x^2) & \text{when } t \in [-1,1]\\ 0 & \text{otherwise}\end{cases} \quad (5.7.2348)$$

Let us now consider some important properties of the empirical density function.

Let us calculate its expected mean (as it is a matter of calculating the expected mean of the density function itself, this will ultimately give us the bias of the expected mean):

$$\text{E}(x) = \text{E}\left(\sum_{i=1}^{n}\frac{1}{nh}K\left(\frac{x-x_i}{h}\right)\right) = \frac{1}{nh}\sum_{i=1}^{n}\text{E}\left(K\left(\frac{x-x_i}{h}\right)\right) \quad (5.7.2349)$$

For example, for a Gaussian kernel we have:

$$\begin{aligned}\text{E}(X) &= \int x\hat{f}(x)\mathrm{d}x = \int \frac{x}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-x_i}{h}\right)\mathrm{d}x\\ &= \int \sum_{i=1}^{n}\frac{x}{nh}K\left(\frac{x-x_i}{h}\right)\mathrm{d}x = \int \sum_{i=1}^{n}\frac{x}{nh\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-x_i}{h}\right)^2}\mathrm{d}x\end{aligned} \quad (5.7.2350)$$

From this point on, the variable of which we take the expectation, that is:

$$\mathrm{E}\left(K\left(\frac{x-x_i}{h}\right)\right) \tag{5.7.2351}$$

is no longer a probability density function but a simple transformed random variable. We will therefore extend to the continuous case and multiply it by its true probability. The very clever idea being that if the bias is zero we should find:

$$\mathrm{E}(\hat{f}(x))=\mathrm{E}(f(x))=f(x) \tag{5.7.2352}$$

and if we don't fall back on this result. The difference will give us the bias we are looking for!

Let us check that:

$$\begin{aligned}\mathrm{E}(\hat{f}(x))&=\int\limits_{-\infty}^{+\infty}\frac{1}{nh}\sum_{i=1}^{n}\mathrm{E}\left(K\left(\frac{x-t}{h}\right)\right)f(t)\mathrm{d}t=\frac{1}{nh}\int\limits_{-\infty}^{+\infty}K\left(\frac{x-t}{h}\right)f(t)\mathrm{d}t\\&=\frac{1}{nh}n\int\limits_{-\infty}^{+\infty}K\left(\frac{x-t}{h}\right)f(t)\mathrm{d}t=\frac{1}{h}\int\limits_{-\infty}^{+\infty}K\left(\frac{x-t}{h}\right)f(t)\mathrm{d}t\end{aligned} \tag{5.7.2353}$$

Let us put to simplify the further developments:

$$z=\frac{x-t}{h}\Leftrightarrow t=x-hz\Rightarrow\frac{\mathrm{d}z}{\mathrm{d}t}=-\frac{1}{h} \tag{5.7.2354}$$

and therefore:

$$\mathrm{E}(\hat{f}(x))=\frac{1}{h}\int\limits_{-\infty}^{+\infty}K(z)f(x-hz)(-h)\mathrm{d}z=-\int\limits_{-\infty}^{+\infty}K(z)f(x-hz)\mathrm{d}z \tag{5.7.2355}$$

Let us develop the function $f(x-hz)$ as a Taylor series development (see section Sequences and Series page 1696) on $z=0$:

$$\begin{aligned}f(x-hz)&\cong f(x)+\frac{1}{1!}f'(x)\left((x-hz)-x\right)+\frac{1}{2!}f'(x)\left((x-hz)-x\right)^2+\mathcal{O}(h^2)\\&=f(x)+f'(x)(-hz)+\frac{1}{2!}f''(x)(-hz)^2+\mathcal{O}(h^2)\\&=f(x)-hzf'(x)+\frac{1}{2!}(hz)^2f''(x)+\mathcal{O}(h^2)\end{aligned} \tag{5.7.2356}$$

Then we have:

$$\begin{aligned}\mathrm{E}(\hat{f}(x))&=-\int\limits_{-\infty}^{+\infty}K(z)f(x-hz)\mathrm{d}z\cong-\int\limits_{-\infty}^{+\infty}\left[f(x)+hzf'(x)+\frac{(hz)^2}{2}f''(x)\right]\mathrm{d}z\\&=-f(x)\int\limits_{-\infty}^{+\infty}K(z)\mathrm{d}z+hf'(x)\int\limits_{-\infty}^{+\infty}zK(z)\mathrm{d}z+\frac{h^2}{2}f''(x)\int\limits_{-\infty}^{+\infty}z^2K(z)\mathrm{d}z\end{aligned} \tag{5.7.2357}$$

But, we know that if the kernel function is symmetric and maximal at zero (unimodal) then we can invert the sign of the first integral:

$$\mathrm{E}(\hat{f}(x))=f(x)\int\limits_{+\infty}^{-\infty}K(z)\mathrm{d}z+hf'(x)\int\limits_{-\infty}^{+\infty}zK(z)\mathrm{d}z+\frac{h^2}{2}f''(x)\int\limits_{-\infty}^{+\infty}z^2K(z)\mathrm{d}z \qquad (5.7.2358)$$

The first integral is equal to 1 since the kernel function is a probability density function. Then it remains:

$$\mathrm{E}(\hat{f}(x))=f(x)+hf'(x)\int\limits_{-\infty}^{+\infty}zK(z)\mathrm{d}z+\frac{h^2}{2}f''(x)\int\limits_{-\infty}^{+\infty}z^2K(z)\mathrm{d}z \qquad (5.7.2359)$$

The second integral calculates the expected mean. Now, since the kernel function is supposed to be symmetric and maximal in zero, it follows that its expected mean is zero and therefore it remains:

$$\mathrm{E}(\hat{f}(x))=f(x)+\frac{h^2}{2}f''(x)\int\limits_{-\infty}^{+\infty}z^2K(z)\mathrm{d}z=f(x)+\frac{h^2}{2}f''(x)\sigma^2 \qquad (5.7.2360)$$

And therefore the bias of the expected mean is:

$$\text{Bias}\left(\mathrm{E}(\hat{f}(x))\right)\cong\frac{h^2}{2}f''(x)\sigma^2 \qquad (5.7.2361)$$

Thus the bias decreases if we decrease h (but we will immediately show that it is the opposite for the variance).

Now therefore let us calculate the variance:

$$\mathrm{V}\left(\hat{f}(x)\right)=\mathrm{V}\left(\sum_{i=1}^{n}\frac{1}{nh}K\left(\frac{x-x_i}{h}\right)\right)=\frac{1}{n^2h^2}\sum_{i=1}^{n}\mathrm{V}\left(K\left(\frac{x-x_i}{h}\right)\right) \qquad (5.7.2362)$$

Because the x_i, $i=1,2,\ldots,n$, are independently distributed. Now using Huygens Theorem:

$$\begin{aligned}\mathrm{V}\left(K\left(\frac{x-x_i}{h}\right)\right)&=\mathrm{E}\left(K\left(\frac{x-x_i}{h}\right)^2\right)-\left(\mathrm{E}\left(\frac{x-x_i}{h}\right)\right)^2\\&=\int K\left(\frac{x-t}{h}\right)^2f(t)\mathrm{d}t-\left(\int K\left(\frac{x-t}{h}\right)f(t)\mathrm{d}t\right)\end{aligned} \qquad (5.7.2363)$$

Therefore:

$$\begin{aligned}\mathrm{V}(\hat{f}(x))&=\frac{1}{n}\int\frac{1}{h^2}K\left(\frac{x-t}{h}\right)^2f(t)\mathrm{d}t-\frac{1}{n}\left(\frac{1}{h}\int K\left(\frac{x-t}{h}\right)f(t)\mathrm{d}t\right)^2\\&=\frac{1}{n}\int\frac{1}{h^2}K\left(\frac{x-t}{h}\right)^2f(t)\mathrm{d}t-\frac{1}{n}\left(f(x)+\text{Bias}\left(\hat{f}(x)\right)\right)\end{aligned} \qquad (5.7.2364)$$

Substituting $z=(x-t)/h$ we get:

$$\mathrm{V}(\hat{f}(x))=\frac{1}{nh}\int K(z)^2f(x-hz)\mathrm{d}z-\frac{1}{n}\left(f(x)+\mathcal{O}(h^2)\right) \qquad (5.7.2365)$$

Applying again a Taylor approximation yields:

$$\mathrm{V}(\hat{f}(x)) \cong \frac{1}{nh}\int K(z)^2\left(f(x)-hzf'(x)+\mathcal{O}(h)\right)-\frac{1}{n}\left(f(x)+\mathcal{O}(h^2)\right)^2 \tag{5.7.2366}$$

Notice that if n becomes large and h becomes small then the above expression become approximately:

$$\mathrm{V}(\hat{f}(x)) \cong \frac{1}{nh}f(x)\int K^2(z)\mathrm{d}z \tag{5.7.2367}$$

we notice that the variance decreases as h increases.

Selection of the bandwidth h of kernel estimator is a subject of considerable research. There exist more than one heuristic popular method but at least they are all more robust than the Sturge rule!

Here are some examples without the script of some Kernel smoothing:

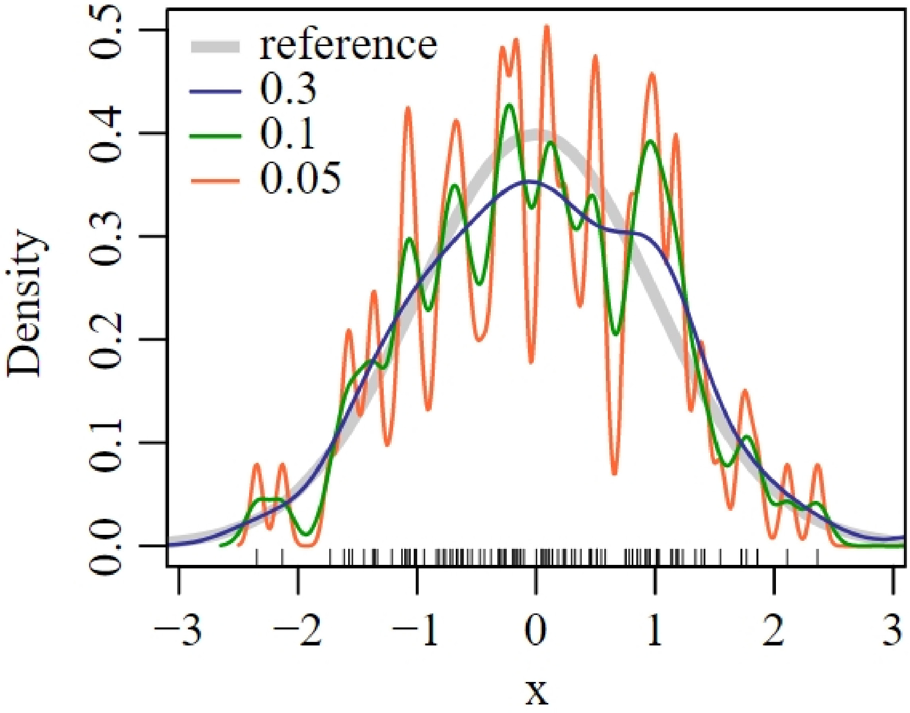


Figure 5.225: **Kernel density estimation of 100 normally distributed random numbers using different smoothing bandwidths (source: Wikipedia)**

and a comparison between a histogram and a Kernel smoothing:

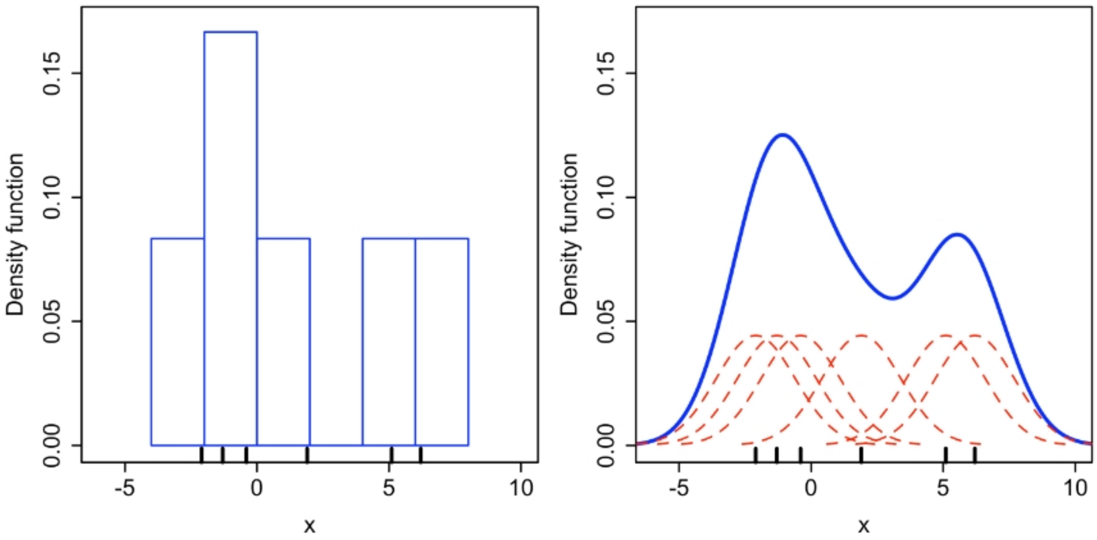


Figure 5.226: Comparison of histogram (left) and kernel density estimate (right) constructed from the same data. The individual 6 kernels are the red dotted curves, the density of the kernel estimates the blue curve. The data points are the small vertical lines the horizontal axis (source: Wikipedia)

And another interesting figure highlighting the issue of histograms [601]:

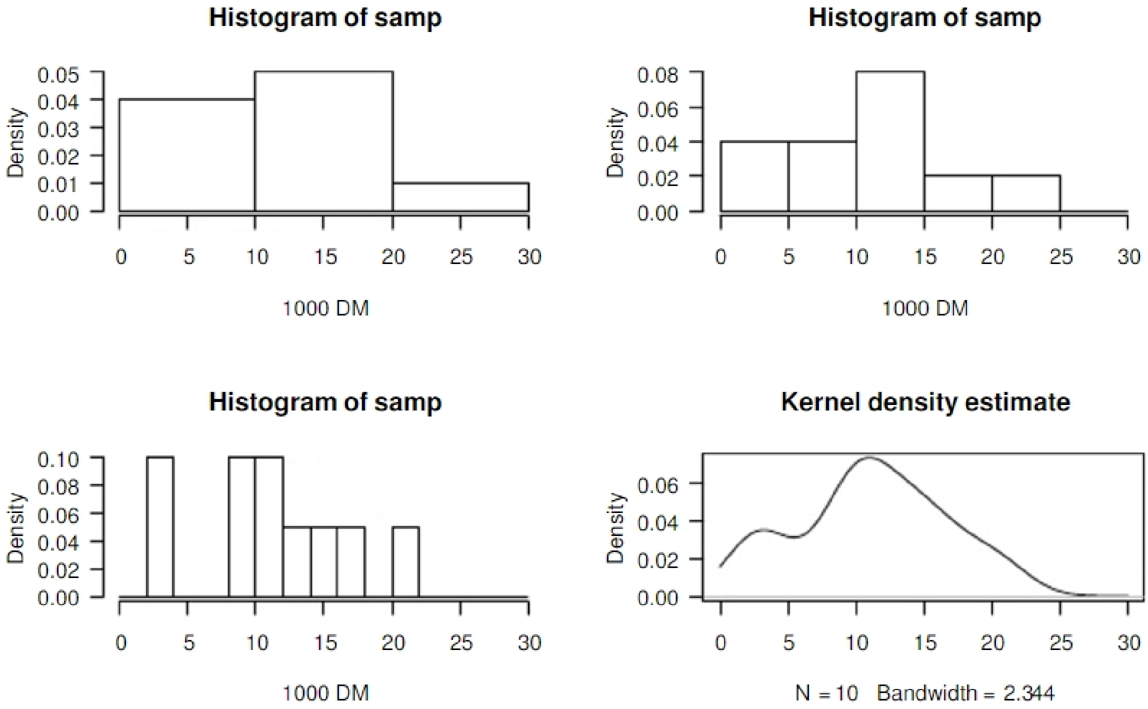


Figure 5.227: Histograms with different bin widths and a kernel estimate of $f(x)$ for the same sample

The reader interested can see our R and MATLAB® companion books to get the details on how to run Kernel smoothing on scientific softwares.

5.7.13.11 Likelihood-ratio Tests

Many statistics software returns, in complementary of classical statistical results, an output that is sometimes named "maximum likelihood test" or "maximum likelihood ratio test" (LRT) for a test whose result is already given. This maximum likelihood test is also sometimes the unique output as classical methods does not provide a computable or accurate output (this is the case of the G-test and Poisson test of means). As we will see just below, the principle is basically ... extremely simple!

Let us recall that we have proved that for a Normal law during our study of likelihood estimators that:

$$\ln\left(L(\mu,\sigma^2)\right)=-\frac{n}{2}\ln(2\pi)-n\ln(\sigma)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \tag{5.7.2368}$$

For example, if we know the standard deviation but we try to estimate the mean then this last relation can be written:

$$\ln\left(L(\hat{\mu})\right)=\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\hat{\mu})^2}{\sigma^2}\right)+c^{te} \tag{5.7.2369}$$

Then nothing prevents us from writing:

$$\ln\left(L(\mu)\right)-\ln\left(L(\hat{\mu})\right)=\left[\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}\right)+c^{te}\right]-\left[\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\hat{\mu})^2}{\sigma^2}\right)+c^{te}\right] \tag{5.7.2370}$$

Therefore after some very small elementary algebra and arithmetic manipulations:

$$-2\ln\left(\frac{L(\mu)}{L(\hat{\mu})}\right)=\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}-\sum_{i=1}^{n}\frac{(x_i-\hat{\mu})^2}{\sigma^2}=\sum_{i=1}^{n}\frac{(x_i-\hat{\mu})^2-(x_i-\mu)^2}{\sigma^2} \tag{5.7.2371}$$

In the case of following hypothesis (two sided Student T-test) with the assumption of homoscedasticity:

$$\begin{aligned}H_0&:\ \mu=\hat{\mu}\\ H_1&:\ \mu\neq\hat{\mu}\end{aligned} \tag{5.7.2372}$$

We have:

$$\frac{(x_i-\hat{\mu})^2-(x_i-\mu)^2}{\sigma^2}=(\mathcal{N}(0,1))^2 \tag{5.7.2373}$$

Therefore under the null hypothesis:

$$-2\ln\left(\frac{L(\mu)}{L(\hat{\mu})}\right)=\sum_{i=1}^{n}\frac{(x_i-\hat{\mu})^2-(x_i-\mu)^2}{\sigma^2}=\sum_{i=1}^{n}Z_i^2=\chi^2_{n-1} \tag{5.7.2374}$$

That is often denoted:

$$\boxed{\Lambda=-2\ln\left(\frac{L(\mu)}{L(\hat{\mu})}\right)=-2\left(\ln(\vec{\theta})-\ln(\hat{\vec{\theta}})\right)=\chi^2_{n-1}} \tag{5.7.2375}$$

and is the "likelihood ratio test for the comparison of the means from a Normal law" or "maximum likelihood goodness-of-test for normality".

Remark

Some softwares use for reasons I was mathematically not able to prove a chi-square distribution χ^2_{n-d-1}, where d is the number of parameters of the implicit assumed distribution law ($d=1$ for a Poisson law or $d=2$ for a Normal distribution, etc.).

Therefore we see already, before continuing, that a likelihood ratio test is a statistical test used for comparing the goodness of fit of two models, one of which (the null model) is a special case of the other (the alternative model). The test is based on the likelihood ratio, which expresses how many times more likely the data are under one model than the other. This likelihood ratio, or equivalently its logarithm, can then be used obviously to compute a p-value, or compared to a critical value to decide whether to reject the null model in favour of the alternative model. When the logarithm of the likelihood ratio is used, the statistic is known as a "log-likelihood ratio statistic", and the probability distribution of this test statistic, assuming that the null model is true, can be approximated using "Wilks' theorem" (approximation by a chi-squared distribution).

Remark

In the case of distinguishing between two models, each of which has no unknown parameters, use of the likelihood ratio test can be justified by the Neyman–Pearson lemma, which demonstrates that such a test has the highest power among all competitors. But this is out of the scope of this textbook (at least actually...).

In practice we find more often the test of the ratio of maximum likelihood with the Poisson's law. We recall that for the latter we have also proved during our study of likelihood estimators:

$$\ln\left(L(\mu)\right)=\ln(\mu)\sum_{i=1}^n x_i-\sum_{i=1}^n \ln(x_i!)-\mu n \tag{5.7.2376}$$

Therefore:

$$\ln\left(L(\mu)\right)=\sum_{i=1}^n \left(x_i\ln(\mu)-\mu\right)+c^{te} \tag{5.7.2377}$$

As for the case with the Normal law, we can write:

$$\ln\left(L(\mu)\right)-\ln\left(L(\hat{\mu})\right)=\left[\sum_{i=1}^n \left(x_i\ln(\mu)-\mu\right)+c^{te}\right]-\left[\sum_{i=1}^n \left(x_i\ln(\hat{\mu})-\hat{\mu}\right)+c^{te}\right] \tag{5.7.2378}$$

Therefore:

$$\ln\left(\frac{L(\mu)}{L(\hat{\mu})}\right)=\sum_{i=1}^n\left[x_i\ln\left(\frac{\mu}{\hat{\mu}}\right)-\mu+\hat{\mu}\right] \tag{5.7.2379}$$

What it is customary to write:

$$-2\ln\left(\frac{L(\mu)}{L(\hat{\mu})}\right)=-2\sum_{i=1}^n\left[n_i\ln\left(\frac{\mu}{\hat{\mu}}\right)-\mu+\hat{\mu}\right] \tag{5.7.2380}$$

Suppose we want to test the fact that our measurements do not deviate too much from the true expectation, that is to say:

$$\begin{aligned} H_0 &: \mu = \hat{\mu} \\ H_1 &: \mu \neq \hat{\mu} \end{aligned} \tag{5.7.2381}$$

We have the under the null hypothesis:

$$\boxed{\Lambda = -2\ln\left(\frac{L(\mu)}{L(\hat{\mu})}\right) = 0} \tag{5.7.2382}$$

So.... arrrrgh. We cannot make inference on such a result. So it's very annoying to deal with such a likelihood ratio ... Fortunately we know that in the case where:

$$\sigma^2 = \mu \qquad \mu \rightarrow +\infty \tag{5.7.2383}$$

the Poisson's law tends to a Normal law (if μ is greater than about 10, then the Normal distribution is a good approximation if an appropriate continuity correction is performed). Then in these conditions we use the previous result of the Normal law!

More generally we can (sadly) found the above final relations in the following forms in various textbooks:

$$\boxed{\Lambda(x) = \frac{L(\theta_0 \,|\, x)}{L(\theta_1 \,|\, x)} = \frac{f(\bigcup_i x_i \,|\, \theta_0)}{f(\bigcup_i x_i \,|\, \theta_1)} = \frac{L(\theta_0 \,|\, x)}{\sup\{\, L(\theta \,|\, x) : \theta \in \{\theta_0, \theta_1\}\}}} \tag{5.7.2384}$$

So many common test statistics such as the Z-test, the F-test, the T-test, the Pearson's chi-squared test and the G-test are tests for nested models and can be phrased as log-likelihood ratios or approximations thereof.

Remark

Some softwares propose another adequation test for discrete distribution, named the "minimum chi-square estimation goodness-of-fit". The idea is quite simple and we will illustrate it with a Poisson distribution. Consider the following sample ($n = 8$ for a total sample size equal to 20):

Value	**Frequency**
0	1
1	2
2	4
3	5
4	3
5	3
6	1
7	0
8	1

The minimum chi-square estimate of the population of parameter λ, is the λ value that minimize:

$$\sum_{k=0}^{n}\frac{(\text{Observed}-\text{Expected})^2}{\text{Expected}}=\sum_{k=0}^{8}\frac{\left((\text{count in cell } k)-20\left(\frac{\lambda^k e^{-\lambda}}{k!}\right)\right)^2}{20\left(\frac{\lambda^k e^{-\lambda}}{k!}\right)} \quad (5.7.2385)$$

Numerical computation shows that the value of λ that minimizes the chi-square statistic is about 3.5242. That is the minimum chi-square estimate of λ. For that value of λ, the chi-square statistic is about 3.062764. There are 10 cells. If the null hypothesis had specified a single distribution, rather than requiring λ to be estimated, then the null distribution of the test statistic would be a chi-square distribution with $10-1=9$ degrees of freedom.

5.7.14 Robustness and Nonparametric Statistics

In the area of inferential statistics and hypothesis testing, robustness is a recurring concept (the banks are compelled to stress testing / crash testing of their risk models for example). We have already mentioned this fact previously.

Definitions (#5.7.153):

D1. A test is named "robust test" if it remains valid while the assumptions for the application are not all met. This can be a somewhat low sample size or a probability distribution (Normal distribution for parametric tests) that is not very well satisfied, or the outliers that influence too much the result of the test. For example, ANOVA is robust with respect to the Normality assumption but not compared to that of homoscedasticity.

D2. An indicator is named "robust indicator" if it is not very sensitive to the presence of outliers (the correlation coefficient or the average, for examples, are not very robust indicators).

D3. More generally, a model is named "robust model" when it allows an extension of the results (in time or for a population). The robustness is equally applicable to a multiple regression, statistical tests or to a scorecard, or project planning.

Therefore, unless to be only descriptive, your studies will have to respect some rules so that their findings are generalisable.

First condition of a good robustness: the data! Intuitively, we all know that we do not transform a special case in generality (which would not fall in the field of statistics but to countertop discussions). Sufficient data builds reliable and solid models. For example, forecasts derived from a time series showing seasonality require at least three or four years of history.

However, the amount is not enough, we need quality (for most manager Quality is a very difficult concept)!! Indeed, it is best to avoid doing a study on unreliable information that can lead to costly decisions. Furthermore, it is recommended to be very careful with the manipulation of outliers. If this is not possible, we turn to more appropriate methods, such as those that using the median rather than the mean, ranks rather than values, or filtering techniques.

Definition (#5.7.154): "Nonparametric statistics" is the branch of statistics that is not based solely on parametrized families of probability distributions (common examples of parameters are the mean and variance).

The term "nonparametric statistics" has been imprecisely defined in the following two ways (among others...):

1. They do not rely on assumptions that the data are drawn from a given probability distribution. We then say that they are "distribution free methods".
2. They do not rely on any arithmetical operation of all the individual data (like "ranks" or "ranges" for example!)

As we will see further below and as we have also already seen some special cases earlier above, there is quite a lot of nonparametric statistics and nonparametric tests (keep in mind that even if their are not based on parametrized families of distributions, they have however other assumptions that are visible only when your study their mathematical proof!!!). Rather than listing them all here, we rather want to answer the most common question from many practitioners that is a non-exhaustive list of the equivalent nonparametric to some parametric tests and statistics. So here is a table that summarize for the most important one that question request:

Parametric Tests	Nonparametric Tests	Main characteristics
1-sample T or Z test	1-sample sign test	Test on the median for data from a nonsymmetric distribution
	1-sample Wilcoxon test	Test on the median for data from a symmetric distribution
2-sample T-test	Mann-Whithney test	Test on two medians using ranks of the sample data
One-Way ANOVA	Kruskal-Wallis test	Test on the equality of medians from two or more populations. More powerful than Mood's median test, but less robust to outliers
	Mood's median test	Test on the equality of medians from two or more populations. More robust to outliers than the Kruskal-Wallis test, but less powerful.
Two-Way Randomized Block ANOVA	Friedman test	Test on medians, using randomized block experiments
Pearson's Correlation Coefficient	Spearman's rank correlation coefficient	Measures statistical dependence between two variables using ranks
Anderson-Darling, Shapiro-Wilk, Kolmogorov-Smirnov GoF	Mann–Whitney U or Wilcoxon rank sum test	Tests whether two samples are drawn from the same distribution, as compared to a given alternative hypothesis.

Table 5.76: **Summary of some naive parametric tests/statistics and their nonparametric equivalent**

For a much more exhaustive list of parametric and non-parametric tests, the reader has to refer to the Data Science mind map at page 6770 or to the both excellent textbooks [304] and [430].

Or for those who prefer a decision tree:

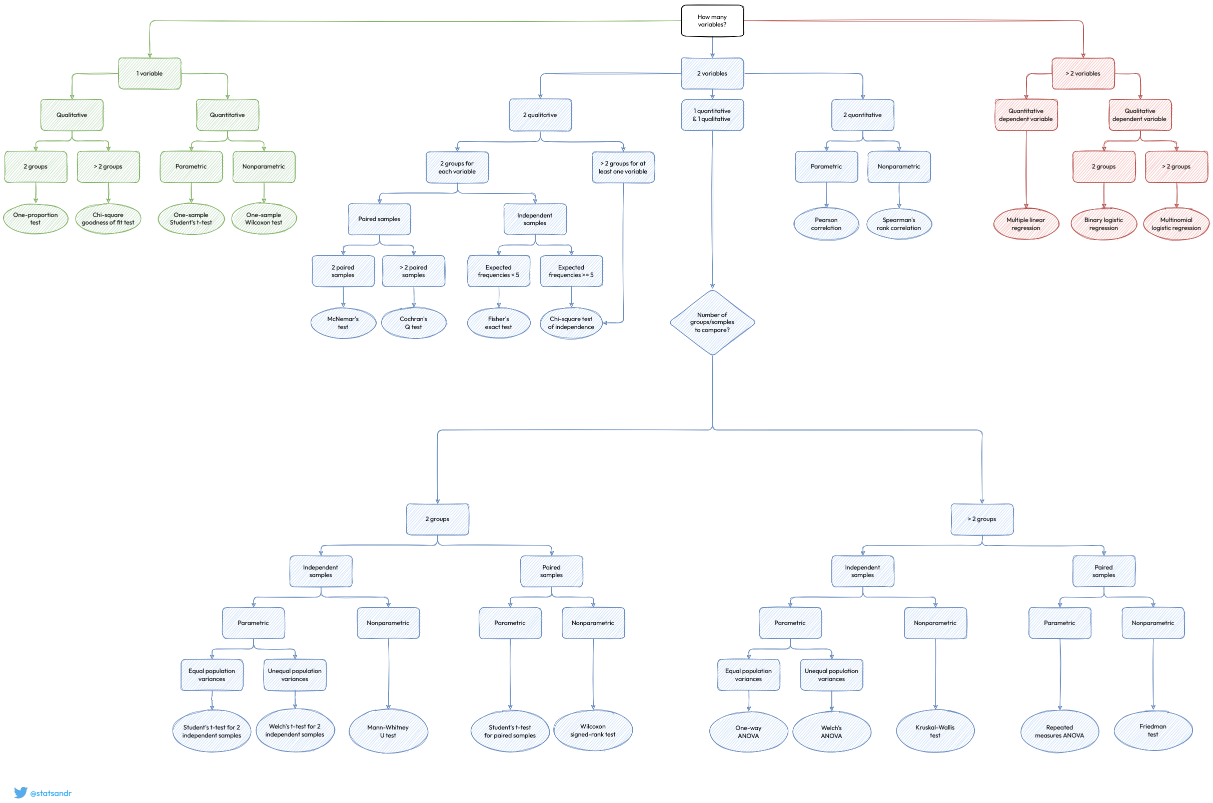

How many variables?
1 variable
Qualitative
Quantitative
2 groups
> 2 groups
Parametric
Nonparametric
One-proportion test
Chi-square goodness of fit test
One-sample Student's t-test
One-sample Wilcoxon test
2 variables
2 qualitative
1 quantitative & 1 qualitative
2 quantitative
2 groups for each variable
> 2 groups for at least one variable
Paired samples
Independent samples
2 paired samples
> 2 paired samples
Expected frequencies < 5
Expected frequencies >= 5
McNemar's test
Cochran's Q test
Fisher's exact test
Chi-square test of independence
Parametric
Nonparametric
Pearson correlation
Spearman's rank correlation
> 2 variables
Quantitative dependent variable
Qualitative dependent variable
2 groups
> 2 groups
Multiple linear regression
Binary logistic regression
Multinomial logistic regression
Number of groups/samples to compare?
2 groups
Independent samples
Paired samples
Parametric
Nonparametric
Parametric
Nonparametric
Equal population variances
Unequal population variances
Student's t-test for 2 independent samples
Welch's t-test for 2 independent samples
Mann-Whitney U test
Student's t-test for paired samples
Wilcoxon signed-rank test
> 2 groups
Independent samples
Paired samples
Parametric
Nonparametric
Parametric
Nonparametric
Equal population variances
Unequal population variances
One-way ANOVA
Welch's ANOVA
Kruskal-Wallis test
Repeated measures ANOVA
Friedman test
@statsandr

Figure 5.228: **Decision tree for naive statistical tests (author: Antoine Soetewey)**

5.7.14.1 Rank Statistics

Definition (#5.7.155): The Rank Statistics, also named "Order Statistics" are defined as all the techniques of statistical calculations and statistical inference that have for main objective of getting rid of the knowledge of a parametric distribution and for using it ranks only (order) of the measurements. This is a very powerful and practically used tool for nonparametric statistics!

As we have already mentioned, we are talking about parametric tests when we stipulate that the data are from a given known distribution. In this case, the characteristics of the data can be summarized using the parameters estimated on the sample, the subsequent test procedure then only focus on these parameters.

Remark

The reader interested to learn more but without detailed is referred to the excellent book of Gopal K. Kanji which contains a summary presentation of the most used 100 parametric and nonparametric test with for each a small practical example. Sadly this textbook should be published in a new edition with the 200 most used statistical tests to be almost complete and perfect.

Nonparametric tests (such as the chi-square tests already seen) eliminate the necessary step consisting to estimate the parameters of the distribution prior to doing the hypothesis test. In general, conclusions drawn from nonparametric tests are not as powerful as the parametric ones. However, as nonparametric tests make fewer assumptions, they are more flexible, more robust, and applicable to non-quantitative data.

When the data are quantitative, most common nonparametric tests transforms values into ranks. The name "ranks test" is then often encountered. When the data are qualitative, only the nonparametric can be used.

5.7.14.1.1 L-Statistics

Before addressing nonparametric distributions and tests, let us give some definitions that the reader may have seen be in the hyperspecialized literature and that we avoided to use so far...

The median, mean and range suggest the use of linear combinations of the components of the order statistics vector.

Definition (#5.7.156): Thus, let us denote by $X_{(1)} \geq X_{(2)}, \ldots, \geq X_{(n)}$, the order statistics (values ordered in descending order by their rank and numbered). We define then the "L-statistic" by:

$$\boxed{L = \sum_{i=1}^{n} a_{i,n} X_{(i)}} \quad (5.7.2386)$$

The first most known "L-estimator" is the arithmetic mean for which:

$$a_{i,n} = \frac{1}{n} \quad (5.7.2387)$$

The second most known L-estimator is the median for which we have when n is odd:

$$a_{i,n}=\begin{cases}1 & i=\dfrac{n+1}{2}\\ 0 & i\neq\dfrac{n+1}{2}\end{cases} \tag{5.7.2388}$$

and when n is even:

$$a_{i,n}=\begin{cases}1 & i=\dfrac{n}{2},\dfrac{n}{2}+1\\ 0 & \text{otherwise}\end{cases} \tag{5.7.2389}$$

The third best L-estimator known is the range for which we have:

$$a_{i,n}=\begin{cases}+1 & i=1\\ -1 & i=n\\ 0 & \text{otherwise}\end{cases} \tag{5.7.2390}$$

There are other empirical L-estimators but we will stop here because the list is quite long.

5.7.14.1.2 Ranks Distribution law

The k-th order statistic is the kth smallest value in sample of n from a random distribution $F(x)$. Let the random variables $X_1,X_2,\ldots,X_n$ be independent and identically distributed. Label the smallest value $X_{(1)}$, the next smallest $X_{(2)}$ and so on up to $X_{(n)}$. Then the value of the kth order statistic is $X_(k)$, for $1\leq k\leq n$.

Let $F(x)$ denote the distribution function of the X_i's and $f(x)$ the density function. Let us first determine the distribution of the kth order statistic, that is:

$$P(X_{(k)}\leq x) \tag{5.7.2391}$$

To have $X_{(k)}\leq x$ we need to have at least k of the n X_i's less than or equal to x. Each of the n independent trials has probability $F(x)$ of being less than or equal to x, thus the number of trials less than or equal to x has a Binomial distribution with n trials and probability $F(x)$. That is:

$$\begin{aligned}P(X_{(k)}\leq x)&=P\left(\text{at least } k \text{ success from a } \mathcal{B}(n,F(x)) \text{ distribution}\right)\\&=\sum_{j=1}^{n}P(j\text{ sucesses})\\&=\sum_{j=1}^{n}\binom{n}{j}\left(F(x)\right)^j\left(1-F(x)\right)^{n-j}\end{aligned} \tag{5.7.2392}$$

To find the density function we take the derivative with respect to x:

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}x}P(X_{(k)}\leq x)\\
&=\frac{\mathrm{d}}{\mathrm{d}x}\left(\sum_{j=1}^{n}\binom{n}{j}(F(x))^j(1-F(x))^{n-j}\right)\\
&=\sum_{j=k}^{n-1}\binom{n}{j}\left[j(F(x))^{j-1}f(x)(1-F(x))^{n-1}-(n-j)(F(x))^j(1-F(x))^{n-j-1}f(x)\right]\\
&+nf(x)(F(x))^{n-1}\\
&=nf(x)\sum_{j=k}^{n-1}\left[\binom{n-1}{j-1}(F(x))^{j-1}(1-F(x))^{n-1}-\binom{n-1}{j}(F(x))^j(1-F(x))^{n-j-1}\right]\\
&+nf(x)(F(x))^{n-1}\\
&=nf(x)\left(\sum_{j=k-1}^{n-2}\binom{n-1}{j}(F(x))^j(1-F(x))^{n-j-1}-\sum_{j=k}^{n-1}(F(x))^j(1-F(x))^{n-j-1}\right)\\
&+nf(x)(F(x))^{n-1}\\
&=nf(x)\binom{n-1}{k-1}(F(x))^{k-1}(1-F(x))^{n-(k-1)-1}-nf(x)\binom{n-1}{n-1}(F(x))^{n-1}(1-F(x))^{n-(n-1)-1}\\
&+nf(x)(F(x))^{n-1}\\
&=nf(x)\binom{n-1}{k-1}(F(x))^{k-1}(1-F(x))^{n-(k-1)-1}-nf(x)(F(x))^{n-1}+nf(x)(F(x))^{n-1}\\
&=nf(x)\binom{n-1}{k-1}(F(x))^{k-1}(1-F(x))^{n-(k-1)-1}
\end{aligned}
\tag{5.7.2393}
$$

Written in this form, it is obvious that the majority of terms will cancel. This leaves:

$$\frac{\mathrm{d}}{\mathrm{d}x}P(X_{(k)}\leq x)=\frac{n!}{(n-k)!(k-1)!}f(x)(F(x))^{k-1}(1-F(x))^{n-k}\tag{5.7.2394}$$

5.7.14.1.3 Distribution free confidence intervals for percentiles

Earlier we learned (among a lot of others stuff) how to calculate a sample percentile as a point estimate of a population (or distribution) percentile. Just as it is a good idea to calculate confidence intervals for other population parameters, such as means and variances, it would be a good idea to learn how to calculate a confidence interval for percentiles of a population. That's what we'll work on just below. As the title of the subject suggests it, we won't make any assumptions about the distribution of the data, that is, other than it being continuous!

As is generally the case, let's motivate the method for calculating a confidence interval for a population median M_e by way of a concrete example[85]. Suppose $Y_1<Y_2<Y_3<Y_4<Y_5$ are the order statistics of a random sample of size $n=5$ from a continuous distribution. Our previous work tells us that Y_3 serves as a good point estimator of the median M_e. Let's see what we

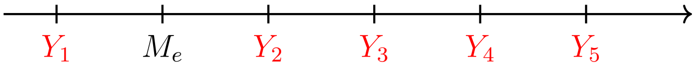
---

[85] It's a well knows example that you can find in numerous textbooks and websites which original author seems impossible to determine.

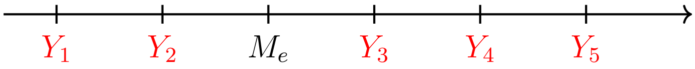
can come up with for a confidence interval given we have these order statistics at our disposal. Well, suppose we suggested that the interval constrained by the first and fifth order statistics, that is, $[Y_1, Y_5]$ would serve as a good interval. How confident can we be that the interval $[Y_1, Y_5]$ would contain the unknown population median M_e? To answer that question, we simply need to calculate the following probability:

$$P(Y_1 < M_e < Y_5) \tag{5.7.2395}$$

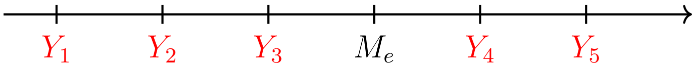
Calculating the probability reduces to a simple binomial calculation once we figure out all the ways in which the population median M_e is sandwiched between Y_1 and Y_5. Well, the population median M_e is sandwiched between Y_1 and Y_5, if the first order statistic is the only order statistic less than the median M_e:



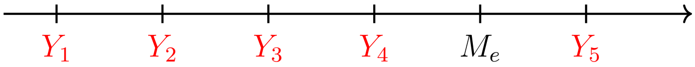
The population median M_e is sandwiched between Y_1 and Y_5, if the first two order statistics are the only order statistics less than the median M_e:



The population median M_e is sandwiched between Y_1 and Y_5, if the first three order statistics are less than the median M_e, and the fourth and fifth order statistics are greater than M_e:



And, the population median M_e is sandwiched between Y_1 and Y_5, if the fifth order statistic is the only order statistic greater than the median M_e:



This means that in order to calculate the probability $P(Y_1 \leq M_e \leq Y_5)$, we need to calculate the probability of each of the above events. Now, if we let W denote the number of $X_i < M_e$, then W is a binomial random variable with n mutually independent trials and probability of success $p = P(X_i < M_e) = 0.5$. And, reviewing the events as depicted above, the desired probability is calculated as:

$$P(Y_1 \leq M_e \leq Y_5) = P(W=1) + P(W=2) + P(W=3) + P(W=4) \tag{5.7.2396}$$

The binomial probability mass function (or, alternatively, the binomial table) makes the calculation straightforward:

$$P(Y_1 \leq M_e \leq Y_5) = \sum_{k=1}^{4} P(W=k) = \sum_{k=1}^{4} \binom{5}{k} (0.5)^k (0.5)^{5-k} = 0.9376 \tag{5.7.2397}$$

So, the probability that the random interval $[Y1, Y5]$ contains the median M_e is 0.9376. We aren't always so lucky with arriving at a decent confidence coefficient on our first try. Sometimes we have to try again aiming to get a confidence coefficient that it as least 90%, but as close to 95% as possible. In this case, the confidence coefficient for the interval $[Y2, Y4]$ is:

$$P(Y_2 \leq M_e \leq Y_4) = \sum_{k=2}^{3} P(W=k) = \sum_{k=2}^{3} \binom{5}{k} (0.5)^k (0.5)^{5-k} = 0.6250 \qquad (5.7.2398)$$

Clearly, we would be better served to stick with the interval $[Y_1, Y_5]$ in this case.

Notice that in general we then have:

$$\boxed{P(Y_i \leq M_e \leq Y_j) = \sum_{k=i}^{j-1} P(W=k) = \sum_{k=i}^{j-1} \binom{n}{k} (0.5)^k (0.5)^{n-k}} \qquad (5.7.2399)$$

The method that we learned for finding a confidence interval for the median of a continuous distribution can be easily extended so that we can find a confidence interval for any percentile π_p. The only thing we have to change is the probability of a success, that is, that X_i is less than π_p:

$$p = P(X_i < \pi_p) \qquad (5.7.2400)$$

Then, the exact confidence coefficient is calculated just as before using the binomial distribution with parameters n and p:

$$\boxed{P(Y_i \leq \pi_p \leq Y_j) = \sum_{k=i}^{j-1} \binom{n}{k} p^k (1-p)^{n-k}} \qquad (5.7.2401)$$

5.7.14.1.4 Wilcoxon Rank Sum Test

The idea of the "Wilcoxon rank sum test" is the following: if we collect two samples of measurement, and that we store the values in order, the alternance of the X_i (of size n_x) and Y_i (of size n_y) should be fairly steady if both samples distribution law F and G respectively follow the same probability distribution. It is therefore like a "fit check".

This is in fact equivalent as saying that the null hypothesis is: «*The distribution of the ranks in the two groups are equal*».

Formally:

$$H_0 : P(X_A > X_B) = 0.5; H_A : P(X_A > X_B) \neq 0.5 \qquad (5.7.2402)$$

In plain language the null hypothesis is that the probability of a random observation from group A exceeding the paired observation drawn from group B is one half (i.e. a random observation in group A has just as much probability of being greater than, as being less than its paired observation in group B).

The practical value of this is hard to see, and thus in many places, including textbooks, the null hypothesis is presented as the “two populations have equal medians”. The actual null hypothesis can be expressed as the latter median hypothesis, but only under the additional assumption that the shapes of the distributions are identical in each group!

In other words, our interpretation of the test as comparing the medians of the distributions requires the location-shift-only alternative to be the case. Since this is rarely true, and never assessed, we should probably use extreme caution in using, and especially in interpreting, the Wilcoxon rank-sum test.

It is therefore not such as Chi-square adequation test to compare measurements at a theoretical law, but to other measurement of a hypothesized same law.

Remark

The Wilcoxon rank sum test is a nonparametric test, used therefore for "Nonparametric Null Hypothesis Significance Tests" (NNHST), because we have no need of use of any measure of dispersion or position of the random variables. Moreover, it is a test say to be "robust" in the sense that it does not assume the Normality of the data.

Let us take an example before tackling the theoretical aspect. Here are two samples of size 10 ($n_x = n_y$) of quantitative variables:

$$\begin{gathered}5.7, 3.2, 8.4, 4.1, 6.9, 5.3, 1.7, 3.2, 2.5, 7.4\\ 8.1, 5.5, 3.4, 7.9, 4.6, 1.6, 8.5, 7.1, 8.7, 5.7\end{gathered} \tag{5.7.2403}$$

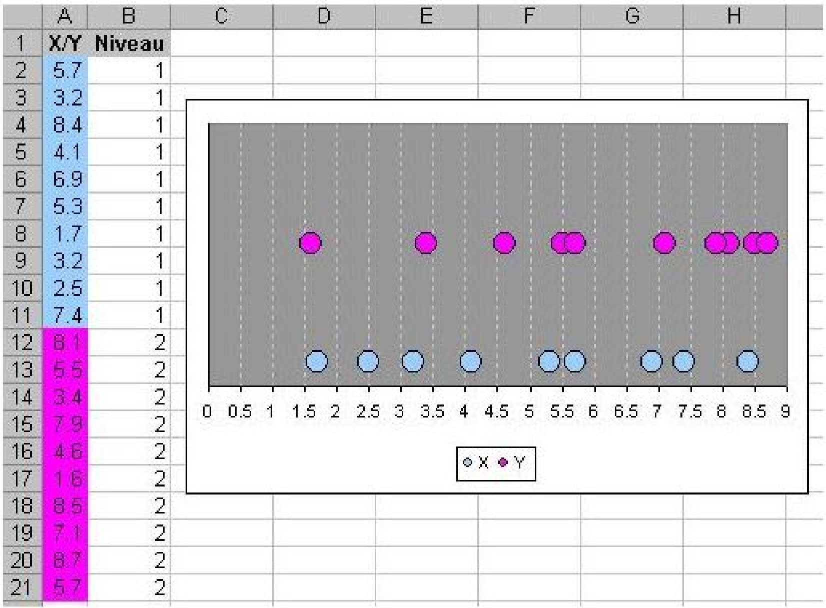
Remark

The Wilcoxon rank sum test may well be used for ordinal variables (but only as long as they are in an acceptable number). Typically, the Wilcoxon rank sum test is also used to analyse the response to business surveys using a Likert scales of 7 points.

Here are the order statistics of the sample of size 20 ($N = n_x + n_y$) regrouped and ordered (the 10 values of the first sample are underlined):

$$1.6, \underline{1.7}, \underline{2.5}, \underline{3.2}, \underline{3.2}, 3.4, \underline{4.1}, 4.6, \underline{5.3}, 5.5, \underline{5.7}, 5.7, \underline{6.9}, 7.1, \underline{7.4}, 7.9, 8.1, \underline{8.4}, 8.5, 8.7 \tag{5.7.2404}$$

The values of the first sample X (named "treatment values") in this example looks like to be smaller than those of the second sample Y (named "control values") that we often represent in a diagram as the following (by cheating a bit with Microsoft® Excel® 11.8346):



	A	B
1	X/Y	Niveau
2	5.7	1
3	3.2	1
4	8.4	1
5	4.1	1
6	6.9	1
7	5.3	1
8	1.7	1
9	3.2	1
10	2.5	1
11	7.4	1
12	8.1	2
13	5.5	2
14	3.4	2
15	7.9	2
16	4.6	2
17	1.6	2
18	8.5	2
19	7.1	2
20	8.7	2
21	5.7	2

Figure 5.229: **Comparing ordered values of two samples in Microsoft® Excel® 11.8346**

Remark

You must be careful about the Wilcoxon test we are studying here and the Kruskal-Wallis test that we will see later. Indeed, many people think that because they are nonparametric they don't have any assumptions... However as we know this is wrong!!! A nonparametric test is indeed independent of distribution parameters but there may have other assumptions. In the case of the Wilcoxon and the Kruskal-Wallis test, the assumption is (quite obviously...) that the scale parameters between the both samples are equal (otherwise it would be a non-sense to compare two things that we already know as being not comparable!).

The idea is then to find out if this trend is statistically significant? That is to say, whether we have a difference of the kind $F < G$ between their respective distributions laws:

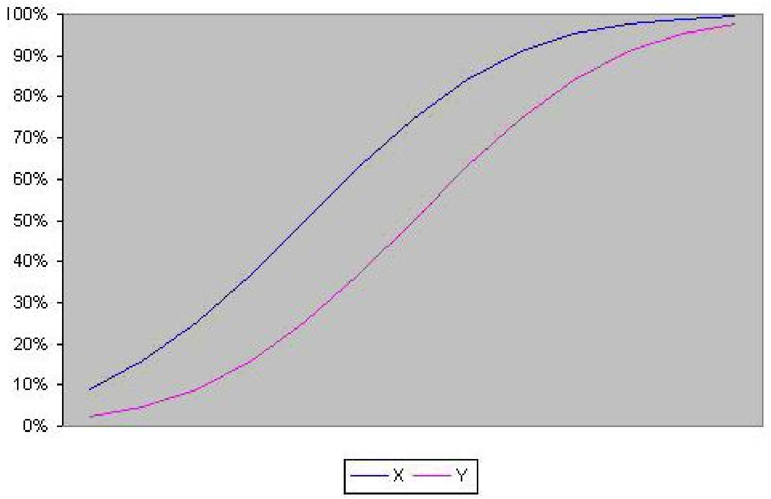


Figure 5.230: **Generic example of the comparison of two distributions in Microsoft® Excel® 11.8346**

or whether they can be considered as identical. For this, we must examine the concept of "rank":

Definition (#5.7.157): Given a random n-sample $X_1, X_2, \ldots, X_n$ of any continuous statistical law, we denote by R_i the rank of the X_i ordered in a sample population. The rank i is a non-zero integer strictly positive and between 1 and n.

✍**Example:**

In:

$$1.6, \underline{1.7}, \underline{2.5}, \underline{3.2}, \underline{3.2}, 3.4, \underline{4.1}, 4.6, \underline{5.3}, 5.5, \underline{5.7}, 5.7, \underline{6.9}, 7.1, \underline{7.4}, 7.9, 8.1, \underline{8.4}, 8.5, 8.7$$

We have respectively the following "order statistics":

$$X: R_2 = 1.7, R_3 = 2.5, R_4 = 3.2, R_5 = 3.2, R_7 = 4.1, \ldots \tag{5.7.2405}$$

$$Y: R_1 = 1.6, R_6 = 3.4, R_8 = 4.6, R_{10} = 5.5, R_{11} = 5.7, \ldots \tag{5.7.2406}$$

Once the concept of "rank" defined and calculated, let us look at the sum in the context of our example with two samples:

$$1.6, \underline{1.7}, \underline{2.5}, \underline{3.2}, \underline{3.2}, 3.4, \underline{4.1}, 4.6, \underline{5.3}, 5.5, \underline{5.7}, 5.7, \underline{6.9}, 7.1, \underline{7.4}, 7.9, 8.1, \underline{8.4}, 8.5, 8.7$$

The sum rank denoted traditionally W_x (W for Wilcoxon) for the first sample is thus:

$$W_x = 2+3+4+5+7+9+12+13+15+18 = 88 \tag{5.7.2407}$$

and for the second sample:

$$W_y = 1+6+8+10+11+14+16+17+19+20 = 122 \tag{5.7.2408}$$

The values W_x, W_y are named "Wilcoxon Statistics".

We can therefore already see that there is indeed a difference that seems a priori significant in terms of ranking between both samples. The problem remains to build a rigorous mathematical tool to infer a fact with some certainty.

For this, let us first introduce the average of the ranks using the result shown in the section of Sequences and Series by considering only one sample:

$$\bar{R} = \frac{1}{n}\sum_{i=1}^{n} R_i = \frac{1}{n}\sum_{i=1}^{n} i = \frac{1}{n}\frac{n+1}{2} \tag{5.7.2409}$$

By calculating this, we notice quickly that this is the expected mean of the uniform discrete distribution study earlier in this section for a discrete random variable with values between 1 and n, which is exactly the definition of rank! Thus, we have the rank that will have for mean and variance for the entire population:

$$\begin{aligned} \mu = \bar{R} &= \frac{n+1}{2} \\ \mathrm{V}(R) &= \frac{n^2-1}{12} \end{aligned} \tag{5.7.2410}$$

For those who find this analogy questionable we give just right below the proof of the variance using the Huygens relation and the sum of squares of positive integers proved in the section of

Sequences and Series:

$$\begin{aligned}\mathrm{V}(R)&=\mathrm{E}(R^2)-\mathrm{E}(R)^2=\sum_{i=1}^{n}p(R_i)R_i^2-\left(\frac{n+1}{2}\right)^2\\&=\frac{1}{n}\frac{n(2n+1)(n+1)}{6}-\frac{(n+1)^2}{4}=\frac{2(2n+1)(n+1)-3(n^2+n+1)}{12}\\&=\frac{2(2n^2+2n+n+1)-3n^2-6n-3}{12}=\frac{4n^2+4n+2n+2-3n^2-6n-3}{12}=\frac{n^2-1}{12}\end{aligned}\tag{5.7.2411}$$

But obviously for a single sample this has no interest! Let us take again our two series X_i, Y_i respectively of equal size $n_x=10, n_y=10$ without distinction:

$$1.6, 1.7, 2.5, 3.2, 3.2, 3.4, 4.1, 4.6, 5.3, 5.5, 5.7, 5.7, 6.9, 7.1, 7.4, 7.9, 8.1, 8.4, 8.5, 8.7\tag{5.7.2412}$$

We then have the statistical indicators of ranks without distinction (you must remember that we still do not know at this level of the mathematical development if this will be helpful or not):

$$\begin{aligned}W_{x,y}&=\sum_{i=1}^{n_x+n_y}R_i=\sum_{i=1}^{n_x+n_y}i=\frac{(n_x+n_y)(n_x+n_y+1)}{2}=210\\\mu_{x,y}&=\bar{R}_{x,y}=\frac{(n_x+n_y)+1}{2}=10.5\\\mathrm{V}(R_{x,y})&=\frac{(n_x+n_y)^2-1}{12}=33.25\end{aligned}\tag{5.7.2413}$$

and the statistical indicators of ranks but this time with distinction:

$$1.6, \underline{1.7}, \underline{2.5}, \underline{3.2}, \underline{3.2}, 3.4, \underline{4.1}, 4.6, \underline{5.3}, 5.5, \underline{5.7}, 5.7, \underline{6.9}, 7.1, \underline{7.4}, 7.9, 8.1, \underline{8.4}, 8.5, 8.7$$

We then have the following local statistical indicators:

$$\begin{aligned}W_x&=\sum R_{x,i}=2+3+4+5+7+9+12+13+15+18=88\\W_y&=\sum R_{y,i}=1+6+8+10+11+14+16+17+19+20=122\\\bar{R}_x&=\frac{1}{n}\sum R_{x,i}=\frac{2+3+4+5+7+9+12+13+15+18}{10}=8.8\\\bar{R}_y&=\frac{1}{n}\sum R_{y,i}=\frac{1+6+8+10+11+14+16+17+19+20}{10}=12.2\\\mathrm{V}(R_x)&=\frac{1}{n_x}\sum(R_{x,i}-\bar{R}_x)^2=27.16\\\mathrm{V}(R_x)&=\frac{1}{n_y}\sum(R_{y,i}-\bar{R}_y)^2=33.56\end{aligned}\tag{5.7.2414}$$

These calculations now done, we have nothing concretely yet rigorous regarding the Wilcoxon's rank sum test which purpose is for reminder to check whether the two samples follow the same law or not (and therefore have the same moments as the expected mean, variance, median, etc.).

To move forward, let us consider the n_x values of the sample X. We know (see section Probabilities page 528) then that there is:

$$C_{n_x}^{n_x+n_y}=\binom{n_x+n_y}{n_x}=\frac{(n_x+n_y)!}{n_x!((n_x+n_y)-n_x)!}=\frac{(n_x+n_y)!}{n_x!n_y!}\tag{5.7.2415}$$

number of possible arrangements of the X_i in the population of the samples and if the Wilcoxon's rank sum test is verified (that is to say, the probability laws are the same for both samples), the various arrangements should be equally likely.

For example, if we take 2 samples with respectively each 2 measures (2 random variables of treatment and 2 control random variables), we have:

$$(n_x+n_y)!=4!=24 \tag{5.7.2416}$$

arrangements all different:

$$\begin{matrix} (1,2,3,4) & (2,1,3,4) & (3,1,2,4) & (4,1,2,3) \\ (1,2,4,3) & (2,1,4,3) & (3,1,4,2) & (4,1,3,2) \\ (1,3,2,4) & (2,3,1,4) & (3,2,1,4) & (4,2,1,3) \\ (1,3,4,2) & (2,3,4,1) & (3,2,4,1) & (4,2,3,1) \\ (1,4,2,3) & (2,4,1,3) & (3,4,1,2) & (4,3,1,2) \\ (1,4,3,2) & (2,4,3,1) & (3,4,2,1) & (4,3,2,1) \end{matrix} \tag{5.7.2417}$$

But... sadly... this is not what we want in our case because we would like already to be able to distinguish the two samples and also do not take into account the arrangements that consist only of a permutation of the variables of the same sample. We then have (see section Probabilities page 528):

$$C_{n_x}^{n_x+n_y}=\binom{n_x+n_y}{n_x}=\frac{4!}{2!2!}=6 \tag{5.7.2418}$$

possible combinations! Effectively with two samples having two treatment variables (X) and two control variables (Y), we have (W_S is the sum of ranks of last column):

Possible ranks (Controls)	**Possible ranks (Treatments)**	W_S
$1,2$	$3,4$	7
$1,3$	$2,4$	6
$1,4$	$2,3$	5
$2,3$	$1,4$	4
$2,4$	$1,3$	4
$3,4$	$1,2$	3

Table 5.77: **Ranks representation of 2 variables of treatment and of control**

If the null hypothesis of the Wilcoxon's sum test rank is not rejected, the 6 rankings are equally likely. We conclude the following table:

Value of W_S	3	4	5	6	7
Probability	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
Cumulative	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{4}{6}$	$\frac{5}{6}$	$\frac{6}{6}$

Table 5.78: **Probabilities associated to the Wilcoxon's rank sum test**

This table being constructed, suppose that we observe for the sum of rank of treatment variables: $W_S = 7$. The threshold of a one-sided test would then give in conformity with the table above:

$$P(W_S \geq 7) = \frac{1}{6} \cong 17\% \qquad (5.7.2419)$$

or if we get $W_S = 3$:

$$P(W_S \geq 3) = \frac{5}{6} \cong 83\% \qquad (5.7.2420)$$

So we would reject the null hypothesis of an identical distribution between the two samples at any upper limit (or lower, respectively) fixed in advance by laboratory policy ... in unilateral or bilateral test (reason why some statistical softwares give unilateral test values + bilateral test values at the same time).

Two very important things you need to notice for what will follow is that:

1. First in the construction of the above table (where we take again the first part here):

Value of W_S	3	4	5	6	7
Probability	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

   we see there is a symmetry at the value 5, which means that the law W_S is symmetrical in this particular case. But if we take another example with two samples respectively including two control variables and three of treatments (two random variables) we would have:

Value of W_S	6	7	8	9	10	11	12
Probability	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$

   the reader can check that whatever the number of samples and the number of variables and control treatment, the probability table above is always balanced (well there is a mathematical proof of this but I find it inelegant). But in fact it is pretty intuitive, like combinations $C_{n_x}^{n_x+n_y}$ are independent of the fact that the ranks are sorted in ascending or descending order, it is therefore necessary that there is a symmetry.

2. Secondly the values of the measured variables do not come into account in this parametric statistics but only the tabulated values of the ranks with their associated probabilities. Indeed, as you may have noticed, we did not need explicit values of the random variables to build the table above!

Now, knowing that the law W_S is symmetrical and discreet we would like to calculate its expected mean.

The smallest possible value of W_S is assuming it is in the sample X (computer algorithms automatically determine in which sample but anyway in practice, samples almost always have the same size):

$$W_{S,\min}=1+\ldots+n_x=\frac{1}{2}n_x(n_x+1) \tag{5.7.2421}$$

The largest possible value is naturally (remember that $N=n_x+n_y$):

$$W_{S,\max}=\frac{1}{2}N(N+1)-\frac{1}{2}n_y(n_y+1) \tag{5.7.2422}$$

The expected mean of the sum of the ranks of one of the two samples is then:

$$\begin{aligned}
\mathrm{E}(W_S)&=W_{S,\min}+\frac{W_{S,\max}-W_{S,\min}}{2}=W_{S,\max}-\frac{W_{S,\max}-W_{S,\min}}{2}\\
&=\frac{1}{2}n_x(n_x+1)+\frac{\frac{1}{2}N(N+1)-\frac{1}{2}n_y(n_y+1)-\frac{1}{2}n_x(n_x+1)}{2}\\
&=\frac{2n_x(n_x+1)+N(N+1)-n_y(n_y+1)-n_x(n_x+1)}{4}\\
&=\frac{n_x(n_x+1)+N(N+1)-n_y(n_y+1)}{4}=\frac{n_x^2+n_x+N^2+N-n_y^2-n_y}{4}\\
&=\frac{n_x^2+n_x+N^2+N-n_y^2-(N-n_x)}{4}=\frac{n_x^2+2n_x+N^2-(N-n_x)^2}{4}\\
&=\frac{n_x^2+2n_x+N^2-(N^2-2Nn_x+n_x^2)}{4}=\frac{2n_x+2Nn_x}{4}=\frac{n_x+Nn_x}{2}\\
&=\frac{1}{2}n_x(N+1)
\end{aligned} \tag{5.7.2423}$$

Then finally:

$$\mathrm{E}(W_S)=\frac{1}{2}n_x(N+1) \tag{5.7.2424}$$

To calculate the variance, which will be useful to us to make if necessary an approximation that we will see further below, appears (unfortunately) the covariance because knowledge of the ranks give partial information about other ranks. So we have:

$$\mathrm{V}(W_S)=\mathrm{V}\left(\sum_{i=1}^{n_x}R_i\right)=\sum_{i=1}^{n_x}\mathrm{V}(R_i)+\sum_i\sum_{i\neq j}\mathrm{cov}(R_i,R_j) \tag{5.7.2425}$$

We already know with what we have just proved above that:

$$\sum_{i=1}^{n_x}\mathrm{V}(R_i)=\sum_{i=1}^{n_x}\frac{N^2-1}{12}=\frac{n_x(N^2-1)}{12}=\frac{n_x\left((n_x+n_y)^2-1\right)}{12} \tag{5.7.2426}$$

The problem remains the term with covariance. For its calculation there are rigorous techniques into several pages and... a tip that is much shorter. The trick is to use the global ranking variable which we denote T_i with $i=1\ldots n+m$. As the sum of the T_i is a constant, then we have:

$$\begin{aligned}\mathrm{V}\left(\sum_{i=1}^{n_x+n_y} T_i\right)&=\mathrm{V}\left(\frac{N(N+1)}{2}\right)=\mathrm{V}\left(\frac{(n_x+n_y)(n_x+n_y+1)}{2}\right)\\&=0=\sum_{i=1}^{n_x+n_y}\mathrm{V}(T_i)+\sum_{i=1}^{n_x+n_y}\sum_{i=1,j\neq i}^{n_x+n_y}\mathrm{cov}(T_i,T_j)\\&=(n_x+n_y)\frac{(n_x+n_y)^2-1}{12}+(n_x+n_y)(n_x+n_y-1)\mathrm{cov}(T_i,T_j)\end{aligned}\tag{5.7.2427}$$

It comes then:

$$\mathrm{cov}(T_i,T_j)=-\frac{(n_x+n_y)^2-1}{12(n_x+n_y-1)}=-\frac{n_x+n_y+1}{12}\tag{5.7.2428}$$

We can then take back the initial calculation by replacing the covariances by their expression, the last relation obtained for the covariances calculated on the T_i applying also (which is not necessarily intuitive ... but the trick works) to the R_i:

$$\begin{aligned}\mathrm{V}\left(\sum_{i=1}^{n_x} R_i\right)&=\frac{n_x((n_x+n_y)^2-1)}{12}+n_x(n_x-1)\left(-\frac{n_x+n_y+1}{12}\right)\\&=\frac{1}{12}\left[n_x((n_x+n_y)^2-1)-n_x(n_x-1)(n_x+n_y+1)\right]\\&=\frac{1}{12}\left[n_x(n_x^2+2n_xn_y+n_y^2-1)-(n_x^2-n_x)(n_x+n_y+1)\right]\\&=\frac{1}{12}\left[n_x^3+2n_x^2n_y+n_xn_y^2-n_x-n_x^2(n_x+n_y+1)+n_x(n_x+n_y+1)\right]\\&=\frac{1}{12}\left[n_x^3+2n_x^2n_y+n_xn_y^2-n_x-n_x^3-n_x^2n_y-n_x^2+n_x^2+n_xn_y+n_x\right]\\&=\frac{1}{12}\left(n_x^2n_y+n_xn_y^2+n_xn_y\right)=\frac{n_xn_y(n_x+n_y+1)}{12}\end{aligned}\tag{5.7.2429}$$

Finally:

$$\mathrm{V}\left(\sum_{i=1}^{n_x} R_i\right)=\frac{n_xn_y(n_x+n_y+1)}{12}\tag{5.7.2430}$$

This is the same result as the rigorous methodology that can be found in some rare references.

✍**Example:**

Let us turn to a practical case for the exact case. So consider 2 samples having 2 treatments variables (X) and 2 control variables (Y) (it is a bit simplistic and absurd as an example but it facilitates the educational aspect ...) we have:

$$\begin{gathered}X:5.7,3.2\\Y:8.1,5.5\end{gathered}$$

Thus (the treatment variable therefore has the ranks 1 and 3 which makes a sum of rank of 4):

$$\underline{3.2}, 5.5, \underline{5.7}, 8.1$$

Therefore:

$$\begin{gathered}X: R_1 = 3.2, R_3 = 5.7\\ Y: R_2 = 5.5, R_4 = 8.1\end{gathered}$$

We have the following table as we have shown above:

Value of W_S	3	4	5	6	7
Probability	$\dfrac{1}{6}$	$\dfrac{1}{6}$	$\dfrac{2}{6}$	$\dfrac{1}{6}$	$\dfrac{1}{6}$

with in this case:

$$W_S = 1+3 = 4 \tag{5.7.2431}$$

$$\text{E}(W_S) = \frac{1}{2}n_x(N+1) = \frac{1}{2}2(4+1) = 5 \tag{5.7.2432}$$

If we choose the traditional bilateral confidence threshold level to 5% we have according to the table above that:

$$\begin{gathered}P(W_S \geq 7) = \frac{1}{6} \cong 17\%\\ P(W_S \leq 3) = \frac{1}{6} \cong 17\%\end{gathered} \tag{5.7.2433}$$

So in other words we see that there is:

$$100\% - 2\frac{1}{6} = 66.\bar{6}\% \tag{5.7.2434}$$

of cumulative probability that W_S is between 3 and 7 (the bar above the 6 means for reminder that this digit is repeated to infinity). So 4 is necessarily included in the bilateral range of 95% ... and we do not the reject the null hypothesis as what the two samples are not different. The corresponding p-value in bilateral is then the half of $33.\bar{3}\%$.

Remark

In fact if we wanted to make an interesting computational example by playing with a bilateral threshold of 5% (or 2.5% on either side) we should have at least 2 samples with 4 random variables, that is to say 70 possible combinations of ranks. Below 4 random variables per sample, it is clear that the two-tailed test at a level of 95% will be such that we will almost never reject the null hypothesis of equality...

If the size of the two samples is large enough (most practitioners consider that each sample must be at least 20 individuals), it has been shown by simulations that we can make the following approximation (used by many statistical software):

$$\mathcal{N}(0,1) \cong \frac{W_S - \frac{1}{2}n_x(N+1)}{\sqrt{\frac{n_x n_y(N+1)}{12}}} \tag{5.7.2435}$$

obviously always determining the p-value in bilateral. With the previous example (with only 4 measurement in total), we have therefore:

$$\mathcal{N}(0,1) \cong \frac{W_S - \frac{1}{2}n_x(N+1)}{\sqrt{\frac{n_x n_y(N+1)}{12}}} = \frac{4-5}{\sqrt{1.291}} \cong -0.7746 \tag{5.7.2436}$$

Which corresponds to a cumulative probability of 21.93%. So the corresponding bilateral p-value is about $(1-22\%)/2 \cong 34\%$ (compared to the value of about 33% with the exact case).

5.7.14.1.5 Mann-Whitney Rank Sum Test

The "Mann-Whitney rank sum test" is also a nonparametric adequation test (see [369]) which can be deduced from the Wilcoxon rank sum test (hence based on the same assumption of identically distributed samples with same variances). Furthermore it is inspired to the point that we sometimes name it in the industry "Wilcoxon-Mann-Whitney test" or "Wilcoxon-Mann-Whitney adequation test" or "MWW test" (without specifying each time in the name that it is based on the sum of ranks).

The purpose of this test, as for the Wilcoxon rank sum test, is to find a way to verify that two independent samples not necessarily of the same size are from the same law or not (verbatim come from a same population or not) but with a different approach!

Remark

Just as the Wilcoxon rank sum test, the Mann-Whitney rank sum test may well be used for ordinal variables (categorical but so long as they are in an acceptable number).

Some softwares also generate confusion because they propose the Wilcoxon rank sum test under the name of Mann-Whitney test... and vice-versa ... and most do not indicate or do not offer always the choice between the exact or approximate version... And furthermore the Wilcoxon rank sum test and that of the Signed rank test that we will see further below are not differentiated .... so be careful! This is typically a problem whose source is the absence of an ISO standard to define the terminology and options that should be available in scientific softwares...

To see what is this test, let us build the rank table using two samples including two control variables and three measurement variables, then we have:

Possible ranks (Controls)	Possible ranks (Treatments)	W_S
$1,2$	$3,4,5$	12
$1,3$	$2,4,5$	11
$1,4$	$2,3,5$	10
$1,5$	$2,3,4$	9
$2,3$	$1,4,5$	10
$2,4$	$1,3,5$	9
$2,5$	$1,3,4$	8
$3,4$	$1,2,5$	8
$3,5$	$1,2,4$	7
$4,5$	$1,2,3$	6

Table 5.79: **Representation of ranks 2 control variables and 3 treatment variables**

from which we deduce the following table:

Value of W_S	6	7	8	9	10	11	12
Probability	$\dfrac{1}{10}$	$\dfrac{1}{10}$	$\dfrac{2}{10}$	$\dfrac{2}{10}$	$\dfrac{2}{10}$	$\dfrac{1}{10}$	$\dfrac{1}{10}$

Now imagine that we have another experience to be analysed using two samples having 3 control variables and 2 treatment variables (therefore the symmetric or the previous case!!!), then we have:

Possible ranks (Controls)	Possible ranks (Treatments)	W_S
$3,4,5$	$1,2$	3
$2,4,5$	$1,3$	4
$2,3,5$	$1,4$	5
$2,3,4$	$1,5$	6
$1,4,5$	$2,3$	5
$1,3,5$	$2,4$	6
$1,3,4$	$2,5$	7
$1,2,5$	$3,4$	7
$1,2,4$	$3,5$	8
$1,2,3$	$4,5$	9

Table 5.80: **Representation of ranks 3 control variables and 2 treatment variables**

from which we deduce the following table (the reader will notice that this is exactly the same table as the previous regarding the probabilities!!):

Value of W_S	3	4	5	6	7	8	9
Probability	$\dfrac{1}{10}$	$\dfrac{1}{10}$	$\dfrac{2}{10}$	$\dfrac{2}{10}$	$\dfrac{2}{10}$	$\dfrac{1}{10}$	$\dfrac{1}{10}$

Well the idea of the Mann-Whitney rank sum test is quite simple. Indeed, rather than tabulate symmetric situations, it is enough to simply subtract at each value W_S, the value $W_{S,\min}$ so that each table is identical and only one of the two is helpful. Let us see this with the first table:

Possible ranks (Controls)	**Possible ranks (Treatments)**	$W_s - \frac{1}{2}n_x(n_x+1)$
$1,2$	$3,4,5$	6
$1,3$	$2,4,5$	5
$1,4$	$2,3,5$	4
$1,5$	$2,3,4$	3
$2,3$	$1,4,5$	4
$2,4$	$1,3,5$	3
$2,5$	$1,3,4$	2
$3,4$	$1,2,5$	2
$3,5$	$1,2,4$	1
$4,5$	$1,2,3$	0

Table 5.81: **Symmetrization (2 control/3 treatments)**

We deduce the following table:

Value of W_{XY}	0	1	2	3	4	5	6
Probability	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$

Now imagine that we have another experiment to analyse using two sample having 3 control variables and 2 treatments variables, we have therefore using the same idea:

Possible ranks (Controls)	**Possible ranks (Treatments)**	$W_s - \frac{1}{2}n_x(n_x+1)$
$3,4,5$	$1,2$	0
$2,4,5$	$1,3$	1
$2,3,5$	$1,4$	2
$2,3,4$	$1,5$	3
$1,4,5$	$2,3$	2
$1,3,5$	$2,4$	3
$1,3,4$	$2,5$	4
$1,2,5$	$3,4$	4
$1,2,4$	$3,5$	5
$1,2,3$	$4,5$	6

Table 5.82: **Symmetrization (3 control/2 treatments)**

From which we deduce this time again exactly the same table as before:

Value of W_{YX}	0	1	2	3	4	5	6
Probability	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$

this is why the literature mention that we can take whatever we want!

So to summarize, the Mann-Whitney variant (in the specific case here it is the variant named "exact Mann-Whitney variant") consists to tabulate for symmetrical situations a variable denoted by W_{XY} naturally defined by:

$$\boxed{W_{XY}=W_S-\frac{1}{2}n_x(n_x+1)} \tag{5.7.2437}$$

Denoted also often in the literature by:

$$U_1=R_1-\frac{1}{2}n_1(n_1+1) \tag{5.7.2438}$$

since $W_{XY}\in\{0,1,2,\ldots\}$ and therefore:

$$P(W_{XY}=k|n_x,n_y)=P(W_{XY}=k|n_y,n_x) \tag{5.7.2439}$$

In the tables that we can find in books, the probabilities are given with the normalized value of U. Thus, if we take our previous example, but with the usual notations in practice (U instead of W_{YX}):

Value of U	0	1	2	3	4	5	6
Probability	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$

We see that the cumulative probability that $U=2$ is of 0.4. The above table is sometimes given in literature as follows:

Factor $n_2=3$			
U/n_1	1	2	3
0	0.250	0.100	0.050
1	0.500	0.200	0.100
2	0.750	**0.400**	0.200
3	1	0.600	0.350
4		0.800	0.500
5		0.900	0.650
6		1	...

Table 5.83: **Classic representation of the Mann-Whitney test**

where we have put in bold the value corresponding to our example ($U=2,n_1=2,n_2=3$).

Then the practitioner has to choose what he wants to do with these tables if he wishes to make a bilateral or unilateral test.

Remarks

R1. It is important to remember that we have showed with an example that we can also take:

$$W_{XY}=W_S-\frac{1}{2}n_x(n_x+1) \quad (5.7.2440)$$

than:

$$W_{YY}=W_S-\frac{1}{2}n_y(n_y+1) \quad (5.7.2441)$$

as they generate the same tables!

R2. W_{XY} is traditionally denoted U by practitioners as we already have seen it, this is why in some books this test is given under the name "Mann-Whitney U test" with the associated tables under the same name. But take care not to make confusion with the "Wilcoxon U test" sometimes named "Wilcoxon inversion text" which is base on the alternation of sample values when grouped (this test will not be developed in this textbook).

To see the approximate version (asymptotic) of the Mann-Whitney U test we need the expression of the expected mean and variance. For this, remember that we have seen that the sum of normalized ranks was given by:

$$U_1=R_1-\frac{1}{2}n_1(n_1+1) \quad (5.7.2442)$$

But we can also use as we saw:

$$U_2=R_2-\frac{1}{2}n_2(n_2+1) \quad (5.7.2443)$$

and as:

$$R_1+R_2=\frac{N(N+1)}{2} \quad (5.7.2444)$$

with for reminder:

$$N=n_1+n_2 \quad (5.7.2445)$$

We therefore have:

$$\begin{aligned}U_1+U_2&=\left(R_1-\frac{n_1(n_1+1)}{2}\right)+\left(R_2-\frac{n_2(n_2+1)}{2}\right)\\&=\frac{N(N+1)}{2}-\frac{n_1(n_1+1)}{2}-\frac{n_2(n_2+1)}{2}=\frac{(n_1+n_2)(N+1)}{2}-\frac{n_1(n_1+1)}{2}-\frac{n_2(n_2+1)}{2}\\&=\frac{n_1(n_1+n_2+1)}{2}+\frac{n_2(n_1+n_2+1)}{2}-\frac{n_1(n_1+1)}{2}-\frac{n_2(n_2+1)}{2}\\&=\frac{n_1n_1+n_1n_2+n_1+n_2n_1+n_2n_2+n_2-n_1n_1-n_1-n_2n_2-n_2}{2}=n_1n_2\end{aligned} \quad (5.7.2446)$$

The average of the both U is therefore the arithmetic average of the sum. Hence:

$$m_U = \frac{n_1 n_2}{2} \tag{5.7.2447}$$

This means that U_1 or U_2 must be different enough of the latter average so that we reject the null hypothesis H_0 that suppose (for recall) that both samples are issued from the same distribution law. But to determine the p-value, we also need the standard deviation. So let us search it!

The standard deviation is the same a for the Wilcoxon rank sum test (as the second term in the expression of U is a constant and therefore is variance is equal to 0). Therefore, it only remains the variance of the sum of ranks that we have already proved earlier as having for value:

$$\sigma_U = \sqrt{\frac{n_1 n_2 (N+1)}{12}} \tag{5.7.2448}$$

✍**Example:**

Let us take the same example as done with the Wilcoxon rank sum test but slightly modified (for the example to be more easy to understand) that is to say:

$$X : 5.7, 3.2$$
$$Y : 8.1, 5.5, 1.2$$

That is to say grouped and sorted:

$$\underline{1.2}, 3.2, \underline{5.5}, 5.7, \underline{8.1}$$

We then have:

$$\begin{aligned} U_1 &= (2+4) - \frac{1}{2}2(2+1) = 3 \\ U_2 &= (1+3+5) - \frac{1}{2}3(3+1) = 3 \end{aligned} \tag{5.7.2449}$$

So we can choose anyone for the test since both U are equal. If we look at the table above, with $(U=3, n_1=2, n_2=3)$, we have therefore a cumulative probability of 60% that U is equal to 3. So we do not reject the null hypothesis H_0 (in unilateral) that the two samples are from the same distribution law.

The approximation by a Normal distribution gives then:

$$Z = \frac{U - m_U}{\sigma_U} = \frac{3 - \dfrac{n_1 n_2}{2}}{\sqrt{\dfrac{n_1 n_2 (N+1)}{12}}} = \frac{3 - \dfrac{3 \cdot 2}{2}}{\sqrt{\dfrac{3 \cdot 2(5+1)}{12}}} = 0 \tag{5.7.2450}$$

Then the cumulative probability is 50% with the Normal approximation which corresponds to a p-value of 50% in unilateral. Again we do not also not reject the null hypothesis H_0 here.

5.7.14.1.5.1 Treatment of equal ranks in Ranks based tests

When we do a Wilcoxon-Mann-Whitney rank sum test Mann-Whitney or other, equality of ranks may occur.

As theoretical introduction let us take an example:

$$N=n_x+n_y=2+3=5 \tag{5.7.2451}$$

with the following data:

Data	17	17	17	19	21
Rank	?	?	?	4	5

Table 5.84: **Example of problem in case of equalities**

A conventional solution (among others ...) is to assign to each "?" the average rank. So in this case we have:

$$?=\frac{1}{3}(1+2+3)=2 \tag{5.7.2452}$$

The table:

Possible ranks (Controls)	**Possible ranks (Treatments)**	W_S
$1,2$	$3,4,5$	12
$1,3$	$2,4,5$	11
$1,4$	$2,3,5$	10
$1,5$	$2,3,4$	9
$2,3$	$1,4,5$	10
$2,4$	$1,3,5$	9
$2,5$	$1,3,4$	8
$3,4$	$1,2,5$	8
$3,5$	$1,2,4$	7
$4,5$	$1,2,3$	6

Table 5.85: **Representation of ranks 2 control variables and 3 treatment variables**

becomes in this case:

Possible ranks (Controls)	**Possible ranks (Treatments)**	W_S^*
$2,2$	$2,4,5$	11
$2,2$	$2,4,5$	11
$2,4$	$2,2,5$	9
$2,5$	$2,2,4$	8
$2,2$	$2,4,5$	11
$2,4$	$2,2,5$	9
$2,5$	$2,2,4$	8
$2,4$	$2,2,5$	9
$2,5$	$2,2,4$	8
$4,5$	$2,2,2$	6

Table 5.86: **Representation of ranks 2 control variables and 3 treatment variables in case of equalities**

where W_S^* (notice the small upper right *!) is the Wilcoxon statistics when we are in the presence of statistical equalities. The law W_S^* can be more or less different from W_S. Indeed:

Wilcoxon statistics	6	7	8	9	10	11	12
Probability of W_S	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$
Probability of W_S^*	$\frac{1}{10}$	0	$\frac{3}{10}$	$\frac{3}{10}$	0	$\frac{3}{10}$	0

Caution! From the mathematical derivation we have seen above, we can claim that at the opposite of the urban legend, the Mann-Whitney test compares the mean ranks - it does not compare medians and does not compare distributions (i.e. stochastic dominance/stochastic superiority)!!! More generally, this test answers this question: What is the probability that a randomly selected value from the population with the larger mean rank is greater than a randomly selected value from the other population? However if the compared distributions are independent and identically distributed and symmetric, then YES it's equivalent as comparing obviously two medians! It is possible, although not very common, for groups to have different rank sums and yet have equal or nearly equal medians (hence the Mann-Whitney test will be significant even if the medians are equals).

5.7.14.1.6 One sample Wilcoxon rank sum signed Test

The purpose of the "Wilcoxon signed rank sum test" also sometimes named "Wilcoxon median test", is to use a nonparametric technique for checking the symmetry or not of a distribution, and therefore make verbatim a hypothesis on the value of the median. The idea is both simple and subtle.

The principle is that if we compare the differences denoted D_i between individuals of a sample relatively to the median, we know that if we have (for example) an odd number of individuals all different (not equal), then we have 50% of the data above and below the median. Then, to control that distribution of the values satisfies a certain symmetry, the idea (simple but clever) then consists in:

1. Calculate the differences $|D_i|$ in absolute values relatively to the median.
2. Order these absolute differences by ascending order and assign them a respective rank.
3. Calculate the sum of the ranks of the differences D_i that are negative.
4. Calculate the sum of the ranks of the differences D_i that are positive.

and if the sample has a symmetric unimodal distribution (so the median is then confused with the average), there should be a sum of negative ranks S_- that is not statistically significantly different from the sum of positive ranks S^+.

We notice therefore that this hypothesis test for it to work is that the statistical distribution is symmetrical and unimodal!!

Remark

For recall, during our study of Wilcoxon or Mann-Whitney tests independent samples seen above (samples which do not necessarily have the same size), we ordered all the values of the two samples and we make a calculation on the ranks of these values. In tests for paired samples (i.e. of the same size), we ordered the differences values (not the original values!) and works on the ranks of the differences!!!

According to the idea (principle) described above, the sum of the rank bearing the sign $-$ has then for average:

$$\mu_-=\bar{R}_-=\mathrm{E}(R_-)=\mathrm{E}(1X_1+2X_2+\ldots+nX_n)=\sum_{i=1}^{n} i\mathrm{E}(X_n) \tag{5.7.2453}$$

But we have already shown that the expected mean of the binomial distribution is:

$$\mathrm{E}(X_n)=Np \tag{5.7.2454}$$

And as in our case N is equal to 1 (only one value ...) and p is equal to $1/2$ (one in two chance of having a negative sign), it comes immediately using the proofs of the section Sequences and Series:

$$\mu_-=\bar{R}_-=\mathrm{E}(R_-)=\frac{1}{2}\sum_{i=1}^{n} i=\frac{1}{2}\frac{n(n+1)}{2}=\frac{n(n+1)}{4} \tag{5.7.2455}$$

and for the variance using also the results of the section Sequences and Series:

$$\mathrm{V}(R_-)=\mathrm{V}\left(\sum_{i=1}^{n} iX_i\right)=\sum_{i=1}^{n}\mathrm{V}(iX_i)=\sum_{i=1}^{n} i^2\mathrm{V}(X_i) \tag{5.7.2456}$$

and again using the variance of the binomial distribution and the results of section Sequences and Series:

$$\mathrm{V}(R_-)=\sum_{i=1}^{n} i^2\mathrm{V}(X_i)=\sum_{i=1}^{n} i^2\frac{1}{4}=\frac{1}{4}\sum_{i=1}^{n} i^2=\frac{1}{4}\frac{n(n+1)(2n+1)}{6}=\frac{n(n+1)(2n+1)}{24} \tag{5.7.2457}$$

Obviously the sum of ranks of negative differences (respectively positive) will be at the minimum equal to zero and at maximum $n(n+1)/2$. Therefore, the expected mean in the case of a bilateral test should not be too close to one of these two extreme values.

In the case where n is large enough (more than thirty), we can use the approximation of the reduced centered Normal distribution for the variable:

$$\mathcal{N}(0,1)=\frac{S_- - \dfrac{n(n+1)}{4}}{\sqrt{\dfrac{n(n+1)(2n+1)}{24}}} \tag{5.7.2458}$$

where S_- is the sum of ranks of negative differences.

Finally let us notice that empirically if some differences from the median are zero, they will not be included in the ranks. If differences are equal we will take an average rank...

✍**Example:**

Let us start with the case with one sample compared to its experimental median (at the opposite with the comparison to a hypothesized a priori median when we consider the distribution symmetric and unimodal). Consider that we measured the following values for the diameter of a piece:

$$39, 20.2, 40, 32.2, 30.5, 26.5, 42.1, 45.6, 42.1, 45.6, 42.1, 29.9, 40.9$$

We wish to know if the calculated experimental median (which value is equal to 40 in this case) of this sample can not be rejected as a main indicator to threshold level of 5% in bilateral (which will be the case if the number of positive and negative differences is fairly balanced). We then construct the following table:

Measurement	**Differences**	**Absolute value**	**Rank**	R_+	R_-
39	-1	1	2		2
20.2	-19.8	19.8	11		11
40	0	0			
32.2	-7.8	7.8	6		6
30.5	-9.5	9.5	8		9
26.5	-13.5	13.5	10		10
42.1	2.1	2.1	4		4
45.6	5.6	5.6	6.5		6.5
42.1	2.1	2.1	4	4	
45.6	5.6	5.6	6.5	6.5	
42.1	2.1	2.1	4	4	
45.6	5.6	5.6	6.5	6.5	
42.1	2.1	2.1	4	4	
29.9	-10.1	10.1	9		9
40.9	0.9	0.9	1	1	
		Sum:		26	46

At first glance the equality of ranks does not look great but nevermind... we will going a little bit further...

Remark

According to some textbooks the sum of ranks does not give the same value as there are several techniques to calculate the ranks of values that are the equals... However, we have chosen the one used by Minitab software that is customary in the scientific community and which corresponds to that which we have already presented the rules earlier.

If we consider that the number of individuals is sufficient ... we use the approximation (even if in this case the conditions are not satisfied):

$$\mathcal{N}(0,1)=\frac{S_{+,-}-\dfrac{n(n+1)}{4}}{\sqrt{\dfrac{n(n+1)(2n+1)}{24}}} \qquad (5.7.2459)$$

Thus with this example:

$$\mathcal{N}(0,1)=\frac{S_{+}-\dfrac{n(n+1)}{4}}{\sqrt{\dfrac{n(n+1)(2n+1)}{24}}}\cong\frac{26-33}{12.747}\cong -0.549 \qquad (5.7.2460)$$

and respectively:

$$\mathcal{N}(0,1)=\frac{S_- - \dfrac{n(n+1)}{4}}{\sqrt{\dfrac{n(n+1)(2n+1)}{24}}}\cong\frac{46-33}{12.747}\cong 1.02 \quad (5.7.2461)$$

The first case corresponds with the approximation using a Normal distribution to a cumulative probability of 29.13% obtained with the English versions of Microsoft® Excel® 14.0.6123 using the function:

`=NORM.S.DIST(-0.549,TRUE)`

and thus obviously a bilateral p-value of about $2\cdot 29.13\%\cong 58.26\%$.

The second case corresponds with the approximation using a Normal distribution to a cumulative probability of 84.62% obtained with the English versions of Microsoft® Excel® 14.0.6123 using the function:

`=NORM.S.DIST(1.02,TRUE)`

which corresponds to a bilateral p-value of about $2\cdot(1-0.8462)/2=30.76\%$. Minitab gives a bilateral p-value of 32% as it does not use the Normal approximation.

Therefore for the both calculations we can prudently (and sadly) do not reject the null hypothesis H_0 as 40 (the median) is in the middle of the confidence interval (also the sign test leads to the same conclusion).

Remark

A software like Minitab even if offering the Wilcoxon median test for 1 sample gives for the median a value of 36.5 and gives for the median confidence interval the values 31.1 and 42.1. If we apply the bootstrapping method presented in detail in section of Theoretical Computing we get for estimated median 40 (and an average of 38.733) and as interval 30.50 and 42.10... Well in any case we anyway not to reject the null hypothesis H_0 but it is still boring not to know how these values are calculated in Minitab...

5.7.14.1.7 Wilcoxon rank sum signed test for two paired samples

The "Wilcoxon rank sum signed test for two paired samples" or simply "Wilcoxon signed rank test" is 100% based on the principle of the test sample with one sample. The only difference is that the null or alternative hypothesis are based on the difference in the median of the data taken in pairs (two by two) of each of the samples. In most cases, the null hypothesis H_0 is that the median of the differences is zero against the alternative hypothesis H_A that it is statistically significantly different from zero.

Like the T-test for paired samples, the Wilcoxon rank sum test applies to two-sample designs involving repeated measures, matched pairs, or "before" and "after" measures!!

As the mathematical developments are the same as for the one sample test let us attack it directly by an example.

First let us just emphasize that by extension, that an hypothesis for that test to work is that the statistical distribution of differences is therefore symmetrical and unimodal!!

Many introductory texts motivate the signed rank test as a test of median difference, or more rarely in our experience, mean difference, without mentioning that two fairly strict assumptions are required for this interpretation:

- The distribution of both groups must have the same shape.
- The variance of both groups must be equal.

If both these assumptions are true, then the signed rank test can validly be interpreted as having a null hypothesis of equal medians (or equal means).

Remark

Notice that in many statistical softwares, R_+ (see below) is denoted by the letter V.

✍**Example:**

We have two different softwares $(S1, S2)$ to compare and we want to submit to 12 tasks $(T1, T2, T3, \ldots, T12)$ of specific but similar calculations for each of the software. We would like to know if the software have a processing time that is statistically significantly different or not and if so which one is better.

We then have the following table where the time is in minutes and where the differences $x_i - y_i$ are denoted d_i:

Task	$S1$	$S2$	d_i	$\lvert d_i\rvert$	**Rank**	R_+	R_-
$T1$	24.0	23.1	0.9	0.9	1	1	
$T2$	16.7	20.4	−3.7	3.7	4		4
$T3$	21.6	17.7	3.9	3.9	5	5	
$T4$	23.7	20.7	3.0	3.0	2.5	2.5	
$T5$	37.5	42.1	−4.6	4.6	6		6
$T6$	31.4	36.1	−4.7	4.7	7		7
$T7$	14.9	21.8	−6.9	6.9	10		10
$T8$	37.3	40.3	−3.0	3.0	2.5		2.5
$T9$	17.9	26.0	−8.1	8.1	11		11
$T10$	15.5	15.5	0.0	0.0	–		
$T11$	29.0	35.4	−6.4	6.4	9		9
$T12$	19.9	25.5	−5.6	5.6	8		8
					Sum:	8.5	57.5

We already see that the software $S1$ is generally faster than $S2$ without using any tables of the Wilcoxon's exact sign test, we can therefore already say that qualitatively the difference is statistically significant.

If we consider that the number of individuals (sample sizes) is sufficient ... we use the Normal approximation (even in this special example the conditions are not satisfied):

$$\mathcal{N}(0,1)=\frac{S_{+,-}-\dfrac{n(n+1)}{4}}{\sqrt{\dfrac{n(n+1)(2n+1)}{24}}} \tag{5.7.2462}$$

That is to say in this example:

$$\mathcal{N}(0,1)=\frac{S_{+}-\dfrac{n(n+1)}{4}}{\sqrt{\dfrac{n(n+1)(2n+1)}{24}}}\cong\frac{8.5-39}{12.747}\cong-2.392 \tag{5.7.2463}$$

and respectively:

$$\mathcal{N}(0,1)=\frac{S_{-}-\dfrac{n(n+1)}{4}}{\sqrt{\dfrac{n(n+1)(2n+1)}{24}}}\cong\frac{57.5-39}{12.747}\cong 1.453 \tag{5.7.2464}$$

The first case corresponds in approximation to a Normal distribution with a cumulative probability of 0.836% obtained with the English version of Microsoft® Excel® 14.0.6123 using the function:

```
=NORM.S.DIST(-2.392,TRUE)
```

and thus a bilateral p-value of $2\cdot 0.836\%\cong 1.68\%$.

The second case corresponds in approximation to a Normal distribution with a cumulative probability of 92.668% obtained with the English version of Microsoft® Excel® 14.0.6123 using the function:

```
=NORM.S.DIST(1.453,TRUE)
```

and thus a bilateral p-value of $2 \cdot (1-0.92774)\% \cong 14.452\%$. With a software like Minitab 15.1.2 which does not offer the Wilcoxon test for paired samples but for which there is a trick to run it anyway, we get a p -value of 3.3% (same result as with the statistical software `R` which, on the other hand, has this Wilcoxon paired test implemented!). Other softwares gives a p-value always less than 5% (but the values differ from one software to another...).

Therefore, with manual calculations and using the Normal approximation, we would reject the null hypothesis H_0 at the threshold of 5% (if we only focus on the first result). It is the same with `R` and Minitab softwares, we would also reject the null hypothesis H_0 at the threshold of 5%!

Remark

As far as we know there are at least two different ways of calculating the variance of the Wilcoxon rank sum signed test for paired samples (and this is without taking into account the continuity correction which adds a $+0.5$ to the numerator). This is why most of the times results differs between statistical softwares.

5.7.14.1.8 Kruskal-Wallis Test

The Kruskal-Wallis test is a nonparametric test often compared (a bit faster ...) to a nonparametric one-way canonical ANOVA with fixed factor to compare if two or more populations have the same median (null hypothesis H_0) except that it does not requires the assumptions necessary for the parametric version of the ANOVA. When several compared populations pass through this test, the KW-test does not say which population is statistically significantly different but only that there is at least one who is. In fact, as we will show, the KW-test is only an extension of the Mann-Whitney U test seen earlier for a number of populations greater than or equal to three.

To study this test, we will assume that we have only two populations and we will afterwards make an intuitive generalization. This approach is that one that would have used Wilcoxon before that Kruskal and Wallis make the rigorous proof of the generalized case.

> Caution! The Mann-Whitney test and Kruskal-Wall tests fail most of the times as tests of medians. Indeed, that's because as already mentioned earlier above, they test something more complex which reduces to median only under STRONG assumption! Remember that they test "stochastic equality" $H_0: F_1(x)=F_2(x)$ (null hypothesis) vs "stochastic superiority" $H_1: F_1(x)<F_2(x)$ (alternative hypothesis). Their result is expressed in pseudo-median of differences if and only if both samples have same variance and same shape (skewness). This reduces to medians if the group-distributions are also symmetric.

To study this test, let us first recall that (relations whose origin and verbatim the proof have already been explained during our study of the Mann-Whitney test seen above) the average of the sum of ranks and the variance of the sum of ranks is given by:

$$\mu_R=\bar{R}=\frac{N+1}{2} \tag{5.7.2465}$$

$$\text{V}(R)=\sigma_R^2=\frac{N^2-1}{12} \tag{5.7.2466}$$

in the case where there is no duplicate values. Under this assumption, remember that $\bar{R}$ can be assimilated to the rank of the median value (in the case of an odd number of values).

Let us recall that the average values of n draws without replacement among N will be close to a Normal distribution when N big enough, and we have already proved at the beginning of this section that:

$$\sigma_{\bar{X}}=\frac{\sigma}{\sqrt{n}} \tag{5.7.2467}$$

and if the population is not very large, the variance must be corrected by the correction factor on finite population that we have already prove earlier:

$$\text{fpc}=\sqrt{\frac{N-n}{N-1}} \tag{5.7.2468}$$

Therefore it comes:

$$\sigma_{\bar{X}}=\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\Leftrightarrow\sigma_{\bar{X}}^2=\frac{\sigma^2}{n}\frac{N-n}{N-1} \tag{5.7.2469}$$

We then have if we are concerned with the ranks (the variance of the ranks being the true variance: there is no estimator!):

$$\sigma_{\bar{R}}^2=\frac{\sigma_R^2}{n}\frac{N-n}{N-1}=\frac{\left(\dfrac{N^2-1}{12}\right)}{n}\frac{N-n}{N-1}=\frac{(N-1)(N+1)}{12n}\frac{N-n}{N-1}=\frac{(N+1)(N-n)}{12n} \tag{5.7.2470}$$

Now, to form a reduced centered Normal variable Z we can center and reduce the random variable $\hat{\bar{R}}$ obtained by sampling by writing:

$$Z\cong\frac{\hat{\bar{R}}-\bar{R}}{\sigma_{\bar{R}}}=\frac{\hat{\bar{R}}-\dfrac{N+1}{2}}{\sqrt{\dfrac{(N+1)(N-n)}{12n}}} \tag{5.7.2471}$$

where $\hat{\bar{R}}$ is the average of the sum of the ranks of a sample of the population. And in fact all the trick of the idea behind the Kruskal-Wallis test is here: the statistical distribution of the average of the sum of the ranks of a large number of samples of N values has approximately a Normal distribution (read again if necessary our study of limits of drawings samples of a population at the beginning of this section)!

Let us take the square:

$$Z^2\cong\frac{12}{(N+1)(N-n)}n\left(\hat{\bar{R}}-\frac{N+1}{2}\right)\cong\chi^2(1) \tag{5.7.2472}$$

The approximation by the chi-square law is only valid if n is large enough as we have already discussed in detail during our study of the test chi-square adjustment.

And so the parenthesis of the first equality is equal to the square of the deviation of the rank to the median. This is why we often say that this is a median test (but that is an abusive shortcut).

Before continuing, let us insist on the fact that the scenario in which we find ourselves is that of a random sample of n items from N, which is equivalent to end up with two samples (one of size n and the other of size $N-n$) of the same law (verbatim from the same population/distribution). It then comes that (relations that we will use later):

$$n_i+n_j=n_1+n_2=n+(N-n)=N \tag{5.7.2473}$$

and by extension of the case with one sample, if we denote by R_i the sum of ranks of the sample numbers i, we also have:

$$n_i\bar{R}_i+n_j\bar{R}_j=n_1\bar{R}_1+n_2\bar{R}_2=n\bar{R}_1+(N-n)\bar{R}_2=N\bar{R}=N\frac{N+1}{2}=\frac{1}{2}N(N+1) \tag{5.7.2474}$$

It follows that if we write for what will follow:

$$\bar{R}_1=\hat{\bar{R}}=\frac{R}{n} \tag{5.7.2475}$$

where R is the sum of the ranks of the sample $i=1$, we have:

$$\bar{R}_2=\frac{\dfrac{1}{2}N(N+1)-n\hat{\bar{R}}}{N-n}=\frac{\dfrac{1}{2}N(N+1)-R}{N-n} \tag{5.7.2476}$$

If we now write the relation proved above:

$$\frac{12}{(N+1)(N-n)}n\left(\hat{\bar{R}}-\frac{N+1}{2}\right)=\frac{12}{(N+1)(N-n)}n\left(\hat{\bar{R}}-\bar{R}\right)\cong\chi^2(1) \tag{5.7.2477}$$

as follows (this is a clever development in reverse order... starting from the third line):

$$\begin{aligned}
&\frac{12}{(N+1)(N-n)}n\left(\hat{\bar{R}}-\frac{N+1}{2}\right)=\frac{\hat{\bar{R}}^2-(N+1)\hat{\bar{R}}+\dfrac{1}{4}(N+1)^2}{\dfrac{(N+1)(N-n)}{12n}}=\frac{\dfrac{R^2}{n^2}-(N+1)\dfrac{R}{n}+\dfrac{1}{4}(N+1)^2}{\dfrac{(N+1)(N-n)}{12n}}\\
&=\frac{12\dfrac{R^2}{n}-12(N+1)R+3n(N+1)^2}{(N+1)(N-n)}=\frac{12NR^2-12Nn(N+1)R+3(N+1)^2Nn^2}{N(N+1)n(N-n)}\\
&=\frac{12NR^2-12Nn(N+1)R+3N^2(N+1)^2n}{N(N+1)n(N-n)}-\frac{3(N+1)N(N+1)n(N-n)}{N(N+1)n(N-n)}\\
&=\frac{12}{N(N+1)}\left(\frac{(N-n)R^2+\dfrac{1}{4}N^2(N+1)^2n-Nn(N+1)R+nR^2}{n(N-n)}\right)-3(N+1)\\
&=\frac{12}{N(N+1)}\left(\frac{R^2}{n}+\frac{\left(\dfrac{1}{2}N(N+1)-R\right)^2}{N-n}\right)-3(N+1)=\frac{12}{N(N+1)}\left(\frac{\bar{R}_1^2}{n_1}+\frac{\bar{R}_2^2}{n_2}\right)-3(N+1)\\
&=\frac{12}{N(N+1)}\left(\sum_{i=1}^{2}\frac{\bar{R}_i^2}{n_i}\right)-3(N+1)
\end{aligned} \tag{5.7.2478}$$

and we find ourselves at the end with the fact that we worked from the beginning with two samples, one of size n and therefore the other (verbatim by sampling) of size $N-n$.

The previous result (which was the one desired from the beginning) can be generalized as follows under the named "Kruskal-Wallis H test" to a given confidence level in unilateral (sometimes this relation is written without the parentheses around the sum which can lead to a bad reading):

$$\boxed{H=\frac{12}{N(N+1)}\left(\sum_{i=1}^{c}\frac{\bar{R}_i^2}{n_i}\right)-3(N+1)\cong\chi^2(c-1)} \tag{5.7.2479}$$

and if all the n_i are equals, we fall back on a well known notation of the previous relation:

$$H=\frac{12}{nN(N+1)}\left(\sum_{i=1}^{c}\bar{R}_i\right)-3(N+1)\cong\chi^2(c-1) \tag{5.7.2480}$$

The approximation according to a chi-square law, however, can be discussed when the sample size (c) is small (refer to our study of the Chi-square law earlier).

✍**Example:**

Let us take the original example of Kruskal-Wallis. We consider that we have three machines that are the same at the origin but two of them have undergone some modifications. We measured the daily production a number of times and got the following table:

Task	**Standard**	**Rank**	**Modified 1**	**Rank**	**Modified 2**	**Rank**
340	5	339	4	347	10	
345	9	333	2	343	7	
330	1	344	8	349	11	
342	6			355	12	
338	3					
			- - - - - - - - - - - - - - - - - - - -			
n	5		3		4	12
R	24		14		40	78
R^2/n	11.5		65.33		400	580.53

Table 5.87: **Example table for the Kruskal-Wallis test**

We then have well:

$$\begin{gathered}\sum n_i = N = 12\\ \sum R_i = \frac{1}{2}N(N+1) = 78\end{gathered} \tag{5.7.2481}$$

and:

$$H = \frac{12}{N(N+1)}\left(\sum_{i=1}^{3}\frac{\bar{R}_i^2}{n_i}\right) - 3(N+1) = \frac{12}{12\cdot 13}580.83 - 3.13 \cong \chi^2(2) \tag{5.7.2482}$$

Be we have:

$$P\left(\chi^2(2) \geq 5.656\right) \cong 0.059 \tag{5.7.2483}$$

That can be obtained easily with Microsoft® Excel® 14.0.7166 in English with the function:

```
=1-CHISQ.DIST(5.656,2,TRUE)
```

In the present case, to a threshold level of 5%, so we are at the limit with the approximation by a chi-square distribution. As shown by Kruskal and Wallis, a Monte Carlo simulation gives a p-value of 0.049.

In short, in this situation it would be better have to reject the null hypothesis as that the productions are similar. And therefore preferred the alternative hypothesis such that these are rather different. A recommendation is to redo the test by pairs to see what is statistically significantly one by one.

5.7.14.1.9 Friedman Test

The Friedman test, recommended by the norm NF ISO 8587 for sensory analysis (ranking test), consider an experiment with two factors (the first being considered as the treatment and the second as the blocks of tests as well as the ANOVA with two-fixed factor without repetition) that is analysed using the ranks, as the measurement values do not satisfy the conditions for the application of ANOVA. However, instead of ANOVA, Friedman test applies to paired data as we shall see now.

Let us associate, as we have already done it several times, the theory to an example starting from the following table where 8 subjects (blocks) denoted by B under hypnosis were subjected to 4 emotions (treatments) denoted by T. Their epidermal electrical potential was measured (millivolts) in each case (and the order of the treatments was randomized):

	Subjects (B)							
Emotions (T)	1	2	3	4	5	6	7	8
Fear	23.1	57.6	10.5	23.6	11.9	54.6	21.0	20.3
Joy	23.1	57.6	10.5	23.6	11.9	54.6	21.0	20.3
Sadness	22.5	53.7	10.8	21.1	13.7	39.2	13.7	16.3
Calm	22.6	53.1	8.3	21.6	13.3	37.0	14.8	14.8

Table 5.88: **Example table for the Friedman test**

The central and subtle idea is not to assign a rank to the entire population of the measures as is the case for the Kruskal-Wallis test (we would then lose the concept of the blocks: verbatim the second factor) but block by block all supposed therefore independent of each other.

Remark

We will not deal (as well as in our study of the Kruskal-Wallis test) the situation where some measurements are equal with some others in the same block, the existing proofs being not really convincing in my point of view.

So, to every value $\{x_{tb}\}_{T\times B}$ of the table we now associate the rank $\{r_{tb}\}_{T\times B}$ corresponding to each treatment. Which will give us:

	Subjects (B)							
Emotions (T)	1	2	3	4	5	6	7	8
Fear	4	4	3	4	1	4	4	3
Joy	3	2	2	1	4	3	1	4
Sadness	1	3	4	2	3	2	2	2
Calm	2	1	1	3	2	1	3	1

Well now that we have built such a kind of nonparametric ANOVA table with two fixed factors without repetitions what do we do? What is the idea? Well the basic idea is the same as the Kruskal-Wallis test: we use the property of the average of the sum of the ranks but while having in mind that this time that the numbering (ranking) was not made on all measures of the table but block by block.

In the context of our particular example we have therefore:

	Subjects (B)									
Emotions (T)	1	2	3	4	5	6	7	8		
Fear	4	4	3	4	1	4	4	3	$R_1 = 27$	$\bar{R}_1 = 3.375$
Joy	3	2	2	1	4	3	1	4	$R_2 = 20$	$\bar{R}_2 = 2.50$
Sadness	1	3	4	2	3	2	2	2	$R_3 = 19$	$\bar{R}_3 = 2.375$
Calm	2	1	1	3	2	1	3	1	$R_4 = 14$	$\bar{R}_4 = 1.75$

and in case of no influence of treatments, we expect to have:

$$R_1 = R_2 = R_3 = R_4 = R_t \tag{5.7.2484}$$

or alternatively (this is equivalent):

$$\bar{R}_1 = \bar{R}_2 = \bar{R}_3 = \bar{R}_4 = \bar{R}_t \tag{5.7.2485}$$

If there is no influence of the treatments this last four values should be equal and fluctuate around:

$$\mu_R = \bar{R} = \frac{T+1}{2} = \frac{4+1}{2} = 2.5 \tag{5.7.2486}$$

We can feel that the fluctuation of the $\bar{R}_t$ around μ_R must follow a Normal centered distribution if there is not any influence (there is a proof of this in the original article of Friedman but it has sometimes some unexplained gaps that refrain us to present it). We can the also reduce the Normal law such as:

$$Z \cong \frac{\bar{R}_r - \bar{R}}{\sigma_{\bar{R}}} = \frac{\bar{R}_r - \bar{R}}{\dfrac{\sigma_R}{\sqrt{B}}} = \frac{\bar{R}_r - \bar{R}}{\dfrac{\sqrt{\dfrac{T^2-1}{12}}}{\sqrt{B}}} \tag{5.7.2487}$$

It is not always intuitive that the standard error is obtained by dividing the root of B (the number of blocks) because the majority of practitioners have intuition to divide by the root of T of the number of treatments when they study the theoretical aspect of Friedman test. But this can be verified with a numerical application or by remembering that the calculation of the variance σ_R is done from the ranks of B for a given treatment, ranks which the values (in the example above these values are 8 time between 1 and 4) are obviously assumed for a given treatment to be independent and identically distributed.

So we have:

$$Z^2 = \frac{(\bar{R}_t - \bar{R})^2}{\sigma_{\bar{R}}^2} = \frac{(\bar{R}_t - \bar{R})^2}{\dfrac{\sigma_R^2}{B}} \tag{5.7.2488}$$

Unlike the Kruskal-Wallis, we don't do any sampling, so we should not correct the deviation with the correction factor on finite population (fpc) to decrease its value.

The idea of Friedmann (at least that is how we will present it) is to say that the standard deviation of the sum of ranks of treatments obtained in the same way as in the Kruskal-Wallis test (which origin has been detailed in our study of the Mann-Whitney test):

$$\hat{\sigma}_R^2=\frac{T^2-1}{12} \tag{5.7.2489}$$

is this time only an estimator of the true standard deviation and we must used the relation between the unbiased and biased estimator to correct this estimate (relation proved during our study of the estimators):

$$\hat{\sigma}_R^2=\frac{T}{T-1}\hat{\sigma}_R^2=\frac{T}{T-1}\frac{T^2-1}{12} \tag{5.7.2490}$$

Therefore:

$$\sum_{t=1}^{T} Z_t^2=\frac{\displaystyle\sum_{t=1}^{T}\left(\bar{R}_t-\frac{T+1}{2}\right)^2}{\displaystyle\frac{T}{T-1}\frac{\hat{\sigma}_R^2}{B}}=\frac{T-1}{T}\frac{\displaystyle\sum_{t=1}^{T}\left(\bar{R}_t-\frac{T+1}{2}\right)^2}{\dfrac{\frac{T^2-1}{12}}{B}} \tag{5.7.2491}$$

where we have removed a degree of freedom to the chi-square for the reason already met many times in this section.

Then after some elementary simplifications we get the "Q Friedman test" (which is therefore a nonparametric test):

$$\boxed{Q=\frac{12B}{T(T+1)}\sum_{t=1}^{T}\left(\bar{R}_t-\frac{T+1}{2}\right)^2\cong\chi^2(T-1)} \tag{5.7.2492}$$

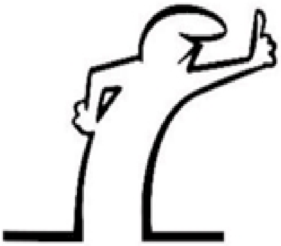
✍**Example:**

Going back to our example, then it comes:

$$\begin{aligned}&\frac{12B}{T(T+1)}\sum_{t=1}^{T}\left(\bar{R}_t-\frac{T+1}{2}\right)^2=\frac{12\cdot 8}{4(4+1)}\sum_{t=1}^{4}\left(\bar{R}_t-\frac{4+1}{2}\right)^2\\&=\frac{24}{5}\left[\left(3.375-\frac{5}{2}\right)^2+\left(2.50-\frac{5}{2}\right)^2+\left(2.375-\frac{5}{2}\right)^2+\left(1.75-\frac{5}{2}\right)^2\right]=6.45\end{aligned} \tag{5.7.2493}$$

The critical value of $\chi^2(T-1)$ at the 5% threshold is 7.81. So we sadly do not reject the null hypothesis H_0 as what treatments have no effect (no difference between the treatments). The cumulative probability corresponding to 6.45 (thus the p-value) is 90%.

And keep in mind that excellent meme by Adrian Olszewski:



- Variable transformation

Box-Cox, Logarithm, Johnson, Manly, etc.

⊳ Change a LOT of things

- Rank-based methods

Mann-Whitney, Kruskal-Wallis, Friedman, etc.

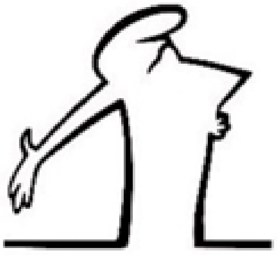
⊳ Totally change H_0; complex interpretation, may fool you about medians

---

- More general (semi-) parametric methods

GLM, GLS, [G]LMM, GEE

⊳ Retain null hypotheses, handle heteroscedasticity, dependency, non-normality



- Computer intensive testing

Permutation tests, Bootstrap & Jackknife resampling

⊳ Retain null hypotheses, handle heteroscedasticity, dependency, non-normality, partially - outliers

Or this table also by Adrian Olszewski:

Subject of comparisons H_0 **vs** H_1	**1 factor**				**2 factors + interactions**		
	2 levels		k **levels**		**of** k **levels**	+ **numerical covariates**	+ **repeated data**
	unpaired	**repeated data**	**unpaired**	**repeated data**			
stochastic equivalence vs stochastic superiority	• Mann-Whitney (Wilcoxon) • Brunner-Munzel	Signed rank (Wilcoxon)	Kruskal-Wallis	• Friedman • Skillings-Mack RM-ANOVA • ANOVA Type Statistics (ATS) • Wald Type Statistics (WTS) • Aligned Ranked Transform (ART) ANOVA	• ATS • Permuted WTS • ART ANOVA • Brunner-Dette-Munk	Ordinal Logistic Regression = Proportional-Odds Model + Wald or Likelihood Ratio Test (LRT) inference	GEE-fit Generalized Estimating Equations Ordinal Logistic Regression with appropriate correlation structure
equal means vs unequal means	• Permutation Welch T • Permutation Yuen-Welch T (outliers)		• Van der Waerden • Permutation 1-way ANOVA • GEE-fit General Linear Model	GEE-fit General Linear Model appropriate correlation structure	• GEE-General Linear Model + Wald inference • GEE-fit Ordinal Logistic Regression + extra work to estimate eCDF • Robust (huberized) Linear Mixed Model + Wald/LRT For known non-Gaussian distribution • Generalized Linear/Additive Mixed Model + Wald/LRT • GEE/fit Generalized Linear Model + Wald		
equal medians vs unequal medians	• Brown-Mood's • MW(W) under ID samples	Wilcoxon on symmetric sample	• Quantile regression (with random effects) + Wald or bootstrap inference • Khmaladze test for pure location shift				

Table 5.89: **Flexible distribution-free & robust testing common hypotheses about numerical data (author: Adrian Olszewski)**

5.7.14.1.10 Spearman's rank correlation coefficient

The "Spearman's rank correlation coefficient", denoted R_S is the correlation coefficient of the sequence $(R(i), S(i))\,, i=1\ldots n$, of the ranks naturally inspired by the linear correlation coefficient of Pearson that we saw at the beginning of this section:

$$R_S=\frac{\mathrm{E}\left[(R_i-\bar{R})(S_i\bar{S})\right]}{\sqrt{\mathrm{V}(R)\mathrm{V}(S)}}=\frac{\mathrm{E}(R\cdot S)-\mathrm{E}(R)\mathrm{E}(S)}{\sqrt{\left(\mathrm{E}(R^2)-\mathrm{E}(R)^2\right)\left(\mathrm{E}(S^2)-\mathrm{E}(S)^2\right)}} \tag{5.7.2494}$$

Let us also take an example before we tackle the theoretical aspect. Consider that the measurement of a sample of size 10 (we took the same values as that taken for the previous nonparametric rank test studies):

X	Y
5.7	8.1
3.2	5.5
8.4	3.4
4.1	7.9
6.9	4.6
5.3	1.6
1.7	8.5
3.2	7.1
2.5	8.7
7.4	5.7

with their respective ranks in accordance with the idea of Kendall (simple idea but that had to be found!):

X	$R(i)$	Y	$S(i)$
5.7	7	8.1	8
3.2	3	5.5	4
8.4	10	3.4	2
4.1	5	7.9	7
6.9	8	4.6	3
5.3	6	1.6	1
1.7	1	8.5	9
3.2	4	7.1	6
2.5	2	8.7	10
7.4	9	5.7	5

Now let us show that the above given relation is simplified drastically because the values of R, such as those of S, browse the first n integers. For this, remember that we have proved in the section of Sequences and Series, that:

$$\sum_{i=1}^{n} i=\frac{n(n+1)}{2} \tag{5.7.2495}$$

Therefore:

$$\mathrm{E}(R)=\mathrm{E}(S)=\frac{1}{n}\frac{n(n+1)}{2}=\frac{n+1}{2} \tag{5.7.2496}$$

Hence:

$$R_S=\frac{\mathrm{E}(R\cdot S)-\left(\dfrac{n+1}{2}\right)^2}{\sqrt{\left(\mathrm{E}(R^2)-\left(\dfrac{n+1}{2}\right)^2\right)\left(\mathrm{E}(S^2)-\left(\dfrac{n+1}{2}\right)^2\right)}} \tag{5.7.2497}$$

We also have proved in the section of Sequences and Series that:

$$\sum_{i=1}^{n} i^2=\frac{n(n+1)(2n+1)}{6} \tag{5.7.2498}$$

therefore:

$$\mathrm{E}(R^2)=\mathrm{E}(S^2)=\frac{1}{n}\frac{n(n+1)(2n+1)}{6}=\frac{(n+1)(2n+1)}{6} \tag{5.7.2499}$$

Therefore it comes:

$$\begin{aligned}
R_S&=\frac{\mathrm{E}(R\cdot S)-\left(\dfrac{n+1}{2}\right)^2}{\dfrac{(n+1)(2n+1)}{6}-\left(\dfrac{n+1}{2}\right)^2}=\frac{12\mathrm{E}(R\cdot S)-12\left(\dfrac{n+1}{2}\right)^2}{2(n+1)(2n+1)-3(n+1)^2}\\
&=\frac{12\mathrm{E}(R\cdot S)-3(n+1)^2}{(2n+2)(2n+1)-3(n^2+2n+1)}=\frac{12\mathrm{E}(R\cdot S)-3(n^2+2n+1)}{4n^2+2n+4n+2-3n^2-6n-3}\\
&=\frac{12\mathrm{E}(R\cdot S)-3n^2-6n-3}{n^2-1}=\frac{12\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}R_iS_i-3n^2-6n-3}{n^2-1}\\
&=\frac{12\displaystyle\sum_{i=1}^{n}R_iS_i}{n^3-n}-\frac{3n^2+6n+3}{n^2-1}=\frac{12\displaystyle\sum_{i=1}^{n}R_iS_i}{n^3-n}-\frac{3(n+1)^2}{n^2-1}=\frac{12\displaystyle\sum_{i=1}^{n}R_iS_i}{n^3-n}-\frac{3(n+1)^2}{(n+1)(n-1)}\\
&=\frac{12\displaystyle\sum_{i=1}^{n}R_iS_i}{n(n^2-1)}-\frac{3(n+1)}{n-1}=\frac{6\displaystyle\sum_{i=1}^{n}2R_iS_i}{n(n^2-1)}-\frac{3(n+1)}{n-1}
\end{aligned} \tag{5.7.2500}$$

Now let us play a little bit more to get a more simplified expression by observing that:

$$\mathrm{E}(R_i-S_i)^2=\sum_{i=1}^{n}R_i^2-2R_iS_i+S_i^2=\sum_{i=1}^{n}R_i^2-2\sum_{i=1}^{n}R_iS_i+\sum_{i=1}^{n}S_i^2 \tag{5.7.2501}$$

Then it comes:

$$2\sum_{i=1}^{n}R_iS_i=\sum_{i=1}^{n}R_i^2+\sum_{i=1}^{n}S_i^2-\sum_{i=1}^{n}(R_i-S_i)^2=\sum_{i=1}^{n}R_i^2+\sum_{i=1}^{n}S_i^2-\sum_{i=1}^{n}d_i^2 \tag{5.7.2502}$$

Then we have:

$$\begin{aligned}R_S&=\frac{6\sum_{i=1}^{n}2R_iS_i}{n(n^2-1)}-\frac{3(n+1)}{n-1}=\frac{6\left(\sum_{i=1}^{n}R_i^2+\sum_{i=1}^{n}S_i^2-\sum_{i=1}^{n}d_i^2\right)}{n(n^2-1)}\frac{3(n+1)}{n-1}\\&=\frac{6\left(\sum_{i=1}^{n}R_i^2+\sum_{i=1}^{n}S_i^2\right)}{n(n^2-1)}-\frac{3(n+1)}{n-1}-\frac{6\sum_{i=1}^{n}d_i^2}{n(n^2-1)}\end{aligned}\tag{5.7.2503}$$

But we proved that:

$$\sum_{i=1}^{n}i^2=\sum_{i=1}^{n}R_i^2=\sum_{i=1}^{n}S_i^2=\frac{n(n+1)(2n+1)}{6}\tag{5.7.2504}$$

Therefore:

$$\begin{aligned}R_S=&\frac{6\left(\sum_{i=1}^{n}R_i^2+\sum_{i=1}^{n}S_i^2\right)}{n(n^2-1)}-\frac{3(n+1)}{n-1}-\frac{6\sum_{i=1}^{n}d_i^2}{n(n^2-1)}\\&=\frac{2n(n+1)(2n+1)}{n(n^2-1)}-\frac{3(n+1)}{n-1}-\frac{6\sum_{i=1}^{n}d_i^2}{n(n^2-1)}=\frac{2(n+1)(2n+1)}{n^2-1}-\frac{3(n+1)(n+1)}{(n-1)(n+1)}-\frac{6\sum_{i=1}^{n}d_i^2}{n(n^2-1)}\\&=\frac{2(n+1)(2n+1)-3(n+1)(n+1)}{(n-1)(n+1)}-\frac{6\sum_{i=1}^{n}d_i^2}{n(n^2-1)}\\&=\frac{2(2n+1)-3(n+1)}{n-1}-\frac{6\sum_{i=1}^{n}d_i^2}{n(n^2-1)}=\frac{4n+2-3n-3}{n-1}-\frac{6\sum_{i=1}^{n}d_i^2}{n(n^2-1)}\\&=\frac{n-1}{n-1}-\frac{6\sum_{i=1}^{n}d_i^2}{n(n^2-1)}=1-\frac{6\sum_{i=1}^{n}d_i^2}{n(n^2-1)}\end{aligned}\tag{5.7.2505}$$

Thus we find the famous relation available in all (good) statistics and Monte Carlo simulation[86] textbooks:

$$\boxed{R_S=1-\frac{6\sum_{i=1}^{n}d_i^2}{n(n^2-1)}}\tag{5.7.2506}$$

The Spearman coefficient takes the same essential properties as the Pearson coefficient namely that:

$$-1\leq R_S\leq +1\tag{5.7.2507}$$

---

[86] Indeed, the Spearman coefficient is quite important for the modelling of random correlated variables as we will see it at page **??** in the section of Numerical Methods.

and is equal to 0 when the variables are correlated (dot not forget the important subtleties already mentioned in our study of the Pearson coefficient!!!).

Note that this coefficient seems to be defined only for a pair of variables (I have never seen a generalization to a multivariate case at this day).

5.7.14.1.11 Kendall's tau correlation Coefficient

The "Kendall's tau-b correlation coefficient", τ_b, is a nonparametric (rank-statistics) measure of association based on the number of concordances and discordances in paired observations.

It's used in many fields of sociology, engineering and quantitative finance (for an example in quantitative finance see the reference [385]) and often compared or communicated at the same time with the Pearson's correlation coefficient and Spearman's correlation coefficient!

Suppose two observations (X_i,Y_i) and (X_j,Y_j). They are said to be *concordant* if they are in the same order with respect to each variable. That is, if:

(1) $X_i<X_j$ and $Y_i<Y_j$, or if

(2) $X_i>X_j$ and $Y_i>Y_j$

They are *discordant* if they are in the reverse ordering for X and Y, or the values are arranged in opposite directions. That is, if:

(1) $X_i<X_j$ and $Y_i>Y_j$, or if

(2) $X_i>X_j$ and $Y_i<Y_j$

The two observations are tied if $X_i=X_j$ and/or $Y_i=Y_j$.

Remark

The Kendall's tau-c τ_c ignore ties. We won't discuss Kendall's tau-c as it's not often used anymore. In the absence of ties, both formulas yield identical results.

The total number of pairs that can be constructed for a sample size of n is:

$$N=\binom{n}{2}=\frac{1}{2}n(n-1) \qquad (5.7.2508)$$

N can be decomposed into these five quantities:

$$N=P+Q+X_0+Y_0+(XY)_0 \qquad (5.7.2509)$$

where P is the number of concordant pairs, Q is the number of discordant pairs, X_0 is the number of pairs tied only on the X variable, Y_0 is the number of pairs tied only on the Y variable, and $(XY)_0$ is the number of pairs tied on both X and Y.

The Kendall's tau-b for measuring order association between variables X and Y is given by the following formula:

$$\boxed{\tau_b=\frac{P-Q}{\sqrt{(P+Q+X_0)(P+Q+Y_0)}}} \tag{5.7.2510}$$

If there are no ties $X_0=Y_0=0$ then we fall back on a common expression that we can find in many textbooks of the Kendall's tau-b:

$$\tau_b=\frac{P-Q}{\sqrt{(P+Q)(P+Q)}}=\frac{P-Q}{P+Q}=\frac{P-Q}{\frac{1}{2}n(n-1)} \tag{5.7.2511}$$

If we use the symmetric difference distance (see first example below), that latter can be written:

$$\tau_b=\frac{P-Q}{\frac{1}{2}n(n-1)}=\frac{\frac{1}{2}n(n-1)-d_\Delta}{\frac{1}{2}n(n-1)}=1-\frac{2d_\Delta}{n(n-1)} \tag{5.7.2512}$$

This value becomes scaled and ranges between -1 (perfect negative monotonous relation among two variables: a lower score on variable X is always associated with a higher score on variable Y) and $+1$ (perfect positive monotonous relation: a lower score on variable X is always associated with a lower score on variable Y).

Unlike Spearman's rank correlation coefficient it does estimate a population variance as:

$$\tau_b \text{ is the sample estimate of } \tau_b=P[\text{concordance}]-P[\text{discordance}] \tag{5.7.2513}$$

Furthermore, if two variables are independent, $\tau_b=0$ but the reverse does not always hold: a curvilinear or other non monotonous relation may still exist.

✍**Examples:**

E1. Suppose that two experts order four wines denoted $\{a,b,c,d\}$ The first expert gives the following order: $E_1=[a,c,b,d]$, which corresponds to the following ranks $R_1=[1,3,2,4]$ and the second expert orders the wines as $E_2=[a,c,d,b]$ which corresponds to the following ranks $R_2=[1,4,2,3]$. The order given by the first expert is composed of the following 6 ordered pairs:

$$\mathrm{P}_1=\{[a,c],[a,b],[a,d],[c,b],[c,d],[b,d]\}$$

The order given by the second expert is composed of the following 6 ordered pairs:

$$\mathrm{P}_2=\{[a,c],[a,b],[a,d],[c,b],[c,d],[d,b]\}$$

The set of pairs which are in only one set of ordered pairs is:

$$\{[b,d][d,b]\}$$

which gives a value of $d_\Delta(\mathrm{P}_1,\mathrm{P}_2)=2$. With this value of the symmetric difference distance we compute the value of the Kendall's rank correlation coefficient between the order given by these two experts as:

$$\tau_b=1-\frac{2d_\Delta(\mathrm{P}_1,\mathrm{P}_2)}{n(n-1)}=1-\frac{2\cdot 2}{12}=1-\frac{1}{3}=\frac{2}{3}\cong 0.67$$

Notice that the result is equal to:

$$\tau_b = \frac{P-Q}{P+Q} = \frac{5-1}{6} = \frac{2}{3} \cong 0.67$$

E2. Two interviewers ranked 12 candidates (A through L) for a position. The results from most preferred to least preferred are:

- Interviewer 1: ABCDEFGHIJKL
- Interviewer 2: ABDCFEHGJILK

we want to calculate the τ_b.

For that purpose we make a table of rankings. The first column, *Candidate* is optional and for reference only. The rankings for *Interviewer 1* should be in ascending order (from least to greatest):

Candidate	**Interviewer 1**	**Interviewer 2**
A	1	1
B	2	2
C	3	4
D	4	3
E	5	6
F	6	5
G	7	8
H	8	7
I	9	10
J	10	9
K	11	12
L	12	11

After we count the number of concordant pairs, using the second column. Concordant pairs are how many larger ranks are below a certain rank. For example, the first rank in the second interviewer's column is a 1, so all 11 ranks below it are larger:

Candidate	Interviewer 1	Interviewer 2	Concordant	Discordant
A	1	1	11	
B	2	2		
C	3	4		
D	4	3		
E	5	6		
F	6	5		
G	7	8		
H	8	7		
I	9	10		
J	10	9		
K	11	12		
L	12	11		

However, going down the list to the third row (a rank of 4), the rank immediately below (3) is smaller, so it doesn't count for a concordant pair:

Candidate	Interviewer 1	Interviewer 2	Concordant	Discordant
A	1	1	11	
B	2	2	10	
C	3	4	8	
D	4	~~3~~		
E	5	6		
F	6	5		
G	7	8		
H	8	7		
I	9	10		
J	10	9		
K	11	12		
L	12	11		

When all concordant pairs have been counted, it looks like this:

Candidate	Interviewer 1	Interviewer 2	Concordant	Discordant
A	1	1	11	
B	2	2	10	
C	3	4	8	
D	4	4	8	
E	5	6	6	
F	6	5	6	
G	7	8	4	
H	8	7	4	
I	9	10	2	
J	10	9	2	
K	11	12	0	
L	12	11		

Now we count the number of discordant pairs and insert them into the next column. The number of discordant pairs is similar to the previous steps, only we are looking for smaller ranks, not larger ones:

Candidate	Interviewer 1	Interviewer 2	Concordant	Discordant
A	1	1	11	0
B	2	2	10	0
C	3	4	8	1
D	4	4	8	0
E	5	6	6	1
F	6	5	6	0
G	7	8	4	1
H	8	7	4	0
I	9	10	2	1
J	10	9	2	0
K	11	12	0	1
L	12	11		

And we sum these two new columns:

Candidate	Interviewer 1	Interviewer 2	Concordant	Discordant
A	1	1	11	0
B	2	2	10	0
C	3	4	8	1
D	4	4	8	0
E	5	6	6	1
F	6	5	6	0
G	7	8	4	1
H	8	7	4	0
I	9	10	2	1
J	10	9	2	0
K	11	12	0	1
L	12	11		
		Total:	61	5

And, inserting the totals into the formula we get:

$$\tau_b=\frac{P-Q}{P+Q}=\frac{61-5}{61+5}=\frac{56}{66}=0.85 \tag{5.7.2514}$$

We don't need to have two judges to compare to each other. For example in sensitivity analysis we may consider that a company wants to bring a new food product to the market. Potential customers taste 10 versions of this product which have different lipid levels. Customers then rate each version. Once aggregated across all tasters, these ratings are not studied as such but are classified so as to obtain an overall order of preferences. We want to know if this classification can be linked to the lipid level:

Lipid concentration	**Preferences**
13	10
14	7
15	9
16	8
17	5
18	6
19	4
20	3
21	1
22	2

Then we apply the same procedure as above comparing both columns and we get $\tau_b = -0.822$.

Remark

Notice that when the Spearman's rank correlation coefficient can be thought of as the regular Pearson's product moment correlation coefficient as computed from ranks, Kendall's tau rather represents a probability. Specifically, it is the difference between the probability that the observed data are in the same order for the two variables versus the probability that the observed data are in different orders for the two variables.

The Kendall's correlation coefficient depends only the order of the pairs, and it can always be computed assuming that one of the rank order serves as a reference point (e.g. with $N=4$ elements we assume arbitrarily that the first order is equal to 1234). Therefore, with two rank orders provided on n objects, there are $n!$ different possible outcomes (each corresponding to a given possible order) to consider for computing the sampling distribution of τ_b. We can compute the probability p associated with each possible value of τ_b. For example, we find that the p-value associated with a one-sided test for a value of $\tau_b = \frac{2}{3}$ is equal to:

$$p = P\left(\tau_b \geq \frac{2}{3}\right) = \frac{\text{Number of } \tau \geq \frac{2}{3}}{\text{Total number of } \tau} = \frac{4}{24} = 0.17 \tag{5.7.2515}$$

Therefore for our Wine example above we cannot reject the null hypothesis, and we cannot conclude that the expert displayed a significant disagreement in their ordering of the wines.

The computation of the sampling distribution is always theoretically possible because it is finite. But this requires computing $n!$ coefficients of correlation, and therefore it becomes practically impossible to implement these computations for even moderately large values of n. This problem, however, is not as drastic as it seems because the sampling distribution of τ_b converges towards a Normal distribution (the convergence is satisfactory for values of n larger than 10), with a mean of 0 and a variance equal to[87]:

$$\sigma_{\tau_b}^2 = \frac{2(2n+5)}{9n(n-1)} \tag{5.7.2516}$$

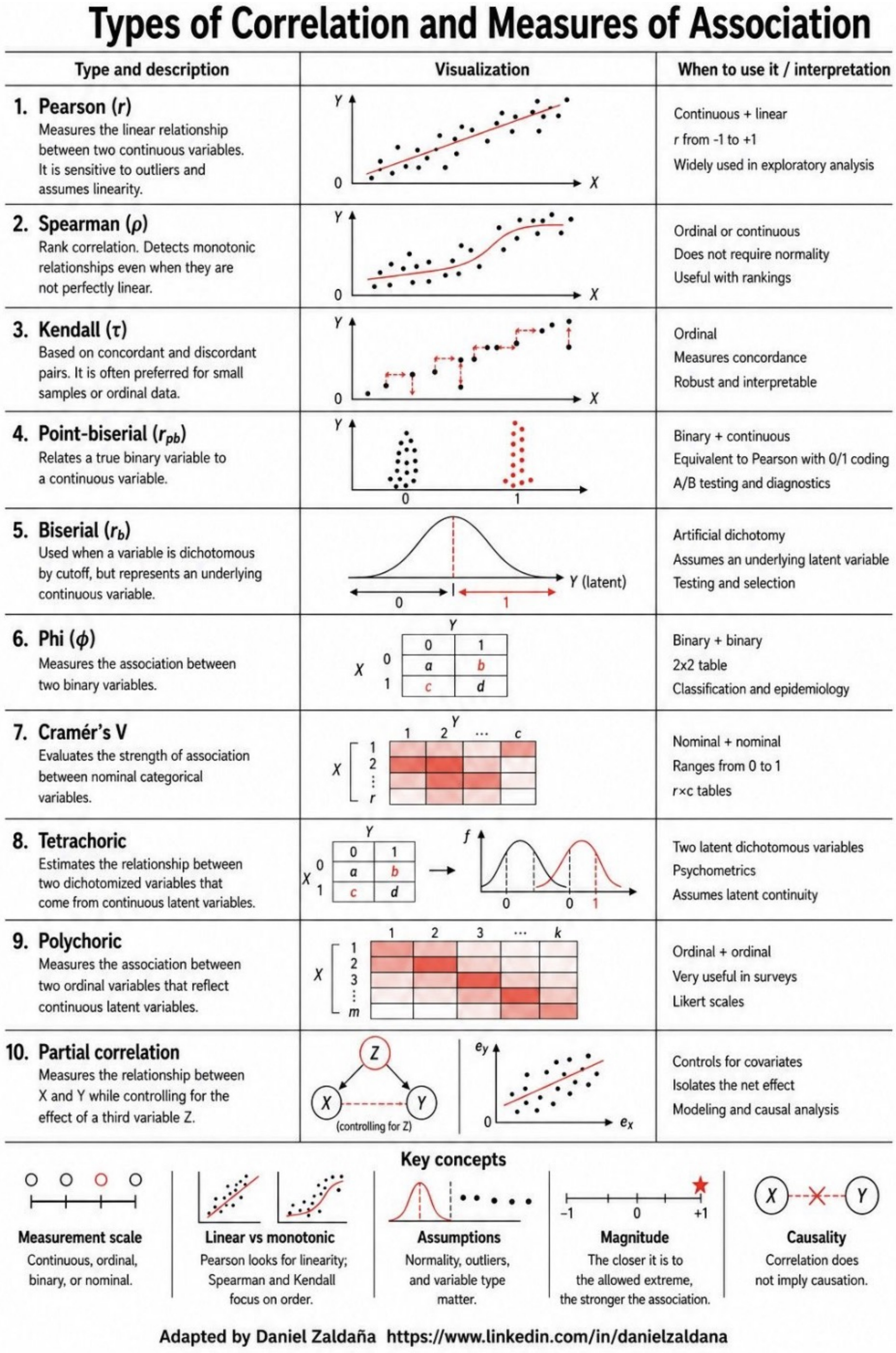
---

[87] We went as far as we could (11933 according to holocene calendar) to find the original detailed proof of this variance but we were not able to find it. Sadly we don't have the time and the fun to rewrite the proof on our own... So without proof - as always in science (!) - use that relation at your own risks!

Therefore, for n larger than 10, a null hypothesis test can be performed by transforming τ_b into a Z value as:

$$Z_{\tau_b} \cong \frac{\hat{\tau}_b}{\sigma_{\tau_b}} = \frac{\hat{\tau}_b}{\sqrt{\frac{2(2n+5)}{9n(n-1)}}} \tag{5.7.2517}$$

This Z value is Normally distributed with a mean of 0 and a standard deviation of 1.



Types of Correlation and Measures of Association

Type and description	Visualization	When to use it / interpretation
1. Pearson (r) Measures the linear relationship between two continuous variables. It is sensitive to outliers and assumes linearity.		Continuous + linear r from -1 to +1 Widely used in exploratory analysis
2. Spearman (ρ) Rank correlation. Detects monotonic relationships even when they are not perfectly linear.		Ordinal or continuous Does not require normality Useful with rankings
3. Kendall (τ) Based on concordant and discordant pairs. It is often preferred for small samples or ordinal data.		Ordinal Measures concordance Robust and interpretable
4. Point-biserial (r_{pb}) Relates a true binary variable to a continuous variable.		Binary + continuous Equivalent to Pearson with 0/1 coding A/B testing and diagnostics
5. Biserial (r_b) Used when a variable is dichotomous by cutoff, but represents an underlying continuous variable.		Artificial dichotomy Assumes an underlying latent variable Testing and selection
6. Phi (ϕ) Measures the association between two binary variables.		Binary + binary 2x2 table Classification and epidemiology
7. Cramér's V Evaluates the strength of association between nominal categorical variables.		Nominal + nominal Ranges from 0 to 1 r×c tables
8. Tetrachoric Estimates the relationship between two dichotomized variables that come from continuous latent variables.		Two latent dichotomous variables Psychometrics Assumes latent continuity
9. Polychoric Measures the association between two ordinal variables that reflect continuous latent variables.		Ordinal + ordinal Very useful in surveys Likert scales
10. Partial correlation Measures the relationship between X and Y while controlling for the effect of a third variable Z.		Controls for covariates Isolates the net effect Modeling and causal analysis

Key concepts

Measurement scale Continuous, ordinal, binary, or nominal.

Linear vs monotonic Pearson looks for linearity; Spearman and Kendall focus on order.

Assumptions Normality, outliers, and variable type matter.

Magnitude The closer it is to the allowed extreme, the stronger the association.

Causality Correlation does not imply causation.

Adapted by Daniel Zaldaña https://www.linkedin.com/in/danielzaldana

Figure 5.231: **Summary of most common coefficients of correlations and measures of association**

5.7.14.2 Range Statistics

The "range statistics" is a very important tool in finance and quality engineering (to mention only the two best-known examples). As the reader will see in what follows below, these statistics are by construction a subdomain of order statistics and the result that we will obtain here will be absolutely useful to us in the context of quality control charts (see section Engineering page 8934) and therefore also in trading signals in finance.

Given $X_1, X_2, \ldots, X_n$ supposed independent and identically distributed random variables from a law of distribution F and density f. Let us recall that we define the order statistic $X_{(i)}$ by:

$$X_{(1)} \leq \ldots \leq X_{(i)} \leq \ldots \leq X_{(n)} \tag{5.7.2518}$$

By writing:

$$\begin{aligned} W_n &= \min(X_{(1)}, X_{(2)}, \ldots, X_{(n)}) \\ M_n &= \max(X_{(1)}, X_{(2)}, \ldots, X_{(n)}) \end{aligned} \tag{5.7.2519}$$

The variables W_n and M_n define the extreme order statistics and their difference:

$$D_n = M_n - W_n \tag{5.7.2520}$$

is named the "extreme deviation".

For what will follow, we will consider as obvious the relation:

$$\min\left(X_{(1)}, X_{(2)}, \ldots, X_{(n)}\right) = \max\left(-X_{(1)}, -X_{(2)}, \ldots, -X_{(n)}\right) \tag{5.7.2521}$$

Let us now determine the distribution function of the maximum M_n:

$$F_M(x) = P\left(\{M_n \leq x\}\right) = P\left(\bigcap_{i=1}^{n} \left(X_{(i)} \leq x\right)\right) \tag{5.7.2522}$$

because writing that $M_n \leq x$ is equivalent to write that for every $X_{(i)}$ we have $X_{(i)} \leq n$ (not easy to guess that you must have this approach...).

We then have as the variables are independent (see section Probabilities page 480):

$$F_M(x) = P\left(\bigcap_{i=1}^{n} \left(X_{(i)} \leq x\right)\right) = \prod_{i=1}^{n} P(X_{(i)} \leq x) = \prod_{i=1}^{n} F(x) = [F(x)]^n \tag{5.7.2523}$$

and therefore we obviously have the distribution function:

$$f_M(x) = nf(x)\left[F(x)\right]^{n-1} = \frac{\partial}{\partial x} F_M(x) = \frac{\partial}{\partial x}\left[F(x)\right]^n \tag{5.7.2524}$$

Respectively based on the same idea:

$$\begin{aligned} F_W(x) &= P\left(\{W_n \leq x\}\right) = 1 - P\left(\{W_n > x\}\right) = 1 - P\left(\bigcap_{i=1}^{n} \left(X_{(i)} > x\right)\right) \\ &= 1 - \prod_{i=1}^{n} P\left(X_{(i)} > x\right) = 1 - \prod_{i=1}^{n} \left(1 - P\left(X_{(i)} < x\right)\right) = 1 - \prod_{i=1}^{n} (1 - F(x)) \\ &= 1 - (1 - F(x))^n \end{aligned} \tag{5.7.2525}$$

and therefore we obviously have the distribution function:

$$f_W(x)=nf(x)\left[1-F(x)\right]^{n-1}=\frac{\partial}{\partial x}F_W(x)=\frac{\partial}{\partial x}\left(1-(1-F(x))^n\right) \tag{5.7.2526}$$

Hence:

$$\begin{gathered}\mathrm{E}(R)=\mathrm{E}(M_n-W_n)=\mathrm{E}(M_n)-\mathrm{E}(W_n)=\int\limits_{-\infty}^{+\infty}xf_M(x)\mathrm{d}x-\int\limits_{-\infty}^{+\infty}xf_W(x)\mathrm{d}x\\=\int\limits_{-\infty}^{+\infty}x\left(f_M(x)-f_W(x)\right)\mathrm{d}x=\int\limits_{-\infty}^{+\infty}xf_D(x)\mathrm{d}x=\int\limits_{-\infty}^{+\infty}\frac{\partial}{\partial x}\left[[F(x)]^n-(1-[1-F(x)]^n)\right]\mathrm{d}x\end{gathered} \tag{5.7.2527}$$

having used the linearity of the expected mean and using the fact that for the two distribution functions we are working on the same random variable.

Using an integration by parts (see section Differential and Integral Calculus page 1532):

$$\begin{aligned}\mathrm{E}(R)=\bar{R}=&\int\limits_{-\infty}^{+\infty}\frac{\partial}{\partial x}\left[[F(x)]^n-(1-[1-F(x)]^n)\right]\mathrm{d}x\\&=x\left[[F(x)]^n-(1-[1-F(x)]^n)\right]\Big|_{-\infty}^{+\infty}-\int\limits_{-\infty}^{+\infty}\left(\left[[F(x)]^n-(1-[1-F(x)]^n)\right]\right)\mathrm{d}x\\&=\int\limits_{-\infty}^{+\infty}\left(1-[F(x)]^n-[1-F(x)]^n\right)\mathrm{d}x\end{aligned} \tag{5.7.2528}$$

remembering that $F(-\infty)$ and $F(+\infty)=1$.

Now let us consider the special case where the distribution function follows a Normal centered reduced distribution such as:

$$F(x)=\Phi\left(\frac{x-\mu}{\sigma}\right) \tag{5.7.2529}$$

Then we have:

$$\bar{R}=\int\limits_{-\infty}^{+\infty}\left(1-[F(x)]^n-[1-F(x)]^n\right)\mathrm{d}x=\int\limits_{-\infty}^{+\infty}\left(1-\Phi\left(\frac{x-\mu}{\sigma}\right)^n-\left[1-\Phi\left(\frac{x-\mu}{\sigma}\right)\right]^n\right)\mathrm{d}x \tag{5.7.2530}$$

Let us do a change of variable:

$$y=\frac{x-\mu}{\sigma}\Rightarrow\frac{\mathrm{d}y}{\mathrm{d}x}=\frac{1}{\sigma}\Rightarrow\mathrm{d}x=\sigma\mathrm{d}y \tag{5.7.2531}$$

Then we have:

$$\bar{R}=\sigma\int\limits_{-\infty}^{+\infty}(1-\Phi(y)^n-[1-\Phi(y)]^n)\,\mathrm{d}y \qquad (5.7.2532)$$

and then we found the relation given (almost 99% of the time without proof) in the (good) books about statistical process control:

$$d_2(n)=\int\limits_{-\infty}^{+\infty}(1-\Phi(y)^n-[1-\Phi(y)]^n)\,\mathrm{d}y \qquad (5.7.2533)$$

named "Hartley constant" (Leonard Henry Caleb Tippett was the first to develop this theory in the year 11925 according to holocene calendar) and therefore:

$$\boxed{\bar{R}=\sigma d_2(n)} \qquad (5.7.2534)$$

This constant is as far as we know not possible to calculate to calculate formally. Either we have to use Taylor series approximations of the terms of the integral, which becomes a nightmare for large n, either through a calculation using Monte Carlo simulation (see section Numerical Methods page 6708). As it is relatively long to implement in a spreadsheet, quality engineers prefer to use tables in which we find for example:

Values of n	**Values of $d_2(n)$ with a Normal distribution**
2	1.128
3	1.693
4	2.059
5	2.326
6	2.534
7	2.704
8	2.847
9	2.970
10	3.078
...	...

Table 5.90: **Tabulated values of the Hartley constant $d_2(n)$**

The previous relation is mainly know by quality engineers under the form of an estimated of the unknown process standard deviation:

$$\sigma=\frac{\bar{R}}{d_2(n)} \qquad (5.7.2535)$$

where n is a sample sub-group sample size.

Let us now see the variance of the range using always the Huygens theorem:

$$\mathrm{V}(R)=\mathrm{E}(R^2)-\mathrm{E}(R)^2=\mathrm{E}(R^2)-\sigma^2 d_2^2(n) \qquad (5.7.2536)$$

The calculation of $\mathrm{E}(R^2)$ is not nice (at least I have found nothing that satisfies the pedagogical goal of this textbook), the smallest complete proof held on three to four A4 pages and formally bring anything because we end with an integral that cannot be calculate by hand (by cons if anyone has a simple, elegant and detailed proof do not hesitate to send it to us!). It is for this reason that after having written:

$$W = \frac{R}{\sigma} \tag{5.7.2537}$$

...if we write now as do many technical books:

$$\mathrm{V}(R) = \left[\mathrm{E}(W^2) - d_2^2(n)\right]\sigma^2 \tag{5.7.2538}$$

Then we have:

$$\sigma_R = \sigma\sqrt{\mathrm{E}(W^2) - d_2^2(n)} = \sigma d_3(n) \tag{5.7.2539}$$

But as we do not know the maximum likelihood unbiased estimator of the standard deviation σ, we will use the proven relation:

$$\bar{R} = \sigma d_2(n) \tag{5.7.2540}$$

To finally have a biased estimator of the variance of the range:

$$\boxed{\hat{\sigma}_R = \frac{d_3(n)\bar{R}}{d_2(n)} = \frac{d_3\bar{R}}{d_2}} \tag{5.7.2541}$$

Here are some tabulated values of $d_3(n)$:

Values of n	**Values of $d_3(n)$ with a Normal distribution**
2	0.852
3	0.888
4	0.879
5	0.864
6	0.848
7	0.833
8	0.819
9	0.807
10	0.797
...	...

Table 5.91: **Tabulated values of the $d_3(n)$ constant**

In a typical statistical process control (SPC) application involving sample ranges, data are collected in subgroups of a constant size, say n. For each subgroup, the range is computed and plotted. The average range, $\bar{R}$, for k such subgroups is used to calculate trial control limits from which statistical control may be judged. If the process has been in a state of statistical control and is approximately normally distributed, an estimate of the process standard deviation, σ, based on $\bar{R}$ can be obtained. This estimator is $\bar{R}/d_2$. The factor d_2 depends only on the subgroup sample size n and that the process is normally distributed.

5.7.14.2.1 Tukey's Range Test

"Tukey's range test", also known as "Tukey's test", "Tukey method", "Tukey's honest significance test", or "Tukey's HSD" (honestly significant difference) test, is a single-step multiple comparison procedure and statistical test developed in 11949 (holocene). It can be used to correctly interpret the statistical significance of the difference between means that have been selected for comparison because of their extreme values.

The method was initially developed and introduced by John Tukey for use in Analysis of Variance (ANOVA), and usually has only been taught in connection with ANOVA (i.e. linear regression) under the name of "Tukey's constrast". However, the studentized range distribution used to determine the level of significance of the differences considered in Tukey's test has vastly broader application: It is useful for researchers who have searched their collected data for remarkable differences between groups, but then cannot validly determine how significant their discovered stand-out difference is using standard statistical distributions used for other conventional statistical tests, for which the data must have been selected at random. Since when stand-out data is compared it was by definition not selected at random, but rather specifically chosen because it was extreme, it needs a different, stricter interpretation provided by the likely frequency and size of the studentized range; the modern practice of "data mining" is a example where it is used.

Let us suppose that we have we have Z_k independent centered reduced random variables. And let us denote by U a random variable following a chi-square law with v degrees of freedom.

Let us now define, for reasons that will seem obvious a little further below, the "Studentized range" (the origin of the name comes from its resemblance with the definition of the Student law) by:

$$Q_{k,v}=\frac{\max(Z_i)-\min(Z_i)}{\sqrt{U/v}}=\frac{R_Z}{\sqrt{U/v}} \tag{5.7.2542}$$

and let us try to determine if this relation follows a known law and has a possible application (the find in the numerator what we defined earlier as the "extreme deviation" but with a different notation).

For this, let us show that we fall on the above definition by considering a somewhat more general case where we have X_i independent random variables that follow a Normal distribution $\mathcal{N}(\mu,\sigma)$ and the standard deviation:

$$S_*^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 \tag{5.7.2543}$$

And let us study the ratio:

$$\frac{\max(Z_i)-\min(Z_i)}{S_*}=\frac{R_X}{S_*} \tag{5.7.2544}$$

Let us now proceed to the classical transformations already seen and proved and used many times since the beginning of this section:

$$Z_i=\frac{X_i-\mu}{\sigma} \tag{5.7.2545}$$

$$U=\chi_v^2=\frac{vS_*^2}{\sigma^2} \tag{5.7.2546}$$

and therefore we have:

$$\frac{\max(Z_i)-\min(Z_i)}{S_*}=\frac{\max\left(\dfrac{X_i-\mu}{\sigma}\right)-\min\left(\dfrac{X_i-\mu}{\sigma}\right)}{\sqrt{\dfrac{U}{v\sigma^2}}}$$

$$=\frac{\max\left(\dfrac{X_i}{\sigma}\right)-\min\left(\dfrac{X_i}{\sigma}\right)}{\sqrt{\dfrac{1}{\sigma}\dfrac{U}{v}}}=\frac{1}{\sigma}\frac{\max\left(X_i\right)-\min\left(X_i\right)}{\dfrac{1}{\sigma}\sqrt{\dfrac{U}{v}}}=\frac{\max\left(X_i\right)-\min\left(X_i\right)}{\sqrt{\dfrac{U}{v}}})=Q_{k,v} \tag{5.7.2547}$$

So that's it already for the first step For the moment, even if we do not know if this definition follows a known distribution law, we can already write the following very interesting definition (notice that the term on the left is always positive):

$$\max\left(X_i\right)-\min\left(X_i\right)=Q_{k,v,1-\alpha}S_* \tag{5.7.2548}$$

or written in another way:

$$R_X=Q_{k,v,1-\alpha}S_* \tag{5.7.2549}$$

and therefore we can calculate what is the cumulative probability of a range of measurements compared to a critical range $R_{X,\text{crit}}$ directly corresponding to a prescribed threshold α. Which brings us to the possibility to write:

$$P\left(R_X\leq R_{X,\text{crit}}\right)=1-\alpha \tag{5.7.2550}$$

Let us now recall that we have seen that the distribution function of the extreme deviation was given by a relation that was to our knowledge not calculable analytically:

$$f_R(x)=\partial_x\left[[F(x)]^2-(1-[1-F(x)]^n)\right] \tag{5.7.2551}$$

So the distribution function $Q_{k,v,1-\alpha}$ can therefore not be assimilate to a known law when $F(x)$ is any law. We must therefore unfortunately tabulate this distribution using Monte Carlo simulations (see section Numerical Methods page 6708) or refer to existing tables.

Now to continue, we make a detour to the one-way fixed factor ANOVA that we studied earlier. Let us first recall that we have shown that for independent and identically Normal distributed random variables we had:

$$\sigma_{\bar{X}}=\frac{\sigma}{\sqrt{n}} \tag{5.7.2552}$$

and since a the ANOVA with fixed factor is also based on the assumption that:

$$\sigma_R^2=\sigma_A^2=\sigma_T^2=\sigma^2 \tag{5.7.2553}$$

this implies that asymptotically the estimators have the same property:

$$S_R^2=S_A^2=\sigma_T^2=S^2 \tag{5.7.2554}$$

We also know that the standard deviation of the mean of a sample of a fixed-factor ANOVA is given in the framework and assumptions of the one-way fixed factor ANOVA by:

$$\sigma_i = \frac{\sigma}{\sqrt{n}} \tag{5.7.2555}$$

But within the framework of one-way fixed factor ANOVA, we have also proved that under the assumptions imposed, we had:

$$S^2 = \text{MSE} = \frac{\sum_{ij}(x_{ij} - \bar{x}_i)^2}{N-k} \tag{5.7.2556}$$

Therefore it comes:

$$S_i = \sqrt{\frac{\text{MSE}}{n}} \tag{5.7.2557}$$

that is an estimator of:

$$\sigma_i = \frac{\sigma}{\sqrt{n}} \tag{5.7.2558}$$

And as we have proved that:

$$\frac{(N-k)\text{MSE}}{\sigma^2} = \chi^2_{N-k} \tag{5.7.2559}$$

Therefore it comes:

$$\frac{(N-k)\dfrac{\text{MSE}}{k}}{\dfrac{\sigma^2}{k}} = \chi^2_{N-k} \tag{5.7.2560}$$

Therefore, we are naturally led to conclude that the relation we defined earlier:

$$\max{(X_i)} - \min{(X_i)} \leq Q_{k,v,1-\alpha} S_* \tag{5.7.2561}$$

may be used in the study of the one-way fixed factor ANOVA under the form:

$$\max{(\bar{x}_i)} - \min{(\bar{x}_i)} \leq Q_{k,N-k,1-\alpha}\sqrt{\frac{\text{MSE}}{n}} \tag{5.7.2562}$$

to do a pre- or posterior-test (post hoc) to a one-way fixed factor ANOVA to check the hypothesis of equality of means and identify which are the aberrant means (multiple comparison test). So the Tukey's test is often accompanied by the Cochran C-test that we have already discussed earlier when we make a one-way fixed factor ANOVA.

Therefore within the ANOVA framework, we must reject the hypothesis of equality of the sample means if:

$$\boxed{\max{(\bar{x}_i)} - \min{(\bar{x}_i)} \geq Q_{k,N-k,1-\alpha}\sqrt{\frac{\text{MSE}}{n}}} \tag{5.7.2563}$$

or written in another way:

$$\max\left(\bar{x}_i\right)-\min\left(\bar{x}_i\right)\geq R_{\mu,\text{crit}} \tag{5.7.2564}$$

In this case, it is then almost immediate we can build the following confidence interval:

$$\boxed{\left(\bar{x}_i-\bar{x}_j\right)-R_{\mu,\text{crit}}\leq\mu_i-\mu_j\leq\left(\bar{x}_i-\bar{x}_j\right)+R_{\mu,\text{crit}}} \tag{5.7.2565}$$

This is the "Tukey's range test" or also simply named "Tukey's test".

It should be known now that there is a Tukey's post-hoc test for the one-way fixed factor ANOVA where the relation:

$$\max\left(\bar{x}_i\right)-\min\left(\bar{x}_i\right)\geq Q_{k,N-k,1-\alpha}\sqrt{\frac{\text{MSE}}{n}} \tag{5.7.2566}$$

will not take the two most extreme means but will compare all the means pairwise (and why not after all!) with the highest average (well we could also have fun making all possible combinations as do some statistical softwares). In this case the relation to use is the same as above except that if we have for example a one-way fixed factor ANOVA with 4 levels then we will have 3 pairwise comparisons (average differences must always be positive!!!). Thus, imagining that the third average is the largest and that in descending order the biggest averages are the 4th, 2nd and 1st (the first is therefore smaller) then it comes:

$$\begin{aligned}\bar{x}_3-\bar{x}_1&\geq Q_{4,N-4,1-\alpha}\sqrt{\frac{\text{MSE}}{n}}\\\bar{x}_3-\bar{x}_2&\geq Q_{3,N-4,1-\alpha}\sqrt{\frac{\text{MSE}}{n}}\\\bar{x}_3-\bar{x}_4&\geq Q_{2,N-4,1-\alpha}\sqrt{\frac{\text{MSE}}{n}}\end{aligned} \tag{5.7.2567}$$

This approach (to extend the basic principle of Tukey's test), is named the "Newman-Keuls test" or "Student-Newman-Keuls test" (SNK).

5.7.14.3 Extreme Value Theory

"Extreme value theory" or "extreme value analysis" (EVA) is a branch of statistics dealing with the extreme deviations from the median of probability distributions. It seeks to assess, from a given ordered sample of a given random variable, the probability of events that are more extreme than any previously observed. Extreme value analysis is widely used in many disciplines, such as structural engineering, finance, Earth sciences, traffic prediction, and geological engineering. For example, EVA might be used in the field of hydrology to estimate the probability of an unusually large flooding event, such as the 100-year flood. Similarly, for the design of a breakwater, a coastal engineer would seek to estimate the 50-year wave and design the structure accordingly.

To finish depending on donations

5.7.15 Multivariate Statistics

Obviously Multivariate statistics is a subdivision of statistics encompassing the simultaneous observation and analysis of more than one random variable.

Before we continue, remember that we have already derived two multivariate distribution earlier above! The Normal multivariate distribution (see page 720) and the Student-T canonical multivariate distribution (see page 789). We have also already introduced the multivariate ANOVA (i.e. MANOVA) earlier above at page 1091.

Remark

Certain types of problem involving multivariate data, for example simple linear regression and multiple regression that we saw earlier above, are not usually considered as special cases of multivariate statistics because the analysis is dealt with by considering the (univariate) conditional distribution of a single outcome variable given the other variables.

There are many different models, each with its own type of analysis that we will try to address in this textbook as always with a maximum of details and as accessible as possible (with softwares examples in the companion book for sure!):

1. Principal Components Analysis (PCA) that creates a new set of orthogonal variables that contain the same information as the original set. It rotates the axes of variation to give a new set of orthogonal axes, ordered so that they summarize decreasing proportions of the variation (see below page 1213).

2. Factor Analysis that is similar to PCA but allows the user to extract a specified number of synthetic variables, fewer than the original set, leaving the remaining unexplained variation as error. The extracted variables are known as latent variables or factors; each one may be supposed to account for covariation in a group of observed variables (see section of Numerical Methods page 6925).

3. Canonical Correlation Analysis (CCA) that finds linear relations among two sets of variables; it is the generalised (i.e. canonical) version of bivariate correlation.

4. Correspondence Analysis (CA), or reciprocal averaging, that finds (like PCA) a set of synthetic variables that summarise the original set. The underlying model assumes chi-squared dissimilarities among records (cases). The fundamental formal difference is that PCA decomposes relations between columns only (e.g. by decomposing their covariance matrix), treating rows as "cases"; while CA decomposes columns and rows simultaneously. Where conventional Factor Analysis determines which variables cluster together, correspondence analysis determines which category values are close together.

5. Canonical (or "constrained") Correspondence Analysis (CCA) for summarising the joint variation in two sets of variables; combination of correspondence analysis and multivariate regression analysis. The underlying model assumes chi-squared dissimilarities among records (cases).

6. Multidimensional Scaling comprises various algorithms to determine a set of synthetic

variables that best represent the pairwise distances between records. The original method is principal coordinates analysis (PCoA based on PCA).

7. Discriminant Analysis (linear or quadratic), or canonical variate analysis, attempts to establish whether a set of variables can be used to distinguish between two or more groups of cases.
8. Simultaneous Equations Models involve more than one regression equation, with different dependent variables, estimated together.
9. Multivariate Analysis of Variance (MANOVA) that extends the analysis of variance to cover cases where there is more than one dependent variable to be analyzed simultaneously (see page 1091).
10. Multivariate Analysis of Covariance (MANCOVA) that is an extension of analysis of covariance (ANCOVA) methods to cover cases where there is more than one dependent variable and where the control of concomitant continuous independent variables (covariates) is required.
11. Mixed Models that contains both fixed effects and random effects. These models are useful in a wide variety of disciplines in the physical, biological and social sciences. They are particularly useful in settings where repeated measurements are made on the same statistical units (longitudinal study).

5.7.15.1 Principal Component Analysis

Principal component analysis (P.C.A.) is a mathematical method of graphical data analysis that consist to look for directions in space that best represent the correlations between n random variables (supposed to have linear relations between them). In other terms it is a dimension reduction process (as for the singular values decomposition process - SVD - proved in the section of Linear Algebra or the Linear Discriminant Analysis proved in the section of Numerical Methods among many others techniques as visible on the Data Science mind map at page 6770) as it gives the possibility to the analyst to choose what are the variables in a model that explain the best the variability.

Simply said, a P.C.A. allows for example to find in a dataset buying behaviour similarities between observed classes.

Even if the PCA is mainly used to visualize data, we must not forget that this is also a way to:

- To decorrelate the data. In the new base, consisting in new axis, the points have a zero correlation (we will prove it).
- To classify data into correlated clusters (in the industry it is mainly this possibility that is interesting!).

Remarks

R1. There are several versions of P.C.A. known under the name of "Karhunen-Loeve transformation" or "Hotelling transformation" and that can also be applied without programming in spreadsheet softwares or in specialized one (where the computing time will be shorter by cons ... and the results most accurate too...).

R2. Depending on the authors and the point of view P.C.A. belong to the field of statistics named "Explanatory statistics".

When we consider only two effects, it is customary to characterize their joint effect via the correlation coefficient. When we stands in two dimension, the available points (the sample of points drawn following the joint distribution) can be represented in a plane. The result of the P.C.A. in this plane is the determination of the two axes that best explain the dispersion of the available points.

When there are more than two effects, for example three effects, there are three coefficients of correlations to be taken into account. The issue that gave rise to the P.C.A. is how to have a quick intuition of the joint effects?

In dimension larger than two, a P.C.A. will always determine the axes that best explain the dispersion of the cloud of available points.

The objective of the P.C.A. is to graphically describe a data array of individual with large quantitative variables:

individuals/variables	**var**$_1, \ldots,$ **var**$_j, \ldots,$ **var**$_n$
ind$_1$	
$\vdots$	
ind$_i$	x_{ij}
$\vdots$	
ind$_n$	

Table 5.92: **Generic representation of a P.C.A. table**

In order not to complicate the presentation of this method and to allow the reader to completely redo the calculations, we will work on the theory with a direct famous example.

Let us consider for example a study of a botanist who measured the dimensions of 15 iris flowers (the P.C.A. is also widely used in finance to determine the elements that most influences the volatility of a portfolio). The three variables ($p = 3$) measured are:

- x_1: Length of sepal
- x_2: Width of sepal
- x_3: Length of the petal

The data are the following:

Flower n°	x_1	x_2	x_3
1	5.1	3.5	1.4
2	4.9	3.0	1.4
3	4.7	3.2	1.3
4	4.6	3.1	1.5
5	5.0	3.6	1.4
6	7.0	3.2	4.7
7	6.4	3.2	4.5
8	6.9	3.1	4.9
9	5.5	2.3	4.0
10	6.5	2.8	4.6
11	6.3	3.3	6.0
12	5.8	2.7	5.1
13	7.1	3.0	5.9
14	6.3	2.9	5.6
15	6.5	3.0	5.8

Table 5.93: **Practical tabulated data example for P.C.A.**

For us, such a data table is simply a real components matrix with n rows (the individuals) with p columns (the variables):

$$X=(x_{ij})_{\substack{i=1\ldots n\\ \text{æ}=1\ldots p}} \tag{5.7.2568}$$

Thereafter the index i will correspond to the line index and therefore to individuals. We will identify the individual i with point line $x_{i.}=(x_{i1},\ldots,x_{ip})$ which will be considered as a point in an affine space (see section Vector Calculus page 1815) of dimension p. The index j will correspond to the index column so the variables. We will identify the variable j with the column vector:

$$\vec{x}_{.j}=\begin{pmatrix}x_{1j}\\ \vdots\\ x_{nj}\end{pmatrix} \tag{5.7.2569}$$

this is therefore a vector in the vector space of dimension n in $\mathbb{R}^n$.

We will place ourselves in what will follow in two perspectives: either we will take the table data as n points an affine space of dimension p, or we will take this table as p points in a vector space of dimension n. We will see that there are dualities between these two perspectives.

The mathematical tool that we will use here is linear algebra (see section Linear Algebra page 1924), with the concepts of dot product, Euclidean norm and Euclidean distance.

To simplify the presentation, we will initially consider that each individual as each variable has the same importance, the same weight. We also consider the case of the Euclidean distance.

We will start by centering the data, that is to say to put origin of the reference frame on the center of gravity of the point cloud. This does not change the appearance of the cloud, but allows us to get the coordinates of a point M, equal to the coordinates of the vector $\overrightarrow{GM}$, and thus to place ourselves in the vector space to be able to do the math! Since we assume throughout what will follow that the weights of the individuals are identical, we will take $m_i = 1/n$ with $i = 1 \ldots n$.

We consider the orthonormal reference frame $(\text{O}, \vec{e}_1, \vec{e}_2, \ldots, \vec{e}_p)$ in the canonical basis $(\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_p)$ of $\mathbb{R}^p$. Given G being therefore the center of gravity of the point cloud. As each variable and each individual is assumed to have the same weight, then G has for coordinates the reference frame $(\text{O}, \vec{e}_1, \vec{e}_2, \ldots, \vec{e}_p)$:

$$\overrightarrow{\text{O}G} = \begin{pmatrix} \bar{x}_{.1} \\ \vdots \\ \bar{x}_{.j} \\ \vdots \\ \bar{x}_{.p} \end{pmatrix} \tag{5.7.2570}$$

with:

$$\bar{x}_{.j} = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \tag{5.7.2571}$$

We then have yet graphically:

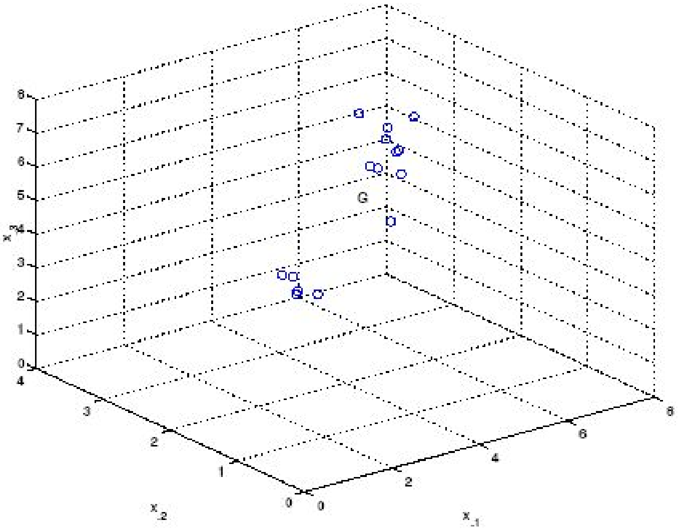


Figure 5.232: **Measurement points and center of gravity**

We name "centered matrix" the matrix:

$$X_c = \begin{bmatrix} x_{11} - \bar{x}_{.1} & \cdots & x_{1j} - \bar{x}_{.j} & \cdots & x_{1p} - \bar{x}_{.p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} - \bar{x}_{.1} & \cdots & x_{ij} - \bar{x}_{.j} & \cdots & x_{ip} - \bar{x}_{.p} \\ \vdots & & \vdots & & \vdots \\ x_{n1} - \bar{x}_{.1} & \cdots & x_{nj} - \bar{x}_{.j} & \cdots & x_{np} - \bar{x}_{.p} \end{bmatrix} \tag{5.7.2572}$$

Remark

The matrix of centered data contains the centered coordinates (which we will denote by xc_{ij}) of the individuals in the reference frame $(G,\vec{e}_1,\vec{e}_2,\ldots,\vec{e}_p)$. We place ourselves for what will follow always in this reference frame for point cloud of the individuals and we will take O $=G$.

For our example we have:

$$G=\vec{x}=(5.91,3.06,3.87) \tag{5.7.2573}$$

and for the centered matrix:

$$X_c=\begin{bmatrix} -0.8067 & 0.440 & -2.4733 \\ -1.0067 & -0.0600 & -2.4733 \\ -1.2067 & 0.1400 & -2.5733 \\ -1.3067 & 0.0400 & -2.3733 \\ -0.9067 & 0.1400 & -2.4733 \\ 1.0933 & 0.5400 & 0.8267 \\ 0.4933 & 0.1400 & 0.6257 \\ 0.933 & 0.0400 & 1.0267 \\ -0.4067 & -0.7600 & 0.1267 \\ 0.5933 & -0.2600 & 0.7267 \\ 0.3933 & 0.2400 & 2.1267 \\ -0.1067 & -0.3600 & 1.2267 \\ 1.1933 & -0.0600 & 2.0267 \\ 0.3933 & -0.1600 & 1.7267 \\ 0.5933 & -0.0600 & 1.9267 \end{bmatrix} \tag{5.7.2574}$$

and graphically:

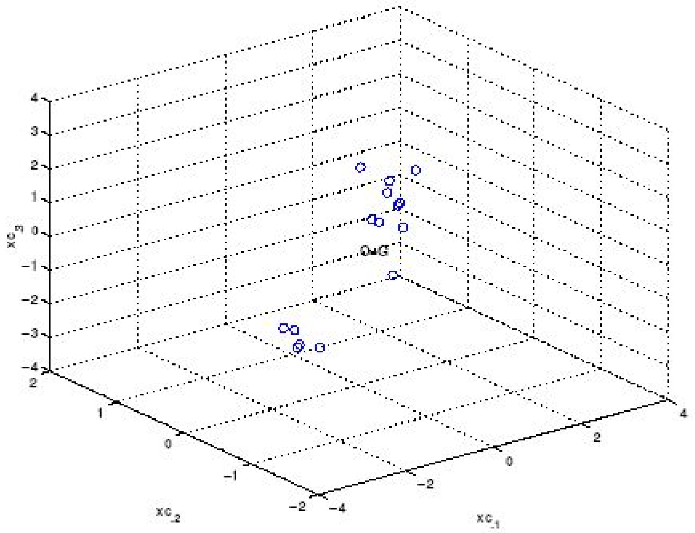


Figure 5.233: Centered measured points

To give an equal importance to each variable so that the type of units of the measurement does not influence the analysis (and also as we have proved it at page 621 that the correlation matrix is equal to the variance-covariance matrix!), we will work with reduced centered data (see section Statistics page 595). For this, we will denote first:

$$\sigma_j^2=\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\bar{x}_{.j})^2=\frac{1}{n}\sum_{i=1}^{n}c_{ij}^2=\frac{1}{n}\|\vec{c}_{.j}\|^2 \qquad (5.7.2575)$$

where the reader may have perhaps noticed that we take the biased variance. But in the reality, we will take obviously the variance estimator and we will therefore divide by $n-1$ rather than by n.

The variance of the sample centered variable is equal near to a factor of $1/n$ to the norm of this same variable but centered. The matrix of reduced centered data (dimensionless) is therefore:

$$Y=\begin{bmatrix}\dfrac{x_{11}-\bar{x}_{.1}}{\sigma_1} & \cdots & \dfrac{x_{1j}-\bar{x}_{.j}}{\sigma_j} & \cdots & \dfrac{x_{1p}-\bar{x}_{.p}}{\sigma_p}\\ \vdots & & \vdots & & \vdots\\ \dfrac{x_{i1}-\bar{x}_{.1}}{\sigma_1} & \cdots & \dfrac{x_{ij}-\bar{x}_{.j}}{\sigma_j} & \cdots & \dfrac{x_{ip}-\bar{x}_{.p}}{\sigma_p}\\ \vdots & & \vdots & & \vdots\\ \dfrac{x_{n1}-\bar{x}_{.1}}{\sigma_1} & \cdots & \dfrac{x_{nj}-\bar{x}_{.j}}{\sigma_j} & \cdots & \dfrac{x_{np}-\bar{x}_{.p}}{\sigma_p}\end{bmatrix} \qquad (5.7.2576)$$

If we denote by $D_{1/\sigma}$ the following diagonal matrix:

$$D_{1/\sigma}=\begin{bmatrix}1/\sigma_1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & 1/\sigma_p\end{bmatrix} \qquad (5.7.2577)$$

Then we have:

$$Y=X_cD_{1/\sigma} \qquad (5.7.2578)$$

Remark

Each component of the matrix Y is therefore of zero mean and unit variance (which is equivalent to say that the norm of the standard reduced variable is equal to the unit as we will prove it immediately).

We define the "matrix of centered normalized data" by (we then speak of "normalized PCA" which is not mandatory but simplifies interpretation):

$$Z=\frac{1}{\sqrt{n}}Y \qquad (5.7.2579)$$

Or also (it is simply the mean squared error that we have introduced earlier in the section):

$$z_{ij}=\frac{x_{ij}-\bar{x}_{.j}}{\sqrt{n}\sigma_j} \qquad (5.7.2580)$$

The terminology comes of course from the fact that the sum of the square of the components of each column of the matrix Z has a unit norm. Indeed:

$$\|\vec{z}_{.j}\|^2=\sum_{i=1}^{n}z_{ij}^2=\sum_{i=1}^{n}\left(\frac{x_{ij}-\bar{x}_{.j}}{\sqrt{n}\sigma_j}\right)^2=\frac{1}{\sigma_j^2}\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\bar{x}_{.j})^2=1 \qquad (5.7.2581)$$

Which gives:

$$Z=\begin{bmatrix} -0.2383 & 0.3572 & -0.3375 \\ -0.2974 & -0.0487 & -0.3375 \\ -0.3565 & 0.1136 & -0.3512 \\ -0.3861 & 0.0324 & -0.3239 \\ -0.2679 & 0.4384 & -0.3375 \\ 0.3230 & 0.1136 & 0.1128 \\ 0.1457 & 0.1136 & 0.0855 \\ 0.2935 & 0.0324 & 0.1401 \\ -0.1201 & -0.6170 & 0.0172 \\ 0.1753 & -0.2110 & 0.0991 \\ 0.1162 & 0.1948 & 0.2902 \\ -0.0315 & -0.2922 & 0.1674 \\ 0.3526 & -0.0487 & 0.2765 \\ 0.1161 & -0.1298 & 0.2356 \\ 0.1753 & -0.0487 & 0.2629 \end{bmatrix} \tag{5.7.2582}$$

We have graphically:

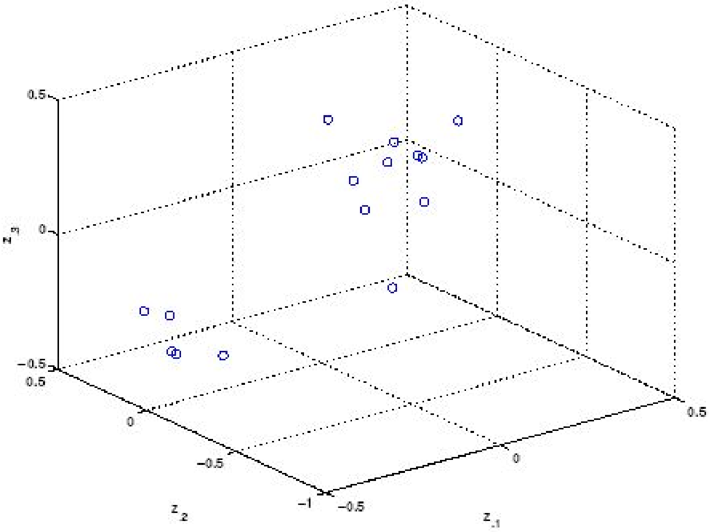


Figure 5.234: **Centered reduced measured points**

Represent the point cloud of the reduced centered data or reduced normalized data don't change the shape of it. Indeed, the difference between the two is only a change of scale.

The interesting information on individuals is the distance between the points! Indeed more this distance is great between two individuals $z_{i.}$ and $z_{i'.}$ plus the two individuals will be different and better we can characterize them. But we must first choose a distance. We will take the Euclidean distance (see section Topology page 2446):

$$d^2(z_{i.},z_{i'.})=||\overrightarrow{z_{i.}z_{i'.}}||^2=\sum_{j=1}^{p}(z_{ij}-z_{i'j})^2 \tag{5.7.2583}$$

The following figures show the orthogonal projections in space of this scatter cloud respectively in the planes $(\text{O},\vec{e}_1,\vec{e}_2)$, $(\text{O},\vec{e}_2,\vec{e}_3)$ and finally in $(\text{O},\vec{u}_1,\vec{u}_3)$ that is the best projection, named "factorial plane" (or sometimes "scores diagram"), in the sense that it best respect the distances between individuals (verbatim, it deforms the least the cloud of points in space). The objective of the principal component analysis is to determine this best plan and we will prove now how!

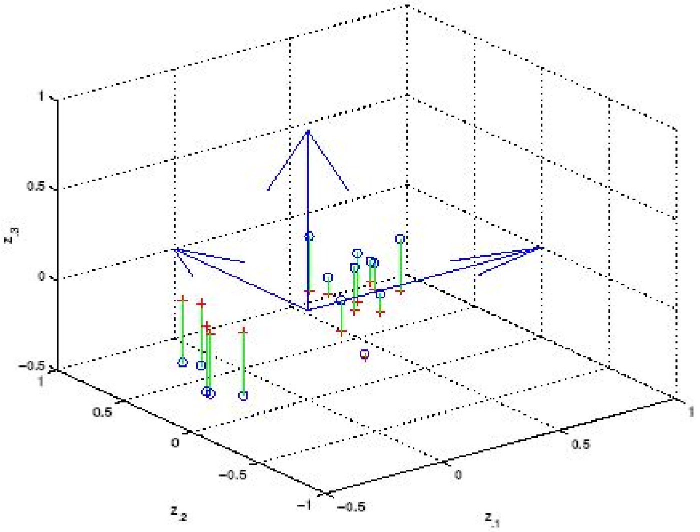


Figure 5.235: Projection of the points on the horizontal plane of the center reduced basis

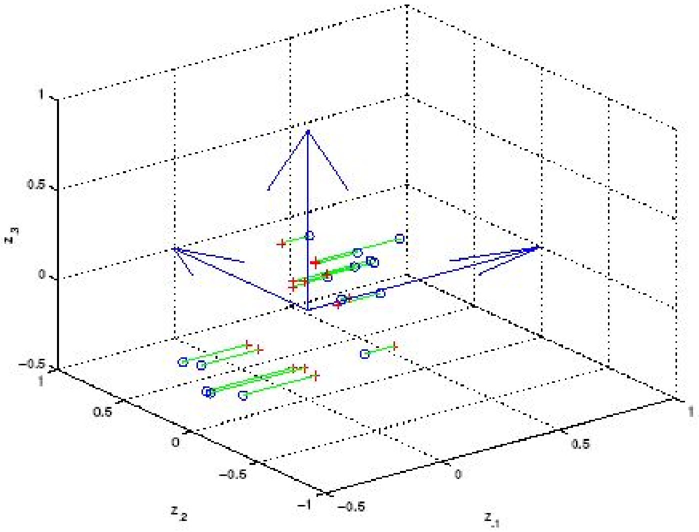


Figure 5.236: Projection of the points on the vertical plane of the center reduced basis

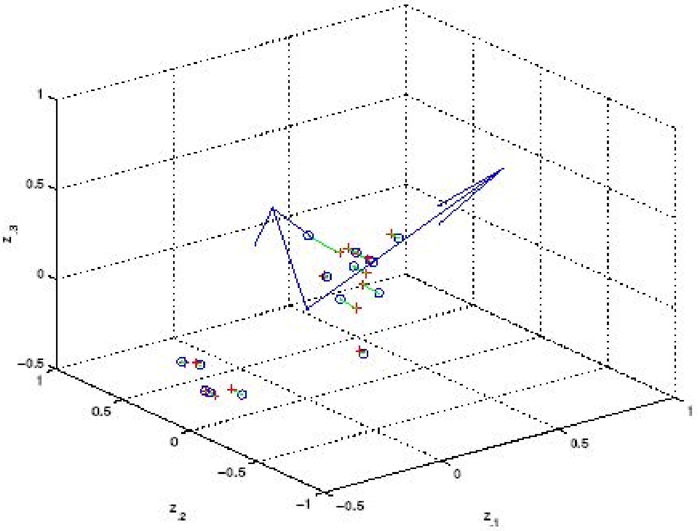


Figure 5.237: **Projection of the points on the factorial plane of the center reduced basis**

And the plane view of each of the projections:

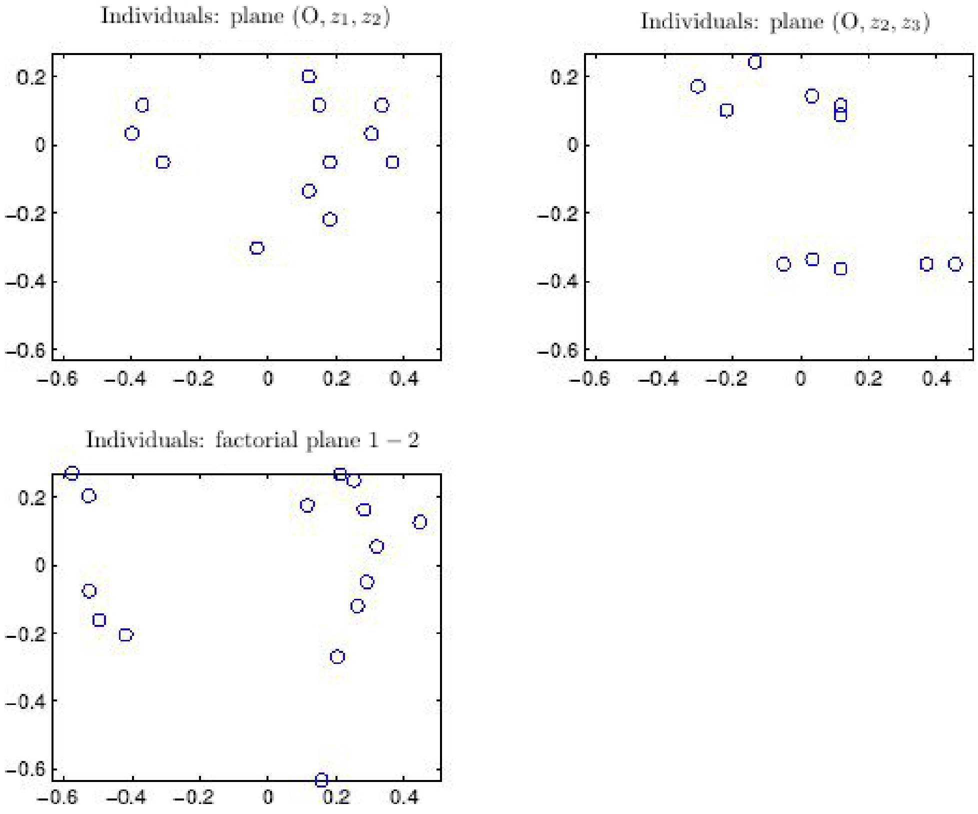


Figure 5.238: **Plane view of each of the projections**

Before determining the factorial plane, we will first now try to detect possible links between variables.

We recall (see section Statistics page 597) that the covariance between two variables $x_{.j}$ and $x_{.j'}$

is given by:

$$\text{cov}(x_j,x_{.j'})=\text{E}\left[(x_{.j}-\bar{x}_{.j})(x_{.j'}-\bar{x}_{.j'})\right]=\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\bar{x}_{.j})(x_{ij'}-.\bar{j}')=\frac{1}{n}\sum_{i=1}^{n}c_{ij}c_{ij'} \tag{5.7.2584}$$

and that the linear correlation coefficient (see section Statistics page 614) is:

$$R(x_{.j},x_{.j'})=\frac{\text{cov}(x_{.j},x_{.j'})}{\sigma_j\sigma_{j'}} \tag{5.7.2585}$$

We will denote for later:

$$c_{j,j'}=\text{cov}(x_{.j},x_{.j'}) \qquad \text{and} \qquad r_{j,j'}=R(x_{.j},x_{.j'}) \tag{5.7.2586}$$

the matrices of variance-covariance and of correlation (both being for recall square and symmetric matrices) with $j=1\ldots p, j'=1\ldots p$.

We see quite easily that the matrix of covariance is at a given coefficient $1/n$ near, the matrix of canonical dot products of vectors of the centered reduced matrix X_c (in other words, each component of the variance-covariance matrix is equal to the dot product of the centered variables). We deduce the following relation:

$$c_{j,j'}=\frac{1}{n}X_c^TX_c \tag{5.7.2587}$$

Remark

For a given matrix A, the expression A^TA plays an important role in statistics. For example, it is an important part of the analytical solution of ordinary least squares. Or, for PCA, its eigenvectors are the principal components of the data. Geometrically, matrix A^TA is named "matrix of scalar products" (as it contains all the inner products of all columns in A). Algebraically, it's named "sum-of-squares-and-cross-products matrix" (SSCP). Dividing SSCP matrix by n, the sample size or number of rows of A, we get what is commonly named the MSCP (mean-square-and-cross-product) matrix. If we center the columns of A the $A^TA/(n-1)$ is the covariance matrix. If we Z-standardized the columns of A the $A^TA/(n-1)$ is the Pearson correlation matrix.

The matrix of variance-covariance (since, as we was it in the section Statistics, the diagonal contains the variances ... for recall!) is a well known interpretation in this textbook. By cons, what is new and we will be very useful to determine the factorial plane is the matrix of linear correlations that can also be written as follows:

$$R=r_{j,j'}=\frac{1}{m}Y^TY=Z^TZ \tag{5.7.2588}$$

Which gives for our example where we have three variables (very easy to calculate using a spreadsheet software like Microsoft® Excel®), the following square matrix (the data are centered or not the components of the matrix are the same):

$$R=\begin{bmatrix}1 & -0.1609 & 0.8854\\ -0.1609 & 1 & -0.3818\\ 0.8854 & -0.3818 & 1\end{bmatrix} \tag{5.7.2589}$$

To continue, always with the aim to determine the factorial plane, let us define the concept of inertia of a point cloud.

Definition (#5.7.158): We name "inertia of a point cloud" the quantity:

$$I=\frac{1}{n}\sum_{i=1}^{n} d^2(G,M_i) \tag{5.7.2590}$$

where G is the center of gravity of the point cloud and M_i a point of $\mathbb{R}^p$ of coordinates x_i^T.

Remark

The square of the distance is taken by anticipation of the developments that will follow.

Then we prove the following relation:

$$\sum_{i=1}^{n}\sum_{i'=1}^{n} d^2(M_i,M'_i)=2nI \tag{5.7.2591}$$

Proof 5.48.8.

$$\begin{aligned}\sum_{i=1}^{n}\sum_{i'=1}^{n} d^2(M_i,M'_i)&=\sum_{i,i'}\|\overrightarrow{M_iM_{i'}}\|^2=\sum_{i,i'}\|\overrightarrow{M_iG}+\overrightarrow{GM_{i'}}\|^2\\&=\sum_{i,i'}\left(\sum_{i}\|\overrightarrow{M_iG}\|^2+\|\overrightarrow{GM_{i'}}\|^2+2\overrightarrow{M_iG}\cdot\overrightarrow{GM_{i'}}\right)\\&=\sum_{i'}\left(\sum_{i}d^2(M_i,G)\right)+\sum_{i}\left(\sum_{i'}d^2(G,M_{i'})\right)+2\sum_{i}\left(\overrightarrow{M_iG}\cdot\vec{0}\right)\\&=\sum_{i'}nI+\sum_{i}nI+1+0\\&=2nI\end{aligned} \tag{5.7.2592}$$

■ Q.E.D.

We will in all the remaining text with normalized centered data, verbatim with the matrix Z. The points M_i will therefore have here for coordinates z_i^T.

The problem now is to find the best affine space of dimension p in the sense that it respects the best the distances between points. For this, we will seek the best vectorial line $\Delta_{\vec{u}}$ that is perfectly determined by the vector $\vec{u}$. Let us denote H_i the orthogonal projection of M_i on the line Δ_u. So our problem is to find the line (verbatim the vector $\vec{u}$) that makes the sum of the squares of the distances between the points H_i is maximized. We write the problem as a quadratic programming problem (see section Numerical Methods page 6662):

$$\begin{aligned}\underset{\vec{u}}{\text{maximize}}\quad&\sum_{i,i'}d^2(H_i,H'_i)\\\text{subject to}\quad&\vec{u}\in\mathbb{R}^p\\&\|\vec{u}\|=1\end{aligned} \tag{5.7.2593}$$

But here we have:

$$\sum_{i,i'} d^2(H_i, H'_i) = 2nI \tag{5.7.2594}$$

Indeed, the projected centroid (barycentre) of the cloud points is also the origin. Consequently, our problem can be written:

$$\begin{aligned} \underset{\vec{u}}{\text{maximize}} \quad & I \\ \text{subject to} \quad & \vec{u} \in \mathbb{R}^p \\ & \|\vec{u}\| = 1 \end{aligned} \tag{5.7.2595}$$

Itself therefore being equivalent to:

$$\begin{aligned} \underset{\vec{u}}{\text{maximize}} \quad & \sum_{i,i'} d^2(\text{O}, H_i,) \\ \text{subject to} \quad & \vec{u} \in \mathbb{R}^p \\ & \|\vec{u}\| = 1 \end{aligned} \tag{5.7.2596}$$

Let us solve this problem:

First, since H_i is the orthogonal projection of the point M_i on $\Delta_{\vec{u}}$ we have $\overrightarrow{\text{O}H}_i = \alpha_i \vec{u}$ for all i with $\alpha_i = \overrightarrow{\text{O}M}_i \circ \vec{u}$. Following this, the coordinates of the points H_i on the straight line $\Delta_{\vec{u}}$ are:

$$\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = Z\vec{u} \tag{5.7.2597}$$

If follows that we have:

$$\sum_i d^2(\text{O}, H_i) = \sum_i \alpha_i^2 = ||Z\vec{u}||^2 = Z\vec{u} \circ Z\vec{u} \tag{5.7.2598}$$

Here we seek the unit vector $\vec{u}$. The matrix Z is perfectly known to us. But, we have:

$$(Z\vec{u}) \circ (Z\vec{u}) = (Z^T Z\vec{u}) \circ \vec{u} = (R\vec{u}) \circ \vec{u} \tag{5.7.2599}$$

The correlation matrix R is symmetrical therefore, following the spectral theorem proved in the section of Linear Algebra page 2004, it is diagonalizable in an orthonormal basis of eigenvectors. Thus, we have proved that in the spectral theorem:

$$\Lambda = S^{-1} R S \tag{5.7.2600}$$

is diagonal (up to us to choose the content) if R is symmetrical and S is orthogonal (which is in our special example a square 3×3 matrix). Then we deduce the following relation (factorization) that it is common to name the "eigenvalue decomposition" of R:

$$R = S^{-1} \Lambda S \tag{5.7.2601}$$

and as S has been proven as being orthogonal (and that there exists a family of eigenvectors for this!), we have (see section Linear Algebra page 1954):

$$S^{-1} = S^T \tag{5.7.2602}$$

Therefore:

$$R=S\Lambda S^T \tag{5.7.2603}$$

where we choose for Λ the diagonal matrix of the eigenvalues sorted in descending order: $\lambda_1\geq\lambda_2\geq\ldots\geq\lambda_p$.

We then have:

$$(R\vec{u})\circ\vec{u}=(S\Lambda S^T\vec{u})\circ\vec{u}=(\Lambda S^T\vec{u})\circ S^T\vec{u})=(\Lambda\vec{w})\circ\vec{w}=\sum_{j=1}^p\lambda_j w_j^2 \tag{5.7.2604}$$

In the literature, this sum is often denoted as follows (in statistical software often named "eigenvalue decomposition"):

$$\sum_{j=1}^p\lambda_j w_j^2=\begin{pmatrix}\sqrt{\lambda_1}w_1 & \sqrt{\lambda_2}w_2 & \ldots & \sqrt{\lambda_p}w_p\end{pmatrix}\begin{pmatrix}\sqrt{\lambda_1}w_1\\ \sqrt{\lambda_2}w_2\\ \vdots\\ \sqrt{\lambda_p}w_p\end{pmatrix} \tag{5.7.2605}$$

But S being orthogonal, we have therefore:

$$\left\{\vec{w}=S^T\vec{u}|\vec{u}\in\mathbb{R}^p\quad\|\vec{u}\|=1\right\}\Rightarrow\{\vec{w}\in\mathbb{R}^p\quad\|\vec{w}\|=1\} \tag{5.7.2606}$$

and this is comes from the fact that the orthogonal matrix is as we proved in the section Linear Algebra an isometry (it conserves the norm!).

Since the eigenvalues are in descending order, we will write:

$$\sum_{j=1}^p\lambda_j w_j^2=\lambda_1\left(w_1^2+\sum_{j=2}^p\frac{\lambda_j}{\lambda_1}w_j^2\right) \tag{5.7.2607}$$

Or the term in brackets is strictly less than or equal to 1 by the previous involvement. Therefore:

$$\lambda_1\left(w_1^2+\sum_{j=2}^p\frac{\lambda_j}{\lambda_1}w_j^2\right)\leq\lambda_1 \tag{5.7.2608}$$

Therefore:

$$\sum_{j=1}^p\lambda_j w_j^2\leq\lambda_1 \tag{5.7.2609}$$

But remember that our objective is to maximize this inequality. In other words to seek w_1 such that the equality is respected. We see quickly enough that will be the case if $w_1=1$ and that the other terms are zero. Thus, a trivial solution to our maximization problem is:

$$\vec{w}=\begin{pmatrix}1\\0\\0\\ \vdots\\0\end{pmatrix} \tag{5.7.2610}$$

either because:

$$\vec{w}=S^T\vec{u}=S^{-1}\vec{u}\Rightarrow=S\vec{w} \tag{5.7.2611}$$

which is then the first eigenvector of the matrix R (since R is diagonalized in this base) associated with the largest eigenvalue λ_1. Hence the fact that this solution is often denoted as:

$$\vec{u}_1=S\vec{w}_1 \tag{5.7.2612}$$

always with $\Lambda=S^{-1}RS$ (it is therefore quite easy to determine S using softwares when R and Λ are known).

Once we have found the first vector line, we seek a second orthogonal line in the subspace to the vector line that maximizes the inertia of the projected point cloud. We prove, and guess, that the solution is given by the vector line directed by the eigenvector associated to the second eigenvalue of the correlation matrix and so on ...

Thus, we obtain a new basis $(\vec{u}_1,\ldots,\vec{u}_p)$ whose one of the plane is the factorial plane. However, we need to know the components of Z in this base. As this base was built under the condition that R is diagonalizable via the matrix S then the latter matrix is the linear application that will allow us to express Z in the base $(\vec{u}_1,\ldots,\vec{u}_p)$ via the relation:

$$\Psi=ZS \tag{5.7.2613}$$

Thus, in our example the three eigenvalues of the correlation matrix R are (see section Linear Algebra page 1997):

$$\lambda_1=2.0303 \qquad \lambda_2=0.8846 \qquad \lambda_3=0.0853 \tag{5.7.2614}$$

and therefore:

$$\Lambda=\begin{pmatrix}2.0303 & 0 & 0\\ 0 & 0.8846 & 0\\ 0 & 0 & 0.0853\end{pmatrix} \tag{5.7.2615}$$

Remark

Some softwares indicate the weights in respective and cumulated $\%$ for each of the eigenvalues. Thus we have in our special case the following respective weights in $\%$ of the total:

$$P_{\lambda_1}=\dfrac{2.0303}{\sum\limits_i\lambda_i}=66.67\% \qquad P_{\lambda_2}=\dfrac{0.8846}{\sum\limits_i\lambda_i}=29.48\% \qquad P_{\lambda_3}=\dfrac{0.0853}{\sum\limits_i\lambda_i}=2.84\% \tag{5.7.2616}$$

So the first component explains 66.67% of the effect. The first two components explain 96.15%, etc. This is why, for example, that in finance that among ten or more components, we will take only those that "explain" for example the 95%.

By having the three eigenvalues, to determine the three eigenvectors $(\vec{u},\vec{u}_2,\vec{u}_3)$ that form the main base, we need to solve the following system of three equations with three unknowns (see section Linear Algebra page 1929) for each eigenvalue:

$$\underbrace{\begin{bmatrix} 1-\lambda_i & -0.1609 & 0.8854 \\ -0.1609 & 1-\lambda_i & -0.3818 \\ 0.8854 & -0.3818 & 1-\lambda_i \end{bmatrix}}_{R}\underbrace{\begin{pmatrix} u_{1,i} \\ u_{2,i} \\ u_{3,i} \end{pmatrix}}_{\vec{u}_i}=\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \tag{5.7.2617}$$

Which gives therefore (expected if someone request the details we will not provide them here as they can be simply made by hand or with any spreadsheet software) for the matrix of eigenvectors:

$$S=(\vec{u}_1,\vec{u}_2,\vec{u}_3)=\begin{bmatrix} 0.6410 & 0.3803 & -0.6666 \\ -0.3527 & 0.9174 & 0.1841 \\ 0.6816 & 0.1171 & 0.7222 \end{bmatrix} \tag{5.7.2618}$$

which satisfies therefore:

$$R=S\Lambda S^T \tag{5.7.2619}$$

or written in another way (following the remark from a reader who wanted to check the calculations and was trapped):

$$R=S(\Lambda S^T) \tag{5.7.2620}$$

We then have for coordinates of the points M_i in the base $(\vec{u}_1,\vec{u}_2,\vec{u}_3)$ using:

$$\Psi=ZS \tag{5.7.2621}$$

The following matrix:

$$\Psi=\begin{bmatrix} -0.5268 & 0.2045 & -0.0198 \\ -0.4178 & -0.2043 & -0.564 \\ -0.5259 & -0.0750 & 0.0052 \\ -0.4966 & -0.1604 & 0.0305 \\ -0.5760 & 0.2699 & 0.0161 \\ 0.2525 & 0.2488 & -0.1169 \\ 0.1156 & 0.1757 & -0.0150 \\ 0.2818 & 0.1634 & -0.0916 \\ 0.1578 & -0.6311 & -0.0218 \\ 0.2634 & -0.1194 & -0.0871 \\ 0.2108 & 0.2660 & 0.1739 \\ 0.2039 & -0.2697 & 0.0912 \\ 0.4469 & 0.1261 & -0.459 \\ 0.2908 & -0.0490 & 0.0712 \\ 0.3196 & 0.0546 & 0.0663 \end{bmatrix} \tag{5.7.2622}$$

The coordinates of the projection of the point cloud in the best plane defined by the vectors $(\vec{u}_1,\vec{u}_2)$ are then the first two columns of the previous matrix (thus corresponding the sepal length and sepals width).

Indeed, we immediately see that these are the two columns that will maximize the sum of the norms in the given plane:

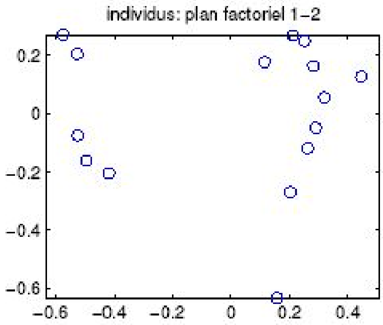
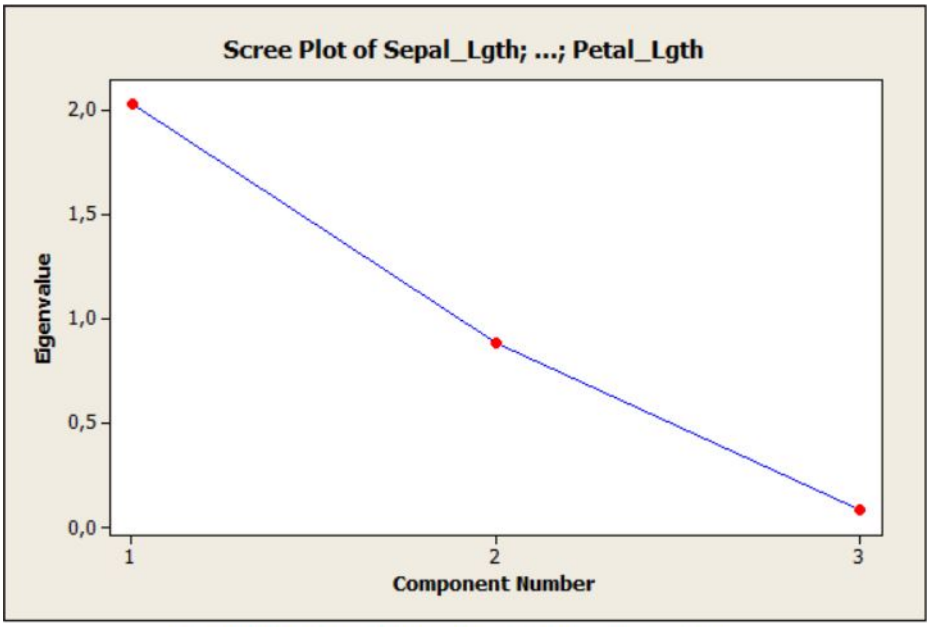


Figure 5.239: **Factorial plane already shown above ...**

A software like Minitab 15.1 (reference in quality management industry) gives the following information for the eigenvalues (not very useful info graphically ... in my opinion):



```
Eigenanalysis of the Correlation Matrix

Eigenvalue  2,0302  0,8846  0,0853
Proportion   0,677   0,295   0,028
Cumulative   0,677   0,972   1,000
```

Figure 5.240: **Eigenvalues for the PCA as given by Minitab 15.1 ("scree plot")**

and the following factorial plane (remains the question how the values are calculated in Minitab because they are not identical to those we got here my manual calculations... but the graphic is corresponding and this is the most important!):

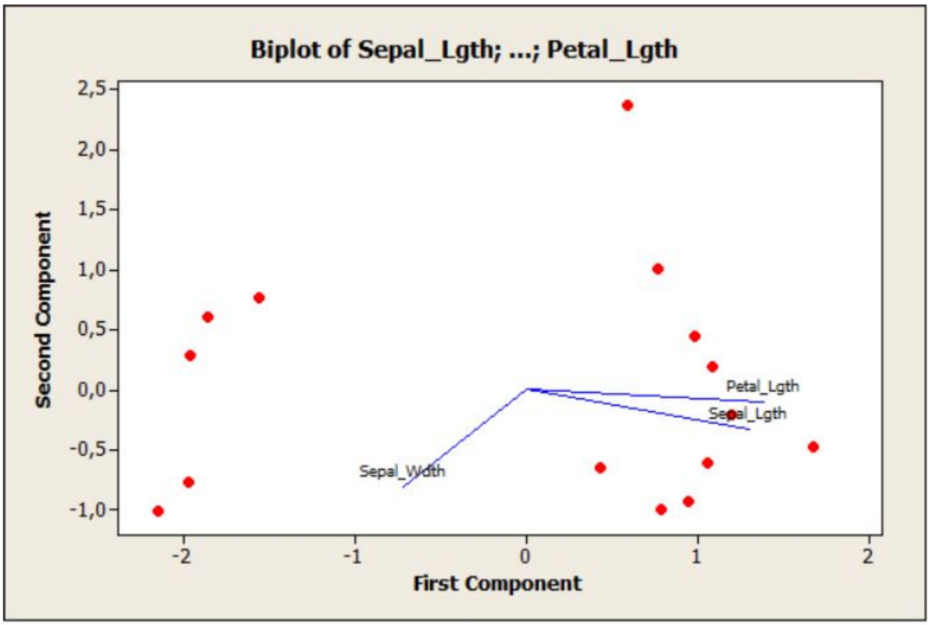


Figure 5.241: **Factorial plane as given by Minitab 15.1**

If the reader has carefully followed all the steps we did so far, he may have notice that we did:

1. Zero-center the data (mean normalization)
2. Reduced the data
3. Decorrelate the data by:
   (a) Calculating the covariance matrix
   (b) Calculating the eigenvectors of the covariance matrix
   (c) Applying the matrix of eigenvectors to the data (this will apply the rotation)

Furthermore some softwares and practitioners scale the uncorrelated matrix in order to obtain a covariance matrix corresponding to the identity matrix (ones on the diagonal and zeros on the other cells). To do that we scale the decorrelated data by dividing each dimension by the square-root of its corresponding eigenvalue.

All the 7 steps enumerated above are named "PCA whitening" or just "Whitening" (also sometimes "sphering data"). So "whitening" data means that we want to transform them in a way to have a covariance matrix that is the identity matrix (1 in the diagonal and 0 for the other cells) and is named "whitening" in reference to "white noise".

To close this subject, the reader must know that many software use the fact that the vectors $z_{.j}$ are of unit norm to do the scalar product which corresponds in this case simply the cosine between vectors such as:

$$z_{.j}\circ z_{.i}=||z_{.j}||\cdot||z_{.i}||\cos(\widehat{z_{.j}z_{.i}})=\cos(\widehat{z_{.j}z_{.i}}) \tag{5.7.2623}$$

and as we have proved earlier above that:

$$R=r_{j,j'}=\frac{1}{n}Y^TY=Z^TZ \tag{5.7.2624}$$

Therefore it comes:

$$R=r_{j,j'}=\cos(\widehat{z_{.j}z_{.i}}) \tag{5.7.2625}$$

and as in our special example we have 3 vectors $z_{.j}$, there is therefore 3 possible dot products if we omit the scalar products of the vectors with themselves. Therefore the matrix:

$$R=\begin{bmatrix} 1 & -0.1609 & 0.8854 \\ -0.1609 & 1 & -0.3818 \\ 0.8854 & -0.3818 & 1 \end{bmatrix} \tag{5.7.2626}$$

also contains the angles (in radian) between the vectors $z_{.j}$.

Finally, let us indicate that the PCA being sensitive to outliers, it is sometimes better to transform the values of the original array in their ranks (see the study of rank statistics earlier above!) and applying exactly the same algorithm or using the Spearman's rank correlation coefficient. We then speak of a "PCA by ranks".

Remark

A user-friendly software like XL-Stats proposes automatically in the PCA menu five "different types" of Principal Component Analysis that are respectively: Pearson (n), Pearson ($n-1$), Spearman, Kendall, Covariance. Obviously the same can be done with other statistical softwares but actually is not visually simple available choices (we have to write very little scripts).

There is also a group of robust PCA methods that have been developed and that are better suited for handling high-dimensional data in the situation where the sample size is lower than the dimension. These methods are: Robust PCA by projection-pursuit (PP-PCA), Spherical PCA (SPCA), Robust PCA (ROBPCA), Robust Sparse PCA (ROSPCA).

We can also compute the variance of the PCA using bootstrapping techniques (see section Numerical Methods page 6740) like for all statistical analysis!

5.7.15.1.1 SVD and PCA

Now we will prove an important theorem that is not quite obvious.

Theorem 5.49. The PCA above (using the spectral theorem, i.e. eigendecomposition) is a special case of the SVD (that latter is therefore less restrictive)!!!

Proof 5.49.1. We have proved before that:

$$R=Z^TZ \tag{5.7.2627}$$

Using the properties of the transposed matrices (see section Linear Algebra page 1951):

$$R=ZZ^T \tag{5.7.2628}$$

And that:

$$R=ZZ^T=S\Lambda S^T \tag{5.7.2629}$$

And we have build in the section of Linear Algebra the singular value decomposition (see section Linear Algebra page 2004):

$$M = UDV^T \tag{5.7.2630}$$

That is in our case here:

$$R = UDV^T \tag{5.7.2631}$$

Therefore:

$$RR^T = (UDV^T)(UDV^T)^T = (UDV^T)(VDU^T) \tag{5.7.2632}$$

and since V is an orthogonal matrix ($V^TV = \mathbb{1}$):

$$\boxed{R = UD^2U^T} \tag{5.7.2633}$$

So if we compare:

$$R = UD^2U^T \qquad \text{and} \qquad R = UDV^T \tag{5.7.2634}$$

the correspondence is quite easily seen!

So that means in order for the spectral theorem of R to be equal to its SVD, we need R to:

- Have orthonormal eigenvectors (meaning all it does is to scale its input along n orthogonal directions, where R is $n \times n$)
- Have positive eigenvalues (colloquially, it must be a real matrix and not "flip" anything)

It turns out that the above conditions are equivalent to the following:

- R must be a symmetric real matrix (this is equivalent to R having real eigenvalues and orthonormal eigenvectors)
- R's (real) eigenvalues must be positive

So, we say: Both are equal when R is symmetric positive-semidefinite (see section Linear Algebra page 1957)! This can be denoted technically as $R \succeq 0 \ \wedge \ R^\dagger = R$.

In facts, using the SVD to perform PCA makes much better sense numerically than forming the covariance matrix to begin with, since the construction of ZZ^T can cause loss of precision. Indeed in some cases the PCA can diverge very quick and give inaccurate results! This is why many statistical softwares gives the option to run a PCA based on SVD (singular value decomposition theorem) or on the spectral theorem.

■ Q.E.D.

A common question on Internet forums is: *Intuitively, what is the difference between Spectral Theorem (eigendecomposition) and Singular Value Decomposition?*. Also, if: *Is one of them more general than the other? Is either a special case of the other?*

First, according to what has been written earlier, it should be quite obvious that eigendecomposition arises from asking in which directions a quadratic form has the greatest impact, when the SVD by construction arises from asking in which directions a linear transformation[88] has the greatest impact!

Now the reader must keep in mind that the spectral theorem describes the effect of a matrix A on a vector as the following 3-steps process $A = Q\Lambda Q^{-1}$:

1. An invertible linear transformation (Q^{-1}, hence Q must be invertible, i.e. square!)
2. A scaling (Λ with $\Lambda = \text{diag}(\vec{\lambda})$)
3. The inverse of the initial transformation (Q)

This decomposition doesn't always exist, but the spectral theorem describes the conditions under which such a decomposition exists.

When for the SVD we have the following 3-steps process $A = U\Sigma V^T$:

1. A first rotation in the input space (V is a rotation matrix as $V^TV = \mathbb{1}$)
2. A simple positive scaling that takes a vector in the input space to the output space (Σ diagonal with positive entries!)
3. And another rotation in the output space (U is a rotation matrix as $U^TU = \mathbb{1}$)

The fundamental theorem of linear algebra says that such a decomposition always exists (especially because no inverse is involved and therefore doesn't apply only to square matrices).

---

[88] Hence the fact that PCA produces linear transformations, thereby capturing linear relationships between the original variables so that we might lose predictive information that are non-linear!

5.7.15.2 Chi-square Correspondence Factor Analysis (C2CFA)

The Chi-square factor correspondence analysis, abbreviated C2CFA is a method of data analysis widely used in biostatistics and survey analysis. The C2CFA technique is mainly used for large tables to compare all data (if possible all expressed in the same unit, like a currency, a dimension, a frequency or any other measurable quantity). It may in particular allow the study of contingency tables (or co-occurrence cross tables) and describe the connection between two variables. It serves to identify and prioritize all the dependencies between the rows and columns of the table.

If more than two variables are to be taken into considerations we speak of Chi-square multiple correspondence analysis (C2MCA).

Let us tackle now directly the theory with an example. For this we consider the following table (with two variables) of the areas of the types of tree that stands in Picardy (France) in 11984 (holocene calendar) in hectares:

	Leafy	**Resinous**	**Mixes**	**Total by Area**
Aisne (A)	$106,500$	$3,380$	$1,470$	$111,350$
Oise (O)	$101,700$	$310,000$	0	$111,700$
Somme (S)	$45,200$	$4,350$	50	$49,600$
Total	$253,400$	$17,730$	$1,520$	$272,650$

Table 5.94: **Contingency table (crosstab) of CFA**

The experts in the domain sometimes name the totals for rows and columns, respectively, the "line margins" and "column margins". When the whole table is put in percentages, relatively to the total of the totals, it is named "joint frequency representation":

	Leafy	**Resinous**	**Mixes**	**Total by Area**
Aisne (A)	39.06%	1.24%	0.54%	40.84%
Oise (O)	37.30%	3.67%	0%	40.97%
Somme (S)	16.58%	1.60%	0.02%	18.19%
Total	92.93%	6.50%	0.56%	100%

Table 5.95: **Joint frequency table of CFA**

We wish to analyse whether there are degrees of similarity and difference between the variables. Let us notice that we are not trying to compare equality of means or variances therefore statistical tools seen in previously are not suitable for this kind of analysis.

If we choose the Euclidean distance (see section Vector Calculus page 1818):

$$d^2(x_i,x_{i'})=\left|\left|\overrightarrow{(x_ix_{i'})}\right|\right|^2=\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 \qquad (5.7.2635)$$

on the raw data to measure the differences between departments, we get the following differences:

$$\begin{aligned}d^2(A,O)&=\sum_{j=1}^{3}(x_{1j}-x_{2j})^2=(106,500-101,700)^2+(3,380-10,000)^2+(1,470-0)^2\\&=69,025,300\Rightarrow d(A,O)=\sqrt{69,025,300}=8,308.1\ [\text{ha}]\end{aligned} \tag{5.7.2636}$$

and so on for the other regions. We then get:

$$\begin{aligned}d(A,O)&=8,308.1\ [\text{ha}]\\d(A,S)&=61,324.1\ [\text{ha}]\\d(O,S)&=56,781.8\ [\text{ha}]\end{aligned} \tag{5.7.2637}$$

We see by looking at the table and before any calculation that the departments of Aisne and Oise seems to be similar and that the department of the Somme differs significantly. The distances put in evidence that observation.

But in the above table the profiles of the Oise and the Somme, with a very small mixed forest, are however very similar in proportion.

In this context, we see that the Euclidean distance transcribed quite well the mass differences between the departments. In other words, the Aisne and the Oise are similar because their surfaces are close. To eliminate the artifact related to the orders of magnitude, we need to transform the data into percentages (percentages of regions). We then get:

	Leafy	**Resinous**	**Mixes**	**% Area**
Aisne (A)	95.6%	3.0%	1.3%	40.8%
Oise (O)	91.0%	9.0%	0.0%	41.0%
Somme (S)	91.1%	8.8%	0.1%	18.2%

Table 5.96: **Transformation into percentages of the contingency table**

where the experts in the field sometimes name the column %Area "marginal profile rows" or "mass" (and respectively when they indicate the rows of the percents for the trees).

If we choose the Euclidean distance on the proportions (relative data), we get:

$$d^2(p_i,p_{i'})=||\overrightarrow{p_ip_{i'}}||^2=\sum_{j=1}^{p}\left(\frac{x_{ij}}{x_{i.}}-\frac{x_{i'j}}{x_{i'.}}\right)^2=\sum_{j=1}^{p}\left(\frac{x_{ij}}{\displaystyle\sum_j x_{ij}}-\frac{x_{i'j}}{\displaystyle\sum_j x_{i'j}}\right)^2 \tag{5.7.2638}$$

Therefore:

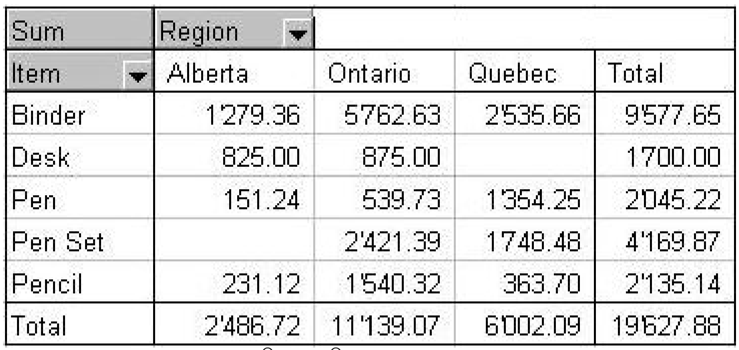
$$\begin{aligned}d(A,O)&=7.6\%\\d(A,S)&=7.4\%\\d(O,S)&=0.2\%\end{aligned} \tag{5.7.2639}$$

This time, the Oise and Somme appear well as have the most similar forests. We see that work with relative data seems more relevant in this case!

Now let us borrow an idea to economists who, when they have tables of the same kind as the previous one, calculate what they named the "index" or "elasticity" (also often name "specificity index" in statistics) and which is given by the ratios between the joint frequency and the marginal frequency:

$$I=\frac{\dfrac{x_{ij}}{x_{i.}}}{\dfrac{x_{.j}}{x_{..}}}=\frac{\dfrac{x_{ij}}{x_{j.}}}{\dfrac{x_{i.}}{x_{..}}} \tag{5.7.2640}$$

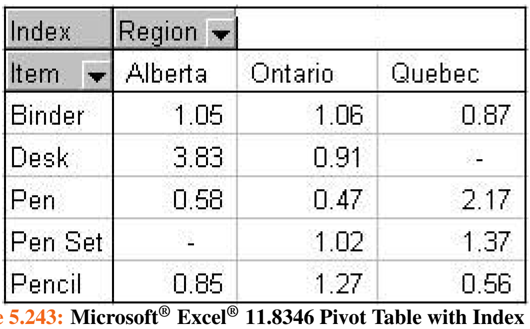
Here is an example obtained with the Pivot Table tool in Microsoft® Excel® 11.8346 which includes natively the Index function. First the starting table:



Sum	Region			
Item	Alberta	Ontario	Quebec	Total
Binder	1'279.36	5762.63	2'535.66	9'577.65
Desk	825.00	875.00		1700.00
Pen	151.24	539.73	1'354.25	2'045.22
Pen Set		2'421.39	1748.48	4'169.87
Pencil	231.12	1'540.32	363.70	2'135.14
Total	2'486.72	11'139.07	6'002.09	19'627.88

Figure 5.242: **Microsoft® Excel® 11.8346 Pivot Table for Index analysis**

and by activating the Index function in the Pivot Table properties we get:



Index	Region		
Item	Alberta	Ontario	Quebec
Binder	1.05	1.06	0.87
Desk	3.83	0.91	-
Pen	0.58	0.47	2.17
Pen Set	-	1.02	1.37
Pencil	0.85	1.27	0.56

Figure 5.243: **Microsoft® Excel® 11.8346 Pivot Table with Index values**

To understand where these values comes from, let us look for example the article *Desk* in the region *Alberta*. It has a return (joint frequency) of:

$$\frac{x_{21}}{x_{2.}}=\frac{825}{1700}\cong 48.52\% \tag{5.7.2641}$$

compared to all regions, which is above the value of 33.33% that this article would have for yield for all regions if there is no region of preference!

The *Alberta* region a yield (marginal rate) of:

$$\frac{x_{1}}{x_{..}}=\frac{2486.72}{19627.88}\cong 12.66\% \tag{5.7.2642}$$

with respect to all regions which is below to the yield of 33.33% that it would have if there were no regions of preference. Thus, this index table shows whether the differences are qualitatively significant!!

The ratio therefore gives:

$$I=\frac{\dfrac{x_{21}}{x_{2.}}}{\dfrac{x_{.1}}{x_{..}}}\cong 3.83 \qquad (5.7.2643)$$

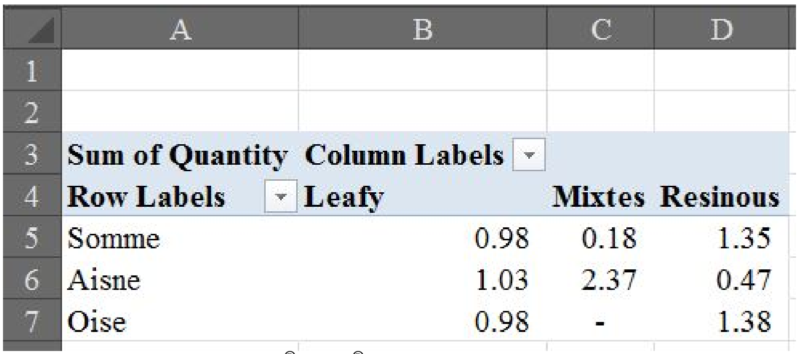
which shows a strong shift between the value obtained and the value we would have if the proportions were respected (over-representation of 283%).

So it is a kind of conformity compliance: if the ratio was equal to 1, it is that the regional sales performance of this particular article would be consistent with respect to all sales of this region with respect to the national market. There would be no anomalies. Let us see this example for our trees where we had seen the effectives:

	Leafy	**Resinous**	**Mixes**	**Total by Area**
Aisne (A)	$106,500$	$3,380$	$1,470$	$111,350$
Oise (O)	$101,700$	$10,000$	0	$111,700$
Somme (S)	$45,200$	$4,350$	50	$49,600$
Total	$253,400$	$17,730$	$1,520$	$272,650$

Table 5.97: **Contingency table (crosstab) of CFA**

and for which we get the following Pivot Table of the effective index in Microsoft® Excel® 14.0.7166:



	A	B	C	D
1				
2				
3	**Sum of Quantity**	**Column Labels**		
4	**Row Labels**	**Leafy**	**Mixtes**	**Resinous**
5	Somme	0.98	0.18	1.35
6	Aisne	1.03	2.37	0.47
7	Oise	0.98	-	1.38

Figure 5.244: **Microsoft® Excel® 14.0.7166 Pivot Table for Index analysis**

and we still see clearly with this table that it is well the Oise and Somme are most alike!

Before continuing, we might ask ourselves the following most important question: What are the theoretical numbers that we would have been obtained if the proportions of the trees in the regions were strictly equivalent to the overall proportions (i.e. so that the index are all unitary)?

Well simply by making the following calculations (it is merely a rule of three calculated in each cell) where the reader must - if possible - understand the meaning without applying this foolishly:

	Leafy	**Resinous**	**Mixes**
Aisne (A)	$=(253,400/272,650)\cdot 111,350$ $=103,488$	$=(17,730/272,650)\cdot 111,350$ $=7,241$	$=(1,520/272,650)\cdot 111,350$ $=621$
Oise (O)	$=(253,400/272,650)\cdot 111,700$ $=103,813$	$=(17,730/272,650)\cdot 111,700$ $=7,264$	$=(1,520/272,650)\cdot 111,700$ $=623$
Somme (S)	$=(253,400/272,650)\cdot 49,600$ $=46,098$	$=(17,730/272,650)\cdot 49,600$ $=3,225$	$=(17,730/272,650)\cdot 49,600$ $=276$
Total	$253,400$	$17,730$	$1,520$

Table 5.98: **Proportions respect of CFA**

Or in a more readable way:

	Leafy	**Resinous**	**Mixes**	**Total by Area**
Aisne (A)	$E_{11}=103,488$	$E_{12}=7,241$	$E_{13}=621$	$r_1=111,350$
Oise (O)	$E_{21}=103,813$	$E_{22}=7,264$	$E_{23}=623$	$r_2=111,700$
Somme (S)	$E_{31}=46,098$	$E_{32}=3,225$	$E_{33}=276$	$r_3=49,600$
Total	$c_1=253,400$	$c_2=17,730$	$c_3=1,520$	$N=272,650$

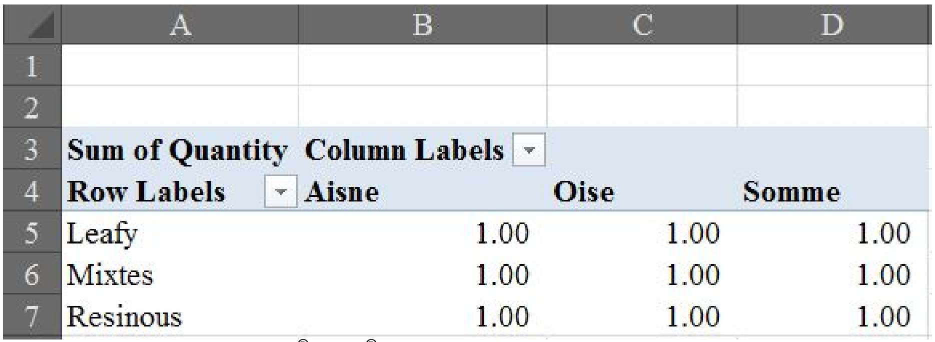
Notice that in the table above, each cell (denoted E_{ij}), is obtained by computing:

$$E_{ij}=\frac{c_j\cdot r_i}{N} \tag{5.7.2644}$$

Remark

Notice that if the cells have to be given in percentage, rather than in counts, we would then have to calculate instead: $p_{ij}=E_{ij}/N=(c_j\cdot r_i)/N^2$. Hence obviously: $E_{ij}=Np_{ij}$.

And we get with these new values of the table the following theoretical index with Microsoft® Excel® 14.0.7166:



	A	B	C	D
1				
2				
3	**Sum of Quantity**	**Column Labels**		
4	**Row Labels**	**Aisne**	**Oise**	**Somme**
5	Leafy	1.00	1.00	1.00
6	Mixtes	1.00	1.00	1.00
7	Resinous	1.00	1.00	1.00

Figure 5.245: **Microsoft® Excel® 14.0.7166 Pivot Table for Theoretical Index analysis**

which shows now that the proportions are now observed! Parenthesis closed (but on which will come back later)!

Well when we want to make a factorial correspondence analysis, our relation:

$$d^2(x_i,x_{i'})=\sum_{j=1}^{p}\left(\frac{x_{ij}}{x_{i.}-\dfrac{x_{i'j}}{x_{i'.}}}\right)^2 \tag{5.7.2645}$$

Therefore becomes:

$$\boxed{d^2(x_i,x_{i'})=\sum_{j=1}^{p}\frac{\left(\dfrac{x_{ij}}{x_{i.}}-\dfrac{x_{i'j}}{x_{i'.}}\right)^2}{\dfrac{x_{.j}}{x_{..}}}} \tag{5.7.2646}$$

hence:

$$\begin{aligned}d(A,O)&=0.30\\ d(A,S)&=0.39\\ d(O,S)&=0.02\end{aligned} \tag{5.7.2647}$$

Again, the Oise and the Somme appear as the most similar.

The distance above is named the "Chi-square metric" because it looks like (but that's all!) to the distance used in the independence test of the same name (see section Statistics page 1238) but here, it only helps to establish a hierarchy in the context of a contingency table and to observe similar the variables in a more easy way!!

Remark

There is another way to calculate a CFA based on an Euclidean distance but by taking care beforehand to transform in a special way to the contingencies table so that the calculation are the same that when we use the Chi-square metric.

5.7.15.3 Chi-squared Test of Independence

The χ^2-test is applied when you have two categorical variables from a single population. It is used to determine whether there is a significant association between the two variables.

For example, in an election survey, voters might be classified by gender (male or female) and voting preference (Democrat, Republican, or Independent). We could use a chi-square test for independence to determine whether gender is related to voting preference.

Suppose that a categorical variable A has r levels, and another categorical variable B has c levels. The null hypothesis H_0 states that knowing the level of variable A does not help you predict the level of variable B. That is, the variables are independent.

$$\begin{aligned}&H_0: \text{Variable } A \text{ and variable } B \text{ are independent}\\ &H_1: \text{Variable } A \text{ and variable } B \text{ are not independent}\end{aligned} \tag{5.7.2648}$$

We will present this test with a companion example because we know from experience that it is more effective for learning and understanding.

Let us recall that during the introduction of the previous method for comparing numbers (values) and detect which were the closest we gave the following observed effectives:

	Leafy	Resinous	Mixes	Total by Area
Aisne (A)	$106,500$	$3,380$	$1,470$	$111,350$
Oise (O)	$101,700$	$3,380$	$1,470$	$111,700$
Somme (S)	$45,200$	$4,350$	50	$49,600$
Total	$253,400$	$17,730$	$1,520$	$272,650$

Table 5.99: **Contingency table (crosstab) for χ^2 analysis**

and we also showed how to find the table of theoretical effectives (rounded to the nearest whole number) where the proportions should have possibly be respected:

	Leafy	Resinous	Mixes	Total by Area
Aisne (A)	$103,488$	$7,241$	621	$111,350$
Oise (O)	$103,813$	$7,264$	623	$111,700$
Somme (S)	$46,098$	$3,225$	276	$49,600$
Total	$253,400$	$17,730$	$1,520$	$272,650$

Table 5.100: **Contingency table (crosstab) for χ^2 analysis with theoretical effective**

We know that building the last table above assumes that the three regions are in identical conditions for everything related to the growth and multiplication of trees and the number of trees is in direct causal relation (!!!!) with regions and there is no other intermediate causes ... what is a strong assumption!

But under this assumption, let us suppose that we would like know if the differences observed between the number of trees and the regions are statistically significant or purely random because of the experimental sample? In other words, we want to know if the number of trees actually depends on the regions in which they grow or if these values are only due to the sample chance? This is why this test is named the "Chi-square test of independence".

So keep in mind that in the Chi-square test of independence, all subjects/units are collected at random from a population, and two categorical variables are observed for each unit.

Remark

The chi-square test of independence is recommended in sensory analysis by the standard ISO 8588:1987 under the name "test A-Not A".

To answer this questions we must first have a reference. And this reference is precisely the assumption of direct causal link (proportions respected) that we have given just previously.

If we consider that each cell of the table of observed effective corresponds to a random variable of unknown law and that each cell of theoretical table is considered as following a random binomial variable (and asymptotically as a Normal distribution) then as we proved it earlier we know that we can use the chi-square test of adjustment (see page 1113):

$$D=\sum_{i=1}^{N}E_i^2=\sum_{i=1}^{N}\frac{(k_i-np_i)^2}{np_i} \tag{5.7.2649}$$

to have a good idea (but still pretty rough considering the assumptions!) if the differences between the observed effective values are due to chance or are significant. But, if D is small, the probability that it is due to chance is great but if D is large so we have a real - significant - difference (so we use the chi-square test of adjustment but in reverse way in fact!).

It remains to determine the number of degrees of freedom of the χ^2 law that follows this sum in this configuration!

In the particular case (but quite easily generalized by recurrence) of a table with two inputs with two categorized variables X with l levels and Y with c we will have respectively the l rows and c columns.

Thus, the table will have obviously $l \cdot c$ cells. In the table of theoretical effectives (where each cell is treated as a random variable) each cell will be entirely determined by the sum of the other so that the number of degrees of freedom will be logically as we have seen during our study of degrees of freedom:

$$N-1=l\cdot c-1 \tag{5.7.2650}$$

Thus, taking our example of forests, the total of total is $272,650$ which allows us to write this last relation and thus determine the value of a possibly empty cell, all other data being known!

A chi-square test on this type of table tests the hypothesis of independence against the alternative hypothesis of non-independence. Under the assumption of independence we believe we need only:

$$(l-1)+(c-1) \tag{5.7.2651}$$

values on the N one to determine all of them (assuming implicitly known the sum of all rows and columns).

So if we have a table of 2 rows by 2 columns, we only need if we know the total rows and columns, to know 2 values (i.e. $(2-1)+(2-1)=2$) to determine the 2 missing values. The reasoning is the same for a table of 3 rows by 3 columns where we just have to know at least 4 values (i.e. $(3-1)+(3-1)=4$) to determine the missing 5 values.

The degrees of freedom for the chi-square is then:

$$\text{df}=(lc-1)-[(l-1)+(c-1)]=lc-1-l+1-c+1=lc-l-c+1=(l-1)(c-1) \tag{5.7.2652}$$

It is this relation that we say us (trivially!) that if in a table of 2 rows by 2 columns so with 4 cells (total of rows and columns being known!) that being known only one single values (df being equal to 1), we can determine the other 3 missing values.

As we know a possible definition of the number of degrees of freedom df is that it is the maximum number of values of the model such that none of them is calculable from the others.

Similarly, for a table of 3 rows by 3 columns with 9 cells as is the case of our example above with forests, the knowledge of only 4 cells allows us thanks to total lines and columns to determine the other 5 that would possibly not be known.

Hence the relation in the context of the application of chi-square for the final equation:

$$\boxed{\chi^2((l-1)(c-1))=\chi^2_{(l-1)(c-1)}=\sum_{i,j}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}} \tag{5.7.2653}$$

by making use of notations used in the industry. The term:

$$\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \tag{5.7.2654}$$

is often named "square of standardized residual". The squared root is more commonly used and named the "standardized residual":

$$\frac{O_{ij}-E_{ij}}{\sqrt{E_{ij}}} \tag{5.7.2655}$$

The standardized residual is a measure of the strength of the difference between observed and expected values. It's a measure of how significant your cells are to the chi-square value. When you compare the cells, the standardized residual makes it easy to see which cells are contributing the most to the value, and which are contributing the least. If your sample is large enough, the standardized residual can be roughly compared, as we already know it, to a Z-score! Standardization can work even if your variables are not Normally distributed.

A general rule of thumb for figuring out what the standardized residual means, is:

- If the standardised residual is less than -2, the cell observed frequency is less than the expected frequency.
- Greater than $+2$ and the observed frequency is greater than the expected frequency.

If your standardised residuals are ± 3, then it means that something extremely unusual is happening. If you get ± 4, its something from the Twilight Zone! This makes sense if you think about the 68 95 99.7 rule: if your data is normally distributed, 95% of your data should be within two standard deviations from the mean. If you have something greater than that, then you're looking at an outlier.

The ratio:

$$\frac{\dfrac{(O_{ij}-E_{ij})^2}{E_{ij}}}{\displaystyle\sum_{i,j}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}} \tag{5.7.2656}$$

is often named the "contribution to the Chi-square of independence".

Remark

To use this test properly, the practitioner should check first that the differences (numerator) follow a Normal distribution or that all terms of the sum follow a Chi-square distribution or approximately (asymptotically) a Normal centered reduced distribution centered and that the effective in each cell are greater than 5 otherwise Monte Carlo simulations have to be used to determine the p-value.

In our example, we have:

$$D=\sum_{i,j}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}=\sum_{i=1}^{3}\sum_{j=1}^{3}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}\cong 5,602.35 \tag{5.7.2657}$$

and the p-value of this value with the chi-square distribution with 4 degrees of freedom:

$$\chi^2((l-1)(c-1))=\chi^2(4)=\chi^2_4 \tag{5.7.2658}$$

is so close to zero (not statistically significant) that we have no chance of being mistaken in asserting that the observed differences in the table are statistically significant between the 3 and so that there is very likely independence.

Remark

Let us recall that there are plenty of statistical tests with the name that contains "chi-square" in it. For example:

- Chi-square Goodness-of-fit test (allow us to determine whether a specified population distribution is valid)
- Chi-square test of association/independence (allow us to determine whether the distribution of one variable has been influenced by another variable)
- Chi-square test of homogeneity (allow us to compare two or more population proportions), also named Chi-square test of heterogeneity
- Chi-square test for outliers (allow us to detect if there is at least one outlier in comparison of a specific distribution)
- Chi-square test for difference of two count data (special case of homogeneity test with only one category!)

Even if all those tests have different purposes, their mathematical framework and procedure are EXACTLY the same and therefore any of these five tests can be run to get at the same time the five conclusions with the same p-values and the same critical chi-square values!!!

5.7.15.3.1 Equivalence of Z-test and χ^2 test for proportions

We have mentioned during our study of the Z-test for the comparison of two proportions that softwares like `R` use the χ^2 test to compare proportions rather than the Z-test. It's time now to see why there is an equivalence between the 2-sample Wald's Z-statistic for comparing proportions with pooled variances and the Pearson's χ^2 (independence) test for a 2×2 contingency table.!

Say now that we have a study of two categorical variables each with only two levels. One of the response levels is considered the *Success* response and the other the *Failure* response (typical of an A/B test as used sensitivity analysis and especially well known in webdesign...). A general 2×2 table of the observed counts would be as follows:

	Success	**Failure**	**Total**
Group 1	$A\,(O_{11})$	$B\,(O_{12})$	$n_1=A+B$
Group 2	$C\,(O_{21})$	$D\,(O_{22})$	$n_2=C+D$

The observed counts in this table represent the following proportions:

	Success	**Failure**
Group 1	$\hat{p}_1=\frac{A}{A+B}=\frac{A}{n_1}$	$1-\hat{p}_1$
Group 2	$\hat{p}_2=\frac{C}{C+D}=\frac{C}{n_2}$	$1-\hat{p}_2$

What is also often denoted in this special case:

	Success	**Failure**	**Total**
Group 1	p_1	q_1	100%
Group 2	p_2	q_2	100%
Total	p	q	100%

Let us now use the table above blindly as our "contingency" table by noticing that if we take the above table as reference, then $N=100\%=1$.

The usual χ^2 of this table written after we substitute proportions instead of frequencies in its formula, looks like this (as $N=100\%=1$ and the total rows are also equal to 100% it simplifies things quite a lot):

$$\begin{aligned}\chi_1^2&=\sum_{i,j}^{n}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}\\&=\frac{(n_1p_1-n_1p/100\%)^2}{n_1p/100\%}+\frac{(n_2p_2-n_2p/100\%)^2}{n_2p/100\%}+\frac{(n_1q_1-n_1q/100\%)^2}{n_1q/100\%}+\frac{(n_2q_2-n_2q/100\%)^2}{n_2q/100\%}\\&=\frac{(n_1p_1-n_1p)^2}{n_1p}+\frac{(n_2p_2-n_2p)^2}{n_2p}+\frac{(n_1q_1-n_1q)^2}{n_1q}+\frac{(n_2q_2-n_2q)^2}{n_2q}\end{aligned}\tag{5.7.2659}$$

But this can be simplified even more. Indeed, if you remember how we computer the p_{ij}, these are the multiplication of the row total and column total. But in our special case above (with a table containing only proportions!), the rows are always equal to $100\%=1$. Therefore only the total column terms need to be written. This bring us to write:

$$\begin{aligned}\chi_1^2&=\frac{(O_{11}-p\cdot 100\%)^2}{p\cdot 100\%}+\frac{(O_{21}-p\cdot 100\%)^2}{p\cdot 100\%}+\frac{(O_{12}-q\cdot 100\%)^2}{q\cdot 100\%}+\frac{(O_{22}-q\cdot 100\%)^2}{q\cdot 100\%}\\&=\frac{(O_{11}-p)^2}{p}+\frac{(O_{21}-p)^2}{p}+\frac{(O_{12}-q)^2}{q}+\frac{(O_{22}-q)^2}{q}\\&=n_1\frac{(p_1-p)^2}{p}+n_2\frac{(p_2-p)^2}{p}+n_1\frac{(q_1-q)^2}{q}+n_2\frac{(q_2-q)^2}{q}\end{aligned}\tag{5.7.2660}$$

We can simplify even more:

$$\begin{aligned}\chi_1^2&=n_1\frac{(p_1-p)^2}{p}+n_1\frac{(q_1-q)^2}{q}+n_2\frac{(p_2-p)^2}{p}+n_2\frac{(q_2-q)^2}{q}\\&=n_1\left[\frac{(p_1-p)^2}{p}+\frac{(q_1-q)^2}{q}\right]+n_2\left[\frac{(p_2-p)^2}{p}+\frac{(q_2-q)^2}{q}\right]\\&=\frac{n_1(p_1-p)^2q}{pq}+\frac{n_1(q_1-q)^2p}{pq}+\frac{n_2(p_2-p)^2q}{pq}+\frac{n_2(q_2-q)^2p}{pq}\\&=\frac{n_1(p_1-p)^2(1-p)+n_1(1-p_1-1+p)^2p+n_2(p_2-p)^2(1-p)+n_2(1-p_2-1+p)^2p}{pq}\\&=\frac{n_1(p_1-p)^2(1-p)+n_1(p-p_1)^2p+n_2(p_2-p)^2(1-p)+n_2(p-p_2)^2p}{pq}\\&=\frac{[n_1(p_1-p)^2][(1-p)+p]+[n_2(p_2-p)^2][(1-p)+p]}{pq}\\&=\frac{n_1(p_1-p)^2+n_2(p_2-p)^2}{pq}\end{aligned}\tag{5.7.2661}$$

Remember that:

$$p=\frac{n_1p_1+n_2p_2}{n_1+n_2}\tag{5.7.2662}$$

Plug-in that latter in the numerator in the above relation, we get:

$$\begin{aligned}\chi_1^2 &= \frac{n_1\left(p_1-p\right)^2+n_2\left(p_2-p\right)^2}{pq}\\ &= \frac{n_1\left(p_1-\frac{n_1p_1+n_2p_2}{n_1+n_2}\right)^2+n_2\left(p_2-\frac{n_1p_1+n_2p_2}{n_1+n_2}\right)^2}{pq}\\ &= \frac{n_1\left(p_1(n_1+n_2)-(n_1p_1+n_2p_2)\right)^2+n_2\left(p_2(n_1+n_2)-(n_1p_1+n_2p_2)\right)^2}{pq(n_1^2+n_2^2)}\\ &= \frac{n_1\left(p_1n_2-n_2p_2\right)^2+n_2\left(p_2n_1-n_1p_1\right)^2}{pq(n_1^2+n_2^2)}\\ &= \frac{n_1\left(p_1^2n_2^2-2p_1n_2n_2p_2+n_2^2p_2^2\right)+n_2\left(p_2^2n_1^2-2p_2n_1n_1p_1+n_1^2p_1^2\right)}{pq(n_1^2+n_2^2)}\\ &= \frac{n_1p_1^2n_2^2-2n_1p_1n_2n_2p_2+n_1n_2^2p_2^2+n_2p_2^2n_1^2-2n_2p_2n_1n_1p_1+n_2n_1^2p_1^2}{pq(n_1^2+n_2^2)}\\ &= \frac{p_1^2(n_1n_2^2+n_2n_1^2)+p_2^2(n_1n_2^2+n_2n_1^2)-2p_1p_2(n_1n_2^2+n_2n_1^2)}{pq(n_1^2+n_2^2)}\\ &= \frac{(p_1-p_2)^2(n_1n_2^2+n_2n_1^2)}{pq(n_1^2+n_2^2)}\end{aligned}\tag{5.7.2663}$$

We divide both numerator and denominator by the $(n_1n_2^2+n_2n_1^2)$ and get:

$$\begin{aligned}\chi_1^2 &= \frac{(n_1n_2^2+n_2n_1^2)}{(n_1n_2^2+n_2n_1^2)}\frac{(p_1-p_2)^2(n_1n_2^2+n_2n_1^2)}{pq(n_1^2+n_2^2)}=\frac{(p_1-p_2)^2}{pq\dfrac{(n_1^2+n_2^2)}{n_1n_2^2+n_2n_1^2}}\\ &= \frac{(p_1-p_2)^2}{pq\dfrac{(n_1+n_2)(n_1-n_2)}{n_1n_2(n_1+n_2)}}=\frac{(p_1-p_2)^2}{pq\dfrac{n_1-n_2}{n_1n_2}}=\frac{(p_1-p_2)^2}{pq\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}\end{aligned}\tag{5.7.2664}$$

and as the χ_1^2 is equal to Z^2 this lead us to:

$$Z^2=\frac{(p_1-p_2)^2}{pq\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}\tag{5.7.2665}$$

that is the squared Z-statistic of the Z-test of proportions (pooled version as seen at page 881) for *Success* response!!

Thus, the 2×2 homogeneity Chi-square statistic (and test) is equivalent to the Z-test of two proportions. The so called expected frequencies computed in the chi-square test in a given column is the weighted (by the group n) average vertical profile (i.e. the profile of the "average group") multiplied by that group's n. Thus, it comes out that chi-square tests the deviation of each of the two groups profiles from this average group profile, - which is equivalent to testing the groups' profiles difference from each other, which is the Z-test of proportions. This is why quite a significant percentage of webdesigner refer to A/B proportion test as Z-test or χ^2 test because they are simply the same (however rigorously we should always mention if we use the pooled or unpooled version with or without Yates correction...)

This result demonstrates of a link between variables association measure (χ^2) and a group difference measure (Z-test statistic). Attribute associations and group differences are (often) the two facets of the same thing.

Recall from our Z-test of two proportions that our null hypothesis is that the two population proportions, p_1 and p_2, were assumed equal while the two-sided alternative hypothesis was that they were not equal.

This null hypothesis would be analogous to the two groups being independent. Also, if the two success proportions are equal, then the two failure proportions would also be equal.

Remark

For a practical (i.e. numerical) example of the equivalent between the pooled proportions Z-test and the pooled 2×2 contingency χ^2 test, the reader can refer to our `R` companion book where we will show that in facts `R` always use the chi-square version whatever the function we use (`prop.test()` function or `chisq.test()` both use the chi-square test anyway!).

5.7.15.3.2 Cramér's V

The "Cramér's V" (sometimes referred to as "Cramér's phi" and denoted as φ_c) is a measure of association between two nominal variables, giving a value between 0 and $+1$ (inclusive). It is based on Pearson's chi-squared statistic and was published by Harald Cramér in 11946 (holocene calendar).

As we will prove it, φ_c is the intercorrelation of two discrete variables and may be used with variables having two or more levels. φ_c is a symmetrical measure, it does not matter which variable we place in the columns and which in the rows. Also, the order of rows/columns doesn't matter, so φ_c may be used with nominal data types or higher (ordered, numerical, etc.)

We have seen just previously that the chi-square test of independence may be used to measure the degree of association of two categorical variables in a contingency table of l rows and c columns:

$$D=\sum_{ij}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \tag{5.7.2666}$$

and that this distance follows a chi-square distribution with $(l-1)(c-1)$ degrees of freedom. We will prove intuitively that the maximum value of the distance D is given by:

$$D_{\max}=N\left(\min(l,c)-1\right) \tag{5.7.2667}$$

and that maximum value is achieved if and only if each row or column contains exactly one non-zero value. Under the latter condition, we can always rearrange the contingency table so to have all non-zero terms on the diagonal of the table.

Obviously if the table is not square as below:

Row/Column	c_1	c_2	c_3	c_4	**Total**
l_1	a_1	0	0	0	a_1
l_2	0	a_2	0	0	a_2
l_3	0	0	a_3	0	a_3
l_4	0	0	0	a_4	a_4
l_5	0	0	0	0	0
Total	a_1	a_2	a_3	a_4	N

Table 5.101: **Rectangular contingency table**

The case that maximize D request diagonal terms on the smallest dimension in row or column (row dimension in the example above) denoted by tradition q. In the special example above we have:

$$q=\min(l,c)=c=4 \tag{5.7.2668}$$

In this special case and obviously theoretical one, the rows that have only zero values can be omitted and therefore the previous table can be reduced to:

Row/Column	c_1	c_2	c_3	c_4	**Total**
l_1	a_1	0	0	0	a_1
l_2	0	a_2	0	0	a_2
l_3	0	0	a_3	0	a_3
...	...	...	...	...	...
l_q	0	0	0	a_q	a_q
Total	a_1	a_2	...	a_q	N

Table 5.102: **Squared rectangular contingency table**

Of course, ignoring the rows or columns that have only zero values bring us to assume that for the distance D of the chi-square we request that:

$$\frac{(0-0)^2}{0}=0 \tag{5.7.2669}$$

which is quite indisputable ... For what will follow, we will need the following relations:

$$N=\sum_{i=1}^{q} a_i \tag{5.7.2670}$$

and:

$$N^2=\left(\sum_{i=1}^{q} a_i\right)^2=\sum_{i=1}^{q} a_i\sum_{j=1}^{q} a_j=\sum_{i=1}^{q}\sum_{j=1}^{q} a_ia_j=\sum_{i=1}^{q} a_i^2+\sum_{i\neq j}^{q} a_ia_j \tag{5.7.2671}$$

Therefore it comes:

$$\begin{aligned}
D_{\max}&=\sum_{ij}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}=\sum_{i}^{q}\frac{\left(a_i-\dfrac{a_i^2}{N}\right)^2}{\dfrac{a_i^2}{N}}+\sum_{i\neq j}^{q}\frac{\left(0-\dfrac{a_ia_j}{N}\right)^2}{\dfrac{a_ia_j}{N}}=\sum_{i}^{q}\frac{\left(\dfrac{Na_i-a_i^2}{N}\right)^2}{\dfrac{a_i^2}{N}}+\sum_{i\neq j}^{q}\frac{a_ia_j}{N}\\
&=\sum_{i}^{q}\frac{(a_i(N-a_i))^2}{a_i^2N}+\frac{1}{N}\left(\sum_{i\neq j}^{q}a_ia_j+\sum_{i}^{q}a_i^2\right)-\frac{1}{N}a_i^2\\
&=\sum_{i}^{q}\frac{(N-a_i)^2}{N}+\frac{1}{N}N^2-\frac{1}{N}\sum_{i}^{q}a_i^2=\frac{1}{N}\sum_{i}^{q}(N-a_i)^2-\frac{1}{N}\sum_{i}^{q}a_i^2+\frac{1}{N}N^2\\
&=\frac{1}{N}\left[\sum_{i}^{q}(N-a_i)^2-\sum_{i}^{q}a_i^2\right]+N=\frac{1}{N}\sum_{i}^{q}\left((N-a_i)^2-a_i^2\right)+N=\sum_{i}^{q}\frac{((N-a_i)^2-a_i^2)}{N}+N\\
&=\sum_{i}^{q}\frac{N^2-2Na_i+a_i^2-a_i^2}{N}+N=\sum_{i}^{q}\frac{N^2-2Na_i}{N}+N=\sum_{i}^{q}(N-2a_i)+N\\
&=qN-2\sum_{i}^{q}a_i+N=qN-2N+N=N(q-1)=N(\min(l,c)-1)
\end{aligned} \tag{5.7.2672}$$

We then define the following value:

$$\boxed{V=\sqrt{V^2}=\sqrt{\frac{\chi^2(l-1)(c-1)}{N(q-1)}}=\sqrt{\frac{\chi^2_{(l-1)(c-1)}}{N(q-1)}}} \tag{5.7.2673}$$

as being the "Cramér's V coefficient" (the majority of software, however, give the value of V squared). The latter is such that it never exceeds 1 and allows a more intuitive interpretation of the degree of association in a contingency table.

In the case where the table size is of 2 rows and 2 columns, the previous relation reduces then immediately to:

$$V=\sqrt{\frac{\chi_1^2}{N}} \tag{5.7.2674}$$

Relation that is traditional to note in this case as follows:

$$\varphi=\sqrt{\frac{\chi_1^2}{N}} \tag{5.7.2675}$$

and to name "Cramér's phi" or "phi coefficient".

✍ **Example:**

Consider the following table (even if the conditions are not met for a chi-square test of independence):

Projects	**Certified Project Manager**	**Non-Certified Project Manager**	**Total**
Deadlines met	8	1	9
Deadlines not met	4	5	9
Total	12	6	18

Table 5.103: **Example of a contingency table for calculating the Cramér's V**

with the theoretical effectives:

Projects	**Certified Project Manager**	**Non-Certified Project Manager**	**Total**
Deadlines met	$(12/18)\cdot 9=6$	$(6/18)\cdot 9=3$	9
Deadlines not met	$(12/18)\cdot 9=6$	$(6/18)\cdot 9=3$	9
Total	12	6	18

Table 5.104: **Contingency table of the example with theoretical effectives**

Then we have:

$$\begin{aligned}D=\chi^2_{(l-1)(c-1)}=\chi_1^2&=\sum_{ij}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}\\&=\frac{(8-6)^2}{6}+\frac{(4-6)^2}{6}+\frac{(1-3)^2}{3}+\frac{(5-3)^2}{3}\\&=0.\bar{6}+0.\bar{6}+1.\bar{3}+1.\bar{3}=4\end{aligned} \tag{5.7.2676}$$

And to a threshold level of 95%, we get with Microsoft® Excel® 14.0.6123:

$$\chi^2_{95\%}(1)=\texttt{=CHISQ.INV(0.95,1)}=3.841 \tag{5.7.2677}$$

or with the corresponding p-value:

$$=\texttt{1-CHISQ.DIST(4,1,TRUE)}=0.045 \tag{5.7.2678}$$

It then comes immediately:

$$V=\sqrt{V^2}=\sqrt{\frac{\chi^2_{(l-1)(c-1)}}{N(q-1)}}=\sqrt{\frac{4}{18(2-1)}}=\sqrt{0.\bar{2}}\cong 0.4714 \tag{5.7.2679}$$

We are a bit at the limit here ... since the p-value is very close to the traditional 0.05. However, take a decision in this case where the effectives are so low would be equivalent to conclude anything and even more that this statistical tool is built on an accumulation of approximations.

5.7.15.4 Pearson's phi coefficient

In statistics, the "Pearson's phi coefficient" is an another measure of association for two binary variables. Introduced by Karl Pearson, this measure is similar to the Pearson correlation coefficient in its interpretation. In fact, a Pearson correlation coefficient estimated for two binary variables (Bernoulli variables) will return the Pearson's phi coefficient.

To introduce this correlation coefficient and proof that it is only a special case of the Pearson correlation coefficient, let us consider the following table:

		Y		
		$Y=1$	$Y=0$	**Total**
X	$X=1$	n_{11}	n_{12}	$n_{1.}$
	$X=0$	n_{21}	n_{22}	$n_{2.}$
	Total	$n_{.1}$	$n_{.2}$	n

Table 5.105: **Dichotomous example of a table of contingencies**

or written in another way:

		Y		
		$Y=1$	$Y=0$	**Total**
X	$X=1$	a	b	$a+b$
	$X=0$	c	d	$c+d$
	Total	$a+c$	$b+d$	1

Let us recall that the Pearson's correlation coefficient:

$$\rho_{X,Y}=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}=\rho_{X,Y}=\frac{\mathrm{E}[XY]-\mathrm{E}[X]\,\mathrm{E}[Y]}{\sqrt{\mathrm{E}[X^2]-[\mathrm{E}[X]]^2}\,\sqrt{\mathrm{E}[Y^2]-[\mathrm{E}[Y]]^2}} \tag{5.7.2680}$$

Now let X and Y be Bernoulli random variables (see page 652). We don't assume independence or identical distribution, but we do assume that all 4 of the following probabilities are non-zero.

Let:

$$a:=P[X=1,Y=1],b:=P[X=1,Y=0],c:=P[X=0,Y=1],d:=P[X=0,Y=0] \tag{5.7.2681}$$

We have first:

$$\begin{aligned}\mathrm{E}[X]&=a+b=p\\ \mathrm{E}[Y]&=a+c=q\end{aligned}\tag{5.7.2682}$$

We also have:

$$\begin{aligned}\mathrm{V}(X)&=\mathrm{E}[(X-\mathrm{E}[X])^2]=\mathrm{E}[(X-p)^2]=p(1-p)^2+(1-p)(-p)^2\\ &=p(1-2p+p^2)+p^2-p^3=p-2p^2+p^3+p^2-p^3\\ &=p-p^2=p(1-p)\end{aligned}\tag{5.7.2683}$$

and:

$$\begin{aligned}\sigma_X&=\sqrt{\mathrm{V}(X)}=\sqrt{p(1-p)}=\sqrt{(a+b)(1-(a+b))}\\ \sigma_Y&=\sqrt{\mathrm{V}(Y)}=\sqrt{q(1-q)}=\sqrt{(a+c)(1-(a+c))}\end{aligned}\tag{5.7.2684}$$

and:

$$\operatorname{cov}(X,Y)=\mathrm{E}[XY]-\mathrm{E}[X]\mathrm{E}[Y]=a-pq=a-(a+b)(a+c)\tag{5.7.2685}$$

Finally, by substitution into the equation for ρ_{XY}:

$$\begin{aligned}\rho_{XY}&=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}\\ &=\frac{a-(a+b)(a+c)}{\sqrt{(a+b)(1-(a+b))}\sqrt{(a+c)(1-(a+c))}}\\ &=\frac{a-(a+b)(a+c)}{\sqrt{(a+b)(1-(a+b))(a+c)(1-(a+c))}}\end{aligned}\tag{5.7.2686}$$

Doing the right substitutions, this lead us to:

$$\boxed{\phi=\frac{nn_{11}-n_{1.}n_{.1}}{\sqrt{n_{1.}(n-n_{1.})n_{.1}(n-n_{.1})}}}\tag{5.7.2687}$$

If we replace n by $n_{11}+n_{12}+n_{22}+n_{21}$ and $n_{1.}$ by $n_{11}+n_{12}$ and $n_{.1}$ by $n_{11}+n_{21}$ then we get after some very easy simplification:

$$\phi=\frac{n_{11}n_{22}-n_{12}n_{21}}{\sqrt{(n_{11}+n_{12})(n_{21}+n_{22})(n_{11}+n_{21})(n_{12}+n_{22})}}\tag{5.7.2688}$$

The interpretation for the phi coefficient is similar to the Pearson correlation coefficient. The range is from -1 to $+1$, where:

- 0 is no relationship
- $+1$ is a perfect positive relationship: most of your data falls along the diagonal cells.
- -1 is a perfect negative relationship: most of your data is not on the diagonal

The Political Science Department at Quinnipiac University posted this useful list of the meaning of Pearson's Correlation coefficients. The same rule of thumb can be used for the Phi coefficient. Note that these are "crude estimates" for interpreting strengths of relationships:

Interval	Interpretation
$0.7 \leq \varphi \leq 1$	Very strong positive relationship
$0.4 \leq \varphi < 0.70$	Strong positive relationship
$0.3 \leq \varphi < 0.4$	Moderate positive relationship
$0.2 \leq \varphi < 0.3$	Weak positive relationship
$0.0 < \varphi < 0.2$	No or negligible positive relationship
$\varphi = 0$	No relationship
$0.0 > \varphi > -0.2$	No or negligible negative relationship
$-0.2 > \varphi > -0.3$	Weak negative relationship
$-0.3 > \varphi > -0.4$	Moderate negative relationship
$-0.4 > \varphi > -0.70$	Strong negative relationship
$-0.7 > \varphi \geq 1$	Very strong negative relationship

Table 5.106: φ **Pearson correlation coefficient**

Although computationally the Pearson correlation coefficient reduces to the Pearson's phi coefficient in the 2×2 case, they are not in general the same. The Pearson correlation coefficient ranges from -1 to $+1$, where ± 1 indicates perfect agreement or disagreement, and 0 indicates no relationship. The phi coefficient has a maximum value that is determined by the distribution of the two variables if one or both variables can take on more than two values.

✍**Example:**

Consider the following table (even if the conditions are not met for a chi-square test of independence):

Projects	**Certified Project Manager**	**Non-Certified Project Manager**	**Total**
Deadlines met	8	1	9
Deadlines not met	4	5	9
Total	12	6	18

Table 5.107: **Squared rectangular contingency table**

Then we have:

$$\phi = \frac{nn_{11} - n_{1.}n_{.1}}{\sqrt{n_{1.}(n-n_{1.})n_{.1}(n-n_{.1})}} = \frac{18 \cdot 8 - 9 \cdot 12}{\sqrt{9(18-9) \cdot 12(18-12)}} \cong 0.471 \qquad (5.7.2689)$$

We can prove that Pearson's phi coefficient is equal to Cramèr's phi coefficient seen just before (that's by the way why the notation is the same!) and that was given for recall:

$$\varphi = \sqrt{\frac{\chi_1^2}{N}} \qquad (5.7.2690)$$

To simplify the notation, let us put:

$$n_{11}=a, n_{12}=b, n_{21}=c, n_{22}=d \tag{5.7.2691}$$

We then have for a contingency table of size 2×2:

$$\chi^2=\sum_{l=1}^{2}\sum_{c=1}^{2}\frac{(n_{lc}-e_{lc})^2}{e_{lc}}=\frac{(a-e_{11})^2}{e_{11}}+\frac{(b-e_{12})^2}{e_{12}}+\frac{(c-e_{21})^2}{e_{21}}+\frac{(d-e_{22})^2}{e_{22}} \tag{5.7.2692}$$

With same as for our study of the Cramèr's V coefficient:

$$e_{11}=\frac{(a+b)(a+c)}{n}, e_{12}=\frac{(a+b)(b+d)}{n}, e_{21}=\frac{(a+c)(c+d)}{n}, e_{22}=\frac{(b+d)(c+d)}{n} \tag{5.7.2693}$$

Which implies:

$$\begin{aligned}\chi^2&=\frac{\left(a-\dfrac{(a+b)(a+c)}{n}\right)^2}{\dfrac{(a+b)(a+c)}{n}}+\frac{\left(b-\dfrac{(a+b)(b+d)}{n}\right)^2}{\dfrac{(a+b)(b+d)}{n}}+\frac{\left(c-\dfrac{(a+c)(c+d)}{n}\right)^2}{\dfrac{(a+c)(c+d)}{n}}+\frac{\left(d-\dfrac{(b+d)(c+d)}{n}\right)^2}{\dfrac{(b+d)(c+d)}{n}}\\&=\frac{\dfrac{1}{n^2}}{\dfrac{1}{n}}\left[\frac{(na-(a+b)(a+c))^2}{(a+b)(a+c)}+\frac{(nb-(a+b)(b+d))^2}{(a+b)(b+d)}+\frac{(nc-(a+c)(c+d))^2}{(a+c)(c+d)}+\frac{nd-(b+d)(c+d))^2}{(b+d)(c+d)}\right]\end{aligned} \tag{5.7.2694}$$

But we have:

$$\begin{aligned}na-(a+b)(a+c)&=(a+b+c+d)a-(a+b)(a+c)\\&=a^2+ab+ac+ad-a^2-ac-ba-bc=ad-bc\end{aligned} \tag{5.7.2695}$$

and:

$$\begin{aligned}nb-(a+b)(b+d)&=(a+b+c+d)b-(a+b)(b+d)\\&=ab+b^2+bc+bd-ab-ad-b^2-bd=bc-ad\end{aligned} \tag{5.7.2696}$$

and:

$$\begin{aligned}nc-(a+c)(c+d)&=(a+b+c+d)c-(a+c)(c+d)\\&=ac+bc+c^2+cd-ac-ad-c^2-cd=bc-ad\end{aligned} \tag{5.7.2697}$$

and:

$$\begin{aligned}nd-(b+d)(c+d)&=(a+b+c+d)d-(b+d)(c+d)\\&=ad+bd+dc+d^2-bc-bd-dc-d^2=ad-bd\end{aligned} \tag{5.7.2698}$$

Hence:

$$\begin{aligned}\chi^2&=\frac{1}{n}\left[\frac{(ad-bc)^2}{(a+b)(a+c)}+\frac{(bc-ad)^2}{(a+b)(b+d)}+\frac{(bc-ad)^2}{(a+c)(c+d)}+\frac{(ad-bc)^2}{(b+d)(c+d)}\right]\\&=\frac{(ad-bc)^2}{n}\left[\frac{1}{(a+b)(a+c)}+\frac{1}{(a+b)(b+d)}+\frac{1}{(a+c)(c+d)}+\frac{1}{(b+d)(c+d)}\right]\\&=\frac{(ad-bc)^2}{n}\frac{(b+d)(c+d)+(a+c)(c+d)+(a+b)(b+d)+(a+c)(a+b)}{(a+c)(a+b)(b+d)(c+d)}\end{aligned} \tag{5.7.2699}$$

Notice that:

$$\begin{aligned}(b+d)(c+d)&+(a+c)(c+d)+(a+b)(b+d)+(a+c)(a+b)\\&=dc+d^2+ac+ad+c^2+cd+ab+ad+b^2+bd+a^2+ab+ca+bc\\&=a^2+b^2+c^2+d^2+2bc+2bd+2dc+2ac+2ab+2ad\\&=(a+b+c+d)^2=n^2\end{aligned}\tag{5.7.2700}$$

This last result bring us to:

$$\begin{aligned}\chi^2&=\frac{(ad-bc)^2}{n}\frac{n^2}{(a+c)(a+b)(b+d)(c+d)}=n\frac{(ad-bc)^2}{(a+c)(a+b)(b+d)(c+d}\\&=n\varphi^2\end{aligned}\tag{5.7.2701}$$

Hence:

$$\phi=\sqrt{\frac{\chi^2}{n}}\tag{5.7.2702}$$

Pearson's phi coefficient is the source of inspiration of "Matthews correlation coefficient" (M.C.C.) that is used in Machine Learning as a measure of the quality of binary (two-class) classifications. Only the vocabulary differs. Indeed, it is defined as:

$$\boxed{\text{MCC}=\frac{\text{TP}\cdot\text{TN}-\text{FP}\cdot\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}}\tag{5.7.2703}$$

While there is no perfect way of describing the confusion matrix $\mathcal{C}$ of true and false positives and negatives by a single number (we will see in the chapter of Numerical Methods what are "confusion matrices"!), the Matthews correlation coefficient is generally regarded as being one of the best such measures in this beginning of the 121st century (holocene calendar).

In this relation, TP is the number of true positives, TN the number of true negatives, FP the number of false positives and FN the number of false negatives. If any of the four sums in the denominator is zero, the denominator can be arbitrarily set to one; this results in a Matthews correlation coefficient of zero, which can be shown to be the correct limiting value.

5.7.15.5 Exact Fisher's Test

When the effective in the contingency table are too small or that the values are really too irregular, the use of chi-square (Pearson's test) is not possible because the application conditions are no longer valid. We will see that the Fisher exact test can be formalized analytically in 2×2 contingency tables (the majority of statistical software only support this particular scenario for the Fisher exact test) otherwise what must be used Monte Carlo simulations.

The principle of the "Fisher's exact test" (usable both in bilateral or unilateral even if this latter is far more common in practice), so based on the crossover frequency, is to determine whether the pattern observed in the table contingency is an extreme situation with respect to all possible situations. As we will prove it, this test has the special property that any cell of the table can be referred for the test because the underlying distributions (marginals) of probability are equivalent.

To study this test, as often in this section, we use directly an example as theoretical companion.

Let us consider the following contingency table (which now is known to us ...):

Projects	**Certified Project Manager**	**Non-Certified Project Manager**	**Total**
Deadlines met	8	1	9
Deadlines not met	4	5	9
Total	12	6	18

Table 5.108: **Squared rectangular contingency table**

which is not really suitable for chi-square test of independence since the content of the cells is less than 10 units and the number of degrees of freedom would be equal to unity.

The same table in percentages will give (even if it is unnecessary for the study of this test):

Projects	**Certified Project Manager**	**Non-Certified Project Manager**	**Total**
Deadlines met	88.88%	11.12%	50%
Deadlines not met	44.44%	55.56%	50%
Total	66.66%	33.34%	100%

The theoretical effectives are given by (even if it is as useless to study this test but statistical software also frequently communicate these values):

Projects	**Certified Project Manager**	**Non-Certified Project Manager**	**Total**
Deadlines met	$(12/18)\cdot 9=6$	$(6/18)\cdot 9=3$	9
Deadlines not met	$(12/18)\cdot 9=6$	$(6/18)\cdot 9=3$	9
Total	12	6	18

The question we will begin to ask ourselves is the following: knowing the totals for each row and each column, what is the probability of having the values present in each cell?!

This question can be reformulated if we change the table in the following generic form:

Projects	**Certified Project Manager**	**Non-Certified Project Manager**	**Total**
Deadlines met	$a=k$	b	$a+b=m$
Deadlines not met	c	d	$c+d$
Total	$a+c=p$	$b+d$	$a+b+c+d=n$

Explicitly and relatively to our example by adopting the notation in use of the hypergeometric law, the question is to know what is the probability of having 8 ($a=k$) projects among the 18

(n) whose deadlines were met by certified project managers knowing that there are 9 projects (m) at total whose deadlines were met and 12 projects (p) at total managed by certified project managers.

We saw earlier that in this case it is an exhaustive drawing, we then have to use the hypergeometric law given by for recall by as proven earlier above by:

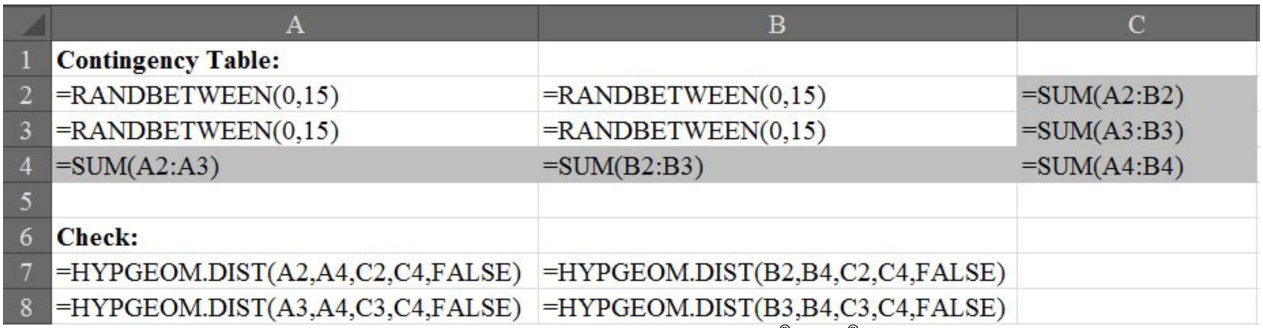
$$P(n,p,m,k)=\mathcal{H}(k)=\frac{C_k^m C_{p-k}^{n-m}}{C_p^n}=\frac{\dfrac{m!}{k!(m-k)!}\dfrac{(n-m)!}{(p-k)!((n-m)-(p-k))!}}{\dfrac{n!}{p!(n-p)!}} \qquad (5.7.2704)$$

Thus with of Microsoft® Excel® 11.8346:

```
=HYPGEOM.DIST(k,p,m,n,FALSE)
=HYPGEOM.DIST(8,12,9,18,FALSE)=0.06108597
```

where for recall, k is the number of successes in the sample, p is the sample size, m the number of successes in the population and n the size of the population.

In fact, as already mentioned, the probabilities are all equal regardless of the selected cell of the contingency table !! This can be checked numerically for sceptics using again a spreadsheet software like Microsoft® Excel® 14.0.6123 by creating the following structure:



	A	B	C
1	**Contingency Table:**		
2	=RANDBETWEEN(0,15)	=RANDBETWEEN(0,15)	=SUM(A2:B2)
3	=RANDBETWEEN(0,15)	=RANDBETWEEN(0,15)	=SUM(A3:B3)
4	=SUM(A2:A3)	=SUM(B2:B3)	=SUM(A4:B4)
5			
6	**Check:**		
7	=HYPGEOM.DIST(A2,A4,C2,C4,FALSE)	=HYPGEOM.DIST(B2,B4,C2,C4,FALSE)	
8	=HYPGEOM.DIST(A3,A4,C3,C4,FALSE)	=HYPGEOM.DIST(B3,B4,C3,C4,FALSE)	

Figure 5.246: **Exact Fisher test symmetry with Microsoft® Excel® 14.0.6123**

and therefore each time that the reader presses the F9 keyboard key he can check that all probabilities are always equal.

This can also be checked formally by selecting a cell of the table and writing:

$$P(n,p,m,k)=\frac{C_k^m C_{p-k}^{n-m}}{C_p^n}=\frac{\dfrac{m!}{k!(m-k)!}\dfrac{(n-m)!}{(p-k)!((n-m)-(p-k))!}}{\dfrac{n!}{p!(n-p)!}} \qquad (5.7.2705)$$

and for another cell of the same column, we will have:

$$P(n,p,n-m,p-k)=\frac{\dfrac{(n-m)!}{(p-k)!((n-m)-(p-k))!}\dfrac{(n-(n-m))!}{(p-(p-k))!((n-(n-m))-(p-(p-k)))!}}{\dfrac{n!}{p!(n-p)!}}$$
$$=\frac{\dfrac{(n-m)!}{(p-k)!((n-m)-(p-k))!}\dfrac{m!}{k!(m-k)!}}{\dfrac{n!}{p!(n-p)!}} \tag{5.7.2706}$$

and therefore:

$$P(n,p,m,k)=P(n,p,n-m,p-k) \tag{5.7.2707}$$

and so on...

Anyway, that being said, so we have in the upper left cell the value 8 while the theoretical effective is 6. The first thing we can answer is whether this value of 8 is unusually large or not compared the theoretical effective. For this, we calculate by example the unilateral cumulative probability of being less than or equal to 8. We have then with Microsoft® Excel® 14.0.6123 and later (the last parameter of 1 - equivalent to `TRUE` - of the function indicate to the software we want the cumulative probability):

```
=HYPGEOM.DIST(8,12,9,18,1)=0.995475113
```

It therefore appears with a threshold of 5% in unilateral that this value is unusually large. We are then in an extreme situation.

By cons, even if the probabilities are equal for all the cells, the cumulative probability is not! Thus we have for example for the value of the lower left cell (to check if it is abnormally small compared to the theoretical effective of 6):

```
=HYPGEOM.DIST(4,12,9,18,1)=0.06561086
```

So this is a value that is not abnormally small. However, we would like to have a test to conclude if the entire table is or is not an extreme configuration. However, by making the calculation cell by cell, we will not get too much...

The idea then is (at least on the "paper") to build all tables whose marginal frequencies are $9;9$ and $12;6$ and to calculate the probability of a given cell (the advantage of this technique is that the conclusion will be the same regardless the chosen cell as reference for the calculations):

Projects	**Certified Project Manager**	**Non-Certified Project Manager**	**Total**
Deadlines met	9	0	9
Deadlines not met	3	6	9
Total	12	6	18

`=HYPGEOM.DIST(9,12,9,18)=0.004524887`

Projects	**Certified Project Manager**	**Non-Certified Project Manager**	**Total**
Deadlines met	8	1	9
Deadlines not met	4	5	9
Total	12	6	18

`=HYPGEOM.DIST(8,12,9,18)=0.06108597`

Projects	**Certified Project Manager**	**Non-Certified Project Manager**	**Total**
Deadlines met	7	2	9
Deadlines not met	5	4	9
Total	12	6	18

`=HYPGEOM.DIST(7,12,9,18)=0.244343891`

Projects	**Certified Project Manager**	**Non-Certified Project Manager**	**Total**
Deadlines met	6	3	9
Deadlines not met	6	3	9
Total	12	6	18

`=HYPGEOM.DIST(6,12,9,18)=0.380090498`

Projects	**Certified Project Manager**	**Non-Certified Project Manager**	**Total**
Deadlines met	5	4	9
Deadlines not met	7	2	9
Total	12	6	18

`=HYPGEOM.DIST(5,12,9,18)=0.244343891`

Projects	**Certified Project Manager**	**Non-Certified Project Manager**	**Total**
Deadlines met	4	5	9
Deadlines not met	8	1	9
Total	12	6	18

`=HYPGEOM.DIST(4,12,9,18)=0.061085973`

Projects	**Certified Project Manager**	**Non-Certified Project Manager**	**Total**
Deadlines met	3	6	9
Deadlines not met	9	0	9
Total	12	6	18

`=HYPGEOM.DIST(3,12,9,18)=0.004524887`

To summarize with the values of k (corresponding to the top left cell):

k	**Probability**
9	0.00452489
8	0.06108597
7	0.24434389
6	0.3800905
5	0.24434389
4	0.06108597
3	0.00452489
Sum:	1

Table 5.109: **Probability of the hypergeometric law corresponding to the combination**

As in the column of the original table which we just worked with the smallest value is 4 and the biggest 8, we will take the tail probabilities to get what is the p-value to be greater or equal 8 and less than or equal 4 (this is therefore a bilateral test). Then we have:

k	**Probability**
9	0.00452489
8	0.06108597
4	0.06108597
3	0.00452489
Sum:	0.131221719

Table 5.110: **Selection of values of interest**

Therefore the p-value is 13.12%. We can therefore not say that our original table is in an extremal configuration if we chose an empirical threshold value of 5%. Many softwares only communicate the p-value.

Remark

The choice of the bounds in this test are subject to debate. Indeed, if we chose for example to focus on the probability to be in the closed bounded interval of 4 and 8 (inclusive!), we would have a result of 99.09% and then we should consider that we are in an extreme configuration if we made the 5% empirical threshold choice. Therefore the choice of the bounds with a discrete distribution like the Hypergeometric law many times something not obvious at the opposite of a test based on a continuous distribution that suffers absolutely not of this kind of problems. The majority of softwares that we know at this date take an open bounder interval (this corresponds therefore to the first calculation we made with the p-value of 13.12%.

Finally, to close this subject, the reader can check that he will find the same results whatever he takes for reference cell.

Let us also indicate that the Fisher exact test can obviously be used in Machine Learning to measure the performance of a binary classification!!!!!

5.7.15.6 Cohen's kappa agreement

If our judgements reflect our thinking, they are rarely in agreement with those of others.

This inter-individual variability beneficial to humans, however, is disadvantageous in many scientific disciplines, where it is often necessary to assess and improve the agreement between similar information applied to the same object in the context of a quality control requirement or in sensory analysis.

The nonparametric Cohen's kappa test allows for example to quantify the binary (dichotomous) agreement between two or more researchers or technicians when the judgements are qualitative.

This kappa statistic is frequently used to test "inter-rater reliability". The importance of rater reliability lies in the fact that it represents the extent to which the data collected in the study are correct representations of the variables measured. Measurement of the extent to which data collectors (raters) assign the same score to the same variable is named "interrater reliability". While there have been a variety of methods to measure interrater reliability, traditionally it was measured as percent agreement, calculated as the number of agreement scores divided by the total number of scores. In 11960 (holocene calendar), Jacob Cohen critiqued use of percent agreement due to its inability to account for chance agreement. He introduced the Cohen's kappa, developed to account for the possibility that raters actually guess on at least some variables due to uncertainty. Like most correlation statistics, the kappa can range from -1 to $+1$. While the kappa is one of the most commonly used statistics to test interrater reliability, it has limitations. Judgements about what level of kappa should be acceptable for health research are questioned. Cohen's suggested interpretation may be too lenient for health related studies because it implies that a score as low as 0.41 might be acceptable. Kappa and percent agreement are compared, and levels for both kappa and percent agreement that should be demanded in healthcare studies are suggested.

Many situations in the healthcare industry rely on multiple people to collect research or clinical laboratory data. The question of consistency, or agreement among the individuals collecting data immediately arises due to the variability among human observers. Well-designed research studies must therefore include procedures that measure agreement among the various data collectors. Interrater reliability is a concern to one degree or another in most large studies due to the fact that multiple people collecting data may experience and interpret the phenomena of interest differently.

Remark

There are a number of statistics that have been used to measure interrater and intrarater reliability. A partial list includes percent agreement, Cohen's kappa (for two raters), the Fleiss kappa (adaptation of Cohen's kappa for three or more raters) the contingency coefficient, the Pearson R and the Spearman rho, the intra-class correlation coefficient, the concordance correlation coefficient, and Krippendorff's alpha (useful when there are multiple raters and multiple possible ratings). Use of correlation coefficients such as Pearson's R may be a poor reflection of the amount of agreement between raters resulting in extreme over or underestimates of the true level of rater agreement.

Let us take the case in the medical field where two or more practitioners examining the same patient or even offer different diagnoses or different therapeutic decisions. In the absence of a reference, this proliferation of advices does not bring the expected security of a perfect diagnostic or therapeutic agreement for the physician and the patient. It is therefore important that the agreement in a team or between teams is the best possible guarantee for the quality and continuity of care.

One solution here is to realize a session of "concordance experiment" between physicians to estimate their rate of agreement using the kappa coefficient and study their disagreements to address them.

To illustrate the concepts, let us consider the very important case of two quality managers that analysed 11 pieces to reject or accept them. They get:

		Bob		
		Rejected	**Accepted**	**Total**
Alice	**Rejected**	3	2	5
	Accepted	1	5	6
	Total	4	7	11

Table 5.111: **Dichotomous example of table of contingencies**

The theoretical frequencies being obtained always by a rule of three (same calculation of rule of three that for the chi-square and Fisher test of independence exact seen above):

		Bob		
		Rejected	**Accepted**	**Total**
Alice	**Rejected**	$(4/11)\cdot 5=1.82$	$(7/11)\cdot 5=3.18$	5
	Accepted	$(4/11)\cdot 6=2.18$	$(7/11)\cdot 6=3.82$	6
	Total	4	7	11

Table 5.112: **Theoretical frequencies**

Cohen's kappa is defined by the ratio:

$$\boxed{\kappa_C=\frac{\text{Observed}_{\text{According}}-\text{Expected}_{\text{According}}}{\text{Trials}-\text{Expected}_{\text{According}}}} \tag{5.7.2708}$$

Therefore in our example:

$$\kappa_C=\dfrac{(3+5)-\left(\dfrac{4\cdot 5}{11}+\dfrac{7\cdot 6}{11}\right)}{11-\left(\dfrac{4\cdot 5}{11}+\dfrac{7\cdot 6}{11}\right)}=\dfrac{(3+5)-(1.82+3.82)}{11-(1.82+3.82)}=0.441 \qquad (5.7.2709)$$

This value of 0.441 indicates a moderate agreement between the two individuals (Alice and Bob).

If instead of having the frequencies, we work in percentages (proportions) of the total, the Kappa is then:

$$\kappa_C=\dfrac{\%\text{Observed}_{\text{According}}-\%\text{Expected}_{\text{According}}}{1-\%\text{Expected}_{\text{According}}}=\dfrac{p_{o,a}-p_{e,a}}{1-p_{e,a}} \qquad (5.7.2710)$$

Which will give obviously in our example:

$$\kappa_C=\dfrac{\dfrac{3+5}{11}-\dfrac{1}{11^2}(4\cdot 5+7\cdot 6)}{1-\dfrac{1}{11^2}(4\cdot 5+7\cdot 6)}=0.441 \qquad (5.7.2711)$$

That it is usage to write in the following condensed form:

$$\boxed{\kappa_C=\dfrac{\sum\limits_i p_{ii}-\sum\limits_i p_{i+}p_{i+}}{1-\sum\limits_i p_{i+}p_{i+}}=\dfrac{(p_{00}+p_{11})-(p_{0+}p_{0+}+p_{1+}p_{1+})}{1-(p_{0+}p_{0+}+p_{1+}p_{1+})}} \qquad (5.7.2712)$$

with:

$$-1\leq\kappa_C\leq+1 \qquad (5.7.2713)$$

where $+1$ corresponds to a perfect agreement and -1 to a perfect non-agreement. Obviously to have a perfect agreement, we must have the cells (Rejected, Rejected) and (Accepted, Accepted) that are equal and that the other cells are zero.

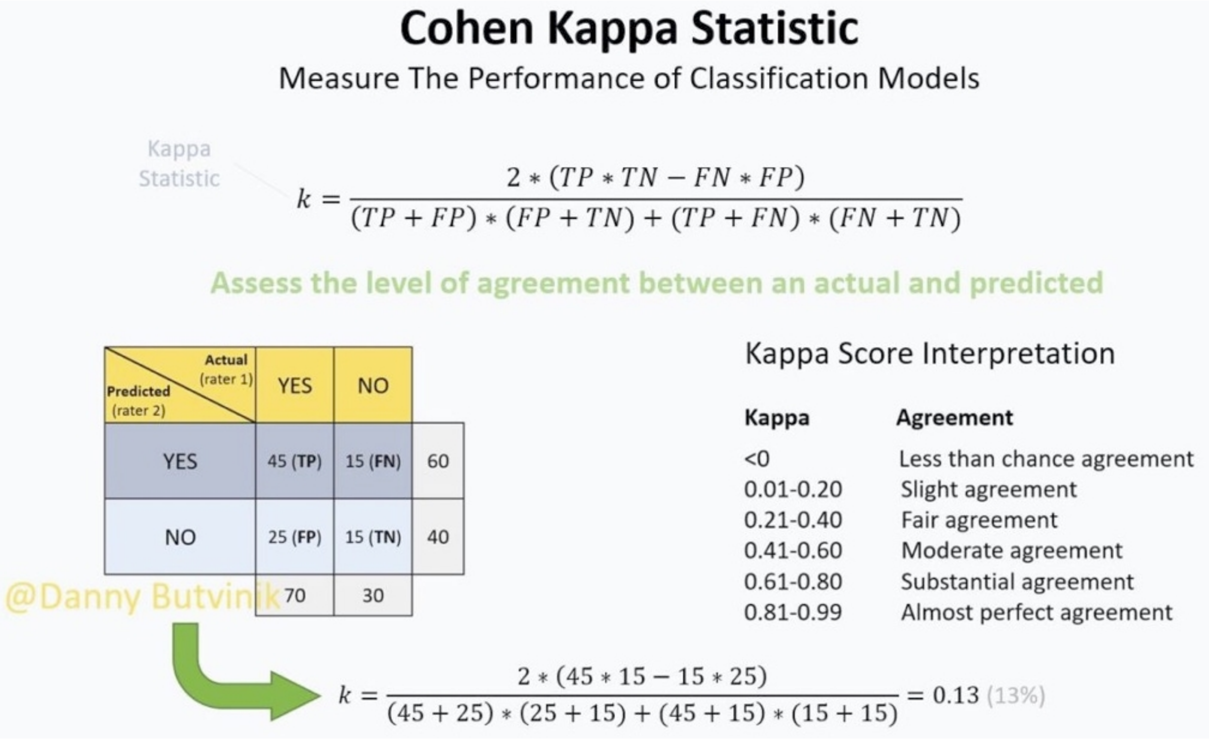
The following table of interpretation of the positive part of the κ_C is of common usage (the negative one having not too much interest...):

Interval	Interpretation
$0.8\leq\kappa\leq 1$	Very good agreement
$0.6\leq\kappa<0.8$	Good agreement
$0.4\leq\kappa<0.6$	Moderated agreement
$0.2\leq\kappa<0.4$	Poor agreement
$0.0\leq\kappa<0.2$	No agreement

Table 5.113: κ_C **usage interpretation table**

However, the practitioner must be very critical (as always!) by using this type of tool. Understanding its construction also helps - also as always - to identify its weaknesses and assumptions that are quite questionable.

Let us finally indicate that the Cohen's Kappa can obviously be used in Machine Learning to measure the performance of a binary classification! A quite well known example in that field is given by the following illustration made by Danny Butvinik:



Cohen Kappa Statistic

Measure The Performance of Classification Models

Kappa Statistic

$$k=\frac{2*(TP*TN-FN*FP)}{(TP+FP)*(FP+TN)+(TP+FN)*(FN+TN)}$$

Assess the level of agreement between an actual and predicted

Predicted (rater 2) \ Actual (rater 1)	YES	NO	
YES	45 (TP)	15 (FN)	60
NO	25 (FP)	15 (TN)	40
	70	30	

@Danny Butvinik

Kappa Score Interpretation

Kappa	Agreement
<0	Less than chance agreement
0.01-0.20	Slight agreement
0.21-0.40	Fair agreement
0.41-0.60	Moderate agreement
0.61-0.80	Substantial agreement
0.81-0.99	Almost perfect agreement

$$k=\frac{2*(45*15-15*25)}{(45+25)*(25+15)+(45+15)*(15+15)}=0.13\ (13\%)$$

Figure 5.247: **Cohen Kappa statistics in binary classification models (author: Danny Butvinik)**

How do we get the relation visible in the figure above?:

$$\kappa=\frac{2(\text{TP}\cdot\text{TN}-\text{FN}\cdot\text{FP})}{(\text{TP}+\text{FP})(\text{FP}+\text{TN})+(\text{TP}+\text{FN})(\text{FN}+\text{TN})} \tag{5.7.2714}$$

where TP are the true positives, FP are the false positives, TN are the true negatives, and FN are the false negatives[89]. It's quite simple (but boring)!

So based on the above table we see that:

$$p_{o,a}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{5.7.2715}$$

and:

$$p_{e,a}=\frac{(\text{TP}+\text{FN})(\text{TP}+\text{FP})+(\text{FP}+\text{TN})(\text{FN}+\text{TN})}{(\text{TP}+\text{TN}+\text{FP}+\text{FN})^2} \tag{5.7.2716}$$

Injecting that into:

$$\kappa_C=\frac{p_{o,a}-p_{e,a}}{1-p_{e,a}} \tag{5.7.2717}$$

---

[89] In this case, Cohen's Kappa is equivalent to the Heidke skill score known in Meteorology. The measure was first introduced by Myrick Haskell Doolittle in 11888 (holocene calendar).

we get the expected result visible in the figure above! Indeed, let us put TP$=a$, TN$=b$, FP$=c$, FN$=d$. Then (on request of reader...):

$$\begin{aligned}
\kappa_C=\frac{p_{o,a}-p_{e,a}}{1-p_{e,a}}&=\frac{\dfrac{a+b}{a+b+c+d}-\dfrac{(a+d)(a+c)+(c+b)(d+b)}{(a+b+c+d)^2}}{1-\dfrac{(a+d)(a+c)+(c+b)(d+b)}{(a+b+c+d)^2}}\\
&=\frac{\dfrac{a+b}{a+b+c+d}-\dfrac{(a^2+ac+ad+cd)+(b^2+bc+bd+cd)}{(a+b+c+d)^2}}{1-\dfrac{(a^2+ac+ad+cd)+(b^2+bc+bd+cd)}{(a+b+c+d)^2}}\\
&=\frac{\dfrac{a+b}{a+b+c+d}-\dfrac{a^2+ac+ad+b^2+bc+bd+2cd}{(a+b+c+d)^2}}{1-\dfrac{a^2+ac+ad+b^2+bc+bd+2cd}{(a+b+c+d)^2}}\\
&=\frac{\dfrac{a+b}{a+b+c+d}-\dfrac{a^2+ac+ad+b^2+bc+bd+2cd}{a^2+2ab+2ac+2ad+b^2+2bc+2bd+c^2+2cd+d^2}}{1-\dfrac{a^2+ac+ad+b^2+bc+bd+2cd}{a^2+2ab+2ac+2ad+b^2+2bc+2bd+c^2+2cd+d^2}}\\
&=\frac{\dfrac{((a+b)\cdot(a+b+c+d))-(a^2+ac+ad+b^2+bc+bd+2cd)}{(a+b+c+d)\cdot(a+b+c+d)}}{\dfrac{(a^2+2ab+2ac+2ad+b^2+2bc+2bd+c^2+2cd+d^2)-(a^2+ac+ad+b^2+bc+bd+2cd)}{a^2+2ab+2ac+2ad+b^2+2bc+2bd+c^2+2cd+d^2}}\\
&=\frac{\dfrac{2ab-2cd}{a^2+2ab+2ac+2ad+b^2+2bc+2bd+c^2+2cd+d^2}}{\dfrac{2ab+ac+ad+bc+bd+c^2+d^2}{a^2+2ab+2ac+2ad+b^2+2bc+2bd+c^2+2cd+d^2}}\\
&=\frac{2ab-2cd}{2ab+ac+ad+bc+bd+c^2+d^2}=\frac{2(ab-cd)}{(a+c)(c+b)+(a+d)(d+b)}
\end{aligned}\tag{5.7.2718}$$

That's it!

5.7.15.7 McNemar's test

McNemar's test may well be calculated along the Cohen's Kappa (the first being a statistical hypothesis test and the second one only an empirical point estimator of concordance). The idea is that under the null hypothesis H_0 (in this case named "symmetry hypothesis"), one of the diagonals of the table should have equal values. In other words in the form of proportions and only by focusing on one of the two diagonals:

$$H_0: p_{11}=p_{22}\tag{5.7.2719}$$

or of frequencies:

$$H_0: n_{11}=n_{22}\tag{5.7.2720}$$

Knowing that:

$$n_{11}+n_{22}=n\tag{5.7.2721}$$

and under the condition that n is large enough, we can write based on a binomial law whose behaviour is asymptotically Normal:

$$Z=\frac{\dfrac{n_{11}}{n}-0.5}{\sqrt{\dfrac{0.5(1-0.5)}{n}}} \tag{5.7.2722}$$

We can realize that this is equivalent to write:

$$Z=\frac{\dfrac{n_{11}}{n}-\dfrac{1}{2}\dfrac{n_{11}+n_{22}}{n}}{\sqrt{\dfrac{\left(\dfrac{n_{11}+n_{22}}{4n}\right)}{n}}}=\frac{n_{11}-\left(\dfrac{n_{11}+n_{22}}{2}\right)}{\dfrac{\sqrt{n_{11}+n_{22}}}{2}} \tag{5.7.2723}$$

In the literature, we often find this last relation in the form:

$$Z=\frac{b-\left(\dfrac{b+c}{2}\right)}{\dfrac{\sqrt{b+c}}{2}}=\frac{b-c}{\sqrt{b+c}} \tag{5.7.2724}$$

To return to our original relation, some take the square and then approximate the square of Z as a chi-square law (... !) with one degree of freedom (but good in reality approximate a Normal centered reduced distribution by the chi-square law with one degree of freedom is bulls... in our point of view...):

$$\boxed{Z^2=\frac{\left(\dfrac{n_{11}}{n}-0.5\right)^2}{\dfrac{0.5(1-0.5)}{n}}=\frac{(b-c)^2}{b+c}\cong\chi_1^2} \tag{5.7.2725}$$

which is often the relation defined in the books (without proof...) as the "McNemar test".

The test is normally conducted bilaterally. The advantage of the McNemar test is the ease with which we can build a confidence interval of the difference in the diagonal. Indeed, starting from the estimator of the difference:

$$\hat{d}=p_{11}-p_{22}=\frac{n_{11}}{n}-\frac{n_{22}}{n} \tag{5.7.2726}$$

Therefore using the variance (and covariance) proved at page 667, we get:

$$\begin{aligned}\text{V}(\hat{d})&=\text{V}\left(\frac{n_{11}}{n}-\frac{n_{22}}{n}\right)=\frac{1}{n^2}\text{V}(n_{11}-n_{22})=\frac{1}{n^2}\left[\text{V}(n_{11})+\text{V}(n_{22})-2\text{cov}(n_{11},n_{22})\right]\\&=\frac{1}{n^2}\left[np_{11}(1-p_{11})+np_{22}(1-p_{22})-2(-np_{11}p_{22})\right]\\&=\frac{1}{n}\left[p_{11}(1-p_{11})+p_{22}(1-p_{22})+2p_{11}p_{22}\right]\\&=\frac{1}{n}\left[\frac{n_{11}}{n}\left(1-\frac{n_{11}}{n}\right)+\frac{n_{22}}{n}\left(1-\frac{n_{22}}{n}\right)+\frac{n_{11}}{n}\frac{n_{22}}{n}\right]\end{aligned} \tag{5.7.2727}$$

and therefore we can do an approximate confidence intervals of the following if the usual conditions are met:

$$\hat{d}-Z_{\alpha/2}\sqrt{\text{V}(\hat{d})}\leq d\leq\hat{d}+Z_{\alpha/2}\sqrt{\text{V}(\hat{d})} \tag{5.7.2728}$$

That is to say a little bit more explicitly:

$$\boxed{\left(\frac{\hat{n}_{11}}{n}-\frac{\hat{n}_{22}}{n}\right)-Z_{\alpha/2}\sqrt{\text{V}(\hat{d})}\leq\frac{n_{11}}{n}-\frac{n_{22}}{n}\leq\left(\frac{\hat{n}_{11}}{n}-\frac{\hat{n}_{22}}{n}\right)+Z_{\alpha/2}\sqrt{\text{V}(\hat{d})}} \tag{5.7.2729}$$

with as we just proved it:

$$\text{V}(\hat{d})=\frac{1}{n}\left[\frac{n_{11}}{n}\left(1-\frac{n_{11}}{n}\right)+\frac{n_{22}}{n}\left(1-\frac{n_{22}}{n}\right)+\frac{n_{11}}{n}\frac{n_{22}}{n}\right] \tag{5.7.2730}$$

✍**Example:**

During a social audit, a survey is conducted on 200 employees about the work organization. After reorganization, the same question is asked. Can we consider that there has been real changes?

	Yes (before)	**No (before)**
Yes (after)	55	38
No (after)	38	82

Table 5.114: **Paired before/after simple survey**

We then have taking arbitrarily the diagonal $(25,38)$ the for analysis:

$$Z^2=\frac{\left(\dfrac{n_{11}}{n}-0.5\right)^2}{\dfrac{0.5(1-0.5)}{n}}=\frac{(b-c)^2}{b+c}=2.684$$

The majority of statistical software will not give you 2.683 as they apply an empirical correction due to the rough approximation of that is the chi-square law. You will then have on the softwares display the value of 2.285 for Z^2.

The p-value it is given (without correction) with a spreadsheet software like the Microsoft® Excel® 14.0.6123:

`= 1-CHISQ.DIST(2.683, 1, 1)` $=10.14\%$

and with the correction use by statistical software we would get about 13%. So in both cases the p-value is greater than the traditional 5% threshold, we therefore can not reject the null hypothesis H_0 as the difference is large.

We will not calculate in this example the confidence interval of the difference because almost all statistical softwares have different methods for calculating this value.

Finally, to close the subject on the McNemar test, let us indicate an empirical indicator often used that is named the "Yule coefficient" or "Yule's Q" and defined by:

$$Q=\frac{P\,(X \text{ and } Y \text{ are concordant})-P\,(X \text{ and } Y \text{ are discordant})}{P\,(X \text{ and } Y \text{ are concordant})+P\,(X \text{ and } Y \text{ are discordant})}=\frac{ad-bc}{ad+bc}$$

Where a,b,c,d must be positive non null integers corresponding to position values of contingency tables similar to:

	Yes	**No**
Positive	a	b
Negative	c	d

Remark

Yule's Q is just the a 2×2 version of the Goodman and Kruskal's gamma coefficient. We can also obviously calculate an odds ratio using $(a\cdot d)/(b\cdot c)$ (and then calculate its confidence interval according to the derivation at page 6496). We can convert also this odds ratio to a -1 to $+1$ scale using $(\text{OR}-1)/(\text{OR}+1)$.

Yule's Q is always a number between -1 and $+1$. Here is a table with possible interpretation of the values of Yule's Q:

Interval	Interpretation
$Q=0$	No association between the variables
$0\leq \lvert Q\rvert \leq 0.29$	A negligible or very small association
$0.30\leq \lvert Q\rvert \leq 0.49$	A moderate association between the variables
$0.50\leq \lvert Q\rvert \leq 0.69$	A substantial association between the variables
$0.70\leq \lvert Q\rvert < 1$	A very strong association
$\lvert Q\rvert = 1$	A perfect association between the events

Table 5.115: **Yule's Q usage interpretation table**

The "Cochran-Armitage test for trend" (see just below), also named "Chi-square test for trend" in many statistical softwares, is used in categorical data analysis when the aim is to assess for the presence of an association between a variable with 2 categories and a variable with k categories. It modifies the Pearson chi-squared test to incorporate a suspected ordering in the effects of the k categories of the second variable. For example, doses of a treatment can be ordered as 'low', 'medium', and 'high', and we may suspect that the treatment benefit cannot become smaller as the dose increases. The trend test is often used as a genotype-based test for case-control genetic association studies.

5.7.15.7.1 Derivation for Cochran-Armitage's trend test for the 2×3 genotype table

We will derive here a special case as the general one is out of the scope of this textbook. Consider for this that we have sampled R cases and S controls:

	Number of disease alleles			
	0	1	2	**Totals**
Cases	r_0	r_1	r_2	R
Controls	s_0	s_1	s_2	S
Totals	n_0	n_1	n_2	N

We want to test H_0 against H_1. The idea is:

- H_0: All entries in the table are proportional, vs.
- H_1: Within a column, the absolute value of the difference between the probability of an observation being classified as "Case" or "Control" increases monotonically across the table.

We'll work with the difference between the values in the column so we first standardize the rows to have the same sums:

	Number of disease alleles			
	0	1	2	**Totals**
Cases	$S\cdot r_0$	$S\cdot r_1$	$S\cdot r_2$	$R\cdot S$
Controls	$R\cdot s_0$	$R\cdot s_1$	$R\cdot s_2$	$R\cdot S$

We choose a set of scores x_1, x_2, and x_3 and form the test statistics:

$$U=\sum_{i=0}^{2}x_i(Sr_i-Rs_i) \tag{5.7.2731}$$

Under H_0:

$$P(\text{Case}|\text{Count}=i)=P(\text{Control}|\text{Count}=i)=\frac{n_i}{N}\qquad\text{and}\qquad \text{E}(U)=0 \tag{5.7.2732}$$

Letting $S=N-R$, $s_i=n_i-r_i$, and choosing scores $x_0=0$, $x_1=1$, and $x_2=2$, the test statistic becomes:

$$U=N(r_1+2r_2)-R(n_1+2n_2) \tag{5.7.2733}$$

Returning to generic scores x_i, we calculate the variance of U as:

$$\begin{aligned}\text{V}(U)=\text{V}\left(\sum_{i=0}^{2}x_i(Sr_i-Rs_i)\right)&=\text{V}\left(S\sum_{i=0}^{2}x_ir_i-R\sum_{i=0}^{2}x_is_i\right)\\ =&S^2\text{V}\left(S\sum_{i=0}^{2}x_ir_i\right)+R^2\text{V}\left(\sum_{i=0}^{2}x_is_i\right)\\ =&S^2\left[\sum_{i=0}^{2}x_i^2\text{V}(r_i)+2\sum_{i=0}^{1}\sum_{j=i+1}^{2}x_ix_j\text{cov}(r_i,r_j)\right]+R^2\left[\sum_{i=0}^{2}x_i^2\text{V}(s_i)+2\sum_{i=0}^{1}\sum_{j=i+1}^{2}x_ix_j\text{cov}(s_i,s_j)\right]\end{aligned} \tag{5.7.2734}$$

Under H_0 (we use the variance and covariance[90] of the Hypergeometric distribution ignoring the finite population correction factor):

$$\begin{aligned}\text{V}(U)\cong S^2&\left[\sum_{i=0}^{2}x_i^2R\left(\frac{n_i}{N}\right)\left(\frac{N-n_i}{N}\right)-2\sum_{i=0}^{1}\sum_{j=i+1}^{2}x_ix_jR\left(\frac{n_i}{N}\right)\left(\frac{n_j}{N}\right)\right]\\ &+R^2\left[\sum_{i=0}^{2}x_i^2S\left(\frac{n_i}{N}\right)\left(\frac{N-n_i}{N}\right)-2\sum_{i=0}^{1}\sum_{j=i+1}^{2}x_ix_jS\left(\frac{n_i}{N}\right)\left(\frac{n_j}{N}\right)\right]\\ &=\frac{S}{N^2}\left[\sum_{i=0}^{2}x_i^2SRn_i(N-n_i)-2\sum_{i=0}^{1}\sum_{j=i+1}^{2}x_ix_jSRn_in_j\right]\\ &+\frac{R}{N^2}\left[\sum_{i=0}^{2}x_i^2SRn_i(N-n_i)-2\sum_{i=0}^{1}\sum_{j=i+1}^{2}x_ix_jSRn_in_j\right]\\ &=\frac{SR}{N}\left[\sum_{i=0}^{2}x_i^2n_i(N-n_i)-2\sum_{i=0}^{1}\sum_{j=i+1}^{2}x_ix_jn_in_j\right]\end{aligned} \tag{5.7.2735}$$

---

[90] If the reader is not able to find where was our proof of that covariance, you can find it here at page 667

Let $S=N-R$. Choose scores $x_0=0$, $x_1=1$, and $x_2=2$ and the latter relation reduces to:

$$\text{V}(U)\cong\frac{(N-R)R}{N}\left[N(n_1+4n_2)-(n_1+2n_2)^2\right] \tag{5.7.2736}$$

For N large (this justifies the fact that we ignored the finite population correction factor earlier), we then have:

$$\frac{U}{\sigma_U}\cong\frac{N(r_1+2r_2)-R(n_1+2n_2)}{\sqrt{\dfrac{(N-R)R}{N}\left[N(n_1+4n_2)-(n_1+2n_2)^2\right]}}\sim\mathcal{N}(0,1) \tag{5.7.2737}$$

So, using Sasieni's (11997 according to holocene calendar) notation:

$$\boxed{\chi_G^2=\frac{N\left[N(r_1+2r_2)-R(n_1+2n_2)\right]^2}{(N-R)R\left[N(n_1+4n_2)-(n_1+2n_2)^2\right]}\sim\chi_1^2} \tag{5.7.2738}$$

Again we notice that the Cochran-Armitage test for trend use a Wald statistics and therefore belongs to the family of Wald tests.

Remark

For a practical example and implementation of that test our reader can refer to the `R` companion book.

5.7.16 Survival Statistics

Survival analysis is a branch of statistics for analysing the expected duration of time until one or more events happen, such as death in biological organisms, failure in mechanical systems, time to product reorder of customers, etc.

This topic is named "reliability theory" or "reliability analysis" in engineering (see section Industrial Engineering page 8716), "duration analysis" or "duration modelling" in economics, and "event history analysis" in sociology. Survival analysis attempts to answer questions such as: what is the proportion of a population which will survive past a certain time? Of those that survive, at what rate will they die or fail? Can multiple causes of death or failure be taken into account? How do particular circumstances or characteristics increase or decrease the probability of survival?

To answer such questions, it is necessary to define "lifetime". In the case of biological survival, death is unambiguous, but for mechanical reliability, failure may not be well-defined, for there may well be mechanical systems in which failure is partial, a matter of degree, or not otherwise localized in time. Even in biological problems, some events (for example, heart attack or other organ failure) may have the same ambiguity. The theory outlined below assumes well-defined events at specific times; other cases may be better treated by models which explicitly account for ambiguous events.

More generally, survival analysis involves the modelling of time to event data; in this context, death or failure is considered an "event" in the survival analysis literature – traditionally only a single event occurs for each subject, after which the organism or mechanism is dead or broken.

In practice we consider:

- The life tables (see section of Population Dynamics page 7299) to describe the survival times of members of a group.
- The Kaplan-Meier method to also describe the survival times of members of a group (see just below).
- The Cochran-Mantel-Haenzel test (Log-Rank test) to compare the survival times of two or more groups (see just below).
- The Cox's proportional[91] hazard model (Cox regression) to describe the effect of categorical or quantitative variables on survival (see just below).
- The Threshold Regression Models in that also describe the effect of categorical or quantitative variables on survival however without the proportional hazard assumption of the Cox model!

Here is a more exhaustive (i.e. however still incomplete) taxonomy of survivals analysis methods:

---

[91] The Cox proportional regression assumes hazards are proportional (PH), which means that the relative hazard remains constant!

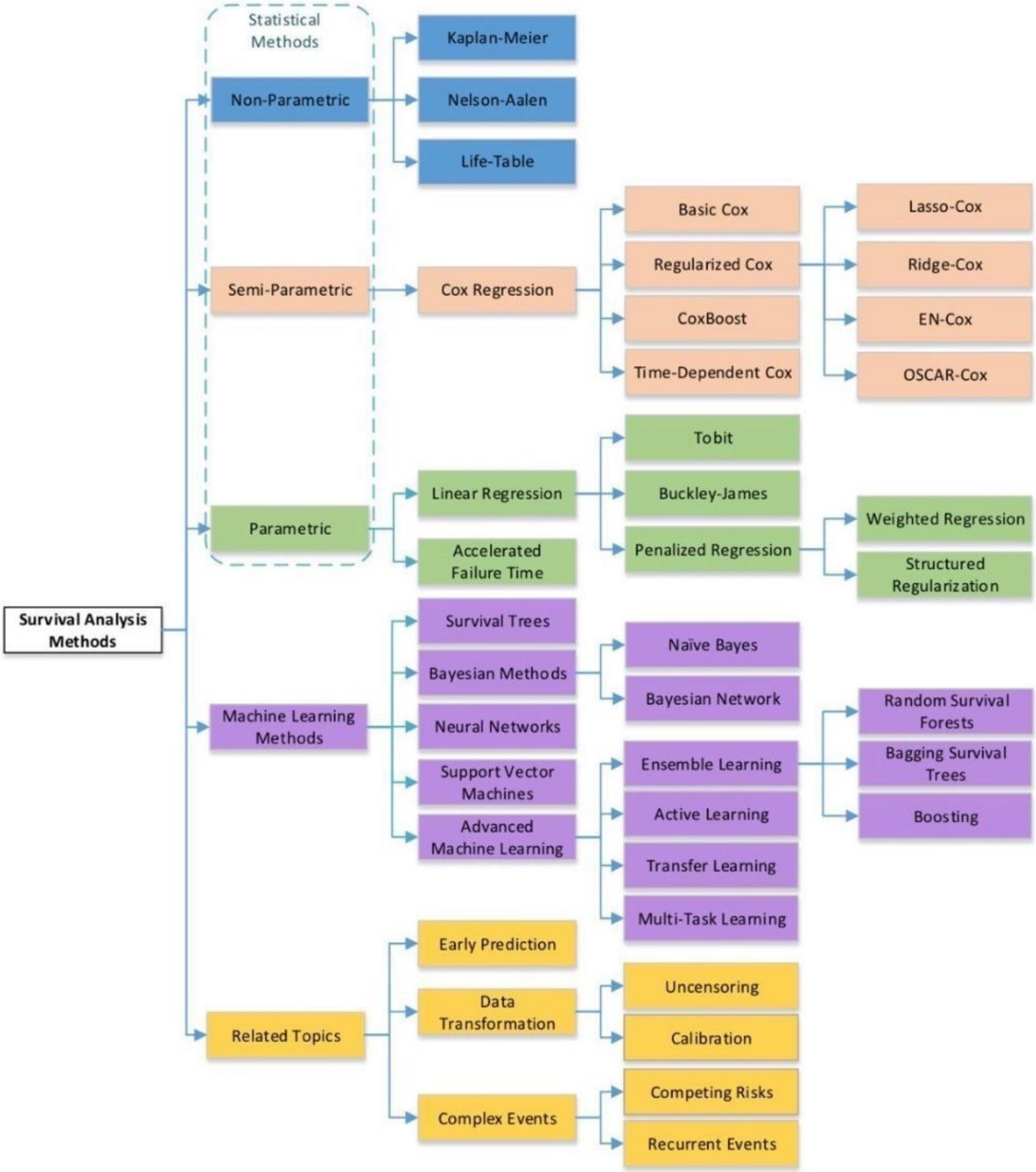

Survival Analysis Methods
Statistical Methods
Non-Parametric
Kaplan-Meier
Nelson-Aalen
Life-Table
Semi-Parametric
Cox Regression
Basic Cox
Regularized Cox
CoxBoost
Time-Dependent Cox
Lasso-Cox
Ridge-Cox
EN-Cox
OSCAR-Cox
Parametric
Linear Regression
Accelerated Failure Time
Tobit
Buckley-James
Penalized Regression
Weighted Regression
Structured Regularization
Machine Learning Methods
Survival Trees
Bayesian Methods
Neural Networks
Support Vector Machines
Advanced Machine Learning
Naïve Bayes
Bayesian Network
Ensemble Learning
Active Learning
Transfer Learning
Multi-Task Learning
Random Survival Forests
Bagging Survival Trees
Boosting
Related Topics
Early Prediction
Data Transformation
Complex Events
Uncensoring
Calibration
Competing Risks
Recurrent Events

Figure 5.248: **Taxonomy of the methods developed for survival analysis (source: ACM Computing Surveys, Vol. 1, No. 1, Article 1, April 12018)**

5.7.16.1 Kaplan-Meier Survival Rate

OK... we don't like to do this in this textbook!!! But for the comfort of our readers and because of some readers feedback let us copy/paste here the theory about the Kaplan-Meier Survival Rate as introduced initially in the section of Industrial Engineering at page 8763.

In areas such as high-level industry, high-level medicine or high-level biology, we are often interested in the:

1. Survival duration after a serious event
2. Duration of remission after treatment or surgery
3. Duration of symptoms after an abnormality
4. Duration of a symptom-less infection

We seek very often to distinguish at least "the event of interest":

1. System shutdown after serious event
2. End of remission
3. End of a symptom after anomaly
4. Beginning of a symptom during an infection

of the variable to explain "duration before the event of interest":

1. Elapsed duration before shutdown
2. Elapsed duration before the end of remission
3. Elapsed duration before the end of the symptom
4. Elapsed duration without symptoms

Definitions (#5.7.159):

D1. We name "remission", the reduction of a disease or a temporarily malfunction.

D2. The "survival time" or "lifetime" T means the time which elapses from an initial time (start of treatment, diagnosis or failure, etc.) until the occurrence of a final event of interest (patient death, relapse, remission, cure, repair, etc.). We say that the object of the study "survives at time t" if at this moment the final interest event has not yet occurred.

Remark

Although this type of study can be associated with preventive maintenance (see section Industrial Engineering page 8716), statisticians associate this type of study rather to the domain of "survival analysis".

We will focus in this textbook on a particular context but that can be easily generalized:

- Cohort/Clinical trial where we study the survival time of each patient (machine).
- All patients (machines) do not have the same observation time (different instants of entry into the study).
- We have information on the survival time of each patient (machine) but we do not know exactly when it happens.

From the last two points, we conclude that the survival time can be censored. So the usual statistical techniques does not apply directly as censored data require special treatment (of course, if we remove the censored data we lose information). It goes without saying that this situation is extremely common and therefore the study of the Kaplan-Meier estimator is very important.

Definition (#5.7.160): The duration T is said to be censored if the duration is not observed completely. The different types of censoring are:

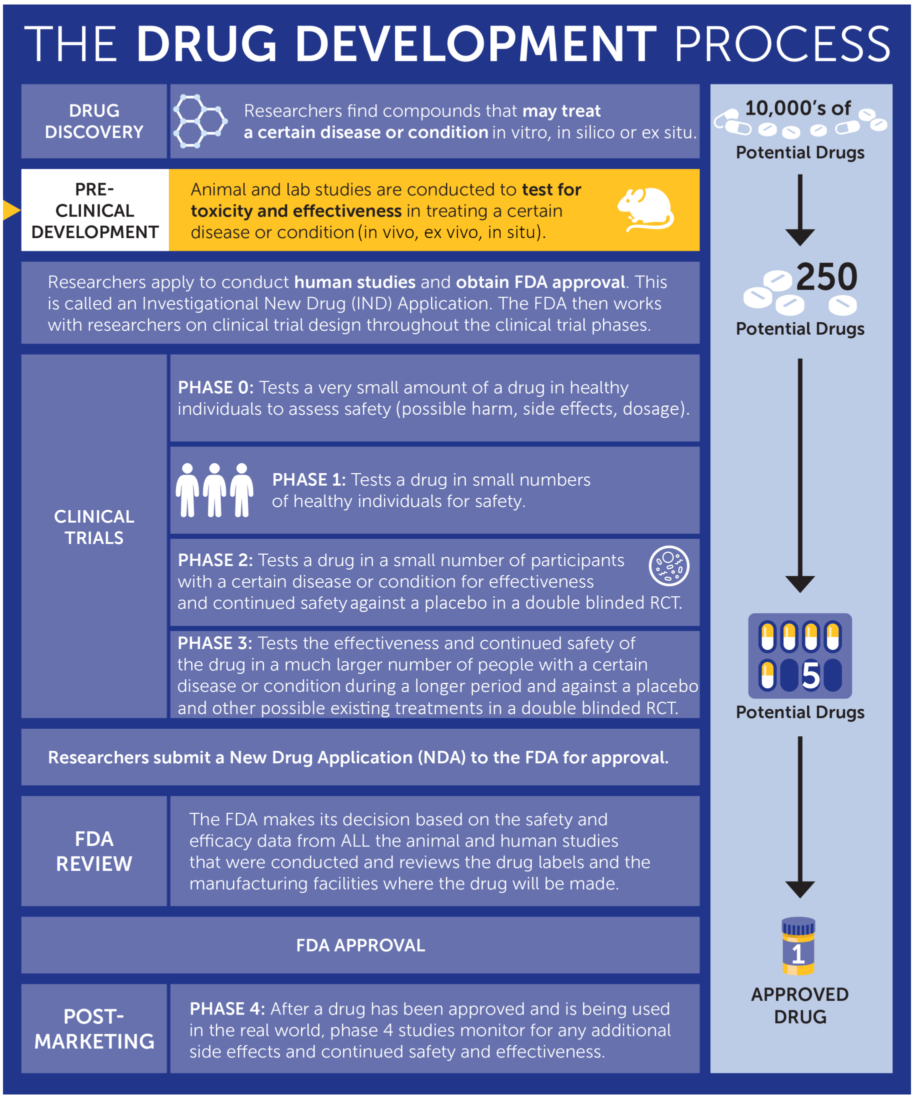
- Type I censoring: fixed right (right censored). In this situation, the time is not observable beyond a maximum, fixed, named "fixed-censoring" and imposed. So either we have the opportunity to observe the real duration of the event of interest for the item if it occurs before the fixed-censoring, or we limit ourselves to the fixed-censoring time if the vent of interest has not occurred before.
- Type I censorship: fixed left (left censored). In this situation, the study duration is not observable before a known date named "waiting censorship" and imposed. So either the event of interest takes place at the time of the wait censorship or after. If applicable later, we will consider the length of time between the date of censorship and the date of the event of interest.
- Type II censoring: wait. In this situation, we observe the lives of n individuals until $m \leq n$ individuals have seen the event to occur (deceased). The duration considered is therefore that of the beginning of the experiment until the event of interest for the m-th.
- Type III censoring: random left. In this situation, we do not know when the event of interest has occurred (because we started to observe the subject of study too late). We can not then deal with "durations" in the measurable sense and we have to limit ourselves to a simple count.
- Type III censoring: random right. In this situation, we do not know when the event will take place (because we stopped to observe the subject before it occurs for any reason). We can not then deal with "durations" in the measurable in the measurable sense and we must limit ourselves to a simple count.

- Type IV censoring: random intervals. In this situation, we have a mixture of random left and right censoring. That is to say that for some study subjects, we do not know when the event of interest has begun, and for others we do not know when it will occur (if any...).

In the machinery industry, we often deal with the type I censoring: fixed right. In the medical field, in clinical trials, we often deal with a type III censoring: random right. In the case of pandemics, we are dealing with type III censoring: random left.

To introduce this subject, rather than doing obscure theory, as always in this textbook we prefer at pragmatic approach. Suppose that the study is a clinical trial involving two groups of patients receiving two types of treatments. Two important questions raised to the physicians in a Phase II clinical trial (phase I is for non-toxicity approval to human and phase 0 for animals):

1. Is one of the two treatments more effective than the other in terms of improving patient survival?
2. Can we highlight prognostic factors, that is to say that improve / deteriorate survival?


THE **DRUG DEVELOPMENT** PROCESS

DRUG DISCOVERY — Researchers find compounds that **may treat a certain disease or condition** in vitro, in silico or ex situ.

PRE-CLINICAL DEVELOPMENT — Animal and lab studies are conducted to **test for toxicity and effectiveness** in treating a certain disease or condition (in vivo, ex vivo, in situ).

Researchers apply to conduct **human studies** and **obtain FDA approval**. This is called an Investigational New Drug (IND) Application. The FDA then works with researchers on clinical trial design throughout the clinical trial phases.

CLINICAL TRIALS

PHASE 0: Tests a very small amount of a drug in healthy individuals to assess safety (possible harm, side effects, dosage).

PHASE 1: Tests a drug in small numbers of healthy individuals for safety.

PHASE 2: Tests a drug in a small number of participants with a certain disease or condition for effectiveness and continued safety against a placebo in a double blinded RCT.

PHASE 3: Tests the effectiveness and continued safety of the drug in a much larger number of people with a certain disease or condition during a longer period and against a placebo and other possible existing treatments in a double blinded RCT.

Researchers submit a New Drug Application (NDA) to the FDA for approval.

FDA REVIEW — The FDA makes its decision based on the safety and efficacy data from ALL the animal and human studies that were conducted and reviews the drug labels and the manufacturing facilities where the drug will be made.

FDA APPROVAL

POST-MARKETING — **PHASE 4:** After a drug has been approved and is being used in the real world, phase 4 studies monitor for any additional side effects and continued safety and effectiveness.

Figure 5.249: **Drug development and clinical trials**

To answer the first question we can develop statistical methods that will allow us to compare the two groups of patients who receive both types of treatment.

To answer the second question we propose a model that links patient survival time for explanatory variables and highlight the prognostic factors.

Let us as always assist the theory with an example. For this consider the following table where two cohorts of patients (we move from mechanical engineering to human engineering...) of same initial size with acute leukemia drug test (6-MP) against a placebo[92] (of course blindly).

We have the following remission times for 21 patients (the table of 21 lines therefore indicates the number of weeks for where patient is considered cured after treatment before falling again ill):

6-mercaptopurine (6-MP) Group	**Placebo Group**
6	1
6	1
6	2
6+	2
7	3
9+	4
10	4
10+	5
11+	5
13	8
16	8
17+	8
19+	8
20+	11
22	11
23	12
25	12
32+	15
32+	17
34+	22
35+	23

Table 5.116: **Two cohort survival analysis with right censoring**

The "+" sign corresponds to patients who left the study for that week. They are therefore censored! For example, the fourth patient was lost of view for any reason after 6 weeks of treatment with 6-MP: it has therefore a duration of remission greater than 6 weeks! So in the study 6-MP, there are 21 patients and 11 of them are censored data.

---

[92] Let us recall that in the science of drugs, a "pure placebo" is a treatment without any active substance; an "impure placebo" is a pharmacologically active product but has no effect on the pathology treated, or whose efficacy has not been sufficiently demonstrated.

Remark

The theoretical model assumes that censoring is independent of the survival time (not informative censoring). But if censoring is due to the discontinuation of treatment, the independence assumption is not valid anymore!

For the placebo group it is simple to make a survival curve. It is sufficient to produce the following table (for the omitted weeks, obviously we impose the number of remission as constants):

Week i	**Total # remission at week i**	**Proportion (probability) of remission at week i**
0	21	100%
1	19	$19/21=90\%$
2	17	$17/21=81\%$
3	16	$16/21=76\%$
4	14	$14/21=67\%$
5	12	$12/21=57\%$
8	8	$8/21=38\%$
11	6	$6/21=29\%$
12	4	$4/21=19\%$
15	3	$3/21=14\%$
17	2	$2/21=10\%$
22	1	$1/20=0.05\%$
23	0	0%

Table 5.117: **Two cohort survival analysis with right censoring**

So if the data are not censored, the survival $S(t)$ can be estimated by the proportion of individuals surviving at time t, that is customary to write under the following mathematical form:

$$\hat{S}_k(t)=P_k(T_k>t)=\frac{1}{n}\sum_{i=1}^{k}1_{T_i>t} \tag{5.7.2739}$$

The idea is therefore to estimate:

$$P(T>t)=P(\text{probability to survive until time } t) \tag{5.7.2740}$$

by the proportion of patients who survived until time t.

If the data are censored, the estimated survival function requires specific tools. Kaplan and Meier have proposed in this particular case the following calculation:

$$\begin{aligned}&P(\text{be in remission a the } i\text{-th week})\\&=P(\text{to not relapse at the } i\text{-th week})\cdot P(\text{be in remission a the } i-1\text{-th week})\end{aligned} \tag{5.7.2741}$$

Let's see it in a slightly more mathematical form. If we denote by $X(1)\leq X(2)\leq\ldots\leq X(n)$ the moments (sorted) where an event occurred (death or censored), then we have:

$$P(T>X(k))=P(T>X(k)|T>X(k-1))P(T>X(k-1)) \tag{5.7.2742}$$

With of course:

$$P(T>0)=1 \tag{5.7.2743}$$

We estimate:

$$\hat{P}_k = \left(1 - \frac{d_k}{R_k}\right) \qquad (5.7.2744)$$

where d_k is the number of deaths (failures) observed in the time corresponding to the event $X(k)$ and R_k is the number of individuals at risk (at risk of death/failure) just before $X(k)$.

We define the "Kaplan-Meier estimator" for any $X(0) \leq t < X(k)$ by:

$$\boxed{\hat{S}_{\text{KM}}(t) = \prod_{k=1}^{n} \hat{P}_k = \prod_{k=1}^{n} \left(1 - \frac{d_k}{R_k}\right) = \prod_{k=1}^{n} \left(1 - \hat{\lambda}_k\right)} \qquad (5.7.2745)$$

where $\hat{\lambda}_k$ is the discrete hazard (i.e. conditional probability of death).

Remark

There is a much simpler estimator named "Nelson-Aalen estimator" and simply defined as:

$$\hat{S}_{\text{NA}} = \sum_{k=1}^{n} \frac{1}{R_k} \qquad (5.7.2746)$$

That latter estimator is just a sum over the observed survival times.

Therefore we get doing now for 6-MP group (not the placebo group!!!!!) the following:

Remission duration (uncensored) observed	**Items in remission at the beginning of k**	**Probability of not relapse at k knowing that we are at at $k-1$ ($\hat{P}_k$)**	**Survival probability of Kaplan-Meier**
0	21	$21/21 = 100\%$	100%
6	21	$18/21 = 1 - 3/21 = 85.7\%$	$100\% \cdot 85.7\% = 85.7\%$
7	17	$16/17 = 1 - 1/17 = 94.1\%$	$85.7\% \cdot 94.1\% = 80.7\%$
10	15	$14/15 = 1 - 1/15 = 93.3\%$	$80.7\% \cdot 93.3\% = 75.3\%$
13	12	$11/12 = 1 - 1/12 = 91.7\%$	$75.3\% \cdot 91.7\% = 69\%$
16	11	$10/11 = 1 - 1/11 = 90.9\%$	$69\% \cdot 90.9\% = 62.7\%$
22	7	$6/7 = 1 - 1/7 = 85.7\%$	$62.7\% \cdot 85.7\% = 53.8\%$
23	6	$5/6 = 1 - 1/6 = 83.3\%$	$53.8\% \cdot 83.3\% = 44.8\%$

Table 5.118: **6-MP survival analysis**

With the corresponding survival curves (including the confidence interval):

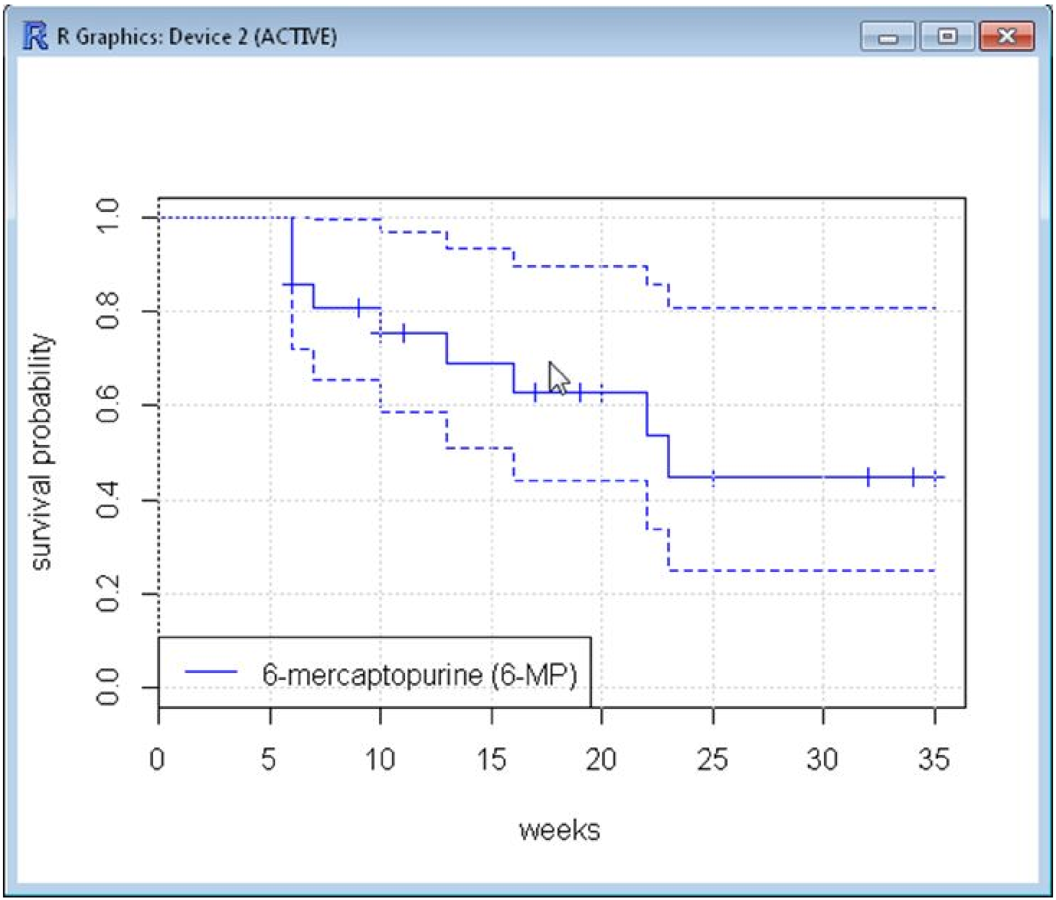


Figure 5.250: Kaplan-Meier Typical survival curves in R software

We thus find the same values as those given for example by Minitab Statistical Software 15.1.1 and R (see the companion books on Minitab or R for the details).

If we assume that the $\hat{\lambda}_k$'s are just binomial proportions given the R_k's we have the following normalized variance and mean of binomial random variable (see page 663):

$$\mathrm{E}(\hat{\lambda}_k) = \hat{\lambda}_k \qquad \mathrm{V}(\hat{\lambda}_k) = \frac{\hat{\lambda}_k(1-\hat{\lambda}_k)}{R_k} \tag{5.7.2747}$$

Remark

Sadly it is a tradition in survival analysis to denote the normalized variance of $\hat{\lambda}_k$ by $\mathrm{V}(\hat{\lambda}_k)$ instead of $\mathrm{V}(\hat{\lambda}_k/r_k)$...

And we will assume also that the $\hat{\lambda}_k$'s are asymptotically independent (...).

Since $\hat{S}_{\text{KM}}(t)$ is a function of the $\hat{\lambda}_k$'s, we can estimate its variance using the univariate Delta method (see section Numerical Methods page 6504) that has for important result for recall that:

$$g(X_n) \underset{n\to+\infty}{\cong} \mathcal{N}\left(0, g'(\mu)^2\sigma^2\right) \tag{5.7.2748}$$

with:

$$\mathrm{V}(g(X_n)) \cong g'(\mu)^2\sigma^2 \tag{5.7.2749}$$

Therefore instead of dealing with $\hat{S}_{\text{KM}}(t)$ directly, we will look at its natural logarithm:

$$\ln(\hat{S}_{\text{KM}}(t))=\sum_{k=1}^{n}\ln(1-\hat{\lambda}_k) \tag{5.7.2750}$$

Thus, by the approximate independence of the $\hat{\lambda}_k$'s and keeping in mind that we just choose $g()$ as the natural logarithm and using the univariate delta method:

$$\begin{aligned}\hat{\text{V}}\left[\ln(\hat{S}_{\text{KM}}(t))\right]&\cong\sum_{k=1}^{n}\hat{\text{V}}\left[\ln(1-\hat{\lambda}_k)\right]\underset{\text{delta method}}{\cong}\sum_{k=1}^{n}\left(\frac{1}{1-\hat{\lambda}_k}\right)^2\hat{\text{V}}(\hat{\lambda}_k)\\&=\sum_{k=1}^{n}\left(\frac{1}{1-\hat{\lambda}_k}\right)^2\frac{\hat{\lambda}_k(1-\hat{\lambda}_k)}{R_k}=\sum_{k=1}^{n}\frac{\hat{\lambda}_k}{(1-\hat{\lambda}_k)R_k}=\sum_{k=1}^{n}\frac{d_k}{(R_k-d_k)R_k}\end{aligned} \tag{5.7.2751}$$

And applying on it the univariate delta method:

$$\text{V}(g(X_n))\cong g'(\mu)^2\sigma^2=g'(\mu)^2\text{V}(X_n)\Rightarrow\text{V}\left(\ln(\hat{S}_{\text{KM}}(t)\right)\cong\frac{1}{\hat{S}^2_{\text{KM}}(t)}\hat{\text{V}}\left(\hat{S}_{\text{KM}}(t)\right) \tag{5.7.2752}$$

then rearranging and injecting the previous result we get:

$$\hat{\text{V}}\left(\hat{S}_{\text{KM}}(t)\right)\cong\hat{S}^2_{\text{KM}}(t)\text{V}\left(\ln(\hat{S}_{\text{KM}}(t)\right)=\hat{S}^2_{\text{KM}}(t)\sum_{k=1}^{n}\frac{d_k}{(R_k-d_k)R_k} \tag{5.7.2753}$$

The following approximation:

$$\boxed{\hat{\text{V}}\left(\hat{S}_{\text{KM}}(t)\right)\cong\hat{S}^2_{\text{KM}}(t)\sum_{k=1}^{n}\frac{d_k}{(R_k-d_k)R_k}} \tag{5.7.2754}$$

is named the "Greenwood's relation". If we take the square root then the result is associated to the standard error (because of the division by R_k):

$$\text{SE}\left(\hat{S}_{\text{KM}}(t)\right)\cong\hat{S}_{\text{KM}}(t)\sqrt{\sum_{k=1}^{n}\frac{d_k}{(R_k-d_k)R_k}}=\hat{S}_{\text{KM}}(t)\frac{\sqrt{\sum_{k=1}^{n}\dfrac{d_k}{(R_k-d_k)}}}{\sqrt{R_k}}=\hat{S}_{\text{KM}}(t)\frac{\sigma_{\hat{S}_{\text{KM}}(t)}}{\sqrt{R_k}} \tag{5.7.2755}$$

In order to construct an error bar for $\hat{S}_{\text{KM}}(t)$ we need to make some sort of distributional assumption. The simplest is to assume that $\hat{S}_{\text{KM}}(t)$ is normally distributed such that we can write:

$$\hat{S}_{\text{KM}}(t)\pm Z_{\alpha/2}\text{SE}\left(\hat{S}_{\text{KM}}(t)\right) \tag{5.7.2756}$$

Or more explicitly:

$$\boxed{\hat{S}_{\text{KM}}(t)-Z_{\alpha/2}\hat{S}_{\text{KM}}(t)\sqrt{\sum_{k=1}^{n}\frac{d_k}{(R_k-d_k)R_k}}\leq S_{\text{KM}}(t)\leq\hat{S}_{\text{KM}}(t)+Z_{\alpha/2}\hat{S}_{\text{KM}}(t)\sqrt{\sum_{k=1}^{n}\frac{d_k}{(R_k-d_k)R_k}}} \tag{5.7.2757}$$

A snag with this, of course is that the distribution of $\hat{S}_{\text{KM}}(t)$ is not really normal. Problems arise when $\hat{S}_{\text{KM}}(t)$ is close to 0 or 1 as indeed the confidence interval can yield values to >1 or <0. A common empirical approach (among others!) is then to take for the confidence interval the complementary log-log transformation:

$$\hat{W}(t)=\ln\left(-\ln\left(\hat{S}_{\text{KM}}(t)\right)\right) \tag{5.7.2758}$$

Since this quantity is unrestricted, the confidence interval will be in the right range when we transform back:

$$\hat{S}_{\text{KM}}(t)=e^{-e^{\hat{W}(t)}} \tag{5.7.2759}$$

To see why this works:

$$0\leq\hat{S}_{\text{KM}}(t)\leq 1 \tag{5.7.2760}$$

$$-\infty\leq\ln\left(\hat{S}_{\text{KM}}(t)\right)\leq 0 \tag{5.7.2761}$$

$$0\leq-\ln\left(\hat{S}_{\text{KM}}(t)\right)\leq+\infty \tag{5.7.2762}$$

$$-\infty\leq\ln\left(-\ln\left(\hat{S}_{\text{KM}}(t)\right)\right)\leq+\infty \tag{5.7.2763}$$

To get a confidence interval we need now also to find the standard error corresponding to this log-log transformation, and for this purpose we need to calculate:

$$\hat{\text{V}}(\hat{W}(t))=\hat{\text{V}}\left(\ln\left(-\ln\left(\hat{S}(t)\right)\right)\right) \tag{5.7.2764}$$

From our previous calculations, we know:

$$\hat{\text{V}}\left[\ln(\hat{S}_{\text{KM}}(t))\right]\cong\sum_{k=1}^{n}\frac{d_k}{(R_k-d_k)R_k} \tag{5.7.2765}$$

Applying the univariate delta method again, we get:

$$\hat{\text{V}}(\hat{W}(t))=\hat{\text{V}}\left(\ln\left(-\ln\left(\hat{S}(t)\right)\right)\right)\cong\frac{1}{\left[\ln\left(\hat{S}(t)\right)\right]^2}\sum_{k=1}^{n}\frac{d_k}{(R_k-d_k)R_k} \tag{5.7.2766}$$

We take the square root of the above to get the standard error:

$$\text{SE}(\hat{W}(t))\cong\frac{1}{\ln\left(\hat{S}(t)\right)}\sqrt{\sum_{k=1}^{n}\frac{d_k}{(R_k-d_k)R_k}} \tag{5.7.2767}$$

And therefore we have:

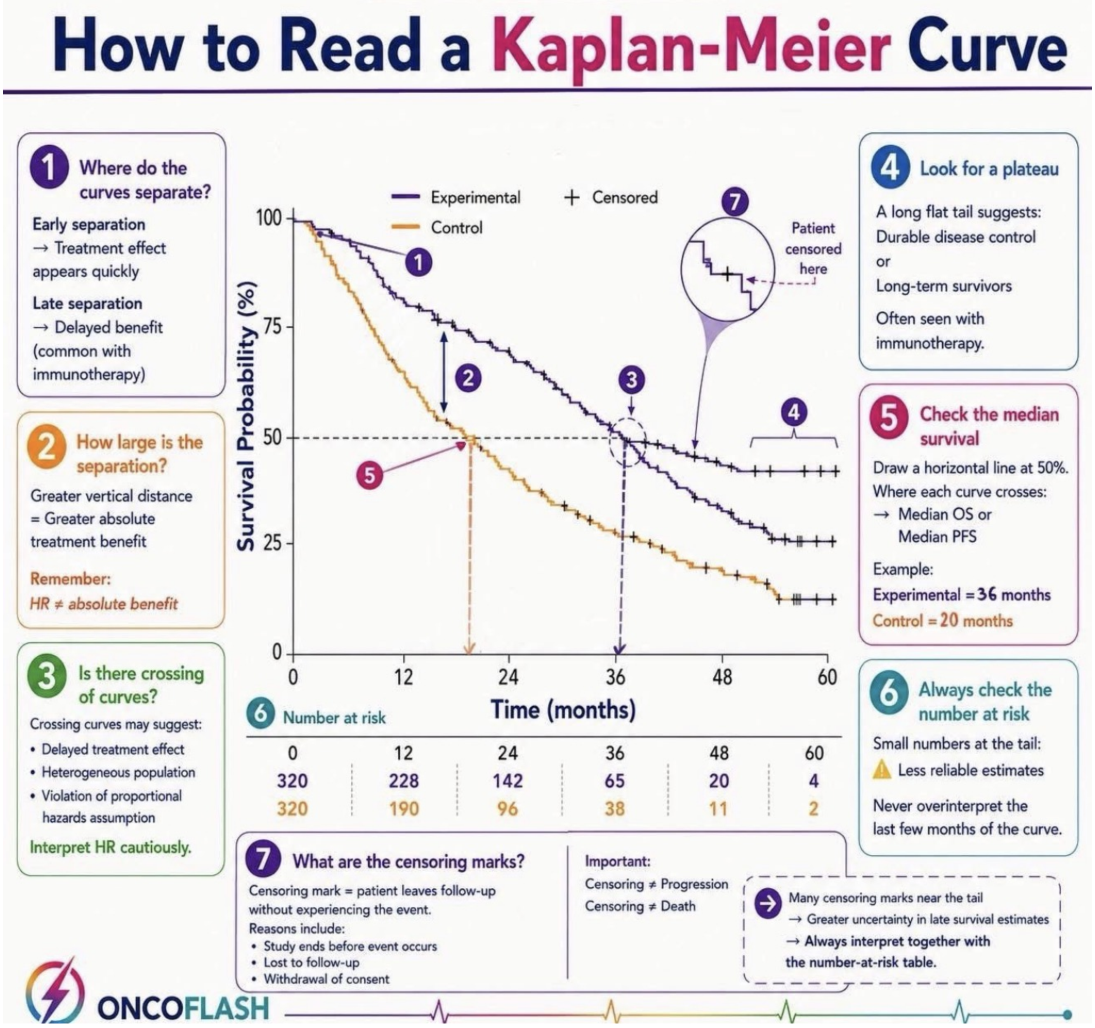
$$\hat{W}(t)-Z_{\alpha/2}\frac{1}{\ln\left(\hat{S}(t)\right)}\sqrt{\sum_{k=1}^{n}\frac{d_k}{(R_k-d_k)R_k}}\leq W(t)\leq\hat{W}(t)+Z_{\alpha/2}\frac{1}{\ln\left(\hat{S}(t)\right)}\sqrt{\sum_{k=1}^{n}\frac{d_k}{(R_k-d_k)R_k}} \tag{5.7.2768}$$

Since $\hat{S}_{\text{KM}}(t)=e^{-e^{\hat{W}(t)}}$, the ("modern") confidence bounds for the confidence interval of $\hat{S}_{\text{KM}}(t)$ are finally (note that the upper and lower bound switch!):

$$\exp\left\{-e^{\hat{W}(t)+Z_{\alpha/2}\dfrac{1}{\ln\left(\hat{S}(t)\right)}\sqrt{\displaystyle\sum_{k=1}^{n}\frac{d_k}{(R_k-d_k)R_k}}}\right\}\leq S_{\text{KM}}(t)\leq\exp\left\{-e^{\hat{W}(t)-Z_{\alpha/2}\dfrac{1}{\ln\left(\hat{S}(t)\right)}\sqrt{\displaystyle\sum_{k=1}^{n}\frac{d_k}{(R_k-d_k)R_k}}}\right\}\tag{5.7.2769}$$

Substituting $\hat{W}(t)=\ln\left(-\left(\hat{S}_{\text{KM}}(t)\right)\right)$ back into the above bounds, we get the definitive "modern" relation for the Kaplan-Meier estimator confidence bound:

$$\boxed{\hat{S}_{\text{KM}}(t)\exp\left\{e^{Z_{\alpha/2}\dfrac{1}{\ln\left(\hat{S}(t)\right)}\sqrt{\displaystyle\sum_{k=1}^{n}\frac{d_k}{(R_k-d_k)R_k}}}\right\}\leq S_{\text{KM}}(t)\leq\hat{S}_{\text{KM}}(t)\exp\left\{e^{-Z_{\alpha/2}\dfrac{1}{\ln\left(\hat{S}(t)\right)}\sqrt{\displaystyle\sum_{k=1}^{n}\frac{d_k}{(R_k-d_k)R_k}}}\right\}}\tag{5.7.2770}$$



How to Read a Kaplan-Meier Curve

① Where do the curves separate?

Early separation
→ Treatment effect appears quickly

Late separation
→ Delayed benefit (common with immunotherapy)

② How large is the separation?

Greater vertical distance = Greater absolute treatment benefit

Remember:
HR ≠ absolute benefit

③ Is there crossing of curves?

Crossing curves may suggest:
- Delayed treatment effect
- Heterogeneous population
- Violation of proportional hazards assumption

Interpret HR cautiously.

④ Look for a plateau

A long flat tail suggests:
Durable disease control
or
Long-term survivors

Often seen with immunotherapy.

⑤ Check the median survival

Draw a horizontal line at 50%.
Where each curve crosses:
→ Median OS or Median PFS

Example:
Experimental = **36** months
Control = **20** months

⑥ Number at risk

0	12	24	36	48	60
320	228	142	65	20	4
320	190	96	38	11	2

⑥ Always check the number at risk

Small numbers at the tail:
⚠ Less reliable estimates

Never overinterpret the last few months of the curve.

⑦ What are the censoring marks?

Censoring mark = patient leaves follow-up without experiencing the event.
Reasons include:
- Study ends before event occurs
- Lost to follow-up
- Withdrawal of consent

Important:
Censoring ≠ Progression
Censoring ≠ Death

Many censoring marks near the tail
→ Greater uncertainty in late survival estimates
→ **Always interpret together with the number-at-risk table.**

ONCOFLASH

5.7.16.2 Cochran–Mantel–Haenszel Tests

In statistics, the Cochran–Mantel–Haenszel tests are a collection of test statistics used in the analysis of stratified categorical data. One of these test statistics is the "Cochran–Mantel–Haenszel (CMH) test", which allows the comparison of two groups on a dichotomous/categorical response.

This test, also named "log rank test" or "Cochran-Mantel(-Haenszel) test with stratified time" or "Mantel-Cox test", or "Cochran-Mantel-Haenszel Chi-square test" or furthermore "multi-strata proportions test" ... has for main objective in practice to test the null hypothesis H_0 as what two survival curves (control group[93] versus test group), such as those visible in the figure below (where the surviving population was normalized on the y-axis), are significantly different or not under the assumptions that:

H1. Each stratum (layer) is independent of the previous one.

H2. For each stratum we are expecting that the expected proportion of survivors / death is the same all things being equal... or "*ceteris paribus*" for those that prefer Latin (see below if this is not clear).

H3. Each stratum is distributed according to the hypergeometric law.

H4. The hypergeometric distribution can be approximated by a Normal distribution (what we will recall, can only be done under certain conditions!!!).

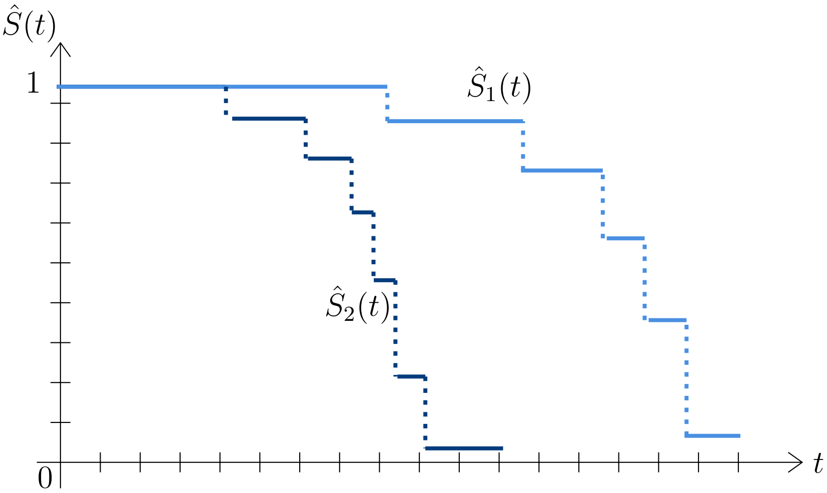


Figure 5.251: **Cox Typical survival curves (without their confidence intervals)**

In the CMH test, the data are arranged in a series of associated 2×2 contingency tables, the null hypothesis is that the observed response is independent of the treatment used in any 2×2 contingency table. The CMH test's use of associated 2×2 contingency tables increases the ability of the test to detect associations (the power of the test is increased).

---

[93] The irony with antivaxers saying they don't want to be part of an experiment is that they don't realize that they are the control group...

In other words, the null hypothesis H_0 of the test is that treatment (medical, mechanical or other) has no influence between the control group and the test group in one or multiple stratum (corresponding to different hospitals, cohorts, laboratories, etc.), this is written as:

$$H_0: \quad S_1(t)=S_2(t)=\ldots=S_n(t) \qquad 0<t<+\infty \tag{5.7.2771}$$

To introduce this test we create a 2×2 contingency table for each stratum, that can correspond to a time interval $[t_i,t_{i+1}]$ where $i=1,2,\ldots,n$ and that can also be assimilated to a different hospital/laboratory for a same clinical/reliability study, which have typically the following structure:

	Observed Deaths	**Observed Survivors**	**Total**
Group I (control)	a_i	b_i	L_{1i}
Group II (test)	c_i	d_i	L_{2i}
Total	C_{1i}	C_{2i}	n_i

Table 5.119: **Typical 2×2 contingency table for the CMH test**

We speak then of "general stratified $2\times 2\times n$ table" or of "contingency table in three dimensions"...

Remark

Let us recall that if we have a single table and that we only want to compare the proportion of survivors or deaths for both groups, a test for differences in proportions will be applied as seen earlier above. We can also do a chi-square test if we always have only one single table (see also earlier above) and the ad hoc conditions are met or also a Fisher exact test if the conditions of the chi-square test are not met. This is why in good statistical software (like Minitab for example), these three tests are available next to each other.

Thus, under the assumption that all other things remaining equal (H2), and this is the core of this test (!!!!), the expected number E of individuals for each cell in the period i will like as for the exact Fisher test or the Kohen kappa agreement coefficient be equal to[94]:

$$\begin{aligned}E_{a,i}&=L_{1i}\frac{C_{1i}}{n_i}=\frac{(a_i+b_i)(a_i+c_i)}{n_i}\\E_{c,i}&=L_{2i}\frac{C_{1i}}{n_i}=\frac{(c_i+d_i)(a_i+c_i)}{n_i}\\E_{b,i}&=L_{1i}\frac{C_{2i}}{n_i}=\frac{(a_i+b_i)(b_i+d_i)}{n_i}\\E_{d,i}&=L_{2i}\frac{C_{2i}}{n_i}=\frac{(c_i+d_i)(b_i+d_i)}{n_i}\end{aligned} \tag{5.7.2772}$$

So it must be well understood that, for example, $E_{a,i}$ then represents the expected number of dead individuals from the control group if it would behaved like this of all individuals of the Control + Test groups. So if the two groups (or survival curves) behave identically, the expected value should be equal to the observed value. To be sure that the reader understand, let us illustrate this concept with a specific example:

---

[94] L states for "Line" and C for "Column"

✍**Example:**

We consider the following observed values:

	Observed Deaths	**Observed Survivors**	**Total**
Group I (control)	200	800	1000
Group II (test)	280	1120	1400
Total	480	1920	2400

Table 5.120: **Special case of observed values for the CMH test**

Here, the expected values gives:

	Observed Deaths	**Observed Survivors**	**Total**
Group I (control)	200	800	1000
Group II (test)	280	1120	1400
Total	480	1920	2400

We see in the above this particular case, the expected values are equal to the observed one (simply for the reason that the ratio $800/1,000$ and $1,120/1,400$ represent in the table the same percentage - the same proportion - of 80%). Obviously, if the proportion of survivors observed for the two groups are all other things remaining equal, it is the same for the observed dead. So what the reader must understand when we do a MCH test is that we are free to choose the column to analyse (because ultimately it is the same!).

Now we can notice that each of the relationships:

$$E_{a,i}=L_{1i}\frac{C_{1i}}{n_i} \qquad E_{c,i}=L_{2i}\frac{C_{1i}}{n_i} \qquad E_{b,i}=L_{1i}\frac{C_{2i}}{n_i} \qquad E_{d,i}=L_{2i}\frac{C_{2i}}{n_i} \tag{5.7.2773}$$

represents in fact the mean of a binomial or hypergeometric distribution (see the proofs earlier above during our study of the corresponding laws) since the two laws have the same expression for the mean. However, as the size of individuals in the cells could be significant relatively to the total of rows or columns, the drawing could not be independent. Then we must come back on the hypergeometric law (that is to say the third hypothesis enumerated before).

We also see that by symmetry of the above relation, that the variable of interest is the column attribute or line attribute does not change anything in reality the result of the calculation as (for example):

$$E_{a,i}=L_{1i}\frac{C_{1i}}{n_i}=C_{1i}\frac{L_{1i}}{n_i} \tag{5.7.2774}$$

The variance for each cell will be that of the hypergeometric law we have already proved during it study and that was given by:

$$\mathrm{V}(X)=kpq\frac{n-k}{n-1}=k\left(\frac{r}{n}\right)\left(\frac{s}{n}\right)\frac{n-k}{n-1}=\frac{krs(n-k)}{n^2(n-1)}=\frac{klrs}{n^2(n-1)} \tag{5.7.2775}$$

with for recall $l=n-k$.

If we adapt the notation of the previous relation to that of our table above, this provides us for all cells (by the analytic form of variance of the hypergeometric law the variance has the same expression for each cell!):

$$\mathrm{V}_i = L_{1i}\frac{L_{2i}C_{1i}C_{2i}}{n_i^2(n_i-1)} = L_{2i}\frac{L_{1i}C_{1i}C_{2i}}{n_i^2(n_i-1)} \tag{5.7.2776}$$

with respectively (for those who want to make the analogy with our detailed study of hypergeometric law):

$$n = n_i \qquad r = C_{2i} \qquad s = C_{1i} \qquad k = L_{1i} \qquad l = n - k = n_i - L_{1i} = L_{2i} \tag{5.7.2777}$$

Here again, we see that whether the variable of interest in either line or column, the calculated value of the variance remains the same!!!

Remark

A common question that arise frequently for those studying this test is to know why we can not simply do the sum of all stratum (layers) in a single table (because it would be much easier, of course...)? Well we can not put together the tables into a single one because they do not necessarily follow the same law (not identically distributed in the sense that the hypergeometric law has different settings from one table to another) and unfortunately, the hypergeometric distribution is not stable by the addition.

And so what...??? How will this help us to compare if the two survival curves are identical through time (or across different hospitals/labs if this is a clinical/reliability test) ?!

Well let us take as companion example the table with the following observed values:

	Observed Deaths	Observed Survivors	Total
Group I (control)	200	900	1100
Group II (test)	280	1120	1400
Total	480	2020	2500

Table 5.121: Special case of observed values for the CMH test

We then get for the expected values:

	Observed Deaths	Observed Survivors	Total
Group I (control)	211.2	888.8	1100
Group II (test)	280	1120	1400
Total	268.8	$1'131.2$	2500

Table 5.122: Expected values of previous table

Thus the differences:

	Observed Deaths	Observed Survivors
Group I (control)	-11.2	11.2
Group II (test)	11.2	-11.2

Table 5.123: Difference between expected values and observed values

We understand then better why it has no influence to choose a particular cell to make the test. We simply take the one that we will arrange us the more (depending on what will follow...).

So taking randomly (since the choice does not affect - at least for now .... - the test result as we have shown in our previous calculations) the column Survivors and particularly the observed group Test. Its expected value is then:

	Observed Survivors
Group II (test)	$E_{d,1}=\dfrac{L_{21}C_{21}}{n_1}=\dfrac{1400\cdot 2020}{2500}=1131.2$

Table 5.124: **Expected value of the Tests-Survivor cell**

The difference between observed and expected gives:

$$1120-1131.2=-11.2 \tag{5.7.2778}$$

That difference will be obviously zero if the observed was equal to the expected! With for variance:

$$\text{V}_1=\frac{L_{1i}L_{2i}C_{1i}C_{2i}}{n_i^2(n_i-1)}=\frac{480\cdot 2020\cdot 1100\cdot 1400}{2500^2(2500-1)}\cong 95.602 \tag{5.7.2779}$$

Now, under the conditions:

$$L_{2i}>9\left(\frac{n_i-C_{2i}}{C_{2i}}\right)\qquad L_{2i}>9\left(\frac{C_{2i}}{n_iC_{2i}}\right)\qquad L_{2i}>\frac{n_i}{10} \tag{5.7.2780}$$

as seen during our study of the hypergeometric law, we can approximate it by a Normal law (hypothesis H4).

So in our case, this approximation is not acceptable (the third condition is disqualifying) and so the test can not be performed (usually we use to take in the table the column and the line that allow this approximation since the choice does not influence on the value of the test but on the authorization to use the mentioned approximation!). But if it had been, we could approximate the hypergeometric distribution by a Normal distribution such that:

$$\mathcal{N}(\mu_i,\sigma_i)=\left(L_{2i}\frac{C_{2i}}{n_i},\sqrt{\frac{L_{2i}L_{1i}C_{1i}C_{2i}}{n_i^2(n_i-1)}}\right) \tag{5.7.2781}$$

Which can be obviously reduced to a centered reduced Normal distribution if we take for the cell of our table:

$$\frac{d_i-L_{2i}\dfrac{C_{2i}}{n_i}}{\sqrt{\dfrac{L_{2i}L_{1i}C_{1i}C_{2i}}{n_i^2(n_i-1)}}}\cong\mathcal{N}(0,1) \tag{5.7.2782}$$

And therefore it is sufficient for a stratum (layer) i to know if we are outside of the confidence interval that we have set. But ... we have several strata (layers)! The idea then is to sum on the T independent strata (then we fall back on the hypothesis H1) such that:

$$\sum_{i=1}^{T}\frac{d_i-L_{2i}\dfrac{C_{2i}}{n_i}}{\sqrt{\dfrac{L_{2i}L_{1i}C_{1i}C_{2i}}{n_i^2(n_i-1)}}}\cong\sum_{i=1}^{T}\mathcal{N}(0,1)=\mathcal{N}\left(0,\sqrt{T}\right) \tag{5.7.2783}$$

which was not quite practical at the time when not everyone had a computer since only $\mathcal{N}(0,1)$ tables we available. This is why we prefer to do the following sum:

$$\frac{\displaystyle\sum_{i=1}^{T}\left(d_i - L_{2i}\frac{C_{2i}}{n_i}\right)}{\sqrt{\displaystyle\sum_{i=1}^{T}\frac{L_{2i}L_{1i}C_{1i}C_{2i}}{n_i^2(n_i-1)}}} \cong \mathcal{N}(0,1) \qquad (5.7.2784)$$

Or in condensed form (again regardless of the choice of the cell!):

$$\frac{\displaystyle\sum_i (d_i - E_{d,i})}{\sqrt{\displaystyle\sum_i V_i}} \cong \mathcal{N}(0,1) \qquad (5.7.2785)$$

And if the differences between observed and expected is not too great across all strata, the value of this expression well be found within a certain confidence interval of the Normal distribution. If it is outside, the hypothesis of equality of the two survival curves will be rejected.

However, the majority of statistical softwares take the square of that relation.

It comes then that the square follows a chi-squared law with one degree of freedom (we have proved this earlier in this section) such that we fall back on the final form such that the log-rank test can be found in many books:

$$\boxed{\frac{\left(\displaystyle\sum_i (d_i - E_{d,i})\right)^2}{\displaystyle\sum_i V_i} \cong \chi^2(1)} \qquad (5.7.2786)$$

So this is a Wald statistic (see earlier above) and therefore the CMH test belongs to the family of Wald tests!

For practical reasons, we add a 0.5 term to the sum, this provides a better approximation to the Normal distribution. Therefore we can see sometimes in books:

$$\frac{\left(\displaystyle\sum_i (d_i - E_{d,i}) + 0.5\right)^2}{\displaystyle\sum_i V_i} \cong \chi^2(1) \qquad (5.7.2787)$$

✍**Example:**

Let us consider the following observed values of two hospitals A and B for a clinical trial:

A	**Observed Deaths**	**Observed Survivors**	**Total**
Group I (control)	288	4	292
Group II (test)	400	50	450
Total	688	54	742

B	**Observed Deaths**	**Observed Survivors**	**Total**
Group I (control)	300	10	310
Group II (test)	450	40	490
Total	750	50	800

We don't want to know if the two hospitals are different or not (this is not the purpose!), but whether the differences between the control group and test group across all hospitals is significantly different or not.

Well we already intuitively guess the result when we see the values ... but still let us do the calculations.

We are already seeing that for hospital A, the column Survivors meets the three conditions for the approximation by a Normal distribution (which is not the case for the column of the Deaths):

$$54 > 9\left(\frac{742-450}{742}\right) \qquad 54 > 9\left(\frac{742}{742-450}\right) \qquad 54 < \frac{742}{10} \tag{5.7.2788}$$

Similarly for the hospital B (and this is fortunate because once a column selected in one stratum, it requires that the approximation condition is applicable to the same column of all other strata!!!):

$$50 > 9\left(\frac{800-490}{800}\right) \qquad 50 > 9\left(\frac{800}{800-490}\right) \qquad 50 < \frac{80}{10} \tag{5.7.2789}$$

A	**Observed Survivors**	**Total**
Group I (control)	21.25	292
Group II (test)	32.75	450
Total	54	742

B	**Observed Survivors**	**Total**
Group I (control)	19.38	310
Group II (test)	30.62	490
Total	50	800

Which gives for the differences:

A	**Observed Survivors**
Group I (control)	-17.25
Group II (test)	17.25

B	**Observed Survivors**
Group I (control)	-9.38
Group II (test)	9.38

We then quickly see why once the column chosen, the choice of the line does not matter anymore (either a sum of negative values or a sum of positive values and as we take the square of the sum, it changes nothing finally!). Let us arbitrarily choose the second line. Then we have:

$$\frac{\left(\sum\limits_{i=1}^{2}(d_i-E_{d,i})\right)^2}{\sum\limits_{i=1}^{2}V_i}=\frac{\left(\sum\limits_{i=1}^{2}\left(d_i-\dfrac{L_{2i}C_{2i}}{n_i}\right)\right)^2}{\sum\limits_{i=1}^{2}\dfrac{L_{2i}L_{1i}C_{1i}C_{2i}}{n_i^2(n_i1)}} \tag{5.7.2790}$$

$$=\frac{(17.25+9.38)^2}{\dfrac{292\cdot 450\cdot 688\cdot 54}{742^2(742-1)}+\dfrac{310\cdot 490\cdot 750\cdot 50}{258^2(258-1)}}\cong\frac{736}{11.966+11.139}\cong 31.845$$

And we have with a spreadsheet software like Microsoft® Excel® 14.0.6123 in left-sided with a threshold risk of 5%:

$$\chi^2_{95\%}(1)=\texttt{CHISQ.INV(95\%,1)}\cong 3.841$$

So on the cumul of the two hospitals (strata) the Control group is significantly different from the Test group. The p-value of the test is typically given with Microsoft® Excel® 14.0.6123:

$$\texttt{=1-CHISQ.DIST(31.845,1,1)}\cong 0.000002\%$$

Now let us quote a famous situation[95]:

«One day when I was a junior medical student, a very important Boston surgeon visited the school and delivered a great treatise on a large number of patients who had undergone successful operations for vascular reconstruction.

At the end of the lecture, a young student at the back of the room timidly asked, "*Do you have any controls?*" Well, the great surgeon drew himself up to his full height, hit the desk, and said, "*Do you mean did I not operate on half the patients?*" The hall grew very quiet then. The voice at the back of the room very hesitantly replied, "*Yes, that's what I had in mind.*" Then the visitor's fist really came down as he thundered, "*Of course not. That would have doomed half of them to their death.*"

God, it was quiet then, and one could scarcely hear the small voice ask, "*Which half?*"»

---

[95] Dr. E. E. Peacock, Jr., quoted in Medical World News (September 1, 11972 according to holocene calendar), p. 45, as quoted in Tufte's 11974 (holocene calendar) book *Data Analysis for Politics and Policy*.

5.7.16.3 Cox (Proportional Hazard) Model

The "Cox proportional hazard model" (not to be confused with "Box-Cox transformations") in continuous time, also known as the "proportional hazards semi-parametric model" or in abbreviated version "Cox PHM" or also "Cox regression model" (RCM), is a model of risk ratio analysis. It expresses the ratio of the instantaneous risk, the "hazard rate", to know the (risky) event studied after a given duration of exposure according to a linear combination of explanatory factors. This model is used extensively in clinical studies, as well as in preventive maintenance and bank credit risk.

This model belongs as the Kaplan-Meier model to the "duration models" or "survival models" and makes use of some of the results demonstrated in the section of Industrial Engineering as part of the study of preventive maintenance (see page 8763).

Remark

The Kaplan-Meier model makes it possible to compare several curves of survival, the Cox's model allows not only to do the same but also to be able to do so on the basis of multiple factors and also to make predictions for values not observed during the experiment.

For the sake of simplification and essentially pedagogical, we will be interested here only in the case where the data are uncensored, the covariates without interactions and themselves independent of time (for example, the covariate "age" would be independent ... of the time...). We will make the presentation more complex as this textbook evolves.

First of all it is necessary to know that the Cox regression model is, as one of its other names indicates it, based on the instantaneous hazard rate (risk) $h(t)$! Why this choice when we could take the survival or reliability functions ??? Simply because the survival functions $R(t)$ and reliability $F(t)$ have constraints that would complicate the construction of the regression model. Whereas the only thing we impose on the hazard rate is that it is between $[0,+\infty[$.

The goal of the Cox model is to be able to quantify and test the effects of individual characteristics such as gender, level of education, social class, nationality, past experience, vaccination, etc. on transition risks (life/death or sick/not sick, etc.).

The basic idea is to be able to say "the risk rate of such group is X times greater than the risk rate of the other group" or in other words "the risk of such a group is in a proportion of X relative to such another group" and a naive first approach and mathematically easy way to manage this is to say that the risk rate is of the form (in the univariate case[96]):

$$h(x)=\beta_0+\beta_1 x \tag{5.7.2791}$$

where x is an explanatory variable most often categorical (and rather of the binary type) and where β_0 is therefore in extenso the risk rate corresponding to $x=0$.

Unlike the "relative risk" (i.e. risk ratio) or "odds ratio", it is a model that is more predictive and is adapted to more than two variables (and not just binary!).

---

[96] So a first assumption is that we assume continuous covariates have a linear form and this should always be checked

However it is a bad design for the moment for the simple reason that the definition domain of $h(x)$ is in actually not limited to $[0,+\infty[$. An idea would be to do as for the logistic model, take the exponential such that:

$$h(x)=e^{\beta_0+\beta_1 x}=e^{\beta_0}e^{\beta_1 x}=h_0 e^{\beta_1 x}>0 \tag{5.7.2792}$$

Obviously, we can have more than one explanatory variable. From then on, as we are already know it, the model would be written:

$$h(\vec{x})=h_0 e^{\vec{\beta}^T\vec{x}} \tag{5.7.2793}$$

where h_0 is obviously the baseline risk when all covariates are zero. In practice, we will see a little further that the determination of this basic risk is of no interest because it is eliminated. This last relation is sometimes written in the following form to highlight the risk part of the characteristic factors of the individuals studied:

$$h(\vec{x})=h_0 e^{\vec{\beta}^T\vec{x}}=h_0 r(x) \tag{5.7.2794}$$

However, we must consider that the hazard rate is time dependent, as we saw earlier in our study of maintenance techniques, such that we have to define the "instantaneous risk":

$$\boxed{h(t,\vec{x})=h_0(t,\vec{\alpha})e^{\vec{\beta}^T\vec{x}}} \tag{5.7.2795}$$

Remark

Therefore the standard situation of the application of this survival method in clinical research projects assumes that a homogeneous population is investigated when subject under different conditions (e.g. experimental treatment and standard treatment). The survival model then assumes that the survival data of the different patients are independent form each other and that each patient's individual survival time distribution is the same (independent and identically distributed failure times)!

However, in the field of clinical trials one observes in many most practical situations that patients differ substantially. The effect of a drug, a treatment or the influence of various explanatory variables may differ greatly between subgroups of patients. To account for such unobserved heterogeneity in the study population Vaupel et al. (11979 according to holocene calendar) introduced "univariate frailty models" into survival analysis (available in R). The key idea is, that individuals possess different frailties, and that those patients who are most frail will die earlier than the others. Consequently, systematic selection of robust individuals (that means patients with low frailty) takes place. We will not introduce that mathematical model in this textbook!

It is important to note that in a simple case we have:

$$h_0(t,\vec{\alpha})=h_0(t)e^{\beta_0} \tag{5.7.2796}$$

that is why the Cox model always has the y-intercept (i.e. β_0) that is zero because it is embedded in the fundamental hazard $h_0(t,\vec{\alpha})$.

In reality, if we take a step back, we could see in the previous expression an analogy with the approach by the method of separation of variables as it is done quite often when solving certain differential equations where the solution is constructed such that each factor depends only on certain parameters.

What is particularly appreciated by the practitioners in this empirical model and which explain the origin of its name is that we do not need to know $h_0(t,\vec{\alpha})$. Indeed, if we consider the ratio of the same model but for different values of covariates at the same time, we have the "hazard ratio":

$$\boxed{\text{H.R.}=\frac{h(t,\vec{x}_i)}{h(t,\vec{x}_j)}=\frac{h_0(t,\vec{\alpha})e^{\vec{\beta}^T\vec{x}_i}}{h_0(t,\vec{\alpha})e^{\vec{\beta}^T\vec{x}_j}}=e^{\vec{\beta}^T\vec{x}_i-\vec{\beta}^T\vec{x}_j}=e^{\vec{\beta}^T(\vec{x}_i-\vec{x}_j)}} \tag{5.7.2797}$$

Most of time the ratio is of the following type:

$$\text{H.R.}=\frac{h(t,\vec{x}_i)}{h(t,\vec{x}_j)}=\frac{\text{Hazard in the intervention group}}{\text{Hazard in the control group}} \tag{5.7.2798}$$

We see therefore that an assumption of this approach is that the two models are assumed to have the same initial properties (assumption often hidden in clinical studies)!!

✍**Examples:**

E1. For a single dichotomous covariate, say with values 0 and 1, the hazard ratio is:

$$\text{H.R.}=\frac{h(t,x=1)}{h(t,x=0)}=\frac{h_0(t)e^{\beta\cdot 1}}{h_0(t)e^{\beta\cdot 0}}=\frac{e^{\beta}}{e^{0}}=e^{\beta} \tag{5.7.2799}$$

Thus, the relative risk of two individuals with different covariate values should be independent of time or constant at all times. This is an inherent assumption of the Cox model (and any other proportional hazards model).

E2. With two covariates we have:

$$\text{H.R.}=\frac{h(t,\vec{x})}{h(t,\vec{x}\,')}=\frac{h_0(t)e^{\vec{\beta}^T\vec{x}}}{h_0(t)e^{\vec{\beta}^T\vec{x}'}}=e^{\vec{\beta}^T(\vec{x}-\vec{x}')}=e^{\beta_1(x_1-x_1')+\beta_2(x_2-x_2')} \tag{5.7.2800}$$

Consider now for example that we keep x_1 fixed such that $x_1'=x_1$ and that we take $x_2'=x_2+1$ (x_2 may be an integer value like the rounded age of the patient for example). Then the previous relation will be written:

$$\text{H.R.}=e^{\beta_1(x_1-x_1')+\beta_2(x_2-x_2')}=e^{\beta_1\cdot 0+\beta_2\cdot 1}=e^{\beta_2} \tag{5.7.2801}$$

And notice that whatever the value of x_2 (and keeping fixed the other covariates), each incrementation of this covariate by $+1$ will multiply the instantaneous risk by a factor e^{β} (this value is often written in percentage obviously).

The consensus is to read the values of the hazard ratio as following:

- If $\text{H.R.}=0.5$: At any particular time, half as many patients in the treatment group are experiencing an event compared to the control group.

- If H.R. $=1.0$: At any particular time, event rates are the same in both groups.
- If H.R. $=2.0$: At any particular time, twice as many patients in the treatment group are experiencing an event compared to the control group.

Assuming that the previous assumption is satisfied, it is thus very interesting in the practice of "having fun" to vary the characteristics of individuals in order to calculate how the difference $(\vec{x}_i-\vec{x}_j)$ increases or decreases the risk ratio in percentages (this is where lies the power of investigation and the fun of the Cox model!).

So not only is the basic risk eliminated when we look at relative risk (this is by assumption!!!), but time dependence too!

So the fundamental assumption of the Cox model, as already mentioned, is that we assume for different groups that the basic hazard is equal!!! This assumption of proportionality is trivially visible when we use qualitatively survival graphs like those shown later below with two groups[97] (i.e. two stratas) and we see that they start at the same point at $t=0$ (otherwise we can use other models likes the Threshold Regression Models).

So for recall, the assumption so far are:

H1. The model is multiplicative (i.e. the link between $h(t)$ and $h_0(t)$ must be of the form $h(t)=c^{te}h_0(t)$)

H2. We assume continuous covariates have a linear form

H3. We assume the covariates to be time independent (hence proportional hazards must hold!)

H4. We assume the initial risk rate to be equal

A classical test for time independence is to transform the survival curves by taking:

$$-\ln\left(-\ln\left(\hat{S}(t,\vec{x})\right)\right) \tag{5.7.2802}$$

and check if the both resulting linearised curve a parallel. Indeed, let us recall that we have[98] (see section Industrial Engineering page **??**):

$$S(t,\vec{x})=e^{-\int\limits_0^t h(u,\vec{x})\mathrm{d}u} \tag{5.7.2803}$$

And the Cox hypothesis is:

$$h(t,\vec{x})=h_0(t)e^{\vec{\beta}^T\vec{x}}=h_0(t)e^{\sum_{j=1}^p\beta_j x_j} \tag{5.7.2804}$$

Therefore:

$$S(t,\vec{x})=e^{-\int\limits_0^t h_0(u)e^{\sum_{j=1}^p\beta_j x_j}\mathrm{d}u}=\left(e^{-\int\limits_0^t h_0(u)\mathrm{d}u}\right)^{e^{\sum_{j=1}^p\beta_j x_j}}=S_0(t)^{e^{\sum_{j=1}^p\beta_j x_j}} \tag{5.7.2805}$$

---

[97] Population under study is considered as consisting of two sub-populations with different risks.

[98] We just change the notation corresponding to that of the medical tradition, with $S(t)=R(t)$ and $h(u)=\lambda(u)$

Taking the natural logarithm:

$$-\ln\left(S(t,\vec{x})\right)=-\ln(S_0(t))e^{\sum_{j=1}^{p}\beta_j x_j} \tag{5.7.2806}$$

and once again:

$$-\ln\left(-\ln\left(S(t,\vec{x})\right)\right)=-\ln\left(-\ln(S_0(t))\right)-\sum_{j=1}^{p}\beta_j x_j \tag{5.7.2807}$$

Therefore for two sets of observations:

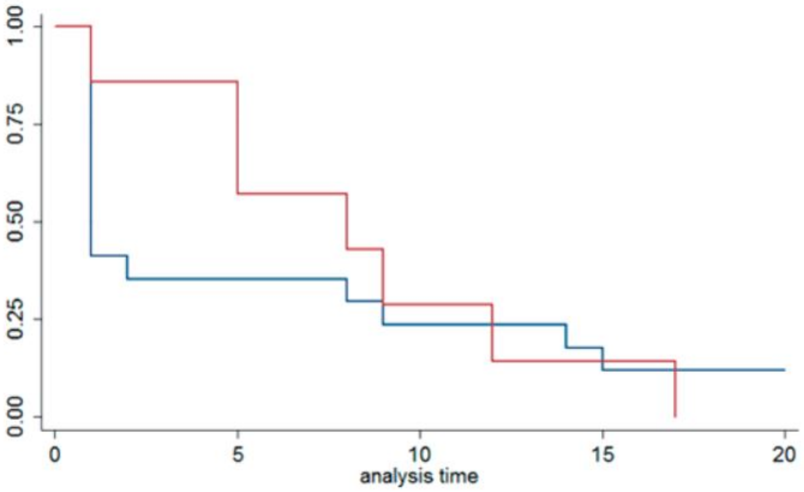
$$-\ln\left(S(t,\vec{x}_2)\right)-\ln\left(S(t,\vec{x}_1)\right)=\sum_{j=1}^{p}\beta_j\left(x_{2,j}-x_{1,j}\right)=\vec{\beta}^T\left(\vec{x}_2-\vec{x}_1\right) \tag{5.7.2808}$$

Then space between the natural logarithm of the two curves should be constant over time!

Remark

In practice a common test to check the assumption of proportionality is by using the "Schoenfeld residuals" against the transformed time.

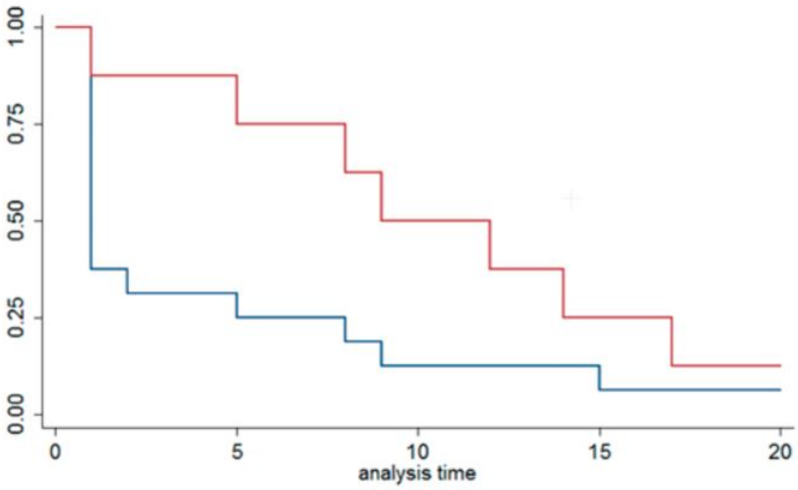
So let us see some example of two strates of Cox PH model regression to make some qualitative judgement based on the plots:



The above plot (regression) violates the following assumptions:

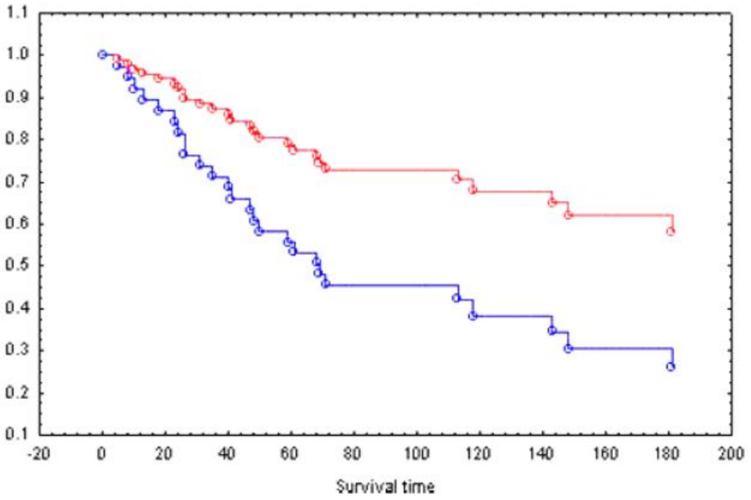
1. The distance between the curves are surely not constant as the curves cross
2. We cannot judge of the underlying Normality of the data
3. The blue curve seem not be based on a linear model (that we take the natural logarithm or not!)
4. Surely the hazard risk ratio is equal at the beginning (this is quite bad) but even worst... there seem on the long term to no have any asymptotic constant hazard risk ratio

The plot below is a bit better relatively to the respect of the assumption of the model:



Indeed the curves don't cross anymore. But sadly the distance doesn't seems to be constant.

The next regression is better than the two previous one (seems to have an asymptotic constant distance[99]):



Well it is nice to be able to compare logically the ratio of the hazard rate at the same time and for different values of covariates but how do determine the coefficients of the model??? A first reflex would be to take the natural logarithm of the model to make a simple linear regression:

$$\ln(h(t,\vec{x})=\ln\left(h_0(t,\vec{\alpha})e^{\vec{\beta}^T\vec{x}}\right)=\ln(h_0(t,\vec{\alpha}))+\vec{\beta}^T\vec{x} \quad (5.7.2809)$$

But we see immediately that the time component will be problematic! Then comes the second reflex: work with maximum likelihood (see above page 818)! But then a question arises ...: the hazard function is not strictly speaking a probability then what will we maximize ???

---

[99] We think that the reader understand now that if we can by vertical translation superpose perfectly the both regression than hypothesis H3 holds!

Well for this we are going to use conditional probabilities (see above page 485), even if at first sight it is not easy to see what connects conditional probabilities to the hazard[100] function but wait a little, denoting by x_i the particular character that identifies a certain class of individuals.

Consider for this E, the set of elements including individuals that survived S and those do not survived denoted $\bar{S}$. Such that for a given identical time:

$$\{S(t),\bar{S}(t)\}\in E(t) \tag{5.7.2810}$$

such that:

$$S(t)\cup\bar{S}(t)=E(t) \tag{5.7.2811}$$

as well by construction:

$$S(t)\cap\bar{S}(t)=\varnothing \tag{5.7.2812}$$

Since the intersection is empty it will be difficult to reveal the hazard function we may think ... Well that's where a trick comes!

It is true that if we take the set $\bar{S}$ of those who have not survived always at the same instant that those who have survived S the two sets will always be disjoint, such as by involving the variable time, we have:

$$S(t)\cap\bar{S}(t)=\varnothing \quad \forall t \tag{5.7.2813}$$

On the other hand, if we mix the times, we have for example:

$$\bar{S}(t+\Delta t)\cap S(t) \tag{5.7.2814}$$

which is not necessarily empty anymore since they can overlap partially when taken at different times!

Since there is a superposition of the two sets, when taken at different times, we can now cleverly write a conditional probability:

$$P(\bar{S}_{t+\Delta t}/S_t)=\frac{P(\bar{S}_{t+\Delta t}\cap S_t)}{P(S_t)} \tag{5.7.2815}$$

Assuming that the number of survivors decrease, that latter can we written:

$$P(\bar{S}_{t+\Delta t}/S_t)=\frac{P(\bar{S}_{\Delta t})}{P(S_t)} \tag{5.7.2816}$$

Or in everyday language:

$$P(\bar{S}_{t+\Delta t}/S_t)=\frac{P(\bar{S}_{\Delta t})}{P(S_t)}=\frac{\text{probability that an individual did not survive between } [t,\Delta t]}{\text{probability that an individual has survived (until) time } t} \tag{5.7.2817}$$

---

[100] The term "hazard" refers to the potential source of harm, danger, or adverse event, while "probability" is the likelihood of that harm or adverse event occurring.

or say differently:

$$P(\bar{S}_{t+\Delta t}/S_t)=\frac{P(\bar{S}_{\Delta t})}{P(S_t)}=\frac{\text{proportion of individuals that did not survive between }[t,\Delta t]}{\text{proportion of individuals that have survived (until) time }t} \tag{5.7.2818}$$

However, we recognize here in the numerator, if we take again the usual notations preventive maintenance (see section Industrial Engineering 8716), $\mathrm{d}F(t)$ the cumulative failure function! Therefore:

$$P(\bar{S}_{t+\Delta t}/S_t)=\frac{F(t+\Delta t)-F(t)}{\text{proportion of individuals that have survived (until) time }t} \tag{5.7.2819}$$

and we recognize $R(t)$, the reliability function, at the denominator such that:

$$P(\bar{S}_{t+\Delta t}/S_t)=\frac{F(t+\Delta t)-F(t)}{R(t)}=\frac{F(t+\Delta t)-F(t)}{1-F(t)} \tag{5.7.2820}$$

Hence:

$$P(\bar{S}_{t+\Delta t}/S_t)=\frac{\Delta F(t)}{R(t)} \tag{5.7.2821}$$

And we recognize here a ratio that we already see during our study of preventive maintenance !!! So we can write:

$$P(\bar{S}_{t+\Delta t}/S_t)=\frac{\Delta F(t)}{R(t)}=\lambda(t)\Delta t=h(t)\Delta t \tag{5.7.2822}$$

We might think that it would be possible to use this conditional probability to apply the maximum likelihood technique, but ... the problem is that there is implicitly in $h(t)$, a function dependent of time that we have to eliminate in order to do something. From then on, the idea is to make a ratio of the hazard functions for the same given instant so the temporal part will be eliminated! And who says ratio of probabilities, says again conditional probability!

Let us recall that we have just got:

$$P(\bar{S}_{t+\Delta t}/S_t)=\frac{\Delta F(t)}{R(t)} \tag{5.7.2823}$$

That is the conditional probability that an individual without any special characteristics does not survive during the given time slot while all the others (having also no special characteristics either) as well as himself survived until time $t+\Delta t$. Now, if we introduce the characteristics, then we have the possibility to introduce another layer of conditional probability. For this, let us take the special case but quite easily generalizable of a single special characteristic (covariate) and an event of interest on an individual denoted i among all the other individuals j (including the individual i) that are also tracked with this same characteristics. The previous conditional probability can then be written:

$$P\left(\{\bar{S}_{t+\Delta t}/S_t\}\right)=h_i(t,i)\Delta t=h_0(t)h(i)\Delta t \tag{5.7.2824}$$

Now, let's compare this conditional probability of a particular individual conditionally to all other individuals (including i himself for recall!) with all the same characteristics:

$$\frac{P\left(\{\bar{S}_{t+\Delta t}/S_t\}_i\cap\{\bar{S}_{t+\Delta t}/S_t\}_j\right)}{P\left(\{\bar{S}_{t+\Delta t}/S_t\}_j\right)} \tag{5.7.2825}$$

Remark

The attentive reader may have noticed that the model as constructed for the moment assumes that non-survivors are unique for a given value of the characteristic. This is why we are talking about: "Cox proportional hazard regression model, ungrouped, uncensored and time-independent covariates". However, we will see below how to deal with the case where two individuals with different or equal characteristics do not survive at the same time.

But in the particular case which concerns us $\{\bar{S}_{t+\Delta t}/S_t\}_i \in \{\bar{S}_{t+\Delta t}/S_t\}_j$. We then have:

$$\frac{P\left(\{\bar{S}_{t+\Delta t}/S_t\}_i \cap \{\bar{S}_{t+\Delta t}/S_t\}_j\right)}{P\left(\{\bar{S}_{t+\Delta t}/S_t\}_j\right)} = \frac{P\left(\{\bar{S}_{t+\Delta t}/S_t\}_i\right)}{P\left(\{\bar{S}_{t+\Delta t}/S_t\}_j\right)} \tag{5.7.2826}$$

But that we can also write more explicitly (don't forget that j represents a set of individuals different from i) as each individual event related is assumed as a disjoint probability (see section Probabilities page 477):

$$\begin{aligned}\frac{P\left(\{\bar{S}_{t+\Delta t}/S_t\}_i\right)}{P\left(\{\bar{S}_{t+\Delta t}/S_t\}_j\right)} &= \frac{P\left(\{\bar{S}_{t+\Delta t}/S_t\}_i\right)}{\sum\limits_{k=1}^{n} P\left(\{\bar{S}_{t+\Delta t}/S_t\}_k\right)} = \frac{h(t,i)\Delta t}{\sum\limits_{k=1}^{n} h(t,x_i)\Delta t} = \frac{h(t,i)}{\sum\limits_{k=1}^{n} h(t,i)}\\ &= \frac{h_0(t)h(i)}{\sum\limits_{k=1}^{n} h_0(t)h(k)} = \frac{h(i)}{\sum\limits_{k=1}^{n} h(k)}\end{aligned} \tag{5.7.2827}$$

But the reader must keep in mind two things:

1. The arrow of time is oriented
2. Hence the events are ordered!

What does that mean, or what are the implications? Just that the event related to an individual i can occur only before some other event have maybe occur to other individual with a rank index small than i. That means that we should write instead:

$$\frac{P\left(\{\bar{S}_{t+\Delta t}/S_t\}_i\right)}{P\left(\{\bar{S}_{t+\Delta t}/S_t\}_j\right)} = \frac{h(i)}{\sum\limits_{k\geq i}^{n} h(k)} \tag{5.7.2828}$$

or more explicitly:

$$\frac{P\left(\{\bar{S}_{t+\Delta t}/S_t\}_i\right)}{P\left(\{\bar{S}_{t+\Delta t}/S_t\}_j\right)} = \frac{h(i)}{\sum\limits_{k\geq i}^{n} h(k)} = \frac{e^{\beta x_i}}{\sum\limits_{k\geq j} e^{\beta x_j}} \tag{5.7.2829}$$

This relation is often written as following in textbooks:

$$\boxed{P([i],t_i) = \frac{P\left(\{\bar{S}_{t+\Delta t}/S_t\}_i\right)}{P\left(\{\bar{S}_{t+\Delta t}/S_t\}_j\right)} = \frac{e^{\beta x_i}}{\sum\limits_{k\geq j} e^{\beta x_j}} = \frac{\phi_i}{\sum\limits_{k\in\mathcal{R}(t_i)} \phi_j}} \tag{5.7.2830}$$

and is the "risk probability" of individual i at time point t_i, relatively to covariate x (unique covariate in the special case!), where individual events are disjoints with proportional initial hazard and where $\mathcal{R}(t_i)$ is the set of all individuals surviving or functioning at time t_i (risk set at t_i).

The last relation explains the probability we will use to maximize likelihood. Since time-dependent basis hazard does not appear, it is customary to speak of "partial likelihood" of n individuals[101]:

$$\text{PL}(\beta,x)=\prod_{i=1}^{n}\left(\frac{e^{\beta x_i}}{\displaystyle\sum_{j\geq i}e^{\beta x_j}}\right) \tag{5.7.2831}$$

Or in some textbooks it is written (including the case of multiple covariates and possible censored data):

$$\boxed{L_p(\vec{\beta},\vec{x})=\prod_{i=1}^{n}\left(\frac{\phi_i}{\displaystyle\sum_{k\in\mathcal{R}(t_i)}\phi_j}\right)^{\delta_i}=\prod_{i=1}^{n}\left(P([i],t_i)\right)^{\delta_i}} \tag{5.7.2832}$$

where the power δ_i means the individuals having the events (failure/death) actually contribute to the likelihood, but not the right-censored cases.

We can think of the partial likelihood as the joint density function for subjects' ranks in terms of event order, if there were no censoring and no tied event times.

Since likelihood can only be maximized with a computer (Newton's method), it is customary not to go further by taking the natural logarithm (but it would not change the result!). But in practice the coefficient are however estimated by minimizing the negative of the log-likelihood:

$$\ln\left(L_p(\vec{\beta},\vec{x})\right)=-\delta_i\sum_{i=1}^{n}\left[\vec{\beta}^T\vec{x}-\ln\left(\sum_{j\geq i}e^{\vec{\beta}^T\vec{x}_j}\right)\right] \tag{5.7.2833}$$

This writing is still not easy to use in practice. For this reason, we will make a numerical example.

✍**Example:**

To see how, consider the data below with a single covariate (deaths in clinical study in weeks after the start of the test based on BMI[a]):

---

[101] Because we are suppressing the actual times of events even though they are known we use the name "partial likelihood"

i	t_i [weeks]	BMI x_i
1	6	31.4
2	98	21.5
3	189	27.1
4	374	22.7

Remark

Since there are in this small example obviously no qualitative variables, we are talking about "non-stratified time-dependent Cox proportional hazard model".

There are therefore four events with the respective data of BMI 31.4, 21.4, 27.1 and 22.7. As we have seen, time no longer intervenes but only the rank i, that is, the order in which the events of interest take place.

It then corresponds four conditional probabilities which we will denote for the sake of simplification of writing P_1, P_2, P_3, P_4. Therefore:

$$\begin{aligned}
P_1 &= \frac{e^{\beta x_1}}{e^{\beta x_1}+e^{\beta x_2}+e^{\beta x_3}+e^{\beta x_4}} = \frac{e^{\beta\cdot 31.4}}{e^{\beta\cdot 31.4}+e^{\beta\cdot 21.5}+e^{\beta\cdot 27.1}+e^{\beta\cdot 22.7}}\\
P_2 &= \frac{e^{\beta x_2}}{e^{\beta x_2}+e^{\beta x_3}+e^{\beta x_4}} = \frac{e^{\beta\cdot 21.5}}{e^{\beta\cdot 21.5}+e^{\beta\cdot 27.1}+e^{\beta\cdot 22.7}}\\
P_3 &= \frac{e^{\beta x_3}}{e^{\beta x_3}+e^{\beta x_4}} = \frac{e^{\beta\cdot 27.1}}{e^{\beta\cdot 27.1}+e^{\beta\cdot 22.7}}\\
P_4 &= \frac{e^{\beta x_4}}{e^{\beta x_4}} = \frac{e^{\beta\cdot 22.7}}{e^{\beta\cdot 22.7}}
\end{aligned} \tag{5.7.2834}$$

We then have the partial likelihood that is given by:

$$\text{PL}(\beta,x)=\frac{e^{\beta\cdot 31.4}}{e^{\beta\cdot 31.4}+e^{\beta\cdot 21.5}+e^{\beta\cdot 27.1}+e^{\beta\cdot 22.7}}\frac{e^{\beta\cdot 21.5}}{e^{\beta\cdot 21.5}+e^{\beta\cdot 27.1}+e^{\beta\cdot 22.7}}\frac{e^{\beta\cdot 27.1}}{e^{\beta\cdot 27.1}+e^{\beta\cdot 22.7}}\frac{e^{\beta\cdot 22.7}}{e^{\beta\cdot 22.7}} \tag{5.7.2835}$$

And a numerical application that maximizes this with the solver of the spreadsheet software Microsoft® Excel® (or other spreadsheet or tool) immediately gives:

$$\hat{\beta}\cong 0.2227 \tag{5.7.2836}$$

This is consistent with what returns statistics software like MedCalc 12.7.5 or R (see the example in the corresponding companion book)! This gives us:

$$h(t,x)\cong h_0(t)e^{0.2227x} \tag{5.7.2837}$$

From there we can easily compare at a given moment the ratio of chance between two individuals with different characteristics.

We can also give the corresponding survival plot (here made with R and including the confidence interval that we will see further below):

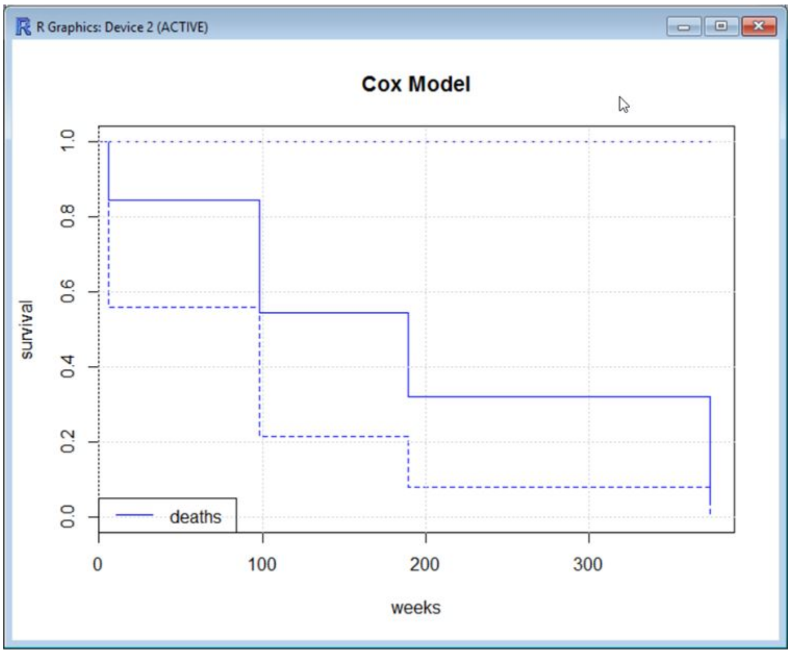


Figure 5.252: **Cox PH typical survival curves in R software**

---

[a] Body Mass Index

We notice after this short overview that at no moment we needed to know or assume the underlying distribution law (Weibull, Exponential or other) and their associated parameters. The Cox model is therefore a nonparametric technique.

It exists (although at the base it is not the goal of the Cox model), estimators of $h_0(t)$, the best known of which is the Breslow estimator, but it is considered unreliable (it would have been proved it is inconsistent) and applicable only for the cumulative rate of hazard and not the instantaneous rate (so beware of softwares that give $h_0(t)$....). And the interest of the Cox model is to be interested in the risk ratio and not in estimating the complete survival/reliability function hazard (Kaplan-Meier is then better adapted). So actually it is not planned to have a detailed introduction of this estimators (and some other one) in this textbook!

Remark

In the R software, the function `basehaz()` defaults to the Nelson-Aalen estimate of the cumulative hazard with a Breslow-type estimate of survival. Alternatively, a Kaplan-Meier estimator for the cumulative hazard can be selected.

Let us now discuss a bit about simultaneous events! In principle, the durations are supposed to be the realizations of a continuous variable and the covariates also (that continuous underlying variable being the time obviously), the probability that two or more individuals know the event at the same moment is theoretically null. In practice, the data are often recorded according to a certain time division (weekly, monthly data, etc.) so that it is common for the same duration of survival to characterize several different individuals.

Basically this does not seriously complicate the previous writing, simply the number of calculations can become such if one uses the exact contribution of these individuals to partial likelihood that approximations have been proposed (available all three in SAS software). However as part of this textbook we will focus on the exact version!

So suppose that for a certain time, two individuals have known the event interest and that at the same time, three other individuals are also at risk. If we start by treating the case of the first individual, then there are 5 individuals at risk. When we treat the case of the second individual, only 4 individuals are risky since the first one has already been taken into account. In this case, the probability of having both events simultaneously is consistent with that of 2 non-incompatible but independent events (see section Probabilities page 477):

$$\frac{e^{\beta x_1}}{e^{\beta x_1}+e^{\beta x_2}+e^{\beta x_3}+e^{\beta x_4}+e^{\beta x_4}}\cdot\frac{e^{\beta x_2}}{e^{\beta x_2}+e^{\beta x_3}+e^{\beta x_4}+e^{\beta x_4}} \tag{5.7.2838}$$

Naturally, the fact of starting with the first individual is arbitrary since one can just as easily admit to starting with the second. In total, the contribution of this duration is two individuals at partial likelihood then is:

$$\frac{e^{\beta x_1}}{e^{\beta x_1}+e^{\beta x_2}+e^{\beta x_3}+e^{\beta x_4}+e^{\beta x_4}}\cdot\frac{e^{\beta x_2}}{e^{\beta x_2}+e^{\beta x_3}+e^{\beta x_4}+e^{\beta x_4}}+\frac{e^{\beta x_2}}{e^{\beta x_1}+e^{\beta x_2}+e^{\beta x_3}+e^{\beta x_4}+e^{\beta x_4}}\cdot\frac{e^{\beta x_1}}{e^{\beta x_2}+e^{\beta x_3}+e^{\beta x_4}+e^{\beta x_4}} \tag{5.7.2839}$$

Let us note (for those who may have a doubt that it is always smaller or equal to unity as required by the axiom of probabilities) than in the case where we have a population with two individuals this reduces to:

$$\frac{e^{\beta x_1}}{e^{\beta x_1}+e^{\beta x_2}}\frac{e^{\beta x_2}}{e^{\beta x_2}}+\frac{e^{\beta x_2}}{e^{\beta x_1}+e^{\beta x_2}}\frac{e^{\beta x_1}}{e^{\beta x_1}}=\frac{e^{\beta x_1}}{e^{\beta x_1}+e^{\beta x_2}}+\frac{e^{\beta x_2}}{e^{\beta x_1}+e^{\beta x_2}}=1 \tag{5.7.2840}$$

which is consistent with the axiom of certain event in the field of probabilities! And when the population is larger, it is immediate that the denominator of each term will also larger, the total probability is then still less or equal to unity!

If we repeat the manoeuvre with three individuals simultaneously, we will have 6 terms that will add up. More generally, if there are k individuals among n which lives the event of interest at the same time, the calculation of the contribution involves $k!$ terms. The calculation time can therefore become very penalizing, for example with k worth 10, we have $3,628,800$ terms...

Now as for the Kaplan-Meier survival estimator, let us see how we can construct the confidence intervals for the hazard ratio.

As earlier we will use again the univariate Delta method (see section Numerical Methods page 6504) to get standard error for $e^{\hat{\beta}}$. Let us recall that this method gives us that:

$$g(X_n)\underset{n\rightarrow+\infty}{\cong}\mathcal{N}\left(0,g'(\mu)^2\sigma^2\right) \tag{5.7.2841}$$

This gives immediately for the variance:

$$\hat{\mathrm{V}}\left(e^{\hat{\beta}}\right)=e^{2\hat{\beta}}\hat{\mathrm{V}}\left(\hat{\beta}\right) \tag{5.7.2842}$$

Hence assuming that the β are Normally distributed (and is a scalar...):

$$e^{\hat{\beta}}-Z_{\alpha/2}\sqrt{e^{2\hat{\beta}}\hat{\mathrm{V}}\left(\hat{\beta}\right)}\leq e^{\beta}\leq e^{\hat{\beta}}+Z_{\alpha/2}\sqrt{e^{2\hat{\beta}}\hat{\mathrm{V}}\left(\hat{\beta}\right)} \tag{5.7.2843}$$

The problem here is that we can have the lower bound of e^{β} that can be negative...

Hence we prefer to use:

$$e^{\hat{\beta}-Z_{\alpha/2}\text{SE}\left(\hat{\beta}\right)} \leq e^{\beta} \leq e^{\hat{\beta}+Z_{\alpha/2}\text{SE}\left(\hat{\beta}\right)} \tag{5.7.2844}$$

If we consider individuals that are identical apart from one covariate x, the hazard ratio becomes as we already know:

$$\frac{h_i}{h_j}=\frac{e^{\beta x_i}}{e^{\beta x_j}}=e^{\beta(x_i-x_j)} \tag{5.7.2845}$$

The confidence interval of the hazard ratio is therefore:

$$\boxed{e^{\left(\hat{\beta}-Z_{\alpha/2}\text{SE}\left(\hat{\beta}\right)\right)(x_i-x_j)} \leq \text{H.R.} \leq e^{\left(\hat{\beta}+Z_{\alpha/2}\text{SE}\left(\hat{\beta}\right)\right)(x_i-x_j)}} \tag{5.7.2846}$$

✍**Example:**

A Cox proportional hazard model was used to model the survival times of cancer patients. Tumour size (in [mm]) was included as a covariate, with coefficient β. The maximum likelihood estimate of β was $\hat{\beta}=0.0176$ with standard error 0.004.

We want to calculate an estimate of the hazard ratio between two individuals with tumours measuring 46 [mm] and 37 [mm] who are identical in other ways and construct a 95% confidence interval for the hazard ratio.

We have for hazard ratio:

$$\frac{e^{46\beta}}{e^{37\beta}}=e^{9\beta} \tag{5.7.2847}$$

Using the estimate:

$$e^{9\hat{\beta}}=e^{0.162}=1.176 \tag{5.7.2848}$$

The 95% confidence interval for β is:

$$0.0176\pm 1.96\cdot 0.004 \tag{5.7.2849}$$

That is:

$$0.01016\leq\beta\leq 0.02584 \tag{5.7.2850}$$

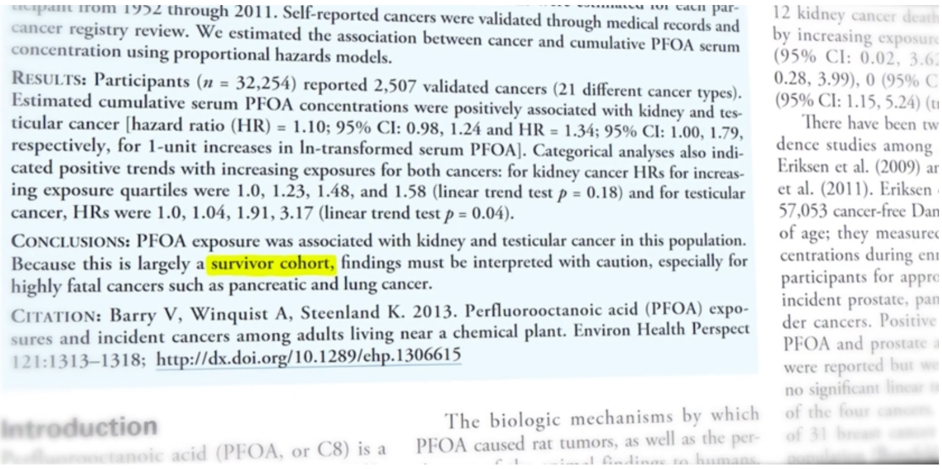
The 95% confidence interval for hazard ratio is therefore:

$$1.096\leq e^{9\beta}\leq 1.262 \tag{5.7.2851}$$

Below we can see a print screen of a famous scientific paper providing evidence using the Hazard Ratio that eternal chemicals like PFOA[102] (all humans and animals around the world are intoxicated with these chemicals since the early 121st century according to holocene calendar) are strongly correlated with multiple types of cancers for people having a dozen or a hundred of part per millions of these chemical in their blood... (this evidence has cost hundred of millions of dollars to their manufacturers):

---

[102] PFOA, or perfluorooctanoic acid, is a synthetic chemical classified as a per- and polyfluoroalkyl substance (PFAS).



from 1952 through 2011. Self-reported cancers were validated through medical records and cancer registry review. We estimated the association between cancer and cumulative PFOA serum concentration using proportional hazards models.

RESULTS: Participants ($n = 32{,}254$) reported 2,507 validated cancers (21 different cancer types). Estimated cumulative serum PFOA concentrations were positively associated with kidney and testicular cancer [hazard ratio (HR) = 1.10; 95% CI: 0.98, 1.24 and HR = 1.34; 95% CI: 1.00, 1.79, respectively, for 1-unit increases in ln-transformed serum PFOA]. Categorical analyses also indicated positive trends with increasing exposures for both cancers: for kidney cancer HRs for increasing exposure quartiles were 1.0, 1.23, 1.48, and 1.58 (linear trend test $p = 0.18$) and for testicular cancer, HRs were 1.0, 1.04, 1.91, 3.17 (linear trend test $p = 0.04$).

CONCLUSIONS: PFOA exposure was associated with kidney and testicular cancer in this population. Because this is largely a survivor cohort, findings must be interpreted with caution, especially for highly fatal cancers such as pancreatic and lung cancer.

CITATION: Barry V, Winquist A, Steenland K. 2013. Perfluorooctanoic acid (PFOA) exposures and incident cancers among adults living near a chemical plant. Environ Health Perspect 121:1313–1318; http://dx.doi.org/10.1289/ehp.1306615

5.7.16.3.1 Stratified Cox Model

Suppose the observations of a variable define sub-categories of the initial population. An obvious example is given by a sex covariate coded 0 or 1.

Suppose again that one is not interested in a direct measure of the impact of this variable on risk or that one thinks that this variable does not verify the proportional risks hypothesis so that we refuse its inclusion in the list of covariates (in other words, instead of assuming that the *sex* binary covariate on the survival risk is constant in time and multiplicative on the risk, we can assume that it is the basis risk that is different by men and women and therefore we have two different stratum!).

It is possible, however, that we want to take into account an effect of the sex of the individuals to measure the impact of the other variables. The solution as already mentioned in the parenthesis before is then to stratify. In this case the explanatory coefficients are constrained to be identical in each of the strata but a partial likelihood function is constructed separately on each of them, the function that is maximized to find the parameter estimators being the product of these partial likelihoods. As a result, obviously, the basic risk differs between strata. For example, instead of having (special simple case with two covariates):

$$h(t,\vec{x})=h_0(t,\vec{\alpha})e^{\beta_1x_1+\beta_2x_2} \tag{5.7.2852}$$

where x_2 would be the *sex* covariate, we have rather have the following system:

$$\begin{gathered}
h(t,\vec{x})=h(t|x_2=0,x_1)=h_{01}(t|x_2=0,x_1)e^{\beta_1x_1}\\
h(t,\vec{x})=h(t|x_2=1,x_1)=h_{02}(t|x_2=1,x_1)e^{\beta_1x_1+\beta_2}=h_{02}(t|x_2=1,x_1)e^{\beta_2}e^{\beta_1x_1}=h'_{02}(t|x_2=1,x_1)e^{\beta_1x_1}
\end{gathered} \tag{5.7.2853}$$

In summary:

$$\begin{aligned}
\text{Stratum 1}: h(t,\vec{x})&=h_{01}(t|x_2=0,x_1)e^{\beta_1x_1}\\
\text{Stratum 2}: h(t,\vec{x})&=h'_{02}(t|x_2=1,x_1)e^{\beta_1x_1}
\end{aligned} \tag{5.7.2854}$$

Thus, for two individuals of the same stratum we will have more chance to have a proportional hazard (parallel survival curves) as the ratio cancels the baseline hazard. On the other hand, it is obvious that the comparison of two individuals of different strata will have very little chance of being independent of time since the ratio of basic hazards will not cancel each other out.

However if they cancel each other out, we have then a kind of an evidence that their are proportional and that it is reasonable to model the stratification term in the regular manner!

Therefore "stratified Cox models":

$$h_k(t,\vec{x}) = h_{0k}(t,\vec{\alpha})e^{\vec{\beta}^T\vec{x}} \tag{5.7.2855}$$

where h_{0k} is the baseline hazard for stratum k are a useful extension of the standard Cox models (where for recall h_0 is common to all the individuals in a study):

$$h(t,\vec{x}) = h_0(t,\vec{\alpha})e^{\vec{\beta}^T\vec{x}} \tag{5.7.2856}$$

to allow for covariates with non-proportional hazards.

The partial likelihood that we have introduce earlier:

$$L_p(\vec{\beta},\vec{x}) = \prod_{i=1}^{n}\left(\dfrac{e^{\vec{\beta}^T\vec{x}_i}}{\sum\limits_{j\geq i} e^{\vec{\beta}^T\vec{x}_j}}\right)^{\delta_i} \tag{5.7.2857}$$

is computed for stratified model as:

$$L_{p,\text{St.}}(\vec{\beta},\vec{x}) = \prod_{k=1}^{K}\prod_{i=1}^{n}\left(\dfrac{e^{\vec{\beta}^T\vec{x}_{kj}}}{\sum\limits_{j\geq i} e^{\vec{\beta}^T\vec{x}_{kj}}}\right)^{\delta_i} \tag{5.7.2858}$$

Finally, in the case of stratification, since the responsible variable of stratification is not explicitly measured, we can not use it as an explanatory variable.

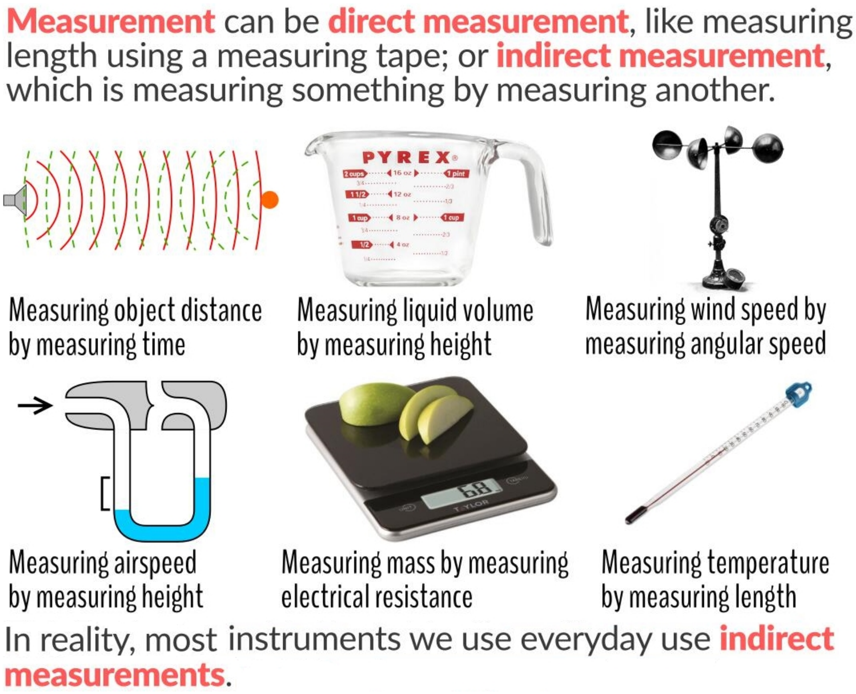
So what we can summarize so far? We have seen three kinds of models with for each at least one special mathematical model (but there are much more in reality when you become an expert in the field as we have already shown it in a figure earlier summarizing all most famous survival analysis methods!):

- Parametric models: Where the baseline hazard and covariates are all assumed to have the same distributions (e.g. Weibull or Exponential models most of the time and this is typically the case of the Threshold Regression Models).
- Nonparametric models: Where the baseline hazard and the covariates have no specific distribution assumed (e.g. Kaplan-Meier model).
- Semi-Parametric models: Where the baseline hazard has no special distribution assumed but the covariates have some distribution assumed (e.g. Cox proportional hazard model).

5.7.17 Propagation of Errors (experimental uncertainty analysis)

It is impossible to know (measure) the exact value of a physical quantity experimentally, it is very important therefore to determine its uncertainty.

We obviously name "error", the difference between the measured value and the exact value. However, since we do not know the exact value, we can not know the error anyway .... The result is still uncertain. That is why we speak of "measurement uncertainty".



Measurement can be **direct measurement**, like measuring length using a measuring tape; or **indirect measurement**, which is measuring something by measuring another.

Measuring object distance by measuring time
Measuring liquid volume by measuring height
Measuring wind speed by measuring angular speed
Measuring airspeed by measuring height
Measuring mass by measuring electrical resistance
Measuring temperature by measuring length

In reality, most instruments we use everyday use **indirect measurements**.

We distinguish two main types of uncertainty:

1. The "systematic errors": they affect the result and this constantly in the same direction (errors of measurement devices, accuracy limits, etc.). It must then be eliminate, or correct the outcome, if possible!
2. The (statistical) "accidental errors": we must the repeat the measurements and calculate the average estimate uncertainty using statistical techniques.

The second type of error makes a very big use of all statistical tools we have presented so far. So we will not repeat them here and then we will focus only on a few new concepts.

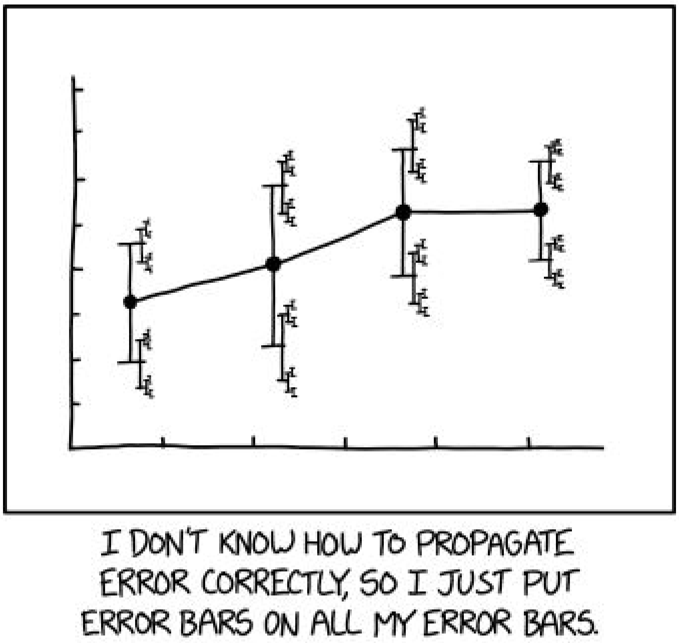

I DON'T KNOW HOW TO PROPAGATE ERROR CORRECTLY, SO I JUST PUT ERROR BARS ON ALL MY ERROR BARS.

Figure 5.253: **Errors propagation (source: xkcd)**

5.7.17.1 Absolute and Relative Uncertainties (direct calculation of bias)

If the true value of a variable is x (theoretically supposed to be known) and the measured value x_0, then δx_0 is the "absolute uncertainty" (uncertainty due to measurement devices) or "absolute error".

The interval of fluctuation is therefore denoted by:

$$x_0 - \Delta x_0 \leq x \leq x_0 + \Delta x_0 \tag{5.7.2859}$$

or:

$$x = x_0 \pm \Delta x \tag{5.7.2860}$$

The "uncertainty" or "relative error" is itself defined by:

$$\frac{\Delta x_0}{x_0} \tag{5.7.2861}$$

The absolute uncertainty is used to find an approximation of the last significant digit thereof. By cons, when we want to compare two measurements having absolute uncertainties to identify which was the largest margin of error, we calculate the relative uncertainty of this number by dividing the absolute uncertainty by the measurement itself, and transform the result typically in percentage.

In other words, the relative uncertainty gives an idea of the accuracy of the measurement in %. If we make a measurement with an absolute uncertainty of 1 [mm], we will not know if this is a good measure or not. It depends if we measured the size of a coin, of our neighbour, or of the Paris-Marseille distance of the Earth-Moon distance. In short, it depends on the relative uncertainty (that is to say the ratio of the absolute uncertainty and of the measurement).

Therefore in the case of a law of the type $f(x,y,z,t)$ for which we seek the error we would calculate the following (one time with everything positive and another with all terms negative):

$$\frac{\Delta f(x,y,z,t)}{f(x,y,z,t)}=\frac{f(x\pm\Delta x,y\pm\Delta y,z\pm\Delta y,t\pm\Delta t)}{f(x,y,z,t)} \tag{5.7.2862}$$

This method is quite boring as it must be calculated for each value (x,y,z,t).

5.7.17.2 Statistical Errors

In most measurement, we can estimate the error due to random phenomena, named "random error", by a series of n measurements $x_1,x_2,\ldots,x_i,\ldots,x_n$ and this opposed to the "systematic error" which is the not random part of the error.

Random error allows to introduce the concepts of:

- Repeatability: This is defined as the closeness of agreement between the results of successive measurements of a same item, made with the same method, by the same operator with the same measuring instruments, in the same laboratory (conditions), and in rather short time intervals (see a little further below a little bit more rigorous definition in line with international standards).
- Reproducibility (sometimes named a "rightness"): which is defined as the closeness of agreement between the results of successive measurements of the same quantity, in the case where individual measurements are made: by different methods, using different instruments extent by different operators in different laboratories!

These two type of errors are almost always grouped under the labels "R&R" or "R&R Study" in the industry. In general, the agreement is less good when it comes to reproducibility.

Remark

There are softwares running two-way fixed factor ANOVA with repetition as Minitab that generate very detailed reports for R&R analysis.

These two types of errors can be illustrated by the target shooting in a more general way:

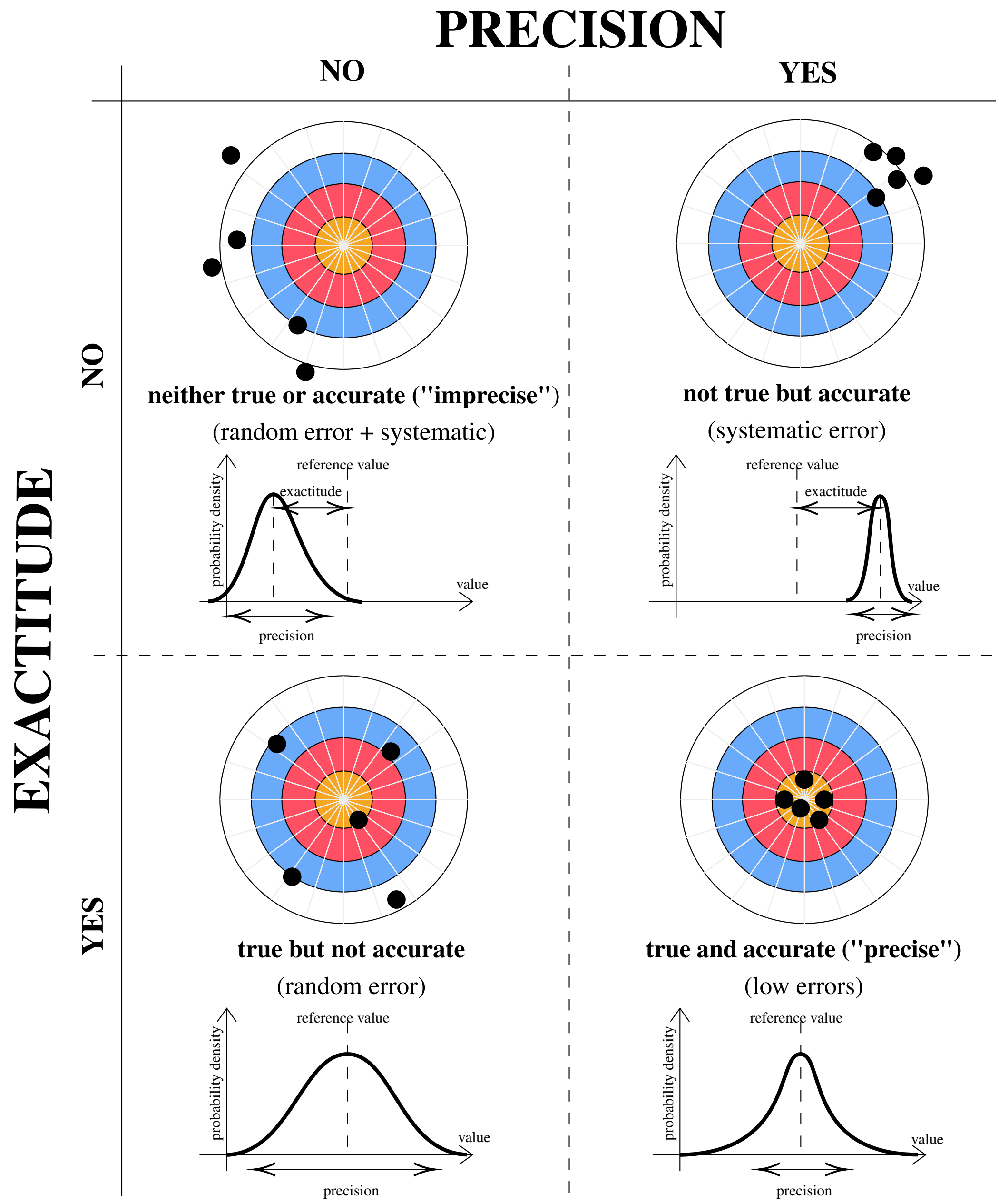

PRECISION
NO
YES
EXACTITUDE
NO
YES
neither true or accurate ("imprecise")
(random error + systematic)
reference value
exactitude
probability density
value
precision
not true but accurate
(systematic error)
reference value
exactitude
probability density
value
precision
true but not accurate
(random error)
reference value
probability density
value
precision
true and accurate ("precise")
(low errors)
reference value
probability density
value
precision

Figure 5.254: Different types of measurement errors

Remark

Measurement error is the difference between a measured value & true value. Measurement error is not a "mistake". Variability is an inherent part of any measurement process and can never be eliminated!

Below an illustration of the same Indonesian Rp 500 coin using three different measuring devices:

3 cm ± 0.5 cm 27 mm ± 0.5 mm 27.32 mm ± 0.005 mm

As we have seen earlier above, the arithmetic mean value will be in the univariate case:

$$\bar{x}=\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i \tag{5.7.2863}$$

and the standard deviation (biased estimator as proved earlier) always in the univariate case (the nth case has already been proved during our detailed study of the variance):

$$\hat{\sigma}^2=\overline{\Delta x}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{5.7.2864}$$

and the unbiased standard deviation (as also proven earlier):

$$\sigma^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{5.7.2865}$$

and we have proved that the standard deviation of the mean was given by (under some assumptions!):

$$\sigma_{\bar{x}}=\frac{\sigma}{\sqrt{n}}=\frac{1}{\sqrt{n}}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}=\sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{5.7.2866}$$

and as we have proved it, after a large number of independent measurements, the distribution of errors on a measure follows a Normal distribution so that we can write for the fluctuation interval (if we do not have enough measurement, we then use the fluctuation interval based on the Student law):

$$\begin{array}{cr}\bar{x}_i-\sigma<x_i<\bar{x}_i+\sigma & 68\% \\ \bar{x}_i-2\sigma<x_i<\bar{x}_i+2\sigma & 95.4\%\end{array} \tag{5.7.2867}$$

In short, we can use all statistical tools seen so far in the field of measurement in laboratories or elsewhere!

The result of a measurement (or estimation and this even in the field of project management!!!) must include rigorously at least 4 elements. For example for a symmetrical error:

$$\underbrace{125.3}_{1}\;\underbrace{[\text{kg}]}_{2}\pm\underbrace{1.7}_{3}\,[\text{kg}]\;\underbrace{(k=2)}_{4} \tag{5.7.2868}$$

or an asymmetrical one:

$$125.3^{+1.7}_{-2.1}\,[\text{kg}](k=2) \tag{5.7.2869}$$

Where we have:

1. The numerical value with the correct number of decimal places.
2. Unit of measurement according to the international measurement system.
3. Expanded uncertainty of $k \cdot \sigma$ (fluctuation interval)
4. The integer value of k used for fluctuation interval.

This method is quite more useful than the previous one as we don't need to recalculate it each time for each experimental value. It's this method that for example the famous CERN (European Organization for Nuclear Research) use[103].

In astrophysics papers we can also found the following notation separating the part of the variation from the measurement tool and from what is measured:

$$125.3^{+1.7}_{-2.1}(k=2) \pm 6 \text{ (systematic) [kg]} \tag{5.7.2870}$$

or (always in astrophysics and cosmology) we can also typically see values given as:

$$0.09 \pm 0.25(1\sigma) \pm 0.49(2\sigma) \tag{5.7.2871}$$

Caution! It happens many many many times... in scientific papers and also in textbooks that a measurement is provided without indication of its approximation. It is then generally admitted in practice that the absolute incertitude isn't greater than a unit of the power order indicated by the last digit.

5.7.17.3 Repeatability

The repeatability r, the likely measurement difference between two measurements of similar objects in the same laboratory under similar operating conditions, is normatively defined (in the norms ISO 5725:1987 and AFNOR NF X 06-041) in the one-dimensional (univariate) case by:

$$P\left(|X_1 - X_2| < r\right) = p \tag{5.7.2872}$$

where p is a high value probability, usually equal to 95% and X_1, X_2 two independent and identically distributed random variables according to a Normal distribution with unknown mean and unknown variance $\mathcal{N}(\mu,\sigma)$. By the stability of the Normal distribution, then it comes:

$$X_1 - X_2 = \mathcal{N}(\mu,\sigma) - \mathcal{N}(\mu,\sigma) = \mathcal{N}\left(0, \sqrt{2}\sigma\right) \tag{5.7.2873}$$

But, we saw earlier in this section in the context of the study of the confidence interval of the mean:

$$P\left(Y \geq \frac{\delta}{\sigma}\right) = 0.025 \tag{5.7.2874}$$

---

[103] Add to this the Six Sigma measurement methodology studied in the Industrial Engineering section

So verbatim:

$$P\left(Y\leq\frac{\delta}{\sigma}\right)=0.975 \tag{5.7.2875}$$

and therefore using the tables, we have:

$$\frac{\delta}{\sigma}\cong 1.96 \tag{5.7.2876}$$

and therefore:

$$\delta\cong 1.96\sigma \tag{5.7.2877}$$

Either with the notation respecting the norms for laboratories:

$$r\cong 1.96\sigma \tag{5.7.2878}$$

But in the present case, we have a double variance. So it comes:

$$\boxed{r\cong 1.96\sqrt{2}\sigma\cong 2.77\sigma} \tag{5.7.2879}$$

So we fall back on the relation available in the norms with the famous 2.77 coefficient. Obviously after it is clear that the value of r should be minimized!

5.7.17.4 Error propagation (linearised approximation)

Given a measurement $x+\delta x$ and $y=f(x)$ a function of x. What is the uncertainty on y it if we know only the uncertainty of a measurement device, but that would not be given as a statistical standard deviation?

In this type of situation we speak of "indirect measurement". This is typically the case if we want to measure an intensity I by measuring it indirectly by making the ratio of the voltage U by the resistance R used for the measurement as $I=U/R$. It is indeed obvious that in the latter situation we can not make the sum of the voltage incertitude and resistance incertitude because the system is not homogeneous at the unit level!!!

When δx is small, $f(x)$ is replaced in the neighbourhood of x by its tangent (it is simply the derivative of course!):

$$\Delta y=\left|\frac{\mathrm{d}f}{\mathrm{d}x}\right|\Delta x \tag{5.7.2880}$$

but if y depends of several variables x,z,t measured with the uncertainties $\delta x,\delta z,\delta t$:

$$y=f(x,z,t) \tag{5.7.2881}$$

the maximum possible error is then the exact total differential (see section of Differential and Integral Calculus page 1466):

$$\mathrm{d}y=\left|\frac{\partial f}{\partial x}\right|\mathrm{d}x+\left|\frac{\partial f}{\partial z}\right|\mathrm{d}z+\left|\frac{\partial f}{\partial t}\right|\mathrm{d}t \tag{5.7.2882}$$

Using first order Taylor expansion series (see section Sequences and Series page 1696) we can write:

$$\Delta y \cong \left|\frac{\partial f}{\partial x}\right| \Delta x+\left|\frac{\partial f}{\partial z}\right| \Delta z+\left|\frac{\partial f}{\partial t}\right| \Delta t \tag{5.7.2883}$$

What we denote often as the sum of the partial derivatives with their respective uncertainty:

$$\Delta y \cong \sum_{i=1}^{n}\left|\frac{\partial y}{\partial x_i}\right| \Delta x_i \tag{5.7.2884}$$

and this works very well as long as the increments Δx are sufficiently small. Even highly curved functions are nearly linear over a small enough region!

The latter relation is the "propagation law" of the studied problem. The partial derivative in factor of the uncertainty is in the science of measurement, named the "uncertainty coefficient".

For sure we can also write to get the relative error:

$$\boxed{\frac{\Delta y}{y} \cong \sum_{i=1}^{n} \frac{1}{y}\left|\frac{\partial y}{\partial x_i}\right| \Delta x_i} \tag{5.7.2885}$$

Which bring us to:

$$\begin{cases} y & = x+z \Longrightarrow \Delta y=\Delta x+\Delta z \\ y & = x-z \Longrightarrow \Delta y=\Delta x+\Delta z \end{cases} \Rightarrow \frac{\Delta y}{y}=\frac{\Delta x+\Delta z}{x \pm y} \tag{5.7.2886}$$

and (it is not intuitive at the first sight that the result for the product is the same as for the ratio):

$$\begin{cases} y & = xz \Longrightarrow \Delta y=z\Delta x+x\Delta z \\ y & = \dfrac{x}{z} \Longrightarrow \Delta y=\dfrac{z\Delta x+x\Delta z}{z^2} \end{cases} \Rightarrow \frac{\Delta y}{y}=\frac{\Delta x}{x}+\frac{\Delta z}{z} \tag{5.7.2887}$$

Therefore the relative incertitude of a product or of a quotient of uncertain values is equal to the sum of the relative incertitudes. It is thus clear that a mathematical operation cannot improve the uncertainty of the data.

For the above we deduce immediately:

$$y=x^n \Rightarrow \frac{\Delta y}{y}=|n|\frac{\Delta x}{x} \tag{5.7.2888}$$

and therefore:

$$g=x^n y^m z^p \ldots \Rightarrow \frac{\Delta g}{g}=|n|\frac{\Delta x}{x}+|m|\frac{\Delta y}{y}+|p|\frac{\Delta z}{z}+\ldots \tag{5.7.2889}$$

Here is a summary of this as a table (with a small change in notations):

$z=$	$x+y$	$x-y$	xy	$\dfrac{x}{y}$
$\Delta z=$	$\Delta x+\Delta y$	$\Delta x+\Delta y$	$x\Delta y+y\Delta x$	$\dfrac{x\Delta y+y\Delta x}{y^2}$
$\dfrac{\Delta z}{z}=$	$\dfrac{\Delta x+\Delta y}{x+y}$	$\dfrac{\Delta x+\Delta y}{x-y}$	$\dfrac{x\Delta y+y\Delta x}{xy}=\dfrac{\Delta x}{x}+\dfrac{\Delta y}{y}$	$\dfrac{\Delta x}{x}+\dfrac{\Delta y}{y}$

Table 5.125: **Absolute and Relative errors common relations**

Remark

The result of multiplication, division, subtraction or addition is rounded to as many significant digits that the data that has the smallest number of them.

Obviously this propagation law (linear) is valid only in the range where the function can be approximated as linear. So be careful in its use! Otherwise you have to take a Taylor series approximation of higher order.

If the uncertainty of the measurement is given in a statistic form (standard deviation), it is evident therefore that we will use the variance properties already seen at the beginning of this section... at least for simple cases.

Once again the pitfall of this method in comparison to the statistical one is that you have the calculate the error for each measurement...

✍**Example:**

We want to calculate the incertitude of the measurement of the density of a cube of a given material. We then know that:

$$\rho=\frac{m}{V}=\frac{m}{a^3}\Rightarrow\frac{\Delta\rho}{\rho}=\frac{\Delta m}{m}+3\frac{\Delta a}{a} \tag{5.7.2890}$$

Now let's deal with another famous case in astronomy, astrophysics and cosmology where we often have to deal with trigonometric functions!

Consider we have measurements in the form $\cos(\phi)\pm\Delta\cos(\phi)$ and we need to find $\phi\pm\Delta\phi$.

Using the above relation we obviously get:

$$\left|\frac{\Delta\cos(\phi)}{\Delta\phi}\right|=\sin(\phi) \tag{5.7.2891}$$

where we assume $\{\phi\neq 0^\circ,90^\circ\}$ so that:

$$|\Delta\cos(\phi)|=\sin(\phi)\Delta\phi \tag{5.7.2892}$$

In experimental setup we normally have $\Delta\cos(\phi)=\hat{\sigma}_{\cos(\phi)}$. Hence:

$$\Delta\phi=\frac{\hat{\sigma}_{\cos(\phi)}}{\sin(\phi)} \tag{5.7.2893}$$

As in astronomy, astrophysics and cosmology we deal with very small angle angles, the computed uncertainty will be large (since the the expression for $\Delta\phi$ becomes large if the denominator approaches zero) so an additional term to the $|\Delta\cos(\phi)|$ expression is added so that we now have the equation (Taylor series expansion about $\phi\cong 0$ - second term is the second derivative of original $\cos$ term):

$$|\Delta\cos(\phi)|\cong\sin(\phi)\Delta\phi+\frac{1}{2}\cos(\phi)(\Delta\phi)^2 \tag{5.7.2894}$$

so that for small angles, i.e., $\cos(\phi)\cong 1$, the quadratic term dominates!

5.7.17.5 Error propagation (statistical approach)

Finally, there is one additional point to discuss: Suppose we measure a quantity u several times, or by several different methods, and for each measurement u_i we estimate its uncertainty σ_i. The σ_i are not necessarily equal; some of the measurements will be better than others, because of larger sample sizes (more repetitions), or because of other factors—like better apparatus. How do we determine our best estimate of u, and how do we find the uncertainty in that estimate?

For example, suppose a length x is measured by one person n_1 times and by another person n_2 times, so that the first person finds:

$$u_1 := \bar{x}_1 = \frac{1}{n_1}\sum_k x_k \qquad \text{with} \qquad \sigma_1^2 = \frac{1}{n_1}\sigma^2 \tag{5.7.2895}$$

while the second person finds:

$$u_2 := \bar{x}_2 = \frac{1}{n_2}\sum_j x_j \qquad \text{with} \qquad \sigma_2^2 = \frac{1}{n_2}\sigma^2 \tag{5.7.2896}$$

Here σ_1 is the uncertainty in u_1, σ_2 is the uncertainty in u_2, and σ^2 is the population variance of the x-values. How should u_1 and u_2 be combined to yield an overall u, and what is the uncertainty in this final $\bar{u}$? Since $n_1 = \sigma^2/\sigma_1^2$ and $n_2 = \sigma^2/\sigma_2^2$, we have:

$$\bar{u} = \frac{1}{n_1+n_2}\left(\sum_k x_k + \sum_j x_j\right) = \frac{1}{n_1+n_2}(n_1u_1+n_2u_2) = \frac{1}{\dfrac{1}{\sigma_1^2}+\dfrac{1}{\sigma_2^2}}\left(\frac{u_1}{\sigma_1^2}+\frac{u_2}{\sigma_2^2}\right) \tag{5.7.2897}$$

with by definition:

$$\sigma_u^2 = \frac{1}{\dfrac{1}{\sigma_1^2}+\dfrac{1}{\sigma_2^2}} \tag{5.7.2898}$$

In general, if there are n values of u, here is the generalized result, in a form that depends only on each u_k and its uncertainty σ_k:

$$\boxed{\bar{u} = \frac{\displaystyle\sum_{k=1}^{n}\frac{u_k}{\sigma_k^2}}{\displaystyle\sum_{k=1}^{n}\frac{1}{\sigma_k^2}} \qquad \sigma_{\bar{x}}^2 = \frac{1}{\displaystyle\sum_{k=1}^{n}\frac{1}{\sigma_k^2}}} \tag{5.7.2899}$$

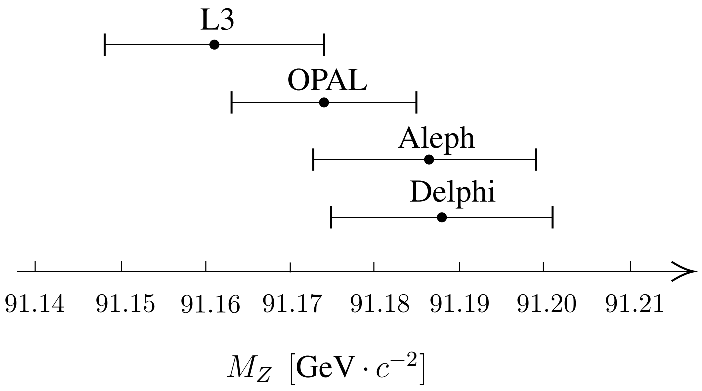
Note how more measurements, or more accurate measurements, reduce the uncertainty by increasing its reciprocal.

✍Example:

The field of particle physics provides numerous situations where the previous relations and also the χ^2 can be applied. A particularly simple example[a] involves measurements of the mass M_Z of the Z^0 boson by experimental groups at CERN. The results of measurements of M_Z made by four different detectors (L3, OPAL, Aleph and Delphi) are as follows[b]:

Detector	**Mass in GeV/c^2**
L3	91.161 ± 0.013
OPAL	91.174 ± 0.011
Aleph	91.186 ± 0.013
Delphi	91.188 ± 0.013

The listed uncertainties are estimates of the σ_i, the standard deviations for each of the measurements. The figure below shows these measurements plotted on a horizontal mass scale (vertically displaced for clarity):



The question arises: Can these data be well described by a single number, namely an estimate of M_Z made by determining the weighted mean of the four measurements?

We know that we can calculate the weighted mean $\bar{M}_Z$, and its standard deviation $\sigma_{\bar{M}_Z}$ like this:

$$\bar{M}_Z = \frac{\sum M_i/\sigma_i^2}{\sum 1/\sigma_i^2} \quad \text{and} \quad \sigma^2_{\bar{M}_Z} = \frac{1}{\sum 1/\sigma_i^2} \tag{5.7.2900}$$

to find:

$$\bar{M}_Z \pm \sigma_{\bar{M}_Z} = 91.177 \pm 0.006 \tag{5.7.2901}$$

Then we form χ^2:

---

[a] This example was provided by Patricia Burchat, a UCSC physics faculty member from 11988-11994 (according to holocene calendar), and now at Stanford.

[b] More recent measurements are listed at http://pdg.lbl.gov/2014/listings/rpp2014-list-z-boson.pdf

$$\chi^2 = \sum_{i}^{4} \frac{M_i - \bar{M}_Z}{\sigma_i^2} \cong 2.78 \tag{5.7.2902}$$

We know that we must expect this value of χ^2 to be drawn from a chi-square distribution with 3 degrees of freedom.

This correspond to a p-value of 0.426803. The result is not significant at $p < 0.05$. We cannot therefore reject the null hypothesis that the four measurements of the Z^0 boson mass are consistent with each other.

5.7.17.6 Significant Numbers

In small schools (and sometimes of higher level and in corporations), it is required to transform a measurement expressed in a certain unit into another unit.

For example, taking the tables, we can have the following conversion:

$$140\,[\text{lb}] = 140 \cdot 0.45349237\,[\text{kg/lb}] = 63.5029318\,[\text{kg}]$$

Then comes the question (that the student or practitioner may have forgotten ...). Starting from a measurement with an accuracy of about 1 [lb] (therefore of the order of 0.5 [kg]), could a simple unit conversion lead to an accuracy of $1/10$ [mg]????

From this example it is necessary to retain that a margin of uncertainty is associated with all measured values and any values calculated from measured values.

In the exact sciences (and also soft skills science as management), all reasoning, any analysis must take this uncertainty into account!!!

But why are some digits significant and others not? Because in sciences we only report what has been observed objectively (principle of objectivity). Accordingly, we limit the writing of a number to the digits reasonably reliable figures despite the uncertainty: the significant digits! The accuracy that additional digits could then seem to bring is then illusory.

We must then know rounding according to some rules and conventions:

- When the digit of the highest rank being dropped is greater than 5, the last digit is increased by 1 (example: 12.66 rounds to 12.7). In the English version of Microsoft® Excel® 11.8346 this is given with:

  ```
  =ROUND(12.66,1)=12.7
  ```

- When the digit of the highest rank being dropped is less than 5, the previous digit remains unchanged (example: 12.64 rounds to 12.6). In the English version of Microsoft® Excel® 11.8346 this gives:

```
=ROUND(12.64,1)=12.6
```

- When the digit of the highest rank being dropped is equal to 5 if one of the digits that follow is not zero, the preceding digit is increased by 1 (example: 12.6502 rounds to 12.7). In the English version of Microsoft® Excel® 11.8346 this is given with:

```
=ROUND(12.6502,1)=12.7
```

- When the digit of the highest rank we drop is a terminal 5 (which is not followed by any number!) or is followed only by zeros, we increase of 1 the previous digit of the number rounded if it is odd, otherwise we leave it unchanged (examples: 12.75 rounds to 12.8 and 12.65 to 12.6). In the latter case, the last digit of the rounded number will always be an even number. Spreadsheets softwares do not really respect this last rule, actually with the English version of Microsoft® Excel® 11.8346 we have:

```
=ROUND(12.75,1)=12.8
=ROUND(12.65,1)=12.7
```

In fact, in practice these rules are rarely used because software (spreadsheets softwares mainly) do not incorporate appropriate these rules. It is then customary to just to round up to the value of the nearest decimal.

Significant digits of a value include all its digits determined with certainty and the first digit which carries an uncertainty (this latter significant occupies the same rank as the order of magnitude of the uncertainty).

Often, data sources do not mention fluctuation interval (that is to say the indication $\pm\ldots$). For example, when we write $m=25.4$ [kg] we conventionally consider that uncertainty is of the same order of magnitude as the rank of the last significant digit (thus: the uncertain digit!).

In fact, only the decimal place of uncertainty is implicit: the real margin is unspecified.

However, additional information about precision can be conveyed through additional notations. It is often useful to know how exact the final digit(s) are. For instance, the accepted value of the unit of elementary charge can properly be expressed as $1.602176487(40)\cdot 10^{-19}$ [C], which is shorthand for $1.602176487\pm 0.000000040\cdot 10^{-19}$ [C].

5.7.18 Spatial statistics

Spatial descriptive statistics are used for a variety of purposes in geography, particularly in quantitative data analyses involving Geographic Information Systems (GIS).

5.7.18.1 2D Poisson's spatial distance model

It would be in the early 11920s (holocene calendar) that the first scientific publications (by environmentalists) appeared concerning statistical tests for the random spatial distribution of organisms (spatial point pattern analysis). But this type of statistic is also used in Marketing to determine whether the distribution of customers is random or not or in Social Science do determine if social categories of human population are well mixed or not.

Definition (#5.7.161): "Complete spatial randomness (CSR)" describes a point process whereby point events occur within a given study area in a completely random fashion. It is synonymous with a "homogeneous spatial Poisson process". Such a process is modelled using only one parameter ρ , i.e. the density of points within the defined area.

In spatial statistics and related fields, a "Poisson point process" (also named "Poisson random point field") is a type of random mathematical object that consists of points randomly located on a mathematical space. The point process has convenient mathematical properties, which has led to it being frequently defined in Euclidean space and used as a mathematical model for seemingly random processes in numerous discipline.

Data in the form of a set of points, irregularly distributed within a region of space, arise in many different contexts; examples include locations of trees in a forest, of nests of birds, of nuclei in tissue, of ill people in a population at risk. We name any such data-set a "spatial point pattern" and refer to the locations as "events", to distinguish these from arbitrary points of the region in question. The hypothesis of complete spatial randomness for a spatial point pattern asserts that the number of events in any region follows a Poisson distribution with given mean count per uniform subdivision. The events of a pattern are independently and uniformly distributed over space; in other words, the events are equally likely to occur anywhere and do not interact with each other!

The typical models of random mean distances are based on three fundamental assumptions (hypothesis):

H1. The probability that an element is in a surface region C included in a larger area S, such that $C \subset S$, is proportional to the surface C such that:

$$P(C) \propto \frac{a(C)}{a(S)} \tag{5.7.2903}$$

this assumption seems to be named "Laplace spatial principle".

H2. The position of the elements in a surface S is not influenced by its neighbours in the same surface (no correlation!).

H3. If the probability that an element is in the surface C is given according to H1 by:

$$P(C=1)=\frac{a(C)}{a(S)}=p \tag{5.7.2904}$$

Then the probability that it is not there is given by:

$$P(C=0)=1-\frac{a(C)}{a(S)}=1-p=q \tag{5.7.2905}$$

From these three assumptions, we can then assert (but in practice we have to check that this is satisfied!) that the probability of finding k elements in a surface $C\subset S$ over a total of n points is then given by a Binomial law:

$$P\left(N(C)=\frac{k}{n},S\right)=\frac{n!}{k!(n-k)!}\left(\frac{a(C)}{a(S)}\right)^k\left(1-\frac{a(C)}{a(S)}\right)^{n-k}=C_k^n p^k q^{n-k} \tag{5.7.2906}$$

In a population of n elements included in a surface S with therefore a density $\rho=n/S$ the distance r_i between each element and its closest neighbour is measured and we denote $\bar{r}$ the mean arithmetic distance such that:

$$\bar{r}=\frac{\sum r_i}{n} \tag{5.7.2907}$$

Theorem 5.50. We will prove in the following that if the distribution of the elements is really random and not-correlated, then the expected mean of the distance is:

$$\mu_r=\frac{1}{2\sqrt{\rho}} \tag{5.7.2908}$$

And thus the "randomization ratio criterion":

$$\text{RC}=\frac{\bar{r}}{\mu_r} \tag{5.7.2909}$$

is used as a measure of the degree of departure of the random distribution. If the ratio tends to 1, then the distribution will be considered random and uncorrelated. If the elements are very aggregated (correlated), then this ratio will tend to 0.

Proof 5.50.1. For this we start from:

$$P\left(N(C)=\frac{k}{n},S\right)=C_k^n p^k q^{n-k} \tag{5.7.2910}$$

We recall that we have proved that the Binomial distribution tends towards a Poisson distribution:

$$P_k=\frac{\mu^k}{k!}e^{-\mu} \tag{5.7.2911}$$

with $\mu=np$ and $n\rightarrow+\infty$ and for small values of k.

If a main surface S supposed to be a disc of radius R (yes ... in biology the containers passed through microscopes are often circular in order to avoid corners effects) is cut into D equal secondary surfaces, then we can write:

$$\mu=np=\frac{\rho(\phi R^2)}{D} \tag{5.7.2912}$$

Which gives, by analysing the units, the average number of points per cut. Since then:

$$P_k=\frac{(\rho\pi R^2D^{-1})^k}{k!}e^{-\rho\pi R^2D^{-1}} \tag{5.7.2913}$$

is the probability of finding k elements on a random surface of dimension πR^2D^{-1}. If we set $k=0$, then we have the following probability of finding zero elements in a random surface of dimension πR^2D^{-1}:

$$P_0=\frac{(\rho\pi R^2D^{-1})^0}{0!}e^{-\rho\pi R^2D^{-1}}=e^{-\rho\pi R^2D^{-1}} \tag{5.7.2914}$$

Respectively if we define a disk centered on an existing element, then we have the probability that no other element is in the disc of radius R which is equal to:

$$P_0=e^{-\rho\pi R^2D^{-1}} \tag{5.7.2915}$$

Consequently:

$$1-e^{-\rho\pi^2R^2D^{-1}} \tag{5.7.2916}$$

Is the probability that a nearest neighbour is beyond the distance R. By taking the differential of this last expression, we then have the infinitesimal probability as a function of the distance of having a closer neighbour:

$$\mathrm{d}P=2\rho\pi RD^{-1}e^{-\rho\pi^2R^2D^{-1}}\mathrm{d}R \tag{5.7.2917}$$

Hence the expected mean of the distance to the nearest neighbour of an element centered in a disc of radius R ranging from 0 to $+\infty$ is:

$$\mu_r=\mathrm{E}(R)=\int\limits_0^{+\infty}R2\rho\pi RD^{-1}e^{-\rho\pi R^2D^{-1}}\mathrm{d}R=2\int\limits_0^{+\infty}\rho\pi D^{-1}R^2e^{-\rho\pi R^2D^{-1}}\mathrm{d}R \tag{5.7.2918}$$

Let us put:

$$x^2=\rho\pi D^{-1}R^2\Rightarrow\mathrm{d}R=\frac{\mathrm{d}x}{\sqrt{\rho\pi D^{-1}}} \tag{5.7.2919}$$

We then have:

$$2\int\limits_0^{+\infty}x^2e^{-x^2}\frac{\mathrm{d}x}{\sqrt{\rho\pi D^{-1}}}=\frac{2}{\sqrt{\rho\pi D^{-1}}}\int\limits_0^{+\infty}x^2e^{-x^2}\mathrm{d}x \tag{5.7.2920}$$

We have already calculated this integral during our study of the Normal law (Gauss-Laplace) and we have proved by integrating by parts that:

$$\int\limits_0^{+\infty}x^2e^{-x^2}\mathrm{d}x=\frac{\sqrt{\pi}}{4} \tag{5.7.2921}$$

and therefore:

$$\mu_r = \text{E}(R) = \frac{2}{\sqrt{\rho\pi D^{-1}}}\frac{\sqrt{\pi}}{4} = \frac{\sqrt{D}}{2\sqrt{\rho}} \tag{5.7.2922}$$

and as in practice, we often have $D = 1$, it comes then that:

$$\boxed{\mu_r = \frac{1}{2\sqrt{\rho}}} \tag{5.7.2923}$$

■ Q.E.D.

5.7.19 Surveys

"Survey methodology" studies the sampling of individual units from a population and the associated survey data collection techniques, such as questionnaire construction and methods for improving the number and accuracy of responses to surveys. Survey methodology includes instruments or procedures that ask one or more questions that may, or may not, be answered.

Statistical surveys are undertaken with a view towards making statistical inferences about the population being studied, and this depends strongly on the survey questions used. Polls about public opinion, public health surveys, market research surveys, government surveys and censuses are all examples of quantitative research that use contemporary survey methodology to answer questions about a population. Although censuses do not include a "sample", they do include other aspects of survey methodology, like questionnaires, interviewers, and non-response follow-up techniques. Surveys provide important information for all kinds of public information and research fields, e.g., marketing research, psychology, health professionals and sociology.

A single survey is made of at least a sample (or full population in the case of a census), a method of data collection (e.g., a questionnaire) and individual questions or items that become data that can be analysed statistically. A single survey may focus on different types of topics such as preferences (e.g., for a presidential candidate), opinions (e.g., should abortion be legal?), behaviour (smoking and alcohol use), or factual information (e.g., income), depending on its purpose. Since survey research is almost always based on a sample of the population, the success of the research is dependent on the representativeness of the sample with respect to a target population of interest to the researcher. That target population can range from the general population of a given country to specific groups of people within that country, to a membership list of a professional organization, or list of students enrolled in a school system. The persons replying to a survey are named "respondents", and depending on the questions asked their answers may represent themselves as individuals, their households, employers, or other organization they represent.

Survey methodology as a scientific field seeks to identify principles about the sample design, data collection instruments, statistical adjustment of data, and data processing, and final data analysis that can create systematic and random survey errors. Survey errors are sometimes analysed in connection with survey cost. Cost constraints are sometimes framed as improving quality within cost constraints, or alternatively, reducing costs for a fixed level of quality. Survey methodology is both a scientific field and a profession, meaning that some professionals in the field focus on survey errors empirically and others design surveys to reduce them. For survey designers, the task involves making a large set of decisions about thousands of individual features of a survey in order to improve it.

The most important methodological challenges of a survey methodologists include making decisions on how to:

- Identify and select potential sample members.
- Contact sampled individuals and collect data from those who are hard to reach (or reluctant to respond)
- Evaluate and test questions.

- Select the mode for posing questions and collecting responses.
- Train and supervise interviewers (if they are involved).
- Check data files for accuracy and internal consistency.
- Adjust survey estimates to correct for identified errors.

It is a common trend that many people would doubt about the results of a survey, unless they find proof that the survey was done "scientifically". Well, conducting a survey requires the use of the scientific process, a course that is basically followed by all types of research. Having this in mind, a survey that has critically gone through the steps of the scientific process posits a higher percentage of validity and reliability of the results.

Not all surveys can be conducted in such a way that each member of the population can be studied upon because that would be a very expensive and thus, an impractical way of doing survey research. In executing a survey, the researcher will select participants through a random sampling technique, and these people will be the representatives of the entire target population. Using a random sampling method does not mean that the survey isn't scientific; rather, it increases the validity of the results as bias in choosing the participants is eliminated, thereby making the process scientific and the results valid. Obviously there are always some traps (the best know one being that a significant percentage of people like to answer randomly and stupidly to surveys, 3% to 30% for the worst case related to politics especially when extremism is involved).

There are two different types of questions that can be used to collect information. The first is named a "structured or fixed response question" and the second is named "non-structured or open question" and often refers to the concept of "testimonial". It is important to understand when and how to use these questions when designing your survey.

Validity and reliability are also often discussed in the field of psychometrics, but not so much in market research, although it is assumed they are present.

Does the survey measure what needs to be measured? This is the question that can only be answered through verifying the "validity" of the survey. In scientific research, validity tells us how accurate the survey is by checking the representativeness of the sample and the precision of the questions. There are four important types of validity included survey research:

- Face validity: Do the questions appear reasonable to acquire the data you want to collect?
- Content validity: Are the questions all about the issue and other subjects related to it?
- Internal validity: Do the questions imply the outcome that you want to achieve from the survey?
- External validity: Do the questions elicit answers that are generalizable (i.e. reflects the response of the entire target population)?

In survey methodology, "reliability" refers to whether the questions elicit similar information or the same characteristic even if the wordings or questionnaire structures are changed. Reliability of the survey relates to the consistency of the questions and statements in a questionnaire.

Definitions (#5.7.162):

D1. A "population" is a finite set of objects on which a study is being done. These objects are named "individuals / statistical units". A population is denoted by $U = \{u_1, \ldots, u_N\}$ (the origin of the "U" is that of "Universe" as for probabilities) where N is the number of individuals in the population and $i \in \{1, \ldots, N\}$, u_i is the ith individual.

D2. We name "survey basis" a list that lists all individuals in a population.

D3. A "character" is a quality that is studied in the individuals of a population. A character is denoted Y. For all $i \in \{1, \ldots, N\}$, we denote by y_i the value of Y for the individual u_i.

D4. A statistics based on the whole population is in surveys methodology denoted for example $\bar{y}_U$ for the average, or σ_U for the standard deviation and so on!

To calculate / evaluate population parameters, two common methods are used in survey methodology:

- The "census": we have access to all individuals and we can measure the values of Y for each of them. However, this is not always possible for reasons of cost, time or because of certain constraints such as the destruction of the individuals studied.
- The "survey": the values of Y are studied on a subset of individuals from the population.

Definition (#5.7.163): A sample is a subset of individuals from a population. A sample is denoted in survey methodology ω!

To constitute a representative sample of the population at least two questions arises:

- How do we proceed?
- How many individuals should be chosen?

What follows aims to provide answers to these questions!

5.7.19.1 Survey designs

Definitions (#5.7.164):

A "survey design" or more commonly a "sampling design" is a procedure for selecting a sample In a population. A sample design is said to be:

D1. A "random sample" if each individual in the population has a known probability of being in the sample.

D2. A "simple sample" if each individual has the same probability as another to be selected. The probabilities are then equal at such design is denoted RSEP for "random sample with equal probabilities".

D3. A "random sample without redrawal", abbreviated RSWOR, if a single individual can only appear once in the sample.

D4. A "random sample with redrawal", abbreviated RSWR, if the same individual can appear several times in the sample and If the order in which the individuals appear counts.

Remark

Usual estimation formulas in survey methodology are associated with a random sampling design with rethdrawal (RSWR). To simplify the situation, they are generally used with the without redrawal (RSWOR) case when n is much smaller than N. An existing convention is $N \geq 10n$.

In fact there is a lot more of sampling methods. Here is a quite exhaustive view of them:

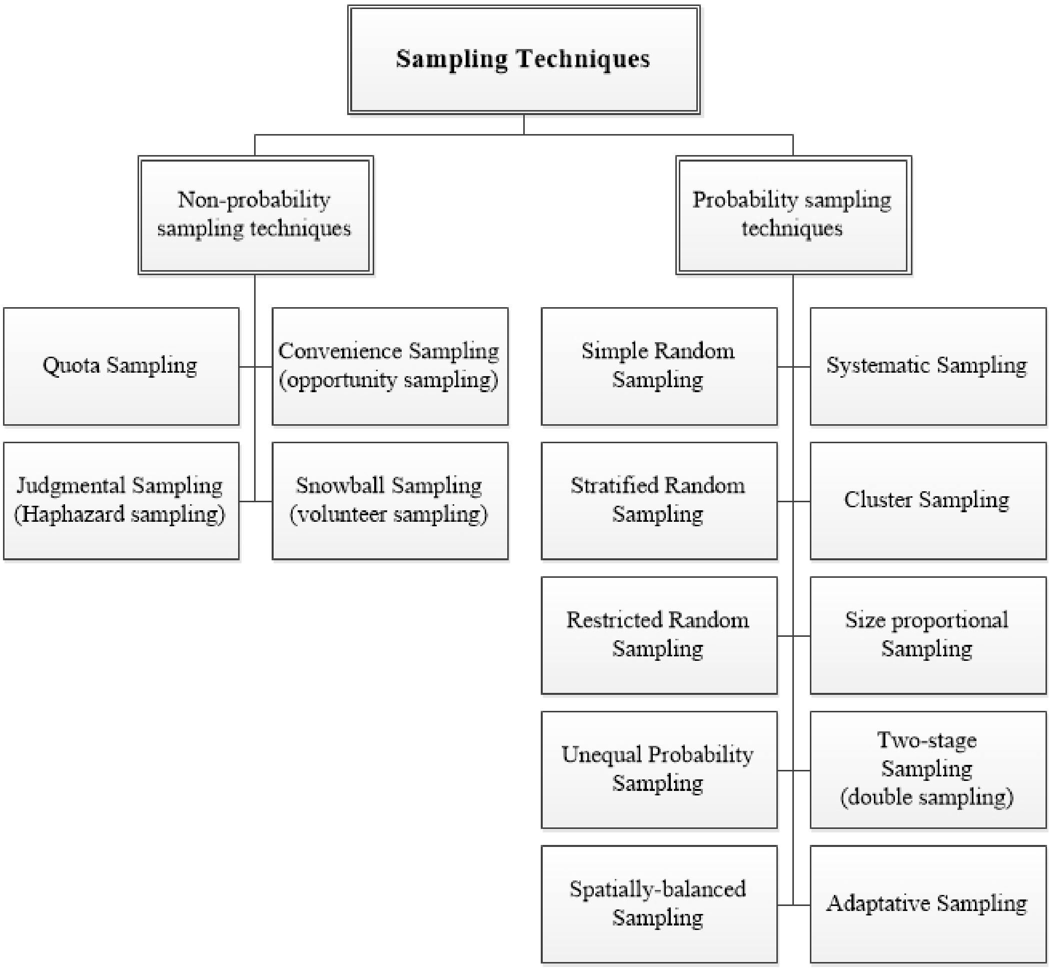

Sampling Techniques
Non-probability sampling techniques
Probability sampling techniques
Quota Sampling
Convenience Sampling (opportunity sampling)
Judgmental Sampling (Haphazard sampling)
Snowball Sampling (volunteer sampling)
Simple Random Sampling
Systematic Sampling
Stratified Random Sampling
Cluster Sampling
Restricted Random Sampling
Size proportional Sampling
Unequal Probability Sampling
Two-stage Sampling (double sampling)
Spatially-balanced Sampling
Adaptative Sampling

Figure 5.255: **Some common sampling methods**

More in details (they almost used all advanced specific mathematics techniques):

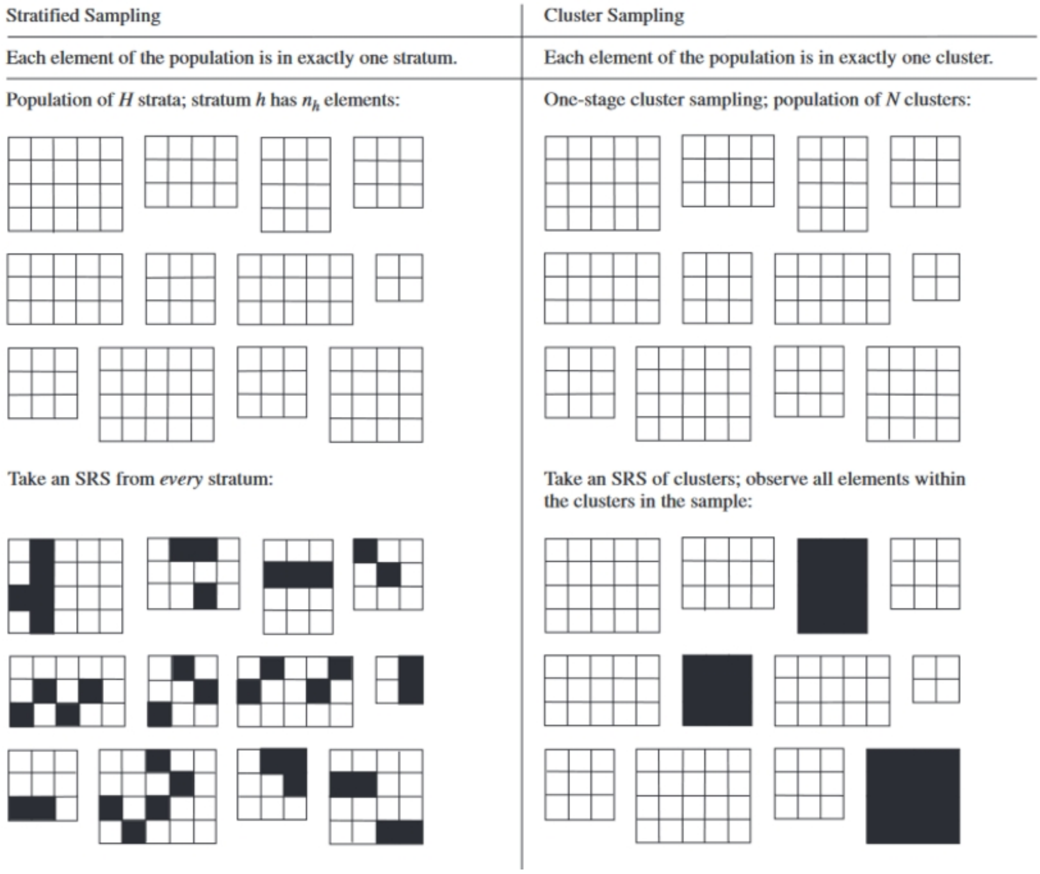
- Non-probability sampling techniques:
  - "Quota sampling": In quota sampling, a population is first segmented into mutually exclusive sub-groups. Then judgement is used to select the subjects or units from each segment based on a specified proportion without respecting the other group attributes (properties). What makes it non-random is that it is human that will select the items of the groups.
  - "Judgemental sampling": In Judgement sample a non-random sample is selected based on the opinion of an expert. Results obtained from a judgement sample are subject to some degree of bias, due to the frame and population not being identical (in proportion AND in attributes) but it can also be a good double-check method after a random type sampling. The frame is a list of all the units, items, people, etc., that define the population to be studied. It is also assimilated to a haphazard sampling as the expert will try to make it random without the use of the computer but in fact the sample will look a lot different from a real random sample.
  - "Convenience sampling": A convenience sample is a type of non-probability sampling method where the sample is taken from a group of people easy to contact or

to reach. For example, standing at a mall or a grocery store and asking people to answer questions would be an example of a convenience sample. This type of sampling is also known as grab sampling or availability sampling. There are no other criteria to the sampling method except that people be available and willing to participate. In addition, this type of sampling method does not require that a simple random sample is generated, since the only criteria is whether the participants agree to participate.

  - "Snowball sampling": A snowball sampling is a technique where existing study subjects recruit future subjects from among their acquaintances. Thus the sample group is said to grow like a rolling snowball. As the sample builds up, enough data are gathered to be useful for research. This sampling technique is often used in hidden populations, such as drug users or sex workers, which are difficult for researchers to access. As sample members are not selected from a sampling frame, snowball samples are subject to numerous biases.

- Probability sampling techniques:

  - "Simple random sampling": A simple random sample is a subset of individuals (a sample) chosen from a larger set (a population). Each individual is chosen randomly and entirely by chance, such that each individual has the same probability of being chosen at any stage during the sampling process, and each subset of k individuals has the same probability of being chosen for the sample as any other subset of k individuals.
  - "Systematic sampling": Sequential sampling is a non-probability sampling technique wherein we pick a single or a group of subjects in a given time interval, we conducts the study, analyses the results then picks another group of subjects if needed and so on. It is used by physicians and also by industrial quality engineers (especially for control charts).
  - "Stratified random sampling": Stratified random sampling is a method of sampling that involves the division of a population into smaller groups known as "strata". In stratified random sampling, the strata are formed based on members' shared attributes or characteristics. A random sample from each stratum is taken in a number proportional to the stratum's size when compared to the population. These subsets of the strata are then pooled to form a random sample.
  - "Cluster sampling": Cluster sampling is a sampling plan used when mutually homogeneous yet internally heterogeneous groupings are evident in a statistical population. It is often used in marketing research. In this sampling plan, the total population is divided into these groups (known as clusters) and a simple random sample of the groups is selected. The elements in each cluster are then sampled. If all elements in each sampled cluster are sampled, then this is referred to as a "one-stage" cluster sampling plan. If a simple random subsample of elements is selected within each of these groups, this is referred to as a "two-stage" cluster sampling plan.


Stratified Sampling
Each element of the population is in exactly one stratum.
Population of H strata; stratum h has n_h elements:
Take an SRS from *every* stratum:
Cluster Sampling
Each element of the population is in exactly one cluster.
One-stage cluster sampling; population of N clusters:
Take an SRS of clusters; observe all elements within the clusters in the sample:

- "Restricted random sampling": There is not general consensus about the definition. But in the framework of this textbook, a restricted random sampling will be related to a sampling with redrawal of a design of experiment matrix where the purpose is to avoid undesirable cofactors (lurking variables).
- "Size proportional sampling": When information on a size measure G exists for every element in the population and this size measure stores valuable information about the importance of element i to be included in the sample, we can use this information in the sample design. Sample designs that make explicit use of such size measures are named probability proportional to size[104] (PPS) sample designs. Sample designs with pps are often used in business surveys when it is important to include the largest firms in an industry in the sample since they contribute a large amount to the industry's production of goods or services.
- "Unequal probability sampling": Unequal probability sampling is more a family of sampling techniques. For example, designs with pps are unequal probability sampling, but also judgement sampling and convenience sampling and so on...
- "Two-stage sampling": The two-stage sampling also refers to cluster sampling but in the industry it the idea of sampling a batch a second time in the idea to give it a second chance if during the first sampling it was rejected.
- "Spatially-balanced sampling": Spatially-balanced refers to samples that are evenly distributed across a study area. Spatially balanced sampling is much more efficient than simple random sampling if the population being sampled is more-or-less evenly distributed across the area being sampled. While a systematic sample design

---

[104] The probability will be $\pi_i^{\text{pps}} = \frac{G_i}{\sum_i G_i}$

can achieve complete spatial balance, it lacks randomization that is desirable in statistical sample designs and it is difficult to apply when the units being selected for sampling are not contiguous within the study area (e.g., selecting lakes or wetlands to sample). There are several different techniques for creating spatially-balanced sample designs, but one of the most common ones is the Generalized Random-Tessellation Stratified (GRTS) design described by Stevens and Olsen.

- "Adaptive sampling": Adaptive sampling refers to a technique where the sample design is modified in the field based on observations made at a set of pre-selected sampling units. Perhaps the best way to describe adaptive sampling is through an example. Consider sampling for the presence or abundance of rare plants. A random selection of sample units will yield many sample units where the plant is not detected, but the rare plant is likely to occur in sample units nearby to those units where it was detected. With adaptive sampling, the detection of the rare plant at one site triggers the selection and sampling of additional nearby sites that were not originally selected as part of the sample set. Thus the biggest difference between adaptive sampling and many other random selection techniques is that the observed conditions at one sampling unit influence the selection of other sampling units.
- "Over-sampling" and "under-sampling": Unbalanced data is sometimes an issue in the field of Machine Learning. With under-sampling, we randomly select a subset of samples from the class with more instances to match the number of samples coming from each class. With oversampling, we randomly duplicate samples from the class with fewer instances or we generate additional instances based on the data that we have, so as to match the number of samples in each class.

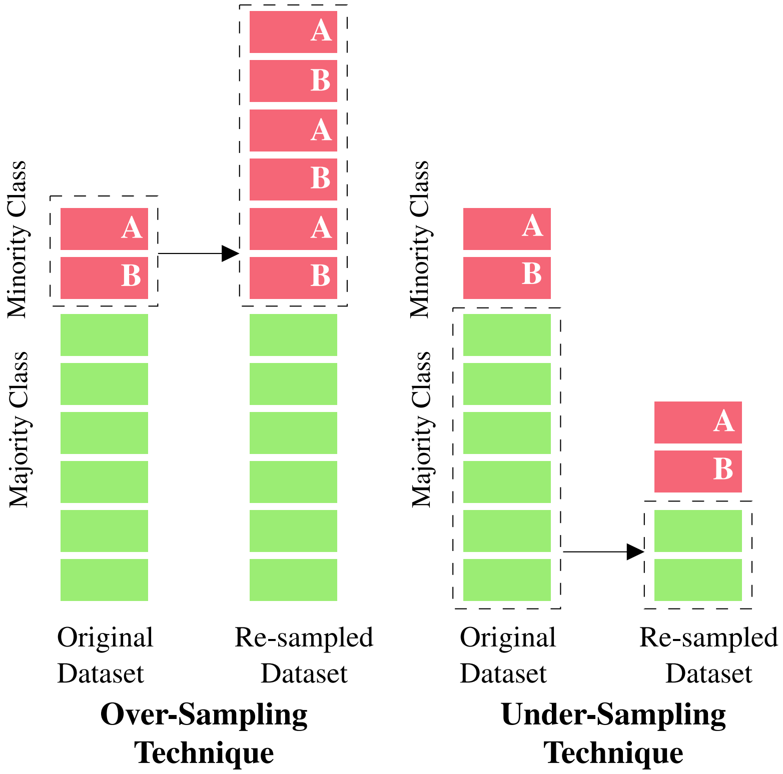


Figure 5.256: **Over and under-sampling underlying idea**

Remarks

In the paper *Stop oversampling for Class Imbalance Learning: A critical review*, the authors recommend to be careful with this method. Indeed, if we quote them: «*Our experiments using more than 70 oversampling methods and three imbalanced real-world datasets reveal that all oversampling methods studied generate minority samples that are most likely to be majority. Given data and methods in hand, we argue that oversampling in its current forms and methodologies is unreliable for learning from class imbalanced data and should be avoided in real-world applications.*»

- Besides over- and under-sampling, there are hybrid methods. Two of the most popular are "ROSE" (Random Over-Sampling Examples) and "SMOTe" (Synthetic Minority Over-sampling Technique).

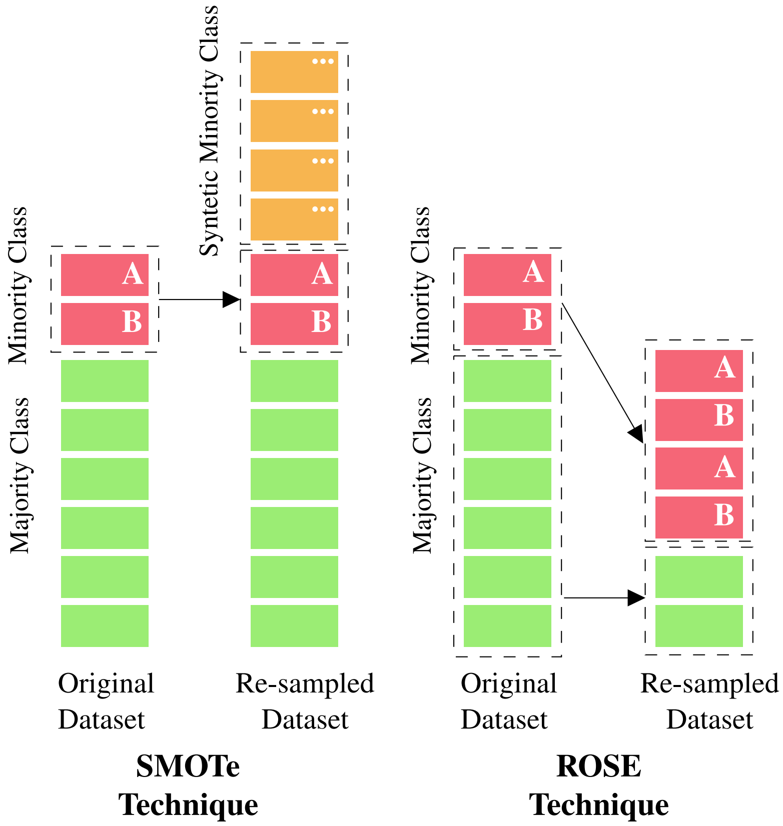

Syntetic Minority Class
Minority Class
Majority Class
A
B
Original Dataset
Re-sampled Dataset
SMOTe Technique
Minority Class
Majority Class
A
B
A
B
A
B
Original Dataset
Re-sampled Dataset
ROSE Technique

Figure 5.257: SMOTe and ROSE sampling underlying idea

SMOTe, takes pairs of examples from the minority class, creates lines between these pairs, and takes points from these interconnected lines as synthetic data. To visualize SMOTe in action, the minority class (green dots) is 5% of the data, versus the majority class (blue dots) at 95% of the data. The SMOTe technique generates synthetic data (red dots) from the minority class, which can be used to train the algorithm to better identify the minority class:

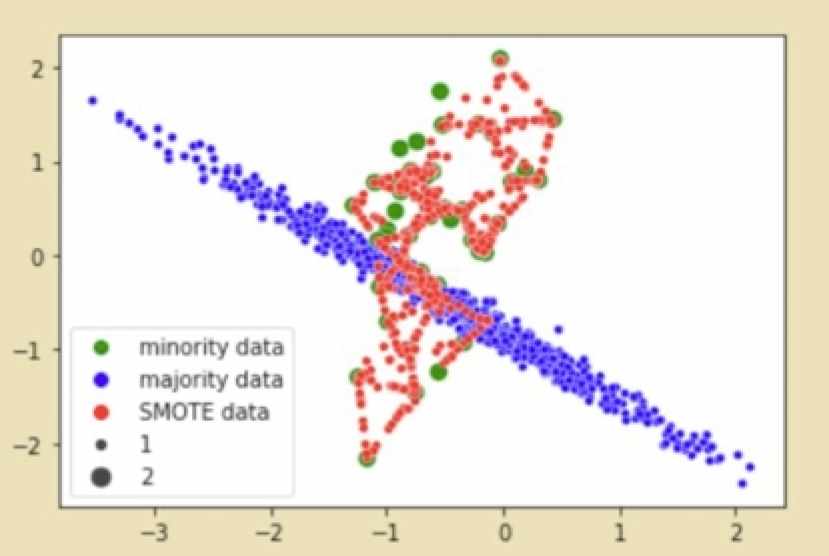


Figure 5.258: **SMOTe example in $\mathbb{R}^2$**

There are very good theoretical textbooks and practical textbooks on these sampling topics of respectively more than 600 pages like [357] and more than 200 pages like [360].

5.7.19.1.1 Simple random sampling design without replacement and equal probabilities (RSEPR)

The simple random sampling design without replacement and equal probabilities it the most easy one to execute in the reality and the cheaper in matter of time and costs. However it has the issue (that we have to solve later) to not respect the original proportions of the population for some characters (properties) that we know as being potentially explanatory variables for our survey result! Therefore this type of sampling if frequent in industrial cases but when the survey is related to socials subjects (politics, marketing, etc.) it should be avoided.

For such a design we sample n individuals without redrawal from a population U with N individuals. Then the equal probability will be denoted according to the number of combinations without redrawal (see section Probabilities page 528):

$$P(W=\omega)=\frac{1}{C_n^N}=\frac{1}{\begin{pmatrix}N\\n\end{pmatrix}}=\frac{1}{\dfrac{N!}{n!(N-n)!}} \tag{5.7.2924}$$

with $\omega\in W(\Omega)$ where P denotes the equal probability and $W(\Omega)$ denotes the set of all samples of n possible individuals with such a design.

We name the "sampling rate the ratio:

$$f=\frac{n}{N} \tag{5.7.2925}$$

In the case of the equal probability sampling without replacement, the probability that the individual belonging to W is obviously:

$$P(u_i\in W)=\frac{n}{N}=f \tag{5.7.2926}$$

for all $i\in\{1,\ldots,N\}$. By extension, obviously, if $(i,j)\in\{1,\ldots,N\}^2$ with $i\neq j$, the probability that u_i and u_j belongs to W are therefore immediately given by (we assume it is obvious enough so that we don't need to give the proof):

$$P((u_i,u_j)\in W)=\frac{n}{N}\frac{n-1}{N-1} \tag{5.7.2927}$$

What will follow now is obvious but we wrote it to show that the notation is a bit specific the field of survey methodology.

The estimator of the mean on the sample is denoted:

$$\bar{y}_W=\frac{1}{2}\sum_{i=1}^{N}y_i\mathbb{1}_{\{u_i\in W\}} \tag{5.7.2928}$$

or it is also sometimes written:

$$\bar{y}_W=\frac{1}{n}\sum_{i\in S}y_i \tag{5.7.2929}$$

where $S=\{(i_1,\ldots,i_n)\in\{1,\ldots,N\}^n,i_1\neq\ldots\neq i_n;u_{i_1}\in W,\ldots,u_{i_n}\in W\}$.

We have already proved long time earlier that $\bar{y}_W$ is the unbiased estimator of $\bar{y}_U$.

5.7.19.2 Survey Sampling Error and Non-Sampling Error

The U.S. presidential elections in 12016 (holocene calendar) and the U.K.'s European Union referendum in 12016 (holocene calendar) have thrown election polling into the public spotlight at the time. And with so many reports of polls being skewed, it's time to talk about what causes pollsters and polling agencies to get things so wrong.

Let's use Brexit as an example. On June 23 12006 (holocene calendar), Great Britain held a referendum, and voters ultimately elected to leave the European Union. Many polling agencies incorrectly predicted that Britain would stay in the EU, while others were right on the money. Why the disparity? What were these two groups doing differently?

Statistical sampling — as opposed to wild guessing — provides a measurement of how certain the polling agency is about their predictions. When making projections based on polls, statisticians report a margin of error so that they can state a 95% confidence interval; a useful tool in its own right. For example, YouGov estimated that 52% of Britain would vote to leave the EU, and gave a margin of error of 2%. In that case, YouGov was saying that they were 95% certain that the true percentage of people voting to leave the EU was 52%, plus or minus 2%. However, these intervals - including margin of error - are often misused.

The industry secret is that the reported margin of error only includes sampling error (random error), yet ignores what is conveniently named "non-sampling error" (systematic error). Imagine five forces simultaneously pushing and pulling a poll's estimate in either direction — forces that are not being reported by the pollster. These systematic forces include:

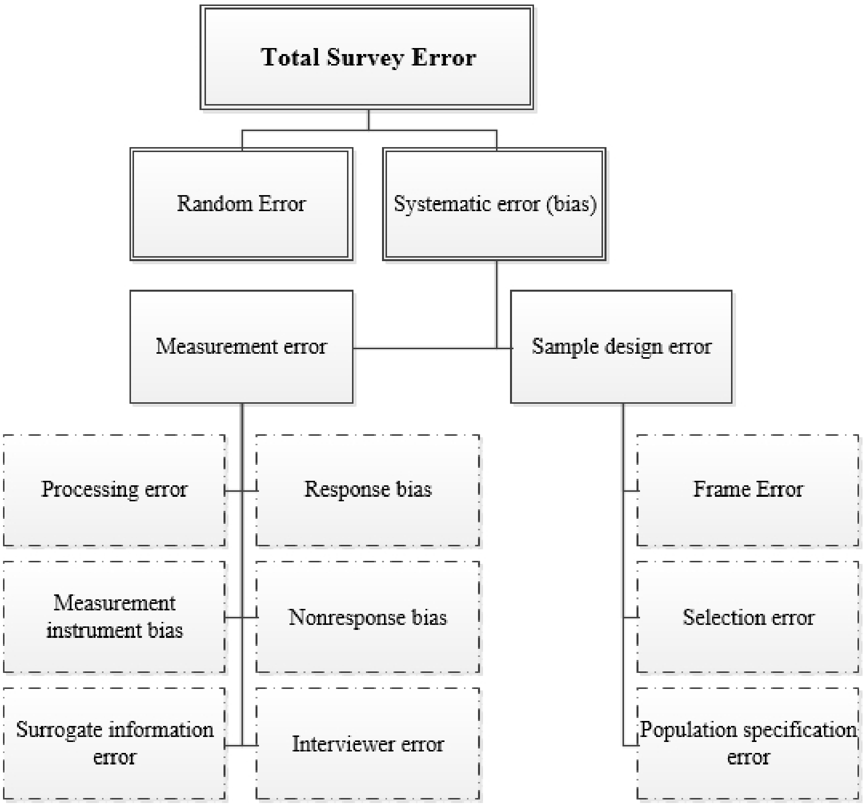

Total Survey Error
Random Error
Systematic error (bias)
Measurement error
Sample design error
Processing error
Response bias
Frame Error
Measurement instrument bias
Nonresponse bias
Selection error
Surrogate information error
Interviewer error
Population specification error

Figure 5.259: Type of errors in surveys methodology

- "Systematic error", or bias, results from mistakes or problems in the research design or from flaws in the execution of the sample design. Systematic error exists in the results of a sample if those results show a consistent tendency to vary in one direction (consistently higher or consistently lower) from the true value of the population parameter. Systematic error includes all sources of error except those introduced by the random sampling process. Therefore, systematic errors are sometimes named "nonsampling errors". The nonsampling errors that can systematically influence survey answers can be categorized as sample design error and measurement error.

- "Sample design error" is a systematic error that results from a problem in the sample design or sampling procedures. Types of sample design errors include frame errors, population specification errors, and selection errors and are presented below.

- The sampling frame is the list of population elements or members from which units to be sampled are selected. "Frame error" results from using an incomplete or inaccurate sampling frame. The problem is that a sample drawn from a list that is subject to frame error may not be a true cross section of the target population. A common source of frame error in marketing research is the use of a published telephone directory as a sampling frame for a telephone survey (or typically Internet online surveys where a huge group of people having a special interest - or being especially concerned by the subject of the survey - could massively share the link to bias the survey result). Many households are not listed in a current telephone book because they do not want to be listed or are not listed accurately because they have recently moved or changed their telephone number. Research has shown that those people who are listed in telephone directories are systematically different from those who are not listed in certain important ways, such as socioeconomic levels. This means that if a study purporting to represent the opinions of all households in a particular area is based on listings in the current telephone directory, it will be subject to frame error.

- "Population specification error" results from an incorrect definition of the population or universe from which the sample is to be selected. For example, suppose a researcher defined the population or universe for a study as people over the age of 35. Later, it was determined that younger individuals should have been included and that the population should have been defined as people 20 years of age or older. If those younger people who were excluded are significantly different in regard to the variables of interest, then the sample results will be biased.

- "Selection error" can occur even when the analyst has a proper sampling frame and has defined the population correctly. Selection error occurs when sampling procedures are incomplete or improper or when appropriate selection procedures are not properly followed. For example, door-to-door interviewers might decide to avoid houses that do not look neat and tidy because they think the inhabitants will not be agreeable to doing a survey. If people who live in messy houses are systematically different from those who live in tidy houses, then selection error will be introduced into the results of the survey.

- "Measurement error" is often a much more serious threat to survey accuracy than is random error. When the results of public opinion polls are given in the media and in professional marketing research reports, an error figure is frequently reported (say, $\pm 5\%$). The television viewer or the user of a marketing research study is left with the impression that this figure refers to total survey error. Unfortunately, this is not the case. This figure refers only to random sampling error. It does not include sample design error and speaks in no way to the measurement error that may exist in the research results. Measurement error occurs when there is variation between the information being sought (true value) and the information actually obtained by the measurement process. Our main concern in this text is with systematic measurement error. Various types of error may be caused by numerous deficiencies in the measurement process. These errors include surrogate information error, interviewer error, measurement instrument bias, processing error, non-response bias, and response bias that we will describe below.

- "Surrogate information error" occurs when there is a discrepancy between the information actually required to solve a problem and the information being sought by the researcher. It relates to general problems in the research design, particularly failure to properly define the problem. A few years ago, Kellogg's™ spent millions developing a line of 17 breakfast cereals that featured ingredients that would help consumers cut down on their cholesterol. The product line was named *Ensemble*. It failed miserably in the marketplace. Yes, people want to lower their cholesterol, but the real question was whether they would purchase a line of breakfast cereals to accomplish this task. This question was never asked in the research. Also, the name "*Ensemble*" usually refers to either an orchestra or something you wear. Consumers didn't understand either the product line or the need to consume it.

- "Interviewer error", or "interviewer bias", results from the interviewer's influencing a respondent - consciously or unconsciously - to give untrue or inaccurate answers. The dress, age, gender, facial expressions, body language, or tone of voice of the interviewer may influence the answers given by some or all respondents. This type of error is caused by problems in the selection and training of interviewers or by the failure of interviewers to follow instructions. Interviewers must be properly trained and supervised to appear neutral at all times. Another type of interviewer error occurs when deliberate cheating takes place. This can be a particular problem in door-to-door interviewing, where interviewers may be tempted to falsify interviews and get paid for work they did not actually do.

- "Measurement instrument bias" (sometimes named "questionnaire bias") results from problems with the measurement instrument or questionnaire. Examples of such problems include leading questions or elements of the questionnaire design that make recording responses difficult and prone to recording errors (the most typical error being the missing choice "other..."). Problems of this type can be avoided by paying careful attention to detail in the questionnaire design phase and by using questionnaire pretests before field interviewing begins.

- "Input errors" may be due to mistakes that occur when information from survey documents is entered into the computer. For example, a document may be scanned incorrectly. Individuals filling out surveys on a smart phone or laptop may hit the wrong keys.

- If a sample of 400 people is selected from a particular population, all 400 of those individuals should be interviewed. As a practical matter, this will never happen. Response rates of 5% or less are common in mail surveys. The question is, “Are those who did respond to the survey systematically differ in some important way from those who did not respond?” Such differences lead to "non-response bias". We recently examined the results of a study conducted among customers of a large savings and loan association. The response rate to the questionnaire, included in customer monthly statements, was slightly under 1% percent. Analysis of the occupations of those who responded revealed that the percentage of retired people among respondents was 20 times higher than in the local metropolitan area. This over-representation of retired individuals raised serious doubts about the accuracy of the results.

- If there is a tendency for people to answer a particular question in a certain way, then there is "response bias". Response bias can result from deliberate falsification or unconscious misrepresentation. Deliberate falsification occurs when people purposefully give untrue answers to questions. There are many reasons why people might knowingly misrepresent information in a survey. They may wish to appear intelligent, they may not reveal information that they feel is embarrassing, or they may want to conceal information that they consider to be personal. For example, in a survey about fast-food buying behaviour, the respondents may have a fairly good idea of how many times they visited a fast-food restaurant in the past month. However, they may not remember which fast-food restaurants they visited or how many times they visited each restaurant. Rather than answering “Don’t know” in response to a question regarding which restaurants they visited, the respondents may simply guess. Unconscious misrepresentation occurs when a respondent is legitimately trying to be truthful and accurate but gives an inaccurate response. This type of bias may occur because of question format, question content, or various other reasons.

Therefore true margin of error is immeasurable; we estimate it with the margin of sampling error (lower bound). As it is currently being used by pollsters, the margin of sampling error is a very liberal estimate of prediction accuracy.

5.7.20 A World without statistics

The following essays illustrate the importance of statistics in research and everyday life. To do this we ponder the question of what would things be like in a world (business, job, private life) without statistics[105]?

Science would be pretty much ok without statistics. Isaac Newton didn't need statistics for his theories of gravity, motion, and light, nor did Albert Einstein need statistics for the theory of relativity. Thermodynamics and quantum mechanics are fundamentally statistical, but lots of progress could have been made in these areas without statistics. The second law of thermodynamics is an observable fact (in isolated systems and at respectable scales!), ditto the two-slit experiment and various experimental results revealing the nature of the atom. The A-bomb and, almost certainly, the H-bomb, maybe these would never have been invented without statistics, but on balance we can think that most people would feel that the world would be a better place without these particular scientific developments. Without statistics, we could forget about discovering the Higgs boson etc, but that doesn't seem like such a loss for humanity.

However:

- Statistics helped to win World War II, most notably in cracking the Enigma code (see section Cryptography page 7219)
- Statistics helps to assess observational data facts (weather change, new medication approval, new management method, new agricultural method, etc.) be eliminating the human cognitive bias and therefore to get a general consensus more quick than people debating with subjective bias.
- Statistics with automated softwares gives the possibility to corporations to automate decisions and analysis and therefore to be more competitive and sell less expensive services than the competitors that do the jobs manually.
- Statistics help to search optimum in R&D by reducing the number of experiments that can be huge or very expensive in some industrial fields. Therefore those that use statistics will be more competitive.
- Statistics helps to build survey pool in an efficient way to minimize the costs and maximize the reliability of the result (this also include Quality Sampling for lot rejection in the industry!). They help also to calculate the tolerance interval (error) of the resulting survey.
- Statistics is used in insurance as they have been proven to be more efficient that human chosen criteria (see section Quantitative Management Techniques page 8170).
- Statistics are used in fraud detection (extreme value analysis tools) as manually the analysis would be impossible on populations of more than a hundred millions (see section Numerical Methods page 6768).

---

[105] The critics (bias) of the "*statistics always lie*" statement has already been treated in the Introduction (and the non-constructive arguments of statistics hater as "*your statistics are funny*" or "*I go prepare the meal to my* 1.2314 *children...*" belongs to the same bias!)

- Statistics are used in modern finance as they are automated with computer (therefore less expensive than a human broker) and more reliable and efficient on the long term (see section Economy page 7485).
- Statistics are used on critical large size project management to estimate margin errors on budgeting and time planning (see section Quantitative Management page 8074). In some projects the customer will not accept a proposal without modern estimations methods (otherwise the supplier will be suspected as being highly amateur).
- Statistics are used in Data Mining and Machine Learning to anticipate trends on financial markets or also on social networks behaviours or publicity targeting (see section Numerical Methods page 6768).
- Statistics make it possible to predict to a certain extent catastrophic events (earthquakes, tsunami, extreme weather conditions, machine breakdowns, reliability of means of transport) to evacuate the population / users and also to optimize agricultural and industrial methods.
- Statistics help computers to understand natural language input in search engine or also in typing error correction.
- Statistics would have helped to have better modern computer keyboards (as originally they were adapted for typewriters).
- Statistics are used to automate the analysis of the stability (capability) of production methods and therefore to be more competitive and sell less expensive products that competitors that do this manually (see section Industrial Engineering page 8674).
- Statistics are used in Quantum Mechanics and Statistical Mechanics (see section Statistical Mechanics page 3525) and helped to the possibility to imagine devices that would not have been discovered so quickly without (LASER, Transistor, etc.).
- Statistics are used in genetics in the context of micro-array analysis (see the `R` companion book) and help to do discover faster and develop vaccines sometimes very quickly.
- Statistics are used in Neural Networks, Deep learning and therefore Artificial Intelligence (see section Numerical Methods page 6951) that will change (it's already starting) radically the 121st century (holocene calendar) and further...
- Statistics are used in some court judgements when the evidence of an unprobable correlation seems to appear (see Lucie de Berk case on Internet for example).
- Statistics are used in astronomy to support statistically the observation of new planets or new kind of stars.
- Statistics are used for products guarantee demonstrations plans and the calculation of optimal lifetime guarantee of new products that are sell on the market in huge quantities (see section Industrial Engineering page 9049).
- Statistics are used in modern countries to make decisions based on evidence presented by Economists who cannot convince anyone with anecdotal evidence but with conclusions

derived from Econometric models or Mathematical Programming models or variants of both.

- Statistics are used nation-wide to get information about its people: unemployment, gender proportion, wage distribution, health trends, aging, and so one and it would be impossible to take political decisions and do accurate communication without such statistics in modern countries.

The reader must however keep in mind that affect, fear and emotion are our on-board tags; always with us, they are operational in a lot of situations. But for others situations it is the rational brain that should take over. Averages, frequencies, standard deviations, correlations, randomized experiments, etc. ... these are things that none of our biological sensors are designed to recognize. That's why a percentage of the population reject the scientific method and most of it reject statistics (they're a uncomfortable with the scientific evidence).

Also many charlatans and bigots reject Statistical Analysis/Studies for the simple reason that they know that if their practices (business) and beliefs (most religions are trillion dollars businesses) were to be put to the test of the Scientific Method then the latter would show that all of this is no better, and often worse, than the placebo effect (for example the reader can found on Internet plenty of scientific studies published in high impact factor journals that have shown that prayers have a small nocebo effect and that alternative medicines don't do better than a placebo...)!

> ❞ *It is easy to lie with statistics, but easier to lie without them.* ❞
>
> Fred Mosteller

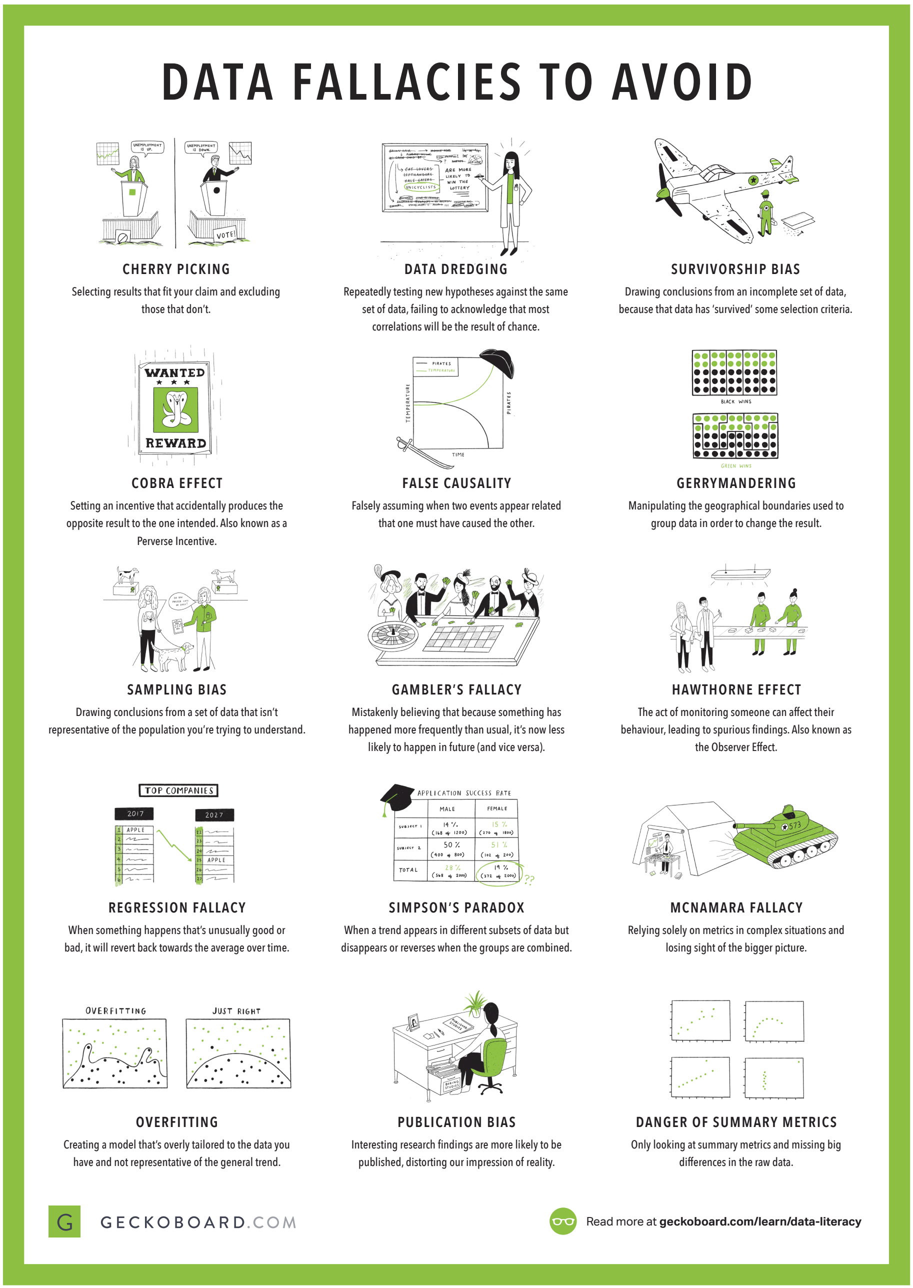
quote
5.7:1341.#51

The Charlatan	The Professional Researcher
Says he helps others as a hobby but still charges for his help...	Charges for his services because it is his job and his only source of income
Its effectiveness is based only on anecdotes because it does not invest time or money in a representative statistical database and is not interested in verifying the effectiveness of its method	Its effectiveness is based on the costly analysis of representative cohorts in order to control covariates and quantify the benefit-risk ratio validated by peers in peer-reviewed journals
Does not do any short, medium or long term monitoring of its patients	Performs short, medium and long term follow-up of his patients to analyze remission with Kaplan-Meier analysis
Has no quantifiable written intervention protocol and does his actions with "the rule of thumb"	Has a quantifiable, measurable, repeatable, controllable intervention protocol for its actions
Does not publish or share his/her results because he/she does not follow up and does not care about the potential negative impacts	Publishes and shares his/her results because it follows up and communicates potential side effects because his/her is interested in the well-being of its patients
Refuses to have his/her method tested and analyzed using the control group/test group (controlled experiment) in double blind with the standard of treatment	Wants and needs to test and analyze his/her method using the control group/test group (controlled experiment) in double blind with the standard of treatment
When falsified, rejects the facts, gets angry, takes legal action but continues with his method	When falsified accept the facts and stop the method

5.7.20.1 Data fallacies

So far we have presented in details some common human problems and interpretation issues in statistics as (in addition to the deplorable way that the mass media communicate scientific results as seen on page 79) the fact that correlation is not always causation, a zero correlation doesn't mean an absence of pattern, the Simpson's paradox, the Anscombe's quartet, the usage of the wrong statistics indicators or tests, p-hacking, etc.

Some of them can be summarized in the following poster (remember that poster for when we will study Data Mining on page 6768):



DATA FALLACIES TO AVOID

CHERRY PICKING
Selecting results that fit your claim and excluding those that don't.

DATA DREDGING
Repeatedly testing new hypotheses against the same set of data, failing to acknowledge that most correlations will be the result of chance.

SURVIVORSHIP BIAS
Drawing conclusions from an incomplete set of data, because that data has 'survived' some selection criteria.

COBRA EFFECT
Setting an incentive that accidentally produces the opposite result to the one intended. Also known as a Perverse Incentive.

FALSE CAUSALITY
Falsely assuming when two events appear related that one must have caused the other.

GERRYMANDERING
Manipulating the geographical boundaries used to group data in order to change the result.

SAMPLING BIAS
Drawing conclusions from a set of data that isn't representative of the population you're trying to understand.

GAMBLER'S FALLACY
Mistakenly believing that because something has happened more frequently than usual, it's now less likely to happen in future (and vice versa).

HAWTHORNE EFFECT
The act of monitoring someone can affect their behaviour, leading to spurious findings. Also known as the Observer Effect.

REGRESSION FALLACY
When something happens that's unusually good or bad, it will revert back towards the average over time.

SIMPSON'S PARADOX
When a trend appears in different subsets of data but disappears or reverses when the groups are combined.

MCNAMARA FALLACY
Relying solely on metrics in complex situations and losing sight of the bigger picture.

OVERFITTING
Creating a model that's overly tailored to the data you have and not representative of the general trend.

PUBLICATION BIAS
Interesting research findings are more likely to be published, distorting our impression of reality.

DANGER OF SUMMARY METRICS
Only looking at summary metrics and missing big differences in the raw data.

G GECKOBOARD.COM

Read more at **geckoboard.com/learn/data-literacy**

Figure 5.260: **Common Data fallacies**
(credit: Geckoboard, source: https://www.geckoboard.com/learn/data-literacy/**)**

❞ *Even a rough estimate, even a false one, is better than no estimate at all.* ❞

quote
5.7:1343.#52

~90 %

★★★★★

125 votes, 62.5%

6

Algebra

Algebra is the science of calculating the quantities or structures represented by letters. (Larousse)

Contents

6.1 Calculus **1350**
6.1.1 Equations and Inequations 1351
6.1.1.1 Equations 1352
6.1.1.2 Inequations 1356
6.1.2 Remarkable Identities 1361
6.1.3 Polynomials 1368
6.1.3.1 General relations between roots and coefficients 1373
6.1.3.2 Multiplication of Polynomials 1374
6.1.3.3 Euclidean Division of Polynomials 1374
6.1.3.4 Factorization Theorem of Polynomials 1377
6.1.3.5 Diophantine equation 1378
6.1.3.6 First order univariate Polynomial and Equations 1379
6.1.3.7 Second order univariate Polynomial and Equations 1382
6.1.3.8 Third order univariate Polynomial and Equations 1390
6.1.3.9 Fourth order univariate Polynomial and Equations 1394
6.1.3.10 Trigonometric Polynomials 1397
6.1.3.11 Cyclotomic Polynomials 1398
6.1.3.12 Legendre Polynomials 1400
6.1.3.13 Laguerre Polynomials 1404
6.2 Set Algebra **1417**
6.2.1 Groups Algebra and Geometry 1417
6.2.1.1 Cyclic Groups 1418

6.2.1.2 Transformations Groups 1421
6.2.1.3 Group of Symmetries 1430
6.2.1.4 Permutations Groups 1435
6.2.2 Galois Theory 1445
6.2.2.1 Elementary symmetric and Invariant Polynomials 1446
6.2.2.2 General Vieta's formulas 1449
6.3 Differential and Integral Calculus 1450
6.3.1 Differential Calculus 1450
6.3.1.1 Differentials 1464
6.3.1.2 Usual Derivatives 1475
6.3.1.3 Implicit Differentiation 1492
6.3.1.4 Smoothness 1496
6.3.2 Integral Calculus 1497
6.3.2.1 Definite Integral 1497
6.3.2.2 Indefinite Integral 1508
6.3.2.3 Double Integral 1518
6.3.2.4 Integration by Substitution 1522
6.3.2.5 Integration by Parts 1532
6.3.2.6 Usual Primitives 1534
6.3.2.7 Leibniz integral rule 1560
6.3.2.8 Integral representation of first kind Bessel's function 1564
6.3.2.9 Dirac Function 1569
6.3.2.10 Gamma Euler Function 1574
6.3.2.11 Curvilinear Integrals 1583
6.3.2.12 Surface Integrals 1588
6.3.2.13 Integrals involving parametric equations 1593
6.3.2.14 Improper Integrals 1595
6.3.2.15 Elliptic Integrals 1599
6.3.3 Differential Equations 1603
6.3.3.1 First order differential equations 1605
6.3.3.2 Linear differential equations 1605
6.3.3.3 Resolution methods of linear differential equations 1607
6.3.3.4 Classification of differential equations 1621
6.3.3.5 Systems of Differential Equations 1627
6.3.3.6 Nonlinear dynamical systems 1644
6.3.3.7 Regular Methods of Perturbations 1645
6.4 Sequences and Series 1651

6.4.1 Sequences . . . 1651
6.4.1.1 Arithmetic Sequences . . . 1653
6.4.1.2 Harmonic Sequences . . . 1655
6.4.1.3 Geometric Sequences . . . 1656
6.4.1.4 Cauchy Sequence . . . 1657
6.4.1.5 Fibonacci Sequence . . . 1661
6.4.1.6 Logic Sequences/Psychologist Sequences . . . 1663
6.4.2 Series . . . 1664
6.4.2.1 Gauss Series . . . 1666
6.4.2.2 Arithmetic Series . . . 1679
6.4.2.3 Geometric Series . . . 1679
6.4.2.4 Telescoping Series . . . 1691
6.4.2.5 Grandi's Series . . . 1691
6.4.2.6 Taylor and Maclaurin Series . . . 1696
6.4.2.7 Fourier Series (trigonometric series) . . . 1727
6.4.2.8 Bessel Series . . . 1768
6.4.3 Convergence Criteria . . . 1775
6.4.3.1 Integral Test . . . 1776
6.4.3.2 D'Alembert Rule . . . 1777
6.4.3.3 Alternating Series Test . . . 1780
6.4.3.4 Fixed Point Theorem . . . 1780
6.4.4 Generating Functions . . . 1783
6.4.4.1 Ordinary Generating Functions (transformation of a sequence into a series) . . . 1783
6.4.4.2 Multivariate Generating Functions . . . 1789
6.4.4.3 Functional Generating Functions . . . 1789
6.5 Vector Calculus . . . 1794
6.5.1 Concept of Arrow . . . 1795
6.5.2 Set of Vectors . . . 1797
6.5.2.1 Pseudo-Vectors . . . 1799
6.5.2.2 Normal vector . . . 1801
6.5.2.3 Multiplication by a scalar . . . 1802
6.5.3 Vector Spaces . . . 1805
6.5.3.1 Linear Combinations . . . 1806
6.5.3.2 Sub-vector spaces . . . 1806
6.5.3.3 Generating families . . . 1807
6.5.3.4 Linear Dependence or Independence . . . 1807
6.5.3.5 Base of a vectorial space . . . 1809

6.5.3.6 Direction Angles . . . 1811
6.5.3.7 Dimensions of a vector space . . . 1812
6.5.3.8 Extension of a free family . . . 1813
6.5.3.9 Rank of a finite family . . . 1813
6.5.3.10 Direct Sums . . . 1814
6.5.3.11 Affine spaces . . . 1815
6.5.4 Euclidean Vector Spaces . . . 1818
6.5.4.1 Scalar Product (Dot Product) . . . 1821
6.5.4.2 Cross Product . . . 1830
6.5.4.3 Mixed Product (triple product) . . . 1838
6.5.5 Vectorial Functional Space . . . 1840
6.5.6 Hermitian Vector Space . . . 1842
6.5.6.1 Hermitian Inner Product . . . 1843
6.5.6.2 Hermitian Cauchy-Schwarz inequality . . . 1844
6.5.6.3 Types of Vectors Spaces . . . 1845
6.5.7 System of Coordinates . . . 1846
6.5.7.1 Cartesian (rectangular) Coordinate System . . . 1846
6.5.7.2 Spherical Coordinate System . . . 1849
6.5.7.3 Cylindrical Coordinate System . . . 1861
6.5.7.4 Polar Coordinate System . . . 1863
6.5.8 Differential Operators . . . 1866
6.5.8.1 Gradients of Scalar Field . . . 1869
6.5.8.2 Gradients of Vector Field . . . 1878
6.5.8.3 Divergences of a Vector Field . . . 1879
6.5.8.4 Rotationals of a Vector Field (Curl) . . . 1888
6.5.8.5 Laplacians of Scalar Fields (Laplace Operator) . . . 1899
6.5.8.6 Laplacians of Vector Fields . . . 1903
6.5.8.7 Remarkable Identities . . . 1911
6.5.8.8 Summary . . . 1918
6.6 Linear Algebra . . . 1924
6.6.1 Linear Systems . . . 1929
6.6.2 Linear Transformations . . . 1934
6.6.3 Matrices . . . 1936
6.6.3.1 Rank of a matrix . . . 1938
6.6.3.2 Matrix Algebra . . . 1945
6.6.3.3 Type of Matrices . . . 1948
6.6.3.4 Determinant . . . 1961

6.6.4 Change of Basis (frames) . . . 1992
6.6.5 Eigenvalues and Eigenvectors . . . 1997
6.6.5.1 Rotation Matrices and Eigenvalues . . . 2001
6.6.6 Spectral Theorem . . . 2004
6.6.7 Matrix Decompositions . . . 2011
6.6.7.1 Singular Value Decomposition (SVD) . . . 2011
6.6.7.2 LU Decomposition . . . 2024
6.6.7.3 Cholesky's Decomposition . . . 2027
6.6.7.4 QR Decomposition . . . 2030
6.6.8 Woodbury Matrix identity and Sherman-Morrison relation . . . 2035
6.7 Tensor Calculus . . . 2038
6.7.1 Tensor . . . 2039
6.7.2 Indicial Notation . . . 2042
6.7.2.1 Summation on multiple index . . . 2044
6.7.2.2 Kronecker Symbol . . . 2044
6.7.2.3 Antisymmetric Symbol (Levi-Civita symbol) . . . 2045
6.7.3 Metric and Signature . . . 2051
6.7.4 Gram's Determinant . . . 2054
6.7.5 Contravariant and Covariant Components . . . 2059
6.7.6 Operation in Basis . . . 2062
6.7.6.1 Gram-Schmidt Orthogonalization Method . . . 2062
6.7.6.2 Change of Basis . . . 2066
6.7.6.3 Reciprocal Basis (Dual Basis) . . . 2067
6.7.7 Euclidean Tensors (cartesian tensor) . . . 2071
6.7.7.1 Fundamental Tensor . . . 2072
6.7.7.2 Tensor product (dyadic) of two vectors and matrices . . . 2073
6.7.7.3 Tensor Spaces . . . 2076
6.7.7.4 Linear combination of tensors . . . 2080
6.7.7.5 Contraction of indices . . . 2080
6.7.8 Special Tensors . . . 2085
6.7.8.1 Symmetric Tensor . . . 2085
6.7.8.2 Antisymmetric Tensor . . . 2088
6.7.8.3 Fundamental Tensor . . . 2091
6.7.9 Curvilinear Coordinates . . . 2091
6.7.9.1 Natural basis in spherical coordinates (curvilinear basis in spherical coordinates) 2097
6.7.9.2 Natural basis in polar coordinates (curvilinear basis in polar coordinates) . . . 2099
6.7.9.3 Natural basis in cylindrical coordinates (cylindrical basis in polar coordinates) . 2099

6.7.10 Christoffel symbols 2103

6.7.11 Ricci Theorem 2112

6.7.12 Riemann-Christoffel symbols 2122

6.7.13 Ricci curvature (Ricci tensor) 2127

6.7.14 Einstein Tensor 2141

6.8 Spinor Calculus 2145

6.8.1 Unit Spinor 2146

6.8.2 Geometric Properties 2152

6.8.2.1 Plane Symmetries 2153

6.8.2.2 Rotations 2155

6.8.2.3 Properties of Pauli Matrices 2161

6.1. Calculus

In the Arithmetic chapter of this textbook we have written extensively on various theorems using abstract numbers in order to extend the scope of the validity of the latter. However we have only discussed a few on how we should handle these abstract numbers. This is what we will see now.

As you may know already know, a number may be considered by making an abstraction of the nature of the objects that constitutes the group that it characterizes and as well as how to codify it (Indo-Arabic numbers, Roman numbers or other system...). We then say that the number is an "abstract number" and when we handle these kinds of objects we say that we are doing "algebra calculus" or also "literal calculation".

Definition (#6.1.165): "Literal calculation" is the fact to calculate with variables (that is to say with letters) as you would with numbers.

For the mathematicians it is often not advantageous to work with numerical values $(1,2,3,\ldots)$ because they represent only specific cases. What look for physicists, engineers and mathematicians are universally applicable relations in a most general framework as possible.

These abstract numbers today commonly named "variable" are often represented by the Latin alphabet (for which the first letters of the Latin alphabet $a,b,c,\ldots$ often denote imposed values and the last $\ldots,x,y,z$ variables values), Greek alphabet is also mostly used to represent more or less complex mathematical operators. Latin and Greek letters are also used to represent "Universal constants" such as the speed of light c, the gravitational constant G, the value of π, the Euler number e, etc.

Remark

It seems that the letters to represent numbers were used for the first time by Viète in the middle of the 116th century (but the notation of exponents did not exist at this time).

A variable is therefore likely to take different numerical values. All these values can vary according to the character of the problem considered. Let us recall (we had already defined this in the section on Numbers of the Arithmetic chapter) that given two numbers a and b such that $a<b$, then:

R1. We name "domain of definition" of a variable, all numerical values it is likely to take between two specified limits (endpoints) or on a set (like $\mathbb{N},\mathbb{R},\mathbb{R}^+$, etc.).

R2. We name "closed interval with endpoints a and b", the set of all numbers x between these two values and we denote as example as follows:

$$[a,b]=\{x\in\mathbb{R}\,|\,a\leq x\leq b\} \tag{6.1.1}$$

R3. We name "open interval with endpoints a and b", the set of all numbers x between these two values not included and we denote it as example as follows:

$$]a,b[=\{x\in\mathbb{R}\,|\,a<x<b\} \tag{6.1.2}$$

R4. We name "interval closed, left open right" or "semi-closed left" the following relation as example:

$$[a,b[=\{x\in\mathbb{R}\,|\,a\leq x<b\} \tag{6.1.3}$$

R5. We name "interval open left, closed right" or "semi-closed right" the following relation as example:

$$]a,b]=\{x\in\mathbb{R}\,|\,a<x\leq b\} \tag{6.1.4}$$

Remark

If the variable x can take all possible negative and positive values we write therefore: $]-\infty,+\infty[$ where the symbol "∞" means "infinite". Obviously there can be combinations of open infinite right intervals with left endpoint and vice versa.

Definition (#6.1.166): We name "neighbourhood of a", any open interval containing a (it's a simple concept that we will use later to define a continuous function). So:

$$\forall\varepsilon>0,]a-\varepsilon,a+\varepsilon[\tag{6.1.5}$$

is a neighbourhood of a.

6.1.1 Equations and Inequations

Elementary algebra consists starting from the definitions of addition, subtraction, multiplication, and power and their properties (associativity, distributivity, commutativity, identity element, inverse, ...) - this is according to which set we are working with a field or a commutative abelian group or not (see section Set Theory page 416) - to handle within a fixed goal "algebraic equations" linking together variables and constants.

We will define afterwards what an equation and an inequality are but first we want to define some of their properties:

Let A and B be any two polynomials (or monomials) - see definitions a little further - the expressions:

$$A>B,A=B,A<B,A\leq B,A\geq B \tag{6.1.6}$$

satisfy the following properties:

P1. We can always add or subtract from the two members of an equation or inequality same polynomial by obtaining an equivalent equation or inequality (i.e. with the same solutions or reductions). We then say that the equality or inequality remain "true" by the operations of addition or subtraction member to member.

P2. If we multiply or if we divide both members of an equation or inequality by the same positive number we also get an equivalent equation or inequality (we have already seen this in the sections below). We then say that the equality or inequality remains "true" by the operation of multiplication or division member to member.

P3. If we multiply or if we divide both sides of an inequality by the same negative number and if we reverse the direction of inequality, then we get an inequality or equivalent equation.

6.1.1.1 Equations

Definition (#6.1.167): An "equation" is a relation of equality between all abstract values (i.e.: two algebraic expressions) or not all abstract (since we're talking about equations with one unknown, two unknowns, three unknowns and some constants...) interconnected by various operators.

The perfect mastery of elementary algebra is fundamental in physics and mathematics and in the industry!!! Since there are an infinite number of types of equations, we will not present them all here. It is the role of the teacher/trainer in classes to train the brain of his audience for several years (2-3 years on average) to solve a lot of different configurations of algebraic equations (exposed as every day, purely mathematical problems or geometric problems) and this so that students manipulate these equations without errors in a logical and rigorous reasoning (practice makes perfect...)!!!

In other words: A teacher/trainer and an ad hoc institution are irreplaceable to acquire knowledge and experience and to have a feedback on experience!!!

Remark

We have attempted below to make a simple generalization of the basic rules of elementary algebra. This generalization will be more easy to understand to the reader that already have the habit of manipulating abstract quantities.

Thus, either $a, b, c, d, e, \ldots, x, y$ abstract numbers can take any numerical value (we stay within the high-school classical numbers...).

Let Ξ (the Greek capital letter ruling "Xi") representing one or more abstract numbers (variables) operating between them in any way as we have or not different and distinguishable algebraic monomials (one abstract number) or polynomials (poly = many). Therefore we do here a kind of abstract of the abstraction or if you prefer a variable of several variables.

Properties (in fact these are more examples that properties...):

P1. We will always have $\Xi=\Xi$ if and only if the term left Ξ of equality is the same term as the one Ξ that is on the right of the equality. If this condition is satisfied then we have:

$$\Xi-\Xi=0 \tag{6.1.7}$$

Otherwise:

$$\Xi-\Xi=\Xi \tag{6.1.8}$$

where we therefore exclude the case where all the Ξ above are identical to each other (otherwise we get again the property P1).

P2. We have if $\Xi \neq 0$:

$$\frac{\Xi}{\Xi}=1 \Leftrightarrow 1=1 \Leftrightarrow \Xi=\Xi \qquad (6.1.9)$$

which verifies the symbolism of the equation $\Xi=\Xi$ in the case only where the elements are identical between them (we obviously exclude the case with zero denominator).

P3. If all the Ξ are identical, then:

$$\Xi\cdot\Xi\cdot\Xi\cdot\ldots\cdot\Xi=\prod\Xi=\Xi^n \qquad (6.1.10)$$

Otherwise we have:

$$\Xi\cdot\Xi\cdot\Xi\cdot\ldots\cdot\Xi \qquad (6.1.11)$$

which cannot be written in a simple condensed form. It may also happen that:

$$\Xi\cdot\Xi\cdot\Xi\cdot\ldots\cdot\Xi=\Xi \qquad (6.1.12)$$

with the Ξ on the right of equality identical to none, one or more Ξ of the left member of equality.

P4. We can have:

$$\frac{\Xi^{\Xi}}{\Xi^{\Xi}}=1 \qquad (6.1.13)$$

without necessarily having the numerator or denominator to be equals (we exclude for sure the case where the denominator is equal to zero).

Otherwise we can also have:

$$\frac{\Xi^{\Xi}}{\Xi^{\Xi}}=\Xi \quad \text{or} \quad \frac{\Xi^{\Xi}}{\Xi^{\Xi}}=\Xi^{\Xi-\Xi} \qquad (6.1.14)$$

But don't forget (see section Operators page 313) that in the general case where the numerator and denominator are not equal that:

$$\frac{\Xi^{\Xi}}{\Xi^{\Xi}}\neq\frac{\Xi^{\Xi}+c^{te}}{\Xi^{\Xi}+c^{te}} \qquad (6.1.15)$$

P5. We have if all denominators are strictly equals:

$$\cfrac{\Xi}{\cfrac{\Xi}{\cfrac{\Xi}{\cfrac{\Xi}{\ldots}}}}=\frac{\Xi}{\Xi^n}=\Xi^{1-n} \qquad (6.1.16)$$

but it is however not impossible to have:

$$\frac{\Xi}{\frac{\Xi}{\frac{\Xi}{\frac{\Xi}{\ldots}}}}=\Xi \tag{6.1.17}$$

with the Ξ at the right of equality identical to any one or more of the left member of the equality or even it is quite possible to have:

$$\frac{\Xi}{\frac{\Xi}{\frac{\Xi}{\frac{\Xi}{\ldots}}}}=\Xi^{\Xi} \tag{6.1.18}$$

P6. Let $\star$ represent exclusively the addition or the subtraction symbol we have (at a given sign variation):

$$\star\Xi\star\Xi\star\ldots\star\Xi=n\cdot\Xi \tag{6.1.19}$$

if all the Ξ are identical to each other or if the combination of an indeterminate number of Ξ are equal to the Ξ on the right of equality.

Otherwise we will have (meaning that the result will give any undetermined monomial or polynomial):

$$\star\Xi\star\Xi\star\ldots\star\Xi=\Xi \tag{6.1.20}$$

P7. We have if all Ξ (based) raised at the power are strictly identical:

$$\Xi^{\Xi}\cdot\Xi^{\Xi}\cdot\ldots\cdot\Xi^{\Xi}=\Xi^{\sum\Xi} \tag{6.1.21}$$

if and only if the bases Ξ are equal (or could be decomposable to be equal) and the powers Ξ are not necessarily equals.

From the knowledge of these 7 basic examples rules/properties, we can solve, simplify or show that a simple equation has solutions or not relatively to a given problem or statement.

Thus, given an operand or a sequence of any operations on one or more abstraction of abstracts and among all, one (or more) of which the numeric values is or are unknown (the others are known, so we could be able to prove that a relation (statement) like:

$$\Psi\Xi=\Psi\Xi \tag{6.1.22}$$

has existing solutions or not (that is to say: is True or False).

In the case of an equation with the absolute value (see section Operators page 324) of the type:

$$|\Psi\Xi|=\Psi\Xi \tag{6.1.23}$$

with the second member being strictly positive (otherwise the previous relations would be a nonsense) this is equivalent of course from the definition of absolute value to write:

$$\Psi\Xi=\Psi\Xi \quad \text{and} \quad \Psi\Xi=-\Psi\Xi \tag{6.1.24}$$

Remarks

R1. The presence of the absolute value in an algebraic equation in which we seek solutions often doubles the number of solutions.

R2. An equation is named "conditional equation" when there are numbers in the set of definitions that are not solutions (which is actually the most common case). Conversely, if any number of the definition set is solution of the equation then the equation is named an "identity equation".

We sometimes have to solve a "system of equations" (see for example the linear system of equations at page 6304).

- What is it?: It is a set of at least two equations to solve (solving is not always equivalent as to simplify an expression!).
- What is the specificity of such a system?: All solutions of the system are the intersection of all the solutions of the equations to be solved (for detailed examples see the sections of Linear Algebra and Numerical Methods).
- What is the usage?: It is endless (see the different chapters of the book), these systems solve problems involving applications of mathematics to other fields (finance, engineering, operational research, etc.).

Because of the unlimited variety of applications, it is difficult to establish precise general rules for solutions and the existing procedures to follow will for sure be useful only if the problem can be formulated into equations and the following procedures can help a little bit at least to avoid some errors:

1. If we have a problem statement already written, we read it several times carefully and we consider the given facts and the amount of unknowns to find and their domain of definition (summarizing statement on a sheet of paper or anywhere else is often useful for large problems!).
2. Select letters that represents the unknown quantities. This is one of the decisive steps in the search for solutions. Sentences containing words such as: find, what, how, where, when should help you to identify the unknown quantity.
3. Make a drawing (in your head or on paper) with captions. This is for sure possible most of the time only with problems having 1, 2 or 3 unknowns.
4. List the known facts and relations about the unknown quantities. A relation can be described by an equation in which appear in one or both sides of the equal sign statements written with normal sentences.
5. After the previous step, write one or more equations that accurately describe what was stated with sentences.
6. Solve the equation or the system of equations formulated in the previous steps using of the multiple heuristic existing techniques.

7. Check the solutions obtained in the previous step by referencing to the initial problem statement (check that the solution is consistent with the conditions of the statement).

Some methods of resolutions of systems of equations are treated in detail in the section of Numerical methods and also in the section Linear Algebra (you will thus therefore better understand the procedure above).

6.1.1.2 Inequations

Previously we have seen that an equation was composed from an equality of various calculations with different terms (with at least one "unknown" or an "abstract number") and that:

1. "Solve an equation" is to process consisting to calculate the value of the unknown to satisfy the equality (when a solution exists!)
2. "Simplify an equation" is the process consisting to mathematically minimize the number of terms (factoring or eliminate...)
3. "Develop an equation" is the process consisting to flatten all the terms.

Why do we need to recall the definition of an equation? Just because for the inequality, it is almost the same intellectual process! The difference ? If the equation is an equality, inequality is an inequality (...): As the equation, the inequality is composed of various calculations with different terms interconnected by any operators with at least one unknown.

Main differences between equality and inequality equations:

1. Equality: Symbolized by the symbol $=$
2. Inequality: Symbolized by the strict or large order relations $<, \leq, \geq, >$

When we solve an inequality, our unknown may have a range of values that satisfy the inequality. We say then that the inequality of the solution is a "set of values". This is the fundamental difference between equality (several solutions) and inequality (range of solutions)!

Let us make a refresh about the signs that we can meet in an inequality:

- $<$: Must be read "strictly inferior to" or "strictly less than". In this case the target numerical value is not included in the range and we can then represent the range (interval) with an open left square bracket $]\ldots$ or right square bracket $\ldots[$ next to the target value.

✍**Example:**

Writing $x < 5$ means $x \in]-\infty, 5[$ and $x < -5$ means $x \in]-\infty, -5[$.

- $>$: Must be "strictly superior to" or "strictly greater than". In this case the numerical target value is also not included in the range (interval) and we can then represent the range with an open left square bracket $]\ldots$ or right square bracket $\ldots[$ next to the target value.

✍**Example:**

Writing $x > 5$ means $x \in]5, +\infty[$ and $x > -5$ means $x \in]-5, +\infty[$.

- $\leq$: Must be read "inferior or equal to" or "less than or equal to". In this case, the numerical target value is in the range (interval) and we can then represent the range with a closed left square bracket $[\ldots$ or right square bracket $\ldots]$.

✍**Example:**

Writing $x \leq 5$ means $x \in]-\infty, 5]$ and $x < -5$ means $x \in]-\infty, -5]$.

- $\geq$: Must be read "superior or equal to" or "greater than or equal to". In this case, the numerical target value is in the range (interval) and we can then represent the range with a closed left square bracket $[\ldots$ or right square bracket $\ldots]$.

✍**Example:**

Writing $x \geq 5$ means $x \in [5, +\infty[$ and $x > -5$ means $x \in [-5, +\infty[$.

The objective of inequalities is most of the time (except aesthetics purpose) to have at least one numeric value that defines the domain of solution (of all the abstract terms - variables - of the inequality) that satisfies the inequality.

There are many ways to represent the domains of definition of variables which satisfy an inequality. We will see through small examples what are some of the most common possibilities:

✍Example:

Given a linear inequality (of first degree) on x with a single unknown on which we impose an arbitrary particular constraint for example (of course, the expression can contain more terms...):

$$x<2 \tag{6.1.25}$$

we have in the above inequation already simplified all the terms that were superfluous. Solve the inequality is like looking for the x values less than 2. Of course, there is not only one solution in $\mathbb{R}$ but a set (interval) of solutions and this is the principle of inequalities!

To solve the inequality, we first observe the type of inequality imposed ("strict" or "equal"). Then, in high-schools classes (and not only sometimes...) we represent the set $\mathbb{R}$ traditionally by a table such as:

$-\infty$	0	$+\infty$
....	.......|.......	....

Table 6.1: **Table for inequality resolution**

We intuitively know that the solution of our inequality includes all values below 2 (2 being itself excluded from the solutions) until $-\infty$. Then we write this interval or domain as follows:

$$S=]-\infty,2[\tag{6.1.26}$$

Then we can represent in a tabular way the set of solutions (it helps to understand and prepares the student for solving systems of equations and inequalities and to the study of functions variations). For this, we take back the template of the previous table and place in it our target value (we only have one in this special example but sometimes there can be several because there is a singularity or roots for certain values of the definition domain), that is to say the value 2:

$-\infty$	0	2	$+\infty$
....	.......|.......	.......[.......	....

Table 6.2: **Augmented table for inequality resolution**

and finally, we delimitate with color (...) the set of solutions to from $-\infty$ to $+2$ excluded:

$-\infty$	0	2	$+\infty$
....	.......|.......	.......[.......	....

Table 6.3: **Highlighted table for inequality resolution**

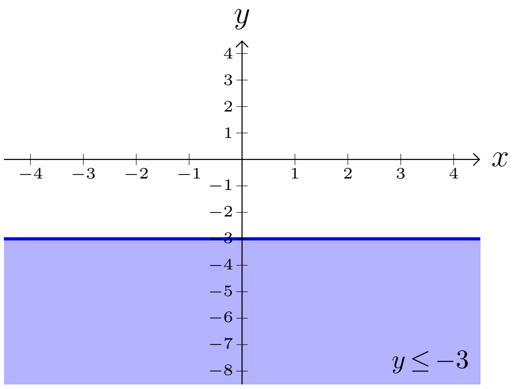
At the value 2, we do not forget to mark the sign $\ldots[$ to show that this value is excluded from the solutions. And that's it, and the concept can be extrapolated to much more complex inequalities.

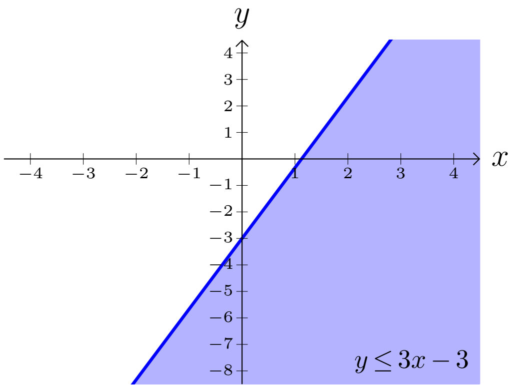
Remarks

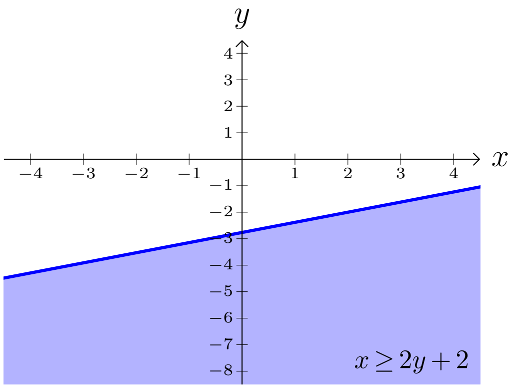
R1. Sometimes instead of representing tables as we have done above, some teachers (this is a completely artistic choice) ask their students to shade the boxes of the table and to draw small circles inside..., or they also use small arrows, or draw the graph of the inequality (the latter method is certainly aesthetic but takes time in complex case but an example is given below...).

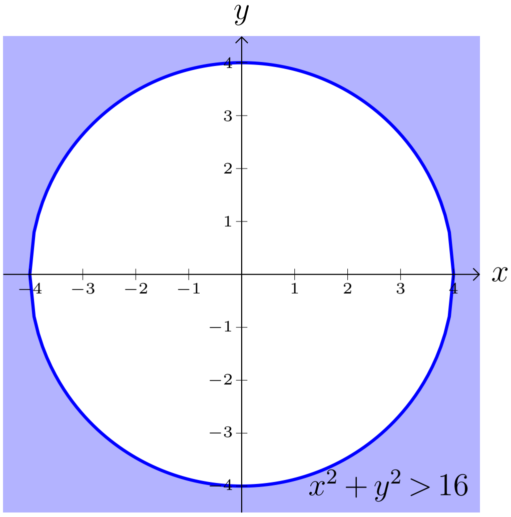
R2. As part of inequalities of degree greater than 1, it is necessary (see later what that means exactly) first too determine the roots of the inequality that determine the intervals and then by trial and error, determine which intervals are to reject or keep.

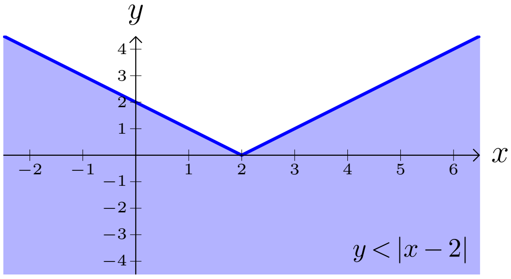
Here are some graphical examples of some very simple univariate inequalities:











We can also (just like equations) sometimes have to solve a "system of inequalities". What is it ?: It is a set of at least two inequalities to solve. The peculiarity of the system ?: The set of solutions of the system is the intersection of all the solutions of every inequality.

For example the system of three inequalities:

$$\begin{aligned} -x_1+2x_2 &\leq 3 \\ 2x_1-x_2 &\leq 5 \\ x_1+x_2 &\geq 3 \end{aligned} \tag{6.1.27}$$

In other words, the method is the same as the previous one, with the difference that our table (representing the areas of solutions) will include an additional line for each additional inequality in the system plus one line of synthesis which is the projection of the possible solutions areas of the system.

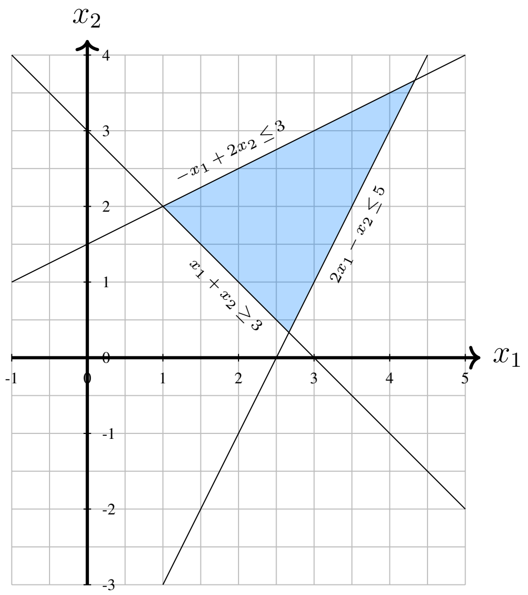
Thus, a system of n inequalities will have obviously a summary table with $n+1$ lines.

Mathematically, the areas (there may be several which are disjoint) can be written as a set of domains:

$$S=\{]\ldots,\ldots],[\ldots,\ldots],\ldots,[\ldots,\ldots[\} \tag{6.1.28}$$

Systems of inequalities are very common in many problems of mathematics, physics, econometrics, etc. It is important to practice solve them during your studies with the help of your teacher.

For example, here is a possible representation of the solution domain of the previous system of inequalities:



6.1.2 Remarkable Identities

The remarkable identities are some kind of magical relations that we use most often for factoring or solving algebraic equations in all fields relative to Applied Mathematics. They have an important place and must be absolutely be mastered by the reader.

Let us recall some notions that have already study in the section of Set Theory and of the chapter Arithmetic (we assume the concept of neutral element known as already defined):

- Commutativity:

$$a+b=b+a \qquad \text{and} \qquad ab=ba \tag{6.1.29}$$

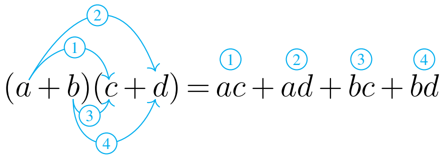
- Associativity:

$$a+(b+c)=(a+b)+c \qquad \text{and} \qquad a(bc)=(ab)c \tag{6.1.30}$$

- Distributivity (for reminder the inverse operation being named the factorization):

$$a(b+c)=ab+ac \tag{6.1.31}$$

Or illustrated:



$$(a+b)(c+d)=\overset{①}{ac}+\overset{②}{ad}+\overset{③}{bc}+\overset{④}{bd}$$

Similar properties are valid with the subtraction operation of course with the adequate domain of definition.

We can check with numerical values (by replacing each abstract number a randomly chosen number), or by development (it would be better, so you are sure you have understood what we were talking about until now), that the following algebraic identities are satisfied (they are the most known one):

1. Second degree identity:

$$(a+b)^2=(a+b)(a+b)=a^2+ab+ba+b^2=a^2+2ab+b^2 \tag{6.1.32}$$

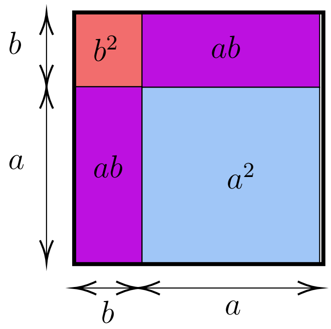


Figure 6.1: Visual of squared binomial identity

A funny mental application of this second degree identity (apart from the fact that the latter is encountered a lot in engineering and physics) is that you can easily found the square of any two-digit number (and maybe impress some of your friends) as any two-digit number xy can be rewritten as $10\cdot x+y$ and therefore the square can be found by easily applying:

$$(10x+y)^2=100x^2+20xy+y^2$$

2. Third degree identity:

$$(a+b)^3=(a+b)(a+b)^2=(a+b)(a^2+2ab+b^2)=a^3+3a^2b+3ab^2+b^3 \tag{6.1.33}$$

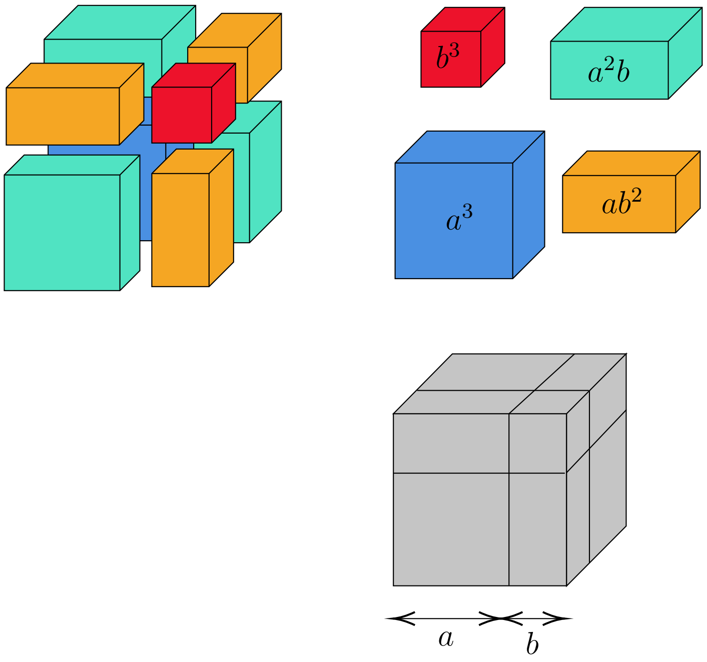


Figure 6.2: Visual of cubic binomial identity

Remark

We can very well put that in practice that $(a+(c+d))^2 := (a+b)^2$ where we obviously putted that $b := c+d$ (we do an "abstract of the abstraction" or more commonly: a "change of variable"). And therefore:

$$(a+b+c)^2 = a^2+2a(b+c)+(b+c)^2 = a^2+b^2+c^2+2ab+2bc+2ac \tag{6.1.34}$$

We can thus notice that in generality, to calculate the development $(a+b)^n$, we use the development of $(a+b)^{n-1}$, that is to say with the identity calculated with the previous value of n.

We can also see a pattern if we put all together:

$$\begin{aligned}(a+b)^0 &= 1\\ (a+b)^1 &= 1a+1b\\ (a+b)^2 &= 1a^2+2ab+1b^2\\ (a+b)^3 &= 1a^3+3a^2b+3ab^2+1b^3\\ &\ldots\end{aligned} \tag{6.1.35}$$

We thus notice the following properties for a and b:

1. The powers of a decreasing from n to 0 ($a^0 = 1$, so it is not noted in the last term).
2. The powers of b increase from 0 to n ($a^0 = 1$, so it is not noted in the first term).
3. In each term, the sum of the powers of a and b is equal to n.
4. The multiplier coefficients in front of each term are calculated by summing the multipliers of two terms of development achieved with the previous value of n (see figure below).

The so named "binomial coefficients" can then be obtained by the construction of the "Pascal's triangle" below:

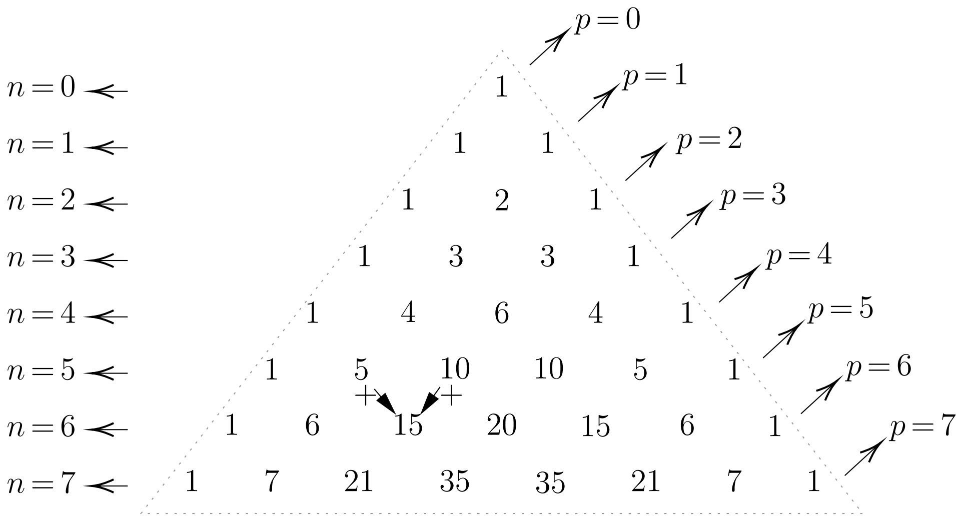


Figure 6.3: **Pascal triangle**

Where each element is given by (see section Probabilities page 522):

$$C_p^n=\frac{n!}{p!(n-p)!} \tag{6.1.36}$$

with $n,p\in\mathbb{N}^+$.

Theorem 6.1. We can then prove that:

$$\boxed{(a+b)^n=\sum_{p=0}^{n}C_p^n a^{n-p}b^p} \tag{6.1.37}$$

which is the famous "Newton's binomial" (that we will reuse many times in different chapters and thus sections of this textbook) or also named "Binomial theorem".

More explicitly:

$$(a+b)^n=\sum_{p=0}^{n}C_p^n a^{n-p}b^p=a^n+\frac{n}{1}a^{n-1}b+\frac{n(n-1)}{1\cdot 2}a^{n-2}b^2+\frac{n(n-1)(n-2)}{1\cdot 2\cdot 3}a^{n-3}b^3+\ldots+b^n \tag{6.1.38}$$

also named "binomial series".

Proof 6.1.1. This relation can be proved quite simply by induction assuming true the previous relation and calculating it for the rank 1:

$$(a+b)^1=a+b=C_0^1a^1b^0+C_1^1a^0b^1=a+b \tag{6.1.39}$$

Then we get:

$$\begin{aligned}(a+b)^{n+1}&=(a+b)(a+b)^n=(a+b)\sum_{p=0}^{n}C_p^n a^{n-p}b^p\\&=\sum_{p=0}^{n}C_p^n a^{n+1-p}b^p+\sum_{p=0}^{n}C_p^n a^{n-p}b^{p+1}\\&=C_0^n a^{n+1}+\sum_{p=1}^{n}C_p^n a^{n+1-p}b^p+\sum_{p=0}^{n-1}C_p^n a^{n-p}b^{p+1}+C_n^n b^{p+1}\\&=C_0^n a^{n+1}+\sum_{p=1}^{n}C_p^n a^{n+1-p}b^p+\sum_{p=1}^{n}C_{p-1}^n a^{n+1-p}b^p+C_n^n b^{p+1}\\&=C_0^n a^{n+1}+\sum_{p=1}^{n}\left(C_p^n+C_{p-1}^n\right)a^{n+1-p}b^p+C_n^n b^{p+1}\\&=C_0^n a^{n+1}+\sum_{p=1}^{n}C_p^{n+1}a^{n+1-p}b^p+C_n^n b^{p+1}\\&=\sum_{p=0}^{n+1}C_p^{n+1}a^{n+1-p}b^p\end{aligned} \tag{6.1.40}$$

The relation is true for rank $n+1$, so it is true for any n.

■ Q.E.D.

By applying the binomial formula twice, we can get for the fun the following trinomial formula:

$$\begin{aligned}(a+b+c)^n&=((a+b)+c)^n\\&=\sum_{k=0}^{n}C_k^n(a+b)^kc^{n-k}=\sum_{k=0}^{n}C_k^n\left(\sum_{l=0}^{k}C_l^ka^lb^{k-l}\right)c^{n-k}\\&=\sum_{k=0}^{n}\sum_{l=0}^{k}C_k^nC_l^ka^lb^{k-l}c^{n-k}=\sum_{k=0}^{n}\sum_{l=0}^{k}\frac{n!}{k!(n-k)!}\cdot\frac{k!}{l!(k-l)!}a^lb^{k-l}c^{n-k}\\&=\sum_{k=0}^{n}\sum_{l=0}^{k}\frac{n!}{l!(k-l)!(n-k)!}a^lb^{k-l}c^{n-k}=\sum_{0\leq l,j,m\leq n,l+j+m=n}\frac{n!}{l!j!m!}a^lb^jc^m\end{aligned}\tag{6.1.41}$$

Regarding remarkable identities with negative values, there is no need to memorize the location of the sign "$-$". Just make a change of variable and once the development is made we replace the variable back (inverse change of variable).

✍**Example:**

$$\begin{aligned}(a-b)^2\underset{c:=-b}{=}(a+c)^2=a^2+2ac+c^2=a^2+2a(-b)+(-b)^2\\=a^2-2ab+(-b)(-b)=a^2-2ab+b^2\end{aligned}\tag{6.1.42}$$

and so on for all finite power n.

We can of course mix genders (...) such as (particularly famous example):

$$(a+b)(a-b)=a(a-b)+b(a-b)=a^2-b^2\tag{6.1.43}$$

and some remarkable additional practical relation that are often used in small classes for exercises:

$$\begin{aligned}a^3+b^3&=(a+b)(a^2-ab+b^2)\\a^3-b^3&=(a-b)(a^2+ab+b^2)\\(x+a)(x+b)&=x^2+(a+b)x+ab\\(x+a)(x-b)&=x^2+(a-b)x-ab\\(x-a)(x-b)&=x^2-(a+b)x+ab\\(ax+b)(cx+d)&=acx^2+(ad+bc)x+bd\end{aligned}\tag{6.1.44}$$

Remark

When, from the right-hand side (in a simplified numerical form) the teacher asks his students as an exercise to get the factoring of the left of the equality, there is no other way but to proceed by successive tests.

For information we also get following famous development that is immediately deducible from what we have seen before:

$$(1+b)^n=\sum_{p=0}^{n}C_p^n 1^{n-p}b^p=\sum_{p=0}^{n}C_p^n b^p=1+nb+\frac{n(n-1)}{2!}b^2+\ldots \tag{6.1.45}$$

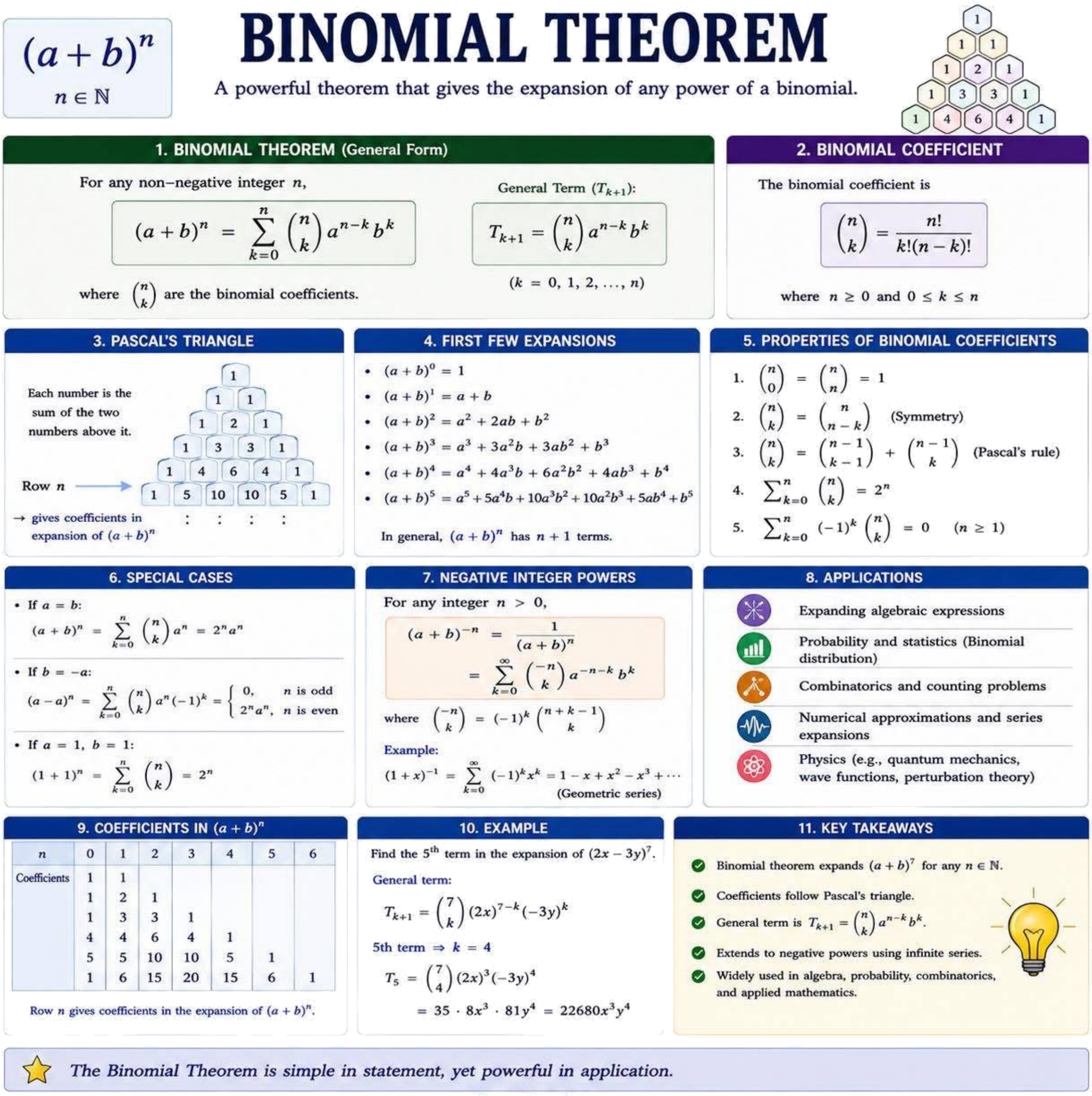
which is valid for any value of b and is named the "binomial expansion".

For small b we could neglect all terms involving b^2 or higher power of b, giving us the approximation relation:

$$(1+b)^n\cong 1+nb \tag{6.1.46}$$

with $b\ll 1$.

Of course there is still a lot of more useful relations with monomial (from which the biggest part is only a generalization of those presented above) that the reader will discover by his own reasoning and according to his practice and in this textbook through the different chapters.



$(a+b)^n$

$n\in\mathbb{N}$

BINOMIAL THEOREM

A powerful theorem that gives the expansion of any power of a binomial.

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1

1. BINOMIAL THEOREM (General Form)

For any non–negative integer n,

$$(a+b)^n=\sum_{k=0}^{n}\binom{n}{k}a^{n-k}b^k$$

where $\binom{n}{k}$ are the binomial coefficients.

General Term (T_{k+1}):

$$T_{k+1}=\binom{n}{k}a^{n-k}b^k$$

$(k=0,1,2,\ldots,n)$

2. BINOMIAL COEFFICIENT

The binomial coefficient is

$$\binom{n}{k}=\frac{n!}{k!(n-k)!}$$

where $n\geq 0$ and $0\leq k\leq n$

3. PASCAL'S TRIANGLE

Each number is the sum of the two numbers above it.

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1
: : : :

Row n →

→ gives coefficients in expansion of $(a+b)^n$

4. FIRST FEW EXPANSIONS

- $(a+b)^0=1$
- $(a+b)^1=a+b$
- $(a+b)^2=a^2+2ab+b^2$
- $(a+b)^3=a^3+3a^2b+3ab^2+b^3$
- $(a+b)^4=a^4+4a^3b+6a^2b^2+4ab^3+b^4$
- $(a+b)^5=a^5+5a^4b+10a^3b^2+10a^2b^3+5ab^4+b^5$

In general, $(a+b)^n$ has $n+1$ terms.

5. PROPERTIES OF BINOMIAL COEFFICIENTS

1. $\binom{n}{0}=\binom{n}{n}=1$
2. $\binom{n}{k}=\binom{n}{n-k}$ (Symmetry)
3. $\binom{n}{k}=\binom{n-1}{k-1}+\binom{n-1}{k}$ (Pascal's rule)
4. $\sum_{k=0}^{n}\binom{n}{k}=2^n$
5. $\sum_{k=0}^{n}(-1)^k\binom{n}{k}=0 \quad (n\geq 1)$

6. SPECIAL CASES

- If $a=b$:

$$(a+b)^n=\sum_{k=0}^{n}\binom{n}{k}a^n=2^na^n$$

- If $b=-a$:

$$(a-a)^n=\sum_{k=0}^{n}\binom{n}{k}a^n(-1)^k=\begin{cases}0, & n \text{ is odd}\\ 2^na^n, & n \text{ is even}\end{cases}$$

- If $a=1,\ b=1$:

$$(1+1)^n=\sum_{k=0}^{n}\binom{n}{k}=2^n$$

7. NEGATIVE INTEGER POWERS

For any integer $n>0$,

$$(a+b)^{-n}=\frac{1}{(a+b)^n}=\sum_{k=0}^{\infty}\binom{-n}{k}a^{-n-k}b^k$$

where $\binom{-n}{k}=(-1)^k\binom{n+k-1}{k}$

Example:

$$(1+x)^{-1}=\sum_{k=0}^{\infty}(-1)^kx^k=1-x+x^2-x^3+\cdots$$

(Geometric series)

8. APPLICATIONS

Expanding algebraic expressions

Probability and statistics (Binomial distribution)

Combinatorics and counting problems

Numerical approximations and series expansions

Physics (e.g., quantum mechanics, wave functions, perturbation theory)

9. COEFFICIENTS IN $(a+b)^n$

n	0	1	2	3	4	5	6
Coefficients	1	1					
	1	2	1				
	1	3	3	1			
	4	4	6	4	1		
	5	5	10	10	5	1	
	1	6	15	20	15	6	1

Row n gives coefficients in the expansion of $(a+b)^n$.

10. EXAMPLE

Find the 5th term in the expansion of $(2x-3y)^7$.

General term:

$$T_{k+1}=\binom{7}{k}(2x)^{7-k}(-3y)^k$$

5th term $\Rightarrow k=4$

$$T_5=\binom{7}{4}(2x)^3(-3y)^4$$

$$=35\cdot 8x^3\cdot 81y^4=22680x^3y^4$$

11. KEY TAKEAWAYS

- ✓ Binomial theorem expands $(a+b)^7$ for any $n\in\mathbb{N}$.
- ✓ Coefficients follow Pascal's triangle.
- ✓ General term is $T_{k+1}=\binom{n}{k}a^{n-k}b^k$.
- ✓ Extends to negative powers using infinite series.
- ✓ Widely used in algebra, probability, combinatorics, and applied mathematics.

The Binomial Theorem is simple in statement, yet powerful in application.

Remark

It is of course possible to multiply polynomials between them and distribute the multiplicative terms. Conversely, it is often asked to students in small classes to do the reverse procedure ("factoring" or "decompose" a polynomial) so they get used to the handling of remarkable identities. Decomposed into a product of factors is an important operation in mathematics, since it is thus possible to reduce the study of complicated expressions to the study of several simpler expressions having interesting properties depending on the context.

We can generalize the binomial theorem (that's useful when we need to prove the epsilon-delta continuity of some power functions) such that for $|x|<1$ and a real number α, we still can write $(1+x)^\alpha$ as the convergent series:

$$(1+x)^\alpha=\sum_{k=0}^{+\infty}\begin{pmatrix}\alpha\\k\end{pmatrix}x^k \tag{6.1.47}$$

Where for reminder:

$$\begin{pmatrix}\alpha\\k\end{pmatrix}=\frac{\alpha!}{k!(\alpha-k)!}=\frac{\alpha(\alpha-1)(\alpha-2)\cdots(\alpha-k+1)}{k!} \tag{6.1.48}$$

For example:

$$\begin{aligned}\begin{pmatrix}1/2\\4\end{pmatrix}&=\frac{\frac{1}{2}\cdot\left(-\frac{1}{2}\right)\cdot\left(-\frac{3}{2}\right)\cdot\left(-\frac{5}{2}\right)}{24}=-\frac{15}{16\cdot 24}\\ \begin{pmatrix}-n\\k\end{pmatrix}&=\frac{-n\cdot(-n-1)\cdots(-n-k+1)}{k!}=(-1)^k\frac{n(n+1)\cdots(n+k-1)}{k!}=\\&=(-1)^k\frac{(n+k-1)!}{(n-1)!k!}=(-1)^k\begin{pmatrix}n+k-1\\k\end{pmatrix}\end{aligned} \tag{6.1.49}$$

Now, to use this effectively when $(a+b)^{-n}$, you need at least one between a and b not to be zero and $|a|\neq|b|$. Then, pick the one with the highest absolute value and factor it out. If it is a, this means writing all as:

$$a^{-n}\left(1+\frac{b}{a}\right)^{-n}=a^{-n}\sum_{k=0}^{\infty}\begin{pmatrix}-n\\k\end{pmatrix}\frac{b^k}{a^k}=\sum_{k=0}^{\infty}\begin{pmatrix}n+k-1\\k\end{pmatrix}(-b)^k a^{-k-n} \tag{6.1.50}$$

This can be even extended further to complex numbers ($\mathbb{C}$) (again for the study of the continuity of complex defined functions) using the Euler Gamma function (see section Integral and Differential Calculus page 1574) by using the fact that (however it is not defined when $x+1$ is equal to zero or to a negative integer):

$$\Gamma(x+1)=x! \tag{6.1.51}$$

Therefore:

$$\begin{pmatrix}\alpha\\k\end{pmatrix}=\frac{\alpha!}{k!(\alpha-k)!}=\frac{\Gamma(\alpha+1)}{\Gamma(k+1)\Gamma(\alpha-k+1)} \tag{6.1.52}$$

6.1.3 Polynomials

Definition (naive version #168): We name "algebraic univariate polynomial $P(x)$" a function of degree $n\in\mathbb{N}$ which is written:

$$\boxed{P_n(x)=P(x)=a_nx^n+a_{n-1}x^{n-1}+\ldots+a_1x^1+a_0} \tag{6.1.53}$$

or in a more condensed manner by:

$$P_n(x)=\sum_{k=0}^{n}a_kx^k \tag{6.1.54}$$

where the "dominant factor" of a polynomial, also named "leading coefficient", is the coefficient of the monomial of highest degree a_n and the "leading term" is simply a_nx^n. If $a_n=1$ is equal to 1 then we speak of "normalized polynomial".

If we consider each coefficient of the polynomial as the component of a vector, then we can manipulate polynomials as vector spaces denoted $P_n(\mathbb{R})$ and of dimension $n+1$.

Remarks

R1. The subscript n of $P(x)$ is most of time omitted as explicitly defined in the expression of the polynomial itself.

R2. The reader who has read the section of Set Theory, probably remember that the set of all polynomial of degree n or lower form a vector space structure!

We will assume without proof and obvious (because we are engineers or physicists but not mathematicians) that if we denote $\text{deg}(P_n(x))=n$ the degree of a polynomial, and that $P_n(x)$ and $P_m(x)$ are both polynomials, that then:

$$\text{deg}(P_n(x)+Q_m(x))\leq\max(\text{deg}(P_n(x)),\text{deg}(Q_m(x))) \tag{6.1.55}$$

and:

$$\text{deg}(P_n(x)\cdot Q_m(x))=\text{deg}(P_n(x))+\text{deg}(Q_m(x)) \tag{6.1.56}$$

Definition (set theory version #169): Let k be a ring (see section Set Theory page 426) and $n\in\mathbb{N}^*$, the "polynomial ring" in n indeterminate (or variables) $k[X_1,\ldots,X_n]$ is constructed from an elementary polynomial, named "monomial" of the form:

$$\lambda X_1^{m_1}\ldots X_m^{m_n} \tag{6.1.57}$$

where $\lambda\in k$ is the "coefficient of the monomial", $m_1,m_2,\ldots,m_n$ are positive non-null integers and where $X_1^{m_1}\ldots X_m^{m_n}$ forms the "literal part of the monomial". Thus, by construction, a polynomial is a sum of a finite number of monomials named the "terms of the polynomial".

Therefore, the common special case used in small classes and presented at the beginning is $k[X]$, that is to say the ring of univariate polynomials with coefficients in k. Indeed most of time engineers and students deals with "ring of univariate polynomials with coefficient in $\mathbb{R}$" and denoted by $\mathbb{R}[X]$. Any element of $k[X]$ is therefore written as:

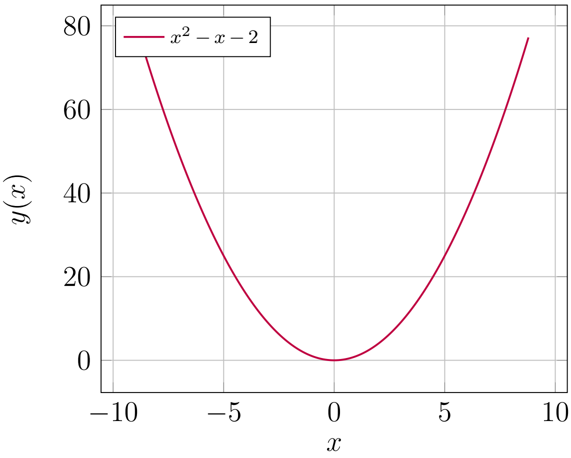
$$\sum_{i=0}^{n} a_i X^i \tag{6.1.58}$$

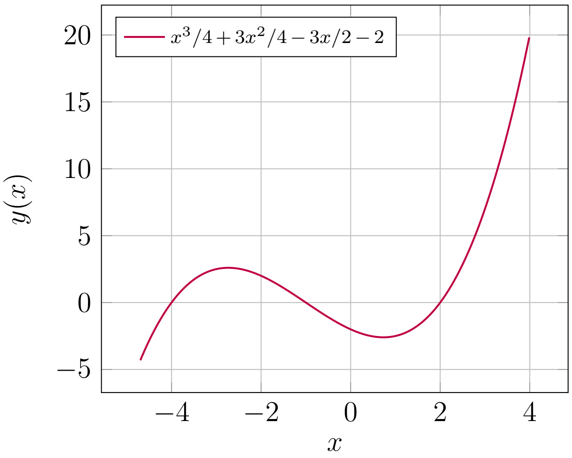
with $a_i \in k$ (most of time $a_i \in \mathbb{R}$), $i = 0 \ldots n$ and $n \in \mathbb{N}$.

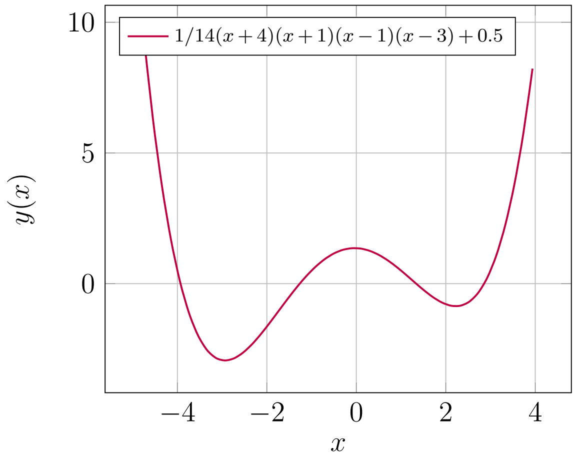
Remarks

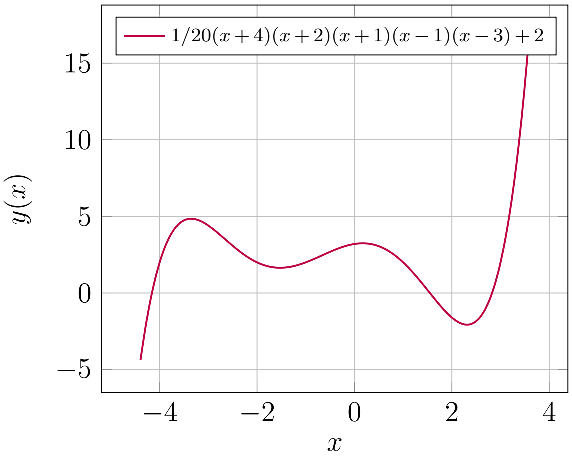
R1. Notice that the powers i are always positive or null in $k[X]$!!!

R2. We say that two polynomials are "similar" if they have the same literal part.









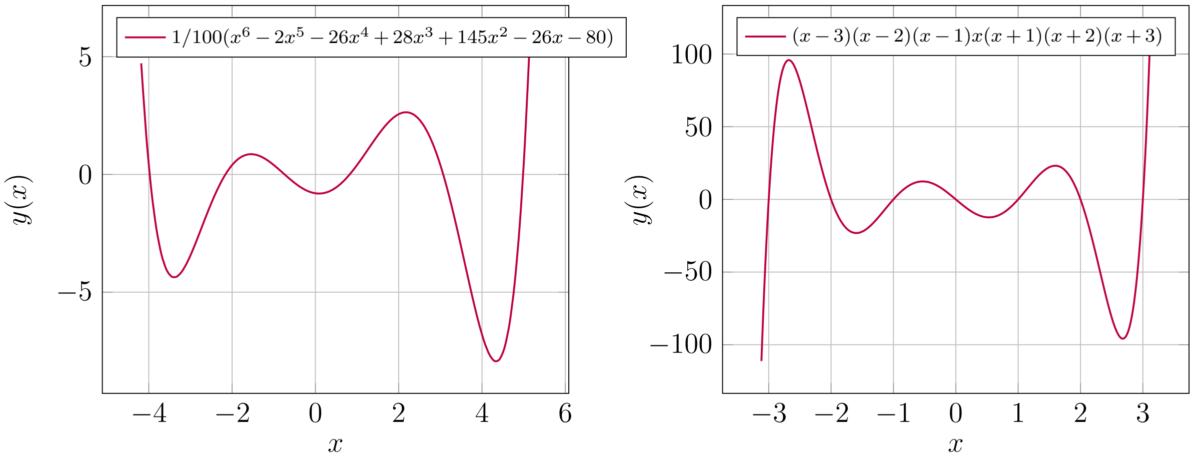


Figure 6.6: Some polynomials plotted with TikZ

The "limiting behaviour" of a function describes what happens to the function as $x \to \pm\infty$. The degree of a polynomial and the sign of its leading coefficient dictates its limiting behaviour. In particular:

- If the degree of a polynomial $f(x)$ is even and the leading coefficient is positive, then $f(x) \to +\infty$ as $x \to \pm\infty$.
- If $f(x)$ is an even degree polynomial with negative leading coefficient, then $f(x) \to -\infty$ as $x \to \pm\infty$.
- If $f(x)$ is an odd degree polynomial with positive leading coefficient, then $f(x) \to -\infty$ as $x \to +\infty$ and $f(x) \to +\infty$ as $x \to +\infty$.
- If $f(x)$ is an odd degree polynomial with negative leading coefficient, then $f(x) \to +\infty$ as $x \to -\infty$ and $f(x) \to -\infty$ as $x \to +\infty$.

These results are summarized in the table below:

Degree of the polynomial	**Leading coefficient**	
	$+$	$-$
Even	$f(x) \to +\infty$ as $x \to \pm\infty$	$f(x) \to -\infty$ as $x \to \pm\infty$
Odd	$f(x) \to -\infty$ as $x \to -\infty$ $f(x) \to -\infty$ as $x \to -\infty$	$f(x) \to +\infty$ as $x \to -\infty$ $f(x) \to -\infty$ as $x \to +\infty$

Table 6.4: Limiting behaviour of polynomials

Definition (#6.1.170): We name "turning point" a point at which the graph changes direction from increasing to decreasing or decreasing to increasing (or differently said: where the derivatives is equal to zero).

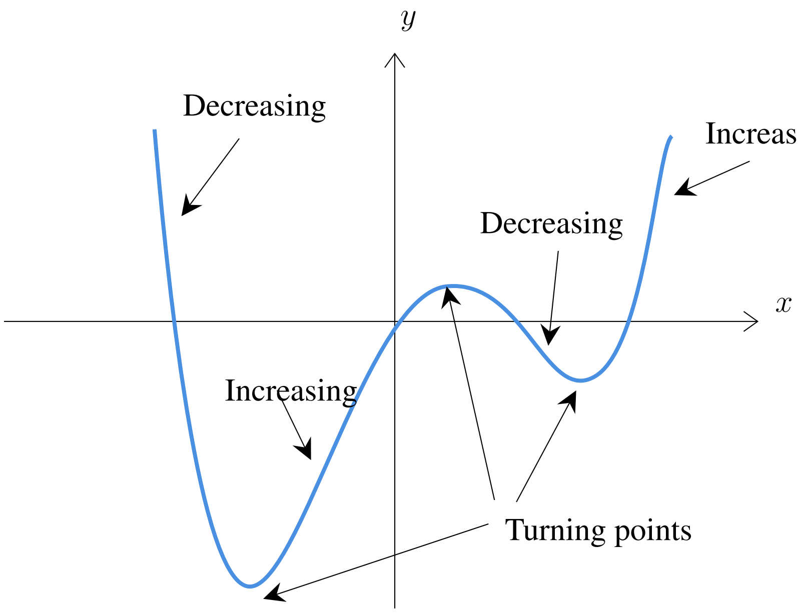


Figure 6.7: **Turning point explicitly highlighted**

The function f above is a 4^{th} degree polynomial and has 3 turning points. The maximum number of turning points of a polynomial function is always one less than the degree of the function.

Definition (#6.1.171): We name "root" or "zero of (univariate) polynomial", the x values such as the "polynomial equation" $P(x)=0$ is satisfied at the condition that at least one of the a_n with $n>0$ is not null.

If the polynomial is normalized[1] and admits one or more roots r_n we can then obviously factorize it as following (we will prove it more rigorously further below):

$$P(x)=(x-r_n)(x-r_{n-1})\ldots(x-r_2)(x-r_1) \qquad (6.1.59)$$

so that when x takes the value of one of the roots r_n, the expression above is zero. This is what we name by convention "factorize a polynomial".

If the polynomial is not normalized then the factorization will be obviously given by:

$$P(x)=a_n(x-r_n)(x-r_{n-1})\ldots(x-r_2)(x-r_1) \qquad (6.1.60)$$

Algebraic identities are particular forms of polynomial functions. Indeed, consider a constant c and a variable x and:

$$(c+x)^n=\sum_{k=0}^{n}C_k^n c^k x^{n-k}=C_0^n c^0 x^n+C_1^n c^1 x^{n-1}+\ldots+C_{n-1}^n c^{n-1}x^1+C_n^n c^n \qquad (6.1.61)$$

We see that if we put:

$$C_0^n c^0=a_n,\; C_1^n c^1=a_{n-1},\; \ldots,\; C_{n-1}^n c^{n-1}=a_1,\; C_n^n c^n=a_0 \qquad (6.1.62)$$

---

[1] Indeed, if you develop the given expression, you will see that the factor in front of the highest order term x^n is 1.

we fall back on:

$$(c+x)^n = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x^1 + a_0 \tag{6.1.63}$$

Definitions (#6.1.172):

D1. A polynomial in one indeterminate is named an "univariate polynomial", a polynomial in more than one indeterminate is named a "multivariate polynomial". A polynomial with two indeterminates is named a "bivariate polynomial".

A famous example in pure mathematics of a multivariate polynomial given many times in undergraduates course is:

$$x^n + y^n = z^n \tag{6.1.64}$$

D2. In the case of polynomials in more than one indeterminate, a polynomial is named "homogeneous of degree n if all its non-zero terms have degree n (the example just above is such a polynomial!).

6.1.3.1 General relations between roots and coefficients

Let us see now an interesting relation between the roots of a polynomial and its coefficients.

For this, let us consider the following polynomial:

$$P = a_n x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \ldots + a_2 x^2 + a_0 \tag{6.1.65}$$

And let us denote its roots by $r_1, r_2, \ldots, r_n$ (n root does not necessarily mean n distinct roots...). We then have:

$$\begin{aligned} P &= a_n x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \ldots + a_2 x^2 + a_0 \\ &= a_n(x-r_1)(x-r_2)\ldots(x-r_{n-1})(x-r_n) \end{aligned} \tag{6.1.66}$$

We then deduce by developing the last equality and by identifying with the coefficients that:

$$\begin{aligned} r_1 + r_2 + r_3 + \ldots + r_{n-1} + r_n &= -\frac{a_{n-1}}{a_n} \\ r_1 r_2 + r_1 r_3 + \ldots + \ldots + r_{n-1} r_n &= \frac{a_{n-2}}{a_n} \\ r_1 r_2 r_3 + r_1 r_2 r_4 + \ldots + r_{n-2} r_{n-1} r_n &= -\frac{a_{n-3}}{a_n} \\ &\ldots \\ r_1 r_2 r_3 r_4 r_5 r_6 r_7 r_8 \ldots r_{n-1} r_n &= (-1)^n \frac{a_0}{a_n} \end{aligned} \tag{6.1.67}$$

We can write it again differently by defining the numbers $\sigma_1, \sigma_2, \ldots, \sigma_n$ as following:

$$\boxed{\begin{aligned} &\sigma_1 = \sum_{i=1}^{n} r_i = r_1 + r_2 + r_3 + \ldots + r_{n-1} + r_n = -\frac{a_{n-1}}{a_n} \\ &\sigma_2 = \sum_{i<j} r_i r_j = r_1 r_2 + r_1 r_3 + \ldots + \ldots + r_{n-1} r_n = \frac{a_{n-2}}{a_n} \\ &\sigma_3 = \sum_{i<j<k} r_i r_j r_k = r_1 r_2 r_3 + r_1 r_2 r_4 + \ldots + r_{n-2} r_{n-1} r_n = -\frac{a_{n-3}}{a_n} \\ &\ldots \\ &\sigma_n = \prod_{i=1}^{n} r_i = (-1)^n \frac{a_0}{a_n} \end{aligned}} \tag{6.1.68}$$

Conversely if the $r_1, r_2, \ldots, r_n$ satisfies the preceding system, then they are the roots of a normalized polynomial of the form:

$$P = a_n\left(x^n - \sigma_1 x^{n-1} + \sigma_2 x^{n-2} + \ldots + (-1)^n \sigma_n\right) \tag{6.1.69}$$

Notice that for the famous case $n = 2$ that we found:

$$\begin{aligned} \sigma_1 &= \sum_{i=1}^{2} r_i = r_1 + r_2 = -\frac{a_1}{a_2} \\ \sigma_2 &= \sum_{i<j} r_i r_j = r_1 r_2 = \frac{a_0}{a_2} \end{aligned} \tag{6.1.70}$$

If we consider individuals that are identical apart from one covariate x, the hazard ratio becomes as we already know:

$$\frac{h_i}{h_j}=\frac{e^{\beta x_i}}{e^{\beta x_j}}=e^{\beta(x_i-x_j)} \tag{6.1.71}$$

The confidence interval of the hazard ratio is therefore:

$$\boxed{e^{\left(\hat{\beta}-Z_{\alpha/2}\text{SE}\left(\hat{\beta}\right)\right)(x_i-x_j)}\leq \text{H.R.}\leq e^{\left(\hat{\beta}+Z_{\alpha/2}\text{SE}\left(\hat{\beta}\right)\right)(x_i-x_j)}} \tag{6.1.72}$$

6.1.3.2 Multiplication of Polynomials

The product of the two polynomials P (of degree n) and Q (of degree m) is obviously given by:

$$(PQ)(x)\equiv P(x)Q(x)=\sum_{i=0}^{n}a_ix^i\sum_{j=0}^{m}b_jx^j=\sum_{i=0}^{n}\sum_{j=0}^{m}a_ib_jx^{i+j} \tag{6.1.73}$$

There is no much things more to say about the multiplication of polynomials for the needs of this textbook.

6.1.3.3 Euclidean Division of Polynomials

Let us put ourselves now in the ring $k[X]$. If $P(x)\in k[X]$, we denote by $\text{deg}(P)$ the degree of the polynomial $P(X)$ with coefficients in a ring k (real or complex ... whatever!)

Remark

By convention:

$$\text{deg}(0)=-\infty \tag{6.1.74}$$

Theorem 6.2. Given:

$$u(X)=\sum_{i=0}^{k}u_iX^i,v(X)=\sum_{i=0}^{m}v_iX^i\in k[X] \tag{6.1.75}$$

with $k,m>0$. Then there are two unique polynomial $q(X),r(X)\in k[X]$ such as:

$$\text{deg}(r)<\text{deg}(v)=m \tag{6.1.76}$$

and:

$$u(X)=q(X)v(X)+r(X) \tag{6.1.77}$$

where $q(X)$ is the "quotient polynomial" and $r(X)$ the "residual polynomial".

Proof 6.2.1. If $u(X)=0$ the result is obvious. Let us suppose that $u(X)\neq 0$ and let us prove the existence by induction on the degree k of $u(X)$.

If $k=0$ then $q(X)=0$ (since $m>0$) and therefore $r(X)=u(X)$ will do the job.

Now let us suppose the statement for any $k\leq n$... (this supposition is completely free of charge...):

Let $u(X)$ be of degree $k=n+1$. If $m>n+1$ then $q(X)=0$ and $r(X)=u(X)$ can also do the job.

Otherwise, if $m\leq n+1$ then by writing (u_{n+1} is the $n+1$-th coefficient of the polynomial $u(X)$ and v_m the m-th coefficient of $v(X)$:

$$u(X)-\frac{u_{n+1}}{v_m}X^{n+1-m}v(X) \tag{6.1.78}$$

we reduce then $u(X)$ to a polynomial of degree $\leq n$ since $v(X)$ is of degree m (and that it exists)!

Indeed, the term:

$$\frac{u_{n+1}}{v_m}X^{n+1-m}v(X) \tag{6.1.79}$$

removed (at least) the term of highest degree $u_{n+1}X^{n+1}$.

By the induction hypothesis, there are $f(X)$ and $g(X)$ such as:

$$u(X)-\frac{u_{n+1}}{v_m}X^{n+1-m}v(X)=f(X)v(X)+g(X) \tag{6.1.80}$$

with $\text{deg}(g)<m$. So after rearranging:

$$u(X)=\left(\frac{u_{n+1}}{v_m}X^{n+1-m}v(X)+f(X)\right)v(X)+g(X) \tag{6.1.81}$$

and therefore:

$$q(X)=\frac{u_{n+1}}{v_m}X^{n+1-m}v(X)+f(X) \qquad r(X)=g(X) \tag{6.1.82}$$

do the job!

So by induction we see that the Euclidean division exists in the polynomial ring $k[X]$.

■ Q.E.D.

Remark

This proof allowed us in the section of Set Theory to show that this ring is "principal".

✍**Example:**

We will see only one example as the idea is always same. We want to divide x^3+x^2 by $x-1$ we get:

$$
\begin{array}{rrrrr|l}
x^3 & +x^2 & & & \cdot(-1) & x \quad -1 \\ \cline{6-6}
-x^3 & +x^2 & & & & x^2 \ +2x \ +2 \\ \cline{1-2}
 & 2x^2 & & & & \\
 & -2x^2 & +2x & & & \\ \cline{2-3}
 & & 2x & & & \\
 & & -2x & +2 & & \\ \cline{3-4}
 & & & +1 & &
\end{array}
\tag{6.1.83}
$$

6.1.3.4 Factorization Theorem of Polynomials

We will now prove an important theorem that is in fact originally illustrated (among others) by the remarkable identities we saw above:

Theorem 6.3. If a polynomial $P(X)\in k[\mathbb{K}]$ with coefficients in k of degree $n\geq 1$ has a root $x=r$ in the ring k, then we can factorize $P(x)$ by $(x-r)$ such that:

$$P(x)=(x-r)Q(x) \tag{6.1.84}$$

where Q is a polynomial of degree $n-1$ (and therefore can be a simple monomial).

In other words, "factorize a polynomial", it is written it as a product of monomials (in the general case: of polynomial). When not only applied to polynomials or also simply to numbers, factorization is an operation that transforms a sum into a product!

Proof 6.3.1. The idea is to perform the Euclidean division of $P(x)$ by $(x-r)$. According to the previous theorem, there exists a pair (Q,R) of polynomials such as:

$$P(x)=(x-r)Q+R \tag{6.1.85}$$

and according to the result of the previous theorem on the Euclidean division:

$$\text{deg}(R)<\text{deg}(x-r) \tag{6.1.86}$$

But $\text{deg}(x-r)=1$, so $\text{deg}(R)=0$ (or $-\infty$ by convention). $R(x)$ is therefore a constant polynomial function. Moreover, by hypothesis, r is a root of $P(x)$. We have therefore:

$$0=P(r)=(r-r)Q(r)+R(r) \tag{6.1.87}$$

So $R(r)=0$. Therefore $R(x)$ is the zero polynomial and the theorem is practically proved. It remains to prove that $\text{deg}(Q)=n-1$, which is an immediate consequence of the relation:

$$\text{deg}(P(x))=\text{deg}(x-r)+\text{deg}(Q(x)) \tag{6.1.88}$$

Hence:

$$n=1+\text{deg}(Q) \tag{6.1.89}$$

■ Q.E.D.

From this property to factorize a polynomial, named sometimes "factorization theorem", we can give a foretaste of a much more important theorem:

Theorem 6.4. Let us show that if we have a polynomial function $P(X)\in k[X]$ of degree $n\in\mathbb{N}$ with coefficients in k, then it has at most a finite number n of roots (some being possibly confused) in k.

Proof 6.4.1. First, because $P(x)$ has a degree (order), $P(x)$ is not a zero polynomial function. Then, let us argue by the absurd:

If the function $P(x)$ has n roots with $p>n$ (more roots than degree....), by denoting these roots $r_1,\ldots,r_p$, we have, by the previous factorization theorem (applied p times):

$$P(x)=(x-r_1)(x-r_2)\ldots(x-r_p)Q(x) \tag{6.1.90}$$

where Q is a polynomial of degree:

$$n-p<0 \tag{6.1.91}$$

Now, since by definition a polynomial is a polynomial if and only if its degree (order) belongs to $\mathbb{N}$, the polynomial Q must be the zero polynomial such that:

$$Q(x)=0 \tag{6.1.92}$$

It follows that:

$$P(x)=0 \tag{6.1.93}$$

This contradicts the initial hypothesis that P is not the zero polynomial, hence:

$$p\leq n \tag{6.1.94}$$

■ Q.E.D.

6.1.3.5 Diophantine equation

If we generalize the concept of univariate polynomial with several variables such as:

$$P=P(x,y,z) \tag{6.1.95}$$

then we name "Diophantine equation" an equation of the form:

$$\boxed{P(x,y,z)=0} \tag{6.1.96}$$

where P is a polynomial with integer (or rational) coefficients for which seeks the radicals (roots) strictly in $\mathbb{N}$ or $\mathbb{Q}$. Conventional Diophantine equations are for examples:

- The linear Diophantine equations:

$$ax+by+c^{te}=0 \tag{6.1.97}$$

- The Pythagorean triples:

$$x^n+y^n=z^n \tag{6.1.98}$$

For the general proof of the latter, the reader will have to wait a little bit the time for the redactors of this textbook to have the time to understand the traditional proof to be able to simplified it (...).

6.1.3.6 First order univariate Polynomial and Equations

Given the linear function:

$$P_1(x) = ax + b = a\left(x + \frac{b}{a}\right) \tag{6.1.99}$$

If $a \neq 0$ then the first the equation:

$$a\left(x + \frac{b}{a}\right) = 0 \tag{6.1.100}$$

has a simple closed formed root given obviously by:

$$r = -\frac{b}{a} \tag{6.1.101}$$

such that $P_1(r) = 0$.

If $b = 0$ this polynomial is named an "affine function".

Remarks

R1. If the coefficients of the univariate polynomial of degree 1 are all such that $a, b \in \mathbb{R}$ then the root also belongs to $\mathbb{R}$.

R2. If one of coefficients of the univariate polynomial of degree 1 belongs to $\mathbb{C}$ then the root also belongs to $\mathbb{C}$.

R3. If the both coefficients of the univariate polynomial of degree 1 belongs to $\mathbb{C}$ then the root also belongs to $\mathbb{C}$ or $\mathbb{R}$.

R4. We say that two polynomial equations are "equivalent" if the admit the same solutions.

Here are also some properties for univariate first order polynomials that we give without proof as they seem very very intuitive to us (except on reader request):

P1. If we add (or respectively subtract) a same number to each member of the equation (left from the "$=$" sign, and also right), we get an equation that has the same solutions as the original equation (and this whatever is its degree).

P2. If we multiply (or respectively divide) each me member of an equation (left from the "$=$" sign, and also right) by a same non-null number, we get an equation that has the same solutions as the original equation (and this whatever is its degree).

The method should be general enough to by applied to all equations of the same kind, be build on the four arithmetic basis operations (addition, subtraction, multiplication and division) and the extraction of roots. We can find the solutions (roots), of an equation thanks to its coefficients, using only the previous operations (that is to say in a "closed form"), we then say that the equation can be solved by "radicals".

6.1.3.6.1 Solving by circular reference

Many executives in the small, medium and international companies no longer know how to solve linear equations of the first degree (and beyond also of course...). So I think i the be interesting to the student of see how the hug majority of the people that are not able to solve such high-school level problems proceed (it's especially interesting when you know that they are paid quite a lot to solve such "complicated" problem according to their board committee) The easiest way to see this is how I see they use the spreadsheet software Microsoft® Excel®. Indeed, consider the following example:

A company makes a turnover (income) of $1,500.-$ and has to pay $1,000.-$ of fixed costs (costs). It has 40% tax payable after deduction. The company would like to make a donation (tax deductible) such that the amount of this donation represents 10% of taxes after deduction.

This is written mathematically simply if we denote by R the turnover, C the fixed costs, D the donation, t the tax rate:

$$D = 10\% \cdot (R - C - D)(1 - t) \tag{6.1.102}$$

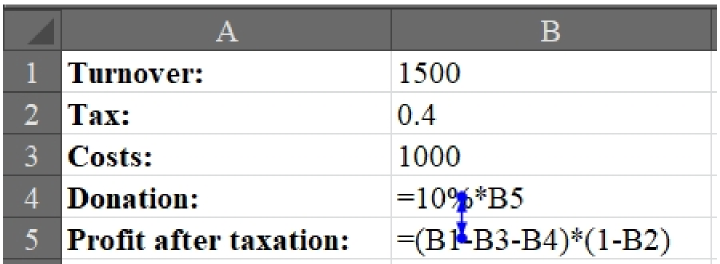
After some small trivial rearrangements we get:

$$D = \frac{10\%(R - C)(1 - t)}{1 + 10\%(1 - t)} = \frac{10\%(1500 - 1000)}{1 + 10\%(1 - 40\%)} \cong 28.301 \tag{6.1.103}$$

This corresponds well to the 10% tax after deduction since:

$$(R - C - D)(1 - t) \cong 283.01.- \tag{6.1.104}$$

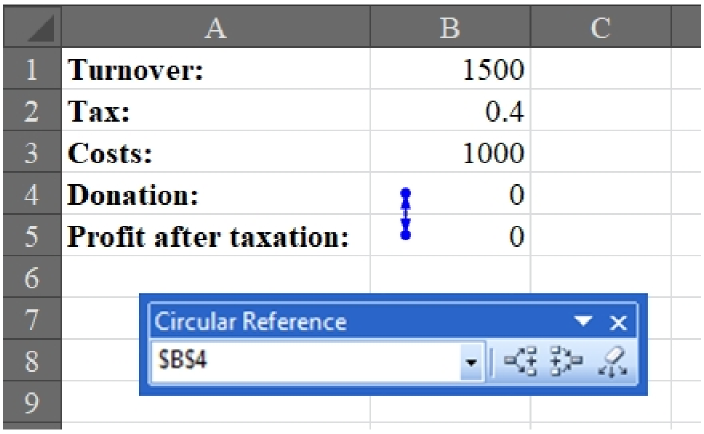
And here's how (bad) executives using a spreadsheet like Microsoft® Excel® and "solve" this (you can trust me I really see people do this in corporation because they are not able to solve problem on the paper anymore). First they write in the spreadsheet software (which automatically detects the inconsistency of the user by indicating it with a blue arrow to highlight the "circular reference"):



	A	B
1	**Turnover:**	1500
2	**Tax:**	0.4
3	**Costs:**	1000
4	**Donation:**	=10%*B5
5	**Profit after taxation:**	=(B1-B3-B4)*(1-B2)

Figure 6.8: **Solving simple equation by circular reference in Microsoft® Excel®**

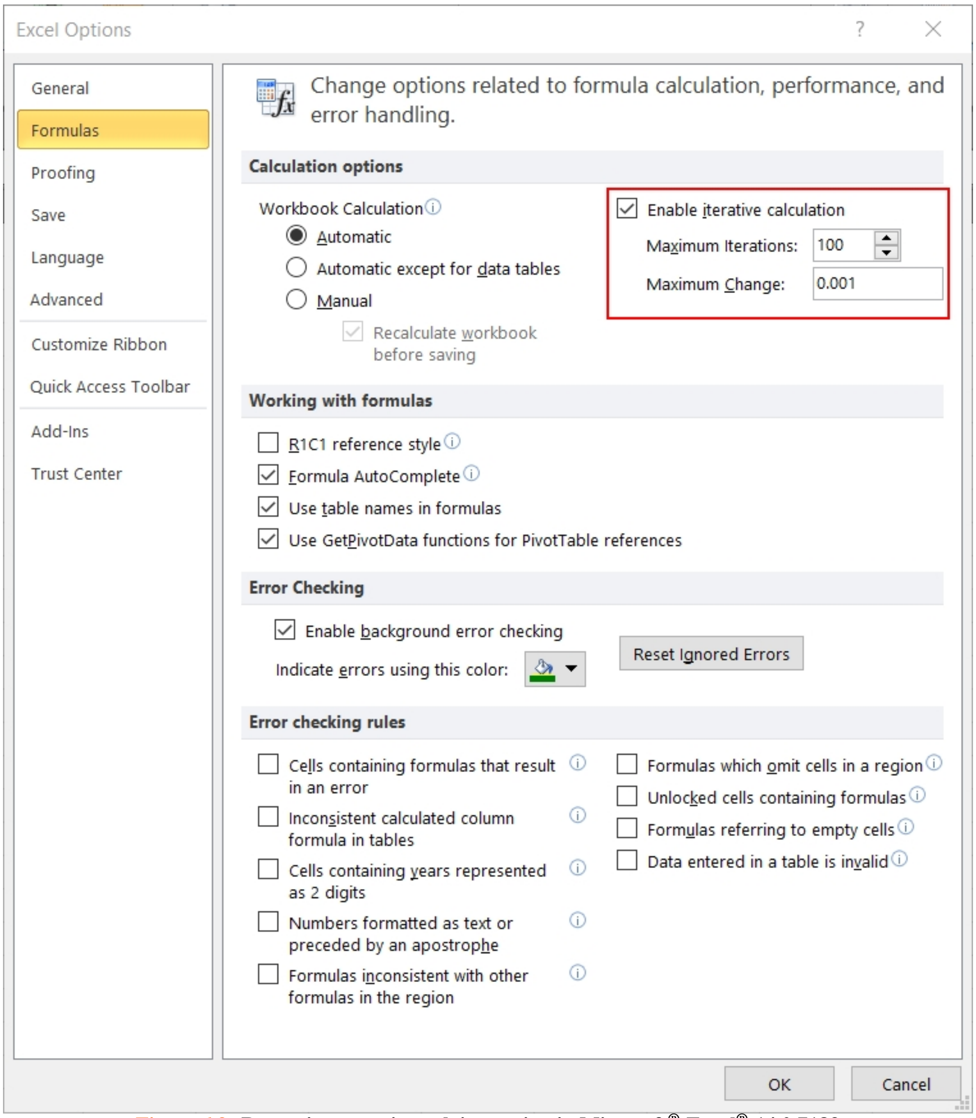
Validated this gives:



	A	B	C
1	**Turnover:**	1500	
2	**Tax:**	0.4	
3	**Costs:**	1000	
4	**Donation:**	0	
5	**Profit after taxation:**	0	
6			
7			
8			
9			

Circular Reference
$B$4

To solve this equation, it is necessary to activate in the software options the authorization to use the iteration (the position of this option depends on the version of the spreadsheet).



Excel Options
?
General
Formulas
Proofing
Save
Language
Advanced
Customize Ribbon
Quick Access Toolbar
Add-Ins
Trust Center

Change options related to formula calculation, performance, and error handling.

Calculation options

Workbook Calculation
● Automatic
○ Automatic except for data tables
○ Manual
☑ Recalculate workbook before saving

☑ Enable iterative calculation
Maximum Iterations: 100
Maximum Change: 0.001

Working with formulas

☐ R1C1 reference style
☑ Formula AutoComplete
☑ Use table names in formulas
☑ Use GetPivotData functions for PivotTable references

Error Checking

☑ Enable background error checking
Indicate errors using this color:

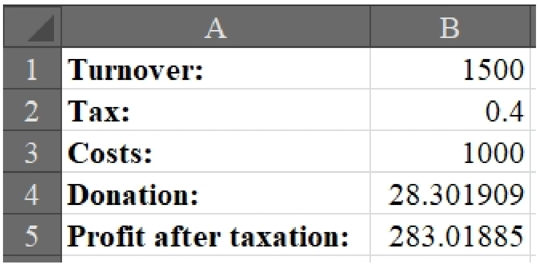
Reset Ignored Errors

Error checking rules

☐ Cells containing formulas that result in an error
☐ Inconsistent calculated column formula in tables
☐ Cells containing years represented as 2 digits
☐ Numbers formatted as text or preceded by an apostrophe
☐ Formulas inconsistent with other formulas in the region
☐ Formulas which omit cells in a region
☐ Unlocked cells containing formulas
☐ Formulas referring to empty cells
☐ Data entered in a table is invalid

Figure 6.9: **Recurring equation solving option in Microsoft® Excel® 14.0.7183**

Which then gives:



	A	B
1	**Turnover:**	1500
2	**Tax:**	0.4
3	**Costs:**	1000
4	**Donation:**	28.301909
5	**Profit after taxation:**	283.01885

So as we can see, this is quite a non-efficient way to solve a simple high-school problem (a shame for managers or business analysts that have MBA...!).

6.1.3.7 Second order univariate Polynomial and Equations

Given the following univariate polynomial with coefficient in $\mathbb{R}$ (trinomial of second degree):

$$P_2(x)=ax^2+bx+c=a\left[\left(x+\frac{b}{2a}\right)^2-\frac{b^2-4ac}{4a^2}\right] \tag{6.1.105}$$

If we represent this univariate polynomial on the plane, this give us:

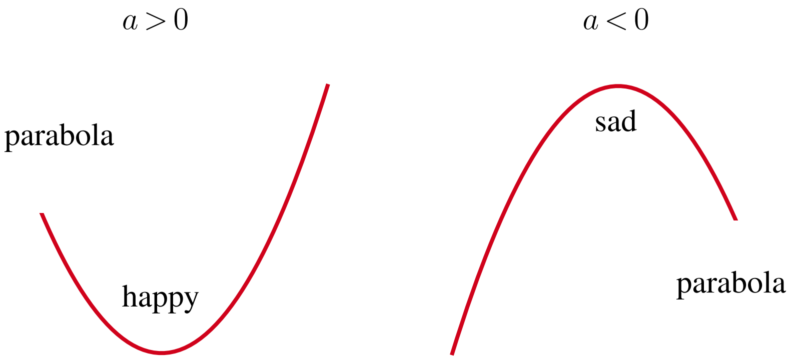
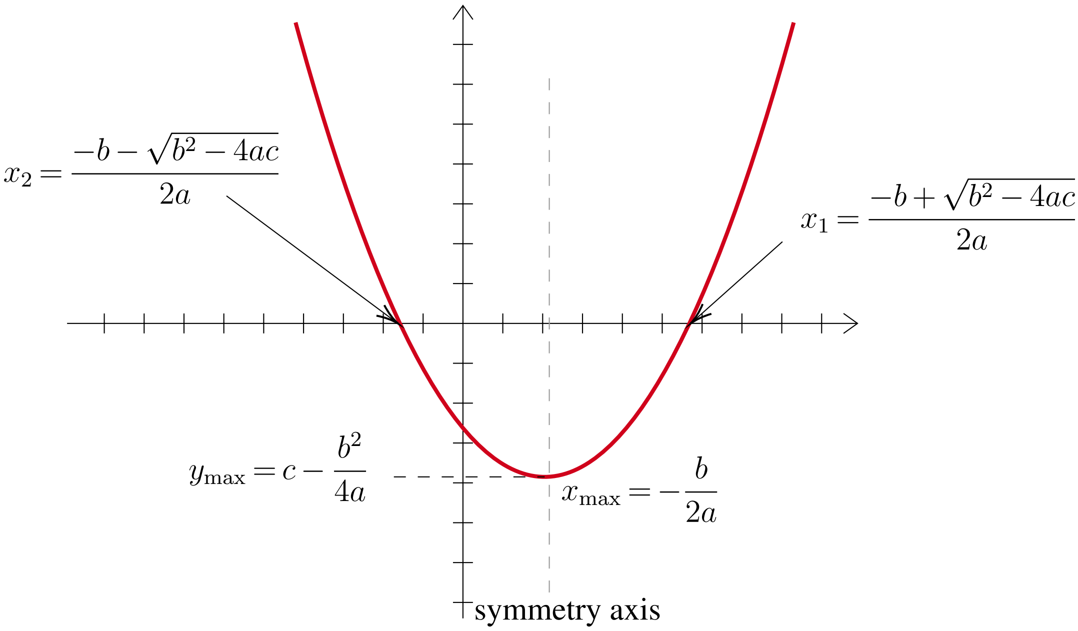


Figure 6.10: **Typical orientation for second degree polynomials**

If we take the derivative of this function (see section Differential and Integral Calculus page 1450) and we search in what point the derivative is equal to zero, we will always found the optimum on the inflection point (see section Differential and Integral Calculus page 1454) of the parabola (which corresponds also to its symmetry axis):



$x_2 = \frac{-b-\sqrt{b^2-4ac}}{2a}$
$x_1 = \frac{-b+\sqrt{b^2-4ac}}{2a}$
$y_{\max} = c - \frac{b^2}{4a}$
$x_{\max} = -\frac{b}{2a}$
symmetry axis

Figure 6.11: **Inflection point of the tangent?**

If $a \neq 0$, then we have:

$$\left(x+\frac{b}{2a}\right)=\frac{b^2-4ac}{4a^2} \Rightarrow x+\frac{b}{2a}=\pm\frac{\sqrt{b^2-4ac}}{2a} \Rightarrow x=\frac{-b\pm\sqrt{b^2-4ac}}{2a} \qquad (6.1.106)$$

We then have a "double root" (or "root of multiplicity 2") that we denote by:

$$r_{1,2}=\frac{-b\pm\sqrt{b^2-4ac}}{2a} \qquad (6.1.107)$$

such that $P(r_{1,2})=0$ and we define a new term named sometimes the "determinant of the polynomial" and most commonly the "discriminant of the polynomial":

$$\Delta = b^2-4ac \qquad (6.1.108)$$

Finally:

$$\boxed{r_{1,2}=\frac{-b\pm\sqrt{b^2-4ac}}{2a}=\frac{-b\pm\sqrt{\Delta}}{2a}} \qquad (6.1.109)$$

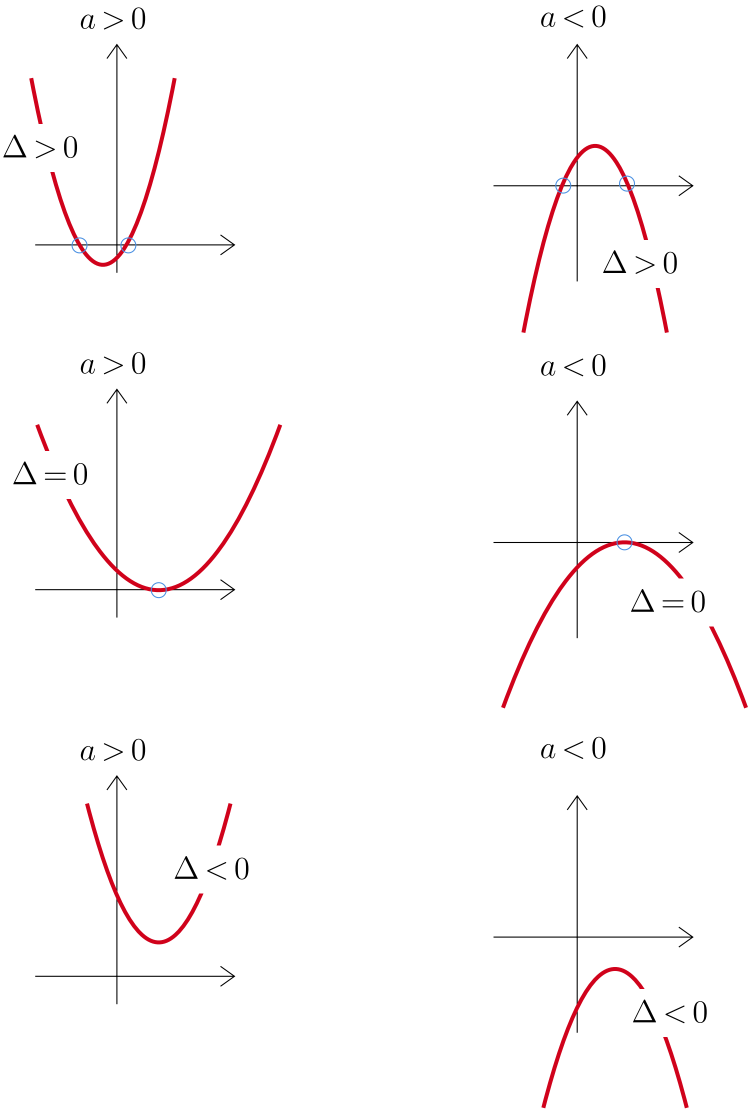
If the second order univariate polynomial on x has two root, we can then factorize it in an irreducible form (following the factorization theorem proved earlier) in the following form:

$$P(x)=a(x-r_1)(x-r_2) \qquad (6.1.110)$$

We also prove easily from the expression of the root by doing simple algebra the "Vieta relations" (on request of the readers we can detail the developments in necessary):

$$r_1+r_2=-\frac{b}{a} \qquad r_1r_2=\frac{c}{a} \qquad (6.1.111)$$

Depending on the sign of $2a$ and this of the discriminant Δ, we have:


$a > 0$
$\Delta > 0$
$a < 0$
$\Delta > 0$
$a > 0$
$\Delta = 0$
$a < 0$
$\Delta = 0$
$a > 0$
$\Delta < 0$
$a < 0$
$\Delta < 0$

Or with some real polynomials:

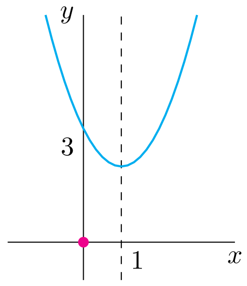
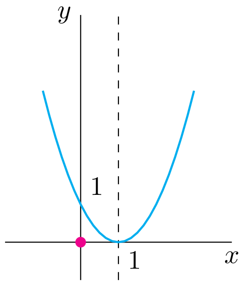
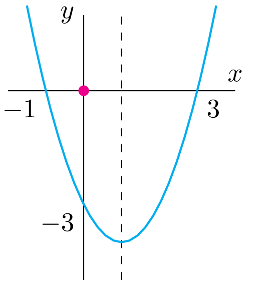
$y = x^2 - 2x + 3$	$y = x^2 - 2x + 1$	$y = x^2 - 2x - 3$
		
$\Delta = b^2 - 4ac$ $= (-2)^2 - 4 \cdot 1 \cdot 3$ $= -8$	$\Delta = b^2 - 4ac$ $= (-2)^2 - 4 \cdot 1 \cdot 1$ $= 0$	$\Delta = b^2 - 4ac$ $= (-2)^2 - 4 \cdot 1 \cdot (-3)$ $= 16$
$\Delta < 0$	$\Delta = 0$	$\Delta > 0$

Figure 6.12: Some real example of second degree polynomials (author: Lê Huy Tien)

Therefore:

- If $\Delta < 0$ our polynomial has no real roots and cannot be factorized in a multiplication of monomial with real ($\mathbb{R}$) factor but with complex one ($\mathbb{C}$). Therefore (it is recommended to have read first the part about Complex Numbers page 243 in the section Numbers of this textbook):
$$r_{1,2} = \frac{-b \pm \sqrt{(-1)(b^2-4ac)}}{2a} = \frac{-b \pm \sqrt{\mathrm{i}^2(b^2-4ac)}}{2a} = \frac{-b \pm \mathrm{i}\sqrt{(b^2-4ac)}}{2a} = 0 \tag{6.1.112}$$
and we know that we can write any complex number in a condensed form (Euler formula) and as the complex roots of a polynomial of the second degree are conjugate (we already know this jargon) we have:
$$r_{1,2} = re^{\pm(\varphi+2k\pi)} \tag{6.1.113}$$
where (recall) r is the module of the complex roots (module that is equal for the both) and φ the argument of the complex roots (equal in absolute value).
- If $\Delta = 0$ the polynomial equation then has a single solution that is obviously:
$$r_{1,2} = \frac{-b \pm \sqrt{0}}{2a} = \frac{-b}{2a} \tag{6.1.114}$$
- If $\Delta > 0$ the polynomial equation then has two solutions defined by the general relations which we have already given above:
$$r_{1,2} = \frac{-b \pm \sqrt{b^2-4ac}}{2a} \tag{6.1.115}$$

About the complex case, let us take as an example the following quadratic polynomial:
$$x^2 + 1 = 0 \tag{6.1.116}$$
that admits only two complex roots that are i and $-$i. In the real plane this polynomial will be represented in Maple™ 4.00b by:

```
>plot(x^2+1,x=-5..5);
```

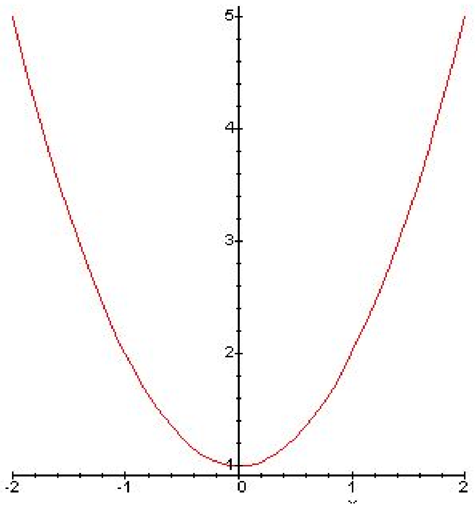


Figure 6.13: **Plot example of a polynomial of degree 2 which admits only complex solutions**

where we see well that there is no real solutions (zeros). While placing us in the complex $\mathbb{C}$, we have:

```
>plot3d(abs((re+I*im)^2+1),re=-2..2,im=0..2,view=[-2..2,-2..2,0..2],
orientation=[-130,70],contours=50,style=PATCHCONTOUR,axes=frame,
grid=[100,100],numpoints=10000);
```

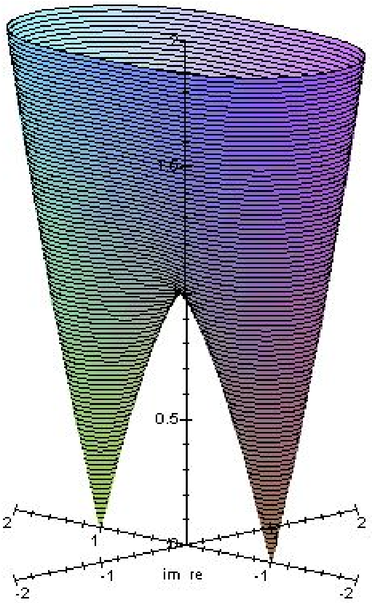


Figure 6.14: **The same polynomial but playing with the complex representation**

where the two zeros are visible on the imaginary axis at -1 and $+1$. Obviously when it's the first time we see a function shown in a figure taking into account the complex values we try to find where is the corresponding parabola of the purely real case. To do this, we simply cut the surface above on two on the imaginary axis and we get then:

```
>plot3d(abs((re+I*im)^2+1),re=-2..2,im=0..2,view=[-2..2,-2..2,0..2],
orientation=[-130,70],contours=50,style=PATCHCONTOUR,axes=frame,
grid=[100,100],numpoints=10000);
```

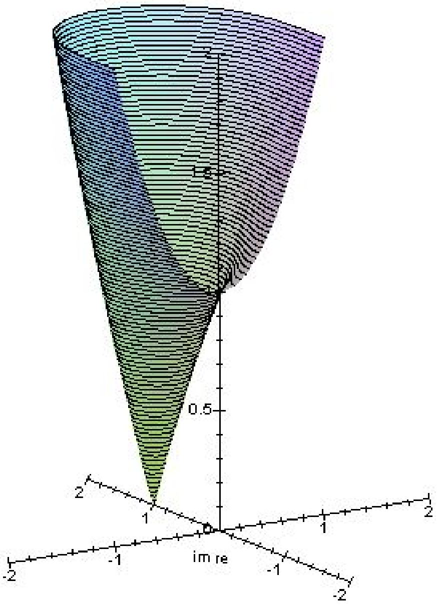


Figure 6.15: **A little zoom on the same polynomial**

where we find again clearly our parabola visible on the cutted surface. So we can ask ourselves whether the complex numbers are a natural extension of our conventional space beyond our physical senses and our common measuring devices...!

6.1.3.7.1 Irrational Equations

The practitioner must always take the habit to verify the solution in the original equation to be sure of the validation of the definition domain of the function. Indeed, there are solutions to the resolution of equations that do not satisfy the original equation and this is what we name "extraneous solutions" and this is typically the case of irrational equations.

Definition (#6.1.173): An "irrational equation" is an equation where the unknown is under a radical (that is to say in typical case: under a square root).

✍**Example:**

Consider the following equation:

$$\sqrt{2+x}+4-\sqrt{10-3x}=0 \tag{6.1.117}$$

To solve it first we can redistribute:

$$\sqrt{2+x}+4=\sqrt{10-3x} \tag{6.1.118}$$

We take the power of 2:

$$2+x+8\sqrt{2+x}+16=10-3x \tag{6.1.119}$$

We simplify little bit:

$$8\sqrt{2+x}=-8-4x \tag{6.1.120}$$

again:

$$2\sqrt{2+x}=-2-x \tag{6.1.121}$$

We take the power of 2 again:

$$4(2+x)=4+4x+x^2 \tag{6.1.122}$$

We simplify a last time:

$$x^2=4 \tag{6.1.123}$$

We get two trivial solutions that are $x_1=-2$ and $x_2=2$. But only the solution $x_1=-2$ satisfy the proposed equation. Indeed, if you put the first solution in the original equation we get:

$$\sqrt{2+(-2)}+4-\sqrt{10-3\cdot(-2)}=\sqrt{0}+4-\sqrt{16}=4-4=0 \tag{6.1.124}$$

But if we put the second one:

$$\sqrt{2+(2)}+4-\sqrt{10-3\cdot(2)}=\sqrt{4}+4-\sqrt{4}=2+4-2=4\neq 0 \tag{6.1.125}$$

6.1.3.7.2 Golden Ratio

There is a second order univariate polynomial whose solution is famous around the world. The solution is a value named the "Golden ratio" or "Divine proportion" (...) and is found in architectural, aesthetic or in phyllotaxis (that is to say, in the arrangement of the leaves around the stem of the plants).

This number is:

$$\boxed{\varphi=\frac{1+\sqrt{5}}{2}=1.6180339887\ldots} \tag{6.1.126}$$

The golden ratio is also an algebraic number (any complex number that is a root of a non-zero polynomial in one variable with rational coefficients) and even an algebraic integer (complex number that is a root of some monic polynomial with coefficients in $\mathbb{Z}$) as it is solution of:

$$x^2-x-1=0 \tag{6.1.127}$$

Historically, the reader should know that two quantities a and b are said to be "in the Golden ratio φ" if their ratio is the same as the ratio of their sum to the larger of the two quantities:

$$\frac{a+b}{a}=\frac{a}{b}=\varphi \tag{6.1.128}$$

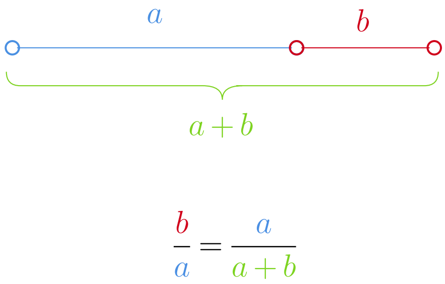


Figure 6.16: **Golden ratio illustration**

One method for finding the value of φ is to start with the left fraction. Through simplifying the fraction and substituting in $b/a=1/\varphi$:

$$\frac{a+b}{a}=1+\frac{b}{a}=1+\frac{1}{\varphi} \tag{6.1.129}$$

Therefore:

$$1+\frac{1}{\varphi}=\varphi \tag{6.1.130}$$

Multiplying by φ gives:

$$\varphi+1=\varphi^2 \tag{6.1.131}$$

which can be rearranged to:

$$\varphi^2-\varphi-1=0 \tag{6.1.132}$$

Using the quadratic formula, two solutions are obtained:

$$\varphi_1=\frac{1+\sqrt{5}}{2}=1.6180339887\ldots \qquad \varphi_2=\frac{1-\sqrt{5}}{2}=-0.6180339887\ldots \tag{6.1.133}$$

Because φ is the ratio between positive quantities φ is necessarily positive:

$$\varphi=\frac{1+\sqrt{5}}{2}=1.6180339887\ldots \tag{6.1.134}$$

Now let us introduce a useful relation for when we will study the ordinary generating function (see page 1783) relatively to the Fibonacci sequence. Let us recall that we started from:

$$1+\frac{1}{\varphi}=\varphi \tag{6.1.135}$$

we then have:

$$-\frac{1}{\varphi}=1-\varphi=1-\frac{1-\sqrt{5}}{2}=\varphi_2 \tag{6.1.136}$$

This is why $-1/\varphi$ is named the "Golden ratio conjugate"!

6.1.3.8 Third order univariate Polynomial and Equations

Even if it is rare to solve such a thing in theoretical physics or in engineering, solving an univariate polynomial of the 3rd degree is quite recreational and shows a good example of an already mature mathematical reasoning (we have these developments thanks Scipione del Ferro and Jerome Cardan mathematicians of the 116th century...).

Given the equation:

$$ax^3+bx^2+cx+d=0 \tag{6.1.137}$$

with the coefficients in $\mathbb{R}$ (to begin...). In a first time, the reader will be able to see that the reasoning that we have applied for the polynomial a smaller order than 3 stuck when the order is greater (excepted for special simple cases obviously...).

We will avoid the problem by using a change of variables subtle but quite justified.

Thus, nothing prevents us from putting that:

$$\exists X \Rightarrow x=X-\frac{b}{3a} \tag{6.1.138}$$

and that by dividing the polynomial of degree 3 by a to write:

$$\begin{aligned}
&x^3+\frac{b}{a}x^2+\frac{c}{a}x+\frac{d}{a}=0\\
&\left(X-\frac{b}{3a}\right)^3+\frac{b}{a}\left(X-\frac{b}{3a}\right)^2+\frac{c}{a}\left(X-\frac{b}{3a}\right)+\frac{d}{a}=0\\
&\left(X^3-\cancel{3}X^2\frac{b}{\cancel{3}a}+\cancel{3}X\frac{b^2}{3\cancel{3}a^2}-\frac{b^3}{27a^3}\right)+\frac{b}{a}\left(X^2-2X\frac{b}{3a}+\frac{b^2}{9a^2}\right)+\frac{c}{a}\left(X-\frac{b}{3a}\right)+\frac{d}{a}=0
\end{aligned} \tag{6.1.139}$$

By grouping the terms of the same order:

$$X^3+X^2\left(-\frac{b}{a}+\frac{b}{a}\right)+X\left(\frac{b^2}{3a^2}-2\frac{b^2}{3a^2}+\frac{c}{a}\right)+\left(-\frac{b^3}{27a^3}+\frac{b^3}{9a^3}-\frac{cb}{3a^2}+\frac{d}{a}\right)=0$$
$$X^3+X\left(\frac{c}{a}-\frac{b^2}{3a^2}\right)+\left(\frac{2b^3}{27a^3}-\frac{cb}{3a^2}+\frac{d}{a}\right)=0 \tag{6.1.140}$$

and let us write (nothing, but really nothing prevent us to do this):

$$\begin{cases}(1) \quad x=X-\dfrac{b}{3a}\\ (2) \quad p=\dfrac{c}{a}-\dfrac{b^2}{3a^2}\\ (3) \quad q=\dfrac{2b^3}{27a^3}-\dfrac{cb}{3a^2}+\dfrac{d}{a}\end{cases} \tag{6.1.141}$$

where (1) is known if and only if X is known and where p,q are anyway unknowns.

The polynomial[2]:

$$P=X^3+pX+q=0 \tag{6.1.142}$$

being of odd degree, it admits (as it can be seen from any visual plot of such a polynomial with real coefficients) at least one real root, named the "certain root" and at maximum three roots. The reader will easily check by himself that with a graphical representation of an odd degree polynomial this is trivial!

The left hand side of this equation is a "monic trinomial" called a "depressed cubic", because the quadratic term has coefficient 0.

Now let us make another subtle change of variable (we have the right to do this):

$$X=u+v \tag{6.1.143}$$

by imposing the condition that u,v must be such that $3uv=-p$ (nothing prevents us from imposing such a constraint), then we have:

$$P=X^3+pX+q=(u+v)^3-3uv(u+v)+q=0\Rightarrow u^3+v^3=-q \tag{6.1.144}$$

Therefore we have:

$$\begin{cases}(1') \quad u^3+v^3=-q\\ (2') \quad u^3v^3=\dfrac{-p^3}{27}\Leftrightarrow 3uv=-p\end{cases} \tag{6.1.145}$$

We can very well make an analogy between the two equations (1') and (2') and the Vieta relations we had obtained for the polynomial of degree 2 which we recall were:

$$r_1+r_2=-\frac{b}{a} \quad \text{and} \quad r_1r_2=\frac{c}{a} \tag{6.1.146}$$

---

[2] The first time I had to solve such a polynomial was to calculate the nutation of a gyroscope and the second time was for the calculation of the horizon of a Black Hole based on the Schwarzschild metric with cosmological constant that was in natural units: $\dfrac{\Lambda}{3}r^3-r-2M=0$

excepted that we have now (we adopt another notation for these intermediate roots):

$$z_1 = u^3 \qquad z_2 = v^3 \tag{6.1.147}$$

which gives us for the polynomial P by imposing (always by analogy) $a = 1$ a new equation:

$$Z^2 + qZ - \frac{p^3}{27} = 0 \tag{6.1.148}$$

for which z_1, z_2 are the roots. The latter equation has for discriminant:

$$\Delta = q^2 + 4\frac{p^3}{27} \tag{6.1.149}$$

Let us take now scenario by scenario:

1. If $\Delta > 0$, the equation on Z admits two solutions z_1, z_2 whose sum will give us indirectly the value of X since by definition $X = u + v$ and $z_1 = u^3$ and $z_2 = v^3$. We see that we have all the ingredients to find the first root of the original equation that will be the certain root (or "certain zero"). So:

   $$X_1 = u + v = \sqrt[3]{z_1} + \sqrt[3]{z_2} = \sqrt[3]{-\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}} + \sqrt[3]{-\frac{q}{2} - \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}} \tag{6.1.150}$$

   as $\Delta > 0$ and the superior roots are cubic we necessarily have $X_1 \in \mathbb{R}$ if all coefficient of the equation are well in $\mathbb{R}$.

2. If $\Delta = 0$, we know it, the equation on Z admit a double root and as the discriminant has a square power of q this means necessarily that p is negative.

   The polynomial P therefore also has a double root and the same for the original equation. We saw also that for a second degree polynomial if the discriminant is zero roots are:

   $$r_{1,2} = -\frac{b}{2a} \tag{6.1.151}$$

   Then by analogy:

   $$X_1 = X_2 = u + v = \sqrt[3]{z_1} + \sqrt[3]{z_2} = 2\sqrt[3]{z_{1,2}} = 2\sqrt[3]{\frac{-q}{2}} \tag{6.1.152}$$

3. If $\Delta < 0$ we must again use complex numbers as we did in our study of the polynomial of degree 2. Thus, we know that the equation Z admits two complex solutions such as:

   $$z_{1,2} = -\frac{q}{2} \pm \mathrm{i}\sqrt{-\frac{q^2}{4} - \frac{p^3}{27}} \tag{6.1.153}$$

   and once again as the roots are conjugated we can write in condensed form:

   $$z_{1,2} = re^{\pm\mathrm{i}(\varphi + 2k\pi)} \tag{6.1.154}$$

   and as:

   $$X = u + v = \sqrt[3]{z_1} + \sqrt[3]{z_2} \tag{6.1.155}$$

we therefore have:

$$\begin{aligned}X_k&=u_k+v_k=u_k+\bar{u}_k=\sqrt[3]{re^{\mathrm{i}(\varphi+2k\pi)}}+\sqrt[3]{re^{-\mathrm{i}(\varphi+2k\pi)}}\\&=\sqrt[3]{r}e^{\frac{\mathrm{i}(\varphi+2k\pi)}{3}}+\sqrt[3]{r}e^{-\frac{\mathrm{i}(\varphi+2k\pi)}{3}}\end{aligned}\tag{6.1.156}$$

As u_k,v_k are conjugated, we have necessarily $X_k\in\mathbb{R}$.

✍**Example:**

Consider the following equation:

$$x^3-15x-4=0\tag{6.1.157}$$

We therefore have:

$$\begin{aligned}x&=X-\frac{b}{3a}=X\\p&=\frac{c}{a}-\frac{b^2}{3a^2}=\frac{c}{a}=-15\\q&=\frac{2b^3}{27a^3}-\frac{cb}{3a^2}+\frac{d}{a}=-4\end{aligned}\tag{6.1.158}$$

and therefore:

$$\Delta=q^2+4\frac{p^3}{27}=16+4\frac{-3375}{27}=-484<0\tag{6.1.159}$$

We therefore have:

$$\begin{aligned}X&=\sqrt[3]{-\frac{q}{2}+\sqrt{\frac{q^2}{4}+\frac{p^3}{27}}}+\sqrt[3]{-\frac{q}{2}-\sqrt{\frac{q^2}{4}+\frac{p^3}{27}}}\\&=\sqrt[3]{2+\sqrt{\frac{15}{4}+\frac{-3375}{27}}}+\sqrt[3]{2-\sqrt{\frac{15}{4}+\frac{-3375}{27}}}\\&=\sqrt[3]{2+\sqrt{-121.25}}+\sqrt[3]{2-\sqrt{-121.25}}=\sqrt[3]{2+\mathrm{i}\sqrt{121.25}}+\sqrt[3]{2-\mathrm{i}\sqrt{121.25}}=4\end{aligned}\tag{6.1.160}$$

The polynomials of degree three are therefore well solvable by radicals.

Notice also another interesting approach due to Viète! We start again from:

$$X^3+pX+q=0\tag{6.1.161}$$

and let us set $X=u\cos(\theta)$. The idea is to choose u to make the previous equation coincide with the following trigonometric identity:

$$4\cos^3(\theta)-3\cos(\theta)-\cos(3\theta)=0\tag{6.1.162}$$

Before continuing let us prove this identity using the trigonometric identities proved in the section of Trigonometry:

$$\begin{aligned}\cos(3\theta) &= \cos(2\theta+\theta) = \cos(2\theta)\cos(\theta) - \sin(2\theta)\sin(\theta)\\ &= (\cos^2(\theta) - \sin^2(\theta))\cos(\theta) - 2\sin(\theta)\cos(\theta)\sin(\theta)\\ &= \cos^3(\theta) - \sin^2(\theta)\cos(\theta) - 2\sin^2(\theta)\cos(\theta)\\ &= \cos(\theta)(\cos^2(\theta) - \sin^2(\theta) - 2\sin^2(\theta)) = \cos(\theta)(\cos^2(\theta) - 3\sin^2(\theta))\\ &= \cos^3(\theta) - 3\sin^2(\theta)\cos(\theta) = \cos^3(\theta) - 3(1-\cos^2(\theta))\cos(\theta)\\ &= \cos^3(\theta) - 3\cos(\theta) + 3\cos^3(\theta) = 4\cos^3(\theta) - 3\cos(\theta)\end{aligned} \tag{6.1.163}$$

In fact, choosing $u = 2\sqrt{-\dfrac{p}{3}}$ and dividing the equation by $\dfrac{u^3}{4}$ we get:

$$4\cos^3(\theta) - 3\cos(\theta) - \frac{3q}{2p}\sqrt{\frac{-3}{p}} = 0 \tag{6.1.164}$$

Comparing with the above identity, we get:

$$\cos(3\theta) = \frac{3q}{2p}\sqrt{\frac{-3}{p}} \tag{6.1.165}$$

and thus the roots of the depressed cubic equation is:

$$X_k = 2\sqrt{-\frac{p}{3}}\cos\left(\frac{1}{3}\arccos\left(\frac{3q}{2p}\sqrt{\frac{-3}{p}}\right) - \frac{2\pi k}{3}\right) \quad \text{for} \quad k = 0,1,2 \tag{6.1.166}$$

6.1.3.9 Fourth order univariate Polynomial and Equations

The univariate polynomial equation to solve here is:

$$ax^4 + bx^3 + cx^2 + dx + e = 0 \tag{6.1.167}$$

with $a \neq 0$.

Remark

We own the method of resolution of 4th degree polynomial to the Italian mathematician Ludovico Ferrari (also) of the 116th century (holocene calendar).

If we divide by a we have:

$$x^4 + b'x^3 + c'x^2 + d'x + e' = 0 \tag{6.1.168}$$

And by putting:

$$x = y - \frac{b'}{4} \tag{6.1.169}$$

the equation will be reduce to:

$$\begin{aligned}
&\left(y'-\frac{b'}{4}\right)^4+b'\left(y'-\frac{b'}{4}\right)^3+c'\left(y'-\frac{b'}{4}\right)^2+d'\left(y'-\frac{b'}{4}\right)+e'\\
&=\left(y^4-4y^3\frac{b'}{4}+6y^2\frac{b'^2}{4^2}-4y\frac{b'^3}{4^3}+\frac{b'^4}{4^4}\right)+b'\left(y^3-3y^3\frac{b'}{4}+3y\frac{b'^2}{4^2}-\frac{b'^3}{4^3}\right)\\
&+c'\left(y^2-2y\frac{b'}{4}+\frac{b'^2}{4^2}\right)+d'\left(y-\frac{b'}{4}\right)+e'\\
&=y^4+(b'-b')y^3+\left(\frac{3}{4}b'^2+c'+\frac{3}{8}b'^2\right)y^2+\left(\frac{3}{16}b'^3-\frac{1}{2}c'b'-\frac{1}{16}b'^3+d'\right)y\\
&+\left(\frac{b'^4}{256}-\frac{1}{4}d'b'-\frac{1}{64}b'^4+\frac{1}{16}c'b'^2+e'\right)
\end{aligned} \tag{6.1.170}$$

where we see that the coefficient in front of the y^3 vanishes. Thus, any polynomial of the type:

$$P(x)=x^4+bx^3+cx^2+dx+1e \tag{6.1.171}$$

can be written in the following form:

$$y^4+c''y^2+d''y+e''=0 \tag{6.1.172}$$

By putting:

$$x=y-\frac{b'}{4} \tag{6.1.173}$$

$$c''=\frac{3}{4}b'^2+c'+\frac{3}{8}b'^2 \tag{6.1.174}$$

$$d''=\frac{3}{16}b'^3-\frac{1}{2}c'b'-\frac{1}{16}b'^3+d' \tag{6.1.175}$$

$$e''=\frac{b'^4}{256}-\frac{1}{4}d'b'-\frac{1}{64}b'^4+\frac{1}{16}c'b'^2+e' \tag{6.1.176}$$

Remark

If $d''=0$ the equation to solve is reality a "bisquare equation". The change of variable $X=y^2$ then allows to focus to a polynomial equation of the second degree (what we know is easily to solve).

We now introduce a parameter t (that we will choose wisely afterwards) and we rewrite the polynomial equation as follows:

$$(x^2+t)^2-\left[(2t-c'')x^2-d''x+(t^2-e'')\right]=0 \tag{6.1.177}$$

Remark

If the reader develop and distribute all the term of the previous relation he will fall back obviously on $x^4+c''x^2+d''x+e''=0$.

The underlying idea is to try to ensure that the bracketed part of the previous expression can be written as a square as:

$$(2t-c'')x^2-d''x+(t^2-e'')=(x-\lambda)^2 \tag{6.1.178}$$

Because in this case, using:

$$a^2-b^2=(a+b)(a-b) \tag{6.1.179}$$

Our polynomial equation can then be written:

$$(x^2+t)^2-(x-\lambda)^2=\big((x^2+t)+(x-\lambda)\big)\big((x^2+t)-(x-\lambda)\big)=0 \tag{6.1.180}$$

and we would have only to solve two polynomial equations of the second degree (what we know already how to do).

But for us to write:

$$(2t-c'')x^2-d''x+(t^2-e'')=(x-\lambda)^2 \tag{6.1.181}$$

the expression of second degree to the left of equality should have only one root. But we saw in our study of polynomial equations of the second degree that meant since the discriminant is zero:

$$\Delta=b^2-4ac=0 \tag{6.1.182}$$

and that the root was given by:

$$x=-\frac{b}{2a} \tag{6.1.183}$$

Which corresponds in our case to:

$$\Delta=d''^2-4(2t-c'')(t^2-e'')=0 \tag{6.1.184}$$

and therefore that:

$$4(2t-c'')(t^2-e'')=d''^2 \tag{6.1.185}$$

with:

$$\lambda=-\frac{-d''}{2(2t-c'')}=\frac{d''}{4t-2c''} \tag{6.1.186}$$

So finally, if t is such that $4(2t-c'')(t^2-e'')=d''^2$, then we have:

$$(2t-c'')x^2-d''x+(t^2-e'')=(2t-c'')(x-\lambda)(x-\lambda)=(2t-c'')\left(x-\frac{d''}{4t-2c''}\right)^2 \tag{6.1.187}$$

as the fundamental polynomial theorem give us for a polynomial of the second degree a unique root:

$$P(x)=ax^2+bx+c=a(x-r_1)(x-r_2)=a(x-r_1)^2 \tag{6.1.188}$$

To conclude, it is enough to find a number t satisfying the following relation:

$$4(2t-c'')(t^2-e'')=d''^2 \tag{6.1.189}$$

which is a degree 3 polynomial that we already know how to solve using the Cardan method.

Such general methods doesn't exist anymore for polynomial of degree higher or equal to 5 as we will prove it during our study of Galois Theory (see section Set Algebra page 1445).

6.1.3.10 Trigonometric Polynomials

Definition (#6.1.174): We name "trigonometric polynomial" of degree N any finite sum of the type:

$$S_N(\theta)=\sum_{n=-N}^{N} c_n e^{\mathrm{i}n\theta} \tag{6.1.190}$$

where $c_n \in \mathbb{C}$.

A trigonometric polynomial can also be written using the usual trigonometric functions with the following changes:

$$S_N(\theta)=c_0+\sum_{n=1}^{N} c_n e^{\mathrm{i}n\theta}+\sum_{n=1}^{N} c_{-n} e^{-\mathrm{i}n\theta} \tag{6.1.191}$$

Either using Euler's formula (see section Numbers page 251):

$$S_N(\theta)=c_0+\sum_{n=1}^{N} c_n(\cos(n\theta)+\mathrm{i}\sin(n\theta))+\sum_{n=1}^{N} c_{-n}(\cos(n\theta)-\mathrm{i}\sin(n\theta)) \tag{6.1.192}$$

What we can also rewrite as:

$$S_N(\theta)=c_0+\sum_{n=1}^{N}(c_n+c_{-n})\cos(n\theta)+\sum_{n+1}^{N}\mathrm{i}(c_n-c_{-n})\sin(n\theta) \tag{6.1.193}$$

By putting:

$$a_0=2c_0 \qquad a_n=c_n+c_{-n} \qquad b_n=\mathrm{i}(c_n-c_{-n}) \tag{6.1.194}$$

It comes:

$$S_N(\theta)=\frac{a_0}{2}+\sum_{i=1}^{N} a_n\cos(n\theta)+\sum_{n=1}^{N} b_n\sin(n\theta) \tag{6.1.195}$$

In engineering, the latter is often denoted:

$$S_N(t)=\frac{a_0}{2}+\sum_{i=1}^{N} a_n\cos(n2\pi f_0 t)+\sum_{n=1}^{N} b_n\sin(n2\pi f_0 t)=\frac{a_0}{2}+\sum_{i=1}^{N} a_n\cos\left(n\frac{2\pi}{T}t\right)+\sum_{n=1}^{N} b_n\sin\left(n\frac{2\pi}{T}t\right) \tag{6.1.196}$$

where f_0 is referred to as the "fundamental frequency".

We will extensively discussed in the section Sequences and Series how to use these polynomials in the context of the study of Fourier series (see page 1727).

6.1.3.11 Cyclotomic Polynomials

Definition (#6.1.175): If n is an integer (belongs to $\mathbb{N}$) and x a complex number (belongs to $\mathbb{C}$), we name "cyclotomic polynomial" that we denote traditionally by Φ_n and that we define as being the product of all monomials:

$$(x-\alpha) \tag{6.1.197}$$

where α is a nth primitive root of $\mathbb{C}$. In other words:

$$\Phi_n=(x-\alpha_0)(x-\alpha_2)(x-\alpha_2)\ldots(x-\alpha_{n-1}) \tag{6.1.198}$$

To recall an nth "unit root" (sometimes named "De Moivre number") is a complex number whose n-th power is equal to 1.

Thus, the set of all n-th unit roots is given by:

$$G_n=\{x\in\mathbb{C}\quad\text{such that}\quad x^n=1\} \tag{6.1.199}$$

which is a cyclic group (see the section Set Theory and also the section Set Algebra page 1417).

Then we name "n-th primitive root of unity" any element of this group generating it.

The elements of G_n are of the type:

$$e^{\frac{2\mathrm{i}k\pi}{n}} \tag{6.1.200}$$

with $k\in\mathbb{Z}$. We write then the set of the G_n in the form:

$$G_n=\left\{e^{\frac{2\mathrm{i}k\pi}{n}}\quad\text{with}\quad k\in\{0,\ldots,n-1\}\right\} \tag{6.1.201}$$

A small example of cyclotomic polynomial (more example will be given below):

$$\Phi_4(x)=(x-\mathrm{i})(x+\mathrm{i})(x-1)(x+1)=x^4-1 \tag{6.1.202}$$

with:

$$G_4=\left\{e^{\frac{2\mathrm{i}0\pi}{4}}=1,e^{\frac{2\mathrm{i}1\pi}{4}}=1,e^{\frac{2\mathrm{i}2\pi}{4}}=-1,e^{\frac{2\mathrm{i}3\pi}{4}}=-\mathrm{i}\right\} \tag{6.1.203}$$

which are therefore the 4th roots of unity (in other words: each of these number set to the power of 4 is equal to 1). They form the group G_4 and this can be generated only by i and $-\mathrm{i}$ (field generator according to what was seen in the section Set Theory).

So a cyclotomic polynomial is the product of factors that is written:

$$\Phi_n(x)=\prod_n\left(x-e^{\frac{2\mathrm{i}k\pi}{n}}\right) \tag{6.1.204}$$

with $k\in\{0,\ldots,n-1\}$.

We will see with the examples below that if n is even then:

$$\Phi_n(x)=x^n-1=\prod_n\left(x-e^{\frac{2\mathrm{i}k\pi}{n}}\right) \tag{6.1.205}$$

and if n is odd:

$$\Phi_n(x)=1+x+x^2+x^{n-1}=\sum_{k=0}^{n-1}x^k=\prod_n\left(x-e^{\frac{2\mathrm{i}k\pi}{n}}\right) \tag{6.1.206}$$

Example:

For n up to 30, the cyclotomic polynomials are:

$$\begin{aligned}
\Phi_1(x)&=x-1\\
\Phi_2(x)&=x+1\\
\Phi_3(x)&=x^2+x+1\\
\Phi_4(x)&=x^2+1\\
\Phi_5(x)&=x^4+x^3+x^2+x+1\\
\Phi_6(x)&=x^2-x+1\\
\Phi_7(x)&=x^6+x^5+x^4+x^3+x^2+x+1\\
\Phi_8(x)&=x^4+1\\
\Phi_9(x)&=x^6+x^3+1\\
\Phi_{10}(x)&=x^4-x^3+x^2-x+1\\
\Phi_{11}(x)&=x^{10}+x^9+x^8+x^7+x^6+x^5+x^4+x^3+x^2+x+1\\
\Phi_{12}(x)&=x^4-x^2+1\\
\Phi_{13}(x)&=x^{12}+x^{11}+x^{10}+x^9+x^8+x^7+x^6+x^5+x^4+x^3+x^2+x+1\\
\Phi_{14}(x)&=x^6-x^5+x^4-x^3+x^2-x+1\\
\Phi_{15}(x)&=x^8-x^7+x^5-x^4+x^3-x+1\\
\Phi_{16}(x)&=x^8+1\\
\Phi_{17}(x)&=x^{16}+x^{15}+x^{14}+x^{13}+x^{12}+x^{11}+x^{10}+x^9+x^8+x^7+x^6+x^5\\
&\quad+x^4+x^3+x^2+x+1\\
\Phi_{18}(x)&=x^6-x^3+1\\
\Phi_{19}(x)&=x^{18}+x^{17}+x^{16}+x^{15}+x^{14}+x^{13}+x^{12}+x^{11}+x^{10}+x^9+x^8+x^7\\
&\quad+x^6+x^5+x^4+x^3+x^2+x+1\\
\Phi_{20}(x)&=x^8-x^6+x^4-x^2+1\\
\Phi_{21}(x)&=x^{12}-x^{11}+x^9-x^8+x^6-x^4+x^3-x+1\\
\Phi_{22}(x)&=x^{10}-x^9+x^8-x^7+x^6-x^5+x^4-x^3+x^2-x+1\\
\Phi_{23}(x)&=x^{22}+x^{21}+x^{20}+x^{19}+x^{18}+x^{17}+x^{16}+x^{15}+x^{14}+x^{13}+x^{12}\\
&\quad+x^{11}+x^{10}+x^9+x^8+x^7+x^6+x^5+x^4+x^3+x^2+x+1\\
\Phi_{24}(x)&=x^8-x^4+1\\
\Phi_{25}(x)&=x^{20}+x^{15}+x^{10}+x^5+1\\
\Phi_{26}(x)&=x^{12}-x^{11}+x^{10}-x^9+x^8-x^7+x^6-x^5+x^4-x^3+x^2-x+1\\
\Phi_{27}(x)&=x^{18}+x^9+1\\
\Phi_{28}(x)&=x^{12}-x^{10}+x^8-x^6+x^4-x^2+1\\
\Phi_{29}(x)&=x^{28}+x^{27}+x^{26}+x^{25}+x^{24}+x^{23}+x^{22}+x^{21}+x^{20}+x^{19}+x^{18}+x^{17}+x^{16}+x^{15}\\
&\quad+x^{14}+x^{13}+x^{12}+x^{11}+x^{10}+x^9+x^8+x^7+x^6+x^5+x^4+x^3+x^2+x+1\\
\Phi_{30}(x)&=x^8+x^7-x^5-x^4-x^3+x+1
\end{aligned}\tag{6.1.207}$$

6.1.3.12 Legendre Polynomials

Definition (#6.1.176): Legendre polynomials are defined by (it is strongly recommended to read the sections of Differential and Integral Calculus page 1450 and also of Functional Analysis 2170 before continuing):

$$P_n(x)=\frac{(-1)^n}{n!2^n}\frac{\mathrm{d}^n}{\mathrm{d}x^n}(1-x^2)^n \tag{6.1.208}$$

where P_n is therefore a polynomial of degree n. We will see again these polynomials in the resolution of differential equations in physics (heat propagation, quantum physics, quantum chemistry, etc.). In most book the Legendre polynomial are written in the following equivalent form:

$$P_n(x)=\frac{1}{n!2^n}\frac{\mathrm{d}^n}{\mathrm{d}x^n}(x^2-1)^n \tag{6.1.209}$$

We will focus here only and uniquely on the properties that are used actually in the other Physics sections of this textbook!!!

Let us prove that following the definition of functional scalar product (see the sections Functional Analysis page 2379 and Vector Calculus 1821) Legendre polynomials are orthogonal. This is a very important property for our study of Quantum Chemistry (see page 6134) later!

Proof 6.4.2. Let P be a polynomial of degree $\leq n-1$. It suffices to prove that $\langle P_n|P\rangle=0$, that is to say that P_n is orthogonal to the space of polynomials of degree less than n. Indeed, we have:

$$\langle P_n|P\rangle=\int\limits_{-1}^{1}P_n(x)P(x)\mathrm{d}x=\frac{(-1)^n}{n!2^n}\int\limits_{-1}^{1}\frac{\mathrm{d}^n}{\mathrm{d}x^n}(1-x^2)^nP(x)\mathrm{d}x \tag{6.1.210}$$

integrating by parts we get:

$$\begin{aligned}\langle P_n|P\rangle&=\underbrace{\frac{(-1)^n}{n!2^n}\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}(1-x^2)^nP(x)\Bigg|_{-1}^{1}}_{=0}-\frac{(-1)^n}{n!2^n}\int\limits_{-1}^{1}\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}(1-x^2)^n\frac{\mathrm{d}}{\mathrm{d}x}P(x)\mathrm{d}x\\&=-\frac{(-1)^n}{n!2^n}\int\limits_{-1}^{1}\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}(1-x^2)^n\frac{\mathrm{d}}{\mathrm{d}x}P(x)\mathrm{d}x\end{aligned} \tag{6.1.211}$$

> Caution!!! For the above zero term, only the term $(1-x^2)^n$ is derived there! So since x is squared, whatever the derivative the value will always be the same. This is what justifies that this term is equal to zero.

Continuing in this way we get after n integration by parts (see section Differential and Integral Calculus page 1532):

$$\langle P_n|P\rangle=-\frac{(-1)^n}{n!2^n}\int\limits_{-1}^{1}(1-x^2)^n\underbrace{\frac{\mathrm{d}^n}{\mathrm{d}x^n}P(x)}_{=0}\,\mathrm{d}x=0 \tag{6.1.212}$$

Remark

The derivative term is zero, since the derivative polynomial is of degree $n-1$.

■ Q.E.D.

Here are some useful properties for the section of Quantum Chemistry of the Legendre polynomials:

P1. We have $P_n(1)=1$:

Proof 6.4.3.

$$P_n(x)=\frac{(-1)^n}{n!2^n}\frac{\mathrm{d}^n}{\mathrm{d}x^n}(1-x^2)^n=\frac{(-1)^n}{n!2^n}\frac{\mathrm{d}^n}{\mathrm{d}x^n}((1+x)^n(1-x)^n) \tag{6.1.213}$$

and using the Leibniz differentiation rule for products (see section Differential and Integral Calculus page 1479) we have:

$$\frac{(-1)^n}{n!2^n}\frac{\mathrm{d}^n}{\mathrm{d}x^n}((1+x)^n(1-x)^n)=\frac{(-1)^n}{n!2^n}\sum_{j=0}^{n}\binom{n}{j}\frac{n!}{(n-j)!}(1+x)^{n-j}\frac{(-1)^{n-j}n!}{j!}(1-x)^j \tag{6.1.214}$$

Therefore:

$$P_n(1)=\frac{(-1)^n}{n!2^n}(-1)^n n!=1 \tag{6.1.215}$$

■ Q.E.D.

P2. We have $P_n(-x)=P_n(-x)$ if n is even:

Proof 6.4.4. If n is even:

$$\frac{\mathrm{d}^n}{\mathrm{d}x^n}(1-x^2)^n \tag{6.1.216}$$

is an even function and therefore:

$$P_n(x)=\frac{(-1)^n}{n!2^n}\frac{\mathrm{d}^n}{\mathrm{d}x^n}(1-x^2)^n \tag{6.1.217}$$

is even.

■ Q.E.D.

P3. We have $P_n(-x)=-P_n(x)$ if n is odd:

Proof 6.4.5. If n is odd:

$$\frac{\mathrm{d}^n}{\mathrm{d}x^n}(1-x^2)^n \tag{6.1.218}$$

is an odd function and therefore:

$$P_n(x)=\frac{(-1)^n}{n!2^n}\frac{\mathrm{d}^n}{\mathrm{d}x^n}(1-x^2)^n \tag{6.1.219}$$

is odd.

■ Q.E.D.

Theorem 6.5. We will now prove the validity of the following recurrence relation for the P_n (relations that we use in physics):

$$(n+1)P_{n+1}(x)=(2n+1)xP_n(x)-nP_{n-1}(x) \tag{6.1.220}$$

for $n\geq 1$.

Proof 6.5.1. $xP_n(x)$ is a polynomial of degree $n+1$, there exists therefore $a_j\in\mathbb{R}$ such that this polynomial can be expressed as a linear combination of the family of polynomials constituting the orthonormal basis (basis that enables us to generate/build the $xP_n(x)$):

$$xP_n(x)=\sum_{j=0}^{n+1}a_jP_j \tag{6.1.221}$$

Therefore we can write:

$$\langle xP_n|P_k\rangle=\left\langle\sum_{j=0}^{n+1}a_jP_j\middle|P_k\right\rangle=\sum_{j=0}^{n+1}a_j\langle P_j|P_k\rangle=a_k\|P_k\|^2 \tag{6.1.222}$$

but if we choose $k\leq n-2$ (because xP_k is therefore of degree $n-1$):

$$\langle xP_n|P_k\rangle=\int\limits_{-1}^{1}xP_n(x)P_k(x)\mathrm{d}x=\langle P_n|xP_k\rangle=0 \tag{6.1.223}$$

Therefore:

$$a_k\|P_k\|^2=0 \tag{6.1.224}$$

that is to say that $a_k=0$. Then it follows:

$$xP_n=a_{n+1}P_{n+1}+a_nP_n+a_{n-1}P_{n-1} \tag{6.1.225}$$

By the properties of the Legendre polynomials proved previously, we can write the equalities:

$$xP_n(x),x=1:\qquad 1=a_{n+1}+a_n+a_{n-1} \tag{6.1.226}$$

and:

$$xP_n(x),x=-1:\qquad 1=a_{n+1}-a_n+a_{n-1} \tag{6.1.227}$$

hence:

$$a_n=0\qquad\text{and}\qquad 1=a_{n+1}+a_{n-1} \tag{6.1.228}$$

The dominant coefficient of P_n denoted by $\text{dom}(P_n(x))$ is defined (for recall) as the coefficient of the monomial of the highest degree. Thus, it is given by:

$$\begin{aligned}P_n(x)&=\frac{(-1)^n}{n!2^n}\frac{\mathrm{d}^n}{\mathrm{d}x^n}(1-x^2)^n\Rightarrow P_n(x)=\frac{(-1)^n}{n!2^n}\frac{\mathrm{d}^n}{\mathrm{d}x^n}x^{2n}\\&\Rightarrow P_n(x)=\frac{(-1)^n}{n!2^n}\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}(2nx^{2n-1})=\frac{(-1)^n}{n!2^n}\frac{\mathrm{d}^{n-n}}{\mathrm{d}x^{n-n}}(2n(2n-1)\ldots x^{2n-n})\\&=\frac{(-1)^n}{n!2^n}(2n(2n-1)\ldots(2n-(n-1))x^n)=\frac{(-1)^n}{n!2^n}(2n(2n-1)\ldots(n+1)x^n)=\frac{(2n)!}{(n!)^22^n}x^n\end{aligned} \tag{6.1.229}$$

Therefore:

$$\text{dom}(P_n(x))=\frac{(2n)!}{(n!)^22^n} \tag{6.1.230}$$

Remark

The reader will verify if needed for a given n that:

$$2n(2n-1)\ldots(n+1)=\frac{(2n)!}{n!} \tag{6.1.231}$$

The relation:

$$xP_n=a_{n+1}P_{n+1}+a_nP_n+a_{n-1}P_{n-1} \tag{6.1.232}$$

we got earlier impose us that the dominant coefficient of the polynomial of the linear combination to be equal to the dominant coefficient of the polynomial xP_n (we have eliminated the $(-1)^n$ which simplifies):

$$\cancel{(-1)^n}\frac{(2n)!}{(n!)^2 2^n}=a_{n+1}\cancel{(-1)^{n+2}}\frac{(2n+2)!}{(n+1)!^2 2^{n+1}} \tag{6.1.233}$$

After simplification we get:

$$a_{n+1}=\frac{(2n)!(n+1)!^2 2^{n+1}}{(n!)^2 2^n(2n+2)!}=\frac{(n+1)\cancel{(n+1)2}}{(2n+1)\cancel{(2n+2)}}=\frac{n+1}{2n+1} \tag{6.1.234}$$

and which finally gives easily:

$$a_{n-1}=\frac{n}{2n+1} \tag{6.1.235}$$

The relation:

$$xP_n=a_{n+1}P_{n+1}+a_nP_n+a_{n-1}P_{n-1} \tag{6.1.236}$$

becomes therefore:

$$(2n+1)xP_n=(n+1)P_{n+1}+nP_{n-1} \tag{6.1.237}$$

■ Q.E.D.

The first Legendre polynomials are:

$$\begin{aligned}
&P_0(x)=1\\
&P_1(x)=x\\
&P_2(x)=\frac{3}{2}x^2-\frac{1}{2}\\
&P_3(x)=\frac{5}{2}x^3-\frac{3}{2}x\\
&P_4(x)=\frac{35}{8}x^4-\frac{15}{4}x^2+\frac{3}{8}\\
&\qquad\ldots
\end{aligned} \tag{6.1.238}$$

The graphs of these polynomials (up to $n=5$) are shown below:

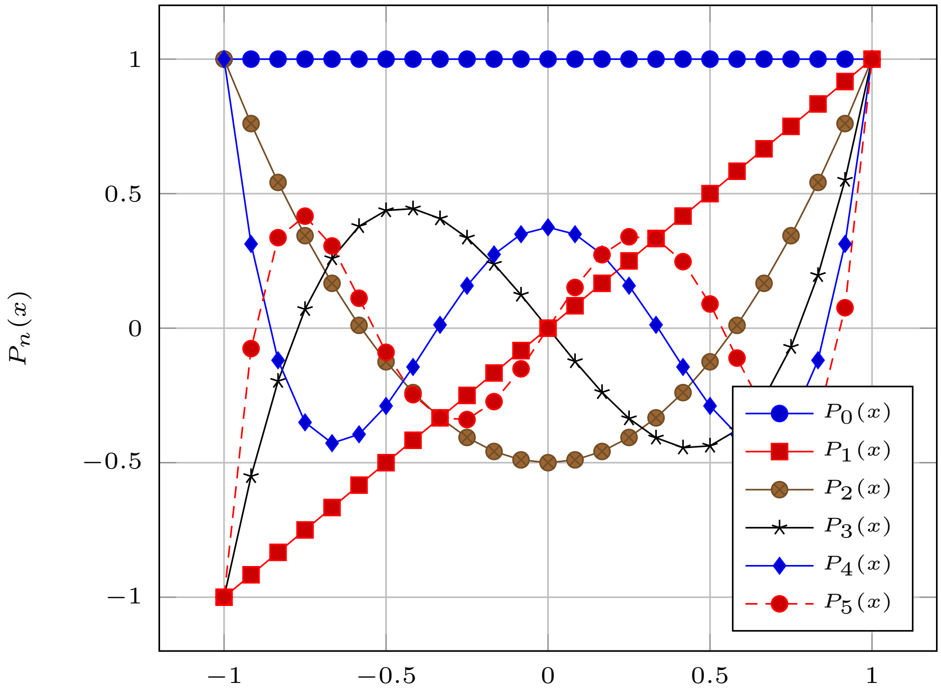


Figure 6.17: Five first Legendre polynomials

6.1.3.13 Laguerre Polynomials

Definition (#6.1.177): Before we deal with the maths of Laguerre polynomials the reader has to known the huge majority of the text below is a simple copy/paste of the content of the excellent textbook [24]. It is by the way strongly recommended to read the sections of Differential and Integral Calculus page 1450 and also of Functional Analysis 2170 before continuing!

"Laguerre's ordinary differential equation" (which derives from the radial partial differential equation of Schrödinger's for the hydrogen atom[3] as we will prove it in the section of Quantum Chemistry at page 6173) is:

$$\boxed{xy''(x)+(1-x)y'(x)+\lambda y(x)=0} \tag{6.1.239}$$

This equation has non-singular solutions only if n is a non-negative integer.

Let us derive a solution using the Frobenius method (see section Sequences and Series page 1773)! For that purpose we first assume that the solution y is given in the form:

$$y=\sum_{n=0}^{+\infty}a_nx^{r+n} \tag{6.1.240}$$

with its first two derivatives:

$$y'=\sum_{n=0}^{+\infty}(r+n)a_nx^{r+n-1} \tag{6.1.241}$$

---

[3] The Laguerre polynomials arise in quantum mechanics, in the radial part of the solution of the Schrödinger equation for a one-electron atom. They also describe the static Wigner functions of oscillator systems in quantum mechanics in phase space. They further enter in the quantum mechanics of the Morse potential and of the 3D isotropic harmonic oscillator.

and:

$$y'' = \sum_{n=0}^{+\infty} (r+n)(r+n-1)a_n x^{r+n-2} \tag{6.1.242}$$

By substituting these expressions in the given differential equation, we get:

$$x\sum_{n=0}^{+\infty}(r+n)(r+n-1)a_n x^{r+n-2} + (1-x)\sum_{n=0}^{+\infty}(r+n)a_n x^{r+n-1} + \lambda\sum_{n=0}^{+\infty} a_n x^{r+n} = 0 \tag{6.1.243}$$

By multiplying the factors with the corresponding summations, we get:

$$\sum_{n=0}^{+\infty}(r+n)(r+n-1)a_n x^{r+n-1} + \sum_{n=0}^{+\infty}(r+n)a_n x^{r+n-1} - \sum_{n=0}^{+\infty}(r+n)a_n x^{r+n} + \lambda\sum_{n=0}^{+\infty} a_n x^{r+n} = 0 \tag{6.1.244}$$

Now we adjust the index of the first two summations so that x in them has exponent $r+n$:

$$\sum_{n=-1}^{+\infty}(r+n+1)(r+n)a_{n+1} x^{r+n} + \sum_{n=-1}^{+\infty}(r+n+1)a_{n+1} x^{r+n} - \sum_{n=0}^{+\infty}(r+n)a_n x^{r+n} + \lambda\sum_{n=0}^{+\infty} a_n x^{r+n} = 0 \tag{6.1.245}$$

We evaluate a few terms so that all summations start at index $n=0$ and get:

$$\begin{gathered} r(r-1)a_0 x^{r-1} + \sum_{n=0}^{+\infty}(r+n+1)(r+n)a_{n+1}x^{r+n} + ra_0 x^{r-1} \\ + \sum_{n=0}^{+\infty}(r+n+1)a_{n+1}x^{r+n} - \sum_{n=0}^{+\infty}(r+n)a_n x^{r+n} + \lambda\sum_{n=0}^{+\infty} a_n x^{r+n} = 0 \end{gathered} \tag{6.1.246}$$

Combining the terms, we have:

$$[r(r-1)+r]a_0 x^r + \sum_{n=0}^{+\infty}\{[(r+n+1)(r+n)+(r+n+1)]a_{n+1} - (r+n-\lambda)a_n\}\, x^{r+n} = 0 \tag{6.1.247}$$

which is, after simplifying:

$$r^2 a_0 x^r + \sum_{n=0}^{+\infty}\left[(r+n+1)^2 a_{n+1} - (r+n-\lambda)a_n\right] x^{r+n} = 0 \tag{6.1.248}$$

Equating the coefficients on both sides, we get:

$$\begin{cases} r^2 a_0 = 0 \\ (r+n+1)^2 a_{n+1} - (r+n-\lambda)a_n = 0, \forall n \in \mathbb{N}_0 \end{cases} \tag{6.1.249}$$

Dividing the last equation by $r+n+1 \neq 0$ and shifting the indices, we get the "recurrence equation":

$$(r+n)^2 a_n - (r+n-1-\lambda)a_{n-1} = 0, \forall n \in \mathbb{N} \tag{6.1.250}$$

Since $a_0 \neq 0$, we get the "indical equation" of the initial equation:

$$r^2 = 0 \tag{6.1.251}$$

with roots $r_{1,2} = 0$.

The recurrence relation can be written as:

$$a_n = \frac{(r+n-1-\lambda)a_{n-1}}{(r+n)^2} = \frac{(n-1-\lambda)a_{n-1}}{n^2} \tag{6.1.252}$$

Therefore:

$$\begin{aligned}
a_0 &:= 1\\
a_1 &= \frac{-\lambda \cdot a_0}{1^2} = \frac{-\lambda \cdot 1}{1} = -\lambda\\
a_2 &= \frac{(1-\lambda)\cdot(-\lambda)}{4} = \frac{-\lambda+\lambda^2}{4} = \frac{\lambda(\lambda-1)}{2^2}\\
a_3 &= \frac{(2-\lambda)\cdot(-\lambda+\lambda^2)}{9\cdot 4} = \frac{(2-\lambda)\cdot\lambda(\lambda-1)}{2^2\cdot 3^2} = \frac{-\lambda(\lambda-1)(\lambda-2)}{2^2\cdot 3^2}\\
a_4 &= \frac{-\lambda(3-\lambda)(\lambda-1)(\lambda-2)}{4^2\cdot 3^2\cdot 2^2} = \frac{\lambda(\lambda-3)(\lambda-2)(\lambda-1)}{4^2\cdot 3^2\cdot 2^2}
\end{aligned} \tag{6.1.253}$$

Therefore, choosing $a_0 = 1$, the solution $y = y_1$ is equal to:

$$\begin{aligned}
y_1 &= x^r \sum_{n=0}^{+\infty} a_n x^n = 1 + \frac{-\lambda}{1^2}x + \frac{(-\lambda)(1-\lambda)}{1^2\cdot 2^2}x^2 + \frac{(-\lambda)(1-\lambda)(2-\lambda)}{1^2\ldots 2^2\ldots 3^2}x^3\\
&+\cdots + \frac{(-\lambda)(1-\lambda)\cdots(n-1-\lambda)}{(n!)^2} + \cdots\\
&= 1 - \lambda x + \frac{\lambda(\lambda-1)}{2^2}x^2 - \frac{\lambda(\lambda-1)(\lambda-2)}{2^2\cdot 3^2}x^3 + \frac{\lambda(\lambda-1)(\lambda-2)(\lambda-3)}{2^2\cdot 3^2\cdot 4^2}x^4 + \ldots
\end{aligned} \tag{6.1.254}$$

We see that this sum is finite as whatever the value of λ at a given moment all factor will have $(\lambda - \ldots)$ that is equal to zero. So all the terms will cancel. Thus the solution y_1 has finitely many terms and is a polynomial.

For $\lambda = n$, this solution becomes:

$$\begin{aligned}
y(t) &= 1 - x + \frac{1}{2}x^2 - \frac{1\cdot 2\cdot 3}{2^2\cdot 3^2}x^3 + \frac{1\cdot 2\cdot 3\cdot 4}{2^2\cdot 3^2\cdot 4^2}x^4 - \ldots\\
&= 1 - x + \frac{1}{2}x^2 - \frac{1}{2\cdot 3}x^3 + \frac{1}{2\cdot 3\cdot 4}x^4 - \ldots\\
&= 10x + \frac{1}{2!}x^2 - \frac{1}{3!}x^3 + \frac{1}{4!}x^4 - \ldots
\end{aligned} \tag{6.1.255}$$

We can see that this solution is polynomial if $\lambda = n$. Now let us come back to:

$$\begin{aligned}
y_1 &= 1 - \lambda x + \frac{\lambda(\lambda-1)}{2^2}x^2 - \frac{\lambda(\lambda-1)(\lambda-2)}{2^2\cdot 3^2}x^3 + \frac{\lambda(\lambda-1)(\lambda-2)(\lambda-3)}{2^2\cdot 3^2\cdot 4^2}x^4 + \ldots\\
&= 1 - \frac{\lambda(\lambda-1)!}{1^2(\lambda-1)!}x^1 + \frac{\lambda(\lambda-1)(\lambda-2)!}{1^2 2^2(\lambda-2)!}x^2 - \frac{\lambda(\lambda-1)(\lambda-2)(\lambda-3)!}{1^2 2^2 3^2(\lambda-3)!}x^3 + \frac{\lambda(\lambda-1)(\lambda-2)(\lambda-3)(\lambda-4)!}{1^2 2^2 3^2 4^2(\lambda-4)!}x^4 + \ldots\\
&= 1 - \frac{\lambda!}{(1!)^2(\lambda-1)!}x^1 + \frac{\lambda!}{(2!)^2(\lambda-2)!}x^2 - \frac{\lambda!}{(3!)^2(\lambda-3)!}x^3 + \frac{\lambda!}{(4!)^2(\lambda-4)!}x^4 + \ldots
\end{aligned} \tag{6.1.256}$$

That last relation bring us to define the Laguerre polynomial as:

$$\boxed{L_n(x)=\sum_{\lambda=0}^{n}\frac{(-1)^{\lambda}n!}{(\lambda!)^2(n-\lambda)!}x^{\lambda}=\sum_{\lambda=0}^{n}\binom{n}{\lambda}\frac{(-x)^{\lambda}}{\lambda!}} \tag{6.1.257}$$

That relation can often be found in textbooks under the following form:

$$\begin{aligned}L_n(x)&=\frac{(-1)^n}{n!}\left[x^n-\frac{n^2}{1!}x^{n-1}+\frac{n^2(n-1)^2}{2!}x^{n-2}-\cdots+(-1)^n n!\right]\\&=\sum_{m=0}^{n}\frac{(-1)^m n!x^m}{(n-m)(m!)^2}=\sum_{s=0}^{n}\frac{(-1)^{n-s}n!x^{n-s}}{((n-s)!)^2 s!}\end{aligned} \tag{6.1.258}$$

We can then build the following table:

n	$L_n(x)$
0	1
1	$-x+1$
2	$\frac{1}{2}(x^2-4x+2)$
3	$\frac{1}{6}(-x^3+9x^2-18x+6)$
4	$\frac{1}{24}(x^4-16x^3+72x^2-96x+24)$
5	$\frac{1}{120}(-x^5+25x^4-200x^3+600x^2-600x+120)$
6	$\frac{1}{720}(x^6-36x^5+450x^4-2400x^3+5400x^2-4320x+720)$
...	...
n	$\frac{1}{n!}((-x)^n+n^2(-x)^{n-1}+\ldots+n(n!)(-x)+n!)$

Table 6.5: **List of some Laguerre polynomials**

And let us give some plots of these Laguerre polynomials:

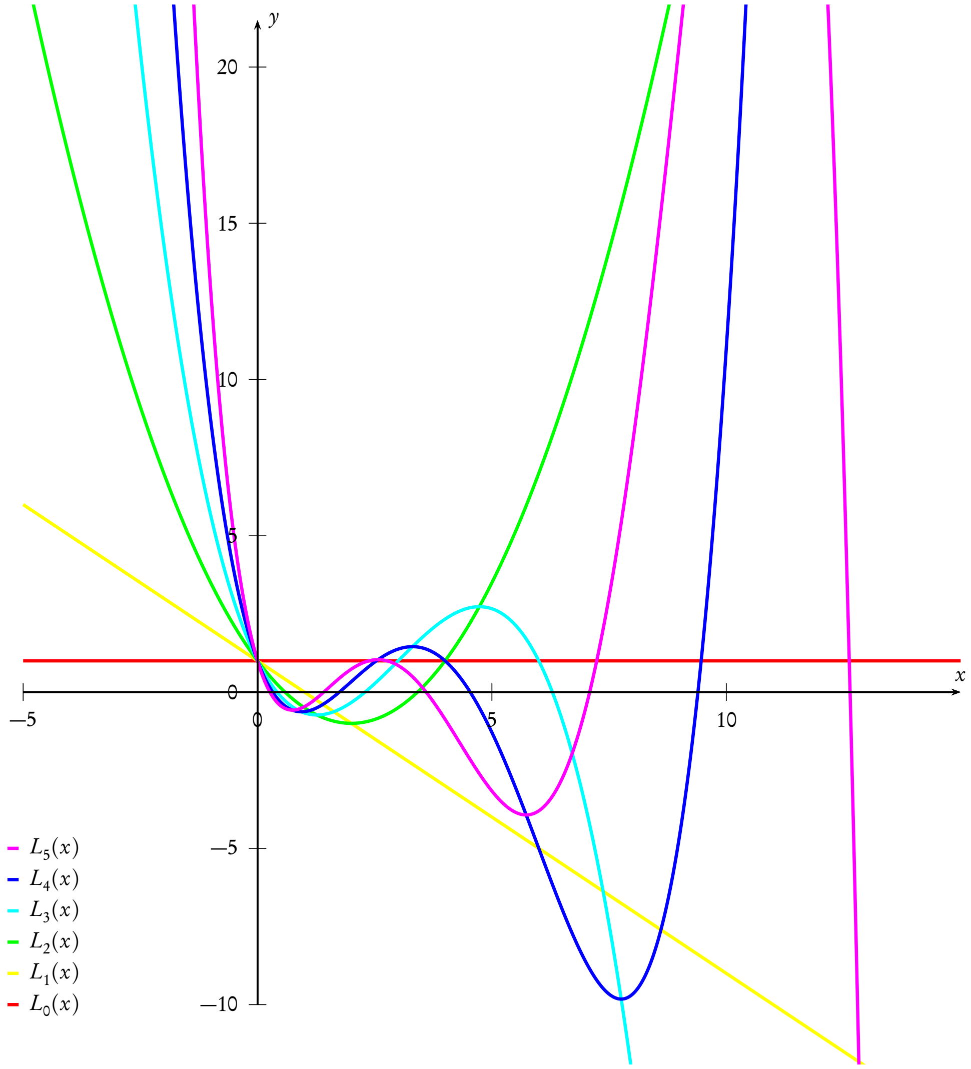


Figure 6.18: **Plot of six first Laguerre polynomials**

Now we can easily prove that:

$$\boxed{L_n(x)=\frac{e^x}{n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left(x^n e^{-x}\right)\qquad \text{(integral } n\text{)}} \tag{6.1.259}$$

named "Rodrigues' formula" for Laguerre polynomials (they are different way to express it and different versions also of it depending on what type of Laguerre polynomial we are dealing with!).

If we use the explicit representation for Laguerre polynomials:

$$L_n(x)=\sum_{k=0}^{n}\left(\begin{array}{c} n \\ k \end{array}\right)\frac{(-x)^k}{k!} \tag{6.1.260}$$

then the result follows easily from the Leibniz differentiation rule for products:

$$(f(x)g(x))^{(n)}=\sum_{k=0}^{n}\left(\begin{array}{c} n \\ k \end{array}\right) f^{(k)}(x)g^{(n-k)}(x) \tag{6.1.261}$$

Just let $f(x)=x^n$ and $g(x)=e^{-x}$. Then $f^{(k)}(x)=x^{n-k}n!/(n-k)!$ and $g^{(n-k)}(x)=(-1)^{n-k}e^{-x}$. Now if we clean it up and redo the sum in reverse order, i.e., $k\rightarrow n-k$ then we fall back on the Rodrigues' formula above.

Now let us prove that the functional generating function (see section Sequences and Series page 1789) of Laguerre polynomials is given by:

$$\boxed{G(x,t)=\sum_{n=0}^{+\infty}L_n(x)t^n,\quad |t|<1} \tag{6.1.262}$$

The short proof is given by (we simply use the Maclaurin series of e^z where $z=-xt/(1-t)$):

$$\begin{aligned}\sum_{n=0}^{+\infty}L_n(x)t^n&=\sum_{n=0}^{+\infty}\sum_{k=0}^{n}(-1)^k\begin{pmatrix}n\\k\end{pmatrix}\frac{x^kt^n}{k!}=\sum_{k=0}^{+\infty}\frac{(-1)^kx^k}{k!}\sum_{n=k}^{+\infty}\begin{pmatrix}n\\k\end{pmatrix}t^n\\&=\sum_{k=0}^{+\infty}\frac{(-1)^kx^k}{k!}\frac{t^k}{(1-t)^{k+1}}=\frac{1}{1-t}\sum_{k=0}^{+\infty}\frac{(-1)^k(xt)^k}{k!(1-t)^k}=\frac{1}{1-t}\sum_{k=0}^{+\infty}\frac{1}{k!}\left(\frac{(-xt)}{1-t}\right)^k\\&=\frac{e^{-xt/(1-t)}}{1-t}\end{aligned} \tag{6.1.263}$$

By differentiating the generating function above with respect to x and t, we get the two recurrence relations for the Laguerre polynomials as follows:

To derive the first recurrence relation we start from:

$$G(x,t)=\sum_{n=0}^{+\infty}L_n(x)t^n=\frac{\exp\{-xt/(1-t)\}}{1-t} \tag{6.1.264}$$

We differentiate both sides with respect to t to get:

$$\sum_{n=0}^{+\infty}L_n(x)nt^{n-1}=\frac{1}{(1-t)^2}e^{\frac{-xt}{1-t}}-\frac{1}{1-t}e^{-\frac{xt}{1-t}}\frac{x}{(1-t)^2} \tag{6.1.265}$$

$$=\frac{1}{1-t}\sum_{n=0}^{+\infty}L_n(x)t^n-\frac{x}{(1-t)^2}L_n(x)t^n \tag{6.1.266}$$

Multiplying both sides by $(1-t)^2$ and simplifying, we get:

$$\begin{aligned}\sum_{n=0}^{+\infty}nL_n(x)t^{n-1}-2\sum_{n=0}^{+\infty}nL_n(x)t^n+\sum_{n=0}^{+\infty}nL_n(x)t^{n+1}&=\sum_{n=0}^{+\infty}L_n(x)t^n\\&-\sum_{n=0}^{+\infty}L_n(x)t^{n+1}-x\sum_{n=0}^{+\infty}L_n(x)t^n\end{aligned} \tag{6.1.267}$$

We now equate the coefficients of t^n from both sides in the above relation to get:

$$\begin{aligned}(n+1)L_{n+1}(x)-2nL_n(x)+(n-1)L_{n-1}(x)=L_n(x)-L_{n-1}(x)-xL_n(x)\\\Rightarrow\quad(n+1)L_{n+1}(x)=(2n+1-x)L_n(x)-nL_{n-1}(x)\end{aligned} \tag{6.1.268}$$

To get the second recursion relation we start again form the generating function:

$$G(x,t)=\sum_{n=0}^{+\infty}L_n(x)t^n=\frac{e^{-xt/(1-t)}}{1-t} \tag{6.1.269}$$

We now differentiate both sides of the above equality with respect to x to get:

$$\sum_{n=0}^{+\infty}L'_n(x)t^n=\frac{1}{1-t}\cdot e^{-\frac{xt}{1-t}}\left[\frac{-t}{1-t}\right] \tag{6.1.270}$$

using the generating function that latter can be rewritten:

$$\sum_{n=0}^{+\infty}L'_n(x)t^n=\frac{-t}{1-t}\sum_{n=0}^{+\infty}L_n(x)t^n \tag{6.1.271}$$

Therefore:

$$\begin{gathered}(1-t)\sum_{n=0}^{+\infty}L'_n(x)t^n=-t\sum_{n=0}^{+\infty}L_n(x)t^n\\ \sum_{n=0}^{+\infty}L'_n(x)t^n-\sum_{n=0}^{+\infty}L'_n(x)t^{n+1}=\sum_{n=0}^{+\infty}L_n(x)t^{n+1}\end{gathered} \tag{6.1.272}$$

Equating the coefficients of t^n from both sides of the above gives:

$$\begin{gathered}L'_n(x)-L'_{n-1}(x)=-L_{n-1}(x)\\ \Rightarrow L'_{n-1}(x)=L'_n(x)+L_{n-1}(x)\end{gathered} \tag{6.1.273}$$

Therefore:

$$L'_n(x)=L'_{n+1}(x)+L_n(x) \tag{6.1.274}$$

and rearranging:

$$L'_{n+1}(x)=L'_n(x)-L_n(x) \tag{6.1.275}$$

From the first recurrence relation:

$$(n+1)L_{n+1}(x)=(2n+1-x)L_n(x)-nL_{n-1}(x) \tag{6.1.276}$$

We differentiate the above with respect to x to get:

$$(n+1)L'_{n+1}(x)=(2n+1-x)L'_n(x)-L_n(x)-nL'_{n-1}(x) \tag{6.1.277}$$

Substituting the values of $L'_{n-1}(x)$ and $L'_{n+1}(x)$ derived earlier in the previous relation, we get:

$$(n+1)\left[L'_n(x)-L_n(x)\right]=(2n+1-x)L'_n(x)-L_n(x)-n\left[L'_n(x)+L_{n-1}(x)\right] \tag{6.1.278}$$

and after simplification we get the second recurrence relation:

$$xL'_n(x)=nL_n(x)-nL_{n-1}(x) \tag{6.1.279}$$

For the proof of the orthogonality of Laguerre polynomials, the reader can refer to the section Analysis page 2386.

6.1.3.13.1 Associated Laguerre Polynomials

Definition (#6.1.178): In many applications, particularly in quantum mechanics (again!), we need the "associated Laguerre polynomials" also named "generalized Laguerre polynomials" defined by:

$$L_n^k(x) = (-1)^k \frac{\mathrm{d}^k}{\mathrm{d}x^k} L_{n+k}(x) \tag{6.1.280}$$

From the series form of $L_n(x)$ we verify that the lowest associated Laguerre polynomials are given by:

$$\begin{aligned} L_0^k(x) &= 1 \\ L_1^k(x) &= -x + k + 1 \\ L_2^k(x) &= \frac{x^2}{2} - (k+2)x + \frac{(k+2)(k+1)}{2} \end{aligned} \tag{6.1.281}$$

Therefore in general:

$$\boxed{L_n^k(x) = (-1)^k \frac{\mathrm{d}^k}{\mathrm{d}x^k} L_{n+k}(x) = \sum_{m=0}^{n} (-1)^m \frac{(n+k)!}{(n-m)!(k+m)!m!} x^m, \quad k > -1} \tag{6.1.282}$$

The first few associated Laguerre polynomials are:

n	$L_n^k(x)$
L_0^k	1
L_1^k	$-x + (k+1)$
L_2^k	$\frac{x^2}{2} - (k+2)x + \frac{(k+1)(k+2)}{2}$
L_3^k	$-\frac{x^3}{6} + \frac{(k+3)x^2}{2} - \frac{(k+2)(k+3)x}{2} + \frac{(k+1)(k+2)(+3)}{6}$
$\ldots$	$\ldots$

Table 6.6: **List of some associated Laguerre polynomials**

And a plot of some of them:

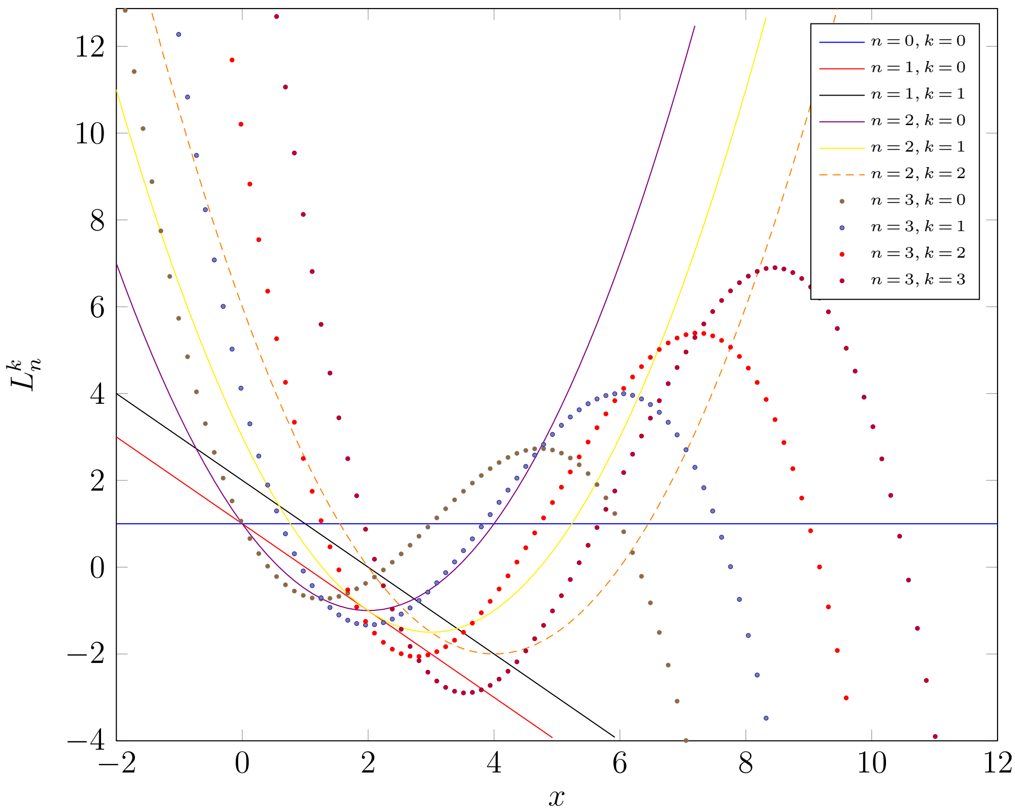


Figure 6.19: **Plot of some associated Laguerre polynomials (source: Wikipedia)**

Or using a recurrence relation for any $k \geq 1$:

$$L_{k+1}^{(\alpha)}(x)=\frac{(2k+1+\alpha-x)L_k^{(\alpha)}(x)-(k+\alpha)L_{k-1}^{(\alpha)}(x)}{k+1} \tag{6.1.283}$$

A generating function may be developed by differentiating the Laguerre generating function k times to yield:

$$\begin{aligned}(-1)^k\frac{\mathrm{d}^k}{\mathrm{d}x^k}\frac{e^{-xz/(1-z)}}{1-z}&=(-1)^k\sum_{n=0}^{+\infty}\frac{\mathrm{d}^k}{\mathrm{d}x^k}L_{n+k}(x)z^{n+k}=\sum_{n=0}^{+\infty}L_n^k(x)z^{n+k}\\&=\left(\frac{z}{1-z}\right)^k\frac{e^{xz/(1-z)}}{1-z}\end{aligned} \tag{6.1.284}$$

From the last two members of this equation, cancelling the common factor z^n, we get:

$$\frac{e^{-xz/(1-z)}}{(1-z)^{k+1}}=\sum_{n=0}^{+\infty}L_n^k(x)z^n,\quad |z|<1 \tag{6.1.285}$$

From this, for $x=0$, the binomial expansion:

$$\frac{1}{(1-z)^{k+1}}=\sum_{n=0}^{+\infty}\left(\begin{array}{c}-k-1\\n\end{array}\right)(-z)^n=\sum_{n=0}^{+\infty}L_n^k(0)z^n \tag{6.1.286}$$

yields:

$$L_n^k(0)=\frac{(n+k)!}{n!k!} \tag{6.1.287}$$

Recurrence relations can be derived from the generating function or by differentiating the Laguerre polynomial recurrence relations. Among the numerous possibilities are:

$$\begin{gathered}(n+1)L_{n+1}^k(x)=(2n+k+1-x)L_n^k(x)-(n+k)L_{n-1}^k(x)\\ x\frac{\mathrm{d}L_n^k(x)}{\mathrm{d}x}=nL_n^k(x)-(n+k)L_{n-1}^k(x)\end{gathered}\tag{6.1.288}$$

Thus, we obtain from differentiating Laguerre's ordinary differential equation once:

$$x\frac{\mathrm{d}L_n''}{\mathrm{d}x}+L_n''-L_n'+(1-x)\frac{\mathrm{d}L_n'}{\mathrm{d}x}+n\frac{\mathrm{d}L_n}{\mathrm{d}x}=0\tag{6.1.289}$$

and eventually from differentiating Laguerre's ordinary differential equation k times:

$$x\frac{\mathrm{d}^k}{\mathrm{d}x^k}L_n''+k\frac{\mathrm{d}^{k-1}}{\mathrm{d}x^{k-1}}L_n''-k\frac{\mathrm{d}^{k-1}}{\mathrm{d}x^{k-1}}L_n'+(1-x)\frac{\mathrm{d}^k}{\mathrm{d}x^k}L_n'+n\frac{\mathrm{d}^k}{\mathrm{d}x^k}L_n=0\tag{6.1.290}$$

Adjusting the index $n\rightarrow n+k$, we have the "associated Laguerre ordinary differential equation":

$$x\frac{\mathrm{d}^2L_n^k(x)}{\mathrm{d}x^2}+(k+1-x)\frac{\mathrm{d}L_n^k(x)}{\mathrm{d}x}+nL_n^k(x)=0\tag{6.1.291}$$

Most known under the following form:

$$\boxed{xy''+(\alpha+1-x)y'+ny=0}\tag{6.1.292}$$

When associated Laguerre polynomials appear in a physical problem it is usually because that physical problem involves the above differential equation. The most important application is the bound states of the hydrogen atom, which are derived from it! A "Rodrigues representation of the associated Laguerre polynomial":

$$\boxed{L_n^k(x)=\frac{e^x x^{-k}}{n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left(e^{-x}x^{n+k}\right)}\tag{6.1.293}$$

may be obtained by:

$$\begin{aligned}L_n^k(x)=&\frac{e^x x^{-k}}{n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left(e^{-x}x^{n+k}\right)=&(-1)^n\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left[L_{n+k}(x)\right]\\ =&\sum_{m=0}^{+\infty}(-1)^m\frac{(n+k)!}{(n-m)!(k+m)!m!}x^m\end{aligned}\tag{6.1.294}$$

The associated Laguerre polynomials shows the same orthogonality properties as for the Laguerre polynomials!

Let us now derive a useful result again for our study of Quantum Chemistry: the normalization constant of the associated Laguerre polynomial!

Using Rodrigue's formula for the associated Laguerre polynomial we get:

$$L_n^k(x)=\frac{x^{-k}e^x}{n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left(x^{n+k}e^{-x}\right)\tag{6.1.295}$$

Therefore:

$$\begin{aligned}\int\limits_0^{+\infty} x^k e^{-x}\left(L_n^k(x)\right)^2\mathrm{d}x&=\frac{1}{n!}\int\limits_0^{+\infty} L_n^k(x)\frac{\mathrm{d}^n}{\mathrm{d}x^n}(x^{n+k}e^{-x})\mathrm{d}x\\&=\frac{(-1)^m}{n!}\int\limits_0^{+\infty}\left(\frac{\mathrm{d}^m}{\mathrm{d}x^m}L_n^k(x)\right)\left(\frac{\mathrm{d}^{n-m}}{\mathrm{d}x^{n-m}}(x^{n+k}e^{-x})\right)\mathrm{d}x\end{aligned}\tag{6.1.296}$$

for $m\leqslant n$ (using induction on m and integration by parts). Taking $m=n$ (recall that $L_n^k(x)$ is a polynomial of degree n with the leading coefficient equal to $(-1)^n/n!$), we get the needed:

$$\int\limits_0^{+\infty} x^k e^{-x}\left(L_n^k(x)\right)^2\mathrm{d}x=\frac{1}{n!}\int\limits_0^{+\infty} x^{n+k}e^{-x}\mathrm{d}x=\frac{(n+k)!}{n!}\tag{6.1.297}$$

6.1.3.13.2 Laguerre functions

By letting:

$$\psi_n^k(x)=e^{-x/2}x^{k/2}L_n^k(x)\tag{6.1.298}$$

$\psi_n^k(x)$ satisfies the self-adjoint ordinary differential equation:

$$x\frac{\mathrm{d}^2\psi_n^k(x)}{\mathrm{d}x^2}+\frac{\mathrm{d}\psi_n^k(x)}{\mathrm{d}x}+\left(-\frac{x}{4}+\frac{2n+k+1}{2}-\frac{k^2}{4x}\right)\psi_n^k(x)=0\tag{6.1.299}$$

The $\psi_n^k(x)$ are sometimes named "Laguerre functions".

A further useful form is given by defining (this corresponds to modifying the function $\psi_n^k(x)$ to eliminate the first derivative in the previous ordinary differential equation):

$$\Phi_n^k(x)=e^{-x/2}x^{(k+1)/2}L_n^k(x)\tag{6.1.300}$$

Substitution into the associated Laguerre equation yields (corresponding to the dimensionless ordinary differential equation of the radial term of the non-rigid rotator that we meet during our study of the hydrogenoid atom in the section of Quantum Chemistry):

$$\boxed{\frac{\mathrm{d}^2\Phi_n^k(x)}{\mathrm{d}x^2}+\left(-\frac{1}{4}+\frac{2n+k+1}{2x}-\frac{k^2-1}{4x^2}\right)\Phi_n^k(x)=0}\tag{6.1.301}$$

Let us prove that claim! To attain the second derivative, we need the first derivative, and use the notation:

$$y=e^{-x/2}x^{(k+1)/2}v\tag{6.1.302}$$

where $v=L_j^k(x)$, because the indices do not change and only serve to add clutter, and we can remember the independent variable is x. The first derivative is then:

$$\begin{aligned}y'&=-\frac{1}{2}e^{-x/2}x^{(k+1)/2}v+e^{-x/2}\left(\frac{k+1}{2}\right)x^{(k-1)/2}v+e^{-x/2}x^{(k+1)/2}v'\\&=\left[-\frac{1}{2}v+\left(\frac{k+1}{2x}\right)v+v'\right]e^{-x/2}x^{(k+1)/2}\\&\Rightarrow\quad\left(e^{x/2}x^{-(k+1)/2}\right)y'=-\frac{1}{2}v+\frac{k+1}{2x}v+v'\end{aligned}\tag{6.1.303}$$

In a similar way we get:

$$\left(e^{x/2}x^{-(k+1)/2}\right)y''=\frac{1}{4}v-\frac{1}{2}\frac{k+1}{2x}v-\frac{1}{2}v'-\frac{1}{2}\frac{k+1}{2x}v+\frac{k+1}{2x}\frac{k-1}{2x}v+\frac{k+1}{2x}v'-\frac{1}{2}v'+\frac{k+1}{2x}v'+v'' \tag{6.1.304}$$

By substituting the second derivative and the function in the associated Laguerre ordinary differential equation:

$$y''+\left(-\frac{1}{4}+\frac{2j+k+1}{2x}-\frac{k^2-1}{4x^2}\right)y=0 \tag{6.1.305}$$

and dividing by the common factor of $e^{-x/2}x^{(k+1)/2}$, the remaining terms are:

$$\begin{aligned}
&\left(\frac{1}{4}v-\frac{1}{2}\frac{k+1}{2x}v-\frac{1}{2}v'-\frac{1}{2}\frac{k+1}{2x}v+\frac{k+1}{2x}\frac{k-1}{2x}v+\frac{k+1}{2x}v'-\frac{1}{2}v'+\frac{k+1}{2x}v'+v''\right)\\
&\qquad\qquad+\left(-\frac{1}{4}+\frac{2j+k+1}{2x}-\frac{k^2-1}{4x^2}\right)v=0\\
&\Rightarrow v''+\cancel{\frac{1}{4}}v-\frac{1}{2}\frac{k+1}{2x}v-\frac{1}{2}v'-\frac{1}{2}\frac{k+1}{2x}v+\cancel{\frac{k^2+1}{4x^2}}v+\frac{k+1}{2x}v'-\frac{1}{2}v'+\frac{k+1}{2x}v'\\
&\qquad\qquad-\cancel{\frac{1}{4}}v+\frac{2j+k+1}{2x}v-\cancel{\frac{k^2+1}{4x^2}}v=0\\
&\Rightarrow v''-\cancel{\frac{k+1}{4x}}v-\frac{1}{2}v'-\cancel{\frac{k+1}{4x}}v+\frac{k+1}{x}v'-\frac{1}{2}v'+\frac{j}{x}v+\cancel{\frac{k+1}{2x}}v=0\\
&\Rightarrow v''-v'+\frac{k+1}{x}v'+\frac{j}{x}=0\\
&\Rightarrow xv''-xv'+(k+1)v'+jv=0\\
&\Rightarrow xv''+(1-x+k)v'+jv=0
\end{aligned} \tag{6.1.306}$$

which is the associated Laguerre equation. Since $v=L_j^k(x)$, and the $L_j^k(x)$ are solutions to the associated Laguerre equation. The above last line is equivalent to:

$$xL_j^{k''}(x)+(1-x+k)L_j^{k'}(x)+jL_j^k(x)=0 \tag{6.1.307}$$

which is the associated Laguerre equation, which we know to be a true statement, so:

$$y=e^{-x/2}x^{(k+1)/2}v=e^{-x/2}x^{(k+1)/2}L_j^k(x) \tag{6.1.308}$$

are solution to it! The proof is finished.

The corresponding normalization integral:

$$\int\limits_0^{+\infty}\left|\Phi_n^k(x)\right|^2\mathrm{d}x \tag{6.1.309}$$

is (without proof):

$$\int\limits_0^{+\infty}e^{-x}x^{k+1}\left[L_n^k(x)\right]^2\mathrm{d}x=\frac{(n+k)!}{n!}(2n+k+1) \tag{6.1.310}$$

Notice that the $\Phi_n^k(x)$ do not form an orthogonal set (except with x^{-1} as a weighting function) because of the x^{-1} in the term $(2n+k+1)/2x$.

To close this section, here is a summary of all polynomials we have seen so far and also of other polynomials we will prove later through this entire book:

Name	**Formulation**
Vanilla polynomials	$f(x)=\sum_n a_x x^n$
Trigonometric polynomials	$S_N(\theta)=\sum_{n=-N}^{N} c_n e^{in\theta}$
Cyclotomic polynomials	$\Phi_n(x)=\prod_n \left(x-e^{\frac{2ik\pi}{n}}\right)$
Legendre polynomials	$P_n(x)=\frac{(-1)^n}{n!2^n}\frac{\mathrm{d}^n}{\mathrm{d}x^n}(1-x^2)^n$
Laguerre polynomials	$L_n(x)=\sum_{\lambda=0}^{n}\frac{(-1)^\lambda n!}{(\lambda!)^2(n-\lambda)!}x^\lambda=\sum_{\lambda=0}^{n}\left(\begin{array}{c} n \\ \lambda \end{array}\right)\frac{(-x)^\lambda}{\lambda!}$
Associated Laguerre polynomials	$L_n^k(x)=\sum_{m=0}^{n}(-1)^m\frac{(n+k)!}{(n-m)!(k+m)!m!}x^m,\quad k>-1$
Bessel polynomials	$y_n(x)=\sum_{k=0}^{n}\frac{(n+k)!}{(n-k)!k!}\left(\frac{x}{2}\right)^k$
Hermite polynomials	$H_n(Q)=(-1)^n e^{Q^2}\frac{\mathrm{d}^n}{\mathrm{d}Q^n}e^{-Q^2}$
Bernstein polynomials	$B_{i,n}(t)=\frac{(n-1)!}{(1-i)!(n-i)!}t^{i-1}(1-t)^{n-i}=C_{i-1}^{n-1}t^{i-1}(1-t)^{n-i}$
Lagrange polynomials	$P(x)=\sum_{j=0}^{n}\left(\prod_{i=0,i\neq j}^{n}\frac{x-x_i}{x_j-x_i}\right)f(x_j)$
Newton polynomial	$N_n(x)=f[x_0]+\sum_{i=1}^{n}f[x_0,\ldots,x_i]\left(\prod_{j=0}^{i-1}(x-x_j)\right)$

Table 6.7: **Most common polynomials**

~90 %

★★★☆☆

70 votes, 56.29%

6.2. Set Algebra

We will approach in this textbook the study of set structures very pragmatically (since you must remember that this textbook is dedicated to engineers). Thus, it will be made use of the minimum of formalism and only the proofs of the elements that we consider as absolutely essential to the engineer will be presented in this. Moreover, numerous proofs will be made by example and we will focus largely on the algebraic theory of groups as it has a prominent place in physics almost more than for other set-structures.

6.2.1 Groups Algebra and Geometry

The symmetries of geometric figures, of crystals and all the other items of macroscopic physics, are subject for centuries of observations and studies. In modern terms, the symmetries of a given object form a "group".

Since the mid-119th century (holocene calendar), the group theory took a huge extension, and its applications to quantum mechanics and the theory of elementary particles have developed throughout the 120th century (holocene calendar).

In a letter of 11877 (holocene calendar) to the mathematician Adolf Meyer, Sophus Lie wrote that he created the Theory of Groups in January 11873 (holocene calendar). It is of the groups he named "continuous groups" and that are named today "Lie groups". Lie sought to expand the use of groups from the domain of algebraic equations, where Galois had introduced them, to the differential equations.

Since 11871 (holocene calendar), the notion of infinitesimal generator of a one-parameter group of transformations appeared in his work. This is the set of infinitesimal generators of the subgroups with one parameter of a continuous group that forms what we name today a "Lie algebra".

It was Weyl and Wigner that show the pre-eminent role of the theory of groups and their representations in particular in the new quantum mechanics that Heisenberg and Dirac were developing. The general idea of representation theory is try to study a group by doing it acting on a vector space in a linear way: we try to see the group as a group of matrices (hence the term "representation "). We can, from relatively well-known properties of the automorphism group of the vector space (see section Set Theory page 443), get to deduce some group properties of main interest.

We can consider the theory of group representations as a vast generalization of the Fourier analysis. Its development is continuous and has, since the mid-120th century (holocene calendar), countless applications in differential geometry, ergodic theory, probability theory, number theory, the theory of automorphic forms, in that of dynamic systems as well as in physics, chemistry, molecular biology and signal processing. Currently, entire branches of mathematics and physics depend on it.

Before we begin, we refer the reader to the section on Set Theory to get a refresh of the structure and fundamental properties that make up a Group and also to the section of Linear Algebra (because we use some results proved in it).

6.2.1.1 Cyclic Groups

The cyclic group (whose definition of has already been given in the section of Set Theory) will serve us a basis for the study of finite groups. Moreover, rather than making generalized expansions we preferred to take specific examples to present the idea of cyclic group (more suitable for an engineering approach).

We will take then the very nice example of the hours of a clock... with three different approaches that will successively (!) and simple address the concept of a cyclic group.

1. First approach:

   Let us imagine a clock with a needle which can take 12 possible positions (but no intermediate positions). We will denote in a special way the 12 possible positions: $\overline{0},\overline{1},\overline{2},\ldots,\overline{11}$ (the line above the numbers is not innocent!).

   Nothing prevents us on all of these positions to define an addition, for example:

   $$\overline{2}+\overline{3}=\overline{5},\;\overline{3}+\overline{2}=\overline{5},\;\overline{5}+\overline{7}=\overline{0},\;\overline{9}+\overline{5}=\overline{2},\;\overline{1}-\overline{8}=\overline{5},\tag{6.2.1}$$

   which is similar to the results we get when we in our daily life we do calculations manipulating time.

2. Second approach (first extension):

   If we observe well a Watch or a Clock, we notice that every time we add 12 (or withdraw 12...) to a value of hours of our Watch/Clock then we fall back on the set of numbers that are well defined that are also in $\mathbb{Z}$. Therefore (obviously as part of a Watch/Clock only the first positive values are of interest to us most of the time but here we do math so we generalize a bit...):

   $$\begin{aligned}\overline{0}&:=0+12\mathbb{Z}=\{\ldots,-24,-12,0,12,24,\ldots\}\\\overline{1}&:=1+12\mathbb{Z}=\{\ldots,-11,1,13,25,\ldots\}\\\overline{2}&:=2+12\mathbb{Z}=\{\ldots,-10,2,14,26,\ldots\}\end{aligned}\tag{6.2.2}$$

   Here we fall back on a concept we had already seen in the section of Numbers Theory. This are congruence classes and all of these classes form the quotient set $\mathbb{Z}/12\mathbb{Z}$. If we endow this quotient set of an addition law, it is normally easy to observe that it is an internal law to the quotient set, that it is associative, that there exist a neutral element and that every element has an inverse (opposite).

   Thus, this quotient set equipped only with the addition law (if adding the multiplication we can form a ring) is a commutative group.

3. Third approach (second and last extension):

   Let us see a third and final approach which is why the quotient group is cyclic.

   If we project the rotation of the hands of our Watch/Clock (all rotations in the set-algebra are traditionally clockwise!) in $\mathbb{C}$ and that we define:

$$x := e^{\frac{2\pi \mathrm{i}}{12}} \tag{6.2.3}$$

   We then have $x^{12} = x^0 = 1$ and:

$$\forall a,k \in \mathbb{Z}: \qquad x = x^{a+12k} \tag{6.2.4}$$

   which is why the quotient group $(\mathbb{Z}/12\mathbb{Z},+)$ is named "cyclic group" (by group isomorphism according to what has been seen in the section Set Theory). Its isomorphic is noted by C_{12} and all elements are of modulus 1. It is common to denote all complex number of modulus 1 as following:

$$\mathbb{U} = \{z \in \mathbb{C} \| z| = 1\} \tag{6.2.5}$$

   If we represent in $\mathbb{C}$ the isomorphic set C_{12} we get then on the unit circle a polygon with n vertices as shown in the figure below:

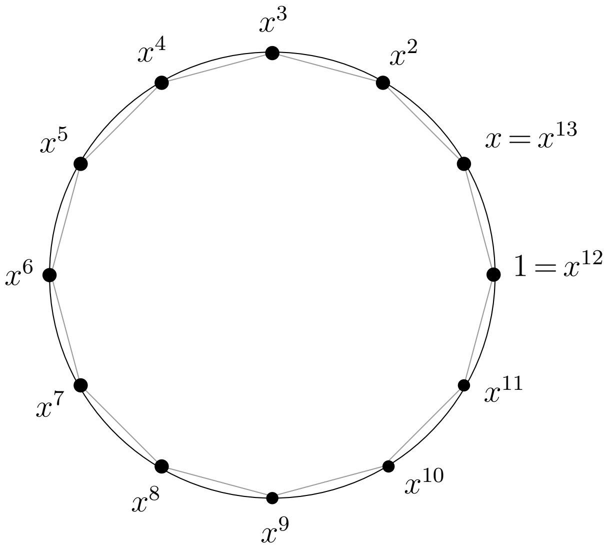


Figure 6.20: C_{12} **cyclic group**

   Furthermore, the number of component elements of $\mathbb{Z}/12\mathbb{Z}$ being finite, $(\mathbb{Z}/12\mathbb{Z},+)$ is finite. Unlike the group $(\mathbb{Z},+)$ which is itself a discrete infinite group.

   This concept of finitude is perhaps most obvious with the example that we will do afterwards with $\mathbb{Z}/4\mathbb{Z}$ where the reader will observe that this set has the same number of elements than C_4.

Remark

Mathematicians name C_n the "group of n-th roots of unity". An n-th root of the unit (sometimes named "De Moivre's number") is a complex number whose n-th power is equal to 1. Moreover, for a given integer n, all n-th roots of unit are located on the unit circle and are the vertices of a regular n-sided polygon having an apex of affix 1.

Definition (#6.2.179): A "finite group" is a mathematical group with a finite number of elements like "permutations groups", "symmetric groups", "cyclic groups", etc.

What interests particularly physicists at first are the representations of finite groups (also the continuous groups that we will see later). Thus, the representative of $\mathbb{Z}/n\mathbb{Z}$ is known to us as the rotation in the complex plane is given as we have shown during our study of complex numbers in the section Numbers:

$$\begin{bmatrix} \cos\left(\dfrac{2k\pi}{n}\right) & -\sin\left(\dfrac{2k\pi}{n}\right) \\ \sin\left(\dfrac{2k\pi}{n}\right) & \cos\left(\dfrac{2k\pi}{n}\right) \end{bmatrix} \tag{6.2.6}$$

with $k\in[0,n-1]$. This representative is a subgroup of the group of rotations $\text{O}(2)$ which will be discussed further. The group of rotations of the plane itself being a subgroup of the linear group $\text{GL}(2)$ (we will give a precise definition and an example further below).

In fact, mathematicians are able to prove that all quotients groups $\mathbb{Z}/n\mathbb{Z}$ are cyclic to an isomorphism with C_n and then they say that $\mathbb{Z}/n\mathbb{Z}$ is a finite quotient of the monogenic group $\mathbb{Z}$...

This approach is perhaps a bit abstract for the Padawan... So if the reader remembers the section Set Theory we saw a very precise definition of what is the cyclicity of a group: A group G is said to be cyclic if G is generated by the power of at least one of its elements $a\in G$ named the "generator" such that:

$$G=\{a^n,n\in\mathbb{Z}\} \tag{6.2.7}$$

Let us check whether this is the case for the group:

$$G=\{\mathbb{Z}/4\mathbb{Z},+\} \tag{6.2.8}$$

which is a school case.

We will denote the elements that make up this finite group:

$$\{0,1,2,3\} \tag{6.2.9}$$

This being done, we should be careful that in the set definition of a cyclic group we speak of "power" if the internal law of the group is multiplication but if the internal law is the addition, then we have:

$$G=\{n\cdot a,n\in\mathbb{Z}\} \tag{6.2.10}$$

The first generator of the group $G=\{\mathbb{Z}/4\mathbb{Z},+\}$ is 1. Indeed:

$$\begin{aligned} &1=1\cdot 1=1\\ &1+1=2\cdot 1=2\\ &1+1+1=3\cdot 1=3\\ &1+1+1+1=4\cdot 1=1 \end{aligned} \tag{6.2.11}$$

The second generating element of the same group is 3:

$$\begin{aligned}3&=1\cdot 3=3\\3+3&=2\cdot 3=2\\3+3+3&=3\cdot 3=1\\3+3+3+3&=4\cdot 3=0\end{aligned}\qquad(6.2.12)$$

For cons, the reader can check that 2 is not a generator of this group!

In fact, regarding the groups $G=\{\mathbb{Z}/n\mathbb{Z},+\}$ mathematicians are able to prove that only the elements of the group items that are prime with n are generators (that is to say elements whose greatest common divisor is 1).

Remark

Somewhat similar to the way prime numbers are the basic building blocks of the integers, there exist some special kind of finite groups that are the basic building groups of all finite common groups. These special kind of groups are named "finite simple groups". The classification of the finite, simple groups is unprecedented in the history of mathematics, for its proof is $15,000$ pages long, needed almost 50 years to be completed and involved a few hundred mathematicians. That's why it's named the "Enormous Theorem".

That's all for our introduction to cyclic groups for engineers. Let us turn now to another group category.

6.2.1.2 Transformations Groups

The transformations groups, that include the rotations groups, are the one that most interests physicists especially in the areas of continuum mechanics, chemistry, quantum physics and art ... Mathematicians appreciate obviously the study of rotations groups in the context of geometry (but not only!) and computer scientists equally linear groups. We have also seen an example of a rotations group just before.

Definition (#6.2.180): We name "linear group of order n" and we note $\text{GL}(n)$ the invertible matrices $n\times n$ or also named "regular matrices" (i.e. their determinant is not zero according to what we saw in the section Linear Algebra) which component are in any field or ring $\mathbb{K}$ (the ring $\mathbb{R}$ or the group $\mathbb{C}$ most of time):

$$\boxed{\text{GL}(n)=\text{GL}(n,\mathbb{K})=\text{GL}_n(\mathbb{K})=\{A\in M_n|\det(A)\neq 0\}}\qquad(6.2.13)$$

The group is so named because the columns of an invertible matrix are linearly independent.

We will consider here as obvious that $\text{GL}(n)$ is a group: the matrix multiplication is associative array and each matrix of $\text{GL}(n)$ has an inverse by definition (as the determinant is not null). On the other hand, the product of two regular matrices gives a regular matrix that is again invertible.

A simple and important example of linear group is that of the sub-"group of affine transformations" of the plane that is traditionally noted (that is intuitive):

$$F:M(x,y)\rightarrow M'(x',y'):\begin{cases}x'=ax+by+\alpha\\y'=cx+dy+\beta\end{cases}\qquad(6.2.14)$$

with $a,b,c,d,\alpha,\beta \in \mathbb{R}$, $ad-bc \neq 0$ (we will see why and how this latter inequality a little further below).

Definition (#6.2.181): The "affine group" or "general affine group" of any affine space A over a field $\mathbb{K}$ is the group denoted $\text{Aff}(A)$ or $\text{Aff}(n,\mathbb{K})$ of all invertible affine transformations from the space into itself. It is a "Lie group" if $\mathbb{K}$ is the real or complex field or quaternions.

Let us take a small practical example:

$$F:M(x,y)\rightarrow M'(x',y'):\begin{cases} x'=0.5x-0.4y+0.5 \\ y'=0.8x+0.2y-1.5 \end{cases} \tag{6.2.15}$$

whic applied to a circle gives:

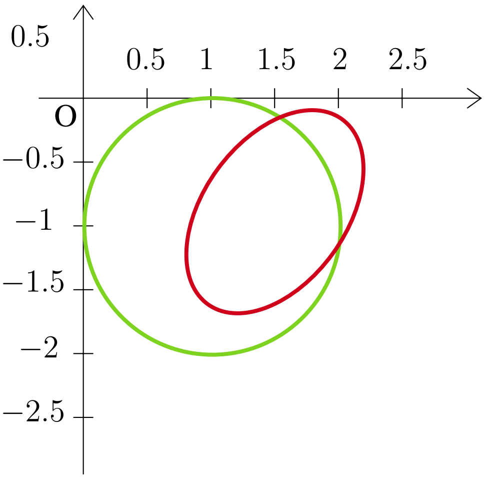


Figure 6.21: **Simple affine transformation on a circle**

This transformation is a way to define the ellipses as images of a circle by an affine transformation.

The coefficients α,β are irrelevant for the shape of the image figure. In fact, they obviously induce only translations to the figures. So we can do without them if we only seek to distort the original figure.

Therefore it remains:

$$F:M(x,y)\rightarrow M'(x',y'):\begin{cases} x'=ax+by \\ y'=cx+dy \end{cases} \tag{6.2.16}$$

which can be written in matrix form:

$$\begin{pmatrix} x' \\ y' \end{pmatrix}=\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{pmatrix} x' \\ y' \end{pmatrix} \tag{6.2.17}$$

The affine transformation is therefore reduced to the matrix:

$$F=\begin{bmatrix} a & b \\ c & d \end{bmatrix} \tag{6.2.18}$$

and as we have seen in the section of Linear Algebra, matrix multiplication is associative but is not commutative, so the linear transformation is not either.

The neutral element is the matrix:

$$I=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} \tag{6.2.19}$$

and the inverse of F is:

$$F^{-1}=\begin{bmatrix}\dfrac{d}{ad-bc} & \dfrac{-b}{ad-bc}\\ \dfrac{-c}{ad-bc} & \dfrac{a}{ad-bc}\end{bmatrix}=\dfrac{1}{ad-bc}\begin{bmatrix}d & -b\\ -c & a\end{bmatrix} \tag{6.2.20}$$

and as we have imposed $ad-bc\neq 0$ any element has therefore an inverse. Thus, the linear affine group is not commutative and... is therefore a group...

As we will see it, all the "classics" Lie groups are subgroups of GL(n).

Definition (#6.2.182): We name "special linear group of order n" and denote by SL(n) the invertible (square $n\times n$) matrices with coefficients in an arbitrary field and with determinant equal to unity:

$$\boxed{\text{SL}(n)=\{A\in M_n|\det(A)=1\}} \tag{6.2.21}$$

This is obviously a subgroup of GL(n).

Returning to the previous example and remembering that the determinant of a square two-dimensional matrix is (see section Linear Algebra 1961):

$$\det(F)=ad-bc=1=\det(F^{-1}) \tag{6.2.22}$$

we notice geometrically well what it means to have a unit determinant in this case! Indeed we will see in the section of Linear Algebra during our geometric interpretation of the determinant that having a determinant is equivalent to have a surface. Thus, having $ad-bc$ equal to the unit allows us regardless of the processing order, to have the area equal to 1. Thus, the special linear is a group of transformation that keeps surfaces value.

Definition (#6.2.183): We name "orthogonal real group of order n" and denote by O(n) the invertible orthogonal (square $n\times n$) matrices (see the section Linear Algebra for a refresh of what are orthogonal matrices):

$$\boxed{\text{O}(n)=\{A\in M_n(\mathbb{R})|A^TA=I_n\}} \tag{6.2.23}$$

Furthermore, we proved in the section Linear Algebra in our study rotation matrices that $A^TA=I_n$ implies that $\det(A)=\pm 1$.

This is the case for example of the O(2) matrix seen previously (it belongs to the orthogonal group but also to the group of rotations that we will see further below):

$$\begin{bmatrix}\cos\left(\dfrac{2k\pi}{n}\right) & -\sin\left(\dfrac{2k\pi}{n}\right)\\ \sin\left(\dfrac{2k\pi}{n}\right) & \cos\left(\dfrac{2k\pi}{n}\right)\end{bmatrix} \tag{6.2.24}$$

which is orthogonal as it is easy to check (just multiple by the transposed to see if you get and identity matrix).

Remark

O(1) is made up of all trivial matrices... $[1][-1]$ that are simply one component vectors or just... simple scalars.

Definition (#6.2.184): If $A\in \text{O}(n)$ and that we have $\det(A)=1$ then we get a subgroup of O(n) named "special real group orthogonal of order n" and therefore defined by:

$$\boxed{\text{SO}(n)=\{A\in M_n(\mathbb{R})|A^TA=I_n,\det(A)=1\}\subset \text{O}(n)} \tag{6.2.25}$$

The rotation matrix given previously is part of this group since its determinant is equal to unity! Furthermore, this group occupies a very special place in physics and we will meet it again during our study of quantum physics.

The subgroup SO(2), also sometimes named "circle group" and denoted S^1 that we had also studied in the section of Euclidean Geometry has a representative given by the matrix:

$$R(\theta)=\begin{bmatrix}\cos(\theta) & -\sin(\theta)\\ \sin(\theta) & \cos(\theta)\end{bmatrix} \tag{6.2.26}$$

and occupies a special place in the family of the SO(n) groups with n greater than unity. Indeed it is the only one that is commutative. Moreover, it is isomorphic to $e^{\mathrm{i}\theta}$ either U(1) the multiplicative group of complex numbers of modulus 1. This is also the proper symmetry group of a circle and the equivalent continuous of C_n.

The subgroup SO(3) is given by the matrix (see section Euclidean Geometry page 2650):

$$R_X(\theta)=\begin{bmatrix}1 & 0 & 0\\ 0 & \cos(\theta) & \sin(\theta)\\ 0 & -\sin(\theta) & \cos(\theta)\end{bmatrix} \tag{6.2.27}$$

for the rotation around the x-axis in the three-dimensional space is not commutative (the rotation matrices in the plane being commutative for recall!). Moreover the quaternions that we have seen in the section Numbers, whose representative is therefore SO(3), form also a non-commutative group (relatively to the multiplication law) as we have seen in the section Numbers.

Compared to a unit vector makes we can relatively easily visually speaking see that SO(3) is a closed subgroup of GL(3), that is to say, the set of linear groups of dimension 3.

Remark

SO(1) consists in the matrix $[1]$.

Definition (#6.2.185): We name "unitary linear group of order n" and we denote it U(n) the matrices whose components are complex (as part of this textbook the most often) or real and which are orthogonal:

$$\boxed{\text{U}(n)=\{A\in M_n(\mathbb{C})|A^\dagger A=I_n\}} \tag{6.2.28}$$

Notice also that any unitary matrix with complex components and of one dimension (thus that belongs to $\text{U}(n)$...) is a complex number of unit module, which can always be written in the form $e^{\text{i}\mathbb{R}}$.

We have already seen an example in this textbook during our study of spinors (see section of the same name page 2145). These are the Pauli matrices (used in the section of Relativistic Quantum Physics) given by:

$$\sigma_x=\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix},\sigma_y=\begin{bmatrix}0 & -\text{i}\\ \text{i} & 0\end{bmatrix},\sigma_z=\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix} \tag{6.2.29}$$

Definition (#6.2.186): We name "special unitary group of order n" and we denote by $\text{SU}(n)$ the matrices whose entries are complex and are orthogonal and whose determinant is unity:

$$\boxed{\text{SU}(n)=\{A\in M_n(\mathbb{C})|A_\dagger A=I_n,\det(A)=1\}\subset\text{U}(n)} \tag{6.2.30}$$

Remark

$\text{U}(1)$ is equal to $\text{SU}(1)$ and it is therefore the complex unit circle equals to $e^{\text{i}\mathbb{R}}$. Moreover, $\text{SO}(2)$ is commutative and isomorphic to $\text{U}(1)$ because it is the set of the rotations of the plane.

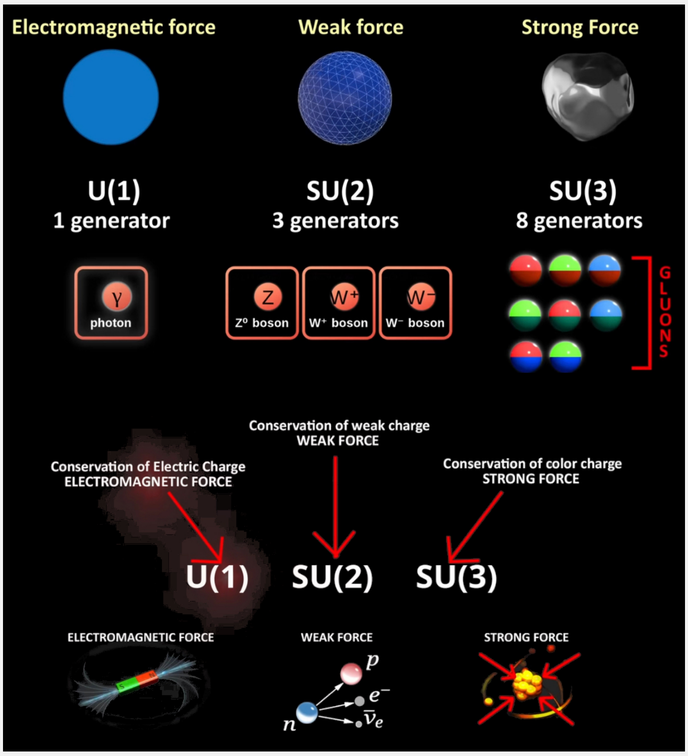
A well known example is still the one of the Pauli matrices but simply written in the form used in Relativistic Quantum Physics (see section of the same name page 4937):

$$\text{i}\sigma_x=\begin{bmatrix}0 & \text{i}\\ \text{i} & 0\end{bmatrix},\text{i}\sigma_y=\begin{bmatrix}0 & 1\\ -1 & 0\end{bmatrix},\text{i}\sigma_z=\begin{bmatrix}\text{i} & 0\\ 0 & -\text{i}\end{bmatrix} \tag{6.2.31}$$

which are part of $\text{SU}(2)$ and as we have shown (implicitly) at the beginning of the section of Spinor Calculus isomorphic to the quaternion group $\text{SO}(3)$ of module 1. Relations that mathematicians name in the present situation a "homomorphism's overlay"...

Remark

The special unitary group has a particular importance in particle physics. If the unitary group $\text{U}(1)$ is the electromagnetic gauge group (think to the complex number appearing in the solutions of the wave equation!), $\text{SU}(2)$ is the group associated with the weak interaction, and $\text{SU}(3)$ the one of the strong interaction. This is for example due to the structure of representations of $\text{SU}(3)$ that Gell-Mann conjectured the existence of quarks!


Electromagnetic force
Weak force
Strong Force
U(1)
1 generator
SU(2)
3 generators
SU(3)
8 generators
γ
photon
Z
Z° boson
W+
W⁺ boson
W−
W⁻ boson
GLUONS
Conservation of weak charge
WEAK FORCE
Conservation of Electric Charge
ELECTROMAGNETIC FORCE
Conservation of color charge
STRONG FORCE
U(1)
SU(2)
SU(3)
ELECTROMAGNETIC FORCE
WEAK FORCE
STRONG FORCE
p
e⁻
ν̄e
n

Let us see with a different approach to that used in the section of Spinor Calculus how to show that the Pauli matrices are the bases of $\text{SU}(2)$?

First, the reader has to know that we proved in the section of Spinor Calculus that any rotation in space of three dimensions could be expressed using the relation (for small angles!):

$$R\left(\vec{L},\frac{\theta}{2}\right)=\cos\left(\frac{\theta}{2}\right)\sigma_0-\mathrm{i}\sin\left(\frac{\theta}{2}\right)\vec{L}\circ\vec{\sigma} \tag{6.2.32}$$

And we saw in the section of Quantum Computing that an explicit decomposed formulation of the previous relation was:

$$\begin{aligned}R_x(\theta)&=\begin{bmatrix}\cos\left(\dfrac{\theta}{2}\right) & -\mathrm{i}\sin\left(\dfrac{\theta}{2}\right)\\ -\mathrm{i}\sin\left(\dfrac{\theta}{2}\right) & \cos\left(\dfrac{\theta}{2}\right)\end{bmatrix}\\ R_y(\theta)&=\begin{bmatrix}\cos\left(\dfrac{\theta}{2}\right) & -\sin\left(\dfrac{\theta}{2}\right)\\ \sin\left(\dfrac{\theta}{2}\right) & \cos\left(\dfrac{\theta}{2}\right)\end{bmatrix}\\ R_z(\theta)&=\begin{bmatrix}e^{-\mathrm{i}\theta/2} & 0\\ 0 & e^{\mathrm{i}\theta/2}\end{bmatrix}\end{aligned} \tag{6.2.33}$$

and therefore that all $\text{SU}(2)$ element is produced from these three matrices which each make describe to the end of a vector a curve on the surface of a sphere!

Now, we notice that these three matrices are equal to when $\theta=0$:

$$R_x(\theta)=R_y(\theta)=R_z(\theta)=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} \tag{6.2.34}$$

We then obtain the identity matrix. So if we search the tangent at this point we can therefore build a base on it (three orthogonal vectors).

Let's look at this:

$$\begin{aligned}
\left.\frac{\mathrm{d}R_x(\theta)}{\mathrm{d}\theta}\right|_{\theta=0} &= \begin{bmatrix} -\dfrac{1}{2}\sin\left(\dfrac{\theta}{2}\right) & -\mathrm{i}\dfrac{1}{2}\cos\left(\dfrac{\theta}{2}\right) \\ -\mathrm{i}\dfrac{1}{2}\cos\left(\dfrac{\theta}{2}\right) & -\dfrac{1}{2}\sin\left(\dfrac{\theta}{2}\right) \end{bmatrix}\Bigg|_{\theta=0} = \begin{bmatrix} 0 & -\dfrac{1}{2}\mathrm{i} \\ -\dfrac{1}{2}\mathrm{i} & 0 \end{bmatrix} = -\frac{\mathrm{i}}{2}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = -\frac{\mathrm{i}}{2}\sigma_x \\
\left.\frac{\mathrm{d}R_y(\theta)}{\mathrm{d}\theta}\right|_{\theta=0} &= \begin{bmatrix} -\dfrac{1}{2}\sin\left(\dfrac{\theta}{2}\right) & -\dfrac{1}{2}\cos\left(\dfrac{\theta}{2}\right) \\ \dfrac{1}{2}\cos\left(\dfrac{\theta}{2}\right) & -\dfrac{1}{2}\sin\left(\dfrac{\theta}{2}\right) \end{bmatrix}_{\theta=0} = \begin{bmatrix} 0 & -\dfrac{1}{2} \\ \dfrac{1}{2} & 0 \end{bmatrix} = -\frac{1}{\mathrm{i}2}\begin{bmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{bmatrix} = -\frac{\mathrm{i}}{2}\sigma_y \\
\left.\frac{\mathrm{d}R_z(\theta)}{\mathrm{d}\theta}\right|_{\theta=0} &= \begin{bmatrix} -\dfrac{1}{2}\mathrm{i}e^{-\mathrm{i}\theta/2} & 0 \\ 0 & \dfrac{1}{2}\mathrm{i}e^{\mathrm{i}\theta/2} \end{bmatrix}_{\theta=0} = \begin{bmatrix} -\dfrac{1}{2}\mathrm{i} & 0 \\ 0 & \dfrac{1}{2}\mathrm{i} \end{bmatrix} = -\frac{\mathrm{i}}{2}\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} = -\frac{\mathrm{i}}{2}\sigma_z
\end{aligned} \tag{6.2.35}$$

Thus, $\mathrm{SU}(2)$ has for basis:

$$\left\{-\frac{\mathrm{i}}{2}\sigma_x, -\frac{\mathrm{i}}{2}\sigma_y, -\frac{\mathrm{i}}{2}\sigma_z\right\} \tag{6.2.36}$$

and are in other words the infinitesimal generators of the group $\mathrm{SU}(2)$. $\mathrm{SU}(2)$ has therefore a basis that is a Lie algebra according to the vocabulary of mathematicians.

This result is quite remarkable ... Since $\mathrm{SU}(2)$ and $\mathrm{SO}(3)$ are isomorphic, then we can get the basis of the Lie algebra $\mathrm{SO}(3)$ while using the same method !!!

Let us see this! We proved in the section of Euclidean Geometry that the rotations matrices were given by (we change the R by a U so to not confuse with the previous matrices):

$$U_x = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta) & \sin(\theta) \\ 0 & -\sin(\theta) & \cos(\theta) \end{bmatrix}, U_y = \begin{bmatrix} \cos(\gamma) & 0 & -\sin(\gamma) \\ 0 & 1 & 0 \\ \sin(\gamma) & 0 & \cos(\gamma) \end{bmatrix}, U_z = \begin{bmatrix} \cos(\phi) & \sin(\phi) & 0 \\ -\sin(\phi) & \cos(\phi) & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{6.2.37}$$

We notice again that on $\theta = \gamma = \phi = 0$ the curve that is generated by the extremity of a vector from the three rotation matrices pass through:

$$U_x = U_y = U_z = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{6.2.38}$$

Then in the same manner as for $\mathrm{SU}(2)$, we calculate the derivatives in these angles to determine the generating base matrices of $\text{SO}(3)$:

$$\begin{aligned}
\left.\frac{\mathrm{d}U_x(\theta)}{\mathrm{d}\theta}\right|_{\theta=0} &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\sin(\theta) & \cos(\theta) \\ 0 & -\cos(\theta) & -\sin(\theta) \end{bmatrix} \Big|_{\theta=0} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix} = X_x \\
\left.\frac{\mathrm{d}U_y(\gamma)}{\mathrm{d}\gamma}\right|_{\gamma=0} &= \begin{bmatrix} -\sin(\gamma) & 0 & -\cos(\gamma) \\ 0 & 0 & 0 \\ \cos(\gamma) & 0 & -\sin(\gamma) \end{bmatrix} \Big|_{\gamma=0} = \begin{bmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} = X_y \\
\left.\frac{\mathrm{d}U_z(\phi)}{\mathrm{d}\phi}\right|_{\phi=0} &= \begin{bmatrix} -\sin(\phi) & \cos(\phi) & 0 \\ -\cos(\phi) & -\sin(\phi) & 0 \\ 0 & 0 & 0 \end{bmatrix} \Big|_{\phi=0} = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = X_z
\end{aligned} \tag{6.2.39}$$

The Lie algebra of $\text{SO}(3)$ has therefore for basis:

$$\{X_x, X_y, X_z\} \tag{6.2.40}$$

In physics, we prefer to work with complex matrices. We then introduce the matrices:

$$\{J_x = \mathrm{i}X_x, J_y = \mathrm{i}X_y, J_z = \mathrm{i}X_z\} \tag{6.2.41}$$

It must then been notice that if we define:

$$J_{\pm} = J_x \pm \mathrm{i}J_y = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \mathrm{i} \\ 0 & -\mathrm{i} & 0 \end{bmatrix} \pm \mathrm{i} \begin{bmatrix} 0 & 0 & -\mathrm{i} \\ 0 & 0 & 0 \\ \mathrm{i} & -0 & 0 \end{bmatrix} \pm \mathrm{i} \begin{bmatrix} 0 & 0 & -\mathrm{i} \\ 0 & 0 & 0 \\ \mathrm{i} & -0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & \pm 1 \\ 0 & 0 & \mathrm{i} \\ \pm 1 & -\mathrm{i} & 0 \end{bmatrix} \tag{6.2.42}$$

We trivially have to the complex conjugate of the transposed matrix:

$$(J_{\pm})^{\dagger} = J_{\mp} \tag{6.2.43}$$

and by the way ... we also have the following non-commutative relations (which we can develop on request):

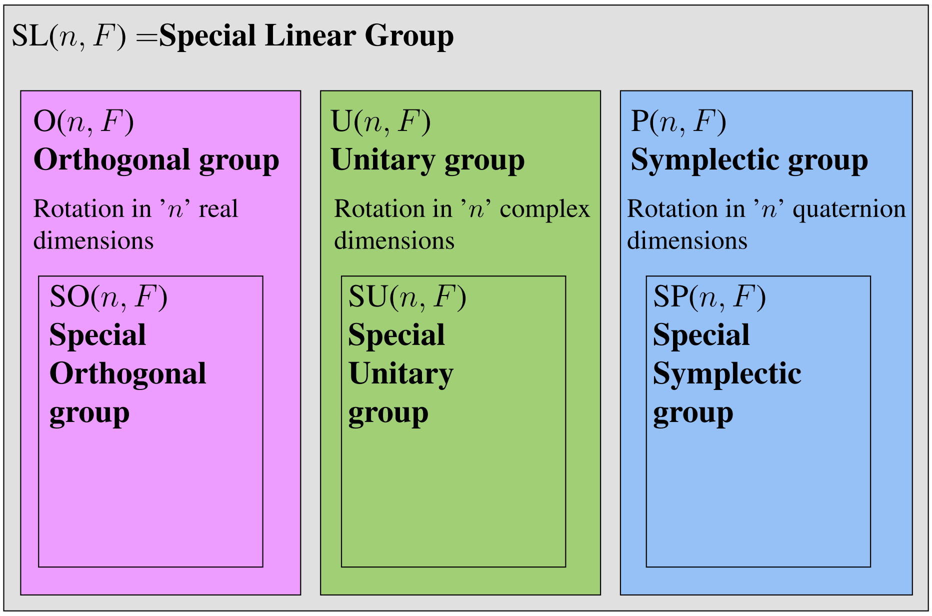
$$[J_x, J_y] = \mathrm{i}J_z \quad \text{(cycl.)} \tag{6.2.44}$$

and also the relation of commutation:

$$\left[J_z, \|\vec{J}\|^2\right] = 0 \quad \text{(cycl.)} \tag{6.2.45}$$

that the Pauli matrices also satisfy and ... for recall (or for information for those who have not yet read the section of Wave Quantum Physics) the J_i are the operators of the total angular momentum of the spin-orbit coupling system!!!

Most of the groups we have seen until so far can be summarised with the following figure:


SL(n, F) =**Special Linear Group**
O(n, F)
Orthogonal group
Rotation in 'n' real dimensions
SO(n, F)
Special Orthogonal group
U(n, F)
Unitary group
Rotation in 'n' complex dimensions
SU(n, F)
Special Unitary group
P(n, F)
Symplectic group
Rotation in 'n' quaternion dimensions
SP(n, F)
Special Symplectic group

The rotations with the quaternions indicated in the figure above are studied in the section Numbers of the chapter Arithmetic (see page 261).

Or also we can summarize what we have seen so far as a unique set or relationships (to accustom the reader to the absence of international standards for notation in mathematics...):

$$\boxed{\begin{aligned}
\mathrm{SL}_n(\mathbb{R}) &= \{X \in \mathrm{GL}_n(\mathbb{R}) : \det(X) = 1\}\\
\mathrm{SL}_n(\mathbb{C}) &= \{X \in \mathrm{GL}_n(\mathbb{C}) : \det(X) = 1\}\\
\mathrm{O}(n) &= \left\{X \in \mathrm{GL}_n(\mathbb{R}) : X^T X = \mathbb{1}_n\right\}\\
\mathrm{SO}(n) &= \mathrm{O}(n) \cap \mathrm{SL}_n(\mathbb{R})\\
\mathrm{O}(p,q) &= \left\{X \in \mathrm{GL}_{p+q}(\mathbb{R}) : X^T \mathbb{1}_{p,q} X = \mathbb{1}_{p,q}\right\}\\
\mathrm{SO}(p,q) &= \mathrm{O}(p,q) \cap \mathrm{SL}_{p+q}(\mathbb{R})\\
\mathrm{U}(n) &= \left\{X \in \mathrm{GL}_n(\mathbb{C}) : \bar{X} X = \mathbb{1}_n\right\}\\
\mathrm{SU}(n) &= \mathrm{U}(n) \cap \mathrm{SL}_n(\mathbb{C})
\end{aligned}} \tag{6.2.46}$$

Now let us talk about the "dimensions" (i.e. free parameters) of these groups. For $\mathrm{GL}_n(\mathbb{R})$ its dimensions are n^2 that's quite obvious and as $\mathrm{GL}_n(\mathbb{C})$ is made of complex numbers each having two real parameters, this doubles the number of dimensions is such that we have $2n^2$. $\mathrm{SL}_n(\mathbb{R})$ has a constraint that is its matrices must have a determinant equal to 1, this constraint reduces the free parameters by one unit, hence it has n^2-1 dimensions. For $\mathrm{SL}_n(\mathbb{C})$ we just multyply it by two again that gives $2n^2-2$. All others follow the same pattern.

6.2.1.3 Group of Symmetries

The symmetry group of an object denoted X (image, signal, etc. in 1D, 2D, 3D or other) is the group of all isometries (an isometry is a transformation that preserves length) under which it is invariant with composition as an operation.

Any group of symmetries which elements have a common fixed point, which is true for all symmetry groups of limited figures, can be represented as a subgroup of the orthogonal group $\text{O}(n)$ by choosing the origin as a fixed point. The proper symmetry group is a subgroup of the special orthogonal group $\text{SO}(n)$, and hence it is also named the "rotations group" of the figure.

In what follows, we will interpret the composition of two operation of symmetries or rotations as the multiplication as well as for permutations.

Let us see first two fundamental definitions!

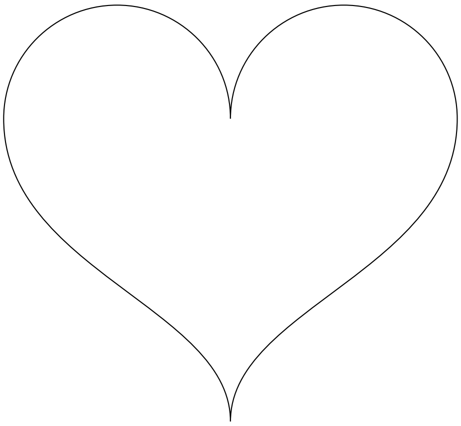
Definitions (#6.2.187):

D1. The "group of symmetries", also named "group of invariants" of X is the set of symmetries of X, generated by the multiplication structure given by the composition that leaves X invariant.

D2. The "order" of a group is the total number of all its symmetries only (including the identity!).

✍**Examples:**

E1. The heart (...):



has a group of symmetries of 2 elements, namely the identity application id and the application r_v that is the reflection in respect to the vertical axis (sub-group of symmetries with 1 item). We observe also that the symmetrical is also provided via the relation:

$$r_v \circ r_v = \text{id} \tag{6.2.47}$$

E2. The letter phi (...):

$$\Phi$$

has a total symmetry group of 4 elements, namely the identity application id, the both reflections r_h and r_v and the rotation of angle π which we denote by t_π (subgroup of rotations of 1 item). This form thus has a group of symmetries of order 3.

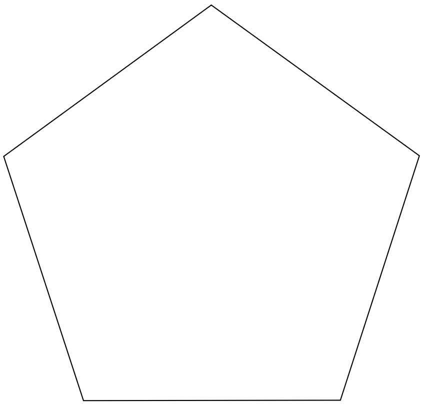
In this group we have:

$$r_h \circ r_v = t_\pi \tag{6.2.48}$$

(which is commutative!), $t_\pi \circ t_\pi$ is the rotation of angle 2π, which is the same application as the identity application, therefore $t_\pi \circ t_\pi = \text{id}$.

Thus, the symmetry group of this letter is indeed commutative and the composition law is internal. It is indeed a group.

E3. The regular pentagon:



has a group of symmetries of 10 elements, namely, more precisely 5 rotations id, $t_{2\pi/5}, t_{4\pi/5}, t_{6\pi/5}, t_{8\pi/5}$ and also 5 reflections along the 5-axis symmetry. It is therefore a group of symmetries of order 5 corresponding to cyclic group $\mathbb{Z}/5\mathbb{Z}$.

Remark

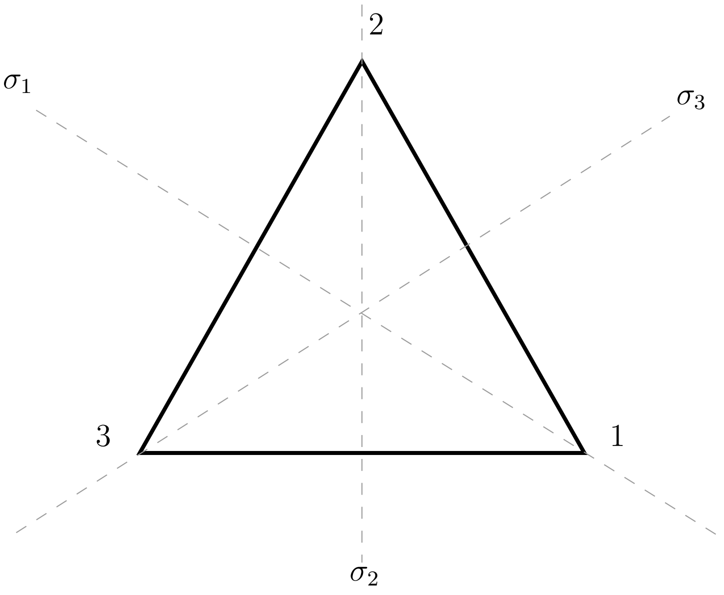
More generally, the group of symmetries of a regular n-gon (where n is odd) has exactly $2n$ elements. This group is named "dihedral group of order n" and is most often denoted by D_{2n} (be careful because some authors do not multiply by a factor of 2 so that the index is then directly the order and not the number of elements).

The pentagon has therefore D_{10} for dihedral group and $\mathbb{Z}/5\mathbb{Z}$ is a "distinct subgroups" (we'll come on this notion of subgroup later below).

E4. The dihedral group D_6 of order 3 of the isometries of an equilateral triangle (regular polygon) has 6 elements which we will denote (so that the writing is less heavy):

$$D_6 = \{\text{id}, t_{2\pi/3}, t_{4\pi/3}, \sigma_1, \sigma_2, \sigma_3\} \tag{6.2.49}$$

where $\sigma_1, \sigma_2, \sigma_3$ are the symmetries relatively to the three bisectors (mediators respectively). The compositions table below, named in general a "Cayley table", of this dihedral group also shows that this group is non-commutative:

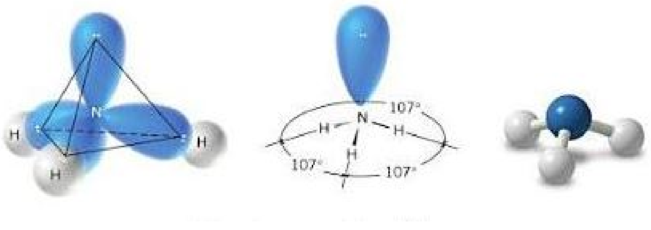


$\nearrow D_6$	id	$t_{2\pi/3}$	$t_{4\pi/3}$	σ_1	σ_2	σ_3
id	id	$t_{2\pi/3}$	$t_{4\pi/3}$	σ_1	σ_2	σ_3
$t_{2\pi/3}$	$t_{2\pi/3}$	$t_{4\pi/3}$	id	σ_3	σ_1	σ_2
$t_{\pi/3}$	$t_{4\pi/3}$	id	$t_{2\pi/3}$	σ_2	σ_3	σ_1
σ_1	σ_1	σ_2	σ_3	id	$t_{2\pi/3}$	$t_{4\pi/3}$
σ_2	σ_2	σ_3	σ_1	$t_{4\pi/3}$	id	$t_{2\pi/3}$
σ_3	σ_3	σ_1	σ_2	$t_{2\pi/3}$	$t_{4\pi/3}$	id

Table 6.8: **Symmetries of dihedral group of order 3**

We will return back later on this example when we will introduce a little further below the concept of a distinct group in our study of permutations groups.

E5. Let's look at one last example applied to chemistry by enumerating the symmetry operations that leave the ammonia molecule NH_3 (tetrahedron) invariant.



The transformation group contains 6 elements: the identity id, C_3 that is the rotation of $2\pi/3$, C_3^2 that is the rotation of $4\pi/3$ (which we will denote after by C_{-3}) both along the z-axis (therefore perpendicular to the xy-plane...) and 3 axes of symmetry $\sigma_1, \sigma_2, \sigma_3$ of symmetries/reflection each passing through the middle of one of the base edges in the middle of the opposite edge as shown in the figure below (pyramid view from above):

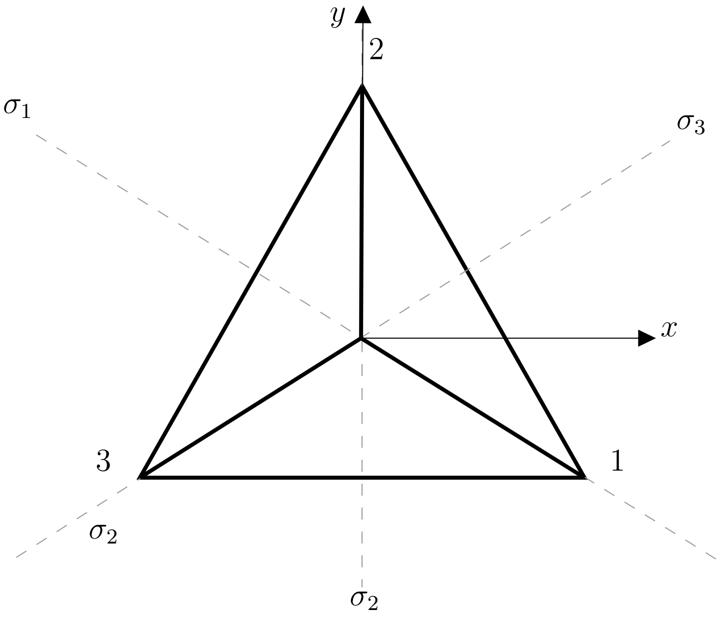


Figure 6.22: **Operations leaving invariant a tetrahedron**

The combination of the various elements of symmetries show that the compositions table is (which proves that the law is internal and we are working effectively in a group):

↗	id	σ_1	σ_2	σ_3	C_3	C_{-3}
id	id	σ_1	σ_2	σ_3	C_3	C_{-3}
σ_1	σ_1	id	C_{-3}	C_3	σ_3	σ_2
σ_2	σ_2	C_3	id	C_{-3}	σ_1	σ_3
σ_3	σ_3	C_{-3}	C_3	id	σ_2	σ_1
C_3	C_3	σ_2	σ_3	σ_1	C_{-3}	id
C_{-3}	C_{-3}	σ_3	σ_1	σ_2	id	C_3

Table 6.9: **Compositions of transformations of the tetrahedron**

Attention to the order of operations in the above table, we first apply the line element and afterwards the column element!

We note that the group is not commutative.

6.2.1.3.1 Orbits and Stabilizers

We will now see two definitions we meet again in crystallography (their name is not innocent!).

Definition (#6.2.188): The orbit of an element x of E is given by:

$$\text{Orb}_x = \{g \cdot x, g \in G\} \tag{6.2.50}$$

The orbit of x is the set of all positions (in E) likely to be occupied by the image of x under the action of G. The orbits obviously form a partition of E.

✍**Example:**

Let us consider a set E on which a group G will act, by:

$$E = \{S_0, S_1, S_2, S_3, S_4, S_5\} \tag{6.2.51}$$

the set of all 6 vertices of a hexagon on which we do act the group $G = \{\text{id}, t_{2\pi/3}, t_{4\pi/3}\}$. We see already trivially see that G is a group!

But now let us, consider an element of E, for example S_0.

Its orbit will be by definition:

$$\text{Orb}_{S_0} = \{\text{id}(S_0)S_0, t_{2\pi/2}(S_0) = S_2, t_{4\pi/3}(S_0) = S_4\} \tag{6.2.52}$$

Definition (#6.2.189): The stabilisator of x of an element of E is the set:

$$\text{Stab}_x = \{g \in G | g \cdot x = x\} \tag{6.2.53}$$

of all elements that let x invariant under their action. It is a subgroup of G.

✍**Example:**

To continue with our previous example. Its stabilisator is reduced to:

$$\text{Stab}_{S_0} = \{\text{id}\} \tag{6.2.54}$$

6.2.1.4 Permutations Groups

Symmetric groups have significant importance in some areas of Quantum Physics but also in mathematics as part of Linear Algebra (for the theory of the determinant as we will see in the corresponding section) and in Galois Theory (see further below) and also in Error Correcting Codes (see corresponding section for the example used with VISA credit cards). So these must be also pay special attention!

Let us first recall (see section Probabilities page 525) that in a set $\{1, \ldots, n\}$ there are $n!$ possible permutations. Mathematicians say, rightly, that there are $n!$ bijections and name this number "order of the permutations of the group".

✍**Example:**

Given for example the set $\{1,2,3\}$. This set has $3!$ possible permutations that are denoted in the context of the permutation groups as follows:

$$\{(1),(1\,2),(1\,3),(2\,3),(1\,2\,3),(1\,3\,2)\} \tag{6.2.55}$$

This must be read in the order: identity application id, 1 takes on 2 or 2 on 1 (in terms of positions!), 1 takes on 3 or 3 on 1, 2 takes on 3 or 3 on 2, 1 takes on 2 that takes on 3 that takes on 1, 1 takes on 3 that takes on 2 that takes on 1.

$$\begin{aligned}
&\begin{pmatrix}1 & 2 & 3\\ 1 & 2 & 3\end{pmatrix}\rightarrow(1) && \begin{pmatrix}1 & 2 & 3\\ 2 & 1 & 3\end{pmatrix}\rightarrow(1,2)\\
&\begin{pmatrix}1 & 2 & 3\\ 3 & 2 & 1\end{pmatrix}\rightarrow(1,3) && \begin{pmatrix}1 & 2 & 3\\ 1 & 3 & 2\end{pmatrix}\rightarrow(2,3)\\
&\begin{pmatrix}1 & 2 & 3\\ 3 & 1 & 2\end{pmatrix}\rightarrow(1,2,3) && \begin{pmatrix}1 & 2 & 3\\ 2 & 3 & 1\end{pmatrix}\rightarrow(1,3,2)
\end{aligned} \tag{6.2.56}$$

We can easily observe that the composition of two permutations is not commutative:

$$\begin{aligned}(1,\ 2)\circ(1,\ 3)&=\{1,2,3\}\\(1,\ 3)\circ(1,\ 2)&=\{3,1,2\}\end{aligned} \tag{6.2.57}$$

and that the composition of two permutations is an internal law:

$$(1\,2)\circ(1\,3\,2)=\{1,3,2\}=(2\,3) \tag{6.2.58}$$

with a neutral element that is indeed the identity id. So we do have well a non-commutative group. Let us recall also the reader that certain elements of the group, if well chosen, can form a subgroup. This is the example of:

$$\{(1),(1\,2)\} \tag{6.2.59}$$

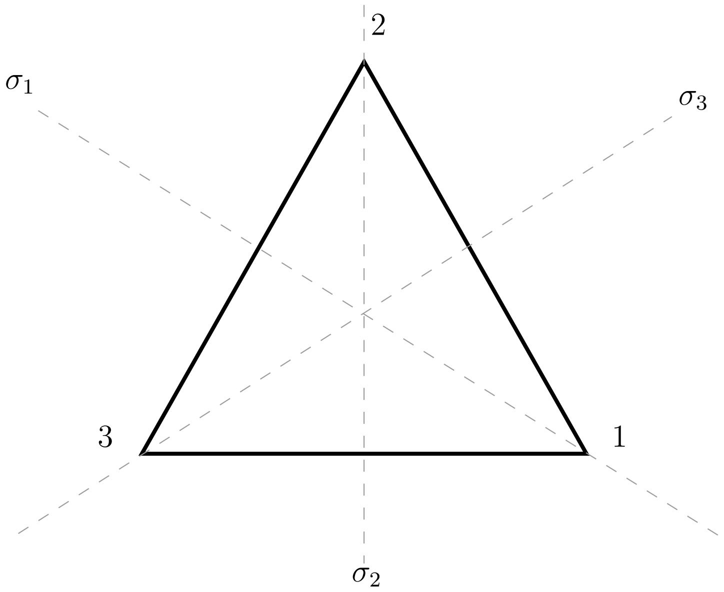
which is a sub-group of S_3 (it must be easy to the reader to check that it has all the properties of a group).

Definition (#6.2.190): A subgroup H of a group G is named "distinct group" if, for every g of G and every h of H, we have that ghg^{-1} is element of H. The mathematicians name this an "inner automorphism"...

Let us first consider interesting introducing geometric example after which we will come back to this definition with S_3.

✍**Example:**

We have seen earlier the elements of the dihedral symmetry group of order 3 of the equilateral triangle. Geometrically they all correspond to displacements in the plane in which the triangle is located. We got for recall of the following table of compositions:



$\nearrow D_6$	*id*	$t_{2\pi/3}$	$t_{4\pi/3}$	σ_1	σ_2	σ_3
id	id	$t_{2\pi/3}$	$t_{4\pi/3}$	σ_1	σ_2	σ_3
$t_{2\pi/3}$	$t_{2\pi/3}$	$t_{4\pi/3}$	id	σ_3	σ_1	σ_2
$t_{\pi/3}$	$t_{4\pi/3}$	id	$t_{2\pi/3}$	σ_2	σ_3	σ_1
σ_1	σ_1	σ_2	σ_3	id	$t_{2\pi/3}$	$t_{4\pi/3}$
σ_2	σ_2	σ_3	σ_1	$t_{4\pi/3}$	id	$t_{2\pi/3}$
σ_3	σ_3	σ_1	σ_2	$t_{2\pi/3}$	$t_{4\pi/3}$	id

First, we easily see using this table that we have:

- The sub-group made of $\{\text{id}\}$ of order 1
- The sub-group made of $\{\text{id}, t_{2\pi/3}, t_{4\pi/3}\}$ of order 3
- The sub-group made of $\{\text{id}, \sigma_1\}$ of order 2
- The sub-group made of $\{\text{id}, \sigma_2\}$ of order 2

- The sub-group made of $\{\text{id}, \sigma_3\}$ of order 2

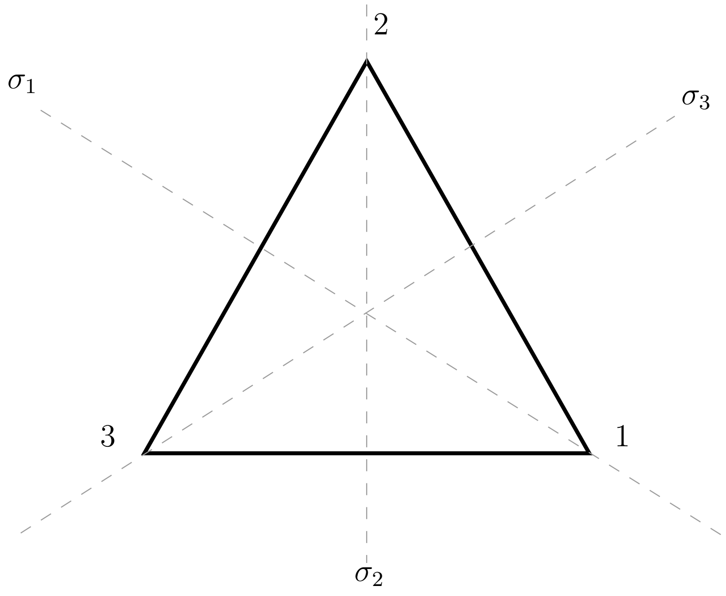
Among these five subgroups, let us see which one are distincted sub-groups (this is relatively easy to see using the table of compositions above):

- The sub-group made of $\{\text{id}\}$
- The sub-group made of $\{\text{id}, t_{2\pi/3}, t_{4\pi/3}\}$

We will now see a remarkable thing! By numbering with $1, 2$ and 3 the vertices of the equilateral triangle and taking the rotations clockwise, we can identify the elements of D_6 to the following elements of S_3:

$$\begin{gathered}
\text{id} = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix} \rightarrow (1) \quad t_{2\pi/3} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix} \rightarrow (1,2,3) \quad t_{4\pi/3} = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix} \rightarrow (1,3,2) \\
\sigma_1 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix} \rightarrow (2,3) \quad \sigma_2 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix} \rightarrow (1,3) \quad \sigma_3 = \begin{pmatrix} \rightarrow 1 & 2 & 3 \\ 2 & 1 & 2 \end{pmatrix} \rightarrow (1,2)
\end{gathered} \tag{6.2.60}$$

and rebuild the same table of compositions (copy of the previous one but just with the notation change ... hehe!):



$\nearrow S_3$	(1)	(1 2 3)	(1 3 2)	(1 2)	(1 3)	(2 3)
(1)	(1)	(1 2 3)	(1 3 2)	(1 2)	(1 3)	(2 3)
(1 2 3)	(1 2 3)	(1 3 2)	(1)	(2 3)	(1 2)	(1 3)
(1 3 2)	(1 3 2)	(1)	(1 2 3)	(1 3)	(2 3)	(1 2)
(1 2)	(1 2)	(1 3)	(2 3)	(1)	(1 2 3)	(1 3 2)
(1 3)	(1 3)	(2 3)	(1 2)	(1 3 2)	(1)	(1 2 3)
(2 3)	(2 3)	(1 2)	(1 3)	(1 2 3)	(1 2 3)	(1)

Well ... this little interlude closed, let us return to the distincted subgroup of S_3 (as it will be important for our introduction to Galois groups) and let us first recall that:

$$G=\{(1),(1\ 2),(1\ 3),(2\ 3),(1\ 2\ 3),(1\ 3\ 2)\} \tag{6.2.61}$$

and we see that the distinct subgroup consists of:

$$H=\{(1),(1\ 2\ 3),(1\ 3\ 2)\} \tag{6.2.62}$$

Definition (#6.2.191): For any subgroup H stable by the inner automorphisms of a group G, we name "index of H in G" the quotient of the order of the group G by the order of the subgroup H and we denote it by $[G/H]$.

✍**Example:**

The index of the subgroup $\{(1),(12)\}$ in the group S_3 is $6/2$ that is to say 3.

This concept will be very helpful to us during our introduction to Galois Theory.

Let us consider now the particular permutation σ to introduce the subject from a different but equivalent angle:

$$\{1,2,3,4\}\xrightarrow{\sigma}\{3,2,4,1\} \tag{6.2.63}$$

As we know the mathematicians are accustomed to note that, at first, in the form:

$$\begin{pmatrix}1&2&3&4\\3&2&4&1\end{pmatrix} \tag{6.2.64}$$

with:

$$\sigma(1)=3,\sigma(2)=2,\sigma(3)=4,\sigma(4)=1 \tag{6.2.65}$$

Hence:

$$\begin{pmatrix}1&2&3&4\\\sigma(1)&\sigma(2)&\sigma(3)&\sigma(4)\end{pmatrix} \tag{6.2.66}$$

Given σ and τ, two permutations, it is natural to look at their composition $\tau\circ\sigma$ (recall that this means that first we apply σ and afterwards τ as for the composition of functions).

Therefore if:

$$\sigma=\begin{pmatrix}1&2&3&4\\3&2&4&1\end{pmatrix}\quad\text{and}\quad\tau=\begin{pmatrix}1&2&3&4\\2&1&4&3\end{pmatrix} \tag{6.2.67}$$

Then:

$$\tau\circ\sigma=\begin{pmatrix}1&2&3&4\\4&1&3&2\end{pmatrix} \tag{6.2.68}$$

and:

$$\sigma\circ\tau=\begin{pmatrix}1&2&3&4\\4&2&1&3\end{pmatrix} \tag{6.2.69}$$

Now the idea is to interpret the composition as a multiplication of permutations. This multiplication is then non-commutative as we have seen in the previous example. We have generally $\sigma\circ\tau\neq\tau\circ\sigma$.

Each bijection has an inverse (reciprocal function). In our example it is obviously:

$$\sigma^{-1}=\begin{pmatrix}1&2&3&4\\4&2&1&3\end{pmatrix}\tag{6.2.70}$$

Geometrically, to calculate the inverse σ^{-1} of an element σ, we just need to take the reflection of the drawing of σ in a horizontal axis as shown in the left side of the figure below:

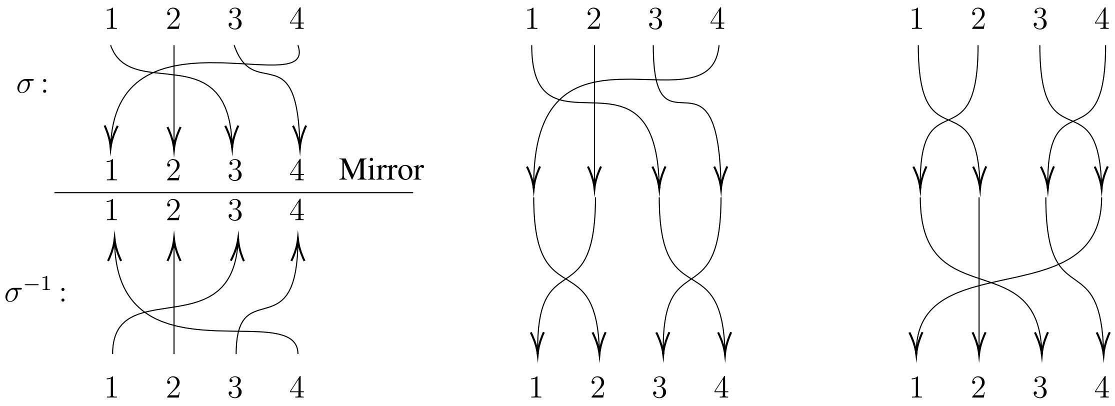


Figure 6.23: **Examples of compositions and inverse of permutations**

We may represent a permutation σ as a permutation matrix, i.e., the matrix that maps:

$$\sigma:\begin{pmatrix}1\\2\\\vdots\\n\end{pmatrix}\mapsto\begin{pmatrix}\sigma(1)\\\sigma(2)\\\vdots\\\sigma(n)\end{pmatrix}\tag{6.2.71}$$

An $n\times n$ matrix is a permutation matrix if and only if all its entries are zeros and ones, each column has exactly one 1, and each row has exactly one 1. In this instance, we have:

$$\sigma=\begin{bmatrix}0&0&1&0\\0&1&0&0\\0&0&0&1\\1&0&0&0\end{bmatrix}\tag{6.2.72}$$

Definitions (#6.2.192):

D1. The set of permutations of a set with n elements, with this structure of multiplication, is named the "group of permutations of order n" or "group of substitutions of order n", and is denoted by S_n or $S(n)$.

D2. We say that an element σ of S_n is a "cycle of order k", or simply a "k-th cycle" if there exists $a_1,a_2,\ldots,a_k\in\{1,\ldots,n\}$ such that:

- σ send a_1 on a_2, a_2 on $a_3,\ldots,a_{k-1}$, and a_k on a_1.

- σ fix all other elements of S_n

and we denote the cycle as following:

$$\sigma=(a_1,a_2,\ldots,a_k) \tag{6.2.73}$$

✍**Example:**

Perhaps, for a better understanding let us our previous example of S_4:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 4 & 1 \end{pmatrix} \tag{6.2.74}$$

This symmetric group is a 3-cycle denoted $\sigma=(1\,3\,4)$ because in order: 1 send on 3, 3 send on 4 and 4 send on 1 (and 2 is not mentioned as it remains fixed). We can also write this in the following equivalent ways: $\sigma=(3\,4\,1)$ or also $\sigma=(4\,1\,3)$.

D3. The order of a k-cycle is k (hence the name!).

✍**Example:**

Indeed if we take again our $\sigma=(1\,3\,4)$, then we have:

$$\sigma^2=(1\,4\,3) \quad \text{and} \quad \sigma^3=\text{id} \tag{6.2.75}$$

D4. We say that a permutation σ is a "cycle" if there exists $k\in\mathbb{N}$ such as that σ is a k-cycle.

Warning! Any permutation must be written as a product of disjoint cycles (that is to say, a number that appears in a cycle should not appear in another cycle).

✍**Example:**

For example, in S_9, we have:

$$\sigma=\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 5 & 2 & 8 & 3 & 7 & 1 & 6 & 4 & 9 \end{pmatrix}=(1\;5\;7\;6)(3\;8\;4) \tag{6.2.76}$$

So this service is a product of a 4-cycle and of a 3-disjoint cycle.

We also let the reader see for himself that the cyclic group generated by σ (which in this case is a subgroup of S_9) is of order 12 (12-cycle)...

Remark

Mathematicians can prove that if σ is an element that has a decomposition into c disjoint cycles of length $n_1,n_2,\ldots,n_c$ then the order of σ is the least common multiple of the orders of all disjoint cycles that compose it.

We also assume intuitive that in the common vocabulary, a 2-cycle into S_n is also named a "transposition".

Let us go a little further. We propose to show by example that the set of all transpositions generates S_n. In other words, any permutation is written as a product of transpositions!

✍**Example:**

Let us return to our example (it is an even permutation):

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 4 & 1 \end{pmatrix} = (1\;3\;4) = (1\;3)(4\;1) \tag{6.2.77}$$

In general, a k-cycle is thus written as the product of $k-1$ transpositions.

Theorem 6.6. As the permutations of a finite set form a group, this means (among other things) that there is therefore always an integer k, such that p applied k time is the identity transformation (that is to say, the operation that does not change anything).

Proof 6.6.1. If G is a finite group and that $g \in G$, we consider the sequence of items (recall that in a group there is only one an operation and therefore the square, cube, etc. means we compose this operation!):

$$g, g^2, g^3, \ldots \tag{6.2.78}$$

For example, in the permutation groups, the operation is the composition of permutations.

Since G is finite and that this sequence is infinite, there are necessarily two elements that are equal in the sequence... So there are two different n and m such that:

$$g^n = g^m \tag{6.2.79}$$

Assuming that $n < m$, the previous equality is simplified and we get:

$$g^{m-n} = e \tag{6.2.80}$$

where e is the neutral element of the group.

■ Q.E.D.

We will see that the permutations being bijective, we can create on finite groups, compositions of permutations operations that always end up returning to the initial state (identity application id)!!!

✍**Examples:**

E1. In a list of 5 items, we exchange the first and the third, and, at the same time, we pass the second in position 4, this that is in position 4 in put in position 5 and the one in position 5 is put on position 2. And we reiterate. This gives:

$$ABCDE \rightarrow CEABD \rightarrow ADCEB \rightarrow CBADE \rightarrow AECBD \rightarrow CDAEB \rightarrow ABCDE \tag{6.2.81}$$

We returned to the starting point after 6 steps.

E2. Let us consider the "Photomaton transformation" of the image of Mona Lisa of size 256 by 256 pixels:

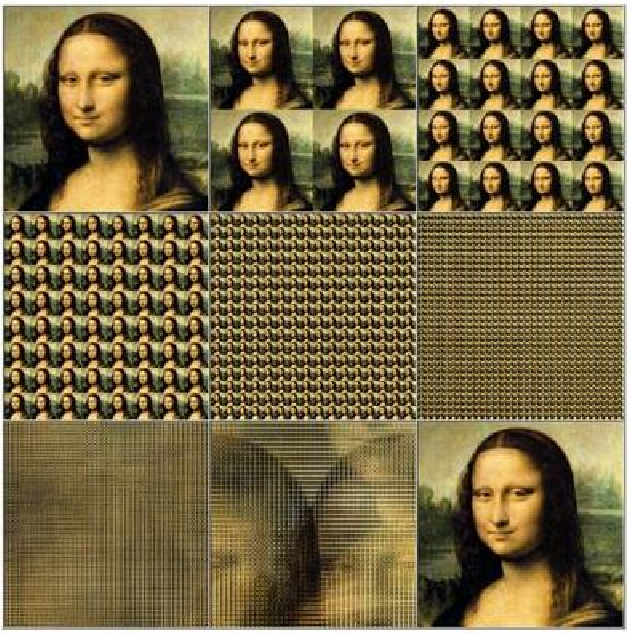


Figure 6.24: **Photomaton transformation**

We could thing that each image was obtained from the previous by reducing the size of the image half, which gave four similar pieces that we have placed in square to obtain an image having the same size as original image. But in fact it is not!!!! The number of pixels has been preserved (no pixel is duplicated!!!) and we actually just moved the pixels by permutation to get four images that do not actually contain all the information of the original image but only a part!

By repeating the procedure 8 times we always fall back on the original image regardless of the original image. The question is to understand why?

Let us consider that the original image is a square with a size of 16 pixels wide by 16 pixels high (but you can also apply what follows with a rectangular image of any size and you will see that it works also!). Each pixel of a line (the process is exactly the same for columns!) is identified by a coordinate along the X axis going from 0 to 15.

Thus we have at the beginning a sequence of numbers where the pixel coordinates correspond to their x coordinate:

$$0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10\ 11\ 12\ 13\ 14\ 15$$

We then do the permutation that consist to denote by k the position of a pixel and to do:

$$\text{permutation:}\begin{cases} x \text{ is even:} & 2k \rightarrow k \\ x \text{ is odd:} & 2k+1 \rightarrow 8+k \end{cases} \tag{6.2.82}$$

This then gives the first permutation:

$$\begin{array}{c} 0\;1\;2\;3\;4\;5\;6\;7\;8\;9\;10\;11\;12\;13\;14\;15 \\ \downarrow \text{Permutation } 1 \\ 0\;2\;4\;6\;8\;10\;12\;14\;1\;3\;5\;7\;9\;11\;13\;15 \\ \downarrow \text{Permutation } 2 \\ 0\;4\;8\;12\;1\;5\;9\;13\;2\;6\;10\;14\;3\;7\;11\;15 \\ \downarrow \text{Permutation } 3 \\ 0\;8\;1\;9\;2\;10\;3\;11\;4\;12\;5\;13\;6\;14\;7\;15 \\ \downarrow \text{Permutation } 4 \\ 0\;1\;2\;3\;4\;5\;6\;7\;8\;9\;10\;11\;12\;13\;14\;15 \end{array}$$

Thus, for an image of 16 by 16 pixels, it takes four permutations, which corresponds to $2^4=16$. So for an image of 256 pixels, we have $256=2^8$, hence the fact that we need 8 permutation to find back the original Mona Lisa with:

$$\text{permutation:}\begin{cases} x \text{ is even:} & 2k \rightarrow k \\ x \text{ is odd:} & 2k+1 \rightarrow 128+k \end{cases} \tag{6.2.83}$$

Thus, in the general case of an image of width L in number of pixels counting from 1, the transformation is:

$$\text{permutation:}\begin{cases} x \text{ is even:} & 2k \rightarrow k \\ x \text{ is odd:} & 2k+1 \rightarrow E^{+}\left[\dfrac{L}{2}\right]+k \end{cases} \tag{6.2.84}$$

where $E^{+}[\ldots]$ is the upper (nearest) integer value in the case where L is odd.

The reader will also have maybe notice something interesting if we return to our example with the image of 16 pixels ... Indeed, let us take the third pixel from the left of coordinate x equal to 2. In binary, its initial position is then 0010. After the first permutation, its x coordinate is equal to 1, or in binary: 0001. After the second permutation, the x coordinate is equal to 8, or in binary: 1000, etc. In fact we see that every permutation can be summarized in binary by shifting the bits to the right.

Definition (#6.2.193): Given $\sigma \in S_n$ a permutation. We say that σ is an "even permutation" if, in a writing of σ as product of transpositions, there is an even number of transpositions. We say the obviously that σ is an "odd permutation" if, in a writing of σ as product of transposition, there is an odd number of transpositions.

Let us end with a small complement... We know that S_3 is a group of permutations of order 3

and therefore with $3!=6$ possible permutations.

If we list the 6 permutations we saw what we get:

$$\{(1),(1\,2),(1\,3),(2\,3),(1\,2\,3),(1\,3\,2)\} \tag{6.2.85}$$

Among these only some can be written as an even product of transpositions:

$$(1\,2\,3)=(1\,2)(3\,1) \quad \text{and} \quad (1\,3\,2)=(1\,3)(2\,1) \tag{6.2.86}$$

The even permutations form with the identity permutation, a subgroup (not commutative!) we name the "alternating group of order n" and that we denote by AN. It is easy to check with the previous example.

6.2.2 Galois Theory

In abstract algebra, Galois theory, named after Évariste Galois, provides a connection between field theory and group theory. Using Galois theory, certain problems in field theory can be reduced to group theory, which is, in some sense, simpler and better understood.

Originally, Galois used permutation groups to describe how the various roots of a given polynomial equation are related to each other.

The birth and development of Galois theory was caused by the following question, whose answer is known as the "Abel–Ruffini theorem": Why is there no formula for the roots of a fifth (or higher) degree polynomial equation in terms of the coefficients of the polynomial, using only the usual algebraic operations (addition, subtraction, multiplication, division) and application of radicals (square roots, cube roots, etc)?.

Galois theory not only provides a beautiful answer to this question, it also explains in detail why it is possible to solve equations of degree four or lower in the above manner, and why their solutions take the form that they do. Further, it gives a conceptually clear, and often practical, means of telling when some particular equation of higher degree can be solved in that manner.

Galois theory originated in the study of symmetric functions: the coefficients of a monic polynomial are (up to sign) the elementary symmetric polynomials in the roots. For instance:

$$(x-a)(x-b)=x^2-(a+b)x+ab \tag{6.2.87}$$

where 1, $a+b$ and ab are the elementary polynomials of degree 0, 1 and 2 in two variables.

We tried to make this part of the book as easy as possible. So we hope our goal is reach if you understand what follows. Let us now begin!!!

6.2.2.1 Elementary symmetric and Invariant Polynomials

Definition (#6.2.194): The "kth elementary symmetric polynomial in n variables", denoted $s_k(x_k,\ldots,x_n)$ is the sum of all possible degree k monomials in n variables with each x_i appearing NO MORE THAN ONCE IN EACH MONOMIAL. Formally, for $k \leq n$:

$$s_k(x_1,\ldots,x_n) = \sum_{1\leq i_1<\ldots<i_k\leq n} x_{i_1}x_{i_2}\ldots x_{i_k} \tag{6.2.88}$$

A polynomial is said "invariant under S_n" if and only if it is a polynomial in the elementary symmetric functions $s_1,\ldots,s_n$.

Therefore:

$$\begin{aligned}
s_0(x_1,x_2,\ldots,x_n) &= 1,\\
s_1(x_1,x_2,\ldots,x_n) &= \sum_{1\leq j\leq n} x_j,\\
s_2(x_1,x_2,\ldots,x_n) &= \sum_{1\leq j<k\leq n} x_jx_k,\\
s_3(x_1,x_2,\ldots,x_n) &= \sum_{1\leq j<k<l\leq n} x_jx_kx_l\\
&\vdots\\
s_n(x_1,x_2,\ldots,x_n) &= x_1x_2\ldots x_n
\end{aligned} \tag{6.2.89}$$

✍**Example:**

For $n=1$:

$$s_1(x_1) = x_1, \tag{6.2.90}$$

For $n=2$:

$$\begin{aligned}
s_1(x_1,x_2) &= x_1+x_2\\
s_2(x_1,x_2) &= x_1x_2
\end{aligned} \tag{6.2.91}$$

For $n=3$:

$$\begin{aligned}
s_1(x_1,x_2,x_3) &= x_1+x_2+x_3,\\
s_2(x_1,x_2,x_3) &= x_1x_2+x_1x_3+x_2x_3,\\
s_3(x_1,x_2,x_3) &= x_1x_2x_3
\end{aligned} \tag{6.2.92}$$

For $n=4$:

$$\begin{aligned}
s_1(x_1,x_2,x_3,x_4) &= x_1+x_2+x_3+x_4\\
s_2(x_1,x_2,x_3,x_4) &= x_1x_2+x_1x_3+x_1x_4+x_2x_3+x_2x_4+x_3x_4\\
s_3(x_1,x_2,x_3,x_4) &= x_1x_2x_3+x_1x_2x_4+x_1x_3x_4+x_2x_3x_4\\
s_4(x_1,x_2,x_3,x_4) &= x_1x_2x_3x_4
\end{aligned} \tag{6.2.93}$$

Now consider the equation:

$$ax^2+bx+c=0 \tag{6.2.94}$$

It can be rewritten as:

$$1\cdot x^2+\frac{b}{a}x+\frac{c}{a}=0 \tag{6.2.95}$$

or as we know the Vieta relations with two roots, the latter can be rewritten:

$$1\cdot x^2-(r_1+r_2)x+r_1r_2=(x-r_1)(x-r_2)=0 \tag{6.2.96}$$

That is to say:

$$s_0\cdot x^2-s_1x+s_2=0 \tag{6.2.97}$$

Similarly, for a third-degree polynomial:

$$\begin{aligned}&ax^3+bx^2+cx+d=\\&a(x-r_1)(x-r_2)(x-r_3)=a[x^3-(r_1+r_2+r_3)x^2+(r_1r_2+r_1r_3+r_2r_3)x-r_1r_2r_3]\\&=a[s_0x^3-s_1x^2+s_2x-s_3]\end{aligned} \tag{6.2.98}$$

and therefore:

$$s_1=r_1+r_2+r_3=-\frac{b}{a}\quad s_2=r_1r_2+r_1r_3+r_2r_3=\frac{c}{a}\quad s_3=r_1r_2r_3=-\frac{d}{a} \tag{6.2.99}$$

The elementary symmetric polynomials appear when we expand a linear factorization of a monic polynomial. We have the identity:

$$\prod_{i=1}^{n}(x-r_i)=x^n-s_1(r_1,\ldots,r_n)x^{n-1}+s_2(r_1,\ldots,r_n)x^{n-2}+\cdots+(-1)^ns_n(r_1,\ldots,r_n) \tag{6.2.100}$$

That is, when we substitute numerical values for the variables $r_1,r_2,\ldots,r_n$, we get the monic univariate polynomial (with variable x) whose roots are the values substituted for $r_1,r_2,\ldots,r_n$ and whose coefficients are up to their sign the elementary symmetric polynomials. These relations between the roots and the coefficients of a polynomial are named "General Vieta's formulas" for which we have already see two special cases in the section of Calculus and that we will generalize further below.

Theorem 6.7. If $r_1,r_2,\ldots,r_n$ are the roots of a degree n polynomial, then:

$$\boxed{\prod_{i=1}^{n}(x-r_i)=x^n+\sum_{i=1}^{n}(-1)^is_i(r_1,\ldots,r_n)x^{n-i}} \tag{6.2.101}$$

Proof 6.7.1. We will prove this by induction on the degree of the polynomial. If our polynomial is of degree $n=1$ with root r, the left hand side is $x-r$, and the right hand side is $x-s_1(r)=$

$x-r$, so the equation holds for $n=1$. Suppose the equation holds for all polynomials of degree n. Let $P(x)$ be of degree $n+1$ with roots $r_1,\dots,r_{n+1}$.

Then, we can write:

$$P(x)=(x-r_{n+1})\prod_{i=1}^{n}(x-a_i)=(x-r_{n+1})\left(x^n+\sum_{i=1}^{n}(-1)^i s_i x^{n-i}\right)\tag{6.2.102}$$

where we let s_i denote $s_i(r_1,\dots,r_n)$ for brevity. By multiplying out the right hand side:

$$\begin{aligned}P(x)&=(x-r_{n+1})(x^n+\sum_{i=1}^{n}(-1)^i s_i x^{n-i})\\&=x^{n+1}+x\sum_{i=1}^{n}(-1)^i s_i x^{n-i}-r_{n+1}x^n-r_{n+1}\sum_{i=1}^{n}(-1)^i s_i x^{n-i}\\&=x^{n+1}+\sum_{i=1}^{n}(-1)^i s_i x^{n-i+1}-r_{n+1}x^n+r_{n+1}\sum_{i=1}^{n}(-1)^{i+1} s_i x^{n-i}\\&=x^{n+1}+\sum_{i=1}^{n}(-1)^i s_i x^{n-i+1}-r_{n+1}x^n+\sum_{i=1}^{n}(-1)^{i+1} s_i r_{n+1} x^{n-i}\\&=x^{n+1}+\sum_{i=1}^{n}(-1)^i s_i x^{n-i+1}-r_{n+1}x^n+\sum_{i=1}^{n-1}(-1)^{i+1} s_i r_{n+1} x^{n-i}+(-1)^{n+1}s_n r_{n+1}\\&=x^{n+1}+(-1)s_1x^n+\sum_{i=1}^{n-1}(-1)^{i+1} s_{i+1} x^{n-i}-r_{n+1}x^n+\sum_{i=1}^{n-1}(-1)^{i+1} s_i r_{n+1} x^{n-i}+(-1)^{n+1}s_n r_{n+1}\\&=x^{n+1}+(-1)s_1x^n-r_{n+1}x^n+\sum_{i=1}^{n-1}(-1)^{i+1}(s_{i+1}+s_i r_{n+1})x^{n-i}+(-1)^{n+1}s_n r_{n+1}\\&=x^{n+1}-(r_{n+1}+s_1)x^n+\sum_{i=1}^{n-1}(-1)^{i+1}(s_{i+1}+s_i r_{n+1})x^{n-i}+(-1)^{n+1}s_n r_{n+1}\end{aligned}\tag{6.2.103}$$

Since:

$$s_1+r_{n+1}=(r_1+\ldots+r_n)+r_{n+1}=s_1(r_1,\dots,r_{n+1})\tag{6.2.104}$$

and:

$$s_n r_{n+1}=(r_1 r_2\ldots r_n)r_{n+1}=s_{n+1}(r_1,\dots,r_{n+1})\tag{6.2.105}$$

So we get:

$$\begin{aligned}P(x)&=x^{n+1}-(r_{n+1}+s_1)x^n+\sum_{i=1}^{n-1}(-1)^{i+1}(s_{i+1}+s_i r_{n+1})x^{n-i}+(-1)^{n+1}s_n r_{n+1}\\&=x^{n+1}-s_1(r_1,\dots,r_{n+1})x^n+\sum_{i=1}^{n-1}(-1)^{i+1}(s_{i+1}+s_i r_{n+1})x^{n-i}+(-1)^{n+1}s_{n+1}(r_1,\dots,r_{n+1})\end{aligned}\tag{6.2.106}$$

Now it remains that if we can prove that:

$$s_{i+1}+s_i a_{n+1}=s_{i+1}(a_1,\ldots,a_{n+1}) \tag{6.2.107}$$

for all the other i, we conclude the equation holds for $n+1$, hence for all n.

Remember that by definition:

$$s_k(x_1,\ldots,x_n)=\sum_{i\leq i_1<\ldots<i_k\leq n} x_{i_1}x_{i_2}\ldots x_{i_k} \tag{6.2.108}$$

Then:

$$s_{i+1}(r_1,\ldots,r_{n+1})=\sum_{1\leq j_1<\ldots<j_{i+1}\leq n+1} r_{j_1}r_{j_2}\ldots r_{j_{n+1}} \tag{6.2.109}$$

By separating the sum with respect to monomials divisible by r_{n+1}, we see the above is equal to (most of time the best is to check this be using one the previous examples given earlier):

$$\begin{gathered}\sum_{1\leq j_1<\ldots<j_{i+1}\leq n+1} r_{j_1}r_{j_2}\ldots r_{j_{i+1}}=\\ \sum_{1\leq j_1<\ldots<j_{i+1}\leq n} r_{j_1}r_{j_2}\ldots r_{j_{i+1}}+r_{n+1}\sum_{1\leq j_1<\ldots<j_i\leq n} r_{j_1}r_{j_2}\ldots r_{j_i}=s_{i+1}(r_1,\ldots,r_{n+1})+r_{n+1}s_i(r_1,\ldots,r_n)\end{gathered} \tag{6.2.110}$$

so it is clear the relation we wanted holds.

■ Q.E.D.

6.2.2.2 General Vieta's formulas

If we write the second degree equation as following:

$$a_2x^2+a_1x+a_0=0 \tag{6.2.111}$$

We already know that:

$$s_1=r_1+r_2=-\frac{a_1}{a_2} \qquad s_2=r_1r_2=\frac{a_0}{a_2} \tag{6.2.112}$$

And also for a third degree polynomial we saw just before in the examples that if we have:

$$a_3x^3+a_2x^2+a_1x+a_0=0 \tag{6.2.113}$$

then:

$$s_1=r_1+r_2+r_3=-\frac{a_2}{a_3} \quad s_2=r_1r_2+r_1r_3+r_2r_3=\frac{a_1}{a_3} \quad s_3=r_1r_2r_3=-\frac{a_0}{a_3} \tag{6.2.114}$$

We can easily see a pattern emerging that is:

$$\boxed{s_k=\sum_{i\leq i_1<\ldots<i_k\leq n} r_{i_1}r_{i_2}\ldots r_{i_k}=(-1)^k\frac{a_{n-k}}{a_k}} \tag{6.2.115}$$

~90 %

★★★★★

16 votes, 81.25%

6.3. Differential and Integral Calculus

DIFFERENTIAL CALCULUS is one of the most exciting and vast field of mathematics and there is colossal literature on the subject. The results initiated by scientists as Fermat, Newton, Leibniz, Euler and company since the late 117th century (holocene calendar) found absolutely implications in all areas of physics, computer science, electronics, chemistry, finance, biology and mathematics itself.

Mathematicians have written such a lot of theorems since the birth of differential calculus in the mid-116th century (holocene calendar) that the validation of a sample of these theorems is sometimes tricky because requiring for some of them a whole life to be covered (it is a problem that the community of mathematicians recognizes) and checked (so that sometimes nobody checks...).

This fact know, we have chosen here to present only the items absolutely necessary to the understanding of the fundamental tools for the engineer. The purists will excuse us for the moment not to present some theorems that seem indispensable to them but we will prepare once we will have more time...

We will mainly study in what follows what mathematicians like to precisely name (and they are right!): the general cases of real functions of one real variable. More complex functions (several real or complex variables, continuous or discrete) will come once this first part is finished. But the reader can already found theorems with complex variables in the section of Complex Analysis.

Remark

We will not prove in this textbook all the primitives and derivatives of all possible functions because as there are an infinite number of possible functions, there is also an infinite number of derivatives and primitives! It is the role of teachers in educational institutions to train students to apply and understand the reasoning of derivations and integration by applications on known functions (the Internet probably never replace the school at this level).

6.3.1 Differential Calculus

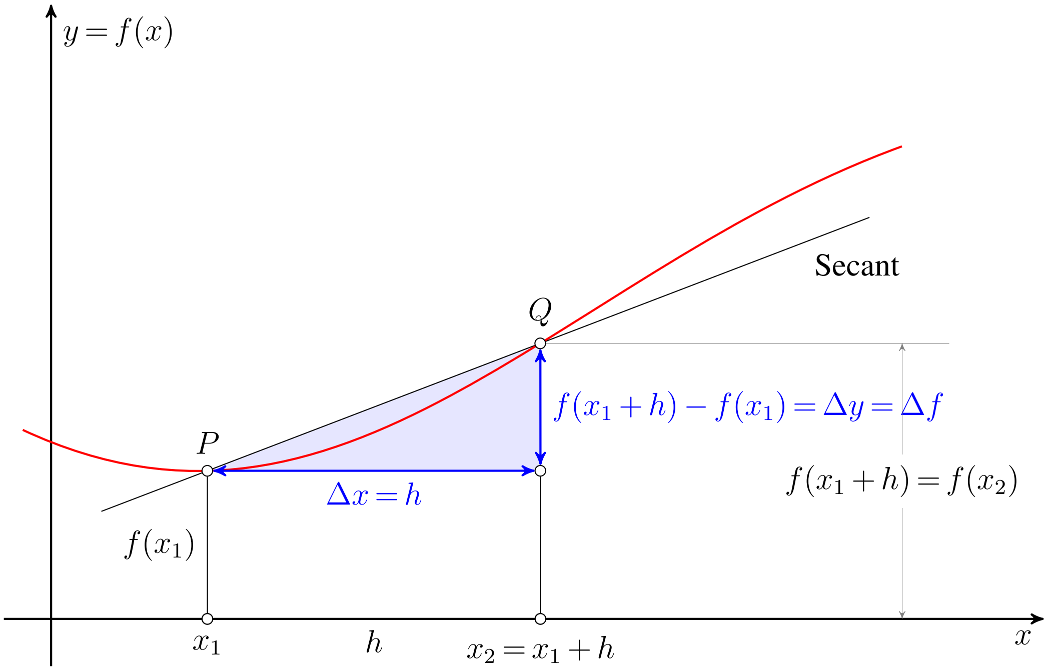
Let f be a real function of one real variable x denoted $f(x)$ (we restrict ourselves to this case for now and we will study the partial derivatives in any number of dimensions later) continue in at least one interval where the horizontal axis contains the value a.

Definitions (#6.3.195):

D1. We name "average slope", or "directing coefficient" the report of the orthogonal projection of two points $x_1 \neq x_2$ of the function f not necessarily continuous on the x-axis and y-axis as:

$$\frac{\Delta f}{\Delta x}=\frac{f(x_2)-f(x_1)}{x_2-x_1}=\frac{f(x_1+h)-f(x_1)}{(x_1+h)-(x_1)}\underset{x_1:=a}{=}\frac{f(a+h)-f(a)}{h} \tag{6.3.1}$$

What can be represented graphically as follows with a specific function:



Remark

Δ named "delta" expresses the fact that we take a difference of the same amount.

We assume as obvious (without proof) that two functions whose slopes are the same in the same interval of definition, are parallel (on a plane).

Remark

We will prove in the chapter of Analytical Geometry page 2804 that two functions whose slopes multiplication is equal -1 are perpendicular.

D2. We name "derivative on a" or "instantaneous slope" or "first derivative", the limit when h tends to 0 (if the limits exists) of the ratio of the orthogonal projection of two points $x_1 \neq x_2$ infinitely close of a continuous function f (in the sense that it does not contain "holes") on the x-axis and y-axis so that:

$$\boxed{\frac{\Delta f}{\Delta x}\underset{h\to 0}{=}\frac{\mathrm{d}f_a}{\mathrm{d}x}=\lim_{h\to 0}=\frac{f(a+h)-f(h)}{h}=f'(a)} \tag{6.3.2}$$

This relation is sometimes named "Newton's difference quotient".

A graphical interpretation gives us that $f'(a)$ is quite obviously the directing coefficient (the slope of the tangent at the point of abscissa a).

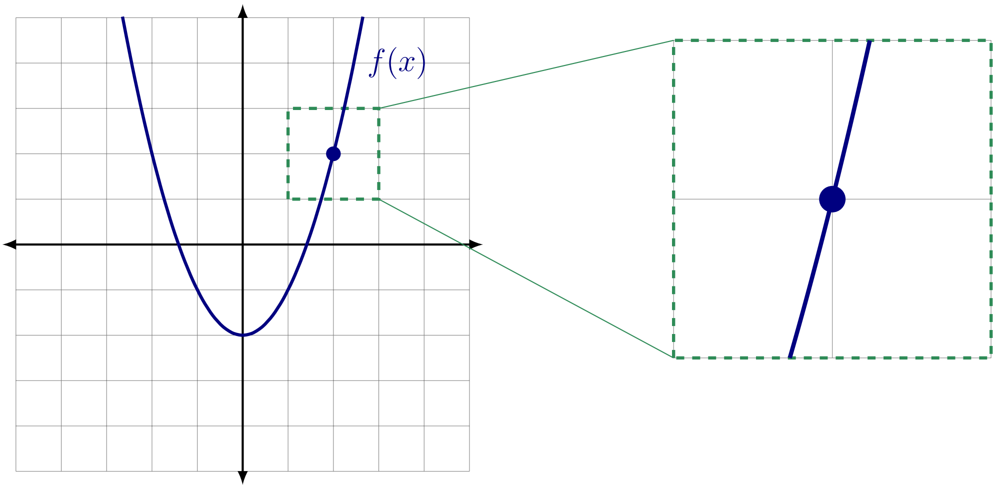


Figure 6.25: A continuous function at the limit looks always like a tangent (author: Lauren K. Williams)

The "tangent" to a curve Γ at a point A is the limit position, when it exists, of the secant line AB when point B of the curve tends towards point A. To be made perfectly rigorous, this definition requires the introduction of notions of topology allowing the calculation of such a limit as we will see in the section of Functional Analysis at page 2226.

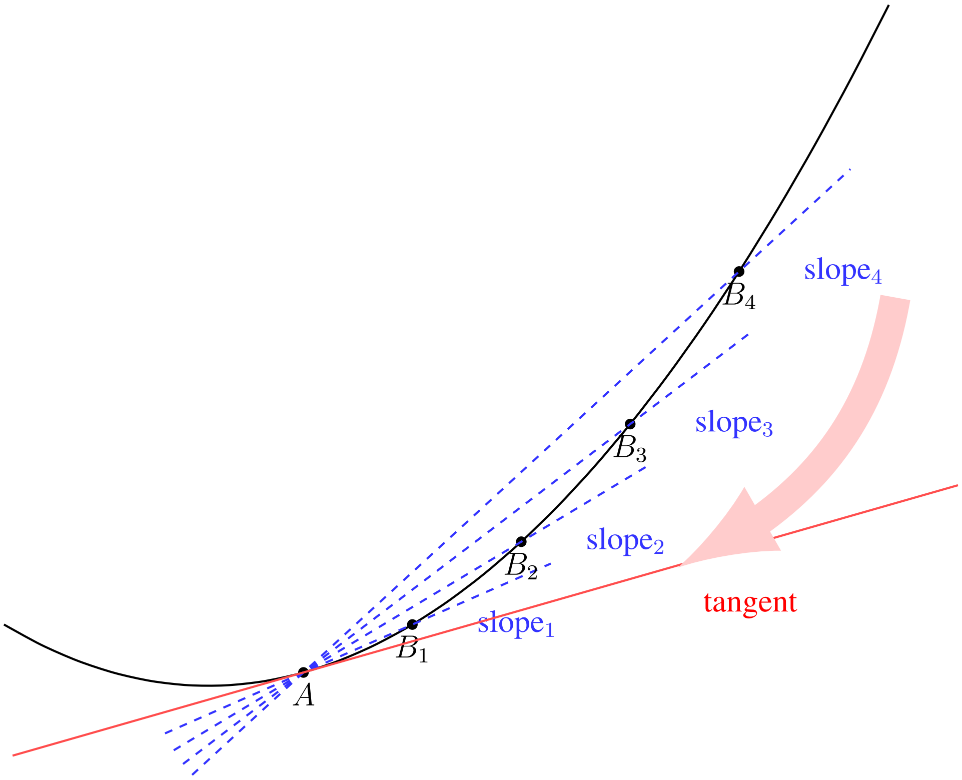


Figure 6.26: Secant vs Tangent

Remarks

R1. The letter d (written upright, not italicized as recommended by the ISO 80000-2 international standard: *Quantities and units - Part 2: Mathematical signs and symbols to be used in the natural sciences and technology*) means here a "differential" and expresses the fact that we take an infinitesimal difference of the same amount.

R2. We refer the reader to the section of Functional Analysis (page 2226) for the definition of a continuous function.

D3. Let f be a function defined on an interval I and differentiable at every point a of I. The function that to any real number of I associates the number $f'(a)$ is named the "derivative function of f on I" and is denoted by f'.

Remark

About the notations of the derivatives... physicists adopt depending on their mood various possible notations. Thus, consider the real function $f(x)$ with one real variable x you can find in the literature and in this textbook the following notations for the first derivative:

$$\frac{\mathrm{d}}{\mathrm{d}x}f(x)=\frac{\mathrm{d}f(x)}{\mathrm{d}x}=\mathrm{d}_x f(x)=\dot{f}(x)=f'(x) \tag{6.3.3}$$

or implicitly assuming that f is a function of x (this gives the opportunity to reduce the size of developments):

$$\boxed{\frac{\mathrm{d}}{\mathrm{d}x}f=\frac{\mathrm{d}f}{\mathrm{d}x}=\mathrm{d}_x f(x)=\dot{f}=f'} \tag{6.3.4}$$

We can in the same way define 2nd order derivatives (derivative of a derivative), the derivatives of order 3 (derivative from a derivative of order 2) and so on. We will frequently meet such derivatives in physics or in pure mathematics for functional analysis.

Note that the derivatives of order two have a very important interpretation in physics and in the field of optimization (see section Numerical Methods page 6673). Indeed, if the sign of the first derivative is positive and then becomes negative (going trough the value of zero) when x increases, then we easily guess we travel through a local maximum of a function (the point where the derivative is zero) and that if the sign of the first derivative becomes negative and positive when x increases, so we travel a local minimum of the function (the point where the derivative is zero). In other words, when the slope changes sign (becomes zero by changing the sign) function passes through an extremum (maximum or minimum) and the tangent is "horizontal" in this point: parallel to the x-axis. We speak then of "turning point":

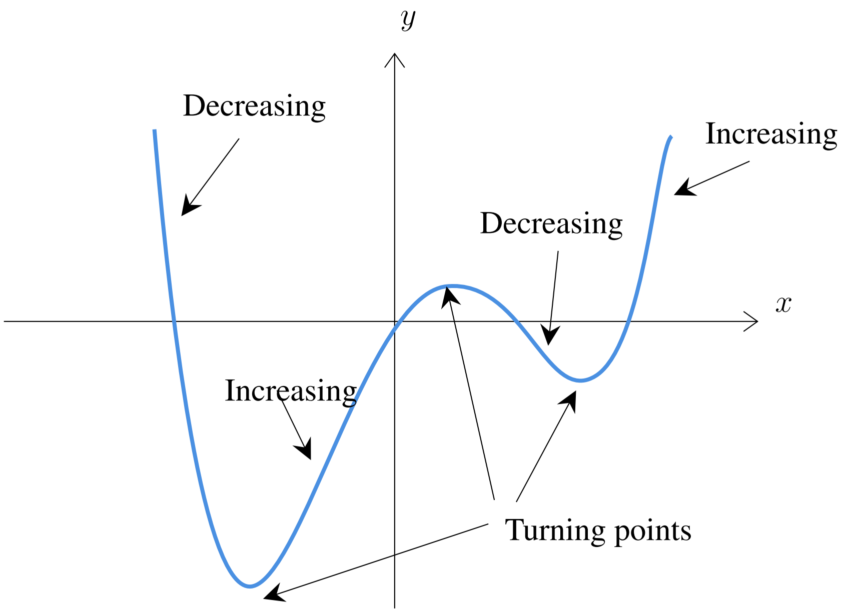


Figure 6.27: **Turning point explicitly highlighted**

But when the 2nd order derivative is zero, that means that the curvature of the function is reversing. We speak then about an "inflection point":

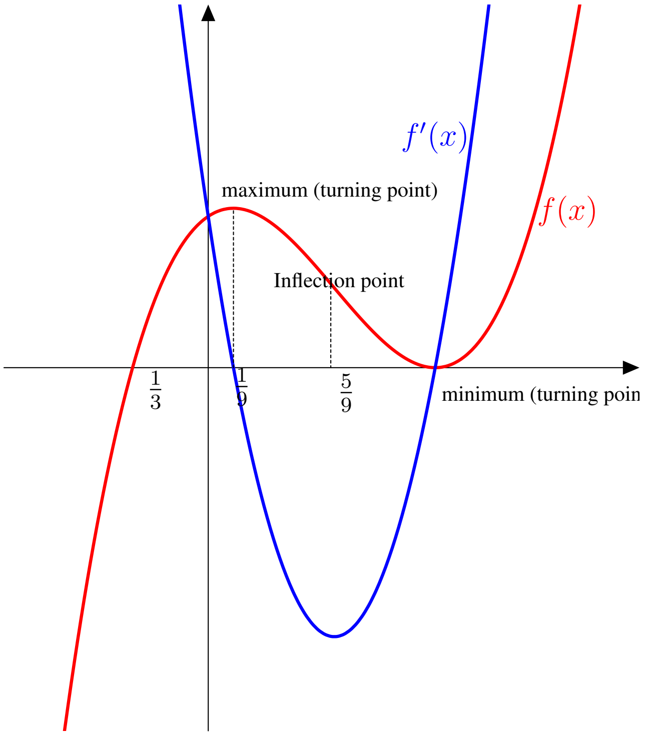
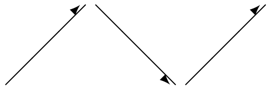


Figure 6.28: **Turning point and inflection point of** $3x^3-5x^2+x+1$

So a very important thing that should always keep in mind (!!!) when you write that the derivative of a function is null, is that we can have a derivative which vanishes at a point without being an extremum (remember that we name this an "inflection point"). To check whether it is really an extremum, we can calculate the second derivative to eliminate the case of an inflection point (because in a inflection point the second derivative will not be null). Otherwise we must use a table of variations to ensure that we are dealing with a maximum or minimum like for example with the function x^3-3x^2+2:

x	$-\infty$		0		2		$+\infty$
f'		$+$	0	$-$	0	$+$	
f							

Which corresponding plot is (with Maple™ 4.00b):

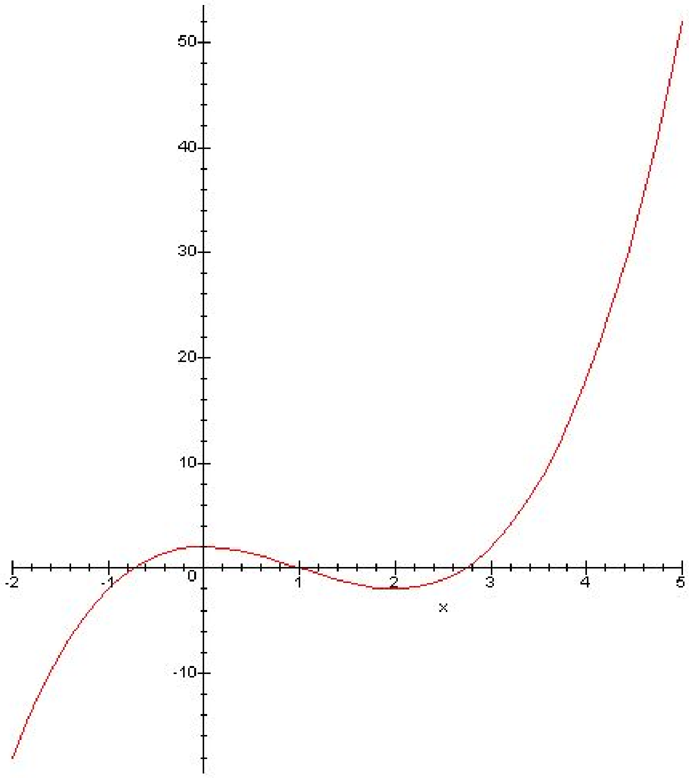


Figure 6.29: **Plot of function** x^3-3x^2+2

Another example with $f(x)=x^4-4x^3+11$ (with Maple™ 4.00b too):

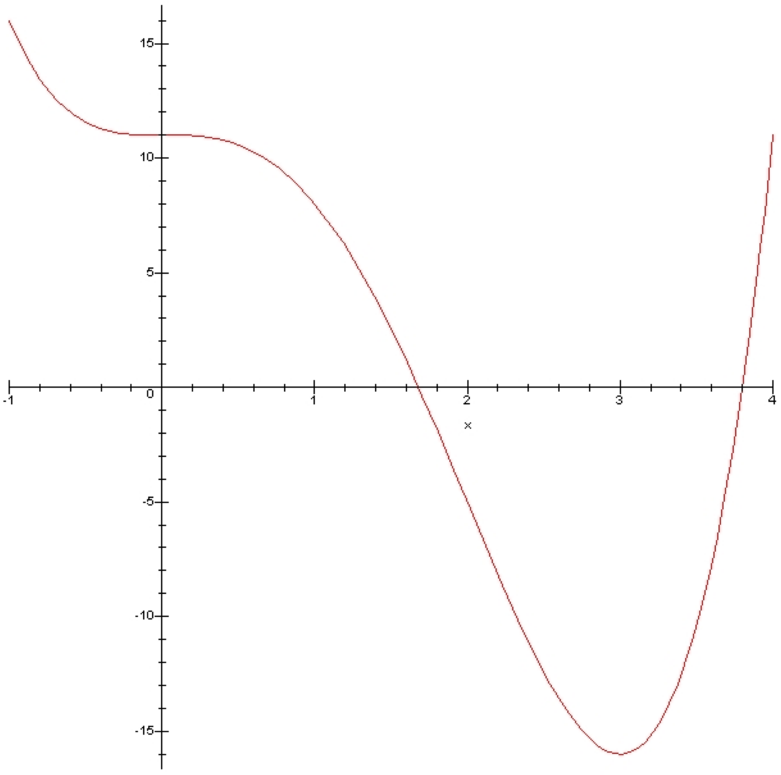


Figure 6.30: **Plot of function** $x^4 - 4x^3 + 11$

With a more detailed variation table:

x	$-\infty$		0		2		3		$+\infty$
$f'(x)$		$-$	0	$-$		$-$	0	$+$	
$f''(x)$		$+$	0	$-$	0	$+$		$+$	
$f(x)$	$+\infty$	↘	11 Σ.K.	↘	-5 Σ.K.	↘	-16 T.E.	↗	$+\infty$

Here is a very entertaining example of a function with its first and second derivatives in Maple™ 4.00:

```
>plot([tanh(x),diff(tanh(x),x),diff(tanh(x),x$2)],x=-5...5,
color=[red,green,blue]);
```

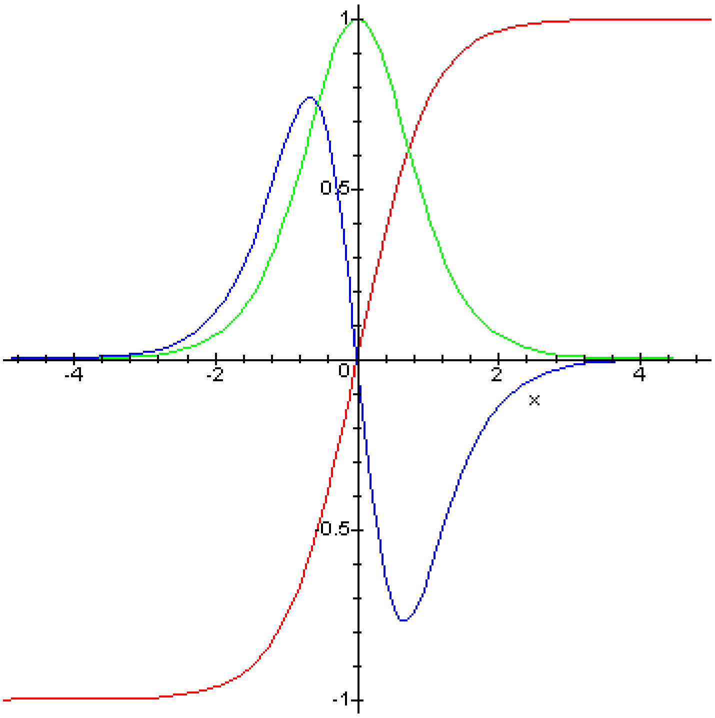


Figure 6.31: **Plot of the hyperbolic tangent function, its first and second derivatives**

The "second derivative" or "second order derivative" is given by:

$$\boxed{f''(x)=\lim_{h\to 0}\dfrac{f(x+h)-2f(x)+f(x-h)}{h^2}} \tag{6.3.5}$$

Indeed, using the Taylor series expansions (see page 1723) of $f(x+h)$ and $f(x-h)$:

$$\begin{aligned}f(x+h)&=f(x)+f'(x)h+f''(x)\frac{h^2}{2}+f'''(x)\frac{h^3}{3!}+\cdots\\ f(x-h)&=f(x)-f'(x)h+f''(x)\frac{h^2}{2}-f'''(x)\frac{h^3}{3!}+\cdots\end{aligned} \tag{6.3.6}$$

Adding the above relations gives:

$$\frac{f(x+h)-2f(x)+f(x-h)}{h^2}=f''(x)+2\frac{f''''(x)}{4!}h^2+\cdots \tag{6.3.7}$$

taking the limit of the above equation as h goes to zero gives the desired result:

$$f''(x)=\lim_{h\to 0}\frac{f(x+h)-2f(x)+f(x-h)}{h^2} \tag{6.3.8}$$

Remark

Two good functions to easy remember the property of the second derivative is $f(x)=x^2$ and $f(x)=-x^2$. As you know the first has a global minimum and the second a global maximum and if you calculate the second derivative for the first one you get a positive constant and a negative constant for the second.

A very important inflection point in business and science is the inflection point of the Normal distribution (see section Statistics page 692).

Given the function of the univariate Gauss distribution:

$$f(x)=\frac{1}{2\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{6.3.9}$$

The points of inflections occur when the second derivative $f''(x)=0$. Then we have to calculate it. First:

$$f'(x)=-\frac{(x-\mu)}{\sigma^2\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{6.3.10}$$

Hence:

$$f''(x)=\frac{(x-\mu-\sigma)(x-\mu+\sigma)}{\sigma^2\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{6.3.11}$$

We put it equal to zero:

$$0=\frac{(x-\mu-\sigma)(x-\mu+\sigma)}{\sigma^2\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{6.3.12}$$

Therefore we have two possible situations:

$$0=(x-\mu-\sigma) \qquad \text{and} \qquad 0=(x-\mu+\sigma) \tag{6.3.13}$$

Hence the two inflection points occur at:

$$x=\mu\pm\sigma \tag{6.3.14}$$

or one standard deviation above and below the mean.

Now, following a problem of understanding of one of our reader in one of the chapters of this textbook, let us specify a technique often used by physicists. Consider a derivative of order 2 as:

$$\ddot{f}=f''=\frac{\mathrm{d}^2f(x)}{\mathrm{d}x}=\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\mathrm{d}f(x)}{\mathrm{d}x}\right)=\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\mathrm{d}}{\mathrm{d}x}f(x)\right) \tag{6.3.15}$$

If we look at the $\dfrac{\mathrm{d}}{\mathrm{d}x}$ as a differential operator (what it is!) we can obviously write:

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\mathrm{d}}{\mathrm{d}x}f(x)\right)=\frac{\mathrm{d}}{\mathrm{d}x}\frac{\mathrm{d}}{\mathrm{d}x}(f(x))=\left(\frac{\mathrm{d}}{\mathrm{d}x}\right)^2f(x) \tag{6.3.16}$$

Finally we have:

$$\frac{\mathrm{d}^2f(x)}{\mathrm{d}x}=\left(\frac{\mathrm{d}}{\mathrm{d}x}\right)^2f(x) \tag{6.3.17}$$

and so it comes after simplification by $f(x)$:

$$\boxed{\frac{\mathrm{d}^2}{\mathrm{d}x}=\left(\frac{\mathrm{d}}{\mathrm{d}x}\right)^2} \tag{6.3.18}$$

otherwise we can not have this equality if the operator acts explicitly on a function in a math or any physical relation.

This may seem obvious to some but sometimes less so for others ... and it is clearly useful to clarify this because it is often used in the sections of Special Relativity, General Relativity, Corpuscular Quantum Physics and Wave Quantum Physics.

Let us now indicate and prove two properties intuitively obvious of derivative which will be essential for us several times some proofs in this textbook (for example in the section of Numerical Methods or just simple in this section...).

Theorem 6.8. Consider first two real numbers $a<b$ and f a continuous real-valued function on the closed interval $[a,b]$ and differentiable on the open interval $]a,b[$ such that $f(a)=f(b)$. So we want to prove that there is obviously at least one element $c\in]a,b[$ such that $f'(c)=0$ (this is typically the case of polynomial functions!).

This property is named "Rolle's theorem" and therefore it explicitly shows that there is at least one element where the derivative of f is zero when we browsing its path we return back to the same value images for two distinct values of the abscissas (pre-images), that is to say that there exists at least one point where the tangent is horizontal.

Proof 6.8.1. First, if f is constant, the result is immediate... Otherwise, as f is continuous on the closed interval $[a,b]$ it admits at least one minimum or maximum considering that we rely on the assumption that $f(a)=f(b)$ and that f is not constant. The extrema is reached at a point c belonging to the open interval $]a,b[$ (the fact of taking an open interval allows in some cases to avoid having an extrema again in a or b).

Suppose as a first case that $f(c)$ is global maximum in the interval. The derivative of the function f between c and a second point a has a known sign:

- For h strictly positive and such that $c+h$ still belongs to the interval $[a,b]$:

  $$\frac{f(c+h)-f(c)}{h}<0 \tag{6.3.19}$$

  Considering the limit when h tends to 0, the derivative valuation on $f'(c)$ is thus negative.

- For h strictly negative and such that $c+h$ still belongs to the interval $[a,b]$:

  $$\frac{f(c+h)-f(c)}{h}>0 \tag{6.3.20}$$

  Considering the limit when h tends to 0, the derivative valuation on $f'(c)$ is thus positive.

Ultimately, the derivative of f is zero at point c.

The proof is analogous if $f(c)$ is a minimum in the interval, with the signs of derivatives that are opposite.

■ Q.E.D.

It should be quite obvious to the reader that in the general case, the change of basis doesn't conserve the norm!

Let us now consider two real numbers a,b and $f(x)$ a continuous function on $[a,b]$ (that is to say $\mathcal{C}([a,b])$) and differentiable on $]a,b[$.

Theorem 6.9. So we propose to show that there is at least one real number $c \in]a,b[$ such that:

$$f(b)-f(a)=f'(c)(b-a) \tag{6.3.21}$$

This can also be written as follows:

$$f(x_0+k)-f(x_0)=f'(x_0+s\cdot k)k \tag{6.3.22}$$

with $s\in]0,1[$.

Since the term on the left represents a finite increase of the term on the right, then this result is named "mean value theorem" or better "theorem of finite increments".

Geometrically this means that on at least one point c of the graph of the function $f(x)$, there is a tangent with a director coefficient of:

$$\frac{f(b)-f(a)}{b-a} \tag{6.3.23}$$

Graphically this gives:

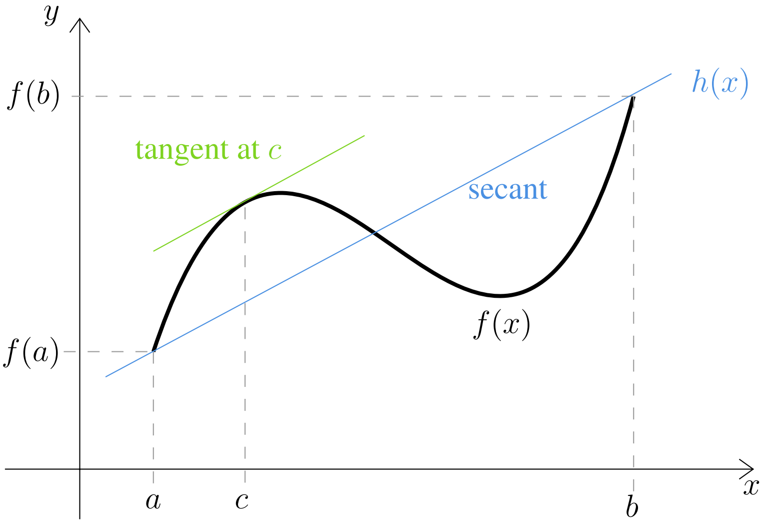


Figure 6.32: **Graphical representation of Rolle's theorem**

Proof 6.9.1. We first have:

$$h(x)=(x-a)\frac{f(b)-f(a)}{b-a}+f(a) \tag{6.3.24}$$

because the slope of $h(x)$ is obviously:

$$\frac{f(b)-f(a)}{b-a} \tag{6.3.25}$$

and as we must have $f(a)$ when $x=a$ it follows the relation given previously.

Then, to show that such a c value exists, the idea is to bring the two points a and b in the same ordinate making this brings us back to Rolle's theorem and for that, we define a function $g(x)$ by:

$$g(x)=f(x)-h(x)=f(x)-\left[(x-a)\frac{f(b)-f(a)}{b-a}+f(a)\right] \tag{6.3.26}$$

which is such that indeed $g(a)=g(b)$ ... is in this case equal to 0 (but this value is not relevant).

Therefore, the Rolle's theorem discussed above indicates that there is a point between a and b where the derivative of $g(x)$ is zero such that $g'(c)=0$. And by seeing that:

$$g'(x)=f'(x)-\frac{f(b)-f(a)}{b-a} \tag{6.3.27}$$

we get:

$$g'(c)=f'(c)-\frac{f(b)-f(a)}{b-a}=0 \tag{6.3.28}$$

Therefore after simplification:

$$f(b)-f(a)=f'(c)(b-a) \tag{6.3.29}$$

■ Q.E.D.

Using this little theorem and mathematical tools introduced earlier, we can build a little theorem useful and powerful physics.

Definition (#6.3.196): We name "Hôpital's rule" (also sometimes named "Bernoulli's rule") the method that uses the derivative in order to determine the boundaries difficult to calculate with most quotients which often appear in physics (the rule is named after the 117th-century - according to holocene calendar - French mathematician Guillaume de l'Hôpital. Although the rule is often attributed to l'Hôpital, the theorem was first introduced to him in 11694 by the Swiss mathematician Johann Bernoulli).

Proof 6.9.2. Consider two functions $f(x)$ and $g(x)$ and such that $f(a)=g(a)=0$ so we can write:

$$\frac{f(x)}{g(x)}=\frac{f(x)}{g(x)}\frac{x-a}{x-a}=\frac{f(x)}{x-a}\frac{x-a}{g(x)}=\frac{f(x)-f(a)}{x-a}\frac{x-a}{g(x)-g(a)} \tag{6.3.30}$$

Then according to the definition of the derivative:

$$\lim_{x\to a}\frac{f(x)}{g(x)}=\lim_{x\to a}\frac{f(x)-f(a)}{x-a}\frac{x-a}{g(x)-g(a)}=\frac{f(a+h)-f(a)}{(a+h)-h}\frac{(a+h)-h}{g(a+h)-g(a)}=\frac{f'(a)}{g'(a)} \tag{6.3.31}$$

■ Q.E.D.

We can generalize this previous result initially based on a little too strong constraint:

$$f(a)=g(a)=0 \tag{6.3.32}$$

Proof 6.9.3. Let us recall that according to the mean value theorem, if $f(x)$ is differentiable on an interval $]a,b[$ and continuous on $[a,b]$ then there is a real c in the interval $[a,b]$ such that:

$$f'(c)=\frac{f(b)-f(a)}{b-c} \tag{6.3.33}$$

If the theorem is true for two functions satisfying the same constraints then we have two functions such as:

$$f'(c)=\frac{f(b)-f(a)}{b-c} \quad \text{and} \quad g'(c)=\frac{f(b)-f(a)}{b-c} \tag{6.3.34}$$

If $g'(c)$ is not zero then we have the right to write the ratio (some name this the "generalized mean value theorem"...):

$$\frac{f'(c)}{g'(c)}=\frac{f(b)-f(a)}{b-a}\frac{b-a}{g(b)-g(a)}=\frac{f(b)-f(a)}{g(b)-g(a)} \tag{6.3.35}$$

which without losing validity as c is in the range $[a,x]$ can be written:

$$\frac{f'(c)}{g'(c)}=\frac{f(x)-f(a)}{g(x)-g(a)} \tag{6.3.36}$$

Therefore, when $x\rightarrow a$ that implies the range $[a,x]$ is always smaller and thus $c\rightarrow a$ we have:

$$\begin{aligned}\lim_{x\rightarrow a}&=\frac{f'(c)}{g'(c)}=\frac{f(a+h)-f(a)}{g(a+h)-g(a)}=\frac{f(a+h)-f(a)}{(a+h)-a}\frac{(a+h)-a}{g(a+h)-g(a)}\\&=\frac{f'(a)}{g'(a)}\end{aligned} \tag{6.3.37}$$

So we just proved now that in the first simplified proof of l'Hôpital's rule of the relation:

$$\frac{f(a+h)-f(a)}{(a+h)-a}\frac{(a+h)-a}{g(a+h)-g(a)}=\frac{f'(a)}{g'(a)} \tag{6.3.38}$$

we had is true in general and that it is not necessary that $f(a)=g(a)=0$ is true for the result to be fair!

■ Q.E.D.

Remark

The result can be generalized, still under the same condition, with:

$$\lim_{x\to a}\frac{f(x)}{g(x)}=\lim_{x\to a}\frac{f'(x)}{g'(x)}=\lim_{x\to a}\frac{f''(x)}{g''(x)}=\ldots=\lim_{x\to a}\frac{f^{(k)}(x)}{g^{(k)}(x)}=\lim_{x\to a}\frac{f^{(k+1)}(x)}{g^{(k+1)}(x)} \tag{6.3.39}$$

✍**Examples:**

E1. We want to calculate:

$$\lim_{x\to 2}\frac{x^2+x-6}{x^2-4} \tag{6.3.40}$$

There is two ways to solve that! First we can factorize:

$$\lim_{x\to 2}\frac{x^2+x-6}{x^2-4}=\lim_{x\to 2}\frac{(x+3)(x-2)}{(x+2)(x-2)}=\lim_{x\to 2}\frac{x+3}{x-2}=\frac{5}{4} \tag{6.3.41}$$

Or using l'Hôpital's rule (we see that the assumption $f(2)=g(2)=0$ is satisfied!):

$$\lim_{x\to 2}\frac{x^2+x-6}{x^2-4}=\lim_{x\to 2}\frac{2x+1}{2x}=\frac{5}{4} \tag{6.3.42}$$

But we see that we can't applied l'Hôpital's rule twice as assumptions would not be met.

E2. We want to calculate:

$$\lim_{x\to +\infty}\frac{x^2-4}{2+x-4x^2} \tag{6.3.43}$$

We obviously can't apply l'Hôpital's rule as $f(+\infty)\neq g(+\infty)\neq 0$. The factorization technique will also fail. We have to use a power trick.

We divide by the highest denominator power:

$$\lim_{x\to +\infty}\frac{x^2-4}{2+x-4x^2}=\lim_{x\to +\infty}\frac{1-\dfrac{4}{x^2}}{\dfrac{2}{x^2}+\dfrac{1}{x}-4}=-\frac{1}{4} \tag{6.3.44}$$

We will also caution here the previous "divide by highest power rule" is applicable in the specific scenario where you are taking a limit at infinity or minus infinity AND the limit expression is of the form $P(x)/Q(x)$ where both P and Q are linear combinations of powers of x and the largest power of x appearing in the entire expression is non-negative.

Caution!!! L'Hôpital's Rule is only for single-variable functions. In multivariable calculus, pathdependence must be checked. This limit depends on the path taken, so the limit does not exist.

6.3.1.1 Differentials

We noted earlier what is a differential d. But there are actually several different types of differential of a function (note that we distinguish the masculine and feminine gender of the word):

1. Differentials
2. Partial differentials
3. Total exact differentials
4. Total inexact differentials

Remember as seen at the beginning of this section hat we name "differential $\mathrm{d}f$" of an univariate function the relation given by:

$$\boxed{\mathrm{d}f = f'\mathrm{d}x} \tag{6.3.45}$$

However, to express the effect of changing all the variables of a multivariate function f, we must use another type of differential which we name the "total differential" (derived into two subfamilies: total exact differential and total inexact differential).

Take, for example, a function $f(x,y)$ of the two variables x and y. The increase $\mathrm{d}f$ of the function f, for a finite increase of x to $x+\Delta x$ and of y to $y+\Delta y$ is obviously given by:

$$\Delta f(x) = f(x+\Delta x),y+\Delta y) - f(x,y) \tag{6.3.46}$$

we can also write:

$$\Delta f(x) = f(x+\Delta x),y+\Delta y) - f(x,y+\Delta y) + f(x,y+\Delta y) - f(x,y) \tag{6.3.47}$$

Or also:

$$\Delta f(x) = \frac{f(x+\Delta x),y+\Delta y) - f(x,y+\Delta y)}{\Delta x}\Delta x + \frac{f(x,y+\Delta y) - f(x,y)}{\Delta y}\Delta y \tag{6.3.48}$$

For infinitely small increments of x and y:

$$\mathrm{d}f = \lim_{\Delta x\to 0}\left(\frac{f(x+\Delta x),y+\Delta y) - f(x,y+\Delta y)}{\Delta x}\Delta x\right) + \lim_{\Delta y\to 0}\left(\frac{f(x,y+\Delta y) - f(x,y)}{\Delta y}\Delta y\right) \tag{6.3.49}$$

Let us therefore focus on the two terms when going to the limit:

$$\frac{f(x+\mathrm{d}x),y+\mathrm{d}y) - f(x,y+\mathrm{d}y)}{\mathrm{d}x} \quad \text{and} \quad \frac{f(x,y+\mathrm{d}y) - f(x,y)}{\mathrm{d}y} \tag{6.3.50}$$

The first term on the left, we see it clearly, finally gives the change in x of the function $f(x,y)$ by having y fixed on the variation. We denote therefore this by (if the fixed variables are trivial we do note write them):

$$\left(\frac{\partial f}{\partial x}\right)_y = \left(\frac{\partial f}{\partial x}\right) = \frac{f(x+\mathrm{d}x),y+\mathrm{d}y) - f(x,y+\mathrm{d}y)}{\mathrm{d}x} \tag{6.3.51}$$

and even:

$$\left(\frac{\partial f}{\partial y}\right)_x = \left(\frac{\partial f}{\partial y}\right) = \frac{f(x,y+\mathrm{d}y) - f(x,y)}{\mathrm{d}y} \tag{6.3.52}$$

Remark

When a variable is fixed to study the variation of the other, some authors and teachers of older generations like to say, "all things being equal f varies in function of ... as ...". In short, it is an expression that can be found in other areas of mathematics (such as multivariate linear regressions) but that is disappearing...

Both expressions:

$$\left(\frac{\partial f}{\partial x}\right)_y \quad \text{and} \quad \left(\frac{\partial f}{\partial y}\right)_x \tag{6.3.53}$$

are what we name "partial differential" or just "partial derivative" (whose practical application case of the simplest and probably the most interesting and pedagogically relevant example available today in this entire book is the supply chain Wilson's model with rupture presented in the section of Quantitative Management).

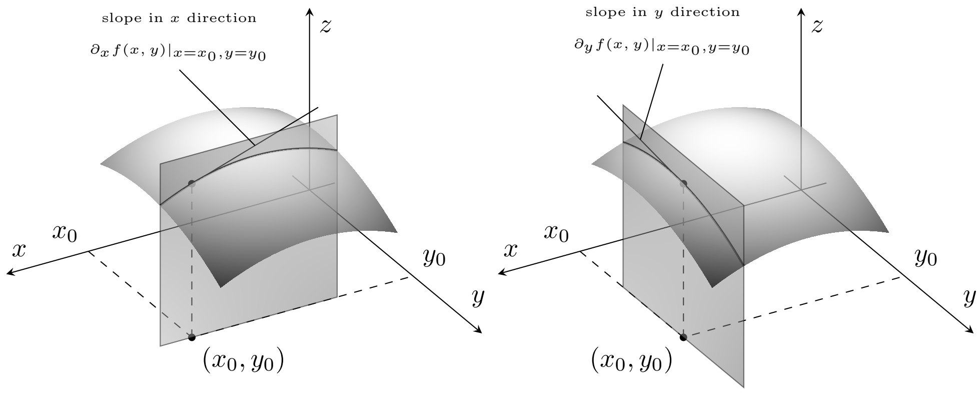


Figure 6.33: **Partial derivative illustrated**

✍**Example:**

Consider $f(x,y)=x^2+y^3+e^{xy^2}$. Computing the partial derivative is equivalent to computing the regular derivative with the other variables treated as constants. e.g.:

$$\left.\frac{\partial f}{\partial x}\right|_y = 2x+y^2e^{xy^2}$$

Second and mixed partial derivatives can also be computed:

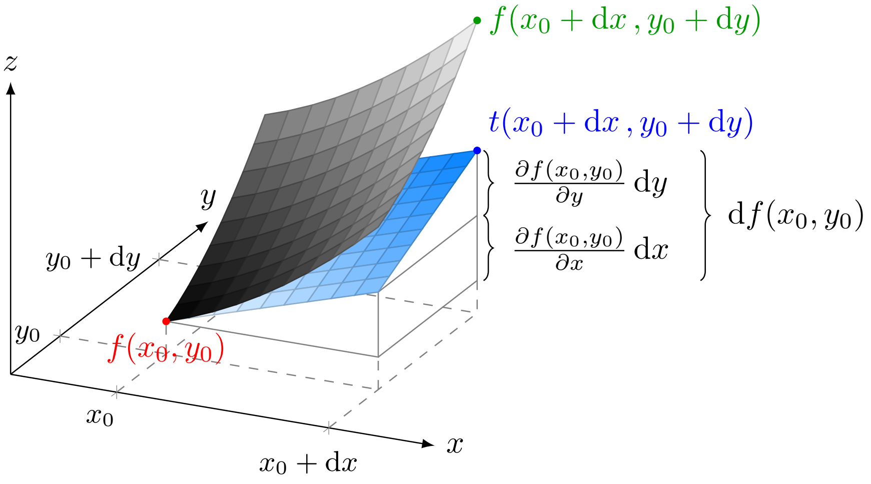
$$\begin{gathered}\frac{\partial^2 f}{\partial x^2}=2+y^4e^{xy^2}\\ \frac{\partial^2 f}{\partial y\partial x}=\frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right)=2ye^{xy^2}+2xy^3e^{xy^2}\end{gathered}$$

It is often cumbersome to write out the variables that are kept constant. If all other variables are being held constant, we simply don't write them out, and just say $\frac{\partial f}{\partial x}$.

We have therefore:

$$\boxed{\mathrm{d}f=\frac{\partial f}{\partial x}\mathrm{d}x+\frac{\partial f}{\partial y}\mathrm{d}y} \tag{6.3.54}$$

that is the "differential of f". The thermodynamicists them often talk of "total exact differential of f" or just "exact differential of f" or even simple "total derivative" and also "exterior derivative".



$f(x_0+\mathrm{d}x\,,y_0+\mathrm{d}y)$
z
$t(x_0+\mathrm{d}x\,,y_0+\mathrm{d}y)$
$\frac{\partial f(x_0,y_0)}{\partial y}\,\mathrm{d}y$
$\frac{\partial f(x_0,y_0)}{\partial x}\,\mathrm{d}x$
$\mathrm{d}f(x_0,y_0)$
y
$y_0+\mathrm{d}y$
y_0
$f(x_0,y_0)$
x_0
x
$x_0+\mathrm{d}x$

Figure 6.34: Total differential illustrated

There is also a shorter way (a matter of notation) to get the same result by considering an arbitrary displacement in any direction $(x,y)\rightarrow(x+\delta x,y+\delta y)$.

We have:

$$\begin{aligned}\delta f&=f(x+\delta x,y+\delta y)-f(x,y)\\&=f(x+\delta x,y+\delta y)-f(x+\delta x,y)+f(x+\delta x,y)-f(x,y)\\&=f_y(x+\delta x,y)\delta y+o(\delta y)+f_x(x,y)\delta x+o(\delta x)\\&=(f_y(x,y)+o(1))\,\delta y+o(\delta y)+f_x(x,y)\delta x+o(\delta x)\\&=\frac{\partial f}{\partial x}\delta x+\frac{\partial f}{\partial y}\delta y+o(\delta x,\delta y)\end{aligned} \tag{6.3.55}$$

If we take the limit as $\delta x,\delta y\rightarrow 0$ we fall back on the above boxed relation.

The previous relation is a special case of what mathematicians name very generally a "differential form":

$$\boxed{\mathrm{d}f=M(x,y)\mathrm{d}x+N(x,y)\mathrm{d}y} \tag{6.3.56}$$

we will come back about this a little further below...

It is customary to write:

$$\vec{F}=\begin{pmatrix}M(x,y)\\N(x,y)\end{pmatrix} \tag{6.3.57}$$

that is to say under vector field way.

It is important to remember the form of the total derivative because we will meet it again almost everywhere in special operators in quantum physics, in fluid mechanics, in electrodynamics, in thermodynamics, in economy, etc.

Geometrically, the partial derivatives can be interpreted as follows: the function $f(x,y)$ defines typically a surface in $\mathbb{R}$ whose intersection with the plane $y=y_0$ is a curve given by $f(x,y_0)$.

The partial derivative $\partial_x f$ is then the slope of this curve at every point x. We then have naturally the following function for the slope at the point (x_0,y_0):

$$f(x,y_0)=f(x_0,y_0)+\frac{\partial f}{\partial x}(x_0,y_0)(x-x_0) \tag{6.3.58}$$

Similarly, the tangent to the curve $f(x_0,y)$ will be given by:

$$f(x_0,y)=f(x_0,y_0)+\frac{\partial f}{\partial y}(x_0,y_0)(y-y_0) \tag{6.3.59}$$

The plane locally tangent to the point (x_0,y_0) determined by its two tangents is then given by:

$$f(x_0,y_0)=f(x_0,y_0)+\frac{\partial f}{\partial x}(x_0,y_0)(x-x_0)+\frac{\partial f}{\partial y}(x_0,y_0)(y-y_0) \tag{6.3.60}$$

Reorganizing the terms as:

$$f(x_0,y_0)-f(x_0,y_0)=\frac{\partial f}{\partial x}(x_0,y_0)(x-x_0)+\frac{\partial f}{\partial y}(x_0,y_0)(y-y_0) \tag{6.3.61}$$

We recognize:

$$\mathrm{d}f=\frac{\partial f}{\partial x}\mathrm{d}x+\frac{\partial f}{\partial y}\mathrm{d}y \tag{6.3.62}$$

Thus, for example, the surface represented by the function:

$$f(x,y)=e^{-x^2-y^2} \tag{6.3.63}$$

is shown below with the two tangents passing through the point:

$$(x_0,y_0)=(0.3,0.1) \tag{6.3.64}$$

and that respective equations are:

$$\begin{aligned}f(x,y_0)=f(x_0,y_0)+\frac{\partial f}{\partial x}(x_0,y_0)(x-x_0)&=e^{-x^2-y^2}\Big|_{0.3,0.1}-2xe^{-x^2-y^2}\Big|_{0.3,0.1}(x-x_0)\\&\cong 0.905-0.543(x-0.3)=-0.543x+1.0679\end{aligned} \tag{6.3.65}$$

and:

$$\begin{aligned}f(x_0,y)=f(x_0,y_0)+\frac{\partial f}{\partial y}(x_0,y_0)(y-y_0)&=e^{-x^2-y^2}\Big|_{0.3,0.1}-2ye^{-x^2-y^2}\Big|_{0.3,0.1}(y-y_0)\\&\cong 0.905-0.181(x-0.1)=-0.181x+0.9231\end{aligned} \tag{6.3.66}$$

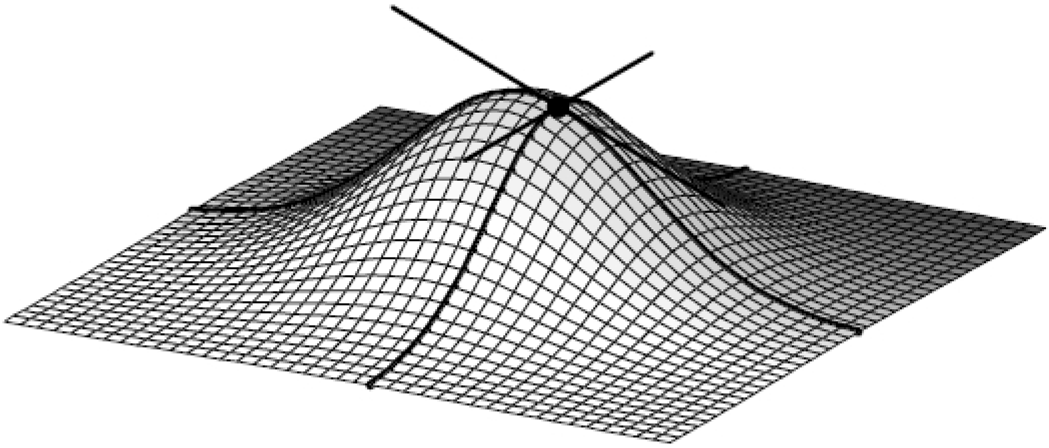


Figure 6.35: **Both partial derivatives tangents of the function at the point of interest**

The "tangent plane" at this point is then given by:

$$\begin{aligned} f(x_0,y_0) &= f(x_0,y_0)+\frac{\partial f}{\partial x}(x_0,y_0)(x-x_0)+\frac{\partial f}{\partial y}(x_0,y_0)(y-y_0) \\ &\cong 0.905-0.543(x-0.3)-0.181(y-0.1)=-0.543x-0.181y+1.086 \end{aligned} \tag{6.3.67}$$

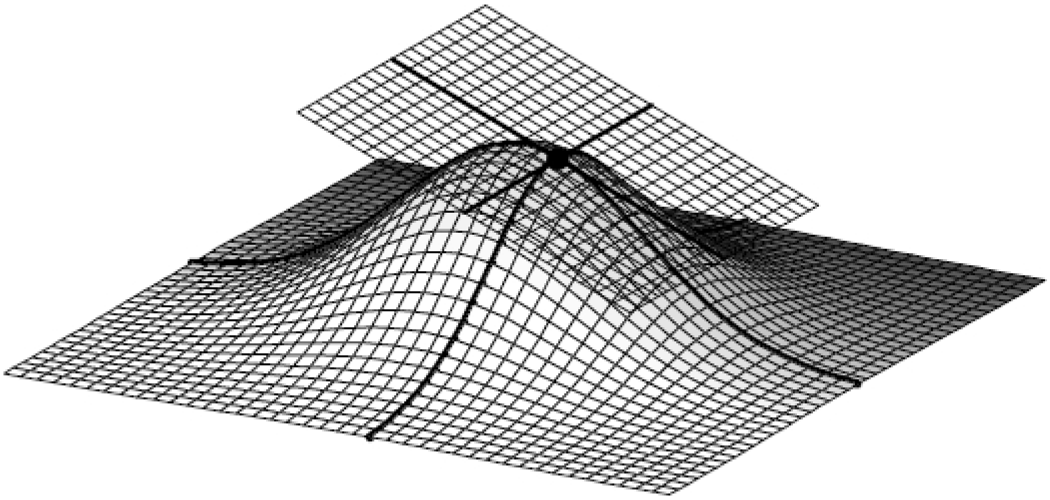


Figure 6.36: **The two tangents of the function at the point of interest with the tangent plane**

Remark

Similarly, for a function of more than two variables, for example $f(x,y,z)$, the total derivative $\mathrm{d}f$ is given by:

$$\mathrm{d}f=\left(\frac{\partial f}{\partial x}\right)_{y,z}\mathrm{d}x+\left(\frac{\partial f}{\partial y}\right)_{x,z}\mathrm{d}y+\left(\frac{\partial f}{\partial z}\right)_{x,y}\mathrm{d}z=\frac{\partial f}{\partial x}\mathrm{d}x+\frac{\partial f}{\partial y}\mathrm{d}y+\frac{\partial f}{\partial z}\mathrm{d}z \tag{6.3.68}$$

In the above equation, the differential $\mathrm{d}f$ was calculated from the expression of the function f. Since there is a function f satisfying an expression $\mathrm{d}f$, the differential $\mathrm{d}f$ is then, as we know, named the "total exact derivative".

We take the opportunity to make an important indication on the use of partial derivatives by physicists (and therefore in many chapter of this textbook). We have seen that if f depends on two variables x,y we have:

$$\mathrm{d}f=\dfrac{\partial f}{\partial x}\mathrm{d}x+\dfrac{\partial f}{\partial y}\mathrm{d}y \tag{6.3.69}$$

and if it depends of only one variable we have then:

$$\mathrm{d}f=\dfrac{\partial f}{\partial x}\mathrm{d}x \tag{6.3.70}$$

And therefore in the univariate case:

$$\dfrac{\mathrm{d}f}{\mathrm{d}x}=\dfrac{\partial f}{\partial x} \tag{6.3.71}$$

this is why many physicists mixed the two notations...

Now you have to know however that there are also total exact derivative that no functions satisfies. In this case, we speak about "total inexact derivative" and for determining whether a total derivative is exact or inexact, we use the properties of partial derivatives (very important case in thermodynamics!!!).

Given the famous general differential form (notation used typically in differential geometry):

$$\mathrm{d}z=M(x,y)\mathrm{d}x+N(x,y)\mathrm{d}y \tag{6.3.72}$$

where $M(x,y)$ and $N(x,y)$ are functions of the variables x and y. If $\mathrm{d}z$ is an exact total differential, then:

$$\mathrm{d}z=\dfrac{\partial z}{\partial x}\mathrm{d}x+\dfrac{\partial z}{\partial y}\mathrm{d}y \tag{6.3.73}$$

this requires verbatim that:

$$M(x,y)=\dfrac{\partial z}{\partial x}\qquad N(x,y)=\dfrac{\partial z}{\partial y} \tag{6.3.74}$$

or, by performing a second partial derivative that:

$$\dfrac{\partial}{\partial y}\dfrac{\partial z}{\partial x}=\dfrac{\partial^2 z}{\partial x\partial y}=\dfrac{\partial}{\partial y}M(x,y)=\dfrac{\partial M}{\partial y}\qquad \dfrac{\partial}{\partial x}\dfrac{\partial z}{\partial y}=\dfrac{\partial^2 z}{\partial x\partial y}=\dfrac{\partial}{\partial x}N(x,y)=\dfrac{\partial N}{\partial x} \tag{6.3.75}$$

so that the differential form is an exact total derivative.

Before continuing, we need a result given by the "Schwarz's theorem" (but which was proved in the late 117th century by one of the Bernoulli brothers) which is the following:

Theorem 6.10. Given a function f, if:

$$f,\dfrac{\partial f}{\partial x},\dfrac{\partial f}{\partial y},\dfrac{\partial^2 f}{\partial x\partial y},\dfrac{\partial^2 f}{\partial y\partial x} \tag{6.3.76}$$

are continuous (we must really check that this assumption is true!) then we get a very important result in practice:

$$\dfrac{\partial^2 f}{\partial x\partial y}(x_0,y_0)=\dfrac{\partial^2 f}{\partial y\partial x}(x_0,y_0) \tag{6.3.77}$$

for every $(x_0,y_0)\in U$ where U is the domain of definition where f is continuous (and therefore assumed to be differentiable).

Proof 6.10.1. We consider the expression:

$$H=f(x_0+k,y_0+h)-f(x_0+k,y_0)-f(x_0,y_0+h)+f(x_0,y_0) \tag{6.3.78}$$

Let us write:

$$\begin{aligned}g(x)&:=f(x,y_0+h)-f(x,y_0)\\ w(x)&:=f(x_0+k,y)-f(x_0,y)\end{aligned} \tag{6.3.79}$$

Then we have:

$$H=g(x_0+k)-g(x_0)=w(y_0+h)-w(y_0) \tag{6.3.80}$$

By the mean value theorem:

$$\begin{aligned}H&=g(x_0+k)-g(x_0)=kg'(x_0+sk)\\ H&=w(y_0+h)-w(y_0)=hw'(y_0+th)\end{aligned} \tag{6.3.81}$$

With $s,t\in]0,1[$. By taking the definitions of w and g we get:

$$\begin{aligned}H&=kg'(x_0+sk)=k\left(\frac{\partial f}{\partial x}(x_0+sk,y_0+h)-\frac{\partial f}{\partial x}(x_0+sk,y_0)\right)\\ H&=hw'(y_0+th)=h\left(\frac{\partial f}{\partial y}(x_0+k,y_0+th)-\frac{\partial f}{\partial y}(x_0,y_0+th)\right)\end{aligned} \tag{6.3.82}$$

by applying the again the mean value theorem to both sides in brackets we find:

$$\begin{aligned}H&=kh\frac{\partial^2 f}{\partial y\partial x}(x_0+sk,y_0+\tilde{s}h)\\ H&=hk\frac{\partial^2 f}{\partial x\partial y}(x_0+\tilde{t}k,y_0+th)\end{aligned} \tag{6.3.83}$$

With $\tilde{s},\tilde{t}\in]0,1[$. To finish we see that we have:

$$\frac{\partial^2 f}{\partial y\partial x}(x_0+sk,y_0+\tilde{s}h)=\frac{\partial^2 f}{\partial x\partial y}(x_0+\tilde{t}k,y_0+th) \tag{6.3.84}$$

and by continuity when $k,h\rightarrow 0$, we have:

$$\frac{\partial^2 f}{\partial y\partial x}(x_0,y_0)=\frac{\partial^2 f}{\partial x\partial y}(x_0,y_0) \tag{6.3.85}$$

More simply written:

$$\boxed{\frac{\partial^2 f}{\partial y\partial x}=\frac{\partial^2 f}{\partial x\partial y}} \tag{6.3.86}$$

So if f is expressed in a total derivative form therefore the cross differentials are equal (the reciprocal is not necessarily true!).

■ Q.E.D.

By induction on the number of variables we can prove in the general case (that is long but it is possible we will do it if it needs to be done and we have the time...).

So finally getting back to our original problem, we have:

$$\frac{\partial^2 z}{\partial x \partial y}=\frac{\partial^2 z}{\partial y \partial x} \tag{6.3.87}$$

Which finally gives us the "Schwarz condition":

$$\frac{\partial M}{\partial y}=\frac{\partial N}{\partial x} \tag{6.3.88}$$

This is the condition that must be met for a differential form to be an exact total differential and the condition that it should not meet to be a non-total exact differential!!! This is a very important property for the study of Thermodynamics (see corresponding section)!

In order not to confuse the two types of differentials, we use the symbol δ to represent a non-total exact differential:

$$\delta z=M\mathrm{d}x+N\mathrm{d}y \tag{6.3.89}$$

and d for a total exact differential:

$$\mathrm{d}z=M\mathrm{d}x+N\mathrm{d}y \tag{6.3.90}$$

The distinction is extremely important because only the total exact differentials that satisfy:

$$\frac{\partial M}{\partial y}=\frac{\partial N}{\partial x} \tag{6.3.91}$$

have an integral that depends only on the limits of integration (since all the variables change at the same time). Therefore non-total exact differentials depend only on the limits of integration, meaning that:

$$\int\limits_{z_1}^{z_2} \delta z \neq z_2-z_1 \tag{6.3.92}$$

and therefore on a closed path we can have:

$$\oint_\gamma \delta z \neq 0 \tag{6.3.93}$$

While for total exact differentials:

$$\int\limits_{z_1}^{z_2} \mathrm{d}z=\Delta z \tag{6.3.94}$$

and therefore (see detailed proof later when we will deal with curvilinear integrals):

$$\oint_\gamma \mathrm{d}z=0 \tag{6.3.95}$$

In other words, the variation of a function whose differential is total exact, does not depend on the path followed, but only of the initial and final states as it can be expressed as the gradient of a function (see the proof by example in the section of Electrostatics when we check that the electrostatic potential difference is independent of the path). We name such a function that satisfies an exact total differential in physics a "state function" and in mathematics a "holomorphic function" (see section Complex Analysis page 2399), that is, i.e. a function whose value depends only on the present and future state, not its history.

This distinction is very important and especially in thermodynamics where it should be determined whether a physical quantity is a total exact differential (a "state function"!) or not to know how systems evolve.

✍**Example:**

An important example of differential form in thermodynamics, is the elementary work δW of a force exerted on a body in motion in the plane Oxy, we have:

$$\delta W = \vec{F} \circ \mathrm{d}\vec{r} = F_x \mathrm{d}x + F_y \mathrm{d}y \tag{6.3.96}$$

The forces F_x, F_y are not necessarily derived from the same potential $U(x,y)$ such that:

$$\frac{\partial U}{\partial x} = \frac{\partial U}{\partial y} \tag{6.3.97}$$

therefore δW is indeed in this special case a non-total exact differential!

Finally something maybe important to keep in mind?!:

- Suppose $y = f(x,w)$, while in turn $x = g(t)$ and $w = h(t)$. How does y change when t changes?

  $$\frac{\mathrm{d}y}{\mathrm{d}t} = \frac{\partial y}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t} + \frac{\partial y}{\partial \omega}\frac{\mathrm{d}\omega}{\mathrm{d}t} \tag{6.3.98}$$

  Notice that if we assume $y = f(x, w(x))$ (a direct dependency!) but with x having no implicit dependencies, the above relation becomes:

  $$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\partial y}{\partial x}\frac{\partial x}{\partial x} + \frac{\partial y}{\partial \omega}\frac{\partial \omega}{\partial x} \tag{6.3.99}$$

- Suppose $y = f(x,w)$, while in turn $x = g(t,s)$ and $w = h(t,s)$. How does y change when t changes?

  $$\begin{aligned}\frac{\partial y}{\partial t} &= \frac{\partial y}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial y}{\partial \omega}\frac{\partial \omega}{\partial t}\\ \frac{\partial y}{\partial s} &= \frac{\partial y}{\partial x}\frac{\partial x}{\partial x} + \frac{\partial y}{\partial \omega}\frac{\partial \omega}{\partial s}\end{aligned} \tag{6.3.100}$$

Notice that the first bullet point is the "total derivative" when the second bullet point is the "partial total derivative".

Remark

For information there are plenty of different types of derivatives in mathematics:

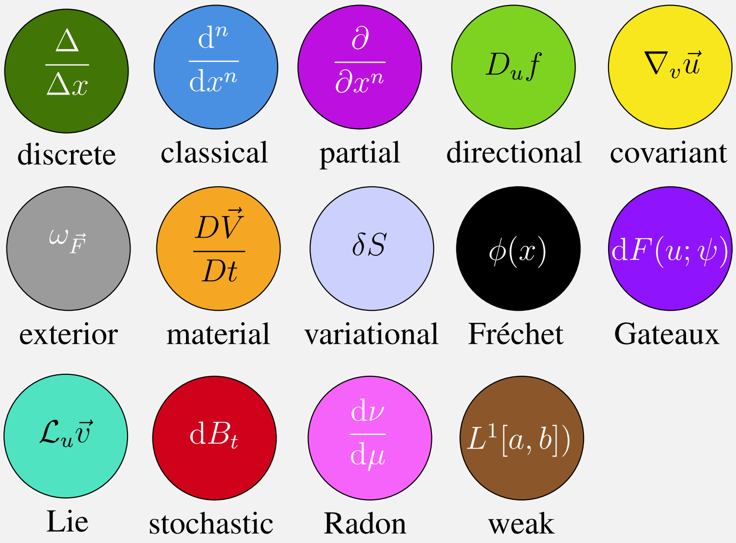

$\frac{\Delta}{\Delta x}$
discrete
$\frac{\mathrm{d}^n}{\mathrm{d}x^n}$
classical
$\frac{\partial}{\partial x^n}$
partial
$D_u f$
directional
$\nabla_v \vec{u}$
covariant
$\omega_{\vec{F}}$
exterior
$\frac{D\vec{V}}{Dt}$
material
δS
variational
$\phi(x)$
Fréchet
$\mathrm{d}F(u;\psi)$
Gateaux
$\mathcal{L}_u \vec{v}$
Lie
$\mathrm{d}B_t$
stochastic
$\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$
Radon
$L^1[a,b])$
weak

Figure 6.37: Most common types of derivatives in mathematics

in this textbook we will use and introduce only a few of them for practical purposes like the discrete, classical, the partial, covariant, stochastic, material and variational derivatives.

6.3.1.1.1 Implicit functions

It frequently happens, particularly in Thermodynamics, that certain variables, let's call them x, y, and z, are not independent, but linked by a relation of the form $g(x,y,z)=0$. In this case, each variable is implicitly defined in terms of the other two, and we can write:

$$x=x(y,z)\quad y=y(x,z)\quad \text{and}\quad z=z(x,y) \tag{6.3.101}$$

Even if we cannot explicitly define these functions, it is possible to relate their partial derivatives. Indeed, let us consider the function $x=x(y,z)$ and write its total differential:

$$\mathrm{d}x=\left.\frac{\partial x}{\partial y}\right|_z \mathrm{d}y+\left.\frac{\partial x}{\partial z}\right|_y \mathrm{d}z \tag{6.3.102}$$

For example, we can deduce from this $\mathrm{d}y$:

$$\mathrm{d}y=\frac{1}{(\partial x/\partial y)_z}\mathrm{d}x-\frac{(\partial x/\partial z)_y}{(\partial x/\partial y)_z}\mathrm{d}z \tag{6.3.103}$$

Let's identify this relationship with the expression for the total differential of $y(x,z)$:

$$\mathrm{d}y=\left.\frac{\partial y}{\partial x}\right|_z \mathrm{d}x+\left.\frac{\partial y}{\partial z}\right|_x \mathrm{d}z \tag{6.3.104}$$

We deduce from this:

$$\left.\frac{\partial y}{\partial x}\right|_z=\frac{1}{\left.\frac{\partial x}{\partial y}\right|_z} \tag{6.3.105}$$

and:

$$\left.\frac{\partial y}{\partial z}\right|_x = -\frac{(\partial x/\partial z)_y}{(\partial x/\partial y)_z} \qquad (6.3.106)$$

therefore:

$$\left.\frac{\partial x}{\partial y}\right|_z \cdot \left.\frac{\partial y}{\partial z}\right|_x \cdot \left.\frac{\partial z}{\partial x}\right|_y = -1 \qquad (6.3.107)$$

named "triangular relation". Note in passing that the fact that this product equals -1 and not $+1$ shows that a partial derivative cannot be treated as a fraction!

✍**Example:**

As we will prove it in the Thermodynamics section, the equation of state of a van der Waals gas is given by:

$$\left(p+\frac{a}{V^2}\right)(V-b) = RT \qquad (6.3.108)$$

where a, b, and R are three positive constants. We wish to calculate the isobaric thermal expansion coefficient of this gas, defined by:

$$\alpha \overset{\text{def}}{=} \left.\frac{1}{V}\frac{\partial V}{\partial T}\right|_p \qquad (6.3.109)$$

The first instinct is to look for the expression $V(T,p)$. However, a quick examination of the equation reveals tedious calculations since V is obtained by solving a cubic equation. It is therefore much more efficient to use the triangular relationship:

$$\left.\frac{\partial V}{\partial T}\right|_p \cdot \left.\frac{\partial T}{\partial p}\right|_V \cdot \left.\frac{\partial p}{\partial V}\right|_T = -1 \qquad (6.3.110)$$

Indeed, the derivatives $\partial T/\left.\partial p\right|_V$ and $\partial p/\left.\partial V\right|_T$ are easily calculated:

$$T = \frac{p(V-b)}{R} + \frac{a(V-b)}{RV^2} \qquad (6.3.111)$$

hence:

$$\left.\frac{\partial T}{\partial p}\right|_V = \frac{V-b}{R} \qquad (6.3.112)$$

and:

$$p = \frac{RT}{V-b} - \frac{a}{V^2} \qquad (6.3.113)$$

so therefore:

$$\left.\frac{\partial p}{\partial V}\right|_T = -\frac{RT}{(V-b)^2} + \frac{2a}{V^3} \qquad (6.3.114)$$

Thus, the triangular relationship allows us to get:

$$\left.\frac{\partial V}{\partial T}\right|_p=\frac{\left.\frac{\partial T}{\partial p}\right|_V}{\left.\frac{\partial p}{\partial V}\right|_T}=\frac{(V-b)V^3}{TV^3-\frac{2a}{R}(V-b)^2} \tag{6.3.115}$$

which gives:

$$\alpha=\frac{(V-b)V^2}{TV^3-\frac{2a}{R}(V-b)^2} \tag{6.3.116}$$

6.3.1.2 Usual Derivatives

We will prove in details here the most common univariate derivatives (around a small thirty) and some of their major properties that we will meet in theoretical physics and mathematics (in fact, we will used them all in respectively the chapters on Mechanics, Engineering, Atomistic, Social Mathematics, etc.). The list below is not exhaustive at this time but the proofs being generalized, they can be applied to many other similar cases (that we will apply / meet throughout almost all this textbook).

1. Derivative of $f(x)=x^n$:

   First start with a particular case, the derivative of x^3:

$$\lim_{h\to 0}\frac{(a+h)^3-a^3}{h}=\lim_{h\to 0}\frac{a^3+3a^2h+3ah^2+h^3-a^3}{h}=\lim_{h\to 0}3a^2+3ah+h^2=3a^2 \tag{6.3.117}$$

   Therefore the derivative of the cubic function is $3a^2$.

   We can generalize this result for any positive or negative integer n and we will see that the function f defined on $\mathbb{R}$ by $f(x)=x^n$ is differentiable and that its derivative f' is given by $f'(x)=nx^{n-1}$.

$$\begin{aligned}\lim_{h\to 0}\frac{(a+h)^n-a^n}{h}&=\lim_{h\to 0}\frac{\sum\limits_{k=0}^{n}C_n^ka^kh^{n-k}-a^n}{h}=\lim_{h\to 0}\frac{\sum\limits_{k=0}^{n-1}C_n^ka^kh^{n-k}+a^n-a^n}{h}\\&=\lim_{h\to 0}\frac{\sum\limits_{k=0}^{n-1}C_n^ka^kh^{n-k}}{h}=\lim_{h\to 0}\sum_{k=0}^{n-1}C_n^ka^kh^{n-k-1}\\&=\lim_{h\to 0}\sum_{k=0}^{n-2}\left(C_n^ka^kh^{n-k-1}+C_n^{n-1}a^{n-1}\right)=C_n^{n-1}a^{n-1}=\frac{n!}{(n-1)!}a^{n-1}\\&=na^{n-1}\end{aligned} \tag{6.3.118}$$

Therefore we have (a few examples can be helpful to understand the scope of this result!):

$$\begin{aligned}
&f(x)=x^2\Rightarrow f'(x)=2x^1=2x\\
&f(x)=x^{-1}\Rightarrow f'(x)=-x^{-2}\\
&f(x)=\sqrt{x}\Rightarrow f'(x)=-\frac{1}{2}x^{-\frac{1}{2}}
\end{aligned}\tag{6.3.119}$$

So we see that having determined the derivative of a function of the form x^n, we also determined the derivative of any function that is written in such a form:

$$f(x)=\frac{1}{x^n}=x^{-n}\quad\text{and}\quad f(x)=\sqrt[n]{x}=x^{1/n}\tag{6.3.120}$$

However, the following functions:

$$f(x)=\frac{1}{x^n}\quad\text{and}\quad f(x)=\sqrt[n]{x}\tag{6.3.121}$$

are not differentiable in the $x=0$ because the function is not more defined on this point (division by zero). Furthermore, relatively to the root function (not integer power), the derivative is not defined in $\mathbb{R}_-^*$.

Remark

It is immediate from the results above that if $f(x)=x^n$ for n being a positive integer then $f^{(k)}(x)=x^{n-k}\,n!/(n-k)!$ for $k\leq n$.

2. Derivative of $f(x)=c^{te}$:

The previous result gives an interesting immediate result for constant functions such as:

$$f(x)=c^{te}=c^{te}x^0\tag{6.3.122}$$

it is then not difficult to determine that the derivative is simply:

$$f'(x)=c^{te}0x^{0-1}=0\tag{6.3.123}$$

So the derivative of any constant function is zero (it is important to remember that result when we will study the properties of integrals) !!!

3. Derivative of $f(x)=\cos(x)$:

Given any real fixed number a, then (be careful it is useful to know the remarkable trigonometric relations that we proved in the section Trigonometry!):

$$\begin{aligned}
f'(x)&=\lim_{h\to 0}\frac{\cos(a+h)-\cos(a)}{h}=\lim_{h\to 0}\frac{-2\sin\left(\dfrac{a+h+a}{2}\right)\sin\left(\dfrac{a+h-a}{2}\right)}{h}\\
&=\lim_{h\to 0}\frac{-2\sin\left(\dfrac{2a+h}{2}\right)\sin\left(\dfrac{h}{2}\right)}{h}=\lim_{h\to 0}\left(-2\sin\left(\frac{2a+h}{2}\right)\right)\frac{1}{2}\lim_{h\to 0}\left(\frac{\sin\left(\dfrac{h}{2}\right)}{\dfrac{h}{2}}\right)\\
&=-2a\sin(a)\frac{1}{2}=-\sin(a)
\end{aligned}\tag{6.3.124}$$

Because using l'Hôpital's rule (or by seeing that $\sin(x)$ can be assimilate by a straight line function $f(x)=x$ near $x=0$):

$$\lim_{h\to 0}\frac{\sin(x)}{x}=1 \tag{6.3.125}$$

So to summarize:

$$f(x)=\cos(x)\Rightarrow f'(x)=-\sin(x) \tag{6.3.126}$$

4. Derivative of $f(x)=\sin(x)$:

   Given any real fixed number a, then (be careful again it is useful to know the remarkable trigonometric relations that we proved in the section Trigonometry!):

$$\begin{aligned} f'(x)&=\lim_{h\to 0}\frac{\sin(a+h)-\sin(a)}{h}=\lim_{h\to 0}\frac{2\cos\left(\dfrac{a+h+a}{2}\right)\sin\left(\dfrac{a+h-a}{2}\right)}{h}\\ &=\lim_{h\to 0}\frac{2\cos\left(\dfrac{2a+h}{2}\right)\sin\left(\dfrac{h}{2}\right)}{h}=\lim_{h\to 0}\left(2\cos\left(\frac{2a+h}{2}\right)\right)\frac{1}{2}\lim_{h\to 0}\left(\frac{\sin\left(\dfrac{h}{2}\right)}{\dfrac{h}{2}}\right)\\ &=2a\cos(a)\frac{1}{2}=\cos(a) \end{aligned} \tag{6.3.127}$$

   Because using l'Hôpital's rule (or by seeing that $\sin(x)$ can be assimilate by a straight line function $f(x)=x$ near $x=0$):

$$\lim_{h\to 0}\frac{\sin(x)}{x}=1 \tag{6.3.128}$$

   So to summarize:

$$f(x)=\sin(x)\Rightarrow f'(x)=\cos(x) \tag{6.3.129}$$

5. Derivative of $f(x)=\log_b(x)$:

   We begin by writing that:

$$y+\Delta y=\log_b(x+\Delta x) \tag{6.3.130}$$

   Therefore:

$$\Delta y=\log_b(x+\Delta x)-y=\log_b(x+\Delta x)-\log_b(x)=\log_b\left(\frac{x+\Delta x}{x}\right)=\log_b\left(1+\frac{\Delta x}{x}\right) \tag{6.3.131}$$

   Therefore:

$$\frac{\Delta y}{\Delta x}=\frac{1}{\Delta x}\log_b\left(1+\frac{\Delta x}{x}\right) \tag{6.3.132}$$

Multiply and divide by x the term in the right member of the last previous equality:

$$\frac{\Delta y}{\Delta x}=\frac{1}{x}\frac{x}{\Delta x}\log_b\left(1+\frac{\Delta x}{x}\right)=\frac{1}{x}\log_b\left(\left(1+\frac{\Delta x}{x}\right)^{\frac{x}{\Delta x}}\right) \tag{6.3.133}$$

Denote the quantity $\dfrac{\Delta x}{x}$ by α. It is obvious that $\alpha\rightarrow 0$ when Δx tends to zero for a given x. Consequently:

$$\frac{\Delta y}{\Delta x}=\frac{1}{x}\log_b\left((1+\alpha)^{\frac{1}{\alpha}}\right) \tag{6.3.134}$$

However, we find here again an another historical origin from Euler's constant (see the section Functional Analysis page 2289 for the proof), that is:

$$\lim_{\alpha\rightarrow 0}(1+\alpha)^{\frac{1}{\alpha}}=e \tag{6.3.135}$$

Therefore:

$$f'(x)=\lim_{\alpha\rightarrow 0}\frac{\Delta y}{\Delta x}=\lim_{\alpha\rightarrow 0}\frac{1}{x}\log_b\left((1+\alpha)^{\frac{1}{\alpha}}\right)=\frac{1}{x}\log_b(e) \tag{6.3.136}$$

An important special case is the case where $b=e$. Then we have the famous result:

$$y=\ln(x)\Rightarrow y'=\frac{1}{x} \tag{6.3.137}$$

6. Derivative of a sum of functions:

   Let us now consider u and v two functions. The sum function $s=u+v$ is derivable over any interval where u and v are differentiable, we will denote by s' the derivative of the sum. Let us now see what is its expression.

   Let a be a real number and u,v two defined and differentiable functions on a:

$$\begin{aligned}s'&=\lim_{h\rightarrow 0}\frac{(u+v)(a+h)-(u+v)(a)}{h}=\lim_{h\rightarrow 0}\frac{u(a+h)+v(a+h)-u(a)-v(a)}{h}\\&=\lim_{h\rightarrow 0}\frac{u(a+h)-u(a)+v(a+h)-v(a)}{h}=\lim_{h\rightarrow 0}\frac{u(a+h)-u(a)}{h}+\lim_{h\rightarrow 0}\frac{v(a+h)-v(a)}{h}\\&=u'(a)+v'(a)\end{aligned} \tag{6.3.138}$$

   So the derivative of a sum is the sum of the derivatives.

   This result can be generalized for a sum of any number of functions.

7. Derivative of a product of functions (named the "product rule"):

   Let us now consider u and v two functions. The product function $p=uv$ is derivable over any interval where u and v are differentiable, we will denote by p' the derivative of the product. Let us now see what is its expression.

   Let a be a real number and u,v two defined and differentiable functions on a:

$$\lim_{h\to 0}\frac{(uv)(a+h)-(uv)(a)}{h}=\lim_{h\to 0}\frac{u(a+h)v(a+h)-u(a)v(a)}{h} \tag{6.3.139}$$

We add to this last relation two terms whose sum is zero such as:

$$\begin{aligned}p' &= \lim_{h\to 0}\frac{u(a+h)v(a+h)-u(a)v(a)}{h}\\ &= \lim_{h\to 0}\frac{u(a+h)v(a+h)-u(a)v(a+h)+u(a)v(a+h)-u(a)v(a)}{h}\\ &= \lim_{h\to 0}\frac{u(a+h)v(a+h)-u(a)v(a+h)}{h}+\lim_{h\to 0}\frac{u(a)v(a+h)-u(a)v(a)}{h}\\ &= \lim_{h\to 0}\frac{u(a+h)-u(a)}{h}v(a+h)+\lim_{h\to 0}\frac{v(a+h)-v(a)}{h}u(a)\\ &= \lim_{h\to 0}\frac{u(a+h)-u(a)}{h}\lim_{h\to 0}v(a+h)+\lim_{h\to 0}\frac{v(a+h)-v(a)}{h}\lim_{h\to 0}u(a)\\ &= u'(a)v(a)+u(a)v'(a)\end{aligned} \tag{6.3.140}$$

Hence the well known result:

$$p=u'v+uv' \tag{6.3.141}$$

But there is a more general formulation than the first derivative of a product:

Theorem 6.11. Consider for this purpose always our two functions u and v, n times differentiable on an interval I. Then the uv product is n times differentiable on I and:

$$\boxed{(uv)^{(n)}=\sum_{k=0}^{n}C_k^n u^{(k)}v^{(n-k)}} \tag{6.3.142}$$

and this constitutes the "Leibniz differentiation rule for products" that we used in the section of Calculus for the study of Legendre polynomials and also Laguerre polynomials (which we are themselves essential to our study of Quantum Chemistry).

The proof of this expression is very similar to that made for the Newton's binomial theorem (see section Calculus page 1364).

Proof 6.11.1. Either:

$$(uv)^{(0)}=uv \tag{6.3.143}$$

On the other hand:

$$(uv)^{(0)}=\sum_{k=0}^{0}C_k^0 u^k v^{(0-k)}=u^{(0)}v^{(0)}=uv \tag{6.3.144}$$

The relation is thus at least well initialized.

The proof is made by induction. Thus, the goal is to show that for $\forall n\geq 0\in\mathbb{N}$ if:

$$(uv)^{(n)}=\sum_{k=0}^{n}C_k^n u^k v^{(n-k)} \tag{6.3.145}$$

then:

$$(uv)^{(n+1)}=\sum_{k=0}^{n+1}C_k^{n+1}u^k v^{(n+1-k)} \tag{6.3.146}$$

We have therefore:

$$\begin{aligned}(uv)^{(n+1)}&=\left((uv)^{(n+1)}\right)'=\sum_{k=0}^{n}C_k^n\left(u^{(k)}v^{(n-k)}\right)'\\&=\sum_{k=0}^{n}C_k^n\left(\left(u^{(k)}\right)'v^{(n-k)}+u^{(k)}\left(v^{(n-k)}\right)'\right)\\&=\sum_{k=0}^{n}C_k^n\left(u^{(k+1)}v^{(n-k)}+u^{(k)}v^{(n-k+1)}\right)\\&=\sum_{k=0}^{n}C_k^n u^{(k+1)}v^{(n-k)}+\sum_{k=0}^{n}C_k^n u^{(k)}v^{(n+1-k)}\end{aligned} \tag{6.3.147}$$

We'll do now a change of variable in the first sum to not have the term in $k+1$ again. We put for this purpose $j=k+1$:

$$\sum_{k=0}^{n}C_k^n u^{(k+1)}v^{(n-k)}=\sum_{j=1}^{n+1}C_{j-1}^n u^{(j)}v^{(n-(j-1))} \tag{6.3.148}$$

If we go back to the letter k, we have:

$$\sum_{k=0}^{n}C_k^n u^{(k+1)}v^{(n-k)}=\sum_{k=1}^{n+1}C_{k-1}^n u^{(k)}v^{(n+1-k)} \tag{6.3.149}$$

So we have:

$$(uv)^{(n+1)}=\sum_{k=1}^{n+1}C_{k-1}^n u^{(k)}v^{(n+1-k)}+\sum_{k=0}^{n}C_k^n u^{(k)}v^{(n+1-k)} \tag{6.3.150}$$

We want to combine this two sums. For this, we discard the terms in excess in each:

$$\begin{aligned}(uv)^{(n+1)}=&C_{n+1-1}^n u^{(n+1)}v^{(n+1-(n+1))}+\sum_{k=1}^{n}C_{k-1}^n u^{(k)}v^{(n+1-k)}\\&+\sum_{k=1}^{n}C_k^n u^{(k)}v^{(n+1-k)}+C_0^n u^{(0)}v^{(n+1-0)}\end{aligned} \tag{6.3.151}$$

This therefore gives:

$$\begin{aligned}(uv)^{(n+1)}&=u^{(n+1)}v^{(0)}+\sum_{k=1}^{n}C_{k-1}^n u^{(k)}v^{(n+1-k)}+\sum_{k=1}^{n}C_k^n u^{(k)}v^{(n+1-k)}+u^{(0)}v^{(n+1)}\\&=u^{(n+1)}v^{(0)}+\sum_{k=1}^{n}(C_{k-1}^n+C_k^n)u^{(k)}v^{(n+1-k)}+u^{(0)}v^{(n+1)}\end{aligned} \tag{6.3.152}$$

According to Pascal's formula (see section Probabilities page 529), we have:

$$C_{k-1}^{n}+C_{k}^{n}=C_{k}^{n+1} \tag{6.3.153}$$

Therefore:

$$(uv)^{(n+1)}=u^{(n+1)}v^{(0)}+\sum_{k=1}^{n}C_{k}^{n+1}u^{(k)}v^{(n+1-k)}+u^{(0)}v^{(n+1)} \tag{6.3.154}$$

But we have at the same time:

$$u^{(n+1)}v^{(0)}=C_{n+1}^{n+1}u^{(n+1)}v^{(n+1-(n+1))} \qquad u^{(0)}v^{(n+1)}=C_{0}^{n+1}u^{(0)}v^{(n+1-0)} \tag{6.3.155}$$

Therefore:

$$\begin{aligned}(uv)^{(n+1)}&=C_{n+1}^{n+1}u^{(n+1)}v^{(n+1-(n+1))}+\sum_{k=1}^{n}C_{k}^{n+1}u^{(k)}v^{(n+1-k)}+C_{0}^{n+1}u^{(0)}v^{(n+1-0)}\\&=\sum_{k=0}^{n+1}C_{k}^{n+1}u^{(k)}v^{(n+1-k)}\end{aligned} \tag{6.3.156}$$

■ Q.E.D.

8. Derivative of a ratio of two differentiable functions (named the "quotient rule"):

   Let $f(x)=g(x)/h(x)$. Applying the definition of the derivative and properties of limits gives the following proof:

$$\begin{aligned}f'(x)&=\lim_{k\to 0}\frac{f(x+k)-f(x)}{k}\\&=\lim_{k\to 0}\frac{\dfrac{g(x+k)}{h(x+k)}-\dfrac{g(x)}{h(x)}}{k}\\&=\lim_{k\to 0}\frac{g(x+k)h(x)-g(x)h(x+k)}{k\cdot h(x)h(x+k)}\\&=\lim_{k\to 0}\frac{g(x+k)h(x)-g(x)h(x+k)}{k}\cdot\lim_{k\to 0}\frac{1}{h(x)h(x+k)}\\&=\left(\lim_{k\to 0}\frac{g(x+k)h(x)-g(x)h(x)+g(x)h(x)-g(x)h(x+k)}{k}\right)\cdot\frac{1}{h(x)^2}\\&=\left(\lim_{k\to 0}\frac{g(x+k)h(x)-g(x)h(x)}{k}-\lim_{k\to 0}\frac{g(x)h(x+k)-g(x)h(x)}{k}\right)\cdot\frac{1}{h(x)^2}\\&=\left(h(x)\lim_{k\to 0}\frac{g(x+k)-g(x)}{k}-g(x)\lim_{k\to 0}\frac{h(x+k)-h(x)}{k}\right)\cdot\frac{1}{h(x)^2}\\&=\frac{g'(x)h(x)-g(x)h'(x)}{h(x)^2}\end{aligned} \tag{6.3.157}$$

   We can also do the proof using implicit differentiation. Indeed, let $f(x)=g(x)/h(x)$, so $g(x)=f(x)h(x)$. The product rule proved just earlier above gives:

$$g'(x)=f'(x)h(x)+f(x)h'(x) \tag{6.3.158}$$

Solving for $f'(x)$ and substituting back for $f(x)$ gives:

$$f'(x)=\frac{g'(x)-f(x)h'(x)}{h(x)}=\frac{g'(x)-\dfrac{g(x)}{h(x)}\cdot h'(x)}{h(x)}=\frac{g'(x)h(x)-g(x)h'(x)}{h(x)^2} \tag{6.3.159}$$

9. Derivative of a composite univariate function:

   Let us consider the composite function $f=g\circ u=g(u(x))$ of two functions g and u differentiable, the first in $u(a)$, the second in a. We have therefore:

$$f'=\lim_{h\to 0}\frac{g\circ u(a+h)-g\circ u(a)}{h}=\lim_{h\to 0}\frac{g(u(a+h))-g(u(a))}{h} \tag{6.3.160}$$

   Let us put now $k=u(a+h)-u(a)$ then we have:

$$\lim_{h\to 0} k=0 \tag{6.3.161}$$

   Let us continue our previous development:

$$\begin{aligned}f'&=\lim_{h\to 0}\frac{g(u(a+h))-g(u(a))}{h}=\lim_{h\to 0}\frac{g(k+u(a)))-g(u(a))}{h}\\&=\lim_{h\to 0}\frac{g(u(a)+k)-g(u(a))}{k}\frac{k}{h}=\lim_{h\to 0}\frac{g(u(a)+k)-g(u(a))}{k}\lim_{h\to 0}\frac{k}{h}\\&=\lim_{h\to 0}\frac{g(u(a)+k)-g(u(a))}{k}\lim_{h\to 0}\frac{u(a+h)-u(a)}{h}=\frac{\mathrm{d}g}{\mathrm{d}u}\frac{\mathrm{d}u}{\mathrm{d}x}\\&=g'(u(a))u'(a)\end{aligned} \tag{6.3.162}$$

   Thus the derivative of a composite function is given by the derivative of the function, multiplied by the "inner derivative". Furthermore, this type of derivation is very important because often used in physics under the name of "(univariate) chain derivation" or simply "(univariate) chain rule".

   Let's see what it is. The previous obtained relation can be rewritten in another more common way:

$$y=f(g(x))\Rightarrow y'=f'g'=\frac{\mathrm{d}f}{\mathrm{d}g}\frac{\mathrm{d}g}{\mathrm{d}x} \tag{6.3.163}$$

   Or typically when we have multiple functions that multiplies each other:

$$y=f(g(x))h(j(x))\Rightarrow y'=\frac{\mathrm{d}f}{\mathrm{d}g}\frac{\mathrm{d}g}{\mathrm{d}x}+\frac{\mathrm{d}h}{\mathrm{d}j}\frac{\mathrm{d}j}{\mathrm{d}x} \tag{6.3.164}$$

   We then speak of "power chain rule".

10. Derivative of a composite bivariate function:

   Theorem 6.12. The t derivative of the composite function $z=f(x(t),y(t))$ is:

$$\frac{\mathrm{d}}{\mathrm{d}t}[f(x(t),y(t))]=\frac{\partial f(x(t),y(t))}{\partial x}\frac{\mathrm{d}x(t)}{\mathrm{d}t}+\frac{\partial f(x(t),y(t))}{\partial y}\frac{\mathrm{d}y(t)}{\mathrm{d}t} \tag{6.3.165}$$

   We assume in this theorem and its applications that $x=x(t)$ and $y=y(t)$ have first derivatives at t and that $z=f(x,y)$ has continuous first order derivatives in an open circle centered at $(x(t),y(t))$.

Proof 6.12.1. We fix t and set $(x,y)=(x(t),y(t))$. We consider non-zero Δt so small that $(x(t+\Delta t),y(t+\Delta t))$ is in the circle where f has continuous first derivatives and set $\Delta x=x(t+\Delta t)-x(t)$ and $\Delta y=y(t+\Delta t)-y(t)$. Then by definition of the derivative:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}[f(x(t),y(t))]&=\lim_{\Delta t\to 0}\frac{f(x(t+\Delta t),y(t+\Delta t))-f(x(t),y(t))}{\Delta t}\\&=\lim_{\Delta t\to 0}\frac{f(x+\Delta x,y+\Delta y)-f(x,y)}{\Delta t}\\&=\lim_{\Delta t\to 0}\frac{[f(x+\Delta x),y)-f(x,y)]+[f(x+\Delta x,y+\Delta y)-f(x+\Delta x,y)]}{\Delta t}\end{aligned}\tag{6.3.166}$$

We can apply the "mean value theorem" that states (see the proof further below during our study of integral calculus):

$$f(b)-f(a)=f'(c)(b-a)\tag{6.3.167}$$

to the expression in the first set of square brackets on the right of the last equality above where y is constant and to the expression in the second set of square brackets where x is constant. We conclude that there is a number c_1 between x and $x+\Delta x$ and a number c_2 between y and $y+\Delta y$ such that:

$$\begin{aligned}f(x+\Delta x,y)-f(x,y)&=\frac{\partial f(x,y)}{\partial x}(c_1,y)\Delta x\\f(x+\Delta x,y+\Delta y)-f(x+\Delta x,y)&=\frac{\partial f(x,y)}{\partial x}(x+\Delta x,c_2)\Delta y\end{aligned}\tag{6.3.168}$$

We add the both relations above and divide by Δt to get:

$$\frac{f(x+\Delta x,y+\Delta y)-f(x,y)}{\Delta t}=\frac{\partial f(x,y)}{\partial x}(c_1,y)\frac{\Delta x}{\Delta t}+\frac{\partial f(x,y)}{\partial x}(x+\Delta x,c_2)\frac{\Delta y}{\Delta t}\tag{6.3.169}$$

The functions $x=x(t)$ and $y=y(t)$ are continuous at t because they have derivatives at that point. Consequently, as $\Delta t\to 0$, the numbers Δx and Δy both tend to zero and the circle including the constants c_i collapses to the point (x,y), Because the partial derivatives of f are continuous, the term $\partial_x f(c_1,y+\Delta y)$ tends to $\partial_x f(x,y)$ and the term $\partial_y f(x,c_2)$ tends to $\partial_y f(x,y)$ as $\Delta t\to 0$. Moreover:

$$\frac{\Delta x}{\Delta t}\to x'(t)\qquad\frac{\Delta y}{\Delta t}\to y'(t)\tag{6.3.170}$$

as $\Delta t\to 0$, so the above relation becomes:

$$\boxed{\frac{\mathrm{d}}{\mathrm{d}t}[f(x(t),y(t))]=\frac{\partial f(x(t),y(t))}{\partial x}\frac{\mathrm{d}x(t)}{\mathrm{d}t}+\frac{\partial f(x(t),y(t))}{\partial y}\frac{\mathrm{d}y(t)}{\mathrm{d}t}}\tag{6.3.171}$$

named the "multivariate chain rule" (but in reality it is only the bivariate case...) and that is veeeeery important for study physics.

The latter relation is sometimes written:

$$\frac{\mathrm{d}}{\mathrm{d}t}[f(x(t),y(t))]=\partial_x f(x(t),y(t))\frac{\mathrm{d}x(t)}{\mathrm{d}t}+\partial_y f(x(t),y(t))\frac{\mathrm{d}y(t)}{\mathrm{d}t}\tag{6.3.172}$$

■ Q.E.D.

11. Derivative of an inverse function:

    Theorem 6.13. If the function f is continuous, strictly monotonic over an interval I, derivable over I, then the reciprocal function f^{-1} is derivable on the interval $f(I)$ and admits for derivative function:

    $$\forall y \in f(I) \qquad (f^{-1})'(y) = \frac{1}{f(f^{-1}(y))'} \tag{6.3.173}$$

    Proof 6.13.1. Indeed we can write:

    $$x \overset{f}{\mapsto} y = f(x) \overset{f^{-1}}{\longmapsto} x = f^{-1}(y) \tag{6.3.174}$$

    That is to say (identity application):

    $$x = f^{-1}(y) = f^{-1}(f(x)) = f^{-1} \circ f(x) \tag{6.3.175}$$

    By application of the derivation of composite functions seen just above we have:

    $$(f^{-1}(f(x)))' = (f^{-1}(y))' f'(x) = (x)' = 1 \tag{6.3.176}$$

    Therefore:

    $$(f^{-1}(y))' = \frac{1}{f'(x)} = \frac{1}{f'(f^{-1}(y))} \tag{6.3.177}$$

    For a variable x, it is more common to write for the derivative of the inverse function as following:

    $$(f^{-1}(x))' = \frac{1}{f'(y)} \tag{6.3.178}$$

■ Q.E.D.

12. Derivative of $|x|$:

    By definition:

    $$\frac{\mathrm{d}}{\mathrm{d}x}|x| = \frac{\mathrm{d}}{\mathrm{d}x}\sqrt{x^2} = \frac{\mathrm{d}}{\mathrm{d}x}\left(x^2\right)^{\frac{1}{2}} \tag{6.3.179}$$

    By applying the chain rule for derivatives, we get:

    $$\frac{\mathrm{d}}{\mathrm{d}x}|x| = \frac{1}{2}\left(x^2\right)^{-\frac{1}{2}} \cdot 2x = \frac{x}{\sqrt{x^2}} \tag{6.3.180}$$

    Therefore:

    $$\frac{\mathrm{d}}{\mathrm{d}x}|x| = \frac{x}{|x|} \tag{6.3.181}$$

    Or in another way (from the definition of derivative):

    $$\begin{aligned}\left.\frac{\mathrm{d}|x|}{\mathrm{d}x}\right|_{x=0} &= \lim_{x\to 0}\frac{|x|-0}{x-0}\\ &= \begin{cases}\lim_{x\to 0^+}\frac{x}{x} & : x > 0\\ \lim_{x\to 0^-}\frac{-x}{x} & : x < 0\end{cases} \qquad \text{Definition of Absolute Value}\\ &= \begin{cases}1 & : x > 0\\ -1 & : x < 0\end{cases}\end{aligned}$$

13. Derivative of $\arccos(x)$:

    Using the previous result of the reciprocal function and the derivative of $\cos(x)$ proved above, we can calculate the derivative of the function $\arccos(x)$:

$$(f^{-1}(x))'=(\arccos(x))=\frac{1}{f'(y)}=\frac{1}{-\sin(y)}=\frac{-1}{\sin(y)}=\frac{1}{\sqrt{1-\cos^2(y)}}=\frac{-1}{\sqrt{1-x^2}} \tag{6.3.182}$$

14. Derivative of $\arcsin(x)$:

    Using the previous result of the reciprocal function and the derivative of $\sin(x)$ proved above, we can calculate the derivative of the function $\arcsin(x)$:

$$(f^{-1}(x))'=(\arcsin(x))=\frac{1}{f'(y)}=\frac{1}{\cos(y)}=\frac{1}{\sqrt{1-\sin^2(y)}}=\frac{1}{\sqrt{1-x^2}} \tag{6.3.183}$$

15. Derivative from a quotient of two functions:

    Consider that function:

$$f=\frac{u}{v} \tag{6.3.184}$$

is derivable over any interval where u and v are differentiable functions and wherein the function v is not null.

The function f can be considered as the product of two functions: the function u and the function $1/v$. A product of two functions is differentiable if each is differentiable, it is necessary that the function u is differentiable and the function $1/v$ is also derivable which is the case when v is differentiable and not null.

$$\left(\frac{u}{v}\right)'=\left(u\frac{1}{v}\right)'=u'\frac{1}{v}+u\left(\frac{-v'}{v^2}\right)=\frac{u'v-uv'}{v^2} \tag{6.3.185}$$

16. Derivative of the function $\tan(x)$:

    By definition (see section Trigonometry page 2508), $\forall x\neq k\dfrac{\pi}{2}, k\in\mathbb{Z}$:

$$f(x)=\tan(x)=\frac{\sin(x)}{\cos(x)} \tag{6.3.186}$$

and then applying the derivative rule for a quotient as proved above, we have:

$$f'(x)=\frac{\cos^2(x)+\sin^2(x)}{\cos^2(x)}=\frac{1}{\cos^2(x)} \tag{6.3.187}$$

or:

$$f'(x)=\frac{\cos^2(x)+\sin^2(x)}{\cos^2(x)}=1+\frac{\sin^2(x)}{\cos^2(x)}=1+\tan^2(x) \tag{6.3.188}$$

17. Derivative of the function $\cot(x)$:

    By definition (see section Trigonometry page 2508),$\forall x \neq k\pi, k \in \mathbb{Z}$:

$$f(x)=\cot(x)=\frac{1}{\tan(x)} \tag{6.3.189}$$

    and therefore (applying once again the rule of the derivative of a quotient as proved previously):

$$f'(x)=\cot'(x)=-\frac{1+\tan^2(x)}{\tan^2(x)}=-\frac{1}{\tan^2(x)}-1=-(\cot^2(x)+1) \tag{6.3.190}$$

    or:

$$\cot'(x)=-\frac{1}{\tan^2(x)}-1=-\frac{\cos^2(x)}{\sin^2(x)}-1=-\frac{\cos^2(x)+\sin^2(x)}{\sin^2(x)}=-\frac{1}{\sin^2(x)} \tag{6.3.191}$$

18. Derivative of the function $\arctan(x)$:

    We use the properties of the derivative of reciprocal functions proved previously:

$$\forall x \in \mathbb{R} \quad f'(x)=\arctan'(x)=\frac{1}{1+\tan^2(\arctan(x))}=\frac{1}{1+x^2} \tag{6.3.192}$$

19. Derivative of the function $\text{arccot}(x)$:

    We also use the properties of the derivative of reciprocal functions proved previously:

$$\forall x \in \mathbb{R} \quad f'(x)=\text{arccot}'(x)=\frac{1}{-(1+\cot^2(\text{arccot}(x)))}=-\frac{1}{1+x^2} \tag{6.3.193}$$

20. Derivative of e^x:

    We will prove in our study of Theoretical Computing (see section of the same name 6301) that the "Euler number" can be calculated from the series:

$$e^x=\sum_{k=0}^{+\infty}\frac{x^k}{k!} \tag{6.3.194}$$

    which converges on $\mathbb{R}$. By derivating term by term this series that converges, we get:

$$\forall x \in \mathbb{R} \quad f'(x)=\sum_{k=0}^{+\infty}k\frac{x^{k-1}}{k!}=\sum_{k=1}^{+\infty}\frac{x^{k-1}}{(k-1)!}=e^x \tag{6.3.195}$$

    Thus the exponential is its own derivative. So now we can afford to study the derivatives of some hyperbolic trigonometric functions (see section Trigonometry page 2532) and many other specific cases (see all other chapters of the book).

21. Derivative of $\sinh(x)$:

    Remember (see section Trigonometry page 2508):

$$\sinh(x)=\frac{1}{2}\left(e^x-e^{-x}\right) \tag{6.3.196}$$

    So trivially:

$$\sinh'(x)=\frac{1}{2}\left(e^x+e^{-x}\right) \tag{6.3.197}$$

22. Derivative of $\cosh(x)$:

    Remember (see section Trigonometry page 2508):

    $$\cosh(x)=\frac{1}{2}\left(e^x+e^{-x}\right) \quad (6.3.198)$$

    So trivially:

    $$\cosh'(x)=\frac{1}{2}\left(e^x-e^{-x}\right)=\sinh(x) \quad (6.3.199)$$

23. Derivative of $\tanh(x)$:

    Remember (see section Trigonometry page 2508):

    $$\tanh(x)=\frac{\sinh(x)}{\cosh(x)} \quad (6.3.200)$$

    Therefore by applying the derivative from a quotient we get:

    $$\tanh'(x)=\frac{\cosh^2(x)-\sinh^2(x)}{\cosh^2(x)}=\frac{1}{\cosh^2(x)} \quad (6.3.201)$$

    or:

    $$\tanh'(x)=\frac{\cosh^2(x)-\sinh^2(x)}{\cosh^2(x)}=1-\frac{\sinh^2(x)}{\cosh^2(x)}=1-\tanh^2(x) \quad (6.3.202)$$

24. Derivative of $\coth(x)$:

    Remember (see section Trigonometry page 2508):

    $$\forall x\neq 0 \qquad \coth(x)=\frac{\cosh(x)}{\sinh(x)}=\frac{1}{\tanh(x)} \quad (6.3.203)$$

    Therefore by applying the derivative from a quotient we get:

    $$\coth'(x)=-\frac{1-\tanh^2(x)}{\tanh^2(x)}=1-\frac{1}{\tanh^2(x)}=1-\coth^2(x) \quad (6.3.204)$$

25. Derivative of arcsinh(x):

    We also use the properties of the derivative of reciprocal functions proved previously:

    $$\text{arcsinh}'(x)=\frac{1}{\cosh(\text{arcsinh}(x))} \quad (6.3.205)$$

    But (see section Trigonometry page 2508):

    $$1=\cosh^2(\text{arcsinh}(x))-\sinh^2(\text{arcsinh}(x))=\cosh^2(\text{arcsinh}(x))-x^2 \quad (6.3.206)$$

    and therefore:

    $$1+x^2=\cosh^2(\text{arcsinh}(x)) \quad (6.3.207)$$

    Since $\cosh(x)$ takes only positive values, we have:

    $$\cosh(\text{arcsinh}(x))=\sqrt{1+x^2} \quad (6.3.208)$$

    Then finally:

    $$\text{arcsinh}'(x)=\frac{1}{\cosh(\text{arcsinh}(x))}=\frac{1}{\sqrt{1+x^2}} \quad (6.3.209)$$

26. Derivative of $\text{arccosh}(x)$:

    We also use the properties of the derivative of reciprocal functions proved previously:

    $$\forall x > 1 \quad \text{arccosh}'(x) = \frac{1}{\sinh(\text{arccosh}(x))} \tag{6.3.210}$$

    But (see section Trigonometry page 2508):

    $$1 = \cosh^2(\text{arccosh}(x)) - \sinh^2(\text{arccosh}(x)) = x^2 - \sinh^2(\text{arccosh}(x)) \tag{6.3.211}$$

    and therefore:

    $$-1 + x^2 = \sinh^2(\text{arccosh}(x)) \tag{6.3.212}$$

    Since $\text{arccosh}(x)$ takes only positive values so do $\sinh(x)$, then we have:

    $$\sinh(\text{arccosh}(x)) = \sqrt{x^2 - 1} \tag{6.3.213}$$

    Then finally:

    $$\forall x > 1 \quad \text{arccosh}'(x) = \frac{1}{\sinh(\text{arccosh}(x))} = \frac{1}{\sqrt{x^2 - 1}} \tag{6.3.214}$$

27. Derivative of $\text{arctanh}(x)$:

    We also use the properties of the derivative of reciprocal functions proved previously:

    $$\forall x \in [-1, 1] \quad \text{arctanh}'(x) = \frac{1}{1 - \tanh^2(\text{arctanh}(x))} = \frac{1}{1 - x^2} \tag{6.3.215}$$

28. Derivative of $\text{arccoth}(x)$:

    We also use the properties of the derivative of reciprocal functions proved previously:

    $$\forall |x| > 1 \quad \text{arccoth}'(x) = \frac{1}{1 - \coth^2(\text{arccoth}(x))} = \frac{1}{1 - x^2} \tag{6.3.216}$$

29. Derivative of a^x with $a > 0$:

    $$f(x) = a^x = e^{\ln(a^x)} = e^{x \ln(a)} \tag{6.3.217}$$

    So (derivative of a composite function):

    $$f'(x) = \ln(a) e^{x \ln(a)} = a^x \ln(a) \tag{6.3.218}$$

30. We are aware of the chain rules for the derivative of a product of function. But what about the power that is quite frequent in physics?

    Indeed, the derivative of $f(x)^{g(x)}$ is considered by many as one of the most versatile derivative!

    First, to make the proof, it should be obvious that:

    $$\frac{\mathrm{d}}{\mathrm{d}x} x^n = n x^{n-1} \tag{6.3.219}$$

and:

$$\frac{\mathrm{d}}{\mathrm{d}x}a^x=\ln(a)a^x \tag{6.3.220}$$

are just special cases of the derivative of $f(x)^{g(x)}$.

To solve:

$$\frac{\mathrm{d}}{\mathrm{d}x}f(g)^{g(x)} \tag{6.3.221}$$

we need to play a little trick to make the differentiation easier. And nothing is as easy as:

$$\frac{\mathrm{d}}{\mathrm{d}x}e^x=e^x \tag{6.3.222}$$

So since $x=e^{\ln(x)}$ we will use that to our advantage.

We start with:

$$\frac{\mathrm{d}}{\mathrm{d}x}f(g)^{g(x)}=\frac{\mathrm{d}}{\mathrm{d}x}\left(e^{\ln\left(f(x)^{g(x)}\right)}\right) \tag{6.3.223}$$

We know that:

$$\ln\left(f(x)^{g(x)}\right)=g(x)\ln(f(x)) \tag{6.3.224}$$

so we plug that in to get this:

$$\frac{\mathrm{d}}{\mathrm{d}x}f(g)^{g(x)}=\frac{\mathrm{d}}{\mathrm{d}x}\left(e^{g(x)\ln(f(x))}\right) \tag{6.3.225}$$

Now all we have to do is differentiate the right-hand side!

Remember that:

$$\frac{\mathrm{d}}{\mathrm{d}x}e^{f(x)}=e^{f(x)}\frac{\mathrm{d}}{\mathrm{d}x}f(x) \tag{6.3.226}$$

Then:

$$\frac{\mathrm{d}}{\mathrm{d}x}e^{g(x)\ln(f(x))}=e^{g(x)\ln(f(x))}\frac{\mathrm{d}}{\mathrm{d}x}\left(g(x)\ln(f(x))\right) \tag{6.3.227}$$

That is equivalent as to write:

$$\frac{\mathrm{d}}{\mathrm{d}x}f(x)^{g(x)}=f(x)^{g(x)}\frac{\mathrm{d}}{\mathrm{d}x}\left(g(x)\ln(f(x))\right) \tag{6.3.228}$$

Now all we have to do is differentiate the right-hand side. This is the simple multiplication rule:

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(g(x)\ln(f(x))\right)=\frac{\mathrm{d}g(x)}{\mathrm{d}x}\ln(f(x))+g(x)\frac{\mathrm{d}\ln(x)}{\mathrm{d}x} \tag{6.3.229}$$

And then we need to differentiate the natural log:

$$\frac{\mathrm{d}\ln(x)}{\mathrm{d}x}=\frac{1}{f(x)}\frac{\mathrm{d}f(x)}{\mathrm{d}x} \tag{6.3.230}$$

So we plug that in to get this:

$$\frac{\mathrm{d}}{\mathrm{d}x}f(x)^{g(x)}=f(x)^{g(x)}\left[\frac{\mathrm{d}g(x)}{\mathrm{d}x}\ln(f(x))+\frac{g(x)}{f(x)}\frac{\mathrm{d}f(x)}{\mathrm{d}x}\right] \tag{6.3.231}$$

And then we distribute $f(x)^{g(x)}$ to get:

$$\frac{\mathrm{d}}{\mathrm{d}x}f(x)^{g(x)}=f(x)^{g(x)}\frac{\mathrm{d}g(x)}{\mathrm{d}x}\ln(f(x))+f(x)^{g(x)}\frac{g(x)}{f(x)}\frac{\mathrm{d}f(x)}{\mathrm{d}x} \tag{6.3.232}$$

Then if we simplify:

$$\boxed{\frac{\mathrm{d}}{\mathrm{d}x}f(x)^{g(x)}=f(x)^{g(x)}\frac{\mathrm{d}g(x)}{\mathrm{d}x}\ln(f(x))+f(x)^{g(x)-1}g(x)\frac{\mathrm{d}f(x)}{\mathrm{d}x}} \tag{6.3.233}$$

So some application examples give us:

$$\frac{\mathrm{d}}{\mathrm{d}x}x^n=nx^{n-1}\frac{\mathrm{d}}{\mathrm{d}x}x \tag{6.3.234}$$

Indeed, all we have to do is set $f(x)=x$ and $g(x)=n$ and we get:

$$\frac{\mathrm{d}}{\mathrm{d}x}x^n=x^n\frac{\mathrm{d}n}{\mathrm{d}x}\ln(x)+x^{n-1}n\frac{\mathrm{d}n}{\mathrm{d}x}=x^n\cdot 0\cdot\ln(x)+x^{n-1}n\cdot 1=x^{n-1}n \tag{6.3.235}$$

Or another application:

$$\frac{\mathrm{d}}{\mathrm{d}x}a^x \tag{6.3.236}$$

all we have to do is set $f(x)=a$ and $g(x)=x$ and we get:

$$\frac{\mathrm{d}}{\mathrm{d}x}a^x=a^x\frac{\mathrm{d}x}{\mathrm{d}x}\ln(a)+a^{x-1}x\frac{\mathrm{d}x}{\mathrm{d}x}=a^2\cdot 1\cdot\ln(a)+a^{x-1}x\cdot 0=a^x\cdot\ln(a) \tag{6.3.237}$$

31. For the fun (we don't use this derivative anywhere in this textbook on any practical case), let us calculate the derivative of x^x.

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\left[x^x\right]&=\frac{\mathrm{d}}{\mathrm{d}x}\left[\mathrm{e}^{\ln(x)x}\right]=\mathrm{e}^{x\ln(x)}\cdot\frac{\mathrm{d}}{\mathrm{d}x}[x\ln(x)]\\&=\left(\frac{\mathrm{d}}{\mathrm{d}x}[x]\cdot\ln(x)+x\cdot\frac{\mathrm{d}}{\mathrm{d}x}[\ln(x)]\right)\mathrm{e}^{x\ln(x)}\\&=\mathrm{e}^{x\ln(x)}\left(1\ln(x)+\frac{1}{x}x\right)=\mathrm{e}^{x\ln(x)}(\ln(x)+1)=x^x(\ln(x)+1)\end{aligned} \tag{6.3.238}$$

To summarize for the most important one (some will be proved later in this textbook):

$f(x)$	$f'(x)$	Condition
c^{te}	0	None
x^n	nx^{n-1}	None
$\cos(x)$	$-\sin(x)$	$x \in \mathbb{R}$
$\sin(x)$	$\cos(x)$	$x \in \mathbb{R}$
$\log_b(x)$	$\dfrac{1}{x}\log_b(e)$	$x > 0, b > 0, b \neq 1$
$u(x)+v(x)$	$u'(x)+v'(x)$	None
$u(x)v(x)$	$u'(x)v(x)+u(x)v'(x)$	None
$u(x)v(x)$	$u'(x)v(x)+u(x)v'(x)$	None
$u(x)/v(x)$	$\dfrac{u'(x)v(x)-u(x)v'(x)}{v(x)^2}$	$v(x) \neq 0, v'(x) \neq 0$
$\lvert x \rvert$	$x/\lvert x \rvert$	
$v \circ u = v(u(x))$	$\dfrac{\mathrm{d}v}{\mathrm{d}u}\dfrac{\mathrm{d}u}{\mathrm{d}x}$	None
$\frac{\mathrm{d}}{\mathrm{d}t}[f(x(t),y(t))]$	$\frac{\partial f(x(t),y(t))}{\partial x}\frac{\mathrm{d}x(t)}{\mathrm{d}t}+\frac{\partial f(x(t),y(t))}{\partial y}\frac{\mathrm{d}y(t)}{\mathrm{d}t}$	None
$f^{-1}(y)$	$\frac{1}{f(f^{-1}(y))'}$	None
$\arccos(x)$	$\frac{-1}{\sqrt{1-x^2}}$	$1 \geq x \geq -1$
$\arcsin(x)$	$\frac{1}{\sqrt{1-x^2}}$	$1 \geq x \geq -1$
$\tan(x)$	$1+\tan^2(x)$	$\forall x \neq k\frac{\pi}{2}, k \in \mathbb{Z}$
$\cot(x)$	$-\left(\cot^2(x)+1\right) = -\frac{1}{\sin^2(x)}$	$\forall x \neq k\pi, k \in \mathbb{Z}$
$\arctan(x)$	$\frac{1}{1+x^2}$	$\forall x \in \mathbb{R}$
$\mathrm{arccot}(x)$	$-\frac{1}{1+x^2}$	$\forall x \in \mathbb{R}$
e^x	e^x	None
$\sinh(x)$	$\frac{1}{2}\left(e^x+e^{-x}\right)$	None
$\cosh(x)$	$\frac{1}{2}\left(e^x-e^{-x}\right) = \sinh(x)$	None
$\coth(x)$	$1-\frac{1}{\tanh^2(x)} = 1-\coth^2(x)$	$\forall x \neq 0$
$\mathrm{arcsinh}(x)$	$\frac{1}{\cosh(\mathrm{arcsinh}(x))} = \frac{1}{\sqrt{1+x^2}}$	None
$\mathrm{arccosh}(x)$	$\frac{1}{\sinh(\mathrm{arccosh}(x))} = \frac{1}{\sqrt{x^2-1}}$	$\forall x > 1$
$\mathrm{arctanh}(x)$	$\frac{1}{1-\tanh^2(\mathrm{arctanh}(x))} = \frac{1}{1-x^2}$	$\forall x \in [-1,1]$
$\mathrm{arccoth}(x)$	$\frac{1}{1-\coth^2(\mathrm{arccoth}(x))} = \frac{1}{1-x^2}$	$\forall \lvert x \rvert > 1$
a^x	$a^x \ln(a)$	$a > 0$
$f(x)^{g(x)}$	$f(x)^{g(z)}\frac{\mathrm{d}g(x)}{\mathrm{d}x}\ln(f(x))+f(x)^{g(z)-1}g(x)\frac{\mathrm{d}f(x)}{\mathrm{d}x}$	None
x^x	$x^x(\ln(x)+1)$	$x \neq 0$
$H(x)$ (Heaviside)	$\delta(x-a)$	None

Table 6.10: **Most common derivatives**

6.3.1.3 Implicit Differentiation

Thus far, the functions we have been concerned with functions that have been defined explicitly. A function is defined explicitly if *the output is given directly in terms of the input*. For instance, in the function:

$$f(x)=4x^2 \tag{6.3.239}$$

the value of $f(x)$ is given explicitly or directly in terms of the input. Just by knowing the input we can immediately find the output. A second type of function that is also useful for us to consider is an "implicitly defined function". A function is defined implicitly if *the output cannot be found directly from the input*. For instance (stupid simple example):

$$\sqrt{f}=2x \tag{6.3.240}$$

is an implicitly defined function, because for each positive x value there is a corresponding f value, but we cannot find it directly from the function. We would need to square both sides, and then we would have the explicitly defined function:

$$f=4x^2 \tag{6.3.241}$$

It is also possible for us to have implicitly defined functions that we cannot rewrite as an explicitly defined function!!!

For instance we might have the function:

$$\sin(f(x))+f(x)=x \tag{6.3.242}$$

For a given x value, there *may* be a corresponding output value $f(x)$ which makes this a true statement. In this way a function $f(x)$ would be defined for all such x where there is a solution. For instance, we have $f(0)=0$, because setting $x=f(x)=0$ in the above equation is a true statement. Right now we don't have the proper tools to solve such an equation, but the important concept here is that we can have a function defined in such a way.

When we speak of functions, we mean that we have a rule which provides us with at most one output for a given input (there is no output for inputs at which the function is not defined). In a more general sense we might want to look at rules that provide us with multiple outputs for a given input. Such an example would be the equation:

$$x^2+y^2=1 \tag{6.3.243}$$

which is the equation for the unit circle (see section Analytical Geometry page 2768). It turns out that this object consists of two functions, namely:

$$y=\pm\sqrt{1-x^2} \tag{6.3.244}$$

The equation of this circle does not define a function (because the output is multi-valued!!), but it does define some type of curve in the x-y plane. In general, we should be able to describe an arbitrary curve as a combination of some number of functions. It is sensible (and useful) to consider the slope of certain points on the curve, which would simply correspond to the slope of the specific function that defines that part of the curve.

To solve the problem of finding the derivative of a function defined in a way such as $\sin(f(x))+f(x)=x$ or by a curve like a circle, we employ the chain rule once again. The way in which we will employ it is named "implicit differentiation". The process works as follows: we differentiate both sides of the equation with respect to x (or the independent variable), and then we solve for the derivative of the dependent variable. Anywhere we find the function f (or the dependent variable), we will use the chain rule to find the derivative. Let us begin with a few simple examples:

✍**Examples:**

E1. We want to find $\mathrm{d}y/\mathrm{d}x$ if $y^2=x$.

One method of solving this problem would be rewriting it in terms of an explicit function of y, and differentiating. Since we have $y=\pm\sqrt{x}$, we actually have two functions, and we would find:

$$\frac{\mathrm{d}y_1}{\mathrm{d}x}=\frac{1}{2\sqrt{x}} \quad \text{and} \quad \frac{\mathrm{d}y_2}{\mathrm{d}x}=-\frac{1}{2\sqrt{x}} \tag{6.3.245}$$

This works sufficiently well in this situation, but what about a function we cannot rewrite explicitly? We would need implicit differentiation. Let's apply implicit differentiation to this situation, as a means of exercise:

$$\begin{aligned} y^2&=x\\ \frac{\mathrm{d}}{\mathrm{d}x}y^2&=\frac{\mathrm{d}}{\mathrm{d}x}x\\ 2y\frac{\mathrm{d}y}{\mathrm{d}x}&=1\\ \frac{\mathrm{d}y}{\mathrm{d}x}&=\frac{1}{2y} \end{aligned} \tag{6.3.246}$$

Were we unable to rewrite y explicitly in terms of x, this is as far as we could go, but we know that $y=\pm\sqrt{x}$, and plugging in this result we find that:

$$\frac{\mathrm{d}y_1}{\mathrm{d}x}=\frac{1}{2\sqrt{x}} \quad \text{and} \quad \frac{\mathrm{d}y_2}{\mathrm{d}x}=-\frac{1}{2\sqrt{x}}$$

once again. In this way, we were able to calculate both derivatives at once.

E2. We want to find $\mathrm{d}f/\mathrm{d}x$ for $\sin(f)+f=x$.

Here we have no choice but to apply implicit differentiation:

$$\begin{aligned} \sin(f)+f&=x\\ \frac{\mathrm{d}}{\mathrm{d}x}(\sin(f)+f)&=\frac{\mathrm{d}x}{\mathrm{d}x}\\ \cos(f)\frac{\mathrm{d}f}{\mathrm{d}x}+\frac{\mathrm{d}f}{\mathrm{d}x}&=1\\ \frac{\mathrm{d}f}{\mathrm{d}x}(\cos(f)+1)&=1\\ \frac{\mathrm{d}f}{\mathrm{d}x}&=\frac{1}{\cos(f)+1} \end{aligned} \tag{6.3.247}$$

A famous case of application in pure mathematics of implicit differentiation (we will see later for physics) is the one at the origin of this technique: the "folium of Descartes[4]" defined by:

$$x^3+y^3-9xy=0 \tag{6.3.248}$$

The curve was first proposed by René Descartes in... 11638 (holocene calendar).

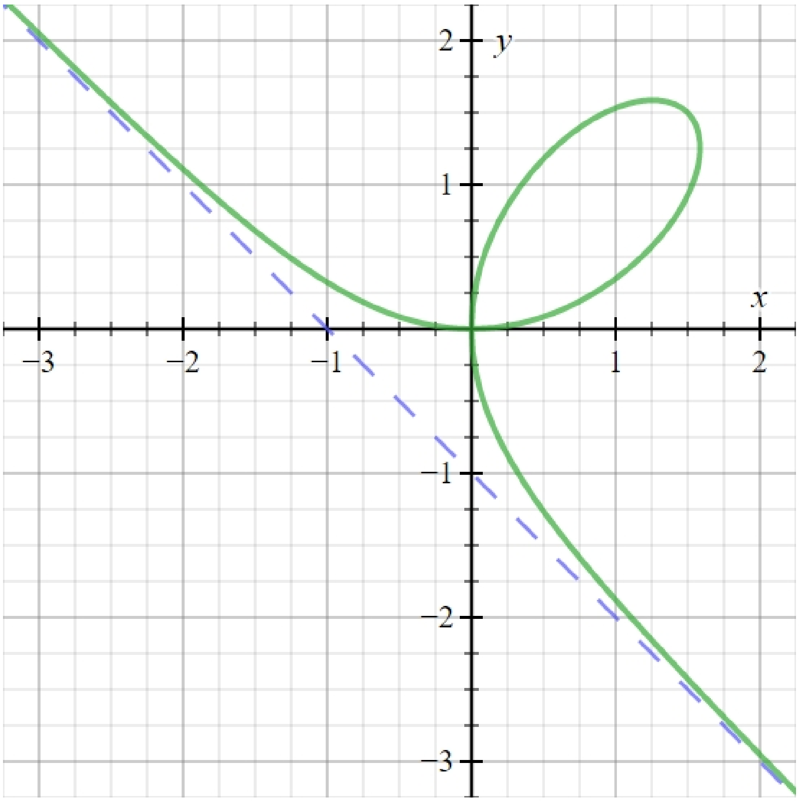


Figure 6.38: The folium of Descartes (green) with asymptote (blue) (source: Wikipedia)

Its claim to fame lies in an incident in the development of calculus. Descartes challenged Pierre de Fermat to find the tangent line to the curve at an arbitrary point since Fermat had recently discovered a method for finding tangent lines. Fermat solved the problem easily, something Descartes was unable to do. Since the invention of calculus, the slope of the tangent line can be found easily using implicit differentiation in any point as we will show it!

Consider now that we want to find the slope of the curve at the point $(2,4)$. Then we use implicit differentiation:

$$\begin{gathered}x^3+y^3-9xy=0\\3x^2+3y^2\frac{\mathrm{d}y}{\mathrm{d}x}-9y-9x\frac{\mathrm{d}y}{\mathrm{d}x}=0\\(3y^2-9y)\frac{\mathrm{d}y}{\mathrm{d}x}=9y-3x^2\\\frac{\mathrm{d}y}{\mathrm{d}x}=\frac{9y-3x^2}{3y^2-9y}=\frac{3y-x^2}{y^2-3y}\end{gathered} \tag{6.3.249}$$

---

[4] The name comes from the Latin word folium which means "leaf".

Evaluating the derivative at the point $(2,4)$ we find:

$$\left.\frac{\mathrm{d}y}{\mathrm{d}x}\right|_{(2,4)}=\left.\frac{3y-x^2}{y^2-3y}\right|_{(2,4)}=\frac{12-4}{16-6}=\frac{4}{5} \tag{6.3.250}$$

Now that we have the slope of the tangent line at the point of interest, we use the point-slope form:

$$\begin{gathered}y_t-y_0=m(x-x_0)\\ y_t-4=\frac{4}{5}(x-4)y_t=\frac{4}{5}(x-2)+4=\frac{4}{5}x+\frac{12}{5}\end{gathered} \tag{6.3.251}$$

Now as far as the line normal to the curve at this point is concerned, we need to find the line perpendicular to the tangent line. This line will cross through the same point, but the slope will be the negative reciprocal of the slope of the tangent line. It follows that:

$$y_n=-\frac{5}{4}(x-2)+4=-\frac{5}{4}x+\frac{13}{2} \tag{6.3.252}$$

Ok now let us focus on an example applied to physics (putting apart the famous case of looking for the tangent of an ellipse). We will prove in the section of Continuum Mechanics the Van der Waals equation (see page 3907):

$$\left(P+\frac{n^2a}{V^2}\right)(V-nb)=nRT \tag{6.3.253}$$

If T remains constant, consider that we want to find the rate of change of the volume with respect to the pressure, that is to say $\mathrm{d}V/\mathrm{d}P$! We can challenge here to calculate this rate to transform this equation into an explicitly defined function... So we know we have to use implicit differentiation. First:

$$\frac{\mathrm{d}}{\mathrm{d}P}\left(\left(P+\frac{n^2a}{V^2}\right)(V-nb)\right)=\frac{\mathrm{d}}{\mathrm{d}P}(nRT) \tag{6.3.254}$$

This gives immediately:

$$\left(1-\frac{n^2a}{V^3}\frac{\mathrm{d}V}{\mathrm{d}P}\right)(V-nb)+\left(P+\frac{n^2a}{V^2}\right)\frac{\mathrm{d}V}{\mathrm{d}P}=0 \tag{6.3.255}$$

Now we distribute:

$$V-nb-\frac{n^2a}{V^2}\frac{\mathrm{d}V}{\mathrm{d}P}+\frac{n^3ab}{V^3}\frac{\mathrm{d}V}{\mathrm{d}P}+P\frac{\mathrm{d}V}{\mathrm{d}P}+\frac{n^2a}{V^2}\frac{\mathrm{d}V}{\mathrm{d}P}=0 \tag{6.3.256}$$

After simplification and rearrangement we get:

$$\left(\frac{n^3ab}{V^3}+P\right)\frac{\mathrm{d}V}{\mathrm{d}P}=nb-V \tag{6.3.257}$$

Finally:

$$\frac{\mathrm{d}V}{\mathrm{d}P}=\frac{nb-V}{\dfrac{n^3ab}{V^3}+P} \tag{6.3.258}$$

The special case of the examples above with the Van der Waals equation and the folium of Descartes is in some text books introduced as following with a two variables case:

$$\frac{\mathrm{d}}{\mathrm{d}x}f(x,y)=\frac{\mathrm{d}}{\mathrm{d}x}(c^{te}) \tag{6.3.259}$$

Therefore:

$$\frac{\partial f}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}x}+\frac{\partial f}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}x}=0 \tag{6.3.260}$$

Reduce to:

$$\frac{\partial f}{\partial x}+\frac{\partial f}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}x}=0 \tag{6.3.261}$$

Finally:

$$\boxed{\frac{\mathrm{d}y}{\mathrm{d}x}=-\frac{\dfrac{\partial f}{\partial x}}{\dfrac{\partial f}{\partial y}}} \tag{6.3.262}$$

That's it for now. We will stop here on this topic as we don't need more techniques or example to our study of physics and engineering as introduced in this textbook.

6.3.1.4 Smoothness

Smoothness has to do with how many derivatives of a function exist and are continuous. The term "smooth function" is often used technically to mean a function that has derivatives of all orders everywhere in its domain.

Definition (#6.3.197): A "differentiability class" is a classification of functions according to the properties of their derivatives. Higher order differentiability classes correspond to the existence of more derivatives.

Consider an open set on the real line $\mathbb{R}$ and a function f defined on that set with real or complex values. Let k be a non-negative integer. The function f is said to be of (differentiability) class $\mathcal{C}^k$ if the derivatives $f',f'',\ldots,f(k)$ exist and are continuous (the continuity is implied by differentiability for all the derivatives except for $f(k)$). The function f is said to be of class $\mathcal{C}^\infty$, or smooth, if it has derivatives of all orders. The function f is said to be of class $\mathcal{C}^\omega$, or simply "analytic", if f is smooth and if it equals its Taylor series expansion around any point in its domain (see section Sequences and Series page 1696). $\mathcal{C}^\omega$ is thus strictly contained in $\mathcal{C}^\infty$.

6.3.2 Integral Calculus

We will discuss here the basic principles of integral calculus in $\mathbb{R}^n$. Most advanced topics will come depending the time that are available to the redactors of this textbook but the reader can already refer to the Complex Analysis section (see page 2390) for integration techniques based on the Residue Theorem that is very powerful and useful especially for some integral in quantum physics.

> ❞ *Differentiation is a science, Integration is an art.* ❞

quote
6.3:1497.#53

For information this textbook in mainly dedicated to the Darboux integral, the Riemann integral and the Lebesgue integral. Only if we have the time and there is a practical necessity for this, we will also in the future introduce the Regulated Integral, the Riemann-Stieltjes integral and the Lebesgue-Stielties integral!

6.3.2.1 Definite Integral

The origin of Integral Calculus seems to come from Archimedes that was fascinated with calculating the areas of various shapes. He used a process that has come to be known as the *method of exhaustion*, which used smaller and smaller shapes, the areas of which could be calculated exactly, to fill an irregular region and thereby obtain closer and closer approximations to the total area. In this process, an area bounded by curves is filled with rectangles, triangles, and shapes with exact area formulas. These areas are then summed to approximate the area of the curved region. This subsection introduces the Definite Integrals.

The first idea of the concept of integral is to calculate the algebraic area (positive area if above the x-axis or negative below) between a curve and its support. See the figure below with a positive area and the notations for the developments that will follow:

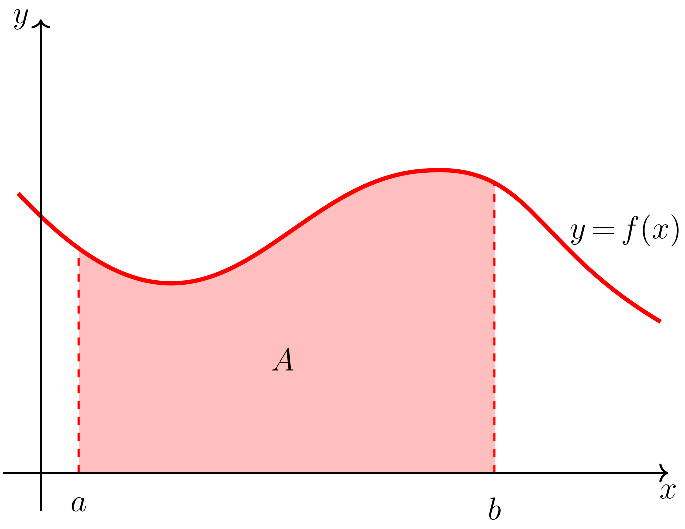


Figure 6.39: **Area A to be calculated in a bounded continuous positive function**

Or with different algebraic areas (the difference between the blue area and the yellow area is named the "net signed area"):

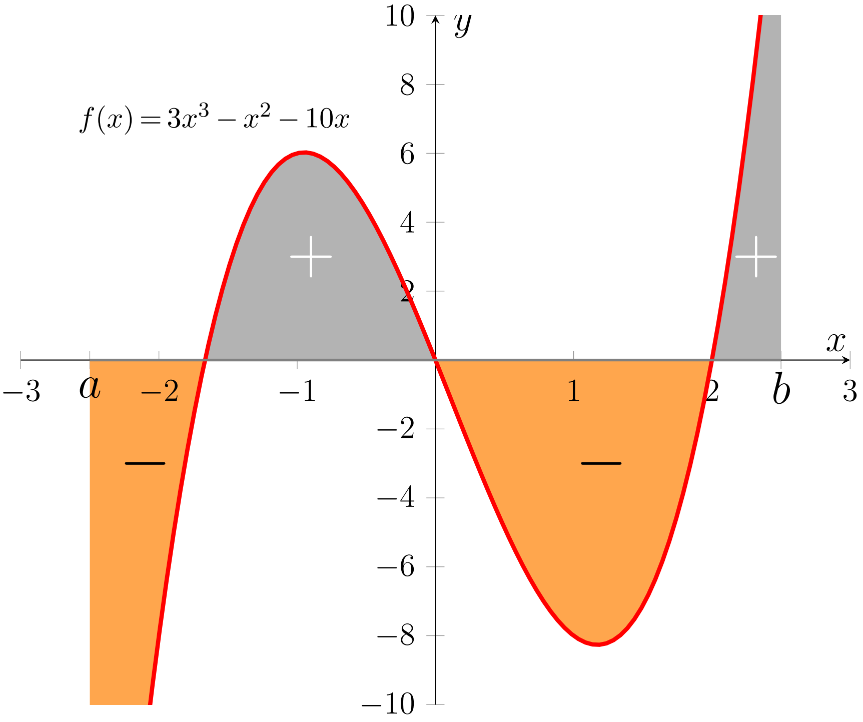


Figure 6.40: **Area A to be calculated in a bounded continuous positive and negative function**

An approximate value of the area under a curve can be achieved by a division into n vertical rectangular bands of the same width (or from different width). In particular (among a lot of other possibilities), we can achieve a framework of this area with the help of a sum of upper bound areas (majorant) A_M (also sometimes denoted S and named "upper Darboux sum") or a lower bound sum (minorant) A_m (also sometimes denoted s and named "lower Darboux sum") for a given cutting:

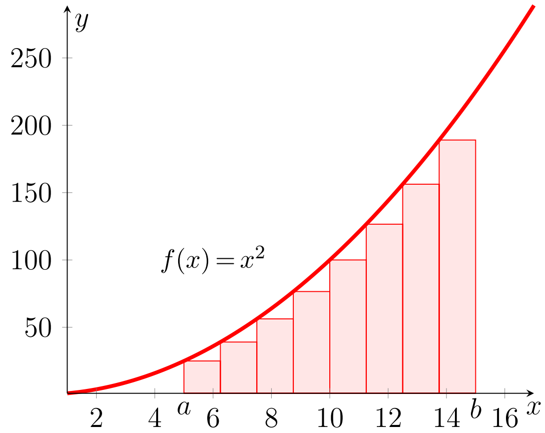


Figure 6.41: **Minorant sum of areas A_m**

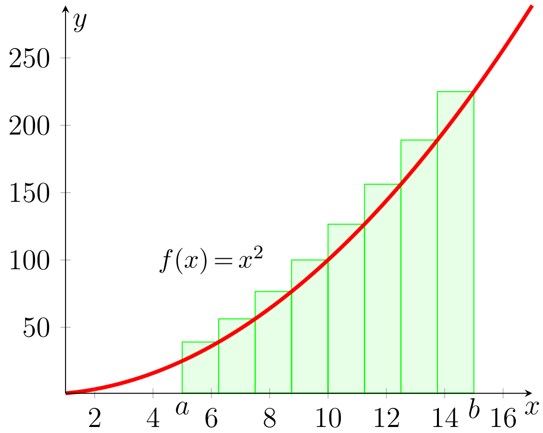


Figure 6.42: **Minorant sum of areas A_M**

Or if we superimposed the both we can see the difference between the minorant and majorant areas and guess that if the number of rectangles tends to infinity, both results converge (it's common representation in many textbooks):

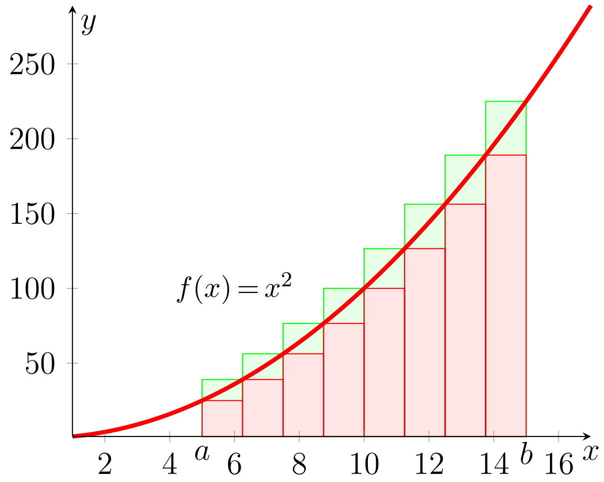


Figure 6.43: **Minorant and majorant sum of areas superimposed**

Suppose now, as mentioned just before, that the number n of bands tends to infinity. As the bands are of the same width, the width of each bands tends to 0 (objectively it is not necessary for the width of the cutting of the subintervals to be the same everywhere).

If the sums A_m and A_M have both a limit when the number n of bands tends to infinity, then the area A under the curve is between these two limits. We write this:

$$\lim_{n\rightarrow+\infty} A_m < A < \lim_{n\rightarrow+\infty} A_M \tag{6.3.263}$$

Obviously if these two limits are equal, their value is that of the area A under the curve.

Hence a first direct definition of the definite integral named also "Riemann integral".

Definition (#6.3.198): Given an interval $[a,b]$, divided into n equal parts, let f be a continuous function on the interval $[a,b]$, consider A_m, the algebraic minorant sum of areas and A_M the algebraic majorant sum. We name "definite integral" of f from a to b, denoted by:

$$\boxed{\int\limits_a^b f(x)\mathrm{d}x} \tag{6.3.264}$$

the number A such that:

$$A=\lim_{n\to+\infty} A_m=\lim_{n\to+\infty} A_M \tag{6.3.265}$$

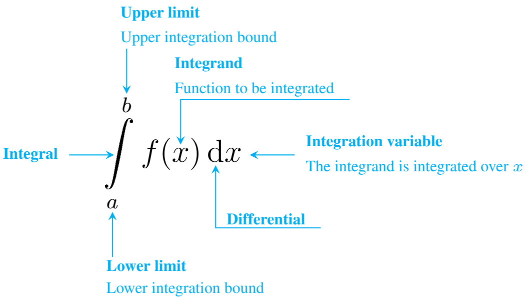
provided that this limit exists. If this limit exists, then we say that f is "integrable" on $[a,b]$ and the definite integral exists.

The symbol:

$$\int_a^b \tag{6.3.266}$$

is only the symbol of a discrete sum $\sum$ but applied to the case of infinitely small increments.

The numbers a and b of the integral (that can also be functions sometimes!) are named the "integration limits" or "integration bounds": a is the "lower bound", b is the "upper bound".


Upper limit
Upper integration bound
Integrand
Function to be integrated
Integral
$\int_a^b f(x)\,\mathrm{d}x$
Integration variable
The integrand is integrated over x
Differential
Lower limit
Lower integration bound

(6.3.267)

Intuitively, it is obvious that when $a=b$ we extend the definition as follows:

$$\int\limits_a^b f(x)\mathrm{d}x=0 \tag{6.3.268}$$

Finally, notice that it is quite possible that the result of the integral to be negative or even complex since it is an algebraic surface! That is to say the result can be in general in $\mathbb{C}$.

Remarks

R1. Other letters rather than x can be used in the evaluation of the definite integral. So if f is integrable on $[a,b]$, then $\int\limits_a^b f(x)\mathrm{d}x=\int\limits_a^b f(t)\mathrm{d}t=\int\limits_a^b f(s)\mathrm{d}s$, etc. This is why the variable x is sometimes named "dummy variable".

R2. As we will see below, it is essential not to confuse between a "definite integral" and a "indefinite integral". Thus, an indefinite integral, denoted $\int f(x)\mathrm{d}x$ is a function, or more precisely, a family of functions also named "primitives of f" (see below) while a definite integral, denoted $\int\limits_a^b f(x)\mathrm{d}x$ is considered as a constant.

Let us present a second approach for the definition of the integral, somewhat more rigorous than the previous one (following the request of several readers). We will use this time, by tradition the notation S surface instead of the area A.

Let f be a bounded function on $[a,b]$. We consider a subdivision σ of its support $[a,b]$ that we note:

$$\sigma=\{x_0=a<x_1<x_2\ldots<x_{n-1}<x_n=b\} \tag{6.3.269}$$

where the intervals are not necessarily of equal sizes.

We write for $i=1,2,3,\ldots,n$:

$$m_i=\inf_{x\in[x_{i-1},x_i]} f(x) \quad \text{and} \quad M_i=\sup_{x\in[x_{i-1},x_i]} f(x) \tag{6.3.270}$$

Definitions (#6.3.199):

D1. We name "lower Darboux sum" (also sometimes named "minorant sum" and denoted A_m) associated with f and σ the surface:

$$s_{[a,b]}(f,\sigma)=\sum_{i=1}^{n} m_i(x_i-x_{i-1}) \tag{6.3.271}$$

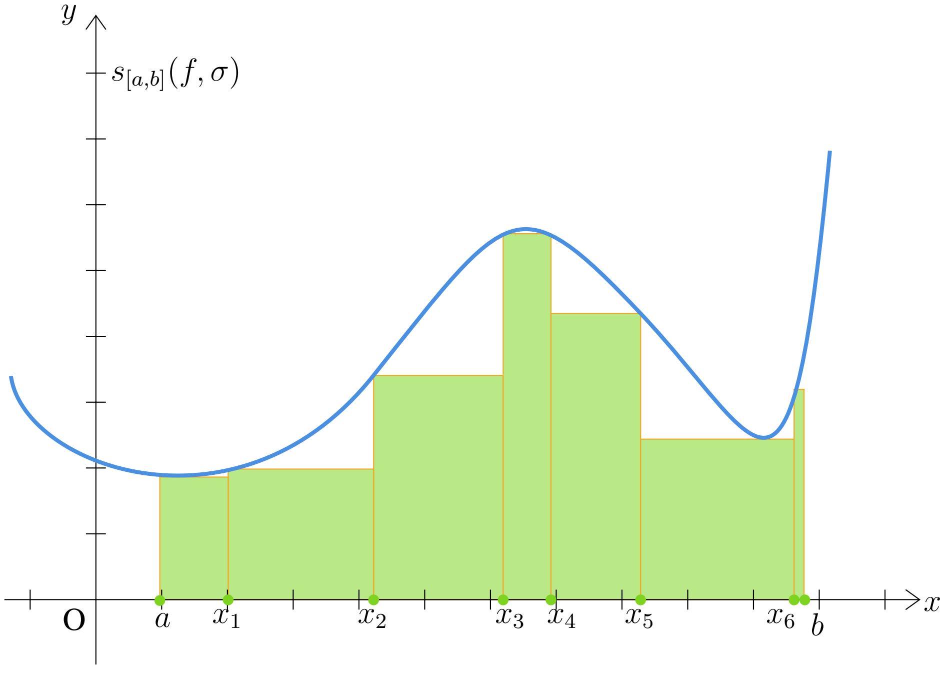


Figure 6.44: **Principle of calculating the lower Darboux sum**

D2. We name "upper Darboux sum" (also sometimes named "majorant sum" and denoted A_M) associated with f and σ the surface:

$$S_{[a,b]}(f,\sigma)=\sum_{i=1}^{n} M_i(x_i-x_{i-1}) \tag{6.3.272}$$

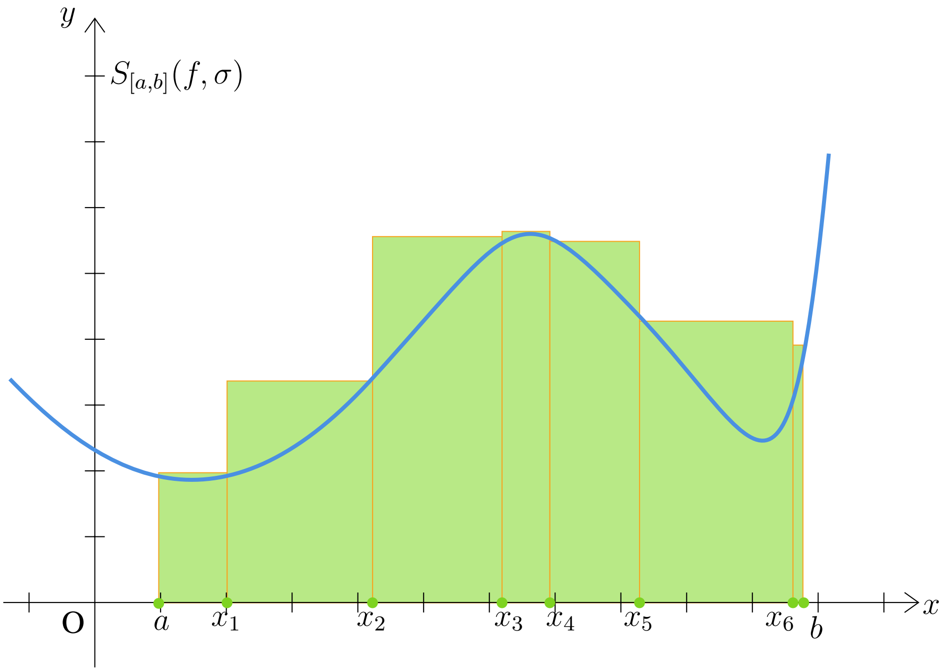


Figure 6.45: Principle of calculating the upper Darboux sum

A function f is said to be "Riemann-integrable on $[a,b]$" if and only if the above two surfaces coincide when the intervals become infinitely small.

$$\int\limits_a^b f(x)\mathrm{d}x \tag{6.3.273}$$

All Riemann-integrable functions on $[a,b]$ are denoted sometimes by $\mathcal{R}_{[a,b]}$.

We can make a quick proof and dirty proof of that equivalence (as it's not very interesting for an engineer because the result is quite intuitive).

Theorem 6.14. Remember that we have the following definition of the Darboux integrability: Let f be defined and bounded on $[a,b]$, then f is Darboux integrable if for all $\varepsilon>0$ there exists a partition P of $[a,b]$ such that $U(f,P)-L(f,P)<\varepsilon$ (where U and L are the upper and lower Riemann sums respectively).

A function is "Riemann integrable" if:

$$\sup_P\{L(f,P)\}=\underline{\int_a^b} f(x)=\overline{\int_a^b} f(x)=\inf_P\{U(f,P)\} \tag{6.3.274}$$

or equivalently:

$$\inf_P\{U(f,P)\}-\sup_P\{L(f,P)\}<\varepsilon \tag{6.3.275}$$

for any $\varepsilon>0$. In this case the supremum of the lower sums is the same as the infimum of the upper sums.

Proof 6.14.1. Let $\varepsilon>0$. There exists P such that $U(f,P)-L(f,P)<\varepsilon$. Then since:

$$L(f,P)\leq\sup_P\{L(f,P)\}\leq\inf_P\{U(f,P)\}\leq U(f,P) \tag{6.3.276}$$

clearly:

$$\inf_P\{U(f,P)\}-\sup_P\{L(f,P)\}<\varepsilon \tag{6.3.277}$$

holds.

■ Q.E.D.

Example:

Suppose we want to show that the function $f(x)=x$ is Darboux-integrable on the interval $[0,1]$ and determine its value. To do this we partition $[0,1]$ into n equally sized subintervals each of length $1/n$. We denote a partition of n equally sized subintervals as P_n.

Now since $f(x)=x$ is strictly increasing on $[0,1]$, the infimum on any particular subinterval is given by its starting point. Likewise the supremum on any particular subinterval is given by its end point. The starting point of the k-th subinterval in P_n is $(k-1)/n$ and the end point is k/n. Thus the lower Darboux sum on a partition P_n is given by:

$$\begin{aligned}L_{f,P_n}&=\sum_{k=1}^{n}f\left(x_{k-1}\right)\left(x_k-x_{k-1}\right)=\sum_{k=1}^{n}\frac{k-1}{n}\cdot\frac{1}{n}=\frac{1}{n^2}\sum_{k=1}^{n}[k-1]\\&=\frac{1}{n^2}\left[\frac{(n-1)n}{2}\right]\end{aligned} \tag{6.3.278}$$

similarly, the upper Darboux sum is given by:

$$\begin{aligned}U_{f,P_n}&=\sum_{k=1}^{n}f\left(x_{k}\right)\left(x_k-x_{k-1}\right)=\sum_{k=1}^{n}\frac{k}{n}\cdot\frac{1}{n}=\frac{1}{n^2}\sum_{k=1}^{n}k\\&=\frac{1}{n^2}\left[\frac{(n+1)n}{2}\right]\end{aligned} \tag{6.3.279}$$

Since:

$$U_{f,P_n}-L_{f,P_n}=\frac{1}{n} \tag{6.3.280}$$

Thus for given any $\varepsilon > 0$, we have that any partition P_n with $n > \frac{1}{\varepsilon}$ satisfies:

$$U_{f,P_n} - L_{f,P_n} < \varepsilon \tag{6.3.281}$$

which shows that f is Darboux integrable. To find the value of the integral note that:

$$\int\limits_0^1 f(x)dx = \lim_{n\to\infty} U_{f,P_n} = \lim_{n\to\infty} L_{f,P_n} = \frac{1}{2} \tag{6.3.282}$$

In other words:

$$\underline{\int_a^b} f(x) - \overline{\int_a^b} f(x) = U(f,P_n) - L(f,P_n) = \mathcal{O}\left(\frac{1}{n}\right) \tag{6.3.283}$$

And as $n \to +\infty$, we have:

$$\underline{\int_a^b} f(x) - \overline{\int_a^b} f(x) = 0 \tag{6.3.284}$$

Hence:

$$\underline{\int_a^b} f(x) = \overline{\int_a^b} f(x) \tag{6.3.285}$$

Therefore the function is also Riemann integrable!

Darboux sums are not very useful for the effective calculation of an integral, for example using a computer, because it is usually quite difficult to find the $\inf$ and $\sup$ on sub-intervals. Rather, we consider:

$$s_{[a,b]}(f) = \sum_{i=1}^{n} (x_i - x_{i-1}) f(x_{i-1}) \quad \text{or} \quad S_{[a,b]}(f) = \sum_{i=1}^{n} (x_i - x_{i-1}) f(x_i) \tag{6.3.286}$$

The "Riemann sum" is defined from the fact that we if denote a "partition" (or "regular partition" if they all have the same width) of the interval $[a,b]$ by:

$$\Delta x_i = x_i - x_{i-1} \tag{6.3.287}$$

and that:

$$f(x_i) = f(\xi_i) \tag{6.3.288}$$

where $\xi_i \in [x_{i-1}, x_i]$, then:

$$\boxed{S \cong s_{[a,b]}(f) = S_{[a,b]}(f) = \sum_{i=1}^{n} f(\xi_i)\Delta x_i} \tag{6.3.289}$$

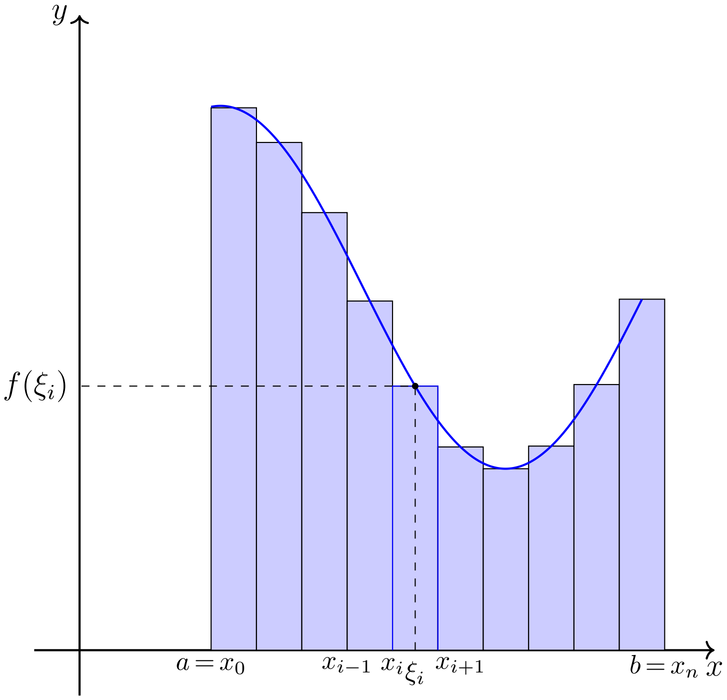

y
$f(\xi_i)$
$a = x_0$
x_{i-1}
x_i
ξ_i
x_{i+1}
$b = x_n$
x

Figure 6.46: **Riemann Sum**

But as we must choose an ξ_i, we often takes either the right or the left one, thus taking randomly the "method of left rectangles":

$$S \cong \sum_{i=1}^{n} f(x_{i-1})\Delta x_i \tag{6.3.290}$$

Which would give us the example below:

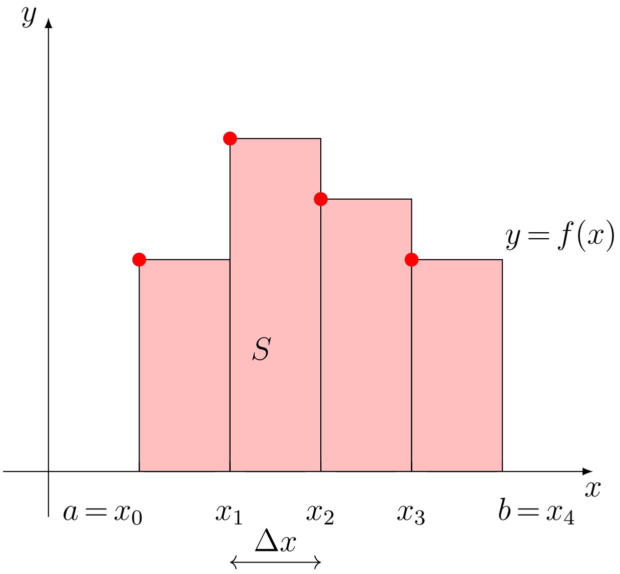
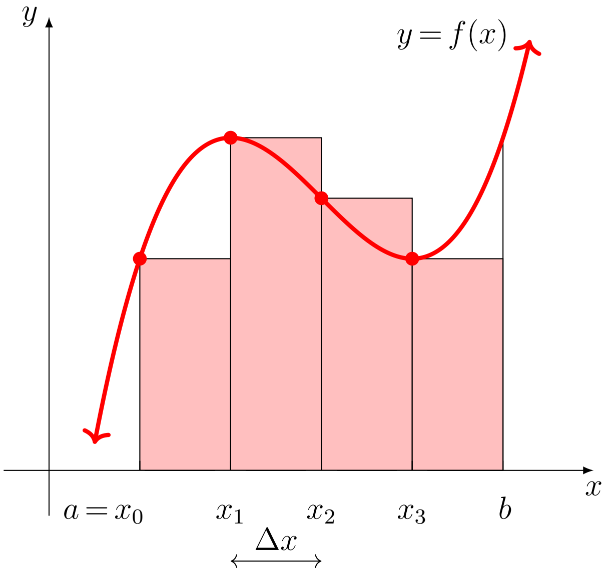

y
$y = f(x)$
S
x
$a = x_0$
x_1
x_2
x_3
$b = x_4$
Δx

Figure 6.47: **Principle of calculating methods using left rectangles**

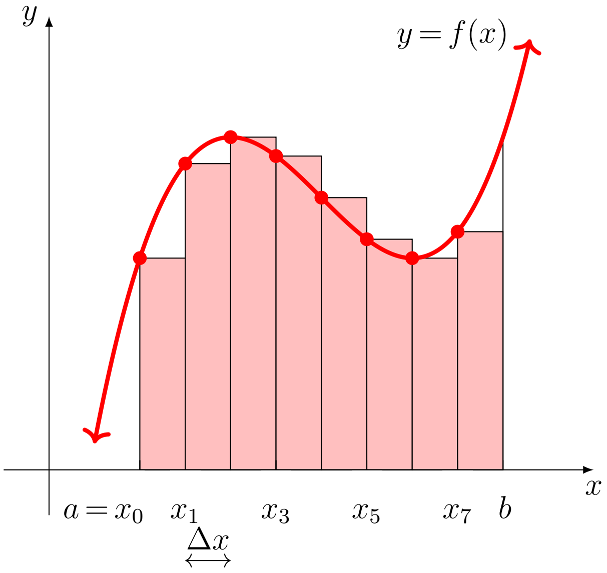
either:

$$\begin{aligned} S &= \sum_{i=1}^{4} f(x_{i-1})\Delta x_i = f(x_0)\Delta x_1 + f(x_1)\Delta x_2 + f(x_2)\Delta x_3 + f(x_3)\Delta x_4 \\ &= f(x_0)(x_1 - x_0) + f(x_1)(x_2 - x_1) + f(x_2)(x_3 - x_2) + f(x_3)(x_4 - x_3) \end{aligned} \tag{6.3.291}$$

But it is easy for a step function... but it is less so for a continuous function for which we will get only an approximate value of the actual surface! The idea is then to take intervals smaller and smaller:



y
$y = f(x)$
x
$a = x_0$
x_1
x_2
x_3
b
Δx



y
$y = f(x)$
x
$a = x_0$
x_1
x_3
x_5
x_7
b
Δx

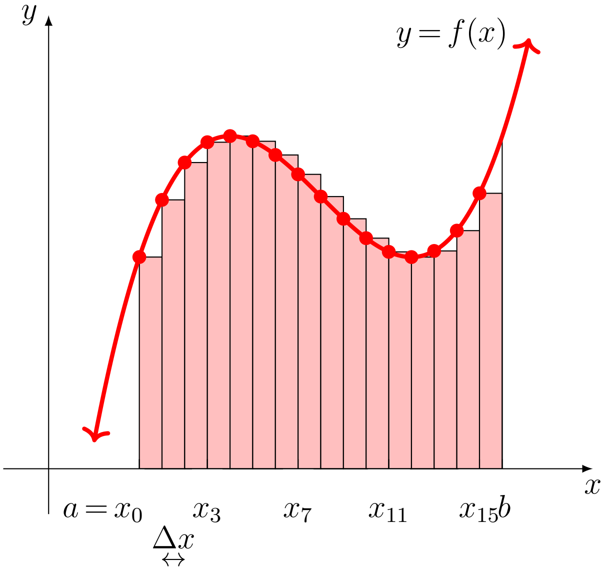


Figure 6.48: **Principle of the calculation of the Riemann integral with left rectangles method**

And then, at the limit, we get the desired quantity:

$$S \underset{\lim \Delta x_i \to 0}{=} \int\limits_a^b f(x)\mathrm{d}x \tag{6.3.292}$$

The fact to search this limit is named "calculate the integral," and more specifically for the chosen method: "calculate the Riemann integral".

✍**Example:**

We want to use the construction of the definite integral to evaluate:

$$\int\limits_0^2 x^2\mathrm{d}x \tag{6.3.293}$$

using a right-endpoint approximation to generate the Riemann sum.

For this purpose we first set up the Riemann sum. Based on the limits of integration, we have $a=0$ and $b=2$. For $i=0,1,2,\ldots,n$, let $P=\{x_i\}$ be a regular partition of $[0,2]$. Then:

$$x_i = x_0 + i\Delta x = 0 + i\left[\frac{2}{n}\right] = \frac{2i}{n} \tag{6.3.294}$$

Thus, the function value at the right endpoint of the interval is:

$$f(x_i) = x_i^2 = \left(\frac{2i}{n}\right)^2 = \frac{4i^2}{n^2} \tag{6.3.295}$$

Then the Riemann sum takes the form:

$$\sum_{i=1}^{n} f(x_i)\Delta x = \sum_{i=1}^{n}\left(\frac{4i^2}{n^2}\right)\frac{2}{n} = \sum_{i=1}^{n}\frac{8i^2}{n^3} = \frac{8}{n^3}\sum_{i=1}^{n} i^2 \tag{6.3.296}$$

Using the Gauss summation relation of $\sum_{i=1}^{n} i^2$ (see section Sequences and Series page 1667), we have:

$$\begin{aligned}\sum_{i=1}^{n} f(x_i)\Delta x &= \frac{8}{n^3}\sum_{i=1}^{n} i^2 = \frac{8}{n^3}\left[\frac{n(n+1)(2n+1)}{6}\right] = \frac{8}{n^3}\left[\frac{2n^3+3n^2+n}{6}\right]\\ &= \frac{16n^3+24n^2+n}{6n^3} = \frac{8}{3}+\frac{4}{n}+\frac{1}{6n^2}\end{aligned} \tag{6.3.297}$$

Now, to calculate the definite integral, we need to take the limit as $n \to +\infty$. We get:

$$\begin{aligned}\int_0^2 x^2 \mathrm{d}x &= \lim_{n\to+\infty}\sum_{i=1}^{n} f(x_i)\Delta x = \lim_{n\to+\infty}\left(\frac{8}{3}+\frac{4}{n}+\frac{1}{6n^2}\right)\\ &= \lim_{n\to+\infty}\left(\frac{8}{3}\right)+\lim_{n\to+\infty}\left(\frac{4}{n}\right)+\lim_{n\to+\infty}\left(\frac{1}{6n^2}\right) = \frac{8}{3}+0+0 = \frac{8}{3}\end{aligned} \tag{6.3.298}$$

Remark

The definition of the Darboux integral has the advantage of being easier to apply in computations or proofs than that of the "real" Riemann integral that doesn't take necessarily equal with interval and therefore defined by:

$$\int_a^b f(t)\mathrm{d}t := \lim_{\substack{p\to 0\\ \sum^*(p)}} \sum_{k=1}^{n} f(t_k)\times\Delta x_k \tag{6.3.299}$$

where $\sum^*(p)$ is the set of subdivisions of $[a,b]$ such that $\max|\Delta x_k| \leq p; t_k \in [x_k, x_{k+1}]$ (disclaimer: of course for Riemann this limit must exist in the $\varepsilon-\delta$ sense!). Consequently, introductory textbooks on calculus and real analysis often develop Riemann integration using the Darboux integral, rather than the true Riemann integral.

6.3.2.2 Indefinite Integral

We have seen before in our study of derivatives the following problem: given a function denoted by $F(x)$, find its derivative $f(x)$, that is to say the algebraic function:

$$F'(x) = f(x) \tag{6.3.300}$$

Definition (#6.3.200): We say that the function $F(x)$ is a "primitive" or "indefinite integral" of the function $f(x)$ on any segment $[a,b]$ if at any point in any segment we have $F'(x) = f(x)$.

Two more explicit and less technical alternative definitions are:

- An "indefinite integral" is a FUNCTION of x (or another variable), while a "definite integral" is a VALUE!

- The collection of all antiderivatives of $f(x)$ is named the "indefinite integral" of f with respect to x.

Another way to see the indefinite integral concept is to go through the "fundamental theorem of integral and differential calculus" also sometimes named "fundamental theorem of calculus" whose two properties are stated as follows:

1. If A (area) is the function defined by $A(X)=\int\limits_a^X f(t)\mathrm{d}t$ for each X in any $[a,b]$, then A is the primitive of f on $[a,b]$ which is zero in a (or in other words: $f(t)$ is the derivative of A).
2. If F is a primitive of f on any $[a,b]$, then:

$$\boxed{\int\limits_a^b f(t)\mathrm{d}t=F(b)-F(a)} \tag{6.3.301}$$

that can also be written obviously as following:

$$\int\limits_a^b F'(t)\mathrm{d}t=\int\limits_a^b \frac{\mathrm{d}F}{\mathrm{d}t}\mathrm{d}t=F(b)-F(a) \tag{6.3.302}$$

Some people then say that the sum of the changes outside of a function (the left side of the equality) is equal to the changes inside the function (the right side of the equality). That's typically a special 1-dimensional case statement of the Stokes (general manifold) theorem!

Let us prove the first property of this fundamental theorem:

Proof 6.14.2. Given the function:

$$A(X)=\int\limits_a^X f(t)\mathrm{d}t \tag{6.3.303}$$

If f is positive, and $h>0$ (the proof in the case where $h<0$ is similar) and as $X>0$, we know that we can think of $A(X)$ as the area under the curve of f from $t=a$ to $t=X$.

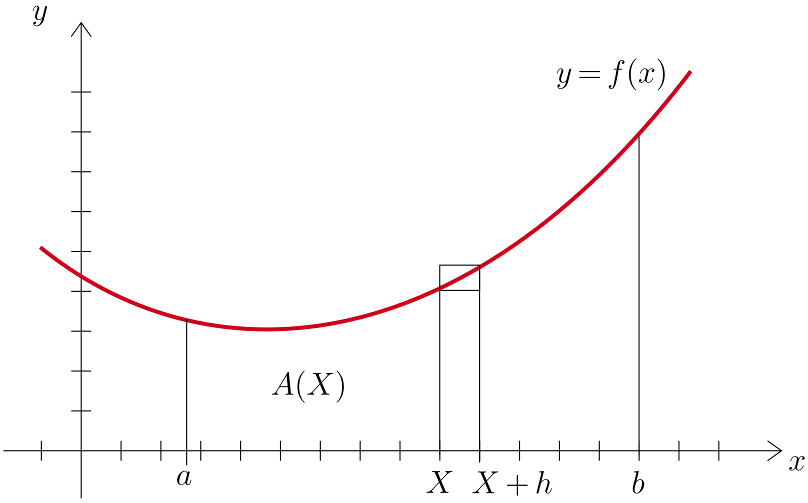


Figure 6.49: **Graphical representation of the area**

To show that A is a primitive of f, we will prove that $A' = f$. According to the definition of the derivative:

$$A'(X) = \lim_{h \to 0} \frac{A(X+h) - A(X)}{h} \tag{6.3.304}$$

Let us study this quotient: $A(x+h) - A(x)$ is represented by the area of the strip of width h, sandwiched between two rectangles of width h.

Given M the maximum of f on the interval $[X, X+h]$ and m the minimum interval of f over the same interval. The respective areas of the two rectangles are Mh and mh.

We then have the following double inequality:

$$mh \leq A(X+h) - A(X) \leq Mh \tag{6.3.305}$$

As h is positive, we can divide by h without changing the meaning of the inequalities:

$$m \leq \frac{A(X+h) - A(X)}{h} \leq M \tag{6.3.306}$$

When $h \to 0^+$ and if f is a continuous function, then M and m have for limit $f(X)$, and the ratio:

$$\frac{A(X+h) - A(X)}{h} \tag{6.3.307}$$

which is between m and M, has effectively for limit $f(X)$.

As $A'(X) = f(X)$ for all X, this shows that the derivative of the area function is f. And therefore A is an indefinite primitive of f. As $A(a) = 0$, A is effectively the primitive of f which vanishes on a.

Another very short and nice way to prove that is to write:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}F(x)&=\lim_{h\to 0}\frac{1}{h}\left[\int\limits_a^{x+h}f(t)\mathrm{d}t-\int\limits_a^{x}f(t)\mathrm{d}t\right]\\&=\lim_{h\to 0}\frac{1}{h}\int\limits_x^{x+h}f(t)\mathrm{d}t\\&=\lim_{h\to 0}\frac{1}{h}\left[f(x)h+\mathcal{O}\left(h^2\right)\right]\\&=f(x)\end{aligned}\tag{6.3.308}$$

■ Q.E.D.

Before starting the proof of the second properties of the fundamental theorem, let us give and prove the following theorem that will we be essential to us: If $F_1(x)$ and $F_2(x)$ are two primitives of $f(x)$ on any segment $[a,b]$, their difference is a constant (this result is very important in physics in terms of the study of what we name the "initial conditions").

Proof 6.14.3. We have after the definition of the concept of Primitive:

$$\begin{gathered}F_1'(x)=f(x)\\F_2'(x)=f(x)\end{gathered}\tag{6.3.309}$$

for $\forall x\in[a,b]$.

Let us write:

$$F_1(x)-F_2(x)=\varphi(x)\tag{6.3.310}$$

We can write:

$$F_1'(x)-F_2'(x)=\left[F_1(x)-F_2(x)\right]'=\varphi'(x)=f(x)-f(x)=0\tag{6.3.311}$$

So it comes from what we saw during our study of derivatives that:

$$\varphi(x)=c^{te}\tag{6.3.312}$$

Then we have:

$$F_1(x)-F_2(x)=c^{te}\tag{6.3.313}$$

■ Q.E.D.

It follows from this theorem that if we know any primitive $F(x)$ of the function $f(x)$ any other primitive of this function will be of the form:

$$F(x)+c^{te}\tag{6.3.314}$$

where c^{te} is named the "constant of integration".

So finally, we name "indefinite integral" of $f(x)$ and we denote by:

$$\int f(x)\mathrm{d}x \tag{6.3.315}$$

any expression of the form $F(x)+c^{te}$ where $F(x)$ is a primitive of $f(x)$. Thus, by writing convention:

$$\boxed{\int f(x)\mathrm{d}x=F(x)+c^{te}} \tag{6.3.316}$$

if and only if $F'(x)=f(x)$.

In this context, $f(x)$ is also named "integrand function" and $f(x)\mathrm{d}x$, the "function under the sum sign".

Remark

An "antiderivative" of a function f is one function F whose derivative is f. The indefinite integral of f is the set of ALL antiderivatives of f. Therefore indefinite integral and antiderivative is not the same as the first is the set of all the seconds! If f and F are as described just now, the indefinite integral of f has the form $\{F+c^{te}|c^{te}\in\mathbb{R}\}$ when an antiderivative is just one element of this set!! But not all teachers are according on the definition of antiderivatives...

Caution!!! Sometimes not every antiderivative is an area function. This is exactly what's happening with the constant function $f(x)=0$. No matter what a we pick, we will always have:

$$\int\limits_a^x 0\mathrm{d}t=0 \tag{6.3.317}$$

for every x. And it's not just the always zero function which does this!

Geometrically, we can consider the indefinite integral as a set (family) of curves as we move from one to another by performing a translation in the positive or negative direction of the axis.

Let us return to the proof of item (2) of the fundamental theorem of integral (and differential) calculus:

Proof 6.14.4. Let F be a primitive of f. Since two primitives differ by a constant, we have well:

$$A(X)=F(X)+c^{te} \tag{6.3.318}$$

that we can also write:

$$\int\limits_a^X f(t)\mathrm{d}t=F(X)+c^{te} \tag{6.3.319}$$

for all X in $[a,b]$. The particular case $X=a$ gives $\int\limits_a^a f(t)\mathrm{d}t$ and therefore $F(a)+c^{te}=0$ and we get obviously $c^{te}=-F(a)$. Substituting, we get:

$$\int\limits_a^X f(t)\mathrm{d}t=F(X)-F(a) \tag{6.3.320}$$

As this identity is valid for all X in the interval $[a,b]$ it is true especially for $X=b$. Therefore:

$$\boxed{\int\limits_a^b f(t)\mathrm{d}t=F(b)-F(a)} \tag{6.3.321}$$

■ Q.E.D.

This last result also shows something useful!: It is not necessary when we evaluate and integral to take into account the constant of the general primitive since it is cancelled by the difference of the two primitives!!

Remarks

R1. The above fundamental theorem that shows the link between primitive and integral led us to use the same symbol $\int$ to write a primitive (introduced by Leibniz in the late 117th century), which is a function, and an integral that it is a number.

R2. We also proved in the section of Analytical Mechanics using an integral how to calculate the full length of a curve in the plane if the function $f(x)$ is explicitly known.

Here are some trivial properties of integration that is good to remember because often used elsewhere in this textbook (if it do not seem obvious to you, contact us and we will give the detailed proof):

P1. The derivative of an indefinite integral is equal to the integrand:

$$\left(\int f(x)\mathrm{d}x\right)'=\left(F(x)+c^{te}\right)'=f(x) \tag{6.3.322}$$

P2. The differential of an indefinite integral is equal to the expression under the integral sign:

$$\mathrm{d}\left(\int f(x)\mathrm{d}x\right)=f(x) \tag{6.3.323}$$

P3. The indefinite integral of the differential of a given function is equal to the sum of this function and an arbitrary constant:

$$\int F(x)\mathrm{d}x=F(x) \tag{6.3.324}$$

P4. The indefinite integral of the sum (or subtraction) of two or more algebraic functions is equal to the algebraic sum of their integral (do not forget that we work with the set of all primitives and not a specific given one primitive!):

$$\int [f_1(x)+f_2(x)]\,\mathrm{d}x=\int f_1(x)\mathrm{d}x+\int f_2(x)\mathrm{d}x \tag{6.3.325}$$

Proof 6.14.5. To prove this, following the request of a reader, we will prove that the derivative of the left-hand side allows us to find the derivative of the right-hand side and vice versa (reciprocal) with the above properties.

According to P1 we have:

$$\left(\int [f_1(x)+f_2(x)]\,\mathrm{d}x\right)'=\left(\int f_1(x)\mathrm{d}x\right)'+\left(\int f_2(x)\mathrm{d}x\right)'=f_1(x)+f_2(x) \tag{6.3.326}$$

Let us check if it is the same with the right-hand side (we assume known the properties of derivatives that we proved earlier in this section):

$$\begin{aligned}\left(\int f_1(x)\mathrm{d}x+\int f_2(x)\mathrm{d}x\right)'&=\left(\int f_1(x)\mathrm{d}x\right)'+\left(\int f_2(x)\mathrm{d}x\right)'\\&=f_1(x)+f_2(x)\end{aligned} \tag{6.3.327}$$

■ Q.E.D.

P5. We can get out a constant factor from the integral sign, that is to say:

$$\int c^{te}f(x)\mathrm{d}x=c^{te}\int f(x)\mathrm{d}x \tag{6.3.328}$$

We justify this equality by deriving the two members (and according to the properties of derivatives):

$$\begin{aligned}\left(\int c^{te}f(x)\mathrm{d}x\right)'&=c^{te}f(x)\\\left(c^{te}\int f(x)\mathrm{d}x\right)'&=c^{te}\left(\int f(x)\mathrm{d}x\right)'=c^{te}f(x)\end{aligned} \tag{6.3.329}$$

P6. We can take out a constant factor in the argument of the integrated function (rather rarely used):

$$\int f(ax)\mathrm{d}x=\frac{1}{a}F(ax)+c^{te} \tag{6.3.330}$$

Indeed, by differentiating both members of equality we have following the properties of derivatives:

$$\left(\int f(ax)\mathrm{d}x\right)'=\left(\frac{1}{a}F(ax)+c^{te}\right)'=f(ax) \tag{6.3.331}$$

P7. The integration of a function whose argument is summed (or subtracted) is algebraically the primitive of the argument summed (respectively subtracted):

$$\int f(x+b)\mathrm{d}x = F(x+b)+c^{te} \tag{6.3.332}$$

This property can be showed identically to the previous one using also the derivatives properties.

P8. The combination of properties P6 and P7 allows us to write:

$$\int f(ax+b)\mathrm{d}x = \frac{1}{a}F(ax+b)+c^{te} \tag{6.3.333}$$

P9. Let f be a continuous function on $[a,b]$, we have for all c belonging to this interval:

$$\int\limits_a^b f(x)\mathrm{d}x = \int\limits_a^c f(x)\mathrm{d}x + \int\limits_c^b f(x)\mathrm{d}x \tag{6.3.334}$$

This theorem, sometimes named "Chasles relation" (by its vector equivalent) follows directly from the definition of the indefinite integral. F being a primitive of f on $[a,b]$ we have:

$$F(c)-F(a)+F(b)-F(c) = F(b)-F(a) \tag{6.3.335}$$

P10. This is a property often used in the section of Statistics (we do not find an easy way to express this property by everyday language so...):

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}y}\int\limits_{g(x)}^{h(y)} f(t)\mathrm{d}t &= \frac{\mathrm{d}}{\mathrm{d}y}\left(F(h(y))-F(g(x))\right)\\ &= f(h(y))\frac{\mathrm{d}h(y)}{\mathrm{d}y}+f(g(x))\frac{\mathrm{d}g(x)}{\mathrm{d}y}\\ &= f(h(y))h'(y)\end{aligned} \tag{6.3.336}$$

Let us see now two properties that will be helpful we sometimes to calculate difficult integrals:

P11. If a function is even (see section Functional Analysis page 2212), the integral on symmetrical bounds is equivalent to:

$$\begin{aligned}\int\limits_{-L}^{+L} f(x)\mathrm{d}x &= \int\limits_{-L}^{0} f(x)\mathrm{d}x + \int\limits_{0}^{+L} f(x)\mathrm{d}x = \int\limits_{-L}^{0} \underset{u=-x}{f(-x)\mathrm{d}x} + \int\limits_{0}^{+L} f(x)\mathrm{d}x\\ &= -\int\limits_{-L}^{0} f(u)\mathrm{d}u + \int\limits_{0}^{+L} f(x)\mathrm{d}x = \int\limits_{0}^{+L} f(u)\mathrm{d}u + \int\limits_{0}^{+L} f(x)\mathrm{d}x = 2\int\limits_{0}^{+L} f(x)\mathrm{d}x\end{aligned} \tag{6.3.337}$$

P12. If a function is odd (see section Functional Analysis page 2212), the integral on symmetrical bounds is equivalent to:

$$\begin{aligned}\int\limits_{-L}^{+L} f(x)\mathrm{d}x &= \int\limits_{-L}^{0} f(x)\mathrm{d}x + \int\limits_{0}^{+L} f(x)\mathrm{d}x = \int\limits_{-L}^{0} -f(-x)\mathrm{d}x + \int\limits_{0}^{+L} f(x)\mathrm{d}x \\ &= \underset{u=-x}{\int\limits_{0}^{+L} f(-x)\mathrm{d}y} + \int\limits_{0}^{+L} f(x)\mathrm{d}x = -\int\limits_{0}^{+L} f(u)\mathrm{d}u + \int\limits_{0}^{+L} f(x)\mathrm{d}x = 0\end{aligned} \tag{6.3.338}$$

P13. The integral of a periodic function is invariant under a shift of its integration. This is a property that we will use further below to finalize the proof of the integral representation of the zero order Bessel function of the first type.

If f is a periodic function of period T we know that for any value a:

$$\int\limits_{a}^{a+T} f(t)\mathrm{d}t = \int\limits_{0}^{T} f(t)\mathrm{d}t \tag{6.3.339}$$

So now consider:

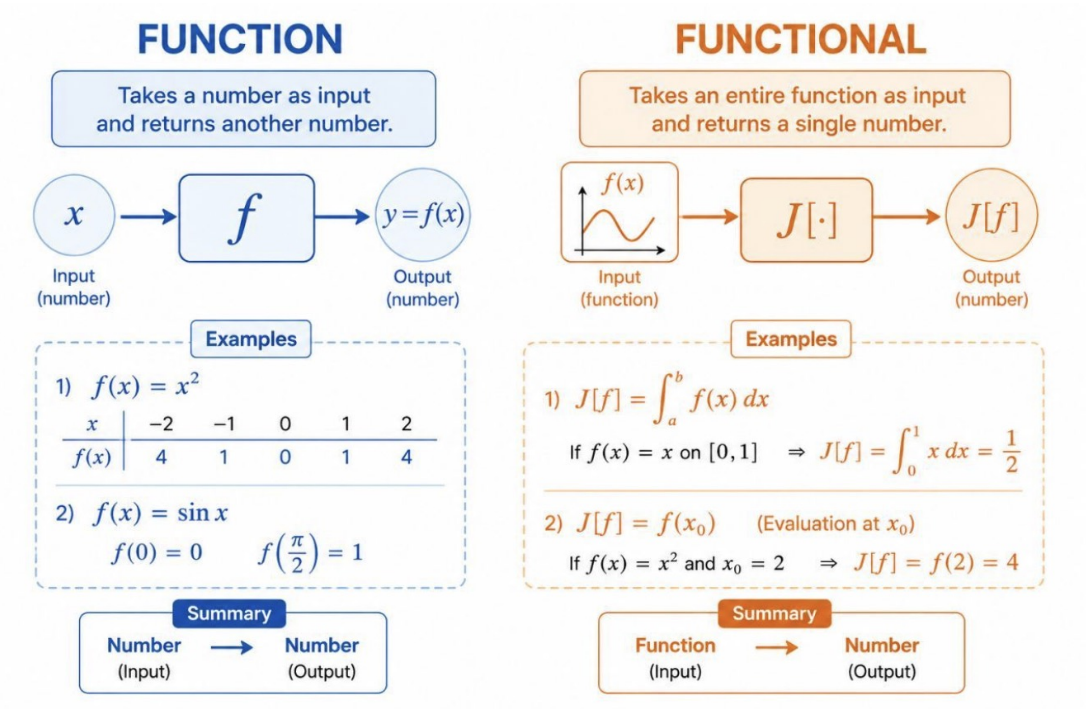
$$\int\limits_{a}^{a+T} f(t)\mathrm{d}t = \int\limits_{0}^{T} f(t)\mathrm{d}t + \int\limits_{T}^{a+T} f(t)\mathrm{d}t \tag{6.3.340}$$

We do the change of variable $y = t - T$, then we have for the last integral:

$$\int\limits_{a}^{a+T} f(t)\mathrm{d}t = \int\limits_{0}^{T} f(y+T)\mathrm{d}y = \int\limits_{0}^{a} f(y)\mathrm{d}y \tag{6.3.341}$$

Therefore:

$$\int\limits_{a}^{a+T} f(t)\mathrm{d}t = \int\limits_{0}^{T} f(t)\mathrm{d}t + \int\limits_{0}^{a} f(y)\mathrm{d}y = \int\limits_{0}^{T} f(t)\mathrm{d}t \tag{6.3.342}$$



FUNCTION

Takes a number as input and returns another number.

$x \rightarrow \boxed{f} \rightarrow y=f(x)$

Input (number) Output (number)

Examples

1) $f(x) = x^2$

x	-2	-1	0	1	2
$f(x)$	4	1	0	1	4

2) $f(x) = \sin x$

$f(0) = 0 \qquad f\left(\frac{\pi}{2}\right) = 1$

Summary

Number (Input) → Number (Output)

FUNCTIONAL

Takes an entire function as input and returns a single number.

$f(x) \rightarrow \boxed{J[\cdot]} \rightarrow J[f]$

Input (function) Output (number)

Examples

1) $J[f] = \int_a^b f(x)\,dx$

If $f(x) = x$ on $[0,1]$ $\Rightarrow J[f] = \int_0^1 x\,dx = \frac{1}{2}$

2) $J[f] = f(x_0)$ (Evaluation at x_0)

If $f(x) = x^2$ and $x_0 = 2$ $\Rightarrow J[f] = f(2) = 4$

Summary

Function (Input) → Number (Output)

6.3.2.3 Double Integral

The idea of double integrals is to measure the volume of the area bounded by the graph of a function of two variables over a domain D of the plane (below D is rectangular):

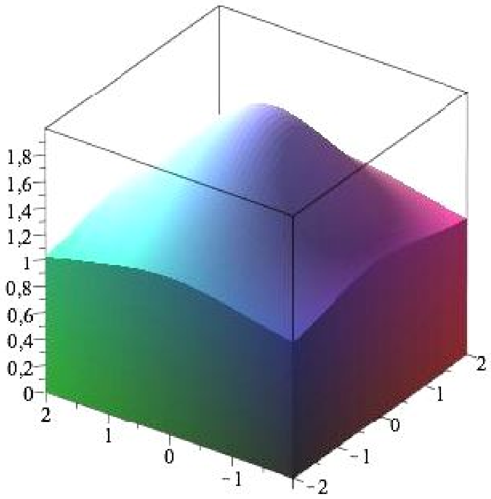


Figure 6.50: **Example of a continuous function of two variables over a squared domain**

It could be obvious to the reader that double integrals are extremely important in the field of Applied Mathematics!

Again, the idea is the same as the definite integral. If we adapt a simplistic approach, we decompose the continuous function like a staircase and the volume to calculate is then reduced to the sum of the volumes of parallelepipeds:

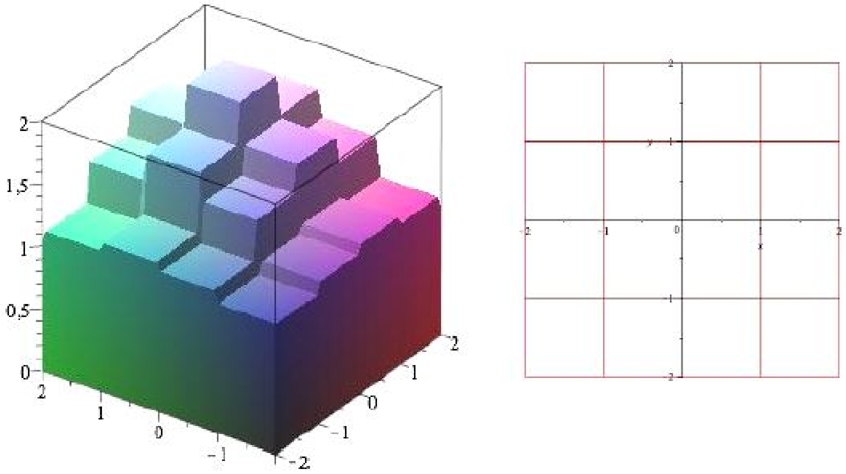


Figure 6.51: **Volume decomposition into big parallelepipeds**

Therefore we have the double sum:

$$V \cong \sum f(x_i,y_i)\Delta x_i \Delta y_i = \sum f(x_i,y_i)\Delta y_i \Delta x_i \qquad (6.3.343)$$

For a continuous function, we proceed by successive approximations: we calculate the Riemann sums for subdivisions always thinner and thinner of the domain D:

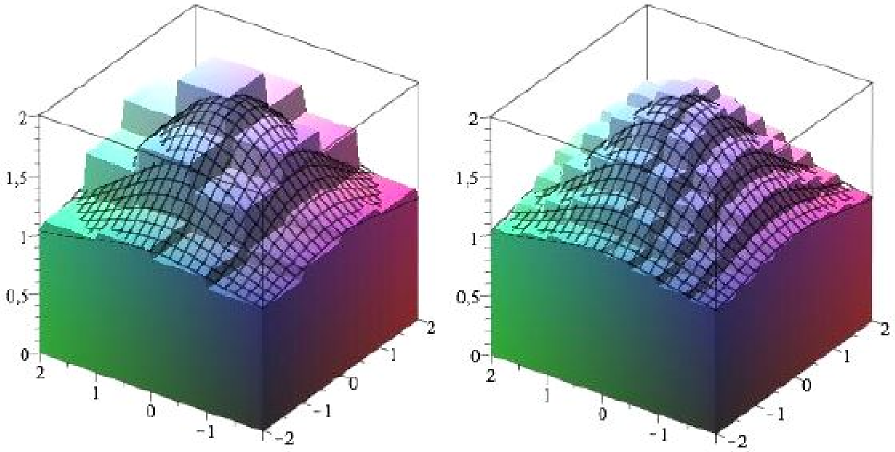


Figure 6.52: Decomposition of the volume in always thinner parallelepipeds

and therefore at the limit:

$$V=\iint_D f(x,y)\mathrm{d}x\mathrm{d}y=\iint_D f(x,y)\mathrm{d}y\mathrm{d}x \tag{6.3.344}$$

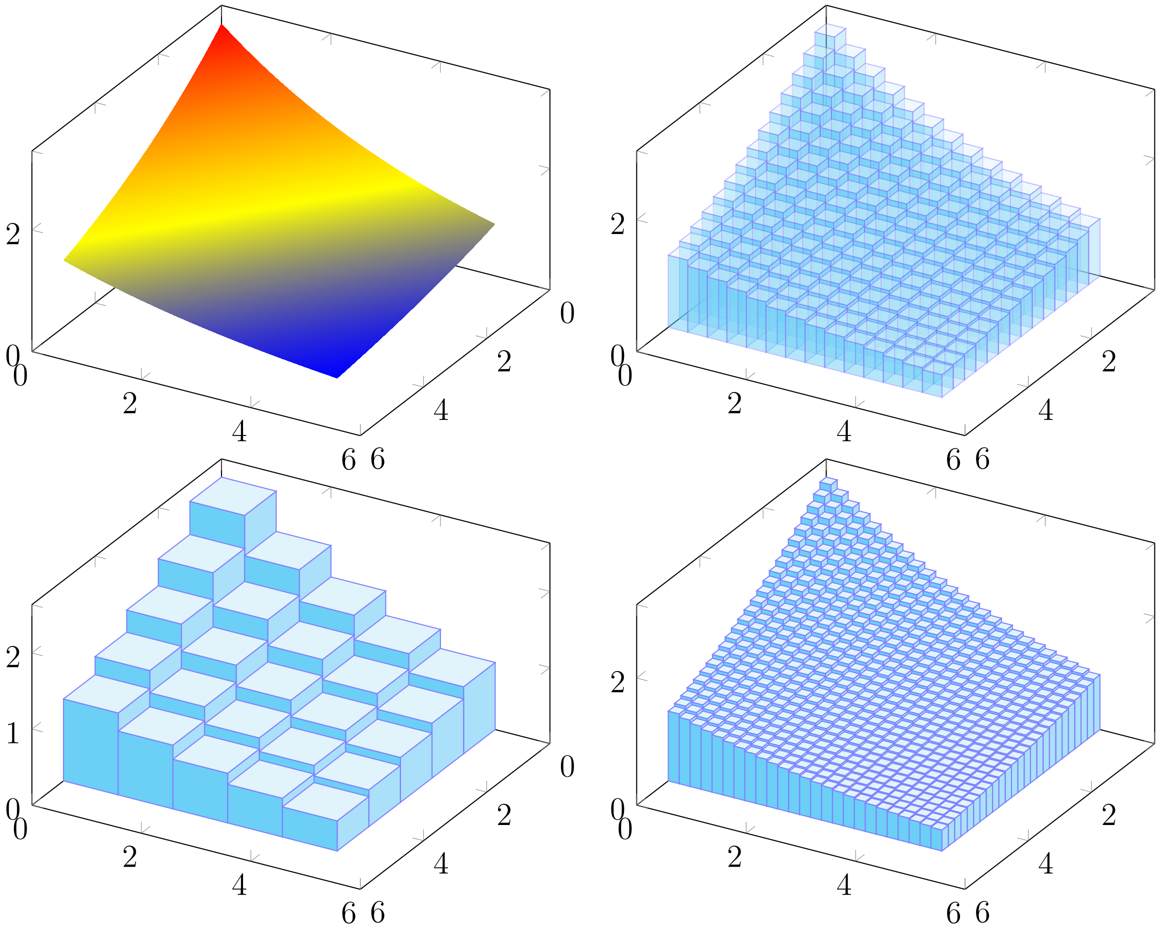


Figure 6.53: Illustration of a double integral on a square domain with TikZ

But..., when we want to integrate an area that is not rectangular, things get a priori more complicated ... Let's see a workaround.

For this purpose, we will build closed bounded domain D as follows:

$$(x,y)\in D\Leftrightarrow\begin{cases}x\in[a,b]\\ y\in[u(x),v(x)]\end{cases} \tag{6.3.345}$$

where the reader will have noticed that the support is the variable x by trough the two functions u and v. This then is what we name a "type I domain" (and therefore if it is y that parametrizes x then it is "type II domain").

This can be illustrated by the figure below:

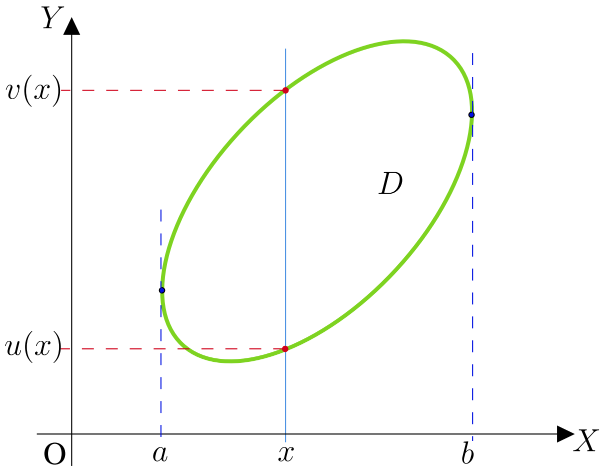


Figure 6.54: **Example of a type I domain**

where we notice that this simplistic approach (there are other possible approaches but that need the use of Measurement Theory) requires that the domain is simply convex[5] (that is to say they are no holes outside the domain D between $u(x)$ and $v(x)$) or decomposed into simply convex disjoint subdomains.

To resume, we can integrate as follows:

$$V=\iint_D f(x,y)\mathrm{d}x\mathrm{d}y=\int\limits_a^b\left(\int\limits_{u(x)}^{v(x)}f(x,y)\mathrm{d}y\right)\mathrm{d}x \tag{6.3.346}$$

So we transform the double integral in two nested simple integrals.

---

[5] As we have seen in the section of Geometric Shapes an area is convex if any couple of points on the bounds of the area with a crossing line has this latter no outside of the area

6.3.2.3.1 Fubini's theorem

We will see now an important theorem used repeatedly in different section of this textbook and which permits to reverse the order of integration.

Remembering that:

$$(x,y)\in D \Leftrightarrow \begin{cases} x\in[a,b] \\ y\in[u(x),v(x)] \end{cases} \tag{6.3.347}$$

we can also use:

$$(x,y)\in D \Leftrightarrow \begin{cases} y\in[c,d] \\ x\in[\alpha(y),\beta(y)] \end{cases} \tag{6.3.348}$$

So with this parametrization we can write:

$$V=\iint_D f(x,y)\mathrm{d}x\mathrm{d}y=\int\limits_a^b\left(\int\limits_{u(x)}^{v(x)} f(x,y)\mathrm{d}y\right)\mathrm{d}x=\int\limits_c^d\left(\int\limits_{\alpha(y)}^{\beta(y)} f(x,y)\mathrm{d}x\right)\mathrm{d}y \tag{6.3.349}$$

We can change the order of integration, the calculation is different, but the result is the same. But that's not what really interests us here!

Consider a function such as (we say that the function can be "variable separable"):

$$f(x,y)=\varphi(x)\psi(y) \tag{6.3.350}$$

Therefore:

$$V=\iint_D f(x,y)\mathrm{d}x\mathrm{d}y=\int\limits_a^b\left(\int\limits_{u(x)}^{v(x)} f(x,y)\mathrm{d}y\right)\mathrm{d}x=\int\limits_a^b\left(\int\limits_{u(x)}^{v(x)} \varphi(x)\psi(y)\mathrm{d}y\right)\mathrm{d}x \tag{6.3.351}$$

Suppose that the domain is a rectangle (we do this simplification otherwise the proof complicates considerably). Meaning:

$$(x,y)\in D \Leftrightarrow \begin{cases} x\in[a,b] \\ y\in[c,d] \end{cases} \tag{6.3.352}$$

Therefore the integral linearity property:

$$\begin{aligned} V&=\int\limits_a^b\left(\int\limits_{u(x)}^{v(x)} \varphi(x)\psi(y)\mathrm{d}y\right)\mathrm{d}x=\int\limits_a^b\left(\int\limits_c^d \varphi(x)\psi(y)\mathrm{d}y\right)\mathrm{d}x=\int\limits_a^b\left(\varphi(x)\int\limits_c^d \psi(y)\mathrm{d}y\right)\mathrm{d}x \\ &=\int\limits_a^b\varphi(x)\left(\int\limits_c^d \psi(y)\mathrm{d}y\right)\mathrm{d}x=\left(\int\limits_c^d \psi(y)\mathrm{d}y\right)\int\limits_a^b\varphi(x)\mathrm{d}x=\int\limits_a^b\varphi(x)\mathrm{d}x\int\limits_c^d \psi(y)\mathrm{d}y \end{aligned} \tag{6.3.353}$$

6.3.2.4 Integration by Substitution

When we can not easily determine the primitive of a given function, we can find sometimes, with a smart change of variables (sometimes very subtle...), bypass the difficulty. It does not work every time (because some functions are not formally integrable) but it is worth a try before taking out your computer.

Again, we give only the general form of the method. It is the role of teachers in schools to train, and train, and train students to understand and master these techniques. In addition, the sections in this textbook that treats of exact sciences (physics, computer science, astrophysics, chemistry, ...) replete with examples using this technique and thus serve implicitly as exercises.

Consider we want to calculate the integral (not bounded for the moment):

$$\int f(x)\mathrm{d}x \tag{6.3.354}$$

although we do not know directly how to calculate the primitive of this function $f(x)$ (at least we imagine being in that situation) we know (in one way or another) that exists (we do treat of improper integrals at this level).

The technique then consists in this integral to perform the following change of variable:

$$x=\varphi(t) \tag{6.3.355}$$

where $\varphi(t)$ is a continuous function and also its derivative, and having an inverse function. Then $\mathrm{d}x=\varphi'(t)\mathrm{d}t$, and let us prove now that in this case the equality:

$$\int f(x)\mathrm{d}x=\int f(\varphi(t))\varphi'(t)\mathrm{d}t \tag{6.3.356}$$

is satisfied.

Proof 6.14.6. We imply here that the variable t will be replaced after integration of the right member by its expression in function of x. To justify this equality in this sense, it suffices to show that the considered amounts where each is a defined only with a difference of a given arbitrary constant have the same derivative with respect to x. The derivative of the left member is:

$$\left(\int f(x)\mathrm{d}x\right)'=f(x) \tag{6.3.357}$$

We derive only the right member with respect to x considering that t is a function of x. We know that:

$$\frac{\mathrm{d}x}{\mathrm{d}t}=\varphi'(t) \tag{6.3.358}$$

We get therefore:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\left(\int f(\varphi(t))\varphi'(t)\mathrm{d}t\right)&=\frac{d}{\mathrm{d}t}\left(\int f(\varphi(t))\varphi'(t)\mathrm{d}t\right)\frac{\mathrm{d}t}{\mathrm{d}x}\\&=f(\varphi(t))\varphi'(t)\frac{1}{\varphi'(t)}\\&=f(\varphi(t))\\&=f(x)\end{aligned} \tag{6.3.359}$$

■ Q.E.D.

Obviously, the function $x=\varphi(t)$ must be chosen so that we know how to calculate the indefinite integral on the right side of equality.

Remark

Sometimes it is preferable to choose the change of variable in the form $t=\varphi(x)$ instead of $x=\psi(t)$ because as this to a large tendency to simplify the length of the equation instead of making it longer.

It is obvious that this theorem will be more explicitly written:

$$\boxed{\int\limits_{\varphi(a)}^{\varphi(b)} f(x)\mathrm{d}x=\int\limits_{a}^{b} f(\varphi(t))\varphi'(t)\mathrm{d}t} \tag{6.3.360}$$

6.3.2.4.1 Jacobian

Consider a domain D of the plane u, v delimited by a curve L. Suppose that the x, y coordinates are functions of the new variables u, v (always in the context of a change of variables!) by the following relations:

$$\begin{gathered} x=\varphi(u,v) \quad y=\phi(u,v) \\ (x,y)\in D \underset{\varphi,\phi}{\longrightarrow} (\varphi,\phi)\in D' \end{gathered} \tag{6.3.361}$$

where the functions $\varphi(x,y)$ and $\phi(u,v)$ are unique, continuous and have continuous derivatives in a given domain D' which we will define later. It corresponds therefore following previous relations to any pair of values u, v only one couple of values x, y and vice versa.

It follows from what was said just before that at any point $P(x,y)$ of the plane Oxy corresponds univocally a point $P'(u,v)$ of the plane Ouv of coordinates u, v defined by the above relations. The numbers v and u will be named "curvilinear coordinates" of P and we will see concrete and schematics examples of these in the section of Vector Calculus.

If in the Oxy plane the set of points P describes a closed curve L defining a domain D, the corresponding set of points describes in Ouv a given domain D'. It corresponds to any point of D' a point of D. Thus, the relations of transformation establish a bi-univocal correspondence between the points of the domains D and D'.

Now consider in D' a straight line of equation $u=c^{te}$. In general, the relations of transformation make it corresponds in the plane Oxy a curved line (or vice versa). Therefore, let us cut the domain D' by the multiple straight lines of equations $u=c^{te}$ and $v=c^{te}$ into small rectangular areas (we will not take into account at the limit, the rectangles on the boundary of D'). The corresponding curves of the domain D then cut this latter it into quadrilateral (defined by curves therefore). Obviously, the reverse applies!

Consider in the plane Ouv the rectangle $\Delta s'$ limited by the straight lines:

$$\begin{gathered} u=c^{te}, u+\Delta u=c^{te} \\ v=c^{te}, v+\Delta v=v^{te} \end{gathered} \tag{6.3.362}$$

and the curvilinear quadrilateral corresponding Δs in the plane Oxy. We will designate areas of these partial domains also by $\Delta s'$ by Δs. We have obviously:

$$\Delta s' = \Delta u \Delta v \tag{6.3.363}$$

The areas Δs and $\Delta s'$ are different in general.

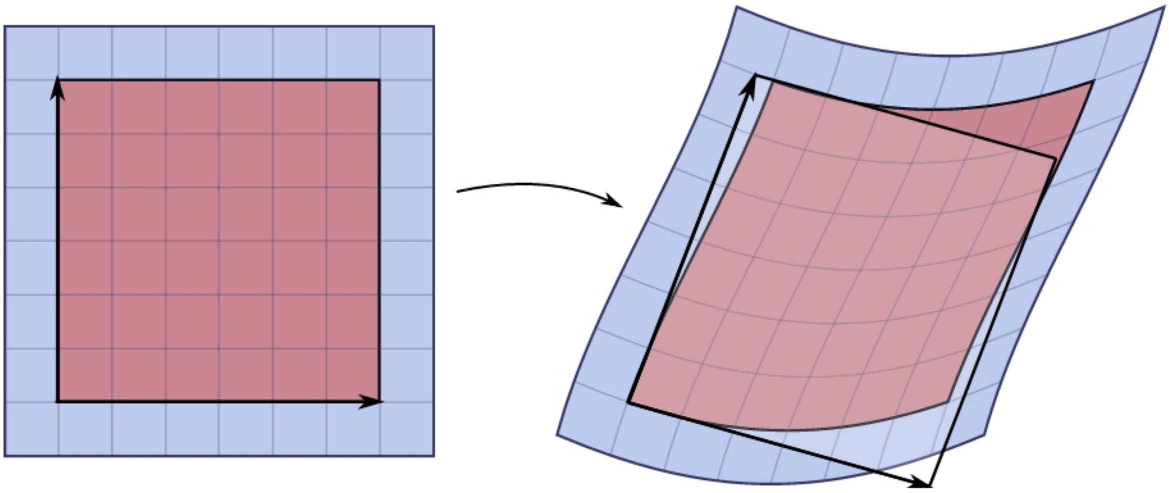


Figure 6.55: **A nonlinear map $f:\mathbb{R}^2 \mapsto \mathbb{R}^2$ sends a small square to a distorted parallelogram**

Suppose in D a continuous function $z=f(x,y)$. It corresponds to any value of this function of the domain D the same value $z=F(u,v)$ (what we want to check) in D', where:

$$F(u,v)=f(x,y)=f\left(\varphi(u,v),\phi(u,v)\right) \tag{6.3.364}$$

Consider the integrals sum of the function z in the domain D. We obviously have the following equality:

$$\sum f(x,y)\Delta s=\sum F(u,v)\Delta s \tag{6.3.365}$$

Let us calculate Δs, that is to say the area of the curvilinear quadrilateral P_1,P_2,P_3,P_4 in the plane Oxy:

We determine the coordinates of its vertices:

$$\begin{array}{lll}
P_1(x_1,y_1) & x_1=\varphi(u,v) & y_1=\phi(u,v)\\
P_2(x_2,y_2) & x_2=\varphi(u+\Delta u,v) & y_2=\phi(u+\Delta u,v)\\
P_3(x_3,y_3) & x_3=\varphi(u+\Delta u,v+\Delta v) & y_3=\phi(u+\Delta u,v+\Delta v)\\
P_4(x_4,y_4) & x_4=\varphi(u,v+\Delta v) & y_4=\phi(u,v+\Delta v)
\end{array}$$

We will assimilate in the calculation of the are of the quadrilateral P_1,P_2,P_3,P_4 the arcs $P_1P_2,P_2P_3,P_3P_4,P_4P_1$ to parallel line segments. We will also replace the increasing values of the functions by their differentials. This means that we ignore the infinitely small differentials of higher order than Δu and Δv. The previous relations then become:

$$\begin{array}{ll}
x_1=\varphi(u,v) & y_1=\phi(u,v)\\
x_2=\varphi(u,v)+\dfrac{\partial\varphi}{\partial u}\Delta u & y_2=\phi(u,v)+\dfrac{\partial\phi}{\partial u}\Delta u\\
x_3=\varphi(u,v)+\dfrac{\partial\varphi}{\partial u}\Delta u+\dfrac{\partial\varphi}{\partial v}\Delta v & y_3=\phi(u,v)+\dfrac{\partial\phi}{\partial u}\Delta u+\dfrac{\partial\phi}{\partial v}\Delta v\\
x_4=\varphi(u,v)+\dfrac{\partial\varphi}{\partial v}\Delta v & y_4=\phi(u,v)+\dfrac{\partial\phi}{\partial v}\Delta v
\end{array} \tag{6.3.366}$$

Under these assumptions, the curvilinear quadrilateral $P_1P_2P_3P_4$ can be likened to a parallelogram. His area is therefore approximately equal to twice the area of the triangle $P_1P_2P_3$, area that we can calculate by using the properties of the determinant (as we will prove it in the section on Linear Algebra, in $\mathbb{R}^2$ the determinant represents a parallelogram while in $\mathbb{R}^3$ it represents the volume of a parallelepiped):

$$\begin{aligned}\Delta s &\cong |\det(\vec{v}_1,\vec{v}_2)| = |\det(M)|\\ &= \left|\det\begin{bmatrix} a_{11} & a_{12}\\ a_{21} & a_{22}\end{bmatrix}\right| = |a_{11}a_{22}-a_{21}a_{12}|\end{aligned} \tag{6.3.367}$$

Such as (this is here that the best choice has to be done so that the final expression is the simplest and most aesthetic, for this purpose we proceed by trials and finally we make the choice below):

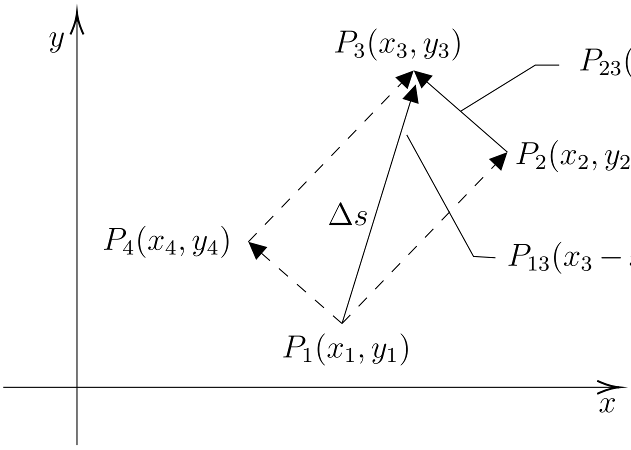


Figure 6.56: **Graphical representation of the determinant**

Thus we have:

$$\begin{aligned}\Delta s &\cong |\det(\vec{v}_1,\vec{v}_2)| = |D(\vec{P}_{13},\vec{P}_{23})|\\ &= \left|\det\begin{bmatrix}(x_3-x_1) & (x_3-x_2)\\ (y_3-y_1) & (y_3-y_2)\end{bmatrix}\right|\\ &= |(x_3-x_1)(y_3-y_2)-(x_3-x_2)(y_3-y_1)|\\ &= \left|\left(\frac{\partial\varphi}{\partial u}\Delta u+\frac{\partial\varphi}{\partial v}\Delta v\right)\frac{\partial\phi}{\partial v}\Delta v-\left(\frac{\partial\phi}{\partial u}\Delta u+\frac{\partial\phi}{\partial v}\Delta v\right)\frac{\partial\varphi}{\partial v}\Delta v\right|\\ &= \left|\frac{\partial\varphi}{\partial u}\frac{\partial\phi}{\partial v}\Delta u\Delta v-\frac{\partial\varphi}{\partial v}\frac{\partial\phi}{\partial u}\Delta u\Delta v\right|\\ &= \left|\frac{\partial\varphi}{\partial u}\frac{\partial\phi}{\partial v}\Delta u\Delta v-\frac{\partial\varphi}{\partial v}\frac{\partial\phi}{\partial u}\Delta u\Delta v\right| = \left|\frac{\partial\varphi}{\partial u}\frac{\partial\phi}{\partial v}-\frac{\partial\varphi}{\partial v}\frac{\partial\phi}{\partial u}\right|\Delta u\Delta v\\ &= |\det(J)|\Delta u\Delta v\\ &= \left|\det\begin{bmatrix}\dfrac{\partial\varphi}{\partial u} & \dfrac{\partial\varphi}{\partial v}\\ \dfrac{\partial\phi}{\partial u} & \dfrac{\partial\phi}{\partial v}\end{bmatrix}\right|\Delta s'\end{aligned} \tag{6.3.368}$$

Therefore the following relation (containing what is usually named the "functional determinant"):

$$\Delta s\cong\left|\det\begin{bmatrix}\dfrac{\partial\varphi}{\partial u} & \dfrac{\partial\varphi}{\partial v}\\ \dfrac{\partial\phi}{\partial u} & \dfrac{\partial\phi}{\partial v}\end{bmatrix}\right|\Delta s' \tag{6.3.369}$$

with:

$$J=\begin{bmatrix}\dfrac{\partial\varphi}{\partial u} & \dfrac{\partial\varphi}{\partial v}\\ \dfrac{\partial\phi}{\partial u} & \dfrac{\partial\phi}{\partial v}\end{bmatrix} \tag{6.3.370}$$

that is the "Jacobian matrix" (its determinant is simply named the "Jacobian" (for short)) of the coordinate transformation $\mathbb{R}^2\rightarrow\mathbb{R}^3$. By applying exactly the same reasoning to $\mathbb{R}^3$, then the Jacobian is written (by changing some notations because otherwise it becomes unreadable):

$$J=\begin{bmatrix}\dfrac{\partial P}{\partial u} & \dfrac{\partial P}{\partial v} & \dfrac{\partial P}{\partial w}\\ \dfrac{\partial Q}{\partial u} & \dfrac{\partial Q}{\partial v} & \dfrac{\partial Q}{\partial w}\\ \dfrac{\partial R}{\partial u} & \dfrac{\partial R}{\partial v} & \dfrac{\partial R}{\partial w}\end{bmatrix} \tag{6.3.371}$$

In short, what it is useful exactly? Well let us come back to our relation:

$$\Delta s\cong|\det(J)|\Delta s' \tag{6.3.372}$$

which is finally only an approximation because in the calculations of the area Δs we neglected the infinitely small differential of higher order. However, more the dimensions of the elementary domains Δs and $\Delta s'$ are small, and more we are approaching the true equality. The equality finally taking place when we go to the limit (finally also in math we make approximations... eh!), the surfaces of the elementary domains tending to zero:

$$|\det(J)|=\lim_{\Delta s\rightarrow 0}\frac{\Delta s}{\Delta s'} \tag{6.3.373}$$

We apply now the equality obtained to calculate the double integral (we can do the same with the triple of course). So we can finally write (this is the only way of putting the thing that makes sense):

$$\begin{gathered}\sum f(x,y)\Delta s=\sum F(u,v)\Delta s\\ \Downarrow\\ \sum f(x,y)\Delta s\cong\sum F(u,v)|\det(J)|\Delta s'\end{gathered} \tag{6.3.374}$$

Passing to the limit, we get the strict equality:

$$\boxed{\iint_D f(x,y)\mathrm{d}x\mathrm{d}y=\iint_{D'} F(u,v)|\det(J)|\mathrm{d}u\mathrm{d}v} \tag{6.3.375}$$

This is the coordinate transformation relation in a double integral! It allows us to reduce the calculation of a double integral in the domain D to domain D', which can simplify the problem. Similarly, for a triple integral, we write:

$$\boxed{\iiint_D f(x,y,z)\mathrm{d}x\mathrm{d}y\mathrm{d}z=\iiint_{D'} F(u,v,w)|\det(J)|\mathrm{d}u\mathrm{d}v\mathrm{d}w} \tag{6.3.376}$$

✍**Examples:**

Let us now determine the Jacobian for the most common coordinate systems (we send the reader back again to Vector Calculus section for more information about these systems):

E1. Polar coordinates $x=r\cos(\phi),y=r\sin(\phi)$:

$$\left|\det\begin{bmatrix}\dfrac{\partial x}{\partial r} & \dfrac{\partial x}{\partial \phi}\\ \dfrac{\partial y}{\partial r} & \dfrac{\partial y}{\partial \phi}\end{bmatrix}\right|=\left|\det\begin{bmatrix}\cos(\phi) & -r\sin(\phi)\\ \sin(\phi) & r\cos(\phi)\end{bmatrix}\right|=|r| \tag{6.3.377}$$

Since r^2 is always positive, we simply write:

$$|\det(J)|=r \tag{6.3.378}$$

E2. In cylindrical coordinates $x=r\cos(\phi),y=r\sin(\phi),z=z$ (see section Linear Algebra page 1961 for the calculation of such a determinant):

$$\begin{aligned}\left|\det\begin{bmatrix}\dfrac{\partial x}{\partial r} & \dfrac{\partial x}{\partial \phi} & \dfrac{\partial x}{\partial z}\\ \dfrac{\partial y}{\partial r} & \dfrac{\partial y}{\partial \phi} & \dfrac{\partial y}{\partial z}\\ \dfrac{\partial z}{\partial r} & \dfrac{\partial z}{\partial \phi} & \dfrac{\partial z}{\partial z}\end{bmatrix}\right| &= \left|\det\begin{bmatrix}\cos(\phi) & -r\sin(\phi) & 0\\ \sin(\phi) & r\cos(\phi) & 0\\ 0 & 0 & 1\end{bmatrix}\right|\\ &=\left|\begin{bmatrix}\cos(\phi)r\cos(\phi)\cdot 1-\cos(\phi)\cdot 0\cdot 0\\ +r\sin(\phi)\sin(\phi)\cdot 1-r\sin(\phi)\cdot 0\cdot 0\\ +0\cdot\sin(\phi)\cdot 0-1\cdot r\cos(\phi)\cdot 0\end{bmatrix}\right|\\ &=|r|\end{aligned} \tag{6.3.379}$$

Since r is always positive, we simply write:

$$|\det(J)|=r \tag{6.3.380}$$

E3. In spherical coordinates $x=r\sin(\theta)\cos(\phi),y=r\sin(\theta)\sin(\phi),z=r\cos(\theta)$ (see section Linear Algebra page 1961 for the calculation of such a determinant):

$$\begin{aligned}\left|\det\begin{bmatrix}\dfrac{\partial x}{\partial r} & \dfrac{\partial x}{\partial \phi} & \dfrac{\partial x}{\partial \theta}\\ \dfrac{\partial y}{\partial r} & \dfrac{\partial y}{\partial \phi} & \dfrac{\partial y}{\partial \theta}\\ \dfrac{\partial z}{\partial r} & \dfrac{\partial z}{\partial \phi} & \dfrac{\partial z}{\partial \theta}\end{bmatrix}\right| &= \left|\det\begin{bmatrix}\sin(\theta)\cos(\phi) & -r\sin(\theta)\sin(\phi) & r\cos(\theta)\cos(\phi)\\ \sin(\theta)\sin(\phi) & r\sin(\theta)\cos(\phi) & r\cos(\theta)\sin(\phi)\\ \cos(\theta) & 0 & -r\sin(\theta)\end{bmatrix}\right|\\ &=\left|\begin{matrix}-\sin(\theta)\cos(\phi)r\sin(\theta)\cos(\phi)r\sin(\theta)-\sin(\theta)\cos(\phi)r\cos(\theta)\sin(\phi)\cdot 0\\ -r\sin(\theta)\sin(\phi)\sin(\theta)\sin(\phi)r\sin(\theta)-r\sin(\theta)\sin(\phi)r\cos(\theta)\sin(\phi)\cos(\theta)\\ +r\cos(\theta)\cos(\phi)\sin(\theta)\sin(\phi)\cdot 0+r\cos(\theta)\cos(\phi)r\sin(\theta)\cos(\phi)r\sin(\theta)\end{matrix}\right|\\ &=\left|\begin{matrix}-(\sin^3(\theta)\cos^2(\phi)r^2+\sin^3(\theta)\sin^2(\phi)r^2)\\ -(r^2\sin(\theta)\sin^2(\phi)\cos^2(\theta)+r^2\cos^2(\theta)\cos^2(\phi)\sin(\theta))\end{matrix}\right|\\ &=|-\sin^3(\theta)r^2-r^2\sin(\theta)\cos^2(\theta)|\\ &=|-r^2\sin(\theta)(\sin^2(\theta)+\cos^2(\theta))|\\ &=|-r^2\sin(\theta)|\end{aligned} \tag{6.3.381}$$

Since r is always positive, we simply write:

$$|\det(J)| = r^2 \sin(\theta) \quad \text{with} \quad \sin(\theta) \geq 0 \tag{6.3.382}$$

The definition of differentiability in multivariable calculus is a bit technical. There are subtleties to watch out for, as one has to remember the existence of the derivative is a more stringent condition than the existence of partial derivatives. But, in the end, if our function is nice enough so that it is differentiable, then the derivative itself isn't too complicated. It's a fairly straightforward generalization of the single variable derivative.

In single variable calculus, you learned that the derivative of a function $f:\mathbb{R}\rightarrow\mathbb{R}$ at a single point is just a real number, the rate of increase of the function (i.e., slope of the graph) at that point. We could think of that number as a 1×1 matrix, so if we like, we could denote the derivative of $f(x)$ at $x=a$ as:

$$\mathrm{D}f(a) = \left[\dfrac{\mathrm{d}f}{\mathrm{d}x}(a)\right] \tag{6.3.383}$$

For a scalar-valued function of multiple variables, such as $f(x,y)$ or $f(x,y,z)$, we can think of the partial derivatives as the rates of increase of the function in the coordinate directions. If the function is differentiable, then the derivative is simply a row matrix containing all of these partial derivatives, which we name the "matrix of partial derivatives" (that corresponds to the Jacobian matrix). For $f:\mathbb{R}^n\rightarrow\mathbb{R}$, viewed as a $f(\vec{x})$, where $\vec{x}=(x_1,x_2,\ldots,x_n)$, the $1\times n$ matrix of partial derivatives at $\vec{x}=\vec{a}$ is:

$$\mathrm{D}f(\mathbf{a}) = \left[\dfrac{\partial f}{\partial x_1}(\vec{a}), \dfrac{\partial f}{\partial x_2}(\vec{a}), \ldots, \dfrac{\partial f}{\partial x_n}(\vec{a})\right] \tag{6.3.384}$$

The last generalization is to vector-valued functions, $f:\mathbb{R}^n\rightarrow\mathbb{R}^m$. Here, $f(\vec{x})$ is a function of the vector $\vec{x}=(x_1,x_2,\ldots,x_n)$ whose output is a vector of m components. We could write f in terms of its components as:

$$f(\vec{x}) = (f_1(\vec{x}), f_2(\vec{x}), \ldots, f_m(\vec{x})) = \begin{bmatrix} f_1(\vec{x}) \\ f_2(\vec{x}) \\ \vdots \\ f_m(\vec{x}) \end{bmatrix} \tag{6.3.385}$$

(Recall that when we view vectors as matrices, we view them as column matrices, so the components are stacked up on top of each other.)

To form the matrix of partial derivatives, we think of $f(\vec{x})$ as column matrix, where each component is a scalar-valued function. The matrix of partial derivatives of each component $f_i(\vec{x})$ would be a $1\times n$ row matrix, as above. We just stack these row matrices on top of each other to form a larger matrix. We get that the full $m\times n$ matrix of partial derivatives at $\vec{x}=\vec{a}$ is:

$$\mathrm{D}f(\vec{a}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\vec{a}) & \frac{\partial f_1}{\partial x_2}(\vec{a}) & \ldots & \frac{\partial f_1}{\partial x_n}(\vec{a}) \\ \frac{\partial f_2}{\partial x_1}(\vec{a}) & \frac{\partial f_2}{\partial x_2}(\vec{a}) & \ldots & \frac{\partial f_2}{\partial x_n}(\vec{a}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\vec{a}) & \frac{\partial f_m}{\partial x_2}(\vec{a}) & \ldots & \frac{\partial f_m}{\partial x_n}(\vec{a}) \end{bmatrix} \tag{6.3.386}$$

Though we should probably refer to the derivative of f as the linear transformation that is associated with the matrix $\mathrm{D}f(\vec{a})$, it's fine at this level to refer to the "matrix of partial derivatives" $Df(\vec{a})$ as "the derivative" of f at the point $\vec{a}$ (assuming that f is differentiable at $\vec{a}$, of course).

✍**Example:**

Let $f(x,y)=x^2+y^2$. We want to find $\mathrm{D}f(1,2)$ and the equation for the tangent plane (using the multivariate Taylor development as seen at page 1714) at $(x,y)=(1,2)$.

We have first:

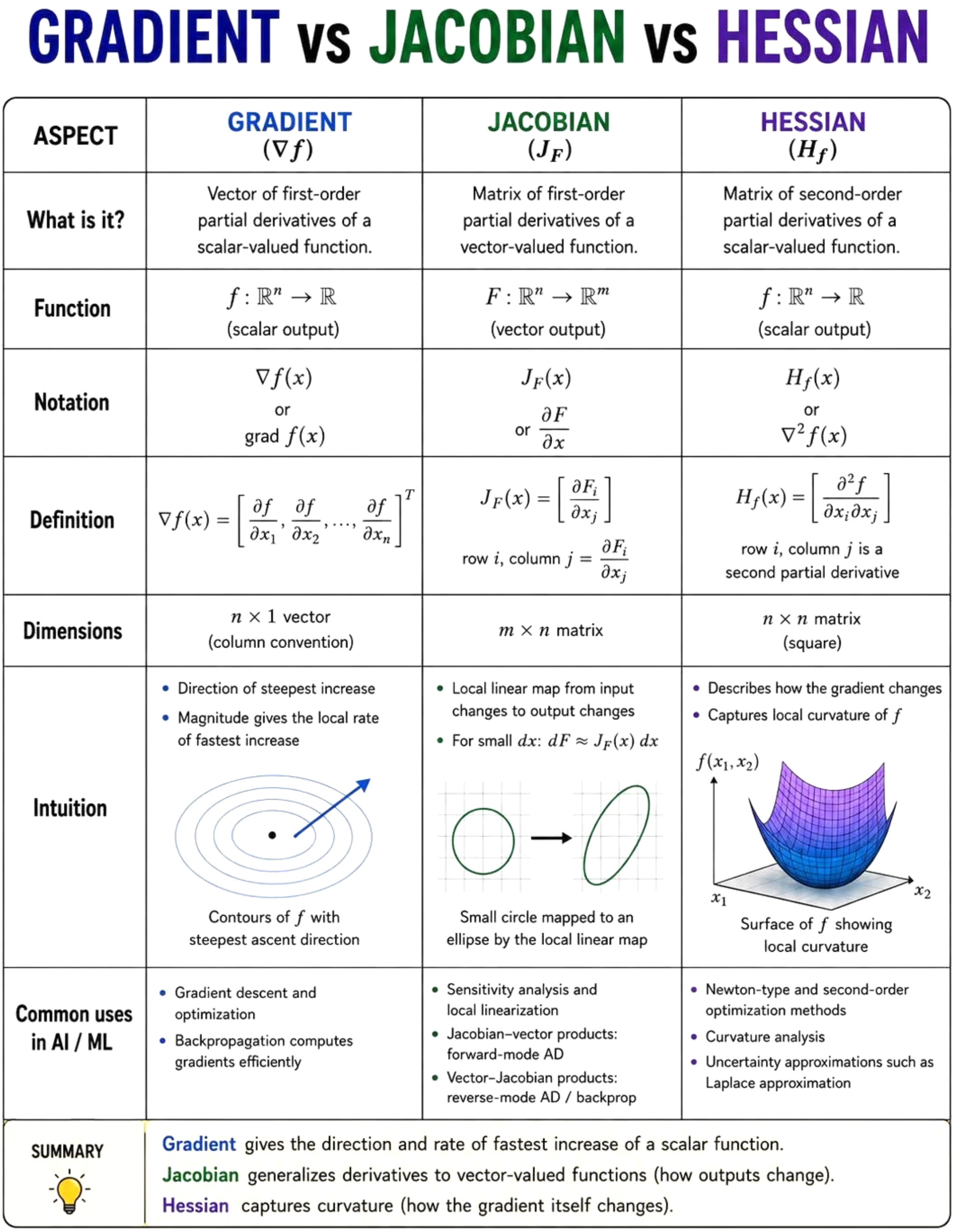
$$\begin{aligned}
&\frac{\partial f}{\partial x}(x,y)=2x\\
&\frac{\partial f}{\partial x}(1,2)=2\\
&\frac{\partial f}{\partial y}(x,y)=2y\\
&\frac{\partial f}{\partial y}(1,2)=4
\end{aligned}\tag{6.3.387}$$

So $\mathrm{D}f(1,2)=[2 \quad 4]$.

Since both partial derivatives $\frac{\partial f}{\partial x}(x,y)$ and $\frac{\partial f}{\partial y}(x,y)$ are continuous functions, we know that $f(x,y)$ is differentiable. Therefore, $\mathrm{D}f(1,2)$ is the derivative of f, and the function has a tangent plane there.

To calculate the equation of the tangent plane, the only additional calculation is the value of f at $(x,y)=(1,2)$, which is $f(1,2)=1^2+2^2=5$. The equation for the tangent plane is therefore (still using Taylor multivariate approximation):

$$\begin{aligned}
z &\cong f(1,2)+\frac{\partial f}{\partial x}(1,2)(x-1)+\frac{\partial f}{\partial y}(1,2)(y-2)\\
&=5+2(x-1)+4(y-2)
\end{aligned}\tag{6.3.388}$$



GRADIENT vs JACOBIAN vs HESSIAN

ASPECT	**GRADIENT** (∇f)	**JACOBIAN** (J_F)	**HESSIAN** (H_f)
What is it?	Vector of first-order partial derivatives of a scalar-valued function.	Matrix of first-order partial derivatives of a vector-valued function.	Matrix of second-order partial derivatives of a scalar-valued function.
Function	$f : \mathbb{R}^n \to \mathbb{R}$ (scalar output)	$F : \mathbb{R}^n \to \mathbb{R}^m$ (vector output)	$f : \mathbb{R}^n \to \mathbb{R}$ (scalar output)
Notation	$\nabla f(x)$ or grad $f(x)$	$J_F(x)$ or $\frac{\partial F}{\partial x}$	$H_f(x)$ or $\nabla^2 f(x)$
Definition	$\nabla f(x) = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n} \right]^T$	$J_F(x) = \left[\frac{\partial F_i}{\partial x_j} \right]$ row i, column $j = \frac{\partial F_i}{\partial x_j}$	$H_f(x) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j} \right]$ row i, column j is a second partial derivative
Dimensions	$n \times 1$ vector (column convention)	$m \times n$ matrix	$n \times n$ matrix (square)
Intuition	- Direction of steepest increase - Magnitude gives the local rate of fastest increase Contours of f with steepest ascent direction	- Local linear map from input changes to output changes - For small dx: $dF \approx J_F(x)\, dx$ Small circle mapped to an ellipse by the local linear map	- Describes how the gradient changes - Captures local curvature of f Surface of f showing local curvature
Common uses in AI / ML	- Gradient descent and optimization - Backpropagation computes gradients efficiently	- Sensitivity analysis and local linearization - Jacobian–vector products: forward-mode AD - Vector–Jacobian products: reverse-mode AD / backprop	- Newton-type and second-order optimization methods - Curvature analysis - Uncertainty approximations such as Laplace approximation

SUMMARY

Gradient gives the direction and rate of fastest increase of a scalar function.
Jacobian generalizes derivatives to vector-valued functions (how outputs change).
Hessian captures curvature (how the gradient itself changes).

6.3.2.5 Integration by Parts

When we seek to make integrations, it is very common that we have to use a tool (or calculation method) named "integration by parts". There are different degrees of use of this tool and we'll start with the most simple and that is the most used in all chapters and sections in this textbook.

First we start from the derivative of the product of two functions proved above:

$$(f(t)g(t))' = f'(t)g(t) + f(t)g'(t) \tag{6.3.389}$$

so we have:

$$f(t)g'(t) = (f(t)g(t))' - f'(t)g(t) \tag{6.3.390}$$

and we get:

$$\int\limits_a^b f(t)g'(t)\mathrm{d}t = \int\limits_a^b \left[(f(t)g(t))' - f'(t)g(t)\right]\mathrm{d}t = \int\limits_a^b (f(t)g(t))'\mathrm{d}t - \int\limits_a^b f'(t)g(t)\mathrm{d}t \tag{6.3.391}$$

after a final simplification we finally get the famous very important equality:

$$\boxed{\int\limits_a^b f(t)g'(t)\mathrm{d}t = f(t)g(t)\Big|_a^b - \int\limits_a^b f'(t)g(t)\mathrm{d}t} \tag{6.3.392}$$

But sometimes we will need the generalization of that relation. We can show that if f and g are two applications (functions) of class $\mathcal{C}^n$ (n times differentiable) on $[a,b]$ in $\mathbb{C}$, then:

$$\int\limits_a^b f(t)g^{(n)}(t)\mathrm{d}t = \left[\sum_{k=0}^{n-1}(-1)^k f^{(k)}\cdot g^{(n-1-k)}\right]_a^b + (-1)^n \int\limits_a^b f^{(n)}(t)g(t)\mathrm{d}t \tag{6.3.393}$$

Proof 6.14.7. Let us proceed by induction on n (beware it is not necessarily easy to understand as often with demonstrations by induction!).

Knowing that the relation is true for $n=1$, we assume it true for n (as given in the above relation!) and we prove it for $n+1$ (so we must laid on the previous relation but with $n+1$ instead of n):

$$\begin{aligned}\int\limits_a^b f(t)\cdot g^{(n+1)}(t)\mathrm{d}t &= f(t)g^{(n)}(t)\Big|_a^b - \int\limits_a^b f'(t)g^{(n)}(t)\mathrm{d}t = f(t)g^{(n)}(t)\Big|_a^b - \left[\sum_{k=0}^{n-1}(-1)^k f^{(k+1)}g^{(n-1-k)}\right]_a^b \\ &\quad + (-1)^{n+1}\int\limits_a^b f^{(n+1)}(t)g(t)\mathrm{d}t \\ &= f(t)g^{(n)}(t)\Big|_a^b - \left[\sum_{k=0}^{n}(-1)^k f^{(k)}g^{(n-k)}\right]_a^b + (-1)^{n+1}\int\limits_a^b f^{(n+1)}(t)g(t)\mathrm{d}t \\ &= \left[\sum_{k=0}^{n}(-1)^k f^{(k)}g^{(n-k)}\right]_a^b + (-1)^{n+1}\int\limits_a^b f^{(n+1)}(t)g(t)\mathrm{d}t\end{aligned} \tag{6.3.394}$$

Remark

The trick (proposed by a reader) in this proof is to see that $-(-1)^n$ gives a minus sign when n is even and plus sign when n is odd and also that $+(-1)^{n+1}$ gives a minus sing when n is even and plus sign when n is odd.

For $n=1$ we fall back on the well known and very often used equality in this textbook:

$$\int\limits_a^b f(t)g'(t)\mathrm{d}t = f(t)g(t)\Big|_a^b - \int\limits_a^b f'(t)g(t)\mathrm{d}t \tag{6.3.395}$$

■ Q.E.D.

6.3.2.6 Usual Primitives

There are in math and physics many primitives or functions defined on integrals we see quite frequently (but not exclusively). Furthermore, all primitive proved below will be used in the various chapters on Mechanics, Engineering, Atomistic, Social Mathematics, etc of this textbook. So, as in any formula booklet, we propose you the most sixty knows primitives but with... the proofs!

> Caution! Because the fundamental theorem of calculus proved earlier above (page 1509) says that the integral of a function is the same as its antiderivative, we will not derive below the integrals of some functions for which the antiderivative can already be found in our list of usual derivatives that can be found at page 1475. Especially since the direct calculation of the integral of some simple functions (like the integral of $1/x$) is unnecessarily laborious (even if it is sometimes considered funny by some readers...).

So as just mentioned will omit the primitives that can be immediately deduced from the usual derivatives we have proved earlier. This means for example that we assume known the following very important primitives (certainly the most used in all pages of this textbook):

$$\begin{aligned}
&\int \frac{1}{x}\mathrm{d}x = \ln(x) + c^{te}\\
&\int \sin(x)\mathrm{d}x = -\cos(x) + c^{te}\\
&\int \cos(x)\mathrm{d}x = \sin(x) + c^{te}\\
&\int e^x\mathrm{d}x = e^x + c^{te}\\
&\ldots
\end{aligned} \tag{6.3.396}$$

Otherwise here is the list of the most common primitives (the reader will meet anyway many others - developed in the details - as he reads this textbook):

1. Primitive of $f(x) = \tan(x)$:

   By definition we have:

   $$\int \tan(x)\mathrm{d}x = \int \frac{\sin(x)}{\cos(x)}\mathrm{d}x \tag{6.3.397}$$

   We use the change of variable $u = \cos(x)$, $\mathrm{d}u = -\sin(x)\mathrm{d}x$ and therefore:

   $$\int \frac{\sin(x)}{\cos(x)}\mathrm{d}x = -\int \frac{1}{u}\mathrm{d}u = -\ln|u| = -\ln|\cos(x)| \tag{6.3.398}$$

   Therefore:

   $$F(x) = -\ln|\cos(x)| \tag{6.3.399}$$

2. Primitive of $f(x)=\cot(x)$:

   By definition we have:

$$\int \cot(x)\mathrm{d}x=\int \frac{\cos(x)}{\sin(x)}\mathrm{d}x \tag{6.3.400}$$

   We use the change of variable $u=\sin(x)$, $\mathrm{d}u=\cos(x)\mathrm{d}x$ and therefore:

$$\int \frac{\cos(x)}{\sin(x)}\mathrm{d}x=\int \frac{1}{u}\mathrm{d}u=\ln|u|=\ln|\sin(x)| \tag{6.3.401}$$

   Therefore:

$$F(x)=\ln|\sin(x)| \tag{6.3.402}$$

3. Primitive of $f(x)=\arcsin(x)$:

   We integrate by parts:

$$\int 1\cdot\arcsin(x)\mathrm{d}x=x\arcsin(x)-\int \frac{x}{\sqrt{1-x^2}}\mathrm{d}x \tag{6.3.403}$$

   If we put $u=1-x^2$, giving us $\mathrm{d}u=-2x\mathrm{d}x$, we get:

$$\int \frac{x}{\sqrt{1-x^2}}\mathrm{d}x=-\int \frac{1}{2\sqrt{u}}\mathrm{d}u=-\sqrt{u}=-\sqrt{1-x^2} \tag{6.3.404}$$

   Therefore:

$$F(x)=x\arcsin(x)+\sqrt{1-x^2} \tag{6.3.405}$$

4. Primitive of $f(x)=\arccos(x)$:

   We integrate by parts again:

$$\int 1\cdot\arccos(x)\mathrm{d}x=x\arccos(x)+\int \frac{x}{\sqrt{1-x^2}}\mathrm{d}x \tag{6.3.406}$$

   If we put $u=1-x^2$, giving us $\mathrm{d}u=-2x\mathrm{d}x$, we get:

$$\int \frac{x}{\sqrt{1-x^2}}\mathrm{d}x=-\int \frac{1}{2\sqrt{u}}\mathrm{d}u=-\sqrt{u}=-\sqrt{1-x^2} \tag{6.3.407}$$

   Therefore:

$$F(x)=x\arccos(x)-\sqrt{1-x^2} \tag{6.3.408}$$

5. Primitive of $f(x)=\arctan(x)$:

   We integrate by parts again:

$$\int 1\cdot\arctan(x)\mathrm{d}x=x\arctan(x)-\int \frac{x}{\sqrt{1+x^2}}\mathrm{d}x \tag{6.3.409}$$

   If we put $u=1+x^2$, giving us $\mathrm{d}u=2x\mathrm{d}x$, we get:

$$\int \frac{x}{\sqrt{1+x^2}}\mathrm{d}x=\int \frac{1}{2u}\mathrm{d}u=\frac{1}{2}\ln|u|=\frac{1}{2}\ln(1+x^2) \tag{6.3.410}$$

   Therefore:

$$F(x)=x\arctan(x)-\frac{1}{2}\ln(1+x^2) \tag{6.3.411}$$

6. Primitive of $f(x)=\text{arccot}(x)$:

   Once again... we integrate by part:

$$\int 1\cdot\text{arccot}(x)\mathrm{d}x=x\mathrm{d}(x)+\int\frac{x}{\sqrt{1+x^2}}\mathrm{d}x \tag{6.3.412}$$

   If we put $u=1+x^2$, giving us $\mathrm{d}u=2x\mathrm{d}x$, we get:

$$\int\frac{x}{\sqrt{1+x^2}}\mathrm{d}x=\int\frac{1}{2u}\mathrm{d}u=\frac{1}{2}\ln(|u|)=\frac{1}{2}\ln(1+x^2) \tag{6.3.413}$$

   Therefore:

$$F(x)=x\text{arccot}(x)+\frac{1}{2}\ln(1+x^2) \tag{6.3.414}$$

7. Primitive of $f(x)=xe^{ax}$ with $a\in\mathbb{R}\{0\}$:

   An integration by part gives:

$$\int xe^{ax}\mathrm{d}x=\frac{x}{a}e^{ax}-\frac{1}{a}\int e^{ax}\mathrm{d}x=\frac{x}{a}e^{ax}-\frac{1}{a^2}e^{ax}=e^{ax}\left(\frac{1}{a}x-\frac{1}{a^2}\right) \tag{6.3.415}$$

   Therefore:

$$F(x)=e^{ax}\left(\frac{1}{a}x-\frac{1}{a^2}\right) \tag{6.3.416}$$

   The generalization is in our point of view immediate and we get the following reduction formula:

$$\int x^ne^{ax}\mathrm{d}x=\frac{x^ne^{ax}}{a}-\frac{n}{a}\int x^{n-1}a^{ax}\mathrm{d}x \tag{6.3.417}$$

Remark

Another very important primitive with the exponential in physics is that we had proved in our study of the law of Gauss-Laplace law (Normal Law) in the section of Statistics (determination of the expected mean).

8. Primitive of $f(x)=\ln(x)$:

   We write:

$$\int\ln(x)\mathrm{d}x=\int 1\cdot\ln(x)\mathrm{d}x \tag{6.3.418}$$

   Integrating by parts we found:

$$\int 1\cdot\ln(x)\mathrm{d}x=x\ln(x)-\int x\frac{1}{x}\mathrm{d}x=x\ln(x)-x=x(\ln(x)-1) \tag{6.3.419}$$

   Finally:

$$F(x)=x(\ln(x)-1) \tag{6.3.420}$$

9. Primitive of $f(x)=x\ln(ax)$ with $a\in\mathbb{R}\{0\}$:

   An integration by part give us:

$$\int x\ln(ax)\mathrm{d}x=\frac{1}{2}x^2\ln(ax)-\frac{1}{2}\int x^2\frac{1}{x}\mathrm{d}x=\frac{1}{2}x^2\ln(ax)-\frac{1}{4}x^2 \tag{6.3.421}$$

   Therefore:

$$F(x)=\frac{1}{2}x^2\ln(ax)-\frac{1}{4}x^2 \tag{6.3.422}$$

10. Primitive of $f(x)=a^x$ for $a>0,a\neq 1$:

    To begin we write:

$$f(x)=a^x\Leftrightarrow f'(x)=a^x\ln(a) \tag{6.3.423}$$

    Therefore we get:

$$\frac{f'(x)}{\ln(a)}=a^x\Rightarrow\int\frac{f'(x)}{\ln(a)}\mathrm{d}x=\int a^x\mathrm{d}x \tag{6.3.424}$$

    and:

$$\frac{1}{\ln(a)}\int f'(x)\mathrm{d}x=\frac{f(x)}{\ln(a)}\quad\text{and}\quad\int a^x\mathrm{d}x=\int f(x)\mathrm{d}x=F(x) \tag{6.3.425}$$

    Finally:

$$F(x)=\frac{f(x)}{\ln(a)}=\frac{a^x}{\ln(a)} \tag{6.3.426}$$

11. Primitive of $f(x)=\log_a(x)$:

    For $a>0,a\neq 1$ knowing that (see the properties of logarithms in the section of Functional Analysis):

$$\log_a(x)=\frac{\ln(x)}{\ln(a)}=\frac{1}{\ln(a)}\ln(x) \tag{6.3.427}$$

    we get using the primitive of $\ln(x)$:

$$F(x)=\frac{x}{\ln(a)}\big(\ln(x)-1\big) \tag{6.3.428}$$

12. Primitive of $f(x)=\tanh(x)$:

    We have:

$$\int\tanh(x)\mathrm{d}x=\int\frac{\sinh(x)}{\cosh(x)}\mathrm{d}x \tag{6.3.429}$$

    we use the change of variables $u=\cosh(x),\mathrm{d}u=\sinh(x)\mathrm{d}x$ and we get:

$$\int\frac{\sinh(x)}{\cosh(x)}\mathrm{d}x=\int\frac{1}{u}\mathrm{d}u=\ln(|u|)=\ln|\cosh(x)| \tag{6.3.430}$$

    Finally:

$$F(x)=\ln\big(|\cosh(x)|\big) \tag{6.3.431}$$

13. Primitive of $f(x)=\coth(x)$:

    We know we have:

    $$\int \coth(x)\mathrm{d}x=\int \frac{\cosh(x)}{\sinh(x)}\mathrm{d}x \tag{6.3.432}$$

    we use the change of variables $u=\sinh(x)$, $\mathrm{d}u=\cosh(x)\mathrm{d}x$ and we get:

    $$\int \frac{\cosh(x)}{\sinh(x)}\mathrm{d}x=\int \frac{1}{u}\mathrm{d}u=\ln|u|=\ln|\sinh(x)| \tag{6.3.433}$$

    Finally:

    $$F(x)=\ln\left(|\sinh(x)|\right) \tag{6.3.434}$$

14. Primitive of $f(x)=\text{arcsinh}(x)$:

    We integrate by parts:

    $$\int 1\cdot\text{arcsinh}(x)\mathrm{d}x=x\text{arcsinh}(x)-\int \frac{x}{\sqrt{1+x^2}}\mathrm{d}x \tag{6.3.435}$$

    If we put $u=1+x^2$, $\mathrm{d}u=2x\mathrm{d}x$ we get:

    $$\int \frac{x}{\sqrt{1+x^2}}\mathrm{d}x=\int \frac{1}{2\sqrt{u}}\mathrm{d}u=\sqrt{u}=\sqrt{1+x^2} \tag{6.3.436}$$

    Finally:

    $$F(x)=x\text{arcsinh}(x)-\sqrt{1+x^2} \tag{6.3.437}$$

15. Primitive of $f(x)=\text{arccosh}(x)$:

    If we integrate by parts as before:

    $$\int 1\cdot\text{arccosh}(x)\mathrm{d}x=x\cdot\text{arccosh}(x)-\int \frac{x}{\sqrt{x^2-1}}\mathrm{d}x \tag{6.3.438}$$

    If we put $u=x^2-1$, $\mathrm{d}u=2x\mathrm{d}x$ we get:

    $$\int \frac{x}{\sqrt{x^2-1}}\mathrm{d}x=\int \frac{1}{2\sqrt{u}}\mathrm{d}u=\sqrt{u}=\sqrt{x^2-1} \tag{6.3.439}$$

    Finally:

    $$F(x)=x\text{arccosh}(x)-\sqrt{x^2-1} \tag{6.3.440}$$

16. Primitive of $f(x)=\text{arctanh}(x)$:

    We integrate by parts:

    $$\int 1\cdot\text{arctanh}(x)\mathrm{d}x=x\text{arctanh}(x)-\int \frac{x}{1-x^2}\mathrm{d}x \tag{6.3.441}$$

    If we put $u=1-x^2$, $\mathrm{d}u=-2x\mathrm{d}x$ we get:

    $$\int \frac{x}{1-x^2}\mathrm{d}x=-\int \frac{1}{2u}\mathrm{d}u=-\frac{1}{2}\ln(|u|)=-\frac{1}{2}\ln\left(|1-x^2|\right) \tag{6.3.442}$$

    Finally:

    $$F(x)=x\text{arctanh}(x)+\frac{1}{2}\ln|1-x^2| \tag{6.3.443}$$

17. Primitive of $f(x)=\text{arccoth}(x)$:

    We integrate by parts:

$$\int 1\cdot\text{arccoth}(x)\mathrm{d}x=\text{arccoth}(x)-\int\frac{x}{1-x^2}\mathrm{d}x \tag{6.3.444}$$

    If we put $u=1-x^2, \mathrm{d}u=-2x\mathrm{d}x$ we get:

$$\int\frac{x}{1-x^2}\mathrm{d}x=-\int\frac{1}{2u}\mathrm{d}u=-\frac{1}{2}\ln(|u|)=-\frac{1}{2}\ln\left(|1-x^2|\right) \tag{6.3.445}$$

    Finally:

$$F(x)=x\cdot\text{arccoth}(x)+\frac{1}{2}\ln\left(|1-x^2|\right) \tag{6.3.446}$$

18. Primitive of $f(x)=\sin^n(x)$ with $n\geq 2$:

    Let us put $I_n=\int\sin^n(x)\mathrm{d}x$. An integration by parts give:

$$I_n=\int\sin(x)\cdot\sin^{n-1}(x)\mathrm{d}x=-\cos(x)\cdot\sin^{n-1}(x)+(n+1)\int\cos^2(x)\cdot\sin^{n-2}(x)\mathrm{d}x \tag{6.3.447}$$

    substituting $\cos^2(x)$ by $1-\sin^2 x$ in the last primitive, we get:

$$I_n=-\cos(x)\cdot\sin^{n-1}(x)+(n-1)(I_{n-2}-I_n) \tag{6.3.448}$$

    and therefore:

$$F(x)=I_n=\frac{1}{n}\left(-\cos(x)\cdot\sin^{n-1}(x)+(n-1)I_{n-2}\right) \tag{6.3.449}$$

    All primitives of the same form (recurrence relation) are named "reduction formulas".

19. Primitive of $f(x)=\cos^n(x)$ with $n\geq 2$:

    In this case we have the recursive formula:

$$\begin{aligned}F(x)&=\int\cos^n(x)\mathrm{d}x=\frac{1}{n}\left(\sin(x)\cdot\cos^{n-1}(x)+(n-1)\int\cos^{n-2}(x)\mathrm{d}x\right)\\&=\frac{1}{n}\left(\sin(x)\cdot\cos^{n-1}(x)+(n-1)I_{n-2}\right)\end{aligned} \tag{6.3.450}$$

    that is proved exactly in the same way as the previous recursive relation (the reader can request the details if required).

20. Primitive of $f(x)=\tan^2(x)$:

    Knowing that $\tan'(x)=1+\tan^2(x)$ we have:

$$\int\tan^2(x)\mathrm{d}x=\int\left(\tan'(x)-1\right)\mathrm{d}x=\tan(x)-x \tag{6.3.451}$$

    Therefore:

$$F(x)=\tan(x)-x \tag{6.3.452}$$

21. Primitive of $f(x)=\cot^2(x)$:

    Knowing that $\cot'(x)=-1-\cot^2(x)$ we have:

$$\int \cot^2(x)\mathrm{d}x=\int \big(-\cot'(x)-1\big)\mathrm{d}x=-\cot(x)-x \tag{6.3.453}$$

    Therefore:

$$F(x)=\int \cot^2(x)\mathrm{d}x=-\cot(x)-x \tag{6.3.454}$$

22. Primitive of $f(x)=\sin^{-2}(x)$:

    Using remarkable trigonometric identities (see section Trigonometry page 2519), we have:

$$\int \frac{1}{\sin^2(x)}\mathrm{d}x=\int \frac{\sin^2(x)+\cos^2(x)}{\sin^2(x)}\mathrm{d}x=\int \big(1+\cot^2(x)\big)\mathrm{d}x=-\cot(x) \tag{6.3.455}$$

    Thanks to the primitive of $\cot^2(x)$. Therefore:

$$F(x)=\int \frac{1}{\sin^2(x)}\mathrm{d}x=-\cot(x) \tag{6.3.456}$$

23. Primitive of $f(x)=\cos^{-2}(x)$:

    Using once again remarkable trigonometric identities (see section Trigonometry page 2519), we have:

$$\int \frac{1}{\cos^2(x)}\mathrm{d}x=\int \frac{\sin^2(x)+\cos^2(x)}{\cos^2(x)}\mathrm{d}x=\int \big(1+\tan^2(x)\big)\mathrm{d}x=\tan(x) \tag{6.3.457}$$

    Thanks to the primitive of $\tan^2(x)$. Therefore:

$$F(x)=\int \frac{1}{\cos^2(x)}\mathrm{d}x=\tan(x) \tag{6.3.458}$$

24. Primitive of $f(x)=\sin^{-1}(x)$:

    We use the substitution $x=2\arctan(t), t=\tan(x/2)$. Knowing that (see section Trigonometry page 2519):

$$\sin(x)=\frac{2\tan\left(\dfrac{x}{2}\right)}{1+\tan^2\left(\dfrac{x}{2}\right)} \tag{6.3.459}$$

    we then get:

$$\sin(x)=\frac{2t}{1+t^2} \quad \text{and} \quad \mathrm{d}x=\frac{2}{1+t^2}\mathrm{d}t \tag{6.3.460}$$

    Therefore:

$$\int \frac{1}{\sin(x)}\mathrm{d}x=\int \frac{1}{t}\mathrm{d}t=\ln(|t|)=\ln\left(\left|\tan\frac{x}{2}\right|\right) \tag{6.3.461}$$

    Finally:

$$F(x)=\int \frac{1}{\sin(x)}\mathrm{d}x=\ln\left(\left|\tan\frac{x}{2}\right|\right) \tag{6.3.462}$$

25. Primitive of $f(x)=\cos^{-1}(x)$:

    Knowing that $\cos(x)=\sin(x+\pi/2)$ (see section Trigonometry page 2514) we have:

$$\int\frac{1}{\cos(x)}\mathrm{d}x=\int\frac{1}{\sin\left(x+\frac{\pi}{2}\right)}\mathrm{d}x \tag{6.3.463}$$

    We do the change of variable $x+\pi/2=u$, $\mathrm{d}u=\mathrm{d}x$:

$$\begin{aligned}\int\frac{1}{\cos(x)}\mathrm{d}x&=\int\frac{1}{\sin\left(x+\frac{\pi}{2}\right)}\mathrm{d}x=\int\frac{1}{\sin(u)}\mathrm{d}u\\&=\ln\left(\left|\tan\left(\frac{u}{2}\right)\right|\right)=\ln\left(\left|\tan\left(\frac{x}{2}+\frac{\pi}{4}\right)\right|\right)\end{aligned} \tag{6.3.464}$$

    thanks to the knowledge of the primitive of $\sin^{-1}(x)$. Finally:

$$F(x)=\int\frac{1}{\cos(x)}\mathrm{d}x=\ln\left(\left|\tan\left(\frac{x}{2}+\frac{\pi}{4}\right)\right|\right) \tag{6.3.465}$$

26. Primitive of $f(x)=\dfrac{1}{1+\cos(x)}$:

    We do the substitution $x=2\arctan(t), t=\tan(x/2)$, knowing that (see section Trigonometry page 2519):

$$\cos(x)=\frac{1-\tan^2\left(\frac{x}{2}\right)}{1+\tan^2\left(\frac{x}{2}\right)} \tag{6.3.466}$$

    we get:

$$\cos(x)=\frac{1-t^2}{1+t^2}\quad\text{and}\quad\mathrm{d}x=\frac{2}{1+t^2}\mathrm{d}t \tag{6.3.467}$$

    Therefore:

$$\int\frac{1}{1+\cos(x)}\mathrm{d}x=\int 1\mathrm{d}t=t=\tan\left(\frac{x}{2}\right) \tag{6.3.468}$$

    Finally:

$$F(x)=\int\frac{1}{1+\cos(x)}\mathrm{d}x=\tan\left(\frac{x}{2}\right) \tag{6.3.469}$$

27. Primitive of $f(x)=\dfrac{1}{1-\cos(x)}$:

    We do again the substitution $x=2\arctan(t), t=\tan(x/2)$. Then we find:

$$\int\frac{1}{1-\cos(x)}\mathrm{d}x=\int\frac{1}{t^2}\mathrm{d}t=-\frac{1}{t}=-\cot\left(\frac{x}{2}\right) \tag{6.3.470}$$

    So that finally:

$$F(x)=\int\frac{1}{1-\cos(x)}\mathrm{d}x=-\cot\left(\frac{x}{2}\right) \tag{6.3.471}$$

28. Primitive of $f(x)=\dfrac{1}{1+\sin(x)}$:

    Knowing that:

    $$\sin(x)=\cos\left(x-\frac{\pi}{2}\right) \tag{6.3.472}$$

    we can write:

    $$\int\frac{1}{1+\sin(x)}\mathrm{d}x=\int\frac{1}{1+\cos\left(x-\frac{\pi}{2}\right)}\mathrm{d}x \tag{6.3.473}$$

    By making the change of variables:

    $$u=x-\frac{\pi}{2}\quad\text{with}\quad\mathrm{d}u=\mathrm{d}x \tag{6.3.474}$$

    we get:

    $$\int\frac{1}{1+\sin(x)}\mathrm{d}x=\int\frac{1}{1+\cos(u)}\mathrm{d}u=\tan\left(\frac{u}{2}\right)=\tan\left(\frac{x}{2}-\frac{\pi}{4}\right) \tag{6.3.475}$$

    Finally:

    $$F(x)=\int\frac{1}{1+\sin(x)}\mathrm{d}x=\tan\left(\frac{x}{2}-\frac{\pi}{4}\right) \tag{6.3.476}$$

29. Primitive of $f(x)=\dfrac{1}{1-\sin(x)}$:

    By the same reasoning as above using the cosine we get:

    $$F(x)=\int\frac{1}{1-\sin(x)}\mathrm{d}x=-\cot\left(\frac{x}{2}-\frac{\pi}{4}\right) \tag{6.3.477}$$

30. Primitive of $f(x)=\sinh^n(x)$ with $n\geq 2$:

    Let us put:

    $$I_n=\int\sinh^n(x)\mathrm{d}x \tag{6.3.478}$$

    An integration by parts gives (we proved during our study of usual derivatives that the primitive of the hyperbolic sine was the hyperbolic cosine):

    $$\begin{aligned}I_n=\int\sinh(x)\sinh^{n-1}(x)\mathrm{d}x&=\cosh(x)\sinh(x)^{n-1}(x)\\&\quad-(n-1)\int\cosh^2(x)\sinh^{n-2}(x)\mathrm{d}x\end{aligned} \tag{6.3.479}$$

    by substituting $\cosh^2(x)$ and $1+\sinh^2(X)$ in the last integral, we get:

    $$I_n=\cosh(x)\sinh(x)^{n-1}(x)-(n-1)(I_{n-2}+I_n) \tag{6.3.480}$$

    and therefore:

    $$F(x)=I_n=\int\sinh^n(x)=\frac{1}{n}\left(\cosh(x)\sinh(x)^{n-1}(x)-(n-1)I_{n-2}\right) \tag{6.3.481}$$

    Therefore we get easily the special case:

    $$F(x)=\int\sinh^2(x)=\frac{1}{2}\left(\cosh(x)\sinh(x)(x)-x\right) \tag{6.3.482}$$

31. Primitive of $f(x)=\cosh^n(x)$ with $n\geq 2$:

    In this case we also have the recurrence relation:

$$I_n=\int\cosh(x)^n x\mathrm{d}x=\frac{1}{n}\big(\sinh(x)\cosh(x)^{n-1}(x)+(n-1)I_{n-2}\big)\tag{6.3.483}$$

    that is proved in the same way as above. So we also get easily the special case:

$$\int\cosh^2(x)\mathrm{d}x=\frac{1}{2}(\cosh(x)\sinh(x)+x)\tag{6.3.484}$$

32. Primitive of $f(x)=\tanh^2(x)$:

    Knowing that (proved during our study of usual derivatives):

$$\tanh'(x)=1-\tanh^2(x)\tag{6.3.485}$$

    we have:

$$\int\tanh^2(x)\mathrm{d}x=\int\big(1-\tanh'(x)\big)\mathrm{d}x=x-\tanh(x)\tag{6.3.486}$$

    Therefore:

$$F(x)=\int\tanh^2(x)\mathrm{d}x=x-\tanh(x)\tag{6.3.487}$$

33. Primitive of $f(x)=\coth^2(x)$:

    Knowing that (proved during our study of usual derivatives):

$$\coth'(x)=1-\coth^2(x)\tag{6.3.488}$$

    we have:

$$\int\coth^2(x)\mathrm{d}x=\int\big(1-\coth'(x)\big)\mathrm{d}x=x-\coth(x)\tag{6.3.489}$$

    Therefore:

$$F(x)=\int\coth^2(x)\mathrm{d}x=x-\coth(x)\tag{6.3.490}$$

34. Primitive of $f(x)=\dfrac{1}{\sinh^2(x)}$:

    We have using the primitive of $f(x)=\coth^2(x)$:

$$\int\sinh^2(x)\mathrm{d}x=\int\frac{\cosh^2(x)-\sinh^2(x)}{\sinh^2(x)}\mathrm{d}x=\int(\coth^2(x)-1)\mathrm{d}x=-\coth(x)\tag{6.3.491}$$

    Therefore:

$$F(x)=\int\sinh^2(x)\mathrm{d}x=\coth(x)\tag{6.3.492}$$

35. Primitive of $f(x)=\dfrac{1}{\cosh^2(x)}$:

    We have using the primitive of $f(x)=\tanh^2(x)$:

$$\int \cosh^2(x)\mathrm{d}x=\int \frac{\cosh^2(x)-\sinh^2(x)}{\cosh^2(x)}\mathrm{d}x=\int (1-\tanh^2(x))\mathrm{d}x=\tanh(x) \tag{6.3.493}$$

    Therefore:

$$F(x)=\int \cosh^2(x)\mathrm{d}x=\tanh(x) \tag{6.3.494}$$

36. Primitive of $f(x)=\dfrac{1}{\sinh(x)}$:

    We do the substitution:

$$t=e^x \quad \text{with} \quad \mathrm{d}x=\frac{\mathrm{d}t}{t} \tag{6.3.495}$$

    We get using the derivative of arctanh(x):

$$\begin{aligned}\int \frac{1}{\sinh(x)}\mathrm{d}x&=\int \frac{2}{e^x-e^{-x}}\mathrm{d}x=\int \frac{2}{t\left(t-\dfrac{1}{t}\right)}\mathrm{d}t=\int \frac{2}{t^2-1}\mathrm{d}t\\&=-2\text{arctanh}(t)=-2\text{arctanh}(e^x)\end{aligned} \tag{6.3.496}$$

    and finally:

$$F(x)=\int \frac{1}{\sinh(x)}\mathrm{d}x=-2\text{arctanh}(e^x) \tag{6.3.497}$$

37. Primitive of $f(x)=\dfrac{1}{\cosh(x)}$:

    We do the substitution:

$$t=e^x \quad \text{with} \quad \mathrm{d}x=\frac{\mathrm{d}t}{t} \tag{6.3.498}$$

    We get using the derivative of $\arctan(x)$:

$$\begin{aligned}\int \frac{1}{\cosh(x)}\mathrm{d}x&=\int \frac{2}{e^x+e^{-x}}\mathrm{d}x=\int \frac{2}{t\left(t+\dfrac{1}{t}\right)}\mathrm{d}t\\&=\int \frac{2}{t^2+1}\mathrm{d}t=2\arctan(t)=2\arctan(e^x)\end{aligned} \tag{6.3.499}$$

    and finally:

$$F(x)=\int \frac{1}{\cosh(x)}\mathrm{d}x=2\arctan(e^x) \tag{6.3.500}$$

38. Primitive of $f(x)=\dfrac{1}{1+\cosh(x)}$:

We do the substitution:

$$t=e^x \text{ with } \mathrm{d}x=\frac{\mathrm{d}t}{t} \tag{6.3.501}$$

We get:

$$\begin{aligned}\int\frac{1}{1+\cosh(x)}\mathrm{d}x&=\int\frac{2}{2+e^x+e^{-x}}\mathrm{d}x=\int\frac{2}{t\left(2+t+\dfrac{1}{t}\right)}\mathrm{d}t\\&=\int\frac{2}{(t+1)^2}\mathrm{d}t=-\frac{2}{1+t}+c^{te}=-\frac{2}{1+e^x}+c^{te}\end{aligned} \tag{6.3.502}$$

Finally we get:

$$F(x)=\int\frac{1}{1+\cosh(x)}\mathrm{d}x=-\frac{2}{1+e^x}+1=\frac{e^x-1}{e^x+1}=\tanh\left(\frac{x}{2}\right) \tag{6.3.503}$$

39. Primitive of $f(x)=\dfrac{1}{1-\cosh(x)}$:

We do the substitution:

$$t=e^x \text{ with } \mathrm{d}x=\frac{\mathrm{d}t}{t} \tag{6.3.504}$$

We get:

$$\begin{aligned}\int\frac{1}{1-\cosh(x)}\mathrm{d}x&=\int\frac{2}{2-e^x-e^{-x}}\mathrm{d}x=\int\frac{2}{t\left(2-t-\dfrac{1}{t}\right)}\mathrm{d}t=\int\frac{2}{(t-1)^2}\mathrm{d}t\\&=\frac{2}{t-1}+c^{te}=\frac{2}{e^x-1}+c^{te}\end{aligned} \tag{6.3.505}$$

Finally:

$$F(x)=\int\frac{1}{1-\cosh(x)}\mathrm{d}x=\frac{2}{e^x-1}+1=\frac{e^x+1}{e^x-1}=\coth\left(\frac{x}{2}\right) \tag{6.3.506}$$

40. Primitive of $f(x)=\dfrac{1}{1+\sinh(x)}$:

We do the substitution:

$$t=e^x \text{ with } \mathrm{d}x=\frac{\mathrm{d}t}{t} \tag{6.3.507}$$

We get:

$$\int\frac{1}{1+\sinh(x)}\mathrm{d}x=\int\frac{2}{2+e^x-e^{-x}}\mathrm{d}x=\int\frac{2}{2t+t^2-1}\mathrm{d}t \tag{6.3.508}$$

But:

$$\begin{aligned}\frac{2}{2t+t^2-1}&=\frac{2}{(t+1-\sqrt{2})(t+1+\sqrt{2})}\\&=\frac{1}{\sqrt{2}}\left(\frac{1}{(t+1-\sqrt{2})}-\frac{1}{(t+1+\sqrt{2})}\right)\end{aligned}\tag{6.3.509}$$

Therefore:

$$\begin{aligned}\int\frac{2}{2t+t^2-1}\mathrm{d}t&=\frac{1}{\sqrt{2}}\left(\int\frac{1}{(t+1-\sqrt{2})}\mathrm{d}t-\int\frac{1}{(t+1+\sqrt{2})}\mathrm{d}t\right)\\&=\frac{1}{\sqrt{2}}\left[\ln\left(\left|t+1-\sqrt{2}\right|\right)-\ln\left(\left|t+1+\sqrt{2}\right|\right)\right]\\&=\frac{1}{\sqrt{2}}\ln\left(\frac{\left|t+1-\sqrt{2}\right|}{\left|t+1+\sqrt{2}\right|}\right)=\frac{1}{\sqrt{2}}\ln\left(\frac{\left|e^x+1-\sqrt{2}\right|}{\left|e^x+1+\sqrt{2}\right|}\right)\end{aligned}\tag{6.3.510}$$

Therefore:

$$F(x)=\int\frac{1}{1+\sinh(x)}\mathrm{d}x=\frac{1}{\sqrt{2}}\ln\left(\frac{\left|e^x+1-\sqrt{2}\right|}{\left|e^x+1+\sqrt{2}\right|}\right)\tag{6.3.511}$$

41. Primitive of $f(x)=\dfrac{1}{1-\sinh(x)}$:

We still do the same substitution:

$$t=e^x\ \text{ with }\ \mathrm{d}x=\frac{\mathrm{d}t}{t}\tag{6.3.512}$$

We get:

$$\int\frac{1}{1-\sinh(x)}\mathrm{d}x=\int\frac{2}{2-e^x+e^{-x}}\mathrm{d}x=-\int\frac{2}{-2t+t^2-1}\mathrm{d}t\tag{6.3.513}$$

But:

$$\begin{aligned}\frac{2}{-2t+t^2-1}&=\frac{2}{(t-1-\sqrt{2})(t-1+\sqrt{2})}\\&=\frac{1}{\sqrt{2}}\left(\frac{1}{(t-1-\sqrt{2})}-\frac{1}{(t-1+\sqrt{2})}\right)\end{aligned}\tag{6.3.514}$$

Therefore:

$$\begin{aligned}-\int\frac{2}{-2t+t^2-1}\mathrm{d}t&=-\frac{1}{\sqrt{2}}\left(\int\frac{1}{(t-1-\sqrt{2})}\mathrm{d}t+\int\frac{1}{(t-1+\sqrt{2})}\mathrm{d}t\right)\\&=\frac{1}{\sqrt{2}}\left[\ln\left(\left|t-1+\sqrt{2}\right|\right)-\ln\left(\left|t-1-\sqrt{2}\right|\right)\right]\\&=\frac{1}{\sqrt{2}}\ln\left(\frac{\left|t-1+\sqrt{2}\right|}{\left|t-1-\sqrt{2}\right|}\right)\frac{1}{\sqrt{2}}\ln\left(\frac{\left|e^x-1+\sqrt{2}\right|}{\left|e^x-1-\sqrt{2}\right|}\right)\end{aligned}\tag{6.3.515}$$

Finally:

$$F(x)=\int\frac{1}{1-\sinh(x)}\mathrm{d}x=\frac{1}{\sqrt{2}}\ln\left(\frac{\left|e^x-1+\sqrt{2}\right|}{\left|e^x-1-\sqrt{2}\right|}\right)\tag{6.3.516}$$

42. Primitive of $f(x)=e^{ax}\sin(bx)$ with $a,b\in\mathbb{R}, a^2+b^2\neq 0$:

    A first integration by parts gives:

$$\int e^{ax}\sin(bx)\mathrm{d}x=\frac{1}{a}e^{ax}\sin(bx)-\frac{b}{a}\int e^{ax}\cos(bx)\mathrm{d}x \tag{6.3.517}$$

    A second integration by parts gives:

$$\int e^{ax}\sin(bx)\mathrm{d}x=\frac{1}{a}e^{ax}\sin(bx)+\frac{b}{a}\left(\frac{1}{a}e^{ax}\cos(bx)-\frac{b}{a}\int e^{ax}\sin(bx)\mathrm{d}x\right) \tag{6.3.518}$$

    So we have the equality:

$$\int e^{ax}\sin(bx)\mathrm{d}x=\frac{1}{a}e^{ax}\sin(bx)-\frac{b}{a^2}e^{ax}\cos(bx)-\frac{b^2}{a^2}\int e^{ax}\sin(bx)\mathrm{d}x \tag{6.3.519}$$

    Therefore redistributing the previous relation:

$$F(x)=\int e^{ax}\sin(bx)\mathrm{d}x=\int e^{ax}\sin(bx)\mathrm{d}x=\frac{e^{ax}}{a^2+b^2}\big(a\sin(bx)-b\cos(bx)\big) \tag{6.3.520}$$

43. Primitive of $f(x)=e^{ax}\cos(bx)$ with $a,b\in\mathbb{R}, a^2+b^2\neq 0$:

    A similar reasoning than before shows that (we can detail on demand as always!):

$$F(x)=\int e^{ax}\cos(bx)\mathrm{d}x=\frac{e^{ax}}{a^2+b^2}\big(a\cos(bx)+b\sin(bx)\big) \tag{6.3.521}$$

44. Primitive of $f(x)=x\sin(ax)$ with $a\in\mathbb{R}^*$:

    And integration by parts gives us:

$$\begin{aligned}F(x)=\int x\sin(ax)\mathrm{d}x&=-\frac{1}{a}x\cos(ax)+\frac{1}{a}\int\cos(ax)\mathrm{d}x\\&=-\frac{1}{a}x\cos(ax)+\frac{1}{a^2}\sin(ax)\end{aligned} \tag{6.3.522}$$

45. Primitive of $f(x)=x\cos(ax)$ with $a\in\mathbb{R}^*$:

    And integration by parts gives us:

$$\begin{aligned}F(x)=\int x\cos(ax)\mathrm{d}x&=\frac{1}{a}x\sin(ax)-\frac{1}{a}\int\sin(ax)\mathrm{d}x\\&=\frac{1}{a}x\sin(ax)+\frac{1}{a^2}\cos(ax)\end{aligned} \tag{6.3.523}$$

46. Primitive of $f(x)=\dfrac{1}{(x-a)(x-b)}$ with $a\neq b$:

    We have the following relation (in integral calculus we name such decomposition a "partial fraction decomposition"):

$$\frac{1}{(x-a)(x-b)}=\frac{1}{a-b}\left(\frac{1}{(x-a)}-\frac{1}{(x-b)}\right) \tag{6.3.524}$$

Therefore:

$$\begin{aligned}\int\frac{1}{(x-a)(x-b)}\mathrm{d}x&=\frac{1}{a-b}\int\left(\frac{1}{(x-a)}-\frac{1}{(x-b)}\right)\mathrm{d}x\\&=\frac{1}{a-b}\left(\int\frac{1}{(x-a)}\mathrm{d}x-\int\frac{1}{(x-b)}\mathrm{d}x\right)\\&=\frac{1}{a-b}(\ln|x-a|-\ln|x-b|)\\&=\frac{1}{a-b}\ln\left(\left|\frac{x-a}{x-b}\right|\right)\end{aligned}\tag{6.3.525}$$

Finally:

$$F(x)=\int\frac{1}{(x-a)(x-b)}\mathrm{d}x=\frac{1}{a-b}\ln\left(\left|\frac{x-a}{x-b}\right|\right)\tag{6.3.526}$$

47. Primitive of $f(x)=\dfrac{1}{a^2-x^2}$ with $a\neq 0$:

We have using the previous result:

$$\begin{aligned}\int\frac{1}{a^2-x^2}\mathrm{d}x=\int\frac{1}{(a+x)(a-x)}\mathrm{d}x&=-\int\frac{1}{(x+a)(x-a)}\mathrm{d}x\\=-\frac{1}{2a}\ln\left(\left|\frac{x-a}{x+a}\right|=\frac{1}{2a}\ln\left|\frac{x+a}{x-a}\right|\right)\end{aligned}\tag{6.3.527}$$

Therefore:

$$F(x)=\int\frac{1}{a^2-x^2}\mathrm{d}x=\frac{1}{2a}\ln\left(\left|\frac{x+a}{x-a}\right|\right)\tag{6.3.528}$$

48. Primitive of $f(x)=\dfrac{1}{a^2+x^2}$ with $a\neq 0$:

Doing a change of variable:

$$x=au\qquad\text{with}\qquad\mathrm{d}x=a\mathrm{d}u\tag{6.3.529}$$

We get using the derivative of $\arctan(x)$:

$$\begin{aligned}F(x)=\int\frac{1}{a^2+x^2}\mathrm{d}x&=\frac{1}{a}\int\frac{1}{1+u^2}\mathrm{d}u\\&=\frac{1}{a}\arctan(u)=\frac{1}{a}\arctan\left(\frac{x}{a}\right)\end{aligned}\tag{6.3.530}$$

49. Given:

$$I_{n+1}=\int\frac{1}{(1-x^2)^{n+1}}\mathrm{d}x\tag{6.3.531}$$

with $n\in\mathbb{N}$. We get:

$$\begin{aligned}I_{n+1}&=\int\frac{1}{(1-x^2)^{n+1}}\mathrm{d}x=\int\left(\frac{1-x^2}{(1-x^2)^{n+1}}+\frac{x^2}{(1-x^2)^{n+1}}\right)\mathrm{d}x\\&=\int\frac{1}{(1-x^2)^{n}}\mathrm{d}x+\int\frac{x^2}{(1-x^2)^{n+1}}\mathrm{d}x=I_n+\int\frac{x^2}{(1-x^2)^{n+1}}\mathrm{d}x\end{aligned}\tag{6.3.532}$$

But this last primitive can be solved by parts:

$$\begin{aligned}\int\frac{x^2}{(1-x^2)^{n+1}}\mathrm{d}x&=\int\frac{x}{(1-x^2)^{n+1}}x\mathrm{d}x\\&=\frac{1}{2n(1-x^2)^n}x-\int\frac{1}{2n(1-x^2)^n}\mathrm{d}x=\frac{1}{2n(1-x^2)^n}x-\frac{1}{2n}I_n\end{aligned} \tag{6.3.533}$$

Therefore:

$$I_{n+1}=I_n+\frac{1}{2n(1-x^2)^n}x-\frac{1}{2n}I_n=\frac{1}{2n(1-x^2)^n}x-\frac{(2n-1)}{2n}I_n \tag{6.3.534}$$

What we find most frequently in the literature under the form:

$$\boxed{2nI_{n+1}=\frac{x}{(1-x^2)^n}+(2n-1)I_n} \tag{6.3.535}$$

Identically to the next development, we have for (the sign change):

$$I_{n+1}=\int\frac{1}{(1+x^2)^{n+1}}\mathrm{d}x \tag{6.3.536}$$

the following relation:

$$\boxed{2nI_{n+1}=\frac{x}{(1+x^2)^n}+(2n-1)I_n} \tag{6.3.537}$$

You can find an application of these two primitives in the Newtonian cosmological model of the Universe in the section of Astrophysics and also in the section of General Relativity in the study of the Shapiro effect!

50. Primitive of $f(x)=\dfrac{1}{(1-x^2)^2}$:

    We have using the primitives of $\dfrac{1}{(1-x^2)^n}$ (proved before) and of $\dfrac{1}{1-x^2}$ (also proved above):

    $$F(x)=\int\frac{1}{(1-x^2)^2}\mathrm{d}x=\frac{1}{2}\left[\frac{x}{1-x^2}+\frac{1}{2}\ln\left(\left|\frac{x+1}{x-1}\right|\right)\right] \tag{6.3.538}$$

51. Primitive of $f(x)=\dfrac{1}{(1+x^2)^2}$:

    We have using the primitives of $\dfrac{1}{(1+x^2)^n}$ (proved before) and of $\dfrac{1}{1+x^2}$ (also proved above):

    $$F(x)=\int\frac{1}{(1+x^2)^2}\mathrm{d}x=\frac{1}{2}\left(\frac{x}{1+x^2}+\arctan(x)\right) \tag{6.3.539}$$

    You can find an application of this primitive in the section of General Relativity for the proper falling time into a non-rotating Black Hole.

52. Primitive of $f(x)=\sqrt{x^2-a^2}$ with $a\in\mathbb{R}^*$ (case relative to the area under a hyperbola):

We can assume without loss of generality $a>0$. Note that the domain of definition of f is $]-\infty,-a]\cup[a,+\infty[$.

We will determine now a primitive of f only on the interval $[a,+\infty[$ (because that is it we will need in some sections of this textbook).

Let us make the change of variable:

$$x=a\cdot\cosh(t) \tag{6.3.540}$$

So with:

$$t=\text{arccosh}\left(\frac{1}{a}x\right),\quad \mathrm{d}x=a\sinh(t)\mathrm{d}t \tag{6.3.541}$$

where we consider the function $\cosh:\mathbb{R}^+\to[1,+\infty[$ with for reciprocal the function $\text{arccosh}:[1,+\infty[\to\mathbb{R}^+$ given by (see section Trigonometry page 2534):

$$\text{arccosh}(y)=\ln\left(y+\sqrt{y^2-1}\right) \tag{6.3.542}$$

We get then using then primitive of $\sinh^2(x)$:

$$\begin{aligned}\int\sqrt{x^2-a^2}\mathrm{d}x=a^2\int\sinh^2(x)\mathrm{d}t&=\frac{a^2}{2}(\cosh(t)\sinh(t)-t)\\&=\frac{a^2}{2}\cdot\left(\frac{1}{a}x\sinh\left(\text{arccosh}\left(\frac{1}{a}x\right)\right)-\text{arccosh}\left(\frac{1}{a}x\right)\right)\end{aligned} \tag{6.3.543}$$

but (see section Trigonometry page 2532) as:

$$\sinh^2(u)=\cosh^2(u)-1 \tag{6.3.544}$$

Therefore:

$$\sinh\left(\text{arccosh}\left(\frac{1}{a}x\right)\right)=\sqrt{\cosh^2\left(\text{arccosh}\left(\frac{1}{a}x\right)\right)-1}=\sqrt{\frac{1}{a^2}x^2-1} \tag{6.3.545}$$

and using another identity proved in the section Trigonometry:

$$\text{arccosh}(x)=\ln\left(x+\sqrt{x^2+1}\right) \tag{6.3.546}$$

we have therefore:

$$\begin{aligned}\int\sqrt{x^2-a^2}\mathrm{d}x&=\frac{a^2}{2}\left(\frac{x}{a}\sqrt{\frac{x^2}{a^2}-1}-\ln\left(\frac{x}{a}+\sqrt{\frac{x^2}{a^2}-1}\right)\right)\\&=\frac{a^2}{2}\left(\frac{x}{a^2}\sqrt{x^2-a^2}-\ln\left(\frac{x}{a}+\frac{1}{a}\sqrt{x^2-a^2}\right)\right)\\&=\frac{a^2}{2}\left(\frac{x}{a^2}\sqrt{x^2-a^2}-\ln\left(x+\sqrt{x^2-a^2}\right)+\ln(a)\right)\end{aligned} \tag{6.3.547}$$

as the primitive are given to a given constant, we can write:

$$F(x)=\int\sqrt{x^2-a^2}\mathrm{d}x=\frac{x}{2}\sqrt{x^2-a^2}-\frac{a^2}{2}\ln\left(x+\sqrt{x^2-a^2}\right) \tag{6.3.548}$$

for $x\geq a$. F is then a primitive of $\sqrt{x^2-a^2}$ on the interval $[a,+\infty[$.

53. Primitive of $f(x)=\sqrt{a^2-x^2}$ with $a\in\mathbb{R}^*$:

    We can assume without loss of generality $a>0$. Note that the domain of definition of f is $[-a,a]$.

    We make the substitution:

$$x=a\sin(t) \qquad \text{with} \qquad \mathrm{d}x=a\cos(t)\mathrm{d}t \tag{6.3.549}$$

    we get:

$$\begin{aligned}\int\sqrt{a^2-x^2}\mathrm{d}x&=a^2\int\sqrt{1-\sin^2(t)}\cos(t)\mathrm{d}t=a^2\int\cos^2(t)\mathrm{d}t\\&=\frac{a^2}{2}\left(t+\sin(t)\cos(t)\right)=\frac{a^2}{2}\left[\arcsin\left(\frac{x}{a}\right)+\frac{x}{a}\cos\left(\arcsin\left(\frac{x}{a}\right)\right)\right]\end{aligned} \tag{6.3.550}$$

    where we used the primitive $\cos^n(x)$ with $n=2$ proved above. Now we have:

$$\cos(\arcsin(y))=\sqrt{1-\sin^2(\arcsin y)}=\sqrt{1-y^2} \tag{6.3.551}$$

    Then:

$$\begin{aligned}\int\sqrt{a^2-x^2}\mathrm{d}x&=\frac{a^2}{2}\left[\arcsin\left(\frac{x}{a}\right)+\frac{x}{a}\cos\left(\arcsin\left(\frac{x}{a}\right)\right)\right]\\&=\frac{a^2}{2}\left[\arcsin\left(\frac{x}{a}\right)+\frac{x}{a}\sqrt{1-\frac{x^2}{a^2}}\right]\\&=\frac{a^2}{2}\arcsin\left(\frac{x}{a}\right)+\frac{x}{2}\sqrt{a^2-x^2}\end{aligned} \tag{6.3.552}$$

    and:

$$F(x)=\int\sqrt{a^2-x^2}\mathrm{d}x=\frac{a^2}{2}\arcsin\left(\frac{x}{a}\right)+\frac{x}{2}\sqrt{a^2-x^2} \tag{6.3.553}$$

54. Primitive of $f(x)=\sqrt{x^2+a^2}$ with $a\in\mathbb{R}^*$:

    We can assume without loss of generality $a>0$.

    Let us make the change of variable:

$$x=a\sinh(t) \tag{6.3.554}$$

    with:

$$t=\operatorname{arcsinh}\left(\frac{1}{a}x\right),\quad \mathrm{d}x=a\cosh(t)\mathrm{d}t \tag{6.3.555}$$

    We get:

$$\int\sqrt{x^2+a^2}\mathrm{d}x=a^2\int\cosh^2(t)\mathrm{d}t=\frac{a^2}{2}\left(t+\sinh(t)\cosh(t)\right) \tag{6.3.556}$$

Therefore:

$$\int \sqrt{x^2+a^2}\mathrm{d}x=\frac{a^2}{2}(t+\sinh(t)\cosh(t))=\frac{a^2}{2}\left(\text{arcsinh}\left(\frac{x}{a}\right)+\frac{x}{a}\cosh\left(\text{arcsinh}\left(\frac{x}{a}\right)\right)\right) \tag{6.3.557}$$

But as we saw in the section on Trigonometry page 2532:

$$\text{arcsinh}\left(\frac{x}{a}\right)=\ln\left(\frac{x}{a}+\frac{1}{a}\sqrt{x^2+a^2}\right)=\ln(x+\sqrt{x^2+a^2})-\ln(a) \tag{6.3.558}$$

and:

$$\cosh\left(\text{arcsinh}\left(\frac{x}{a}\right)\right)=\sqrt{1+\sinh^2(\text{arcsinh}\left(\frac{x}{a}\right))}=\sqrt{1+\frac{x^2}{a^2}}=\frac{1}{a}\sqrt{a^2+x^2} \tag{6.3.559}$$

Finally we have:

$$F(x)=\int \sqrt{x^2+a^2}\mathrm{d}x=\frac{a^2}{2}\ln\left(x+\sqrt{x^2+a^2}\right)+\frac{x}{2}\sqrt{a^2+x^2} \tag{6.3.560}$$

where $ln(a)$ has been omitted because the primitives are given to a given constant.

55. Primitive of $f(x)=\dfrac{1}{\sqrt{a^2-x^2}}$ with $a\in\mathbb{R}^*$:

    We can assume without loss of generality $a>0$.

    We do the substitution:

$$x=a\sin(t)\qquad\text{with}\qquad \mathrm{d}x=a\cos(t)\mathrm{d}t \tag{6.3.561}$$

We get:

$$\begin{aligned}F(x)&=\int\frac{1}{\sqrt{a^2-x^2}}\mathrm{d}x=\int\frac{1}{\sqrt{a^2-a^2\sin^2(t)}}a\cos(t)\mathrm{d}t\\&=\int\frac{1}{a\sqrt{1-\sin^2(t)}}a\cos(t)\mathrm{d}t=\int\frac{1}{a\cos(t)}a\cos(t)\mathrm{d}t=\int 1\mathrm{d}t=t\\&=\arcsin\left(\frac{x}{a}\right)\end{aligned} \tag{6.3.562}$$

56. Primitive of $f(x)=\dfrac{1}{\sqrt{a^2+x^2}}$ with $a\in\mathbb{R}^*$:

    We can assume without loss of generality $a>0$.

    We do the substitution:

$$x=a\sinh(t)\qquad\text{with}\qquad t=\text{arcsinh}\left(\frac{1}{a}x\right),\quad \mathrm{d}x=a\cosh(t)\mathrm{d}t \tag{6.3.563}$$

We get in the same manner as the previous usual integrals:

$$\int\frac{1}{\sqrt{(x^2+a^2)}}\mathrm{d}x=\int 1\mathrm{d}x=t=\text{arcsinh}\left(\frac{x}{a}\right) \tag{6.3.564}$$

and knowing that (see section Trigonometry page 2534):

$$\text{arcsinh}(y)=\ln\left(y+\sqrt{y^2+1}\right) \tag{6.3.565}$$

We then get finally the following important primitive:

$$\boxed{F(x)=\int\frac{1}{\sqrt{x^2+a^2}}\mathrm{d}x=\ln(x+\sqrt{x^2+a^2})} \tag{6.3.566}$$

Proceeding in the same way, but using the hyperbolic cosine instead of hyperbolic sine, we get obviously (we can detail on demand as always):

$$\boxed{F(x)=\int\frac{1}{\sqrt{x^2-a^2}}\mathrm{d}x=\ln(x+\sqrt{x^2-a^2})} \tag{6.3.567}$$

We will reuse these last two relation in important practical cases of the section of Analytical Mechanics, Civil Engineering (where the constant a is equal 1, $\ln(a)$ will be equal to 0) and General Relativity (where a will be non-zero and therefore it will not be possible to omit the constant $\ln(a)$).

57. Primitive of $f(x)=\sqrt{\dfrac{1}{x}-1}$:

So the problem is:

$$\int\sqrt{\frac{1}{x}-1}\mathrm{d}x \tag{6.3.568}$$

Substitute:

$$u=\sqrt{\frac{1}{x}-1}\longrightarrow\mathrm{d}u=-\frac{1}{2\sqrt{\frac{1}{x}-1}x^2}\mathrm{d}x=-2\int\frac{u^2}{\left(u^2+1\right)^2}\,\mathrm{d}u \tag{6.3.569}$$

Now solving:

$$\int\frac{u^2}{\left(u^2+1\right)^2}\mathrm{d}u \tag{6.3.570}$$

Write u^2 as u^2+1-1 and split:

$$\int\frac{u^2}{\left(u^2+1\right)^2}\mathrm{d}u=\int\left(\frac{u^2+1}{\left(u^2+1\right)^2}-\frac{1}{\left(u^2+1\right)^2}\right)\mathrm{d}u=\int\left(\frac{1}{u^2+1}-\frac{1}{\left(u^2+1\right)^2}\right)\mathrm{d}u \tag{6.3.571}$$

We apply linearity:

$$\int\frac{u^2}{\left(u^2+1\right)^2}\mathrm{d}u=\int\frac{1}{u^2+1}\,\mathrm{d}u-\int\frac{1}{\left(u^2+1\right)^2}\mathrm{d}u \tag{6.3.572}$$

We know from earlier above that the first term is:

$$\int\frac{1}{u^2+1}\mathrm{d}u=\arctan(u) \tag{6.3.573}$$

Now solving:

$$\int \frac{1}{\left(u^2+1\right)^2}\mathrm{d}u \tag{6.3.574}$$

We also have proved this integral earlier above. Then:

$$\int \frac{1}{\left(u^2+1\right)^2}\mathrm{d}u == \frac{\arctan(u)}{2}+\frac{u}{2\left(u^2+1\right)} \tag{6.3.575}$$

Plug in solved integrals:

$$\int \frac{1}{u^2+1}\mathrm{d}u - \int \frac{1}{\left(u^2+1\right)^2}\mathrm{d}u = \frac{\arctan(u)}{2}-\frac{u}{2\left(u^2+1\right)} \tag{6.3.576}$$

Plug in solved integrals:

$$-2\int \frac{u^2}{\left(u^2+1\right)^2}\mathrm{d}u = \frac{u}{u^2+1}-\arctan(u) \tag{6.3.577}$$

We undo the substitution:

$$u=\sqrt{\frac{1}{x}-1} \tag{6.3.578}$$

This gives:

$$\sqrt{\frac{1}{x}-1}x-\arctan\left(\sqrt{\frac{1}{x}-1}\right) \tag{6.3.579}$$

Hence the problem is solved:

$$\int \sqrt{\frac{1}{x}-1}\mathrm{d}x = \sqrt{\frac{1}{x}-1}x-\arctan\left(\sqrt{\frac{1}{x}-1}\right)+c^{te} \tag{6.3.580}$$

58. In mathematics, there are several integrals known as the "Dirichlet integral". One of those is the improper integral of the sinc function over the positive real line:

$$\int\limits_0^{+\infty}\frac{\sin(\omega)}{\omega}\mathrm{d}\omega=\int\limits_0^{+\infty}\operatorname{sinc}(\omega)\mathrm{d}\omega \tag{6.3.581}$$

By Fubini's Theorem (see page 1521):

$$\int\limits_0^{+\infty}\left(\int\limits_0^{+\infty}e^{-xy}\sin(x)\mathrm{d}y\right)\mathrm{d}x=\int\limits_0^{+\infty}\left(\int\limits_0^{+\infty}e^{-xy}\sin(x)\mathrm{d}x\right)\mathrm{d}y \tag{6.3.582}$$

Then using the primitive of e^{ax}:

$$\int\limits_0^{+\infty}e^{-xy}\sin(x)\mathrm{d}y=-e^{-xy}\frac{\sin(x)}{x}\bigg|_0^{+\infty}=\frac{\sin(x)}{x} \tag{6.3.583}$$

and using the primitive of $e^{ax}\sin(bx)$:

$$\int_0^{+\infty} e^{-xy}\sin(x)\mathrm{d}x=\left.\frac{-e^{-xy}(y\sin(x)+\cos(x))}{y^2+1}\right|_0^{+\infty}=\frac{1}{y^2+1} \tag{6.3.584}$$

Hence:

$$\int_0^{+\infty}\frac{\sin(x)}{x}\mathrm{d}x=\int_0^{+\infty}\frac{1}{y^2+1}\mathrm{d}y \tag{6.3.585}$$

Using the primitive of $\frac{1}{x^2+a^2}$ seen earlier:

$$\int_0^{+\infty}\frac{\sin(x)}{x}\mathrm{d}x=\int_0^{+\infty}\frac{1}{y^2+1}\mathrm{d}y=\arctan(y)\Big|_0^{+\infty} \tag{6.3.586}$$

As:

$$\lim_{y\to+\infty}\arctan(y)=\frac{\pi}{2} \tag{6.3.587}$$

Then:

$$\int_0^{+\infty}\frac{\sin(x)}{x}\mathrm{d}x=\arctan(y)\Big|_0^{+\infty}=\frac{\pi}{2} \tag{6.3.588}$$

59. Let us consider an integral of the following form that we can use to improve the Stirling formula (see section Numerical Methods page 6301) and also that we need absolutely for the study of the circular Fresnel aperture diffraction (see section Wave Optics page 4470):

$$I=\int_a^b e^{-\lambda g(y)}h(y)\mathrm{d}y \tag{6.3.589}$$

where:

- λ is large;
- $g(y)$ is a smooth function which has a local minimum at y^* in the interior of the interval $[a,b]$;
- $h(y)$ is smooth.

The integral can be the characteristic generating function of the distribution of $g(Y)$ when Y has density h (see section Statistics page 926), it could be a posterior expectation of $h(Y)$, or just a "simple" integral.

When λ is large, the contribution to this integral is essentially entirely by construction originating from a neighbourhood around y^*.

We formalize this by Taylor expansion of the function g around y^*:

$$g(y)=g(y^*)+g'(y^*)(y-y^*)+\frac{1}{2}g''(y^*)(y-y^*)^2+\ldots \tag{6.3.590}$$

Since y^* is a local minimum, we have:

$$g'(y^*)=0 \qquad g''(y^*)>0 \tag{6.3.591}$$

and therefore:

$$g(y)=g(y^*)+\frac{1}{2}g''(y^*)(y-y^*)^2+\ldots \tag{6.3.592}$$

Therefore:

$$\begin{aligned}I&=\int\limits_a^b e^{-\lambda g(y)}h(y)\mathrm{d}y\\ &\cong\int\limits_a^b e^{-\lambda\left(g(y^*)+\frac{1}{2}g''(y^*)(y-y^*)^2\right)}h(y)\mathrm{d}y\\ &=e^{-\lambda g(y^*)}\int\limits_{-\infty}^{+\infty}e^{-\lambda g''(y^*)(y-y^*)^2/2}h(y)\mathrm{d}y\end{aligned} \tag{6.3.593}$$

The fact that above the bounds have change from $[a,b]$ to $]-\infty,+\infty[$ is that we assume that the area of interest is around y^* and that because λ is soooo large that veeeery quickly a bit away from y^* we can consider the curve as negligible!

If we approximate $h(y)$ linearly around y^*, that is to say:

$$h'(y^*)=\frac{h(y)-h(y^*)}{y-y^*} \tag{6.3.594}$$

such that:

$$h(y)=h'(y^*)(y-y^*)+h(y^*) \tag{6.3.595}$$

Therefore:

$$\begin{aligned}I&\cong e^{-\lambda g(y^*)}\int\limits_{-\infty}^{+\infty}e^{-\lambda g''(y^*)(y-y^*)^2/2}h(y)\mathrm{d}y\\ &=e^{-\lambda g(y^*)}\int\limits_{-\infty}^{+\infty}e^{-\lambda g''(y^*)(y-y^*)^2/2}\left(h'(y^*)(y-y^*)+h(y^*)\right)\mathrm{d}y\\ &=e^{-\lambda g(y^*)}h(y^*)\int\limits_{-\infty}^{+\infty}e^{-\lambda g''(y^*)(y-y^*)^2/2}\mathrm{d}y+e^{-\lambda g(y^*)}h'(y^*)\int\limits_{-\infty}^{+\infty}(y-y^*)e^{-\lambda g''(y^*)(y-y^*)^2/2}\mathrm{d}y\end{aligned} \tag{6.3.596}$$

The reader must not forget that here $\lambda g''(y^*)$ is a constant and that if we put:

$$\lambda g''(y^*)=\frac{1}{\sigma^2} \tag{6.3.597}$$

Then for the first integral above we see that we fall back on the integral of something very similar to the Gauss distribution (with mean y^* and standard deviation $\lambda g''(y^*)$) and then it comes almost immediately (see section Statistics page 695):

$$\int\limits_{-\infty}^{+\infty} e^{-\lambda g''(y^*)(y-y^*)^2/2}\mathrm{d}y=\sqrt{\frac{2\pi}{\lambda g''(y^*)}} \tag{6.3.598}$$

For the second integral:

$$\int\limits_{-\infty}^{+\infty} (y-y^*)e^{-\lambda g''(y^*)(y-y^*)^2/2}\mathrm{d}y \tag{6.3.599}$$

the change of variable $y-y^*=x$ give us:

$$\int\limits_{-\infty}^{+\infty} xe^{-\lambda g''(y^*)x^2/2}\mathrm{d}x \tag{6.3.600}$$

It is immediate that primitive is of the form:

$$c^{te}e^{-\lambda g''(y^*)x^2/2} \tag{6.3.601}$$

Therefore by symmetry the second integral is zero! We have finally:

$$\boxed{I\cong e^{-\lambda g(y^*)}h'(y^*)\sqrt{\frac{2\pi}{\lambda g''(y^*)}}} \tag{6.3.602}$$

This calculation is named "Laplace's Method of Integration" or simply "Laplace Integration".

Remark

People interesting in further reading on common and beautiful integrals can take a look to the very good book of P.J. Nahin: *Inside Interesting Integrals* [408].

To summarize here is a table with the most important primitives (some will be proved later in this textbook):

Function $f(x)$	**Primitive** $F(x)$
0	c^{te}
c^{te}	$c^{te}x$
$\frac{1}{x}$	$\ln(x)+c^{te}$
$\sin(x)$	$-\cos(x)+c^{te}$
$\cos(x)$	$\sin(x)+c^{te}$
e^x	e^x+c^{te}
$\tan(x)$	$-\ln\lvert\cos(x)\rvert+c^{te}$

Function $f(x)$	**Primitive** $F(x)$
$\cot(x)$	$\ln\lvert\sin(x)\rvert + c^{te}$
$\arcsin(x)$	$x\arcsin(x) + \sqrt{1-x^2} + c^{te}$
$\arctan(x)$	$x\arctan(x) - \frac{1}{2}\ln\left(1+x^2\right) + c^{te}$
$\text{arccot}(x)$	$x\,\text{arccot}(x) + \frac{1}{2}\ln\left(1+x^2\right) + c^{te}$
xe^{ax}	$e^{ax}\left(\frac{1}{a}x - \frac{1}{a^2}\right) + c^{te}$
$\ln(x)$	$x(\ln(x)-1) + c^{te}$
$x\ln(ax)$	$\frac{1}{2}x^2\ln(ax) - \frac{1}{4}x^2 + c^{te}$
a^x	$\dfrac{a^x}{\ln(a)} + c^{te}$
$\log_a(x)$	$\dfrac{x}{\ln(a)}(\ln(x)-1) + c^{te}$
$\tanh(x)$	$\ln(\lvert\cosh(x)\rvert) + c^{te}$
$\coth(x)$	$\ln(\lvert\sinh(x)\rvert) + c^{te}$
$\text{arcsinh}(x)$	$x\,\text{arcsinh}(x) - \sqrt{1+x^2} + c^{te}$
$\text{arccosh}(x)$	$x\,\text{arccosh}(x) - \sqrt{x^2-1} + c^{te}$
$\text{arctanh}(x)$	$x\,\text{arctanh}(x) + \frac{1}{2}\ln\lvert 1-x^2\rvert + c^{te}$
$\text{arccoth}(x)$	$x\cdot\text{arccoth}(x) + \frac{1}{2}\ln\left(\lvert 1-x^2\rvert\right) + c^{te}$
$\sin^n(x)$	$\dfrac{1}{n}\left(-\cos(x)\cdot\sin^{n-1}(x) + (n-1)I_{n-2}\right) + c^{te}$
$\cos^n(x)$	$\dfrac{1}{n}\left(\sin(x)\cdot\cos^{n-1}(x) + (n-1)I_{n-2}\right) + c^{te}$
$\tan^2(x)$	$\tan(x) - x + c^{te}$
$\cot^2(x)$	$-\cot(x) - x + c^{te}$
$\sin^{-2}(x)$	$-\cot(x) + c^{te}$
$\cos^{-2}(x)$	$\tan(x) + c^{te}$
$\sin^{-1}(x)$	$\ln\left(\left\lvert\tan\dfrac{x}{2}\right\rvert\right) + c^{te}$
$\cos^{-1}(x)$	$\ln\left(\left\lvert\tan\left(\dfrac{x}{2}+\dfrac{\pi}{4}\right)\right\rvert\right) + c^{te}$
$\dfrac{1}{1+\cos(x)}$	$\ln\left(\left\lvert\tan\left(\dfrac{x}{2}+\dfrac{\pi}{4}\right)\right\rvert\right) + c^{te}$
$\dfrac{1}{1-\cos(x)}$	$-\cot\left(\dfrac{x}{2}\right) + c^{te}$
$\dfrac{1}{1+\sin(x)}$	$\tan\left(\dfrac{x}{2}-\dfrac{\pi}{4}\right) + c^{te}$
$\dfrac{1}{1-\sin(x)}$	$-\cot\left(\dfrac{x}{2}-\dfrac{\pi}{4}\right) + c^{te}$
$\sinh^n(x)$	$\cosh(x)\sinh(x)^{n-1}(x) - (n-1)\left(I_{n-2}+I_n\right) + c^{te}$
$\cosh^n(x)$	$\dfrac{1}{n}\left(\sinh(x)\cosh(x)^{n-1}(x) + (n-1)I_{n-2}\right) + c^{te}$
$\tanh^2(x)$	$x - \tanh(x) + c^{te}$
$\coth^2(x)$	$x - \coth(x) + c^{te}$
$\sinh^{-2}(x)$	$\coth(x) + c^{te}$
$\cosh^{-2}(x)$	$\tanh(x) + c^{te}$
$\sinh^{-1}(x)$	$-2\,\text{arctanh}\left(e^2\right) + c^{te}$
$\cosh^{-1}(x)$	$2\arctan\left(e^x\right) + c^{te}$
$\dfrac{1}{1+\cosh(x)}$	$\tanh\left(\dfrac{x}{2}\right) + c^{te}$

Function $f(x)$	**Primitive** $F(x)$
$\dfrac{1}{1-\cosh(x)}$	$\coth\left(\dfrac{x}{2}\right)+c^{te}$
$\dfrac{1}{1+\sinh(x)}$	$\dfrac{1}{\sqrt{2}}\ln\left(\dfrac{\lvert e^x+1-\sqrt{2}\rvert}{\lvert e^x+1+\sqrt{2}\rvert}\right)+c^{te}$
$\dfrac{1}{1-\sinh(x)}$	$\dfrac{1}{\sqrt{2}}\ln\left(\dfrac{\lvert e^x-1+\sqrt{2}\rvert}{\lvert e^x-1-\sqrt{2}\rvert}\right)+c^{te}$
$e^{ax}\sin(bx)$	$\dfrac{e^{ax}}{a^2+b^2}(a\sin(bx)-b\cos(bx))+c^{te}$
$e^{ax}\cos(bx)$	$\dfrac{e^{ax}}{a^2+b^2}(a\cos(bx)+b\sin(bx))+c^{te}$
$x\sin(ax)$	$-\dfrac{1}{a}x\cos(ax)+\dfrac{1}{a^2}\sin(ax)+c^{te}$
$x\cos(ax)$	$\dfrac{1}{a}x\sin(ax)+\dfrac{1}{a^2}\cos(ax)+c^{te}$
$\dfrac{1}{(x-a)(x-b)}$	$\dfrac{1}{a-b}\ln\left(\left\lvert\dfrac{x-a}{x-b}\right\rvert\right)+c^{te}$
$\dfrac{1}{a^2-x^2}$	$\dfrac{1}{2a}\ln\left(\left\lvert\dfrac{x+a}{x-a}\right\rvert\right)+c^{te}$
$\dfrac{1}{a^2+x^2}$	$\dfrac{1}{a}\arctan\left(\dfrac{x}{a}\right)+c^{te}$
$\dfrac{1}{(1-x^2)^2}$	$\dfrac{1}{2}\left[\dfrac{x}{1-x^2}+\dfrac{1}{2}\ln\left(\left\lvert\dfrac{x+1}{x-1}\right\rvert\right)\right]+c^{te}$
$\dfrac{1}{(1+x^2)^2}$	$\dfrac{1}{2}\left(\dfrac{x}{1+x^2}+\arctan(x)\right)+c^{te}$
$\sqrt{x^2-a^2}$	$\dfrac{x}{2}\sqrt{x^2-a^2}-\dfrac{a^2}{2}\ln\left(x+\sqrt{x^2-a^2}\right)+c^{te}$
$\sqrt{a^2-x^2}$	$\dfrac{a^2}{2}\arcsin\left(\dfrac{x}{a}\right)+\dfrac{x}{2}\sqrt{a^2-x^2}+c^{te}$
$\sqrt{x^2+a^2}$	$\dfrac{a^2}{2}\ln\left(x+\sqrt{x^2+a^2}\right)+\dfrac{x}{2}\sqrt{a^2+x^2}+c^{te}$
$\left(\sqrt{a^2-x^2}\right)^{-1}$	$\arcsin\left(\dfrac{x}{a}\right)+c^{te}$
$\dfrac{1}{\sqrt{a^2+x^2}}$	$\ln\left(x+\sqrt{x^2-a^2}\right)+c^{te}$
$\sqrt{\dfrac{1}{x}-1}$	$\sqrt{\tfrac{1}{x}-1}x-\arctan\left(\sqrt{\tfrac{1}{x}-1}\right)+c^{te}$
$\delta(x-a)$	$\int\limits_{-\infty}^{\infty}\delta(x-a)\mathrm{d}x=1$
$\delta(x)$	$\int\limits_{-\infty}^{x}\delta(x)\mathrm{d}x=H(x)$ (Heaviside)

Table 6.11: **Most common primitives**

6.3.2.7 Leibniz integral rule

In Calculus, "Leibniz's integral rule", named after Gottfried Leibniz, states that for an integral of the form:

$$I=\int\limits_{a(x)}^{b(x)} f(x,t)\mathrm{d}t \tag{6.3.603}$$

where $-\infty<a(x),b(x)<+\infty$, the derivative of this integral is expressible as:

$$\boxed{\frac{\mathrm{d}}{\mathrm{d}x}\left(\int\limits_{a(x)}^{b(x)} f(x,t)\mathrm{d}t\right)=f\left(x,b(x)\right)\frac{\mathrm{d}}{\mathrm{d}x}b(x)-f\left(x,a(x)\right)\frac{\mathrm{d}}{\mathrm{d}x}a(x)+\int\limits_{a(x)}^{b(x)}\frac{\partial}{\partial x}f(x,t)\mathrm{d}t} \tag{6.3.604}$$

This relation can be of use when evaluating certain definite integrals. When used in this context, the Leibniz rule for differentiating under the integral sign is also known as "Feynman's trick".

Notice that if $a(x)$ and $b(x)$ are constants rather than functions of x, we have a special case of Leibniz's rule:

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\int\limits_{a}^{b} f(x,t)\mathrm{d}t\right)=\int\limits_{a}^{b}\frac{\partial}{\partial x}f(x,t)\mathrm{d}t \tag{6.3.605}$$

Besides, if $a(x)=a$ and $b(x)=x$, which is a common situation as well (for example, in the proof of Cauchy's repeated integration formula), we have:

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\int_{a}^{x} f(x,t)\mathrm{d}t\right)=f\left(x,x\right)+\int_{a}^{x}\frac{\partial}{\partial x}f(x,t)\mathrm{d}t \tag{6.3.606}$$

Proof 6.14.8. First let us write:

$$I(\alpha)=\int\limits_{a(\alpha)}^{b(\alpha)} f(x,\alpha)\mathrm{d}x \tag{6.3.607}$$

We start from the very definition of the derivative:

$$\frac{\mathrm{d}I}{\mathrm{d}\alpha}=\lim_{\Delta\alpha\to 0}\frac{I(\alpha+\Delta\alpha)-I(\alpha)}{\Delta\alpha} \tag{6.3.608}$$

Now, since the integration limits depend (in general) on α, then a $\Delta\alpha$ will cause a Δa and a Δb

and so we have to write:

$$\begin{aligned}
I(\alpha+\Delta\alpha)-I(\alpha) &= \int\limits_{a+\Delta a}^{b+\Delta b} f(x,\alpha+\Delta\alpha)\mathrm{d}x-\int\limits_{a}^{b} f(x,\alpha)\mathrm{d}x\\
&= \left(\int\limits_{a+\Delta a}^{a}+\int\limits_{a}^{b}+\int\limits_{b}^{\mathrm{b}+\Delta b}\right) f(x,\alpha+\Delta\alpha)\mathrm{d}x-\int\limits_{a}^{b} f(x,\alpha)\mathrm{d}x\\
&= \int\limits_{a}^{b}\{f(x,\alpha+\Delta\alpha)-f(x,\alpha)\}\mathrm{d}x+\int\limits_{b}^{b+\Delta b} f(x,\alpha+\Delta\alpha)\mathrm{d}x\\
&\quad -\int\limits_{a}^{a+\Delta a} f(x,\alpha+\Delta\alpha)\mathrm{d}x
\end{aligned} \tag{6.3.609}$$

As $\Delta\alpha\rightarrow 0$ we have $\Delta a\rightarrow 0$ and $\Delta b\rightarrow 0$, too, and so:

$$\begin{aligned}
&\lim_{\Delta\alpha\to 0}\{I(\alpha+\Delta\alpha)-I(\alpha)\}\\
&\quad=\lim_{\Delta\alpha\to 0}\int\limits_{a}^{b}\{f(x,\alpha+\Delta\alpha)-f(x,\alpha)\}\mathrm{d}x+f(b,\alpha)\Delta b-f(a,\alpha)\Delta a
\end{aligned} \tag{6.3.610}$$

where the last two terms follow because as $\Delta a\rightarrow 0$ and $\Delta b\rightarrow 0$ the value of x over the entire integration interval remains practically unchanged at $x=a$ or at $x=b$ respectively. Thus:

$$\begin{aligned}
\frac{\mathrm{d}I}{\mathrm{d}\alpha}&=\lim_{\Delta\alpha\to 0}\frac{I(\alpha+\Delta\alpha)-I(\alpha)}{\Delta\alpha}=\lim_{\Delta\alpha\to 0}\frac{1}{\Delta\alpha}\int_a^b\{f(x,\alpha+\Delta\alpha)-f(x,\alpha)\}\mathrm{d}x\\
&\quad+\lim_{\Delta\alpha\to 0}f(b,\alpha)\frac{\Delta b}{\Delta\alpha}-\lim_{\Delta\alpha\to 0}f(a,\alpha)\frac{\Delta a}{\Delta\alpha}
\end{aligned} \tag{6.3.611}$$

or, taking the $1/\Delta\alpha$ inside the integral (the Riemann integral itself is defined as a limit, so what we are doing is reversing the order of two limiting operations, something a pure mathematician would want to justify but, as usual in this textbook, we won't worry about it!):

$$\frac{\mathrm{d}I}{\mathrm{d}\alpha}=\int\limits_{a}^{b}\frac{\partial f}{\partial\alpha}\mathrm{d}x+f(b,\alpha)\frac{\mathrm{d}b}{\mathrm{d}\alpha}-f(a,\alpha)\frac{\mathrm{d}a}{\mathrm{d}\alpha} \tag{6.3.612}$$

■ Q.E.D.

Differentiation under the integral sign is mentioned in the late physicist Richard Feynman's best-selling memoir *Surely You're Joking, Mr. Feynman!*in the chapter *A Different Box of Tools*. He describes learning it, while in high school, from an old text, *Advanced Calculus* (11926 according to holocene calendar), by Frederick S. Woods (who was a professor of mathematics in the Massachusetts Institute of Technology). The technique was not often taught when Feynman later received his formal education in Calculus, but using this technique, Feynman was able to solve otherwise difficult integration problems upon his arrival at graduate school at Princeton University:

«One thing I never did learn was contour integration. I had learned to do integrals by various methods shown in a book that my high school physics teacher Mr. Bader had given me. One day he told me to stay after class. "Feynman," he said, "you talk too much and you make too much noise. I know why. You're bored. So I'm going to give you a book. You go up there in the back, in the corner, and study this textbook, and when you know everything that's in this textbook, you can talk again." So every physics class, I paid no attention to what was going on with Pascal's Law, or whatever they were doing. I was up in the back with this textbook: "Advanced Calculus", by Woods. Bader knew I had studied "Calculus for the Practical Man" a little bit, so he gave me the real works. It was for a junior or senior course in college. It had Fourier series, Bessel functions, determinants, elliptic functions, all kinds of wonderful stuff that I didn't know anything about. That book also showed how to differentiate parameters under the integral sign, it's a certain operation. It turns out that's not taught very much in the universities; they don't emphasize it. But I caught on how to use that method, and I used that one damn tool again and again. So because I was self-taught using that book, I had peculiar methods of doing integrals. The result was, when guys at MIT or Princeton had trouble doing a certain integral, it was because they couldn't do it with the standard methods they had learned in school. If it was contour integration, they would have found it; if it was a simple series expansion, they would have found it. Then I come along and try differentiating under the integral sign, and often it worked. So I got a great reputation for doing integrals, only because my box of tools was different from everybody else's, and they had tried all their tools on it before giving the problem to me.»

✍**Example:**

Let us compute the following integral with variables limits:

$$I(\alpha)=\int\limits_{a(x)}^{b(x)} f(x,t)\mathrm{d}t=\int\limits_{0}^{\pi/2}\frac{\ln(1+\cos(\alpha)\cos(x))}{\cos(x)}\mathrm{d}x \qquad 0<\alpha<\pi \tag{6.3.613}$$

As both $a(x)$ and $b(x)$ are constants rather than functions of x, we will use the special case of Leibniz's integral rule:

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\int\limits_{a}^{b} f(x,t)\mathrm{d}t\right)=\int\limits_{a}^{b}\frac{\partial}{\partial x}f(x,t)\mathrm{d}t \tag{6.3.614}$$

Therefore differentiating under the integral with respect to α , we have (at the end we use one of the previous usual primitive and trigonometric relations):

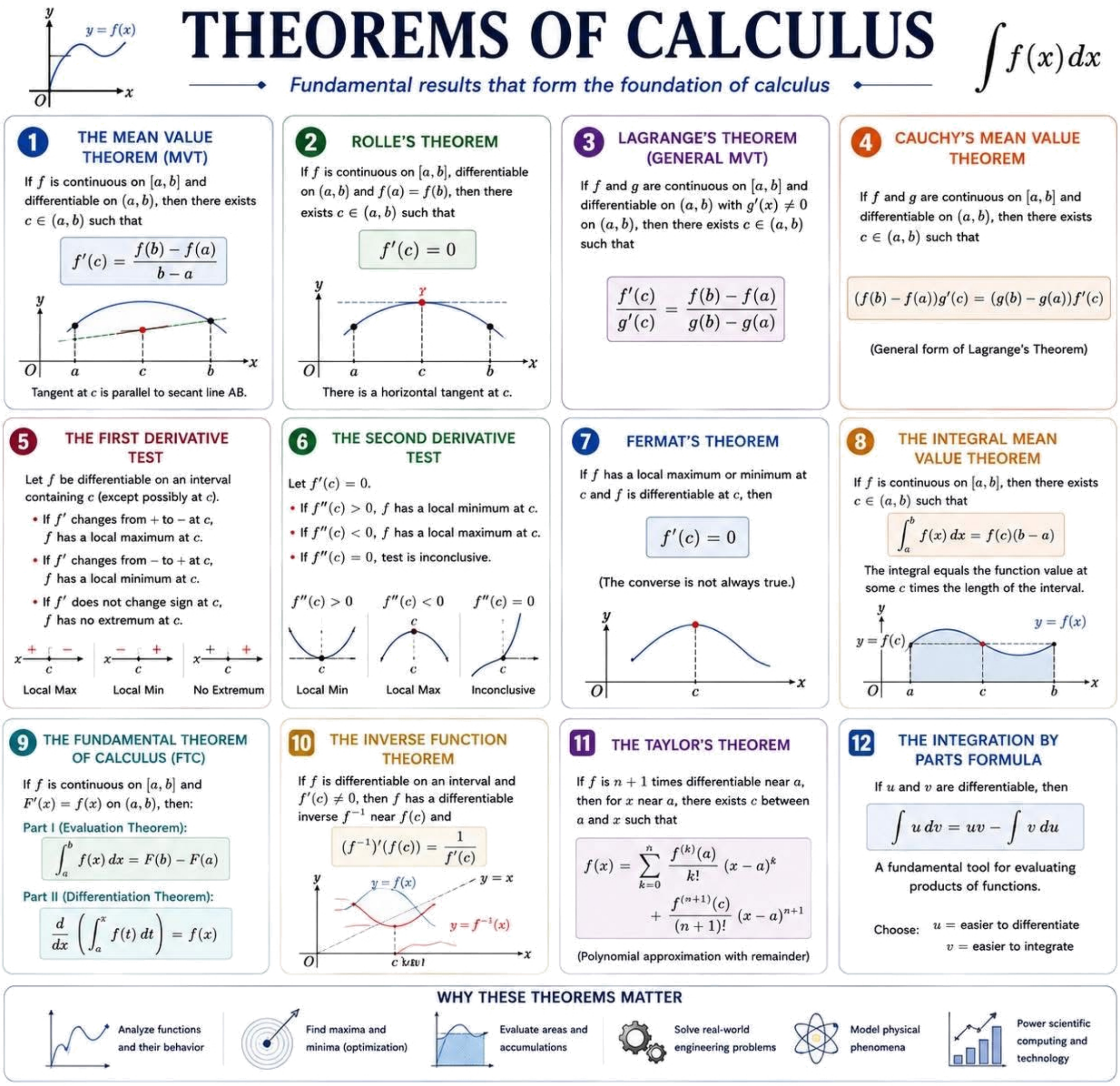
$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\alpha}I(\alpha) &= \int_0^{\pi/2}\frac{\partial}{\partial\alpha}\left(\frac{\ln(1+\cos(\alpha)\cos(x))}{\cos(x)}\right)\mathrm{d}x = -\int_0^{\pi/2}\frac{\sin(\alpha)}{1+\cos(\alpha)\cos(x)}\mathrm{d}x\\
&= -\int_0^{\pi/2}\frac{\sin(\alpha)}{\left(\cos^2\left(\frac{x}{2}\right)+\sin^2\left(\frac{x}{2}\right)\right)+\cos(\alpha)\left(\cos^2\left(\frac{x}{2}\right)-\sin^2\left(\frac{x}{2}\right)\right)}\mathrm{d}x\\
&= -\frac{\sin(\alpha)}{1-\cos(\alpha)}\int_0^{\pi/2}\frac{1}{\cos^2\left(\frac{x}{2}\right)}\frac{1}{\frac{1+\cos(\alpha)}{1-\cos(\alpha)}+\tan^2\left(\frac{x}{2}\right)}\mathrm{d}x\\
&= -\frac{2\sin(\alpha)}{1-\cos(\alpha)}\int_0^{\pi/2}\frac{\frac{1}{2}\sec^2\left(\frac{x}{2}\right)}{\frac{2\cos^2\left(\frac{\alpha}{2}\right)}{2\sin^2\left(\frac{\alpha}{2}\right)}+\tan^2\left(\frac{x}{2}\right)}\mathrm{d}x\\
&= -\frac{2\left(2\sin\left(\frac{\alpha}{2}\right)\cos\left(\frac{\alpha}{2}\right)\right)}{2\sin^2\left(\frac{\alpha}{2}\right)}\int_0^{\pi/2}\frac{1}{\cot^2\left(\frac{\alpha}{2}\right)+\tan^2\left(\frac{x}{2}\right)}\mathrm{d}\left(\tan\left(\frac{x}{2}\right)\right)\\
&= -2\cot\left(\frac{\alpha}{2}\right)\int_0^{\pi/2}\frac{1}{\cot^2\left(\frac{\alpha}{2}\right)+\tan^2\left(\frac{x}{2}\right)}\mathrm{d}\left(\tan\left(\frac{x}{2}\right)\right)\\
&= -2\arctan\left(\frac{\tan\left(\frac{x}{2}\right)}{\cot\left(\frac{\alpha}{2}\right)}\right)\Bigg|_0^{\pi/2} = -2\arctan\left(\tan\left(\frac{x}{2}\right)\tan\left(\frac{\alpha}{2}\right)\right)\Bigg|_0^{\pi/2}\\
&= -2\arctan\left(\tan\left(\frac{\alpha}{2}\right)\right) = -\alpha
\end{aligned} \tag{6.3.615}$$

Therefore:

$$I(\alpha) = -\frac{\alpha^2}{2}+c^{te} \tag{6.3.616}$$

But $I(\pi/2)=0$ by definition, so we must have $c^{te}=\pi^2/8$ and finally:

$$I(\alpha) = -\frac{\alpha^2}{2}+\frac{\pi^2}{8} \tag{6.3.617}$$



THEOREMS OF CALCULUS

Fundamental results that form the foundation of calculus

$\int f(x)\,dx$

① THE MEAN VALUE THEOREM (MVT)

If f is continuous on $[a,b]$ and differentiable on (a,b), then there exists $c\in(a,b)$ such that

$$f'(c)=\frac{f(b)-f(a)}{b-a}$$

Tangent at c is parallel to secant line AB.

② ROLLE'S THEOREM

If f is continuous on $[a,b]$, differentiable on (a,b) and $f(a)=f(b)$, then there exists $c\in(a,b)$ such that

$$f'(c)=0$$

There is a horizontal tangent at c.

③ LAGRANGE'S THEOREM (GENERAL MVT)

If f and g are continuous on $[a,b]$ and differentiable on (a,b) with $g'(x)\neq 0$ on (a,b), then there exists $c\in(a,b)$ such that

$$\frac{f'(c)}{g'(c)}=\frac{f(b)-f(a)}{g(b)-g(a)}$$

④ CAUCHY'S MEAN VALUE THEOREM

If f and g are continuous on $[a,b]$ and differentiable on (a,b), then there exists $c\in(a,b)$ such that

$$(f(b)-f(a))g'(c)=(g(b)-g(a))f'(c)$$

(General form of Lagrange's Theorem)

⑤ THE FIRST DERIVATIVE TEST

Let f be differentiable on an interval containing c (except possibly at c).

- If f' changes from $+$ to $-$ at c, f has a local maximum at c.
- If f' changes from $-$ to $+$ at c, f has a local minimum at c.
- If f' does not change sign at c, f has no extremum at c.

Local Max Local Min No Extremum

⑥ THE SECOND DERIVATIVE TEST

Let $f'(c)=0$.

- If $f''(c)>0$, f has a local minimum at c.
- If $f''(c)<0$, f has a local maximum at c.
- If $f''(c)=0$, test is inconclusive.

$f''(c)>0$ $f''(c)<0$ $f''(c)=0$

Local Min Local Max Inconclusive

⑦ FERMAT'S THEOREM

If f has a local maximum or minimum at c and f is differentiable at c, then

$$f'(c)=0$$

(The converse is not always true.)

⑧ THE INTEGRAL MEAN VALUE THEOREM

If f is continuous on $[a,b]$, then there exists $c\in(a,b)$ such that

$$\int_a^b f(x)\,dx=f(c)(b-a)$$

The integral equals the function value at some c times the length of the interval.

⑨ THE FUNDAMENTAL THEOREM OF CALCULUS (FTC)

If f is continuous on $[a,b]$ and $F'(x)=f(x)$ on (a,b), then:

Part I (Evaluation Theorem):

$$\int_a^b f(x)\,dx=F(b)-F(a)$$

Part II (Differentiation Theorem):

$$\frac{d}{dx}\left(\int_a^x f(t)\,dt\right)=f(x)$$

10 THE INVERSE FUNCTION THEOREM

If f is differentiable on an interval and $f'(c)\neq 0$, then f has a differentiable inverse f^{-1} near $f(c)$ and

$$(f^{-1})'(f(c))=\frac{1}{f'(c)}$$

11 THE TAYLOR'S THEOREM

If f is $n+1$ times differentiable near a, then for x near a, there exists c between a and x such that

$$f(x)=\sum_{k=0}^{n}\frac{f^{(k)}(a)}{k!}(x-a)^k+\frac{f^{(n+1)}(c)}{(n+1)!}(x-a)^{n+1}$$

(Polynomial approximation with remainder)

12 THE INTEGRATION BY PARTS FORMULA

If u and v are differentiable, then

$$\int u\,dv=uv-\int v\,du$$

A fundamental tool for evaluating products of functions.

Choose: $u=$ easier to differentiate
$v=$ easier to integrate

WHY THESE THEOREMS MATTER

Analyze functions and their behavior

Find maxima and minima (optimization)

Evaluate areas and accumulations

Solve real-world engineering problems

Model physical phenomena

Power scientific computing and technology

6.3.2.8 Integral representation of first kind Bessel's function

A particularly useful and powerful way of treating Bessel functions employs their integral representation as we will see in the section of Wave Optics page 4521.

Remember that in the section of Sequences and Series we have proved that the generating function of Bessel's function was (see page 1791):

$$\begin{aligned}
e^{(x/2)(t-1/t)}G(x,t)&=\sum_{n=-\infty}^{+\infty}\sum_{s=\max(0,-n)}^{+\infty}\frac{(-1)^s}{(n+s)!s!}\left(\frac{x}{2}\right)^{n+2s}t^n\\
&=\sum_{n=-\infty}^{+\infty}\left[\sum_{s=\max(0,-n)}^{+\infty}\frac{(-1)^s}{(n+s)!s!}\left(\frac{x}{2}\right)^{n+2s}\right]t^n
\end{aligned}\qquad(6.3.618)$$

That is:

$$G(x,t)=\sum_{n=-\infty}^{+\infty}J_n(x)t^n \tag{6.3.619}$$

Now remember that (see section Numbers page 251):

$$e^{\mathrm{i}\theta}=\cos(\theta)+\mathrm{i}\sin(\theta) \tag{6.3.620}$$

So if we return to the generating function, and substitute $t=e^{\mathrm{i}\theta}$, we get:

$$\begin{aligned}e^{(x/2)(t-1/t)}&=e^{(x/2)(e^{\mathrm{i}\theta}-e^{-\mathrm{i}\theta})})=e^{(x/2)2\mathrm{i}\sin(x)}=e^{\mathrm{i}x\sin(x)}=\sum_{n=-\infty}^{+\infty}J_n(x)\left(e^{\mathrm{i}\theta}\right)^n=\sum_{n=-\infty}^{+\infty}J_n(x)e^{n\mathrm{i}\theta}\\&=J_0(x)+2\left[J_2(x)\cos(2\theta)+J_4(x)\cos(4\theta)+\ldots\right]+2\mathrm{i}\left[J_1(x)\sin(\theta)+J_3(x)\sin(3\theta)+\ldots\right]\end{aligned} \tag{6.3.621}$$

In which to condensate the result we have used first the property proved during our study of the generating function of Bessel's functions:

$$J_n(x)=(-1)^nJ_{-n}(x) \tag{6.3.622}$$

and also:

$$J_1(x)e^{\mathrm{i}\theta}+J_{-1}(x)e^{-\mathrm{i}\theta}J_1(x)\left(e^{\mathrm{i}\theta}-e^{-\mathrm{i}\theta}\right)=2\mathrm{i}J_1(x)\sin(\theta)J_2(x)e^{2\mathrm{i}\theta}+J_{-2}(x)e^{-2\mathrm{i}\theta}=2J_2(x)\cos(2\theta) \tag{6.3.623}$$

and so on...

Now remember that:

$$e^{\mathrm{i}x\sin(x)}=\cos(x\sin(\theta))+\mathrm{i}\sin(x\sin(\theta)) \tag{6.3.624}$$

Therefore identifying real and imaginary part, we get:

$$\begin{aligned}\cos(x\sin(\theta))&=J_0(x)+2\left[J_2(x)\cos(2\theta)+J_4(x)\cos(4\theta)+\ldots\right]=J_0(x)+2\sum_{n=1}^{+\infty}J_{2n}(x)\cos(2n\theta)\\\sin(x\sin(\theta))&=2\left[J_1(x)\sin(\theta)+J_3(x)\sin(3\theta)+\ldots\right]=2\sum_{n=1}^{+\infty}J_{2n-1}\sin((2n-1)\theta)\end{aligned} \tag{6.3.625}$$

Remember also that we have proved during our study of Fourier series in the section of Sequences and Series (see page 1727) that:

$$(\mathrm{I})\begin{cases}\displaystyle\int_0^T\cos(nx)\cos(kx)\mathrm{d}x=0\\\displaystyle\int_0^T\cos(nx)\sin(kx)\mathrm{d}x=0\\\displaystyle\int_0^T\sin(nx)\sin(kx)\mathrm{d}x=0\end{cases}\qquad(\mathrm{II})\begin{cases}\displaystyle\int_0^T\cos^2(kx)\mathrm{d}x=\frac{T}{2}\\\displaystyle\int_0^T\cos(kx)\sin(kx)\mathrm{d}x=0\\\displaystyle\int_0^T\sin^2(kx)\mathrm{d}x=\frac{T}{2}\end{cases} \tag{6.3.626}$$

with $n,k\in\mathbb{N}$ and $n\neq k$ with $n,k\in\mathbb{N}$ and $n=k$

That is:

$$\int\limits_0^{\pi}\cos(n\theta)\cos(m\theta)\mathrm{d}\theta=\frac{\pi}{2}\delta_{nm}\qquad\int\limits_0^{\pi}\sin(n\theta)\sin(m\theta)\mathrm{d}\theta=\frac{\pi}{2}\delta_{nm}\tag{6.3.627}$$

where δ_{nm} is the Kronecker symbol (see section Tensor Calculus page 2044).

Now let us write:

$$\begin{gathered}\int\limits_0^{\pi}\left(J_0(x)+2\sum_{n=1}^{+\infty}J_{2n}(x)\cos(2n\theta)\right)\cos(n\theta)\mathrm{d}\theta=\int\limits_0^{\pi}\cos(x\sin(\theta))\cos(n\theta)\mathrm{d}\theta\\ \int\limits_0^{\pi}\left(2\sum_{n=1}^{+\infty}J_{2n-1}\sin((2n-1)\theta)\right)\sin(n\theta)\mathrm{d}\theta=\int\limits_0^{\pi}\sin(x\sin(\theta))\sin(n\theta)\mathrm{d}\theta\end{gathered}\tag{6.3.628}$$

Let us focus on the first integral:

$$\int\limits_0^{\pi}J_0(x)\cos(n\theta)\mathrm{d}\theta+2\left(\int\limits_0^{\pi}J_2(x)\cos(2\theta)\cos(n\theta)\mathrm{d}\theta+\int\limits_0^{\pi}J_4(x)\cos(4\theta)\cos(n\theta)\mathrm{d}\theta+\ldots\right)\tag{6.3.629}$$

So we see above that whatever the value for any n, except of $n=0$:

$$\int\limits_0^{\pi}J_0(x)\cos(n\theta)\mathrm{d}\theta=J_0(x)\int\limits_0^{\pi}\cos(n\theta)\mathrm{d}\theta=0\tag{6.3.630}$$

Therefore it only remains for $n>0$:

$$2\left(\int\limits_0^{\pi}J_2(x)\cos(2\theta)\cos(n\theta)\mathrm{d}\theta+\int\limits_0^{\pi}J_4(x)\cos(4\theta)\cos(n\theta)\mathrm{d}\theta+\int\limits_0^{\pi}J_6(x)\cos(6\theta)\cos(n\theta)\mathrm{d}\theta+\ldots\right)\tag{6.3.631}$$

and we see above that if $n>0$ is odd all the integrals vanish but if $n>0$ is even, only the corresponding $J_n(x)$ remains!

Exactly the same analysis can be done for:

$$\int\limits_0^{\pi}\left(2\sum_{n=1}^{+\infty}J_{2n-1}\sin((2n-1)\theta)\right)\sin(n\theta)\mathrm{d}\theta=\int\limits_0^{\pi}\sin(x\sin(\theta))\sin(n\theta)\mathrm{d}\theta\tag{6.3.632}$$

We have therefore for each of the integrals above, especially for each the left term that (recall that $n=0,2,4,\ldots$ is even and $n=1,3,5,\ldots$ is odd):

$$\begin{gathered}\frac{1}{\pi}\int\limits_0^{\pi}\cos(x\sin(\theta))\cos(n\theta)\mathrm{d}\theta=\begin{cases}J_n(x) & n\text{ is even}\\ 0 & n\text{ is odd}\end{cases}\\ \frac{1}{\pi}\int\limits_0^{\pi}\sin(x\sin(\theta))\sin(n\theta)\mathrm{d}\theta=\begin{cases}0 & n\text{ is even}\\ J_n(x) & n\text{ is odd}\end{cases}\end{gathered}\tag{6.3.633}$$

If these two equations are added together we have using trigonometric identities (see section Trigonometry page 2519):

$$\begin{aligned} J_n(x) &= \frac{1}{\pi}\int_0^{\pi}(\cos(x\sin(\theta))\cos(n\theta)+\sin(x\sin(\theta))\sin(n\theta))\mathrm{d}\theta \\ &= \frac{1}{\pi}\cos(n\theta - x\sin(\theta))\mathrm{d}\theta \end{aligned} \tag{6.3.634}$$

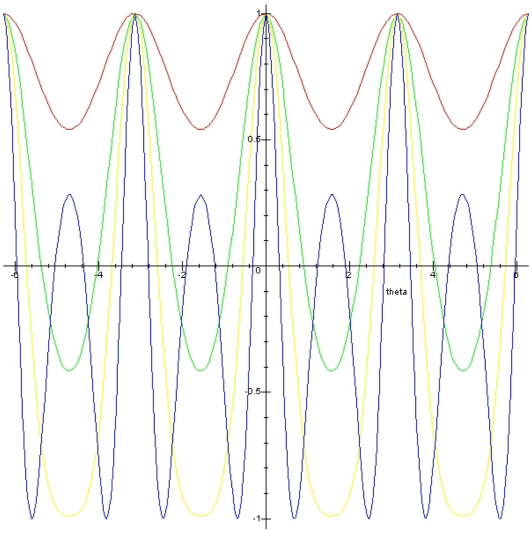
for $n = 0, 1, 2, 3, \ldots$.

If we put $n = 0$ in the above relation, we get:

$$J_0(x) = \frac{1}{\pi}\cos(-x\sin(\theta))\mathrm{d}\theta = \frac{1}{\pi}\cos(x\sin(\theta))\mathrm{d}\theta \tag{6.3.635}$$

If we plot $\cos(x\sin(\theta))$ we see that it repeats itself in all four quadrants (it's an even function):

```
>plot([cos(sin(theta)),cos(2*sin(theta)),cos(3*sin(theta)),cos(5*sin(theta))]
,theta=-2*Pi..2*Pi);
```



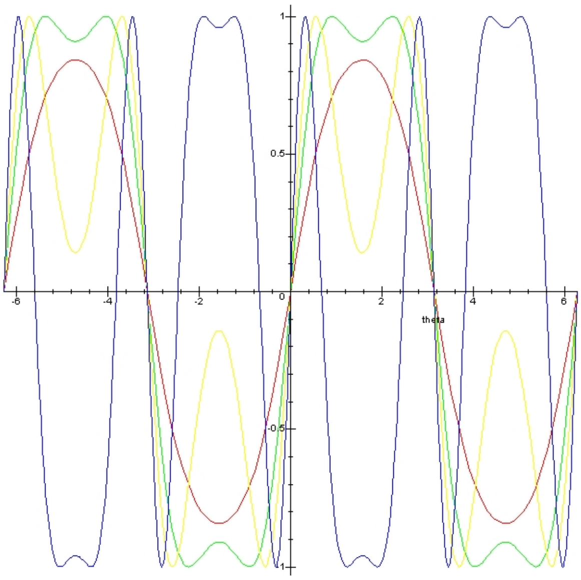
So we can write:

$$\boxed{J_0(x) = \frac{1}{2\pi}\int_0^{2\pi}\cos(x\sin(\theta))\mathrm{d}\theta} \tag{6.3.636}$$

This is the real integral representation of the zero order Bessel function of the first type.

But in many developments we don't use the above expression as there is not phasor that is visible. So the trick is to notice that $\sin(x\sin(\theta))$ reverses its sign in the third a fourth quadrant (it's and odd function):

```
>plot([sin(sin(theta)),sin(2*sin(theta)),sin(3*sin(theta)),sin(5*sin(theta))],
theta=-2*Pi..2*Pi);
```



So we have:

$$\frac{1}{2\pi}\int\limits_0^{2\pi}\sin(x\sin(\theta))\mathrm{d}\theta=0 \tag{6.3.637}$$

Adding the both relations by multiplying the second by i, we have:

$$\begin{aligned}J_0(x)&=\frac{1}{\pi}\int\limits_0^{2\pi}\cos(x\sin(\theta))\mathrm{d}\theta+0=\frac{1}{\pi}\int\limits_0^{2\pi}\cos(x\sin(\theta))\mathrm{d}\theta+\mathrm{i}\frac{1}{2\pi}\int\limits_0^{2\pi}\sin(x\sin(\theta))\mathrm{d}\theta\\&=\frac{1}{2\pi}\int\limits_0^{2\pi}\left(\cos(x\sin(\theta))+\mathrm{i}\sin(x\sin(\theta))\right)\mathrm{d}\theta\end{aligned} \tag{6.3.638}$$

Finally we get the complex representation of the zero order Bessel function of the first type:

$$J_0(x)=\frac{1}{2\pi}\int\limits_0^{2\pi}e^{\mathrm{i}x\sin(\theta)}\mathrm{d}\theta \tag{6.3.639}$$

Let us do a change of variable $\theta=\varphi+\pi/2$, then:

$$J_0(x)=\frac{1}{2\pi}\int\limits_{-\pi/2}^{3/4\pi}e^{\mathrm{i}x\sin(\varphi+\pi/2)}\mathrm{d}\varphi=\frac{1}{2\pi}\int\limits_{-\pi/2}^{3/4\pi}e^{\mathrm{i}x\cos(\varphi)}\mathrm{d}\varphi \tag{6.3.640}$$

But we have proved earlier that for any periodic function:

$$\int\limits_a^{T+a}f(t)\mathrm{d}t=\int\limits_0^T f(t)\mathrm{d}t \tag{6.3.641}$$

Therefore:

$$\boxed{J_0(x)=\frac{1}{2\pi}\int\limits_0^{2\pi} e^{\mathrm{i}x\cos(\varphi)}\mathrm{d}\varphi}\tag{6.3.642}$$

This integral representation my be obtained in various ways but this one seems the most easy one to us. Many other integral representation exists.

6.3.2.9 Dirac Function

The "Dirac function", also named "Dirac peak" or "delta function", plays a very important practical role both in electronic and computer also wave mechanics and quantum field theory (this allows to discretize a continuum!) as well as in the field of civil engineering (see the sections of the same name for examples).

Before going further we could notice that it is wrong to speak about a "function" because a function is an application of a start set (usually the set of real or complex number with one or more dimensions) in an arrival set (usually the set of real or complex numbers in one or more dimensions). While the domain of definition of the Dirac function is not a set of numbers but strictly speaking a set of functions!

More technically the Dirac delta function, or δ function, is a generalized function, or distribution, on the real number line that is zero everywhere except at zero, with an integral of one over the entire real line. The delta function is sometimes thought of as an infinitely high, infinitely thin spike at the origin, with total area one under the spike, and physically represents the density of an idealized point mass or point charge. It was introduced by theoretical physicist Paul Dirac. In the context of signal processing it is often referred to as the "unit impulse symbol" (or function). Its discrete analogue is the "Kronecker delta function", which is usually defined on a discrete domain and takes values 0 and 1.

As always in this textbook we will focus here only on the properties we will need to study Applied Mathematics stuffs of other sections of the book.

To represent mentally in an easy way this function, first consider the function defined by:

$$\begin{aligned}f(x)&=\frac{1}{a}\quad\text{if}\quad -\frac{a}{2}<x<\frac{a}{2}\\ f(x)&=0\quad\text{if}\qquad\quad |x|>\frac{a}{2}\end{aligned}\tag{6.3.643}$$

The representation of $y=f(x)$ above is a rectangle of width a, and of height $1/a$ and unit surface. The Dirac function can be considered as the boundary when $a\rightarrow 0$ of the $f(x)$. So we have:

$$\boxed{\begin{aligned}\delta(x)&=0 &&\text{if}\quad x\neq 0\\ \delta(x)&=+\infty &&\text{if}\quad x=0\end{aligned}}\tag{6.3.644}$$

That is to say (the fact that the arrow stops at 1 is a tradition but its misleading because in reality it should stop at infinity!!! what you see below is the integral of the Dirac function as we will discuss later below):

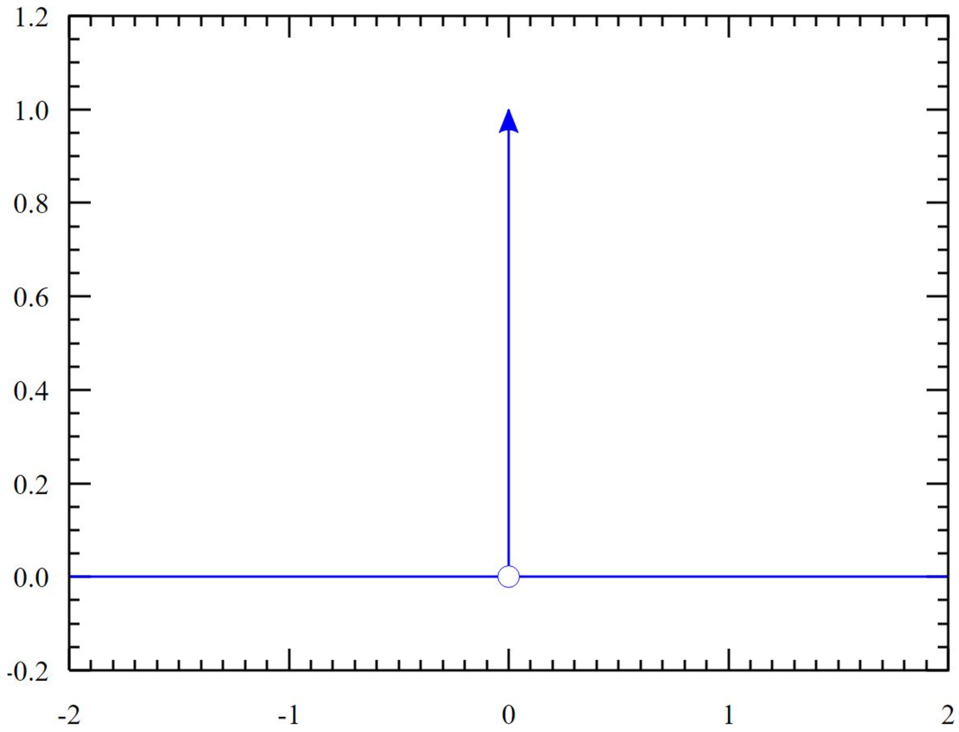


Figure 6.57: Schematic representation of the Dirac delta function by a line surmounted by an arrow (source: Wikipedia)

with:

$$\int_{-\infty}^{+\infty}\delta(x)\mathrm{d}x=\int_{-\varepsilon}^{+\varepsilon}\delta(x)\mathrm{d}x=1 \tag{6.3.645}$$

where ε is a number greater than 0 and as small as we want.

Remark

As the reader will have probably noticed it when we introduced the initial function $f(x)$, the resulting delta Dirac function has therefore the dimension of the inverse of a length!

For a function $g(x)$ continuous on $x=0$, we have:

$$\delta(g)=\int_{-\infty}^{+\infty}g(x)\delta(x)\mathrm{d}x=g(0) \tag{6.3.646}$$

By extension we have:

$$\boxed{\begin{array}{ll}\delta(x-x_0)=0 & \text{if}\quad x\neq x_0\\ \delta(x-x_0)=+\infty & \text{if}\quad x=x_0\end{array}} \tag{6.3.647}$$

and for a function $g(x)$ continue on x_0:

$$\int_{-\infty}^{+\infty}g(x)\delta(x-x_0)\mathrm{d}x=g(x_0) \tag{6.3.648}$$

Or more generally we have the following "regularized delta function":

$$\int_{\alpha}^{\beta} g(x)\delta(x-x_0)\mathrm{d}x = \begin{cases} g(x_0) & \text{if } \alpha < x_0 < \beta \\ 0 & \text{otherwise} \end{cases} \tag{6.3.649}$$

For example:

$$\int_{1}^{6} (3x-1)\delta(x-2)\mathrm{d}x = 5 \tag{6.3.650}$$

It is then relatively easy to define the Dirac function in 3-dimensional space by:

$$\delta(\vec{r}-\vec{r}_0) = \delta(x-x_0)\delta(y-y_0)\delta(y-y_0) \tag{6.3.651}$$

As already mentioned earlier, we will prove the properties of the Dirac function only if we will need them in other sections of this textbook.

To see now the integral representation of the Dirac function (very useful in Quantum Field Theory) let us recall the Fourier transform of a function $f(t)$ (see section Sequences and Series page 1758) using another common notation and convention used by physicists and mathematicians[6] (some of them invert the definition... i.e. the Fourier transform is designated as the Inverse Fourier transform...):

$$\tilde{f}(s) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(t)e^{\mathrm{i}t\omega}\mathrm{d}t \tag{6.3.652}$$

This transform is reversible, i.e., you can go back from $\tilde{f}(s)$ to $f(t)$ by:

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \tilde{f}(\omega)e^{-\mathrm{i}t\omega}\mathrm{d}\omega \tag{6.3.653}$$

If we set $f(t) = \delta(t)$ in the above equations, we find:

$$\tilde{\delta}(s) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \delta(t)e^{\mathrm{i}t\omega}\mathrm{d}t = \frac{1}{\sqrt{2\pi}} \tag{6.3.654}$$

$$\delta(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\mathrm{i}t\omega}\mathrm{d}\omega = \int_{-\infty}^{+\infty} \frac{e^{-\mathrm{i}t\omega}}{2\pi}\mathrm{d}\omega \tag{6.3.655}$$

In other words, the delta function and a constant $1/\sqrt{2\pi}$ are Fourier-transform of each other.

---

[6] Instead of using the limit of ever-narrowing rectangular pulse of unit area when defining delta function, any similar functions can be used, provided their integral is unity and their amplitude increase as their pulse-like property narrows.

Another way to see the integral representation of the delta function is again using the limits. For example, using the limit of the Gaussian (and the fact that the Fourier transform of a Gaussian is another Gaussian as proven at page 1765):

$$\delta(t)=\lim_{\sigma\to 0}\frac{1}{\sqrt{2\pi}\sigma}e^{-t^2/2\sigma^2}=\lim_{\sigma\to 0}\int\limits_{-\infty}^{+\infty}\frac{1}{2\pi}e^{-\omega^2\sigma^2/2}e^{-\mathrm{i}\omega t}\mathrm{d}\omega=\int\limits_{-\infty}^{+\infty}\frac{e^{-\mathrm{i}\omega t}}{2\pi}\mathrm{d}\omega \qquad (6.3.656)$$

The first equality is often denoted:

$$\delta(x-a)=\lim_{k\to 0}\frac{1}{k\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-a}{k}\right)^2} \qquad (6.3.657)$$

and depicted with the following figure (for $a=0$):

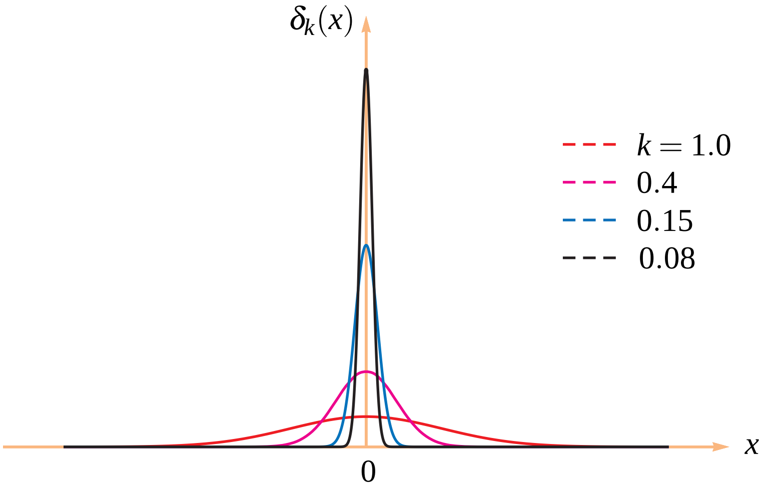


Figure 6.58: **Regularized Gaussian delta Dirac function**

Or more generally:

$$\boxed{\delta(t-x_0)=\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty}e^{-\mathrm{i}\omega(t-x_0)}\mathrm{d}\omega} \qquad (6.3.658)$$

Notice that this definition leads immediately to $\delta(0)=+\infty$.

6.3.2.9.1 Heaviside function

Remember that we have seen just earlier the following property:

$$\int\limits_{\alpha}^{\beta}g(x)\delta(x-x_0)\mathrm{d}x=\begin{cases}g(x_0) & \text{if } \alpha<x_0<\beta\\ 0 & \text{otherwise}\end{cases} \qquad (6.3.659)$$

Notice that the above finite bounded integral leads to:

$$\int_a^b g(x)\delta(x)\mathrm{d}x=\begin{cases}g(0) & \text{if } 0\in[a,b]\\ 0 & \text{if } 0\notin[a,b]\end{cases} \tag{6.3.660}$$

and also to (we put $g(x)=1$):

$$\int_a^b \delta(x)\mathrm{d}x=\begin{cases}1 & \text{if } 0\in[a,b]\\ 0 & \text{if } 0\notin[a,b]\end{cases} \tag{6.3.661}$$

Now, let us put $a=-\infty$ and $b=x$, therefore for every $x\in\mathbb{R}$:

$$\int_{-\infty}^x \delta(x)\mathrm{d}x=\begin{cases}1 & \text{if } 0\in]-\infty,x]\\ 0 & \text{if } 0\notin]-\infty,x]\end{cases} \tag{6.3.662}$$

the latter is equivalent to write:

$$\int_{-\infty}^x \delta(x)\mathrm{d}x=\begin{cases}1 & \text{if } x>0\\ 0 & \text{if } x<0\end{cases} \tag{6.3.663}$$

And now we can do something clever, as $\mathbb{R}$ is dense, we can therefore define a new function named the "Heaviside function" as a function for which every value is computed as the semi-bounded Dirac integral such that:

$$\boxed{H(x)=\int_{-\infty}^x \delta(x-a)dx} \tag{6.3.664}$$

So each valuated point of the Heaviside function $H(x)$ is in fact the evaluation of a semi-bounded Dirac function estimated on point x.

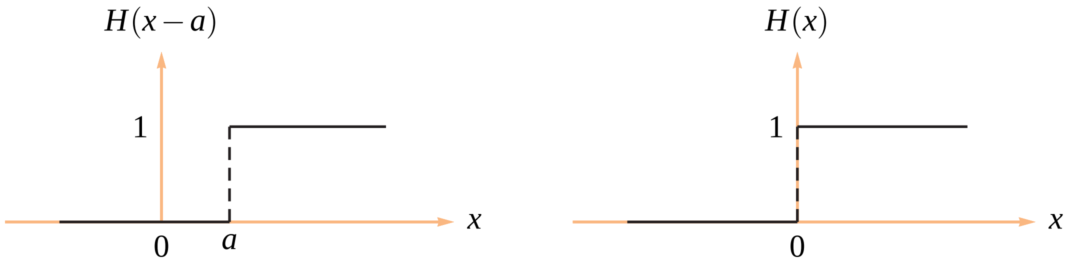


Figure 6.59: **Plot of the Heaviside function**

Differentiation of the above definition with respect to x, yields the following relation:

$$\frac{\mathrm{d}H(x)}{\mathrm{d}x}=\delta(x) \tag{6.3.665}$$

6.3.2.10 Gamma Euler Function

We define the Euler Gamma function (Eulerian integral of the second kind), more commonly named "upper incomplete Gamma function, by the following integral:

$$\boxed{\Gamma_{-1}(x)=\int\limits_0^{+\infty} t^{x-1}e^{-t}\mathrm{d}t} \tag{6.3.666}$$

with x belonging to the set of complex numbers whose real part is positive and non-zero (thus the positive real number are also included in the domain of definition)! Indeed, if we take complex numbers with a zero or negative real part, the integral diverges and is then undefined!

Remark

We have already met this integral and some of its properties (which will be proved here) in our study of the Beta, Gamma, Chi-square, Fisher and Student statistical distribution functions (see section Statistics page 647). We will also use this integral in maintenance (see section Industrial Engineering page 8716), in String theory and other engineering and physics fields (see the corresponding chapters) as for the canonical negative binomial generalized linear regression (see section Numerical Methods page 6316).

Consider we want to solve a quite common related practical case:

$$\int\limits_0^{+\infty} x^{t-1}\,e^{-c^{te}x}\mathrm{d}x \tag{6.3.667}$$

Set:

$$y=c^{te}x \qquad \mathrm{d}y=c^{te} \qquad \mathrm{d}x=\frac{y}{c^{te}} \tag{6.3.668}$$

Hence:

$$\int\limits_0^{+\infty}\left(\frac{y}{c^{te}}\right)^{t-1}\,e^{-y}\,\frac{\mathrm{d}y}{c^{te}} \tag{6.3.669}$$

Arranging we get:

$$\frac{1}{(c^{te})^t}\int\limits_0^{+\infty} y^{t-1}\,e^{-y}\,\mathrm{d}y \tag{6.3.670}$$

That is:

$$\int\limits_0^{+\infty} x^{t-1}\,e^{-c^{te}x}\mathrm{d}x=\frac{\Gamma_{-1}(t)}{(c^{te})^t} \tag{6.3.671}$$

Here is a graphical plot of the module of the Euler Gamma function $\Gamma_{-1}(x)$ for x browsing an interval of real numbers (take care in Maple™ 4.00b to write GAMMA capitalized!!!):

```
>with(plots):
>plot(GAMMA(x),x=-Pi..Pi,y=-5..5);
```

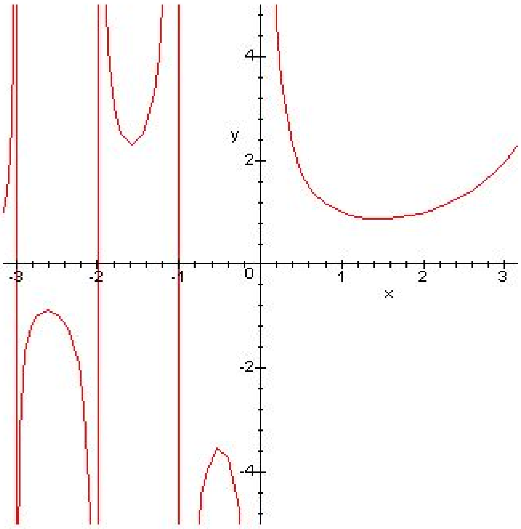
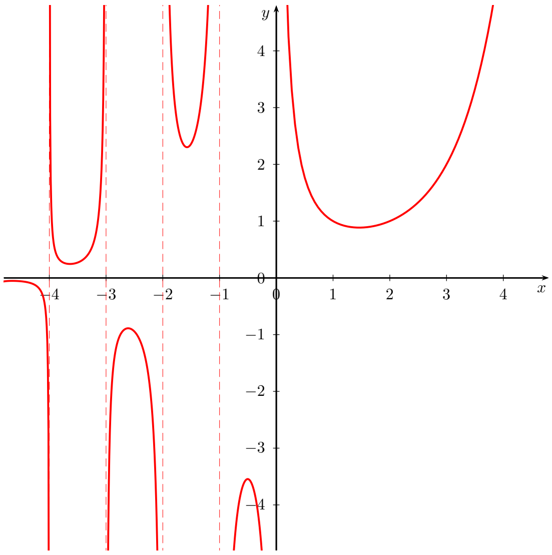


Figure 6.60: **Plot of the Euler Gamma function in Maple™ 4.00b**

Or in a much better quality with LuaLaTeX/PST:



and always the same function in Maple™ 4.00b but now in the complex plane and always with in ordinate the module of the Gamma Euler function (we can see that there are singularities every time x is a negative integer or zero!):

```
>with(plots):
>plot3d(abs(GAMMA(x+y*I)),x=-Pi..Pi,y=-Pi..Pi,view=0..5,
grid=[30,30],orientation=[-120,45],axes=frame,style=patchcontour);
```

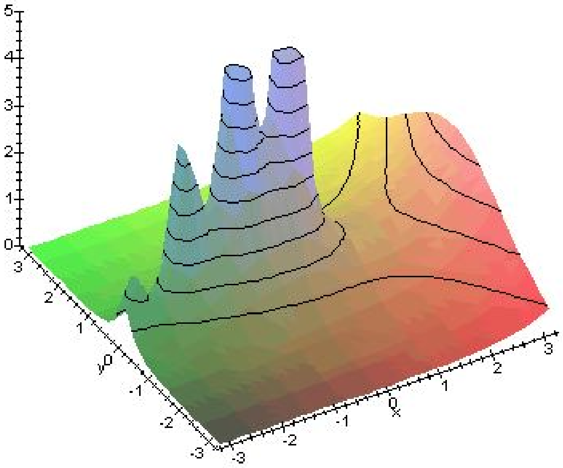


Figure 6.61: **Plot of the Euler Gamma function in the complex plane in Maple™ 4.00b**

This function is interesting if we impose the variable x to belong to the set of integer numbers and that we write it as follows:

$$\Gamma_0(x)=\int\limits_0^{+\infty} t^x e^{-t}\mathrm{d}t \tag{6.3.672}$$

Let us integrate by part the latter function:

$$\begin{aligned}\Gamma_0(x)&=\int\limits_0^{+\infty} t^x e^{-t}\mathrm{d}t=-t^x e^{-t}\Big|_0^{+\infty}+x\int\limits_0^{+\infty} t^{x-1}e^{-t}\mathrm{d}t\\&=-t^x e^{-t}\Big|_0^{+\infty}+x\Gamma_0(x-1)\end{aligned} \tag{6.3.673}$$

Since the exponential function decreases much faster than t^x then we have:

$$\Gamma_0(x)=x\Gamma_0(x-1) \tag{6.3.674}$$

In literature, we frequently find the following notations (there are confusing):

$$\begin{aligned}\Gamma_{-1}(x)=\Gamma_0(x-1)=\Gamma(x)&=\int\limits_0^{+\infty} t^{x-1}e^{-t}\mathrm{d}t\\ \Gamma_0(x)=\Gamma_0(x+1)=\Gamma(x)&=\int\limits_0^{+\infty} t^x e^{-t}\mathrm{d}t\end{aligned} \tag{6.3.675}$$

Which brings us to write the result in a more traditional form:

$$\boxed{\Gamma(x+1)=x\Gamma(x)} \tag{6.3.676}$$

From the relation $\Gamma_0(x)=x\Gamma_0(x-1)$, it comes by induction:

$$\Gamma_0(x)=x(x-1)\Gamma_0(x-2)=\ldots=\Gamma_0(0)x! \tag{6.3.677}$$

But:

$$\Gamma_0(0)=\Gamma(1)=\int\limits_0^{+\infty} t^0e^{-t}\mathrm{d}t=\int\limits_0^{+\infty} e^{-t}\mathrm{d}t=-e^{-t}\Big|_0^{+\infty}=-e^{-\infty}-(-e^{-0})=1 \tag{6.3.678}$$

That gives:

$$\Gamma_0(x)=x\cdot(x-1)\cdot\Gamma_0(x-2)=\ldots=\Gamma_0(0)x!=1\cdot x!=x! \tag{6.3.679}$$

Therefore:

$$\Gamma_0(x+1)=x! \tag{6.3.680}$$

or written in another way for $x\in\mathbb{N}^*$:

$$\boxed{\Gamma(x+1)=x!=\int\limits_0^{+\infty} t^x e^{-t}\mathrm{d}t} \tag{6.3.681}$$

Notice that for $x=0$, we have:

$$\Gamma(1)=0!=\int\limits_0^{+\infty} t^0e^{-t}\mathrm{d}t=\int\limits_0^{+\infty} e^{-t}\mathrm{d}t=-e^{-t}\Big|_0^{+\infty}=-e^{-\infty}-(-e^{-0})=1 \tag{6.3.682}$$

Let us see for the fun what is the value of i! using the Euler formula of complex numbers:

$$\begin{aligned}\mathrm{i}!&=\int\limits_0^{+\infty} t^{\mathrm{i}}e^{-t}\mathrm{d}t=\int\limits_0^{+\infty} e^{\ln\left(t^{\mathrm{i}}\right)}e^{-t}\mathrm{d}t\\&=\int\limits_0^{+\infty} e^{\mathrm{i}\ln(t)}e^{-t}\mathrm{d}t=\int\limits_n^{+\infty}(\cos(\ln(t))+\mathrm{i}\sin(\ln(t)))e^{-t}\mathrm{d}t\\&=\int\limits_0^{+\infty}\cos(\ln(t))e^{-t}\mathrm{d}t+\mathrm{i}\int\limits_0^{+\infty}\sin(\ln(t))e^{-t}\mathrm{d}t\\&\cong 0.498-0.155\mathrm{i}\end{aligned} \tag{6.3.683}$$

where the last integrals have to be computed using numerical methods.

Another interesting and useful result of the Euler gamma function is obtained when we replace t by y^2 and calculate this latter for $x=0.5$.

First we have:

$$\frac{\mathrm{d}t}{\mathrm{d}y}=\frac{\mathrm{d}y^2}{\mathrm{d}y}=2y\Rightarrow \mathrm{d}t=2y\mathrm{d}y \tag{6.3.684}$$

and after:

$$\Gamma_{-1}\left(\frac{1}{2}\right)=\int\limits_0^{+\infty} t^{\frac{1}{2}-1}e^{-t}\mathrm{d}t=\int\limits_0^{+\infty} y^{-1}e^{-y^2}2y\mathrm{d}y=2\int\limits_0^{+\infty} e^{-y^2}\mathrm{d}y \tag{6.3.685}$$

But, as we have proved it in the section Statistics during our study of the Normal distribution (see page 695), we recognize here the Gauss integral that is equal for recall to:

$$\Gamma_{-1}\left(\frac{1}{2}\right)=2\int\limits_{0}^{+\infty}e^{-y^2}\mathrm{d}y=\sqrt{\pi} \tag{6.3.686}$$

Various kinds of relations can be derived using the recurrence relation:

$$\Gamma(x+1)=x\Gamma(x) \tag{6.3.687}$$

and from the previous result. For example the Gamma function for $n+\frac{1}{2}$, where n in an integer will be very useful to us in the section Statistics for the derivation of the non-central chi-square distribution and can be derived as follows:

$$\begin{aligned}\Gamma\left(n+\frac{1}{2}\right)&=\left(n-\frac{1}{2}\right)\Gamma\left(n-\frac{1}{2}\right)\\&=\left(n-\frac{1}{2}\right)\left(n-\frac{3}{2}\right)\Gamma\left(n-\frac{3}{2}\right)\\&=\left(n-\frac{1}{2}\right)\left(n-\frac{3}{2}\right)\ldots\left(\frac{3}{2}\right)\left(\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right)\\&=\left(\frac{2n-1}{2}\right)\left(\frac{2n-3}{2}\right)\ldots\left(\frac{3}{2}\right)\left(\frac{1}{2}\right)\sqrt{\pi}\\&=\frac{(2n-1)(2n-3)\ldots(3)(1)}{2^n}\sqrt{\pi}\\&=\frac{(2n-1)(2n-2)(2n-3)(2n-4)\ldots(3)(2)(1)}{2^n(2n-2)(2n-4)\ldots2}\\&=\frac{(2n-1)(2n-2)(2n-3)(2n-4)\ldots(3)(2)(1)}{2^n2^{n-1}(n-1)(n-2)\ldots1}\\&=\frac{(2n-1)!}{2^{2n-1}(n-1)!}\sqrt{\pi}\end{aligned} \tag{6.3.688}$$

6.3.2.10.1 Incomplete regularized Gamma function

The "(upper) incomplete regularized Gamma function", useful in the study of the A/B testing for count data (see section Statistics page 999), is defined by:

$$\boxed{Q(\alpha,\beta\lambda)=\frac{\Gamma(\alpha,t)}{\Gamma(\alpha)}=\int\limits_{x}^{+\infty}\frac{\beta^{\alpha}\lambda^{\alpha-1}e^{-\beta\lambda}}{\Gamma(\alpha)}\mathrm{d}\lambda} \tag{6.3.689}$$

Obviously the non-regularized version is defined as:

$$Q(\alpha,\beta\lambda)=\int\limits_{x}^{+\infty}\beta^{\alpha}\lambda^{\alpha-1}e^{-\beta\lambda}\mathrm{d}\lambda \tag{6.3.690}$$

And factoring one β it is also sometimes defined as (with the conventional notation) and putting $t=\beta\lambda$:

$$Q(\alpha,\beta\lambda)=\int\limits_{x}^{+\infty}(\beta\lambda)^{\alpha-1}e^{-\beta\lambda}\beta\mathrm{d}\lambda=Q(\alpha,t)=\int\limits_{x}^{+\infty}t^{\alpha-1}e^{-t}\mathrm{d}t \tag{6.3.691}$$

So whatever which version we choose for the next developments, the results remains the same!

Remark

We also define the "(lower) incomplete regularized Gamma function" by:

$$q(\alpha,\beta\lambda)=\int\limits_{0}^{x}\frac{\beta^{\alpha}\lambda^{\alpha-1}e^{-\beta\lambda}}{\Gamma(\alpha)}\mathrm{d}\lambda \tag{6.3.692}$$

These functions were first investigated by the mathematician Friedrich Prym in 11877 (holocene calendar), and $Q(\alpha,\beta\lambda)$ has also been named "Prym's function".

Before continuing consider the following special case that will be useful to further below:

$$Q(1,\beta\lambda)=\int\limits_{x}^{+\infty}\frac{\beta^{1}\lambda^{1-1}e^{-\beta\lambda}}{\Gamma(1)}\mathrm{d}\lambda=\beta\int\limits_{x}^{+\infty}e^{-\beta\lambda}\mathrm{d}\lambda=\beta\left(-\frac{e^{-\beta\lambda}}{\beta}\bigg|_{x}^{+\infty}\right)=-e^{-\beta\lambda}\bigg|_{x}^{+\infty}=e^{-\beta x} \tag{6.3.693}$$

And we need also to prove as intermediate result a special recursive relation. For this thanks to an integration by part we prove the following relation:

$$\Gamma(\alpha+1,t)=\alpha\Gamma(\alpha,t)+t^{\alpha}e^{-t} \tag{6.3.694}$$

Indeed:

$$\begin{aligned}\Gamma(\alpha,t)&=\int\limits_{t}^{+\infty}x^{\alpha-1}e^{-x}\mathrm{d}x=\frac{x^{\alpha}}{\alpha}e^{-x}\bigg|_{t}^{+\infty}+\frac{1}{\alpha}\int_{t}^{+\infty}x^{\alpha}e^{-x}\mathrm{d}x\\&=-\frac{t^{\alpha}}{\alpha}e^{-t}+\frac{1}{\alpha}\Gamma(\alpha+1,t)\end{aligned} \tag{6.3.695}$$

and multiplying by α in both side of the equal we proved the previous equality!

Knowing that $\Gamma(\alpha+1)=\alpha\Gamma(\alpha)$ and that by definition $Q(\alpha,t)=\Gamma(\alpha,t)/\Gamma(\alpha)$ and dividing:

$$\Gamma(\alpha+1,t)=\alpha\Gamma(\alpha,t)+t^{\alpha}e^{-t} \tag{6.3.696}$$

by $\Gamma(\alpha+1)$ we get:

$$Q(\alpha+1,t)=\alpha\frac{\Gamma(\alpha,t)}{\Gamma(\alpha+1)}+\frac{t^{\alpha}e^{-t}}{\Gamma(\alpha+1)}=Q(\alpha,t)+\frac{t^{\alpha}e^{-t}}{\Gamma(\alpha+1)}$$

From the above relation, we can calculate $Q(\alpha+2,t)$:

$$\begin{aligned}Q(\alpha+2,t)&=Q(\alpha+1,t)+\frac{t^{\alpha+1}e^{-t}}{\Gamma(\alpha+2)}\\&=Q(\alpha,t)+\frac{t^{\alpha}e^{-t}}{\Gamma(\alpha+1)}+\frac{t^{\alpha+1}e^{-t}}{\Gamma(\alpha+2)}\\&=Q(\alpha,t)+t^{\alpha}e^{-t}\left(\frac{1}{\Gamma(\alpha+1)}+\frac{t}{\Gamma(\alpha+2)}\right)\end{aligned} \tag{6.3.697}$$

and $Q(\alpha+3,t)$ for the fun...:

$$\begin{aligned}Q(\alpha+3,t)&=Q(\alpha+2,t)+\frac{t^{\alpha+2}e^{-t}}{\Gamma(\alpha+3)}\\&=Q(\alpha,t)+t^{\alpha}e^{-t}\left(\frac{1}{\Gamma(\alpha+1)}+\frac{t}{\Gamma(\alpha+2)}\right)+\frac{t^{\alpha+2}e^{-t}}{\Gamma(\alpha+3)}\\&=Q(\alpha,t)+t^{\alpha}e^{-t}\left(\frac{1}{\Gamma(\alpha+1)}+\frac{t}{\Gamma(\alpha+2)}+\frac{t^2}{\Gamma(\alpha+3)}\right)\end{aligned} \tag{6.3.698}$$

We show easily by recurrence the following power recursive relation:

$$\boxed{Q(\alpha+n,t)=Q(\alpha,t)+t^{\alpha}e^{-t}\sum_{k=0}^{n-1}\frac{t^k}{\Gamma(\alpha+k+1)}} \tag{6.3.699}$$

Using the identity $Q(1,t)=e^{-t}$ and the previous recursive relationship we can express Q as:

$$Q(n,t)=e^{-t}\sum_{k=0}^{n-1}\frac{t^k}{\Gamma(k+1)} \tag{6.3.700}$$

Indeed, let us put first $\alpha=1$:

$$\begin{aligned}Q(n+1,t)&=Q(1,t)+te^{-t}\sum_{k=0}^{n-1}\frac{t^k}{\Gamma(k+2)}=Q(1,t)+e^{-t}\sum_{k=0}^{n-1}\frac{t^{k+1}}{\Gamma(k+2)}\\&=e^{-t}+e^{-t}\sum_{k=0}^{n-1}\frac{t^{k+1}}{\Gamma(k+2)}\end{aligned} \tag{6.3.701}$$

We put $n=n+1$ (hence we require now that $n>1$), therefore:

$$Q(n,t)=e^{-t}+e^{-t}\sum_{k=0}^{n-2}\frac{t^{k+1}}{\Gamma(k+2)} \tag{6.3.702}$$

We do another change of variable $k=k+1$:

$$Q(n,t)=e^{-t}+e^{-t}\sum_{k=1}^{n-1}\frac{t^k}{\Gamma(k+1)}=e^{-t}\sum_{k=0}^{n-1}\frac{t^k}{\Gamma(k+1)} \tag{6.3.703}$$

6.3.2.10.2 Euler-Mascheroni Constant

This small text is a just curiosity regarding to Euler's constant e and to almost every Differential and Integral calculus tools that we have seen until now. This is a very nice example (almost artistic) of what we can do with mathematics as soon as we have enough tools at our disposal.

Moreover, this constant is useful in some differential equations which we see later.

Remember that we saw in the section of Functional Analysis that the Euler constant e is defined by the limit:

$$e=\lim_{\alpha\to+\infty}\left(1+\frac{1}{\alpha}\right)^{\alpha} \tag{6.3.704}$$

In a more general case we can easily demonstrate in the same way that (you can ask us the details if needed):

$$\lim_{n\to+\infty}\left(1-\frac{t}{n}\right)^{n}=e^{-t} \tag{6.3.705}$$

This obviously suggests:

$$\Gamma_{-1}(x)=\int\limits_{0}^{+\infty}t^{x-1}e^{-t}\mathrm{d}t=\lim_{n\to+\infty}\int\limits_{0}^{+\infty}t^{x-1}\left(1-\frac{t}{n}\right)^{n}\mathrm{d}t \tag{6.3.706}$$

by a change of variable $t=nu$ we write:

$$\Gamma_{-1}(x)=\lim_{n\to+\infty}\int\limits_{0}^{1}(nu)^{x-1}(1-u)^{n}n\mathrm{d}u=\lim_{n\to+\infty}n^{x}\int\limits_{0}^{1}u^{x-1}(1-u)^{n}\mathrm{d}u \tag{6.3.707}$$

And we use the definition of the Beta function (see section Statistics page 761):

$$\begin{aligned}B(x-1,n)&=\int\limits_{0}^{1}u^{x-1}(1-u)^{n}\mathrm{d}u\\&=\frac{u^{x}}{x}(1-u)^{n}\Bigg|_{0}^{1}+\frac{1}{x}n\int\limits_{0}^{1}u^{x}(1-u)^{n-1}\mathrm{d}u\\&=\frac{n}{x}B(x,n-1)=\frac{n(n-1)}{x(x+1)}B(x+1,n-2)\\&=\frac{n!}{x(x+1)\ldots(x+n)}\end{aligned} \tag{6.3.708}$$

Therefore:

$$\Gamma_{-1}(x)=\lim_{n\to+\infty}\frac{n!n^{x}}{x(x+1)\ldots(x+n)} \tag{6.3.709}$$

To transform this expression we can write:

$$\begin{aligned} n^x &= e^{\ln(n^x)} = e^{x \cdot \ln(n)} \\ &= e^{x\left(\cdot \ln(n) - 1 + 1 - \frac{1}{2} + \frac{1}{2} - \ldots - \frac{1}{n} + \frac{1}{n}\right)} \\ &= e^{-x \cdot (1 + \frac{1}{2} + \ldots + \frac{1}{n} - \ln(n))} \cdot e^x \cdot e^{\frac{x}{2}} \cdot \ldots \cdot e^{\frac{x}{n}} \end{aligned} \tag{6.3.710}$$

But the quantity:

$$1 + \frac{1}{2} + \ldots + \frac{1}{n} - \ln(n) \tag{6.3.711}$$

tends to the limit $\gamma = 0.5772$, named "Euler-Mascheroni constant" or also "Euler Gamma constant" when n tends to infinity.

Therefore:

$$\Gamma_{-1}(x) = \lim_{n \to +\infty} \frac{n! \cdot e^{-\gamma x} \cdot e^x \cdot e^{\frac{x}{2}} \cdot \ldots \cdot e^{\frac{x}{n}}}{x(x+1)\ldots(x+n)} = \lim_{n \to +\infty} \frac{n! \cdot e^{-\gamma x} \prod\limits_{k=1}^{n} e^{\frac{x}{k}}}{x(x+1)\ldots(x+n)} \tag{6.3.712}$$

We divide each term of the product $(x+1)\ldots(x+n)$ by the corresponding integer taken into $n!$, so we get (according to a reader request we have put a maximum of details!):

$$\begin{aligned} \frac{1}{\Gamma_{-1}(x)} &= \frac{1}{\lim\limits_{n \to +\infty} \dfrac{n! e^{\lambda x} \prod_{k=1}^{n} e^{\frac{x}{k}}}{x(x+1)(x+2)\ldots(x+n)}} = x \cdot e^{\gamma x} \lim_{n \to +\infty} \frac{(x+1)(x+2)\ldots(x+n)}{n!} \prod_{k=1}^{n} e^{-\frac{x}{k}} \\ &= x \cdot e^{\gamma x} \lim_{n \to +\infty} \frac{(x+1)(x+2)\ldots(x+n)}{1 \cdot 2 \ldots \cdot n} \prod_{k=1}^{n} e^{-\frac{x}{k}} = x \cdot e^{\gamma x} \lim_{n \to +\infty} \frac{(x+1)}{1} \frac{(x+2)}{2} \ldots \frac{(x+n)}{n} \prod_{k=1}^{n} e^{-\frac{x}{k}} \\ &= x \cdot e^{\gamma x} \lim_{n \to +\infty} (x+1) \left(\frac{x}{2} + 1\right) \ldots \left(\frac{x}{n} + 1\right) \prod_{k=1}^{n} e^{-\frac{x}{k}} = x \cdot e^{\gamma x} \lim_{n \to +\infty} \prod_{k=1}^{n} \left(\frac{x}{k} + 1\right) \prod_{k=1}^{n} e^{-\frac{x}{k}} \\ &= x \cdot e^{\gamma x} \lim_{n \to +\infty} \prod_{k=1}^{n} \left(1 + \frac{x}{k}\right) e^{-\frac{x}{k}} = x \cdot e^{\gamma x} \prod_{k=1}^{+\infty} \left(1 + \frac{x}{k}\right) e^{-\frac{x}{k}} \end{aligned} \tag{6.3.713}$$

6.3.2.11 Curvilinear Integrals

The line integrals (curvilinear integrals) are also very important in physics. The reader will thus find see them again in the section of Classical Mechanics, Electrodynamics Magnetostatics and to calculate the work of a force or the "flow field", or in the chapter of Euclidean Geometry to calculate the center of gravity of curves (functions), or in the section of Geometric Shapes to calculate the surface of some bodies of revolution but also in Corpuscular Quantum Physics for the famous "path integral" (which is only the term used by physicists to say "line integral") or for the specific calculation of integrals using the residue theorem proved in the section of Complex Analysis or for transformation states in the section of Thermodynamics. This is why there will not be here examples of line integrals because they are already so many applications in the other chapters of this textbook.

With the definition of these integrals, we can prove two very important results detailed in section of Vector Calculus that are respectively the Green's theorem and Stokes' theorem or even the theorem of residues proved in the section of Complex Analysis and already mentioned in the preceding paragraph (this is important enough to mention it twice!).

More technically a line integral is an integral where the function to be integrated is evaluated along a curve. The terms "path integral", "curve integral", and "curvilinear integral" are also used; "contour integral" as well, although that is typically reserved for line integrals in the complex plane.

6.3.2.11.1 Curvilinear Integral of a scalar field

Consider a parametrized curve C (see section Differential Geometry page 2850) by a vector function $\vec{r}(t)$ with $t\in[a,b]$ of class $\mathcal{C}^1$ piecewise (this condition is necessary so that we can integrate the curve without problems).

Definitions (#6.3.201):

D1. The curve is said to be a "closed curve" if $\vec{r}(a)=\vec{r}(b)$

D2. The curve is said to be a "smooth curve" if $\forall t\in[a,b]\;\vec{r}\,'(t)\neq 0$

Recall that a parametric curve can be written as follows (all vector function can be written in this form):

$$\vec{r}(t)=x(t)\vec{i}+y(t)\vec{j}=\begin{pmatrix}x(t)\\y(t)\end{pmatrix} \tag{6.3.714}$$

Consider a function or a "scalar field" $f(x,y)$ defined in a neighbourhood of C. We subdivide $[a,b]$ into n subintervals of equal length Δt as:

$$t_0=a<t_1<\ldots<t_n=b \tag{6.3.715}$$

We choose on each subinterval a point $t_i^*\in[t_i,t_{i+1}]$. Given δs_i the length of the arc C connecting the point $(x(t_i),y(t_i))$ and $(x(t_{i+1}),y(t_{i+1})$, the integral of f along C is defined as the "line integral" or "path integral":

$$\int\limits_C f(x,y)\mathrm{d}s=\lim_{n\rightarrow+\infty}\sum_{i=0}^{n} f(x(t_i^*),y(t_i^*))\Delta s_i \tag{6.3.716}$$

Which as we know, can be written (see section of Differential Geometry page 2856 of Analytical Mechanics page 3172 and many others):

$$\int\limits_C f(x(t),y(t))\mathrm{d}s=\int\limits_a^b \sqrt{\left(\frac{\mathrm{d}x(t)}{\mathrm{d}t}\right)^2+\left(\frac{\mathrm{d}y(t)}{\mathrm{d}t}\right)^2}\mathrm{d}t \tag{6.3.717}$$

and that can obviously be immediately extended to the case to 3 variables and more.

Or in vector form:

$$\int\limits_C f(x,y)\mathrm{d}s=\int\limits_a^b \left\|\frac{\mathrm{d}\vec{r}}{\mathrm{d}t}\right\|\mathrm{d}t \tag{6.3.718}$$

The line integral is linear, that is to say, if $C=C_1\cup C_2$ and that $C_1\cap C_2$ is a point, then (without going into the strict definition of the union of two curves...):

$$\int\limits_C f(x,y)\mathrm{d}s=\int\limits_{C_1} f(x,y)\mathrm{d}s+\int\limits_{C_2} f(x,y)\mathrm{d}s \tag{6.3.719}$$

6.3.2.11.2 Curvilinear Integral of a vector field

Consider a vector field (e.g. a force field) as:

$$\vec{F}=\begin{pmatrix}F_x(x,y))\\F_y(x,y)\end{pmatrix}=F_x(x,y,)\vec{i}+F_y(x,y,)\vec{j} \tag{6.3.720}$$

and an infinitesimal element of a curve (path) $\mathcal{C}^1$ piecewise as:

$$\mathrm{d}\vec{r}=\begin{pmatrix}\mathrm{d}x\\\mathrm{d}y\end{pmatrix}=\mathrm{d}x\vec{i}+\mathrm{d}y\vec{j} \tag{6.3.721}$$

The idea is then to consider that the dot product (vector field projection on the path element) represents the work along the differential element:

$$\vec{F}\circ\mathrm{d}\vec{r} \tag{6.3.722}$$

Therefore the work on all the path will be given by (using the property of linearity of the curvilinear integral):

$$W=\int_C \vec{F}\circ\mathrm{d}\vec{r}=\int_C (F_x(x,y)\mathrm{d}x+F_y(x,y)\mathrm{d}y)=\int_C F_x(x,y)\mathrm{d}x+\int_C F_y(x,y)\mathrm{d}y \tag{6.3.723}$$

This can obviously be generalized to n dimensions. Let us indicate that when the line integral (path integral) of a vector field is extended to a closed curve, then we speak of "circulation of the vector field".

As:

$$\frac{\mathrm{d}\vec{r}}{\mathrm{d}t}=\vec{r}\,'(t) \tag{6.3.724}$$

We then have write a fairly common notation:

$$\int_C \vec{F}\circ\mathrm{d}\vec{r}=\int_C \vec{F}(\vec{r}(t))\circ\vec{r}\,'(t)\mathrm{d}t=\int_C \big(F_x(x(t),y(t))r_x'(t)+F_y(x(t),y(t))r_y'(t)\big)\,\mathrm{d}t \tag{6.3.725}$$

Remark

In physics many times problems are often in the plane and require the transition to polar coordinates because many academic physic problems are centro-symmetric, which also facilitates the calculations of path integrals.

✍**Example:**

Let us calculate the work of the force of gravity moving a mass M of the point $M_1(a_1,b_1,c_1)$ to the point $M_2(a_2,b_2,c_2)$ along an arbitrary path C. The projections of the force of gravity on the coordinate axes are:

$$F_x=0,F_y=0,F_z=-mg \tag{6.3.726}$$

The work accomplish is then:

$$\begin{aligned}W&=\int_C \vec{F}\circ\mathrm{d}\vec{r}=\int_C (F_x(x,y,z)\mathrm{d}x+F_y(x,y,z)\mathrm{d}y+F_z(x,y,z)\mathrm{d}y)\\&=-\int_C mg\mathrm{d}z=-mg\Delta z\end{aligned} \tag{6.3.727}$$

so we find a very known result of the section of Classical Mechanics.

A line integral of a vector $\vec{F}$ field along a curve C_1 is independent of the path of integration if:

$$\int\limits_{C_1}\vec{F}\circ\mathrm{d}\vec{r}=\int\limits_{C_2}\vec{F}\circ\mathrm{d}\vec{r} \tag{6.3.728}$$

for any non-null curve C_2 having only the same points of departure and arrival. Furthermore if the vector field satisfied (where U in physics is typically a potential):

$$\left.\begin{aligned}F_x(x,y)&=\frac{\partial U}{\partial x}\\F_y(x,y)&=\frac{\partial U}{\partial y}\end{aligned}\right\}\Rightarrow\vec{F}=\overrightarrow{\mathrm{grad}}(U)=\vec{\nabla}(U) \tag{6.3.729}$$

as (the reader will recognize an exact total differential form):

$$\vec{F}\circ\mathrm{d}\vec{r}=F_x(x,y)\mathrm{d}x+F_y(x,y)\mathrm{d}y=\frac{\partial U}{\partial x}\mathrm{d}x+\frac{\partial U}{\partial y}\mathrm{d}y \tag{6.3.730}$$

Then the integral path on an arbitrary curve only depends only on the difference of the values the function U at the two ends! This is the "fundamental theorem for line integrals" or "gradient theorem for line integrals".

Proof 6.14.9. If the differential form of the vector field satisfies a total exact differential, we have:

$$\begin{aligned} W_{A,B} &= \int\limits_A^B \left(\frac{\partial U}{\partial x}\mathrm{d}x + \frac{\partial U}{\partial y}\mathrm{d}y\right) = \int\limits_A^B \left(\frac{\partial U}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t} + \frac{\partial U}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t}\right)\mathrm{d}t = \int\limits_{t_1}^{t_2} \frac{\mathrm{d}}{\mathrm{d}t}U(x(t),y(t))\mathrm{d}t \\ &= U(t)|_{t_1}^{t_2} = U(x(t_2),y(t_2)) - U(x(t_1),y(t_1)) = U(B) - U(A) \end{aligned} \tag{6.3.731}$$

That is:

$$\boxed{\int\limits_\Gamma \vec{F}\circ\mathrm{d}\vec{r} \Leftrightarrow \int\limits_\Gamma \vec{\nabla}(U)\circ\mathrm{d}\vec{r}: \quad \int\limits_\Gamma \vec{\nabla}(U)\circ\mathrm{d}\vec{r} = U(b) - U(a)} \tag{6.3.732}$$

■ Q.E.D.

So the line integral of an exact total differential does not depend on the path of integration but only the ends! We also conclude that if $\vec{F}$ is derived from a scalar potential and $A = B$, the line integral is then zero.

In physics this result is interpreted by saying that the work provided by a force $\vec{F}$ derived from a scalar potential acting on an elementary particle in a finite displacement does not depend on the path followed.

Definitions (#6.3.202):

D1. When the curve (path) C is closed and the path integral has a result independent result of the direction in which this path is travelled, we use the following notation (the letter below the symbol representing the path can of course vary...):

$$\oint_C \tag{6.3.733}$$

If this closed integral is always zero, we say that the integrated vector field is a "conservative vector field" and "derives from a scalar potential" (and therefore satisfies the Schwarz's theorem to be written as an exact total differential) since this stems from the proof given already just above.

D2. When the value of the integral of a closed path depends on the orientation (clockwise not equal to counter-clockwise) we use the following notation (the letter below the symbol representing the path can of course vary...):

$$\ointclockwise_C \quad \text{or} \quad \ointctrclockwise_C \tag{6.3.734}$$

Thus if the direction is direct (that is to say "counter-clockwise" or "trigonometric") as the notation on the above, its sign will be positive; if on the contrary the direction is clockwise his sign will be negative (see the proof in the section of Complex Analysis). Therefore we often speak about respectively "negative direction" or "positive direction".

Thus, to summarize, a line integral (path integral) is fully defined by the expression under the symbol of the integral, the form of integration path and the direction of the integration.

Remark

The reader will find some the proofs of very important properties of curvilinear integrals in the section of Vector Calculus as the Green-Riemann theorem or a particular application to study holomorphic functions in the section of Complex Analysis.

6.3.2.12 Surface Integrals

Surface integrals are a natural generalization of line (curvilinear) integrals: instead of integrating over a curve, we integrate over a surface in three space. Such integrals are important in any of the subjects that deal with continuous media (solids, fluids, gases), as well as subjects that deal with force fields, like electromagnetic or gravitational fields.

Though most of our work will be spent seeing how surface integrals can be calculated and what they are used for, we first want to indicate briefly how they are defined. The surface integral of the (continuous) function $f(x,y,z)$ over the surface S is denoted by:

$$\iint_S f(x,y,z)\mathrm{d}S \qquad (6.3.735)$$

You can think of $\mathrm{d}S$ as the area of an infinitesimal piece of the surface S. To define the integral above, we subdivide the surface S into small pieces having area ΔS_i, pick a point (x_i,y_i,z_i) in the i-th piece, and form the Riemann sum:

$$\sum f(x_i,y_i,z_i)\,\Delta S_i \qquad (6.3.736)$$

As the subdivision of S gets finer and finer, the corresponding sums above approach a limit which does not depend on the choice of the points or how the surface was subdivided. The previous surface integral is defined to be this limit (the surface has to be smooth and not infinite in extent, and the subdivisions have to be made reasonably, otherwise the limit may not exist, or it may not be unique).

The most important type of surface integral is the one which calculates the flux of a vector field across S. Earlier, we calculated the flux of a plane vector field $\vec{F}(x,y)$ across a directed curve in the xy-plane. What we are doing now is the analog of this in space.

We assume that S is oriented: this means that S has two sides and one of them has been designated to be the "positive side". At each point of S there are two unit normal vectors, pointing in opposite directions; the positively directed unit normal vector, denoted by $\vec{n}$, is the one standing with its base (i.e., tail) on the positive side. If S is a closed surface, like a sphere or cube - that is, a surface with no boundaries, so that it completely encloses a portion of 3-space - then by convention it is oriented so that the outer side is the positive one, i.e., so that $\vec{n}$ always points towards the outside of S:

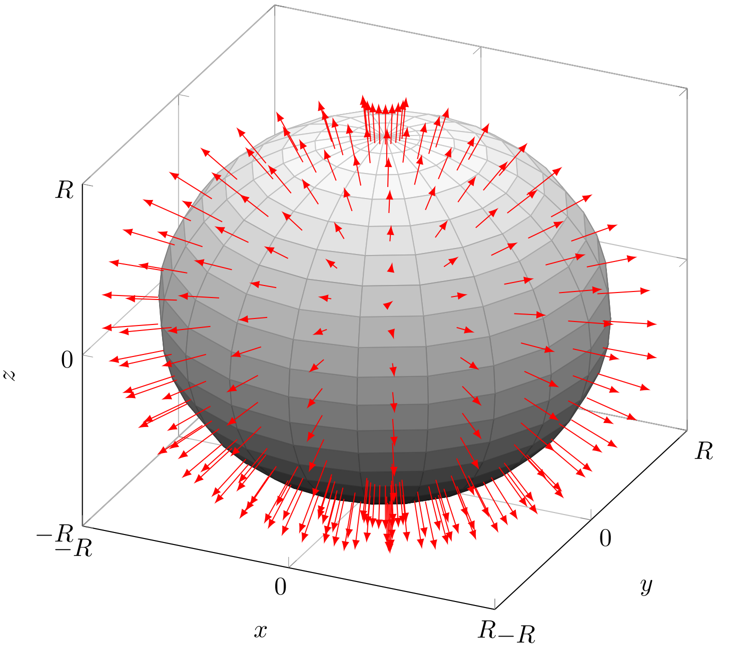


Figure 6.62: **Outgoing flux for surface integral of an entire sphere**

Let $\vec{F}(x,y,z)$ be a continuous vector field in space, and S an oriented surface. We define:

$$\text{flux of } \vec{F} \text{ through } S = \iint_S (\vec{F}\circ\vec{n})dS = \iint_S \vec{F}\circ\mathrm{d}\vec{S} \tag{6.3.737}$$

the two integrals are the same, but the second is written using the common and suggestive abbreviation $\mathrm{d}\vec{S}=\vec{n}\mathrm{d}S$.

If $\vec{F}$ represents the velocity field for the flow of an incompressible fluid of density 1, then $\vec{F}\circ\vec{n}$ represents the component of the velocity in the positive perpendicular direction to the surface, and $\vec{F}\circ\vec{n}\mathrm{d}S$ represents the flow rate across the little infinitesimal piece of surface having area $\mathrm{d}S$. The integral above adds up these flows across the pieces of surface, so that we may interpret it as saying:

$$\text{flux of } \vec{F} \text{ through } S = \text{ net flow rate across } S \tag{6.3.738}$$

where we count flow in the direction of $\vec{n}$ as positive, flow in the opposite direction as negative. More generally, if the fluid has varying density, then the right side of the above relation is the net mass transport rate of fluid across S (per unit area, per time unit).

If $\vec{F}$ is a force field, then nothing is physically flowing, and one just uses the term "flux" to denote the surface integral, as in:

$$\text{flux of } \vec{F} \text{ through } S = \iint_S (\vec{F}\cdot\vec{n})dS = \iint_S \vec{F}\circ\mathrm{d}\vec{S} \tag{6.3.739}$$

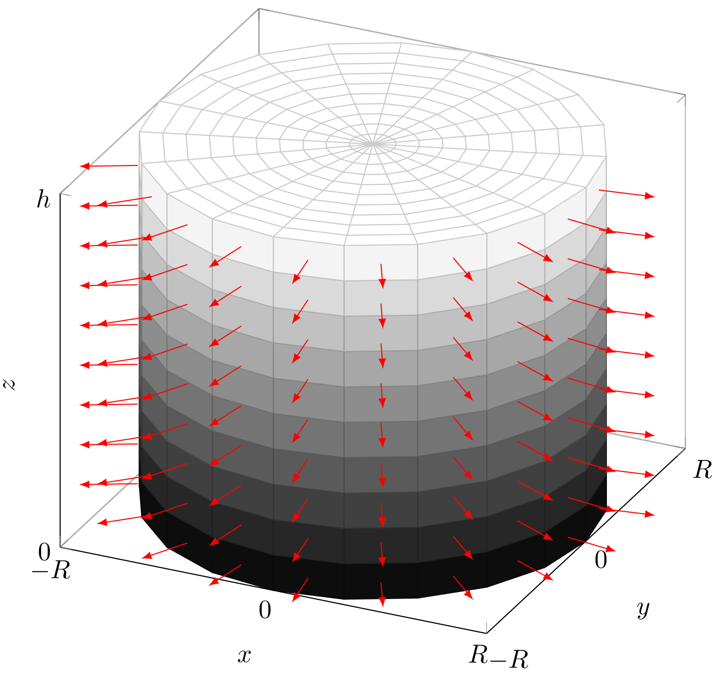


Figure 6.63: **Outgoing flux for surface integral of the side of a cylinder**

✍**Example:**

E1. Let us find the flux of $\vec{F}=z\vec{i}+x\vec{j}+y\vec{k}$ outward through the portion of the cylinder $x^2+y^2=a^2$ in the first octant and below the plane $z=h$.

The piece of cylinder is pictured. The word "outward" suggests that we orient the cylinder so that $\vec{n}$ points outward (as the figure above), i.e., away from the z axis. Since by inspection $\vec{n}$ is radially outward and horizontal:

$$\vec{n}=\frac{x\vec{i}+y\vec{j}}{a} \tag{6.3.740}$$

This is the outward normal to the circle $x^2+y^2=a^2$ in the xy-plane; $\vec{n}$ has no z-component since it is horizontal. We divide by a to make its length 1.

To get $\mathrm{d}S$, the infinitesimal element of surface area, we use cylindrical coordinates to parametrize the cylinder:

$$x=a\cos(\theta),\quad y=a\sin(\theta)\quad z=z \tag{6.3.741}$$

As the parameters θ and z vary, the whole cylinder is traced out ; the piece we want satisfies $0\leq\theta\leq\pi/2, 0\leq z\leq h$. The natural way to subdivide the cylinder is to use little pieces of curved rectangle, bounded by two horizontal arc of circles and two vertical lines on the surface. Its area $\mathrm{d}S$ is the product of its height and width:

$$\mathrm{d}S=\mathrm{d}z\cdot a\mathrm{d}\theta \tag{6.3.742}$$

Having obtained $\vec{n}$ and $\mathrm{d}S$, the rest of the work is routine. We express the integrand of our surface integral:

$$\Phi_S=\iint_S(\vec{F}\circ\vec{n})\mathrm{d}S=\iint_S\vec{F}\circ\mathrm{d}\vec{S} \tag{6.3.743}$$

in terms of z and θ :

$$\begin{aligned}\vec{F}\circ\vec{n}\mathrm{d}S&=\dfrac{zx+xy}{a}\cdot a\mathrm{d}z\mathrm{d}\theta\\&=\left(az\cos(\theta)+a^2\sin(\theta)\cos(\theta)\right)\mathrm{d}z\mathrm{d}\theta\end{aligned}\tag{6.3.744}$$

This last step is essential, since the $\mathrm{d}z$ and $\mathrm{d}\theta$ tell us the surface integral will be calculated in terms of z and θ, and therefore the integrand must use these variables also. We can now calculate the flux through S :

$$\begin{aligned}\iint_S \vec{F}\circ\vec{n}\mathrm{d}S&=\int\limits_0^{\pi/2}\int\limits_0^{h}\left(az\cos(\theta)+a^2\sin(\theta)\cos(\theta)\right)\mathrm{d}z\mathrm{d}\theta\\ \text{inner integral }&=\dfrac{ah^2}{2}\cos(\theta)+a^2h\sin(\theta)\cos(\theta)\\ \text{outer integral }&=\left[\dfrac{ah^2}{2}\sin(\theta)+a^2h\dfrac{\sin^2(\theta)}{2}\right]_0^{\pi/2}=\dfrac{ah}{2}(a+h)\end{aligned}\tag{6.3.745}$$

We can notice quickly that on the whole lateral surface of the cylinder the surface integral is equal to zero.

E2. Let us found the flux of $\vec{F}=xz\vec{i}+yz\vec{j}+z^2\vec{k}$ outward through that part of the sphere $x^2+y^2+z^2=a^2$ lying in the first octant $(x,y,z,\geq 0)$.

Once again, we begin by finding $\vec{n}$ and $\mathrm{d}S$ for the sphere. We take the outside of the sphere as the positive side, so $\vec{n}$ points radially outward from the origin; we see by inspection therefore that:

$$\vec{n}=\dfrac{x\vec{i}+y\vec{j}+z\vec{k}}{a}\tag{6.3.746}$$

where we have divided by a to make $\vec{n}$ a unit vector.

To do the integration, we use spherical coordinates ρ,ϕ,θ. On the surface of the sphere of radius $\rho=a$, so the coordinates are just the two angles ϕ and θ. The area element $\mathrm{d}S$ is most easily found using the famous spherical volume element:

$$\mathrm{d}V=\rho^2\sin(\phi)\mathrm{d}\rho\mathrm{d}\phi\mathrm{d}\theta=\mathrm{d}S\cdot\mathrm{d}\rho=\text{ area }\cdot\text{ thickness}\tag{6.3.747}$$

so that dividing by the thickness $\mathrm{d}\rho$ and setting $\rho=a$, we get:

$$\mathrm{d}S=a^2\sin(\phi)\mathrm{d}\phi\mathrm{d}\theta\tag{6.3.748}$$

Finally since the area element $\mathrm{d}S$ is expressed in terms of ϕ and θ, the integration will be done using these variables, which means we need to express x,y,z in terms of ϕ and θ. We use the formulas expressing Cartesian in terms of spherical coordinates (setting $\rho=a$ since (x,y,z) is on the sphere):

$$x=a\sin(\phi)\cos(\theta),\quad y=a\sin(\phi)\sin(\theta),\quad z=a\cos(\phi)\tag{6.3.749}$$

We can now calculate the flux integral:

$$\Phi_S=\iint_S(\vec{F}\circ\vec{n})\mathrm{d}S=\iint_S\vec{F}\circ\mathrm{d}\vec{S} \tag{6.3.750}$$

By:

$$\vec{n}=\frac{x\vec{i}+y\vec{j}+z\vec{k}}{a} \tag{6.3.751}$$

and:

$$\mathrm{d}S=a^2\sin(\phi)\mathrm{d}\phi\mathrm{d}\theta \tag{6.3.752}$$

the integrand is:

$$\vec{F}\circ\vec{n}\mathrm{d}S=\frac{1}{a}\left(x^2z+y^2z+z^2z\right)\cdot a^2\sin(\phi)\mathrm{d}\phi\mathrm{d}\theta \tag{6.3.753}$$

Using:

$$x=a\sin(\phi)\cos(\theta),\quad y=a\sin(\phi)\sin(\theta),\quad z=a\cos(\phi) \tag{6.3.754}$$

and noting that $x^2+y^2+z^2=a^2$, the integral becomes (we put $u=\sin(\phi)\longrightarrow\mathrm{d}u=\cos(\phi)\mathrm{d}\phi$):

$$\begin{aligned}\iint_S\vec{F}\circ\vec{n}\mathrm{d}S&=a^4\int\limits_0^{\pi/2}\int\limits_0^{\pi/2}\cos(\phi)\sin(\phi)\mathrm{d}\phi\mathrm{d}\theta=a^4\frac{\pi}{2}\int\limits_0^{\pi/2}\int\limits_0^{\pi/2}\cos(\phi)\sin(\phi)\mathrm{d}\phi\\&=a^4\frac{\pi}{2}\int\limits_0^{\pi/2}u\mathrm{d}u=a^4\frac{\pi}{2}\frac{1}{2}\sin^2(\phi)\Big]_0^{\pi/2}=\frac{\pi a^4}{4}\end{aligned} \tag{6.3.755}$$

6.3.2.13 Integrals involving parametric equations

Now that we have seen how to calculate the derivative of a plane curve, the next question is this: How do we find the area under a curve defined parametrically?

To derive an expression for the area under a parametric curve defined by the functions:

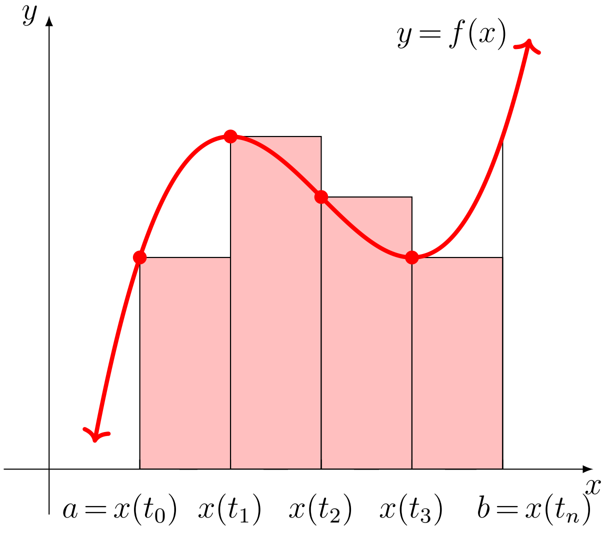
$$x = x(t) \qquad y = y(t) \tag{6.3.756}$$

with $a \leq t \leq b$, we assume that $x(t)$ is differentiable and start with an equal partition of the interval $a \leq t \leq b$.

Let us suppose:

$$t_0 = a < t_1 < t_2 < \ldots < t_n = b \tag{6.3.757}$$

and consider the following figure:



We use rectangles to approximate the area under the curve. The height of a typical rectangle in this parametrization is $y(x(\overline{t_i}))$ for some value $\overline{t_i}$ in the i-th subinterval, and the width can be calculated as $x(t_i) - x(t_{i-1})$. Thus the area of the i-th rectangle is given by:

$$S_i = y(x(\overline{t_i}))(x(t_i) - x(t_{i-1})) \tag{6.3.758}$$

Then a Riemann sum for the area is:

$$S_n = \sum_{i=1}^{n} y(x(\overline{t_i}))(x(t_i) - x(t_{i-1})) \tag{6.3.759}$$

Multiplying and dividing each area by $t_i - t_{i-1}$ gives:

$$S_n = \sum_{i=1}^{n} y(x(\overline{t_i}))\frac{(x(t_i) - x(t_{i-1}))}{x(t_i) - x(t_{i-1})}(x(t_i) - x(t_{i-1})) = \sum_{i=1}^{n} y(x(\overline{t_i}))\frac{(x(t_i) - x(t_{i-1}))}{\Delta t}\Delta t \tag{6.3.760}$$

Taking the limit as n approaches infinity gives:

$$\boxed{S=\lim_{n\rightarrow+\infty}=\int\limits_a^b y(t)x'(t)\mathrm{d}t} \tag{6.3.761}$$

And it is obvious that applying Pythagoras's theorem on an infinitesimal length of the parametric curve we have:

$$\boxed{L=\int\limits_{t_1}^{t_2}\sqrt{\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2+\left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2}\mathrm{d}t} \tag{6.3.762}$$

with $x=x(t)$, $y=y(t)$ and $t_1<t<t_2$. This gives the arc length of the parametric curve between two points on the curve.

It comes also immediately:

$$L=\int\limits_{t_1}^{t_2}\frac{\mathrm{d}x}{\mathrm{d}t}\sqrt{1+\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2}\mathrm{d}t=\int\limits_{t_1}^{t_2}x'(t)\sqrt{1+y'(t)^2}\mathrm{d}t \tag{6.3.763}$$

The chain rule gives:

$$\mathrm{d}x=x'(t)\mathrm{d}t \tag{6.3.764}$$

Therefore:

$$\boxed{L=\int\limits_a^b\sqrt{1+y'(t)^2}\mathrm{d}x} \tag{6.3.765}$$

We will meet again this relation in the section of Analytical Mechanics.

In astronomy we often have to deal with closed curved and calculated the distance travel on that curve. But working in cartesian coordinates is not always is the best. This is why it is better to change in polar coordinates.

The idea is to suppose that we are able to express our curve of interest in the following form:

$$r=f(\theta) \tag{6.3.766}$$

where $\alpha\leq\theta\leq\beta$. In order to adapt the arc length relation for a polar curve, we use the relations:

$$x=r\cos(\theta)=f(\theta)\cos(\theta)\qquad y=r\sin(\theta)=f(\theta)\sin(\theta) \tag{6.3.767}$$

and we replace the parameter t by θ. Then:

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}\theta}&=f'(\theta)\cos(\theta)-f(\theta)\sin(\theta)\\ \frac{\mathrm{d}y}{\mathrm{d}\theta}&=f'(\theta)\sin(\theta)+f(\theta)\cos(\theta)\end{aligned} \tag{6.3.768}$$

We replace $\mathrm{d}t$ by $\mathrm{d}\theta$, and the lower and upper limits of integration are α and β respectively. Then the arc length formula becomes:

$$\begin{aligned}L&=\int\limits_a^b\sqrt{\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2+\left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2}\mathrm{d}t=\int\limits_\alpha^\beta\sqrt{\left(\frac{\mathrm{d}x}{\mathrm{d}\theta}\right)^2+\left(\frac{\mathrm{d}y}{\mathrm{d}\theta}\right)^2}\mathrm{d}t\\&=\int\limits_\alpha^\beta\sqrt{(f'(\theta)\cos(\theta)-f(\theta)\sin(\theta))^2+(f'(\theta)\sin(\theta)+f(\theta)\cos(\theta))^2}\mathrm{d}\theta\\&=\int\limits_\alpha^\beta\sqrt{(f'(\theta))^2(\cos^2(\theta)+\sin^2(\theta))+(f(\theta))^2(\cos^2(\theta)+\sin^2(\theta))}\mathrm{d}\theta\\&=\int\limits_\alpha^\beta\sqrt{(f'(\theta))^2+(f(\theta))^2}\mathrm{d}\theta=\int\limits_\alpha^\beta\sqrt{r^2+\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2}\mathrm{d}\theta\end{aligned}\tag{6.3.769}$$

So finally in polar coordinates:

$$\boxed{L=\int\limits_\alpha^\beta\sqrt{r^2+\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2}\mathrm{d}\theta}\tag{6.3.770}$$

6.3.2.14 Improper Integrals

Improper integrals are definite integrals where one or both of the boundaries is at infinity, or where the integrand has a vertical asymptote in the interval of integration. As crazy as it may sound, we can actually calculate some improper integrals using some clever methods that involve limits.

By abuse of notation, improper integrals are often written symbolically just like standard definite integrals, perhaps with infinity among the limits of integration. When the definite integral exists (in the sense of either the Riemann integral or the more advanced Lebesgue integral), this ambiguity is resolved as both the proper and improper integral will coincide in value and this is what will the most occur through all applications of integrals in physics, chemistry and engineering through this textbook!

For the Riemann integral (or the Darboux integral, which is equivalent to it as we have seen earlier above), improper integration is necessary both for unbounded intervals (since one cannot divide the interval into finitely many subintervals of finite length) and for unbounded functions with finite integral (since, supposing it is unbounded above, then the upper integral will be infinite, but the lower integral will be finite)!

An "improper integral" of a function $f(x)>0$ is given basically by:

$$\int\limits_a^{+\infty}f(x)\mathrm{d}x=\lim_{N\rightarrow+\infty}\int\limits_a^N f(x)\mathrm{d}x\tag{6.3.771}$$

And we say that the improper integral converge if this limit exists and diverges otherwise.

More generally, improper integrals are given by the following pairs of possibilities:

$$\boxed{\lim_{b\rightarrow+\infty}\int\limits_a^b f(x)\mathrm{d}x \qquad \lim_{a\rightarrow-\infty}\int\limits_a^b f(x)\mathrm{d}x} \tag{6.3.772}$$

or:

$$\boxed{\lim_{c\rightarrow b^-}\int\limits_a^c f(x)\mathrm{d}x \qquad \lim_{c\rightarrow a^+}\int\limits_c^b f(x)\mathrm{d}x} \tag{6.3.773}$$

in which one takes a limit in one or the other (or sometimes both) endpoints.

Remark

By abuse of notation, improper integrals are often written symbolically just like standard definite integrals, perhaps with infinity among the limits of integration. When the definite integral exists (in the sense of either the Riemann integral or the more advanced Lebesgue integral), this ambiguity is resolved as both the proper and improper integral will coincide in value.

Geometrically then the improper integral represents the total area under a curve stretching to infinity. If the integral $\int_a^{+\infty} f(x)\mathrm{d}x$ converges the total area under the curve is finite; otherwise it's infinite.

But how can an area that extends to infinity be finite? Obviously the area between a and N (i.e. $\int_a^N f(x)\mathrm{d}x$) is finite. As N goes to infinity this quantity will either grow without bound or it will converge to some finite value.

The domains where improper integrals are the most used, without even be noticeable most of time by students or engineering practitioners, are respectively:

- Statistics (see corresponding section) when we normalize or check the condition of convergence to a cumulated probability of 1 of a density function (most of time in statistics one or the both boundaries are equal to infinity)
- Wave Quantum Physics (see corresponding section) where we deal sometimes with free propagating particles from infinity to infinity (this also happens sometimes in General Relativity)
- Differential equations, especially when we solve them by using Fourier Transform or Laplace transform (we have many examples using this transforms across the book) for practical application in physics and high level financial engineering.
- In astrophysics or electrostatics when dealing with any punctual potential field source of the type $1/r^2$ for which we want to calculate the work necessary to bring an object from infinity to that source (calculation of the type $\int_{+\infty}^{r} f(r)\mathrm{d}r$).

✍**Examples:**

E1. We want compute the very important integral (to introduce Laplace Transforms):

$$I=\int\limits_{0}^{+\infty} e^{-at}\mathrm{d}t \tag{6.3.774}$$

with $a\in\mathbb{R}$.

Following the definition above we need to first compute a definite integral and the take a limit. So from the definition:

$$I=\int\limits_{0}^{+\infty} e^{-at}\mathrm{d}t=\lim_{N\rightarrow+\infty}\int\limits_{0}^{N} e^{-at}\mathrm{d}t \tag{6.3.775}$$

We first compute the definite integral. We start with the case $a=0$:

$$I_{a=0}=\lim_{N\rightarrow+\infty}\int\limits_{0}^{N}\mathrm{d}t=\lim_{N\rightarrow+\infty} t\Big|_{0}^{N}=\lim_{N\rightarrow+\infty} N=+\infty \tag{6.3.776}$$

therefore for $a=0$ the improper integral I does not exist. When $a\neq 0$ we have:

$$I_{a\neq 0}=\lim_{N\rightarrow+\infty}-\frac{1}{a}\left(e^{-aN}-1\right) \tag{6.3.777}$$

In the case $a<0$, that is $a=-|a|$, we have that:

$$I_{a<0}=\lim_{N\rightarrow+\infty}-\frac{1}{a}\left(e^{-aN}-1\right)=-\infty \tag{6.3.778}$$

therefore for $a<0$ the improper integral I does not exist. In the case $a>0$ we have:

$$I_{a>0}=\lim_{N\rightarrow+\infty}-\frac{1}{a}\left(e^{-aN}-1\right)=-\frac{1}{a}(0-1)=\frac{1}{a} \tag{6.3.779}$$

E2. We want to compute:

$$I=\int\limits_{-2}^{3}\frac{1}{x^3}\mathrm{d}x \tag{6.3.780}$$

The integrand is not continuous at $x=0$ and so we will need to split the integral up at that point:

$$I=\int\limits_{-2}^{3}\frac{1}{x^3}\mathrm{d}x=\int\limits_{-2}^{0}\frac{1}{x^3}\mathrm{d}x+\int\limits_{0}^{3}\frac{1}{x^3}\mathrm{d}x \tag{6.3.781}$$

Now we need to look at each of these integrals and see if they are convergent:

$$\int\limits_{-2}^{0}\frac{1}{x^3}\mathrm{d}x=\lim_{t\rightarrow 0^-}\int\limits_{-2}^{t}\frac{1}{x^3}\mathrm{d}x=\lim_{t\rightarrow 0^-}\left(-\frac{1}{2x^2}\right)\Bigg|_{-2}^{t}=\lim_{t\rightarrow 0^-}\left(-\frac{1}{2t^2}+\frac{1}{8}\right)=-\infty \tag{6.3.782}$$

At this point we're done. One of the integrals is divergent that means the integral that we were asked to look at is divergent. We don't even need to bother with the second integral.

On a side note, notice that the area under the curve of this infinite interval was not infinity as the reader may have suspected it to be (perhaps). In fact, it was a surprisingly small number. Of course this won't always be the case, but it is important enough to point out that not all areas on an infinite interval will yield infinite areas.

Let's now get some definitions out of the way. We will name these integrals "convergent integrals" if the associated limit exists and is a finite number (i.e. it's not plus or minus infinity) and "divergent integrals" if the associated limit either doesn't exist or is equal to plus or minus infinity.

Notice that the similar integral:

$$\int_{-1}^{+1} \frac{\mathrm{d}x}{x} \tag{6.3.783}$$

cannot be assigned a value in the previous way, as the integrals above and below zero do not independently converge. However there is another possibility that we will see in the section of Functional Analysis during our study of the Hilbert transform (see page 2376).

6.3.2.15 Elliptic Integrals

"Elliptic integrals" originally arose in connection with the problem of giving the arc length of an ellipse. They were first studied by Giulio Fagnano and Leonhard Euler. Modern mathematics defines an "elliptic integral" as any function f which can be expressed in the form:

$$f(x)=\int\limits_{c}^{x} R\left(t,\sqrt{P(t)}\right)\mathrm{d}t \qquad (6.3.784)$$

where R is a rational function of its two arguments, P is a polynomial of degree 2 to 4 with no repeated roots, and c is a constant.

In general, integrals in this form cannot be expressed in terms of elementary functions. Exceptions to this general rule are when P has repeated roots, or when $R(x,y)$ contains no odd powers of t. However, with the appropriate reduction formula, every elliptic integral can be brought into a form that involves integrals over rational functions and the three Legendre canonical forms (i.e. the elliptic integrals of the first, second and third kind).

Let us inform now the reader that we will in this textbook, encounter the following Elliptic Integrals:

- In the section of Non-Euclidean Geometry at page 2678 (complete elliptic integral of second kind):

$$E(k)=\int\limits_{0}^{\pi/2}\sqrt{1-k^2\sin^2(\tau)}\mathrm{d}x \qquad (6.3.785)$$

- In the section of Geometric Shapes at page 2919 (second-order elliptic integral):

$$J(a,b)=\int\limits_{0}^{\pi/2}\sqrt{a^2\cos^2(\theta)+b^2\sin^2(\theta)}\mathrm{d}\theta \qquad (6.3.786)$$

- In the section of Classical Mechanics for the study of the pendulum 3412 (elliptic integral of the first kind):

$$T=4\sqrt{\frac{L}{g}}\int\limits_{0}^{\pi/2}\frac{\mathrm{d}\theta}{\sqrt{1-p^2\sin^2(\theta)}}=4\sqrt{\frac{L}{g}}F\left(\frac{\pi}{2},p^2\right) \qquad (6.3.787)$$

6.3.2.15.1 Incomplete Elliptic Integrals

It is common first to consider three types integrals that we name "incomplete[7] elliptic integrals".

Definitions (#6.3.203):

D1. The "incomplete elliptic integral of the first kind" depends of two parameters, that are the amplitude ϕ and the angle α:

$$F(\phi,\alpha)=\int_0^{\phi}\frac{\mathrm{d}\theta}{\sqrt{1-\sin^2(\alpha)\sin^2(\theta)}} \tag{6.3.788}$$

We can do the following simple change of variables:

$$\begin{aligned} t&=\sin(\theta)\\ \mathrm{d}t&=\cos(\theta)\mathrm{d}\theta=\sqrt{1-t^2}\mathrm{d}\theta \end{aligned} \tag{6.3.789}$$

Hence:

$$F(\phi,\alpha)=\int_0^{\sin(\phi)}\frac{\mathrm{d}t}{\sqrt{(1-\sin^2(\alpha)t^2)(1-t^2)}} \tag{6.3.790}$$

We then use also sometimes instead the parameter $k=\sin(\alpha)$, $(\leq m\leq 1)$, by writing:

$$F(\phi,k)=\int_0^{\phi}\frac{\mathrm{d}\theta}{\sqrt{1-k^2\sin^2(\theta)}}=\int_0^{\sin(\phi)}\frac{\mathrm{d}t}{\sqrt{(1-k^2t^2)(1-t^2)}} \tag{6.3.791}$$

named the "Jacobi's form" of the elliptic integral.

✍**Example:**

Using (see section Trigonometry page 2519):

$$\cos(x)=\cos(\tfrac{1}{2}x+\tfrac{1}{2}x)=\cos^2(\tfrac{1}{2}x)-\sin^2(\tfrac{1}{2}x)=1-2\,\sin^2(\tfrac{1}{2}x) \tag{6.3.792}$$

Then:

$$\int_0^{x}\frac{1}{\sqrt{\cos(x)}}\mathrm{d}x=\int_0^{x}\frac{\mathrm{d}x}{\sqrt{1-2\sin^2(\frac{1}{2}x)}}=2\int_0^{x/2}\frac{\mathrm{d}\theta}{\sqrt{1-2\sin^2(\theta)}}=2F\left(\tfrac{1}{2}x,\sqrt{2}\right) \tag{6.3.793}$$

D2. The "incomplete elliptic integral of the second type" is define we the same both parameters:

$$E(\phi,\alpha)=\int_0^{\phi}\sqrt{1-\sin^2(\alpha)\sin^2(\theta)}\mathrm{d}\theta \tag{6.3.794}$$

---

[7] The term "incomplete integral" seems to be an old fashioned way to name a primitive function...

We can also do the following simple change of variable:

$$\begin{aligned} t &= \sin(\theta) \\ \mathrm{d}t &= \cos(\theta)\mathrm{d}\theta = \sqrt{1-t^2}\mathrm{d}\theta \end{aligned} \tag{6.3.795}$$

and $k = \sin(\alpha)$ to get same as previously:

$$E(\phi,\alpha) = \int\limits_0^{\phi} \sqrt{1-k^2\sin^2(\theta)}\mathrm{d}\theta = \int\limits_0^{x} \frac{\sqrt{1-k^2t^2}}{\sqrt{1-t^2}}\mathrm{d}t \tag{6.3.796}$$

D3. The "incomplete elliptic integral of the third kind":

$$\Pi(\phi,\alpha,n) = \int\limits_0^{\phi} \frac{\mathrm{d}\theta}{(1-n\sin^2(\theta))\sqrt{1-\sin^2(\alpha)\sin^2(\theta)}} \tag{6.3.797}$$

or:

$$\Pi(\phi,k,n) = \int\limits_0^{\phi} \frac{\mathrm{d}\theta}{(1-n\sin^2(\theta))\sqrt{1-k^2\sin^2(\theta)}} \tag{6.3.798}$$

for $n > 0$. We can do the same type of change of variable, and therefore:

$$\Pi(\phi,k,n) = \int\limits_0^{\sin(\phi)} \frac{\mathrm{d}\theta}{(1-nt^2)\sqrt{(1-t^2)(1-kt^2)}} \tag{6.3.799}$$

6.3.2.15.2 Complete Elliptic Integrals

When the amplitude is equal to $\pi/2$ (hence $t = 1$), we write the following "complete elliptic integrals".

Definitions (#6.3.204):

D1. The "complete elliptic integral of the first kind" may thus be defined from the incomplete elliptic integral of the first kind:

$$K(k) = F(\pi/2,k) = \int\limits_0^{\pi/2} \frac{\mathrm{d}\theta}{\sqrt{1-k^2\sin^2(\theta)}} = \int\limits_0^{1} \frac{\mathrm{d}t}{\sqrt{(1-t^2)(1-k^2t^2)}} \tag{6.3.800}$$

D2. The "complete elliptic integral of the second kind" may thus be defined from the incomplete elliptic integral of the second kind also:

$$E(k) = E(\pi/2,k) = \int\limits_0^{\frac{\pi}{2}} \sqrt{1-k^2\sin^2(\theta)}\mathrm{d}\theta = \int\limits_0^{1} \frac{\sqrt{1-k^2t^2}}{\sqrt{1-t^2}}\mathrm{d}t \tag{6.3.801}$$

D3. The "complete elliptic integral of the third kind" may thus be defined from the incomplete elliptic integral of the third kind also:

$$\Pi(k,n)=\Pi(\pi/2,k,n)=\int\limits_0^{\pi/2}\frac{\mathrm{d}\theta}{(1-n\sin^2(\theta))\sqrt{1-\sin^2(\alpha)\sin^2(\theta)}}=\int\limits_0^{1}\frac{\mathrm{d}\theta}{(1-nt^2)\sqrt{(1-t^2)(1-kt^2)}} \tag{6.3.802}$$

6.3.3 Differential Equations

Definition (#6.3.205): In mathematics, a "differential equation" (D.E.) is a relation between one or more unknown functions and their derivatives up to order n. The "order" of a differential equation corresponding to the maximum degree of differentiation which one of the functions is subjected.

Compared to our goal of trying to see how the math describes the sensible reality, differential equations are a great success but are also the source of many troubles. First there are modelling difficulties (see for example the differential equation system of General Relativity in the corresponding section of the book...) resolution difficulties (there is no general method even with numerical computer methods as you can see in the corresponding section!), then their are proper mathematical difficulties (that's why some D.E. have million dollar price in case or resolution), finally difficulties related to the fact that certain differential equations are unstable by nature and give chaotic solutions (see the section Population Dynamics or Meteorology for flagrant simple examples!).

Remark

The differential equations are used to construct mathematical models of physical or biological phenomena, such as for the study of radioactivity, celestial mechanics, electronic circuits, population development or even financial stochastic process. Therefore, differential equations represent a vast field of study, both in pure and Applied Mathematics.

> ❞ *Since Newton, mankind has come to realize that the laws of physics are always expressed in the language of differential equations.* ❞
>
> Steven Strogatz

quote 6.3:1603.#54

The most general differential equation of order n can always be written in the form:

$$F\left(x,y,y',\ldots,y^{(n)}\right)=0 \qquad (6.3.803)$$

We consider in this textbook only the cases where x and y have their values in $\mathbb{R}$. A solution to such a D.E. on the interval $I\in\mathbb{R}$ is a function $y\in\mathcal{C}^n(I,\mathbb{R})$ (a function $y:I\rightarrow\mathbb{R}$ which is n times continuously differentiable) such that for any $x\in I$, we have:

$$F\left(x,y(x),y'(x),\ldots,y^{(n)}(x)\right)=0 \qquad (6.3.804)$$

Name differential equation	Notation	Parameters involved
Laplace	$\nabla^2 f = 0$	$f(x,y,z)$
Diffusion (conduction)	$\frac{\partial f}{\partial t} = D\nabla^2 f$	Diffusion coefficient D
Helmholtz	$\nabla^2 f + \alpha^2 f = 0$	Constant α
Wave	$\nabla^2 f = \frac{1}{c^2}\frac{\partial^2 f}{\partial t^2}$	Wave speed c
Legendre	$\frac{\mathrm{d}}{\mathrm{d}x}\left[(1-x^2)\frac{\mathrm{d}y}{\mathrm{d}x}\right] + l(l+1)y = 0$	Integer l
Associated Legendre	$\frac{\mathrm{d}}{\mathrm{d}x}\left[(1-x^2)\frac{\mathrm{d}y}{\mathrm{d}x}\right] + \left[l(l+1) - \frac{m^2}{1-x^2}\right]y = 0$	Integers l, m
Bessel	$x^2\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} + x\frac{\mathrm{d}y}{\mathrm{d}x} + (x^2 - m^2)y = 0$	Constant m
Hermite	$\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} - 2x\frac{\mathrm{d}y}{\mathrm{d}x} + 2\alpha y = 0$	Constant α
Laguerre	$x\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} + (1-x)\frac{\mathrm{d}y}{\mathrm{d}x} + \alpha y = 0$	Constant α
Associated Laguerre	$x\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} + (1+k-x)\frac{\mathrm{d}y}{\mathrm{d}x} + \alpha y = 0$	Integer k , constant α
Chebyshev	$(1-x^2)\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} - x\frac{\mathrm{d}y}{\mathrm{d}x} + n^2 y = 0$	Integer n
Euler (or Cauchy)	$x^2\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} + ax\frac{\mathrm{d}y}{\mathrm{d}x} + by = f(x)$	Constants a, b
Bernoulli	$\frac{\mathrm{d}y}{\mathrm{d}x} + p(x)y = q(x)y^a$	
Airy	$\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} = xy$	
Blach & Scholes	$\frac{\partial F}{\partial t} + \frac{1}{2}\sigma^2 S^2\frac{\partial^2 F}{\partial S^2} + rS\frac{\partial F}{\partial S} - rF = 0$	Constant σ, S, r
Fuchs	$x^n y^{(n)} + p_1(x)x^{n-1}y^{(n-l)} + \ldots + p_n(x)y = 0$	p_i holomorph

Table 6.12: **Some of the most famous differential equations we will build and use in this textbook in practical cases**

Remarks

R1. For reasons that will be developed later, we also say "integrate a D.E." instead of saying "finding a solution to the D.E.". The first expression is particularly found in the American literature.

R2. Since all this textbook is full examples of differential equations with initial conditions (we speak then about a "Cauchy problem") and methods of resolutions in the section of Classical Mechanics, Atomic Physics, Cosmology, Econometry, Sequences and series, Industrial Engineering, Statistics, etc., we will not give any application examples here and will focus only on the minimum theoretical useful aspect.

6.3.3.1 First order differential equations

A differential equation of the first order is therefore a D.E. which involves only the first derivative y'.

Definition (#6.3.206): A first order differential equation is named "D.E. of order 1 with separate variables" if it can be written as:

$$f(y)y' = g(x) \tag{6.3.805}$$

Such a differential equation can be easily integrated. Indeed, we write:

$$y' = \frac{\mathrm{d}y}{\mathrm{d}x} \tag{6.3.806}$$

Then symbolically:

$$f(y)\mathrm{d}y = g(x)\mathrm{d}x \Leftrightarrow \int f(y)\mathrm{d}y = \int g(x)\mathrm{d}x + c^{te} \tag{6.3.807}$$

Remark

We write explicitly here the arbitrary integration constant $c^{te} \in \mathbb{R}$ (which is normally implicitly present in the indefinite integrals) to not forget it!

So the purpose is first to find the primitives F and G of f and g, and then to express y in terms of x (and of c^{te}):

$$F(y) = G(x) + c^{te} \Leftrightarrow y = F^{-1}\left(G(x) + c^{te}\right) \tag{6.3.808}$$

The integration constant is fixed when asked for a given $x = x_0$, we get a particular value of $y(x) = y(x_0) = y_0$. We speak then of "initial value problem" or of "initial conditions" also abbreviated: IC. So, in other words, initial conditions are values of the solution and/or its derivative(s) at specific points. As we will see eventually, solutions to ... nice enough ... differential equations are unique and hence only one solution will meet the given conditions. The number of initial conditions that are required for a given differential equation will depend upon the order of the differential equation as we will see.

6.3.3.2 Linear differential equations

Definition (#6.3.207): A differential equation of order n is named "linear differential equation L.D.E." if and only if it is of the form:

$$L(y) = f(x) \tag{6.3.809}$$

with:

$$L(y) = a_0(x)y + a_1(x)y' + a_2(x)y'' + \ldots + a_n(x)y^{(n)} \tag{6.3.810}$$

Hence a differential equation is linear if the dependent variable $(y, y', y'', \ldots)$ appears only linearly.

Let us now see a property that may seem insignificant at first glance but which will become very important later!

We will prove now that L is a linear application:

$$L(y+z)=\sum_{i=0}^{n}a_i(x)(y+z)^{(i)}=\sum_{i=0}^{n}a_i(x)y^{(i)}+\sum_{i=0}^{n}a_i(x)z^{(i)}=L(y)+L(z) \quad (6.3.811)$$

and for all $\lambda\in\mathbb{R}$:

$$L(\lambda y)=\sum_{i=0}^{n}a_i(x)(\lambda y)^{(i)}=\lambda\sum_{i=0}^{n}a_i(x)y^{(i)}=\lambda L(y) \quad (6.3.812)$$

Then we say that the linear D.E. represent a linear model if the multiples of this function (or any linear combination) are also a solution. Thus, in physics, for a linear system, the amplification of the cause involves an amplification of the effect (the systems are often linear in high-school problems but in reality they are rather the exception!).

For example, the ordinary differential equation of order 2 of the simple pendulum proved in the section of Classical Mechanics is not linear because it contains a sine term that is not separable.

Definition (#6.3.208): The linear differential equation (which is the most common in physics):

$$L(y)=0 \quad (6.3.813)$$

is named "homogeneous equation H.E." or " equation without second member E.W.S.M." (and sometimes "complementary equation") associated to:

$$L(y)=f(x) \quad (6.3.814)$$

In other words, a differential equation is homogeneous if $y=0$ is a solution.

Theorem 6.15. We will now prove an important property of H.E.: the set $\{S_0\}$ of solutions of the H.E. is the kernel of the linear application L (which means for refresh: $L(S_0)=0$) and the set $\{S\}$ of solutions to $L(y)=f(x)$ is given by:

$$S=y_p+S_0=\left\{y_p+y_h\middle|y_h\in\{S_0\}\right\} \text{ with } L(y_p)=f(x) \quad (6.3.815)$$

that is to say that the solutions of the form:

$$y=y_p+y_h \quad (6.3.816)$$

where y_p is a "particular/specific solution" to $L(y)=f(x)$ and y_h the "homogeneous solution" give all the D.E. solutions.

Proof 6.15.1. The first statement will be assumed obvious.

As regards to the second part, any function of the form y_p+y_h is solution of $L(y)=f(x)$.

Indeed it is trivial and it follows from the definition of the kernel concept (see section Set Theory page 404):

$$L(y_p+y_h)=L(y_p)+L(y_h)=f(x)+0=f(x) \quad (6.3.817)$$

■ Q.E.D.

What is important also to understand with the linear D.E. with second member, it is that if we find solutions to $L(y)$ with a second given member and solutions to the same D.E. with another different second member, then the sum of all these solutions will be a solution of the D.E. with the sum of the second members!!!

6.3.3.3 Resolution methods of linear differential equations

There are many ways to solve accurately or approximately linear or non-linear differential equations. Let us give the list of the few methods we will analyse further below by the example (but who are already many, many times in the chapters of Mechanics, Cosmology, Social Sciences and Quantum Physics):

- The "method of characteristic polynomial of D.E." (see below) used a bit in every section of the chapter of Mechanics/Quantum Physics/Cosmology/Chemistry and Social Sciences of this textbook.
- The "method of integrating factor" (see also below) for general knowledge but used to this date for practical cases in this textbook.
- The "method of variation of the constant" (see below) and used to this date only in the section of Industrial Engineering in this textbook.
- The "method of disturbances of D.E." (see below) useful for the wave quantum physics and quantum field theory.

Note also other widely used methods (classical high-school techniques) but that are treated case by case in the individual sections of this textbook because solving approaches are too numerous and specific to each problem:

- The "separation of variables method of D.E." as we will see a very specific and simple case just further below but for which it is best to refer to the mentioned sections (the heat equation in the section of Thermodynamics, wave equation in the section Marine & Weather Engineering, Schrödinger evolution equation in the section of Wave Quantum Physics, vibration of a drum in the section of Wave Mechanics, etc.) for concrete examples.
- The "matrix method for solving D.E." and "trivial solution of D.E." (Lotka-Volterra model in the section of Populations Dynamics, electron or nuclear spin resonance in the section of Relativistic Quantum Physics, Lorenz model in the section of Marine & Weather Engineering, etc.).
- The "spectral method" using the spectral theorem proved in the section of Linear Algebra page 2004 (see the section of Industrial Engineering for the calculation of system reliability by Markov chains for a concrete example).

- The "method of the Fourier transform of the D.E." or "method of the Laplace transform of the D.E." (see the heat equation in the section Thermodynamics, the resolution of the Black & Scholes equation in the section of Economy, the beam equation under point load in the section of Civil Engineering for practical applications).
- "Numerical methods for D.E." to solve the differential equations using computer when the D.E. have no known analytic solutions or when they have but we need a visual three dimensional view of the solutions (heat equation in the section of Numerical Methods).
- The "Frobenius method" named after Ferdinand Georg Frobenius and also "power series solutions", that is a way to find an infinite series solution for a second-order ordinary differential equation of a special form. We will use this technique in the section of Sequences and Series for our study of the Bessel series and also introduce Bessel series by solving in the section of Mechanical Engineering the problem of the self-buckling column with the power series solutions.

Remark

The first differential equations were solved around the end of the 117th century (holocene calendar) and beginning of the 118th century. By the middle of the 118th century people realized that the first methods we listed above had reached a dead end. One reason was the lack of functions to write the solutions of differential equations. The elementary functions we use in calculus, such as polynomials, quotient of polynomials, trigonometric functions, exponentials, and logarithms, were simply not enough. People even started to think of differential equations as sources to find new functions. It was matter of little time before mathematicians started to use power series expansions to find solutions of differential equations. Convergent power series define functions far more general than the elementary functions from calculus.

6.3.3.3.1 Method of characteristic polynomial

Solving simple differential equations (with constant coefficients and without second member most of the time...) uses a technique using a characteristic polynomial of the differential equation which we will see the details in the developments that follow on few special cases very frequent in physics.

It is a relatively simple method to implement when we seek solutions to the homogeneous differential equation without second member (E.W.S.M.). In the contrary case, the presence of a second member, we add the solutions of the homogeneous equation to the particular solutions.

6.3.3.3.1.1 Resolution of the H.E. of the first order L.D.E. with constant coefficients

Consider the following L.D.E. with constant coefficients:

$$\frac{\mathrm{d}y}{\mathrm{d}x}+ay=f(x) \tag{6.3.818}$$

which is a simplified version of the following general L.D.E. with constant coefficients:

$$a'\frac{\mathrm{d}y}{\mathrm{d}x}+b'y=f(x) \tag{6.3.819}$$

where:

$$a=\frac{b'}{a'} \tag{6.3.820}$$

We write its associated homogeneous equation (E.W.S.M.):

$$\frac{\mathrm{d}y}{\mathrm{d}x}+ay=0 \tag{6.3.821}$$

Which can be written:

$$\frac{\mathrm{d}y}{y}=-a\mathrm{d}x \tag{6.3.822}$$

Therefore:

$$\ln(y)=-ax+c^{te}\Rightarrow y_h=e^{-ax+c^{te}}=e^{c^{te}}a^{-ax}=Ce^{-ax} \tag{6.3.823}$$

There is behind this homogeneous solution infinite solutions: to each value given to the constant C there is a solution.

We still need to add to this homogeneous solution the particular solution y_p and for that we have a collection of recipes, depending on the type of the function $f(x)$ of the second member of the differential equation. We will see in each case in the various chapters this textbook as already mentioned.

6.3.3.3.1.2 Resolution of the H.E. of the first order L.D.E. with non-constant coefficients

The general solution of homogeneous linear differential equations (E.W.S.M.) of order 1 with non constant coefficients:

$$a(x)y'+b(x)y=0 \tag{6.3.824}$$

can always be reduced to the following form:

$$y'+g(x)y=0 \tag{6.3.825}$$

where:

$$g(x)=\frac{b(x)}{a(x)} \tag{6.3.826}$$

Well obviously there is the solution $y=0$... but let us try to do better. So we have:

$$\frac{y'}{y}=-g(x) \tag{6.3.827}$$

It comes therefore:

$$\ln(y)=-G(x)+c^{te} \tag{6.3.828}$$

where $G(x)$ is a primitive of $g(x)$. Since then:

$$y=e^{c^{te}}e^{-G(x)}=Ae^{-G(x)} \tag{6.3.829}$$

It is also common to find these developments in another notation a little bit more explicit.

So we start again of the differential equation without second member with non-constant coefficients:

$$a(x)\frac{\mathrm{d}y}{\mathrm{d}x}+b(x)y=0 \tag{6.3.830}$$

after rearrangement:

$$\frac{\mathrm{d}y}{y}=-\frac{b(x)}{a(x)}\mathrm{d}x \tag{6.3.831}$$

And then:

$$\ln(y_0)+c^{te}=-\int\frac{b(x)}{a(x)}\mathrm{d}x \tag{6.3.832}$$

Therefore:

$$y_0=c^{te}e^{-\int\frac{b(x)}{a(x)}\mathrm{d}x} \tag{6.3.833}$$

This result will be very useful to calculate the Fourier transform of a Gaussian function (see section Sequences and Series page 1765), Fourier transform, which is essential to resolve in a fairly general way the heat equation (see section Thermodynamics page 3684) resolution that will finally allow us to prove the Black & Scholes equation (see section Economy page 7804).

6.3.3.3.1.3 Resolution of the H.E. of the second order L.D.E. with constant coefficients

Consider the L.D.E. with constant coefficients:

$$\frac{\mathrm{d}^2y}{\mathrm{d}x^2}+a\frac{\mathrm{d}y}{\mathrm{d}x}+by=f(x) \tag{6.3.834}$$

which is a simplified version of the following general L.D.E. with constant coefficients:

$$a'\frac{\mathrm{d}^2y}{\mathrm{d}x^2}+b'\frac{\mathrm{d}y}{\mathrm{d}x}+c'y=f(x) \tag{6.3.835}$$

where:

$$a=\frac{b'}{a'} \qquad b=\frac{c'}{a'} \tag{6.3.836}$$

We write its homogeneous associated equation (E.W.S.M.):

$$\frac{\mathrm{d}^2y}{\mathrm{d}x^2}+a\frac{\mathrm{d}y}{\mathrm{d}x}+by=0 \tag{6.3.837}$$

wherein the function of the second member is zero. We can quite quickly consider a solution of the type (inspiring of the form of the solutions of the first order differential equations):

$$y=\lambda e^{Kx+\tau} \tag{6.3.838}$$

where τ is a constant. Which gives us therefore:

$$\lambda K^2e^{Kx+\tau}+a\lambda Ke^{Kx+\tau}+b\lambda e^{Kx+\tau}=0 \tag{6.3.839}$$

What we can simplify by:

$$K^2+aK+b=0 \tag{6.3.840}$$

If our starting assumption is correct, we only have to solve in K this "characteristic equation (CHARE)" or "characteristic polynomial" of the homogeneous equation to find the homogeneous solution:

$$K_{1,2}=\frac{-a\pm\sqrt{a^2-4b}}{2} \tag{6.3.841}$$

whose solutions depend on the sign of the discriminant of the characteristic polynomial:

$$\Delta=a^2-4b \tag{6.3.842}$$

- If the discriminant is strictly positive, that is to say $\Delta>0$:

  So we know that the characteristic polynomial has two distinct roots and then we have:

  $$y_h=Ae^{K_1(x+\tau)}+Be^{K_2(x+\delta)} \tag{6.3.843}$$

  where $K_1\tau=c^{te}$ and $K_2\delta=c^{te'}$. Then we say that the solution is "delayed" or "advanced" by the values of these constants. But the key is to note that if $y_h(x)$ is a solution, then $y_h(x\pm\Delta x)$ is always a solution!

  We then speak of "general solution of the homogeneous equation". There is behind this result an infinity of solutions: to each value given to the constants A,B corresponds a solution.

  Physicists also write sometimes this in a particular form by putting:

  $$\alpha=\frac{K_1+K_2}{2}\qquad\beta=\frac{K_1-K_2}{2} \tag{6.3.844}$$

  with then:

  $$\alpha+\beta=K_1\qquad\alpha-\beta=K_2 \tag{6.3.845}$$

  And using the hyperbolic trigonometric functions (see section Trigonometry page 2532):

  $$\begin{aligned}\cosh(\beta x)&=\frac{e^{\beta x}+e^{-\beta x}}{2}\\ \sinh(\beta x)&=\frac{e^{\beta x}-e^{-\beta x}}{2}\end{aligned} \tag{6.3.846}$$

where finally the possibility to write the homogeneous solution in the form (when we omit the advance or delay $\delta=\tau=0$):

$$y_h=e^{\alpha x}\left(C_1\cosh(\beta t)+C_2\sinh(\beta t)\right) \tag{6.3.847}$$

In addition, let us show that the solutions of the E.W.S.M. form a vector space of dimension 2 (corresponding to the order of our differential equation)!

Indeed:

- The zero function: $y=0$ is a solution of the E.W.S.M. (this is unnecessary to prove because obvious...!)
- The sum or subtraction of solutions remains a solution (this we have already proved it before)
- The elements of the basis of a vector space (the solutions of the E.W.S.M.) are linearly independent (that's an interesting property that we will need later!)

Let us put:

$$y_h=\lambda_1 f_1(x)+\lambda_2 f_2(x) \tag{6.3.848}$$

Then:

$$\begin{aligned}y'_h&=\lambda_1 f'_1+\lambda_2 f'_2\\ y''_h&=\lambda_1 f''_1+\lambda_2 f''_2\end{aligned} \tag{6.3.849}$$

These relations injected into the E.W.S.M. in generalized form:

$$a'\frac{\mathrm{d}^2y}{\mathrm{d}x^2}+b'\frac{\mathrm{d}y}{\mathrm{d}x}+c'y=0 \tag{6.3.850}$$

Then gives:

$$\begin{aligned}a'y''_h+b'y'_h+c'y_h&=a'\left[\lambda_1 f''_1+\lambda_2 f''_2\right]+b'\left[\lambda_1 f'_1+\lambda_2 f'_2\right]+c'\left[\lambda_1 f_1+\lambda_2 f_2\right]\\ &=\lambda_1\left(a'f''_1+b'f'_1+c'f_1\right)+\lambda_2\left(a'f''_2+b'f'_2+c'f_2\right)=\lambda_1\cdot 0+\lambda_2\cdot 0=0\end{aligned} \tag{6.3.851}$$

We do have indeed a vector space structure.

Let us recall that conversely two functions are linearly dependent if:

$$\forall t\quad f_1(x)=\lambda f_2(x) \tag{6.3.852}$$

- If the discriminant is strictly zero, that is to say $\Delta=0$:

  The characteristic equation has a real double root K.

  By going a little fast we would say then:

$$y_h=A^{K(x+\tau)} \tag{6.3.853}$$

and that it is over... but that it is forget that the vector base must be formed of two independent solutions!

So the second option is probably... of the form:

$$y_{h2}=z(x)A^{K(x+\tau)} \tag{6.3.854}$$

Then:

$$\begin{aligned}y'_{h2}&=Ke^{K(x+\tau)}z(x)+z'(x)e^{K(x+\tau)}\\ y''_{h2}&=K^2e^{K(x+\tau)}z(x)+Ke^{K(x+\tau)}z'(x)+Kz'(x)e^{K(x+\tau)}+z''(x)e^{K(x+\tau)}\\ &=K^2e^{K(x+\tau)}z(x)+2Ke^{K(x+\tau)}z'(x)+z''(x)e^{K(x+\tau)}\end{aligned} \tag{6.3.855}$$

If we inject it into the E.W.S.M. in generalized form:

$$a'\frac{\mathrm{d}^2y}{\mathrm{d}x^2}+b'\frac{\mathrm{d}y}{\mathrm{d}x}+c'y=0 \tag{6.3.856}$$

Then:

$$\begin{aligned}&a'\left(K^2e^{K(x+\tau)}z+2Ke^{K(x+\tau)}z'+z''e^{K(x+\tau)}\right)+b'\left(Ke^{K(x+\tau)}z+z'e^{K(x+\tau)}\right)+c'ze^{K(x+\tau)}=0\\ &e^{K(x+\tau)}\left(a'K^2+b'K+c'\right)z+e^{K(x+\tau)}(2a'K+b')z'+a'e^{K(x+\tau)}z''=0\\ &\Rightarrow\left(a'K^2+b'K+c'\right)z+(2a'K+b')z'+a'z''=0\end{aligned} \tag{6.3.857}$$

That is to say in our case:

$$(K^2+aK+b)z+(2K+a)z'+z''=0 \tag{6.3.858}$$

But, both actual real values of K are precisely solutions of:

$$K^2+aK+b=0 \tag{6.3.859}$$

The prior-previous relation is reduced to:

$$(2K+a)z'+z''=0 \tag{6.3.860}$$

and as we are in the case of study where the discriminant is zero, we have:

$$K_{1,2}=\frac{-a\pm\Delta}{2}=\frac{-a}{2} \tag{6.3.861}$$

Therefore and finally the prior-previous relation reduces to:

$$z''=0 \tag{6.3.862}$$

We deduce from it that:

$$z''=0\Rightarrow z'=c^{te}\Rightarrow z=Bx+c^{te} \tag{6.3.863}$$

Therefore finally:

$$y_{h2}=(Bx+c^{te})e^{K(x+c^{te})} \tag{6.3.864}$$

Which gives for the general solution of the E.W.S.M.:

$$y_h=(C_1x+C_2)e^{K(x+\tau)} \tag{6.3.865}$$

- If the discriminant is strictly negative, that is to say $\Delta < 0$:

  The characteristic equation has two complex conjugate roots (see section Calculus page 1383):

  $$K_1 = \alpha + \mathrm{i}\beta \qquad K_2 = \alpha - \mathrm{i}\beta \tag{6.3.866}$$

  Therefore:

  $$\begin{aligned} y_h &= Ae^{(\alpha+\mathrm{i}\beta)(x+\tau)} + Be^{(\alpha-\mathrm{i}\beta)(x+\delta)} \\ &= Ae^{\alpha(x+\tau)}e^{\mathrm{i}\beta(x+\tau)} + Be^{\alpha(x+\delta)}e^{-\mathrm{i}\beta(x+\delta)} = e^{\alpha(x+\tau)}Ae^{\mathrm{i}\beta(x+\tau)} + e^{\alpha(x+\delta)}B^{-\mathrm{i}\beta(x+\delta)} \\ &= Ae^{\alpha(x+\tau)}\left(\cos(\beta x + \tau\beta) + \mathrm{i}\sin(\beta x + \tau\beta)\right) + Be^{\alpha(x+\delta)}\left(\cos(\beta x + \delta\beta) + \mathrm{i}\sin(\beta x + \delta\beta)\right) \\ &= y_{h1} + y_{h2} \end{aligned} \tag{6.3.867}$$

  But if we look instead for real solutions, we can always put A and B as being equal such as:

  $$\begin{aligned} y'_h &= \frac{1}{2}(y_{h1} + y_{h2}) + \frac{1}{2\mathrm{i}}(y_{h1} - y_{h2}) = e^{\alpha(x+\tau)}\cos(\beta x + \tau\beta) + e^{\alpha(x+\delta)}\sin(\beta x + \delta\beta) \\ &= e^{\alpha x}e^{\alpha\tau}\cos(\beta x + \tau\beta) + e^{\alpha x} + e^{\alpha\delta}\sin(\beta x + \delta\beta) = e^{\alpha x}\left(A'\cos(\beta x + \tau\beta) + B'\sin(\beta x + \delta\beta)\right) \end{aligned} \tag{6.3.868}$$

  And if we set the delay and advance respectively as being zero ($\delta = \tau = 0$), so we find the relation available in most books without proof:

  $$y'_h = e^{\alpha x}\left(A'\cos(\beta x) + B'\sin(\beta x)\right) \tag{6.3.869}$$

  where A' and B' are any two real constants. There is another important form of this last relation (often used in electronics, for example). Indeed, it is possible for any A' and $B' \in \mathbb{R}$, to find C' and ϕ also in $\mathbb{R}$ such as the following equality holds:

  $$A'\cos(\beta x) + B'\sin(\beta x) = C'\cos(\beta t - \phi) \tag{6.3.870}$$

  We put:

  $$C' = \sqrt{A'^2 + B'^2} \tag{6.3.871}$$

  Then:

  $$A'\cos(\beta x) + B'\sin(\beta x) = C'\left[\frac{A'}{M}\cos(\beta x) + \frac{B'}{M}\sin(\beta x)\right] \tag{6.3.872}$$

  It is then possible to find ϕ such as:

  $$\frac{A'}{M} = \cos(\phi) \qquad \frac{B'}{M} = \sin(\phi) \tag{6.3.873}$$

  Therefore our initial expression (proposition) can be written as:

  $$A'\cos(\beta x) + B'\sin(\beta x) = C'\left[\cos(\phi)\cos(\beta x) + \sin(\phi)\sin(\beta x)\right] = C'\cos(\beta t - \phi) \tag{6.3.874}$$

  Finally:

  $$y_h = C'e^{\alpha x}\cos(\beta t - \phi) \tag{6.3.875}$$

So we can make the following summary:

Discriminant	Roots	Homogeneous solution
$\Delta > 0$	$K_{1,2} = \dfrac{-a \pm \sqrt{a^2 - 4b}}{2}$	$y_h = Ae^{K_1(x+\tau)} + Be^{K_2(x+\delta)}$
$\Delta = 0$	$K_{1,2} = \dfrac{-a \pm \sqrt{a^2 - 4b}}{2}$	$y_h = (C_1 x + C_2)e^{K(x+\tau)}$
$\Delta < 0$	$K_{1,2} = \dfrac{-a \pm \sqrt{a^2 - 4b}}{2}$	$y_h = e^{\alpha x}(A' \cos(\beta x) + B' \sin(\beta x))$

Table 6.13: **Typical solutions of H.E. of The L.D.E. with constant coefficients**

6.3.3.3.2 Integrating Factor Method (Euler's Method)

The technique of integrating factor is useful when it comes to solve differential equations of the form:

$$y' + f(x)y = g(x) \tag{6.3.876}$$

We will us that technique during our study of the RC circuit response to a linear increase in voltage (see section Electric Engineering page 8471).

The basic idea is to find a function $M(x)$, named "integration factor", by which can be multiplied by our differential equation to bring the left-hand side of equality to a simple derivative. For example, for a linear differential equation as the one above, we choose often the following integration factor (but this is by far not the only possibility and this choice does not solve everything!):

$$\text{IF} = e^{F(x)} = e^{\int f(x)\mathrm{d}x} \tag{6.3.877}$$

Therefore we have:

$$e^{F(x)}(y' + f(x)y) = g(x)e^{F(x)} \tag{6.3.878}$$

or by distributing:

$$y'e^{F(x)} + f(x)ye^{F(x)} = g(x)e^{F(x)} \tag{6.3.879}$$

Which can therefore be seen as:

$$y'e^{F(x)} + F'(x)ye^{F(x)} = g(x)e^{F(x)} \tag{6.3.880}$$

or even better (and therein lies the whole trick)...:

$$\left(ye^{F(x)}\right)' = g(x)e^{F(x)} \tag{6.3.881}$$

We can then take the primitive with respect to x:

$$\int \left(ye^{F(x)}\right)' \mathrm{d}x = \int g(x)e^{F(x)}\mathrm{d}x \tag{6.3.882}$$

and trivially (!) we have the left primitive that is immediate:

$$ye^{F(x)} = \int g(x)e^{F(x)}\mathrm{d}x \tag{6.3.883}$$

Which is sometimes written as:

$$y = \frac{1}{e^{F(x)}} \int g(x)e^{F(x)}\mathrm{d}x = \frac{1}{e^{\int f(x)\mathrm{d}x}} \int g(x)e^{\int f(x)\mathrm{d}x}\mathrm{d}x \tag{6.3.884}$$

✍**Example:**

Consider the following differential equation:

$$y'+\frac{y}{x}-4=0 \tag{6.3.885}$$

That we will rewrite as:

$$\frac{\mathrm{d}y}{\mathrm{d}x}+\frac{1}{x}y=4 \tag{6.3.886}$$

We see then that (assuming x is positive):

$$f(x)=\frac{1}{x}\Rightarrow F(x)=\ln(x) \tag{6.3.887}$$

Then we have:

$$e^{\ln(x)}\left(\frac{\mathrm{d}y}{\mathrm{d}}x+\frac{1}{x}y\right)=e^{\ln(x)}4 \tag{6.3.888}$$

Hazard making sometimes things good (the example is purposely very simple), we have this equality that simplified as:

$$e^{\ln(x)}=x \tag{6.3.889}$$

in:

$$x\left(\frac{\mathrm{d}y}{\mathrm{d}}x+\frac{1}{x}y\right)=4x \tag{6.3.890}$$

Which may be condense in:

$$(xy)'=4x \tag{6.3.891}$$

By integrating:

$$\int(xy)'\mathrm{d}x=\int 4x\mathrm{d}x \tag{6.3.892}$$

It then comes immediately:

$$xy=2x^2+c^{te} \tag{6.3.893}$$

Finally:

$$y=2x+\frac{x}{c^{te}} \tag{6.3.894}$$

6.3.3.3.3 Method of separation of variables

The method of separation of variables (also known as the "Fourier method") is any of several methods for solving ordinary and partial differential equations, in which algebra allows one to rewrite an equation so that each of two variables occurs on a different side of the equation.

In mathematics, a "partial differential equation PDE" is a differential equation that contains unknown multivariable functions and their partial derivatives (a special case are ordinary differential equations, which deal with functions of a single variable and their derivatives). PDEs are used to formulate problems involving functions of several variables, and are either solved by hand, or used to create a relevant computer model.

PDEs can be used to describe a wide variety of phenomena such as sound, heat, electrostatics, electrodynamics, fluid dynamics, elasticity, or quantum mechanics. These seemingly distinct physical phenomena can be formalized similarly in terms of PDEs. Just as ordinary differential equations often model one-dimensional dynamical systems, partial differential equations often model multidimensional systems. PDEs find their generalization in stochastic partial differential equations.

A partial differential equation (PDE) for the function $U(x_1,\cdots,x_n)$ is an equation of the form:

$$\boxed{f\left(x_1,\ldots,x_n,U,\frac{\partial U}{\partial x_1},\ldots,\frac{\partial U}{\partial x_n},\frac{\partial^2 U}{\partial x_1\partial x_1},\ldots,\frac{\partial^2 U}{\partial x_1\partial x_n},\ldots\right)=0} \qquad (6.3.895)$$

If f is a linear function of U and its derivatives, then the PDE is named a "linear partial differential equation". Common examples of linear PDEs include the heat equation, the wave equation, Laplace's equation, Helmholtz equation, Klein–Gordon equation, and Poisson's equation (see the chapters of Mechanics, Electrodynamics and Quantum Physics for the study of most of them!).

The method of separation of variables is a very common technique used in physics when we have second-order differential equations. Many useful examples and very detailed are already in the various chapters already mentioned above. Here we will just present a special case just for doing things good but with the minimum subsistence level!

Consider the common case of physical partial differential equation of the type:

$$\frac{\partial^2 U}{\partial x^2}-\frac{\partial U}{\partial y}=0 \qquad (6.3.896)$$

The solution of this equation therefore requires finding a function U which depends on x and y such that:

$$U=U(x,y) \qquad (6.3.897)$$

In physics, the idea is then to put that we can always find a said a "separable" solution of the form:

$$U(x,y)=X(x)Y(y) \qquad (6.3.898)$$

Thus, the differential equation can be written as:

$$\frac{\partial^2 X(x)Y(y)}{\partial x^2}-\frac{\partial X(x)Y(y)}{\partial y}=0 \qquad (6.3.899)$$

Which can be rewritten as:

$$Y(y)\frac{\partial^2 X(x)}{\partial x^2}=X(x)\frac{\partial Y(y)}{\partial y} \qquad (6.3.900)$$

Or:

$$\frac{1}{X(x)}\frac{\partial^2 X(x)}{\partial x^2}=\frac{1}{Y(y)}\frac{\partial Y(y)}{\partial y} \qquad (6.3.901)$$

After rearrangement it is use in physics to note this last equality in condensed form:

$$\frac{X''}{X}=\frac{Y''}{Y} \qquad (6.3.902)$$

This equality can only take place if each term is a constant since X depends only on x and Y only on y. It comes then:

$$\frac{X''}{X}=c^{te} \qquad \frac{Y''}{Y}=c^{te} \qquad (6.3.903)$$

And each differential equation can then be solved independently of the other and once the solutions found we multiplied them determine the expression of U.

Remark

Probably the most beautiful example is the section of Quantum Chemistry.

How do we know that this technique of separation of variables is valid? When the differential operators in various variables are additive in the partial differential equation, that is, when there are no products of differential operators in different variables, the separation method usually works. We are proceeding in the spirit of *let's try and see if it works*... If our attempt succeeds, then this technique will be justified. If it does not succeed, we shall find out soon enough and then we shall try another attack, such as Green's functions, integral transforms, or brute-force numerical analysis!

6.3.3.3.4 Method of constant variation

The variation of constants, is a general method to solve inhomogeneous linear ordinary differential equations.

For first-order inhomogeneous linear differential equations it is usually possible to find solutions via integrating factors or undetermined coefficients with considerably less effort, although those methods leverage heuristics that involve guessing and don't work for all inhomogeneous linear differential equations.

Variation of parameters extends to linear partial differential equations as well, specifically to inhomogeneous problems for linear evolution equations like the heat equation, wave equation, and vibrating plate equation. In this setting, the method is more often known as "Duhamel's principle", named after Jean-Marie Duhamel who first applied the method to solve the inhomogeneous heat equation.

The idea of the method of variation of the constant is as follows: if we have a particular solution affected by constants, we know that depending of the initial conditions thereof are well determined. The idea is then to generalize by putting that these constants are functions. In some cases obviously mathematical developments will show that the functions are necessarily constant.

The underlying idea behind this method is to say that the solutions of the (linear) differential equation with the second member will look like the solutions of the homogeneous equation. As the term on the right will disrupt this solution, we vary only constants (which will therefore no longer be constants), but we remain on the "base" of homogeneous solutions, to seek close solutions. After we check that this "physicist reasoning" gives out all the solutions of the differential equation.

Let us see before studying to the general case a simple example by considering the following differential equation:

$$y'(t)+by(t)=1 \tag{6.3.904}$$

for which the particular solution of the homogeneous equation (E.W.S.M.) is (if you need the details not hesitate to ask!):

$$y_0(t)=Ae^{-bt} \tag{6.3.905}$$

The method of variation of the constant consists then to put that:

$$A=A(t) \tag{6.3.906}$$

and therefore:

$$y_0'(t)=A'e^{-bt}-bA(t)e^{-bt} \tag{6.3.907}$$

But by the differential equation with the second member, we have:

$$y_0'(t)+by_0(t)=1\Rightarrow y_0'(t)=1-by_0(t)=1-bA^{-bt} \tag{6.3.908}$$

So using the last two relations it comes:

$$\left.\begin{array}{l} y_0'(t)=A'(t)e^{-bt}-bA(t)e^{-bt} \\ y_0'(t)=1-bAe^{-bt} \end{array}\right\}\Rightarrow A'(t)=e^{bt} \tag{6.3.909}$$

and it follows that:

$$A(t)=\frac{1}{b}e^{bt} \tag{6.3.910}$$

where we eliminated the integration constant because what we want is a particular solution! The particular general solution (pg) is then the sum of the particular solution and the homogeneous one and this with the variation of the constant:

$$y_{0,\mathbf{pg}}(t)=Ae^{-bt}+A'(t)e^{-bt}=A^{-bt}+\frac{1}{b}e^{bt}e^{-bt}=Ae^{-bt}+\frac{1}{b} \tag{6.3.911}$$

Thus, generalizing the previous example, so we have a differential equation of the form:

$$y'(t)+by(t)=c \tag{6.3.912}$$

The general particular solution will then be:

$$C_0y_0(t):=C_0(t)y_0(t) \tag{6.3.913}$$

Then we have:

$$y_0'(t)=C_0'y_0(t)+C_0(t)y_0'(t) \tag{6.3.914}$$

hence injected into the original differential equation:

$$\underbrace{C_0'y_0(t)+C_0(t)y_0'(t)}_{y'(t)}+b(t)\underbrace{C_0(t)y_0(t)}_{y(t)}=c(t) \tag{6.3.915}$$

Therefore after factoring similar terms:

$$C_0'y_0(t)+C_0(t)\left(y_0'(t)+b(t)C_0(t)y_0(t)\right)=c(t) \tag{6.3.916}$$

So we have the above relation and the particular solution to the homogeneous differential equation (therefore without second member):

$$\left.\begin{array}{l} C_0'y_0(t)+C_0(t)\left(y_0'(t)+b(t)C_0(t)y_0(t)\right)=c(t) \\ y_0'(t)=-b(t)C_0(t)y_0(t) \end{array}\right\}\Rightarrow C_0'y_0(t)=c(t) \tag{6.3.917}$$

Therefore we find:

$$C_0'=\frac{c(t)}{y_0(t)} \tag{6.3.918}$$

and it sufficient then to integrate this equation to find $C_0(t)$. Then, the particular general solution (pg) will be the sum of the particular homogeneous solution and that with the variation of the constant.

6.3.3.4 Classification of differential equations

Before we begin, the reader may wonders what classifying differential equations can be used for, well, here are the two main arguments for the usefulness of a classification in the order of importance:

- Many books have authors who systematically speak of a differential equation by categorizing it, so it is more pleasant to know what they are talking about.
- Some finite element modelling software (including MATLAB®) requires that the differential equation category be chosen before anything else can be done as a calculation.

First a reminder of what we have seen so far:

- **Ordinary Differential Equation:** It is a differential equation that involves one or more ordinary derivatives but without having partial derivatives. Like for example:

$$m\frac{\mathrm{d}^2x}{\mathrm{d}t^2}=f(x)$$

- **Partial Differential Equation:** It is a differential equation that involves partial derivatives. It has two or more independent variables. Like for example:

$$\frac{\partial^2 u}{\partial x^2}+4xy\frac{\partial^2 u}{\partial y^2}+u=2\left(\frac{\partial^2 u}{\partial x^2}\right)^2+4\frac{\partial^3 u}{\partial x\partial y^2}=10x$$

- **Linear Differential Equation:** It is first degree with respect to the dependent variable(s) and its derivatives that can be expressed in the form:

$$\frac{\mathrm{d}y}{\mathrm{d}x}+p(x)y=q(x)$$

- **Non-Linear Differential Equation:** It is second degree or higher differential equation with respect to dependent variables and its derivatives. Like for example

$$\frac{\mathrm{d}^2y}{\mathrm{d}x^2}-\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2+12y=\cos(x)$$

- **Non-homogeneous Differential Equation:** It is a differential equation whose right-hand side is not equal to zero. A second order non-homogeneous differential equation can be written in this form:

$$y''+f(x)y'+g(x)y=r(x)$$

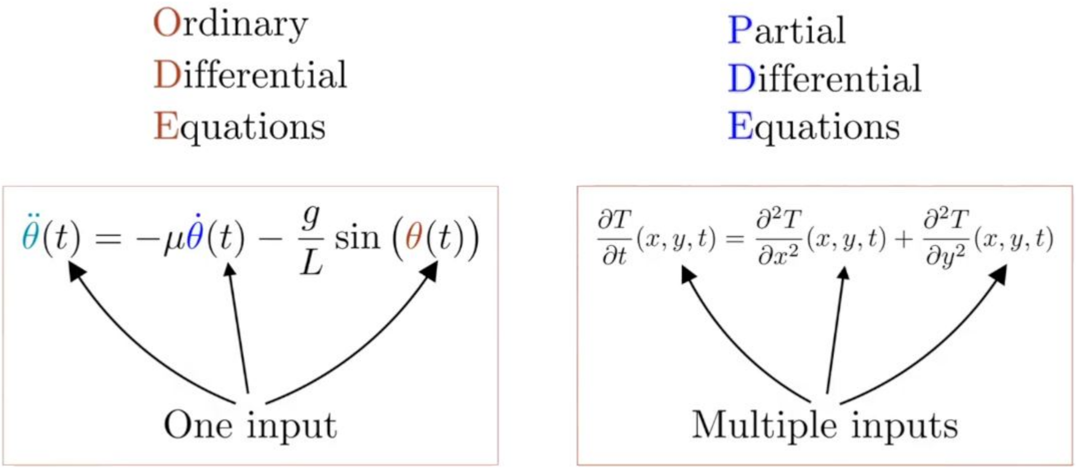
So this being said, formally, we name "partial differential equation PDE" of order less than or equal to 2 in a domain $\Omega\subset\mathbb{R}^n$ and of unknown:

$$\Phi(x)=\Phi(x_1,x_2,\ldots,x_N):\Omega\mapsto\mathbb{R} \qquad (6.3.919)$$

an equation of the following general type:

$$\boxed{\sum_{i,j}^{n}\left(a_{ij}(x)\frac{\partial^2\Phi(x)}{\partial x_i\partial x_j}\right)+\sum_{i}^{n}\left(b_i(x)\frac{\partial\Phi(x)}{\partial x_i}\right)+c(x)\Phi(x)=s(x)} \tag{6.3.920}$$

where $s(x)$ is often named a "source term" in analogy with the main situation in physics concerned.


Ordinary
Differential
Equations
$\ddot{\theta}(t)=-\mu\dot{\theta}(t)-\frac{g}{L}\sin\left(\theta(t)\right)$
One input
Partial
Differential
Equations
$\frac{\partial T}{\partial t}(x,y,t)=\frac{\partial^2 T}{\partial x^2}(x,y,t)+\frac{\partial^2 T}{\partial y^2}(x,y,t)$
Multiple inputs

It is now important to generalize the latter equation in vector form. Thus, we introduce the following notations:

- $A=[a_{ij}]$ the $n\times n$ symmetric matrix of the coefficients in front of the terms of order 2
- $B=(f_i(x))$ the vector of size n of the coefficients in front of the terms of order 1
- $[H\Phi(x)]$ the $n\times n$ symmetric Hessian matrix (see section Sequences and Series page 1715) of Φ:

$$H\Phi(x)_{ij}=\frac{\partial^2\Phi(x)}{\partial x_i\partial x_j} \tag{6.3.921}$$

- $\vec{\nabla}(\Phi(x))$ the vector of size n of Φ:

$$\vec{\nabla}(\Phi(x))=\left(\frac{\partial\Phi(x)}{\partial x_i}\right) \tag{6.3.922}$$

- The notation (named "Frobenius (matrix) dot product"):

$$A:B=\text{tr}\left(A^TB\right)=\sum_{i,j}^{n}a_{ij}b_{ij} \tag{6.3.923}$$

also sometimes denoted as the classical dot product: $\langle A,B\rangle$.

Remark

The "Frobenius norm" (very useful to derive the generalised singular value decomposition used massively in Data Science) of a matrix $\mathbb{R}^{m\times n}$ is given by:

$$\|X\|_F=\left(\text{tr}\left(X^TX\right)\right)^{1/2}=\left(\sum_{i=1}^{m}\sum_{j=1}^{n}x_{ij}^2\right)^{1/2}\tag{6.3.924}$$

More explicitly let A be an $m\times n$ matrix in $\mathbb{C}$ (i.e. $A_{mn}\mathbb{C}()$) and denote by a_{ij} the entry of A at row i and column j. Hence, we have:

$$A^TA=\begin{bmatrix}\sum_{i=1}^m|a_{i1}|^2 & \sum_{i=1}^m\bar{a}_{i1}a_{i2} & \ldots & \sum_{i=1}^m\bar{a}_{i1}a_{in}\\ \sum_{i=1}^m\bar{a}_{i2}a_{i1} & \sum_{i=1}^m|a_{i2}|^2 & \ldots & \sum_{i=1}^m\bar{a}_{i2}a_{in}\\ \vdots & \vdots & \ddots & \vdots\\ \sum_{i=1}^m\bar{a}_{n1}a_{i1} & \sum_{i=1}^m\bar{a}_{n1}a_{i2} & \ldots & \sum_{i=1}^m|a_{in}|^2\end{bmatrix}$$

Then obviously:

$$\text{tr}\left(A^TA\right)=\sum_{j=1}^{n}\sum_{i=1}^{m}|a_{ij}|^2=\left(\sqrt{\sum_{j=1}^{n}\sum_{i=1}^{m}|a_{ij}|^2}\right)^2=\|A\|_F^2$$

With this, the previous relation:

$$\sum_{i,j}^{n}\left(a_{ij}(x)\frac{\partial^2\Phi(x)}{\partial x_i\partial x_j}\right)+\sum_{i}^{n}\left(b_i(x)\frac{\partial\Phi(x)}{\partial x_i}\right)+c(x)\Phi(x)=s(x)\tag{6.3.925}$$

can be rewritten as:

$$\boxed{A(x):[H\Phi(x)]+B(x)\circ(\vec{\nabla}(\Phi(x)))+c(x)\Phi(x)=s(x)}\tag{6.3.926}$$

What is often (abusively) written in a very condensed way:

$$AH\Phi+B\vec{\nabla}\Phi+c\Phi=s\tag{6.3.927}$$

✍**Example:**

E1. Let us consider the following PDE:

$$\frac{\partial^2\Phi}{\partial x^2}(x,y)+\frac{\partial^2\Phi}{\partial x}+\frac{\partial\Phi}{\partial x}(x,y)=\sin(x,y)\tag{6.3.928}$$

First, it is easy to determine (because it is a definition) that:

$$[H\Phi(x)]=\begin{bmatrix}\dfrac{\partial^2\Phi}{\partial^2x}(x,y) & \dfrac{\partial^2\Phi}{\partial x\partial y}(x,y)\\ \dfrac{\partial^2\Phi}{\partial y\partial x}(x,y) & \dfrac{\partial^2\Phi}{\partial^2y}(x,y)\end{bmatrix}\qquad\vec{\nabla}(\Phi(x))=\begin{pmatrix}\dfrac{\partial\Phi}{\partial x}(x,y)\\ \dfrac{\partial\Phi}{\partial y}(x,y)\end{pmatrix}\tag{6.3.929}$$

It is also trivial that:

$$c(x,y)=0 \tag{6.3.930}$$

and that:

$$s(x,y)=\sin(x+y) \tag{6.3.931}$$

It is also trivial that:

$$B=\begin{pmatrix}1\\0\end{pmatrix} \tag{6.3.932}$$

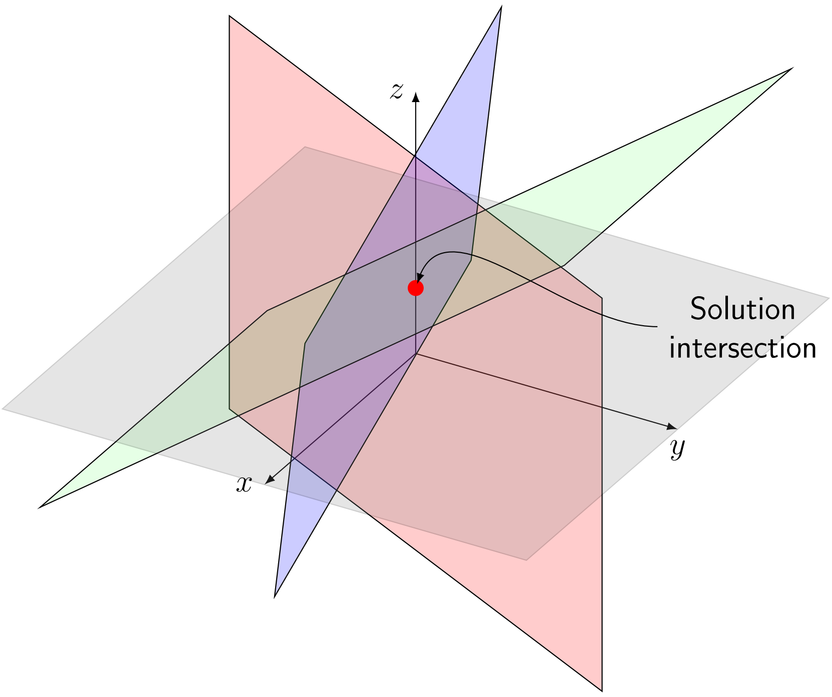
Finally, the minor difficulties are to find:

$$a=\begin{pmatrix}1 & \frac{1}{2}\\ \frac{1}{2} & 0\end{pmatrix} \tag{6.3.933}$$

E2. Let us consider the following system:

$$\begin{cases} 2x-\ \ y \quad\quad =0\\ -x+2y-\ \ z=-1\\ \quad -3y+4z=4\end{cases} \tag{6.3.934}$$

That has a unique solution that can be depicted by the following crossing of planes (this time made with TikZ rather than Maple™):



For linear PDE of order 2, the matrix A is non-zero and is symmetric. It is therefore diagonalizable with real eigenvalues (see section Linear Algebra page 1997) and their study provides elements of classification of linear PDE systems of order 2 under the denomination of "elliptic", "hyperbolic" or "parabolic" PDE (see section Analytical Geometry page 2794).

It is often customary to say in physics that the elliptics PDE characterize problems of equilibrium or stationarity, that hyperbolics PDE characterize problems of wave propagations (the Einstein field equations can in some special cases be reduced to this class of PDE) and finally that parabolic PDE characterize diffusion problems (see examples further below).

That latter terminology comes from the fact that when the matrix A is constant, the curves:

$$x^TAx=c^{te} \tag{6.3.935}$$

are respectively ellipsoids, hyperboloid and paraboloid (see section Analytical Geometry page 2798).

Indeed, we have proved in the section of Analytical Geometry that following the determinant of A, we had:

- If $\det(A)=ac-b^2>0$, the curve Γ is either empty, reduced to a point, or an ellipse.
- If $\det(A)=ac-b^2<0$, the curve Γ is the union of two intersecting lines, that is to say a hyperbola.
- If $\det(A)=ac-b^2=0$, the curve Γ is either empty, a line, or two distinct parallel lines, or a parabola.

More explicitly what we have seen above can be reformulated as following. If we have the following PDE:

$$a\frac{\partial^2\Phi}{\partial x^2}+b\frac{\partial^2\Phi}{\partial x\partial y}+c\frac{\partial^2\Phi}{\partial y^2}+d\frac{\partial\Phi}{\partial x}+e\frac{\partial\Phi}{\partial y}+fu=0 \tag{6.3.936}$$

where a, b, c, d, e and f are real constants, that latter is said to be:

- An "elliptical partial differential equation" if $ac-b^2>0$ that is to say if the eigenvalues are all positive or all negative (see section Analytical Geometry page 2798).
- A "hyperbolic partial differential equation" if $ac-b^2<0$ that is to say there is only one negative eigenvalue and all the others are positive, or there is only one positive eigenvalue and all the others are negative (see section Analytical Geometry page 2798).
- A "parabolic partial differential equation" if $ac-b^2=0$ that is to say if the eigenvalues are all positive or all negative, save one that is zero (see section Analytical Geometry page 2798).

✍**Example:**

The Laplace's equation:

$$\nabla^2\phi=0 \tag{6.3.937}$$

is an elliptic PDE.

The wave equation:

$$\frac{\partial^2 U}{\partial t^2}=v^2\frac{\partial^2 U}{\partial x^2} \tag{6.3.938}$$

is a hyperbolic PDE.

The heat equation:

$$\frac{\partial U}{\partial t}=k\frac{\partial^2 U}{\partial x^2} \tag{6.3.939}$$

is a parabolic PDE.

6.3.3.5 Systems of Differential Equations

Let us study now special developments that will also be useful for the study of quantum physics or for the resolution of particular systems of differential equations (see corresponding section in this textbook for the details on these examples) and especially one which is well known in Chaos Theory!

Let us first indicate to the reader before going further that more complex inhomogeneous case (with second member) and with unknown coefficients is treated directly by an example in the section of Industrial Engineering during the study of the reliability of a repairable system as a Markov chain with resolution using the determinants and eigenvalues/eigenvectors.

To start this first approach, we will have to introduce the concept of exponentiation of a matrix!

6.3.3.5.1 Matrix exponentiation method

The set of matrices $n \times n$ with coefficients in $\mathbb{C}$ denoted $M_n(\mathbb{C})$ is a vector space for the addition of matrices and multiplication by a scalar. We will as always denote by $\mathbb{1}_n$ the identity matrix of dimension n.

We will admit that a sequence of matrices A_n converges to a matrix A if and only if the sequences of coefficients of the matrices A_n converges towards the corresponding coefficients of A.

✍**Example:**

In $M_2(\mathbb{C})$ the sequence of matrices:

$$\begin{pmatrix} \dfrac{1}{n} & 2 \\ 0 & \dfrac{n+1}{n} \end{pmatrix} \tag{6.3.940}$$

converges to:

$$\begin{pmatrix} 0 & 2 \\ 0 & 1 \end{pmatrix} \tag{6.3.941}$$

when $n \rightarrow +\infty$. If $x \in \mathbb{C}$, we saw in our study of complex numbers (see section on Numbers page 251) that the series:

$$\sum_{k=0}^{+\infty} \frac{x^k}{k!} \tag{6.3.942}$$

does converge and its limit is denoted by e^x. In fact here there is no difficulty in replacing x by a matrix A since we know (we have proved it during our study of complex numbers) that any complex number can be written as follows (the body of complex numbers is isomorphic to the field of real matrices of square dimensions 2 having this form):

$$x = a + \mathrm{i}b = a\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} b = \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \tag{6.3.943}$$

and that a complex number is equivalent to put his matrix form also to the square:

$$x^2=(a+\mathrm{i}b)(a+\mathrm{i}b)=a^2+aib+aib-b^2=(a^2-b^2)+2ab\mathrm{i}=\begin{bmatrix} a^2-b^2 & 2ab \\ 2ab & a^2-b^2 \end{bmatrix} \tag{6.3.944}$$

Indeed:

$$\begin{bmatrix} a & -b \\ b & a \end{bmatrix}^2=\begin{bmatrix} a & -b \\ b & a \end{bmatrix}\begin{bmatrix} a & -b \\ b & a \end{bmatrix}=\begin{bmatrix} a^2-b^2 & 2ab \\ 2ab & a^2-b^2 \end{bmatrix} \tag{6.3.945}$$

We then define the "exponential of a matrix" $A\in M_n(\mathbb{C})$ as the matrix limit of the sequence:

$$e^A=\sum_{k=0}^{+\infty}\frac{A^k}{k!} \tag{6.3.946}$$

If the matrix A is diagonal obviously its exponential is easy to calculate. Indeed, if:

$$A=\begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}\Rightarrow A^k=\begin{bmatrix} \lambda_1^k & & 0 \\ & \ddots & \\ 0 & & \lambda_n^k \end{bmatrix} \tag{6.3.947}$$

It follows:

$$e^A=\sum_{k=0}^{+\infty}\frac{1}{k!}\begin{bmatrix} \lambda_1^k & & 0 \\ & \ddots & \\ 0 & & \lambda_n^k \end{bmatrix}=\begin{bmatrix} \displaystyle\sum_{k=0}^{+\infty}\frac{\lambda_1^k}{k!} & & 0 \\ & \ddots & \\ 0 & & \displaystyle\sum_{k=0}^{+\infty}\frac{\lambda_n^k}{k!} \end{bmatrix}=\begin{bmatrix} e^{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & e^{\lambda_n} \end{bmatrix} \tag{6.3.948}$$

This property of diagonal matrices will also be very useful to us in the section of General Relativity during our study of the variation of the metric determinant.

However, it is clear that a non-diagonal matrix will be much more complicated to deal with! We will then use a diagonalization technique or an endomorphisms reduction[8] (see section Linear Algebra page 1934) to simplify our problem.

So notice that if $S\in M_n(\mathbb{C})$ is reversible and if $A\in M_n(\mathbb{C})$ then:

$$e^{SAS^{-1}}=Se^AS^{-1} \tag{6.3.949}$$

This follows from the fact that (think to the change of basis of a linear application as it has been studied in the section of Linear Algebra page 1992... perhaps it may help):

$$\left(SAS^{-1}\right)^k=SA\underbrace{S^{-1}S}_{=1}A\underbrace{S^{-1}S}_{=1}AS^{-1}\ldots\underbrace{S^{-1}S}_{=1}AS^{-1}=SA^kS^{-1} \tag{6.3.950}$$

---

[8] In linear algebra, an "endomorphism reduction" aims to express matrices and endomorphisms in a simpler form, for example to facilitate calculations. This consists essentially in finding a decomposition of the vector space into a direct sum of stable subspaces on which the induced endomorphism is simpler. Less geometrically, this corresponds to finding a basis of the space in which the endomorphism is simply expressed like for example the spectral theorem (see section Linear Algebra page 2004).

Therefore:

$$e^{SAS^{-1}}=\sum_{k=0}^{+\infty}\frac{SA^kS^{-1}}{k!}=S\left(\sum_{k=0}^{+\infty}\frac{A^k}{k!}\right)S^{-1} \tag{6.3.951}$$

This development will enable us to bring the computing of the exponential of a diagonalizable matrix in search of its eigenvalues and its eigenvectors.

✍Examples:

Let us calculate e^A where:

$$A=\begin{pmatrix}3 & -4\\ -6 & 1\end{pmatrix} \tag{6.3.952}$$

The eigenvalues of A are $\lambda_1=-3$, $\lambda_2=7$, and associated eigenvectors are:

$$\vec{v}_1=\begin{pmatrix}2\\ 3\end{pmatrix}\qquad \vec{v}_2=\begin{pmatrix}1\\ -1\end{pmatrix} \tag{6.3.953}$$

Indeed:

$$\begin{pmatrix}3 & -4\\ -6 & 1\end{pmatrix}\begin{pmatrix}2\\ 3\end{pmatrix}=-3\begin{pmatrix}2\\ 3\end{pmatrix}\text{ and }\begin{pmatrix}3 & -4\\ -6 & 1\end{pmatrix}\begin{pmatrix}1\\ -1\end{pmatrix}=7\begin{pmatrix}1\\ -1\end{pmatrix} \tag{6.3.954}$$

By putting:

$$S=\begin{pmatrix}2 & 1\\ 3 & -1\end{pmatrix} \tag{6.3.955}$$

We get:

$$A=S\begin{pmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{pmatrix}S^{-1}=S\begin{pmatrix}-3 & 0\\ 0 & 7\end{pmatrix}S^{-1} \tag{6.3.956}$$

with:

$$S^{-1}=S\begin{pmatrix}\frac{1}{5} & \frac{1}{5}\\ \frac{3}{5} & -\frac{2}{5}\end{pmatrix} \tag{6.3.957}$$

Therefore:

$$\begin{aligned}e^A&=e^{S\begin{pmatrix}-3 & 0\\ 0 & 7\end{pmatrix}S^{-1}}=Se^{\begin{pmatrix}-3 & 0\\ 0 & 7\end{pmatrix}}S^{-1}=S\begin{pmatrix}e^{-3} & 0\\ 0 & e^7\end{pmatrix}S^{-1}\\ &=\frac{1}{5}\begin{pmatrix}2e^{-3}+3e^7 & 2e^{-3}-2e^7\\ 2e^{-3}-3e^7 & 2e^{-3}+2e^7\end{pmatrix}\end{aligned} \tag{6.3.958}$$

Now, let us recall that in the case of real numbers we know that if $x,y\in\mathbb{R}$ then:

$$e^{x+y}=e^xe^y \tag{6.3.959}$$

In the case of matrices we can prove that if $A,B\in M_n(\mathbb{C})$ are two matrices that commute with one another, that is to say such that $AB=BA$, then:

$$e^{A+B}=e^Ae^B \tag{6.3.960}$$

The condition of commutativity comes from the fact that the addition in the exponential is itself commutative. The proof is therefore intuitive.

An important corollary of this proposition is that for any matrix $A\in M_n(\mathbb{C})$, e^A is reversible. Indeed the matrices A and $-A$ commute and therefore:

$$e^Ae^{-A}=e^{A-A}=e^0=I_n=\mathbb{1} \tag{6.3.961}$$

We recall that a matrix A with complex coefficients is unitary if:

$$AA^\dagger=I_n \tag{6.3.962}$$

The following theorem will serve us later:

Theorem 6.16. Let us prove that if A is a Hermitian matrix (also named "self-adjoint") (see section Linear Algebra page 1953) then for any $t\in\mathbb{R}$, $e^{\mathrm{i}tA}$ is unitary.

Proof 6.16.1.

$$\left(e^{\mathrm{i}tA}\right)^\dagger=e^{(\mathrm{i}tA)^\dagger}=e^{-\mathrm{i}tA^\dagger}=e^{-\mathrm{i}tA} \tag{6.3.963}$$

Therefore:

$$\left(e^{\mathrm{i}tA}\right)^\dagger\cdot e^{\mathrm{i}tA}=e^{-\mathrm{i}tA}e^{\mathrm{i}tA}=e^{\mathrm{i}t(A-A)}=I \tag{6.3.964}$$

■ Q.E.D.

Remember that this condition for a self-adjoint matrix is linked to the definition of linear group of order n (see section Set Algebra page 1424).

One of the first applications of the exponentation of matrices is the resolution of ordinary differential equations. Indeed, from the linear differential equation below using as initial condition $y(0)=0$ and where A is a matrix:

$$y'(t)+Ay(t)=0 \tag{6.3.965}$$

the solution is given by (as seen previously):

$$y(t)=Ce^{-At} \tag{6.3.966}$$

We frequently find that kind of systems of differential equations in biology (population dynamics), astrophysics (study of plasmas) or in fluid mechanics (chaos theory) and in classical mechanics (coupled systems), astronomy (coupled orbits), in electrical engineering, etc.

✍**Example:**

Suppose we have the following homogeneous system of differential equations (without constant terms):

$$\begin{aligned} x' &= 3x - 4y \\ y' &= -6x + y \end{aligned} \tag{6.3.967}$$

The associated matrix is then:

$$A = \begin{pmatrix} 3 & 4 \\ -6 & 1 \end{pmatrix} \tag{6.3.968}$$

and its exponential (see developments made above):

$$e^{-At} = \frac{1}{5}\begin{pmatrix} 2e^{-3t} + 3e^{7t} & 2e^{-3t} - 2e^{7t} \\ 2e^{-3t} - 3e^{7t} & 2e^{-3t} + 2e^{7t} \end{pmatrix} \tag{6.3.969}$$

The general solution of the system is:

$$\begin{pmatrix} x \\ y \end{pmatrix} = C_1 \begin{pmatrix} 2e^{-3t} + 3e^{7t} \\ 2e^{-3t} - 3e^{7t} \end{pmatrix} + C_2 \begin{pmatrix} 2e^{-3t} - 2e^{7t} \\ 2e^{-3t} + 2e^{7t} \end{pmatrix} \tag{6.3.970}$$

So we have:

$$\begin{aligned} x &= C_1(2e^{-3t} + 3e^{7t}) + C_2(2e^{-3t} - 2e^{7t}) = A'e^{-3t} + B'e^{7t} \\ y &= C_1(2e^{-3t} - 3e^{7t}) + C_2(2e^{-3t} + 2e^{7t}) = D'e^{-3t} + E'e^{7t} \end{aligned} \tag{6.3.971}$$

By calculating the derivative of the previous relations and comparing to:

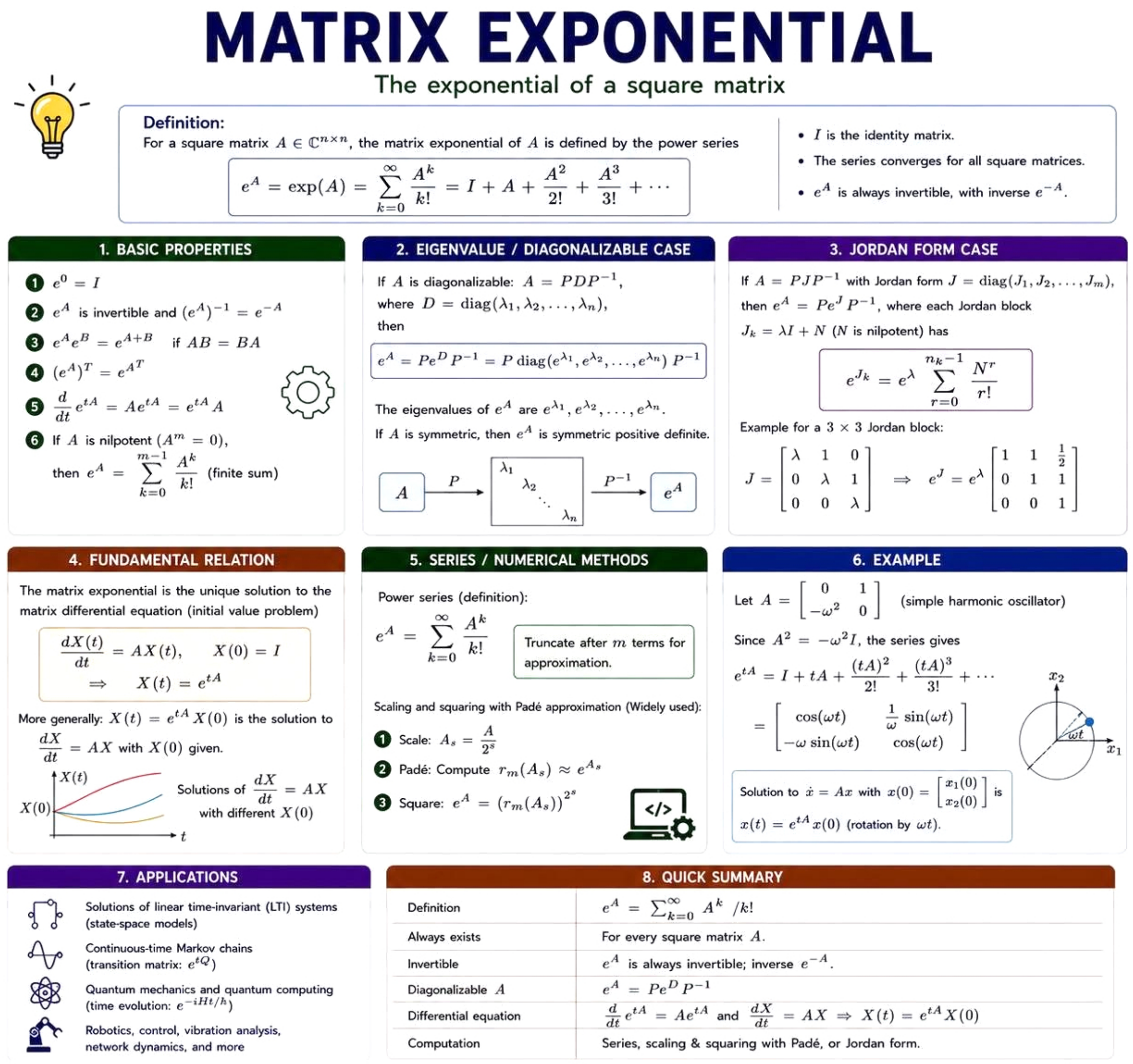
$$\begin{aligned} x' &= 3x - 4y \\ y' &= -6x + y \end{aligned} \tag{6.3.972}$$

we easily determine the constants to get:

$$\begin{aligned} E' - 6B' = 7E' \Rightarrow -6E' = 6B' \Rightarrow E' &= -B' \\ -6A' + F' = -3F' \Rightarrow 4F' = 6A' \Rightarrow F' &= \frac{3}{2}A' \end{aligned} \tag{6.3.973}$$

which finally gives us:

$$\begin{aligned} x &= A'e^{-3t} + B'e^{7t} \\ y &= \frac{3}{2}A'e^{-3t} - B'e^{7t} \end{aligned} \tag{6.3.974}$$



MATRIX EXPONENTIAL

The exponential of a square matrix

Definition:

For a square matrix $A \in \mathbb{C}^{n\times n}$, the matrix exponential of A is defined by the power series

$$\boxed{e^A = \exp(A) = \sum_{k=0}^{\infty} \frac{A^k}{k!} = I + A + \frac{A^2}{2!} + \frac{A^3}{3!} + \cdots}$$

- I is the identity matrix.
- The series converges for all square matrices.
- e^A is always invertible, with inverse e^{-A}.

1. BASIC PROPERTIES

❶ $e^0 = I$

❷ e^A is invertible and $(e^A)^{-1} = e^{-A}$

❸ $e^A e^B = e^{A+B}$ if $AB = BA$

❹ $(e^A)^T = e^{A^T}$

❺ $\frac{d}{dt} e^{tA} = Ae^{tA} = e^{tA}A$

❻ If A is nilpotent ($A^m = 0$),
then $e^A = \sum_{k=0}^{m-1} \frac{A^k}{k!}$ (finite sum)

2. EIGENVALUE / DIAGONALIZABLE CASE

If A is diagonalizable: $A = PDP^{-1}$,
where $D = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$,
then

$$\boxed{e^A = Pe^D P^{-1} = P \operatorname{diag}(e^{\lambda_1}, e^{\lambda_2}, \ldots, e^{\lambda_n})\, P^{-1}}$$

The eigenvalues of e^A are $e^{\lambda_1}, e^{\lambda_2}, \ldots, e^{\lambda_n}$.

If A is symmetric, then e^A is symmetric positive definite.

$$\boxed{A} \xrightarrow{P} \begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \ddots \ \lambda_n \end{bmatrix} \xrightarrow{P^{-1}} \boxed{e^A}$$

3. JORDAN FORM CASE

If $A = PJP^{-1}$ with Jordan form $J = \operatorname{diag}(J_1, J_2, \ldots, J_m)$, then $e^A = Pe^J P^{-1}$, where each Jordan block $J_k = \lambda I + N$ (N is nilpotent) has

$$\boxed{e^{J_k} = e^{\lambda} \sum_{r=0}^{n_k-1} \frac{N^r}{r!}}$$

Example for a 3×3 Jordan block:

$$J = \begin{bmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{bmatrix} \Rightarrow e^J = e^{\lambda} \begin{bmatrix} 1 & 1 & \frac{1}{2} \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

4. FUNDAMENTAL RELATION

The matrix exponential is the unique solution to the matrix differential equation (initial value problem)

$$\boxed{\begin{aligned} &\frac{dX(t)}{dt} = AX(t), \qquad X(0) = I \\ &\Rightarrow \qquad X(t) = e^{tA} \end{aligned}}$$

More generally: $X(t) = e^{tA}X(0)$ is the solution to $\frac{dX}{dt} = AX$ with $X(0)$ given.

Solutions of $\frac{dX}{dt} = AX$ with different $X(0)$

5. SERIES / NUMERICAL METHODS

Power series (definition):

$$e^A = \sum_{k=0}^{\infty} \frac{A^k}{k!}$$

Truncate after m terms for approximation.

Scaling and squaring with Padé approximation (Widely used):

❶ Scale: $A_s = \frac{A}{2^s}$

❷ Padé: Compute $r_m(A_s) \approx e^{A_s}$

❸ Square: $e^A = (r_m(A_s))^{2^s}$

6. EXAMPLE

Let $A = \begin{bmatrix} 0 & 1 \\ -\omega^2 & 0 \end{bmatrix}$ (simple harmonic oscillator)

Since $A^2 = -\omega^2 I$, the series gives

$$e^{tA} = I + tA + \frac{(tA)^2}{2!} + \frac{(tA)^3}{3!} + \cdots$$

$$= \begin{bmatrix} \cos(\omega t) & \frac{1}{\omega}\sin(\omega t) \\ -\omega \sin(\omega t) & \cos(\omega t) \end{bmatrix}$$

Solution to $\dot{x} = Ax$ with $x(0) = \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix}$ is $x(t) = e^{tA}x(0)$ (rotation by ωt).

7. APPLICATIONS

- Solutions of linear time-invariant (LTI) systems (state-space models)
- Continuous-time Markov chains (transition matrix: e^{tQ})
- Quantum mechanics and quantum computing (time evolution: $e^{-iHt/\hbar}$)
- Robotics, control, vibration analysis, network dynamics, and more

8. QUICK SUMMARY

Definition	$e^A = \sum_{k=0}^{\infty} A^k / k!$
Always exists	For every square matrix A.
Invertible	e^A is always invertible; inverse e^{-A}.
Diagonalizable A	$e^A = Pe^D P^{-1}$
Differential equation	$\frac{d}{dt} e^{tA} = Ae^{tA}$ and $\frac{dX}{dt} = AX \Rightarrow X(t) = e^{tA}X(0)$
Computation	Series, scaling & squaring with Padé, or Jordan form.

6.3.3.5.2 Phase space

If we are given a general nth order differential equation of the form:

$$a_n(x)y^{(n)}+a_{n-1}y^{(n-1)}+\cdots+a_1(x)y'+a_0(x)y=f(x) \tag{6.3.975}$$

and we have a solution y, then we can plot a graph of y versus x, and see how y evolves with x.

However, one problem is that for such an equation, the solution is not just determined by the initial condition $y(x_0)$, but also $y'(x_0)$, $y''(x_0)$ etc. So if we just have a snapshot of the value of y at a particular point x_0, we have completely no idea how it would evolve in the future.

So how much information do we actually need? At any point x_0, if we are given the first $n-1$ derivatives, i.e. $y(x_0)$, $y'(x_0)$, $\ldots$, $y^{(n-1)}(x_0)$, we can then get the nth derivative and also any higher derivatives from the differential equation.

Thus we are led to consider the "solution vector":

$$\vec{Y}(x)=\left(y(x),y'(x),\cdots,y^{n-1}(x)\right) \tag{6.3.976}$$

We say such a vector lies in the "phase space", which is an n-dimensional space (obviously this tool is quite limited for humans as we have huge difficulties to deal with spaces having more than three dimensions...). So for each x, we thus get a point $\vec{Y}(x)$ lying in the n-dimensional space. Moreover, given any point in the phase space, if we view it as the initial conditions for our differential equation, we get a unique trajectory in the phase space.

Remark

Honestly as authors of this textbook we have never seen the subject treated and the next one on "stability theory" used in real application cases in engineering nowadays. Maybe it was useful before computers? However the topic leads to quite beautiful concepts so why not...

✍**Example:**

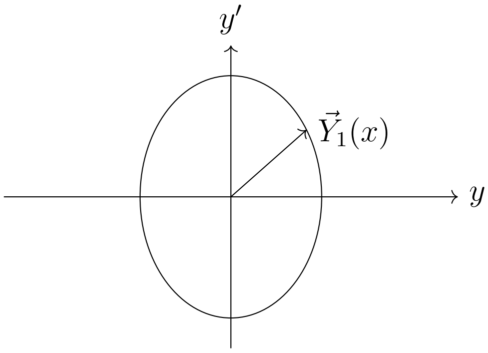
Consider:

$$y''+4y=0$$

Suppose we have an initial condition of:

$$y_1(0)=1,y_1'(0)=0$$

Then we can solve the equation to obtain $y_1(x)=\cos(2x)$. Thus the initial solution vector is $\vec{Y}_1(0)=(1,0)$, and the trajectory as x varies is given by $\vec{Y}_1(x)=(\cos(2x),-2\sin(2x))$. Thus as x changes, we trace out an ellipse in the clockwise direction:



Another possible initial condition is $y_2(0)=0, y_2'(0)=2$. In this case, we obtain the solution $y(x)=\sin(2x)$, with a solution vector $\vec{Y}_2(x)=(\sin(2x), 2\cos(2x))$.

Note that as vectors, our two initial conditions $(1,0)$ and $(0,2)$ are independent (we will see further below what is exactly means). Moreover, as x changes, the two solution vectors $\vec{Y}_1(x), \vec{Y}_2(x)$ remain independent.

In general, for a 2nd order equation, the phase space is a 2-dimensional space, and we can take the two complementary functions $\vec{Y}_1$ and $\vec{Y}_2$ as basis vectors for the phase space at each particular value of x. Of course, we need the two solutions to be linearly independent.

Definition (#209): Given a differential equation with solutions y_1, y_2, the "Wronskian" is the determinant (this definition will be motivated soon below!):

$$W(x)=\begin{vmatrix} y_1 & y_2 \\ y_1' & y_2' \end{vmatrix}=y_1y_2'-y_2y_1' \tag{6.3.977}$$

Intuitively, as we will see further below the Wronskian checks that y_1 and y_2 are linearly independent (i.e. that y_1 is not a multiple of y_2 other than trivially when $x_1=x_2=0$), that the vector (y_1,y_2) is linearly independent from $(\partial_x y_1, \partial_x y_2)$ and equivalently that $\partial_x y_1$ is not a multiple of $\partial_x y_2$.

The Wronskian requirement is in fact not any stricter than linear independence. We can show that linear dependence implies vanishing of the Wronskian. If y_1 and y_2 are two linearly dependent solutions, then there exist nonzero scalars a_1, a_2 such that:

$$a_1y_1+a_2y_2=0 \tag{6.3.978}$$

Differentiating, we get:

$$a_1\partial_x y_1+a_2\partial_x y_2=0 \tag{6.3.979}$$

so that:

$$a_2\partial_x y_2=-a_1\partial_x y_1 \tag{6.3.980}$$

Multiplying the first equation by $a_2\partial_x y_2$ gives:

$$\begin{aligned} 0&=a_1a_2y_1\partial_x y_2+a_2^2y_2\partial_x y_2\\ &=a_1a_2\left(y_1y_xy_2-y_2\partial_x y_1\right)\\ &\rightarrow y_1\partial_x y_2-y_2\partial_x y_1=0 \end{aligned} \tag{6.3.981}$$

In other words: To show that the two functions φ_1 and φ_2 are linearly independent you make:

$$Ay_1(x)+By_2(x)=0 \tag{6.3.982}$$

and you want to prove that A an B are null.

But you need another condition to force A and B to be zero. And you get it by deriving:

$$Ay_1'(x)+By_2'(x)=0 \tag{6.3.983}$$

This is a linear homogeneous system:

$$\begin{cases} Ay_1(x)+By_2(x)=0 \\ Ay_1'(x)+By_2'(x)=0 \end{cases} \tag{6.3.984}$$

which have only the solution $A=B=0$ if the determinant is non-null. But this determinant is precisely the Wronskian!!

Definition (#210): Two solutions $y_1(x)$ and $y_2(x)$ are "independent solutions" solutions of the differential equation if and only if $\vec{Y}_1$ and $\vec{Y}_2$ are linearly independent as vectors in the phase space for some x, i.e. if and only if the Wronskian is non-zero for some x.

In both cases, the Wronskian is never zero. Is it possible that it is zero for some x while non-zero for others? The answer is: No! And the proof given below is named the "Abel's Theorem".

Theorem 6.17. Given a differential equation $y''+p(x)y'+q(x)y=0$ (the theorem can be generalized to nth order linear ordinary differential equations), either $W=0$ for all x, or $W\neq 0$ for all x. i.e. if two solutions are independent for some particular x, then they are independent for all x.

Proof 6.17.1. If y_1 and y_2 are both solutions, then:

$$\begin{gathered} y_2(y_1''+py_1'+qy_1)=0 \\ y_1(y_2''+py_2'+qy_2)=0 \end{gathered} \tag{6.3.985}$$

Subtracting the two equations, we have:

$$y_1y_2''-y_2y_1''+p(y_1y_2'-y_2y_1')=0 \tag{6.3.986}$$

Note that (this is what motivates the definition of the Wronskian!):

$$W=y_1y_2'-y_2y_1' \tag{6.3.987}$$

and:

$$W'=y_1y_2''+y_1'y_2'-(y_2'y_1'+y_2y_1'')=y_1y_2''-y_2y_1'' \tag{6.3.988}$$

There we these definition we can write:

$$W'+pW=0 \tag{6.3.989}$$

And the solution is immediate:

$$W(x)=W_0e^{-\int p\mathrm{d}x} \tag{6.3.990}$$

Where $W_0=c^{te}$. Since the exponential function is never zero, either $W_0=0$, in which case $W=0$, or $W_0\neq 0$ and $W\neq 0$ for any value of x!

In other words: The Wronskian is always either identically zero, always positive, or always negative.

■ Q.E.D.

6.3.3.5.2.1 Stability theory

In studying systems of differential equations, it is often useful to study the of solutions without obtaining an algebraic form for the solution. This is done by exploring equilibrium solutions and solutions nearby equilibrium solutions. Such techniques will be seen to be useful in studying nonlinear systems.

"Stability theory" addresses the stability of solutions of differential equations and of trajectories of dynamical systems under small perturbations of initial conditions. The heat equation, for example, is a stable partial differential equation because small perturbations of initial data lead to small variations in temperature at a later time.

Many parts of the qualitative theory of differential equations and dynamical systems deal with asymptotic properties of solutions and the trajectories - what happens with the system after a long period of time. The simplest kind of behaviour is exhibited by equilibrium points, or fixed points, and by periodic orbits. If a particular orbit is well understood, it is natural to ask next whether a small change in the initial condition will lead to similar behaviour. Stability theory addresses the following questions: Will a nearby orbit indefinitely stay close to a given orbit? Will it converge to the given orbit? In the former case, the orbit is said to be "stable"; in the latter case, it is said to be "asymptotically stable" and the given orbit is said to be "attracting".

Ok now let us dive into the maths! Consider two dependent variables $y_1(t), y_2(t)$ related by the following "nonautonomous system[9]" of a couple system of first order differential equations in the plane (as there are two differentials equations it's also named a "planar system"):

$$\begin{aligned}\dot{y}_1 &= P(y_1,y_2,t) = ay_1+by_2+f_1(t)\\ \dot{y}_2 &= Q(y_1,y_2,t) = cy_1+dy_2+f_2(t)\end{aligned} \tag{6.3.991}$$

An autonomous system being by definition explicitly independent of time:

$$\begin{aligned}\dot{y}_1 &= ay_1+by_2\\ \dot{y}_2 &= cy_1+dy_2\end{aligned} \tag{6.3.992}$$

We can write this in vector notation by:

$$\begin{pmatrix}\dot{y}_1\\ \dot{y}_2\end{pmatrix} = \begin{bmatrix}a & b\\ c & d\end{bmatrix}\begin{pmatrix}y_1\\ y_2\end{pmatrix}+\begin{pmatrix}f_1\\ f_2\end{pmatrix} \tag{6.3.993}$$

or:

$$\dot{\vec{Y}} = M\vec{Y}+\vec{F} \tag{6.3.994}$$

We can convert this to a higher-order equation by:

$$\begin{aligned}\ddot{y}_1 &= a\dot{y}_1+b\dot{y}_2+\dot{f}_1\\ &= a\dot{y}_1+b(cy_1+dy_2+f_2)+\dot{f}_1\\ &= a\dot{y}_1+bcy_1+d(\dot{y}_1-ay_1-f_1)+bf_2+\dot{f}_1\end{aligned} \tag{6.3.995}$$

---

[9] An "autonomous system" or "autonomous differential equation" is a system of ordinary differential equations which does not explicitly depend on the independent variable. When the variable is time, they are also called "time-invariant systems". Many laws in physics, where the independent variable is usually assumed to be time, are expressed as autonomous systems because it is assumed the laws of nature which hold now are identical to those for any point in the past or future.

so:

$$\ddot{y}_1-(a+d)\dot{y}_1+(ad-bc)y_1=bf_2-df_1+\dot{f}_1 \tag{6.3.996}$$

However, this actually complicates the equation...

So what we usually do is the other way round: if we have a high-order equation, we can do this to change it to a system of first-order equations!

If:

$$\ddot{y}+a\dot{y}+by=f \tag{6.3.997}$$

write $y_1=y$ and $y_2=\dot{y}$. Then let:

$$\vec{Y}=\begin{pmatrix}y\\ \dot{y}\end{pmatrix} \tag{6.3.998}$$

Our system of equations becomes:

$$\begin{aligned}\dot{y}_1&=y_2\\ \dot{y}_2&=f-ay_2-by_1\end{aligned} \tag{6.3.999}$$

or:

$$\dot{\vec{Y}}=\begin{bmatrix}0&1\\-b&-a\end{bmatrix}\begin{pmatrix}y_1\\y_2\end{pmatrix}\begin{pmatrix}0\\f\end{pmatrix} \tag{6.3.1000}$$

Now consider the general equation:

$$\dot{\vec{Y}}=M\vec{Y}+\vec{F} \tag{6.3.1001}$$

Rearranging:

$$\dot{\vec{Y}}-M\vec{Y}=\vec{F} \tag{6.3.1002}$$

We first look for a complementary solution $\vec{Y}_c=\vec{v}e^{\lambda t}$, where $\vec{v}$ is a constant vector. So we get:

$$\lambda\vec{v}-M\vec{v}=\vec{0} \tag{6.3.1003}$$

We can write this as:

$$M\vec{v}=\lambda\vec{v} \tag{6.3.1004}$$

So λ is the eigenvalue of M and $\vec{v}$ is the corresponding eigenvector!!!

We can solve this by solving the characteristic equation:

$$\det(M-\lambda I)=0 \tag{6.3.1005}$$

Then for each λ, we find the corresponding $\vec{v}$. The corresponding eigenvalues equation is therefore:

$$\lambda^2-(a+d)\lambda+(ad-bc)=0 \tag{6.3.1006}$$

The eigenvalues are not determined by the four numbers (a,b,c,d), but only two numbers $(a+d, ad-bc)$.

For a matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$, the trace is $\text{tr}(M) = a+d$, and the determinant is $\det(M) = ad-bc$.

If λ_1 and λ_2 are roots of:

$$\lambda^2 - \text{tr}(M)\lambda + \det(M) = (\lambda-\lambda_1)(\lambda-\lambda_2) = 0 \tag{6.3.1007}$$

then:

$$\lambda_1+\lambda_2 = \text{tr}(M), \lambda_1\lambda_2 = \det(M), \lambda_{1,2} = \frac{\text{tr}(M) \pm \sqrt{\text{tr}(M)^2 - 4\det(M)}}{2} \tag{6.3.1008}$$

and as a reminder, we have:

- $\text{tr}(M)^2 - 4\det(M) > 0$, two real roots (related to "overdamped systems" since the system leads to exponential decay). Then the solution is commonly written as ($\vec{v}_1$ and $\vec{v}_2$ are the eigenvectors):

$$\vec{Y} = A\vec{v}_1 e^{\lambda_1 t} + B\vec{v}_2 e^{\lambda_2 t} \tag{6.3.1009}$$

- $\text{tr}(M)^2 - 4\det(M) = 0$, one (repeated) root (related to "critically damped systems" - there are no oscillations). Then the solution is commonly written as:

$$\vec{Y} = A\vec{v}_1 e^{\lambda t} + B\vec{v}_2 e^{\lambda t}(\vec{v}_2 + t\vec{v}_1) \tag{6.3.1010}$$

- $\text{tr}(M)^2 - 4\det(M) < 0$, two complex conjugated roots (related to "underdamped oscillating systems" consisting of decaying exponential time oscillations). Then the solution is commonly written as:

$$\vec{Y} = A\Re\left(e^{\alpha t}\vec{v}(\cos(\beta t) + \mathrm{i}\sin(\beta t)\right) + B\Im\left(e^{\alpha t}\vec{v}(\cos(\beta t) + \mathrm{i}\sin(\beta t)\right) \tag{6.3.1011}$$

Let us see an example:

✍**Example:**

Consider the following system:

$$\dot{\vec{Y}} = \begin{bmatrix} -4 & 24 \\ 1 & -2 \end{bmatrix} \vec{Y} + \begin{pmatrix} 4 \\ 1 \end{pmatrix} e^t$$

The characteristic equation of M is:

$$\begin{vmatrix} -4-\lambda & 24 \\ 1 & -2-\lambda \end{vmatrix} = 0$$

which gives:

$$(\lambda+8)(\lambda-2) = 0$$

and therefore:

$$\lambda_{1,2} = 2, -8$$

When $\lambda = 2$, $\vec{v}$ satisfies:

$$\begin{bmatrix} -6 & 24 \\ 1 & -4 \end{bmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \vec{0}$$

and we get $\vec{v}_1 = \begin{pmatrix} 4 \\ 1 \end{pmatrix}$.

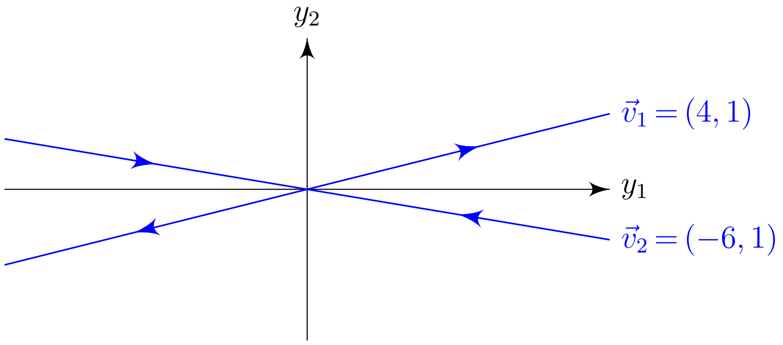
When $\lambda = -8$, we have:

$$\begin{pmatrix} 4 & 24 \\ 1 & 6 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \vec{0}$$

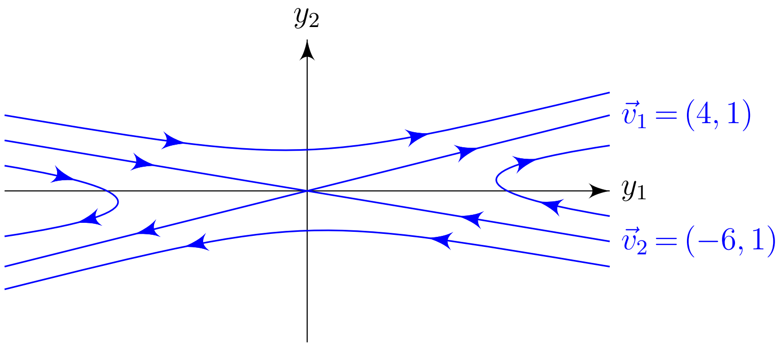
and there $\vec{v}_2 = \begin{pmatrix} -6 \\ 1 \end{pmatrix}$. So the complementary solution is:

$$\vec{Y} = A \begin{pmatrix} 4 \\ 1 \end{pmatrix} e^{2t} + B \begin{pmatrix} -6 \\ 1 \end{pmatrix} e^{-8t}$$

To plot the phase-space trajectories, we first consider the cases where $\vec{Y}$ is an eigenvector. If $\vec{Y}$ is an integer multiple of $\begin{pmatrix} 4 \\ 1 \end{pmatrix}$, then it will keep moving outwards in the same direction. Similarly, if it is an integer multiple of $\begin{pmatrix} -6 \\ 1 \end{pmatrix}$, it will move towards the origin.



We can now add more (curved) lines based on these two:



To find the particular solution, we try $\vec{Y}_p = \vec{u}e^t$. Then:

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} - \begin{bmatrix} -4 & 24 \\ 1 & -2 \end{bmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} 4 \\ 1 \end{pmatrix}$$

$$\begin{bmatrix} 5 & -24 \\ -1 & 3 \end{bmatrix}\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}=\begin{pmatrix} 4 \\ 1 \end{pmatrix}$$
$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}=-\frac{1}{9}\begin{bmatrix} 3 & 24 \\ 1 & 5 \end{bmatrix}\begin{pmatrix} 4 \\ 1 \end{pmatrix}=\begin{pmatrix} -4 \\ -1 \end{pmatrix}$$

So the general solution is:

$$\vec{Y}=A\begin{pmatrix} 4 \\ 1 \end{pmatrix}e^{2t}+B\begin{pmatrix} -6 \\ 1 \end{pmatrix}e^{-8t}-\begin{pmatrix} 4 \\ 1 \end{pmatrix}e^{t}=\begin{bmatrix} 4e^{2t} & -6e^{-8t} \\ 1e^{2t} & 1e^{-8t} \end{bmatrix}\begin{pmatrix} A \\ B \end{pmatrix}-\begin{pmatrix} 4 \\ 1 \end{pmatrix}e^{t}$$

Often denoted in the general case:

$$\vec{Y}=\Phi(t)\vec{C}+\vec{x}_p(t)$$

Remark

Notice that if the system above was autonomous, i.e. independent of time such that:

$$\frac{\mathrm{d}y_1}{\mathrm{d}t}=-4y_1+42y_2$$
$$\frac{\mathrm{d}y_2}{\mathrm{d}t}=1y_1-2y_2$$

then we can get the slopes of the phase diagram by just computing:

$$\frac{\mathrm{d}y_2}{\mathrm{d}y_1}=\frac{\frac{\mathrm{d}y_2}{\mathrm{d}t}}{\frac{\mathrm{d}y_1}{\mathrm{d}t}}=\frac{1y_1-2y_2}{-4y_1+42y_2}$$

Consider the general solution of a autonomous system known as being like (as seen in the example above):

$$\vec{Y}=\Phi(t)\vec{C} \qquad (6.3.1012)$$

We are now ready to solve the initial value problem:

$$\dot{\vec{Y}}=M\vec{Y}, \quad \vec{Y}(t_0)=\vec{Y}_0 \qquad (6.3.1013)$$

Starting with the general solution, we have that:

$$\vec{Y}_0=\vec{Y}(t_0)=\Phi(t_0)\vec{C} \qquad (6.3.1014)$$

As usual, we need to solve for the c_k (i.e. the components of $\vec{C}$). Using matrix methods, this is very easy. Since the Wronskian is not zero, then we can invert Φ at any value of t. So, we have:

$$\vec{C}=\Phi^{-1}(t_0)\vec{Y}_0 \qquad (6.3.1015)$$

Putting $\vec{C}$ back into the general solution, we get the solution to the initial value problem:

$$\vec{Y}(t) = \Phi(t)\Phi^{-1}(t_0)\,\vec{Y}_0 \qquad (6.3.1016)$$

As always we can (and should always!) easily verify that this is a solution of the system and satisfies the initial condition at $t = t_0$.

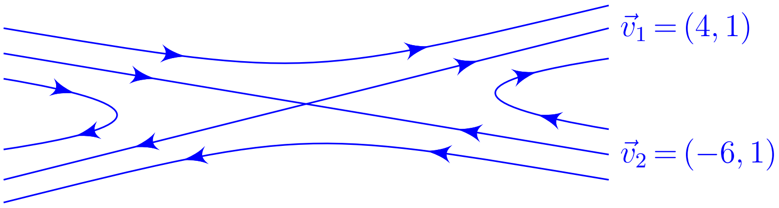
The matrix combination $\Phi(t)\Phi^{-1}(t_0)$ is useful. So, we will define the resulting product to be the "principal matrix solution", denoting it by:

$$\Psi(t) = \Phi(t)\Phi^{-1}(t_0) \qquad (6.3.1017)$$

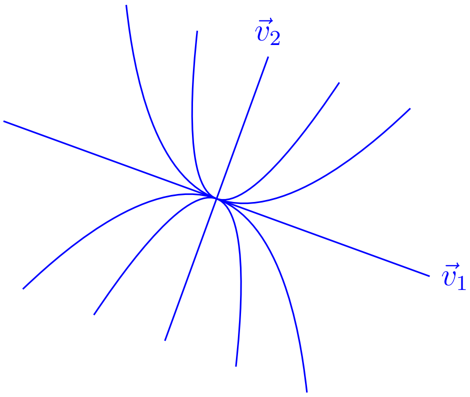
Thus, the solution of the initial value problem is then denoted $\vec{Y}(t) = \Psi(t)\vec{Y}_0$.

In general, there are basically three possible cases of $\dot{\vec{Y}} = M\vec{Y}$ corresponding to three different possible eigenvalues of M:

1. If both λ_1, λ_2 are real with opposite sign ($\lambda_1\lambda_2 < 0$). Without loss of generality assume $\lambda_1 > 0$. Then there is a saddle as above:

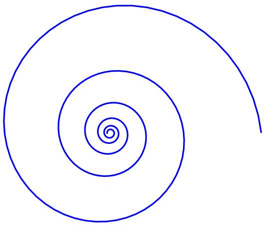


2. If λ_1, λ_2 are real with $\lambda_1\lambda_2 > 0$. Without loss of generality assume $|\lambda_1| \geq |\lambda_2|$. Then the phase portrait is:



If both $\lambda_1, \lambda_2 < 0$, then the arrows point towards the intersection and we say there is a stable node. If both are positive, they point outwards and there is an unstable node.

3. If λ_1, λ_2 are complex conjugates, then we get a spiral:



If $\Re(\lambda_{1,2})<0$, then it spirals inwards. If $\Re(\lambda_{1,2})>0$, then it spirals outwards. If $\Re(\lambda_{1,2})=0$, then we have ellipses with common centers instead of spirals. We can determine whether the spiral is positive (as shown above), or negative (mirror image of the spiral above) by considering the eigenvectors.

Coming back to:

$$\lambda_1+\lambda_2=\text{tr}(M),\lambda_1\lambda_2=\det(M),\lambda_{1,2}=\frac{\text{tr}(M)\pm\sqrt{\text{tr}(M)^2-4\det(M)}}{2}=\frac{\text{tr}(M)\pm\sqrt{\Delta}}{2} \tag{6.3.1018}$$

We then have as reminder:

- $\text{tr}(M)^2-4\det(M)>0$, two real roots (related to "overdamped systems" since the system leads to exponential decay). Then the solution is commonly written as ($\vec{v}_1$ and $\vec{v}_2$ are the eigenvectors):

$$\vec{Y}=A\vec{v}_1e^{\lambda_1 t}+B\vec{v}_2e^{\lambda_2 t} \tag{6.3.1019}$$

- $\text{tr}(M)^2-4\det(M)=0$, one (repeated) root (related to "critically damped systems" - there are no oscillations). Then the solution is commonly written as:

$$\vec{Y}=A\vec{v}_1e^{\lambda t}+B\vec{v}_2e^{\lambda t}(\vec{v}_2+t\vec{v}_1) \tag{6.3.1020}$$

- $\text{tr}(M)^2-4\det(M)<0$, two complex conjugated roots (related to "underdamped oscillating systems" consisting of decaying exponential times oscillations). Then the solution is commonly written as:

$$\vec{Y}=A\Re\left(e^{\alpha t}\vec{v}(\cos(\beta t)+\mathrm{i}\sin(\beta t)\right)+B\Im\left(e^{\alpha t}\vec{v}(\cos(\beta t)+\mathrm{i}\sin(\beta t)\right) \tag{6.3.1021}$$

But more in details:

- If $\text{tr}(M)>0,\det(M)>0$, and $\text{tr}(M)^2-4\det(M)>0$, it is a source (unstable system!);
- If $\text{tr}(M)<0,\det(M)>0$, and $\text{tr}(M)^2-4\det(M)>0$, it is a sink (stable system!);
- If $\text{tr}(M)>0,\det(M)>0$, and $\text{tr}(M)^2-4\det(M)<0$, it is a spiral source (unstable system!);
- If $\text{tr}(M)<0,\det(M)>0$, and $\text{tr}(M)^2-4\det(M)<0$, it is a spiral sink (stable system!);
- If $\det(M)<0$, then it is a saddle (unstable system!). In this case, $\text{tr}(M)^2-4\det(M)>0$ always holds and it does not matter whether $\text{tr}(M)>0$ or $\text{tr}(M)<0$;
- If $\text{tr}(M)=0$ or $\det(M)=0$, it is degenerate (neutral system!)

If represented as a table:

Type	Eigenvalues	Stability
Node	Real λ, same signs	$\lambda < 0$, stable $\lambda > 0$, unstable
Saddle	Real λ opposite signs	Mostly Unstable
Center	λ pure imaginary	-
Focus/Spiral	Complex $\lambda, \Re(\lambda) \neq 0$	$\Re(\lambda) < 0$, stable $\Re(\lambda) > 0$, unstable
Degenerate Node	Repeated roots	$\lambda > 0$, stable
Lines of Equilibria	One zero eigenvalue	$\lambda < 0$, stable

Table 6.14: **List of typical behaviours in planar systems**

Here is a visual summary based on the matrix and the determinant (based on a manual drawing by Douglas R. Hundley):

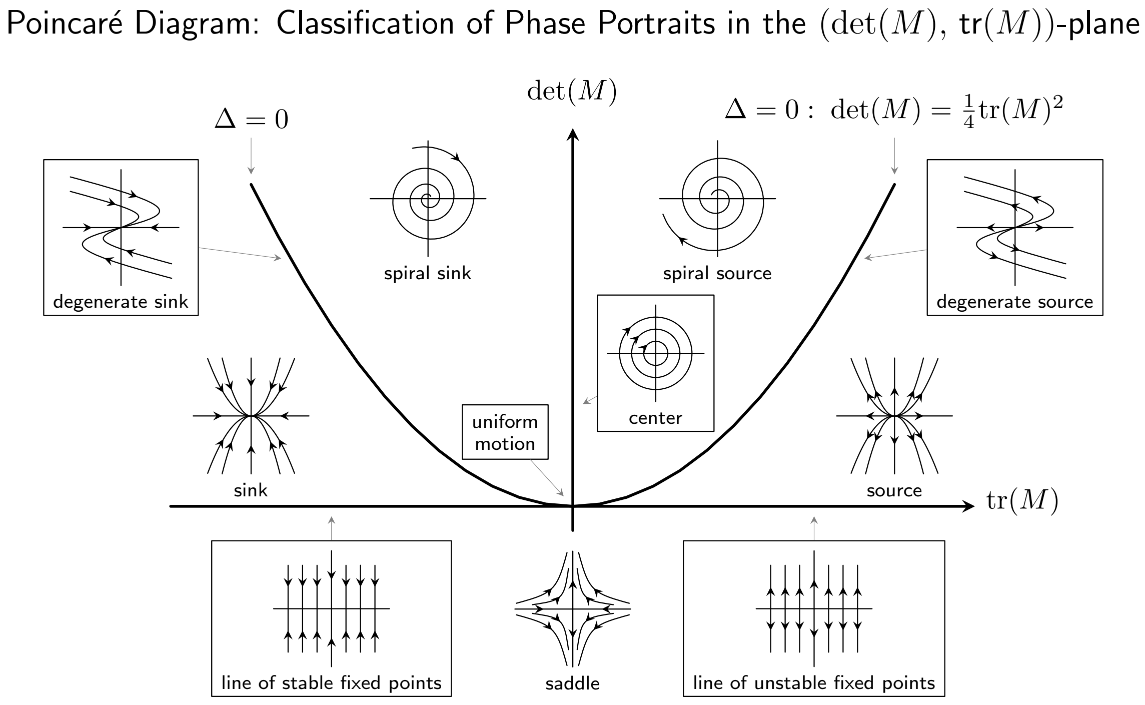

Poincaré Diagram: Classification of Phase Portraits in the $(\det(M), \text{tr}(M))$-plane
$\Delta = 0$
$\det(M)$
$\Delta = 0: \det(M) = \frac{1}{4}\text{tr}(M)^2$
degenerate sink
spiral sink
spiral source
degenerate source
center
uniform motion
sink
source
$\text{tr}(M)$
line of stable fixed points
saddle
line of unstable fixed points

Figure 6.64: **Poincaré Diagram - Classification of Phase Portraits for a 2×2 matrix (author: Gernot Salzer)**

In the trace-determinant plane above, the points with coordinates (T, D) (i.e. Trace/Determinant) correspond to systems with repeated eigenvalues. Furthermore, points above the parabola correspond to systems with two complex eigenvalues which are complex conjugates with each other. Lastly, points below the parabola correspond to planar systems with two distinct eigenvalues.

Caution! The above Poincaré diagram is really only for a 2×2 matrix A. For higher dimensions it can't be used anymore!

Remark

If the matrix M is symmetric we get another typical type of pattern ("star" pattern) that is not represented in the figure above.

6.3.3.6 Nonlinear dynamical systems

In mathematics and science, a nonlinear system is a system in which the change of the output is not proportional to the change of the input. Nonlinear problems are of interest to engineers, biologists, physicists, mathematicians, and many other scientists since most systems are inherently nonlinear in nature. Nonlinear dynamical systems, describing changes in variables over time, may appear chaotic, unpredictable, or counterintuitive, contrasting with much simpler linear systems. They are hard to solve or unsolvable in some situations. In most cases they cannot be solved by hand but by numerical simulations.

quote
6.3:1644.#55

> ❞ *Mathematics takes place at different timescales. If you can solve a problem in 5 minutes that others need an hour to solve, you can probably get a good job. If you can solve a problem in a month that others might need a year to solve, you will probably do well as a graduate student. But if you can solve a problem in 10 years that nobody else can solve in a lifetime, you could be a great mathematician.* ❞
>
> Robert Israel

A system of differential equations is said to be nonlinear if it is not a system of linear equations. Problems involving nonlinear differential equations are extremely diverse, and methods of solution or analysis are problem dependent. Examples of nonlinear differential equations treated in quite in details in this textbook are:

- The double pendulum (see page 3440)
- The Navier–Stokes equations in fluid dynamics (see page 3811)
- The Lotka–Volterra equations in biology (see page 7350)
- The Einstein's field equations in General Relativity (see page 5766)
- The Lorenz equation in meteorology (see page 8318)
- etc.

So the reader interested can pick up one of the subject above to get a quite detailed mathematical treatment of a special case of nonlinear dynamical systems.

6.3.3.7 Regular Methods of Perturbations

Very frequently in physics (high level physics) or financial engineering, a mathematical problem can not be solved exactly. Even if the solution is known sometimes there are such a dependency of parameters that the solution is difficult to use as such.

Sometimes, however, it happens that an identified parameter of the differential equation, which we denote by tradition with the Greek letter ε, is such that the solution is available and reasonably simple for $\varepsilon = 0$.

The problem then is to know how the solution is altered for a non-zero ε but still small. This study is the center of "perturbation theory" that we will use, for example in the section of General Relativity to calculate the precession of the perihelion of Mercury.

As the perturbation theory within the general framework is too complex for this textbook purpose, we propose an approach by example with first a simple algebraic equation and then with what interests us: a differential equation.

6.3.3.7.1 Perturbation theory for algebraic equations

Consider the following polynomial equation:

$$x^2 - 1 = \varepsilon x \qquad (6.3.1022)$$

We know from our study of the section Functional Analysis, that this polynomial equation has two roots that are trivially:

$$x_{1,2} = \frac{\varepsilon \pm \sqrt{\varepsilon^2 + 4}}{2} = \frac{\varepsilon}{2} \pm \sqrt{1 + \frac{\varepsilon^2}{4}} \qquad (6.3.1023)$$

For small ε, these roots can be approximated by the first term of the Taylor series expansion (see section Sequences and Series page 1696):

$$x_1 = 1 + \frac{\varepsilon}{2} + \frac{\varepsilon^2}{8} + \mathcal{O}(\varepsilon^3) \quad x_2 = -1 + \frac{\varepsilon}{2} - \frac{\varepsilon^2}{8} + \mathcal{O}(\varepsilon^3) \qquad (6.3.1024)$$

The question is whether we can get the two previous relation without a priori knowledge of the exact solution of the initial polynomial equation? The answer is obviously: YES with the help of perturbation theory.

The technique is based on four steps:

1. In the first step, we assume that the solution of the polynomial equation is an expression of the type Taylor series on ε. Then we have:

$$x = X_0 + \varepsilon X_1 + \varepsilon^2 X_2 + \mathcal{O}(\varepsilon^3) \qquad (6.3.1025)$$

   where X_1, X_2, X_3 are obviously to be determined.

2. In the second step, we inject the solution in our hypothetical polynomial equation:

$$\left(X_0+\varepsilon X_1+\varepsilon^2 X_2+\mathcal{O}(\varepsilon^3)\right)^2-1=\varepsilon\left(X_0+\varepsilon X_1+\varepsilon^2 X_2+\mathcal{O}(\varepsilon^3)\right) \tag{6.3.1026}$$

As:

$$\begin{aligned}
x^2 =&X_0^2+2\left(\varepsilon X_1+\varepsilon^2 X_2+\mathcal{O}(\varepsilon^3)\right)X_0+\left(\varepsilon X_1+\varepsilon^2 X_2+\mathcal{O}(\varepsilon^3)\right)^2\\
=&X_0^2+2\left(\varepsilon X_1+\varepsilon^2 X_2+\mathcal{O}(\varepsilon^3)\right)X_0+\left(\left(\varepsilon X_1+\varepsilon^2 X_2\right)^2+2\left(\varepsilon X_1+\varepsilon^2 X_2\right)\mathcal{O}(\varepsilon^3)\varepsilon^3\right)\\
\cong&X_0^2+2\left(\varepsilon X_1+\varepsilon^2 X_2\right)X_0+\left(\varepsilon X_1+\varepsilon^2 X_2\right)^2+\mathcal{O}(\varepsilon^3)\\
=&X_0^2+2\left(\varepsilon X_1+\varepsilon^2 X_2\right)X_0+\left(\varepsilon^2 X_1^2+2\varepsilon^3 X_1X_2+\varepsilon^4 X_2^2\right)+\mathcal{O}(\varepsilon^3)\\
\cong&X_0^2+2\left(\varepsilon X_1+\varepsilon^2 X_2\right)X_0+\left(\varepsilon^2 X_1^2\right)+\mathcal{O}(\varepsilon^3)\\
=&X_0^2+2\varepsilon X_0X_1+\varepsilon^2\left(X_1^2+2X_0X_2\right)+\mathcal{O}(\varepsilon^3)
\end{aligned} \tag{6.3.1027}$$

and:

$$\varepsilon x=\varepsilon X_0+\varepsilon^2 X_1+\mathcal{O}(\varepsilon^3) \tag{6.3.1028}$$

It finally comes that the polynomial equation can be written as:

$$X_0^2-1+\varepsilon(2X_0X_1-X_0)+\varepsilon^2\left(X_1^2+2X_0X_2-X_1\right)+\mathcal{O}(\varepsilon^3)=0 \tag{6.3.1029}$$

3. In the third step we successively equalize the terms with 0 such as:

$$\begin{aligned}
&\mathcal{O}(\varepsilon^0):X_0^2-1=0\\
&\mathcal{O}(\varepsilon^1):2X_0X_1-X_0=0\\
&\mathcal{O}(\varepsilon^2):X_1^2+2X_0X_2-X_1=0\\
&\mathcal{O}(\varepsilon^3):\ \ldots
\end{aligned} \tag{6.3.1030}$$

4. Fourth and last step, we solve successively the polynomial equations above to get:

$$X_0=\pm 1\Rightarrow X_1=\frac{1}{2}\Rightarrow X_2=\pm\frac{1}{8} \tag{6.3.1031}$$

By injecting these results in the hypothetical solution:

$$x=X_0+\varepsilon X_1+\varepsilon^2 X_2+\mathcal{O}(\varepsilon^3) \tag{6.3.1032}$$

it is obvious to observe that we fall back on the certain solution:

$$x_1=1+\frac{\varepsilon}{2}+\frac{\varepsilon^2}{8}+\mathcal{O}(\varepsilon^3)\qquad x_2=-1+\frac{\varepsilon}{2}-\frac{\varepsilon^2}{8}+\mathcal{O}(\varepsilon^3) \tag{6.3.1033}$$

6.3.3.7.2 Perturbation theory of differential equations

Perturbation theory is therefore also often used to resolve numerous differential equations. This is the case for example in fluid mechanics, in General Relativity or quantum physics.

Again, rather than doing a super abstract and general theory, we will see the concept with an example as previously.

Consider the following ordinary differential equation with second member and constant coefficients:

$$y''=-\varepsilon y'-1 \tag{6.3.1034}$$

Or written in another way:

$$y''+\varepsilon y'=-1 \tag{6.3.1035}$$

with the boundary conditions:

$$y(0)=0, y'(0)=1 \tag{6.3.1036}$$

The exact resolution is relatively easy to get!

First we start with the homogeneous equation:

$$y''+\varepsilon y'=0 \tag{6.3.1037}$$

So it is a linear differential equation of order 2 with constant coefficients, equation that it is relatively easy to solve in the general case. Given the equation:

$$y''+p\varepsilon y'+qy=0 \tag{6.3.1038}$$

Assume that the function y which satisfies this differential equation is the of the form $y=e^{Kx}$ where K may be a complex number. Then we have:

$$K^2e^{Kx}+pKe^{Kx}+qe^{Kx}=0 \quad \text{or} \quad K^2+pK+q=0 \tag{6.3.1039}$$

provided, of course, that $e^{Kx}\neq 0$. This last relation is the auxiliary quadratic equation of the differential equation (characteristic polynomial in other words). It has two solutions/roots (it's a simple resolution of a polynomial of the second degree) which we denote in the general case: K_1,K_2. Which means that:

$$K_1^2e^{K_1x}+pK_1e^{K_1x}+qK_1e^{K_1x}=0 \quad \text{and} \quad K_2^2e^{K_2x}+pK_2e^{K_2x}+qK_2e^{K_2x}=0 \tag{6.3.1040}$$

are satisfied for the two roots. If we do the sum, since both are equal to the same constant:

$$\left(K_1^2e^{K_1x}+K_2^2e^{K_2x}\right)+p\left(K_1e^{K_1x}+K_2e^{K_2x}\right)+q\left(e^{K_1x}+e^{K_2x}\right)=0 \tag{6.3.1041}$$

Thus, it is immediate that the general solution of the homogeneous equation is of the type:

$$y_h=Ae^{K_1x}+Be^{K_2x} \tag{6.3.1042}$$

where A, B are obviously constants to be determined. Now we solve the characteristic polynomial:

$$K^2+\varepsilon K=0 \tag{6.3.1043}$$

It comes immediately that:

$$K_{1,2}=\frac{-\varepsilon\pm\sqrt{\varepsilon^2}}{2}\Rightarrow\begin{cases}K_1=-\varepsilon\\K_2=0\end{cases} \tag{6.3.1044}$$

Therefore:

$$y_h=Ae^{-\varepsilon x} \tag{6.3.1045}$$

Now a particular solution to:

$$y''+\varepsilon y'=-1 \tag{6.3.1046}$$

is relatively trivially a solution of the type:

$$y_p=Bx \tag{6.3.1047}$$

where B is of course a constant to be determined and which is equal (once injected into the differential equation):

$$y''+\varepsilon y'=-1\Rightarrow 0+\varepsilon B=1\Rightarrow B=-\frac{1}{\varepsilon} \tag{6.3.1048}$$

Therefore:

$$y_p=-\frac{x}{\varepsilon}+c^{te} \tag{6.3.1049}$$

Hence finally the general solution:

$$y=y_p+y_h=Ae^{-ex}-\frac{x}{\varepsilon}+c^{te} \tag{6.3.1050}$$

Then, with the initial conditions that are a for reminder:

$$y(0)=0,y'(0)=1 \tag{6.3.1051}$$

it is very easy to find A:

$$y(0)=0\Rightarrow Ae^{-e0}-\frac{0}{\varepsilon}+c^{te}=A+c^{te}=0 \tag{6.3.1052}$$

$$\Rightarrow c^{te}=-A \tag{6.3.1053}$$

We also have:

$$y'(0)=1\Rightarrow-\varepsilon Ae^{-\varepsilon 0}-\frac{1}{\varepsilon}+c^{te}=-\varepsilon A-\frac{1}{\varepsilon}+c^{te}=1\Rightarrow A=-\frac{1}{\varepsilon}\left(1-c^{te}+\frac{1}{\varepsilon}\right)=-\left(\frac{\varepsilon-\varepsilon c^{te}+1}{\varepsilon^2}\right) \tag{6.3.1054}$$

We are free to choose that $c^{te}=0$ which gives us:

$$A=-\left(\frac{\varepsilon+1}{\varepsilon^2}\right) \tag{6.3.1055}$$

Then:

$$y=y_p+y_h=Ae^{-\varepsilon x}-\frac{x}{\varepsilon}+c^{te} \tag{6.3.1056}$$

becomes:

$$\begin{aligned} y&=-\frac{\varepsilon+1}{\varepsilon^2}e^{-\varepsilon x}-\frac{x}{\varepsilon}+c^{te}\\ &=-\frac{1+\varepsilon}{\varepsilon^2}e^{-\varepsilon x}-\frac{x}{\varepsilon}+\frac{1+\varepsilon}{\varepsilon^2}\\ &=\frac{1+\varepsilon}{\varepsilon^2}\left(1-e^{-\varepsilon x}\right)-\frac{x}{\varepsilon} \end{aligned} \tag{6.3.1057}$$

Now that we have the general solution if ε is small we can take the development of order 4 in Maclaurin series of the exponential (see section Sequences and Series page 1703). Such as:

$$e^{-\varepsilon x}=1-\varepsilon x+\frac{1}{2}\varepsilon^2x^2-\frac{1}{6}\varepsilon^3x^3+\mathcal{O}(\varepsilon^4x^4) \tag{6.3.1058}$$

Injected into y this gives (you will notice that we sometimes express explicitly... the term of order 5 by anticipation...):

$$\begin{aligned} y&=\frac{1+\varepsilon}{\varepsilon^2}\left(1-e^{-\varepsilon x}\right)-\frac{x}{\varepsilon}\\ &=\frac{1+\varepsilon}{\varepsilon^2}\left(\varepsilon x-\frac{1}{2}\varepsilon^2x^2+\frac{1}{6}\varepsilon^3x^3-\mathcal{O}(\varepsilon^4x^4)\right)-\frac{x}{\varepsilon}\\ &=\frac{1}{\varepsilon^2}\left(\varepsilon x-\frac{1}{2}\varepsilon^2x^2+\frac{1}{6}\varepsilon^3x^3-\mathcal{O}(\varepsilon^4x^4)\right)+\frac{1}{\varepsilon}\left(\varepsilon x-\frac{1}{2}\varepsilon^2x^2+\frac{1}{6}\varepsilon^3x^3-\mathcal{O}(\varepsilon^4x^4)\right)-\frac{x}{\varepsilon}\\ &=\frac{1}{\varepsilon}x-\frac{1}{2}x^2+\frac{1}{6}\varepsilon x^3-\frac{1}{24}\varepsilon^2x^4+x-\frac{1}{2}\varepsilon x^2+\frac{1}{6}\varepsilon^2x^3-\mathcal{O}(\varepsilon^3x^4)-\frac{x}{\varepsilon}\\ &=x-\frac{1}{2}x^2+\varepsilon\left(-\frac{1}{2}x^2+\frac{1}{6}x^3\right)+\varepsilon^2\left(\frac{1}{6}x^3-\frac{1}{24}x^4\right)-\mathcal{O}(\varepsilon^3x^4) \end{aligned} \tag{6.3.1059}$$

Now that we have this development, what we want to show is that from a perturbative expansion we can find the same result in series and this without any prior knowledge on the solution.

Again, the development is done in 4 steps:

1. In the first step, we assume that the solution of differential equation is an expression of the type Taylor series on ε. Then we have:

   $$y=y_0+\varepsilon y_1+\varepsilon^2y_2+\mathcal{O}(\varepsilon^3) \tag{6.3.1060}$$

   where y_0,y_1,y_2 are obviously to be determined.

2. In the second step, we inject the hypothetical solution of our differential equation in itself with the initial conditions and we develop the whole.

   First the differential equation:

$$\begin{aligned} y''+\varepsilon y'+1 &= \left(y_0''+\varepsilon y_1''+\varepsilon^2 y_2''+\mathcal{O}(\varepsilon^3)\right)+\varepsilon\left(y_0'+\varepsilon y_1'+\varepsilon^2 y_2'+\mathcal{O}(\varepsilon^3)\right)+1 \\ &= (y_0''+1)+\varepsilon(y_1''+y_0')+\varepsilon^2(y_2''+y_1')+\mathcal{O}(\varepsilon^3)=0 \end{aligned} \tag{6.3.1061}$$

   then the initial conditions:

$$\begin{aligned} y(0) &= y_0(0)+\varepsilon y_1(0)+\varepsilon^2 y_2(0)+\mathcal{O}(\varepsilon^3)=0 \\ y'(0) &= \left(y_0'(0)-1\right)+\varepsilon y_1'(0)+\varepsilon^2 y_2'(0)+\mathcal{O}(\varepsilon^3)=0 \end{aligned} \tag{6.3.1062}$$

3. In the third step we equalize successively the terms with 0 such as:

$$\begin{array}{llll} \mathcal{O}(\varepsilon^0): & y_0''+1=0 & y_0(0)=0 & y_0'(0)-1=0 \\ \mathcal{O}(\varepsilon^1): & y_1''+y_0'=0 & y_1(0)=0 & y_1'(0)=0 \\ \mathcal{O}(\varepsilon^2): & y_2''+y_1'=0 & y_2(0)=0 & y_2'(0)=0 \\ \mathcal{O}(\varepsilon^3): & \cdots & & \end{array} \tag{6.3.1063}$$

4. In the fourth step we solve the differential equations listed above (if you do not see how we solve them do not hesitate to contact us!):

$$\begin{aligned} y_0 &= x-\frac{1}{2}x^2 \\ y_1 &= -\frac{1}{2}x^2+\frac{1}{6}x^3 \\ y_2 &= \frac{1}{6}x^3-\frac{1}{24}x^4 \end{aligned} \tag{6.3.1064}$$

   By injecting relations in the supposed solution developed in Taylor series and injected into the differential equation:

$$y=y_0+\varepsilon y_1+\varepsilon^2 y_2+\mathcal{O}(\varepsilon^3) \tag{6.3.1065}$$

   We fall back on:

$$y=x-\frac{1}{2}x^2+\varepsilon\left(-\frac{1}{2}x^2+\frac{1}{6}x^3\right)+\varepsilon^2\left(\frac{1}{6}x^3-\frac{1}{24}x^4\right)-\mathcal{O}(\varepsilon^3x^4) \tag{6.3.1066}$$

quote
6.3:1650.#56

> ❝ *Science is a differential equation. Religion is a boundary condition.* ❞
>
> Alan Turing

~95 %

★★★★☆

119 votes, 75.45%

6.4. Sequences and Series

SEQUENCES AND SERIES have a great importance in Applied Mathematics and that is why we devote to them a whole section. We will also see them often in various sections of the Mechanics chapters when we need to make some minor approximations (...) as well as in the sections of Economy and Quantitative Management Techniques. The reader should try to not confuse in what follows the concept of "sequences" with that of "series" which, while being similar in substance, are not always analyse mathematically in the same way.

We wanted to study in this section simple things without going to far within the topological concepts of sequences and series. However, those interested in more rigorous definitions can read the sections Fractals (see chapter of Theoretical Computing page 7038) and Topology (see chapter Topology page 2446) where many concepts about series are (supremum, infimum, subsequence, Bolzano-Weierstrass' theorem, etc.).

6.4.1 Sequences

Definition (#6.4.211): A "sequence" of a set is a family of elements indexed by the set of natural numbers (see section Numbers page 215) or by a part of it. In a vulgarized way we say that a sequence is a list of objects put in order, each with a order number. We typically write a sequence as:

$$(u_n) \quad \text{or} \quad (u_n)_{n\in\mathbb{N}} \tag{6.4.1}$$

where indexing is sometimes (by tradition...) without the 0.

For some sequences, we provide the first term u_1 (if indexing starts with 1 instead of 0), and a formula for any term u_{n+1} from the previous term u_n regardless $n\geq 1$. We name such a formulation a "recurring definition" and the sequence is defined "recursively" (and even if it is indexed from 0 instead of 1).

Before seeing some examples of sequences families that will be used in the various sections of the book (Population Dynamic, Economy, Nuclear Physics, etc.) let us see a small set of definitions as it is the tradition in mathematics...

Definitions (#6.4.212):

D1. Numbers (as sequence) are in "arithmetic progression" if the difference of two consecutive terms is equal to a constant r named the "reason".

D2. Numbers (as sequence) are in "geometric progression" if the ratio of two consecutive terms is equal to a constant r also named the "reason".

D3. Numbers (as sequence) are in "harmonic progression" if the inverse of two consecutive terms are in arithmetic progression.

Therefore, a number b is respectively the arithmetic mean, geometric, harmonic of a and c if the numbers a, b, c are respectively in an arithmetic, geometric or harmonic progression.

Remark

For the definitions of the averages listed above see the section Statistics.

Definitions (#6.4.213):

D1. A "majorated sequence" is a sequence such that there is a real number M such that: $\forall n \in \mathbb{N},\ u_n \leq M$

D2. A "minorated sequence" is a sequence such that there is a real number m such that: $\forall n \in \mathbb{N},\ u_n \geq m$

D3. A "bounded sequence" is a sequence that is both majorated and minorated.

D4. A sequence (u_n) is named "increasing sequence" if $\forall n \in \mathbb{N},\ u_{n+1} - u_n > 0$

D5. A sequence (u_n) is named "decreasing sequence" if $\forall n \in \mathbb{N},\ u_{n+1} - u_n < 0$

Remark

If a sequence is increasing or decreasing, we sometimes just say it is a "monotonous sequence" (without specifying if its increasing or decreasing).

D6. A sequence (u_n) is named "constant sequence" if $\forall n \in \mathbb{N},\ u_{n+1} = u_n$

We will now see some practical important arithmetic and geometric sequences that will be used later in other sections of this textbook.

But keep in mind during the lecture that we can express sequences most of time in two ways:

- A "closed formula" or "closed relation" for a sequence $(u_n)_{n\in\mathbb{N}}$ is a relation for u_n using a fixed finite number of operations on n only (and nothing else!). This is what we normally think of as a function of n only,
- A "recursive definition" for a sequence $(u_n)_{n\in\mathbb{N}}$ consists of a recurrence relation: a relation relating a term of the sequence to previous terms (terms with smaller index n) and an initial condition.

6.4.1.1 Arithmetic Sequences

Definition (#6.4.214): We say that numbers or "terms" are in an "arithmetic sequence" when the difference of their sequential value is equal to a constant r named the "reason" of the sequence so that (recursive relation):

$$u_{k+1}=u_k+r \tag{6.4.2}$$

where r is the "reason" of the progression. We then obviously have if the indexing starts from 0:

$$\boxed{u_k=u_0+k\cdot r} \tag{6.4.3}$$

✍**Examples:**

E1. The sequence:

$$n,n+1,n+2,n+3,\ldots,n+k \tag{6.4.4}$$

where n is a constant and the reason is equal 1.

E2. The sequence:

$$n+0\cdot x,n+1\cdot x,n+2\cdot x,n+3\cdot x,\ldots,n+k\cdot x \tag{6.4.5}$$

is an arithmetic sequence of reason x.

Thus, if we write u_n any term of the sequence (u_n) of reason r, we have:

$$u_0=\text{c}^{te},u_1=u_0+r,u_2=u_1+r=u_0+2\cdot r,\ldots,u_n=u_0+n\cdot r \tag{6.4.6}$$

We have the following properties for this type of sequences:

P1. A term whose rank is the average of the ranks of the other two terms is the arithmetic mean of these two terms.

Proof 6.17.2. Consider now (u_n) an arithmetic sequence of reason r given by the previous development:

$$u_n=u_0+n\cdot r \tag{6.4.7}$$

and $a,b,k\in\mathbb{N}$ such as $a+b=2k$, then we have:

$$\begin{aligned}u_a+u_b&=u_0+(a)r+u_0+(b)r=2u_0+2\left(\frac{a+b}{2}\right)r\\&=2\left(u_0+\left(\frac{a+b}{2}\right)r\right)=2u_{\frac{a+b}{2}}=2u_k\end{aligned} \tag{6.4.8}$$

and so:

$$u_k=\frac{u_a+u_b}{2} \tag{6.4.9}$$

with $k=\dfrac{a+b}{2}$.

■ Q.E.D.

P2. For three consecutive terms u_n, u_{n+1}, u_{n+2} in an arithmetic sequence of reason r, the second term is the arithmetic mean of the other two.

Proof 6.17.3. Let us write:

$$\frac{u_n+u_{n+2}}{2}=\frac{u_n+(u_n+2r)}{2}=2\frac{u_n+r}{2}=u_n+r=u_{n+1} \tag{6.4.10}$$

■ Q.E.D.

P3. If $u_1, u_2, u_3, \ldots, u_n, \ldots$ is an arithmetic sequence of ratio r, then the n-th partial sum S_n (that is to say, the sum of the first n terms to the power of 1) is given by:

$$S_n=\frac{n}{2}(2u_1+n\cdot r)=\frac{n}{2}(u_1+u_n) \tag{6.4.11}$$

when indexing is starts from 1.

Proof 6.17.4. We can write the sequence:

$$\begin{aligned}S_n&=u_1+u_2+u_3+u_4+\ldots+u_n\\&=u_1+(u_1+r)+(u_1+2r)+(u_1+3r)+\ldots+(u_1+(n-1)r)\end{aligned} \tag{6.4.12}$$

Playing with the second line, we get:

$$S_n=(u_1+u_1+u_1+u_1+\ldots+u_1)+(r+2r+3r+4r+\ldots+(n-1)r) \tag{6.4.13}$$

What can be simplified even more:

$$S_n=nu_1+r(1+2+3+4+\ldots+(n-1)) \tag{6.4.14}$$

Considering that we will prove a little bit later that the simple following Gauss series:

$$1+2+3+4+\ldots+n \tag{6.4.15}$$

is equal to:

$$S_1=\frac{n(n+1)}{2} \tag{6.4.16}$$

We then have for:

$$1+2+3+4+\ldots+(n-1) \tag{6.4.17}$$

the following relation:

$$S_1=\frac{(n-1)n}{2} \tag{6.4.18}$$

We thus get:

$$S_n=nu_1+r\frac{(n-1)n}{2}=\frac{n}{2\left[2u_1+r(n-1)\right]} \tag{6.4.19}$$

We see with the latter relation that if $u_1=r=1$ we fall back on the simple Gauss series.

As:

$$u_n=u_1+r(n-1) \tag{6.4.20}$$

when the indexation starts from 1 we thus get:

$$S_n=\frac{n}{2}\left[u_1+u_n\right] \tag{6.4.21}$$

■ Q.E.D.

We will see other types of summations a little bit further below during our study of series!

6.4.1.2 Harmonic Sequences

Definition (#6.4.215): We say that numbers $\frac{1}{a},\frac{1}{b},\frac{1}{c},\ldots$ generates an "harmonic sequence" when their inverses are in arithmetical progression (also with a "reason" r. We represent this progress by:

$$\boxed{h_{k+1}=\frac{1}{u_{k+1}}=\frac{1}{u_k+r}} \tag{6.4.22}$$

We then obviously if the indexing starts from 0:

$$\frac{1}{u_k}=\frac{1}{u_0+k\cdot r} \tag{6.4.23}$$

Moreover, we assume, in what follows, that there is no zero denominator.

By sharing this type of sequences successively in groups containing 2^n terms, we observe that each of the group is bigger than the next one. For example:

$$\frac{1}{1}>\frac{1}{2}+\frac{1}{3}>\frac{1}{4}+\frac{1}{5}+\frac{1}{6}+\frac{1}{7}>\ldots \tag{6.4.24}$$

We can also see that each term is the harmonic mean of the previous and consecutive one:

Proof 6.17.5.

$$\begin{aligned}\frac{1}{\frac{1}{h_{k-1}}+\frac{1}{h_{k+1}}}&=\frac{1}{\frac{1}{\frac{1}{u_0+(k-1)r}}+\frac{1}{\frac{1}{u_0+(k+1)r}}}=\frac{1}{(u_0+(k-1)r)+(u_0+(k+1)r)}\\&=\frac{1}{2u_0+2kr}=\frac{1}{2(u_0+kr)}=\frac{1}{2}\frac{1}{h_k}\end{aligned} \tag{6.4.25}$$

Thus:

$$\frac{1}{h_k}=\frac{2}{\frac{1}{h_{k-1}}+\frac{1}{h_{k+1}}} \tag{6.4.26}$$

So finally:

$$h_k=\frac{\frac{1}{h_{k-1}}+\frac{1}{h_{k+1}}}{2} \tag{6.4.27}$$

■ Q.E.D.

6.4.1.3 Geometric Sequences

Definition (#6.4.216): A "geometric sequence" is a sequence of numbers such that each of them is equal to the previous n multiplied by a constant number q that we also name the "reason" or "common ratio" of the sequence. What will denote by:

$$n, qn, q^2n, q^3n, \ldots \tag{6.4.28}$$

Thus, if we denote by u_n any term of the sequence (u_n), we have (trivial):

$$\boxed{u_k = u_0 q^k} \tag{6.4.29}$$

Here are some properties for such a type of sequence (without proof until now... except if some readers ask for them because most are really trivial):

P1. (trivial) The quotient of two terms of the same sequence is a power of the reason q whose exponent equals the difference in rank of the two terms chosen (simple ratio of two same bases with different powers).

P2. (trivial) If we multiply or divide term by term two geometric sequences, we get a third geometric sequence whose raison equal the product (respectively the quotient) of the reasons of the two chosen sequences (simple operation with the reasons of the two original sequences).

P3. In a geometric sequence, a term whose rank is the average of the ranks of the other two terms is the geometric mean (see section Statistics page 565) of these two terms (reread many times if needed...).

Let us prove the property P3:

Proof 6.17.6. Given a geometric sequence with real positive reason q, we have for recall:

$$u_n = u_0 q^n \tag{6.4.30}$$

Let a, b be the ranks of two terms of the geometric sequence, then we have:

$$u_a \cdot u_b = u_0^2 q^a q^b = \left(u_0 q^{\frac{a+b}{2}}\right)^2 = u_{\frac{a+b}{2}}^2 = u_k^2 \tag{6.4.31}$$

and thus:

$$u_k = \sqrt{u_a \cdot u_b} \tag{6.4.32}$$

■ Q.E.D.

Corollary 6.17.1. We have as corollary that for three consecutive terms $n, n+1, n+2$ in a geometric progression, the second term is the geometric mean of the other two.

Proof 6.17.7. We have:

$$u_n u_{n+2} = u_0^2 q^n q^{n+2} = u_0^2 q^{n+(n+2)} = u_0^2 q^{2n+2} \tag{6.4.33}$$

Thus:

$$\sqrt{u_n u_{n+2}} = \sqrt{u_0^2 q^{2n+2}} = u_0 q^{n+1} = u_{n+1} \tag{6.4.34}$$

■ Q.E.D.

However, there are some special sequences that have special properties that we find very frequently in mathematical or theoretical physics in this textbook. Without going into too much detail, here's a partial list of with the important proofs that we will have to use later:

6.4.1.4 Cauchy Sequence

It is often interesting for the mathematician, as much as for the physicist, to know the properties of a sequence with a given type of progression. The most important property is the limit to which it tends.

> **Remark**
>
> The reader who is not comfortable with topology can skip the text that follows... and whoever wants to know more about Cauchy sequences may read the section Topology or also particularly the section on Fractals at page 7038.

Definition (#6.4.217): Let (X,d) a metric space (see section Topology page 2447), we say that the sequence:

$$(x_n)_{n\in\mathbb{N}},\quad x_n\in X\quad \forall n\in\mathbb{N} \tag{6.4.35}$$

converges to $x\in X$ if (by definition!):

$$\forall\varepsilon>0,\exists N\in\mathbb{N}\quad\text{as}\quad n\geq N\Rightarrow d(x_n,x)<\varepsilon \tag{6.4.36}$$

In other words, more we go far in the sequence, the more points are close (in the sense of the metric d) to each other.

If we chose a particular metric (the Euclidean one for example) and a discrete sequence the above definition will look like this:

$$\forall\varepsilon>0,\exists N\in\mathbb{N}\quad\text{as}\quad n\geq N\Rightarrow|a_n-a|<\varepsilon \tag{6.4.37}$$

where the convergence point is therefore a and we have:

$$\lim_{n\rightarrow+\infty}a_n=a \tag{6.4.38}$$

In the example of the figure below where the sequence seems to converge to 1.13 we observe that for a given non-zero positive ε, there is a particular n which we denote N ($n=17$ in the figure below) from which the sequence converges:

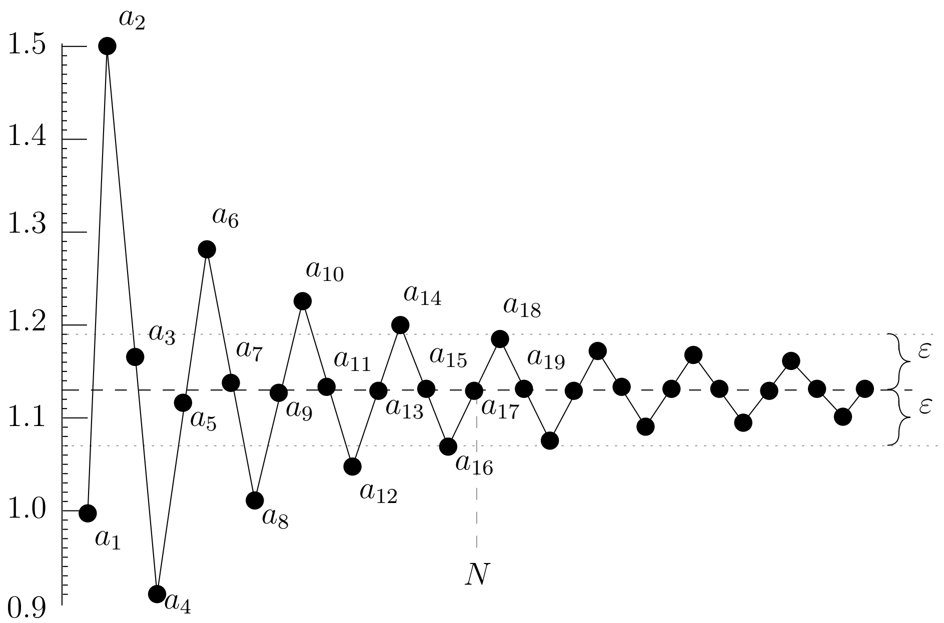


Figure 6.65: **Illustration of the principle of convergence of a sequence**

However, the above definition of the convergence makes problem because the number x should be known. In most cases of interest x is unfortunately not known. To break this deadlock, Cauchy had the idea to propose the following definition:

Definition (#6.4.218): We say by definition that a sequence $(x_n)_{n\in\mathbb{N}}$ of elements of X is a "Cauchy sequence" if:

$$\forall\varepsilon>0, \exists N\in\mathbb{N}\,\text{such that}\,\forall m,n\geq N\,\text{we have}\,d(x_n,x_m)\leq\varepsilon \qquad (6.4.39)$$

The reader must notice that it is not sufficient for each term to become arbitrarily close to the preceding term. This is why require that $|a_{N+1}-a_N|<\varepsilon$ is not sufficient!

It is almost obvious then that any convergent sequence is a Cauchy sequence (well there are some subtleties that we will not reference for now).

Remark

This criterion therefore facilitates some proofs because it helps to show the existence of a limit without involving its value, generally unknown.

Theorem 6.18. Let us now prove that theorem that asses that any convergent sequence is a Cauchy sequence.

Proof 6.18.1. Consider a sequence u_n converging to the value l (which is unknown to us) and $\varepsilon>0$ (randomly selected). Then there exists according to the definition of a convergent sequence, $n\in\mathbb{N}$ such that:

$$n\geq N\Rightarrow d(u_n,l)\leq\frac{\varepsilon}{2} \qquad (6.4.40)$$

The choice to write $\varepsilon/2$ is completely arbitrary but in fact we anticipate the final result of the demonstration so that it is more aesthetic...

Therefore for $p,q>N$ (in fact know the value of N is irrelevant, since it should work for any value... well don't forget that N depends on $\varepsilon/2$) we have using the triangle inequality (see section Vector Calculus page 1828):

$$d(u_p,u_q)\leq d(u_p,l)+d(l,u_q) \tag{6.4.41}$$

and because $d(u_n,l)\leq \dfrac{\varepsilon}{2}$ we can write:

$$d(u_p,u_q)\leq \frac{\varepsilon}{2}+\frac{\varepsilon}{2} \tag{6.4.42}$$

Therefore:

$$d(u_p,u_q)\leq \varepsilon \tag{6.4.43}$$

That may be a bit abstract so let's see an example with the harmonic sequence as an example to close the proof:

$$u_n=\sum_{k=1}^{n}\frac{1}{k} \tag{6.4.44}$$

First, nothing it is not forbidden to us to take $n\geq 2$ (otherwise it will be hard to make a difference between two terms...).

Therefore we take the Euclidean distance:

$$d(u_p,u_q)=d(u_{2n},u_n)=u_{2n}-u_n=\sum_{k=1}^{2n}\frac{1}{k}-\sum_{k=1}^{n}\frac{1}{k}=\sum_{k=n+1}^{2n}\frac{1}{k} \tag{6.4.45}$$

First, the reader will note that in all cases since $k\leq 2n$ is between $n+1$ and $2n$. Which brings us to write:

$$\frac{1}{k}\geq\frac{1}{2n}\qquad \forall\, n+1\leq k\leq 2n \tag{6.4.46}$$

So from this inequality it comes automatically that each term of the sum on the left below will be greater than each term of the sum on the right:

$$\sum_{k=n+1}^{2n}\frac{1}{k}\geq\sum_{k=n+1}^{2n}\frac{1}{2n} \tag{6.4.47}$$

With (just do a particular example):

$$\sum_{k=n+1}^{2n}\frac{1}{2n}=\frac{1}{2} \tag{6.4.48}$$

Therefore:

$$\sum_{k=n+1}^{2n}\frac{1}{k}\geq\frac{1}{2} \tag{6.4.49}$$

Now the idea is to see if the sum on the left is therefore greater than or equal to $\varepsilon=1/2$ and this for any n. Thus the suite is not convergent!

Thus, the idea is that we found an ε for which the Cauchy criterion is deficient.

■ Q.E.D.

So it is not because the points are always closer to each other that they converge to a given point, because this point may not exist.

The best example is probably the following (it is also a little bit stupid example but...):

✍**Example:**

Let us take $X=\mathbb{Q}$ and the absolute difference as distance:

$$d(p,q)=|p-q| \tag{6.4.50}$$

Given z an irrational number and $q_j\in\mathbb{Q}$ with $j\in\mathbb{N}^*$ such that:

$$|z-q_j|\leq\frac{1}{j} \tag{6.4.51}$$

The idea is that greater is j, more the rational number q_j is near the irrational z and we know we can find such a sequence. Let us show that the q_1,q_2 we could be able to build form a Cauchy sequence! Indeed using triangle inequality:

$$|q_m-q_n|\leq|z-q_m|-\leq|z-q_n|\leq\frac{1}{m}-\frac{1}{m} \tag{6.4.52}$$

and therefore is a Cauchy sequence if and only if $|q_m-q_n|\leq\varepsilon$ if:

$$m,n\geq\frac{1}{2\varepsilon} \tag{6.4.53}$$

We have thus found a N (equal to $\dfrac{1}{2\varepsilon}$) which satisfies our definition of a Cauchy sequence. But this sequence does not converge in $\mathbb{Q}$ otherwise z would be rational. You can check this with π and the sequence:

$$\begin{array}{ll}
q_1=\dfrac{3}{1}=3 & |\pi-q_1|\cong 0.141592\leq\dfrac{1}{1}\\
q_2=\dfrac{13}{4}=3.25 & |\pi-q_1|\cong 0.108407\leq\dfrac{1}{2}\\
q_3=\dfrac{19}{6}=3.1\bar{6} & |\pi-q_3|\cong 0.02507401\leq\dfrac{1}{3}\\
q_4=\dfrac{22}{7}=3.1428571428571428 & |\pi-q_4|\cong 0.001264489267\leq\dfrac{1}{4}\\
\ldots &
\end{array} \tag{6.4.54}$$

Remark

Mathematicians use such results to define the set of irrational and also by using some additional topological concepts.

We have just seen that a Cauchy sequence is not necessarily a convergent sequence in X. The inverse is however true: any convergent sequence is a Cauchy sequence!!

6.4.1.5 Fibonacci Sequence

If we calculate a sequence of numbers starting with 0 and 1, such that each term is equal to the sum of the two previous ones, we can form the following sequence:

$$0,1,1,2,3,5,8,13,21,34,55,89,144,\ldots \tag{6.4.55}$$

therefore, if we designate the different terms by:

$$u_0,u_1,u_2,\ldots \tag{6.4.56}$$

We build therefore the following sequence law:

$$u_{n+2}=u_n+u_{n+1} \tag{6.4.57}$$

More often written in the following form:

$$\boxed{F_{n+2}=F_n+F_{n+1}} \tag{6.4.58}$$

The Fibonacci sequence has many interesting strong properties, which will be developed later. However, it seems to be the first "recurring sequence" known in history (hence the fact that we are talking about it in this textbook).

The origin of this sequence seem to come from a rabbit problem asked to Fibonacci in 11202 (holocene calendar). Starting with a couple of rabbits, how much couples of rabbits will we get after a given number of months knowing that each pair produces a new pair every month (and no couples die...), which becomes productive only after two months. Therefore we have:

- Beginning: We have nothing (0)
- 1st month: We buy a couple of baby rabbits (1).
- 2nd month: The couple of rabbits are now adults (1).
- 3rd month: We have the couple of rabbits that make a new couple of baby rabbits. We have two couples (2).
- 4th month: We have two couple of adults with a new couple of babies. We have three couples (3).
- 5th month: We have three couples of adults rabbits and two new couples of baby rabbits. We have five couples (5)
- and so on...

Let us take now a "in real life" example (this is typically a biased scientific example because you will always finish to find in Nature what your are looking for to argue your theories with a least on particular example...): the heart of some flowers! The scales of a pineapple or pinecone form two families of spirals wound in opposite directions. On a pine cone, you will count 5 spiral in one direction and 8 in the other, on pineapples, 8 and 13, on sunflowers 21 and 34. Each time we get Fibonacci numbers!

A famous illustration of this is to do draw the following simple figure (named"Fibonacci Spiral") which reproduces the Fibonacci numbers on a grid plan with squares and corners connect with arc circles:

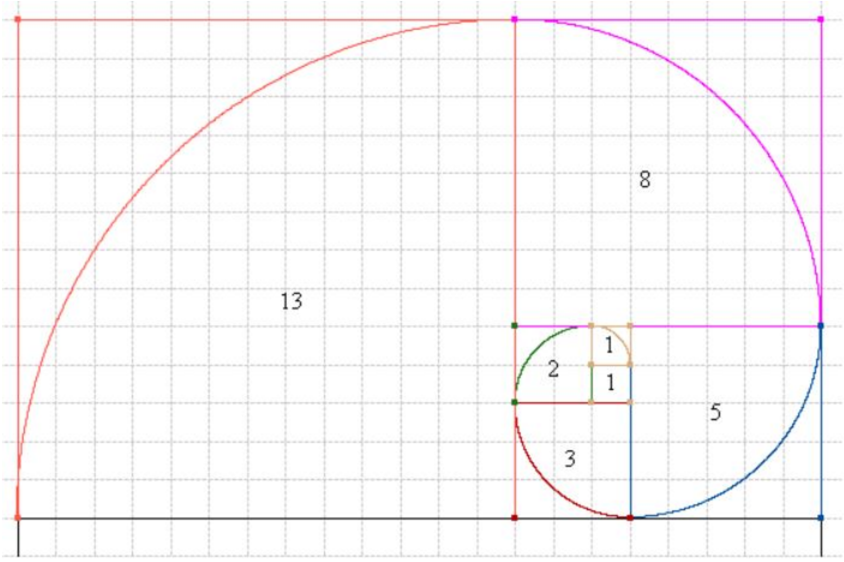


Figure 6.66: **Fibonacci spiral**

Remark

The Fibonacci spiral (quarter-circles tangent to the interior of each square) must not be confused with the "Golden Spiral" (a logarithmic spiral whose growth factor is the Golden ratio φ)!

Let us now consider the limit of the Fibonacci sequence, let:

$$\lim_{n\to+\infty}\frac{F_n}{F_{n-1}}=L \qquad (6.4.59)$$

by the properties of limits:

$$\lim_{n\to+\infty}\frac{F_n}{F_{n-1}}=\lim_{n\to+\infty}\frac{F_{n+1}}{F_n}=L \qquad (6.4.60)$$

As the construction of the Fibonacci sequence is given by:

$$F_{n+1}=F_n+F_{n-1} \qquad (6.4.61)$$

Using this equation to substitute, we get:

$$\lim_{n\to+\infty}\frac{F_{n+1}}{F_n}=\lim_{n\to+\infty}\frac{F_n+F_{n-1}}{F_n}=1+\frac{1}{L} \qquad (6.4.62)$$

and so we get the equivalent well known equation (see page 1389):

$$1+\frac{1}{L}=L\Leftrightarrow L^2-L-1=0 \tag{6.4.63}$$

and finally solving for L using the quadratic formula yields to (by keeping the only solution that makes sense):

$$\boxed{\lim_{n\rightarrow+\infty}\frac{F_n}{F_{n-1}}=L=\frac{1+\sqrt{5}}{2}=\varphi} \tag{6.4.64}$$

that is the "Golden ratio"!

We will come back later (see page 1783) on this sequence to prove another relation it has with the Golden ratio (see page 1389) through the direct determination of the n term value!

6.4.1.6 Logic Sequences/Psychologist Sequences

Definition (#6.4.219): Psychologists name "logical sequences", sequences that they write with an idea in mind, and they name "logical" people who find their idea, although there are other possibilities mathematically speaking (but psychologist don't know anything about real logic).

For example if you have to find the next number X to the logic sequence:

$$4,5,10,50,400,3500,X \tag{6.4.65}$$

In fact, you make the difference between the last and prior-previous number then you multiply by 10 therefore the next number is $X=(3500-400)\cdot 10=31000$.

From a mathematical point of view, any number is suitable to replace X, also exists for each value of X, a polynomial in n that takes the values $4,5,10,50,400,3500$ for $n=0,1,2\ldots,6$.

For the example above we can take for example:

$$P(n)=-\frac{1}{120}(x-6)(2646x^5-25735x^4+88490x^3-124945x^2+59584x+80) \tag{6.4.66}$$

and is such that $P(0)=4,P(1)=5,P(2)=10,\ldots,P(6)=0$.

6.4.2 Series

The physicists often needs to simply and formally solve problems, to approximate some given "terms" of their equations. For this purpose, they will use the properties of some given series. Also statisticians and financial analysts often face to series they need to simplify.

Definition (#6.4.220): Let be given an infinite number sequence:

$$u_1, u_2, u_3, \ldots, u_n, \ldots \tag{6.4.67}$$

The expression:

$$u_1 + u_2 + u_3 + \ldots + u_n + \ldots \tag{6.4.68}$$

is named a "numeric series".

Definition (#6.4.221): The partial sum of the first n terms of the series is named "partial sum" and denoted by:

$$\boxed{S_n = u_1 + u_2 + u_3 + \ldots + u_n} \tag{6.4.69}$$

If the following limit denoted S exists and is finite:

$$S = \lim_{n \to +\infty} S_n \tag{6.4.70}$$

we name it "sum of the series" and we say that then that the series is a "convergent series" (it is therefore a Cauchy series!). However, if the limit does not exist, we say that the series is a "divergent series" and has no sum (for details see further below when we will deal with some empirical convergence criteria).

Theorem 6.19. Also let us prove for fun (because it is almost trivial) that if $\sum\limits_{k \geq 0} u_k$ is a convergent numerical series then:

$$\lim_{k \to +\infty} u_k = 0 \tag{6.4.71}$$

But the opposite is not necessarily true!! In fact remember the example during our study above of Cauchy sequences with the harmonic series $\sum_{k=1}^{n} \dfrac{1}{k}$ that is not convergent even if the terms tends to zero when $k \to +\infty$.

Proof 6.19.1. We assume first that $\sum\limits_{k \geq 0} u_k$ is a convergent series and denote its limit by S. Let:

$$S_n = \sum_{k=0}^{n} u_k \qquad S_{n-1} = \sum_{k=0}^{n-1} u_k \tag{6.4.72}$$

Therefore:

$$S_n - S_{n-1} = u_n \tag{6.4.73}$$

However, if the series is really convergent:

$$\lim_{n \to +\infty} S_n = \lim_{n \to +\infty} S_{n-1} = S \tag{6.4.74}$$

So finally:

$$\lim_{n \to +\infty} S_n - \lim_{n \to +\infty} S_{n-1} = \lim_{n \to +\infty} (S_n - S_{n-1}) = \lim_{n \to +\infty} u_n = S - S = 0 \tag{6.4.75}$$

■ Q.E.D.

Following a reader request, let us prove in detail (not only assuming that it is intuitive) that the harmonic series:

$$H=\sum_{n=1}^{+\infty}\frac{1}{n}=1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8}+\ldots \tag{6.4.76}$$

diverges.

Proof 6.19.2. Suppose the series converges to H, i.e.:

$$H=1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8}+\ldots \tag{6.4.77}$$

Then:

$$H\geq 1+\frac{1}{2}+\underbrace{\frac{1}{4}+\frac{1}{4}}+\underbrace{\frac{1}{6}+\frac{1}{6}}+\underbrace{\frac{1}{8}+\frac{1}{8}}+\ldots \tag{6.4.78}$$

Calculating the underbraces, we get:

$$H\geq 1+\frac{1}{2}+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\ldots=\frac{1}{2}+1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\ldots=\frac{1}{2}+H \tag{6.4.79}$$

This contradiction concludes the proof.

■ Q.E.D.

Remark

Many people are struggling understanding intuitively why the harmonic series diverges but the p-harmonic series converges. For sure there are methods and applications to prove convergence and divergence, but many people have trouble understanding intuitively why it is. We know we must never trust too much or intuition, but this is hard for many to grasp. In both cases, the terms of the series are getting smaller, hence are approaching zero, but they both result in different answers!

$$\begin{aligned}\sum_{n=1}^{+\infty}\frac{1}{n}&=1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\cdots=\text{diverges}\\ \sum_{n=1}^{+\infty}\frac{1}{n^2}&=1+\frac{1}{2^2}+\frac{1}{3^2}+\frac{1}{4^2}+\cdots=\text{converges}\end{aligned} \tag{6.4.80}$$

Perhaps a convincing way is to make an analytical prolongation to change the sum into an integral (see section Differential and Integral Calculus page 1534):

$$\int_1^{+\infty}\frac{1}{x^2}\mathrm{d}x=-\frac{1}{x}\bigg|_1^{+\infty}=-\frac{1}{+\infty}+1=1 \tag{6.4.81}$$

converges but:

$$\int_{1}^{+\infty} \frac{1}{x}\mathrm{d}x = \ln(x)\Big|_{1}^{+\infty} = +\infty \qquad (6.4.82)$$

In the next subsection we will see how to calculate the partial sum of some classic series that are important in physics, statistics and finance. We will start with Gauss arithmetic series that are an expression of the sum of the n first non-zero integers raised to a given power k in a condensed form. The application of this condensed form of a series has an important practical use in physics, statistics and finance when we wish to simplify the expression of certain results.

6.4.2.1 Gauss Series

Gauss arithmetic series are an expression of the sum of the n first non-zero integers raised to a given power k in a condensed form S_k. The application of this condensed form of a series has an important practical use in physics, statistics and finance when we wish to simplify the expression of certain results.

It is said that Gauss have found an attractive method in 11786 (holocene calendar) to determine the arithmetic sum of the first n integers at the power of 1 when he was nine years old (...):

$$\begin{array}{rl} S_1 = & 1+2+3+\ldots+(n-1)+n \\ +S_1 = & n+(n-1)+(n-2)+\ldots+2+1 \\ \hline 2S_1 = & (n+1)+(n+1)+(n+1)+\ldots+(n+1)+(n+1) \end{array} \qquad (6.4.83)$$

After simplifying we find easily the following closed formula (notice that this is not a recursive relation!):

$$\boxed{S_n = \frac{n(n+1)}{2}} \qquad (6.4.84)$$

for $n \geq 0$. Let us indicate that each intermediate sum of the series (1, 3, 6, 10, 15, etc.) is named "triangular number" since it is possible to represent it in the following way:

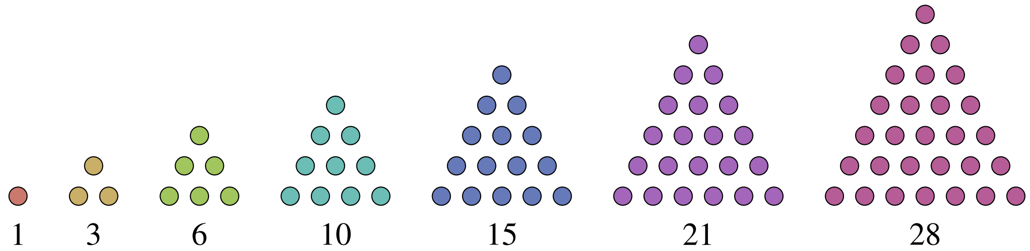


Figure 6.67: **Triangular numbers (author: Lauren K. Williams)**

We can continue with higher powers but not as exercises because these relations are very useful!

Now let us calculate the very important case that we find ourselves in a number of other sections (Economy, Quantum Wave Physics, Industrial Engineering, etc.) and that is the sum of the first n square integer numbers (still non-zero!).

Let us write for this:

$$\begin{aligned}S_1&=1+2+3+\ldots+n\\S_2&=1^2+2^2+3^2+\ldots+n^2\end{aligned}\qquad(6.4.85)$$

We know from Newton's binomial theorem (see section Calculus page 1364):

$$(k+1)^3-k^3=3k^2+3k+1\qquad(6.4.86)$$

So we can write and add member to member the n following equalities:

$$\begin{array}{r}2^3-1^3=3\cdot 1^2+3\cdot 1+1\\3^3-2^3=3\cdot 2^2+3\cdot 2+1\\4^3-3^3=3\cdot 3^2+3\cdot 3+1\\\ldots\qquad\qquad\\(n+1)^3-n^3=3n^3+3n+1\\\hline(n+1)^3-1=3(1^2+2^2+\ldots+n^2)+3(1+2+\ldots+n)+n\end{array}\qquad(6.4.87)$$

And the sum can be simplified as:

$$(n+1)^3-1=3S_2+3S_1+1n\qquad(6.4.88)$$

After some elementary algebra manipulations we get:

$$\begin{aligned}3S_2&=(n+1)^3-1-3S_1-n=(n+1)^3-1-3\frac{n(n+1)}{2}-n\\&=(n+1)^3-(n+1)-3\frac{n(n+1)}{2}=\frac{n(n+1)}{2}\left[2(n+1)^2-2-3n\right]\end{aligned}\qquad(6.4.89)$$

Therefore:

$$S_2=\frac{n(n+1)}{6}\left[2n^2+4n+2-2-3n\right]=\frac{n+1}{6}(2n^2+n)\qquad(6.4.90)$$

Finally:

$$\boxed{S_2=\frac{n(n+1)(2n+1)}{6}}\qquad(6.4.91)$$

We continue with the sum of the first n cubes (non-zero) integers. The principle is the same as before, we write:

$$\begin{aligned}S_1&=1+2+\ldots+n\\S_2&=1^2+2^2+\ldots+n^2\\S_3&=1^3+2^3+\ldots+n^3\end{aligned}\qquad(6.4.92)$$

We know from Newton's binomial theorem (see section Calculus page 1364):

$$(k+1)^4-k^4=4k^3+6k^2+4k+1\qquad(6.4.93)$$

We get by varying k from 1 to n, n relations that we can add a member to member:

$$\begin{array}{l} 2^4-1^4=4\cdot 1^3+6\cdot 1^2+4\cdot 1+1 \\ 3^4-2^4=4\cdot 2^3+6\cdot 2^2+4\cdot 2+1 \\ \quad\ldots \\ \underline{(n+1)^4-n^4=4n^3+6n^2+4n+1 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad} \\ (n+1)^4-1=4(1^3+2^3+\ldots+n^3)+6(1^2+2^2+\ldots+n^2)+4(1+2+\ldots+n)+n \end{array} \tag{6.4.94}$$

And the sum can be simplified as:

$$(n+1)^4-1^4=4S_3+6S_2+4S_1+n \tag{6.4.95}$$

Giving after development:

$$(n+1)^4-1^4=4S_3+6\frac{n(n+1)(2n+1)}{6}+4\frac{n(n+1)}{2}+n \tag{6.4.96}$$

And after a first simplification:

$$(n+1)^4-1^4=4S_3+n(n+1)(2n+1)+2n(n+1)+n \tag{6.4.97}$$

And a second simplification:

$$\begin{aligned} 4S_3&=-n(n+1)(2n+1)-2n(n+1)-n+(n+1)^4-1\\ &=-\left((n^2+n)(2n+1)\right)-(2n^2+2n)-n+(n^4+4n^3+6n^2+4n+1)-1\\ &=-(2n^3+n^2+2n^2+n)+n^4+4n^3+4n^2+n\\ &=n^4+2n^3+n^2=n^2(n^2+2n+1)=n^2(n+1)^2 \end{aligned} \tag{6.4.98}$$

The result is therefore:

$$S_3=\left(\frac{n(n+1)}{2}\right)^2 \tag{6.4.99}$$

or written differently:

$$\boxed{S_3=\frac{n^2(n+1)^2}{4}} \tag{6.4.100}$$

For sure, we can continue like this during a long time, but from a certain value of the power things get a bit more complicated (furthermore the method is a little bit boring). Thus, one of the members of the Bernoulli family (it was a family of very talented mathematicians... as you can see int the Biographies chapter) founded a general relation working for any power by defining what we name the "Bernoulli polynomial" (see further below).

Let us conclude with a last case that we will need during our study of Fourier series. We put:

$$S_n=x+x^3+x^5+\ldots+x^{2n-1} \tag{6.4.101}$$

We want express this expression (series) as rational fraction. To do this, we multiply all by x^2. So we have two expressions:

$$\begin{aligned} S_n&=x+x^3+x^5+\ldots+x^{2n-1}\\ x^2S_n&=x^3+x^5+x^7+\ldots+x^{2n+1} \end{aligned} \tag{6.4.102}$$

We subtract the first from the second:

$$S_n(x^2-1)=x^{2n+1}-x \tag{6.4.103}$$

Finally:

$$\boxed{S_n=\frac{x^{2n+1}-x}{x^2-1}} \tag{6.4.104}$$

Most of times, to indicate that this is for odd powers we prefer to write:

$$S_{2n+1}=\frac{x^{2n+1}-x}{x^2-1} \tag{6.4.105}$$

Similarly, for the needs of the section of Economy, we have:

$$\begin{aligned}S_n&=1+x^2+x^4+\ldots+x^{2n}\\ x^2S_n&=x^2+x^4+x^6+\ldots+x^{2n+2}\end{aligned} \tag{6.4.106}$$

Therefore:

$$S_n(1-x^2)=1-x^{2n+2} \tag{6.4.107}$$

Finally:

$$\boxed{S_n=\frac{1-x^{2n+2}}{1-x^2}} \tag{6.4.108}$$

Most of times, to indicate that this is for even powers we prefer to write:

$$S_{2n}=\frac{1-x^{2n+2}}{1-x^2} \tag{6.4.109}$$

6.4.2.1.1 Bernoulli's Numbers and Polynomials

As we have seen above, it is possible to express the sum of the first n non-zero integers to a given power (the first four have been proved previously) following the below relations where we put now $n:=n+1$ as we want now n to be the number of terms for which we want the sum including 0 (hence the negative sign in the relations below that we did not have earlier):

$$\begin{aligned}&S_0=n\\ &S_1=\frac{1}{2}n^2-\frac{1}{2}n=\frac{n(n-1)}{2}\\ &S_2=\frac{1}{3}n^3-\frac{1}{2}n^2+\frac{1}{6}n=\frac{n(n-1)(2n-2)}{6}\\ &S_3=\frac{1}{4}n^4-\frac{1}{2}n^3+\frac{1}{4}n^2=\frac{n^2(n-1)^2}{4}\\ &S_4=\frac{1}{5}n^5-\frac{1}{2}n^4+\frac{1}{3}n^3-\frac{1}{30}n=\frac{n(n-1)(2n-1)(3n^2-3n-1)}{30}\\ &S_5=\frac{1}{6}n^6-\frac{1}{2}n^5+\frac{5}{12}n^4-\frac{1}{12}n^2=\frac{n^2(2n^2-2n-1)(n-1)^2}{12}\\ &\quad\ldots\end{aligned} \tag{6.4.110}$$

It is said that Jacob Bernoulli then noticed that the polynomials S_p had the form:

$$S_p=\frac{1}{p+1}n^{p+1}-\frac{1}{2}n^p+\frac{p}{12}n^{p-1}+0\cdot n^{p-2}+\ldots \tag{6.4.111}$$

In this expression, the numbers $(1,-1/2,1/12,0,\ldots)$ seem not to depend on p. More generally, after trial and error we see that the polynomial can be written as:

$$\begin{aligned}S_p(n)&=\sum_{k=0}^{p}\frac{B_k}{k!}\frac{p!}{(p+1-k)!}n^{p+1-k}=\frac{1}{p+1}\sum_{k=0}^{p}C_k^{p+1}n^{p+1-k}\\&=\frac{B_0}{0!}\frac{n^{p+1}}{p+1}+\frac{B_1}{1!}n^p+\frac{B_2}{2!}pn^{p-1}+\frac{B_3}{3!}p(p-1)n^{p-2}+\ldots+\frac{B_p}{1!}n\end{aligned} \tag{6.4.112}$$

Giving by identification the "Bernoulli's numbers":

$$B_0=1,B_1=-\frac{1}{2},B_2=\frac{1}{6},B_3=0,\ldots \tag{6.4.113}$$

Theorem 6.20. Thereafter, it seems that mathematicians in their research fell randomly (???) on the fact that the Bernoulli numbers could be expressed by the series:

$$G(z)=\frac{z}{e^z-1}=\sum_{k=0}^{+\infty}B_k\frac{z^k}{k!} \tag{6.4.114}$$

with $|z|<2\pi$. In other words, the generating function of Bernoulli numbers would be $G(z)$. If we expand the first terms of this series:

$$G(z)=\left(\frac{e^z-1}{z}\right)^{-1}=\left(1+\frac{z}{2!}+\frac{z^2}{3!}+\frac{z^3}{4!}+\ldots\right)^{-1}=1-\frac{z}{2}+\frac{z^2}{12}-\frac{z^4}{720}+\frac{z^6}{30240}-\frac{z^8}{1209600}+\ldots \tag{6.4.115}$$

Proof 6.20.1. We have seen during our study of complex numbers (see section Numbers page 251) that:

$$e^z=1+\frac{z}{1!}+\frac{z^2}{2!}+\ldots \tag{6.4.116}$$

Therefore:

$$G(z)=\left(\frac{z+\dfrac{z^2}{2!}+\dfrac{z^3}{3!}+\ldots}{z}\right)^{-1}=\left(1+\frac{z}{2!}+\frac{z^2}{3!}+\ldots\right)^{-1} \tag{6.4.117}$$

Let us write now:

$$G(z):=a_0+a_1z+a_2z^2+\ldots \tag{6.4.118}$$

Then we must have:

$$\begin{aligned}G(z)\frac{1}{G(z)}&=(a_0+a_1z+a_2z^2+\ldots)\frac{1}{\left(1+\dfrac{z}{2!}+\dfrac{z^2}{3!}+\ldots\right)^{-1}}\\&=(a_0+a_1z+a_2z^2+\ldots)\left(1+\frac{z}{2!}+\frac{z^2}{3!}+\ldots\right)=1\end{aligned} \tag{6.4.119}$$

We see (by distributing) that:

$$(a_0+a_1z+a_2z^2+\ldots)\left(1+\frac{z}{2!}+\frac{z^2}{3!}+\ldots\right)=a_0+\left(a_1+\frac{a_0}{2!}\right)z+\left(a_2+\frac{a_1}{2!}+\frac{a_0}{3!}\right)z^2+\ldots \tag{6.4.120}$$

for all this to be equal to unity we must have:

$$\begin{aligned} a_0&=1\\ a_1+\frac{a_0}{2!}&=0\\ a_2+\frac{a_1}{2!}+\frac{a_0}{3!}&=0\\ &\ldots \end{aligned} \tag{6.4.121}$$

From the second equation we get:

$$a_1=-\frac{1}{2} \tag{6.4.122}$$

and from the third equation we get:

$$a_2=-\frac{1}{6}+\frac{1}{4}=\frac{1}{12} \tag{6.4.123}$$

etc. Continuing this way we show that:

$$a_3=0,\, a_4=-\frac{1}{720},\ldots \tag{6.4.124}$$

It is obvious that this method allows us to calculate by hand only the first terms of this series.

Thus, based on (as we will see later, this is the expression of an "ordinary generating function"):

$$G(z)=\frac{z}{e^z-1}=\sum_{k=0}^{+\infty}B_k\frac{z^k}{k!} \tag{6.4.125}$$

we find that the first Bernoulli numbers are:

k	B_k
0	1
1	-1/2
2	1/6
3	0
4	-1/30
5	0
6	1/42
7	0
8	-1/30
9	0
10	5/66
11	0
12	-691/2730
13	0
14	7/6
...	...

Table 6.15: **Bernoulli numbers**

The reader will have noticed that $B_k = 0$ when k is odd and different from 1. That's why sometimes the above relation is given by (remember that 0 is an even number!):

$$G(z) = \frac{z}{e^z - 1} = \sum_{k \geq 0\,(\text{even})} B_k \frac{z^k}{k!} \tag{6.4.126}$$

■ Q.E.D.

We see easily that the values of the Bernoulli numbers can not be described in a simple way. In fact, they are essentially values of the zeta Riemann function (see below page 1682) for negative integer values of the variable, and these numbers are associated with profound theoretical properties that go beyond the study of this textbook. Furthermore, the Bernoulli numbers also appear in the Taylor series expansion of circular and hyperbolic trigonometric tangent functions in the Euler-Maclaurin formula (see below).

With a small modification it is possible to define the "Bernoulli polynomials" $B_k(x)$ by:

$$G(z,x) = \frac{ze^{zx}}{e^z - 1} = \sum_{k=0}^{+\infty} B_k(x) \frac{z^k}{k!} \tag{6.4.127}$$

with:

$$G(z,0) = G(z) \tag{6.4.128}$$

Theorem 6.21. Furthermore, it is normally easy to observe that:

$$\frac{\partial G(z,x)}{\partial x} = z\frac{ze^{zx}}{e^z - 1} = zG(z,x) = \sum_{k=0}^{+\infty} \frac{\mathrm{d}B_k(x)}{\mathrm{d}x} \frac{z^k}{k!} \tag{6.4.129}$$

and therefore it normally easy to deduce that:

$$\sum_{k=0}^{+\infty} \frac{\mathrm{d}B_k(x)}{\mathrm{d}x} \frac{z^k}{k!} = \sum_{k=1}^{+\infty} B_{k-1}(x) \frac{z^k}{(k-1)!} \tag{6.4.130}$$

Proof 6.21.1. On one side we have:

$$zG(z,x) = z\sum_{k=0}^{+\infty} B_k(x) \frac{z^k}{k!} = \sum_{k=0}^{+\infty} B_k(x) \frac{z^{k+1}}{k!} = \sum_{k=1}^{+\infty} B_{k-1}(x) \frac{z^k}{(k-1)!} \tag{6.4.131}$$

and another we have:

$$zG(z,x) = \sum_{k=0}^{+\infty} \frac{\mathrm{d}B_k(x)}{\mathrm{d}x} \frac{z^k}{k!} \tag{6.4.132}$$

So:

$$\sum_{k=0}^{+\infty} \frac{\mathrm{d}B_k(x)}{\mathrm{d}x} \frac{z^k}{k!} = \sum_{k=1}^{+\infty} B_{k-1}(x) \frac{z^k}{(k-1)!} \tag{6.4.133}$$

■ Q.E.D.

And by identification of the coefficients we deduce:

$$\frac{\mathrm{d}B_0(x)}{\mathrm{d}x}=0 \tag{6.4.134}$$

and for $k\geq 1$:

$$\frac{\mathrm{d}B_k(x)}{\mathrm{d}x}=kB_{k-1}(x) \tag{6.4.135}$$

It is then easy to deduce that the polynomials $B_k(x)$ are of degree k:

$$\begin{aligned}
&B_0(x)=1\\
&B_1(x)=x-\frac{1}{2}\\
&B_2(x)=x^2-x+\frac{1}{6}\\
&B_3(x)=x^3-\frac{3}{2}x^2+\frac{1}{2}x\\
&B_4(x)=x^4-2x^3+x^2-\frac{1}{30}\\
&B_5(x)=x^5-\frac{5}{2}x^4+\frac{5}{3}x^3-\frac{1}{6}x\\
&\ldots
\end{aligned} \tag{6.4.136}$$

Here is a plot of these polynomials:

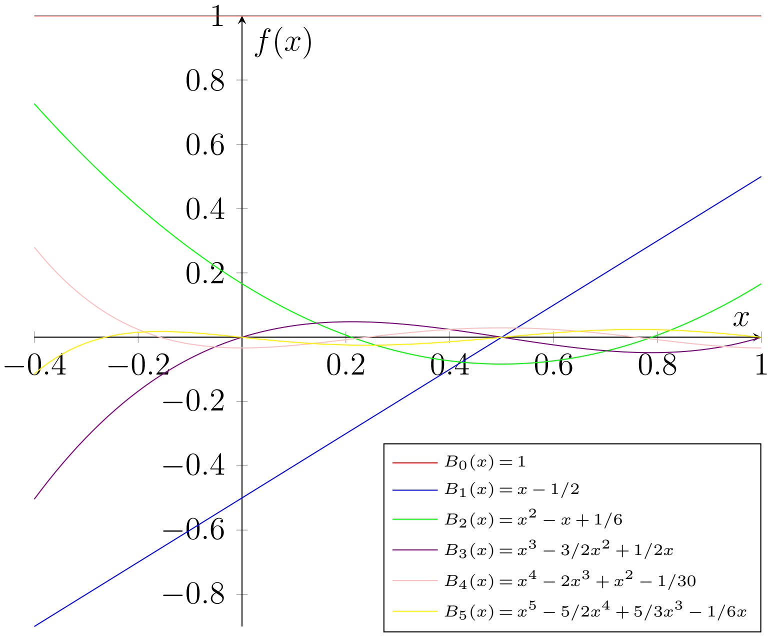


Figure 6.68: Some Bernoulli polynomials

What is remarkable is that using the Bernoulli polynomials, we see that it is possible to write the S_p as follows after some trials:

$$\boxed{S_p(n)=\sum_{k=0}^{n-1}k^p=\frac{1}{p+1}\big(B_{p+1}(n)-B_{p+1}(0)\big)} \qquad (6.4.137)$$

Some write this relations even otherwise. Indeed, from previous relation, we can write:

$$(p+1)\sum_{k=0}^{n-1}k^p=B_{p+1}(n)-B_{p+1}(0) \qquad (6.4.138)$$

using:

$$\frac{\mathrm{d}B_k(x)}{\mathrm{d}x}=kB_{k-1}(x) \qquad (6.4.139)$$

We have:

$$\begin{aligned}&\frac{\mathrm{d}B_{p+1}(x)}{\mathrm{d}x}=(p+1)B_p(x)\\ \Rightarrow &B_p(x)=\frac{1}{p+1}B'_{p+1}(x)\\ \Rightarrow &\int\limits_0^k B_p(x)\mathrm{d}x=\frac{1}{p+1}\int\limits_0^k B'_{p+1}(x)\mathrm{d}x=\frac{1}{p+1}B_{p+1}(x)\Big|_0^k=\frac{1}{p+1}\left[B_{p+1}(x)-B_{p+1}(0)\right]\end{aligned} \qquad (6.4.140)$$

So we just demonstrated in an engineer way that:

$$\boxed{S_p(n)=\sum_{k=0}^{n-1}k^p=\int\limits_0^k B_p(x)\mathrm{d}x} \qquad (6.4.141)$$

However, we can now ask ourselves what happens to the partial sum of arithmetic and geometric sequences as presented earlier in this section.

6.4.2.1.2 Euler-Maclaurin formula

The Euler–Maclaurin formula is a relation for the difference between an integral and a closely related sum. It can be used to approximate integrals by finite sums, or conversely to evaluate finite sums and infinite series using integrals and the machinery of calculus. For example, many asymptotic expansions are derived from the formula[10].

If m and n are natural numbers and $f(x)$ is a real or complex valued continuous function for real numbers x in the interval $[m,n]$, then the integral:

$$I(n)=\int\limits_m^n f(x)\mathrm{d}x \qquad (6.4.142)$$

---

[10] It will be useful to us during our study of the electromagnetic effect in the section of Wave Quantum Physics.

can be approximated by the sum (or vice versa):

$$S(n) = f(m+1) + \cdots + f(n-1) + f(n) \tag{6.4.143}$$

(see rectangle method page 6623). The Euler-Maclaurin formula provides expressions for the difference between the sum and the integral in terms of the higher derivatives $f^{(k)}(x)$ evaluated at the endpoints of the interval, that is to say when $x = m$ and $x = n$.

Explicitly, for p a positive integer and a function $f(x)$ that is p times continuously differentiable on the interval $[m,n]$, we have:

$$S(n) - I(n) = \sum_{k=1}^{p} \frac{B_k}{k!}\left(f^{(k-1)}(n) - f^{(k-1)}(m)\right) + R_p \tag{6.4.144}$$

where B_k is the k th Bernoulli number (with $B_1 = 1/2$) and R_p, is an error term which depends on n, m, p, and f and is usually small for suitable values of p. The formula is often written with the subscript taking only even values, since the odd Bernoulli numbers are zero except for B_1. In this case we have:

$$\sum_{i=m}^{n} f(i) - \int\limits_{m}^{n} f(x)\mathrm{d}x = \frac{f(n)+f(m)}{2} + \sum_{k=1}^{\lfloor p/2 \rfloor} \frac{B_{2k}}{(2k)!}\left(f^{(2k-1)}(n) - f^{(2k-1)}(m)\right) + R_p \tag{6.4.145}$$

or alternatively:

$$\sum_{i=m+1}^{n} f(i) - \int\limits_{m}^{n} f(x)\mathrm{d}x = \frac{f(n)-f(m)}{2} + \sum_{k=1}^{\lfloor p/2 \rfloor} \frac{B_{2k}}{(2k)!}\left(f^{(2k-1)}(n) - f^{(2k-1)}(m)\right) + R_p \tag{6.4.146}$$

Proof 6.21.2. One of the fundamental concepts of calculus is the correspondence between sums and integrals. Given any sufficiently well-behaved continuous function $f(x)$, consider the sum the integral defined by:

$$S(n) = \sum_{k=0}^{n} f(k) \qquad I(n) = \int\limits_{x=0}^{n} f(x)\mathrm{d}x \tag{6.4.147}$$

These two functions obviously have some similarities, but they also have significant differences. For example, the summation is taken over the values of $f(k)$ at discrete integer arguments $k = 0, 1, 2, \ldots, n$, whereas the integral is taken over the values of $f(x)$ for arguments x varying continuously from 0 to n. Strictly speaking, the function $S(n)$ is defined only for integer arguments, while $I(m)$ is defined for any non-negative real value of n. Nevertheless, we can derive an interesting and useful relation between these two functions.

First, notice that $I(n)$ can be written in the form:

$$I(n) = \int\limits_{x=0}^{1} [f(0+x) + f(1+x) + f(2+x) + \ldots + f(n-1+x)]\mathrm{d}x \tag{6.4.148}$$

The Taylor series expansion of each of the individual terms of the integrand is of the form:

$$f(k+x)=\frac{f^{(0)}(k)}{0!}+\frac{f^{(1)}(k)}{1!}x+\frac{f^{(2)}(k)}{2!}x^2+\frac{f^{(3)}(k)}{3!}x^3+\ldots \tag{6.4.149}$$

where $f^{(j)}$ denotes the jth derivative of f. The integral of this term from $x=0$ to 1 is:

$$\int\limits_{x=0}^{1} f(k+x)\mathrm{d}x=\frac{f^{(0)}(k)}{1!}+\frac{f^{(1)}(k)}{2!}+\frac{f^{(2)}(k)}{3!}+\frac{f^{(3)}(k)}{4!}+\ldots \tag{6.4.150}$$

Inserting all these expressions into the original integral, we get:

$$I(n)=\frac{S(0)}{1!}+\frac{S(1)}{2!}+\frac{S(2)}{3!}+\frac{S(3)}{4!}+\ldots \tag{6.4.151}$$

where:

$$S(j)=\sum_{k=0}^{n-1} f^{(j)}(k) \tag{6.4.152}$$

Now, if we had started with the derivative of $I(n)$ in place of simply $I(n)$ itself, we would have arrived at the same expression:

$$I(n)=\frac{S(0)}{1!}+\frac{S(1)}{2!}+\frac{S(2)}{3!}+\frac{S(3)}{4!}+\ldots \tag{6.4.153}$$

except that each $f^{(j)}$ would be replaced with $f^{(j+1)}$, and hence each $S(j)$ would be replaced with $S(j+1)$, and $I(n)$ would be replaced with $I^{(1)}(n)$ where:

$$I^{(j)}(n)=\int\limits_{x=0}^{n} f^{j}(x)\mathrm{d}x=f^{(j-1)}(n)-f^{(j-1)}(0) \tag{6.4.154}$$

Thus beginning with successively higher derivatives of f, we get the infinite sequence of relations:

$$\begin{aligned}
I^{(0)}(n)&=\frac{S(1)}{1!}+\frac{S(2)}{2!}+\frac{S(3)}{3!}+\frac{S(4)}{4!}+\ldots\\
I^{(1)}(n)&=\frac{S(2)}{1!}+\frac{S(3)}{2!}+\frac{S(4)}{3!}+\frac{S(5)}{4!}+\ldots\\
I^{(2)}(n)&=\frac{S(3)}{1!}+\frac{S(4)}{2!}+\frac{S(5)}{3!}+\frac{S(6)}{4!}+\ldots\\
&\text{etc.}
\end{aligned} \tag{6.4.155}$$

Each of these equations has (in general) infinitely many terms, but we can solve the system of equations for any finite numbers of terms by taking just a restricted portion of these equations. For example, taking just the terms up to $S(6)$ from the first four equations, we have (in matrix form):

$$\begin{bmatrix}
\frac{1}{1!} & \frac{1}{2!} & \frac{1}{3!} & \frac{1}{4!} & \frac{1}{5!} & \frac{1}{6!}\\
0 & \frac{1}{1!} & \frac{1}{2!} & \frac{1}{3!} & \frac{1}{4!} & \frac{1}{5!}\\
0 & 0 & \frac{1}{1!} & \frac{1}{2!} & \frac{1}{3!} & \frac{1}{4!}\\
0 & 0 & 0 & \frac{1}{1!} & \frac{1}{2!} & \frac{1}{3!}\\
0 & 0 & 0 & 0 & \frac{1}{1!} & \frac{1}{2!}\\
0 & 0 & 0 & 0 & 0 & \frac{1}{1!}
\end{bmatrix}
\begin{bmatrix} S(0)\\ S(1)\\ S(2)\\ S(3)\\ S(4)\\ S(5)\\ S(6) \end{bmatrix}
=
\begin{bmatrix} I^{(0)}\\ I^{(1)}\\ I^{(2)}\\ I^{(3)}\\ I^{(4)}\\ I^{(5)}\\ I^{(6)} \end{bmatrix} \tag{6.4.156}$$

Solving this system of equations gives:

$$\begin{bmatrix} S(1) \\ S(2) \\ S(3) \\ S(4) \\ S(5) \\ S(6) \end{bmatrix} = \begin{bmatrix} \frac{B_0}{0!} & \frac{B_1}{1!} & \frac{B_2}{2!} & \frac{B_3}{3!} & \frac{B_4}{4!} & \frac{B_5}{5!} \\ 0 & \frac{B_0}{0!} & \frac{B_1}{1!} & \frac{B_2}{2!} & \frac{B_3}{3!} & \frac{B_4}{4!} \\ 0 & 0 & \frac{B_0}{0!} & \frac{B_1}{1!} & \frac{B_2}{2!} & \frac{B_3}{3!} \\ 0 & 0 & 0 & \frac{B_0}{0!} & \frac{B_1}{1!} & \frac{B_2}{2!} \\ 0 & 0 & 0 & 0 & \frac{B_0}{0!} & \frac{B_1}{1!} \\ 0 & 0 & 0 & 0 & 0 & \frac{B_0}{0!} \end{bmatrix} \tag{6.4.157}$$

where:

$$B_0 = 1 \quad B_1 = -\frac{1}{2} \quad B_2 = \frac{1}{6} \quad B_3 = 0 \quad B_4 = -\frac{1}{30} \quad B_5 = 0 \quad B_6 = \frac{1}{42} \tag{6.4.158}$$

and so on are the Bernoulli numbers. Thus we have:

$$\mathrm{S}(1) = \frac{B_0}{0!}I(0) + \frac{B_1}{1!}I(1) + \frac{B_2}{2!}I(2) + \frac{B_3}{3!}I(3) + \ldots \tag{6.4.159}$$

Inserting the expressions for each $I^{(j)}$ from relation:

$$I^{(j)}(n) = \int\limits_{x=0}^{n} f^{(j)}(x)\mathrm{d}x = f^{(j-1)}(n) - f^{(j-1)}(0) \tag{6.4.160}$$

this gives:

$$S(0) = \frac{B_0}{0!}\int\limits_{x=0}^{n} f(x)\mathrm{d}x + \frac{B_1}{1!}[f(n) - f(0)] + \frac{B_2}{2!}\left[f^{(1)}(n) - f^{(1)}(0)\right] + \ldots \tag{6.4.161}$$

Recall that $S(0)$ is defined as the sum of $f(k)$ for $k = 0, 1, 2, \ldots, n-1$. Therefore, if we wish to express this in terms of a summation from $k = 0$ to n, we must add $f(n)$ to both sides. Noting that $B_1 = -1/2$, this leads to the result:

$$\boxed{\sum_{x=0}^{n} f(x) = \int\limits_{x=0}^{n} f(x)\mathrm{d}x + \frac{1}{2}[f(n) + f(0)] + \frac{B_2}{2!}[f(1)(n) - f(1)(0)] + \ldots} \tag{6.4.162}$$

This is named the "Euler-Maclaurin formula", a generalization of Bernoulli's formula for the sum of powers of the first n integers. Often used in the following form:

$$\sum_{x=0}^{n} f(x) - \int\limits_{x=0}^{n} f(x)\mathrm{d}x = \frac{1}{2}[f(n) + f(0)] + \frac{B_2}{2!}[f(1)(n) - f(1)(0)] + \ldots \tag{6.4.163}$$

or using a condensed notation:

$$S - I = \sum_{k=1}^{p} \frac{B_k}{k!}\left(f^{(k-1)}(n) - f^{(k-1)}(m)\right) + R_p \tag{6.4.164}$$

We should note that this series is divergent for most applications, but the error is less than the first neglected term, so the Euler-Maclaurin formula is often a useful method of approximation, relating the series summation to the continuous integral of a function.

■ Q.E.D.

✍**Example:**

The Bernoulli numbers from B_1 to B_7 are $1/2, 1/6, 0, -1/30, 0, 1/42, 0$. Therefore the low-order cases of the Euler-Maclaurin formula are:

$$\begin{aligned}
\sum_{i=m}^{n} f(i) - \int\limits_{m}^{n} f(x)\mathrm{d}x &= \frac{f(m)+f(n)}{2} + \int\limits_{m}^{n} f'(x)P_1(x)\mathrm{d}x \\
&= \frac{f(m)+f(n)}{2} + \frac{1}{6}\frac{f'(n)-f'(m)}{2!} - \int\limits_{m}^{n} f''(x)\frac{P_2(x)}{2!}\mathrm{d}x \\
&= \frac{f(m)+f(n)}{2} + \frac{1}{6}\frac{f'(n)-f'(m)}{2!} + \int\limits_{m}^{n} f'''(x)\frac{P_3(x)}{3!}\mathrm{d}x \\
&= \frac{f(m)+f(n)}{2} + \frac{1}{6}\frac{f'(n)-f'(m)}{2!} - \frac{1}{30}\frac{f'''(n)-f'''(m)}{4!} \\
&- \int\limits_{m}^{n} f^{(4)}(x)\frac{P_4(x)}{4!}\mathrm{d}x \\
&= \frac{f(m)!f(n)}{2} + \frac{1}{6}\frac{f'(n)-f'(m)}{2!} - \frac{1}{30}\frac{f'''(n)-f'''(m)}{4!} \\
&+ \int\limits_{m}^{n} f^{(5)}(x)\frac{P_5(x)}{5!}\mathrm{d}x \\
&= \frac{f(m)+f(n)}{2} + \frac{1}{6}\frac{f'(n)-f'(m)}{2!} - \frac{1}{30}\frac{f'''(n)-f'''(m)}{4!} \\
&+ \frac{1}{42}\frac{f^{(5)}(n)-f^{(5)}(m)}{6!} - \int\limits_{m}^{n} f^{(6)}(x)\frac{P_6(x)}{6!}\mathrm{d}x \\
&= \frac{f(m)+f(n)}{2} + \frac{1}{6}\frac{f'(n)-f'(m)}{2!} - \frac{1}{30}\frac{f'''(n)-f'''(m)}{4!} \\
&+ \frac{1}{42}\frac{f^{(5)}(n)-f^{(5)}(m)}{6!} + \int\limits_{m}^{n} f^{(7)}(x)\frac{P_7(x)}{7!}\mathrm{d}x
\end{aligned} \tag{6.4.165}$$

6.4.2.2 Arithmetic Series

We have shown above that the partial sum of a Gauss series (analogous to the sum of the terms of an arithmetic progression of reason $r = 1$) was given by:

$$S_1 = \frac{n(n+1)}{2} \tag{6.4.166}$$

If we not denote the value of the n-th term by u_n instead of n, the development that we made for the series of Gauss then brings us to:

$$S_n = \frac{n(u_n+1)}{2} \tag{6.4.167}$$

And if we denote the first term 1 of the Gauss series u_0 then we have:

$$\boxed{S_n = \frac{n(u_n+u_0)}{2}} \tag{6.4.168}$$

Which gives us simple the partial sum of the n-terms of an arithmetic sequence of reason r.

✍**Examples:**

E1. A simple Gauss series of reason 1 starting at 4, finishing at 6:

$$S_3 = 4+5+6 = \frac{n(u_n+u_0)}{2} = \frac{3(6+4)}{2} = 15 \tag{6.4.169}$$

E2. Now an arithmetic partial sum series of reason 2 starting at 4, finishing at 8:

$$4+6+8 = \frac{n(u_n+u_0)}{2} = \frac{3(8+4)}{2} = 18 \tag{6.4.170}$$

Remark

The reader will have observed that the reason r does not appear in the latter relation. Indeed, by taking (always) the same development that for the Gauss series, the term r is simplified and vanish!

6.4.2.3 Geometric Series

Similarly, with a geometrical sum where we have for recall:

$$u_n = u_0 q^n \tag{6.4.171}$$

we have therefore:

$$\begin{array}{rr} & S_n = u_0 + u_1 + \cdots + u_{n-1} + u_n \\ & qS_n = u_1 + u_2 + \cdots + u_n + u_{n+1} \\ - & \overline{s_n - qS_n = u_0 + u_1 - u_1 + u_2 - u_2 + \cdots + u_n - u_n - u_{n+1}} \end{array} \tag{6.4.172}$$

The last relation is written (after simplification):

$$S_n - qS_n = (1-q)S_n = u_0 - u_{n+1} \tag{6.4.173}$$

and if $q \neq 1$ we get:

$$S_n = \frac{u_0 - u_{n+1}}{(1-q)} \tag{6.4.174}$$

which can be written by factoring u_0:

$$\boxed{S_n = u_0 \frac{(1-q^{n+1})}{(1-q)}} \tag{6.4.175}$$

If q is positive and less than 1, as n approaches infinity we have the result that will be used extensively in the section Economy:

$$S_{n \to +\infty} \underset{q<1}{=} u_0 \frac{1}{(1-q)} \tag{6.4.176}$$

That we found often in the following form (special case):

$$S_{n \to +\infty} \underset{x<1}{=} = 1 + x + x^2 + x^3 + \ldots + x^n = \frac{1}{1-x} \tag{6.4.177}$$

✍**Examples:**

E1. Consider the following geometric series of reason $q = 2$:

$$1 \cdot 2^0 + 1 \cdot 2^1 + 1 \cdot 2^2 + 1 \cdot 2^3 + \cdots + 1 \cdot 2^n = 1 + 2 + 4 + 8 + \cdots + 2^n \tag{6.4.178}$$

to calculate the sum of the first four terms $\{1, 2, 3, 4\}$, we take the power of 2 equivalent of $n = 2$ (zero is not taken into account). We then get well: $S_3 = 15$.

E2. Consider the quite famous geometric series of reason $q = 1/2$. We then have:

$$S_n = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots = \frac{1}{2} \frac{1 - \left(\frac{1}{2}\right)^{n+1}}{1 - \frac{1}{2}} = 1 - \frac{1}{2^{n+1}} \tag{6.4.179}$$

On the request of a reader let deal with the famous case $1 = 0.99999\ldots$ whose rigorous derivation use geometric series! First let use see three naive proof of this equality.

First way that doesn't involve geometric series, we start from:

$$\frac{1}{3} = 0.33333\ldots \tag{6.4.180}$$

We multiply by 3 on both sides. That gives the obvious result (as there is no room for a number between 1 and $1 = 0.99999\ldots$):

$$1 = 0.99999\ldots \tag{6.4.181}$$

Second way that also don't involve geometric series, we start from:

$$a=0.99999\ldots \tag{6.4.182}$$

We multiply by 10 on both sides:

$$10\cdot a=9.99999\ldots \tag{6.4.183}$$

We subtract a from both sides:

$$\underbrace{10\cdot a-a}_{9a}=\underbrace{9.99999\ldots-0.99999\ldots}_{9} \tag{6.4.184}$$

Therefore:

$$\begin{gathered}9\cdot a=9\Longrightarrow a=1\\0.99999\ldots=1\end{gathered} \tag{6.4.185}$$

Finally the last naive proof still not using geometric series is a very short argument deduced from the following fact (so-named "Archimedean property of the set of real numbers"):

If two real numbers are different, then there is at least a third number between them, different from the other two. This third number can be the average between the two. Now, as we cannot insert a number between $0.99999...$ and 1; they are therefore necessarily equal.

Now that we have seen the three common naive proof, let us see the rigorous one using geometric series properties!

We know that:

$$0.99999\ldots=0.9+0.09+0.009+\ldots \tag{6.4.186}$$

We define $0.9999...$ as a geometric series, that is to say a sum in which each term is equal to the previous one multiplied by a constant, here $1/10$.

We say that it is a geometric series with a ratio of $1/10$, and we write:

$$0.99999\ldots=0.9+0.09+0.009+\ldots=\frac{9}{10}+\frac{9}{10^2}+\frac{9}{10^3}+\cdots=\sum_{k=1}^{+\infty}\frac{9}{10^k} \tag{6.4.187}$$

But using the proof of the sum of the first n terms of a geometric series, we get:

$$S_n=u_0\frac{(1-q^{n+1})}{(1-q)}=0.9\cdot\frac{1-\frac{1}{10^n}}{1-\frac{1}{10}} \tag{6.4.188}$$

Hence:

$$S_n=1-\frac{1}{10^n} \tag{6.4.189}$$

But:

$$0.9999\ldots=\lim_{n\to+\infty}S_n=\lim_{n\to+\infty}1-\frac{1}{10^n}=1 \tag{6.4.190}$$

Therefore:

$$0.99999\ldots=1 \tag{6.4.191}$$

6.4.2.3.1 Zeta function and Euler's product formula

The German mathematician Riemann named "zeta" a function already studied before him, but that he examined when the value is a complex number (see section Numbers page 243). This function is represented as a series of inverse powers of integers. This how typically the series looks like for a special case:

$$\zeta(5)=\frac{1}{1}+\frac{1}{2^5}+\frac{1}{3^5}+\cdots=\sum_{n=1}^{+\infty}\frac{1}{n^5} \tag{6.4.192}$$

This may seems to the reader as pure abstract mathematics. However as we will see in the next chapters of this textbook, this function will appear for example during our study of the Black body (Stefan-Boltzmann law), or during our study of the Casimir force in Quantum Physics, but also during our study of the recombination of the Cosmological Microwave Background.

Remarks

R1. It is traditional to denote by s the variable upon which this series depends.

R2. The Riemann zeta function is a special case of the "zeta function" defined by:

$$\zeta_A(s)=\sum_{n=1}^{+\infty}\frac{1}{a_n^s} \tag{6.4.193}$$

where $s=\Re(s)+\mathrm{i}\Im(s)=\sigma+\mathrm{i}t$ but when $a_n=n$.

This series has an interesting property but if we remain within the framework of positive integer powers here is how we can introduce its origin:

$$\frac{1}{x^0}+\frac{1}{x^1}+\cdots+\frac{1}{x^n}+\cdots=u_0\frac{(1-q^{n+1})}{(1-q)}=1\frac{1-\left(\dfrac{1}{x^{n+1}}\right)}{1-\dfrac{1}{x}} \tag{6.4.194}$$

when $n\longrightarrow+\infty$ then we have:

$$\sum_{n=0}^{+\infty}=\frac{1}{x^n}=\frac{1}{x^0}+\frac{1}{x^1}+\cdots+\frac{1}{x^n}+\cdots=\frac{1}{1-\dfrac{1}{x}} \tag{6.4.195}$$

If we put $x=2^s$, we get the sum of the inverse of powers of 2 and similarity with $x=3^s$ such that:

$$\begin{gathered}1+\frac{1}{2^s}+\frac{1}{4^5}+\frac{1}{8^5}+\cdots+\frac{1}{2^{a5}}+\cdots=\frac{1}{1-\dfrac{1}{2^5}}\\ 1+\frac{1}{3^5}+\frac{1}{9^5}+\frac{1}{27^5}+\cdots+\frac{1}{3^{b5}}+\cdots=\frac{1}{1-\dfrac{1}{3^5}}\end{gathered} \tag{6.4.196}$$

If we do the product of these two expressions, we get the sum of the powers of all fractions whose denominator is a product of 2 and 3:

$$\frac{1}{1-\dfrac{1}{2^s}}\cdot\frac{1}{1-\dfrac{1}{3^s}}=1+\frac{1}{2^s}+\frac{1}{3^s}+\frac{1}{4^s}+\frac{1}{5^s}+\frac{1}{6^s}+\cdots+\frac{1}{(2^a3^b)^s}+\cdots \tag{6.4.197}$$

If we take all primes left, we'll get on the right all integers, since every integer is the product of prime numbers according to the fundamental theorem of arithmetic (see section Number Theory page 352), and this is Euler fundamental identity: what we now name "Riemann zeta function" is both a finished product involving prime numbers and an infinite sum of inverse powers of all integers:

$$\zeta(s)=\frac{1}{1-\dfrac{1}{2^s}}\cdot\frac{1}{1-\dfrac{1}{3^s}}\cdot\frac{1}{1-\dfrac{1}{5^s}}=1+\frac{1}{2^s}+\frac{1}{3^s}+\frac{1}{4^s}+\frac{1}{5^s}+\frac{1}{6^s}+\cdots+\frac{1}{(2^a3^b5^c)^s} \tag{6.4.198}$$

In condensed notation, the "Euler's product formula" of the Riemann zeta function is given by:

$$\boxed{\zeta(s)=\sum_{n=1}^{+\infty}\frac{1}{n^s}=\prod_{p\,(\text{prime})}\frac{1}{1-\dfrac{1}{p^s}}} \tag{6.4.199}$$

where p as indicated below the product symbol are the prime numbers.

Remark

The zeta Riemann function is a special case of "Dirichlet series" that are defined by:

$$\mathfrak{D}(s)=\sum_{n=1}^{+\infty}\frac{a_n}{n^s} \tag{6.4.200}$$

When the real part of s is greater than 1, the Dirichlet series converges. On the other hand, the Dirichlet series diverges when the real part of s is less than or equal to 1, so, in particular, the series $1+2+3+4+\ldots$ that results from setting $s=-1$ does not converge!

We now recommend most readers to skip what follows on the Riemann zeta function and return back here once the Fourier series and the Gamma-Euler function presented later in this section are mastered and understood...

We assume in what follows that the Fourier series and Gamma-Euler function are known and mastered and that the Parseval's equality (see below page 1756) was studied (since it is also proved further below).

We will seek to determine first the Riemann zeta function for two positive values (s respectively having values 2 and 4) that will be useful in the valuation of integrals in some sections of the chapter about Mechanics. As the two values are positive the series converge and there won't be too many issues evaluating the Riemann zeta function[11].

Once done we will determine the value of the Riemann zeta function for any negative integer value! And here obviously the series seems to be divergent. In mathematics and theoretical physics for that kind of situation, we use the "zeta function regularization" that is a type of regularization or summability method that assigns finite values to divergent sums or products, and that in particular can be used to define determinants and traces of some self-adjoint operators. The technique is now commonly applied to problems in physics, but has its origins in attempts to give precise meanings to ill-conditioned sums appearing in Number Theory. As we will see just below and during our study of Grandi's series (see page 1691) there are several different summation methods named "zeta function regularization" for defining the sum of a possibly divergent series!

---

[11] For the section of Cosmology we should need also the value of the Riemann zeta function for $s=3$ but however the result give an irrational constant (i.e. the result cannot be written as a fraction p/q where p and q are integers) named the "Apéry's constant". There is a theorem that proves that $\zeta(3)$ is irrational named the "Apéry's theorem".

So let's start!

- To determine the value of $\zeta(2)$, we will express the function:

  $$f(x) \qquad \forall x \in [-\pi, \pi] \tag{6.4.201}$$

  in Fourier series form (see a little further below in this section at page 1727). During our study of Fourier series we will see that there are two traditional ways to define a Fourier series and we have done here the choice of the definition of the most commonly used among physicists and engineers:

  $$f(x) = \frac{a_0}{2} + \sum_{n=1}^{+\infty} (a_n \cos(nx) + b_n \sin(nx)) \tag{6.4.202}$$

  As we prove it in our study of Fourier series, Fourier coefficients a_0, a_n, b_n are obtained by solving:

  $$a_n = \frac{2}{T} \int\limits_{-T/2}^{T/2} f(x) \cos(nx) \mathrm{d}x \qquad b_n = \frac{2}{T} \int\limits_{-T/2}^{T/2} f(x) \sin(nx) \mathrm{d}x \tag{6.4.203}$$

  and using the integration by parts (see section Differential and Integral Calculus page 1532) we have:

  $$\begin{aligned}
  a_0 &= \frac{2}{T} \int\limits_{-T/2}^{+T/2} f(x) \mathrm{d}x = \frac{2}{2\pi} \int\limits_{-\pi}^{\pi} x \mathrm{d}x = \frac{1}{\pi} \int\limits_{-\pi}^{\pi} x \mathrm{d}x = \frac{1}{2\pi} \left(x^2 \big|_{-\pi}^{\pi} \right) = 0 \\
  a_n &= \frac{2}{T} \int\limits_{-T/2}^{+T/2} f(x) \cos(nx) \mathrm{d}x = \frac{2}{2\pi} \int\limits_{-\pi}^{\pi} x \cos(nx) \mathrm{d}x = \frac{1}{\pi} \left(\frac{x \sin(nx)}{n} \bigg|_{-\pi}^{\pi} - \int\limits_{-\pi}^{\pi} \frac{\sin(nx)}{n} \mathrm{d}x \right) \\
  &= \frac{1}{\pi} \left(\frac{\cos(nx)}{n^2} \bigg|_{-\pi}^{\pi} \right) \\
  b_n &= \frac{2}{T} \int\limits_{-T/2}^{+T/2} f(x) \sin(nx) \mathrm{d}x = \frac{2}{2\pi} \int\limits_{-\pi}^{\pi} x \sin(nx) \mathrm{d}x = \frac{1}{\pi} \int\limits_{-\pi}^{\pi} x \sin(nx) \mathrm{d}x \\
  &= \frac{1}{\pi} \left(-\frac{x \cos(nx)}{n} \bigg|_{-\pi}^{\pi} - \int\limits_{-\pi}^{\pi} \frac{\cos(nx)}{n} \mathrm{d}x \right) = \frac{1}{\pi} \left(-\frac{x \cos(nx)}{n} \bigg|_{-\pi}^{\pi} - \frac{\sin(nx)}{n} \bigg|_{-\pi}^{\pi} \right) \\
  &= \frac{1}{\pi} \left(-\frac{x \cos(nx)}{n} \bigg|_{-\pi}^{\pi} \right) = (-1)^{n+1} \frac{2}{n}
  \end{aligned} \tag{6.4.204}$$

It comes then:

$$f(x)=x=\frac{a_0}{2}+\sum_{n=1}^{+\infty}(a_n\cos(nx)+b_n\sin(nx))=\sum_{n=1}^{+\infty}b_n\sin(nx)=\sum_{n=1}^{+\infty}(-1)^{n+1}\frac{2}{n}\sin(nx) \tag{6.4.205}$$

But the Parseval's theorem that we will prove in our study of Fourier series a little bit further below gives us too (depending on the choice of the definition of the Fourier series and associated coefficients, the Parseval's theorem is expressed a little bit differently!):

$$f(x)=x=\sum_{n=1}^{+\infty}(-1)^{n+1}\frac{2}{n}\sin(nx)=\frac{a_0^2}{4}+\frac{1}{2}\sum_{n=1}^{+\infty}(a_n^2+b_n^2) \tag{6.4.206}$$

Therefore we get immediately:

$$f(x)=x=\sum_{n=1}^{+\infty}(-1)^{n+1}\frac{2}{n}\sin(nx)=\frac{1}{2}\sum_{n=1}^{+\infty}b_n^2=\frac{1}{2}\sum_{n=1}^{+\infty}\left((-1)^{n+1}\frac{2}{n}\right)^2=\frac{1}{2}\sum_{n=1}^{+\infty}\frac{4}{n^2}=2\sum_{n=1}^{+\infty}\frac{1}{n^2} \tag{6.4.207}$$

But we will also see during our proof of Parseval's theorem that:

$$\frac{1}{T}\int\limits_{-T/2}^{T/2}f(t)\bar{f}(t)\mathrm{d}t=\frac{a_0^2}{4}+\frac{1}{2}\sum_{n=1}^{+\infty}(a_n^2+b_n^2) \tag{6.4.208}$$

Therefore it comes in our case:

$$\frac{1}{2\pi}\int\limits_{-\pi}^{\pi}x^2\mathrm{d}x=2\sum_{n=1}^{+\infty}\frac{1}{n^2} \tag{6.4.209}$$

Therefore:

$$\frac{1}{2\pi}\frac{x^3}{3}\bigg|_{-\pi}^{\pi}=2\sum_{n=1}^{+\infty}\frac{1}{n^2}\Rightarrow\frac{1}{2\pi}\left[\frac{\pi^3}{3}-\frac{(-\pi)^3}{3}\right]=2\sum_{n=1}^{+\infty}\frac{1}{n^2}\Rightarrow\frac{\pi^3}{3\pi}=2\sum_{n=1}^{+\infty}\frac{1}{n^2} \tag{6.4.210}$$

and finally:

$$\zeta(2)=\sum_{n=1}^{+\infty}\frac{1}{n^2}=\frac{\pi^2}{6} \tag{6.4.211}$$

The problem of finding the value this series converges to is known as the "Basel problem". Leonhard Euler found the solution in 11735 (holocene calendar)...

- To determine the value of $\zeta(4)$, we will do the same, but with the function:

$$f(x)=x^2 \qquad \forall x\in[-\pi,+\pi] \tag{6.4.212}$$

in the form of Fourier series:

$$f(x)=x^2=\frac{a_0}{2}+\sum_{n=1}^{+\infty}(a_n\cos(nx)+b_n\sin(nx)) \tag{6.4.213}$$

For this purpose, we will calculate Fourier coefficients using the integration by parts (see section Differential and Integral Calculus page 1532).

Then we have:

$$\begin{aligned}
a_0&=\frac{2}{T}\int\limits_{-T/2}^{+T/2} f(x)\mathrm{d}x=\frac{2}{2\pi}\int\limits_{-\pi}^{\pi}x^2\mathrm{d}x=\frac{1}{3\pi}\left(x^3|_{-\pi}^{\pi}\right)=\frac{2\pi^2}{3}\\
a_n&=\frac{2}{T}\int\limits_{-T/2}^{+T/2} f(x)\cos(nx)\mathrm{d}x=\frac{2}{2\pi}\int\limits_{-\pi}^{\pi}x^2\cos(nx)\mathrm{d}x=\frac{1}{\pi}\left(\frac{x^2\sin(nx)}{n}\Bigg|_{-\pi}^{\pi}-\int\limits_{-\pi}^{\pi}-2x\frac{\sin(nx)}{n}\mathrm{d}x\right)\\
&=-\frac{2}{\pi n}\left(-x\frac{\cos(nx)}{n}\Bigg|_{-\pi}^{\pi}+\frac{\sin(nx)}{n^2}\Bigg|_{-\pi}^{\pi}\right)=\frac{2}{\pi n}x\frac{\cos(nx)}{n}\Bigg|_{-\pi}^{\pi}=(-1)^{n+1}\frac{4}{n^2}\\
b_n&=\frac{2}{T}\int\limits_{-T/2}^{+T/2} f(x)\sin(nx)\mathrm{d}x=\frac{2}{2\pi}\int\limits_{-\pi}^{\pi}x^2\sin(nx)\mathrm{d}x=\frac{1}{\pi}\int\limits_{-\pi}^{\pi}x^2\sin(nx)\mathrm{d}x\\
&=\frac{1}{\pi}\left(-x^2\frac{\cos(nx)}{n}\Bigg|_{-\pi}^{\pi}+\int\limits_{-\pi}^{\pi}2x\frac{\cos(nx)}{n}\mathrm{d}x\right)=\frac{2}{\pi n}\left(x\frac{\sin(nx)}{n}\Bigg|_{-\pi}^{\pi}+\int\limits_{-T/2}^{+T/2}\frac{\sin(nx)}{x}\mathrm{d}x\right)\\
&=-\frac{2}{\pi n}\left(\int\limits_{-T/2}^{+T/2}\frac{\sin(nx)}{x}\mathrm{d}x\right)=-\frac{2}{\pi n^2}\left(\frac{-\cos(nx)}{n}\Bigg|_{-\pi}^{\pi}\right)=0
\end{aligned}\tag{6.4.214}$$

Therefore we have:

$$f(x)=x^2=\frac{a_0}{2}+\sum_{n=1}^{+\infty}a_n\cos(nx)=\frac{\pi^2}{3}+\sum_{n=1}^{+\infty}(-1)^{n+1}\frac{4}{n^2}\cos(nx)\tag{6.4.215}$$

But the Parseval's theorem that we will prove below gives us also:

$$f(x)=x^2=\frac{\pi^2}{3}+\sum_{n=1}^{+\infty}(-1)^{n+1}\frac{4}{n^2}\cos(nx)=\frac{a_0^2}{4}+\frac{1}{2}\sum_{n=1}^{+\infty}(a_n^2+b_n^2)\tag{6.4.216}$$

It then comes immediately:

$$f(x)=x^2=\frac{\pi^2}{3}+\sum_{n=1}^{+\infty}(-1)^{n+1}\frac{4}{n^2}\cos(nx)=\frac{\pi^2}{9}+8\sum_{n=1}^{+\infty}\frac{1}{n^4}=\frac{\pi^4}{9}+8\zeta(4)\tag{6.4.217}$$

But we will see also see later below during our Parseval's theorem proof that:

$$\frac{1}{T}\int\limits_{-T/2}^{+T/2}f(t)\bar{f}(t)\mathrm{d}t=\frac{a_0^2}{4}+\frac{1}{2}\sum_{n=1}^{+\infty}(a_n^2+b_n^2)\tag{6.4.218}$$

Then it comes in our case:

$$\frac{1}{2\pi}\int\limits_{-\pi}^{+\pi}x^4\mathrm{d}x=\frac{\pi^4}{9}+8\zeta(4)\tag{6.4.219}$$

Therefore:

$$\frac{1}{2\pi}\frac{x^5}{5}\bigg|_{-\pi}^{\pi}=\frac{\pi^4}{9}+8\zeta(4)\Rightarrow\frac{1}{2\pi}\left[\frac{\pi^5}{5}-\frac{(-\pi)^5}{5}\right]=\frac{\pi^4}{9}+8\zeta(4) \tag{6.4.220}$$

That is to say:

$$\frac{\pi^4}{5}-\frac{\pi^4}{9}=8\zeta(4)\Rightarrow\frac{9\pi^4}{45}-\frac{5\pi^4}{45}=8\zeta(4)\Rightarrow\frac{4\pi^4}{45}=8\zeta(4) \tag{6.4.221}$$

Finally:

$$\zeta(4)=\sum_{n=1}^{+\infty}\frac{1}{n^4}=\frac{\pi^4}{90} \tag{6.4.222}$$

- Now let us focus on the first divergent case $\zeta(-1)$.

  There are a few ways to prove that:

$$\zeta(s=-1)=\sum_{n=1}^{+\infty}\frac{1}{n^{-1}}=\sum_{n=1}^{+\infty}n=1+2+3+\ldots=-\frac{1}{12} \tag{6.4.223}$$

  One method, along the lines of Euler's reasoning, uses the relationship between the Riemann zeta function and what we define as the "Dirichlet eta function" or "alternating zeta function[12]":

$$\eta(s)=\sum_{n=1}^{+\infty}\frac{(-1)^{n-1}}{n^s}=\frac{1}{1^s}-\frac{1}{2^s}+\frac{1}{3^s}-\frac{1}{4^s}+\cdots \tag{6.4.224}$$

  This Dirichlet series is the alternating sum corresponding to the Dirichlet series expansion of the Riemann zeta function, $\zeta(s)$.

  Where both Dirichlet series converge, one has the identities:

$$\begin{aligned}\zeta(s)&=1^{-s}+2^{-s}+3^{-s}+4^{-s}+5^{-s}+6^{-s}+\cdots\\ 2\cdot2^{-s}\zeta(s)&=\phantom{1^{-s}+{}}2\cdot2^{-s}\phantom{+3^{-s}}+2\cdot4^{-s}\phantom{+5^{-s}}+2\cdot6^{-s}+\cdots\\ \left(1-2^{1-s}\right)\zeta(s)&=1^{-s}-2^{-s}+3^{-s}-4^{-s}+5^{-s}-6^{-s}+\cdots=\eta(s)\end{aligned} \tag{6.4.225}$$

  The identity$(1-2^{1-s})\zeta(s)=\eta(s)$ continues to hold when both functions are extended by analytic continuation to include values of s for which the above series diverge. Substituting $s=-1$, one gets $-3\zeta(-1)=\eta(-1)$. Now, computing $\eta(-1)$ is an easier task, as the eta function is equal to the Abel sum of its defining series, which is a one-sided limit:

$$-3\zeta(-1)=\eta(-1)=\lim_{x\to1^-}\left(1-2x+3x^2-4x^3+\cdots\right)=\lim_{x\to1^-}\frac{1}{(1+x)^2}=\frac{1}{4} \tag{6.4.226}$$

  Dividing both sides by -3, one gets:

$$\zeta(-1)=-\frac{1}{12} \tag{6.4.227}$$

---

[12] Yes there are a lot of different kind (variations) of the zeta function. You can find a list of the 34 most important of them on Wikipedia.

The first example in which zeta function regularization is available appears in the Casimir effect (see section page Wave Quantum Physics page 4825), which is in a flat space the bulk contributions of the quantum field in three space dimensions. In this case we must calculate the value of Riemann zeta function at -3:

$$\zeta(s=-3)=\sum_{n=1}^{+\infty}\frac{1}{n^s}=\sum_{n=1}^{+\infty}n^3=\zeta(-3)=1+8+27+64+\ldots \qquad (6.4.228)$$

which diverges explicitly. However, it can be analytically continued to $s=-3$ where hopefully there is no pole (see section Complex Analysis page 2440), thus giving a finite value to the expression.

Remark

The reason why we see $\zeta(-1)$ in many YouTube videos and physics textbook, rather than $\zeta(-3)$ is that when you imagine the Casimir effect as happening in one dimension (along a line rather than in 3D), the energy density you calculate is $\zeta(-1)$ rather than $\zeta(-3)$.

Let us see another and more rigorous way to prove that $\zeta(-1)=-\dfrac{1}{12}$.

We need a result first:

$$\begin{aligned}\int_0^{+\infty}\frac{xt^{x-1}}{e^t+1}\mathrm{d}t&=\int_0^{+\infty}\frac{xt^{x-1}}{1+e^{-t}}e^{-t}\mathrm{d}t=x\sum_{k=1}^{+\infty}(-1)^{k-1}\int_0^{+\infty}t^{x-1}e^{-kt}\mathrm{d}t\\&=x\sum_{k=1}^{+\infty}(-1)^{k-1}k^{-x}\int_0^{+\infty}t^{x-1}e^{-t}\,\mathrm{d}t=x\eta(x)\Gamma(x)=(1-2^{1-x})\zeta(x)\Gamma(x+1)\end{aligned} \qquad (6.4.229)$$

Let us now multiply the above result by $x+1$, then integrate by parts twice, to we get:

$$\begin{aligned}(1-2^{1-x})\zeta(x)\Gamma(x+2)&=\int_0^{+\infty}\frac{(x+1)xt^{x-1}}{e^t+1}\mathrm{d}t=\int_0^{+\infty}\frac{(x+1)t^xe^t}{(e^t+1)^2}\mathrm{d}t\\&=\int_0^{+\infty}\frac{t^{x+1}(e^{2t}-e^t)}{(e^t+1)^3}\mathrm{d}t\end{aligned} \qquad (6.4.230)$$

Now we can plug in $x=-1$ into the above result to get:

$$\begin{aligned}(1-2^2)\zeta(-1)\Gamma(1)&=\int_0^{+\infty}\frac{e^{2t}-e^t}{(e^t+1)^3}\mathrm{d}t=\int_1^{+\infty}\frac{u-1}{(u+1)^3}\mathrm{d}u\\&=\int_1^{+\infty}\left(\frac{1}{(u+1)^2}-\frac{2}{(u+1)^3}\right)\mathrm{d}u=\frac{1}{4}\end{aligned} \qquad (6.4.231)$$

Since $(1-2^2)\Gamma(1)=-3$ the above result can be rearrange to:

$$\zeta(-1)=-\frac{1}{12} \qquad (6.4.232)$$

We will see a third method just below to get that same result and even a fourth one during our study later of Grandi's series (see page 1691).

- Let us prove now an elegant and fast way to determine the value of the zeta function for all negative integers (there are other methods but involving heavy and boring mathematics for most engineers).

  For that purpose we will first need to prove the following lemma:

  Lemma 6.21.1. The Riemann zeta function in the critical strip is given by:

  $$\zeta(s)=\frac{1}{s-1}+1-s\int_1^{+\infty}\frac{x-[x]}{x^{s+1}}\mathrm{d}x \tag{6.4.233}$$

  where $[x]$ denotes the integer part of x.

  For the proof we need first for $\Re(s)>1$:

  $$\begin{aligned}\zeta(s)&=\sum_{n=1}^{+\infty}\frac{1}{n^s}=\sum_{n=1}^{1}\frac{n}{n^s}+\sum_{n=2}^{\infty}\frac{1}{n^s}=\sum_{n=1}^{1}\frac{n}{n^s}+\sum_{n=2}^{+\infty}\frac{n-(n-1)}{n^s}\\&=\sum_{n=1}^{1}\frac{n}{n^s}+\sum_{n=2}^{+\infty}\frac{n}{n^s}-\sum_{n=2}^{+\infty}\frac{n-1}{n^s}=\sum_{n=1}^{+\infty}\frac{n}{n^s}-\sum_{n=2}^{+\infty}\frac{n-1}{n^s}\\&=\sum_{n=1}^{+\infty}\frac{n}{n^s}-\sum_{n=1}^{+\infty}\frac{n}{(n+1)^s}=\sum_{n=1}^{+\infty}n\left[\frac{1}{n^s}-\frac{1}{(n+1)^s}\right]\\&=s\sum_{n=1}^{+\infty}n\int_n^{n+1}x^{-s-1}\mathrm{d}x\end{aligned} \tag{6.4.234}$$

  Since $[x]=n$ for any x in the interval $[n,n+1]$, we have:

  $$\begin{aligned}&=s\sum_{n=1}^{+\infty}\int_n^{n+1}[x]x^{-s-1}\mathrm{d}x=s\int_1^{+\infty}[x]x^{-s-1}\mathrm{d}x\\&=s\left[\int_1^{+\infty}x^{-s}\mathrm{d}x\right]-s\int_1^{+\infty}\{x\}x^{-s-1}\mathrm{d}x\quad\text{(because by definition:}[x]=x-\{x\})\\&=s\left[\left.\frac{x^{-s+1}}{-s+1}\right|_1^{+\infty}\right]-s\int_1^{+\infty}\{x\}x^{-s-1}\mathrm{d}x\end{aligned} \tag{6.4.235}$$

  allowing the following simplification:

  $$\zeta(s)=\frac{s}{s-1}-s\int_1^{+\infty}\frac{\{x\}}{x^{s+1}}\mathrm{d}x\quad\Re(s)>1 \tag{6.4.236}$$

Because $0 \leq \{x\} < 1$, the last integral converges and is holomorphic (see section Complex Analysis page 2399) on $\Re(s) > 0$. But that means the full equation is meromorphic (see section Complex Analysis page 2402) on $\Re(s) > 0$, and thus provides an analytic continuation of $\zeta(s)$ on the half plane $\Re(s) > 0$. The $s/(s-1)$ term gives a simple pole at $s = 1$ with residue 1 (see section Complex Analysis page 2440).

Injecting again back $[x] = x - \{x\} \Leftrightarrow \{x\} = x - [x]$, we get:

$$\zeta(s) = \frac{s}{s-1} - s\int\limits_1^{+\infty} \frac{x-[x]}{x^{s+1}}\mathrm{d}x = \frac{1}{s-1} + 1 - s\int\limits_1^{+\infty} \frac{x-[x]}{x^{s+1}}\mathrm{d}x \quad \Re(s) > 1 \tag{6.4.237}$$

That finish the proof of the lemma.

Now remember that we have proved just earlier above that:

$$\zeta(s) = s\int\limits_1^{+\infty} \frac{[x]}{x^{s+1}}\mathrm{d}x = \frac{s}{s-1} - s\int\limits_1^{+\infty} \frac{\{x\}}{x^{s+1}}\mathrm{d}x \tag{6.4.238}$$

where $\{x\}$ denotes the fractional part of x. This gives the analytic continuation of $\zeta(s)$ for $\Re(s) > 0$. We can now proceed inductively. Writing:

$$\int\limits_1^{+\infty} \frac{\{x\}}{x^{s+1}}\mathrm{d}x = \sum_{n=1}^{\infty} \int\limits_n^{n-1} \frac{x-n}{x^{s+1}}\mathrm{d}x = \sum_{n=1}^{+\infty} \int\limits_0^1 \frac{u\mathrm{d}u}{(u+n)^{n+1}} \tag{6.4.239}$$

and integrating the last integral by parts, we get:

$$\sum_{n=1}^{+\infty} \int\limits_0^1 \frac{u\mathrm{d}u}{(u+n)^{s+1}} = \frac{1}{2}\sum_{n=1}^{+\infty} \frac{1}{(n+1)^{s+1}} + \frac{(s+1)}{2}\int\limits_1^{+\infty} \frac{\{x\}^2}{x^{s+2}}\mathrm{d}x = \frac{1}{2}(\zeta(s+1)-1) + \frac{(s+1)}{2}\int\limits_1^{+\infty} \frac{\{x\}^2}{x^{s+2}}\mathrm{d}x \tag{6.4.240}$$

and the latter integral converges for $\Re(s) > -1$. That is:

$$\zeta(s) = \frac{s}{s-1} - \frac{s}{2}(\zeta(s+1)-1) - \frac{s(s+1)}{2}\int_1^x \frac{\{x\}^2}{x^{s+2}}\mathrm{d}x \tag{6.4.241}$$

Thus, inductively, we deduce:

$$\begin{aligned}\zeta(s) = 1 + \frac{1}{s-1} - \sum_{r=1}^{m} \frac{s(s+1)\cdots(s+r-1)}{(r+1)!}(\zeta(s+r)-1) \\ - \frac{s(s+1)\cdots(s+m)}{(m+1)!}\sum_{n=1}^{\infty}\int_0^1 \frac{u^{m+1}\mathrm{d}u}{(u+n)^{s+m+1}}\end{aligned} \tag{6.4.242}$$

and the infinite sum on the right hand side converges for $\Re(s) > -m$.

If in the relation above we put $s = 1-m$, and note that for $r = m$ that $\zeta(s+m)$ has a simple pole at $s = 1-m$, we get the recurrence:

$$\zeta(1-m) = 1 - \frac{1}{m} + \frac{(-1)^m}{m(m+1)} - \sum_{r=1}^{m-1}(-1)^r \begin{pmatrix} m-1 \\ r \end{pmatrix} \frac{1}{r+1}(\zeta(1-m+r)-1) \tag{6.4.243}$$

The first few values at non positive integers are immediate and given by:

$$\zeta(0) = -\frac{1}{2}, \zeta(-1) = -\frac{1}{12}, \zeta(-2) = 0, \zeta(-3) = \frac{1}{120} \tag{6.4.244}$$

And that's it!

6.4.2.4 Telescoping Series

A "telescoping series" is a series in which most of terms cancel in each of the partial sums, leaving only some of the first terms and some of the last terms.

For example, the series:

$$\sum_{n=1}^{+\infty}\frac{1}{n(n+1)} \tag{6.4.245}$$

simplifies as:

$$\begin{aligned}\sum_{n=1}^{+\infty}\frac{1}{n(n+1)}&=\sum_{n=1}^{+\infty}\left(\frac{1}{n}-\frac{1}{n+1}\right)\\&=\lim_{N\rightarrow+\infty}\sum_{n=1}^{N}\left(\frac{1}{n}-\frac{1}{n+1}\right)\\&=\lim_{N\rightarrow+\infty}\left[\left(1-\frac{1}{2}\right)+\left(\frac{1}{2}-\frac{1}{3}\right)+\cdots+\left(\frac{1}{N}-\frac{1}{N+1}\right)\right]\\&=\lim_{N\rightarrow+\infty}\left[1+\left(-\frac{1}{2}+\frac{1}{2}\right)+\left(-\frac{1}{3}+\frac{1}{3}\right)+\cdots+\left(-\frac{1}{N}+\frac{1}{N}\right)-\frac{1}{N+1}\right]\\&=\lim_{N\rightarrow+\infty}\left[1-\frac{1}{N+1}\right]=1.\end{aligned} \tag{6.4.246}$$

We will encounter such a series for business purposes (management) in our study of Queuing Theory in the section of Quantitative Management!!!

6.4.2.5 Grandi's Series

The "Grandi's series" (after Italian mathematician, philosopher, and priest Guido Grandi, who gave a memorable treatment of the series in 11703 according to holocene calendar) is defined as the following arithmetic series:

$$\sum_{n=0}^{+\infty}(-1)^n=1-1+1-1+\cdots \tag{6.4.247}$$

It is a very famous series in mathematics and physics because:

- It highlights in a very simple way the fact (see below) that it is dangerous to manipulate infinite series
- Its result seems completes non-intuitive but in fact it opens the door to a more general definition of what is a "sum"
- It is a beautiful example of a series that seems useless and purely mathematics but that has in fact important application in quantum physics (Casimir Effect as seen in the section of Wave Quantum Physics at page 4825) and String Theory (number of dimensions as seen in the section of String Theory).

and this is why we dedicate to it a special subsection in this textbook!

It seems quite obvious at a first glance that it is a divergent series, meaning that it lacks a sum in the usual sense (the sequence of partial sums of Grandi's series clearly does not approach any number). But the other hand, its Cesàro sum is $1/2$!!?? So what the hell is a Cesàro sum?

Definition (#6.4.222): In mathematical analysis a "Cesàro sum assigns values to some infinite sums that are not convergent in the usual sense. The Cesàro sum is defined as the limit of the arithmetic mean of the partial sums of the series.

Let $\{a_n\}$ be a sequence, and let:

$$s_k = a_1 + \cdots + a_k = \sum_{n=1}^{k} a_n \tag{6.4.248}$$

be the kth partial sum of the series:

$$\sum_{n=1}^{+\infty} a_n \tag{6.4.249}$$

The series $\sum_{n=1}^{+\infty} a_n$ is say to be "Cesàro summable", with Cesàro sum $S \in \mathbb{R}$, if the average value of its partial sums s_k tends to S:

$$\lim_{n\rightarrow\infty} \frac{1}{n}\sum_{k=1}^{n} s_k = S \tag{6.4.250}$$

In other words, the Cesàro sum of an infinite series is the limit of the arithmetic mean (average) of the first n partial sums of the series, as n goes to infinity. If a series is convergent, then it is Cesàro summable and its Cesàro sum is the usual sum. For any convergent sequence, the corresponding series is Cesàro summable and the limit of the sequence coincides with the Cesàro sum.

One obvious method to attack the Grandi's series:

$$\sum_{n=0}^{+\infty} (-1)^n = 1-1+1-1+1-1+1-1+\ldots \tag{6.4.251}$$

is to treat it like a telescoping series and perform the subtractions in place:

$$\sum_{n=0}^{+\infty} (-1)^n = (1-1)+(1-1)+(1-1)+\ldots = 0+0+0+\ldots = 0 \tag{6.4.252}$$

On the other hand, a similar bracketing procedure leads to the apparently contradictory result:

$$\sum_{n=0}^{+\infty} (-1)^n = 1+(-1+1)+(-1+1)+(-1+1)+\ldots = 1+0+0+0+\ldots = 1 \tag{6.4.253}$$

Thus, by applying parentheses to Grandi's series in different ways, one can get either 0 or 1 as a "value". It can be shown that it is not valid to perform many seemingly innocuous operations on a series, such as reordering individual terms, unless the series is absolutely convergent. Otherwise these operations can alter the result of summation.

Treating Grandi's series as a divergent geometric series we may use the same algebraic methods that evaluate convergent geometric series to get a third value:

$$S = 1 - 1 + 1 - 1 + \ldots \tag{6.4.254}$$

so:

$$1 - S = 1 - (1 - 1 + 1 - 1 + \ldots) = 1 - 1 + 1 - 1 + \ldots = S \tag{6.4.255}$$

Therefore:

$$1 - S = S \tag{6.4.256}$$

Finally:

$$S = \sum_{n=0}^{+\infty} (-1)^n = \frac{1}{2} \tag{6.4.257}$$

The same conclusion results from calculating $-S$, subtracting the result from S, and solving $2S = 1$.

The above manipulations do not consider what the sum of a series actually means. Still, to the extent that it is important to be able to bracket series at will, and that it is more important to be able to perform arithmetic with them, one can arrive at two conclusions:

- The series $1 - 1 + 1 - 1 + \ldots$ has no sum
- ...but its sum should be $1/2$ (see further below)

In fact, both of these statements can be made precise and formally proven, but only using well-defined mathematical concepts that arise in the 119th century (holocene calendar). After the late 117th century introduction of calculus in Europe, but before the advent of modern rigour, the tension between these answers fuelled what has been characterized as an "endless" and "violent" dispute between mathematicians. The funniest is that the violent discussions still continue today... a YouTube video on this subject has more than $5,000$ comments... and a blog post more than 200 comments and a forum thread more than 600... (and by the way the comments show the worst of what the human behaviour and respect can be...). So this remains quite a hot topic!

Let us also recall that at the beginning of our study of geometric series we have proved that:

$$S_n = u_0 q^0 + u_0 q^1 + u_0 q^2 + u_0 q^3 + \ldots = u_0 \frac{1 - q^{n+1}}{1 - q} \tag{6.4.258}$$

Therefore if $u_0 = 1$ this reduce to:

$$S_n = 1 + q^1 + q^2 + q^3 + \ldots = \frac{1 - q^{n+1}}{1 - q} \tag{6.4.259}$$

where as n goes to infinity, the absolute value of $|q|$ must be less than one for the series to converge!

Now notice that if $q=-1$ we fall back on Grandi's series and therefore that latter is a special case of the geometric series $1+q^1+q^2+q^3+\ldots$ and then we would perhaps write a bit too quick:

$$S_n=1+(-1)^1+(-1)^2+(-1)^3+\ldots=1-1+1-1+\ldots=\frac{1-(-1)^{n+1}}{1-(-1)} \tag{6.4.260}$$

But as we have just mention it, we are not authorized to write the latter fraction if $q=\pm 1$ otherwise the series diverge excepted... if we work in a special axiomatic framework. This is almost as chocking as when imaginary numbers were introduced in the 116th century (holocene calendar), as well as negative numbers that were poorly understood and regarded by some as fictitious or useless, much as zero once was...

Remark

There have been very interesting studies about the reaction of high-school level students to Grandi's series presentation. The reactions and analysis are very interesting and I can personally only recommend every teacher to introduce this series in classes but without giving the result in a first time!

But there is a "one more thing"... We will now calculate a sum, thinking it really gives infinity as a result:

$$B=\sum_{n=1}^{+\infty} n=1+2+3+4+5+\ldots \tag{6.4.261}$$

To do this, let's do another trick of mathematical magician:

$$B=1+2+3+4+\ldots \tag{6.4.262}$$

Therefore:

$$4B=4+8+\ldots \tag{6.4.263}$$

So we can compute:

$$-3B=B-4B=1-2+3-4+\ldots=\sum_{n=1}^{+\infty}(-1)^{n+1}n=C \tag{6.4.264}$$

Our first concrete result, squared, can be rewritten as follows:

$$S^2=(-1+1-1+1-1+\ldots)^2 \tag{6.4.265}$$

Or well:

$$S^2=S\cdot S \tag{6.4.266}$$

Explicitly:

$$(-1+1-1+1-1+\ldots)$$

$$
\begin{array}{r}
\underline{\times(-1+1-1+1-1+\ldots)} \\
=1-1+1-1+1-1+1-1+\cdots \\
-1+1-1+1-1+1-1+\cdots \\
+1-1+1-1+1-1+\cdots \\
\ldots
\end{array}
$$

Summing each column we see that we fall back on:

$$1-2+3-4+5-6+\ldots=C \tag{6.4.267}$$

Therefore:

$$S^2=C \tag{6.4.268}$$

But as $S=1/2$, then:

$$S^2=C=(1/2)^2=\frac{1}{4} \tag{6.4.269}$$

Therefore:

$$B=-\frac{C}{3}=\frac{1}{12} \tag{6.4.270}$$

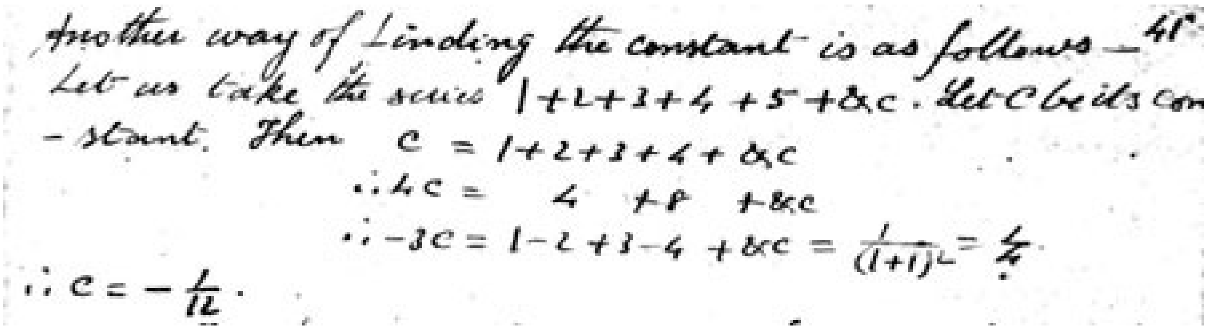
That is (to freak out a last time), we have shown that:

$$\boxed{\sum_{n=1}^{+\infty} n=1+2+3+4+5+\ldots=-\frac{1}{12}} \tag{6.4.271}$$

Quite interesting! But don't forget that we get this result in a special framework that we have build and that generalize the conventional sum that we learn at high-school. It seems that all generalizations of the conventional sum (as mathematicians do definitions with coffee as we know...) lead to the same result of $-1/12$...

It is also quite chocking that as the set of natural numbers $\mathbb{N}$ is closed under addition that the sum of all naturals is $-1/12$...

Furthermore, all this stuff is not new! It was known by many people long time before (few hundred years) and it was especially Srinivasa Ramanujan and later Godfrey Harold Hardy in a book titled *Divergent Series* that formalized the subject in a more elegant and technical way than the one you have here above.



Another way of finding the constant is as follows — 41
Let us take the series $1+2+3+4+5+\&c$. Let C be its con
-stant. Then $C = 1+2+3+4+\&c$

$\therefore 4C = \quad 4 \quad +8 \quad +\&c$

$\therefore -3C = 1-2+3-4+\&c = \frac{1}{(1+1)^2} = \frac{1}{4}$

$\therefore C = -\frac{1}{12}$.

Figure 6.69: Piece of Ramanujan's first notebook about $-1/12$

6.4.2.6 Taylor and Maclaurin Series

Taylor and Maclaurin series provide a convenient and powerful tool to simplify theoretical models and computer calculations (fluid modelling or fields in space). They are used heavily in all fields of physics but they are also found in the industry including engineering (design of experiments, numerical methods, quality management), statistics (integral approximations), finance (stochastic processes), complex analysis... We strongly advise the reader to consult carefully the developments that follow. There is also a serious funny quote on the subject:

quote
6.4:1696.#57

> ❝ *Everything is just some form of first order Taylor Expansion!* ❞

To start, consider a polynomial (with one variable/univariate):

$$P_n(x)=\sum_{k=0}^{n} a_k x^k \qquad (6.4.272)$$

We trivially have for this latter:

$$P(0)=a_0 \qquad (6.4.273)$$

Given now the derivative of the polynomial $P(x)$:

$$P(x)=\sum_{k=1}^{n} k a_k x^{k-1} \qquad (6.4.274)$$

Therefore:

$$P'(0)=1\cdot a_1 \qquad (6.4.275)$$

and so on with $P''(x), P'''(x), \ldots$ such that:

$$\begin{aligned} P''(x)&=\sum_{k=2}^{n} k(k-1)a_k x^{k-2} \Rightarrow P''(0)=1\cdot 2\cdot a_2\\ P'''(x)&=\sum_{k=3}^{n} k(k-1)(k-2)a_k x^{k-3} \Rightarrow P''(0)=1\cdot 2\cdot 3\cdot a_3\\ &\ldots\\ \frac{\mathrm{d}^n P(x)}{\mathrm{d}x^n}&=\sum_{k=n}^{n} k(k-1)\cdots(k-(n-1))a_k x^{k-n} \Rightarrow P''(0)=n!\cdot a_n \end{aligned} \qquad (6.4.276)$$

Then:

$$a_0=P(0), a_1=\frac{1}{1!}\frac{\mathrm{d}P(0)}{\mathrm{d}x}, a_2=\frac{1}{2!}\frac{\mathrm{d}^2P(0)}{\mathrm{d}x^2}, a3=\frac{1}{3!}\frac{\mathrm{d}^3P(0)}{\mathrm{d}x^3}, \cdots, a_n=\frac{1}{n!}\frac{\mathrm{d}^nP(0)}{\mathrm{d}x^n} \qquad (6.4.277)$$

Therefore:

$$\boxed{P_n(x)=\sum_{k=0}^{n}\frac{1}{k!}\frac{\mathrm{d}^kP(0)}{\mathrm{d}x^k}x^k} \qquad (6.4.278)$$

relation that we name "limited Maclaurin series" or simply "Maclaurin series" of order $k+1$.

Remark

In practice, as we will see in many other sections of this textbook, we often use limited developments of order 1 (also named "affine approximations", or "affine tangent approximations"), which can facilitate the calculations, when we do not expect too much precision in the final result.

Now by applying the same reasoning but by centring the value of the polynomial on $x = x_0$, we have:

$$P_n(x) = \sum_{k=0}^{n} a_k (x - x_0)^k \tag{6.4.279}$$

and so the previous development becomes more general:

$$\boxed{P_n(x) = \sum_{k=0}^{n} \frac{1}{k!} \frac{\mathrm{d}^k P(x_0)}{\mathrm{d}x^k} (x - x_0)^k} \tag{6.4.280}$$

which is no other than the general expression of a polynomial expression in a form named "limited Taylor series" of order $k+1$. This function can be assimilate to a polynomial as n is finite. But if n is infinite, as we shall see later, this series converges to the function we are seeking the representation in the form of a sum of terms.

Thus, some functions $f(x)$ of class $\mathcal{C}^n$ that can be approximated by a polynomial $P(x)$ (a sum of powers in other words...) centered on the value x_0 can be expressed as:

$$f(x) = \sum_{k=0}^{n} \frac{1}{k!} \frac{\mathrm{d}^k f(x_0)}{\mathrm{d}x^k} (x - x_0)^k = \sum_{k=0}^{n} \frac{\mathrm{d}^{(k)} f(x_0)}{k!} (x - x_0)^k \tag{6.4.281}$$

Result often referred to as "Taylor's theorem".

But this last relation is not correct for all functions that can not be expressed as a polynomial. Therefore we say that the series is not convergent for them. We will see an example later below.

The latter relation is sometimes also written ... more conventionally:

$$f(x) = \sum_{k=0}^{n} \frac{\mathrm{D}^k f(x_0)}{k!} (x - x_0)^k \tag{6.4.282}$$

In finance (and not only!), we will often use the following rearrangement:

$$\boxed{\begin{gathered} f(x) = f(x_0) + \sum_{k=1}^{n} \frac{\mathrm{D}^k f(x_0)}{k!} (x - x_0)^k \\ \Rightarrow f(x) - f(x_0) = \sum_{k=1}^{n} \frac{\mathrm{D}^k f(x_0)}{k!} (x - x_0)^k \\ \lim_{x - x_0 \to \mathrm{d}x} \mathrm{d}x = \sum_{k=1}^{n} \frac{\mathrm{D}^k f(x_0)}{k!} \mathrm{d}x^k \end{gathered}} \tag{6.4.283}$$

Let us return briefly to the approximation of $f(x)$ near and centered in x_0:

$$f(x)=\sum_{k=0}^{n}\frac{1}{k!}\frac{\mathrm{d}^k f(x_0)}{\mathrm{d}x^k}(x-x_0)^k=\sum_{k=0}^{n}\frac{1}{k!}\frac{\mathrm{d}^k f(x_0)}{\mathrm{d}x^k}(\Delta x)^k \tag{6.4.284}$$

Another very common notation in physics and financial engineering of the above relation is:

$$\widetilde{f}(x)=f(x_0)+\sum_{k=1}^{n}\frac{1}{k!}\frac{\mathrm{d}^k f(x_0)}{\mathrm{d}x^k}(\Delta x)^k \Leftrightarrow \mathrm{d}f(x)=\sum_{k=1}^{n}\frac{1}{k!}\frac{\mathrm{d}^k f(x_0)}{\mathrm{d}x^k}(\Delta x)^k \tag{6.4.285}$$

Some people do not like using this formulation because we have the risk to forget that the approximation for a few terms is only good as long as we are not too far from x_0 with x. This is why it often happens that we write:

$$x=x_0+\Delta h=x_0+h \tag{6.4.286}$$

with x_0 fixed and a h variable but small (!) and so it then comes a common form of notation of Taylor series(!):

$$f(x_0+h)=\sum_{k=0}^{n}\frac{1}{k!}\frac{\mathrm{d}^k f(x_0)}{\mathrm{d}x^k}(x_0+h-x_0)^k=\sum_{k=0}^{n}\frac{h^k}{k!}\frac{\mathrm{d}^k f}{\mathrm{d}x^k}(x_0) \tag{6.4.287}$$

Let's see an application example with Maclaurin series (with x_0 being zero) of the function $\sin(x)$ and Maple™ 4.00b:

```
>p[n](x) = sum((D@@i)(f)(a)/i!*(x-a)^i,i=0..n);
>p11:= taylor(sin(x),x=0,12);
>p11:= convert(p11,polynom);
>with(plots):
>tays:= plots[display](sinplot):
for i from 1 by 2 to 11 do
tpl:= convert(taylor(sin(x), x=0,i),polynom):
tays:= tays,plots[display]([sinplot,plot(tpl,x=-Pi..2*Pi,y=-2..2,
color=black,title=convert(tpl,string))]) od:
>plots[display]([tays],view=[-Pi..2*Pi,-2..2]);
```

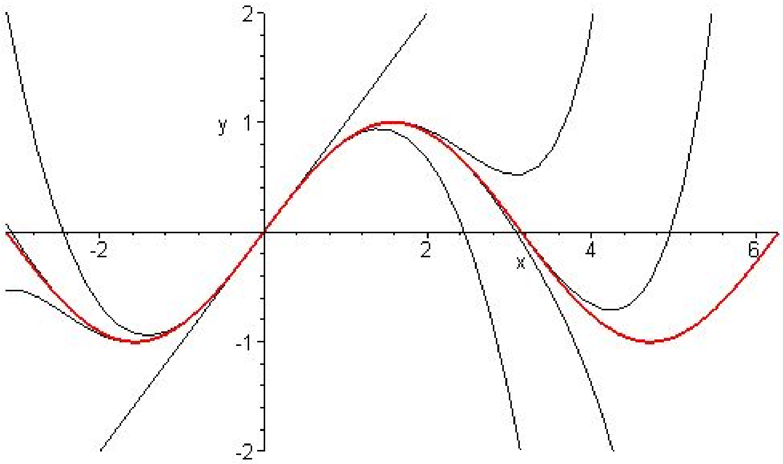


Figure 6.70: **Approximation of the sine function by a Maclaurin development in Maple™ 4.00b**

Or in better quality with TikZ:

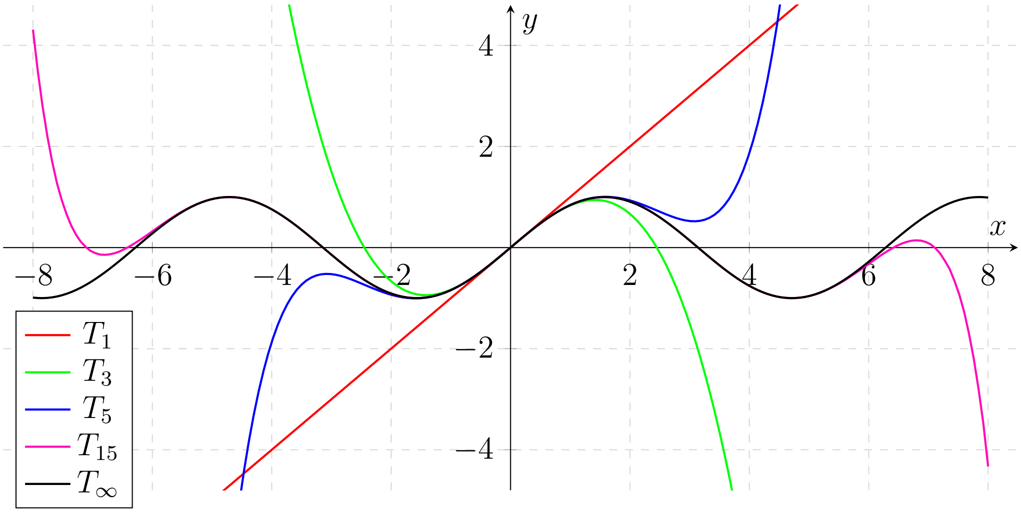


Figure 6.71: **Approximation of the sine function by a Maclaurin development with TikZ**

We see well in this example that the Maclaurin series only allows to approach a function at a point with a limited number of points. But more terms we take (put 100 terms in the Maple™ code above) more the validity is big on the whole domain of definition of the function. In fact it is possible to prove that the function $sin(x)$ is exactly expressible in Maclaurin series when the number of terms is infinite. We say then that its "rest" is zero.

But this is not true for all functions! For example the function:

$$f(x) = \frac{1}{1-x^2} \tag{6.4.288}$$

```
>p[n](x) = sum((D@@i)(f)(a)/i!*(x-a)^i,i=0..n);
>p10:= taylor(1/(1-x^2),x=0,10);
>p10:= convert(p10,polynom);
>with(plots):
>tays:= plots[display](xplot):
for i from 1 by 2 to 10 do
tpl:= convert(taylor(1/(1-x^2), x=0,i),polynom):
tays:= tays,plots[display]([xplot,plot(tpl,x=-2..2,y=-2..2,
color=black,title=convert(tpl,string))]) od:
>plots[display]([tays],view=[-2..2,-2..2]);
```

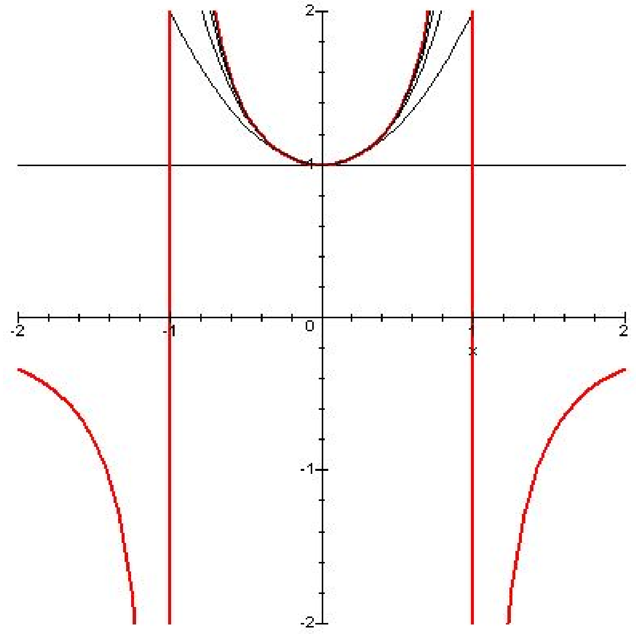


Figure 6.72: **Example of non-convergent Maclaurin series in Maple™ 4.00b**

We see above that regardless of the number of terms that we take the Maclaurin series converges only in one area of definition between $]-1,1[$. This interval is named the "radius of convergence" and it determination (the singularity) is crucial in many areas of engineering, physics and analysis. We will return in more detail on this example in the section of Complex Analysis.

But we can shift the Maclaurin series of the previous function to approximate the function with a Taylor series in other non-singular point such as in x_0 having for value 2:

```
>p[n](x) = sum((D@@i)(f)(a)/i!*(x-a)^i,i=0..n);
>p10:= taylor(1/(1-x^2),x=2,10);
>p10:= convert(p10,polynom);
>with(plots):
>tays:= plots[display](xplot):
for i from 1 by 2 to 10 do
tpl:= convert(taylor(1/(1-x^2), x=2,i),polynom):
tays:= tays,plots[display]([xplot,plot(tpl,x=0..5,y=-2..2,
color=black,title=convert(tpl,string))]) od:
>plots[display]([tays],view=[-0..5,-2..2]);
```

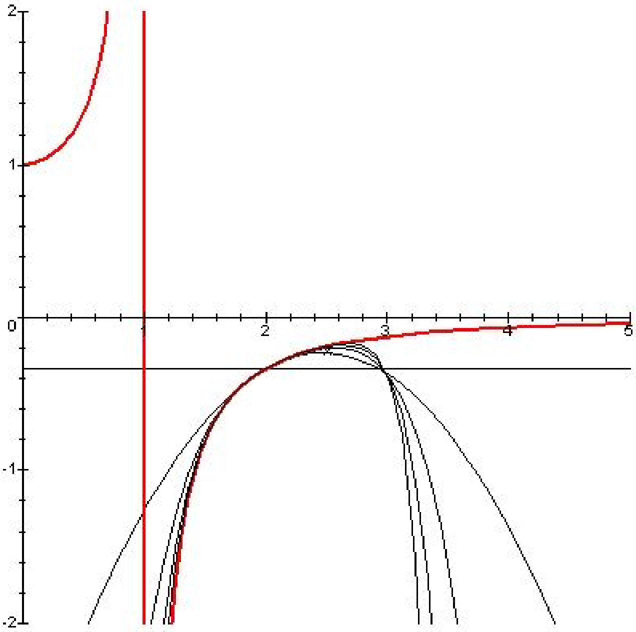
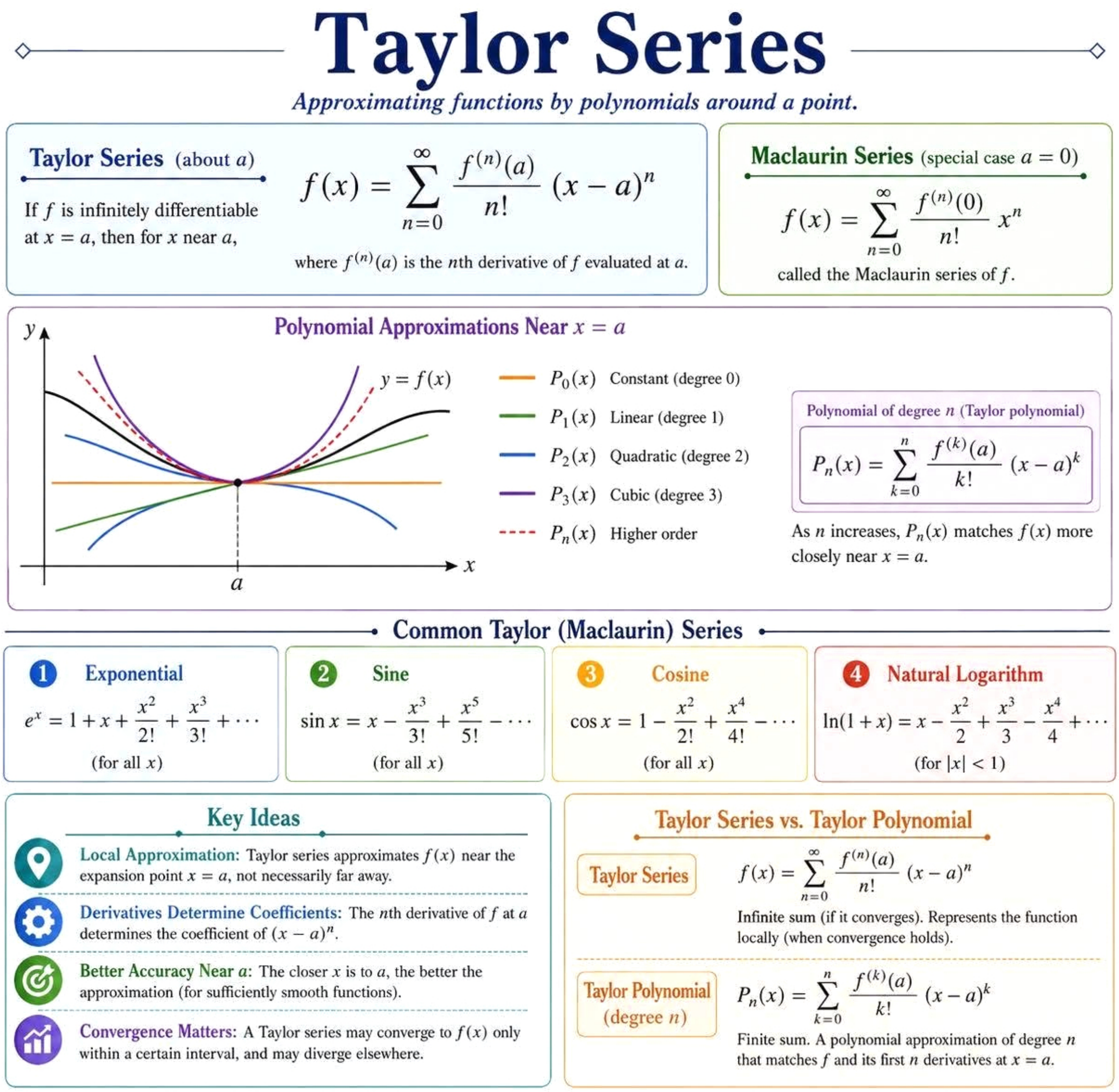


Figure 6.73: **Shift possibility of Maclaurin series in Maple™ 4.00b**

We will study a generalization to the complex plane of Taylor series in the section of Complex Analysis to get a veeeeery powerful result for physicists to calculate complicated curvilinear integrals.



Taylor Series

Approximating functions by polynomials around a point.

Taylor Series (about a)

If f is infinitely differentiable at $x = a$, then for x near a,

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n$$

where $f^{(n)}(a)$ is the nth derivative of f evaluated at a.

Maclaurin Series (special case $a = 0$)

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!}x^n$$

called the Maclaurin series of f.

Polynomial Approximations Near $x = a$

—	$P_0(x)$	Constant (degree 0)
—	$P_1(x)$	Linear (degree 1)
—	$P_2(x)$	Quadratic (degree 2)
—	$P_3(x)$	Cubic (degree 3)
----	$P_n(x)$	Higher order

Polynomial of degree n (Taylor polynomial)

$$P_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k$$

As n increases, $P_n(x)$ matches $f(x)$ more closely near $x = a$.

Common Taylor (Maclaurin) Series

① **Exponential**

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

(for all x)

② **Sine**

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots$$

(for all x)

③ **Cosine**

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots$$

(for all x)

④ **Natural Logarithm**

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots$$

(for $|x| < 1$)

Key Ideas

Local Approximation: Taylor series approximates $f(x)$ near the expansion point $x = a$, not necessarily far away.

Derivatives Determine Coefficients: The nth derivative of f at a determines the coefficient of $(x-a)^n$.

Better Accuracy Near a: The closer x is to a, the better the approximation (for sufficiently smooth functions).

Convergence Matters: A Taylor series may converge to $f(x)$ only within a certain interval, and may diverge elsewhere.

Taylor Series vs. Taylor Polynomial

Taylor Series

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n$$

Infinite sum (if it converges). Represents the function locally (when convergence holds).

Taylor Polynomial (degree n)

$$P_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k$$

Finite sum. A polynomial approximation of degree n that matches f and its first n derivatives at $x = a$.

6.4.2.6.1 Usual Maclaurin developments

We will prove here the most frequent Maclaurin developments (about ten) to the second order that we can meet in theoretical and mathematical physics (in fact we have developed here only use almost everywhere in the book). The list is not exhaustive for the time being but as the proof below are generalized, they can be applied to many other cases (that we will apply/meet throughout this textbook).

> **Remark**
>
> The Taylor expansions (that is to say elsewhere than on zero) are very rare (there are one or two in this entire book but they are detailed in their respective sections), we will omit them.

1. Taylor-Maclaurin development of $f(x)=e^x$:

   First remember that we have proved in the section of Differential and Integral Calculus that:

$$\begin{aligned}f'(x)&=\frac{\mathrm{d}e^x}{\mathrm{d}x}=e^x\\ f''(x)&=\frac{\mathrm{d}^2e^x}{\mathrm{d}x^2}=e^x\\ f'''(x)&=\frac{\mathrm{d}^3e^x}{\mathrm{d}x^3}=e^x\end{aligned} \tag{6.4.289}$$

   Therefore we have:

$$\begin{aligned}&f_3(x)\cong e^x\\ &\cong\sum_{k=0}^{3}\frac{1}{k!}\frac{\mathrm{d}^kP(0)}{\mathrm{d}x^k}x^k=\frac{1}{0!}\frac{\mathrm{d}^0P(0)}{\mathrm{d}x^0}x^0+\frac{1}{1!}\frac{\mathrm{d}^1P(0)}{\mathrm{d}x^1}x^1+\frac{1}{2!}\frac{\mathrm{d}^2P(0)}{\mathrm{d}x^2}x^2+\frac{1}{3!}\frac{\mathrm{d}^3P(0)}{\mathrm{d}x^3}x^3\\ &=\frac{1}{0!}e^x|_{x=0}+\frac{1}{1!}\frac{\mathrm{d}e^x}{\mathrm{d}x}\bigg|_{x=0}x^1+\frac{1}{2!}\frac{\mathrm{d}^2e^x}{\mathrm{d}x^2}\bigg|_{x=0}x^2+\frac{1}{3!}\frac{\mathrm{d}^3e^x}{\mathrm{d}x^3}\bigg|_{x=0}x^3\\ &=\frac{1}{0!}e^x|_{x=0}+\frac{1}{1!}e^x|_{x=0}x+\frac{1}{2!}e^x|_{x=0}x^2+\frac{1}{3!}e^x|_{x=0}x^3\\ &=1+\frac{1}{1!}e^0x+\frac{1}{2!}e^0x^2+\frac{1}{3!}e^0x^3\\ &=1+\frac{1}{1!}x+\frac{1}{2!}x^2+\frac{1}{3!}x^3=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}\end{aligned} \tag{6.4.290}$$

   More generally:

$$e^x=1+\sum_{i=1}^{+\infty}\frac{1}{i!}x^i=\sum_{i=0}^{+\infty}\frac{1}{i!}x^i \tag{6.4.291}$$

   And therefore we have the famous result for $x=1$:

$$e=\sum_{n=0}^{+\infty}\frac{1}{n!} \tag{6.4.292}$$

   that is sometimes named the "exponential sequence".

2. Taylor-Maclaurin development of $f(x)=\sin(x)$:

   First remember that we have proved in the section of Differential and Integral Calculus that:

$$\begin{aligned}f'(x)&=\frac{\mathrm{d}\sin(x)}{\mathrm{d}x}=\cos(x)\\ f''(x)&=\frac{\mathrm{d}^2\sin(x)}{\mathrm{d}x^2}=\frac{\mathrm{d}\cos(x)}{\mathrm{d}x}=-\sin(x)\\ f'''(x)&=\frac{\mathrm{d}^3\sin(x)}{\mathrm{d}x^3}=\frac{\mathrm{d}(-\sin(x))}{\mathrm{d}x}=-\cos(x)\end{aligned}\tag{6.4.293}$$

   Therefore we have:

$$\begin{aligned}f_3(x)&\cong\sin(x)\\ &\cong\sum_{k=0}^{3}\frac{1}{k!}\frac{\mathrm{d}^kP(0)}{\mathrm{d}x^k}x^k=\frac{1}{0!}\frac{\mathrm{d}^0P(0)}{\mathrm{d}x^0}x^0+\frac{1}{1!}\frac{\mathrm{d}^1P(0)}{\mathrm{d}x^1}x^1+\frac{1}{2!}\frac{\mathrm{d}^2P(0)}{\mathrm{d}x^2}x^2+\frac{1}{3!}\frac{\mathrm{d}^3P(0)}{\mathrm{d}x^3}x^3\\ &=\frac{1}{0!}\sin(x)|_{x=0}+\frac{1}{1!}\left.\frac{\mathrm{d}\sin(x)}{\mathrm{d}x}\right|_{x=0}x^1+\frac{1}{2!}\left.\frac{\mathrm{d}^2\sin(x)}{\mathrm{d}x^2}\right|_{x=0}x^2+\frac{1}{3!}\left.\frac{\mathrm{d}^3\sin(x)}{\mathrm{d}x^3}\right|_{x=0}x^3\\ &=\frac{1}{0!}\sin(x)|_{x=0}+\frac{1}{1!}\cos(x)|_{x=0}x-\frac{1}{2!}\sin(x)|_{x=0}x^2-\frac{1}{3!}\cos(x)|_{x=0}x^3\\ &=\frac{1}{1!}x-\frac{1}{3!}x^3=x-\frac{x^3}{3!}\end{aligned}\tag{6.4.294}$$

3. Taylor-Maclaurin development of $f(x)=\cos(x)$:

   First remember that we have proved in the section of Differential and Integral Calculus that:

$$\begin{aligned}f'(x)&=\frac{\mathrm{d}\cos(x)}{\mathrm{d}x}=-\sin(x)\\ f''(x)&=\frac{\mathrm{d}^2\cos(x)}{\mathrm{d}x^2}=\frac{\mathrm{d}(-\sin(x))}{\mathrm{d}x}=-\cos(x)\\ f'''(x)&=\frac{\mathrm{d}^3\cos(x)}{\mathrm{d}x^3}=\frac{\mathrm{d}(-\cos(x))}{\mathrm{d}x}=\sin(x)\end{aligned}\tag{6.4.295}$$

   Therefore we have:

$$\begin{aligned}f_3(x)&\cong\cos(x)\\ &\cong\sum_{k=0}^{3}\frac{1}{k!}\frac{\mathrm{d}^kP(0)}{\mathrm{d}x^k}x^k=\frac{1}{0!}\frac{\mathrm{d}^0P(0)}{\mathrm{d}x^0}x^0+\frac{1}{1!}\frac{\mathrm{d}^1P(0)}{\mathrm{d}x^1}x^1+\frac{1}{2!}\frac{\mathrm{d}^2P(0)}{\mathrm{d}x^2}x^2+\frac{1}{3!}\frac{\mathrm{d}^3P(0)}{\mathrm{d}x^3}x^3\\ &=\frac{1}{0!}\cos(x)|_{x=0}+\frac{1}{1!}\left.\frac{\mathrm{d}\cos(x)}{\mathrm{d}x}\right|_{x=0}x^1+\frac{1}{2!}\left.\frac{\mathrm{d}^2\cos(x)}{\mathrm{d}x^2}\right|_{x=0}x^2+\frac{1}{3!}\left.\frac{\mathrm{d}^3\cos(x)}{\mathrm{d}x^3}\right|_{x=0}x^3\\ &=\frac{1}{0!}\cos(x)|_{x=0}-\frac{1}{1!}\sin(x)|_{x=0}x-\frac{1}{2!}\cos(x)|_{x=0}x^2+\frac{1}{3!}\sin(x)|_{x=0}x^3\\ &=\frac{1}{0!}-\frac{1}{2!}x^2=1-\frac{x^2}{2!}\end{aligned}\tag{6.4.296}$$

4. Taylor-Maclaurin development of $f(x)=\tan(x)$:

   First remember that we have proved in the section of Differential and Integral Calculus that:

$$\begin{aligned}
f'(x)&=\frac{\mathrm{d}\tan(x)}{\mathrm{d}x}=\frac{1}{\cos^2(x)}=1+\tan^2(x)\\
f''(x)&=\frac{\mathrm{d}^2\tan(x)}{\mathrm{d}x^2}=\frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{\cos^2(x)}=2\frac{\sin(x)}{\cos^3(x)}\\
f'''(x)&=\frac{\mathrm{d}^3\tan(x)}{\mathrm{d}x^3}=2\frac{\mathrm{d}\left(\dfrac{\sin(x)}{\cos^3(x)}\right)}{\mathrm{d}x}=2\frac{\mathrm{d}\left(\sin(x)\dfrac{1}{\cos^3(x)}\right)}{\mathrm{d}x}\\
&=\frac{2}{\cos^2(x)}+2\sin(x)3\frac{\sin(x)}{\cos^4(x)}=\frac{2}{\cos^2(x)}+6\frac{\sin^2(x)}{\cos^4(x)}
\end{aligned}\tag{6.4.297}$$

   Therefore we have:

$$\begin{aligned}
&f_3(x)\cong\tan(x)\\
&\cong\sum_{k=0}^{3}\frac{1}{k!}\frac{\mathrm{d}^kP(0)}{\mathrm{d}x^k}x^k=\frac{1}{0!}\frac{\mathrm{d}^0P(0)}{\mathrm{d}x^0}x^0+\frac{1}{1!}\frac{\mathrm{d}^1P(0)}{\mathrm{d}x^1}x^1+\frac{1}{2!}\frac{\mathrm{d}^2P(0)}{\mathrm{d}x^2}x^2+\frac{1}{3!}\frac{\mathrm{d}^3P(0)}{\mathrm{d}x^3}x^3\\
&=\frac{1}{0!}\tan(x)|_{x=0}+\frac{1}{1!}\frac{\mathrm{d}\tan(x)}{\mathrm{d}x}\bigg|_{x=0}x^1+\frac{1}{2!}\frac{\mathrm{d}^2\tan(x)}{\mathrm{d}x^2}\bigg|_{x=0}x^2+\frac{1}{3!}\frac{\mathrm{d}^3\tan(x)}{\mathrm{d}x^3}\bigg|_{x=0}x^3\\
&=\frac{1}{0!}\tan(0)x^0+\frac{1}{1!}\left(1+\frac{\sin^2(x)}{\cos^2(x)}\right)\bigg|_{x=0}x^1+\frac{1}{2!}\left(2\frac{\sin(x)}{\cos(x)}+2\frac{\sin^3(x)}{\cos^3(x)}\right)\bigg|_{x=0}x^2\\
&+\frac{1}{3!}\left(2+8\frac{\sin^2(x)}{\cos^2(x)}+6\frac{\sin^4(x)}{\cos^4(x)}\right)\bigg|_{x=0}x^3\\
&=\frac{1}{1!}x-\frac{2}{3!}x^3=x+\frac{x^3}{3}
\end{aligned}\tag{6.4.298}$$

Remark

The reader must remember that for physicists and engineers the three following approximations are very useful and famous for small θ:

$$\sin(\theta)\cong\theta\qquad\cos(\theta)\cong 1\qquad\tan(\theta)\cong\theta$$

5. Taylor-Maclaurin development of $f(x)=\arctan(x)$:

   First remember that we have proved in the section of Differential and Integral Calculus that:

$$\begin{aligned}
f'(x)&=\frac{\mathrm{d}\arctan(x)}{\mathrm{d}x}=\frac{1}{1+x^2}\\
f''(x)&=\frac{\mathrm{d}^2\arctan(x)}{\mathrm{d}x^2}=\frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{1+x^2}=-2\frac{x}{(1+x^2)^2}\\
f'''(x)&=\frac{\mathrm{d}^3\arctan(x)}{\mathrm{d}x^3}=\frac{\mathrm{d}\left(-2\dfrac{x}{(1+x^2)^2}\right)}{\mathrm{d}x}=-2\frac{\mathrm{d}\left(x\dfrac{1}{(1+x^2)^2}\right)}{\mathrm{d}x}\\
&=\frac{-2}{(1+x^2)^2}+(-2)x\frac{(-2)2x}{(1+x^2)^3}=\frac{-2}{(1+x^2)^2}+\frac{8x^2}{(1+x^2)^3}
\end{aligned}\tag{6.4.299}$$

Therefore we have:

$$\begin{aligned}
&f_3(x) \cong \arctan(x)\\
&\cong \sum_{k=0}^{3} \frac{1}{k!}\frac{\mathrm{d}^k P(0)}{\mathrm{d}x^k}x^k = \frac{1}{0!}\frac{\mathrm{d}^0 P(0)}{\mathrm{d}x^0}x^0 + \frac{1}{1!}\frac{\mathrm{d}^1 P(0)}{\mathrm{d}x^1}x^1 + \frac{1}{2!}\frac{\mathrm{d}^2 P(0)}{\mathrm{d}x^2}x^2 + \frac{1}{3!}\frac{\mathrm{d}^3 P(0)}{\mathrm{d}x^3}x^3\\
&= \frac{1}{0!}\arctan(x)|_{x=0} + \frac{1}{1!}\frac{\mathrm{d}\arctan(x)}{\mathrm{d}x}\bigg|_{x=0} x^1 + \frac{1}{2!}\frac{\mathrm{d}^2\arctan(x)}{\mathrm{d}x^2}\bigg|_{x=0} x^2 + \frac{1}{3!}\frac{\mathrm{d}^3\arctan(x)}{\mathrm{d}x^3}\bigg|_{x=0} x^3\\
&= \frac{1}{0!}\arctan(0)x^0 + \frac{1}{1!}\frac{1}{1+x^2}\bigg|_{x=0} x^1 + \frac{1}{2!}\frac{-2x}{(1+x^2)^2}\bigg|_{x=0} x^2\\
&+ \frac{1}{3!}\left(\frac{-2}{(1+x^2)^2} + \frac{8x^2}{(1+x^2)^3}\right)\bigg|_{x=0} x^3\\
&= \frac{1}{1!}x - \frac{2}{3!}x^3 = x - \frac{x^3}{3}
\end{aligned} \tag{6.4.300}$$

6. Taylor-Maclaurin development of $f(x) = \dfrac{1}{1+x}$:

First remember that we have proved in the section of Differential and Integral Calculus that:

$$\begin{aligned}
f'(x) &= \frac{\mathrm{d}\frac{1}{1+x}}{\mathrm{d}x} = -\frac{1}{1+x}^{2}\\
f''(x) &= \frac{\mathrm{d}^2\frac{1}{1+x}}{\mathrm{d}x^2} = \frac{\mathrm{d}}{\mathrm{d}x} - \frac{1}{1+x}^{2} = \frac{2}{(1+x)^3}\\
f'''(x) &= \frac{\mathrm{d}^3\frac{1}{1+x}}{\mathrm{d}x^3} = \frac{\mathrm{d}\left(\dfrac{2}{(1+x)^3}\right)}{\mathrm{d}x} = -\frac{6}{(1+x)^4}
\end{aligned} \tag{6.4.301}$$

Therefore we have:

$$\begin{aligned}
&f_3(x) \cong \frac{1}{1+x}\\
&\cong \sum_{k=0}^{3} \frac{1}{k!}\frac{\mathrm{d}^k P(0)}{\mathrm{d}x^k}x^k = \frac{1}{0!}\frac{\mathrm{d}^0 P(0)}{\mathrm{d}x^0}x^0 + \frac{1}{1!}\frac{\mathrm{d}^1 P(0)}{\mathrm{d}x^1}x^1 + \frac{1}{2!}\frac{\mathrm{d}^2 P(0)}{\mathrm{d}x^2}x^2 + \frac{1}{3!}\frac{\mathrm{d}^3 P(0)}{\mathrm{d}x^3}x^3\\
&= \frac{1}{0!}\frac{1}{1+x}|_{x=0} + \frac{1}{1!}\frac{\mathrm{d}\frac{1}{1+x}}{\mathrm{d}x}\bigg|_{x=0} x^1 + \frac{1}{2!}\frac{\mathrm{d}^2\frac{1}{1+x}}{\mathrm{d}x^2}\bigg|_{x=0} x^2 + \frac{1}{3!}\frac{\mathrm{d}^3\frac{1}{1+x}}{\mathrm{d}x^3}\bigg|_{x=0} x^3\\
&= \frac{1}{0!}1x^0 - \frac{1}{1!}\frac{1}{(1+x)^2}\bigg|_{x=0} x^1 + \frac{1}{2!}\frac{2}{(1+x)^3}\bigg|_{x=0} x^2 - \frac{1}{3!}\left(\frac{6}{(1+x)^4}\right)\bigg|_{x=0} x^3\\
&= 1 - \frac{1}{1!}x - \frac{2}{2!}x^2 - \frac{6}{3!}x^3 = 1 - x + x^2 - x^3
\end{aligned} \tag{6.4.302}$$

It then follows immediately another Taylor series we will also meet again many number of times:

$$f_2(x) \cong \frac{1}{(1+x)^2} = -\frac{\mathrm{d}}{x}\frac{1}{(1+x)} \cong -\frac{\mathrm{d}}{x}\left(1 - x + x^2 - x^3\right) = 1 - x^2 + 3x^2 \tag{6.4.303}$$

For our study of the Cosmological Microwave Background in the section of Cosmology (especially recombination era), we will need a result related to the two series above named the "Binomial theorem for negative integer exponents". Given the binomial coefficient given for recall by:

$$\begin{pmatrix} \alpha \\ k \end{pmatrix}=\frac{\alpha!}{k!(\alpha-k)!} \tag{6.4.304}$$

Let n be a positive integer, then (without proof but using "engineer intuition") we have:

$$\boxed{\frac{1}{(1+x)^n}=1-nx+\frac{(-n)(-n-1)}{2}x^2+\cdots=\sum_{k=0}^{+\infty}\begin{pmatrix} n+k-1 \\ k \end{pmatrix}(-1)^k x^k} \tag{6.4.305}$$

for $|x|<1$.

7. Taylor-Maclaurin development of $f(x)=\sqrt{1+x}$:

   First remember that we have proved in the section of Differential and Integral Calculus that:

$$\begin{aligned}
f'(x)&=\frac{\mathrm{d}\sqrt{1+x}}{\mathrm{d}x}=\frac{1}{2}\frac{1}{\sqrt{1+x}}\\
f''(x)&=\frac{\mathrm{d}^2\sqrt{1+x}}{\mathrm{d}x^2}=\frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{2}\frac{1}{\sqrt{1+x}}=-\frac{1}{4}\frac{1}{(1+x)^{3/2}}\\
f'''(x)&=\frac{\mathrm{d}^3\sqrt{1+x}}{\mathrm{d}x^3}=\frac{\mathrm{d}\left(-\frac{1}{4}\frac{1}{(1+x)^{3/2}}\right)}{\mathrm{d}x}=\frac{3}{8}\frac{1}{(1+x)^{5/2}}
\end{aligned} \tag{6.4.306}$$

   Therefore we have:

$$\begin{aligned}
f_3(x)&\cong\sqrt{1+x}\\
&\cong\sum_{k=0}^{3}\frac{1}{k!}\frac{\mathrm{d}^kP(0)}{\mathrm{d}x^k}x^k=\frac{1}{0!}\frac{\mathrm{d}^0P(0)}{\mathrm{d}x^0}x^0+\frac{1}{1!}\frac{\mathrm{d}^1P(0)}{\mathrm{d}x^1}x^1+\frac{1}{2!}\frac{\mathrm{d}^2P(0)}{\mathrm{d}x^2}x^2+\frac{1}{3!}\frac{\mathrm{d}^3P(0)}{\mathrm{d}x^3}x^3\\
&=\frac{1}{0!}\sqrt{1+x}|_{x=0}+\frac{1}{1!}\left.\frac{\mathrm{d}\sqrt{1+x}}{\mathrm{d}x}\right|_{x=0}x^1+\frac{1}{2!}\left.\frac{\mathrm{d}^2\sqrt{1+x}}{\mathrm{d}x^2}\right|_{x=0}x^2+\frac{1}{3!}\left.\frac{\mathrm{d}^3\sqrt{1+x}}{\mathrm{d}x^3}\right|_{x=0}x^3\\
&=\frac{1}{0!}\sqrt{1+x}\Big|_{x=0}x^0+\frac{1}{1!}\frac{1}{2}\frac{1}{\sqrt{1+x}}\Big|_{x=0}x^1-\frac{1}{2!}\frac{1}{4}\frac{1}{(1+x)^{3/2}}\Big|_{x=0}x^2+\frac{1}{3!}\left(\frac{3}{8}\frac{1}{(1+x)^{5/2}}\right)\Big|_{x=0}x^3\\
&=\frac{1}{0!}+\frac{1}{1!}\frac{1}{2}x-\frac{1}{2!}\frac{1}{4}x^2+\frac{1}{3!}\frac{3}{8}x^3=1+\frac{1}{2}x-\frac{1}{2\cdot 4}x^2+\frac{1\cdot 3}{2\cdot 4\cdot 6}x^3
\end{aligned} \tag{6.4.307}$$

   It then also follows immediately another Taylor series we will also meet again many number of times:

$$\begin{aligned}
f_2(x)&\cong\frac{1}{\sqrt{1+x}}=2\frac{\mathrm{d}}{\mathrm{d}x}\sqrt{1+x}\cong 2\frac{\mathrm{d}}{\mathrm{d}x}\left(1+\frac{1}{2}x-\frac{1}{2\cdot 4}x^2+\frac{1\cdot 3}{2\cdot 4\cdot 6}x^3\right)\\
&=1-\frac{1}{2}x+\frac{1\cdot 3}{2\cdot 4}x^2
\end{aligned} \tag{6.4.308}$$

8. Taylor-Maclaurin development of $f(x)=\sqrt{1-x^2}$ (that will be useful to us in the section of Astronomy):

Using the usual derivatives we have proved in the section of Differential and Integral Calculus we have:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\left[\sqrt{1-x^2}\right]&=\frac{1}{2}\left(1-x^2\right)^{\frac{1}{2}-1}\cdot\frac{\mathrm{d}}{\mathrm{d}x}\left[1-x^2\right]=\frac{\frac{\mathrm{d}}{\mathrm{d}x}[1]-\frac{\mathrm{d}}{\mathrm{d}x}\left[x^2\right]}{2\sqrt{1-x^2}}\\&=\frac{0-2x}{2\sqrt{1-x^2}}=-\frac{x}{\sqrt{1-x^2}}\end{aligned}\tag{6.4.309}$$

And for the second derivative:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\left[-\frac{x}{\sqrt{1-x^2}}\right]&=-\frac{\mathrm{d}}{\mathrm{d}x}\left[\frac{x}{\sqrt{1-x^2}}\right]=-\frac{\frac{\mathrm{d}}{\mathrm{d}x}[x]\cdot\sqrt{1-x^2}-x\cdot\frac{\mathrm{d}}{\mathrm{d}x}\left[\sqrt{1-x^2}\right]}{\left(\sqrt{1-x^2}\right)^2}\\&=-\frac{1\sqrt{1-x^2}-x\cdot\frac{1}{2}\left(1-x^2\right)^{\frac{1}{2}-1}\cdot\frac{\mathrm{d}}{\mathrm{d}x}\left[1-x^2\right]}{1-x^2}\\&=-\frac{\sqrt{1-x^2}-\frac{x\left(\frac{\mathrm{d}}{\mathrm{d}x}[1]-\frac{\mathrm{d}}{\mathrm{d}x}\left[x^2\right]\right)}{2\sqrt{1-x^2}}}{1-x^2}=-\frac{\sqrt{1-x^2}-\frac{x(0-2x)}{2\sqrt{1-x^2}}}{1-x^2}\\&=-\frac{\sqrt{1-x^2}+\frac{x^2}{\sqrt{1-x^2}}}{1-x^2}=-\frac{1}{\sqrt{1-x^2}}-\frac{x^2}{\left(1-x^2\right)^{\frac{3}{2}}}=-\frac{1}{\left(1-x^2\right)^{\frac{3}{2}}}\end{aligned}\tag{6.4.310}$$

and for the third derivative:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\left[-\frac{1}{\left(1-x^2\right)^{\frac{3}{2}}}\right]&=-\left(-\frac{3}{2}\right)\left(1-x^2\right)^{-\frac{3}{2}-1}\cdot\frac{\mathrm{d}}{\mathrm{d}x}\left[1-x^2\right]\\&=\frac{3\left(\frac{\mathrm{d}}{\mathrm{d}x}[1]-\frac{\mathrm{d}}{\mathrm{d}x}\left[x^2\right]\right)}{2\left(1-x^2\right)^{\frac{5}{2}}}=\frac{3(0-2x)}{2\left(1-x^2\right)^{\frac{5}{2}}}=-\frac{3x}{\left(1-x^2\right)^{\frac{5}{2}}}\end{aligned}\tag{6.4.311}$$

and for the fourth and last derivative:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}x}\left[-\frac{3x}{(1-x^2)^{\frac{5}{2}}}\right] &= -3\cdot\frac{\mathrm{d}}{\mathrm{d}x}\left[\frac{x}{(1-x^2)^{\frac{5}{2}}}\right] \\
&= -3\cdot\frac{\frac{\mathrm{d}}{\mathrm{d}x}[x]\cdot(1-x^2)^{\frac{5}{2}}-x\cdot\frac{\mathrm{d}}{\mathrm{d}x}\left[(1-x^2)^{\frac{5}{2}}\right]}{\left((1-x^2)^{\frac{5}{2}}\right)^2} \\
&= -\frac{3\left(1\,(1-x^2)^{\frac{5}{2}}-\frac{5}{2}\,(1-x^2)^{\frac{5}{2}-1}\cdot\frac{\mathrm{d}}{\mathrm{d}x}\,[1-x^2]\cdot x\right)}{(1-x^2)^5} \\
&= -\frac{3\left((1-x^2)^{\frac{5}{2}}-\frac{5\left(\frac{\mathrm{d}}{\mathrm{d}x}[1]-\frac{\mathrm{d}}{\mathrm{d}x}[x^2]\right)x\left(1-x^2\right)^{\frac{3}{2}}}{2}\right)}{(1-x^2)^5} \\
&= -\frac{3\left((1-x^2)^{\frac{5}{2}}-\frac{5(0-2x)x\left(1-x^2\right)^{\frac{3}{2}}}{2}\right)}{(1-x^2)^5} \\
&= -\frac{3\left((1-x^2)^{\frac{5}{2}}+5x^2\,(1-x^2)^{\frac{3}{2}}\right)}{(1-x^2)^5} = -\frac{3}{(1-x^2)^{\frac{5}{2}}}-\frac{15x^2}{(1-x^2)^{\frac{7}{2}}} \\
&= -\frac{12x^2+3}{(1-x^2)^{\frac{7}{2}}}
\end{aligned} \tag{6.4.312}$$

Therefore based on the above results:

$$\begin{aligned}
f'(x) &= \frac{\mathrm{d}\sqrt{1-x^2}}{\mathrm{d}x} = -\frac{x}{\sqrt{1-x^2}} \\
f''(x) &= -\frac{1}{(1-x^2)^{\frac{3}{2}}} \\
f'''(x) &= -\frac{3x}{(1-x^2)^{\frac{5}{2}}} \\
f''''(x) &= -\frac{12x^2+3}{(1-x^2)^{\frac{7}{2}}}
\end{aligned} \tag{6.4.313}$$

Therefore we have:

$$\begin{aligned}
f_4(x) &\cong \sqrt{1+x}\\
&\cong \sum_{k=0}^{4}\frac{1}{k!}\frac{\mathrm{d}^kP(0)}{\mathrm{d}x^k}x^k=\frac{1}{0!}\frac{\mathrm{d}^0P(0)}{\mathrm{d}x^0}x^0+\frac{1}{1!}\frac{\mathrm{d}^1P(0)}{\mathrm{d}x^1}x^1+\frac{1}{2!}\frac{\mathrm{d}^2P(0)}{\mathrm{d}x^2}x^2+\frac{1}{3!}\frac{\mathrm{d}^3P(0)}{\mathrm{d}x^3}x^3+\frac{1}{4!}\frac{\mathrm{d}^4P(0)}{\mathrm{d}x^4}x^4\\
&=\frac{1}{0!}\sqrt{1-x^2}|_{x=0}+\frac{1}{1!}\left.\frac{\mathrm{d}\sqrt{1-x^2}}{\mathrm{d}x}\right|_{x=0}x^1+\frac{1}{2!}\left.\frac{\mathrm{d}^2\sqrt{1-x^2}}{\mathrm{d}x^2}\right|_{x=0}x^2+\frac{1}{3!}\left.\frac{\mathrm{d}^3\sqrt{1-x^2}}{\mathrm{d}x^3}\right|_{x=0}x^3\\
&+\frac{1}{4!}\left.\frac{\mathrm{d}^4\sqrt{1-x^2}}{\mathrm{d}x^4}\right|_{x=0}x^4\\
&=\frac{1}{0!}\left.\sqrt{1-x^2}\right|_{x=0}x^0-\frac{1}{1!}\left.\frac{x}{\sqrt{1-x^2}}\right|_{x=0}x^1-\frac{1}{2!}\left.\frac{1}{(1-x^2)^{\frac{3}{2}}}\right|_{x=0}x^2-\frac{1}{3!}\left.\frac{3x}{2\,(1-x^2)^{\frac{7}{4}}}\right|_{x=0}x^3\\
&-\frac{1}{4!}\left.\frac{12x^2+3}{(1-x^2)^{\frac{7}{2}}}\right|_{x=0}x^4\\
&=\frac{1}{0!}-\frac{1}{2!}x^2-\frac{1}{4!}3x^4=1-\frac{x^2}{2}-\frac{x^3}{8}
\end{aligned}\tag{6.4.314}$$

9. Taylor-Maclaurin development of $f(x)=\ln(1+x)$:

   First remember that we have proved in the section of Differential and Integral Calculus that:

$$\begin{aligned}
f'(x)&=\frac{\mathrm{d}\ln(1+x)}{\mathrm{d}x}=\frac{1}{1+x}\\
f''(x)&=\frac{\mathrm{d}^2\ln(1+x)}{\mathrm{d}x^2}=\frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{1+x}=-\frac{1}{(1+x)^2}\\
f'''(x)&=\frac{\mathrm{d}^3\ln(1+x)}{\mathrm{d}x^3}=\frac{\mathrm{d}\left(-\dfrac{1}{(1+x)^2}\right)}{\mathrm{d}x}=\frac{2}{(1+x)^3}
\end{aligned}\tag{6.4.315}$$

   Therefore we have:

$$\begin{aligned}
f_3(x)&\cong \ln(1+x)\\
&\cong\sum_{k=0}^{3}\frac{1}{k!}\frac{\mathrm{d}^kP(0)}{\mathrm{d}x^k}x^k=\frac{1}{0!}\frac{\mathrm{d}^0P(0)}{\mathrm{d}x^0}x^0+\frac{1}{1!}\frac{\mathrm{d}^1P(0)}{\mathrm{d}x^1}x^1+\frac{1}{2!}\frac{\mathrm{d}^2P(0)}{\mathrm{d}x^2}x^2+\frac{1}{3!}\frac{\mathrm{d}^3P(0)}{\mathrm{d}x^3}x^3\\
&=\frac{1}{0!}\ln(1+x)|_{x=0}+\frac{1}{1!}\left.\frac{\mathrm{d}\ln(1+x)}{\mathrm{d}x}\right|_{x=0}x^1+\frac{1}{2!}\left.\frac{\mathrm{d}^2\ln(1+x)}{\mathrm{d}x^2}\right|_{x=0}x^2+\frac{1}{3!}\left.\frac{\mathrm{d}^3\ln(1+x)}{\mathrm{d}x^3}\right|_{x=0}x^3\\
&=\frac{1}{0!}\left.\ln(1+x)\right|_{x=0}x^0+\frac{1}{1!}\left.\frac{1}{1+x}\right|_{x=0}x^1-\frac{1}{2!}\left.\frac{1}{(1+x)^2}\right|_{x=0}x^2+\frac{1}{3!}\left.\frac{2}{(1+x)^3}\right|_{x=0}x^3\\
&=\frac{1}{1!}x+\frac{1}{2!}x^2+\frac{2}{3!}x^3=x-\frac{1}{2}x^2+\frac{1}{3}x^3
\end{aligned}\tag{6.4.316}$$

10. Now let us consider the important case for the Langevin model of paramagnetism that is the approximated Taylor expansion of the hyperbolic cotangent function (see section Trigonometry page 2532), that is for refresh defined by the relation:

$$\coth(x)=\frac{e^x+e^{-x}}{e^x-e^{-x}}\tag{6.4.317}$$

For this, we will use the Landau notation, with expressions like $\mathcal{O}(x^n)$ remembering that we proved a little before above:

$$e^x = 1 + x + \frac{x^2}{2!} + \mathcal{O}(x^3) \tag{6.4.318}$$

when $x \to 0$. For the hyperbolic cotangent we have then:

$$\begin{aligned}\coth(x) = \frac{e^x + e^{-x}}{e^x - e^{-x}} &= \frac{1 + x + \frac{x^2}{2} + \mathcal{O}(x^3) + 1 - x + \frac{x^2}{2} + \mathcal{O}(x^3)}{1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \mathcal{O}(x^4) - 1 + x - \frac{x^2}{2} + \frac{x^3}{3!} + \mathcal{O}(x^4)} \\ &= \frac{2 + x^2 + \mathcal{O}(x^3)}{2x + \frac{x^3}{3} + \mathcal{O}(x^4)} = \frac{2 + x^2 + \mathcal{O}(x^3)}{2x} \underbrace{\frac{1}{1 + \frac{x^2}{6} + \mathcal{O}(x^3)}}_{:=A}\end{aligned} \tag{6.4.319}$$

Now we must remember as we have just proved a little earlier that:

$$\frac{1}{1+x} = 1 - x + x^2 - x^3 + \ldots \tag{6.4.320}$$

for $|x| < 1$. Therefore:

$$A = \frac{1}{1 + \left(\frac{x^2}{6} + \mathcal{O}(x^3)\right)} = 1 - \frac{x^2}{6} + \mathcal{O}(x^4) \tag{6.4.321}$$

and finally replacing this in the previous expression we find:

$$\begin{aligned}\coth(x) = \frac{e^x + e^{-x}}{e^x - e^{-x}} &= \frac{2 + x^2 + \mathcal{O}(x^3)}{2x}\left(1 - \frac{x^2}{6} + \mathcal{O}(x^4)\right) \\ &= \frac{1}{2x}\left(2 + x^2 - 2\frac{x^2}{6} + \mathcal{O}(x^4)\right) = \frac{1}{x} + \frac{x}{3} + \mathcal{O}(x^4)\end{aligned} \tag{6.4.322}$$

11. Another famous Maclaurin series used thousand of times in the world for business application is the computation of the numerical values of the Normal distribution:

$$\mathcal{N}(\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \mathrm{d}x \tag{6.4.323}$$

So first to simplify this integral, we typically let:

$$z = \frac{x - \mu}{\sigma} \tag{6.4.324}$$

that we know already (see section Statistics page 595) as being the z score of a data value. With this simplification, the integral above becomes:

$$\mathcal{N}(0, 1) = \frac{1}{\sqrt{2\pi}} \int_{(a-\mu)/\sigma}^{(b-\mu)/\sigma} e^{-z^2} \mathrm{d}z \tag{6.4.325}$$

The Maclaurin series for $e^{-x^2/2}$ is given by:

$$\begin{aligned} e^{-x^2/2} &= \sum_{n=0}^{+\infty} \frac{\left(-\dfrac{x^2}{2}\right)^2}{n!} = 1 - \frac{x^2}{2^1 1!} + \frac{x^4}{2^2 2!} - \frac{x^6}{2^3 3!} + \ldots + (-1)^n \frac{x^{2n}}{2^n n!} + \ldots \\ &= \sum_{n=0}^{+\infty} (-1)^n \frac{x^{2n}}{2^n n!} \end{aligned} \tag{6.4.326}$$

Therefore:

$$\begin{aligned} &\frac{1}{\sqrt{2\pi}} \int e^{-z^2/2} \mathrm{d}z = \frac{1}{\sqrt{2\pi}} \int \left(1 - \frac{z^2}{2^1 1!} + \frac{z^4}{2^2 2!} - \frac{z^6}{2^3 3!} + \ldots + (-1)^n \frac{z^{2n}}{2^n n!} + \ldots\right) \mathrm{d}z \\ &= \frac{1}{\sqrt{2\pi}} \left(c^{te} + z - \frac{z^3}{3 \cdot 2^1 \cdot 1!} + \frac{z^5}{5 \cdot 2^2 \cdot 2!} - \frac{z^7}{7 \cdot 2^3 \cdot 3!} + \ldots + (-1)^n \frac{z^{2n+1}}{(2n+1) \cdot 2^n \cdot n!} + \ldots\right) \end{aligned} \tag{6.4.327}$$

Therefore:

$$\frac{1}{\sqrt{2\pi}} \int_{a'}^{b'} e^{-z^2/2} \mathrm{d}z = \frac{1}{\sqrt{2\pi}} \left(c^{te} + z - \frac{z^3}{3 \cdot 2^1 \cdot 1!} + \frac{z^5}{5 \cdot 2^2 \cdot 2!} - \ldots + (-1)^n \frac{z^{2n+1}}{(2n+1) \cdot 2^n \cdot n!} + \ldots\right)\Bigg|_{a'}^{b'} \tag{6.4.328}$$

It is obvious that the constant will eliminate itself. Therefore!

$$\frac{1}{\sqrt{2\pi}} \int_{a'}^{b'} e^{-z^2/2} \mathrm{d}z = \frac{1}{\sqrt{2\pi}} \left(z - \frac{z^3}{3 \cdot 2^1 \cdot 1!} + \frac{z^5}{5 \cdot 2^2 \cdot 2!} - \ldots + (-1)^n \frac{z^{2n+1}}{(2n+1) \cdot 2^n \cdot n!} + \ldots\right)\Bigg|_{a'}^{b'} \tag{6.4.329}$$

and in the common case in business where $a = 0$ we get (with two terms only):

$$\begin{aligned} &\frac{1}{\sqrt{2\pi}} \int_{-\mu/\sigma}^{(b-\mu)/\sigma} e^{-z^2/2} \mathrm{d}z \\ &= \frac{1}{\sqrt{2\pi}} \left(\left(\frac{b-\mu}{\sigma}\right) - \frac{\left(\dfrac{b-\mu}{\sigma}\right)^3}{3 \cdot 2^1 \cdot 1!} + \frac{\left(\dfrac{b-\mu}{\sigma}\right)^5}{5 \cdot 2^2 \cdot 2!} \right) \\ &- \frac{1}{\sqrt{2\pi}} \left(\left(\frac{-\mu}{\sigma}\right) - \frac{\left(\dfrac{-\mu}{\sigma}\right)^3}{3 \cdot 2^1 \cdot 1!} + \frac{\left(\dfrac{-\mu}{\sigma}\right)^5}{5 \cdot 2^2 \cdot 2!} \right) \end{aligned} \tag{6.4.330}$$

To summarize here is a table with the most important Maclaurin series:

Function	**Maclaurin approximation**
e^x	$\sum\limits_{n=0}^{+\infty}\dfrac{x^n}{n!}=1+\dfrac{x}{1!}+\dfrac{x^2}{2!}+\dfrac{x^3}{3!}+\cdots$
$\sin(x)$	$\sum_{n=0}^{\infty}(-1)^n\frac{x^{2n+1}}{(2n+1)!}=x-\frac{x^3}{3!}+\frac{x^5}{5!}-\frac{x^7}{7!}+\cdots$
$\cos(x)$	$\sum\limits_{n=0}^{+\infty}(-1)^n\dfrac{x^{2n}}{(2n)!}=1-\dfrac{x^2}{2!}+\dfrac{x^4}{4!}-\dfrac{x^6}{6!}+\cdots$
$\tan(x)$	$\sum\limits_{n=1}^{+\infty}\dfrac{(-1)^{n-1}2^{2n}\left(2^{2n}-1\right)B_{2n}x^{2n-1}}{(2n)!}=x+\dfrac{1}{3}x^3+\dfrac{2}{15}x^5+\dfrac{17}{315}x^7+\cdots$
$\tan^{-1}(x)$	$\sum\limits_{k=0}^{+\infty}\dfrac{(-1)^k x^{2k+1}}{2k+1}=x-\dfrac{x^3}{3}+\dfrac{x^5}{5}-\cdots+\dfrac{(-1)^k x^{2k+1}}{2k+1}+\cdots$, for $\lvert x\rvert\leq 1$
$(1+x)^k$	$\sum\limits_{n=0}^{+\infty}\begin{pmatrix}k\\n\end{pmatrix}x^n=1+kx+\dfrac{k(k-1)}{2!}x^2+\dfrac{k(k-1)(k-2)}{3!}x^3+\cdots$
$\ln(1+x)$	$\sum\limits_{k=1}^{+\infty}\dfrac{(-1)^{k+1}x^k}{k}=x-\dfrac{x^2}{2}+\dfrac{x^3}{3}-\cdots+\dfrac{(-1)^{k+1}x^k}{k}+\cdots$, for $-1<x\leq 1$
$-\ln(1-x)$	$\sum\limits_{k=1}^{+\infty}\dfrac{x^k}{k}=x+\dfrac{x^2}{2}+\dfrac{x^3}{3}+\cdots+\dfrac{x^k}{k}+\cdots$, for $-1\leq x<1$
$\dfrac{1}{1-x}$	$\sum\limits_{k=0}^{+\infty}x^k=1+x+x^2+\cdots+x^k+\cdots$, for $\lvert x\rvert<1$
$\dfrac{1}{1+x}$	$\sum\limits_{k=0}^{+\infty}(-1)^k x^k=1-x+x^2-\cdots+(-1)^k x^k+\cdots$, for $\lvert x\rvert<1$
$\sinh(x)$	$\sum\limits_{k=0}^{+\infty}\dfrac{x^{2k+1}}{(2k+1)!}=x+\dfrac{x^3}{3!}+\dfrac{x^5}{5!}+\cdots+\dfrac{x^{2k+1}}{(2k+1)!}+\cdots$, for $\lvert x\rvert<+\infty$
$\cosh(x)$	$\sum\limits_{k=0}^{+\infty}\dfrac{x^{2k}}{(2k)!}=1+\dfrac{x^2}{2!}+\dfrac{x^4}{4!}+\cdots+\dfrac{x^{2k}}{(2k)!}+\cdots$, for $\lvert x\rvert<+\infty$
$\sqrt{1+x}$	$1+\dfrac{1}{2}x-\dfrac{1}{2\cdot 4}x^2+\dfrac{1\cdot 3}{2\cdot 4\cdot 6}x^3+\cdots$
$\dfrac{1}{\sqrt{1+x}}$	$1-\dfrac{1}{2}x+\dfrac{1\cdot 3}{2\cdot 4}x^2+\cdots$
$\coth(x)$	$\dfrac{1}{x}+\dfrac{x}{3}-\dfrac{x^3}{45}+\dfrac{2x^5}{945}+\cdots+\dfrac{(-1)^{n-1}2^{2n}B_n x^{2n-1}}{(2n)!}+\cdots$, for $0<\lvert x\rvert<\pi$
$\dfrac{1}{\sqrt{2\pi}}\displaystyle\int_{\alpha'}^{\sigma'}e^{-z^2/2}\mathrm{d}z$	$\dfrac{1}{\sqrt{2\pi}}\left(z-\dfrac{z^3}{3\cdot 2^1\cdot 1!}+\dfrac{z^5}{5\cdot 2^2\cdot 2!}-\ldots+(-1)^n\dfrac{z^{2n+1}}{(2n+1)\cdot 2^n\cdot n!}+\ldots\right)\Bigg\rvert_{\alpha'}^{b}$

Table 6.16: **Most common Maclaurin series**

6.4.2.6.2 Taylor series of bivariate functions (multivariate Taylor series)

We will see now how to approach a function $f(x,y)$ of two real variables by a sum of powers (Taylor series). This type of approximation is widely used in many fields of engineering (see sections of Industrial Engineering page 8786 and Numerical Methods page 6673).

We are looking for an approximation of $f(x,y)$ at point $f(x_0+h,y_0+h)$. For this, let us write (a priori nothing prevents us from doing so) that:

$$x=x(t)=x_0+ht \qquad y=y(t)=y_0+kt \tag{6.4.331}$$

Then we have:

$$f(t)=f(x(t),y(t)) \tag{6.4.332}$$

The value of (the trick is here!):

$$f(1)=f(x(1),y(1))=f(x_0+h,y_0+k) \tag{6.4.333}$$

can be approximated using a Taylor series around the value 0 such that:

$$\begin{aligned}f(1)&=\sum_{k=0}^{n}\frac{1}{k!}\frac{{}^k f(t_0)}{\mathrm{d}t^k}(t-t_0)^k=\sum_{k=0}^{n}\frac{1}{k!}\frac{{}^k f(0)}{\mathrm{d}t^k}(1-0)^k=\sum_{k=0}^{n}\frac{1}{k!}\frac{{}^k f(0)}{\mathrm{d}t^k}\\&=f(0)+1\cdot\frac{\mathrm{d}f}{\mathrm{d}t}(0)+\frac{1}{2!}\frac{\mathrm{d}^2f}{\mathrm{d}t^2}(0)+\ldots+\frac{1}{n!}\frac{\mathrm{d}^nf}{\mathrm{d}t^n}(0)\\&=f(x_0,y_0)+1\cdot\frac{\mathrm{d}f(x(t),y(t))}{\mathrm{d}t}(0)+\frac{1}{2!}\frac{\mathrm{d}^2f(x(t),y(t))}{\mathrm{d}t^2}(0)+\ldots+\frac{1}{n!}\frac{\mathrm{d}^nf(x(t),y(t))}{\mathrm{d}t^n}(0)\end{aligned} \tag{6.4.334}$$

But we have:

$$\begin{aligned}\frac{\mathrm{d}f(x(t),y(t))}{\mathrm{d}t}&=\frac{\dfrac{\partial f}{\partial x}\mathrm{d}x+\dfrac{\partial f}{\partial y}\mathrm{d}y}{\mathrm{d}t}=\frac{\dfrac{\partial f}{\partial x}\mathrm{d}(x_0+ht)+\dfrac{\partial f}{\partial y}\mathrm{d}(y_0+kt)}{\mathrm{d}t}=\frac{\dfrac{\partial f}{\partial x}h\mathrm{d}t+\dfrac{\partial f}{\partial y}k\mathrm{d}t}{\mathrm{d}t}\\&=\frac{\partial f}{\partial x}h+\frac{\partial f}{\partial y}k=\left(h\frac{\partial}{\partial x}+k\frac{\partial}{\partial y}\right)f\end{aligned} \tag{6.4.335}$$

and:

$$\begin{aligned}\frac{\mathrm{d}^2f(x(t),y(t))}{\mathrm{d}t^2}&=\frac{\mathrm{d}\left(\mathrm{d}f(x(t),y(t))\right)}{\mathrm{d}t^2}=\frac{\dfrac{\partial(\mathrm{d}x)}{\partial x}\mathrm{d}x+\dfrac{\partial(\mathrm{d}f)}{\partial y}\mathrm{d}f}{\mathrm{d}t^2}\\&=\frac{\partial\left(\dfrac{\partial f}{\partial x}\mathrm{d}x+\dfrac{\partial f}{\partial y}\mathrm{d}y\right)+\mathrm{d}y\dfrac{\partial}{\partial y}\left(\dfrac{\partial f}{\partial x}\mathrm{d}x+\dfrac{\partial f}{\partial y}\mathrm{d}y\right)}{\mathrm{d}t^2}\\&=\frac{\left[(\mathrm{d}x)^2\dfrac{\partial^2f}{\partial x^2}+\mathrm{d}x\mathrm{d}y\dfrac{\partial^2f}{\partial x\partial y}\right]+\left[\mathrm{d}x\mathrm{d}y\dfrac{\partial^2f}{\partial y\partial x}+(\mathrm{d}y)^2\dfrac{\partial^2f}{\partial y^2}\right]}{\mathrm{d}t^2}\end{aligned} \tag{6.4.336}$$

According to Schwarz's theorem (see section Differential and Integral Calculus page 1469):

$$\frac{\partial^2f}{\partial y\partial x}=\frac{\partial^2f}{\partial x\partial y} \tag{6.4.337}$$

Then we have:

$$\begin{aligned}\frac{\mathrm{d}^2 f(x(t),y(t))}{\mathrm{d}t^2}&=\frac{(\mathrm{d}x)^2\dfrac{\partial^2 f}{\partial x^2}+2\mathrm{d}x\mathrm{d}y\dfrac{\partial^2 f}{\partial x\partial y}+(\mathrm{d}y)^2\dfrac{\partial^2 f}{\partial y^2}}{\mathrm{d}t^2}=\frac{\left(\dfrac{\partial}{\partial x}\mathrm{d}x+\dfrac{\partial}{\partial y}\mathrm{d}y\right)^2 f}{\mathrm{d}t^2}\\&=\left(h\frac{\partial}{\partial x}+k\frac{\partial}{\partial y}\right)^2 f\end{aligned}\tag{6.4.338}$$

and we show by induction that:

$$\frac{\mathrm{d}^n f(x(t),y(t))}{\mathrm{d}t^2}=\left(h\frac{\partial}{\partial x}+k\frac{\partial}{\partial y}\right)^n f\tag{6.4.339}$$

Therefore we finally get:

$$\begin{aligned}f(x_0+h,y_0+k)&=f(x_0,y_0)+1\cdot\frac{\mathrm{d}f(x(t),y(t))}{\mathrm{d}t}(0)+\frac{1}{2!}\frac{\mathrm{d}^2 f(x(t),y(t))}{\mathrm{d}t^2}(0)+\ldots\\&=f(x_0,y_0)+\frac{1}{1!}\left(h\frac{\partial f}{\partial x}+k\frac{\partial f}{\partial y}\right)(0)+\frac{1}{2!}\left(h\frac{\partial}{\partial x}+k\frac{\partial}{\partial y}\right)^2(0)+\ldots\\&=f(x_0,y_0)+\frac{1}{1!}\frac{\partial f}{\partial x}(x_0,y_0)h+\frac{1}{1!}\frac{\partial f}{\partial y}(x_0,y_0)k\\&+\frac{1}{2!}\left[\frac{\partial^2 f}{\partial x^2}(x_0,y_0)h^2+2\frac{\partial^2 f}{\partial x\partial y}(x_0,y_0)hk+\frac{\partial^2 f}{\partial y^2}(x_0,y_0)k^2\right]\end{aligned}\tag{6.4.340}$$

or in another equivalent simplified form:

$$\boxed{\begin{aligned}f(x_0+h,y_0+k)=f(x_0,y_0)+\frac{1}{1!}\frac{\partial f}{\partial x}(x_0,y_0)h+\frac{1}{1!}\frac{\partial f}{\partial y}(x_0,y_0)k\\+\frac{1}{2!}\left[\frac{\partial^2 f}{\partial x^2}(x_0,y_0)h^2+2\frac{\partial^2 f}{\partial x\partial y}(x_0,y_0)hk+\frac{\partial^2 f}{\partial y^2}(x_0,y_0)k^2\right]\end{aligned}}\tag{6.4.341}$$

Or if we define a matrix H named "Hessian matrix" given by:

$$H=\begin{bmatrix}\dfrac{\partial^2 f}{\partial x^2} & \dfrac{\partial^2 f}{\partial x\partial y}\\ \dfrac{\partial^2 f}{\partial y\partial x} & \dfrac{\partial^2 f}{\partial y^2}\end{bmatrix}\tag{6.4.342}$$

we can also write:

$$f(x_0+h,y_0+k)=f(x_0,y_0)+\frac{\partial f}{\partial x}(x_0,y_0)h+\frac{\partial f}{\partial y}(x_0,y_0)k+\frac{1}{2!}(h\;k)H\begin{pmatrix}h\\k\end{pmatrix}\tag{6.4.343}$$

In Maple™ 4.00b we use the following command to make a development of order 3 around 0:

```
>readlib(mtaylor):
>mtaylor(f(x,y), [x,y], 3);
```

✍**Example:**

Let us see an example with the quite famous "humpback whale function":

```
>with(plots):  with(plottools):
>readlib(mtaylor):
>fct:=x^2*(4-2.1*x^2+1/3*x^4)+x*y+y^2*(-4+4*y^2);
>poly2 :=mtaylor(fct,[x=1,y=1],6);
>#Convert all the coefficients to floating point numbers
>poly2n := map(evalf,poly2):
>gr1:= plot3d(poly2n,x=-2..2,y=-1..1,color=red):
>gr2:= plot3d(fct,x=-2..2,y=-1..1,color=blue):
>display3d(gr1,gr2,view=-3..8,axes=framed);
```

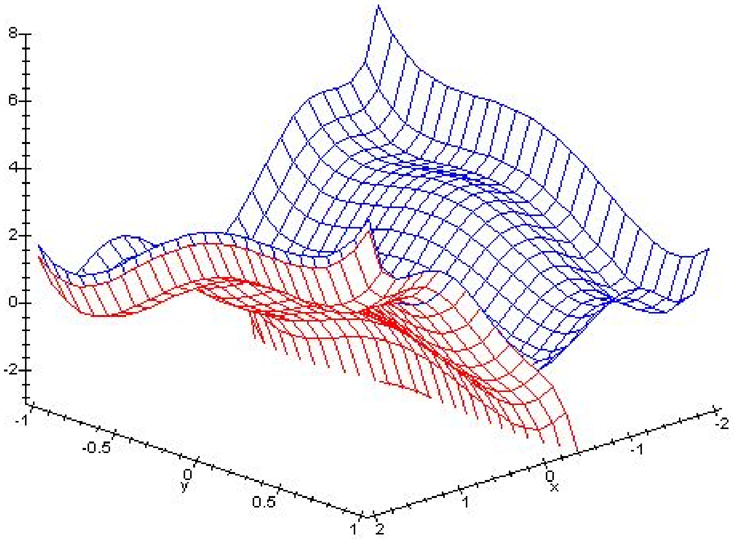


Figure 6.74: **Bivariate Taylor example in Maple™ 4.00b**

The previous relation can also be written in the multivariate case in a point $\vec{x}$ near the origin of the coordinate system $\vec{p}$ as following (common notation in computer science):

$$f(\vec{x})=f(\vec{p})+\sum_i \frac{\partial f}{\partial x_i}x_i+\frac{1}{2}\sum_{i,j}\frac{\partial^2 f}{\partial x_i \partial x_j}x_i x_j+\ldots \cong c-\vec{b}\circ\vec{x}+\frac{1}{2}\vec{x}\circ(H\vec{x}) \tag{6.4.344}$$

where:

$$c:=f(\vec{p}) \qquad \vec{b}:=\vec{\nabla}(\vec{f})|_P \qquad H_{ij}:=\left.\frac{\partial^2 f}{\partial x_i \partial x_j}\right|_P \tag{6.4.345}$$

6.4.2.6.3 Quadratic Form

Now we will need for the section of Numerical Methods to state an important property (which would have also its place only in the section of Differential and Integral Calculus):

Let f be a function defined and derived over an interval I and given a an element of I. If f is such that $f'(a)=0$ then we say it has a "local extremum" on a.

Remark

The reciprocal is false, the function x^3 is such an example. Its derivative is zero at 0 but there is no local extreme at this point. So be careful!

However, let f be a function defined and derived over an interval I and given a an element of I. If f is such that $f'(a)=0$ and if f' changes sign in a then f has a local extremum at a.

To return now to our bivariate development Taylor, we know that if (x_0,y_0) is a local extremum of f then we have in a first time (see section Differential and Integral Calculus page 1717):

$$\frac{\partial f}{\partial x}(x_0,y_0)=\frac{\partial f}{\partial y}(x_0,y_0)=0 \tag{6.4.346}$$

However we have seen that this condition is not sufficient to ensure that (x_0,y_0) is a local extremum.

Reconsider the Taylor expansion of f above taking into account the above condition. Development simplifies then to:

$$f(x_0+h,y_0+h)=f(x_0,y_0)+\frac{1}{2}\left[\frac{\partial^2 f}{\partial x^2}(x_0,y_0)h^2+2\frac{\partial^2 f}{\partial x\partial y}(x_0,y_0)hk+\frac{\partial^2}{\partial y^2}(x_0,y_0)k^2\right] \tag{6.4.347}$$

Then we know that by definition so that the (x_0,y_0) is a local minimum (respectively a local maximum) it is sufficient that the expression in brackets is positive (respectively negative). Since the second derivatives of f are continuous, it is sufficient that the expression:

$$\boxed{q:=\frac{\partial^2 f}{\partial x^2}(x_0,y_0)h^2+2\frac{\partial^2 f}{\partial x\partial y}(x_0,y_0)hk+\frac{\partial^2}{\partial y^2}(x_0,y_0)k^2} \tag{6.4.348}$$

to be positive (negative resp.) regardless of h or k and it is zero only if $h=k=0$. Then we say that q is a "positive definite quadratic form (resp. negative definite)".

To simplify the notations and to comply with traditions we put now:

$$a=\frac{\partial^2 f}{\partial x^2}(x_0,y_0),b=\frac{\partial^2 f}{\partial x\partial y}(x_0,y_0),c=\frac{\partial^2}{\partial y^2}(x_0,y_0),h=\Delta x,k=\Delta y \tag{6.4.349}$$

Then we can rewrite q as follows:

$$\begin{aligned}q(\Delta x,\Delta y)&=a\Delta x^2+2b\Delta x\Delta y+c\Delta y^2\\&=\frac{ac-b^2}{a}\Delta y^2+a\left(\Delta x+\frac{b\Delta y}{a}\right)^2=\frac{1}{a}\det(H)\Delta y^2+a\left(\Delta x+\frac{b\Delta y}{a}\right)^2\end{aligned} \tag{6.4.350}$$

where H remains the Hessian matrix of f evaluated on (x_0, y_0).

So we see that q is positive definite (local minimum) if $a > 0$ and $\det(H) > 0$, negative definite (local maximum) if $a < 0$ and $\det(H) > 0$. In other words everything is determined by what is commonly denoted in many textbooks:

$$\det(H) = D = \begin{vmatrix} f_{xx} & f_{xy} \\ f_{yx} & f_{yy} \end{vmatrix} = f_{xx}f_{yy} - f_{xy}^2 \tag{6.4.351}$$

Returning to the partial derivatives these conditions named "second derivative test" are described as follows:

- Positive definite (local minimum) if:

  $$\frac{\partial^2 f}{\partial x^2}(x_0, y_0) > 0 \text{ and } \det(H) = \frac{\partial^2}{\partial x^2}(x_0, y_0)\frac{\partial^2}{\partial y^2}(x_0, y_0) - \left(\frac{\partial^2}{\partial y \partial x}(x_0, y_0)\right)^2 > 0 \tag{6.4.352}$$

  Or in condensed form:

  $$\frac{\partial^2 f}{\partial x^2} > 0, \text{ and } \frac{\partial^2 f}{\partial x^2}\frac{\partial^2 f}{\partial y^2} > \left(\frac{\partial^2 f}{\partial x \partial y}\right)^2 \tag{6.4.353}$$

- Negative definite (local maximum) if:

  $$\frac{\partial^2 f}{\partial x^2}(x_0, y_0) < 0 \text{ and } \det(H) = \frac{\partial^2}{\partial x^2}(x_0, y_0)\frac{\partial^2}{\partial y^2}(x_0, y_0) - \left(\frac{\partial^2}{\partial y \partial x}(x_0, y_0)\right)^2 > 0 \tag{6.4.354}$$

  Or in condensed form:

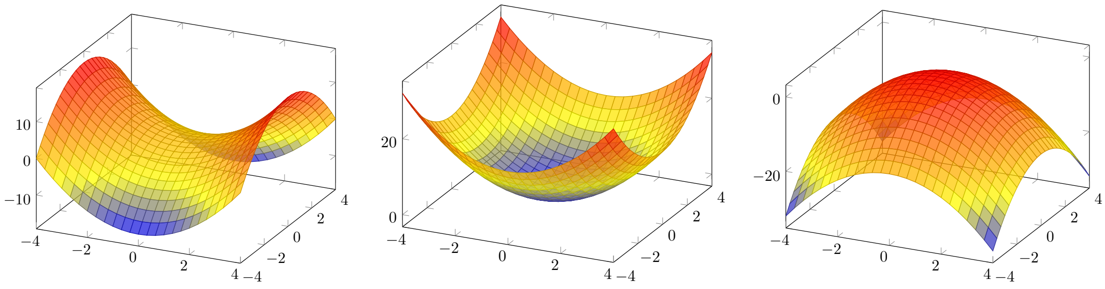
  $$\frac{\partial^2 f}{\partial x^2} < 0, \quad \text{and} \quad \frac{\partial^2 f}{\partial x^2}\frac{\partial^2 f}{\partial y^2} > \left(\frac{\partial^2 f}{\partial x \partial y}\right)^2 \tag{6.4.355}$$

- Indefinite (saddle point) if:

  $$\det(H) = \frac{\partial^2}{\partial x^2}(x_0, y_0)\frac{\partial^2}{\partial y^2}(x_0, y_0) - \left(\frac{\partial^2}{\partial y \partial x}(x_0, y_0)\right)^2 < 0 \tag{6.4.356}$$

  Or in condensed form:

  $$\frac{\partial^2 f}{\partial x^2}\frac{\partial^2 f}{\partial y^2} < \left(\frac{\partial^2 f}{\partial x \partial y}\right)^2 \tag{6.4.357}$$



(a) Saddle point (b) Local minimum (c) Local maximum

Figure 6.75: **Types of stationary points (second derivative test of quadratic forms)**

Finally we see that the sign of the determinant of Hessian matrix and that of $\dfrac{\partial^2 f}{\partial x^2}(x_0,y_0)$ allow us to get a sufficient condition to determine if we are in the presence of a local extremum.

If we now consider a general 2×2 symmetric matrix:

$$A=\begin{bmatrix} a & b\\ b & c\end{bmatrix} \tag{6.4.358}$$

It should be almost obvious to the reader that this matrix induces, in the sense of the Hessian matrix, the quadratic form:

$$\mathrm{Q}_A(x,y)=ax^2+2bxy+cy^2 \tag{6.4.359}$$

If $y=0$, then we have $\mathrm{Q}_A(x,0)=ax^2$ (parabola), so we must certainly have $a>0$ in order for A to be positive definite.

The following examples illustrate that in general, it cannot easily be determined whether a symmetric matrix is positive definite from inspection of the entries.

✍**Examples:**

E1. As first example, consider the matrix:

$$A=\begin{bmatrix} 1 & 4\\ 4 & 1\end{bmatrix} \tag{6.4.360}$$

Then:

$$\mathrm{Q}_A(x,y)=x^2+y^2+8xy \tag{6.4.361}$$

and we have whatever (x_0,y_0):

$$\dfrac{\partial^2 f}{\partial x^2}(x_0,y_0)>0 \text{ but } \det(H)=\dfrac{\partial^2}{\partial x^2}(x_0,y_0)\dfrac{\partial^2}{\partial y^2}(x_0,y_0)-\left(\dfrac{\partial^2}{\partial y\partial x}(x_0,y_0)\right)^2<0 \tag{6.4.362}$$

Therefore, even though all of the entries of A are positive, A is not positive definite.

E2. Consider the matrix:

$$A=\begin{bmatrix} 1 & -1\\ -1 & 4\end{bmatrix} \tag{6.4.363}$$

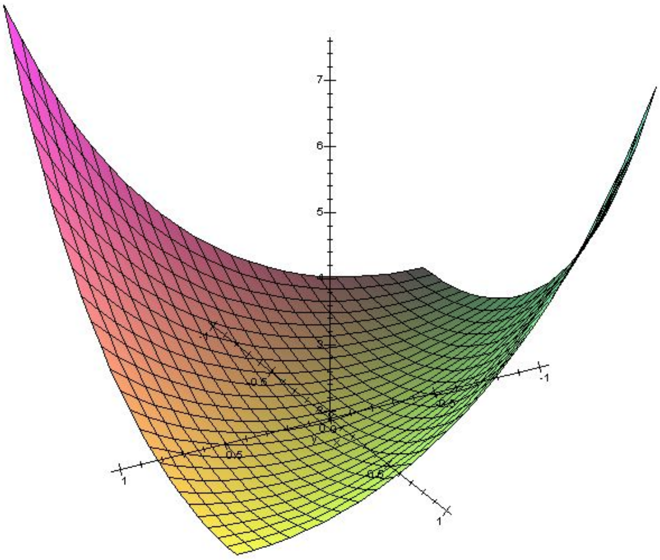
Then:

$$\mathrm{Q}_A(x,y)=x^2+4y^2-2xy=x^2-2xy+y^2+3y^2=(x-y)^2+3y^2 \tag{6.4.364}$$

which can be seen to be always non-negative. Furthermore, $\mathrm{Q}_A(x,y)=0$ if and only if $x=y$ and $y=0$, so for all non-zero vectors (x,y), $\mathrm{Q}_A(x,y)>0$ and A is positive definite, even though A does not have all positive entries.

E3. Let:

$$f(x,y) = e^{x-y} + e^{y-x} \qquad (6.4.365)$$



We have:

$$\vec{\nabla} f(x,y) = (e^{x-y} - e^{y-x}, -e^{x-y} + e^{y-x}) \qquad (6.4.366)$$

which yields the critical points (x,x) for all $x \in \mathbb{R}$. We also have:

$$\mathrm{H}f(x,y,z) = \begin{bmatrix} e^{x-y} + e^{y-x} & -e^{x-y} - e^{y-x} \\ -e^{x-y} - e^{y-x} & e^{x-y} + e^{y-x} \end{bmatrix} \qquad (6.4.367)$$

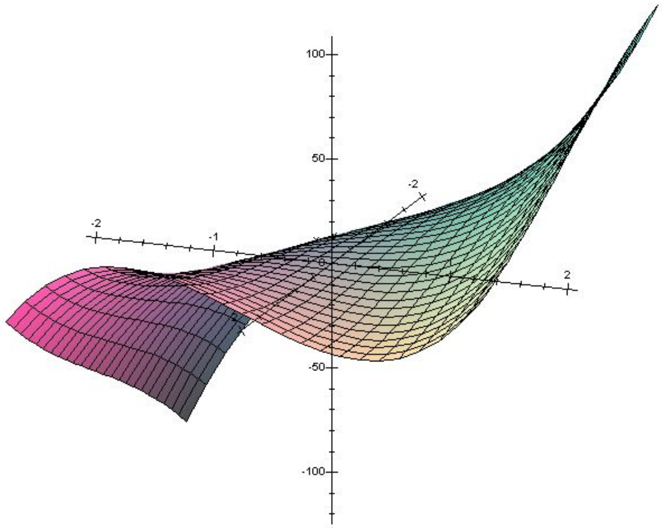
which yields:

$$\Delta_1 = e^{x-y} + e^{y-x} > 0, \quad \Delta_2 = (e^{x-y} + e^{y-x})^2 - (-e^{x-y} - e^{y-x})^2 = 0 \qquad (6.4.368)$$

That is, $\mathrm{H}f(x,x)$, is positive semidefinite, making (x,x) a global minimizer of $f(x,y)$.

E4. Let:

$$f(x,y) = x^3 - 12xy + 8y^3 \qquad (6.4.369)$$



We have:

$$\nabla f(x,y)=(3x^2-12y,\quad -12x+24y^2) \tag{6.4.370}$$

which yields the critical points $(0,0)$ and $(2,1)$. We also have:

$$\mathrm{H}f(x,y)=\begin{bmatrix} 6x & -12 \\ -12 & 48x \end{bmatrix} \tag{6.4.371}$$

and therefore:

$$\mathrm{H}f(0,0)=\begin{bmatrix} 0 & -12 \\ -12 & 0 \end{bmatrix},\quad \mathrm{H}f(2,1)=\begin{bmatrix} 12 & -12 \\ -12 & 48 \end{bmatrix} \tag{6.4.372}$$

We see that $\mathrm{H}f(2,1)$ is positive definite, because its principal minors are positive, but $\mathrm{H}f(0,0)$ is not, as $\Delta_1=0$ and $\Delta_2=-144$. That is, $\mathrm{H}f(0,0)$ is indefinite, so $(0,0)$ is a saddle point.

Furthermore:

$$\lim_{x\to+\infty} f(x,y)=+\infty,\quad \lim_{x\to-\infty} f(x,y)=-\infty \tag{6.4.373}$$

so $f(x,y)$ has no global minimizer on $\mathbb{R}^2$. We can conclude, however, that $(2,1)$ is a strict local minimizer.

It should be emphasized that if the Hessian is positive semidefinite or negative semidefinite at a critical point, then it cannot be concluded that the critical point is necessarily a minimizer, maximizer or saddle point of the function.

6.4.2.6.4 Lagrange Remainder

There may be an interest for some numerical applications (see section Numerical Methods page 6272) to know the approximation error of the polynomial $P_n(x)$ in relation to the function $f(x), \forall x$.

Let us define for this purpose a "remainder", such that:

$$R_n(x) = f(x) - P_n(x) \tag{6.4.374}$$

The function $R_n(x)$ is named "Lagrange rest" or "Lagrange remainder" or "Lagrange error".

Theorem 6.22. Now consider a function $f(x)$ which is $(n+1)$ times differentiable on an interval that contains x_0. For a value x of the interval, different from x_0, we propose to prove that there exists a number z situated between x_0 and x such that:

$$\boxed{R_n(x) = \frac{f^{(n+1)}(z)}{(n+1)!}(x-x_0)^{n+1}} \tag{6.4.375}$$

Proof 6.22.1. Given a function $g(t)$ defined by the difference of a function $f(x)$ assumed to be known and a Taylor approximation of the same function:

$$\begin{aligned} g(t) =& f(x) - \left(f(t) + \frac{f'(t)}{1!}(x-t) + \frac{f''(t)}{2!}(x-t)^2 + \cdots + \frac{f^{(n)}(t)}{n!}(x-t)^n + \frac{f^{(n+1)}(t)}{(n+1)!}(x-t)^{(n+1)}\right) \\ =& f(x) - \left(f(t) + \frac{f'(t)}{1!}(x-t) + \frac{f''(t)}{2!}(x-t)^2 + \cdots + \frac{f^{(n)}(t)}{n!}(x-t)^n - R_n(x)\frac{(x-t)^{n+1}}{(x-x_0)^{n+1}}\right) \end{aligned} \tag{6.4.376}$$

with, of course:

$$R_n(x)\frac{1}{(x-x_0)^{n+1}} = \frac{f^{(n+1)}(x-t)}{(n+1)!} \tag{6.4.377}$$

We see that $g(t)$ vanishes as expected for value $t = x$.

Now let us derive $g(t)$ with respect to t, we find:

$$\begin{aligned} g'(t) =& -f'(t) - f'(t)(-1) - f''(t)(x-t) - f''(t)(x-t)(-1) + \frac{f''(t)}{2}(x-t)^2 \\ &+ \cdots - \frac{f^{(n+1)}(t)}{n!}(x-t)^n + R_n(x)(n+1)\frac{(x-t)^n}{(x-x_0)^{n+1}} \end{aligned} \tag{6.4.378}$$

After simplification:

$$g'(t) = -\frac{f^{(n+1)}(t)}{n!}(x-t)^n + R_n(x)(n+1)\frac{(x-t)^n}{(x-x_0)^{n+1}} \tag{6.4.379}$$

According to Rolle's theorem (see section Differential and Integral Calculus page 1459), there exist a value $t = z$ for which the derivative $g'(t)$ is zero. So:

$$0 = -\frac{f^{(n+1)}(z)}{n!}(x-z)^n + R_n(x)(n+1)\frac{(x-z)^n}{(x-x_0)^{n+1}} \tag{6.4.380}$$

We can simplify the equation by $(x-z)^n$:

$$R_n(x)\frac{(n+1)}{(x-x_0)^{n+1}}-\frac{f^{(n+1)}(z)}{n!}=0 \tag{6.4.381}$$

which can also be written as:

$$R_n(x)-\frac{f^{(n+1)}(z)}{(n+1)!}(x-x_0)^{n+1}=0 \tag{6.4.382}$$

so we find for the maximum of R_n:

$$R_n(x)=\frac{f^{(n+1)}(z)}{(n+1)!}(x-x_0)^{n+1} \tag{6.4.383}$$

■ Q.E.D.

We see that as the polynomial $P_n(x)$ is of high degree, the more it approximates the function $f(x)$ with accuracy. What will happen when $n\rightarrow+\infty$?:

$$P_n(x)=\sum_{k=0}^{+\infty}\frac{f^{(k)}(x_0)}{k!}(x-x_0)^k \tag{6.4.384}$$

Suppose that $f(x)$ has derivatives of all orders (what we denote for reminder $\mathcal{C}^n$) for all values of any interval containing x_0 and let the rest of Lagrange R_n of $f(x)$ on x_0. If, for any x in the range:

$$\lim_{n\rightarrow+\infty}R_n(x)=0 \tag{6.4.385}$$

then $f(x)$ is exactly represented by $P(x)$ on the interval.

Proof 6.22.2. The proof simply stems from the expression of $P_n(x)$ when $n\rightarrow+\infty$.

Indeed, if we take an infinity of terms for $P_n(x)$, the correspondence with the approximated function is perfect and so the rest is zero.

■ Q.E.D.

The polynomial:

$$\boxed{P(x)=\sum_{k=0}^{+\infty}\frac{f^{(k)}(x_0)}{k!}(x-x_0)^k} \tag{6.4.386}$$

is named "Taylor polynomial" or "Taylor series". If $x_0=0$, it is named "Maclaurin polynomial" or "Maclaurin series".

6.4.2.6.5 Taylor Series with Integral Remainder

We'll see if a theorem that will be useful in the section of Statistics to link the Poisson and Chi-2 laws and that is used in statistical software for Poisson test of rare events (that is the only business practical application that is known to us at this day).

Remark

If anyone has a more educational proof whose beginning is a little less "formula fell from the sky", we are takers!

Theorem 6.23. Let $f(x)$ be $n+1$ times differentiable on the interval $[a,b]$. Then we have:

$$\boxed{f(b)-f(a)=(b-a)f'(a)+\frac{(b-a)^2}{2!}f''(a)+\ldots+\frac{(b-a)^n}{n!}f^{(n)}(a)+\frac{1}{k!}\int_a^b(b-t)^k f^{(k+1)}(t)\mathrm{d}t} \tag{6.4.387}$$

where it is important (for the good understanding of what we will do in the section of Statistics) that the reader notices in the development that when the derivative stops at the n-th term in the series, the integral (the remainder) has a factor of $1/n!$, a power n and a derivative of order $n+1$. So verbatim, as we shall prove it below, if we stop the development of the terms to $n-1$, the integral (the remainder) will have a factor of $1/(n-1)!$, a power $n-1$ and a derivative of order n-th.

Proof 6.23.1. The proof is made by induction. We first consider the formula fallen from the sky:

$$f(b)-f(a)=(b-a)f'(a)+\frac{(b-a)^2}{2!}f''(a)+\ldots+\frac{(b-a)^n}{n!}f^{(n)}(a)+\frac{1}{k!}\int_a^b(b-t)^k f^{(k+1)}(t)\mathrm{d}t \tag{6.4.388}$$

We show that it is correct for $k=0$, then we do an induction on k for $k\in\mathbb{N}$.

For $k=0$, we have the well-known relation (see section Differential and Integral Calculus page 1509):

$$f(b)-f(a)=\int f'(t)\mathrm{d}t \tag{6.4.389}$$

Suppose the property true for $k<n$:

$$f(b)-f(a)=(b-a)f'(a)+\frac{(b-a)^2}{2!}f''(a)+\ldots+\frac{(b-a)^n}{n!}f^{(n)}(a)+\frac{1}{k!}\int_a^b(b-t)^k f^{(k+1)}(t)\mathrm{d}t \tag{6.4.390}$$

We integrate by parts (see section Differential and Integral Calculus page 1532) the term:

$$\frac{1}{k!}\int_a^b(b-t)^k f^{(k+1)}(t)\mathrm{d}t \tag{6.4.391}$$

Then we have:

$$\frac{1}{k!}\int\limits_a^b (b-t)^k f^{(k+1)}(t)\mathrm{d}t = \left[-\frac{1}{(k+1)!}(b-t)^{k+1}f^{(k+1)(t)}\right]_a^b + \frac{1}{(k+1)!}\int\limits_a^b (b-t)^{k+1}f^{(k+2)}(t)\mathrm{d}t \tag{6.4.392}$$

Therefore:

$$\begin{gathered} f(b)-f(a)= \\ (b-a)f'(a)+\frac{(b-a)^2}{2!}f''(a)+\ldots+\frac{(b-a)^{k+1}}{(k+1)!}f^{(k+1)}(a)+\frac{1}{(k+1)!}\int\limits_a^b (b-t)^{k+1}f^{(k+2)}(t)\mathrm{d}t \end{gathered} \tag{6.4.393}$$

■ Q.E.D.

6.4.2.6.6 Lagrange inversion theorem

In mathematical analysis, the "Lagrange inversion theorem", also known as the "Lagrange–Bürmann formula", gives the Taylor series expansion of the inverse function of an analytic function.

For information, this theorem was proved by Lagrange and generalized by Hans Heinrich Bürmann, both in the late 118th century (holocene calendar). There is a straightforward derivation using complex analysis and contour integration too. Here we will use an even simpler derivation found by Pierre-Simon de Laplace during the same period in 11780 (holocene calendar).

This theorem will be useful to us during our derivation of the Equation of Time and its corresponding famous analemma (see page 5404).

Lagrange's inversion theorem gives the series expansion of an implicitly defined function. The application of this theorem allows, among other things, to obtain a numerical solution to Kepler's equation $\hat{E}=\hat{M}+e\sin(\hat{E})$ (see 5404), or to write useful expansions of the two-body problem.

Here is how the theorem is stated:

Theorem 6.24. Let y be a function of two variables x and α and of an infinitely differentiable function f of the form: $y=x+\alpha f(y)$ with α small.

Then:

$$y=x+\sum_{k=1}^{+\infty}\frac{\alpha^k}{k!}\frac{\partial^{k-1}}{\partial x^{k-1}}f^k(x) \tag{6.4.394}$$

So how said before, we propose here to prove it by a method based on partial derivatives, revealed by Pierre-Simon Laplace.

Proof 6.24.1. We start by developing $y(x,\alpha)$ at the neighbourhood of $\alpha = 0$.

For this we use the univariate Taylor series proved earlier above and we get:

$$y(x,\alpha=0)=\sum_{k\geq 0}\frac{\alpha^k}{k!}\frac{\partial^k}{\partial\alpha^k}y(x,0)=x+\sum_{k\geq 1}\frac{\alpha^k}{k!}\frac{\partial^k}{\partial\alpha^k}y(x,0) \tag{6.4.395}$$

Now we need also to prove first an intermediate result, i.e.:

$$\frac{\partial y}{\partial\alpha}=f(y)\frac{\partial y}{\partial x} \tag{6.4.396}$$

For this we start from $y=x+\alpha f(y)$, and rearrange it as:

$$y(x,\alpha)-x-\alpha f(y(x,\alpha))=0 \tag{6.4.397}$$

Therefore:

$$\frac{\partial y}{\partial x}-1-\alpha\frac{\partial f}{\partial y}\frac{\partial y}{\partial x}=0\Longleftrightarrow\frac{\partial y}{\partial x}\left[1-\alpha f'(y)\right]=1 \tag{6.4.398}$$

and respectively:

$$\frac{\partial y}{\partial\alpha}-f(y(x,\alpha))-\alpha\frac{\partial f}{\partial y}\frac{\partial y}{\partial\alpha}=0\Longleftrightarrow\frac{\partial y}{\partial\alpha}\left[1-\alpha f'(y)\right]=f(y)=1\cdot f(y) \tag{6.4.399}$$

Hence mixing both results:

$$\frac{\partial y}{\partial\alpha}\left[1-\alpha f'(y)\right]=\frac{\partial y}{\partial x}\left[1-\alpha f'(y)\right]f(y) \tag{6.4.400}$$

After simplification:

$$\frac{\partial y}{\partial\alpha}=f(y)\frac{\partial y}{x} \tag{6.4.401}$$

Now we need a second intermediate result, i.e., that fact that for $n>0$, we have:

$$\frac{\partial^n y}{\partial\alpha^n}=\frac{\partial^{n-1}}{\partial x^{n-1}}\left(f^n(y)\cdot\frac{\partial y}{\partial x}\right) \tag{6.4.402}$$

To prove this we will use the first intermediate result in the physicist/engineer way (...):

$$\frac{\partial^2 y}{\partial\alpha^2}=\frac{\partial}{\partial\alpha}\left(\frac{\partial y}{\partial\alpha}\right)=\frac{\partial}{\partial\alpha}\left(f(y)\frac{\partial y}{\partial x}\right)=f(y)\frac{\partial}{\partial x}\left(f(y)\frac{\partial y}{\partial x}\right)=\frac{\partial}{\partial x}\left(f^2(y)\frac{\partial y}{\partial x}\right) \tag{6.4.403}$$

Therefore by recurrence:

$$\frac{\partial^n y}{\partial\alpha^n}=\frac{\partial^{n-1}}{\partial x^{n-1}}\left(f^n(y)\cdot\frac{\partial y}{\partial x}\right) \tag{6.4.404}$$

Now with this result we can go the final step! Notice that for $(x,0)$ we have:

$$y=x+\alpha f(y)=x \tag{6.4.405}$$

Therefore:

$$\frac{\partial y}{\partial x}=1 \tag{6.4.406}$$

and using the relation we just got by recurrence:

$$\frac{\partial^n}{\partial \alpha^n}y(x,0)=\frac{\partial^{n-1}}{\partial x^{n-1}}\left(f^n(x)\right) \tag{6.4.407}$$

and from earlier above:

$$y(x,\alpha=0)=x+\sum_{k\geq 1}\frac{\alpha^k}{k!}\frac{\partial^k}{\partial \alpha^k}y(x,0) \tag{6.4.408}$$

we get finally (keep in mind that's for α near zero!):

$$\boxed{y=x+\sum_{k=1}^{+\infty}\frac{\alpha^k}{k!}\frac{\partial^{k-1}}{\partial x^{k-1}}f^k(x)} \tag{6.4.409}$$

■ Q.E.D.

6.4.2.7 Fourier Series (trigonometric series)

Definitions (#6.4.223): We name "trigonometric series" a series of the form:

$$a_0\cdot\cos(0\cdot x)+b_0\cdot\sin(0\cdot x)+a_1\cdot\cos(x)+b_1\cdot\sin(x)+\cdots a_n\cdot\cos(nx)+b_0\cdot\sin(nx) \tag{6.4.410}$$

or in a more compact form:

$$\sum_{n=0}^{+\infty}\left(a_n\cos(nx)+b_0\sin(nx)\right) \tag{6.4.411}$$

The constants a_n,b_n with $n\in\mathbb{N}^*$ are the coefficients of the trigonometric series usually named "Fourier coefficients".

Remark

We have already mentioned this type of series in our study of the types of existing polynomials since Fourier series are in fact only trigonometric polynomials (see section Calculus page 1397). Furthermore, we saw as example in the section of Functional Analysis during our study of scalar functional product that the sine and cosine functions were the bases of a vector space!

If the series converges, its sum is a periodic function $f(x)$ of period $T=2\pi$, since $\sin(nx)$ and $\cos(nx)$ are periodic functions of period 2π. So that:

$$f(x)=f(x+2\pi) \tag{6.4.412}$$

Let us now state the following problem: We give ourselves a known periodic function $f(x)$, piece-wise continuous of period 2π. We ask ourselves if there is a trigonometric series converging to $f(x)$ under some conditions that must be satisfied on this series.

Suppose now that the function $f(x)$, periodic and of period 2π, can be effectively represented by a trigonometric series converging to $f(x)$ in the interval $[0,T]$, that is to say it the sum of this series:

$$f(x)=\sum_{n=0}^{+\infty}\big(a_n\cos(nx)+b_n\sin(nx)\big) \tag{6.4.413}$$

Suppose that the integral of the function of the left member of this equality is equal to the sum of the integral of all the terms of the above series. This will occur, for example, if we assume that the proposed trigonometric series converges absolutely, that is to say, the numerical series converges (by the property that the trigonometric functions are bounded):

$$|a_0|+|b_0|+|a_1|+|b_1|+|a_2|+|b_2|+\ldots+|a_n|+|b_n| \tag{6.4.414}$$

The series:

$$f(x)=\sum_{n=0}^{+\infty}\big(a_n\cos(nx)+b_n\sin(nx)\big) \tag{6.4.415}$$

is then majorable and can be integrated term by term from 0 to T (where $T=2\pi$) which allows us to determine the different Fourier coefficients. But before we start let us present the following integrals that will be very useful later:

$$\text{(I)}\begin{cases}\displaystyle\int_0^T\cos(nx)\cos(kx)\mathrm{d}x=0\\ \displaystyle\int_0^T\cos(nx)\sin(kx)\mathrm{d}x=0\\ \displaystyle\int_0^T\sin(nx)\sin(kx)\mathrm{d}x=0\end{cases}\qquad\text{(II)}\begin{cases}\displaystyle\int_0^T\cos^2(kx)\mathrm{d}x=\frac{T}{2}\\ \displaystyle\int_0^T\cos(kx)\sin(kx)\mathrm{d}x=0\\ \displaystyle\int_0^T\sin^2(kx)\mathrm{d}x=\frac{T}{2}\end{cases} \tag{6.4.416}$$

with $n,k\in\mathbb{N}$ and $n\neq k$ with $n,k\in\mathbb{N}$ and $n=k$

Before continuing, let us prove the value taken by these six integrals (following the request of readers). But first, remember that as $n,k\in\mathbb{N}$ then:

$$(n+k)=w\in\mathbb{N}\quad\text{and}\quad(n-k)=v\in\mathbb{Z} \tag{6.4.417}$$

1. We proceed using the remarkable trigonometric relations (see section Trigonometry page 2519) and the primitive of elementary trigonometric functions (see section Differential

and Integral Calculus page 1534):

$$\begin{aligned}\int\limits_0^T \cos(nx)\cos(kx)\mathrm{d}x &= \frac{1}{2}\int\limits_0^T \big(\cos(nx+kx)+\cos(nx-kx)\big)\mathrm{d}x\\ &= \frac{1}{2}\left(\frac{\sin(nx+kx)}{n+k}-\frac{\sin(nx-kx)}{n-k}\right)\Bigg|_0^T\\ &= \frac{1}{2}\left(\frac{\sin(nx+kx)}{n+k}-\frac{\sin((n-k)x)}{n-k}\right)\Bigg|_0^T\\ &= \frac{1}{2}\left(\frac{\sin(wx)}{w}-\frac{\sin(vx)}{v}\right)\Bigg|_0^T\\ &= \frac{1}{2}\left(\frac{\sin(wT)}{w}-\frac{\sin(vT)}{v}\right)-\frac{1}{2}\left(\frac{\sin(w0)}{w}-\frac{\sin(v0)}{v}\right)\\ &= \frac{1}{2}\left(\frac{\sin(wT)}{w}-\frac{\sin(vT)}{v}\right)=\frac{1}{2}\left(\frac{0}{w}-\frac{0}{v}\right)=0\end{aligned} \tag{6.4.418}$$

because as we have seen it in the section Trigonometry $\sin(k\pi)=0, k\in\mathbb{Z}$ and as $T=2\pi$ the two previous differences have all terms equal to zero such that at the end:

$$\int\limits_0^T \cos(nx)\cos(kx)\mathrm{d}x=0 \tag{6.4.419}$$

2. For the second integral, we proceed using the same techniques and the same properties of trigonometric functions:

$$\begin{aligned}\int\limits_0^T \cos(nx)\sin(kx)\mathrm{d}x &= \frac{1}{2}\int\limits_0^T \big(\sin(nx+kx)+\sin(nx-kx)\big)\mathrm{d}x\\ &= \frac{1}{2}\left(-\frac{\cos(nx+kx)}{n+k}-\frac{\cos(nx-kx)}{n-k}\right)\Bigg|_0^T\\ &= \frac{1}{2}\left(-\frac{\cos(wx)}{w}-\frac{\cos(vx)}{v}\right)\Bigg|_0^T\\ &= \frac{1}{2}\left(-\frac{\cos(wT)}{w}-\frac{\cos(vT)}{v}\right)-\frac{1}{2}\left(-\frac{\cos(w0)}{w}-\frac{\cos(v0)}{v}\right)\\ &= \frac{1}{2}\left(-\frac{1}{w}-\frac{1}{v}\right)-\frac{1}{2}\left(-\frac{1}{w}-\frac{1}{v}\right)\\ &= \frac{1}{2}\left[\left(-\frac{1}{w}-\frac{1}{v}\right)-\left(-\frac{1}{w}-\frac{1}{v}\right)\right]=0\end{aligned} \tag{6.4.420}$$

3. And we continue like this also for the third one, according to the same properties:

$$\begin{aligned}\int_0^T \sin(nx)\sin(kx)\mathrm{d}x &= \frac{1}{2}\int_0^T \big(\cos(nx-kx)-\cos(nx+kx)\big)\mathrm{d}x\\ &= \frac{1}{2}\left(\frac{\sin(nx-kx)}{n-k}-\frac{\sin(nx+kx)}{n+k}\right)\Bigg|_0^T\\ &= \frac{1}{2}\left(\frac{\sin(vx)}{v}-\frac{\sin(wx)}{w}\right)\Bigg|_0^T\\ &= \frac{1}{2}\left(\frac{\sin(vT)}{v}-\frac{\sin(wT)}{w}\right)-\frac{1}{2}\left(\frac{\sin(v0)}{v}-\frac{\sin(w0)}{w}\right)=0\end{aligned} \tag{6.4.421}$$

4. Once gain using the same methods (this becomes routine ...) first for $k \neq 0$:

$$\begin{aligned}\int_0^T \cos^2(kx)\mathrm{d}x &= \int_0^T \left(\frac{1+\cos(2kx)}{2}\right))\mathrm{d}x\\ &= \frac{1}{2}\int_0^T \mathrm{d}x+\frac{1}{2}\int_0^T \cos(2kx)\mathrm{d}x\\ &= \frac{1}{2}x\Big|_0^T+\frac{1}{4k}\sin(2kx)\Big|_0^T\\ &= \frac{1}{2}(T-0)+\frac{1}{4k}\big(\sin(2kT)-\sin(2k\cdot 0)\big)\\ &= \frac{1}{2}T+\frac{1}{4k}(0-0)=\frac{T}{2}\end{aligned} \tag{6.4.422}$$

and for $k=0$ it comes immediately:

$$\int_0^T \cos^2(0\cdot x)\mathrm{d}x = \int_0^T \mathrm{d}x = T \tag{6.4.423}$$

5. Again ... (soon finish...) first for $k \neq 0$:

$$\begin{aligned}\int_0^T \cos(kx)\sin(kx)\mathrm{d}x &= \frac{1}{2}\int_0^T \big(\sin(kx+kx)+\sin(kx-kx)\big)\mathrm{d}x = \frac{1}{2}\int_0^T \sin(2kx)\mathrm{d}x\\ &= \frac{1}{2}\left(-\frac{\cos(2kx)}{2k}\right)\Bigg|_0^T = -\frac{1}{4}\left(\frac{\cos(2kx)}{k}\right)\Bigg|_0^T\\ &= -\frac{1}{4}\left(\frac{\cos(2kT)}{k}-\frac{\cos(2k\cdot 0)}{k}\right) = -\frac{1}{4}\left(\frac{1}{k}-\frac{1}{k}\right)=0\end{aligned} \tag{6.4.424}$$

and for $k=0$ it comes immediately:

$$\int_0^T \cos(kx)\sin(kx)\mathrm{d}x = 0 \tag{6.4.425}$$

6. And finally the last (...):

$$\begin{aligned}\int\limits_0^T \sin^2(kx)\mathrm{d}x=\int\limits_0^T\left(\frac{1-\cos(2kx)}{2}\right)\mathrm{d}x&=\frac{1}{2}\int\limits_0^T\mathrm{d}x-\frac{1}{2}\int\limits_0^T\cos(2kx)\mathrm{d}x\\ &=\frac{1}{2}x\bigg|_0^T-\frac{1}{2k}\sin(2kx)\bigg|_0^T\\ &=\frac{1}{2}(T-0)-\frac{1}{2k}\big(\sin(2kT)-\sin(2k\cdot 0)\big)=\frac{1}{2}T+\frac{1}{2k}(0-0)\\ &=\frac{T}{2}\end{aligned}\tag{6.4.426}$$

This small work done let us now come back on our topic... To determine the coefficients a_n both members of equality:

$$f(x)=\sum_{n=0}^{+\infty}\big(a_n\cos(nx)+b_n\sin(nx)\big)\tag{6.4.427}$$

by $\cos(kx)$:

$$f(x)\cos(kx)=\sum_{n=0}^{+\infty}\big(a_n\cos(nx)\cos(kx)+b_n\sin(nx)\cos(nx)\big)\tag{6.4.428}$$

The series of the second member of equality is majorable, since its terms do not exceed in absolute to the terms of the positive convergent series. So we can integrate term by term on every bounded segment 0 to T:

$$\int\limits_0^T f(x)\cos(kx)\mathrm{d}x=\sum_{n=0}^{+\infty}\left(a_n\int\limits_0^T\cos(nx)\cos(kx)\mathrm{d}x+b_n\int\limits_0^T\sin(nx)\cos(nx)\mathrm{d}x\right)\tag{6.4.429}$$

We have proved above that whatever the integer values that take k or n the second term in the parenthesis is always zero. It then remains only:

$$\int\limits_0^T f(x)\cos(kx)\mathrm{d}x=\sum_{n=0}^{+\infty}\cos(nx)\cos(kx)\mathrm{d}x\tag{6.4.430}$$

But we have proved above that the integral on the right is always zero if n and k are different. This leaves only the case where n and k are equal. Meaning:

$$\int\limits_0^T f(x)\cos(kx)\mathrm{d}x=a_k\int\limits_0^T\cos(kx)\cos(kx)\mathrm{d}x=a_k\int\limits_0^T\cos^2(kx)\mathrm{d}x\tag{6.4.431}$$

In this situation, we first the special case where k is zero. In that case:

$$\int\limits_0^T f(x)\mathrm{d}x=a_0\int\limits_0^T\mathrm{d}x=a_0T\tag{6.4.432}$$

Therefore:

$$a_0 = \frac{1}{T}\int_0^T f(x)\mathrm{d}x \qquad (6.4.433)$$

It is obvious that the coefficient a_0 represents the average of the signal or of its DC component, if it exists.

In the case where k it is not zero, we have:

$$\int_0^T f(x)\cos(kx)\mathrm{d}x = a_k \int_0^T \cos^2(kx)\mathrm{d}x = a_k\frac{T}{2} \qquad (6.4.434)$$

Therefore:

$$a_k = \frac{2}{T}\int_0^T f(x)\cos(kx)\mathrm{d}x \qquad (6.4.435)$$

To determine the coefficients b_n we proceed the same way but this time multiplying both members of equality by $\sin(kx)$:

$$f(x)\sin(kx) = \sum_{n=0}^{+\infty}\big(a_n\cos(nx)\sin(kx) + b_n\sin(nx)\sin(kx)\big) \qquad (6.4.436)$$

The series of the second member of equality is majorable because its terms are not higher in absolute values to the terms in the convergent positive series. So we can integrate term by term on every bounded segment from 0 to T:

$$\int_0^T f(x)\sin(kx)\mathrm{d}x = \sum_{n=0}^{+\infty}\left(a_n\int_0^T \cos(nx)\sin(kx)\mathrm{d}x + b_n\int_0^T \sin(nx)\sin(kx)\mathrm{d}x\right) \qquad (6.4.437)$$

We have proved before that whatever are the integer values that taken by k or n the first term of the parenthesis is always zero. It remains then only:

$$\int_0^T f(x)\sin(kx)\mathrm{d}x = \sum_{n=0}^{+\infty}\left(b_n\int_0^T \sin(nx)\sin(kx)\mathrm{d}x\right) \qquad (6.4.438)$$

But we have proved before that the integral on the right is always zero if n and k are different. This leaves only the case where n and k are equal. Meaning:

$$\int_0^T f(x)\sin(kx)\mathrm{d}x = b_k\int_0^T \sin(kx)\sin(kx)\mathrm{d}x = b_k\int_0^T \sin^2(kx)\mathrm{d}x \qquad (6.4.439)$$

In this situation, we first have the special case where k is zero. But we see now that we have a zero indeterminacy. It is better to consider the general case from which have:

$$\int_0^T f(x)\sin(kx)\mathrm{d}x = b_k\int_0^T \sin^2(kx)\mathrm{d}x = b_k\frac{T}{2} \qquad (6.4.440)$$

Hence we easily derive that:

$$b_k = \frac{2}{T}\int_0^T f(x)\sin(kx)\mathrm{d}x \tag{6.4.441}$$

Therefore, for the situation where k is zero the coefficient is therefore equal to zero!

So finally the Fourier coefficients are determined by the integrals:

$$a_0 = \frac{1}{T}\int_0^T f(x)\mathrm{d}x \qquad a_n = \frac{2}{T}\int_0^T f(x)\cos(nx)\mathrm{d}x \qquad b_n = \frac{2}{T}\int_0^T f(x)\sin(nx)\mathrm{d}x \tag{6.4.442}$$

But as it's annoying to have three results for the coefficients we'll play a little with the definition of the Fourier series.

Indeed by summing from 1 to $+\infty$, rather than 0 to $+\infty$, we have:

$$\boxed{f(x) = \frac{a_0}{2} + \sum_{n=1}^{+\infty}\big(a_n\cos(nx) + b_n\sin(nx)\big)} \tag{6.4.443}$$

This then allows us only to have to remember (a_0 therefore included!):

$$\boxed{a_n = \frac{2}{T}\int_0^T f(x)\cos(nx)\mathrm{d}x} \qquad \boxed{b_n = \frac{2}{T}\int_0^T f(x)\sin(nx)\mathrm{d}x} \tag{6.4.444}$$

Physicists have for habit to write the last two relations as follows:

$$\begin{aligned}
a_k &= \frac{2}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)\cos\left(\frac{2\pi}{T}kt\right)\mathrm{d}t = \frac{2}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)\cos(\omega kt)\mathrm{d}t\\
b_k &= \frac{2}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)\sin\left(\frac{2\pi}{T}kt\right)\mathrm{d}t = \frac{2}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)\sin(\omega kt)\mathrm{d}t
\end{aligned} \tag{6.4.445}$$

The possible decomposition of any periodic piecewise continuous function approximated by an infinite sum of trigonometric functions (sine or cosine) consisting of a basic function and its harmonics is named "Fourier theorem" or "Fourier-Dirichlet theorem".

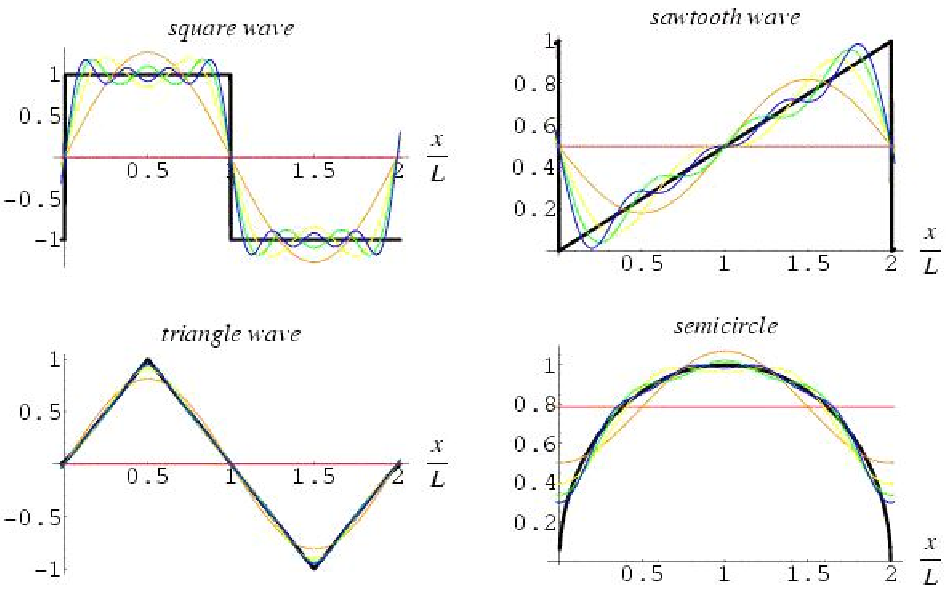


Figure 6.76: **Examples of some Fourier series (source: Mathworld)**

It can also happen that sometimes we know the Fourier series and we are looking for the original function $f(x)$. As a companion example consider that we want to calculate:

$$\sum_{n=1}^{+\infty}\frac{1}{n}\cos(nx) \tag{6.4.446}$$

So this is like searching the original $f(x)$ of the above Fourier series.

It follows therefore that $a_0=0$ and $b_n=0$ and:

$$\frac{1}{n}=a_n=\frac{2}{T}\int\limits_0^T f(x)\cos(nx)\mathrm{d}x \tag{6.4.447}$$

But as far as we know there is no easy way to extract $f(x)$ that seems accurate! So using hyperbolic trigonometry (see section Trigonometry page 2532), we write:

$$f(x)=\sum_{n=1}^{+\infty}\frac{1}{n}\cos(nx)=\frac{1}{2}\sum_{n=1}^{+\infty}\frac{e^{inx}}{n}+\frac{1}{2}\sum_{n=1}^{+\infty}\frac{e^{-inx}}{n} \tag{6.4.448}$$

Now these power series may be identifies as Maclaurin expansions of $-\ln(1-z)$ (see proof above) with $z=e^{\mathrm{i}x}$ for the first term and $z=e^{-\mathrm{i}x}$ for the second term.

Therefore:

$$\begin{aligned}f(x)=\sum_{n=1}^{+\infty}\frac{1}{n}\cos(nx)&=-\frac{1}{2}\left[\ln(1-e^{\mathrm{i}x})+\ln(1-e^{-\mathrm{i}x})\right]\\&=-\ln\left((1-e^{\mathrm{i}x})(1+e^{-\mathrm{i}x})\right)^{1/2}=-\ln\left(1-e^{-\mathrm{i}x}-e^{\mathrm{i}x}+e^{\mathrm{i}x}e^{-\mathrm{i}x}\right)^{1/2}\\&=-\ln\left(2-e^{-\mathrm{i}x}-e^{\mathrm{i}x}\right)^{1/2}=-\ln\left(2-2\cos(x)\right)^{1/2}=-\ln\left(2\sin\left(\frac{x}{2}\right)\right)\end{aligned} \tag{6.4.449}$$

The Fourier series allows implicitly to represent all the frequencies in a periodical signal whose function is known mathematically (closed form). We can wonder why talk about Fourier series when, in practice, we do not really know the mathematical representation of a signal? This will bring us to a better understanding of the concept of the Fourier transform in discrete-time that we will see a little further, which does not need a mathematical representation of a continuous and periodic signal.

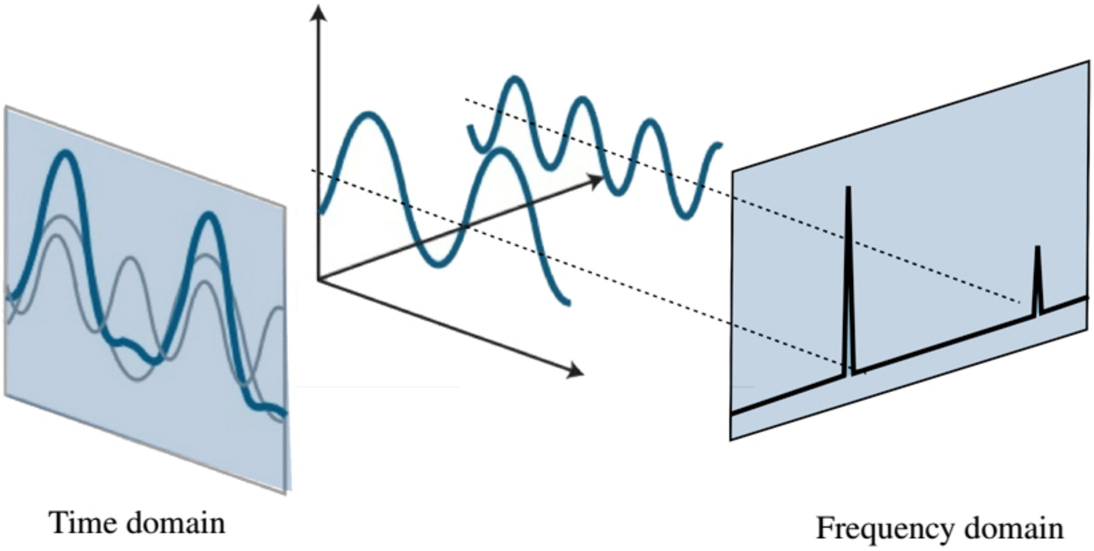


Figure 6.77: Time and frequency domain

We notice also that if $f(x)$, that is to say the periodic function of which we seek expression in trigonometric Fourier series, is even then the series will also be even and thus contain only cosine terms (the cosine function being an even function) implying that $b_n=0$ and otherwise in the case of an odd function $a_n=0$ (the sine being for reminder an odd function)!

It should be noted, and this is important for what will follow, that as we have seen in the section of Calculus during our study of trigonometric polynomials (see page 1397), Fourier series could be written in the following complex form (by changing some notations and passing the sum to infinity):

$$f(x)=\sum_{n=-\infty}^{+\infty} c_n e^{inx} \qquad (6.4.450)$$

and we have seen that (always in the section of Calculus) that:

$$a_0=2c_0 \qquad a_n=c_n+c_{-n} \qquad b_n=\mathrm{i}(c_n-c_{-n}) \qquad (6.4.451)$$

Therefore:

$$c_n=\frac{a_n-\mathrm{i}b_n}{2} \qquad (6.4.452)$$

This gives us:

$$\begin{aligned}c_n &= \frac{\dfrac{2}{T}\displaystyle\int_0^T f(x)\cos(nx)\mathrm{d}x - \dfrac{2}{T}\mathrm{i}\int_0^T f(x)\sin(nx)\mathrm{d}x}{2}\\ &= \frac{1}{T}\int_0^T f(x)\big(\cos(nx)-\mathrm{i}\sin(nx)\big)\mathrm{d}x\\ &= \frac{1}{T}\int_0^T f(x)e^{-\mathrm{i}nx}\mathrm{d}x\end{aligned} \tag{6.4.453}$$

Therefore:

$$c_n = \frac{1}{T}\int_0^T f(x)e^{-\mathrm{i}nx}\mathrm{d}x \tag{6.4.454}$$

Or more generally:

$$\boxed{c_n = \frac{1}{T}\int_{t_0}^{T+t_0} f(x)e^{-\mathrm{i}nx}\mathrm{d}x} \tag{6.4.455}$$

So if we take the famous case where $t_0 = -T/2$ we get:

$$c_n = \frac{1}{T}\int_{-T/2}^{T/2} f(x)e^{-\mathrm{i}nx}\mathrm{d}x \tag{6.4.456}$$

Obviously the two relation above are the Fourier coefficient in the "time space" (or "time domain") point of view. We will see in the section of Electrical Engineering that there is an equivalent if we work with Fourier series in the frequency space (or "frequency domain").

✍**Examples:**

E1. Upon decomposition of a continuous signal, we say (improperly at our point of view) that the coefficients a_n, b_n are each (implicitly) a separate frequency associated with an amplitude that we visualize on a graph by vertical lines. This graph shows the "frequency spectrum" of the decomposed signal. We can also add another representation which is named "phase spectrum". This spectrum gives us the phase of the harmonic signal (in phase advance or delay).

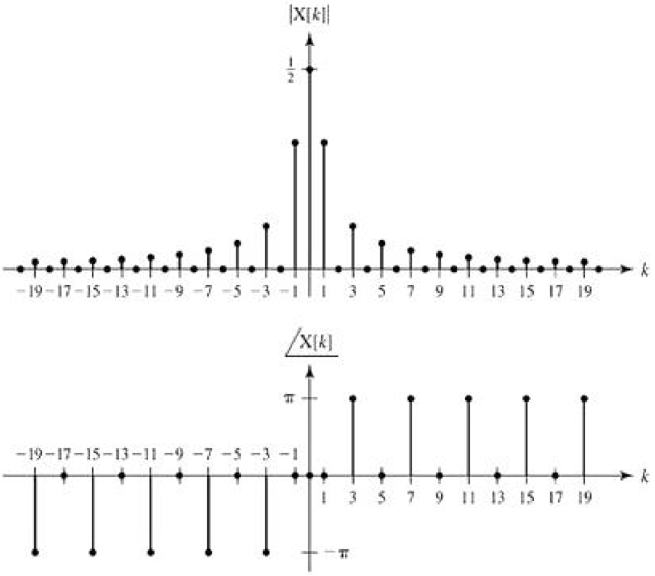
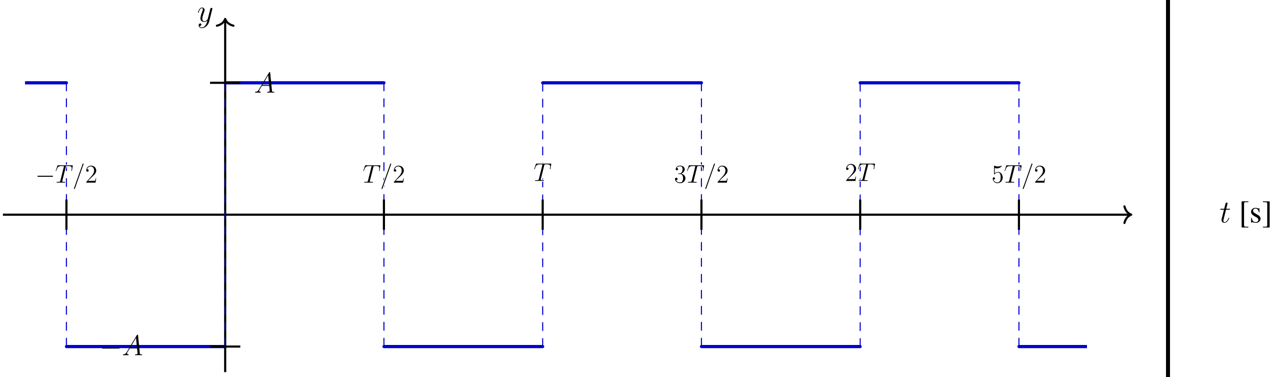


Figure 6.78: Example of amplitudes and frequencies associated to the different coefficients

Let us see now how to decompose a known periodic signal into several distinct amplitudes and frequencies signals.

Let us take for example, a periodic square wave (or "rectangular wave") signal defined over a period $T=2$ and of amplitude A such that:

$$f(t)=\begin{cases}-A & -1<t<0\\ A & 0<t<1\end{cases} \tag{6.4.457}$$



At period $T=2$ corresponds as we know a pulsation:

$$\omega=\frac{2\pi}{T}=\pi \tag{6.4.458}$$

Let us calculate first the coefficients c_k thanks to the integral that determine the coefficients (the choice of the bounds of the integral is therefore assumed that the signal is periodic by construction!):

$$\begin{aligned}c_k&=\frac{1}{T}\int\limits_{T} f(t)e^{-\mathrm{i}kwt}\mathrm{d}t=\frac{1}{2}\int\limits_{-1}^{1} f(t)e^{-\mathrm{i}k\pi t}\mathrm{d}t=\frac{1}{2}\left(\int\limits_{-1}^{0}-Ae^{-\mathrm{i}k\pi t}\mathrm{d}t+\int\limits_{0}^{1}Ae^{-\mathrm{i}k\pi t}\mathrm{d}t\right)\\&=\frac{-A}{2}\left(\int\limits_{-1}^{0}e^{-\mathrm{i}k\pi t}\mathrm{d}t-\int\limits_{0}^{1}e^{-\mathrm{i}k\pi t}\mathrm{d}t\right)=\frac{-A}{2}\left(\frac{-1}{ik\pi}\right)\left[e^{-\mathrm{i}k\pi t}\Big|_{-1}^{0}\right]+\frac{A}{2}\left(\frac{-1}{\mathrm{i}k\pi}\right)\left[e^{-\mathrm{i}k\pi t}\Big|_{0}^{1}\right]\\&=\frac{A}{2\mathrm{i}k\pi}\left[1-e^{-\mathrm{i}k\pi}\right]-\frac{A}{2\mathrm{i}k\pi}\left[e^{-\mathrm{i}k\pi}-1\right]=\frac{A}{2}\left[\frac{1-e^{-\mathrm{i}k\pi}}{\mathrm{i}k\pi}+\frac{e^{-\mathrm{i}k\pi}-1}{-\mathrm{i}k\pi}\right]\\&=\frac{A}{\mathrm{i}k\pi}\left[1-\frac{1}{2}\left(e^{\mathrm{i}k\pi}+e^{-\mathrm{i}k\pi}\right)\right]=\frac{A}{\mathrm{i}k\pi}\left[1-\cos(k\pi)\right]\end{aligned}\tag{6.4.459}$$

Taking $k=2$, we have:

$$c_2=\frac{A}{i2\pi}\left[1-\frac{1}{2}\left(e^{\mathrm{i}2\pi}+e^{-\mathrm{i}2\pi}\right)\right]=0\tag{6.4.460}$$

Similarly for $k=4,6,8$ and for any even number.

About odd numbers, we will have:

$$c_1=\frac{A}{\mathrm{i}\pi}\left[1-\frac{1}{2}\left(e^{\mathrm{i}\pi}+e^{-\mathrm{i}\pi}\right)\right]=\frac{2A}{\mathrm{i}\pi},\qquad c_3=\frac{A}{\mathrm{i}3\pi}\left[1-\frac{1}{2}\left(e^{\mathrm{i}3\pi}+e^{-\mathrm{i}3\pi}\right)\right]=\frac{2A}{\mathrm{i}3\pi}\tag{6.4.461}$$

The coefficients will then be:

$$c_k=\begin{cases}0 & k\text{ even}\\ \dfrac{2A}{\mathrm{i}k\pi} & k\text{ odd}\end{cases}\tag{6.4.462}$$

There is only problem in this relations, the coefficient c_0 cannot be calculated according to this relation because you can see that if $k=0$ in the result above, we have an infinite value and it is at least impossible. The coefficient is null or not null but never infinite (at least in physics because this implies and infinite energy).

To find the coefficient c_0, we must calculate the integral for $k=0$. The coefficient c_0 is then determined by:

$$\begin{aligned}c_0&=\frac{1}{T}\int\limits_{T} f(t)e^{-\mathrm{i}k\omega t}\mathrm{d}t=\frac{1}{2}\int\limits_{-1}^{1} f(t)\mathrm{d}t=\frac{1}{2}\int\limits_{-1}^{0} f(t)\mathrm{d}t+\frac{1}{2}\int\limits_{0}^{1} f(t)\mathrm{d}t\\&=\frac{1}{2}\int\limits_{-1}^{0}(-A)\mathrm{d}t+\frac{1}{2}\int\limits_{0}^{1}(A)\mathrm{d}t=0\end{aligned}\tag{6.4.463}$$

The "frequency" spectrum (caution to the abuse of language!) and amplitude will be of the following form for $k=-5\ldots+5$ and $A=1$ null frequencies not being shown:

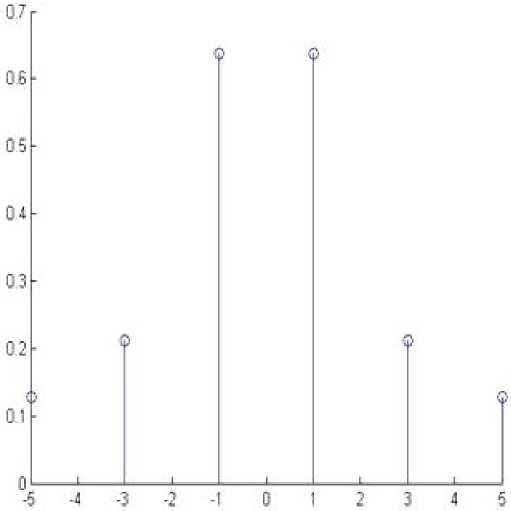
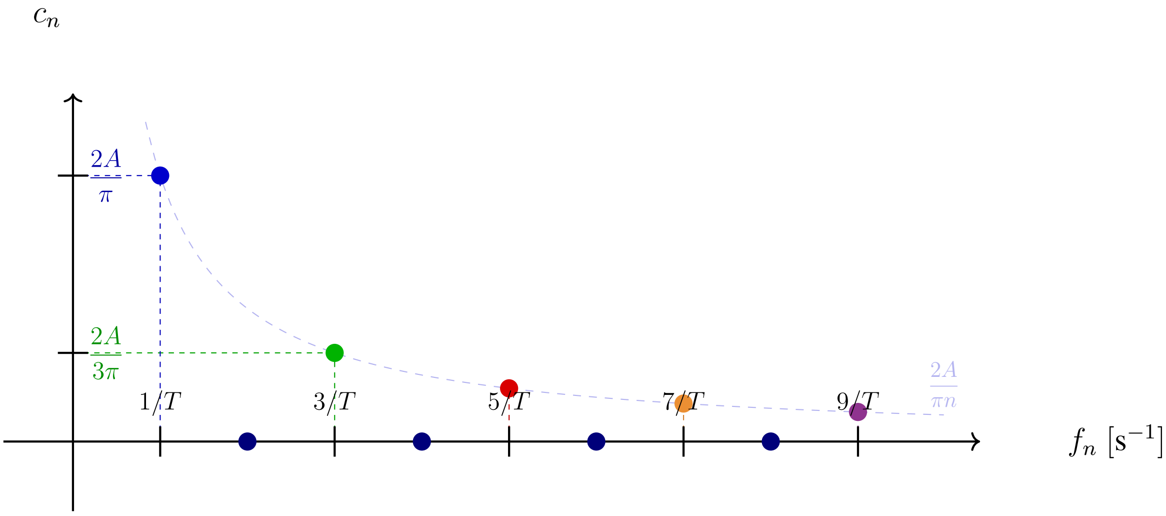
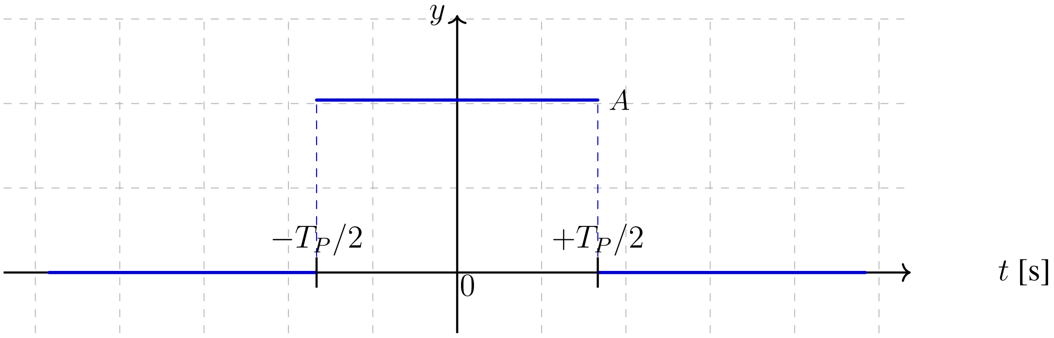


Figure 6.79: Frequency spectrum of the example Fourier series coefficients

Or as the frequency spectrum is symmetric it is also often represented as following:


c_n
$\frac{2A}{\pi}$
$\frac{2A}{3\pi}$
$1/T$
$3/T$
$5/T$
$7/T$
$9/T$
$\frac{2A}{\pi n}$
f_n [s^{-1}]

E2. Let us consider another famous case! The signal decomposition of the square of amplitude A and width T_p (this function is sometimes denoted $\Pi_P(t)$) and of period T:


y
A
$-T_P/2$
$+T_P/2$
0
t [s]

$$f(t)=\begin{cases}A & -T_P/2\geq t\geq +T_P/2\\ 0 & |t|>T_P/2\end{cases}\tag{6.4.464}$$

and:

$$\begin{aligned}c_k&=\frac{1}{T}\int_T f(t)e^{-\mathrm{i}k\omega t}\mathrm{d}t=\frac{1}{T}\int\limits_{-T_P/2}^{+T_P/2}Ae^{-\mathrm{i}k\omega t}\mathrm{d}t=\frac{A}{-\mathrm{i}k\omega T}\left[e^{-\mathrm{i}\omega t}\Big|_{-T_P/2}^{+T_P/2}\right]\\&=\frac{A}{-\mathrm{i}k2\pi fT}\left[e^{-\mathrm{i}k\omega\frac{T_P}{2}}-e^{\mathrm{i}k\omega\frac{T_P}{2}}\right]=\frac{A}{-\mathrm{i}k2\pi fT}\left[e^{-\mathrm{i}k2\pi f\frac{T_P}{2}}-e^{\mathrm{i}k2\pi f\frac{T_P}{2}}\right]\\&=\frac{A}{k\pi}\left[\frac{e^{\mathrm{i}k\pi fT_P}-e^{-\mathrm{i}k\pi fT_P}}{2\mathrm{i}}\right]=\frac{A}{k\pi}\sin(k\pi fT_P)\\&=AfT_p\frac{\sin(k\pi fT_P)}{k\pi fT_P}=\frac{A}{k\pi}\mathrm{sinc}(kfT_p)\end{aligned}\tag{6.4.465}$$

So we fall back on a discrete version of the famous cardinal sine function (see section Trigonometry page 2517)!

The abuse to talk about "frequencies" for Fourier coefficients thus leads us to have negative frequencies on the x-axis... but it's only a question of vocabulary (there is no direct relation with the real frequencies) with which you must be familiar.

The amplitude spectrum and phase is calculated according to the following relations:

$$|c_k|=\sqrt{(\Re(c_k))^2+(\Im(c_k))^2}\quad\text{and}\quad\angle\theta=\tan^{-1}\left(\frac{\Im(c_k)}{\Re(c_k)}\right)\tag{6.4.466}$$

It is then relatively easy to notice that if T tends to a larger and larger number, the spectrum peaks approach increasingly. So when T tends to infinity the spectrum becomes continuous!!!

The phase spectrum of the above example will give the following for the odd values:

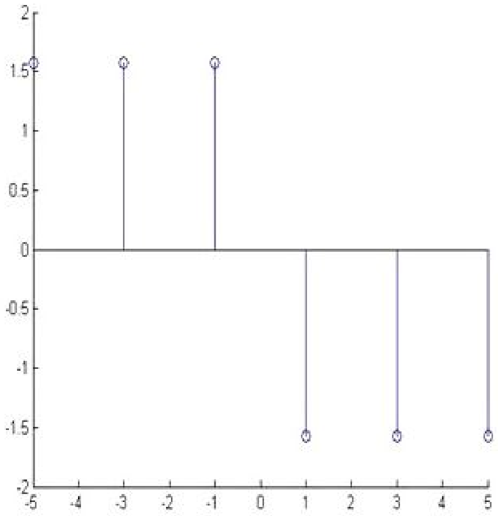
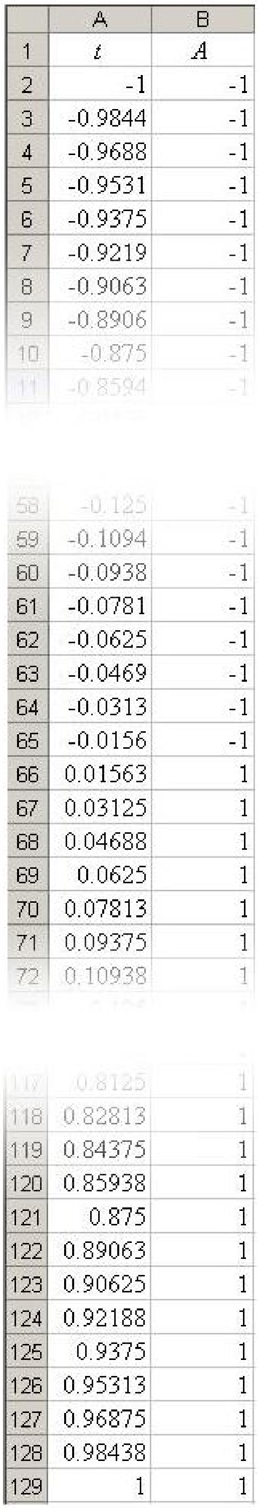


Figure 6.80: **Phase spectrum for the Fourier Series**

It is even possible for example to get relatively easily the frequency spectrum in a software like Microsoft® Excel® 11.8346 (the reader will find an example much more detailed and interesting on the companion exercise server in the section Sequences and Series) !!!

Indeed, it is enough for this purpose to sample for example our signal 128 times (Microsoft® Excel® 11.8346 needs 2^n samples and works only under this condition!). Then we divide the interval $-1<t<0$ in 64 samples and ditto for the interval $0>t>1$:



	A	B
1	*t*	*A*
2	-1	-1
3	-0.9844	-1
4	-0.9688	-1
5	-0.9531	-1
6	-0.9375	-1
7	-0.9219	-1
8	-0.9063	-1
9	-0.8906	-1
10	-0.875	-1
11	-0.8594	-1
58	-0.125	-1
59	-0.1094	-1
60	-0.0938	-1
61	-0.0781	-1
62	-0.0625	-1
63	-0.0469	-1
64	-0.0313	-1
65	-0.0156	-1
66	0.01563	1
67	0.03125	1
68	0.04688	1
69	0.0625	1
70	0.07813	1
71	0.09375	1
72	0.10938	1
117	0.8125	1
118	0.82813	1
119	0.84375	1
120	0.85938	1
121	0.875	1
122	0.89063	1
123	0.90625	1
124	0.92188	1
125	0.9375	1
126	0.95313	1
127	0.96875	1
128	0.98438	1
129	1	1

Figure 6.81: **Signal sample**

Which gives in graphical form (be careful because for the discrete Fourier transform works well in Microsoft® Excel® 11.8346, it is necessary that the sampling frequency - corresponding to the number of measurements in a second - is at least 100 times higher than the frequency of the original signal otherwise the result can be absurd!):

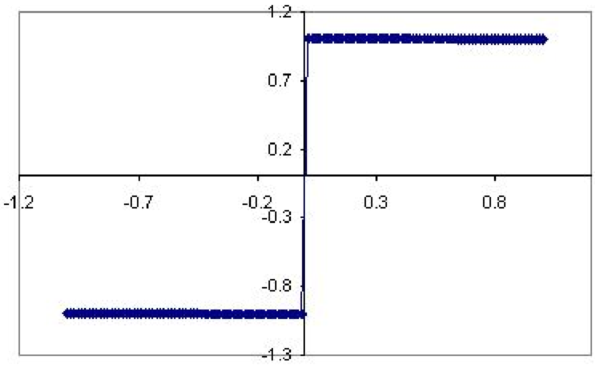


Figure 6.82: Graphical representation of the data series in Microsoft® Excel® 11.8346

Afterwards in Microsoft® Excel® you simply go to the menu **Tools/Utility Analysis** and choose the **Fourier Analysis** option:

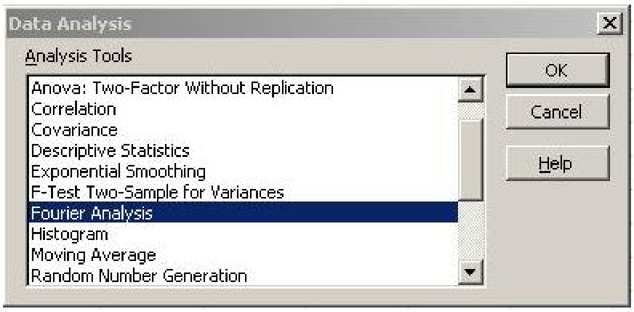
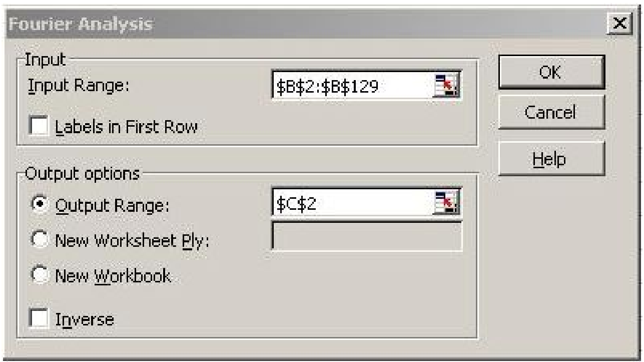

Data Analysis
Analysis Tools
Anova: Two-Factor Without Replication
Correlation
Covariance
Descriptive Statistics
Exponential Smoothing
F-Test Two-Sample for Variances
Fourier Analysis
Histogram
Moving Average
Random Number Generation
OK
Cancel
Help

Figure 6.83: Screenshot of Utility Analysis dialog box of Microsoft® Excel® 11.8346

Then comes the following dialog box that must be fill-in as indicated below (we see that the x-axis does not matter!):


Fourier Analysis
Input
Output options
OK
Cancel
Help

Input Range: `$B$2:$B$129`

☐ Labels in First Row

● Output Range: `$C$2`

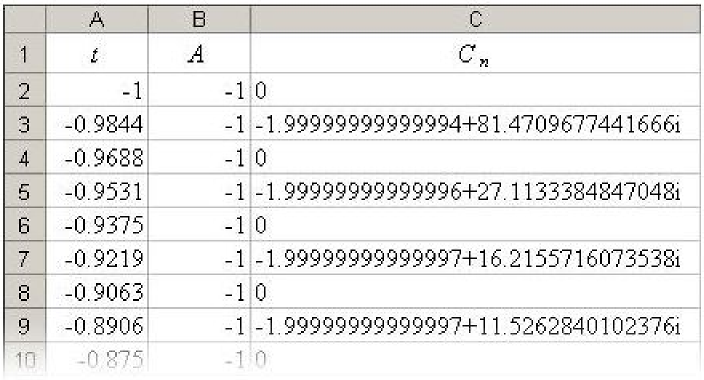
○ New Worksheet Ply:

○ New Workbook

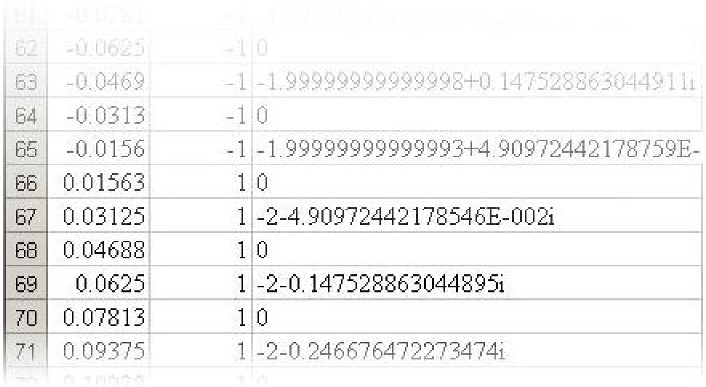
☐ Inverse

Figure 6.84: Parameters of the Fourier Analysis tool in Microsoft® Excel® 11.8346

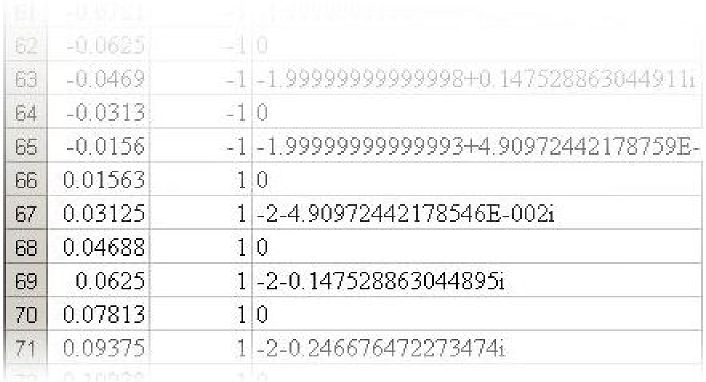
Then comes the following generated list for the coefficients:



	A	B	C
1	*t*	*A*	C_n
2	-1	-1	0
3	-0.9844	-1	-1.99999999999994+81.4709677441666i
4	-0.9688	-1	0
5	-0.9531	-1	-1.99999999999996+27.1133384847048i
6	-0.9375	-1	0
7	-0.9219	-1	-1.99999999999997+16.2155716073538i
8	-0.9063	-1	0
9	-0.8906	-1	-1.99999999999997+11.5262840102376i
10	-0.875	-1	0



61	-0.0781	-1	[illegible]
62	-0.0625	-1	0
63	-0.0469	-1	-1.99999999999998+0.147528863044911i
64	-0.0313	-1	0
65	-0.0156	-1	-1.99999999999993+4.90972442178759E-
66	0.01563	1	0
67	0.03125	1	-2-4.90972442178546E-002i
68	0.04688	1	0
69	0.0625	1	-2-0.147528863044895i
70	0.07813	1	0
71	0.09375	1	-2-0.246676472273474i

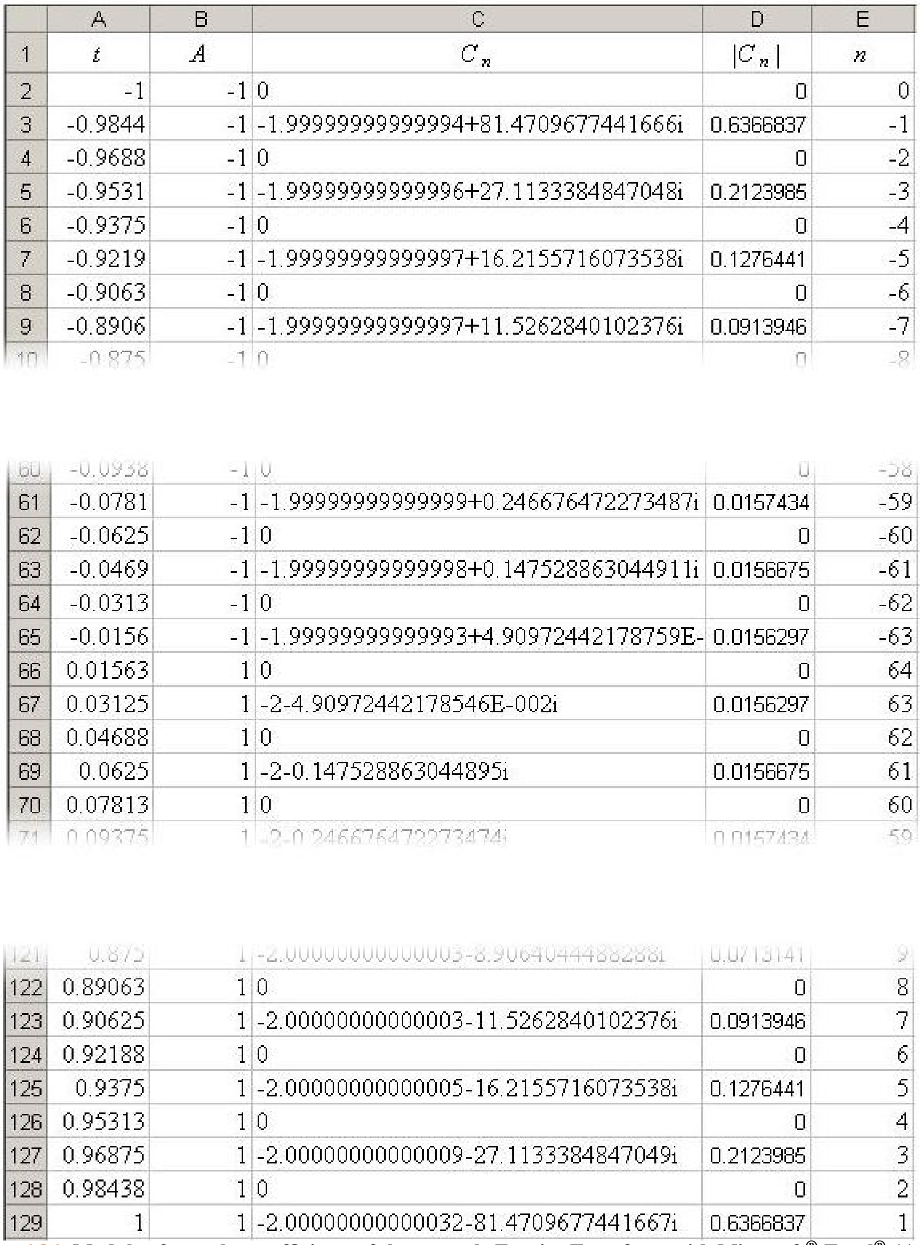


61	-0.0781	-1	[illegible]
62	-0.0625	-1	0
63	-0.0469	-1	-1.99999999999998+0.147528863044911i
64	-0.0313	-1	0
65	-0.0156	-1	-1.99999999999993+4.90972442178759E-
66	0.01563	1	0
67	0.03125	1	-2-4.90972442178546E-002i
68	0.04688	1	0
69	0.0625	1	-2-0.147528863044895i
70	0.07813	1	0
71	0.09375	1	-2-0.246676472273474i

Figure 6.85: **Corresponding Fourier coefficients to sampled signal with Microsoft® Excel® 11.8346**

It remains to calculate the module of the complex numbers with the native function `IMABS( )` function in Microsoft® Excel® 11.8346 and divide the result by 128 for each of the coefficients c_n but we already see that each pair even coefficient is zero and this match well the theoretical result obtained previously.

We then have putting the index n in front of each module:



	A	B	C	D	E
1	t	A	C_n	$\|C_n\|$	n
2	-1	-1	0	0	0
3	-0.9844	-1	-1.99999999999994+81.4709677441666i	0.6366837	-1
4	-0.9688	-1	0	0	-2
5	-0.9531	-1	-1.99999999999996+27.1133384847048i	0.2123985	-3
6	-0.9375	-1	0	0	-4
7	-0.9219	-1	-1.99999999999997+16.2155716073538i	0.1276441	-5
8	-0.9063	-1	0	0	-6
9	-0.8906	-1	-1.99999999999997+11.5262840102376i	0.0913946	-7
10	-0.875	-1	0	0	-8
60	-0.0938	-1	0	0	-58
61	-0.0781	-1	-1.99999999999999+0.246676472273487i	0.0157434	-59
62	-0.0625	-1	0	0	-60
63	-0.0469	-1	-1.99999999999998+0.147528863044911i	0.0156675	-61
64	-0.0313	-1	0	0	-62
65	-0.0156	-1	-1.99999999999993+4.90972442178759E-	0.0156297	-63
66	0.01563	1	0	0	64
67	0.03125	1	-2-4.90972442178546E-002i	0.0156297	63
68	0.04688	1	0	0	62
69	0.0625	1	-2-0.147528863044895i	0.0156675	61
70	0.07813	1	0	0	60
71	0.09375	1	-2-0.246676472273474i	0.0157434	59
121	0.875	1	-2.00000000000003-8.9064044488288i	0.0713141	9
122	0.89063	1	0	0	8
123	0.90625	1	-2.00000000000003-11.5262840102376i	0.0913946	7
124	0.92188	1	0	0	6
125	0.9375	1	-2.00000000000005-16.2155716073538i	0.1276441	5
126	0.95313	1	0	0	4
127	0.96875	1	-2.00000000000009-27.1133384847049i	0.2123985	3
128	0.98438	1	0	0	2
129	1	1	-2.00000000000032-81.4709677441667i	0.6366837	1

Figure 6.86: Module of complex coefficients of the example Fourier Transform with Microsoft® Excel® 11.8346

By plotting a customized scatter diagram (still with Microsoft® Excel® 11.8346) of columns `D` and `E`, we finally get (we restricted the x-axis to $[-5, +5]$ for easier reading):

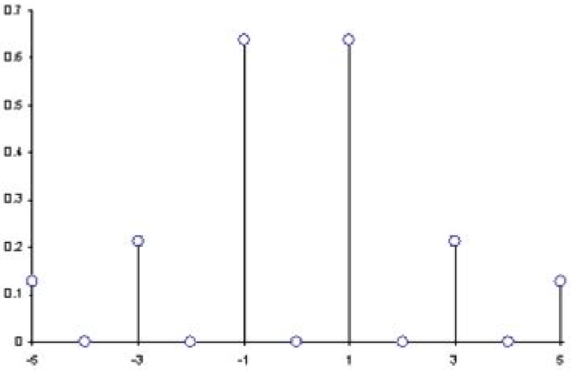


Figure 6.87: Frequency spectrum of the transformed with Microsoft® Excel® 11.8346

To compare with the theoretical calculations (chart already presented previously) ...:

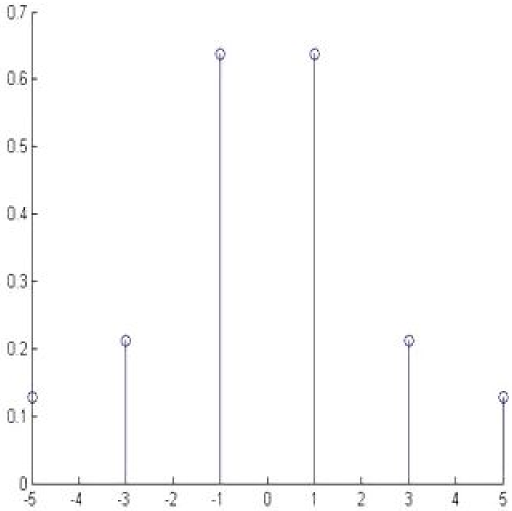


Figure 6.88: **Theoretical frequency spectrum of the example Fourier series coefficients**

Or in a more general and pedagogical way:

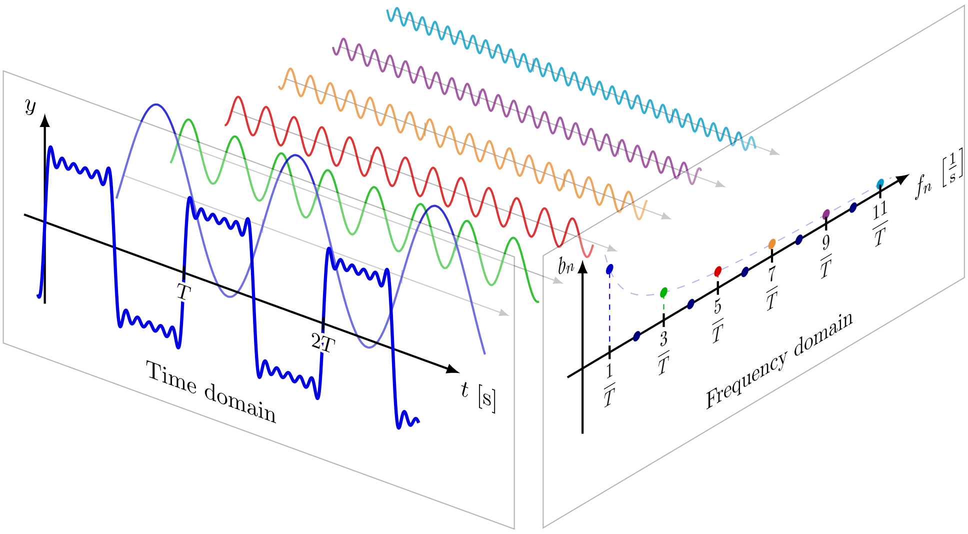

y
T
2T
t [s]
Time domain
b_n
f_n $[\frac{1}{s}]$
$\frac{1}{T}$
$\frac{3}{T}$
$\frac{5}{T}$
$\frac{7}{T}$
$\frac{9}{T}$
$\frac{11}{T}$
Frequency domain

Figure 6.89: **Pedagogical way to see the Fourier Transform (author: Izaak Neutelings)**

✍**Example:**

Let us now consider another example identical to the previous with a different approach. We define a periodic function of period $T=2\pi$ as follows:

$$f(x)=\begin{cases}-1 & -\pi<x<0\\ +1 & 0\leq x\leq\pi\end{cases} \tag{6.4.467}$$

Let us calculate the Fourier coefficients (we translate the bounds of the integral since the function is periodic, this does not change anything but makes the calculations easier!):

$$\begin{aligned}a_n&=\frac{2}{2\pi}\int\limits_0^{2\pi}f(x)\cos(nx)\mathrm{d}x=\frac{1}{\pi}\left(\int\limits_0^{\pi}1\cos(nx)\mathrm{d}x+\int\limits_{\pi}^{2\pi}-1\cos(nx)\mathrm{d}x\right)\\&=\frac{1}{\pi}\left(\int\limits_0^{\pi}\cos(nx)\mathrm{d}x-\int\limits_{\pi}^{2\pi}\cos(nx)\mathrm{d}x\right)=\frac{1}{\pi}\left(\left.\frac{\sin(nx)}{n}\right|_0^{\pi}-\left.\frac{\sin(nx)}{n}\right|_{\pi}^{2\pi}\right)=0\end{aligned}\tag{6.4.468}$$

and:

$$\begin{aligned}b_n&=\frac{2}{2\pi}\int\limits_0^{2\pi}f(x)\sin(nx)\mathrm{d}x=\frac{1}{\pi}\left(\int\limits_0^{\pi}1\sin(nx)\mathrm{d}x+\int\limits_{\pi}^{2\pi}-1\sin(nx)\mathrm{d}x\right)\\&=\frac{1}{\pi}\left(\int\limits_0^{\pi}\sin(nx)\mathrm{d}x-\int\limits_{\pi}^{2\pi}\sin(nx)\mathrm{d}x\right)=\frac{1}{\pi}\left(-\left.\frac{\cos(nx)}{n}\right|_0^{\pi}+\left.\frac{\cos(nx)}{n}\right|_{\pi}^{2\pi}\right)\\&=\frac{1}{\pi}\left(\left.\frac{\cos(nx)}{n}\right|_{\pi}^{2\pi}-\left.\frac{\cos(nx)}{n}\right|_0^{\pi}\right)=\frac{1}{n\pi}\left(\cos(nx)\Big|_{\pi}^{2\pi}-\cos(nx)\Big|_0^{\pi}\right)\\&=\frac{1}{n\pi}\big(\cos(2\pi n)-\cos(\pi n)-\cos(\pi n)+1\big)=\frac{1}{n\pi}\big(\cos(2\pi n)-2\cos(\pi n)+1\big)\end{aligned}\tag{6.4.469}$$

We notice that b_n is equal to 0 for n even and equal to $4\pi/n$ when n is odd. The Fourier series of the function under consideration is thus written:

$$\begin{aligned}f(x)&=\sum_{n=0}^{+\infty}\big(a_n\cos(nx)+b_n\sin(nx)\big)=\sum_{n=0}^{+\infty}b_n\sin(nx)\\&=4\pi\left[\frac{\sin(x)}{1}+\frac{\sin(3x)}{3}+\frac{\sin(5x)}{5}+\ldots+\frac{\sin(2x+1)}{2x+1}+\ldots\right]\end{aligned}\tag{6.4.470}$$

What in Maple™ 4.00b will be written:

```
>S:=(4/Pi)*Sum(sin((2*n+1)*x)/(2*n+1),n=0..N);
```

and that we can plot using the command:

```
>plot(subs(N=4,S),subs(N=8,S),subs(N=16,S),x=-Pi..Pi,
color=[red,green,blue],numpoints=200);
```

What gives three plots for 4, 8 and 16 of the series in red, green and blue:

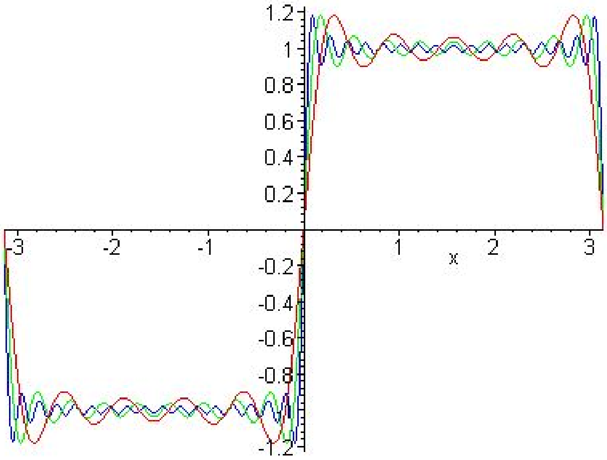


Figure 6.90: **Example of Fourier series in Maple™ 4.00b with 4, 8 and 16 terms**

For 50 terms we get:

```
> plot(subs(N=50,S),x=-Pi..Pi,numpoints=800);
```

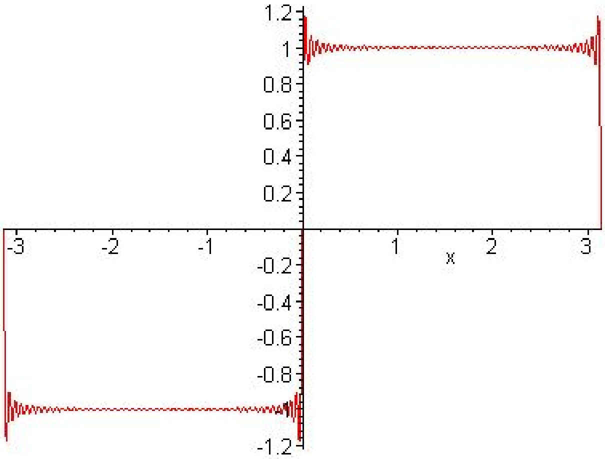


Figure 6.91: **Example of Fourier series in Maple™ 4.00b with 50 terms**

We see on the example above the side effects named "Gibbs phenomenon". It is possible to prove they occur to the value of the abscissa corresponding to $x=\pi/2n$ and matching equation and the peak rises to ± 1.179 for all values of n. Let's see this!

We proved just before for our example that:

$$\begin{aligned} f(x) &= \sum_{n=0}^{+\infty}\left(a_n\cos(nx)+b_n\sin(nx)\right)=\sum_{n=0}^{+\infty} b_n\sin(nx) \\ &= 4\pi\left[\frac{\sin(x)}{1}+\frac{\sin(3x)}{3}+\frac{\sin(5x)}{5}+\ldots\frac{\sin(2x+1)}{2x+1}+\ldots\right] \end{aligned} \tag{6.4.471}$$

Which can be written:

$$f(x)=\frac{4}{\pi}\int_0^x\left[\cos(t)+\cos(3t)+\ldots+\cos(2n+1)\right]\mathrm{d}t=\frac{4}{\pi}\int_0^x\Re\left[e^{\mathrm{i}t}+e^{3\mathrm{i}t}+\ldots+e^{\mathrm{i}(2n+1)t}\right]\mathrm{d}t \tag{6.4.472}$$

using the proof made much more earlier at the beginning of this section as that:

$$S_n = x + x^3 + x^5 + \ldots + x^{2n-1} = \frac{x^{2n+1} - x}{x^2 - 1} \tag{6.4.473}$$

Then we have:

$$\begin{aligned}
f(x) &= \frac{4}{\pi}\int\limits_0^x \Re\left[e^{\mathrm{i}t} + e^{3\mathrm{i}t} + \ldots + e^{\mathrm{i}(2n+1)t}\right]\mathrm{d}t \\
&= \frac{4}{\pi}\int\limits_0^x \Re\left[\frac{e^{(2n+1)\mathrm{i}t} - e^{\mathrm{i}t}}{e^{2\mathrm{i}t} - 1}\right]\mathrm{d}t = \frac{4}{\pi}\int\limits_0^x \Re\left[\frac{e^{\mathrm{i}t\left(e^{2n\mathrm{i}t}-1\right)}}{e^{\mathrm{i}t}\left(e^{\mathrm{i}t} - e^{-\mathrm{i}t}\right)}\right]\mathrm{d}t = \frac{4}{\pi}\int\limits_0^x \Re\left[\frac{e^{2n\mathrm{i}t} - 1}{e^{\mathrm{i}t} - e^{-\mathrm{i}t}}\right]\mathrm{d}t
\end{aligned} \tag{6.4.474}$$

Remember that during our study of complex numbers (see section Numbers page 243) we proved that:

$$\sin(x) = \frac{e^{\mathrm{i}x} - e^{-\mathrm{i}x}}{2\mathrm{i}} \tag{6.4.475}$$

Which brings us to:

$$\begin{aligned}
f(x) &= \frac{4}{\pi}\int\limits_0^x \Re\left[\frac{e^{2n\mathrm{i}t} - 1}{e^{\mathrm{i}t} - e^{-\mathrm{i}t}}\right]\mathrm{d}t = \frac{4}{\pi}\int\limits_0^x \Re\left[\frac{e^{2n\mathrm{i}t} - 1}{2\mathrm{i}\sin(t)}\right]\mathrm{d}t \\
&= \frac{4}{\pi}\int\limits_0^x \Re\left[\frac{\cos(2nt) + \mathrm{i}\sin(2nt) - 1}{2\mathrm{i}\sin(t)}\right]\mathrm{d}t = \frac{2}{\pi}\int\limits_0^x \Re\left[\frac{\cos(2nt)}{\mathrm{i}\sin(t)} + \frac{\mathrm{i}\sin(2nt)}{\mathrm{i}\sin(t)} - \frac{1}{\mathrm{i}\sin(t)}\right]\mathrm{d}t \\
&= \frac{2}{\pi}\int\limits_0^x \Re\left[-\frac{\mathrm{i}\cos(2nt)}{\sin(t)} + \frac{\sin(2nt)}{\sin(t)} + \frac{\mathrm{i}}{\sin(t)}\right]\mathrm{d}t = \frac{2}{\pi}\int\limits_0^x \Re\left[\frac{\sin(2nt)}{\sin(t)} - \mathrm{i}\frac{\cos(2nt) - 1}{\sin(t)}\right]\mathrm{d}t \\
&= \frac{2}{\pi}\int\limits_0^x \frac{\sin(2nt)}{\sin(t)}\mathrm{d}t
\end{aligned} \tag{6.4.476}$$

We will now focus on small values of x. So we can then make a Maclaurin development of first order at the denominator (but not the numerator because of the presence of the n):

$$f(x) \cong \frac{2}{\pi}\int\limits_0^x \frac{\sin(2nt)}{t}\mathrm{d}t \tag{6.4.477}$$

We make a change of variable:

$$f(x) \cong \frac{2}{\pi}\int\limits_0^{2nx} \frac{\sin(s)}{\dfrac{s}{2n}}\frac{\mathrm{d}s}{2n} = \int\limits_0^{2nx} \frac{\sin(s)}{s}\mathrm{d}s = \int\limits_0^{2nx} \operatorname{sinc}(s)\mathrm{d}s \tag{6.4.478}$$

where we used the traditional notation of the "sine cardinal" in the last relation as defined in the section Trigonometry at page 2517 (remember that this fraction is common in physics this is why it has a specific notation).

As what interest us is to determine the maximum of the Gibbs phenomenon (the disturbance), we see that it takes place in this particular case that we presented (see figure above) for each multiple of π and as the denominator of the expression of the integral will decrease as the multiple is higher, it follows that the greatest maximum is at the point where $2nx=\pi$ (the point 0 at the opposite will cancel the integral therefore we must this latter of our choice). Then we have:

$$f_{\text{Gibbs}}(x)\cong\int\limits_0^\pi\frac{\sin(s)}{s}\mathrm{d}s \tag{6.4.479}$$

the evaluation of this integral can be done only numerically as far as we know, therefore we get:

$$f_{\text{Gibbs}}(x)\cong 1.179 \tag{6.4.480}$$

That is about 18% above the expected threshold value.

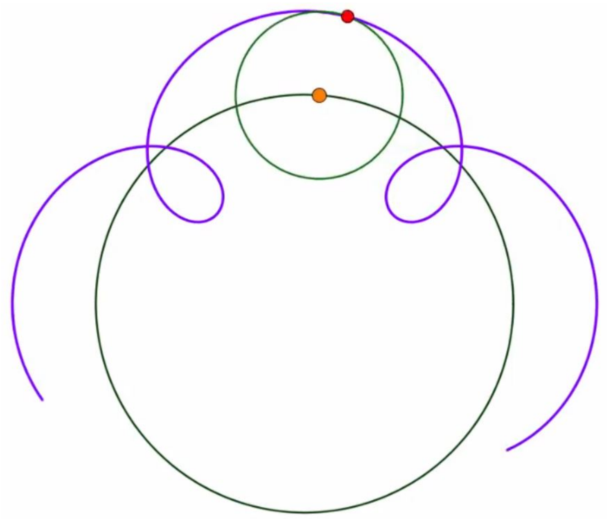
6.4.2.7.1 Fourier series generative art

Before continuing on hard mathematical topics about Fourier series, let us open a parenthesis where the purpose here take the science out of this quote:

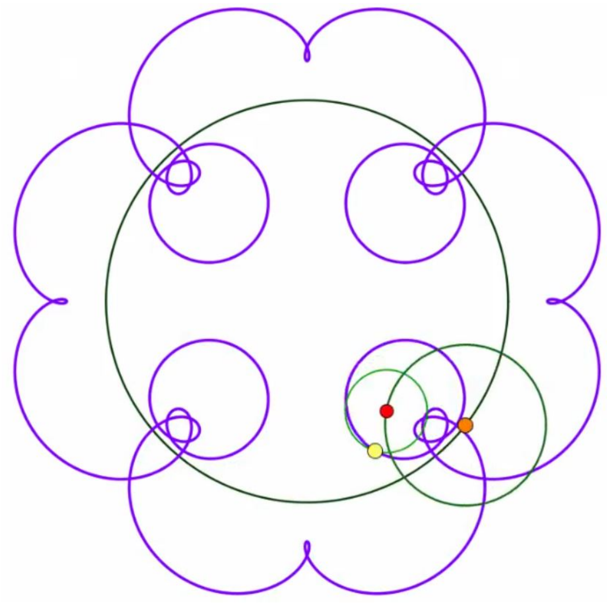
> ❞ *With four parameters I can fit an elephant.* ❞
>
> J. Von Neumann

quote 6.4:1749.#58

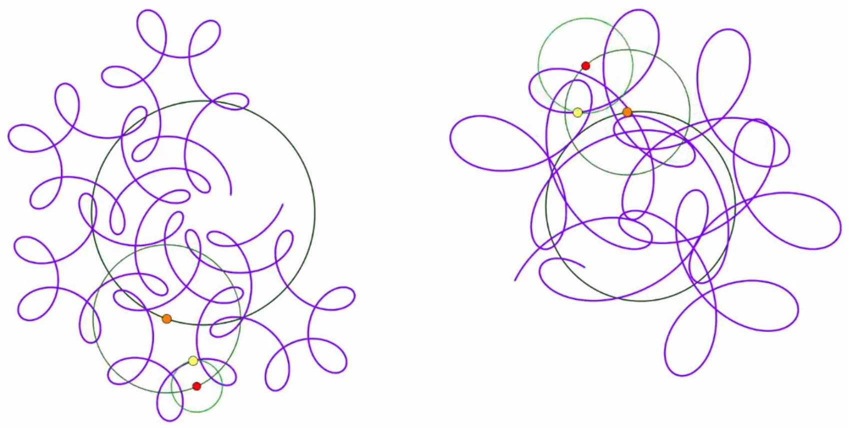
With computer it is easy to drawing stuffs using points moving on a circle, that are themselves centered on another circle like illustrated below (an epicycloid):



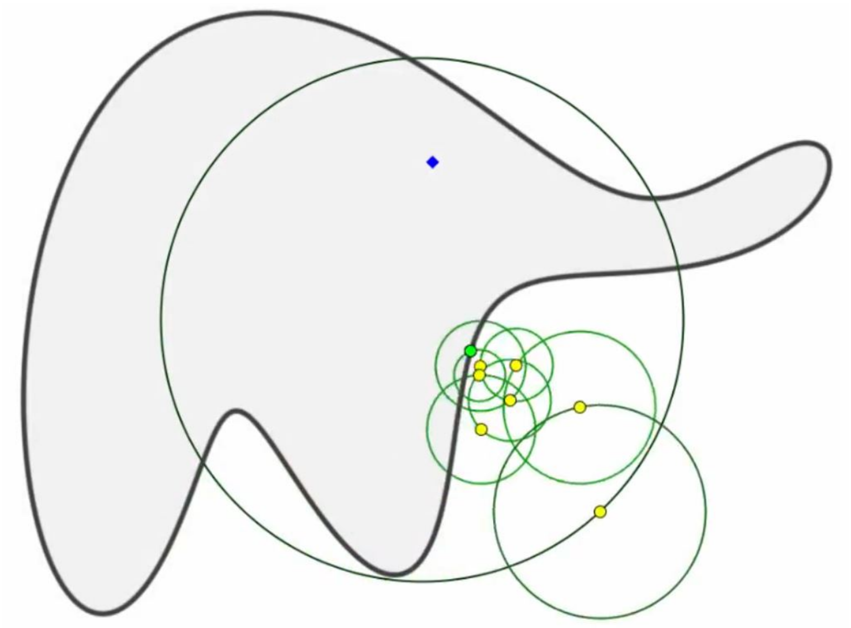
or a more complicated example:



Obviously as the examples above, the results change vary depending on the rotating speed of each point on its circle:

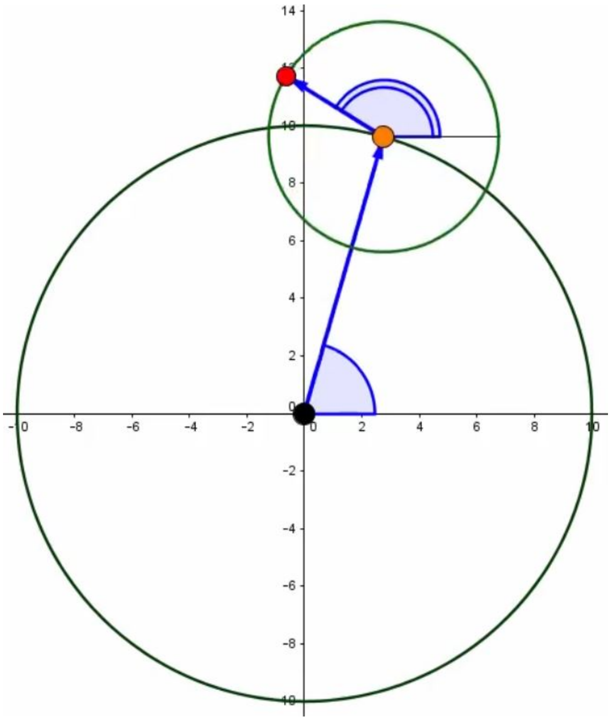


or a case with eight circles (each of them have a point moving at a specific rotation speed):



is based in fact on Fourier series!

To understand why consider the following example:



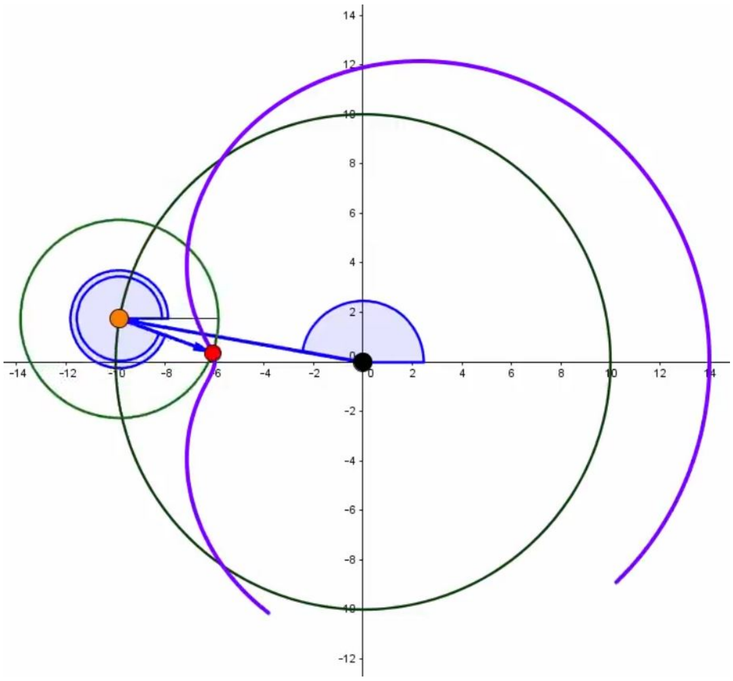
The first point is described obviously by the parametric equation:

$$\begin{pmatrix} 10\cos(t) \\ 10\sin(t) \end{pmatrix} \qquad (6.4.481)$$

and the second point therefore by (if it turns two time faster than the angle in the main circle):

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \begin{pmatrix} 10\cos(t)+4\cos(2t) \\ 10\sin(t)+4\sin(2t) \end{pmatrix} \qquad (6.4.482)$$

This is the parametric equation of the following epicycloid:



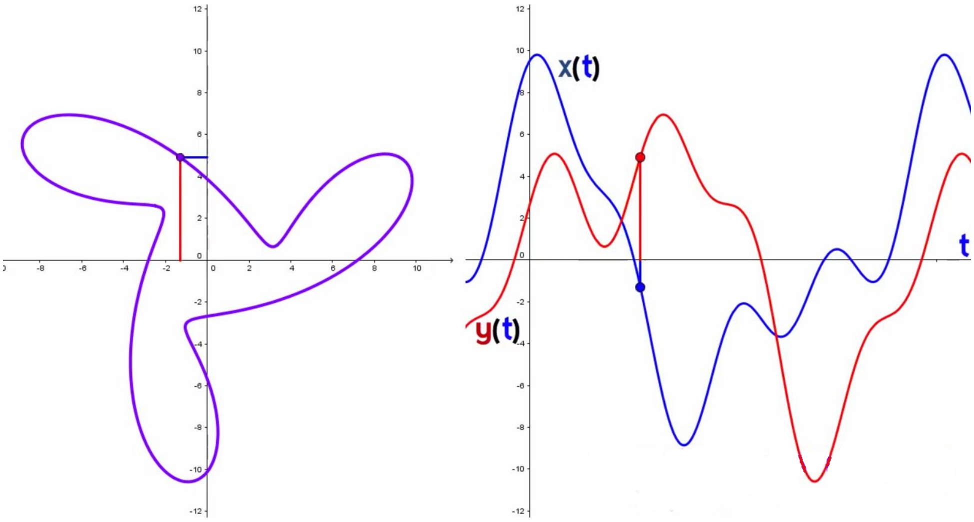
In the general case we have:

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \begin{pmatrix} a_0 + a_1\cos(n_1 t + \varphi_{x1}) + a_2\cos(n_2 t + \varphi_{x2}) + \ldots + a_N\cos(n_N t + \varphi_{x3}) \\ a_0 + a_1\sin(n_1 t + \varphi_{y1}) + a_2\sin(n_2 t + \varphi_{y2}) + \ldots + a_N\sin(n_N t + \varphi_{y3}) \end{pmatrix} \tag{6.4.483}$$

where a_i are the radius of the circles and n are the various circles speed (pulsation).

But the question is rather the opposite. From a given curve, how can find the values of the parameters?

The idea is quite simple. It's to follow respectively each coordinate in function of another one as illustrated below, where the closed shape is transformed into two curves:

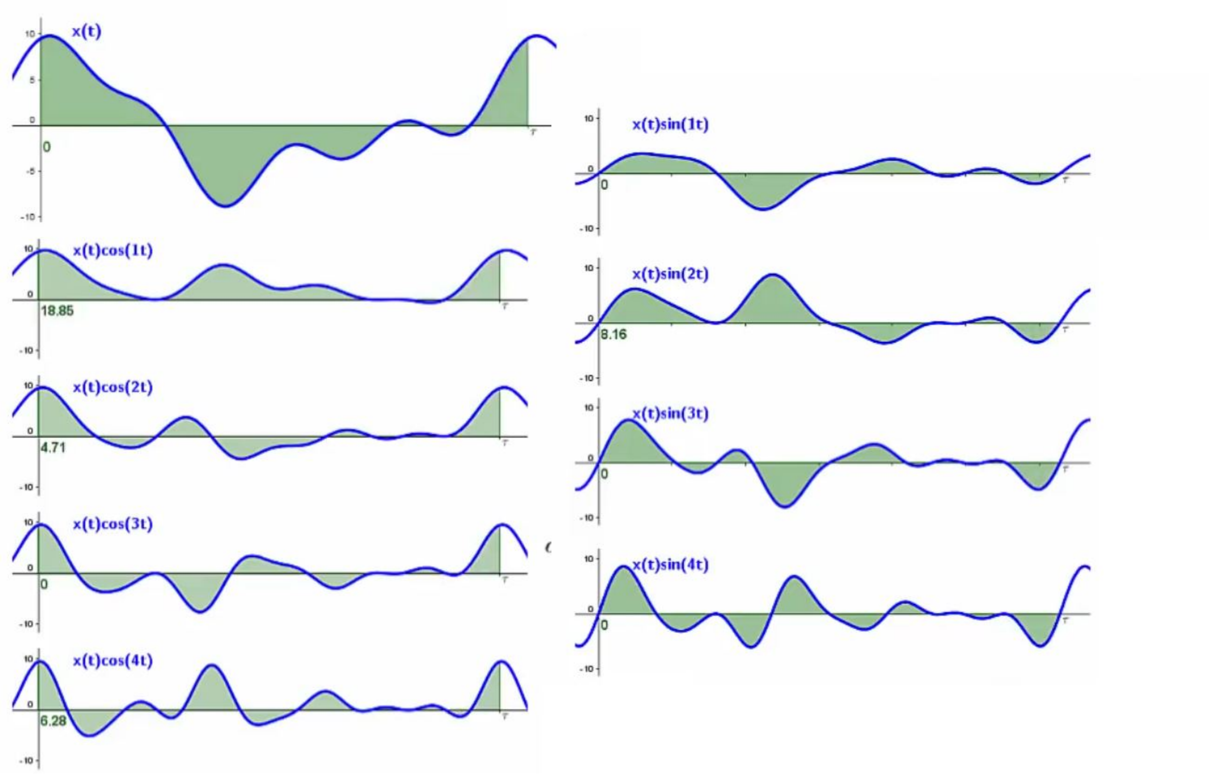


and as seen during our study of the Fourier series, we know how to find the Fourier series of such curves!

We know that the Fourier coefficients are given by:

$$\begin{aligned}
&\forall k\neq 0, a_k=\frac{2}{\tau}\int\limits_0^{\tau} x(t)\cos(kt)\mathrm{d}t\\
&\forall k\neq 0, b_k=\frac{2}{\tau}\int\limits_0^{\tau} x(t)\sin(kt)\mathrm{d}t\\
&\text{and } a_0=\frac{1}{\tau}\int\limits_0^{\tau} x(t)\mathrm{d}t
\end{aligned}\tag{6.4.484}$$

If we do that numerically for example with $x(t)$:



That gives:

$$\begin{aligned}
a_0 &= \frac{1}{\tau}\int_0^\tau x(t)\mathrm{d}t = 0\\
a_1 &= \frac{2}{\tau}\int_0^\tau x(t)\cos(1t)\mathrm{d}t = 6\\
b_1 &= \frac{2}{\tau}\int_0^\tau x(t)\sin(1t)\mathrm{d}t = 0\\
a_2 &= \frac{2}{\tau}\int_0^\tau x(t)\cos(2t)\mathrm{d}t = 1.5\\
b_2 &= \frac{2}{\tau}\int_0^\tau x(t)\sin(2t)\mathrm{d}t = 2.6\\
a_3 &= \frac{2}{\tau}\int_0^\tau x(t)\cos(3t)\mathrm{d}t = 0\\
b_3 &= \frac{2}{\tau}\int_0^\tau x(t)\sin(3t)\mathrm{d}t = 0\\
a_4 &= \frac{2}{\tau}\int_0^\tau x(t)\cos(4t)\mathrm{d}t = 2\\
b_4 &= \frac{2}{\tau}\int_0^\tau x(t)\sin(4t)\mathrm{d}t = 0
\end{aligned} \tag{6.4.485}$$

Therefore:

$$x(t) \cong 6\cos(t) + 1.5\cos(2t) + 2\cos(4t) + 2.6\sin(2t) \tag{6.4.486}$$

and using the same procedure, we could find:

$$\begin{aligned}
a_2 = \frac{2}{\tau}\int_0^\tau y(t)\cos(2t)\mathrm{d}t = 2.6 \quad & b_1 = \frac{2}{\tau}\int_0^\tau y(t)\sin(1t)\mathrm{d}t = 6\\
b_2 = \frac{2}{\tau}\int_0^\tau y(t)\sin(2t)\mathrm{d}t = -1.5 \quad & b_4 = \frac{2}{\tau}\int_0^\tau y(t)\sin(4t)\mathrm{d}t = 2
\end{aligned} \tag{6.4.487}$$

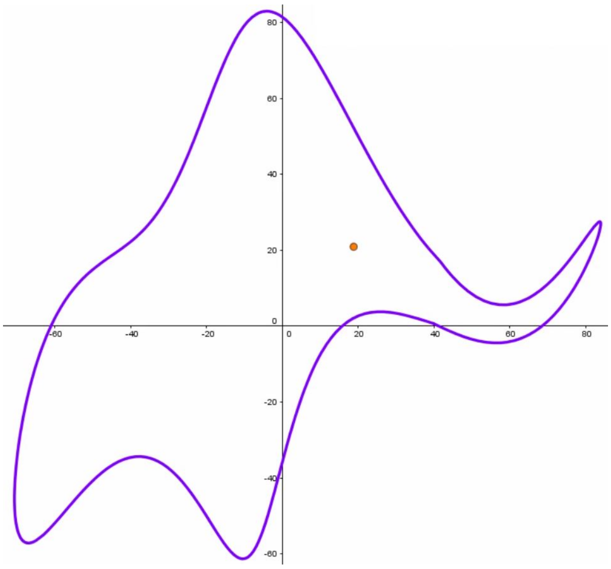
Therefore:

$$y(t) \cong 2.6\cos(2t) + 6\sin(t) - 1.5\sin(2t) + 2\sin(4t) \tag{6.4.488}$$

We see however that $x(t)$ doesn't have only cosine terms and that $y(t)$ doesn't have only sine terms. Then playing a bit with elementary trigonometric identities we can find (it's not absolutely a necessary transformation but like this it will be in adequation with the general expression given earlier above):

$$\begin{aligned}
x(t) &= 6\cos(t) + 3\cos\left(2t - 60^\circ\right) + 2\cos(4t)\\
y(t) &= 6\sin(t) + 3\sin\left(2t + 120^\circ\right) + 2\sin(4t)
\end{aligned} \tag{6.4.489}$$

Now let us come back on the four parameter elephant challenge. A satisfying result is:



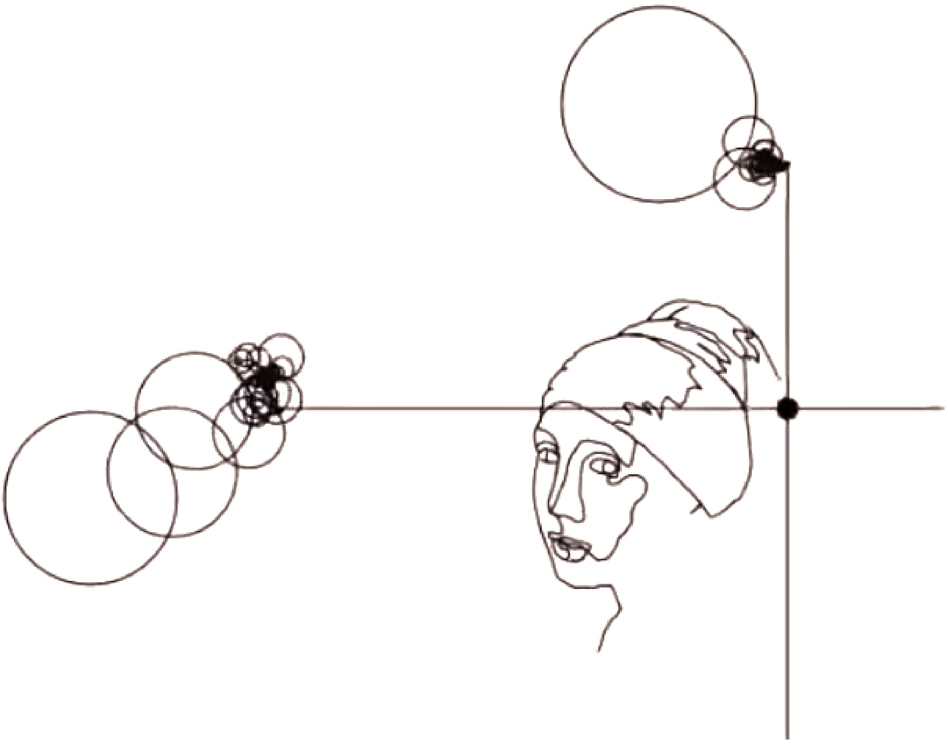
with:

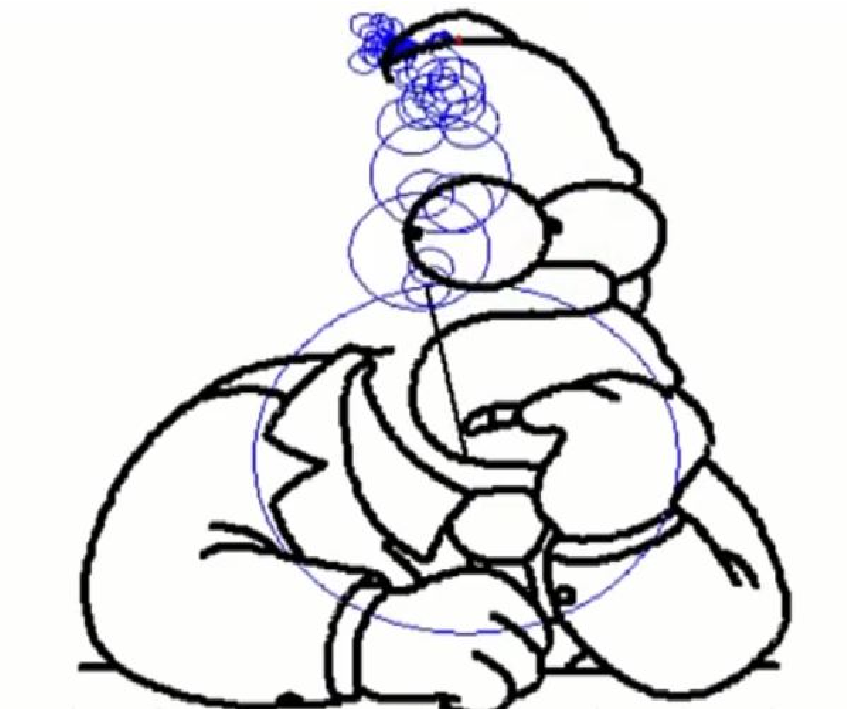
$$\begin{aligned}x(t)&=\Re(a)\cos(t)+\Re(b)\sin(t)+\Re(c)\sin(2t)+\Re(d)\sin(3t)\\y(t)&=\Im(a)\sin(t)+\Im(b)\sin(2t)+\Im(c)\cos(3t)+\Im(d)\cos(5t)\end{aligned}\tag{6.4.490}$$

with the four parameters:

$$\begin{aligned}a&=-60+50\mathrm{i}\\b&=30+18\mathrm{i}\\c&=-8-12\mathrm{i}\\d&=10+14\mathrm{i}\end{aligned}\tag{6.4.491}$$

Some people have fun doing other things:





Remark

If the reader want to see how we plot the Fermi's elephant, he can refer to our `R` companion book.

6.4.2.7.2 Power of a signal

A periodic signal has an infinite energy and null average power (see section Electrical Engineering page 8451). Its average power over a period is then defined by:

$$P_f=\frac{1}{T}\int\limits_{T}|f(t)|^2\mathrm{d}t \tag{6.4.492}$$

If we develop this equation, we have:

$$\begin{aligned}P_f&=\frac{1}{T}\int\limits_{T}f(t)\bar{f}(t)\mathrm{d}t=\frac{1}{T}\int\limits_{T}f(t)\left(\sum_{k=-\infty}^{+\infty}\bar{c}_k e^{-\mathrm{i}k\omega_0 t}\right)\mathrm{d}t\\&=\frac{1}{T}\sum_{k=-\infty}^{+\infty}\bar{c}_k\int\limits_{T}f(t)e^{-\mathrm{i}k\omega_0 t}\mathrm{d}t=\sum_{k=-\infty}^{+\infty}\bar{c}_k c_k=\sum_{k=-\infty}^{+\infty}|c_k|^2\end{aligned} \tag{6.4.493}$$

This means that the power of a continuous-time periodic signal is equal to the sum of the squared Fourier coefficients. This is what we name the "Parseval's theorem". This means that if we have any signal that can be decomposed in Fourier series, we can know the power of that signal using only the spectral coefficients.

In reality, we can't mathematically determine the expression of this signal, we use therefore discretization or sampling and then we use a discrete Fourier transform, we can calculate the power of this signal using only the spectral coefficients. This gives us a characteristic of the signal.

Let us also indicate the following result which will be very useful to us in the section of Thermodynamics for the study of the black-body and is also very closely related to important properties of the Riemann zeta function:

$$\begin{aligned}\sum_{k=-\infty}^{+\infty}|c_k|^2=\sum_{k=-\infty}^{+\infty}\left|\frac{a_k-\mathrm{i}b_k}{2}\right|=\frac{1}{4}\sum_{k=-\infty}^{+\infty}\left(\sqrt{a_k^2+b_k^2}\right)^2\\=\frac{1}{4}\sum_{k=-\infty}^{+\infty}a_k^2+b_k^2=\frac{1}{4}\sum_{k=-\infty}^{-1}\left(a_k^2+b_k^2\right)+\frac{1}{4}\left(a_0^2+b_0^2\right)+\frac{1}{4}\sum_{k=1}^{+\infty}\left(a_k^2+b_k^2\right)\\=\frac{1}{4}a_0^2+2\frac{1}{4}\sum_{k=1}^{+\infty}\left(a_k^2+b_k^2\right)=\frac{a_0^2}{4}+\frac{1}{2}\sum_{k=1}^{+\infty}\left(a_k^2+b_k^2\right)\end{aligned}\tag{6.4.494}$$

The following relation:

$$\boxed{\sum_{k=-\infty}^{+\infty}|c_k|^2=\frac{a_0^2}{4}+\frac{1}{2}\sum_{k=1}^{+\infty}\left(a_k^2+b_k^2\right)}\tag{6.4.495}$$

is named "Parseval's equality".

According to the definition of the Fourier series and the definition of the coefficient a_0 which follows immediately, we also have frequently in the literature:

$$\sum_{k=-\infty}^{+\infty}|c_k|^2=a_0+\frac{1}{2}\sum_{k=1}^{+\infty}\left(a_k^2+b_k^2\right)\tag{6.4.496}$$

6.4.2.7.3 Fourier Transform

Fourier series are a very powerful tool for the analysis of periodic signals for example, but the set of periodic functions is small compared to all the functions that we encounter in physical and engineering problems. So, will we introduce a new extremely powerful analytical tool that extends to a class of more general functions that have very important applications in signal processing, image processing, sound processing, statistics, finance (fourier option pricing techniques) and markets advanced statistics!!!

Remark

Many teachers and authors associate Fourier Series and Fourier Transform in the field of Functional Analysis. This is right in fact but it seemed to us more appropriate to put the study of the basics of this two subjects in this section because closely related to Sequences and Series. However, the subjects that follows normally the study of Fourier Transform, that is to say: Laplace Transforms, Hilbert Transform and others will be given in the section of Functional Analysis of this textbook (see page 2305). The Fast Fourier Transform study can be found in the section of Numerical Methods.

The Fourier transform (FT) is then used for both periodic signals and for aperiodic signals.

For this, we start from study of Fourier series with the complex notation of a periodic function of period T by considering that the period is becoming increasingly big to such that $T\rightarrow+\infty$. Therefore the spectral lines gradually approach to turn into a continuous spectrum.

Therefore, let us resume the expressions proved just earlier:

$$f(x)=\sum_{n=-\infty}^{+\infty}c_n e^{\mathrm{i}nx} \qquad c_n=\frac{1}{T}\int\limits_{-\frac{T}{2}}^{\frac{T}{2}}f(x)e^{-\mathrm{i}nx}\mathrm{d}x \tag{6.4.497}$$

that we can write equivalently in the following traditional form (wherein it is customary to put the factor $1/T$ rather in $f(t)$):

$$f(t)=\frac{1}{T}\sum_{n=-\infty}^{+\infty}c_k e^{\mathrm{i}\frac{2\pi n}{T}t} \qquad c_n=\int\limits_{-\frac{T}{2}}^{\frac{T}{2}}f(t)e^{-\mathrm{i}\frac{2\pi n}{T}t}\mathrm{d}t \tag{6.4.498}$$

and let us write this for future needs in the following form:

$$f(t)=\frac{1}{2\pi}\sum_{n=-\infty}^{+\infty}\frac{2\pi}{T}c_n e^{\mathrm{i}\frac{2\pi n}{T}t} \qquad c_n=\int\limits_{-\frac{T}{2}}^{\frac{T}{2}}f(t)e^{-\mathrm{i}\frac{2\pi n}{T}t}\mathrm{d}t \tag{6.4.499}$$

and let us put naturally that:

$$\omega_n=\frac{2\pi n}{T} \tag{6.4.500}$$

Thus, when $T\rightarrow+\infty$, the pulsation tends to zero and we have $\omega_n\rightarrow\omega$ because we move from discrete values in continuous values that browse through the set of real number $\mathbb{R}$ (for all k).

Therefore:

$$f(t)=\frac{1}{2\pi}\sum_{n=-\infty}^{+\infty}\frac{2\pi}{T}c_k e^{\mathrm{i}\omega_n t} \qquad c_n=\int\limits_{-\frac{T}{2}}^{\frac{T}{2}} f(t)e^{-\mathrm{i}\omega_n t}\mathrm{d}t \tag{6.4.501}$$

we pass to the limit that is to say:

$$T\rightarrow+\infty \qquad \omega_n\rightarrow\omega \tag{6.4.502}$$

This implies that:

$$\frac{2\pi}{T}=\frac{2\pi(n+1)}{T}-\frac{2\pi n}{T}=\frac{2\pi\Delta n}{T}\underset{T\rightarrow+\infty}{=}\mathrm{d}\omega \tag{6.4.503}$$

Therefore we get for the coefficients (we change the notation because the previous one is inadequate):

$$F(\omega)=\int\limits_{-\infty}^{+\infty} f(t)e^{-\mathrm{i}\omega t}\mathrm{d}t \tag{6.4.504}$$

and for the infinite series (the sum becomes an integral)

$$f(t)=\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} F(\omega)e^{\mathrm{i}\omega t}\mathrm{d}\omega \tag{6.4.505}$$

Caution!!! To make the difference between the given function and its equivalent in which we seek expression in infinite sum, we will denote them differently now. Thus, we get:

$$f(t)=\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} F(\omega)e^{\mathrm{i}\omega t}\mathrm{d}\omega \qquad F(\omega)=\int\limits_{-\infty}^{+\infty} f(t)e^{-\mathrm{i}\omega t}\mathrm{d}t \tag{6.4.506}$$

Thus the discrete Fourier series becomes a continuous function.

Definitions (#6.4.224):

D1. We name "Fourier Transform (FT)" of f, or more rigorously the "Continuous Time Fourier Transform (CTFT)", the relation:

$$\boxed{\mathcal{F}\big(f(t)\big)(\omega)=F(\omega)=\int\limits_{-\infty}^{+\infty} f(t)e^{-\mathrm{i}\omega t}\mathrm{d}t} \tag{6.4.507}$$

sometimes also denoted as follows:

$$\hat{f}(\omega)=\int f(t)e^{-\mathrm{i}\omega t}\mathrm{d}t \tag{6.4.508}$$

sometimes also named "spectral density amplitude".

D2. We name "Inverse Fourier Transform (IFT)", or more rigorously the Continuous Time Inverse Fourier Transform (CTIFT)", of F the relation:

$$\boxed{\mathcal{F}^{-1}\big(F(\omega)\big)(t)=f(t)=\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty}F(\omega)e^{\mathrm{i}\omega t}\mathrm{d}\omega} \tag{6.4.509}$$

Any such transformation technique (as there are many as we will see in the section of Functional Analysis!) is named an "integral transformation".

Remark

There are many ways of writing the Fourier transform according to the choice of the initial value of T! The reader must also know that it is quite common in physics to invert the definitions...the Fourier transform may be designated as the Inverse Fourier transform and reciprocally...

Some physicist and engineers prefer to make the two previous relations symmetrical by putting the same coefficient in both directions, which will be for example $1/\sqrt{2}$. This will give:

$$\begin{aligned}\mathcal{F}\big(f(t)\big)(\omega)=F(\omega)&=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}f(t)e^{-\mathrm{i}\omega t}\mathrm{d}t\\ \mathcal{F}^{-1}\big(F(\omega)\big)(t)=f(t)&=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}F(\omega)e^{\mathrm{i}\omega t}\mathrm{d}\omega\end{aligned} \tag{6.4.510}$$

Let us also give the corresponding three-dimensional version that will serve us many times in wave mechanics, electrodynamics, wave optics or in the various sections of quantum physics of this textbook:

$$\mathcal{F}(\vec{k})=\frac{1}{(2\pi)^{\frac{3}{2}}}\iiint f(\vec{r})e^{-\mathrm{i}\vec{k}\circ\vec{r}}\mathrm{d}^3r \qquad f(\vec{r})=\frac{1}{(2\pi)^{\frac{3}{2}}}\iiint \mathcal{F}(\vec{k})e^{\mathrm{i}\vec{k}\circ\vec{r}}\mathrm{d}^3k \tag{6.4.511}$$

To make things perhaps clearer (at least we hope so), let us prove generally that the previous Fourier transform $\mathcal{F}$ is isometric (retains the "norm" - or "modulus" if you prefer...)

Theorem 6.25. For any functions f,g we have the functional inner product:

$$f\circ\mathcal{F}(g)=\mathcal{F}^{-1}(f)\circ g \tag{6.4.512}$$

But since the functions are in the complex space, as we saw in the section of Vector Calculus, then we must use the notation of the hermitian product:

$$\langle f|\mathcal{F}(g)\rangle=\langle\mathcal{F}^{-1}(f)|g\rangle \tag{6.4.513}$$

Remember that:

$$\langle f|g\rangle=\int\limits_{\mathbb{R}^n}f(\vec{r})\overline{g(\vec{r})}\mathrm{d}^nr \tag{6.4.514}$$

Proof 6.25.1. Then we want to prove the equality:

$$\langle f|g\rangle=\int\limits_{\mathbb{R}^n} f(\vec{r})\overline{g(\vec{r})}\mathrm{d}^n r \qquad (6.4.515)$$

Explicitly:

$$\langle f|\mathcal{F}(g)\rangle=\int\limits_{\mathbb{R}^n} f(\vec{r})\overline{\mathcal{F}(g)}\mathrm{d}^n r \qquad (6.4.516)$$

But the variable to integrate but must be the same and for $\mathcal{F}(g)$ to be implicitly dependent of $\vec{r}$ it is necessary to take the Fourier transform on $\vec{k}$. Such as:

$$\mathcal{F}(\vec{r})=\frac{1}{(2\pi)^{\frac{3}{2}}}\iiint f(\vec{k})e^{-\mathrm{i}\vec{k}\circ\vec{r}}\mathrm{d}^3 k \qquad (6.4.517)$$

Therefore:

$$\begin{aligned}\langle f(\vec{r})|\mathcal{F}(g)(\vec{r})\rangle&=\int\limits_{\mathbb{R}^n} f(\vec{r})\overline{\mathcal{F}(g)}\mathrm{d}^n r\int\limits_{\mathbb{R}^n} f(\vec{r})\overline{\left(\frac{1}{(2\pi)^{\frac{n}{2}}}\int\limits_{\mathbb{R}^n} g(\vec{k})e^{-\mathrm{i}\vec{k}\circ\vec{r}}d^n k\right)}\mathrm{d}^n r\\&=\int\limits_{\mathbb{R}^n} f(\vec{r})\left(\frac{1}{(2\pi)^{\frac{n}{2}}}\int\limits_{\mathbb{R}^n} \overline{g(\vec{k})}e^{-\mathrm{i}\vec{k}\circ\vec{r}}\mathrm{d}^n k\right)\mathrm{d}^n r\end{aligned} \qquad (6.4.518)$$

Therefore using the Fubini's theorem (see section Differential and Integral page 1521):

$$\begin{aligned}\langle f|\mathcal{F}(g)\rangle&=\int\limits_{\mathbb{R}^n} \overline{g(\vec{k})}\left(\frac{1}{(2\pi)^{\frac{n}{2}}}\int\limits_{\mathbb{R}^n} f(\vec{r})e^{\mathrm{i}\vec{k}\circ\vec{r}}\mathrm{d}^n r\right)\mathrm{d}^n k\\&=\int\limits_{\mathbb{R}^n} \overline{g(\vec{k})}\mathcal{F}^{-1}(f)(\vec{k})\mathrm{d}^n k=\langle\mathcal{F}^{-1}(f),g\rangle\end{aligned} \qquad (6.4.519)$$

Thanks to this result, we have also proved (this is immediate):

$$\langle\mathcal{F}(f),\mathcal{F}(g)\rangle=\langle\mathcal{F}^{-1}(\mathcal{F}(f)),g\rangle=\langle f,g\rangle \qquad (6.4.520)$$

We have not specified the bounds: they are infinite in every definition (we include all possible $\vec{k}$ or $\vec{r}$).

■ Q.E.D.

Let us now see and prove three interesting properties of the Fourier transform:

P1. If the function f is an even function (see section Functional Analysis page 2212), it comes a simplification of the Fourier Transform such that:

$$\begin{aligned}\mathcal{F}\big(f(x)\big)(\omega)&=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} f(t)e^{-\mathrm{i}\omega t}\mathrm{d}t=\frac{1}{\sqrt{2\pi}}\left(\int\limits_{-\infty}^{0} f(t)e^{-\mathrm{i}\omega t}\mathrm{d}t+\int\limits_{0}^{+\infty} f(t)e^{-\mathrm{i}\omega t}\mathrm{d}t\right)\\&=\frac{1}{\sqrt{2\pi}}\left(-\int\limits_{-\infty}^{0} f(t)e^{\mathrm{i}\omega t}\mathrm{d}t+\int\limits_{0}^{+\infty} f(t)e^{-\mathrm{i}\omega t}\mathrm{d}t\right)\\&=\frac{1}{\sqrt{2\pi}}\left(\int\limits_{0}^{+\infty} f(t)e^{\mathrm{i}\omega t}\mathrm{d}t+\int\limits_{0}^{+\infty} f(t)e^{-\mathrm{i}\omega t}\mathrm{d}t\right)\\&=\frac{1}{\sqrt{2\pi}}\int\limits_{0}^{+\infty} f(t)\left(e^{\mathrm{i}\omega t}+e^{-\mathrm{i}\omega t}\right)\mathrm{d}t=\frac{2}{\sqrt{2\pi}}\int\limits_{0}^{+\infty} f(t)\cos(\omega t)\mathrm{d}t\end{aligned} \tag{6.4.521}$$

P2. If f is odd, we proceed in the same manner as above, and we get:

$$\mathcal{F}\big(f(x)\big)(\omega)=-\frac{2\mathrm{i}}{\sqrt{2\pi}}\int\limits_{0}^{+\infty} f(t)\sin(\omega t)\mathrm{d}t \tag{6.4.522}$$

P3. Very important property of the Fourier transforms which will be useful to us in finance (see section Economy page 7485), and also as part of the study of the heat equation (see section of Thermodynamics page 3694).

First remember that the Fourier transform is given by:

$$\mathcal{F}\big(f(t)\big)(\omega)=F(\omega)=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} f(t)e^{-\mathrm{i}\omega t}\mathrm{d}t \tag{6.4.523}$$

We want to see what happens if:

$$\mathcal{F}\big(\partial_t f(t)\big)(\omega)=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} \partial_t f(t)e^{-\mathrm{i}\omega t}\mathrm{d}t \tag{6.4.524}$$

By doing an integration by parts (see section Integral and Differential page 1532):

$$\int\limits_{a}^{b} f(t)g'(t)\mathrm{d}t=f(t)g(t)|_a^b+\int\limits_{a}^{b} f'(t)g(t)\mathrm{d}t \tag{6.4.525}$$

we get:

$$\begin{aligned}\mathcal{F}\big(\partial_t f(t)\big)(\omega)&=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} e^{-\mathrm{i}\omega t}\partial_t f(t)\mathrm{d}t=\frac{1}{\sqrt{2\pi}}e^{-\mathrm{i}\omega t}f(t)|_{-\infty}^{+\infty}-\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} -\mathrm{i}\omega^{-\mathrm{i}\omega t}f(t)\mathrm{d}t\\&=\mathrm{i}\omega\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} e^{-\mathrm{i}\omega t}f(t)\mathrm{d}t=\mathrm{i}\omega\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} f(t)e^{-\mathrm{i}\omega t}\mathrm{d}t=\mathrm{i}\omega\mathcal{F}(\omega)\end{aligned} \tag{6.4.526}$$

where we put ourselves in the situation where:

$$\lim_{t\to\pm\infty} f(t)=0 \tag{6.4.527}$$

Therefore:

$$\boxed{\mathcal{F}\big(\partial_t f(t)\big)=\mathrm{i}\omega\mathcal{F}(\omega)} \tag{6.4.528}$$

More generally:

$$\boxed{\mathcal{F}\big(\partial_t^n f(t)\big)=(\mathrm{i}\omega)^n\mathcal{F}(\omega)} \tag{6.4.529}$$

This property can be useful for the following scheme:

$$\begin{matrix}\text{origin}\\ \text{space}\end{matrix}\left\{\begin{tikzcd} \begin{matrix}\text{ODE}\\ \text{in } t\end{matrix} \arrow[r, "\mathcal{F}"] \arrow[d, dashed] & \begin{matrix}\text{eqn.}\\ \text{in } \omega\end{matrix} \arrow[d] \\ f(t) & F(\omega) \arrow[l, "\mathcal{F}^{-1}"'] \end{tikzcd}\right\}\begin{matrix}\text{transform}\\ \text{space}\end{matrix}$$

Remark

The branch of "harmonic analysis", or "2D Fourier analysis", is the branch of mathematics that studies the representation of functions or signals as a superposition of basic waves. It deepens and generalizes the notions of Fourier series and Fourier transform. The basic waves are named "harmonics", hence the name of discipline. During the last two centuries it has had numerous applications in physics and economics under the name "spectral analysis" and knows recent applications including signal processing, quantum mechanics, neuroscience, stratigraphy, statistics, etc.

✍**Examples:**

E1. Let us see now a first example (among the three fundamentals one) of a Fourier transform that we will use again in the sections about Quantum Physics as well as in Wave Optics. We want to calculate the Fourier transform (spectrum) of the following function (rectangular pulse):

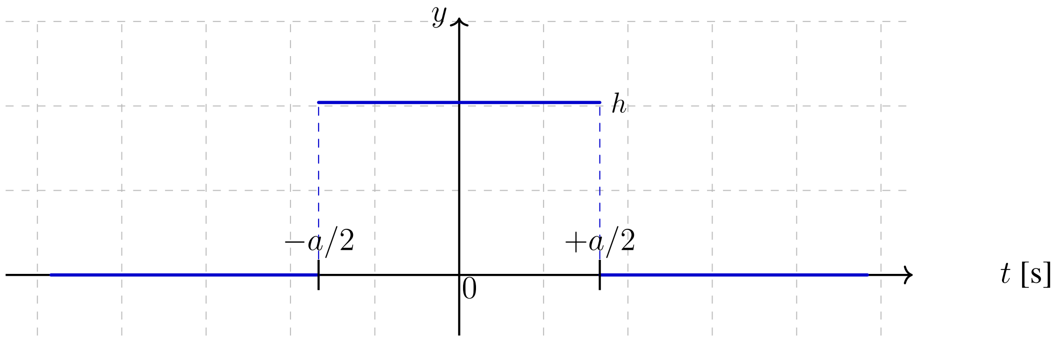
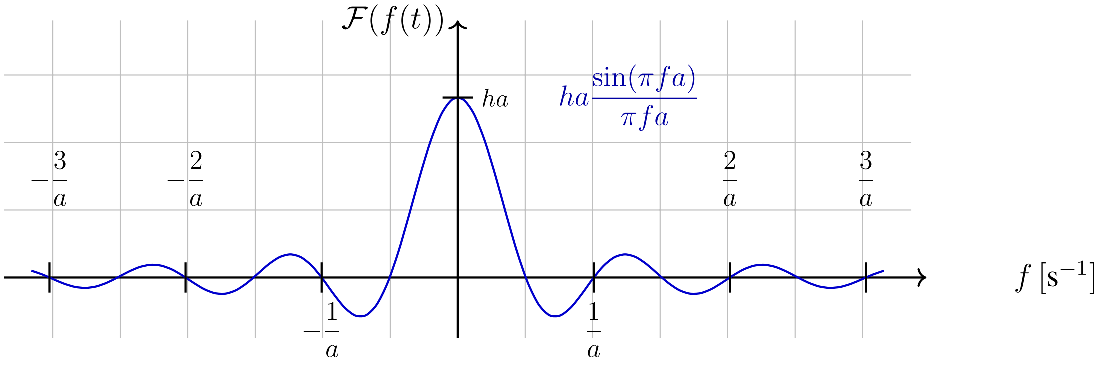


Figure 6.92: **Rectangular pulse example for Fourier Transform (author: Izaak Neutelings)**

We have therefore:

$$\begin{aligned}\mathcal{F}(f(t)) &= \int_{-\infty}^{+\infty} f(t)e^{-\mathrm{i}\omega t}\mathrm{d}t = \int_{-\infty}^{+\infty} f(t)e^{-\mathrm{i}2\pi f t}\mathrm{d}t = \int_{-\frac{a}{2}}^{+\frac{a}{2}} h e^{-\mathrm{i}2\pi f t}\mathrm{d}t \\ &= h\int_{-\frac{a}{2}}^{+\frac{a}{2}} e^{-\mathrm{i}2\pi f t}\mathrm{d}t = -\frac{h}{\mathrm{i}2\pi f}\left[e^{-\mathrm{i}2\pi f t}\right]_{-\frac{a}{2}}^{+\frac{a}{2}} = -\frac{h}{\mathrm{i}2\pi f}\left(e^{-\mathrm{i}\pi f a} - e^{\mathrm{i}\pi f a}\right) \\ &= -\frac{h}{\mathrm{i}2\pi f}\left(e^{\mathrm{i}(-\pi f a)} - e^{\mathrm{i}(\pi f a)}\right) = -\frac{h}{\mathrm{i}2\pi f}(-2\mathrm{i}\sin(\pi f a)) = h\frac{\sin(\pi f a)}{\pi f} \\ &= ha\frac{\sin(\pi f a)}{\pi f a} = ha\mathrm{sinc}(\pi f a)\end{aligned} \tag{6.4.530}$$

where sinc is the sine cardinal as we already know (see section Trigonometry page 2517).



So we fall back on the sinc and if we take the squared modulus squared we therefore get the decomposition of a theoretical monochromatic wave diffracted by a rectangular slot!!! Thus, it seems possible to study the diffraction phenomena using the Fourier transform and this field is named "Fourier optics". We'll come back on this later in the section of Functional Analysis when we deepen the Fourier transforms.

We know that the spectrum (described by the sinc function) crosses zero every time that the sine function is zero, that is to say, every time the frequency is a multiple of $1/a$.

The spectrum of this pulse illustrates two important points regarding the limited time signals:

P1. A short signal has a broadband spectrum.

P2. To a narrow spectrum correspond a long-term signal.

E2. The Fourier transform of an integrable function f is given as we know now by:

$$F(\omega) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} f(t)e^{-\mathrm{i}\omega t}\mathrm{d}t \tag{6.4.531}$$

Consider the integrable Gaussian function of the type :

$$f(t)=e^{-at^2} \tag{6.4.532}$$

with $a>0$ defined on $\mathbb{R}$.

We want to compute its Fourier transform because it is a very important case and particularly useful for solving the heat equation that we will treat in the section of Thermodynamics and also to solve the differential equation of Black & Scholes in the section Economy.

The brilliant trick, if we want to avoid making complex analysis on 3 A4 pages, is to notice that $F(\omega)$ is the solution of the following linear differential equation:

$$y'(\omega)+\frac{\omega}{2a}y=0 \tag{6.4.533}$$

where y is a function of ω.

Indeed deriving $F(\omega)$ we get:

$$\begin{aligned}F'(\omega)&=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}f(t)e^{-\mathrm{i}\omega t}\mathrm{d}t=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}e^{-at^2}\frac{\mathrm{d}}{\mathrm{d}\omega}e^{-\mathrm{i}\omega t}\mathrm{d}t\\&=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}e^{-at^2}\left(-\mathrm{i}te^{-\mathrm{i}\omega t}\right)\mathrm{d}t=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}\left(-\mathrm{i}te^{-at^2}\right)e^{-\mathrm{i}\omega t}\mathrm{d}t\end{aligned} \tag{6.4.534}$$

Integration by parts gives us:

$$\begin{aligned}F'(\omega)&=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}e^{-\mathrm{i}\omega t}\left(-\mathrm{i}te^{-at^2}\right)\mathrm{d}t\\&=\frac{1}{\sqrt{2\pi}}\left[e^{-\mathrm{i}\omega t}\frac{\mathrm{i}e^{-at^2}}{2a}\bigg|_{-\infty}^{+\infty}-\int\limits_{-\infty}^{+\infty}\left(-\mathrm{i}\omega e^{-\mathrm{i}\omega t}\right)\left(\frac{\mathrm{i}e^{-at^2}}{2a}\right)\mathrm{d}t\right]\\&=\frac{-\omega}{2a}\frac{1}{\sqrt{2a}}\int\limits_{-\infty}^{+\infty}e^{-\mathrm{i}\omega t}e^{-at^2}\mathrm{d}t=-\frac{\omega}{2a}\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}e^{-at^2}e^{-\mathrm{i}\omega t}\mathrm{d}t\end{aligned} \tag{6.4.535}$$

We recognize the expression of the Fourier transform of f. Therefore:

$$F'(\omega)=-\frac{\omega}{2a}F(\omega) \tag{6.4.536}$$

This shows that the $F(\omega)$ is solution of the differential equation above.

We have proved in the section of Differential and Integral Calculus that the general solution of this differential equation is given by:

$$y=Ae^{-G(\omega)} \tag{6.4.537}$$

where $A \in \mathbb{R}$. And as in the present case:

$$g(x)=\frac{\omega}{2a} \tag{6.4.538}$$

The primitive $G(x)$ is therefore easy to calculate and we get:

$$y=Ae^{-\frac{\omega^2}{4a}} \tag{6.4.539}$$

Therefore:

$$F(\omega)=Ae^{-\frac{\omega^2}{4a}} \tag{6.4.540}$$

To determine the constant A it suffices to note that:

$$F(0)=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} e^{-at^2}e^{-\mathrm{i}0t}\mathrm{d}t=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} e^{-at^2}\mathrm{d}t=\frac{1}{\sqrt{2\pi}}\sqrt{\frac{\pi}{a}}=\frac{1}{\sqrt{2a}} \tag{6.4.541}$$

and therefore:

$$F(\omega)=\frac{1}{\sqrt{2a}}e^{-\frac{\omega^2}{4a}} \tag{6.4.542}$$

It is then usual to say that the Fourier transform of a Gaussian is another Gaussian!!

E3. This last example is very important to understand the Nyquist sampling (see section Electrical Engineering page 8506). We consider a square pulse in the frequency domain:

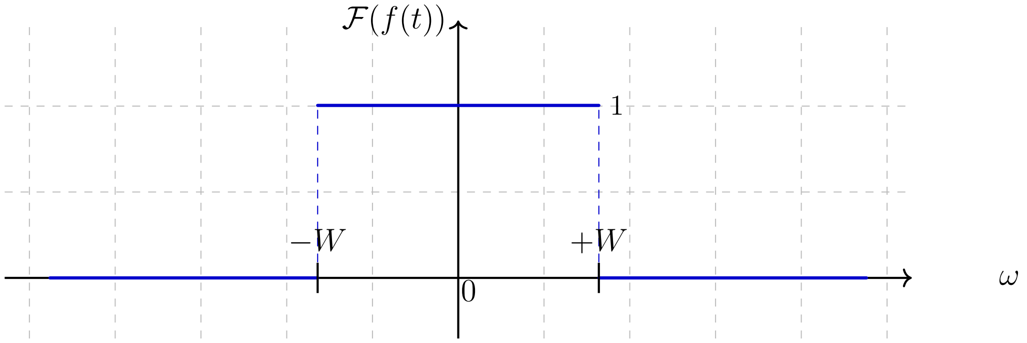


Figure 6.93: **Rectangular pulse example for Inverse Fourier Transform (author: Izaak Neutelings)**

Let us calculate its inverse Fourier transform:

$$\begin{aligned}\mathcal{F}^{-1}(F(\omega))(t)&=\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} F(\omega)e^{\mathrm{i}\omega t}\mathrm{d}\omega=\frac{1}{2\pi}\int\limits_{-W}^{+W} 1\cdot e^{\mathrm{i}\omega t}\mathrm{d}\omega\\&=\frac{1}{2\pi}\frac{1}{\mathrm{i}t}\left[e^{\mathrm{i}\omega t}\Big|_{-W}^{+W}\right]=\frac{1}{2\pi}\frac{1}{\mathrm{i}t}\left(e^{\mathrm{i}\omega t}-e^{-\mathrm{i}\omega t}\right)=\frac{1}{\pi t}\sin(Wt)=\frac{W}{\pi}\mathrm{sinc}(Wt)\end{aligned} \tag{6.4.543}$$

There is something important to notice here (or to remember whatever...)!!! First, we know that sinc is equal to 1 at $t=0$. But we also know that every time $Wt=k\pi$ with $k\in\mathbb{N}^*$ we have $\text{sinc}=0$. Therefore, every time $t=k\pi/W$ we have:

$$\mathcal{F}^{-1}(F(\omega))(t)=\frac{W}{\pi}\text{sinc}(Wt)=0 \tag{6.4.544}$$

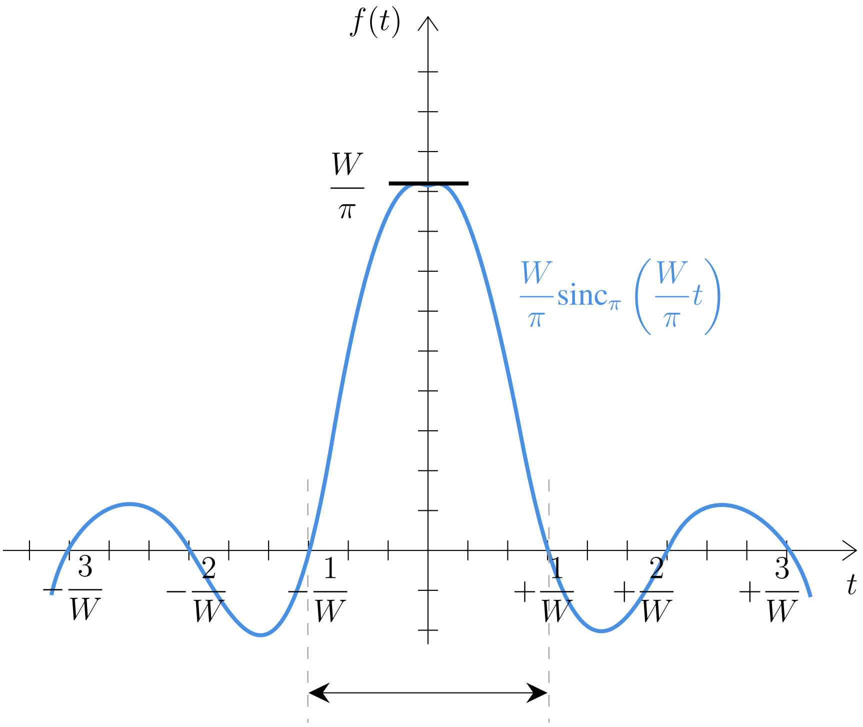
To make things more nice, signal processing engineers like to write rather the previous result into the form:

$$\mathcal{F}^{-1}(F(\omega))(t)=\frac{W}{\pi}\text{sinc}\left(\pi\frac{W}{\pi}t\right)=\frac{W}{\pi}\text{sinc}_\pi\left(\frac{W}{\pi}t\right) \tag{6.4.545}$$

and as you have almost surely noticed... this introduce a new type of sine cardinal (sic...!). You may ask yourself why....????? That's quite easy. Written this way the sinc_π cancels for:

$$t=\frac{k}{W}\qquad\forall k\in\mathbb{N}^* \tag{6.4.546}$$

This is prettier.... and even has a name: the "normalized sinc function" and result in the following well known plot in signal processing for the study of the Nyquist sampling theorem:



width of the main lobe is equal to $2/W$

We will see plenty of other Fourier transform in the section of Function Analysis page 2319.

6.4.2.8 Bessel Series

Bessel functions are very useful in many advanced fields of physics involving delicate differential equations to solve. The areas in which we find them most often are calorimetry (heat conduction), nuclear physics (physics of reactors), optics and fluid mechanics.

This series are still not study too much in the graduate curriculum and it is often the role of the student to seek the additional information it needs on this subject in the library of his school. We wanted to present here the developments that avoid this approach while staying at home in front of our computer (furthermore books on the subject are quite rare...).

> **Remark**
>
> We usually speak by abuse of language of "Bessel functions" instead of "Bessel series".

There is a significant amount of Bessel functions but we will restrict ourselves to the study of those most used one in physics.

6.4.2.8.1 Zero order Bessel's Functions

The function known as the "Zero order Bessel's function", is defined by the power series:

$$\boxed{J_0(x)=1-\frac{x^2}{2^2}+\frac{x^4}{2^2\cdot 4^2}-\frac{x^6}{2^2\cdot 4^2\cdot 6^2}+\ldots} \tag{6.4.547}$$

It is during the study of the properties of derivation and integration that Friedrich Bessel found that this series of power is a solution to a differential equation that is found frequently in physics. That is why it bears his name.

If u_r represents the r-th term of the series, we easily see that:

$$\frac{u_{r+1}}{u_r}=\frac{-x^2}{(2r)^2} \tag{6.4.548}$$

which tends to zero as $r\rightarrow +\infty$, regardless of the value of x. This has for consequence that the series converges for all values of x. Since this is a series of positive power, the function $J_0(x)$ and all its derivatives are continuous function for all values of x, real or complex.

6.4.2.8.2 n order Bessel's Functions

The function $J_n(x)$, known as the "n order Bessel's function", is defined, when n is a positive integer, by the power series:

$$\boxed{J_n(x)=\frac{x^n}{2^n n!}\left(1-\frac{x^2}{2\cdot(2n+2)}+\frac{x^4}{2\cdot 4(2n+2)(2n+4)}-\ldots\right)} \tag{6.4.549}$$

which converges for all values of x, real or complex.

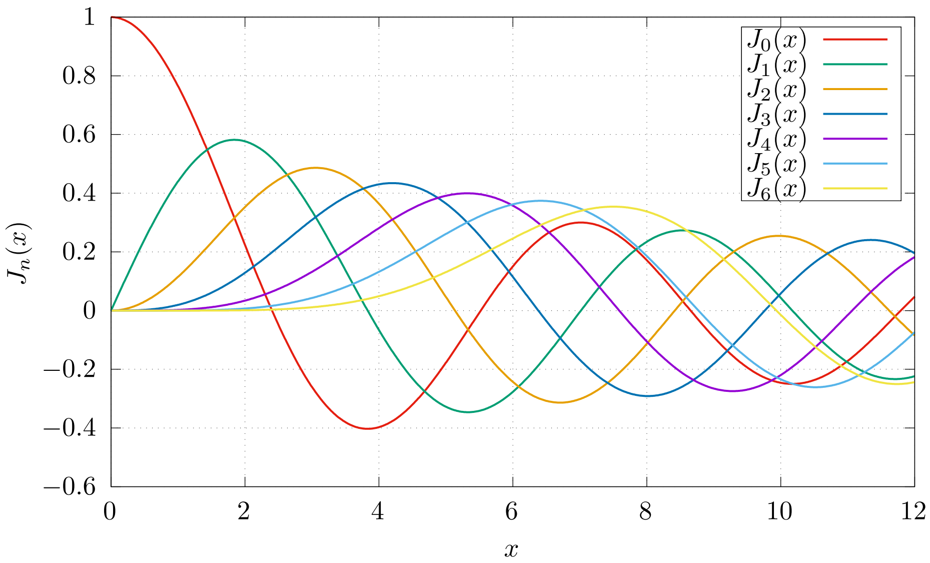


Figure 6.94: **Plot of a few Bessel functions (source: Wikipedia)**

and in Microsoft® Excel® or Maple™ 4.00b the previous function can be found under the name `BESSELJ( )`. For example for the previous graph in Maple™, we just write:

```
>plot([BesselJ(0,x),BesselJ(1,x),BesselJ(2,x),BesselJ(3,x)],x=0..20);
```

Let us see in particular, that for $n=1$ we have:

$$J_1(x)=\frac{x}{2}-\frac{x^3}{2^2\cdot 4}+\frac{x^5}{2^2\cdot 4^2\cdot 6}-\frac{x^7}{2^2\cdot 4^2\cdot 6^2\cdot 8}+\ldots \qquad (6.4.550)$$

and when $n=2$:

$$J_2(x)=\frac{x^2}{2\cdot 4}-\frac{x^4}{2^2\cdot 4\cdot 6}+\frac{x^6}{2^2\cdot 4^2\cdot 6\cdot 8}-\frac{x^8}{2^2\cdot 4^2\cdot 6^2\cdot 8\cdot 10}+\ldots \qquad (6.4.551)$$

We can notice that $J_n(x)$ is an even function of x when n is even and odd if n is odd (see section Functional Analysis page 2212).

If we play to do engineering maths we notice by trial and errors that:

$$\begin{aligned}J_n(x)&=\frac{x^n}{2^n n!}\left(1-\frac{x^2}{2\cdot(2n+2)}+\frac{x^4}{2\cdot 4(2n+2)(2n+4)}-\ldots\right)\\&=\left(\frac{x}{2}\right)^n\frac{1}{n!}\sum_{k=0}^{+\infty}\left((-1)^k\cdot(???)\right)\left(\frac{x}{2}\right)^{2k}\end{aligned} \qquad (6.4.552)$$

Based on this trial and error approach we have using the above expression by factorizing the $(x/2)^{2k}$ term only:

$$\begin{aligned}J_0&=1-\frac{1}{1}\frac{x^2}{2^2}+\frac{1}{2}\frac{x^4}{2^4}-\frac{1}{36}\frac{x^6}{2^6}+\ldots\\J_1&=\frac{x}{2\cdot 1!}\left(1-\frac{x^2}{2^2\cdot 2}+\frac{x^4}{2^2\cdot 4^2\cdot 3}-\frac{x^6}{2^2\cdot 4^2\cdot 6^2\cdot 4}+\ldots\right)\\&=\frac{x}{2}\left(\frac{1}{1}-\frac{1}{2}\frac{x^2}{2^2}+\frac{1}{2^2\cdot 3}\frac{x^4}{2^4}-\frac{1}{6^2\cdot 4}\frac{x^6}{2^6}+\ldots\right)\\J_2&=\ldots\end{aligned} \qquad (6.4.553)$$

So finally we see after trial and errors again (instead than doing 5 pages of mathematical developments as do mathematicians) that:

$$\boxed{J_n=\left(\frac{x}{2}\right)^n\sum_{k=0}^{+\infty}\frac{(-1)^k}{k!(n+k)!}\left(\frac{x}{2}\right)^{2k}=\sum_{k=0}^{+\infty}\frac{(-1)^k}{k!(n+k)!}\left(\frac{x}{2}\right)^{2k+n}} \tag{6.4.554}$$

More generally for n that is non-integer we use the Euler Gamma function (see section Differential and Integral Calculus page 1574):

$$J_n=\sum_{k=0}^{+\infty}\frac{(-1)^k}{k!\Gamma(n+k+1)}\left(\frac{x}{2}\right)^{2k+n} \tag{6.4.555}$$

That is also sometimes written by pure mathematicians (...):

$$J_n=\sum_{k=0}^{+\infty}\frac{(-1)^k}{\Gamma(k+1)\Gamma(n+k+1)}\left(\frac{x}{2}\right)^{2k+n} \tag{6.4.556}$$

Now by differentiating the function $J_0(x)$ and comparing the result with the series $J_1(x)$, we see that without much pain that:

$$\frac{\mathrm{d}J_0(x)}{\mathrm{d}x}=-J_1(x) \tag{6.4.557}$$

We also find without too much difficulty, the following relation:

$$\frac{\mathrm{d}}{\mathrm{d}x}xJ_1(x)=xJ_0(x) \tag{6.4.558}$$

Remark

In general by recursive reasoning we get:

$$\boxed{\frac{\mathrm{d}}{\mathrm{d}x}(x^nJ_n(x))=x^nJ_{n-1}(x)} \tag{6.4.559}$$

Therefore:

$$\boxed{x^nJ_n(x)=\int_0^x y^nJ_{n-1}(y)\mathrm{d}y} \tag{6.4.560}$$

This last relation will be useful to us in the section of Wave Optics for our study of the circular aperture diffraction (Airy disk).

Using the fact that:

$$\frac{\mathrm{d}J_0(x)}{\mathrm{d}x}=-J_1(x) \tag{6.4.561}$$

and including it in the previous relation, we find:

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(x\frac{\mathrm{d}J_0(x)}{\mathrm{d}x}\right)+xJ_0(x)=0 \tag{6.4.562}$$

written in another way:

$$x\frac{\mathrm{d}^2J_0(x)}{\mathrm{d}x^2}+\frac{\mathrm{d}J_0(x)}{\mathrm{d}x}+xJ_0(x)=0 \tag{6.4.563}$$

$y=J_0(x)$ is therefore a solution of the differential equation of the second order:

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(x\frac{\mathrm{d}y}{\mathrm{d}x}\right)+xy=0 \tag{6.4.564}$$

written otherwise:

$$x\frac{\mathrm{d}^2y}{\mathrm{d}x^2}+\frac{\mathrm{d}y}{\mathrm{d}x}+xy=0 \tag{6.4.565}$$

or even:

$$\frac{\mathrm{d}^2y}{\mathrm{d}x^2}+\frac{1}{x}\frac{\mathrm{d}y}{\mathrm{d}x}+y=0 \tag{6.4.566}$$

A solution to an equation of parameter n which is not a multiple of $J_n(x)$ is named "Bessel function of the second kind". Let us suppose now that u is such a function and let us put $v=J_0(x)$; then according to the relation:

$$x\frac{\mathrm{d}^2y}{\mathrm{d}x^2}+\frac{\mathrm{d}y}{\mathrm{d}x}+xy=0 \tag{6.4.567}$$

we have:

$$xu''+u'+xu=0 \quad \text{and} \quad xv''+v'+xv=0 \tag{6.4.568}$$

Multiplying the first relation by v and the second by u and after subtracting, we get:

$$x(u''v-uv'')+u'v-uv'=0 \tag{6.4.569}$$

we therefore also have:

$$\frac{\mathrm{d}}{\mathrm{d}x}(u'v-uv')=u''v+u'v'-u'v'-uv''=u''v-uv'' \tag{6.4.570}$$

we can therefore write:

$$\frac{\mathrm{d}}{\mathrm{d}x}[x(u'v-uv')]=0 \tag{6.4.571}$$

Indeed, because if we develop, we find:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}[x(u'v-uv')]&=1\cdot u'v+xu''v+xu'v'-1\cdot uv'-xu'v'-xuv''\\&=x(u''v+u'v'-u'v'-uv'')+u'v-uv'\\&=x(u''v-uv'')+u'v-uv'=0\end{aligned} \tag{6.4.572}$$

For the equality:

$$\frac{\mathrm{d}}{\mathrm{d}x}[x(u'v-uv')]=0 \tag{6.4.573}$$

is satisfied, we have:

$$x(u'v-uv')=c^{te}=B \tag{6.4.574}$$

Dividing by xv^2, we have:

$$\frac{(u'v-uv')}{v^2}=\frac{B}{xv^2} \tag{6.4.575}$$

which is equivalent to:

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{u}{v}\right)=\frac{B}{xv^2} \tag{6.4.576}$$

immediately, by integrating it comes:

$$\frac{u}{v}=A+B\int\frac{1}{xv^2}\mathrm{d}x \tag{6.4.577}$$

where A is a constant. Consecutively we have since $v=J_0(x)$:

$$u=AJ_0(x)+BJ_0(x)\int\frac{1}{xJ_0^2(x)}\mathrm{d}x \tag{6.4.578}$$

where we recall, A and B are constants, and $B\neq 0$ if u is not a multiple of $J_0(x)$ by definition!

If in the last relation, $J_0(x)$ is replaced by its expression in terms of series we have:

$$\frac{1}{xJ_0^2(x)}=\frac{1}{x}+\frac{x}{2}+\frac{5x^3}{32}+\ldots \tag{6.4.579}$$

For those who want to check this last relation (I do not like this kind of algebraic calculations) in Maple™ 4.00b just write:

```
>1/x*taylor(1/(series(BesselJ(0,x),x))^2,x=0,5);
```

Therefore:

$$\begin{aligned}J_0(x)\int\frac{1}{xJ_0^2(x)}\mathrm{d}x&=J_0(x)\left[\log(x)+\frac{x^2}{4}+\frac{5x^4}{128}+\ldots\right]\\&=J_0(x)\log(x)+\left(1-\frac{x^2}{2^2}+\ldots\right)\left(\frac{x^2}{4}+\frac{5x^4}{128}+\ldots\right)\\&=J_0(x)\log(x)+\frac{x^2}{4}-\frac{3x^4}{128}+\ldots\end{aligned} \tag{6.4.580}$$

consecutively if we put:

$$\boxed{Y_0(x)=J_0(x)\log(x)+\frac{x^2}{4}-\frac{3x^4}{128}+\ldots} \tag{6.4.581}$$

where $Y_0(x)$ is a particular Bessel function of the second type named "Bessel-Neumann function of the second kind of zero order".

Identically to the fact that when $J_0(x) \to 0$ when $x \to 0$, the expression $Y_0(x)$ because of the term $\log(x)$ when x is small approaches $Y_0(x) \to -\infty$ when $x \to +0$.

Finally, it comes from what we have seen that $J_0(x)$ and $Y_0(x)$ are independent solutions of the differential equation:

$$\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} + \frac{1}{x}\frac{\mathrm{d}y}{\mathrm{d}x} + y = 0 \tag{6.4.582}$$

The general solution being therefore:

$$y = AJ_0(x) + BY_0(x) \tag{6.4.583}$$

where A, B are arbitrary constants and $x > 0$ so that $Y_0(x)$ is real.

If we replace x by kx, where k is a constant, the differential equation becomes:

$$\frac{1}{k^2}\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} + \frac{1}{kx}\frac{\mathrm{d}y}{k\mathrm{d}x} + y = 0 \tag{6.4.584}$$

by multiplying the whole by k^2 we find the general form of the differential equation:

$$\boxed{\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} + \frac{1}{x}\frac{\mathrm{d}y}{\mathrm{d}x} + k^2 y = 0} \tag{6.4.585}$$

whose general solution is:

$$\boxed{y = AJ_0(kx) + BY_0(kx)} \tag{6.4.586}$$

where $k > 0$ such that $Y_0(kx)$ is real when $x > 0$.

In fact, the Bessel functions are solutions of the differential equation previously studied and solved by the "Frobenius method". Indeed, let us write:

$$L_y \equiv \left(x\frac{\mathrm{d}^2}{\mathrm{d}x^2} + \frac{\mathrm{d}}{\mathrm{d}x} + x \right) y \tag{6.4.587}$$

and let us make the substitution:

$$y = x^\rho \sum_{k=0}^{+\infty} c_k x^k = x^\rho + c_1 x^{\rho+1} + c_2 x^{\rho+2} + c_3 x^{\rho+3} + \ldots \tag{6.4.588}$$

substituting in Ly, we get:

$$Ly = \rho^2 x^{\rho-1} + (\rho+1)^2 c_1 x^\rho + \left[(\rho+2)^2 c_2 + 1\right] x^{\rho+1} + \left[(\rho+3)^2 c_3 + c_1\right] x^{\rho+1} \tag{6.4.589}$$

Now let us choose the c_i to satisfy the differential equation such as:

$$\begin{aligned} (\rho+1)^2 c_1 &= 0 \\ (\rho+2)^2 c_2 + 1 &= 0 \\ (\rho+3)^2 c_3 + c_1 &= 0 \\ &\ldots \end{aligned} \tag{6.4.590}$$

Therefore, unless ρ is a negative integer, we have:

$$\begin{aligned}
&c_1 = c_3 = c_5 = c_7 = \ldots = 0\\
&c_2 = -\frac{1}{(\rho+2)^2}\\
&c_4 = -\frac{c_2}{(\rho+4)^2} = \frac{1}{(\rho+2)^2(\rho+4)^2}\\
&\ldots
\end{aligned} \tag{6.4.591}$$

By substituting these values in the relation:

$$y = x^\rho + c_1 x^{\rho+1} + c_2 x^{\rho+2} + c_3 x^{\rho+3} + \ldots \tag{6.4.592}$$

we get:

$$\boxed{y = x^\rho \left(1 - \frac{x^2}{(\rho+2)^2} + \frac{1}{(\rho+2)^2(\rho+4)^2} - \cdots\right)} \tag{6.4.593}$$

Therefore:

$$x\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} + \frac{\mathrm{d}y}{\mathrm{d}x} + xy = Ly = \rho^2 x^{\rho-1} \tag{6.4.594}$$

If we put $\rho =$ in the prior-previous relation, we get:

$$y = x^\rho \left(1 - \frac{x^2}{(\rho+2)^2} + \frac{1}{(\rho+2)^2(\rho+4)^2} - \cdots\right) = J_0(x) \tag{6.4.595}$$

6.4.2.8.3 Bessel's Differential Equations of order n

We have defined the Bessel series as:

$$J_n(x) = \frac{x^n}{2^n n!}\left(1 - \frac{x^2}{2(2n+2)} + \frac{x^4}{2\cdot 4\cdot(2n+2)(2n+4)} - \cdots\right) \tag{6.4.596}$$

Let us put:

$$y = J_n(x)2^n n! = x^n - \frac{x^{n+2}}{2(2n+2)} + \frac{x^{n+4}}{2\cdot 4(2n+2)(2n+4)} - \cdots \tag{6.4.597}$$

and let us derivate as follows:

$$x\frac{\mathrm{d}}{\mathrm{d}x}\left(x\frac{\mathrm{d}y}{\mathrm{d}x}\right) = n^2 x^n - \frac{(n+2)^2 x^{n+2}}{2(2n+2)} + \frac{(n+4)^2 x^{n+4}}{2\cdot 4(2n+2)(2n+4)} - \cdots \tag{6.4.598}$$

But we also have:

$$n^2 y = n^2 x^n - \frac{n^2 x^{n+2}}{2(2n+2)} + \frac{n^2 x^{n+4}}{2\cdot 4(2n+2)(2n+4)} - \cdots \tag{6.4.599}$$

By subtraction:

$$x\frac{\mathrm{d}}{\mathrm{d}x}\left(x\frac{\mathrm{d}y}{\mathrm{d}x}\right) - n^2 y = -x^{n+2} + \frac{x^{n+4}}{2(2n+2)} - \cdots = -x^2 y \tag{6.4.600}$$

Which finally gives:

$$x\frac{\mathrm{d}}{\mathrm{d}x}\left(x\frac{\mathrm{d}y}{\mathrm{d}x}\right)+(x^2-n^2)y=0 \tag{6.4.601}$$

This is also written:

$$\boxed{x^2\frac{\mathrm{d}^2y}{\mathrm{d}x}+x\frac{\mathrm{d}y}{\mathrm{d}x}+(x^2-n^2)y=0} \tag{6.4.602}$$

which is named the "Bessel differential equation of order n" or more simply "Bessel equation". In fact, most schools or Internet sites give this differential equation as a definition but now it is clear that there is rigorous reasoning behind this equation.

The solution is therefore of the type:

$$y=2^n n!J_n(x) \tag{6.4.603}$$

which is still sometimes written using the gamma Euler function (see section Differential and Integral Calculus page 1574):

$$y=2^n\Gamma_0(n)J_n(x)=c^{te}J_n(x) \tag{6.4.604}$$

It follows that:

$$x^2\frac{\mathrm{d}^2J_n(x)}{\mathrm{d}x^2}+x\frac{\mathrm{d}J_n(x)}{\mathrm{d}x}+(x^2-n^2)J_n(x)=0 \tag{6.4.605}$$

and therefore that $y=J_n(x)$ is the solution of this differential equation.

We will fall back on such a differential equation during our study of the wave equation of a circular drum (see section Wave Mechanics page 3500), during our study of the physics of nuclear reactors (see section Nuclear Physics) and finally during our study of self-buckling (see section Mechanical Engineering page 8415).

6.4.3 Convergence Criteria

When we study a series, one of the fundamental questions is that of the convergence or divergence of this series.

If a series converges, its general term approaches zero as n approaches infinity:

$$\lim_{n\to+\infty} a_n=0 \tag{6.4.606}$$

or obviously generally:

$$\lim_{n\to+\infty} a_n=c^{te} \tag{6.4.607}$$

This criterion is necessary but insufficient to establish the convergence of a series. By cons, if this criterion is not met, we are absolutely sure that the series does not converge (so it diverges!).

Three methods are proposed to deepen the convergence criteria:

1. The integral test
2. The d'Alembert rule
3. The Cauchy rule

In the following paragraphs, we will assume the series with positive terms. The case of alternating series will be seen later.

6.4.3.1 Integral Test

The integral test for convergence is a method used to test infinite series of non-negative terms for convergence. It was developed by Colin Maclaurin and Augustin-Louis Cauchy and is sometimes known as the "Maclaurin–Cauchy test".

Given the series with decreasing positive (monotone decreasing) terms:

$$a_1+a_2+a_3+\cdots+a_n+\cdots \tag{6.4.608}$$

That is to say:

$$a_1\geq a_2\geq a_3\geq\cdots\geq a_n\geq\cdots \tag{6.4.609}$$

and given a continuous decreasing function such that:

$$f(1)=a_1,f(2)=a_2,\ldots,f(n)=a_n \tag{6.4.610}$$

Then the infinite series:

$$\sum_{n=1}^{+\infty} f(n) \tag{6.4.611}$$

converges to a real number if and only if the improper integral:

$$\int\limits_1^{+\infty} f(x)\mathrm{d}x \tag{6.4.612}$$

is finite. In other words, if the integral diverges, then the series diverges as well.

Remark

In no case does the integral gives the value of the sum of the series! The full test only gives an indication of the convergence of the series. Before making the test of the integral, it is important to check that the terms of the series are strictly decreasing to fill the condition $a_1\geq a_2\geq a_3\geq\cdots\geq a_n\geq\cdots$.

✍**Example:**

The harmonic series:

$$\sum_{n=1}^{+\infty}\frac{1}{n} \tag{6.4.613}$$

diverges because (see section Differential and Integral Calculus page 1534):

$$\int\limits_{1}^{+\infty}\frac{1}{n}\mathrm{d}n=\ln(n)\Big|_{1}^{+\infty}=\ln(+\infty)\rightarrow+\infty \tag{6.4.614}$$

So this harmonic series does not converge.

6.4.3.2 D'Alembert Rule

The "ratio test" is also a test (or "criterion") for the convergence of a series of the type:

$$\sum_{n=1}^{+\infty}a_n \tag{6.4.615}$$

where each term is a real or complex number and a_n is non-zero when n is large. The test was first published by Jean le Rond d'Alembert and is sometimes known as "d'Alembert's ratio test" or as the "Cauchy ratio test".

The usual form of the test makes use of the limit:

$$L=\lim_{n\rightarrow+\infty}\left|\frac{a_{n+1}}{a_n}\right| \tag{6.4.616}$$

The ratio test states that:

1. if $L<1$ then the series converges absolutely;
2. if $L>1$ then the series does not converge;
3. if $L=1$ or the limit fails to exist, then the test is inconclusive, because there exist both convergent and divergent series that satisfy this case

and we define the "radius of convergence" by:

$$R=\frac{1}{L} \tag{6.4.617}$$

For the proof suppose that:

$$L=\lim_{n\rightarrow+\infty}\left|\frac{a_{n+1}}{a_n}\right|<1 \tag{6.4.618}$$

Remark

In reality this rule is normally without the absolute value. The case with the absolute value as above is named the "absolute convergence test" and applied for the more general case of an alternate series such that:

$$a_1+|-a_2|+a_3+|-a_4|+\ldots+|-a_n|+\ldots \qquad (a_n>0) \tag{6.4.619}$$

If an alternating series of terms is absolutely convergent, the absolute series that follows also converge.

Therefore the absolute convergence test is a generalization of the d'Alembert rule but most of time we don't make any distinctions between the both.

We can then show that the series converges absolutely by showing that its terms will eventually become less than those of a certain convergent geometric series. To do this, let:

$$r=\frac{L+1}{2} \tag{6.4.620}$$

Then r is strictly between L and 1, and:

$$|a_{n+1}|<r|a_n| \tag{6.4.621}$$

for sufficiently large n (say, n greater than N).

Hence:

$$|a_{n+i}|<r^i|a_n| \tag{6.4.622}$$

for each $n>N$ and $i>0$, and so:

$$\sum_{i=N+1}^{+\infty}|a_i|=\sum_{i=1}^{+\infty}|a_{N+i}|<\sum_{i=1}^{+\infty}r^i|a_{N+1}|=|a_{N+1}|\sum_{i=1}^{+\infty}r^i=|a_{N+1}|\frac{r}{1-r}<+\infty \tag{6.4.623}$$

That is, the series converges absolutely.

On the other hand, if $L>1$, then:

$$|a_{n+1}|>|a_n| \tag{6.4.624}$$

for sufficiently large n, so that the limit of the sum is non-zero. Hence the series diverges.

✍**Example:**

Given the following geometric series:

$$\sum_{n=1}^{+\infty}\left(\frac{1}{2}\right)^n=\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\cdots \tag{6.4.625}$$

We get the quotient:

$$L=\frac{a_{n+1}}{a_n}=\frac{0.5^{n+1}}{0.5^n}=0.5 \tag{6.4.626}$$

Therefore the series converge!

Obviously some practical applications will (can) give for example:

$$L=\lim_{n\rightarrow+\infty}\sqrt[n]{a_n} \tag{6.4.627}$$

and some practitioners name this the "Cauchy convergence rule"... (do not confuse with the Cauchy convergence test that will be study in the section Fractals at page 7038).

✍**Example:**

Let us do now a last important example. We studied earlier above Bessel series. But it can be not obvious that these series converge. Let us prove that this is indeed the case for J_0, that is to say for:

$$J_0=\sum_{k=0}^{+\infty}\frac{(-1)^k}{k!(n+k)!}\left(\frac{x}{2}\right)^{2k+n}=\sum_{k=0}^{+\infty}\frac{(-1)^k x^{2k}}{2^{2k}k!^2} \tag{6.4.628}$$

Therefore:

$$\begin{aligned}L&=\lim_{k\rightarrow+\infty}\left|\frac{a_{n+1}}{a_n}\right|=\lim_{k\rightarrow+\infty}\left|\frac{(-1)^{k+1}x^{2(k+1)}}{2^{2(k+1)}[(k+1)!]^2}\cdot\frac{2^{2k}k!^2}{(-1)^k x^{2k}}\right|\\&=\lim_{k\rightarrow+\infty}\left|\frac{x^{2k+2}}{2^{2k+2}[(k+1)!]^2}\cdot\frac{2^{2k}(k!)^2}{x^{2k}}\right|\\&=\lim_{k\rightarrow+\infty}\left|\frac{x^{2k}x^2}{2^{2k}2^2(k+1)!(k+1)!}\cdot\frac{2^{2k}k!k!}{x^{2k}}\right|=\lim_{n\rightarrow+\infty}\left|\frac{x^2k!k!}{4(k+1)k!(k+1)k!}\right|\\&=\lim_{k\rightarrow+\infty}\left|\frac{x^2}{4(k+1)(k+1)}\right|=\lim_{k\rightarrow+\infty}\frac{1}{(k+1)^2}\left|\frac{x^2}{4}\right|=\lim_{n\rightarrow+\infty}\frac{x^2}{4(k+1)^2}\\&=0\end{aligned} \tag{6.4.629}$$

6.4.3.3 Alternating Series Test

The alternating series test is a method used to prove that an alternating series with terms that decrease in absolute value is a convergent series. The test was used by Gottfried Leibniz and is sometimes known as "Leibniz's test", "Leibniz's rule, or the "Leibniz criterion".

A series of the form:

$$\sum_{n=0}^{+\infty}(-1)^n a_n = a_0 - a_1 + a_2 - a_3 + \cdots \tag{6.4.630}$$

where either all a_n are positive or all a_n are negative, is named an "alternating series".

The alternating series test then says: if a_n decreases monotonically and that:

$$\lim_{n\rightarrow+\infty} a_n = 0 \tag{6.4.631}$$

then the alternating series converges!

There's a lot of other tests of the same kind as the Raabe's test, the Bertrand's test, the Gauss's test, the Kummer's test, etc.

6.4.3.4 Fixed Point Theorem

The fixed point theorem is not really useful in physics and for the engineers (but implicitly it is essential but physicists and engineers often use math tools whose properties have already been approved in advance by mathematicians), however we find it in Chaos Theory and in Theoretical Computing (see the section on Fractals at page 7038 especially the topic on the Sierpinski triangle). We can therefore only recommend the reader to take the time to read and understand the explanations and developments that follow.

Let (X,d), be a complete metric space (see sections Topology page 2447 or Fractals page 7054) and $T:X\mapsto X$ an application strictly contracting of constant L (see the Lipschitz functions in the section Topology page 2457), then there exists a unique point $\omega\in X$ such that:

$$T(\omega)=\omega \tag{6.4.632}$$

ω is then named the "fixed point" of T (think to the case $\cos(x)=x$). Furthermore, if denote by:

$$T^n(x)=\underbrace{T(T(T\ldots T}_{n \text{ times}}(x)\ldots)) \tag{6.4.633}$$

the image of x by the n-th iterate of T, then we have:

$$\omega=\lim_{t\rightarrow\infty}\left(T^n(x)\right) \quad \forall x\in X \tag{6.4.634}$$

and the convergence speed can also be estimated by:

$$\mathrm{d}\left(\omega,T^n(x)\right)\leq \mathrm{d}\left(x,T(x)\right)\frac{L^n}{1-L} \tag{6.4.635}$$

By the fact that we restrict our study to iterating a function, we speak of "Banach fixed-point theorem" that gives a general criterion guaranteeing that, if it is satisfied, the procedure of iterating a function yields a fixed point.

Remark

You can have fun with your pocket calculator or your operating system by choosing a random number and taking the cosine iteratively. You will find you will tend to 0.74 and therefore it is verbatim the solution of $\cos(x)=x$.

Proof 6.25.2. Given $x\in X$. We consider the following sequence $(T^n(x))_{n\in\mathbb{N}}$ as defined above. First we will prove that this sequence is a Cauchy sequence (see above what is a Cauchy sequence).

Applying the triangle inequality (see section Vector Calculus page 1828) several times we have:

$$\begin{aligned}\mathrm{d}\big(T^{n+k}(x),T^n(x)\big) \leq &\mathrm{d}\big(T^{n+k}(x),T^{n+k-1}(x)\big)+\mathrm{d}\big(T^{n+k-1}(x),T^{n+k-2}(x)\big)\\ &+\cdots+\mathrm{d}\big(T^{n+1}(x),T^n(x)\big)\end{aligned}\tag{6.4.636}$$

But:

$$\mathrm{d}\big(T^{n+1}(x),T^n(x)\big)\leq L\cdot\mathrm{d}\big(T^n(x),T^{n-1}(x)\big)\leq\cdots\leq L^n\cdot\mathrm{d}\big(T(x),x\big)\tag{6.4.637}$$

Therefore:

$$\mathrm{d}\big(T^{n+k}(x),T^n(x)\big)\leq(L^{n+k-1}+L^{n+k-2}+\cdots+L^n)\cdot\mathrm{d}\big(T(x),x\big)=L^n\cdot d\big(T(x),x\big)\cdot(1+L+L^2+\ldots+L^{k-1})\tag{6.4.638}$$

and as:

$$(1+L+L^2+\ldots+L^{k-1})\leq\frac{1-L^k}{1-L}\leq\frac{1}{1-L}\tag{6.4.639}$$

Therefore:

$$\mathrm{d}\big(T^{n+k}(x),T^n(x)\big)\leq\frac{L^n}{1-L}\cdot\mathrm{d}\big(T(x),x\big)\tag{6.4.640}$$

To finish:

$$0\leq L\leq 1\Rightarrow\lim_{n\rightarrow+\infty}\frac{L^n}{1-L}=0\tag{6.4.641}$$

that is to say, that in a first time $(T^n(x))_{n\in\mathbb{N}}$ converge, and we put:

$$\omega=\lim_{n\rightarrow+\infty}T^n(x)\tag{6.4.642}$$

Now we check that ω is a fixed point of T. Indeed T is uniformly continuous (as Lipschitz - see section Topology page 2457) therefore a fortiori continue:

$$\omega=\lim_{n\rightarrow+\infty}T^n(x)=\lim_{n\rightarrow+\infty}T\big(T^{n-1}(x)\big)\underbrace{=}_{\text{as }T\text{ is continuous}}T\left(\lim_{n\rightarrow+\infty}T^{n-1}(x)\right)=T(\omega)\tag{6.4.643}$$

It remains to check that ω is the only fixed point (therefore we will have proved that ω does not depend on the choice of x). Suppose that we also have $T(y)=y$ then:

$$\mathrm{d}(y,\omega)=\big(T(y),T(\omega)\big)\leq L\cdot\mathrm{d}(y,\omega)\Rightarrow\mathrm{d}(y,\omega)=0\Rightarrow y=\omega \tag{6.4.644}$$

An estimate of the speed of convergence is given by:

$$\mathrm{d}\big(T^{n+k}(x),T^n(x)\big)\leq\frac{L^n}{1-L}\cdot\mathrm{d}\big(T(x),x\big) \tag{6.4.645}$$

$\mathrm{d}(,)$ is continuous with respect to each variable so:

$$\lim_{k\rightarrow+\infty}\mathrm{d}\big(T^{n+k}(x),T^n(x)\big)=\mathrm{d}\left(\lim_{k\rightarrow\infty}T^{n+k}(x),T^n(x)\right)=\mathrm{d}\big(\omega,T^n(x)\big) \tag{6.4.646}$$

and the limits preserve the inequalities (not strict one) thus:

$$\mathrm{d}\big(\omega,T^n(x)\big)\leq\frac{L^n}{1-L}\mathrm{d}\big(T(x),x\big) \tag{6.4.647}$$

■ Q.E.D.

6.4.4 Generating Functions

For some mathematical financial risk management models (see section Economy page 7485) and also integration transformation of Bessel functions (see section Differential and Integral Calculus page 1564) we will need in this textbook a gentle introduction to generating functions.

6.4.4.1 Ordinary Generating Functions (transformation of a sequence into a series)

Remember first that we have proved earlier above that the general Maclaurin expansion of a function was given by:

$$f(x)=\sum_{k=0}^{+\infty}\frac{f^{(n)}(0)}{n!}x^n \tag{6.4.648}$$

That is for $x\cong 0$:

$$f(x)=f(0)+f'(0)x+\frac{f''(0)}{2}x^2+\frac{f'''(0)}{6}x^3+\ldots \tag{6.4.649}$$

In the case of our study, let us write the latter relation as:

$$\boxed{G(x)=a_0+a_1x+a_2x^2+a_3x^3+\ldots=\sum_k a_kx^k} \tag{6.4.650}$$

We say then that the above relation is the "ordinary generating function" of the sequences of numbers $a_0,a_1,a_2,a_3,\ldots$ in the formal parameter x. And we don't care if it diverges or not!

Because the generating function is an algebraic expression that encodes the sequence and allows us to manipulate it in ways that are not possible in other forms. Many times if the sequence you are looking at is "interesting" (and this word has lots of interpretations), the generating function has a short simple form.

The generating function allows us to derive formulas for the sequence, identities involving the sequence, estimate the values and so much more as we will see further below.

One thing to never forget: The generating function is not the sequence and the sequence is not the generating function. They are not the same thing. One is a sequence, the other is an algebraic expression!

If you have a sequence you can say "the generating function of the sequence" to refer to the algebraic object. If you have a generating function you might say "the sequence of coefficients of the generating function" in order to refer to the sequence.

Let us try now a companion example, the sequence consisting of all 1:

$$1,1,1,1,1,1,1,\ldots \tag{6.4.651}$$

The generating function is therefore the geometric series that we have proved earlier above implicitly (explicitly you can see the proof at page 2419):

$$1+1x+1x^2+1x^3+\ldots=1+x+x^2+x^3+\ldots=\frac{1}{1-x} \tag{6.4.652}$$

The next simple companion example would be the positive integers:

$$1,2,3,4,5,\ldots \tag{6.4.653}$$

This has for generating function:

$$1+2x+3x^2+4x^3+\ldots \tag{6.4.654}$$

Now we observe that the derivative of this generating function is the generating function itself:

$$1+x+x^2+x^3+\ldots=\frac{1}{1-x} \tag{6.4.655}$$

Indeed:

$$\frac{\mathrm{d}}{\mathrm{d}x}(1+x+x^2+x^3+\ldots)=1+2x+3x^2+4x^3+\ldots \tag{6.4.656}$$

That is we have therefore:

$$\frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{1-x}=1+2x+3x^2+4x^3+\ldots \tag{6.4.657}$$

We conclude that:

$$\frac{1}{(1-x)^2}=1+2x+3x^2+4x^3+\ldots=\sum_{k=0}^{+\infty}(k+1)x^k \tag{6.4.658}$$

We can't use the exactly same trick to figure out the generating function for the sequence:

$$1,4,9,16,25 \tag{6.4.659}$$

because if we take the derivative of:

$$1+2x+3x^2+4x^3+\ldots \tag{6.4.660}$$

then we do not quite have the square integers. But the attentive reader will notice that if we first multiply the latter relation by x and then take the derivative then we have:

$$\begin{aligned}1+4x+9x^2+16x^3+\ldots&=1^2+2^2x+3^2x+4^2x^3+\ldots\\&=\frac{\mathrm{d}}{\mathrm{d}x}(x+2x^2+3x^3+4x^4+\ldots)\\&=\frac{\mathrm{d}}{\mathrm{d}x}(x(1+2x+3x^2+4x^3+\ldots))\\&=\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{x}{(1-x)^2}\right)\end{aligned} \tag{6.4.661}$$

Therefore:

$$\sum_{k=0}^{+\infty}(k+1)^2x^k=\frac{1+x}{(1-x)^3} \tag{6.4.662}$$

It is easy with a software like Maple™ to control that the Maclaurin series of $(1+x)/(1-x)^3$ is equal to the generating function and therefore gives us the coefficients a_i of the sequence.

Let us see now a non-trivial example... It's the Fibonacci sequence given for recall (see page 1661) by (each term is equal to the sum of the previous two):

$$1,1,2,3,5,8,13,21,34,55,\ldots \tag{6.4.663}$$

We will give the generating function for this sequence a name $F(x)$ so then:

$$F(x)=\sum_{k=0}^{+\infty}=1+x+2x^2+3x^3+5x^4+\ldots \tag{6.4.664}$$

where $F_0=F_1=1$ and $F_k=F_{k-1}+F_{k-2}$ for $k\geq 2$. Then we can see that:

$$\begin{aligned}F(x)&=1+x+\sum_{k=2}^{+\infty}F_kx^k=1+x+\sum_{k=2}^{+\infty}(F_{k-1}+F_{k-2})x^k\\&=1+x+\sum_{k=2}^{+\infty}F_{k-1}x^k+\sum_{k=2}^{+\infty}F_{k-2}x^k\\&=1+x+(x^2+2x^3+3x^4+\ldots)+(x^2+x^3+2x^4+3x^5+\ldots)\\&=1+xF(x)+x^2F(x)\end{aligned} \tag{6.4.665}$$

Since we have figured out that:

$$F(x)=1+xF(x)+x^2F(x) \tag{6.4.666}$$

then:

$$F(x)-xF(x)-x^2F(x)=1 \tag{6.4.667}$$

and this can be rewritten as:

$$F(x)(1-x-x^2)=1 \tag{6.4.668}$$

and hence:

$$F(x)=\frac{1}{1-x-x^2} \tag{6.4.669}$$

It is always surprising that the generating function for the Fibonacci numbers has such a compact formula. But once again, using for example a software like Maple™ and doing the Maclaurin expansion of the above function, you will get a series whose coefficients a_k corresponds to the Fibonacci sequence!

But let us continue to investigate...! We have therefore the denominator:

$$1-x-x^2 \tag{6.4.670}$$

that can be factorized assuming the equation:

$$1-x-x^2=0 \tag{6.4.671}$$

has solutions! We quickly found that the roots are given by:

$$x_{1,2}=-\frac{1\pm\sqrt{5}}{2} \tag{6.4.672}$$

We recognize here the negative values of the Golden ratio φ and its conjugate (see earlier above page 1389).

Therefore the denominator can be written as (using the relation between the Golden ratio and its conjugate):

$$\begin{aligned}1-x-x^2&=(x-(-\varphi_1))(x-(-\varphi_2))=(x+\varphi_1)(x+\varphi_2))=\varphi_1\varphi_2\left(\frac{x}{\varphi_1}+1\right)\left(\frac{x}{\varphi_2}+1\right)\\&=\varphi\left(-\frac{1}{\varphi}\right)\left(\frac{x}{\varphi_1}+1\right)\left(\frac{x}{\varphi_2}+1\right)=-\left(\frac{x}{\varphi_1}+1\right)\left(\frac{x}{\varphi_2}+1\right)\\&=-\left(-x\varphi_2+1\right)\left(-x\varphi_1+1\right)=-\left(1-x\varphi_1\right)\left(1-x\varphi_2\right)\end{aligned}\tag{6.4.673}$$

Thus we can write:

$$F(x)=\frac{1}{1-x-x^2}=-\frac{1}{\left(1-x\varphi_1\right)\left(1-x\varphi_2\right)}\tag{6.4.674}$$

Let us decompose this fraction (the $-$ sign is now the in the constants A and B):

$$F(x)=-\frac{1}{(1-x\varphi_1)(1-x\varphi_2)}=\frac{A}{(1-x\varphi_1)}+\frac{B}{(1-x\varphi_2)}\tag{6.4.675}$$

where A, B need to be found out. Clearly we have:

$$1=A(1-x\varphi_2)+B(1-x\varphi_1)\tag{6.4.676}$$

By identification, we have the following linear system:

$$\begin{aligned}A+B&=1\\-Ax\varphi_2-Bx\varphi_1&=-x(A\varphi_2+B\varphi_1)=0\end{aligned}\tag{6.4.677}$$

Hence:

$$A+B=1\tag{6.4.678}$$
$$A\varphi_2+B\varphi_1=0\tag{6.4.679}$$

Or written explicitly:

$$A+B=1\tag{6.4.680}$$
$$-\frac{A}{\varphi}+B\varphi=0\tag{6.4.681}$$

The solution to this linear system is straightforward and given by:

$$A=\frac{\varphi^2}{\varphi^2+1}=\frac{\varphi_1}{\varphi_1-\varphi_2}\qquad B=\frac{1}{\varphi^2+1}=\frac{-\varphi_2}{\varphi_1-\varphi_2}\tag{6.4.682}$$

Thus we have:

$$\begin{aligned}F(x)&=\frac{A}{(1-x\varphi_1)}+\frac{B}{(1-x\varphi_2)}=\frac{1}{\varphi_1-\varphi_2}\left(\frac{\varphi_1}{(1-x\varphi_1)}-\frac{\varphi_2}{(1-x\varphi_2)}\right)\\&=\frac{1}{\varphi_1-\varphi_2}\left(\varphi_1\frac{1}{(1-x\varphi_1)}-\varphi_2\frac{1}{(1-x\varphi_2)}\right)\end{aligned}\tag{6.4.683}$$

Next we use another quite famous result, the following Taylor development:

$$\frac{1}{1-x}=1+x+x^2+x^3+\ldots \tag{6.4.684}$$

for $|x|<1$. Using this result we can see that:

$$\sum_{k=0}^{+\infty} F_k x^k = F(x) = \frac{1}{\varphi_1-\varphi_2}\left(\sum_{k=0}^{+\infty}(\varphi_1^{k+1}-\varphi_2^{k+1})x^k\right) \tag{6.4.685}$$

Therefore:

$$F_k=\frac{\varphi_1^{k+1}-\varphi_2^{k+1}}{\varphi_1-\varphi_2}=\frac{\varphi_1^{k+1}-\varphi_2^{k+1}}{\sqrt{5}}=\frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^{k+1}-\left(\frac{1-\sqrt{5}}{2}\right)^{k+1}\right] \tag{6.4.686}$$

Hence:

$$\boxed{F_k=\frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^{k+1}-\left(\frac{1-\sqrt{5}}{2}\right)^{k+1}\right]} \tag{6.4.687}$$

This result is sometimes named the "Binet's Fibonacci formula".

6.4.4.1.1 Composition of Generating functions

Now if we have two generating functions:

$$G_1(x)=\sum_{k=0}^{+\infty} a_k x^k \qquad \text{and} \qquad G_2(x)=\sum_{k=0}^{+\infty} b_k x^k \tag{6.4.688}$$

for two sequences of integers $a_0,a_1,a_2,a_3,\ldots$ and $b_0,b_1,b_2,b_3,\ldots$ then there are several ways that we can combine these both sequences and get generating functions for the new corresponding sequence.

- Sum: If we add the generating functions we have that:

  $$A(x)+B(x)=\sum_{k=0}^{+\infty}(a_k+b_k)x^k \tag{6.4.689}$$

  is a generating function for the series:

  $$a_0+b_0,a_1+b_1,a_2+b_2,a_3+b_3,\ldots \tag{6.4.690}$$

- Product: However if we multiply the two generating function, we ha that:

$$\begin{aligned}G_1(x)G_2(x)&=(a_0+a_1x+a_2x^2+a_3x^3+\ldots)(b_0+b_1x+b_2x^2+b_3x^3+\ldots)\\&=a_0b_0+(a_1b_0+a_0b_1)x+(a_2b_0+a_1b_1+a_0b_2)x^2+\ldots\\&=\sum_{k=0}^{+\infty}(a_nb_0+a_{n-1}b_1+\ldots+a_0b_n)x^n\end{aligned} \tag{6.4.691}$$

This can be summarized in the expression:

$$G_1(x)G_2(x)=\sum_{k=0}^{+\infty}\left(\sum_{i=0}^{n}a_{k-i}b_i\right) \quad (6.4.692)$$

Another special case of the product of generating function is the product:

$$\frac{1}{1-x}G(x) \quad (6.4.693)$$

It is the product of two generating functions, the first one being the generating function:

$$\frac{1}{1-x}=1+x+x^2+x^3+\ldots \quad (6.4.694)$$

By:

$$G_1(x)G_2(x)=\sum_{k=0}^{+\infty}\left(\sum_{i=0}^{n}a_{k-i}b_i\right) \quad (6.4.695)$$

The product of these is a generating function for the series (as all $a_i=1$):

$$b_0,b_0+b_1,b_0+b_1+b_2,b_0+b_1+b_2+b_3,\ldots \quad (6.4.696)$$

- Derivative: We have already seen a couple examples of the use of the derivative in previous examples.

Remember now that we have proved that:

$$\frac{1}{(1-x)^2}=1+2x+3x^2+4x^3\ldots=\sum_{k=0}^{+\infty}(k+1)x^k \quad (6.4.697)$$

Therefore by the fact that:

$$\frac{1}{1-x}G(x) \quad (6.4.698)$$

has for generating sequence:

$$b_0,b_0+b_1,b_0+b_1+b_2,b_0+b_1+b_2+b_3,\ldots \quad (6.4.699)$$

we have then that:

$$\frac{1}{1-x}\frac{1}{(1-x)^2} \quad (6.4.700)$$

is then a generating function for the sequence of the sum of the first n positive integers:

$$1,1+2,1+2+3,1+2+3+4,\ldots \quad (6.4.701)$$

In particular, the coefficient of x^k is:

$$1+2+3+\ldots+k+(k+1) \quad (6.4.702)$$

We also know by taking the derivative of $1/(1-x)^2$ we have:

$$\frac{\mathrm{d}}{\mathrm{d}x}=\frac{2}{(1-x^2)}=2\cdot 1+3\cdot 2x+4\cdot 3x^2+5\cdot 4x^3+\ldots \tag{6.4.703}$$

Therefore if we divide this equation by two we have:

$$\frac{1}{(1-x)^3}=\sum_{k=0}^{+\infty}\frac{(k+1)(k+1)}{2}x^k \tag{6.4.704}$$

It must be known that the coefficient of x^k in $1/(1-x)^3$ is equal to $(k+1)(k+2)/2$ and it is equal to the sum of the first $k+1$ integers, so:

$$1+2+3+\ldots+n+(n+1)=\frac{(n+1)(n+2)}{2} \tag{6.4.705}$$

6.4.4.2 Multivariate Generating Functions

Remember that in the section of Calculus we have proved that:

$$\begin{aligned}
&(a+b)^0=1\\
&(a+b)^1=1a+1b\\
&(a+b)^2=1a^2+2ab+1b^2\\
&(a+b)^3=1a^3+3a^2b+3ab^2+1b^3\\
&\qquad\ldots
\end{aligned} \tag{6.4.706}$$

can be expressed by a condensed form that involves the binomial coefficient (see page 1364):

$$(a+b)^n=\sum_{p=0}^{n}C_p^n a^{n-p}b^p \tag{6.4.707}$$

then this is what we name a "multivariate generating function". It works just as the other generating functions we have previously worked with except that it has two parameters.

6.4.4.3 Functional Generating Functions

So far we have seen generating functions that gives scalar values. But a more general family are the "functional generating functions" that give functions instead of simple scalars.

A generating function for a sequences of functions $\{f_n(x)\}$ is obviously a power series of the type:

$$\boxed{G(x,t)=\sum_n f_n(x)t^n} \tag{6.4.708}$$

whose coefficient are now functions of x.

Let us see the both examples that will be useful to us in the sections of Economy, Wave Optics and also Differential and Integral Calculus!.

Let us start with the most important example involving probabilities and especially the Poisson distribution as used in the First Boston Credit Risk Metric model in the section Economy!

Let us recall that the Poisson distribution mass, that is, the probability that N is equal to n, is given by (see section Statistics page 686):

$$P(n)=\frac{\mu^n e^{-\mu}}{n!} \tag{6.4.709}$$

for $n=0,1,2,\ldots$.

The probability generating function is equal to:

$$P_N(t)=P(n=0)+P(n=1)t+P(n=2)t^2+\ldots=\sum_{n=0}^{+\infty}\frac{\mu^n e^{-\mu}}{n!}t^n \tag{6.4.710}$$

We know also the following Maclaurin infinite series expansion:

$$e^u=1+u+\frac{u^2}{2!}+\frac{u^3}{3!}+u\frac{u^4}{4!}+\ldots= \tag{6.4.711}$$

And if we rewrite:

$$P_N(t)=\sum_{n=0}^{+\infty}\frac{\mu^n e^{-\mu}}{n!}t^n \tag{6.4.712}$$

in the following way:

$$P_N(t)=e^{-\mu}\sum_{n=0}^{+\infty}\frac{(\mu t)^n}{n!} \tag{6.4.713}$$

And further assuming the above Maclaurin expansion we can write:

$$\sum_{n=0}^{+\infty}\frac{(\mu t)^n}{n!}=e^{\mu t} \tag{6.4.714}$$

Which implies:

$$\boxed{P_N(t)=e^{-\mu}\sum_{n=0}^{+\infty}\frac{(\mu t)^n}{n!}=e^{-mu}e^{\mu t}=e^{\mu(t-1)}} \tag{6.4.715}$$

This is the probability generating function of a Poisson distribution that we will use for our study of the CreditRisk model.

And now for our study of Wave Optics (especially the Fresnel diffraction for a circular aperture!) let us found the functional generating function of the first kind Bessel functions!

Theorem 6.26. The generating function for the sequence of Bessel function of first kind, on integer order, is:

$$G(x,t)=e^{(x/2)(t-1/t)}=\sum_{n=-\infty}^{+\infty}J_n(x)t^n \tag{6.4.716}$$

Proof 6.26.1. To get an expression for $J_n(x)$, we use the Maclaurin series for e^x to get:

$$\begin{aligned}G(x,t)=e^{xt/2}e^{-x/(2t)}&=\sum_{r=0}^{+\infty}\frac{1}{r!}\left(\frac{x}{2}\right)^r t^r\sum_{s=0}^{+\infty}\frac{(-1)^s}{s!}\left(\frac{x}{2t}\right)^s=\sum_{r=0}^{+\infty}\frac{1}{r!}\left(\frac{x}{2}\right)^r t^r\sum_{s=0}^{+\infty}\frac{(-1)^s}{s!}\left(\frac{x}{2}\right)^s t^{-s}\\&=\sum_{r=0}^{+\infty}\sum_{s=0}^{+\infty}\frac{(-1)^2}{r!s!}\left(\frac{x}{2}\right)^{r+s}t^{r-s}\end{aligned} \tag{6.4.717}$$

Now let us make the change of variable: $n=r-s$.

Therefore the expression in the sum becomes:

$$\frac{(-1)^s}{(n+s)!s!}\left(\frac{x}{2}\right)^{n+2s}t^n \tag{6.4.718}$$

Then it is obvious that as the range of r and s is $]-\infty,+\infty[$ we have for the range of n also: $]-\infty,+\infty[$. So the first sum is easy to determine:

$$\sum_{n=-\infty}^{+\infty}\frac{(-1)^s}{(n+s)!s!}\left(\frac{x}{2}\right)^{n+2s}t^n \tag{6.4.719}$$

But as we can see in the expression in the sum above, we don't get rid of the summation variable s. So there is obviously a second missing sum on the variable s! The interval of summation for s is not obvious at a first glance...

What is sure is that s is still in the range $[0,+\infty]$ (positive values) if we look at the term $s!$ in the denominator. But what makes problem is the lower bound of the sum. So if we look closely to the relation in the sum above we have a term $(n+s)!$. Obviously $n+s$ must never be negative! It comes therefore that if $n<0$, we must have $n+s\geq 0$, that is when $n<0$, s must start $-n$. Finally:

$$G(x,t)=\sum_{n=-\infty}^{+\infty}\sum_{s=\max(0,-n)}^{+\infty}\frac{(-1)^s}{(n+s)!s!}\left(\frac{x}{2}\right)^{n+2s}t^n=\sum_{n=-\infty}^{+\infty}\left[\sum_{s=\max(0,-n)}^{+\infty}\frac{(-1)^s}{(n+s)!s!}\left(\frac{x}{2}\right)^{n+2s}\right]t^n \tag{6.4.720}$$

That is:

$$G(x,t)=\sum_{n=-\infty}^{+\infty}J_n(x)t^n \tag{6.4.721}$$

where the J_n are the Bessel function of the first kind.

■ Q.E.D.

Therefore, for $n \geq 0$:

$$J_n(x)=\sum_{s=\max(0,-n)}^{+\infty} \frac{(-1)^s}{(n+s)!s!}\left(\frac{x}{2}\right)^{n+2s} \tag{6.4.722}$$

And for $n<0$:

$$J_{-n}(x)=\sum_{s=n}^{+\infty} \frac{(-1)^s}{((-n)+s)!s!}\left(\frac{x}{2}\right)^{(-n)+2s}=\sum_{s=n}^{+\infty} \frac{(-1)^s}{(s-n)!s!}\left(\frac{x}{2}\right)^{-n+2s} \tag{6.4.723}$$

Using an index shift, we get:

$$J_{-n}(x)=\sum_{s=0}^{+\infty} \frac{(-1)^{s+n}}{((s+n)-n)!(s+n)!}\left(\frac{x}{2}\right)^{-n+2(s+n)}=\sum_{s=0}^{+\infty} \frac{(-1)^{s+n}}{s!(s+n)!}\left(\frac{x}{2}\right)^{n+2s}=(-1)^n J_n(x) \tag{6.4.724}$$

There is also a very useful recursive relationship when implementing first kind Bessel functions in computer code that we can derive thanks to the generating function!

First, let us remind that we have just seen that from Generating Function for Bessel Function of the First Kind of Order n of x we have:

$$\exp\left(\frac{x}{2}\left(t-\frac{1}{t}\right)\right)=\sum_{n=-\infty}^{+\infty} J_n(x)t^n \tag{6.4.725}$$

Differentiating both sides of the equation with respect to t we get:

$$\frac{x}{2}\left(1+\frac{1}{t^2}\right)\exp\left(\frac{x}{2}\left(t-\frac{1}{t}\right)\right)=\sum_{n=-\infty}^{+\infty} nJ_n(x)t^{n-1} \tag{6.4.726}$$

That can be obviously rewritten as:

$$\frac{x}{2}\left(1+\frac{1}{t^2}\right)\sum_{n=-\infty}^{+\infty} J_n(x)t^n=\sum_{n=-\infty}^{+\infty} nJ_n(x)t^{n-1} \tag{6.4.727}$$

Distributing:

$$\frac{x}{2}\left(1\cdot\sum_{n=-\infty}^{+\infty} J_n(x)t^n+\frac{1}{t^2}\cdot\sum_{n=-\infty}^{+\infty} J_n(x)t^n\right)=\sum_{n=-\infty}^{+} +\infty nJ_n(x)t^{n-1} \tag{6.4.728}$$

and simplifying:

$$\frac{x}{2}\left(\sum_{n=-\infty}^{+\infty} J_n(x)t^n+\sum_{n=-\infty}^{+\infty} J_n(x)t^{n-2}\right)=\sum_{n=-\infty}^{+\infty} nJ_n(x)t^{n-1} \tag{6.4.729}$$

Translating the index variable of summation we can write:

$$\frac{x}{2}\left(\sum_{n=-\infty}^{+\infty} J_{n-1}(x)t^{n-1}+\sum_{n=-\infty}^{+\infty} J_{n+1}(x)t^{n-1}\right)=\sum_{n=-\infty}^{+\infty} nJ_n(x)t^{n-1} \tag{6.4.730}$$

By comparing coefficients we that we have for each n:

$$\frac{x}{2}\left(J_{n-1}(x)+J_{n+1}(x)\right)=nJ_n(x) \tag{6.4.731}$$

Rearranging:

$$J_{n-1}(x)+J_{n+1}(x)=\frac{2n}{x}J_n(x) \tag{6.4.732}$$

Finally we get the useful result we were looking for:

$$J_{n+1}(x)=\frac{2n}{x}J_n(x)-J_{n-1}(x) \tag{6.4.733}$$

Where for reminder:

$$\begin{aligned}J_0(x)&=\sum_{k=0}^{\infty}\frac{(-1)^k}{(k!)^2}\left(\frac{x}{2}\right)^{2k}\\&=1-\frac{x^2}{2^2}+\frac{x^4}{2^2\cdot 4^2}-\frac{x^6}{2^2\cdot 4^2\cdot 6^2}+\cdots\end{aligned} \tag{6.4.734}$$

~95 %

★★★☆☆

44 votes, 64.09%

6.5. Vector Calculus

VECTOR CALCULUS or "vector analysis" is a branch of mathematics that studies the scalar or vector fields that are sufficiently regular of Euclidean spaces (see definition further below).

The importance of vector calculus comes from its extensive use in physics and in the engineering sciences. It is from this perspective that we will present it, and this is why we limit ourselves mostly to the case of the usual three-dimensional spaces. In this context, a vector field associates to each point of the space a vector (with three real components), while a scalar field associates just a unique real number to such a point.

✍**Example:**

For example, imagine the water from a lake. The temperature data at each point forms a scalar field, that of his speed at each point, a vector field (see definition further below).

Physical concepts such as strength or speed are characterized by a direction, an orientation and an intensity. This triple character is highlighted by arrows. These are the source of the concept of vector and are the most suggestive example. Although their nature is essentially geometric, it is their ability to bind to each other, so their algebraic behaviour, which mostly retain our attention. Split into equivalence classes their set represents the classic model of a "vector space" (see section Set Theory page 686). One of our primary goals here is the detailed description of this model.

Remarks

R1. Before reading what follows, the reader is advised to have at least read diagonally the section on Set Theory in the Arithmetic chapter. We define there what is a "vector space" using the tools of Set Theory. Even if this concept is although not absolutely essential it is still interesting to see how two areas of mathematics fit together and also just for the reason... of introducing vector stuffs with a least a little bit rigour.

R2. Vectorial analysis contains many terms and definitions that must be learned by heart. This work is hard but unfortunately necessary...

6.5.1 Concept of Arrow

Definition (#6.5.225): We denote by U the ordinary space of elementary geometry and $P, Q, \ldots$ its points. We will name "arrow" all directed line segment (in space). The arrow of origin P (origin point) and extremity Q (terminal point) will be denoted $\overrightarrow{PQ}$ or abbreviated by a single letter (Latin or Greek) arbitrarily chosen as example: $\vec{F}$.

Remark

In the norm ISO 80000-2:2009 *Mathematical signs and symbols to be used in the natural sciences and technology* it is authorized to represent the vectors with a letter in bold.

We will consider as obvious that any arrow is characterized by its direction, its orientation (because of a given direction it can point in both directions), its intensity or magnitude (length) and its origin.

In vector (or multivariable) calculus, we will deal with functions of two or three variables (usually x, y or x, y, z, respectively). The graph of the arrow of coordinates (x, y), lies in Euclidean space, which in the Cartesian coordinate system consists of all ordered doublets of real numbers (a, b). Since the Euclidean space can be 3-dimensional (and more or less for sure!), we denote it by $\mathbb{R}^3$.

The graph of the arrow consists of the points (a, b, c) in $\mathbb{R}^3$. The 3-dimensional coordinate system of Euclidean space can be represented on a flat surface, such as this page or a blackboard, only by giving the illusion of three dimensions, in the manner shown in the figure below:

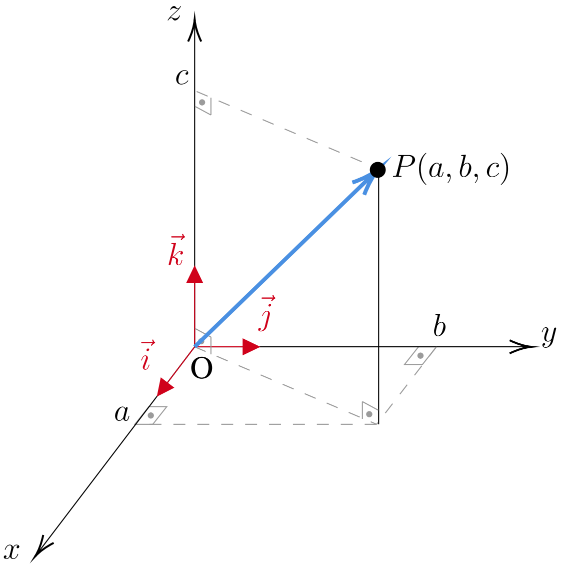
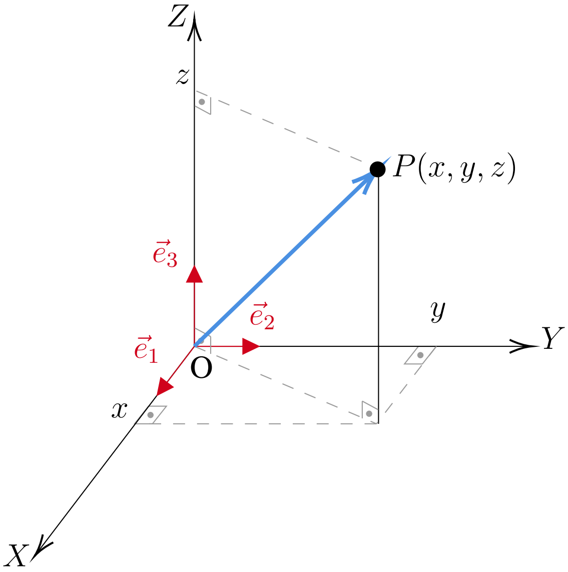


Figure 6.95: Example of arrow in $\mathbb{R}^3$ euclidean space

Or the same with another quite common notation:



Euclidean space has three mutually perpendicular coordinate axes (x, y and z), and three mutually perpendicular coordinate planes: the xy-plane, yz-plane and xz-plane:

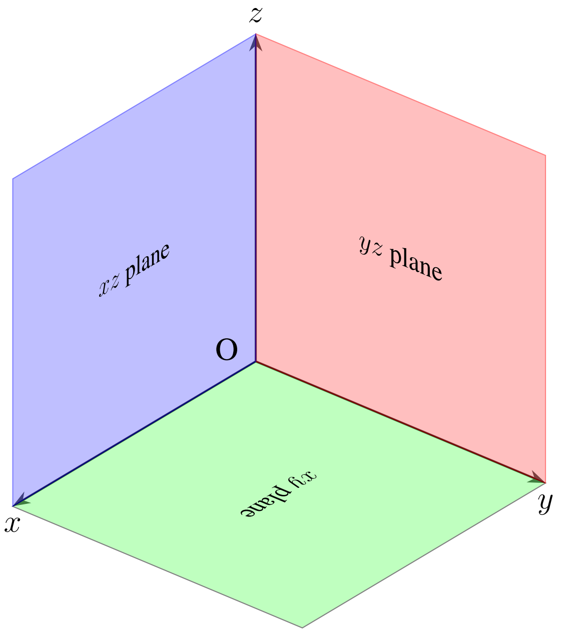


Figure 6.96: **Mutually perpendicular planes in $\mathbb{R}^3$**

The coordinate system shown above is also sometimes stupidly known as a "right-handed coordinate system" (...), because it is possible, using the right hand, to point the index finger in the positive direction of the x-axis, the middle finger in the positive direction of the y-axis, and the thumb in the positive direction of the z-axis, as below:

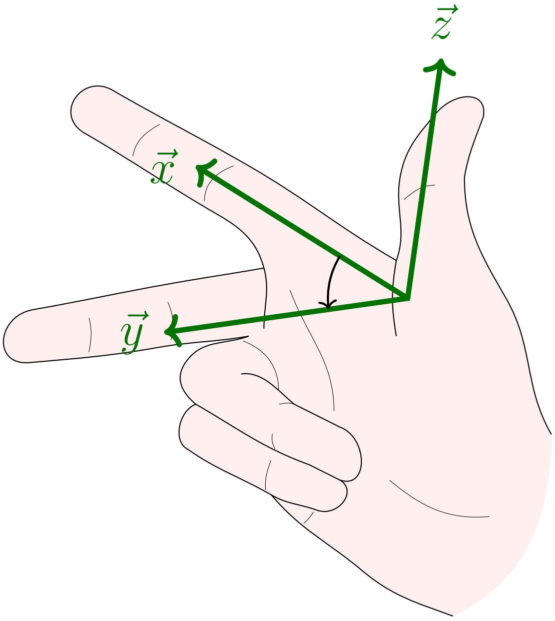


Figure 6.97: **Right-handed coordinate system (author: Izaak Neutelings)**

6.5.2 Set of Vectors

Definitions (#6.5.226):

D1. We say that two arrows are "equivalent arrows" if they have the same direction, the same orientation and the same intensity.

D2. We say that two arrows are "colinear arrows" only if they have the same direction.

Let us now split the set of all arrows in equivalence classes: two arrows belong to the same class if and only if they are equivalent.

So:

Definitions (#6.5.227):

D1. Each equivalence class of arrows whose origin point and terminal point are distinct is a "vector"" or rather a "free vector" because its origin is not taken into account (if its origin is well defined, then we have a "bounded vector").

D2. Degenerated arrows (that is to say of the form $\overrightarrow{PP}$) are named "zero vector" and written $\vec{0}$ when they have an undefined direction and orientation and zero intensity (origin and terminal point are not distinct).

The set of vectors will be commonly referred by $\mathbb{V}$. Note that the elements of $\mathbb{V}$ are (equivalence) classes arrows and not individual arrows. It is however clear that any arrow is sufficient to determine the class of equivalence to which it belongs and it is natural to name the corresponding class: "representative class" of the vector.

Let us now draw a representative of a vector $\vec{y}$ from the end of a representative vector $\vec{x}$. The arrow whose origin is that of the representative $\vec{x}$ and the end of representative $\vec{y}$ determines a new vector which we write: $\vec{x}+\vec{y}$. The operation that combines any two vectors by their sum is named "vector addition".

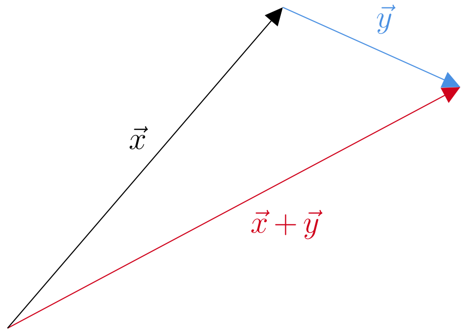


Figure 6.98: **Example of a sum of two vectors**

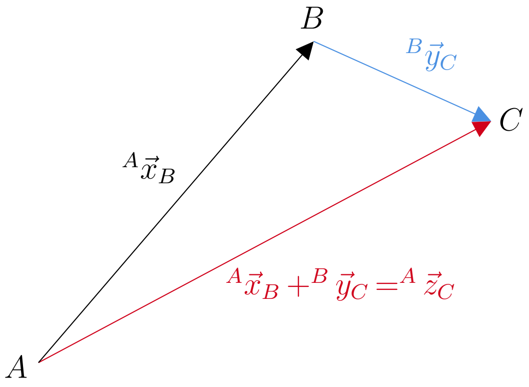


Figure 6.99: **Example of a sum of two vectors with robot dynamics notation**

Using a figure, it is easy to show that the operation of vector addition is associative and commutative, i.e. that:

$$(\vec{x}+\vec{y})+\vec{z}=\vec{x}+(\vec{y}+\vec{z}) \tag{6.5.1}$$

and:

$$\vec{x}+\vec{y}=\vec{y}+\vec{x} \tag{6.5.2}$$

It is also evident that the zero vector $\vec{0}$ is the neutral element of the vector addition. Formally:

$$\vec{x}+\vec{0}=\vec{x} \quad \text{and} \quad \vec{x}+(-\vec{x})=\vec{0} \tag{6.5.3}$$

where $-\vec{x}$ means the opposite of vector $\vec{x}$, that is to say the vector whose representatives have the same direction and the same intensity as those of $\vec{x}$, but the opposite orientation.

Definitions (#6.5.228):

D1. Two vectors whose sum is zero are then named "opposed vectors" since the only thing that differentiates them is their orientation...

D2. It follows that if two or more vectors have the same direction, the same intensity and the same orientation then we say that they are "equal vectors".

As we can see the reverse operation of vector addition is the vector subtraction. Subtract a vector is equivalent to adding the opposite vector.

Remarks

R1. The addition extends, by induction, to the case of any finite family of vectors. Under associativity, these successive additions can be performed in any order, which justifies the writing without brackets.

R2. The multiplication between two vectors is a concept that does not exist. But, as we shall see it a little further, we can multiply the vectors by some other vector properties that which we name the "norm" or simply by scalars and still by some other things...

6.5.2.1 Pseudo-Vectors

In physics, in the statement named the "Curie's principle" (see section Principia page 3144), physicists mention of what they name "pseudo-vectors". This is the simple vocabulary to talk about something equally trivial but basically only a few people actually do use. But it can still be useful to present what it is.

In fact, vectors and pseudo-vectors are transformed in the same way for a rotation or translation (we will see in our study of Linear Algebra how mathematically perform this type transformations). It is not the same in symmetry with respect to a plane or at one point. In these transformations we have by definition the following properties:

P1. A vector is transformed into its symmetrical.

P2. A pseudo-vector is converted into the opposite of its symmetrical.

Here is a figure with typical examples (the choice of letters representing the vectors and pseudo-vectors is not due to chance; they are a wink to the properties of electric and magnetic fields as studied in the Electromagnetism chapter):

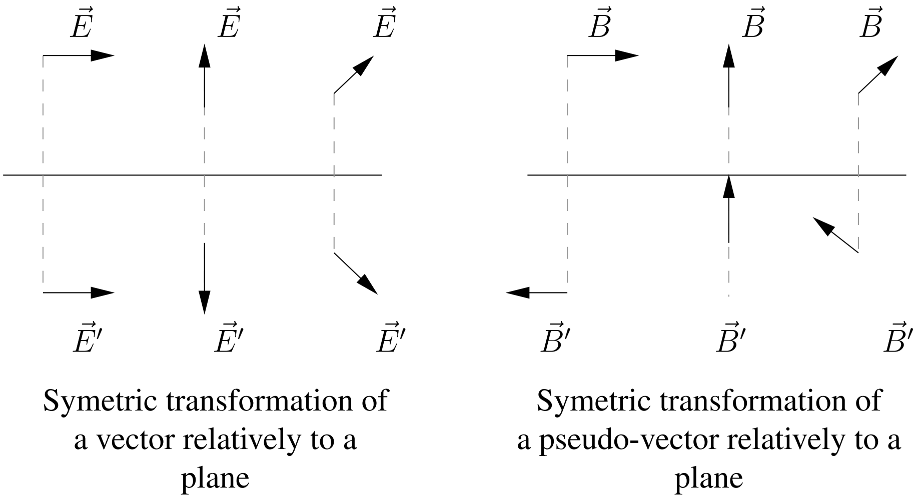

$\vec{E}$
$\vec{E}$
$\vec{E}$
$\vec{B}$
$\vec{B}$
$\vec{B}$
$\vec{E}'$
$\vec{E}'$
$\vec{E}'$
$\vec{B}'$
$\vec{B}'$
$\vec{B}'$
Symetric transformation of a vector relatively to a plane
Symetric transformation of a pseudo-vector relatively to a plane

Figure 6.100: Differences of transformations between a vector and a pseudo-vector

A well known practical example is a pseudo-vector that we will study in detail much further below and resulting of an operation named "cross product" (see page 1830):

$$\vec{x}\times\vec{y}=\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}\times\begin{pmatrix}y_1\\y_2\\y_3\end{pmatrix}=\begin{pmatrix}x_2y_3-x_3y_2\\x_3y_1-x_1y_3\\x_1y_2-x_2y_1\end{pmatrix}\tag{6.5.4}$$

And to see why the result is a pseudo-vector, consider the special simple case:

$$\vec{x}\times\vec{y}=\begin{pmatrix}1\\0\\0\end{pmatrix}\times\begin{pmatrix}0\\1\\0\end{pmatrix}=\begin{pmatrix}0\\0\\1\end{pmatrix}\tag{6.5.5}$$

Now if we do a symmetric operation of the XOZ plane we get:

$$\vec{x}\times\vec{y}=\begin{pmatrix}1\\0\\0\end{pmatrix}\times\begin{pmatrix}0\\-1\\0\end{pmatrix}=\begin{pmatrix}0\\0\\-1\end{pmatrix}\tag{6.5.6}$$

So as we can see the vector resulting of the cross product is a pseudo-vector because under transformation of the plane it's orientation change!

A vector resulting of a mathematical operation of symmetry that does not change its orientation is named a "polar vector" but in fact almost everybody say just "vector".

Now that we have an idea of what vectors are, we can start to perform some of the usual algebraic operations on them and this is what we name "vector algebra".

6.5.2.2 Normal vector

The "normal vector", denoted $\vec{N}$ or $\vec{n}$, to a surface is a vector which is perpendicular to the surface or curve at a given point. When normal vectors are considered on closed surfaces or closed curves, the inward-pointing normal (pointing towards the interior of the surface) and outward-pointing normal are usually distinguished.

The unit vector obtained by normalizing the normal vector (i.e., dividing a non-zero normal vector by its vector norm) is the unit normal vector, often known simply as the "unit normal"[13].

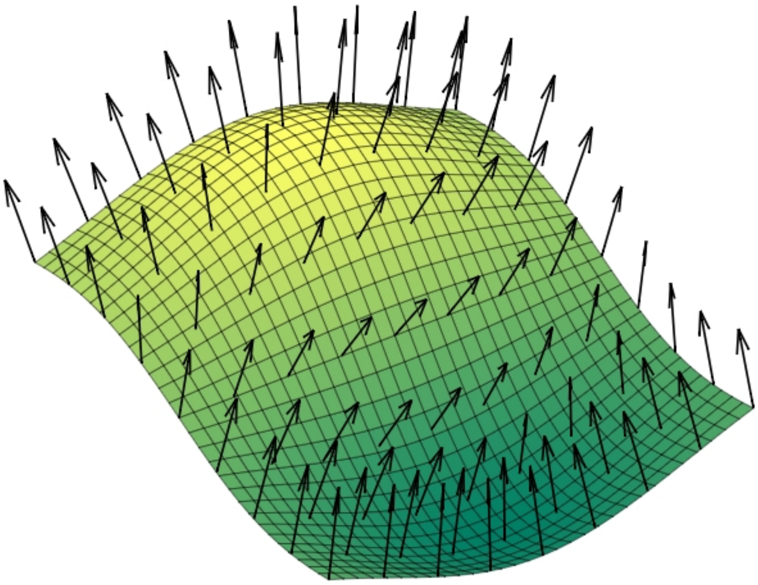


Figure 6.101: A vector field of normals to a surface

In the three-dimensional case a "surface normal", or simply a "normal", to a surface at a point P is a vector that is perpendicular to the tangent plane to that surface at P. The word "normal" is also used as an adjective: a line *normal* to a plane, the *normal* component of a force, the *normal* vector, etc.

We will see and prove in details in this section and many others, how to calculate explicitly various normal vectors to curves and surfaces.

For example that the normal vector to two vectors $\vec{a}$ and $\vec{b}$ is given by the cross product (see further below page 1830):

$$\vec{n}=\vec{a}\times\vec{b} \tag{6.5.7}$$

Or the normal to a parametric curve in Differential Geometry page 2865:

$$\vec{N}=R\frac{\mathrm{d}\vec{T}}{\mathrm{d}s} \tag{6.5.8}$$

---

[13] Care should be taken to not confuse the terms "vector norm" (length of vector), "normal vector" (perpendicular vector) and "normalized vector" (unit-length vector)

Or the normal vector of a plane of equation $ax+by+cz+d=0$ in the section of Analytical Geometry page 2802:

$$\vec{n}=\begin{pmatrix}a\\b\\c\end{pmatrix} \tag{6.5.9}$$

Or the gradient that is perpendicular to the isolines of a surface given by the cartesian function $f(x,y,z)=0$ of smoothness $\mathcal{C}^1$ as we will see further below page 1869:

$$\overrightarrow{\text{grad}}(f)=\begin{pmatrix}\dfrac{\partial f}{\partial x}\\\dfrac{\partial f}{\partial y}\\\dfrac{\partial f}{\partial z}\end{pmatrix} \tag{6.5.10}$$

and so on...

6.5.2.3 Multiplication by a scalar

The vector expression $\alpha\cdot\vec{x}$ named "product of vector $\vec{x}$ by scalar α" is defined as follows:

Definition (#6.5.229): Take a representative arrow $\vec{x}$ and construct a same direction arrow in the same or opposite orientation , depending on whether α (scalar) is positive or negative, and of intensity $|\alpha|$ times the intensity of the initial arrow. The arrow thus obtained is a representative of the vector of relation:

$$\alpha\cdot\vec{x} \tag{6.5.11}$$

If $\alpha=0$ or $\vec{x}=0$ we write:

$$\alpha\cdot\vec{x}=\vec{0} \tag{6.5.12}$$

The operation consisting of performing the product of a scalar by a vector is named "scalar multiplication".

We easily check that the scalar multiplication is associative and distributive with respect to the vector numerical addition, formally:

$$\begin{gathered}\alpha\cdot(\beta\cdot\vec{x})=(\alpha\cdot\beta)\cdot\vec{x}\\(\alpha+\beta)\cdot\vec{x}=\alpha\cdot\vec{x}+\beta\cdot\vec{x}\end{gathered} \tag{6.5.13}$$

The multiplication of a vector by non-null scalar doesn't change its direction if the scalar is positive but if the scalar is negative the vector will still have the same direction but it orientation will be opposite.

From this definition he have that two vectors $\vec{v}$ and $\vec{w}$ are parallel (denoted by $\vec{v}\,||\,\vec{w}$)) if one is a scalar multiple of the other. You can think of scalar multiplication of a vector as stretching or shrinking the vector, and as flipping the vector in the opposite direction if the scalar is a negative number.

Let's see a concrete example worldwide known of use of vectors with scalars (probably also the simplest example):

6.5.2.3.1 Rule of three

Let us go back to the "rule of three" (sometimes also named "rules of ratios and proportions" or "unit reduction method") often define in small classes (middle-school) intuitively but without nice proof. This rule is probably the most widely used algorithm in the world that identifies a fourth number when are given three and the four numbers are linearly dependent.

The rule of three is derived most of time in two versions:

- V1. Simple a direct if the magnitudes are directly proportional.
- V2. Simple and reverse if the quantities are inversely proportional.

and when two variables X and Y are proportional we note for recall:

$$X \propto Y \tag{6.5.14}$$

Theorem 6.27. Suppose now that X can take the values x_1, x_2. Y will take the values linearly dependent y_1, y_2 then the following proportional relation applies:

$$\boxed{\frac{x_1}{x_2}=\frac{y_1}{y_2}=c^{te}} \tag{6.5.15}$$

is named "simple and direct ratio".

Proof 6.27.1. Given two collinear vectors $\vec{x}=(x_1,x_2), \vec{y}=(y_1,y_2)$ and therefore proportional to a given factor λ such that:

$$\vec{x}=\lambda\vec{y} \Leftrightarrow \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}=\lambda\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \Leftrightarrow \lambda=\frac{x_1}{y_1}=\frac{x_2}{y_2} \Leftrightarrow \frac{y_1}{y_2}=\frac{x_1}{x_2} \tag{6.5.16}$$

This is often denoted in small classes:

$$\begin{gathered} x_1\,(a) \longrightarrow y_1\,(b) \\ x_2\,(c) \longrightarrow y_2\,(d) \end{gathered} \tag{6.5.17}$$

Therefore:

$$y_2=\frac{y_1\cdot x_2}{x_1} \Leftrightarrow d=\frac{b\cdot c}{a} \tag{6.5.18}$$

■ Q.E.D.

Remark

If this ratio is not equal (thus: not proportional), then we must switch to other tools such as simple and inverse ratio, or regression techniques and verbatim: extrapolation.

✍**Example:**

E1. In Lausanne (Switzerland), in 12011 (holocene calendar), garbage bags are taxed and the following rates apply: a bag of 17 [L] is 1.- and the bag of 110 [L] is 3.80.-. Reported to 17 [L] the price of the garbage bag of 110 [L] is thus:

$$\frac{3.80}{110}\cdot 17 \cong 0.58.\text{-} \tag{6.5.19}$$

That is to say approximately 60% of the price of the bag of 17 [L] (then go search for an explanation... ???).

E2. If the Earth has a radius of approximately $6,371$ kilometres and the Moon a radius of approximately $1,737.4$ kilometres and that their distance is actually in average of approximately $384,400$ kilometres, what would be the Earth-Moon distance and the size of the Moon is the Earth was reduce to the radius of a soccer ball of 11.25 centimetres.

The Moon radius would therefore be (everything has been converted into centimetres):

$$\frac{11.25}{6,371\cdot 10^6}\cdot 1,737.4\cdot 10^6 \cong 3[\text{cm}]$$

and the Earth-Moon distance will become:

$$\frac{11.25}{6,371\cdot 10^6}\cdot 384,400\cdot 10^6 \cong 678[\text{cm}]$$

Theorem 6.28. The following proportional relation:

$$\boxed{\frac{x_1}{x_2}=\frac{y_2}{y_1}=c^{te}} \tag{6.5.20}$$

is named "simple and inverse ratio".

Proof 6.28.1. Given two collinear vectors $\vec{x}=(x_1,x_2),\vec{y}=(y_1,y_2)$ and therefore proportional to a given factor λ such that:

$$\vec{x}=\lambda\vec{y}\Leftrightarrow\begin{pmatrix}x_1\\x_2\end{pmatrix}=\lambda\begin{pmatrix}\dfrac{1}{y_1}\\\dfrac{1}{y_2}\end{pmatrix}\Leftrightarrow\lambda=y_1x_1=y_2x_2\Leftrightarrow\frac{x_1}{x_2}=\frac{y_2}{y_1} \tag{6.5.21}$$

■ Q.E.D.

Remark

R1. If this ratio is not equal (thus: not inverse proportional), then we must switch to other tools such as simple and direct ratio, or regression techniques and verbatim: extrapolation.

R2. We also name "simple or inverse joint rule", a series of direct or inverse rule of three.

Basically, it is enough that we knew three of the four variables to solve this simple equation of the first degree.

Conversions of currencies or units of measurement are made using the direct or indirect simple rule of three. Parity calculations (provisional calculation made by an importer of a certain country having his own units of measurement and currency, who searches among several foreign offers - in systems of units of measurement and currencies which differ from the importer - which is the most advantageous or vice versa) are also done with the rule of three.

In such calculations, the agents of the exchange market have noticed that most of the time the ratio values were close to unity. They were thus naturally led to define the "percentage" as the proportion of a quantity or magnitude relative to another, measured with hundred (at least most of time...). Remember (see section Numbers page 234):

- Given a scalar $x\in\mathbb{R}$ then expressed in percentage it will denoted by:

$$x\%=x\cdot 100 \tag{6.5.22}$$

- Given a scalar $x\in\mathbb{R}$ then expressed in per-thousand it will denoted by:

$$x‰=x\cdot 1,000 \tag{6.5.23}$$

6.5.3 Vector Spaces

Definition (#6.5.230): We name "vector space" a set E of elements designated by $\vec{x},\vec{y},\ldots$ and named (as we know) "vectors", with a "vector algebraic structure" defined by the operation of vector addition (and thus vector subtraction) and scalar multiplication. These two operations satisfy the laws of associativity, commutativity, distributivity, neutral element and opposing element as we have already seen in the section of Set Theory.

For more information about what a vector space set is exactly the reader will have therefore to refer to the section of Set Theory at page 1805 where this concept is defined more strictly (it would be redundant to repeat it here an anyway it is not crucial because the properties are intuitive).

For every positive integer n, a_i means all the n-tuples of numbers arranged in a vector column:

$$\vec{v}=\begin{pmatrix}a_1\\a_2\\\vdots\\a_n\end{pmatrix} \tag{6.5.24}$$

or as line vector (vector column that has been **T**ransposed):

$$\vec{v}^T=(a_1\,a_2\,\cdots\,a_n) \tag{6.5.25}$$

and $\mathbb{R}^n$ provides clearly a vector space structure. The vectors of this space will be named as we already know: "vectors". They are often denoted more briefly by:

$$\vec{v}=(a_i)_{1\leq i\leq n} \tag{6.5.26}$$

or even more briefly by:

$$(a_i) \tag{6.5.27}$$

The number a_i is sometimes name "term" or "component of index i" of (a_i).

Now, unless stated otherwise, the vectors will always be the elements of a vector space E.

6.5.3.1 Linear Combinations

Definition (#6.5.231): We name "linear combination" of vectors $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n$ any vector relation of the form:

$$\vec{v} = \alpha_1\vec{x}_1 + \alpha_2\vec{x}_2 + \ldots + \alpha_n\vec{x}_n \tag{6.5.28}$$

where $\alpha_1, \alpha_2, \ldots, \alpha_k$ are numbers named "coefficients of the linear combination".

When a vector can be expressed in the above way we say that the vector is in "component form".

The null vector $\vec{0}$ is a linear combination of the $\alpha_1\vec{x}_1 + \alpha_2\vec{x}_2 + \ldots + \alpha_n\vec{x}_n$ with all coefficients equal to zero. We speak therefore of "trivial linear combination".

Definition (#6.5.232): We name "convex combination", any linear combination whose coefficients are non-negative and sum equal to 1. The set of convex combinations of two points P and Q of a punctual space P_0 (with an origin) is the line segment P and Q, traditionally denoted $\overline{PQ}$. To realize this, we just write:

$$\alpha P + (1 - \alpha Q) = Q + \alpha(P - Q) \tag{6.5.29}$$

and we make α vary from 0 to 1 and to find that all the points of the segment are thereby obtained.

If the vector $\vec{v}$ is a linear combination of $\alpha_1\vec{x}_1 + \alpha_2\vec{x}_2 + \ldots + \alpha_n\vec{x}_n$ and each of these vectors $\vec{x}_i$ is a linear combination of a set of independent vectors $\vec{y}_1, \vec{y}_2, \ldots, \vec{y}_k$, then it could be obvious that $\vec{v}$ is also a linear combination of $\vec{y}_1, \vec{y}_2, \ldots, \vec{y}_k$.

Definition (#6.5.233): A number n of non-zero vector are "coplanar" if one of them is a linear combination of the others. For example, three vectors are coplanar if one of them is in the plane defined by the two others.

6.5.3.2 Sub-vector spaces

Definition (#6.5.234): We name "vectorial subspace V of E" any subset of E that is itself a vector space for the operations of addition and scalar multiplication defined in E.

A vectorial sub-space V, as a vectorial space, can not be empty as it includes at least one vector, i.e. its zero vector, this being also necessarily also the zero vector of E. In addition, together with the vectors $\vec{x}$ and $\vec{y}$ (if it contains other vectors than the zero vector), it also includes all their linear combinations $\alpha\vec{x} + \beta\vec{y}$.

Conversely, as soon as we see any subset having these properties is a vectorial subspace. We have thus established the following proposition:

A subset V of E is a subspace of E if and only if V is not empty and $\alpha\vec{x}+\beta\vec{y}$ belongs to V for every pair $(\vec{x},\vec{y})$ of V and all any pair $(\alpha,\beta)\in\mathbb{R}$.

6.5.3.3 Generating families

It follows that if we have a family of vectors $(\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_k)$ the set of linear combinations of $\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_k$ with $k<n$ can be a subspace V of E, more specifically the smallest subspace of E consisting of $\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_k$.

The $\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_k$ vectors that satisfy the above condition are named "generators" of V and the family $(\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_k)$ the "generating family" of V. We also say that these vectors or family "generate S".

Remark

The subspace generated by a non-zero vector consists of all multiples of this vector. We name such a subspace a "vector line". A subspace generated by two vector non multiple of each other is named a "vector map" or "vector plane".

6.5.3.4 Linear Dependence or Independence

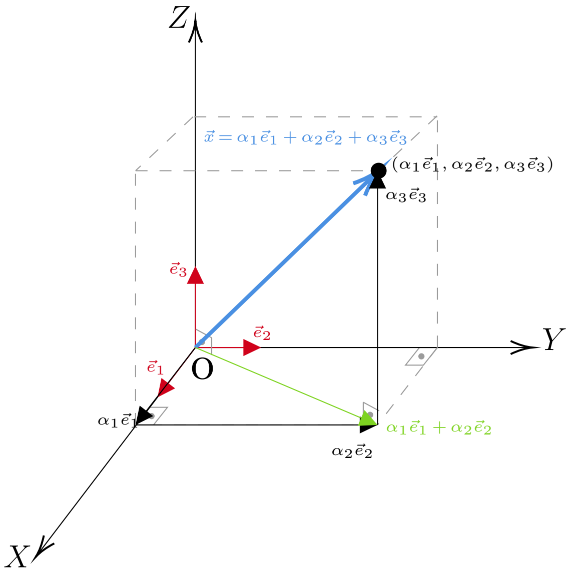
What follows is very important in physics: we advise future physicists or engineer really take the time to read the developments below.

If $(\vec{e}_1,\vec{e}_2,\vec{e}_3)$ are three vectors of E^3 whose representatives are not parallel to the same plane (by convention a zero-vector is parallel to any plane), so any vector $\vec{x}$ of E^3 can be written by the linear combination:

$$\vec{x}=\alpha_1\vec{e}_1+\alpha_2\vec{e}_2+\alpha_3\vec{e}_3 \tag{6.5.30}$$

where $\alpha_1,\alpha_2,\alpha_3$ belongs typically to $\mathbb{R}$.

For example the above vector $\vec{x}$ (but can also be get for different values of α_i!):



Z
$\vec{x}=\alpha_1\vec{e}_1+\alpha_2\vec{e}_2+\alpha_3\vec{e}_3$
$(\alpha_1\vec{e}_1,\alpha_2\vec{e}_2,\alpha_3\vec{e}_3)$
$\alpha_3\vec{e}_3$
$\vec{e}_3$
$\vec{e}_2$
Y
$\vec{e}_1$
O
$\alpha_1\vec{e}_1$
$\alpha_1\vec{e}_1+\alpha_2\vec{e}_2$
$\alpha_2\vec{e}_2$
X

Figure 6.102: **Example of a construction of a vector in a three-dimensional space**

In particular, the only possibility to get the zero vector $\vec{0}$ as a linear combination of $(\vec{e}_1,\vec{e}_2,\vec{e}_3)$ is to assign the trivial value 0 to the coefficients $\alpha_1,\alpha_2,\alpha_3$.

Conversely, if for three vectors $\vec{e}_1,\vec{e}_2,\vec{e}_3$ of E^3 the relation:

$$\vec{x}=\alpha_1\vec{e}_1+\alpha_2\vec{e}_2+\alpha_3\vec{e}_3=\vec{0} \tag{6.5.31}$$

implies $\alpha_1=\alpha_2=\alpha_3=0$, none of the vectors may be linear combination of the other two, in other words, their representatives are not parallel to the same plane.

Based on these observations, we will extend the notion of absence of parallelism to a same plane in the case of any number of vectors of a given vector space E.

We say that the vectors $\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_k$ are "linearly independent" if the relation:

$$\alpha_1\vec{e}_1+\alpha_2\vec{e}_2+\ldots+\alpha_k\vec{e}_k=0 \tag{6.5.32}$$

necessarily implies $\alpha_1=\alpha_2=\ldots=\alpha_k=0$, in other words, if the trivial linear combination is the only linear combination of $\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_k$ which is zero. Otherwise, we say that the vectors $\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_k$ are "linearly dependent".

If the intention is fixed on the family $(\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_k)$ rather than the terms of which it is made, we say that the latter is a "free family" or "linked family" following that the vectors are linearly independent or dependent.

6.5.3.5 Base of a vectorial space

Definition (#6.5.235): We say that a family of finite vectors is a basis of E if and only if:

1. If it is free.
2. It generates E.

Following this definition, every free family $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_k$ is a basis of the subspace it generates.

✍**Example:**

If we consider $\mathbb{C}$ as a $\mathbb{R}$-vector space (see section Set Theory page 1805), then since all the elements of $\mathbb{C}$ are written $a + ib$, the elements that generate $\mathbb{C}$ are 1 and i (both are free).

A base of $\mathbb{C}$ (which is 2-dimensional) as a $\mathbb{R}$-vector space is therefore the free finite set $\{1, \mathrm{i}\}$.

For a family of vectors $(\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_n)$ to be a basis of E, then it is necessary and sufficient that every vector $\vec{x}$ of E is expressed uniquely as a linear combination of the vectors $(\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_n)$:

$$\vec{x} = x_1\vec{e}_1 + x_2\vec{e}_2 + \cdots + x_n\vec{e}_n \tag{6.5.33}$$

The above relation is decomposition of $\vec{x}$ following the base $(\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_n)$ where the coefficients $x_1, x_2, \ldots, x_n$ are the components of $\vec{x}$ in this base. In the presence of a base, each vector is determined entirely by its components.

Proposition:

If $x_1, x_2, \ldots, x_n$ are the components of $\vec{x}$ and $y_1, y_2, \ldots, y_n$ those of $\vec{y}$ then:

$$x_1 + y_1, x_2 + y_2, \cdots, x_n + y_n \tag{6.5.34}$$

are the components of $\vec{x} + \vec{y}$.

In other words, add two vectors is equivalent to add their components and multiply a vector by a scalar obviously equivalent to multiplying its components by the same scalar. The basis is an important tool because it allows you to perform operations on vectors through operations on numbers.

✍**Example:**

The following column vectors of $\mathbb{R}^n$:

$$\begin{pmatrix}1\\0\\\cdot\\\cdot\\\cdot\\0\end{pmatrix},\begin{pmatrix}0\\1\\\cdot\\\cdot\\\cdot\\0\end{pmatrix},\ldots,\begin{pmatrix}0\\0\\\cdot\\\cdot\\\cdot\\1\end{pmatrix} \tag{6.5.35}$$

generate a base that we name "canonical basis" of $\mathbb{R}^n$ (we will work in complex spaces in another section of this textbook).

Remark

As part of the three-dimensional space, bases are very often treated as a triad (actually if you connect the ends of the three vectors by features you will get an imaginary triad).

6.5.3.6 Direction Angles

It is clear that only one standard angle cannot describe the direction of a vector in space (in an n-dimensional space, at least $n-1$ angles are required to describe the direction of a vector). We then use the concept of "direction angles". This is to measure the angle of the vector $\vec{U}$ with respect to each of the positive axis of the base:

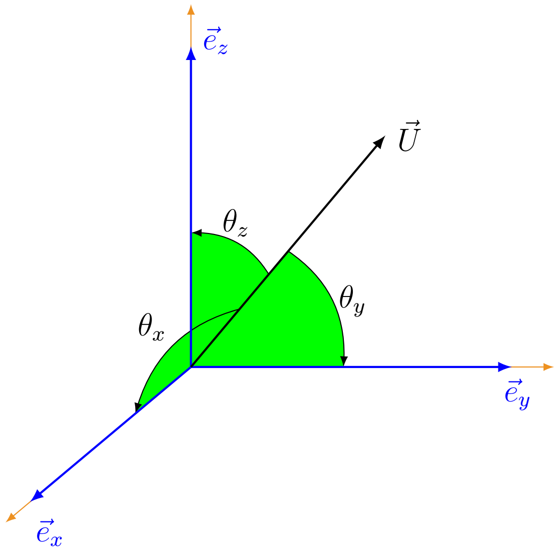


Figure 6.103: Representation of direction angles

If:

$$\vec{U}=(u_1,u_2,u_3)=u_1\vec{i}+u_2\vec{j}+u_3\vec{k} \tag{6.5.36}$$

Then by definition:

$$\begin{aligned}u_1=x&=\|\vec{U}\|\cos(\theta_x)\\u_2=y&=\|\vec{U}\|\cos(\theta_y)\\u_3=z&=\|\vec{U}\|\cos(\theta_z)\end{aligned} \tag{6.5.37}$$

The values:

$$\begin{aligned}\cos(\theta_x)&=\frac{x}{\sqrt{x^2+y^2+z^2}}\\\cos(\theta_y)&=\frac{y}{\sqrt{x^2+y^2+z^2}}\\\cos(\theta_z)&=\frac{z}{\sqrt{x^2+y^2+z^2}}\end{aligned} \tag{6.5.38}$$

are named the "cosines directions" of $\vec{x}$.

The three angles mentioned are not completely independent. Indeed, two are enough to completely determine the direction of a vector in space, the third can be deduced from the following equality (obtained from the calculation of the sum of squares of previous relations):

$$\cos^2(\theta_x)+\cos^2(\theta_y)+\cos^2(\theta_z)=1 \tag{6.5.39}$$

And it is this natural constraint that means, in reality, that in an n-dimensional space we have $n-1$ angles (degrees of freedom) sufficient to describe the direction of the vector.

Therefore the direction cosines are the scalar components of a unit standard vector $\vec{u}$ having the same direction as $\vec{U}$:

$$\frac{\vec{U}}{\|\vec{U}\|}=\vec{u}=\big(\cos(\theta_x),\cos(\theta_y),\cos(\theta_z)\big) \tag{6.5.40}$$

6.5.3.7 Dimensions of a vector space

We say that a basis E is of "finite size" if it is generated by a finite family of vectors. Otherwise, we say that E is of "infinite dimension" (we'll discuss this type of spaces in another section). Any finite dimensional vector space and not reduced to the zero vector has a basis. In fact, from any generating family of such a vector space we can extract a basis.

The dimension of a vector space is denoted by:

$$\dim(E) \tag{6.5.41}$$

Any vector space E of non-zero finite dimension n can be mapped in one to one correspondence (that is to say in bijection) with $\mathbb{R}^n$. We just need to choose a basis of E and to match to any vector $\vec{x}$ of E the column vector whose terms are the components of $\vec{x}$ in the chosen basis (this is mathematician blah blah but it will be useful when we will discuss more complex spaces):

$$\begin{array}{cc} E & \mathbb{R}^n \\ \vec{x}(x_1,x_2,\ldots,x_n)\Leftrightarrow & \begin{pmatrix}0\\0\\ \cdot\\ \cdot\\ \cdot\\ 1\end{pmatrix}\end{array} \tag{6.5.42}$$

This correspondence preserves the operations of addition and multiplication by a scalar as we have already seen; in other words, it can perform operations on vectors by operations on numbers.

Remark

Then we say that E and $\mathbb{R}^n$ are "isomorphic" or that the correspondence is an isomorphism (see section Set Theory page 443).

6.5.3.8 Extension of a free family

Theorem 6.29. Given $(\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_k)$ a free family and $(\vec{v}_1,\vec{v}_2,\ldots,\vec{v}_m)$ a generating family of E. If $(\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_k)$ is not a basis of E, we can extract a subfamily

$$(\vec{v}_{i1},\vec{v}_{i2},\ldots,\vec{v}_{il})$$

of $(\vec{v}_1,\vec{v}_2,\ldots,\vec{v}_m)$ so that the family $(\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_k,\vec{v}_{i1},\vec{v}_{i2},\ldots,\vec{v}_{il})$ is a basis of E.

Remark

Such a theorem is useful when going from a mathematical space passage having given properties to another space with different mathematical properties.

Proof 6.29.1. We assume that at least one of the vectors $\vec{v}_i$ is not a linear combination of vectors $(\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_k)$, otherwise $(\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_k)$ would generate E and would therefore be a possible basis of E. Let us note that vector $\vec{v}_{i1}$. The family $(\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_k,\vec{v}_{i1})$ is then a free family. Indeed, the relation:

$$\alpha_1\vec{x}_1+\cdots+\alpha_k\vec{x}_k+\beta_1\vec{v}_{i1}=\vec{0} \tag{6.5.43}$$

then implies first that $\beta_1=0$, otherwise $\vec{v}_{i1}$ would be a linear combination of the vectors $\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_k$, and then all $\alpha_i=0$ since the vectors $\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_k$ are linearly independent.

If the family $(\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_k,\vec{v}_{i1})$ generates E, it is then a possible base for E and the theorem is proved. Otherwise, the same reasoning ensures the existence of another vector $\vec{v}_{i2}$ .... If the new resulting family is not a basis of E, then the extraction process vectors $\vec{v}_i$ of $(\vec{v}_1,\vec{v}_2,\ldots,\vec{v}_m)$ continues. When it stops, we will get an "extension" of $(\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_k)$ in a free family generating E, that is to say a base of E.

■ Q.E.D.

It returns a corollary:

Every finite dimensional vector space and not reduced to zero vector has a basis! In fact, from any generating family of such a space, we can extract a base.

6.5.3.9 Rank of a finite family

Definition (#6.5.236): We name "rank of a family of vectors" and denote by rk(S) the dimension of the subspace S of E it generates.

Theorem 6.30. The rank of a family of vector $(\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_k)$ is less than or equal to k and is equal to k if and only if the family is free.

Proof 6.30.1. Let us set aside the trivial first case where the rank of the family $(\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_k)$ is zero. By the previous corollary, then we can extract from this family a base of the subspace it generates. The rank is than less or equal to k following that $(\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_k)$ is a linked family or not.

■ Q.E.D.

6.5.3.10 Direct Sums

Definition (#6.5.237): We say that the sum $S+T$ of two subspaces S and T of E is a "direct sum" if (special case applied to a 2 dimensional space!):

$$S \cap T = \left\{\vec{0}\right\} \tag{6.5.44}$$

In this case, we note it:

$$S \oplus T \tag{6.5.45}$$

In other words, the sum of two vector subspaces S and T of E is direct if the decomposition of all element $S+T$ into a sum of an element of S and of T is unique.

✍**Example:**

For example, the XY-plane, a two-dimensional vector space, can be thought of as the direct sum of two one-dimensional vector spaces, namely the X and Y axes. In this direct sum, the x and y axes intersect only at the origin (the zero vector). Addition is defined coordinate-wise, that is:

$$(x_1, y_1) + (x_2, y_2) = (x_1 + x_2, y_1 + y_2) \tag{6.5.46}$$

which is the same as vector addition.

This concept of trivial decomposition will be very useful in some theorems, the most important in this textbook is definitely the spectral theorem (see section of Linear Algebra page 2004) that has important implications in statistics!!!

From the direct sum we can introduce the concept of "complementary subspace" also named "subspace" (depending on countries ...):

Theorem 6.31. Suppose that E is of finite dimensions. For any subspace S of E, there exists a subspace T (not unique) of E such that E is the direct sum of S and T. We say then that T is a "supplementary subspace" of S into E.

Proof 6.31.1. First let us set aside the trivial case where $S = \left\{\vec{0}\right\}$ and $S = E$. The subspace S admits a basis $(\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_k)$, where k is less than the dimension n of E. By the theorem of extension of a free family, this basis can be extended in a basis $(\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_k, \vec{e}_{k+1}, \ldots, \vec{e}_n)$ of E. Let T be the subspace vector generated by the family $(\vec{e}_{k+1}, \ldots, \vec{e}_n)$. If $\vec{x}$ is any vector of E, then $\vec{x} = \vec{s} + \vec{t}$, where $\vec{s}$ is a vector of S and $\vec{t}$ a vector of T. In addition $S \cap T = \left\{\vec{0}\right\}$, because no vector, excepted the zero vector may be a linear combination of the vectors $\vec{e}_1, \ldots, \vec{e}_k$ and of the vectors $\vec{e}_{k+1}, \ldots, \vec{e}_n$. We therefore conclude that:

$$E = S \oplus T \tag{6.5.47}$$

■ Q.E.D.

6.5.3.11 Affine spaces

In mathematics, an affine space $G = \mathrm{A}\mathbb{R}^n$ is a geometric structure that is independent of the concepts of distance and measure of angles, keeping only the properties related to parallelism and ratio of lengths for parallel line segments as their is not origin point =$(0,0)$.

The space G of elementary geometry is both common and the source of the concept of "affine space" that we will introduce because when middle-school student begin to learn geometry they learn it (the huge majority of time) without any reference point =$(0,0)$.

In an affine space, there is therefore no distinguished point that serves as an origin. Hence, no vector has a fixed origin and no vector can be uniquely associated to a point. In an affine space, there are instead "displacement vectors", also named "translation vectors" or simply translations, between two points of the space.

This space G is associated with the "geometric vector space" V by the correspondence between vectors and arrows studied so far! The following definition is only to highlight the main common points of this correspondence:

Definition (#6.5.238): Let U be a non-empty set of elements that we name "points" and let us denote them by the letters $P,Q,\ldots$; given also E a vector space. Suppose that to any two points (P,Q) corresponds a vector denoted $\overrightarrow{PQ}$ (typically the point P is chosen as fictive origin). We say then that U is an "affine space" of directed space E if the following conditions are met:

C1. For any fixed point P, the correspondence between pairs (P,Q) and vectors $\vec{x}$ defined by only one point plus the origin point is bijective, i.e., for every vector $\vec{x}$ it exists a point Q such that we can define a vector $\overrightarrow{PQ}$.

C2. For each triple of points (P,Q,R):

$$\overrightarrow{PQ}+\overrightarrow{QR}=\overrightarrow{PR} \tag{6.5.48}$$

This is the famous "Chasles relation" (which we will see later a pseudo-equivalent in the section of Differential and Integral Calculus).

C3. If P is a point and $\vec{x}$ a vector, to express that Q is the unique point such as $\vec{x}=\overrightarrow{PQ}$, we write:

$$Q = P+\vec{x} \tag{6.5.49}$$

Although being a bit excessive, this writing is consistent with the usage and suggests well the idea of the operation it designates.

✍**Example:**

Below an "artistic" example of an affine space $U = \mathbb{AR}^2$ where there is no origin and any extremity of a line can be considered as the origin of a vector:

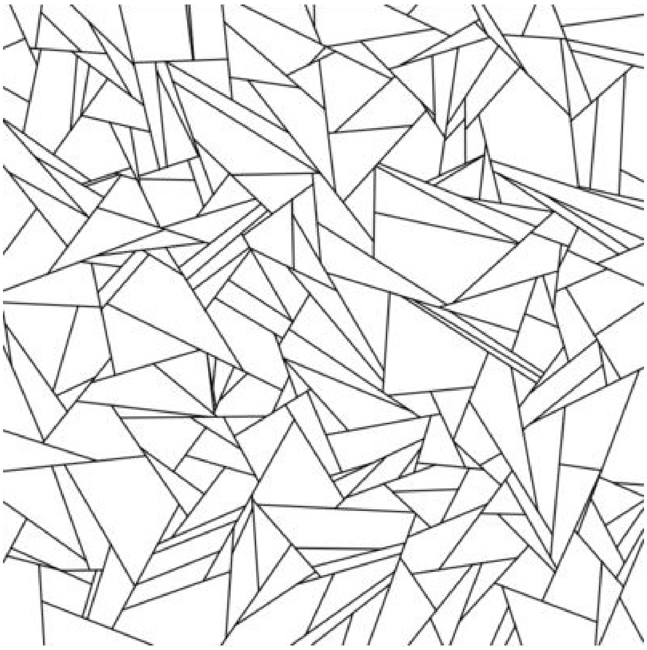


Figure 6.104: **Artistic but real example of an $G = \mathbb{AR}^2$ affine space**

The following properties follow directly from the definition of affine space:

P1. For any point P, $P + (\vec{x} + \vec{y}) = (P + \vec{x}) + \vec{y}$

P2. For any point P, $\overrightarrow{PP} = \vec{0}$. This results from the $\overrightarrow{PQ} + \overrightarrow{QR} = \overrightarrow{PR}$ provided in the case where we have $P = Q = R$.

P3. $\overrightarrow{PQ} = -\overrightarrow{QP}$. Just put $R = P$ in the De Chasles relation $\overrightarrow{PQ} + \overrightarrow{QR} = \overrightarrow{PR}$.

P4. Parallelogram rule:

Given the polygon with the vertices (clockwise) P, P', Q, Q' and edges $\overrightarrow{PP'}, \overrightarrow{P'Q'}, \overrightarrow{QQ'}, \overrightarrow{PQ}$:

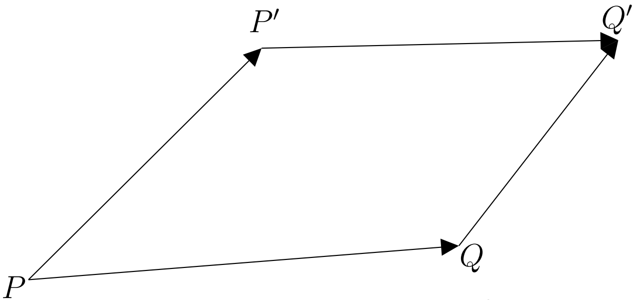


Figure 6.105: **Vectorial polygon in $\mathbb{AR}^2$**

We have:

$$\overrightarrow{PP'} = \overline{QQ'} \tag{6.5.50}$$

if and only if:

$$\overrightarrow{P'Q'} = \overline{PQ} \tag{6.5.51}$$

which would then give a parallelogram!

Indeed, replacing R with Q' in the Chasles relation we have:

$$\overrightarrow{PQ} + \overrightarrow{QQ'} = \overrightarrow{PQ'} \tag{6.5.52}$$

and by doing the same but replacing R by Q' and Q by P' we get:

$$\overrightarrow{PP'} + \overrightarrow{P'Q'} = \overrightarrow{PQ'} \tag{6.5.53}$$

We then have by equating the last two relations:

$$\overrightarrow{PQ} + \overrightarrow{QQ'} = \overrightarrow{PP'} + \overrightarrow{P'Q'} \tag{6.5.54}$$

Which forces the aforementioned equality that we wanted to prove.

Earlier we saw what made that a space G could be provided with a vector space structure (we saw then that it was therefore "vectorialized"). In the general case of an affine space G, the process is the same:

We choose any point O of G. The correspondence between pairs (O, P) and vectors of director space E being therefore biunivocal we define then the addition of points and multiplication of a point by a scalar by the corresponding operations on the vectors of E. Armed with these two operations, G becomes a vector space, named "vectorialized space G regarding to O". We denote this space by V and we name the point O "origin".

Given how operations have been defined, it follows that V is isomorphic to the space vector E:

$$P \overset{V}{=} \text{O} + \vec{x} \leftrightarrow \overset{E}{\vec{x}} \tag{6.5.55}$$

However, this isomorphism depends on the choice of the origin O and in practice this origin is selected on the basis of the data inherent to the studied problem. For example, if an affine transformation allows an invariant point (which does not move), it is advantageous to select that point as the origin.

Remarks

R1. When we talk about dimension of an affine space, we talk about the size of its director space.

R2. The space G of elementary geometry is an affine space of type $\text{A}\mathbb{R}^2$. Indeed, its direction is the geometrical space G and the conditions of definition of affine space are met.

R3. An affine space is a set of elements with a difference function. This difference is a binary function, which takes two points P and Q (both in G) and yields an element (a vector) $\vec{v}$ of a vector space E. We write $\vec{v}=P-Q$. Additionally, this difference function must ensure that, for any point P in E, it holds $P-P=\vec{0}$, where $\vec{0}$ is the null vector of E.

For the people who want to learn more about affine spaces (and related subjects), we recommend the lecture of the following excellent french textbook [8].

6.5.4 Euclidean Vector Spaces

Before defining what is an Euclidean vector space, let us first define some mathematical tools and some concepts.

We can, by choosing a unit length, measure the intensity of each arrow, in other words, determine its length. We can also measure the angular distance of two arrows (or vectors) of any common origin (not necessarily distinct) taking as the unit of angle measurement for example the radian (see section Trigonometry page 2503). The measurement of this difference is then a number between 0 and π named "angle" of the two arrows (see the section of Euclidean Geometry for more details). If both arrows have same direction and orientation, their angle is zero and if they have same direction and the opposite orientation, this same angle is π.

The representing arrows of a same vector $\vec{x}$ all have the same length. We denote this length (distance), named also "norm" or "module", by the notation[14]:

$$\boxed{\|\vec{x}\|} \tag{6.5.56}$$

given in the three-dimensional case explicitly by the following Euclidean distance if the origin of the vector coincides with the origin O of the vector basis:

$$\|\vec{x}\|=\left\|\begin{pmatrix}x\\y\\z\end{pmatrix}\right\|=\sqrt{x^2+y^2+z^2} \tag{6.5.57}$$

---

[14] This is actually a simplified notation because, as we will see much later in this textbook, there are several types of standards!

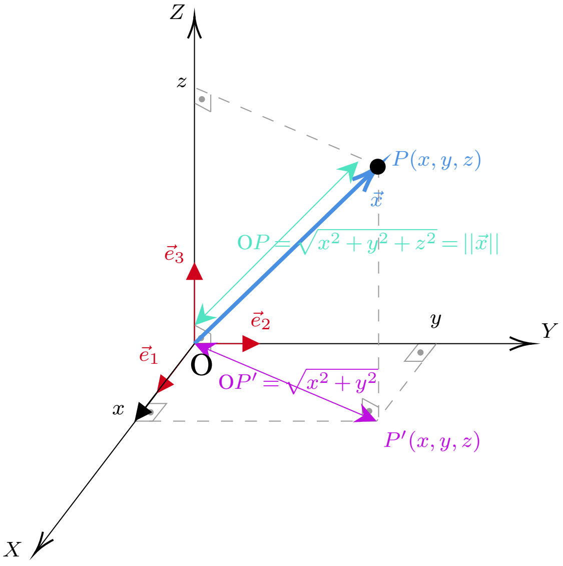


Figure 6.106: **Details of the calculation of the norm in an orthogonal coordinate system $\mathbb{R}^3$**

If $\vec{x}$ is a non-zero vector we can build an unit norm vector $\vec{u}$ of same direction and orientation (colinear) by the following operation that is used a lot in physics:

$$\boxed{\vec{u}=\frac{\vec{x}}{\|\vec{x}\|}} \tag{6.5.58}$$

We will name "non-zero angle of vectors $\vec{x}$ and $\vec{y}$" the angle of two arrows of common origin representing one being $\vec{x}$ and the other $\vec{y}$.

However, more strictly speaking a "norm" is defined on a real vector space (or complex) E, so that we speak then of "normalized vector space" is an application:

$$f\,(\|\;\|):\mathbb{R}^n\mapsto\mathbb{R}^+ \tag{6.5.59}$$

satisfying the following properties:

P1. Positivity:

$$\forall\vec{x}\in E\quad\|\vec{x}\|\geq 0 \tag{6.5.60}$$

P2. Linearity:

$$\forall\lambda\in\mathbb{R},\forall\vec{x}\in E\quad\|\lambda\vec{x}\|=|\lambda|\cdot\|\vec{x}\| \tag{6.5.61}$$

where we take the modulus of the constant if this is not in the set of real numbers $\mathbb{R}$ but in the set of complex numbers $\mathbb{C}$.

P3. Nullity (often associated with the property P1):

$$\forall\vec{x}\in E\quad\|\vec{x}\|=0\Leftrightarrow\vec{x}=0 \tag{6.5.62}$$

P4. Minkowski inequality (triangle inequality):

$$\forall\vec{x},\vec{y}\in E\quad\|\vec{x}+\vec{y}\|\leq\|\vec{x}\|+\|\vec{y}\| \tag{6.5.63}$$

That we will prove further below.

Remarks

R1. These properties are mainly imposed by our intuitive approach of Euclidean space (vector space of finite dimension over the field of real number $\mathbb{R}$ and with a scalar product that we will see later) and its geometric interpretation (through the fact that it is also an affine space $\mathrm{A}\mathbb{R}^n$).

R2. We will prove a little further below the property P4 under the name of "triangle inequality" and we will do a little more general study of this inequality under the name "Minkowski inequality" in the section Topology.

6.5.4.1 Scalar Product (Dot Product)

Definition (#6.5.239): An "Euclidean vector space" of dimension n is a vector space (real and of finite dimensional for the purists) with a specific operation, the "scalar product" also named "dot product" which we denote (notation specific to this textbook) regarding to the special case of vectors:

$$\|\vec{x}\| \cdot \|\vec{y}\| \cos(\theta) = \vec{x} \circ \vec{y} = x_1 y_1 + x_2 y_2 + \ldots + x_n y_n \tag{6.5.64}$$

Remarks

R1. We find in some books (for information) the notation $(\vec{x}|\vec{y})$ or $\langle\vec{x}|\vec{y}\rangle$ in the generalization of this definition as we shall see a little further below. According to the standard ISO 80000-2:2009 *Mathematical signs and symbols to be used in the natural sciences and technology* we should write the dot product as $a \cdot b$. Using the Linear Algebra notation that we will see later, the dot product will be written: $\vec{x}^T\vec{y}$.

R2. The scalar product has a huge importance in the whole field of mathematics and physics; and we will see it again in all following chapters of this textbook. It is therefore necessary to carefully understand what follows.

R3. The scalar product may be viewed as a projection of the length of a vector along the length of another one as we will see later.

This scalar product has the following properties (most of which stem from the definition itself) in a Euclidean space:

P1. Commutativity: $\vec{x} \circ \vec{y} = \vec{y} \circ \vec{x}$

P2. Associativity: $\alpha(\vec{x} \circ \vec{y}) = (\alpha\vec{x}) \circ \vec{y} = \vec{x} \circ (\alpha\vec{y})$

P3. Distributivity: $\vec{x} \circ (\vec{y} + \vec{z}) = \vec{x} \circ \vec{y} + \vec{x} \circ \vec{z}$

P4. Non-degenerated: $\vec{x} \circ \vec{y} = 0$ then $\forall\vec{x} \neq \vec{0} \Rightarrow \vec{y} = \vec{0}$

P5. Squared scalar: $\|\vec{x}\|^2 = \vec{x} \circ \vec{x}$ and $\vec{x} \circ \vec{x} > 0$ if $\vec{x} \neq \vec{0}$

P6. Bi-linearity: $(\alpha\vec{x} + \beta\vec{y}) \circ \vec{z} = \alpha(\vec{x} \circ \vec{z}) + \beta(\vec{y} \circ \vec{z})$

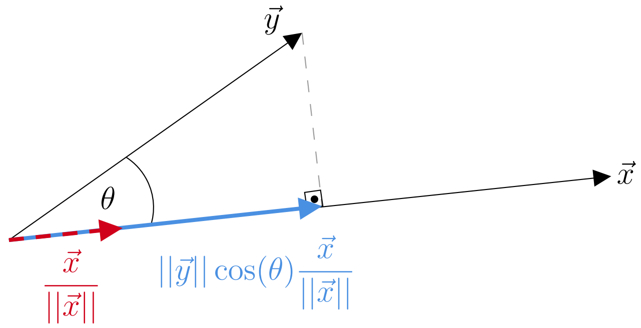
Only the latter property requires perhaps a proof (and one of the results of the proof will be useful to us later to prove another very important property of the scalar product):

Proof 6.31.2. Given:

$$\text{proj}_x\vec{y} = \|\vec{y}\| \cos(\theta) \frac{\vec{x}}{\|\vec{x}\|} = \cos(\theta) \left(\frac{\|\vec{y}\| \cdot \|\vec{x}\|}{\|\vec{x}\|^2}\right) \vec{x} = \cos(\theta) \left(\frac{\|\vec{y}\| \cdot \|\vec{x}\|}{\vec{x} \circ \vec{x}}\right) \vec{x} \tag{6.5.65}$$

which is the "orthogonal projection vector" (the x on index of project meaning "the vector $\vec{x}$") of the vector $\vec{y}$ of standardization at the unit of vector $\vec{x}$.

Using the scalar (dot) product, the vector $\text{proj}_x\vec{y}$ can be expressed otherwise, we just need to take the relation that we have seen above:



$\vec{y}$
$\vec{x}$
θ
$\dfrac{\vec{x}}{||\vec{x}||}$
$||\vec{y}||\cos(\theta)\dfrac{\vec{x}}{||\vec{x}||}$

Figure 6.107: **Geometrical representation of the dot product (projection)**

$$\|\vec{x}\|\cdot\|\vec{y}\|\cos(\theta)=\vec{x}\circ\vec{y}\Leftrightarrow\|\vec{x}\|\cdot\|\vec{y}\|\cos(\theta)=\vec{x}\circ\vec{y} \tag{6.5.66}$$

and introduce it into $\text{proj}_x\vec{y}$ to get:

$$\text{proj}_x\vec{y}=\cos(\theta)\left(\frac{\|\vec{y}\|\cdot\|\vec{x}\|}{\vec{x}\circ\vec{x}}\right)\vec{x}=\frac{\vec{x}\circ\vec{y}}{\vec{x}\circ\vec{x}}\vec{x} \tag{6.5.67}$$

The norm of $\text{proj}_x\vec{y}$ is written:

$$\|\text{proj}_x\vec{y}\|=\frac{|\vec{x}\circ\vec{y}|}{\|\vec{x}\|\cdot\|\vec{x}\|}\|\vec{x}\|=\frac{|\vec{x}\circ\vec{y}|}{\|\vec{x}\|} \tag{6.5.68}$$

If $\vec{x}$ has a unit norm, the relations of previous projections are simplified and become obviously:

$$\text{proj}_x\vec{y}=(\vec{x}\circ\vec{y})\vec{x}\quad\text{and}\quad\|\text{proj}_x\vec{y}\|=|\vec{x}\circ\vec{y}| \tag{6.5.69}$$

By elementary geometric considerations (distributivity of the scalar product), it is easy to realize that:

$$\text{proj}_v(\vec{x}+\vec{y})=\text{proj}_v\vec{x}+\text{proj}_v\vec{y}\quad\text{and}\quad\text{proj}_v\alpha\vec{x}=\alpha\text{proj}_v\vec{x} \tag{6.5.70}$$

If we now return to the proof of the bi-linearity property:

$$(\alpha\vec{x}+\beta\vec{y})\circ\vec{z}=\alpha(\vec{x}\circ\vec{z})+\beta(\vec{y}\circ\vec{z}) \tag{6.5.71}$$

We have in a first time:

$$\text{proj}_v((\alpha\vec{x}+\beta\vec{y}))=\text{proj}_v\alpha\vec{x}+\text{proj}_v\beta\vec{y}=\alpha\text{proj}_v\vec{x}+\beta\text{proj}_v\vec{y} \tag{6.5.72}$$

and, from the definition of the property of the orthogonal projection, it comes immediately by a one-to-one correspondence:

$$\begin{gathered}\text{proj}_v(\alpha\vec{x}+\beta\vec{y})=\text{proj}_v\alpha\vec{x}+\text{proj}_v\beta\vec{y}=\alpha\text{proj}_v\vec{x}+\beta proj_v\vec{y}\\ \frac{(\alpha\vec{x}+\beta\vec{y})\circ\vec{z}}{\vec{z}\circ\vec{z}}\vec{z}=\alpha\frac{\vec{x}\circ\vec{z}}{\vec{z}\circ\vec{z}}\vec{z}+\beta\frac{\vec{y}\circ\vec{z}}{\vec{z}\circ\vec{z}}\vec{z}\end{gathered} \tag{6.5.73}$$

hence the property P6 that follows by multiplying the two members of equality by $\vec{z}\circ\vec{z}$ and after by simplification by $\vec{z}$.

■ Q.E.D.

Definitions (#6.5.240):

D1. A vector space E is said to be an "proper Euclidean vector space" if $\forall \vec{x} \in E \qquad \|\vec{x}\| > 0$.

D2. We say that the vectors $\vec{x}$ and $\vec{y}$ are "orthogonal vectors" if they are non-null and that their scalar product is equal to zero (their angle is equal to $\pi/2$).

D3. A basis of vectors $(\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_n)$ is said to be an "orthonormal basis" if all the vectors $\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_n$ are pairwise orthogonal and their norm is equal to the unit (thus constituting a: free family).

Remark

We will see in the section Tensor Calculus (we could have done here too but we don't use it for a practical case in this section therefore...) how from a set of independent vectors build an orthogonal basis. This is what the reader will find under the name "(Gram-)Schmidt orthogonalization method" (see page 2062).

By a simple geometric argument, we see that every vector is the sum of its orthogonal projections on the vectors of an orthonormal basis, that is, if $(\vec{e}_1, \vec{e}_2, \vec{e}_2) = (\vec{u}, \vec{v}, \vec{w})$ is an orthonormal basis in $\mathbb{R}^3$ for example:

$$\vec{x} = (\vec{x} \circ \vec{e}_1)\vec{e}_1 + (\vec{x} \circ \vec{e}_2)\vec{e}_2 + (\vec{x} \circ \vec{e}_3)\vec{e}_3 = \alpha_1 \vec{e}_1 + \alpha_2 \vec{2} + \alpha_3 \vec{e}_3 = \vec{a}_x + \vec{a}_y + \vec{a}_z \tag{6.5.74}$$

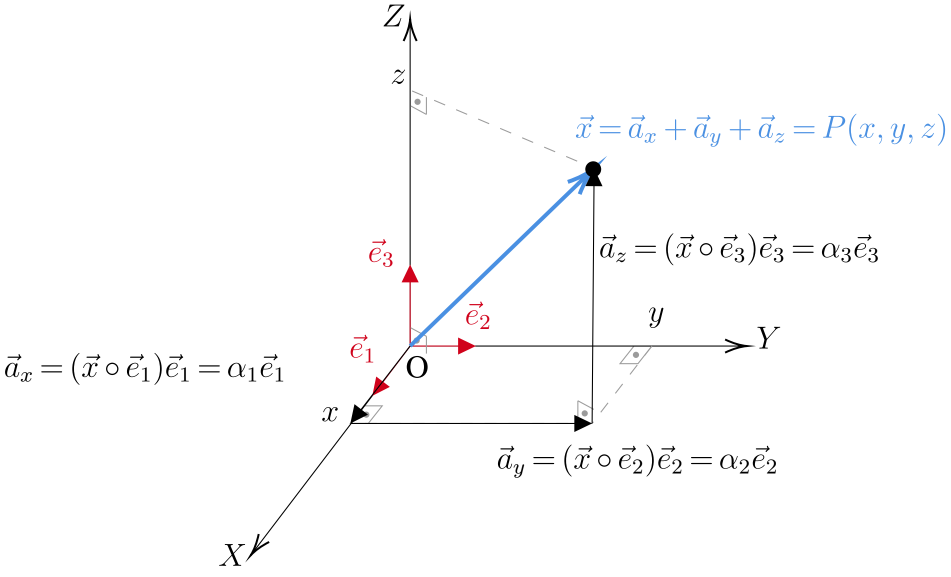

Z
z
$\vec{x} = \vec{a}_x + \vec{a}_y + \vec{a}_z = P(x, y, z)$
$\vec{e}_3$
$\vec{a}_z = (\vec{x} \circ \vec{e}_3)\vec{e}_3 = \alpha_3 \vec{e}_3$
$\vec{e}_2$
y
Y
$\vec{a}_x = (\vec{x} \circ \vec{e}_1)\vec{e}_1 = \alpha_1 \vec{e}_1$
$\vec{e}_1$
O
x
$\vec{a}_y = (\vec{x} \circ \vec{e}_2)\vec{e}_2 = \alpha_2 \vec{e}_2$
X

Figure 6.108: Orthonormal basis projection example

This decomposition is also obtained by the P6 property of the scalar product. Indeed, consider the components $(\vec{x}_1, \vec{x}_2, \vec{x}_3)$ of a vector $\vec{x}$ in our orthonormal basis:

$$\begin{aligned}\vec{x} \circ \vec{e}_1 &= (x_1 \circ \vec{e}_1 + x_2 \circ \vec{e}_2 + x_3 \circ \vec{e}_3) \circ \vec{e}_1 \\ &= x_1(\vec{e}_1 \circ \vec{e}_1) + x_2(\vec{e}_2 \circ \vec{e}_1) + x_3(\vec{e}_3 \circ \vec{e}_1) \\ &= x_1\end{aligned} \tag{6.5.75}$$

since $\vec{e}_1 \circ \vec{e}_1 = 1$ and $\vec{e}_1 \circ \vec{e}_2 = 0$. Therefore we get immediately:

$$\vec{x} \circ \vec{e}_2 = x_2 \qquad \text{and} \qquad \vec{x} \circ \vec{e}_3 = x_3 \tag{6.5.76}$$

hence the decomposition.

Given the respective components (x_1, x_2, x_3) and (y_1, y_2, y_3) of the vectors $\vec{x}$ and $\vec{y}$ in a canonical orthonormal basis $(\vec{e}_1, \vec{e}_2, \vec{e}_3)$ we know now that we can write the scalar product in the form:

$$\vec{x} \circ \vec{y} = (x_1\vec{e}_1 + x_2\vec{e}_2 + x_3\vec{e}_3) \circ (y_1\vec{e}_1 + y_2\vec{e}_2 + y_3\vec{e}_3) \tag{6.5.77}$$

by the property P6 of the scalar product:

$$\vec{x} \circ \vec{y} = x_1\vec{e}_1 \circ (y_1\vec{e}_1 + y_2\vec{e}_2 + y_3\vec{e}_3) + x_2\vec{e}_2 \circ (y_1\vec{e}_1 + y_2\vec{e}_2 + y_3\vec{e}_3) + x_3\vec{e}_3 \circ (y_1\vec{e}_1 + y_2\vec{e}_2 + y_3\vec{e}_3) \tag{6.5.78}$$

using the properties P1 and P6 again:

$$\begin{gathered}\vec{x} \circ \vec{y} = x_1(y_1\vec{e}_1 + y_2\vec{e}_2 + y_3\vec{e}_3) \circ \vec{e}_1 + x_2(y_1\vec{e}_1 + y_2\vec{e}_2 + y_3\vec{e}_3) \circ \vec{e}_2 + x_3(y_1\vec{e}_1 + y_2\vec{e}_2 + y_3\vec{e}_3) \circ \vec{e}_3 \\ \Rightarrow \\ (y_1\vec{e}_1 + y_2\vec{e}_2 + y_3\vec{e}_3) \circ \vec{e}_1 = y_1\vec{e}_1 \circ \vec{e}_1 + y_2\vec{e}_2 \circ \vec{e}_1 + y_3\vec{e}_3 \circ \vec{e}_1 = y_1\|\vec{e}_1\|^2 = y_1 \\ (y_1\vec{e}_1 + y_2\vec{e}_2 + y_3\vec{e}_3) \circ \vec{e}_2 = y_1\vec{e}_1 \circ \vec{e}_2 + y_2\vec{e}_2 \circ \vec{e}_2 + y_3\vec{e}_3 \circ \vec{e}_2 = y_2\|\vec{e}_2\|^2 = y_2 \\ (y_1\vec{e}_1 + y_2\vec{e}_2 + y_3\vec{e}_3) \circ \vec{e}_3 = y_1\vec{e}_1 \circ \vec{e}_3 + y_2\vec{e}_2 \circ \vec{e}_3 + y_3\vec{e}_3 \circ \vec{e}_3 = y_3\|\vec{e}_3\|^2 = y_3\end{gathered} \tag{6.5.79}$$

Which finally gives us the very famous and important decomposition:

$$\boxed{\vec{x} \circ \vec{y} = x_1y_1 + x_2y_2 + x_3y_3} \tag{6.5.80}$$

This is one of the most important relation in the field of vector calculus, which we name "canonical scalar product" or "canonical dot product".

Remark

The angle θ of the dot product is sometimes denoted by:

$$\theta = \angle(\vec{x}, \vec{y}) \tag{6.5.81}$$

Now let us prove with a simple two dimensional case a property that physicist like a lot to characterize an orthogonal linear application (see section Linear Algebra page 1954): that the dot product is invariant under any orthogonal transformation (then abusively said "invariant under basis change...").

For this let us consider a vector $\vec{x} = (x_1, x_2)$ and the 2D rotation matrix (see section Numbers page 2648):

$$\begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \tag{6.5.82}$$

Now let us calculate:

$$\begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1\cos(\theta) - x_2\sin(\theta) \\ x_1\sin(\theta) + x_2\cos(\theta) \end{pmatrix} \tag{6.5.83}$$

So now let us consider two vectors $\vec{a}$ and $\vec{b}$ we have proved just above that their dot product is given by:

$$\vec{a}\circ\vec{b}=a_1b_1+a_2b_2+a_3b_3 \tag{6.5.84}$$

And after the chosen orthogonal transformation we get:

$$\begin{aligned}\vec{a}'\circ\vec{b}'=&\begin{pmatrix}a_1\cos(\theta)-a_2\sin(\theta)\\a_1\sin(\theta)+a_2\cos(\theta)\end{pmatrix}\circ\begin{pmatrix}b_1\cos(\theta)-b_2\sin(\theta)\\b_1\sin(\theta)+b_2\cos(\theta)\end{pmatrix}\\&=(a_1\cos(\theta)-a_2\sin(\theta))\,(b_1\cos(\theta)-b_2\sin(\theta))\\&+(a_1\sin(\theta)+a_2\cos(\theta))\,(b_1\sin(\theta)+b_2\cos(\theta))\\&=a_1b_1\cos^2(\theta)-a_2b_1\sin(\theta)\cos(\theta)-a_1b_2\cos(\theta)\sin(\theta)+a_2b_2\sin^2(\theta)\\&+a_1b_1\sin^2(\theta)+a_2b_1\cos(\theta)\sin(\theta)+a_1b_2\sin(\theta)\cos(\theta)+a_2b_2\cos^2(\theta)\\&=a_1b_1+a_2b_2=\vec{a}\circ\vec{b}\end{aligned} \tag{6.5.85}$$

So the dot product is indeed invariant under this rotation that is a special 2D case of orthogonal transformation and in fact under any other orthogonal transformation.

Remark

The dot product, must not be confused with the "element-wise multiplication" of two vectors defined by:

$$\boxed{\vec{x}\diamond\vec{y}=\begin{pmatrix}x_1y_1\\x_2y_2\\\vdots\\x_ny_n\end{pmatrix}} \tag{6.5.86}$$

6.5.4.1.1 Cauchy–Schwarz inequality

In mathematics, the Cauchy–Schwarz inequality is a useful inequality encountered in many different settings, such as linear algebra, analysis, probabilities, statistics and other areas (just read this textbook entirely to have an idea...). It is considered to be one of the most important inequalities in all of mathematics!!!

The relation:

$$(\vec{x}\circ\vec{y})\circ(\vec{x}\circ\vec{y})=\vec{x}\circ\vec{x}+\vec{x}\circ\vec{y}+\vec{y}\circ\vec{x}+\vec{y}\circ\vec{y}=\vec{x}\circ\vec{x}+\vec{y}\circ\vec{y}+2\vec{x}\circ\vec{y} \tag{6.5.87}$$

can also be trivially written as follows if we use the concept of the norm and the definition of the scalar product:

$$(\vec{x}\circ\vec{y})\circ(\vec{x}\circ\vec{y})=\|\vec{x}+\vec{y}\|^2=\|\vec{x}\|^2+\|\vec{y}\|^2+2\vec{x}\circ\vec{y} \tag{6.5.88}$$

It is interesting to notice that if both $\vec{x}$ and $\vec{y}$ are orthogonal vectors, we fall back on the result of a famous theorem: the Pythagorean theorem!

Indeed, therefore we have if the two vectors are orthogonal:

$$\vec{x}\circ\vec{y}=0 \tag{6.5.89}$$

This gives us:

$$\boxed{\|\vec{x}+\vec{y}\|^2=\|\vec{x}\|^2+\|\vec{y}\|^2} \tag{6.5.90}$$

This relations is very important in physics and mathematics. It must be remembered!

Theorem 6.32. We name "Cauchy-Schwarz inequality", the inequality, valid for any choice of the vectors $\vec{x}$ and $\vec{y}$, the relation:

$$\boxed{|\vec{x}\circ\vec{y}|\leq\|\vec{x}\|\|\vec{y}\|} \tag{6.5.91}$$

Which can also be written as:

$$|\vec{x}\circ\vec{y}|^2\leq(\vec{x}\circ\vec{x})\cdot(\vec{y}\circ\vec{y}) \tag{6.5.92}$$

First we will consider as obvious that equality only occurs when the two vectors are collinear.

Proof 6.32.1. We put ourself in the case where $\vec{x},\vec{y}\neq\vec{0}$. So then $\lambda\in\mathbb{R}$ we have obviously according to the properties of the scalar product:

$$(\vec{x}+\lambda\vec{y})\circ(\vec{x}+\lambda\vec{y})=\vec{x}\circ\vec{x}+2\lambda\vec{x}\circ\vec{y}+\lambda^2\vec{y}\circ\vec{y}\geq 0 \tag{6.5.93}$$

So this is a simple equation of the second degree where variable is λ. Remembering what we saw in our study of polynomials of second degree (see section Calculus page 1368), the previous relation (that it is always greater than or equal to zero) is satisfied if the discriminant b^2-4ac is negative or zero. In other words, if:

$$4(\vec{x}\circ\vec{y})^2-4(\vec{x}\circ\vec{x})(\vec{y}\circ\vec{y})\leq 0 \tag{6.5.94}$$

Thus after simplification:

$$(\vec{x}\circ\vec{y})^2-(\vec{x}\circ\vec{x})(\vec{y}\circ\vec{y})\leq 0\Rightarrow(\vec{x}\circ\vec{y})^2\leq(\vec{x}\circ\vec{x})(\vec{y}\circ\vec{y}) \tag{6.5.95}$$

■ Q.E.D.

When E is in $\mathbb{R}^n$, the Cauchy-Schwarz inequality is written with the vector components:

$$\sum_{i=1}^{n} a_i b_i \leq \sqrt{\sum_{i=1}^{n} a_i^2 \sum_{i=1}^{n} b_i^2} \tag{6.5.96}$$

In the particular case where $\forall i, b_i = 1$ it becomes:

$$\left(\sum_{i=1}^{n} a_i\right)^2 \leq n \sum_{i=1}^{n} a_i^2 \tag{6.5.97}$$

or even:

$$\left(\frac{1}{n}\sum_{i=1}^{n} a_i\right)^2 \leq \frac{1}{n} \sum_{i=1}^{n} a_i^2 \tag{6.5.98}$$

which shows that the square of the arithmetic mean is less than or equal to the arithmetic mean of the squares. This result is important for the study of Statistics!

Furthermore, using the property of the cosine and the Cauchy-Schwarz inequality we can write immediately:

$$-1 \leq \frac{\vec{x} \circ \vec{y}}{\|\vec{x}\| \cdot \|\vec{y}\|} \leq 1 \tag{6.5.99}$$

relation that we will see again in the context of the study of Statistics (see section Statistics page 613).

Notice that the Cauchy-Schwarz inequality is a special case of the "Hölder's inequality for sums defined by:

$$\sum_{k=1}^{n} |a_k b_k| \leq \left(\sum_{k=1}^{n} |a_k|^p\right)^{1/p} \left(\sum_{k=1}^{n} |b_k|^q\right)^{1/q} \tag{6.5.100}$$

Indeed, plugging $p = q = 2$ into Holder's inequality we get:

$$\sum_{k=1}^{n} |a_k b_k| \leq \left(\sum_{k=1}^{n} |a_k|^2\right)^{1/2} \left(\sum_{k=1}^{n} |b_k|^2\right)^{1/2} \tag{6.5.101}$$

Squaring either side we get exactly Cauchy's inequality;

$$\left(\sum_{k=1}^{n} |a_k b_k|\right)^2 \leq \left(\sum_{k=1}^{n} |a_k|^2\right) \left(\sum_{k=1}^{n} |b_k|^2\right) \tag{6.5.102}$$

6.5.4.1.2 Triangle Inequalities

By majoring $2\vec{x}\circ\vec{y}$ by $2\|\vec{x}\|\cdot\|\vec{y}\|$ (using the Cauchy-Schwarz inequality!) in the relation already establish previously:

$$\|\vec{x}+\vec{y}\|^2=\|\vec{x}\|^2+\|\vec{y}\|^2+2\vec{x}\circ\vec{y} \tag{6.5.103}$$

we get:

$$\|\vec{x}+\vec{y}\|^2\leq\|\vec{x}\|^2+\|\vec{y}\|^2+2\|\vec{x}\|\cdot\|\vec{y}\|=(\|\vec{x}\|+\|\vec{y}\|)^2 \tag{6.5.104}$$

which takes us immediately by taking the root square the "triangle inequality" that is very useful for the study of Sequences and Series and also in Topology:

$$\boxed{\|\vec{x}+\vec{y}\|\leq\|\vec{x}\|+\|\vec{y}\|} \tag{6.5.105}$$

Also sometimes denoted:

$$\|\vec{z}\|\leq\|\vec{x}\|+\|\vec{y}\| \tag{6.5.106}$$

The concept can be illustrated also as following:

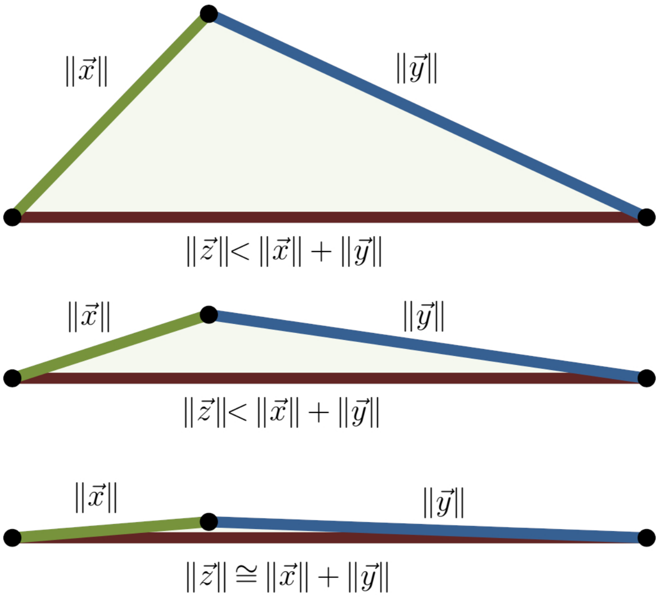

$\|\vec{x}\|$
$\|\vec{y}\|$
$\|\vec{z}\|<\|\vec{x}\|+\|\vec{y}\|$
$\|\vec{x}\|$
$\|\vec{y}\|$
$\|\vec{z}\|<\|\vec{x}\|+\|\vec{y}\|$
$\|\vec{x}\|$
$\|\vec{y}\|$
$\|\vec{z}\|\cong\|\vec{x}\|+\|\vec{y}\|$

Figure 6.109: **Illustration of the triangular inequality**

And obviously in the limit case (when all vectors are on the same line) we have: $\|\vec{z}\|=\|\vec{x}\|+\|\vec{y}\|$.

Remark

The generalization of this inequality relatively to the choice of the norm (that is to say: the way we define a distance) as we will see in the section of Topology, gives what we name the "Minkowski inequality".

By applying one time the triangle inequality to the vectors $\vec{x}$ and $(\vec{y}-\vec{x})$ and another time to vectors $(\vec{y})$ and $(\vec{x}-\vec{y})$ we get the variant:

$$\|\vec{x}\|-\|\vec{y}\|\leq\|\vec{x}-\vec{y}\| \tag{6.5.107}$$

6.5.4.1.3 General Scalar/Dot Product

Let us see now another and little more general, formal and abstract way to define the dot product while trying to stay as simple as possible (caution! in the general case the notation of the scalar product changes!).

First the reader must know that, as we will see it in the section of Functional Analysis, all the concepts studied until now in this section can also be applied to a special category of functions! Yeeesss!!! Add, subtract functions like we do it with vectors is obvious but you must know that some more or less complicated functions are colinear or orthogonal (think to affine functions!) and furthermore there are not limited to $\mathbb{R}$ but can be extended to $\mathbb{C}$ easily and hence the scalar product departure set.

To make thinks no too much complicate we will focus here only on a gentle generalization of the dot product to vectors (we will come back on the case of functions in the section of Functional Analysis later).

Definition (#6.5.241): Let E be a real vector space (once again we focus here only on simple vectors for the moment). A "positive symmetric bilinear form" on E also named "inner product", is an application:

$$\phi:\begin{cases}E\times E\mapsto\mathbb{R}\\(x,y)\mapsto\langle x|y\rangle\end{cases}\tag{6.5.108}$$

which satisfies by definition:

P1. Positivity:

$$\forall\vec{x}\in E:\qquad\phi(\vec{x},\vec{x})=\langle\vec{x}|\vec{x}\rangle\geq 0\in\mathbb{R}\tag{6.5.109}$$

P2. Nullity (defined):

$$\forall\vec{x}\in E:\qquad\phi(\vec{x},\vec{x})=\langle\vec{x}|\vec{x}\rangle=0\Rightarrow\vec{x}=0\tag{6.5.110}$$

P3. Symmetry (defined):

$$\forall(\vec{x},\vec{y})\in E^2:\qquad\phi(l\vec{x},\vec{y})=\phi(\vec{y},\vec{x})\Leftrightarrow\langle l\vec{x}|\vec{y}\rangle=\langle\vec{y}|\vec{x}\rangle\tag{6.5.111}$$

P4. The bilinearity (bilinear form) with, in order, the "linearity on the left" and "linearity on the right":

$$\forall(\vec{x},\vec{y},\vec{z})\in E^3,\forall(\lambda,\mu)\in\mathbb{R}^2:\begin{cases}\langle\lambda\vec{x}+\mu\vec{y}|\vec{z}\rangle=\lambda\langle\vec{x}|\vec{z}\rangle+\mu\langle\vec{y}|\vec{z}\rangle\\\langle\vec{x}|\lambda\vec{y}+\mu\vec{z}\rangle=\lambda\langle\vec{x}|\vec{y}\rangle+\mu\langle\vec{x}|\vec{z}\rangle\end{cases}\tag{6.5.112}$$

P... And so on... we have the same six properties as the scalar product as the both are the same if we focus only on vectors in $\mathbb{R}$.

Remark

Again, these properties are mainly imposed by our intuitive approach of the Euclidean space and its geometric interpretation.

Definition (#6.5.242): A space E provided with a scalar product is named in general (with the departure set in $\mathbb{C}$) a "pre-Hilbert space" or "inner product space". If E is of finite dimension, then we speak of "Euclidean space".

Remark

If E is a Euclidean space, then if the determinant $\det(\vec{e}_1,\vec{e}_2,\ldots,\vec{e}_n)$ is equal ± 1 (see section Linear Algebra page 1961) we speak of "oriented Euclidean space".

We will see in our study of Topology (see section of the same name page 2446) that the properties of the scalar product are the foundation bricks to set a norm and therefore a distance in a metric space. This distance will be given according to what we will see in the section of Topology:

$$\forall \vec{x},\vec{y}\in E \quad d(\vec{x},\vec{y}):=\|\vec{x}-\vec{y}\| \tag{6.5.113}$$

Definition (#6.5.243): We say that a space E having a dot product (inner product) $\langle\cdot|\cdot\rangle$ is a "Hilbert space" if this space is complete for the metric defined above.

In other words, having a metric space provided with a distance generated by a scalar product is one thing. Then having a measurable distance is another one!!! A Hilbert space has thus distances measurable in the topological sense because the set we are working on is continuous and any point can be approached indefinitely (imagine having a rule and you can not approach on this rule the points that define the dimensions of your object... it would be embarrassing...). So without complete space a lot of theorems of functional analysis (that is strongly linked to vector calculus) could not be used in the study of vector spaces and this would be very embarrassing in Quantum Wave Physics for example...

Formally, remember that a metric space is complete if all Cauchy sequences (see section Sequences and Series page 1657) of this space are converging (see section Fractals page 7046) in a metric space (see section Topology page 2447).

6.5.4.2 Cross Product

The cross product of two vectors is a proper operation to the dimension 3. To introduce it, it is first necessary to orient the space intended to receive it. The orientation is defined by the concept of "determinant", therefore we will begin with a brief introduction to the study of this concept. This study will be repeated later in more details in the analysis of linear systems in the section of Linear Algebra.

Definition (#6.5.244): We name basically "determinant" of two column vectors of $\mathbb{R}^2$ (for the general form of the determinant see the section of Linear Algebra page 1961):

$$\begin{pmatrix}a_1\\a_2\end{pmatrix},\begin{pmatrix}b_1\\b_2\end{pmatrix} \tag{6.5.114}$$

and we denote it:

$$\det([\vec{a},\vec{b}])=\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} \tag{6.5.115}$$

the number:

$$\det([\vec{a},\vec{b}])=a_1b_2-a_2b_1 \tag{6.5.116}$$

We name determinant of three column vectors of $\mathbb{R}^3$ (once again see the section Linear Algebra for a generalization):

$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix},\begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix},\begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} \tag{6.5.117}$$

and we denote it:

$$\det([\vec{a},\vec{b},\vec{c}])=\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix} \tag{6.5.118}$$

the number:

$$\begin{aligned}\det([\vec{a},\vec{b},\vec{c}])&=a_1\begin{vmatrix} b_2 & c_2 \\ b_3 & c_3 \end{vmatrix}-a_2\begin{vmatrix} b_1 & c_1 \\ b_3 & c_3 \end{vmatrix}-a_3\begin{vmatrix} b_1 & c_1 \\ b_2 & c_2 \end{vmatrix}\\ &=a_1b_2c_3+a_2b_3c_1+a_3b_1c_2-a_1b_3c_2-a_2b_1c_3-a_3b_2c_1\end{aligned} \tag{6.5.119}$$

Thus, the function that associates to each pair of column vectors of $\mathbb{R}^2$ (or respectively to each triplet of column vectors of $\mathbb{R}^3$) has a determinant named "determinant of order 2" (respectively "determinant of order 3")

As we will prove it in the section of Linear Algebra the determinant has the property of being multiplied by -1 if one of the column vectors is replaced by its opposite or two of its column vectors are exchanged. In addition, the determinant is non-zero if and only if its column vectors are linearly independent (the proof - that has a great importance in Applied Mathematics - is a few lines further below and a generalization can be found in the section of Linear Algebra).

Definition (#6.5.245): Given x_1,x_2,x_3 and y_1,y_2,y_3 the respective components of the vectors $\vec{x}$ and $\vec{y}$ in the orthonormal basis $(\vec{e}_1,\vec{e}_2,\vec{e}_3)$. We name "cross-product" of $\vec{x}$ and $\vec{y}$, and we denote it in most books by:

$$\vec{x}\times\vec{y} \tag{6.5.120}$$

and in a minority of books:

$$\vec{x}\wedge\vec{y} \tag{6.5.121}$$

the vector:

$$\vec{x}\times\vec{y}=(x_2y_3-x_3y_2)\vec{e}_1+(x_3y_1-x_1y_3)\vec{e}_2+(x_1y_2-x_2y_1)\vec{e}_3 \tag{6.5.122}$$

or as components:

$$\boxed{\vec{x}\times\vec{y}=\begin{pmatrix} x_2y_3-x_3y_2 \\ x_3y_1-x_1y_3 \\ x_1y_2-x_2y_1 \end{pmatrix}=\begin{pmatrix} 0 & -x_3 & x_2 \\ x_3 & 0 & -x_1 \\ -x_2 & x_1 & 0 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}} \tag{6.5.123}$$

The matrix form above will be very useful to us in the section Mechanics for the construction of the Inertial Matrix.

Remarks

R1. The first notation is the international notation due to Gibbs (which we will use throughout this textbook), the second is the French notation due to Burali-Forti (quite annoying because confusing with the notation of the operator AND in Proof Theory or Logical Systems).

R2. It is usually quite annoying to remember by heart the relations that form the cross product. But fortunately there are at least three good mnemonics techniques:

1. The first and probably the fastest method is to remember by heart only one of the expressions of the components of the cross product and after by decrement of the indices (by starting again from 3 when we reaches 0) get all the other components. But we must still find a simple way to remember by heart one of the components... A good way is the following mathematical property of two collinear vectors giving an easy with to find back the third component (the one along the z-axis):

   Given two colinear vectors in the same plane, then:

   $$\vec{x}=\lambda\vec{y}\Leftrightarrow\begin{pmatrix}x_1\\x_2\end{pmatrix}=\lambda\begin{pmatrix}y_1\\y_2\end{pmatrix}\Leftrightarrow\frac{x_1}{y_1}=\frac{x_2}{y_2}\Leftrightarrow x_1y_2-x_2y_1=0 \qquad (6.5.124)$$

   We fall back on the expression of the third component of the cross product of two vectors.

   Or if you want to remember only the first component given in letters by $z_x=x_yy_z-x_zy_y$ the indices gives "xyzzy" (like a name of a person...). The second and third equations can be obtained from the first by simply vertically rotating the subscripts, $x\rightarrow y\rightarrow z\rightarrow x$.

2. The second method that we will see in details during our study of the section of Tensor Calculus is to use the Levi-Civita antisymmetric symbol. This method is certainly the most aesthetic of all but not necessarily the fastest to develop and the easiest to remember. We give here just the expression without explanations at the moment as we will study this later (but t is also useful to get the general expression of the determinant):

   $$(\vec{x}\times\vec{y})_i=\varepsilon_{ijk}x_jy_k \qquad (6.5.125)$$

3. The latter method is quite simple and trivial but it implicitly uses the first method as you must remember how the calculate at least a 2×2 determinant. The idea is the following: the i-th component of $\vec{x}\times\vec{y}$ is the determinant of the two column vectors from which we have removed the i-th term, the second determinant is, however, with a "$-$" sign such that:

   $$\vec{x}\times\vec{y}=\begin{vmatrix}x_2 & y_2\\x_3 & y_3\end{vmatrix}\vec{e}_1-\begin{vmatrix}x_1 & y_1\\x_3 & y_3\end{vmatrix}\vec{e}_2+\begin{vmatrix}x_1 & y_1\\x_2 & y_2\end{vmatrix}\vec{e}_3 \qquad (6.5.126)$$

It is important, even if it is relatively simple to remember, that the different cross products for orthogonal basis vectors are (and especially for the canonical basis) first:

$$\vec{e}_1\times\vec{e}_1=\vec{e}_2\times\vec{e}_2=\vec{e}_3\times\vec{e}_3=\vec{0} \tag{6.5.127}$$

and also:

$$\vec{e}_1\times\vec{e}_2=\vec{e}_3,\quad \vec{e}_2\times\vec{e}_3=\vec{e}_1,\quad \vec{e}_3\times\vec{e}_1=\vec{e}_2 \tag{6.5.128}$$

These relations can be condensed using the Dirac function symbol (see section Tensor Calculus page 2044):

$$\vec{e}_i\times\vec{e}_j=\delta_{ij} \tag{6.5.129}$$

The cross product also has the following properties that we will prove just now:

P1. Antisymmetry:

$$\vec{x}\times\vec{y}=-(\vec{y}\times\vec{x}) \tag{6.5.130}$$

P2. Linearity:

$$(\alpha\vec{x}+\beta\vec{y})\times\vec{z}=\alpha(\vec{x}\times\vec{z})+\beta(\vec{y}\times\vec{z}) \tag{6.5.131}$$

P3. If and only if $\vec{x}$ and $\vec{y}$ are linearly independent (very important!):

$$\vec{x}\times\vec{y}\neq\vec{0} \tag{6.5.132}$$

P4. Non associativity:

$$\vec{x}\times(\vec{y}\times\vec{z})\neq(\vec{x}\times\vec{y})\times\vec{z} \tag{6.5.133}$$

P5. Distributivity over the sum:

$$\vec{x}\times(\vec{a}+\vec{b})=\vec{x}\times\vec{a}+\vec{x}\times\vec{b} \tag{6.5.134}$$

The first two properties directly derived from the definition and the property P4 is easily verified by developing the components and comparing the results.

Then let us prove the third property which is very important in linear algebra (next section) and the fifth one (because requested by a reader).

Theorem 6.33. If and only if $\vec{x}$ and $\vec{y}$ are linearly independent (very important!):

$$\vec{x}\times\vec{y}\neq\vec{0} \tag{6.5.135}$$

Proof 6.33.1. Given two vectors $\vec{x}(x_1,x_2,x_3)$ and $\vec{y}(y_1,y_2,y_3)$. If the two vectors are linearly dependent then there exists an $\alpha\in\mathbb{R}$ such that we can write:

$$\vec{y}=\alpha\vec{x} \tag{6.5.136}$$

If we develop the cross product of the two vectors that are dependent to a given factor α, we get:

$$\begin{aligned}\vec{x}\times\vec{y}&=\vec{x}\times\alpha\vec{x}\\&=(x_2\alpha x_3-x_3\alpha x_2)\vec{e}_1+(x_3\alpha x_1-x_1\alpha x_3)\vec{e}_2+(x_1\alpha x_2-x_2\alpha x_1)\vec{e}_3\end{aligned} \tag{6.5.137}$$

It goes without saying that the above result is equal to the zero vector $\vec{0}$ if indeed the two vectors are linearly dependent.

■ Q.E.D.

Theorem 6.34. The cross product is distributive over the sum.

Proof 6.34.1.

$$\begin{aligned}\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}\times\left(\begin{pmatrix}a_1\\a_2\\a_3\end{pmatrix}+\begin{pmatrix}b_1\\b_2\\b_3\end{pmatrix}\right)=\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}\times\begin{pmatrix}a_1+b_1\\a_2+b_2\\a_3+b_3\end{pmatrix}&=\begin{pmatrix}x_2(a_3+b_3)-x_3(a_2+b_2)\\x_3(a_1+b_1)-x_1(a_3+b_3)\\x_1(a_2+b_2)-x_2(a_1+b_1)\end{pmatrix}\\&=\begin{pmatrix}x_2a_3-x_3a_2\\x_3a_1-x_1a_3\\x_1a_2-x_2a_1\end{pmatrix}+\begin{pmatrix}x_2b_3-x_3b_2\\x_3b_1-x_1b_3\\x_1b_2-x_2b_1\end{pmatrix}=\vec{x}\times\vec{a}+\vec{x}\times\vec{b}\end{aligned}\tag{6.5.138}$$

■ Q.E.D.

If we now assume that both vector $\vec{x}$ and $\vec{y}$ are linearly independent and non-zero vectors, we must prove that the following two cross product properties:

Theorem 6.35. The resultant of a cross product of two vectors $\vec{x}$ and $\vec{y}$ is a vector orthogonal (perpendicular) to $\vec{x}$ and $\vec{y}$ if they are not null.

Proof 6.35.1. To prove this we simply write the development using the dot product:

$$\vec{x}\circ(\vec{x}\times\vec{y})=x_1\begin{vmatrix}x_2&y_2\\x_3&y_3\end{vmatrix}-x_2\begin{vmatrix}x_1&y_1\\x_3&y_3\end{vmatrix}+x_3\begin{vmatrix}x_1&y_1\\x_2&y_2\end{vmatrix}=\begin{vmatrix}x_1&x_1&y_2\\x_2&x_2&y_2\\x_3&x_3&y_3\end{vmatrix}=0\tag{6.5.139}$$

This equation shows that the vector $\vec{x}$ is perpendicular to the resulting vector of cross product between $\vec{x}$ and $\vec{y}$.

■ Q.E.D.

Theorem 6.36. The cross product has for norm (module):

$$\|\vec{x}\|\cdot\|\vec{y}\|\sin(\theta)\tag{6.5.140}$$

where θ is the angle between $\vec{x}$ and $\vec{y}$.

Proof 6.36.1. To prove this we simply write the development of the norm of the cross product:

$$\begin{aligned}\|\vec{x}\times\vec{y}\|^2&=(x_2y_3-x_3y_2)^2+(x_3y_1-x_1y_3)^2+(x_1y_2-x_2y_1)^2\\&=(x_1^2+x_2^2+x_3^2)(y_1^2+y_2^2+y_3^2)-(x_1y_1+x_2y_2+x_3y_3)^2\\&=\|\vec{x}\|^2\|\vec{y}\|^2-\|\vec{x}\|^2\|\vec{y}\|^2\cos^2(\theta)=\|\vec{x}\|^2\|\vec{y}\|^2(1-\cos^2(\theta))=\|\vec{x}\|^2\|\vec{y}\|^2\sin^2(\theta)\end{aligned}\tag{6.5.141}$$

Finally:

$$\boxed{\|\vec{x}\times\vec{y}\|=\|\vec{x}\|\|\vec{y}\|\sin(\theta)}\tag{6.5.142}$$

■ Q.E.D.

We then notice that in the case where E is the Euclidean vector space, the norm of the cross product is the area (surface) of the parallelogram constructed on representatives of vectors $\vec{x}$ and $\vec{y}$ of common origin:

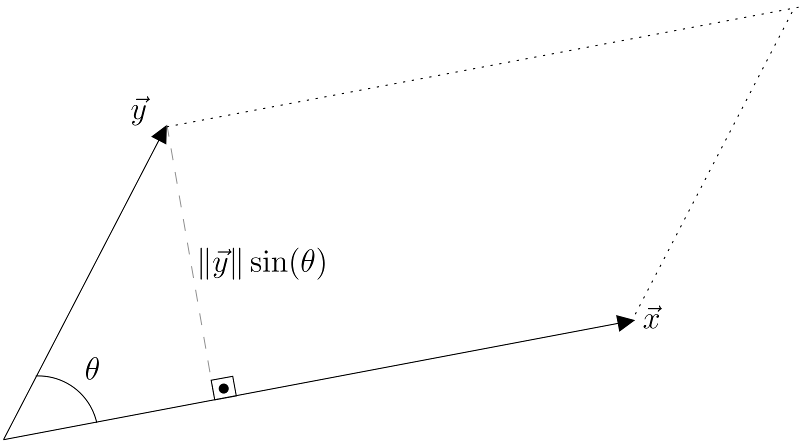


Figure 6.110: Geometrical representation of the cross product

If $\vec{x}$ and $\vec{y}$ are linearly independent, the triplet $(\vec{x},\vec{y},\vec{x}\times\vec{y})$ and also the triplet $(\vec{x},\vec{y},\vec{y}\times\vec{x})$ are direct.

Indeed, $(\vec{z}_1,\vec{z}_2,\vec{z}_3)$ being the components of $\vec{x}\times\vec{y}$ (in the basis $(\vec{e}_1,\vec{e}_2,\vec{e}_3)$), the determinant of passage of $(\vec{e}_1,\vec{e}_2,\vec{e}_3)$ to $(\vec{x}\times\vec{y},\vec{x},\vec{y})$ (for example) will be written:

$$\begin{vmatrix} z_1 & x_1 & y_1 \\ z_2 & x_2 & y_2 \\ z_3 & x_3 & y_3 \end{vmatrix} = z_1^2+z_2^2+z_3^2 \tag{6.5.143}$$

This determinant is positive, as at least one of the z_i is not zero, according to the third property of linear independence of the cross product.

Here are a few very important properties of practical utility of the cross product (particularly in the different sections of physics in this textbook!) that are trivial to check whether the developments with explicit components are done (we can make them on request if needed!):

P1. $\vec{x}\times(\vec{y}\times\vec{z})=(\vec{x}\circ\vec{z})\vec{y}-(\vec{x}\circ\vec{y})\vec{z}$

Remark

The latter relation is sometimes named the "Grassmann rule", or more commonly "dual vector product", or even by physicists "triple cross product", and it is important to note that without the parentheses the result is not unique!

Let us prove this identity! First the left part is equal to:

$$\begin{aligned}\vec{x}\times(\vec{y}\times\vec{z})=\vec{x}\times\left(\begin{pmatrix}y_1\\y_2\\y_3\end{pmatrix}\times\begin{pmatrix}z_1\\z_2\\z_3\end{pmatrix}\right)&=\vec{x}\times\begin{pmatrix}y_2z_3-y_3z_2\\y_3z_1-y_1z_3\\y_1z_2-y_2z_1\end{pmatrix}=\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}\times\begin{pmatrix}y_2z_3-y_3z_2\\y_3z_1-y_1z_3\\y_1z_2-y_2z_1\end{pmatrix}\\&=\begin{pmatrix}x_2(y_1z_2-y_2z_1)-x_3(y_3z_1-y_1z_3)\\x_3(y_2z_3-y_3z_2)-x_1(y_1z_2-y_2z_1)\\x_1(y_3z_1-y_1z_3)-x_2(y_2z_3-y_3z_2)\end{pmatrix}\end{aligned}\tag{6.5.144}$$

And for the right part:

$$\begin{aligned}(\vec{x}\circ\vec{z})\vec{y}-(\vec{x}\circ\vec{y})\vec{z}&=(x_1z_1+x_2z_2+x_3z_3)\vec{y}-(x_1y_1+x_2y_2+x_3y_3)\vec{z}\\&=\begin{pmatrix}(x_1z_1+x_2z_2+x_3z_3)y_1\\(x_1z_1+x_2z_2+x_3z_3)y_2\\(x_1z_1+x_2z_2+x_3z_3)y_3\end{pmatrix}-\begin{pmatrix}(x_1y_1+x_2y_2+x_3y_3)z_1\\(x_1y_1+x_2y_2+x_3y_3)z_2\\(x_1y_1+x_2y_2+x_3y_3)z_3\end{pmatrix}\\&=\begin{pmatrix}x_1z_1y_1+x_2z_2y_1+x_3z_3y_1-x_1y_1z_1-x_2y_2z_1-x_3y_3z_1\\x_1z_1y_2+x_2z_2y_2+x_3z_3y_2-x_1y_1z_2-x_2y_2z_2-x_3y_3z_2\\x_1z_1y_3+x_2z_2y_3+x_3z_3y_3-x_1y_1z_3-x_2y_2z_3-x_3y_3z_3\end{pmatrix}\\&=\begin{pmatrix}x_2z_2y_1+x_3z_3y_1-x_2y_2z_1-x_3y_3z_1\\x_1z_1y_2+x_3z_3y_2-x_1y_1z_2-x_3y_3z_2\\x_1z_1y_3+x_2z_2y_3-x_1y_1z_3-x_2y_2z_3\end{pmatrix}\end{aligned}\tag{6.5.145}$$

and then we see that there is indeed equality!

P2. $(\vec{x}\times\vec{y})\circ(\vec{z}\times\vec{v})=(\vec{x}\circ\vec{z})(\vec{y}\circ\vec{v})-(\vec{x}\circ\vec{v})(\vec{y}\circ\vec{z})$

P3. $(\vec{x}\times\vec{y})\circ\vec{z}=-(\vec{x}\times\vec{z})\circ\vec{y}$

P4. $(\vec{x}\times\vec{y})\circ\vec{z}=\vec{x}\circ(\vec{y}\times\vec{z})$

Indeed, let:

$$\vec{x}=(x_1,x_2,x_3)\,,\quad \vec{y}=(y_1,y_2,y_3)\,,\quad \vec{z}=(z_1,z_2,z_3\tag{6.5.146}$$

First compute $\vec{y}\times\vec{z}$:

$$\vec{y}\times\vec{z}=\begin{pmatrix}y_2z_3-y_3z_2\\y_3z_1-y_1z_3\\y_1z_2-y_2z_1\end{pmatrix}\tag{6.5.147}$$

Taking the dot product with $\vec{x}$:

$$\begin{aligned}\vec{x}\circ(\vec{y}\times\vec{z})&=x_1\,(y_2z_3-y_3z_2)+x_2\,(y_3z_1-y_1z_3)+x_3\,(y_1z_2-y_2z_1)\\&=x_1y_2z_3-x_1y_3z_2+x_2y_3z_1-x_2y_1z_3+x_3y_1z_2-x_3y_2z_1\end{aligned}\tag{6.5.148}$$

Now we compute $\vec{x}\times\vec{y}$:

$$\vec{x}\times\vec{y}=\begin{pmatrix}x_2y_3-x_3y_2\\x_3y_1-x_1y_3\\x_1y_2-x_2y_1\end{pmatrix}\tag{6.5.149}$$

Taking the dot product with $\vec{z}$:

$$\begin{aligned}(\vec{x}\times\vec{y})\circ\vec{z} &= (x_2y_3-x_3y_2)\,z_1+(x_3y_1-x_1y_3)\,z_2+(x_1y_2-x_2y_1)\,z_3\\ &= x_2y_3z_1-x_3y_2z_1+x_3y_1z_2-x_1y_3z_2+x_1y_2z_3-x_2y_1z\end{aligned} \tag{6.5.150}$$

Reordering the terms of of both intermediate results above gives exactly the equality:

$$\vec{x}\circ(\vec{y}\times\vec{z})=(\vec{x}\times\vec{y})\circ\vec{z} \tag{6.5.151}$$

P5. $\|\vec{x}\times\vec{y}\|^2=(\vec{x}\circ\vec{x})^2(\vec{y}\circ\vec{y})^2-(\vec{x}\circ\vec{y})^2$

The last identity is related to the Pythagorean theorem (see section Euclidean Geometry page 2609). Indeed, we see it better by rewriting:

$$\|\vec{x}\times\vec{y}\|^2+(\vec{x}\circ\vec{y})^2=\|\vec{x}\|^2\|\vec{y}\|^2 \tag{6.5.152}$$

This can be seen from the definitions of the cross product and dot product, as:

$$\vec{x}\times\vec{y}=ab\vec{n}\sin(\theta) \tag{6.5.153}$$

$$\vec{x}\circ\vec{y}=xy\cos(\theta) \tag{6.5.154}$$

6.5.4.3 Mixed Product (triple product)

We can extend the definition of the cross product to another type of mathematical tool we name the "mixed product".

Definition (#6.5.246): We name "mixed product", also sometimes named "triple scalar product", of vectors $\vec{x},\vec{y},\vec{z}$ the double product:

$$\vec{x}\circ(\vec{y}\times\vec{z}) \tag{6.5.155}$$

often condensed under the following notation:

$$[\vec{x},\vec{y},\vec{z}] \tag{6.5.156}$$

From what we saw in the definition of the dot and cross product, mixed product can also be written:

$$\boxed{\vec{x}\circ(\vec{y}\times\vec{z})=[\vec{x},\vec{y},\vec{z}]=\|\vec{x}\|\cdot\|\vec{y}\times\vec{z}\|\cos(\alpha)=\|\vec{x}\|\cdot\|\vec{y}\|\cdot\|\vec{z}\|\sin(\theta)\cos(\alpha)} \tag{6.5.157}$$

We note that in the case where E is the Euclidean vector space $\mathbb{R}^3$, the absolute value of the mixed product symbolize the oriented volume of the parallelepiped, built on the representatives $\vec{x},\vec{y},\vec{z}$ of common origin:

$$\vec{x}\circ(\vec{y}\times\vec{z})=\underbrace{\|\vec{y}\times\vec{z}\|}_{\text{basis}}\cdot\underbrace{\|\vec{x}\|\cos(\alpha)}_{\text{height}}=V \tag{6.5.158}$$

It is quite trivial that the mixed product is an extension to the three-dimensional case of the cross product. Indeed, in the expression of the mixed product, the cross product is the base surface of the parallelepiped and the scalar product project the vectors on the resulting vector from the cross product which gives the height h of the parallelepiped.

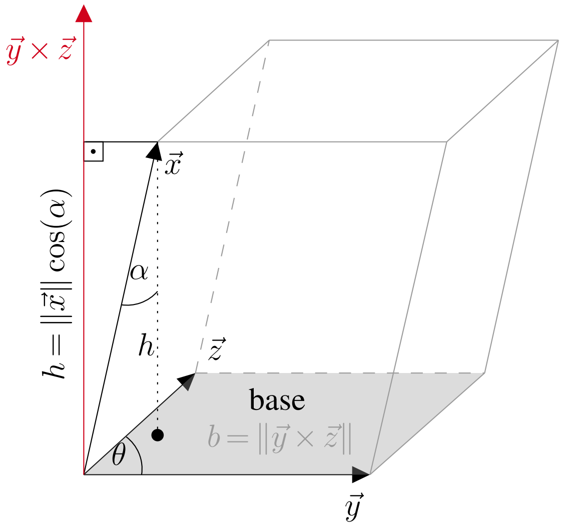


Figure 6.111: **Mixed product illustration**

If we develop:

$$[\vec{x},\vec{y},\vec{z}]=\vec{x}\begin{pmatrix}y_2z_3-y_3z_2\\y_3z_1-y_1z_3\\y_1z_2-y_2z_1\end{pmatrix}=x_1(y_2z_3-y_3z_2)+x_2(y_3z_1-y_1z_3)+x_3(y_1z_2-y_2z_1) \tag{6.5.159}$$

so triple product can also be understood as the determinant of a 3×3 matrix (thus also its inverse) having the three vectors either as its rows or its columns (see section Linear Algebra page 1967):

$$[\vec{x},\vec{y},\vec{z}]=\begin{vmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{vmatrix} \qquad (6.5.160)$$

By the commutative properties of the scalar product, we have:

$$(\vec{x}\times\vec{y})\circ\vec{z}=\vec{z}\circ(\vec{x}\times\vec{y}) \qquad (6.5.161)$$

and the reader will check without any trouble (we can write the details on request) that by developing components we have:

$$(\vec{x}\times\vec{y})\circ\vec{z}=\vec{x}\circ(\vec{y}\times\vec{z}) \qquad (6.5.162)$$

The triple product has also the following properties - that we don't use actually anywhere in this textbook - that the reader could be able to check easily by developing just the components of each expression excepted perhaps the third one (we can detailed as always on request if needed):

P1. $[\vec{x},\vec{y},\vec{z}]=[\vec{z},\vec{x},\vec{y}]=[\vec{y},\vec{z},\vec{x}]=-[\vec{y},\vec{x},\vec{z}]=-[\vec{z},\vec{y},\vec{x}]=-[\vec{x},\vec{z},\vec{y}]$

P2. $[\alpha\vec{x}+\beta\vec{y},\vec{z},\vec{v}]=\alpha[\vec{x},\vec{z},\vec{v}]+\beta[\vec{y},\vec{z},\vec{v}]$

P3. $[\vec{x},\vec{y},\vec{z}]\neq 0$ if and only if $\vec{x},\vec{y},\vec{z}$ are independent.

P4. $(\vec{x}\times\vec{y})\times(\vec{z}\times\vec{v})=[\vec{x},\vec{y},\vec{v}]\cdot\vec{z}-[\vec{x},\vec{y},\vec{z}]\cdot\vec{v}$

The first property in named "property of cyclicity" and the second one obviously "property of linearity" of the mixed product. We don't know if the two last properties have a name...

Obviously the third property is quite easy to interpret: The mixed product is zero if any two of the vectors, $\vec{x},\vec{y}$ and $\vec{z}$ are parallel, or if $\vec{x},\vec{y}$ and $\vec{z}$ are coplanar. That is, when the given three vectors are coplanar the altitude of the parallelepiped is zero and thus the mixed product is such that $V=0$.

Remark

We will come back on the triple product during our study on tensor calculus as it gives the opportunity to get a very interesting result concerning a future application in General Relativity.

6.5.5 Vectorial Functional Space

Given $\mathcal{C}^k_{[a,b]}$ the set of real functions that can be k-times derivates (see section Differential and Integral Calculus page 1496) in the closed bounded interval $[a,b]$. We will designate the elements of this set by the letters $\vec{f},\vec{g},\ldots$.

The value of $\vec{f}$ at the point t will be obviously denoted by $\vec{f}(t)$. Say that $\vec{f}=\vec{g}$ is therefore equivalent than to say that:

$$\vec{f}=\vec{g}, \quad \forall t\in[a,b] \tag{6.5.163}$$

In a condensed way some practitioners denote this $\vec{f}(t)\equiv\vec{g}(t)$, the symbol $\equiv$ indicating obviously that the two members are equals for any t in the bounded interval $[a,b]$.

Consider the two following operations:

- $\vec{f}(t)+\vec{g}(t)$ defined by the relation $(\vec{f}+\vec{g})(t)\equiv\vec{f}(t)+\vec{g}(t)$
- $\alpha\vec{f}$ defined by the relation $(\alpha\vec{f})(t)=\alpha\vec{f}(t)$

These two operations satisfy to all conditions of the vectors of a vector space as we have already defined at the beginning of this section (associativity, commutativity, null vector, opposed vector, distributivity, neutral element) and therefore gives us the possibility to assign to $\mathcal{C}^k_{[a,b]}$ of a vector space structure! Le null vector of this space being obviously the null function (equal to zero everywhere) and the opposite of $\vec{f}$ being $-\vec{f}$.

It is interesting to notice that $\mathcal{C}^k_{[a,b]}$ as a vector space is a generalization of $\mathbb{R}^n$ to the continuous case. Indeed, we can consider any vector $\vec{v}=(a_i)$ of $\mathbb{R}^n$ in the form of a real function defined on the set $\{1,2,\ldots,n\}$: the value of this function at the point i is simply a_i.

Remark

The polynomials of order n with one unknown form as an example of functional vector space of dimension $n+1$ such that for each coefficient of the polynomial corresponds a vector component such that:

$$P(x)=a_nx^n+a_{n-1}x^{n-1}+\ldots+a_1x+a_0\mapsto\begin{pmatrix}a_n\\a_{n-1}\\\vdots\\a_1\\a_0\end{pmatrix} \tag{6.5.164}$$

The preferred application field of the abstract theory of the dot product (inner product) is formed by the functional vector spaces. We name therefore "canonical scalar product" in $\mathcal{C}_{[a,b]}(\mathbb{R}^2)$ the operation defined by the relation:

$$\vec{f}\circ\vec{g}=\int\limits_a^b\vec{f}(t)\vec{g}(t)\mathrm{d}t \tag{6.5.165}$$

This operation defines indeed a scalar product, the properties of the latter being verified (on reader request we can add the proof if necessary), and furthermore, the integral:

$$\vec{f}\circ\vec{f}=\|\vec{f}\|^2=\int\limits_a^b \vec{f}^2(t)\mathrm{d}t \tag{6.5.166}$$

is positive if the continuous function $\vec{f}$ is not identically zero.

Technically the latter relation is written when in $\mathbb{R}^2$:

$$\|f\|_{L^2}=\sqrt{\vec{f}\circ\vec{f}}=\int\limits_a^b |f|^2(t)\mathrm{d}t \tag{6.5.167}$$

We will give more precision about this norm and its associated scalar product and with example in the section of Functional Analysis.

6.5.6 Hermitian Vector Space

The purpose of what will follow is, as always in this textbook, not to give a detailed study about vector spaces in $\mathbb{C}$ but just to give the minimal knowledge and vocabulary necessary to the lecture of some theoretical models in physics and especially those presented in this textbook in the section of Wave Quantum Physics.

When the scalars that appears in the definition of vector spaces are complex numbers (in $\mathbb{C}$ as seen in the section Numbers at page 243), and not only real numbers, then we speak obviously about "complex vectorial spaces".

Remark

Rigorously in the common communication, people should always precise if we speak of real vectorial space or complex vectorial space...

Let us give some examples of famous complex vectorial spaces (as many people think they are useless):

✍**Examples:**

E1. The space vector $\mathbb{C}^n$ of column-vectors with n components ($\mathbb{C}^1$ being obviously identified to $\mathbb{C}$). We will meet, among others, such vector space in the section of Relativistic Quantum Physics.

E2. The vectorial space of univariate polynomial with coefficient in $\mathbb{C}$. We will meet such spaces in the section of Wave Quantum Physics or even Quantum Chemistry.

E3. The vectorial space of complex functions of one real or complex variables continuous or not. We will meet such vector spaces frequently in the section of Wave Mechanics and especially in the section of Electrodynamics.

The purpose here is to adapt what we have seen so far to complex vectorial space. The following example, show us that we can transpose as it the previous definitions. Indeed, let us consider the vector space $\mathbb{C}^n$. As for $\mathbb{R}^n$, we could have the tentation to define a dot product (i.e. inner product) on $\mathbb{C}^n$ by:

$$\forall \vec{x},\vec{y}\in\mathbb{C}^n, \vec{x}=\begin{pmatrix}x_1\\ \vdots\\ x_n\end{pmatrix}, \vec{y}=\begin{pmatrix}y_1\\ \vdots\\ y_n\end{pmatrix} : \quad \langle\vec{x}|\vec{y}\rangle = x_1y_1+\ldots+x_ny_n \tag{6.5.168}$$

with $x_i,y_i\in\mathbb{C}$.

Sadly, we see that this definition is not satisfactory because we could have therefore:

$$\langle\vec{x}|\vec{x}\rangle = x_1^2+\ldots+x_n^2 \tag{6.5.169}$$

and this quantity is in general not a real number in a complex vector space (take for example components complex exponentials) and violates the property of positivity of the dot product and therefore prevent us to introduce and use the concept of distance. What is obviously a big problem in our actual perception of the world.

We could therefore not define anymore a norm in $\mathbb{C}^n$ by writing:

$$\|\vec{x}\|=\sqrt{\langle\vec{x}|\vec{x}\rangle}=\sqrt{x_1^2+\ldots+x_n^2} \tag{6.5.170}$$

For $\langle\vec{x}|\vec{x}\rangle$ to be a positive real number we see that it would be better to define the scalar product like this:

$$\boxed{\forall\vec{x},\vec{y}\in\mathbb{C}^n:\quad\langle\vec{x}|\vec{y}\rangle=\vec{x}\circ\bar{\vec{y}}=x_1\bar{y}_1+\ldots+x_n\bar{y}_n} \tag{6.5.171}$$

In this case we have therefore:

$$\langle\vec{x}|\vec{x}\rangle=x_1\bar{x}_1+\ldots+x_n\bar{x}_n=|x_1|^2+\ldots+|x_n|^2 \tag{6.5.172}$$

which is well a positive real number. From there, we can once again define a norm for complex vector space $\mathbb{C}^n$ by putting:

$$\|\vec{x}\|=\sqrt{\langle\vec{x}|\vec{x}\rangle}=\sqrt{|x_1|^2+\ldots+|x_n|^2} \tag{6.5.173}$$

We will now show how to define an inner product on a complex vector space in the general case.

6.5.6.1 Hermitian Inner Product

Definition (#6.5.247): Let $\mathcal{H}$ be a complex vector space (!). We name "scalar product" or more accurately "Hermitian inner product" on $\mathcal{H}$ (that is to say: a dot product in complex space...), an application:

$$\phi:\begin{cases}\mathcal{H}\times\mathcal{H}\to\mathbb{C}\\(\vec{x},\vec{y})\quad\mapsto\langle\vec{x}|\vec{y}\rangle\end{cases} \tag{6.5.174}$$

That satisfies (they are more properties but we will focus only about what we need for practical application in this textbook and especially quantum physics):

P1. Positivity:

$$\forall\vec{x}\in\mathcal{H}:\quad\phi(\vec{x},\vec{x})=\langle\vec{x}|\vec{x}\rangle\geq 0\in\mathbb{R}^+ \tag{6.5.175}$$

P2. Nullity:

$$\forall\vec{x}\in\mathcal{H}:\quad\phi(\vec{x},\vec{x})=\langle\vec{x}|\vec{x}\rangle=0\Leftrightarrow\vec{x}=0 \tag{6.5.176}$$

P3. Hermitian symmetry:

$$\forall(\vec{x},\vec{y})\in\mathcal{H}^2:\quad\phi(\vec{x},\vec{y})=\overline{\phi(\vec{y},\vec{x})}\Leftrightarrow\langle\vec{x}|\vec{y}\rangle=\overline{\langle\vec{y}|\vec{x}\rangle} \tag{6.5.177}$$

P4. Bilinearity (bilinear form) changes a little bit too ... so that we speak now of "sesquilinearity". We speak then, in order, of left anti-linearity and of right linearity such as:

$$\forall(\vec{x},\vec{y},\vec{z})\in\mathcal{H}^3,\forall(\lambda,\mu)\in\mathbb{C}^2\begin{cases}\langle\lambda\vec{x}+\mu\vec{y}|\vec{z}\rangle=\bar{\lambda}\langle\vec{x}|\vec{z}\rangle+\bar{\mu}\langle\vec{y}|\vec{z}\rangle\\\langle\vec{x}|\lambda\vec{y}+\mu\vec{z}\rangle=\lambda\langle\vec{x}|\vec{y}\rangle+\mu\langle\vec{x}|\vec{z}\rangle\end{cases} \tag{6.5.178}$$

Remarks

R1. It seems that some mathematicians put the anti-linearity on the right. It's just a matter of agreement that does not matter and exists because of a lack of international norms in mathematics.

R2. The reader may notice easily that if the elements of the above definitions are all in the set $\mathbb{R}$ then the sesquilinearity is reduced to the bilinearity and the hermitian character to a simple symmetry. So the hermitian inner product reduces to the scalar product.

R3. We want to give for now only the minimum on the vast subject that is complex vector spaces so that the reader can read without too much trouble the beginning of the section of Wave Quantum Physics.

When we join to a complex vector space a scalar product then just as a real vector space becomes an Euclidean vector space or Prehilbertien vector space, the complex vector space becomes what we name a "Hermitian vector space" (term often used in the section of Wave Quantum Physics).

Definition (#6.5.248): Again, we say that a space $\mathcal{H}$ provided with an Hermitian product $\langle\vec{x}|\lambda\vec{y}+\mu\vec{z}\rangle$ is a "Hilbert space" if this space is complete for the metric defined above.

Thus, Hilbert spaces is a generalization of spaces including dot products and Hermitian dot product of Euclidean and Prehilbertien spaces.

6.5.6.2 Hermitian Cauchy-Schwarz inequality

Let us now consider the generalization of the Cauchy-Schwarz inequality that we have already proved only for the special case of real vectors at page 1826.

Consider $\vec{x}$ and $\vec{y}$ two complex vectors. Then for $\vec{y}\neq\vec{0}$ and for any complex number λ, we have by definition of the hermitian inner product (remember that $z\bar{z}=|z|^2$):

$$0\leq\langle\vec{x}-\lambda\vec{y},\vec{x}-\lambda\vec{y}\rangle=\langle\vec{x},\vec{x}\rangle-\lambda\langle\vec{y},\vec{x}\rangle-\bar{\lambda}\langle\vec{x},\vec{y}\rangle+|\lambda|^2\langle\vec{y},\vec{y}\rangle \tag{6.5.179}$$

Now let us take:

$$\lambda=\frac{\langle\vec{x},\vec{y}\rangle}{\langle\vec{y},\vec{y}\rangle} \tag{6.5.180}$$

Then the hermitian Cauchy-Schwarz inequality easily follows from:

$$0\leq\langle\vec{x},\vec{x}\rangle-\frac{\langle\vec{x},\vec{y}\rangle}{\langle\vec{y},\vec{y}\rangle}\langle\vec{y},\vec{x}\rangle-\frac{\langle\vec{y},\vec{x}\rangle}{\langle\vec{y},\vec{y}\rangle}\langle\vec{x},\vec{y}\rangle+\left|\frac{\langle\vec{x},\vec{y}\rangle}{\langle\vec{y},\vec{y}\rangle}\right|^2\langle\vec{y},\vec{y}\rangle \tag{6.5.181}$$

as simplified it gives:

$$0\leq\langle\vec{x},\vec{x}\rangle-\frac{|\langle\vec{x},\vec{y}\rangle|^2}{\langle\vec{y},\vec{y}\rangle} \tag{6.5.182}$$

Hence rearranged:

$$\boxed{|\langle\vec{x},\vec{y}\rangle|^2\leq\langle\vec{x},\vec{y}\rangle\langle\vec{y},\vec{y}\rangle} \tag{6.5.183}$$

Equality is when $\vec{y}=\lambda\vec{x}$ (λ complex). Then trivially:

$$|\langle\vec{x},\vec{y}\rangle|^2=|\lambda|^2\langle\vec{x},\vec{x}\rangle=\langle\vec{x},\vec{x}\rangle\langle\lambda\vec{x},\lambda\vec{x}\rangle=\langle\vec{x},\vec{x}\rangle\langle\vec{y},\vec{y}\rangle \tag{6.5.184}$$

This complex version of the Cauchy-Schwarz inequality will be useful to us to derive the complex version of the Cauchy-Schwarz inequality for integrals (see page 2388) that has an important place in Complex Analysis and therefore in Quantum Physics!

6.5.6.3 Types of Vectors Spaces

To sum it all up:

- We name "pre-Hilbert space" (real or complex) any vector space of finite dimension or not, provided with a dot (scalar) product.
- We name "Hilbert space" (real or complex) any complete prehilbertian space (as space provide with a norm).
- We name "Euclidean space" any real vector space of finite dimension with a dot (scalar) product and denoted by $\mathcal{E}^n$.
- We name Hermitian space any complex vector space of finite dimension with a dot (scalar) product and denote it by $\mathcal{H}^n$.

6.5.7 System of Coordinates

We will address here the aspect of coordinates changes of vector components not from a basis to another one (for that you need to go see the section of Linear Algebra) but from one coordinate system to another. That means that in any case we will stay in an Euclidean space. This type of transformation has strong implication in physical (and a little bit less in pure mathematics) when we want to simplify the study of physical systems whose equations become easier to handle in other coordinate systems.

Definition (#6.5.249): In mathematics, a "coordinate system" is used to match to each point of an n-dimensional space, a m-tuple of scalars.

> **Remark**
>
> In many cases, the considered scalars are real numbers, but it is possible to use complex numbers or elements of any field (see section Set Theory page 429). More generally, coordinates can come from a ring or other related algebraic structure.

Although we are in a chapter and in a section of this textbook that is supposed to be purely maths oriented..., we will allow ourselves in what follows to make a direct connection with physics relatively the terms of the speed and acceleration in different coordinate systems (sorry for the "math skills only" people...). Indeed, our teaching experience has shown us that this helps the readers (most of time students) to better understand the various abstract concepts.

6.5.7.1 Cartesian (rectangular) Coordinate System

We do not want to take too much time on this system as it is well known to everyone usually. However, let us recall that most of the time, in physics, the Cartesian systems in which we are working are in $\mathbb{R}^2$(two real spatial dimensions), or $\mathbb{R}^3$ (three real spatial dimensions) or even $\mathbb{R}^4$ or $\mathbb{C}^4$ (three spatial dimensions and one of time) when we work in relativity. The number of dimensions can be higher as for example with the Kaluza-Klein theory (five dimensions) merging General Relativity and Electromagnetism or much more with String Theory (above 20 dimensions!!!).

In $\mathbb{R}^3$ (the most common case in undergraduate physics), there are three basis vectors traditionally denoted by:

$$\vec{e}_1, \vec{e}_2, \vec{e}_3 \tag{6.5.185}$$

Or more explicitly:

$$\boxed{\vec{e}_x = \begin{pmatrix}1\\0\\0\end{pmatrix}, \vec{e}_y = \begin{pmatrix}0\\1\\0\end{pmatrix}, \vec{e}_z = \begin{pmatrix}0\\0\\1\end{pmatrix}} \tag{6.5.186}$$

In this system, the position of a point P (identifiable by a vector $\vec{x}$ for example) is defined by the three numbers named "coordinates" (more generally "components") denoted (typically in Tensor Calculus):

$$(x_1, x_2, x_3) \tag{6.5.187}$$

and in physics denoted more conventionally by:

$$(x,y,z) \tag{6.5.188}$$

where usually the component (z) represents the height (vertical), the component (x) is the width and the component (y) is the length (obviously these are completely arbitrary choice).

This point P can be spotted by a vector arbitrarily designated $\vec{r}$ in the basis $\vec{e}_i$ by the relation (using Tensor notation):

$$\vec{x}=x^i\vec{e}_i \tag{6.5.189}$$

and if the basis is canonical (orthonormal) such that:

$$(\vec{e}_1,\vec{e}_2,\vec{e}_3)=\left\{\begin{pmatrix}1\\0\\0\end{pmatrix},\begin{pmatrix}0\\1\\0\end{pmatrix},\begin{pmatrix}0\\0\\1\end{pmatrix}\right\} \tag{6.5.190}$$

we write:

$$\vec{r}=x^i\vec{e}_i=x_1\begin{pmatrix}1\\0\\0\end{pmatrix}+x_2\begin{pmatrix}0\\1\\0\end{pmatrix}+x_3\begin{pmatrix}0\\0\\1\end{pmatrix}=\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}:=\begin{pmatrix}x\\y\\z\end{pmatrix} \tag{6.5.191}$$

In physics, if we work with coordinates, it's in the purpose to always be able to determine the location of an item as depicted in the figure below:

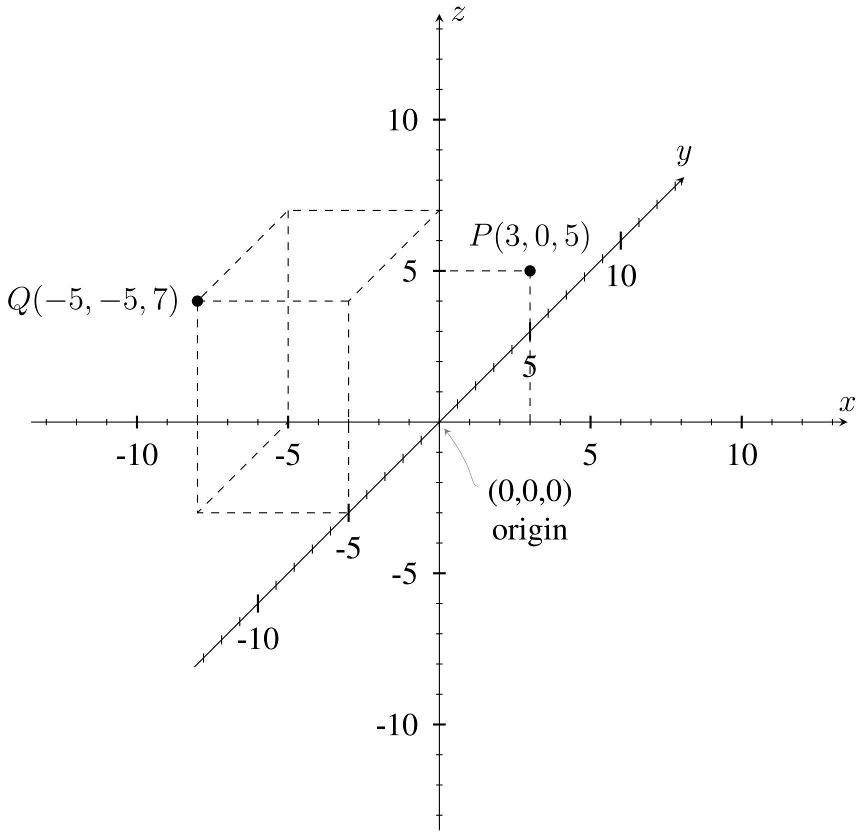


Figure 6.112: **Localisation of a point in a simple 3D coordinate system**

Or, as we shall see it more rigorously in the section of Analytical Mechanics, the physicist works with the following concepts (each element being often time-dependent):

- Position: $\vec{r}=\Big(x(t),y(t),z(t),t\Big)$
- Velocity: $\dfrac{\mathrm{d}\vec{r}}{\mathrm{d}t}=\dot{\vec{r}}=\vec{v}=\Big(\dot{x}(t),\dot{y}(t),\dot{z}(t),t\Big)$
- Acceleration: $\dfrac{\mathrm{d}\vec{v}}{\mathrm{d}t}=\dot{\vec{v}}=\vec{a}=\Big(\ddot{x}(t),\ddot{y}(t),\ddot{z}(t),t\Big)$

Now let us see how the different concepts are expressed in systems such as spherical, cylindrical and polar coordinates (remember that we remains for all of them in a flat Euclidean space!!!).

6.5.7.2 Spherical Coordinate System

The choice to start with this coordinate system is not a coincidence. It has the advantage of being a generalization of cylindrical and polar systems that we will meet thereafter and will help us easier to determine the expressions of position, velocity and acceleration[15].

We traditionally represent (in Switzerland ... and in agreement with the standard ISO 31-11) a spherical coordinate system as follows:

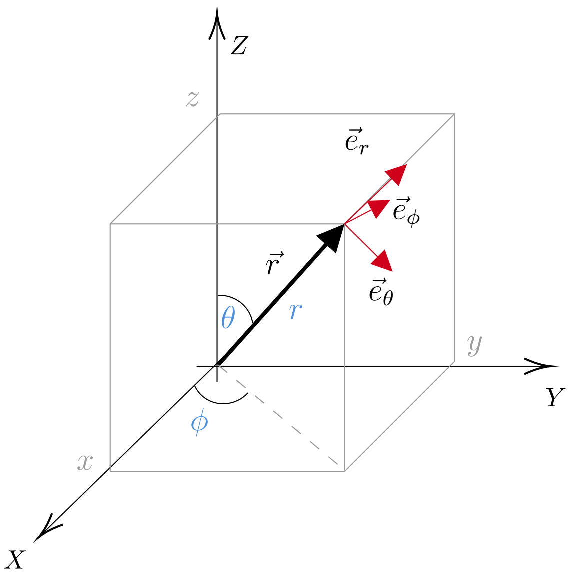


Figure 6.113: **Representation of the spherical coordinate system**

We see very clearly if we know the basic trigonometric relations and identities (see the section of the same name in the Geometry chapter) we have the transformations:

$$\begin{aligned}
&r=\sqrt{x^2+y^2+z^2}, \quad 0\leq r<+\infty\\
&\theta=\arccos\left(\frac{z}{r}\right)=\arccos\left(\frac{z}{\sqrt{x^2+y^2+z^2}}\right), \quad 0\leq\theta<\pi, r\neq 0\\
&\phi=\arccos\left(\frac{x}{\sqrt{x^2+y^2}}\right)=\arccos\left(\frac{x}{r\sin\theta}\right)=\arcsin\left(\frac{y}{r\sin\theta}\right), \quad 0\leq\phi<2\pi, r\neq 0
\end{aligned} \tag{6.5.192}$$

where the two angles θ, ϕ are respectively the colatitude ($\pi/2-\theta$ is the latitude) and longitude:

---

[15] Two commonly-used sets of "orthogonal curvilinear coordinates" are cylindrical polar coordinates and spherical polar coordinates.

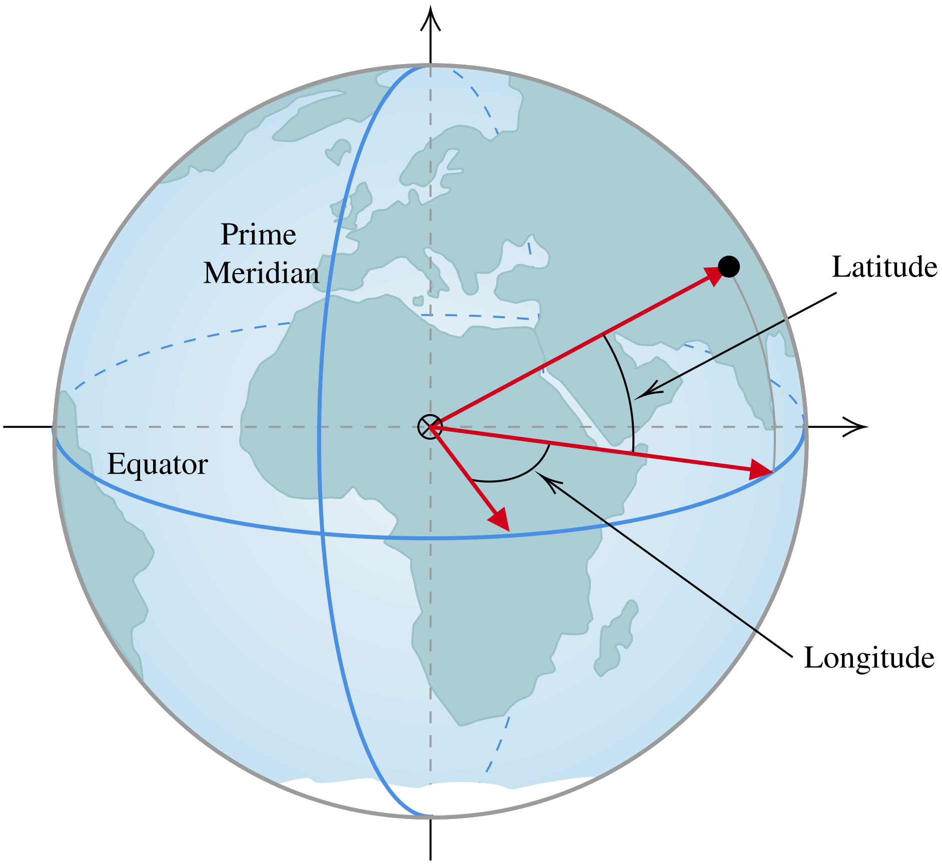


Figure 6.114: **Latitude and Longitude concepts illustrated**

We have inversely:

$$x = r\sin(\theta)\cos(\phi) \quad y = r\sin(\theta)\sin(\phi) \quad z = r\cos(\theta) \tag{6.5.193}$$

Caution! Depending on the angle convention in the prior-previous figure the reader can find different expression for x, y, z in other textbooks. A typical other convention is for example:

$$x = r\cos(\theta)\cos(\phi) \quad y = r\sin(\theta)\cos(\phi) \quad z = r\sin(\phi)$$

Now we must find the expressions that connect the vectors of the spherical basis that we choose to denote by $\vec{e}_r,\vec{e}_\theta,\vec{e}_\phi$ with the vectors of the Cartesian basis $\vec{e}_x,\vec{e}_y,\vec{e}_z$:

$$\begin{aligned}
\vec{e}_r =&\frac{\vec{r}}{r}=\frac{x\vec{e}_x+y\vec{e}_y+z\vec{e}_z}{r}=\frac{r\sin(\theta)\cos(\phi)\vec{e}_x+r\sin(\theta)\sin(\phi)\vec{e}_y+r\cos(\theta)\vec{e}_z}{r}\\
=&\sin(\theta)\cos(\phi)\vec{e}_x+\sin(\theta)\sin(\phi)\vec{e}_y+\cos(\theta)\vec{e}_z\\
\vec{e}_\phi =&\frac{\vec{e}_z\times\vec{e}_r}{\sin(\theta)}=\frac{1}{\sin(\theta)}\begin{pmatrix}0\\0\\1\end{pmatrix}\times\begin{pmatrix}\sin(\theta)\cos(\phi)\\ \sin(\theta)\sin(\phi)\\ \cos(\phi)\end{pmatrix}=\frac{1}{\sin(\theta)}\begin{pmatrix}-\sin(\theta)\sin(\phi)\\ \sin(\theta)\cos(\phi)\\ 0\end{pmatrix}\\
=&\frac{1}{\sin(\theta)}\begin{pmatrix}-\sin(\theta)\sin(\phi)\\0\\0\end{pmatrix}+\frac{1}{\sin(\theta)}\begin{pmatrix}0\\ \sin(\theta)\cos(\phi)\\0\end{pmatrix}=-\vec{e}_x\sin(\phi)+\vec{e}_y\cos(\phi)\\
\vec{e}_\theta =&\vec{e}_\phi\times\vec{e}_r=\begin{pmatrix}-\sin(\phi)\\ \cos(\phi)\\ 0\end{pmatrix}\times\begin{pmatrix}\sin(\theta)\cos(\phi)\\ \sin(\theta)\sin(\phi)\\ \cos(\theta)\end{pmatrix}=\begin{pmatrix}\cos(\phi)\cos(\theta)-0\cdot\sin(\theta)\sin(\phi)\\ 0\cdot\sin(\theta)\cos(\phi)-(-\sin(\phi))\cos(\theta)\\ -\sin(\phi)\sin(\theta)\sin(\phi)-\cos(\phi)\sin(\theta)\cos(\phi)\end{pmatrix}\\
=&\begin{pmatrix}\cos(\phi)\cos(\theta)\\ \sin(\phi)\cos(\theta)\\ -\sin(\phi)\sin(\theta)\sin(\phi)-\cos(\phi)\sin(\theta)\cos(\phi)\end{pmatrix}=\vec{e}_x\cos(\theta)\cos(\phi)+\vec{e}_y\cos(\theta)\sin(\phi)-\vec{e}_z\sin(\theta)
\end{aligned}\tag{6.5.194}$$

Let us indicate that by dividing by $\sin(\theta)$ the second basis vector $\vec{e}_\phi$, we make sure that by the properties of the norm of the cross product that:

$$\|\vec{e}_\phi\|=\|\vec{e}_z\|\cdot\|\vec{e}_r\|\sin(\phi)=1\tag{6.5.195}$$

will be well normalized to unity as expected (as we know from start that as we take a direct orthogonal coordinate system the product of norms of the basis vectors must be equal to 1)!

For later purposes, let us determine the partial differentials of each of these coordinates:

$$\begin{aligned}
&\frac{\partial\vec{e}_r}{\partial r}=\vec{0}\\
&\frac{\partial\vec{e}_r}{\partial\theta}=\vec{e}_x\cos(\theta)\cos(\phi)+\vec{e}_y\cos(\theta)\sin(\phi)-\vec{e}_z\sin(\theta)=\vec{e}_\theta\\
&\frac{\partial\vec{e}_r}{\partial\phi}=-\vec{e}_x\sin(\theta)\sin(\phi)+\vec{e}_y\sin(\theta)\cos(\phi)=(-\vec{e}_x\sin(\theta)+\vec{e}_y\cos(\phi))\sin(\theta)=\vec{e}_\phi\sin(\theta)\\
&\frac{\partial\vec{e}_\phi}{\partial r}=\vec{0}\\
&\frac{\partial\vec{e}_\phi}{\partial\theta}=\vec{0}\\
&\frac{\partial\vec{e}_\phi}{\partial\phi}=-\vec{e}_x\cos(\phi)-\vec{e}_y\sin(\phi)=-(\vec{e}_r\sin(\theta)+\vec{e}_\theta\cos(\theta))\\
&\frac{\partial\vec{e}_\theta}{\partial r}=\vec{0}\\
&\frac{\partial\vec{e}_\theta}{\partial\theta}=-\vec{e}_x\sin(\theta)\cos(\phi)-\vec{e}_y\sin(\theta)\sin(\phi)-\vec{e}_z\cos(\theta)=-\vec{e}_r\\
&\frac{\partial\vec{e}_\theta}{\partial\phi}=-\vec{e}_x\cos(\theta)\sin(\phi)+\vec{e}_y\cos(\theta)\cos(\phi)=\vec{e}_\phi\cos(\theta)
\end{aligned}\tag{6.5.196}$$

We will also use later (for the study of vector operators further below and the geodesic of the sphere in the section of Analytical Mechanics) the variation $\mathrm{d}\vec{r}$ expressed in spherical coordinates:

$$\begin{aligned}\mathrm{d}\vec{r}&=\mathrm{d}(r\vec{e}_r)=\vec{e}_r\mathrm{d}r+r\mathrm{d}\vec{e}_r=\vec{e}_r\mathrm{d}r+r\left(\frac{\partial\vec{e}_r}{\partial r}\mathrm{d}r+\frac{\partial\vec{e}_r}{\partial\theta}\mathrm{d}\theta+\frac{\partial\vec{e}_r}{\partial\phi}\mathrm{d}\phi\right)\\&=\vec{e}_r\mathrm{d}r+\vec{e}_\theta r\mathrm{d}\theta+\vec{e}_\phi r\sin(\theta)\mathrm{d}\phi\end{aligned}\qquad(6.5.197)$$

Or more explicitly:

$$\boxed{\vec{e}_r=\begin{pmatrix}\sin(\theta)\cos(\phi)\\\sin(\theta)\sin(\phi)\\\cos(\theta)\end{pmatrix},\vec{e}_\phi=\begin{pmatrix}-\sin(\phi)\\\cos(\phi)\\0\end{pmatrix},\vec{e}_\theta=\begin{pmatrix}\cos(\theta)\cos(\phi)\\\cos(\theta)\sin(\phi)\\-\sin(\theta)\end{pmatrix}}\qquad(6.5.198)$$

All this can be put in a famous matrix form as (this also applies automatically to the component of the unitary vector):

$$\begin{pmatrix}\vec{e}_r\\\vec{e}_\theta\\\vec{e}_\phi\end{pmatrix}=\begin{bmatrix}\sin(\theta)\cos(\phi)&\sin(\theta)\sin(\phi)&\cos(\theta)\\\cos(\theta)\cos(\phi)&\cos(\theta)\sin(\phi)&-\sin(\theta)\\-\sin(\phi)&\cos(\phi)&0\end{bmatrix}\begin{pmatrix}\vec{e}_x\\\vec{e}_y\\\vec{e}_z\end{pmatrix}\qquad(6.5.199)$$

To express the velocity and acceleration in spherical coordinates, we will also need the derivatives with respect to time:

$$\begin{aligned}\dot{\vec{e}}_r&=\frac{\partial\vec{e}_r}{\partial r}\dot{r}+\frac{\partial\vec{e}_r}{\partial\theta}\dot{\theta}+\frac{\partial\vec{e}_r}{\partial\phi}\dot{\phi}=\vec{e}_\theta\dot{\theta}+\vec{e}_\phi\dot{\phi}\sin(\theta)\\\dot{\vec{e}}_\theta&=\frac{\partial\vec{e}_\theta}{\partial r}\dot{r}+\frac{\partial\vec{e}_\theta}{\partial\theta}\dot{\theta}+\frac{\partial\vec{e}_\theta}{\partial\phi}\dot{\phi}=-\vec{e}_r\dot{\theta}+\vec{e}_\phi\dot{\phi}\cos(\theta)\\\dot{\vec{e}}_\phi&=\frac{\partial\vec{e}_\phi}{\partial r}\dot{r}+\frac{\partial\vec{e}_\phi}{\partial\theta}\dot{\theta}+\frac{\partial\vec{e}_\phi}{\partial\phi}\dot{\phi}=-(\vec{e}_r\sin(\dot{\theta})+\vec{e}_\theta\cos(\dot{\theta}))\dot{\phi}\end{aligned}\qquad(6.5.200)$$

So if we do now a little bit of physics, we have:

$$\begin{aligned}\vec{v}=\dot{\vec{r}}&=\begin{pmatrix}\mathrm{d}_t(r\sin(\theta)\cos(\phi))\\\mathrm{d}_t(r\sin(\theta)\sin(\phi))\\\mathrm{d}_t(r\cos(\theta))\end{pmatrix}=\begin{pmatrix}\dot{r}\sin(\theta)\cos(\phi)+r\cos(\theta)\cos(\phi)\dot{\theta}-r\sin(\theta)\sin(\phi)\dot{\phi}\\\dot{r}\sin(\theta)\sin(\phi)+r\cos(\theta)\sin(\phi)\dot{\theta}+r\sin(\theta)\cos(\phi)\dot{\phi}\\\dot{r}\cos(\theta)-r\sin(\theta)\dot{\theta}\end{pmatrix}\\&=\begin{pmatrix}\sin(\theta)\cos(\phi)\\\sin(\theta)\sin(\phi)\\\cos(\theta)\end{pmatrix}\dot{r}+\begin{pmatrix}r\cos(\theta)\cos(\phi)\\r\cos(\theta)\sin(\phi)\\-r\sin(\theta)\end{pmatrix}\dot{\theta}+\begin{pmatrix}-r\sin(\theta)\sin(\phi)\\r\sin(\theta)\cos(\phi)\\0\end{pmatrix}\dot{\phi}\\&=\vec{e}_r\dot{r}+r\begin{pmatrix}\cos(\theta)\cos(\phi)\\\cos(\theta)\sin(\phi)\\-\sin(\theta)\end{pmatrix}\dot{\theta}+r\sin(\theta)\begin{pmatrix}-\sin(\phi)\\\cos(\phi)\\0\end{pmatrix}\dot{\phi}=\vec{e}_r\dot{r}+r\vec{e}_\theta\dot{\theta}+r\sin(\theta)\vec{e}_\phi\dot{\phi}\end{aligned}\qquad(6.5.201)$$

This brings us to (we will need this relation mainly in the section of Astrophysics):

$$\boxed{\vec{v}=\dot{r}\vec{e}_r+\vec{e}_\theta r\dot{\theta}+\vec{e}_\phi r\dot{\phi}\sin(\theta)}\qquad(6.5.202)$$

It is interesting to notice that we get the same result through the following method that may be less intuitive:

$$\vec{v}=\dot{\vec{r}}=\dot{\vec{e}}r+\dot{r}\vec{e}_r\qquad(6.5.203)$$

and substituting the derivative obtained above:

$$\dot{\vec{e}}_r=\frac{\partial\vec{e}_r}{\partial r}\dot{r}+\frac{\partial\vec{e}_r}{\partial\theta}\dot{\theta}+\frac{\partial\vec{e}_r}{\partial\phi}\dot{\phi}=\vec{e}_\theta\dot{\phi}+\vec{e}_\phi\dot{\phi}\sin(\theta) \tag{6.5.204}$$

Concerning the acceleration we get:

$$\begin{aligned}
\vec{a}=\dot{\vec{v}}=\begin{pmatrix}\ddot{x}\\ \ddot{y}\\ \ddot{z}\end{pmatrix}&=\begin{pmatrix}\mathrm{d}_t\left(\dot{r}\sin(\theta)\cos(\phi)+r\cos(\theta)\cos(\phi)\dot{\theta}-r\sin(\theta)\sin(\phi)\dot{\phi}\right)\\ \mathrm{d}_t\left(\dot{r}sin(\theta)\sin(\phi)+r\cos(\theta)\sin(\phi)\dot{\theta}+r\sin(\theta)\cos(\phi)\dot{\theta}\right)\\ \mathrm{d}_t\left(\dot{r}\cos(\theta)-r\sin(\theta)\dot{\theta}\right)\end{pmatrix}\\
&=\begin{pmatrix}\left(\ddot{r}\sin(\theta)\cos(\phi)+\dot{r}\cos(\theta)\cos(\phi)\dot{\theta}-\dot{r}\sin(\theta)\sin(\phi)\dot{\phi}\right)+\\ \left(\dot{r}\cos(\theta)\cos(\phi)\dot{\theta}-r\sin(\theta)\cos(\phi)\dot{\theta}\dot{\theta}+r\cos(\theta)\cos(\phi)\ddot{\theta}-r\cos(\theta)\sin(\phi)\dot{\theta}\dot{\phi}\right)\\ -\left(\dot{r}\sin(\theta)\sin(\phi)\dot{\phi}+r\cos(\theta)\sin(\phi)\dot{\phi}\dot{\theta}+r\sin(\theta)\cos(\phi)\dot{\phi}\dot{\phi}+r\sin(\theta)\sin(\phi)\ddot{\phi}\right)\\ \\ \left(\ddot{r}\sin(\theta)\sin(\phi)+\dot{r}\cos(\theta)\sin(\phi)\dot{\theta}+\dot{r}\sin(\theta)\sin(\phi)\dot{\phi}\right)\\ +\left(\dot{r}\cos(\theta)\sin(\phi)\dot{\theta}-r\sin(\theta)\sin(\phi)\dot{\theta}\dot{\theta}+r\cos(\theta)\sin(\phi)\ddot{\theta}+r\cos(\theta)\cos(\phi)\dot{\theta}\dot{\phi}\right)\\ +\left(\dot{r}\sin(\theta)\cos(\phi)\dot{\phi}+r\cos(\theta)\cos(\phi)\dot{\phi}\dot{\theta}-r\sin(\theta)\sin(\phi)\dot{\phi}\dot{\phi}+r\sin(\theta)\sin(\phi)\ddot{\phi}\right)\\ \\ \left(\ddot{r}\cos(\theta)-\dot{r}\sin(\theta)\dot{\theta}\right)-\left(\dot{r}\sin(\theta)\dot{\theta}+r\cos(\theta)\dot{\theta}\dot{\theta}+r\sin(\theta)\dot{\theta}\right)\end{pmatrix}\\
&=\left(\ddot{r}-r\dot{\theta}\right)\begin{pmatrix}\sin(\theta)\cos(\phi)\\ \sin(\theta)\sin(\phi)\\ \cos(\theta)\end{pmatrix}+\left(2\dot{r}\dot{\theta}+r\ddot{\theta}\right)\begin{pmatrix}\cos(\theta)\cos(\phi)\\ \cos(\theta)\sin(\phi)\\ -\sin(\theta)\end{pmatrix}\\
&+\left(2\dot{r}\sin(\theta)\dot{\phi}+2r\cos(\theta)\dot{\theta}\dot{\phi}+r\sin(\theta)\ddot{\phi}\right)\begin{pmatrix}-\sin(\phi)\\ \cos(\phi)\\ 0\end{pmatrix}+r\sin(\theta)\dot{\phi}^2\begin{pmatrix}-\cos(\phi)\\ \sin(\phi)\\ 0\end{pmatrix}\\
&=\left(\ddot{r}-r\dot{\theta}\right)\vec{e}_r+\left(2\dot{r}\dot{\theta}+r\ddot{\theta}\right)\vec{e}_\theta+\left(2\dot{r}\sin(\theta)\dot{\phi}+2r\cos(\theta)\dot{\theta}\dot{\phi}+r\sin(\theta)\ddot{\phi}\right)\vec{e}_\phi-r\sin(\theta)\dot{\phi}^2\begin{pmatrix}\cos(\phi)\\ \sin(\phi)\\ 0\end{pmatrix}
\end{aligned} \tag{6.5.205}$$

But we have:

$$\begin{aligned}
\begin{pmatrix}\cos(\phi)\\ \sin(\phi)\\ 0\end{pmatrix}&=\begin{pmatrix}\cos(\phi)\sin^2(\theta)+\cos(\phi)\cos^2(\theta)\\ \sin(\phi)\sin^2(\theta)+\sin(\phi)\cos^2(\theta)\\ 0\end{pmatrix}\\
&=\sin(\theta)\begin{pmatrix}\sin(\theta)\cos(\phi)\\ \sin(\theta)\sin(\phi)\\ \cos(\theta)\end{pmatrix}+\cos(\theta)\begin{pmatrix}\cos(\theta)\cos(\phi)\\ \cos(\theta)\sin(\phi)\\ -\sin(\theta)\end{pmatrix}\\
&=\sin(\theta)\vec{e}_r+\cos(\theta)\vec{e}_\theta
\end{aligned} \tag{6.5.206}$$

Therefore it comes:

$$\begin{aligned}
\vec{a}=&\left(\ddot{r}-r\dot{\theta}^2\right)\vec{e}_r+\left(2\dot{r}\dot{\theta}+r\ddot{\theta}\right)\vec{e}_\theta+\left(2\dot{r}\sin(\theta)\dot{\phi}+2r\cos(\theta)\dot{\theta}\dot{\phi}+r\sin(\theta)\ddot{\phi}\right)\vec{e}_\phi\\
&-r\sin(\theta)\dot{\phi}^2\left(\sin(\theta)\vec{e}_r+\cos(\theta)\vec{e}_\theta\right)
\end{aligned} \tag{6.5.207}$$

Thus finally:

$$\boxed{\begin{aligned}\vec{a} =&\vec{e}_r(\ddot{r}-r\dot{\theta}^2-r\dot{\phi}^2\sin^2(\theta))+\vec{e}_\theta(r\ddot{\theta}+2\dot{r}\dot{\theta}-r\dot{\phi}^2\sin(\theta)\cos(\theta))\\ &+\vec{e}_\phi(r\ddot{\phi}\sin(\theta)+2r\dot{\theta}\dot{\phi}\cos(\theta)+2\dot{r}\dot{\phi}\sin(\theta))\end{aligned}} \tag{6.5.208}$$

6.5.7.2.1 Orthodromic distance

We will now see a powerful and simple application of vector calculus (especially of the dot product with spherical coordinates!). The reader will see later in the section of Analytical Mechanics that the geodesic of the sphere is the equation of the great circles (circles on the sphere whose centers coincide with the center of the sphere).

But without knowing this result, and even without needing it, we want to calculate the "great-circle distance" or "orthodromic distance" that is the shortest distance between two points on the surface of a sphere (to not confuse with the "loxodromic distance" that is the straight line that we draw on a Mercator Earth's projection plane! - a line of constant compass bearing - not the shortest path on a sphere), measured along the surface of the sphere:

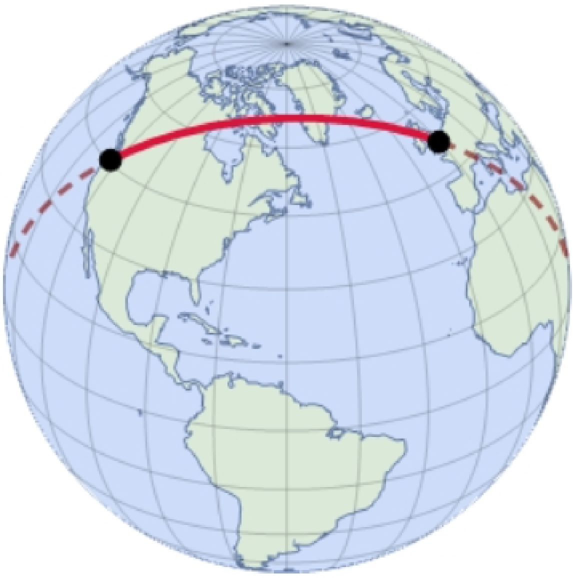


Figure 6.115: Orthodromic distance (great-circle distance)

Through any two points on a sphere that are not directly opposite each other, there is a unique great circle. The two points separate the great circle into two arcs. The length of the shorter arc is the great-circle distance between the points. A great circle endowed with such a distance is named a "Riemannian circle" in Riemannian geometry.

Let us consider two points on the surface of a sphere P_1 and P_2 and their respective latitude and longitude such that they are defined by:

$$P_1=(\text{latitude},\text{longitude})=(\theta_1,\phi_1) \qquad P_2=(\text{latitude},\text{longitude})=(\theta_2,\phi_2) \tag{6.5.209}$$

Therefore in spherical coordinates:

$$\begin{aligned} P_1 &= (R\sin(\theta_1)\cos(\phi_1), R\sin(\theta_1)\sin(\phi_1), R\cos(\theta_1)) \\ P_2 &= (R\sin(\theta_2)\cos(\phi_2), R\sin(\theta_2)\sin(\phi_2), R\cos(\theta_2)) \end{aligned} \quad (6.5.210)$$

Hence taking the vectors having for origin the center of the sphere:

$$\begin{aligned} \vec{P}_1 &= R\sin(\theta_1)\cos(\phi_1)\vec{i} + R\sin(\theta_1)\sin(\phi_1)\vec{j} + R\cos(\theta_1)\vec{k} \\ \vec{P}_2 &= R\sin(\theta_2)\cos(\phi_2)\vec{i} + R\sin(\theta_2)\sin(\phi_2)\vec{j} + R\cos(\theta_2)\vec{k} \end{aligned} \quad (6.5.211)$$

The Earth is nearly spherical, so great-circle distance formulas give the distance between points on the surface of the Earth (as the crow flies) correct to within 0.5% or so.

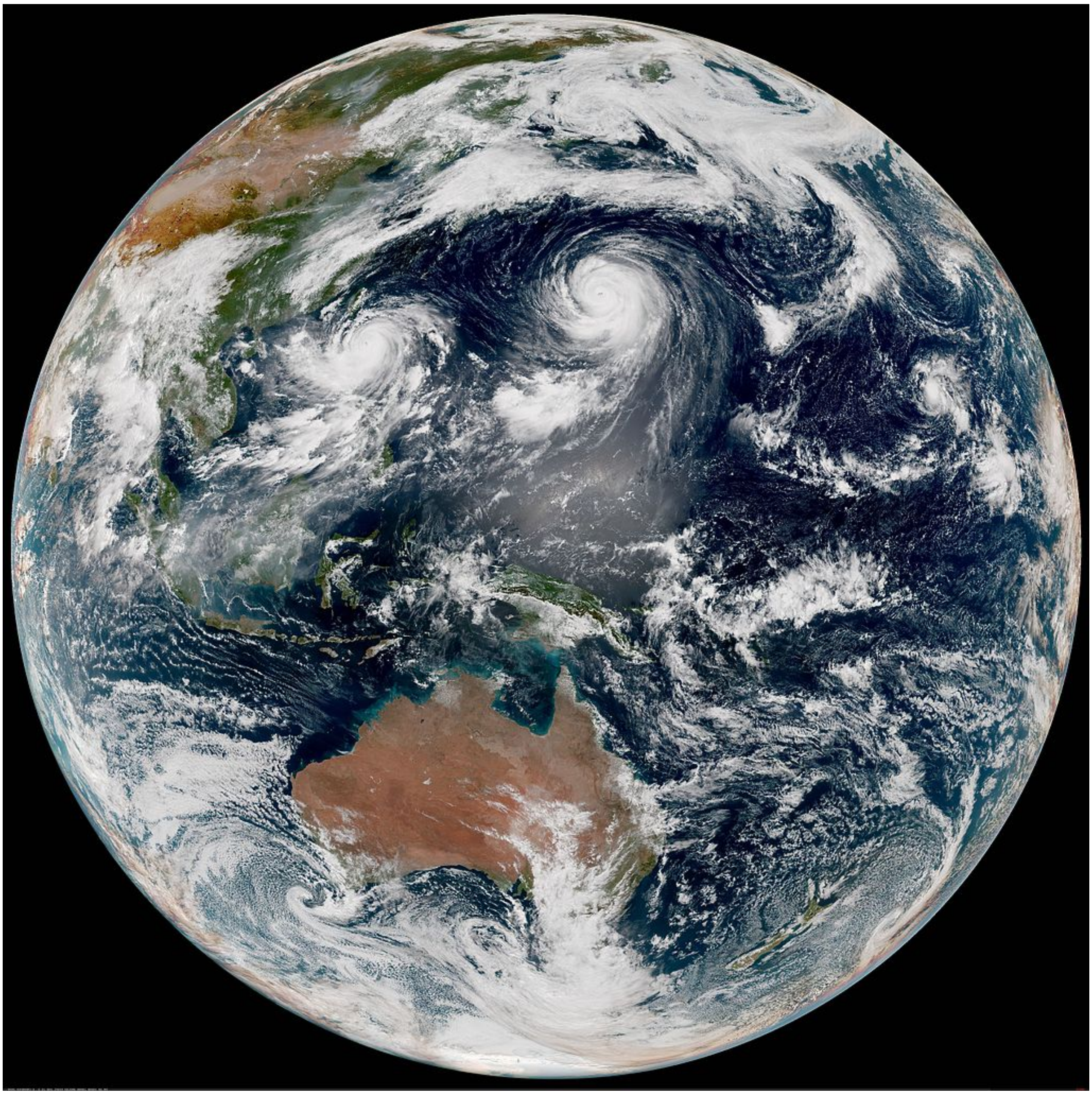


Figure 6.116: Japanese Himawari 8 satellite's Earth, Rayleigh-corrected, true-color image captured from the AHI sensor (year 12015 according to holocene calendar)

We do now the dot product and use trigonometric identity:

$$\begin{aligned}\vec{P}_1 \circ \vec{P}_2 &= RR\left(\sin(\theta_1)\sin(\theta_2)\cos(\phi_1)\cos(\phi_2)+\sin(\theta_1)\sin(\theta_2)\sin(\phi_1)\sin(\phi_2)+\cos(\theta_1)\cos(\theta_2)\right)\\ &= R^2\left[\sin(\theta_1)\sin(\theta_2)\left(\cos(\phi_1)\cos(\phi_2)+\sin(\theta_1)\sin(\theta_2)\right)+\cos(\theta_1)\cos(\theta_2)\right]\\ &= R^2\left[\sin(\theta_1)\sin(\theta_2)\cos(\phi_2-\phi_1)+\cos(\theta_1)\cos(\theta_2)\right]\end{aligned} \tag{6.5.212}$$

In function of the angle that we will denote α we know that the dot product has also for expression:

$$\vec{P}_1 \circ \vec{P}_2 = R^2\cos(\alpha) \tag{6.5.213}$$

By identification of that latter relation with prior-previous one we get:

$$R^2\cos(\alpha) = R^2\left[\sin(\theta_1)\sin(\theta_2)\cos(\phi_2-\phi_1)+\cos(\theta_1)\cos(\theta_2)\right] \tag{6.5.214}$$

Hence after simplification:

$$\alpha = \arccos\left(\sin(\theta_1)\sin(\theta_2)\cos(\phi_2-\phi_1)+\cos(\theta_1)\cos(\theta_2)\right) \tag{6.5.215}$$

As we know that the circumference of circle is given by $P=2\pi R$ then the arc lengths by unit of angle is given by:

$$l = R\alpha \tag{6.5.216}$$

So finally we get the great circle distance relation:

$$\boxed{l = R\arccos\left(\sin(\theta_1)\sin(\theta_2)\cos(\phi_2-\phi_1)+\cos(\theta_1)\cos(\theta_2)\right)} \tag{6.5.217}$$

That can be found in the literature into different forms because of the spherical angles convention as already mentioned during our introduction of spherical coordinates at page 1849.

The relation above works well (we will do a numerical application later!) and we can see many people copying it in spreadsheet softwares or business intelligence softwares to run some kind of geostatistics analytics. However you may encounter another relation in practice (again used in various spreadsheet softwares) that is quite old and gives the same result. The reason both are used is because people copy/paste the first relationship the found thanks to Internet search engines without investigating if there are multiple ways to calculate a same thing. The relation we will see now exist because at a time where computer didn't exist it was easier to be used thanks to tables written by some humans...

Remark

For reasons that are a mixture of valid science and historical accident, there is no one agreed 'latitude and longitude' coordinate system. There are many different meridians of zero longitude (prime meridians) and many different circles of zero latitude (equators), although the former generally pass somewhere near Greenwich, and the latter is always somewhere near the rotational equator.

The result is that different systems of latitude and longitude in common use today can disagree on the coordinates of a point by more than 200 meters! For any application where an error of this size would be significant, it's important to know which system is being used and exactly how it is defined.

The figure below shows three points that all have the same latitude and longitude (approximately N50 56 18, W001 28 11), in three different coordinate systems (OSGB36, WGS84 and ED50). Each one of these coordinate systems is widely used and fit for its purpose, and none of them is wrong. The differences between them are just a result of the fact that any system of 'absolute coordinates' is always arbitrary. Standard conventions ensure only that different coordinate systems tend to agree to within half a kilometer or so, but there is no fundamental reason why they should agree at all.

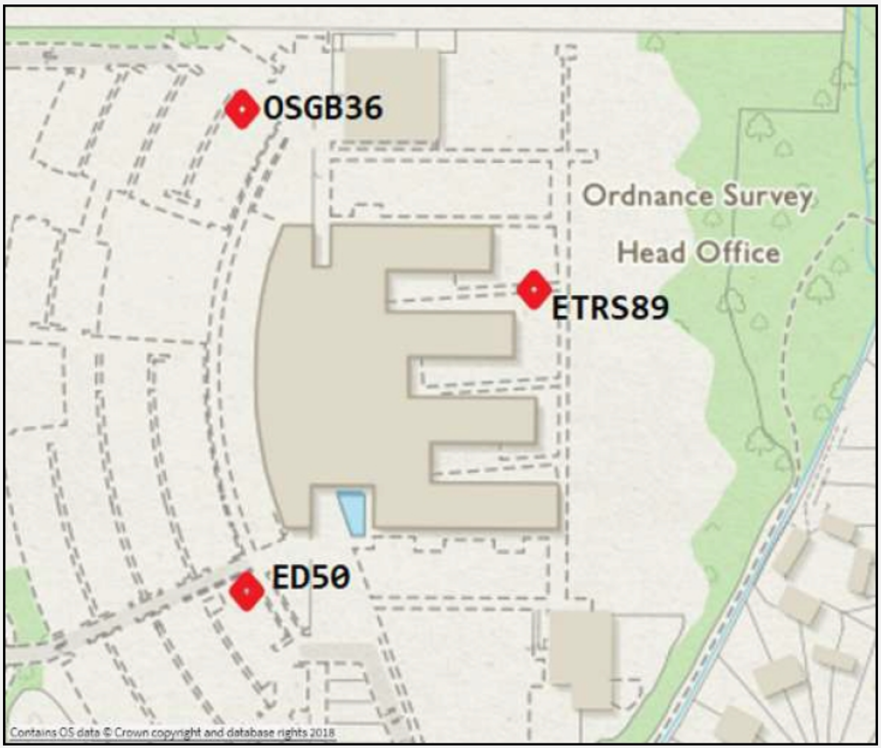


Figure 6.117: **Three points with the same latitude and longitude in three different coordinate systems (the map extract is approximately 300m wide)**

Furthermore, the true coordinates of a ground point do change due to the continuous deforming motions of the Earth. The relative motion of two continents can be 10 centimeters a year, which is significant for mapping because it is constant year after year – after 50 years a region of the earth may have moved by 5 meters relative to a neighbouring continent.

6.5.7.2.2 Haversine distance

The term "haversine" apparently comes from "half versed sine". To derive the haversine distance we assume that we are given two points P_1 and P_2 determined by their respective latitude-longitude pairs: $P_1(\theta_1,\phi_1), P_2(\theta_2,\phi_2)$. In cartesian coordinates we have $P_1=P_1(x_1,y_1,z_1)$ and $P_2=P_2(x_2,y_2,z_2)$, where x,y and z are determined by the spherical coordinates through the other typical convention:

$$\begin{aligned}x&=R\cos(\theta)\cos(\phi)\\y&=R\cos(\theta)\sin(\phi)\\z&=R\sin(\theta)\end{aligned}\tag{6.5.218}$$

The euclidean distance d between P_1 and P_2 is given by the three-dimensional Pythagorean theorem:

$$d^2=(x_1-x_2)^2+(y_1-y_2)^2+(z_1-z_2)^2\tag{6.5.219}$$

The bulk of our work will be in computing this distance in terms of the spherical coordinates. Converting the cartesian coordinates to spherical coordinates, we get:

$$\begin{aligned}d^2&=(R\cos(\theta_1)\cos(\phi_1)-R\cos(\theta_2)\cos(\phi_2)^2+(R\cos(\theta_1)\sin(\phi_1)-R\cos(\theta_2)\sin(\phi_2))^2\\&+(R\sin(\theta_1)-R\sin(\theta_2))^2\\&=R^2\left[\cos^2(\theta_1)\cos^2(\phi_1)-2\cos(\theta_1)\cos(\phi_1)\cos(\theta_2)\cos(\phi_2)+\cos^2(\theta_2)\cos^2(\phi_2)\right.\\&+\cos^2(\theta_1)\sin^2(\phi_1)-2\cos(\theta_1)\sin(\phi_1)\cos(\theta_2)\sin(\phi_2)+\cos^2(\theta_2)\sin^2(\phi_2)\\&\left.+\sin^2(\theta_1)-2\sin(\theta_1)\sin(\theta_2)+\sin^2(\theta_2)\right]\\&=R^2\left(2-2\cos(\theta_1)\cos(\theta_2)\cos(\phi_1-\phi_2)-2\sin(\theta_1)\sin(\theta_2)\right)\end{aligned}\tag{6.5.220}$$

Hence:

$$\frac{d^2}{R^2}=2-2\cos(\theta_1)\cos(\theta_2)\cos(\phi_1-\phi_2)-2\sin(\theta_1)\sin(\theta_2)\tag{6.5.221}$$

The distance is for historical (computational) reasons often represented in therms of the so-named "haversine function" defined by:

$$\text{haversine}(A)=\sin^2\left(\frac{A}{2}\right)=\frac{1-\cos(A)}{2}\tag{6.5.222}$$

This says that $\cos(A)=1-2\text{haversine}(A)$.

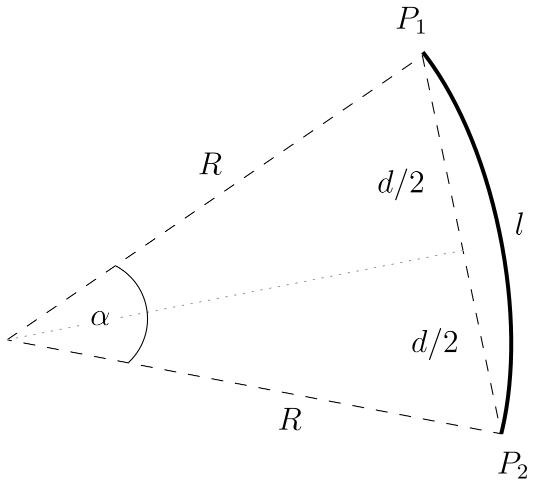
Returning to the relation for d^2/R^2 above, we continue, feeding in the new notation and using trigonometric identities:

$$\begin{aligned}\frac{d^2}{R^2}&=2-2\cos(\theta_1)\cos(\theta_2)\cos(\phi_1-\phi_2)-2\sin(\theta_1)\sin(\theta_2)\\&=2-2\cos(\theta_1)\cos(\theta_2)\left(1-2\text{haversine}(\phi_1-\phi_2)\right)-2\sin(\theta_1)\sin(\theta_2)\\&=2-2\cos(\theta_1-\theta_2)+4\cos(\theta_1)\cos(\theta_2)\text{haversine}(\phi_1-\phi_2)\\&=4\text{haversine}(\theta_1-\theta_2)+4\cos(\theta_1)\cos(\theta_2)\text{haversine}(\phi_1-\phi_2)\end{aligned}\tag{6.5.223}$$

Therefore:

$$\left(\frac{d}{2R}\right)^2=\text{haversine}(\theta_1-\theta_2)+\cos(\theta_1)\cos(\theta_2)\text{haversine}(\phi_1-\phi_2) \tag{6.5.224}$$

Now, based on the figure below:



we have:

$$\sin\left(\frac{\alpha}{2}\right)=\frac{d}{2R}\Leftrightarrow\sin^2\left(\frac{\alpha}{2}\right)=\text{haversine}(\alpha)=\left(\frac{d}{2R}\right)^2 \tag{6.5.225}$$

Therefore:

$$\text{haversine}(\alpha)=\text{haversine}(\theta_1-\theta_2)+\cos(\theta_1)\cos(\theta_2)\text{haversine}(\phi_1-\phi_2) \tag{6.5.226}$$

But remember that we are looking for l in the above figure. So as $l=R\alpha$, and therefore $\alpha=l/R$, we have:

$$\text{haversine}\left(\frac{l}{R}\right)=\text{haversine}(\theta_1-\theta_2)+\cos(\theta_1)\cos(\theta_2)\text{haversine}(\phi_1-\phi_2) \tag{6.5.227}$$

Remark

On Wikipedia, the latter relation is denoted:

$$\text{hav}\left(\frac{d}{r}\right)=\text{hav}(\varphi_2-\varphi_1)+\cos(\varphi_1)\cos(\varphi_2)\,\text{hav}(\lambda_2-\lambda_1)$$

Explicitly:

$$\sin^2\left(\frac{l}{2R}\right)=\text{haversine}(\theta_1-\theta_2)+\cos(\theta_1)\cos(\theta_2)\text{haversine}(\phi_1-\phi_2) \tag{6.5.228}$$

Therefore:

$$l=2R\arcsin\left(\sqrt{\text{haversine}(\theta_1-\theta_2)+\cos(\theta_1)\cos(\theta_2)\text{haversine}(\phi_1-\phi_2)}\right) \tag{6.5.229}$$

Or even more explicitly we have then the "haversine distance":

$$\boxed{l = 2R\arcsin\left(\sqrt{\sin^2\left(\frac{\theta_1-\theta_2}{2}\right)+\cos(\theta_1)\cos(\theta_2)\sin^2\left(\frac{\phi_1-\phi_2}{2}\right)}\right)} \qquad (6.5.230)$$

Working out the great circle distance between two points is so important in navigation that people in the old days produced tables giving the values of the haversine function and also of the inverse of the haversine function. This made seafarers' lives a lot easier. Working out the distance l only involved looking up two cosine values and two haversine values, adding and multiplying them in the correct way, looking up the inverse of the haversine function and multiplying by R— done!

The reason why the haversine function has come out of fashion is that with the help of calculators and computers it's easy enough to work previous calculation methods. That's why you don't find a haversine button on your average calculator.

✍**Example:**

Let us compare the result of the two relations established earlier above! That is:

$$l = 2R\arcsin\left(\sqrt{\sin^2\left(\frac{\theta_1-\theta_2}{2}\right)+\cos(\theta_1)\cos(\theta_2)\sin^2\left(\frac{\phi_1-\phi_2}{2}\right)}\right)$$

and:

$$l = R\arccos\left(\sin(\theta_1)\sin(\theta_2)\cos(\phi_2-\phi_1)+\cos(\theta_1)\cos(\theta_2)\right)$$

Let's give it a go. Liverpool has latitude 53.4° and longitude -3°, and New York has latitude 40.71° and longitude -74°. These are measured in degrees. Converting them into radians (multiplying by $\pi/180$) gives $\theta_\mathrm{L}=0.932^\circ$ and $\phi_\mathrm{L}=-0.052^\circ$ for Liverpool, and $\theta_\mathrm{NY}=0.71^\circ$ and $\phi_\mathrm{NY}=-1.291^\circ$ for New York (rounded to three decimal places).

Plugging these into the haversine relation, with the radius of the Earth $R=6371$, gives:

$$l = 2\cdot 6371\cdot\arcsin\left(\sqrt{\sin^2\left(\frac{0.932-(-0.052)}{2}\right)+\cos(0.932)\cos(0.71)\sin^2\left(\frac{(-0.052)-(-1.291)}{2}\right)}\right)\cong 5323\,[\mathrm{km}]$$

and with the second method we get the same result (remember we establised it using another convention for the definition of spherical coordinates that's why we have to take the complementary angle):

$$l = R\arccos\left(\sin(\pi/2-0.932)\sin(\pi/2-0.71)\cos((-1.291)-(-0.052))+\cos(\theta_1)\cos(\theta_2)\right)\cong 5323\,[\mathrm{km}]$$

6.5.7.3 Cylindrical Coordinate System

The cylindrical coordinate system (very useful in the study of helical motion systems) is quite similar to spherical coordinates as it can be seen as a slice of the sphere.

Given the figure:

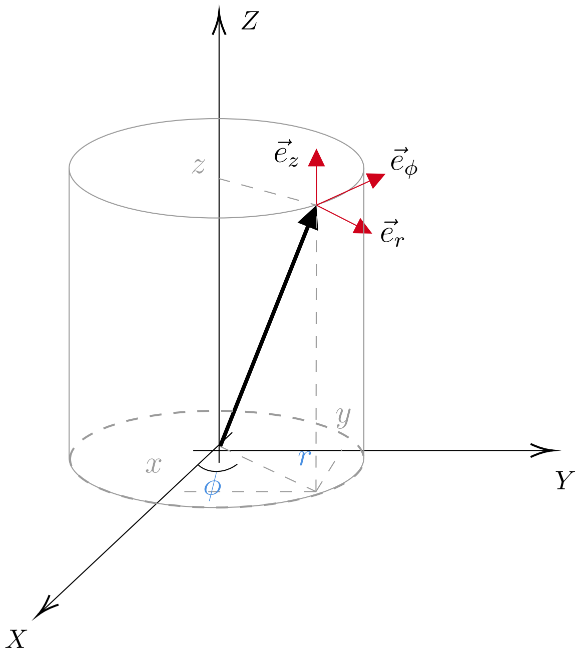


Figure 6.118: **Representation of the cylindrical coordinate system**

Warning!!! The vector $\vec{r}$ is unlike the previous spherical case defined only in the XY plane or a plane which is parallel to it!

It comes easily in cylindrical coordinates for $r>0$:

$$r=\sqrt{x^2+y^2} \quad \phi=\arccos\left(\frac{x}{r}\right) \quad \text{or} \quad \phi=\arcsin\left(\frac{y}{r}\right) \quad \text{or} \quad \phi=\arctan\left(\frac{x}{y}\right) \tag{6.5.231}$$

and vice versa:

$$x=r\cos(\phi) \quad y=r\sin(\phi) \quad z=z \tag{6.5.232}$$

Now we must find the expressions that connect the vectors of the cylindrical base that we choose to denote by $\vec{e}_r, \vec{e}_\phi, \vec{e}_z$ (instead of $\vec{r}, \vec{\phi}, \vec{z}$ as it is done traditionally) with the vectors of the Cartesian base $\vec{e}_x, \vec{e}_y, \vec{e}_z$. We have identically to what we did for the spherical coordinates:

$$\begin{aligned}
\vec{e}_r &= \frac{\vec{r}}{r}=\frac{x\vec{e}_x+y\vec{e}_y}{r}=\vec{e}_x\cos(\phi)+\vec{e}_y\sin(\phi)\\
\vec{e}_\phi &= \frac{\vec{e}_z\times\vec{e}_r}{\sin(\phi)}=-\vec{e}_x\sin(\phi)+\vec{e}_y\cos(\phi)\\
\vec{e}_z &= \vec{e}_z
\end{aligned} \tag{6.5.233}$$

Or more explicitly:

$$\boxed{\vec{e}_r=\begin{pmatrix}\cos(\phi)\\ \sin(\phi)\\ 0\end{pmatrix},\vec{e}_\phi=\begin{pmatrix}-\sin(\phi)\\ \cos(\phi)\\ 0\end{pmatrix},\vec{e}_z=\begin{pmatrix}0\\ 0\\ 1\end{pmatrix}}\tag{6.5.234}$$

All this can be put in a famous matrix form as (this also applies automatically to the component of the unitary vector):

$$\begin{pmatrix}\vec{e}_r\\ \vec{e}_\phi\\ \vec{e}_z\end{pmatrix}=\begin{bmatrix}\cos(\phi) & \sin(\phi) & 0\\ -\sin(\phi) & \cos(\phi) & 0\\ 0 & 0 & 1\end{bmatrix}\begin{pmatrix}\vec{e}_x\\ \vec{e}_y\\ \vec{e}_z\end{pmatrix}\tag{6.5.235}$$

Let us indicate that by dividing by $\sin(\phi)$ the second vector base $\vec{e}_\phi$, we ensure us that by the properties of the norm of the cross product we have:

$$\|\vec{e}_\phi\|=\|\vec{e}_z\|\cdot\|\vec{e}_r\|\sin(\phi)=1\tag{6.5.236}$$

will be well normalized to unity as expected (as we know from start that as we take a direct orthogonal coordinate system the product of norms of the basis vectors must be equal to 1)!! In the case of cylindrical coordinates the angle being anyway right, we would not be obliged to indicate this division, but we have made this choice for consistency with previous developments...

Remark

It is important to notice that the cross product of two basis vectors always gives the third perpendicular basis vector (like the Cartesian and spherical coordinates so!).

For future needs, let us determine the partial differential of each of these coordinates:

$$\begin{gathered}\frac{\partial\vec{e}_r}{\partial r}=0\quad;\quad\frac{\partial\vec{e}_r}{\partial\phi}=-\vec{e}_x\sin(\phi)+\vec{e}_y\cos(\phi)=\vec{e}_\phi\quad;\quad\frac{\partial\vec{e}_r}{\partial z}=0\\ \frac{\partial\vec{e}_\phi}{\partial r}=0\quad;\quad\frac{\partial\vec{e}_\phi}{\partial\phi}=-\vec{e}_x\cos(\phi)-\vec{e}_y\sin(\phi)=-\vec{e}_r\quad;\quad\frac{\partial\vec{e}_\phi}{\partial z}=0\\ \frac{\partial\vec{e}_z}{\partial r}=0\quad;\quad\frac{\partial\vec{e}_z}{\partial\phi}=0\quad;\quad\frac{\partial\vec{e}_z}{\partial z}=0\end{gathered}\tag{6.5.237}$$

We will also use later (for the study of vector operators) the variation $\mathrm{d}\vec{r}$ expressed in cylindrical coordinates:

$$\mathrm{d}\vec{r}=\mathrm{d}(r\vec{e}_r)=\vec{e}_r\mathrm{d}r+r\mathrm{d}\vec{e}_r=\vec{e}_r\mathrm{d}r+r\left(\frac{\partial\vec{e}_r}{\partial r}\mathrm{d}r+\frac{\partial\vec{e}_r}{\partial\phi}\mathrm{d}\phi+\frac{\partial\vec{e}_r}{\partial z}\mathrm{d}z\right)=\vec{e}_r\mathrm{d}r+r\vec{e}_\phi\mathrm{d}\phi\tag{6.5.238}$$

To express the speed and acceleration in cylindrical coordinates, we will also need the derivatives with respect to time:

$$\begin{aligned}\dot{\vec{e}}_r&=\frac{\partial\vec{e}_r}{\partial r}\dot{r}+\frac{\partial\vec{e}_r}{\partial\phi}\dot{\phi}+\frac{\partial\vec{e}_r}{\partial z}\dot{z}=\vec{e}_\phi\dot{\phi}\\ \dot{\vec{e}}_\phi&=\frac{\partial\vec{e}_\phi}{\partial r}\dot{r}+\frac{\partial\vec{e}_\phi}{\partial\phi}\dot{\phi}+\frac{\partial\vec{e}_r}{\partial z}\dot{z}=-\vec{e}_r\dot{\phi}\\ \dot{\vec{e}}_z&=\frac{\partial\vec{e}_z}{\partial r}\dot{r}+\frac{\partial\vec{e}_z}{\partial\phi}\dot{\phi}+\frac{\partial\vec{e}_z}{\partial z}\dot{z}=\vec{0}\end{aligned}\tag{6.5.239}$$

So if we now do a little bit physics, we get (let us recall that the z component is independent of other cylindrical components):

$$\vec{v}=\dot{\vec{r}}+\dot{\vec{z}}=\dot{\vec{e}}_r r+\dot{r}\vec{e}_r+\vec{e}_z\dot{z}+\dot{\vec{e}}_z z \tag{6.5.240}$$

which brings us to:

$$\boxed{\vec{v}=\dot{r}\vec{e}_r+\vec{e}_\phi r\dot{\phi}+\vec{e}_z\dot{z}} \tag{6.5.241}$$

For acceleration we get (exactly the same approach as for the expression of the speed):

$$\boxed{\vec{a}=\vec{e}_r\left(\ddot{r}-r\dot{\phi}^2\right)+\vec{e}_\phi\left(r\ddot{\phi}+2\dot{r}\dot{\phi}\right)+\vec{e}_z\ddot{z}} \tag{6.5.242}$$

6.5.7.4 Polar Coordinate System

The polar coordinate system is very similar to the cylindrical coordinates as it can be seen as an entrenchment of one dimension (the height) of the cylindrical system (we will often encounter this system in the section of Classical Mechanics, Corpuscular Quantum Physics and Astronomy).

Given the figure:

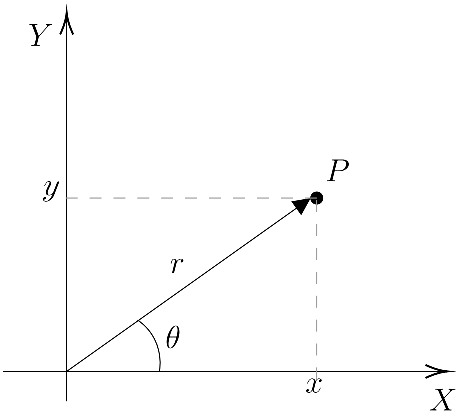


Figure 6.119: **Representation of the polar coordinate system**

Thus, it comes easily in polar coordinates for $r>0$:

$$r=\sqrt{x^2+y^2} \qquad \phi=\arccos\left(\frac{x}{r}\right) \qquad \phi=\arcsin\left(\frac{y}{r}\right) \quad \text{and} \quad \phi=\arctan\left(\frac{y}{x}\right) \tag{6.5.243}$$

and vice versa:

$$x=r\cos(\phi) \qquad y=r\sin(\phi) \tag{6.5.244}$$

Remark

In physics we frequently use the previous parametric equation of the circle in the following way involving time t and pulsation (as a constant) ω:

$$\vec{r}(t)=r\cos(\omega t)\vec{e}_x+r\sin(\omega t)\vec{e}_y \tag{6.5.245}$$

Hence:

$$\|\vec{r}(t)\|=\sqrt{r^2\cos^2(\omega t)+r^2\sin^2(\omega t)} \tag{6.5.246}$$

Now we must find the expressions that connect the vectors of the polar base that we choose to denote by $\vec{e}_r,\vec{e}_\phi$ (instead of $\vec{r},\vec{\phi}$ as it is done traditionally) with the vectors of the Cartesian base $\vec{e}_x,\vec{e}_y$. We have identically to what we did for the spherical coordinates:

$$\begin{aligned}\vec{e}_r&=\frac{\vec{r}}{r}=\frac{x\vec{e}_x+y\vec{e}_y}{r}=\vec{e}_x\cos(\phi)+\vec{e}_y\sin(\phi)\\ \vec{e}_\phi&=\frac{\vec{e}_z\times\vec{e}_r}{\sin(\theta)}=-\vec{e}_x\sin(\phi)+\vec{e}_y\cos(\phi)\end{aligned} \tag{6.5.247}$$

Or more explicitly:

$$\boxed{\vec{e}_r=\begin{pmatrix}\cos(\phi)\\ \sin(\phi)\end{pmatrix},\vec{e}_\phi=\begin{pmatrix}-\sin(\phi)\\ \cos(\phi)\end{pmatrix}} \tag{6.5.248}$$

Once again, dividing by $\sin(\theta)$ the second basis vector $\vec{e}_\phi$, we ensure the properties of the norm of the vector product that:

$$\|\vec{e}_\phi\|=\|\vec{e}_z\|\cdot\|\vec{e}_r\|\sin(\theta)=1 \tag{6.5.249}$$

will be well normalized to unity (as we know from start that as we take a direct orthogonal coordinate system the product of norms of the basis vectors must be equal to 1)!. In the case of polar coordinate the angle being a anyway right, we would not be obliged to indicate this division, but we have made this choice for consistency with the previous developments.

For future needs, Let us determine the partial differential of each of these coordinates:

$$\begin{aligned}\frac{\partial\vec{e}_r}{\partial r}&=\vec{0}\qquad&\frac{\partial\vec{e}_r}{\partial\phi}&=-\vec{e}_x\sin(\phi)+\vec{e}_y\cos(\phi)=\vec{e}_\phi\\ \frac{\partial\vec{e}_\phi}{\partial r}&=0\qquad&\frac{\partial\vec{e}_\phi}{\partial\phi}&=-\vec{e}_x\cos(\phi)-\vec{e}_y\sin(\phi)=-\vec{e}_r\end{aligned} \tag{6.5.250}$$

We will also use later (for the study of vector operators) the variation $\mathrm{d}\vec{r}$ expressed in polar coordinates:

$$\begin{aligned}\mathrm{d}\vec{r}&=\mathrm{d}(r\vec{e}_r)=\vec{e}_r\mathrm{d}r+r\mathrm{d}\mathrm{d}\vec{e}_r=\vec{e}_r\mathrm{d}r+r\left(\frac{\partial\vec{e}_r}{\partial r}\mathrm{d}r+\frac{\partial\vec{e}_r}{\partial\phi}\mathrm{d}\phi\right)\\ &=\vec{e}_r\mathrm{d}r+\vec{e}_\phi r\mathrm{d}\phi\end{aligned} \tag{6.5.251}$$

To express the speed and acceleration in polar coordinates, we will also need the derivatives with respect to time:

$$\begin{aligned}\dot{\vec{e}}_r&=\frac{\partial\vec{e}_r}{\partial r}\dot{r}+\frac{\partial\vec{e}_r}{\partial\phi}\dot{\phi}=\vec{e}_\phi\dot{\phi}\\ \dot{\vec{e}}_\phi&=\frac{\partial\vec{e}_\phi}{\partial r}\dot{r}+\frac{\partial\vec{e}_\phi}{\partial\phi}\dot{\phi}=-\vec{e}_r\dot{\phi}\end{aligned} \tag{6.5.252}$$

So if we now do a little bit physics, we have:

$$\vec{v}=\dot{\vec{r}}=\dot{\vec{e}}_r r+\dot{r}\vec{e}_r \tag{6.5.253}$$

and therefore:

$$\boxed{\vec{v}=\dot{r}\vec{e}_r+\vec{e}_\phi r\dot{\phi}} \tag{6.5.254}$$

where the first term is the radial velocity component and the second term the tangential component of the (angular) velocity. The velocity expression in polar coordinates is very important in astronomy as it allows quite easily calculate the calculation of the kinetic energy:

$$E_c=\frac{1}{2}m\|\vec{v}\|^2=\frac{1}{2}m\left(\sqrt{\dot{r}^2+r^2\dot{\phi}}\right)^2=\frac{1}{2}\left(\dot{r}^2+r^2\dot{\phi}\right) \tag{6.5.255}$$

For the acceleration we get:

$$\boxed{\vec{a}=\vec{e}_r(\ddot{r}-\dot{\phi}^2 r)+\vec{e}_\phi(2\dot{r}\dot{\phi}+r\ddot{\phi})} \tag{6.5.256}$$

where the first term is the radial acceleration, the second term the centripetal acceleration, the third the Coriolis acceleration and finally the fourth one the tangential acceleration.

6.5.8 Differential Operators

Definition (#6.5.250): Define a scalar field, vector field or tensor field in a volume V, it is define an application that for any point $\vec{r}$ of this volume V associates respectively a scalar, a vector or a tensor.

Thus, the application f that at any point $\vec{r}$ of V of spatial coordinates (x,y,z) associates the scalar value $f(\vec{r})=f(x,y,z)$ is a scalar field in V.

At each point of a volume traversed by a moving fluid, the vector that coincides at every moment with the speed of the particle which pass through this point at this same time defines a 3D vector field, optionally variable in time. The fields thus defined are a basic mathematical tool in physics.

Remark

When we plot a scalar field, the set of continuous dots of equal value is so-named "iso-lines" or more commonly "contours".

Vector fields are especially an important tool for describing many physical concepts, such as gravitation and electromagnetism, which affect the behaviour of objects over a large region of a plane or of space. They are also useful for dealing with large-scale behaviour such as atmospheric storms or deep-sea ocean currents.

For example the figure below shows a gravitational field exerted by two astronomical objects, such as a star and a planet or a planet and a moon. At any point in the figure, the vector associated with a point gives the net gravitational force[16] exerted by the two objects on an object of unit mass. The vectors of largest magnitude in the figure are the vectors closest to the larger object. The larger object has greater mass, so it exerts a gravitational force of greater magnitude than the smaller object.

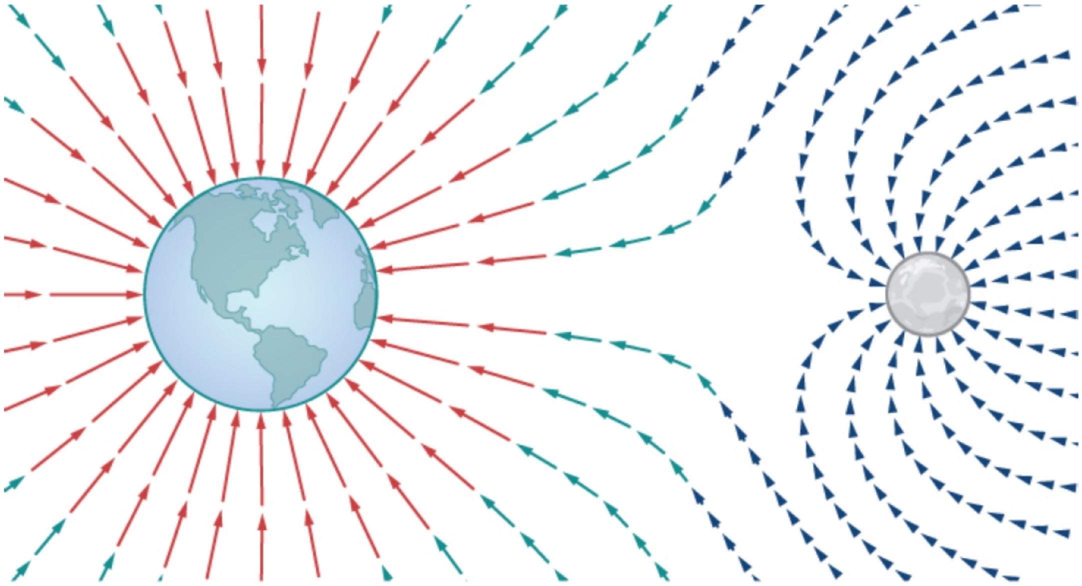


Figure 6.120: **Gravitation vector field example (source: OpenStax)**

---

[16] We will see it during our study of General Relativity that the gravitational force is fictional and we should us the expression "gravitational interaction" instead.

Another example is the figure below that shows the velocity of a river at points on its surface. The vector associated with a given point on the river's surface gives the velocity of the water at that point. Since the vectors to the left of the figure are small in magnitude, the water is flowing slowly on that part of the surface. As the water moves from left to right, it encounters some rapids around a rock. The speed of the water increases, and a whirlpool occurs in part of the rapids.

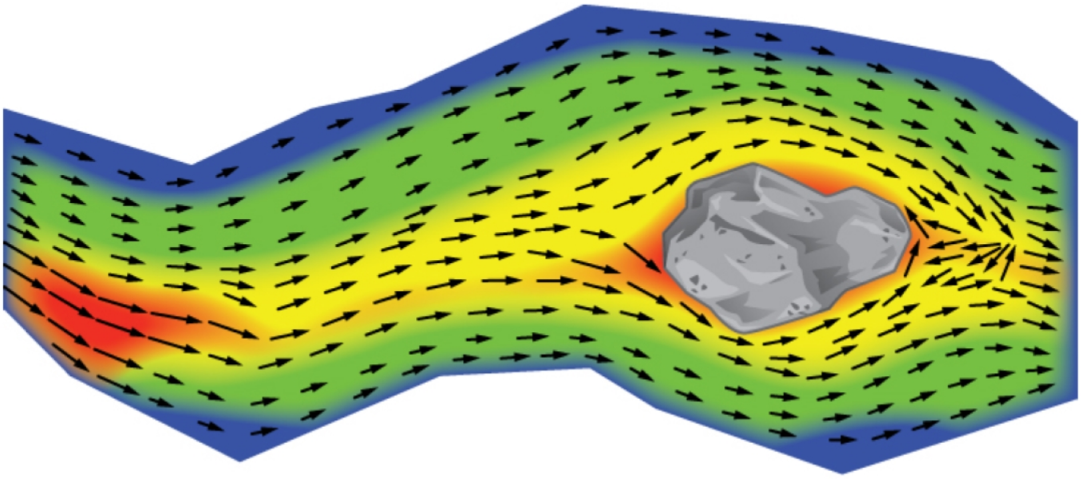
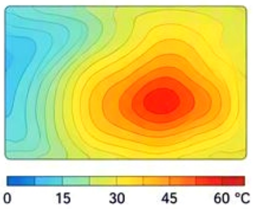
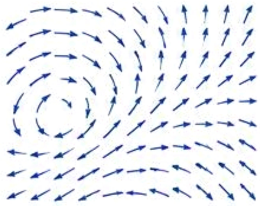
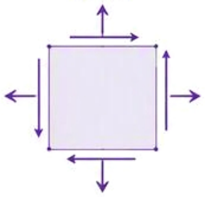


Figure 6.121: **Velocity vector field example (source: OpenStax)**

The gradient, divergence and the rotational are the three main linear differential operators of the first order that will be introduced here. This means they only involve the partial first derivatives (or simply "differentials") of the fields, unlike, for example, the Laplace operator which involves partial derivatives of order 2.

SCALAR, VECTOR, AND TENSOR FIELDS

A COMPARISON OF THEIR MAIN CHARACTERISTICS AND APPLICATIONS

	SCALAR FIELD	VECTOR FIELD	TENSOR FIELD (ORDER 2)
NATURE	Associates a real number with each point in space.	Associates a vector with each point in space.	Associates a second-order tensor (matrix) with each point in space.
TYPICAL SYMBOLS	$\varphi,\ f,\ T,\ \rho,\ \dots$	$\vec{F},\ \vec{v},\ \vec{A},\ \vec{B},\ \dots$	$T,\ \sigma,\ \varepsilon,\ C,\ \dots$
VALUE AT A POINT	A single number (e.g., $\varphi(P) = 3.14$)	A vector (magnitude and direction) (e.g., $\vec{F}(P) = (2, -1, 3)$)	A matrix (tensor) (e.g., $T(P) = \begin{pmatrix} 1 & 2 & 0 \\ 2 & -1 & 3 \\ 0 & 3 & 4 \end{pmatrix}$)
VISUAL REPRESENTATION	Temperature map  	Wind field 	Stress tensor (in 2D) 
PHYSICAL EXAMPLES	Temperature Pressure Electric potential Density Concentration	Fluid velocity Electric field Magnetic field Acceleration Flux (e.g., heat flux)	Stress tensor (mechanics) Strain tensor (deformations) Inertia tensor Permittivity tensor Conductivity tensor
NUMBER OF INDEPENDENT COMPONENTS IN 3D	**1**	**3** (x, y, z)	**9** in general (6 if symmetric)
BEHAVIOR UNDER COORDINATE TRANSFORMATION	Invariant (the value does not change)	Transforms as a vector (rotation + change of basis)	Transforms according to the law of second-order tensors (rotation + change of basis)
COMMON OPERATIONS	Grad, div, $\nabla\varphi$	div $\vec{F}$, curl $\vec{F}$, $\nabla \cdot \vec{F}$, $\nabla \times \vec{F}$	Contraction, trace, product $\nabla \cdot T,\ T : S,\ \mathrm{tr}(T), \dots$
PRACTICAL APPLICATIONS	**Weather & Climate** Temperature, pressure in atmospheric models. **Electrostatics** Electric potential in devices and materials. **Medical Imaging** Intensity in MRI, CT, ultrasound images.	**Fluid Dynamics** Airflow around aircraft, ocean currents. **Electromagnetism** Electric and magnetic fields in motors, antennas. **Mechanics** Velocity and acceleration of bodies and structures.	**Structural Mechanics** Stress analysis in bridges, buildings, and materials. **Continuum Mechanics** Strain and deformation of materials. **Anisotropic Materials** Modeling of crystals, composites, and tissues.

IN SUMMARY

● **SCALAR FIELD**
A number at each point: intensity without direction.

→

→ **VECTOR FIELD**
A vector at each point: intensity with direction.

→

▦ **TENSOR FIELD**
A tensor at each point: directional relationships with greater complexity.

6.5.8.1 Gradients of Scalar Field

The gradient is an operator that applies to a scalar field and transforms it into a vector field. Intuitively, the gradient indicates the direction of the greater variation of the scalar field, and the intensity of this variation. For example, the gradient of the altitude is directed along the maximum slope line and its norm increases with the slope.

Given a three-dimensional scalar field $f(x,y,z)$, wherein x and y and z are the Cartesian coordinates of a point M in space. When M moves in space according to the $\mathrm{d}\vec{r}$ of components $\mathrm{d}x,\mathrm{d}y$ and $\mathrm{d}z$, the scalar field f varies according to the total differential $\mathrm{d}f$:

$$\mathrm{d}f=\frac{\partial f}{\partial x}\mathrm{d}x+\frac{\partial f}{\partial y}\mathrm{d}y+\frac{\partial f}{\partial z}\mathrm{d}z=\begin{pmatrix}\dfrac{\partial f}{\partial x}\\\dfrac{\partial f}{\partial y}\\\dfrac{\partial f}{\partial z}\end{pmatrix}\circ\begin{pmatrix}\mathrm{d}x & \mathrm{d}y & \mathrm{d}z\end{pmatrix} \tag{6.5.257}$$

From this relation, we can define the "gradient operator", also named "gradient vector", of a scalar field such as:

$$\mathrm{d}f=\overrightarrow{\text{grad}}(f)\circ\mathrm{d}\vec{r}=\vec{\nabla}(f)\circ\mathrm{d}\vec{r} \tag{6.5.258}$$

where:

$$\overrightarrow{\text{grad}}(f) \tag{6.5.259}$$

is a vector operator named "gradient of the scalar field f". To condense the writing, we sometimes use the symbol $\vec{\nabla}$ named the "nabla of the scalar field f".

The vector obtained by the gradient calculation has the following three fundamental properties:

P1. The components of the gradient represent by construction the variation (slope) of the function f in the different directions of space.

P2. The gradient is perpendicular to the isolines of the function f.

P3. The direction of the gradient (and therefore its norm) is the maximum variation of f.

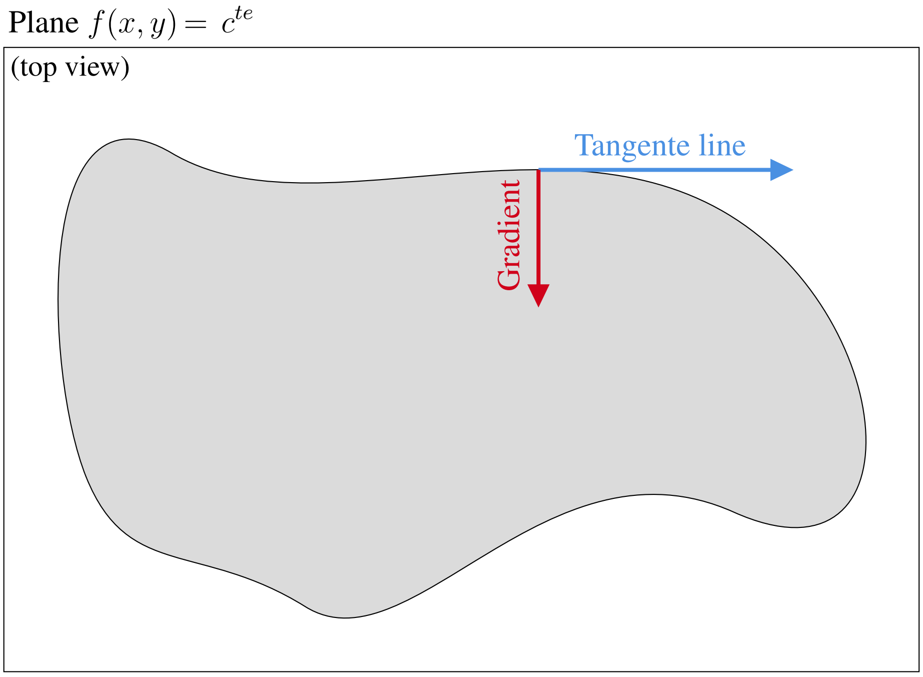
P4. The direction of the gradient shows the values where f increases.

Following the request of some readers let us prove some of these properties.

Given $t\mapsto C(t)$ an isoline. Then $t\mapsto f(c(t))=c^{te}$ and therefore:

$$0=\frac{\mathrm{d}}{\mathrm{d}t}f(c(t))=\vec{\nabla}_{c(t)}\left(f(t)\right)\circ c'(t) \tag{6.5.260}$$

which proves that the gradient is orthogonal to the tangent of the isoline (property P2). This can be illustrated as following:



Remark

In other words: The gradient $\vec{\nabla}_{x_0,y_0}f$ of a function $f(x,y)$ at x_0,y_0 is perpendicular to the tangent of the level-set $f(x,y)=0$ of the function at x_0,y_0.

Let us come back on:

$$\vec{\nabla}(f)\circ\mathrm{d}\vec{r} \tag{6.5.261}$$

That it is tradition to write as:

$$D_uf=\|\vec{\nabla}(f)\|\cdot\|\mathrm{d}\vec{r}\|\cos(\theta) \tag{6.5.262}$$

named the "directional derivative". Its value is maximum obviously if $\theta=0$, that is to say that the gradient is colinear to the variation $\mathrm{d}\vec{r}$. Hence, the direction (and therefore the norm) of greatest increase of f, often named direction of "deepest ascent", is the same direction as the gradient vector!! Thus we proved the property P3.

Obviously the directional derivative takes on its greatest negative value if $\theta=\pi$ (or 180 degrees). Hence, the direction of greatest decrease of f is the direction opposite to the gradient vector and often named direction of "deepest descent".

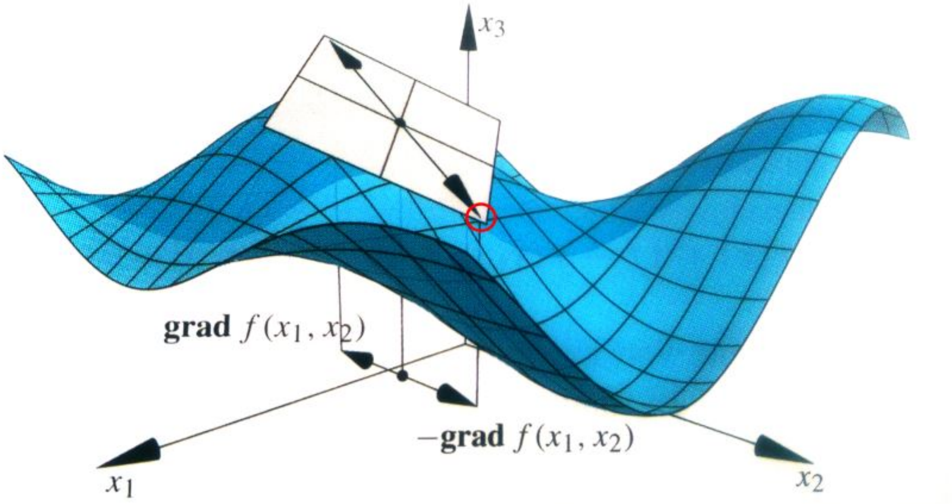


Figure 6.122: Gradient deepest ascent and deepest descent

If $\theta = \pi/2$ then there is no change in altitude in this direction and we are on an "isocline".

The property P4 can be explained without formalism as following:

- If the function is decreasing in one variable, then the partial derivative is negative, so the component vector of the gradient for that variable points in the negative direction - which means increasing function value.
- If the function is increasing in one variable, then the partial derivative is positive, so the component vector of the gradient for that variable points in the positive direction - which means increasing function value.

Then is doesn't matter how the function profile is, the gradient, by definition, points in the increasing direction. Indeed, when $f(x,y)$ is decreasing in x, the function decreases as you move forward in x. But because the partial derivative with respect to x is negative, the x-component of the gradient points towards origin (move backward in x), that is to say in the direction which makes f to increase.

From the definition and from the total differential, we get:

$$\frac{\partial f}{\partial x}\mathrm{d}x+\frac{\partial f}{\partial y}\mathrm{d}y+\frac{\partial f}{\partial z}\mathrm{d}z=\vec{\nabla}(f)_x\circ\mathrm{d}x+\vec{\nabla}(f)_y\mathrm{d}y+\vec{\nabla}(f)_z\mathrm{d}z \tag{6.5.263}$$

This leads us to put that:

$$\vec{\nabla}(f)_x=\frac{\partial f}{\partial x}\mathrm{d}x \qquad \vec{\nabla}(f)_y=\frac{\partial f}{\partial y}\mathrm{d}y \qquad \vec{\nabla}(f)_z=\frac{\partial f}{\partial z}\mathrm{d}z \tag{6.5.264}$$

and so that finally the operator of the "gradient in Cartesian coordinates" is given by:

$$\boxed{\vec{\nabla}=\left(\frac{\partial}{\partial x},\frac{\partial}{\partial y},\frac{\partial}{\partial z}\right)} \tag{6.5.265}$$

Finally we see that the gradient of a scalar field $f(x,y,z)$ is the vector field whose components at each point are the three derivatives of the scalar field f with respect to the three-dimensional coordinates, denoted here by x,y,z and that by its direction, and its norm, the gradient vector of a scalar field at a point therefore includes indications on how the field varies around this point.

Remark

One of the necessary and sufficient conditions for a vector field to be the gradient of a scalar field f is that this vector field is irrotational (see below the rotational operator of a vector field).

✍**Example:**

Let us find the direction for which the directional of:

$$f(x,y,)=3x^2-4xy+2y^2 \tag{6.5.266}$$

at $(-2,3)$ is a maximum and what is its maximum value?

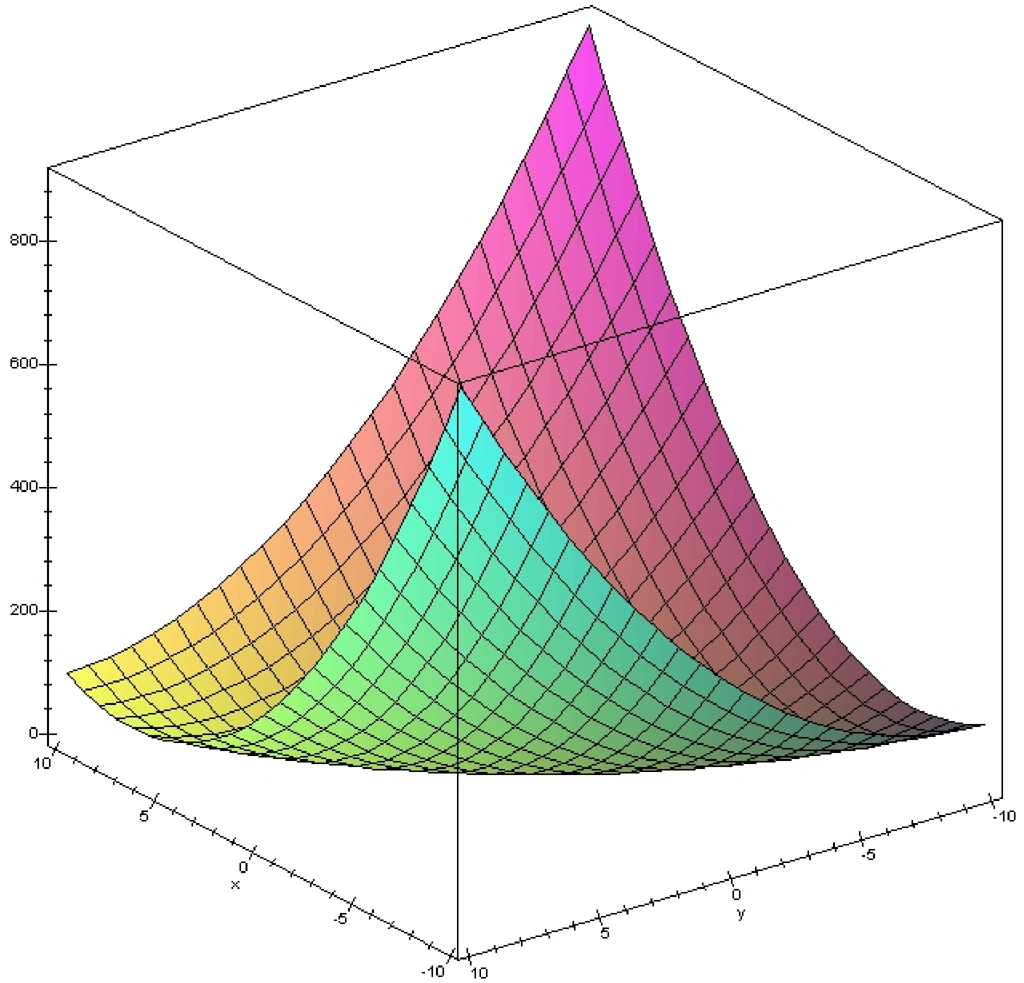


Figure 6.123: **Maple™ plot of** $f(x,y,)=3x^2-4xy+2y^2$

The maximum value of the directional derivative occurs as we have just proved it when

$\vec{\nabla} f$ and the unit vector $\mathrm{d}\vec{r}$ point in the same direction.

Therefore we start by calculating $\vec{\nabla} f(x,y)$:

$$\vec{\nabla} f(x,y) = \begin{pmatrix} \partial_x \\ \partial_y \end{pmatrix} (3x^2 - 4xy + 2y^2) = \begin{pmatrix} 6x - 4y \\ -4x + 4y \end{pmatrix} \tag{6.5.267}$$

Next we evaluate the gradient at $(-2,3)$:

$$\vec{\nabla} f(-2,3) = \begin{pmatrix} 6(-2) - 4(3) \\ -4(-2) + 4(3) \end{pmatrix} = \begin{pmatrix} -24 \\ 20 \end{pmatrix} \tag{6.5.268}$$

We need to find a unit vector that points in the same direction as $\vec{\nabla} f(-2,3)$, so the next step is to divide $\vec{\nabla} f(-2,3)$ by its norm (the value of that norm being also the maximum value of the directional derivative at point $(-2,3)$), which gives:

$$\frac{\vec{\nabla} f(-2,3)}{\|\vec{\nabla} f(-2,3)\|} = \frac{1}{\sqrt{(-24)^2 + (20)^2}} \begin{pmatrix} -24 \\ 20 \end{pmatrix} = \frac{1}{\sqrt{976}} \begin{pmatrix} -24 \\ 20 \end{pmatrix} \cong \begin{pmatrix} -0.768 \\ 0.640 \end{pmatrix} \tag{6.5.269}$$

As the unit vector above is a vector in the plane, nothing prevents us to calculate the angle it does with the axis. By applying elementary trigonometry, we get if we denote this angle by θ:

$$\cos(\theta) = \frac{-0.768}{1} = -0.768 \quad \text{or} \quad \sin(\theta) = \frac{0.640}{1} = 0.640 \tag{6.5.270}$$

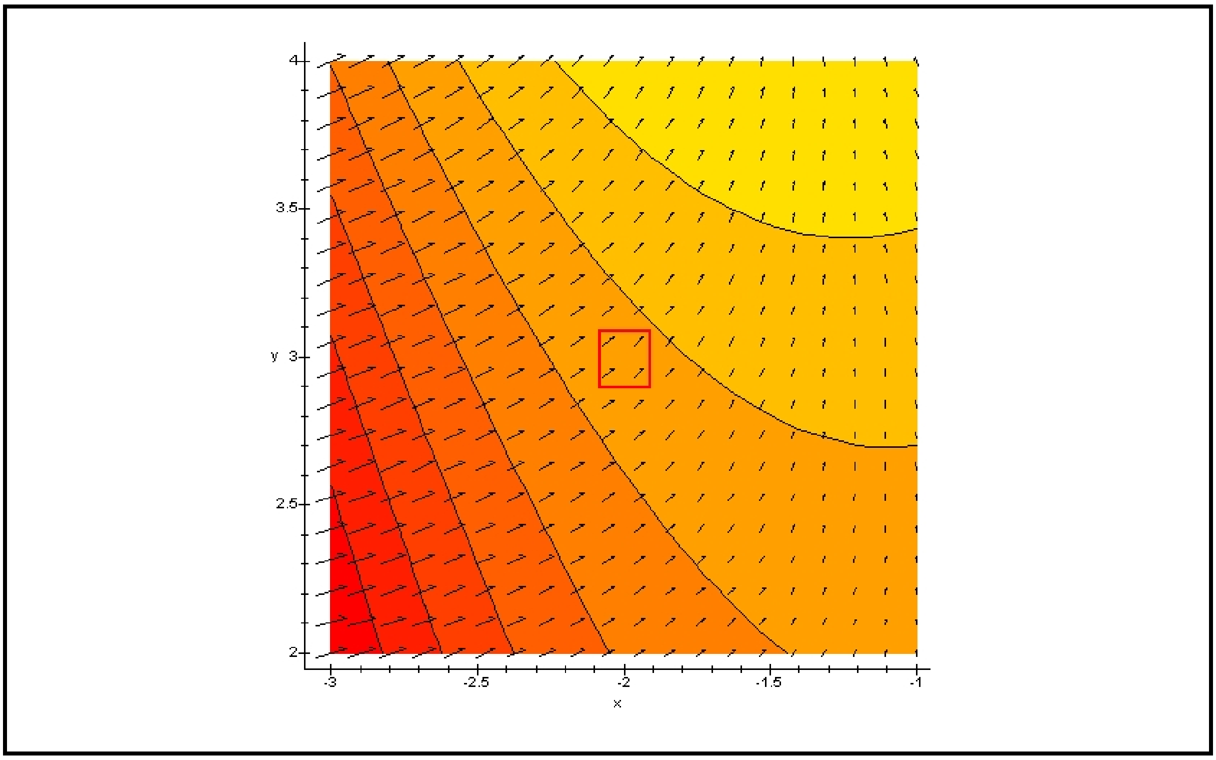
Since cosine is negative and sine is positive, the angle must be in the second quadrant. Therefore:

$$\theta = \pi - \arcsin(0.640) \cong 2.45\,[\text{rad}] \tag{6.5.271}$$

that is to say approximately 114.5916 degrees (or seen from the point of view of the first quadrant: 39.792 degrees).

With Maple™ 4.00b we can have a more detailed and general investigation of the calculation we just did:

```
>with(plots):
>with(linalg):
>Pa:=contourplot(3*x^3-4*x*y+2*y^2,x=-3..-1,y=2..4,filled=true):
>Pb:=fieldplot(grad(3*x^3-4*x*y+2*y^2,vector([x,y])),x=-3..-1,y=2..4):
>display(Pa,Pb);
```



After having defined the gradient in Cartesian coordinates x,y,z we have to address the expression of this operator in other coordinate systems. It is common in physics to have to use cylindrical, polar and spherical coordinates to simplify the formal study of physical systems. So if we refer to the previous study of coordinate systems, we have (recall) first in polar coordinates:

$$\mathrm{d}\vec{r}=\vec{e}_r\mathrm{d}r+\vec{e}_\phi r\mathrm{d}\phi \tag{6.5.272}$$

But, with the definition of gradient in Cartesian coordinates, in polar coordinates we have the following definition:

$$\mathrm{d}f=\vec{\nabla}(f)\circ\mathrm{d}\vec{r} \tag{6.5.273}$$

If we express the total exact differential (see section of Differential and Integral Calculus page 1466) of $\mathrm{d}f$ we get the following relation:

$$\mathrm{d}f=\frac{\partial f}{\partial r}\mathrm{d}r+\frac{\partial f}{\partial \phi}\mathrm{d}\phi=\vec{\nabla}(f)\circ\mathrm{d}\vec{r}=\vec{\nabla}(f)\circ(\vec{e}_r\mathrm{d}r+\vec{e}_\phi r\mathrm{d}\phi) \tag{6.5.274}$$

This allows us to get the relation:

$$\mathrm{d}f=\frac{\partial f}{\partial r}\mathrm{d}r+\frac{\partial f}{\partial \phi}\mathrm{d}\phi=\vec{\nabla}(f)\circ(\vec{e}_r\mathrm{d}r+\vec{e}_\phi r\mathrm{d}\phi)=\vec{\nabla}(f)_r\mathrm{d}r+\vec{\nabla}(f)_\phi r\mathrm{d}\phi \tag{6.5.275}$$

therefore:

$$\mathrm{d}f=\frac{\partial f}{\partial r}\mathrm{d}r+\frac{\partial f}{\partial \phi}\mathrm{d}\phi=\vec{\nabla}(f)_r\mathrm{d}r+\vec{\nabla}(f)_\phi r\mathrm{d}\phi \tag{6.5.276}$$

which bring us to:

$$\vec{\nabla}(f)_r=\frac{\partial f}{\partial r}\quad;\quad\vec{\nabla}(f)_\phi=\frac{1}{r}\frac{\partial f}{\partial \phi} \tag{6.5.277}$$

Thus the "gradient in polar coordinates" is expressed as:

$$\boxed{\vec{\nabla}=\left(\frac{\partial}{\partial r},\frac{1}{r}\frac{\partial f}{\partial \phi}\right)} \tag{6.5.278}$$

Let us now tackle the expression of the gradient in cylindrical coordinates. Let us recall that during our study of different coordinate systems we obtained for cylindrical coordinates:

$$\mathrm{d}\vec{r}=\vec{e}_r\mathrm{d}r+\vec{e}_\phi r\mathrm{d}\phi+\vec{e}_z\mathrm{d}z \tag{6.5.279}$$

So we already know that the expression of the gradient in cylindrical coordinates will be the same in polar coordinates with the exception of the addition of the vertical z component that is independent of other coordinates. Thus we get the "gradient in cylindrical coordinates":

$$\boxed{\vec{\nabla}=\left(\frac{\partial}{\partial r},\frac{1}{r}\frac{\partial f}{\partial \phi},\frac{\partial}{\partial z}\right)} \tag{6.5.280}$$

Let us now tackle on the expression of the gradient in spherical coordinates. Let us recall that during our study of the different coordinate systems we obtained for the spherical coordinates:

$$\mathrm{d}\vec{r}=\vec{e}_r\mathrm{d}r+\vec{e}_\theta r\mathrm{d}\theta+\vec{e}_\phi r\sin(\theta)\mathrm{d}\phi \tag{6.5.281}$$

But, with the definition of gradient in Cartesian coordinates, we have in spherical coordinates the following definition:

$$\mathrm{d}f=\vec{\nabla}(f)\circ\mathrm{d}\vec{r} \tag{6.5.282}$$

If we express the total differential of $\mathrm{d}f$ we get the following relations:

$$\mathrm{d}f=\frac{\partial f}{\partial r}\mathrm{d}r+\frac{\partial f}{\partial \phi}\mathrm{d}\phi+\frac{\partial f}{\partial \theta}\mathrm{d}\theta=\vec{\nabla}(f)\circ\mathrm{d}\vec{r}=\vec{\nabla}(f)\circ\mathrm{d}(\vec{e}_r\mathrm{d}r+\vec{e}_\theta r\mathrm{d}\theta+\vec{e}_\phi r\sin(\theta)\mathrm{d}\phi) \tag{6.5.283}$$

This allows us to get the relation (we now use the notation that uses the operator "nabla"):

$$\begin{aligned}\frac{\partial f}{\partial r}\mathrm{d}r+\frac{\partial f}{\partial \phi}\mathrm{d}\phi+\frac{\partial f}{\partial \theta}\mathrm{d}\theta&=\vec{\nabla}(f)\circ d(\vec{e}_r\mathrm{d}r+\vec{e}_\theta r\mathrm{d}\theta+\vec{e}_\phi r\sin(\theta)\mathrm{d}\phi)\\&=\vec{\nabla}(f)_r\mathrm{d}r+\vec{\nabla}(f)_\theta r\mathrm{d}\theta+\vec{\nabla}(f)_\phi r\sin(\theta)\mathrm{d}\phi\end{aligned} \tag{6.5.284}$$

The relation:

$$\frac{\partial f}{\partial r}\mathrm{d}r+\frac{\partial f}{\partial \phi}\mathrm{d}\phi+\frac{\partial f}{\partial \theta}\mathrm{d}\theta=\vec{\nabla}(f)_r\mathrm{d}r+\vec{\nabla}(f)_\theta r\mathrm{d}\theta+\vec{\nabla}(f)_\phi r\sin(\theta)\mathrm{d}\phi \tag{6.5.285}$$

requires that:

$$\vec{\nabla}(f)_r=\frac{\partial f}{\partial r}\quad;\quad\vec{\nabla}(f)_\theta=\frac{1}{r}\frac{\partial f}{\partial \phi}\quad;\quad\vec{\nabla}(f)_\phi=\frac{1}{r\sin(\theta)}\frac{\partial f}{\partial \theta} \tag{6.5.286}$$

Thus the "gradient in spherical coordinates" is expressed as:

$$\boxed{\vec{\nabla}=\left(\frac{\partial}{\partial r},\frac{1}{r}\frac{\partial f}{\partial \phi},\frac{1}{r\sin(\theta)}\frac{\partial}{\partial \theta}\right)} \tag{6.5.287}$$

So we finally saw all the expressions of the gradient operator in the Cartesian, polar, cylindrical and spherical systems.

✍**Example:**

Let us see now a visual example of the previous developments in Maple™ 4.00 and a special case function $f(x,y)=\sin(x)\sin(y)$.

```
> with(linalg):
> with(plots):
> plot3d(sin(x)*sin(y),x=-3..3,y=-3..3,axes=framed);
```

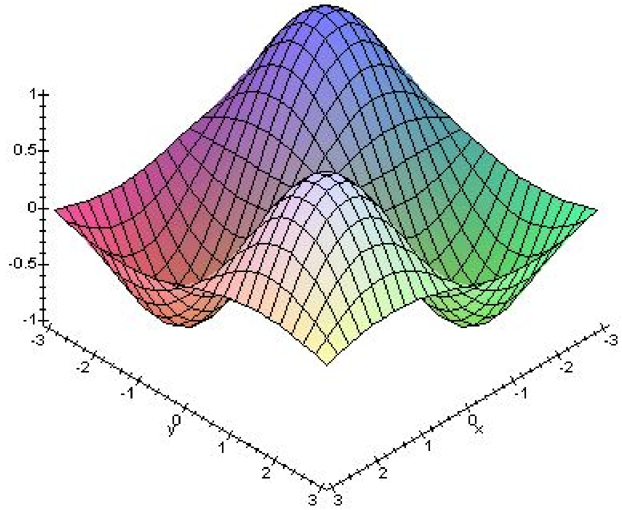


Figure 6.124: **Plot of the function taken as example**

And now we show the isolines:

```
>contourplot3d(sin(x)*sin(y),x=-3..3,y=-3..3,filled=true,
axes=framed,coloring=[red,blue],style=patch);
```

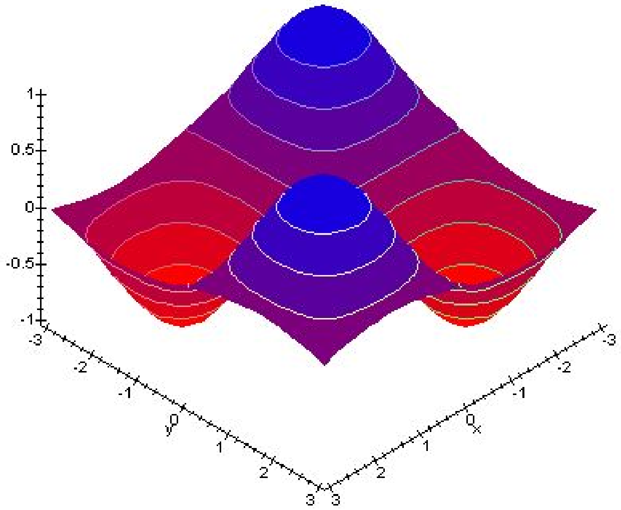


Figure 6.125: **Function with is isolines**

And now we plot the gradient vector with a plane projection and we see they are indeed perpendicular to the isolines:

```
>Pa:=contourplot(sin(x)*sin(y),x=-3..3,y=-3..3,contours=10,
coloring=[red,blue],filled=true):
>Pb:=fieldplot(grad(sin(x)*sin(y),vector([x,y])),x=-3..3,y=-3..3,
arrows=THICK):
>display(Pa,Pb);
```

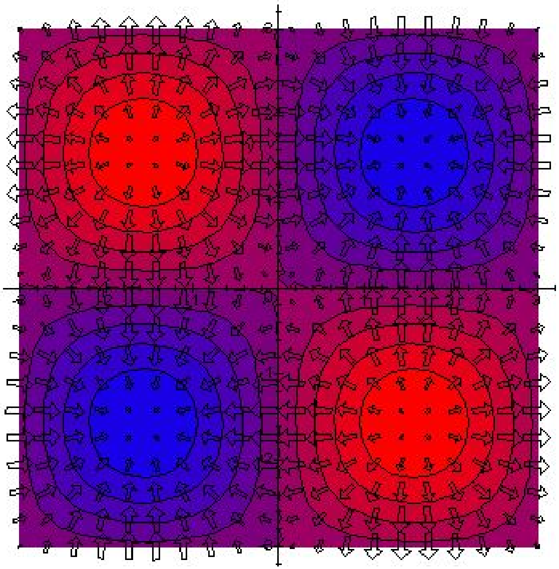


Figure 6.126: Function with its isolines and projected gradient

And with a 3D perspective:

```
>campo:=fieldplot3d([diff(sin(x)*sin(y),x),diff(sin(x)*sin(y),
y),0],x=-3..3,y=-3..3, z=-3..3,axes=framed,arrows=THICK);
>superf:=plot3d(sin(x)*sin(y),x=-3..3,y=-3..3):
> display(campo,superf);
```

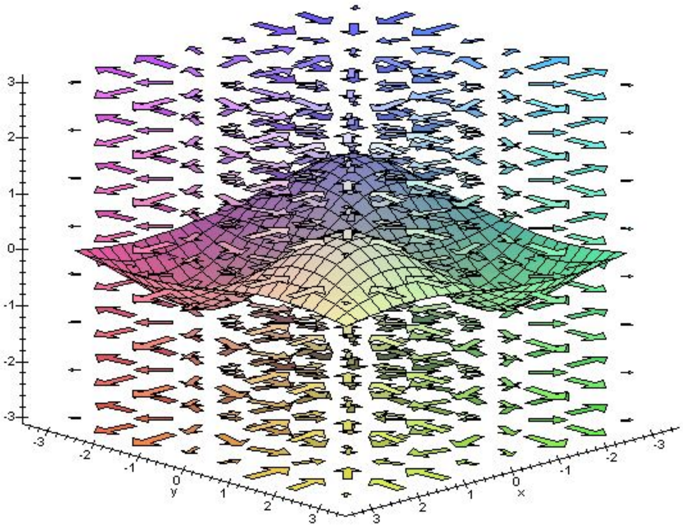


Figure 6.127: Function with its isolines and gradient in 3D

Seen from above with a small rotation:

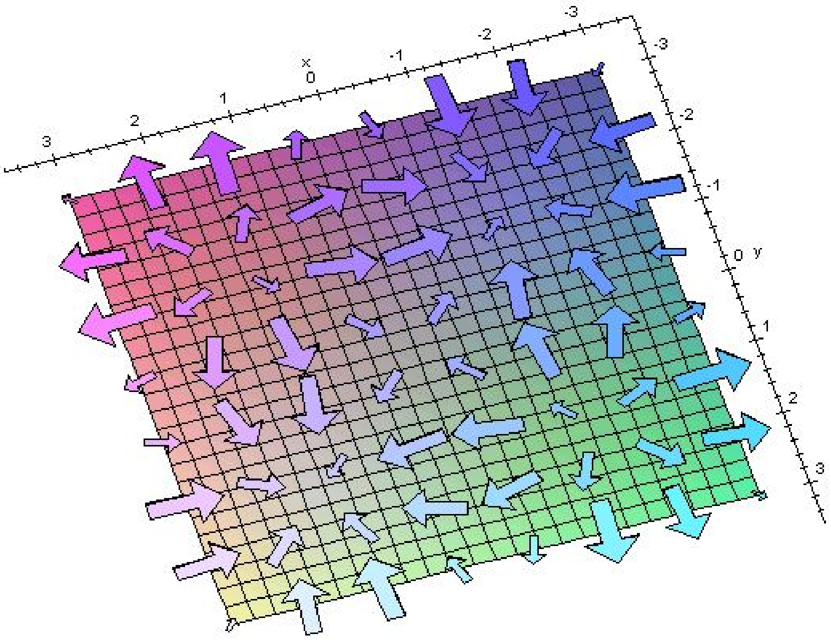


Figure 6.128: **Function rotation with its isolines and gradient**

As we will see during our study of the electromagnetic retarded potentials (see page 4168), if we define $\partial_t \vec{z} = \vec{z}$ then we get two results that may not be very intuitive for some readers. The first result is that:

$$\vec{\nabla}(z) = \frac{\vec{z}}{z}$$

and:

$$\frac{\vec{z}}{z} v = \vec{v}$$

6.5.8.2 Gradients of Vector Field

The gradient of a vector field $\vec{f}(x,y,z) : \mathbb{R}^3 \mapsto \mathbb{R}^3$ is the field named "tensor field" defined by the following nine relations in Cartesian coordinates:

$$\vec{\nabla}(\vec{f}) = \frac{\partial f_i}{\partial x_j} = \begin{bmatrix} \dfrac{\partial f_1}{\partial x} & \dfrac{\partial f_1}{\partial y} & \dfrac{\partial f_1}{\partial z} \\ \dfrac{\partial f_2}{\partial x} & \dfrac{\partial f_2}{\partial y} & \dfrac{\partial f_2}{\partial z} \\ \dfrac{\partial f_3}{\partial x} & \dfrac{\partial f_3}{\partial y} & \dfrac{\partial f_3}{\partial z} \end{bmatrix} \tag{6.5.288}$$

We will use such a gradient in our study in the section of Marine & Weather Engineering of the Papillon effect whose origin comes from the determination of the Navier-Stokes equations of the section of Continuum Mechanics and we will also use this type of gradient in the section of Numerical Methods in our study of the Gauss-Newton optimization method.

We have the following 4 components in polar coordinates:

$$\vec{\nabla}(\vec{f}) = \begin{bmatrix} \dfrac{\partial f_1}{\partial r} & \dfrac{1}{r}\dfrac{\partial f_1}{\partial \phi} \\ \dfrac{\partial f_2}{\partial r} & \dfrac{1}{r}\dfrac{\partial f_2}{\partial \phi} \end{bmatrix} \tag{6.5.289}$$

We have the following 9 components in cylindrical coordinates:

$$\vec{\nabla}(\vec{f})\begin{bmatrix}\dfrac{\partial f_1}{\partial r} & \dfrac{1}{r}\dfrac{\partial f_1}{\partial \phi} & \dfrac{\partial f_1}{\partial z}\\ \dfrac{\partial f_2}{\partial r} & \dfrac{1}{r}\dfrac{\partial f_2}{\partial \phi} & \dfrac{\partial f_2}{\partial z}\\ \dfrac{\partial f_3}{\partial r} & \dfrac{1}{r}\dfrac{\partial f_3}{\partial \phi} & \dfrac{\partial f_3}{\partial z}\end{bmatrix} \tag{6.5.290}$$

We have the following 9 components in spherical coordinates:

$$\vec{\nabla}(\vec{f})\begin{bmatrix}\dfrac{\partial f_1}{\partial r} & \dfrac{1}{r}\dfrac{\partial f_1}{\partial \theta} & \dfrac{1}{r\sin(\theta)}\dfrac{\partial f_1}{\partial \phi}\\ \dfrac{\partial f_2}{\partial r} & \dfrac{1}{r}\dfrac{\partial f_2}{\partial \theta} & \dfrac{1}{r\sin(\theta)}\dfrac{\partial f_2}{\partial \phi}\\ \dfrac{\partial f_3}{\partial r} & \dfrac{1}{r}\dfrac{\partial f_3}{\partial \theta} & \dfrac{1}{r\sin(\theta)}\dfrac{\partial f_3}{\partial \phi}\end{bmatrix} \tag{6.5.291}$$

So we finally saw all the expressions of a gradient vector field in Cartesian , polar, cylindrical and spherical system.

6.5.8.3 Divergences of a Vector Field

The divergence is applied to a vector field and turns it into a scalar field. Therefore it is an application from $\mathbb{R}^3 \mapsto \mathbb{R}$. Intuitively, and in the most common case, the divergence of a vector field expressed its tendency to come from or converge to some points.

Remark

Non-initiated people often confuse the gradient and divergence operator. To make the difference we must remember that the divergence of a vector is a number and that the gradient is a vector! The gradient indicates the direction in which the change is the most important and its amplitude. The divergence simply say what comes in or out from a given point.

However, we must distinguish two contributions to the divergence that we will rigorously define a little further below: one due to the variations named the "directional divergence" and the other due to variations in modules (norm) named the "modular divergence". Thus, for simple fields, we can imagine cases where the divergence would only be modular and others, where it would only be directional. We could also build a field where the two types of divergence coexist, but having adverse effects.

Let us consider for example a vector $\vec{f}$ of space and we make it pass through any surface S. Physicists then assimilate the quantity $\vec{f}$ which moves along the normal vector to the surface as a flow of $\vec{f}$ through S.

To be convinced of this analogy we can imagine a fluid flowing on a flat surface, the flow through the surface is obviously zero in this case, by cons if the fluid flows vertically through a horizontal surface the flow will be maximal. It is then immediate that we want to represent the flow by the scalar product of $\vec{f}$ with the normal $\vec{n}$ to the surface S.

Remark

We must always pay attention to the direction of $\vec{n}$ because at any point of a surface S we have in general two normal vectors $\vec{n}$ that are colinear but of opposite directions.

If the surface is planar then the normal $\vec{n}$ is the same everywhere, but if it changes from place to place, then we will look at a small surface element $\mathrm{d}S$.

If a small flow element is defined by:

$$\mathrm{d}\Phi = (\vec{f} \circ \vec{n})\mathrm{d}S \tag{6.5.292}$$

then the total flow will be given by:

$$\Phi = \int_S (\vec{f} \circ \vec{n})\mathrm{d}S \tag{6.5.293}$$

which is sometimes written (it's a little bit abusive but why not...):

$$\Phi = \int_S \vec{f}\mathrm{d}\vec{S} \tag{6.5.294}$$

Let us now suppose that our vector $\vec{f}$ moves a point $M(x,y,z)$ in space to $M'(x+\mathrm{d}x, y+\mathrm{d}y, z+\mathrm{d}z)$ through a rectangular parallelepiped of sides of $\mathrm{d}x, \mathrm{d}y$ and $\mathrm{d}z$:

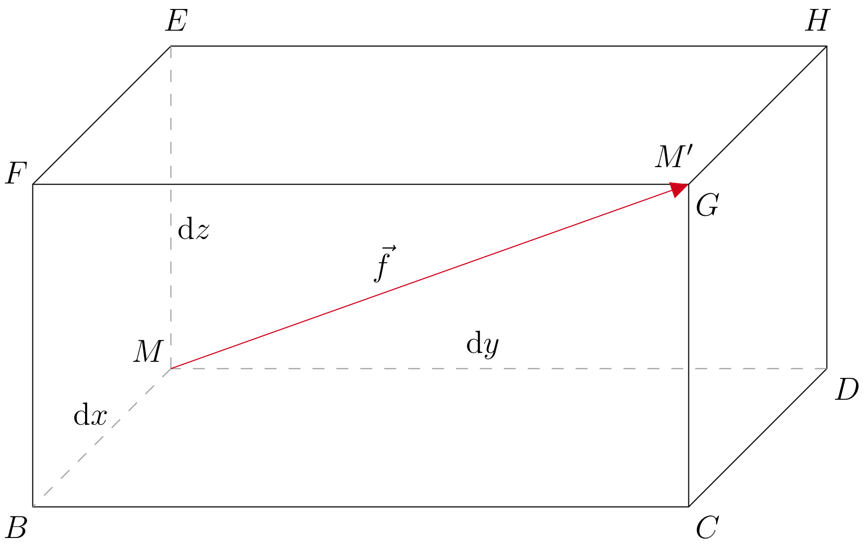


Figure 6.129: **Move of a vector through a parallelepiped**

We can decompose the movement (flow) through each face of the parallelepiped (decompositions in the orthonormal basis). For example, if we are interested at the decomposed part of the flow through the face $(\mathrm{d}y, \mathrm{d}z)$ described by the peaks vertices $BCFG$ we have obviously $\vec{n} = (1,0,0)$.

We still need to determine how to represent the flow $\vec{f}$ for this direction. As the flow is a function, that is to say that each of its components may be dependent of the three components of the space (if we take the case of a function in $\mathbb{R}^3$) we have:

$$\vec{f}=\begin{pmatrix} f_x(x,y,z)\\ f_y(x,y,z)\\ f_z(x,y,z)\end{pmatrix} \tag{6.5.295}$$

Remark

Those who are not convinced can go read the beginning of section Electrodynamics where we take the electric field as an (excellent) example (see page 3969).

While the variation of the flow according to x is given by:

$$\vec{f}_x=\begin{pmatrix} f_x(x+\mathrm{d}x,y,z)\\ f_y(x+\mathrm{d}x,y,z)\\ f_z(x+\mathrm{d}y,y,z)\end{pmatrix}-\begin{pmatrix} f_x(x,y,z)\\ f_y(x,y,z)\\ f_z(x,y,z)\end{pmatrix} \tag{6.5.296}$$

which give us:

$$\begin{aligned}
\mathrm{d}\Phi_x&=((\vec{f})_x\circ\vec{n}_x)\mathrm{d}y\mathrm{d}z=\Big(f_x(x+\mathrm{d}x,y,z)-f_x(x,y,z)\Big)\mathrm{d}y\mathrm{d}z=\frac{\partial f_x}{\partial x}\mathrm{d}x\mathrm{d}y\mathrm{d}z\\
\mathrm{d}\Phi_y&=((\vec{f})_y\circ\vec{n}_y)\mathrm{d}x\mathrm{d}z=\Big(f_y(x,y+\mathrm{d}y,z)-f_x(x,y,z)\Big)\mathrm{d}x\mathrm{d}z=\frac{\partial f_y}{\partial y}\mathrm{d}x\mathrm{d}y\mathrm{d}z\\
\mathrm{d}\Phi_z&=((\vec{f})_z\circ\vec{n}_z)\mathrm{d}x\mathrm{d}y=\Big(f_z(x,y,z+\mathrm{d}z)-f_x(x,y,z)\Big)\mathrm{d}x\mathrm{d}y=\frac{\partial f_z}{\partial z}\mathrm{d}x\mathrm{d}y\mathrm{d}z
\end{aligned} \tag{6.5.297}$$

therefore by summing:

$$\mathrm{d}\Phi=\left(\frac{\partial f_x}{\partial x}+\frac{\partial f_y}{\partial y}+\frac{\partial f_z}{\partial z}\right)\mathrm{d}x\mathrm{d}y\mathrm{d}z=\mathrm{div}(\vec{f})\mathrm{d}V=\vec{\nabla}\circ\vec{f}\mathrm{d}V \tag{6.5.298}$$

Compared to the first expression of Φ, the term $\mathrm{d}x\mathrm{d}y\mathrm{d}z$ is then a volume element and not more of surface. We also have an interesting result:

$$\Phi=\int\limits_V\left(\vec{\nabla}\circ\vec{f}\right)\mathrm{d}V=\int\limits_S\left(\vec{f}\circ\vec{n}\right)\mathrm{d}S \tag{6.5.299}$$

whose more explicit and rigorous writing should be (to highlight well that the considered closed surface is the boundary of the closed studied volume):

$$\Phi=\iiint_V(\vec{\nabla}\circ\vec{f})\mathrm{d}V=\oiint_S(\vec{f}\circ\vec{n})\mathrm{d}x\mathrm{d}y \tag{6.5.300}$$

or more commonly written:

$$\boxed{\Phi=\iiint_V(\vec{\nabla}\circ\vec{f})\mathrm{d}V=\oiint_S\vec{f}\circ\mathrm{d}\vec{S}} \tag{6.5.301}$$

Remark

See the practical cases in the section Electrodynamics where for example the electric field divergence is zero for a free spherical charge as the vectors point in different directions (directional divergence) and where the norms decrease as the inverse of the square of the radius (modular convergence). Both contributions are in opposition and so the total divergence is zero.

The development above is named "Ostrogradsky theorem" or "Gauss-Ostrogradsky theorem" or more simply "divergence theorem" and actually defines the total divergence of $\vec{f}$ in a volume as the flow $\vec{f}$ through the "walls" of the closed volume (Gauss closed surface), which expresses well the name "divergence".

Now reconsider the previous relation but extracting a unitary vector from the vector field $\vec{f}$ such that:

$$\iiint_V (\vec{\nabla}\circ(f\vec{u}))\mathrm{d}V=\oiint_S (f\vec{u})\circ\mathrm{d}\vec{S} \tag{6.5.302}$$

where now f is a scalar field. This can be rewritten obviously using chain rule derivatives and dot product commutativity:

$$\iiint_V (\vec{\nabla}(f)\circ\vec{u}+f\vec{\nabla}\circ\vec{u})\mathrm{d}V=\left(\oiint_S f\mathrm{d}\vec{S}\right)\circ\vec{u} \tag{6.5.303}$$

In the special case of a uniform field we have:

$$\vec{\nabla}\circ\vec{u}=0 \tag{6.5.304}$$

Then it remains:

$$\iiint_V \vec{\nabla}(f)\circ\vec{u}\mathrm{d}V=\left(\oiint_S f\mathrm{d}\vec{S}\right)\circ\vec{u} \tag{6.5.305}$$

The dot product being distributive on the sum of vectors we can rewrite this as:

$$\left(\iiint_V \vec{\nabla}(f)\mathrm{d}V\right)\circ\vec{u}=\left(\oiint_S f\mathrm{d}\vec{S}\right)\circ\vec{u} \tag{6.5.306}$$

and therefore we get the "gradient theorem[17]":

$$\boxed{\iiint_V \vec{\nabla}(f)\mathrm{d}V=\oiint_S f\mathrm{d}\vec{S}} \tag{6.5.307}$$

We define the operator "divergence" by the following relation (the tensor notation has been used to shorten the writing) in an n-dimensional space:

$$\mathrm{div}(\vec{f})=\vec{\nabla}\circ\vec{f}=\sum_{i=0}^{n}\frac{\partial f_i}{\partial x_i} \tag{6.5.308}$$

---

[17] As it is applicable only for uniform field it is not used a lot in practice. Its most well-known application is to prove the Archimedes principle in the most general case.

Thus we have for the operator "divergence in Cartesian coordinates":

$$\boxed{\vec{\nabla}=\left(\frac{\partial}{\partial x},\frac{\partial}{\partial y},\frac{\partial}{\partial z}\right)} \tag{6.5.309}$$

If the divergence of a vector field is identically zero in all the points of an Eulerian frame[18], the triple integral flux of this field through a volume V will be:

$$\Phi=\iiint_V(\vec{\nabla}\circ\vec{f})\mathrm{d}V=0 \tag{6.5.310}$$

It follows that the flow of this vector field through the edges of the volume is zero, that is to say that the incoming flow compensates the output flow. We say that such a field vectors having a null divergence has a "conservation flow".

To determine the expression of the divergence operator in polar coordinates let us recall the relations proved earlier above:

$$\begin{aligned}\vec{e}_r&=\frac{\vec{r}}{r}=\frac{x\vec{e}_x+y\vec{e}_y}{r}=\vec{e}_x\cos(\phi)+\vec{e}_y\sin(\phi)\\ \vec{e}_\phi&=\frac{\vec{e}_z\times\vec{e}_r}{\sin(\theta)}=-\vec{e}_x\sin(\phi)+\vec{e}_y\cos(\phi)\end{aligned} \tag{6.5.311}$$

Given now a vectorial function $\vec{f}:\mathbb{R}^2\rightarrow\mathbb{R}^2$. We have:

$$\vec{f}=f_x\vec{e}_x+f_y\vec{e}_y=f_r\vec{e}_r+f_\theta\vec{e}_\theta \tag{6.5.312}$$

Knowing the expression of $\vec{e}_r,\vec{e}_\phi$ depending on $\vec{e}_x,\vec{e}_y$, from the expression above we deduce:

$$\begin{aligned}f_x&=\cos(\phi)f_r-\sin(\phi)f_\phi\\ f_y&=\sin(\phi)f_r+\cos(\phi)f_\phi\end{aligned} \tag{6.5.313}$$

The divergence of $\vec{f}$ being defined in the two dimensional case by:

$$\vec{\nabla}\circ\vec{f}=\frac{\partial f_x}{\partial x}+\frac{\partial f_y}{\partial y} \tag{6.5.314}$$

we then have:

$$\vec{\nabla}\circ\vec{f}=\vec{\nabla}(f_x)\circ\vec{e}_x+\vec{\nabla}(f_y)\circ\vec{e}_y \tag{6.5.315}$$

The first term is (application of the gradient in polar coordinates!):

$$\begin{aligned}\vec{\nabla}(f_x)\circ\vec{e}_x&=\left(\frac{\partial f_x}{\partial r}\vec{e}_r+\frac{1}{r}\frac{\partial f_x}{\partial\phi}\vec{e}_\phi\right)\circ\vec{e}_x\\ &=\frac{\partial f_x}{\partial r}\vec{e}_r\circ\vec{e}_x+\frac{1}{r}\frac{\partial f_x}{\partial\phi}\vec{e}_\phi\circ\vec{e}_x\\ &=\frac{\partial f_x}{\partial r}\cos(\phi)-\frac{1}{r}\frac{\partial f_x}{\partial\phi}\sin(\phi)\end{aligned} \tag{6.5.316}$$

---

[18] The Eulerian specification of the flow field is a way of looking at fluid motion that focuses on specific locations in the space through which the fluid flows as time passes. This can be visualized by sitting on the bank of a river and watching the water pass the fixed location.

in the same way we get for the second term (we can as always give more details on request):

$$\begin{aligned}\vec{\nabla}(f_y)\circ\vec{e}_y&=\frac{\partial f_y}{\partial r}\vec{e}_r\circ\vec{e}_y+\frac{1}{r}\frac{\partial f_y}{\partial\phi}\vec{e}_\phi\circ\vec{e}_y\\&=\frac{\partial f_y}{\partial r}\sin(\phi)+\frac{1}{r}\frac{\partial f_y}{\partial\phi}\cos(\phi)\end{aligned}\tag{6.5.317}$$

By adding the two terms and expressing the partial derivatives of the functions f_x, f_y in function of the partial derivatives of the functions f_r, f_ϕ using the relations:

$$\begin{aligned}f_x&=\cos(\phi)f_r-\sin(\phi)f_\phi\\f_y&=\sin(\phi)f_r+\cos(\phi)f_\phi\end{aligned}\tag{6.5.318}$$

We get:

$$\begin{aligned}\vec{\nabla}\circ\vec{f}=&\frac{\partial f_x}{\partial r}\cos(\phi)-\frac{1}{r}\frac{\partial f_x}{\partial(\phi)}\sin(\phi)+\frac{\partial f_y}{\partial r}\sin(\phi)+\frac{1}{r}\frac{\partial f_y}{\partial\phi}\cos(\phi)\\=&\cos^2(\phi)\frac{\partial f_r}{\partial r}-\cos(\phi)\sin(\phi)\frac{\partial f_\phi}{\partial r}-\frac{1}{r}\sin(\phi)\left(-\sin(\phi)f_r+\cos(\phi)\frac{\partial f_r}{\partial\phi}-\cos(\phi)f_\phi-\sin(\phi)\frac{\partial f_\phi}{\partial\phi}\right)\\&+\sin^2(\phi)\frac{\partial f_r}{\partial r}+\sin(\phi)\cos(\phi)\frac{\partial f_\phi}{\partial r}+\frac{1}{r}\cos(\phi)\left(\cos(\phi)f_r+\sin\phi\frac{\partial f_r}{\partial\phi}-\sin(\phi)f_\phi+\cos(\phi)\frac{\partial f_\phi}{\partial\phi}\right)\end{aligned}\tag{6.5.319}$$

After simplification:

$$\vec{\nabla}\circ\vec{f}=\frac{\partial f_r}{\partial r}+\frac{1}{r}f_r+\frac{1}{r}\frac{\partial f_\phi}{\partial\phi}=\frac{1}{r}\frac{\partial(rf_r)}{\partial r}+\frac{1}{r}\frac{\partial f_\phi}{\partial\phi}\tag{6.5.320}$$

The expression of the operator "divergence in polar coordinates" is then:

$$\boxed{\vec{\nabla}=\left(\frac{1}{r}\frac{\partial(r}{\partial r},\frac{1}{r}\frac{\partial}{\partial\phi}\right)}\tag{6.5.321}$$

To determine the expression of the divergence operator in cylindrical coordinates let us recall the relations:

$$\begin{aligned}\vec{e}_r&=\frac{\vec{r}}{r}=\frac{x\vec{e}_x+y\vec{e}_y}{r}=\vec{e}_x\cos(\phi)+\vec{e}_y\sin(\phi)\\\vec{e}_\phi&=\frac{\vec{e}_z\times\vec{e}_r}{\sin(\phi)}=-\vec{e}_x\sin(\phi)+\vec{e}_y\cos(\phi)\\\vec{e}_z&=\vec{e}_z\end{aligned}\tag{6.5.322}$$

Given now a vector function $\vec{f}:\mathbb{R}^3\rightarrow\mathbb{R}^3$. We have:

$$\vec{f}=f_x\vec{e}_x+f_y\vec{e}_y+f_z\vec{e}_z=f_r\vec{e}_r+f_\phi\vec{e}_\phi+f_z\vec{e}_z\tag{6.5.323}$$

As we know the expressions of $\vec{e}_r,\vec{e}_\phi,\vec{e}_z$ in function of the $\vec{e}_x,\vec{e}_y,\vec{e}_z$, from the above expressions we deduce:

$$\begin{aligned}f_x&=\cos(\phi)f_r-\sin(\phi)f_\phi\\f_y&=\sin(\phi)f_r+\cos(\phi)f_\phi\\f_z&=f_z\end{aligned}\tag{6.5.324}$$

The divergence of $\vec{f}$ being defined in the three dimensional case by:

$$\vec{\nabla}\circ\vec{f}=\frac{\partial f_x}{\partial x}+\frac{\partial f_y}{\partial y}+\frac{\partial f_z}{\partial z} \tag{6.5.325}$$

we then have:

$$\vec{\nabla}\circ\vec{f}=\vec{\nabla}(f_x)\circ\vec{e}_x+\vec{\nabla}(f_y)\circ\vec{e}_y+\vec{\nabla}(f_z)\circ\vec{e}_z \tag{6.5.326}$$

The first term is equal to (application of the gradient in cylindrical coordinates):

$$\begin{aligned}\vec{\nabla}(f_x)\circ\vec{e}_x&=\left(\frac{\partial f_x}{\partial r}\vec{e}_r+\frac{1}{r}\frac{\partial f_x}{\partial\phi}\vec{e}_\phi+\frac{\partial f_x}{\partial z}\vec{e}_z\right)\circ\vec{e}_x\\&=\frac{\partial f_x}{\partial r}\vec{e}_r\circ\vec{e}_x+\frac{1}{r}\frac{\partial f_x}{\partial\phi}\vec{e}_\phi\circ\vec{e}_x+\frac{\partial f_x}{\partial z}\vec{e}_z\circ\vec{e}_x\\&=\frac{\partial f_x}{\partial r}\cos(\phi)-\frac{1}{r}\frac{\partial f_x}{\partial\phi}\sin(\phi)\end{aligned} \tag{6.5.327}$$

in the same way we get for the second component (we can give the details on request):

$$\begin{aligned}\vec{\nabla}(f_y)\circ\vec{e}_y&=\frac{\partial f_y}{\partial r}\vec{e}_r\circ\vec{e}_y+\frac{1}{r}\frac{\partial f_y}{\partial\phi}\vec{e}_\phi\circ\vec{e}_y+\frac{\partial f_y}{\partial z}\vec{e}_z\circ\vec{e}_y\\&=\frac{\partial f_y}{\partial r}\sin(\phi)+\frac{1}{r}\frac{\partial f_y}{\partial\phi}\cos(\phi)\end{aligned} \tag{6.5.328}$$

and for the last one:

$$\begin{aligned}\vec{\nabla}(f_z)\circ\vec{e}_z&=\frac{\partial f_z}{\partial r}\vec{e}_r\circ\vec{e}_z+\frac{1}{r}\frac{\partial f_z}{\partial\phi}\vec{e}_\phi\circ\vec{e}_z+\frac{\partial f_z}{\partial z}\vec{e}_z\circ\vec{e}_z\\&=\frac{\partial f_z}{\partial z}\end{aligned} \tag{6.5.329}$$

By adding the three terms and expressing the partial derivatives of the functions f_x,f_y,f_z in function of the partial derivatives of the functions f_r,f_ϕ,f_z using the relations:

$$\begin{aligned}f_x&=\cos(\phi)f_r-\sin(\phi)f_\phi\\f_y&=\sin(\phi)f_r+\cos(\phi)f_\phi\\f_z&=f_z\end{aligned} \tag{6.5.330}$$

we get:

$$\begin{aligned}&\vec{\nabla}\circ\vec{f}=\frac{\partial f_x}{\partial r}\cos(\phi)-\frac{1}{r}\frac{\partial f_x}{\partial\phi}\sin(\phi)+\frac{\partial f_y}{\partial r}\sin(\phi)+\frac{1}{r}\frac{\partial f_y}{\partial\phi}\cos(\phi)+\frac{\partial f_z}{\partial z}\\&=\cos^2(\phi)\frac{\partial f_r}{\partial r}-\cos(\phi)\sin(\phi)\frac{\partial f_\phi}{\partial r}-\frac{1}{r}\sin(\phi)\left(-\sin(\phi)f_r+\cos(\phi)\frac{\partial f_r}{\partial\phi}-\cos(\phi)f_\phi-\sin(\phi)\frac{\partial f_\phi}{\partial\phi}\right)\\&+\sin^2(\phi)\frac{\partial f_r}{\partial r}+\sin(\phi)\cos(\phi)\frac{\partial f_\phi}{\partial r}+\frac{1}{r}\cos(\phi)\left(\cos\phi f_r+\sin(\phi)\frac{\partial f_r}{\partial\phi}-\sin(\phi)f_\phi+\cos(\phi)\frac{\partial f_\phi}{\partial\phi}\right)+\frac{\partial f_z}{\partial z}\end{aligned} \tag{6.5.331}$$

After simplification:

$$\vec{\nabla}\circ\vec{f}=\frac{f_r}{\partial r}+\frac{1}{r}f_r+\frac{1}{r}\frac{f_\phi}{\partial\phi}+\frac{f_z}{\partial z}=\frac{1}{r}\frac{\partial(rf_r)}{\partial r}+\frac{1}{r}\frac{f_\phi}{\partial\phi}+\frac{f_z}{\partial z} \tag{6.5.332}$$

The expression of the operator "divergence in cylindrical coordinates" is then:

$$\boxed{\vec{\nabla}=\left(\frac{1}{r}\frac{\partial(r}{\partial r},\frac{1}{r}\frac{\partial}{\partial\phi},\frac{\partial}{\partial z}\right)} \tag{6.5.333}$$

To find the expression of the divergence in spherical coordinates, let us recall the relations:

$$\begin{aligned}\vec{e}_r&=\sin(\theta)\cos(\phi)\vec{e}_x+\sin(\theta)\sin(\phi)\vec{e}_y+\cos(\theta)\vec{e}_z\\ \vec{e}_\phi&=-\vec{e}_x\sin(\phi)+\vec{e}_y\cos(\phi)\\ \vec{e}_\theta&=-\vec{e}_x\cos(\theta)\cos(\phi)-\vec{e}_y\cos(\theta)\sin(\phi)-\vec{e}_z\sin(\theta)\end{aligned} \tag{6.5.334}$$

Given now a vector function $\vec{f}:\mathbb{R}^3\rightarrow\mathbb{R}^3$. We have:

$$\vec{f}=f_x\vec{e}_x+f_y\vec{e}_y+f_z\vec{e}_z=f_r\vec{e}_r+f_\theta\vec{e}_\theta+f_\phi\vec{e}_\phi \tag{6.5.335}$$

Knowing the expression of $\vec{e}_r,\vec{e}_\theta,\vec{e}_\phi$ depending on $\vec{e}_x,\vec{e}_y,\vec{e}_z$, from the expression above we deduce:

$$\begin{aligned}f_x&=\sin(\theta)\cos(\phi)f_r+\cos(\theta)\sin(\phi)f_\theta-\sin(\phi)f_\phi\\ f_y&=\sin(\theta)\sin(\phi)f_r+\cos(\theta)\sin(\phi)f_\theta+\cos(\phi)f_\phi\\ f_z&=\cos(\theta)f_r-\sin(\theta)f_\theta\end{aligned} \tag{6.5.336}$$

The divergence of $\vec{f}$ being defined in the three dimensional case by:

$$\vec{\nabla}\circ\vec{f}=\frac{\partial f_x}{\partial x}+\frac{\partial f_y}{\partial y}+\frac{\partial f_z}{\partial z} \tag{6.5.337}$$

we then have:

$$\vec{\nabla}\circ f=\vec{\nabla}(f_x)\circ\vec{e}_x+\vec{\nabla}(f_y)\circ\vec{e}_y+\vec{\nabla}(f_z)\circ\vec{e}_z \tag{6.5.338}$$

The first component is equal to (application of the gradient in spherical coordinates):

$$\begin{aligned}\vec{\nabla}(f_x)\circ\vec{e}_x&=\left(\frac{\partial f_x}{\partial r}\vec{e}_r+\frac{1}{r}\frac{\partial f_x}{\partial\theta}\vec{e}_\theta+\frac{1}{r\sin(\theta)}\frac{\partial f_x}{\partial\phi}\vec{e}_\phi\right)\circ\vec{e}_x\\ &=\frac{\partial f_x}{\partial r}\vec{e}_r\circ\vec{e}_x+\frac{1}{r}\frac{\partial f_x}{\partial\theta}\vec{e}_\theta\circ\vec{e}_x+\frac{1}{r\sin(\theta)}\frac{\partial f_x}{\partial\phi}\vec{e}_\phi\circ\vec{e}_x\\ &=\frac{\partial f_x}{\partial r}\sin(\theta)\cos(\phi)+\frac{1}{r}\frac{\partial f_x}{\partial\theta}\cos(\theta)\cos(\phi)-\frac{\sin(\phi)}{r\sin(\theta)}\frac{\partial f_x}{\partial\phi}\end{aligned} \tag{6.5.339}$$

in the same way we get for the second component (we can give the details on request):

$$\vec{\nabla}(f_y)\circ\vec{e}_y=\frac{\partial f_y}{\partial r}\sin(\theta)\sin(\phi)+\frac{1}{r}\frac{\partial f_y}{\partial\theta}\cos(\theta)\sin(\phi)+\frac{\cos(\phi)}{r\sin\theta}\frac{\partial f_y}{\partial\phi} \tag{6.5.340}$$

and finally for the third and last one:

$$\vec{\nabla}(f_z)\circ\vec{e}_z=\frac{\partial f_z}{\partial r}\cos(\theta)-\frac{1}{r}\frac{\partial f_z}{\partial\theta}\sin(\theta) \tag{6.5.341}$$

By adding the three terms and expressing the partial derivatives of the functions f_x,f_y,f_z in function of the partial derivatives of the functions f_r,f_θ,f_ϕ using the relations:

$$\begin{aligned}
f_x&=\sin(\theta)\cos(\phi)f_r+\cos(\theta)\cos(\phi)f_\theta-\sin(\phi)f_\phi\\
f_y&=\sin(\theta)\sin(\phi)f_r+\cos(\theta)\sin(\phi)f_\theta+\cos(\phi)f_\phi\\
f_z&=\cos(\theta)f_r-\sin(\theta)f_\theta
\end{aligned}\tag{6.5.342}$$

we get (we can develop the intermediate details on request):

$$\vec{\nabla}\circ\vec{f}=\frac{1}{r\sin(\theta)}\frac{\partial f_\phi}{\partial\phi}+\frac{\cos(\theta)}{r\sin(\theta)}f_\theta+\frac{1}{r}\frac{\partial f_\theta}{\partial\theta}+\frac{\partial f_r}{\partial r}+\frac{2}{r}f_r\tag{6.5.343}$$

We notice that we can regroup terms depending on the same variable using the property of the derivative, so we get for the expression of the divergence in spherical coordinates:

$$\vec{\nabla}\circ\vec{f}=\frac{1}{r\sin(\theta)}\frac{\partial f_\phi}{\partial\phi}+\frac{1}{r\sin(\theta)}\frac{\partial(f_\theta\sin(\theta))}{\partial\theta}+\frac{1}{r^2}\frac{\partial(r^2f_r)}{\partial r}\tag{6.5.344}$$

and therefore the operator of "divergence in spherical coordinates" is:

$$\boxed{\vec{\nabla}=\left(\frac{1}{r^2}\frac{\partial(r^2}{\partial r},\frac{1}{r\sin(\theta)}\frac{\partial(\sin(\theta)}{\partial\theta},\frac{1}{r\sin(\theta)}\frac{\partial}{\partial\phi}\right)}\tag{6.5.345}$$

So we finally saw all the expressions of the divergence operator of a vector field in Cartesian, polar, cylindrical and spherical systems.

Keep in mind that way to "see" divergence on a vector field plot is to look at what happens to the magnitude of vectors as you move along the flow of the vector field. If the vector field is increasing in magnitude as you move along the flow of a vector field, then the divergence is positive. If the vector field is decreasing in magnitude as you move along the flow of a vector field, then the divergence is negative. If the vector field does not change in magnitude as you move along the flow of the vector field, then the divergence is zero. Also, remember that the divergence of a vector field is often a variable quantity and will change depending on location!

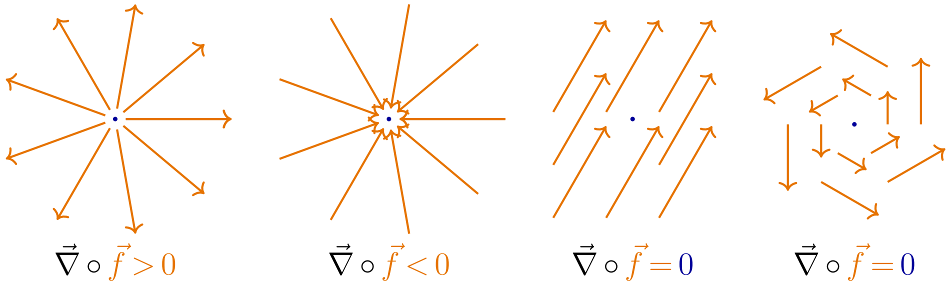

$\vec{\nabla}\circ\vec{f}>0$
$\vec{\nabla}\circ\vec{f}<0$
$\vec{\nabla}\circ\vec{f}=0$
$\vec{\nabla}\circ\vec{f}=0$

Figure 6.130: **Various configurations of divergence (author: Izaak Neutelings)**

6.5.8.4 Rotationals of a Vector Field (Curl)

The "curl" or "rotationnal" of a vector field can be seen (this is a simplification!) as the vector field whose field lines are perpendicular to those we have calculated the rotational as shown in the special example below (we will see more academic detailed further below):

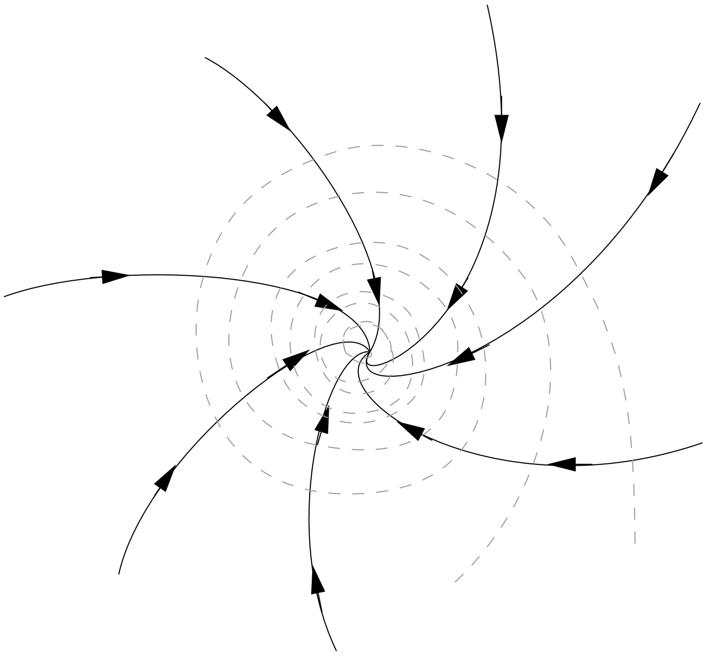


Figure 6.131: **Example of rotational (curl) of a vector field**

In a little bit more technical way the rotational is a vector operator that describes the infinitesimal rotation of a 3-dimensional vector field. At every point in the field, the rotational of that point is represented by a vector. The attributes of this vector (length and direction) characterize the rotation at that point. he direction of the rotational is the axis of rotation, as determined by the right-hand rule, and the magnitude of the rotational is the magnitude of rotation.

The rotational transforms a vector field in another vector field. For most people it is more difficult to accurately represent than the gradient and divergence, it intuitively reflects the tendency of a field to rotate around a point (the way it is twisted).

Let us give before tackling with the mathematical stuff and also mathematical examples two every-day life examples:

✍**Examples:**

E1. In a tornado, the wind turns around the eye of the storm and the wind velocity vector field has a non-zero rotational around the eye.

E2. The rotational of the velocity field of a disc that rotates at a constant speed is constant, directed along the axis of rotation and oriented such that the rotation takes place, in relation to it, in the direct sense.

A vector field is said to be "irrotational" when the rotational of this field is identically zero at all points of space. Otherwise, we say it is a "vortex".

In the usual case where $\mathrm{d}x$ is an element of length, the measurement unit of the rotational is then the unit of the considerated field divided by a unit length. For example, in fluid mechanics: the unity of the rotational of a velocity field is radians per unit time, as an angular velocity as we divide a velocity ([ms^{-1}] by the length [m]!

The divergence gives some indication of the behaviour of a vector or a vector field: how it moves in relation to the normal and how it crosses the surface, but it is insufficient. Take a field which would have the shape of a cylinder and another field which have a helicoidal form of the same diameter as the cylinder. If the move in the same direction the divergence will be the same even if the movements are quite different This requires that we determine how the field is bent as it passes through a surface: this will be determined by the circulation (as the work of a force, for example) of the vector along a closed curve, obtained with the sum of dot products $\vec{f}\circ\mathrm{d}\vec{r}$ on the closed contour (see section Differential and Integral Calculus page 1583):

$$\oint \vec{f}\circ\mathrm{d}\vec{r} \tag{6.5.346}$$

In fact it's the same to look at how twisted is the vector with respect to the normal vector of the surface which leads us to define the "rotational" or "swirl vector" by writing:

$$\oint \vec{f}\circ\mathrm{d}\vec{r}=\int\limits_S \mathrm{rot}(\vec{f})\circ\mathrm{d}\vec{S} \tag{6.5.347}$$

that thus establishes a relation between the line integral and the surface integral (we then transform a line integral on a closed path in a surface integral delimited by the given path).

Theorem 6.37. In other words, the rotational is calculated by using the fact that the flow around a closed basic path of a vector field is equal to the flux of its rotational through the immediate elementary surface generated by this path.

This is the "Stokes' theorem" (which is more rigorously demonstrable with a heavy mathematical formalism) which is in fact a definition of the rotational operator which we will seek the explicit mathematical expression right now!

Proof 6.37.1. Given $\vec{f}$ a vector field defined in a given space. We want to calculate the circulation of $\vec{f}$ around a closed path (contour) C:

$$\oint \vec{f}\circ\mathrm{d}\vec{r} \tag{6.5.348}$$

We choose for contour C the edge of an infinitesimal rectangle $(\mathrm{d}x,\mathrm{d}y)$ that is into $\mathbb{R}^3$ and parallel to the xy-plane (note that we travel the contour so as to always have the surface to our left!):

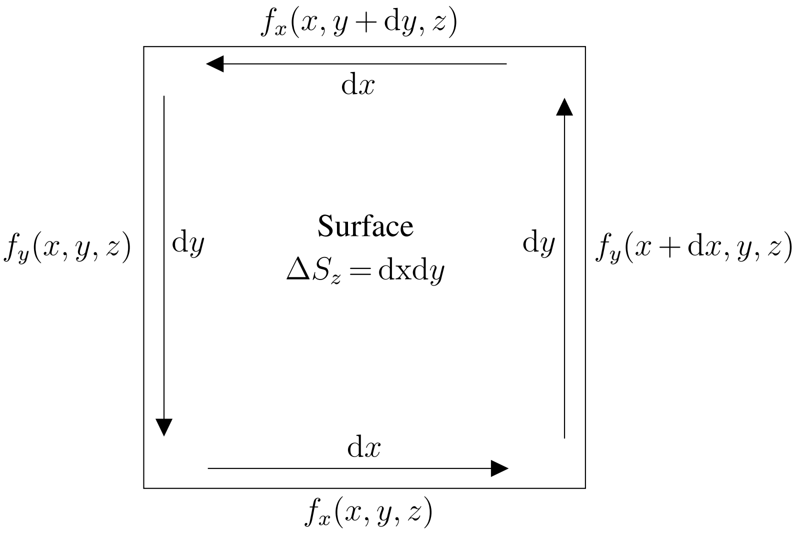


Figure 6.132: **Contour (path) of integration**

For the two horizontal sides (edges), the contribution to the circulation is:

$$f_x(x,y,z)\mathrm{d}x-f_x(x,y+\mathrm{d}y,z)\mathrm{d}x \tag{6.5.349}$$

Which authorizes us to write:

$$\frac{f_x(x,y,z)-f_x(x,y+\mathrm{d}y,z)}{\mathrm{d}y}\mathrm{d}x\mathrm{d}y=-\frac{\partial f_x}{\partial y}\mathrm{d}S_z \tag{6.5.350}$$

Same for the vertical sides (edges) we have also:

$$\frac{f_y(x+\mathrm{d}x,y,z)-f_y(x,y,z)}{\mathrm{d}x}\mathrm{d}x\mathrm{d}y=-\frac{\partial f_y}{\partial x}\mathrm{d}S_z \tag{6.5.351}$$

Therefore we have the circulation following z:

$$(\vec{f}\circ\mathrm{d}\vec{r})_z=\left(\frac{\partial f_y}{\partial x}-\frac{\partial f_x}{\partial y}\right)\mathrm{d}S_z \tag{6.5.352}$$

Which can also be written in the following more general and important form:

$$\boxed{\oint\limits_C(f_x\mathrm{d}x+f_y\mathrm{d}y)=\int\limits_S\left(\frac{\partial f_y}{\partial x}-\frac{\partial f_x}{\partial y}\right)\mathrm{d}S_z} \tag{6.5.353}$$

and that is nothing less than the famous "Green's theorem" or also known as the "Green-Riemann theorem" that we will see again in the section of Complex Analysis.

Green's theorem is a special case of Stokes (general manifold) theorem, but a generalisation to 2-dimension of the fundamental theorem of calculus. But stated in another way: The total spin outside of a function (closed curve) is equal to the sum of little spins on the inside of the function (closed curve).

And that we will write in the situation that interest us:

$$\int\limits_S\left(\frac{\partial f_y}{\partial x}-\frac{\partial f_x}{\partial y}\right)\mathrm{d}S_z:=\int\limits_S\mathrm{rot}(\vec{f})_z\mathrm{d}S_z \tag{6.5.354}$$

By circular permutation we then get:

$$\text{rot}(\vec{f}) = \left(\frac{\partial f_z}{\partial y} - \frac{\partial f_y}{\partial z}\right)\vec{e}_x + \left(\frac{\partial f_x}{\partial z} - \frac{\partial f_z}{\partial x}\right)\vec{e}_y + \left(\frac{\partial f_y}{\partial x} - \frac{\partial f_x}{\partial y}\right)\vec{e}_z \tag{6.5.355}$$

Either in vector condensed form:

$$\boxed{\text{rot}(\vec{f}) = \begin{pmatrix} \dfrac{\partial f_z}{\partial y} - \dfrac{\partial f_y}{\partial z} \\ \dfrac{\partial f_x}{\partial z} - \dfrac{\partial f_z}{\partial x} \\ \dfrac{\partial f_y}{\partial x} - \dfrac{\partial f_x}{\partial y} \end{pmatrix}} \tag{6.5.356}$$

This allows us to better understand the notation, or the non intuitive definition of the rotational in many books and that is:

$$\text{rot}(\vec{f}) = \vec{\nabla} \times \vec{f} \tag{6.5.357}$$

that is to say the cross product of the gradient operator by the vector field!

So finally we have proved the Stokes theorem or also named in this form "curl theorem" that gives well:

$$\boxed{\oint \vec{f} \circ \mathrm{d}\vec{r} = \int\limits_S \left(\vec{\nabla} \times \vec{f}\right) \circ \mathrm{d}\vec{S}} \tag{6.5.358}$$

and at the same time the rotational in Cartesian coordinates.

The above classical Stokes theorem can be stated in one sentence: The line integral of a vector field over a loop is equal to the flux of its curl through the enclosed surface.

■ Q.E.D.

Remark

The above relation is in fact not the "Stokes theorem" but a special case of it named the "Kelvin–Stokes theorem". The stokes theorem is given by (without proof):

$$\int_{\partial\Omega} \omega = \int_{\Omega} \mathrm{d}\omega \tag{6.5.359}$$

and says that the integral of a differential form ω over the boundary of some orientable manifold Ω is equal to the integral of its exterior derivative $\mathrm{d}\omega$ over the whole of Ω. It is the generalization of the fundamental theorem of calculus to N dimensions.

This is the most general and conceptually pure form of Stokes' theorem, of which the fundamental theorem of calculus (1D case), the fundamental theorem of line integrals, Green's theorem (2D case), Stokes' (original) theorem (3D case) and the divergence theorem are all special cases!

✍**Examples:**

E1. Let us take the vector field, which depends on x and y linearly:

$$\vec{f}(x,y,z)=y\vec{e}_x-x\vec{e}_y \tag{6.5.360}$$

Its plot look like this in Maple™ 4.00b:

```
 >with(DEtools):  with(plots):
>fieldplot([y, -x], x=-5..5, y=-5..5,arrows = medium,
color = sqrt(x ^2 + y^2),thickness=2,labels=[‘x‘,‘y‘],
title=‘Simple vector field‘);
```

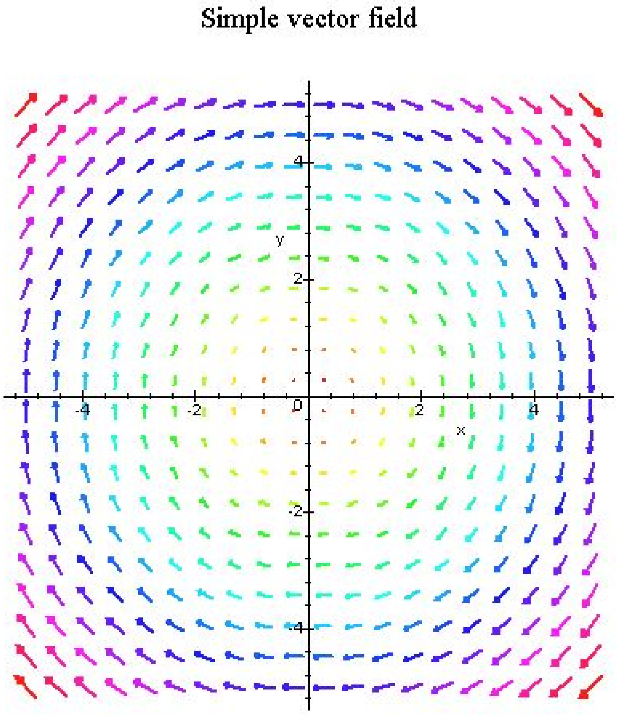


Figure 6.133: Vector field example in Maple™ 4.00b

Simply by visual inspection, we can see that the field is rotating. If we place a paddle wheel anywhere, we see immediately its tendency to rotate clockwise. Using the right-hand rule, we expect the rotational to be into the page. If we are to keep a right-handed coordinate system, into the page will be in the negative z direction. The lack of x and y directions is analogous to the cross product operation.

If we calculate the rotational:

$$\vec{\nabla}\times\vec{f}=0\cdot\vec{e}_x+0\cdot\vec{e}_y+\left(\frac{\partial}{\partial x}(-x)-\frac{\partial}{\partial y}y\right)\vec{e}_z=-2\vec{e}_z \tag{6.5.361}$$

With Maple™ 4.00b we get this algebraic result with the following commands:

```
>with(linalg):
>f:=[y,-x,0];v:=[x,y,z];
>curl(f,v);
```

As we did yet not have the time to find an easy way to plot the resulting rotational vector field in the release 4.00b of Maple™ we will take the picture provided by Wikipedia:

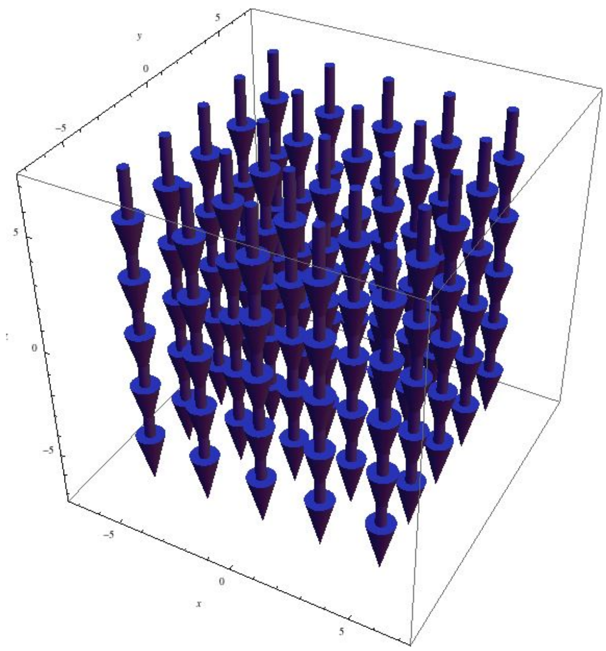


Figure 6.134: **Rotational of previous vector field (source: Wikipedia)**

E2. Suppose we now consider a slightly more complicated vector field:

$$\vec{f}(x,y,z)=-x^2\vec{e}_y \tag{6.5.362}$$

Its plot:

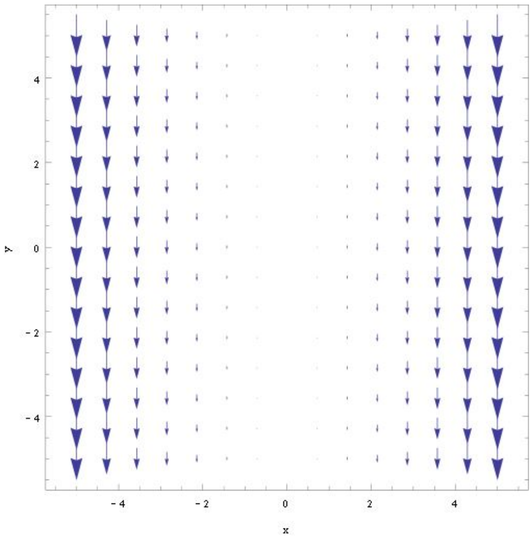


Figure 6.135: **Second vector field example (source: Wikipedia)**

We might not see any rotation initially, but if we closely look at the right, we see a larger field at, say, $x=4$ than at $x=3$. Intuitively, if we placed a small paddle wheel there, the larger "current" on its right side would cause the paddle-wheel to rotate clockwise, which corresponds to a rotational in the negative z direction. By contrast, if we look at a point on the left and placed a small paddle wheel there, the larger "current" on its left side would cause the paddle-wheel to rotate counter-clockwise, which corresponds to a rotational in the positive z direction. Let's check out our guess by doing the math:

$$\vec{\nabla}\times\vec{f}=0\cdot\vec{e}_x+0\cdot\vec{e}_y+\frac{\partial}{\partial x}(-x^2)\vec{e}_z=-2x\vec{e}_z \qquad (6.5.363)$$

Indeed the rotational is in the positive z direction for negative x and in the negative z direction for positive x, as expected. Since this rotational is not the same at every point, its plot is a bit more interesting:

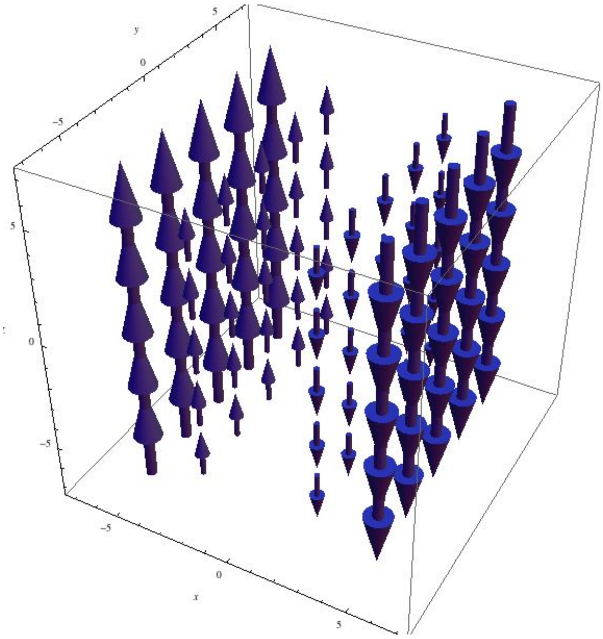


Figure 6.136: **Rotational of previous vector field (source: Wikipedia)**

Let us now determine the expression of the rotational in cylindrical coordinates (the rotational in polar coordinates is not defined in $\mathbb{R}^2$).

Using the same technique as for the rotational in Cartesian coordinates, we write the circulation of $\vec{f}$ along a contour corresponding to a small piece $P_1P_2P_3P_4$ of an orthogonal cylinder (oriented in the direction of the z-axis):

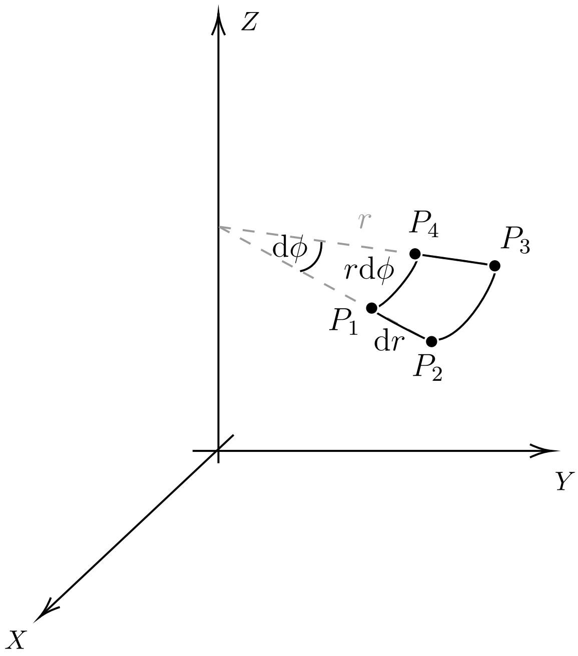


Figure 6.137: **Representation of the cylinder piece** $P_1P_2P_3P_4$

We have then by fixing z (caution! the $\mathrm{d}\vec{r}$ has nothing to do with the cylinder radius r... the notation can be confusing I'm sorry!):

$$\begin{aligned}(\vec{f}\circ\mathrm{d}\vec{r})_{P_1P_2} &= f_r(r,\phi,z)\mathrm{d}r\\(\vec{f}\circ\mathrm{d}\vec{r})_{P_2P_3} &= f_\phi(r+\mathrm{d}r,\phi,z)(r+\mathrm{d}r)\mathrm{d}\phi\\(\vec{f}\circ\mathrm{d}\vec{r})_{P_3P_4} &= -f_r(r,\phi+\mathrm{d}\phi,z)\mathrm{d}r\\(\vec{f}\circ\mathrm{d}\vec{r})_{P_4P_1} &= -f_\phi(r,\phi,z)r\mathrm{d}\phi\end{aligned} \tag{6.5.364}$$

The total circulation thus gives after regrouping terms:

$$\begin{aligned}&\vec{f}\circ\mathrm{d}\vec{r}= f_r(r,\phi,z)\mathrm{d}r+f_\phi(r+\mathrm{d}r,\phi,z)(r+\mathrm{d}r)\mathrm{d}\phi-f_r(r,\phi+\mathrm{d}phi,z)\mathrm{d}r-f_\phi(r,\phi,z)r\mathrm{d}\phi\\&=r\left[f_\phi(r+\mathrm{d}r,\phi,z)-f_\phi(r,\phi,z)\right]\mathrm{d}\phi+f_\phi(r+\mathrm{d},\phi,z)\mathrm{d}\phi\mathrm{d}r-\left[f_r(r,\phi+\mathrm{d}\phi,z)\right]\mathrm{d}r\end{aligned} \tag{6.5.365}$$

We can not at this stage compare with the rotational because it is difficult to us to make appearing the differential of the surface if we look at the differentials that currently appear in circulation. The best is then to divide everything by $r\mathrm{d}\phi\mathrm{d}r$:

$$\begin{aligned}&\frac{1}{r\mathrm{d}\phi\mathrm{d}r}\vec{f}\circ\mathrm{d}\vec{r}\\&=\frac{f_\phi(r+\mathrm{d}r,\phi,z)-f_\phi(r,\phi,z)}{\mathrm{d}r}-\frac{1}{r}\frac{f_r(r,\phi+\mathrm{d}\phi,z)-f_r(r,\phi,z)}{\mathrm{d}\phi}+\frac{1}{r}f_\phi(r+\mathrm{d}r,\phi,z)\\&\underset{\substack{\mathrm{d}r\to 0\\ \mathrm{d}\phi\to 0}}{=}\left(\frac{\partial f_\phi}{\partial r}+\frac{1}{r}f_\phi(r,\phi,z)-\frac{1}{r}\frac{\partial f_r}{\partial\phi}\right)=\frac{1}{r}\frac{\partial(rf_\phi)}{\partial r}-\frac{1}{r}\frac{\partial f_r}{\partial\phi}\end{aligned} \tag{6.5.366}$$

Therefore:

$$\text{rot}(\vec{f})_z = \frac{1}{r}\frac{\partial(rf_\phi)}{\partial r} - \frac{1}{r}\frac{\partial f_r}{\partial \phi} \tag{6.5.367}$$

Now we determine the rotational by fixing ϕ. The problem is like having a rectangle in the space that we travel to determine the circulation. But we already know what is the result of the rotational for a rectangle in Cartesian coordinates following the z-axis:

$$\text{rot}(\vec{f})_z = \frac{\partial f_y}{\partial x} - \frac{\partial f_x}{\partial y} \tag{6.5.368}$$

except that in cylindrical coordinates we have to replace z by ϕ, x by y, y by r and f_y by f_r and finally f_x by f_z (this choice is always appropriate simply because the circulation is such that the surface is always on our left). This gives us:

$$\text{rot}(\vec{f})_\phi = \frac{\partial f_r}{\partial z} - \frac{\partial f_z}{\partial r} \tag{6.5.369}$$

It therefore only remains to us to find the component of the rotational on r (therefore when r is fixed). The calculation is more difficult as we have to follow (positively always!) a curved surface by the variation of the angle ϕ.

We then have by fixing r:

$$\begin{aligned}(\vec{f}\circ \mathrm{d}\vec{r})_{P_1P_2} &= f_\phi(r,\phi,z)r\mathrm{d}\phi\\(\vec{f}\circ \mathrm{d}\vec{r})_{P_2P_3} &= f_z(r,\phi+\mathrm{d}\phi,z)\mathrm{d}z\\(\vec{f}\circ \mathrm{d}\vec{r})_{P_3P_4} &= -f_\phi(r,\phi,z+\mathrm{d}z)r\mathrm{d}\phi\\(\vec{f}\circ \mathrm{d}\vec{r})_{P_4P_1} &= -f_z(r,\phi,z)\mathrm{d}z\end{aligned} \tag{6.5.370}$$

the total circulation thus gives after regrouping terms:

$$\begin{aligned}\vec{f}\circ \mathrm{d}\vec{r} &= f_z(r,\phi+\mathrm{d}\phi,z)\mathrm{d}z - f_z(r,\phi,z)\mathrm{d}z - rf_\phi(r,\phi,z+\mathrm{d}z)\mathrm{d}\phi + rf_\phi(r,\phi,z)\mathrm{d}\phi\\&= \left[f_z(r,\phi+\mathrm{d}\phi,z) - f_z(r,\phi,z)\right]\mathrm{d}z - r\left[f_\phi(r,\phi,z+\mathrm{d}z) - f_\phi(r,\phi,z)\right]\mathrm{d}\phi\end{aligned} \tag{6.5.371}$$

We can not at this stage compare with the rotational because it is difficult to us to make appearing the differential of the surface if we look at the differentials that currently appear in circulation. The best is then to divide everything by $r\mathrm{d}\phi\mathrm{d}z$:

$$\frac{1}{r\mathrm{d}\phi\mathrm{d}z}\vec{f}\circ \mathrm{d}\vec{r} = \frac{1}{r}\frac{f_z(r,\phi+\mathrm{d}\phi,z) - f_z(r,\phi,z)}{\mathrm{d}\phi} - \frac{f_\phi(r,\phi,z+\mathrm{d}z) - f_\phi(r,\phi,z)}{\mathrm{d}z} = \frac{1}{r}\frac{\partial f_z}{\partial \phi} - \frac{\partial f_\phi}{\partial z} \tag{6.5.372}$$

Then finally:

$$\text{rot}(\vec{f})_r = \frac{1}{r}\frac{\partial f_z}{\partial \phi} - \frac{\partial f_\phi}{\partial z} \tag{6.5.373}$$

And finally we have the "rotational in cylindrical coordinates" given in its globality by:

$$\boxed{\text{rot}(\vec{f}) = \begin{pmatrix}\dfrac{1}{r}\dfrac{\partial f_z}{\partial \phi} - \dfrac{\partial f_\phi}{\partial z}\\ \dfrac{\partial f_r}{\partial z} - \dfrac{\partial f_z}{\partial r}\\ \dfrac{1}{r}\dfrac{\partial(rf_\phi)}{\partial r} - \dfrac{1}{r}\dfrac{\partial f_r}{\partial \phi}\end{pmatrix}} \tag{6.5.374}$$

The reader can check verify that this result is simply the gradient in cylindrical coordinates applied to the vector field $\vec{f}$.

To be convinced, let us now show directly the expression of the rotational in spherical coordinates by showing this through the cross product of the gradient in spherical coordinates with the vector field $\vec{f}$.

First let us recall that we have obtained for the gradient in spherical coordinates:

$$\vec{\nabla}=\left(\frac{\partial}{\partial r},\frac{1}{r}\frac{\partial}{\partial\theta},\frac{1}{r\sin(\theta)}\frac{\partial}{\partial\phi}\right) \tag{6.5.375}$$

Therefore we have:

$$\vec{\nabla}\times\vec{f}=\begin{pmatrix}\dfrac{\partial}{\partial r}\\ \dfrac{1}{r}\dfrac{\partial}{\partial\theta}\\ \dfrac{1}{r\sin(\theta)}\dfrac{\partial}{\partial\phi}\end{pmatrix}\times\begin{pmatrix}f_r\\ f_\theta\\ f_\phi\end{pmatrix} \tag{6.5.376}$$

what we can write using the decomposition in basis vectors:

$$\begin{aligned}\vec{\nabla}\times\vec{f}=&\left(\vec{e}_r\frac{\partial}{\partial r}+\frac{\vec{e}_\theta}{r}\frac{\partial}{\partial\theta}+\frac{\vec{e}_\phi}{r\sin(\theta)}\frac{\partial}{\partial\phi}\right)\times(f_r\vec{e}_r+f_\theta\vec{e}_\theta+f_\phi\vec{e}_\phi)\\ &\left(\vec{e}_r\frac{\partial}{\partial r}+\frac{\vec{e}_\theta}{r}\frac{\partial}{\partial\theta}+\frac{\vec{e}_\phi}{r\sin(\theta)}\frac{\partial}{\partial\phi}\right)\times\vec{f}=\vec{e}_r\times\frac{\partial\vec{f}}{\partial r}+\frac{\vec{e}_\theta}{r}\times\frac{\partial\vec{f}}{\partial\theta}+\frac{\vec{e}_\phi}{r\sin(\theta)}\times\frac{\partial\vec{f}}{\partial\phi}\\ &=\vec{e}_r\times\left(\frac{\partial f_r}{\partial r}\vec{e}_r+\frac{\partial f_\theta}{\partial r}\vec{e}_\theta+\frac{\partial f_\phi}{\partial r}\vec{e}_\phi+f_r\frac{\partial\vec{e}_r}{\partial r}+f_\theta\frac{\partial\vec{e}_\theta}{\partial r}+f_\phi\frac{\partial\vec{e}_\phi}{\partial r}\right)\\ &+\frac{\vec{e}_\theta}{r}\times\left(\frac{\partial f_r}{\partial\theta}\vec{e}_r+\frac{\partial f_\theta}{\partial\theta}\vec{e}_\theta+\frac{\partial f_\phi}{\partial\theta}\vec{e}_\phi+f_r\frac{\partial\vec{e}_r}{\partial\theta}+f_\theta\frac{\partial\vec{e}_\theta}{\partial\theta}+f_\phi\frac{\partial\vec{e}_\phi}{\partial\theta}\right)\\ &+\frac{\vec{e}_\phi}{r\sin(\theta)}\times\left(\frac{\partial f_r}{\partial\phi}\vec{e}_r+\frac{\partial f_\theta}{\partial\phi}\vec{e}_\theta+\frac{\partial f_\phi}{\partial\phi}\vec{e}_\phi+f_r\frac{\partial\vec{e}_r}{\partial\phi}+f_\theta\frac{\partial\vec{e}_\theta}{\partial\phi}+f_\phi\frac{\partial\vec{e}_\phi}{\partial\phi}\right)\end{aligned} \tag{6.5.377}$$

Thanks to the partial derivatives that we proved earlier during our introduction to the spherical coordinates, it comes:

$$\begin{aligned}&\vec{\nabla}\times\vec{f}=\\ &=\vec{e}_r\times\left(\frac{\partial f_r}{\partial r}\vec{e}_r+\frac{\partial f_\theta}{\partial r}\vec{e}_\theta+\frac{\partial f_\phi}{\partial r}\vec{e}_\phi+\vec{0}+\vec{0}+\vec{0}\right)\\ &+\frac{\vec{e}_\theta}{r}\times\left(\frac{\partial f_r}{\partial\theta}\vec{e}_r+\frac{\partial f_\theta}{\partial\theta}\vec{e}_\theta+\frac{\partial f_\phi}{\partial\theta}\vec{e}_\phi+f_r\vec{e}_\theta+f_\theta(-\vec{e}_r)+\vec{0}\right)\\ &+\frac{\vec{e}_\phi}{r\sin(\theta)}\times\left(\frac{\partial f_r}{\partial\phi}\vec{e}_r+\frac{\partial f_\theta}{\partial\phi}\vec{e}_\theta+\frac{\partial f_\phi}{\partial\phi}\vec{e}_\phi+f_r\sin(\theta)\vec{e}_\phi+f_\theta\cos(\theta)\vec{e}_\phi+f_\phi\left[-(\vec{e}_r\sin(\theta)+\vec{e}_\theta\cos(\theta))\right]\right)\end{aligned} \tag{6.5.378}$$

The cross products with the collinear vectors cancelled. Therefore it remains:

$$\begin{aligned}\vec{\nabla}\times\vec{f}=&\vec{e}_r\times\left(\frac{\partial f_\theta}{\partial r}\vec{e}_\theta+\frac{\partial f_\phi}{\partial r}\vec{e}_\phi\right)\\ &+\frac{\vec{e}_\theta}{r}\times\left(\frac{\partial f_r}{\partial\theta}\vec{e}_r+\frac{\partial f_\phi}{\partial\theta}\vec{e}_\phi+f_\theta(-\vec{e}_r)\right)\\ &+\frac{\vec{e}_\phi}{r\sin(\theta)}\times\left(\frac{\partial f_r}{\partial\phi}\vec{e}_r+\frac{\partial f_\theta}{\partial\phi}\vec{e}_\theta+f_\phi\left[-(\vec{e}_r\sin(\theta)+\vec{e}_\theta\cos(\theta))\right]\right)\end{aligned} \tag{6.5.379}$$

As the cross product of two basis vectors give the corresponding orthogonal vector (positively or negatively) then we have:

$$\begin{aligned}\vec{\nabla}\times\vec{f}=&\left(\frac{\partial f_\theta}{\partial r}\vec{e}_\phi-\frac{\partial f_\phi}{\partial r}\vec{e}_\theta\right)\\&+\left(-\frac{1}{r}\frac{\partial f_r}{\partial\theta}\vec{e}_\phi+\frac{1}{r}\frac{\partial f_\phi}{\partial\theta}\vec{e}_r-\frac{f_\theta}{r}\vec{e}_\phi\right)\\&+\left(\frac{1}{r\sin(\theta)}\frac{\partial f_r}{\partial\phi}\vec{e}_\theta-\frac{1}{r\sin(\theta)}\frac{\partial f_\theta}{\partial\phi}\vec{e}_r-\frac{f_\phi}{r}\vec{e}_\theta+\frac{f_\phi\vec{e}_r\cos(\theta)}{r\sin(\theta)}\right)\end{aligned}\qquad(6.5.380)$$

By regrouping the terms it comes:

$$\begin{aligned}\vec{\nabla}\times\vec{f}=\vec{e}_r&\left(\frac{1}{r}\frac{\partial f_\phi}{\partial\theta}-\frac{1}{r\sin(\theta)}\frac{\partial f_\theta}{\partial\phi}+\frac{f_\phi\cos(\theta)}{r\sin(\theta)}\right)+\vec{e}_\theta\left(-\frac{f_\phi}{\partial r}+\frac{1}{r\sin(\theta)}\frac{\partial f_r}{\partial\phi}-\frac{f_\phi}{r}\right)\\&+\vec{e}_\phi\left(\frac{\partial f_\theta}{\partial r}-\frac{1}{r}\frac{\partial f_r}{\partial\theta}-\frac{f_\theta}{r}\right)\end{aligned}\qquad(6.5.381)$$

Thus by simplifying:

$$\vec{\nabla}\times\vec{f}=\frac{\vec{e}_r}{r\sin(\theta)}\left(\frac{\partial}{\partial\theta}(\sin(\theta)f_\phi)-\frac{\partial f_\theta}{\partial\phi}\right)+\frac{\vec{e}_\theta}{r\sin(\theta)}\left(\frac{\partial f_r}{\partial\phi}-\sin(\theta)\frac{\partial}{\partial r}(rf_\phi)\right)+\frac{e_p\vec{h}i}{r}\left(\frac{\partial}{\partial r}(rf_\theta)-\frac{\partial f_r}{\partial\theta}\right)\qquad(6.5.382)$$

Thus finally we get the "rotational in spherical coordinates":

$$\boxed{\vec{\nabla}\times\vec{f}=\begin{pmatrix}\dfrac{1}{r\sin(\theta)}\dfrac{\partial}{\partial\theta}(\sin(\theta)f_\phi)-\dfrac{1}{r\sin(\theta)}\dfrac{\partial f_\theta}{\partial\phi}\\\dfrac{1}{r\sin(\theta)}\dfrac{\partial f_r}{\partial\phi}-\dfrac{1}{r}\dfrac{\partial}{\partial r}(rf_\phi)\\\dfrac{1}{r}\dfrac{\partial}{\partial r}(rf_\theta)-\dfrac{1}{r}\dfrac{\partial f_r}{\partial\theta}\end{pmatrix}}\qquad(6.5.383)$$

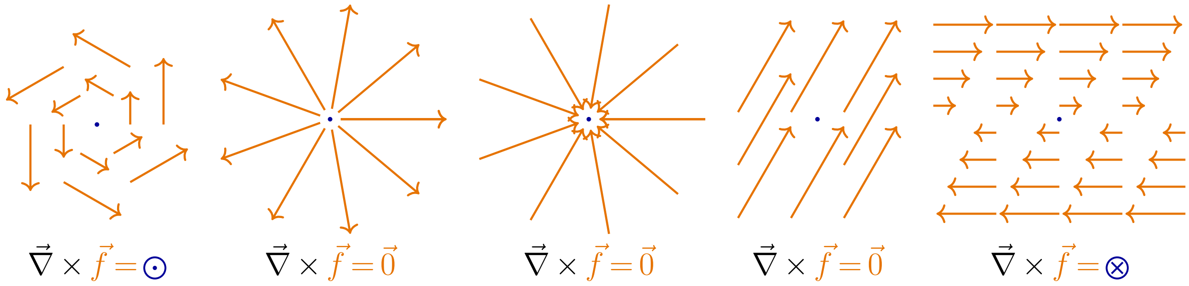

$\vec{\nabla}\times\vec{f}=\odot$
$\vec{\nabla}\times\vec{f}=\vec{0}$
$\vec{\nabla}\times\vec{f}=\vec{0}$
$\vec{\nabla}\times\vec{f}=\vec{0}$
$\vec{\nabla}\times\vec{f}=\otimes$

Figure 6.138: **Various configurations of rotational (author: Izaak Neutelings)**

6.5.8.5 Laplacians of Scalar Fields (Laplace Operator)

The Laplacian of a scalar field $\phi(x_1,x_2,x_3)$ give also a scalar field that gives the difference between the value of the function ϕ on one point and it average around this point. In other words: the second partial derivative measure the variations of the slope on the study point in its immediate neighbourhood and following one direction at a time. If the second partial derivative is null following x, then the slope is constant in its immediate neighbourhood and following this dimension (direction), this implies that the value of the function at the study point is the average of its neighbourhood (following one dimension).

The reader will be able to see again major practical applications of this differential operator in the sections of Complex Analysis, Quantum Chemistry, Astronomy, Electrodynamics, Weather & Marine Engineering, Wave Mechanics, Wave Quantum Physic and Quantum Field Theory.

This operator is defined from the divergence and the gradient and we denote it by (tensor notations):

$$\boxed{\Delta\cdot f=\vec{\nabla}^2 f=\mathrm{div}(\overrightarrow{\mathrm{grad}}(f))=\vec{\nabla}\circ(\vec{\nabla}\cdot\phi)=\sum_{i=1}^{n}\frac{\partial^2 f}{\partial x_i^2}} \tag{6.5.384}$$

The Laplacian is null, or quite small, when the function varies. The functions satisfying the "Laplace equation":

$$\Delta\cdot\phi=0 \tag{6.5.385}$$

are named "harmonic function".

Thus the "scalar Laplacian operator in cartesian coordinates" is by this definition, given by:

$$\vec{\nabla}^2=\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial y^2}\right) \tag{6.5.386}$$

The Laplacian of a scalar field in other coordinate systems is a little bit more hard to get that for the other differential operators. There are more than one possible proof but among the existing one we have try to choose (as always) this that seem to us the most interesting in the point of view of the tools used (and not of simplicity!).

Given the Laplacian in cartesian coordinates in $\mathbb{R}^2$ of a scalar field f:

$$\vec{\nabla}^2 f=\left(\frac{\partial^2 f}{\partial x^2}+\frac{\partial^2 f}{\partial y^2}\right) \tag{6.5.387}$$

To determine this expression in polar coordinates, we will use the total exact differential and the multivariate chain rule in polar coordinates (see section Differential and Integral Calculus page 1483):

$$\mathrm{d}f=\frac{\partial f}{\partial r}\mathrm{d}r+\frac{\partial f}{\partial\phi}\mathrm{d}\phi\Rightarrow\frac{\mathrm{d}f}{\mathrm{d}x}=\frac{\partial f}{\partial r}\frac{\mathrm{d}r}{\mathrm{d}x}+\frac{\partial f}{\partial\phi}\frac{\mathrm{d}\phi}{\mathrm{d}x} \tag{6.5.388}$$

therefore for a second derivative:

$$\begin{aligned}\frac{\mathrm{d}^2 f}{\mathrm{d}x^2} &= \frac{d}{\mathrm{d}x}\left(\frac{\partial f}{\partial r}\frac{\mathrm{d}r}{\mathrm{d}x}+\frac{\partial f}{\partial \phi}\frac{\mathrm{d}\phi}{\mathrm{d}x}\right)=\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial r}\right)\frac{\mathrm{d}r}{\mathrm{d}x}+\frac{\partial f}{\partial r}\frac{\mathrm{d}^2 r}{\mathrm{d}x^2}+\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial \phi}\right)\frac{\mathrm{d}\phi}{\mathrm{d}x}+\frac{\partial f}{\partial \phi}\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2}\\ &=\left(\frac{\partial^2 f}{\partial r^2}\frac{\partial r}{\partial x}+\frac{\partial^2 f}{\partial r\partial\phi}\right)\frac{\mathrm{d}r}{\mathrm{d}x}+\frac{\partial f}{\partial r}\frac{\mathrm{d}^2 r}{\mathrm{d}x^2}+\left(\frac{\partial^2 f}{\partial\phi\partial r}\frac{\partial r}{\partial x}+\frac{\partial^2 f}{\partial\phi^2}\frac{\partial\phi}{\partial x}\right)\frac{\mathrm{d}\phi}{\mathrm{d}x}+\frac{\partial f}{\partial\phi}\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2}\end{aligned}\tag{6.5.389}$$

but we know that we have in polar coordinates:

$$r=\sqrt{x^2+y^2}\quad\text{and}\quad\phi=\arctan\left(\frac{y}{x}\right)\tag{6.5.390}$$

hence for the first derivative:

$$\frac{\mathrm{d}r}{\mathrm{d}x}=\frac{x}{\sqrt{x^2+y^2}}=\frac{x}{r}\qquad\text{and}\qquad\frac{\mathrm{d}\phi}{\mathrm{d}x}=\frac{-y}{x^2+y^2}=-\frac{y}{r^2}\tag{6.5.391}$$

and for the second derivative:

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}=\frac{r-\frac{\mathrm{d}r}{\mathrm{d}x}}{r^2}=\frac{r-\left(\frac{x}{r}\right)x}{r^2}=\frac{y^2}{r^3}\quad\text{and}\quad\frac{\phi^2}{\mathrm{d}x^2}=\frac{\mathrm{d}}{\mathrm{d}x}\left(-\frac{y}{r^2}\right)=-y\frac{\mathrm{d}}{\mathrm{d}x}\left(r^2\right)=-y\left(-\frac{2x}{r^4}\right)=\frac{2xy}{r^4}\tag{6.5.392}$$

therefore:

$$\begin{aligned}\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}=\frac{\partial^2 f}{\partial r^2}\left(\frac{x}{r}\right)^2+\frac{\partial^2 f}{\partial r\partial\phi}\left(\frac{x}{r}\right)\left(-y\frac{y}{r^2}\right)+\frac{\partial f}{\partial r}\left(\frac{y^2}{r^3}\right)+\frac{\partial^2 f}{\partial\phi\partial r}\left(\frac{x}{r}\right)\left(-\frac{y}{r^2}\right)\\+\frac{\partial^2 f}{\partial\phi^2}\left(-\frac{y}{r^2}\right)^2+\frac{\partial f}{\partial\phi}\left(\frac{2xy}{r^4}\right)\end{aligned}\tag{6.5.393}$$

and given that the second partial derivatives are continuous, then the cross derivatives are equal according to the Schwarz's theorem (see section Differential and Integral Calculus page 1469):

$$\frac{\partial^2 f}{\partial r\partial r}=\frac{\partial^2 f}{\partial\phi\partial r}\tag{6.5.394}$$

Therefore:

$$\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}=\frac{\partial^2 f}{\partial r^2}\left(\frac{x^2}{r^2}\right)+\frac{\partial^2 f}{\partial r\partial\phi}\left(-\frac{2xy}{r^3}\right)+\frac{\partial f}{\partial r}\left(\frac{y^2}{r^3}\right)+\frac{\partial^2 f}{\partial\phi^2}\left(\frac{y^2}{r^4}\right)+\frac{\partial f}{\partial\phi}\left(\frac{2xy}{r^4}\right)\tag{6.5.395}$$

Similarly, we will have:

$$\frac{\mathrm{d}^2 f}{\mathrm{d}y^2}=\frac{\partial^2 f}{\partial r^2}\left(\frac{y^2}{r^2}\right)+\frac{\partial^2 f}{\partial r\partial\phi}\left(\frac{2xy}{r^3}\right)+\frac{\partial f}{\partial r}\left(\frac{x^2}{r^3}\right)+\frac{\partial^2 f}{\partial\phi^2}\left(\frac{x^2}{r^4}\right)+\frac{\partial f}{\partial\phi}\left(-\frac{2xy}{r^4}\right)\tag{6.5.396}$$

hence the expression of the Laplace operator in polar coordinates by adding the last two expressions:

$$\vec{\nabla}^2 f=\frac{\partial^2 f}{\partial r^2}+\frac{1}{r}\frac{\partial f}{\partial r}+\frac{1}{r^2}\frac{\partial^2 f}{\partial\phi^2}\tag{6.5.397}$$

Therefore the "scalar Laplacian in polar coordinates" is finally given by:

$$\boxed{\vec{\nabla}^2 = \left(\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{1}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial \phi^2}\right)} \tag{6.5.398}$$

To find the expression of the scalar Laplacian operator in spherical coordinates, we will use the intuition of the physicist and the concepts of similarity.

We will first of all help us with the below figure to find out what we mean:

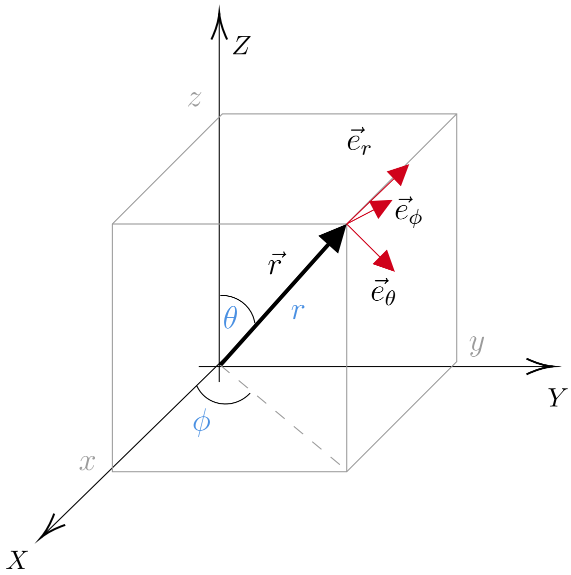


Figure 6.139: **Recall of the spherical coordinate system representation**

Recall that the relation between cartesian and spherical coordinates are given by the relations:

$$\begin{aligned} r &= \sqrt{x^2+y^2+z^2} \\ \tan(\phi) &= \frac{y}{x} \\ \cos(\theta) &= \frac{z}{x^2+y^2+z^2} \end{aligned} \tag{6.5.399}$$

We will now consider the following similarities:

1. Cylindrical coordinates:

$$x = \rho\cos(\phi) \quad \text{and} \quad y = \rho\sin(\phi) \tag{6.5.400}$$

2. Spherical coordinates:

$$z = r\cos(\theta) \quad \text{and} \quad \rho = r\sin(\theta) \tag{6.5.401}$$

Let us build a correspondence table:

$$\begin{array}{lcccc} \text{Cyl:} & x & \rho & \phi & y \\ & \downarrow & \downarrow & \downarrow & \downarrow \\ \text{Sph:} & z & r & \theta & \rho \end{array} \tag{6.5.402}$$

The goal is to play with this correspondence with in a first time the Laplacian in cylindrical coordinates where we have subtracted from both sides of the equality the term $\dfrac{\partial^2 f}{\partial z^2}$. Therefore:

$$\frac{\partial^2 f}{\partial x^2}+\frac{\partial^2 f}{\partial y^2}=\frac{\partial^2 f}{\partial \rho^2}+\frac{1}{\rho}\frac{\partial f}{\partial \rho}+\frac{1}{\rho^2}\frac{\partial^2 f}{\partial \phi^2} \tag{6.5.403}$$

Let us now use our small correspondence table and we get:

$$\frac{\partial^2 f}{\partial z^2}+\frac{\partial^2 f}{\partial \rho^2}=\frac{\partial^2 f}{\partial r^2}+\frac{1}{r}\frac{\partial f}{\partial r}+\frac{1}{r^2}\frac{\partial^2 f}{\partial \theta^2} \tag{6.5.404}$$

The second term of the equality of the latter relation is the spherical equivalent of the term #1 of the Laplacian in cylindrical coordinates:

$$\frac{\partial^2 f}{\partial x^2}+\frac{\partial^2 f}{\partial y^2}+\frac{\partial^2 f}{\partial z^2}=\underbrace{\frac{\partial^2 f}{\partial \rho^2}+\frac{\partial^2 f}{\partial z^2}}_{\#1}+\frac{1}{\rho}\frac{\partial f}{\partial \rho}+\frac{1}{\rho^2}\frac{\partial f}{\partial \phi^2} \tag{6.5.405}$$

Now let us examine and focus on the term: $\dfrac{1}{\rho}\dfrac{\partial f}{\partial \rho}$

Identically as when we determined the relation:

$$\mathrm{d}f=\frac{\partial f}{\partial \rho}\mathrm{d}\rho+\frac{\partial f}{\partial \phi}\mathrm{d}\phi\Rightarrow\frac{\mathrm{d}f}{\mathrm{d}x}=\frac{\partial f}{\partial \rho}\frac{\mathrm{d}\rho}{\mathrm{d}x}+\frac{\partial f}{\partial \phi}\frac{\mathrm{d}\phi}{\mathrm{d}x} \tag{6.5.406}$$

we get:

$$\mathrm{d}u=\frac{\partial u}{\partial rho}\mathrm{d}\rho+\frac{\partial u}{\partial \phi}\mathrm{d}\phi\Rightarrow\frac{\mathrm{d}u}{\mathrm{d}y}=\frac{\partial u}{\partial \rho}\frac{\mathrm{d}\rho}{\mathrm{d}y}+\frac{\partial u}{\partial phi}\frac{\mathrm{d}\phi}{\mathrm{d}y} \tag{6.5.407}$$

with:

$$\frac{\mathrm{d}\rho}{\mathrm{d}y}=\frac{y}{\sqrt{x^2+y^2}}=\frac{y}{\rho}\quad\text{and}\quad\frac{\mathrm{d}\phi}{\mathrm{d}y}=\frac{x}{x^2+y^2}=\frac{x}{\rho^2} \tag{6.5.408}$$

Which gives us the possibility to write:

$$\frac{\mathrm{d}f}{\mathrm{d}y}=\frac{y}{\rho}\frac{\partial f}{\partial \rho}+\frac{x}{\rho^2}\frac{\partial f}{\partial \rho} \tag{6.5.409}$$

If we play again with our small correspondence table we get:

$$\frac{\mathrm{d}f}{\mathrm{d}\rho}=\frac{\rho}{r}\frac{\partial f}{\partial r}+\frac{z}{r^2}\frac{\partial f}{\partial \theta} \tag{6.5.410}$$

We divide the latter relation by ρ and we get:

$$\frac{1}{\rho}\frac{\mathrm{d}f}{\mathrm{d}\rho}=\frac{1}{r}\frac{\partial f}{\partial r}+\frac{z}{\rho r^2}\frac{\partial f}{\partial \theta}=\frac{1}{r}\frac{\partial f}{\partial r}+\frac{\cot(\theta)}{r^2}\frac{\partial f}{\partial \theta} \tag{6.5.411}$$

We have therefore above the spherical equivalent of the second term #2 of the Laplacian in cylindrical coordinates:

$$\frac{\partial^2 f}{\partial x^2}+\frac{\partial^2 f}{\partial y^2}+\frac{\partial^2 f}{\partial z^2}=\underbrace{\frac{\partial^2 f}{\partial \rho^2}+\frac{\partial^2 f}{\partial z^2}}_{\#1}+\underbrace{\frac{1}{\rho}\frac{\partial f}{\partial \rho}}_{\#2}+\frac{1}{\rho^2}\frac{\partial^2 f}{\partial \phi^2} \tag{6.5.412}$$

The last and third term is quite simple to determine. We just have to replace ρ by $r\sin(\theta)$ to get:

$$\frac{1}{\rho^2}\frac{\partial^2 f}{\partial\phi^2} \rightarrow \frac{1}{r^2\sin^2(\theta)}\frac{\partial^2 u}{\partial\phi^2} \tag{6.5.413}$$

By bringing together all terms obtained previously, we finally get the extended form of the Laplacian in spherical coordinates used so much in physics (see corresponding sections of this textbook):

$$\vec{\nabla}^2 = \frac{\partial^2 f}{\partial r^2} + \frac{1}{r}\frac{\partial f}{\partial r} + \frac{1}{r^2}\frac{\partial^2 f}{\partial\theta^2} + \frac{1}{r}\frac{\partial f}{\partial r} + \frac{\cot(\phi)}{r^2}\frac{\partial f}{\partial\theta} + \frac{1}{r^2\sin^2(\theta)}\frac{\partial^2 f}{\partial\phi^2} \tag{6.5.414}$$

We can shorten this expression by factoring the terms:

$$\vec{\nabla}^2 = \frac{\partial^2 f}{\partial r^2} + 2\frac{1}{r}\frac{\partial f}{\partial r} + \frac{1}{r^2}\frac{\partial^2 f}{\partial\theta^2} + \frac{\cot(\phi)}{r^2}\frac{\partial f}{\partial\theta} + \frac{1}{r^2\sin^2(\theta)}\frac{\partial^2 f}{\partial\phi^2} \tag{6.5.415}$$

If we condense even a little bit more, we get the final expression of the "scalar Laplacian in spherical coordinates" named also "spherical Laplacian":

$$\boxed{\vec{\nabla}^2 = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right) + \frac{1}{r^2\sin(\theta)}\frac{\partial}{\partial\theta}\left(\sin(\theta)\frac{\partial}{\partial\theta}\right) + \frac{1}{r^2\sin^2(\theta)}\frac{\partial^2}{\partial\phi^2}} \tag{6.5.416}$$

6.5.8.6 Laplacians of Vector Fields

As to the Laplacian of a scalar field, the Laplacian of a vector field is only a very convenient notation system for condensing the writing of the components of a vector field.

The reader will also find practical applications of this operator in the sections of Electrokinetics, Electrodynamics and Continuum Mechanics.

Thus, the vector Laplacian is often defined by (introducing at the same time the different existing notations):

$$\Delta\vec{f} = \vec{\nabla}^2\vec{v} = \vec{\nabla}\left(\vec{\nabla}\circ\vec{f}\right) - \vec{\nabla}\times\left(\vec{\nabla}\times\vec{f}\right) \tag{6.5.417}$$

We also prove that in the specific case of the Cartesian coordinates, the Laplacian of a vector field has for components the scalar Laplacian of each of the components.

So we have therefore in Cartesian coordinates:

$$\begin{aligned}
\Delta\vec{f} &= \vec{\nabla}\left(\vec{\nabla}\circ\vec{f}\right) - \vec{\nabla}\times\left(\vec{\nabla}\times\vec{f}\right) = \vec{\nabla}\left(\partial_x f_1 + \partial_y f_2 + \partial_z f_3\right) - \vec{\nabla}\times\begin{pmatrix}\partial_y f_3 - \partial_z f_2\\ \partial_z f_1 - \partial_x f_3\\ \partial_x f_2 - \partial_y f_1\end{pmatrix}\\
&= \begin{pmatrix}\partial_x\partial_x f_1 + \partial_x\partial_y f_2 + \partial_x\partial_z f_3\\ \partial_y\partial_x f_1 + \partial_y\partial_y f_2 + \partial_y\partial_z f_3\\ \partial_z\partial_x f_1 + \partial_z\partial_y f_2 + \partial_z\partial_z f_3\end{pmatrix} - \begin{pmatrix}\partial_y(\partial_x f_2 - \partial_y f_1) - \partial_z(\partial_z f_1 - \partial_x f_3)\\ \partial_z(\partial_y f_3 - \partial_z f_2) - \partial_x(\partial_x f_2 - \partial_y f_1)\\ \partial_x(\partial_z f_1 - \partial_x f_3) - \partial_y(\partial_y f_3 - \partial_z f_3)\end{pmatrix}\\
&= \begin{pmatrix}\partial_x\partial_x f_1 + \partial_x\partial_y f_2 + \partial_x\partial_z f_3 - \partial_y\partial_x f_2 + \partial_y\partial_y f_1 + \partial_z\partial_z f_1 - \partial_z\partial_x f_3\\ \partial_y\partial_x f_1 + \partial_y\partial_y f_2 + \partial_y\partial_z f_3 - \partial_z\partial_y f_3 + \partial_z\partial_z f_2 + \partial_x\partial_x f_2 - \partial_x\partial_y f_1\\ \partial_z\partial_x f_1 + \partial_z\partial_y f_2 + \partial_z\partial_z f_3 - \partial_x\partial_z f_1 + \partial_x\partial_x f_3 - \partial_y\partial_y f_3 + \partial_y\partial_z f_2\end{pmatrix}\\
&= \begin{pmatrix}\partial_x\partial_x f_1 + \partial_y\partial_y f_1 + \partial_z\partial_z f_1\\ \partial_x\partial_x f_1 + \partial_y\partial_y f_2 + \partial_z\partial_z f_2\\ \partial_x\partial_x f_1 + \partial_y\partial_y f_2 + \partial_z\partial_z f_3\end{pmatrix}
\end{aligned} \tag{6.5.418}$$

and thus the "vector Laplacian of a vector field in Cartesian coordinates" is indeed the scalar Laplacian of each component:

$$\boxed{\Delta\vec{f}=\begin{pmatrix}\displaystyle\sum_{i=1}^{3}\frac{\partial^2 f_1}{\partial x_i^2}\\ \displaystyle\sum_{i=1}^{3}\frac{\partial^2 f_2}{\partial x_i^2}\\ \displaystyle\sum_{i=1}^{3}\frac{\partial^2 f_3}{\partial x_i^2}\end{pmatrix}} \tag{6.5.419}$$

Or more explicitly:

$$\vec{\nabla}=\begin{pmatrix}\dfrac{\partial^2 f_1}{\partial x^2}+\dfrac{\partial^2 f_1}{\partial y^2}+\dfrac{\partial^2 f_1}{\partial z^2}\\ \dfrac{\partial^2 f_2}{\partial x^2}+\dfrac{\partial^2 f_2}{\partial y^2}+\dfrac{\partial^2 f_2}{\partial z^2}\\ \dfrac{\partial^2 f_3}{\partial x^2}+\dfrac{\partial^2 f_3}{\partial y^2}+\dfrac{\partial^2 f_3}{\partial z^2}\end{pmatrix} \tag{6.5.420}$$

The Laplacian of a vector field, frequently named "vectorial laplacian", in other coordinates systems is quite simple to get once we know the Laplacian of a scalar field in the same coordinates!

We have first in cylindrical coordinates:

$$\begin{gathered}\Delta\vec{f}=\vec{\nabla}(\vec{\nabla}\circ\vec{f})-\vec{\nabla}\times(\vec{\nabla}\times\vec{f})=\vec{\nabla}\left(\frac{1}{r}\partial_r r f_r+\frac{1}{r}\partial_\phi f_\phi+\partial_z f_z\right)-\nabla\times\begin{pmatrix}\frac{1}{r}\partial_\phi f_z-\partial_z f_\phi\\ \partial_z f_r-\partial_r f_z\\ \frac{1}{r}\partial_r r f_\phi-\frac{1}{r}\partial_\phi f_r\end{pmatrix}\\ =\begin{pmatrix}\partial_r\left(\frac{1}{r}\partial_r r f_r\right)+\partial_r\left(\frac{1}{r}\partial_\phi f_\phi\right)+\partial_r(\partial_z f_z)\\ \frac{1}{r}\partial_\phi\left(\frac{1}{r}\partial_r r f_r\right)+\frac{1}{r}\partial_\phi\left(\frac{1}{r}\partial_\phi f_\phi\right)+\frac{1}{r}\partial_\phi(\partial_z f_z)\\ \partial_z\left(\frac{1}{r}\partial_r r f_r\right)+\partial_z\left(\frac{1}{r}\partial_\phi f_\phi\right)+\partial_z\partial_z f_z\end{pmatrix}\\ -\begin{pmatrix}\frac{1}{r}\partial_\phi\left(\frac{1}{r}\partial_r r f_\phi-\frac{1}{r}\partial_\phi f_r\right)-\partial_z(\partial_z f_r-\partial_r f_z)\\ \partial_z\left(\frac{1}{r}\partial_\phi f_z-\partial_z f_\phi\right)-\partial_r\left(\frac{1}{r}\partial_r r f_\phi-\frac{1}{r}\partial_\phi f_r\right)\\ \frac{1}{r}\partial_r r(\partial_z f_r-\partial_r f_z)-\frac{1}{r}\partial_\phi\left(\frac{1}{r}\partial_\phi f_z-\partial_z f\phi\right)\end{pmatrix}\end{gathered} \tag{6.5.421}$$

To simplify (because of a lack of space) let us focus first on the first line:

$$\begin{aligned}
&\partial_r\left(\frac{1}{r}\partial_r r f_r\right)+\partial_r\left(\frac{1}{r}\partial_\phi f_\phi\right)+\partial_r(\partial_z f_z)-\frac{1}{r}\partial_\phi\left(\frac{1}{r}\partial_r r f_\phi-\frac{1}{r}\partial_\phi f_r\right)+\partial_z(\partial_z f_r-\partial_r f_z)\\
&=\partial_r\left(\frac{1}{r}\partial_r r f_r\right)+\partial_r\left(\frac{1}{r}\partial_\phi f_\phi\right)-\frac{1}{r}\partial_\phi\left(\frac{1}{r}\partial_r r f_\phi-\frac{1}{r}\partial_\phi f_r\right)+\partial_z^2 f_r\\
&=-\frac{1}{r^2}\partial_r r f_r+\frac{1}{r}\partial_r^2 r f_r-\frac{1}{r^2}\partial_\phi f_\phi+\frac{1}{r}\partial_r\partial_\phi f_\phi-\frac{1}{r^2}\partial_\phi\partial_r r f_\phi+\frac{1}{r^2}\partial_\phi^2 f_r+\partial_z^2 f_r\\
&=-\frac{1}{r^2}f_r-\frac{1}{r}\partial_r f_r+\frac{1}{r}\partial_r^2 r f_r-\frac{1}{r^2}\partial_\phi f_\phi+\frac{1}{r}\partial_r\partial_\phi f_\phi-\frac{1}{r^2}\partial_\phi\partial_r f_\phi+\frac{1}{r^2}\partial_\phi^2 f_r+\partial_z^2 f_r\\
&=-\frac{1}{r^2}f_r-\frac{1}{r}\partial_r f_r+\frac{1}{r}\partial_r^2 r f_r-\frac{2}{r^2}\partial_\phi f\phi+\frac{1}{r^2}\partial_\phi^2 f_r+\partial_z^2 f_r\\
&=-\frac{1}{r^2}f_r-\partial 1 r\partial_r f_r+\frac{1}{r}\partial_r f_r+\frac{1}{r}\partial_r r\partial_r f_r-\frac{2}{r^2}\partial_\phi f_\phi+\frac{1}{r^2}\partial_\phi^2 f_r+\partial_z^2 f_r\\
&=-\frac{1}{r^2}f_r+\frac{1}{r}\partial_r r\partial_r f_r-\frac{2}{r^2}\partial_\phi f_\phi+\frac{1}{r^2}\partial_\phi^2 f_r+\partial_z^2 f_r\\
&=-\frac{1}{r^2}f_r+\frac{1}{r}\partial_r f_r+\partial_r^2 f_r-\frac{2}{r^2}\partial_\phi f_\phi+\frac{1}{r^2}\partial_\phi^2 f_r+\partial_z^2 f_r
\end{aligned}\tag{6.5.422}$$

and after for the second line:

$$\begin{aligned}
&=\frac{1}{r}\partial_\phi\left(\frac{1}{r}\partial_r r f_r\right)+\frac{1}{r}\partial_\phi\left(\frac{1}{r}\partial_\phi f_\phi\right)+\frac{1}{r}\partial_\phi(\partial_z f_z)-\partial_z\left(\frac{1}{r}\partial_\phi f_z-\partial_z f_\phi\right)+\partial_r\left(\frac{1}{r}\partial_r r f_\phi-\frac{1}{r}\partial_\phi f_r\right)\\
&=\frac{1}{r}\partial_\phi\left(\frac{1}{r}\partial_r r f_r\right)+\frac{1}{r}\partial_\phi\left(\frac{1}{r}\partial_\phi f_\phi\right)+\frac{1}{r}\partial_\phi\partial_z f_z-\partial_z\frac{1}{z}\partial_\phi f_z+\partial_z^2 f_\phi+\partial_r\frac{1}{r}\partial_r r f_\phi-\partial_r\frac{1}{r}\partial_\phi f_r\\
&=\frac{1}{r^2}\partial_\phi\partial_r r f_r+\frac{1}{r^2}\partial_\phi^2 f_\phi+\frac{1}{r}\partial_\phi\partial_z f_z-\frac{1}{r}\partial_z\partial_\phi f_z+\partial_z^2 f_\phi+\partial_r\frac{1}{r}\partial_r r f_\phi-\partial_r\frac{1}{r}\partial_\phi f_r\\
&=\frac{1}{r^2}\partial_\phi\partial_r r f_r+\frac{1}{r^2}\partial_\phi^2 f_\phi+\partial_z^2 f_\phi+\partial_r\frac{1}{r}r f_\phi-\partial_r\frac{1}{r}\partial_\phi f_r\\
&=\frac{1}{r^2}\partial_\phi f_r+\frac{1}{r^2}\partial_\phi r\partial_r f_r+\frac{1}{r^2}\partial_\phi^2 f_\phi+\partial_z^2 f_\phi-\frac{1}{r^2}\partial_r r f_\phi+\frac{1}{r}\partial_r^2 r f_\phi+\frac{1}{r^2}\partial_\phi f_r-\frac{1}{r}\partial_r\partial_\phi f_r\\
&=\frac{2}{r^2}\partial_\phi f_r+\frac{1}{r}\partial_\phi\partial_r f_r+\frac{1}{r^2}\partial_\phi^2 f_\phi f_\phi+\partial_z^2 f_\phi-\frac{1}{r^2}f_\phi-\frac{1}{r}f_\phi+\frac{1}{r}\partial_r^2 r f_\phi-\frac{1}{r}\partial_r\partial_\phi f_r\\
&=\frac{2}{r^2}\partial_\phi f_r+\frac{1}{r^2}\partial_\phi^2 f_\phi+\partial_z^2 f_\phi+\partial_z^2 f_\phi-\frac{1}{r^2}f_\phi+\frac{1}{r}\partial_r r\partial_r f_\phi\\
&=\frac{2}{r^2}\partial_\phi f_r+\frac{1}{r^2}\partial_\phi^2 f_\phi+\partial_z^2 f_\phi+\partial_z^2 f_\phi-\frac{1}{r^2}f_\phi+\frac{1}{r}\partial_r f_\phi+\partial_r^2 f_\phi
\end{aligned}\tag{6.5.423}$$

and finally the third one:

$$\begin{aligned}
&=\partial_z\left(\frac{1}{r}\partial_r r f_r\right)+\partial_z\left(\frac{1}{r}\partial_\phi f_\phi\right)+\partial_z\partial_z f_z-\frac{1}{r}\partial_r r(\partial_z f_r-\partial_r f_z)+\frac{1}{r}\partial_\phi\left(\frac{1}{r}\partial_\phi f_z-\partial_z f_\phi\right)\\
&=\frac{1}{r}\partial_z\partial_r r f_r+\partial 1r\partial_z\partial_\phi f_\phi+\partial_z^2 f_z-\partial 1r\partial_r r\partial_z f_r+\frac{1}{r}\partial_r r\partial_r f_z+\frac{1}{r^2}\partial_p hi^2 f_z-\frac{1}{r}\partial_\phi\partial_z f_\phi\\
&=\frac{1}{r}\partial_z\partial_r r f_r+\partial_z^2 f_z-\frac{1}{r}\partial_r r\partial f_r+\frac{1}{r}\partial_r r\partial_r f_z+\frac{1}{r^2}\partial_\phi^2 f_z\\
&=\frac{1}{r}\partial_z f_r+\frac{1}{r}\partial_z r\partial_r f_r+\partial_z^2 f_z-\frac{1}{r}\partial_z f_r-\partial_r\partial_z f_r+\frac{1}{r}\partial_r f_z+\partial_r^2 f_z+\partial 1r^2\partial_\phi^2 f_z\\
&=\frac{1}{r}\partial_z r\partial_r f_r+\partial_z^2 f_z-\partial_r\partial_z f_r+\frac{1}{r}\partial_r f_z+\partial_r^2 f_z+\frac{1}{2}\partial_\phi^2 f_z\\
&=\partial_z\partial_r f_r+\partial_z^2 f_z-\partial_r\partial_z f_r+\frac{1}{r}\partial_r f_z+\partial_r^2 f_z+\frac{1}{r^2}\partial_\phi^2 f_z\\
&=\partial_z^2 f_z+\frac{1}{r}\partial_r f_z+\partial_r^2 f_z+\frac{1}{r^2}\partial_\phi^2 f_z
\end{aligned}\tag{6.5.424}$$

So what gives for us for the "vector Laplacian in cylindrical coordinates" as we can find it in tables or forms:

$$\boxed{\Delta\vec{f}=\begin{pmatrix}-\dfrac{1}{r^2}f_r+\dfrac{1}{r}\partial_r f_r+\partial_r^2 f_r-\dfrac{2}{r^2}\partial_\phi f_\phi+\dfrac{1}{r^2}\partial_\phi^2 f_r+\partial_z^2 f_r\\ \dfrac{2}{r^2}\partial_\phi f_r+\dfrac{1}{r^2}\partial_\phi^2 f_\phi+\partial_z^2 f_\phi+\partial_z^2 f_\phi-\dfrac{1}{r^2}f_\phi+\dfrac{1}{r}\partial_r f_\phi+\partial_r^2 f_\phi\\ \partial_z^2 f_z+\dfrac{1}{r}\partial_r f_z+\partial_r^2 f_z+\dfrac{1}{r^2}\partial_\phi^2 f_z\end{pmatrix}}\tag{6.5.425}$$

To finish and in the joy (...) let us make the merry and detailed calculations of the vector Laplacian operator in spherical coordinates (it's quite long but it's just to make sure that we fall back on what is in tables and forms):

$$\begin{aligned}
\Delta\vec{f}&=\vec{\nabla}(\vec{\nabla}\circ\vec{f})-\vec{\nabla}\times(\vec{\nabla}\times\vec{f})=\vec{\nabla}\left(\frac{1}{r^2}\partial_r(r^2 f_r)+\frac{1}{r\sin(\theta)}\partial_\theta(\sin(\theta)f_\theta)+\frac{1}{r\sin(\theta)}\partial_\phi f_\phi\right)\\
&-\vec{\nabla}\times\begin{pmatrix}\dfrac{1}{r\sin(\theta)}\partial_\theta(\sin(\theta)f_\phi)-\dfrac{1}{r\sin(\theta)}\partial_\phi f_\theta\\ \dfrac{1}{r\sin(\theta)}\partial_\phi f_r-\dfrac{1}{r}\partial_r(rf_\phi)\\ \dfrac{1}{r}\partial_r(rf_\theta)-\dfrac{1}{r}\partial_\theta f_r\end{pmatrix}\\
&=\begin{pmatrix}\partial_r\left(\dfrac{1}{r^2}\partial_r(r^2 f_r)\right)+\partial_r\left(\dfrac{1}{r\sin(\theta)}\partial_\theta(\sin(\theta)f_\theta)\right)+\partial_r\left(\dfrac{1}{r\sin(\theta)}\partial_\phi f_\phi\right)\\ \dfrac{1}{r}\partial_\theta\left(\dfrac{1}{r^2}\partial_r(r^2 f_r)\right)+\dfrac{1}{r}\partial_\theta\left(\dfrac{1}{r\sin(\theta)}\partial_\theta(\sin(\theta)f_\theta)\right)+\dfrac{1}{r}\partial_\theta\left(\dfrac{1}{r\sin(\theta)\partial_\phi f_\phi}\right)\\ \dfrac{1}{r\sin(\theta)}\partial_\phi\left(\dfrac{1}{r^2}\partial_r(r^2 f_r)\right)+\dfrac{1}{r\sin(\theta)}\partial_\phi\left(\dfrac{1}{r\sin(\theta)}\partial_\theta(\sin(\theta)f_\theta)\right)+\dfrac{1}{r\sin(\theta)}\partial_\phi\left(\dfrac{1}{r\sin(\theta)}\right)\end{pmatrix}\\
&-\begin{pmatrix}\dfrac{1}{r\sin(\theta)}\partial_\theta\left(\sin(\theta)\left(\dfrac{1}{r}\partial_r(rf_\theta)-\dfrac{1}{r}\partial_\theta f_r\right)\right)-\dfrac{1}{r\sin(\theta)}\partial_\phi\left(\dfrac{1}{r}\partial_\phi f_r-\dfrac{1}{r}\partial_r(rf_\phi)\right)\\ \dfrac{1}{r\sin(\theta)}\partial_\phi\left(\dfrac{1}{r\sin(\theta)}\partial_\theta(\sin(\theta)f_\phi)-\dfrac{1}{r\sin(\theta)}\partial_\phi f_\theta\right)-\dfrac{1}{r}\partial_r\left(r\left(\dfrac{1}{r}\partial_r(rf_\theta)-\dfrac{1}{r}\partial_\theta f_r\right)\right)\\ \dfrac{1}{r}\partial_r\left(r\left(\dfrac{1}{r\sin(\theta)}\partial_\phi f_r-\dfrac{1}{r}\partial_r(rf_\phi)\right)\right)-\dfrac{1}{r}\partial_\theta\left(\dfrac{1}{r\sin(\theta)}\partial_\theta(\sin(\theta f_\phi)-\dfrac{1}{r\sin(\theta)}\partial_\phi f_\theta\right)\end{pmatrix}
\end{aligned}\tag{6.5.426}$$

Let us focus on the first line (caution! this will be quite long...):

$$
\begin{aligned}
&-\frac{1}{r\sin(\theta)}\cos(\theta)\partial_r f_\theta-\frac{1}{r\sin(\theta)}\sin(\theta)\partial_\theta\partial_r f_\theta+\frac{1}{r\sin(\theta)}\frac{\cos(\theta)}{r}\partial_\theta f_r+\frac{1}{r\sin(\theta)}\frac{\sin(\theta)}{r}\partial_\theta\partial_\theta f_\theta\\
&+\frac{1}{r^2\sin^2(\theta)}\partial_\phi^2 f_r-\frac{2}{r^2\sin(\theta)}\partial_\phi f_\phi\\
&=\partial_r\left(\frac{1}{r^2}\partial_r(r^2 f_r)\right)-\frac{1}{r^2\sin(\theta)}\partial_\theta(\sin(\theta)f_\theta)+\frac{1}{r\sin(\theta)}\partial_r\partial_\theta(\sin(\theta)f_\theta)\\
&-\frac{\cot(\theta)}{r^2}f_\theta-\frac{1}{r^2}\partial_\theta f_\theta-\frac{\cot(\theta)}{r}\partial_r f_\theta-\frac{1}{r}\partial_\theta\partial_r f_\theta+\frac{\cot(\theta)}{r^2}\partial_\theta f_r+\frac{1}{r^2}\partial_\theta^2 f_r\\
&+\frac{1}{r^2\sin^2(\theta)}\partial_\phi^2 f_r-\frac{2}{r^2\sin(\theta)}\partial_\phi f_\phi\\
&=\partial_r\left(\frac{1}{r^2}\partial_r(r^2 f_r)\right)-\frac{1}{r^2\sin(\theta)}\cos(\theta)f_\theta-\frac{1}{r^2\sin(\theta)}\partial_\theta f_\theta+\frac{1}{r\sin(\theta)}\partial_r f_\theta+\frac{1}{r\sin(\theta)}\partial_r\partial_\theta f_\theta\\
&-\frac{\cot(\theta)}{r^2}f_\theta-\frac{1}{r^2}\partial_\theta f_\theta-\frac{\cot(\theta)}{r}\partial_r f_\theta-\frac{1}{r}\partial_\theta\partial_r f_\theta+\frac{\cot(\theta)}{r^2}\partial_\theta f_r+\frac{1}{r^2}\partial_\theta^2 f_r\\
&+\frac{1}{r^2\sin^2(\theta)}\partial_\phi^2 f_r-\frac{2}{r^2\sin(\theta)}\partial_\phi f_\phi\\
&=\partial_r\left(\frac{1}{r^2}\partial_r(r^2 f_r)\right)-\frac{\cot(\theta)}{r^2}f_\theta-\frac{1}{r^2}\partial_\theta f_\theta+\frac{\cot(\theta)}{r}\partial_r f_\theta+\frac{1}{r}\partial_r\partial_\theta f_\theta\\
&-\frac{\cot(\theta)}{r^2}f_\theta-\frac{1}{r^2}\partial_\theta f_\theta-\frac{\cot(\theta)}{r}\partial_r f_\theta-\frac{1}{r}\partial_\theta\partial_r f_\theta+\frac{\cot(\theta)}{r^2}\partial_\theta f_r+\frac{1}{r^2}\partial_\theta^2 f_r\\
&+\frac{1}{r^2\sin^2(\theta)}\partial_\phi^2 f_r-\frac{2}{r^2\sin(\theta)}\partial_\phi f_\phi\\
&=\partial_r\left(\frac{1}{r^2}\partial_r(r^2 f_r)\right)-\frac{1}{r^2\sin^2(\theta)}\partial_\phi^2 f_r-\frac{2}{r^2\sin(\theta)}\partial_\phi f_\phi\\
&=-\frac{2}{r^3}\partial_r(r^2 f_r)+\left(\frac{1}{r^2}\partial_r^2(r^2 f_r)\right)\\
&-\frac{2\cot(\theta)}{r^2}f_\theta-\frac{2}{r^2}\partial_\theta f_\theta\frac{\cot(\theta)}{r^2}\partial_\theta f_r+\frac{1}{r^2}\partial_\theta^2 f_r+\frac{1}{r^2\sin^2(\theta)}\partial_\phi f_r-\frac{2}{r^2\sin(\theta)}\partial_\phi f_\phi\\
&=-\frac{2}{r^3}2rf_r-\frac{2}{r^3}r^2\partial_r f_r+\left(\frac{1}{r^2}\partial_r(2rf_2+r^2\partial_r f_r)\right)\\
&-\frac{2\cot(\theta)}{r^2}f_\theta-\frac{2}{r^2}\partial_\theta f_\theta+\frac{\cot(\theta)}{r^2}\partial_\theta f_r+\frac{1}{r^2}\partial_\theta^2 f_r+\frac{1}{r^2\sin^2(\theta)}\partial_\phi^2 f_r-\frac{2}{r^2\sin(\theta)}\partial_\phi f_\phi\\
&=-\frac{4}{r^2}f_r-\frac{2}{r}\partial_r f_r+\frac{1}{r^2}2f_r+\frac{1}{r^2}2r\partial_r f_r+\frac{1}{r^2}2r\partial_r^2 f_r\\
&-\frac{2\cot(\theta)}{r^2}f_\theta-\frac{2}{r^2}\partial_\theta f_\theta+\frac{\cot(\theta)}{r^2}\partial_\theta f_r+\frac{1}{r^2}\partial_\theta^2 f_r+\frac{1}{r^2\sin^2(\theta)}\partial_\phi^2 f_r-\frac{2}{r^2\sin(\theta)}\partial_\phi f_\phi\\
&=-\frac{2}{r^2}f_r+\frac{2}{r}\partial_r f_r+\underbrace{\partial_r^2 f_r}_{=\frac{1}{r}\partial_r^2(rf_r)}-\frac{2\cot(\theta)}{r^2}f_\theta-\frac{2}{r^2}\partial_\theta f_\theta+\frac{\cot(\theta)}{r^2}\partial_\theta f_r+\frac{1}{r^2}\partial_\theta^2 f_r\\
&+\frac{1}{r^2\sin^2(\theta)}\partial_\phi^2 f_r-\frac{2}{r^2\sin(\theta)}\partial_\phi f_\phi
\end{aligned}
\tag{6.5.427}
$$

$$
\begin{aligned}
&=-\frac{2}{r^2}f_r+\frac{1}{r}\partial_r^2(rf_r)\\
&-\frac{2\cot(\theta)}{r^2}f_\theta-\frac{2}{r^2}\partial_\theta f_\theta+\frac{\cot(\theta)}{r^2}\partial_\theta f_r+\frac{1}{r^2}\partial_\theta^2 f_r+\frac{1}{r^2\sin^2(\theta)}\partial_\phi^2 f_r-\frac{2}{r^2\sin(\theta)}\partial_\phi f_\phi
\end{aligned}
\tag{6.5.428}
$$

That's it for the first line ... Let us get on to the second line always with joy...:

$$
\begin{aligned}
&=\frac{1}{r}\left(\frac{1}{r^2}\partial_r(r^2f_r)\right)+\frac{1}{r}\partial_\theta\left(\frac{1}{r\sin(\theta)}\partial_\theta(\sin(\theta)f_\theta)\right)+\frac{1}{r}\partial_\theta\left(\frac{1}{r\sin(\theta)}\partial_\phi f_\phi\right)\\
&-\frac{1}{r\sin(\theta)}\partial_\phi\left(\frac{1}{\sin(\theta)}\partial_\theta(\sin(\theta f_\phi)-\frac{1}{r\sin(\theta)}\partial_\phi f_\theta\right)+\frac{1}{r}\partial_r\left(r\left(\frac{1}{r}\partial_r(rf_\theta)-\frac{1}{r}\partial_\theta f_r\right)\right)\\
&=\frac{1}{r^3}\partial_\theta\partial_r r^2 f_r-\frac{1}{r^2}\frac{\cos(\theta)}{\sin^2(\theta)}\partial_\phi f_\phi+\frac{1}{r^2\sin(\theta)}\partial_\theta\partial_\phi f_\phi-\frac{1}{r^2}\frac{\cos(\theta)}{\sin^2(\theta)}\partial_\phi f_\phi+\frac{1}{r^2\sin(\theta)}\partial_\theta\partial_\phi f_\phi\\
&+\frac{1}{r^2\sin(\theta)}\partial_\theta\partial_\phi f_\phi-\frac{1}{r^2\sin^2(\theta)}\partial_\phi\partial_\theta\sin(\theta)f_\phi+\frac{1}{r^2\sin^2(\theta)}\partial_\phi^2 f_\theta+\frac{1}{r}\partial_r(\partial_r(rf_\theta)-\partial_\theta f_r)\\
&=\frac{1}{r^3}\partial_\theta 2rf_r+\frac{1}{r^3}\partial_\theta r^2\partial_r f_r-\frac{1}{r^2}\frac{\cot(\theta)}{\sin(\theta)}\partial_\theta\sin(\theta)f_\theta+\frac{1}{r^2\sin(\theta)}\partial_\theta^2(\sin(\theta)f_\theta)-\frac{1}{r^2}\frac{\cot(\theta)}{\sin(\theta)}\partial_\phi f_\phi\\
&+\frac{1}{r^2\sin(\theta)}\partial_\theta\partial_\phi f_\phi-\frac{1}{r^2\sin^2(\theta)}\partial_\phi\partial_\theta\sin(\theta)f_\phi+\frac{1}{r^2\sin(\theta)}\partial_\phi^2 f_\theta+\frac{1}{r}\partial_r(f_\theta+r\partial_r f_\theta-\partial_\theta f_r)\\
&=\frac{2}{r^2}\partial_\theta f_r+\frac{1}{r}\partial_\theta\partial_r f_r-\frac{1}{r^2}\frac{\cot(\theta)}{\sin(\theta)}\cos(\theta)f_\theta-\frac{1}{r^2}\frac{\cot(\theta)}{\sin(\theta)}\sin(\theta)\partial_\theta f_\theta+\frac{1}{r^2\sin(\theta)}\partial_\theta^2(\sin(\theta)f_\theta)\\
&-\frac{1}{r^2}\frac{\cot(\theta)}{\sin(\theta)}\partial_\phi f_\phi+\frac{1}{r^2\sin(\theta)}\partial_\theta\partial_\phi f_\phi-\frac{1}{r^2\sin^2(\theta)}\partial_\phi\cos(\theta)f_\phi-\\
&\frac{1}{r^2\sin^2(\theta)}\partial_\phi\sin(\theta)\partial_\theta f_\phi+\frac{1}{r^2\sin^2(\theta)}\partial_\phi^2 f_\theta+\frac{1}{r}\partial_r f_\theta+\frac{1}{r}\partial_r r\partial_r f_\theta-\frac{1}{r}\partial_r\partial_\theta f_r\\
&=\frac{2}{r^2}\partial_\theta f_r-\frac{1}{r^2}\cot^2(\theta)f_\theta-\frac{\cot(\theta)}{r^2}\partial_\theta f_\theta+\frac{1}{r^2\sin(\theta)}\partial_\theta\cos(\theta)f_\theta+\frac{1}{r^2\sin(\theta)}\partial_\theta\sin(\theta)\partial_\theta f_\theta\\
&-\frac{2}{r^2}\frac{\cot(\theta)}{\sin(\theta)}\partial_\phi f_\phi+\frac{1}{r^2\sin(\theta)}\partial_\theta\partial_\phi f_\phi-\frac{1}{r^2\sin(\theta)}\partial_\phi\partial_\theta f_\phi+\frac{1}{r^2\sin^2(\theta)}\partial_\phi^2 f_\theta+\frac{1}{r}\partial_r^2(rf_\theta)\\
&=\frac{2}{r^2}\partial_\theta f_r-\frac{1}{r^2}\cot^2(\theta)f_\theta-\frac{\cot(\theta)}{r^2}\partial_\theta f_\theta-\frac{1}{r^2\sin(\theta)}\sin(\theta)f_\theta+\frac{1}{r^2\sin(\theta)}\cos(\theta)\partial_\theta f_\theta\\
&+\frac{1}{r^2\sin(\theta)}\cos(\theta)\partial_\theta f_\theta+\frac{1}{r^2\sin(\theta)}\sin(\theta)\partial_\theta^2 f_\theta-\frac{2}{r^2}\frac{\cot(\theta)}{\sin(\theta)}\partial_\phi f_\phi+\frac{1}{r^2\sin(\theta)}\partial_\theta\partial_\phi f_\phi\\
&-\frac{1}{r^2\sin(\theta)}\partial_\phi\partial_\theta f_\phi+\frac{1}{r^2\sin^2(\theta)}\partial_\phi^2 f_\theta+\frac{1}{r}\partial_r^2(rf_\theta)\\
&=\frac{2}{r^2}\partial_\theta f_r-\frac{\cot^2(\theta)}{r^2}f_\theta+\frac{\cot(\theta)}{r^2}\partial_\theta f_\theta-\frac{1}{r^2}f_\theta+\frac{1}{r^2}\partial_\theta^2 f_\theta-\frac{2}{r^2}\frac{\cot(\theta)}{\sin(\theta)}\partial_\phi f_\phi+\frac{1}{r^2\sin^2(\theta)}\partial_\phi^2 f_\theta+\frac{1}{r}\partial_r^2(rf_\theta)\\
&=\frac{2}{r^2}\partial_\theta f_r-\frac{1}{r^2}(\cot^2(\theta)-1)f_\theta+\frac{\cot(\theta)}{r^2}\partial_\theta f_\theta+\frac{1}{r^2}\partial_\theta^2 f_\theta-\frac{2}{r^2}\frac{\cot(\theta}{\sin(\theta)}\partial_\phi f_\phi+\frac{1}{r^2\sin^2(\theta)}\partial_\phi^2 f_\theta+\frac{1}{r}\partial_r^2(rf_\theta)\\
&=\frac{2}{r^2}\partial_\theta f_r-\frac{1}{r^2\sin^2(\theta)}f_\theta+\frac{\cot(\theta)}{r^2}\partial_\theta f_\theta+\frac{1}{r^2}\partial_\theta^2 f_\theta-\frac{2}{r^2}\frac{\cot(\theta)}{\sin(\theta)}\partial_\phi f_\phi+\frac{1}{r^2\sin^2(\theta)}\partial_\phi^2 f_\theta+\frac{1}{r}\partial_r^2(rf_\theta)
\end{aligned}
\tag{6.5.429}
$$

and finally a last effort for the third and last line:

$$
\begin{aligned}
&=\frac{1}{r\sin(\theta)}\partial_\phi\left(\frac{1}{r^2}\partial_r(r^2f_r)\right)+\frac{1}{r\sin(\theta)}\partial_\phi\left(\frac{1}{r\sin(\theta)}\partial_\theta(\sin(\theta)f_\theta)\right)+\frac{1}{r\sin(\theta)}\partial_\phi\left(\frac{1}{r\sin(\theta)}\partial_\phi f_\phi\right)\\
&-\frac{1}{r}\partial_r\left(r\left(\frac{1}{r\sin(\theta)}\partial_\phi f_r-\frac{1}{r}\partial_r(rf_\phi)\right)\right)+\frac{1}{r}\partial_\theta\left(\frac{1}{r\sin(\theta)}\partial_\theta(\sin(\theta)f_\phi)-\frac{1}{r\sin(\theta)}\partial_\phi f_\theta\right)\\
&=\frac{1}{r^3\sin(\theta)}\partial_\phi\partial_r r^2f_r+\frac{1}{r^2\sin^2(\theta)}\partial_\phi\partial_\theta\sin(\theta)f_\theta+\frac{1}{r^2\sin^2(\theta)}\partial^2_\phi f_\phi\\
&-\frac{1}{r}\partial_r\left(\frac{1}{\sin(\theta)}\partial_\phi f_r-\partial_r rf_\phi\right)+\frac{1}{r}\partial_\theta\frac{1}{r\sin(\theta)}\partial_\theta f_\theta\\
&=\frac{1}{r^3\sin(\theta)}\partial_\phi 2rf_r+\frac{1}{r^3\sin(\theta)}\partial_\phi r^2\partial_r f_r+\frac{1}{r^2\sin^2(\theta)}\partial_\phi+\cos(\theta)f_\theta+\frac{1}{r^2\sin^2(\theta)}\partial_\phi\sin(\theta)\partial_\theta f_\theta\\
&+\frac{1}{r^2\sin^2(\theta)}\partial^2_\phi f_\phi-\frac{1}{r\sin(\theta)}\partial_r\partial_\phi f_r+\frac{1}{r}\partial_r\partial_r f_\phi\\
&-\frac{1}{r}\frac{\cos(\theta)}{r\sin^2(\theta)}\partial_\theta(\sin(\theta)f_\phi)+\frac{1}{r^2\sin(\theta)}\partial_\theta\partial_\theta(\sin(\theta)f_\phi)+\frac{1}{r}\frac{\cos(\theta)}{r\sin^2(\theta)}\partial_\phi f_\theta-\frac{1}{r^2\sin(\theta)}\partial_\theta\partial_\phi f_\theta\\
&=\frac{2}{r^2\sin(\theta)}\partial_\phi f_r+\frac{1}{r\sin(\theta)}\partial_\phi\partial_r f_r+\frac{\cot(\theta)}{r^2\sin(\theta)}\partial_\phi f_\theta+\frac{1}{r^2\sin(\theta)}\partial_\phi\partial_\theta f_\theta\\
&=\frac{2}{r^2\sin(\theta)}\partial_\phi f_r+\frac{1}{r\sin(\theta)}\partial_\phi\partial_r f_r+\frac{\cot(\theta)}{r^2\sin(\theta)}\partial_\phi f_\theta+\frac{1}{r^2\sin(\theta)}\partial_\phi\partial_\theta f_\theta\\
&+\frac{1}{r^2\sin^2(\theta)}\partial^2_\phi-\partial 1r\sin(\theta)\partial_r\partial_\phi f_r+\frac{1}{r}\partial_r\partial_r rf_\phi\\
&=-\frac{\cot(\theta)}{r^2\sin(\theta)}\partial_\theta(\sin(\theta)f_\phi)+\frac{1}{r^2\sin(\theta)}\partial_\theta\partial_\theta(\sin(\theta)f_\phi)+\frac{\cot(\theta)}{r^2\sin(\theta)}\partial_\phi f_\theta-\frac{1}{r^2\sin(\theta)}\partial_\theta\partial_\phi f_\theta\\
&=\frac{2}{r^2\sin(\theta)}\partial_\phi f_r+\frac{\cot(\theta)}{r^2\sin(\theta)}\partial_\phi f_\theta+\frac{1}{r^2\sin(\theta)}\partial_\phi\partial_\theta f_\theta+\frac{1}{r^2\sin^2(\theta)}\partial^2_\phi f_\phi+\frac{1}{r}\partial_r\partial_r rf_\phi\\
&-\frac{\cot(\theta)}{r^2\sin(\theta)}\cos(\theta)f_\phi-\frac{\cot(\theta)}{r^2\sin(\theta)}\sin(\theta)\partial_\theta f_\phi+\frac{1}{r^2\sin(\theta)}\partial_\theta\cos(\theta)f_\phi+\frac{1}{r^2\sin(\theta)}\partial_\theta\sin(\theta)\partial_\theta f_\phi\\
&+\frac{\cot(\theta)}{r^2\sin(\theta)}\partial_\phi f_\theta-\frac{1}{r^2\sin(\theta)}\partial_\theta\partial_\phi\partial_\theta\\
&=\frac{2}{r^2\sin(\theta)}\partial_\phi f_r+\frac{2\cot(\theta)}{r^2\sin(\theta)}\partial_\phi f_\theta+\frac{1}{r^2\sin^2(\theta)}\partial^2_\phi f_\phi+\frac{1}{r}\partial_r\partial_r rf_\phi\\
&-\frac{\cot(\theta)}{r^2\sin(\theta)}\cos(\theta)f_\phi-\frac{\cot(\theta)}{r^2\sin(\theta)}\sin(\theta)\partial_\theta f_\phi+\frac{1}{r^2\sin(\theta)}\partial_\theta\cos(\theta)f_\phi+\frac{1}{r^2\sin(\theta)}\partial_\theta\sin(\theta)\partial_\theta f_\phi\\
&=\frac{2}{r^2\sin(\theta)}\partial_\phi f_r+\frac{2\cot(\theta)}{r^2\sin(\theta)}\partial_\phi f_\theta+\frac{1}{r^2\sin^2(\theta)}\partial^2_\phi f_\phi+\frac{1}{r}\partial_r\partial_r f_\phi+\frac{1}{r}\partial_r r\partial_r f_\phi\\
&-\frac{\cot^2(\theta)}{r^2}f_\phi-\frac{\cot(\theta)}{r^2}\partial_\theta f_\phi+\frac{1}{r^2\sin(\theta)}\partial_\theta\cos(\theta)f_\phi+\frac{1}{r^2\sin(\theta)}\partial_\theta\sin(\theta)\partial_\theta f_\phi\\
&=\frac{2}{r^2\sin(\theta)}\partial_\phi f_r+\frac{2\cot(\theta)}{r^2\sin(\theta)}\partial_\phi f_\theta+\frac{1}{r^2\sin^2(\theta)}\partial^2_\phi f_\phi+\frac{1}{r}\partial_r\partial_r f_\phi+\frac{1}{r}\partial_r f_\phi+\partial^2_r f_\phi-\frac{\cot^2(\theta)}{r^2}f_\phi\\
&-\frac{\cot(\theta)}{r^2}\partial_\theta f_\phi-\frac{1}{r^2\sin(\theta)}\sin(\theta)f_\phi+\frac{1}{r^2\sin(\theta)}\cos(\theta)\partial_\theta f_\phi+\frac{1}{r^2\sin(\theta)}\cos(\theta)\partial_\theta f_\phi\\
&+\frac{1}{r^2\sin(\theta)}\sin(\theta)\partial^2_\theta f_\phi
\end{aligned}
\tag{6.5.430}
$$

$$\begin{aligned}
&=\frac{2}{r^2\sin(\theta)}\partial_\phi f_r+\frac{2\cot(\theta)}{r^2\sin(\theta)}\partial_\phi f_\theta+\frac{1}{r^2\sin^2(\theta)}\partial_\phi^2 f_\phi+\underbrace{\frac{2}{r}\partial_r f_\phi+\frac{1}{r}\partial_r^2 f_\phi}=\frac{1}{r}\partial_r^2 r f_\phi\\
&-\frac{\cot^2(\theta)}{r^2}f_\phi-\frac{\cot(\theta)}{r^2}\partial_\theta f_\phi-\frac{1}{r^2}f_\phi+\frac{\cot(\theta)}{r^2}\partial_\theta f_\phi+\frac{\cot(\theta)}{r^2}\partial_\theta f_\phi+\frac{1}{r^2}\partial_\theta^2 f_\phi\\
&=\frac{2}{r^2\sin(\theta)}\partial_\phi f_r+\frac{2\cot(\theta)}{r^2\sin(\theta)}\partial_\phi f_\theta+\frac{1}{r^2\sin^2(\theta)}\partial_\phi^2 f_\phi+\frac{1}{r}\partial_r^2 r f_\phi-\frac{1}{r^2}f_\phi(\cot^2(\theta)+1)\\
&+\frac{\cot(\theta)}{r^2}\partial_\theta f_\phi+\frac{1}{r^2}\partial_\theta^2 f_\phi
\end{aligned}\tag{6.5.431}$$

So what gives for the "vector Laplacian in spherical coordinates" as we can find it in tables or forms:

$$\boxed{\Delta\vec{f}=\begin{pmatrix}
-\dfrac{2f_r}{r^2}+\dfrac{\partial_r^2 r f_r}{r^2}-\dfrac{2\cot(\theta)f_\theta}{r^2}-\dfrac{2\partial_\theta f_\theta}{r^2}+\dfrac{\cot(\theta)\partial_\theta f_r}{r^2}+\dfrac{\partial_\theta^2 f_r}{r^2}+\dfrac{\partial_\phi^2 f_r}{r^2\sin^2(\theta)}-\dfrac{2\partial_\phi f_\phi}{r^2\sin^2(\theta)}\\
\dfrac{2\partial_\theta f_r}{r^2}-\dfrac{f_\theta}{r^2\sin^2(\theta)}+\dfrac{\cot(\theta)\partial_\theta f_\theta}{r^2}+\dfrac{\partial_\theta^2 f_\theta}{r^2}-\dfrac{2\cot(\theta)\partial_\phi f_\phi}{r^2\sin(\theta)}+\dfrac{\partial_\phi^2 f_\theta}{r^2\sin^2(\theta)}+\dfrac{\partial_r^2 r f_\theta}{r}\\
\dfrac{2\partial_\phi f_r}{r^2\sin(\theta)}+\dfrac{2\cot(\theta)\partial_\phi f_\theta}{r^2\sin(\theta)}+\dfrac{\partial_\phi^2 f_\phi}{r^2\sin^2(\theta)}+\dfrac{\partial_r^2 r f_\phi}{r}-\dfrac{f_\phi}{r^2\sin^2(\theta)}+\dfrac{\cot(\theta)\partial_\theta f_\phi}{r^2}+\dfrac{\partial_\theta^2 f_\phi}{r^2}
\end{pmatrix}}\tag{6.5.432}$$

that's it... for the sceptics...

6.5.8.7 Remarkable Identities

The scalar and vector differential operators have some very simple remarkable identities that we will find very often in physics in this textbook for practical purposes.

Let us first see the relations that make no sense (in case you would fall on them without purpose...):

$$\text{rot}(\text{div}(\vec{f}))\quad\text{or}\quad\vec{\nabla}\times(\vec{\nabla}\circ\vec{f})\tag{6.5.433}$$

For the relation above, the rotational (curl) of a divergence does not exist since the rotational operator applies to a vector field while the divergence is a scalar!

$$\text{rot}(\Delta\cdot f)\quad\text{or}\quad\vec{\nabla}\times(\Delta\cdot f)\tag{6.5.434}$$

For the above relation the rotational (curl) of a scalar Laplacian does not exist since the rotational operator applies to a vector field while by construction, the Laplacian is a scalar.

This done...! Let us now see some remarkable identities without proof for the majority (if there is a proof this is because one reader did the request to have all the details...):

1. By construction the scalar Laplacian is the divergence of the gradient of the scalar field:

$$\text{div}\left(\overrightarrow{\text{grad}}(f)\right)=\vec{\nabla}\circ\vec{\nabla}(f)=\vec{\nabla}^2 f=\Delta\cdot f\tag{6.5.435}$$

2. The rotational (curl) of the gradient is equal to zero:

$$\text{rot}\left(\overrightarrow{\text{grad}}(f)\right)=\vec{\nabla}\times(\vec{\nabla}f)=\vec{0} \tag{6.5.436}$$

Therefore if the rotational (curl) of a vector variable (vector field) is zero, this same variable can be expressed as the gradient of a scalar potential!!!!!!!!!!!! This is a veeeeerrrry important property (or trick depending of the point of view...) in Electromagnetics and Fluid Mechanics and Quantum Physics!

Proof 6.37.2.

$$\vec{\nabla}\times(\vec{\nabla}f)=\vec{\nabla}\times\begin{pmatrix}\dfrac{\partial f}{\partial x}\\\dfrac{\partial f}{\partial y}\\\dfrac{\partial f}{\partial z}\end{pmatrix}=\begin{pmatrix}\dfrac{\partial}{\partial y}\left(\dfrac{\partial f}{\partial z}\right)-\dfrac{\partial}{\partial z}\left(\dfrac{\partial f}{\partial y}\right)\\\dfrac{\partial}{\partial z}\left(\dfrac{\partial f}{\partial x}\right)-\dfrac{\partial}{\partial x}\left(\dfrac{\partial f}{\partial z}\right)\\\dfrac{\partial}{\partial x}\left(\dfrac{\partial f}{\partial y}\right)-\dfrac{\partial}{\partial y}\left(\dfrac{\partial f}{\partial x}\right)\end{pmatrix}=\vec{0} \tag{6.5.437}$$

■ Q.E.D.

3. The curl of a cross product, relation used sometimes in fluid mechanics (we will use it also!), is equal to:

$$\vec{\nabla}\times\left(\vec{f}\times\vec{g}\right)=\left(\vec{g}\circ\vec{\nabla}\right)\vec{f}-\vec{g}\left(\vec{\nabla}\circ\vec{f}\right)-\left(\vec{f}\circ\vec{\nabla}\right)\vec{g}+\vec{f}\left(\vec{\nabla}\circ\vec{g}\right) \tag{6.5.438}$$

Let us derive this:

Proof 6.37.3.

$$\begin{aligned}
\vec{\nabla}\times(\vec{f}\times\vec{g}) =& \nabla\times\left((f_yg_z-f_zg_y)\,\vec{i}+(f_zg_x-f_xg_z)\,\vec{j}+(f_xg_y-f_yg_x)\,\vec{k}\right)\\
=&\left(\frac{\partial}{\partial y}(f_xg_y-f_yg_x)-\frac{\partial}{\partial z}(f_zg_x-f_xg_z)\right)\vec{i}\\
&+\left(\frac{\partial}{\partial z}(f_yg_z-f_zg_y)-\frac{\partial}{\partial x}(f_xg_y-f_yg_x)\right)\vec{j}\\
&+\left(\frac{\partial}{\partial x}(f_zg_x-f_xg_z)-\frac{\partial}{\partial y}(f_yg_z-f_zg_y)\right)\vec{k}\\
=&\left(\frac{\partial f_xg_y}{\partial y}-\frac{\partial f_yg_x}{\partial y}-\frac{\partial f_zg_x}{\partial z}+\frac{\partial f_xg_z}{\partial z}\right)\vec{i}\\
&+\left(\frac{\partial f_yg_z}{\partial z}-\frac{\partial f_zg_y}{\partial z}-\frac{\partial f_xg_y}{\partial x}+\frac{\partial f_yg_x}{\partial x}\right)\vec{j}\\
&+\left(\frac{\partial f_zg_x}{\partial x}-\frac{\partial f_xg_z}{\partial x}-\frac{\partial f_yg_z}{\partial y}+\frac{\partial f_zg_y}{\partial y}\right)\vec{k}\\
=&\left(f_x\frac{\partial g_y}{\partial y}+\frac{\partial f_x}{\partial y}g_y-f_y\frac{\partial g_x}{\partial y}-\frac{\partial f_y}{\partial y}g_x-f_z\frac{\partial g_x}{\partial z}-\frac{\partial f_z}{\partial z}g_x+f_x\frac{\partial g_z}{\partial z}+\frac{\partial f_x}{\partial z}g_z\right)\vec{i}\\
&+\left(f_y\frac{\partial g_z}{\partial z}+\frac{\partial f_y}{\partial z}g_z-f_z\frac{\partial g_y}{\partial z}-\frac{\partial f_z}{\partial z}g_y-f_x\frac{\partial g_y}{\partial x}-\frac{\partial f_x}{\partial x}g_y+f_y\frac{\partial g_x}{\partial x}+\frac{\partial f_y}{\partial x}g_x\right)\vec{j}\\
&+\left(f_z\frac{\partial g_x}{\partial x}+\frac{\partial f_z}{\partial x}g_x-f_x\frac{\partial g_z}{\partial x}-\frac{\partial f_x}{\partial x}g_z-f_y\frac{\partial g_z}{\partial y}-\frac{\partial f_y}{\partial y}g_z+f_z\frac{\partial g_y}{\partial y}+\frac{\partial f_z}{\partial y}g_y\right)\vec{k}\\
=&\left(f_x\left(\frac{\partial g_y}{\partial y}+\frac{\partial g_z}{\partial z}\right)-\left(f_y\frac{\partial g_x}{\partial y}+f_z\frac{\partial g_x}{\partial z}\right)-\left(\frac{\partial f_y}{\partial y}+\frac{\partial f_z}{\partial z}\right)g_x+\left(\frac{\partial f_x}{\partial y}g_y+\frac{\partial f_x}{\partial z}g_z\right)\right)\vec{i}\\
&+\left(f_y\left(\frac{\partial g_x}{\partial x}+\frac{\partial g_z}{\partial z}\right)-\left(f_x\frac{\partial g_y}{\partial x}+f_z\frac{\partial g_y}{\partial z}\right)-\left(\frac{\partial f_x}{\partial x}+\frac{\partial f_z}{\partial z}\right)g_y+\left(\frac{\partial f_y}{\partial x}g_x+\frac{\partial f_y}{\partial z}g_z\right)\right)\vec{j}\\
&+\left(f_z\left(\frac{\partial g_x}{\partial x}+\frac{\partial g_y}{\partial y}\right)-\left(f_x\frac{\partial g_z}{\partial x}+f_y\frac{\partial g_z}{\partial y}\right)-\left(\frac{\partial f_x}{\partial x}+\frac{\partial f_y}{\partial y}\right)g_z+\left(\frac{\partial f_z}{\partial x}g_x+\frac{\partial f_z}{\partial y}g_y\right)\right)\vec{k}\\
=&\left(f_x\left(\frac{\partial g_x}{\partial x}+\frac{\partial g_y}{\partial y}+\frac{\partial g_z}{\partial z}\right)-\left(f_x\frac{\partial g_x}{\partial x}+f_y\frac{\partial g_x}{\partial y}+f_z\frac{\partial g_x}{\partial z}\right)-\left(\frac{\partial f_x}{\partial x}+\frac{\partial f_y}{\partial y}+\frac{\partial f_z}{\partial z}\right)g_x+\left(\frac{\partial f_x}{\partial x}g_x+\frac{\partial f_x}{\partial y}g_y+\frac{\partial f_x}{\partial z}g_z\right)\right)\vec{i}\\
&+\left(f_y\left(\frac{\partial g_x}{\partial x}+\frac{\partial g_y}{\partial y}+\frac{\partial g_z}{\partial z}\right)-\left(f_x\frac{\partial g_y}{\partial x}+f_y\frac{\partial g_y}{\partial y}+f_z\frac{\partial g_y}{\partial z}\right)-\left(\frac{\partial f_x}{\partial x}+\frac{\partial f_y}{\partial y}+\frac{\partial f_z}{\partial z}\right)g_y+\left(\frac{\partial f_y}{\partial x}g_x+\frac{\partial f_y}{\partial y}g_y+\frac{\partial f_y}{\partial z}g_z\right)\right)\vec{j}\\
&+\left(f_z\left(\frac{\partial g_x}{\partial x}+\frac{\partial g_y}{\partial y}+\frac{\partial g_z}{\partial z}\right)-\left(f_x\frac{\partial g_z}{\partial x}+f_y\frac{\partial g_z}{\partial y}+f_z\frac{\partial g_z}{\partial z}\right)-\left(\frac{\partial f_x}{\partial x}+\frac{\partial f_y}{\partial y}+\frac{\partial f_z}{\partial z}\right)g_z+\left(\frac{\partial f_z}{\partial x}g_x+\frac{\partial f_z}{\partial y}g_y+\frac{\partial f_z}{\partial z}g_z\right)\right)\vec{k}\\
=&\left(g_x\frac{\partial}{\partial x}+g_y\frac{\partial}{\partial y}+g_z\frac{\partial}{\partial z}\right)\left(f_x\vec{i}+f_y\vec{j}+f_z\vec{k}\right)-\left(g_x\vec{i}+g_y\vec{j}+g_z\vec{k}\right)\left(\frac{\partial f_x}{\partial x}+\frac{\partial f_y}{\partial y}+\frac{\partial f_z}{\partial z}\right)\\
&-\left(f_x\frac{\partial}{\partial x}+f_y\frac{\partial}{\partial y}+f_z\frac{\partial}{\partial z}\right)\left(g_x\vec{i}+g_y\vec{j}+g_z\vec{k}\right)+\left(f_x\vec{i}+f_y\vec{j}+f_z\vec{k}\right)\left(\frac{\partial g_x}{\partial x}+\frac{\partial g_y}{\partial y}+\frac{\partial g_z}{\partial z}\right)\\
=&\left(\vec{g}\circ\vec{\nabla}\right)\vec{f}-\vec{g}\left(\vec{\nabla}\circ\vec{f}\right)-\left(\vec{f}\circ\vec{\nabla}\right)\vec{g}+\vec{f}\left(\vec{\nabla}\circ\vec{g}\right)
\end{aligned}\tag{6.5.439}$$

■ Q.E.D.

4. The dot product of two rotational is equal to something boring to say just with words...:

$$(\vec{\nabla}\times\vec{x})\circ(\vec{\nabla}\times\vec{y})=\vec{y}\circ(\vec{\nabla}\times(\vec{\nabla}\times\vec{x}))+\vec{\nabla}\circ(\vec{y}\times(\vec{\nabla}\times\vec{x}))\tag{6.5.440}$$

5. The divergence of the rotational (curl) of a vector field is always equal to zero:

$$\text{div}\left(\text{rot}(\vec{f})\right)=0\quad\text{or}\quad\vec{\nabla}\circ(\vec{\nabla}\times\vec{f})=0\tag{6.5.441}$$

Proof 6.37.4.

$$\begin{aligned}
&\vec{\nabla}\circ(\vec{\nabla}\times\vec{f})=\vec{\nabla}\circ\left(\begin{pmatrix}\partial_x\\\partial_y\\\partial_z\end{pmatrix}\times\begin{pmatrix}f_1\\f_2\\f_3\end{pmatrix}\right)=\vec{\nabla}\circ\begin{pmatrix}\partial_yf_3-\partial_zf_2\\\partial_zf_1-\partial_xf_3\\\partial_xf_2-\partial_yf_1\end{pmatrix}\\
&=\begin{pmatrix}\partial_x\\\partial_y\\\partial_z\end{pmatrix}\circ\begin{pmatrix}\partial_yf_3-\partial_zf_2\\\partial_zf_1-\partial_xf_3\\\partial_xf_2-\partial_yf_1\end{pmatrix}=\partial_x(\partial_yf_3-\partial_zf_2)+\partial_y(\partial_zf_1-\partial_xf_3)+\partial_z(\partial_xf_2-\partial_yf_1)\\
&=\bcancel{\partial_x\partial_yf_3}-\cancel{\partial_x\partial_zf_2}+\xcancel{\partial_y\partial_z\partial_zf_1}-\bcancel{\partial_y\partial_xf_3}+\cancel{\partial_z\partial_xf_2}-\xcancel{\partial_z\partial_yf_1}=0
\end{aligned}\tag{6.5.442}$$

■ Q.E.D.

6. The rotational (curl) of the rotational of a vector field is equal to the gradient of the divergence of this vector field less its vector Laplacian:

$$\text{rot}(\text{rot}(\vec{f})) = \overrightarrow{\text{grad}}(\text{div}(\vec{f})) - \Delta\vec{f} \quad \text{or} \quad \vec{\nabla} \times (\vec{\nabla} \times \vec{f}) = \vec{\nabla}(\vec{\nabla} \circ \vec{f}) - \Delta\vec{f} \tag{6.5.443}$$

or also sometimes denoted:

$$\vec{\nabla} \times (\vec{\nabla} \times \vec{f}) = \vec{\nabla}(\vec{\nabla} \circ \vec{f}) - \nabla^2\vec{f} \tag{6.5.444}$$

That under the following rearrangement:

$$\nabla^2\vec{f} = \vec{\nabla} \times (\vec{\nabla} \times \vec{f}) - \vec{\nabla}(\vec{\nabla} \circ \vec{f}) \tag{6.5.445}$$

is named the "Hemholtz decomposition".

Proof 6.37.5.

$$\begin{aligned}\vec{\nabla} \times (\vec{\nabla} \times \vec{f}) &= \vec{\nabla} \times \left(\begin{pmatrix}\partial_x \\ \partial_y \\ \partial_z\end{pmatrix} \times \begin{pmatrix}f_1 \\ f_2 \\ f_3\end{pmatrix}\right) = \vec{\nabla} \times \begin{pmatrix}\partial_y f_3 - \partial_z f_2 \\ \partial_z f_1 - \partial_x f_3 \\ \partial_x f_2 - \partial_y f_1\end{pmatrix} = \begin{pmatrix}\partial_x \\ \partial_y \\ \partial_z\end{pmatrix} \times \begin{pmatrix}\partial_y f_3 - \partial_z f_2 \\ \partial_z f_1 - \partial_x f_3 \\ \partial_x f_2 - \partial_y f_1\end{pmatrix} \\ &= \begin{pmatrix}\partial_y(\partial_x f_2 - \partial_y f_1) - \partial_z(\partial_z f_1 - \partial_x f_3) \\ \partial_z(\partial_y f_3 - \partial_z f_2) - \partial_x(\partial_x f_2 - \partial_y f_1) \\ \partial_x(\partial_z f_1 - \partial_x f_3) - \partial_y(\partial_y f_3 - \partial_z f_2)\end{pmatrix}\end{aligned} \tag{6.5.446}$$

It is then easy to check that this last equality is equal to:

$$\begin{aligned}\vec{\nabla}(\vec{\nabla} \circ \vec{f}) - \Delta\vec{f} &= \vec{\nabla}\,(\partial_x f_1 + \partial_y f_2 + \partial_z f_3) - \Delta\vec{f} = \begin{pmatrix}\partial_x(\partial_x f_1 + \partial_y f_2 + \partial_z f_3) \\ \partial_y(\partial_x f_1 + \partial_y f_2 + \partial_z f_3) \\ \partial_z(\partial_x f_1 + \partial_y f_2 + \partial_z f_3)\end{pmatrix} - \Delta\vec{f} \\ &= \begin{pmatrix}\partial_x(\partial_x f_1 + \partial_y f_2 + \partial_z f_3) \\ \partial_y(\partial_x f_1 + \partial_y f_2 + \partial_z f_3) \\ \partial_z(\partial_x f_1 + \partial_y f_2 + \partial_z f_3)\end{pmatrix} - \begin{pmatrix}\dfrac{\partial^2 f_1}{\partial x^2} + \dfrac{\partial^2 f_1}{\partial y^2} + \dfrac{\partial^2 f_1}{\partial z^2} \\ \dfrac{\partial^2 f_2}{\partial x^2} + \dfrac{\partial^2 f_2}{\partial y^2} + \dfrac{\partial^2 f_2}{\partial z^2} \\ \dfrac{\partial^2 f_3}{\partial x^2} + \dfrac{\partial^2 f_3}{\partial y^2} + \dfrac{\partial^2 f_3}{\partial z^2}\end{pmatrix} \\ &= \begin{pmatrix}\partial_x(\partial_x f_1 + \partial_y f_2 + \partial_z f_3) - \dfrac{\partial^2 f_1}{\partial x^2} - \dfrac{\partial^2 f_1}{\partial y^2} - \dfrac{\partial^2 f_1}{\partial z^2} \\ \partial_y(\partial_x f_1 + \partial_y f_2 + \partial_z f_3) - \dfrac{\partial^2 f_2}{\partial x^2} - \dfrac{\partial^2 f_2}{\partial y^2} - \dfrac{\partial^2 f_2}{\partial z^2} \\ \partial_z(\partial_x f_1 + \partial_y f_2 + \partial_z f_3) - \dfrac{\partial^2 f_3}{\partial x^2} - \dfrac{\partial^2 f_3}{\partial y^2} - \dfrac{\partial^2 f_3}{\partial z^2}\end{pmatrix} \\ &= \begin{pmatrix}\partial_x(\partial_y f_2 + \partial_z f_3) - \dfrac{\partial^2 f_1}{\partial y^2} - \dfrac{\partial^2 f_1}{\partial z^2} \\ \partial_y(\partial_x f_1 + \partial_z f_3) - \dfrac{\partial^2 f_2}{\partial x^2} - \dfrac{\partial^2 f_2}{\partial z^2} \\ \partial_z(\partial_x f_1 + \partial_y f_2 +) - \dfrac{\partial^2 f_3}{\partial x^2} - \dfrac{\partial^2 f_3}{\partial y^2}\end{pmatrix} = \begin{pmatrix}\partial_y(\partial_x f_2 - \partial_y f_1) - \partial_y(\partial_z f_1 - \partial_x f_3) \\ \partial_z(\partial_y f_3 - \partial_z f_2) - \partial_x(\partial_x f_2 - \partial_y f_1) \\ \partial_x(\partial_z f_1 - \partial_x f_3) - \partial_y(\partial_y f_3 - \partial_z f_2)\end{pmatrix}\end{aligned} \tag{6.5.447}$$

■ Q.E.D.

7. The multiplication of the nabla operator by the dot product of two vectors is equal to ... (see below), which provides a very useful relation in Fluid Mechanics:

$$\vec{\nabla}\cdot(\vec{x}\circ\vec{y})=\vec{x}\times(\vec{\nabla}\times\vec{y})+\vec{y}\times(\vec{\nabla}\times\vec{x})+(\vec{y}\circ\vec{\nabla})\cdot\vec{x}+(\vec{x}\circ\vec{\nabla})\cdot\vec{y} \tag{6.5.448}$$

8. The scalar product of the rotational (curl) of a vector is the difference of the commutated operators such that (we can provide the detail proof on request):

$$\vec{x}\circ(\vec{\nabla}\times\vec{y})=\vec{y}\circ(\vec{\nabla}\times\vec{x})-\vec{\nabla}\circ(\vec{x}\times\vec{y}) \tag{6.5.449}$$

   We will use this last relation in our study of electromagnetic radiation pressure in the section of Electrodynamics (among others...).

9. The gradient of a cross product is the difference of the commutated operators such that (we can also provide the detail proof on request):

$$\vec{\nabla}(\vec{x}\times\vec{y})=\vec{y}\cdot(\text{rot}(\vec{x}))-\vec{x}\cdot(\text{rot}(\vec{y})) \tag{6.5.450}$$

   We will use this last relation in our study of superconductors in the section of Electrokinetics page 4307.

10. Let us consider a vector $\vec{v}$ which each component depends on x,y,z. We have (the left term is named the "material derivative", see section Continuum Mechanics page 3829):

$$(\vec{v}\circ\vec{\nabla})\vec{v}=\frac{1}{2}\vec{\nabla}\vec{v}^2-\vec{v}\times(\vec{\nabla}\times\vec{v}) \tag{6.5.451}$$

   Thus explicitly:

$$\left(\begin{pmatrix}v_x\\v_y\\v_z\end{pmatrix}\circ\begin{pmatrix}\partial_x\\\partial_y\\\partial_z\end{pmatrix}\right)\cdot\begin{pmatrix}v_x\\v_y\\v_z\end{pmatrix}=\frac{1}{2}\begin{pmatrix}\partial_x\\\partial_y\\\partial_z\end{pmatrix}\left\|\begin{pmatrix}v_x\\v_y\\v_z\end{pmatrix}\right\|^2-\begin{pmatrix}v_x\\v_y\\v_z\end{pmatrix}\times\left(\begin{pmatrix}\partial_x\\\partial_y\\\partial_z\end{pmatrix}\times\begin{pmatrix}v_x\\v_y\\v_z\end{pmatrix}\right) \tag{6.5.452}$$

   Or even more explicit:

$$\left(\begin{pmatrix}v_x\\v_y\\v_z\end{pmatrix}\circ\begin{pmatrix}\partial_x\\\partial_y\\\partial_z\end{pmatrix}\right)\cdot\begin{pmatrix}v_x\\v_y\\v_z\end{pmatrix}=\frac{1}{2}\begin{pmatrix}\partial_x\\\partial_y\\\partial_z\end{pmatrix}(v_x^2+v_y^2+v_z^2)-\begin{pmatrix}v_x\\v_y\\v_z\end{pmatrix}\times\left(\begin{pmatrix}\partial_x\\\partial_y\\\partial_z\end{pmatrix}\times\begin{pmatrix}v_x\\v_y\\v_z\end{pmatrix}\right) \tag{6.5.453}$$

   So now to prove that the equality holds, we will develop the left part and right part. So let us begin with the left part:

$$\left(\begin{pmatrix}v_x\\v_y\\v_z\end{pmatrix}\circ\begin{pmatrix}\partial_x\\\partial_y\\\partial_z\end{pmatrix}\right)\cdot\begin{pmatrix}v_x\\v_y\\v_z\end{pmatrix}=(v_x\partial_x+v_y\partial_y+v_z\partial_z)\begin{pmatrix}v_x\\v_y\\v_z\end{pmatrix} \tag{6.5.454}$$

Now let us develop the right part:

$$\begin{aligned}
&\frac{1}{2}\begin{pmatrix}\partial_x\\ \partial_y\\ \partial_z\end{pmatrix}(v_x^2+v_y^2+v_z^2)-\begin{pmatrix}v_x\\ v_y\\ v_z\end{pmatrix}\times\left(\begin{pmatrix}\partial_x\\ \partial_y\\ \partial_z\end{pmatrix}\times\begin{pmatrix}v_x\\ v_y\\ v_z\end{pmatrix}\right)\\
&=\begin{pmatrix}v_x\partial_x v_x+v_y\partial_x v_y+v_z\partial_x v_z\\ v_x\partial_y v_x+v_y\partial_y v_y+v_z\partial_y v_z\\ v_x\partial_z v_x+v_y\partial_z v_y+v_z\partial_z v_z\end{pmatrix}-\begin{pmatrix}v_x\\ v_y\\ v_z\end{pmatrix}\times\left(\begin{pmatrix}\partial_y v_z-\partial_z v_y\\ \partial_z v_x-\partial_x v_z\\ \partial_x v_y-\partial_y v_x\end{pmatrix}\right)\\
&=\begin{pmatrix}v_x\partial_x v_x+v_y\partial_x v_y+v_z\partial_x v_z\\ v_x\partial_y v_x+v_y\partial_y v_y+v_z\partial_y v_z\\ v_x\partial_z v_x+v_y\partial_z v_y+v_z\partial_z v_z\end{pmatrix}-\begin{pmatrix}v_y(\partial_x v_y-\partial_y v_x)-v_z(\partial_z v_x-\partial_x v_z)\\ v_z(\partial_y v_z-\partial_z v_y)-v_x(\partial_x v_y-\partial_y v_x)\\ v_x(\partial_z v_x-\partial_x v_z)-v_y(\partial_y v_z-\partial_z v_y)\end{pmatrix}\\
&=\begin{pmatrix}v_x\partial_x v_x+v_y\partial_x v_y+v_z\partial_x v_z-v_y\partial_x v_y+v_y\partial_y v_x+v_z\partial_z v_x-v_z\partial_x v_z\\ v_x\partial_y v_x+v_y\partial_y v_y+v_z\partial_y v_z-v_z\partial_y v_z+v_z\partial_z v_y+v_x\partial_x v_y-v_x\partial_y v_x\\ v_x\partial_z v_x+v_y\partial_z v_y+v_z\partial_z v_z-v_x\partial_z v_x+v_x\partial_x v_z+v_y\partial_y v_z-v_y\partial_z v_y\end{pmatrix}\\
&=\begin{pmatrix}v_x\partial_x v_x+v_y\partial_y v_x+v_z\partial_z v_x\\ v_y\partial_y v_y+v_z\partial_z v_y+v_x\partial_x v_y\\ v_z\partial_z v_z+v_x\partial_x v_z+v_y\partial_y v_z\end{pmatrix}
\end{aligned}\tag{6.5.455}$$

and then we see that the equality holds!

11. Now let us consider a chain rule sometimes useful in the upcoming sections (again about physics topics!). For this consider we start with:

$$\text{rot}\left(\frac{\vec{f}(x,y,z)}{g(x,y,z)}\right)=\begin{pmatrix}\dfrac{\partial(f_z/g)}{\partial y}-\dfrac{\partial(f_y/g)}{\partial z}\\ \dfrac{\partial(f_x/g)}{\partial z}-\dfrac{\partial(f_z/g)}{\partial x}\\ \dfrac{\partial(f_y/g)}{\partial x}-\dfrac{\partial(f_x/g)}{\partial y}\end{pmatrix}\tag{6.5.456}$$

Applying the quotient chain rule derived in the section of Differential and Integral Calculus:

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{f(x)}{g(x)}\right)=\frac{g(x)f'(x)-f(x)g'(x)}{[g(x)]^2}=\frac{g\partial_x f-f\partial_x g}{g^2}\tag{6.5.457}$$

we then get:

$$\begin{aligned}
\text{rot}\left(\frac{\vec{f}(x,y,z)}{g(x,y,z)}\right)&=\begin{pmatrix}\dfrac{\partial(f_z/g)}{\partial y}-\dfrac{\partial(f_y/g)}{\partial z}\\ \dfrac{\partial(f_x/g)}{\partial z}-\dfrac{\partial(f_z/g)}{\partial x}\\ \dfrac{\partial(f_y/g)}{\partial x}-\dfrac{\partial(f_x/g)}{\partial y}\end{pmatrix}=\frac{1}{g^2}\begin{pmatrix}(g\partial_y f_z-f_z\partial_y g)-(g\partial_z f_y-f_y\partial_z g)\\ (g\partial_z f_x-f_x\partial_z g)-(g\partial_x f_z-f_z\partial_x g)\\ (g\partial_x f_y-f_y\partial_x g)-(g\partial_y f_x-f_x\partial_y g)\end{pmatrix}\\
&=\frac{1}{g}\begin{pmatrix}\partial_y f_z-\partial_z f_y\\ \partial_z f_x-\partial_x f_z\\ \partial_x f_y-\partial_y f_x\end{pmatrix}+\frac{1}{g^2}\begin{pmatrix}-f_z\partial_y g+f_y\partial_z g\\ -f_x\partial_z g+f_z\partial_x g\\ -f_y\partial_x g+f_x\partial_y g\end{pmatrix}=\frac{1}{g}\vec{\nabla}\times\vec{f}-\frac{1}{g^2}\vec{\nabla}(g)\times\vec{f}\\
&=\frac{1}{g}\vec{\nabla}\times\vec{f}+\vec{\nabla}\left(\frac{1}{g}\right)\times\vec{f}
\end{aligned}\tag{6.5.458}$$

So finally:

$$\text{rot}\left(\frac{\vec{f}(x,y,z)}{g(x,y,z)}\right)=\frac{1}{g}\vec{\nabla}\times\vec{f}+\vec{\nabla}(\frac{1}{g})\times\vec{f} \tag{6.5.459}$$

or written:

$$\vec{\nabla}\times\left(\frac{\vec{f}(x,y,z)}{g(x,y,z)}\right)=\frac{1}{g}\vec{\nabla}\times\vec{f}+\vec{\nabla}(\frac{1}{g})\times\vec{f} \tag{6.5.460}$$

6.5.8.8 Summary

As part of this textbook, we will use the different notations presented and summarized in the table below. Their usage gives us the possibility in the context of the different theories to avoid confusion with other mathematics being (tools). It's annoying but we have to do with it.

Type	Expression
Gradient of a scalar field $A:\mathbb{R}\mapsto\mathbb{R}^3$	$\vec{\nabla}(f)\quad \vec{\nabla}\cdot f\quad \vec{\nabla}f\quad \overrightarrow{\text{grad}}(f)$
Gradient of a vector field $A:\mathbb{R}^n\mapsto\mathbb{R}^n\times\mathbb{R}^n$	$\vec{\nabla}(\vec{f})$
Divergence of a vector field $A:\mathbb{R}^n\mapsto\mathbb{R}$	$\vec{\nabla}\circ\vec{f}\quad \text{div}(\vec{f})$
Rotational of a vector field $A:\mathbb{R}^3\mapsto\mathbb{R}^3$	$\vec{\nabla}\times\vec{f}\quad \text{rot}(\vec{f})$
Laplacian of a scalar field $A:\mathbb{R}^n\mapsto\mathbb{R}$	$\Delta\cdot f\quad \vec{\nabla}^2 f\quad \vec{\nabla}\circ(\vec{\nabla}\cdot f)$
Laplacian of a vector field $A:\mathbb{R}^3\mapsto\mathbb{R}^3$	$\Delta\vec{f}$

Table 6.17: Summary of vector differential operators

Now let us do a quick summary of main differential operators:

- The gradient can be assimilated to the "slope" (example: the electric field is the "slope" of the electrostatic potential).

  The various expressions of the gradient operator (placed under the form of the nabla operator) in Cartesian, polar, cylindrical, spherical coordinates are the following:

$$\begin{aligned}
\vec{\nabla}&=\left(\frac{\partial}{\partial x},\frac{\partial}{\partial y},\frac{\partial}{\partial z}\right)\\
\vec{\nabla}&=\left(\frac{\partial}{\partial r},\frac{1}{r}\frac{\partial f}{\partial \phi}\right)\\
\vec{\nabla}&=\left(\frac{\partial}{\partial r},\frac{1}{r}\frac{\partial f}{\partial \phi},\frac{\partial}{\partial z}\right)\\
\vec{\nabla}&=\left(\frac{\partial}{\partial r},\frac{1}{r}\frac{\partial f}{\partial \phi},\frac{1}{r\sin\theta}\frac{\partial}{\partial \theta}\right)
\end{aligned}\tag{6.5.461}$$

- The divergence characterizes a flow of something that comes from somewhere, a source, or who goes to it. If the divergence is different from zero, it means that there is concentration around a point, so the density increases (or decreases, it depends on the sign). It could be the density of electric charges or the mass density. Hence the famous theorem that says that the flow (that which passes through a surface) is equal to the integral of the divergence (what remains).

  The various expressions of the divergence operator (placed under the form of the nabla operator) in Cartesian, polar, cylindrical, spherical coordinates are the following:

$$\begin{aligned}
\vec{\nabla} &= \left(\frac{\partial}{\partial x},\frac{\partial}{\partial y},\frac{\partial}{\partial z}\right)\\
\vec{\nabla} &= \left(\frac{1}{r}\frac{\partial(r}{\partial r},\frac{1}{r}\frac{\partial}{\partial\phi}\right)\\
\vec{\nabla} &= \left(\frac{1}{r}\frac{\partial(r}{\partial r},\frac{1}{r}\frac{\partial}{\partial\phi},\frac{\partial}{\partial z}\right)\\
\vec{\nabla} &= \left(\frac{1}{r^2}\frac{\partial(r^2}{\partial r},\frac{1}{r\sin(\theta)}\frac{\partial(\sin(\theta)}{\partial\theta},\frac{1}{r\sin(\theta)}\frac{\partial}{\partial\phi}\right)
\end{aligned}\tag{6.5.462}$$

- The rotational characterizes the existence of a vortex (widely used in fluid mechanics). If there is a whirlwind, we can follow a flow line on a closed curve (closed: in the differential point of view, not in the geometrical one!) without it change of direction: the circulation will the not be equal to zero (it is equal to integral of the rotational (curl)).

  The various expressions of the rotational (curl) operator (placed under the form of the nabla operator) in Cartesian, cylindrical, spherical coordinates are the following:

$$\begin{aligned}
\text{rot}(\vec{f}) = \vec{\nabla}\times\vec{f} &= \begin{pmatrix}\dfrac{\partial f_z}{\partial y}-\dfrac{\partial f_y}{\partial z}\\ \dfrac{\partial f_x}{\partial z}-\dfrac{\partial f_z}{\partial x}\\ \dfrac{\partial f_y}{\partial x}-\dfrac{\partial f_x}{\partial y}\end{pmatrix}\\
\text{rot}(\vec{f}) = \vec{\nabla}\times\vec{f} &= \begin{pmatrix}\dfrac{1}{r}\dfrac{\partial f_z}{\partial\phi}-\dfrac{\partial f_\phi}{\partial z}\\ \dfrac{\partial f_r}{\partial z}-\dfrac{\partial f_z}{\partial r}\\ \dfrac{1}{r}\dfrac{\partial(rf_\phi)}{\partial r}-\dfrac{1}{r}\dfrac{\partial f_r}{\partial\phi}\end{pmatrix}\\
\text{rot}(\vec{f}) = \vec{\nabla}\times\vec{f} &= \begin{pmatrix}\dfrac{1}{r\sin(\theta)}\dfrac{\partial}{\partial\theta}(\sin(\theta)f_\phi)-\dfrac{1}{r\sin(\theta)}\dfrac{\partial f_\theta}{\partial\phi}\\ \dfrac{1}{r\sin(\theta)}\dfrac{\partial f_r}{\partial\phi}-\dfrac{1}{r}\dfrac{\partial}{\partial r}(rf_\phi)\\ \dfrac{1}{r}\dfrac{\partial}{\partial r}(rf_\theta)-\dfrac{1}{r}\dfrac{\partial f_r}{\partial\theta}\end{pmatrix}
\end{aligned}\tag{6.5.463}$$

- The Laplacian of a scalar field gives a scalar field that measures the difference between the value of the function at a point and its average around that point. In other words, the partial second derivative measure the variations of the slope at the point examined in the immediate surroundings and in one dimension at a time. If the partial second derivative is zero in one direction, then the slope is constant in the immediate surroundings and according to this dimension, this means that the value of the function on the study is equal to the average of his neighbourhood (in one dimension).

  The different expressions of the scalar Laplacian operator (placed under the form of the nabla operator) in Cartesian, polar and spherical coordinates are:

$$\begin{aligned}\vec{\nabla}^2 &= \left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial y^2}\right)\\ \vec{\nabla}^2 &= \left(\frac{\partial^2}{\partial r^2}+\frac{1}{r}\frac{1}{\partial r}+\frac{1}{r^2}\frac{\partial^2}{\partial \phi^2}\right)\\ \vec{\nabla}^2 &= \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right)+\frac{1}{r^2\sin(\theta)}\frac{\partial}{\partial\theta}\left(\sin(\theta)\frac{\partial\theta}{\partial\theta}\right)+\frac{1}{r^2\sin^2(\theta)}\frac{\partial^2}{\partial\phi^2}\end{aligned} \tag{6.5.464}$$

- As for the Laplacian of a scalar field, the Laplacian of a vector field is only a very convenient notation system for condensing the writing of the components of a vector field.

  The different expressions of the vectorial Laplacian operator (placed under the form of the nabla operator) in Cartesian, polar and spherical coordinates are:

$$\begin{aligned}\Delta\vec{f} &= \begin{pmatrix}\displaystyle\sum_{i=1}^{3}\frac{\partial^2 f_1}{\partial x_i^2}\\ \displaystyle\sum_{i=1}^{3}\frac{\partial^2 f_2}{\partial x_i^2}\\ \displaystyle\sum_{i=1}^{3}\frac{\partial^2 f_3}{\partial x_i^2}\end{pmatrix}\\ \Delta\vec{f} &= \begin{pmatrix}-\frac{1}{r^2}f_r+\frac{1}{r}\partial_r f_r+\partial_r^2 f_r-\frac{2}{r^2}\partial_\phi f_\phi+\frac{1}{r^2}\partial_\phi^2 f_r+\partial_z^2 f_r\\ \frac{2}{r^2}\partial_\phi f_r+\frac{1}{r^2}\partial_\phi^2 f_\phi+\partial_z^2 f_\phi+\partial_z^2 f_\phi-\frac{1}{r^2}f_\phi+\frac{1}{r}\partial_r f_\phi+\partial_r^2 f_\phi\\ \partial_z^2 f_z+\frac{1}{r}\partial_r f_z+\partial_r^2 f_z+\frac{1}{r^2}\partial_\phi^2 f_z\end{pmatrix}\\ \Delta\vec{f} &= \begin{pmatrix}-\frac{2f_r}{r^2}+\frac{\partial_r^2 r f_r}{r^2}-\frac{2\cot(\theta)f_\theta}{r^2}-\frac{2\partial_\theta f_\theta}{r^2}+\frac{\cot(\theta)\partial_\theta f_r}{r^2}+\frac{\partial_\theta^2 f_r}{r^2}+\frac{\partial_\phi^2 f_r}{r^2\sin^2(\theta)}-\frac{2\partial_\phi f_\phi}{r^2\sin^2(\theta)}\\ \frac{2\partial_\theta f_r}{r^2}-\frac{f_\theta}{r^2\sin^2(\theta)}+\frac{\cot(\theta)\partial_\theta f_\theta}{r^2}+\frac{\partial_\theta^2 f_\theta}{r^2}-\frac{2\cot(\theta)\partial_\phi f_\phi}{r^2\sin(\theta)}+\frac{\partial_\phi^2 f_\theta}{r^2\sin^2(\theta)}+\frac{\partial_r^2 r f_\theta}{r}\\ \frac{2\partial_\phi f_r}{r^2\sin(\theta)}+\frac{2\cot(\theta)\partial_\phi f_\theta}{r^2\sin(\theta)}+\frac{\partial_\phi^2 f_\phi}{r^2\sin^2(\theta)}+\frac{\partial_r^2 r f_\phi}{r}-\frac{f_\phi}{r^2\sin^2(\theta)}+\frac{\cot(\theta)\partial_\theta f_\phi}{r^2}+\frac{\partial_\theta^2 f_\phi}{r^2}\end{pmatrix}\end{aligned} \tag{6.5.465}$$

And also we have the following list of remarkable identities:

$$\text{rot}\left(\overrightarrow{\text{grad}}(f)\right)=\vec{\nabla}\times(\vec{\nabla}f)=\vec{0} \tag{6.5.466}$$

$$(\vec{\nabla}\times\vec{x})\circ(\vec{\nabla}\times\vec{y})=\vec{y}\circ(\vec{\nabla}\times(\vec{\nabla}\times\vec{x}))+\vec{\nabla}\circ(\vec{y}\times(\vec{\nabla}\times\vec{x})) \tag{6.5.467}$$

$$\text{div}\left(\text{rot}(\vec{f})\right)=0 \quad \text{or} \quad \vec{\nabla}\circ(\vec{\nabla}\times\vec{f})=0 \tag{6.5.468}$$

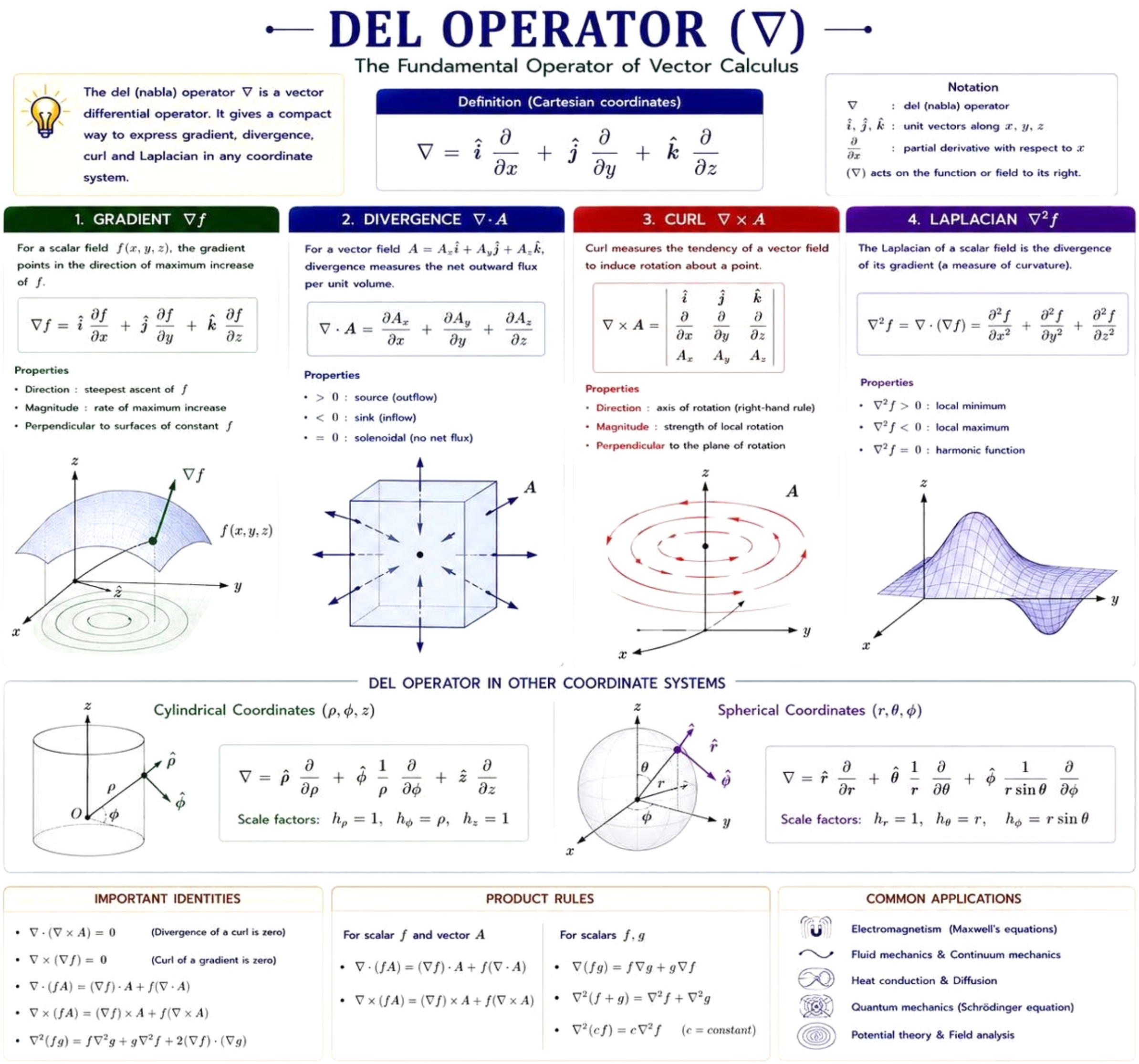
$$\text{rot}(\text{rot}(\vec{f}))=\overrightarrow{\text{grad}}(\text{div}(\vec{f}))-\Delta\vec{f} \quad \text{or} \quad \vec{\nabla}\times(\vec{\nabla}\times\vec{f})=\vec{\nabla}(\vec{\nabla}\circ\vec{f})-\Delta\vec{f} \tag{6.5.469}$$

$$\vec{\nabla}\cdot(\vec{x}\circ\vec{y})=\vec{x}\times(\vec{\nabla}\times\vec{y})+\vec{y}\times(\vec{\nabla}\times\vec{x})+(\vec{y}\circ\vec{\nabla})\cdot\vec{x}+(\vec{x}\circ\vec{\nabla})\cdot\vec{y} \tag{6.5.470}$$

$$\vec{x}\circ(\vec{\nabla}\times\vec{y})=\vec{y}\circ(\vec{\nabla}\times\vec{x})-\vec{\nabla}\circ(\vec{x}\times\vec{y}) \tag{6.5.471}$$

$$\vec{\nabla}(\vec{x}\times\vec{y})=\vec{y}\cdot(\text{rot}(\vec{x}))-\vec{x}\cdot(\text{rot}(\vec{y})) \tag{6.5.472}$$

$$(\vec{v}\circ\vec{\nabla})\vec{v}=\frac{1}{2}\vec{\nabla}\vec{v}^2-\vec{v}\times(\vec{\nabla}\times\vec{v}) \tag{6.5.473}$$



DEL OPERATOR (∇)

The Fundamental Operator of Vector Calculus

The del (nabla) operator ∇ is a vector differential operator. It gives a compact way to express gradient, divergence, curl and Laplacian in any coordinate system.

Definition (Cartesian coordinates)

$$\nabla = \hat{\imath}\,\frac{\partial}{\partial x} + \hat{\jmath}\,\frac{\partial}{\partial y} + \hat{k}\,\frac{\partial}{\partial z}$$

Notation

∇ : del (nabla) operator
$\hat{\imath},\hat{\jmath},\hat{k}$: unit vectors along x, y, z
$\frac{\partial}{\partial x}$: partial derivative with respect to x
(∇) acts on the function or field to its right.

1. GRADIENT ∇f

For a scalar field $f(x,y,z)$, the gradient points in the direction of maximum increase of f.

$$\nabla f = \hat{\imath}\,\frac{\partial f}{\partial x} + \hat{\jmath}\,\frac{\partial f}{\partial y} + \hat{k}\,\frac{\partial f}{\partial z}$$

Properties

- Direction : steepest ascent of f
- Magnitude : rate of maximum increase
- Perpendicular to surfaces of constant f

2. DIVERGENCE $\nabla\cdot A$

For a vector field $A = A_x\hat{\imath} + A_y\hat{\jmath} + A_z\hat{k}$, divergence measures the net outward flux per unit volume.

$$\nabla\cdot A = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z}$$

Properties

- > 0 : source (outflow)
- < 0 : sink (inflow)
- $= 0$: solenoidal (no net flux)

3. CURL $\nabla\times A$

Curl measures the tendency of a vector field to induce rotation about a point.

$$\nabla\times A = \begin{vmatrix} \hat{\imath} & \hat{\jmath} & \hat{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ A_x & A_y & A_z \end{vmatrix}$$

Properties

- **Direction** : axis of rotation (right-hand rule)
- **Magnitude** : strength of local rotation
- **Perpendicular** to the plane of rotation

4. LAPLACIAN $\nabla^2 f$

The Laplacian of a scalar field is the divergence of its gradient (a measure of curvature).

$$\nabla^2 f = \nabla\cdot(\nabla f) = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}$$

Properties

- $\nabla^2 f > 0$: local minimum
- $\nabla^2 f < 0$: local maximum
- $\nabla^2 f = 0$: harmonic function

DEL OPERATOR IN OTHER COORDINATE SYSTEMS

Cylindrical Coordinates (ρ, ϕ, z)

$$\nabla = \hat{\rho}\,\frac{\partial}{\partial \rho} + \hat{\phi}\,\frac{1}{\rho}\,\frac{\partial}{\partial \phi} + \hat{z}\,\frac{\partial}{\partial z}$$

Scale factors: $h_\rho = 1,\ h_\phi = \rho,\ h_z = 1$

Spherical Coordinates (r, θ, ϕ)

$$\nabla = \hat{r}\,\frac{\partial}{\partial r} + \hat{\theta}\,\frac{1}{r}\,\frac{\partial}{\partial \theta} + \hat{\phi}\,\frac{1}{r\sin\theta}\,\frac{\partial}{\partial \phi}$$

Scale factors: $h_r = 1,\ h_\theta = r,\ h_\phi = r\sin\theta$

IMPORTANT IDENTITIES

- $\nabla\cdot(\nabla\times A) = 0$ (Divergence of a curl is zero)
- $\nabla\times(\nabla f) = 0$ (Curl of a gradient is zero)
- $\nabla\cdot(fA) = (\nabla f)\cdot A + f(\nabla\cdot A)$
- $\nabla\times(fA) = (\nabla f)\times A + f(\nabla\times A)$
- $\nabla^2(fg) = f\nabla^2 g + g\nabla^2 f + 2(\nabla f)\cdot(\nabla g)$

PRODUCT RULES

For scalar f and vector A

- $\nabla\cdot(fA) = (\nabla f)\cdot A + f(\nabla\cdot A)$
- $\nabla\times(fA) = (\nabla f)\times A + f(\nabla\times A)$

For scalars f, g

- $\nabla(fg) = f\nabla g + g\nabla f$
- $\nabla^2(f+g) = \nabla^2 f + \nabla^2 g$
- $\nabla^2(cf) = c\nabla^2 f$ (c = constant)

COMMON APPLICATIONS

- Electromagnetism (Maxwell's equations)
- Fluid mechanics & Continuum mechanics
- Heat conduction & Diffusion
- Quantum mechanics (Schrödinger equation)
- Potential theory & Field analysis

Finally let us finish this summary with all the theorems that we have obtained so far in this section that are named "first order integral theorems":

- Gradient theorem (only for uniform field):

$$\iiint_V \vec{\nabla}(f)\mathrm{d}V = \oiint_S f\mathrm{d}\vec{S} \tag{6.5.474}$$

- Ostrogradsky theorem (divergence theorem):

$$\Phi = \iiint_V (\vec{\nabla}\circ\vec{f})\mathrm{d}V = \oiint_S \vec{f}\circ\mathrm{d}\vec{S} \tag{6.5.475}$$

- Green theorem (Green-Riemann theorem):

$$\oint_C (f_x\mathrm{d}x + f_y\mathrm{d}y) = \int_S \left(\frac{\partial f_y}{\partial x} - \frac{\partial f_x}{\partial y}\right)\mathrm{d}S_z \tag{6.5.476}$$

- Stokes theorem (curl theorem):

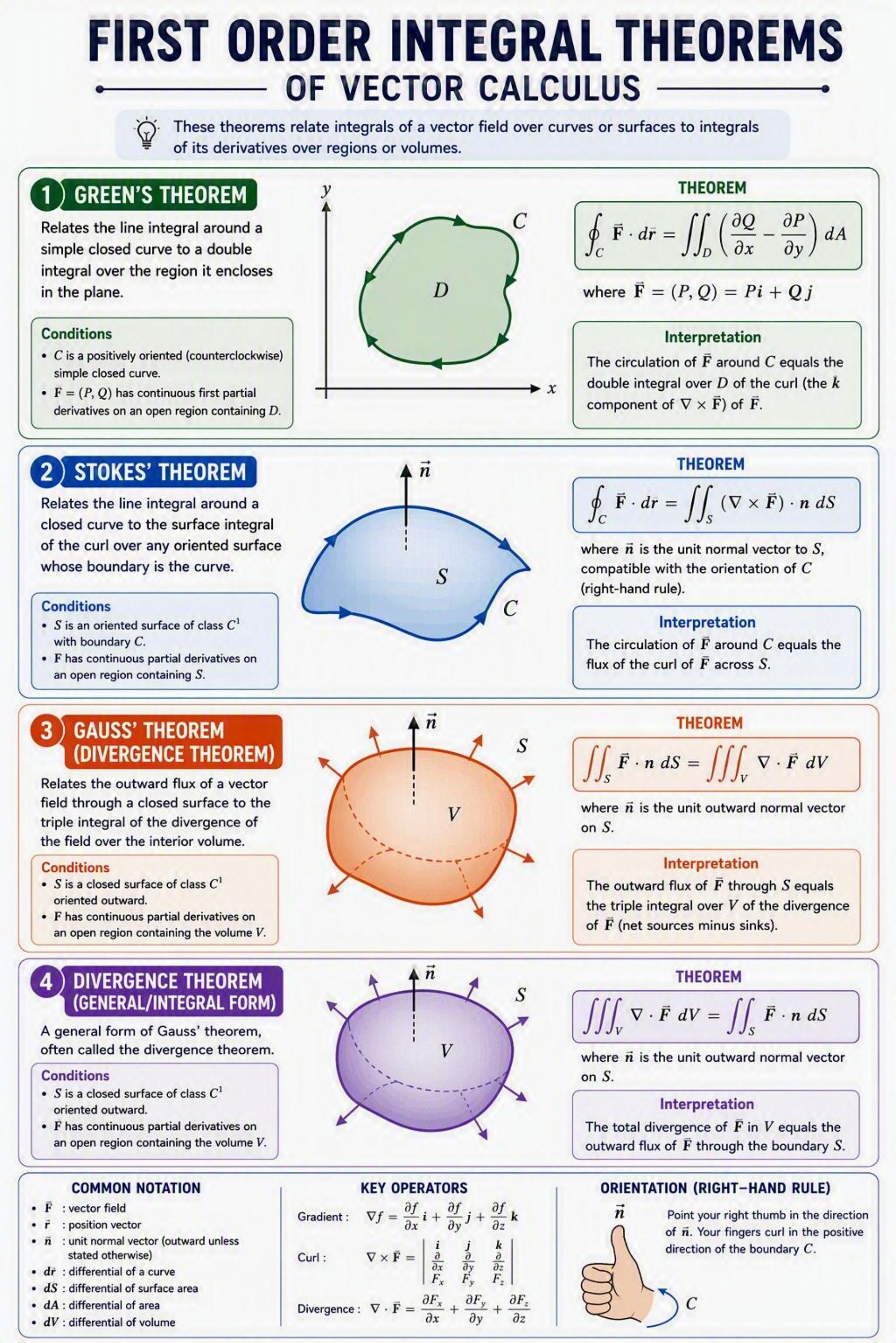
$$\oint \vec{f} \circ \mathrm{d}\vec{r} = \int\limits_S \left(\vec{\nabla} \times \vec{f}\right) \circ \mathrm{d}\vec{S} \qquad (6.5.477)$$

Keep in mind that they are all special case of the Stokes theorem (general manifold theorem[19]) including the fundamental theorem of calculus!

Then the fundamental theorem of calculus (1D), Green's theorem (2D) and the divergence theorem (3D) (i.e. Stokes original theorem) can all be stated respectively by the following parametrised sentence (still special case of Stokes general manifold theorem): *The total [change, spin, flow] outside is equal to the sum of little [changes, spins, flows] on the inside*.

---

[19] If one day we have the time to write the LaTeXto provide a detailed example we will! But actually that doesn't bring too much interest for an engineer.



FIRST ORDER INTEGRAL THEOREMS
OF VECTOR CALCULUS

These theorems relate integrals of a vector field over curves or surfaces to integrals of its derivatives over regions or volumes.

1 GREEN'S THEOREM

Relates the line integral around a simple closed curve to a double integral over the region it encloses in the plane.

Conditions
- C is a positively oriented (counterclockwise) simple closed curve.
- $\mathbf{F} = (P, Q)$ has continuous first partial derivatives on an open region containing D.

THEOREM

$$\oint_C \vec{\mathbf{F}} \cdot d\vec{r} = \iint_D \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA$$

where $\vec{\mathbf{F}} = (P, Q) = P\boldsymbol{i} + Q\,\boldsymbol{j}$

Interpretation

The circulation of $\vec{\mathbf{F}}$ around C equals the double integral over D of the curl (the $\boldsymbol{k}$ component of $\nabla \times \vec{\mathbf{F}}$) of $\vec{\mathbf{F}}$.

2 STOKES' THEOREM

Relates the line integral around a closed curve to the surface integral of the curl over any oriented surface whose boundary is the curve.

Conditions
- S is an oriented surface of class C^1 with boundary C.
- $\mathbf{F}$ has continuous partial derivatives on an open region containing S.

THEOREM

$$\oint_C \vec{\mathbf{F}} \cdot d\vec{r} = \iint_S (\nabla \times \vec{\mathbf{F}}) \cdot \boldsymbol{n}\, dS$$

where $\vec{\boldsymbol{n}}$ is the unit normal vector to S, compatible with the orientation of C (right-hand rule).

Interpretation

The circulation of $\vec{\mathbf{F}}$ around C equals the flux of the curl of $\vec{\mathbf{F}}$ across S.

3 GAUSS' THEOREM (DIVERGENCE THEOREM)

Relates the outward flux of a vector field through a closed surface to the triple integral of the divergence of the field over the interior volume.

Conditions
- S is a closed surface of class C^1 oriented outward.
- $\vec{\mathbf{F}}$ has continuous partial derivatives on an open region containing the volume V.

THEOREM

$$\iint_S \vec{\mathbf{F}} \cdot \boldsymbol{n}\, dS = \iiint_V \nabla \cdot \vec{\mathbf{F}}\, dV$$

where $\vec{\boldsymbol{n}}$ is the unit outward normal vector on S.

Interpretation

The outward flux of $\vec{\mathbf{F}}$ through S equals the triple integral over V of the divergence of $\vec{\mathbf{F}}$ (net sources minus sinks).

4 DIVERGENCE THEOREM (GENERAL/INTEGRAL FORM)

A general form of Gauss' theorem, often called the divergence theorem.

Conditions
- S is a closed surface of class C^1 oriented outward.
- $\vec{\mathbf{F}}$ has continuous partial derivatives on an open region containing the volume V.

THEOREM

$$\iiint_V \nabla \cdot \vec{\mathbf{F}}\, dV = \iint_S \vec{\mathbf{F}} \cdot \boldsymbol{n}\, dS$$

where $\vec{\boldsymbol{n}}$ is the unit outward normal vector on S.

Interpretation

The total divergence of $\vec{\mathbf{F}}$ in V equals the outward flux of $\vec{\mathbf{F}}$ through the boundary S.

COMMON NOTATION
- $\vec{\mathbf{F}}$: vector field
- $\vec{r}$: position vector
- $\vec{\boldsymbol{n}}$: unit normal vector (outward unless stated otherwise)
- $d\vec{r}$: differential of a curve
- dS : differential of surface area
- dA : differential of area
- dV : differential of volume

KEY OPERATORS

Gradient : $\nabla f = \frac{\partial f}{\partial x}\boldsymbol{i} + \frac{\partial f}{\partial y}\boldsymbol{j} + \frac{\partial f}{\partial z}\boldsymbol{k}$

Curl : $\nabla \times \vec{\mathbf{F}} = \begin{vmatrix} \boldsymbol{i} & \boldsymbol{j} & \boldsymbol{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ F_x & F_y & F_z \end{vmatrix}$

Divergence : $\nabla \cdot \vec{\mathbf{F}} = \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + \frac{\partial F_z}{\partial z}$

ORIENTATION (RIGHT–HAND RULE)

Point your right thumb in the direction of $\vec{\boldsymbol{n}}$. Your fingers curl in the positive direction of the boundary C.

~95 %

★★★★★

99 votes, 84.44%

6.6. Linear Algebra

THERE are several approaches to learn Linear Algebra. First a pragmatic way (we'll start with this one because our experience has shown us that it is the one that seemed to work best for students) and a more formal way that we will present in after. We should first warn the reader that linear algebra is a powerful calculation tool that we use enormous in economic and industrial practice in the following areas (see the respective chapters of the book for specific examples): Statistics , Electrical Engineering, Finance market, Numerical optimization methods, Optics, Quantum Physics, Electrodynamics, Relativity, Fluid mechanics, etc. It is then necessary to attach a special attention to this subject.

Remark

Linear algebra quietly underpins a surprising amount of what accountants and even low level financial analysts do because spreadsheet tools like Microsoft® Excel® rely on matrix-style (and vector) operations to process data efficiently! Array formulas, vectorized calculations, and functions such as the famous `SUMPRODUCT( )` essentially perform dot products between matrices and vectors. Understanding the linear algebra beneath these tools helps employees to design typical intermediate level spreadsheets calculations without constantly relying on quite expensive consultants that have an engineering or university degree with a linear algebra background.

First let us answer at two questions from a reader: Why Linear Algebra is named like this? And is there a Non-Linear Algebra?

Here are my answers:

1. This is named "Linear Algebra" because it was first necessary to choose a name... and also because it is a generalization of the scalar algebra but with vectors where applications are no longer scalar functions but matrix applications whose effect is to act as a linear sum of a base vectors (at least this can be interpreted as such).
2. Officially and to my knowledge there is no non-linear algebra in the same philosophy as what we will see in this section. It seem that they are mathematicians who have created "non linear algebra" but these have nothing to do with matrices.

Now, remember that we studied in section Calculus how to determine the intersection (if any exist) of the equation of two lines in $\mathbb{R}^2$ (we can extend the problem obviously to more than two lines), as it is equivalent to solve a polynomial of order 1, given by:

$$y_1 = a_1 x + b_1 \quad \text{and} \quad y_2 = a_2 x + b_2 \tag{6.6.1}$$

where $a_i, b_i \in \mathbb{R}$.

Thus seeking the value of x for which:

$$y_1=y_2 \tag{6.6.2}$$

leads us to write:

$$a_1x+b_1=a_2x+b_2 \Leftrightarrow x=\frac{b_1-b_2}{a_2-a_1} \tag{6.6.3}$$

However, there is another way of presenting the problem as we have seen in the Numerical Methods section (Theoretical Computer chapter). Indeed, we can write the problem in the form of a block of equations:

$$\begin{aligned} y_1&=a_1x+b_1\\ y_2&=a_2x+b_2 \end{aligned} \tag{6.6.4}$$

and as we seek $y_1=y_2=y$, we have:

$$\begin{aligned} y-a_1x&=b_1\\ y-a_2x&=b_2 \end{aligned} \tag{6.6.5}$$

This writing is named as we have seen in the chapter of Theoretical Computing (see section of Numerical Methods page 6304) a "linear system" that we can solve by subtracting or adding the lines between them (all the solutions are always equal), which gives us:

$$\begin{array}{r} y-a_1x=b_1\\ +y-a_2x=b_2\,|\cdot(-1)\\ \hline -a_1x+a_2x=b_1-b_2 \end{array} \tag{6.6.6}$$

and we see that we fall back on the solution:

$$x=\frac{b_1-b_2}{a_2-a_1} \tag{6.6.7}$$

So there are two ways to present a problem of intersection of lines:

1. In the form of an equation
2. In the form of a system

We will focus in a part of this section to the second method that will allow us with tools seen in the section of Vector Calculus to resolve not only the intersections of one or more straight lines but one or more straight, plans, hyperplanes, etc. in respectively $\mathbb{R},\mathbb{R}^2,\ldots,\mathbb{R}^n$.

Obviously we will see that Linear Algebra is not only use for this purpose but can also be used the generalized some physics mathematical models or to express geometrical transformations of vectors or figures in 2D or 3D or also express Markov Chains, special properties of Multivariate Statistics, calculation of some differential equation (for reliability engineering for example) and much more! Many examples are given in this textbook about the application of Linear Algebra in real life.

Before attacking the theoretical part, let us introduce a very interesting example but that requires a concept - the determinant - that we will prove rigorously much further in detail in this section (it seemed to us more pedagogical to approach this subject now rather than the reader must wait to browse dozens of pages of mathematical developments before reaching the rigorous definition of the determinant).

Consider the system of two linear equations with two unknowns (system of intersection of planes):

$$\begin{aligned} a_1x_1+b_1x_2&=c_1\\ a_2x_1+b_2x_2&=c_2 \end{aligned} \tag{6.6.8}$$

If we solve this, we quickly get (technique named "substitution method"):

$$\left.\begin{aligned} x_1&=\frac{c_1-b_1x_2}{a_1}\\ x_2&=\frac{c_2-a_2x_1}{b_2} \end{aligned}\right\} \Rightarrow x_1=\frac{c_1-b_1\left(\dfrac{c_2-a_2x_1}{b_2}\right)}{a_1}=\frac{b_2c_1-b_1c_2+b_1a_2x_1}{a_1b_2} \tag{6.6.9}$$

It comes then:

$$a_1b_2x_1-b_1a_2x_1=x_1(a_1b_2-b_1a_2)=b_2c_1-b_1c_2 \tag{6.6.10}$$

and so at the end:

$$x_1=\frac{b_2c_1-b_1c_2}{a_1b_2-b_1a_2} \tag{6.6.11}$$

and if we define a little bit fast something that is named the "determinant", that we will introduce further below rigorously, as follows:

$$\begin{aligned} D\left(\vec{c},\vec{b}\right)&=D\left(\begin{bmatrix} c_1 & b_1\\ c_2 & b_2\end{bmatrix}\right)=c_1b_2-c_2b_1\\ D\left(\vec{a},\vec{b}\right)&=D\left(\begin{bmatrix} a_1 & b_1\\ a_2 & b_2\end{bmatrix}\right)=a_1b_2-b_1a_2 \end{aligned} \tag{6.6.12}$$

or with another very more common notation:

$$\begin{aligned} \det\left(\vec{c},\vec{b}\right)&=|\vec{c},\vec{b}|=c_1b_2-c_2b_1\\ \det\left(\vec{a},\vec{b}\right)&=|\vec{a},\vec{b}|=a_1b_2-b_1a_2 \end{aligned} \tag{6.6.13}$$

we thus have:

$$x_1=\frac{c_1b_2-c_2b_1}{a_1b_2-a_2b_1}=\frac{\det\left(\vec{c},\vec{b}\right)}{\det\left(\vec{a},\vec{b}\right)} \tag{6.6.14}$$

And by proceeding in the same way we get:

$$\left.\begin{aligned} x_1&=\frac{c_1-b_1x_2}{a_1}\\ x_2&=\frac{c_2-a_2x_1}{b_2} \end{aligned}\right\} \Rightarrow x_2=\frac{c_2-a_2\left(\dfrac{c_1-b_1x_2}{a_1}\right)}{b_2}=\frac{a_1c_2-a_2c_1+a_2b_1x_2}{b_2a_1} \tag{6.6.15}$$

It comes then:

$$b_2a_1x_2-a_2b_1x_2=x_2(b_2a_1-a_2b_1)=a_1c_2-a_2c_1 \tag{6.6.16}$$

and so finally we get:

$$x_1=\frac{c_1b_2-c_2b_1}{a_1b_2-a_2b_1}=\frac{\det\left(\vec{c},\vec{b}\right)}{\det\left(\vec{a},\vec{b}\right)} \quad \text{and} \quad x_2=\frac{a_1c_2-a_2c_1}{a_1b_2-a_2b_1}=\frac{\det\left(\vec{a},\vec{c}\right)}{\det\left(\vec{a},\vec{b}\right)} \tag{6.6.17}$$

It then appears clear that if:

$$\det\left(\vec{a},\vec{b}\right)=\det\left(\vec{c},\vec{b}\right)=\det\left(\vec{a},\vec{c}\right)=0 \tag{6.6.18}$$

the system has infinitely many solutions. In contrast, the system has no solution if:

$$\det\left(\vec{a},\vec{b}\right)=0 \qquad \text{and} \qquad \left[\det\left(\vec{c},\vec{b}\right)\neq 0 \qquad \text{or} \qquad \det\left(\vec{a},\vec{c}\right)\neq 0\right] \tag{6.6.19}$$

And if the reader repeats (happily...) the procedure for a system of three equations with three unknowns of the type (intersection of hyperplanes):

$$\begin{aligned}a_1x_1+b_1x_2+c_1x_3&=d_1\\a_2x_1+b_2x_2+c_2x_3&=d_2\\a_3x_1+b_3x_2+c_3x_3&=d_3\end{aligned} \tag{6.6.20}$$

We then get (after some basic boring algebraic operations):

$$x_1=\frac{\det\left(\vec{d},\vec{b},\vec{c}\right)}{\det\left(\vec{a},\vec{b},\vec{c}\right)},x_2=\frac{\det\left(\vec{a},\vec{d},\vec{c}\right)}{\det\left(\vec{a},\vec{b},\vec{c}\right)},x_3=\frac{\det\left(\vec{a},\vec{b},\vec{d}\right)}{\det\left(\vec{a},\vec{b},\vec{c}\right)} \tag{6.6.21}$$

with:

$$\begin{aligned}\det\left(\vec{a},\vec{b},\vec{c}\right)&=a_1b_2c_3-a_1c_2b_3-b_1a_2c_3+b_1c_2a_3+c_1a_3b_3-c_1b_3a_3\\\det\left(\vec{d},\vec{b},\vec{c}\right)&=d_1b_2c_3-d_1c_2b_3-b_1d_2c_3+b_1c_2d_3+c_1d_2b_3-c_1b_2d_3\\\det\left(\vec{a},\vec{d},\vec{c}\right)&=a_1d_2c_3-a_1c_2d_3-b_1d_2c_3+b_1c_2d_3+c_1d_2b_3-c_1b_2d_3\\\det\left(\vec{a},\vec{b},\vec{d}\right)&=a_1b_2d_3-a_1d_2b_3-b_1a_2d_2+b_1d_2a_3+d_1a_2b_3-d_1b_2a_3\end{aligned} \tag{6.6.22}$$

It then appears clear that if:

$$\det\left(\vec{a},\vec{b},\vec{c}\right)=\det\left(\vec{d},\vec{b},\vec{c}\right)=\det\left(\vec{a},\vec{d},\vec{c}\right)=\det\left(\vec{a},\vec{b},\vec{d}\right)=0 \tag{6.6.23}$$

the system has infinitely many solutions. In contrast, the system has no solution if:

$$\det\left(\vec{a},\vec{b},\vec{c}\right)=0\text{ and }\left[\det\left(\vec{d},\vec{b},\vec{c}\right)\neq 0\text{ or }\det\left(\vec{a},\vec{d},\vec{c}\right)\neq 0\right] \tag{6.6.24}$$

and so on for n equations with n unknowns.

However, there was a condition to satisfy: as we have seen in the previous example, we could not solve a system of equations with two unknowns if we have only one equation. That is why it is necessary and sufficient for a system of equations with n unknowns to have at least n equations. Thus, we speak of "systems with n equations in n unknowns". We also prove that it is necessary and sufficient that the determinant is non-zero for a linear system, whose matrix equivalent is square, to have a unique solution (the concept of "determinant" and "matrix" will be defined further below robustly) and therefore that the corresponding matrix to the system is invertible (non-singular).

6.6.1 Linear Systems

Definitions (#6.6.251):

D1. We name "linear system", or simply "system" any family of equations of the form:

$$\begin{gathered}a_{11}x_1+a_{12}x_2+\ldots+a_{1n}x_n=b_1\\ a_{21}x_1+a_{22}x_2+\ldots+a_{2n}x_n=b_2\\ \ldots\\ a_{m1}x_1+a_{m2}x_2+\ldots+a_{mn}x_n=b_m\end{gathered} \tag{6.6.25}$$

where each line represents the equation of a line, plane or hyperplane (see section Analytical Geometry page 2800), and a_{mn} are the "system coefficients", the b_m the "coefficients of the second member" and the x_n the "unknowns of the system".

D2. If the system has n unknown and n equations and has a unique solution, we then name it a "Cramer system" (11750 according to holocene calendar).

D3. If the coefficients of the second member are all zero, then we say that the system is a "homogeneous system" so it has at least the trivial solution where all x_n are equal to zero. Otherwise we say that we have an "inhomogeneous system".

D4. We name "homogeneous system associated to the system", the system of equations we get by substituting zeros to the coefficients of the second member (b_m).

Let us now recall the following items:

- The linear equation (see section Functional Analysis page 2195) is given by:

$$a_1x_1+a_2x_2=b \tag{6.6.26}$$

  by defining $x=x_1$ and $y=x_2$.

- The equation of a plane (see section Analytical Geometry page 2800) is given by:

$$a_1x_1+a_2x_2+a_3x_3=b \tag{6.6.27}$$

  by defining $x=x_1,y=x_2,z=x_3$.

- The equation of a hyperplane is very easily (if you don't see how let us know we will specify it) generalizable from the demonstration of that of the plane and we then get:

$$a_1x_1+a_2x_2+\ldots+a_nx_n=b \tag{6.6.28}$$

  by defining $x=x_1,y=x_2,z=x_3,\ldots$

We often write a linear system in the following condensed form:

$$\boxed{\sum_{j=1}^{n} a_{ij}x_j = b_i \quad i = 1, 2, 3, \ldots, m} \tag{6.6.29}$$

We name "system solution" or "vector system solution" any t-uple $(x_1^0, x_2^0, \ldots, x_n^0)$ such as:

$$\sum_{j=1}^{n} a_{ij}x_j^0 = b_i \quad i = 1, 2, 3, \ldots, m \tag{6.6.30}$$

Solve a system means finding all the solutions of this system (we find many such systems in economic, operational research or design of experiments). Two system with n unknowns are named "equivalent systems" if all of the solution of one system is the solution of the other, i.e., if they have the same set of solutions. We sometimes say that the equations of a system are "compatible equations" or "incompatible equations", depending on whether the system has at least one solution or not admit any.

We can also give for sure a geometric interpretation to these systems. Suppose that the first members of the equations of the system are not zero. So we know that each of these equations represents a hyperplane of an affine space (see section Vector Calculus page 1815) of dimension n. Therefore, all solutions of the system, considered as set of n-tuples of coordinates is a finite intersection of hyperplanes.

✍**Example:**

Let us consider the following system of linear equations:

$$\left\{\begin{array}{rl} x_1 \qquad\quad - 3x_3 &= -3 \\ 2x_1 - 5x_2 - \;\; x_3 &= -2 \\ x_1 + 2x_2 - 5x_3 &= 1 \end{array}\right. \tag{6.6.31}$$

denoted conventionally in high-school classes in the form:

$$\left\{\begin{array}{rl} x \qquad\quad - 3z &= -3 \\ 2x - 5y - \;\; z &= -2 \\ x + 2y - 5z &= 1 \end{array}\right. \tag{6.6.32}$$

This system has for solutions the points representing the intersection of the three planes defined by the three equations. But as we can see it visually in Maple™ 4.00b using the following commands:

```
>with(plots):
>implicitplot3d(x-3*z=-3,2*x-5*y-z=-2,x+2*y-5*z=1,x=-3..3,
y=-3..3,z=-3..3);
```

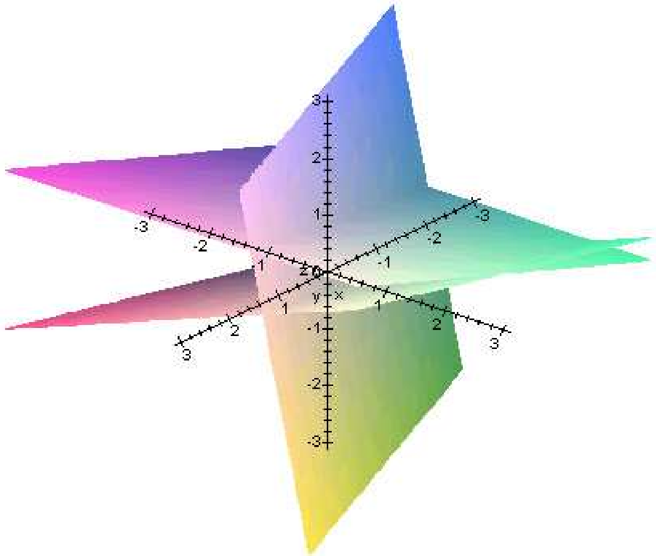


Figure 6.140: Graphical representation of the linear system with Maple™ 4.00b

There is no solutions. This can easily be checked by hand or in Maple™ 4.00b by writing:

```
>solve(x-3*z=-3,2*x-5*y-z=-2,x+2*y-5*z=1,x,y,z);
```

Remark

For "classic" resolution methods of such systems, we refer the readers to the section on Numerical Methods of the chapter on Computing Science.

Finally, note an important case in practice: "overdetermined systems" where we have more equations than unknowns.

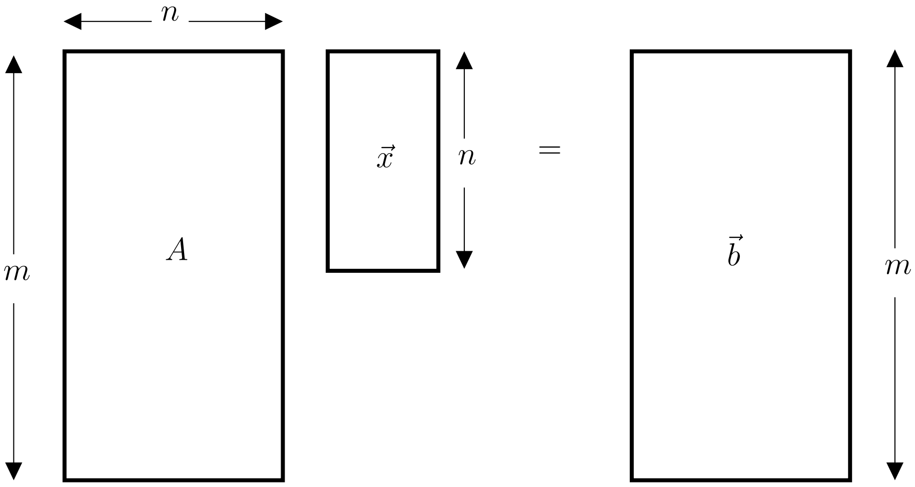


Figure 6.141: Overdetermined system

The first situation of this type dates from the 118th century (holocene calendar) through the study of the lunar oscillations but we also find this frequently in R&D laboratory in the context of design of experiments (see section Industrial Engineering page 8780) or in structural equation models (see section Numerical Methods).

An overdetermined system may be written as:

$$X\vec{b}\approx\vec{y} \tag{6.6.33}$$

where X is $n\times m$ and $\text{rank}(X\,|\,\vec{y})>m$; that is, the system is not consistent. We have changed the notation slightly from the consistent system $Ax=b$ that we have been using because now we have in mind statistical applications, and in those the notation $y\approx X\beta$ is more common. The problem is to determine a value of b that makes the approximation close in some sense. In applications of linear systems, we refer to this as "fitting" the system, which is referred to as a "model".

✍**Example:**

Consider the special case but very telling following example of three equations with two unknowns:

$$\begin{cases} 2x_1+x_2=-1 \\ -3x_1+x_2=-2 \\ -x_1+x_2=1 \end{cases} \tag{6.6.34}$$

System that will be written in the form of matrix and vector as following:

$$\begin{bmatrix} 2 & 1 \\ -3 & 1 \\ -1 & 1 \end{bmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}=\begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix} \tag{6.6.35}$$

Or said in an "augmented form" as follows:

$$\left[\begin{array}{cc|c} 2 & 1 & -1 \\ -3 & 1 & -1 \\ -1 & 0 & 1 \end{array}\right] \tag{6.6.36}$$

and in Maple™ 4.00b:

```
>with(plots):
>implicitplot(2*x+3*y=-1,-3*x+y=-2,-x+y=1,x=-3..3,y=-3..3);
```

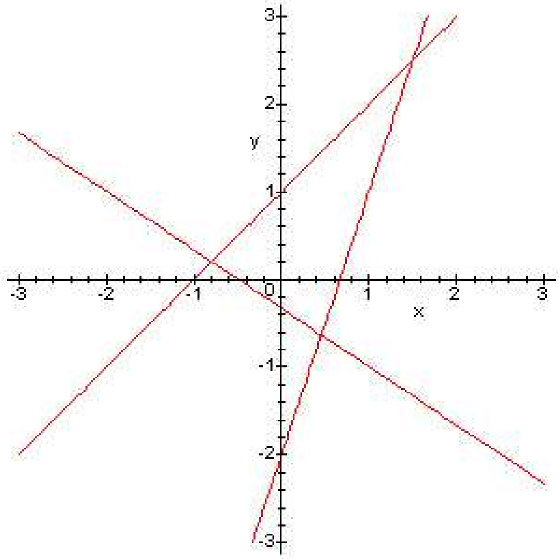


Figure 6.142: **Representation of a system of 3 equations with two unknowns in Maple™ 4.00b**

We can see on the chart above that this particular system has no complete solution but thus has a solution if we take only the problem by pair of equations... which does not necessarily help in facts...

Notice, that we have just seen that the system could be written as following:

$$\begin{bmatrix} 2 & 1 \\ -3 & 1 \\ -1 & 1 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix} \tag{6.6.37}$$

This looks like a multiple linear regression system (see section of Numerical Methods page 6304) whose column vector of unknowns can be viewed as the coefficient vector $\vec{\beta}$ of the regression line such that:

$$\begin{bmatrix} 2 & 1 \\ -3 & 1 \\ -1 & 1 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix} \Leftrightarrow X\vec{\beta}=\vec{y} \tag{6.6.38}$$

We then proved in detail in the section of Numerical Methods that:

$$\boxed{\hat{\vec{\beta}}=\left(X^TX\right)^{-1}X^T\vec{y}} \tag{6.6.39}$$

at the condition that the square matrix X^TX is, as we will see further below, invertible (non-singular). Since the works of Gauss on the subject it is considered that this approach (hence implicitly based on ordinary least squares regression[20]) remains even today the best approach to solve such systems!

If the invertible property is satisfied, we find a "pseudo-solution" (this is the official terminology...) by making calculations quickly by hand (or with a spreadsheet software like Microsoft®

---

[20] Such that we found $\vec{\beta}$ that minimize $\|X\vec{\beta}-\vec{y}\|^2$

Excel®):

$$\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix}=\begin{pmatrix}x_1\\ x_2\end{pmatrix}=\begin{pmatrix}-0.052\\ -0.368\end{pmatrix} \tag{6.6.40}$$

and injecting these values in the initial overdetermined system, the reader will quickly understand why we talk about "pseudo-solution"...

Remark

A reader asked us why we don't use for the example above just the relation $\vec{\beta}=\vec{y}X^{-1}$? The answer is simple! Because X is not a square matrix or in other words... it is in our example over-determined this is why we can only found a "pseudo-solution".

This was the pragmatic way of looking at it ... now let us turn to the second slightly more mathematical way ... (but still relatively simple):

6.6.2 Linear Transformations

Definition (#6.6.252): A "linear transformation" or "linear application" A is a mapping of a vector space E to a vector space F such that with K being in $\mathbb{R}$ or in $\mathbb{C}$:

$$\begin{aligned}\forall\vec{x}_1,\vec{x}_2\in E: &\quad A(\vec{x}_1+\vec{x}_2)=A\vec{x}_1+A\vec{x}_2\\ \forall\vec{x}\in E, a\in K: &\quad Aa\vec{x}=aA\vec{x}\end{aligned} \tag{6.6.41}$$

this constitutes, for recall, an endomorphism (see section Set Theory page 443).

The first property specifies that the transformation of a sum of vectors must be equal to the sum of the transformed vectors, so that it is linear. The second property specifies that the transformation of a vector to which we applied a scale factor (scaling) must also be equal to the same factor applied on the transformation of the original vector. If either of these two properties is not met, the transformation is therefore not linear.

We now show that any linear transformation can be represented by a matrix!

Given the basis vectors $(\vec{v}_1,\vec{v}_2,\ldots,\vec{v}_n)$ of E and $(\vec{u}_1,\vec{u}_2,\ldots,\vec{u}_n)$ for those of F with $m\geq n$. With these bases, we can represent any vectors $\vec{x}\in E,\vec{y}\in F$ with the following linear combinations (see section Vector Calculus page 1806):

$$\vec{x}=\sum_{j=1}^{n}x_j\vec{v}_j \qquad \vec{y}=\sum_{i=1}^{m}y_i\vec{u}_i \tag{6.6.42}$$

Consider the linear transformation A which applies E to F ($A:E\mapsto F$). So:

$$A(\vec{x})=\vec{y} \tag{6.6.43}$$

that we can rewrite as follows:

$$A\left(\sum_{j=1}^{n}x_j\vec{v}_j\right)=\sum_{i=1}^{m}y_i\vec{u}_i \tag{6.6.44}$$

But since A is a linear operator by definition, we can also write:

$$\sum_{j=1}^{n} x_j A(\vec{v}_j)=\sum_{i=1}^{m} y_i\vec{u}_i \qquad (6.6.45)$$

Considering now that the vectors $A(\vec{v}_j)$ are elements of F, we can rewrite them as a linear combination of its basis vectors:

$$A(\vec{v}_j)=\sum_{i=1}^{m} a_{ij}\vec{u}_i \qquad (6.6.46)$$

Therefore, we get:

$$\sum_{j=1}^{n} x_j A(\vec{v}_j)=\sum_{j=1}^{n} x_j\sum_{i=1}^{m} a_{ij}\vec{u}_i=\sum_{i=1}^{m} y_i\vec{u}_i \qquad (6.6.47)$$

By reversing the order of summations, we can write:

$$\sum_{i=1}^{m} \vec{u}_i\sum_{j=1}^{n} a_{ij}x_j=\sum_{i=1}^{m} y_i\vec{u}_i \qquad (6.6.48)$$

and rearranging the latter relation, we produce the result:

$$\sum_{i=1}^{m} \vec{u}_i\left(\sum_{j=1}^{n} a_{ij}x_j-y_i\right)=\vec{0} \qquad (6.6.49)$$

Finally, remembering that the basis vectors $\vec{u}_i$ must be independent, we can conclude that their coefficients must necessarily be zero, so:

$$\sum_{j=1}^{n} a_{ij}x_j=y_i \qquad (6.6.50)$$

Which corresponds to the "matrix product":

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad (6.6.51)$$

That we can write:

$$A\vec{x}=\vec{y} \qquad (6.6.52)$$

In other words, any linear transformation can be described by a matrix A that is multiplied with the vector that we want to transform, to get the vector resulting from the transformation.

6.6.3 Matrices

So we name a "matrix" with m rows and n columns, or "type $m \times n$ matrix" (the first term always correspond to the rows and the second to columns to the second, to remember this there is a good mnemonic trick: President Lincoln - abbreviation of Lin(e) and Col(um)n ...), any array of numbers in a ring $\mathbb{K}$ (that is most of times $\mathbb{R}$):

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & a_{ij} & \cdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mj} & \cdots & a_{mn} \end{bmatrix} \tag{6.6.53}$$

For example:

$$\begin{array}{cc} & \text{\# columns} \\ \text{\# rows} & \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \end{pmatrix} \end{array} \tag{6.6.54}$$

is a 5×5 matrix.

We often denote a matrix of type $m \times n$ briefly by:

$$(a_{ij})_{\substack{1\leq i\leq m \\ 1\leq j\leq n}} \in \mathbb{K} \tag{6.6.55}$$

or simply (a_{ij}). In a more formal way:

$$\mathbb{K}^{m\times n} = \left\{ \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & a_{ij} & \cdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mj} & \cdots & a_{mn} \end{bmatrix} \middle| a_{ij} \in \mathbb{K} \forall i,j \right\} \tag{6.6.56}$$

The number a_{ij} is named "term or coefficient of index i,j". The index i is named the "line index" and the index j the "column index".

We denote by $M_{mn}(\mathbb{K})$ all matrices $m \times n$ whose coefficients take values in $\mathbb{K}$ (typically $\mathbb{R}$ or $\mathbb{C}$ for example).

When $m = n$, we say that (a_{ij}) is a "square matrix of order n":

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & a_{ij} & \cdots & \vdots \\ a_{m1} & a_{n2} & \cdots & a_{nj} & \cdots & a_{nn} \end{bmatrix} \tag{6.6.57}$$

In this case, the terms $a_{11}, a_{22}, \ldots, a_{nn}$ are named "diagonal terms" denoted by:

$$\text{diag}(a_1, \ldots, a_n) \tag{6.6.58}$$

We will assign to the matrices special symbols, i.e. uppercase Latin letters: $A, B, \ldots$ for matrices and for columns-matrices symbols that will be vectorial lowercase letters $\vec{a}, \vec{b}, \ldots$.

We also name a matrix with a single row a "line-matrix" and a matrix with a single column a "column-matrix". It is clear that a matrix column is nothing but a "column vector" or simply a "vector" (see section Vector Calculus page 1797). Thereafter, the rows of a matrix will be treated as lines-matrices and the columns to columns-matrices.

More formally, let $M \in \mathbb{R}^{m\times n}$ be a matrix, then a "submatrix" of M is a matrix which is made of some rectangle of elements in M. Rows and columns are submatrices! In particular:

- An $m \times 1$ submatrix is named a "column vector" of M. The j-th column vector is denoted $\text{col}_j(M)$ or $\vec{M}_j$ and $\vec{M}_{ji} = M_{ij}$ for $1 \leq j \leq n$. In other words:

$$\text{col}_j(M)=\vec{M}_j=\begin{pmatrix} a_{1j}\\ a_{2j}\\ \vdots\\ a_{mj}\end{pmatrix}\Rightarrow M=\begin{bmatrix} a_{11} & a_{21} & \cdots & a_{1n}\\ a_{21} & a_{22} & \cdots & a_{2n}\\ \vdots & \ldots & \ddots & \vdots\\ a_{m1} & a_{m1} & \cdots & a_{mn}\end{bmatrix}=[\text{col}_1(A)|\text{col}_2(A)|\ldots|\text{col}_n(A)] \tag{6.6.59}$$

- An $1 \times n$ submatrix of M is named a "row vector" of M. The i-th row vector is denoted $\text{row}_i(M)$ and $(\text{row}_i(M))_j = M_{ij}$ for $1 \leq i \leq m$ in other words:

$$\text{row}_i(M)=[a_{i1}\; a_{i2}\;\cdots\; a_{in}]\Rightarrow M=\begin{bmatrix} a_{11} & a_{21} & \cdots & a_{1n}\\ a_{21} & a_{22} & \cdots & a_{2n}\\ \vdots & \ldots & \ddots & \vdots\\ a_{m1} & a_{m1} & \cdots & a_{mn}\end{bmatrix}=\begin{pmatrix}\text{row}_1(M)\\ \text{row}_2(M)\\ \ldots\\ \text{row}_m(M)\end{pmatrix} \tag{6.6.60}$$

We name "zero matrix", and we denote by 0_{mn} or simply $\mathbf{0}$ or $\mathbb{O}$ (this last notation is preferred in this textbook!) any matrix in which each term is zero:

$$0_{mn}=\mathbf{0}=\mathbb{O}=\begin{bmatrix} 0 & 0 & 0 & \cdots & 0\\ 0 & 0 & 0 & \cdots & 0\\ 0 & 0 & 0 & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \cdots & 0\end{bmatrix} \tag{6.6.61}$$

Null columns-matrices are designated by the symbol vector: $\vec{0}$.

We name "identity matrix of order n" or "unit matrix of order n", and denote it by I_n, or simply $\mathbb{1}$, the following square matrix of order n:

$$I_n=\mathbb{1}=\begin{bmatrix} 1 & 0 & 0 & \cdots & 0\\ 0 & 1 & 0 & \cdots & 0\\ 0 & 0 & 1 & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \cdots & 1\end{bmatrix} \tag{6.6.62}$$

It can also be written using the Kronecker delta notation δ_{ij} (see section Tensors Calculus page 2044).

Caution! When we work with matrices having complex coefficients we must always use the term "identity matrix" rather than "unitary matrix" because in the field of complex numbers the unitary matrix is another mathematical object that should not be confused!

We will see later that the zero matrix acts as a neutral element of the matrix addition and the unit matrix as neutral element for the matrix multiplication.

The ij-th "standard basis matrix" for $\mathbb{R}^{m\times n}$ is denoted E_{ij} for $1\leq i\leq m$ and $1\leq j\leq n$. The matrix E_{ij} is zero in all entries except for the (i,j)-th slot where it has a 1. In other words, we define $(E_{ij})_{kl}=\delta_{ik}\delta_{jl}$. Hence every matrix in $\mathbb{R}^{m\times n}$ is a linear combination of the E_{ij}.

Indeed, let $M\in\mathbb{R}^{m\times n}$ then:

$$M=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n}\\ a_{21} & a_{22} & \cdots & a_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ a_{m1} & a_{m2} & \cdots & a_{mn}\end{bmatrix}=a_{11}\begin{bmatrix} 1 & 0 & \cdots & 0\\ 0 & 0 & \cdots & 0\\ \vdots & \vdots & \ddots & 0\\ 0 & 0 & \cdots & 0\end{bmatrix}+a_{12}\begin{bmatrix} 0 & 1 & \cdots & 0\\ 0 & 0 & \cdots & 0\\ \vdots & \vdots & \ddots & 0\\ 0 & 0 & \cdots & 0\end{bmatrix}+\ldots+a_{mn}\begin{bmatrix} 0 & 0 & \cdots & 0\\ 0 & 0 & \cdots & 0\\ \vdots & \vdots & \ddots & 0\\ 0 & 0 & \cdots & 1\end{bmatrix}$$
$$=a_{11}E_{11}+a_{12}E_{12}+\ldots+a_{mn}E_{mn}=\sum_{i=1}^{m}\sum_{j=1}^{n}a_{ij}E_{ij} \tag{6.6.63}$$

The purpose of the concept of matrix will appear throughout the texts that will follow but the immediate reason for this notion is simply to allow some finite families of numbers to be designed as a rectangular array and to generalize physics theorems to multidimensional spaces.

6.6.3.1 Rank of a matrix

We will now briefly review the definition of "rank of a finite family" that we saw in the section of Vector Calculus at page 1813.

As a reminder we name "rank" of a free family of vectors the dimension (positive integer number) of the vector subspace S of E that it generates.

Definition (#6.6.253): Given $(\vec{a}_1,\vec{a}_2,\ldots,\vec{a}_n)$ the columns of a matrix A, we name "rank of A", and denote by $\text{rk}(A)$, the rank of the family of vectors $(\vec{a}_1,\vec{a}_2,\ldots,\vec{a}_n)$. More precisely, in linear algebra, the rank of a matrix A is the dimension of the vector space generated (or spanned) by its columns. This is the same as the dimension of the space spanned by its rows as we will see it later. It is a measure of the "nondegenerateness" of the system of linear equations and linear transformation encoded by A (see earlier above!). There are multiple equivalent definitions of rank. A matrix's rank is one of its most fundamental characteristics!

In a slightly more familiar language (...) the rank of a matrix is given by the number of columns-matrices that can't be expressed by the combination and scalar multiplication of other columns-matrices of the same matrix!!

Given this definition it is almost obvious that the rank of a matrix is zero if and only if the matrix is the zero matrix $\mathbb{O}$.

Before we enter in more formal calculations let us see already some introducing examples:

✍**Examples:**

E1. The matrix:

$$\begin{bmatrix} 1 & 2 & 1 \\ -2 & -3 & 1 \\ 3 & 5 & 0 \end{bmatrix} \tag{6.6.64}$$

has rank 2: the first two rows are linearly independent, so the rank is at least 2, but all three rows are linearly dependent (the first is equal to the sum of the second and third) so the rank must be less than 3.

E2. The matrix:

$$A = \begin{bmatrix} 1 & 1 & 0 & 2 \\ -1 & -1 & 0 & -2 \end{bmatrix} \tag{6.6.65}$$

has rank 1: there are non-zero columns, so the rank is positive, but any pair of columns is linearly dependent. Similarly, the transpose:

$$A^T = \begin{bmatrix} 1 & -1 \\ 1 & -1 \\ 0 & 0 \\ 2 & -2 \end{bmatrix} \tag{6.6.66}$$

of A has rank 1. Indeed, since the column vectors of A are the row vectors of the transpose of A, the statement that the column rank of a matrix equals its row rank is equivalent to the statement that the rank of a matrix is equal to the rank of its transpose, i.e.:

$$\boxed{\text{rk}(A) = \text{rk}(A^T)} \tag{6.6.67}$$

The above examples show the statement that the column rank of a matrix equals its row rank is equivalent to the statement that the rank of a matrix is equal to the rank of its transpose.

To summarize:

- Rank-0 Matrices. There is only one rank-zero matrix of a given size, namely the all-zero matrix.
- Rank-1 Matrices. A rank-one matrix is precisely a non-zero matrix of the type assumed in the example above - all rows are multiples of each other.

  An equivalent definition of a rank-1 $m \times n$ matrix is as the outer product $\vec{u}\vec{v}^T$ of an m-

vector $\vec{u}$ and an n-vector $\vec{v}$:

$$A=\vec{u}\vec{v}^T=\begin{bmatrix} - & u_1\vec{v}^T & - \\ - & u_2\vec{v}^T & - \\ & \vdots & \\ - & u_m\vec{v}^T & - \end{bmatrix}=\begin{bmatrix} | & | & & | \\ v_1\vec{u} & v_2\vec{u} & \cdots & v_n\vec{u} \\ | & | & & | \end{bmatrix} \tag{6.6.68}$$

Note that each row is a multiple of $\vec{v}^T$, and each column is multiple of $\vec{u}$.

- Rank-2 Matrices. A rank-two matrix is just a superposition (i.e., sum) of two rank-1 matrices:

$$A=\vec{u}\vec{v}^T+\vec{w}\vec{z}^\top=\begin{bmatrix} - & u_1\vec{v}^T+w_1\vec{z}^T & - \\ - & u_2\vec{v}^T+w_2\vec{z}^T & - \\ & \vdots & \\ - & u_m\vec{v}^T+w_m\vec{z}^T & - \end{bmatrix}=\begin{bmatrix} | & | \\ \vec{u} & \vec{w} \\ | & | \end{bmatrix}\cdot\begin{bmatrix} - & \vec{v}^T & - \\ - & \vec{z}^T & - \end{bmatrix} \tag{6.6.69}$$

- Rank-k Matrices. The general definition of matrix rank should now be clear: a matrix A has rank k if it can be written as the sum of k rank-one matrices, and cannot be written as the sum of $k-1$ or fewer rank-one matrices.

  Rephrased in terms of matrix multiplication, an equivalent definition is that A can be written as, or "factored into," the product of a long and skinny $(m\times k)$ matrix Y and a short and long $(k\times n)$ matrix Z^T (and that A cannot be likewise factored into the product of $m\times(k-1)$ and $(k-1)\times n$ matrices).

  There are many equivalent definitions of the rank of a matrix A. The following two conditions are equivalent to each other and to the definition above (any one of the three conditions implies the other two):

  1. The largest linearly independent subset of columns of A has size k. That is, all n columns of A arise as linear combinations of only k of them.
  2. The largest linearly independent subset of rows of A has size k. That is, all m rows of A arise as linear combinations of only k of them.

  It is clear that the first condition implies the second and third: if $A=YZ^T$, then all of A's columns are linear combinations of the k columns of Y, and all of A's rows are linear combinations of the k rows of Z^T.

Before continuing we would like to indicate to the reader that later we will prove that if the rows of a matrix are independent, its determinant (computable only for $n\times n$ square matrices) is non zero $\det(A)\neq 0$ and therefore $\text{rk}(A)=n$ and obviously $\text{rk}(A)=1$ when $\det(A)=0$.

Definition (#6.6.254): A matrix is said to have "full rank" if its rank equals the largest possible for a matrix of the same dimensions, which is the lesser of the number of rows and columns (i.e. $\text{rk}(A)=\min(m,n)$). A matrix is said to be "rank deficient" if it does not have full rank.

Remark

If there is difficulty in determining the rank of a matrix there is a technique named "matrix scaling" that we will see just below that can do this work very quickly.

Definition (#6.6.255): We name "system associated matrix":

$$\begin{aligned} a_{11}x_1+a_{12}x_2+\ldots+a_{1n}x_n&=b_1\\ a_{21}x_1+a_{22}x_2+\ldots+a_{2n}x_n&=b_2\\ &\ldots\\ a_{m1}x_1+a_{m2}x_2+\ldots+a_{mn}x_n&=b_m \end{aligned} \tag{6.6.70}$$

the mathematical object defined by:

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n}\\ a_{21} & a_{22} & \cdots & a_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \tag{6.6.71}$$

that is to say, the matrix A in which the terms are the coefficients of the system. We name "matrix of the second member of the linear system", or simply "second member of the system", the matrix-column $\vec{b}=(b_i)$ whose terms are the coefficients of the second member of this system. We also name "augmented matrix associated to a system" the matrix A obtained by adding $\vec{b}=(b_i)$ as the $(n+1)$-th column.

If we now consider an associated system matrix A and of second member $\vec{b}$. Let us always denote the columns of A by $(\vec{a}_1,\vec{a}_2,\ldots,\vec{a}_n)$. The system can then be written equivalently as a linear vector equation:

$$x_1\vec{a}_1+x_2\vec{a}_2+\ldots+x_n\vec{a}_n=\vec{b} \tag{6.6.72}$$

Now remember a theorem that we saw and proved in the section of Vector Calculus: for the rank of a family of vectors $(\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_n)$ to be the equal to rank of augmented family $(\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_n,\vec{y})$, it is necessary and sufficient that the vector $\vec{y}$ is a linear combination of the vectors $\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_n)$.

It follows that our linear system in a vector form has at least one solution $(\vec{x}_1^0,\vec{x}_2^0,\ldots,\vec{x}_n^0)$ if the rank of the family $(\vec{a}_1,\vec{a}_2,\ldots,\vec{a}_n)$ is equal to the rank of augmented family $(\vec{a}_1,\vec{a}_2,\ldots,\vec{a}_n,\vec{b})$ and this solution is unique if and only if the rank of the family $(\vec{a}_1,\vec{a}_2,\ldots,\vec{a}_n)$ is equal to n.

Thus, for a linear system matrix of associated matrix A and of second member $\vec{b}$ admits at least one solution, it is necessary and sufficient that the rank of A is equal to the rank of the augmented matrix $(A|\vec{b})$. If this condition is met, the system admits a unique solution if and only if the rank of A is equal to the number of unknowns in other words: if the columns of A are linearly independent!!!

We say that a matrix is "staggered" if its rows (lines) meet the following two conditions:

C1. Any null line is followed by lines full of zeros

C2. The leading coefficient (the first non-zero number from the left, also named the "pivot") of a non-zero row is always strictly to the right of the leading coefficient of the row above it

These two conditions imply that all entries in a column below a leading coefficient are zeros.

A non-zero row echelon matrix is therefore of the form (by adding and subtracting rows between them):

$$\begin{bmatrix} \left|\underline{a_{1j_1}}\right. & \cdots & \cdots & a_{1n} \\ 0 & \left|\underline{a_{2j_2}}\right. & \cdots & a_{2n} \\ 0 & 0 & \left|\underline{a_{1j_r}}\right. & a_{rn} \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{6.6.73}$$

where $j_1 < j_2 < \ldots < j_r$ and $a_{1j_1}, a_{2j_2}, \ldots, a_{rj_r}$ are non-zero terms. The terminal zero lines may be missing.

The columns of index $j_1, j_2, \ldots, j_r$ of an echelon matrix are clearly linearly independent. Considered as vectors-columns of $\mathbb{R}^2$, so they form a basis of this vector space. Considering the other columns also as vectors-columns of $\mathbb{R}^n$, we deduce that they are necessarily linear combination of those of index $j_1, j_2, \ldots, j_r$ and therefore that the rank of the echelon matrix A is $\text{rk}(A) = r$.

We will note that r is also the number of non-zero lines of the echelon matrix and also the rank of the lines of this matrix, since the non-zero lines are therefore clearly independent (we proved in the section Vector Calculus that the rank of lines and columns has same value with the same properties of independence).

We can therefore allow ourselves to do a given number of elementary (extra) operations on the lines of matrices that we will be very useful, without changing its rank:

P1. We can swap the lines.

Remark

As we know the matrix can be seen just an aesthetic graphic representation of a linear system. So swap two lines does not change the system.

P2. Multiply a row by a non-zero scalar.

Remark

This obviously not altering the linear independence of the vectors lines.

P3. Adding to an original line, a multiple of another line.

Remark

The original line will disappear in favour of the new that is independent of all (former) others. The system thus remains linearly independent.

Any matrix can be transformed into a row echelon form by a finite sequence of the previous properties here's how:

1. Pivot the matrix
    (a) Find the pivot, the first non-zero entry in the first column of the matrix.
    (b) Interchange rows, moving the pivot row to the first row.
    (c) Multiply each element in the pivot row by the inverse of the pivot, so the pivot equals 1.
    (d) Add multiples of the pivot row to each of the lower rows, so every element in the pivot column of the lower rows equals 0.
2. To get the matrix in row echelon form, repeat the pivot.
    (a) Repeat the procedure from Step 1 above, ignoring previous pivot rows.
    (b) Continue until there are no more pivots to be processed.
3. To get the matrix in reduced row echelon form, process non-zero entries above each pivot.
    (a) Identify the last row having a pivot equal to 1, and let this be the pivot row.
    (b) Add multiples of the pivot row to each of the upper rows, until every element above the pivot equals 0.
    (c) Moving up the matrix, repeat this process for each row.

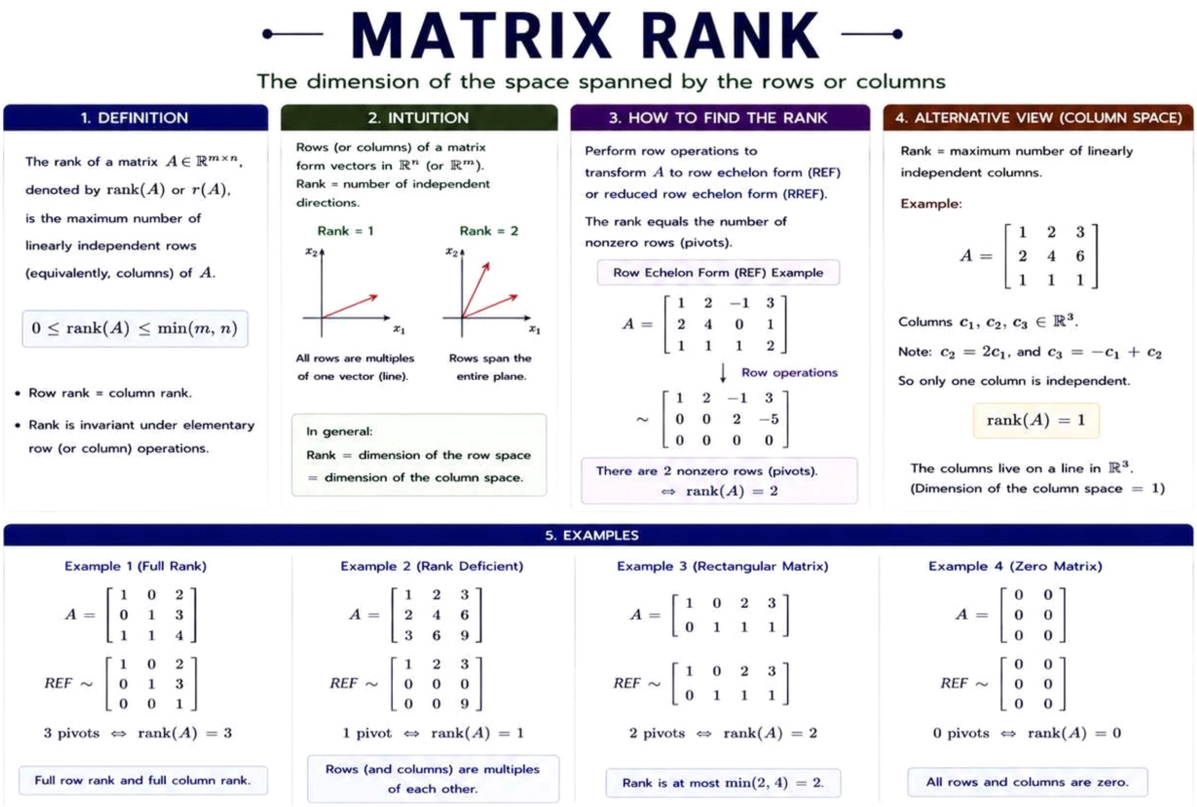
It is therefore obvious that the elementary operations on the rows of a matrix do not change the rank of the rows of the matrix. However, we know that the rank of lines of a row matrix is equal to the rank of the columns, that is to say to the rank of this matrix (once again see the section Vector Calculus for the proof). We conclude that the column rank of any matrix of the type $m\times n$ is also equal to the rank of the rows of this matrix.

As a corollary of this conclusion, it appears that:

$$\text{rk}(A)=\min(m,n) \tag{6.6.74}$$

When solving linear systems of m equation with n unknowns it appears, as we have already noted at the beginning of this section (and with practical example in the section of Numerical Methods), there must be at least an equal number of equations than unknowns or more rigorously: the number of unknowns must be less or equal to the number of equations such as:

$$\boxed{n=\text{rk}(A|\vec{b})\leq m} \tag{6.6.75}$$



MATRIX RANK

The dimension of the space spanned by the rows or columns

1. DEFINITION

The rank of a matrix $A \in \mathbb{R}^{m\times n}$, denoted by $\operatorname{rank}(A)$ or $r(A)$, is the maximum number of linearly independent rows (equivalently, columns) of A.

$$0 \le \operatorname{rank}(A) \le \min(m, n)$$

- Row rank = column rank.
- Rank is invariant under elementary row (or column) operations.

2. INTUITION

Rows (or columns) of a matrix form vectors in $\mathbb{R}^n$ (or $\mathbb{R}^m$). Rank = number of independent directions.

Rank = 1

All rows are multiples of one vector (line).

Rank = 2

Rows span the entire plane.

In general:
Rank = dimension of the row space
= dimension of the column space.

3. HOW TO FIND THE RANK

Perform row operations to transform A to row echelon form (REF) or reduced row echelon form (RREF).

The rank equals the number of nonzero rows (pivots).

Row Echelon Form (REF) Example

$$A = \begin{bmatrix} 1 & 2 & -1 & 3 \\ 2 & 4 & 0 & 1 \\ 1 & 1 & 1 & 2 \end{bmatrix}$$

$$\Big\downarrow \text{ Row operations}$$

$$\sim \begin{bmatrix} 1 & 2 & -1 & 3 \\ 0 & 0 & 2 & -5 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

There are 2 nonzero rows (pivots).
$\Leftrightarrow \operatorname{rank}(A) = 2$

4. ALTERNATIVE VIEW (COLUMN SPACE)

Rank = maximum number of linearly independent columns.

Example:

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 1 & 1 & 1 \end{bmatrix}$$

Columns $c_1, c_2, c_3 \in \mathbb{R}^3$.

Note: $c_2 = 2c_1$, and $c_3 = -c_1 + c_2$

So only one column is independent.

$$\operatorname{rank}(A) = 1$$

The columns live on a line in $\mathbb{R}^3$.
(Dimension of the column space $= 1$)

5. EXAMPLES

Example 1 (Full Rank)

$$A = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 3 \\ 1 & 1 & 4 \end{bmatrix}$$

$$REF \sim \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 3 \\ 0 & 0 & 1 \end{bmatrix}$$

$3 \text{ pivots} \Leftrightarrow \operatorname{rank}(A) = 3$

Full row rank and full column rank.

Example 2 (Rank Deficient)

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{bmatrix}$$

$$REF \sim \begin{bmatrix} 1 & 2 & 3 \\ 0 & 0 & 0 \\ 0 & 0 & 9 \end{bmatrix}$$

$1 \text{ pivot} \Leftrightarrow \operatorname{rank}(A) = 1$

Rows (and columns) are multiples of each other.

Example 3 (Rectangular Matrix)

$$A = \begin{bmatrix} 1 & 0 & 2 & 3 \\ 0 & 1 & 1 & 1 \end{bmatrix}$$

$$REF \sim \begin{bmatrix} 1 & 0 & 2 & 3 \\ 0 & 1 & 1 & 1 \end{bmatrix}$$

$2 \text{ pivots} \Leftrightarrow \operatorname{rank}(A) = 2$

Rank is at most $\min(2, 4) = 2$.

Example 4 (Zero Matrix)

$$A = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$REF \sim \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$0 \text{ pivots} \Leftrightarrow \operatorname{rank}(A) = 0$

All rows and columns are zero.

6.6.3.2 Matrix Algebra

Remember that we saw during our study of Vector Calculus that the algebraic operations of multiplication of a vector by a scalar, addition or subtraction of vectors between them and the operation of scalar product formed in the context of set theory a "vector space" (see section Set Theory page 1805) possessing therefore a "vector algebraic structure". This under the condition that the vectors of course have the same dimensions (this observation, for recall, is not valid if instead of the scalar product we take the cross product).

Just as vectors, we can multiply a matrix by a scalar and add (subtract) them together (as long as they have the same dimensions..) but in addition, we can also multiply two matrices together under certain conditions which we will define below. This will also make the set of matrices in the set-sense, a vector space on K (being most of time $\mathbb{R}$) having also therefore a "vector algebraic structure".

Thus, a vector may also be viewed as a particular matrix of dimension $m \times n$ and operate in the vector space of matrices. Basically..., vector calculus is only a special case of linear algebra!!! This is way at school people learn (after Calculus) Vector Calculus first and Linear Algebra later (and some will lean after Tensor Calculus).

D1. Given $A,B \in M_{mn}(\mathbb{R})$. We name "sum of A and B" the matrix $C \in M_{mn}(\mathbb{R})$ whose coefficients are:

$$c_{ij} = a_{ij} + b_{ij} \qquad 1 \leq i \leq m, 1 \leq j \leq n \tag{6.6.76}$$

That is to say explicitly:

$$\begin{aligned}\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mn} \end{bmatrix} &= \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mn} \end{bmatrix} \\ &= \begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} & \cdots & a_{1n}+b_{1n} \\ a_{21}+b_{21} & a_{22}+b_{22} & \cdots & a_{2n}+b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}+b_{m1} & a_{m2}+b_{m2} & \cdots & b_{mm}+b_{mn} \end{bmatrix}\end{aligned} \tag{6.6.77}$$

D2. Given $A \in M_{mn}(\mathbb{R})$ a matrix and a $\lambda \in \mathbb{R}$ a scalar (we can also take in $\mathbb{C}$ if we want). We name "product of A by λ" the matrix whose coefficients are:

$$c_{ij} = \lambda a_{ij} \qquad 1 \leq i \leq m, 1 \leq j \leq n \tag{6.6.78}$$

That is to say explicitly:

$$\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mn} \end{bmatrix} = \lambda \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} = \begin{bmatrix} \lambda a_{11} & \lambda a_{12} & \cdots & \lambda a_{1n} \\ \lambda a_{21} & \lambda a_{22} & \cdots & \lambda a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda a_{m1} & \lambda a_{m2} & \cdots & \lambda a_{mn} \end{bmatrix} \tag{6.6.79}$$

With the two previous definitions we can actually conclude that the space / set of matrices is a vector space and thus has a vector algebraic structure.

D3. Let E, F, G be three vector spaces of basis respectively $\mathcal{E}, \mathcal{F}, \mathcal{G}$ and two linear application f and g (see the section Set Theory for a refresh).

Let us denote by A the matrix of f with respect to basis $\mathcal{E}, \mathcal{F}$, and B the matrix of g with respect to basis $\mathcal{F}, \mathcal{G}$. Then the matrix C of $g \circ f$ (see the definition of a composite function in the section of Functional Analysis) relating to the basis $\mathcal{E}, \mathcal{G}$ is equal to the product of B by A denoted simply by BA.

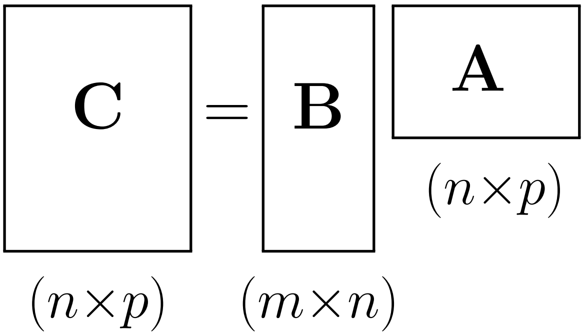
$$\begin{aligned} E \overset{f}{\mapsto} F \overset{g}{\mapsto} G \quad &\text{and} \quad E \overset{g\circ f}{\mapsto} G \\ E \overset{A}{\mapsto} F \overset{B}{\mapsto} G \quad &\text{and} \quad E \overset{BA=C}{\mapsto} G \end{aligned} \tag{6.6.80}$$

So let $B \in M_{mn} \in \mathbb{R}$ and $A \in M_{np} \in \mathbb{R}$, we name "matrix product" or "matrix multiplication" of A and B and we denote it by BA, the matrix $C \in M_{mp}\mathbb{R}$ whose components are:

$$\boxed{c_{ki} = \sum_{j=1}^{n} b_{kj} a_{ji} \qquad 1 \leq i \leq p, 1 \leq k \leq m} \tag{6.6.81}$$

It is important to notice that at the opposite to the addition, A and B may have different dimensions. However! the number of rows of A must be equal to the number of columns of B, as indicated by the index n of the two matrices. So in the product BA, if B is a matrix $m \times n$, A must be a matrix $n \times p$, for any p!!!

Schematically:



or even more explicitly (Falk's scheme):

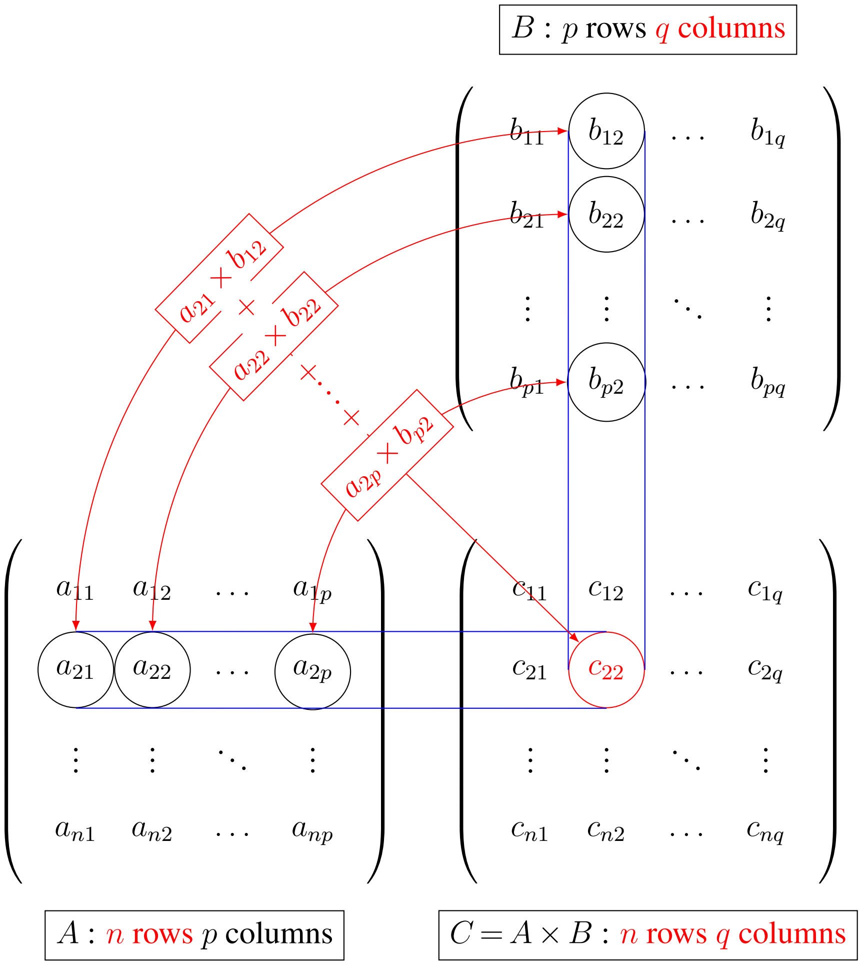

B : p rows q columns
b_{11} b_{12} … b_{1q}
b_{21} b_{22} … b_{2q}
b_{p1} b_{p2} … b_{pq}
$a_{21}\times b_{12}$
+
$a_{22}\times b_{22}$
+…+
$a_{2p}\times b_{p2}$
a_{11} a_{12} … a_{1p}
a_{21} a_{22} … a_{2p}
a_{n1} a_{n2} … a_{np}
c_{11} c_{12} … c_{1q}
c_{21} c_{22} … c_{2q}
c_{n1} c_{n2} … c_{nq}
A : n rows p columns
$C=A\times B$: n rows q columns

Figure 6.143: **Matrix product Falk's scheme (credit: Alain Matthes)**

By denoting with a capital Latin letters matrices and with lowercase Greek letters scalars, the reader can easily verify with what we have seen until now (we can add proofs on request) the following properties of matrix algebra (the matrices are assumed to have adequate dimensions):

P1. Left distributivity: $A(B+C)=AB+AC$

P2. Right distributivity: $(A+B)C=AC+BC$

P3. Scaling association: $(\lambda A)B=A(\lambda A)$

P4. Associativity: $(AB)C=A(BC)$

P5. Non-commutativity: $BA\neq AB$

P6. Absorbing-Element: $A\mathbb{O}=\mathbb{O}$

P7. Neutral Element for addition: $A+\mathbb{O}=A$

It is especially important to remember the property P5 that shows that the multiplication is obviously not commutative (for dimensions greater than 1 of course!) and also the property P4 such that the matrix multiplication is associative.

Concerning the general proof that the assertion of commutativity if false we must pass through a numerical example (because even the general case without replacing algebraic terms by numerical values will not show you much in our point of view...).

Remark

The set of square matrices $M_{nn} \in \mathbb{R}$ of order n with components in $\mathbb{R}$ provided with the sum and the usual matrix multiplication forms a ring (see section Set Theory page 426). This is true more generally if the coefficients of the matrices are taken in any ring: for example, all the matrices $M_{nn} \in \mathbb{Z}$ with integer components is a ring.

One reader asked us to prove the property of associative. So let us begin!

Let $A \in M_{kl}(\mathbb{C}), B \in M_{lm}(\mathbb{C}), C \in M_{mn}(\mathbb{C})$, then for $1 \leq i \leq k, 1 \leq j \leq n$, we have well (we use the explicit expression of matrix components multiplication seen above many times as you can see):

$$\begin{aligned}(A(BC))_{ij} &= \sum_{p=1}^{l} A_{ip}(BC)_{pj} = \sum_{p=1}^{l} A_{ip}\left(\sum_{q=1}^{m} B_{pq}C_{qi}\right) = \sum_{p=1}^{l}\sum_{q=1}^{m} A_{ip}B_{pq}C_{qi}\\ &= \sum_{q=1}^{m}\left(\sum_{p=1}^{l} A_{ip}B_{pq}\right)C_{qj} = \sum_{q=1}^{m}(AB)_{iq}C_{qj} = \sum_{q=1}^{m}(AB)_{iq}C_{qj} = ((AB)C)_{ij}\end{aligned} \tag{6.6.82}$$

6.6.3.3 Type of Matrices

To simplify the notations and length of calculations we introduce now the most common types of matrices that the reader will encounter throughout his reading of this textbook (and not just in the chapters on pure mathematics!).

Some definitions will only be recalls!

We denote by $M_{mn}(\mathbb{K})$ all matrices $m \times n$ whose coefficients take values in K (typically $\mathbb{R}$ or $\mathbb{C}$ for example).

Definitions (#6.6.256):

D1. When $m = n$, we say that $M_{nn}(\mathbb{K}) = (a_{ij})$ is a "square matrix of order n":

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & a_{ij} & \cdots & \vdots \\ a_{m1} & a_{n2} & \cdots & a_{nj} & \cdots & a_{nn} \end{bmatrix} \tag{6.6.83}$$

D2. We name "zero matrix", and we denote by 0_{mn} or simply $\mathbf{0}$ or $\mathbb{O}$ (this last notation is preferred in this textbook!) any matrix in which each term is zero:

$$0_{mn}=\mathbf{0}=\mathbb{O}=\begin{bmatrix}0&0&0&\cdots&0\\0&0&0&\cdots&0\\0&0&0&\cdots&0\\\vdots&\vdots&\vdots&\ddots&\vdots\\0&0&0&\cdots&0\end{bmatrix} \tag{6.6.84}$$

Null columns-matrices are designated by the symbol vector: $\vec{0}$.

D3. We name "identity matrix of order n" or "unit matrix of order n", and denote it by I, or simply $\mathbb{1}$, the square matrix of order n:

$$I_n=\mathbb{1}=\begin{bmatrix}1&0&0&\cdots&0\\0&1&0&\cdots&0\\0&0&1&\cdots&0\\\vdots&\vdots&\vdots&\ddots&\vdots\\0&0&0&\cdots&1\end{bmatrix} \tag{6.6.85}$$

It can also be written using the Kronecker delta notation δ_{ij} (see section Tensors Calculus page 2044).

Remark

Even if we use it actually for only two practical applications in this textbook (Principal Component Analysis and Multidimensional Scaling), and that we don't found it often in other textbooks, the latter definition must not be confuse we the matrix named "matrix of ones", and defined for a square $n\times n$ matrix by:

$$\mathbb{J}=\begin{bmatrix}1&1&1&\cdots&1\\1&1&1&\cdots&1\\1&1&1&\cdots&1\\\vdots&\vdots&\vdots&\ddots&\vdots\\1&1&1&\cdots&1\end{bmatrix} \tag{6.6.86}$$

D4. We name "diagonal matrix" any square matrix $A\in M_{nn}\in\mathbb{C}$ which only the diagonal has non-null elements:

$$A=\begin{bmatrix}a_{11}&0&0&\cdots&0\\0&a_{22}&0&\cdots&0\\0&0&a_{33}&\cdots&0\\\vdots&\vdots&\vdots&\ddots&\vdots\\0&0&0&\cdots&a_{nn}\end{bmatrix} \tag{6.6.87}$$

Formally:

$$\forall i,j\in\{1,n\}:\quad i\neq j\Rightarrow a_{ij}=0 \tag{6.6.88}$$

The usual notation of a diagonal matrix is:

$$D=\text{diag}\,\{a_{11},a_{22},\ldots,a_{nn}\}:=\text{diag}\,\{d_1,d_2,\ldots,d_n\} \tag{6.6.89}$$

D5. We name "lower triangular matrix" a square matrix if all the entries above the main diagonal are zero:

$$\begin{bmatrix} a_{11} & 0 & 0 & 0 & 0 & 0 \\ a_{21} & a_{22} & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & a_{ij} & 0 & 0 \\ a_{m1} & a_{n2} & \cdots & a_{nj} & \cdots & a_{nn} \end{bmatrix} \tag{6.6.90}$$

Formally:

$$\forall i,j \in \{1,n\}: \quad i<j \Rightarrow a_{ij}=0 \tag{6.6.91}$$

Similarly, a square matrix is named "upper triangular matrix" if all the entries below the main diagonal are zero:

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2j} & \cdots & a_{2n} \\ 0 & 0 & 0 & a_{ij} & \cdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & a_{nn} \end{bmatrix} \tag{6.6.92}$$

Formally:

$$\forall i,j \in \{1,n\}: \quad j<i \Rightarrow a_{ij}=0 \tag{6.6.93}$$

A "triangular matrix" is one that is either lower triangular or upper triangular. A matrix that is both upper and lower triangular is named a "diagonal matrix".

If an upper triangular matrix was obtained by a "staggered" matrix we will write it as following:

$$\begin{bmatrix} \left|\underline{a_{1j_1}}\right. & \cdots & \cdots & a_{1n} \\ 0 & \left|\underline{a_{2j_2}}\right. & \cdots & a_{2n} \\ 0 & 0 & \left|\underline{a_{1j_r}}\right. & a_{rn} \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{6.6.94}$$

D6. A "Toeplitz matrix" (named after Otto Toeplitz) or "diagonal-constant matrix", is a matrix in which each descending diagonal from left to right is constant. For instance, the following matrix is a Toeplitz matrix:

$$\begin{bmatrix} a & b & c & d & e \\ f & a & b & c & d \\ g & f & a & b & c \\ h & g & f & a & b \\ i & h & g & f & a \end{bmatrix} \tag{6.6.95}$$

Any $n \times n$ matrix A of the form:

$$A = \begin{bmatrix} a_0 & a_{-1} & a_{-2} & \cdots & \cdots & a_{-(n-1)} \\ a_1 & a_0 & a_{-1} & \ddots & & \vdots \\ a_2 & a_1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & a_{-1} & a_{-2} \\ \vdots & & \ddots & a_1 & a_0 & a_{-1} \\ a_{n-1} & \cdots & \cdots & a_2 & a_1 & a_0 \end{bmatrix} \tag{6.6.96}$$

is a Toeplitz matrix. If the i,j element of A is denoted $A_{i,j}$, then we have:

$$A_{i,j}=A_{i+1,j+1} \tag{6.6.97}$$

We will meet such a matrix in our study of time series analysis and especially when we will prove the Yule-Walker equations (see section Social Sciences page **??**).

D7. Given M_{nn} a square matrix. The matrix M_{nn} is named "invertible matrix" or "regular matrix" or "non-singular matrix" if and only if M_{nn}^{-1} is such that:

$$M_{nn}M_{nn}^{-1}=I_n=\mathbb{1} \tag{6.6.98}$$

If this is not the case, we say that M_{nn} is a "singular matrix". We will prove later that for a square matrix to invertible (non-singular) on sufficient condition is that its determinant is equal to zero.

This definition is fundamental, it has extremely important consequences in all linear algebra and also in physics (solving linear systems, determining, eigenvectors and eigenvalues, etc.), statistics and finance and it is appropriate to remember it.

Let us prove a useful property of invertible matrices associated with the property of associativity of matrix multiplication:

$$(AB)^{-1}=B^{-1}A^{-1} \tag{6.6.99}$$

Indeed:

$$(AB)\left(B^{-1}A^{-1}\right)=\left(A\left(BB^{-1}\right)\right)A^{-1}=(A\mathbb{1})A^{-1}=AA^{-1}=\mathbb{1} \tag{6.6.100}$$

D8. Given a matrix $A_{mn}:=A$:

$$A=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n}\\ a_{21} & a_{22} & \cdots & a_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ a_{m1} & a_{m2} & \cdots & a_{mn}\end{bmatrix} \tag{6.6.101}$$

We name "transposed matrix" of $A=A_{mn}$, the matrix denoted by $A^T=A_{nm}$ (the superscript T is depending of the books and teachers uppercase or lowercase and either left or right but the standard ISO 80000-2:2009 *Mathematical signs and symbols to be used in the natural sciences and technology* recommends the capital and superscript on the top right[21]) for which we transpose the rows into columns and the into rows:

$$A^T=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m}\\ a_{21} & a_{22} & \cdots & a_{2m}\\ \vdots & \vdots & \ddots & \vdots\\ a_{n1} & a_{n2} & \cdots & a_{nm}\end{bmatrix} \tag{6.6.102}$$

Here are some interesting properties of the transpose of a matrix (which we will be useful to us later in this section for a famous theorem and also in the study of the multiple linear regression methods in the section of Numerical Methods!):

---

[21] Sadly many textbooks, especially from USA denoted the transpose with a prime: A'

P1. $(A^T)^T$
P2. $(\lambda A+B)^T=\lambda A^T+B^T, \lambda\in\mathbb{R}, \lambda\in\mathbb{R}$
P3. $(AB)^T=B^TA^T$
P4. $(A^{-1})^T=(A^T)^{-1}, \exists A^{-1}$
P5. $A\vec{x}\circ\vec{y}=\vec{x}A^T\vec{y}$

The transposed matrix is very important in physics, statistics and finance and obviously in the field of mathematics for example in the context of the theory of groups and symmetries! So it is also worth remembering its definition.

As the third property is the most used one in the various sections of this textbook let us demonstrate it by considering $A\in M_{lm}\in\mathbb{C}, B\in M_{mn}\in\mathbb{C}$:

Proof 6.37.6. Remembering the explicit relation of matrix multiplication seen earlier:

$$(B^TA^T)_{ij}=\sum_{i=1}^{m}(B^T)_{ik}(A^T)_{kj}=\sum_{k=1}^{m}B_{ki}A_{jk} \tag{6.6.103}$$

But in this last equality, we note that we browse B by line and A in column for a i and a j fixed and this we know then corresponds to the matrix multiplication AB, therefore:

$$(B^TA^T)_{ij}=\sum_{k=1}^{m}B_{ki}A_{jk}=(AB)_{ji}=((AB)^T)_{ij} \tag{6.6.104}$$

Finally we have well:

$$((AB)^T)_{ij}=(B^TA^T)_{ij} \tag{6.6.105}$$

■ Q.E.D.

And for the same reasons let us prove the before last property.

Proof 6.37.7. First, it is trivial that if A is invertible:

$$AA^{-1}=I_n \tag{6.6.106}$$

and taking the transpose on both sides of the equality we find (we use the property proved just before):

$$(A^{-1})^TA^T=I_n \tag{6.6.107}$$

The latter equality shows obviously that $(A^{-1})^T$ is the inverse of A^T, that is to say:

$$(A^{-1})^T=(A^T)^{-1} \tag{6.6.108}$$

■ Q.E.D.

Finally, a last and important property of transposed matrices, is that given a matrix A (square or not, symmetric or not), the multiplication by it's own transposed gives a symmetric matrix.

The proof is quite straightforward and very useful for the SVD (Singular Value Decomposition) theorem:

$$(AA^T)^T=(A^T)^TA^T=AA^T \tag{6.6.109}$$

D9. Given:

$$A=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \qquad (6.6.110)$$

a matrix of $M_{mn}(\mathbb{C})$. We name "adjoint matrix" of A, the matrix of $M_{mn}(\mathbb{C})$ defined by:

$$\bar{A}^T=\begin{bmatrix} \bar{a}_{11} & \bar{a}_{12} & \cdots & \bar{a}_{1n} \\ \bar{a}_{21} & \bar{a}_{22} & \cdots & \bar{a}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{a}_{m1} & \bar{a}_{m2} & \cdots & \bar{a}_{mn} \end{bmatrix} \qquad (6.6.111)$$

which is the complex conjugate of the transposed matrix or if you prefer ... the transposed matrix of the complex conjugate (in the case of real components... we obviously don't need to take the conjugate!). To simplify the notations we simply note this matrix $A^\dagger$ (notation frequently use in Quantum Physics and Set Algebra).

Remark

Trivial relation (which is often used in Quantum Field Physics) already proved just before and obviously correct when the components are in $\mathbb{C}$:

$$(AB)^\dagger=B^\dagger A^\dagger \qquad (6.6.112)$$

D10. By definition, a matrix is named "Hermitian matrix" or "self-adjoint matrix"... if it is equal to its own adjoint (conjugate transpose matrix) such that:

$$A=A^\dagger \qquad (6.6.113)$$

D11. Given A as square matrix $M_{nn}(\mathbb{R})$, the "trace" of A denoted $\text{tr}(A)$ is defined by the sum of the terms of the diagonal (very useful in some statistical techniques):

$$\boxed{\text{tr}(A)=\sum_{i=1}^{m} a_{ij}} \qquad (6.6.114)$$

Some useful related relations (we can add the detailed proof on the demand of the readers):

$$\text{tr}(\lambda A+B)=\lambda\text{tr}(A)+\text{tr}(B) \qquad (6.6.115)$$

and:

$$\text{tr}(AB)=\text{tr}(BA) \quad \forall A\in M_{mn}(\mathbb{R}),\forall B\in M_{nm}(\mathbb{R}) \qquad (6.6.116)$$

D12. A matrix A is named "nilpotent matrix" if by multiplying it successively by itself it can give zero. Explicitly, if there exists an integer k such that:

$$A^k=0 \qquad (6.6.117)$$

If the matrix A multiplied by itself gives A ... then we talk about an "idempotent matrix".

Such matrices are for example very common in Markov Chains when the transition matrix contains probabilities (see the sections of Probabilities and Graph Theory).

Remark

To remember this name, we can decompose it into "nil" that means "zero" and "potent" that means "potential". So something "nilpotent" is therefore something that is potentially zero....

D13. A matrix A is named "orthogonal matrix" if its elements are real and if it obeys to:

$$A^{-1}=A^T \tag{6.6.118}$$

which can be translate into (where δ_{ij} the Kronecker symbol):

$$A^TA=AA^T=\mathbb{1}=[\delta_{ij}] \tag{6.6.119}$$

The matrix column vectors A_i are thus orthogonal to each other as the resulting operation above can be seen as a row-column dot product such that for a $n\times n$ squared matrix:

$$\sum_{i=1}^{n}A_{ik}A_{in}=\delta_{kn} \qquad 1\leq k\leq n \tag{6.6.120}$$

Therefore an orthogonal matrix represents also an orthogonal basis!

Remark

When numerical data are stored in the usual way in a matrix X, the matrix X^TX, and its properties often plays an important role in statistical analysis and especially in regression techniques. A matrix of this form is named a "Gramian matrix".

A typical mathematical example is the matrix of the canonical orthonormal basis (see section Vector Calculus page 1810):

$$A=\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix} \tag{6.6.121}$$

or a well known matrix in quantum physics (see section Quantum Computing page 7291):

$$A=\begin{bmatrix}-\dfrac{1}{\sqrt{2}}&\dfrac{1}{\sqrt{2}}\\\dfrac{1}{\sqrt{2}}&\dfrac{1}{\sqrt{2}}\end{bmatrix} \tag{6.6.122}$$

Remarks

R1. Therefore this is typically the case of the canonical basis matrix, or any diagonalized matrix.

R2. If instead of just taking a matrix with real coefficients, we take complex coefficients with its complex transposed matrix (adjoint matrix), then we say (sadly... because it makes confusion with the name of another matrix already defined) that A is a "unitary matrix" if it satisfies the previous relation!

We will come back later, after having introduced the concepts of eigenvectors and eigenvalues, a particular and very important case of orthogonal matrices (named "translations matrices").

Let us also mention another important property in geometry, physics and statistics of orthogonal matrices.

Theorem 6.38. Given $f(\vec{x})=A\vec{x}+\vec{b}$, where A is an orthogonal matrix and $\vec{b}\in\mathbb{R}^n$. Then f (respectively A) is an isometry. That is to say:

$$\|f(\vec{x})-f(\vec{y})\|=\|\vec{x}-\vec{y}\| \quad \forall\vec{x},\vec{y}\in\mathbb{R}^n \tag{6.6.123}$$

So in other words: Orthogonal matrices are linear mappings which preserve the norm (the distance)!!!

Proof 6.38.1. Remembering that $\vec{x}\circ\vec{y}$ is the same in linear algebra notation as $\vec{x}^T\vec{y}$ we then have:

$$\begin{aligned}\|f(\vec{x})-f(\vec{y})\|^2&=(f(\vec{x})-f(\vec{y}))\circ(f(\vec{x})-f(\vec{y}))=(A\vec{x}-A\vec{y})\circ(A\vec{x}-A\vec{y})\\&=A(\vec{x}-\vec{y})\circ A(\vec{y}-\vec{y})=(A^TA(\vec{x}-\vec{y}))\circ(\vec{x}-\vec{y})=(\vec{x}-\vec{y})\circ(\vec{x}-\vec{y})\end{aligned} \tag{6.6.124}$$

and we have well:

$$\|f(\vec{x})-f(\vec{y})\|=\|\vec{x}-\vec{y}\| \tag{6.6.125}$$

■ Q.E.D.

D14. Given a square matrix $A\in M_{nn}$. The matrix A is named "symmetric matrix" if and only if:

$$A^T=A\Leftrightarrow a_{ij}=a_{ji} \tag{6.6.126}$$

We will meet this definition again in the section of Tensor Calculus.

D15. Given a square matrix $A\in M_{nn}$. The matrix A is named "antisymmetric matrix" (also sometimes named "skew-symmetric matrix"...) if and only if:

$$a_{ij}=-a_{ji} \tag{6.6.127}$$

which requires that:

$$a_{11}=a_{22}=\ldots=a_{mm}=0 \tag{6.6.128}$$

In electromagnetism the electromagnetic field tensor has components which can be written as an antisymmetric 4×4 matrix. In classical mechanics, a solid's propensity to spin in various directions is described by the inertia tensor which is symmetric. The energy-momentum tensor from electrodynamics is also symmetric. Most metric in Einstein field equations in General Relativity are symmetric but we can also build antisymmetric metrics. Matrices are everywhere if look for them.

Remark

We will prove in the section of Tensor Calculus (page 2088) that every square matrix are formed by the sum of a symmetric and antisymmetric matrix.

D16. Let E be a vector space of dimension n and $\mathcal{B},\mathcal{B}'$ two basis of E:

$$\mathcal{B}(\vec{e}_1,\vec{e}_2,\ldots,\vec{e}_n),\mathcal{B}'(\vec{e}_1',\vec{e}_2',\ldots,\vec{e}_n') \tag{6.6.129}$$

We name "transition matrix" of the basis $\mathcal{B}$ to the basis $\mathcal{B}'$, and we denote by P the square matrix of $M_{nn}(\mathbb{K})$ which columns are formed of components of vectors of the basis $\mathcal{B}'$ on the basis $\mathcal{B}$ (see further below the detailed treatment of basis changes for more information).

We consider the vector $\vec{x}(x_1,x_2,\ldots,x_n)$ of E which is written in the basis $\mathcal{B}(\vec{e}_1,\vec{e}_2,\ldots,\vec{e}_n)$ and $\mathcal{B}'(\vec{e}_1',\vec{e}_2',\ldots,\vec{e}_n')$ following the relations:

$$\mathcal{B}:\vec{x}=\sum_{i=1}^{n}x_i\vec{e}_i \qquad \mathcal{B}':\vec{x}=\sum_{i=1}^{n}x_i'\vec{e}_i' \tag{6.6.130}$$

With:

$$\vec{X}=\begin{pmatrix}x_1\\x_2\\\ldots\\x_n\end{pmatrix},\vec{X}'=\begin{pmatrix}x_1'\\x_2'\\\ldots\\x_n'\end{pmatrix} \tag{6.6.131}$$

the vector of K^n formed of the components of $\vec{x}$ in the basis $\mathcal{B}$ and of the vector $\vec{x}$ formed of the components in the basis $\mathcal{B}'$. So:

$$X'=P^{-1}X \tag{6.6.132}$$

relation for which the detailed proof will be given later in our study of basis changes. We also have obviously:

$$X=PX' \tag{6.6.133}$$

Remarks

R1. When a vector is given and its basis is not specified, remember that it is therefore implicitly in the canonical basis:

$$(\vec{e}_1,\vec{e}_2,\ldots,\vec{e}_n)=\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix} \tag{6.6.134}$$

which remains invariant by the multiplication by any vector and when the basis used is denoted by $(\vec{e}_i)$ and is not specified, then it is also that of the canonical basis.

R2. If a vector is given relative to the canonical basis, its components are named "covariant", if they are expressed in another non-canonical base, then we say that the components are "contravariant" (for details on the subject see the section of Tensor Calculus).

D17. A matrix is named "positive-definite matrix" (which will be useful in the section of Numerical Methods for some important engineering techniques and in quantitative finance for qualitative estimation of the correlation matrix) if:

$$\forall \vec{x} \in \mathbb{R}^n: \quad \vec{x}^T A\vec{x} \geq 0 \land \vec{x}^T A\vec{x} = 0 \Longrightarrow \vec{x} = \vec{0} \tag{6.6.135}$$

or more simply if $\vec{x} \neq \vec{0}$:

$$\vec{x}^T A\vec{x} > 0 \tag{6.6.136}$$

and "semi-positive matrix" or "positive-semidefinite matrix" if:

$$\forall \vec{x} \in \mathbb{R}^n: \quad \vec{x}^T A\vec{x} \geq 0 \tag{6.6.137}$$

We have already proved that a semi-positive matrix has its eigenvalues which are all positive OR null, while if it is positive definite its eigenvalues are all positive AND not null (see section Analytical Geometry page 2794).

Notice that this implies that if a matrix A is invertible, then $B = A^T A$ is positive definite.

Indeed (even if we have already indirectly seen this proof in the section of Statistics at page 607), B is positive definite if for every non-zero $\vec{x}$ we have:

$$\vec{x}^T B\vec{x} > 0 \tag{6.6.138}$$

Now if we have A such that $B = A^T A$, then using the properties of transposed matrices:

$$\vec{x}^T B\vec{x} = \vec{x}^T (A^T A)\vec{x} = (A\vec{x})^T (A\vec{x}) = \|A\vec{x}\| \tag{6.6.139}$$

But:

$$\|A\vec{x}\| \geq 0 \tag{6.6.140}$$

If A in invertible, $\|A\vec{x}\| = 0$ only when $\vec{x} = 0$.

Let us prove also that the inverse of a positive definite matrix is also positive definite (important property for the Mahalonobis distance seen at page 2466 for example!).

First, let us recall that if A is positive definite, then we have proved (see section Statistics page 608) that it's determinant is never null and that therefore it's always invertible.

Secondly, we will prove further below at page 2004 during our study of spectral theorem that for a definite positive matrix M we have:

$$\det(M) = \prod_i \lambda_i \tag{6.6.141}$$

where λ_i are the eigenvalues of i (positive and non-null for definite positive matrices!). But also further below during our study of determinant at page 1970, we will prove for a matrix M that:

$$\det(M^{-1}) = \frac{1}{\det(M)} \tag{6.6.142}$$

Therefore:

$$\det(M^{-1}) = \frac{1}{\prod_i \lambda_i} = \prod_i \frac{1}{\lambda_i} \tag{6.6.143}$$

And as for a definite positive matrix we have all $\lambda_i > 0$, then M^{-1} is also positive definite since all $1/\lambda_i > 0$ and hence $\det(M^{-1})$ is also strictly positive.

We can extend the above definition so that in general a matrix M is said to be:

- "positive definite" if for any vector $\vec{x} \neq \vec{0}$, $\vec{x}^T M \vec{x} > 0$. This is sometimes denoted $M \succ \vec{0}$
- "positive semidefinite" if $\vec{x}^T M \vec{x} \geq 0$. This is sometimes denoted $M \succeq \vec{0}$
  - "non-negative definite" if it is either positive definite or positive semi definite[22]
- "negative definite" if for any vector $\vec{x} \neq \vec{0}$, $\vec{x}^T M \vec{x} < \vec{0}$. This is sometimes denoted $M \prec \vec{0}$
- "negative semidefinite" if $\vec{x}^T M \vec{x} \leq \vec{0}$. This is sometimes denoted $M \preceq \vec{0}$
  - "non-positive definite" if it is either negative definite or negative semi definite
- "indefinite" if it is nothing of those.

D18. The "adjugate", "classical adjoint", or "adjunct" of a square matrix is the transpose of its cofactor (see further below page 1971) matrix:

$$\text{adj}(A) = C^T = \left((-1)^{i+j} \det(A_{ij})\right)^T_{1 \leq i,j \leq n} \tag{6.6.144}$$

✍**Example:**

Consider a 3×3 matrix:

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \tag{6.6.145}$$

Its cofactor matrix is:

$$C = \begin{pmatrix} +\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} & -\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} & +\begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix} \\ -\begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix} & +\begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix} & -\begin{vmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{vmatrix} \\ +\begin{vmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{vmatrix} & -\begin{vmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{vmatrix} & +\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \end{pmatrix} \tag{6.6.146}$$

where:

$$\begin{vmatrix} a_{im} & a_{in} \\ a_{jm} & a_{jn} \end{vmatrix} = \det \begin{pmatrix} a_{im} & a_{in} \\ a_{jm} & a_{jn} \end{pmatrix} \tag{6.6.147}$$

---

[22] Non-negative definite and positive semidefinite are the same but as you can read the expression from time to time in some textbooks...

Its adjugate is the transpose of its cofactor matrix:

$$\text{adj}(A)=C^T=\begin{pmatrix} +\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} & -\begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix} & +\begin{vmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{vmatrix} \\ \\ -\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} & +\begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix} & -\begin{vmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{vmatrix} \\ \\ +\begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix} & -\begin{vmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{vmatrix} & +\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \end{pmatrix} \tag{6.6.148}$$

D19. A symmetric square matrix having all it components being positive and only zeros on the diagonal is named a "distance matrix" or even sometimes a "pairwise similarity matrix" (we will meet several times this type of matrix in Data Mining techniques during our study of the Numerical Methods section). Most of times this matrix contains the distances, taken pairwise, between the elements of a set (indexed by the row and column number of the corresponding matrix). Depending upon the application involved, the distance being used to define this matrix may or may not be a metric.

A famous type of such matrix is the Euclidean distance matrix. If D_x^2 is a Euclidean distance matrix and the points $x_1,x_2,\ldots,x_n$ are defined on m-dimensional space, then the elements of A are given by:

$$D_x^2=[d_{ij}^2]=\|x_i-x_j\|_2^2 \tag{6.6.149}$$

where $\|\ldots\|_2$ denotes the 2-norm on $\mathbb{R}^m$ and:

$$D_x^2=\begin{bmatrix} 0 & d_{12}^2 & d_{13}^2 & \ldots & d_{1n}^2 \\ d_{21}^2 & 0 & d_{23}^2 & \ldots & d_{2n}^2 \\ d_{31}^2 & d_{32}^2 & 0 & \ldots & d_{3n}^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{n1}^2 & d_{n2}^2 & d_{n3}^2 & \ldots & 0 \end{bmatrix} \tag{6.6.150}$$

D20. A matrix is named "sparse matrix" if it contains a significant number of null values. In numerical methods, there are algorithms that use this specificity to optimize the storage of this type of matrices (used in OLAP[23] cubes and financial engineering).

$$\begin{bmatrix} 11 & 22 & 0 & 0 & 0 & 0 & 0 \\ 0 & 33 & 44 & 0 & 0 & 0 & 0 \\ 0 & 0 & 55 & 66 & 77 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 88 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 99 \end{bmatrix} \tag{6.6.151}$$

The above sparse matrix contains only 9 non-zero elements, with 26 zero elements. Its sparsity is 74%, and its density is 26%.

D21. A "block matrix"(also named "partitioned matrix") is a matrix of the kind:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \tag{6.6.152}$$

---

[23] Acronym of OnLine Analytical Processing.

where A, B, C and D are matrices, named "blocks", such that:

- A and B have the same number of rows
- C and D have the same number of rows
- A and C have the same number of columns
- B and D have the same number of columns

Ideally, a block matrix is obtained by cutting a matrix two times: one vertically and one horizontally. Each of the four resulting pieces is a block.

✍**Examples:**

E1. The matrix:

$$\begin{bmatrix} 1 & 2 & 1 & 0 \\ 0 & 0 & 3 & 0 \\ 2 & 5 & 0 & 7 \\ 2 & 2 & 2 & 2 \end{bmatrix} \tag{6.6.153}$$

can be written as a block matrix:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \tag{6.6.154}$$

where:

$$\begin{aligned} A &= \begin{bmatrix} 1 & 2 \\ 0 & 0 \end{bmatrix} & B &= \begin{bmatrix} 1 & 0 \\ 3 & 0 \end{bmatrix} \\ C &= \begin{bmatrix} 2 & 5 \\ 2 & 2 \end{bmatrix} & D &= \begin{bmatrix} 0 & 7 \\ 2 & 2 \end{bmatrix} \end{aligned} \tag{6.6.155}$$

E2. The matrix:

$$\begin{bmatrix} 1 & 0 & 3 \\ 2 & 5 & 0 \\ 8 & 2 & 2 \end{bmatrix} \tag{6.6.156}$$

can be written as a block matrix:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \tag{6.6.157}$$

where:

$$\begin{aligned} A &= \begin{bmatrix} 1 \end{bmatrix} & B &= \begin{bmatrix} 0 & 3 \end{bmatrix} \\ C &= \begin{bmatrix} 2 \\ 8 \end{bmatrix} & D &= \begin{bmatrix} 5 & 0 \\ 2 & 2 \end{bmatrix} \end{aligned} \tag{6.6.158}$$

An important fact about block matrices is that their multiplication can be carried out as if their blocks were scalars, by using the standard rule for matrix multiplication:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} E & F \\ G & H \end{bmatrix}=\begin{bmatrix} AE+BG & AF+BH \\ CE+DG & CF+DH \end{bmatrix} \tag{6.6.159}$$

The only caveat is that all the blocks involved in a multiplication (e.g., AE, BG, CE) must be conformable. For example, the number of columns of A and the number of rows of E must coincide.

Many proofs in linear algebra are greatly simplified if one can easily deal with the determinants of block matrices, that is, matrices that are subdivided into blocks that are themselves matrices.

Finally, to close this enumeration of types of matrices, remember that we saw the "correlation matrix" on page 621, the "quasi-correlation matrix" on page 607 and also the Fisher information matrix at page 839.

6.6.3.4 Determinant

We will look at determinants in the point of view of the physicist or of the engineer (the mathematician point of view is rather off-putting ...). In physics (whether in classical mechanics and quantum field physics), chemistry or engineering, we frequently have to solve linear systems. But we have now seen that a linear system:

$$\begin{aligned} a_{11}x_1+a_{12}x_2+\ldots+a_{1n}x_n&=b_1\\ a_{21}x_1+a_{22}x_2+\ldots+a_{2n}x_n&=b_2\\ &\ldots\\ a_{m1}x_1+a_{m2}x_2+\ldots+a_{mn}x_n&=b_m \end{aligned} \tag{6.6.160}$$

can be written as:

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}=\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \Leftrightarrow AX=B \tag{6.6.161}$$

and we know that the only soluble linear systems (in the sense that they have a unique solution !!!) are those that have as many equations as unknowns and their determinant is not zero! Thus, the matrix A must be a square matrix M_{mm}.

If a solution exists, then there is a column matrix (or "vector") X such that $AX=B$ which involves:

$$X=A^{-1}B \tag{6.6.162}$$

What imposed this relation? Well this is relatively simple, but at the same time very important: for a linear system to have a unique solution, it is necessary that the matrix is invertible (non-singular)! What relation with the concept of "determinant" then? It's simple: mathematicians have sought how to write inverse matrices of linear systems for which they knew there was a unique solution and they arrived after trial and error to determine a kind of formula to assess if the matrix is invertible (non-singular) or not. Once this formula found, they formalized (as they know so well how to do it...), with a great rigour, the concept surrounding this formula that they named the "determinant". They did it so well that in fact we sometimes forgot that they have found it by trial and error...

Remark

If a matrix of a linear system is not invertible (non-singular), this has the consequence that there is no solution or an infinity of solutions (as usual what ...)

We below first focus on how to build the determinant by defining a particular type of application. Then, after seeing a simple and readable example of the calculation of a determinant, we will focus on determining the formula of it in the general case. Finally, once this is done, we will see what is the relation between the inverse of a matrix and the determinant!!!

In what will follow all vector spaces will be considered of finite dimension and the field of complex numbers $\mathbb{C}$ (those who prefer the can take $\mathbb{R}$ as basis field, in fact we could take any field).

First of all we will do a little bit of pure math (a bit off putting) before moving on concrete stuff.

Given V a vector space, we will write as usual V^n instead of $V\times V\times\ldots\times V$. $(\vec{e}_1,\vec{e}_2,\ldots,\vec{e}_3)$ designate the canonical basis of $\mathbb{R}^n$. $M_n(\mathbb{R})$ is the set of square matrices $n\times n$ with coefficients in $\mathbb{R}$.

Definitions (#6.6.257):

D1. A "multi-linear application" on a space V is defined by an $\varphi:V^n\rightarrow\mathbb{R}$ which is linear in each of its components. Meaning:

$$\begin{aligned}&\varphi(\vec{x}_1,\ldots,\vec{x}_{j-1},\lambda\vec{u}+\mu\vec{v},\vec{x}_{j+1},\ldots,\vec{x}_n)\\&=\lambda\varphi(\vec{x}_1,\ldots,\vec{x}_{j-1},\vec{u},\vec{x}_{j+1},\ldots,\vec{x}_n)+\mu\varphi(\vec{x}_1,\ldots,\vec{x}_{j-1},\vec{v},\vec{x}_{j+1},\ldots,\vec{x}_n)\end{aligned}\tag{6.6.163}$$

for any $\lambda,\mu\in\mathbb{K}$ and $\vec{x}_i,\vec{u},\vec{v}\in V$ where the $\vec{x}_i$ are vectors.

Remark

A non-null multi-linear application is not a linear application of the space V^n in $\mathbb{R}^n$. Excepted if $n=1$. Indeed, this can be verified by the definition of a linear application versus the one of the multi-linear application:

$$\lambda\varphi(\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_n)\neq\varphi(\lambda\vec{x}_1,\lambda\vec{x}_2,\ldots,\lambda\vec{x}_n)\tag{6.6.164}$$

$$\varphi(\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_n)+\varphi(\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_n)\neq\varphi(2\vec{x}_1,2\vec{x}_2,\ldots,2\vec{x}_n)\tag{6.6.165}$$

D2. An "alternated multi-linear application" on V is by definition a multi-linear application that satisfies:

$$\varphi(\vec{x}_1,\ldots,\vec{x}_j,\vec{x}_{j+1},\vec{x}_{j+2},\ldots,\vec{x}_n)=-\varphi(\vec{x}_1,\ldots,\vec{x}_{j+1},\vec{x}_j,\vec{x}_{j+2},\ldots,\vec{x}_n)\tag{6.6.166}$$

for any $j=1\ldots n, \vec{x}_j \in V$. Therefore the permutation of two vectors that follows change the sign of φ.

Theorem 6.39. Therefore, if φ is a multi-linear application, then φ is alternated if and only if $\forall \vec{j} \in V, j=1\ldots n$ we have:

$$\varphi(\vec{x}_1,\ldots,\vec{x}_j,\vec{x}_j,\vec{x}_{j+2},\ldots,\vec{x}_n)=0 \tag{6.6.167}$$

or in a more simple case:

$$\varphi(\vec{x}_j,\vec{x}_j)=0 \tag{6.6.168}$$

Proof 6.39.1. If φ is alternated we have by definition:

$$\varphi(\vec{x}_1,\ldots,\vec{x}_j,\vec{x}_{j+1},\vec{x}_{j+2},\ldots,\vec{x}_n)=-\varphi(\vec{x}_1,\ldots,\vec{x}_{j+1},\vec{x}_j,\vec{x}_{j+2},\ldots,\vec{x}_n) \tag{6.6.169}$$

Therefore by rearranging:

$$\varphi(\vec{x}_1,\ldots,\vec{x}_j,\vec{x}_{j+1},\vec{x}_{j+2},\ldots,\vec{x}_n)+\varphi(\vec{x}_1,\ldots,\vec{x}_{j+1},\vec{x}_j,\vec{x}_{j+2},\ldots,\vec{x}_n)=0 \tag{6.6.170}$$

And if φ is a multi-linear application we can write:

$$2\varphi(\vec{x}_1,\ldots,\vec{x}_j,\vec{x}_{j+1},\vec{x}_{j+2},\ldots,\vec{x}_n)=0 \tag{6.6.171}$$

■ Q.E.D.

Now comes the interesting stuff:

D3. A "determinant" is by definition a multi-linear alternated application:

$$D:\underbrace{\mathbb{R}^n\times\ldots\times\mathbb{R}^n}_{n\text{ times}}\mapsto\mathbb{R} \tag{6.6.172}$$

satisfying as well:

$$D(\vec{e}_1,\ldots,\vec{e}_n)=\det(\vec{e}_1,\ldots,\vec{e}_n)=1 \tag{6.6.173}$$

Remark

The columns of a matrix form n vectors and we then see that a determinant D on $\mathbb{R}^n$ induce an application D of $M_n(\mathbb{R})\mapsto\mathbb{R}$ (where as we know $M_n(\mathbb{R})$ is the set of square matrices $n\times n$ with components in $\mathbb{R}$) defined by:

$$D(M)=D(\vec{m}_1,\vec{m}_2,\ldots,\vec{m}_n) \tag{6.6.174}$$

where $\vec{m}_i$ is the i-th column of M.

Let us study the case $n=2$. If D is a determinant, for any vector:

$$\vec{x}=\lambda\vec{e}_1+\mu\vec{e}_2 \qquad \vec{y}=\alpha\vec{e}_1+\beta\vec{e}_2 \tag{6.6.175}$$

we have:

$$D(\vec{x},\vec{y})=D(\lambda\vec{e}_1+\mu\vec{e}_2,\alpha\vec{e}_1+\beta\vec{e}_2) \tag{6.6.176}$$

As D is multi-linear, we have:

$$\begin{aligned}D(\vec{x},\vec{y})&=D(\lambda\vec{e}_1+\mu\vec{e}_2,\alpha\vec{e}_1+\beta\vec{e}_2)=\lambda D(\vec{e}_1,\alpha\vec{e}_1+\beta\vec{e}_2)+\mu D(\vec{e}_2,\alpha\vec{e}_1+\beta\vec{e}_2)\\&=\lambda(\alpha D(\vec{e}_1,\vec{e}_1)+\beta D(\vec{e}_1,\vec{e}_2))+\mu(\alpha D(\vec{e}_2,\vec{e}_1)+\beta D(\vec{e}_2,\vec{e}_2))\\&=\lambda\alpha D(\vec{e}_1,\vec{e}_1)+\lambda\beta D(\vec{e}_1,\vec{e}_2)+\mu\alpha D(\vec{e}_2,\vec{e}_1)+\mu\beta D(\vec{e}_2,\vec{e}_2)\end{aligned}\tag{6.6.177}$$

as it is alternated:

$$D(\vec{e}_1,\vec{e}_1)=0 \qquad D(\vec{e}_2,\vec{e}_2)=0\tag{6.6.178}$$

it remains:

$$D(\lambda\vec{e}_1+\mu\vec{e}_2,\alpha\vec{e}_1+\beta\vec{e}_2)=\lambda\beta D(\vec{e}_1,\vec{e}_2)+\mu\alpha D(\vec{e}_2,\vec{e}_1)\tag{6.6.179}$$

and we have:

$$D(\lambda\vec{e}_1+\mu\vec{e}_2,\alpha\vec{e}_1+\beta\vec{e}_2)=\lambda\beta D(\vec{e}_1,\vec{e}_2)-\mu\alpha D(\vec{e}_1,\vec{e}_2)\tag{6.6.180}$$

And finally:

$$D(\lambda\vec{e}_1+\mu\vec{e}_2,\alpha\vec{e}_1+\beta\vec{e}_2)=\lambda\beta-\mu\alpha\tag{6.6.181}$$

In fact, we just proved that if the a determinant exists, it is unique and of the form indicated previously, we should also check that the defined application satisfy the properties of a determinant, but the latter is immediate.

Thus, if:

$$M=\begin{bmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{bmatrix}\tag{6.6.182}$$

is a matrix we have then (notice the three common notations for the determinant that you can find in various textbooks and even in the actual one depending on the context and of the traditions):

$$D(M)=\det(M)=|M|=a_{11}a_{22}-a_{21}a_{12}\tag{6.6.183}$$

Let us give now a geometric interpretation of the determinant. Given $\vec{v}_1,\vec{v}_2$ two vectors of $\mathbb{R}^2$:

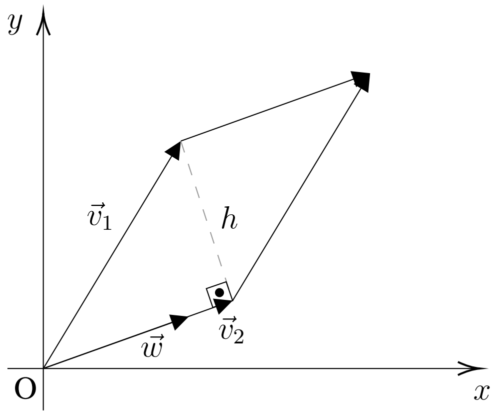


Figure 6.144: Geometrical interpretation of the determinant

The vector $\vec{w}$ is obtained by projecting $\vec{v}_1$ on $\vec{v}_2$ and we have therefore:

$$\vec{w} = \frac{\vec{v}_1 \circ \vec{v}_2}{\|\vec{v}_2\|^2}\vec{v}_2 \text{ and } \|\vec{h}_2\|^2 = \|\vec{v}_1\|^2 - \|\vec{v}_2\|^2 = \|\vec{v}_1\|^2 - \frac{(\vec{v}_1 \circ \vec{v}_2)^2}{\|\vec{v}_2\|^4}\|\vec{v}_2\|^2 = \|\vec{v}_1\|^2 - \frac{(\vec{v}_1 \circ \vec{v}_2)^2}{\|\vec{v}_2\|^2} \tag{6.6.184}$$

The area of the above parallelogram is therefore:

$$S = \|\vec{v}_2\| \cdot \|\vec{h}\| = \|\vec{v}_2\| \left(\|\vec{v}_1\|^2 - \frac{(\vec{v}_1 \circ \vec{v}_2)^2}{\|\vec{v}_2\|^2}\right)^{1/2} = \left(\|\vec{v}_1\|^2\|\vec{v}_2\|^2 - (\vec{v}_1 \circ \vec{v}_2)^2\right)^{1/2} \tag{6.6.185}$$

if:

$$\vec{v}_1 = \lambda\vec{e}_1 + \mu\vec{e}_2 \quad \text{and} \quad \vec{v}_2 = \alpha\vec{e}_1 + \beta\vec{e}_2 \tag{6.6.186}$$

then:

$$S^2 = \left(\lambda^2 + \mu^2\right)\left(\alpha^2 + \beta^2\right) = \lambda^2\beta^2 + \mu^2\alpha^2 - 2\lambda\alpha\mu\beta = (\lambda\beta - \mu\alpha)^2 \tag{6.6.187}$$

and finally:

$$S = |\lambda\beta - \mu\alpha| = |D(\vec{v}_1, \vec{v}_2)| = \left|\det\left(\begin{bmatrix}\lambda & \alpha \\ \mu & \beta\end{bmatrix}\right)\right| \tag{6.6.188}$$

Therefore the determinant, in absolute value, represents the area of the parallelogram defined by the vectors $\vec{v}_1, \vec{v}_2$ when these vectors are linearly independent. We can generalize this result to a n-dimensional space, in particular, for $n = 3$, the determinant of three linearly independent vectors represents the volume of the parallelepiped defined by these as we already proved it during our study of the mixed product in the section of Vector Calculus page 1838.

The more general case of the expression of the determinant is a little trickier to ascertain. This requires that we define a particular bijective application but simple that we have already met in the section Statistics.

Definition (#6.6.258): Given $F_n = \{1, 2, \ldots, n\}, n \in \mathbb{N}^*$ we name "permutation" of F_n any bijective application of F_n in F_n:

$$\sigma : \{1, 2, \ldots, n\} \rightarrow \{1, 2, \ldots, n\} \tag{6.6.189}$$

Given $\mathcal{S}_n$ the set of possible permutations (bijective applications) of $\{1, 2, \ldots, n\}$. $\mathcal{S}_n$ obviously contains ... (see our study of Combinatorics in the section of Probabilities) $n!$ elements. The information an element σ of $\mathcal{S}_n$ is defined by the successive information of:

$$\sigma(1)\ (\sigma(1) \in F_n) \tag{6.6.190}$$

$$\sigma(2)\ (\sigma(2) \in F_n - \{\sigma(1)\}) \tag{6.6.191}$$

$$\ldots\sigma(n)\ (\sigma(n) \in F_n - \{\sigma(1), \sigma(2), \ldots, \sigma(n-1)\}) \tag{6.6.192}$$

Given an ordered sequence of elements (ascending) $\{1, 2, \ldots, n\}, n \in \mathbb{N}^*$ we name "inversion", any permutation of elements in the ordered sequence (so the result will not be ordered at all...). We denote by $I(\sigma)$ the number of inversions.

We say that the permutation σ is even (odd) if $I(\sigma)$ is even (odd). We name "signature" of σ, the number $\varepsilon(\sigma)$ defined by $\varepsilon(\sigma)=(-1)^{I(\sigma)}$, that is to say:

$$\varepsilon(\sigma)=\begin{cases}+1 & \text{if the permutation } \sigma \text{ is even}\\ -1 & \text{if the permutation } \sigma \text{ is odd}\end{cases}\tag{6.6.193}$$

We now have the necessary tools to set up the general relation of the determinant!

Definition (#6.6.259): Given:

$$A=(a_{ij})_{1\leq i,j\leq n}\in M_{nn}(\mathbb{K})\tag{6.6.194}$$

We name "determinant of a square matrix A" of dimension n, and we denote by $\det(A)$ (or sometimes by $D(A)$ or especially in the field of Statistics by $|A|$), the scalar defined by (we'll see a few examples further below):

$$\boxed{\det(A)=\sum_{\sigma\in\mathcal{S}_n}\varepsilon(\sigma)a_{1\sigma(1)}a_{2\sigma(2)}\ldots a_{n\sigma(n)}=\sum_{\sigma\in\mathcal{S}_n}\varepsilon(\sigma)\prod_{i=1}^{n}a_{i\sigma(i)}}\tag{6.6.195}$$

sometimes named "Leibniz formula of determinants" or "Laplace's formula". This relation seems to have been obtained in the past by trial and error and by induction for larger dimensions.

✍**Examples:**

E1. Given $A=(a_{ij})_{1\leq i,j\leq 2}\in M_{22}(\mathbb{K})$, let us consider the $2!=2$ permutations of the second indices (of integers $1,2$) taken in their whole:

$$12 \qquad 21$$

We calculate the signature of σ. Here is the scheme of this rule (recall: we say that there is an "inversion", if in a permutation, a greater integer precede a smallest integer):

$\sigma(1)\sigma(2)$	$\underbrace{12}_{\sigma(1)\sigma(2)}$	$\underbrace{21}_{\sigma(1)\sigma(2)}$
Number of inversions	0	1
Permutation	even	odd
$\varepsilon(\sigma)=(-1)^{I(\sigma)}$	$+1$	-1

Table 6.18: **Inversions and permutations of a determinant of order 2**

Therefore we have:

$$\begin{aligned}\det\begin{bmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{bmatrix}&=\sum_{\sigma\in\mathcal{S}_n}\varepsilon(\sigma)a_{1\sigma(1)}a_{2\sigma(2)}\\ &=\underbrace{a_{11}a_{22}}_{\varepsilon(\sigma)=+1\begin{cases}\sigma(1)=1\\ \sigma(2)=2\end{cases}}-\underbrace{a_{12}a_{21}}_{\varepsilon(\sigma)=-1\begin{cases}\sigma(1)=2\\ \sigma(2)=1\end{cases}}\end{aligned}\tag{6.6.196}$$

This corresponds well to what we saw initially. Remember also on the way that we will soon prove that the determinant of a square matrix must be zero so that the matrix is invertible (non-singular)!

E2. Given $A=(a_{ij})_{1\leq i,j\leq 3}\in M_{33}(\mathbb{K})$, let us consider the $3!=6$ permutations of the second indices (integers $1,2,3$) taken in their whole:

$$123 \quad 132 \quad 213 \quad 31 \quad 312 \quad 321$$

We calculate the signatures of σ. Here is a scheme of this rule (recall: we say that there is an "inverse", if in a permutation, a greatest integer precedes a lower integer):

$\sigma(1)\sigma(2)$	123	132	213	231	312	321
Number of inversions	0	1	1	2	2	3
Permutation	even	odd	odd	even	even	odd
$\varepsilon(\sigma)=(-1)^{I(\sigma)}$	$+1$	-1	-1	$+1$	$+1$	-1

Table 6.19: **Inversions and permutations of a determinant of order** 3

Therefore we have:

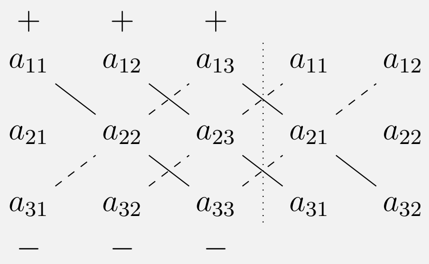
$$\begin{aligned}\det\begin{bmatrix}a_{11} & a_{12} & a_{13}\\ a_{21} & a_{22} & a_{23}\\ a_{31} & a_{32} & a_{33}\end{bmatrix} &= \sum_{\sigma\in\mathcal{S}_n}\varepsilon(\sigma)a_{1\sigma(1)}a_{2\sigma(2)}a_{3\sigma(3)}\\ &=a_{11}\det\begin{pmatrix}a_{22} & a_{23}\\ a_{32} & a_{33}\end{pmatrix}-a_{12}\det\begin{pmatrix}a_{21} & a_{23}\\ a_{31} & a_{33}\end{pmatrix}+a_{13}\det\begin{pmatrix}a_{21} & a_{22}\\ a_{31} & a_{32}\end{pmatrix}\\ &=a_{11}a_{22}a_{33}-a_{11}a_{23}a_{32}-a_{12}a_{21}a_{33}+a_{12}a_{23}a_{31}+a_{13}a_{21}a_{32}-a_{13}a_{22}a_{31}\end{aligned}\tag{6.6.197}$$

From the above formula, each term is in the form of $a_{1j_1}a_{2j_2}a_{3j_3}$ where the second indices j_1,j_2,j_3 is a permutation of $1,2,3$. For example, the first term has $1,2,3$ as the second indices; the second term has $1,3,2$; the third term has $2,1,3$, etc.

The sign of each term is the sign of the permutation. For example, the sign of $1,2,3$ is clearly 1. For the second term, $1,3,2$ is obtained from $1,2,3$ by one transposition $2,3\rightarrow 3,2$, so the sign is -1, etc.

Remark

Some people learn by heart a method named "rule of Sarrus" to calculate the determinants of order three as the previous one given by:


+ + +
a_{11} a_{12} a_{13} a_{11} a_{12}
a_{21} a_{22} a_{23} a_{21} a_{22}
a_{31} a_{32} a_{33} a_{31} a_{32}
− − −

We prefer in this textbook the general formulation of the determinant because applicable to all orders.

Let us see some properties and corollaries of this formulation of the determinant:

P1. Given a square matrix $M_n(K)$ of order n, we do not change the value of the determinant of M_n:

  (a) By performing an elementary operation on the columns of M_n.
  (b) By performing an elementary operation on the rows of M_n.

Proof 6.39.2. If $M_n=(a_{ij})_{i,j=1\dots n}$ then M_n is composed of n column vectors:

$$\vec{v}_n=\begin{pmatrix} a_{1j}\\ a_{2j}\\ \vdots \\ a_{nj}\end{pmatrix} \tag{6.6.198}$$

Doing an elementary operation on the columns of M_n is equivalent to add $\lambda v_i, i\in\{1,\dots,n\}$ to one of the columns v_n of M_n. Given M_n' the matrix obtained by adding λv to the j-th column of M_n, we get:

$$\det(M_n')=\det(v_1,\dots,v_i,\dots,v_j+\lambda v_i,\dots,v_n) \tag{6.6.199}$$

By multi-linearity (finally the proof in not really difficult):

$$\det(M')=\det(v_1,\dots,v_i,\dots,v_j,\dots,v_n)+\lambda\det(v_1,\dots,v_i,\dots,v_i,\dots,v_n) \tag{6.6.200}$$

and as the determinant is alternated:

$$\lambda\det(v_1,\dots,v_i,\dots,v_i,\dots,v_n)=0\Leftarrow\det(M_n')=\det(v_1,\dots,v_i,\dots,v_j,\dots,v_n)=\det(M_n) \tag{6.6.201}$$

About the elementary operations on the rows we just need to consider the transpose (that is to cry such it is simple, but we had to thing about this trick).

■ Q.E.D.

P2. Given $M_n(\mathbb{K})$ a square matrix of order n and given $\lambda\in\mathbb{K}$:

$$\boxed{\det(\lambda M_n)=\lambda^n\det(M_n)} \tag{6.6.202}$$

Proof 6.39.3. As before, it is enough to simply noticed that if $v_1,\dots,v_n$ are the column vectors forming the matrix M_n then $\lambda v_1,\dots,\lambda v_n$ are those that constitute λM_n and:

$$\det(\lambda M_n)=\det(\lambda v_1,\dots,\lambda v_n) \tag{6.6.203}$$

The application being n-linear, we arrive at the equality:

$$\det(\lambda M_n)=\lambda^n\det(v_1,\dots,v_n)=\lambda^n\det(M_n) \tag{6.6.204}$$

■ Q.E.D.

P3. Given $M_n(\mathbb{K})$ a square matrix of order n. We change the sign of the determinant of M_n if:

- We permute two of its columns
- We permute two of its rows

Proof 6.39.4. M_n is constituted by n vectors $v_1,..,v_n$. The determinant of M_n is equal to the determinant of these n. Permute two columns of M_n is same as permuting the two corresponding vectors. Let us suppose that the permuted vectors are the i-th and j-th, the determinant being an alternate application, we have:

$$\det(v_1,\ldots,v_i,\ldots,v_j,\ldots,v_n)=-\det(v_1,\ldots,v_j,\ldots,v_i,\ldots,v_n) \tag{6.6.205}$$

About the rows, we just have to consider the transposed of M_n to arrive to the same result!

■ Q.E.D.

P4. Given $A,B\in M_n\in(\mathbb{C})$ then:

$$\boxed{\det(A\cdot B)=\det(A)\cdot\det(B)} \tag{6.6.206}$$

As far as we know the proof can be done in at least two ways, the first is rather indigestible and abstract... so we will let it to mathematicians (...) even if it has the advantage of being general, the second one (much more easier) is to check this assertion for various squared matrices (the way engineers would do it...):

Proof 6.39.5.

$$\begin{aligned}\det(A\cdot B)=\det\left(\begin{bmatrix}a_{11} & a_{22}\\ a_{21} & a_{22}\end{bmatrix}\begin{bmatrix}b_{11} & b_{22}\\ b_{21} & b_{22}\end{bmatrix}\right)&=\det\left(\begin{bmatrix}a_{11}b_{11}+a_{12}b_{21} & a_{11}b_{12}+a_{12}b_{22}\\ a_{21}b_{11}+a_{22}b_{11} & a_{21}b_{12}+a_{22}b_{22}\end{bmatrix}\right)\\ &=(a_{11}b_{11}+a_{12}b_{21})(a_{21}b_{11}+a_{22}b_{11})-(a_{11}b_{12}+a_{12}b_{22})(a_{21}b_{11}+a_{22}b_{11})\\ &=\cancel{a_{11}b_{11}a_{21}b_{12}}+a_{11}b_{11}a_{22}b_{22}+a_{12}b_{21}a_{21}b_{12}+\cancel{a_{12}b_{21}a_{22}b_{22}}\\ &-\cancel{a_{11}b_{12}a_{21}b_{11}}-a_{11}b_{12}a_{22}b_{21}-a_{12}b_{22}a_{21}b_{11}-\cancel{a_{12}b_{22}a_{22}b_{21}}\\ &=a_{11}b_{11}a_{22}b_{22}-a_{11}b_{12}a_{22}b_{21}-a_{12}b_{22}a_{21}b_{11}+a_{12}b_{21}a_{21}b_{12}\\ &=a_{11}a_{22}b_{11}b_{22}-a_{11}a_{22}b_{12}b_{21}-a_{12}a_{21}b_{22}b_{11}+a_{12}a_{21}b_{21}b_{12}\end{aligned} \tag{6.6.207}$$

and:

$$\begin{aligned}\det(A)\det(B)&=(a_{11}a_{22}-a_{12}a_{21})(b_{11}b_{22}-b_{12}b_{21})\\ &=a_{11}a_{22}b_{11}b_{22}-a_{11}a_{22}b_{12}b_{21}-a_{12}a_{21}b_{22}b_{11}+a_{12}a_{21}b_{21}b_{12}\end{aligned} \tag{6.6.208}$$

The calculations therefore produce results that are identical. We can check for square matrices of higher dimensions.

■ Q.E.D.

P5. A square matrix $A\in M_n(\mathbb{C})$ is invertible (non-singular) if and only if $\det(A)\neq 0$ (this is the most important property among all).

Proof 6.39.6. If A is invertible (non-singular), we have:

$$\boxed{\det\left(A\cdot A^{-1}\right)=\det(\mathbb{1})=\det(A)\det\left(A^{-1}\right)\Rightarrow\det(A)\neq 0} \tag{6.6.209}$$

■ Q.E.D.

Notice that as $\det(\mathbb{1})=1$, if follow immediately that :

$$\det(A^{-1})=\frac{1}{\det(A)} \tag{6.6.210}$$

As we already said it, this is the most important property of matrices as part of theoretical physics because if A is a linear system, the calculation of the determinant indicates whether it has unique solutions or not. Otherwise, as we have already mentioned and study it, the system has no solution, or an infinity of solutions!

We must consider an important special case! Given the following system:

$$A\cdot B=0\Longrightarrow B=0 \tag{6.6.211}$$

where $A\in M_n(\mathbb{K})\neq 0$ and $B\in M_n(\mathbb{K})\neq 0$ are to be determined. It is obvious that A is invertible (non-singular) or not, the trivial solution is $A\cdot B=0$. However, let us imagine a case of theoretical physics where we have $A\cdot B=0$ but for which we know that $A\in M_n(\mathbb{K})\neq 0$ for which we impose $B\in M_n(\mathbb{K})\neq 0$. In this case, we must eliminate the trivial solution $B=0$. Furthermore, calculate the inverse (if it exists) of the matrix A will bring us to nothing concrete except that $B=0$ which obviously does not satisfy us. The only solution is then to play such that the coefficients a_{ij} of the matrix A are such that its determinant is zero and therefore the matrix in invertible! The advantage? Just to have an infinite number of possible solutions (of B then!) that satisfy $A\cdot B=0$. We will need this methodology in the section of Wave Quantum Physics, when we will determine the existence of antiparticles through the linearised Dirac equation. It must therefore be remember.

P6. Two "conjugated" matrices (be careful! it is not the "conjugate" in the complex sense) have the same determinant.

Proof 6.39.7. Let $A\in M_n(\mathbb{C})$, and $P\in M_n(\mathbb{C})$ be a transition matrix from one base to another (see further our study of base changes), then we have:

$$\det(B)=\det\left(P^{-1}AP\right)=\underbrace{\det(P^{-1})}_{=(\det(P))^{-1}}\det(A)\det(P)=\det(A) \tag{6.6.212}$$

$$\det(\overline{A})=\overline{\det(A)} \tag{6.6.213}$$

■ Q.E.D.

P7. For any matrix $A\in M_n(\mathbb{C})$:

Proof 6.39.8.

$$\det(\overline{A})=\sum_{\sigma\in\mathcal{S}_n}\varepsilon(\sigma)\cdot\overline{a_{1,\sigma(1)}}\cdot\ldots\cdot\overline{a_{n,\sigma(n)}} \tag{6.6.214}$$

But as (trivial... simple product of all coefficients):

$$\overline{a_{1,\sigma(1)}}\cdot\ldots\cdot\overline{a_{n,\sigma(n)}}=\overline{a_{\sigma-1(1),1}}\cdot\ldots\cdot\overline{a_{\sigma-1(n),n}} \tag{6.6.215}$$

As (trivial) $\varepsilon(\sigma^-1)=\varepsilon(\sigma)$ and that $x,y\in\mathbb{C}:\bar{x}\cdot\bar{y}=\overline{x\cdot y}$ (see section Numbers page 247) then we can write:

$$\begin{aligned}\det(\bar{A})&=\sum_{\sigma\in\mathcal{S}_n}\varepsilon(\sigma)\cdot\overline{a_{\sigma^{-1}(1),1}}\cdot\ldots\cdot\overline{a_{\sigma^{-1}(n),n}}=\overline{\sum_{\sigma^{-1}\in\mathcal{S}_n}\varepsilon(\sigma^{-1})\cdot a_{\sigma^{-1}(1),1}\cdot\ldots\cdot a_{\sigma^{-1}(n),n}}\\&=\overline{\det(A)}\end{aligned} \tag{6.6.216}$$

■ Q.E.D.

P8. For any matrix $A\in M_n(\mathbb{R})$:

$$\det(A^T)=\det(A) \tag{6.6.217}$$

Proof 6.39.9. Well... it's the same as the previous property but without the conjugate values... In fact, we prove in the same way, the same property for $A\in M_n(\mathbb{C})$.

■ Q.E.D.

P9. Given a matrix $A=(a_{ij})\in M_n(\mathbb{C})$ we denote by A_{ij} the matrix obtained from A by removing the i-th row and the j-th column (very important notation to remember for what will follow!!!). The A_{ij} belongs therefore to $M_{n-1}(\mathbb{C})$. Then for any $i=1\ldots n$:

$$\boxed{\det(A)=\sum_{j=1}^{n}(-1)^{i+j}\cdot a_{ij}\cdot\det(A_{ij})} \tag{6.6.218}$$

where the term:

$$C_{ij}=(-1)^{i+j}\cdot\det(A_{ij}) \tag{6.6.219}$$

is named the "cofactor[24]" or "partial determinant", $\det(A_{ij})$ is named the "minor", and the whole boxed relation above is named the "cofactor expansion theorem".

Proof 6.39.10. For the proof let us define the application:

$$\varphi:M_n(\mathbb{C})\mapsto\mathbb{C},\varphi(A):=\sum_{j=1}^{n}(-1)^{i+j}a_{ij}\det(A_{ij}) \tag{6.6.220}$$

It could be almost easy to see that φ is multi-linear (you just have to consider that $(-1)^{i+j}a_{ij}$ as a simple constant and after by extension of the definition of the determinant... too easy...).

---

[24] From all the cofactors we can build a "cofactor matrix" containing all C_{ij}.

Let us show however that this application is alternated (in this case it is a determinant hat has all the properties of a... determinant!).

Given a_k, a_{k+1} two column vectors A that follows each other. Let us suppose that $a_k = a_{k+1}$, we have to show in this case that $\varphi(A) = 0$ (which comes from the definition of an alternate application).

We have first (it is mandatory by the definition itself) if we don't erase any of the columns j being k or $k+1$:

$$\det(A_{ij}) = 0 \quad \text{if } j \neq k, k+1 \tag{6.6.221}$$

and we have obviously if we remove respectively the column k and the column $k+1$:

$$\det(A_{ik}) = \det(A_{i(k+1)}) \quad \text{and} \quad A_{i,k} = A_{i,k+1} \tag{6.6.222}$$

Therefore:

$$\varphi(A) = (-1)^{i+k} a_{ik} \det(A_{ik}) + (-1)^{i+(k+1)} a_{i(k+1)} \det(A_{i(k+1)}) = 0 \tag{6.6.223}$$

It is therefore OK. The application φ is alternated and multi-linear, it is indeed well a determinant!

We have just proved that φ is a determinant and by unicity we have $\varphi(A) = \det(A)$ for any $A \in M_n(\mathbb{C})$.

■ Q.E.D.

Remark

So finally we have seen now in details the concepts of cofactor that should permit the reader to understand the concept of "adjugate matrix" (or "comatrix") that we had introduce at the beginning of this section at the page 1958.

If we take the following 2×2 matrix:

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \tag{6.6.224}$$

we get:

$$\text{adj}(A) = C^T = \left((-1)^{i+j} \det(A_{ij})\right)^T_{1 \leq i,j \leq n} = \text{adj}(A) = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \tag{6.6.225}$$

Or if we take the following 3×3 matrix (already used in an example earlier!):

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \tag{6.6.226}$$

Then the cofactor matrix is given by:

$$C = \begin{pmatrix} +\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} & -\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} & +\begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix} \\ \\ -\begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix} & +\begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix} & -\begin{vmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{vmatrix} \\ \\ +\begin{vmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{vmatrix} & -\begin{vmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{vmatrix} & +\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \end{pmatrix} \tag{6.6.227}$$

And we notice something that can be generalized: if A is symmetric, then $C=C^T$.

✍**Example:**

Let us see an example of this method by calculating the determinant of:

$$A=\begin{pmatrix}1 & 2 & 3\\ 2 & 1 & 0\\ 1 & -1 & \mathrm{i}\end{pmatrix}\tag{6.6.228}$$

Let us develop the second line ($i=2$). We get:

$$\begin{aligned}\det(A)&=\begin{bmatrix}1 & 2 & 3\\ 2 & 1 & 0\\ 1 & -1 & \mathrm{i}\end{bmatrix}\\ &=\sum_{j=1}^{n}(-1)^{i+j}\cdot a_{ij}\cdot\det(A_{ij})=\sum_{j=1}^{3}(-1)^{2+j}\cdot a_{2j}\cdot\det(A_{2j})\\ &=(-1)^3\cdot 2\cdot\det\begin{pmatrix}2 & 3\\ -1 & \mathrm{i}\end{pmatrix}+(-1)^4\cdot 1\cdot\det\begin{pmatrix}1 & 3\\ 1 & \mathrm{i}\end{pmatrix}+(-1)^5\cdot 0\cdot\det\begin{pmatrix}1 & 2\\ 1 & -1\end{pmatrix}\\ &=-2\cdot(2\mathrm{i}+3)+\mathrm{i}-3=-3\mathrm{i}-9\end{aligned}\tag{6.6.229}$$

Let us develop following the first column for verification (we never know...):

$$\begin{aligned}\det(A)=\begin{bmatrix}1 & 2 & 3\\ 2 & 1 & 0\\ 1 & -1 & \mathrm{i}\end{bmatrix}&=\det\begin{bmatrix}1 & 0\\ -1 & \mathrm{i}\end{bmatrix}-2\cdot\det\begin{bmatrix}2 & 3\\ -1 & \mathrm{i}\end{bmatrix}+\det\begin{bmatrix}2 & 3\\ 1 & 0\end{bmatrix}\\ &=\mathrm{i}-2(2\mathrm{i}+3)-3=-3\mathrm{i}-9\end{aligned}\tag{6.6.230}$$

Or the reader can check in Maple™ 4.00b by using:

```
>with(linalg):
>det(array([[1,2,3],[2,1,0],[1,-1,I]]));
```

The calculation determined above is therefore "exponential" as if for example we must calculate the determinant of a square matrix of order (dimension) $n=10$, then the determinant will be developed in a sum of 10 terms, which each contains the determinant of a matrix of dimension $n=9$, which is a cofactor of the starting matrix. If we develop any of this determinant, we get a sum of 9 determinants where each contains the determinant of a matrix of dimension $n=8$. At this level, there is therefore 90 determinants of matrices of dimension 8 to calculate. The process could continue until it remains only determinant of order 2. And therefore we guess that the number of determinants of order 2 is important!

P10. As we will prove it during our study of the spectral theorem further below at page 2004, for a positive definite matrix M we have:

$$\det(M)=\prod_i \lambda_i>0 \tag{6.6.231}$$

where the λ_i are the eigenvalues of the matrix M. And for a definite semi-positive matrix M we have:

$$\det(M)=\prod_i \lambda_i \geq 0 \tag{6.6.232}$$

Definition (#6.6.260): Given m,n two positive integers and A a $m\times n$ matrix with coefficients in $\mathbb{C}$. For any $k\leq \min(m,n)$ a "minor of order k" is a determinant of A of the type:

$$\det\begin{bmatrix} a_{i_1,j_1} & \cdots & a_{i_1,j_k} \\ \vdots & \ddots & \vdots \\ a_{i_k,j_1} & \cdots & a_{i_k,j_k} \end{bmatrix} \tag{6.6.233}$$

with $1\leq i_1<\ldots\leq m$ and $1\leq j_1<\ldots<j_k\leq n$.

In the particular case of a matrix of order $n>1$ the definition is simpler: the minor M_{ij} of the element a_{ij} is the determinant of the matrix of order $n-1$ that we get by remove the row i and the column j. Therefore, to calculate the minor of an element, we remove the line and the column to which the element belongs to, and we calculate the determinant of the remaining square matrix.

Remark

The Leibniz formula of determinants is not the way you computer libraries calculate determinants. It is more of a theoretical tool, in fact it is probably one of the slowest ways to calculate the determinant. Gaussian Elimination (see section Numerical Methods page 6305) is a fast and relatively simple way to do such a calculation!

6.6.3.4.1 Derivative of a Determinant relatively to a parameter

Let us see now a result that will be quite useful to us in the section General Relativity:

Given a square matrix $n\times n$ with functions $g_{ij}:\mathbb{R}\mapsto\mathbb{R}$ that can be derivate at least one time. Let us put $g:=\det(G)$ with $G=(g_{ij})$. We want to calculate $\mathrm{d}_t g$. Given g_i the i-th column vector of the matrix G. Let us use the formula:

$$g=\det(g_1,\ldots,g_n)=\sum_{\sigma\in\mathcal{S}_n}\varepsilon(\sigma)g_{\sigma_{\sigma(1),1}}\cdot\ldots\cdot g_{\sigma(n),n} \tag{6.6.234}$$

Knowing that the derivative of $g_{\sigma(1),1}\cdot\ldots\cdot g_{\sigma(n),n}$ is (derivative of n products):

$$\begin{aligned}&\frac{\mathrm{d}g_{\sigma(1),1}}{\mathrm{d}t}\cdot g_{\sigma(2),2}\cdots g_{\sigma(n),n}+g_{\sigma(1),1}\cdot\frac{\mathrm{d}g_{\sigma(2),2}}{\mathrm{d}t}\cdot g_{\sigma(3),3}\cdot\ldots\cdot g_{\sigma(n),n}+\ldots\\&\quad+g_{\sigma(1),1}\cdot\ldots\cdot g_{\sigma(n-1),n-1}\cdot\frac{\mathrm{d}g_{\sigma(n),n}}{\mathrm{d}t}\end{aligned} \tag{6.6.235}$$

Therefore we have:

$$\begin{aligned}\frac{\mathrm{d}g}{\mathrm{d}t}=&\sum_{\sigma\in\mathcal{S}_n}\varepsilon(\sigma)\cdot\frac{\mathrm{d}g_{\sigma(1),1}}{\mathrm{d}t}\cdot g_{\sigma(2),2}\cdot\ldots\cdot g_{\sigma(n),n}+\sum_{\sigma\in\mathcal{S}_n}\varepsilon(\sigma)\cdot g_{\sigma(1),1}\cdot\frac{\mathrm{d}g_{\sigma(2),2}}{\mathrm{d}t}\cdot g_{\sigma(3),3}\cdot\ldots\cdot g_{\sigma(n),n}\\&+\ldots+\sum_{\sigma\in\mathcal{S}_n}\varepsilon(\sigma)\cdot g_{\sigma(1),1}\cdot\ldots\cdot g_{\sigma(n-1),n-1}\cdot\frac{\mathrm{d}g_{\sigma(n),n}}{\mathrm{d}t}\end{aligned}\tag{6.6.236}$$

If we take a look closely to the first above sum, we notice that:

$$\sum_{\sigma\in S_n}\varepsilon(\sigma)\cdot\frac{\mathrm{d}g_{\sigma(1),1}}{\mathrm{d}t}\cdot g_{\sigma(2),2}\cdot\ldots\cdot g_{\sigma(n),n}=\det(g'_1,g_2,\ldots,g_n)\tag{6.6.237}$$

where g'_1 is the derivative of the vector g_1. Same for the following sums. Therefore:

$$\frac{\mathrm{d}g}{\mathrm{d}t}=\det(g'_1,g_2,\ldots,g_n)+\det(g_1,g'_2,\ldots,g_n)+\ldots+\det(g_1,g_2,\ldots,g'_n)\tag{6.6.238}$$

Let us develop again. Let us consider the term $\det(g'_1,g_2,\ldots,g_n)$ above. If we develop it relatively to the first column, we get:

$$\det(g'_1,g_2,\ldots,g_n)=\sum_{i=1}^{n}(-1)^{1+i}\cdot\frac{\mathrm{d}g_{i,1}}{\mathrm{d}t}\cdot\det(G_{i,1})\tag{6.6.239}$$

Also, by developing the j-term of the above sum relatively to the j-th column, we get:

$$\det(g_1,\ldots,g'_j,\ldots,g_n)=\sum_{i=1}^{n}(-1)^{j+i}\cdot\frac{\mathrm{d}g_{i,j}}{\mathrm{d}t}\cdot\det(G_{i,j})\tag{6.6.240}$$

If we put:

$$\alpha_{ij}:=(-1)^{i+j}\cdot\det(G_{i,j})\tag{6.6.241}$$

We get:

$$\frac{\mathrm{d}g}{\mathrm{d}t}=\sum_{j=1}^{n}\sum_{i=1}^{n}\alpha_{ij}\cdot\frac{\mathrm{d}g_{i,j}}{\mathrm{d}t}\tag{6.6.242}$$

Which is written in tensor notation (see section Tensor Calculus page 2039):

$$\frac{\mathrm{d}g}{\mathrm{d}t}=\alpha_{ij}\cdot\frac{\mathrm{d}g_{i,j}}{\mathrm{d}t}\tag{6.6.243}$$

We also have:

$$b_{ji}=\frac{1}{\det(G)}\cdot(-1)^{i+j}\cdot\det(G_{i,j})=\frac{\alpha_{ij}}{\det(G)}=\frac{\alpha_{ij}}{g}\tag{6.6.244}$$

where b_{ji} is the coefficient being at the row j-th, column i-th of the matrix G^{-1}. Let us change the notation such that:

$$g^{ij}=b_{ji}=\frac{\alpha_{ij}}{g}\tag{6.6.245}$$

The expression of the derivative is then finally:

$$\frac{\mathrm{d}g}{\mathrm{d}t}=g\cdot\sum_{j=1}^{n}\sum_{i=1}^{n}g^{ij}\cdot\frac{\mathrm{d}g_{i,j}}{\mathrm{d}t} \tag{6.6.246}$$

which is written in tensor notation:

$$\frac{\mathrm{d}g}{\mathrm{d}t}=g\cdot g^{ij}\cdot\frac{\mathrm{d}g_{i,j}}{\mathrm{d}t} \tag{6.6.247}$$

This result, finally quite simple, we will be helpful to us in the section of Tensor Calculus to build the tools necessary for the study of General Relativity and in the context of the determination of Einstein field equations. It is therefore appropriate to remember it!

6.6.3.4.2 Derivative of logarithm of a determinant

Let $M\in\mathbb{R}^{n\times n}$ be a square matrix. For a function $f:\mathbb{R}^{n\times n}\mapsto\mathbb{R}$, define its derivative f' as an $n\times n$ matrix where the entry in row i and column j is $\partial f/\partial m_{ij}$.

For some functions f, the derivative f' has a nice form. We will show now a matrix derivation that we will meet in the section of Numerical Methods during our study of Gaussian Mixture Model and Factor Analysis:

$$[\log(\det(M))]'=\frac{\mathrm{d}\log(\det(M))}{\mathrm{d}M}=(M^{-1})^T \tag{6.6.248}$$

Here, we restrict the domain of the function to M with positive determinant (typically the case of variance-covariance matrices!). Notice that it is not intuitive that the derivative of a scalar function of a matrix is equal to a matrix...

Before we get move to the proof, we need to recall some terms. For a matrix M:

- The (i,j) minor of M, denoted $\det(M_{ij})$, is the determinant of the $(n-1)\times(n-1)$ matrix that remains after removing the ith row and jth column from M
- The cofactor matrix of M, denoted C, is an $n\times n$ matrix such that $C_{ij}=(-1)^{i+j}\det(M_{ij})$
- The adjugate matrix of M, denoted $\text{adj}(M)=C^T=((-1)^{i+j}\det(M_{ij}))^T_{1\leq i,j\leq n}$, is simply the transpose of C.

These terms are useful because they related to both matrix determinants and inverses. If M is invertible, then as we have proved earlier above:

$$M^{-1}=\frac{1}{\det(M)}\text{adj}(M)=\frac{1}{\det(M)}C^T \tag{6.6.249}$$

so:

$$\left(M^{-1}\right)^T_{ij}=\frac{1}{\det(M)}C_{ij} \tag{6.6.250}$$

On the other hand, by the cofactor expansion theorem of the determinant proved earlier above (as for a symmetric matrix $C=C^T$):

$$\det(M)=\sum_{j=1}^{n}(-1)^{i+j}\cdot m_{ij}\cdot\det(M_{ij})=\sum_{j=1}^{n}m_{ij}C_{ij} \tag{6.6.251}$$

So by the derivative product rule:

$$\frac{\partial(\det(M))}{\partial m_{ij}}=\sum_{j=1}^{n}\frac{\partial m_{ij}}{\partial m_{ik}}C_{ij}+m_{ij}\frac{\partial C_{ij}}{\partial m_{ik}} \tag{6.6.252}$$

If $k\neq j$, then $\frac{\partial m_{ij}}{\partial X_{ik}}=0$, otherwise it is equal to 1. This means that the first term above reduces to C_{ik}. For any k, the elements of M which affect C_{ij} are those which do not lie on row i or column j. Hence, $\partial C_{ij}/\partial m_{ik}=0$ for all j! Therefore:

$$\frac{\partial(\det(M))}{\partial m_{ij}}=C_{ij} \tag{6.6.253}$$

Putting all this together with an application of the derivative chain rule, we get (using the for the last equality the relation 6.6.3.4.2):

$$[\log(\det(M))]'_{ij}=\frac{1}{\det(M)}\frac{\partial(\det(M))}{\partial m_{ij}}=\frac{1}{\det(M)}C_{ij}=\left(M^{-1}\right)^T_{ij} \tag{6.6.254}$$

as required!

6.6.3.4.3 Cramer's rule

Let us first solve a simple general 2 by 2 linear system using substitution to see how Cramer's rule out pops (this will be considered as an informal proof obviously!).

For this purpose we start with the following system (don't forget that such a system is traditionally written $A\vec{x}=\vec{b}$):

$$\begin{aligned}a_1x+b_1y&=c_1\\a_2x+b_2y&=c_2\end{aligned} \tag{6.6.255}$$

Multiplying both sides of the first equation by a_2, and both sides of the second equation by a_1, then subtracting, we find that:

$$(a_1b_2-b_1a_2)y=a_1c_2-c_1a_2 \tag{6.6.256}$$

Assuming that $a_1b_2-b_1a_2$ is not 0, we find that:

$$y=\frac{a_1c_2-c_1a_2}{a_1b_2-b_1a_2}=\frac{\begin{vmatrix}a_1&c_1\\a_2&c_2\end{vmatrix}}{\begin{vmatrix}a_1&b_1\\a_2&b_2\end{vmatrix}} \tag{6.6.257}$$

The formula for x can be derived similarly so that we have:

$$x=\frac{\begin{vmatrix}c_1 & b_1\\ c_2 & b_2\end{vmatrix}}{\begin{vmatrix}a_1 & b_1\\ a_2 & b_2\end{vmatrix}}=\frac{c_1b_2-b_1c_2}{a_1b_2-b_1a_2},\qquad y=\frac{\begin{vmatrix}a_1 & c_1\\ a_2 & c_2\end{vmatrix}}{\begin{vmatrix}a_1 & b_1\\ a_2 & b_2\end{vmatrix}}=\frac{a_1c_2-c_1a_2}{a_1b_2-b_1a_2} \tag{6.6.258}$$

If we denote x by x_1 and y by x_2, then we see that we have:

$$x_k=\frac{\det(A_k)}{\det(A)} \tag{6.6.259}$$

where we define A_k to be the $n\times n$ matrix obtained by replacing the k-th column of A by the inhomogeneous term $\vec{b}$.

The rules for 3×3 matrices are similar. Given:

$$\begin{aligned}a_1x+b_1y+c_1z&=d_1\\ a_2x+b_2y+c_2z&=d_2\\ a_3x+b_3y+c_3z&=d_3\end{aligned} \tag{6.6.260}$$

which in matrix notation is:

$$\begin{bmatrix}a_1 & b_1 & c_1\\ a_2 & b_2 & c_2\\ a_3 & b_3 & c_3\end{bmatrix}\begin{pmatrix}x\\ y\\ z\end{pmatrix}=\begin{pmatrix}d_1\\ d_2\\ d_3\end{pmatrix} \tag{6.6.261}$$

Then the values of x, y and z (after some boring algebra) can be found as follows:

$$x=\frac{\begin{vmatrix}d_1 & b_1 & c_1\\ d_2 & b_2 & c_2\\ d_3 & b_3 & c_3\end{vmatrix}}{\begin{vmatrix}a_1 & b_1 & c_1\\ a_2 & b_2 & c_2\\ a_3 & b_3 & c_3\end{vmatrix}},\quad y=\frac{\begin{vmatrix}a_1 & d_1 & c_1\\ a_2 & d_2 & c_2\\ a_3 & d_3 & c_3\end{vmatrix}}{\begin{vmatrix}a_1 & b_1 & c_1\\ a_2 & b_2 & c_2\\ a_3 & b_3 & c_3\end{vmatrix}},\quad z=\frac{\begin{vmatrix}a_1 & b_1 & d_1\\ a_2 & b_2 & d_2\\ a_3 & b_3 & d_3\end{vmatrix}}{\begin{vmatrix}a_1 & b_1 & c_1\\ a_2 & b_2 & c_2\\ a_3 & b_3 & c_3\end{vmatrix}} \tag{6.6.262}$$

Then denoting again x by x_1, y by x_2 and z by x_3 we have again (don't forget that such a system is traditionally written $A\vec{x}=\vec{b}$):

$$x_k=\frac{\det(A_k)}{\det(A)} \tag{6.6.263}$$

where we define A_k to be the $n\times n$ matrix obtained by replacing the k-th column of A by the inhomogeneous term $\vec{b}$.

But now let us deal with the general proof. But before... some readers may argue that it is useless! In fact not really! They are some quite important applications in Tensor Calculus (for example the divergence of a tensor field that has some application in advanced General Relativity[25]), or in computing derivatives implicitly[26]. But this are all quite abstract applications. Indeed! So the reader has to know that the most well known practical application is the use in forecasting and times series analysis in business (more generally in any advanced finance field) for the study of the partial autocorrelation coefficient of autoregressive process (through the Yule-Walker equations).

---

[25] Don't look for it in this textbook!

[26] Actually also not needed to read this textbook entirely!

Theorem 6.40. If $A\vec{x}=\vec{b}$ is a linear system of equations with $\vec{x}=[x_1\,x_2\,\ldots\,x_n]^T$ and $A\in\mathbb{R}^{n\times n}$ such that $\det(A)\neq 0$ then we find the solutions:

$$x_i=\frac{\det(A_i)}{\det(A)} \tag{6.6.264}$$

where we define A_i to be the $n\times n$ matrix obtained by replacing the i-th column of A by the inhomogeneous term $\vec{b}$.

As there is a lot of different ways to make the proof of that theorem, the best an easiest one according to us is that provided by PlanetMath.org and written by someone under the pseudonym `rmilson` the 12013-03-22 (holocene calendar) for which we provide here a simple copy paste (as everything seems perfect to us and it's at the same time a very clever way to do the proof...!):

Proof 6.40.1. Since we assume that $\det(A)\neq 0$, by the properties of the determinant we know that A is invertible.

We claim that this implies that the equation $A\vec{x}=\vec{b}$ has a unique solution. Note that $A^{-1}\vec{b}$ is a solution since:

$$A(A^{-1}\vec{b})=(AA^{-1})\vec{b}=\vec{b} \tag{6.6.265}$$

so we know that a solution exists.

Let $\vec{s}$ be an arbitrary solution to the equation, so $A\vec{s}=\vec{b}$. But then:

$$\vec{s}=(A^{-1}A)\vec{s}=A^{-1}(A\vec{s})=A^{-1}\vec{b} \tag{6.6.266}$$

so we see that $A^{-1}\vec{b}$ is the only solution.

For each integer i, $1\leq i\leq n$, let a_i denote the ith column of A, let e_i denote the ith column of the identity matrix I_n, and let X_i denote the matrix obtained from $\mathbb{1}_n$ by replacing column i with the column vector $\vec{x}$.

We know that for any matrices A, B that the kth column of the product AB is simply the product of A and the kth column of B. Also observe that:

$$Ae_k=a_k \tag{6.6.267}$$

for $k=1\ldots n$. Thus, by multiplication, we have:

$$\begin{aligned}AX_i&=A(\vec{e}_1,\ldots,\vec{e}_{i-1},\vec{x},\vec{e}_{i+1},\ldots,\vec{e}_n)\\&=(A\vec{e}_1,\ldots,A\vec{e}_{i-1},A\vec{x},A\vec{e}_{i+1},\ldots,A\vec{e}_n)\\&=M_i\end{aligned} \tag{6.6.268}$$

Since X_i is $\mathbb{1}_n$ with column i replaced with $\vec{x}$, we will compute the determinant of X_i with the cofactor expansion method. For this, let us recall the cofactor expansion relation that we have just proved earlier and that was for recall:

$$\det(A_i)=\sum_{j=1}^{n}(-1)^{i+j}a_{ij}\det(A_{ij}) \tag{6.6.269}$$

can be used to calculated the determinant only for the cofactor (row) $j=i$ such that for X_i:

$$\det(X_i)=\sum_{j=1}^{n}(-1)^{i+j}x_{ij}\det(X_{ij}) \tag{6.6.270}$$

If we choose to calculate this sum only for a given row i, we know that $x_{ij}\neq 0$ only for $i=j$, then the previous relation reduce to:

$$\det(X_i)=(-1)^{i+i}x_{ii}\det(X_{ii})=(-1)^{2i}x_i\det(X_i)=1\cdot x_i\det(X_i)=x_i\det(\mathbb{1}_{n-1})=x_i\cdot 1=x_i \tag{6.6.271}$$

Thus by the multiplicative property of the determinant:

$$\det(M_i)=\det(AX_i)=\det(A)\det(X_i)=\det(A)x_i \tag{6.6.272}$$

we get:

$$x_i=\frac{\det(M_i)}{\det(A)} \tag{6.6.273}$$

and if we denote M_i by A_i we get the famous "Cramer's rule":

$$\boxed{x_i=\frac{\det(A_i)}{\det(A)}} \tag{6.6.274}$$

as required!

■ Q.E.D.

6.6.3.4.4 Determinant Cofactor and Matrix Inverse

Let us finish our study of the determinants with the "icing on the cake" by giving a very important relation in many fields of engineering, physics and mathematics that connects the inverse of the coefficients of a matrix with miners of order n (we will use this relation further below).

Given $A\in M_n(\mathbb{C})$ an invertible matrix (non singular). Let us write $A=(a_{ij})$ and $A^{-1}=(b_{ij})$. Then:

$$\boxed{b_{ij}=\frac{1}{\det(A)}\cdot(-1)^{i+j}\cdot\det(A_{ji})} \tag{6.6.275}$$

Proof 6.40.2. Let us denote by a_k the k-th column vector of the matrix A. Knowing that $A\cdot A^{-1}=\mathbb{1}$ (under known assumptions), we have (trivial):

$$e_j=\sum_{k=1}^{n}b_{kj}a_k \tag{6.6.276}$$

Let us calculate now $\det(a_1,\ldots,a_{k-1},e_j,a_{k+1},\ldots,a_n)$. First by developing relatively to the k-th column we found (as only one of the coefficient of e_j is not null and that the unique non-null one is equal to the unit):

$$\det(a_1,\ldots,a_{k-1},e_j,a_{k+1},\ldots,a_n)=(-1)^{j+k}\det(A_{jk}) \quad (6.6.277)$$

Furthermore (properties of the determinant):

$$\begin{aligned}&\det(a_1,\ldots,a_{k-1},e_j,a_{k+1},\ldots,a_n)=\det\left(1,\ldots,a_{k-1},\sum_{l=1}^{n}b_{lj}a_l,a_{k+1},\ldots,a_n\right)\\&=\sum_{l=1}^{n}b_{lj}\underbrace{\det(a_1,\ldots,a_{k-1},a_l,a_{k+1},\ldots,a_n)}_{=0\text{ if } l\neq k}=b_{kj}\det(A)\end{aligned} \quad (6.6.278)$$

Therefore:

$$(-1)^{j+k}\det(A_{jk})=b_{kj}\det(A) \quad (6.6.279)$$

That is to say:

$$\boxed{b_{kj}=\frac{1}{\det(A)}(-1)^{j+k}\det(A_{jk})} \quad (6.6.280)$$

That later relation is sometimes denoted:

$$A^{-1}=\frac{1}{\det(A)}\text{adj}(A) \quad (6.6.281)$$

where "adj" means adjacent.

■ Q.E.D.

✍**Examples:**

E1. Inverse of a general 2×2 matrix:

$$\begin{bmatrix}a & b\\ c & d\end{bmatrix}^{-1}=\frac{1}{ad-bc}\begin{bmatrix}d & -b\\ -c & a\end{bmatrix} \quad (6.6.282)$$

We want to compute the inverse of the following 2×2 matrix:

$$A=\begin{bmatrix}2 & 4\\ 3 & 5\end{bmatrix} \quad (6.6.283)$$

using the cofactor method proved above!

First we compute the determinant:

$$\det(A)=|A|=2\cdot 5-4\cdot 3=10-12=-2 \quad (6.6.284)$$

and after the cofactors! We start with the first cofactor (it's very easy in the case of a 2×2 matrix as the minor is just as scalar and therefore the determinant of the minor is equal to the scalar itself!):

$$\begin{bmatrix} 2 & 4 \\ 3 & 5 \end{bmatrix}$$

$$C_{11} = (-1)^{1+1} \cdot \det(A_{11}) = (-1)^2 \cdot 5 = 5$$

Let's move to the next cofactor:

$$\begin{bmatrix} 2 & 4 \\ 3 & 5 \end{bmatrix}$$

$$C_{12} = (-1)^{1+2} \cdot \det(A_{21}) = (-1)^3 \cdot 3 = -3$$

Let's move again to the next cofactor:

$$\begin{bmatrix} 2 & 4 \\ 3 & 5 \end{bmatrix}$$

$$C_{21} = (-1)^{2+1} \cdot \det(A_{12}) = (-1)^3 \cdot 4 = -4$$

And now the last cofactor:

$$\begin{bmatrix} 2 & 4 \\ 3 & 5 \end{bmatrix}$$

$$C_{22} = (-1)^{2+2} \cdot \det(A_{22}) = (-1)^4 \cdot 2 = 2$$

To summarize all cofactors:

$$\begin{aligned} C_{11} &= (-1)^2 \cdot 5 = 5 \\ C_{12} &= (-1)^3 \cdot 3 = -3 \\ C_{21} &= (-1)^3 \cdot 4 = -4 \\ C_{22} &= (-1)^4 \cdot 2 = 2 \end{aligned} \tag{6.6.285}$$

Finally as:

$$b_{ij} = \frac{1}{\det(A)} \cdot (-1)^{i+j} \cdot \det\left(A_{ji}\right) \tag{6.6.286}$$

We have:

$$B = A^{-1} = -\frac{1}{2} \cdot \begin{bmatrix} 5 & -4 \\ -3 & 2 \end{bmatrix} = \begin{bmatrix} -5/2 & 2 \\ 3/2 & -1 \end{bmatrix} = \begin{bmatrix} -2.5 & 2 \\ 1.5 & -1 \end{bmatrix} \tag{6.6.287}$$

We see therefore that the cofactor matrix inversion method has an algorithmic complexity of order $\mathcal{O}(n!)$. It's a very expensive approach!

E2. Inverse of a general 3×3 matrix:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}^{-1} = \frac{\begin{bmatrix} a_{22}a_{33}-a_{23}a_{32} & a_{13}a_{32}-a_{12}a_{33} & a_{12}a_{23}-a_{13}a_{22} \\ a_{23}a_{31}-a_{21}a_{33} & a_{11}a_{33}-a_{13}a_{31} & a_{13}a_{21}-a_{11}a_{23} \\ a_{21}a_{32}-a_{22}a_{31} & a_{12}a_{31}-a_{11}a_{32} & a_{11}a_{22}-a_{12}a_{21} \end{bmatrix}}{a_{11}a_{22}a_{33}+a_{21}a_{32}a_{13}+a_{31}a_{12}a_{23}-a_{11}a_{32}a_{23}-a_{31}a_{22}a_{13}-a_{21}a_{12}a_{33}} \tag{6.6.288}$$

E3. Inverse of a general 4×4 matrix:

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} \tag{6.6.289}$$

and:

$$\begin{aligned} \det(A) = & a_{11}a_{22}a_{33}a_{44}+a_{11}a_{23}a_{34}a_{42}+a_{11}a_{24}a_{32}a_{43} \\ & +a_{12}a_{21}a_{34}a_{43}+a_{12}a_{23}a_{31}a_{44}+a_{12}a_{24}a_{33}a_{41} \\ & +a_{13}a_{21}a_{32}a_{44}+a_{13}a_{22}a_{34}a_{41}+a_{13}a_{24}a_{31}a_{42} \\ & +a_{14}a_{21}a_{33}a_{42}+a_{14}a_{22}a_{31}a_{43}+a_{14}a_{23}a_{32}a_{41} \\ & -a_{11}a_{22}a_{34}a_{43}-a_{11}a_{23}a_{32}a_{44}-a_{11}a_{24}a_{33}a_{42} \\ & -a_{12}a_{21}a_{33}a_{44}-a_{12}a_{23}a_{34}a_{41}-a_{12}a_{24}a_{31}a_{43} \\ & -a_{13}a_{21}a_{34}a_{42}-a_{13}a_{22}a_{31}a_{44}-a_{13}a_{24}a_{32}a_{41} \\ & -a_{14}a_{21}a_{32}a_{43}-a_{14}a_{22}a_{33}a_{41}-a_{14}a_{23}a_{31}a_{42} \end{aligned} \tag{6.6.290}$$

then there exists an inverse matrix of A, and it is:

$$A^{-1} = \frac{1}{\det(A)} \begin{pmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \\ b_{31} & b_{32} & b_{33} & b_{34} \\ b_{41} & b_{42} & b_{43} & b_{44} \end{pmatrix} \tag{6.6.291}$$

where:

$$\begin{aligned}
b_{11} &= a_{22}a_{33}a_{44}+a_{23}a_{34}a_{42}+a_{24}a_{32}a_{43}-a_{22}a_{34}a_{43}-a_{23}a_{32}a_{44}-a_{24}a_{33}a_{42}\\
b_{12} &= a_{12}a_{34}a_{43}+a_{13}a_{32}a_{44}+a_{14}a_{33}a_{42}-a_{12}a_{33}a_{44}-a_{13}a_{34}a_{42}-a_{14}a_{32}a_{43}\\
b_{13} &= a_{12}a_{23}a_{44}+a_{13}a_{24}a_{42}+a_{14}a_{22}a_{43}-a_{12}a_{24}a_{43}-a_{13}a_{22}a_{44}-a_{14}a_{23}a_{42}\\
b_{14} &= a_{12}a_{24}a_{33}+a_{13}a_{22}a_{34}+a_{14}a_{23}a_{32}-a_{12}a_{23}a_{34}-a_{13}a_{24}a_{32}-a_{14}a_{22}a_{33}\\
b_{21} &= a_{21}a_{34}a_{43}+a_{23}a_{31}a_{44}+a_{24}a_{33}a_{41}-a_{21}a_{33}a_{44}-a_{23}a_{34}a_{41}-a_{24}a_{31}a_{43}\\
b_{22} &= a_{11}a_{33}a_{44}+a_{13}a_{34}a_{41}+a_{14}a_{31}a_{43}-a_{11}a_{34}a_{43}-a_{13}a_{31}a_{44}-a_{14}a_{33}a_{41}\\
b_{23} &= a_{11}a_{24}a_{43}+a_{13}a_{21}a_{44}+a_{14}a_{23}a_{41}-a_{11}a_{23}a_{44}-a_{13}a_{24}a_{41}-a_{14}a_{21}a_{43}\\
b_{24} &= a_{11}a_{23}a_{34}+a_{13}a_{24}a_{31}+a_{14}a_{21}a_{33}-a_{11}a_{24}a_{33}-a_{13}a_{21}a_{34}-a_{14}a_{23}a_{31}\\
b_{31} &= a_{21}a_{32}a_{44}+a_{22}a_{34}a_{41}+a_{24}a_{31}a_{42}-a_{21}a_{34}a_{42}-a_{22}a_{31}a_{44}-a_{24}a_{32}a_{41}\\
b_{32} &= a_{11}a_{34}a_{42}+a_{12}a_{31}a_{44}+a_{14}a_{32}a_{41}-a_{11}a_{32}a_{44}-a_{12}a_{34}a_{41}-a_{14}a_{31}a_{42}\\
b_{33} &= a_{11}a_{22}a_{44}+a_{12}a_{24}a_{41}+a_{14}a_{21}a_{42}-a_{11}a_{24}a_{42}-a_{12}a_{21}a_{44}-a_{14}a_{22}a_{41}\\
b_{34} &= a_{11}a_{24}a_{32}+a_{12}a_{21}a_{34}+a_{14}a_{22}a_{31}-a_{11}a_{22}a_{34}-a_{12}a_{24}a_{31}-a_{14}a_{21}a_{32}\\
b_{41} &= a_{21}a_{33}a_{42}+a_{22}a_{31}a_{43}+a_{23}a_{32}a_{41}-a_{21}a_{32}a_{43}-a_{22}a_{33}a_{41}-a_{23}a_{31}a_{42}\\
b_{42} &= a_{11}a_{32}a_{43}+a_{12}a_{33}a_{41}+a_{13}a_{31}a_{42}-a_{11}a_{33}a_{42}-a_{12}a_{31}a_{43}-a_{13}a_{32}a_{41}\\
b_{43} &= a_{11}a_{23}a_{42}+a_{12}a_{21}a_{43}+a_{13}a_{22}a_{41}-a_{11}a_{22}a_{43}-a_{12}a_{23}a_{41}-a_{13}a_{21}a_{42}\\
b_{44} &= a_{11}a_{22}a_{33}+a_{12}a_{23}a_{31}+a_{13}a_{21}a_{32}-a_{11}a_{23}a_{32}-a_{12}a_{21}a_{33}-a_{13}a_{22}a_{31}
\end{aligned} \tag{6.6.292}$$

Here is the pseudocode of the cofactor matrix inversion method:

Algorithm 2: Matrix Inversion using Determinant and Cofactor Method

```
Input: Square Matrix $A$
Output: Inverted Matrix $A^{-1}$
Function MatrixInverse($A$):
    detA ← Determinant($A$)
    if detA = 0 then return "Matrix is not invertible"

    $n$ ← number of rows/columns in $A$ as global variable
    $C$ ← matrix of size $n \times n$ to store the inverse
    for $j \leftarrow 1$ to $n$ do
        for $i \leftarrow 1$ to $n$ do
            minorMat ← Minor($A$, $j$, $i$)
            $C[i][j] \leftarrow (-1)^{i+j}$ · Determinant(minorMat)/detA
        end
    end
    return $C$
Function Determinant($A$):
    if $n = 1$ then
        return $A[1][1]$
    else if $n = 2$ then
        return $A[1][1] \cdot A[2][2] - A[1][2] \cdot A[2][1]$
    detA ← 0
    for $j \leftarrow 1$ to $n$ do
        detA ← detA + $(-1)^{1+j} \cdot A[1][j]$ · Determinant(Minor ($A$, 1, $j$))
    end
    return detA
Function Minor($A$, row, col):
    $C$ ← create a new matrix of size $(n-1) \times (n-1)$
    $r, c \leftarrow 0$
    for $i \leftarrow 1$ to $n$ do
        if $i = row$ then jump iteration

        for $j \leftarrow 1$ to $n$ do
            if $j = col$ then jump iteration

            if $i < row$ then
                $r \leftarrow i$
            else if $i >= row$ then
                $r \leftarrow i - 1$
            if $j < col$ then
                $c \leftarrow j$
            else if $j >= col$ then
                $c \leftarrow j - 1$
            $C[r][c] \leftarrow A[i][j]$
        end
    end
    return $C$
```

For a simple, detailed, and important practical application in the industry (because otherwise in this entire book we will rarely inverse small matrices), the reader can refer to the section of Numerical Methods in the part concerning the multiple linear regression.

Let us also indicate the following important properties where A and B are both square invertible matrices $M_n(\mathbb{C})$ and $\lambda \in \mathbb{C}$ (the first should be obvious, the second has already been presented earlier but unproven and the third one is important for the proof of the variance inflation factor that we will prove in the section of Numerical Methods on page 6398):

$$\begin{gathered}(\lambda A)^{-1}=\lambda^{-1}A^{-1}\\(A^T)^{-1}=(A^{-1})^T\\(AB)^{-1}=B^{-1}A^{-1}\end{gathered}\tag{6.6.293}$$

Let us prove the last property using the property of associativity:

Proof 6.40.3.

$$(AB)(B^{-1}A^{-1})=A(BB^{-1})A^{-1}=AA^{-1}=I_n\tag{6.6.294}$$

Which proves that $B^{-1}A^{-1}$ is indeed the inverse of AB where I_n (also denoted $\mathbb{1}$ for recall) is a diagonal matrix (also square) of dimension n.

■ Q.E.D.

6.6.3.4.5 Determinant of block-diagonal or block-triangular matrices

For our study of Factor Analysis we will need to compute the determinant of two particular type of block-triangular matrices.

First we will need to compute the determinant of block-upper-triangular matrix of the form:

$$\Gamma=\begin{bmatrix}A & B\\ \mathbb{O} & D\end{bmatrix}\tag{6.6.295}$$

where A and D are square matrices. And secondly of a block-lower-triangular matrix of the form:

$$\Gamma=\begin{bmatrix}A & \mathbb{O}\\ C & D\end{bmatrix}\tag{6.6.296}$$

However to calculate the determinant of these matrices we should first prove how to calculate the determinant of a diagonal block matrix of the following type:

$$\Gamma=\begin{bmatrix}A & \mathbb{O}\\ \mathbb{O} & \mathbb{1}\end{bmatrix}\tag{6.6.297}$$

where $\mathbb{1}$ is in this special case a 1×1, that is, $\mathbb{1}=1$. Suppose A is $k\times k$. Then Γ is of dimension $(k+1)(k+1)$.

For the proof, we will use the relation derived earlier above:

$$\det(A)=\sum_{\sigma\in\mathcal{S}_n}\varepsilon(\sigma)\prod_{i=1}^{n}a_{i\sigma(i)}\tag{6.6.298}$$

Proof 6.40.4. So we need to compute:

$$\det(\Gamma)=\sum_{\sigma\in\mathcal{S}_{k+1}}\varepsilon(\sigma)\prod_{i=1}^{k+1}\Gamma_{i\sigma(i)} \tag{6.6.299}$$

where $\mathcal{S}$ is the set of all permutations of the first $k+1$ natural numbers.

In the special case above we can rewrite this as:

$$\det(\Gamma)=\sum_{\sigma\in\mathcal{S}_{k+1}}\varepsilon(\sigma)\underbrace{\Gamma_{(k+1)\sigma(k+1)}}_{=1}\underbrace{\prod_{i=1}^{k}\Gamma_{i\sigma(i)}}_{\prod_{i=1}^{k}A_{i\sigma(i)}}=\sum_{\sigma\in\mathcal{S}_{k}}\varepsilon(\sigma)\prod_{i=1}^{k}A_{i\sigma(i)}=\det(A) \tag{6.6.300}$$

Remark

Let us recall to help to understand what we just did that we are typically in a situation similar to the following below where $a_{33}=1, a_{31}=a_{13}=a_{23}=a_{32}=0$:

$$\begin{aligned}\det(A)&=\det\begin{pmatrix}a_{11}&a_{12}&a_{13}\\a_{21}&a_{22}&a_{23}\\a_{31}&a_{32}&a_{33}\end{pmatrix}=\sum_{\sigma\in\mathcal{S}_4}\varepsilon(\sigma)\prod_{i=1}^{4}a_{i\sigma(i)}\\&=a_{11}\det\begin{pmatrix}a_{22}&a_{23}\\a_{32}&a_{33}\end{pmatrix}-a_{12}\det\begin{pmatrix}a_{21}&a_{23}\\a_{31}&a_{33}\end{pmatrix}+a_{13}\det\begin{pmatrix}a_{21}&a_{22}\\a_{31}&a_{32}\end{pmatrix}\\&=a_{11}a_{22}a_{33}-a_{11}a_{23}a_{32}-a_{12}a_{21}a_{33}+a_{12}a_{23}a_{31}+a_{13}a_{21}a_{32}-a_{13}a_{22}a_{31}\end{aligned} \tag{6.6.301}$$

From the above formula, each term is in the form of $a_{1j_1}a_{2j_2}a_{3j_3}$ where the second indices j_1,j_2,j_3 is a permutation of $1,2,3$. For example, the first term has $1,2,3$ as the second indices; the second term has $1,3,2$; the third term has $2,1,3$, etc.

The sign of each term is the sign of the permutation. For example, the sign of $1,2,3$ is clearly 1. For the second term, $1,3,2$ is obtained from $1,2,3$ by one transposition $2,3\rightarrow 3,2$, so the sign is -1, etc.

We see obviously that in the case $a_{33}=1, a_{31}=a_{13}=a_{23}=a_{32}=0$, the above relation reduces to:

$$\begin{aligned}\det(A)&=\det\begin{pmatrix}a_{11}&a_{12}&0\\a_{21}&a_{22}&0\\0&0&1\end{pmatrix}=\sum_{\sigma\in\mathcal{S}_4}\varepsilon(\sigma)\prod_{i=1}^{4}a_{i\sigma(i)}\\&=a_{11}\det\begin{pmatrix}a_{22}&0\\0&1\end{pmatrix}-a_{12}\det\begin{pmatrix}a_{21}&0\\0&1\end{pmatrix}+0\det\begin{pmatrix}a_{21}&a_{22}\\0&0\end{pmatrix}\\&=a_{11}a_{22}\cdot 1-a_{11}\cdot 0\cdot 0-a_{12}a_{21}\cdot 1+a_{12}\cdot 0\cdot 0+0\cdot a_{21}\cdot 0-0\cdot a_{22}\cdot 0\\&=a_{11}a_{22}-a_{12}a_{21}\end{aligned} \tag{6.6.302}$$

The result for the case in which $\mathbb{1}$ is not 1×1 is proved recursively. For example, if $\mathbb{1}$ is 2×2,

we have:

$$\det\begin{bmatrix} A & \mathbb{O} \\ \mathbb{O} & \mathbb{1} \end{bmatrix} = \det\begin{bmatrix} \begin{bmatrix} A & 0 \\ 0 & 1 \end{bmatrix} & 0 \\ 0 & 1 \end{bmatrix} = \det\begin{bmatrix} A & 0 \\ 0 & 1 \end{bmatrix} = \det(A) \tag{6.6.303}$$

and analogously for larger dimensions.

■ Q.E.D.

The proof for the case in which:

$$\Gamma = \begin{bmatrix} \mathbb{1} & \mathbb{O} \\ \mathbb{O} & \mathbb{A} \end{bmatrix} \tag{6.6.304}$$

is similar to the one just provided.

Ok now that we have this result, let us prove the following determinant:

$$\det(\Gamma) = \det\begin{bmatrix} A & \mathbb{O} \\ C & D \end{bmatrix} = \det(A)\det(D) \tag{6.6.305}$$

Proof 6.40.5. Let us assume that A is $k \times k$ and D is $l \times l$, so that C is $l \times k$ and $\mathbb{O}$ is $k \times l$. In what follows, we will denote by $\mathbb{1}_k$ a $k \times k$ identity matrix and by $\mathbb{O}_{lk}$ an $l \times k$ zero matrix. Note that (in the purpose to decompose the above matrix Γ in product of triangular block matrices!):

$$\begin{aligned}
&\begin{bmatrix} A & \mathbb{O}_{kl} \\ \mathbb{O}_{lk} & \mathbb{1}_l \end{bmatrix} \begin{bmatrix} \mathbb{1}_k & \mathbb{O}_{kl} \\ C & \mathbb{1}_l \end{bmatrix} \begin{bmatrix} \mathbb{1}_k & \mathbb{O}_{kl} \\ \mathbb{O}_{lk} & D \end{bmatrix} \\
&= \begin{bmatrix} A\mathbb{1}_k + \mathbb{O}_{kl}C & A\mathbb{O}_{kl} + \mathbb{O}_{kl}\mathbb{1}_l \\ \mathbb{O}_{lk}\mathbb{1}_k + \mathbb{1}_l C & \mathbb{O}_{lk}\mathbb{O}_{kl} + \mathbb{1}_l\mathbb{1}_l \end{bmatrix} \begin{bmatrix} \mathbb{1}_k & \mathbb{O}_{kl} \\ \mathbb{O}_{lk} & D \end{bmatrix} \\
&= \begin{bmatrix} A & \mathbb{O}_{kl} \\ C & \mathbb{1}_l \end{bmatrix} \begin{bmatrix} \mathbb{1}_k & \mathbb{O}_{kl} \\ \mathbb{O}_{lk} & D \end{bmatrix} \\
&= \begin{bmatrix} A\mathbb{1}_k + \mathbb{O}_{kl}\mathbb{O}_{lk} & A\mathbb{O}_{kl} + \mathbb{O}_{kl}D \\ C\mathbb{1}_k + \mathbb{1}_l\mathbb{O}_{kl} & C\mathbb{O}_{kl} + \mathbb{1}_l D \end{bmatrix} \\
&= \begin{bmatrix} A & \mathbb{O} \\ C & D \end{bmatrix}
\end{aligned} \tag{6.6.306}$$

Thus, similarly to the previous proof with the diagonal block matrices and triangular block matrices:

$$\begin{aligned}
\det(\Gamma) = \det\begin{bmatrix} A & \mathbb{O} \\ C & D \end{bmatrix} &= \det\left(\begin{bmatrix} A & \mathbb{O}_{kl} \\ \mathbb{O}_{lk} & \mathbb{1}_l \end{bmatrix} \begin{bmatrix} \mathbb{1}_k & \mathbb{O}_{kl} \\ C & \mathbb{1}_l \end{bmatrix} \begin{bmatrix} \mathbb{1}_k & \mathbb{O}_{kl} \\ \mathbb{O}_{lk} & D \end{bmatrix}\right) \\
&= \det\left(\begin{bmatrix} A & \mathbb{O}_{kl} \\ \mathbb{O}_{lk} & \mathbb{1}_l \end{bmatrix}\right) \det\left(\begin{bmatrix} \mathbb{1}_k & \mathbb{O}_{kl} \\ C & \mathbb{1}_l \end{bmatrix}\right) \det\left(\begin{bmatrix} \mathbb{1}_k & \mathbb{O}_{kl} \\ \mathbb{O}_{lk} & D \end{bmatrix}\right) \\
&= \det(A)\det(\mathbb{1}_l)\det(\mathbb{1}_k)\det(\mathbb{1}_l)\det(\mathbb{1}_l)\det(D) \\
&= \det(A)\det(D)
\end{aligned} \tag{6.6.307}$$

■ Q.E.D.

6.6.3.4.6 Inverse and Determinant of a partitioned symmetric matrix

We will prove here a set of relations that we will need (once again...) for our study of Factor Analysis in the section of Statistics.

Let us divide an $n \times n$ symmetric matrix A into four blocks:

$$A=\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \tag{6.6.308}$$

The inverse matrix $B=A^{-1}$ can also be divided into four blocks:

$$B=A^{-1}=\begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \tag{6.6.309}$$

Here we assume the dimensionalities of these blocks are:

- A_{11} and B_{11} are $p \times p$
- A_{22} and B_{22} are $q \times q$
- $A_{12}=A_{21}^T$ and $B_{12}=B_{21}^T$ are $p \times q$

with $p+q=n$.

Then what we want to prove first (and after we will be able to focus on the inverse determinant problem) is that the components of the inverse of a partitioned symmetric matrix are given by:

$$\begin{aligned}
B_{11}&=\left(A_{11}-A_{12}A_{22}^{-1}A_{21}\right)^{-1}=A_{11}^{-1}+A_{11}^{-1}A_{12}\left(A_{22}-A_{21}A_{11}^{-1}A_{12}\right)^{-1}A_{21}A_{11}^{-1}\\
B_{22}&=\left(A_{22}-A_{21}A_{11}^{-1}A_{12}\right)^{-1}=A_{22}^{-1}+A_{22}^{-1}A_{21}\left(A_{11}-A_{12}A_{22}^{-1}A_{21}\right)^{-1}A_{12}A_{22}^{-1}\\
B_{21}&=-A_{22}^{-1}A_{21}\left(A_{11}-A_{12}A_{22}^{-1}A_{21}\right)^{-1}\\
B_{12}&=-A_{11}^{-1}A_{12}\left(A_{22}-A_{21}A_{11}^{-1}A_{12}\right)^{-1}
\end{aligned} \tag{6.6.310}$$

ie (it is in this form that we found it in most textbooks):

$$\boxed{B=A^{-1}=\begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}=\begin{bmatrix} \left(A_{11}-A_{12}A_{22}^{-1}A_{21}\right)^{-1} & -A_{11}^{-1}A_{12}\left(A_{22}-A_{21}A_{11}^{-1}A_{12}\right)^{-1} \\ -A_{22}^{-1}A_{21}\left(A_{11}-A_{12}A_{22}^{-1}A_{21}\right)^{-1} & \left(A_{22}-A_{21}A_{11}^{-1}A_{12}\right)^{-1} \end{bmatrix}} \tag{6.6.311}$$

Remark

The term $A_{11}-A_{12}A_{22}^{-1}A_{21}$ seems to be named the "Schur complement" of A_{22} in A. Same for the term $A_{22}-A_{21}A_{11}^{-1}A_{12}$ that seems also be named "Schur complement" of A_{11} in A.

Proof 6.40.6. We start with:

$$\begin{aligned}
\mathbb{1}_n=AA^{-1}=AB&=\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}\begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}\\
&=\begin{bmatrix} A_{11}B_{11}+A_{12}B_{21} & A_{11}B_{12}+A_{12}B_{22} \\ A_{21}B_{11}+A_{22}B_{21} & A_{21}B_{12}+A_{22}B_{22} \end{bmatrix}=\begin{bmatrix} \mathbb{1}_p & 0 \\ 0 & \mathbb{1}_q \end{bmatrix}
\end{aligned} \tag{6.6.312}$$

Let us equate each of the four blocks to get:

$$\begin{array}{lll} A_{11}B_{11}+A_{12}B_{21}=\mathbb{1}_p & \text{then} & B_{11}=A_{11}^{-1}-A_{11}^{-1}A_{12}B_{21} \\ A_{11}B_{12}+A_{12}B_{22}=\mathbb{0} & \text{then} & B_{12}=-A_{11}^{-1}A_{12}B_{22} \\ A_{21}B_{11}+A_{22}B_{21}=\mathbb{0} & \text{then} & B_{21}=-A_{22}^{-1}A_{21}B_{11} \\ A_{21}B_{12}+A_{22}B_{22}=\mathbb{1}_q & \text{then} & B_{22}=A_{22}^{-1}-A_{22}^{-1}A_{21}B_{12} \end{array} \tag{6.6.313}$$

Plug B_{21} into B_{11} to get:

$$B_{11}=A_{11}^{-1}+A_{11}^{-1}A_{12}A_{22}^{-1}A_{21}B_{11} \tag{6.6.314}$$

Solve for B_{11} to get:

$$B_{11}=\left(A_{11}-A_{12}A_{22}^{-1}A_{21}\right)^{-1} \tag{6.6.315}$$

Applying the Woodburry matrix identity (see below page 2035) given for recall by:

$$(A+UBV^T)^{-1}=A^{-1}-A^{-1}U(B^{-1}+V^TA^{-1}U)^{-1}V^TA^{-1} \tag{6.6.316}$$

to this expression we also get the other expression in the theorem, i.e.:

$$A_{11}^{-1}+A_{11}^{-1}A_{12}\left(A_{22}-A_{21}A_{11}^{-1}A_{12}\right)^{-1}A_{21}A_{11}^{-1} \tag{6.6.317}$$

Similarly we can get:

$$\begin{aligned} B_{22}&=\left(A_{22}-A_{21}A_{11}^{-1}A_{12}\right)^{-1} \\ B_{21}&=-A_{22}^{-1}A_{21}\left(A_{11}-A_{12}A_{22}^{-1}A_{21}\right)^{-1} \\ B_{12}&=-A_{11}^{-1}A_{12}\left(A_{22}-A_{21}A_{11}^{-1}A_{12}^{-1}\right)^{-1} \end{aligned} \tag{6.6.318}$$

■ Q.E.D.

Let us see now how to calculate the determinant of a partitioned symmetric matrix, given by:

$$\det(A)=\det\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22}\end{bmatrix}=\det(A_{22})\det(A_{11}-A_{12}A_{22}^{-1}A_{21})=\det(A_{11})\det(A_{22}-A_{21}A_{11}^{-1}A_{12}) \tag{6.6.319}$$

Proof 6.40.7. First, let us consider a matrix of the form:

$$\begin{bmatrix} A & B \\ C & D\end{bmatrix} \tag{6.6.320}$$

where A,B,C,D have size $n\times n$. We start from considering multiplication on the left by a matrix such that:

$$\begin{bmatrix} A^{-1} & \mathbb{0} \\ \mathbb{0} & \mathbb{1}\end{bmatrix}\begin{bmatrix} A & B \\ C & D\end{bmatrix}=\begin{bmatrix} \mathbb{1} & A^{-1}B \\ C & D\end{bmatrix} \tag{6.6.321}$$

Now let us consider the following left multiplication:

$$\begin{bmatrix} \mathbb{1} & \mathbb{0} \\ -C & \mathbb{1}\end{bmatrix}\begin{bmatrix} \mathbb{1} & A^{-1}B \\ C & D\end{bmatrix}=\begin{bmatrix} \mathbb{1} & A^{-1}B \\ \mathbb{0} & D-CA^{-1}B\end{bmatrix} \tag{6.6.322}$$

Therefore (starting from the beginning) in one line, that means we have:

$$\left(\begin{bmatrix}\mathbb{1} & \mathbb{0}\\ -C & \mathbb{1}\end{bmatrix}\begin{bmatrix}A^{-1} & \mathbb{0}\\ \mathbb{0} & \mathbb{1}\end{bmatrix}\right)\begin{bmatrix}A & B\\ C & D\end{bmatrix}=\begin{bmatrix}\mathbb{1} & A^{-1}B\\ \mathbb{0} & D-CA^{-1}B\end{bmatrix} \tag{6.6.323}$$

So that means after rearranging:

$$\begin{bmatrix}A & B\\ C & D\end{bmatrix}=\left(\begin{bmatrix}\mathbb{1} & \mathbb{0}\\ -C & \mathbb{1}\end{bmatrix}\begin{bmatrix}A^{-1} & \mathbb{0}\\ \mathbb{0} & \mathbb{1}\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbb{1} & A^{-1}B\\ \mathbb{0} & D-CA^{-1}B\end{bmatrix} \tag{6.6.324}$$

And it is therefore quite immediate that:

$$\left(\begin{bmatrix}\mathbb{1} & \mathbb{0}\\ -C & \mathbb{1}\end{bmatrix}\begin{bmatrix}A^{-1} & \mathbb{0}\\ \mathbb{0} & \mathbb{1}\end{bmatrix}\right)^{-1}=\begin{bmatrix}A & \mathbb{0}\\ C & \mathbb{1}\end{bmatrix} \tag{6.6.325}$$

So that finally:

$$\begin{bmatrix}A & B\\ C & D\end{bmatrix}=\begin{bmatrix}A & \mathbb{0}\\ C & \mathbb{1}\end{bmatrix}\begin{bmatrix}\mathbb{1} & A^{-1}B\\ \mathbb{0} & D-CA^{-1}B\end{bmatrix} \tag{6.6.326}$$

Similarly we get:

$$\begin{bmatrix}A & B\\ C & D\end{bmatrix}=\begin{bmatrix}\mathbb{1} & B\\ \mathbb{0} & D\end{bmatrix}\begin{bmatrix}A-BD^{-1} & \mathbb{0}\\ D^{-1}C & \mathbb{1}\end{bmatrix} \tag{6.6.327}$$

To summarize:

$$\boxed{\begin{bmatrix}A & B\\ C & D\end{bmatrix}=\begin{bmatrix}A & \mathbb{0}\\ C & \mathbb{1}\end{bmatrix}\begin{bmatrix}\mathbb{1} & A^{-1}B\\ \mathbb{0} & D-CA^{-1}B\end{bmatrix}=\begin{bmatrix}\mathbb{1} & B\\ \mathbb{0} & D\end{bmatrix}\begin{bmatrix}A-BD^{-1} & \mathbb{0}\\ D^{-1}C & \mathbb{1}\end{bmatrix}} \tag{6.6.328}$$

This is a result we needed before attacking the real deal!

Indeed, according to the result above, this means that we can write:

$$\begin{aligned}A=\begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix}&=\begin{bmatrix}A_{11} & 0\\ A_{21} & \mathbb{1}\end{bmatrix}\begin{bmatrix}\mathbb{1} & A_{11}^{-1}A_{12}\\ \mathbb{0} & A_{22}-A_{21}A_{11}^{-1}A_{12}\end{bmatrix}\\ &=\begin{bmatrix}\mathbb{1} & A_{12}\\ \mathbb{0} & A_{22}\end{bmatrix}\begin{bmatrix}A_{11}-A_{12}A_{22}^{-1}A_{21} & \mathbb{0}\\ A_{22}^{-1}A_{21} & \mathbb{1}\end{bmatrix}\end{aligned} \tag{6.6.329}$$

As seen above (page 1969), we have:

$$\boxed{\det(A\cdot B)=\det(A)\cdot\det(B)} \tag{6.6.330}$$

and (page 1986):

$$\det\begin{bmatrix}B & \mathbb{0}\\ X & C\end{bmatrix}=\det\begin{bmatrix}B & X\\ \mathbb{0} & C\end{bmatrix}=\det(B)\det(C) \tag{6.6.331}$$

Therefore the equality:

$$\det(A)=\det\begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix}=\det(A_{22})\det(A_{11}-A_{12}A_{22}^{-1}A_{21})=\det(A_{11})\det(A_{22}-A_{21}A_{11}^{-1}A_{12}) \tag{6.6.332}$$

is proved.

■ Q.E.D.

6.6.4 Change of Basis (frames)

A basis for a vector space of dimension n is a sequence of n vectors $(\vec{e}_1,\ldots,\vec{e}_n)$ with the property that every vector in the space can be expressed uniquely as a linear combination of the basis vectors (see section Vector Calculus page 1809). The matrix representations of operators are also determined by the chosen basis! Since it is often desirable to work with more than one basis for a vector space, it is of fundamental importance in linear algebra to be able to easily transform coordinate-wise representations of vectors and operators taken with respect to one basis to their equivalent representations with respect to another basis. Such a transformation is named a "change of basis".

Let us now suppose that we move from a frame $\mathcal{E}=(\vec{e}_1,\vec{e}_2,\ldots,\vec{e}_n)$ of a space V^n to another space $\mathcal{F}=(\vec{f}_1,\vec{f}_2,\ldots,\vec{f}_n)$ of this same space sharing the same origin O. Thus in two dimension:

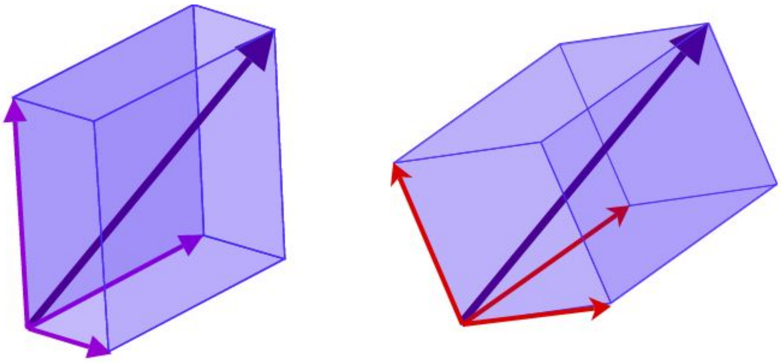


Figure 6.145: A vector represented in two different bases (purple and red arrows) (source: Wikipedia)

Let us decompose the $\vec{f}_i$ in the basis $\mathcal{E}$:

$$\begin{aligned}
\vec{f}_1 &= p_{11}\vec{e}_1+p_{21}\vec{e}_2+\ldots+p_{n1}\vec{e}_n\\
\vec{f}_2 &= p_{12}\vec{e}_1+p_{22}\vec{e}_2+\ldots+p_{n2}\vec{e}_n\\
\ldots\\
\vec{f}_n &= p_{1n}\vec{e}_1+p_{2n}\vec{e}_2+\ldots+p_{nn}\vec{e}_n
\end{aligned}\tag{6.6.333}$$

Definition (#6.6.261): We name "transition matrix" the matrix (linear application) that allows to pass from $\mathcal{E}\mapsto\mathcal{F}$ given by:

$$P=\begin{array}{c}\begin{pmatrix}p_{11} & p_{12} & \cdots & p_{1n}\\ p_{21} & \ddots & \vdots & p_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ p_{n1} & p_{n2} & \cdots & p_{nn}\end{pmatrix}\\ \underbrace{\hspace{5cm}}\\ \begin{array}{cccc}\vec{f}_1 & \vec{f}_1 & \cdots & \vec{f}_n\end{array}\end{array}\tag{6.6.334}$$

Theorem 6.41. Now let us consider the vector given by (such a notation will be more clear when the reader will deal with covariant and contravariant components in the section of Tensor Calculus at page 2059):

$$\vec{v}=x^i\vec{e}_i\tag{6.6.335}$$

So we intend to prove that the components of $y_1, y_2, \ldots, y_n$ of $\vec{v}$ in the basis $\mathcal{F}$ are given by:

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = P \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad X = PY \tag{6.6.336}$$

Thus explicitly:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = P^{-1} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad Y = P^{-1}X \tag{6.6.337}$$

Remark

The matrix P is invertible (non-singular), because its columns are linearly independent (they are the vectors $\vec{f}_i$ decomposed in the basis $\mathcal{E}$ and the $\vec{f}_i$ base and the $\vec{f}_i$ are linearly independent as they form a base!).

Proof 6.41.1. Let us take the case to simplify the case $n=2$ (the proof being quite easily generalized) with $\mathcal{E} = (\vec{e}_1, \vec{e}_2)$ and $\mathcal{F} = (\vec{f}_1, \vec{f}_2)$. Then we have:

$$\begin{aligned} \vec{f}_1 &= p_{11}\vec{e}_1 + p_{21}\vec{e}_2 \\ \vec{f}_2 &= p_{12}\vec{e}_1 + p_{22}\vec{e}_2 \end{aligned} \tag{6.6.338}$$

We therefore have $\vec{v} = x^i\vec{e}_i$ and we seek to express $\vec{v}$ in the basis $\mathcal{F}$ as $\vec{v} = y^i\vec{f}_i$. We'll search the linear application that link these two relation such that:

$$x^i\vec{e}_i = y^i\vec{f}_i \tag{6.6.339}$$

Thus written in an explicit way:

$$\begin{aligned} \vec{v} &= y_1\vec{f}_1 + y_2\vec{f}_2 = y_1(p_{11}\vec{e}_1 + p_{21}\vec{e}_2) + y_2(p_{12}\vec{e}_1 + p_{22}\vec{e}_2) \\ &= p_{11}y_1\vec{e}_1 + p_{21}y_1\vec{e}_2 + p_{12}y_2\vec{e}_1 + p_{22}y_2\vec{e}_2 = (p_{11}y_1 + p_{12}y_2)\vec{e}_1 + (p_{12}\vec{y}_1 + p_{22}y_2)\vec{e}_2 \\ &= x_1\vec{e}_1 + x_2\vec{e}_2 \end{aligned} \tag{6.6.340}$$

Therefore:

$$\begin{aligned} x_1 &= p_{11}y_1 + p_{12}y_2 \\ x_2 &= p_{21}y_1 + p_{22}y_2 \end{aligned} \tag{6.6.341}$$

That is to say:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \tag{6.6.342}$$

So P (if it exists) is indeed the matrix that can express the components of a vector of a basis in those of another basis such that we write in vector notation:

$$\vec{x} = P\vec{y} \tag{6.6.343}$$

■ Q.E.D.

Theorem 6.42. Let us now consider a linear application $g: V^n \mapsto V^n$. Let A be the matrix in the basis $\mathcal{E}$, and B its matrix in the basis $\mathcal{F}$ (of same dimension). Then we might have:

$$\boxed{B = P^{-1}AP} \tag{6.6.344}$$

which is equivalent:

$$BP = AP \tag{6.6.345}$$

or even:

$$A = PBP^{-1} \tag{6.6.346}$$

If there exists such a matrix P satisfying these relations, we say that A and B are "similar matrices".

Remark

Such basis changes are very useful in physics in the resolution of some systems of differential equations putted in matrix form. The change of basis can in some cases considerably simplify the resolution and by keeping at the same time the qualitative solution. The simplification is especially significant when the matrix A can be put into a diagonal form named "Weyr canonical form" and in another one named the (upper triangular) named "Jordan canonical form". It is important for theoretical reasons to know that one can always find the Jordan canonical form of a square matrix. It simplifies many abstract proofs to assume a matrix in the proof is in Jordan canonical form. The proof of that claim is actually out of the scoop of this textbook (especially because the Jordan canonical form is not well-suited to solve system of differential equation because of numerical instabilities).

Proof 6.42.1. Let us take back the fact that we proved that it was eventually possible to build a transition matrix P from the fact that:

$$\vec{v} = x^i\vec{e}_i = y^i\vec{f}_i \tag{6.6.347}$$

and let us put:

$$g(\vec{v}) = x'^i\vec{e}_i = y'^i\vec{f}_i \tag{6.6.348}$$

We have then a function that bring us to write:

$$\vec{x}\,' = A\vec{x} \qquad \vec{y}\,' = B\vec{y} \tag{6.6.349}$$

On the other hand, we have (that we proved earlier):

$$\vec{x} = P\vec{y} \qquad \vec{x}\,' = P\vec{y} \tag{6.6.350}$$

Therefore:

$$\vec{y}\,' = P^{-1}\vec{x}\,' = P^{-1}(A\vec{x}) = P^{-1}(A(P\vec{y})) = (P^{-1}AP)\vec{y} = B\vec{y} \tag{6.6.351}$$

hence:

$$B=P^{-1}AP \tag{6.6.352}$$

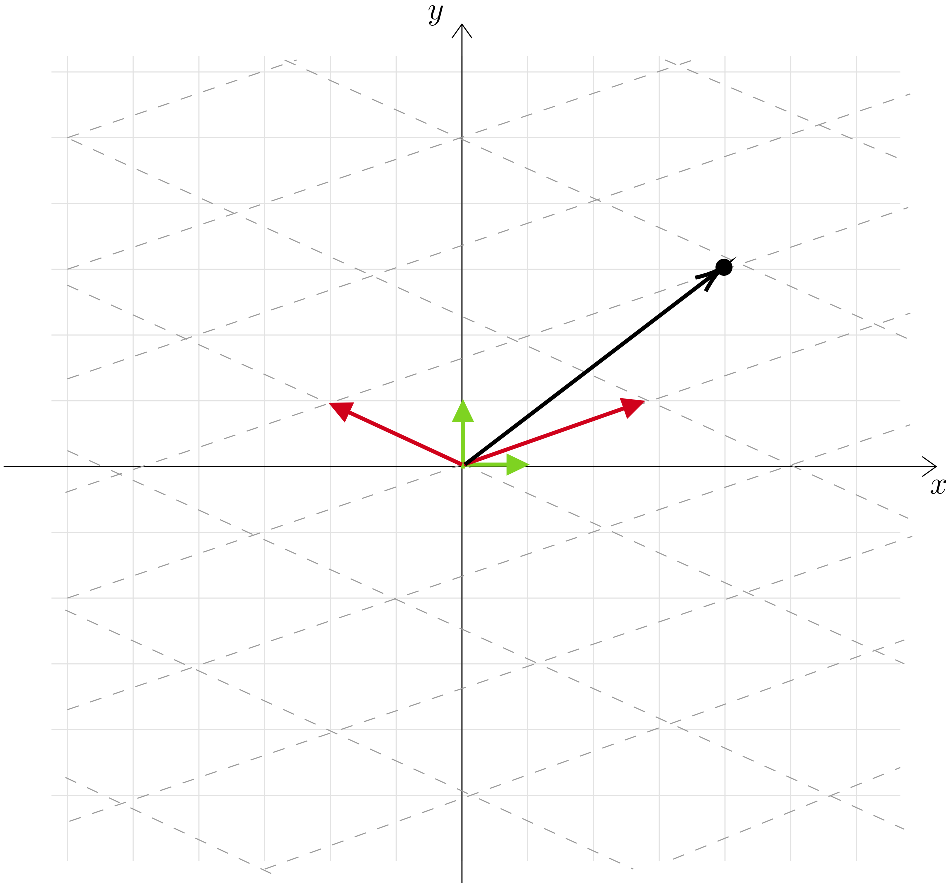
and as we saw it in our study of the determinant, the determinants of A, B are equal and therefore invariant. We will return later back on a similar formulation in our study of the Spectral Theorem below.

■ Q.E.D.

At the vocabulary level we say when we are in the presence of a such a matrix relation that: the matrix A is "conjugated" to the matrix B.

✍**Example:**

Let $B=\left\{\begin{pmatrix}1\\0\end{pmatrix},\begin{pmatrix}0\\1\end{pmatrix}\right\}$ in green in the figure below and $B'=\left\{\begin{pmatrix}3\\1\end{pmatrix},\begin{pmatrix}-2\\1\end{pmatrix}\right\}$ in red in the figure below:



The change of basis matrix form B' to B is:

$$P=\begin{bmatrix}3 & -2\\1 & 1\end{bmatrix}$$

The vector $\vec{v}$ with coordinates $\vec{v}_{B'} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ relative to the basis B' has coordinates:

$$\vec{v}_B = \begin{bmatrix} 3 & -2 \\ 1 & 1 \end{bmatrix} \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 4 \\ 3 \end{pmatrix}$$

relative to the basis B. Since:

$$P^{-1} = \begin{bmatrix} \frac{1}{5} & \frac{2}{5} \\ -\frac{1}{5} & \frac{3}{5} \end{bmatrix}$$

we can check that:

$$\vec{v}_{B'} = \begin{bmatrix} \frac{1}{5} & \frac{2}{5} \\ -\frac{1}{5} & \frac{3}{5} \end{bmatrix} \begin{pmatrix} 4 \\ 3 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

which is what we started with.

6.6.5 Eigenvalues and Eigenvectors

Definition (#6.6.262): An "eigenvalue" is by definition (we will find again this definition in the introduction to quantum algebra in the section of Wave Quantum Physics) a value λ belonging to a field $\mathbb{K}$ such that given a squared matrix $A\in M_{mm}(\mathbb{K})$ we have:

$$\exists\vec{X}\in M_{m1}(\mathbb{K})\quad \vec{X}\neq\vec{0}\text{ and }A\vec{X}=\lambda\vec{X} \tag{6.6.353}$$

and conversely a vector $\vec{X}\in M_{m1}(\mathbb{K})$ is an "eigenvector" if and only if:

$$\vec{X}\neq\vec{0},\exists\lambda\in\mathbb{K},A\vec{X}=\lambda\vec{X} \tag{6.6.354}$$

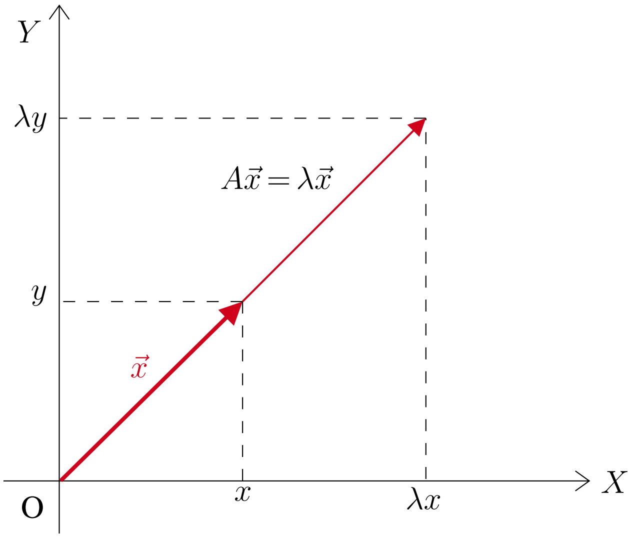


Figure 6.146: **Geometrical interpretation of eigenvector and eigenvalue in $\mathbb{R}^2$**

The major advantage of these concepts will be able the possibility to study a linear application, or any other item linked to a matrix representation, in a simple representation through a basis change on which the restriction of A is a single homothetic transformation (typically solving simple systems of differential equations).

Remark

This definition can be generalized to functional analysis (therefore it does not only applied to Matrix Algebra) too as by defining the "spectrum of an endomorphism" as the set:

$$\text{Sp}(f)=\{\lambda\in\mathbb{K},\exists x\in E\backslash\{0\}:f(x)=\lambda x\} \tag{6.6.355}$$

In other words: when a transformation (application of a matrix) acts on a vector, it modifies the direction of this vector except for certain particular matrices which have eigenvalues!

Thus, all the eigenvalues of a matrix $A\in M_{mm}(\mathbb{K})$ is named "spectrum of A" and satisfies the following homogeneous system named a "matrix-eigenvalue equation":

$$\boxed{(A-\lambda I_n)\vec{X}=\vec{0}} \tag{6.6.356}$$

or (whatever it is the same!):

$$(\lambda I_n - A)\vec{X} = \vec{0} \tag{6.6.357}$$

where I_n (also sometimes denoted by the symbol: $\mathbb{1}$) is for recall a diagonal unit matrix (and therefore also square) of dimension n. This system we know (proved above) has non-trivial solutions, therefore $\vec{X} \neq \vec{0}$ or $(\lambda I_n - A) \neq \vec{0}$, if and only if (we'll see many examples in various section related to physics in this textbook):

$$\det(A - \lambda I_n) = \begin{vmatrix} a_{11}-\lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22}-\lambda & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn}-\lambda \end{vmatrix} = 0 \tag{6.6.358}$$

that is to say that the matrix $A - \lambda I_n$ is not invertible (singular).

The determinant $\det(A - \lambda I_n)$ is a polynomial on λ of degree n and can have at maximum n solutions/eigenvalues as we have proved it in our study of polynomials (see section Calculus page 1368) and is named "characteristic polynomial" of A and the equation $\det(A - \lambda I_n)$ is named "characteristic equation" of A or "eigenvalues equation".

For the small parenthesis, it is nice to notice that we always have in the development of $\det(A - \lambda I_n)$ the trace of the matrix $\text{tr}(A)$ and the determinant $\det(A)$ that appear. Let us see two examples of this:

✍**Examples:**

E1. Let us begin with the case $n = 2$:

$$\det(A - \lambda I_2) = \begin{vmatrix} a_{11}-\lambda & a_{12} \\ a_{21} & a_{22}-\lambda \end{vmatrix} = \lambda^2 - \underbrace{(a_{11}+a_{22})}_{\text{tr(A)}}\lambda + \underbrace{(a_{11}a_{22}-a_{21}a_{12})}_{\text{det(A)}} \tag{6.6.359}$$

Therefore for a square matrix of dimension 2, the eigenvalues are (simple resolution of a polynomial of the 2nd degree):

$$\begin{aligned} \lambda_1 &= \frac{a_{11}+a_{22}+\sqrt{(a_{11}-a_{22})^2+4(a_{11}a_{22}-a_{12}a_{21}}}{2} = \frac{\text{tr}(A)+\sqrt{(\text{tr}(A)^2+4\det(A)}}{2} \\ \lambda_2 &= \frac{a_{11}+a_{22}-\sqrt{(a_{11}-a_{22})^2+4a_{12}a_{21}}}{2} = \frac{\text{tr}(A)-\sqrt{\text{tr}(A)^2+4\det(A)}}{2} \end{aligned} \tag{6.6.360}$$

E2. For a matrix of dimension $n = 3$, we have:

$$\begin{aligned} \det(A - \lambda I_3) = \begin{vmatrix} a_{11}-\lambda & a_{22} & a_{13} \\ a_{21} & a_{22}-\lambda & a_{23} \\ a_{31} & a_{32} & a_{33}-\lambda \end{vmatrix} &= -\lambda^3 + \underbrace{(a_{11}+a_{22}+a_{33})}_{\text{tr(A)}}\lambda^2 \\ &- (a_{11}a_{33}+a_{22}a_{33}+a_{11}a_{22}-a_{32}a_{23}-a_{12}a_{21}-a_{31}a_{13})\lambda \\ &+ \underbrace{a_{11}a_{22}a_{33}-a_{11}a_{32}a_{23}-a_{12}a_{21}a_{33}+a_{31}a_{32}a_{12}+a_{13}a_{21}a_{32}-a_{13}a_{31}a_{22}}_{\det(A)} \end{aligned} \tag{6.6.361}$$

and here... the final solution (roots) are quite less easy in the general case...

On the path let us notice (we will generalize the result coming from this during our study of the spectral theorem) that as multiplying the homogeneous system:

$$(\lambda I_n - A)\vec{X} = \vec{0} \tag{6.6.362}$$

by -1 on the both sides of the equality doesn't change anything to the problem, then we get:

$$\begin{aligned}
&\det(A-\lambda I_2)=0 \Rightarrow\\
&\det(\lambda I_2 - A) = \begin{vmatrix} \lambda - a_{11} & -a_{12} \\ -a_{21} & \lambda - a_{22} \end{vmatrix} = (\lambda - a_{11})(\lambda - a_{22}) - (-a_{12})(-a_{21})\\
&= \lambda^2 - (a_{11}+a_{22})\lambda + a_{11}a_{22} - a_{12}a_{21} = \lambda^2 - \underbrace{(a_{11}+a_{22})}_{\text{tr}(A)}\lambda + (a_{11}a_{22}-a_{12}a_{21}) = 0
\end{aligned} \tag{6.6.363}$$

Since we know that the eigenvalues are solutions of each of the above polynomials, we then have:

$$\lambda^2 - \underbrace{(a_{11}-a_{22})}_{\text{tr}(A)}\lambda + (a_{11}a_{22} - a_{21}a_{12}) = 0 \Rightarrow (\lambda-\lambda_1)(\lambda-\lambda_2) = \lambda^2 - (\lambda_1+\lambda_2)\lambda + \lambda_1\lambda_2 = 0 \tag{6.6.364}$$

Thus by a term by term correspondence it comes the very important result in Statistics (and also in Numerical Methods!) that we will prove later in a more general way with the spectral theorem:

$$\boxed{\text{tr}(A) = \sum_i \lambda_i} \tag{6.6.365}$$

If we look at $(\lambda I_n - A)$ as a linear application f, as it is non-trivial solutions that interest us, we can say that the eigenvalues are the elements λ such that:

$$\ker(f) = \ker(\lambda I_n - A) \neq \{\vec{0}\} \tag{6.6.366}$$

and that the kernel constitutes the "eigenspace" of A of the eigenvalue λ from which non-null elements are the eigenvectors. This is in some textbooks denoted by:

$$E_\lambda = \ker(f - \lambda Id) = \{x \in E; f(x) = \lambda x\} \tag{6.6.367}$$

And by the way... the set of all eigenvectors is named the "spectrum" of f.

It corresponds to the study of the main axes, according to which the application behaves like an expansion (homothetic application) multiplying the vectors by the same constant. This homothetic ratio is then the "eigenvalue", the vectors to which it applies the "eigenvectors" are assembled in a "eigenspace".

Another way of looking at it:

- A vector is said to be an "eigenvector" by a linear application if it is not zero and if the application does only change its size (norm) without changing its direction.

- An "eigenvalue", associated to an "eigenvector", is the size modification factor (homothetic ratio), i.e. the number by which we must multiply the vector to get its image. This factor can be negative (reverse direction of the vector) or zero (vector transformed into a vector of zero length).

  We can say that therefore that the eigenvalue λ "scalar" the application A for the eigenvector $\vec{X}$.

- An "eigenspace" associated to an "eigenvalue" is the set of eigenvectors that have the same eigenvalue value and a zero vector. They suffer all from the multiplication by the same factor.

Remark

In mechanics, we study the eigenfrequencies and eigenmodes of oscillating systems (see section Wave Mechanics page 3492). In Functional Analysis, an eigenfunction is an eigenvector for a linear operator, that is to say a linear application acting on a space of function (see section Functional Analysis page 2170). In geometry and optics, we speak of eigendirections to take into account the curvature of the surfaces (see section Non-Euclidean Geometry page 2663). In graph theory, an eigenvalue is simply an eigenvector of the adjacency matrix of the graph (see section Graph Theory page 3026).

So as the determinant of $\det(A-\lambda I_n)$ is a polynomial on λ then the λ_i are also the roots of the characteristic polynomial:

$$\begin{aligned}\det(A-\lambda I-n)&=(-1)^n(\lambda-\lambda_1)(\lambda-\lambda_2)\ldots(\lambda-\lambda_n)\\&=(-1)(\lambda-\lambda_1)(-1)(\lambda-\lambda_2)\ldots(-1)(\lambda-\lambda_n)\\&=(\lambda_1-\lambda)(\lambda_2-\lambda)\ldots(\lambda_n-\lambda)\end{aligned}\tag{6.6.368}$$

Therefore:

$$\boxed{\det(A-\lambda I_n)=\prod_{i=1}^{n}(\lambda_i-\lambda)}\tag{6.6.369}$$

This is a relation used sometimes in some statistical models (for example the MANOVA!).

Before closing this short introduction to the eigenvalues and eigenvectors (we will discussed this further below), let us indicate that since an eigenvector must satisfy the homogeneous system:

$$(A-\lambda I_n)\vec{X}=\vec{0}\tag{6.6.370}$$

Nothing prevents us then from multiplying the eigenvector by a constant k which normalizes it to the unit (technique often used in statistics and numerical methods to improve the floating precision of the algorithms) since:

$$(A-\lambda I_n)k\vec{X}=(A-\lambda I_n)\vec{U}=k\cdot\vec{0}=\vec{0}\tag{6.6.371}$$

Thus in practice it is customary that if the eigenvector is given for example by:

$$\vec{x}=\begin{pmatrix}x\\y\\z\end{pmatrix}\tag{6.6.372}$$

To normalize it at the unit by writing:

$$\vec{U}=\frac{\vec{X}}{\|\vec{X}\|}=\begin{pmatrix}\dfrac{x}{\sqrt{x^2+y^2+z^2}}\\ \dfrac{y}{\sqrt{x^2+y^2+z^2}}\\ \dfrac{z}{\sqrt{x^2+y^2+z^2}}\end{pmatrix} \tag{6.6.373}$$

For the section of Wave Quantum Physics and more especially for the study of the angular momentum and spin we need to prove that given an operator acting on an eigenvector, then if we square the operator, this result in squaring the eigenvalue.

Proof 6.42.2. Given:

$$Ax=\lambda x \tag{6.6.374}$$

Then:

$$A^2x=A(Ax)=A(\lambda x)=\lambda(Ax)=\lambda(\lambda x)=\lambda^2x \tag{6.6.375}$$

■ Q.E.D.

6.6.5.1 Rotation Matrices and Eigenvalues

Now that we have seen what was an eigenvalue and an eigenvector, let us come back on a particular type of orthogonal matrices that we will be particularly useful to us in our study of quaternions (see section Numbers page 261), of groups and symmetries (see section Set Algebra page 1417) and particle physics (see section Elementary Particle Physics page 5194).

We denote, as what has been seen in the section of Set Algebra, $\text{O}(n)$ the set of $n\times n$ (square) orthogonal matrices with coefficients in $\mathbb{R}$, that is to say, satisfying:

$$A^TA=\mathbb{1} \tag{6.6.376}$$

That will denote also sometimes for recall sometimes as:

$$A^TA=I \tag{6.6.377}$$

The columns and rows of an orthogonal matrix the form the orthonormal basis of the usual space $\mathbb{R}^2$ for the usual dot product.

The determinant of an orthogonal matrix is equal to ±1 (rotation conserves angles and volumes), indeed $A^TA=I$ leads to:

$$\det(A^T)\det(A)=\det(A^TA)=\det(I)=1 \tag{6.6.378}$$

A rotation matrix with determinant $+1$ is a "proper rotation", and one with a negative determinant -1 is an "improper rotation", that is a reflection combined with a proper rotation.

✍**Example:**

Let us calculate now explicitly the determinant of a 2×2 rotation matrix (see section Numbers page 258) and 3×3 rotation matrix (see section Euclidean Geometry page 2648) as it is ask by many student on various Internet forums.

So first we consider the 2×2 rotation matrix and using the relation of the determinant proved earlier, we get:

$$\begin{aligned}\det(A)&=\begin{bmatrix}\cos(\alpha) & -\sin(\alpha)\\ \sin(\alpha) & \cos(\alpha)\end{bmatrix}\\ &=\cos(\alpha)\cos(\alpha)-(-\sin(\alpha)\sin(\alpha))=\cos^2(\alpha)+\sin^2(\alpha)=1\end{aligned} \tag{6.6.379}$$

And for one randomly chosen rotation matrix of the three 3×3 rotation matrices (see section Euclidean Geometry page 2648), we get:

$$\begin{aligned}\det(R_Z(\phi))&=\left|\begin{bmatrix}\cos(\phi) & \sin(\phi) & 0\\ -\sin(\phi) & \cos(\phi) & 0\\ 0 & 0 & 1\end{bmatrix}\right|\\ &=a_{11}a_{22}a_{33}-a_{11}a_{23}a_{32}-a_{12}a_{21}a_{33}+a_{12}a_{23}a_{31}+a_{13}a_{21}a_{32}-a_{13}a_{22}a_{31}\\ &=\cos(\phi)\cos(\phi)\cdot 1-\cos(\phi)\cdot 0\cdot 0-\sin(\phi)(-\sin(\phi))\cdot 1+\sin(\phi)\cdot 0\cdot 0\\ &+1\cdot(-\sin(\phi))\cdot 0-0\cdot 0\cos(\phi)\cdot 0\\ &=\cos(\phi)\cos(\phi)-\sin(\phi)(-\sin(\phi))=\cos^2(\phi)+\sin^2(\phi)=1\end{aligned} \tag{6.6.380}$$

We denote by $\text{SO}(n)$ the set of orthogonal matrices of determinant 1 (for more details see the section of Set Algebra page 1417). Let us show in three points that if $A\in\text{SO}(3,\mathbb{R})$ then A is the rotation matrix relative to an axis passing through the origin.

1. Any eigenvalue of a rotation matrix A (real or complex) is of module 1. In other words it conserve the norm:

   Indeed, if λ is an eigenvalue of eigenvector $\vec{X}$, we have:

   $$|\lambda|^2\|\vec{X}\|^2=\|A\vec{X}\|^2=\langle A\vec{X}|A\vec{X}\rangle=\langle\vec{X}|A^TA\vec{X}\rangle=\langle\vec{X}|\vec{X}\rangle=\|\vec{X}\|^2 \tag{6.6.381}$$

   or noting the dot product with the book usual notation:

   $$|\lambda|^2\|\vec{X}\|^2=\|A\vec{X}\|^2=A\vec{X}\circ A\vec{X}=\vec{X}\circ A^TA\vec{X}=\vec{X}\circ\vec{X}=\|\vec{X}\|^2 \tag{6.6.382}$$

2. It exists a straight line in the space that is used a rotation axes and any vector on this line is not modified by any rotation.

   Let us denote by $\vec{X}$ an eigenvector of eigenvalue 1 (that is to say such that $A\vec{X}=\vec{X}$). As the reader may have perhaps already understand it (read until the end please!), the straight line generated by $\vec{x}$ that we will denote by $\langle\vec{X}\rangle$ constitutes our rotation axes.

   Indeed, any vector $\langle\vec{X}\rangle$ is send on itself by the application A. In this case, the orthonormal space denoted by $\langle\vec{X}\rangle^{\perp}$ that is of dimension 2 is the perpendicular plane to the rotation axes.

3. Any vector perpendicular to the rotation axes remains, after rotation, perpendicular to this axes. In other words, $\langle\vec{X}\rangle$ is invariant through the application of A.

   Indeed, if $\vec{w}\in\langle\vec{x}\rangle$ the, $w=A^TAw=A^Tw$ and for all $\vec{y}\in\langle\vec{X}\rangle^{\perp}$:

   $$A\vec{y}\circ\vec{w}=\vec{y}\circ A^T\vec{w}=\vec{y}\circ\vec{w}=0 \qquad (6.6.383)$$

   that is to say $A\vec{y}\langle\vec{X}\rangle^{\perp}$. Therefore $\langle\vec{X}\rangle^{\perp}$ is invariant by A.

   Finally, the restriction of A to the space $\langle\vec{X}\rangle^{\perp}$ is a rotation!

✍**Example:**

Given $e^{\mathrm{i}\alpha}$ (see the section Numbers where the rotation by the complex number is proven) an eigenvalue (which module is 1 as we proved it during our study of complex numbers) of A restraint to $\langle\vec{X}\rangle^{\perp}$.

Let us write $\vec{w}=\vec{u}+\mathrm{i}\vec{v}$ an eigenvector with $\vec{u},\vec{v}\in\mathbb{R}^2$ such as:

$$A\vec{w}=e^{\mathrm{i}\alpha}\vec{w} \qquad (6.6.384)$$

with (as we already proved it in our study of complex numbers):

$$A=\begin{pmatrix}\cos(\alpha) & -\sin(\alpha)\\ \sin(\alpha) & \cos(\alpha)\end{pmatrix} \qquad (6.6.385)$$

where we know by our study of complex numbers, that the vectors $\vec{u},\vec{v}$ generate an orthogonal basis (not necessarily normalized at the unit!) of $\langle\vec{X}\rangle^{\perp}$.

Remark

We think that it could by easy at this level of the reader to check that this matrix is orthogonal (if it not the case contact us and this will be detailed!).

6.6.6 Spectral Theorem

Let us now see a very important theorem relatively to the eigenvalues and eigenvectors which is named the "spectral theorem" which will be very useful to us for the various sections of physics of this textbook and also the section Statistics as well as in the section of Numerical Methods and Industrial Engineering.

To summarize, mathematicians say in their language that the spectral theorem give the possibility to affirm the diagonalization of endomorphism[27] (of matrices) and also justify the decomposition in eigenvalues (also named "singular value decomposition" abbreviated S.V.D.) that we will see further below during our presentation of some matrices decomposition techniques.

Remark

The singular value decomposition theorem (S.V.D.) is however very general, in the sense that it applies to any rectangular matrices. The eigenvalue decomposition (see further below the details), however, only works for some square matrices.

To simplify the proof, we will deal here only real matrices (component in $\mathbb{R}$) and also avoiding up the language of mathematicians.

We will denote in a first time $M_n(\mathbb{R})$ the set of all $n \times n$ (square) matrices with real coefficients.

We will confuse the matrix $M \in M_n(\mathbb{R})$ with the linear application on the vector space $\mathbb{R}^n$ by:

$$\vec{v} \mapsto M\vec{v} \tag{6.6.386}$$

with $\vec{v} \in \mathbb{R}^n$.

Reminder: We have seen above during our study of basis changes that if $(\vec{c}_1, \ldots, \vec{c}_n)$ is a basis of $\mathbb{R}^n$ and $M \in M_n(\mathbb{R})$ then the matrix of the linear map M in the basis $(\vec{c}_1, \ldots, \vec{c}_n)$ is:

$$S^{-1}MS \tag{6.6.387}$$

where S is the matrix formed by the column vectors $\vec{c}_1, \ldots, \vec{c}_n$.

First, we simply check that if A is a symmetric matrix then (this should be trivial but it can be verified with an example of dimension 2 very quickly):

$$(A\vec{u}) \circ \vec{v} = \vec{u} \circ A^T\vec{v} = \vec{u} \circ A\vec{v} \tag{6.6.388}$$

We now propose to study the following properties of a symmetric matrix M:

P1. All eigenvalues of M are reals.

Proof 6.42.3. Given:

$$\vec{z} = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix} \in \mathbb{C}^n \tag{6.6.389}$$

---

[27] What they also name an "endomorphism reduction"

an a priori complex eigenvector of the eigenvalue $\lambda \in \mathbb{C}$. Let us denote:

$$\vec{\bar{z}} = \begin{pmatrix} \bar{z}_1 \\ \vdots \\ \bar{z}_n \end{pmatrix} \tag{6.6.390}$$

the conjugate vector of $\vec{z}$. Then we have:

$$\vec{\bar{z}}^T M \vec{z} = \vec{\bar{z}}^T \lambda \vec{z} = \lambda \|\vec{z}\|^2 \tag{6.6.391}$$

On the other hand since $M = \overline{M}$ we have:

$$\vec{\bar{z}}^T M \vec{z} = \vec{\bar{z}}^T \overline{M} \vec{z} = \overline{\vec{z}^T M \vec{z}} = \overline{(M\vec{z})^T} \vec{z} = \overline{\lambda \vec{z}^T} \vec{z} = \bar{\lambda} \|\vec{z}\|^2 \tag{6.6.392}$$

As $\vec{z} \neq \vec{0}$, we have $\lambda = \vec{\lambda}$ and therefore, $\lambda \in \mathbb{R}$.

■ Q.E.D.

P2. Two eigenspaces of M relative to different eigenvalues are orthogonal (in other words, the eigenvectors are independent).

Proof 6.42.4. Let λ, μ be two distinct eigenvalues of corresponding eigenvectors $\vec{u}, \vec{v}$. We have (don't forget that M is symmetric!):

$$\lambda(\vec{u} \circ \vec{v}) = (\lambda \vec{u}) \circ \vec{v} = (M\vec{u}) \circ \vec{v} = \vec{u} \circ (M\vec{v}) = \vec{u} \circ (\mu \vec{v}) = \mu(\vec{u} \circ \vec{v}) \tag{6.6.393}$$

therefore:

$$(\lambda - \mu) \cdot (\vec{u} \circ \vec{v}) = 0 \tag{6.6.394}$$

which leads to:

$$(\vec{u} \circ \vec{v}) = 0 \tag{6.6.395}$$

■ Q.E.D.

Before going further, we also have to prove that if $M \in M_n(\mathbb{R})$ is a symmetrical matrix and V and vectorial subspace of $\mathbb{R}^n$ invariant relatively to M (that is to say that satisfies for any $\vec{v} \in V : M\vec{v} \in V$) then we have the following properties:

P3. The orthogonal of V denoted obviously by $V^\perp$ (obtained by applying the Gram-Schmidt method seen in the section of Vector Calculus page 2062) is also invariant through M.

Proof 6.42.5. Given $\vec{v} \in V$ and $\vec{w} \in V^\perp$ then:

$$(M\vec{w}) \circ \vec{v} = \vec{w} \circ \left(\underbrace{M\vec{v}}_{\in V} \right) = 0 \tag{6.6.396}$$

this shows well that $M\vec{w} \in V^\perp$.

■ Q.E.D.

P4. If $(\vec{w}_1,\ldots,\vec{w}_k)$ is an orthonormal basis of $\vec{V}^{\perp}$ then the restriction matrix of M to $V^{\perp}$ in the basis $(\vec{w}_1,\ldots,\vec{w}_k)$ is also symmetrical.

Proof 6.42.6. Let us denote $A=(a_{ij})_{1\leq i,j\leq k}$ the matrix of the restriction of M to $V^{\perp}$ in the basis $(\vec{w}_1,\ldots,\vec{w}_k)$. We have by definition for any $j=1\ldots k$ (as the vector resulting of a linear application such as M can be express in its basis):

$$M\vec{w}_j=\sum_{i=1}^{k}a_{ij}\vec{w}_i \tag{6.6.397}$$

Or:

$$(M\vec{w}_j)\circ\vec{w}_m=\left(\sum_{i=1}^{k}a_{ij}\vec{w}_i\right)\circ\vec{w}_m=\sum_{i=1}^{k}a_{ij}(\vec{w}_i\circ\vec{w}_m)=a_{mj} \tag{6.6.398}$$

as:

$$\vec{w}_i\circ\vec{w}_m=0 \tag{6.6.399}$$

if $i\neq m$ in the orthonormal basis.

On another side:

$$(M\vec{w}_j)\circ\vec{w}_m=\vec{w}_j\circ(M\vec{w}_m)=\vec{w}_j\circ\left(\sum_{i=1}^{k}a_{im}\vec{w}_i\right)=\sum_{i=1}^{k}a_{im}(\vec{w}_j\circ\vec{w}_i)=a_{jm} \tag{6.6.400}$$

Therefore:

$$a_{mj}=a_{jm} \tag{6.6.401}$$

This shows that:

$$A=A^T \tag{6.6.402}$$

■ Q.E.D.

Theorem 6.43. We will now be able to show that any symmetric matrix $M\in M_n(\mathbb{R})$ is diagonalizable. That is to say that there is an invertible matrix S such that the result of the calculation:

$$S^{-1}MS \tag{6.6.403}$$

gives a diagonal matrix!

Remark

In fact we will see, to be more precise, that there exists an orthogonal matrix S such that $S^{-1}MS$ is diagonal.

Reminder: Say that S is "orthogonal" means that $SS^T=I$ (where I is the identity matrix) which is equivalent to say that the columns of S form an orthonormal basis of $\mathbb{R}^n$.

Proof 6.43.1. We prove the assertion by induction on n. If $n=1$ there is nothing to prove. Let us suppose that the assertion is satisfied for $k\leq n$ and let us prove it for $k=n+1$. Then given $M\in M_{n+1}(\mathbb{R})$ a symmetric matrix and λ an eigenvalue of M.

We easily verify that the eigenspace:

$$W=\text{ker}(\lambda I_{n+1}-M) \tag{6.6.404}$$

is invariant by M (just take any numerical application) and that by the proof seen earlier, that $W^{\perp}$ is also invariant by M. Moreover, we know (see section Vector Calculus page 1814) that $\mathbb{R}^{n+1}$ can be decomposed into a direct sum:

$$\mathbb{R}^{n+1}=W\oplus W^{\perp} \tag{6.6.405}$$

If:

$$\dim(W^{\perp})=0 \tag{6.6.406}$$

then:

$$\mathbb{R}^{n+1}=W \tag{6.6.407}$$

and it is sufficient to take an orthonormal basis of W to diagonalise M. Indeed, if $(\vec{w}_1,\ldots,\vec{w}_{n+1})$ is such a basis, the matrix S formed by the column vectors $\vec{w}_j$ $(j=1\ldots n+1)$ is orthogonal and satisfies:

$$S^{-1}MS=S^{-1}\lambda S=\lambda I_{n+1} \tag{6.6.408}$$

and $S^{-1}MS$ is indeed diagonal.

Let us now suppose that $\dim(W^{\perp})>0$ and given $(\vec{u}_1,\ldots,\vec{u}_m)$ with $m\leq n$ an orthonormal basis of $W^{\perp}$. Let us denote by A the restriction matrix of M to $W^{\perp}$ in the basis $(\vec{u}_1,\ldots,\vec{u}_m)$. A is also symmetric (as proved in one of the preceding properties).

By induction hypothesis there exists an orthogonal matrix $H\in M_m(\mathbb{R})$ such that $H^{-1}AH$ is diagonal.

Let us denote by $(\vec{w}_1,\ldots,\vec{w}_{n+1-m})$ an orthonormal basis of W and G the matrix formed by the column vectors: $\vec{w}_1,\ldots,\vec{w}_{n+1-m},\vec{u}_1,\ldots\vec{u}_m$. So we can write that:

$$G^{-1}MG=\begin{bmatrix}\lambda I_{n+1-m} & 0\\ 0 & A\end{bmatrix} \tag{6.6.409}$$

and G is also orthogonal by construction.

Let us consider the following block matrix (matrix of matrices):

$$L=\begin{bmatrix}I_{n+1-m} & 0\\ 0 & H\end{bmatrix}\in M_{n+1}(\mathbb{R}) \tag{6.6.410}$$

and let us put:

$$S=GL \tag{6.6.411}$$

It is almost obvious that S is orthogonal as G and L are also orthogonal. Indeed, if:

$$GG^T = I \qquad \text{and} \qquad LL^T = I \tag{6.6.412}$$

then (remember that matrix multiplication is associative !!!):

$$(GL)(GL)^T = (GL)(L^TG^T) = G(LL^T)G^T = GIG^T = GG^T = I \tag{6.6.413}$$

Also S satisfies:

$$\begin{aligned}S^{-1}MS &= L^{-1}(G^{-1}MG)L\\ &= \begin{bmatrix} I_{n+1-m} & 0 \\ 0 & H \end{bmatrix}^{-1} \cdot \begin{bmatrix} \lambda I_{n+1-m} & 0 \\ 0 & A \end{bmatrix} \cdot \begin{bmatrix} I_{n+1-m} & 0 \\ 0 & H \end{bmatrix} = \begin{bmatrix} \lambda I_{n+1-m} & 0 \\ 0 & H^{-1}AH \end{bmatrix}\end{aligned} \tag{6.6.414}$$

and then:

$$\begin{bmatrix} \lambda I_{n+1-m} & 0 \\ 0 & H^{-1}AH \end{bmatrix} \tag{6.6.415}$$

is indeed diagonal.

■ Q.E.D.

Finally here is finally the famous "spectral theorem" (real case):

Theorem 6.44. Given $M \in M_n(\mathbb{R})$ a symmetric matrix, then there exists an orthonormal basis made of eigenvectors of M.

Proof 6.44.1. So we have seen in the preceding paragraphs that there exists an orthogonal matrix S such that $S^{-1}MS$ is diagonal if M is symmetric! Let denote by $\vec{c}_1, \ldots, \vec{c}_n$ the columns of S. The basis $(\vec{c}_1, \ldots, \vec{c}_n)$ is an orthonormal basis of $\mathbb{R}^2$ as S is orthogonal. Denoting the $\vec{e}_i$ the i-th vector of the canonical basis of $\mathbb{R}^n$ and λ_i and the i-th diagonal coefficient of S^1MS we have without directly supposing that λ_i is an eigenvalue for now:

$$\lambda_i\vec{e}_i = (S^{-1}MS)\vec{e}_i = S^{-1}M\vec{c}_i \tag{6.6.416}$$

by multiplying by S on both sides of the equality we have:

$$\lambda_i S\vec{e}_i = M\vec{c}_i \tag{6.6.417}$$

and therefore:

$$\lambda_i\vec{c}_i = M\vec{c}_i \tag{6.6.418}$$

which shows that $\vec{c}_1, \ldots, \vec{c}_n$ are eigenvectors and λ_i the corresponding eigenvalues.

■ Q.E.D.

To finish about the spectral theorem in this textbook, let us so reprove a result we get earlier but that was presented in a quite ugly way and poorly rigorous (the sum of the eigenvalues equals the trace of a matrix):

$$\text{tr}(A)=\sum_i \lambda_i \tag{6.6.419}$$

Remember that spectral theorem therefore tells us that for any symmetric matrix M, there exists an orthogonal matrix S such that:

$$S^{-1}MS \tag{6.6.420}$$

is diagonal. Nothing prevents us to choose the resulting diagonal matrix as a matrix of eigenvalues in the diagonal. What we denote usually:

$$\Lambda=S^{-1}MS \tag{6.6.421}$$

and as S is a real orthogonal matrix and that by definition we have that a matrix is orthogonal if and only if $A^{-1}=A^T$, then we find the following relation:

$$\boxed{\Lambda=S^TMS} \tag{6.6.422}$$

Commonly named a "spectral decomposition" or "eigendecomposition".

So obviously we will have to find S if M is known or vice versa. Anyway, let us come back on our topic and take track of this relation:

$$\text{tr}(\Lambda)=\sum_i \lambda_i=\text{tr}(S^{-1}MS) \tag{6.6.423}$$

Then by using the property of the trace tr, of the associativity of the matrix multiplication, and the orthogonality of S we have:

$$\sum_{i=1}^n \lambda_i=\text{tr}\left((S^{-1}M)S\right)=\text{tr}\left(S(S^{-1}M)\right)=\text{tr}\left((SS^{-1})M\right)=\text{tr}(I_nM)=\text{tr}(M) \tag{6.6.424}$$

This reprove the results seen earlier above with a condition that was not trivial at this time: the matrix must be symmetrical (or symmetrizable)!

We also have by extension:

$$M=S\Lambda S^{-1} \tag{6.6.425}$$

and therefore by using the proven property relatively to the determinant (during our proofs of the main determinant properties) and the conjugated matrices we get:

$$\det(M)=\det(\Lambda S^{-1})=\det(S)\det(\Lambda)\det(S^{-1})=\det(S)\det(\Lambda)\det(S)^{-1}=\det(\Lambda) \tag{6.6.426}$$

and therefore if M is symmetric we have the important property:

$$\boxed{\det(M)=\det(\Lambda)=\prod \lambda_i} \tag{6.6.427}$$

Remark

This result is in this text a particular form of the more general case (thus also applicable to rectangular or also non-orthogonal matrices) and that seems to be named the "Eckart-Young theorem".

✍**Example:**

We want to show that:

$$M=\begin{bmatrix}1 & 3\\ 3 & 1\end{bmatrix} \tag{6.6.428}$$

we assume that we know that the eigenvalue-eigenvector pairs are:

$$\lambda_1=4 \qquad \vec{v}_1=\begin{pmatrix}1\\ 1\end{pmatrix} \qquad \text{and} \qquad \lambda_2=-2 \qquad \vec{v}_2=\begin{pmatrix}-1\\ 1\end{pmatrix} \tag{6.6.429}$$

We therefore introduce S and Λ as follows:

$$S=[\vec{v}_1,\vec{v}_2]=\begin{bmatrix}1 & -1\\ 1 & 1\end{bmatrix}\Rightarrow S^{-1}=\frac{1}{2}\begin{bmatrix}1 & 1\\ -1 & 1\end{bmatrix} \qquad \Lambda=\begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix}=\begin{bmatrix}4 & 0\\ 0 & -2\end{bmatrix} \tag{6.6.430}$$

We must show that $M=S\Lambda S^{-1}$. This is indeed the case, since:

$$S\Lambda S^{-1}=\begin{bmatrix}1 & -1\\ 1 & 1\end{bmatrix}\begin{bmatrix}4 & 0\\ 0 & -2\end{bmatrix}\frac{1}{2}\begin{bmatrix}1 & 1\\ -1 & 1\end{bmatrix}=\begin{bmatrix}4 & 2\\ 4 & -2\end{bmatrix}\frac{1}{2}\begin{bmatrix}1 & 1\\ -1 & 1\end{bmatrix}=\begin{bmatrix}2 & 1\\ 2 & -1\end{bmatrix}\begin{bmatrix}1 & 1\\ -1 & 1\end{bmatrix}=\begin{bmatrix}1 & 3\\ 3 & 1\end{bmatrix}=M \tag{6.6.431}$$

Therefore M is indeed diagonalizable.

6.6.7 Matrix Decompositions

In the mathematical discipline of linear algebra, a "matrix decomposition" or "matrix factorization" is a factorization of a matrix into a product of matrices in the purpose to simplify problems (typically to simplify the resolution of linear systems especially when the matrix to invert are huge!) or to make emerge some interesting properties in various cases.

There are many different matrix decompositions. Each finds use among a particular class of problems. There is almost two dozens of matrix decomposition techniques as far as we know. However in the texts that will follow below we will focus only on techniques that are used for practical explicit applications in this textbook (mainly in the field of Statistics and by extensions in financial engineering and also Machine Learning).

6.6.7.1 Singular Value Decomposition (SVD)

The singular value decomposition of a matrix M $(m\times n)$ is the factorization of M into the product of three matrices:

$$M=UDV^T \tag{6.6.432}$$

where the columns of U $(m\times r)$ and V $(r\times n)$ are orthonormal such that[28]:

$$\begin{gathered}\sum_{i=1}^{m}U_{ik}U_{ij}=\delta_{kj}\\ \sum_{i=1}^{r}V_{ik}V_{ij}=\delta_{kj}\end{gathered} \tag{6.6.433}$$

and the matrix D $(r\times r)$ is diagonal with positive real entries (r is the rank of M, i.e. $K\leq\min\{m,n\}$).

Remark

If V is orthogonal, then as we have seen at page 1954, we have $V^{-1}=V^T$, then that latter relation can also be written:

$$M=UDV^{-1} \tag{6.6.434}$$

The SVD is useful in many tasks as Data Mining, Image Processing and Advanced Numerical Methods.

To gain insight into the SVD, we treat the rows of an $m\times n$ matrix M as m points in a n-dimensional space and consider the problem of finding the best k-dimensional subspace with respect to the set of points. Here "best" means minimize the sum of the squares of the perpendicular distances of the points to the subspace.

---

[28] In other words, U and V are rotation matrices such that $U^TU=\mathbb{1}_{m\times m}$ and $V^TV=\mathbb{1}_{r\times r}$.

Let us begin with a special case of the problem where the subspace is 1-dimensional: a line through the origin. We will see later that the best-fitting k-dimensional subspace can be found by k applications of the best fitting line algorithm. Finding the best fitting line through the origin with respect to a set of points $\{x_i|1 \leq i \leq m\}$ in the plane means minimizing the sum of the squared distances of the points to the line. Here distance is measured perpendicular to the line. The problem is then named as we know: the "best least squares fit" (see section Numerical Methods page 6324).

Consider projecting a point $\vec{x}_1$ onto a line through the origin:

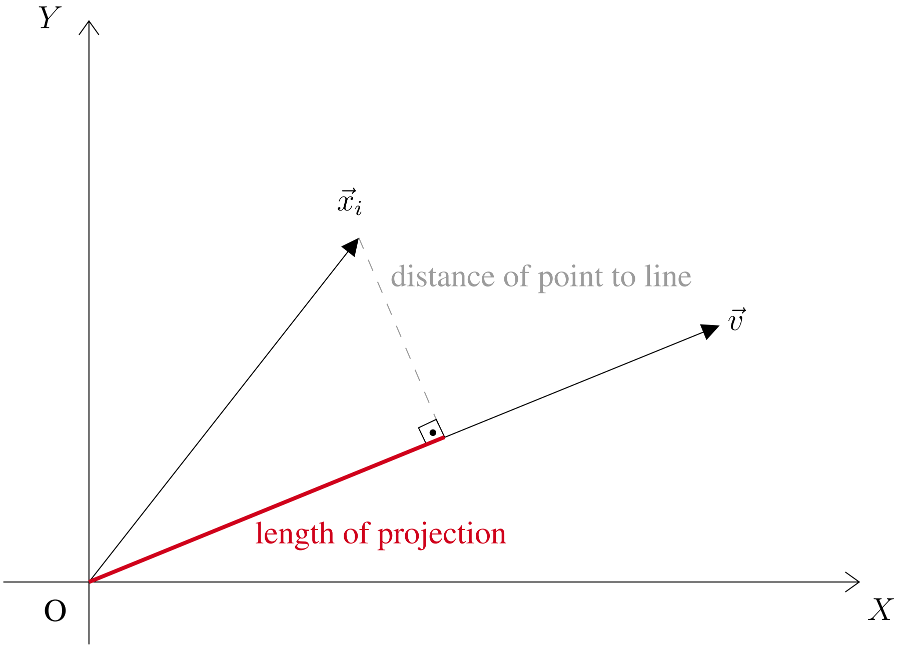


Figure 6.147: The projection of the point $\vec{x}_i$ onto the line through the origin in the direction of $\vec{v}$

Then using Pythagorean theorem we get:

$$\sqrt{x_{i1}^2+x_{i2}^2+\ldots+x_{in}^2}^{\,2} = (\text{lenght of projection})^2 + (\text{distance of point to line})^2 \tag{6.6.435}$$

That is:

$$x_{i1}^2+x_{i2}^2+\ldots+x_{in}^2 = (\text{lenght of projection})^2 + (\text{distance of point to line})^2 \tag{6.6.436}$$

Therefore (see figure):

$$(\text{distance of point to line})^2 = x_{i1}^2+x_{i2}^2+\ldots+x_{in}^2 - (\text{lenght of projection})^2 \tag{6.6.437}$$

To minimize the sum of squares for the distance to the line, one could minimize:

$$\sum_{i=1}^{m}(x_{i1}^2+x_{i2}^2+\ldots+x_{in}^2) \tag{6.6.438}$$

minus the sum of the square of the lengths of the projections of the points to the line. However, $\sum_{i=1}^{m}(x_{i1}^2+x_{i2}^2+\ldots+x_{in}^2)$ is constant! (independent of the line), so minimizing the sum of the squares of the distances is equivalent to maximizing the sum of the squares of the lengths of the projections onto the line. Similarly for best-fit subspaces, we could maximize the sum of the squared lengths of the projections onto the subspace instead of minimizing the sum of squared distances to the subspace.

6.6.7.1.1 Singular Vectors and Values

We will now build the "singular vectors" (and also "singular values") of a $m \times n$ matrix M.

Consider the rows of M as m points in a d-dimensional space. Consider the best fit line through the origin. Let $\vec{v}$ be a unit vector along this line.

The length of the projection of $\vec{x}_i$, the i-th row M, onto $\vec{v}$ is (see section Vector Calculus page 1821):

$$\|\text{proj}_v \vec{x}_i\| \tag{6.6.439}$$

That we will denote for what follows as (see section Vector Calculus page 1821):

$$|\vec{x}_i \circ \vec{v}| = \|\text{proj}_v \vec{x}_i\| \tag{6.6.440}$$

So in our case:

$$x_{i1}^2 + x_{i2}^2 + \ldots + x_{in}^2 = |\vec{x}_i \circ \vec{v}| \tag{6.6.441}$$

From this we denote the sum of all lengths of the projections by:

$$\|\text{proj}_v M\| = \|M\vec{v}\|^2 \tag{6.6.442}$$

The best fit line is the on maximizing $\|M\vec{v}\|^2$ (i.e.: $\|M\vec{v}\|$) and hence minimizing the sum of the squared distances of the points to the line.

With this in mind, we define the "first singular vector $\vec{v}_1$", of M, which is a vector, as the best fit through the origin for the m points in n-space that are the rows of M. Thus:

$$\vec{v}_1 = \max_{\|\vec{v}\|=1} \|M\vec{v}\| \tag{6.6.443}$$

The scalar value:

$$\sigma_1(M) = \|M\vec{v}_1\| \tag{6.6.444}$$

is named the "first singular value" of M. Notice that σ_1^2 is therefore implicitly the sum of the squares of the projections of the points to the line determined by $\vec{v}_1$.

The greedy approach to find this time not the best fit 1-dimensions but 2-dimensional subspace for a matrix M, takes $\vec{v}_1$ as the first basis vector for the 2-dimensional subspace and finds the best 2-dimensional subspace containing $\vec{v}_2$.

Thus, instead of looking for the best 2-dimensional subspace containing $\vec{v}_1$, look for a unit vector, denoted $\vec{v}_2$, perpendicular to $\vec{v}_1$ that maximizes $\|M\vec{v}\|^2$ among all such unit vectors.

Using the same strategy to find the best three and higher dimensional subspaces, defines $\vec{v}_3, \vec{v}_4, \ldots$ in similar manner.

The "second singular vector $\vec{v}_2$", is defined by the best fit line perpendicular to $\vec{v}_1$:

$$\vec{v}_2 = \max_{\vec{v} \perp \vec{v}_1, \|\vec{v}\|=1} \|M\vec{v}\| \tag{6.6.445}$$

The value:

$$\sigma_2(M)=\|M\vec{v}_2\| \tag{6.6.446}$$

is named the "second singular value" of M.

The "third singular vector $\vec{v}_3$" is defined similarly by:

$$\vec{v}_3=\max_{\vec{v}\perp\vec{v}_1,\vec{v}_2,\|\vec{v}\|=1}\|M\vec{v}\| \tag{6.6.447}$$

and so on...

The process stop theoretically when we have found $\vec{v}_1,\vec{v}_2,\ldots,\vec{v}_r$ as singular vector that satisfies:

$$\vec{v}_r=\max_{\vec{v}\perp\vec{v}_1,\ldots,\vec{v}_r,\|\vec{v}\|=1}\|M\vec{v}\| \tag{6.6.448}$$

As the $\vec{v}_i$ are perpendiculars, if we apply M on all this vectors, the resulting vectors will also be perpendicular between them!

Therefore we build the vectors:

$$\vec{u}_i=\frac{1}{\sigma_i(M)}M\vec{v}_i \tag{6.6.449}$$

that are all perpendiculars vectors between them as already mentioned and named "left singular vectors" of M when the $\vec{v}_i$ will be named "right singular vectors". The SVD theorem will fully explain the reason for these terms.

Theorem 6.45. Let M be an $m\times n$ matrix with right singular vector $\vec{v}_1,\ldots,\vec{v}_r$, left singular vectors $\vec{u}_1,\ldots,\vec{u}_r$, and corresponding singular values $\sigma_1,\ldots,\sigma_n$. Then the "singular value decomposition theorem" states that:

$$\boxed{M=\sum_{i=1}^{r}\sigma_i\vec{u}_i\vec{v}_i^T} \tag{6.6.450}$$

Proof 6.45.1. We start naturally from:

$$\sum_{i=1}^{r}\sigma_i\vec{u}_i\vec{v}_i^T=\sum_{i=1}^{r}\sigma_i\left(\frac{1}{\sigma_i}M\vec{v}_i\right)\vec{v}_i^T=\sum_{i=1}^{r}M\vec{v}_i\vec{v}_i^T=M\sum_{i=1}^{r}\vec{v}_i\vec{v}_i^T \tag{6.6.451}$$

Remark

Don't forget that $\vec{x}^T\vec{x}$ in Linear Algebra gives a scalar that is equivalent to the "dot product" ("inner product"), when instead $\vec{x}\vec{x}^T$ gives a square matrix named the "outer product", or sometimes "Gram matrix", defined by:

$$\vec{u}\vec{v}^T=\begin{pmatrix}u_1\\u_2\\u_3\\u_4\end{pmatrix}\begin{pmatrix}v_1 & v_2 & v_3\end{pmatrix}=\begin{bmatrix}u_1v_1 & u_1v_2 & u_1v_3\\u_2v_1 & u_2v_2 & u_2v_3\\u_3v_1 & u_3v_2 & u_3v_3\\u_4v_1 & u_4v_2 & u_4v_3\end{bmatrix} \tag{6.6.452}$$

Now let us consider a special case with (as I don't like the general proof):

$$\vec{v}_1 = \begin{pmatrix} v_{11} \\ v_{21} \\ v_{31} \end{pmatrix} \qquad \vec{v}_2 = \begin{pmatrix} v_{12} \\ v_{22} \\ v_{32} \end{pmatrix} \qquad \vec{v}_3 = \begin{pmatrix} v_{13} \\ v_{23} \\ v_{33} \end{pmatrix} \tag{6.6.453}$$

That gives:

$$\begin{aligned}\sum_{i=1}^{3} \vec{v}_i \vec{v}_i^T &= \begin{bmatrix} v_{11}^2 & v_{11}v_{21} & v_{11}v_{31} \\ v_{21}v_{11} & v_{21}^2 & v_{21}v_{31} \\ v_{31}v_{11} & v_{31}v_{21} & v_{31}^2 \end{bmatrix} + \begin{bmatrix} v_{12}^2 & v_{12}v_{22} & v_{12}v_{32} \\ v_{22}v_{12} & v_{22}^2 & v_{22}v_{32} \\ v_{32}v_{12} & v_{32}v_{22} & v_{32}^2 \end{bmatrix} + \begin{bmatrix} v_{13}^2 & v_{13}v_{23} & v_{13}v_{33} \\ v_{23}v_{13} & v_{23}^2 & v_{23}v_{33} \\ v_{33}v_{13} & v_{33}v_{23} & v_{33}^2 \end{bmatrix} \\ &= \begin{bmatrix} v_{11}^2+v_{12}^2+v_{13}^2 & v_{11}v_{21}+v_{12}v_{22}+v_{13}v_{23} & v_{11}v_{31}+v_{12}v_{32}+v_{13}v_{33} \\ v_{21}v_{11}+v_{22}v_{12}+v_{23}v_{13} & v_{21}^2+v_{22}^2+v_{23}^2 & v_{21}v_{31}+v_{22}v_{32}+v_{23}v_{33} \\ v_{31}v_{11}+v_{32}v_{12}+v_{33}v_{13} & v_{31}v_{21}+v_{32}v_{22}+v_{33}v_{23} & v_{31}^2+v_{32}^2+v_{33}^2 \end{bmatrix}\end{aligned} \tag{6.6.454}$$

So if we look closely the result is the same as if we define the matrix:

$$V = [\vec{v}_1, \vec{v}_2, \vec{v}_3] = \begin{bmatrix} v_{11} & v_{12} & v_{13} \\ v_{21} & v_{22} & v_{23} \\ v_{31} & v_{32} & v_{33} \end{bmatrix} \tag{6.6.455}$$

Therefore we can see that:

$$VV^T = \begin{bmatrix} v_{11} & v_{12} & v_{13} \\ v_{21} & v_{22} & v_{23} \\ v_{31} & v_{32} & v_{33} \end{bmatrix} \begin{bmatrix} v_{11} & v_{21} & v_{31} \\ v_{12} & v_{22} & v_{32} \\ v_{13} & v_{23} & v_{33} \end{bmatrix} \tag{6.6.456}$$

leads to the same result. Therefore we can write:

$$\sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^T = M \sum_{i=1}^{r} \vec{v}_i \vec{v}_i^T = M(VV^T) \tag{6.6.457}$$

But we must not forget that if V is an orthogonal matrix, then it represents an orthonormal basis and then we have proved already earlier above that in this case:

$$V^T V = VV^T = \mathbb{1} = [\delta_{ij}] \tag{6.6.458}$$

Therefore:

$$\sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^T = M(VV^T) = M\mathbb{1} = M \tag{6.6.459}$$

And this finish the proof!

■ Q.E.D.

The reader should also notice that:

$$M = \sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^T \tag{6.6.460}$$

and:

$$\boxed{\underset{(m\times n)}{M}=\underset{(m\times r)}{U}\underset{(r\times r)}{D}\underset{(r\times n)}{V^T}} \tag{6.6.461}$$

also often denoted $M=U\Sigma V^T$ (as Σ is the upper case letter of σ), are equivalent notation for the same thing (just develop the last one explicitly and you will see you fall back on the same result[29])! The difference is that the notation with the sum is most used in Data Mining and that with the matrices in Statistics.

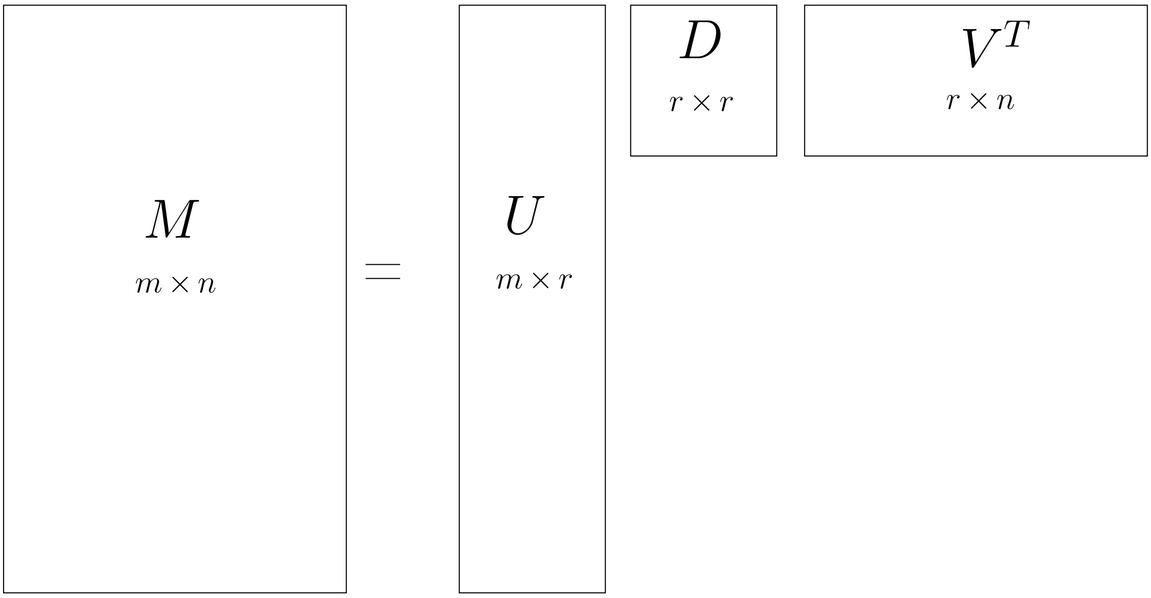


Figure 6.148: SVD decomposition of a $m\times n$ matrix

It is usage to build the matrix D such that the diagonal is in descending order of amplitude and to order the vectors $\vec{v}_i$ in the corresponding order. The reason is quite easy to understand as you can see in the example further below.

This is important to know that this not the only possible decomposition of a matrix. The are many other one but we will focus in this textbook only the decomposition that are directly useful for engineering topics presented in this textbook.

Remark

Some authors prefers to work with the norm of the singular values, that is, with $\sqrt{\sigma_i}$, therefore the left singular value are defined as:

$$\vec{u}_i=\frac{1}{\sqrt{\sigma_i}}M\vec{v}_i \tag{6.6.462}$$

Without that it changes our previous proof result that will just be:

$$M=\sum_{i=1}^{r}\sqrt{\sigma_i}\vec{u}_i\vec{v}^T \tag{6.6.463}$$

---

[29] On request we can write the details

That in Europe is frequently written (but can bring to confusion with the notation of eigenvalues...):

$$M = \sum_{i=1}^{r} \sqrt{\lambda_i} \vec{u}_i \vec{v}^T \quad (6.6.464)$$

or in matrix form (...):

$$M = U\Lambda^{1/2}V^T \quad (6.6.465)$$

Let us take a practical example in image processing of SVD (made by Jason Liu in MATLAB®). The image below is an image made of 400 unique row vectors (the reader can find the equivalent example in our `R` companion book):

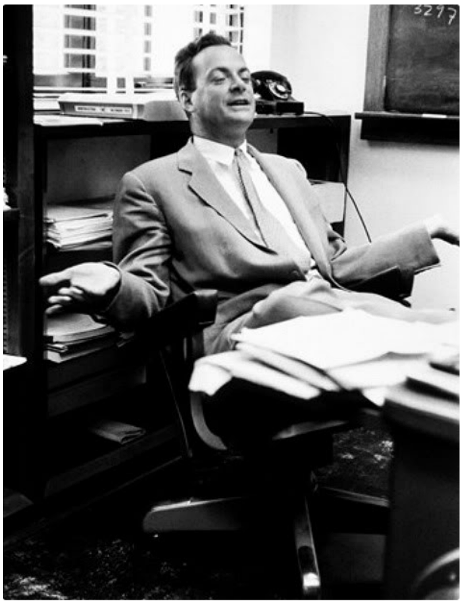


Figure 6.149: SVD in MATLAB® example original image

What happens if in the sum:

$$M = \sum_{i=1}^{r} \sqrt{\sigma_i} \vec{u}_i \vec{v}^T \quad (6.6.466)$$

we take only the first biggest singular vector?:

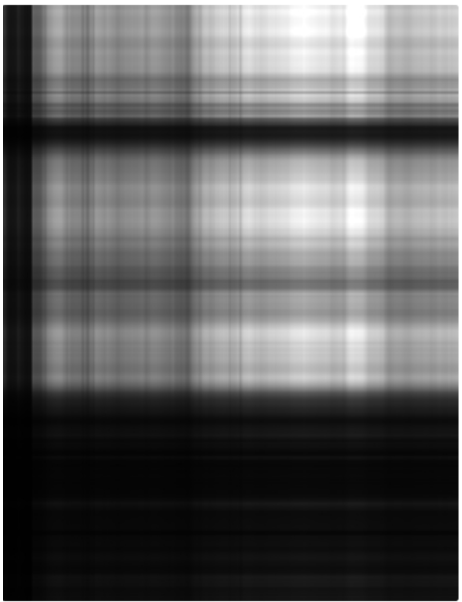


Figure 6.150: **MATLAB® SVD with $r=1$**

What happens if we take the first two singular vectors?:

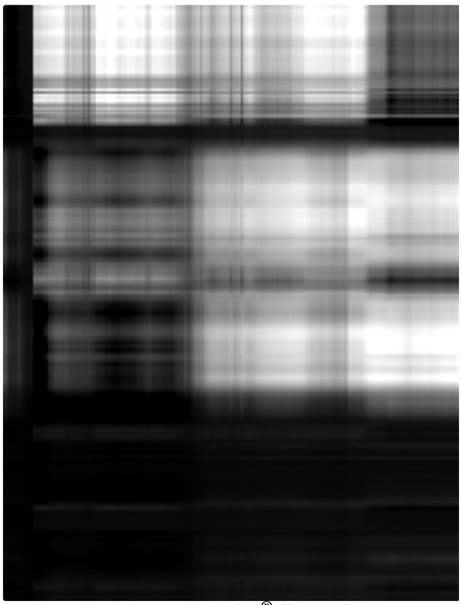


Figure 6.151: **MATLAB® SVD with $r=2$**

...and if we take the first ten singular vectors?:

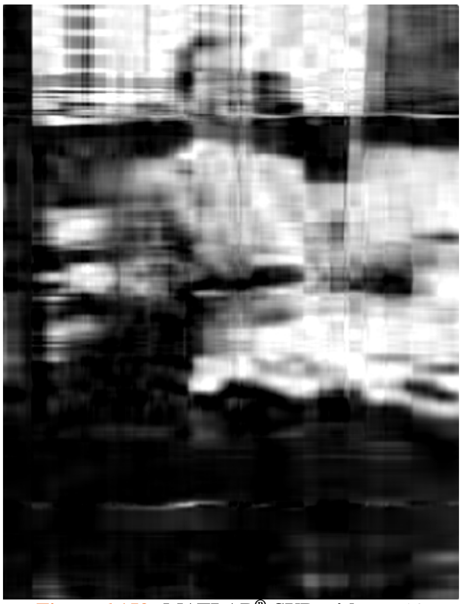


Figure 6.152: **MATLAB® SVD with** $r=10$

...and if we take the first fifty singular vectors?:

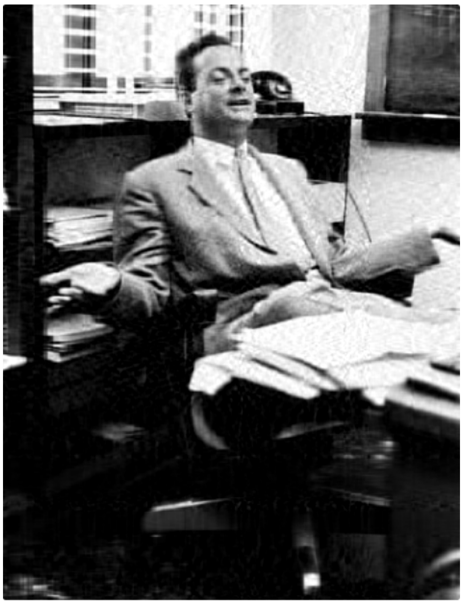


Figure 6.153: **MATLAB® SVD with** $r=50$

There we have it! Using 50 singular vectors we get a decent representation of what 400 vectors look like.

So as we can see SVD is a great space reduction technique!

Remark

Let $U\Sigma V^T$ be the singular value decomposition of a matrix A. Then U is an $m\times m$ unitary matrix (i.e. $U^TU=I$), Σ is a diagonal $m\times n$ matrix ($\Sigma_{ii}=\sigma_i(A)$), and V is an

$n\times n$ unitary matrix (i.e. $V^TV=I$). Using the Frobenius norm (see page 1623) we have:

$$\begin{aligned}\|A\|_F^2&=\text{tr}\left(A^TA\right)=\text{tr}\left(\left(U\Sigma V^T\right)^T U\Sigma V^T\right)=\text{tr}\left(V\Sigma^TU^TU\Sigma V^T\right)\\&=\text{tr}\left(\Sigma^T\Sigma\right)=\sum_{i=1}^{\min\{m,n\}}\sigma_i^2(A)\end{aligned}$$

where we use the property of unitary matrix and the cyclic property of trace.

6.6.7.1.2 Special case of SVD for symmetric matrices

The SVD for symmetric matrices is the unique case that interest us for practical purposes in this textbook (for the PCA: Principal Component Analysis). Therefore we will do a special focus on this latter case.

We will prove here in two steps that first, for a symmetric matrix, the eigenvectors are orthogonal and secondly that the Singular vectors and Singular values are respectively equal to the eigenvectors and eigenvalues for symmetric matrices.

Let us see the first step! In general, for any matrix, it's eigenvectors (when the exist) are NOT always orthogonal to each other. But for a special type of matrix, symmetric matrix, the eigenvalues are always real and the corresponding eigenvectors are always orthogonal!

Proof 6.45.2. Let us take a real symmetric matrix M and two distinct eigenvalues of M, λ_1 and λ_2, such that $M\vec{x}_1=\lambda_1\vec{x}_1$ and:

$$M\vec{x}_1=\lambda_1\vec{x}_1 \quad \text{and} \quad M\vec{x}_2=\lambda_2\vec{x}_2 \tag{6.6.467}$$

where $\vec{x}_1$ and $\vec{x}_2$ are obviously eigenvectors.

From $M\vec{x}_1=\lambda_1\vec{x}_1$, we get:

$$\vec{x}_2^TM\vec{x}_1=\lambda_1\vec{x}_2^T\vec{x}_1 \tag{6.6.468}$$

From $M\vec{x}_2=\lambda_2\vec{x}_2$, we similarly get:

$$\vec{x}_1^TM\vec{x}_2=\lambda_2\vec{x}_1^T\vec{x}_2 \tag{6.6.469}$$

But since M is symmetric:

$$\vec{x}_2^TM\vec{x}_1=\vec{x}_2^TM^T\vec{x}_1=\left(M\vec{x}_2\right)^T\vec{x}_1=\vec{x}_1^T\left(M\vec{x}_2\right)=\vec{x}_1^TM\vec{x}_2 \tag{6.6.470}$$

Also, clearly $\vec{x}_1^T\vec{x}_2=\vec{x}_2^T\vec{x}_1$. Thus:

$$\lambda_1\vec{x}_1^T\vec{x}_2=\lambda_2\vec{x}_1^T\vec{x}_2\,(\lambda_1-\lambda_2)\left(\vec{x}_1^T\vec{x}_2\right)=0 \tag{6.6.471}$$

But $\lambda_1-\lambda_2\neq 0$. Hence:

$$\vec{x}_1^T\vec{x}_2=\vec{x}_1^T\circ\vec{x}_2=0 \tag{6.6.472}$$

as required.

■ Q.E.D.

Now we treat the second step, for this let us consider A be a symmetric $n\times n$ matrix. Then the maximum value of $|A\vec{x}|$, where $\vec{x}$ ranges over unit vectors in $\mathbb{R}^n$, is the largest singular value σ_1, and this is achieved when $\vec{x}$ is an eigenvector of A^TA with eigenvalue σ_1^2.

Proof 6.45.3. Let $\vec{v}_1,\ldots,\vec{v}_n$ be an orthonormal basis for $\mathbb{R}^n$ consisting of eigenvectors of A^TA (indeed as A^TA is symmetric as $(AA^T)^T=(A^T)^TA^T=AA^T$ - this even if A is not symmetric itself - then according to the result just seen below, the eigenvectors are orthogonal) with eigenvalues σ_i^2. If $\vec{x}\in\mathbb{R}^n$, then we can expand $\vec{x}$ in this basis as:

$$\vec{x}=c_1\vec{v}_1+\ldots+c_n\vec{v}_n \tag{6.6.473}$$

for scalars $\{c_1,\ldots,c_n\}\in\mathbb{R}$ since $\vec{x}$ is a unit vector, $\|\vec{x}\|^2=1$, which (since the vectors $\vec{v}_1,\ldots,\vec{v}_n$ are orthonormal) means that:

$$c_1^2+\ldots+c_n^2=1 \tag{6.6.474}$$

On the other hand:

$$\|A\vec{x}\|^2=(A\vec{x})\circ(A\vec{x})=(A\vec{x})^T\circ(A\vec{x})=\vec{x}^T\circ A^TA\vec{x}=\vec{x}\circ\left(A^TA\vec{x}\right) \tag{6.6.475}$$

By:

$$\vec{x}=c_1\vec{v}_1+\ldots+c_n\vec{v}_n \tag{6.6.476}$$

since the $\vec{v}_i$ are eigenvectors of A^TA with eigenvalues σ_i^2, we have:

$$A^TA\vec{x}=c_1\sigma_1^2\vec{v}_1+\ldots+c_n\sigma_n^2\vec{v}_n \tag{6.6.477}$$

Taking the dot product with:

$$\vec{x}=c_1\vec{v}_1+\ldots+c_n\vec{v}_n \tag{6.6.478}$$

and using the fact that the vectors $\vec{v}_1,\ldots,\vec{v}_n$ are orthonormal, we get:

$$\vec{x}\circ\left(A^TA\vec{x}\right)=\|A\vec{x}\|^2=\sigma_1^2c_1^2+\ldots+\sigma_n^2c_n^2 \tag{6.6.479}$$

Since σ_1 is the largest singular value, we get:

$$\|A\vec{x}\|^2\leq\sigma_1^2\left(c_1^2+\ldots+c_n^2\right) \tag{6.6.480}$$

Equality holds when $c_1=1$ and $c_2=\ldots=c_n=0$. Thus the maximum value of $\|A\vec{x}\|^2$ for a unit vector $\vec{x}$ is the eigenvalue σ_1^2, which is achieved when $\vec{x}=\vec{v}_1$ (equality between the vector and one of the eigenvector of A^TA).

■ Q.E.D.

One can similarly show that σ_2 is the maximum of $\|A\vec{x}\|$ where $\vec{x}$ ranges over unit vectors that are orthogonal to $\vec{v}_1$. Likewise, σ_3 is the maximum of $\|A\vec{x}\|$ where $\vec{x}$ ranges over unit vectors that are orthogonal to $\vec{v}_1$ and $\vec{v}_2$; and so forth.

So that's why in the case of symmetric (obviously square) matrices, the SVD traditionally denoted:

$$A = U\Sigma V^T \tag{6.6.481}$$

is then denoted (remember that Λ is the upper case letter of λ, that latter being the traditional notation of eigenvalues):

$$\boxed{A = U\Lambda V^T} \tag{6.6.482}$$

A common question that we can find on Internet forums is: *Why are singular values always non-negative?*

To answer let us take some, non zero singular value σ_i. We can reverse the sign if it is positive. That is, $-\sigma_i = -\sqrt{\lambda_i^2} = -\lambda_i$ where λ_i is an eigenvalue of A^TA corresponding to an eigenvector v_i. That is $A^TAv_i = \lambda_i^2 v_i$.

Who can stop us to write instead $A^TA(-v_i) = \lambda_i^2(-v_i)$? What this means is that we can reverse the sign of a singular value, but then we need to go to the matrix V and reverse the sign of its corresponding eigenvector column.

Hence, there is not a unique way to write $A = UDV^T$. But if we decide that all σ_i are non-negative, then "yes" there is a unique way to write $A = UDV^T$. Of course all σ_i are sorted from largest to smallest (otherwise there would be a bunch of possibilities by permuting any two columns of U and V and their corresponding eigenvalues).

6.6.7.1.3 Orthogonal decomposition (eigenvalue spectral decomposition)

Let us come back to:

$$M = UDV^T \tag{6.6.483}$$

If M is symmetric, than it is a square matrix. Then obviously U and V are also square matrices have the same dimensions as those of M.

If M symmetric, then it is also obviously equal to it's transpose M^T, so if we do the SVD of the transpose we should get the same decomposition as M^T such that:

$$M = UDV^T = M^T = (UDV^T)^T \tag{6.6.484}$$

But as we have proved it during our study of matrix transposition (see page 1951), we have:

$$(UDV^T)^T = (V^T)^T D^T U^T \tag{6.6.485}$$

But we should have the equality:

$$(V^T)^T D^T U^T = UDV^T \tag{6.6.486}$$

Therefore:

$$U = V \tag{6.6.487}$$

So that for symmetric matrices M we have:

$$\boxed{M=UDV^T=UDU^T=VDV^T} \tag{6.6.488}$$

As before V (so was U) is orthogonal but now it is also a square matrix. In this case, as a reminder by definition (the inverse of an orthogonal matrix is by definition equal to its transpose):

$$U^{-1}=U^T \tag{6.6.489}$$

So that finally for symmetric matrices M, we get the special case of SVD, named the "orthogonal decomposition" or "eigenvalue spectral decomposition":

$$\boxed{M=UDV^T=UDU^T=UDU^{-1}} \tag{6.6.490}$$

Also often denoted for the obvious reason already mentioned above:

$$M=U\Lambda V^T=U\Lambda U^T=U\Lambda U^{-1} \tag{6.6.491}$$

Be careful to not make a confusion between Spectral Value Decomposition (SVD), also named Singular Value Decomposition, and Eigenvalue Spectral Decomposition (EVD) that is the special case of SVD but for square symmetric matrices! So that means in order for the SVD of a matrix M to be equal to its eigendecomposition we need M to:

- Have orthonormal eigenvectors (hence be a symmetric real matrix)
- Have positive eigenvalues (colloquially, it must be a real matrix and not "flip" anything)

Notice that also the square root of M (written here as $M^{1/2}$) - such that $M^{1/2}M^{1/2}=M$ - can be easily found to be:

$$M^{1/2}=U\Lambda^{1/2}U^T \tag{6.6.492}$$

Indeed, we can easily check that this equality is correct:

$$M^{1/2}M^{1/2}=U\Lambda^{1/2}U^TU\Lambda^{1/2}U^T=U\Lambda^{1/2}\mathbb{1}\Lambda^{1/2}U^T=U\Lambda U^T=M \tag{6.6.493}$$

Another useful result for later (study of the statistical distribution of the Mahalanobis distance) is the eigenvalue decomposition of the inverse of a matrix M using the relation about invertible matrices proved at page 1951 and the associative property of the matrix product:

$$M^{-1}=\left(U\Lambda U^{-1}\right)^{-1}=\left(U^{-1}\right)^{-1}\Lambda^{-1}U^{-1}=U\Lambda^{-1}U^{-1}=U\Lambda^{-1}U^T=\sum_{k=1}^{n}\lambda^{-1}\vec{u}_i\vec{u}_i^T \tag{6.6.494}$$

6.6.7.2 LU Decomposition

As an alternative to Gaussian elimination, a non-singular matrix A may be written in the form $A=LU$ where L is a lower triangular matrix (all entries above the main diagonal are zero) and all the entries on the main diagonal are unity:

$$\boxed{A=\begin{bmatrix}a_{11}&a_{12}&\cdots&a_{1n}\\a_{21}&a_{22}&\cdots&a_{2n}\\\vdots&\vdots&\ddots&\vdots\\a_{m1}&a_{m2}&\cdots&a_{mn}\end{bmatrix}=LU=\begin{bmatrix}1&0&\cdots&0\\l_{21}&1&\cdots&0\\\vdots&\vdots&1&\vdots\\l_{m1}&l_{m2}&\cdots&1\end{bmatrix}\begin{bmatrix}u_{11}&u_{12}&\cdots&u_{1n}\\0&u_{22}&\cdots&u_{2n}\\\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&u_{mn}\end{bmatrix}}\qquad(6.6.495)$$

This is known as the "LU decomposition" or "LU factorization" or "triangular triangularization".

Remark

LU decomposition is not unique: if $A=LU$, then $A=LDD^{-1}U=(LD)(D^{-1}U)=L'U'$ is again an LU decomposition, if D is a diagonal matrix. An additional assumption $l_{ii}=1$, $\forall i=1,\ldots,n$, (impose that L is a lower triangular unit matrix) guarantees the uniqueness.

For example (three by three case):

$$L=\begin{bmatrix}1&0&0\\l_{21}&1&0\\l_{31}&l_{32}&1\end{bmatrix}\qquad(6.6.496)$$

U is an upper triangular matrix (all entries below the main diagonal are zero). For example (three by three case):

$$U=\begin{bmatrix}u_{11}&u_{12}&u_{13}\\0&u_{22}&u_{23}\\0&0&u_{33}\end{bmatrix}\qquad(6.6.497)$$

The method of LU decomposition involves writing the system $A\vec{x}=\vec{b}$ as $LU\vec{x}=\vec{b}$. If $U\vec{x}$ is written as $\vec{y}$, the system becomes $L\vec{y}=\vec{b}$. Due to the form of L (triangular matrices make the solution straightforward), $\vec{y}$ may be found quickly. Once this is done, the equation $U\vec{x}=\vec{y}$ may be solved for $\vec{x}$.

The tricky bit is probably finding L and U. For a 3×3 matrix A.

Let:

$$\begin{aligned}A=\begin{bmatrix}a_{11}&a_{12}&a_{13}\\a_{21}&a_{22}&a_{23}\\a_{31}&a_{32}&a_{33}\end{bmatrix}&=\begin{bmatrix}1&0&0\\l_{21}&1&0\\l_{31}&l_{32}&1\end{bmatrix}\begin{bmatrix}u_{11}&u_{12}&u_{13}\\0&u_{22}&u_{23}\\0&0&u_{33}\end{bmatrix}\\&=\begin{bmatrix}u_{11}&u_{12}&u_{13}\\u_{11}l_{21}&u_{12}l_{21}+u_{22}&u_{13}l_{21}+u_{23}\\u_{11}l_{31}&u_{12}l_{31}+u_{22}l_{32}&u_{13}l_{31}+u_{23}l_{32}+u_{33}\end{bmatrix}\end{aligned}\qquad(6.6.498)$$

Equating coefficients:

$$\begin{array}{rlrlrl} a_{11}=u_{11} & [1] & a_{12}=u_{12} & [4] & a_{13}=u_{13} & [7] \\ a_{21}=l_{21}u_{11} & [2] & a_{22}=l_{21}u_{12}+u_{22} & [5] & a_{23}=l_{21}u_{13}+u_{23} & [8] \\ a_{31}=l_{31}u_{11} & [3] & a_{32}=l_{31}u_{12}+l_{32}u_{22} & [6] & a_{33}=l_{31}u_{13}+l_{32}u_{23}+u_{33} & [9] \end{array} \tag{6.6.499}$$

These equations may be solved as follows:

- Solve [1] for u_{11}
- Solve [2] for l_{21}
- Solve [3] for l_{31}
- Solve [4] for u_{12}
- Solve [7] for u_{13}
- Solve [5] for u_{22}
- Solve [6] for l_{32}
- Solve [8] for u_{23}
- Solve [9] for u_{33}

This procedure seems to be named the "Crout's algorithm".

As companion example let's solve:

$$\underbrace{\begin{bmatrix} -1 & 1 & -1 \\ -2 & 4 & -2 \\ 1 & 1 & 4 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}}_{\vec{x}} = \underbrace{\begin{bmatrix} -2 \\ -6 \\ 9 \end{bmatrix}}_{\vec{b}} \tag{6.6.500}$$

Let:

$$A = \begin{bmatrix} -1 & 1 & -1 \\ -2 & 4 & -2 \\ 1 & 1 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix} \tag{6.6.501}$$

i.e.:

$$\begin{bmatrix} -1 & 1 & -1 \\ -2 & 4 & -2 \\ 1 & 1 & 4 \end{bmatrix} = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ l_{21}u_{11} & l_{21}u_{12}+u_{22} & l_{21}u_{13}+u_{23} \\ l_{31}u_{11} & l_{31}u_{12}+l_{32}u_{22} & l_{31}u_{13}+l_{32}u_{23}+u_{33} \end{bmatrix} \tag{6.6.502}$$

so, comparing coefficients:

$$\begin{array}{lll} u_{11}=-1 & u_{12}=1 & u_{13}=-1 \\ l_{21}=\frac{-2}{-1}=2 & u_{22}=4-2\cdot 1=2 & u_{23}=-2-2\cdot(-1)=0 \\ l_{31}=\frac{1}{-1}=-1 & l_{32}=\frac{1-(-1)\cdot 1}{2}=1 & u_{33}=4-(-1)\cdot(-1)-1\cdot 0=3 \end{array} \tag{6.6.503}$$

Or written in a more general way for U:

$$\begin{aligned} u_{11} &= a_{11} \\ u_{12} &= a_{12} \\ u_{13} &= a_{13} \\ u_{22} &= a_{22} - u_{12}l_{21} \\ u_{23} &= a_{23} - u_{13}l_{21} \\ u_{33} &= a_{33} - (u_{13}l_{31} + u_{23}l_{32}) \end{aligned} \tag{6.6.504}$$

and for L:

$$\begin{aligned} l_{21} &= \frac{1}{u_{11}}a_{21} \\ l_{31} &= \frac{1}{u_{11}}a_{31} \\ l_{32} &= \frac{1}{u_{22}}(a_{32} - u_{12}l_{31}) \end{aligned} \tag{6.6.505}$$

We see that there is a calculation pattern, which can be expressed as the following relations:

$$\boxed{u_{ij} = a_{ij} - \sum_{k=1}^{i-1} u_{kj}l_{ik} \qquad l_{ij} = \frac{1}{u_{jj}}(a_{ij} - \sum_{k=1}^{j-1} u_{kj}l_{ik})} \tag{6.6.506}$$

Remark

We notice in the second relation that to get the l_{ij} below the diagonal, we have to divide by the diagonal element (pivot) u_{jj} , so we get problems when u_{jj} is either 0 or very small, which leads to numerical instability.

Therefore:

$$A = LU = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} -1 & 1 & -1 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \tag{6.6.507}$$

The equation $LU\vec{x} = \vec{b}$ becomes $L\vec{y} = \vec{b}$ i.e.:

$$\begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} -2 \\ -6 \\ 9 \end{bmatrix} \tag{6.6.508}$$

So (this procedure is named "forward substitution"):

$$\begin{array}{ccccccc} y_1 & & & & & = & -2 \\ 2y_1 & + & y_2 & & & = & -6 \\ -y_1 & + & y_2 & + & y_3 & = & 9 \end{array} \qquad \begin{aligned} y_1 &= -2 \\ y_2 &= -2 \\ y_3 &= 9 \end{aligned} \tag{6.6.509}$$

Where in general we can see that the pattern is:

$$\begin{aligned} y_1 &= \frac{b_1}{l_{11}} \\ y_i &= \frac{1}{l_{ii}}\left(b_i - \sum_{j=1}^{i-1} l_{ij}y_i\right) \qquad i = 2, 3, \ldots, n \end{aligned} \tag{6.6.510}$$

The equation $U\vec{x}=\vec{y}$ becomes:

$$\begin{bmatrix} -1 & 1 & -1 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}=\begin{bmatrix} -2 \\ -2 \\ 9 \end{bmatrix} \tag{6.6.511}$$

So (this procedure is named "backward substitution"):

$$\begin{array}{rcrcrcr} -x_1 & + & x_2 & - & x_3 & = & -2 \\ & & 2x_2 & & & = & -2 \\ & & & & x_3 & = & 9 \end{array} \qquad \begin{array}{l} x_1=-2 \\ x_2=-1 \\ x_3=3 \end{array} \tag{6.6.512}$$

Where in general we can see that the pattern is:

$$\begin{aligned} x_n &= \frac{y_n}{u_{nn}} \\ x_i &= \frac{1}{u_{ii}}\left(y_i-\sum_{j=i+1}^{n} u_{ij}x_j\right) \qquad i=n-1,n-2,\ldots,1 \end{aligned} \tag{6.6.513}$$

For large matrices, this procedure involves fewer operations than Gaussian elimination. However, Gaussian Elimination is probably easier to program.

The matrix U is identical to the triangular matrix found in Gaussian Elimination; $\vec{y}$ is identical to the right hand side immediately before the back-substitution.

LU decomposition can be viewed as the matrix form of Gaussian elimination. Computers usually solve square systems of linear equations using LU decomposition, and it is also a key step when inverting a matrix or computing the determinant of a matrix. LU decomposition was introduced by mathematician Tadeusz Banachiewicz in 11938 (holocene calendar).

6.6.7.3 Cholesky's Decomposition

The Cholesky decomposition is another way of solving systems of linear equations. It can be significantly faster and uses less memory than the LU decomposition by exploiting the property of symmetric matrices. However, it is required that the matrix being decomposed be Hermitian (or real-valued symmetric and thus square) and positive definite.

For many symmetric matrices (strictly speaking for positive definite matrices[30] $\vec{x}^T A\vec{x}>0$ for all $\vec{x}\neq 0$ as detailed at page 1957), LU decomposition can be carried out using $U=L^T$. Therefore:

$$\boxed{A=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}=LU=L(L^T)=LL^T=\begin{bmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \cdots & 0 \\ \vdots & \vdots & l_{33} & \vdots \\ l_{m1} & l_{m2} & \cdots & l_{mn} \end{bmatrix}\begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1n} \\ 0 & l_{22} & \cdots & l_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & l_{mn} \end{bmatrix}} \tag{6.6.514}$$

This is known as the "Cholesky's decomposition". While the Cholesky decomposition only works for symmetric, positive definite matrices, the more general LU decomposition works for any square matrix!

---

[30] Don't forget that the variance-covariance matrix is a positive definite matrix!

As companion example, let us use Cholesky's method to solve:

$$\underbrace{\begin{bmatrix} 4 & -2 & 2 \\ -2 & 17 & -1 \\ 2 & -1 & 10 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}}_{\vec{x}} = \underbrace{\begin{bmatrix} 14 \\ -55 \\ 7 \end{bmatrix}}_{\vec{b}} \tag{6.6.515}$$

Let:

$$A = \begin{bmatrix} 4 & -2 & 2 \\ -2 & 17 & -1 \\ 2 & -1 & 10 \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix} \tag{6.6.516}$$

i.e.:

$$\begin{bmatrix} 4 & -2 & 2 \\ -2 & 17 & -1 \\ 2 & -1 & 10 \end{bmatrix} = \begin{bmatrix} l_{11}^2 & l_{11}l_{21} & l_{11}l_{31} \\ l_{11}l_{21} & l_{21}^2 + l_{22}^2 & l_{21}l_{31} + l_{22}l_{32} \\ l_{11}l_{31} & l_{21}l_{31} + l_{22}l_{32} & l_{31}^2 + l_{32}^2 + l_{33}^2 \end{bmatrix} \tag{6.6.517}$$

So:

$$\begin{aligned} l_{11}^2 &= 4 & [1] \\ l_{11}l_{21} &= -2 & [2] \\ l_{11}l_{31} &= 2 & [3] \\ l_{21}^2 + l_{22}^2 &= 17 & [4] \\ l_{21}l_{31} + l_{22}l_{32} &= -1 & [5] \\ l_{31}^2 + l_{32}^2 + l_{33}^2 &= 10 & [6] \end{aligned} \tag{6.6.518}$$

Plus duplicates of [2], [3] and [5]:

- From [1], $l_{11} = 2$,
- From [2], $l_{21} = -1$,
- From [3], $l_{31} = 1$,
- From [4], $l_{22} = 4$,
- From [5], $l_{32} = 0$,
- From [6], $l_{33} = 3$

So:

$$L = \begin{bmatrix} 2 & 0 & 0 \\ -1 & 4 & 0 \\ 1 & 0 & 3 \end{bmatrix} \tag{6.6.519}$$

The system becomes:

$$\begin{bmatrix} 2 & 0 & 0 \\ -1 & 4 & 0 \\ 1 & 0 & 3 \end{bmatrix} \begin{bmatrix} 2 & -1 & 1 \\ 0 & 4 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 14 \\ -55 \\ 7 \end{bmatrix} \tag{6.6.520}$$

i.e:

$$\begin{bmatrix} 2 & 0 & 0 \\ -1 & 4 & 0 \\ 1 & 0 & 3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 14 \\ -55 \\ 7 \end{bmatrix} \tag{6.6.521}$$

so:

$$\begin{array}{ccccccccc} 2y_1 & & & & & = & 14 & \qquad & y_1=7 \\ -y_1 & + & 4y_2 & & & = & -55 & & y_2=-12 \\ y_1 & & & + & 3y_3 & = & 7 & & y_3=0 \end{array} \tag{6.6.522}$$

$\vec{x}$ satisfies i.e.:

$$\begin{bmatrix} 2 & -1 & 1 \\ 0 & 4 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ -12 \\ 0 \end{bmatrix} \tag{6.6.523}$$

$$\begin{array}{ccccccccc} 2x_1 & - & x_2 & + & x_3 & = & 7 & \qquad & x_1=2 \\ & & 4x_2 & & & = & -12 & & x_2=-3 \\ & & & & 3x_3 & = & 0 & & x_3=2 \end{array} \tag{6.6.524}$$

We can see that from the above example that the diagonal elements l_{kk} of L have a calculation pattern:

$$l_{11}=\sqrt{a_{11}}\quad l_{22}=\sqrt{a_{22}-l_{21}^2}\quad l_{33}=\sqrt{a_{33}-(l_{31}^2+l_{32}^2)} \tag{6.6.525}$$

For the elements below the diagonal (l_{ik} , where $i>k$) there is also a calculation pattern:

$$l_{21}=\frac{1}{l_{11}}a_{21}\quad l_{31}=\frac{1}{l_{11}}a_{31}\quad l_{32}=\frac{1}{l_{22}}(a_{32}-l_{31}l_{21}) \tag{6.6.526}$$

We see that the both cases have a pattern that can be written respectively:

$$\boxed{l_{kk}=\sqrt{a_{kk}-\sum_{j=1}^{k-1}l_{kj}^2}\qquad l_{ik}=\frac{1}{l_{kk}}\left(a_{ik}-\sum_{j=1}^{k-1}l_{ij}l_{kj}\right)} \tag{6.6.527}$$

We can now understand for what the positive definiteness was necessary! Indeed, the positive definiteness of A guarantees that all square roots are real. We can also notice from the above developments that L needs to have positive diagonal entries[31]!

Suppose now a matrix A factors as $A=L^TL$. Then:

$$\begin{aligned} \vec{x}^TA\vec{x} &= \vec{x}^TL^TL\vec{x} \\ &= (L\vec{x})^T(L\vec{x}) \\ &= \|L\vec{x}\|^2 \\ &\geq 0 \end{aligned} \tag{6.6.528}$$

---

[31] The reader should however not forget that from what we have seen during our study of positive definite matrices, the fact that all entries are positive and that the matrix is symmetric is not enough to make it positive definite

This shows that A is positive semidefinite.

If we further assume that L is square and triangular with positive real diagonal entries, then L is invertible, so $L\vec{x}=\vec{0}\Leftrightarrow\vec{x}=\vec{0}$. In this case, we see that A is positive definite.

Some textbooks use Cholesky decomposition to define what makes a matrix positive definite. Indeed, such textbooks only say that any matrix A that can uniquely be written under a Cholesky decomposition $A=L^TL$ are positive definite matrices.

> **Remark**
>
> The above algorithms show that every positive definite matrix A has a Cholesky decomposition. This result can be extended to the positive semi-definite matrix! However for positive semi-definite, it is possible but not for sure the Cholesky decomposition is not unique.

6.6.7.4 QR Decomposition

A "QR decomposition", also named a "QR factorization", of a matrix is a decomposition of any real matrix A into a product $A=QR$ of an orthogonal matrix Q and an upper triangular matrix R:

$$\boxed{\begin{aligned}A=\begin{bmatrix}a_{11}&a_{12}&\cdots&a_{1n}\\a_{21}&a_{22}&\cdots&a_{2n}\\\vdots&\vdots&\ddots&\vdots\\a_{m1}&a_{m2}&\cdots&a_{mn}\end{bmatrix}=QR=\begin{bmatrix}q_{11}&q_{12}&\cdots&q_{1n}\\q_{21}&q_{22}&\cdots&q_{2n}\\\vdots&\vdots&\ddots&\vdots\\q_{m1}&q_{m2}&\cdots&q_{mn}\end{bmatrix}\begin{bmatrix}r_{11}&r_{12}&\cdots&r_{1n}\\0&r_{22}&\cdots&l_{2n}\\\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&r_{mn}\end{bmatrix}\\=[\text{col}_1(Q)|\text{col}_2(Q)|\ldots|\text{col}_n(Q)][\text{col}_1(R)|\text{col}_2(R)|\ldots|\text{col}_n(R)]\end{aligned}} \tag{6.6.529}$$

with $\text{col}_i(Q)\circ\text{col}_j(Q)=0,\forall i\neq j$ and (we will show if further below) $r_{ij}=\text{col}_i(Q)\circ\text{col}_j(A),\forall 1\leq 1\leq i\leq n$.

There are actually several methods to compute the QR decomposition, but we will focus here on the Gram–Schmidt process whose idea is described in details at page 2062.

So let us consider the Gram–Schmidt process applied to the columns of the matrix A and let us recall the result that we have derived during our study of the Gram–Schmidt process:

$$\begin{aligned}\vec{e}_1&=\vec{x}_1\\\vec{e}_2&=\vec{x}_2-\frac{\vec{e}_1\circ\vec{x}_2}{\|\vec{e}_1\|^2}\vec{e}_1=\vec{x}_2-\text{proj}_{e_1}\vec{x}_2\\\vec{e}_3&=-\frac{\vec{e}_1\circ\vec{x}_3}{\|\vec{e}_1\|^2}\vec{e}_1-\frac{\vec{e}_2\circ\vec{x}_3}{\|\vec{e}_2\|^2}\vec{e}_2+\vec{x}_3=-\text{proj}_{e_1}\vec{x}_3-\text{proj}_{e_2}\vec{x}_2+\vec{x}_3\end{aligned} \tag{6.6.530}$$

That we will rewrite in our case:

$$\begin{aligned}
&\vec{u}_1=\text{col}_1(A)\\
&\vec{u}_2=\text{col}_2(A)-\frac{\vec{u}_1\circ\text{col}_2(A)}{\|\vec{u}_1\|^2}\vec{u}_1=\text{col}_2(A)-\text{proj}_{u_1}\text{col}_2(A)\\
&\vec{u}_3=-\frac{\vec{u}_1\circ\text{col}_3(A)}{\|\vec{u}_1\|^2}\vec{e}_1-\frac{\vec{u}_2\circ\text{col}_3(A)}{\|\vec{u}_2\|^2}\vec{u}_2+\text{col}_3(A)=-\text{proj}_{u_1}\text{col}_3(A)-\text{proj}_{u_2}\text{col}_2(A)+\text{col}_3(A)\\
&\quad\vdots\\
&\vec{u}_k=\text{col}_k(A)-\sum_{j=1}^{k-1}\text{proj}_{u_j}\text{col}_k(A)
\end{aligned}\tag{6.6.531}$$

And in the QR decomposition it is traditional to take:

$$\text{col}_i(Q)=\vec{e}_i=\frac{\vec{u}_i}{\|\vec{u}_i\|}\tag{6.6.532}$$

Therefore with this tradition we have (Q is then not only orthogonal but orthonormal!):

$$Q=[\vec{e}_1|\vec{e}_2|\ldots|\vec{e}_n]\tag{6.6.533}$$

Before we continue, let us show a companion example. Let us consider the matrix:

$$A=\begin{bmatrix}-1&3\\1&5\end{bmatrix}\tag{6.6.534}$$

Let us carry out the Gram-Schmidt process with the columns $\text{col}_1(A)$ and $\text{col}_2(A)$:

$$\begin{aligned}
&\vec{u}_1=\text{col}_1(A)=\begin{bmatrix}-1\\1\end{bmatrix}\\
&\vec{e}_1=\frac{\vec{u}_1}{\|\vec{u}_1\|}=\frac{1}{\sqrt{2}}\begin{bmatrix}-1\\1\end{bmatrix}\\
&\vec{u}_2=\text{col}_2(A)-\text{proj}_{u_1}=\text{col}_2(A)-\frac{\vec{u}_1\circ\text{col}_2(A)}{\|\vec{u}_1\|^2}\vec{u}_1=\text{col}_2(A)-(\vec{e}_1\circ\text{col}_2(A))\vec{e}_1\\
&\quad=\begin{bmatrix}3\\5\end{bmatrix}-\frac{(-3+5)}{\sqrt{2}}\frac{1}{\sqrt{2}}\begin{bmatrix}-1\\1\end{bmatrix}=\begin{bmatrix}3\\5\end{bmatrix}-\sqrt{2}\frac{1}{\sqrt{2}}\begin{bmatrix}-1\\1\end{bmatrix}=\begin{bmatrix}3\\5\end{bmatrix}-\begin{bmatrix}-1\\1\end{bmatrix}=\begin{bmatrix}4\\4\end{bmatrix}\\
&\vec{e}_2=\frac{\vec{u}_2}{\|\vec{u}_2\|}=\frac{1}{4\sqrt{2}}\begin{bmatrix}4\\4\end{bmatrix}=\frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}
\end{aligned}\tag{6.6.535}$$

That is how the Gram-Schmidt process produces the matrix Q. Here are one of the methods to find R.

For this the reader must understand (and see!) that at each step of Gram-Schmidt procedure, the operations on the vectors correspond to column operations A, which correspond to multiplying by elementary matrices on the right. Let us write all those matrices:

$$\begin{aligned}
A &= \begin{bmatrix} -1 & 3 \\ 1 & 5 \end{bmatrix} \\
A \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & 1 \end{bmatrix} &= \begin{bmatrix} -\frac{1}{\sqrt{2}} & 3 \\ \frac{1}{\sqrt{2}} & 5 \end{bmatrix} \\
A \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -\sqrt{2} \\ 0 & 1 \end{bmatrix} &= \begin{bmatrix} -\frac{1}{\sqrt{2}} & 4 \\ \frac{1}{\sqrt{2}} & 4 \end{bmatrix} \\
A \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -\sqrt{2} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{4\sqrt{2}} \end{bmatrix} &= \begin{bmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} = Q
\end{aligned} \tag{6.6.536}$$

From this we get (using the properties on the inverse matrices):

$$\begin{aligned}
A &= Q \left(\begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -\sqrt{2} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{4\sqrt{2}} \end{bmatrix} \right)^{-1} \\
&= Q \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{4\sqrt{2}} \end{bmatrix}^{-1} \begin{bmatrix} 1 & -\sqrt{2} \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & 1 \end{bmatrix}^{-1} \\
&= Q \begin{bmatrix} 1 & 0 \\ 0 & 4\sqrt{2} \end{bmatrix} \begin{bmatrix} 1 & \sqrt{2} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 \\ 0 & 1 \end{bmatrix} \\
&= Q \begin{bmatrix} \sqrt{2} & \sqrt{2} \\ 0 & 4\sqrt{2} \end{bmatrix} = QR
\end{aligned} \tag{6.6.537}$$

We have found:

$$R = \begin{bmatrix} \sqrt{2} & \sqrt{2} \\ 0 & 4\sqrt{2} \end{bmatrix} = \sqrt{2} \begin{bmatrix} 1 & 1 \\ 0 & 4 \end{bmatrix} \tag{6.6.538}$$

Let us check the factorization:

$$QR = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} \sqrt{2} \begin{bmatrix} 1 & 1 \\ 0 & 4 \end{bmatrix} = \begin{bmatrix} -1 & 3 \\ 1 & 5 \end{bmatrix} = A \tag{6.6.539}$$

Let us rewrite the equalities using symbols, in order to get the general relation:

$$\begin{aligned}
A &= [\text{col}_1(A), \text{col}_2(A)] = [\vec{u}_1, \text{col}_2(A)] \\
A \begin{bmatrix} \frac{1}{\|\vec{u}_1\|} & 0 \\ 0 & 1 \end{bmatrix} &= \begin{bmatrix} \frac{\vec{u}_1}{\|\vec{u}_1\|} & \text{col}_2(A) \end{bmatrix} = [\vec{e}_1, \text{col}_2(A)] \\
A \begin{bmatrix} \frac{1}{\|\vec{u}_1\|} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -(\vec{e}_1 \circ \text{col}_2(A)) \\ 0 & 1 \end{bmatrix} &= [\vec{e}_1, \text{col}_2(A) - (\vec{e}_1 \circ \text{col}_2(A))\vec{e}_1] = [\vec{e}_1, \vec{u}_1] \\
A \begin{bmatrix} \frac{1}{\|\vec{u}_1\|} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -(\vec{e}_1 \circ \text{col}_2(A)) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{\|\vec{u}_2\|} \end{bmatrix} &= \begin{bmatrix} \vec{e}_1, \frac{\vec{u}_2}{\|\vec{u}_2\|} \end{bmatrix} = [\vec{e}_1, \vec{e}_2] = Q
\end{aligned} \tag{6.6.540}$$

From this we get:

$$\begin{aligned}A&=Q\left(\begin{bmatrix}\dfrac{1}{\|\vec{u}_1\|} & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}1 & -(\vec{e}_1\circ\text{col}_2(A))\\ 0 & 1\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & \dfrac{1}{\|\vec{u}_2\|}\end{bmatrix}\right)^{-1}\\&=Q\begin{bmatrix}1 & 0\\ 0 & \dfrac{1}{\|\vec{u}_2\|}\end{bmatrix}^{-1}\begin{bmatrix}1 & -(\vec{e}_1\circ\text{col}_2(A))\\ 0 & 1\end{bmatrix}^{-1}\begin{bmatrix}\dfrac{1}{\|\vec{u}_1\|} & 0\\ 0 & 1\end{bmatrix}^{-1}\\&=Q\begin{bmatrix}1 & 0\\ 0 & \|\vec{u}_2\|\end{bmatrix}\begin{bmatrix}1 & \vec{e}_1\circ\text{col}_2(A)\\ 0 & 1\end{bmatrix}\begin{bmatrix}\|\vec{u}_1\| & 0\\ 0 & 1\end{bmatrix}\\&=Q\begin{bmatrix}\|\vec{u}_1\| & \vec{e}_1\circ\text{col}_2(A)\\ 0 & \|\vec{u}_2\|\end{bmatrix}=QR\end{aligned}\qquad(6.6.541)$$

This relation for R generalizes to any value of n. For example, if n were 3, R would be given by:

$$R=\begin{bmatrix}\|\vec{u}_1\| & \vec{e}_1\circ\text{col}_2(A) & \vec{e}_1\circ\text{col}_3(A)\\ 0 & \|\vec{u}_2\| & \vec{e}_2\circ\text{col}_2(A)\\ 0 & 0 & \|\vec{u}_3\|\end{bmatrix}\qquad(6.6.542)$$

Knowing this, there is no need to rewrite the whole computation every time. We can just build the matrix R as we go along the Gram-Schmidt process. So more generally:

$$R=\begin{bmatrix}\|\vec{u}_{11}\| & \vec{e}_1\circ\text{col}_2(A) & \cdots & \vec{e}_1\circ\text{col}_n(A)\\ 0 & \|\vec{u}_{22}\| & \cdots & \vec{e}_2\circ\text{col}_n(A)\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \|\vec{u}_{mn}\|\end{bmatrix}\qquad(6.6.543)$$

There is also another way to get R. Indeed, the fact that Q has orthonormal columns can be restated as we know as $Q^TQ=\mathbb{1}$. In particular, Q has a left inverse, namely Q^T. From this find R:

$$A=QR\Rightarrow Q^TA=Q^TQR=R\qquad(6.6.544)$$

In other words, the relation:

$$R=Q^TA\qquad(6.6.545)$$

holds, no matter what the dimensions of the matrix.

If we take back our example, let us recall that we had:

$$A=\begin{bmatrix}-1 & 3\\ 1 & 5\end{bmatrix}\qquad Q=\frac{1}{\sqrt{2}}\begin{bmatrix}-1 & 1\\ 1 & 1\end{bmatrix}\qquad(6.6.546)$$

Then:

$$R=Q^TA=\frac{1}{\sqrt{2}}\begin{bmatrix}-1 & 1\\ 1 & 1\end{bmatrix}\begin{bmatrix}-1 & 3\\ 1 & 5\end{bmatrix}=\frac{1}{\sqrt{2}}\begin{bmatrix}2 & 2\\ 0 & 8\end{bmatrix}=\sqrt{2}\begin{bmatrix}1 & 1\\ 0 & 4\end{bmatrix}\qquad(6.6.547)$$

So to summarize:

Factorization	Restrictions	Properties of Factors
SVD $A_{nm} = U_{nn} D_{nm} V^T_{mm}$	none	U orthogonal V orthogonal D non-negative
	variations: for symmetric A, $A = VCV^T$	
LU $A_{nn} = L_{nn} U_{nn}$	A square, (others)	L full-rank lower triangular U upper triangular
variations:	with partial pivoting, $A = LP$ with full pivoting, $P_1 A P_2 = LU$ $A = LDU$, with D diagonal and $U_{ii} = 1$	
QR $A_{nm} = Q_{nn} R_{nm}$	none	Q orthogonal R upper triangular
	variations: skinny QR for $n > m$, $A = Q_1 R_1$	
Cholesky $A_{nn} = L_{nn} U_{nn}$	A non-negative definite	L full-rank lower triangular U upper triangular
Diagonal (i.e. orthogonal) $A_{nn} = V_{nn} C_{nn} V^T_{nn}$	A symmetric	V orthogonal C diagonal
Square root $A_{nn} = (A^{1/2}_{nn})^2$	A non-negative definite	$A^{1/2}_{nn}$ non-negative definite

Table 6.20: **Matrix Factorizations**

6.6.8 Woodbury Matrix identity and Sherman-Morrison relation

The "Woodbury matrix identity" gives the inverse of an $n \times n$ square matrix A modified by a perturbation term UBV^T:

$$(A+UBV^T)^{-1}=A^{-1}-A^{-1}U(B^{-1}+V^TA^{-1}U)^{-1}V^TA^{-1} \tag{6.6.548}$$

It is a very useful result for our study on the inverse and determinant of a partitioned symmetric matrix (itself useful for the Factor Analysis statistical tool). It has also other practical application like for PPCA (Probabilistic Principal Component Analysis) or for hierarchical linear models!

The proof is straightforward:

Proof 6.45.4.

$$\begin{aligned}
&(A+UBV^T)\left[A^{-1}-A^{-1}U(B^{-1}+V^TA^{-1}U)^{-1}V^TA^{-1}\right]\\
&=\left[\mathbb{1}-U\left(B^{-1}+V^TA^{-1}U\right)^{-1}V^TA^{-1}\right]+\left[UBV^TA^{-1}-UBV^TA^{-1}U\left(B^{-1}+V^TA^{-1}U\right)^{-1}V^TA^{-1}\right]\\
&=\mathbb{1}+UBV^TA^{-1}-\left[U\left(B^{-1}+V^TA^{-1}U\right)^{-1}V^TA^{-1}+UBV^TA^{-1}U\left(B^{-1}+V^TA^{-1}U\right)^{-1}VA^{-1}\right]\\
&=\mathbb{1}+UBV^TA^{-1}-\left(U+UBV^TA^{-1}U\right)\left(B^{-1}+V^TA^{-1}U\right)^{-1}V^TA^{-1}]\\
&=\mathbb{1}+UBV^TA^{-1}-UB\left(B^{-1}+V^TA^{-1}U\right)\left(B^{-1}+V^TA^{-1}U\right)^{-1}V^TA^{-1}\\
&=\mathbb{1}+UBV^TA^{-1}-UBV^TA^{-1}=\mathbb{1}
\end{aligned} \tag{6.6.549}$$

■ Q.E.D.

Consider some special cases:

- If $U=V=\mathbb{1}$, then we get:

$$(A+B)^{-1}=A^{-1}-A^{-1}(A^{-1}+B^{-1})^{-1}A^{-1} \tag{6.6.550}$$

- If $B=\mathbb{1}$, and let $U=[\vec{u}_1,\ldots,\vec{u}_m]$ and $V=[\vec{v}_1,\ldots,\vec{v}_m]$, then we get the inverse of a rank-n modified matrix:

$$\left(A+UV^T\right)^{-1}=\left(A+\sum_{i=1}^{m}\vec{u}_i\vec{v}_i^T\right)^{-1}=A^{-1}-A^{-1}U\left(\mathbb{1}+V^TA^{-1}U\right)^{-1}V^TA^{-1} \tag{6.6.551}$$

- More specially, when $m=1$, $U=\vec{u}$ and $V=\vec{v}$ we get the inverse of rang-1 modified matrix:

$$(A+\vec{u}\vec{v}^T)^{-1}=A^{-1}-\frac{A^{-1}\vec{u}\vec{v}^TA^{-1}}{1+\vec{v}^TA^{-1}\vec{u}} \tag{6.6.552}$$

Let us see two possible derivations of that later case! First possible proof:

Proof 6.45.5. We first show the following is an identity ($A=\mathbb{1}$):

$$(\mathbb{1}+\vec{w}\vec{v}^T)^{-1}=\mathbb{1}-\frac{\vec{w}\vec{v}^T}{1+\vec{v}^T\vec{w}} \tag{6.6.553}$$

Pre-multiplying $\mathbb{1}+\vec{w}\vec{v}^T$, the right side becomes $\mathbb{1}$ as well as the left side:

$$\mathbb{1}+\vec{w}\vec{v}^T-\frac{\vec{w}\vec{v}^T+\vec{w}\vec{v}^T\vec{w}\vec{v}^T}{1+\vec{v}^T\vec{w}}=\mathbb{1}+\vec{w}\vec{v}^T-\frac{\vec{w}\left(1+\vec{v}^T\vec{w}\right)\vec{v}^T}{1+\vec{v}^T\vec{w}}=\mathbb{1}+\vec{w}\vec{v}^T-\vec{w}\vec{v}^T=\mathbb{1} \tag{6.6.554}$$

We next let $\vec{u}=A\vec{w}$ in:

$$(A+\vec{u}\vec{v}^T)^{-1}=A^{-1}-\frac{A^{-1}\vec{u}\vec{v}^TA^{-1}}{1+\vec{v}^TA^{-1}\vec{u}} \tag{6.6.555}$$

and the left side of the relation to be proven becomes:

$$\begin{aligned}\left(A+\vec{u}\vec{v}^T\right)^{-1}&=\left(A+A\vec{w}\vec{v}^T\right)^{-1}=\left(A\left(\mathbb{1}+\vec{w}\vec{v}^T\right)\right)^{-1}=\left(\mathbb{1}+\vec{w}\vec{v}^T\right)^{-1}A^{-1}\\&=\left(\mathbb{1}-\frac{\vec{w}\vec{v}^T}{1+\vec{v}^T\vec{w}}\right)A^{-1}\end{aligned} \tag{6.6.556}$$

Substituting $\vec{w}=A^{-1}\vec{u}$ we get the "Sherman-Morrison relation":

$$\left(A+\vec{u}\vec{v}^T\right)^{-1}=\left(\mathbb{1}-\frac{A^{-1}\vec{u}\vec{v}^T}{1+\vec{v}^TA^{-1}\vec{u}}\right)A^{-1}=A^{-1}-\frac{A^{-1}\vec{u}\vec{v}^TA^{-1}}{1+\vec{v}^TA^{-1}\vec{u}} \tag{6.6.557}$$

■ Q.E.D.

And second possible proof:

Proof 6.45.6. This second proof is based on the following problem. Assuming a linear equation system $A\vec{y}=\vec{b}$ is solve to get $\vec{y}=A^{-1}\vec{b}$, we want to solve this system:

$$(A+\vec{u}\vec{v}^T)\vec{x}=\vec{b} \tag{6.6.558}$$

We first pre-multiply both sides of this equation by A^{-1} to get:

$$A^{-1}\left(A+\vec{u}\vec{v}^T\right)\vec{x}=\left(\mathbb{1}+A^{-1}\vec{u}\vec{v}^T\right)\vec{x}=\vec{x}+A^{-1}\vec{u}\vec{v}^T\vec{x}=A^{-1}\vec{b}=\vec{y} \tag{6.6.559}$$

If we define $\vec{w}=A^{-1}\vec{u}$ and $\alpha=\vec{v}^T\vec{x}$, the above equation can be written as:

$$\vec{x}+\vec{w}\alpha=\vec{y} \tag{6.6.560}$$

Pre-multiplying both sides by $\vec{v}^T$, we get:

$$\vec{v}^T\vec{x}+\vec{v}^T\vec{w}\alpha=\vec{v}^T\vec{y},\qquad\text{i.e.}\qquad\alpha+\vec{v}^TA^{-1}\vec{u}\alpha=\vec{v}^T\vec{y} \tag{6.6.561}$$

Solving for α we get:

$$\alpha=\frac{\vec{v}^T\vec{y}}{1+\vec{v}^TA^{-1}\vec{u}} \tag{6.6.562}$$

Substituting α into the previous equation for $\vec{x}$ we get:

$$\vec{x}=\vec{y}-\vec{w}\alpha=\vec{y}-\frac{\vec{w}\vec{v}^T\vec{y}}{1+\vec{v}^TA^{-1}\vec{u}}=A^{-1}\vec{b}-\frac{A^{-1}\vec{u}\vec{v}^TA^{-1}\vec{b}}{1+\vec{v}^TA^{-1}\vec{u}}=\left(A^{-1}-\frac{A^{-1}\vec{u}\vec{v}^TA^{-1}}{1+\vec{v}^TA^{-1}\vec{u}}\right)\vec{b} \tag{6.6.563}$$

But solving the equation $(\vec{A}+\vec{u}\vec{v}^T)\vec{x}=\vec{b}$ we also get:

$$\vec{x}=(A+\vec{u}\vec{v}^T)^{-1}\vec{b} \tag{6.6.564}$$

we therefore have:

$$(A+\vec{u}\vec{v}^T)^{-1}=A^{-1}-\frac{A^{-1}\vec{u}\vec{v}^TA^{-1}}{1+\vec{v}^TA^{-1}\vec{u}} \tag{6.6.565}$$

■ Q.E.D.

~95 %

★★★★☆

36 votes, 79.94%

6.7. Tensor Calculus

THE conventional vector calculus is a simple and effective technique that adapts perfectly to the study of mechanical and physical properties of matter in an Euclidean space of three dimensions. However, in many fields of physics, it appears experimental quantities that can't be easily represented by simple column vectors of Euclidean vector spaces. This is the case for example in continuum mechanics (fluids or solids), electromagnetism, General Relativity, etc.

Thus, since the late 119th century (holocene calendar), the analysis of forces acting within a continuous medium led to highlight the physical quantities characterized by nine numbers representing the pressure forces or internal stress (see section Continuum Mechanics page 3734 for the details). The representation of these quantities required the introduction of a new mathematical tool who was named "tensor", by reference to its physical origin. Subsequently, starting the years 11900s (holocene calendar), it was R. Ricci and T. Levi-Civita who developed the tensor calculus; then the study of tensor allowed a deepening of the theory of vector spaces and contributed to the development of differential geometry (see section of the same name page 2850).

Tensor calculus, also sometimes named "absolute differential geometry" also has the advantage to free itself from all coordinate systems and the results of the mathematical developments are thus invariant (huge simplification in calculations but in compensation we have a huge increase in abstraction and notation complexity). We therefore don't need to be concerned in what type of referential frame we work and this is very interesting in General Relativity.

We advise strongly the reader to very well master the basics of vector calculus and linear algebra as they have been presented before in previous sections (especially because linear algebra forms the skeleton of tensor calculus!). If necessary, we have chosen when writing this section to come back on certain points seen in the section of Vector Calculus and Linear Algebra (covariant components, contravariant components, etc.).

Furthermore, if the reader has already covered the study of constraints in solids (see section Continuum Mechanics page 3735) or of the Faraday tensor (see section Electrodynamics page 4121) or the energy-momentum tensor (see section General Relativity page 5754) this will be a practical advantage before reading what follows. Furthermore, the redaction of the above items (tensors) was made so that the concept of tensor is introduced if possible (...) intuitively.

We will only do a very few practical examples in this section. Indeed the examples, you have probably already guess it..., will come when we will study the continuum mechanics, General Relativity, quantum field theory, electrodynamics, etc.

A tip perhaps: you saw several times in the Statistics section that writing with vectors and then thinking with matrices was a powerful way to generalize certain important results. For this section on tensors, remember that the idea is the same but we write with matrices and we think with tensors (you will understand this little adage better once you have finished reading this entire section)!

6.7.1 Tensor

Definition (simplistic #263): The "tensors" are mathematical objects often wrongly represented (but this is for pedagogical reasons in engineering courses) as generalizing the notions of vectors and matrices. They were initially introduced in physics to represent the state of stress and deformation of a volume subjected to forces, hence their name (tensions).

The rigorous definition, on the other hand, requires (at least we personally think so...) having first read this section in its entirety. But for the impatient, know that in fact a tensor is roughly like the determinant of a square matrix (see section Linear Algebra page 1961) or like the dot product of vectors (see section Linear Algebra page 1821)... Yes! It is simply a multilinear map T on a space of given dimension (corresponding to the number of columns of the matrix/tensors) which finally gives a scalar (of a given field). For example in the bilinear case a tensor (then named "tensor of order 2") is defined by:

$$\begin{aligned}
&T(\vec{u},\vec{v})\in\mathbb{R}\\
&T(\lambda\vec{u},\mu\vec{v})=\lambda\mu T(\vec{u},\vec{v})\\
&T(\vec{u}+\vec{w},\vec{v})=T(\vec{u},\vec{v})+T(\vec{w},\vec{v})
\end{aligned}\tag{6.7.1}$$

For example for two vectors of dimension two we have in an orthonormal basis using the property of linearity and multiplication by a scalar of the definition of a tensor:

$$\begin{aligned}
T(\vec{u},\vec{v})&=T\left(u^x\vec{e}_x+u^y\vec{e}_y,v^x\vec{e}_x+v^y\vec{e}_y\right)\\
&=T\left(u^x\vec{e}_x,v^x\vec{e}_x+v^y\vec{e}_y\right)+T\left(u^y\vec{e}_y,v^x\vec{e}_x+v^y\vec{e}_y\right)\\
&=T\left(u^x\vec{e}_x,v^x\vec{e}_x\right)+T\left(u^x\vec{e}_x,v^y\vec{e}_y\right)+T\left(u^y\vec{e}_y,v^x\vec{e}_x\right)+T\left(u^x\vec{e}_y,v^y\vec{e}_y\right)\\
&=u^xv^x\underbrace{T(\vec{e}_x,\vec{e}_x)}_{T_{xx}=1}+u^xv^y\underbrace{\cancel{T(\vec{e}_x,\vec{e}_y)}}_{T_{xy}=0}+u^yv^x\underbrace{T(\vec{e}_y,\vec{e}_x)}_{T_{yx}=0}+u^yv^y\underbrace{T(\vec{e}_y,\vec{e}_y)}_{T_{yy}=1}\\
&=u^xv^x
\end{aligned}\tag{6.7.2}$$

And we see that we fall back on the expression of the dot product. And in practice it is unfortunately the matrix (in the bilinear case otherwise it is no longer a matrix!):

$$\begin{bmatrix}T_{xx}&T_{xy}\\T_{yx}&T_{yy}\end{bmatrix}$$

which is often wrongly named (for language simplification purposes in engineering courses) the "tensor" instead of the "matrix associated with the tensorial application T".

For example, we have proved in the section of Continuum Mechanics that normal and tangential forces in a fluid were given by the relation:

$$\begin{bmatrix}N_x\\N_y\\N_z\end{bmatrix}\begin{bmatrix}\sigma_{xx}&\tau_{xy}&\tau_{xz}\\\tau_{yx}&\sigma_{yy}&\tau_{yz}\\\tau_{zx}&\tau_{zy}&\sigma_{zz}\end{bmatrix}\begin{bmatrix}l\\m\\n\end{bmatrix}\tag{6.7.3}$$

what was noted in the traditional condensed form as following (where we no longer distinguish what is tangential to what is normal so there is a loss of clarity):

$$N_i=\sigma_{ij}n_j \tag{6.7.4}$$

We thus make appear a mathematical quantity σ_{ij} with 9 components, while a vector in the same space $\mathbb{R}^3$ has 3 components.

This notion is also much used in the section of General Relativity where we have proved that the energy-momentum tensor in a particularly simple case was given by:

$$T_{ij}=\begin{bmatrix} T^{00} & T^{01} & T^{02} & T^{03} \\ T^{10} & T^{11} & T^{12} & T^{13} \\ T^{20} & T^{21} & T^{22} & T^{23} \\ T^{30} & T^{31} & T^{32} & T^{33} \end{bmatrix}=\begin{bmatrix} \rho\gamma^2c^2 & \rho\gamma^2cv^x & \rho\gamma^2cv^y & \rho\gamma^2cv^z \\ \rho\gamma^2cv^x & \rho\gamma^2v^xv^x & \rho\gamma^2v^xv^y & \rho\gamma^2v^xv^z \\ \rho\gamma^2cv^y & \rho\gamma^2v^yv^x & \rho\gamma^2v^yv^y & \rho\gamma^2v^yv^z \\ \rho\gamma^2cv^z & \rho\gamma^2v^zv^x & \rho\gamma^2v^zv^y & \rho\gamma^2v^zv^z \end{bmatrix} \tag{6.7.5}$$

and satisfies the non-less important equation of conservation:

$$\nabla_\mu T^{\mu\nu}=0 \tag{6.7.6}$$

Or otherwise, still in the section of General Relativity, we have shown that the tensor of the Schwarzschild metric was given by:

$$g^{\mu\nu}=\begin{bmatrix} 1/A(r) & 0 & 0 & 0 \\ 0 & -1/B(r) & 0 & 0 \\ 0 & 0 & -1/r^2 & 0 \\ 0 & 0 & 0 & -1/(r^2\sin^2(\theta)) \end{bmatrix} \tag{6.7.7}$$

and therefore gives us the "equation of the metric" or "invariant interval" (see section Special Relativity page 5704):

$$\mathrm{d}s^2=g_{\mu\nu}\mathrm{d}x^\mu\mathrm{d}x^\nu \tag{6.7.8}$$

Note also that in the section of Special Relativity we have shown that the Lorentz transformation tensor is given by:

$$\begin{pmatrix} x'^0 \\ x'^1 \\ x'^2 \\ x'^3 \end{pmatrix}=\begin{bmatrix} \gamma & -\gamma\beta & 0 & 0 \\ -\gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}\begin{pmatrix} x^0 \\ x^1 \\ x^2 \\ x^3 \end{pmatrix} \tag{6.7.9}$$

which in a condensed form gives the following components transformation:

$$x'^\mu=L^\mu_\nu x^\nu \tag{6.7.10}$$

As regards the transformation of the electromagnetic field we have also proved that the Faraday tensor is given by:

$$F^{\mu\nu}=\begin{bmatrix} 0 & -E_x & -E_y & -E_z \\ E_x & 0 & -B_z & B_y \\ E_y & B_z & 0 & -B_x \\ E_z & -B_y & B_x & 1 \end{bmatrix} \tag{6.7.11}$$

and therefore permits switch from one repository frame to another using the relation:

$$F'^{\mu\nu}=L_{\alpha}^{\mu}L_{\beta}^{\nu}F^{\alpha\beta} \tag{6.7.12}$$

But these are very simple tensor that can be represented in the form of matrices. You should also remember that it is not because you are reading a variable with indices suggests that we are dealing with a tensor that it is necessarily one. For example, the famous relation (widely used in the section of General Relativity and we will prove far further below):

$$\Gamma_{kj}^{i}=g^{il}\Gamma_{klj}=\frac{1}{2}g^{il}\left(\partial_k g_{lj}+\partial_j g_{kl}-\partial_l g_{kj}\right) \tag{6.7.13}$$

might suggest that the first member of the far left is a tensor but in fact it is not... this is just a symbol... hence its name: Christoffel symbol (not: Christoffel tensor).

The interest of tensors in physics is that their characteristics are independent of the chosen coordinates. Thus, a relation between tensors in a base will be true regardless of the base used thereafter. This is a fundamental and powerful characteristic of General Relativity (among others)!

6.7.2 Indicial Notation

We will use thereafter many mathematical symbols: coordinates, components of vectors and tensors, matrix components, etc., whose number in each category is large or indeterminate. To distinguish the various symbols of a category we use indices. For example, instead of the traditional variables x, y, z we will use the variables x_1, x_2, x_3 (as we have already done in the section of Linear Algebra). This rating becomes essential when we an undetermined number of variables.

Thus, if we have n variables, we denote them by $x_1, x_2, \ldots, x_n$.

We will also use superscripts, when required; eg x^1, x^2, x^3. To avoid confusion with writing powers, the quantity x^i to the power p will be written $(x^i)^p$. When the context eliminates any potential ambiguity, the use of parentheses however is not fundamentally necessary.

In tensor calculus there is a summation convention using the fact that the repeated index, below for example the index i, will become itself an indication of the summation. We write then, with this convention:

$$\sum_{i=1}^{n} x^i y^i = x^i y^i = x^1 y^1 + x^2 y^2 + \ldots + x^n y^n \tag{6.7.14}$$

thereby this condense relatively well the notations!

Thus, to represent the linear system:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \end{aligned} \tag{6.7.15}$$

we will write (notice carefully how are written the components of the associated matrix!):

$$a_{ij}x_j = b_i \tag{6.7.16}$$

specifying that's for $n = 3$.

We see in this example, how the summation convention allows a condensed and thus powerful writing.

The summation convention covers all the mathematical symbols having repeating indices. Thus the decomposition of a vector $\vec{x}$ on a basis $(\vec{e}_1, \vec{e}_2, \vec{e}_3)$ will be therefore written for $n = 3$:

$$\vec{x} = x_1\vec{e}_1 + x_2\vec{e}_2 + x_3\vec{e}_3 = x^i\vec{e}_i \tag{6.7.17}$$

In summary, any term that has a repeated index represents a sum over all possible values of the repeated index.

Remark

We name, for obvious reasons we will detail below, the x^i "contravariant component" of the vector $\vec{x}$.

6.7.2.1 Summation on multiple index

The summation convention (due to Einstein) extends to the case where we have, in general, several repeated indices in upper and lower positions so-named "dumb indices" in the same monomial (often physicists omit the rule of set their position opposite as it will be the case often on this textbook too!). Thus, for example, the quantity $A_i^j x^i x^j$, represents the following sum for i and j taking the values from 1 to 2:

$$A_i^j x^i x^j = A_1^1 x^1 x^1 + A_2^1 x^1 x^2 + A_1^2 x^2 x^1 + A_2^2 x^2 x^2 \qquad (6.7.18)$$

Thus we see easily that an expression with two summation indices that take values respectively $1, 2, \ldots, n$, will have n^2; n^3 if there are three summation indices, etc.

However, we must be careful to substitutions with this kind of notation because if we assume that we have the relation:

$$A = a_{ij} x^i y_j \quad \text{and} \quad x^i = c_{ij} y^j \qquad (6.7.19)$$

then to get the expression of A only in function of the variables y^j we cannot write:

$$A = a_{ij} c_{ij} y^i y_j \qquad (6.7.20)$$

because it does not return to the same expression as the dumb indices after development are systematically sum in an identical and rigid way (we leave to the reader make a little application case to see this, if need you can contact us and we will do an example). In other words, a dumb index can not be repeated more than 2 times.

6.7.2.2 Kronecker Symbol

This symbol introduced by the mathematician Kronecker, is the following (often used in physics and in many other fields):

$$\delta_{ij} = \delta^{ij} = \delta_i^j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \qquad (6.7.21)$$

This symbol is named "Kronecker symbol". It conveniently allows you to write, for example, the dot product of two vectors $\vec{e}_1$ and $\vec{e}_2$, of unit norm and orthogonal to each other, in the form:

$$\vec{e}_i \circ \vec{e}_j = \delta_{ij} \qquad (6.7.22)$$

We will find this symbol in many examples of theoretical physics in this textbook (wave quantum physics, quantum field theory, general relativity, fluid mechanics, etc.).

It should be noted that there is a generalized version of the Kronecker symbol:

$$\delta_{mn..p}^{ij..l} = \begin{bmatrix} \delta_m^i & \delta_n^i & \cdots & \delta_p^i \\ \delta_m^j & \delta_n^j & \cdots & \delta_p^j \\ \vdots & \vdots & \ddots & \vdots \\ \delta_m^l & \delta_n^l & \cdots & \delta_p^l \end{bmatrix} \qquad (6.7.23)$$

We have also, for example:

$$\delta^{ijk}_{mnp}=\begin{bmatrix}\delta^i_m & \delta^i_n & \delta^i_p\\ \delta^j_m & \delta^j_n & \delta^j_p\\ \delta^k_m & \delta^k_n & \delta^k_p\end{bmatrix}=\varepsilon^{ijk}\varepsilon_{mnp} \qquad (6.7.24)$$

where ε_{ijk} is the Levi-Civita symbol that will define right now:

6.7.2.3 Antisymmetric Symbol (Levi-Civita symbol)

Another useful symbol is the "symbol of antisymmetry" or also named "antisymmetry tensor" that we will find in the sections of Electrodynamics, General Relativity and Relativistic Quantum Physics in this textbook.

In mathematics, particularly in linear algebra, tensor analysis, and differential geometry, the Levi-Civita symbol represents a collection of numbers; defined from the sign of a permutation of the natural numbers $1,2,\ldots,n$, for some positive integer n.

The values of the Levi-Civita symbol are independent of any metric tensor and coordinate system. Also, the specific term "symbol" emphasizes that it is not a tensor because of how it transforms between coordinate systems, however it can be interpreted as an antisymmetric tensor (a tensor is antisymmetric on (or with respect to) an index subset if it alternates sign (+/-) when any two indices of the subset are interchanged).

In the case $n=2$ the Levi-Civita symbol is defined by:

$$\varepsilon_{ij}=\begin{cases}+1 & \text{if } (i,j) \text{ is } (1,2)\\ -1 & \text{if } (i,j) \text{ is } (2,1)\\ 0 & \text{if } i=j\end{cases} \qquad (6.7.25)$$

The values can be arranged into a 2×2 antisymmetric matrix (we can see we fall back on the definition of an antisymmetric tensor):

$$\begin{pmatrix}\varepsilon_{11} & \varepsilon_{12}\\ \varepsilon_{21} & \varepsilon_{22}\end{pmatrix}=\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix} \qquad (6.7.26)$$

Use of the 2D symbol is relatively uncommon, although in certain specialized topics like supersymmetry and twistor theory it appears in the context of 2-spinors. The 3D and higher-dimensional Levi-Civita symbols are used more commonly.

In three dimensions, the Levi-Civita symbol is defined as follows:

$$\varepsilon_{ijk}=\begin{cases}+1 & \text{if } (i,j,k) \text{ is } (1,2,3),(2,3,1) \text{ or } (3,1,2),\\ -1 & \text{if } (i,j,k) \text{ is } (3,2,1),(1,3,2) \text{ or } (2,1,3),\\ 0 & \text{if } i=j \text{ or } j=k \text{ or } k=i\end{cases} \qquad (6.7.27)$$

i.e. ε_{ijk} is 1 if (i,j,k) is an even permutation of $(1,2,3)$ or in the natural order $(1,2,3)$, -1 if it is an odd permutation, and 0 if any index is repeated. In three dimensions only, the cyclic permutations of $(1,2,3)$ are all even permutations, similarly the anti-cyclic permutations are all odd permutations. This means in 3D it is sufficient to take cyclic or anti-cyclic permutations of $(1,2,3)$ and easily obtain all the even or odd permutations.

It can also be express with the Kronecker symbol:

$$\varepsilon_{ijk}=\begin{bmatrix}\delta_{i1} & \delta_{i2} & \delta_{i3}\\ \delta_{j1} & \delta_{j2} & \delta_{j3}\\ \delta_{k1} & \delta_{k2} & \delta_{k3}\end{bmatrix} \tag{6.7.28}$$

An illustrative representation gives:

$$\varepsilon_{ijk}=$$

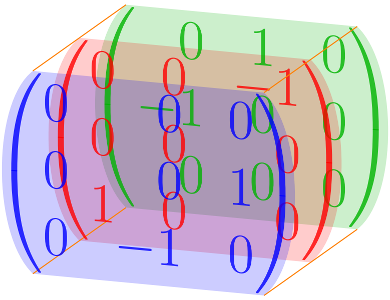


Figure 6.154: **3D Levy-Civita symbol illustration**

By using this symbol, a determinant of order two (see section Linear Algebra page 1961) is then written in the advantageous form:

$$\det(A)=\det(a_{ij})=\varepsilon_{ij}a_1^ia_2^j \tag{6.7.29}$$

and the vector cross product:

$$\boxed{\vec{x}\times\vec{y}=\varepsilon_{ijk}x_jy_k} \tag{6.7.30}$$

where of course, j and k are summed and where the dummy index i is the line number of the resulting vector (if requested we will make the developments). In particular, the rotational (curl) of a vector field (see section Vector Calculus page 1888) is then:

$$\boxed{\vec{\nabla}\times\vec{A}=\varepsilon_{ijk}\frac{\partial A_k}{\partial x_j}=\varepsilon_{ijk}\partial_jA_k} \tag{6.7.31}$$

As an example, let us calculate in index notation the double vector product $\vec{A}\times\vec{B}\times\vec{C}$:

$$\begin{aligned}\vec{A}\times\vec{B}\times\vec{C}&=\varepsilon_{ijk}\varepsilon_{kmn}A_jB_mC_n=(\delta_{im}\delta_{jn}-\delta_{in}\delta jm)A_jB_mC_n=B_i(A_jC_j)-C_i(A_jB_j)\\&=\vec{B}(\vec{A}\circ\vec{C})-\vec{C}(\vec{A}\circ\vec{B})\end{aligned} \tag{6.7.32}$$

where again, the dummy index i is the line number of the resulting vector. Let us see detailed demonstration of these equalities (the order of the equalities below does not need to follow the sequence of equalities of the previous relation).

Proof 6.45.7. We have proved in the section of Vector Calculus the following identity:

$$\vec{A}\times\vec{B}\times\vec{C}=\vec{B}(\vec{A}\circ\vec{C})-\vec{C}(\vec{A}\circ\vec{B}) \tag{6.7.33}$$

Remark

The latter relation is sometimes, and for recall, named the "Grassmann rule" or more commonly "dual vector product".

To prove the relation:

$$\vec{A}\times\vec{B}\times\vec{C}=\varepsilon_{ijk}\varepsilon_{kmn}A_jB_mC_n \tag{6.7.34}$$

to a change of indices let us first prove that:

$$(\vec{B}\times\vec{C})_i=\varepsilon_{ijk}B_jC_k \tag{6.7.35}$$

which give us:

$$\begin{pmatrix}B_2C_3-B_3C_2\\B_3C_1-B_1C_3\\B_1C_2-B_2C_1\end{pmatrix}=\begin{pmatrix}\varepsilon_{1jk}B_jC_k\\\varepsilon_{2jk}B_jC_k\\\varepsilon_{3jk}B_jC_k\end{pmatrix} \tag{6.7.36}$$

We do the development only for the first line (this is already bor... euh long enough...):

$$\begin{aligned}&B_2C_3-B_3C_2=\varepsilon_{1jk}B_jC_k=\varepsilon_{11k}B_1C_k+\varepsilon_{12k}B_2C_k+\varepsilon_{13k}B_3C_k\\&=\varepsilon_{111}B_1C_1+\varepsilon_{112}B_1C_2+\varepsilon_{113}B_1C_3+\varepsilon_{121}B_2C_1+\varepsilon_{122}B_2C_2+\varepsilon_{123}B_2C_3+\varepsilon_{131}B_3C_1\\&+\varepsilon_{132}B_3C_2+\varepsilon_{133}B_3C_3=\varepsilon_{123}B_2C_3+\varepsilon_{132}B_3C_2=1B_2C_3-1B_3C_2=B_2C_3-B_3C_2\end{aligned} \tag{6.7.37}$$

This is the first step that was necessary to be proven.

Now let us prove that for the nth line we have well:

$$(\vec{A}\times(\vec{B}\times\vec{C})_i)_l=\varepsilon_{lni}A_n(\varepsilon_{ijk}B_jC_k)_i \tag{6.7.38}$$

with the help of a result obtained in the section of Vector Calculus (vector product of three different vectors) we have the first term (the first line of the vector resulting from the calculation):

$$\begin{aligned}&(A_2(B_1C_2-B_2C_1)_3-A_3(B_3C_1-B_1C_2)_2)_1=\varepsilon_{1ni}A_n(\varepsilon_{ijk}B_jC_k)_i\\&=\varepsilon_{1n1}A_n(B_2C_3-B_3C_2)_1+\varepsilon_{1n2}A_n(B_3C_1-B_1C_3)_2+\varepsilon_{1n3}A_n(B_1C_2-B_2C_1)_3\\&=\varepsilon_{111}A_1(B_2C_2-B_3C_2)_1+\varepsilon_{121}A_2(B_2C_2-B_3C_2)_1+\varepsilon_{131}A_3(B_2C_3-B_3C_2)_1\\&+\varepsilon_{112}A_1(B_3C_1-B_1C_3)_2+\varepsilon_{122}A_2(B_3C_1-B_1C_3)_2+\varepsilon_{132}A_3(B_3C_1-B_1C_3)_2\\&+\varepsilon_{113}A_1(B_1C_2-B_2C_1)_3+\varepsilon_{123}A_2(B_1C_2-B_2C_1)_3+\varepsilon_{133}A_3(B_1C_2-B_2C_1)_3\\&=\varepsilon_{123}A_2(B_1C_2-B_2C_1)_3+\varepsilon_{132}A_3(B_3C_1-B_1C_3)_2\\&=1A_2(B_1C_2-B_2C_1)_3-1A_3(B_3C_1-B_1C_3)_2=A_2(B_1C_2-B_2C_1)-A_3(B_3C_1-B_1C_3)\end{aligned} \tag{6.7.39}$$

It is immediate that (i being equal to 1):

$$\begin{aligned}&\vec{A}\times\vec{B}\times\vec{C}=(\vec{A}\times\vec{B}\times\vec{C})_i=B_i(A_jC_j)-C_i(A_jB_j)=\\&=B_1(A_1C_1+A_2C_2+A_3C_3)-C_1(A_1B_1+A_2B_2+A_3B_3)\\&=B_1A_1C_1+B_1A_2C_2+B_1A_3C_3-C_1A_1B_1-C_1A_2B_2-C_1A_3B_3\\&=A_2B_1C_2-A_2B_2C_1-A_3B_3C_1+A_3B_1C_3\\&=A_2(B_1C_2-B_2C_1)-A_3(B_3C_1-B_1C_3)\end{aligned} \tag{6.7.40}$$

Let us show now that for i equal 1 we also have:

$$\begin{aligned}\vec{A}\times\vec{B}\times\vec{C}&=(\vec{A}\times\vec{B}\times\vec{C})_i=(\delta_{im}\delta_{jn}-\delta_{in}\delta_{jm})A_jB_mC_n\\&=A_2(B_1C_2-B_2C_1)-A_3(B_3C_1-B_1C_3)\end{aligned}\tag{6.7.41}$$

Indeed:

$$\begin{aligned}&(\vec{A}\times\vec{B}\times\vec{C})_1=(\delta_{1m}\delta_{jn}-\delta_{1n}\delta_{jm})A_jB_mC_n\\&=(\delta_{1m}\delta_{j1}-\delta_{11}\delta_{jm})A_jB_mC_1+(\delta_{1m}\delta_{j2}-\delta_{12}\delta_{jm})A_jB_mC_2+(\delta_{1m}\delta_{j3}-\delta_{13}\delta_{jm})A_jB_mC_3\\&=(\delta_{1m}\delta_{11}-\delta_{11}\delta_{1m})A_1B_mC_1+(\delta_{1m}\delta_{21}-\delta_{11}\delta_{2m})A_2B_mC_1+(\delta_{1m}\delta_{31}-\delta_{11}\delta_{3m})A_3B_bC_1\\&+(\delta_{1m}\delta_{12}-\delta_{12}\delta_{1m})A_1B_mC_2+(\delta_{1m}\delta_{22}-\delta_{12}\delta_{2m})A_2B_mC_2+(\delta_{1m}\delta_{32}-\delta_{12}\delta_{3m})A_3B_mC_2\\&+(\delta_{1m}\delta_{13}-\delta_{13}\delta_{1m})A_1B_mC_3+(\delta_{1m}\delta_{23}-\delta_{13}\delta_{2m})A_2B_mC_2+(\delta_{1m}\delta_{33}-\delta_{13}\delta_{3m})A_3B_mC_3\\&=(\delta_{1m}-\delta_{1m})A_1B_mC_1-\delta_{2m}A_2B_mC_1-\delta_{3m}A_3B_mC_1+\delta_{1m}A_2B_mC_2+\delta_{1m}A_3B_mC_3\\&=-\delta_{2m}A_2B_mC_1-\delta_{3m}A_3B_mC_1+\delta_{1m}A_2B_mC_2+\delta_{1m}A_3B_mC_3\\&=\delta_{21}A_2B_1C_1-\delta_{22}A_2B_2C_1-\delta_{23}A_2B_3C_1-\delta_{31}A_3B_1C_1-\delta_{32}A_3B_2C_1-\delta_{33}A_3B_3C_1\\&+\delta_{11}A_2B_1C_2+\delta_{21}A_2B_2C_2+\delta_{13}A_2B_3C_2+\delta_{11}A_3B_1C_3+\delta_{12}A_3B_2C_3+\delta_{13}A_3B_3C_3\\&=-A_2B_2C_1-A_3B_3C_1+A_2B_1C_2+A_2B_1C_3=A_2(B_1C_2-B_2C_1)-A_3(B_3C_1-B_1C_3)\end{aligned}\tag{6.7.42}$$

■ Q.E.D.

As a second example, let us prove that the divergence of a curl vanishes:

$$\vec{\nabla}\circ(\vec{\nabla}\times\vec{A})=\partial_i(\varepsilon_{ijk}\partial_jA_k)=\varepsilon_{ijk}\partial_i\partial_jA_k\tag{6.7.43}$$

By the Schwarz's theorem (see section Differential and Integral Calculus page 1469) $\partial_i\partial_j$ is symmetric (so invert the indices has no impact) in the indices and that ε_{ijk} is antisymmetric (by definition) in the same indices, the sum on i and j must necessarily cancel. For example, the contribution to the sum of the $i=1, j=2$ is the opposite of this with $i=2, j=1$.

Remarks

R1. The symbol of antisymmetry is often named "Levi-Civita tensor" in the literature. In fact, although it is a tensor in the form of its notation, it's more of a mathematical tool that a mathematical "being" hence the preference of some physicists to name it "symbol" rather than "tensor". But it's up to you ...

R2. By abuse of writing we do not write the basic vector but rigorously, and to avoid forgetting it, remember that in order to balance the members of the equality and in order to clarify that the vectors are expressed in the same base, we should write:

$$\vec{A}\times\vec{B}=\varepsilon_{ijk}A_jB_k\vec{e}_i\qquad\vec{\nabla}\times\vec{A}=\varepsilon_{ijk}\partial_jA_k\vec{e}_i\tag{6.7.44}$$

Let us now see the first simple practical applications of this index notation using the example of the base change that we have seen in the section of Vector Calculus:

Given two bases $(\vec{e}_1,\vec{e}_2,\ldots,\vec{e}_n)$ and $(\vec{e}_1',\vec{e}_2',\ldots,\vec{e}_n')$ of an Euclidean vector space $\mathcal{E}^n$. Each vector of a base can be decomposed on the other base in the form of a linear application (base change matrix - see section Linear Algebra page 1992):

$$\vec{e}_i = A'^k_i \vec{e}'_k \qquad \text{and} \qquad \vec{e}'_k = A^i_k \vec{e}_i \tag{6.7.45}$$

where we obviously use the summation convention for $i,k=1,2,\ldots,n$.

Let us recall that the base change matrix (or "transformation matrix") should have as many columns as the basic vector have lines (dimensions or components). Small example with three dimensions gives:

$$\begin{pmatrix} x'_1 \\ x'_2 \\ x'_3 \end{pmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \begin{bmatrix} x_1a_{11}+x_2a_{12}+x_3a_{13} \\ x_2a_{21}+x_2a_{22}+x_3a_{23} \\ x_1a_{31}+x_2a_{32}+x_3a_{33} \end{bmatrix} \tag{6.7.46}$$

and obviously it is much more funny to write this as:

$$x'_i = A^k_i x_k \tag{6.7.47}$$

so where on A, we have the k that represents the column of the matrix and i the row of the matrix.

Any vector $\vec{x}$ of $\mathcal{E}^n$ can be decomposed (we have already prove this in the section of Vector Calculus) on each basis $\mathcal{E}^n$ under the form:

$$\vec{x} = x^i \vec{e}_i = x'^k \vec{e}'_k \tag{6.7.48}$$

If we seek for the relations between the components x^i and x'^k it is enough to take again the relations prove in the section of Linear Algebra and we have then:

$$\vec{x} = x^i \vec{e}_i = x^i A'^k_i \vec{e}'_k = x'^k A^i_k \vec{e}_i \tag{6.7.49}$$

Immediately by the uniqueness of the decomposition of a vector on a base, we can equalize the components of the basis vectors and we get (the must be careful to rearrange again the order of the terms because the matrix multiplication is in general, not commutative as we already know!):

$$x'^k = A'^k_i x^i \quad \text{and} \quad x^i = A^i_k x'^k \tag{6.7.50}$$

By construction we also have the trivial relation:

$$A'^i_k A^j_i = \delta^j_k \tag{6.7.51}$$

As:

$$x^i = A^i_k x'^k = A^i_k (A'^k_i x^i) = (A^i_k A'^k_i) x^i \tag{6.7.52}$$

A way to prove in a quite general way the previous relation using tensor calculus notation is to remember the following result proved in the section of Linear Algebra:

$$\vec{x} = x^i \vec{e}_i = x^j A'^k_i = \vec{e}'_k = x'^k \vec{e}'_k = x'^k A^i_k \vec{e}_i \tag{6.7.53}$$

and by using:

$$\vec{e}'_k = A^i_k \vec{e}_i \tag{6.7.54}$$

We therefore have:

$$\vec{e}_i = A'^k_i \vec{e}'_k = A'^k_i A^j_k \vec{e}_j \tag{6.7.55}$$

The basis vectors being linearly independent, this last relation implies that when $i \neq j$:

$$A'^k_i A^j_k = 0 \tag{6.7.56}$$

and when $i = j$:

$$A'^k_i A^j_k = 1 \tag{6.7.57}$$

Therefore it comes:

$$A'^k_i A^j_k = \delta^j_i \tag{6.7.58}$$

And for the dot product, the results obtained with the index notation are very interesting and extremely powerful. We have already defined the scalar product in the section of Vector Calculus (see page 1821) but let us see how we handle this with the index notation!

Let us consider an Euclidean vector space $\mathcal{E}^n$ reported any basis $\vec{e}_i$. We already know that vectors are written on this basis:

$$\vec{x} = x^i \vec{e}_i \qquad \vec{y} = y^i \vec{e}_j \tag{6.7.59}$$

The scalar product with respect to its properties and the index notation is then written:

$$\boxed{\vec{x} \circ \vec{y} = (x^i \vec{e}_i) \circ (y^i \vec{e}_j) = x^i y^j (\vec{e}_i \circ \vec{e}_j) = g_{ij} x^i y^j} \tag{6.7.60}$$

This is a fundamental relation for advanced physics (General Relativity and String Theory) that makes appear the "metric covariant tensor" (see section Non-Euclidean Geometry page 2673):

$$\boxed{g_{ij} = \vec{e}_i \circ \vec{e}_j} \tag{6.7.61}$$

and to satisfy the commutative property of the dot product (see section Vector Calculus page 1821) we must obviously have the equality (at least in Euclidean space or approximated as...):

$$g_{ij} = g_{ji} \tag{6.7.62}$$

The prior-previous relation is sometimes written in the form:

$$[g] = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1n} \\ g_{21} & g_{22} & \cdots & g_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ g_{m1} & g_{m2} & \cdots & g_{mn} \end{bmatrix} = \begin{bmatrix} \vec{e}_1 \circ \vec{e}_1 & \vec{e}_1 \circ \vec{e}_2 & \cdots & \vec{e}_1 \circ \vec{e}_n \\ \vec{e}_2 \circ \vec{e}_1 & \vec{e}_2 \circ \vec{e}_2 & \cdots & \vec{e}_2 \circ \vec{e}_n \\ \vdots & \vdots & \ddots & \vdots \\ \vec{e}_m \circ \vec{e}_1 & \vec{e}_m \circ \vec{e}_2 & \cdots & \vec{e}_m \circ \vec{e}_n \end{bmatrix} \tag{6.7.63}$$

Remark

When the basis vectors $\vec{e}_i$ form an orthogonal vector space (not necessarily orthonormal) the quantities:

$$g_{ij}=\vec{e}_i\circ\vec{e}_j \quad (6.7.64)$$

are obviously zero when $i\neq j$. The dot product of two vectors $\vec{x}$ and $\vec{y}$ is then reduce to:

$$\vec{x}\circ\vec{y}=g_{11}x^1y^1+g_{22}g^2y^2+\ldots+g_{nn}x^ny^n \quad (6.7.65)$$

We then have in this particular case:

$$g_{ij}=\delta_{ij}(\vec{e}_i\circ\vec{e}_j) \quad (6.7.66)$$

and therefore when the basis vectors form an orthonormal vector space it is clear that g_{ij} is equal to the Kronecker symbol alone such that:

$$\vec{x}\circ\vec{y}=\|\vec{x}\circ\vec{y}\|^2=\delta_{ij}x^iy^j \quad (6.7.67)$$

6.7.3 Metric and Signature

As we have seen in it in the section of Vector Calculus (and Topology), the dot product of a vector $\vec{x}$ can be used to define the concept of norm of a vector (and also the concept of distance).

Let us recall that we have by definition the norm of a vector which is given by (see section Vector Calculus page 1821):

$$\vec{x}\circ\vec{x}=g_{ij}x^ix^j=\|\vec{x}\|^2 \quad (6.7.68)$$

where the numbers g_{ij} define somehow a "measure" of the vectors; we then say in the language of tensor calculus that they are the "metrics" of the selected vector space.

In the space of classical geometry, the norm is a number that is always strictly positive and which becomes zero if the measured vector is also zero. By cons the previous expression of the norm of a vector, may eventually be negative for any numbers $g_{11},g_{12},\ldots,g_{nn}$ (complex spaces for example). So we can distinguish two kinds pre-Euclidean vector spaces (Euclidean space in which we define a scalar product for recall) depending on the fact that the norm is positive or not. However when in theoretical physics we want to make the analogy with a vector space structure we need that the condition:

$$g=\det[g_{ij}]\neq 0 \quad (6.7.69)$$

is satisfied (g_{ij} can be written as a matrix, nothing prevents us to do it).

Explanations: We know that the dot product must satisfy the commutative property such that:

$$\vec{x}\circ\vec{y}=g_{ij}x^iy^j=\vec{y}\circ\vec{x}=g_{ji}x^jy^i \quad (6.7.70)$$

On the other hand, if for any non-zero y^j we have:

$$g_{ij}x^iy^j=0 \quad (6.7.71)$$

this implies $x^i=0$ (that is one of the properties of the norm we saw in the section of Vector Calculus). We can then write:

$$g_{ij}x^i=0 \tag{6.7.72}$$

We are here simply with a system of n equations with n unknowns (having to admit by hypothesis that only the solution $x^i=0$), it is necessary and sufficient for this that the determinant of the system, denoted g, to be is different from zero (see section Linear Algebra page 1980). So we must have:

$$g=\det[g_{ij}]\neq 0 \tag{6.7.73}$$

It is one of the condition for an expression comparable to a norm under a tensor index notation form in the context of theoretical physics vector space of the states of the system !!

Remarks

R1. The number of $+$ and $-$ signs found in the expression of the dot product is a is a characteristic of a given vector space E^n. It is named the "signature of the vector space".

R2. A practical application of calculation of the metric is presented in details in the section of General Relativity.

From the coefficients of the covariant metric tensor g_{ij} defining the metric of the space E^n, we can introduce the coefficients of the "contravariant metric tensor g^{ij}" defining the metric of a "dual space" E_*^n by the relation:

$$\boxed{g_{ik}g^{kj}=\delta_i^j=\begin{cases}1 & \text{if } i=j\\ 0 & \text{if } i\neq j\end{cases}} \tag{6.7.74}$$

In other words, the metric tensor twice covariant is its own inverse by its equivalent twice contravariant tensor. We will prove it explicitly later by showing during our study of the of the Gram's determinant that contravariant and covariant components of a Euclidean space are equal and both space have the same number of dimensions.

A well known special case that meets the above equality is the Minkowski's metric tensor of (see section Special Relativity page 5705), where we have:

$$\eta_{\mu\nu}=\eta^{\mu\nu}=\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & -1 & 0 & 0\\ 0 & 0 & -1 & 0\\ 0 & 0 & 0 & -1\end{bmatrix} \tag{6.7.75}$$

Remark

The space E^n is also named "primal space" and if it is of Euclidean type let us recall that it is denoted $\mathcal{E}^n$.

The dual space is underpinned by n basis vectors $\vec{e}^{\,i}$ constructed from the vectors $\vec{e}_i$ such that:

$$\boxed{\vec{e}^{\,i}=g^{ij}\vec{e}_j} \tag{6.7.76}$$

It is therefore easy to see that the scalar product of the vectors $\vec{e}^i$ defines the metric g^{ij} of the dual space:

$$\vec{e}^i \circ \vec{e}^j = (g^{ik}\vec{e}_k) \circ (g^{jl}\vec{e}_l) = g^{ik}g^{jl}(\vec{e}_k \circ \vec{e}_l) = g^{ik}(g^{jl}g_{kl}) = g^{ik}\delta_k^j = g^{ij} \qquad (6.7.77)$$

while the vectors $\vec{e}^i$ (contravariant) and $\vec{e}_j$ (covariates) are orthogonal:

$$\vec{e}^i \circ \vec{e}_j = (g^{ik}\vec{e}_k) \circ \vec{e}_j = g^{ik}(\vec{e}_k \circ \vec{e}_j) = g^{ik}g_{kj} = \delta_j^i \qquad (6.7.78)$$

We can also express a vector in the dual base by the next writing by noting that obviously the position of the dummy indices is reversed:

$$\vec{x} = x_1\vec{e}^1 + x_2\vec{e}^2 + x_3\vec{e}^3 = x_i\vec{e}^i \qquad (6.7.79)$$

Remark

The components x_i (orthogonal projections of the vector on the axes) are named, for reasons that we will see further below, the "covariant" components.

So we finally have the possibility to change the vectors of a base to another one:

$$\vec{x} = x_i\vec{e}^i = x_i(g^{ik}\vec{e}_k) = (x_i g^{ik})\vec{e}_k = x^k\vec{e}_k \qquad (6.7.80)$$

where it is important to remember that to make a contravariant a covariant component, we bring up the index:

$$\boxed{x^k = g^{ik}x_i} \qquad (6.7.81)$$

and conversely, to make it covariant:

$$\boxed{x_i = g_{ik}x^k} \qquad (6.7.82)$$

A "covariant vector" or "cotangent vector" (often abbreviated as "covector") has components that co-vary with a change of basis. That is, the components must be transformed by the same matrix as the change of basis matrix. The components of covectors (as opposed to those of vectors) are said to be "covariant".

So, still in the case of the example of the Minkowski metric, if we consider the contravariant four-vector:

$$x^\mu = (x^0, x^i) = (x^0, \vec{x}) \qquad (6.7.83)$$

Then we have:

$$x_\mu = \eta_{\mu\nu}x^\nu = \eta_{\mu\nu}(x^0, \vec{x}) = (x_0, -x_i) = (x_0, -\vec{x}) \qquad (6.7.84)$$

6.7.4 Gram's Determinant

Let us see another approach to determine the base vectors of the dual space that can allow also a better understanding of the concept and will allow us to get an interesting result that we will use during certain calculations of General Relativity and String Theory (mainly its study using to the Lagrangian formalism).

So we have for $i=j=1$:

$$\vec{e}^{\,1}\circ\vec{e}_1=1 \tag{6.7.85}$$

This scalar product can be seen as a normalization condition for the two bases and two scalar products $\vec{e}^{\,2}\circ\vec{e}_1=0, \vec{e}^{\,2}\circ\vec{e}_1=0$ as orthogonalization conditions. Thus, as $\vec{e}_1$ is perpendicular to $\vec{e}^{\,2},\vec{e}^{\,3}$ we can write:

$$\vec{e}^{\,1}=c^{te}(\vec{e}^{\,2}\times\vec{e}^{\,3}) \tag{6.7.86}$$

where c^{te} is a constant of proportionality. Now let us play around with the prior previous relation:

$$\vec{e}^{\,1}\circ\vec{e}_1=c^{te}\vec{e}^{\,1}\circ(\vec{e}^{\,2}\times\vec{e}^{\,3})=1 \tag{6.7.87}$$

Then we get:

$$c^{te}=\frac{1}{\vec{e}^{\,1}\circ(\vec{e}^{\,2}\times\vec{e}^{\,3})}=\frac{1}{[\vec{e}^{\,1},\vec{e}^{\,2},\vec{e}^{\,3}]} \tag{6.7.88}$$

where we see appear the mixed product as we had defined it in the section of Vector Calculus.

Thus we get very easily:

$$\vec{e}_i=\frac{\vec{e}^{\,j}\times\vec{e}^{\,k}}{[\vec{e}^{\,1},\vec{e}^{\,2},\vec{e}^{\,3}]} \tag{6.7.89}$$

and even for contravariant vectors (without proof as may be too obvious?):

$$\vec{e}^{\,i}=\frac{\vec{e}_j\times\vec{e}_k}{[\vec{e}_1,\vec{e}_2,\vec{e}_3]} \tag{6.7.90}$$

Remarks

R1. The reader will have perhaps noticed that the relations above are only valid for a three-dimensional space.

R2. The notation of the two previous relations is mathematically a bit unfair because in reality it is not an equality between two vectors but an application of a vector space in the other one!

R3. As in physics is very frequently considered Cartesian , cylindrical and spherical orthonormal base and that denominator of the two previous relations is always equal to 1 in these bases than the contravariant basis vectors are identified with covariant basis vectors (and vice versa). So the covariant coordinates are equal to the contravariant coordinates for these special cases!!!!!!!!!!

Now let us come back on something that will seem very old of us... In the section of Vector Calculus, we have defined and studied what were the cross product and mixed product. We will now see another way of representing them and see that this representation provides a result for the less quite relevant!

We saw in the section of Vector Calculus that the vector product was given by:

$$\vec{x}\times\vec{y}=(x_2y_3-x_3y_2)\vec{e}_1+(x_3y_1-x_1y_3)\vec{e}_2+(x_1y_2-x_2y_1)\vec{e}_3 \tag{6.7.91}$$

But what we did not see and we will now that we can trivially this expression is only the vector determinant of the following matrices:

$$\vec{x}\times\vec{y}=\det\begin{bmatrix}\vec{e}_1 & \vec{e}_2 & \vec{e}_3\\ x_1 & x_2 & x_3\\ y_1 & y_2 & y_3\end{bmatrix}=\det\begin{bmatrix}\vec{e}_2\vec{e}_3 & \vec{e}_3\times\vec{e}_1 & \vec{e}_1\times\vec{e}_2\\ x_1 & x_2 & x_3\\ y_1 & y_2 & y_3\end{bmatrix} \tag{6.7.92}$$

Yes ... so the result does not give a scalar! It is just a usage rating.

But as we do the tensor calculus, we must now properly distinguish covariant and contravariant components. We'll rewrite it properly with contravariant components:

$$\vec{x}\times\vec{y}=\det\begin{bmatrix}\vec{e}_1 & \vec{e}_2 & \vec{e}_3\\ x^1 & x^2 & x^3\\ y^1 & y^2 & y^3\end{bmatrix} \tag{6.7.93}$$

Similarly, the mixed product can be written using this relation and notation:

$$[\vec{x},\vec{y},\vec{z}]=\vec{x}\circ(\vec{y}\circ\vec{z})=\vec{x}\circ\det\begin{bmatrix}\vec{e}_1 & \vec{e}_2 & \vec{e}_3\\ y^1 & y^2 & y^3\\ z^1 & z^2 & z^3\end{bmatrix}=(x^1\vec{e}_1+x^2\vec{e}_2+x^3\vec{e}_3)\circ\det\begin{bmatrix}\vec{e}_1 & \vec{e}_2 & \vec{e}_3\\ y^1 & y^2 & y^3\\ z^1 & z^2 & z^3\end{bmatrix} \tag{6.7.94}$$

Or, looking at the expression of the determinant we see quite easily, without having to do developments, that:

$$[\vec{x},\vec{y},\vec{z}]=\vec{e}_1\circ(\vec{e}_2\times\vec{e}_3)\det\begin{bmatrix}x^1 & x^2 & x^3\\ y^1 & y^2 & y^3\\ z^1 & z^2 & z^3\end{bmatrix} \tag{6.7.95}$$

Indeed (we calculate the determinant making use of the demonstration of the determinant with three components proved in the section of Linear Algebra):

$$\begin{aligned}
[\vec{x},\vec{y},\vec{z}]=\vec{x}\circ(\vec{y}\times\vec{z})&=\vec{e}_1\circ(\vec{e}_2\times\vec{e}_3)\det\begin{bmatrix}x^1 & x^2 & x^3\\ y^1 & y^2 & y^3\\ z^1 & z^2 & z^3\end{bmatrix}\\
&=\vec{e}_1\circ(\vec{e}_2\times\vec{e}_3)(x^1y^2z^3-x^1y^3z^2-x^2y^1z^3+x^2y^3z^1+x^2y^3z^1+x^3y^1z^2-x^3y^2z^1)\\
&=\vec{e}_1\circ\left(\begin{pmatrix}0\\1\\0\end{pmatrix}\times\begin{pmatrix}0\\0\\1\end{pmatrix}\right)(x^1y^2z^3-x^1y^3z^2-x^2y^1z^3+x^2y^3z^1+x^2y^3z^1+x^3y^1z^2-x^3y^2z^1)\\
&=\vec{e}_1\circ\begin{pmatrix}1\\0\\0\end{pmatrix}(x^1y^2z^3-x^1y^3z^2-x^2y^1z^3+x^2y^3z^1+x^2y^3z^1+x^3y^1z^2-x^3y^2z^1)\\
&=\begin{pmatrix}1\\0\\0\end{pmatrix}\circ\begin{pmatrix}1\\0\\0\end{pmatrix}(x^1y^2z^3-x^1y^3z^2-x^2y^1z^3+x^2y^3z^1+x^2y^3z^1+x^3y^1z^2-x^3y^2z^1)\\
&=(x^1y^2z^3-x^1y^3z^2-x^2y^1z^3+x^2y^3z^1+x^2y^3z^1+x^3y^1z^2-x^3y^2z^1)\\
&=((x^1y^2z^3-x^1y^3z^2)-(x^2y^1z^3-x^2y^3z^1)+(x^2y^3z^1+x^3y^1z^2-x^3y^2z^1))\\
&=(x^1(y^2z^3-y^3z^2)+x^2(y^3z^1-y^1z^3)+x^3(y^1z^2-y^2z^1))\\
&=\begin{pmatrix}x^1\\x^2\\x^3\end{pmatrix}\circ\begin{pmatrix}y^2z^3-y^3z^2\\y^3z^1-y^1z^3\\y^1z^2-y^2z^1\end{pmatrix}=\begin{pmatrix}x^1\\x^2\\x^3\end{pmatrix}\circ\left(\begin{pmatrix}y^1\\y^2\\y^3\end{pmatrix}\times\begin{pmatrix}z^1\\z^2\\z^3\end{pmatrix}\right)=\vec{x}\circ(\vec{y}\times\vec{z})
\end{aligned}\tag{6.7.96}$$

The prior-previous relation is also is frequently written:

$$[\vec{x},\vec{y},\vec{z}]=\sqrt{g}\det\begin{bmatrix}x^1 & x^2 & x^3\\ y^1 & y^2 & y^3\\ z^1 & z^2 & z^3\end{bmatrix}\tag{6.7.97}$$

with obviously:

$$\sqrt{g}=\vec{e}_1\circ(\vec{e}_2\times\vec{e}_3)=[\vec{e}_1,\vec{e}_2,\vec{e}_3]\tag{6.7.98}$$

named "Euclidean volume" (indeed let us recall that the mixed product is a volume as we show it in the section of Vector Calculus!)

Remark

Let us recall again that if the basis vectors are orthonormal, whether they are expressed in Cartesian , cylindrical or spherical coordinates then:

$$\sqrt{g}=\vec{e}_1\circ(\vec{e}_2\times\vec{e}_3)=1\tag{6.7.99}$$

Moreover, we also have the important relation:

$$\sqrt{g}=\vec{e}_1\circ(\vec{e}_2\times\vec{e}_3)=\left(\frac{\vec{e}^2\times\vec{e}^3}{[\vec{e}^1,\vec{e}^2,\vec{e}^3]}\right)\circ(\vec{e}_2\times\vec{e}_3)=\frac{(\vec{e}^2\times\vec{e}^3)(\vec{e}_2\times\vec{e}_3)}{[\vec{e}^1,\vec{e}^2,\vec{e}^3]}\tag{6.7.100}$$

Indeed, using the relation seen in the section of Vector Calculus:

$$(\vec{e}^2\times\vec{e}^3)\circ(\vec{e}_2\times\vec{e}_3)=(\vec{e}^2\circ\vec{e}_2)(\vec{e}^3\circ\vec{e}_3)-(\vec{e}^2\circ\vec{e}_3)(\vec{e}^3\circ\vec{e}_2)\tag{6.7.101}$$

But we have seen previously that $\vec{e}^i \circ \vec{e}_j = \delta^i_j$, therefore:

$$(\vec{e}^2 \circ \vec{e}_2)(\vec{e}^3 \circ \vec{e}_3) - (\vec{e}^2 \circ \vec{e}_3)(\vec{e}^3 \circ \vec{e}_2) = 1 \cdot 1 - 0 \cdot 0 \tag{6.7.102}$$

and thus finally:

$$\sqrt{g} = [\vec{e}_1, \vec{e}_2, \vec{e}_3] = \frac{1}{[\vec{e}^1, \vec{e}^2, \vec{e}^3]} \tag{6.7.103}$$

This having been done let us come back our relation of the vector product:

$$\vec{x} \times \vec{y} = x^i \vec{e}_i \times y^j \vec{e}_j = \det \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \\ x^1 & x^2 & x^3 \\ y^1 & y^2 & y^3 \end{bmatrix} = \det \begin{bmatrix} \vec{e}_2 \times \vec{e}_3 & \vec{e}_3 \times \vec{e}_1 & \vec{e}_1 \times \vec{e}_2 \\ x^1 & x^2 & x^3 \\ y^1 & y^2 & y^3 \end{bmatrix} \tag{6.7.104}$$

and let us express the components of the 1st line of the determinant in their dual basis (in contravariant coordinates):

$$\begin{aligned} \vec{x} \times \vec{y} &= \det \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \\ x^1 & x^2 & x^3 \\ y^1 & y^2 & y^3 \end{bmatrix} = \det \begin{bmatrix} \dfrac{\vec{e}^2 \times \vec{e}^3}{[\vec{e}^1, \vec{e}^2, \vec{e}^3]} & \dfrac{\vec{e}^3 \times \vec{e}^1}{[\vec{e}^1, \vec{e}^2, \vec{e}^3]} & \dfrac{\vec{e}^1 \times \vec{e}^2}{[\vec{e}^1, \vec{e}^2, \vec{e}^3]} \\ x^1 & x^2 & x^3 \\ y^1 & y^2 & y^3 \end{bmatrix} \\ &= \frac{1}{[\vec{e}^1, \vec{e}^2, \vec{e}^3]} \begin{bmatrix} \vec{e}^1 & \vec{e}^2 & \vec{e}^3 \\ x^1 & x^2 & x^3 \\ y^1 & y^2 & y^3 \end{bmatrix} = \sqrt{g} \begin{bmatrix} \vec{e}^1 & \vec{e}^2 & \vec{e}^3 \\ x^1 & x^2 & x^3 \\ y^1 & y^2 & y^3 \end{bmatrix} \end{aligned} \tag{6.7.105}$$

Obviously, if the cross product is expressed in covariant components then we have:

$$\begin{aligned} \vec{x} \times \vec{y} &= x_i \vec{e}^i \times y_j \vec{e}^j = \det \begin{bmatrix} \vec{e}^1 & \vec{e}^2 & \vec{e}^3 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{bmatrix} = \det \begin{bmatrix} \dfrac{\vec{e}_2 \times \vec{e}_3}{[\vec{e}_1, \vec{e}_2, \vec{e}_3]} & \dfrac{\vec{e}_3 \times \vec{e}_1}{[\vec{e}_1, \vec{e}_2, \vec{e}_3]} & \dfrac{\vec{e}_1 \times \vec{e}_2}{[\vec{e}_1, \vec{e}_2, \vec{e}_3]} \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{bmatrix} \\ &= \frac{1}{[\vec{e}_1, \vec{e}_2, \vec{e}_3]} \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{bmatrix} = \frac{1}{\sqrt{g}} \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{bmatrix} \end{aligned} \tag{6.7.106}$$

Now let us apply the mixed product:

$$[\vec{x}, \vec{y}, \vec{z}] = x^i \vec{e}_i \circ (y^j \vec{e}_j \times z^k \vec{e}_k) = x^i \vec{e}_i \sqrt{g} \circ \det \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \\ x^1 & x^2 & x^3 \\ y^1 & y^2 & y^3 \end{bmatrix} \tag{6.7.107}$$

knowing the expression of the determinant of a square 3×3 matrix (see section Linear Algebra page 1967) it comes immediately (we can detail on request as always in this textbook):

$$[\vec{x}, \vec{y}, \vec{z}] = x^i \vec{e}_i \sqrt{g} \circ \det \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \\ x^1 & x^2 & x^3 \\ y^1 & y^2 & y^3 \end{bmatrix} = \sqrt{g} \det \begin{bmatrix} x^1 & x^2 & x^3 \\ y^1 & y^2 & y^3 \\ z^1 & z^2 & z^3 \end{bmatrix} \tag{6.7.108}$$

Conversely, it comes almost immediately:

$$[\vec{x}, \vec{y}, \vec{z}] = x_i \vec{e}^i \circ (y_j \vec{e}^j \times z_k \vec{e}^k) = \frac{1}{\sqrt{g}} \det \begin{bmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ z_1 & z_2 & z_3 \end{bmatrix} \tag{6.7.109}$$

But, we have proved in the section of Vector Calculus that $x_i = \vec{x} \circ \vec{e}_i$. It comes then:

$$\begin{aligned}[\vec{x},\vec{y},\vec{z}] = x_i \vec{e}^{\,i} \circ (y_j \vec{e}^{\,j} \times z_k \vec{e}^{\,k}) &= \frac{1}{\sqrt{g}} \det \begin{bmatrix} \vec{x}\circ\vec{e}_1 & \vec{x}\circ\vec{e}_2 & \vec{x}\circ\vec{e}_3 \\ \vec{y}\circ\vec{e}_1 & \vec{y}\circ\vec{e}_2 & \vec{y}\circ\vec{e}_3 \\ \vec{z}\circ\vec{e}_1 & \vec{z}\circ\vec{e}_2 & \vec{z}\circ\vec{e}_3 \end{bmatrix} \\ &= \frac{1}{[\vec{e}_1,\vec{e}_2,\vec{e}_3]} \det \begin{bmatrix} \vec{x}\circ\vec{e}_1 & \vec{x}\circ\vec{e}_2 & \vec{x}\circ\vec{e}_3 \\ \vec{y}\circ\vec{e}_1 & \vec{y}\circ\vec{e}_2 & \vec{y}\circ\vec{e}_3 \\ \vec{z}\circ\vec{e}_1 & \vec{z}\circ\vec{e}_2 & \vec{z}\circ\vec{e}_3 \end{bmatrix}\end{aligned} \tag{6.7.110}$$

and therefore:

$$\boxed{[\vec{x},\vec{y},\vec{z}][\vec{e}_1,\vec{e}_2,\vec{e}_3] = \det \begin{bmatrix} \vec{x}\circ\vec{e}_1 & \vec{x}\circ\vec{e}_2 & \vec{x}\circ\vec{e}_3 \\ \vec{y}\circ\vec{e}_1 & \vec{y}\circ\vec{e}_2 & \vec{y}\circ\vec{e}_3 \\ \vec{z}\circ\vec{e}_1 & \vec{z}\circ\vec{e}_2 & \vec{z}\circ\vec{e}_3 \end{bmatrix}} \tag{6.7.111}$$

The latter relation is often named "Gram determinant". A special very interesting case gives us (we use the relation between the metric components and the dot products of the vector basis we proved during our study of the metric just earlier above):

$$[\vec{e}_1,\vec{e}_2,\vec{e}_3][\vec{e}_1,\vec{e}_2,\vec{e}_3] = [\vec{e}_1,\vec{e}_2,\vec{e}_3]^2 = \det \begin{bmatrix} \vec{e}_1\circ\vec{e}_1 & \vec{e}_1\circ\vec{e}_2 & \vec{e}_1\circ\vec{e}_3 \\ \vec{e}_2\circ\vec{e}_1 & \vec{e}_1\circ\vec{e}_2 & \vec{e}_2\circ\vec{e}_3 \\ \vec{e}_3\circ\vec{e}_1 & \vec{e}_3\circ\vec{e}_2 & \vec{e}_3\circ\vec{e}_3 \end{bmatrix} = \det \begin{bmatrix} g_{11} & g_{12} & g_{13} \\ g_{21} & g_{22} & g_{23} \\ g_{31} & g_{32} & g_{33} \end{bmatrix} \tag{6.7.112}$$

written in another way:

$$\boxed{(\sqrt{g})^2 = g = \det(g_{ij})} \tag{6.7.113}$$

Thus, the Euclidean volume is given by what name the "functional determinant" of the system (expression that we will see again and use in the section of General relativity to calculate the real volume and also in the section of String Theory):

$$\sqrt{g} = \sqrt{\det(g_{ij})} \tag{6.7.114}$$

that is without units (so you have to multiply it by a factor of elementary volume to get volume units). If we note in another way the determinant:

$$\sqrt{g} = \sqrt{|g|} \tag{6.7.115}$$

We get the common relation we can find in many books on General Relativity and String Theory but given without proof and named the "Riemannian volume" form or simply "volume form":

$$\boxed{V = \int\limits_V \sqrt{g} = \int\limits_V \sqrt{|g|}\mathrm{d}x^1\mathrm{d}x^2\mathrm{d}x^3} \tag{6.7.116}$$

or written in the following "invariant volume element" form:

$$\boxed{\mathrm{d}V = \sqrt{g}\mathrm{d}^4x} \tag{6.7.117}$$

The reader can verify normally easily enough that for the orthonormal Cartesian reference frame we fall back on the volume of a cube and that for the spherical case we fall back well on the expression of the infinitesimal volume of the sphere as used in the section of Geometric Shapes (but on request we can add the details here).

If we use the result we get in the section of Differential Geometry we have therefore for any surface patch:

$$V=\int\limits_V \sqrt{\det\begin{bmatrix}E & F\\ F & G\end{bmatrix}}\mathrm{d}x^1\mathrm{d}x^2\mathrm{d}x^3=\int\limits_V \sqrt{EF-F^2}\mathrm{d}x^1\mathrm{d}x^2\mathrm{d}x^3 \tag{6.7.118}$$

6.7.5 Contravariant and Covariant Components

So far we wrote the dummy indices arbitrarily on superscript or subscript at our discretion. However, this is not always allowed and sometimes the fact that a dummy index is in superscript or subscript has a special significance! This is often the major difficulty in the study of some theorems, because if we do not study those indices from the beginning, we do not really know how to interpret the position of the dummy indices. The reader should then be extremely careful at this level.

For an Euclidean vector space $\mathcal{E}^n$ reported to any base $(\vec{e}_i)$, the scalar product of a vector $\vec{x}=x^i\vec{e}_i$ by a vector $(\vec{e}_j)$ of its base is written as we know by (remember that this is equivalent as projecting the components on the axis corresponding to $\vec{e}_j$):

$$\vec{x}\circ\vec{e}_j=(x^j\vec{e}_i)\vec{e}_j=x^i g_{ij}=x_j \tag{6.7.119}$$

Therefore:

$$\boxed{x_j=x^i g_{ij}} \tag{6.7.120}$$

This relation is of major importance in theoretical physics and tensor calculus. It is important to remember it when we will study the contraction of indexes later (you can observe in the previous relation that we have "lowered" in the left side the index of the component of the right member of the equality).

These scalar products denoted x_j, are named "covariant components" in the base $(\vec{e}_i)$ of the vector $\vec{x}$. These components are therefore defined by:

$$x_j=\vec{x}\circ\vec{e}_j \tag{6.7.121}$$

They will denoted by lower indices !!! We will see later that these components are naturally introduced for some vectors of physics, for example the gradient vector. Moreover, the notion of covariant component is essential for tensors.

Remarks

R1. Never forget that this is therefore the projection of a vector on a vector of its own base!!!!

R2. The basic vectors always have their indices noted down because they are their own covariant components (they project on themselves by scalar product). This is the main trick used by beginners to remember when to put lower indices (and therefore they know when to put the upper one...)!

Conversely, the "contravariant components" (in other words: the non-projected components) can be calculated by solving with respect to the n unknowns, the system of n equations:

$$x_j = x^i g_{ij} \tag{6.7.122}$$

The previous relations show that the covariant components x_j are related to the conventional components x^i and that the contravariant components x^i are therefore numbers such that:

$$\vec{x} = x^i \vec{e}_i \tag{6.7.123}$$

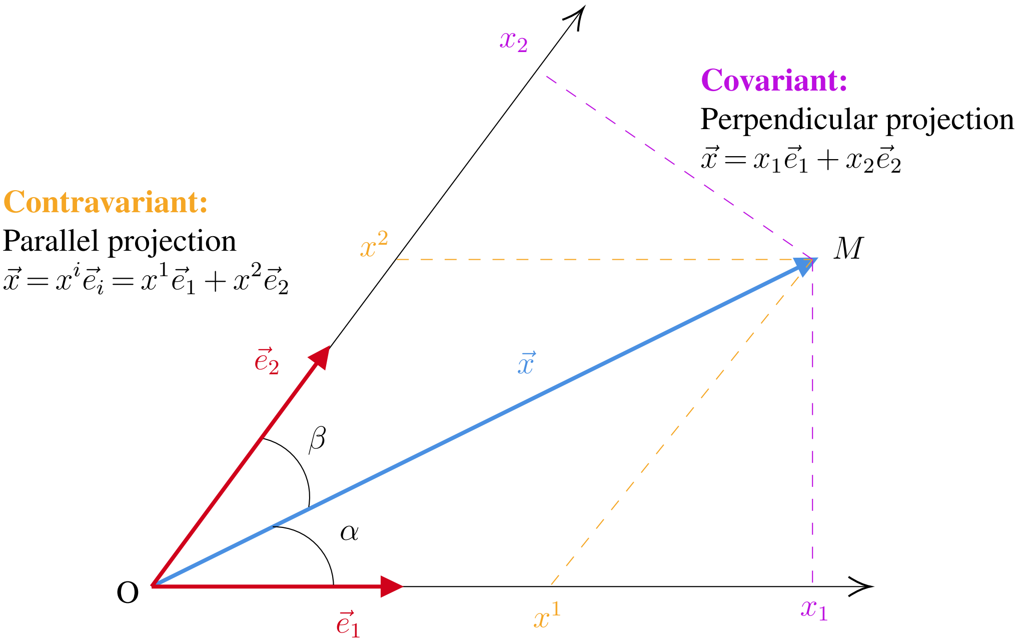

x_2
Covariant:
Perpendicular projection
$\vec{x}=x_1\vec{e}_1+x_2\vec{e}_2$
Contravariant:
Parallel projection
$\vec{x}=x^i\vec{e}_i=x^1\vec{e}_1+x^2\vec{e}_2$
x^2
M
$\vec{e}_2$
$\vec{x}$
β
α
O
$\vec{e}_1$
x^1
x_1

Figure 6.155: **Covariants and Contravariants coordinates**

They will be indicated with superscripts !! The study of the basis changes later below will further justify the appellation of these different components.

In a canonical orthonormal basis (very special case and corresponding to the classical cartesian, polar, cylindric and spherical coordinates), the covariant and contravariant components are the same as we already know after our study of Gram's determinant. Indeed:

$$x_j = \vec{x} \circ \vec{e}_j = (x^i \vec{e}_i) \circ \vec{e}_j = x^i (\vec{e}_j \circ \vec{e}_j) = x^i \delta_{ij} = x^j \tag{6.7.124}$$

Remark

We see above, that the incessant writing of dummy superscript or subscript indices can sometimes lead to some confusion and serious headaches ...

Incidentally, when we refer to a vector (or, more generally, a tensor) as being either contravariant or covariant we're abusing the language slightly, because those terms really just signify two different conventions for interpreting the components of the object with respect to a given coordinate system, whereas the essential attributes of a vector or tensor are independent of the particular coordinate system in which we choose to express it. In general, any given vector or tensor (in a metrical manifold) can be expressed in both contravariant and covariant form with respect to any given coordinate system!

Furthermore It may seem that the naming convention is backwards, because the "contra" components go with the axes, whereas the "cova" components go against the axes, but historically these names were given on the basis on the transformation laws that apply to these two different interpretations.

Most of the physical definitions refer to the contravariant components. But we must be aware in physics that we if we have for example a singularity in a tensor, if may be a real physical singularity or one only due to the choice of coordinate system! There is nothing inherent in the contravariant, covariant, or mixed components for example on stress-energy tensor (in General Relativity) that cannot be changed with a coordinate transformation, EXCEPT for the scalars.

6.7.6 Operation in Basis

The interest of physicist for the tensor calculus, is passing parameters from one base to another for some given reasons (often the aim is to simplify the study of problems or simply because the studied states depend - or may depend - on the geometry of the space in question). It is therefore necessary to introduce the main tools relating thereto. We will also take this opportunity to present the developments that could have been already addressed in the section of Vector Calculus.

Remark

As I like to say... Tensor Calculus is to physics with is XML to computing science. A background independent language!

6.7.6.1 Gram-Schmidt Orthogonalization Method

The "Schmidt orthogonalization method" (also named "Gram-Schmidt orthogonalization method") allows the actual determination of an orthogonal basis for any pre-Euclidean vector space $\mathcal{E}^n$ (we could introduce this method in the section of Vector Calculus but it seemed more interesting to us to present this method in a general and aesthetic framework using tensor calculus).

For this, let us consider a set of n linearly independent vectors $(\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_n)$ of $\mathcal{E}^n$ and suppose that for each vector we have the dot product (square of the norm):

$$\vec{x}_i\circ\vec{x}_i\neq 0 \tag{6.7.125}$$

Let us seek n vectors $\vec{e}_i$ orthogonal between them. Let us start for this with $\vec{e}_1=\vec{x}_1$ and let us seek for $\vec{e}_2$ orthogonal to $\vec{e}_1$ under the form (this is a choice!!!):

$$\vec{e}_2=\lambda_1\vec{e}_1+\vec{x}_2 \tag{6.7.126}$$

The mental visualization of the process is not quit easy so the reader has to trust (anyway...) the mathematical results (if once we have the time we will draw the process of the classical three dimension case).

The coefficient λ_1 is calculated by writing the orthogonality relation:

$$\vec{e}_1\circ\vec{e}_2=\vec{x}_1\circ(\lambda_1\vec{x}_1+\vec{x}_2)=0 \tag{6.7.127}$$

We deduce without too much troubles:

$$\lambda_1=-\frac{\vec{x}_1\circ\vec{x}_2}{\vec{x}_1\circ\vec{x}_1}=-\frac{\vec{x}_1\circ\vec{x}_2}{\|\vec{x}_1\|^2}=-\frac{\vec{e}_1\circ\vec{x}_2}{\|\vec{e}_1\|^2} \tag{6.7.128}$$

The parameter λ_1 being determined, we get the vector $\vec{e}_2$ that is orthogonal to $\vec{e}_1$ and not zero, since the system $(\vec{e}_1,\vec{x}_2,\ldots,\vec{x}_n)$ is linearly independent.

Thus so far we have:

$$\vec{e}_1=\vec{x}_1 \qquad \vec{e}_2=\vec{x}_2-\frac{\vec{e}_1\circ\vec{x}_2}{\|\vec{e}_1\|^2}\vec{e}_1=\vec{x}_2-\text{proj}_{e_1}\vec{x}_2 \tag{6.7.129}$$

The parameter λ_1 being determined, we get the vector $\vec{e}_2$ that is orthogonal to $\vec{e}_1$ and not zero, since the system $(\vec{e}_1,\vec{x}_2,\ldots,\vec{x}_n$ is linearly independent.

The vector $\vec{e}_2$ is sought in the form:

$$\vec{e}_3=\mu_1\vec{e}_1+\mu_2\vec{e}_2+\vec{x}_3 \tag{6.7.130}$$

The two relations of orthogonality: $\vec{e}_1\circ\vec{e}_e=0$ and $\vec{e}_2\circ\vec{e}_3=0$, enables the calculation of the coefficients μ_1 and μ_2. We get therefore:

$$\mu_1=-\frac{\vec{e}_1\circ\vec{x}_3}{\|\vec{e}_1\|^2}\qquad\mu_2=-\frac{\vec{e}_2\circ\vec{x}_3}{\|\vec{e}_2\|^2} \tag{6.7.131}$$

what determines the vector $\vec{e}_3$, orthogonal to $\vec{e}_1$ and $\vec{e}_2$, and not zero, since the $(\vec{e}_1,\vec{e}_2,\vec{x}_3,\ldots,\vec{x}_n)$ are independent. By continuing the same type of calculation, we get step by step a system of orthogonal vectors $(\vec{e}_1,\vec{e}_2,\vec{x}_3,\ldots,\vec{e}_n)$ between them and none of them are zero.

So finally:

$$\boxed{\begin{aligned}&\vec{e}_1=\vec{x}_1\\&\vec{e}_2=\vec{x}_2-\frac{\vec{e}_1\circ\vec{x}_2}{\|\vec{e}_1\|^2}\vec{e}_1=\vec{x}_2-\text{proj}_{e_1}\vec{x}_2\\&\vec{e}_3=-\frac{\vec{e}_1\circ\vec{x}_3}{\|\vec{e}_1\|^2}\vec{e}_1-\frac{\vec{e}_2\circ\vec{x}_3}{\|\vec{e}_2\|^2}\vec{e}_2+\vec{x}_3\end{aligned}} \tag{6.7.132}$$

In case where some vectors are such like $\vec{x}_i\circ\vec{x}_i=0$ (their norm is zero), we replace then $\vec{x}_i$ by $\vec{x'}_i+\lambda\vec{x}_j$, choosing a vector $\vec{x}_j$ so that we get $\vec{x'}_i\circ\vec{x'}_i\neq 0$.

We therefore conclude that any pre-Euclidean space admits orthogonal bases!

✍**Example:**

Suppose:

$$\vec{x}_1=\begin{pmatrix}1\\1\\1\end{pmatrix}\qquad\vec{x}_2=\begin{pmatrix}1\\1\\3\end{pmatrix}\qquad\vec{x}_3=\begin{pmatrix}0\\0\\3\end{pmatrix} \tag{6.7.133}$$

We want to find an orthonormal set of vector that spans $\{\vec{x}_1,\vec{x}_2,\vec{x}_3\}$. For this we use the Gram-Schmidt method followed by a normalisation, let $\vec{e}_1=\vec{x}_1=(1,1,1)$ then we calculate:

$$\vec{e}_2=\vec{x}_2-\frac{\vec{e}_1\circ\vec{x}_2}{\|\vec{e}_1\|^2}\vec{e}_1=\begin{pmatrix}1\\2\\3\end{pmatrix}-\frac{\begin{pmatrix}1\\1\\1\end{pmatrix}\circ\begin{pmatrix}1\\2\\3\end{pmatrix}}{3}\begin{pmatrix}1\\1\\1\end{pmatrix}=\begin{pmatrix}1\\2\\3\end{pmatrix}-\begin{pmatrix}2\\2\\2\end{pmatrix}=\begin{pmatrix}-1\\0\\1\end{pmatrix} \tag{6.7.134}$$

As a quick check we have indeed:

$$\vec{e}_1\circ\vec{e}_2=0 \tag{6.7.135}$$

Next:

$$\vec{e}_3 = -\frac{\vec{e}_1 \circ \vec{x}_3}{\|\vec{e}_1\|^2}\vec{e}_1 - \frac{\vec{e}_2 \circ \vec{x}_3}{\|\vec{e}_2\|^2}\vec{e}_2 + \vec{x}_3 = -\frac{\begin{pmatrix}1\\1\\1\end{pmatrix} \circ \begin{pmatrix}0\\0\\3\end{pmatrix}}{3}\begin{pmatrix}1\\1\\1\end{pmatrix} - \frac{\begin{pmatrix}-1\\0\\1\end{pmatrix} \circ \begin{pmatrix}0\\0\\3\end{pmatrix}}{2}\begin{pmatrix}-1\\0\\1\end{pmatrix} + \begin{pmatrix}0\\0\\3\end{pmatrix}$$
$$= -\frac{3}{3}\begin{pmatrix}1\\1\\1\end{pmatrix} - \frac{3}{2}\begin{pmatrix}-1\\0\\1\end{pmatrix} + \begin{pmatrix}0\\0\\3\end{pmatrix} = \begin{pmatrix}1/2\\-1\\1/2\end{pmatrix} \tag{6.7.136}$$

again it's good to check that $\vec{e}_1 \circ \vec{e}_2 = 0$ and $\vec{e}_1 \circ \vec{e}_3 = 0$ as we desire. Finally we notice that:

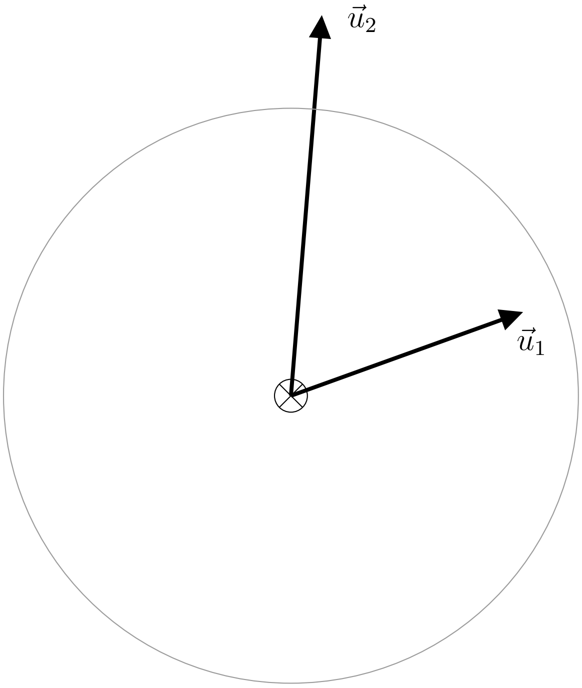
$$\|\vec{e}_1\| = \sqrt{3} \qquad \|\vec{e}_2\| = \sqrt{2} \qquad \|\vec{e}_3\| = \sqrt{3/2} \tag{6.7.137}$$

Hence:

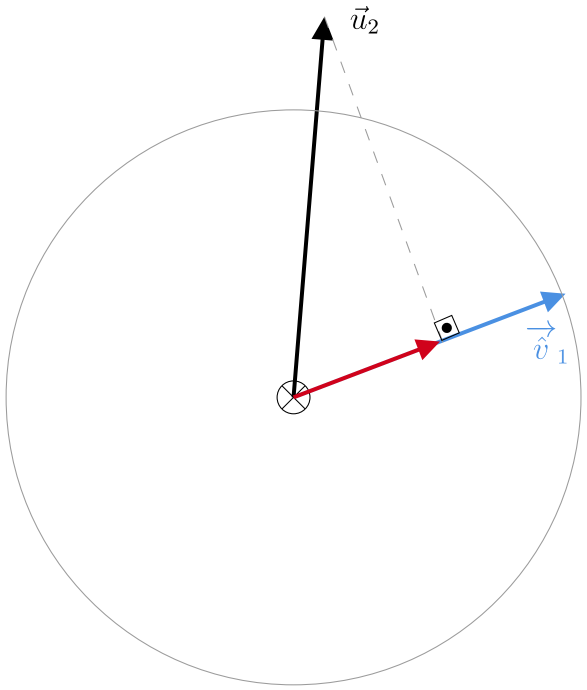
$$\vec{e}_1 = \frac{1}{\sqrt{3}}\begin{pmatrix}1\\1\\1\end{pmatrix} \qquad \vec{e}_2 = \frac{1}{\sqrt{2}}\begin{pmatrix}-1\\0\\1\end{pmatrix} \qquad \vec{e}_3 = \frac{1}{\sqrt{3/2}}\begin{pmatrix}1/2\\-1\\1/2\end{pmatrix} \tag{6.7.138}$$

are orthonormal vectors.

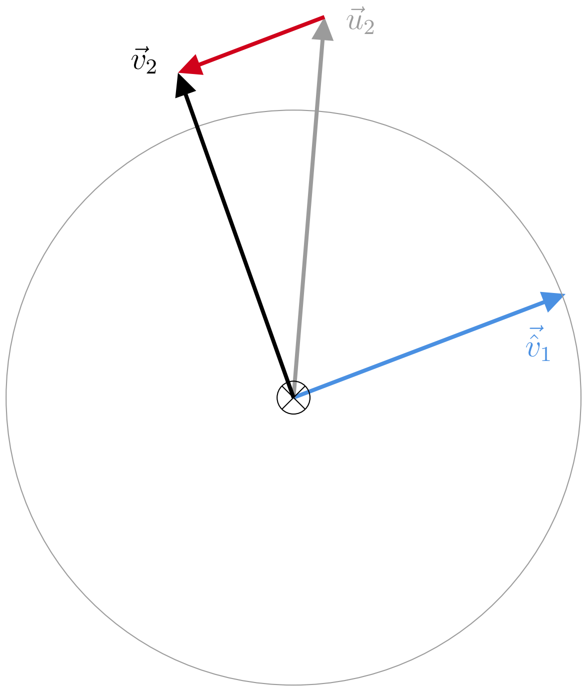
Perhaps pictorials will help! For simplicity, we will use only two vectors $\vec{u}_1$ and $\vec{u}_2$ against the unit circle:



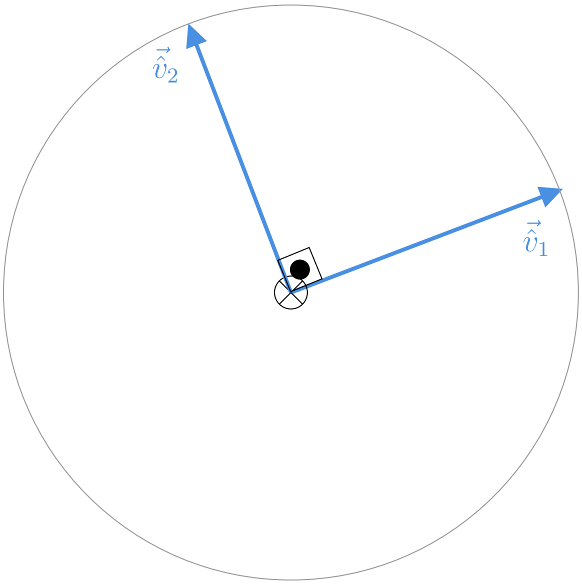
We normalize the first vector $\vec{u}_1$ and denote it $\hat{\vec{v}}_1$. Afterwards we compute the projection of the vector $\vec{u}_2$ onto the vector $\hat{\vec{v}}_1$. The projection is the red segment:


$\vec{u}_2$
$\vec{\hat{v}}_1$

We subtract afterwards the projection (red arrow) from $\vec{u}_2$. This removes the $\vec{u}_2$ components in the span of $\vec{\hat{v}}_1$. The new vector, $\vec{v}_2$, is orthogonal to $\vec{\hat{v}}_1$:


$\vec{u}_2$
$\vec{v}_2$
$\vec{\hat{v}}_1$

The last step is the normalization $\vec{\hat{v}}_2 = \vec{v}_2 / \|v_2\|$:



And for those who likes pseudocodes:

Algorithm 3: Gram-Schmidt orthogonalization

```
Input: A set of linearly independent vectors $\{\vec{x}_1, \vec{x}_2, \ldots \vec{x}_n\}$
Output: An orthonormal basis for the subspace spanned by $\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\}$
for $i = 1$ to $n$ do
    $\vec{u}_i \leftarrow \vec{x}_i$
    for $j = 1$ to $i - 1$ do
        $\vec{u}_i \leftarrow \vec{u}_i - \frac{\vec{u}_i \circ \vec{u}_j}{\|\vec{u}_j\|^2}\vec{u}_j$
    $\vec{e}_i \leftarrow \frac{\vec{u}_i}{\|\vec{u}_i\|}$
```

This system of calculation of bases is of primary importance! It can be used to study physical systems from a pre-Euclidean repository whose properties change over time. Which is typical in General Relativity.

The reader interested in computer science and numerical methods can refer to the QR matrix decomposition earlier above (see page 2030) to see another application of the Gram-Schmidt orthogonalization method.

6.7.6.2 Change of Basis

Given two bases $(\vec{e}_1, \ldots, \vec{e}_n)$ and $(\vec{e}_1{}', \ldots, \vec{e}_n{}')$ of a vector space $\mathcal{E}^n$. Each vector of a base can be decomposed on the other basis as follows (we have already prove it in the section of Linear Algebra at page 1992):

$$\vec{e}_i = A'^k_i \vec{e}_k{}' \qquad \vec{e}_l{}' = A^i_k \vec{e}_i \tag{6.7.139}$$

A vector $\vec{x}$ of $\mathcal{E}^n$, when its contravariant components known, can be decomposed in each base in the form:

$$\vec{x} = x^i \vec{e}_i = x'^k \vec{e}_k{}' \tag{6.7.140}$$

and we have already proved that:

$$x^i = A^i_k x'^k \qquad x'^k = A'^k_i x^i \tag{6.7.141}$$

We notice that the transformation relations of the components of a contravariant vector are the opposite of those of the basis vectors, the quantities A and A' being permuted, hence also the origin of the name "contra"-"variants" of these components!

Let x_i and x'_k be the covariant components of a vector $\vec{x}$ respectively in the bases $(\vec{e}_i)$ and $(\vec{e}'_k)$. Let us replace the basis vectors expressed by the relations:

$$\vec{e}_i = A'^k_i \vec{e}'_k \quad \text{and} \quad \vec{e'}_k = A^k_i \vec{e}_i \tag{6.7.142}$$

in the expression of the definition of the covariant components, therefore we get:

$$x_i = \vec{x} \circ \vec{e}_i = \vec{x} \circ (A'^k_i \vec{e}'_k) = A'^k_i (\vec{x} \circ \vec{e}'_k) = A'^k_i x'_k \tag{6.7.143}$$

Hence the relation between the covariant components in each base:

$$x_i = A'^k_i x'_k \tag{6.7.144}$$

We get also:

$$x'_k = A^k_i x_i \tag{6.7.145}$$

We notice that the covariant components transform as the basis vectors, hence also the name of these components!

Once again, unless the basis is orthonormal, never forget that the covariant and contravariant components are different!!!

6.7.6.3 Reciprocal Basis (Dual Basis)

Now let us come back on the concept of dual space but as seen in the vector calculation point of view. This second approach can perhaps help some readers to better understand the concepts seen previously but against hides the underlying reasoning for the origin of the names "covariant" and "contravariant". But it is still the most common presentation used in the literature...

Given a basis $(\vec{e}_i)$ of an Euclidean vector space $\mathcal{E}^n$. By definition, n vectors $\vec{e}^k$ which satisfy the following relations:

$$\vec{e}_i \circ \vec{e}^k = \delta_{ik} \tag{6.7.146}$$

are named "reciprocal vectors" of the vectors $\vec{e}_i$. They will be denoted with higher indices. By definition, each reciprocal vector $\vec{e}^k$ must therefore be orthogonal to all the vectors $\vec{e}_i$, except for $k = i$.

Let us first show that the reciprocal vectors $\vec{e}^k$ of a given base $(\vec{e}_i)$ are linearly independent. For this, we must show that a linear combination $\lambda \vec{e}^k$ gives a zero vector if and only if each coefficient λ_k is zero.

Proof 6.45.8. Given $\vec{e}=x^i\vec{e}_i$ any vector of $\mathcal{E}_n$. Let us make a dot product by $\vec{x}$ the previous linear combination $\lambda_k\vec{e}^k$, we get:

$$(\lambda_k\vec{e}^k)\circ\vec{x}=(\lambda_k\vec{e}^k)\circ(x^i\vec{e}_i)=\lambda_k x^i(\vec{e}^k\circ\vec{e}_i)=\lambda_k x^i\delta_{ii}=\lambda_i x^i=0 \tag{6.7.147}$$

The latter equality must be verified whatever the x^i, it is therefore necessary that each λ_i is zero and thus the vectors $\vec{e}^k$ are indeed linearly independent vectors.

■ Q.E.D.

The system of n reciprocal vectors forms a basis named the "reciprocal basis" (which is just the dual basis) of the vector space $\mathcal{E}^n$.

✍**Example:**

Given three vectors $\vec{e}_1,\vec{e}_2,\vec{e}_3$ forming a basis (not necessarily orthonormal!) of an euclidean space. We decide to define first:

$$v=\vec{e}_1\circ(\vec{e}_2\times\vec{e}_3) \tag{6.7.148}$$

where, for recall, the symbol × represents the cross product (see section Vector Calculus page 1830) and the whole is the mixed product as also seen in the section Vector calculus and represents an oriented volume.

The following vectors:

$$\boxed{\vec{e}^1=\frac{\vec{e}_2\times\vec{e}_3}{v}\qquad \vec{e}^2=\frac{\vec{e}_3\times\vec{e}_1}{v}\qquad \vec{e}^3=\frac{\vec{e}_1\times\vec{e}_2}{v}} \tag{6.7.149}$$

form the dual basis!

Remark

We recognize here the relations we have just proved earlier above during our study of the Gram's determinant!!!

Now, let us consider a vector on the original base $\vec{e}_1,\vec{e}_2,\vec{e}_3$ that we will denote by (as seen above):

$$\vec{x}=x^i\vec{e}_i \tag{6.7.150}$$

with therefore by definition the contravariant components of the vector that appear as we defined earlier above (and that we had at the same time explained the origin of the name). We also saw above that each contravariant component will also (naturally and by extension) be given by:

$$x^i=\vec{x}\circ\vec{e}^i \tag{6.7.151}$$

Similarly, so we have the covariant components that appear:

$$\vec{x}=x_i\vec{e}^i \tag{6.7.152}$$

$$x_i = \vec{x} \circ \vec{e}_i \tag{6.7.153}$$

In this approach, we then define the contravariant metric tensor and respectively covariant:

$$\vec{e}_i \circ e_j = g_{ij} = g_{ji} \tag{6.7.154}$$

$$\vec{e}^{\,i} \circ e^j = g^{ij} = g^{ji} \tag{6.7.155}$$

It comes therefore for example for the contravariant components (in the case of a three-dimensional space), knowing that the approach is the same for the covariant components:

$$\begin{aligned} x^j &= \vec{x} \circ \vec{e}^{\,i} = (x_1\vec{e}^{\,1} + x_2\vec{e}^{\,2} + x_3\vec{e}^{\,3}) \circ \vec{e}^{\,i} = x_1(\vec{e}^{\,1} \circ \vec{e}^{\,i}) + x_2(\vec{e}^{\,2} \circ \vec{e}^{\,i}) + x_3(\vec{e}^{\,3} \circ \vec{e}^{\,i}) \\ &= x_1 g^{1i} + x_2 g^{2i} + x_3 g^{3i} = x_j g^{ji} \end{aligned} \tag{6.7.156}$$

And so we find the transformation relations between the contravariant and covariant components already seen above with the difference that it seems coming out of a hat by successive definitions and that therefore hides the origin of the name of these components (at least in our point of view). But perhaps some readers prefer this approach...???

✍**Example:**

As an example, consider the basis:

$$\vec{e}_1 = 2\vec{i}_1 + \vec{i}_2 \qquad \vec{e}_2 = \vec{i}_1 + 2\vec{i}_2 \qquad \vec{e}_3 = \vec{i}_3 \tag{6.7.157}$$

Notice that it is not orthogonal because $\vec{e}_1 \circ \vec{e}_2 = 4 \neq 0$.

In this case we have by applying the previous Gram's relations:

$$\begin{aligned} \vec{e}^{\,1} &= \frac{2}{3}\vec{i}_1 + \frac{1}{3}\vec{i}_2 \\ \vec{e}^{\,2} &= -\frac{1}{3}\vec{i}_1 + \frac{2}{3}\vec{i}_2 \\ \vec{e}^{\,3} &= \vec{e}_3 = \vec{i}_3 \end{aligned} \tag{6.7.158}$$

As we can see in the figure below where $\vec{e}_1$ and $\vec{e}_2$ are shown in green, and the reciprocal vectors $\vec{e}^{\,1}$ and $\vec{e}^{\,2}$ are shown in blue:

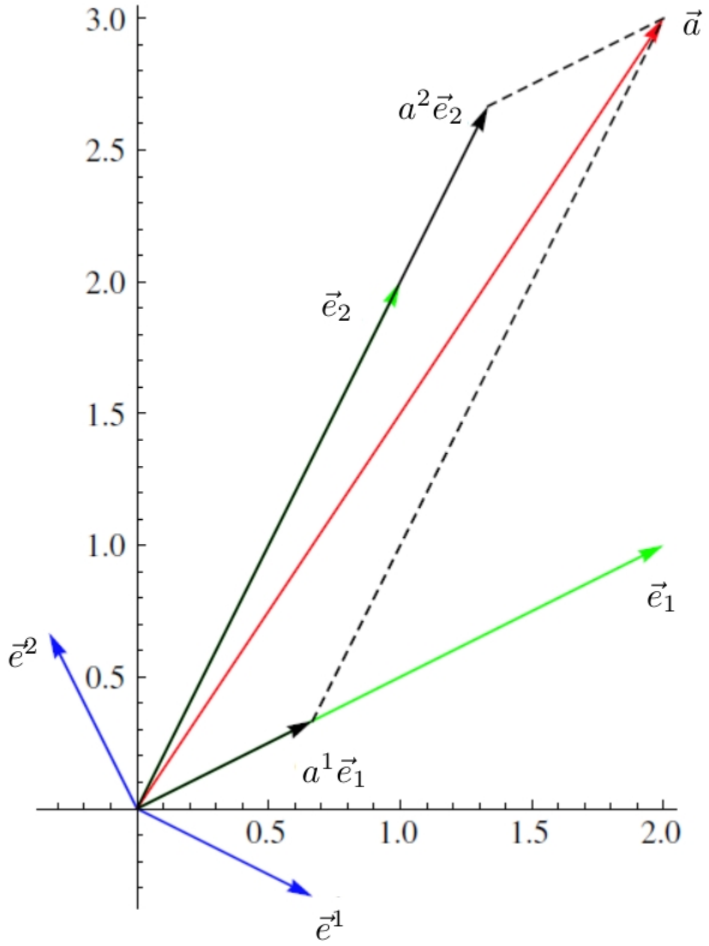


Figure 6.156: **Basis vectors, reciprocal basis vectors and contravariant components**

Notice that by construction we have indeed that $\vec{e}^1$ is orthogonal to $\vec{e}_2$ and $\vec{e}^2$ is orthogonal to $\vec{e}_1$!

For a given vector $\vec{a}$, say:

$$\vec{a} = 2\vec{\imath}_1 + 3\vec{\imath}_2 \qquad (6.7.159)$$

we can use the relation proved earlier to find its contravariant components in the basis $(\vec{e}_1, \vec{e}_2, \vec{e}_3)$. We get:

$$a^1 = \vec{a} \circ \vec{e}^1 = \frac{1}{3} \qquad a^2 = \vec{a} \circ \vec{e}^2 = \frac{4}{3} \qquad a^3 = \vec{a} \circ \vec{e}^3 = 0 \qquad (6.7.160)$$

So that (see figure above):

$$\vec{a} = \frac{1}{3}\vec{e}_1 + \frac{4}{3}\vec{e}_2 \qquad (6.7.161)$$

Now observe that the original basis vectors are reciprocal of the reciprocal ones. Thus we can just as well expand the same vector $\vec{a}$ along the basis vectors:

$$\vec{a} = a_1\vec{e}^1 + a_2\vec{e}^2 + a_3\vec{e}^3 \qquad (6.7.162)$$

with $a_i = \vec{a} \circ \vec{e}_i$.

The components a_i are the covariant components of $\vec{a}$. In our example, we get:

$$a_1 = \vec{a} \circ \vec{e}_1 = 7 \qquad a_2 = \vec{a} \circ \vec{e}_2 = 8 \qquad a_3 = \vec{a} \circ \vec{e}_3 = 0 \tag{6.7.163}$$

so we have:

$$\vec{a} = 7\vec{e}^1 + 8\vec{e}^2 \tag{6.7.164}$$

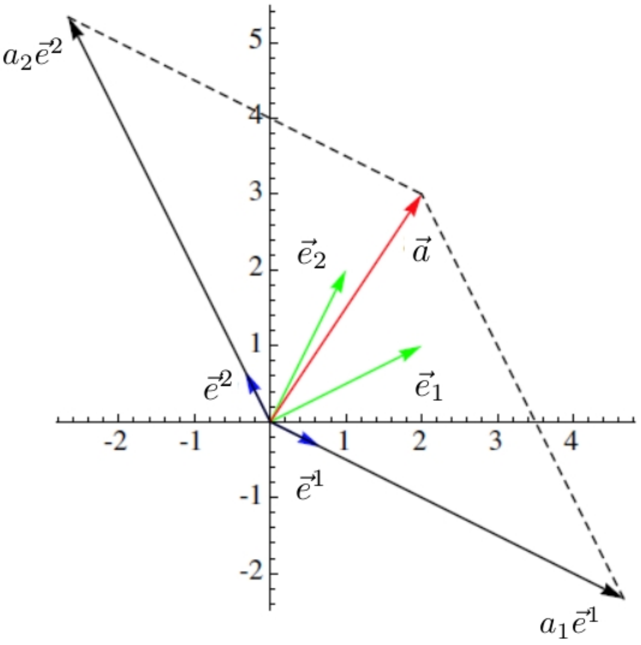


Figure 6.157: **Basis vectors, reciprocal basis vectors and covariant components**

6.7.7 Euclidean Tensors (cartesian tensor)

The generalization of the concept of vector has led us to the study of vector spaces to n dimensions. Tensors are also one-dimensional vectors but possess additional properties compared to vectors.

For the theoretical physicist, tensor calculus is interesting primarily in how the components of the tensor are transformed during a change of basis vector spaces from which they come. We will begin to study them vis-à-vis the properties of bases changes (because it is the most interesting case).

A tensor is in practice often only defined and used in the form of its components. These can be expressed in covariant or contravariant form like any vector. But a new type of components will appear in the tensor, it is the "mixed components". These three types of components are decomposition of Euclidean tensor on different bases.

Definition (#6.7.264): A "Cartesian tensor" uses an orthonormal basis to represent a tensor in a Euclidean space in the form of components.

Use of Cartesian tensors occurs in physics and engineering, such as with the Cauchy stress tensor (see section Continuum Mechanics page 3815) and the moment of inertia tensor in rigid body dynamics (see section Classical Mechanics page 3310). Sometimes general curvilinear coordinates are convenient, as in high-deformation continuum mechanics, or even necessary, as in General Relativity (see section General Relativity page 5720).

6.7.7.1 Fundamental Tensor

During the theory view earlier above, we used the quantities g_{ij}, defined from the scalar product of the basis vectors $(\vec{e}_i)$ of a n-dimensional pre-Euclidean vector space $\mathcal{E}^n$, by:

$$g_{ij} = \vec{e}_i \circ \vec{e}_j \tag{6.7.165}$$

These n^2 quantities are the covariant components of a tensor named the "fundamental tensor" or as we already know the "metric tensor".

Let us study now how vary the quantities g_{ij} when we make a basis change!

Given $(\vec{e}'_k)$ another base linked to the previous by the known relation:

$$\vec{e}_i = A'^k_i \vec{e}'_k \quad \text{and} \quad \vec{e}'_k = A^i_k \vec{e}_i \tag{6.7.166}$$

Substituting the relation $\vec{e}_i = A'^k_i \vec{e}'_k$ in the expression of g_{ij}, it come (we change the indices as it should be done during a substitution):

$$g_{ij} = (A'^k_i \vec{e}'_k) \circ (A'^l_j \vec{e}'_l) = (A'^k_i A'^l_j)(\vec{e}'_k \circ \vec{e}'_l) \tag{6.7.167}$$

In the new base $(\vec{e'}_k)$, the dot products of the basis vectors are therefore quantities such that:

$$g'_{kl} = \vec{e}'_k \circ \vec{e}'_l \tag{6.7.168}$$

So we finally have for the expression of the covariant components g_{ij} in a basis change:

$$\boxed{g_{ij} = (A'^k_i A'^l_j) g'_{kl}} \tag{6.7.169}$$

Identically we have:

$$\boxed{g'_{ij} = (A^k_i A^l_j) g_{kl}} \tag{6.7.170}$$

In general, a sequence of n^2 quantities t_{ij} that transforms, during a base change of $\mathcal{E}^n$, according to the two previous relations, namely:

$$t_{ij} = (A'^k_i A'^l_j) g'_{kl} \qquad \text{and} \qquad t'_{ij} = (A^k_i A^l_j) t_{kl} \tag{6.7.171}$$

are, by definition, the "covariant components of a tensor of order two" (with two indices) on $\mathcal{E}^n$.

We can therefore manipulate quantities expressing the intrinsic properties of bases as standard tensors!

6.7.7.2 Tensor product (dyadic) of two vectors and matrices

Let us consider an Euclidean vector space $\mathcal{E}^n$ of base $(\vec{e}_i)$ and given two vector of $\mathcal{E}^n$:

$$\vec{x}=x^i\vec{e}_i \quad \text{and} \quad \vec{y}=y^j\vec{e}_j \tag{6.7.172}$$

Let us form the two by two products of contravariant components x^i and y^j, namely:

$$u^{ij}=x^iy^j \tag{6.7.173}$$

We thus get n^2 quantities, if the two vectors have the same number of components, which are also the contravariant components of a tensor of order two named the "tensor product" of the vector $\vec{x}$ by the vector $\vec{y}$.

For example for $\vec{x}$ of dimension 2 and $\vec{y}$ of dimension 3 we have:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \otimes \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} x_1y_1 \\ x_1y_2 \\ x_1y_3 \\ x_2y_1 \\ x_2y_2 \\ x_2y_3 \end{pmatrix} \tag{6.7.174}$$

Remark

In quantum physics a "Fock space" defined as the tensor product of all n-body Hilbert spaces. Suppose we have a system described by a Hilbert space H, for example a single particle. The Hilbert space of two non-interacting particles of the same type as that described by H is simply the tensor product:

$$H^2:=H\otimes H \tag{6.7.175}$$

More generally, for a system of N particles as above, the Hilbert space is:

$$H^N:=\underbrace{H\otimes\cdots\otimes H}_{N\text{ times}} \tag{6.7.176}$$

We can also tensorally multiply two matrices A and B. Then, the matrix describing the tensor product $A\otimes B$ is the "Kronecker product" of the two matrices as we use it in Relativistic Quantum Physics.

For example, if:

$$A=\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \qquad B=\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \tag{6.7.177}$$

Then:

$$A\otimes B=\begin{bmatrix}a_{11}&a_{12}\\a_{21}&a_{22}\end{bmatrix}\otimes\begin{bmatrix}b_{11}&b_{12}\\b_{21}&b_{22}\end{bmatrix}=\begin{bmatrix}a_{11}\begin{bmatrix}b_{11}&b_{12}\\b_{21}&b_{22}\end{bmatrix}&a_{12}\begin{bmatrix}b_{11}&b_{12}\\b_{21}&b_{22}\end{bmatrix}\\a_{21}\begin{bmatrix}b_{11}&b_{12}\\b_{21}&b_{22}\end{bmatrix}&a_{22}\begin{bmatrix}b_{11}&b_{12}\\b_{21}&b_{22}\end{bmatrix}\end{bmatrix}$$
$$=\begin{bmatrix}a_{11}b_{11}&a_{11}b_{12}&a_{12}b_{11}&a_{12}b_{12}\\a_{11}b_{21}&a_{11}b_{22}&a_{12}b_{21}&a_{12}b_{22}\\a_{21}b_{11}&a_{21}b_{12}&a_{22}b_{11}&a_{22}b_{12}\\a_{21}b_{21}&a_{21}b_{22}&a_{22}b_{21}&a_{22}b_{22}\end{bmatrix}\qquad(6.7.178)$$

The most famous case is the covariant tensor of rank 2 in a space of 4 dimensions as it is the most used one in tensor calculus:

$$T_{ij}=\vec{x}\otimes\vec{y}=\begin{pmatrix}x_1\\x_2\\x_3\\x_4\end{pmatrix}\otimes\begin{pmatrix}y_1\\y_2\\y_3\\x_4\end{pmatrix}=\begin{pmatrix}x_1y_1&x_1y_2&x_1y_3&x_1y_4\\x_2y_1&x_2y_2&x_2y_3&x_2y_4\\x_3y_1&x_3y_2&x_3y_3&x_3y_4\\x_4y_1&x_4y_2&x_4y_3&x_4y_4\end{pmatrix}\qquad(6.7.179)$$

The reader can also now better understand the origin of the name of the gradient of a vector field (giving a "tensor field") as we saw in the section of Vector Calculus because we can rewrite it now:

$$\vec{\nabla}(\vec{f})=\frac{\partial f_i}{\partial x_j}=\begin{bmatrix}\dfrac{\partial f_1}{\partial x}&\dfrac{\partial f_1}{\partial y}&\dfrac{\partial f_1}{\partial z}\\\dfrac{\partial f_2}{\partial x}&\dfrac{\partial f_2}{\partial y}&\dfrac{\partial f_2}{\partial z}\\\dfrac{\partial f_3}{\partial x}&\dfrac{\partial f_3}{\partial y}&\dfrac{\partial f_3}{\partial z}\end{bmatrix}=\vec{\nabla}\otimes\vec{f}\qquad(6.7.180)$$

The reader will have certainly notice that through the examples above, the tensor product is non-commutative. That is:

$$\vec{x}\otimes\vec{y}\neq\vec{y}\otimes\vec{x}\qquad(6.7.181)$$

We can obviously build tensor products of order three (thus with n^3 terms) such as with the following tensor three times contravariant vectors:

$$u^{ijk}=x^iy^jz^k\qquad(6.7.182)$$

etc.

Let us study the properties of the basis changes of these components. Let us use for the basis changes relations of contravariant components of a vector, namely:

$$x^i=A_k^ix'^k\quad\text{and}\quad x'^k=A'^k_ix^i\qquad(6.7.183)$$

Let us replace in the relation $u^{ij}=x^iy^j$ the components x^i and y^i by their basis change expression, we get:

$$u^{ij}=x^iy^j=(A_k^ix'^k)(A_l^jy'^l)=A_k^iA_l^j(x'^ky'^l)=(A_k^iA_l^j)u'^{kl}\qquad(6.7.184)$$

The quantities u'^{kl} are the new components:

$$u'^{kl}=x'^{k}y'^{l} \tag{6.7.185}$$

The transformation formula of the n^2 quantities u^{ij} on a change of basis change of $\mathcal{E}^n$ is finally (very similar to metric tensor):

$$\boxed{u^{ij}=A_k^iA_l^ju'^{kl}} \tag{6.7.186}$$

Such a change basis relation characterizes the contravariant components of a tensor of order two. Conversely, we get:

$$\boxed{u'^{ij}=A'^{k}_{k}A'^{l}_{j}u^{ij}} \tag{6.7.187}$$

So the n^2 quantities are the "contravariant components of a tensor of order two".

We can then build the same products by pairs for covariant components x_i and y_i of the vectors $\vec{x}$ and $\vec{y}$ thus:

$$u_{ij}=x_iy_i \tag{6.7.188}$$

The formulas of basis change of the covariant components of the vectors are given by the following relations that we have already proved previously:

$$x_i=A'^{k}_{i}x'_{k}\quad \text{and}\quad x'_{k}=A_k^ix_i \tag{6.7.189}$$

Substituting the first relation in the product $u_{ij}=x_iy_i$, we get:

$$u_{ij}=(A'^{k}_{i}x'_{k})(A'^{l}_{j},y'_{l})=A'^{k}_{i}A'^{l}_{j}(x'_{k}y'_{l})=(A'^{k}_{i}A'^{l}_{j})u'_{kl} \tag{6.7.190}$$

This is the basis change relation of covariant components of a tensor of order two. We also easily check that we have:

$$u'_{kl}=A_k^iA_l^ju_{ij} \tag{6.7.191}$$

Identically we have of course $u'_{kl}=x'_{k}y'_{l}$ since $u_{ij}=x_iy_i$.

So the n^2 quantities are then the "covariant components of a tensor of order two".

Let us now create n^2 quantities my multiplying two by two the covariant components of a vector $\vec{x}$ by contravariant components of a vector $\vec{y}$, we get:

$$u_i^j=x_iy^j \tag{6.7.192}$$

Let us perform a basis change in this last relation taking into account the expressions $x_i=A'^{k}_{i}x'_{k}$ and $x^i=A_k^ix'^{k}$, we get:

$$u_i^j=x_iy^j(A'^{k}_{i}x'_{k})(A_l^jy'^{l})=A'^{k}_{i}A_l^j(x'_{k}y'^{l})=(A'^{k}_{i}A_l^j)u'^{l}_{k} \tag{6.7.193}$$

This basis change relation characterizes the "mixed components" of an order two tensor. Conversely, we can verify that we have:

$$u'^{l}_{k}=A_k^iA'^{l}_{j}u_i^j \tag{6.7.194}$$

These mixed components also constitutes the components of a tensor product of $\vec{x}$ by $\vec{y}$, according to a given basis.

In general, a sequence of n^2 quantities that transforms, during a basis of $\mathcal{E}^n$, just as previously established relation are therefore, by definition, "mixed components of a tensor of order two".

6.7.7.3 Tensor Spaces

In the previous study, we used as system of n^2 number, created from a vector space $\mathcal{E}^n$. When these numbers satisfy some basis change relations, we name these quantities, by definition, "components of a tensor".

We have seen that any linear combination of these components constitutes the components of a new tensor. We can therefore add together the components of the tensor and multiply by a scalar, to get other components of a new tensor. These addition and multiplication properties mean that we can use these tensors quantities as vector components.

To clarify how we define a tensor on a base, let us study the particular case of a tensor product of two vectors formed by triplets of numbers (that is to say in $\mathbb{R}^3$ typically). Consider therefore the Euclidean vector space $\mathcal{E}^3$ whose vectors are triplets of number of the form $\vec{x}=(x_1,x_2,x_3)$. The canonical orthonormal basis of $\mathcal{E}^3$ consists of three vectors that we know very well but written in tensor calculus as:

$$\vec{e}_i=(\delta_{1i},\delta_{2i},\delta_{3i}) \tag{6.7.195}$$

with $i=1,2,3$ (nice way to write simple things isn't it...).

Vectors of $\mathcal{E}^3$ gives the possibility to form the nine quantities that we have named the "components of the tensor product" of the vectors $\vec{x}$ and $\vec{y}$.

If we make all possible tensor products between vectors of $\mathcal{E}^3$, we get sequences of 9 numbers that can be used to define the following vector:

$$\vec{x}\otimes\vec{y}=\vec{U}=\left(x^1y^1,x^1y^2,x^1y^3,\ldots,x^3y^3\right) \tag{6.7.196}$$

Remark

We see immediately with the above relation and the previous relation that the tensor product is therefore not commutative.

We are left then with the elements of a vector space $\mathcal{E}^9$ with nine-dimensional space, whose elements all combinations by pairs of three numbers.

We then say that $\mathcal{E}^9$ has a "tensor product structure" which is denoted obviously and in standard calculus by $\mathcal{E}^9:\mathcal{E}^3\otimes\mathcal{E}^3$ or sometimes $\mathcal{E}_3^{(2)}$.

These vectors can be decomposed, for example, on an orthonormal canonical basis:

$$\vec{e}_k=(\delta_{1k},\delta_{2k},\ldots,\delta_{9k}) \tag{6.7.197}$$

with $k=1,2,\ldots,9$.

If we rewrite the quantities x^iy^j according to their place in the expression of $\vec{U}$, i.e.:

$$x^iy^j=u^{ij}=u^k \tag{6.7.198}$$

with $k=1,2,\ldots,9$ and $i,j=1,2,3$, the vectors $\vec{U}$ are then written:

$$\vec{U}=u^k\vec{e}_k \tag{6.7.199}$$

and is an example as we know of tensor of order 2 (obviously we can generalize this approach).

How do these tensor differ from ordinary vectors? Although they are identical to some vectors of $\mathcal{E}^9$ in our example but were formed by the vectors $\vec{x}$ and $\vec{y}$ of $\mathcal{E}^3$. To remember this fact, we write then as we already know:

$$\boxed{T_{ij}=\vec{U}=\vec{x}\otimes\vec{y}} \tag{6.7.200}$$

and they are named as we already know "tensor products of order two" of the vectors $\vec{x}$ and $\vec{y}$. The symbol $\otimes$ is defined in the way we have formed the quantities $x^iy^j=u^{ij}$ and in the order in which they were classified them to form the vector $\vec{U}$.

To recall the dependence between a quantity $x^iy^j=u^{ij}$ and the basis vector $\vec{e}_i$ to which he is assigned, let us rewrite these vectors by putting in place of the index k the two indices i and j, relative to the components, namely:

$$\vec{e}_k=\vec{e}_{ij} \tag{6.7.201}$$

The latter can be written in the form:

$$\boxed{\vec{e}_k=\vec{e}_{ij}=\vec{e}_i\otimes\vec{j}} \tag{6.7.202}$$

The vectors $\vec{e}_i\otimes\vec{j}$ generates a basis of $\mathcal{E}^9$ in the case of our example that is named the "tensor associated basis".

✍**Example:**

Consider:

$$\vec{e}_1=\begin{pmatrix}1\\0\\0\end{pmatrix},\ \vec{e}_2=\begin{pmatrix}0\\1\\0\end{pmatrix},\ \vec{e}_3=\begin{pmatrix}0\\0\\1\end{pmatrix} \tag{6.7.203}$$

We then have for example:

$$\vec{e}_{12}=\vec{e}_1\otimes\vec{e}_2=\begin{bmatrix}1\cdot 0 & 1\cdot 1 & 0\cdot 0\\ 0\cdot 0 & 0\cdot 1 & 0\cdot 0\\ 0\cdot 0 & 0\cdot 0 & 0\cdot 0\end{bmatrix}=\begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix} \tag{6.7.204}$$

That is to say:

$$\vec{e}_1:=\vec{e}_{12}=\begin{pmatrix}0\\1\\0\\0\\0\\0\\0\\0\\0\end{pmatrix} \tag{6.7.205}$$

It follows that as element of a space $\mathcal{E}^n \otimes \mathcal{E}^n$, a tensor $\vec{U}$ is a vector of the general form:

$$\boxed{\vec{U} = u^{ij}(\vec{e}_i \otimes \vec{e}_j)} \tag{6.7.206}$$

Let us study its properties vis-à-vis a base change of $\mathcal{E}^n$ such that:

$$\vec{e}_i = {A'}^k_i \vec{e}\,'_k = A^i_k \vec{e}_i \tag{6.7.207}$$

During such a base change, the base $(\vec{e}_i \otimes \vec{e}_j)$ associated to $\vec{e}_i$ becomes another base $(\vec{e}\,'_k \otimes \vec{e}\,'_l)$ associated to $\vec{e}\,'_k$, that is:

$$\vec{e}\,'_k \otimes \vec{e}\,'_l = (A^i_k \vec{e}_i) \otimes (A^j_l \vec{e}_j) = (A^i_k A^j_l) \otimes \vec{e}_j \tag{6.7.208}$$

It follows that the tensor product $\vec{U}$ has for components in the new basis:

$$\vec{U} = {u'}^{kl}(\vec{e}\,'_k \otimes \vec{e}\,'_l = {u'}^{kl}(A^i_k A^j_l)\vec{e}_i \otimes \vec{e}_j = ({A'}^k_i {A'}^l_j)u^{ij}(A^i_k A^j_l)\vec{e}_i \otimes \vec{e}_j \tag{6.7.209}$$

We have the following properties for the tensor product given:

$$\vec{x} \otimes \vec{y} = (x^i \vec{e}_i) \otimes (y^j \vec{e}_j) = x^i y^j (\vec{e}_i \otimes \vec{e}_j) \tag{6.7.210}$$

P1. Right/Left distributivity relatively to the addition of vectors:

$$\begin{aligned}(\vec{x}+\vec{y}) \otimes \vec{z} &= \vec{x} \otimes \vec{z} + \vec{y} \otimes \vec{z}\\ \vec{x} \otimes (\vec{y}+\vec{z}) &= \vec{x} \otimes \vec{y} + \vec{x} \otimes \vec{z}\end{aligned} \tag{6.7.211}$$

The proof of these relations is simply deduced from the definition of the tensor product. Indeed, we have for example:

$$\begin{aligned}(\vec{x}+\vec{y}) \otimes \vec{z} &= (x^i\vec{e}_i + y^i\vec{e}_i) \otimes z^k\vec{e}_k = (x^i + y^i)\vec{e}_i \otimes z^k\vec{e}_k = (x^i + y^i)z^k\vec{e}_i \otimes \vec{e}_k\\ &= x^i z^k \vec{e}_i \otimes \vec{e}_k + y^i z^k \vec{e}_i \otimes \vec{e}_k = \vec{x} \otimes \vec{z} + \vec{y} \otimes \vec{z}\end{aligned} \tag{6.7.212}$$

P2. Associativity with multiplication by a scalar:

$$(\lambda\vec{x}) \otimes \vec{y} = \vec{x} \otimes (\lambda\vec{y}) = \lambda(\vec{x} \otimes \vec{y}) \tag{6.7.213}$$

Indeed, we have:

$$(\lambda\vec{x}) \otimes \vec{y} = (\lambda x^i \vec{e}_i)\vec{x} \otimes y^i\vec{e}_i = \lambda x^i y^j (\vec{e}_i \otimes \vec{e}_j) = \lambda(\vec{x} \otimes \vec{y}) \tag{6.7.214}$$

P3. When we choose a base in each of the vector spaces $(\vec{e}_i)$ for $\mathcal{E}^n$, $(\vec{f}_i)$ for $\mathcal{F}^m$, the $n \cdot m$ elements of G_{nm} that we denote by $\vec{e}_i \otimes \vec{f}_i$ also form a basis of G_{nm}.

Proof 6.45.9. Already made in the particular example we used earlier above.

■ Q.E.D.

Remark

In practice, we often have to use tensor formed from vectors belonging to the same vector spaces $\mathcal{E}^n$.

We can of course generalize the tensor product to any number of vectors. Gradually, given the property P1, we can consider p vectors $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_p$ each belonging to different vector spaces $\mathcal{E}^{n_1}, \mathcal{E}^{n_2}, \ldots, \mathcal{E}^{n_p}$. If we have:

$$\vec{x}_1 = x^{i1}\vec{e}_{i1}, \vec{x}_2 = x^{i2}\vec{e}_{i2}, \ldots \tag{6.7.215}$$

we can form the tensor product:

$$\vec{x}_1 \otimes \vec{x}_2 \otimes \ldots \otimes \vec{x}_p = x^{i1}x^{i2}\ldots x^{ip}(\vec{e}_{i1} \otimes \vec{e}_{i2} \otimes \ldots \otimes \vec{e}_{ip}) \tag{6.7.216}$$

with $i_1 = \{1, \ldots, n_1\}, i_2 = \{1, \ldots, n_2\}, \ldots, i_p = \{1, \ldots, n_p\}$.

We build thus tensor products of order p belonging to the vector space $\mathcal{E}^{n_1} \otimes \mathcal{E}^{n_2} \otimes \ldots \otimes \mathcal{E}^{n_p}$, space that has a product structure tensor. The elements of this space are by definition tensor of order p.

In order to unify the classification, the elementary vector spaces, which can not be fitted with a tensor product structure can be regarded as having components of a tensor of order 1. In general, we name these elements "vectors", reserving the term "tensor" to elements of tensor spaces of order equal or greater than 2!

Remark

It is of usage to name "tensor of order zero" scalar quantities. It is also rare to meet tensor of order tensor greater than 2.

It is quite obvious and we will not do the proof that we absolutely can redefine all the concepts (base, decomposition on a base, reciprocal basis, dot product, tensor product) that we have seen so far considering tensor of order 1 as a vector (we should therefore rewrite everything that was already written above... which is useless in our point of view).

It is also quite possible to repeat all these definitions for higher order tensor and thus generalize the concept of space tensor for all dimensions.

From these considerations, we can state the "tensoriality criterion":

So that the elements of a sequence of n^p, relative to a base of a vector space $\mathcal{E}^n_{(p)}$, can be considered as the components of a tensor, it is necessary and sufficient that these quantities to be linked together, in two different bases of $\mathcal{E}^n_{(p)}$, by the relations of transformation of the components.

Remark

A vector can be represented in any base by a sequence of n components. However, we can not conclude that any sequence of n numbers is a vector. Indeed, when we put ourselves in another base of space, the components must also change to represent the same object, then we say that the vector is an intrinsic object (whose existence does not depend on the choice of the base). It remains then to know that a vector is a tensor of order 1.

6.7.7.4 Linear combination of tensors

We can form other tensor by combining together the components of various tensor products defined using vectors of the same vector space. For example, let us consider the contravariant components of the tensor products of the vectors $\vec{x},\vec{y}$ and $\vec{w},\vec{z}$:

$$u^{ij}=x^iy^j \qquad v^{ij}=w^iz^j \tag{6.7.217}$$

Let us form the following quantities:

$$t^{ij}=u^{ij}+\lambda v^{ij} \tag{6.7.218}$$

The n^2 quantities t^{ij} also satisfied the general formulas for basis change. We have indeed by substituting the relations of transformation of the contravariant components of a tensor product in the previous expression:

$$t^{ij}=A_k^iA_l^ju'^{kl}+\lambda A_k^iA_l^jv'^{kl}=A_k^iA_l^jt'^{kl} \tag{6.7.219}$$

The n^2 quantities t^{ij}, satisfying the relations of basis change also constitutes components of a tensor of order two.

6.7.7.5 Contraction of indices

Let us consider the mixed tensor product of two vectors $\vec{x}$ and $\vec{y}$ of respective contravariant x^i and covariant y_j components. The mixed components of the tensor product $\vec{V}$ of these two vectors are:

$$v_j^i=x^iy_j \tag{6.7.220}$$

Let us perform the addition of the various components of the tensor $\vec{V}$ such as $i=j$, i.e.:

$$v=x^iy_i \tag{6.7.221}$$

We thus get the expression of the dot product of vectors $\vec{x}$ and $\vec{y}$. The quantity v is a scalar (tensor of order zero). Such an addition on different variance indices constitutes, by definition, the operation of "contraction of indices" of the tensor $\vec{V}$. This operation allowed us to move from a tensor of order two to a tensor of order zero. The tensor $\vec{V}$ has been amputated and of a covariance and a contravariance.

Let us also take the example of a tensor $\vec{U}$ whose mixed components are one time covariant and one two times contravariant u_k^{ij} (caution ... it is not a three-dimensional matrix but simply an indication that the components of this tensor are expressed from three other variables!!!). Let us consider some of its components such as $k=j$, that is the components u_j^{ij} and let us perform the addition of the latter. We then get:

$$v^i=u_1^{i1}+u_2^{i2}+\ldots+u_n^{in}=\delta_j^ku_k^{ij} \tag{6.7.222}$$

These new quantities v^i form the components of a tensor $\vec{V}$ of order one (thus a vector!) and constitute what we name then the "contracted components" of the tensor $\vec{U}$ and of course meet the basis relations change (on request we can prove it but you have to know that it is similar to the one we made for vectors). So we have indeed change form a tensor of order three to a tensor of order one!

So we can see that the underlying idea of the contraction is to allow us to facilitate the resolution of a purely mathematical problem and depending on the situation it may be good to raise or reduce the order of a tensor. This is often a choice that is made by trial and error based on a specific context or that naturally arises from a purely mathematical or mathematical-physical development (as we will see examples further below).

6.7.7.5.1 Raising and lowering indices

If we start with a contravariant or covariant components tensor , we can lower / raise one or more indices by multiplication (if repeated) by g_{ij} or g^{ij} (unitary diagonal and positive signature: of canonical type) to get mixed components on which we can then perform contraction operations.

Let us consider an Euclidean tensor $\vec{U}$ of contravariant components $u^{i_1 i_2 i_3 \ldots i_p}$.

If we want to perform a contraction on this tensor, we will first need to transform it into a mixed tensor. This transformation we be done using a fundamental tensor.

Let us write $\vec{U}$ in mixed components by lowering at the covariant position the index i_1 by example (it is therefore equivalent to express this in contravariant component in covariant component). So:

$$u_{j_1}^{i_2 i_3 \ldots i_p} = g_{i_1, j_1} u^{i_1 i_2 i_3 \ldots i_p} \tag{6.7.223}$$

We see well that in this case to take down a contravariant index in a tensor using a fundamental tensor, we must first go search in the covariants components of the fundamental tensor the one who is itself in contravariant position in the original tensor and replace its position (but this time in covariance) by the other index of the fundamental tensor (it is the same idea when it we desire to operate a contraction on a covariant tensor).

Indeed, let us recall that we have proved that:

$$x_j = g_{ij} x^i \tag{6.7.224}$$

Also remember that raising and then lowering the same index (or conversely) are inverse operations, which is reflected in the covariant and contravariant metric tensors being inverse to each other:

$$g^{ij} g_{jk} = g_{kj} g^{ji} = \delta_k^i \tag{6.7.225}$$

Now that we got a mixed tensor components, we can very well getting contract the indices. Let us choose for example the index i_2 and let us perform the contraction with the index j_1, let us put $i_2 = j_1 = k$ (we are then concerned only to some specific terms), then just writing the whole process from the beginning:

$$g_{i_1 j_1} u^{i_1 i_2 i_3 \ldots i_p} = u_{j_1}^{i_2 i_3 \ldots i_p} = u_k^{k i_3 \ldots i_p} = v^{i_3 \ldots i_p} \tag{6.7.226}$$

So we get after lowering the index and one contraction, a tensor of order $p-2$.

✍**Examples:**

E1. Let us see a first example of raising down and contracting the covariant 4-position first order tensor given by (see section Special Relativity page 5639):

$$X_\mu = (-ct, x, y, z) \tag{6.7.227}$$

in components:

$$X_0 = -ct \qquad X_j = x_j \tag{6.7.228}$$

(where x_j are the usual Cartesian coordinates) and the Minkowski metric tensor with signature $(-+++)$ given by (see section Special Relativity page 5705):

$$\eta_{\mu\nu} = \eta^{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{6.7.229}$$

in components:

$$\eta_{00} = -1 \qquad \eta_{i0} = \eta_{0i} = 0 \qquad \eta_{ij} = \delta_{ij} \tag{6.7.230}$$

To raise the index, multiply by the tensor and contract:

$$X^\lambda = \eta^{\lambda\mu} X_\mu = \eta^{\lambda 0} X_0 + \eta^{\lambda i} X_i \tag{6.7.231}$$

Then for $\lambda = 0$:

$$X^0 = \eta^{00} X_0 + \eta^{0i} X_i = -X_0 \tag{6.7.232}$$

and for $\lambda = j = 1, 2, 3$:

$$X^j = \eta^{j0} X_0 + \eta^{ji} X_i = \delta^{ji} X_i = X_j \tag{6.7.233}$$

So the index-raised contravariant 4-position is:

$$X^\mu = (ct, x, y, z) \tag{6.7.234}$$

E2. For a second order tensor example let us consider the contravariant electromagnetic tensor in the $(+---)$ signature is given by (see section Electrodynamics page 4121):

$$F^{\alpha\beta} = \begin{pmatrix} 0 & -E_x/c & -E_y/c & -E_z/c \\ E_x/c & 0 & -B_z & B_y \\ E_y/c & B_z & 0 & -B_x \\ E_z/c & -B_y & B_x & 0 \end{pmatrix} \tag{6.7.235}$$

in components:

$$F^{0i} = -F^{i0} = -E^i/c \qquad F^{ij} = -\varepsilon^{ijk} B_k \tag{6.7.236}$$

To get the covariant tensor $F_{\alpha\beta}$, we multiply by the metric tensor and contract:

$$\begin{aligned}F_{\alpha\beta}&=\eta_{\alpha\gamma}\eta_{\beta\delta}F^{\gamma\delta}\\&=\eta_{\alpha0}\eta_{\beta0}F^{00}+\eta_{\alpha i}\eta_{\beta0}F^{i0}+\eta_{\alpha0}\eta_{\beta i}F^{0i}+\eta_{\alpha i}\eta_{\beta j}F^{ij}\end{aligned}\qquad(6.7.237)$$

and since $F^{00}=0$ and $F^{0i}=-F^{i0}$, this reduces to:

$$F_{\alpha\beta}=(\eta_{\alpha i}\eta_{\beta0}-\eta_{\alpha0}\eta_{\beta i})F^{i0}+\eta_{\alpha i}\eta_{\beta j}F^{ij}\qquad(6.7.238)$$

Now for $\alpha=0,\beta=k=1,2,3$:

$$\begin{aligned}F_{0k}&=(\eta_{0i}\eta_{k0}-\eta_{00}\eta_{ki})F^{i0}+\eta_{0i}\eta_{kj}F^{ij}\\&=(0-(-\delta_{ki}))F^{i0}+0\\&=F^{k0}=-F^{0k}\end{aligned}\qquad(6.7.239)$$

and by antisymmetry, for $\alpha=k=1,2,3$, $\beta=0$:

$$F_{0k}=-F_{k0}\qquad(6.7.240)$$

then finally for $\alpha=k=1,2,3$, $\beta=\ell=1,2,3$:

$$\begin{aligned}F_{k\ell}&=(\eta_{ki}\eta_{\ell0}-\eta_{k0}\eta_{\ell i})F^{i0}+\eta_{ki}\eta_{\ell j}F^{ij}\\&=0+\delta_{ki}\delta_{\ell j}F^{ij}\\&=F^{k\ell}\end{aligned}\qquad(6.7.241)$$

The (covariant) lower indexed tensor is then:

$$F_{\alpha\beta}=\begin{pmatrix}0&E_x/c&E_y/c&E_z/c\\-E_x/c&0&-B_z&B_y\\-E_y/c&B_z&0&-B_x\\-E_z/c&-B_y&B_x&0\end{pmatrix}\qquad(6.7.242)$$

E3. We will also see further below an example where we contract a tensor of order 1 (one of the contravariant components of the vectors of the spherical base) having already a lowered index:

$$\left.\begin{aligned}\omega_3^3&=\frac{\mathrm{d}r}{r}+\cot(\theta)\mathrm{d}\theta\\g_{ij}&=\begin{bmatrix}1&0&0\\0&r^2&0\\0&0&r^2\sin^2(\theta)\end{bmatrix}\end{aligned}\right\}\Rightarrow\begin{aligned}\omega_{33}&=g_{33}\omega_3^3=r^2\sin^2(\theta)\left(\frac{\mathrm{d}r}{r}+\cot(\theta)\mathrm{d}\theta\right)\\&=r^2\sin^2(\theta)\mathrm{d}r+r^2\sin^2(\theta)\cos(\theta)\mathrm{d}r\mathrm{d}\theta\end{aligned}\qquad(6.7.243)$$

Remarks

R1. In the relation:

$$g_{i_1j_1}u^{i_1i_2i_3\ldots i_p}=u_{j_1}^{i_2i_3\ldots i_p}=u_k^{ki_3\ldots i_p}=v^{i_3\ldots i_p}\qquad(6.7.244)$$

the equality is an abusive notation that we can find in some books (because strictly speaking we should do the calculation in two steps).

R2. As a result of the symmetry of the quantities g_{ij} (the dot product is commutative), this latter tensor is identical to what we would get by lowering to the covariant position the index i_2, and then by doing the contraction of the index i_1 with the index j_2.

Let see this:

The symmetry $g_{ij} = g_{ji}$ takes the form (this may seem confusing but let us remember that the number of the component i indicates the place of this component):

$$g_{i_1 j_1} = g_{i_2 j_2} \tag{6.7.245}$$

Therefore it comes:

$$g_{j_2 i_2} u^{i_1 i_2 i_3 \ldots i_p} = u_{j_2}^{i_1 i_3 \ldots i_p} \tag{6.7.246}$$

and putting $i_1 = j_2 = k$:

$$v^{i_3 i_4 \ldots i_p} = u_k^{k i_3 \ldots i_p} \tag{6.7.247}$$

In general, the contraction of a tensor allows to form a tensor of order $p-2$ from a tensor of order p. We can of course repeat the operation of contraction. Thus, an even tensor, $2p$, will become a scalar after p contractions and an odd order tensor $2p+1$, will become a vector.

We can extend after this definition of the contraction of indices, the tensoriality criterion. We have seen until now two ways to recognize the character of a tensor of a sequence of quantities:

- The first is to show that these quantities are formed by the tensor product of component vectors or by a sum of tensor products;
- The second is to study how these quantities are converted during a basis change and to check the conformity of the relations of transformation;
- The third and new one that brings to put that for a set of n^{p+1} quantities, having p upper and q lower indexes to be a tensor, it is necessary and sufficient that their product fully contracted by the contravariant components of any p vectors and the covariates components of any q vectors, to be a quantity (the norm in fact...) that remains invariant under basis change.

6.7.8 Special Tensors

We may face in theoretical physics and engineering with tensors that have interesting properties. To avoid redundant work in each case, we will list here and proved the various existing properties used in this textbook in other sections and discuss their possible implications briefly (the detailed analysis being reserved for their application in the other same sections of the book).

6.7.8.1 Symmetric Tensor

Consider a tensor $\vec{T}$ of order two of contravariant components T^{ij}. Let us suppose that, following a base $(\vec{e}_i)$, all these components satisfy the relations:

$$T^{ij}=T^{ji} \tag{6.7.248}$$

On another base $(\vec{e}\,'_j)$, related to the previous by the known transformation relations, the new components of T'^{lm} satisfy the relation:

$$T'^{lm}=A'^{l}_{i}A'^{m}_{j}T^{ij}=A'^{m}_{j}A'^{l}_{i}T^{ji} \tag{6.7.249}$$

We see that the property $T^{ij}=T^{ji}$ is therefore an intrinsic characteristic of the tensor $\vec{T}$, independent of the base! We then say that the tensor is a "symmetric tensor" (we will come back again on this concept a little further below) also named "totally invariant tensor" (implicitly meaning: by base change).

The symmetry property is also true for the covariant components of a symmetric tensor since we have:

$$T_{lk}=g_{li}g_{kj}T^{ij}=g_{kj}g_{li}T^{ji}=T_{kl} \tag{6.7.250}$$

Conversely, the symmetry of the covariant components implies that of the contravariant components.

For higher order tensor, the symmetry may be partial, being valid only for two covariant indices or two contravariant indices. Thus, a fourth order tensor, of mixed components T_l^{ijk}, may also be symmetrical in i and j, for example, given:

$$T_l^{ijk}=T_l^{jik} \tag{6.7.251}$$

We check, as above, that such a property is intrinsic.

A tensor is said to be "completely symmetrical tensor" if any transposition of two indices with the same variance, changes the corresponding component into itself. For example, for a tensor of order three T^{ijk}, completely symmetric, we have the following components that are equal:

$$T^{ijk}=T^{jik}=T^{kji}=T^{ikj} \tag{6.7.252}$$

There are many examples of symmetric tensors. Some include:

- the metric tensor $g_{\mu\nu}$ (see section General Relativity page 5731)

- the Einstein tensor $G_{\mu\nu}$ (see further below)
- the Ricci tensor $R_{\mu\nu}$ (see also further below)
- the stress and strain tensor for fluids or solids σ_{ij} (see section Continuum Mechanics page 3815)
- the Lorentz boost tensor Γ_{ν}^{μ} (see section Special Relativity page 5638)
- ...

We can also (very interesting curiosity) obtain a geometric representation of the values of the components of a symmetric tensor of order two!!

For this let us, consider in the ordinary geometric space coordinates x^i, the following equation:

$$a_{ij}x^ix^j=1 \tag{6.7.253}$$

where, for recall, x^ix^j can be seen as a tensor with $i,j=1,2,3$ and where the a_{ij} are real given coefficients. Let us suppose that coefficients are such that:

$$a_{ij}=a_{ji} \tag{6.7.254}$$

The above equation is then:

$$a_{11}x_1^2+a_{22}x_2^2+a_{33}x_3^2+2a_{12}x_1x_2+2a_{23}x_2x_3+2a_{31}x_3x_1=1 \tag{6.7.255}$$

Here we fall back on the equation of a surface of the second degree of a quadratic similar to of the plan that we saw in the section of Analytical Geometry (see page 2765). We know by extension into three dimensions that these surfaces are ellipsoids or hyperboloids, depending to the values of the quantities a_{ij}.

Let us study how the quantities a_{ij} transform when we make a change of coordinates such as:

$$x'^k={A'}^k_ix^i \qquad \text{and} \qquad x^i=A^i_kx'^k \tag{6.7.256}$$

The equation of the quadric is written in this new coordinate system:

$$a_{ij}A^i_kA^j_mx'^kx'^m={a'}_{km}x'^kx'^m=1 \tag{6.7.257}$$

Hence the expression of the coefficients in the new system of axes:

$${a'}_{km}=A^i_kA^j_ma_{ij} \tag{6.7.258}$$

The coefficients a_{ij} therefore transform as the covariant components of a tensor of order two. Conversely, if the quantities a_{ij} are the components of a symmetric tensor, these components define the coefficients of a quadric!! There is therefore a certain equivalence between a symmetric tensor and the coefficients of a quadratic...!!! We say then that the equation of the quadric is the "quadric representation of a symmetric tensor" or "representation surface".

So the representation surface (or representation quadric) is a geometrical representation of a second rank symmetric tensor and is useful for giving us a visual image of the tensor as well as being useful for example in calculating magnitudes of material properties described by second rank symmetric tensors!!

We know from our study of quadrics in the section of Analytical Geometry (by extending it to the three-dimensional case) that we can always find a coordinate system relative to which the equation of a quadratic takes a simpler form:

$$b_1x_1^2+b_2x_2^2+b_3x_3^2=1 \tag{6.7.259}$$

In this case, the basis vectors are supported by the main axes of the quadric. In this coordinate system, the components of the tensor equation reduce to:

$$a_{11}=b_1 \qquad a_{22}=b_2 \qquad a_{33}=b_3 \tag{6.7.260}$$

and $a_{ij}=0$ for the other components. The quantity b_i equation are named the "principal values" of the tensor a_{ij}.

If the quantities b_1,b_2,b_3 are positive, the surface is an ellipsoid, if two quantities are strictly positive and the third strictly negative, we have a one sheet hyperboloid, if two quantities are strictly negative and the third positive, we have two sheets hyperboloid (see section Analytical Geometry page 2844).

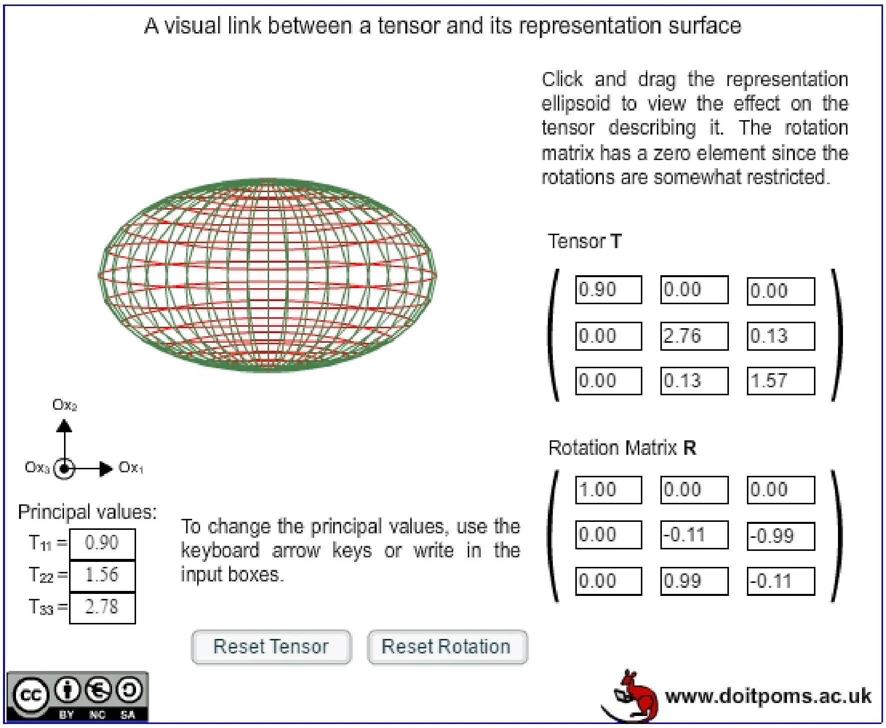
Comparing the expression of the quadric obtained previously with the classic equation:

$$\frac{x_1^2}{a^2}+\frac{x_2^2}{b^2}+\frac{x_3^2}{c^2}=1 \tag{6.7.261}$$

where a, b, c are the semi-axes of an ellipsoid shows that we have:

$$a=\frac{1}{\sqrt{b_1}} \qquad b=\frac{1}{\sqrt{b_2}} \qquad c=\frac{1}{\sqrt{b_b}} \tag{6.7.262}$$

Below we can see a screen shot of an interactive tool of Cambridge University to play with the ellipticity (only!) representation quadric of a rank two symmetric tensor:



A visual link between a tensor and its representation surface

Click and drag the representation ellipsoid to view the effect on the tensor describing it. The rotation matrix has a zero element since the rotations are somewhat restricted.

Tensor **T**

$$\begin{pmatrix} 0.90 & 0.00 & 0.00 \\ 0.00 & 2.76 & 0.13 \\ 0.00 & 0.13 & 1.57 \end{pmatrix}$$

Rotation Matrix **R**

$$\begin{pmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & -0.11 & -0.99 \\ 0.00 & 0.99 & -0.11 \end{pmatrix}$$

Principal values:

T_{11} = 0.90
T_{22} = 1.56
T_{33} = 2.78

To change the principal values, use the keyboard arrow keys or write in the input boxes.

Figure 6.158: **Visual link between a tensor and its representative surface (source: http://www.doitpoms.ac.uk)**

6.7.8.2 Antisymmetric Tensor

When the contravariant or covariant components of a tensor of order two, satisfy the property:

$$T^{ij} = -T^{ji} \tag{6.7.263}$$

we then say that the tensor is a "antisymmetric tensor" (the same applies if the components are covariant). In other words a tensor is antisymmetric on (or with respect to) an index subset if it alternates sign $(+/-)$ when any two indices of the subset are interchanged.

It is an intrinsic property of the tensor which is proved as for the symmetric tensors, to a given "-" sign. A contravariant tensor of order two, thus verifying the following relation, will be named a "symmetric tensor" (we have already mentioned it a bunch of times in the previous paragraphs).

It follows from this definition that an antisymmetric tensor must obviously satisfy the fact that its diagonal components are all zero, such as for example for a rank 2 tensor:

$$T^{11} = T^{22} = \ldots = T^{nn} = 0 \tag{6.7.264}$$

A well known antisymmetric tensor is the electromagnetic tensor $F_{\mu\nu}$ (see section Electrodynamics page 4121).

For example a covariant tensor of order three T_{ijk} will be say to be symmetric in i and k if for all values that can take the indices, we have:

$$T_{ijk} = T_{kji} \tag{6.7.265}$$

Or the fourth order covariant tensor T_{ijkl} will be said in antisymmetric on i and l if for all values that can take the indices, we have:

$$T_{ijkl} = -T_{ljki} \tag{6.7.266}$$

A tensor will "totally antisymmetric" if any transposition of index of same variance (covariant/contravariant) changes the corresponding component into its opposite.

A tensor T_{ij} can be put in the form of a sum of a symmetric tensor and an antisymmetric tensor. Indeed, we have:

$$\boxed{T_{ij} = \frac{1}{2}(T_{ij} + T_{ji}) + \frac{1}{2}(T_{ij} - T_{ji})} \tag{6.7.267}$$

The first term of the sum above is a symmetric tensor and the second, an antisymmetric tensor.

Proof 6.45.10. Consider first that T_{ij} is symmetric, then we have:

$$T_{ij} = \frac{1}{2}(T_{ij} + T_{ji}) + \frac{1}{2}\underbrace{(T_{ij} - T_{ji})}_{=0} = \frac{1}{2}(T_{ij} + T_{ji}) = \frac{1}{2}(T_{ij} + T_{ij}) = \frac{1}{2}T_{ij} = T_{ij} \tag{6.7.268}$$

So this proves that the left term is indeed symmetric for this special case.

Consider secondly that T_{ij} is symmetric, then we have:

$$T_{ij}=\frac{1}{2}\underbrace{(T_{ij}+T_{ji})}_{=0}+\frac{1}{2}(T_{ij}-T_{ji})=\frac{1}{2}(T_{ij}-T_{ji})=\frac{1}{2}2T_{ij}=T_{ij} \tag{6.7.269}$$

So this proves that the left term is indeed antisymmetric for this special case.

Now if T_{ij} is neither symmetric or antisymmetric we have:

$$T_{ij}=\frac{1}{2}(T_{ij}+T_{ji})+\frac{1}{2}(T_{ij}-T_{ji})=\frac{1}{2}T_{ij}+\frac{1}{2}T_{ji}+\frac{1}{2}T_{ij}-\frac{1}{2}T_{ji}=\frac{1}{2}2T_{ij}+\frac{1}{2}(T_{ji}-T_{ji})=\frac{1}{2}T_{ij} \tag{6.7.270}$$

■ Q.E.D.

A tensor T_l^{ijk} will be "partially antisymmetric if for example we have:

$$T_l^{ijk}=-T_l^{jik} \tag{6.7.271}$$

That is to say antisymmetric only for a subset of indices.

Let us now consider two vectors $\vec{x}=x^i\vec{e}_i$ and $\vec{y}=y^j\vec{e}_j$ of a vector space $\mathcal{E}^n$. Let us form the following antisymmetric quantities (we can see in it two tensor products):

$$T^{ij}=x^iy^j-x^jy^i \tag{6.7.272}$$

where we see immediately that the components T^{ij} are those of an antisymmetric tensor $\vec{T}$ by construction as:

$$x^iy^j-x^jy^i=-(x^jy^i-x^iy^j) \tag{6.7.273}$$

The decomposition of the vector $\vec{T}$ in the base $\vec{e}_i\vec{e}_j$ is:

$$\vec{T}=(\vec{x}\times\vec{y})=(x^iy^j-x^jy^i)\vec{e}_i\otimes\vec{e}_j=\vec{x}\otimes\vec{y}-\vec{y}\otimes\vec{x} \tag{6.7.274}$$

The tensor $\vec{x}\otimes\vec{y}$ (written as it in analogy to the vectors cross product for $n=3$) is named the "outer product" of the vectors $\vec{x}$ and $\vec{y}$. We say that this tensor is a "bivector".

The outer product is an antisymmetric tensor which satisfies the following properties:

P1. Anti-commutativity:

$$\vec{x}\times\vec{y}=-\vec{y}\times\vec{x} \tag{6.7.275}$$

the result is:

$$\vec{x}\times\vec{x}=\vec{0} \tag{6.7.276}$$

P2. Left and right distributivity for the vector addition:

$$\begin{aligned}\vec{x}\times(\vec{y}+\vec{z})&=\vec{x}\times\vec{y}+\vec{x}\times\vec{z}\\(\vec{x}+\vec{y})\times\vec{y}&=\vec{x}\times\vec{z}+\vec{y}\times\vec{z}\end{aligned} \tag{6.7.277}$$

P3. Associativity for the multiplication by a scalar:

$$\alpha\vec{x}\times\vec{y}=\vec{x}\times\alpha\vec{y}=\alpha(\vec{x}\times\vec{y}) \tag{6.7.278}$$

P4. Outer products:

$$\vec{e}_i\times\vec{e}_j=\vec{e}_i\otimes\vec{e}_j-\vec{e}_j\otimes\vec{e}_i \tag{6.7.279}$$

constitute a base for all bivectors.

Proof 6.45.11. An antisymmetric tensor $\vec{T}$ of order two, element of $\mathcal{E}^2_{(2)}$ can, as we have proved it earlier, be written as:

$$\vec{T}=T^{ij}\vec{e}_i\otimes\vec{e}_j=\sum_{i<j}T^{ij}\vec{e}_i\otimes\vec{e}_j+\sum_{i\geq j}\vec{e}_i\otimes\vec{e}_j \tag{6.7.280}$$

Exchanging, in the last sum of the above relation, the name of the indices and considering that $T^{ij}=-T^{ji}$, we get:

$$\vec{U}=\sum_{i<j}u^{ij}(\vec{e}_i\otimes\vec{e}_j-\vec{e}_j\otimes\vec{e}_i) \tag{6.7.281}$$

The elements:

$$\vec{e}_i\otimes\vec{e}_j-\vec{e}_j\otimes\vec{e}_i \tag{6.7.282}$$

are linearly independent vectors since the vectors $\vec{e}_i\otimes\vec{e}_j$ also are it. These elements constitute a base on which the antisymmetric tensor can be decomposed.

■ Q.E.D.

The number of distinguishable vectors $\vec{e}_i\otimes\vec{e}_j-\vec{e}_j\otimes\vec{e}_i$ is equal to the number of combinations of vectors taken two by two and distinguishable among n such that (see section Probabilities page 528):

$$C_2^n=\frac{n!}{2!(n-2)!}=\frac{n(n-1)}{2}<n^2 \tag{6.7.283}$$

Indeed among the n^2 components, n components are equal to zero and $n(n-1)$ other components have opposed values to two by two. So we can consider that half of the latter is sufficient to characterize the tensor.

In the context of the outer tensor product where we have:

$$T^{ij}=x^iy^j-x^jy^i \tag{6.7.284}$$

the number of distinguishable components is also equal to $n(n-1)/2$ and they are named "strict components".

We notice that for $n=3$, the strict number of components of the outer product of two vectors is also equal to three. This allows to form, with the components of the bivector, the components of a cross product $\vec{z}$.

Thus, a cross product therefore exists only for a subspace of bivectors whose number of dimensions is equal to 3 and the pre-images that are antisymmetric tensors.

If all these conditions are satisfied, we say that the vector $\vec{z}$ is the "adjoint tensor" of tensor $\vec{T}$.

6.7.8.3 Fundamental Tensor

We saw at the beginning of our study of Tensor Calculus the definition of the components of the fundamental covariant tensor g_{ij}, that is for recall:

$$g_{ij}=\vec{e}_i\circ\vec{e}_j \tag{6.7.285}$$

These quantities are involved, as we know it (see previous topics), in the expression of the dot product of two vectors $\vec{x}$ and $\vec{y}$ of contravariant components x^i and y^i, given by the relation (for recall):

$$\vec{x}\circ\vec{y}=g_{ij}x^iy^i \tag{6.7.286}$$

Let us use the general test of tensoriality to highlight the tensor character of the g_{ij}. The previous expression is a product fully contracted of the quantities g_{ij} with the contravariant components x^iy^i of an arbitrary tensor. As the dot product is an invariant quantity (in this case a scalar) with respect to the base changes, it follows that the n^2 quantities g_{ij} are the covariant components of a tensor.

This tensor is also symmetrical as a result of the symmetry property of the dot product of the basis vectors such that:

$$g_{ij}=g_{ji} \tag{6.7.287}$$

We have the same for the contravariant components of the fundamental tensor:

$$g^{ij}=g^{ji} \tag{6.7.288}$$

If we denote by g_j^i the mixed components of a fundamental tensor to himself:

$$g_j^i=g^{ik}g_{jk} \tag{6.7.289}$$

obviously with the canonical basis:

$$g_j^i=\delta_j^i \tag{6.7.290}$$

6.7.9 Curvilinear Coordinates

The conventional concepts of coordinate system can be generalized as we know to any specific n-dimensional punctual space (see section Principia page 3145). We name "coordinate system" in $\mathcal{E}^n$, any mode of definition of a point M in the considered system.

Definitions (#6.7.265):

D1. For a given coordinates system (Cartesian, spherical, cylindrical, polar ...) system we name "coordinate line" the "place" of the points M when a only single coordinate of M varies, the other being keep as constant.

D2. A "curvilinear coordinates" are a coordinate system for Euclidean space in which the coordinate lines may be curved. These coordinates may be derived from a set of Cartesian coordinates by using a transformation that is locally invertible (a one-to-one map) at each point (see section Vector Calculus page 1846).

This means that one can convert a point given in a Cartesian coordinate system to its curvilinear coordinates and back. The purpose as we already know is that depending on the application, a curvilinear coordinate system may be simpler to use than the Cartesian coordinate system. For instance, a physical problem with spherical symmetry defined (for example, motion of particles under the influence of central forces) is usually easier to solve in spherical polar coordinates than in Cartesian coordinates.

Well-known examples of curvilinear coordinate systems are as we know in three-dimensional Euclidean space the Cartesian, cylindrical and spherical polar coordinates.

Let us first study the generalization of a coordinate system relative to a fixed reference frame (we urge the reader to consult first the subsection about Coordinate Systems in the section of Vector Calculus and the subsection of Analytical Mechanics in the section Principia).

Let us consider a punctual space $\mathcal{E}^n$ and $(\vec{e}_i)$ a reference frame of that space. Given x_i the rectilinear coordinates of a point M of $\mathcal{E}^n$ relatively to this reference frame. Any curvilinear coordinate system u^k, with $k=1,2,\ldots,n$, is obtained by giving n arbitrary functions f^i of parameters u^k, such that:

$$x^i = f^i(u^1,u^2,\ldots,u^n) \tag{6.7.291}$$

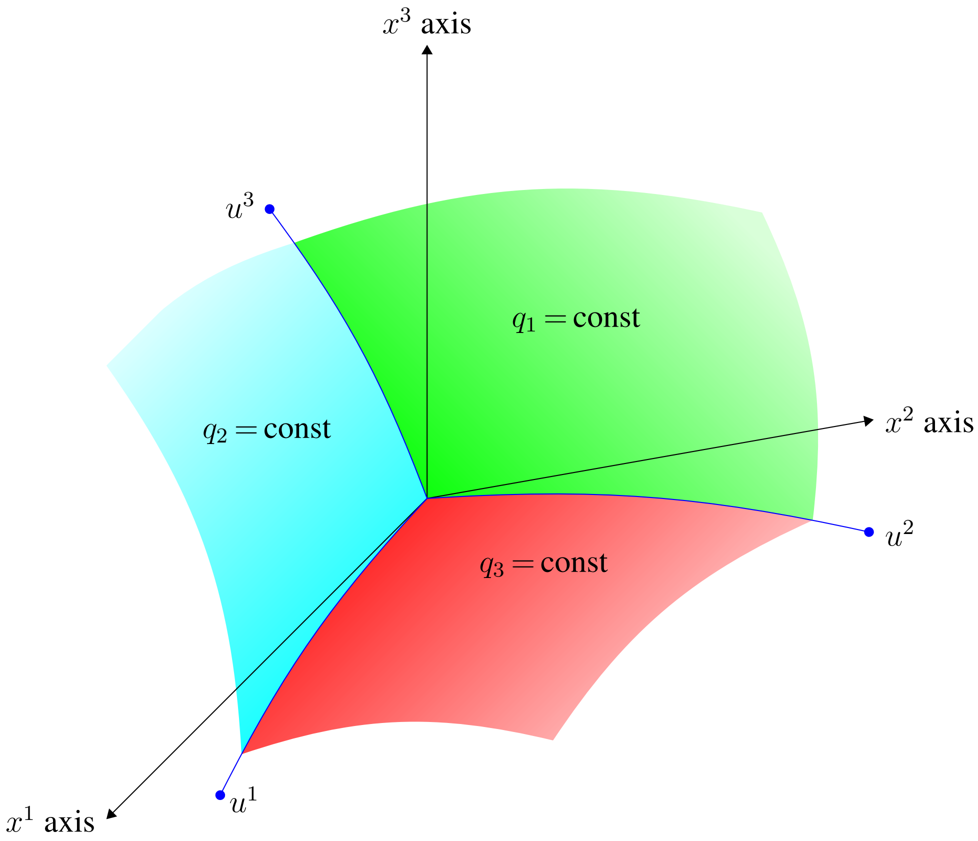


Figure 6.159: **Coordinate surfaces, coordinate lines and coordinate axes of general curvilinear coordinates (source: Wikipedia, author: Lyudmil Antonov)**

We will assume thereafter that these n functions satisfy the following three properties:

P1. They must be of class at least $\mathcal{C}^2$ (differentiable at least twice for the needs of physics: speed and acceleration). This assumption implies, that at any point where it is satisfied, that we have the permutation of derivations (with respect to the two partial derivatives as seen in the section of Differential and Integral Calculus):

$$\partial_{kl} f^i = \partial_{lk} f^i \tag{6.7.292}$$

where for reminder:

$$\frac{\partial f^a}{\partial x^b} = \lim_{h \to 0} \frac{V^a\left(\ldots, x^b + h, \ldots\right) - V^a\left(\ldots, x^b, \ldots\right)}{h} = \partial_b V^a \tag{6.7.293}$$

with $V^a = \vec{V} \cdot \vec{e}^a$.

P2. These functions are such that we can solve the system of n equations of coordinate system change relatively to the variables u^k and express them in function of the parameters x^i, thus:

$$u^k = g^k(x^1, x^2, \ldots, x^n) \tag{6.7.294}$$

still with $k = 1, 2, \ldots, n$.

P3. When the variables u^k vary in a domain Δ, the variables x^i vary in a domain Δ' (think to cartesian↔ spherical coordinates where in cartesian the variable range is infinite when in spherical both of them are limited a typical 2π width range).

P4. The Jacobian of the functions $x^i=f^i(u^1,u^2,\ldots,u^n)$, defined by (see section Differential and Integral Calculus page 1524):

$$\mathrm{D}(\partial_k x^i)=\begin{bmatrix}\partial_1 x^1 & \partial_1 x^2 & \ldots & \partial_1 x^n\\ \partial_2 x^1 & \partial_2 x^2 & \ldots & \partial_2 x^n\\ \ldots & \ldots & \ldots & \ddots\\ \partial_n x^1 & \partial_n x^2 & \ldots & \partial_n x^n\end{bmatrix} \tag{6.7.295}$$

will be assumed different from zero in the domain Δ (and also the Jacobian $\mathrm{D}(\partial_i u^k)$ of the functions $u^k=g^k(x^1,x^2,\ldots,x^n)$) and is the inverse of the Jacobian of $u^k=g^k(x^1,x^2,\ldots,x^n)$. If the jacobians exist, they are not zero as a result primarily of the second property above and implicitly the first.

As we have already mention if we fix $(n-1)$ parameters u^k by varying only one parameter, u^i for example, we get the coordinates $x^i_{(1)}$ of a set of points M that constitute a "coordinate line line". In general, the coordinate lines are not straight but curved as we know. These coordinates u^k are named for this reasons the "curvilinear coordinates". On a point M of $\mathcal{E}^n$ intersect elsewhere n coordinate lines.

We proved in the section of Analytical Mechanics, during our study of punctual spaces (see page 3145), that partial derivatives of a vector $\overrightarrow{\mathrm{O}M}$ of $\mathcal{E}^n$ are independent of the point O (origin) of a given reference frame. If $\mathcal{E}^n$ is assimilated to a system of curvilinear coordinate u^k, we write:

$$\boxed{\vec{e}_k=\frac{\partial\overrightarrow{\mathrm{O}M}}{\partial u^k}=\partial_k\vec{M}} \tag{6.7.296}$$

✍**Example:**

A famous example of curvilinear coordinates u^k, where each u^k is a uniform function of the rectilinear coordinates x^k, the u^k being moreover continuous functions at the point M, is that of the spherical coordinates where we have (see section Vector Calculus page 1849):

$$u^1=r \qquad u^2=\theta \qquad u^3=\phi \tag{6.7.297}$$

and where for recall:

$$x_1=r\sin(\theta)\cos(\phi) \qquad x_2=r\sin(\theta)\sin(\phi) \qquad x_3=r\cos(\theta) \tag{6.7.298}$$

Let us also recall that during our study of the system of spherical coordinates in the section of Vector Calculus we obtained after normalization of the basis vector:

$$\mathrm{d}\vec{M}=\vec{e}_r\mathrm{d}r+\vec{e}_\theta r\mathrm{d}\theta+\vec{e}_\phi r\sin(\theta)\mathrm{d}\phi \tag{6.7.299}$$

Therefore:

$$\frac{\partial\vec{M}}{\partial r}=\vec{e}_r \qquad \frac{\partial\vec{M}}{\partial\theta}=\vec{e}_\theta r \qquad \frac{\partial\vec{M}}{\partial\phi}=\vec{e}_\phi r\sin(\theta) \tag{6.7.300}$$

with:

$$\left\|\frac{\partial\vec{M}}{\partial r}\right\|=1 \qquad \left\|\frac{\partial\vec{M}}{\partial\theta}\right\|=r \qquad \left\|\frac{\partial\vec{M}}{\partial\phi}\right\|=r\sin(\theta) \tag{6.7.301}$$

In a non-Euclidean space, we can not define a unique valid basis over the whole space. Thus, we construct a base at each point separately and for this purpose we use the curvilinear coordinates such that at each point M, the base vectors $\vec{e}_k$ are tangent to the corresponding coordinate line equation via the relation given above:

$$\vec{e}_k=\frac{\partial\overrightarrow{\text{O}M}}{\partial u^k}=\partial_k\vec{M} \tag{6.7.302}$$

Given now $u^1,u^2,\ldots,u^n$ the curvilinear coordinates of the point M with respect to a Cartesian basis $(\vec{e}_i^{\,0})$. In this reference frame, we obviously have:

$$\overrightarrow{\text{O}M}=x^i\vec{e}_i^{\,0} \tag{6.7.303}$$

where the Cartesian coordinates are functions $x^i=x^i(u^1,u^2,\ldots,u^n)$.

The vector $\vec{e}_k$ can therefore also be expressed by:

$$\vec{e}_k=\partial_k(x^i\vec{e}_i^{\,0})=(\partial_k x^i)\vec{e}_i^{\,0} \tag{6.7.304}$$

The reader can check with the example of spherical coordinates (by looking to the explicit version of the corresponding vectors in the section of Vector Calculus) that this relation is correct but at the condition that we work with the non-normalized version of the orthogonal vectors $\vec{e}_r$, $\vec{e}_\theta$, $\vec{e}_\phi$!

From the components $\partial_k x^i$ of the vector $\vec{e}_k$, we can form a determinant $\det(\partial_k x^i)$ which is precisely the Jacobian of the of the functions x^i we have defined previously. Since this determinant is different from zero (at least imposed as such), it follows that the n vectors $\vec{e}_k$ (as functions) are linearly independent (we have proved in the section of Differential and Integral Calculus that this Jacobian is in absolute value equal to $r^2\sin(\theta)$ for the spherical coordinates).

Remark

Let us recall that in the section of Differential and Integral Calculus we have proved that the determinant of the Jacobian appears when calculating the surface of a parallelogram in a non-euclidean space. Obviously is the determinant is zero, then the surface is zero as it means that the vector basis of the parallelogram are all collinear (linearly dependants). Hence the fact that if the determinant is not null then some of the vectors (or all) are independent!

These n vectors, defined by the relation:

$$\vec{e}_k=\partial_k\vec{M} \tag{6.7.305}$$

are named the "natural basis" at the point M of the vector space $\mathcal{E}^n$. They are collinear to the tangents of the n coordinated lines which intersect at the point M where they are defined.

We will not insist on the quite obvious fact that at each system of curvilinear coordinates there is an associated natural basis whose base is expressed by these same coordinates (see section Vector Calculus page 1846).

✍**Example:**

In spherical coordinates, the vectors of the natural basis are those that we obtained in our study of the spherical coordinate system in the section of Vector Calculus and that are orthogonal but not orthonormal.

Let us associate at the point M of $\mathcal{E}^n$ a reference frame formed by the point M and by the vectors of the natural basis[32]. This reference frame is named the "natural reference frame" on M of the coordinate system u^k. It will be denoted by:

$$(M,\vec{e}_k) \qquad \text{or} \qquad (M,\partial_k\vec{M}) \tag{6.7.306}$$

The differential of the vector $\overrightarrow{\mathrm{O}M}$ is then expressed as:

$$\mathrm{d}\vec{M}=\partial_k\vec{M}\mathrm{d}u^k=\vec{e}_k\mathrm{d}u^k \tag{6.7.307}$$

The quantities $\mathrm{d}u^k$ are (obviously) the contravariant components of the vector $\mathrm{d}\vec{M}$ in the natural reference frame $(M,\vec{e}_k)$ of the coordinate system u^k.

Let us now consider any two curvilinear coordinate systems u^i and u^k (think for example to the spherical and cylindrical coordinate systems), linked between them by the relations:

$$u^i=u^i\left(u'^1,u'^2,\ldots,u'^n\right) \qquad u'^k=u'^k\left(u^1,u^2,\ldots,u^n\right) \tag{6.7.308}$$

where the functions $u^i=u^i(u'^1,u'^2,\ldots,u'^n)$ are assumed as we already know to be several times continuously differentiable with respect to the u'^k and same for the functions $u'^k=u'^k(u^1,u^2,\ldots,u^n)$ with respect the coordinates u^i. When we move from one coordinate system to another, we say that we make a "change of curvilinear coordinates".

We saw that in the sections of Differential Geometry and General Relativity that the squared distance $\mathrm{d}s$ between two points M and M' infinitely close was given by the relation:

$$\boxed{\mathrm{d}s^2=g_{ij}\mathrm{d}x^i\mathrm{d}x^j} \tag{6.7.309}$$

where the $\mathrm{d}x^i$ are the components of the vector $\mathrm{d}\vec{M}=\overrightarrow{MM}$, assimilated to a fixed reference frame of a punctual space $\mathcal{E}^n$. When this space is assimilated to a system of curvilinear coordinates u^i, we have seen that the relation:

$$\mathrm{d}\vec{M}=\partial_k\vec{M}\mathrm{d}u^k=\vec{e}_k\mathrm{d}u^k \tag{6.7.310}$$

shows that the vector $\mathrm{d}\vec{M}$ has for curvilinear contravariant components the quantities $\mathrm{d}u^i$ with respect to the natural base $(M,\vec{e}_i$. The square of the distance $\mathrm{d}s^2$ (i.e. the "line-elements") is then written in the natural reference frame:

$$\mathrm{d}s^2=g_{ij}\mathrm{d}u^i\mathrm{d}u^j \tag{6.7.311}$$

---

[32] For recall that natural basis in an Euclidean space is the set of unit vectors pointing in the direction of the axes of a Cartesian coordinate system

Where the quantities $g_{ij}=\vec{e}_i\circ\vec{e}_j$ are the components of the "fundamental tensor" or of "metric tensor" defined using a natural base. The previous expression is named the "linear element of the point space" $\mathcal{E}^n$ or sometimes the "metric" of this space.

The vectors $\vec{e}_i$ of the natural reference frame generally vary from one point to another. This is the case, for example, of the spherical coordinates whose quantities g_{ij} (we show it afterwards with a detailed example) are variable since depending on the parameters r, θ or ϕ!

A curve Γ of $\mathcal{E}^n$ can be defined by the data of the curvilinear coordinates $u^i(\alpha)$ of the locus of the points $M(\alpha)$ as a function of a parameter α (see section Differential Geometry page 2850). The elementary distance $\mathrm{d}s$ on this curve Γ is then written:

$$\mathrm{d}s=\left[g_{ij}\frac{\partial u^i}{\partial\alpha}\frac{\partial u^j}{\partial\alpha}\right]^{1/2}\mathrm{d}\alpha \tag{6.7.312}$$

Let us consider a tangent vector $\vec{v}$ of Γ and denote its components v^i on a basis $\vec{e}_i$. On another basis $\vec{e}_i'$ we denote the components v'^i, so:

$$\vec{v}=v^i\vec{e}_i=v'^i\vec{e}_i' \tag{6.7.313}$$

in which:

$$v^i=\frac{\mathrm{d}x^i}{\mathrm{d}\alpha}\text{ and }v'^i=\frac{\mathrm{d}x^{ii}}{\mathrm{d}\lambda} \tag{6.7.314}$$

If we express the new components in terms of the old ones, then:

$$v'^i=\frac{\mathrm{d}x'^i}{\mathrm{d}\alpha}=\frac{\partial x'^i}{\partial x^j}\frac{\mathrm{d}x^j}{\mathrm{d}\alpha}=\frac{\partial x'^i}{\partial x^j}v^j \tag{6.7.315}$$

This is the explicit form of a transformation named the "contravariant transformation". In order to distinguish them such vectors from the covariant (tangent) vectors, the index is placed on top.

A famous example of a contravariant transformation is given by a differential form $\mathrm{d}f$. For f as a function of coordinates x^i, $\mathrm{d}f$ can be expressed in terms of $\mathrm{d}x^i$ (i.e. the total derivative). The differentials $\mathrm{d}x^i$ transform according to the contravariant rule since:

$$\boxed{\mathrm{d}x'^i=\frac{\partial x'^i}{\partial x^j}\mathrm{d}x^j} \tag{6.7.316}$$

6.7.9.1 Natural basis in spherical coordinates (curvilinear basis in spherical coordinates)

Let us determine the natural basis of the vector space E^3 associated with the point space $\mathcal{E}^3$ of ordinary geometry, in spherical coordinates. Let us write the expression of the vectors $\overrightarrow{\text{O}M}$ in a fixed cartesian basis $(\vec{e}_i^{\,0})$ which is by definition (see the section Vector Calculus for more details):

$$\overrightarrow{\text{O}M}=x^i\vec{e}_i^{\,0}=r\sin(\theta)\cos(\phi)\vec{e}_1^{\,0}+r\sin(\theta)\sin(\phi)\vec{e}_2^{\,0}+r\cos(\theta)\vec{e}_3^{\,0} \tag{6.7.317}$$

The vectors of the natural base are given by:

$$\vec{e}_k=\partial_k\vec{M}\Rightarrow\begin{cases}\vec{e}_1=\partial_1\vec{M}=\partial_r\vec{M}\\ \vec{e}_2=\partial_2\vec{M}=\partial_\theta\vec{M}\\ \vec{e}_3=\partial_3\vec{M}=\partial_\phi\vec{M}\end{cases}\tag{6.7.318}$$

Therefore we have:

$$\vec{e}_1=\partial_1\vec{M}=\sin(\theta)\cos(\phi)\vec{e}_1^{\,0}+\sin(\theta)\sin(\phi)\vec{e}_2^{\,0}+\cos(\theta)\vec{e}_3^{\,0}\tag{6.7.319}$$

The derivative of $\overrightarrow{\mathrm{O}M}$ with respect to θ gives the vector $\vec{e}_2$:

$$\vec{e}_2=\partial_2\vec{M}=r\cos(\theta)\cos(\phi)\vec{e}_1^{\,0}+r\cos(\theta)\sin(\phi)\vec{e}_2^{\,0}-r\sin(\theta)\vec{e}_3^{\,0}\tag{6.7.320}$$

The derivative with respect to φ gives the vector $\vec{e}_3$:

$$\vec{e}_3=\partial_3\vec{M}=-r\sin(\theta)\sin(\phi)\vec{e}_1^{\,0}+r\sin(\theta)\cos(\phi)\vec{e}_2^{\,0}\tag{6.7.321}$$

These three vectors are orthogonal to each other as we can easily verify it by performing the dot products $\vec{e}_i\circ\vec{e}_j$. When this is the case, we say that the coordinates are "orthogonal curvilinear coordinates" (see section Vector Calculus page 1849).

We thus find the same result as in the section of Vector Calculus! These vectors, however, are not all normalized (they're norm is not equal to 1) since we have:

$$g_{11}=\vec{e}_1\circ\vec{e}_1=1\quad g_{22}=\vec{e}_2\circ\vec{e}_2=r^2\quad g_{33}=\vec{e}_3\circ\vec{e}_3=r^2\sin^2(\theta)\tag{6.7.322}$$

The natural basis in spherical coordinates is thus formed by vectors that are variable in direction and in modulus at each point of M. The quantities g_{ij} constitute an example of a metric tensor attached to each of the points M of the space $\mathcal{E}^3$.

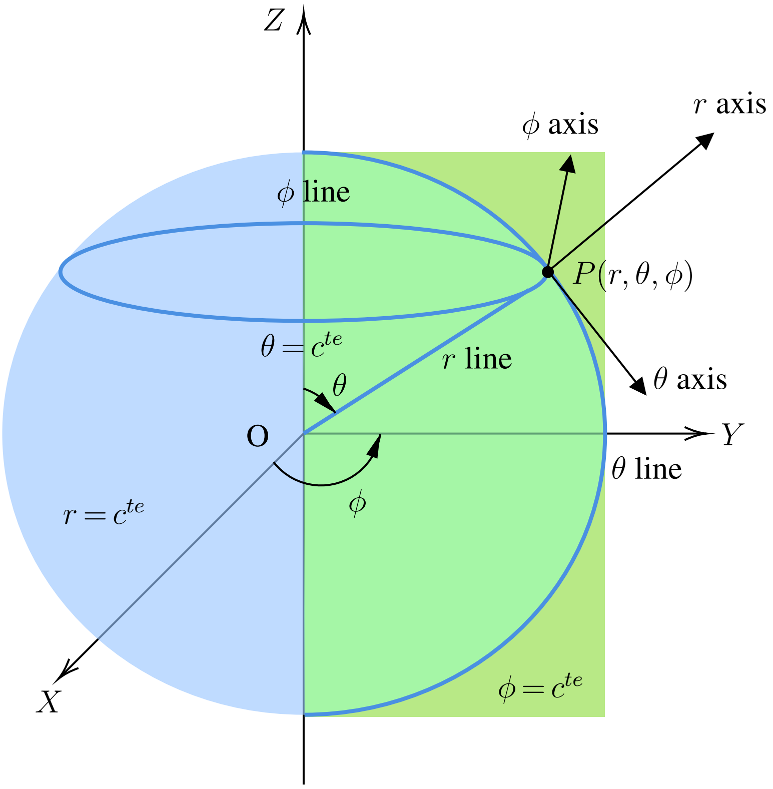


Figure 6.160: Coordinate surfaces, coordinate lines, and coordinate axes of spherical coordinates

The linear element of the surface is then given by (the details of the calculations can be found in the section of General Relativity):

$$\boxed{\mathrm{d}s^2 = \mathrm{d}r^2 + r^2\mathrm{d}\theta^2 + r^2\sin^2(\theta)\mathrm{d}\phi^2} \tag{6.7.323}$$

6.7.9.2 Natural basis in polar coordinates (curvilinear basis in polar coordinates)

Let us determine the natural basis of the vector space E^2 associated with the point space $\mathcal{E}^2$ of ordinary geometry, in polar coordinates. Let us write the expression of the vectors $\overrightarrow{\mathrm{O}M}$ in a fixed cartesian basis $(\vec{e}_i^{\,0})$ which is by definition (see the section Vector Calculus for more details):

$$\overrightarrow{OM} = x^i\vec{e}_i^{\,0} = r\cos(\phi)\vec{e}_1^{\,0} + r\sin(\phi)\vec{e}_2^{\,0} \tag{6.7.324}$$

The vectors of the natural basis are given by:

$$\vec{e}_k = \partial_k\vec{M} \Rightarrow \begin{cases} \vec{e}_1 = \partial_1\vec{M} = \partial_r\vec{M} \\ \vec{e}_2 = \partial_2\vec{M} = \partial_\phi\vec{M} \end{cases} \tag{6.7.325}$$

We have:

$$\vec{e}_1 = \partial_1\vec{M} = \cos(\phi)\vec{e}_1^{\,0} + \sin(\phi)\vec{e}_2^{\,0} \tag{6.7.326}$$

The derivative of $\overrightarrow{\mathrm{O}M}$ with respect to ϕ gives the vector $\vec{e}_2$:

$$\vec{e}_2 = \partial_2\vec{M} = -r\sin(\phi)\vec{e}_1^{\,0} + r\cos(\phi)\vec{e}_2^{\,0} \tag{6.7.327}$$

These two vectors are orthogonal to each other as we can easily verify by performing the dot products $\vec{e}_1 \circ \vec{e}_2$. We thus find the same result as in the section of Vector Calculus.

We then have:

$$g_{11} = \cos^2(\phi) + \sin^2(\phi) \quad g_{21} = g_{12} = 0 \quad g_{22} = r^2 \tag{6.7.328}$$

The linear element of the plane is then given by (the details of the calculations can be found in the section of General Relativity):

$$\boxed{\mathrm{d}s^2 = \mathrm{d}r^2 + r^2\mathrm{d}\phi^2} \tag{6.7.329}$$

6.7.9.3 Natural basis in cylindrical coordinates (cylindrical basis in polar coordinates)

Let us determine the natural basis of the vector space E^3 associated with the point space $\mathcal{E}^3$ of ordinary geometry, in cylindrical coordinates. Let us write the expression of the vectors $\overrightarrow{\mathrm{O}M}$ in a fixed cartesian basis $(\vec{e}_i^{\,0})$ which is by definition (see the section Vector Calculus for more details):

$$\overrightarrow{OM} = x^i\vec{e}_i^{\,0} = r\cos(\varphi)\vec{e}_1^{\,0} + r\sin(\varphi)\vec{e}_2^{\,0} + z\vec{e}_3^{\,0} \tag{6.7.330}$$

The vectors of the natural basis are given by:

$$\vec{e}_k=\partial_k\vec{M}\Rightarrow\begin{cases}\vec{e}_1=\partial_1\vec{M}=\partial_r\vec{M}\\ \vec{e}_2=\partial_2\vec{M}=\partial_\varphi\vec{M}\\ \vec{e}_3=\partial_3\vec{M}=\partial_z\vec{M}\end{cases}\qquad(6.7.331)$$

We have:

$$\vec{e}_1=\partial_1\vec{M}=\cos(\varphi)\vec{e}_1^{\,0}+\sin(\varphi)\vec{e}_2^{\,0}\qquad(6.7.332)$$

The derivative of $\overrightarrow{\mathrm{O}M}$ with respect to φ gives the vector $\vec{e}_2$:

$$\vec{e}_2=\partial_2\vec{M}=-r\sin(\phi)\vec{e}_1^{\,0}+r\cos(\phi)\vec{e}_2^{\,0}\qquad(6.7.333)$$

and finally:

$$\vec{e}_3=\partial_3\vec{M}=\vec{e}_3^{\,0}\qquad(6.7.334)$$

These three vectors are orthogonal to each other as we can easily verify by performing the dot products $\vec{e}_i\circ\vec{e}_j$. We thus find the same result as in the section of Vector Calculus.

We have:

$$g_{11}=1\quad g_{22}=r^2\quad g_{33}=1\qquad(6.7.335)$$

The linear element of the surface is then given by (the details of the calculations can be found in the section of General Relativity):

$$\boxed{\mathrm{d}s^2=\mathrm{d}r^2+r^2\mathrm{d}\phi^2+\mathrm{d}z^2}\qquad(6.7.336)$$

Notice now that:

$$(\mathrm{d}s)^2=g_{\mu\nu}\mathrm{d}x^\mu\mathrm{d}x^\nu\qquad(6.7.337)$$

expresses something about the intrinsic metrical relations of the space, but it does so in terms of a specific coordinate system. If we considered the metrical relations at the same point in terms of a different system of coordinates (such as changing from Cartesian to polar coordinates), the coefficients $g_{\mu\nu}$ would be different. Fortunately there is a simple way of converting the $g_{\mu\nu}$ from one system of coordinates to another, based on the fact that they describe a purely localistic relation among differential quantities. Suppose we are given the metrical coefficients $g_{\mu\nu}$ for the coordinates x^α, and we are also given another system of coordinates y^α that are defined in terms of the x^α by some arbitrary continuous functions:

$$\begin{matrix}y^0=f_0(x^0,x^1,x^2,x^3) & y^1=f_1(x^0,x^1,x^2,x^3)\\ y^2=f_2(x^0,x^1,x^2,x^3) & y^3=f_3(x^0,x^1,x^2,x^3)\end{matrix}\qquad(6.7.338)$$

Assuming the Jacobian of this transformation isn't zero, we know that it's invertible, and so we can just as well express the original coordinates as continuous functions (at this point) of the new coordinates:

$$\begin{matrix}x^0=F_0(y^0,y^1,y^2,y^3) & x^1=F_1(y^0,y^1,y^2,y^3)\\ x^2=F_2(y^0,y^1,y^2,y^3) & x^3=F_3(y^0,y^1,y^2,y^3)\end{matrix}\qquad(6.7.339)$$

Now we can evaluate the total derivatives of the original coordinates in terms of the new coordinates. For example, $\mathrm{d}x^0$ can be written as:

$$\mathrm{d}x^0=\frac{\partial x^0}{\partial y^0}\mathrm{d}y^0+\frac{\partial x^0}{\partial y^1}\mathrm{d}y^1+\frac{\partial x^0}{\partial y^2}\mathrm{d}y^2+\frac{\partial x^0}{\partial y^3}\mathrm{d}y^3 \tag{6.7.340}$$

and similarly for the $\mathrm{d}x^1, \mathrm{d}x^2$, and $\mathrm{d}x^3$. The product of any two of these differentials, $\mathrm{d}x^\mu$ and $\mathrm{d}x^\nu$, is of the form:

$$\mathrm{d}x^\mu \mathrm{d}x^\nu=\frac{\partial x^\mu}{\partial y^\alpha}\frac{\partial x^\nu}{\partial y^\beta}\mathrm{d}y^\alpha \mathrm{d}y^\beta \tag{6.7.341}$$

(remembering the summation convention giving 16 components). Substituting these expressions for the products of x differentials in the metric formula:

$$(\mathrm{d}s)^2=g_{\mu\nu}\mathrm{d}x^\mu \mathrm{d}x^\nu \tag{6.7.342}$$

gives:

$$(\mathrm{d}s)^2=g_{\mu\nu}\frac{\partial x^\mu}{\partial y^\alpha}\frac{\partial x^\nu}{\partial y^\beta}\mathrm{d}y^\alpha \mathrm{d}y^\beta \tag{6.7.343}$$

The first three factors on the right hand side obviously represent the coefficient of $\mathrm{d}y^\alpha \mathrm{d}y^\beta$ in the metric formula with respect to the y coordinates, so we've shown that the array of metric coefficients transforms from the x to the y coordinate system according to the relation:

$$\boxed{\left(g_{\alpha\beta}\right)_y=\frac{\partial x^\mu}{\partial y^\alpha}\frac{\partial x^\nu}{\partial y^\beta}\left(g_{\mu\nu}\right)_x} \tag{6.7.344}$$

Also often denoted:

$$g_{\alpha\beta}^{(y)}=\frac{\partial x^\mu}{\partial y^\alpha}\frac{\partial x^\nu}{\partial y^\beta}g_{\mu\nu}^{(x)} \tag{6.7.345}$$

and more rarely (sadly) denoted:

$$\boxed{\left(g_{\alpha\beta}\right)_y=\Lambda_\alpha^\mu\Lambda_\beta^\nu\left(g_{\mu\nu}\right)_x} \tag{6.7.346}$$

Notice that each component of the new metric array is a linear combination of the old metric components, and the coefficients are the partials of the old coordinates with respect to the new. Arrays that transform to each other in the way:

$$A_{\alpha\beta}=\frac{\partial x^\mu}{\partial y^\alpha}\frac{\partial x^\nu}{\partial y^\beta}A_{\mu\nu} \tag{6.7.347}$$

are named "(second order) covariant[33] tensors".

When we speak of an array being transformed from one system of coordinates to another, it's clear that the array must have a definite meaning independent of the system of coordinates. We could, for example, have an array of scalar quantities, whose values are the same at a given point, regardless of the coordinate system. However, the components of the array might still be required to change for different systems.

---

[33] Keep in mind that when a tensor expressed in terms of Cartesian coordinates, there is no distinction between covariant and contravariant components

Now we define arrays that transform to each other in the way:

$$A^{\mu\nu}=\frac{\partial x^{\mu}}{\partial y^{\alpha}}\frac{\partial x^{\nu}}{\partial y^{\beta}}A_{\alpha\beta} \tag{6.7.348}$$

as "(second order) contravariant tensors".

We will need this important result for our study of the Stress-Energy tensor in the section of General Relativity.

Now the reader must keep in mind that in General Relativity, we don't equip space-time (for a naive and technical definition of "space-time" see the section of Special Relativity at page 5708) itself with a vector space structure, which means that the "four-position" is not a four-vector! As a result, the object x_{μ} whose components are given by $x_{\mu}=g_{\mu\nu}x^{\nu}$ as proved earlier is not a covector! One can show this by its transformation properties.

If we change the coordinate systems from x to y, then:

$$\begin{gathered}g_{\mu\nu}^{(x)}\rightarrow g_{\mu\nu}^{(y)}=\frac{\partial x^{\alpha}}{\partial y^{\mu}}\frac{\partial x^{\beta}}{\partial y^{\nu}}g_{\alpha\beta}^{(x)}\\ x^{\nu}\rightarrow y^{\nu}\end{gathered} \tag{6.7.349}$$

and so:

$$x_{\mu}=g_{\mu\nu}^{(x)}x^{\nu}\rightarrow g_{\mu\nu}^{(y)}y^{\nu}=\frac{\partial x^{\alpha}}{\partial y^{\mu}}\frac{\partial x^{\beta}}{\partial y^{\nu}}g_{\alpha\beta}^{(x)}y^{\nu}=\frac{\partial x^{\alpha}}{\partial y^{\mu}}g_{\alpha\beta}^{(x)}\left(\frac{\partial x^{\beta}}{\partial y^{\nu}}y^{\nu}\right) \tag{6.7.350}$$

If this object was a covector, it would transform as proven earlier using the covariant transformation:

$$x_{\mu}\rightarrow y_{\mu}=\frac{\partial x^{\alpha}}{\partial y^{\mu}}x_{\alpha}=\frac{\partial x^{\alpha}}{\partial y^{\mu}}g_{\alpha\beta}^{(x)}x^{\beta} \tag{6.7.351}$$

Comparing this to what we found above, this is only true if:

$$\frac{\partial x^{\beta}}{\partial y^{\nu}}y^{\nu}=x^{\beta} \tag{6.7.352}$$

i.e. if the coordinate transformation is linear. Since coordinate transformations are, in general, not linear, then this equality does not hold, and x_{ν} are not the components of a covector.

Finally another obvious identity that you can also sometimes found in some rare textbooks is (I never see a practical application of that latter):

$$\frac{\partial x^{\lambda}}{\partial x^{\mu}}\frac{\partial x^{\mu}}{\partial x^{\nu}}=\delta_{\mu}^{\lambda}\delta_{\nu}^{\mu}=\delta_{\nu}^{\lambda}=g_{\nu}^{\lambda} \tag{6.7.353}$$

6.7.10 Christoffel symbols

The study of tensor fields constitutes, for the physicist, the essential element of the tensorial analysis. The generic tensor $\vec{U}$ of this field is a function of the point M and we will denote it simply by:

$$\vec{U}(M) \tag{6.7.354}$$

If the tensor $\vec{U}$ is a function only of M, the field considered is named a "fixed field". If $\vec{U}$ is moreover a function of one or more equation parameters other than the coordinates of M, then we say that this is a "variable field" and we note it:

$$\vec{U}(M,\alpha) \tag{6.7.355}$$

The different algebraic operations on the tensors $\vec{U}(M)$ associated with a same point M do not generates any particular difficulty. The derivative of $\vec{U}(M)$ with respect to a parameter α leads to the use of the classical results relating to the derivation of the vectors.

However, a difficulty appears when we try to calculate the derivative of a tensor $\vec{U}(M)$ with respect to the curvilinear coordinates. Indeed, the components of the tensor are defined at each point M with respect to a natural coordinate system which varies from one point to another.

Consequently, the calculation of the elementary variation, named "elementary transport":

$$\vec{U}(M')-\vec{U}(M) \tag{6.7.356}$$

When we pass from a point M to an infinitely neighbouring point M' this can be done in physics only if we use to the same basis. In order to compare the tensors $\vec{U}(M')$ and $\vec{U}(M)$ each other, we are led to study how a natural coordinate system for a given coordinate system varies when we pass from a point M to an infinitely close point M'.

For a system of curvilinear coordinates u^i given a punctual space $\mathcal{E}^n$, a fundamental problem of tensor analysis consists in determining, with respect to the natural reference point $(M,\vec{e}_k)$ at the point M, the natural reference point $(M',\vec{e}_k')$ at the infinitely close point M'. We then say that we are looking for an "affine connection".

On the one hand, the point M' will be perfectly defined with respect to M if we determine the vector $\mathrm{d}\vec{M}$ such as $\overrightarrow{MM'}=\mathrm{d}\vec{M}$. For curvilinear coordinates u^k, the decomposition of an elementary vector $\mathrm{d}\vec{M}$ is given by the relation that we have previously proved:

$$\mathrm{d}\vec{M}=\partial_k\vec{M}\mathrm{d}u^k=\vec{e}_k\mathrm{d}u^k \tag{6.7.357}$$

the quantities $\mathrm{d}u^k$ being for recall the contravariant components of the vector $\mathrm{d}\vec{M}$ on the natural basis $(\vec{e}_k)$.

And now to make things physically comparable, we must guarantee that vectors of the both basis are also parallel! So the idea is that the derivative we are looking for allows one to transport vectors of a manifold (surface) along curves so that they stay parallel with respect to the connection (derivative). Such an idea is named in physics "parallel transport".

Therefore, the idea is to determine the vector $\vec{e}'_k$ with elementary variations $\mathrm{d}\vec{e}_k$ of the vector $\vec{e}_k$ relatively to the natural reference frame $(M,\vec{e}_k)$, when we go from M to M'. We then have:

$$\mathrm{d}\vec{e}'_k=\vec{e}_k+\mathrm{d}\vec{e}_k \tag{6.7.358}$$

The computation of the vectors $\mathrm{d}\vec{e}_k$ then remains the main problem to solve. We will first consider an example of this type of calculation in spherical coordinates for pedagogical purposes as it can help to understand what will follows.

For this, let us return the expression of the vectors $\vec{e}_k$ of the natural base in spherical coordinates, that is:

$$\begin{aligned}\vec{e}_1&=\partial_1\vec{M}=\sin(\theta)\cos(\phi)\vec{\imath}+\sin(\theta)\sin(\phi)\vec{\jmath}+\cos(\theta)\vec{k}\\ \vec{e}_2&=\partial_2\vec{M}=r\cos(\theta)\cos(\phi)\vec{\imath}+r\cos(\theta)\sin(\phi)\vec{\jmath}-r\sin(\theta)\vec{k}\\ \vec{e}_3&=\partial_3\vec{M}=-r\sin(\theta)\sin(\phi)\vec{\imath}+r\sin(\theta)\cos(\phi)\vec{\jmath}\end{aligned} \tag{6.7.359}$$

Since the basis vectors $\vec{\imath},\vec{\jmath},\vec{k}$ of the fixed Cartesian reference being constant in modulus and in direction, the differential of the vector $\vec{e}_1$ is written:

$$\mathrm{d}\vec{e}_1=(\cos(\theta)\cos(\phi)\vec{\imath}+\cos(\theta)\sin(\phi)\vec{\jmath}-\sin(\theta)\vec{k})\mathrm{d}\theta+(-\sin(\theta)\sin(\phi)+\sin(\theta)\cos(\phi)\vec{\jmath})\mathrm{d}\phi \tag{6.7.360}$$

We notice that the terms in parentheses represent respectively the vectors $\vec{e}_2/r$ and $\vec{e}_3/r$, hence:

$$\mathrm{d}\vec{e}_1=\left(\frac{\mathrm{d}\theta}{r}\right)\vec{e}_2+\left(\frac{\mathrm{d}\phi}{r}\right)\vec{e}_3 \tag{6.7.361}$$

We also compute, by differentiating the vector $\vec{e}_2$:

$$\begin{aligned}\mathrm{d}\vec{e}_2=&\,\mathrm{d}r\cos(\theta)\cos(\phi)\vec{\imath}+\mathrm{d}r\cos(\theta)\sin(\phi)\vec{\jmath}-\mathrm{d}r\sin(\theta)\vec{k}\\ &-r\sin(\theta)\mathrm{d}\theta\cos(\phi)\vec{\imath}-r\sin(\theta)\mathrm{d}\theta\sin(\phi)\vec{\jmath}-r\cos(\theta)\mathrm{d}\theta\vec{k}\\ &-r\cos(\theta)\sin(\phi)\mathrm{d}\phi\vec{\imath}+r\cos(\theta)\cos(\phi)\mathrm{d}\phi\vec{\jmath}\\ &=(\cos(\theta)\cos(\phi)\vec{\imath}+\cos(\theta)\sin(\phi)\vec{\jmath}-\sin(\theta)\vec{k})\mathrm{d}r+(-r\sin(\theta)\cos(\phi)\vec{\imath}-r\sin(\theta)\sin(\phi)\vec{\jmath}-r\cos(\theta)\vec{k})\mathrm{d}\theta\\ &+(-r\cos(\theta)\sin(\phi)\vec{\imath}+r\cos(\theta)\cos(\phi)\vec{\jmath})\mathrm{d}\phi\\ &=\frac{\vec{e}_2}{r}\mathrm{d}r-r\vec{e}_1\mathrm{d}\theta+(-r\cos(\theta)\sin(\phi)\vec{\imath}+r\cos(\theta)\cos(\phi)\vec{\jmath})\mathrm{d}\phi\end{aligned} \tag{6.7.362}$$

with:

$$\cot(\theta)=\frac{\cos(\theta)}{\sin(\theta)} \tag{6.7.363}$$

we have:

$$\vec{e}_3\cot(\theta)=\vec{e}_3\frac{\cos(\theta)}{\sin(\theta)}=-r\cos(\theta)\sin(\phi)\vec{\imath}+r\cos(\theta)\cos(\phi)\vec{\jmath} \tag{6.7.364}$$

So finally:

$$\mathrm{d}\vec{e}_2=\frac{\vec{e}_2}{r}\mathrm{d}r-r\vec{e}_1\mathrm{d}\theta+\vec{e}_3\cot(\theta)\mathrm{d}\phi \tag{6.7.365}$$

And:

$$\begin{aligned}\mathrm{d}\vec{e}_3=&-\mathrm{d}r\sin(\theta)\sin(\phi)\vec{i}+\mathrm{d}r\sin(\theta)\cos(\phi)\vec{j}\\&-r\cos(\theta)\mathrm{d}\theta\sin(\phi)\vec{i}+r\cos(\theta)\mathrm{d}\theta\cos(\phi)\vec{j}\\&-r\sin(\theta)\cos(\phi)\mathrm{d}\phi\vec{i}-r\sin(\theta)\sin(\phi)\mathrm{d}\phi\vec{j}\end{aligned} \quad (6.7.366)$$

After a few elementary and very tricky algebraic operations (...), we get:

$$\mathrm{d}\vec{e}_3=(-r\sin^2(\theta)\mathrm{d}\phi)\vec{e}_1+(-\sin(\theta)\cos(\theta)\mathrm{d}\phi)\vec{e}_2+\left(\frac{\mathrm{d}r}{r}+\cot(\theta)\mathrm{d}\theta\right)\vec{e}_3 \quad (6.7.367)$$

The differentials $\mathrm{d}\vec{e}_k$ are thus decomposed on the natural basis $(\vec{e}_k)$. If we denote by ω_i^k the contravariant components of the vector $\mathrm{d}\vec{e}_1$, that latter is written:

$$\boxed{\mathrm{d}\vec{e}_i=\omega_i^k\vec{e}_k} \quad (6.7.368)$$

The components ω_i^k of the vector $\mathrm{d}\vec{e}_i$ are differential forms (linear combinations of differentials). We have, for example:

$$\omega_1^2=\frac{\mathrm{d}\theta}{r} \qquad \omega_3^3=\frac{\mathrm{d}r}{r}+\cot(\theta)\mathrm{d}\theta \quad (6.7.369)$$

If we denote in a general way by u^i the spherical coordinates, we have:

$$u^1=r \qquad u^2=\theta \qquad u^3=\phi \quad (6.7.370)$$

The coordinate differentials are then denoted:

$$\mathrm{d}u^1=\mathrm{d}r \qquad \mathrm{d}u^2=\mathrm{d}\theta \qquad \mathrm{d}u^3=\mathrm{d}\phi \quad (6.7.371)$$

and the components ω_i^j are then written in a general way:

$$\boxed{\omega_i^j=\Gamma_{ki}^j\mathrm{d}u^k} \quad (6.7.372)$$

Where the quantities Γ_{ki}^j are functions of r,θ,ϕ that will be explicitly obtained by identifying each component ω_i^j. For example, the component ω_3^3 is written with the notation of the previous relation:

$$\omega_3^3=\frac{\mathrm{d}r}{r}+\cot(\theta)\mathrm{d}\theta=\Gamma_{13}^3\mathrm{d}u^1+\Gamma_{23}^3\mathrm{d}u^2+\Gamma_{33}^3\mathrm{d}u^3 \quad (6.7.373)$$

Identifying the coefficients of the differentials, it comes:

$$\Gamma_{13}^3=\frac{1}{r} \qquad \Gamma_{23}^3=\cot(\theta) \qquad \Gamma_{33}^3=0 \quad (6.7.374)$$

By doing the same with the 9 components ω_i^j, we get the 27 (...) terms for which the detailed calculations for the $3\cdot 9=27$ are given much further below in the text. For any curvilinear coordinate system, these quantities Γ_{ki}^j are named "Christoffel symbols of the second kind" or "Euclidean functions of affine connection".

Thus, for a punctual space $\mathcal{E}^3$ and a system of any curvilinear coordinates u^j, the differential $\vec{e}_i=\omega_i^k\vec{e}_k$ of the vectors $\vec{e}_i$ of the natural basis is written on this basis:

$$\mathrm{d}\vec{e}_k=\omega_i^k\vec{e}_k=\Gamma_{ji}^k\mathrm{d}u^j\vec{e}_k \quad (6.7.375)$$

where Γ_{ji}^k in this context is often named "Levi-Civita connection".

Remark

Therefore this also means that given a vector $\vec{v}$ such that:

$$\vec{v}=v^i\vec{e}_i \tag{6.7.376}$$

Thus:

$$\mathrm{d}\vec{v}=\mathrm{d}v^k\vec{e}_k+v^k\mathrm{d}\vec{e}_k \tag{6.7.377}$$

But as:

$$\mathrm{d}\vec{e}_k=\Gamma_{ji}^k\mathrm{d}u^j\vec{e}_k \tag{6.7.378}$$

we have then:

$$\boxed{\mathrm{d}\vec{v}=\left(\mathrm{d}v^k+\Gamma_{ji}^k\mathrm{d}u^j\right)\vec{e}_k} \tag{6.7.379}$$

We have just seen, on the example of the spherical coordinates, that a direct calculation makes it possible, by identification, to obtain explicitly the quantities Γ_{ki}^j. We shall see that we can also obtain the expression of these quantities as a function of the components of g_{ij}.

The calculation of the quantities Γ_{ki}^j as a function of the g_{ij} will lead us to introduce other Christoffel symbols. For this, let us write the covariant components, denoted ω_{ji}, of the differentials $\mathrm{d}\vec{e}_i$, thus (it's kind a definition):

$$\omega_{ji}=\vec{e}_j\circ d\vec{e}_i \tag{6.7.380}$$

✍**Example:**

With our example above:

$$\vec{e}_3=\partial_3\vec{M}=-r\sin(\theta)\sin(\phi)\vec{i}+r\sin(\theta)\cos(\phi)\vec{j} \tag{6.7.381}$$

and:

$$\mathrm{d}\vec{e}_3=(-r\sin^2(\theta)\mathrm{d}\phi)\vec{e}_1+(-\sin(\theta)\cos(\theta)\mathrm{d}\phi)\vec{e}_2+\left(\frac{\mathrm{d}r}{r}+\cot(\theta)\mathrm{d}\theta\right)\vec{e}_3 \tag{6.7.382}$$

We get (most dot products are orthogonal vectors, therefore they vanish):

$$\begin{aligned}\omega_{33}=\vec{e}_3\circ d\vec{e}_3&=\vec{e}_3\circ\left((-r\sin^2(\theta)\mathrm{d}\phi)\vec{e}_1+(-\sin(\theta)\cos(\theta)\mathrm{d}\phi)\vec{e}_2+\left(\frac{\mathrm{d}r}{r}+\cot(\theta)\mathrm{d}\theta\right)\vec{e}_3\right)\\&=\left(\frac{\mathrm{d}r}{r}+\cot(\theta)\mathrm{d}\theta\right)\vec{e}_3\circ\vec{e}_3=\left(\frac{\mathrm{d}r}{r}+\cot(\theta)\mathrm{d}\theta\right)r^2\sin^2(\theta)=r\sin^2(\theta)\mathrm{d}r+r^2\cot(\theta)\mathrm{d}\theta\end{aligned} \tag{6.7.383}$$

The covariant components are also a linear combinations of the differentials $\mathrm{d}u^k$ that we can write in the form:

$$\omega_{ji}=\Gamma_{kji}du^k \tag{6.7.384}$$

The quantities Γ_{kji} are named the "Christoffel symbols of the first kind".

We see quite clearly by going through the definitions and examples of the Christoffel symbols above again that:

1. For the Christoffel symbols of the second kind, they are symmetrical with respect to their lower indices and therefore if the metric is symmetric, we have:

$$\Gamma_{ji}^{k}=\Gamma_{ij}^{k} \tag{6.7.385}$$

2. For the Christoffel symbols of the first kind, they are also symmetrical with respect to their extremal indices and therefore if the metric is symmetric, we have:

$$\Gamma_{kij}=\Gamma_{jik} \tag{6.7.386}$$

Indeed (following the request of a reader), since we have:

$$\mathrm{d}\vec{e}_k=\omega_i^k\vec{e}_k=\Gamma_{ji}^k\mathrm{d}u^j\vec{e}_k \tag{6.7.387}$$

it then comes:

$$\frac{\partial\vec{e}_k}{\partial u^j}=\vec{e}_k\Gamma_{ji}^k \tag{6.7.388}$$

and by swapping the indices i and j:

$$\frac{\partial\vec{e}_k}{\partial u^j}=\vec{e}_k\Gamma_{ij}^k \tag{6.7.389}$$

The term-to-term identification of the development on a concrete case of the last two relations will give (necessarily) the equality:

$$\Gamma_{ij}^k=\Gamma_{ji}^k \tag{6.7.390}$$

that we wanted to prove.

Since the covariant components are related to the contravariant components by the relations (contraction of indices):

$$\omega_{ji}=g_{jl}\omega_i^l=g_{jl}\Gamma_{ki}\mathrm{d}u^k \tag{6.7.391}$$

We get the expression linking the Christoffel symbols of each kind:

$$\boxed{\Gamma_{kji}=g_{jl}\Gamma_{ki}^l} \tag{6.7.392}$$

Conversely:

$$\boxed{\Gamma_{ki}^j=g^{jl}\Gamma_{kli}} \tag{6.7.393}$$

Remark

Various notations are used to represent the Christoffel symbols. The most common are:

- Christoffel symbol of the first kind:

$$\Gamma_{kji}=[ki,j] \tag{6.7.394}$$

- Christoffel symbol of the second kind:

$$\Gamma^{j}_{ki}=[k^j\, i] \tag{6.7.395}$$

Let us consider now a punctual space $\mathcal{E}^n$ and given a linear element $\mathrm{d}s^2$ of this space:

$$\mathrm{d}s^2=g_{ij}\mathrm{d}u^i\mathrm{d}u^j \tag{6.7.396}$$

Starting from:

$$g_{ij}=\vec{e}_i\circ\vec{e}_j \tag{6.7.397}$$

we get by differentiation:

$$\mathrm{d}g_{ij}=\vec{e}_i\circ\mathrm{d}\vec{e}_j+\vec{e}_j\circ\mathrm{d}\vec{e}_i \tag{6.7.398}$$

By injecting in it the expression of the differential $\mathrm{d}\vec{e}_j=\omega^l_j\vec{e}_l$ this gives us:

$$\mathrm{d}g_{ij}=(\vec{e}_i\circ\vec{e}_l)\omega^l_j+(\vec{e}_j\circ\vec{e}_l)\omega^l_i=g_{il}\omega^l_j+g_{jl}\omega^l_i \tag{6.7.399}$$

where the term represents ω^l_j represents the mixed component of the vector $\mathrm{d}\vec{e}_j$. We can make this component covariant by multiplying it by the metric tensor g_{il} so as to form an ω_{ij} quantity which can in turn be expressed by means of the Christoffel symbols as follows as we already know:

$$\omega_{ij}=g_{il}\omega^l_j=g_{il}\Gamma^l_{kj}\mathrm{d}u^k \tag{6.7.400}$$

Substituting the relation $\Gamma_{kji}=g_{jl}\Gamma^l_{ki}$ in the preceding expression (the indices used in this relation are not those of the expression in question, but mutatis mutandis it is equivalent), we then get:

$$\omega_{ij}=\Gamma_{kij}\mathrm{d}u^k \tag{6.7.401}$$

The differential $\mathrm{d}g_{ij}$ is then written:

$$\boxed{\mathrm{d}g_{ij}=\omega_{ij}+\omega_{ji}=\left(\Gamma_{kij}+\Gamma_{kji}\right)\mathrm{d}u^k} \tag{6.7.402}$$

On the other hand, the differential of the function g_{ij} is then written:

$$\boxed{\mathrm{d}g_{ij}=\frac{\partial g_{ij}}{\partial u^k}\mathrm{d}u^k=\partial_k g_{ij}\mathrm{d}u^k} \tag{6.7.403}$$

hence by identifying the coefficients of the differentials $\mathrm{d}u^k$ in these two last expressions (much further below in this section there is a detailed example of all the relations which will follow with several coordinate systems):

$$\Gamma_{kij}+\Gamma_{kji}=\partial_k g_{ij} \tag{6.7.404}$$

Relation that the reader can (if he doubt about its veracity) check with the detailed practical examples which are given much further below.

As we have (in the case of a symmetric metric):

$$\Gamma_{kij}=\Gamma_{kji} \tag{6.7.405}$$

where it is strongly recommended to the reader to remember that the permutation of the indices respecting this last relation generally only works on the extreme indices!

We can therefore write the prior-previous relation as:

$$\Gamma_{jik}+\Gamma_{kji}=\partial_k g_{ij} \tag{6.7.406}$$

and by performing a circular permutation on the indices (hence it is not a permutation of the extreme indices!), we get:

$$\Gamma_{kji}+\Gamma_{ikj}=\partial_i g_{jk}\Gamma_{ikj}+\Gamma_{jik}=\partial_j g_{kj} \tag{6.7.407}$$

By making the sum:

$$\begin{array}{cr} & \Gamma_{jik}+\Gamma_{kji}=\partial_k g_{ij}\\ & \Gamma_{kji}+\Gamma_{ikj}=\partial_i g_{jk}\\ \hline + & \Gamma_{jik}+\Gamma_{kji}+\Gamma_{kji}+\Gamma_{ikj}=\partial_k g_{ij}+\partial_i g_{jk} \end{array} \tag{6.7.408}$$

and by subtracting:

$$\begin{array}{cr} & \Gamma_{jik}+\Gamma_{kji}+\Gamma_{kji}+\Gamma_{ikj}=\partial_k g_{ij}+\partial_i g_{jk}\\ & \Gamma_{ikj}+\Gamma_{jik}=\partial_j g_{kj}\\ \hline - & \Gamma_{jik}+\Gamma_{kji}+\Gamma_{kji}+\Gamma_{ikj}-\Gamma_{ikj}-\Gamma_{jik}=\partial_k g_{ij}+\partial_i g_{jk}-\partial_j g_{ki} \end{array} \tag{6.7.409}$$

Simplifying it comes:

$$2\Gamma_{kji}=\partial_k g_{ij}+\partial_i g_{jk}-\partial_j g_{ki} \tag{6.7.410}$$

therefore:

$$\boxed{\Gamma_{kji}=\frac{1}{2}(\partial_k g_{ij}+\partial_i g_{jk}-\partial_j g_{ki})} \tag{6.7.411}$$

It is the expression of the Christoffel symbols of the first kind as a function of the partial derivatives of the components g_{ij} of the fundamental tensor named "first Christoffel identity". We thus understand why in a locally inertial frame (of the Minkowski type) the Christoffel symbols are all zero (given that the metric is constant).

We get those of the second kind from the following relation (by definition) named "fundamental theorem of Riemannian geometry" or "Levi-Civita connection" or "first Christoffel identity":

$$\boxed{\Gamma_{kj}^{i}=g^{il}\Gamma_{klj}=\frac{1}{2}g^{il}(\partial_k g_{lj}+\partial_j g_{kl}-\partial_l g_{kj})} \tag{6.7.412}$$

The last two expressions listed above allow the actual calculation of the Christoffel symbols for a given metric (hence an enormous gain in calculations). When the quantities g_{ij} are given a priori, we can study the properties of the punctual space defined by the data of this metric, which is the case of the Riemann spaces we will see later.

For our study of Cosmology we will need to calculate:

$$\Gamma_{ki}^{i}=g^{im}\Gamma_{mki}=\frac{1}{2}g^{im}(\partial_i g_{mk}+\partial_k g_{mi}-\partial_m g_{ki}) \tag{6.7.413}$$

For this purpose, notice that:

$$\begin{aligned} g^{im}\partial_i g_{mk} &= g^{im}\partial_i g_{km} && \text{(symmetry of } g\text{)}\\ &= g^{mi}\partial_m g_{ki} && \text{(relabeling dummy indices)}\\ &= g^{im}\partial_m g_{ki} && \text{(symmetry of } g\text{)} \end{aligned}$$

Therefore:

$$\Gamma_{ki}^{i}=\frac{1}{2}g^{im}(\partial_i g_{mk}+\partial_k g_{mi}-\partial_m g_{ki})=\frac{1}{2}g^{im}\partial_k g_{mi} \tag{6.7.414}$$

✍**Example:**

We propose to calculate the Christoffel symbols of the second kind Γ_{kj}^{i} corresponding to the polar coordinate system in the plane (it will be sufficiently long...) that we will write this time (at the opposite of the section of Vector Calculus) in index notation:

$$\begin{array}{ccc} x^1=y^1\cos\left(y^2\right) & & y^1=\sqrt{\left(x^1\right)^2+\left(x^2\right)^2} \\ & \text{with} & \\ x^2=y^1\sin\left(y^2\right) & & y^2=\arctan\left(\dfrac{x^2}{x^1}\right) \end{array} \tag{6.7.415}$$

We will calculate the Christoffel symbols of the second kind from our last relation:

$$\Gamma_{kj}^{i}=g^{il}\Gamma_{klj}=\frac{1}{2}g^{il}(\partial_j g_{kl}+\partial_k g_{lj}-\partial_l g_{kj}) \tag{6.7.416}$$

Let us determine the components of the metric. By the way, they are the same as those we had calculated for the cylindrical coordinates above with the obvious difference that g_{33} does not exist. Therefore, we have:

$$g_{11}=1 \qquad g_{12}=g_{21}=0 \qquad g_{22}=\left(y^1\right)^2 \tag{6.7.417}$$

Let us then calculate the g_{ji}. In this example it is rather trivial, it is enough to apply the relation demonstrated at the beginning of this section:

$$g_{ik}g^{kj}=\delta_i^j \tag{6.7.418}$$

Or by dealing as with standard matrices (see section Linear Algebra page 1981):

$$g^{ji}=(g_{ik})^{-1}=\frac{1}{g_{11}g_{22}-g_{12}g_{21}}\begin{bmatrix} g_{22} & -g_{12}\\ -g_{21} & g_{11}\end{bmatrix} \tag{6.7.419}$$

We then have immediately:

$$g^{11}=1 \qquad g^{12}=g^{21}=0 \qquad g^{22}=\frac{1}{\left(y^1\right)^2} \tag{6.7.420}$$

Now let us develop the Christoffel symbols writing for these coordinates:

$$\begin{aligned}
\Gamma_{11}^1&=\frac{1}{2}g^{1l}\left(\frac{\partial g_{1l}}{\partial y^1}+\frac{\partial g_{l1}}{\partial y^1}-\frac{\partial g_{11}}{\partial y^l}\right) \qquad l=1,2\\
&=\frac{1}{2}g^{11}\left(\frac{\partial g_{11}}{\partial y^1}+\frac{\partial g_{11}}{\partial y^1}-\frac{\partial g_{11}}{\partial y^1}\right)+\frac{1}{2}g^{12}\left(\frac{\partial g_{12}}{\partial y^1}+\frac{\partial g_{21}}{\partial y^1}-\frac{\partial g_{11}}{\partial y^2}\right)=0\\
\Gamma_{11}^2&=\frac{1}{2}g^{2l}\left(\frac{\partial g_{1l}}{\partial y^1}+\frac{\partial g_{l1}}{\partial y^1}-\frac{\partial g_{11}}{\partial y^l}\right) \qquad l=1,2\\
&=\frac{1}{2}g^{21}\left(\frac{\partial g_{11}}{\partial y^1}+\frac{\partial g_{11}}{\partial y^1}-\frac{\partial g_{11}}{\partial y^1}\right)+\frac{1}{2}g^{22}\left(\frac{\partial g_{12}}{\partial y^1}+\frac{\partial g_{21}}{\partial y^1}-\frac{\partial g_{11}}{\partial y^2}\right)=0\\
\Gamma_{12}^1&=\frac{1}{2}g^{1l}\left(\frac{\partial g_{1l}}{\partial y^2}+\frac{\partial g_{l2}}{\partial y^1}-\frac{\partial g_{21}}{\partial y^l}\right) \qquad l=1,2\\
&=\frac{1}{2}g^{11}\left(\frac{\partial g_{11}}{\partial y^2}+\frac{\partial g_{12}}{\partial y^1}-\frac{\partial g_{21}}{\partial y^1}\right)+\frac{1}{2}g^{12}\left(\frac{\partial g_{12}}{\partial y^2}+\frac{\partial g_{22}}{\partial y^1}-\frac{\partial g_{21}}{\partial y^2}\right)=0\\
\Gamma_{12}^2&=\frac{1}{2}g^{2l}\left(\frac{\partial g_{1l}}{\partial y^2}+\frac{\partial g_{l2}}{\partial y^1}-\frac{\partial g_{21}}{\partial y^l}\right) \qquad l=1,2\\
&=\frac{1}{2}g^{21}\left(\frac{\partial g_{11}}{\partial y^2}+\frac{\partial g_{12}}{\partial y^1}-\frac{\partial g_{11}}{\partial y^1}\right)+\frac{1}{2}g^{22}\left(\frac{\partial g_{12}}{\partial y^2}+\frac{\partial g_{22}}{\partial y^1}-\frac{\partial g_{21}}{\partial y^2}\right)\\
&=\frac{1}{2}g^{22}\frac{\partial g_{22}}{\partial y^1}=\frac{1}{2}\frac{1}{\left(y^1\right)^2}2y^1=\frac{1}{y^1}
\end{aligned} \tag{6.7.421}$$

Hence due to the properties of symmetry:

$$\begin{aligned}
\Gamma_{21}^1&=\Gamma_{12}^1=0\\
\Gamma_{21}^2&=\Gamma_{12}^2=\frac{1}{y^1}
\end{aligned} \tag{6.7.422}$$

In the same way:

$$\begin{aligned}
\Gamma_{22}^1&=\frac{1}{2}g^{1l}\left(\frac{\partial g_{2l}}{\partial y^2}+\frac{\partial g_{l2}}{\partial y^2}-\frac{\partial g_{22}}{\partial y^l}\right) \qquad l=1,2\\
&=\frac{1}{2}g^{11}\left(\frac{\partial g_{21}}{\partial y^2}+\frac{\partial g_{12}}{\partial y^2}-\frac{\partial g_{22}}{\partial y^1}\right)+\frac{1}{2}g^{12}\left(\frac{\partial g_{22}}{\partial y^2}+\frac{\partial g_{22}}{\partial y^2}-\frac{\partial g_{22}}{\partial y^2}\right)\\
&=-\frac{1}{2}g^{11}\frac{\partial g_{22}}{\partial y^1}=-\frac{1}{2}\cdot 1\cdot 2y^1=-y^1\\
\Gamma_{22}^2&=\frac{1}{2}g^{2l}\left(\frac{\partial g_{2l}}{\partial y^2}+\frac{\partial g_{l2}}{\partial y^2}-\frac{\partial g_{22}}{\partial y^l}\right) \qquad l=1,2\\
&=\frac{1}{2}g^{21}\left(\frac{\partial g_{21}}{\partial y^2}+\frac{\partial g_{12}}{\partial y^2}-\frac{\partial g_{22}}{\partial y^1}\right)+\frac{1}{2}g^{22}\left(\frac{\partial g_{22}}{\partial y^2}+\frac{\partial g_{22}}{\partial y^2}-\frac{\partial g_{22}}{\partial y^2}\right)\\
&=\frac{1}{2}g^{22}\frac{\partial g_{22}}{\partial y^2}=0
\end{aligned} \tag{6.7.423}$$

To sum up:

$$\begin{array}{cccc} \Gamma_{11}^{1}=0 & \Gamma_{12}^{1}=0 & \Gamma_{21}^{1}=0 & \Gamma_{22}^{1}=-y^{1} \\ \Gamma_{11}^{2}=0 & \Gamma_{12}^{2}=\dfrac{1}{y^{1}} & \Gamma_{21}^{2}=\dfrac{1}{y^{1}} & \Gamma_{22}^{2}=0 \end{array} \tag{6.7.424}$$

6.7.11 Ricci Theorem

Before reading what follows ... We want to remind the reader that the writing of this section is not finished (as most chapters in the book)! Thus, we still have to illustrate the abstract notions that will follow with concrete practical examples but that are examples outside of the special case of General Relativity!

This being said, we have seen in the section of General Relativity that geodesics are the shortest distances between two points in any type of space. What will interest us now is to study the variations of a vector during such a displacement. Let us first recall that the geodesics equation for any curvilinear coordinate system y^i of the punctual space $\mathcal{E}^n$ (see section Principles page 3145) is given by (see section General Relativity page 5747):

$$\frac{\mathrm{d}^2y^i}{\mathrm{d}s^2}+\Gamma_{kj}^{i}\frac{\mathrm{d}y^i}{\mathrm{d}s}\frac{\mathrm{d}y^k}{\mathrm{d}s}=0 \tag{6.7.425}$$

Let us consider now a vector $\vec{v}$ of $\mathcal{E}^n$ of covariant components v_i and let us form the dot product of the vector $\vec{v}$ and $\vec{n}=\mathrm{d}y/\mathrm{d}s$ (the latter vector, denoted here abusively with the indices, gives the components tangent to the geodesic on which the first vector flows), we then have the following quantity:

$$\vec{v}\circ\vec{n}=v_i\frac{\mathrm{d}y^i}{\mathrm{d}s} \tag{6.7.426}$$

When moving along the geodesic from a point M to an infinitely near point M', the scalar undergoes the variation:

$$\mathrm{d}\left(v_i\frac{\mathrm{d}y^i}{\mathrm{d}s}\right)=\mathrm{d}v_k\frac{\mathrm{d}y^k}{\mathrm{d}s}+v_i\mathrm{d}\left(\frac{\mathrm{d}y^i}{\mathrm{d}s}\right) \tag{6.7.427}$$

and as:

$$\frac{\mathrm{d}\left(\dfrac{\mathrm{d}y^i}{\mathrm{d}s}\right)}{\mathrm{d}s}=\frac{\mathrm{d}^2y^i}{\mathrm{d}s^2}\Rightarrow\mathrm{d}\left(\frac{\mathrm{d}y^i}{\mathrm{d}s}\right)=\frac{\mathrm{d}^2y^i}{\mathrm{d}s^2}\mathrm{d}s \tag{6.7.428}$$

Hence:

$$\mathrm{d}\left(v_i\frac{\mathrm{d}y^i}{\mathrm{d}s}\right)=\mathrm{d}v_k\frac{\mathrm{d}y^k}{\mathrm{d}s}+v_i\frac{\mathrm{d}^2y^i}{\mathrm{d}s^2}ds \tag{6.7.429}$$

Let us replace in this last expression, on the one hand, the differential of $\mathrm{d}v_K$ by its exact total differential that we rewrite a bit:

$$\mathrm{d}v_k=\frac{\partial v_k}{\partial y_j}\mathrm{d}y^i=\partial_j v_k\frac{\mathrm{d}s}{\mathrm{d}s}dy^i=\partial_j v_k\frac{\mathrm{d}y^i}{\mathrm{d}s}\mathrm{d}s \qquad (6.7.430)$$

and on the other hand, the second derivative $\mathrm{d}^2y^i/\mathrm{d}s$ by its expression taken from the equation of geodesics. We are getting:

$$\mathrm{d}\left(v_i\frac{\mathrm{d}y^i}{\mathrm{d}s}\right)=\partial_j v_k\frac{\mathrm{d}y^i}{\mathrm{d}s}\frac{\mathrm{d}y^k}{\mathrm{d}s}\mathrm{d}s-v_i\Gamma_{kj}^i\frac{\mathrm{d}y^j}{\mathrm{d}s}\frac{\mathrm{d}y^k}{\mathrm{d}s}\mathrm{d}s=(\partial_i v_k-v_i\Gamma_{kj}^i)\frac{\mathrm{d}y^j}{\mathrm{d}s}\frac{\mathrm{d}y^k}{\mathrm{d}s}\mathrm{d}s \qquad (6.7.431)$$

which can also be written:

$$\mathrm{d}(\vec{v}\circ\vec{n})=(\partial_i v_k-v_i\Gamma_{kj}^i)\mathrm{d}y^j\frac{\mathrm{d}y^k}{\mathrm{d}s}=Dv_k\frac{\mathrm{d}y^k}{\mathrm{d}s} \qquad (6.7.432)$$

Where we have put:

$$\mathrm{D}v_k=(\partial_j v_k-v_i\Gamma_{kj}^i)\mathrm{d}y^i \qquad (6.7.433)$$

which are by definition the absolute differentials of the covariant components of the vector $\vec{v}$. We also define the "covariant derivative for 1-forms" (also named "connection" or "affine connection") by the relation:

$$\boxed{\nabla_j v_k=\partial_j v_k-v_i\Gamma_{kj}^i} \qquad (6.7.434)$$

Remark

In ancient or American textbooks this is often written in the form (which we will not use in this textbook):

$$v_{k;j}=v_{k,j}-v_i\Gamma_{kj}^i \qquad (6.7.435)$$

making use of the ";" to denote the covariant derivative and the "," for the partial differential.

Since the derivative of the product of two functions is the sum of the partial derivatives, then we also have:

$$\begin{aligned}\nabla_a(t_br_c)&=r_c\nabla_a t_b+t_b\nabla_a r_c=r_c(\partial_a t_b-t_e\Gamma_{ab}^e)+t_b(\partial_a r_c-r_e\Gamma_{ac}^e)\\&=r_c\partial_a t_b-r_ct_e\Gamma_{ab}^e+t_b\partial_a r_c-t_br_e\Gamma_{ac}^e=r_c\partial_a t_b+t_b\partial_a r_c-r_ct_e\Gamma_{ab}^e-t_br_e\Gamma_{ac}^e\\&=\partial_a(t_br_c)-r_ct_e\Gamma_{ab}^e-t_br_e\Gamma_{ac}^e\end{aligned} \qquad (6.7.436)$$

If we put $t_br_c\equiv\nabla_j v_i$ then we have (result which we will use after having proved the Ricci theorem to further determine the Einstein tensor necessary in the section of General Relativity):

$$\boxed{\nabla_k(\nabla_j v_i)=\partial_k(\nabla_j v_i)-(\nabla_j v_r)\Gamma_{ik}^r-(\nabla_r v_i)\Gamma_{jk}^r} \qquad (6.7.437)$$

In curvilinear coordinates, in order for the differential of a vector to be a vector, the two vectors of which we take the difference must be in the same point of space. In other words, one of the two infinitely close vectors must be transported in one way or another to the point where the second is located, and only after making the difference between the two vectors which are now in one and only one point of the same punctual space. The "parallel transport" operation must be defined in such a way that in cartesian coordinates (for the small example...), the difference of the components coincides with the ordinary difference $\mathrm{d}v_k$.

Thus, we have indeed in Cartesian coordinates:

$$\mathrm{D}v_k=(\partial_j v_k-v_i\Gamma_{kj}^i)\mathrm{d}y^i=\mathrm{d}v_k \tag{6.7.438}$$

since in this system: $\Gamma_{kj}^i=0$.

Thus, in curvilinear coordinates, the difference of the components of the two vectors after the transport of one of them to the point where the other is denoted δv_k such that we have:

$$\mathrm{D}v_k=(\partial_i v_k-v_i\Gamma_{kj}^i)\mathrm{d}y^j=\mathrm{d}v_k-\delta v_k \tag{6.7.439}$$

This brings us by identification to write:

$$\delta v_k=v_i\Gamma_{kj}^i\mathrm{d}y^i \tag{6.7.440}$$

But also to write the principle of least action (variational principle) in the tensorial form:

$$\mathrm{D}v_k=dv_k-\delta v_k=0 \tag{6.7.441}$$

Let us consider now a tensor of order two, product of two tensors of order one such that (as we have seen in our study of tensor compositions):

$$\begin{aligned}\delta(A_iB_j)=\delta A_{ij}=A_i\delta B_j+B_j\delta A_i&=A_iB_k\Gamma_{jk}^k\mathrm{d}y^h+B_jA_k\Gamma_{ih}^k\mathrm{d}y^h\\&=(A_iB_k\Gamma_{jh}^k+B_jA_k\Gamma_{ih}^k)\mathrm{d}y^k=\delta A_{ij}=(A_{ik}\Gamma_{jh}^k+B_{jk}\Gamma_{ih}^k)\mathrm{d}y^h\end{aligned} \tag{6.7.442}$$

Therefore:

$$\frac{\delta v_{ij}}{\mathrm{d}y^k}=\partial_h v_{ij}=(v_{ik}\Gamma_{jh}^k+v_{jk}\Gamma_{ih}^k) \tag{6.7.443}$$

hence (we take out the last two equalities just for aesthetics!):

$$\partial_k v_{ij}=(v_{ik}\Gamma_{jh}^k+v_{jk}\Gamma_{ih}^k) \tag{6.7.444}$$

A parallel transport is therefore an operation that takes a tangent vector and moves it along a path in space without turning it (relative to the space) or changing its length. In flat space we can say that the transported vector is parallel to the original vector at every point along the path. In curved space we cannot say such a thing. Let's use the spherical surface of Earth to show this! We start at the equator at longitude 0° holding an arrow pointing north:

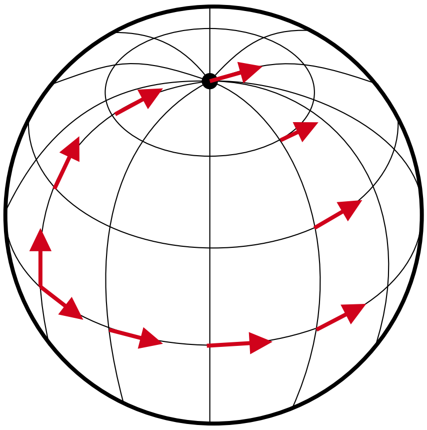


Figure 6.161: **Parallel transport illustration on a sphere**

We go along longitude 0° up to the North Pole, keeping our arrow parallel to the ground and pointing forward all the time. When we get to the North Pole our arrow is pointing in the direction of longitude 180°. Now we go South along longitude 90° East keeping our arrow perpendicular to our path as it was at the North Pole. When we get to the equator our arrow is pointing to the east. Finally, we go West along the equator until we get back to our starting point. We keep the arrow pointing backwards all the way. Though we did not turn the arrow all the way, we are now at the starting point with our arrow turned 90° relative to its original position. We surely can't say that it is now parallel to the original position. This means that the term "direction" cannot be defined globally in a curved space. We can only compare the direction of two vectors if they are at the same point.

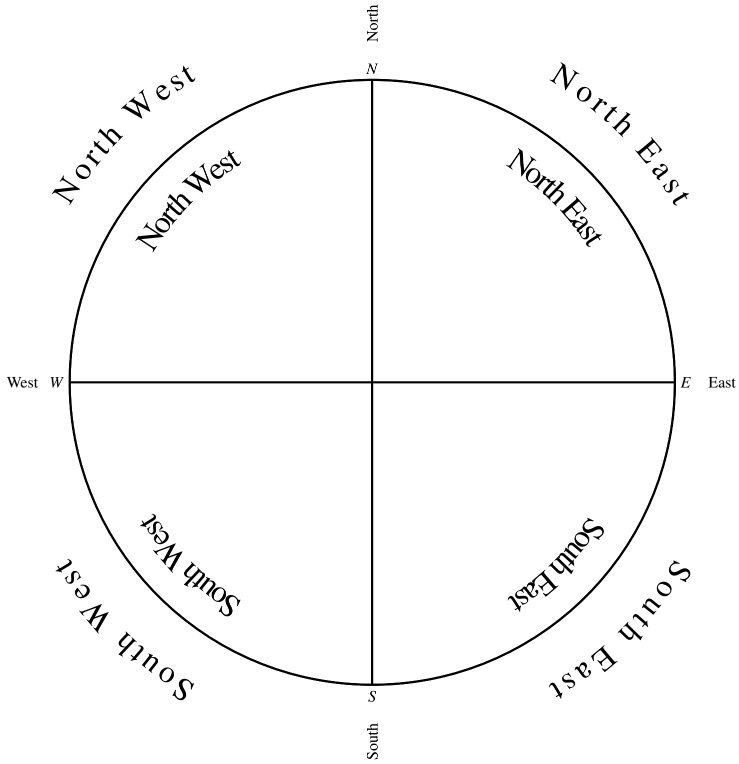


Figure 6.162: Reminder of cardinals names

Definitions (#6.7.266):

D1. "East" is one of the four cardinal directions or points of a compass. It is the opposite direction from West and is the direction from which the Sun rises on the Earth (at least at the day we write these lines...). The word "East" shouldn't be used in the field of geometry excepted when we deal with celestial objects that have a dipolar magnetic field or a sunrise and sunset however even some scientists sadly still associate the "East" with the concept of "Right"...

D2. "West" is one of the four cardinal directions or points of a compass. It is the opposite direction from East and is the direction from which the Sun set on the Earth (at least at the day we write these lines...). The word "West" shouldn't be used in the field of geometry excepted when we deal with celestial objects that have a dipolar magnetic field or a sunrise and sunset however even some scientists sadly still associated the "West" with the concept of "Left"...

D3. "North" is one of the four cardinal directions or points of a compass. It is the opposite of South and is perpendicular to East and West. The word "North" shouldn't be used in the field of geometry excepted when we deal with celestial objects that have a dipolar magnetic field however even some scientists sadly still associated the "North" with the concept of "Top/Up"...

D4. "South" is one of the four cardinal directions or points of a compass. It is the opposite of North and is perpendicular to East and West. The word "North" shouldn't be used in the field of geometry excepted when we deal with celestial objects that have a dipolar magnetic field however even some scientists sadly still associated the "South" with the concept of "Bottom/Down"... The South is differentiate from the North based on the physicist Ampere right hand rule (see section Magnetostatics 4010).

The fact that parallel transport along a closed loop changes the direction of a vector in a curved space but not in a flat space may lead to the idea of using it as a way to measure curvature. It turns out that if we choose a loop that is small enough around a point in a curved space, the amount of change in the direction of a vector that is parallel transported along it is proportional to the area enclosed by the loop. So, the ratio between the area of the loop and the amount of change in the direction of the vector (whatever way we chose to measure it) can be used as a measure to the curvature of the surface that includes the loop. Actually we define curvature by the value of this ratio.

The previous relation leads us to be able to write the metric in its variational form named "Ricci identity":

$$\boxed{\partial_h g_{ij}=g_{ik}\Gamma_{jh}^k+g_{jk}\Gamma_{ih}^k} \qquad (6.7.445)$$

But we also have since $g_{ij}=\vec{e}_i\circ\vec{e}_j$:

$$\mathrm{d}g_{ij}=\vec{e}_i\circ\mathrm{d}\vec{e}_j+\mathrm{d}\vec{e}_i\circ\vec{e}_j=\vec{e}_i\circ\omega_j^k\vec{e}_k+\omega_i^k\vec{e}_k\circ\vec{e}_j=\omega_j^k(\vec{e}_i\circ\vec{e}_k)+\omega_i^k(\vec{e}_k\circ\vec{e}_j) \qquad (6.7.446)$$

hence the identity:

$$\mathrm{d}g_{ij}=\omega_j^k g_{ik}+\omega_i^k g_{jk} \qquad (6.7.447)$$

With the both relations:

$$\mathrm{d}g_{ij}=\omega_j^k g_{ik}+\omega_i^k g_{jk} \qquad \text{and} \qquad \delta g_{ij}=(g_{ik}\Gamma_{jh}^k+g_{jk}\Gamma_{ih}^k)\mathrm{d}y^h \qquad (6.7.448)$$

and the absolute differential (which is simply generalized for a tensor of order two):

$$\mathrm{D}g_{ij}=\mathrm{d}g_{ij}-\delta g_{ij} \qquad (6.7.449)$$

we get:

$$\mathrm{D}g_{ij}=\omega_j^k g_{ik}+\omega_i^k g_{jk}-g_{ik}\Gamma_{jh}^k\mathrm{d}y^h-g_{jk}\Gamma_{ih}^k\mathrm{d}y^h \qquad (6.7.450)$$

Now, let us recall that we have by definition:

$$\omega_j^k=\Gamma_{jh}^k \mathrm{d}y^h \qquad \text{and} \qquad \omega_i^k=\Gamma_{ih}^k \mathrm{d}y^h \tag{6.7.451}$$

Hence finally:

$$\boxed{\mathrm{D}g_{ij}=0} \tag{6.7.452}$$

The absolute differential on a geodesic in the approximation of an infinitesimal transport of the fundamental tensor is therefore (as we could expect) zero. This is the "Ricci theorem". Some theoretical physicists then say that «*the covariant derivative kills the metric*» in the sense that the metric does not change on a space differential.

Finally, we also see that for a tensor of order two (the metric in particular) we have:

$$\mathrm{D}g_{ij}=\mathrm{d}g_{ij}-\delta g_{ij}=\mathrm{d}g_{ij}-g_{ik}\Gamma_{jh}^k \mathrm{d}y^h-g_{jk}\Gamma_{ih}^k \mathrm{d}y^h \tag{6.7.453}$$

We can therefore write the absolute differential which in this particular case is zero:

$$\mathrm{D}g_{ij}=(\partial_h g_{ij}-g_{ik}\Gamma_{jh}^k-g_{jk}\Gamma_{ih}^k)\mathrm{d}y^h=0 \tag{6.7.454}$$

And therefore the "covariant derivative of a $(0,2)$-form" of the metric is null:

$$\boxed{\frac{\mathrm{D}g_{ij}}{\mathrm{d}y^h}=\nabla_h g_{ij}=\partial_h g_{ij}-g_{ik}\Gamma_{jh}^k-g_{jk}\Gamma_{ih}^k=0} \tag{6.7.455}$$

Remark

The reader will have to remember for the definition of the Einstein tensor that:

$$\nabla_k g_{ij}=0 \qquad \text{and} \qquad \nabla_k g_{ij}=\partial_h g_{ij}-g_{ik}\Gamma_{jh}^k-g_{jk}\Gamma_{ih}^k=0 \tag{6.7.456}$$

and that this is another way of expressing that an infinitesimal variation on a geodesic according to the principle of least action kills the metric. We will therefore work from now on (as before) with non-linear differential equations that must be integrated to find the behaviour of matter in a given space.

In general, if T is an (r,s) tensor field (we will need that during our study of General Relativity), then ∇T is an $(r,s+1)$-tensor field given by:

$$(\nabla T)(\eta_1,\ldots,\eta_r,X,Y_1,\ldots,Y_s)=(\nabla_X T)(\eta_1,\ldots,\eta_r,Y_1,\ldots,Y_s) \tag{6.7.457}$$

where the latter expression is intrinsically defined by the equation:

$$\begin{aligned}\nabla_X(T(\eta_1,\ldots,\eta_r,Y_1,\ldots,Y_s))=&(\nabla_X T)(\eta_1,\ldots,\eta_r,Y_1,\ldots,Y_s)\\&+\sum_{i=1}^{r}T(\eta_1,\ldots,\eta_{i-1},\nabla_X\eta_i,\eta_{i+1},\ldots,\eta_r,Y_1,\ldots,Y_s)\\&+\sum_{j=1}^{s}T(\eta_1,\ldots,\eta_r,Y_1,\ldots,Y_{j-1},\nabla_X Y_j,Y_{j+1},\ldots,Y_s)\end{aligned} \tag{6.7.458}$$

In particular, if T is a $(1,1)$-tensor, then $(\nabla_X T)(\eta, Y)$ is intrinsically defined by the equation:

$$\nabla_X(T(\eta,Y)) = (\nabla_X T)(\eta,Y) + T(\nabla_X \eta, Y) + T(\eta, \nabla_X Y) \tag{6.7.459}$$

and therefore the correct expression for $(\nabla_X T)(\eta, Y)$ is:

$$(\nabla_X T)(\eta,Y) := \nabla_X(T(\eta,Y)) - T(\nabla_X \eta, Y) - T(\eta, \nabla_X Y) \tag{6.7.460}$$

This will help us to derive a special case of:

$$\begin{aligned} T^{\mu_1 \ldots \mu_r}{}_{\nu_1 \ldots \nu_s;\rho} = T^{\mu_1 \ldots \mu_r}{}_{\nu_1 \ldots \nu_s,\rho} &+ \Gamma^{\mu_1}_{\sigma\rho} T^{\sigma \ldots \mu_r}{}_{\nu_1 \ldots \nu_s} + \ldots + \Gamma^{\mu_r}_{\sigma\rho} T^{\mu_1 \ldots \sigma}{}_{\nu_1 \ldots \nu_s} \\ &- \Gamma^{\sigma}_{\nu_1\rho} T^{\mu_1 \ldots \mu_r}{}_{\sigma \ldots \nu_s} - \ldots - \Gamma^{\sigma}_{\nu_s\rho} T^{\mu_1 \ldots \mu_r}{}_{\nu_1 \ldots \sigma} \end{aligned} \tag{6.7.461}$$

for the case of a $(1,1)$-tensor (as we will have during our Study of General Relativity).

So, by the definition above in the case of a $(1,1)$-tensor we have $T^{\mu}{}_{\nu;\rho} := \nabla_{e_\rho} T(f^\mu, e_\nu)$, where we are using f^μ and e_v for $\mathrm{d}x^\mu$ and $\partial / \partial x^\nu$.

Hence we have (it's very not obvious so maybe in the future we will detail the process more or found another approach):

$$\nabla_{e_\rho} T(f^\mu, e_\nu) := \nabla_{e_\rho}(T(f^\mu, e_\nu)) - T(\nabla_{e_\rho} f^\mu, e_\nu) - T(f^\mu, \nabla_{e_\rho} e_\nu) \tag{6.7.462}$$

The first and third terms are easily evaluated:

$$\nabla_{e_\rho}(T(f^\mu, e_\nu)) = e_\rho(T(f^\mu, e_\nu)) = e_\rho(T^\mu{}_\nu) = T^\mu{}_{\nu,\rho} \tag{6.7.463}$$

$$T(f^\mu, \nabla_{e_\rho} e_\nu) = T(f^\mu, \Gamma^\sigma_{\nu\rho} e_\sigma) = \Gamma^\sigma_{\nu\rho} T(f^\mu, e_\sigma) = \Gamma^\sigma_{\nu\rho} T^\mu{}_\sigma \tag{6.7.464}$$

For the second term, since f^μ is a $(0,1)$-tensor field, ∇f^μ is a $(0,2)$-tensor field and hence $\nabla_{e_\rho} f^\mu$, for fixed e_ρ, is a $(0,1)$-tensor field. We evaluate:

$$\begin{aligned} (\nabla_{e_\rho} f^\mu)_\lambda = \nabla_{e_\rho} f^\mu(e_\lambda) &:= \nabla_{e_\rho}(f^\mu(e_\lambda)) - f^\mu(\nabla_{e_\rho} e_\lambda) \\ &= e_\rho(\delta^\mu_\lambda) - f^\mu(\Gamma^\sigma_{\lambda\rho} e_\sigma) = 0 - \Gamma^\sigma_{\lambda\rho} f^\mu(e_\sigma) = -\Gamma^\sigma_{\lambda\rho}\delta^\mu_\sigma = -\Gamma^\mu_{\lambda\rho} = -\Gamma^\mu_{\sigma\rho}\delta^\sigma_\lambda \\ &= -\Gamma^\mu_{\sigma\rho} f^\sigma(e_\lambda) \end{aligned} \tag{6.7.465}$$

Therefore $\nabla_{e_\rho} f^\mu = -\Gamma^\mu_{\sigma\rho} f^\sigma$ and hence:

$$T(\nabla_{e_\rho} f^\mu, e_\nu) = T(-\Gamma^\mu_{\sigma\rho} f^\sigma, e_\nu) = -\Gamma^\mu_{\sigma\rho} T(f^\sigma, e_\nu) = -\Gamma^\mu_{\sigma\rho} T^\sigma{}_\nu \tag{6.7.466}$$

Thus, we have:

$$T^\mu{}_{\nu;\rho} = T^\mu{}_{\nu,\rho} + \Gamma^\mu_{\sigma\rho} T^\sigma{}_\nu - \Gamma^\sigma_{\nu\rho} T^\mu{}_\sigma \tag{6.7.467}$$

Sometimes denoted in some textbooks as:

$$D_a T^b_c = \partial_a T^b_c + \Gamma^b_{fa} T^f_c - \Gamma^h_{ca} T^b_h \tag{6.7.468}$$

✍**Example:**

Let's try to verify this by calculating one component of the covariant differentiation in the spherical coordinates.

Let us recall first that the diagonal metric of the 2-sphere surface $\mathcal{S}^2$ expression is (see section Differential Geometry 2884):

$$\begin{gathered}g_{ij}=\begin{bmatrix}g_{\theta\theta} & g_{\theta\phi}\\ g_{\phi\theta} & g_{\phi\phi}\end{bmatrix}=\begin{bmatrix}R^2 & 0\\ 0 & R^2\sin^2(\theta)\end{bmatrix}\\ g^{ij}=\begin{bmatrix}g^{\theta\theta} & g^{\theta\phi}\\ g^{\phi\theta} & g^{\phi\phi}\end{bmatrix}=\begin{bmatrix}\frac{1}{R^2} & 0\\ 0 & \frac{1}{R^2\sin^2(\theta)}\end{bmatrix}\end{gathered}\tag{6.7.469}$$

If we were to calculate the component $g_{\phi\phi;\theta}$, we should then write:

$$\nabla_\theta g_{\phi\phi}=\frac{\partial g_{\phi\phi}}{\partial\theta}-g_{\alpha\phi}\Gamma_{\phi\theta}^{\alpha}-g_{\phi\alpha}\Gamma_{\phi\theta}^{\alpha}\tag{6.7.470}$$

But $g_{ij}\neq 0$ only if $i=j$, based on the above expressions, so we can simplify this relation:

$$\nabla_\theta g_{\phi\phi}=\frac{\partial g_{\phi\phi}}{\partial\theta}-g_{\phi\phi}\Gamma_{\phi\theta}^{\phi}-g_{\phi\phi}\Gamma_{\phi\theta}^{\phi}=\frac{\partial g_{\phi\phi}}{\partial\theta}-2g_{\phi\phi}\Gamma_{\phi\theta}^{\phi}\tag{6.7.471}$$

As (see section Differential and Integral Calculus page 1475):

$$\frac{\partial g_{\phi\phi}}{\partial\theta}=\frac{\partial R^2\sin^2(\theta)}{\partial\theta}=2R^2\sin(\theta)\cos(\theta)\tag{6.7.472}$$

Further below (see page 2137) we will prove that for $\mathcal{S}^2$ we have:

$$\Gamma_{ij}^{\theta}=\begin{bmatrix}0 & 0\\ 0 & -\sin(\theta)\cos(\theta)\end{bmatrix}\qquad\Gamma_{ij}^{\phi}=\begin{bmatrix}0 & \cot(\theta)\\ \cot(\theta) & 0\end{bmatrix}=\begin{bmatrix}0 & \frac{\cos(\theta)}{\sin(\theta)}\\ \frac{\cos(\theta)}{\sin(\theta)} & 0\end{bmatrix}\tag{6.7.473}$$

So that:

$$g_{\phi\phi}\Gamma_{\phi\theta}^{\phi}=R^2\sin^2(\theta)\frac{\cos(\theta)}{\sin(\theta)}=R^2\sin(\theta)\cos(\theta)\tag{6.7.474}$$

Finally, we confirm that this component of the covariant derivative with respect to θ equals also zero in a polar coordinates system, as expected:

$$\nabla_\theta g_{\phi\phi}=\frac{\partial g_{\phi\phi}}{\partial\theta}-2g_{\phi\phi}\Gamma_{\theta\phi}^{\phi}=2R^2\sin(\theta)\cos(\theta)-2R^2\sin(\theta)\cos(\theta)=0\tag{6.7.475}$$

Let us now determine an expression which will be very useful in General Relativity when we will determine the Einstein field equations (another way of expressing that the covariant derivative of the metric is null)!

First let us compute (we switch the notation so that reader can take the habit to see the different common one):

$$D_a g_{bc} = \partial_a g_{bc} - \Gamma_{ba}^{f} g_{fc} - \Gamma_{ca}^{h} g_{hc} \tag{6.7.476}$$

As we have:

$$g_{bc} = \vec{e}_b \circ \vec{e}_c \Rightarrow \partial_a g_{bc} = \frac{\partial \vec{e}_b}{\partial x^a} \circ \vec{e}_c + \vec{e}_b \circ \frac{\partial \vec{e}_c}{\partial x^a} \tag{6.7.477}$$

Therefore:

$$\partial_a g_{bc} = \Gamma_{ba}^{f} \underbrace{\vec{e}_f \circ \vec{e}_c}_{g_{fc}} - \Gamma_{ca}^{h} \underbrace{\vec{e}_h \circ \vec{e}_c}_{g_{hc}} \tag{6.7.478}$$

Hence:

$$D_a g_{bc} = \partial_a g_{bc} - \Gamma_{ba}^{f} g_{fc} - \Gamma_{ca}^{h} g_{hc} = 0 \tag{6.7.479}$$

Let us perform now the contracted multiplication of:

$$\nabla_k g_{ij} = \partial_h g_{ij} - g_{ik} \Gamma_{jh}^{k} - g_{jk} \Gamma_{ih}^{k} = 0 \tag{6.7.480}$$

by g^{ij}, it then comes by using the relation $g^{ij} g_{jl} = \delta_i^l$ (which we had proved much earlier above) that:

$$g^{ij} \partial_h g_{ij} - g^{ij} g_{ik} \Gamma_{jh}^{k} - g^{ij} g_{jk} \Gamma_{ih}^{k} = g^{ij} \partial_h g_{ij} - \delta_j^k \Gamma_{jh}^{k} - \delta_i^k \Gamma_{ih}^{k} = 0 \tag{6.7.481}$$

hence the relation:

$$g^{ij} \partial_h g_{ij} - \Gamma_{jh}^{j} - \Gamma_{ih}^{i} = 0 \tag{6.7.482}$$

The quantities Γ_{jh}^{j} and Γ_{ih}^{h} representing the same sums, we then have:

$$g^{ij} \partial_h g_{ij} = 2\Gamma_{ih}^{i} \tag{6.7.483}$$

Theorem 6.46. Let us consider now the determinant g of the quantities g_{ij}. The derivation of the determinant gives us:

$$\partial_h g = g g_{ij} \partial_h g_{ij} \Leftrightarrow \partial_h g_{ij} = \frac{\partial_k g}{g g^{ij}} \tag{6.7.484}$$

Proof 6.46.1. Given any variable that we choose here as being the time t only to simplify the notations of the calculations that will follow. When the main part of the development is completed, the result can be adapted to any other variable! We will write for what will follow g_j the elements of the j-th column of g_{ij}.

For the following developments, we define the notations:

$$g = \det(g_1, g_2, \ldots, g_n) \qquad \dot{g} = d_t g \qquad \dot{g}_j = \mathrm{d}_t g_j \qquad \dot{g}_{ij} = \mathrm{d}_t g_{ij} \tag{6.7.485}$$

The rule of derivation of a functional determinant is given for recall (see section Linear Algebra page 1974) by:

$$\dot{g} = \det(\dot{g}_1, g_2, \ldots, g_n) + \det(g_1, \dot{g}_2, \ldots, g_n) + \ldots + \det(g_1, g_2, \ldots, \dot{g}_n) \tag{6.7.486}$$

By considering the first determinant, by using the minors (see section Linear Algebra page 1974) for the development of its first column:

$$\alpha_{i1}\dot{g}_{i1} \tag{6.7.487}$$

For the j-th determinant, it comes:

$$\alpha_{ij}\dot{g}_{ij} \tag{6.7.488}$$

Thus:

$$\dot{g} = \alpha_{ij}\dot{g}_{ij} \qquad \text{or} \qquad \dot{g} = \alpha_{ij}g_{ij} \tag{6.7.489}$$

Now, we have demonstrated much earlier that the metric tensor is its own inverse. Therefore:

$$g^{ij} = g_{ij}^{-1} \tag{6.7.490}$$

This allows us to write:

$$gg^{ij} = \alpha_{ij} \tag{6.7.491}$$

and therefore:

$$\dot{g} = g\dot{g}_{ij}g^{ij} \tag{6.7.492}$$

what is also written (following the conventions defined at the beginning of the proof):

$$\mathrm{d}_t g = g\mathrm{d}_t g_{ij}g^{ij} \tag{6.7.493}$$

where the reader must therefore be careful not to read badly, typically thinking that the derivative d_t in the right-hand term derives everything $(g_{ij}g^{ij})$... while it only derives g_{ij}.

However, we can adopt another variable. Let h be this other variable:

$$\partial_h g = gg^{ij}\partial_h g_{ij} \tag{6.7.494}$$

Either by rearranging:

$$\boxed{\partial_h g_{ij} = \left(gg^{ij}\right)^{-1}\partial_h g = \frac{1}{g}\left(g^{ij}\right)^{-1}\partial_h g} \tag{6.7.495}$$

This is what we wanted to prove.

■ Q.E.D.

Now by combining:

$$\Gamma^i_{ih} = \frac{1}{2}g^{ij}\partial_h g_{ij} \tag{6.7.496}$$

Demonstrated earlier above and the result we have just proved:

$$\partial_h g_{ij} = \frac{1}{g}\left(g^{ij}\right)^{-1}\partial_h g \tag{6.7.497}$$

it comes:

$$\Gamma^i_{ih}=\frac{1}{2}g^{ij}\partial_h g_{ij}=\frac{1}{2}g^{ij}\left(\frac{1}{g}(g_{ij})^{-1}\partial_h g\right)=\frac{1}{2g}\partial_h g \qquad (6.7.498)$$

Therefore we have:

$$g^{ij}\partial_h g_{ij}=\frac{\partial_h g}{g}=2\Gamma^i_{ih} \qquad (6.7.499)$$

Let us prove that it is possible to derive this last relation from the following important equality:

$$\boxed{\Gamma^i_{ih}=\frac{1}{\sqrt{g}}\partial_h\sqrt{g}} \qquad (6.7.500)$$

Indeed:

$$\Gamma^i_{ih}=\frac{1}{\sqrt{g}}\partial_h\sqrt{g}=\frac{1}{\sqrt{g}}\left(\frac{1}{2\sqrt{g}}\partial_h g\right)=\frac{1}{2g}\partial_h g \qquad (6.7.501)$$

This relation does not mean much until we will make a more explicit use of it in our study of General Relativity (see section General Relativity page 5720).

Let us now determine the second covariant derivative of the metric tensor. Let us remember before we go further (because it is important) that we had obtained:

$$\nabla_h g_{ij}=\partial_h g_{ij}-g_{ik}\Gamma^k_{jh}-g_{jk}\Gamma^k_{ih} \qquad (6.7.502)$$

6.7.12 Riemann-Christoffel symbols

Before reading what follows ... We want to remind the reader again that the writing of this section is not finished (as most chapters in the book)! Thus, we still have to illustrate the abstract notions that will follow with concrete practical examples but that are example outside of the special case of General Relativity!

Let us recall that we have demonstrated earlier above that:

$$\nabla_k(\nabla_j v_i)=\partial_k(\nabla_j v_i)-(\nabla_j v_r)\Gamma^r_{ik}-(\nabla_r v_i)\Gamma^r_{jk} \qquad (6.7.503)$$

This relation means, exception of an interpretation error from the writer of those linear as sometimes interpreting Tensor Calculus result is a pain in the a.., that the covariant derivative of a tensor of order two - such as the metric - on a geodesic path in two perpendicular directions (the second covariant derivative making it possible to the "perpendicular geodesic" between the two geodesics infinitely close to the first covariant derivative). We know already that such a shift is "parallel transport".

By substituting in it:

$$\nabla_j v_i=\partial_j v_i-\Gamma^l_{ji}v_l \qquad (6.7.504)$$

We then have:

$$\begin{aligned}\nabla_k(\nabla_j v_i)&=\partial_k(\partial_j v_i-\Gamma^l_{ji}v_l)-(\partial_j v_r-\Gamma^l_{jr}v_l)\Gamma^r_{ik}-(\partial_r v_i-\Gamma^l_{ri}v_l)\Gamma^r_{jk}\\&=\partial_{kj}v_i-(\partial_k\Gamma^l_{ji})v_l-\Gamma^l_{ji}(\partial_k v_l)-\Gamma^r_{ik}\partial_j v_r+\Gamma^l_{jr}\Gamma^r_{ik}v_l-\Gamma^r_{jk}\partial_r v_i+\Gamma^r_{jk}\Gamma^l_{ri}v_l\end{aligned}\tag{6.7.505}$$

Let us now switch the indices j and k in the previous expression to have a differential with respect to another path:

$$\nabla_j(\nabla_k v_i)=\partial_{jk}v_i-(\partial_j\Gamma^l_{ki})v_l-\Gamma^l_{ki}(\partial_j v_l)-\Gamma^r_{ij}\partial_k v_r+\Gamma^l_{kr}\Gamma^r_{ij}v_l-\Gamma^r_{kj}\partial_r v_i+\Gamma^r_{kj}\Gamma^l_{ri}v_l\tag{6.7.506}$$

Assuming that the components satisfy the classical properties $\partial_{kj}v_i=\partial_{jk}v_i$, we get by subtraction of the two previous expressions:

$$\begin{array}{rl}&\nabla_k(\nabla_j v_i)=\partial_{kj}v_i-(\partial_k\Gamma^l_{ji})v_l-\Gamma^l_{ji}(\partial_k v_l)-\Gamma^r_{ik}\partial_j v_r+\Gamma^l_{jr}\Gamma^r_{ik}v_l-\Gamma^r_{jk}\partial_r v_i+\Gamma^r_{jk}\Gamma^l_{ri}v_l\\-&\nabla_j(\nabla_k v_i)=\partial_{jk}v_i-(\partial_j\Gamma^l_{ki})v_l-\Gamma^l_{ki}(\partial_j v_l)-\Gamma^r_{ij}\partial_k v_r+\Gamma^l_{kr}\Gamma^r_{ij}v_l-\Gamma^r_{kj}\partial_r v_i+\Gamma^r_{kj}\Gamma^l_{ri}v_l\\\hline&\nabla_k(\nabla_j v_i)-\nabla_j(\nabla_k v_i)=\underbrace{(\partial_{kj}v_i-\partial_{jk}v_i)}_{=0}+(\partial_j\Gamma^l_{ki}-\partial_k\Gamma^l_{ji})v_l+(\Gamma^l_{ki}\partial_j-\Gamma^l_{ji}\partial_k)v_l\\&\qquad+(\Gamma^r_{ij}\partial_k-\Gamma^r_{ik}\partial_j)v_r+(\Gamma^l_{jr}\Gamma^r_{ik}-\Gamma^l_{kr}\Gamma^r_{ij})v_l+(\Gamma^r_{kj}-\Gamma^r_{jk})\partial_r v_i+(\Gamma^r_{jk}\Gamma^l_{ri}-\Gamma^r_{kj}\Gamma^l_{ri})v_l\end{array}\tag{6.7.507}$$

And since we have proved that in the case of a symmetric metric we have:

$$\Gamma^r_{jk}=\Gamma^r_{kj}\tag{6.7.508}$$

We have therefore:

$$\begin{aligned}\nabla_k(\nabla_j v_i)-\nabla_j(\nabla_k v_i)=&(\partial_j\Gamma^l_{ki}-\partial_k\Gamma^l_{ji})v_l+(\Gamma^l_{ki}\partial_j-\Gamma^l_{ji}\partial_k)v_l\\&+(\Gamma^r_{ij}\partial_k-\Gamma^r_{ik}\partial_j)v_r+(\Gamma^l_{jr}\Gamma^r_{ik}-\Gamma^l_{kr}\Gamma^r_{ij})v_l+\underbrace{(\Gamma^r_{kj}-\Gamma^r_{jk})}_{=0}\partial_r v_i+\underbrace{(\Gamma^r_{jk}\Gamma^l_{ri}-\Gamma^r_{kj}\Gamma^l_{ri})}_{=0}v_l\end{aligned}\tag{6.7.509}$$

Remark

The fact of having in the case of a symmetric metric $\Gamma^r_{jk}=\Gamma^r_{kj}$ and which vanishes in the preceding relation, leads many practitioners to define what we name the "torsion tensor" or "torsor" (but in reality it is a particular case of a more general relation which belongs to the domain of differential geometry). Thus, we define the torsion tensor as:

$$T^k_{ih}=\Gamma^k_{ij}-\Gamma^k_{ji}\tag{6.7.510}$$

and in the case of a symmetric metric (Euclidean space), the torsion is zero by extension as we have already seen it! In fact, the Einstein field equations which we shall prove later implicitly implies a symmetric metric with zero torsion. However, it is possible to proved that a non-symmetric tensor can always be decomposed into a symmetric and non-symmetric tensor (this is trivial because it is like separating a complete matrix in the sum of a matrix having a diagonal of non-zero component that are and another matrix having the diagonal zero).

It then remains:

$$\begin{aligned}\nabla_k(\nabla_j v_i)-\nabla_j(\nabla_k v_i)=&(\partial_j\Gamma^l_{ki}-\partial_k\Gamma^l_{ji})v_l+(\Gamma^l_{ki}\partial_j-\Gamma^l_{ji}\partial_k)v_l\\&+(\Gamma^r_{ij}\partial_k-\Gamma^r_{ik}\partial_j)v_r+(\Gamma^l_{jr}\Gamma^r_{ik}-\Gamma^l_{kr}\Gamma^r_{ij})v_l\end{aligned}\tag{6.7.511}$$

Since parallel transport takes place on infinitely close geodesic paths, we take the limit:

$$\partial_j v_l, \partial_k v_l, \partial_k v_r, \partial_j v_r \longrightarrow 0 \tag{6.7.512}$$

Which mainly underlies the fact that the velocity field is almost equal in two infinitely close parallel points.

It then remains:

$$\boxed{\nabla_k(\nabla_j v_i) - \nabla_j(\nabla_k v_i) \cong (\partial_j \Gamma_{ki}^l - \partial_k \Gamma_{ji}^l + \Gamma_{jr}^l \Gamma_{ik}^r - \Gamma_{kr}^l \Gamma_{ij}^r) v_l} \tag{6.7.513}$$

This relation expresses the fact that, like gravity, the curvature of space-time causes a mutual acceleration between the geodesics! Moreover, it is easy to see that the mutual acceleration between the geodesics is zero if the Riemann-Christoffel tensors are null (typically in Cartesian coordinates, in extenso it means for a flat time-space). This is exactly what we expect of gravity: if we do not observe any acceleration, the curvature (we shall now define what it is) is zero and if the curvature is zero, we observe no acceleration. Morale of the story: the gravity is curvature and the curvature is gravity !!

We see that the quantity in parentheses is a tensor of order four that we will write on this site as following (because there are several traditions in the way to write it...):

$$\boxed{R_{i,jk}^l = \partial_j \Gamma_{ik}^l - \partial_k \Gamma_{ij}^l + \Gamma_{ik}^r \Gamma_{jr}^l - \Gamma_{ij}^r \Gamma_{kr}^l} \tag{6.7.514}$$

and which summarizes itself the parallel transport and the fact that gravity and geometry of space are linked together. Obviously, if the metric is of Minkowski type (or tends to a metric of Minkowski under certain conditions), then $R_{i,jk}^l$ is zero! Very rare authors write this last equality in the (unhappy ....) form:

$$R_{i,jk}^l = 2\Gamma_{i[k,j]}^l + 2\Gamma_{i[k}^r \Gamma_{j]r}^l \tag{6.7.515}$$

The tensor $R_{i,jk}^l$ is named the "Riemann-Christoffel tensor" or "Riemannian space tensor". The curvature of a Riemannian space can also be characterized using this tensor.

A fourth-rank tensor has $4^4 = 256$ components as we know. However, because of that fact that in many situations the metric is symmetric, the number of algebraically independent components of Riemann-Christoffel tensor is only 20.

If we multiply the tensor $R_{i,rs}^k$ by g_{jk}, then we have the covariant components of this tensor such that:

$$\boxed{R_{ij,rs} = g_{jk} R_{i,rs}^k} \tag{6.7.516}$$

and given the following relations that we proved earlier above:

$$g_{jk} \Gamma_{rl}^k = \Gamma_{rjl} \qquad g_{jk} \Gamma_{sl}^k = \Gamma_{sjl} \tag{6.7.517}$$

Therefore we get:

$$\begin{aligned} g_{jk} R_{i,rs}^k &= g_{jk} \partial_r \Gamma_{is}^k - g_{jk} \partial_s \Gamma_{ir}^k + g_{jk} \Gamma_{is}^l \Gamma_{rl}^k - g_{jk} \Gamma_{ir}^l \Gamma_{sl}^k \\ &= g_{jk} \partial_r \Gamma_{is}^k - g_{jk} \partial_s \Gamma_{ir}^k + \Gamma_{is}^l \Gamma_{rjl} - \Gamma_{ir}^l \Gamma_{sjl} \end{aligned} \tag{6.7.518}$$

and let us replace the quantities $g_{jk}\partial_r\Gamma^k_{is}$ by $\partial_r\left(g_{jk}\Gamma^k_{is}\right)-\Gamma^k_{is}\partial_r g_{jk}$. We then get:

$$\begin{aligned}g_{jk}R^k_{i,rs}&=(\partial_r(g_{jk}\Gamma^k_{is})-\Gamma^k_{is}\partial_r g_{jk})-(\partial_s(g_{jk}\Gamma^k_{ir})-\Gamma^k_{ir}\partial_s g_{jk})+\Gamma^l_{is}\Gamma_{rjl}-\Gamma^l_{ir}\Gamma_{sjl}\\&=\partial_r(g_{jk}\Gamma^k_{is})-\partial_s(g_{jk}\Gamma^k_{ir})+(\Gamma^l_{is}\Gamma_{rjl}-\Gamma^k_{is}\partial_r g_{jk})-(\Gamma^l_{ir}\Gamma_{sjl}-\Gamma^k_{ir}\partial_s g_{jk})\\&=\partial_r(g_{jk}\Gamma^k_{is})-\partial_s(g_{jk}\Gamma^k_{ir})+(\Gamma^l_{is}\Gamma_{rjl}-\Gamma^l_{is}\partial_r g_{jl})-(\Gamma^l_{ir}\Gamma_{sjl}-\Gamma^l_{ir}\partial_s g_{jl})\\&=\partial_r(g_{jk}\Gamma^k_{is})-\partial_s(g_{jk}\Gamma^k_{ir})+\Gamma^l_{is}(\Gamma_{rjl}-\partial_r g_{jl})-\Gamma^l_{ir}(\Gamma_{sjl}-\partial_s g_{jl})\end{aligned}\tag{6.7.519}$$

We also proved earlier before that:

$$\partial_k g_{ij}=\Gamma_{kji}+\Gamma_{kij}\Rightarrow\Gamma_{kij}-\partial_k g_{ij}=-\Gamma_{kji}\tag{6.7.520}$$

Hence:

$$g_{jk}R^k_{i,rs}=\partial_r(g_{jk}\Gamma^k_{is})-\partial_s(g_{jk}\Gamma^k_{ir})-\Gamma^l_{is}\Gamma_{rlj}+\Gamma^l_{ir}\Gamma_{slj}\tag{6.7.521}$$

and as:

$$g_{jk}\Gamma^k_{is}=\Gamma_{ijs}\qquad\qquad g_{jk}\Gamma^k_{ir}=\Gamma_{ijr}\tag{6.7.522}$$

we get:

$$g_{jk}R^k_{i,rs}=\partial_r\Gamma_{ijs}-\partial_s\Gamma_{ijr}-\Gamma^l_{is}\Gamma_{rlj}+\Gamma^l_{ir}\Gamma_{slj}\tag{6.7.523}$$

and we also proved that:

$$\Gamma_{kji}=\frac{1}{2}\left(\partial_k g_{ij}+\partial_i g_{jk}-\partial_j g_{ki}\right)\tag{6.7.524}$$

Hence:

$$\partial_r\Gamma_{ijs}=\frac{1}{2}\left(\partial_{ri}g_{sj}+\partial_{rs}g_{ji}-\partial_{rj}g_{is}\right)\tag{6.7.525}$$

And by reporting them into the prior-previous relation, we get:

$$\begin{aligned}R_{ij,rs}=g_{jk}R^k_{i,rs}&=\frac{1}{2}\left(\partial_{ri}g_{sj}+\cancel{\partial_{rs}g_{ji}}-\partial_{rj}g_{is}\right)-\frac{1}{2}\left(\partial_{si}g_{rj}+\cancel{\partial_{sr}g_{ji}}-\partial_{sj}g_{ir}\right)\\-\Gamma^l_{is}\Gamma_{rlj}+\Gamma^l_{ir}\Gamma_{slj}&=\frac{1}{2}\left(\partial_{ri}g_{sj}+\partial_{sj}g_{ir}-\partial_{rj}g_{is}-\partial_{si}g_{rj}\right)-\Gamma^l_{is}\Gamma_{rlj}+\Gamma^l_{ir}\Gamma_{slj}\end{aligned}\tag{6.7.526}$$

Finally, we get for the covariant expression of the Riemann-Christoffel tensor:

$$\boxed{R_{ij,rs}=\frac{1}{2}\left(\partial_{ri}g_{sj}+\partial_{sj}g_{ir}-\partial_{rj}g_{is}-\partial_{si}g_{rj}\right)-\Gamma^l_{is}\Gamma_{rlj}+\Gamma^l_{ir}\Gamma_{slj}}\tag{6.7.527}$$

It should be noticed that the Riemann-Christoffel tensor is therefore antisymmetric:

$$\boxed{R_{ij,rs}=-R_{ij,sr}}\tag{6.7.528}$$

and that in the parenthesis of the prior-previous relation we have only double partial derivatives, while outside of the parenthesis the Christoffel symbols contain only simple partial derivatives!

Finally, the permutation of the indices ij and rs as a block gives us as a consequence of the symmetry of the g_{ij} and by inverting their derivation order:

$$\boxed{R_{ij,rs}=R_{rs,ij}} \tag{6.7.529}$$

Let us now perform a circular permutation on the indices j, r, s in the expression (obtained just above)

$$R_{ij,rs}=\frac{1}{2}\left(\partial_{ri}g_{sj}+\partial_{sj}g_{ir}-\partial_{rj}g_{is}-\partial_{si}g_{rj}\right)-\Gamma_{is}^{l}\Gamma_{rlj}+\Gamma_{ir}^{l}\Gamma_{slj} \tag{6.7.530}$$

then we get:

$$\begin{aligned}R_{ir,sj}&=\frac{1}{2}\left(\partial_{jr}g_{is}+\partial_{is}g_{jr}-\partial_{rs}g_{ij}-\partial_{ij}g_{rs}\right)-\Gamma_{ij}^{l}\Gamma_{slr}+\Gamma_{is}^{l}\Gamma_{jlr}\\R_{is,jr}&=\frac{1}{2}\left(\partial_{rs}g_{ij}+\partial_{ij}g_{rs}-\partial_{js}g_{ir}-\partial_{ir}g_{js}\right)-\Gamma_{ir}^{l}\Gamma_{jls}+\Gamma_{ij}^{l}\Gamma_{rls}\end{aligned} \tag{6.7.531}$$

and we get (it is very simple to control by summing the three lines above):

$$\boxed{R_{ij,rs}+R_{ir,sj}+R_{is,jr}=0} \tag{6.7.532}$$

The previous identity is named the "first Bianchi identity" or also "Bianchi algebraic identity" and it highlights the cyclicity property of the Riemann-Christoffel tensor. In reality, we should not use the word "identity" since it is verified only (at least to my knowledge) in the case of a symmetric metric tensor (otherwise, the torsion is not zero for recall!).

The reader will observe that it is immediate that this last relation is satisfied in the case of the Minkowski metric, since if at all points the partial derivative of the metric is zero we have:

$$\partial_{\sigma}g_{\mu\nu}=0\Rightarrow\Gamma_{\mu\nu}^{\rho}=0\Rightarrow R_{\sigma\mu,\nu}^{\rho}=0 \tag{6.7.533}$$

And we will see in the section of General Relativity that this first identity will serve as a basis for the construction of the Schwarzschild metric.

If the metric is of the Minkowski type (we change the notations of the indices to be more in conformity with the usual writings in general relativity) then it is immediate that we also have:

$$\boxed{\partial_{\lambda}R_{\alpha\beta,\mu\nu}+\partial_{\nu}R_{\alpha\beta,\lambda\mu}+\partial_{\mu}R_{\alpha\beta,\nu\lambda}=0} \tag{6.7.534}$$

But in the case where the metric is not of the Minkowski type, this latter relation can be satisfied and has an interest only if and only if the chosen metric is decomposable in Taylor series whose first partial derivatives are zero at 0 (see the section of Differential Geometry for such Taylor series!).

This relation is valid only in the case of a "locally inertial frame (LIF)" in which all Christoffel symbols cancel each other but not their derivatives.

By extension:

$$\boxed{\nabla_{\lambda}R_{\alpha\beta,\mu\nu}+\nabla_{\nu}R_{\alpha\beta,\lambda\mu}+\nabla_{\mu}R_{\alpha\beta,\nu\lambda}=0} \tag{6.7.535}$$

Let us recall that implicitly, this relation, named "Bianchi second identity" or "Bianchi differential identity", always expresses simply (if one may say ...) the fact that gravity and geometry of space are linked together.

Following a reader request let us detail much more how to derivative this identity!

The identity is easiest to derive at the origin of a locally inertial frame (LIF) as already mention, where the first derivatives of the metric tensor, and thus the Christoffel symbols, are all zero. At this point, we have:

$$R_{nj\ell m}=\frac{1}{2}\left(\partial_\ell\partial_j g_{mn}+\partial_m\partial_n g_{j\ell}-\partial_\ell\partial_n g_{jm}-\partial_m\partial_j g_{\ell n}\right) \tag{6.7.536}$$

If the Christoffel symbols are all zero, then the covariant derivative becomes the ordinary derivative:

$$\nabla_j A^k\equiv\partial_j A^k+A^i\Gamma_{ij}^k=\partial_j A^k \tag{6.7.537}$$

Therefore, we get, at the origin of a LIF:

$$\begin{aligned}\nabla_k R_{nj\ell m}&=\partial_k R_{nj\ell m}\\&=\frac{1}{2}\left(\partial_k\partial_\ell\partial_j g_{mn}+\partial_k\partial_m\partial_n g_{j\ell}-\partial_k\partial_\ell\partial_n g_{jm}-\partial_k\partial_m\partial_j g_{\ell n}\right)\end{aligned} \tag{6.7.538}$$

By cyclically permuting the index of the derivative with the last two indices of the tensor, we get:

$$\begin{aligned}\nabla_\ell R_{njmk}&=\partial_\ell R_{njmk}\\&=\frac{1}{2}\left(\partial_\ell\partial_m\partial_j g_{kn}+\partial_\ell\partial_k\partial_n g_{jm}-\partial_\ell\partial_m\partial_n g_{jk}-\partial_\ell\partial_k\partial_j g_{mn}\right)\\\nabla_m R_{njk\ell}&=\partial_m R_{njk\ell}\\&=\frac{1}{2}\left(\partial_m\partial_k\partial_j g_{\ell n}+\partial_m\partial_\ell\partial_n g_{jk}-\partial_m\partial_k\partial_n g_{j\ell}-\partial_m\partial_\ell\partial_j g_{kn}\right)\end{aligned} \tag{6.7.539}$$

By adding up all terms with the covariant derivative and using the commutativity of partial derivatives, we see that all terms cancel in pairs, so we get:

$$\nabla_k R_{nj\ell m}+\nabla_\ell R_{njmk}+\nabla_m R_{njk\ell}=0 \tag{6.7.540}$$

As usual we can use the argument that since we can set up a LIF with its origin at any non-singular point in space-time, this equation is true everywhere and since the covariant derivative is a tensor, this is a tensor equation and is thus valid in all coordinate systems.

6.7.13 Ricci curvature (Ricci tensor)

Before we can see the consequences of second Bianchi's identity, we need to define the "Ricci tensor":

$$R_{\alpha\beta}=R_{\alpha\mu\beta}^{\mu} \tag{6.7.541}$$

which is therefore simply the contraction of the first and third indices of the Riemann-Christoffel tensor which we have given above:

$$R^l_{i,jk}=\partial_j\Gamma^l_{ik}-\partial_k\Gamma^l_{ij}+\Gamma^r_{ik}\Gamma^l_{jr}-\Gamma^r_{ij}\Gamma^l_{kr} \tag{6.7.542}$$

in other words it is just a more condensed notation ... and then the letters for the upper or lower indices as well as the presence of the comma are at the free choice of the writer (according to the mood and especially if the context makes it possible to avoid any confusion).

For example with the Riemann-Christoffel tensor we have just given, the Ricci tensor could be written in the following two ways (we keep the indices with latin letters):

$$\begin{gathered}R_{ik}=R^l_{i,jk}\\R_{ik}=R^l_{ijk}\end{gathered} \tag{6.7.543}$$

The Ricci tensor can be taken as the trace of the Riemann tensor, hence it is of lower rank, and has fewer components. If you have a small geodesic ball in free fall, then (ignoring shear and vorticity) the Ricci tensor tells you the rate at which the volume of that ball begins to change, whereas the Riemann tensor contains information not only about its volume, but also about its shape.

Other contractions of other indices may also be possible but because $R_{\alpha\beta,\mu\nu}$ is antisymmetric on α,β and μ,ν then the contraction on these indices are equivalent to have $\pm R_{\alpha\beta}$.

Remark

There is no widely accepted convention for the sign of the Riemann curvature tensor, or the Ricci tensor, so check the sign conventions of whatever book you are reading!

Similarly, we define the "Ricci scalar" (also sometimes named "Riemann scalar") by the relation:

$$\boxed{R=g^{\alpha\beta}R_{\alpha\beta}=g^{\alpha\beta}g^{\mu\upsilon}R_{\mu\alpha\upsilon\beta}} \tag{6.7.544}$$

which has the following properties:

- If the space is flat, the Ricci scalar is zero
- If space is curved like a sphere, the Ricci scalar is positive
- If the space is curved like a horse saddle, the Ricci scalar is negative

Remark

It is not necessarily obvious at first sight without computing the Ricci scalar to know whether a given metric is curved of flat. We can take a perfectly flat space-time and express it in some bizarre coordinate system, in which the metric has non-constant off-diagonal terms!

Either explicitly (by changing the notation for the indices in order to insist on the fact that it has no impact!):

$$\begin{aligned} R=g^{ij}R_{ij}=g^{ij}R^{\mu}_{i,\mu j}&=g^{ij}\left(\partial_\mu\Gamma^{\mu}_{ij}-\partial_j\Gamma^{\mu}_{i\mu}+\Gamma^{r}_{ij}\Gamma^{\mu}_{\mu\sigma}-\Gamma^{r}_{i\mu}\Gamma^{\mu}_{jr}\right)\\ &=g^{ij}\begin{pmatrix}\partial_\mu\left(\dfrac{1}{2}g^{\mu l}(\partial_i g_{lj}+\partial_j g_{il}-\partial_l g_{ij})\right)-\partial_j\left(\dfrac{1}{2}g^{\mu l}(\partial_i g_{l\mu}+\partial_\mu g_{il}-\partial_l g_{i\mu})\right)\\ +\left(\dfrac{1}{2}g^{rl}(\partial_i g_{lj}+\partial_j g_{il}-\partial_l g_{ij})\right)\left(\dfrac{1}{2}g^{\mu l}(\partial_\mu g_{lr}+\partial_r g_{\mu l}-\partial_l g_{\mu r})\right)\\ -\left(\dfrac{1}{2}g^{rl}(\partial_i g_{l\mu}+\partial_\mu g_{il}-\partial_l g_{i\mu})\right)\left(\dfrac{1}{2}g^{\mu l}(\partial_j g_{lr}+\partial_r g_{jl}-\partial_l g_{jr})\right)\end{pmatrix}\end{aligned} \tag{6.7.545}$$

The Ricci scalar is the trace of the Ricci tensor, and it is a measure of scalar curvature. It can be taken as a way to quantify how the volume of a small geodesic ball (or alternatively its surface area) is different from that of a reference ball in flat space.

We will have concrete practical examples in the section of General Relativity for the first two cases, but let us look at simplified examples for the first two cases (we will not, however, prove the reciprocal).

✍**Examples:**

E1. Let us start with the metric of flat space (without the temporal component). We have (see section General Relativity page 5731):

$$g_{ij}=\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix} \tag{6.7.546}$$

By taking the definition of Ricci's scalar in explicit form:

$$R=g^{ij}\begin{pmatrix}\partial_\mu\left(\dfrac{1}{2}g^{\mu l}(\partial_i g_{lj}+\partial_j g_{il}-\partial_l g_{ij})\right)-\partial_j\left(\dfrac{1}{2}g^{\mu l}(\partial_i g_{l\mu}+\partial_\mu g_{il}-\partial_l g_{i\mu})\right)\\ +\left(\dfrac{1}{2}g^{rl}(\partial_i g_{lj}+\partial_j g_{il}-\partial_l g_{ij})\right)\left(\dfrac{1}{2}g^{\mu l}(\partial_\mu g_{lr}+\partial_r g_{\mu l}-\partial_l g_{\mu r})\right)\\ -\left(\dfrac{1}{2}g^{rl}(\partial_i g_{l\mu}+\partial_\mu g_{il}-\partial_l g_{i\mu})\right)\left(\dfrac{1}{2}g^{\mu l}(\partial_j g_{lr}+\partial_r g_{jl}-\partial_l g_{jr})\right)\end{pmatrix} \tag{6.7.547}$$

It is immediate that R is zero since the partial derivatives will all be zero. So a flat space has a null Ricci scalar.

E2. Let us now look at the metric of the plane expressed in spherical coordinates (without the temporal component). We have (see section General Relativity page 5732):

$$g_{ij}=\begin{bmatrix}1&0&0\\0&r^2&0\\0&0&r^2\sin^2(\theta)\end{bmatrix} \tag{6.7.548}$$

and:

$$g^{ij}=\begin{bmatrix}1 & 0 & 0\\ 0 & \dfrac{1}{r^2} & 0\\ 0 & 0 & \dfrac{r^2}{\sin^2(\theta)}\end{bmatrix} \tag{6.7.549}$$

with:

$$\mathrm{d}s^2=\mathrm{d}r^2+r^2\mathrm{d}\theta^2+r^2\sin^2(\theta)\mathrm{d}\phi^2 \tag{6.7.550}$$

We know that to compute the Ricci scalar (or Ricci curvature), we must therefore compute the contraction of the Riemann-Christoffel tensor (that is, the Ricci tensor) which itself depends on the Christoffel symbols of the second kind which themselves depend on the symbols of Christoffel of the first kind (argh!). We should therefore begin with the lowest level, that is to say by determining all the Christoffel symbols of the first kind given for recall by:

$$\Gamma_{kji}=\frac{1}{2}(\partial_k g_{ij}+\partial_i g_{jk}-\partial_j g_{ki}) \tag{6.7.551}$$

So we have 3^3 that is 27 possible Christoffel symbols of the first kind! Even if some symbols are equal (we have proved it!), We will still calculate everything.

Let's start with joy and good humour...:

$$\begin{aligned}
\Gamma_{111} &= \frac{1}{2}\left(\partial_1 g_{11} + \partial_1 g_{11} - \partial_1 g_{11}\right) = \frac{1}{2}\partial_1 g_{11} = 0\\
\Gamma_{112} &= \frac{1}{2}\left(\partial_1 g_{21} + \partial_2 g_{11} - \partial_1 g_{12}\right) = \frac{1}{2}\partial_2 g_{11} = 0\\
\Gamma_{113} &= \frac{1}{2}\left(\partial_1 g_{31} + \partial_3 g_{11} - \partial_1 g_{13}\right) = \frac{1}{2}\partial_3 g_{11} = 0\\[1em]
\Gamma_{121} &= \frac{1}{2}\left(\partial_1 g_{12} + \partial_1 g_{21} - \partial_2 g_{11}\right) = -\frac{1}{2}\partial_2 g_{11} = 0\\
\Gamma_{122} &= \frac{1}{2}\left(\partial_1 g_{22} + \partial_2 g_{21} - \partial_2 g_{12}\right) = \frac{1}{2}\partial_1 g_{22} = r\\
\Gamma_{123} &= \frac{1}{2}\left(\partial_1 g_{32} + \partial_3 g_{21} - \partial_2 g_{13}\right) = 0\\[1em]
\Gamma_{131} &= \frac{1}{2}\left(\partial_1 g_{13} + \partial_1 g_{31} - \partial_3 g_{11}\right) = -\frac{1}{2}\partial_3 g_{11} = 0\\
\Gamma_{132} &= \frac{1}{2}\left(\partial_1 g_{23} + \partial_2 g_{31} - \partial_3 g_{12}\right) = 0\\
\Gamma_{133} &= \frac{1}{2}\left(\partial_1 g_{33} + \partial_3 g_{31} - \partial_3 g_{13}\right) = \frac{1}{2}\partial_1 g_{33} = r\sin^2(\theta)\\[1em]
\Gamma_{211} &= \frac{1}{2}\left(\partial_2 g_{11} + \partial_1 g_{12} - \partial_1 g_{21}\right) = \frac{1}{2}\partial_2 g_{11} = 0\\
\Gamma_{212} &= \frac{1}{2}\left(\partial_2 g_{21} + \partial_2 g_{12} - \partial_1 g_{22}\right) = -\frac{1}{2}\partial_1 g_{22} = -r\\
\Gamma_{213} &= \frac{1}{2}\left(\partial_2 g_{13} + \partial_3 g_{12} - \partial_1 g_{23}\right) = 0\\[1em]
\Gamma_{221} &= \frac{1}{2}\left(\partial_2 g_{12} + \partial_1 g_{22} - \partial_2 g_{21}\right) = \frac{1}{2}\partial_1 g_{22} = r\\
\Gamma_{222} &= \frac{1}{2}\left(\partial_2 g_{22} + \partial_2 g_{22} - \partial_2 g_{22}\right) = \frac{1}{2}\partial_2 g_{22} = 0\\
\Gamma_{223} &= \frac{1}{2}\left(\partial_2 g_{32} + \partial_3 g_{22} - \partial_2 g_{23}\right) = \frac{1}{2}\partial_3 g_{22} = 0\\[1em]
\Gamma_{231} &= \frac{1}{2}\left(\partial_2 g_{13} + \partial_1 g_{32} - \partial_3 g_{21}\right) = 0\\
\Gamma_{232} &= \frac{1}{2}\left(\partial_2 g_{23} + \partial_2 g_{32} - \partial_3 g_{22}\right) = -\frac{1}{2}\partial_3 g_{22} = 0\\
\Gamma_{233} &= \frac{1}{2}\left(\partial_2 g_{33} + \partial_3 g_{32} - \partial_3 g_{23}\right) = \frac{1}{2}\partial_2 g_{33} = r^2\sin(\theta)\cos(\theta)
\end{aligned}\tag{6.7.552}$$

$$\begin{aligned}
\Gamma_{311} &= \frac{1}{2}\left(\partial_3 g_{11}+\partial_1 g_{13}-\partial_1 g_{31}\right)=\frac{1}{2}\partial_3 g_{11}=0\\
\Gamma_{312} &= \frac{1}{2}\left(\partial_3 g_{21}+\partial_2 g_{13}-\partial_1 g_{32}\right)=0\\
\Gamma_{313} &= \frac{1}{2}\left(\partial_3 g_{31}+\partial_3 g_{13}-\partial_1 g_{33}\right)=-\frac{1}{2}\partial_1 g_{33}=-r\sin^2(\theta)\\
\\
\Gamma_{321} &= \frac{1}{2}\left(\partial_3 g_{12}+\partial_1 g_{23}-\partial_2 g_{31}\right)=0\\
\Gamma_{322} &= \frac{1}{2}\left(\partial_3 g_{22}+\partial_2 g_{23}-\partial_2 g_{32}\right)=\frac{1}{2}\partial_3 g_{22}=0\\
\Gamma_{323} &= \frac{1}{2}\left(\partial_3 g_{32}+\partial_3 g_{23}-\partial_2 g_{33}\right)=-\frac{1}{2}\partial_2 g_{33}=-r^2\sin(\theta)\cos(\theta)\\
\\
\Gamma_{331} &= \frac{1}{2}\left(\partial_3 g_{13}+\partial_1 g_{33}-\partial_3 g_{31}\right)=\frac{1}{2}\partial_1 g_{33}=r\sin^2(\theta)\\
\Gamma_{332} &= \frac{1}{2}\left(\partial_3 g_{23}+\partial_2 g_{33}-\partial_3 g_{32}\right)=\frac{1}{2}\partial_2 g_{33}=r^2\sin(\theta)\cos(\theta)\\
\Gamma_{333} &= \frac{1}{2}\left(\partial_3 g_{33}+\partial_3 g_{33}-\partial_3 g_{33}\right)=\frac{1}{2}\partial_3 g_{33}=0
\end{aligned} \tag{6.7.553}$$

Let us now calculate all the symbols of Christoffel symbols of the second in the details:

$$\Gamma_{kj}^{i}=g^{il}\Gamma_{klj}=\frac{1}{2}g^{il}\left(\partial_k g_{lj}+\partial_j g_{kl}-\partial_l g_{kj}\right) \tag{6.7.554}$$

Again, as the metric tensor is diagonal, this will simplify the calculations!

We have then:

$$\begin{aligned}
\Gamma_{12}^{2} &= g^{22}\Gamma_{122}=\frac{1}{r^2}\cdot r=\frac{1}{r}\\
\Gamma_{13}^{3} &= g^{33}\Gamma_{133}=\frac{1}{r^2\sin^2(\theta)}\cdot r\sin^2(\theta)=\frac{1}{r}\\
\Gamma_{22}^{1} &= g^{11}\Gamma_{212}=1\cdot(-r)=-r\\
\Gamma_{21}^{2} &= g^{22}\Gamma_{221}=\frac{1}{r^2}\cdot r=\frac{1}{r}\\
\Gamma_{23}^{3} &= g^{33}\Gamma_{233}=\frac{1}{r^2\sin^2(\theta)}\cdot r^2\sin(\theta)\cos(\theta)=\frac{\cos(\theta)}{\sin(\theta)}=\frac{1}{\tan(\theta)}=\cot(\theta)\\
\Gamma_{33}^{1} &= g^{11}\Gamma_{313}=1\cdot(-r\sin^2(\theta))=-r\sin^2(\theta)\\
\Gamma_{33}^{2} &= g^{22}\Gamma_{323}=\frac{1}{r^2}\cdot(-r^2\sin(\theta)\cos(\theta))=-\sin(\theta)\cos(\theta)\\
\Gamma_{31}^{3} &= g^{33}\Gamma_{331}=\frac{1}{r^2\sin^2(\theta)}\cdot r\sin^2(\theta)=\frac{1}{r}\\
\Gamma_{32}^{3} &= g^{33}\Gamma_{332}=\frac{1}{r^2\sin^2(\theta)}\cdot r^2\sin(\theta)\cos(\theta)=\frac{\cos(\theta)}{\sin(\theta)}=\frac{1}{\tan(\theta)}=\cot(\theta)
\end{aligned} \tag{6.7.555}$$

Let us now calculate the 9 components of the Ricci tensor in details according to:

$$R_{\alpha\beta}=R_{\alpha\mu\beta}^{\mu}=\partial_\mu\Gamma_{\alpha\beta}^{\mu}-\partial_\beta\Gamma_{\alpha\beta}^{\mu}+\Gamma_{\alpha\beta}^{r}\Gamma_{\mu\sigma}^{\mu}-\Gamma_{\alpha\mu}^{r}\Gamma_{\beta r}^{\mu} \tag{6.7.556}$$

We then have (we calculate them all, even if we know that subsequently those which do not have $\alpha=\beta$ will be useless by the fact that the metric is diagonal):

$$\begin{aligned}
R_{11}=R^{\mu}_{1\mu 1}&=\partial_{\mu}\Gamma^{\mu}_{11}-\partial_1\Gamma^{\mu}_{1\mu}+\Gamma^{r}_{11}\Gamma^{\mu}_{\mu\sigma}-\Gamma^{r}_{1\mu}\Gamma^{\mu}_{1r}=-\partial_1\Gamma^{\mu}_{1\mu}-\Gamma^{r}_{1\mu}\Gamma^{\mu}_{1r}\\
&=-\left(\partial_1\Gamma^{1}_{11}+\partial_1\Gamma^{2}_{12}+\partial_1\Gamma^{3}_{13}\right)-\left(\Gamma^{r}_{11}\Gamma^{1}_{1r}+\Gamma^{r}_{12}\Gamma^{2}_{1r}+\Gamma^{r}_{13}\Gamma^{3}_{1r}\right)\\
&=-\left(\partial_1\Gamma^{2}_{12}+\partial_1\Gamma^{3}_{13}\right)-\left(\Gamma^{r}_{12}\Gamma^{2}_{1r}+\Gamma^{r}_{13}\Gamma^{3}_{1r}\right)\\
&=-\left(\partial_1\Gamma^{2}_{12}+\partial_1\Gamma^{3}_{13}\right)-\left((\Gamma^{1}_{12}\Gamma^{2}_{11}+\Gamma^{2}_{12}\Gamma^{2}_{12}+\Gamma^{3}_{12}\Gamma^{2}_{13})+\Gamma^{r}_{13}\Gamma^{3}_{1r}\right)\\
&=-\left(\partial_1\Gamma^{2}_{12}+\partial_1\Gamma^{3}_{13}\right)-\left((\Gamma^{2}_{12}\Gamma^{2}_{12})+\Gamma^{r}_{13}\Gamma^{3}_{1r}\right)\\
&=-\left(\partial_1\Gamma^{2}_{12}+\partial_1\Gamma^{3}_{13}\right)-\left(\Gamma^{2}_{12}\Gamma^{2}_{12}+(\Gamma^{1}_{13}\Gamma^{3}_{11}+\Gamma^{2}_{13}\Gamma^{3}_{12}+\Gamma^{3}_{13}\Gamma^{3}_{13})\right)\\
&=-\left(\partial_1\Gamma^{2}_{12}+\partial_1\Gamma^{3}_{13}\right)-\left(\Gamma^{2}_{12}\Gamma^{2}_{12}+\Gamma^{3}_{13}\Gamma^{3}_{13}\right)=-\left(\partial_r\frac{1}{r}+\partial_r\frac{1}{r}\right)-\left(\frac{1}{r}\frac{1}{r}+\frac{1}{r}\frac{1}{r}\right)\\
&=-\left(-\frac{1}{r^2}-\frac{1}{r^2}\right)-\frac{2}{r^2}=\frac{2}{r^2}-\frac{2}{r^2}=0
\end{aligned}$$

$$\begin{aligned}
R_{12}=R^{\mu}_{1\mu 2}&=\partial_{\mu}\Gamma^{\mu}_{12}-\partial_2\Gamma^{\mu}_{1\mu}+\Gamma^{r}_{12}\Gamma^{\mu}_{\mu r}-\Gamma^{r}_{1\mu}\Gamma^{\mu}_{2r}\\
&=\left(\partial_1\Gamma^{1}_{12}-\partial_2\Gamma^{1}_{11}+\Gamma^{r}_{12}\Gamma^{1}_{1r}-\Gamma^{r}_{11}\Gamma^{1}_{2r}\right)+\left(\partial_2\Gamma^{2}_{12}-\partial_2\Gamma^{2}_{12}+\Gamma^{r}_{12}\Gamma^{2}_{2r}-\Gamma^{r}_{12}\Gamma^{2}_{2r}\right)\\
&+\left(\partial_3\Gamma^{3}_{12}-\partial_2\Gamma^{3}_{13}+\Gamma^{r}_{12}\Gamma^{3}_{3r}-\Gamma^{r}_{13}\Gamma^{3}_{2r}\right)=\left(-\partial_2\Gamma^{3}_{13}+\Gamma^{r}_{12}\Gamma^{3}_{3r}-\Gamma^{r}_{13}\Gamma^{3}_{2r}\right)\\
&=\left(-\partial_2\Gamma^{3}_{13}+\left(\Gamma^{1}_{12}\Gamma^{3}_{31}+\Gamma^{2}_{12}\Gamma^{3}_{32}+\Gamma^{3}_{12}\Gamma^{3}_{33}\right)-\Gamma^{r}_{13}\Gamma^{3}_{2r}\right)\\
&=\left(-\partial_2\Gamma^{3}_{13}+\left(\Gamma^{2}_{12}\Gamma^{3}_{32}\right)-\Gamma^{r}_{13}\Gamma^{3}_{2r}\right)=\left(-\partial_2\Gamma^{3}_{13}+\left(\Gamma^{2}_{12}\Gamma^{3}_{32}\right)-\left(\Gamma^{1}_{13}\Gamma^{3}_{21}+\Gamma^{2}_{13}\Gamma^{3}_{22}+\Gamma^{3}_{13}\Gamma^{3}_{23}\right)\right)\\
&=\left(-\partial_2\Gamma^{3}_{13}+\left(\Gamma^{2}_{12}\Gamma^{3}_{32}\right)-\left(\Gamma^{3}_{13}\Gamma^{3}_{23}\right)\right)=-\partial_2\Gamma^{3}_{13}+\Gamma^{2}_{12}\Gamma^{3}_{32}-\Gamma^{3}_{13}\Gamma^{3}_{23}\\
&=-\partial_{\theta}\frac{1}{r}+\frac{1}{r}\cot(\theta)-\frac{1}{r}\cot(\theta)=0
\end{aligned}\qquad(6.7.557)$$

$$\begin{aligned}
R_{13} &= R^{\mu}_{1\mu 3} = \partial_{\mu}\Gamma^{\mu}_{13} - \partial_3\Gamma^{\mu}_{1\mu} + \Gamma^{r}_{13}\Gamma^{\mu}_{\mu r} - \Gamma^{r}_{1\mu}\Gamma^{\mu}_{3r}\\
&= \left(\partial_1\Gamma^{1}_{13} - \partial_3\Gamma^{1}_{11} + \Gamma^{r}_{13}\Gamma^{1}_{1r} - \Gamma^{r}_{11}\Gamma^{1}_{3r}\right) + \left(\partial_2\Gamma^{2}_{13} - \partial_3\Gamma^{2}_{12} + \Gamma^{r}_{13}\Gamma^{2}_{2r} - \Gamma^{r}_{12}\Gamma^{2}_{3r}\right)\\
&+ \left(\partial_3\Gamma^{3}_{13} - \partial_3\Gamma^{3}_{13} + \Gamma^{r}_{13}\Gamma^{3}_{3r} - \Gamma^{r}_{13}\Gamma^{3}_{3r}\right) = \left(-\partial_3\Gamma^{2}_{12} + \Gamma^{r}_{13}\Gamma^{2}_{2r} - \Gamma^{r}_{12}\Gamma^{2}_{3r}\right)\\
&= \left(-\partial_3\Gamma^{2}_{12} + \left(\Gamma^{1}_{13}\Gamma^{2}_{21} + \Gamma^{2}_{13}\Gamma^{2}_{22} + \Gamma^{3}_{13}\Gamma^{2}_{23}\right) - \Gamma^{r}_{12}\Gamma^{2}_{3r}\right)\\
&= \left(-\partial_3\Gamma^{2}_{12} + \left(\Gamma^{3}_{13}\Gamma^{2}_{23}\right) - \Gamma^{r}_{12}\Gamma^{2}_{3r}\right) = \left(-\partial_3\Gamma^{2}_{12} + \left(\Gamma^{3}_{13}\Gamma^{2}_{23}\right) - \left(\Gamma^{1}_{12}\Gamma^{2}_{31} + \Gamma^{2}_{12}\Gamma^{2}_{32} + \Gamma^{3}_{12}\Gamma^{2}_{33}\right)\right)\\
&= -\partial_3\Gamma^{2}_{12} = -\partial_{\phi}\frac{1}{r} = 0\\
\\
R_{21} &= R^{\mu}_{2\mu 1} = \partial_{\mu}\Gamma^{\mu}_{21} - \partial_1\Gamma^{\mu}_{2\mu} + \Gamma^{r}_{21}\Gamma^{\mu}_{\mu r} - \Gamma^{r}_{2\mu}\Gamma^{\mu}_{1r}\\
&= \left(\partial_1\Gamma^{1}_{21} - \partial_1\Gamma^{1}_{21} + \Gamma^{r}_{21}\Gamma^{1}_{1r} - \Gamma^{r}_{21}\Gamma^{1}_{1r}\right) + \left(\partial_2\Gamma^{2}_{21} - \partial_1\Gamma^{2}_{22} + \Gamma^{r}_{21}\Gamma^{2}_{2r} - \Gamma^{r}_{22}\Gamma^{2}_{1r}\right)\\
&+ \left(\partial_3\Gamma^{3}_{21} - \partial_1\Gamma^{3}_{23} + \Gamma^{r}_{21}\Gamma^{3}_{3r} - \Gamma^{r}_{23}\Gamma^{3}_{1r}\right)\\
&= \left(\partial_2\Gamma^{2}_{21} + \Gamma^{r}_{21}\Gamma^{2}_{2r} - \Gamma^{r}_{22}\Gamma^{2}_{1r}\right) + \left(-\partial_1\Gamma^{3}_{23} + \Gamma^{r}_{21}\Gamma^{3}_{3r} - \Gamma^{r}_{23}\Gamma^{3}_{1r}\right)\\
&= \left(\partial_2\Gamma^{2}_{21} + (\Gamma^{1}_{21}\Gamma^{2}_{21} + \Gamma^{2}_{21}\Gamma^{2}_{22} + \Gamma^{3}_{21}\Gamma^{2}_{23}) - \Gamma^{r}_{22}\Gamma^{2}_{1r}\right) + \left(-\partial_1\Gamma^{3}_{23} + \Gamma^{r}_{21}\Gamma^{3}_{3r} - \Gamma^{r}_{23}\Gamma^{3}_{1r}\right)\\
&= \left(\partial_2\Gamma^{2}_{21} - \Gamma^{r}_{22}\Gamma^{2}_{1r}\right) + \left(-\partial_1\Gamma^{3}_{23} + \Gamma^{r}_{21}\Gamma^{r}_{3r} - \Gamma^{r}_{23}\Gamma^{3}_{1r}\right)\\
&= \left(\partial_2\Gamma^{2}_{21} - (\Gamma^{1}_{22}\Gamma^{2}_{11} + \Gamma^{2}_{22}\Gamma^{2}_{12} + \Gamma^{3}_{22}\Gamma^{2}_{13})\right) + \left(-\partial_1\Gamma^{3}_{23} + \Gamma^{r}_{21}\Gamma^{3}_{3r} - \Gamma^{r}_{23}\Gamma^{3}_{1r}\right)\\
&= \left(\partial_2\Gamma^{2}_{21}\right) + \left(-\partial_1\Gamma^{3}_{23} + \Gamma^{r}_{21}\Gamma^{3}_{3r} - \Gamma^{r}_{23}\Gamma^{3}_{1r}\right)\\
&= \partial_2\Gamma^{2}_{21} + \left(-\partial_1\Gamma^{3}_{23} + \left(\Gamma^{1}_{21}\Gamma^{3}_{31} + \Gamma^{2}_{21}\Gamma^{3}_{32} + \Gamma^{3}_{21}\Gamma^{3}_{33}\right) - \Gamma^{r}_{23}\Gamma^{3}_{1r}\right)\\
&= \partial_2\Gamma^{2}_{21} + \left(-\partial_1\Gamma^{3}_{23} + \left(\Gamma^{2}_{21}\Gamma^{3}_{32}\right) - \Gamma^{r}_{23}\Gamma^{3}_{1r}\right)\\
&= \partial_2\Gamma^{2}_{21} + \left(-\partial_1\Gamma^{3}_{23} + \Gamma^{2}_{21}\Gamma^{3}_{32} - \left(\Gamma^{1}_{23}\Gamma^{3}_{11} + \Gamma^{2}_{23}\Gamma^{3}_{12} + \Gamma^{3}_{23}\Gamma^{3}_{13}\right)\right)\\
&= \partial_2\Gamma^{2}_{21} + \left(-\partial_1\Gamma^{3}_{23} + \Gamma^{2}_{21}\Gamma^{3}_{32} - \left(\Gamma^{3}_{23}\Gamma^{3}_{13}\right)\right) = \partial_2\Gamma^{2}_{21} - \partial_1\Gamma^{3}_{23} + \Gamma^{2}_{21}\Gamma^{3}_{32} - \Gamma^{3}_{23}\Gamma^{3}_{13}\\
&= \partial_{\theta}\frac{1}{r} - \partial_r\cot(\theta) + \frac{1}{r}\cot(\theta) - \frac{1}{r}\cot(\theta) = 0
\end{aligned} \tag{6.7.558}$$

$$\begin{aligned}
R_{22} &= R^{\mu}_{2\mu 2} = \partial_{\mu}\Gamma^{\mu}_{22} - \partial_2\Gamma^{\mu}_{2\mu} + \Gamma^{r}_{22}\Gamma^{\mu}_{\mu\sigma} - \Gamma^{r}_{2\mu}\Gamma^{\mu}_{2r}\\
&= \left(\partial_1\Gamma^{1}_{22} - \partial_2\Gamma^{1}_{21} + \Gamma^{r}_{22}\Gamma^{1}_{1r} - \Gamma^{r}_{21}\Gamma^{1}_{2r}\right) + \left(\partial_2\Gamma^{2}_{22} - \partial_2\Gamma^{2}_{22} + \Gamma^{r}_{22}\Gamma^{2}_{2r} - \Gamma^{r}_{22}\Gamma^{2}_{2r}\right)\\
&+ \left(\partial_3\Gamma^{3}_{22} - \partial_2\Gamma^{3}_{23} + \Gamma^{r}_{22}\Gamma^{3}_{3r} - \Gamma^{r}_{23}\Gamma^{3}_{2r}\right)\\
&= \left(\partial_1\Gamma^{1}_{22} - \Gamma^{r}_{21}\Gamma^{1}_{2r}\right) + \left(-\partial_2\Gamma^{3}_{23} + \Gamma^{r}_{22}\Gamma^{3}_{3r} - \Gamma^{r}_{23}\Gamma^{3}_{2r}\right)\\
&= \left(\partial_1\Gamma^{1}_{22} - \left(\Gamma^{1}_{21}\Gamma^{1}_{21} + \Gamma^{2}_{21}\Gamma^{1}_{22} + \Gamma^{3}_{21}\Gamma^{1}_{23}\right)\right) + \left(-\partial_2\Gamma^{3}_{23} + \Gamma^{r}_{22}\Gamma^{3}_{3r} - \Gamma^{r}_{23}\Gamma^{3}_{2r}\right)\\
&= \left(\partial_1\Gamma^{1}_{22} - \Gamma^{2}_{21}\Gamma^{1}_{22}\right) + \left(-\partial_2\Gamma^{3}_{23} + \Gamma^{r}_{22}\Gamma^{3}_{3r} - \Gamma^{r}_{23}\Gamma^{3}_{2r}\right)\\
&= \left(\partial_1\Gamma^{1}_{22} - \Gamma^{2}_{21}\Gamma^{1}_{22}\right) + \left(-\partial_2\Gamma^{3}_{23} + \left(\Gamma^{1}_{22}\Gamma^{3}_{31} + \Gamma^{2}_{22}\Gamma^{3}_{32} + \Gamma^{3}_{22}\Gamma^{3}_{33}\right) - \Gamma^{r}_{23}\Gamma^{3}_{2r}\right)\\
&= \left(\partial_1\Gamma^{1}_{22} - \Gamma^{2}_{21}\Gamma^{1}_{22}\right) + \left(-\partial_2\Gamma^{3}_{23} + \Gamma^{1}_{22}\Gamma^{3}_{31} - \Gamma^{r}_{23}\Gamma^{3}_{2r}\right)\\
&= \left(\partial_1\Gamma^{1}_{22} - \Gamma^{2}_{21}\Gamma^{1}_{22}\right) + \left(-\partial_2\Gamma^{3}_{23} + \Gamma^{1}_{22}\Gamma^{3}_{31} - \left(\Gamma^{1}_{23}\Gamma^{3}_{21} + \Gamma^{2}_{23}\Gamma^{3}_{22} + \Gamma^{3}_{23}\Gamma^{3}_{23}\right)\right)\\
&= \left(\partial_1\Gamma^{1}_{22} - \Gamma^{2}_{21}\Gamma^{1}_{22}\right) + \left(-\partial_2\Gamma^{3}_{23} + \Gamma^{1}_{22}\Gamma^{3}_{31} - \Gamma^{3}_{23}\Gamma^{3}_{23}\right)\\
&= \partial_1\Gamma^{1}_{22} - \Gamma^{2}_{21}\Gamma^{1}_{22} - \partial_2\Gamma^{3}_{23} + \Gamma^{1}_{22}\Gamma^{3}_{31} - \Gamma^{3}_{23}\Gamma^{3}_{23}\\
&= \partial_r(-r) - \frac{1}{r}\cdot(-r) - \partial_{\theta}\cot(\theta) + (-r)\frac{1}{r} - \cot^2(\theta)\\
&= -1 + 1 - \left(-1 - \cot^2(\theta)\right) - 1 - \cot^2(\theta) = 0
\end{aligned}$$

$$\begin{aligned}
R_{23} &= R^{\mu}_{2\mu 3} = \partial_{\mu}\Gamma^{\mu}_{23} - \partial_3\Gamma^{\mu}_{2\mu} + \Gamma^{r}_{23}\Gamma^{\mu}_{\mu r} - \Gamma^{r}_{2\mu}\Gamma^{\mu}_{3r}\\
&= \left(\partial_1\Gamma^{1}_{23} - \partial_3\Gamma^{1}_{21} + \Gamma^{r}_{23}\Gamma^{1}_{1r} - \Gamma^{r}_{21}\Gamma^{1}_{3r}\right) + \left(\partial_2\Gamma^{2}_{23} - \partial_3\Gamma^{2}_{22} + \Gamma^{r}_{23}\Gamma^{2}_{2r} - \Gamma^{r}_{22}\Gamma^{2}_{3r}\right)\\
&+ \left(\partial_3\Gamma^{3}_{23} - \partial_3\Gamma^{3}_{23} + \Gamma^{r}_{23}\Gamma^{3}_{3r} - \Gamma^{r}_{23}\Gamma^{3}_{3r}\right) = \left(-\Gamma^{r}_{21}\Gamma^{1}_{3r}\right) + \left(\Gamma^{r}_{23}\Gamma^{2}_{2r} - \Gamma^{r}_{22}\Gamma^{2}_{3r}\right)\\
&= -\left(\Gamma^{1}_{21}\Gamma^{1}_{31} + \Gamma^{2}_{21}\Gamma^{1}_{32} + \Gamma^{3}_{21}\Gamma^{1}_{33}\right) + \left(\Gamma^{r}_{23}\Gamma^{2}_{2r} - \Gamma^{r}_{22}\Gamma^{2}_{3r}\right)\\
&= \left(\Gamma^{r}_{23}\Gamma^{2}_{2r} - \Gamma^{r}_{22}\Gamma^{2}_{3r}\right) = \left(\Gamma^{1}_{23}\Gamma^{2}_{21} + \Gamma^{2}_{23}\Gamma^{2}_{22} + \Gamma^{3}_{23}\Gamma^{2}_{23}\right) - \Gamma^{r}_{22}\Gamma^{2}_{3r}\\
&= -\Gamma^{r}_{22}\Gamma^{2}_{3r} = -\left(\Gamma^{1}_{22}\Gamma^{2}_{31} + \Gamma^{2}_{22}\Gamma^{2}_{32} + \Gamma^{3}_{22}\Gamma^{2}_{33}\right) = 0
\end{aligned}$$

(6.7.559)

$$\begin{aligned}
R_{31} &= R^{\mu}_{3\mu 1} = \partial_{\mu}\Gamma^{\mu}_{31} - \partial_{1}\Gamma^{\mu}_{3\mu} + \Gamma^{r}_{31}\Gamma^{\mu}_{\mu r} - \Gamma^{r}_{3\mu}\Gamma^{\mu}_{1r} \\
&= \left(\partial_1\Gamma^1_{13} - \partial_1\Gamma^1_{31} + \Gamma^r_{31}\Gamma^1_{1r} - \Gamma^r_{31}\Gamma^1_{1r}\right) + \left(\partial_2\Gamma^2_{31} - \partial_1\Gamma^2_{32} + \Gamma^r_{31}\Gamma^2_{2r} - \Gamma^r_{32}\Gamma^2_{1r}\right) \\
&+ \left(\partial_3\Gamma^3_{31} - \partial_1\Gamma^3_{33} + \Gamma^r_{31}\Gamma^3_{3r} - \Gamma^r_{33}\Gamma^3_{1r}\right) \\
&= \left(\Gamma^r_{31}\Gamma^2_{2r} - \Gamma^r_{32}\Gamma^2_{1r}\right) + \left(\partial_3\Gamma^3_{31} + \Gamma^r_{31}\Gamma^3_{3r} - \Gamma^r_{33}\Gamma^3_{1r}\right) \\
&= \left(\left(\Gamma^1_{31}\Gamma^2_{21} + \Gamma^2_{31}\Gamma^2_{22} + \Gamma^3_{31}\Gamma^2_{23}\right) - \Gamma^r_{32}\Gamma^2_{1r}\right) + \left(\partial_3\Gamma^3_{31} + \Gamma^r_{31}\Gamma^3_{3r} - \Gamma^r_{33}\Gamma^3_{1r}\right) \\
&= \left(-\Gamma^r_{32}\Gamma^2_{1r}\right) + \left(\partial_3\Gamma^3_{31} + \Gamma^r_{31}\Gamma^3_{3r} - \Gamma^r_{33}\Gamma^3_{1r}\right) \\
&= -\left(\Gamma^1_{32}\Gamma^2_{11} + \Gamma^2_{32}\Gamma^2_{12} + \Gamma^3_{32}\Gamma^2_{13}\right) + \left(\partial_3\Gamma^3_{31} + \Gamma^r_{31}\Gamma^3_{3r} - \Gamma^r_{33}\Gamma^3_{1r}\right) \\
&= \left(\partial_3\Gamma^3_{31} + \Gamma^r_{31}\Gamma^3_{3r} - \Gamma^r_{33}\Gamma^3_{1r}\right) = \left(\partial_3\Gamma^3_{31} + \left(\Gamma^1_{31}\Gamma^3_{31} + \Gamma^2_{31}\Gamma^3_{32} + \Gamma^3_{31}\Gamma^3_{33}\right) - \Gamma^r_{33}\Gamma^3_{1r}\right) \\
&= \partial_3\Gamma^3_{31} - \Gamma^r_{33}\Gamma^3_{1r} = \partial_3\Gamma^3_{31} - \left(\Gamma^1_{33}\Gamma^3_{11} + \Gamma^2_{33}\Gamma^3_{12} + \Gamma^3_{33}\Gamma^3_{13}\right) \\
&= \partial_3\Gamma^3_{31} = \partial_\theta\frac{1}{r} = 0 \\
\\
R_{32} &= R^{\mu}_{3\mu 2} = \partial_{\mu}\Gamma^{\mu}_{32} - \partial_{2}\Gamma^{\mu}_{3\mu} + \Gamma^{r}_{32}\Gamma^{\mu}_{\mu r} - \Gamma^{r}_{3\mu}\Gamma^{\mu}_{2r} \\
&= \left(\partial_1\Gamma^1_{32} - \partial_2\Gamma^1_{31} + \Gamma^r_{32}\Gamma^1_{1r} - \Gamma^r_{31}\Gamma^1_{2r}\right) + \left(\partial_2\Gamma^2_{32} - \partial_2\Gamma^2_{32} + \Gamma^r_{32}\Gamma^2_{2r} - \Gamma^r_{32}\Gamma^2_{2r}\right) \\
&+ \left(\partial_3\Gamma^3_{32} - \partial_2\Gamma^3_{33} + \Gamma^r_{32}\Gamma^3_{3r} - \Gamma^r_{33}\Gamma^3_{2r}\right) \\
&= \left(-\Gamma^r_{31}\Gamma^1_{2r}\right) + \left(\partial_3\Gamma^3_{32} + \Gamma^r_{32}\Gamma^3_{3r} - \Gamma^r_{33}\Gamma^3_{2r}\right) \\
&= -\left(\Gamma^1_{31}\Gamma^1_{21} + \Gamma^2_{31}\Gamma^1_{22} + \Gamma^3_{31}\Gamma^1_{23}\right) - \left(\partial_3\Gamma^3_{32} + \Gamma^r_{32}\Gamma^3_{3r} - \Gamma^r_{33}\Gamma^3_{2r}\right) \\
&= \left(\partial_3\Gamma^3_{32} + \Gamma^r_{32}\Gamma^3_{3r} - \Gamma^r_{33}\Gamma^3_{2r}\right) = \partial_3\Gamma^3_{32} + \left(\Gamma^1_{32}\Gamma^3_{31} + \Gamma^2_{32}\Gamma^3_{32} + \Gamma^3_{32}\Gamma^3_{33}\right) - \Gamma^r_{33}\Gamma^3_{2r} \\
&= \partial_3\Gamma^3_{32} - \Gamma^r_{33}\Gamma^3_{2r} = \partial_3\Gamma^3_{32} - \left(\Gamma^1_{33}\Gamma^3_{21} + \Gamma^2_{33}\Gamma^3_{22} + \Gamma^3_{33}\Gamma^3_{23}\right) \\
&= \partial_3\Gamma^3_{32} = \partial_\phi\cot(\theta) = 0
\end{aligned} \tag{6.7.560}$$

$$
\begin{aligned}
R_{33}=R^{\mu}_{3\mu 3}&=\partial_\mu\Gamma^{\mu}_{33}-\partial_3\Gamma^{\mu}_{3\mu}+\Gamma^{r}_{33}\Gamma^{\mu}_{\mu r}-\Gamma^{r}_{3\mu}\Gamma^{\mu}_{3r}\\
&=\left(\partial_1\Gamma^1_{33}-\partial_3\Gamma^1_{31}+\Gamma^r_{33}\Gamma^1_{1r}-\Gamma^r_{31}\Gamma^1_{3r}\right)+\left(\partial_2\Gamma^2_{33}-\partial_3\Gamma^2_{32}+\Gamma^r_{33}\Gamma^2_{2r}-\Gamma^r_{32}\Gamma^2_{3r}\right)\\
&+\left(\partial_3\Gamma^3_{33}-\partial_3\Gamma^3_{33}+\Gamma^r_{33}\Gamma^3_{3r}-\Gamma^r_{33}\Gamma^3_{3r}\right)\\
&=\left(\partial_1\Gamma^1_{33}-\Gamma^r_{31}\Gamma^1_{3r}\right)+\left(\partial_2\Gamma^2_{33}+\Gamma^r_{33}\Gamma^2_{2r}-\Gamma^r_{32}\Gamma^2_{3r}\right)\\
&=\left(\partial_1\Gamma^1_{33}-\left(\Gamma^1_{31}\Gamma^1_{31}+\Gamma^2_{31}\Gamma^1_{32}+\Gamma^3_{31}\Gamma^1_{33}\right)\right)+\left(\partial_2\Gamma^2_{33}+\Gamma^r_{33}\Gamma^2_{2r}-\Gamma^r_{32}\Gamma^2_{3r}\right)\\
&=\left(\partial_1\Gamma^1_{33}-\Gamma^3_{31}\Gamma^1_{33}\right)+\left(\partial_2\Gamma^2_{33}+\Gamma^r_{33}\Gamma^2_{2r}-\Gamma^r_{32}\Gamma^2_{3r}\right)\\
&=\left(\partial_1\Gamma^1_{33}-\Gamma^3_{31}\Gamma^1_{33}\right)+\left(\partial_2\Gamma^2_{33}+\left(\Gamma^1_{33}\Gamma^2_{21}+\Gamma^2_{33}\Gamma^2_{22}+\Gamma^3_{33}\Gamma^2_{23}\right)-\Gamma^r_{32}\Gamma^2_{3r}\right)\\
&=\left(\partial_1\Gamma^1_{33}-\Gamma^3_{31}\Gamma^1_{33}\right)+\left(\partial_2\Gamma^2_{33}+\Gamma^1_{33}\Gamma^2_{21}-\Gamma^r_{32}\Gamma^2_{3r}\right)\\
&=\left(\partial_1\Gamma^1_{33}-\Gamma^3_{31}\Gamma^1_{33}\right)+\left(\partial_2\Gamma^2_{33}+\Gamma^1_{33}\Gamma^2_{21}-\left(\Gamma^1_{32}\Gamma^2_{31}+\Gamma^2_{32}\Gamma^2_{32}+\Gamma^3_{32}\Gamma^2_{33}\right)\right)\\
&=\left(\partial_1\Gamma^1_{33}-\Gamma^3_{31}\Gamma^1_{33}\right)+\left(\partial_2\Gamma^2_{33}+\Gamma^1_{33}\Gamma^2_{21}-\Gamma^3_{32}\Gamma^2_{33}\right)\\
&=\partial_r\left(-r\sin^2(\theta)\right)-\frac{1}{r}\left(-r\sin^2(\theta)\right)+\partial_\theta\left(-\sin(\theta)\cos(\theta)\right)+\left(-r\sin^2(\theta)\right)\frac{1}{r}-\cot(\theta)\left(-\sin(\theta)\cos(\theta)\right)\\
&=-\sin^2(\theta)+\sin^2(\theta)-\partial_\theta\left(\sin(\theta)\cos(\theta)\right)-\sin^2(\theta)+\cot(\theta)\sin(\theta)\cos(\theta)\\
&=-\left(-\sin^2(\theta)+\cos^2(\theta)\right)-\sin^2(\theta)+\frac{\cos(\theta)}{\sin(\theta)}\sin(\theta)\cos(\theta)\\
&=\left(\sin^2(\theta)-\cos^2(\theta)\right)-\sin^2(\theta)+\cos^2(\theta)=0
\end{aligned}
\tag{6.7.561}
$$

Let us now calculate the Ricci scalar:

$$R=g^{\alpha\beta}R_{\alpha\beta}\tag{6.7.562}$$

we then have:

$$
\begin{aligned}
R=g^{\alpha\beta}R_{\alpha\beta}&=g^{11}R_{11}+g^{22}R_{22}+g^{33}R_{33}\\
&=1\cdot R_{11}+\frac{1}{r^2}R_{22}+\frac{1}{r^2\sin^2(\theta)}R_{33}\\
&=1\cdot 0+\frac{1}{r^2}\cdot 0+\frac{1}{r^2\sin^2(\theta)}\cdot 0=0
\end{aligned}
\tag{6.7.563}
$$

The Ricci scalar is therefore also zero. This result may be surprising, but in reality it is logical since we have only calculated the scalar curvature of a flat space expressed in spherical coordinates.

E3. Let us now impose ourselves the diagonal metric of the 2-sphere surface $\mathcal{S}^2$ (without the temporal component). We have then in agreement with what we have seen in the section of Differential Geometry page 2884:

$$g_{ij}=\begin{bmatrix} r^2 & 0\\ 0 & r^2\sin^2(\theta)\end{bmatrix}\tag{6.7.564}$$

and:

$$g^{ij}=\begin{bmatrix}\dfrac{1}{r^2} & 0\\ 0 & \dfrac{1}{r^2\sin^2(\theta)}\end{bmatrix} \tag{6.7.565}$$

Therefore (see section Differential Geometry page 2884):

$$\mathrm{d}s^2=r^2\mathrm{d}\theta^2+r^2\sin^2(\theta)\mathrm{d}\phi^2 \tag{6.7.566}$$

often written as:

$$\mathrm{d}s^2=r^2\left(\mathrm{d}\theta^2+\sin^2(\theta)\mathrm{d}\phi^2\right)=r^2\mathrm{d}\Omega^2 \tag{6.7.567}$$

where r is a constant!

We shall therefore begin with the lowest level, that is, by determining all the Christoffel symbols of the first kind given for recall by:

$$\Gamma_{kji}=\frac{1}{2}\left(\partial_k g_{ij}+\partial_i g_{jk}-\partial_j g_{ki}\right) \tag{6.7.568}$$

We have therefore 2^3, that is, 8 possible Christoffel symbols of the first kind! Even if some symbols are equal (we have already proved it!), we will still calculate everything:

$$\begin{aligned}
\Gamma_{111}&=\frac{1}{2}\left(\partial_1 g_{11}+\partial_1 g_{11}-\partial_1 g_{11}\right)=\frac{1}{2}\partial_1 g_{11}=0\\
\Gamma_{112}&=\frac{1}{2}\left(\partial_1 g_{21}+\partial_2 g_{11}-\partial_1 g_{12}\right)=\frac{1}{2}\partial_2 g_{11}=0\\
\\
\Gamma_{121}&=\frac{1}{2}\left(\partial_1 g_{12}+\partial_1 g_{21}-\partial_2 g_{11}\right)=-\frac{1}{2}\partial_2 g_{11}=0\\
\Gamma_{122}&=\frac{1}{2}\left(\partial_1 g_{22}+\partial_2 g_{21}-\partial_2 g_{12}\right)=\frac{1}{2}\partial_1 g_{22}=r^2\sin(\theta)\cos(\theta)\\
\\
\Gamma_{211}&=\frac{1}{2}\left(\partial_2 g_{11}+\partial_1 g_{12}-\partial_1 g_{21}\right)=\frac{1}{2}\partial_2 g_{11}=0\\
\Gamma_{212}&=\frac{1}{2}\left(\partial_2 g_{21}+\partial_2 g_{12}-\partial_1 g_{22}\right)=-\frac{1}{2}\partial_1 g_{22}=-r^2\sin(\theta)\cos(\theta)
\end{aligned} \tag{6.7.569}$$

$$\begin{aligned}
\Gamma_{221}&=\frac{1}{2}\left(\partial_2 g_{12}+\partial_1 g_{22}-\partial_2 g_{21}\right)=\frac{1}{2}\partial_1 g_{22}=r^2\sin(\theta)\cos(\theta)\\
\Gamma_{222}&=\frac{1}{2}\left(\partial_2 g_{22}+\partial_2 g_{22}-\partial_2 g_{22}\right)=\frac{1}{2}\partial_2 g_{22}=0
\end{aligned} \tag{6.7.570}$$

Let us now calculate all the Christoffel symbols of the second kind in the details:

$$\Gamma_{kj}^{i}=g^{il}\Gamma_{klj}=\frac{1}{2}g^{il}\left(\partial_k g_{lj}+\partial_j g_{kl}-\partial_l g_{kj}\right) \tag{6.7.571}$$

Again, as the metric tensor is diagonal, this will simplify the calculations!

We have then:

$$\begin{aligned}\Gamma_{12}^{2}&=g^{22}\Gamma_{122}=\frac{1}{r^2\sin^2(\theta)}\cdot r^2\sin(\theta)\cos(\theta)=\cot(\theta)\\ \Gamma_{22}^{1}&=g^{11}\Gamma_{212}=\frac{1}{r^2}\cdot\left(-r^2\sin(\theta)\cos(\theta)\right)=-\sin(\theta)\cos(\theta)\\ \Gamma_{21}^{2}&=g^{22}\Gamma_{221}=\frac{1}{r^2\sin^2(\theta)}\cdot r^2\sin(\theta)\cos(\theta)=\cot(\theta)\end{aligned}\tag{6.7.572}$$

Let us now calculate the 4 components of the Ricci tensor in the details according to:

$$R_{\alpha\beta}=R_{\alpha\mu\beta}^{\mu}=\partial_\mu\Gamma_{\alpha\beta}^{\mu}-\partial_\beta\Gamma_{\alpha\mu}^{\mu}+\Gamma_{\alpha\beta}^{r}\Gamma_{\mu r}^{\mu}-\Gamma_{\alpha\mu}^{r}\Gamma_{\beta r}^{\mu}\tag{6.7.573}$$

We then have (we calculate them all, even if we know that subsequently those which do not have $\alpha=\beta$ will be useless by the fact that the metric is diagonal):

$$\begin{aligned}R_{11}&=R_{1\mu 1}^{\mu}=\partial_\mu\Gamma_{11}^{\mu}-\partial_1\Gamma_{1\mu}^{\mu}+\Gamma_{11}^{r}\Gamma_{\mu r}^{\mu}-\Gamma_{1\mu}^{r}\Gamma_{1r}^{\mu}\\ &=\left(\partial_1\Gamma_{11}^{1}-\partial_1\Gamma_{11}^{1}+\Gamma_{11}^{r}\Gamma_{1r}^{1}-\Gamma_{11}^{r}\Gamma_{1r}^{1}\right)+\left(\partial_2\Gamma_{11}^{2}-\partial_1\Gamma_{12}^{2}+\Gamma_{11}^{r}\Gamma_{2r}^{2}-\Gamma_{12}^{r}\Gamma_{1r}^{2}\right)\\ &=\left(\partial_2\Gamma_{11}^{2}-\partial_1\Gamma_{12}^{2}+\Gamma_{11}^{r}\Gamma_{2r}^{2}-\Gamma_{12}^{r}\Gamma_{1r}^{2}\right)=\left(-\partial_1\Gamma_{12}^{2}-\Gamma_{12}^{r}\Gamma_{1r}^{2}\right)\\ &=\left(-\partial_1\Gamma_{12}^{2}-\Gamma_{12}^{1}\Gamma_{11}^{2}-\Gamma_{12}^{2}\Gamma_{12}^{2}\right)=\left(-\partial_1\Gamma_{12}^{2}-\Gamma_{12}^{2}\Gamma_{12}^{2}\right)\\ &=\left(-\left(-1-\cot^2(\theta)\right)\right)-\cot^2(\theta)=1\\ \\ R_{12}&=R_{1\mu 2}^{\mu}=\partial_\mu\Gamma_{12}^{\mu}-\partial_2\Gamma_{1\mu}^{\mu}+\Gamma_{12}^{r}\Gamma_{\mu r}^{\mu}-\Gamma_{1\mu}^{r}\Gamma_{2r}^{\mu}\\ &=\left(\partial_1\Gamma_{12}^{1}-\partial_2\Gamma_{11}^{1}+\Gamma_{12}^{r}\Gamma_{1r}^{1}-\Gamma_{11}^{r}\Gamma_{2r}^{1}\right)+\left(\partial_2\Gamma_{12}^{2}-\partial_2\Gamma_{12}^{2}+\Gamma_{12}^{r}\Gamma_{2r}^{2}-\Gamma_{12}^{r}\Gamma_{2r}^{2}\right)\\ &=\left(\partial_1\Gamma_{12}^{1}-\partial_2\Gamma_{11}^{1}+\Gamma_{12}^{r}\Gamma_{1r}^{1}-\Gamma_{11}^{r}\Gamma_{2r}^{1}\right)=\Gamma_{12}^{r}\Gamma_{1r}^{1}=\Gamma_{12}^{1}\Gamma_{11}^{1}+\Gamma_{12}^{2}\Gamma_{12}^{1}=0\end{aligned}\tag{6.7.574}$$

$$\begin{aligned}
R_{21}&=R^{\mu}_{2\mu 1}=\partial_{\mu}\Gamma^{\mu}_{21}-\partial_1\Gamma^{\mu}_{2\mu}+\Gamma^{r}_{21}\Gamma^{\mu}_{\mu r}-\Gamma^{r}_{2\mu}\Gamma^{\mu}_{1r}\\
&=\left(\partial_1\Gamma^1_{21}-\partial_1\Gamma^1_{21}+\Gamma^r_{21}\Gamma^1_{1r}-\Gamma^r_{21}\Gamma^1_{1r}\right)+\left(\partial_2\Gamma^2_{21}-\partial_1\Gamma^2_{22}+\Gamma^r_{21}\Gamma^2_{2r}-\Gamma^r_{22}\Gamma^2_{1r}\right)\\
&=\partial_2\Gamma^2_{21}-\partial_1\Gamma^2_{22}+\Gamma^r_{21}\Gamma^2_{2r}-\Gamma^r_{22}\Gamma^2_{1r}=\Gamma^r_{21}\Gamma^2_{2r}-\Gamma^r_{22}\Gamma^2_{1r}\\
&=\left(\Gamma^1_{21}\Gamma^2_{21}+\Gamma^2_{21}\Gamma^2_{22}\right)-\left(\Gamma^1_{22}\Gamma^2_{11}+\Gamma^1_{22}\Gamma^2_{12}\right)=0\\
\\
R_{22}&=R^{\mu}_{2\mu 2}=\partial_{\mu}\Gamma^{\mu}_{22}-\partial_2\Gamma^{\mu}_{2\mu}+\Gamma^{r}_{22}\Gamma^{\mu}_{\mu r}-\Gamma^{r}_{2\mu}\Gamma^{\mu}_{2r}\\
&=\left(\partial_1\Gamma^1_{22}-\partial_2\Gamma^1_{21}+\Gamma^r_{22}\Gamma^1_{1r}-\Gamma^r_{21}\Gamma^1_{2r}\right)+\left(\partial_2\Gamma^2_{22}-\partial_2\Gamma^2_{22}+\Gamma^r_{22}\Gamma^2_{2r}-\Gamma^r_{22}\Gamma^2_{2r}\right)\\
&=\left(\partial_1\Gamma^1_{22}-\partial_2\Gamma^1_{21}+\Gamma^r_{22}\Gamma^1_{1r}-\Gamma^r_{21}\Gamma^1_{2r}\right)\\
&=\left(\partial_1\Gamma^1_{22}+\Gamma^r_{22}\Gamma^1_{1r}-\Gamma^r_{21}\Gamma^1_{2r}\right)=\left(\partial_1\Gamma^1_{22}+\Gamma^1_{22}\Gamma^1_{11}+\Gamma^2_{22}\Gamma^1_{12}-\Gamma^1_{21}\Gamma^1_{21}-\Gamma^2_{21}\Gamma^1_{22}\right)\\
&=\left(\partial_1\Gamma^1_{22}-\Gamma^2_{21}\Gamma^1_{22}\right)=\left(\sin^2(\theta)-\cos^2(\theta)\right)-\cot(\theta)\cdot\left(-\sin(\theta)\cos(\theta)\right)\\
&=\sin^2(\theta)-\cos^2(\theta)+\cot(\theta)\sin(\theta)\cos(\theta)=\sin^2(\theta)-\cos^2(\theta)+\cos(\theta)\cos(\theta)=\sin^2(\theta)
\end{aligned}\tag{6.7.575}$$

Let us now calculate the Ricci scalar:

$$R=g^{\alpha\beta}R_{\alpha\beta}\tag{6.7.576}$$

Remark

Notice that $R_{\alpha\beta}=1/r^2g^{\alpha\beta}$ and for information we can also prove that (its aesthetics but so far it has no practical application in this textbook):

$$R^{\alpha}_{\mu\beta\nu}=\frac{\delta^{\alpha}_{\beta}g_{\mu\nu}-\delta^{\alpha}_{\nu}g_{\mu\beta}}{r^2}\tag{6.7.577}$$

We then have:

$$\begin{aligned}
R&=g^{\alpha\beta}R_{\alpha\beta}=g^{11}R_{11}+g^{22}R_{22}=\frac{1}{r^2}\cdot R_{11}+\frac{1}{r^2\sin^2(\theta)}R_{22}\\
&=\frac{1}{r^2}\cdot 1+\frac{1}{r^2\sin^2(\theta)}\cdot\sin^2(\theta)=\frac{1}{r^2}+\frac{1}{r^2}=\frac{2}{r^2}
\end{aligned}\tag{6.7.578}$$

We notice that:

1. The Ricci scalar is a constant. This means that the hyper-surface has a constant curvature at all points of the surface (we know that the sphere by symmetry has a constant curvature at all points). It thus possesses a form of symmetry, with respect to its curvature. We are then dealing with a "maximally symmetrical variety".

2. This scalar is positive which describes a domed space (ball, sphere) and if $r\rightarrow+\infty$ the curvature is zero!

Remark

Be careful not to confuse the value of the Ricci curvature with that of the Gauss curvature!!!

6.7.14 Einstein Tensor

Let us apply a contraction to the second Bianchi identity (valid for recall with the "+" only if the metric is positive):

$$g^{\alpha\beta}\left(\nabla_\lambda R_{\alpha\beta,\mu\nu}+\nabla_\nu R_{\alpha\beta,\lambda\mu}+\nabla_\mu R_{\alpha\beta,\nu\lambda}\right)=0 \tag{6.7.579}$$

Let us recall that $\nabla_\lambda g_{\alpha\mu}=0$ and similarly by extension that $\Delta_\lambda g^{\alpha\mu}=0$. So finally this leads us to write by the property of the derivatives (product in sum):

$$\nabla_\lambda g^{\alpha\mu}R_{\alpha\beta,\mu\nu}+\nabla_\nu g^{\alpha\mu}R_{\alpha\beta,\lambda\mu}+\nabla_\mu g^{\alpha\beta}R_{\alpha\mu,\nu\lambda}=0 \tag{6.7.580}$$

and therefore to get:

$$\nabla_\lambda R^{\mu}_{\beta,\mu\nu}+\nabla_\nu R^{\mu}_{\beta,\lambda\mu}+\nabla_\mu R^{\mu}_{\beta,\nu\lambda}=0 \tag{6.7.581}$$

Using the antisymmetry property of the Riemann-Christoffel tensor we write:

$$\nabla_\lambda R^{\mu}_{\beta,\mu\nu}-\nabla_\nu R^{\mu}_{\beta,\mu\lambda}+\nabla_\mu R^{\mu}_{\beta,\nu\lambda}=\nabla_\lambda R_{\beta\nu}-\nabla_\nu R^{\mu}_{\beta,\mu\lambda}+\nabla_\mu R^{\mu}_{\beta,\nu\lambda}=0 \tag{6.7.582}$$

What ultimately brings us to write from the definition of Ricci's tensor:

$$\nabla_\lambda R_{\beta\nu}-\nabla_\nu R_{\beta\lambda}+\nabla_\mu R^{\mu}_{\beta,\nu\lambda}=0 \tag{6.7.583}$$

This last relationship being named "contracted Bianchi identity".

Let us contract this relation once more:

$$\begin{aligned}g^{\beta\nu}\left(\nabla_\lambda R_{\beta\nu}-\nabla_\nu R_{\beta\lambda}+\nabla_\mu R^{\mu}_{\beta,\nu\lambda}\right)&=\nabla_\lambda R^{\nu}_{\nu}-\nabla_\nu R^{\nu}_{\lambda}+\nabla_\mu R^{\mu\nu}_{\nu\lambda}=\nabla_\lambda R^{\nu}_{\nu}-\nabla_\nu R^{\nu}_{\lambda}-\nabla_\mu R^{\mu\nu}_{\lambda\nu}\\&=\nabla_\lambda R-\nabla_\nu R^{\nu}_{\lambda}-\nabla_\mu R^{\mu}_{\lambda}=0\end{aligned} \tag{6.7.584}$$

That which is identical to write using the properties of the Einstein summation (which allows to freely change the indices):

$$\nabla_\lambda R-\nabla_\mu R^{\mu}_{\lambda}-\nabla_\mu R^{\mu}_{\lambda}=\nabla_\lambda R-2\nabla_\mu R^{\mu}_{\lambda}=0 \tag{6.7.585}$$

Which is equivalent to:

$$\nabla_\mu R^{\mu}_{\lambda}-\frac{1}{2}\nabla_\lambda R=0 \tag{6.7.586}$$

As $\nabla_\lambda=g^{\mu}_{\lambda}\nabla_\mu R$, we have:

$$\nabla_\mu\left(R^{\mu}_{\lambda}-\frac{1}{2}g^{\mu}_{\lambda}R\right)=0 \tag{6.7.587}$$

By raising the index λ by multiplication with $g^{\nu\mu}$, we get the "Einstein's identity":

$$\nabla_{\mu}\left(R^{\mu\nu}-\frac{1}{2}g^{\mu\nu}R\right)=0 \tag{6.7.588}$$

The "Einstein tensor" (tensor of order two and contravariant in the present case) which is therefore a constant in a given Riemannian space is therefore defined by:

$$\boxed{G^{\mu\nu}=R^{\mu\nu}-\frac{1}{2}g^{\mu\nu}R} \tag{6.7.589}$$

And expresses in the shortest possible way, parallel transport under all assumptions seen so far.

Identically, we can get the covariant form:

$$\boxed{G_{\mu\nu}=R_{\mu\nu}-\frac{1}{2}g_{\mu\nu}R} \tag{6.7.590}$$

The tensor is therefore constructed for a Riemannian metric only (which nevertheless makes a lot of possible spaces ...), and is automatically non-divergent:

$$\nabla_{\mu}G_{\mu\nu}=0 \tag{6.7.591}$$

It must be remembered, however, that a large part of the latest developments consider a symmetrical metric. This is why some speak of "symmetrical gravitational theory" when we deal with General Relativity.

We shall find this tensor naturally in the section of General Relativity when, by making use of the variational principle, we decompose the action into two terms:

- the action of mass in the gravitational field
- the action of the gravitational field in the absence of mass

By expressing the whole in a Riemannian space we will then get the no less famous Einstein field equations (without further explanations in this section):

$$R_{\mu\nu}-\frac{1}{2}Rg_{\mu\nu}=\frac{8\pi G}{c^4}T_{\mu\nu} \tag{6.7.592}$$

the details being given in the section of General Relativity.

Remark

As we see, we can very well add a constant term to the expression of Einstein's tensor, without changing the nullity of its divergence. This fact, used in Astrophysics, makes it possible to construct models of particular Universes that we will deal with in the section of Cosmology.

✍**Example:**

Let us calculate the order 2 covariant Einstein tensor:

$$G_{\mu\nu}=R_{\mu\nu}-\frac{1}{2}g_{\mu\nu}R \tag{6.7.593}$$

based on the diagonal metric surface of the 2-sphere $\mathcal{S}^2$ (without the temporal component):

$$g_{\mu\nu}=\begin{bmatrix} r^2 & 0 \\ 0 & r^2\sin^2(\theta) \end{bmatrix} \tag{6.7.594}$$

Since the metric is diagonal, we have proved earlier above and in detail by the example that:

$$R_{12}=R_{21}=0 \tag{6.7.595}$$

And as in the present case, we also have:

$$g_{12}=g_{21}=0 \tag{6.7.596}$$

It comes:

$$G_{12}=G_{21}=0 \tag{6.7.597}$$

So we have to focus only on two components:

$$\begin{aligned} G_{11}&=R_{11}-\frac{1}{2}g_{11}R=1-\frac{1}{2}r^2\frac{2}{r^2}=1-\frac{1}{2}r^2\frac{2}{r^2}=0 \\ G_{22}&=R_{22}-\frac{1}{2}g_{22}\frac{2}{r^2}=\sin^2(\theta)-\frac{1}{2}r^2\sin^2(\theta)\frac{2}{r^2}=0 \end{aligned} \tag{6.7.598}$$

which confirms what we have said earlier above.

The complexity of the Einstein Tensor expression can be shown using the formula for the Ricci tensor in terms of Christoffel symbols:

$$\begin{aligned} G_{\alpha\beta}&=R_{\alpha\beta}-\frac{1}{2}g_{\alpha\beta}R \\ &=R_{\alpha\beta}-\frac{1}{2}g_{\alpha\beta}g^{\gamma\zeta}R_{\gamma\zeta} \\ &=(\delta_\alpha^\gamma\delta_\beta^\zeta-\frac{1}{2}g_{\alpha\beta}g^{\gamma\zeta})R_{\gamma\zeta} \\ &=(\delta_\alpha^\gamma\delta_\beta^\zeta-\frac{1}{2}g_{\alpha\beta}g^{\gamma\zeta})(\Gamma^{\varepsilon}_{\gamma\zeta,\varepsilon}-\Gamma^{\varepsilon}_{\gamma\varepsilon,\zeta}+\Gamma^{\varepsilon}_{\varepsilon\sigma}\Gamma^{\sigma}_{\gamma\zeta}-\Gamma^{\varepsilon}_{\zeta\sigma}\Gamma^{\sigma}_{\varepsilon\gamma}), \end{aligned} \tag{6.7.599}$$

where δ_β^α is the Kronecker tensor and the Christoffel symbol ${\Gamma^\alpha}_{\beta\gamma}$ is given for recall by:

$${\Gamma^\alpha}_{\beta\gamma}=\frac{1}{2}g^{\alpha\varepsilon}(g_{\beta\varepsilon,\gamma}+g_{\gamma\varepsilon,\beta}-g_{\beta\gamma,\varepsilon}) \tag{6.7.600}$$

Before cancellations, this formula results in $2\times(6+6+9+9)=60$ individual terms. Cancellations bring this number down somewhat.

~95 %

★★★★☆

40 votes, 82.5%

6.8. Spinor Calculus

As we will see first in relativistic quantum physics, spinors play a major role in quantum theory, and therefore in all contemporary physics (quantum field theory, standard model, string theory, etc.). The actual purpose of this section on the spinors is only to give the tools to the reader that are necessary to a deep understanding of what will be done in the chapter about Atomistic.

It was starting 11927 (holocene calendar) that the physicists Pauli, and after Dirac introduced spinors for the representation of the wave functions (see section Relativistic Quantum Physics page 4932). However, in their mathematical form, spinors were discovered by Elie Cartan in 11913 (holocene calendar) during his research on the representations of the groups following the general theory of Clifford spaces (introduced by the mathematician W.K. Clifford in 11876 according to holocene calendar). He showed, as we will see it, that in spinors provide in fact a linear representation of the group of rotations of a space with any number of dimensions. Thus, spinors are closely related to geometry but they are often introduce in an abstract way without intuitive geometric meaning. Thus, we will try (as always on this textbook) in this section to introduce this tool in the most simple and intuitive possible way with a maximum of details.

The spinor formalism is not interest only of interest for quantum physics and its related developments, among others, Roger Penrose showed that the spinor theory was an extremely fruitful approach to the theory of General Relativity. Even if the most commonly used tool for the treatment of General Relativity is the tensor calculus, Penrose seems to have shown that in the specific case of the four-dimensional space and in the metric Lorentz the formalism of two components spinors was more appropriate.

The theory of spinors named, "spinor calculus" or sometimes "spin geometry" is extremely broad but as we know this textbook aims to address the physicists and engineers therefore we will limit ourselves to spinors properties useful in quantum physics (at least actually).

Remark

We strongly recommend the reader to have previously read the subsubsection on quaternions (see section Numbers page 261), the subsubsection on rotations in space (see section Euclidean Geometry page 2647) and finally, if possible, for a physical practical example, the section on Relativistic Quantum Physics page 4919.

6.8.1 Unit Spinor

We will give here a first simplified and special definition (or example) of spinors. Thus, we will show that it is possible from such a tool to represent a vector equation of a space E^3 of three components using a two-component spinor. The method is extremely simple and the reader who has already read the part of the section on Quantum Wave Physics dealing with the Dirac equation and the section on Quantum Computing will see a rather beautiful analogy.

Consider to start the sphere of radius of the following equation (see section Analytical Geometry page 2964):

$$x^2+y^2+z^2=1 \tag{6.8.1}$$

And consider the following figure:

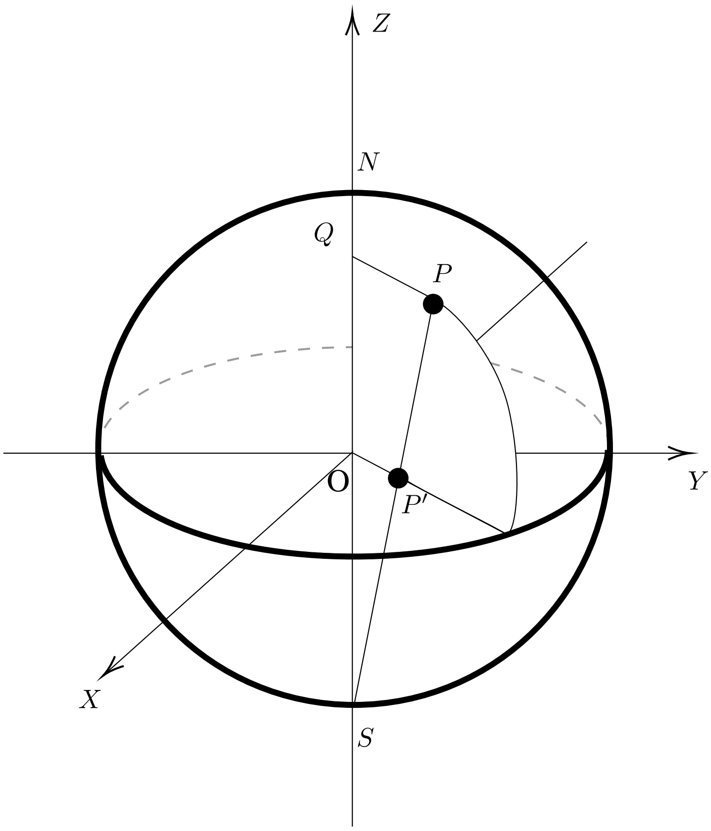


Figure 6.163: **Unit Sphere**

Let us put on this sphere of center on O and unit radius a point P of coordinates (x,y,z) and denote by N (north) and S (south) the points of the sphere intersecting with the Z axis.

The point S will have by convention for coordinates:

$$S=\begin{pmatrix}x\\y\\z\end{pmatrix}=\begin{pmatrix}0\\0\\-1\end{pmatrix} \tag{6.8.2}$$

We obtain a projection so-named "stereographic projection " P' of the point P by tracing the straight line $\overline{SP}$ that passes through the complex equatorial plane $x\text{O}y$ (yes! we chose $\mathbb{C}$ as plane!) at point P' of coordinates (x',y',z').

The similar triangles $SP'\text{O}$ and SPQ (with Q being the orthogonal projection onto the Z-axis of the point P) gives us the following relations by simply applying Thales' theorem (see section Euclidean Geometry page 2610):

$$\triangle SP'\text{O} \sim \triangle SPQ \Leftrightarrow \frac{SP}{SP'} = \frac{x}{x'} = \frac{y}{y'} = \frac{1}{1+z} = \frac{x'+\mathrm{i}y'}{x+\mathrm{i}y} = \frac{x'-\mathrm{i}y}{x-\mathrm{i}y} \tag{6.8.3}$$

Remark

The last two relation (ratios) are obtained by simply applying Thales' theorem (see section Euclidean Geometry) in the complex equatorial plane.

Let us write now to simplify the notations:

$$\xi = x' + \mathrm{i}y' \qquad \bar{\xi} = x' - \mathrm{i}y' \tag{6.8.4}$$

We have from the prior-previous relation that:

$$\frac{1}{1+z} = \frac{x'+\mathrm{i}y'}{x+\mathrm{i}y} \tag{6.8.5}$$

Taking the squared modulus (see the study of complex numbers in the section on Numbers):

$$\frac{1}{|1+z|^2} = \frac{1}{(1+z)^2} = \frac{|x'+\mathrm{i}y'|^2}{|x+\mathrm{i}y|^2} = \frac{\xi\bar{\xi}}{x^2+y^2} \tag{6.8.6}$$

and from the equation of the sphere it follows:

$$x^2+y^2 = 1-z^2 = (1+z)(1-z) \tag{6.8.7}$$

we finally get:

$$\frac{1}{(1+z)^2} = \frac{\xi\bar{\xi}}{x^2+y^2} = \frac{\xi\bar{\xi}}{(1+z)(1-z)} \tag{6.8.8}$$

Let us write now the complex number ξ under the form:

$$\xi = x' + \mathrm{i}y' = \frac{\phi}{\psi} \tag{6.8.9}$$

where ϕ, ψ are two complex numbers that we can always impose verify the following condition of unitarity (nothing prevents us to do that but for theoretical physics purpose this choice suit us well...):

$$\phi\bar{\phi} + \psi\bar{\psi} = 1 \tag{6.8.10}$$

Remark

The following complex numbers satisfy for example (hazard!!??) the above condition:

$$\psi = e^{\mathrm{i}\beta}\cos(\alpha) \qquad \phi = e^{\mathrm{i}\gamma}\sin(\alpha) \tag{6.8.11}$$

Remember before continuing that we have proved in our study of complex numbers (see section Numbers page 247) that:

$$\xi=\frac{\phi}{\psi},\bar{\xi}=\overline{\left(\frac{\phi}{\psi}\right)}=\frac{\bar{\phi}}{\bar{\psi}} \tag{6.8.12}$$

Therefore it comes by injecting the last two relations in the equation given above:

$$\frac{1}{(1+z)^2}=\frac{\dfrac{\phi}{\psi}\dfrac{\bar{\phi}}{\bar{\psi}}}{(1+z)(1-z)} \tag{6.8.13}$$

So we get:

$$\psi\bar{\psi}=\phi\bar{\phi}\frac{1+z}{1-z} \tag{6.8.14}$$

Rearranged:

$$\psi\bar{\psi}(1-z)=\phi\bar{\phi}(1+z) \tag{6.8.15}$$

Therefore:

$$\psi\bar{\psi}-z\psi\bar{\psi}=\phi\bar{\phi}+z\psi\bar{\psi}\Leftrightarrow\psi\bar{\psi}-\phi\bar{\phi}=z\left(\psi\bar{\psi}+\phi\bar{\phi}\right) \tag{6.8.16}$$

Finally we have a simple expression for the z coordinate of the point P because in the last equality above you must remember that $\psi\bar{\psi}+\phi\bar{\phi}=1$ then:

$$z=\psi\bar{\psi}-\phi\bar{\phi} \tag{6.8.17}$$

As we have:

$$x+\mathrm{i}y=\xi(1+z)\quad\text{and}\qquad x-\mathrm{i}y=\bar{\xi}(1+z) \tag{6.8.18}$$

Then by summing and respectively by substracting the two above relations and using previous results we get:

$$\begin{aligned}x&=\frac{1}{2}\left(\xi+\bar{\xi}\right)(1+z)=\frac{1}{2}\left(\frac{\phi}{\psi}+\frac{\bar{\phi}}{\bar{\psi}}\right)((\psi\bar{\psi}+\phi\bar{\phi})+(\psi\bar{\psi}-\phi\bar{\phi}))=\psi\bar{\phi}+\bar{\psi}\phi\\y&=-\frac{\mathrm{i}}{2}\left(\xi-\bar{\xi}\right)(1+z)=\frac{\mathrm{i}}{2}\left(\frac{\phi}{\psi}-\frac{\bar{\phi}}{\bar{\psi}}\right)((\psi\bar{\psi}+\phi\bar{\phi})+(\psi\bar{\psi}-\phi\bar{\phi}))=\mathrm{i}\left(\psi\bar{\phi}-\bar{\psi}\phi\right)\end{aligned} \tag{6.8.19}$$

To resume for the point P we have with our choices:

$$\begin{aligned}x&=\psi\bar{\phi}+\bar{\psi}\phi\\y&=\mathrm{i}\left(\psi\bar{\phi}-\bar{\psi}\phi\right)\\z&=\psi\bar{\psi}-\phi\bar{\phi}\end{aligned} \tag{6.8.20}$$

Thus, for any point P on the sphere of radius unity, we can match a pair of complex numbers satisfying the imposed unitary identity!

Therefore in complete and explicit form we finally have using what we know about complex numbers (see section Numbers page 243) and remarkable trigonometric functions (see section Trigonometry page 2519):

$$\begin{aligned}
\overrightarrow{\mathrm{O}P}=\begin{pmatrix}x\\ y\\ z\end{pmatrix}&=\begin{pmatrix}\psi\bar{\phi}+\bar{\psi}\phi\\ \mathrm{i}\left(\psi\bar{\phi}-\bar{\psi}\phi\right)\\ \psi\bar{\psi}-\phi\bar{\phi}\end{pmatrix}=\begin{pmatrix}e^{\mathrm{i}(\beta-\gamma)}\cos(\alpha)\sin(\alpha)+e^{\mathrm{i}(\gamma-\beta)}\cos(\alpha)\sin(\alpha)\\ \mathrm{i}\left(e^{\mathrm{i}(\beta-\gamma)}\cos(\alpha)\sin(\alpha)-e^{\mathrm{i}(\gamma-\beta)}\cos(\alpha)\sin(\alpha)\right)\\ \cos^2(\alpha)-\sin^2(\alpha)\end{pmatrix}\\
&=\begin{pmatrix}\left(e^{\mathrm{i}(\beta-\gamma)}+e^{-\mathrm{i}(\beta-\gamma)}\right)\cos(\alpha)\sin(\alpha)\\ \mathrm{i}\left(e^{\mathrm{i}(\beta-\gamma)}-e^{-\mathrm{i}(\beta-\gamma)}\right)\cos(\alpha)\sin(\alpha)\\ \cos^2(\alpha)-\sin^2(\alpha)\end{pmatrix}\\
&=\begin{pmatrix}(\cos(\beta-\gamma)+\mathrm{i}\sin(\beta-\gamma)+\cos(-(\beta-\gamma))+\mathrm{i}\sin(-(\beta-\gamma)))\cos(\alpha)\sin(\alpha)\\ \mathrm{i}(\cos(\beta-\gamma)+\mathrm{i}\sin(\beta-\gamma)-\cos(-(\beta-\gamma))-\mathrm{i}\sin(-(\beta-\gamma)))\cos(\alpha)\sin(\alpha)\\ \cos^2(\alpha)-\sin^2(\alpha)\end{pmatrix}\\
&=\begin{pmatrix}(2\cos(\beta-\gamma)+\mathrm{i}\sin(\beta-\gamma)-\mathrm{i}\sin(\beta-\gamma))\cos(\alpha)\sin(\alpha)\\ \mathrm{i}(\mathrm{i}\sin(\beta-\gamma)+\mathrm{i}\sin(\beta-\gamma))\cos(\alpha)\sin(\alpha)\\ \cos^2(\alpha)-\sin^2(\alpha)\end{pmatrix}\\
&=\begin{pmatrix}2\cos(\beta-\gamma)\cos(\alpha)\sin(\alpha)\\ -2\sin(\beta-\gamma)\cos(\alpha)\sin(\alpha)\\ \cos^2(\alpha)-\sin^2(\alpha)\end{pmatrix}=\begin{pmatrix}2\cos(\beta-\gamma)\dfrac{1}{2}\sin(2\alpha)\\ -2\sin(\beta-\gamma)\dfrac{1}{2}\sin(2\alpha)\\ \cos^2(\alpha)-\sin^2(\alpha)\end{pmatrix}\\
&=\begin{pmatrix}\cos(\beta-\gamma)\sin(2\alpha)\\ -\sin(\beta-\gamma)\sin(2\alpha)\\ \cos(2\alpha)\end{pmatrix}=\begin{pmatrix}\cos(\beta-\gamma)\sin(2\alpha)\\ \sin(\gamma-\beta)\sin(2\alpha)\\ \cos(2\alpha)\end{pmatrix}=\begin{pmatrix}\cos(\beta-\gamma)\sin(2\alpha)\\ -\sin(\beta-\gamma)\sin(2\alpha)\\ \cos(2\alpha)\end{pmatrix}
\end{aligned}\tag{6.8.21}$$

We notice easily that the norm of this vector is equal to 1.

This last relation also indicates us that 2α is the angle between $\mathrm{O}z$ and $\overrightarrow{\mathrm{O}P}$ (since the hypotenuse of vector's angle has a unit norm) and therefore by deduction $\gamma-\beta$ represents the angle between $\mathrm{O}x$ and the plane $(\mathrm{O}z,\overrightarrow{\mathrm{O}P})$:

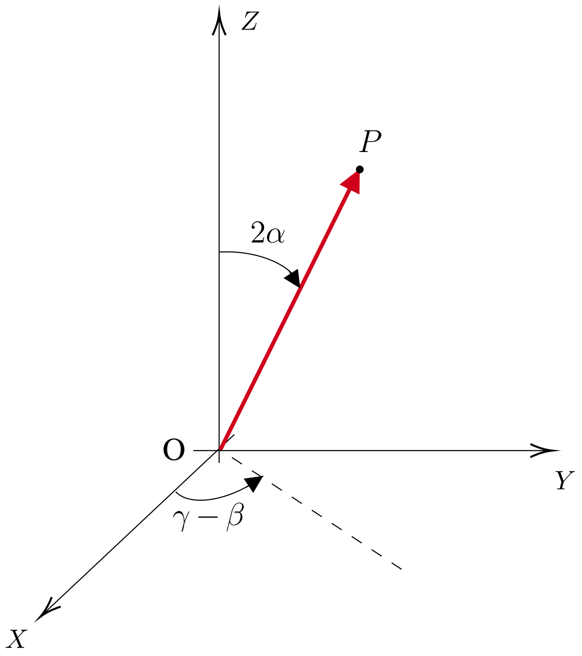


Figure 6.164: **Representation of the rotation**

The pair of complex numbers:

$$x=\psi\bar{\phi}+\bar{\psi}\phi \qquad y=\mathrm{i}\left(\psi\bar{\phi}-\bar{\psi}\phi\right) \tag{6.8.22}$$

is by definition a "unitary spinor" (it contains also all the information about z). Thus, as we have seen, a unitary spinor can also be expressed in the form:

$$\psi=e^{\mathrm{i}\beta}\cos(\alpha) \qquad \phi=e^{\mathrm{i}\gamma}\sin(\alpha) \tag{6.8.23}$$

As well any spinor can be written in the more general form:

$$\psi=\rho e^{\mathrm{i}\beta}\cos(\alpha) \qquad \phi=\rho e^{\mathrm{i}\gamma}\sin(\alpha) \tag{6.8.24}$$

The spin is essentially measured from the oriented z-axis as we have seen it yet with the previous figure.

The stereographic projection led us to represent certain vectors of the Euclidean space $\mathcal{E}^3$ with the elements of a complex two-dimensional vector space that is the "space of spinors".

Remark

This representation is not unique because the arguments of complex number are (in trigonometric form) determined at a given offset constant!

The reader who has already read a little bit the section on Wave Quantum Physics will probably noticed the strange (not innocent) similarity of the following identity and relations:

$$\psi\bar{\psi}+\phi\bar{\phi}=1 \qquad x=\psi\bar{\phi}+\bar{\psi}\phi \qquad y=\mathrm{i}\left(\psi\bar{\phi}-\bar{\psi}\phi\right) \tag{6.8.25}$$

compared to the de Broglie normalization condition (the integral over the entire space of the sum of the products of the complex wave functions and its conjugate is equal to the unit) and to the developments determining the continuity equation also in the section of Wave Quantum Physics .

Let us now see that for future needs, we can find two new vectors $\vec{v}_1, \vec{v}_2$ of Euclidean space $\mathcal{E}^3$ associated with a unitary spinor (ψ, ϕ) determined on the unit sphere. These vectors will have to be orthogonal to each other and with unit norm, each orthogonal to the vector $\overrightarrow{\mathrm{O}P}$.

To simplify the notations let us write $\vec{v}_3 = \overrightarrow{\mathrm{O}P}$ and $\vec{v}_i = (x_i, y_i, z_i), i \in \{1, 2, 3\}$.

The $\vec{v}_1, \vec{v}_2, \vec{v}_3$ vectors are of course bounded by the cross product (see section Vector Calculus page 1830):

$$\vec{v}_3 = \vec{v}_1 \times \vec{v}_2 \tag{6.8.26}$$

hence taking into account the expression of $\overrightarrow{\mathrm{O}P}$ components based on its associated spinor, and the fact that $\psi\bar{\psi} + \phi\bar{\phi} = 1$, we get:

$$\begin{aligned} x_3 &= y_1 z_2 - y_2 z_1 = 1 \cdot (\psi\bar{\phi} + \bar{\psi}\phi) = (\psi\bar{\psi} + \phi\bar{\phi})(\psi\bar{\phi} + \bar{\psi}\phi) \\ y_3 &= z_1 x_2 - z_2 x_1 = 1 \cdot \mathrm{i}(\psi\bar{\phi} - \bar{\psi}\phi) = \mathrm{i}(\psi\bar{\psi} + \phi\bar{\phi})(\psi\bar{\phi} - \bar{\psi}\phi) \\ z_3 &= x_1 y_2 - x_2 y_1 = 1 \cdot (\psi\bar{\psi} - \phi\bar{\phi}) \end{aligned} \tag{6.8.27}$$

Writing the orthogonality of vectors we get them obviously six additional equations. However the orientation of vectors $\vec{v}_1, \vec{v}_2$ being not fixed, there is some uncertainty in the values of their components. Let us select values such that:

$$\begin{aligned} x_1 + \mathrm{i}x_2 &= \psi^2 - \phi^2 = e^{2\mathrm{i}\beta}\cos^2(\alpha) - e^{2\mathrm{i}\gamma}\sin^2(\alpha) \\ y_1 + \mathrm{i}y_2 &= \mathrm{i}(\psi^2 + \phi^2) = \mathrm{i}\left(e^{2\mathrm{i}\beta}\cos^2(\alpha) + e^{2\mathrm{i}\gamma}\sin^2(\alpha)\right) \\ z_1 + \mathrm{i}z_2 &= -2\psi\phi = -2e^{\mathrm{i}(\beta+\gamma)}\cos(\alpha)\sin(\alpha) \end{aligned} \tag{6.8.28}$$

Taking the complex conjugate quantities of previous relation and summing, to have only real parts we have to write:

$$\begin{aligned} x_1 &= \frac{1}{2}(x_1 + \bar{x}_1) = \frac{1}{2}(\psi^2 + \bar{\psi}^2 - \phi^2 - \bar{\phi}^2) \\ y_1 &= \frac{1}{2}(y_1 + \bar{y}_1) = \frac{\mathrm{i}}{2}(\psi^2 - \bar{\psi}^2 + \phi^2 - \bar{\phi}^2) \\ z_1 &= -(\psi\phi + \bar{\psi}\bar{\phi}) \end{aligned} \tag{6.8.29}$$

We can control the norm is equal to the unit. Just check with the squared norm:

$$\begin{aligned} x_1^2 + y_1^2 + z_1^2 &= \frac{1}{4}(\psi^2 + \bar{\psi}^2 - \phi^2 - \bar{\phi}^2)^2 - \frac{1}{4}(\psi^2 - \bar{\psi}^2 + \phi^2 - \bar{\phi}^2)^2 + (\psi\phi + \bar{\psi}\bar{\phi})^2 \\ &= |\psi|^4 - \psi^2\phi^2 - \bar{\psi}^2\bar{\phi}^2 + |\phi|^4 + \psi^2\phi^2 + 2\psi\phi\bar{\psi}\bar{\phi} + \bar{\psi}^2\bar{\phi}^2 \\ &= |\psi|^4 + |\phi|^4 + 2\psi\phi\bar{\psi}\bar{\phi} = |\psi|^4 + |\phi|^4 + 2|\psi|^2|\phi|^2 \\ &= \cos^4(\alpha) + \sin^4(\alpha) + 2\cos^2(\alpha)\sin^2(\alpha) = \cos^4(\alpha) + \sin^4(\alpha) + 2\cos^2(\alpha)(1 - \cos^2(\alpha)) \\ &= \cos^4(\alpha) + \sin^4(\alpha) + 2\cos^2(\alpha) - 2\cos^4(\alpha) = -\cos^4(\alpha) + \sin^4(\alpha) + 2\cos^2(\alpha) \\ &= -\cos^4(\alpha) + (\sin^2(\alpha))^2 + 2\cos^2(\alpha) = -\cos^4(\alpha) + (1 - \cos^2(\alpha))^2 + 2\cos^2(\alpha) = 1 \end{aligned} \tag{6.8.30}$$

In the same way we get:

$$\begin{aligned}x_2&=\frac{\mathrm{i}}{2}(-\psi^2+\bar{\psi}^2+\phi^2-\bar{\phi}^2)\\y_2&=\frac{1}{2}(\psi^2+\bar{\psi}^2+\phi^2+\bar{\phi}^2)\\z_3&=\mathrm{i}(\psi\phi-\bar{\psi}\bar{\phi})\end{aligned}\tag{6.8.31}$$

We can easily check that these values restore well the relations of the vector cross product of $\vec{v}_2$. At any unitary spinor (ψ,ϕ) we can therefore associate three vectors $\vec{v}_1,\vec{v}_2,\vec{v}_3$. We can directly check that the vectors thus calculated are mutually orthogonal and with unit norm.

A reader has make us the request to show in much details as possible this affirmation for x_3. So let's go:

$$\begin{aligned}x_3&=y_1z_2-y_2z_1\\&=\left(\frac{\mathrm{i}}{2}(\psi^2-\bar{\psi}^2+\phi^2-\bar{\phi}^2)\right)(\mathrm{i}(\psi\phi-\bar{\psi}\bar{\phi}))-\left(\frac{1}{2}(\psi^2+\bar{\psi}^2+\phi^2+\bar{\phi}^2)\right)(-(\psi\phi+\bar{\psi}\bar{\phi}))\\&=-(\psi^2-\bar{\psi}^2+\phi^2-\bar{\phi}^2)(\psi\phi-\bar{\psi}\bar{\phi})+\frac{1}{2}(\psi^2+\bar{\psi}^2+\phi^2+\bar{\phi}^2)(\psi\phi+\bar{\psi}\bar{\phi})\\&=\frac{1}{2}\left[(\psi^2+\bar{\psi}^2+\phi^2+\bar{\phi}^2)(\psi\phi+\bar{\psi}\bar{\phi})-(\psi^2-\bar{\psi}^2+\phi^2-\bar{\phi}^2)(\psi\phi-\bar{\psi}\bar{\phi})\right]\\&=\frac{1}{2}\begin{bmatrix}\psi^2\psi\phi+\psi^2\bar{\psi}\bar{\phi}+\bar{\psi}^2\psi\phi+\bar{\psi}^2\bar{\psi}\bar{\phi}+\phi^2\psi\phi+\phi^2\bar{\psi}\bar{\phi}+\bar{\phi}^2\psi\phi+\bar{\phi}^2\bar{\psi}\bar{\phi}\\-\psi^2\psi\phi+\psi^2\bar{\psi}\bar{\phi}+\bar{\psi}^2\psi\phi-\bar{\psi}^2\bar{\psi}\bar{\phi}-\phi^2\psi\phi+\phi^2\bar{\psi}\bar{\phi}+\bar{\phi}^2\psi\phi-\bar{\phi}^2\bar{\psi}\bar{\phi}\end{bmatrix}\\&=\frac{1}{2}\left[2\psi^2\bar{\psi}\bar{\phi}+2\bar{\psi}^2\psi\phi+2\phi^2\bar{\psi}\bar{\phi}+2\bar{\phi}^2\psi\phi\right]=\psi^2\bar{\psi}\bar{\phi}+\bar{\psi}^2\psi\phi+\phi^2\bar{\psi}\bar{\phi}+\bar{\phi}^2\psi\phi\\&=\psi\bar{\psi}\psi\bar{\phi}+\bar{\psi}\psi\bar{\psi}\phi+\phi\bar{\psi}\phi\bar{\phi}+\bar{\phi}\psi\bar{\phi}\phi=\psi\bar{\psi}\left(\psi\bar{\phi}+\bar{\psi}\phi\right)+\phi\bar{\phi}\left(\phi\bar{\psi}+\bar{\phi}\psi\right)\\&=\psi\bar{\psi}\left(\bar{\psi}\phi+\psi\bar{\phi}\right)+\phi\bar{\phi}\left(\phi\bar{\psi}+\bar{\phi}\psi\right)=\psi\bar{\psi}\left(\phi\bar{\psi}+\bar{\phi}\psi\right)+\phi\bar{\phi}\left(\phi\bar{\psi}+\bar{\phi}\psi\right)\\&=\left(\psi\bar{\psi}+\phi\bar{\phi}\right)\left(\phi\bar{\psi}+\bar{\phi}\psi\right)=\left(\psi\bar{\psi}+\phi\bar{\phi}\right)\left(\bar{\psi}\phi+\psi\bar{\phi}\right)=\left(\psi\bar{\psi}+\phi\bar{\phi}\right)\left(\psi\bar{\phi}+\bar{\psi}\phi\right)\end{aligned}\tag{6.8.32}$$

6.8.2 Geometric Properties

We will study the transformations of vectors associated to a spinor to derive the corresponding properties of spinor transformation. As we know some special (trivial) rotations in space can always be expressed as the product of two plane symmetries, therefore we begin by studying these latter.

6.8.2.1 Plane Symmetries

Let us consider first the plan symmetry of a vector:

During a symmetry relative to a plane $\mathcal{P}$, any vector $\overrightarrow{\text{O}M}$ is transformed into a vector $\overrightarrow{\text{O}M'}$. Let us determine a matrix S representing this symmetry with respect to this plane!

Given $\overrightarrow{\text{O}A}$ a unit vector normal to the plane $\mathcal{P}$ and H the root of the perpendicular projection from any given point M of space on the plane $\mathcal{P}$:

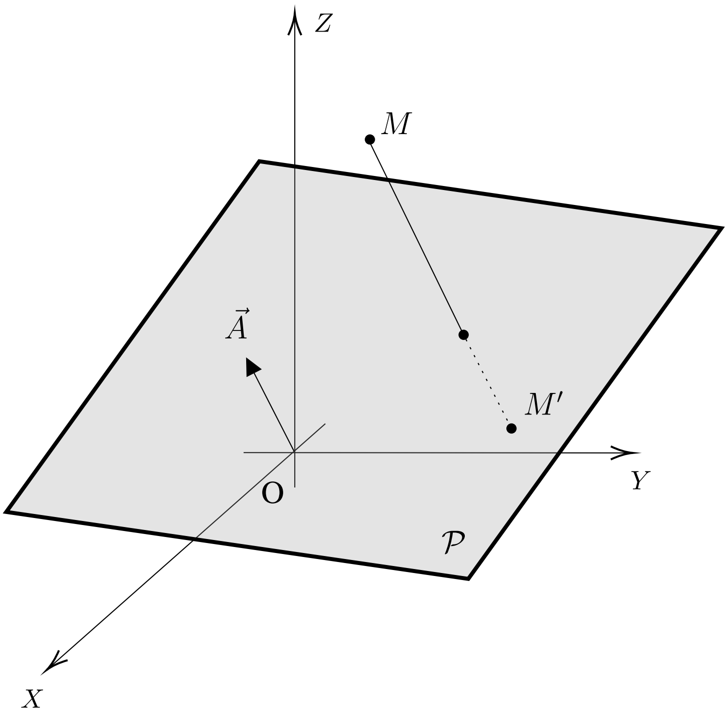


Figure 6.165: **Plane symmetry of a vector relatively to a plane**

Let M' be the symmetric point M with respect to the plane $\mathcal{P}$, we have:

$$\overrightarrow{\text{O}M'} = \overrightarrow{\text{O}M} + \overrightarrow{MM'} = \overrightarrow{\text{O}M} + 2\overrightarrow{MH} = \overrightarrow{\text{O}M} - 2\overrightarrow{\text{O}A}(\overrightarrow{\text{O}M} \circ \overrightarrow{\text{O}A}) \tag{6.8.33}$$

Given a_1, a_2, a_3 the cartesian components of $\overrightarrow{\text{O}A}$ and (x, y, z), (x', y', z') the respective components of the vectors $\overrightarrow{\text{O}M}, \overrightarrow{\text{O}M'}$, the above equation gives us the linear relations:

$$\begin{aligned} x' &= x - 2a_1(a_1x + a_2y + a_3z) \\ y' &= y - 2a_2(a_1x + a_2y + a_3z) \\ z' &= z - 2a_3(a_1x + a_2y + a_3z) \end{aligned} \tag{6.8.34}$$

The matrix S that takes us from vector $\vec{X}(x, y, z)$ to the vector $\vec{X}'(x, y, z)$ has therefore the following expression:

$$S = \begin{bmatrix} 1 - 2a_1^2 & -2a_1a_2 & -2a_1a_3 \\ -2a_2a_1 & 1 - 2a_2^2 & -2a_2a_3 \\ -2a_3a_1 & -2a_3a_2 & 1 - 2a_3^2 \end{bmatrix} \tag{6.8.35}$$

We keep in mind this result and let us now consider two vectors $(\vec{X}_1,\vec{X}_2)$ orthogonal to each other and unitary, defining as we have above seen a unitary spinor (ψ,ϕ) (we used the notation $\vec{v}_1,\vec{v}_2$ before). A symmetry with respect to a plane $\mathcal{P}$ transforms the vectors $\vec{X}_1,\vec{X}_2$ into vectors $\vec{X'}_1,\vec{X'}_2$ which are associated the spinor (ψ',ϕ').

Theorem 6.47. We will now show that the following transformation of the spinor (ψ,ϕ) into spinor $(\psi',\phi')=(x_3',y_3',z_3')$ is:

$$\boxed{\begin{pmatrix}\psi'\\ \phi'\end{pmatrix}=\begin{bmatrix}a_3 & a_1-\mathrm{i}a_2\\ a_1+\mathrm{i}a_2 & -a_3\end{bmatrix}=\mathcal{A}\begin{pmatrix}\psi\\ \phi\end{pmatrix}} \tag{6.8.36}$$

and transforms the vectors $\vec{X}_1,\vec{X}_2$ into vectors $\vec{X}_1',\vec{X}_2'$, these vectors being deduced respectively - as we will just show it - from each other by a single plane symmetry and the matrix $\mathcal{A}$ represents well the sought transformation.

The previous relation gives us therefore:

$$\begin{aligned}\psi'&=a_3\psi+(a_1-\mathrm{i}a_2)\phi\\ \phi'&=(a_1+\mathrm{i}a_2)\phi-a_3\phi\end{aligned} \tag{6.8.37}$$

In all we have so far the previous set of relations and:

$$\vec{X}_3'=(\phi',\psi')\begin{cases}x_1'+\mathrm{i}x_2'=\phi'^2+\psi'^2\\ y_1'+\mathrm{i}y_2'=\mathrm{i}(\psi'^2+\phi'^2)\\ z_1'+\mathrm{i}z_2'=-2\psi'\phi'\end{cases} \tag{6.8.38}$$

Therefore we can deduce:

$$\begin{aligned}x_3'&=\psi'^2-\phi'^2=(a_3\psi+(a_1-\mathrm{i}a_2)\phi)^2-((a_1+\mathrm{i}a_2)\psi-a_3\phi)^2\\
&=(a_3^2\psi^2+2a_3\psi(a_1-\mathrm{i}a_2)\phi+(a_1-\mathrm{i}a_2)^2\phi^2)-((a_1+\mathrm{i}a_2)^2\phi^2-a(a_1+\mathrm{i}a_2)\psi a_3\phi+a_3^2\phi^2)\\
&=(a_3^2\psi^2+2a_3\psi(a_1-\mathrm{i}a_2)\phi+(a_1^2-2a_1\mathrm{i}a_2-a_2^2)\phi^2)\\
&-((a_1^2+2a_1\mathrm{i}a_2-a_2^2)\phi^2-2(a_1+\mathrm{i}a_2)\psi a_3\phi+a_3^2\phi^2)\\
&=a_3^2\psi^2+2a_1a_3\psi\phi-2\mathrm{i}a_2a_3\psi\phi+a_1^2\phi^2-2a_1\mathrm{i}a_2\phi^2-a_2^2\phi^2\\
&-a_1^2\psi^2-2a_1\mathrm{i}a_2\psi^2+a_2^2\psi^2+2a_1a_3\psi\phi+2\mathrm{i}a_2a_2\psi\phi-a_3^2\phi^2\\
&=\left[-a_1^2+a_2^2+a_3^2\right](\psi^2-\phi^2)-2a_1a_2\mathrm{i}(\psi^2+\phi^2)-2a_1a_2(-2\psi\phi)\end{aligned}$$

$$\begin{aligned}y_3'&=\mathrm{i}(\psi'^2+\phi'^2)=\mathrm{i}\left[(a_3\psi+(a_1-\mathrm{i}a_2)\phi)^2+((a_1+\mathrm{i}a_2)\psi-a_3\phi)^2\right]\\
&=\mathrm{i}\left[(a_3^2\psi^2+2a_3\psi(a_1-\mathrm{i}a_2)\phi+(a_1-\mathrm{i}a_2)^2\phi^2)+((a_1+\mathrm{i}a_2)^2\psi^2-2(a_1+\mathrm{i}a_2)\psi a_3\phi+a_3^2\phi^2)\right]\\
&=\mathrm{i}(a_3^2\psi^2+2a_3\psi(a_1-\mathrm{i}a_2)\phi+(a_1^2-2a_1\mathrm{i}a_2-a_2^2)\phi^2)\\
&+\mathrm{i}((a_1^2+2a_1\mathrm{i}a_2-a_2^2)\psi^2-2(a_1+\mathrm{i}a_2)\psi a_2\phi+a_3^2\phi^2)\\
&=\mathrm{i}a_3^2\psi^2+2\mathrm{i}a_1a_2\psi\phi+2a_2a_3\psi\phi+\mathrm{i}a_1^2\phi^2+2a_1a_2\phi^2-\mathrm{i}a_2^2\phi^2\\
&+\mathrm{i}a_1^2\psi^2-2a_1a_2\psi^2-\mathrm{i}a_2^2\psi^2-2a_1a_3\psi\phi+2a_2a_3\psi\phi+\mathrm{i}a_3^2\phi^2\\
&=-2a_1a_2(\psi^2-\phi^2)-2a_2a_3(-2\psi\phi)+(a_1^2-a_2^2+a_3^2)\mathrm{i}(\psi^2+\phi^2)\end{aligned}$$

$$\begin{aligned}z_3'&=-2\psi'\phi'=-2(a_2\psi+(a_1-\mathrm{i}a_2)\phi)((a_1+\mathrm{i}a_2)\psi-a_3\phi)\\
&=-2(a_3\psi+a_1\phi-\mathrm{i}a_2\phi)(a_1\psi+\mathrm{i}a_2\psi-a_3\phi)\\
&=-2a_3a_1(\psi^2-\phi^2)-2a_2a_3\mathrm{i}(\psi^2+\phi^2)+(a_1^2+a_2^2-a_3^2)(-2\psi\phi)\end{aligned} \tag{6.8.39}$$

After, using the fact that $\|\vec{A}\|=a_1^2+a_2^2+a_3^2=1$, we get:

$$\begin{aligned}x_3'&=[1-2a_1^2]x_3-2a_1a_2y_3-2a_1a_3z_3\\y_3'&=-2a_1a_2x_3+[1-2a_2^2]y_3-2a_2a_3z_3\\z_3'&=-2a_3a_1x_3-2a_3a_2y_3+[1-2a_3^2]z_3\end{aligned}\tag{6.8.40}$$

So we fall well back on the symmetry matrix:

$$S=\begin{bmatrix}1-2a_1^2 & -2a_1a_2 & -2a_1a_3\\-2a_2a_1 & 1-2a_2^2 & -2a_2a_3\\-2a_3a_1 & -2a_3a_2 & 1-2a_3^2\end{bmatrix}\tag{6.8.41}$$

Thus, the matrix that we will see again in the section of Relativistic Quantum Physics:

$$\boxed{\mathcal{A}=\begin{bmatrix}a_3 & a_1-\mathrm{i}a_2\\a_1+\mathrm{i}a_2 & -a_3\end{bmatrix}}\tag{6.8.42}$$

$\mathcal{A}$ therefore generates the transformation of a spinor (ψ,ϕ) into a spinor (ψ',ϕ') such that the associated vectors $(\vec{X}_1,\vec{X}_2)$ can be deduced respectively from $(\vec{X}_1',\vec{X}_2')$ by a planar symmetry.

6.8.2.2 Rotations

As we have saw it in the section of Euclidean Geometry, it is possible to rotate a vector in the plane or in space using matrices. Similarly, by extension, it is clear that the multiplication of two rotations is a rotation (that is the elementary linear algebra - at least we consider it as is).

Consider therefore two planes P,Q whose intersection generates a line L and let us denote $\vec{A}(a_1,a_2,a_3)$ and $\vec{B}(b_1,b_2,b_2)$ the unit vectors carried by the respective normal vectors (see section Vector Calculus page 1801) to these two intersecting planes in L:

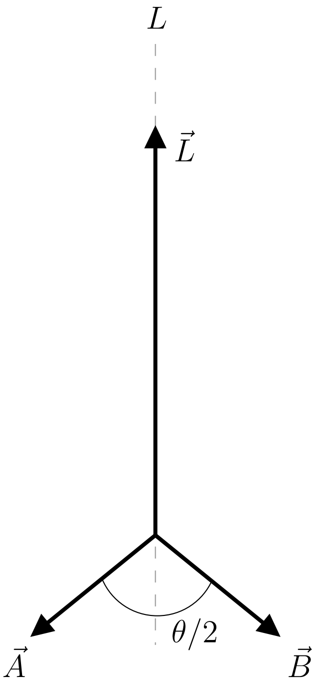


Figure 6.166: **Illustrated intersection of two planes**

Let us denote by $\theta/2$ the angle of the vectors $\vec{A},\vec{B}$ between them (the reason for this notation comes from our study of quaternions (see section Numbers page 261)). Given $\vec{L}$ the unit direction vector carried by the line L resulting from the intersection of planes P,Q and such that:

$$\vec{A}\times\vec{B}=\vec{L}\sin\left(\frac{\theta}{2}\right) \tag{6.8.43}$$

Explanations: $\vec{A},\vec{B}$ are unitary but not necessarily perpendicular and we still need to ensure that $\vec{L}$ is a unit vector (the norm equal to unity!). Therefore, the above relation ensures that:

$$\|\vec{A}\|\cdot\|\vec{B}\|\cancel{\sin\left(\frac{\theta}{2}\right)}=\|\vec{L}\|\cancel{\sin\left(\frac{\theta}{2}\right)}\Rightarrow\|\vec{A}\|\cdot\|\vec{B}\|=\|\vec{L}\|=L_1^2+L_2^2+L_3^2=1 \tag{6.8.44}$$

The previous vector product gives us for the components of $\vec{L}$:

$$L_1\sin\left(\frac{\theta}{2}\right)=a_2b_3-a_3b_2 \tag{6.8.45}$$

$$L_2\sin\left(\frac{\theta}{2}\right)=a_3b_1-a_1b_3 \tag{6.8.46}$$

$$L_3\sin\left(\frac{\theta}{2}\right)=a_1b_2-a_2b_1 \tag{6.8.47}$$

On the other hand, the scalar product can be written:

$$\vec{A}\circ\vec{B}=a_1b_1+a_2b_2+a_3b_3=\|\vec{A}\|\cdot\|\vec{B}\|\cos\left(\frac{\theta}{2}\right)=\cos\left(\frac{\theta}{2}\right) \tag{6.8.48}$$

Remark

We will use these two planes as symmetry planes for our rotations.

As we have noticed previously, a rotation in $\mathcal{E}^3$ can always be done with more than two plane symmetries. Thus, a rotation can be denoted by the application (multiplication) of two matrices of symmetry according to the results obtained previously:

$$R\left(\vec{L},\frac{\theta}{2}\right)=\begin{bmatrix}b_3 & b_1-\mathrm{i}b_2\\ b_1+\mathrm{i}b_2 & -b_3\end{bmatrix}\begin{bmatrix}a_3 & a_1-\mathrm{i}a_2\\ a_1+\mathrm{i}a_2 & -a_3\end{bmatrix} \tag{6.8.49}$$

Developing the product of these two matrices and taking into account the relations arising from vector and dot product we get:

$$R\left(\vec{L},\frac{\theta}{2}\right)=\begin{bmatrix}\cos\left(\frac{\theta}{2}\right)-\mathrm{i}L_3\sin\left(\frac{\theta}{2}\right) & -(\mathrm{i}L_1+L_2)\sin(\left(\frac{\theta}{2}\right)\\ (-\mathrm{i}L_1+L_2)\sin\left(\frac{\theta}{2}\right) & \cos\left(\frac{\theta}{2}\right)+\mathrm{i}L_3\sin(\left(\frac{\theta}{2}\right)\end{bmatrix} \tag{6.8.50}$$

Thus, we can write the transformation of a spinor (ψ,ϕ) and a spinor (ψ',ϕ') with a matrix of the form:

$$R=\begin{bmatrix}a & b\\ -\bar{b} & \bar{a}\end{bmatrix} \tag{6.8.51}$$

whose parameters are named "Cayley-Klein parameters".

The matrix $R\left(\vec{L},\frac{\theta}{2}\right)$ can be written in another form if we do a limited development for infinitely small rotations $\theta/2=\varepsilon/2$ (that's where the physics comes back...):

$$\cos\left(\frac{\varepsilon}{2}\right)=1-\frac{1}{2}\left(\frac{\varepsilon}{2}\right)^2+\ldots \qquad \sin\left(\frac{\varepsilon}{2}\right)=\frac{\varepsilon}{2}-\frac{1}{6}\left(\frac{\varepsilon}{2}\right)^3+\ldots \tag{6.8.52}$$

Using only the first order terms, the rotation matrix is finally written:

$$\boxed{R\left(\vec{L},\frac{\theta}{2}\right)=\begin{bmatrix}1-\mathrm{i}L_3\frac{\varepsilon}{2} & -(\mathrm{i}L_1+L_2)\frac{\varepsilon}{2}\\ (-\mathrm{i}L_1+L_2)\frac{\varepsilon}{2} & 1+\mathrm{i}L_3\frac{\varepsilon}{2}\end{bmatrix}=\begin{bmatrix}1&0\\0&1\end{bmatrix}+\frac{\varepsilon}{2}\begin{bmatrix}-\mathrm{i}L_3 & -(\mathrm{i}L_1+L_2)\\ (-\mathrm{i}L_1+L_2) & \mathrm{i}L_3\end{bmatrix}} \tag{6.8.53}$$

This matrix is the limited development of the matrix of rotation in the neighbourhood of the identity matrix, the latter obviously corresponding to the zero rotation. We note also the latter in the form:

$$R\left(\vec{L},\frac{\theta}{2}\right)=\sigma_0+\varepsilon\chi(\vec{L}) \tag{6.8.54}$$

where the matrix σ_0 is the identity matrix of order two and $\chi(\vec{L})$ is named the "infinitesimal rotation matrix". Now, if we put $L_1=1,L_2=L_3=0$ in $\chi(\vec{L})$ we get:

$$\chi(\vec{L})=\chi(\mathrm{O}x)=\frac{1}{2}\begin{bmatrix}0&-\mathrm{i}\\-\mathrm{i}&0\end{bmatrix}=-\frac{\mathrm{i}}{2}\begin{bmatrix}0&1\\1&0\end{bmatrix} \tag{6.8.55}$$

How to interpret this result? Well it's quite simple, choose $L_1=1,L_2=L_3=0$ gives us a collinear vector $\vec{L}$ to the axis Ox. Therefore, we can very well imagine the planes generating the axis Ox that carries $\vec{L}$. As $\varepsilon/2$ (verbatim $\theta/2$) is generated by the vectors $\vec{A},\vec{B}$ perpendicular to $\vec{L}$ and thus to Ox, then the angle $\varepsilon/2$ (or its variation) represents a variation of the direction of the normal planes to $\vec{A},\vec{B}$ which by symmetry are used to construct the rotation (recall that $\vec{A},\vec{B}$are not necessarily mutually orthogonal). So by extension, having $L_1=1,L_2=L_3=0$ allows us then only to make rotations (symmetries) around the x-axis.

Similarly, a rotation about the y-axis corresponds to $L_2=1,L_1=L_3=0$, which gives:

$$\chi(\vec{L})=\chi(\mathrm{O}y)=\frac{1}{2}\begin{bmatrix}0&-1\\1&0\end{bmatrix}=-\frac{\mathrm{i}}{2}\begin{bmatrix}0&-\mathrm{i}\\\mathrm{i}&0\end{bmatrix} \tag{6.8.56}$$

and the same with $L_3=1,L_1=L_2=0$ we finally have:

$$\chi(\vec{L})=\chi(\mathrm{O}z)=\frac{1}{2}\begin{bmatrix}-\mathrm{i}&0\\0&\mathrm{i}\end{bmatrix}=-\frac{\mathrm{i}}{2}\begin{bmatrix}1&0\\0&-1\end{bmatrix} \tag{6.8.57}$$

The three matrices:

$$\begin{aligned}\chi(\vec{L})&=\chi(\mathrm{O}x)=\frac{1}{2}\begin{bmatrix}0&-\mathrm{i}\\-\mathrm{i}&0\end{bmatrix}=-\frac{\mathrm{i}}{2}\begin{bmatrix}0&1\\1&0\end{bmatrix}\\ \chi(\vec{L})&=\chi(\mathrm{O}y)=\frac{1}{2}\begin{bmatrix}0&-1\\1&0\end{bmatrix}=-\frac{\mathrm{i}}{2}\begin{bmatrix}0&-\mathrm{i}\\\mathrm{i}&0\end{bmatrix}\\ \chi(\vec{L})&=\chi(\mathrm{O}z)=\frac{1}{2}\begin{bmatrix}-\mathrm{i}&0\\0&\mathrm{i}\end{bmatrix}=-\frac{\mathrm{i}}{2}\begin{bmatrix}1&0\\0&-1\end{bmatrix}\end{aligned} \tag{6.8.58}$$

are rotation matrices in the space of "two-dimensional spinors". Physicists and mathematicians say that these matrices are an irreducible representation of dimension two of the group "$\mathrm{SU}(2)$" or named "special group of spatial rotations $\mathrm{SU}(2)$" (see section Set Algebra page 1425).

The previous infinitesimal matrices therefore show us in a skilful way the following matrices:

$$\boxed{\sigma_1=\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix},\sigma_2=\begin{bmatrix}0 & -\mathrm{i}\\ \mathrm{i} & 0\end{bmatrix},\sigma_3=\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}} \tag{6.8.59}$$

These matrices are named "Pauli matrices" and we will find them again in the section of Relativistic Quantum Physics as part of the study of the Dirac equation and the determination of its explicit solutions (using spinors).

Using Pauli matrices, the infinitesimal rotations matrix can finally be written:

$$R\left(\vec{L},\frac{\theta}{2}\right)=\sigma_0+\varepsilon\chi(\vec{L})=\sigma_0-\left(\mathrm{i}\frac{\varepsilon}{2}\right)(L_1\sigma_1+L_2\sigma_2+L_3\sigma_3) \tag{6.8.60}$$

Let us define a vector $\vec{\sigma}$, named "Pauli vector", whose components are the Pauli matrices:

$$\vec{\sigma}=(\sigma_1,\sigma_2,\sigma_3) \tag{6.8.61}$$

The expression $L_1\sigma_1+L_2\sigma_2+L_3\sigma_3$ can be written as a sort of dot product which represents a sum of matrices (the arrow above the sigma is sometimes omitted if no confusion is possible):

$$\boxed{\vec{L}\circ\vec{\sigma}=L_1\sigma_1+L_2\sigma_2+L_3\sigma_3} \tag{6.8.62}$$

The limited development is then written:

$$R\left(\vec{L},\frac{\theta}{2}\right)=\sigma_0+\varepsilon\chi(\vec{L})=\sigma_0-\left(\mathrm{i}\frac{\varepsilon}{2}\right)(L_1\sigma_1+L_2\sigma_2+L_3\sigma_3)=\sigma_0-\left(\mathrm{i}\frac{\varepsilon}{2}\right)\vec{L}\circ\vec{\sigma} \tag{6.8.63}$$

The rotations matrix:

$$R\left(\vec{L},\frac{\theta}{2}\right)=\begin{bmatrix}\cos\left(\frac{\theta}{2}\right)-\mathrm{i}L_3\sin\left(\frac{\theta}{2}\right) & -(\mathrm{i}L_1+L_2)\sin(\left(\frac{\theta}{2}\right)\\ (-\mathrm{i}L_1+L_2)\sin\left(\frac{\theta}{2}\right) & \cos\left(\frac{\theta}{2}\right)+\mathrm{i}L_3\sin(\left(\frac{\theta}{2}\right)\end{bmatrix} \tag{6.8.64}$$

can using Pauli matrices be written in the remarkable form under the assumption of small angles:

$$\boxed{R\left(\vec{L},\frac{\theta}{2}\right)=\cos\left(\frac{\theta}{2}\right)\sigma_0-\mathrm{i}\sin\left(\frac{\theta}{2}\right)\vec{L}\circ\vec{\sigma}} \tag{6.8.65}$$

Expression that we will a lot in the section of Quantum Computing to express the R matrices explicitly and also in the section of Set Algebra page 1417.

What is written sometimes:

$$R\left(\vec{L},\theta\right)=\cos\left(\frac{\theta}{2}\right)\mathbb{1}-\mathrm{i}\sin\left(\frac{\theta}{2}\right)\vec{L}\circ\vec{\sigma} \tag{6.8.66}$$

Which can also be written in an extensive form:

$$\boxed{R\left(\vec{L},\frac{\theta}{2}\right)=\cos\left(\frac{\theta}{2}\right)\sigma_0-L_1\mathrm{i}\sin\left(\frac{\theta}{2}\right)\sigma_1-L_2\mathrm{i}\sin\left(\frac{\theta}{2}\right)\sigma_2-L_3\mathrm{i}\sin\left(\frac{\theta}{2}\right)\sigma_3} \tag{6.8.67}$$

which has the form of a quaternion of angle θ (don't forget that this is only for small angles!) and of axis $\vec{L}$. Hence the reason that have from the beginning chosen the notation $\varepsilon/2$.

It is clear, so that the analogy with quaternions to be stronger, that the 2×2 Pauli matrices are a set of four linearly independent matrices! As the canonical basis for quaternions!

If we denote by $\vec{L}(L_1,L_2,L_3)=\vec{X}(x,y,z)$ then the "spinor product" is finally defined by:

$$\boxed{L_1\sigma_1+L_2\sigma_2+L_3\sigma_3=x\sigma_1+y\sigma_2+z\sigma_3=\vec{X}\circ\vec{\sigma}=\begin{bmatrix}z & x-\mathrm{i}y\\ x+\mathrm{i}y & -z\end{bmatrix}} \tag{6.8.68}$$

This matrix constitutes as we have already mentioned, to the limited development of the rotation matrix in the neighbourhood of the identity matrix, the components of $\vec{x}$ being associated with a spinor whose rotation is through the double symmetry defined by two planes whose intersection is defined by the vector $\vec{X}$.

We can also notice the interesting consequence that a rotation of 2π (360°) rotation does not restore the object to its original position!!!

Indeed:

$$\begin{gathered}R(\vec{L},0)=\cos(0)\mathbb{1}-\mathrm{i}\sin(0)\vec{L}\circ\vec{\sigma}=\mathbb{1}\\ R(\vec{L},2\pi)=\cos(\pi)\mathbb{1}-\mathrm{i}\sin(\pi)\vec{L}\circ\vec{\sigma}=-\mathbb{1}\\ R(\vec{L},4\pi)=\cos(2\pi)\mathbb{1}-\mathrm{i}\sin(2\pi)\vec{L}\circ\vec{\sigma}=\mathbb{1}\end{gathered} \tag{6.8.69}$$

Therefore we need a rotation of 4π (720°) to make a full turn! This corresponds to the spin of $^1\!/_2$. It takes two turns to find that the object reappears equivalently (this is counter intuitive and can make thing we are dealing with object having higher dimensions that what we expected!). We then say that the representation of rotations is "bivaluated".

Schematically this can be represented as:

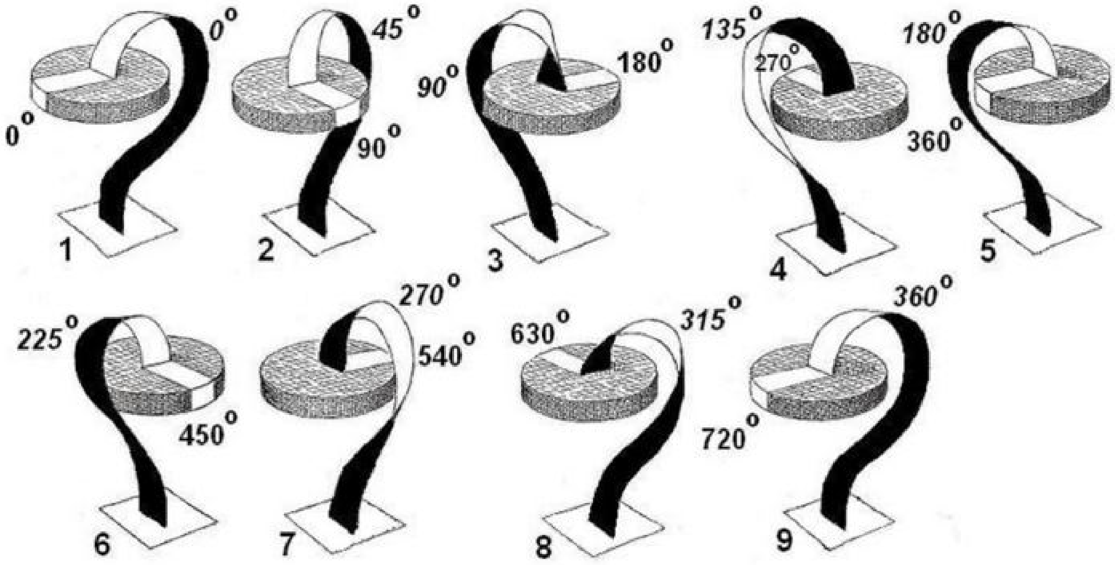

0°
0°
1
45°
90°
2
90°
180°
3
135°
270°
4
180°
360°
5
225°
450°
6
270°
540°
7
630°
315°
720°
8
360°
9

Figure 6.167: **Spinor full rotation (special example)**

Or you can consider the analogy that is to hold a tea cup on the palm of your hand and you turn your hand by maintaining it flat to regain its (the hand, not the cup!) original position. You will see that your hand has to do two laps!

6.8.2.3 Properties of Pauli Matrices

The reader can easily check (if this is not the case he can always contact us and we will write the details) the following main properties of the Pauli matrices, some of which will be used in the section of Relativistic Quantum Physics.

First notice that all three of the Pauli matrices can be compacted into a single expression:

$$\sigma_j=\begin{pmatrix} \delta_{j3} & \delta_{j1}-\mathrm{i}\delta_{j2} \\ \delta_{j1}+\mathrm{i}\delta_{j2} & -\delta_{j3} \end{pmatrix} \tag{6.8.70}$$

where δ_{jk} is the Kronecker delta, which equals $+1$ if $j=k$ and 0 otherwise.

So let's go now for the properties:

P1. Unitarity:

$$\sigma_1^2=\sigma_2^2=\sigma_3^2=\sigma_0 \tag{6.8.71}$$

P2. Anticommutativity:

$$\sigma_i\sigma_j=-\sigma_j\sigma_i \tag{6.8.72}$$

or $i\neq j$ and $i,j=1,2,3$.

The last two properties gives us:

$$\sigma_i\sigma_j+\sigma_j\sigma_i=2\delta_{ij}\sigma_0 \tag{6.8.73}$$

with $i,j=1,2,3$.

P3. Cyclicity:

$$\sigma_1\sigma_2=\mathrm{i}\sigma_3 \quad \sigma_2\sigma_3=\mathrm{i}\sigma_1 \quad \sigma_3\sigma_1=\mathrm{i}\sigma_2 \tag{6.8.74}$$

P4. Commutativity:

$$\begin{aligned}\sigma_1\sigma_2-\sigma_2\sigma_1&=2\mathrm{i}\sigma_3\\ \sigma_3\sigma_1-\sigma_1\sigma_3&=2\mathrm{i}\sigma_2\\ \sigma_2\sigma_3-\sigma_3\sigma_2&=2\mathrm{i}\sigma_1\end{aligned} \tag{6.8.75}$$

P5. Vector product:

Given the square of the different σ by noting abusively by "1" the unitary matrix (we change the indices to give you the habit to use other common notations):

$$\begin{aligned}\sigma_x\sigma_x&=\begin{bmatrix}0&1\\1&0\end{bmatrix}\begin{bmatrix}0&1\\1&0\end{bmatrix}=\begin{bmatrix}1&0\\0&1\end{bmatrix}=1\\ \sigma_y\sigma_y&=\begin{bmatrix}0&-\mathrm{i}\\\mathrm{i}&0\end{bmatrix}\begin{bmatrix}0&-\mathrm{i}\\\mathrm{i}&0\end{bmatrix}=\begin{bmatrix}1&0\\0&1\end{bmatrix}=1\\ \sigma_z\sigma_z&=\begin{bmatrix}1&0\\0&-1\end{bmatrix}\begin{bmatrix}1&0\\0&-1\end{bmatrix}=\begin{bmatrix}1&0\\0&1\end{bmatrix}=1\end{aligned} \tag{6.8.76}$$

Leading us to write that (square norm of the Pauli vector):

$$\sigma=\sigma_x^2+\sigma_y^2+\sigma_z^2=3\cdot\mathbb{1}=3\begin{bmatrix}1&0\\0&1\end{bmatrix} \tag{6.8.77}$$

Let us consider now the following products:

$$\begin{gathered}
\sigma_x\sigma_y=\begin{bmatrix}0&1\\1&0\end{bmatrix}\begin{bmatrix}0&-\mathrm{i}\\\mathrm{i}&0\end{bmatrix}=\begin{bmatrix}\mathrm{i}&0\\0&-\mathrm{i}\end{bmatrix}=\mathrm{i}\begin{bmatrix}1&0\\0&-1\end{bmatrix}=\mathrm{i}\sigma_z\\
\sigma_x\sigma_y=\mathrm{i}\sigma_z\\
\sigma_x\sigma_z=\begin{bmatrix}0&1\\1&0\end{bmatrix}\begin{bmatrix}1&0\\0&-1\end{bmatrix}=\begin{bmatrix}0&-1\\1&0\end{bmatrix}=\begin{bmatrix}0&\mathrm{i}\cdot\mathrm{i}\\-\mathrm{i}\cdot\mathrm{i}&0\end{bmatrix}=\mathrm{i}\begin{bmatrix}0&\mathrm{i}\\-\mathrm{i}&0\end{bmatrix}=-\mathrm{i}\begin{bmatrix}0&-\mathrm{i}\\\mathrm{i}&0\end{bmatrix}=-\mathrm{i}\sigma_y\\
\sigma_x\sigma_z=-\mathrm{i}\sigma_y\\
\sigma_z\sigma_x=\begin{bmatrix}1&0\\0&-1\end{bmatrix}\begin{bmatrix}0&1\\1&0\end{bmatrix}=\begin{bmatrix}0&1\\-1&0\end{bmatrix}=\begin{bmatrix}0&-\mathrm{i}\cdot\mathrm{i}\\\mathrm{i}\cdot\mathrm{i}&0\end{bmatrix}=\mathrm{i}\begin{bmatrix}0&-\mathrm{i}\\\mathrm{i}&0\end{bmatrix}=\mathrm{i}\sigma_y\\
\sigma_z\sigma_x=\mathrm{i}\sigma_y\\
\sigma_y\sigma_z=\begin{bmatrix}0&-\mathrm{i}\\\mathrm{i}&0\end{bmatrix}\begin{bmatrix}1&0\\0&-1\end{bmatrix}=\begin{bmatrix}0&\mathrm{i}\\\mathrm{i}&0\end{bmatrix}=\mathrm{i}\begin{bmatrix}0&1\\1&0\end{bmatrix}=\mathrm{i}\sigma_x\\
\sigma_y\sigma_z=\mathrm{i}\sigma_x\\
\sigma_z\sigma_y=\begin{bmatrix}1&0\\0&-1\end{bmatrix}\begin{bmatrix}0&-\mathrm{i}\\\mathrm{i}&0\end{bmatrix}=\begin{bmatrix}0&-\mathrm{i}\\-\mathrm{i}&0\end{bmatrix}=-\mathrm{i}\begin{bmatrix}0&1\\1&0\end{bmatrix}=-\mathrm{i}\sigma_x\\
\sigma_z\sigma_y=-\mathrm{i}\sigma_x
\end{gathered} \tag{6.8.78}$$

All these relations can be summarized into a unique one (!):

$$\boxed{\sigma_i\sigma_j=\delta_{ij}\mathbb{1}+\mathrm{i}\varepsilon_{ijk}\sigma_k} \tag{6.8.79}$$

where for recall (see section Tensor Calculus page 2044) the Kronecker symbol is defined by:

$$\delta_{ij}=\delta^{ij}=\delta_i^j=\begin{cases}1&\text{if } i=j\\0&\text{if } i\neq j\end{cases} \tag{6.8.80}$$

In three dimensions, the Levi-Civita symbol is defined as follows:

$$\varepsilon_{ijk}=\begin{cases}+1&\text{if }(i,j,k)\text{ is }(1,2,3),(2,3,1)\text{ or }(3,1,2),\\-1&\text{if }(i,j,k)\text{ is }(3,2,1),(1,3,2)\text{ or }(2,1,3),\\\ \ 0&\text{if }i=j\text{ or }j=k\text{ or }k=i\end{cases} \tag{6.8.81}$$

i.e. ε_{ijk} is 1 if (i,j,k) is an even permutation of $(1,2,3)$ or in the natural order $(1,2,3)$, -1 if it is an odd permutation, and 0 if any index is repeated. In three dimensions only, the cyclic permutations of $(1,2,3)$ are all even permutations, similarly the anti-cyclic permutations are all odd permutations. This means in 3D it is sufficient to take cyclic or anti-cyclic permutations of $(1,2,3)$ and easily obtain all the even or odd permutations.

The boxed relation above can be presented as a Cayley table:

×	σ_x	σ_y	σ_z
σ_x	$\mathbb{1}$	$i\sigma_z$	$-\mathrm{i}\sigma_y$
σ_y	$-\mathrm{i}\sigma_z$	$\mathbb{1}$	$\mathrm{i}\sigma_x$
σ_z	$\mathrm{i}\sigma_y$	$-\mathrm{i}\sigma_x$	$\mathbb{1}$

(6.8.82)

We see therefore that the Pauli matrices are involutory (a square matrix is its own inverse):

$$\sigma_1^2=\sigma_2^2=\sigma_3^2=-i\sigma_1\sigma_2\sigma_3=\begin{pmatrix}1&0\\0&1\end{pmatrix}=\mathbb{1} \tag{6.8.83}$$

We also have:

$$\begin{aligned}\sigma_x\sigma_y-\sigma_y\sigma_x&=2\mathrm{i}\sigma_z\\ \sigma_y\sigma_z-\sigma_z\sigma_y&=2\mathrm{i}\sigma_x\\ \sigma_z\sigma_x-\sigma_x\sigma_z&=2\mathrm{i}\sigma_y\end{aligned} \tag{6.8.84}$$

That can be summarized into:

$$[\sigma_j,\sigma_k]=2\mathrm{i}\varepsilon_{jkl}\sigma_l \tag{6.8.85}$$

We fall back here on the components of the vector product:

$$\vec{\sigma}\times\vec{\sigma}=2\mathrm{i}\vec{\sigma} \tag{6.8.86}$$

Now let us develop an important spinor identity which will be useful to us in the section of Relativistic Quantum Physics:

$$\begin{aligned}(\vec{\sigma}\circ\vec{a})(\vec{\sigma}\circ\vec{b})&=(\sigma_1a_1+\sigma_2a_2+\sigma_3a_3)(\sigma_1b_1+\sigma_2b_2+\sigma_3b_3)\\ &=\sum_{i,j}a_i\sigma_ib_j\sigma_j\\ &=\begin{bmatrix}a_z&a_x-\mathrm{i}a_y\\a_x+\mathrm{i}a_y&-a_z\end{bmatrix}\begin{bmatrix}b_z&b_x-\mathrm{i}b_y\\b_x+\mathrm{i}b_y&-b_z\end{bmatrix}\\ &=\begin{bmatrix}a_zb_z+(b_x+\mathrm{i}b_y)(a_x-\mathrm{i}a_y)&a_z(b_x-\mathrm{i}b_y)-b_z(a_x-\mathrm{i}a_y)\\b_z(a_x+\mathrm{i}a_y)-a_z(b_x+\mathrm{i}b_y)&(b_x-\mathrm{i}b_y)(a_x+\mathrm{i}a_y)+a_zb_z\end{bmatrix}\\ &=\begin{bmatrix}a_zb_z+b_xa_x+b_ya_y+\mathrm{i}(b_ya_x-b_xa_y)&a_zb_x-\mathrm{i}a_zb_y-b_za_x+b_z\mathrm{i}a_y\\b_za_x+b_z\mathrm{i}a_y-a_zb_x-\mathrm{i}a_zb_y&a_zb_z+b_xa_x+b_ya_y+\mathrm{i}(b_xa_y-b_ya_x)\end{bmatrix}\\ &=\begin{bmatrix}a_xb_x+a_yb_y+a_zb_z&0\\0&a_xb_x+a_yb_y+a_zb_z\end{bmatrix}\\ &+\begin{bmatrix}\mathrm{i}(b_ya_x-b_xa_y)&a_zb_x-b_za_x+\mathrm{i}(b_za_y-a_zb_y)\\b_za_x-a_zb_x+\mathrm{i}(b_za_y-a_zb_y)&\mathrm{i}(b_xa_y-b_ya_x)\end{bmatrix}\\ &=(\vec{a}\circ\vec{b})\mathbb{1}+\begin{bmatrix}\mathrm{i}(b_ya_x-b_xa_y)&a_zb_x-b_za_x+\mathrm{i}(b_za_y-a_zb_y)\\b_za_x-a_zb_x+\mathrm{i}(b_za_y-a_zb_y)&\mathrm{i}(b_xa_y-b_ya_x)\end{bmatrix}\end{aligned} \tag{6.8.87}$$

But we also have:

$$\begin{aligned}\mathrm{i}\vec{\sigma}\circ(\vec{a}\times\vec{b})&=\mathrm{i}\vec{\sigma}\circ\begin{pmatrix}a_yb_z-a_zb_y\\a_zb_x-a_xb_z\\a_xb_y-a_yb_x\end{pmatrix}\\ &=\mathrm{i}\begin{bmatrix}a_xb_y-a_yb_x&(a_yb_z-a_zb_y)-\mathrm{i}(a_zb_x-a_xb_z)\\(a_yb_z-a_zb_y)+\mathrm{i}(a_zb_x-a_xb_z)&(a_xb_y-a_yb_x)\end{bmatrix}\\ &=\begin{bmatrix}\mathrm{i}(b_ya_x-b_xa_y)&a_zb_x-b_za_x+\mathrm{i}(b_za_y-a_zb_y)\\b_za_x-a_zb_x+\mathrm{i}(b_za_y-b_ya_z)&\mathrm{i}(a_yb_x-a_xb_y)\end{bmatrix}\end{aligned} \tag{6.8.88}$$

So finally:

$$\boxed{(\vec{\sigma}\circ\vec{a})(\vec{\sigma}\circ\vec{b})=(\vec{a}\circ\vec{b})\mathbb{1}+\mathrm{i}\vec{\sigma}\circ(\vec{a}\times\vec{b})} \tag{6.8.89}$$

P6. We note that these matrices are also hermitian (let us recall that a Hermitian matrix is a transposed matrix followed by its complex conjugate according to what we saw in the section of Linear Algebra) such that:

$$\sigma_x = \sigma_x^\dagger \qquad \sigma_y = \sigma_y^\dagger \qquad \sigma_z = \sigma_z^\dagger \tag{6.8.90}$$

It is therefore in the language of quantum physics: Hermitian operators!!!

P7. The following traces can be derived using the commutation and anticommutation relations:

$$\begin{aligned}\operatorname{tr}(\sigma_j) &= 0\\ \operatorname{tr}(\sigma_j\sigma_k) &= 2\delta_{jk}\\ \operatorname{tr}(\sigma_j\sigma_k\sigma_\ell) &= 2\mathrm{i}\varepsilon_{jk\ell}\\ \operatorname{tr}(\sigma_j\sigma_k\sigma_\ell\sigma_m) &= 2\left(\delta_{jk}\delta_{\ell m} - \delta_{j\ell}\delta_{km} + \delta_{jm}\delta_{k\ell}\right)\end{aligned} \tag{6.8.91}$$

If the matrix $\sigma_0 = \mathbb{1}$ is also considered, these relationships become:

$$\begin{aligned}\operatorname{tr}(\sigma_\alpha) &= 2\delta_{0\alpha}\\ \operatorname{tr}(\sigma_\alpha\sigma_\beta) &= 2\delta_{\alpha\beta}\\ \operatorname{tr}(\sigma_\alpha\sigma_\beta\sigma_\gamma) &= 2\sum_{(\alpha\beta\gamma)}\delta_{\alpha\beta}\delta_{0\gamma} - 4\delta_{0\alpha}\delta_{0\beta}\delta_{0\gamma} + 2\mathrm{i}\varepsilon_{0\alpha\beta\gamma}\\ \operatorname{tr}(\sigma_\alpha\sigma_\beta\sigma_\gamma\sigma_\mu) &= 2\left(\delta_{\alpha\beta}\delta_{\gamma\mu} - \delta_{\alpha\gamma}\delta_{\beta\mu} + \delta_{\alpha\mu}\delta_{\beta\gamma}\right) + 4\left(\delta_{\alpha\gamma}\delta_{0\beta}\delta_{0\mu} + \delta_{\beta\mu}\delta_{0\alpha}\delta_{0\gamma}\right) - 8\delta_{0\alpha}\delta_{0\beta}\delta_{0\gamma}\delta_{0\mu} + 2\mathrm{i}\sum_{(\alpha\beta\gamma\mu)}\varepsilon_{0\alpha\beta\gamma}\delta_{0\mu}\end{aligned} \tag{6.8.92}$$

where Greek indices α, β, γ and μ assume values from $\{0, x, y, z\}$ and the notation $\sum_{(\alpha\ldots)}$ is used to denote the sum over the cyclic permutation of the included indices.

Let us now see what are the eigenvectors and eigenvalues of the Pauli matrices because this result is very useful for the section of Relativistic Quantum Physics and of Quantum Computing!

Let us recall that when a transformation (application of a matrix) act on a vector, it changes the direction of the vector except for specific matrices that have eigenvalues. In this case, the direction is conserved but not their length. This property is exploited in quantum mechanics (and not only as we will see in many other sections of this textbook).

Let us determine in a first time, the associated eigenvectors and eigenvalues (see section Linear Algebra page 1997) using the most common method:

The eigenvalues equation (see section Linear Algebra page 1998) is thus written:

$$\det(\sigma_x - \lambda I_n) = \det\begin{bmatrix}0-\lambda & 1\\ 1 & 0-\lambda\end{bmatrix} = 0 \tag{6.8.93}$$

Which gives us as characteristic equation:

$$\det(\sigma_x - \lambda I_n) = \lambda^2 - 1 = 0 \tag{6.8.94}$$

hence the eigenvalues $\lambda = \pm 1$. Which gives us the possibility to determine the eigenvectors as following:

$$\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix} = \begin{pmatrix}x\\ y\end{pmatrix} \Rightarrow \begin{pmatrix}y\\ x\end{pmatrix} = \lambda\begin{pmatrix}x\\ y\end{pmatrix} \tag{6.8.95}$$

Therefore for $\lambda=1$:

$$\begin{pmatrix} y \\ x \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \tag{6.8.96}$$

This impose that $y=x$. The eigenvector is therefore:

$$\begin{pmatrix} x \\ y=x \end{pmatrix} = x\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \forall x \tag{6.8.97}$$

whatever is the value of x.

Conclusion: The proper direction of the vector is conserved but not its length (norm) because it depends on the value of x.

For $\lambda=-1$:

$$\begin{pmatrix} y \\ x \end{pmatrix} = -1\begin{pmatrix} x \\ x \end{pmatrix} = \begin{pmatrix} -x \\ -y \end{pmatrix} \tag{6.8.98}$$

This impose that $y=-x$ and therefore that the eigenvector is equal to:

$$\begin{pmatrix} x \\ y=-x \end{pmatrix} = x\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \forall x \tag{6.8.99}$$

The previous eigenvectors written with the Dirac formalism (see section Relativistic Quantum Physics page 4729) give for $\lambda=1$:

$$|v\rangle = x\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} x \\ x \end{pmatrix} \tag{6.8.100}$$

with a norm (1 since we formalize to the unit):

$$\langle v|v\rangle = x^2+x^2=2x^2=1 \Rightarrow x=\pm\frac{1}{\sqrt{2}} \tag{6.8.101}$$

Remark

In the formalism of Dirac $\langle v|$ is the is named a "Bra" and $|v\rangle$ a "Ket".

This being only valid for components that are real numbers. The normalized eigenvector has therefore for expression:

$$|v\rangle = x = \pm\frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{6.8.102}$$

For $\lambda=-1$, we have:

$$|v\rangle = x\begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} x \\ -x \end{pmatrix} \tag{6.8.103}$$

and:

$$\langle v|v\rangle = x^2+(-x)^2=2x^2=1 \Rightarrow x=\pm\frac{1}{\sqrt{2}} \tag{6.8.104}$$

and the normalized eigenvector has thus for expression:

$$|v\rangle=\pm\frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\end{pmatrix}\tag{6.8.105}$$

Let us now determine the eigenvectors and eigenvalues associated to σ_y by following the same procedure.

So we have for the eigenvalues:

$$\det(\sigma_y-\lambda I_n)=\det\begin{bmatrix}0-\lambda & -\mathrm{i}\\ \mathrm{i} & 0-\lambda\end{bmatrix}=0\Rightarrow\lambda^2-(-\mathrm{i}\cdot\mathrm{i})=0\Rightarrow\lambda=\pm 1\tag{6.8.106}$$

The eigenvectors are determined as following:

$$\begin{pmatrix}0 & -\mathrm{i}\\ \mathrm{i} & 0\end{pmatrix}=\begin{pmatrix}x\\y\end{pmatrix}=y\begin{pmatrix}x\\y\end{pmatrix}\Rightarrow\begin{pmatrix}-\mathrm{i}y\\ \mathrm{i}x\end{pmatrix}=\lambda\begin{pmatrix}x\\y\end{pmatrix}\tag{6.8.107}$$

and therefore for $\lambda=1$:

$$\begin{pmatrix}-\mathrm{i}y\\ \mathrm{i}x\end{pmatrix}=\begin{pmatrix}x\\y\end{pmatrix}\Rightarrow x=\frac{y}{\mathrm{i}}=-\mathrm{i}y\tag{6.8.108}$$

The eigenvector is therefore:

$$\begin{pmatrix}x\\ y=\mathrm{i}x\end{pmatrix}=\begin{pmatrix}x\\ \mathrm{i}x\end{pmatrix}=x\begin{pmatrix}1\\ \mathrm{i}\end{pmatrix},\forall x\tag{6.8.109}$$

The associated norm:

$$\langle v|v\rangle=x^2(1,-\mathrm{i})\begin{pmatrix}1\\1\end{pmatrix}=x^2\left[1+(-\mathrm{i}\cdot\mathrm{i})\right]=2x^2=1\Rightarrow x=\pm\frac{1}{\sqrt{2}}\tag{6.8.110}$$

The normalized vector is therefore expressed as:

$$|v\rangle=x=\pm\frac{1}{\sqrt{2}}\begin{pmatrix}1\\ \mathrm{i}\end{pmatrix}\tag{6.8.111}$$

For $\lambda=-1$:

$$\left(\begin{array}{c}-\mathrm{i}y\\ \mathrm{i}x\end{array}\right)=-\left(\begin{array}{c}x\\ y\end{array}\right)=\left(\begin{array}{c}-x\\ -y\end{array}\right)\Rightarrow y=\frac{1}{\mathrm{i}}x=-\mathrm{i}x\tag{6.8.112}$$

The eigenvector is therefore:

$$\left(\begin{array}{c}x\\ y=-\mathrm{i}x\end{array}\right)=\left(\begin{array}{c}x\\ -\mathrm{i}x\end{array}\right)=x\left(\begin{array}{c}1\\ -\mathrm{i}\end{array}\right),\forall x\tag{6.8.113}$$

The associated norm:

$$\langle v\,|\,v\rangle=x^2[1,\mathrm{i}]\left(\begin{array}{c}1\\ -\mathrm{i}\end{array}\right)=x^2\left[1^2+(-\mathrm{i}\cdot\mathrm{i})\right]=2x^2=1\Rightarrow x=\pm\frac{1}{\sqrt{2}}\tag{6.8.114}$$

The normalized vector has then for expression:

$$|v\rangle = x = \frac{\pm 1}{\sqrt{2}}\begin{pmatrix} 1 \\ -\mathrm{i} \end{pmatrix} \tag{6.8.115}$$

Let us now determine the eigenvectors and eigenvalues associated to σ_z by doing the same again.

We have therefore:

$$\begin{aligned} &\det(\sigma_z - \lambda I_n) = \det\begin{bmatrix} 1-\lambda & 0 \\ 0 & -1-\lambda \end{bmatrix} = (1-\lambda)(-1-\lambda) = -(\lambda^2 - 1) = 0 \\ &\lambda^2 - 1 = 0 \Rightarrow \lambda = \pm 1 \end{aligned} \tag{6.8.116}$$

The eigenvectors are then for $\lambda = 1$:

$$\begin{pmatrix} x \\ -y \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} \tag{6.8.117}$$

Which makes problem to us for say anything ... the only possibility is to choose $y = 0$ and therefore:

$$\begin{pmatrix} x \\ 0 \end{pmatrix} = x\begin{pmatrix} 1 \\ 0 \end{pmatrix} \tag{6.8.118}$$

and the associated norm:

$$\langle v|v\rangle = x^2 + 0^2 = x^2 = 1 \Rightarrow x = \pm 1 \tag{6.8.119}$$

The normalized vector has then for expression:

$$|v\rangle = \pm\begin{pmatrix} 1 \\ 0 \end{pmatrix} \tag{6.8.120}$$

and for $\lambda = -1$ we will have the same choice to do by choosing this time $x = 0$ and therefore:

$$|v\rangle = \begin{pmatrix} 0 \\ -y \end{pmatrix} = -y\begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{6.8.121}$$

hence the associated norm:

$$\langle v|v\rangle = 0^2 + (-y)^2 = y^2 = 1 \Rightarrow y = \pm 1 \tag{6.8.122}$$

The normalized eigenvector has then for expression:

$$|v\rangle = \pm\begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{6.8.123}$$

Therefore the normalized eigenvectors of σ_z are on the directions of the Cartesian coordinate axes. It is for this particular reason that the eigenvectors of σ_z are denoted in quantum computing by:

$$|+\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad |-\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{6.8.124}$$

and the reader should also know that we write also:

$$\sigma_z|\pm\rangle = \sigma_3|\pm\rangle = \pm|\pm\rangle \tag{6.8.125}$$

~100 %

★★★★☆

18 votes, 84.44%

7

Analysis

Analysis is the rigorous formulation of calculus. (Wikipedia)

Contents

7.1 Functional Analysis . . . 2170
  7.1.1 Representations . . . 2171
    7.1.1.1 Tabular Representations . . . 2171
    7.1.1.2 Graphical Representations . . . 2177
    7.1.1.3 Analytical Representations . . . 2206
  7.1.2 Functions . . . 2209
    7.1.2.1 Limits and Continuity of functions . . . 2226
    7.1.2.2 Concavity/Convexity of a function . . . 2280
    7.1.2.3 Euler theorem for homogeneous functions . . . 2282
  7.1.3 Logarithms . . . 2284
  7.1.4 Convolutions . . . 2293
    7.1.4.1 Continuous and Discrete Linear Convolution Product . . . 2293
    7.1.4.2 Matrix Convolution . . . 2299
  7.1.5 Integral Transforms . . . 2303
    7.1.5.1 Fourier Transform . . . 2305
    7.1.5.2 Laplace Transform . . . 2336
    7.1.5.3 $\mathcal{Z}$-Transform . . . 2360
    7.1.5.4 Hilbert Transform . . . 2376
  7.1.6 Functional dot product (inner product) . . . 2379
    7.1.6.1 Cauchy-Schwarz inequality for integrals . . . 2388
7.2 Complex Analysis . . . 2390
  7.2.1 Linear Applications . . . 2390
  7.2.2 Holomorphic Functions . . . 2399

7.2.2.1 Orthogonality of real and imaginary iso-curves . . . 2405
7.2.3 Complex Logarithm . . . 2407
7.2.4 Complex Integral Calculus . . . 2411
7.2.4.1 Convergence of a complex series . . . 2418
7.2.5 Path Decomposition . . . 2427
7.2.5.1 Inverse Path . . . 2428
7.2.6 Laurent Series . . . 2430
7.2.7 Singularities . . . 2437
7.2.8 Residue Theorem . . . 2440
7.2.8.1 Pole at infinity . . . 2444
7.3 Topology . . . 2446
7.3.1 General Topology . . . 2447
7.3.1.1 Topological Spaces . . . 2448
7.3.2 Metric Space and Distance . . . 2450
7.3.2.1 Equivalent Distances . . . 2457
7.3.2.2 Lipschitz Functions . . . 2457
7.3.2.3 Continuity and Uniform Continuity . . . 2460
7.3.3 Opened and Closed Set . . . 2463
7.3.3.1 Balls . . . 2464
7.3.3.2 Partitions . . . 2467
7.3.3.3 Formal Ball . . . 2470
7.3.3.4 Diameter . . . 2470
7.3.4 Varieties . . . 2473
7.3.4.1 Subvariety . . . 2475
7.3.4.2 Surfaces Homeomorphism . . . 2476
7.3.4.3 Differential Varieties . . . 2479
7.4 Measure Theory . . . 2481
7.4.1 Measurable Spaces . . . 2482
7.4.1.1 Monotone Classes . . . 2491
7.4.2 Lebesgue measure in probabilities . . . 2493
7.4.2.1 Lebesgue Measure of the rational and irrational numbers . . . 2496

7.1. Functional Analysis

FUNCTIONAL ANALYSIS is the branch of mathematics and specifically of the analysis that is related to the study of function spaces. It takes its historical roots E as neutral "additive" elementin the study of transformations such as the Fourier transform and in the study of differential equations and integrals. As such it encompasses so many areas that it is difficult to justify that it be a section of this textbook because it is rather a field of study. Moreover, it is because of this difficulty to accurately identify the area it covers that the reader will find the Fundamental Theorem of Calculus in the chapter of Integral and Differential Calculus rather than here...

Why do we use the term "analysis" in the particular case of functions? The reason lies in the historical study of various phenomena of nature and resolution of various technical problems and therefore mathematics, which often lead us to consider the variation of a parameter correlated with the variation of another or several other variables. To study these variations, many tools are available to each of us:

- The engineer, for example, frequently use charts (in cartesian, polar or logarithmic coordinate system... concepts which are discussed further in more detail) to determine the mathematical relations (or "law") linking variables between them. Certainly, this kind of method is (sometimes ...) aesthetic but students know well how it is sometimes painful to transcribe measures points on a sheet of paper or on a computer and consultants know how dangerous can be a chart when not build in a scientific way. This is unfortunately a necessary step (but should avoid an abusive usage) to understand how our predecessors worked and got the results that help us today in our advances in theoretical physics.
- The mathematician and theoretical physicist usually hate to use the paper-pencil-scrawl methods. Nevertheless, the role of the mathematician or physicist is to develop new theories with mathematical axioms or principles which should require no usage of graphical representation nor access to experimental measures that are often attached to it.

Remark

Before starting to read what follows, it may be useful to remind the reader that the definition of the concept of "function" (and the basic properties thereto) is given in the section on Set Theory at page 403.

Function analysis is also strongly linked to Vector Calculus (and not only...). Thus for people who want to increase their knowledge about the fundamentals of function analysis we strongly recommend the reader to have a look to the Vector Calculus section at page 1794.

7.1.1 Representations

We will see in what follows, firstly, how to represent different values related by tables and charts (yes! We must because it helps to understand more complicated stuff) and secondly how to mathematically analyse the properties of these representations only by using abstract mathematical tools.

Definition (#7.1.267): A function is named "univalent function" or "unary function" if the number of its arguments (parameters or variables) is equal to one. In the case of a function of two arguments, we speak about a "bivalent function" or "bivalent function", etc. Formally a function is n-ary if:

$$f: A_1 \times A_2 \times \ldots \times A_n \mapsto B \tag{7.1.1}$$

7.1.1.1 Tabular Representations

Among the possible visual representation of functions, the most intuitive and the oldest is the one where we have in the columns or the rows of a table in an orderly way the values of the independent variable $x_1, x_2, \ldots, x_n$ and the corresponding values, namely the "transformed variables" of the function $y_1, y_2, \ldots, y_n$ in another column or aligned row:

x	$y = f(x)$
x_1	$y_1 = f(x_1)$
x_2	$y_2 = f(x_2)$
...	...
x_n	$y_n = f(x_n)$

In the expression:

$$y = f(a_1, a_2, \ldots, a_n) \tag{7.1.2}$$

we say that the $a_1, a_2, \ldots, a_n$ are the "arguments" of f and such a table where every variable/argument is stored in a column is named a "long format table".

Such are, for example, tables of trigonometric functions, logarithmic tables, etc. and during the experimental study of certain phenomena tables which express the existing functional dependence between the measured physical quantities such as the readings of the temperature of the air stored in a meteorological station during one day.

Of course, this concept can be generalized to any multivariate function regardless its definition domain.

However, this method is laborious and does not permit to directly see the behaviour of the function and therefore a simple and attractive visual analysis of its qualitative properties for humans. It still has the advantage of not requiring any special tools or advanced mathematics and easily tractable by computers[1].

Another type of tabular representation are "crossed-tables", also named "pivot-tables" or "wide format",where we have categorical/ordinal variables names in the first column AND in the first row:

	Express Passenger	**Ordinary Passenger**	**Total**
Arriva Trains Wales	3,079	830	3,909
Cross Country	22,865	63	22,928
London Midland	14,487	33,792	48,279
Virgin Trains	8,594		8,594
Total	49,025	34,685	83,710

And here another kind of such table but with multiple categorical column variables (the same can be done with by having multiple row variables):

	Express Passenger		**Ordinary Passenger**		**Total**	
	# Trains	**Max. Speed**	**# Trains**	**Max. Speed**	**# Trains**	**Max. Speed**
Arriva Trains Wales	3,079	90	830	90	3,909	90
Cross Country	22,865	125	63	100	22,928	125
London Midland	14,487	110	33,792	100	48,279	110
Virgin Trains	85,94	125			8,594	125
Total	49,025	125	34,685	100	83,710	125

We can find more complex versions of the above example, like the generic multi-dimensional crossed-table with subtotals visible at page 1045.

When we have a collection of repeated observations of the same subjects, taken from a given population, over some time, we speak then of "longitudinal table". Typically (the years are given based on the holocene calendar):

	January 12001	**February 12001**	**Total**
Arriva Trains Wales	3,079	830	3,909
Cross Country	22,865	63	22,928
London Midland	14,487	33,792	48,279
Virgin Trains	8,594		8,594
Total	49,025	34,685	83,710

We can have obviously also "longitudinal crossed tables", also named "longitudinal pivot tables":

---

[1] According to the US Open Data Act. of January 12019 (holocene calendar): «*machine readable data is a format that can be easily processed by a computer without human intervention while ensuring no semantic meaning is lost.*»

	January 12001		February 12001		Total	
	# Trains	Max. Speed	# Trains	Max. Speed	# Trains	Max. Speed
Arriva Trains Wales	3, 079	90	830	90	3, 909	90
Cross Country	22, 865	125	63	100	22, 928	125
London Midland	14, 487	110	33, 792	100	48, 279	110
Virgin Trains	8, 594	125			8, 594	125
Total	49, 025	125	34, 685	100	83, 710	125

But all crossed and longitudinal tables (i.e. "wide formats") above are not tractable in the field of science and engineering and therefore should never be used to store datas!!! So it is common to unpivot the columns to get a normalized version of the datas (including the dates at the ISO 8601 format and removing the totals and subtotals and all fancy formatting on numerical values). The previous table, for example, will once unpivoted become the following "long format" table (i.e. "flat table"):

Type	Date	# Trains	Max Speed
Arriva Trains Wales	12001-01-31	3079	90
Arriva Trains Wales	12001-02-28	830	90
Cross Country	12001-01-31	22865	125
Cross Country	12001-02-28	63	100
London Midland	12001-01-31	14487	110
London Midland	12001-02-28	33792	100
Virgin Trains	12001-01-31	8594	125
Virgin Trains	12001-02-28		

Table 7.1: **Example of transformation of a wide crossed table to a long format table (i.e. flat table)**

Remark

Even if the above table is tractable in the field of Science and Engineering (also of Business), it is however absolutely not optimal in a Computer Science storage point of view! The above table should be fractioned as a fact table in a snowflake schema (or star schema) having only relationships with dimension tables. For more see information about these concepts, see [126].

We name "Complete Disjunctive Table" (CDT), the matrix of size $n \times p$ defined by (see example just below):

$$K_{IJ} = k_{ij} \tag{7.1.3}$$

where:

$$k_{ij} = \begin{cases} 1 & \text{if the individual } i \text{ has the modality } j \\ 0 & \text{otherwise} \end{cases} \tag{7.1.4}$$

This is a fundamental tool (i.e. step) in the calculation for Multiple Correspondence Analysis (see section Numerical Methods page ??).

✍**Example:**

Here is an example of complete disjunctive table. From a data table with two category variables, with respectively two and three modalities, a complete disjunctive table is created containing five dummy variables:

	Q1	Q2
Obs. 1	A	C
Obs. 2	B	D
Obs. 3	B	E
Obs. 4	A	D

The complete disjunctive table K_{IJ} is then given by:

	Q1 $-A$	Q1 $-B$	Q2 $-C$	Q2 $-D$	Q2 $-E$
Obs. 1	1	0	1	0	0
Obs. 2	0	1	0	1	0
Obs. 3	0	1	0	0	1
Obs. 4	1	0	0	1	0

Now that the reader has seen the simple definition and an example, more formally let us assume that we have at our disposal $m>2$ qualitative variables $x^1,\ldots,x^m$ and that we observe n individuals, then for each $k\in\{1,\ldots,m\}$, the variable x^k can take c_k distinct values denoted by $1,\ldots,c_k$ and c denotes the total number of modes:

$$c=\sum_{k=1}^{m}c_k \tag{7.1.5}$$

For each $i\in\{1,\ldots,n\}$, $x_i^k\in\{1,\ldots,c_k\}$ is the value of the variable x^k observed for the i-th individual.

For each mode of each variable, we consider the associated indicator vector, i.e. for each $k\in\{1,\ldots,m\}$ and each $\ell\in\{1,\ldots,c_k\}$:

$$\mathbb{I}_{m\ell}=\begin{pmatrix}\mathbf{1}_{x_1^k=\ell}\\ \vdots\\ \mathbf{1}_{x_n^k=\ell}\end{pmatrix}\in\{0,1\}^n \tag{7.1.6}$$

If we put these vectors in the columns of a matrix T of size $n\times c$, you we get the complete disjunctive table:

$$K_{IJ}=[\mathbb{I}_{11}\cdots\mathbb{I}_{1c_1}\cdots\mathbb{I}_{21}\cdots\mathbb{I}_{2c_2}\cdots\mathbb{I}_{m1}\cdots\mathbb{I}_{mc_m}] \tag{7.1.7}$$

We name "Burt table" the matrix of size $p\times p$ defined by (see example just below):

$$B=b_{jj'} \tag{7.1.8}$$

where $b_{jj'}=\sum_{i=1}^{n}k_{ij}k_{ij'}$ (the k_{ij} are the components of the complete disjunctive table defined just above).

Therefore the Burt table can also be defined by K_{IJ}:

$$B=\{b_{jj'}\}_{jj'}=K_{IJ}^TK_{IJ}=\sum_{i=1}^{n}k_{ij}k_{ij'} \tag{7.1.9}$$

✍**Example:**

Using the previous example, we then get:

$$B=\{b_{jj'}\}_{jj'}=K_{IJ}^TK_{IJ}=\begin{bmatrix}2&0&1&1&0\\0&2&0&1&1\\1&0&1&0&0\\1&1&0&2&0\\0&1&0&0&1\end{bmatrix} \tag{7.1.10}$$

It is possible to see $B=\{b_{jj'}\}_{1\leqslant j,j'\leqslant p}$ as a matrix of $p\times p$ blocks such that:

- If $p\neq p'$ then $b_{jj'}$ is the contingency table relative to the variables x^j et $x^{j'}$.
- If $j=j'$ then b_{jj} is a diagonal matrix that contains the marginal totals of the variable x^j denoted by $n_1^j,\ldots,n_{c_j}^j$

✍**Example:**

Consider the following table:

	Gender (G)	Color (C)	Age (A)
1	Woman	Red	20
2	Man	Blue	25
3	Man	Green	30
4	Woman	Red	35

The corresponding complete disjunctive table K_{IJ} is therefore:

	G.Woman	G.Men	C.Blue	C.Red	C.Green	A.20	A.25	A.30	A.35
1	1	0	0	1	0	1	0	0	0
2	0	1	1	0	0	0	1	0	0
3	0	1	0	0	1	0	0	1	0
4	1	0	0	1	0	0	0	0	1

and the corresponding Burt table B:

	G.Woman	G.Man	C.Blue	C.Red	C.Green	A.20	A.25	A.30	A.35
G.Woman	2	0	0	2	0	1	0	0	1
G.Man	0	2	1	0	1	0	1	1	0
C.Blue	0	1	1	0	0	0	1	0	0
C.Red	2	0	0	2	0	1	0	0	1
C.Green	0	1	0	0	1	0	0	1	0
A.20	1	0	0	1	0	1	0	0	0
A.25	0	1	1	0	0	0	1	0	0
A.30	0	1	0	0	1	0	0	1	0
A.35	1	0	0	1	0	0	0	0	1

We name "Complete Disjunctive Table" (CDT), the matrix of size $n \times p$ defined by (see example just below):

$$K_{IJ} = k_{ij} \tag{7.1.11}$$

where:

$$k_{ij} = \begin{cases} 1 & \text{if the individual } i \text{ has the modality } j \\ 0 & \text{otherwise} \end{cases} \tag{7.1.12}$$

This is a fundamental tool (i.e. step) in the calculation for Multiple Correspondence Analysis (see section Numerical Methods page **??**).

✍**Example:**

Here is an example of complete disjunctive table. From a data table with two category variables, with respectively two and three modalities, a complete disjunctive table is created containing five dummy variables:

	Q1	Q2
Obs. 1	A	C
Obs. 2	B	D
Obs. 3	B	E
Obs. 4	A	D

The complete disjunctive table K_{IJ} is then given by:

	Q1 – A	Q1 – B	Q2 – C	Q2 – D	Q2 – E
Obs. 1	1	0	1	0	0
Obs. 2	0	1	0	1	0
Obs. 3	0	1	0	0	1
Obs. 4	1	0	0	1	0

7.1.1.2 Graphical Representations

The natural, relative, real or purely imaginary numbers (see section Numbers page 215) can all be represented as simply by points on a numerical infinite axis (straight line).

To this purpose, we choose on this axis:

1. A point O named "origin"
2. A positive direction, that we indicate by a horizontal arrow
3. A unit of measure (usually represented by small vertical lines: the "graduation")

Such that:

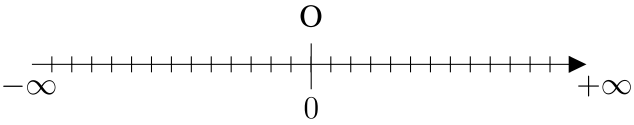


Figure 7.1: Typical representation example of an oriented infinite axis with origin

In most cases we put (traditionally) the axis horizontally and choose the direction from left to right (at least when there is only one axis...).

> **Remark**
>
> The point (letter) **O**, frequently represents the number zero in mathematics but we might very well choose to put the origin elsewhere. For example, in physics, the point **O** is often positioned at the location of the centroid of a system.

It is obvious that the fact that the sets of numbers that we discussed in the section Numbers are ordered implies that every number is represented by a single point on this axis. Thus, two distinct real numbers correspond two different points on the axis.

Thus, there is a correspondence between all numbers and all the points of the axis (in the case of real or complex numbers, it corresponds not a number to each graduation, but a number at each point of the axis!). Thus, each number represents a point or a unique graduation and back to each point or graduation is a single number which is the image.

7.1.1.2.1 2D representations

There are besides the one dimensional representations, other of higher dimensions (phew!...) like the "planar representation" that allow us to draw much more than simple points on a one-dimensional straight line but functions of one variable (but also points!). Let's see what this is and looks like:

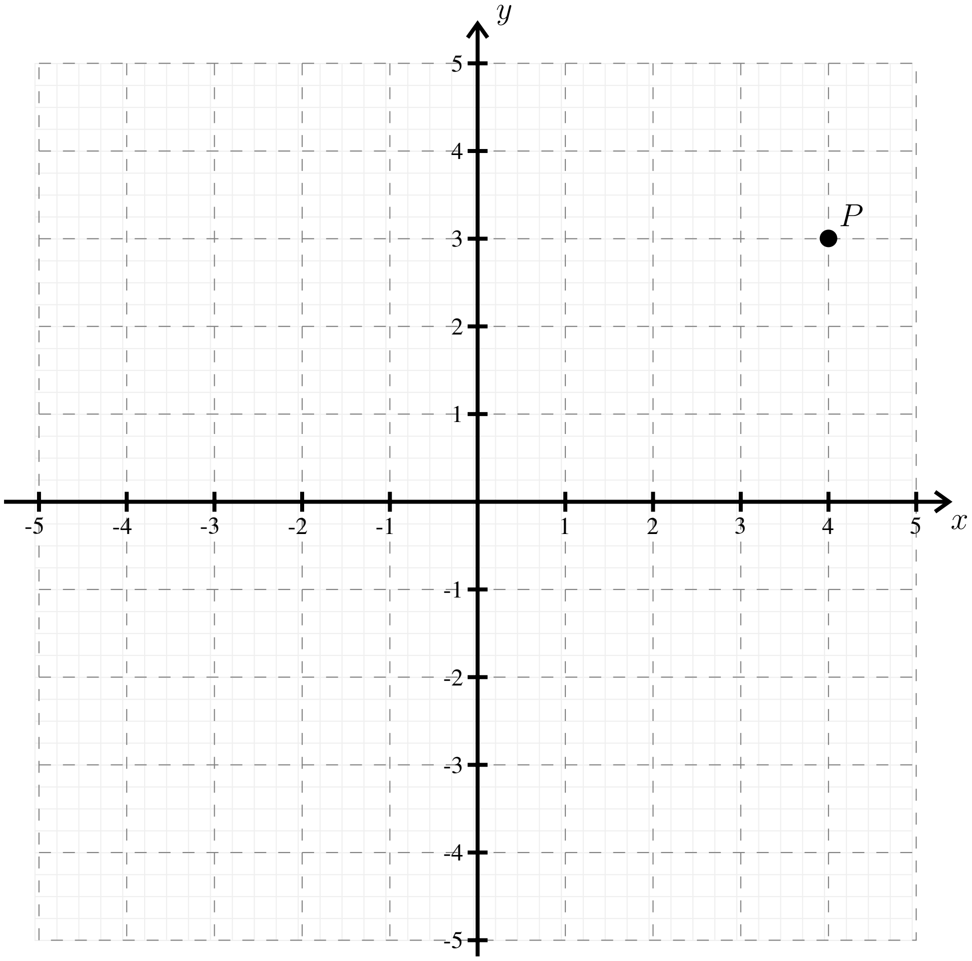


Figure 7.2: **Position of a point in a cartesian plane**

In the above figure we have added two graduation that helps us to identify uniquely the position of the point P given by $P(x,y)=P(+4,+3)=P(4,3)$. We then speak of "cartesian coordinates of a point".

In single-variable calculus, the functions that one encounters are functions of a variable (usually x or t) that varies over some subset of the real number line (which we denote by $\mathbb{R}$). For such a function, say, $y=f(x)$, the graph of the function f consists of the points:

$$(x,y)=(x,f(x)) \tag{7.1.13}$$

These points lie in the Euclidean plane, which, in the Cartesian or rectangular coordinate system, consists of all ordered pairs of real numbers (a,b). We use the word "Euclidean" to denote a system in which all the usual rules of Euclidean geometry hold (see section Euclidean Geometry page 2631). We denote the Euclidean plane by $\mathbb{R}^2$ where the exponent "2" represents the number of dimensions of the plane.

Thus we can for each of a variable x on a horizontal axis, named commonly "x-axis" match a value y through a function f such that:

$$y = f(x) \tag{7.1.14}$$

plotted on a vertical axis, named commonly the "y-axis" which passes through the junction defined by the origin O such tat (arbitrary example):

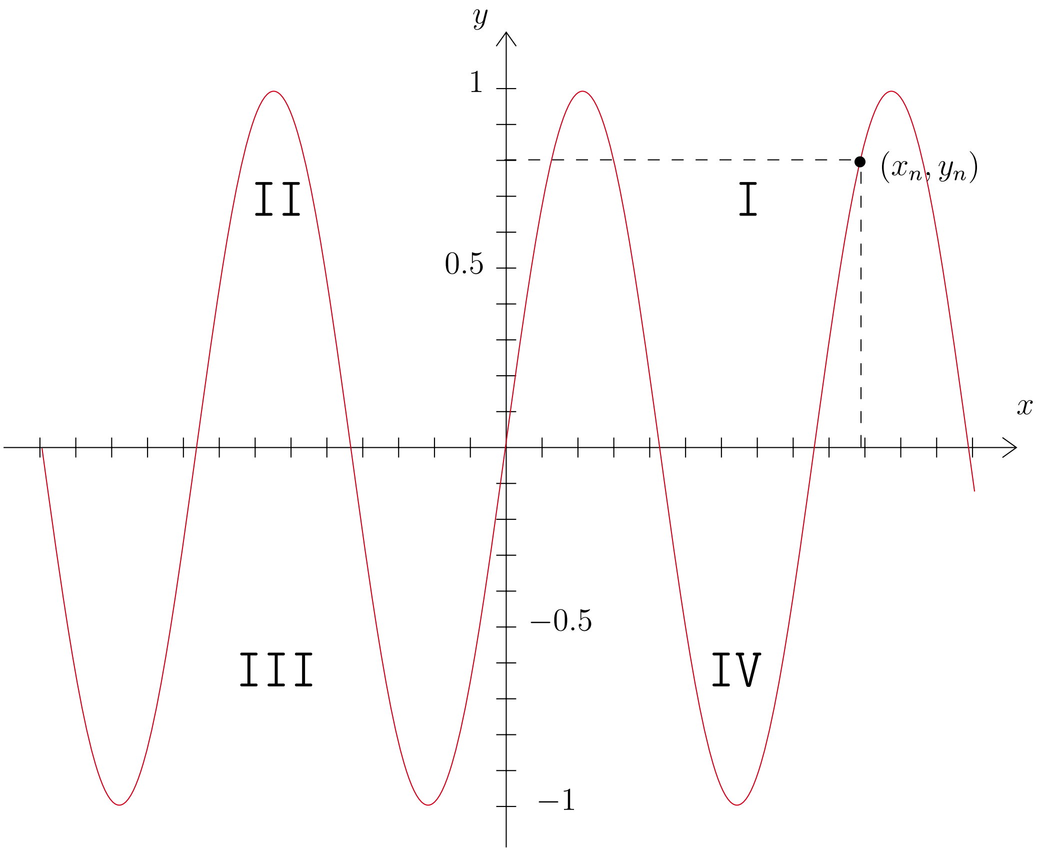
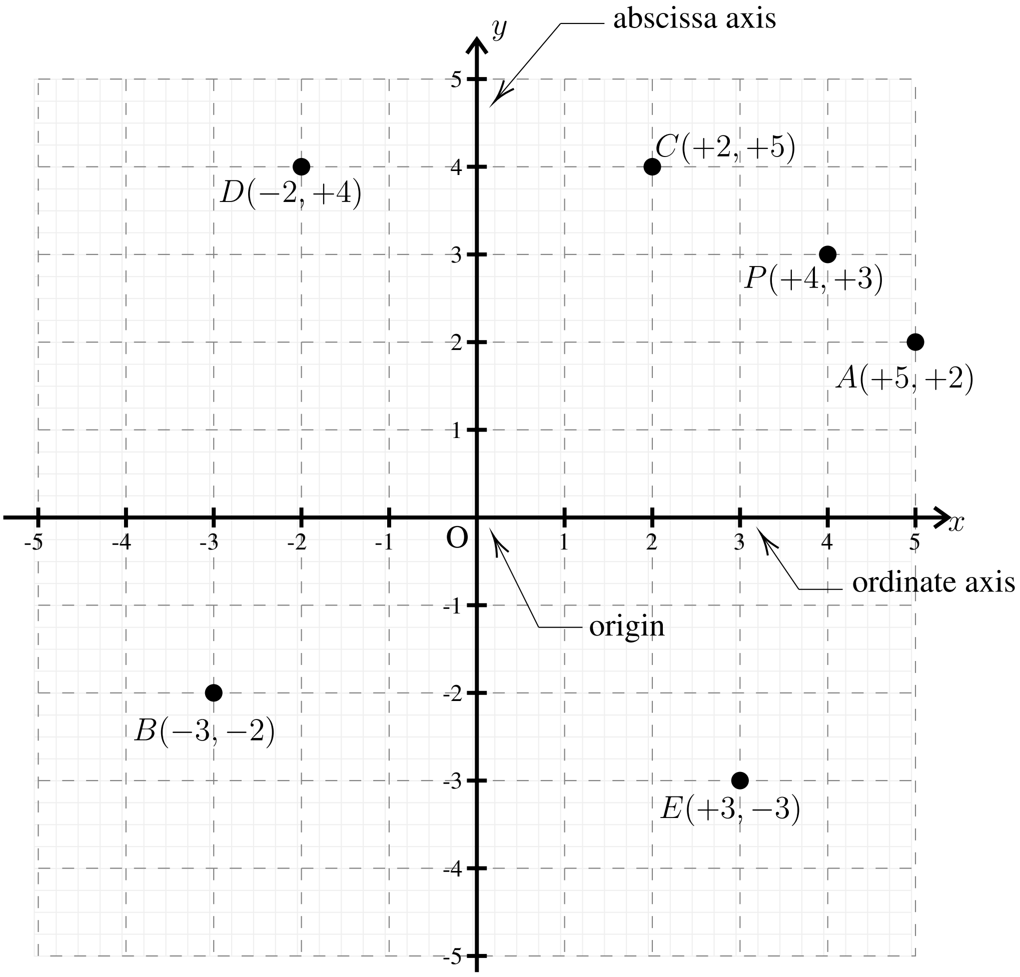


Figure 7.3: **Typical example of a planar representation with orthogonal axes, origin O and the 4 quadrants**

All points of the plane (that latter being denoted with the variations XOY, XY or xOy, Oxy, xy) have for "abscissa" traditionally the x-values corresponding to the independent variable (horizontal axes by tradition) of the function and for "ordinate" the corresponding value of the function (vertical axes by tradition). All these generated what is named the "planar graph" of the function. If there is no confusion, we just say "graph".



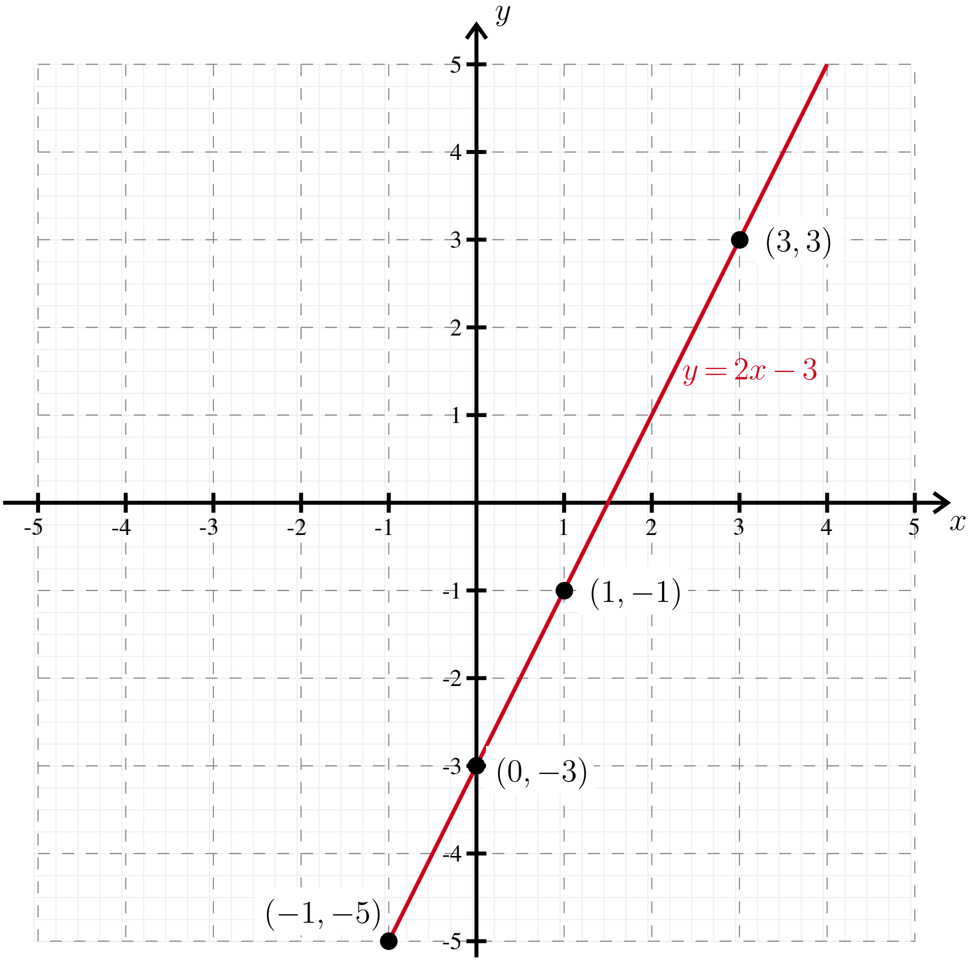
abscissa axis
y
x
$C(+2,+5)$
$D(-2,+4)$
$P(+4,+3)$
$A(+5,+2)$
$B(-3,-2)$
$E(+3,-3)$
O
origin
ordinate axis
5
4
3
2
1
-1
-2
-3
-4
-5
-5
-4
-3
-2
-1
1
2
3
4
5

Figure 7.4: **Vocabulary planar graph**

An interesting detailed example for middle schools students that may help is the planar representation of the following line equation:

$$y = f(x) = 2x - 3 \quad (7.1.15)$$

That gives in the range $x \in [-3.3, +3.3]$ the following graph:



with its equivalent tabular representation for some arbitrary chosen points:

x	$2x-3$	**Point** $(x,2x-3)$
-1	$2\cdot(-1)-3=-5$	$(-1,-5)$
0	$2\cdot(0)-3=-3$	$(-1,-5)$
$+1$	$2\cdot(+1)-3=-1$	$(-1,-5)$
$+3$	$2\cdot(+3)-3=+3$	$(-1,-5)$

Table 7.2: **Values and corresponding transformed variables**

In the case of representation by a rectangular coordinate system (cartesian, polar or logarithmic) as the figure above, we can see that the entire coordinate plane is divided into four areas that by tradition we name "quadrants" as already mentioned just earlier.

> **Remark**
>
> When we wish to highlight a particular point on the graph representing the function, we draw most of time a small round as presented in the prior-previous figure for the point of coordinates (x_n,y_n).

Another classic case of plane graph representation known by a large number of students is the plot of polynomials (see section Calculus page 1368) with real coefficients or trigonometric functions (see section Trigonometry page 2503).

Indeed, to solve polynomial equations of the second degree (see section Calculus page 1382), it is common in small classes that the teacher asks his students in addition to give an algebraic expression of the roots of:

$$y=f(x)=ax^2+bx+c=0 \tag{7.1.16}$$

given by for recall (see section Calculus page 1383 for the proof):

$$x_{1,2}=\frac{-b\pm\sqrt{b^2-4ac}}{2a} \tag{7.1.17}$$

a graphics resolution where the two roots (in the case where there are two distinct real roots) are given by the intersection of the parabola with the x-axis (of course, if the equation has no solution, there are no intersections...):

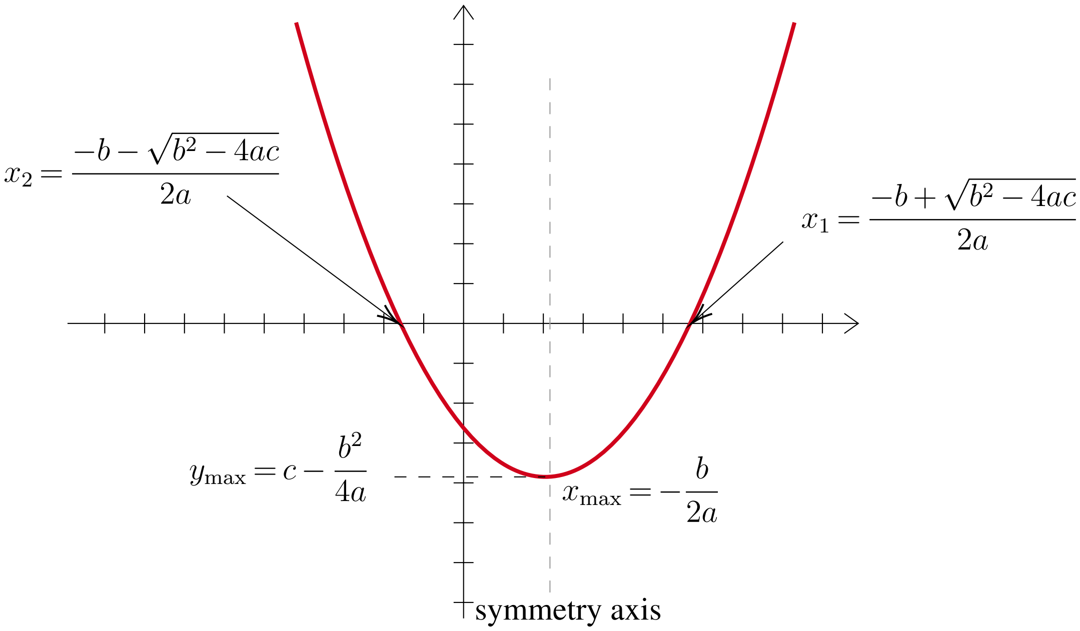


Figure 7.5: **Representation of roots on a planar graph**

The graphical representation can be generalized to polynomial equations of the 3rd, 4th and 5th degree (we will prove much further, using Galois theory that it is not possible to get a general algebraic expression of the roots of a polynomial equation of the 5th degree and higher).

There is another well-known and interesting example of special planar graph because when most young people think that after high-school they will never do maths again, in Switzerland many employees are faced to calculate in spreadsheet softwares (like Microsoft® Excel®) what we name the "coordinate wage" (CW) that is a "step-wise function" defined in year 12013 (holocene calendar) by the swiss government as:

$$\text{CW}:=\begin{cases} 0 & S<\frac{3}{4}R \\ \frac{R}{8} & \frac{3}{4}R\leq S<R \\ S-\frac{7}{8}R & R<S<3R \\ \frac{17}{8}R & 3R\leq S \end{cases} \tag{7.1.18}$$

where R is a minimal value defined also by the swiss government as being equal to $25,800.-$ in 12013 (holocene calendar) and the wage is denoted by the letter S (for **S**alary).

When we plot such a stepwise function with for example Maple™ 4.00b we get:

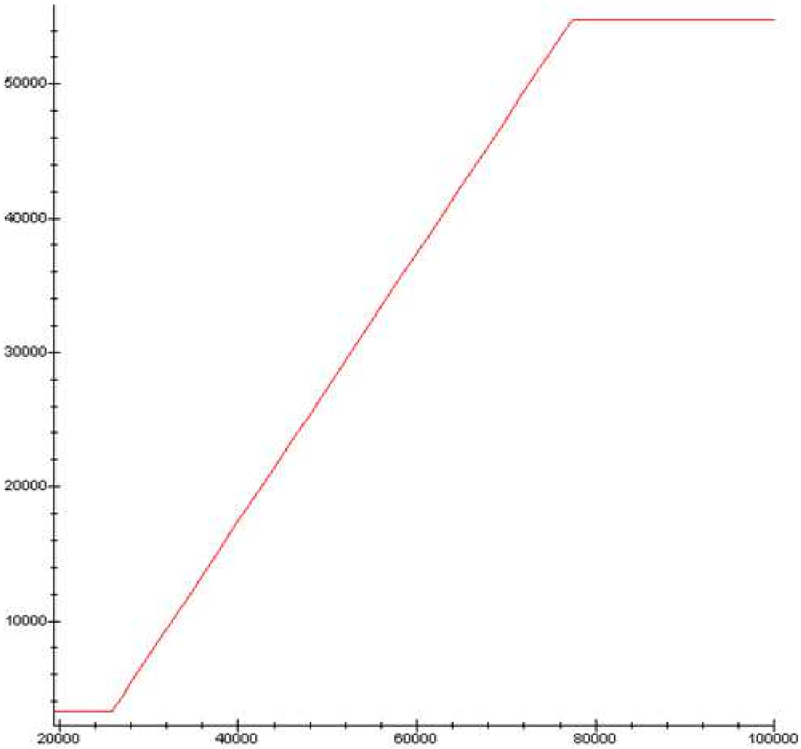


Figure 7.6: Example of step-wise function for swiss coordinate wage in Maple™ 4.00b

And therefore it is obvious thanks to this chart representation that the previous definition can be simplified as:

$$\text{CW} := \begin{cases} 0 & S < \dfrac{3}{4}R \\ \min\left\{\max\left[\dfrac{R}{8}, S - \dfrac{7}{8}R\right], \dfrac{17}{8}R\right\} & S \geq \dfrac{3}{4}R \end{cases} \tag{7.1.19}$$

That is much easier to write in any spreadsheet software or also in Maple™ 4.00b:

```
>R:=258000;
>plot(min(max(R/8,S-7/8*R),17/8*R,S=3/4*R..100000);
```

Also, graphs are as we know powerful qualitative tools in the field of statistics (see section Statistics page 542) but also of data mining (see section Numerical Methods page 6768) as a starting point for data analysis (histograms, cheese, box plots, radar, scatter plots, etc.). The assumptions and ideas that are generated by graphical analysis can be investigated with advanced statistical tools (for a few hundred of examples see the `R` or MATLAB® softwares companion book).

Below for example, a graph (histogram) taken from the Industrial Engineering section that is very common in the field of statistics and project management in the global industry:

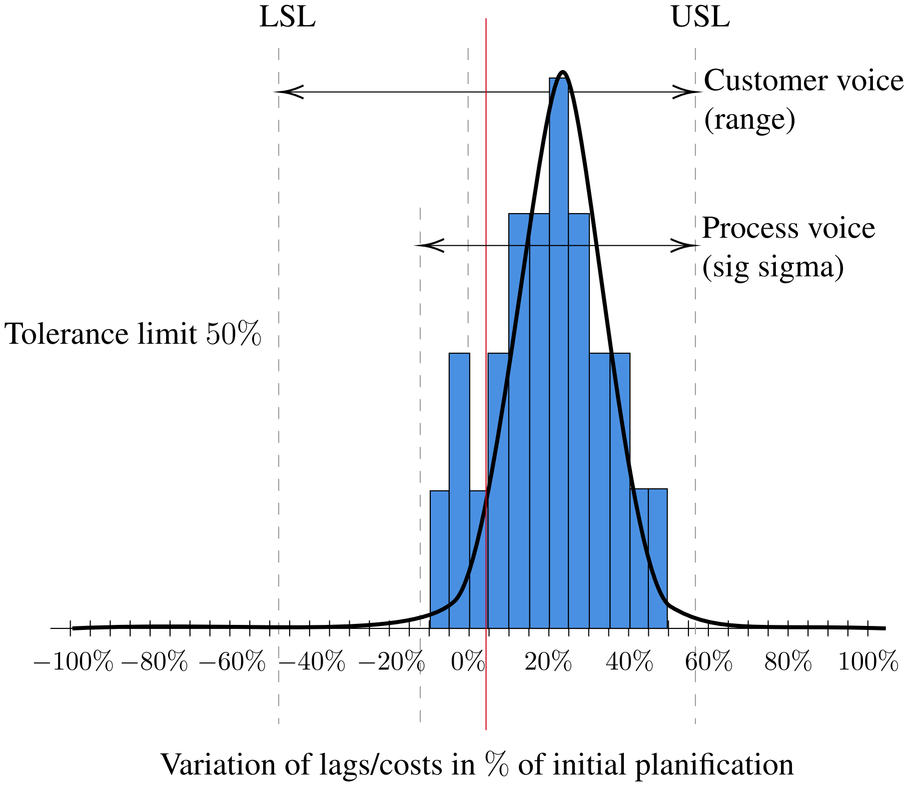


Figure 7.7: **Example of typical histogram in engineering companies (Six Sigma)**

Histograms allow to observe distributions and determine qualitatively if it fits a particular theoretical model.

Graphics can also be used to observe changes over time (time series, control charts, residual analysis, etc.):

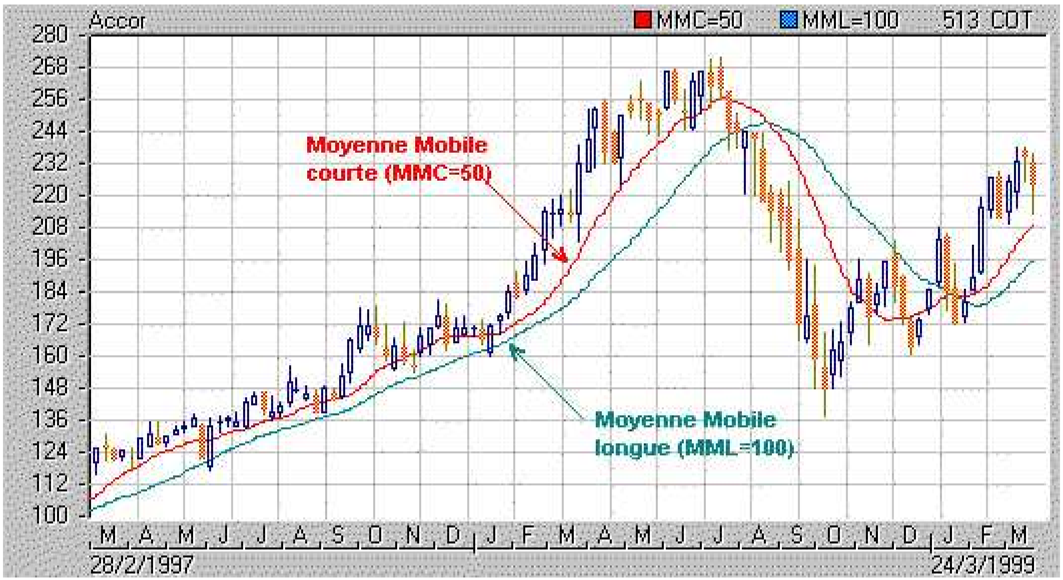
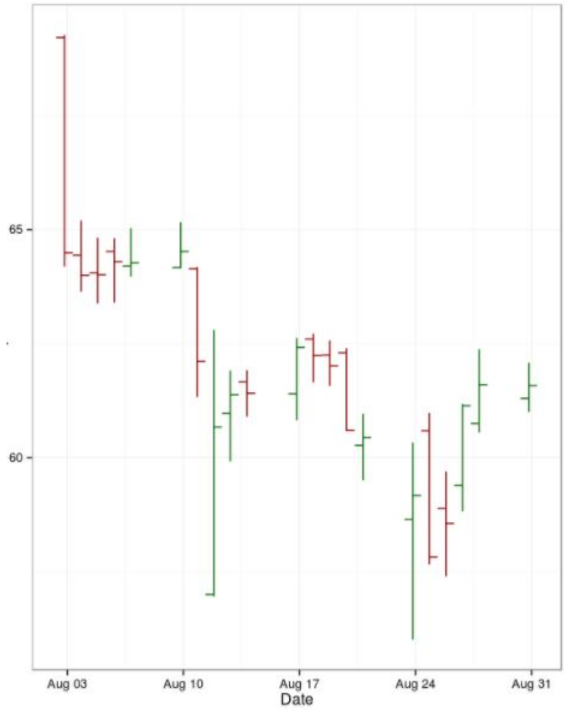
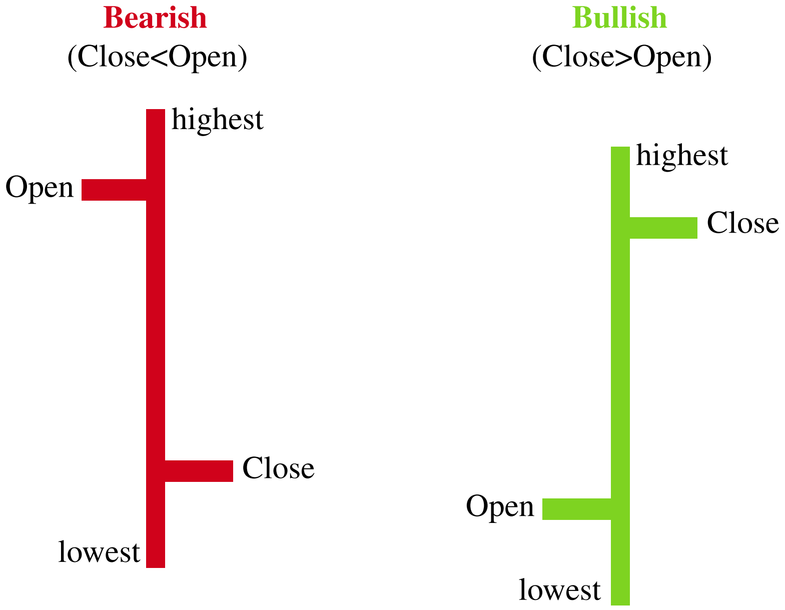


Figure 7.8: **Example of OHLC (candlestick plot) time series with moving averages in financial trading**

or another type of OHLC (Open-High-Low-Close) trading plot:



There are different rules for the colors of an OHLC plot! We can first define a color depending if the close price is lower than the open price of the same symbol/candlestick (the open price is always on the left and the closure price always on the right):


Bearish
(Close<Open)
highest
Open
Close
lowest
Bullish
(Close>Open)
highest
Close
Open
lowest

A common question related to this chart is: *Why doesn't a candlestick open when another one closes?*

This is because the close price is defined by the last tick of clock to arrive before the end of the candles time period, this may be several seconds before the end of the candles time period. The next candles open price is generated by the first tick of the clock to arrive after the beginning of its time period, this tick will reflect a change in price since the tick that generated the close price of the last candle, this first tick of the new candle may be several seconds after the beginning of the candles time period.

And still many other charts... that we have already seen and others that we will see throughout the pages of this textbook.

7.1.1.2.2 3D representations

Of course, in the case of a trivalent function (three-dimensional), that is to say a parameter which depends on two other, the idea is the same as for 2D except that the number of quadrants doubles:

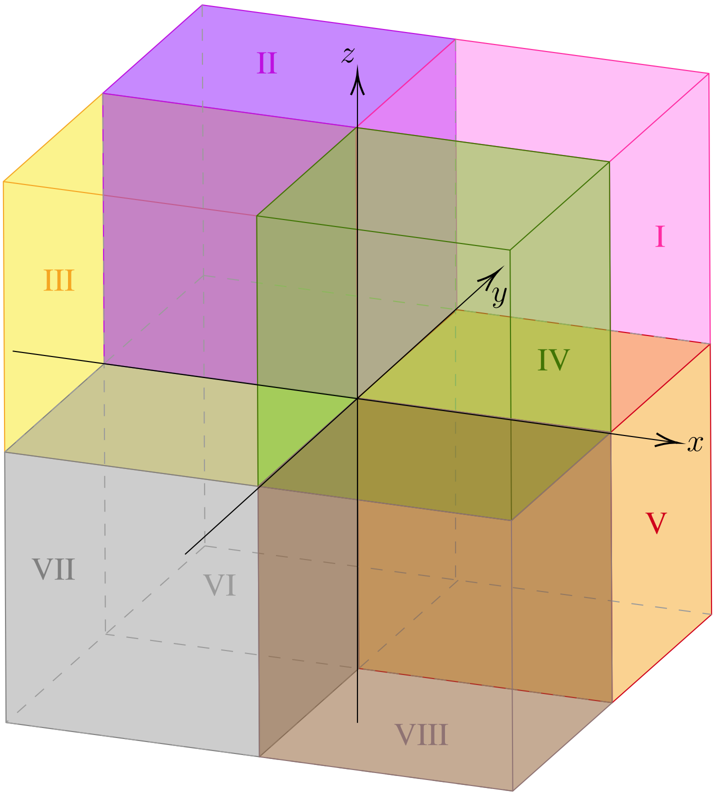


Figure 7.9: **Quadrants in a 3D orthogonal system (source: Wikipedia)**

This 3D method of representation and analysis of a trivalent function was time consuming at the beginning of the 120th century (holocene calendar) but with the help of computers in the end of the 120th century this time consuming problem was almost solved...

In 3D functional analysis, we will deal with functions of two or three variables (usually x, y, z, respectively). The graph of the arrow of coordinates $(x, y, z) = (x, y, f(x, y))$, lies in the Euclidean space. Since the Euclidean space can be 3-dimensional (and more or less for sure!), we denote it by $\mathbb{R}^3$.

Euclidean space has three mutually perpendicular coordinate axes (x, y and z), and three mutually perpendicular coordinate planes: the xy-plane, yz-plane and xz-plane.

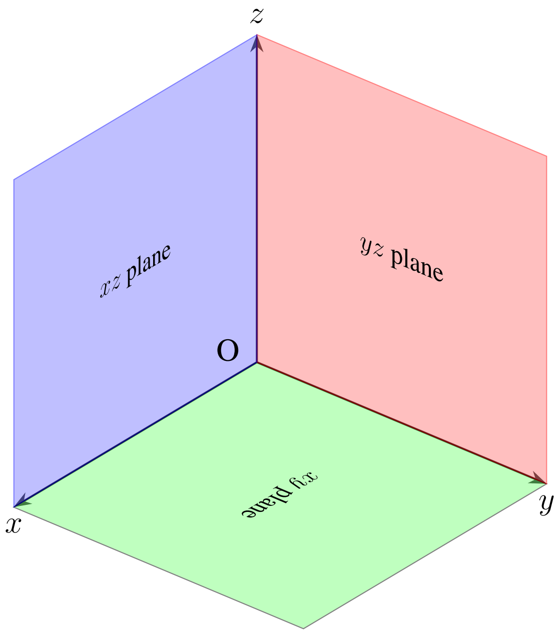


Figure 7.10: **Mutually perpendicular planes in $\mathbb{R}^3$**

The coordinate system shown above is known as a right-handed coordinate system, because it is possible, using the right hand, to point the index finger in the positive direction of the x-axis, the middle finger in the positive direction of the y-axis, and the thumb in the positive direction of the z-axis, as below:

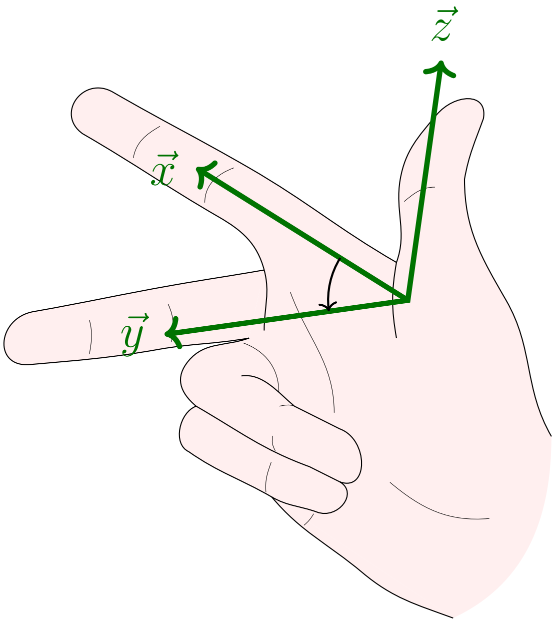


Figure 7.11: **Right-handed coordinate system (author: Izaak Neutelings)**

What we are going to represent now further below (special example), purists mathematicians would denote it as follows (it's nice to have seen at least once this notation as you could meet it in other books):

$$\begin{gathered}f:\mathbb{R}\times\mathbb{R}\mapsto\mathbb{R}\\ z=f(x,y)=\frac{12x}{1+x^2+y^2}\\ (x,y)\in[-10,+10]\times[-5,+5]\end{gathered}\tag{7.1.20}$$

and let us see what it gives with Maple™ 4.00b:

```
>restart:
>with(plots):
>f:=(x,y)->12*x/(1+x^ 2+y^ 2);
>xrange:=-10..10;yrange:=-5..5;
>plot3d(f,xrange,yrange);
```

This will give:

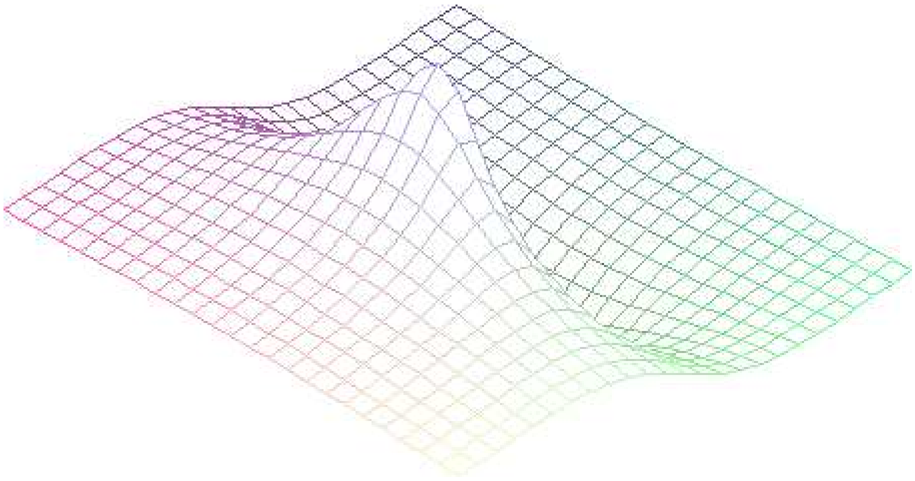


Figure 7.12: **Grid representation of a 3D function in Maple™ 4.00b**

Let us improve the visual by adding a shading interpolation color with warm color to high positions and cold colors to low positions:

```
>plot3d(f,xrange,yrange, style=patchnogrid, grid=[80,50], shading=ZHUE,
axes=FRAME, tickmarks=[3,3,3], labels=[‘x‘,‘y‘,‘f(x,y)‘],
labelfont=[TIMES,BOLD,12], title=‘Graphique rempli‘,
titlefont=[TIMES,BOLD,12], scaling=unconstrained, orientation=[-107,68]);
```

This will give:

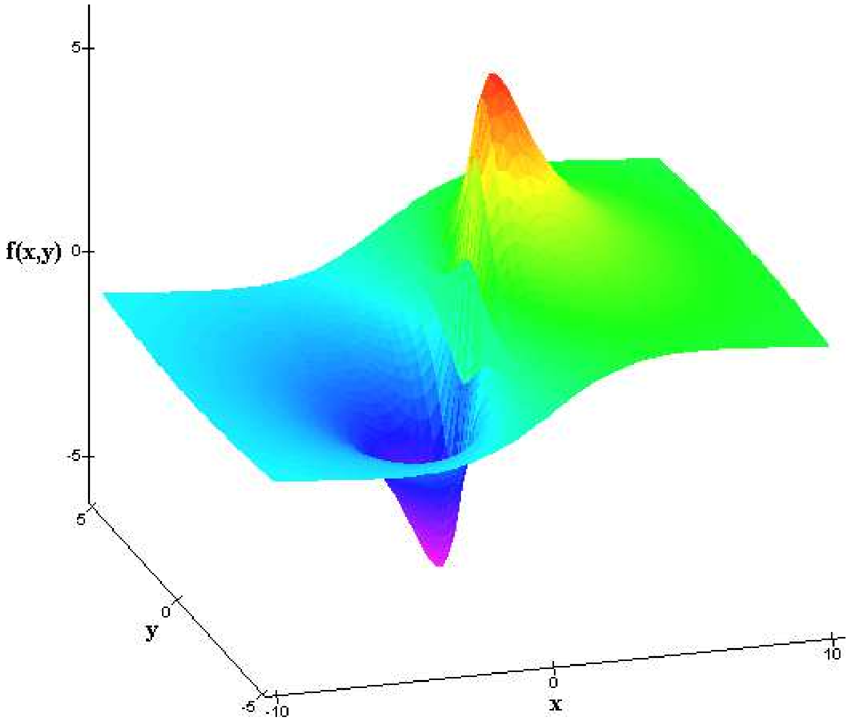


Figure 7.13: **Shading interpolation representation of a 3D function in Maple™ 4.00b**

Let us plot now the "contour lines", also named "isoline" (or "isoquant in Econometry), that represents lines of the same height on the function surface[2] (see section of Differential Geometry page 2857 for a rigorous definition):

```
>plot3d(f,xrange,yrange,style=patchcontour);
```

This will give:

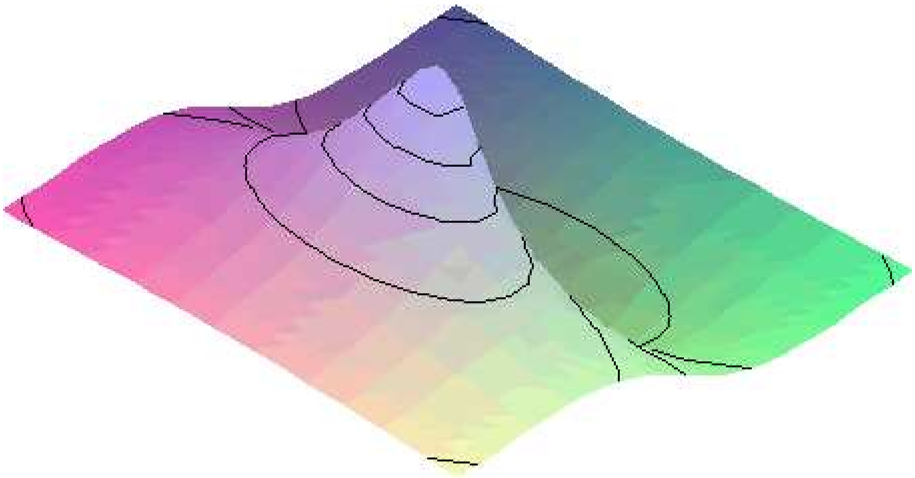


Figure 7.14: **Isolines representation of a 3D function in Maple™ 4.00b**

---

[2] It is a cross-section of the three-dimensional graph of the function $f(x, y)$ parallel to the x, y plane. In cartography, a contour line (often just named a "contour") joins points of equal elevation (height) above a given level, such as mean sea level. A contour map is a map illustrated with contour lines, for example a topographic map, which thus shows valleys and hills, and the steepness of slopes. The contour interval of a contour map is the difference in elevation between successive contour lines.

It's not very nice so let us improve this a little bit:

```
>plot3d(f,xrange,yrange,style=patchcontour,contours=[seq(-7+k/4,k=0..60)],
grid=[80,50],shading=ZHUE,axes=FRAME, tickmarks=[3,3,3],
scaling=unconstrained,orientation=[-107,68]);
```

This will give:

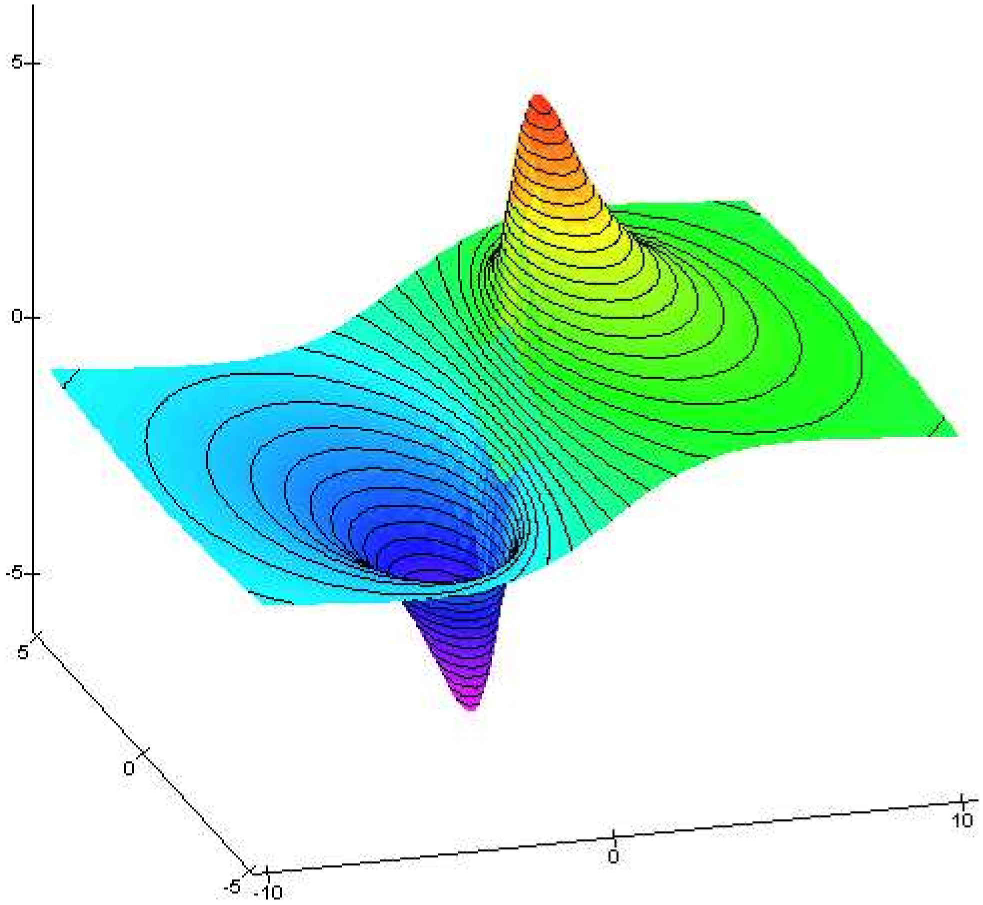


Figure 7.15: **Better representation of a 3D function in Maple™ 4.00b**

With a small rotation to view from above:

```
>plot3d(f,xrange,yrange, style=patchcontour, contours=[seq(-7+k/4,k=0..60)],
grid=[80,50], shading=ZHUE, axes=FRAME, tickmarks=[3,3,3],
scaling=unconstrained, orientation=[-90,0]);
```

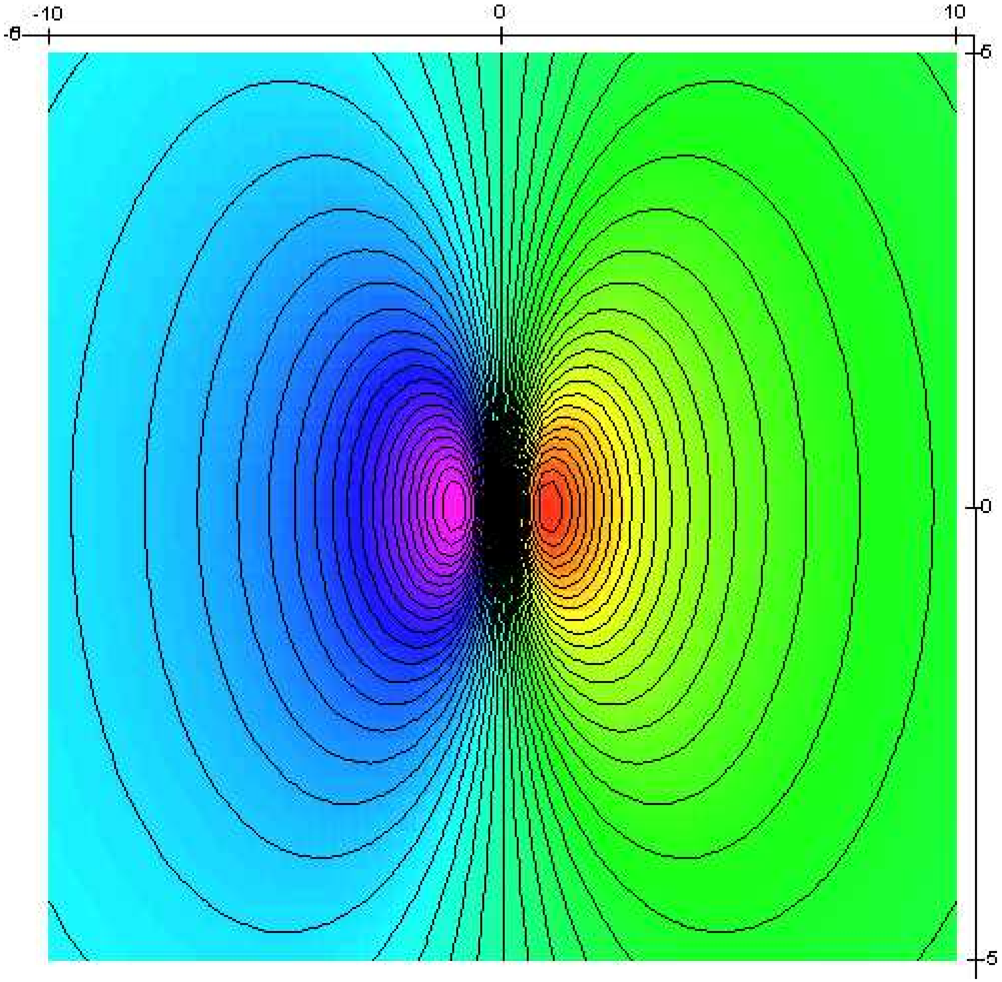


Figure 7.16: **Above representation of a 3D function in Maple™ 4.00b**

And in section view (side view):

```
>plot(f(x,2),x=xrange);
```

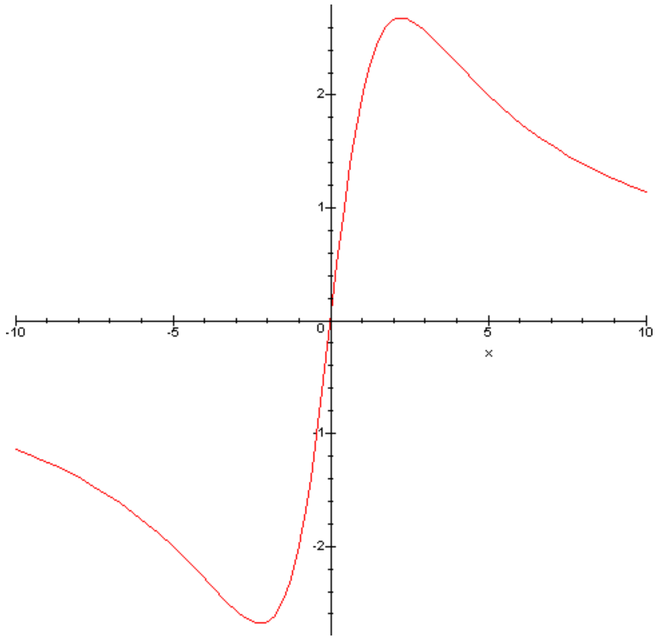


Figure 7.17: **Representation of a section of the pseudo-3D surface**

Or with multiple sections views:

```
>display([seq(plot(f(x,y),x=xrange),y=yrange)]);
```

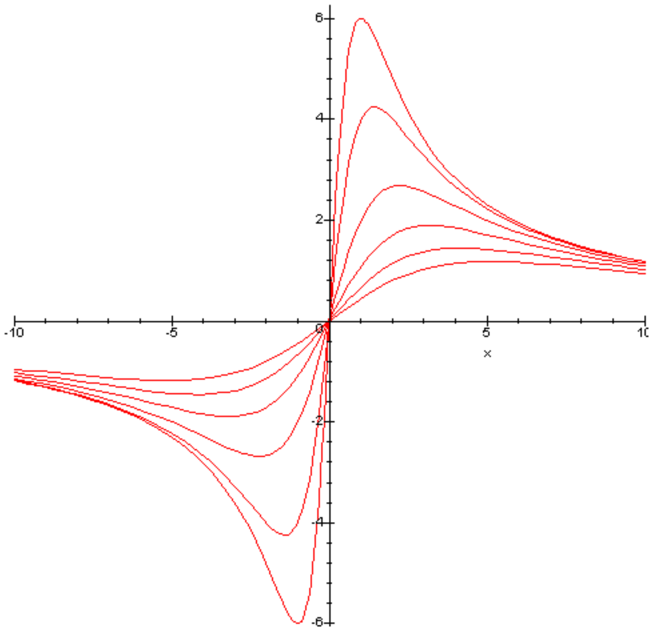


Figure 7.18: **Representation of multiple sections of the pseudo-3D surface**

The reader can also animate the graph above with the following command:

```
>display([seq(display([plot(f(x,k/5),x=xrange),
textplot([6,5,cat('y=',convert(evalf(k/5,2),string))],font=[TIMES,BOLD,16])])
,k=-25..25)],insequence=true, title='Animation',titlefont=[TIMES,BOLD,18]);
```

That's all for this typical and simple example of standard manipulations of an engineer hired in a company and using graphics (in practice it will instead use MATLAB® instead of Maple™ but the reader can refer to the free companion book on MATLAB® that has a few hundreds of pages about graphics).

7.1.1.2.3 2D Vector representations

It is also frequently made use of graphic representations in the context of analytical geometry to simplify analysis or to prove theorems with the help of visual representations (do not abuse of this method!).

Thus, we can easily introduce the concept of "norm" (see section Vector Calculus page 1818) in a very easy way by plotting the distance between two points (in 2D or in 3D) and applying the Pythagorean theorem that will be assumed to be known (see section Euclidean Geometry page 2609).

The main idea of a planar vector representation in physics and engineering labs is that a point P_1 of coordinates (x_1, y_1) that has some physical properties (typically a velocity) will be after a given time at the point P_2 of coordinates (x_2, y_2) supposed to be always in the same plane. In this way, the straight line between P_1 and P_2 is a visualization of the "intensity" of the velocity (and implicitly of the force). When doing that for many points we get a planar representation of a planar vector field (for more example see the companion book on MATLAB®):

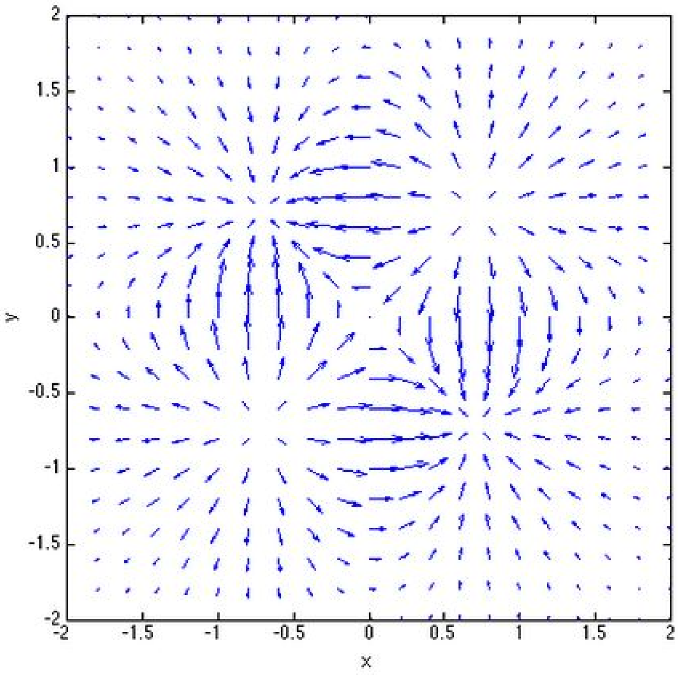


Figure 7.19: **Typical planar vector field with MATLAB®**

Now let us represent three points P_1, P_2, P_3 on a planar graph in which has been defined a referential as presented below:

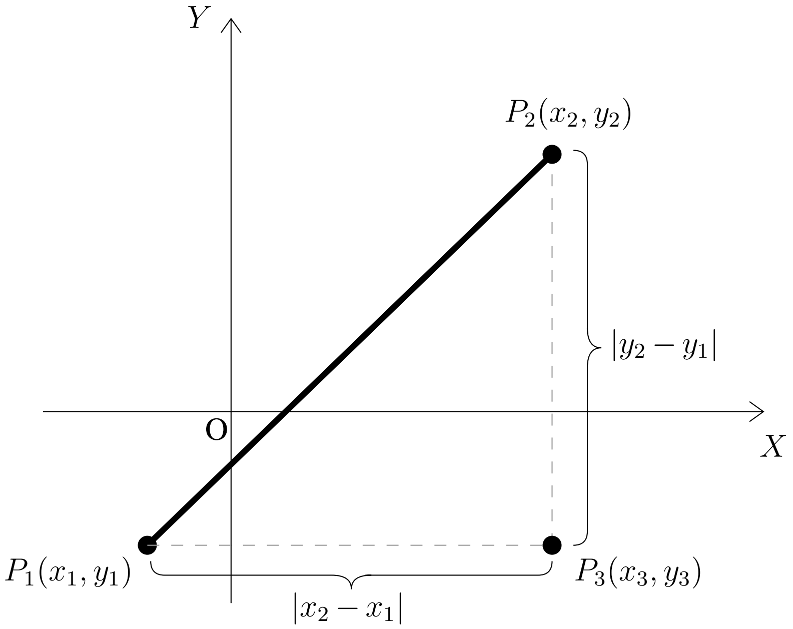


Figure 7.20: **Scenario of thee points in a plane**

We can consider the straight line $\overline{P_1P_2}$ as a vector but not translated at the origin O of the referential.

If $x_1 \neq x_2$ and $y_1 \neq y_2$ (as in the figure above), the points P_1, P_2, P_3 are the vertices of a perpendicular triangle. By applying the Pythagorean theorem (see section Euclidean Geometry page 2609) we can easily calculate the metric distance d between the points P_1 and P_2 as being equal to:

$$[d(P_1,P_2)]^2=[d(P_1,P_3)]^2+[d(P_3,P_2)]^2 \tag{7.1.21}$$

On the figure, we see that:

$$d(P_1,P_3)=|x_2-x_1| \quad \text{and} \quad d(P_3,P_2)=|y_2-y_1| \tag{7.1.22}$$

Since $\forall x\in\mathbb{R}\ |x|^2=x^2$, we can write:

$$d(P_1,P_2)=\sqrt{[d(P_1,P_3)]^2+[d(P_3,P_2)]^2}=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2} \tag{7.1.23}$$

If $x_1=y_1=0$, we end up with a relation named "norm", "module" or "distance" that we have already defined as part of our study of Vector Calculus when the origin of the vector is translated on the origin of the referential (see section of the corresponding name page 1818):

$$d(P_1,P_2)=\sqrt{x_2^2+y_2^2}=\|\vec{v}\| \tag{7.1.24}$$

Theorem 7.1. Obviously, if we consider two points $P_1(x_1,y_1)$, $P_2(x_2,y_2)$, we can determine if a third point $P_3(x_3,y_3)$ is on the mediator (see section Euclidean Geometry page 2597) of the first two and for this that it is obviously sufficient that (by definition of the mediator!):

$$d(P_1,P_3)=d(P_2,P_3) \tag{7.1.25}$$

Remark

We hesitated to put this proof in the section of Analytical Geometry but at then end we have decided that it was a nice example of showing how visual representation can help readers to better understand some subjects.

Proof 7.1.1. As $(x_1,y_1),(x_2,y_2)$ are known, we can easily express an "analytic expression" property of the mediator that is that for each point on the mediator we have:

$$\begin{aligned}
&\sqrt{(x_1-x_3)^2+(y_1-y_3)^2}=\sqrt{(x_2-x_3)^2+(y_2-y_3)^2}\\
&(x_1-x_3)^2+(y_1-y_3)^2=(x_2-x_3)^2+(y_2-y_3)^2\\
&x_1^2-2x_1x_3+\cancel{x_3^2}+y_1^2-2y_1y_3+\cancel{y_3^2}=x_2^2-2x_2x_3^2+\cancel{x_3^2}+y_2^2-2y_2y_3+\cancel{y_3^2}\\
&2y_2y_3-2y_1y_3=x_2^2-2x_2x_3+y_2^2-y_1^2+2x_1x_3-x_1^2\\
&y_32(y_2-y_1)=2(x_1-x_2)x_3+(x_2^2-x_1^2+y_2^2-y_1^2)\\
&y_3=\frac{x_1-x_2}{y_2-y_1}x_3+\frac{x_2^2-x_1^2+y_2^2-y_1^2}{2(y_2-y_1)}=ax_3+b=f(x_3)
\end{aligned} \tag{7.1.26}$$

where a,b are therefore constants and wherein any point that satisfies this relation, which is in this case the equation of a straight line, lies on the mediator.

■ Q.E.D.

Furthermore, it is easy to see that the midpoint of the line segment that coincide with the mediator is given by:

$$M=\left(\frac{x_1+x_2}{2},\frac{y_1+y_2}{2}\right) \tag{7.1.27}$$

So we see that with a simple visual representation, we can achieve results that are sometimes (...) more obvious for students.

Let us use this example to define some concepts on that we will come back further and do some reminders.

Definition (#7.1.268): Any function of the form of a polynomial (see section Calculus page 1368) of degree 1 with constant real coefficients:

$$y=f(x)=ax+b \tag{7.1.28}$$

is the analytic expression of what we name a "straight line" or "linear equation" of "slope" a and "intercept" b (when $x=0$).

Obviously, if:

$$y=f(x)=b \tag{7.1.29}$$

the line is horizontal if we graphically represent it since y is constant for all x and is equal to b. Conversely, if:

$$x=f(y)=a \tag{7.1.30}$$

the straight line will be vertical in the $x\text{O}y$ referential.

7.1.1.2.4 Properties of visual representations

Depending on the type of graph we visualize (especially graphics planes) it is possible to extract some basic properties. Let us see the most important one to know for univariate functions:

$$y = f(x) \tag{7.1.31}$$

P1. The graph of a function is "symmetrical about the y-axis" if the change in from x to $-x$ in the function does not change the value of y such that:

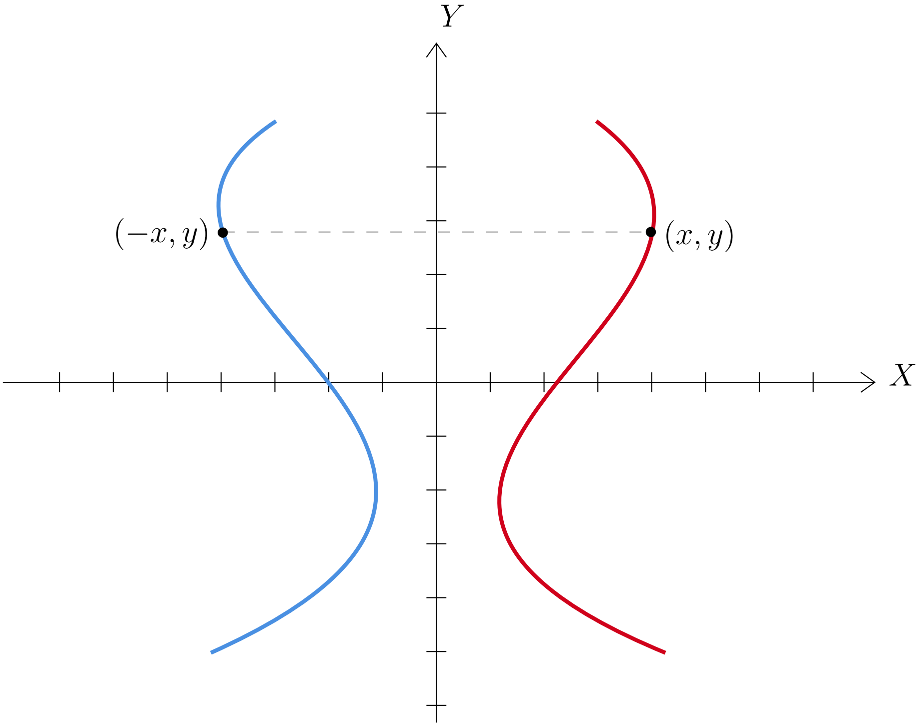


Figure 7.21: **Example of symmetry through the y-axis of a function**

P2. The graph of a function is "symmetrical about the x-axis" if the change from y to $-y$ does not change the value of x such that:

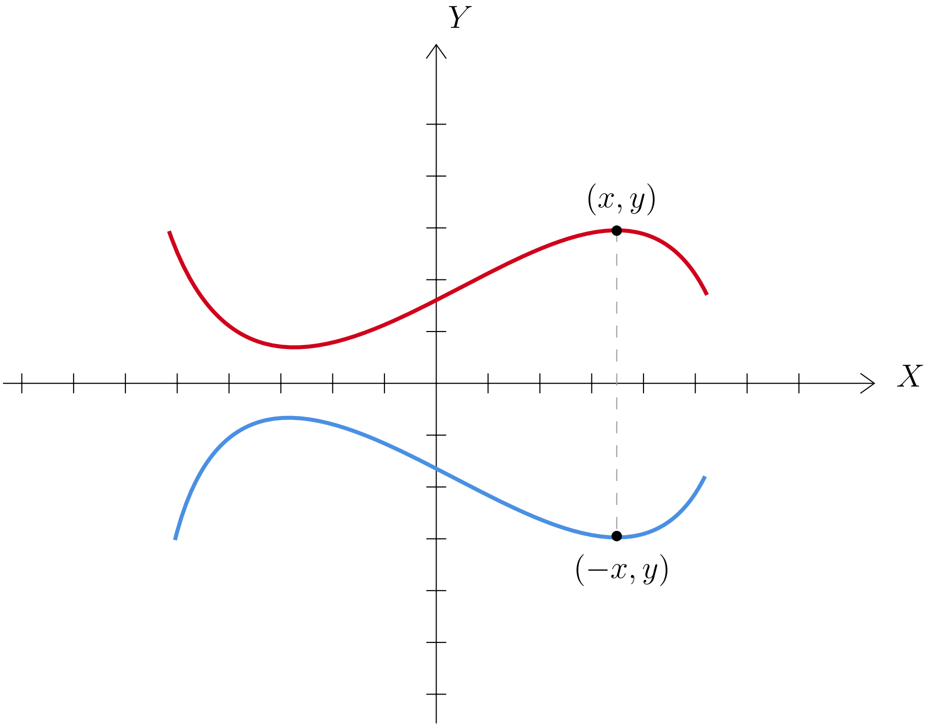


Figure 7.22: **Example of symmetry through the x-axis of a function**

P3. The graph of a function is "symmetrical about the origin O" if the simultaneous change of y to $-y$ and from x to $-x$ gives the following result (that is to say that the change in the sign of one variable change the sign of the other):

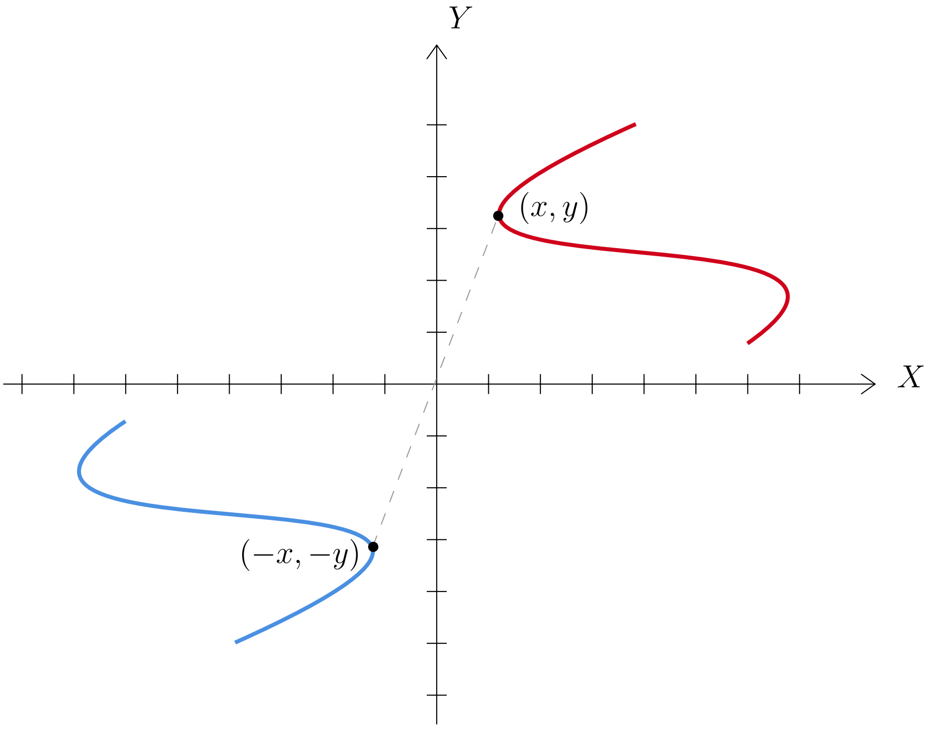


Figure 7.23: **Example of symmetry through the origin O of a function**

P4. Given a function $y=f(x)$, if we add a constant $c^{te}\geq 0$ to this function as:

$$y=f(x)+c^{te} \tag{7.1.32}$$

then the function $f(x)$ is shifted (or "translated") vertically upwards of a distance c^{te} as presented in the figure below:

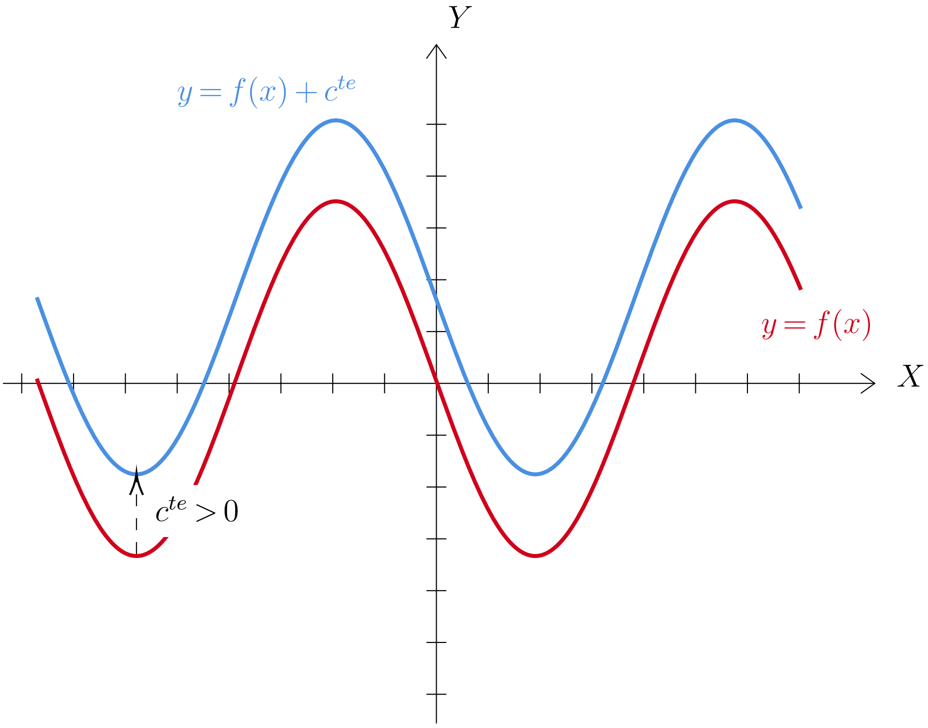


Figure 7.24: Example of a positive vertical translation of a function

And conversely if $c^{te}\geq 0$ but:

$$y=f(x)-c^{te} \tag{7.1.33}$$

then the function $f(x)$ is obviously translated vertically downwards:

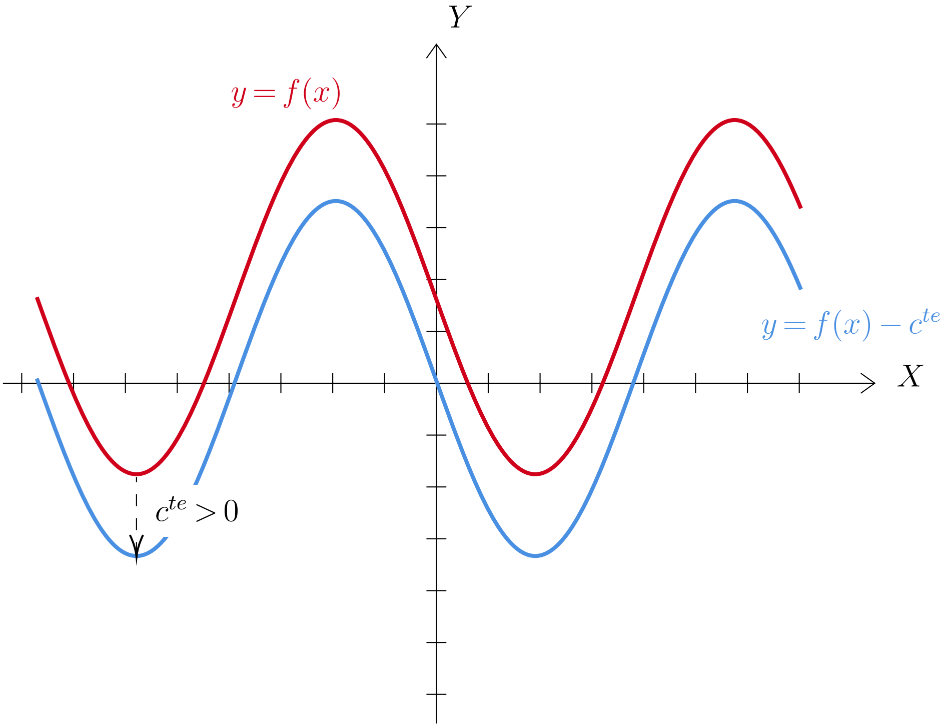


Figure 7.25: Example of a negative vertical translation of a function

We can also consider horizontal translations of functions. Specifically, if we have still $c^{te} > 0$, then $y = f(x)$ is translated horizontally to the right if we write:

$$y = f(x - c^{te}) \tag{7.1.34}$$

which graphically is represented by:

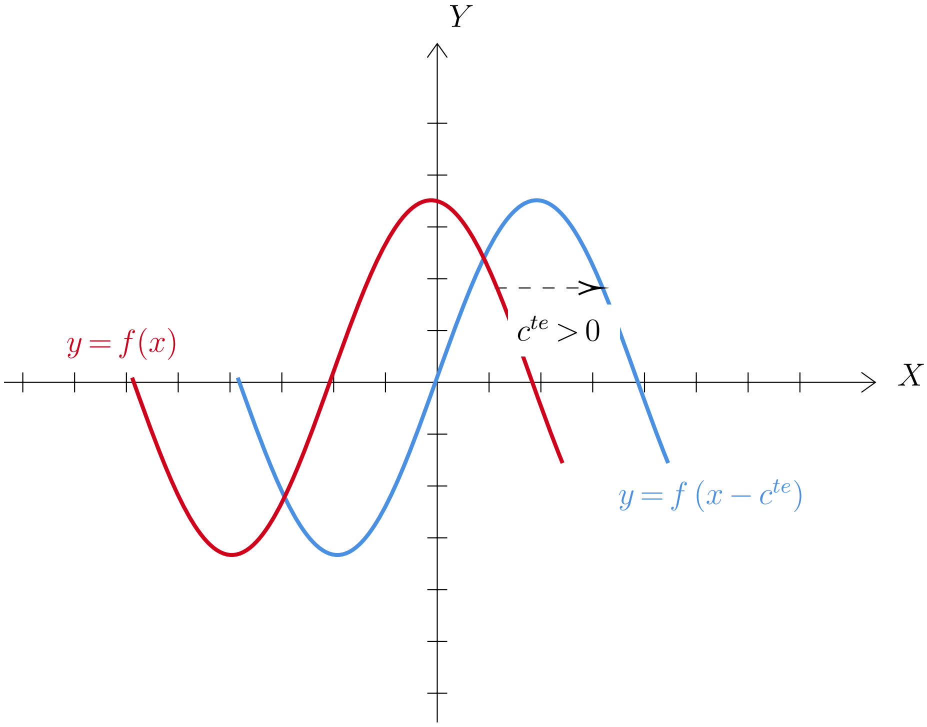


Figure 7.26: Example of negative horizontal translation of a function

and conversely, translated horizontally to the left, if we write:

$$y = f(x + c^{te}) \tag{7.1.35}$$

as shown in the graph below:

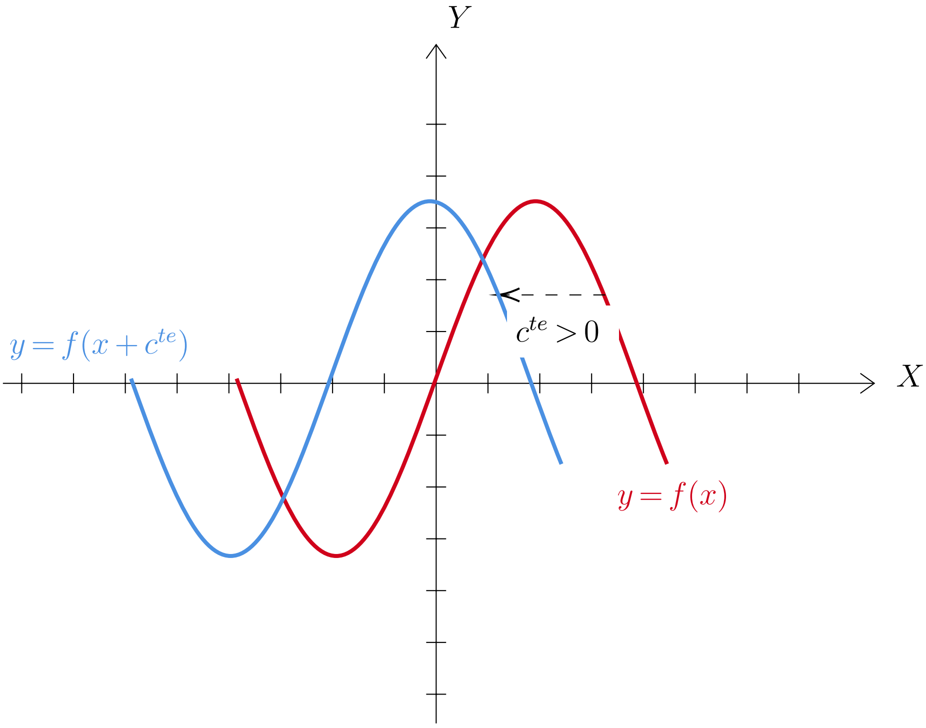


Figure 7.27: **Example of positive horizontal translation of a function**

To stretch or compress vertically a function, we simply multiply $y=f(x)$ by a constant $c^{te}>1$ and respectively $0\leq c^{te}<1$ as:

$$y=c^{te}f(x) \tag{7.1.36}$$

and don't forget that if a function is linear then we have the special property $f(\lambda x)=\lambda f(x)$.

This is graphically represented for the case $c^{te}>1$ by:

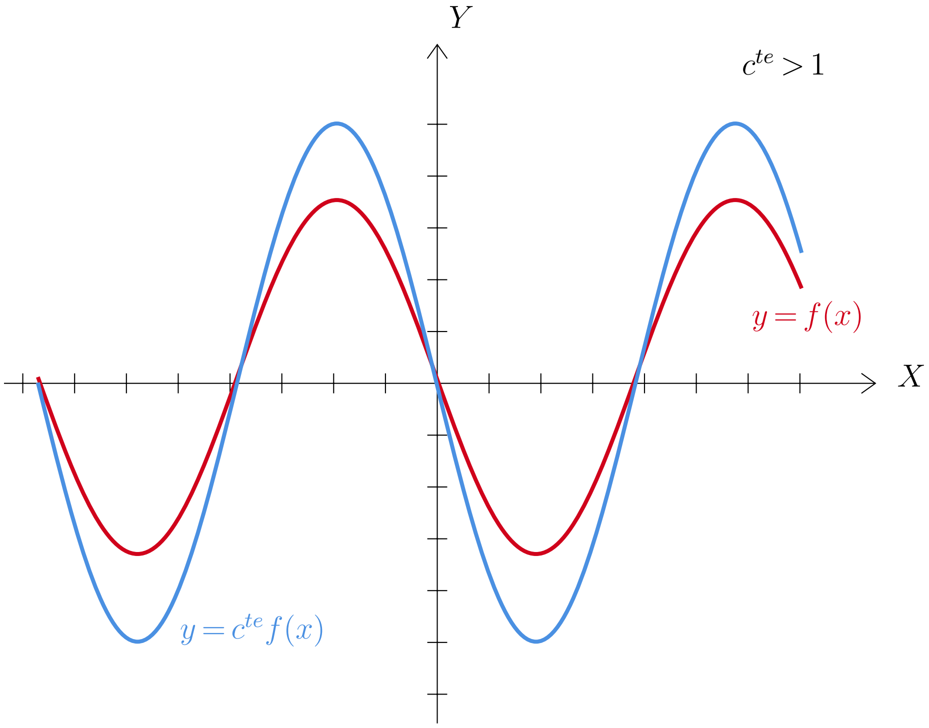


Figure 7.28: **Example of vertical stretch of a function**

and when $0 \leq c^{te} < 1$ by:

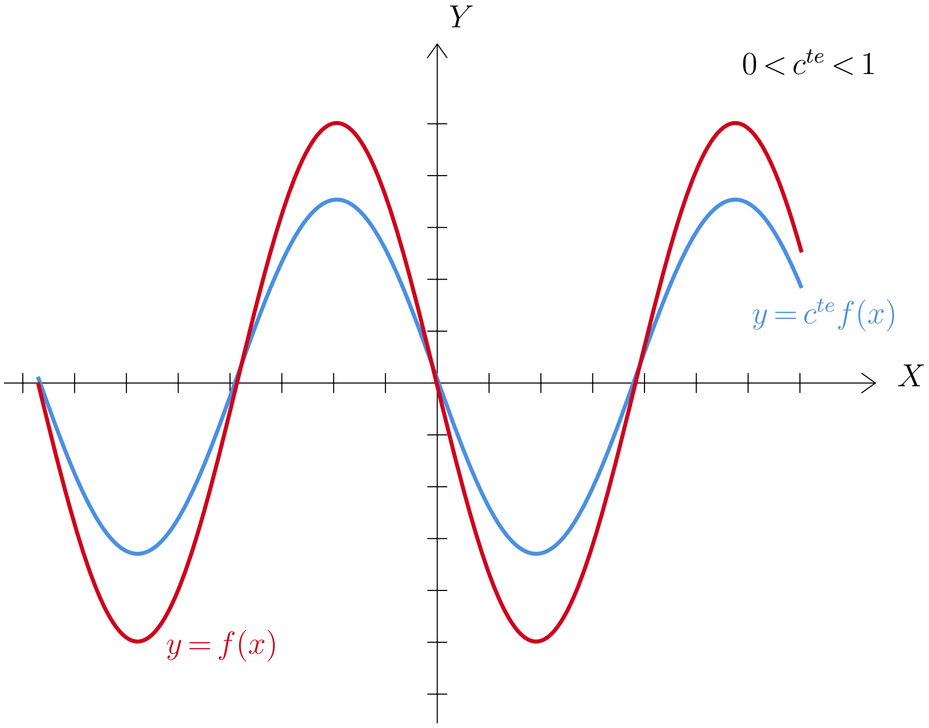


Figure 7.29: **Example of vertical compression of a function**

To stretch or compress a function horizontally, by the same way, we just need to multiply the variable x by a constant $c^{te} > 1$ and respectively $0 < c^{te} < 1$ as:

$$y = f(c^{te} x) \tag{7.1.37}$$

This is graphically represented for the case $0 < c^{te} < 1$ by:

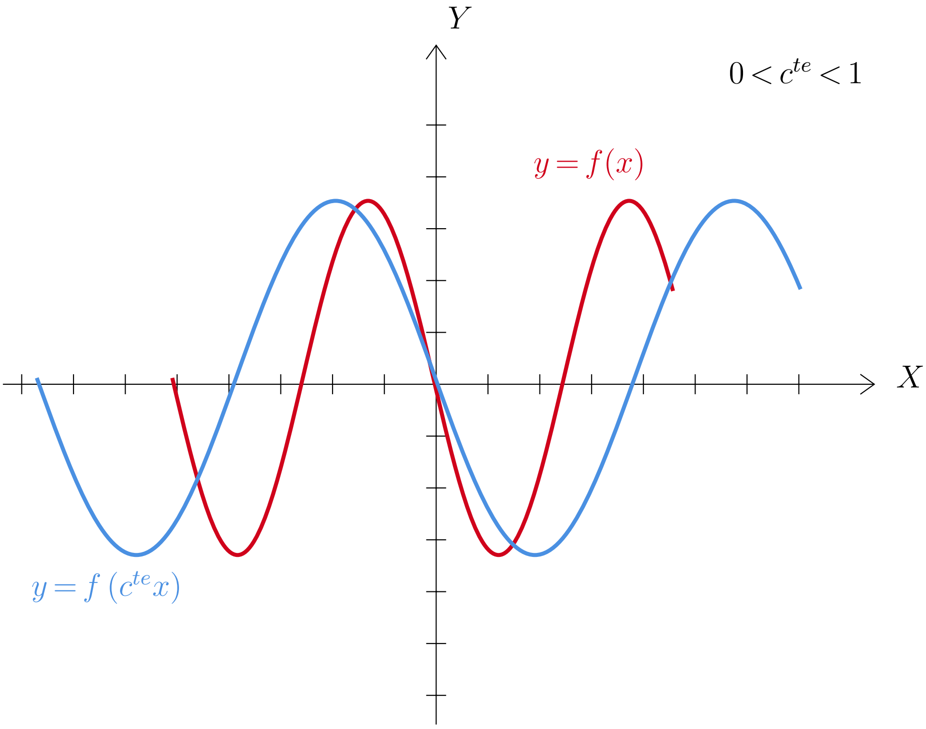


Figure 7.30: **Example of horizontal stretch of a function**

and when $c^{te} > 1$ by:

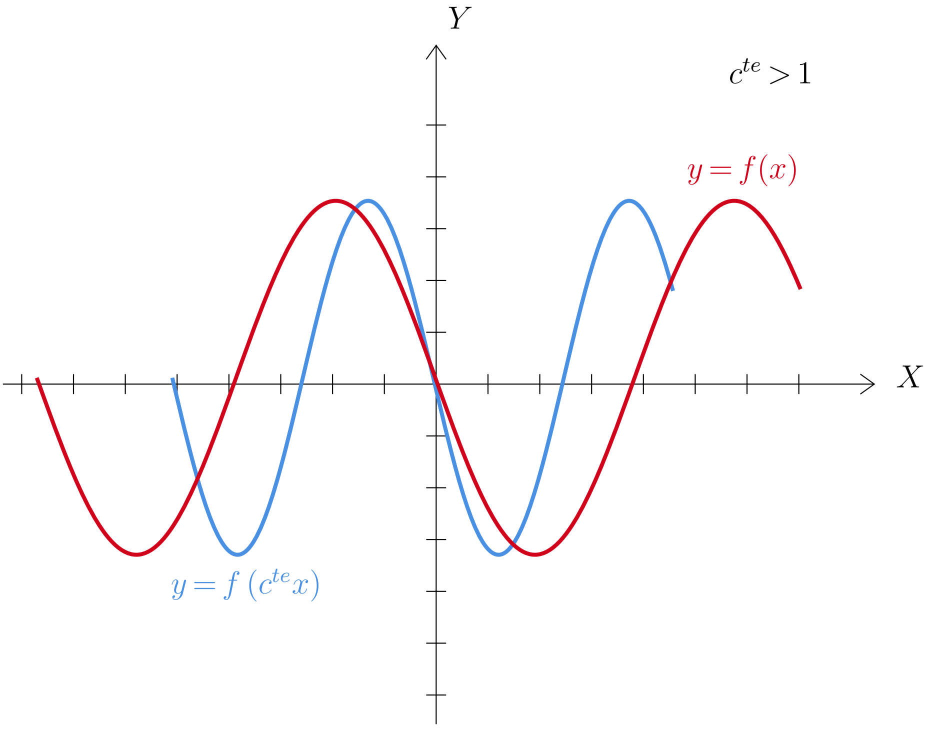


Figure 7.31: **Example of horizontal downscale of a function**

Remark

Translate, stretch, compress a function or apply it a symmetry is transforming it. The plot resulting from these transformations is named the "transformed" from the initial plot.

Definitions (#7.1.269): We say that a function f is (we simplify the definition using an univariate function):

D1. A "constant function" on an interval I if for each pair (x_1, x_2) of elements of I such that $x_1 \neq x_2$, we have $f(x_1) = f(x_2)$. What we denote in a condensed manner by:

$$\forall (x, y) \in I^2 : \quad f(x) = f(y) = c^{te} \tag{7.1.38}$$

D2. An "increasing function" or an "increasing function in the broadest sense" on the interval I if for each pair (x_1, x_2) of elements of I such that $x_1 \leq x_2$, we have $f(x_1) \leq f(x_2)$. What we denote in a condensed manner by:

$$\forall (x, y) \in I^2 \wedge (x \leq y) : \quad f(x) \leq f(y) \tag{7.1.39}$$

D3. A "decreasing function" or an "decreasing function in the broadest sense" on the interval I if for each pair (x_1, x_2) of elements of I such that $x_1 \leq x_2$, we have $f(x_1) \geq f(x_2)$. What we denote in a condensed manner by:

$$\forall (x, y) \in I^2 \wedge (x \leq y) : \quad f(x) \geq f(y) \tag{7.1.40}$$

Remark

A function is a "monotonic function" or "monotonic function in the broadest sense" on an interval I if it is increasing or decreasing in this interval.

D4. A "strictly increasing function" on the interval I if for each pair (x_1, x_2) of elements of I such that $x_1 \leq x_2$, we have $f(x_1) < f(x_2)$. What we denote in a condensed manner by:

$$\forall (x, y) \in I^2 \wedge (x < y) : \quad f(x) < f(y) \tag{7.1.41}$$

D5. A "strictly decreasing function" on the interval I if for each pair (x_1, x_2) of elements of I such that $x_1 \leq x_2$, we have $f(x_1) > f(x_2)$. What we denote in a condensed manner by:

$$\forall (x, y) \in I^2 \wedge (x < y) : \quad f(x) > f(y) \tag{7.1.42}$$

Remark

A function is a "strictly monotonic function" on an interval I if it is strictly increasing or decreasing in this interval.

Here is a visual summary of some of the monotonicity types defined above:

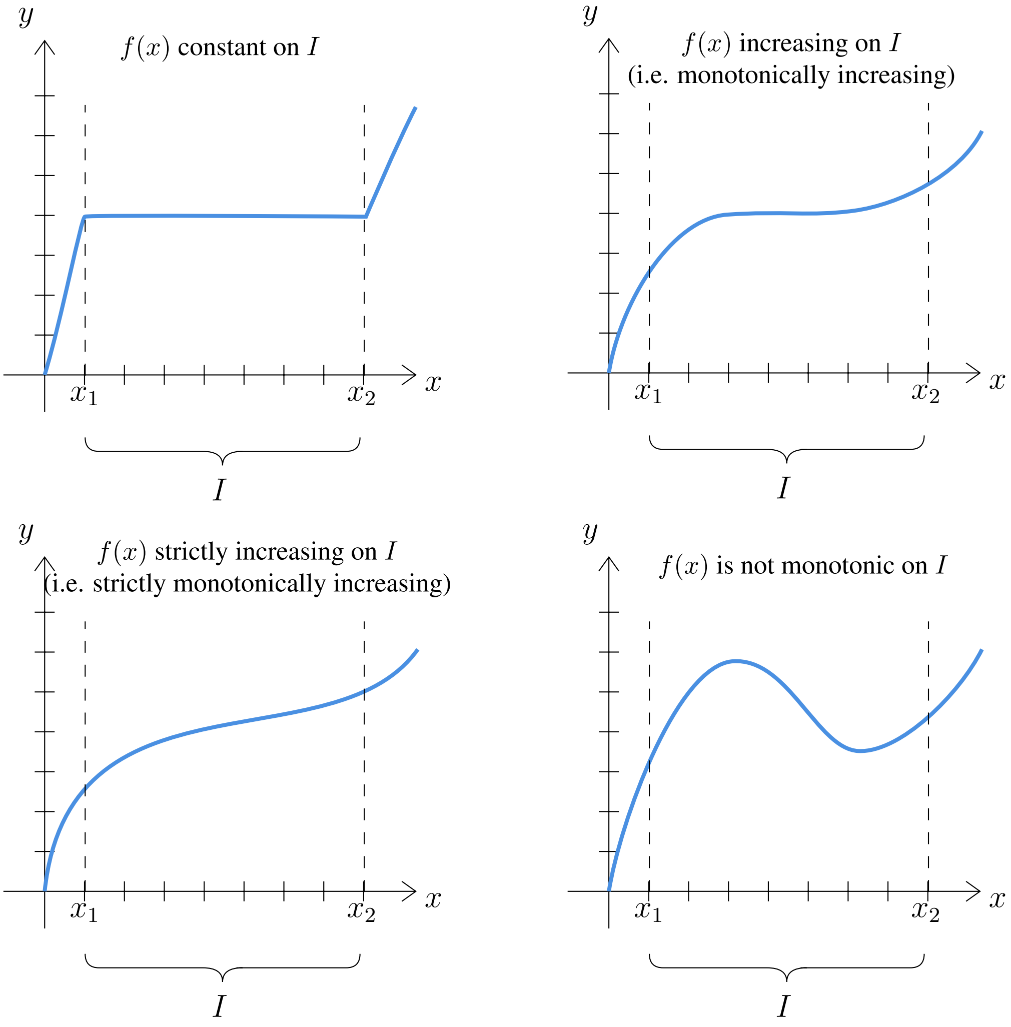


Figure 7.32: **Summary of some of the types of monocities of univariate functions**

There are plenty of other existing visuals/charts representations. You can find in the thumbnail gallery below 206 of them (excepted for two of them, they were all made with the `R` software and the code to create most of them is available in the corresponding companion book):

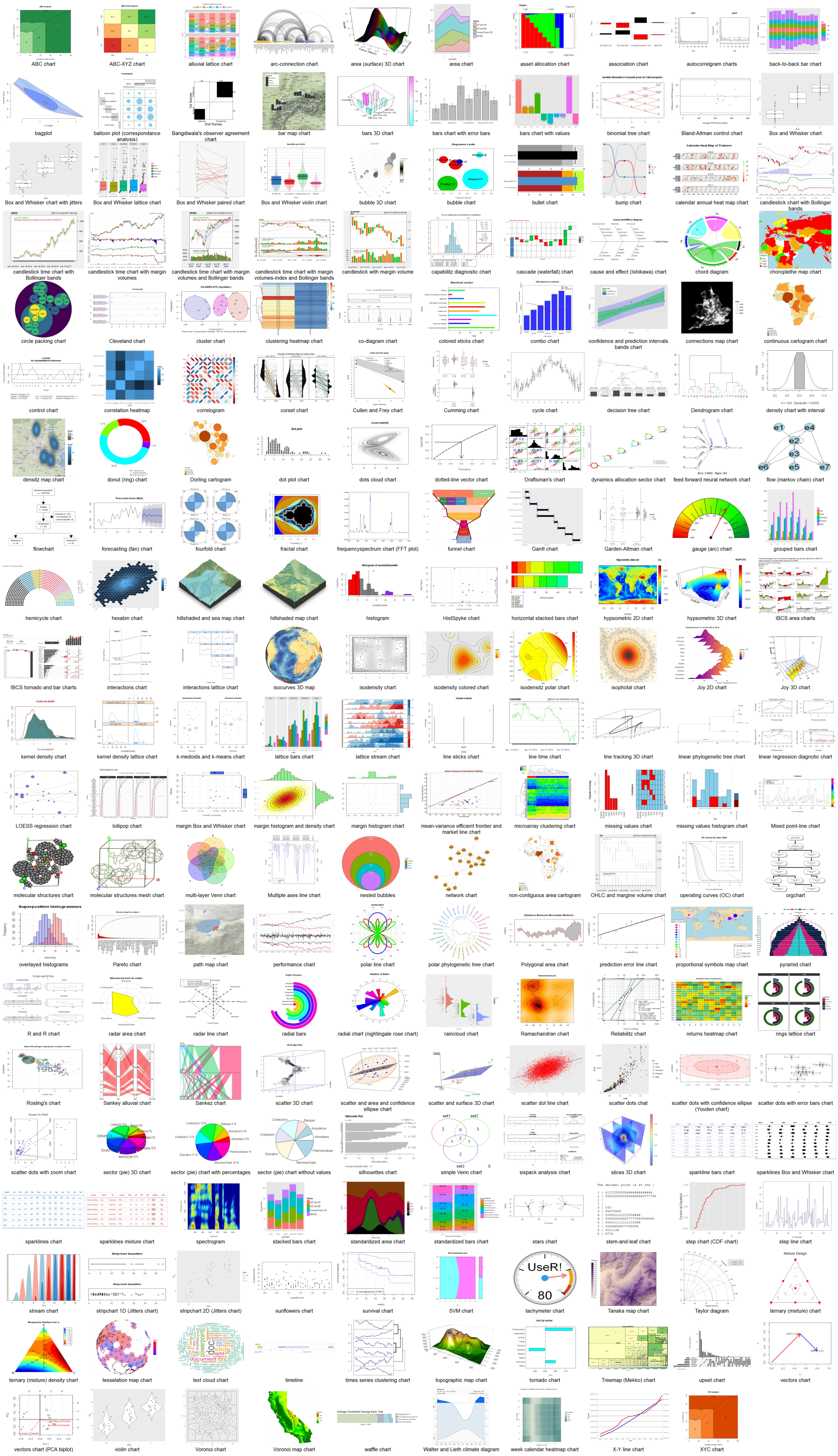

ABC chart
ABC-XYZ chart
alluvial lattice chart
arc-connection chart
area (surface) 3D chart
area chart
asset allocation chart
association chart
autocorrelgram charts
back-to-back bar chart
bagplot
balloon plot (correspondance analysis)
Bangdiwala's observer agreement chart
bar map chart
bars 3D chart
bars chart with error bars
bars chart with values
binomial tree chart
Bland-Altman control chart
Box and Whisker chart
Box and Whisker chart with jitters
Box and Whisker lattice chart
Box and Whisker paired chart
Box and Whisker violin chart
bubble 3D chart
bubble chart
bullet chart
bump chart
calendar annual heat map chart
candlestick chart with Bollinger bands
candlestick time chart with Bollinger bands
candlestick time chart with margin volumes
candlestick time chart with margin volumes and Bollinger bands
candlestick time chart with margin volumes-index and Bollinger bands
candlestick with margin volume
capabilitiy diagnostic chart
cascade (waterfall) chart
cause and effect (Ishikawa) chart
chord diagram
choropleth map chart
circle packing chart
Cleveland chart
cluster chart
clustering heatmap chart
co-diagram chart
colored sticks chart
combo chart
confidence and prediction intervals bands chart
connections map chart
continuous cartogram chart
control chart
correlation heatmap
correlogram
corset chart
Cullen and Frey chart
Cumming chart
cycle chart
decision tree chart
Dendrogram chart
density chart with interval
densitiy map chart
donut (ring) chart
Dorling cartogram
dot plot chart
dots cloud chart
dotted-line vector chart
Draftsman's chart
dynamics allocation sector chart
feed forward neural network chart
flow (markov chain) chart
flowchart
forecasting (fan) chart
fourfold chart
fractal chart
frequencyspectrum chart (FFT plot)
funnel chart
Gantt chart
Garden-Altman chart
gauge (arc) chart
grouped bars chart
hemicycle chart
hexabin chart
hillshaded and sea map chart
hillshaded map chart
histogram
HistSpyke chart
horizontal stacked bars chart
hypsometric 2D chart
hypsometric 3D chart
IBCS area charts
IBCS tornado and bar charts
interactions chart
interactions lattice chart
isocurves 3D map
isodensity chart
isodensity colored chart
isodensitiz polar chart
isophotal chart
Joy 2D chart
Joy 3D chart
kernel density chart
kernel density lattice chart
k-medoids and k-means chart
lattice bars chart
lattice stream chart
line sticks chart
line time chart
line tracking 3D chart
linear phylogenetic tree chart
linear regression diagnotic chart
LOESS regression chart
lollipop chart
margin Box and Whisker chart
margin histogram and density chart
margin histogram chart
mean-variance efficient frontier and market line chart
microarray clustering chart
missing values chart
missing values histogram chart
Mixed point-line chart
molecular structures chart
molecular structures mesh chart
multi-layer Venn chart
Multiple axes line chart
nested bubbles
network chart
non-contiguous area cartogram
OHLC and margine volume chart
operating curves (OC) chart
orgchart
overlayed histograms
Pareto chart
path map chart
performance chart
polar line chart
polar phylogenetic tree chart
Polygonal area chart
prediction error line chart
proportional symbols map chart
pyramid chart
R and R chart
radar area chart
radar line chart
radial bars
radial chart (nightingale rose chart)
raincloud chart
Ramachandran chart
Reliabilitz chart
returns heatmap chart
rings lattice chart
Rosling's chart
Sankey alluvial chart
Sankez chart
scatter 3D chart
scatter and area and confidence ellipse chart
scatter and surface 3D chart
scatter dot line chart
scatter dots chart
scatter dots with confidence ellipse (Youden chart)
scatter dots with error bars chart
scatter dots with zoom chart
sector (pie) 3D chart
sector (pie) chart with percentages
sector (pie) chart without values
silhouettes chart
simple Venn chart
sixpack analysis chart
slices 3D chart
sparkline bars chart
sparklines Box and Whisker chart
sparklines chart
sparklines mixture chart
spectrogram
stacked bars chart
standardized area chart
standardized bars chart
stars chart
stem-and-leaf chart
step chart (CDF chart)
step line chart
stream chart
stripchart 1D (Jitters chart)
stripchart 2D (Jitters chart)
sunflowers chart
survival chart
SVM chart
tachymeter chart
Tanaka map chart
Taylor diagram
ternary (mixture) chart
ternary (mixture) density chart
tesselation map chart
text cloud chart
timeline
times series clustering chart
topographic map chart
tornado chart
Treemap (Mekko) chart
upset chart
vectors chart
vectors chart (PCA biplot)
violin chart
Voronoi chart
Voronoi map chart
waffle chart
Walter and Lieth climate diagram
week calendar heatmap chart
X-Y line chart
XYC chart

Figure 7.33: List of 206 typical types of charts

However there is a few other existing charts used in practice and the reader interested in a more exhaustive list can take a look at the *Type of Charts/Plots/Visuals* branch of the Data Science Mind Map at page 6770.

7.1.1.3 Analytical Representations

The analytical method of representation is by far the most used and consists of representing any function in an "analytic expression" or "closed form" which is a symbolic and abstract mathematical notation of all known mathematical operations that must be applied in a certain order to numbers and letters expressing constants or variables that we seek to analyse.

Note that by "all known mathematical operations", we consider not only the mathematical operations seen in the chapter of Arithmetics (addition, subtraction, root extraction, etc.) but also all the operations that will be defined later in this textbook.

If the functional dependence $y=f(x)$ is such that f is an analytic expression, then we say that the "function y of x" is "given analytically ".

Here are some examples of simple analytical expressions:

$$y=x^4-2, y=\frac{x+1}{x-1}, y=\sqrt{1-x^2} \tag{7.1.43}$$

When we have determined the equation of the mediator (see page 2195), we have obtained an analytical expression of the visual straight line that characterize it as a function of the type:

$$\boxed{f(x)=y=ax+b} \tag{7.1.44}$$

which we recall, is the analytical expression of the equation of a straight line, also named "linear equation" or "affine function", on a plane where two points are known $P_1(x_1,y_1), P_2(x_2,y_2)$, the slope is given by the ratio of vertical growth on the horizontal growth as:

$$a=\frac{y_2-y_1}{x_2-x_1}\equiv\frac{\Delta y}{\Delta x} \tag{7.1.45}$$

A friendly and trivial application is to prove analytically that two non-vertical lines are parallel if and only if they have the same slope. Thus, given two lines with the equations:

$$y_1=ax_1+b_1 \qquad y_2=a_2+b_2 \tag{7.1.46}$$

The lines intersect at a point (x,y) if and only if values of y are equal for a certain x, that is to say:

$$a_1x+b_1=a_2x+b\Leftrightarrow(a_1-a_2)x=b_2-b_1 \tag{7.1.47}$$

The last equation can be solved with respect to x if and only if $a_2-a_2\neq 0$. We have therefore proved that the lines y_1,y_2 intersect if and only if $a_1\neq a_2$. Therefore, they do not intersect (are parallel) if and only if $a_1=a_2$.

In a quite simple way by applying the Pythagorean theorem, it is not difficult (see section Analytical Geometry page 2764) to determine that the equation of a circle with center $C(h,k)$

has for equation (it is of use in mathematics not explain y for the equation of the circle therefore the equation of the latter is much more visually aesthetic and speaking):

$$r=\sqrt{(x-h)^2+(y-k)^2} \tag{7.1.48}$$

In these examples the functions are expressed analytically by a single formula (equality between two analytical expressions) which defines at the same time the "natural domain of definition" of the functions.

Definition (#7.1.270): The "natural domain of definition" of a function given by an analytical expression is the set of x values for which the expression on the right-hand side has a definite value.

For example the function:

$$y=\frac{x+1}{x-1} \tag{7.1.49}$$

is defined for all values of x except the value $x=1$ where we have a singularity (division by zero).

Remark

There are an infinite number of functions and we can not expose them all here, however we will meet more than a thousand on this entire book and should amply suffice to get an idea of their study.

And we have the famous following "table of variations" that is also considered as an analytical tool and also used by some teachers to study the basics of the derivative f' of a function f (see section Differential and Integral Calculus page 1450). For example with the function x^3-3x^2+2 (already seen in the previous mentioned section):

x	$-\infty$		0		2		$+\infty$
f'		$+$	0	$-$	0	$+$	
f		↗		↘		↗	

Whose corresponding plot is:

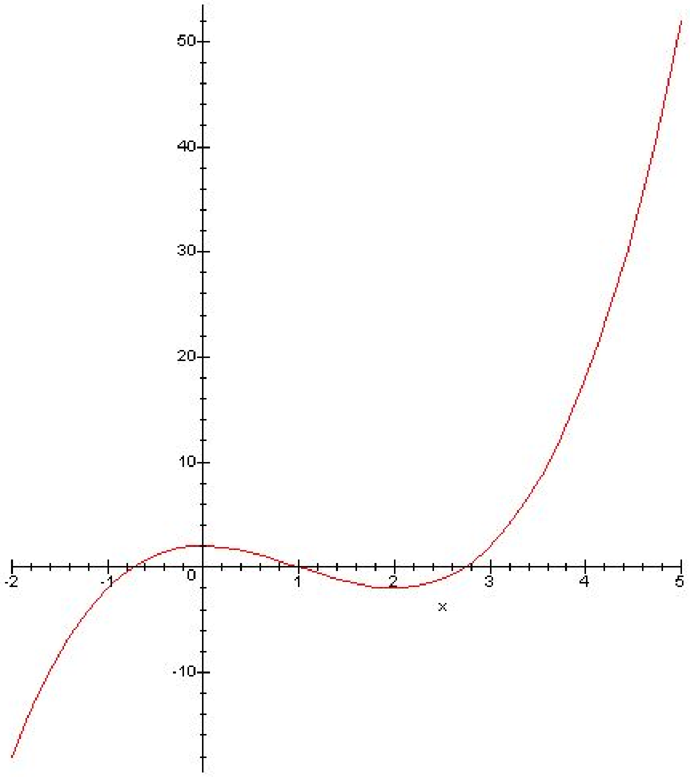


Figure 7.34: **Plot of function** $x^3 - 3x^2 + 2$

7.1.2 Functions

In mathematics, a function is a relation between a set of inputs and a set of permissible outputs with the property that each input is related to exactly one output.

Functions of various kinds are the central objects of investigation in most fields of modern mathematics. There are many ways to describe or represent a function. Some functions may be defined by a formula or algorithm that tells how to compute the output for a given input. Others are given by a picture, named the "graph" of the function. In science, functions are sometimes defined by a table that gives the outputs for selected inputs. A function could be described implicitly, for example as the inverse to another function or as a solution of a differential equation.

First remember the definitions already given earlier during our study of graphical representation of functions:

Definitions (#7.1.271): We say that a function f is (we simplify the definition using an univariate function):

D1. A function is a "constant function" on an interval I if for each pair (x_1,x_2) of elements of I such that $x_1\neq x_2$, we have $f(x_1)=f(x_2)$. What we denote in a condensed manner by:

$$\forall (x,y)\in I^2: \quad f(x)=f(y)=c^{te} \tag{7.1.50}$$

D2. A function is an "increasing function" or an "increasing function in the broadest sense" on the interval I if for each pair (x_1,x_2) of elements of I such that $x_1\leq x_2$, we have $f(x_1)\leq f(x_2)$. What we denote in a condensed manner by:

$$\forall (x,y)\in I^2\wedge(x\leq y): \quad f(x)\leq f(y) \tag{7.1.51}$$

D3. A function is an "decreasing function" or an "decreasing function in the broadest sense" on the interval I if for each pair (x_1,x_2) of elements of I such that $x_1\leq x_2$, we have $f(x_1)\geq f(x_2)$. What we denote in a condensed manner by:

$$\forall (x,y)\in I^2\wedge(x\leq y): \quad f(x)\geq f(y) \tag{7.1.52}$$

Remark

A function is a "monotonic function" or "monotonic function in the broadest sense" on an interval I if it is increasing or decreasing in this interval.

D4. A function is a "strictly increasing function" on the interval I if for each pair (x_1,x_2) of elements of I such that $x_1\leq x_2$, we have $f(x_1)<f(x_2)$. What we denote in a condensed manner by:

$$\forall (x,y)\in I^2\wedge(x<y): \quad f(x)<f(y) \tag{7.1.53}$$

D5. A function is an "strictly decreasing function" on the interval I if for each pair (x_1,x_2) of elements of I such that $x_1\leq x_2$, we have $f(x_1)>f(x_2)$. What we denote in a condensed manner by:

$$\forall (x,y)\in I^2\wedge(x<y): \quad f(x)>f(y) \tag{7.1.54}$$

Remark

A function is a "strictly monotonic function" on an interval I if it is strictly increasing or decreasing in this interval.

And let us add now complementary definitions:

Definitions (#7.1.272):

D1. We say that y is a function of x and we will write $y=f(x), y=\varphi(x)$, etc., if for every value of the variable x belonging to a certain domain of definition (set) D, corresponds a value of the variable y in another target domain of definition (set) E. What we denote in various ways (the third one being the most recommended):

$$D \xrightarrow{f} E \qquad f:D\rightarrow E \qquad f:D\mapsto E \tag{7.1.55}$$

The variable x is named "independent variable" or "input variable" or even "exogenous variable" and y the "dependent variable" or "endogenous variable".

The dependence between the variables x and y is named a "functional dependency". The letter f, which in the symbolic notation of functional dependence, indicates that we need to apply some operations to x to obtain the corresponding y value.

Sometimes we write:

$$y=y(x) \tag{7.1.56}$$

rather than:

$$y=f(x) \tag{7.1.57}$$

In the latter case the letter y expresses at the same time the value of the function and the symbol of operations applied to x.

Remark

As we saw it during our study in the section Set Theory, an application (or function) may be injective, surjective or bijective:

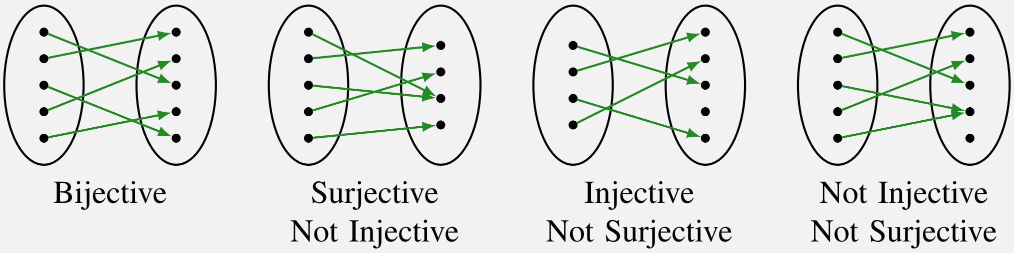


Figure 7.35: **Quick summary of applications/functions types**

It is therefore necessary that the reader for whom these concepts are unknown goes in priority read these definitions.

D2. The set of x values (inputs) for which the value of the function y is given by the function $f(x)$ is named the "range of existence" of the function or "domain of definition" of the function and denoted in this textbook by the letter D.

The set of outputs of $f(x)$ is named the "image" or sometimes the "codomain" and denoted in this textbook by the letter E. When study of the point of view of the knowledge of the output values only, the set of x is named the "pre-image".

D3. A function $f(x)$ is named a "periodic function" if there is a constant c^{te} such that the function's value does not change when we add (or subtract) that constant to the independent variable such as:

$$y=f(x\pm c^{te})=f(x) \tag{7.1.58}$$

which corresponds to a translation along the x-axis. The smallest constant satisfying this condition is named the "period" of the function. It is frequently denoted by the letter T in physics.

The most common periodic functions known by students and engineers are the trigonometric functions (see section of the corresponding name page 2503):

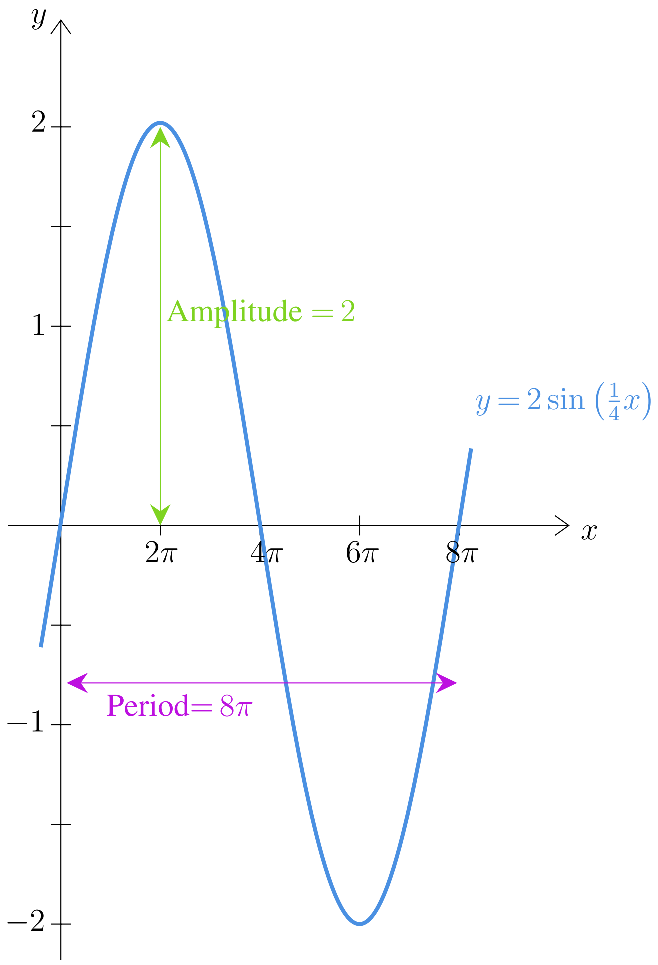


Figure 7.36: **Example of a very simple periodic function with period and amplitude**

Or let us consider a little bit more complicated example:

Remark

The sum of two periodic functions with different periods is not necessarily periodic and there is no general formula to get the period of a function that is the sum of n other functions!

D4. In differential calculus (see section Differential and Integral Calculus page 1450), the expression:

$$\frac{f(x+h)-f(x)}{h} \tag{7.1.59}$$

with $h\neq 0$ is of particular interest. We name it a "growth quotient" (we discuss this in much more detail in our study of differential and integral calculus).

D5. We use certain properties of functions for easy graphical representation and analysis or mathematical simplifications. In particular, a function $f(x)$ is named "even function" if:

$$f(x)=f(-x) \tag{7.1.60}$$

for all x in its definition domain.

That is to say as we already seen previously, it is symmetric relatively with the y-axis:

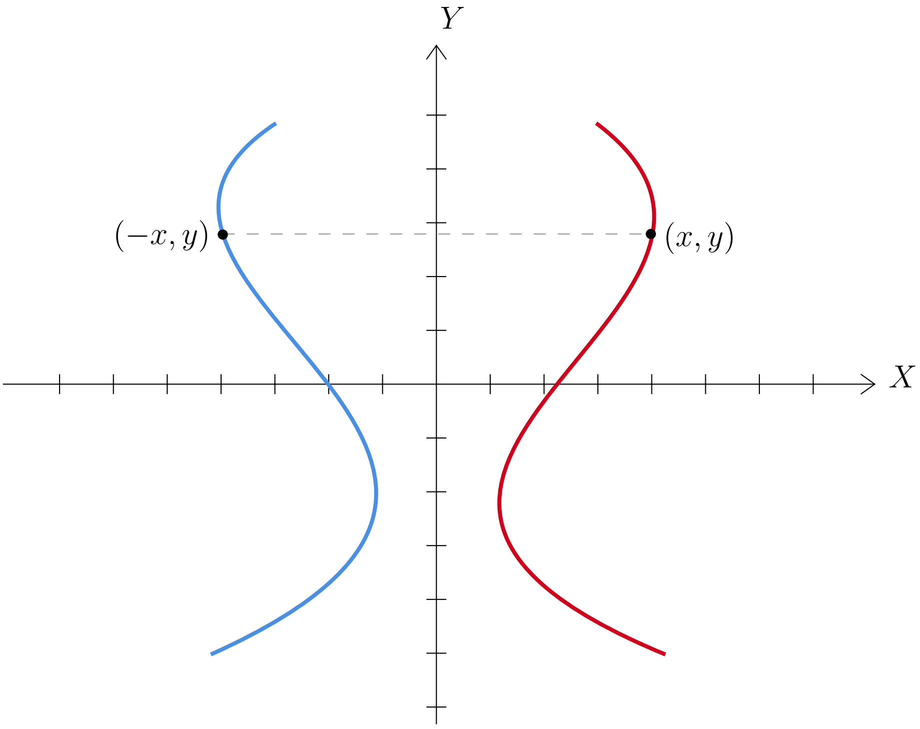


Figure 7.37: **Example of an even function**

A function $f(x)$ is named "odd function" if:

$$f(-x)=-f(-x) \tag{7.1.61}$$

for all x in its definition domain.

That is to say as we already seen previously, it is symmetric relatively with the origin:

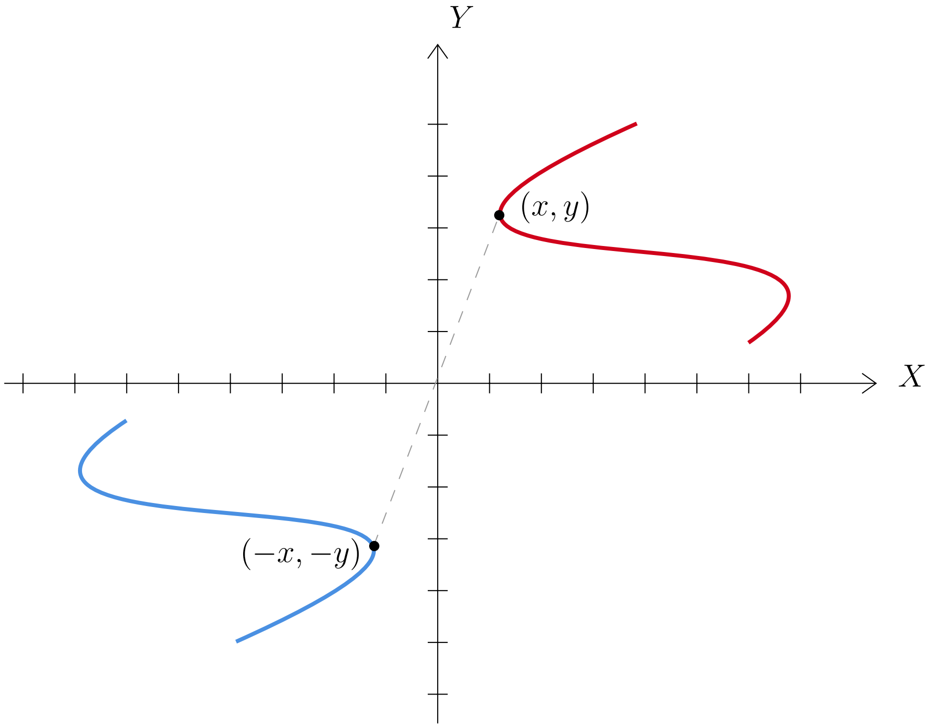


Figure 7.38: **Example of an odd function**

So, to summarize, an even function is a function that is independent of the sign of the variable and an odd function change of sign when we change the sign of the variable (the spiral of Cornus in the section Civil Engineering is a good practical example of odd function). This concept will be very useful to us to simplify some very useful developments in physics (such as Fourier transforms of odd or even functions for example, or the calculation of certain integrals!).

Theorem 7.2. Let us now show that any function $f(x)$ is the sum of an even function $g(x)$ and an odd function $h(x)$.

Remark

Remember that this type of theorem linking a general concept to a particular case and its opposite is often found in mathematics. We will see such examples in tensor calculus with the symmetric and antisymmetric tensor (see section Tensor Calculus page 2085 and page 2088) or in quantum physics with the Hermitian or non-Hermitian operators (see section Wave Quantum Physics page 4695 and page 4716).

Proof 7.2.1. Let us write:

$$f(x)=g(x)+h(x) \tag{7.1.62}$$

Then:

$$f(-x)=g(x)-h(x) \tag{7.1.63}$$

If we sum then we get:

$$g(x)=\frac{f(x)+f(-x)}{2} \tag{7.1.64}$$

and by subtracting:

$$h(x)=\frac{f(x)-f(-x)}{2} \tag{7.1.65}$$

So there is really and odd and even decomposition of any function!!!

■ Q.E.D.

Finally, it is important to note that:

- The product of two even functions is an even function
- The product of two odd functions is an even function
- The product of an even and odd function is an odd function

Let us see a short proof of the last property because we will need it in the chapter on Geometry.

Proof 7.2.2. Let $g(x)$ be an even function and $h(x)$ an odd function such as:

$$g(-x)=g(x) \tag{7.1.66}$$
$$h(-x)=-h(x) \tag{7.1.67}$$

Therefore:

$$q(-x)=g(-x)h(-x)=g(x)(-h(x))=-g(x)h(x)=-q(x) \tag{7.1.68}$$

■ Q.E.D.

D6. In general, if $f(x)$ and $g(x)$ are arbitrary functions, we use the terminology and notations given in the following table:

Terminology	**Value of the function**
Sum $f+g$	$(f+g)(x)=f(x)+g(x)$
Difference $f-g$	$(f-g)(x)=f(x)-g(x)$
Product $f\cdot g$	$(f\cdot g)(x)=f(x)g(x)$
Quotient $\dfrac{f}{g}$	$\left(\dfrac{f}{g}\right)(x)=\dfrac{f(x)}{g(x)}$

Table 7.3: **Terminology about functions**

The definition domains of $f+g$, $f-g$, $f\cdot g$ are the intersection I of the definition domain of $f(x)$ and $g(x)$, that is to say, the numbers which are common to both domains of definition. The domain of definition of f/g is meanwhile the subset I of all x such that $g(x)\neq 0$.

D7. Let y be a function of f of u such that $y=f(u)$ and u a function g of x such that $u=g(x)$, then y depends on x and we have what we name a "composite function" that we denote:

$$\boxed{y=f(g(x))=(f\circ g)(x)} \tag{7.1.69}$$

The last equality should be read "f round g" and not confuse with the "round" symbol with the notation of the dot product that we have seen during our study of the section Vector Calculus page 1821.

The domain of definition of the composite function is either identical to the entire domain of definition of the function $u=g(x)$ or the part of the domain in which the values of u are such that the corresponding values $f(u)$ belong to the domain of definition of this function.

Obviously the principle of composite function can be applied not only once, but an arbitrary number of times such that $y=f(g(h(t)))$ and so on...

In computing science a function may compose with itself a given number of times n, such that $f(f(f(f(f\ldots)))))=f^n$ that must not be confuse with the power notation $f^n(x)$.

If u does not depend on another variable (or it is not itself a composite function), then we say that $f(u)$ is an "elementary function".

Obviously there are an infinite number of elementary functions but most can be classified into families whose expression is similar to one of the following (here they are only univariate functions):

- "Linear functions":

  $$f(x)=ax \tag{7.1.70}$$

  They are simply functions representing straight lines of slope a passing through the origin of the axis.

- "Affine functions":

  $$f(x)=ax+b \tag{7.1.71}$$

  They are simply functions representing straight lines of slope a passing through the origin of the axis or not (linear function with a translation).

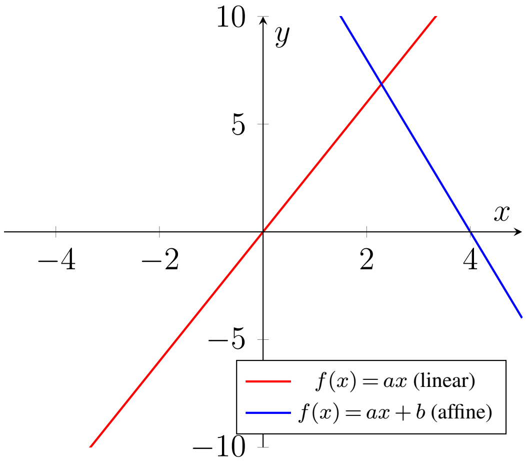


Figure 7.39: **Univariate affine and linear function**

- "Power functions":

  $$f(x)=x^m \tag{7.1.72}$$

where $m \in \mathbb{R}$. Functions involving roots are often named "radical functions".

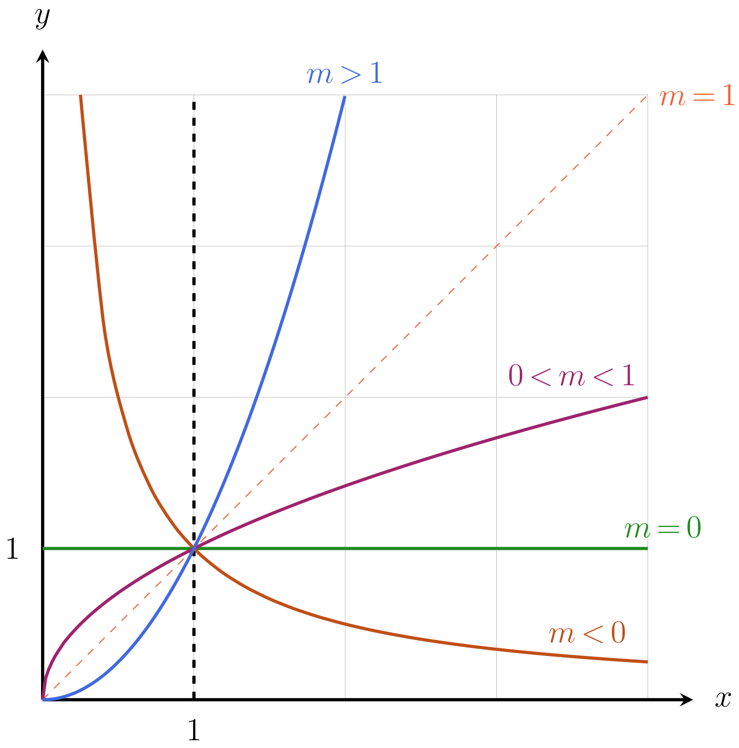


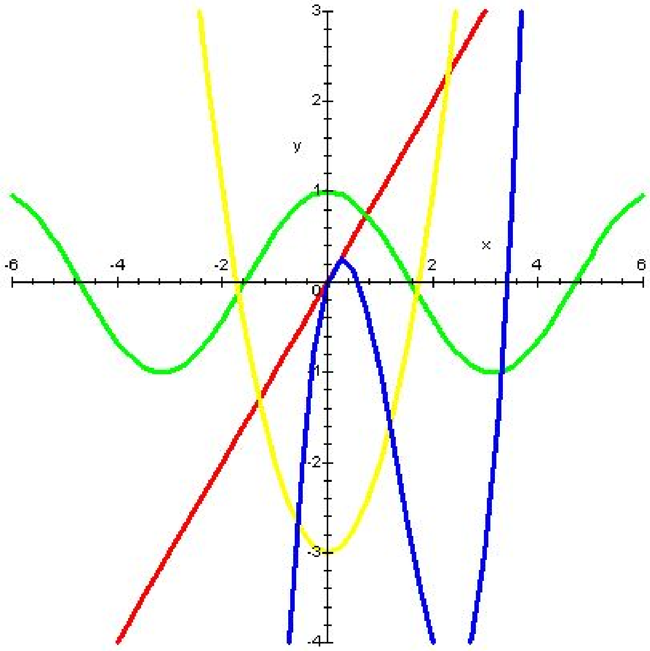
Figure 7.40: **Different plots of simple power functions**

- "Absolute value functions" (see section Arithmetic Operators page 324 for the definition and the study of the "absolute value"):

$$f(x) = |x| \tag{7.1.73}$$

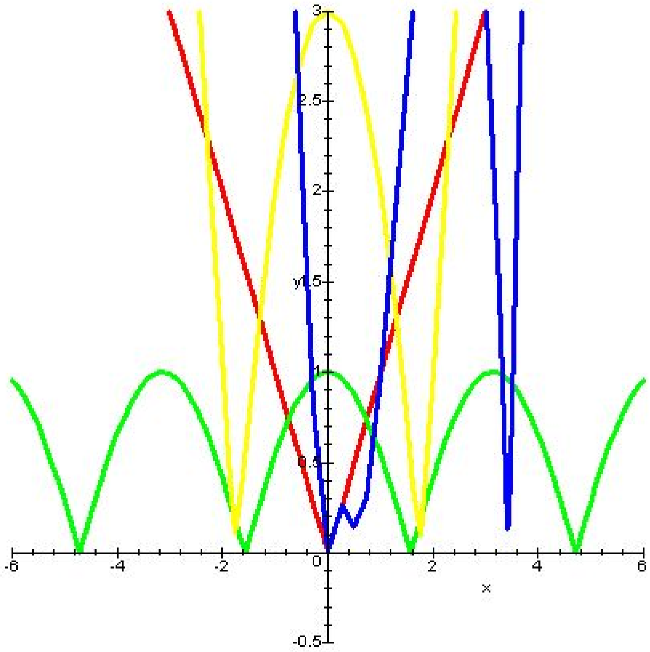
For example the plots with Maple™ 4.00b that we get with the command:

```
>plot([(x),(cos(x)),(x^2-3),(x^3-4*x^2+2*x)],x=-6..6,y=-4..3,
thickness=3);
```



and taking the absolute value:

```
>plot([abs(x),abs(cos(x)),abs(x^2-3),abs(x^3-4*x^2+2*x)]
,x=-6..6,y=-0.5..3,thickness=3);
```



- "Exponential functions":

$$f(x) = a^x \tag{7.1.74}$$

where the famous e^x (see section Functional Analysis page 2289) is only a special case and also $a \in \mathbb{R}$.

When $a \geq 0$ we have typically:

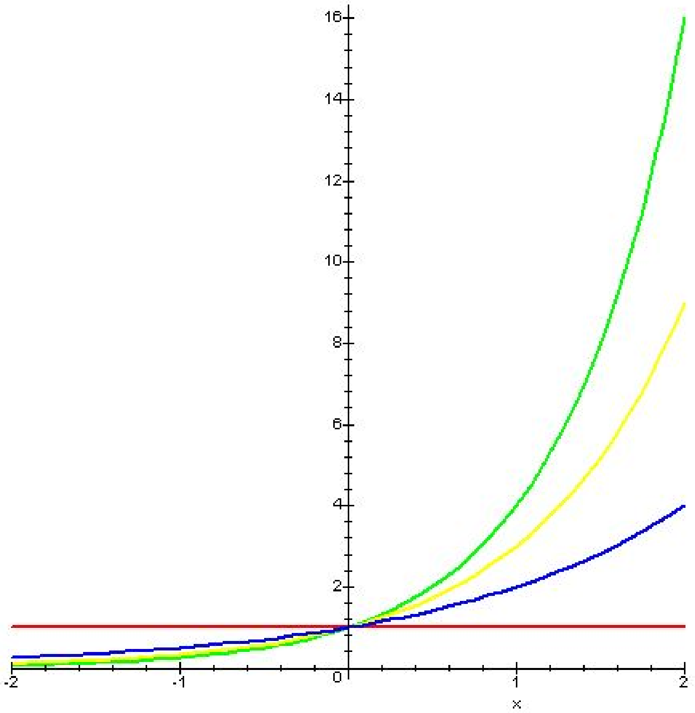


Figure 7.41: **Different plots of simple exponential functions** $(1^x, 2^x, 3^x, 4^x)$ **in Maple™ 4.00b**

If $a < 0, x \in \mathbb{R}$ the function is not defined. Indeed for $(-1)^{0.5} = \{\mathrm{i}, -\mathrm{i}\}$ therefore it is an application from $f: \mathbb{R} \mapsto \mathbb{C}^2$ and as far as we know there is no nice way to represent it visually and anyway this is not a function in the traditional way.

- "Logarithmic functions":

$$f(x) = \log_a(x) \tag{7.1.75}$$

with $a \in \mathbb{R}^+$ and that by construction of the logarithm (see further below) are of the type $f: \mathbb{R}^+ \mapsto \mathbb{R}$.

We have typically:

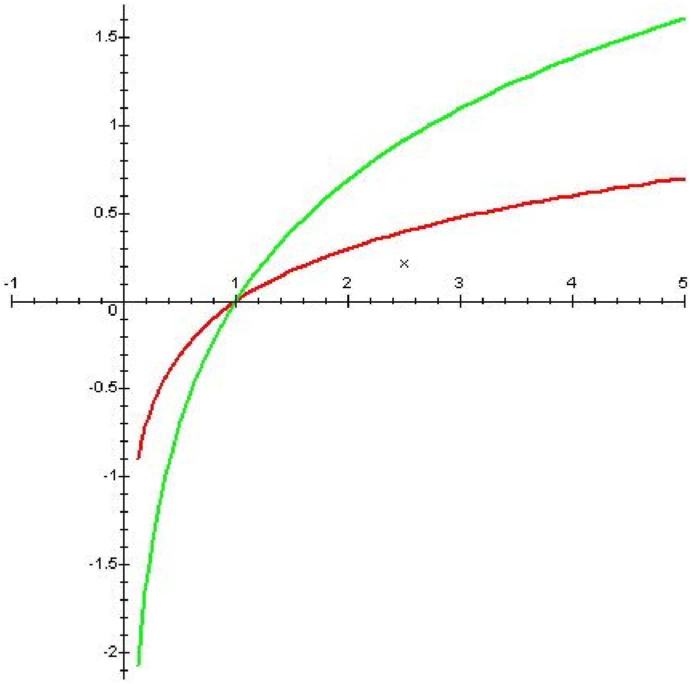


Figure 7.42: **Different plots of logarithm** $\ln(x)=\ln_e(x)$ **in green and** $\log_{10}(x)$ **in red with Maple™ 4.00b**

- "Periodic/Trigonometric functions":

$$y=\sin(x), y=\cos(x), y=\tan(x), y=\cot(x), y=\sec(x), y=\text{cosec}(x), \ldots \tag{7.1.76}$$

We already defined previously what is a periodic function. For the trigonometric functions the reader can see below a plot of the main one but for more details it is strongly recommended to read the section Trigonometry page 2503:

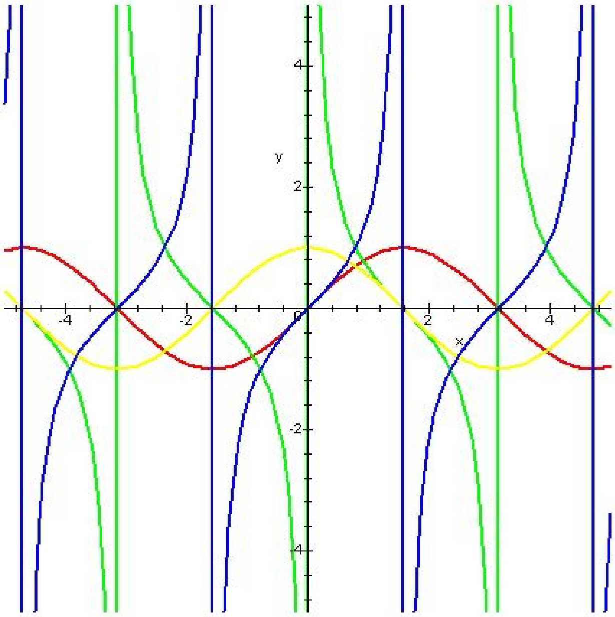


Figure 7.43: **Different plots of trigonometric functions in Maple™ 4.00b**

- "Polynomial functions":

$$f(x)=P(x)=a_n x^n+a_{n-1}x^{n-1}+\cdots+a_1x^1+a_0 \tag{7.1.77}$$

where as we already know $a_0, a_1, \ldots, a_n$ are constant numbers named "coefficients" and n is a positive integer that we name "degree of the polynomial" (voir section Calculus page 1368). Obviously this function is defined for all values of x, that is to say, it is define on an infinite interval.

If follows that power functions of the type x^m and linear functions of the type $f(x) = x$ are a subclass of polynomial for $m \in \mathbb{N}$.

We have already studied more deeply polynomials in the section Calculus with their main properties but let us give again the plot of some of them as recall:

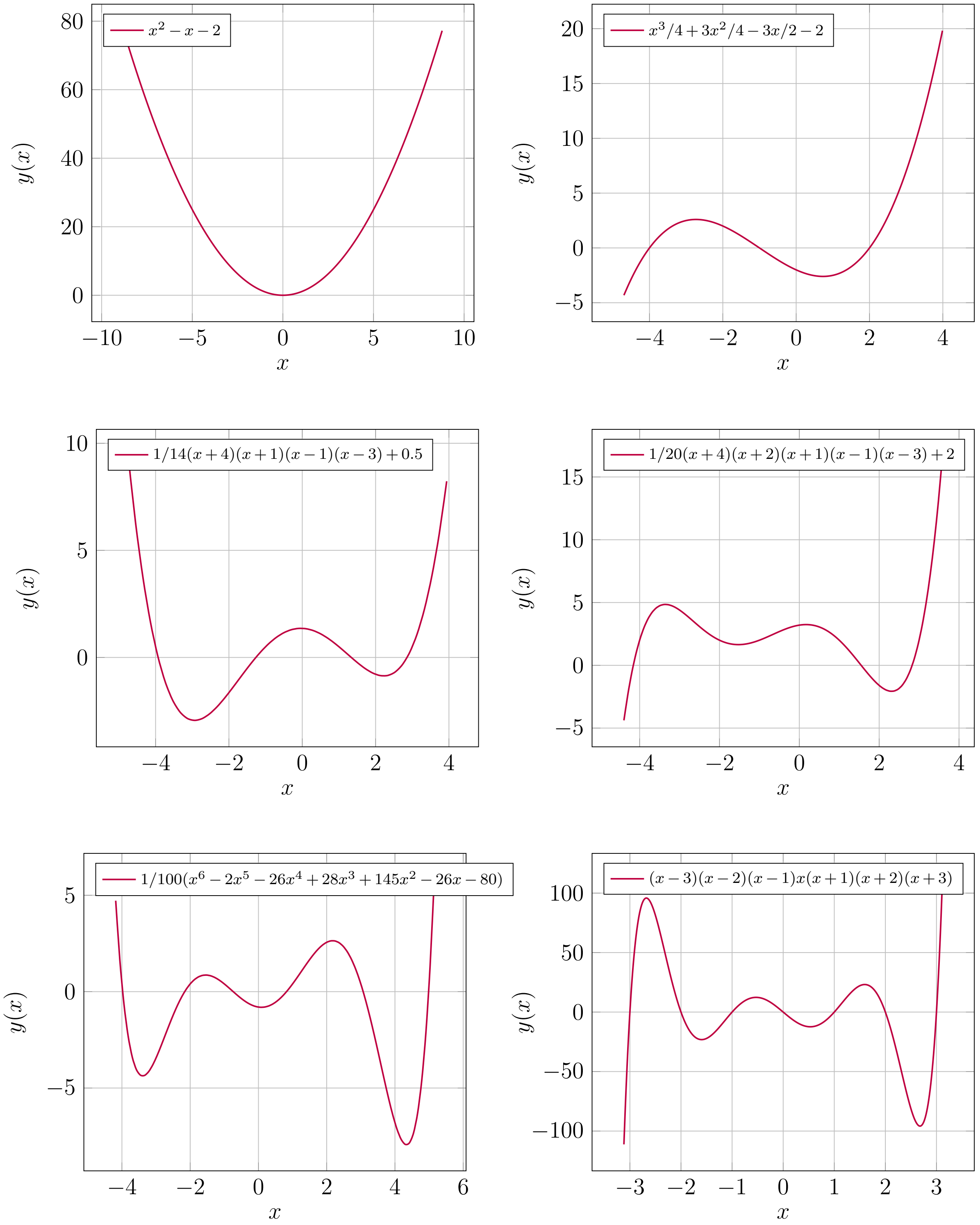


Figure 7.46: **Some polynomials plotted with TikZ**

We will see and study in this textbook some famous polynomials as the: Legendre polynomials (see section Quantum Chemistry page 6140), Bernoulli polynomials (see section Sequences and Series page 1672), Bernstein polynomials (see section Numerical Methods page 6570), Hermite polynomials (see section Functional Analysis page 2385), ...

- "Rational fractions" are polynomials divisions (see section Calculus page 1374):

$$f(x)=\frac{P(x)}{Q(x)}=\frac{a_nx^n+a_{n-1}x^{n-1}+\cdots+a_1x^1+a_0}{b_nx^n+b_{n-1}x^{n-1}+\cdots+b_1x^1+b_0} \tag{7.1.78}$$

Remark

Obviously two rational fractions are equal, if one is obtained from the other by multiplying the numerator and denominator by the same polynomial.

The rational function:

$$f(x)=\frac{x^3-2x}{2(x^2-5)} \tag{7.1.79}$$

is not defined at $x^2=5 \Leftrightarrow x=\pm\sqrt{5}$. It is asymptotic (see further below) to $\frac{x}{2}$ as x approaches infinity:

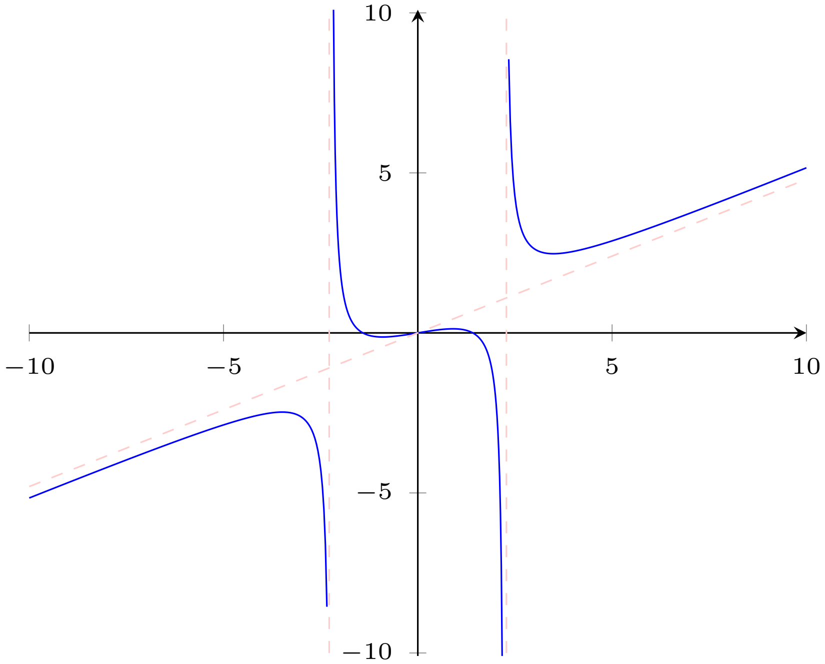


Figure 7.47: **Example of a rational function $f(x)=\frac{x^3-2x}{2(x^2-5)}$**

The rational function:

$$f(x)=\frac{x^2+2}{x^2+1} \tag{7.1.80}$$

is defined for all real numbers, but not for all complex numbers, since if x were a square root of -1 (i.e. the imaginary unit or its negative), then formal evaluation would lead to division by zero!

Remarks

A constant function is a rational function since constants are polynomials! Every polynomial function $f(x)=P(x)$ is a rational function with $Q(x)=1$. The power functions $f(x)=x^m$ are also rational functions when $m\in\mathbb{N}$.

- "Algebraic functions" are defined by the fact that the function $f(x)$ is the result of addition, subtraction, multiplication, division, of variables put to an integer or non-integer power. Therefore most of the functions defined previously can be included in this definition: linear functions, affine functions, power functions, polynomial functions, rational functions.
- "Piecewise functions" is a function defined by different formulas on different parts of its domain. The absolute value is a famous example of a piecewise-defined function because the formula changes with the sign of x:

$$f(x)=|x|:\begin{cases}-x & x<0\\ +x & x\geq 0\end{cases} \qquad (7.1.81)$$

  Another well-known example is the function that calculates pension contributions based on age. All countries (or companies) that have a pension concept have contribution amounts based on a piecewise continuous function.

- "Step functions" $f:[a,b]\in\mathbb{R}$ is defined if and only if there exists a subdivision $(a_i)_{0\leq i\leq n}$ of $[a,b]$ such that $a_0=a$ and $a_n=b$ and $(\lambda_0,\ldots,\lambda_n)\in\mathbb{R}^n$ such as:

$$\forall\in\{0,\ldots,n-1\}\,,\forall x\in]a_i,a_{i+1}[\Leftrightarrow f(x)=\lambda_i \qquad (7.1.82)$$

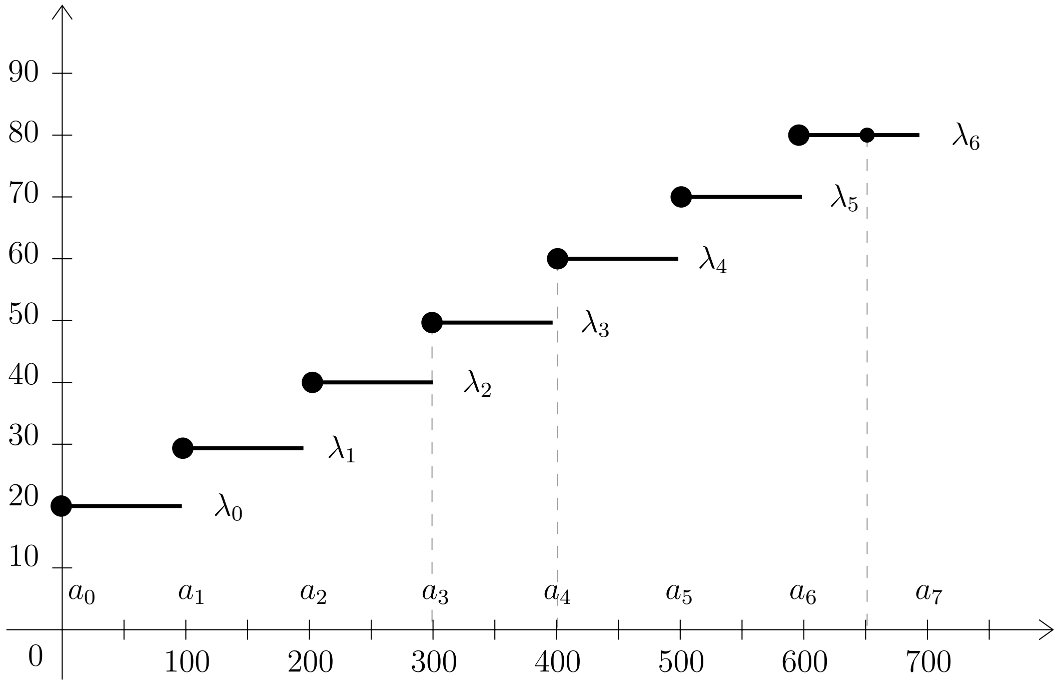


Figure 7.48: **Example of a step function**

Such functions can be found in signal processing and also in statistics for survival analysis.

However, there are a very large number of other elementary functions that will meet in the individual sections of this textbook. Examples include the "Bessel functions" (see section Sequences and Series page 1768), the "Lipschitz functions" (see section Topology page 2457), the "Dirac functions" (see section Differential and Integral Calculus page 1569), the "distribution functions" (see section Statistics page 633), the "Euler gamma function" (see section Differential and Integral Calculus page 1574), etc. The reader will notice that the Dirac function also belongs to the family of distribution functions.

Here is a quite good summary of most elementary functions (non exhaustive but still good!):

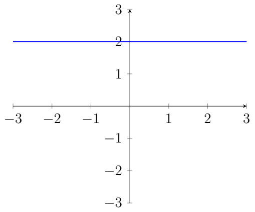
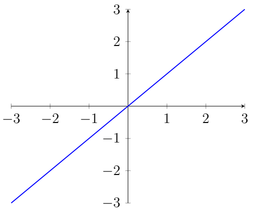
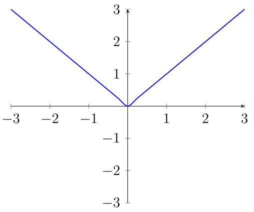
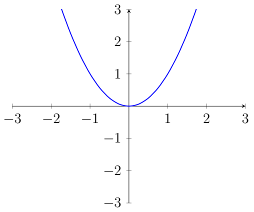
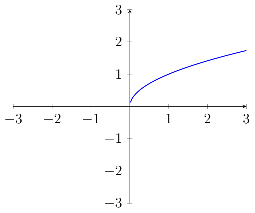
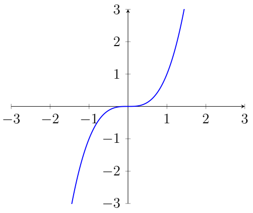
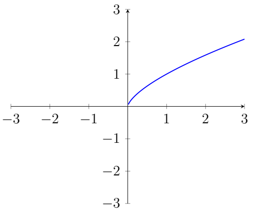
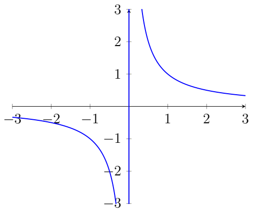
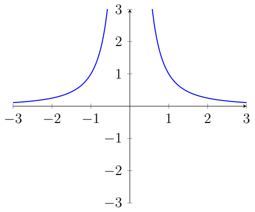
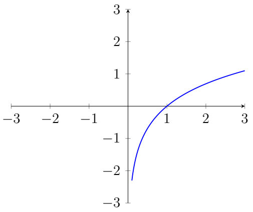
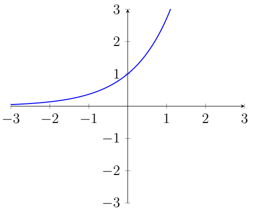
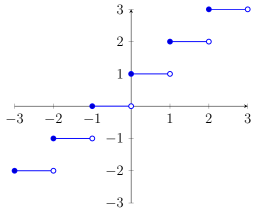
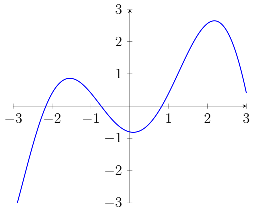
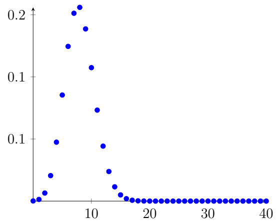
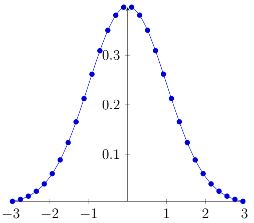
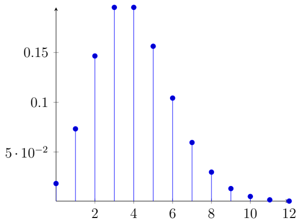
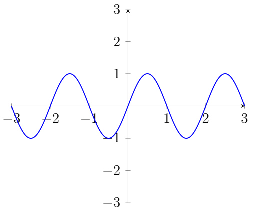
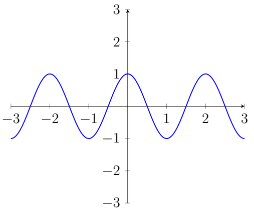
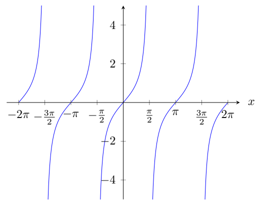
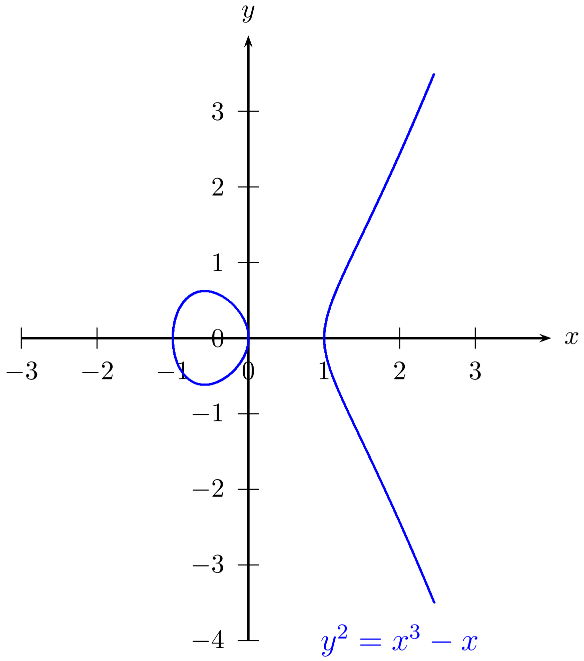
Constant  $f(x)=c^{te}$	Linear  $f(x)=x$	Absolute Value  $f(x)=\|x\|$	Quadratic  $f(x)=x^2$
Square root  $f(x)=\sqrt{x}$	Cubic  $f(x)=x^3$	Cube root  $f(x)=\sqrt[3]{x}$	Inverse  $f(x)=\frac{1}{x}$
Rational  $f(x)=\frac{1}{x^2}$	Logarithmic  $f(x)=\ln(x)$	Exponential  $f(x)=e^x$	Greatest Integer (step function)  $f(x)=[x]$
Polynomial  $f(x)=1/100(x^6-2x^5-26x^4+\ldots$	Binomial distribution  $f(x)=\mathcal{B}(N,k)$	Gaussian distribution  $f(x)=e^x$	Poisson distribution  $f(x)=\mathcal{P}(\mu,k)$
Trigonometric Functions $\rightarrow$	Sine  $f(x)=\sin(x)$	Cosine  $f(x)=\cos(x)$	Tangent  $f(x)=\tan(x)$

Table 7.4: **A summary of some elementary college level functions**

If once you see the plot like that one of an elliptic curve (mathematical tool used in cryptography):



You have to avoid to name that a "function". Indeed, by definition, a "function" has to pass the "vertical line test": The vertical line test is a method that is used to determine whether a given relation is a function or not. The approach is rather simple. Draw a vertical line cutting through the graph of a relation, and then observe the points of intersections. The vertical line test supports the definition of a function. That is, every x-value of a function must be paired to a single y-value. If a vertical line intersects the graph of a relation at exactly one point, it implies that a single x-value is only paired to a unique value of y.

The definition of an "elliptic curve" is the set of all points such that $y^2 = x^3 + ax + b$. That is, $E_{a,b,c} \equiv \{(x, y) : y^2 = x^3 + ax + b\}$.

Let a solution $(x_0, y_0) \in E_{a,b,c}$. then by definition $y_0^2 = x_0^3 + ax_0 + b$. So, $(-y_0)^2 = y_0^2 = x_0^3 + ax_0 + b$. Therefore, $(x_0, -y_0) \in E_{a,b,c}$. Hence, this will fail the vertical line test at $x = x_0$.

That is to say, an elliptic curve cannot be a function!

The graph of relation $yRx \Leftrightarrow y^2 = x^3 - x$ above (as we can't use the word "function" in this case...) is said to be "disconnected"; more specifically, it consists of two "connected components". This is a general topological concept that can be applied to all sets, not just to graphs of relations.

7.1.2.1 Limits and Continuity of functions

We will now consider ordered variables of a special type, which we define by the relation "the variable tends to a limit." In what will follow, the concept of limit of a variable will play a fundamental role, being intimately related to the basic notions of mathematical analysis, derivatives, integrals, etc.

Definition (#7.1.273): The number a is named the "limit" of variable magnitude x, if for any arbitrarily small positive number ε we have:

$$|x-a|<\varepsilon \tag{7.1.83}$$

If the number a is the limit of the variable x, we say that "x tends to the limit a".

We can also define the concept of limit from geometrical considerations (this can help to better understand ... but not always ...):

The constant number a is the limit of the variable x, if for any given neighbourhood, no matter how small, of center a and of radius ε, we can find a value x such that all the points corresponding to the following values of the variable belong to this neighbourhood (notions that we defined earlier). We represent geometrically this as:

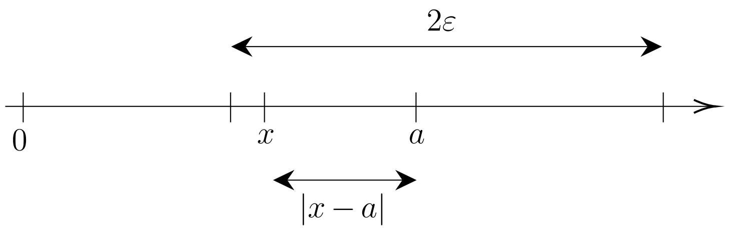


Figure 7.49: **Geometric concept of limit in $\mathbb{R}^3$**

Remarks

R1. It should be trivial that the limit of a constant value is equal to this constant, since the inequality $|x-c^{te}|=|c^{te}-c^{te}|=0<\varepsilon$ is always satisfied for an arbitrary $\varepsilon>0$.

R2. Not all variable have limits. For example $y=\sin(x)$ as this trigonometric function fluctuates between $[-1,+1]$ from $[-\infty,+\infty]$.

Definition (#7.1.274): A variable x tends to infinity if for any positive chosen M, we indicate one value of x from which all successive values of the variable x (values in the neighbourhood of the previous chosen value) satisfy the inequality $|x|>M$. Formally:

$$|x|\rightarrow+\infty,\forall M>0: \quad |x|>M \tag{7.1.84}$$

- A variable x "tends to $+\infty$" if for any positive chosen $M>0$, we indicate one value of x from which all the successive values of the variables x satisfy the inequality $M<x$.

  It is typically the type of consideration that we have for divergent sequences (divergent to infinity) where for a given term of value M of the sequence all the other terms are greater than M:

$$x \to +\infty, \forall M > 0: \quad x > M \tag{7.1.85}$$

- A variable x "tends to $-\infty$" if for any negative chosen $M < 0$, we indicate one value of x from which all the successive values of the variable x satisfy the inequality $x < M$:

$$x \to -\infty, \forall M < 0: \quad x < M \tag{7.1.86}$$

Definition (#7.1.275): Given $y = f(x)$ a function defined in a neighbourhood of a or on some point of this neighbourhood. The function $y = f(x)$ tends to the limit L (that is to say $y \to L$) when x tends to a (that is to say $x \to a$) if for any positive number ε as small as possible, we can indicate a positive number δ such that all x different from a satisfying the inequality $0 < |x - L| < \delta$ also satisfy $|f(x) - L| < \varepsilon$. Formally a function has a limit L on L when in a domain E if:

$$\boxed{\lim_{x \to a} = L \Leftrightarrow \forall x \in E, \exists L, \forall \delta > 0, \forall \varepsilon > 0: \quad 0 < |x - a| < \delta \Rightarrow |f(x) - L| < \varepsilon} \tag{7.1.87}$$

The relation above is named the "epsilon-delta definition of limit or "(ε, δ) definition of limit".

The inequality $0 < |x - a| < \delta$ gives the possibility to have the distance from which we come with our x without taking care of the direction (left or right) as we take for measurement of distance the absolute values. Indeed on a system of axis representing ordinates values, we can, for a given value, coming from the left or from the right (if necessary you can imagine a bus coming to a bus stop that can from the left or from the right only since the absolute distance from it to the bus stop is less than or equal to δ).

If L is the limit of the function $f(x)$ when $x \to a$ we then write in this textbook in any case:

$$\lim_{x \to L} f(x) = L \tag{7.1.88}$$

Obviously the above definition is available when $a = \pm\infty$ or/and $L = \pm\infty$!

To define the direction from which we come from by applying the limit, we use a special notation (recall that this will give us the information of which side of the road comes our bus from...). Thus, if $f(x)$ tends towards the limit L_1 when x approaches a number a by taking only values smaller than a, then we write:

$$\lim_{x \to a_-} f(x) = L_1 \tag{7.1.89}$$

(notice the small $-$ subscript) and we name L_1 the "left limit" of the function $f(x)$ at point a (because remember that the horizontal axis goes from left to right from $-\infty$ to $+\infty$, so small values compared to a given value, are on the left).

If x takes values greater than a, then we will write:

$$\lim_{x \to a_+} f(x) = L_2 \tag{7.1.90}$$

(notice the small $+$ subscript) and we name L_2 the "right limit" of the function $f(x)$ at point a.

In the figure below we have for example:

$$\lim_{x\rightarrow -3_+} f(x)=-\infty \qquad \lim_{x\rightarrow -3_-} f(x)=+\infty \tag{7.1.91}$$

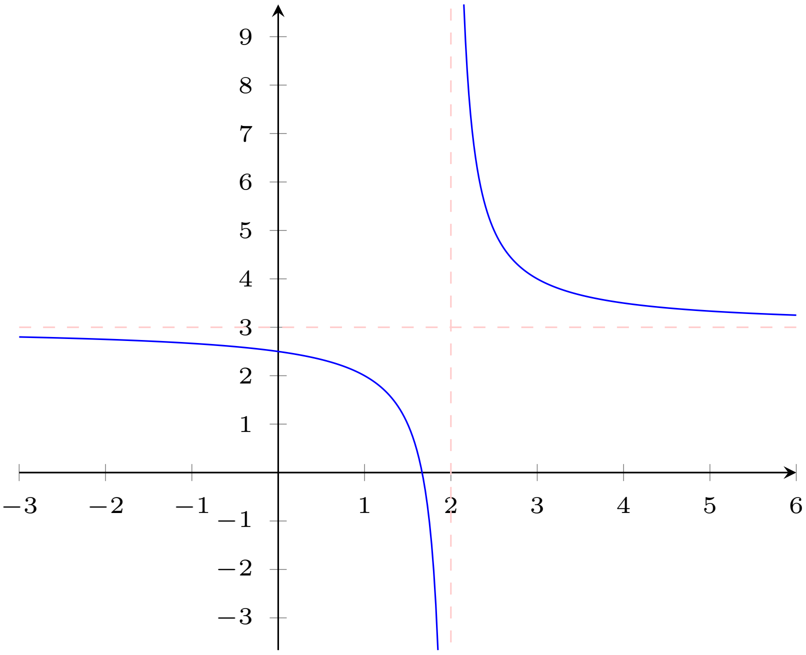


Figure 7.50: Left and right limit examples

It is not always easy (or even possible) to calculate limits of some functions. Let us see some typical examples:

✍**Examples:**

E1. Let us prove that:

$$\lim_{x\rightarrow +\infty}\dfrac{x+1}{x}=1 \tag{7.1.92}$$

is true. For this purpose we have to prove that for any small ε the inequality:

$$\left|1+\dfrac{1}{x}-1\right|<\varepsilon \tag{7.1.93}$$

will be satisfied as soon as $|x|>M$ where M is defined by the choice of ε. The previous inequality is obviously equal to:

$$\left|\dfrac{1}{x}\right|<\varepsilon \tag{7.1.94}$$

which is satisfied if we have x:

$$\left|x\right|>\dfrac{1}{\varepsilon}=M \tag{7.1.95}$$

We admit that the example and the method can be discussed.... But in fact it is only an application of the l'Hôpital's rule (ratio of the derivatives) already proved in the section of Differential and Integral Calculus at page 1461. The reader must also know that we will see also other techniques to determine limits further below with better examples.

E2. Now using Taylor series and change of variables consider we want to calculate:

$$\lim_{x\to+\infty}\sqrt{(x^2+5x)}-x \tag{7.1.96}$$

The method is quite intuitive. Indeed, first we do a change of variable:

$$\begin{aligned}\lim_{x\to+\infty}\sqrt{(x^2+5x)}-x&=\lim_{x\to+\infty}x\sqrt{1+\frac{5}{x}}-x\\&=\lim_{x\to+\infty}x\left(\sqrt{1+\frac{5}{x}}-1\right)=\lim_{y\to 0}\frac{1}{y}\left(\sqrt{1+5y}-1\right)\end{aligned} \tag{7.1.97}$$

Now consider the Taylor series about $x=0$ for the function $f(x)=\sqrt{1+ax}$. We have:

$$f'(x)=\frac{1}{2}a(1+ax)^{-1/2}\qquad f''(x)=-\frac{1}{4}(1+ax)^{-3/2} \tag{7.1.98}$$

Which gives:

$$f(x)\cong 1+\frac{1}{2}ax+\mathcal{O}(x^2) \tag{7.1.99}$$

as a Taylor expansion about $x=0$. Applying this to our limit we see that:

$$\lim_{y\to 0}\frac{1}{y}\left(\sqrt{(1+5y)}-1\right)\cong\lim_{y\to 0}\frac{1}{y}\left(1+\frac{5}{2}y+\mathcal{O}(y^2)-1\right)=\lim_{y\to 0}\frac{1}{y}\left(\frac{5}{2}+\mathcal{O}(y)\right)=\frac{5}{2} \tag{7.1.100}$$

E3. We want to calculate the limit of:

$$\lim_{x\to+\infty}\arctan(x) \tag{7.1.101}$$

How can we deal with something like this? An idea is the to remember that it also implicitly means:

$$\arctan(+\infty)=\theta \tag{7.1.102}$$

Hence:

$$+\infty=\tan(\theta) \tag{7.1.103}$$

And to answer what is the value of θ, we refer to the plot of the function $\tan(\theta)$ and we see then that the first corresponding value is $\pi/2$, therefore:

$$\lim_{x\to+\infty}\arctan(x)=\frac{\pi}{2} \tag{7.1.104}$$

The signification of the symbols $x \to -\infty$ and $x \to +\infty$ makes obvious the signification of the expressions:

$$f(x) \text{ tends to } L \text{ when } x \to +\infty \tag{7.1.105}$$

and:

$$f(x) \text{ tends to } L \text{ when } x \to -\infty \tag{7.1.106}$$

that we denote formally by:

$$\lim_{x \to +\infty} f(x) = L \quad \text{and} \quad \lim_{x \to -\infty} f(x) = L \tag{7.1.107}$$

We have defined the case where the function $f(x)$ tends to a certain limit L when $x \to a_{+,-}$ or $x \to \pm\infty$. Now let us consider the case where the function tend to infinity when the variable x change in a certain way.

We then have typically and obviously:

$$\lim_{x \to a} f(x) = +\infty \qquad \text{or} \qquad \lim_{x \to a} f(x) = +\infty \tag{7.1.108}$$

Or when we need to indicate the direction:

$$\lim_{x \to a_{+,-}} f(x) = +\infty \qquad \text{or} \qquad \lim_{x \to a_{+,-}} f(x) = +\infty \tag{7.1.109}$$

If the function $f(x) \to +\infty$ when $a \to +\infty$ then we write:

$$\lim_{x \to +\infty} f(x) = +\infty \tag{7.1.110}$$

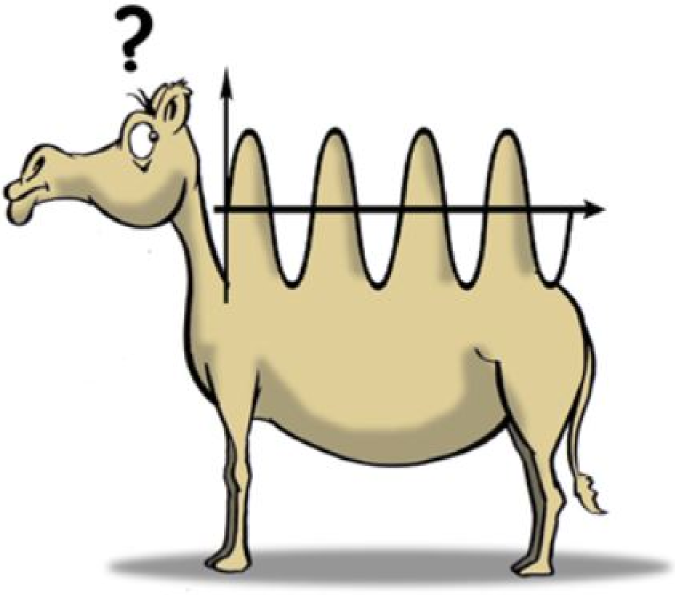
And as we have four possibilities for the sign, we write:

$$\lim_{x \to \pm\infty} f(x) = \pm\infty \tag{7.1.111}$$

that is to say the four following possibilities:

$$\lim_{x \to +\infty} f(x) = +\infty, \; \lim_{x \to -\infty} f(x) = +\infty, \; \lim_{x \to +\infty} f(x) = -\infty, \; \lim_{x \to -\infty} f(x) = -\infty \tag{7.1.112}$$

And once again don't forget as we already mentioned before that some function such as for example $f(x) = \sin(x)$ don't have any finite limit when $x \to \pm\infty$. Then we say that the function is just "bounded" (see section Set Theory page 282).



Now that we've roughly an overview of the concept of limit, we will give an extremely important definition that has a very important place of many areas of high-level mathematics, theoretical physics and computing science (numerical methods).

Definition (#7.1.276): Given a function $f(x)$ and one of its subdomain (or whole one) E (most of time $E \subseteq \mathbb{R}$) and $x_0 \in E$, we say that we have a "continuous function" on x_0 if and only if:

$$\lim_{x \to x_0} f(x) = f(x_0) \tag{7.1.113}$$

That is to say more formally (you have to be able to read the fact that we are going close in an infinitely small way of a limit and this allows the continuity):

$$\boxed{\forall x_0 \in E \forall \varepsilon > 0 \exists \delta > 0 \forall x \in E: \quad |x - x_0| < \delta \Rightarrow |f(x) - f(x_0)| < \varepsilon} \tag{7.1.114}$$

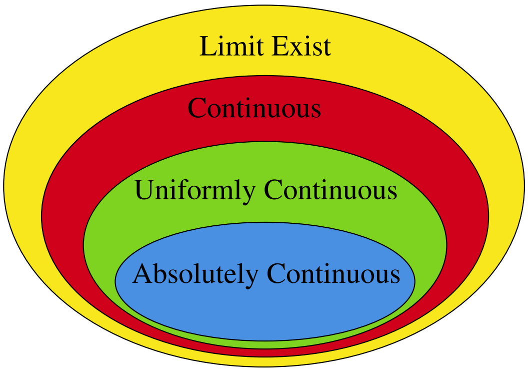
named the "epsilon-delta continuity" or "(ε, δ) continuity".

Remember that $|x - x_0| < \delta \Rightarrow |f(x) - f(x_0)| < \varepsilon$ says when the distance of the input relative to x_0 is small, the distance of the output relative to $f(x_0)$ must be small. How small? Arbitrarily small! If $|a - b| < \varepsilon$ for all $\varepsilon > 0$, then $a = b$! The only way $|x - x_0|$ can be small is if x was actually really close to x_0. That is the purpose of δ! To measure how close we are to x_0. So when we move close to x_0, we should expect to move closer to $f(x_0)$!

In other words: a function is "continuous" if for every point x_0 in the domain E, we can make the images of that point ($f(x_0)$) and another point ($f(x)$) arbitrarily close (of a distance ε) if we move the other point (x) close enough (distance δ) to our given point (x_0).

Caution! Some people have trouble understanding the difference between the definition of a limit of a function and definition of continuity. The difference is that for the limit $0 < |x - L| < \delta$ means that we are not imposing any conditions on the function for $x = L$, while in the definition of continuity, we have $|x - x_0| < \delta$. The difference is that in the definition of continuity at x_0, we insist that the function also be defined at $x = x_0$, but we don't in the definition of limit!

The latter relation will be generalized a little bit in the section of Topology and completed with the concept of... "uniform continuity" (that can itself be generlazied by "absolute continuity" but that's out of the scope of this textbook)!



Remarks

R1. f is "continuous on the left" or respectively "continuous on the right", if we add to the definition above the condition $x > x_0$, respectively $x < x_0$.

R2. A continuous function with a continuous inverse function is named a "homeomorphism".

R3. Instead of saying when it is necessary that a function is not continuous on x_0 or on a given domain, some practitioners prefer to say that the function has an "oscillation".

We have the following trivial corollaries:

C1. $f(x)$ is continuous on x_0 if and only if $f(x)$ is continuous on the left and on the right of x_0.

C2. $f(x)$ is continuous on E if and only if $f(x)$ is continuous on any point of E.

Let $f(x)$ and $g(x)$ be two functions that are continuous at x_0, then we have the following properties (we will prove some of them only on readers requests!):

P1. $f(x)+g(x)$ (also denoted $(f+g)(x)$) is also epsilon-delta continuous at x_0.

As a reader requested the proof of that claim, let's do it!

If $f(x)$ and $g(x)$ are epsilon-delta continuous at a, then $(f+q)(x)$ (shortcut notation for $f(g)+g(x)$) is also epsilon-delta continuous at a.

So far a given $\varepsilon>0$ we have to find $\delta>0$ so if $0|x-a|<\delta$, then $|(f+q)(x)-(f+g)(a)|<\varepsilon$.

First we have obviously:

$$|f(x)+g(x)-(f(a)+g(a))|=|f(x)+g(x)-f(a)-g(a)| \tag{7.1.115}$$

But because of the triangle inequality of absolute value we have:

$$|f(x)-f(a)+g(x)-g(a)|\leq|f(x)-f(a)|+|g(x)-g(a)| \tag{7.1.116}$$

As $f(x)$ and $g(x)$ are assumed epsilon-delta continuous, i.e. $|f(x)-f(a)|<\varepsilon$ when $<|x-a|<\delta_f$ and $|g(x)-f(a)|<\varepsilon$ when $|x-a|<\delta_g$ we have therefore for the previous relation:

$$\underbrace{|f(x)-f(a)|}_{\leq\varepsilon}+\underbrace{|g(x)-g(a)|}_{\leq\varepsilon}\leq 2\varepsilon \tag{7.1.117}$$

Ok you have understand so far how we get that last relation and especially the 2ε, let us now reverse the reasoning! We have from the latter:

$$\begin{gathered}|f(x)-f(a)|+|g(x)-g(a)|<2\varepsilon\\ |f(x)-f(a)+g(x)-g(a)|<2\varepsilon\end{gathered} \tag{7.1.118}$$

and let δ_h be defined as $\min(\delta_g, \delta_f)$ and $h(x)$ be defined as $f(x)+g(x)$. Therefore:

$$|h(x)-h(a)|<2\varepsilon \tag{7.1.119}$$

We can obviously replace δ_f and δ_g by δ_h to get:

$$|f(x)-f(a)|<\varepsilon \text{ when } |x-a|<\delta_h \text{ and } |g(x)-g(a)|<\varepsilon \text{ when } |x-a|<\delta_h \tag{7.1.120}$$

Since:

$$|f(x)-f(a)|+|g(x)-g(a)|<2\varepsilon \tag{7.1.121}$$

that we have derived just before is true when $0<|x-a|<\delta_h$, we can find a δ_h for every value of 2ε.

Finally that means that for every $2\varepsilon>0$ there is a $\delta>0$ such that $|x-a|<\delta \Longrightarrow |h(x)-h(a)|<2\varepsilon$ and for every 2ε there exists an ε so for every ε there exists a δ.

P2. $f(x)-g(x)$ (also denoted $(f-g)(x)$) is also epsilon-delta continuous at x_0.

The proof is immediate from the previous proof as we just have to put $-g(x)=h(x)$ (we will admit that the negative sign of a epsilon-delta continuous function is still epsilon-delta continuous without proof) and therefore we fall back on a sum $f(x)+h(x)$.

P3. $c^{te}f(x)$ is also epsilon-delta continuous at x_0.

P4. $f(x)\cdot g(x)$ (also denoted $fg(x)$) is also epsilon-delta continuous at x_0.

Let $f,g:\mathbb{R}\mapsto\mathbb{R}$ be given such that f and g are epsilon-delta continuous. Let $u:\mathbb{R}\mapsto\mathbb{R}$ be defined by $u(x)=f(x)g(x)$. We want to show: u is epsilon-delta continuous, that is, for all $x_0\in\mathbb{R}$, for every $\varepsilon>0$, there exists some $\delta>0$ such that for all $x\in\mathbb{R}$ with $|x-x_0|<\delta$, $|u(x)-u(x_0)|<\varepsilon$.

We have first using a firs time the properties of the triangle inequality and of the multiplicativeness of the absolute value:

$$\begin{aligned}|f(x)g(x)-f(x_0))g(x_0))|&=|f(x)g(x)-f(x_0))g(x)+f(x_0))g(x)-f(x_0)g(x_0))|\\&\leq|f(x)g(x)-f(x_0))g(x)|+|f(a)g(x)-f(x_0)g(x_0)|\\&=|\,(g(x)[f(x)-f(x_0)]|+|f(x_0)[g(x)-g(x_0)]\,|\\&=|g(x)|\cdot|f(x)-f(x_0)|+|f(x_0)|\cdot|g(x)-g(x_0)|\end{aligned} \tag{7.1.122}$$

Now the $|f(x)-f(x_0)|$ and $|g(x)-g(x_0)|$ terms are good, since $f(x)$ and $g(x)$ are epsilon-delta continuous at x_0. Moreover, the $|g(x_0)|$ term is good since it is constant!

The problem is $g(x)$. Let's see if we can do something clever to get rid of it?! We can try to rewrite is using the triangular inequality (again...) of the absolute value:

$$|g(x)|=|g(x)-g(x_0)+g(x_0)|\leq|g(x)-g(x_0)|+|g(x_0)| \tag{7.1.123}$$

Now remember that $g(x)$ is continuous, therefore $\forall x\in\mathbb{R}$ with $|x-x_0|<\delta$ we have:

$$|g(x)-g(x_0)|<\varepsilon_g \tag{7.1.124}$$

Therefore we can rewrite the prior-previous inequality as:

$$|g(x)| \leq \underbrace{|g(x)-g(x_0)|}_{<\varepsilon_g}+|g(x_0)|<\varepsilon_g+|g(x_0)| \tag{7.1.125}$$

So we get finally a relationship that may help us to get rid of the $g(x)$ that mades us problem earlier:

$$|g(x)|<\varepsilon_g+|g(x_0)| \tag{7.1.126}$$

And we can inject it in the step that was making us troubles, that is:

$$\begin{aligned}|f(x)g(x)-f(x_0)g(x_0)| &\leq |g(x)|\cdot|f(x)-f(x_0)|+|f(x_0)|\cdot|g(x)-g(x_0)|\\ &<(\varepsilon_g+|g(x_0)|)\cdot|f(x)-f(x_0)|+|f(x_0)|\cdot|g(x)-g(x_0)|\end{aligned} \tag{7.1.127}$$

ie:

$$|f(x)g(x)-f(x_0)g(x_0)<(|g(x_0)|+\varepsilon_g)\cdot|f(x)-f(x_0)|+|f(x_0)|\cdot|g(x)-g(x_0)| \tag{7.1.128}$$

But as $f(x)$ and $g(x)$ are epsilon-delta continuous we know that we have $\varepsilon_f,\varepsilon_g$ that exist such that:

$$|f(x)-f(x_0)|<\varepsilon_f \qquad \text{and} \qquad |g(x)-g(x_0)|<\varepsilon_g \tag{7.1.129}$$

Therefore:

$$\begin{aligned}|f(x)g(x)-f(x_0)g(x_0)| &<(|g(x_0)|+\varepsilon_g)\cdot|f(x)-f(x_0)|+|f(x_0)|\cdot|g(x)-g(x_0)|\\ &=(|g(x_0)|+\varepsilon_g)\cdot\varepsilon_f+|f(x_0)|\cdot\varepsilon_g\end{aligned} \tag{7.1.130}$$

We may stop here and the proof is finished as we have only constant terms that can be infinitely small. However mathematicians like to finish[3] the proof with only one term on the right hand side of the equality that looks like a single ε. So obviously as we have two terms that adds together we should have something that looks like $\varepsilon/2+\varepsilon/2=\varepsilon$. For that purpose, mathematicians get rid of the ugly $(|g(x_0)|+\varepsilon_g)\cdot\varepsilon_f$ by defining (they absolutely can its not forbidden):

$$\varepsilon_f=\frac{\varepsilon}{2(|g(x_0)|+\varepsilon_g)} \tag{7.1.131}$$

Therefore we have:

$$\begin{aligned}|f(x)g(x)-f(x_0)g(x_0)| &<(|g(x_0)|+\varepsilon_g)\cdot\varepsilon_f+|f(x_0)|\cdot\varepsilon_g\\ &=(|g(x_0)|+\varepsilon_g)\cdot\frac{\varepsilon}{2(|g(x_0)|+\varepsilon_g)}+|f(x_0)|\cdot\varepsilon_g\\ &=\frac{\varepsilon}{2}+|f(x_0)|\cdot\varepsilon_g\end{aligned} \tag{7.1.132}$$

---

[3] There is a deep reason for that! If we can't finish the proof that way the function will be sure "continuous" but not "uniform continuous" (that's a property most engineers don't care about). For example x^2 or e^x are two functions that are epsilon-delta continuous but not uniform continuous. The impact is out of the scope of this textbook dedicated for reminder to engineers.

Ok it looks better however for mathematicians its still ugly because of the term $|f(x_0)|\cdot\varepsilon_g$. So you may thing we should simple put:

$$\varepsilon_g=\frac{\varepsilon}{2|f(x_0)|} \tag{7.1.133}$$

and it's done as the sum of the two terms would give ε! But it's not!!! Because if in the special case where we may have $|f(x_0)|=0$ then we would have a division by zero! So an idea would be to add a value, that we will denote $\Delta\neq 0$ such that the denominator can never be zero. Let us put for example:

$$\varepsilon_g=\frac{\varepsilon}{2|f(x_0)|+\Delta} \tag{7.1.134}$$

However if we do that we don't have:

$$|f(x_0)|\frac{\varepsilon}{2|f(x_0)|+\Delta}\neq\frac{\varepsilon}{2} \tag{7.1.135}$$

We see obviously that the only solution to get the expected (pretty) result would be to add a term in the firs factor such that:

$$\left(|f(x_0)|+\underbrace{\frac{\Delta}{2}}_{\text{we add that}}\right)\frac{\varepsilon}{2|f(x_0)|+\Delta}=\frac{\varepsilon}{2} \tag{7.1.136}$$

However... can we do that? In other words, can we write something like this (if we come some steps backward):

$$|f(x)g(x)-f(x_0)g(x_0)|\overset{?}{<}(|g(x_0)|+\varepsilon_g)\cdot\varepsilon_f+\left(|f(x_0)|+\frac{\Delta}{2}\right)\cdot\varepsilon_g \tag{7.1.137}$$

And we see obviously that the answer is YES if and only if $\Delta>0$.

Therefore to summarize all we went through, we have:

$$\begin{aligned}|f(x)g(x)-f(x_0)g(x_0)|&=|f(x)g(x)-f(x_0)g(x)+f(x_0)g(x)-f(x_0)g(x_0)|\\&\leq|f(x)g(x)-f(x_0)g(x)|+|f(x_0)g(x)-f(x_0)g(x_0)|\\&=|g(x)(f(x)-f(x_0)|+|f(x_0)(g(x)-g(x_0))|\\&=|g(x)|\cdot|f(x)-f(x_0)|+|f(x_0)|\cdot|g(x)-g(x_0)|\\&<(|g(x_0)|+\varepsilon)\cdot\frac{\varepsilon}{2(|g(x_0)|+\varepsilon)}+|f(x_0)|\cdot|g(x)-g(x_0)|\\&\leq\frac{\varepsilon}{2}+|f(x_0)|\cdot\frac{\varepsilon}{2|f(x_0)|+\Delta}\\&<\frac{\varepsilon}{2}+\left(|f(x_0)|+\frac{\Delta}{2}\right)\cdot\frac{\varepsilon}{2|f(x_0)|+\Delta}\\&=\frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\varepsilon\end{aligned} \tag{7.1.138}$$

And that's the pretty final result mathematicians like to have!

P5. $|f(x)|$ is also epsilon-delta continuous at x_0.

Let $\varepsilon > 0$ be given. Then, since $f(x)$ is continuous at x_0, there is $\delta > 0$ such that if $|x - x_0| < \delta$, then $|f(x) - f(x_0)| < \varepsilon$.

With that same δ, if $|x - x_0| < \delta$, then by the reverse triangle inequality (see section Arithmetic Operators page 326), which says $|a-b| \geq ||a| - |b||$, we have:

$$||f(x)| - |f(x_0)|| \leq |f(x) - f(x_0)| < \varepsilon \tag{7.1.139}$$

Therefore $|f(x)|$ is epsilon-delta continuous at x_0

P6. $1/f(x)$ is also epsilon-delta continuous at x_0.

This time we need to estimate:

$$\left|\frac{1}{f(x)} - \frac{1}{f(x_0)}\right| = \left|\frac{f(x_0) - f(x)}{f(x)f(x_0)}\right| = \frac{|f(x) - f(x_0)|}{|f(x)||f(x_0)|} \tag{7.1.140}$$

The $|f(x) - f(x_0)|$ term is good, and the $|f(x_0)|$ term is good as well (since it is constant).

The only term we need to control is the $|f(x)|$ term.

Since we want $\frac{1}{|f(x)|} <$ something, we need $|f(x)| >$ something!

Using the inverse triangular inequality of the absolute value, we know that we have:

$$|f(x) - f(x_0)| \geq ||f(x)| - |f(x_0)|| \tag{7.1.141}$$

But as $|a| \geq -a$ for all a, then we can write:

$$|f(x) - f(x_0)| \geq ||f(x)| - |f(x_0)|| \geq -(|f(x)| - |f(x_0)|) = |f(x_0)| - |f(x)| \tag{7.1.142}$$

Therefore we get:

$$|f(x_0)| - |f(x)| \leq |f(x) - f(x_0)| \tag{7.1.143}$$

But now keep in mind that as $f(x)$ is continuous there is a δ_1 such that if $|x - x_0| < \delta_1$, then $|f(x) - f(x_0)| < \varepsilon_1$. So let us choose empirically:

$$\varepsilon_1 = \frac{|f(x_0)|}{2} \tag{7.1.144}$$

Then with this choice (keep in mind that this inequality is valid if and only if $|x - x_0|$ is small enough and $f(x) \neq 0$ and that we may choose something else than 2 at the denominator but we made that choice because we will have a nice result at the end!):

$$|f(x_0)| - |f(x)| \leq |f(x) - f(x_0)| < \frac{|f(x_0)|}{2} \tag{7.1.145}$$

And therefore by rearranging:

$$|f(x)| > |f(x_0)| - \frac{|f(x_0)|}{2} = \frac{|f(x_0)|}{2} > 0 \tag{7.1.146}$$

Hence:

$$\frac{1}{|f(x)|}<\frac{2}{|f\left(x_0\right)|} \tag{7.1.147}$$

And that good it's want we wanted remember! We wanted $\frac{1}{|f(x)|}<$ something, we needed $|f(x)|>$ something! So now we have it!

Therefore:

$$\begin{aligned}\left|\frac{1}{f(x)}-\frac{1}{f\left(x_0\right)}\right| &= \frac{|f(x)-f\left(x_0\right)|}{|f(x)|\,|f\left(x_0\right)|}\\ &\leq \frac{|f(x)-f\left(x_0\right)|}{|f\left(x_0\right)|}\left(\frac{2}{|f\left(x_0\right)|}\right)\\ &= |f(x)-f\left(x_0\right)|\left(\frac{2}{|f\left(x_0\right)|^2}\right)\\ &?<\varepsilon\end{aligned} \tag{7.1.148}$$

So are we stuck? No! Because for the last line above we see that it would be just enough to have:

$$|f(x)-f\left(x_0\right)|<\frac{\varepsilon}{2}|f\left(x_0\right)|^2 \tag{7.1.149}$$

And this is something we can also choose as $f(x)$ is epsilon-delta continuous and $f(x)\neq 0$.

Therefore we can write the complete proof starting from here now that we have all the tools! :

Let $\varepsilon>0$ be given.

Then, since $f(x)$ is epsilon-delta continuous at x_0, there is $\delta_1>0$ such that if $|x-x_0|<\delta_1$, then $|f(x)-f\left(x_0\right)|<\frac{|f(x_0)|}{2}$, which implies $|f(x)|>\frac{|f(x_0)|}{2}$, and therefore $\frac{1}{|f(x)|}<\frac{2}{|f(I_0)|}$.

Moreover, since $f(x)$ is epsilon-delta continuous at x_0, there is $\delta_2>0$ such that if $|x-x_0|<\delta_2$, then $|f(x)-f\left(x_0\right)|<\frac{\varepsilon}{2}|f\left(x_0\right)|^2$.

Let $\delta=\min\left\{\delta_1,\delta_2\right\}>0$, then, if $|x-x_0|<\delta$, then:

$$\begin{aligned}\left|\frac{1}{f(x)}-\frac{1}{f\left(x_0\right)}\right| &= \frac{|f(x)-f\left(x_0\right)|}{|f(x)|\,|f\left(x_0\right)|}\\ &\leq \left(\frac{|f(x)-f\left(x_0\right)|}{|f\left(x_0\right)|}\right)\left(\frac{2}{|f\left(x_0\right)|}\right)\\ &= |f(x)-f\left(x_0\right)|\left(\frac{2}{|f\left(x_0\right)|^2}\right)\\ &< \left(\frac{\varepsilon|f\left(x_0\right)|^2}{2}\right)\left(\frac{2}{|f\left(x_0\right)|^2}\right)\\ &= \varepsilon\end{aligned} \tag{7.1.150}$$

And that finish the proof! If $f(x)$ is epsilon-delta continuous then we have indeed $\frac{1}{f}$ that is also epsilon-delta continuous at x_0.

P7. $f(x)/g(x)$ (also denoted $f/g(x)$) is also epsilon-delta continuous at x_0.

This is proven immediately using the property of epsilon-delta continuity of the inverse $(1/f(x))$ and of the multiplication $(f(x)\cdot 1/g(x))$.

P8. $f^n(x)$ is also epsilon-delta continuous at x_0.

This is immediate as it is product n times of $f\cdot f\cdot f\ldots$ as it follows from the property P4.

P9. $\sqrt[n]{x}$ is also epsilon-delta continuous at x_0.

The proof is given in the list of usual limits and continuities further below. What is much more useful is to prove the little bit more general case $\sqrt{f(x)}$ is epsilon-delta continuous!

Suppose you have to prove that $\sqrt{f(x)}$ is a continuous function (where $f(x)$ is continuous). We go by the definition of continuity. Take any $\varepsilon>0$. Then:

$$|\sqrt{f(x)}-\sqrt{f(y)}|^2=f(x)+f(y)-2\sqrt{f(x)f(y)} \tag{7.1.151}$$

Since $f(x)$ should be non-negative:

$$\begin{gathered}|\sqrt{f(x)}-\sqrt{f(y)}|^2<f(x)+f(y)=|f(x)+f(y)|\\ \Longrightarrow|\sqrt{f(x)}-\sqrt{f(y)}|<\sqrt{|f(x)+f(y)}\end{gathered} \tag{7.1.152}$$

Since $f(x)$ is continuous we know that there exist a $\delta>0$ such that:

$$|f(x)+f(y)|<\varepsilon^2 \text{ whenever } |x-y|<\delta \tag{7.1.153}$$

Hence:

$$|\sqrt{f(x)}-\sqrt{f(y)}|<\sqrt{|f(x)+f(y)}|<\varepsilon \text{ whenever } |x-y|<\delta \tag{7.1.154}$$

and therefore $\sqrt{f(x)}$ is epsilon-delta continuous!

P10. $g\circ f$ (also denoted $g(f(x))$ is also epsilon-delta continuous at x_0.

First a small reminder may be useful!:

If A,B,C are subsets of $\mathbb{R}$ and $f:A\rightarrow B$ and $g:B\rightarrow C$ are functions, then the composition $g\circ f:A\rightarrow C$ is defined by:

$$(g\circ f)(x)=g(f(x)) \tag{7.1.155}$$

Let's show that the composition of two epsilon-delta continuous functions is epsilon-delta continuous! In other words: If $f(x)$ is epsilon-delta continuous at x_0 and g is epsilon-delta continuous at $f(x_0)$, then $g\circ f$ is epsilon-delta continuous at x_0.

Let $\varepsilon>0$ be given.

Since g is epsilon-delta continuous at $f(x_0)$, there is $\delta'>0$ such that:

$$|x-f(x_0)|<\delta'\Rightarrow|g(x)-g(f(x_0))|<\varepsilon \tag{7.1.156}$$

Therefore there is $\delta'>0$ such that:

$$|f(x)-f(x_0)|<\delta'\Rightarrow|g(f(x))-g(f(x_0))|<\varepsilon \tag{7.1.157}$$

But now, since $f(x)$ is epsilon-delta continuous at x_0, by the definition of epsilon-delta continuity (but with δ' instead of ε), there is $\delta > 0$ such that if $|x-x_0|<\delta$, then $|f(x)-f(x_0)|<\delta'$.

Therefore, with δ as above, if $|x-x_0|<\delta$, then $|f(x)-f(x_0)|<\delta'$ and therefore:

$$|(g\circ f)(x)-(g\circ f)(x_0)| = |g(f(x))-g(f(x_0))| < \varepsilon \tag{7.1.158}$$

Therefore $g\circ f$ is epsilon-delta continuous at x_0.

7.1.2.1.1 Squeeze theorem

The "squeeze theorem" (also known as the "sandwich theorem", among other names[4]) is a theorem regarding the limit of a function that is trapped between two other functions.

The squeeze theorem is used in analysis, typically to confirm the limit of a function via comparison with two other functions whose limits are known. It was first used geometrically by the mathematicians Archimedes and Eudoxus in an effort to compute π, and was formulated in modern terms by Carl Friedrich Gauss. We will use it further below (in association with the definition of continuity) to compute/prove some very usual limits.

The squeeze theorem is formally stated as follow:

Theorem 7.3. Let I be an interval containing the point a. Let $g(x)$, $f(x)$, and $h(x)$ be functions defined on I, except possibly at a itself. Suppose that for every x in I not equal to a, we have:

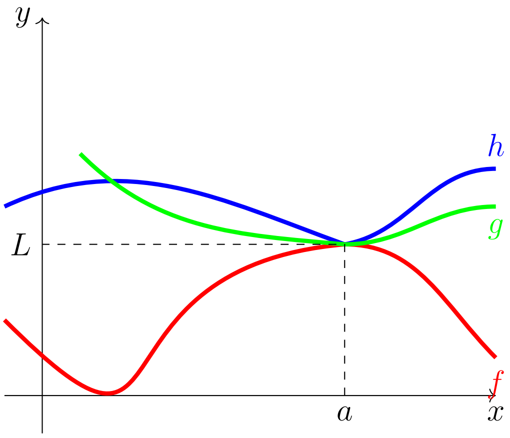
$$g(x)\leq f(x)\leq h(x) \tag{7.1.159}$$

and also suppose that:

$$\lim_{x\to a} g(x) = \lim_{x\to a} h(x) = L \tag{7.1.160}$$

Then:

$$\lim_{x\to a} f(x) = L \tag{7.1.161}$$



---

[4] Also known as the pinching theorem, the sandwich rule, the police theorem, the between theorem and sometimes the squeeze lemma. In Italy, the theorem is also known as the theorem of carabinieri.

The functions $g(x)$ and $h(x)$ are said to be lower and upper bounds (respectively) of $f(x)$ and converges to the limit a.

Here, a is not required to lie in the interior of I. Indeed, if a is an endpoint of I, then the above limits are left- or right-hand limits.

Remark

A similar statement holds for infinite intervals: for example, if $I=[0,+\infty[$, then the conclusion holds, taking the limits as $x\rightarrow+\infty$. This theorem is also valid for sequences. Let $(a_n),(c_n)$ be two sequences converging to ℓ, and (b_n) a sequence. If $\forall n\geq N, N\in\mathbb{N}$ we have $a_n\leq b_n\leq c_n$, then (b_n) also converges to ℓ.

According to the above hypotheses we have, taking the limit inferior and superior:

$$L=\lim_{x\rightarrow a} g(x)\leq\liminf_{x\rightarrow a} f(x)\leq\limsup_{x\rightarrow a} f(x)\leq\lim_{x\rightarrow a} h(x)=L \tag{7.1.162}$$

so all the inequalities are indeed equalities, and the thesis immediately follows.

Proof 7.3.1. A direct proof, using the (ε,δ)-definition of limit above, would be to prove that for all real $\varepsilon>0$ there exists a real $\delta>0$ such that for all x with $|x-a|<\delta$, we have $|f(x)-L|<\varepsilon$. Symbolically:

$$\forall\varepsilon>0,\exists\delta>0:\forall x,(|x-a|<\delta\Rightarrow|f(x)-L|<\varepsilon) \tag{7.1.163}$$

Let us assume that we know (because it is proven in the context of a special case) that:

$$\lim_{x\rightarrow a} g(x)=L \tag{7.1.164}$$

This means that:

$$\forall\varepsilon>0,\exists\delta_1>0:\forall x\,(|x-a|<\delta_1\Rightarrow|g(x)-L|<\varepsilon) \tag{7.1.165}$$

And let us also assume that we also know (because it is proven in the context of a special case) that:

$$\lim_{x\rightarrow a} h(x)=L \tag{7.1.166}$$

This means that:

$$\forall\varepsilon>0,\exists\delta_2>0:\forall x\,(|x-a|<\delta_2\Rightarrow|h(x)-L|<\varepsilon) \tag{7.1.167}$$

Then we have:

$$g(x)\leq f(x)\leq h(x) \tag{7.1.168}$$

and also:

$$g(x)-L\leq f(x)-L\leq h(x)-L \tag{7.1.169}$$

We can choose $\delta := \min\{\delta_1, \delta_2\}$. Then, if $|x-a| < \delta$, combining:

$$\begin{gathered}\forall \varepsilon > 0, \exists \delta_1 > 0 : \forall x\, (|x-a| < \delta_1 \Rightarrow |g(x) - L| < \varepsilon) \\ \text{and} \\ \forall \varepsilon > 0, \exists \delta_2 > 0 : \forall x\, (|x-a| < \delta_2 \Rightarrow |h(x) - L| < \varepsilon)\end{gathered} \tag{7.1.170}$$

We have:

$$-\varepsilon < g(x) - L \leq f(x) - L \leq h(x) - L < \varepsilon \tag{7.1.171}$$

Hence:

$$-\varepsilon < f(x) - L < \varepsilon \tag{7.1.172}$$

Which completes the proof!

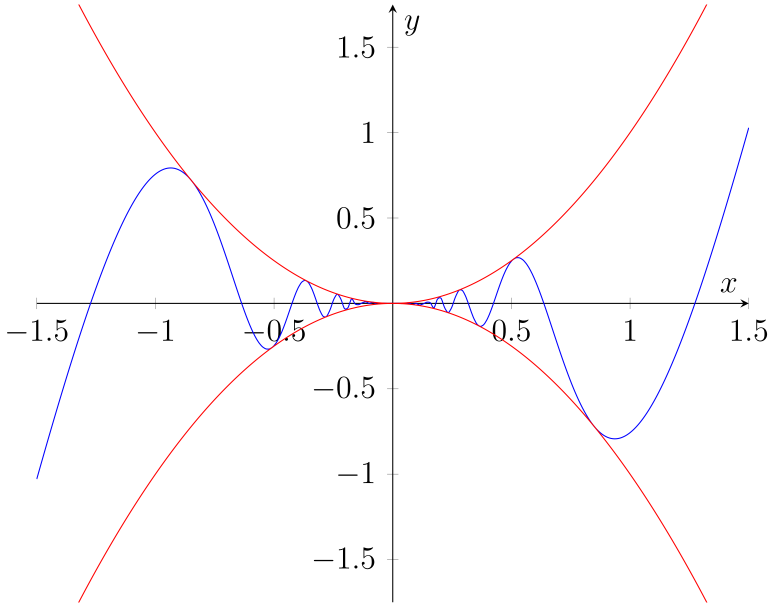
■ Q.E.D.

✍**Example:**

The limit:

$$\lim_{x \to 0} x^2 \sin\left(\frac{1}{x}\right)$$

illustrated below:



Cannot be determined through the limit law:

$$\lim_{x \to a} (f(x) \cdot g(x)) = \lim_{x \to a} f(x) \cdot \lim_{x \to a} g(x)$$

because:

$$\lim_{x \to 0} \sin\left(\frac{1}{x}\right)$$

does not exist.

However, by the definition of the sine function:

$$-1 \leq \sin\left(\frac{1}{x}\right) \leq 1$$

It follows that:

$$-x^2 \leq x^2 \sin\left(\frac{1}{x}\right) \leq x^2$$

Since (assuming that its continuity is proven somewhere):

$$\lim_{x\rightarrow 0} -x^2 = \lim_{x\rightarrow 0} x^2 = 0$$

by the squeeze theorem, $\lim_{x\rightarrow 0} x^2 \sin\left(\frac{1}{x}\right)$ must also be 0.

7.1.2.1.2 Usual Limits (and Continuities)

Delta-epsilon proofs are first found in the works of Augustin-Louis Cauchy (11789–11867 according to holocene calendar).

Here are some usual limits to know (obviously for each of them we will also prove if they are epsilon-delta continuous or not just for curiosity[5]):

1. For any real numbers a and c:

$$\lim_{x\rightarrow a} c = c \tag{7.1.173}$$

   - First we prove the epsilon-delta continuity:
     Let us apply the definition of epsilon-delta continuity. We pick an arbitrary real number x_0. Then for a given $\varepsilon > 0$, we are trying to find a $\delta > 0$ such that for all x:

$$|x - x_0| < \delta \Longrightarrow |f(x) - f(x_0)| = |k - k| = 0 < \varepsilon \tag{7.1.174}$$

     Since $0 < \varepsilon$ is true by hypothesis, we can choose an arbitrary δ like $\delta = \pi^{\pi^{\pi^{\pi}}}$. And this indeed satisfies the condition of epsilon-delta continuity at x_0.
     Since the choice for x_0 was arbitrary, this inequality holds for all real numbers and since the choice for k was arbitrary, this argument holds for all constant functions.
     Therefore, all constant functions are epsilon-delta continuous!
   - Second we prove the epsilon-delta limit:
     This is by definition. Don't need any proof.

2. For any real number a and $n \in \mathbb{N}^*$:

$$\lim_{x\rightarrow a} x^n = a^n \tag{7.1.175}$$

---

[5] Don't forget that the epsilon-delta continuity on a set or subset implies that the limit inside the same set or subset exist too but the contrary is not true!

- First we prove the epsilon-delta continuity:
  So we have to show $\forall\varepsilon>0$ there is a $\delta>0$ such that:

$$|x-x_0|<\delta \Longrightarrow |x^n-x_0^n|<\varepsilon \tag{7.1.176}$$

  We have to show that there is a δ that exists who satisfies the above relation.
  Let ε be any positive number. Denote $x_0+\delta=x$ and $x^n-x_0^n=\Delta$. Then identically:

$$\Delta=(x-x_0)\left(x^{n-1}+x^{n-2}x_0+\ldots+x_0^{n-1}\right) \tag{7.1.177}$$

  Taking the absolute value and using its triangle inequality (see section Arithmetics page 326) gives:

$$|\Delta|=|\delta|\cdot\left|x^{n-1}+x^{n-2}x_0+\ldots+x_0^{n-1}\right|\leq|\delta|\cdot\left(\left|x^{n-1}\right|+\left|x^{n-2}x_0\right|+\ldots+\left|x_0^{n-1}\right|\right) \tag{7.1.178}$$

  But since from the triangular inequality $|x|=|x_0+\delta|\leq|x_0|+|\delta|$ and also $|x_0|\leq|x_0|+|\delta|$, therefore each summand in the parentheses is at most equal to $(|x_0|+|\delta|)^{n-1}$, and since there are n summands, the sum is at most equal to $n\left(|x_0|+|\delta|\right)^{n-1}$. Thus we get:

$$|\Delta|\leqq n|\delta|\left(|x_0|+|h|\right)^{n-1} \tag{7.1.179}$$

  We may choose $|\delta|<1$. This implies:

$$|\Delta|\leq n|\delta|\left(|x_0|+1\right)^{n-1} \tag{7.1.180}$$

  The right hand side of this inequality is less than ε as soon as we still require:

$$|\delta|<\frac{\varepsilon}{n\left(|x_0|+1\right)^{n-1}} \tag{7.1.181}$$

  As we are able to find a δ means that the power function $x\mapsto x^n$ is epsilon-delta continuous at the point x_0 (however as the final relation depends on x_0 notice that it means the function is not uniform continuous!).
- Second we prove the epsilon-delta limit:
  The proof of the epsilon-delta limit is straightforward thanks to the previous result.
  As $x_0=a$, we have:

$$|\delta|<\frac{\varepsilon}{n\left(|a|+1\right)^{n-1}} \tag{7.1.182}$$

  And therefore we can always found a given ε such that the above inequality is satisfied and the limit exist on a. Then we can indeed write:

$$\lim_{x\to a}x^n=a^n \tag{7.1.183}$$

  Or here is another more detailed argument for those who prefer it:

$$\begin{aligned}&f(x)=x^n \text{ is continuous at } x_0>0\\&\Longleftrightarrow \lim_{x\to x_0}x^n=x_0^n\\&\Longleftrightarrow \forall\varepsilon>0,\exists\delta>0:\forall 0<|x-x_0|<\delta,|x^n-x_0^p|<\varepsilon\\&\Longleftrightarrow \forall\varepsilon>0,\exists\delta>0:\forall 0<\left|\frac{x}{x_0}-1\right|<\frac{\delta}{x_0},\left|\left(\frac{x}{x_0}\right)^n-1\right|<\frac{\varepsilon}{x_0^n}\end{aligned} \tag{7.1.184}$$

Let:

$$x'=\frac{x}{x_0},\delta'=\frac{\delta}{x_0},\varepsilon'=\frac{\varepsilon}{x_0^p}$$

Therefore:

$$\begin{aligned}&f(x)=x^n \text{ is continuous at } x_0>0\\ &\Longleftrightarrow \forall\varepsilon'>0,\exists\delta'>0:\forall 0<|x'-1|<\delta',|x'^n-1|<\varepsilon'\\ &\Longleftrightarrow \forall\varepsilon'>0,\exists\delta'>0:\forall 1-\delta<x'<1+\delta,x'\neq 1,1-\varepsilon'<x'^n<1+\varepsilon'\\ &\Longleftrightarrow \forall\varepsilon'>0,\exists\delta'>0:\forall 1-\delta'<x'<1+\delta',x'\neq 1,(1-\varepsilon')^{\frac{1}{n}}<x'<(1+\varepsilon')^{\frac{1}{n}}\end{aligned}\tag{7.1.185}$$

It is obvious to see that, we only need to put:

$$\delta=\min\left\{1-(1-\varepsilon')^{\frac{1}{n}},(1+\varepsilon')^{\frac{1}{n}}-1\right\}\tag{7.1.186}$$

in order the rightmost part of the final step to be satisfied. Therefore the epsilon-delta limit is proved to exist!

3. For any real number $a>0$:

$$\lim_{n\to+\infty}a^{1/n}=\lim_{n\to+\infty}\sqrt[n]{a}=1\tag{7.1.187}$$

- First we prove the epsilon-delta continuity:

  A function is epsilon-delta continuous at x_0 if, for any $\varepsilon>0$, there exists a $\delta>0$ so that, if $|x-x_0|<\delta$, then $|f(x)-f(x_0)|<\varepsilon$.

  Therefore:

$$\begin{aligned}|f(x)-f(x_0)|<\varepsilon\Leftrightarrow&\left|x^{1/n}-x_0^{1/n}\right|<\varepsilon\\ \Leftrightarrow& x_0^{1/n}-\varepsilon<x^{1/n}<x_0^{1/n}+\varepsilon\\ \Leftrightarrow&\left(x_0^{1/n}-\varepsilon\right)^n<x<\left(x_0^{1/n}+\varepsilon\right)^n\end{aligned}\tag{7.1.188}$$

  Thus, if we choose:

$$\delta=\min\left\{\left(c^{1/n}+\varepsilon\right)^n-c,c-\left(c^{1/n}-\varepsilon\right)^n\right\}\text{ for } c\neq 0\tag{7.1.189}$$

  then:

$$|x-x_0|<\delta\Rightarrow|f(x)-f(x_0)|<\varepsilon\tag{7.1.190}$$

  By the way that proves that if $f(x)>0$ is epsilon-delta continuous, then $\sqrt[n]{f(x)}$ is also epsilon-delta continuous!

  If you are not convinced of the last claim there is another way to show that! Further below we will prove in details that the exponential function e^x (see section Functional Analysis page 2289) and logarithm function $\log(x)$ are both epsilon-delta continuous. And as we can write:

$$\sqrt[n]{x}=e^{\ln(x)/n}\tag{7.1.191}$$

and we have proven earlier that if $f(x),g(x)$ are both epsilon-delta continuous then $g(f(x))$ is epsilon-delta continuous, then:

$$\sqrt[n]{x} \tag{7.1.192}$$

is epsilon-delta continuous too! And obviously we can replace x by a epsilon-delta continuous function $f(x)$ and the same reasoning would apply.

- Second we prove the epsilon-delta limit:
  Set $x_n=a^{1/n}$. We have:

$$\ln(x_n)=\frac{1}{n}\cdot\ln(a) \tag{7.1.193}$$

Clearly, we have:

$$\lim_{n\rightarrow+\infty}\ln(x_n)=0 \tag{7.1.194}$$

Hence:

$$\lim_{n\rightarrow+\infty}e^{\ln(x_n)}=e^0=1 \tag{7.1.195}$$

which translates into:

$$\lim_{n\rightarrow+\infty}a^{1/n}=1 \tag{7.1.196}$$

4. For any real numbers a and n:

$$\lim_{x\rightarrow a}\frac{x^n-a^n}{x-a}=na^{n-1} \tag{7.1.197}$$

- First we prove the epsilon-delta continuity:
  We have already proved earlier that if $f(x)$ is epsilon-delta continuous then $f^n(x)$ is epsilon-delta continuous and also that a^n is therefore also epsilon-delta continuous.
  We also have proven earlier that if $f(x),g(x)$ are epsilon-delta continuous then $f(x)-g(x)$ is also epsilon-delta continuous. Therefore we know that x^n-a^n and $x-a$ are epsilon-delta continuous.
  We also have proven earlier that if $f(x),g(x)$ are epsilon-delta continuous then $f(x)/g(x)$ is also epsilon-delta continuous if $g(x)\neq 0$. Therefore we know that:

$$\frac{x^n-a^n}{x-a} \tag{7.1.198}$$

is epsilon-delta continuous!

- Second we prove the epsilon-delta limit:
  We simply use l'Hôpital's rule (see section Differential and Integral Calculus page 1461)! That is:

$$\lim_{x\rightarrow a}\frac{x^n-a^n}{x-a}=\lim_{x\rightarrow a}\frac{\dfrac{\mathrm{d}}{\mathrm{d}x}(x^n-a^n)}{\dfrac{\mathrm{d}}{\mathrm{d}x}(x-a)}=\lim_{x\rightarrow a}nx^{n-1}=na^{n-1} \tag{7.1.199}$$

5. For $a>0$ and n positive integer:

$$\lim_{x\to a}\frac{x-a}{x^n-a^n}=\frac{1}{na^{n-1}} \tag{7.1.200}$$

- First we prove the epsilon-delta continuity:

  The proof is straightforward as it is just the inverse of the previous example!

- Second we prove the epsilon-delta limit:

  It's just the inverse of the previous example so there is nothing special to prove!

6. We have:

$$\lim_{x\to 0}(1+x)^{1/x}=e \tag{7.1.201}$$

- First we prove the epsilon-delta continuity:

  We have already proved earlier that a constant is epsilon-delta continuous and x is also epsilon-delta continuous.

  We have also proved earlier that if $f(x),g(x)$ are epsilon-delta continuous then $f(x)+g(x)$ is epsilon-delta continuous and therefore $1+x$ is epsilon-delta continuous.

  We have also proved earlier that if $f(x)>0$ is epsilon-delta continuous that $f(x)^{1/x}$ is also epsilon-delta continuous. Therefore:

$$(1+x)^{1/x} \tag{7.1.202}$$

  is epsilon-delta continuous!

- Second we prove the epsilon-delta limit:

  This limit is already proven at page 2289.

7. For any real number a :

$$\lim_{x\to+\infty}\left(1+\frac{a}{x}\right)^x=e^a \tag{7.1.203}$$

This limit is well known by students in the special case where $a=1$.

- First we prove the epsilon-delta continuity:

  We have already proved earlier that a constant is epsilon-delta continuous. Therefore 1 is epsilon-delta continuous.

  We have also proved earlier that if $f(x),g(x)\neq 0$ are epsilon-delta continuous then $f(x)/g(x)$ is also epsilon-delta continuous and therefore a/x is epsilon-delta continuous.

  We have also proved earlier that if $f(x),g(x)$ are epsilon-delta continuous then $f(x)+g(x)$ is epsilon-delta continuous and therefore $1+(a/x)$ is epsilon-delta continuous.

  We have also proved earlier that if $f(x)$ is epsilon-delta continuous that $f(x)^n$ is also epsilon-delta continuous. Therefore:

$$\left(1+\frac{a}{x}\right)^x \tag{7.1.204}$$

  is epsilon-delta continuous!

- Second we prove the epsilon-delta limit:

  Let us note that it is equivalent to show that:

  $$\lim_{n\to+\infty}\frac{\ln\left(1+\frac{a}{n}\right)}{1/n}=a \tag{7.1.205}$$

  Do not worry about the domain of $\ln$, since for large n, the expression $1+\frac{a}{n}$ will be a positive number (close to 1). Consider the function:

  $$f(x)=\frac{\ln(1+x)}{x},\quad \text{for } x>0 \tag{7.1.206}$$

  and $f(0)=1$. Using the definition of the derivative of $\ln(1+x)$, we see that $f(x)$ is epsilon-delta continuous at 0, that is, $\lim_{x\to 0}f(x)=1=f(0)$. Hence, for any sequence $\{x_n\}$ which converges to 0, we have:

  $$\lim_{n\to+\infty}f(x_n)=f(0)=1 \tag{7.1.207}$$

  Now, set:

  $$x_n=\frac{a}{n} \tag{7.1.208}$$

  Clearly we have $x_n\to 0$. Therefore, we have:

  $$\lim_{n\to+\infty}\frac{\ln\left(1+\frac{a}{n}\right)}{\frac{a}{n}}=1 \tag{7.1.209}$$

  But, we have:

  $$\frac{\ln\left(1+\frac{a}{n}\right)}{\frac{1}{n}}=a\cdot\frac{\ln\left(1+\frac{a}{n}\right)}{\frac{a}{n}} \tag{7.1.210}$$

  which clearly implies:

  $$\lim_{n\to+\infty}\frac{\ln\left(1+\frac{a}{n}\right)}{1/n}=a \tag{7.1.211}$$

  Since:

  $$\left(1+\frac{a}{n}\right)^n=\exp\left(\frac{\ln\left(1+\frac{a}{n}\right)}{1/n}\right) \tag{7.1.212}$$

  we get:

  $$\lim_{n\to+\infty}\left(1+\frac{a}{n}\right)^n=e^a \tag{7.1.213}$$

8. We have:

$$\lim_{x\to 0}\frac{e^x-1}{x}=1 \tag{7.1.214}$$

- First we prove the epsilon-delta continuity:
  Suppose $|x-x_0|<\delta$ and fix $\varepsilon>0$.
  Consider $|e^x-e^{x_0}|<\varepsilon$, then using an engineer approach[6]:

$$\begin{gathered}-\varepsilon<e^x-e^{x_0}<\varepsilon\\ e^{x_0}-\varepsilon<e^x<e^{x_0}+\varepsilon\\ \ln\left(e^{x_0}-\varepsilon\right)<x<\ln\left(e^{x_0}+\varepsilon\right)\\ \ln\left(e^{x_0}-\varepsilon\right)-x_0<x-x_0<\ln\left(e^{x_0}+\varepsilon\right)-x_0\end{gathered}\tag{7.1.215}$$

  Hence e^x is epsilon-delta continuous at x_0 as long as:

$$\delta=\min\left\{x_0-\ln\left(e^{x_0}-\varepsilon\right),\ln\left(e^{x_0}+\varepsilon\right)-x_0\right\}\tag{7.1.216}$$

  Since x_0 is arbitrarily chosen then e^x is epsilon-delta continuous and we are done.

  Now that we have proven that the exponential is epsilon-delta continuous. We know that the numerator is epsilon-delta continuous and also that the denominator is epsilon-delta continuous. As we have proven earlier that if $f(x)$ and $g(x)$ are epsilon-delta continuous, then $f(x)/g(x)$ is epsilon-delta continuous too, this finish the proof!

- Second we prove the epsilon-delta limit:
  That latter is immediate using l'Hôpital's rule (see section Differential and Integral Calculus page 1461)! We have indeed:

$$\lim_{x\to 0}\frac{e^x-1}{x}=\lim_{x\to 0}\frac{\dfrac{\mathrm{d}}{\mathrm{d}x}(e^x-1)}{\dfrac{\mathrm{d}}{\mathrm{d}x}x}=\lim_{x\to 0}\frac{e^x}{1}=1\tag{7.1.217}$$

9. We have :

$$\lim_{x\to 0}\sin(x)=0\tag{7.1.218}$$

This limit must not be confused with the Maclaurin series that gives for small x that $\sin(x)\cong x$.

- First we prove the epsilon-delta continuity:
  Let $\varepsilon>0$ and $x,y\in\mathbb{R}$. We want:

$$|f(x)-f(y)|<\varepsilon\Longrightarrow|\sin(x)-\sin(y)|<\varepsilon\tag{7.1.219}$$

  Using trigonometric identities this is equivalent to:

$$\left|2\cos\left(\frac{x+y}{2}\right)\sin\left(\frac{x-y}{2}\right)\right|<\varepsilon\tag{7.1.220}$$

---

[6] Because for mathematicians this proof, although correct, relies on several facts about the logarithm, which (assuming it is defined as the inverse of the exponential) is usually shown to be a total function using the intermediate value theorem and hence assuming that the exponential is continuous... We also have to be careful with taking logs if $e^{x_0}-\varepsilon<0$...

Because $|\cos(\alpha)| \leqslant 1, \forall \alpha \in \mathbb{R}$ we can simplify:

$$\left|2\cos\left(\frac{x+y}{2}\right)\sin\left(\frac{x-y}{2}\right)\right| \leq 2\left|\sin\left(\frac{x-y}{2}\right)\right| \tag{7.1.221}$$

it suffices:

$$2\left|\sin\left(\frac{x-y}{2}\right)\right| < \varepsilon \tag{7.1.222}$$

when:

$$|x-y| < \delta \Longrightarrow \left|\frac{x-y}{2}\right| < \delta \tag{7.1.223}$$

Since $|\sin(x)| \leq |x|$:

$$2\left|\sin\left(\frac{x-y}{2}\right)\right| \leq 2\left|\frac{x-y}{2}\right| < 2\delta \tag{7.1.224}$$

Choosing $\delta = \varepsilon/2 > 0$ will do the trick. Because δ doesn't depend on x, y, the continuity is uniform!

- Second we prove the epsilon-delta limit:
  If we use the analytical definition of the $\sin(x)$ function (power series):

$$\sin(x) = \sum_{n=0}^{+\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!} \tag{7.1.225}$$

Therefore calculating:

$$\lim_{x\to 0} \sin(x) = 0 \tag{7.1.226}$$

is equivalent to:

$$\lim_{x\to 0} \sum_{n=0}^{+\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!} \tag{7.1.227}$$

And we have already proven earlier before that:

$$\lim_{x\to a} x^n = a^n \tag{7.1.228}$$

Therefore:

$$\lim_{x\to 0} \sum_{n=0}^{+\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!} = 0 \tag{7.1.229}$$

Hence:

$$\lim_{x\to 0} \sin(x) = 0 \tag{7.1.230}$$

10. We have:

$$\lim_{x\to 0}\cos(x)=1 \tag{7.1.231}$$

The method will be exactly similar to that of $\sin(x)$ seen just earlier!

- First we prove the epsilon-delta continuity:
  Let $\varepsilon>0$ and $x,y\in\mathbb{R}$. We want:

  $$|f(x)-f(y)|<\varepsilon \Longrightarrow |\cos(x)-\cos(y)|<\varepsilon \tag{7.1.232}$$

  Using trigonometric identities this is equivalent to:

  $$\left|2\sin\left(\frac{x+y}{2}\right)\sin\left(\frac{x-y}{2}\right)\right|<\varepsilon \tag{7.1.233}$$

  Because $|\sin(\alpha)|\leqslant 1,\forall\alpha\in\mathbb{R}$ we can simplify:

  $$\left|2\sin\left(\frac{x+y}{2}\right)\sin\left(\frac{x-y}{2}\right)\right|\leq 2\left|\sin\left(\frac{x-y}{2}\right)\right| \tag{7.1.234}$$

  it suffices:

  $$2\left|\sin\left(\frac{x-y}{2}\right)\right|<\varepsilon \tag{7.1.235}$$

  when:

  $$|x-y|<\delta \Longrightarrow \left|\frac{x-y}{2}\right|<\delta \tag{7.1.236}$$

  Since $|\sin(x)|\leq|x|$:

  $$2\left|\sin\left(\frac{x-y}{2}\right)\right|\leq 2\left|\frac{x-y}{2}\right|<2\delta \tag{7.1.237}$$

  Choosing $\delta=\varepsilon/2>0$ will do the trick. Because δ doesn't depend on x,y, the continuity is uniform.

- Second we prove the epsilon-delta limit:
  If we use the analytical definition of the $\cos(x)$ function (power series):

  $$\cos(x)=\sum_{n=0}^{+\infty}\frac{(-1)^n}{(2n)!}x^{2n} \tag{7.1.238}$$

  Therefore calculating:

  $$\lim_{x\to 0}\cos(x)=1 \tag{7.1.239}$$

  is equivalent to:

  $$\lim_{x\to 0}\sum_{n=0}^{+\infty}\frac{(-1)^n}{(2n)!}x^{2n}=\lim_{x\to 0}\left(1+\sum_{n=1}^{+\infty}\frac{(-1)^n}{(2n)!}x^{2n}\right) \tag{7.1.240}$$

And we have already proven earlier before that:

$$\lim_{x\to a} x^n = a^n \tag{7.1.241}$$

Therefore:

$$\lim_{x\to 0}\left(1+\sum_{n=1}^{+\infty}\frac{(-1)^n}{(2n)!}x^{2n}\right)=1+0=1 \tag{7.1.242}$$

Hence:

$$\lim_{x\to 0}\cos(x)=1 \tag{7.1.243}$$

11. We have:

$$\lim_{x\to 0}\tan(x)=0 \tag{7.1.244}$$

- First we prove the epsilon-delta continuity:
  The proof is straightforward! As by definition:

  $$\tan(x)=\frac{\sin(x)}{\cos(x)} \tag{7.1.245}$$

  He have proved earlier that $\sin(x)$ but also $\cos(x)$ are epsilon-delta continuous.
  But we have also proved that if $f(x),g(x)\neq 0$ are epsilon-delta continuous then $f(x)/g(x)$ is epsilon-delta continuous.
  Therefore $\tan(x)$ is epsilon-delta continuous!
- Second we prove the epsilon-delta limit:
  Notice that by the geometric definition we have:

  $$\lim_{x\to 0}\tan(x)=\lim_{x\to 0}\frac{\sin(x)}{\cos(x)}=\lim_{x\to 0}\sin(x)\lim_{x\to 0}\frac{1}{\cos(x)} \tag{7.1.246}$$

  And we have proved earlier above:

  $$\lim_{x\to 0}\sin(x)=0 \quad \text{and} \quad \lim_{x\to 0}\cos(x)=1 \tag{7.1.247}$$

  Therefore:

  $$\lim_{x\to 0}\tan(x)=\lim_{x\to 0}\sin(x)\lim_{x\to 0}\frac{1}{\cos(x)}=0 \tag{7.1.248}$$

12. We have:

$$\lim_{x\to 0}\frac{\sin(x)}{x}=1 \tag{7.1.249}$$

There are multiple ways more or less rigorous to prove that limit (geometric based, derivation based, series based, numerical based, squeeze theorem based, l'Hôpital theorem based, etc.) and that depend also on the definition of the function $\sin(x)$ itself...

- First we prove the epsilon-delta continuity:

  We have proven earlier that $\sin(x)$ is epsilon-delta continuous and also that x is epsilon-delta continuous. And we have proven earlier in the general case that if $f(x), g(x)$ are epsilon-delta continuous functions, assuming $g(x) \neq 0$, then $f(x)/g(x)$ is also epsilon-delta continuous!

  So that finish our proof!

- Second we prove the epsilon-delta limit:

  We have:

$$\lim_{x\to 0} \frac{\sin(x)}{x} = 1 \tag{7.1.250}$$

  if and only if for every $\varepsilon > 0$, there exists a $\delta > 0$ such that $|f(x) - 1| < \varepsilon$ whenever $0 < |x - 0| < \delta$.

  Let $\varepsilon > 0$ be an arbitrary real number.

  Note that:

$$\sin(x) = \sum_{n=0}^{+\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!} \tag{7.1.251}$$

  If $x \neq 0$, we have:

$$\frac{\sin(x)}{x} = \sum_{n=0}^{+\infty} (-1)^n \frac{x^{2n}}{(2n+1)!} = 1 + \sum_{n=1}^{+\infty} (-1)^n \frac{x^{2n}}{(2n+1)!} \tag{7.1.252}$$

  We thus have:

$$|f(x) - 1| = \left|\frac{\sin(x)}{x} - 1\right| = \left|\sum_{n=1}^{+\infty} (-1)^n \frac{x^{2n}}{(2n+1)!}\right| \leq \left|\sum_{n=1}^{+\infty} \frac{x^{2n}}{(2n+1)!}\right| \leq \sum_{n=1}^{+\infty} \left|\frac{x^{2n}}{(2n+1)!}\right| \tag{7.1.253}$$

  Therefore we have:

$$|f(x) - 1| \leq \sum_{n=1}^{+\infty} \left|\frac{x^{2n}}{(2n+1)!}\right| \leq \sum_{n=1}^{+\infty} \left|x^{2n}\right| = \sum_{n=1}^{+\infty} \left|x^2\right|^n \tag{7.1.254}$$

  If $0 < |x| < 1$, then $0 < |x^2| < 1$, and the infinite series $\sum_{n=1}^{+\infty} |x^2|^n$ converges to $\frac{x^2}{1-x^2}$. Choose $\delta := \sqrt{\frac{\varepsilon}{1+\varepsilon}}$. Then $0 < |x - 0| < \delta$ implies that $0 < |x| < \sqrt{\frac{\varepsilon}{1+\varepsilon}} < 1$, and hence $x^2 < \varepsilon - \varepsilon x^2$. But $x^2 < \varepsilon - \varepsilon x^2$ implies that $\frac{x^2}{1-x^2} < \varepsilon$.

  We therefore have $\sum_{n=1}^{+\infty} |x^2|^n < \varepsilon$ whenever $0 < |x - 0| < \delta$. But since $|f(x) - 1| \leq \sum_{n=1}^{+\infty} |x^2|^n$, we have $|f(x) - 1| < \varepsilon$ whenever $0 < |x - 0| < \delta$.

  Since ε was arbitrary, we have:

$$\lim_{x\to 0} \frac{\sin(x)}{x} = 1 \tag{7.1.255}$$

13. We have:

$$\lim_{x\to 0} \frac{\log(1+x)}{x} = 1 \tag{7.1.256}$$

- First we prove the epsilon-delta continuity:

  Let us put $z=1+x$ and we want to prove the epsilon-delta continuity near $x_0=0$ hence near $\log_b(1)$ for $b>1$.

  Let $\varepsilon>0$ and $z_0>0$. We want to find a small enough δ such that for $|z-z_0|<\delta$, $|\log_b(z)-\log_b(z_0)|<\varepsilon$. We have:

  $$|\log_b(z)-\log_b(z_0)|=\left|\log_b\left(\frac{z}{z_0}\right)\right| \tag{7.1.257}$$

  What we want is for z/z_0 to be close to 1. A better condition would be:

  $$b^{-\varepsilon}<\frac{z}{z_0}<b^{\varepsilon} \tag{7.1.258}$$

  In that case, we would have, by the strictly increasing nature of the $\log$ function:

  $$-\varepsilon<\log_b\left(\frac{z}{z_0}\right)<\varepsilon \tag{7.1.259}$$

  Therefore:

  $$\left|\log_b\left(\frac{z}{z_0}\right)\right|<|\varepsilon| \tag{7.1.260}$$

  So we are able to find such a ε but we see that the later is dependent of z_0 and we can't get rid of the latter. Hence $\log_b(z)$ is epsilon-delta continuous on z_0 but not uniformly continuous.

  Notice that in the literature it's very common to rearrange:

  $$b^{-\varepsilon}<\frac{z}{z_0}<b^{\varepsilon} \tag{7.1.261}$$

  as following:

  $$z_0b^{-\varepsilon}<z<z_0b^{\varepsilon} \tag{7.1.262}$$

  and to subtract z_0 everywhere such that:

  $$z_0b^{-\varepsilon}-z_0<z-z_0<z_0b^{\varepsilon}-z_0 \tag{7.1.263}$$

  As we wish to have $|z-z_0|<\delta$, he have finally:

  $$\delta=z_0\left\{1-b^{-\varepsilon},b^{-\varepsilon}-1\right\} \tag{7.1.264}$$

  When the base is the natural exponential the latter is written:

  $$\delta=z_0\left\{1-e^{-\varepsilon},e^{-\varepsilon}-1\right\} \tag{7.1.265}$$

- Second we prove the epsilon-delta limit:

  We simply use l'Hôpital's rule (see section Differential and Integral Calculus page 1461)! That is:

  $$\lim_{x\to 0}\frac{\log(1+x)}{x}=\lim_{x\to 0}\frac{\dfrac{\mathrm{d}}{\mathrm{d}x}\log(1+x)}{\dfrac{\mathrm{d}}{\mathrm{d}x}x}=\lim_{x\to 0}\frac{1}{1+x}=1 \tag{7.1.266}$$

14. We have:

$$\lim_{x\to 0}\frac{\sin^{-1}(x)}{x}=\frac{\arcsin(x)}{x}=1 \tag{7.1.267}$$

- First we prove the epsilon-delta continuity:

  Don't try to use the derivative or integral definition of $\arcsin(x)$ as they assume you already know that the function is epsilon-delta continuous! We want here to avoid using the "inverse theorem" as the latter is quite abstract and the proof not yet provided in this textbook (and very likely will never be!).

  The trick will be to used another expression for the $\arcsin(x)$ that uses the continuity proven earlier for $\sin(x)$. Let us remind for that purpose that we have proved in the section Numbers at page 252 the following de Moivre and Euler formula:

  $$\sin(x)=\frac{e^{\mathrm{i}x}-e^{-\mathrm{i}x}}{2\mathrm{i}} \tag{7.1.268}$$

  Let us put that as equal to y such that:

  $$y=\sin(x)=\frac{e^{\mathrm{i}x}-e^{-\mathrm{i}x}}{2\mathrm{i}} \tag{7.1.269}$$

  Hence:

  $$x=\arcsin(y) \tag{7.1.270}$$

  So we just need to solve for x. It will be a little bit easier if we do the substitution $e^{\mathrm{i}x}=z$. Therefore we get:

  $$2\mathrm{i}y=z-\frac{1}{z} \tag{7.1.271}$$

  And after rearranging:

  $$z^2-2\mathrm{i}yz-1=0 \tag{7.1.272}$$

  We apply the quadratic formula and get:

  $$z=\mathrm{i}y\pm\sqrt{1-y^2} \tag{7.1.273}$$

  One solution equals z the other solution equals $1/z$ pick one. And lets reverse the substitution:

  $$e^{\mathrm{i}x}=\mathrm{i}y+\sqrt{1-y^2} \tag{7.1.274}$$

  Therefore:

  $$x=-\mathrm{i}\ln\left(\mathrm{i}y+\sqrt{1-y^2}\right) \tag{7.1.275}$$

  So we have finally the interesting substitution:

  $$\arcsin(y)=-\mathrm{i}\ln\left(\mathrm{i}y+\sqrt{1-y^2}\right) \tag{7.1.276}$$

And we can rewrite our limit as:

$$\lim_{x\to 0}=\frac{\arcsin(x)}{x}=\frac{-\mathrm{i}\ln\left(\mathrm{i}x+\sqrt{1-x^2}\right)}{x}=1 \tag{7.1.277}$$

So we have to prove that:

$$\frac{\arcsin(x)}{x}=\frac{-\mathrm{i}\ln\left(\mathrm{i}x+\sqrt{1-x^2}\right)}{x} \tag{7.1.278}$$

is epsilon-delta continuous for $0<x<1$! And now that will be straightforward as we have proven that the sum of two epsilon-delta continuous function is epsilon-delta continuous (the content of the logarithm above). But also that the square root is epsilon-delta continuous (still the content of the logarithm above) and just previously that the logarithm is also epsilon-delta continuous! Hence the whole numerator above is epsilon-delta continuous for $0<x<1$!

And as the denominator (x) is also epsilon-delta continuous and that we have proven if $f(x),g(x)\neq 0$ are epsilon-delta continuous then $f(x)/g(x)$ is also epsilon-delta continuous. Then:

$$\frac{\arcsin(x)}{x}=\frac{-\mathrm{i}\ln\left(\mathrm{i}x+\sqrt{1-x^2}\right)}{x} \tag{7.1.279}$$

is indeed epsilon-delta continuous for $0<x<1$!

- Second we prove the epsilon-delta limit:

  We convert the inverse trigonometric sine function by the mathematical relationship between the trigonometric and inverse trigonometric functions. That is $y=\sin^{-1}(x)$, then $x=\sin(y)$.

  Therefore, the quotient of $\arcsin(x)$ by x can be written as the ratio of y to $\sin(y)$ mathematically:

$$\frac{\sin^{-1}(x)}{x}=\frac{y}{\sin(y)} \tag{7.1.280}$$

  Therefore, if x approaches zero, then y tends to $\sin^{-1}(0)$. More explicitly:

$$\lim_{x\to 0}\frac{\arcsin(x)}{x}=\lim_{y\to 0}\frac{y}{\sin(y)}=\lim_{y\to 0}\frac{1}{\frac{\sin(y)}{y}}=\frac{1}{1}=1 \tag{7.1.281}$$

15. We have:

$$\lim_{x\to 0}\frac{\tan(x)}{x}=1 \tag{7.1.282}$$

- First we prove the epsilon-delta continuity:

  We have already proved earlier that $\tan(x)$ is epsilon-delta continuous and same for x.

  We have also proved earlier that if $f(x),g(x)\neq 0$ are epsilon-delta continuous then $f(x)/g(x)$ is also epsilon-delta continuous.

  Therefore:

$$\frac{\tan(x)}{x} \tag{7.1.283}$$

  is epsilon-delta continuous!

- Second we prove the epsilon-delta limit:

  Once we have proved:

$$\lim_{x\to 0}\frac{\sin(x)}{x}=1 \quad \text{and} \quad \lim_{x\to 0}\cos(x)=1 \tag{7.1.284}$$

  We can prove quite easily that:

$$\lim_{x\to 0}\frac{\tan(x)}{x}=1 \tag{7.1.285}$$

  Indeed:

$$\lim_{x\to 0}\frac{\tan(x)}{x}=\lim_{x\to 0}\frac{\sin(x)}{x\cdot\cos(x)}=\lim_{x\to 0}\frac{\sin(x)}{x}\lim_{x\to 0}\frac{1}{\cos(x)}=1\cdot 1=1 \tag{7.1.286}$$

16. We have:

$$\lim_{x\to 0}\frac{\tan^{-1}(x)}{x}=\frac{\arctan(x)}{x}=1 \tag{7.1.287}$$

- First we prove the epsilon-delta continuity:

  Don't try to use the derivative or integral definition of $\arctan(x)$ as they assume you already know that the function is epsilon-delta continuous! We want here to avoid using the "inverse theorem" as the latter is quite abstract and the proof not yet provided in this textbook (and very likely will never be!).

  The trick will be to used another expression for the $\arctan(x)$ that uses the continuity proven earlier for $\sin(x)$ and $\cos(x)$. Let us remind for that purpose that we have proved in the section Numbers at page 252 the following de Moivre and Euler formula:

$$\sin(\varphi)=\frac{e^{\mathrm{i}\varphi}-e^{-\mathrm{i}\varphi}}{2\mathrm{i}},\quad \cos(\varphi)=\frac{e^{\mathrm{i}\varphi}+e^{-\mathrm{i}\varphi}}{2} \tag{7.1.288}$$

  which leads to:

$$\tan(\varphi)=\frac{\sin(\varphi)}{\cos(\varphi)}=\frac{e^{\mathrm{i}\varphi}-e^{-\mathrm{i}\varphi}}{\mathrm{i}\left(e^{\mathrm{i}\varphi}+e^{-\mathrm{i}\varphi}\right)} \tag{7.1.289}$$

  Let $\tan(\varphi)=x$, then:

$$\frac{e^{\mathrm{i}\varphi}-e^{-\mathrm{i}\varphi}}{e^{\mathrm{i}\varphi}+e^{-\mathrm{i}\varphi}}=\mathrm{i}x \tag{7.1.290}$$

  and solving it we find that:

$$e^{\mathrm{i}\varphi}=\sqrt{\frac{1+\mathrm{i}x}{1-\mathrm{i}x}} \tag{7.1.291}$$

  which leads us to:

$$\arctan(x)=\varphi=\frac{1}{\mathrm{i}}\ln\sqrt{\frac{1+\mathrm{i}x}{1-\mathrm{i}x}}=\frac{\mathrm{i}}{2}[\ln(1-\mathrm{i}x)-\ln(1+\mathrm{i}x)] \tag{7.1.292}$$

  when $\forall x\in\mathbb{C}\backslash(\mathrm{i}(]-\infty,-1]\cup[1,+\infty[))$.

Therefore we can rewrite the limit above as:

$$\lim_{x\to 0}\frac{\tan^{-1}(x)}{x}=\frac{\arctan(x)}{x}=\frac{\frac{\mathrm{i}}{2}[\ln(1-\mathrm{i}x)-\ln(1+\mathrm{i}x)]}{x}=1 \tag{7.1.293}$$

As we have proven that a constant is epsilon-delta continuous and x (and therefore $\mathrm{i}x$) is also epsilon-delta continuous then $1-\mathrm{i}x$ and $1+\mathrm{i}x$ are both epsilon-delta continuous.

But we have also proved earlier above that the logarithm is epsilon-delta continuous. Therefore as we have proven that if $f(x),g(x)$ are epsilon-delta continuous then $g(f(x))$ is also epsilon-delta continuous that lead us to the result that $\ln(1-\mathrm{i}x)$ and $\ln(1+\mathrm{i}x)$ are both epsilon-delta continuous!

As we have also proved that if $f(x),g(x)\neq 0$ are epsilon-delta continuous then $f(x)/g(x)$ is also epsilon-delta continuous then the ratio of of $\ln(1-\mathrm{i}x)-\ln(1+\mathrm{i}x)$ by x is also epsilon-delta continuous.

Finally we have proven that:

$$\frac{\tan^{-1}(x)}{x}=\frac{\arctan(x)}{x}=\frac{\frac{\mathrm{i}}{2}[\ln(1-\mathrm{i}x)-\ln(1+\mathrm{i}x)]}{x} \tag{7.1.294}$$

is indeed epsilon-delta continuous $\forall x\in\mathbb{C}\backslash(\mathrm{i}(]-\infty,-1]\cup[1,+\infty[))$.

- Second we prove the epsilon-delta limit:

  This is straightforward. Again we just use l'Hospital rule (see section Differential and Integral Calculus page 1461) and this leads to:

$$\lim_{x\to 0}\frac{\arctan(x)}{x}=\lim_{x\to 0}\frac{\frac{\mathrm{d}}{\mathrm{d}t}\arctan(x)}{\frac{\mathrm{d}}{\mathrm{d}t}x}=\lim_{x\to 0}\frac{\frac{1}{1+x^2}}{1}=\lim_{x\to 0}\frac{1}{1+x^2}=1 \tag{7.1.295}$$

17. We have:

$$\lim_{x\to 0}\frac{a^x-1}{x}=\ln(a) \tag{7.1.296}$$

- First we prove the epsilon-delta continuity:

  This is straightforward to prove! We have proven earlier that a constant (like 1) is epsilon-delta continuous.

  We have also proven that x is epsilon-delta continuous.

  We have proven that if $f(x),g(x)$ are epsilon-delta continuous then $f(x)+g(x)$ and $f(x)-g(x)$ are also epsilon-delta continuous.

  We have proven earlier that if $f(x)$ is epsilon-delta continuous then $f(x)^n$ is epsilon-delta continuous. Therefore a^x-1 is epsilon-delta continuous.

  We have proven that if $f(x),g(x)\neq 0$ are epsilon-delta continuous then $f(x)/g(x)$ is also epsilon-delta continuous.

  Therefore:

$$\frac{a^x-1}{x}=\ln(a) \tag{7.1.297}$$

is epsilon-delta continuous!

- Second we prove the epsilon-delta limit:

  $$\lim_{x\to 0}\frac{a^x-1}{x} \tag{7.1.298}$$

  Let $y=a^x-1$, then $1+y=a^x$.
  Consider the relation:

  $$1+y=a^x \tag{7.1.299}$$

  Using the logarithm on both sides, we have:

  $$\begin{aligned}&\ln(1+y)=\ln(a^x)\\ \Rightarrow&\ln(1+y)=x\ln(a)\\ \Rightarrow&x=\frac{\ln(1+y)}{\ln(a)}\end{aligned} \tag{7.1.300}$$

  Also:

  $$\lim_{x\to 0} y=\lim_{x\to 0}(a^x-1)=a^0-1=1-1=0 \tag{7.1.301}$$

  This shows that $y\to 0$ as $x\to 0$. Therefore, the given limit can be written as:

  $$\begin{aligned}&\lim_{x\to 0}\frac{a^x-1}{x}=\lim_{y\to 0}\frac{y}{\frac{\ln(1+y)}{\ln(a)}}\\ \Rightarrow&\lim_{x\to 0}\frac{a^x-1}{x}=\lim_{y\to 0}\frac{\ln(a)}{\frac{1}{y}\ln(1+y)}\\ \Rightarrow&\lim_{x\to 0}\frac{a^x-1}{x}=\lim_{y\to 0}\frac{\ln a}{\ln(1+y)^{\frac{1}{y}}}\\ \Rightarrow&\lim_{x\to 0}\frac{a^x-1}{x}=\frac{\ln(a)}{\lim_{y\to 0}\ln(1+y)^{\frac{1}{y}}}\\ \Rightarrow&\lim_{x\to 0}\frac{a^x-1}{x}=\frac{\ln(a)}{\ln\left[\lim_{y\to 0}(1+y)^{\frac{1}{y}}\right]}\end{aligned} \tag{7.1.302}$$

  Using the relation $\lim_{x\to 0}(1+x)^{\frac{1}{x}}=e$, we have:

  $$\Rightarrow\lim_{x\to 0}\frac{a^x-1}{x}=\frac{\ln(a)}{\ln(e)} \tag{7.1.303}$$

  But $\ln(e)=1$, we have:

  $$\Rightarrow\lim_{x\to 0}\frac{a^x-1}{x}=\frac{\ln(a)}{1}=\ln(a) \tag{7.1.304}$$

18. We have:

$$\lim_{x\to 0}\frac{(1+x)^m-1}{x}=m \tag{7.1.305}$$

- First we prove the epsilon-delta continuity:

  This is straightforward to prove! We have proven earlier that a constant (like 1) is epsilon-delta continuous.

  We have also proven that x is epsilon-delta continuous.

  We have proven that if $f(x), g(x)$ are epsilon-delta continuous then $f(x)+g(x)$ and $f(x)-g(x)$ are also epsilon-delta continuous. Therefore $1+x$ is epsilon-delta continuous.

  We have proven earlier that if $f(x)$ is epsilon-delta continuous then $f(x)^n$ is epsilon-delta continuous. Therefore $(1+x)^m-1$ is epsilon-delta continuous.

  We have proven that if $f(x), g(x) \neq 0$ are epsilon-delta continuous then $f(x)/g(x)$ is also epsilon-delta continuous. Therefore:

  $$\frac{(1+x)^m-1}{x} \tag{7.1.306}$$

  is epsilon-delta continuous!

- Second we prove the epsilon-delta limit:

  This is straightforward. Again we just use l'Hospital rule (see section Differential and Integral Calculus page 1461) and this leads to:

  $$\lim_{x\rightarrow 0}\frac{(1+x)^m-1}{x}=\lim_{x\rightarrow 0}\frac{\dfrac{\mathrm{d}}{\mathrm{d}t}(1+x)^m-1}{\dfrac{\mathrm{d}}{\mathrm{d}t}x}=\lim_{x\rightarrow 0}\frac{m(x+1)}{1}=\lim_{x\rightarrow 0}m(x+1)=m \tag{7.1.307}$$

19. We have:

$$\lim_{x\rightarrow 0}\frac{1-\cos(x)}{x}=0 \tag{7.1.308}$$

- First we prove the epsilon-delta continuity:

  The proof is straightforward! We have already proved earlier that a constant is epsilon-delta continuous therefore 1 has the same behaviour.

  We have already proved also earlier that $\cos(x)$ is epsilon-delta continuous !

  And we have proved that x is epsilon-delta continuous.

  We have proved that if $f(x), g(x)$ are epsilon-delta continuous then $f(x)-g(x)$ is epsilon-delta continuous. Therefore $1-\cos(x)$ is epsilon-delta continuous.

  We have proved that if $f(x), g(x) \neq 0$ are epsilon-delta continuous then $f(x)/g(x)$ is epsilon-delta continuous. Therefore:

  $$\frac{1-\cos(x)}{x} \tag{7.1.309}$$

  is epsilon-delta continuous!

- Second we prove the epsilon-delta limit:

$$\begin{aligned}\lim_{x\to 0}\frac{1-\cos(x)}{x}&=\lim_{x\to 0}\frac{(1-\cos(x))(1+\cos(x))}{x(1+\cos(x))}=\lim_{x\to 0}\frac{1-\cos(x)+\cos(x)-\cos^2(x)}{x(1+\cos(x))}\\&=\lim_{x\to 0}\frac{1-\cos^2(x)}{x(1+\cos(x))}=\lim_{x\to 0}\frac{\sin^2(x)}{x(1+\cos(x))}=\lim_{x\to 0}\frac{\sin(x)}{x}\cdot\frac{\sin(x)}{1+\cos(x)}\\&=\left(\lim_{x\to 0}\frac{\sin(x)}{x}\right)\left(\lim_{x\to 0}\frac{\sin(x)}{1+\cos(x)}\right)\end{aligned}\tag{7.1.310}$$

  We have proven earlier above that:

$$\lim_{x\to 0}\frac{\sin(x)}{x}=1\tag{7.1.311}$$

  Hence:

$$\begin{aligned}\lim_{x\to 0}\frac{1-\cos(x)}{x}&=\lim_{x\to 0}\frac{(1-\cos(x))(1+\cos(x))}{x(1+\cos(x))}=\left(\lim_{x\to 0}\frac{\sin(x)}{x}\right)\left(\lim_{x\to 0}\frac{\sin(x)}{1+\cos(x)}\right)\\&=1\cdot\lim_{x\to 0}\frac{\sin(x)}{1+\cos(x)}\\&=\frac{\sin(0)}{1+\cos(0)}=\frac{0}{1+1}=0\end{aligned}\tag{7.1.312}$$

20. We have:

$$\lim_{x\to+\infty}x^{1/x}=1\tag{7.1.313}$$

often denoted:

$$\lim_{n\to+\infty}n^{1/n}=1\tag{7.1.314}$$

- First we prove the epsilon-delta continuity:
  The proof is straightforward as:

$$f(x)=x^{\frac{1}{x}}=e^{\frac{1}{x}\ln(x)}\tag{7.1.315}$$

  Indeed, we have proven in the previous examples that for $x>0$, we had $1/x$ and $\ln(x)$ were epsilon-delta continuous.
  And as we have proven in the general case that if $f(x),g(x)$ are epsilon-delta continuous then $f(x)\cdot g(x)$ was also epsilon-delta continuous, then $\frac{1}{x}\ln(x)$ is epsilon-delta continuous!
  We have also proven in the example above that the exponential e is epsilon-delta continuous. And because we have proven in the general case that if $f(x),g(x)$ are epsilon-delta continuous then $g(f(x))$ is also epsilon-delta continuous, then:

$$e^{\frac{1}{x}\ln(x)}\tag{7.1.316}$$

  is epsilon-delta continuous and so is $x^{1/x}$!

- Second we prove the epsilon-delta limit:
  Set $x_n = n^{1/n}$. We have:

$$\ln\left(x_n\right) = \frac{\ln(n)}{n} \qquad (7.1.317)$$

  Clearly, we have (from above):

$$\lim_{n\to+\infty} \ln\left(x_n\right) = 0 \qquad (7.1.318)$$

  Hence:

$$\lim_{n\to+\infty} e^{\ln(x_n)} = e^0 = 1 \qquad (7.1.319)$$

  which translates into:

$$\lim_{n\to+\infty} n^{1/n} = 1 \qquad (7.1.320)$$

21. We have:

$$\lim_{x\to0^+} x\ln(x) = 0 \qquad (7.1.321)$$

- First we prove the epsilon-delta continuity:
  The proof is straightforward as we have proven that $\log(x)$ is epsilon-delta continuous (therefore $\ln(x)$ has the same behaviour) and that x is also epsilon-delta continuous.
  And we have proved earlier that if $f(x), g(x)$ are epsilon-delta continuous then $f(x)\cdot g(x)$ is also epsilon-delta continuous.
  Hence:

$$x\ln(x) \qquad (7.1.322)$$

  is epsilon-delta continuous!
- Second we prove the epsilon-delta limit:
  If we evaluate the limit of the numerator and denominator separately we'll find that as $\ln(x)$ goes to $+\infty$ as x goes to $+\infty : \ln(+\infty) = +\infty$ and x goes to $+\infty$.
  Therefore we have a ratio of two infinities $\frac{+\infty}{+\infty}$ meaning that we will have to apply l'Hôpital's rule (see section Differential and Integral Calculus page 1461):

$$\lim_{x\to+\infty} \frac{\ln(x)}{x} = \lim_{x\to+\infty} \frac{\frac{\mathrm{d}}{\mathrm{d}x}(\ln(x))}{\frac{\mathrm{d}}{\mathrm{d}x}(x)} = \lim_{x\to+\infty} \frac{\frac{1}{x}}{1} = \lim_{x\to+\infty} \frac{1}{x} = 0 \qquad (7.1.323)$$

  The limit approaches 0 because 1 divided over something approaching $+\infty$ becomes closer and closer to 0.

22. We have:

$$\lim_{x\to+\infty} \frac{\ln(x)}{x} = 0 \qquad (7.1.324)$$

- First we prove the epsilon-delta continuity:

  We have already proved earlier that $\log(x)$ is epsilon-delta continuous.

  We have also already proved earlier that x is epsilon-delta continuous.

  We have proved earlier that if we have $f(x), g(x) \neq 0$ that are epsilon-delta continuous then $f(x)/g(x)$ is epsilon-delta continuous.

  Therefore:

  $$\frac{\ln(x)}{x} \tag{7.1.325}$$

  is epsilon-delta continuous!

- Second we prove the epsilon-delta limit:

  The proof is straightforward using l'Hôpital's rule (see section Differential and Integral Calculus page 1461)! Indeed:

  $$\lim_{x\to+\infty}\frac{\ln(x)}{x}=\lim_{x\to+\infty}\frac{\dfrac{\mathrm{d}}{\mathrm{dx}}\ln(x)}{\dfrac{\mathrm{d}}{\mathrm{dx}}x}=\lim_{x\to+\infty}\frac{1}{x}=0 \tag{7.1.326}$$

23. We have:

$$\lim_{x\to 0} x^x=1 \tag{7.1.327}$$

- First we prove the epsilon-delta continuity:

  The proof follows exactly the same pattern as the case with $x^{1/x}$ done earlier above and leads to the same conclusion: x^x is epsilon-delta continuous for $x>0$.

- Second we prove the epsilon-delta limit:

  The proof uses a result we have obtained in one of the previous example. That is:

  $$\lim_{x\to 0} x\ln(x)=0 \tag{7.1.328}$$

  How? That's easy! We just rewrite:

  $$\lim_{x\to 0} x^x=e^{x\ln(x)} \tag{7.1.329}$$

  And therefore the result is immediate:

  $$\lim_{x\to 0} x^x=e^{x\ln(x)}=e^0=1 \tag{7.1.330}$$

24. We have:

$$\lim_{x\to 0}(1+\sin(x))^{1/x}=e \tag{7.1.331}$$

- First we prove the epsilon-delta continuity:

  That will be also straightforward thank to all previous examples!

  First we have proven that a constant like 1 is epsilon-delta continuous.

  We have also proven that $\sin(x)$ is epsilon-delta continuous.

We have also proven that in the general case if $f(x), g(x)$ are epsilon-delta continuous then $f(x)+g(x)$ is epsilon-delta continuous. Therefore $1+\sin(x)$ is epsilon-delta continuous.

We have proven earlier that if $x>0$ then $x^{1/x}$ was epsilon-delta continuous. Therefore we conclude (in the engineering way...) that:

$$(1+\sin(x))^{1/x} \tag{7.1.332}$$

is epsilon-delta continuous!

- Second we prove the epsilon-delta limit:

  Here we use exactly the same trick as before...! That is we rewrite the limit as:

  $$\lim_{x\to 0} e^{\frac{1+\sin(x)}{x}} \tag{7.1.333}$$

  Therefore we just have to compute:

  $$\lim_{x\to 0} \frac{1+\sin(x)}{x} \tag{7.1.334}$$

  We use l'Hôpital's rule (see section Differential and Integral Calculus page 1461)! Therefore we get:

  $$\lim_{x\to 0} \frac{1+\sin(x)}{x} = \frac{\dfrac{\mathrm{d}}{\mathrm{d}x}(1+\sin(x))}{\dfrac{\mathrm{d}}{\mathrm{d}x}x} = \lim_{x\to 0} \cos(x) = 1 \tag{7.1.335}$$

  Therefore:

  $$\lim_{x\to 0}(1+\sin(x))^{1/x} = \lim_{x\to 0} e^{\frac{1+\sin(x)}{x}} = e^1 = e \tag{7.1.336}$$

25. We have for any real a:

$$\lim_{x\to +\infty} x-\sqrt{x^2-a^2} = 0 \tag{7.1.337}$$

- First we prove the epsilon-delta continuity:

  The proof is straightforward using again the results of the previous examples and the general properties of epsilon-delta continuity!

  Indeed, we have already proven that a constant (like a or like a^2 that is still a constant too) is epsilon-delta continuous.

  We have also proven earlier above that x^2 is epsilon-delta continuous.

  And as we have proven the property that if $f(x), g(x)$ are epsilon-delta continuous then $f(x)-g(x)$ is epsilon-delta continuous. That means x^2-a^2 is epsilon-delta continuous.

  We have proven earlier also that if $f(x)>0$ is epsilon-delta continuous then $\sqrt{f(x)}$ is also epsilon-delta continuous. Therefore for $x^2-a^2>0$ (i.e. $x^2>a^2$, or $|x|>|a|$), we have that $\sqrt{x^2-a^2}$ is epsilon-delta continuous too!

- Second we prove the epsilon-delta limit:
  We use a trick:

  $$\lim_{x\rightarrow+\infty} x-\sqrt{x^2-a^2}=\lim_{x\rightarrow+\infty}\left(x-\sqrt{x^2-a^2}\right)\frac{x+\sqrt{x^2-a^2}}{x+\sqrt{x^2-a^2}} \tag{7.1.338}$$

  Hence after simplification:

  $$\lim_{x\rightarrow+\infty}\frac{a^2}{x+\sqrt{x^2-a^2}}=a^2\lim_{x\rightarrow+\infty}\frac{1}{x+\sqrt{x^2-a^2}} \tag{7.1.339}$$

  It's obvious now as $x>a$ and therefore that $\sqrt{x^2-a^2}\geq 0$ that:

  $$\lim_{x\rightarrow+\infty} x+\sqrt{x^2-a^2}=+\infty \tag{7.1.340}$$

  Hence:

  $$\lim_{x\rightarrow+\infty} x-\sqrt{x^2-a^2}=a^2\lim_{x\rightarrow+\infty}\frac{1}{x+\sqrt{x^2-a^2}}=\frac{1}{+\infty}=0 \tag{7.1.341}$$

26. We have:

  $$\lim_{x\rightarrow 0}\frac{\tan(x)-\sin(x)}{x^3}=\frac{1}{2} \tag{7.1.342}$$

  - First we prove the epsilon-delta continuity:
    This is quite straightforward thanks to everything we have proven so far! Indeed, we have proven earlier above that $\tan(x)$ and $\sin(x)$ are epsilon-delta continuous. But we have also proven earlier that if $f(x),g(x)$ are epsilon-delta continuous then $f(x)-g(x)$ is epsilon-delta continuous. Therefore $\tan(x)-\sin(x)$ is epsilon-delta continuous.
    And we have also proven earlier that if $f(x)$ is epsilon-delta continuous then $f(x)^n$ is also epsilon-delta continuous and therefore x^3 at the denominator is also epsilon-delta continuous.
    And finally as we have proven earlier that if $f(x),g(x)\neq 0$ are epsilon-delta continuous then $f(x)/g(x)$ is also epsilon-delta continuous and therefore:

    $$\frac{\tan(x)-\sin(x)}{x^3} \tag{7.1.343}$$

    is epsilon-delta continuous!
  - Second we prove the epsilon-delta limit:
    Let us do that without l'Hôpital's rule (personal choice!). So we use the definitions of trigonometric functions:

    $$\begin{aligned}\frac{\tan(x)-\sin(x)}{x^3}&=\frac{\dfrac{\sin(x)}{\cos(x)}-\sin(x)}{x^3}=\frac{\left(\dfrac{\sin(x)-\sin(x)\cos(x)}{\cos(x)}\right)}{x^3}\\&=\frac{\sin(x)-\sin(x)\cos(x)}{x^3\cos(x)}=\frac{\sin(x)(1-\cos(x))}{x^3\cos(x)}=\frac{\sin(x)}{x}\cdot\frac{1-\cos(x)}{x^2\cos(x)}\\&=\frac{\sin(x)}{x}\frac{1-\cos(x)}{x^2\cos(x)}\frac{(1+\cos(x))}{(1+\cos(x))}=\frac{\sin(x)}{x}\frac{1-\cos^2(x)}{x^2\cos(x)(1+\cos(x))}\\&=\frac{\sin(x)}{x}\frac{\sin^2(x)}{x^2\cos(x)(1+\cos(x))}=\frac{\sin(x)}{x}\left(\frac{\sin(x)}{x}\right)^2\frac{1}{\cos(x)(1+\cos(x))}\end{aligned} \tag{7.1.344}$$

Using the property that:

$$\lim_{x\rightarrow 0} f(x)g(x)=\lim_{x\rightarrow 0} f(x)\cdot\lim_{x\rightarrow 0} g(x) \tag{7.1.345}$$

We get:

$$\begin{aligned}\lim_{x\rightarrow 0}\frac{\tan(x)-\sin(x)}{x^3}&=\lim_{x\rightarrow 0}\frac{\sin(x)}{x}\cdot\left(\lim_{x\rightarrow 0}\frac{\sin(x)}{x}\right)^2\cdot\lim_{x\rightarrow 0}\frac{1}{\cos(x)(1+\cos(x))}\\&=1\cdot 1^2\cdot\frac{1}{1\cdot(1+1)}=\frac{1}{2}\end{aligned} \tag{7.1.346}$$

27. For $q>0$:

$$\lim_{x\rightarrow+\infty}\frac{\log^p(x)}{x^q}=0 \tag{7.1.347}$$

- First we prove the epsilon-delta continuity:

  The proof will be straightforward again thanks to everything we have proven so far! Indeed, we have proven earlier that the logarithm is epsilon-delta continuous, therefore $\log(x)$ is epsilon-delta continuous.

  We have also proven that x is epsilon-delta continuous.

  We have also proven that if $f(x)$ is an epsilon-delta continuous function then $f^n(x)$ is also epsilon-delta continuous. Hence $\log^p(x)$ and x^q are epsilon-delta continuous.

  And finally as we have proven that if $f(x),g(x)\neq 0$ are epsilon-delta continuous then $f(x)/g(x)$ is epsilon-delta continuous that lead us to the conclusion that:

  $$\frac{\log^p(x)}{x^q} \tag{7.1.348}$$

  is epsilon-delta continuous!

- Second we prove the epsilon-delta limit:

  We simply use l'Hôpital's rule (see section Differential and Integral Calculus page 1461)! That is:

  $$\lim_{x\rightarrow+\infty}\frac{\log^p(x)}{x^q}=\frac{\dfrac{\mathrm{d}}{\mathrm{d}x}\log^p(x)}{\dfrac{\mathrm{d}}{\mathrm{d}x}x^q}=\lim_{x\rightarrow+\infty}\frac{1/x}{qx^{q-1}}=\lim_{x\rightarrow+\infty}\frac{1}{qx^q}=0 \tag{7.1.349}$$

28. We have:

$$\lim_{x\rightarrow+\infty} n(\sqrt[n]{x}-1)=\log(x) \tag{7.1.350}$$

This formula is due to the English scientist and mathematician Edmond Halley (11656 to 11742 according to holocene calendar) and was known also by Leonhard Euler, who used it for defining the natural logarithm.

- First we prove the epsilon-delta continuity:

  The proof is straightforward we everything we have seen so far! Indeed, in the first examples above we have proven that $\sqrt[n]{x}$ is epsilon-delta continuous and also that a constant (like 1) is also epsilon-delta continuous.

  But we have proven earlier that if $f(x),g(x)$ are the both epsilon-delta continuous then $f(x)-g(x)$ is also epsilon-delta continuous. Therefore $\sqrt[n]{x}-1$ is epsilon-delta continuous.

  And finally we have proven that if $f(x)$ is epsilon-delta continuous then that same function multiplied by a constant is also epsilon-delta continuous. Therefore:

  $$n(\sqrt[n]{x}-1) \tag{7.1.351}$$

  is epsilon-delta continuous!

- Second we prove the epsilon-delta limit:

  We change the n-th root to power of e and use the power series expansion of exponential function:

  $$\begin{aligned}n(\sqrt[n]{x}-1)&=n\left(e^{\frac{\ln(x)}{n}}-1\right)=n\left(\sum_{m=0}^{+\infty}\frac{(\ln(x)/n)^m)}{m!}-1\right)\\&=\sum_{m=1}^{+\infty}\frac{n(\ln(x)/n)^m}{m!}=\ln(x)+\frac{1}{n}\sum_{m=2}^{+\infty}\frac{(\ln(x))^m}{m!n^{m-2}}\end{aligned} \tag{7.1.352}$$

  And finally:

  $$\lim_{x\rightarrow+\infty}n(\sqrt[n]{x}-1)=\lim_{x\rightarrow+\infty}\ln(x)+\frac{1}{n}\sum_{m=2}^{+\infty}\frac{(\ln(x))^m}{m!n^{m-2}}=\ln(x) \tag{7.1.353}$$

29. We want to show if:

$$\lim_{n\rightarrow+\infty}\left(1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{n}\right)=+\infty \tag{7.1.354}$$

There are many ways to handle this sequence. Let us use calculus techniques again. Consider the function:

$$f(x)=\frac{1}{x},\qquad \text{for } x\geq 1 \tag{7.1.355}$$

We have:

$$\int\limits_{n}^{n+1}\frac{1}{x}\mathrm{d}x=[\ln(x)]_n^{n+1}=\ln(n+1)-\ln(n) \tag{7.1.356}$$

and:

$$\int\limits_{1}^{n+1}\frac{1}{x}\mathrm{d}x=[\ln(x)]_1^{n+1}=\ln(n+1) \tag{7.1.357}$$

for any $n \geq 1$. Note that for any $x \in [n, n+1]$, we have:

$$\frac{1}{n+1} \leq \frac{1}{x} \leq \frac{1}{n} \qquad (7.1.358)$$

hence:

$$\int_{n}^{n+1} \frac{1}{n+1}\mathrm{d}x \leq \int_{n}^{n+1} \frac{1}{x}\mathrm{d}x \leq \int_{n}^{n+1} \frac{1}{n}\mathrm{d}x \qquad (7.1.359)$$

which gives:

$$\frac{1}{n+1} \leq \int_{n}^{n+1} \frac{1}{x}\mathrm{d}x \leq \frac{1}{n} \qquad (7.1.360)$$

Since:

$$\int_{1}^{n+1} \frac{1}{x}\mathrm{d}x = \int_{1}^{2} \frac{1}{x}\mathrm{d}x + \int_{2}^{3} \frac{1}{x}\mathrm{d}x + \ldots + \int_{n-1}^{n} \frac{1}{x}\mathrm{d}x + \int_{n}^{n+1} \frac{1}{x}\mathrm{d}x \qquad (7.1.361)$$

we get:

$$\frac{1}{2} + \ldots + \frac{1}{n+1} \leq \int_{1}^{n+1} \frac{1}{x}\mathrm{d}x \leq 1 + \frac{1}{2} + \ldots + \frac{1}{n} \qquad (7.1.362)$$

In particular, we have:

$$\ln(n+1) = \int_{1}^{n+1} \frac{1}{x}\mathrm{d}x \leq 1 + \frac{1}{2} + \ldots + \frac{1}{n} \qquad (7.1.363)$$

Therefore, since $\lim_{n\rightarrow+\infty} \ln(n+1) = +\infty$, we must have:

$$\lim_{n\rightarrow+\infty} 1 + \frac{1}{2} + \ldots + \frac{1}{n} = +\infty \qquad (7.1.364)$$

We will stop here the list of common epsilon-delta limits and epsilon-delta continuities as it is almost without end! And the reader must keep in mind that we have focused only on univariate functions here above and that most of the examples were only for real valued functions (only two were in the complex planes)...

Most engineers don't like this topic but don't forget that it's the only rigorous way to show that a function can be differentiate or integrate (among other important application for professional mathematicians and theoretical physicists)!

To summarize here is a table with the most important limits:

Most common limits		
$\lim\limits_{z\to a} c^{te}=c^{te}$	$\lim\limits_{x\to a} x^n=a^n$	$\lim\limits_{x\to 0} x^2\sin\left(\dfrac{1}{x}\right)$
$\lim\limits_{z\to +\infty}\left(1+\dfrac{a}{x}\right)^z=e^a$	$\lim\limits_{x\to 0}\dfrac{e^x-1}{x}=1$	$\lim\limits_{x\to 0}\sin(x)=0$
$\lim\limits_{x\to 0}\tan(x)=0$	$\lim\limits_{z\to 0}(1+x)^{1/x}=e$	$\lim\limits_{x\to a}\dfrac{x-a}{x^n-a^n}=\dfrac{1}{na^{n-1}}$
$\lim\limits_{x\to 0}\dfrac{\sin(x)}{x}=1$	$\lim\limits_{x\to 0}\dfrac{\log(1+x)}{x}=1$	$\lim\limits_{x\to 0}\dfrac{\arcsin(x)}{x}=1$
$\lim\limits_{z\to 0}\dfrac{\arctan(x)}{x}=1$	$\lim\limits_{z\to 0}\dfrac{a^x-1}{x}=\ln(a)$	$\lim\limits_{x\to 0}\dfrac{1-\cos(x)}{x}=0$
$\lim\limits_{x\to 0^+} x\ln(x)=0$	$\lim\limits_{z\to +\infty}\dfrac{\ln(x)}{x}=0$	$\lim\limits_{x\to 0}(1+\sin(x))^{1/x}=e$
$\lim\limits_{x\to +\infty} x-\sqrt{x^2-a^2}=0$	$\lim\limits_{x\to 0}\dfrac{\tan(x)-\sin(x)}{x^3}=\dfrac{1}{2}$	$\lim\limits_{x\to +\infty}\dfrac{\log^p(x)}{x^q}=0$
$\lim\limits_{n\to +\infty}\left(1+\dfrac{1}{2}+\dfrac{1}{3}+\ldots+\dfrac{1}{n}\right)=+\infty$	$\lim\limits_{n\to +\infty}\sqrt[n]{a}=1$	$\lim\limits_{x\to 0}\cos(x)=1$
$\lim\limits_{x\to a}\dfrac{x^n-a^n}{x-a}=na^{n-1}$	$\lim\limits_{z\to 0}\dfrac{\tan(x)}{x}=1$	$\lim\limits_{z\to +\infty} x^{1/x}=1$
$\lim\limits_{x\to 0} x^x=1$	$\lim\limits_{z\to 0}\dfrac{\tan(x)}{x}=1$	$\lim\limits_{x\to +\infty} n(\sqrt[n]{x}-1)=\log(x)$

Table 7.5: **Most common limits**

Remark

In some of the proofs of usual limits seen earlier above the reader may have noticed that sometimes that arguments turn out to involve not just ε but a constant multiple of it. This may occur for instance when the limit involves a sum or several arithmetic processes.

But is $K\varepsilon$ just as good as ε? If we en upd with a proof with 2ε, or 22ε, or 0.001ε, that's just as good as coming out with ε! Indeed if ε is an arbitrary small number, so is 22ϵ and 0.001ε. Therefore, if we can prove something is less than $K\varepsilon$, you have shown that it can be made as small as desired.

We formulate this as a general principle, the "K-ε principle". This isn't a standard term in analysis, so don't use it when you go to your next mathematics congress, but it is useful to name an idea that will recur often.

7.1.2.1.3 Limit laws

We now take a look at the "limit laws", the individual properties limits in the univariate case. The proofs will be omitted as it is quite intuitive but any reader can request us the proof of one of them if needed!

Let $f(x)$ and $g(x)$ be defined for all $x\neq a$ over some open interval containing a. Assume that L and M are real numbers such that:

$$\lim_{x\to a} f(x)=L \qquad \text{and} \qquad \lim_{x\to a} g(x)=M \tag{7.1.365}$$

Let c^{te} be a constant. Then, each of the following statements holds:

- The sum law for limits gives:

$$\lim_{x\rightarrow a}(f(x)+g(x))=\lim_{x\rightarrow a}f(x)+\lim_{x\rightarrow a}g(x)=L+M \tag{7.1.366}$$

- The difference law for limits gives:

$$\lim_{x\rightarrow a}(f(x)-g(x))=\lim_{x\rightarrow a}f(x)-\lim_{x\rightarrow a}g(x)=L-M \tag{7.1.367}$$

- Constant multiple law for limits:

$$\lim_{x\rightarrow a}c^{te}f(x)=c^{te}\lim_{x\rightarrow a}f(x)=c^{te}L \tag{7.1.368}$$

- Product law for limits:

$$\lim_{x\rightarrow a}f(x)g(x)=\lim_{x\rightarrow a}f(x)\lim_{x\rightarrow a}g(x)=L\cdot M \tag{7.1.369}$$

- Quotient law for limits:

$$\lim_{x\rightarrow a}\frac{f(x)}{g(x)}=\frac{\lim\limits_{x\rightarrow a}f(x)}{\lim\limits_{x\rightarrow a}g(x)}=\frac{L}{M} \tag{7.1.370}$$

  for $M\neq 0$.

- Power law for limits:

$$\lim_{x\rightarrow a}(f(x))^n=\left(\lim_{x\rightarrow a}f(x)\right)^n=L^n \tag{7.1.371}$$

  for every positive integer n.

- Root law for limits:

$$\lim_{x\rightarrow a}\sqrt[n]{f(x)}=\sqrt[n]{\lim_{x\rightarrow a}f(x)}=\sqrt[n]{L} \tag{7.1.372}$$

  for all L if n is odd and for $L\geq 0$ if n is even.

7.1.2.1.4 Asymptotes

The term "asymptote" is used in mathematics to precise possible properties of an infinite branch of curve which growth tends to an infinitesimal value.

In analytic geometry, an asymptote of a curve is simply said to be a line such that the distance between the curve and the line approaches zero as they tend to infinity. In some contexts, such as algebraic geometry, an asymptote is defined as a line which is tangent to a curve at infinity.

The study of asymptotic behavior is particularly developed in the studies of functions. In the scientific field, it is common to study time-dependent functions (evolution of populations, chemical or nuclear reactions, temperature graphs, oscillation of a shock absorber). One of the researcher's objectives is then to know the state at the end of the experiment, that is to say when a large interval of time has passed. The objective is then not to know the intermediate variations, but to determine the stable, infinite behavior of the measured phenomenon. The researcher therefore studies the asymptotic behavior of his function with the tools that mathematics offers him.

Remark

The word "asymptote" is derived from the Greek and means "not falling together".

Definitions (#7.1.277):

D1. When the limit of a function $f(x)$ tends to a constant c^{te} when $x \to \pm\infty$, then the graphical representation of this function leads us to draw a horizontal line that we name "horizontal asymptote" which equation satisfies:

$$\lim_{x\to\pm\infty} f(x) - c^{te} = f(x_0) \tag{7.1.373}$$

D2. When the limit of a function $f(x)$ tends to $\pm\infty$ when $x \to a_{+,-}$, then the graphical representation of this function leads us to draw a vertical line that we name "vertical asymptote" which equation satisfies:

$$\lim_{x\to a_{+,-}} f(x) \pm \infty = 0 \tag{7.1.374}$$

Vertical asymptotes is the typical symptom of a division by zero in a fraction and has a very important place in physics. The syndrome is also named a "singularity".

✍**Example:**

The graph of the function:

$$f(x) = \frac{1}{x-1} \tag{7.1.375}$$

has the straight lines of $x = 1$ as vertical asymptote and $y = 0$ as horizontal asymptote:

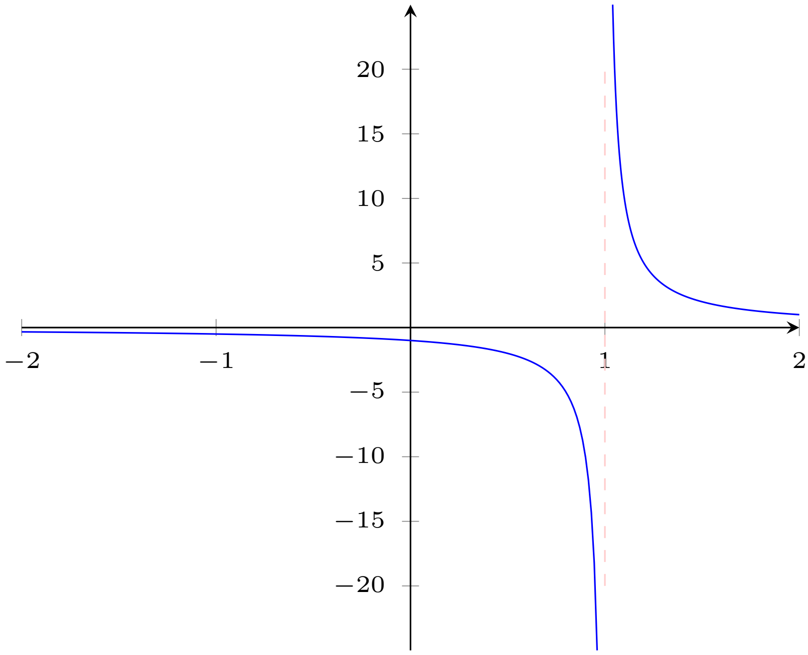


Figure 7.51: **Graphical representation of a horizontal and vertical asymptote**

D3. The straight line of equation $y=ax+b$ is an "oblique asymptote" or "slant asymptote" of a curve of function $f(x)$ if:

$$\lim_{x\to\pm\infty} f(x)-(ax+b)=0 \tag{7.1.376}$$

the values of a and b can be easily found using the following relations:

$$\begin{gathered}\lim_{x\to\pm\infty}\frac{f(x)}{x}\\ \lim_{x\to\pm\infty} f(x)-ax\end{gathered} \tag{7.1.377}$$

> Caution! A curve may have two distinct oblique asymptotes in $+\infty$ and $-\infty$.

To find a possible oblique asymptote, one must already be certain that the function $f(x)$ admits an infinite limit in $+\infty$ or $-\infty$ then only we look for the limits at $-\infty$ and $+\infty$ of $f(x)/x$ and $f(x)-ax$.

Three typical cases can be considered for oblique asymptotes:

(a) The representative curve of $f(x)$ has for asymptotical direction the affine equation $y=ax$:

$$\lim_{x\to\pm\infty}\frac{f(x)}{x}=a \tag{7.1.378}$$

✍**Example:**

The graph of the function:

$$f(x) = x + \frac{1}{x} \tag{7.1.379}$$

has the straight line of $y = x$ as oblique asymptote:

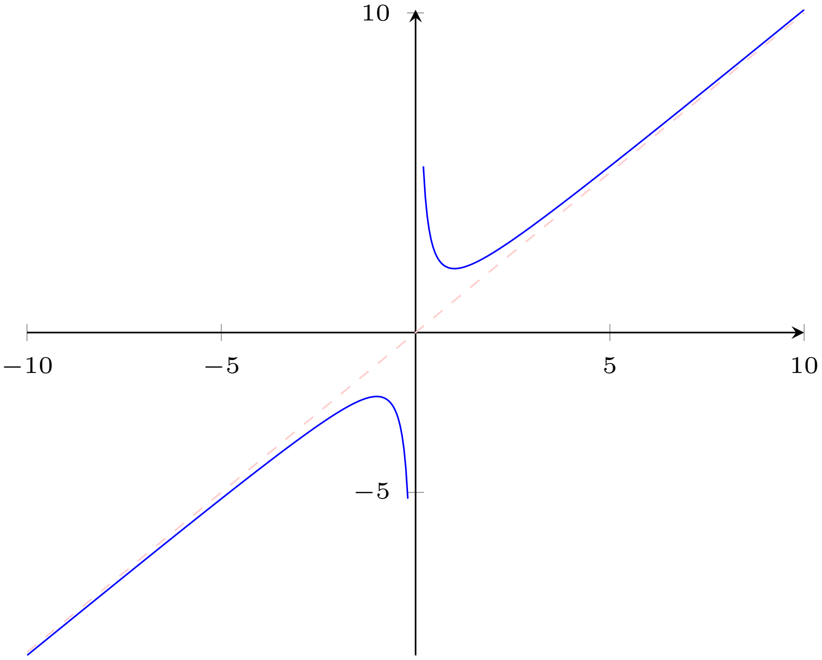


Figure 7.52: **Graphical representation of an oblique affine asymptote**

(b) The representative curve of $f(x)$ has an infinite branch (this branch has not close form asymptote) and the only one thing we can say is that x-axis is the direction of this asymptote. Such an asymptote exists when:

$$\lim_{x \to \pm\infty} \frac{f(x)}{x} = 0 \tag{7.1.380}$$

✍**Example:**

The functions $f(x)=\sqrt{x}$ (in red) or $\ln(x)$ (in green) have a limit $f(x)/x$ equal to 0 and both have a "parabolic branch" of direction following the x-axis:

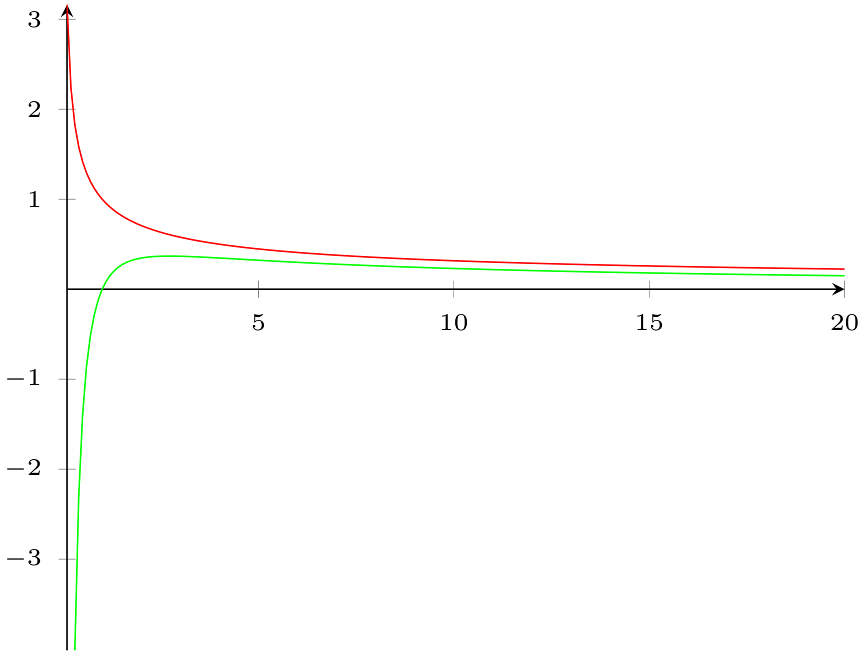


Figure 7.53: **Graphical representation of a parabolic branch example following x-axis**

(c) The representative curve of $f(x)$ has an infinite branch (this branch has not close form asymptote) and the only one thing we can say is that y-axis is the direction of this asymptote (we then also speak of "parabolic branch"):

$$\lim_{x\to\pm\infty}\frac{f(x)}{x}=\pm\infty \tag{7.1.381}$$

(d) A function $f(x)$ is say to have a "curvilinear asymptote" if it satisfies:

$$\lim_{x\to\pm\infty} f(x)=P_n(x) \tag{7.1.382}$$

for $n>1$ where for recall $P_n(x)$ is a polynomial of degree n.

✍**Example:**

The function :

$$f(x)=\frac{x^3+2x^2+3x+4}{x} \tag{7.1.383}$$

has a curvilinear asymptote that is:

$$x^2+2x^2+3x \tag{7.1.384}$$

Indeed:

$$\lim_{x\to\pm\infty} f(x)=\frac{x^3+2x^2+3x+4}{x}=x^2+2x^2+3x+\frac{4}{x} \tag{7.1.385}$$

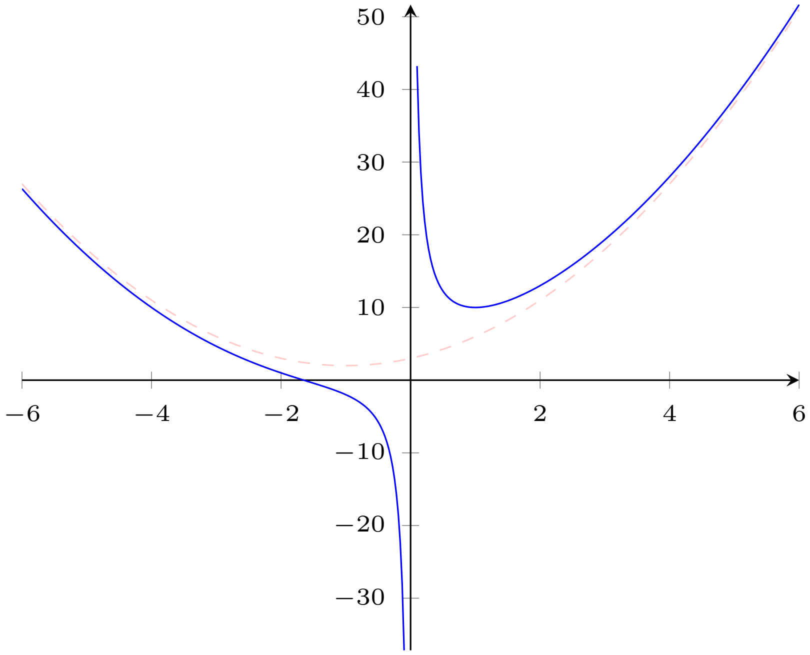


Figure 7.54: **Graphical representation of a curvilinear asymptote of equation** x^2+2x+3

There are many techniques for finding limits that apply in various conditions. It's important to know all these techniques, but it's also important to know when to apply which technique. Some basic techniques who doesn't involve derivation are:

Calculating $\lim\limits_{x\to 0} f(x)$

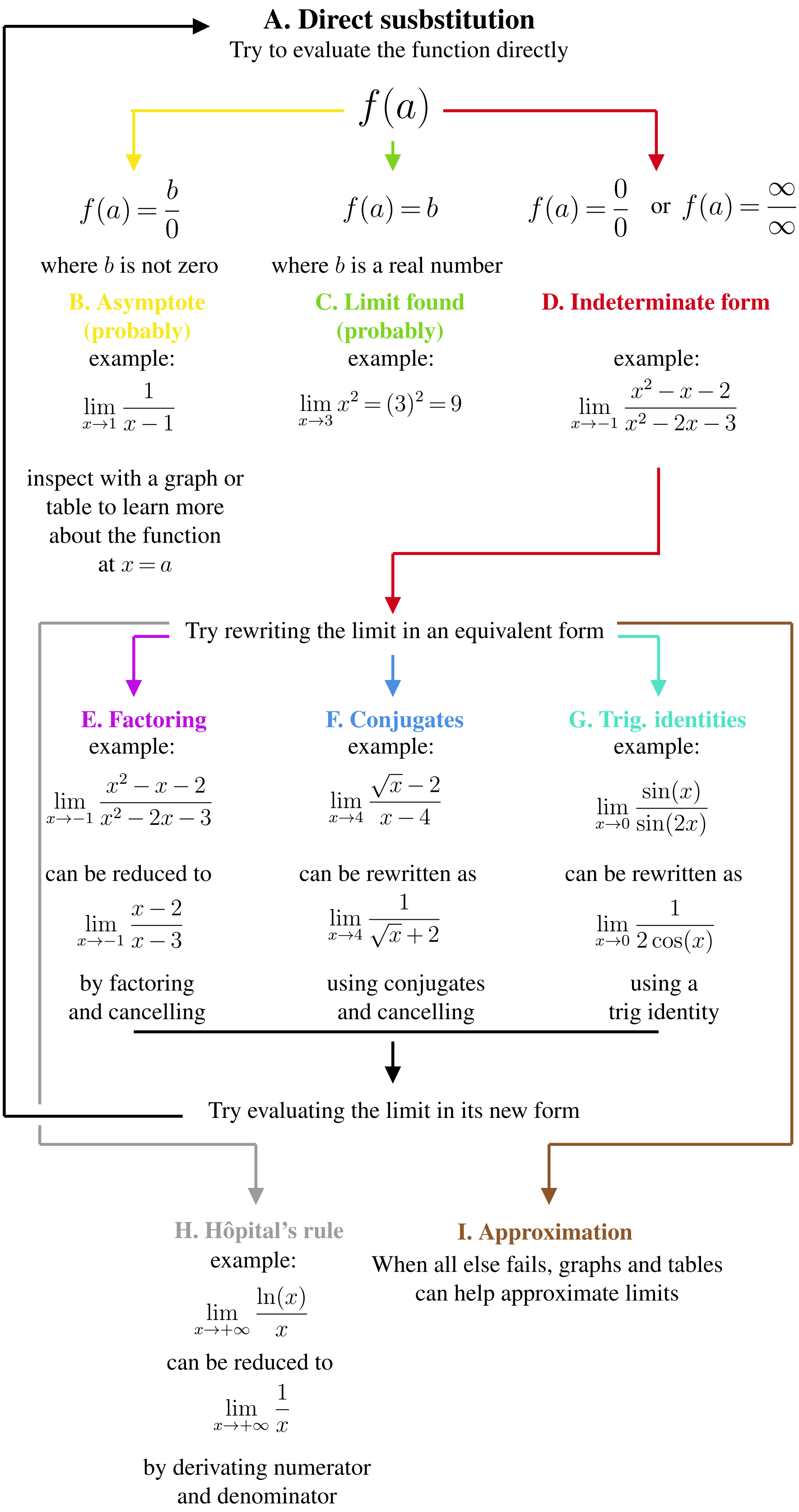

A. Direct susbstitution
Try to evaluate the function directly
$f(a)$
$f(a)=\frac{b}{0}$
where b is not zero
B. Asymptote (probably)
example:
$\lim\limits_{x\to 1}\frac{1}{x-1}$
inspect with a graph or table to learn more about the function at $x=a$
$f(a)=b$
where b is a real number
C. Limit found (probably)
example:
$\lim\limits_{x\to 3} x^2=(3)^2=9$
$f(a)=\frac{0}{0}$ or $f(a)=\frac{\infty}{\infty}$
D. Indeterminate form
example:
$\lim\limits_{x\to -1}\frac{x^2-x-2}{x^2-2x-3}$
Try rewriting the limit in an equivalent form
E. Factoring
example:
$\lim\limits_{x\to -1}\frac{x^2-x-2}{x^2-2x-3}$
can be reduced to
$\lim\limits_{x\to -1}\frac{x-2}{x-3}$
by factoring and cancelling
F. Conjugates
example:
$\lim\limits_{x\to 4}\frac{\sqrt{x}-2}{x-4}$
can be rewritten as
$\lim\limits_{x\to 4}\frac{1}{\sqrt{x}+2}$
using conjugates and cancelling
G. Trig. identities
example:
$\lim\limits_{x\to 0}\frac{\sin(x)}{\sin(2x)}$
can be rewritten as
$\lim\limits_{x\to 0}\frac{1}{2\cos(x)}$
using a trig identity
Try evaluating the limit in its new form
H. Hôpital's rule
example:
$\lim\limits_{x\to +\infty}\frac{\ln(x)}{x}$
can be reduced to
$\lim\limits_{x\to +\infty}\frac{1}{x}$
by derivating numerator and denominator
I. Approximation
When all else fails, graphs and tables can help approximate limits

Figure 7.55: **Basic techniques for finding limits**

7.1.2.1.5 Discontinuities

Continuous functions are of utmost importance in mathematics! However, not all functions are continuous. If a function is not continuous at a point in its domain, one says that it has a "discontinuity" there. The set of all points of discontinuity of a function may be a discrete set, a dense set, or even the entire domain of the function.

In practice we consider the four following types of discontinuities (more details and examples follow after the figure!):

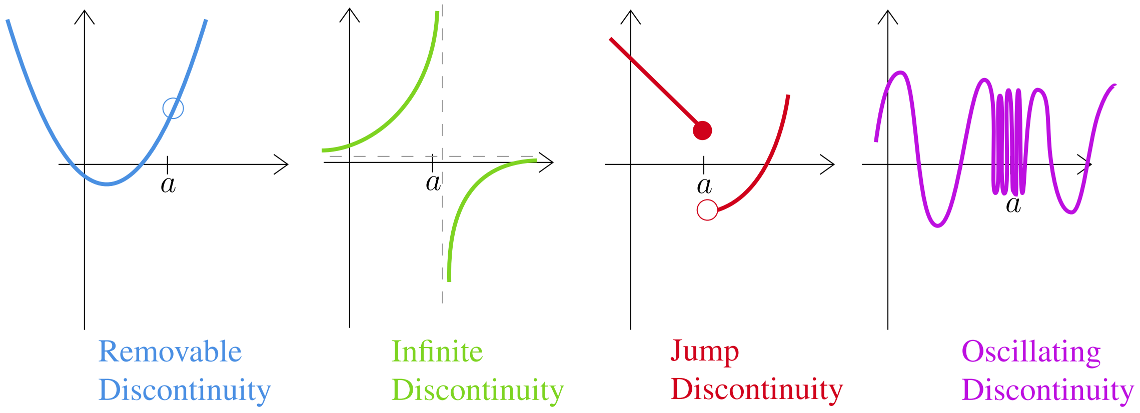


Figure 7.56: **Common types of discontinuities**

For each of the following, consider a real valued function $f(x)$ of a real variable x, defined in a neighborhood of the point x_0 at which $f(x)$ is discontinuous!

Definitions (#7.1.278):

D1. Removable discontinuity:

Consider the following fraction:

$$f(x)=\frac{x^3-3x^2+2x}{x-1} \tag{7.1.386}$$

It seems to have a discontinuity (singularity) in $x_0=1$, but in facts it's removable one (or an "apparent singularity"). Indeed we can easily remove it:

$$f(x)=\frac{x^3-3x^2+2x}{x-1}=\frac{x(x^2-3x+2)}{x-1}=\frac{x(x-1)(x-2)}{x-1}=x(x-2) \tag{7.1.387}$$

And that's it! No discontinuity anymore is visible nor exist. And obviously using the techniques we have seen earlier, it is immediate that this function is epsilon-delta continuous!

The one-sided limit from the negative direction:

$$L^-=\lim_{x\to x_0^-}f(x) \tag{7.1.388}$$

and the one-sided limit from the positive direction:

$$L^+=\lim_{x\to x_0^+}f(x) \tag{7.1.389}$$

at x_0 both exist, are finite, and are equal to $L = L^- = L^+$. In other words, since the two one-sided limits exist and are equal, the limit L of $f(x)$ as x approaches x_0 exists and is equal to this same value. If the actual value of $f\left(x_0\right)$ is not equal to L, then x_0 is named a "removable discontinuity".

D2. Infinite discontinuity/Essential discontinuity:

Consider the function:

$$f(x) = \frac{1}{x} \tag{7.1.390}$$

Notice that as the value of x approaches 0 from the left, the function grows infinitely smaller! Likewise, on the right of the y-axis, as the value of x approaches 0, the function increases infinitely!

How do we prove that this function doesn't have a limit at $x_0 = 0$ using epsilon-delta discontinuity?

That's quite simple! Remember that a function is epsilon-delta continuous if:

$$\forall x \in D \forall \varepsilon > 0 \exists \delta > 0 \forall x_0 \in D : |x - x_0| < \delta \Rightarrow |f(x) - f(x_0)| < \varepsilon \tag{7.1.391}$$

So if we choose $x_0 = 0$ that means obviously if a limit L would exist at this point that:

$$\forall x \in D \forall \varepsilon > 0 \exists \delta > 0 \forall x_0 \in D : |x| < \delta \Rightarrow \left| \frac{1}{x} - L \right| < \varepsilon \tag{7.1.392}$$

Hence:

$$\forall x \in D \forall \varepsilon > 0 \exists \delta > 0 \forall x_0 \in D : 0 < x < \delta \Rightarrow -\varepsilon < \frac{1}{x} - L < +\varepsilon \tag{7.1.393}$$

Let us choose arbitrarily $\varepsilon = 1$ and focus on the right part of the x axis (hence $L > 0$). Therefore if L exist, we should have:

$$\frac{1}{x} - L < +\varepsilon \Rightarrow \frac{1}{x} - L < 1 \tag{7.1.394}$$

ie:

$$\frac{1}{x} < L + 1 \tag{7.1.395}$$

And as we should be able to choose any $0 < x < \delta$ this contradicts obviously the previous inequality as we may always choose a x that gives $1/x > L + 1$. Hence the contradiction, therefore the limit L doesn't exist and $f(x) = 1/x$ is not epsilon-delta continuous at $x_0 = 0$!

Hence a function has an infinite discontinuity if:

$$\lim_{x \to x_0} f(x) = \pm\infty \tag{7.1.396}$$

In this example, both L^- and L^+ do not exist in $\mathbb{R}$, thus satisfying the condition of essential discontinuity. So x_0 is an "essential discontinuity", an "infinite discontinuity" or a "discontinuity of the second kind".

D3. Jump discontinuity:

Consider the function:

$$f(x) = \begin{cases} 2x+3, x \geq 1 \\ 2x-2, x < 1 \end{cases} \tag{7.1.397}$$

Then, the point $x_0 = 1$ is a jump discontinuity.

In this case, a single limit does not exist because the one-sided limits, $L^-$and L^+exist and are finite, but are not equal: since, $L^- \neq L^+$, the limit L does not exist. Then, x_0 is called a jump discontinuity, step discontinuity, or discontinuity of the first kind. For this type of discontinuity, the function f may have any value at x_0.

Let us prove that with the epsilon-delta continuity!

We have $f(1) = 5$. So to show that $f(x)$ is not continuous at $x = 1$, it is enough to show that it is not true that $\lim_{x\to 1} f(x) = 5$.

Suppose to the contrary that the limit exists and is equal to 5. Then for any $\varepsilon > 0$, there is a $\delta > 0$ such that if $|x-1| < \delta$, then $|f(x)-5| < \varepsilon$.

Pick $\varepsilon = 1/2$. We show there is no δ with the required property.

If $x < 1$, then $f(x) < 1$. In particular, if $x < 1$, then $|f(x)-5| > 4$. It follows that there is no δ such that $|x-1| < \delta$ guarantees that $|f(x)-5| < \varepsilon$.

D4. Oscillating discontinuity:

Consider the function:

$$f(x) = \sin\left(\frac{1}{x}\right) \tag{7.1.398}$$

with it's corresponding plot:

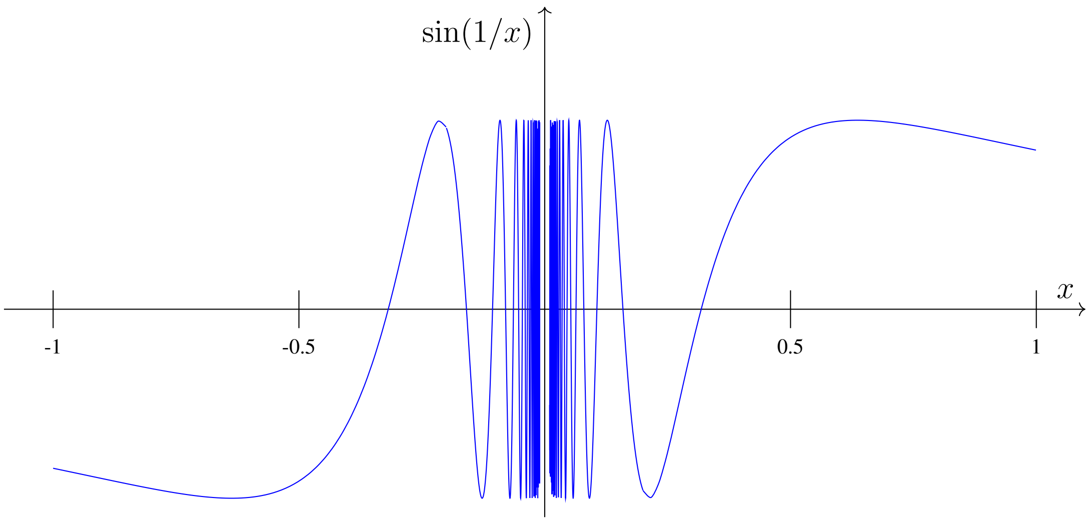


Figure 7.57: **Example of a famous oscillating discontinuity**

Then, the point $x_0 = 0$ is an oscillating discontinuity.

Or a more formal approach: Let $f(x) = \sin(1/x)$.

If $f(x)$ attains a limit L then there exists a $\delta > 0$ for every $\varepsilon > 0$ such that $|f(x) - L| < \varepsilon$ such that $0 < |x - 0| < \delta$.

Considering it from the right we have $0 < x < \delta$ or $+\infty < 1/x < 1/\delta$. Which means There are infinite number of $1/x$ of the form $2n\pi$. Which means the function is oscillating in the interval and the limit does not exist!

It's quite a fascinating result to imagine that for such a simple deterministic function (and that's not chaos physics nor quantum physics stuff but pure mathematics!) leads to a non-deterministic result... That's a nice example that even in pure mathematics not everything leads to perfect known result!

Oscillating discontinuities are bounded. In other words, their oscillations stay between certain lines. For example, the function might be bounded between a high point of $y = +3$ and a low point of $y = -1$. If the function is unbounded at one or both sides, it's an infinite discontinuity. In other words, the function must be completely bounded at all points in order for there to be an oscillating discontinuity. In addition, the one sided limits do not exist at all.

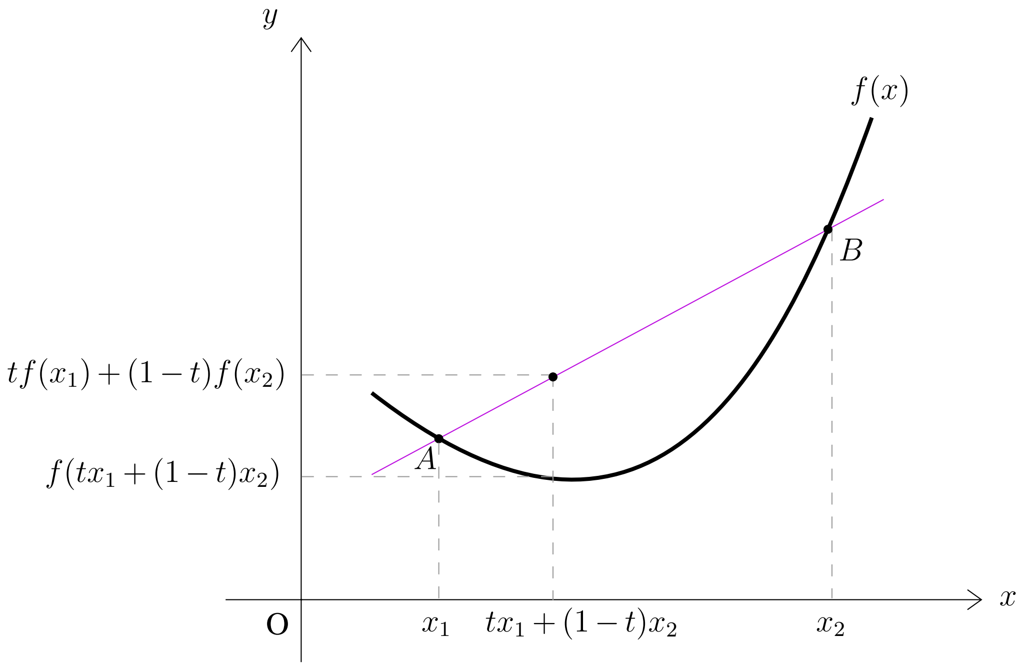
Remark

Oscillating discontinuities are considered by mathematicians as a sub-type of "essential discontinuities" or "non-repairable discontinuities".

7.1.2.2 Concavity/Convexity of a function

We will define now a property which at first sight may seem of no interest as it is so trivial but which we shall find in the section of Statistics for the demonstration of an important relation named "Jensen inequality" and which is of major importance Finance and Insurance for the valuation of options and premiums (see section Economy page 7674) and especially for the application of the Jensen's inequality (see section Statistics page 576).

Consider the following figure:



Definition (#7.1.279): In mathematics, a real function of a real variable is say to a "convex function" if, viewed from below, its graph is convex (in bump); we mean that if A and B are two points of the graph of the function, the segment $\overline{AB}$ is entirely situated above the graph. It is the same to say that the "epigraph" (the set of points above the graph) is a concave set. Conversely, a function whose graph, as seen from below, is seen as a cave, is say to be a "concave function". It is the same to say that the "hypograph" (the set of points below) is a convex set.

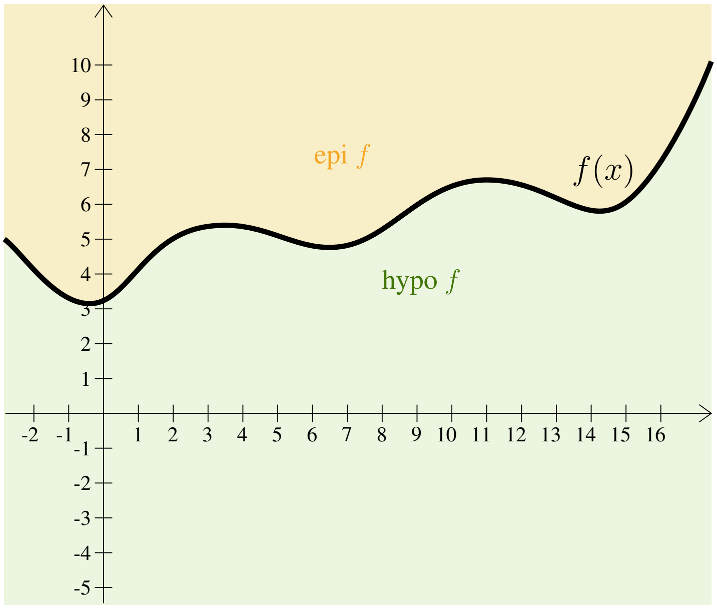


Figure 7.58: **Epigraph and Hypograph of a function**

By specifying by the values of the function what are the points A and B above, we get often an equivalent definition of the convexity of a function: a function defined on a real interval I is convex when, for any x_1 and x_2 of I and all t in $[0,1]$ we have:

$$\boxed{f(tx_1+(1-t)x_2)\leq tf(x_1)+(1-t)f(x_2)} \tag{7.1.399}$$

When the inequality is strict, then we obviously speak of a "strictly convex function".

Remark

Without proof, just by looking to the above chart, we will assume quite obvious that convexity implies $f''(x)\geq 0$ for all x. Just as before, strict convexity occurs when the inequality is strict.

By extension (common sense from our point of view), a function f is concave if $-f$ is convex (which is trivial with the payoff function - see section Economy page 7674 - profile of options from seller or buyer point of view).

✍**Examples:**

E1. Let us consider $f(x)=x^2$. The first derivative of $f(x)$ is given by $\frac{\mathrm{d}}{\mathrm{d}x}f=2x$ and its second derivative by $\frac{\mathrm{d}^2}{\mathrm{d}x^2}f=2$. Since this is always strictly greater than 0, we have proven that $f(x)=x^2$ is strictly convex.

E2. Let us consider $f(x)=\log(x)$. The first derivative is $\frac{\mathrm{d}}{\mathrm{d}x}f=\frac{1}{x}$ and its second derivative is given by $\frac{\mathrm{d}^2}{\mathrm{d}x^2}f=-\frac{1}{x^2}$. Since this is negative for all $x>0$, we have proven that $\log(x)$ is a concave function over $\mathbb{R}_+$.

7.1.2.3 Euler theorem for homogeneous functions

We have to introduce now a definition and a theorem that will be very important for quite advanced concepts in our study of Physics later in this textbook.

Definition (#7.1.280): A "homogeneous function" is one with multiplicative scaling behaviour: if all its arguments are multiplied by a factor, then its value is multiplied by some power of this factor. For example, a homogeneous real-valued function of two variables x and y is a real-valued function that satisfies the condition $f(tx,ty)=t^k f(x,y)$ for some constant k and all real numbers $t\in\mathbb{R}^*$. The constant k is named the "degree of homogeneity".

More generally, if $f:\mathbb{R}^n\mapsto\mathbb{R}$ is a function between two vector spaces over a field F, and k is an integer, then f is said to be homogeneous of degree k if:

$$f(tx,ty,ty)=t^k f(x,y,z) \tag{7.1.400}$$

or more commonly written:

$$f(t\vec{x})=t^k f(\vec{x}) \tag{7.1.401}$$

✍**Example:**

The function $f(x,y)=x^2+y^2$ is homogeneous of degree 2. Indeed:

$$f(tx,ty)=(tx)^2+(ty)^2=t^2(x^2+y^2)=t^2 f(x,y) \tag{7.1.402}$$

Now let us suppose $f:\mathbb{R}^n\mapsto\mathbb{R}$ is continuously differentiable on $\mathbb{R}$. We know that a function is homogeneous of degree k if:

$$f(t\vec{x})=t^k f(\vec{x}) \tag{7.1.403}$$

Differentiating both sides with respect to t, we get:

$$\frac{\mathrm{d}f(t\vec{x})}{\mathrm{d}t}=\sum_{i=1}^{n}\frac{\partial f(tx)}{\partial x_i}x_i=kt^{k-1}f(x) \tag{7.1.404}$$

by the total exact differential chain rule of f.

So the "Euler theorem for homogeneous functions" can be summarized as following:

$$\boxed{f(tx,ty,ty)=t^k f(x,y,z)\Leftrightarrow\sum_{i=1}^{n}\frac{\partial f(tx)}{\partial x_i}x_i=kt^{k-1}f(x)} \tag{7.1.405}$$

Remark

Remember that we have proved in the section of Differential and Integral Calculus (see page 1466) that:

$$\mathrm{d}f=\frac{\partial f}{\partial x}\mathrm{d}x+\frac{\partial f}{\partial y}\mathrm{d}y \tag{7.1.406}$$

Therefore:

$$\frac{\mathrm{d}f}{\mathrm{d}t}=\frac{\partial f}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t}+\frac{\partial f}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t}=\frac{\partial f}{\partial x}\frac{\mathrm{d}(tx)}{\mathrm{d}t}+\frac{\partial f}{\partial y}\frac{\mathrm{d}(ty)}{\mathrm{d}t}=\frac{\partial f}{\partial x}x+\frac{\partial f}{\partial y}y \tag{7.1.407}$$

Then if we choose to set with the special case $k=1$ (it's the case that will interest us the most in Physics) then the above becomes:

$$\sum_{i=1}^{n}\frac{\partial f(x)}{\partial x_i}x_i=f(x) \tag{7.1.408}$$

or explicitly for $n=2$:

$$x\frac{\partial f}{\partial x}+y\frac{\partial f}{\partial y}=f(x,y) \tag{7.1.409}$$

if $f(tx)$ is homogeneous of degree 1.

This is an important result we will need in Lagrangian Mechanics that will be useful for the study of the Lagrangian of a free particle in Special Relativity and in Quantum Cosmology.

7.1.3 Logarithms

We hesitated to put the definition of logarithms in the section Calculus. After a moment of reflection, we decided it was better to put it in this section because to understand it well, we must be aware of the concept of limits, of definition domain and of the power function. We hope that our choice will suit you best.

Given the power (bijective) function of any base where $a \in \mathbb{R}_{+}^{*}/1$ (we exclude 1 otherwise it is not bijective) and denoted for recall by:

$$f(x)=a^x \tag{7.1.410}$$

for which it corresponds to each real number x, exactly one positive number a^x (the image set of the function is in $\mathbb{R}$) such as the powers calculations rules are applicable (see section Calculus page 1353).

We know that for such a function that if $a>1$, then $f(x)$ is an increasing and positive (monotone) in $\mathbb{R}$, and if $0<a<1$, then $f(x)$ is positive and decreasing (monotone) in $\mathbb{R}$.

Remarks

R1. If $a>1$, when x decreases to negative values, the graph of $f(x)$ approaches the x-axis. Thus, the x axis is a horizontal asymptote. When x increases in positive values, the graph rises quickly. This type of change is characteristic of the "law of exponential of growth" and $f(x)$ is sometimes named "growing function"... If $0<a<1$, when x increases, the graph tends asymptotically to the x-axis. This type of variation is known as an "exponential decay".

R2. By studying a^x, we exclude the case where $a \leq 0$ and $a=1$. Notice that if $a<0$, then a^x is not a real number for many values of x (we recall that the whole image set is forced to $\mathbb{R}$ in our previous definition). If $a=0$, the $a^0=0$ is not defined. Finally, if $a=1$, then $a^x=1$ for all x and the graph of $f(x)$ is a horizontal line.

As the power function $f(x)$ is bijective then there exists an inverse function $f^{-1}(x)$ and is named "logarithm function" of base a and is denoted by:

$$f^{-1}(y)=\log_a(y) \tag{7.1.411}$$

and therefore:

$$x=\log_a(y) \tag{7.1.412}$$

if and only if $y=a^x$.

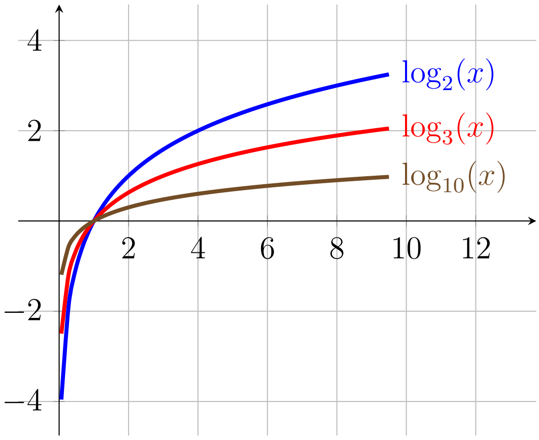
More generally it is defined by:

$$\boxed{\log_a(y)=x \Leftrightarrow a^x=y \qquad a>0, a\neq 1} \tag{7.1.413}$$

Considering $\log_a(x)$ as an exponent, we have the following properties:

Properties	**Justification**
$\log_a 1 = 0$	$a^0 = 1$
$\log_a a = a$	$a^1 = a$
$\log_a a^x$	$a^x = a^x$
$a^{\log_a(x)} = x$	$a^{\log_a(x)} = a^y = x$

Table 7.6: **Properties of the logarithm in base a**



Remarks

R1. The word "logarithm" means "number of logos", "logos" meaning "reason" or "ratio".

R2. The logarithm and power functions are defined by their bases (the number a). When using a power of 10 as a base $(10, 100, 1000, \ldots)$ then we speak of "common system" because they have for $\log$ successive integers.

R3. The integer part of the logarithm is named the "characteristic".

There are two types of logarithms that we find almost exclusively in mathematics and physics: the logarithm of base 10, logarithm of base e (the latter often named "natural logarithm") and logarithm of base 2 for information theory.

First the one in base 10 (the most used on graphical representations):

$$f(x) = \log_{10}(x) \tag{7.1.414}$$

abusively noted:

$$f(x) = \log(x) \tag{7.1.415}$$

and the base (Eulerian) e:

$$f(x) = \log_e(x) \tag{7.1.416}$$

historically noted:

$$\boxed{f(x) = \ln(x)} \tag{7.1.417}$$

the "n" meaning "Napierian".

Remark

Historically, it is John Napier (11550-11617 according to holocene calendar) whose name was Latinized "Napier" that we own the study of logarithms and the name of "natural logarithms" which aimed to facilitate greatly the time for manual calculations.

In English for the logarithm function in base-10 logarithmic we need to calculate:

$$f(x)=\log(x) \tag{7.1.418}$$

ask the following question: at what power $n \in \mathbb{R}$ should we raise 10 to get x?

Formally, this consist to solve the equation:

$$10^n=x \tag{7.1.419}$$

or written in another way:

$$10=\sqrt[n]{x} \tag{7.1.420}$$

with x being known and therefore in base 10:

$$\log(x)=n \tag{7.1.421}$$

To work with exponents of different bases, you also need its opposite operation, the logarithm we have just defined!. So if you are looking to (very useful for pedagogical reasons):

$$2^x=10^y \tag{7.1.422}$$

applying a $\log_{10}$ to the previous equation:

$$y=\log_{10} 2^x=x\cdot\log_{10} 2=0.30103x \tag{7.1.423}$$

Therefore we typically we have an order of $2^{10}\approx 10^3$.

Therefore if we remember that the probability of our universe as being as it is, is approximately of the order of $10^{-10^{500}}$ (see page 474) therefore this probability is equivalent to get a given outcome when throwing approximately $10^{10^{1500}}$ coins.

The logarithm in base 10 is used a lot in graphical representations in the scientific perspective when we look at amplitudes variations. For example with Maple™ 4.00b we have for two sine function having respectively for their respective mean the same amplitude variation of 50% visible below that do not highlights necessarily this fact trivially:

```
>plot(10+0.5*10*sin(x),100+100*0.5*sin(x),x=1..10);
```

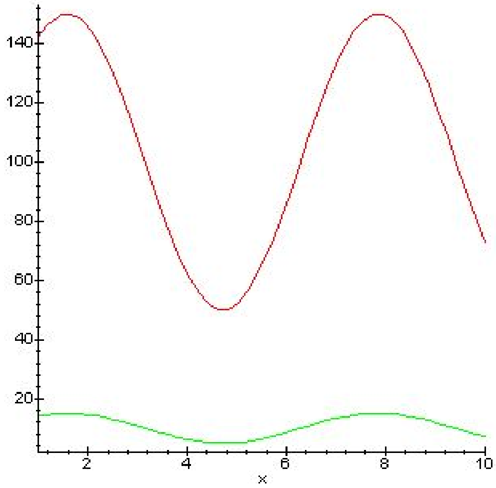


Figure 7.59: Plot with Maple™ 4.00b with two sine functions having same amplitude change compared to their average

While in logarithmic scale, this gives:

```
>with(plots):
>logplot(10+0.5*10*sin(x),100+100*0.5*sin(x),x=1..10);
```

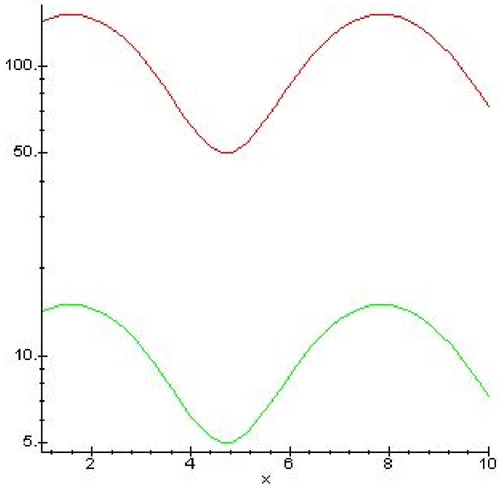


Figure 7.60: Same plot with Maple™ 4.00b but with the y-axis in logarithm (base 10) scale

The logarithmic scale is more well-known for the following kind of plot grid:

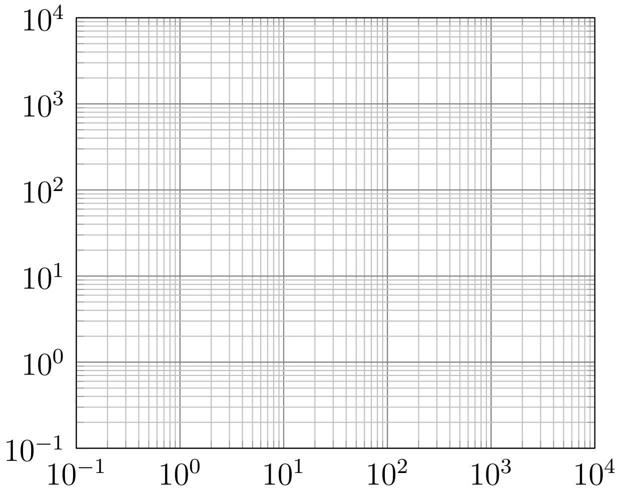


Figure 7.61: **Double logarithmic grid**

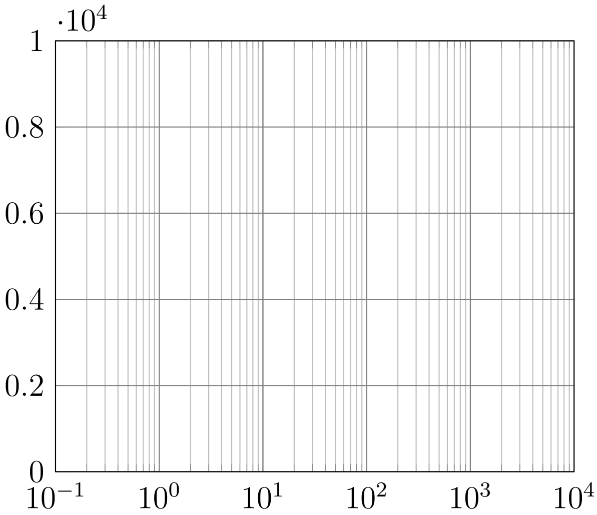


Figure 7.62: **Simple horizontal logarithmic grid**

For the logarithmic function in Eulerian base e it is necessary to calculate:

$$f(x)=\ln(x) \tag{7.1.424}$$

to ask ourselves the following question: at what power $n\in\mathbb{R}$ we must raise the number e to get x?

Formally this consists to solve the equation:

$$e^n=x \tag{7.1.425}$$

with x being known and therefore:

$$\ln(x)=n \tag{7.1.426}$$

Technically, we say that the exponential function (see below for details):

$$\boxed{f(x)=e^x} \tag{7.1.427}$$

is the inverse bijection of the $\ln(x)$ function.

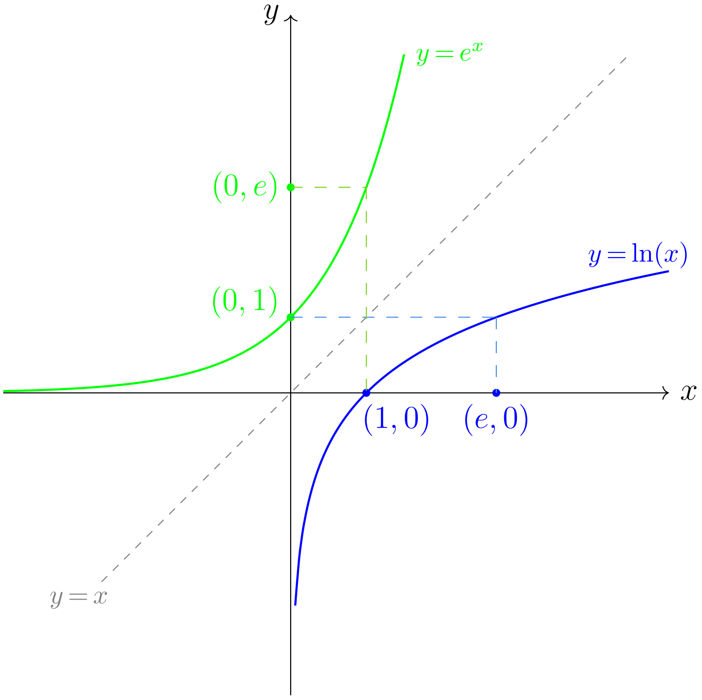


Figure 7.63: Graphical representation of the correspondence between the natural logarithm and the exponential

But what is that "Eulerian" number also named "Euler number"? Why do we find so often in physics and mathematics? Let us first determine the origin of its value:

$$\sqrt[n]{1+n}=(1+n)^{\frac{1}{n}}=\left(1+\frac{1}{\alpha}\right)^{\alpha} \tag{7.1.428}$$

with $\alpha \in \mathbb{N}$ and when $\alpha \rightarrow +\infty$.

Remark

The second term of the equality is typically the type of expression that we find in compound interest in finance (see section Economy page 7557) or in any other type of identical increase factor. And what interests us in this case is when this type of increase tends to infinity.

The interest we have to pose the problem as in this way is that if we do tend $\alpha \rightarrow +\infty$ the function written above tends to e and this function has the special property of being calculable more or less easily for historical reasons using Newton's binomial.

So according to the development of the Newton binomial (see section Calculus page 1364) we can write:

$$\begin{aligned}\left(1+\frac{1}{\alpha}\right)^{\alpha}=&1+\frac{\alpha}{1!}\frac{1}{\alpha}+\frac{\alpha(\alpha-1)}{2!}\frac{1}{\alpha^2}+\frac{\alpha(\alpha-1)(\alpha-2)}{3!}\frac{1}{\alpha^3}\\&+\cdots+\frac{\alpha(\alpha-1)\ldots(\alpha-(\alpha-1))}{\alpha!}\frac{1}{\alpha^{\alpha}}\end{aligned} \tag{7.1.429}$$

This development is similar to the Taylor expansion (see section Sequences and Series page 1696) of some given functions for particular cases of development values (hence the reason why we find this eulerian number in many fields related to applied mathematics as we will see later).

By performing some algebraic transformations that should now be obvious to the reader, we find:

$$\begin{aligned}\left(1+\frac{1}{\alpha}\right)^{\alpha}=1+1+\frac{1}{2!}\left(1-\frac{1}{\alpha}\right)+\frac{1}{3!}\left(1-\frac{1}{\alpha}\right)\left(1-\frac{2}{\alpha}\right)\\+\cdots+\frac{1}{\alpha!}\left(1-\frac{1}{\alpha}\right)\left(1-\frac{2}{\alpha}\right)\ldots\left(1-\frac{\alpha-1}{\alpha}\right)\end{aligned}\tag{7.1.430}$$

We see in this last equality that the function $\left(1+\frac{1}{\alpha}\right)^{\alpha}$ is increasing when α increases. Indeed, when we move from α to the value $\alpha+1$ each term of this sum increases:

$$\frac{1}{2!}\left(1-\frac{1}{\alpha}\right)<\frac{1}{2!}\left(1-\frac{1}{\alpha+1}\right)\text{, etc.}\tag{7.1.431}$$

Let us prove now that the variable $\left(1+\frac{1}{\alpha}\right)^{\alpha}$ is bounded. By seeing that:

$$\left(1-\frac{1}{\alpha}\right)<1,\;\left(1-\frac{1}{\alpha}\right)\left(1-\frac{2}{\alpha}\right)<1\text{, etc.}\tag{7.1.432}$$

So we get by analogy with the extended expression of Newton binomial determined just previously the following order relation:

$$\left(1-\frac{1}{\alpha}\right)^{\frac{1}{\alpha}}<1+\frac{1}{2!}+\frac{1}{3!}+\cdots+\frac{1}{\alpha!}\tag{7.1.433}$$

On the other hand:

$$\frac{1}{3!}<\frac{2}{2!},\;\frac{1}{4!}<\frac{2}{3!},\;\frac{1}{\alpha!}<\frac{2}{2^{\alpha-1}}\tag{7.1.434}$$

We then can write the inequality:

$$\left(1-\frac{1}{\alpha}\right)^{\frac{1}{\alpha}}<1+\underline{1+\frac{1}{2}+\frac{1}{2^2}+\cdots+\frac{1}{2^{\alpha-1}}}\tag{7.1.435}$$

The underlined terms constitute a geometric sequence of reason $q=1/2$ (see section Sequences and Series page 1656) and whose first term is 1. If follows using the result obtained in the section of Sequences ans Series, that we can write:

$$\left(1+\frac{1}{\alpha}\right)^{\alpha}<1+\left[1+\frac{1}{2}+\frac{1}{2^2}+\cdots+\frac{1}{2^{\alpha-1}}\right]=1+\frac{1-(1/2)^{\alpha}}{1-(1/2)}=1+\left[2-\left(\frac{1}{2}\right)^{\alpha-1}\right]<3\tag{7.1.436}$$

Therefore, we have:

$$2\leq\left(1+\frac{1}{\alpha}\right)^{\alpha}<3\tag{7.1.437}$$

We have therefore proved that the function $\left(1+\frac{1}{\alpha}\right)^{\alpha}$ is bounded.

The limit:

$$\boxed{e=\lim_{\alpha\rightarrow\infty}\left(1+\frac{1}{\alpha}\right)^{\alpha}} \tag{7.1.438}$$

tends to this limited value that is the number e whose value is:

$$e=2.7182818284\ldots \tag{7.1.439}$$

The prior previous relation is also know in the following form after an obvious change of variable:

$$e=\lim_{\alpha\rightarrow 0}(1+\alpha)^{1/\alpha} \tag{7.1.440}$$

Remark

As we have proved it in the section Numbers, this number is irrational.

We can then define the "natural exponential function" (reciprocal of the natural logarithm function) by (for those who want to see a kind of circular proof instead than a definition... see page 2246):

$$\boxed{f(x)=e^x=\lim_{\alpha\rightarrow +\infty}\left(1+\frac{x}{\alpha}\right)^{\alpha}} \tag{7.1.441}$$

also sometimes denoted by:

$$f(x)=\exp(x) \tag{7.1.442}$$

The number e and the function that determines it are very useful. We find them in all areas of mathematics and physics and thus in almost all the chapters of this textbook.

As we have proved it in the section of Differential and Integral Calculus the functions e^x has for remarkable property that its derivative is equal to itself:

$$\frac{\mathrm{d}e^x}{\mathrm{d}x}=e^x \tag{7.1.443}$$

and this is used a lot for the resolution of differential equations in physics and finance.

Logarithms have several properties. Here are the most important one in our point of view (we are referring to a given base X) and that are very useful in physics, electronics, chemistry and so on...

Let us begin. First:

$$\log_X(X^mX^n)=\log_X(X^{m+n})=\log_X(X^m)+\log_X(X^n)=m+n \tag{7.1.444}$$

If we put $X^m=a$ and $X^n=b$ we get:

$$\log_X(a\cdot b)=\log_X(a)+\log_X(b) \tag{7.1.445}$$

If we have the special case when $a=b$ then:

$$\log_X(a\cdot a)=\log_X(a)+\log_X(a)=2\log_X(a) \tag{7.1.446}$$

Now let us try to express:

$$\log_X\left(\frac{a}{b}\right) \tag{7.1.447}$$

in another way. For this we put first:

$$t=\frac{a}{b}\Leftrightarrow a=tb \tag{7.1.448}$$

which leads us to the development:

$$\begin{gathered}\log_X(a)=\log_X(t\cdot b)=\log_X(t)+\log_X(b)\\ \Rightarrow\log_X(t)=\log_X(a)-\log_X(b)=\log_X\left(\frac{a}{b}\right)\end{gathered} \tag{7.1.449}$$

Now let us try to express:

$$\log_X\left(x^{1/n}\right) \tag{7.1.450}$$

with $n\in\mathbb{N}^*$ in another way. For this we put first:

$$t=x^{\frac{1}{n}}\Leftrightarrow x=t^n \tag{7.1.451}$$

which leads us to the development:

$$\log_X(x)=\log_X(t^n)=n\cdot\log_X(t)\Rightarrow\log_X(t)=\log_X\left(x^{\frac{1}{n}}\right)=\frac{1}{n}\log_X(x) \tag{7.1.452}$$

There is also another relation used a lot of time in physics in respect to the change of logarithm basis. The first relation is trivial and follows from the algebraic properties of logarithms:

$$a^x=e^{x\ln(a)}=10^{x\log_{10}(a)}=b^{x\log_b(a)} \tag{7.1.453}$$

the second one:

$$\log_a(u)=\frac{\ln(u)}{\ln(a)}=\frac{\log_{10}(u)}{\log_{10}(a)}=\frac{\log_b(u)}{\log_b(a)} \tag{7.1.454}$$

is a bit less trivial and requires perhaps a proof (we used it for our study of continued fractions in the section Number Theory).

Proof 7.3.2. We first use the equivalent equations (of the first relation above):

$$w=\log_a(u)\quad\text{and}\quad a^w=u \tag{7.1.455}$$

and we proceed as follows:

$$\begin{gathered}a^w=u\\ \log_b(a^w)=\log_b(u)\\ w\cdot\log_b(a)=\log_b(u)\end{gathered} \tag{7.1.456}$$

What finally brings us to:

$$w=\frac{\log_b(u)}{\log_b(a)}=\log_a(u) \tag{7.1.457}$$

■ Q.E.D.

7.1.4 Convolutions

A convolution is a mathematical operation on two functions to produce a third function, that is typically viewed as a modified version of one of the original functions, giving the integral of the point-wise multiplication of the two functions as a function of the amount that one of the original functions is translated.

There are different type of convolutions and as always, we will focus in this textbook only on the one that are actually used in other sections of this textbook.

7.1.4.1 Continuous and Discrete Linear Convolution Product

Definition (#7.1.281): The "continuous convolution" of two continuous signals $x(t)$ and $h(t)$ is defined as:

$$\boxed{y(t)=h(t)*x(t)\triangleq\int\limits_{-\infty}^{+\infty}x(\tau)h(t-\tau)\mathrm{d}\tau} \tag{7.1.458}$$

Remark

The convolution is also sometimes denoted with different symbols:

$$y(t)=h(t)\star x(t)\triangleq\int\limits_{-\infty}^{+\infty}x(\tau)h(t-\tau)\mathrm{d}\tau \tag{7.1.459}$$

or:

$$y(t)=h(t)\circledast x(t)\triangleq\int\limits_{-\infty}^{+\infty}x(\tau)h(t-\tau)\mathrm{d}\tau \tag{7.1.460}$$

We will now prove some properties of the convolution by focusing only on those use in the other sections of this textbook!

P1. Convolution is commutative:

$$x(t)*h(t)=h(t)*x(t) \tag{7.1.461}$$

Proof 7.3.3. By making the change of variable $\lambda=t-\tau$, in one form of the definition of convolution:

$$y(t)=h(t)*x(t)\triangleq\int\limits_{-\infty}^{+\infty}x(\tau)h(t-\tau)\mathrm{d}\tau \tag{7.1.462}$$

it becomes:

$$x(t)*h(t)=\int\limits_{-\infty}^{+\infty} x(t-\lambda)h(\lambda)\mathrm{d}\lambda=\int\limits_{-\infty}^{+\infty} h(\lambda)x(t-\lambda)\mathrm{d}\lambda=h(t)*x(t) \tag{7.1.463}$$

proving that convolution is commutative.

■ Q.E.D.

P2. Convolution is associative:

$$(f*g)(t)=\int\limits_{0}^{t} f(s)g(t-s)\mathrm{d}s \tag{7.1.464}$$

Proof 7.3.4. The proof is easier to understand if we consider a limited integral (but you can change the bounds to the infinity one and you will fall back on the general result):

$$\begin{aligned}((f*g)*h)(t)&=\int_0^t (f*g)(s)h(t-s)\mathrm{d}s\\&=\int_{s=0}^t\left(\int_{u=0}^s f(u)g(s-u)\mathrm{d}u\right)h(t-s)\mathrm{d}s\\&=\int_{s=0}^t\int_{u=0}^s f(u)g(s-u)h(t-s)\mathrm{d}u\mathrm{d}s\\&=\int_{u=0}^t\int_{s=u}^t f(u)g(s-u)h(t-s)\mathrm{d}s\mathrm{d}u\\&=\int_{u=0}^t\int_{s=0}^{t-u} f(u)g(s)h(t-s-u)\mathrm{d}s\mathrm{d}u\\&=\int_{u=0}^t f(u)\left(\int_{s=0}^{t-u} g(s)h(t-u-s)\mathrm{d}s\right)\mathrm{d}u\\&=\int_{u=0}^t f(u)(g\star h)(t-u)\mathrm{d}u\\&=(f*(g*h))(t)\end{aligned} \tag{7.1.465}$$

The proof may not be obvious for many readers. So we recommend to see the equivalent proof for the discrete version further below.

■ Q.E.D.

P3. Convolution is distributive:

$$x(t)*[h_1(t)+h_2(t)]=x(t)*h_1(t)+x(t)*h_2(t) \tag{7.1.466}$$

Proof 7.3.5.

$$\begin{aligned}x(t)*[h_1(t)+h_2(t)]&=\int\limits_{-\infty}^{+\infty} x(t)\left[h_1(t-\tau)+h_2(t-\tau)\right]\mathrm{d}\tau=\int\limits_{-\infty}^{+\infty} x(t)h_1(t-\tau)\mathrm{d}\tau+\int\limits_{-\infty}^{+\infty} x(t)h_1(t-\tau)\mathrm{d}\tau\\&=x(t)*h_1(t)+x(t)*h_2(t)\end{aligned} \tag{7.1.467}$$

■ Q.E.D.

P4. Relation with differentiation:

$$(f*g)'=f'*g=f*g' \tag{7.1.468}$$

Proof 7.3.6.

$$\begin{gathered}(f*g)'=\frac{\mathrm{d}}{\mathrm{d}t}\int\limits_{-\infty}^{+\infty}f(\tau)g(t-\tau)\mathrm{d}\tau=\int_{-\infty}^{+\infty}f(\tau)\frac{\partial}{\partial t}g(t-\tau)\mathrm{d}\tau=\int\limits_{-\infty}^{+\infty}f(\tau)\frac{\partial}{\partial t}g(t-\tau)\mathrm{d}\tau\\=\int\limits_{-\infty}^{+\infty}f(\tau)g'(t-\tau)\mathrm{d}\tau=f*g'=\int\limits_{-\infty}^{+\infty}f(\tau)g'(t-\tau)\mathrm{d}\tau=f*g'\end{gathered} \tag{7.1.469}$$

■ Q.E.D.

We will assume as obvious that these properties also apply to the discrete convolution that we will introduce now!

Typically, $y(t)$ is the output of a system characterized by its impulse response function $h(t)$ with input $x(t)$.

Definition (#7.1.282): The "discrete convolution" of two discrete signals $x[n]$ and $h[n]$ is defined as:

$$\boxed{y[n]=\sum_{m=-\infty}^{+\infty}x[n-m]\,h[m]=\sum_{m=-\infty}^{+\infty}h[n-m]\,x[m]} \tag{7.1.470}$$

If $h[m]$ is finite, e.g.:

$$h[m]=\begin{cases}h[m] & |m|\leq k\\ 0 & |m|>k\end{cases} \tag{7.1.471}$$

the convolution becomes:

$$y[n]=\sum_{m=-k}^{k}x[n-m]\,h[m] \tag{7.1.472}$$

If the system in question were a causal system in time domain:

$$h(n)=h[n]u[n],\quad \text{i.e.,}\quad h[n]=0\quad \text{if } n<0 \tag{7.1.473}$$

the above would become:

$$y[n]=\sum_{m=0}^{k}x[n-m]\,h[m] \tag{7.1.474}$$

This is also written:

$$\boxed{(f*g)[n]=\sum_{k=0}^{n}f[k]g[n-k]} \tag{7.1.475}$$

✍**Examples:**

E1. Consider:

$$f[n]=3\delta(n)+2\delta(n-1)+1\delta(n-2) \tag{7.1.476}$$

$$g[n]=2\delta(n)+1\delta(n-1)+3\delta(n-2) \tag{7.1.477}$$

We get visually (it's not obvious!):

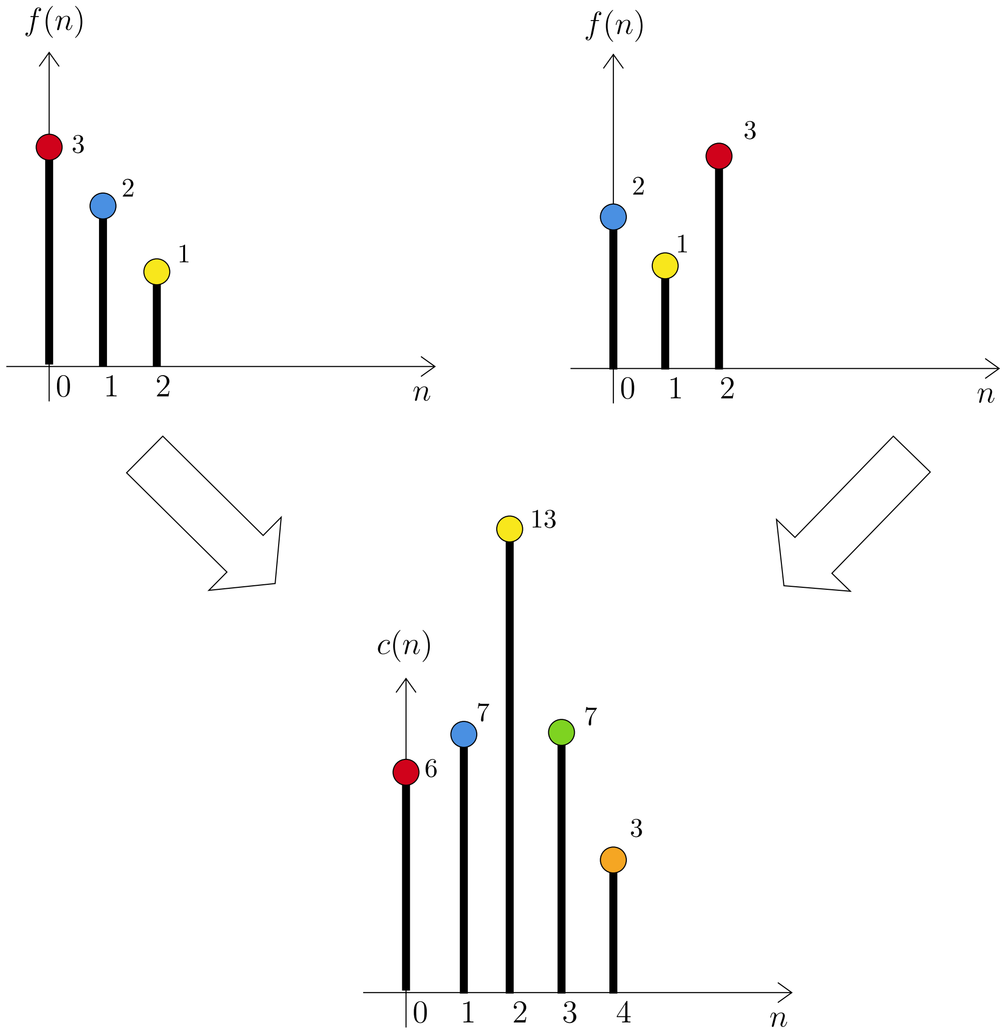


Figure 7.64: Discrete convolution example

The details steps are given by first considering that:

$$(f*g)[n]=c[n]=\sum_{m=0}^{2}g[n-m]f[m] \tag{7.1.478}$$

And we have for recall:

$$\begin{matrix} f(m=0)=3 & g(k=0)=2 \\ f(m=1)=2 & g(k=1)=1 \\ f(m=2)=1 & g(k=2)=3 \end{matrix} \tag{7.1.479}$$

And in tabular form:

n:	0	1	2	3	4
$m=0$	$2\cdot 3=6$	$1\cdot 3=3$	$3\cdot 3=9$	0	0
$m=1$	0	$2\cdot 2=4$	$1\cdot 2=2$	$3\cdot 2=6$	0
$m=2$	0	0	$2\cdot 1=2$	$1\cdot 1=1$	$3\cdot 1=3$
$c[n]$	6	7	13	7	3

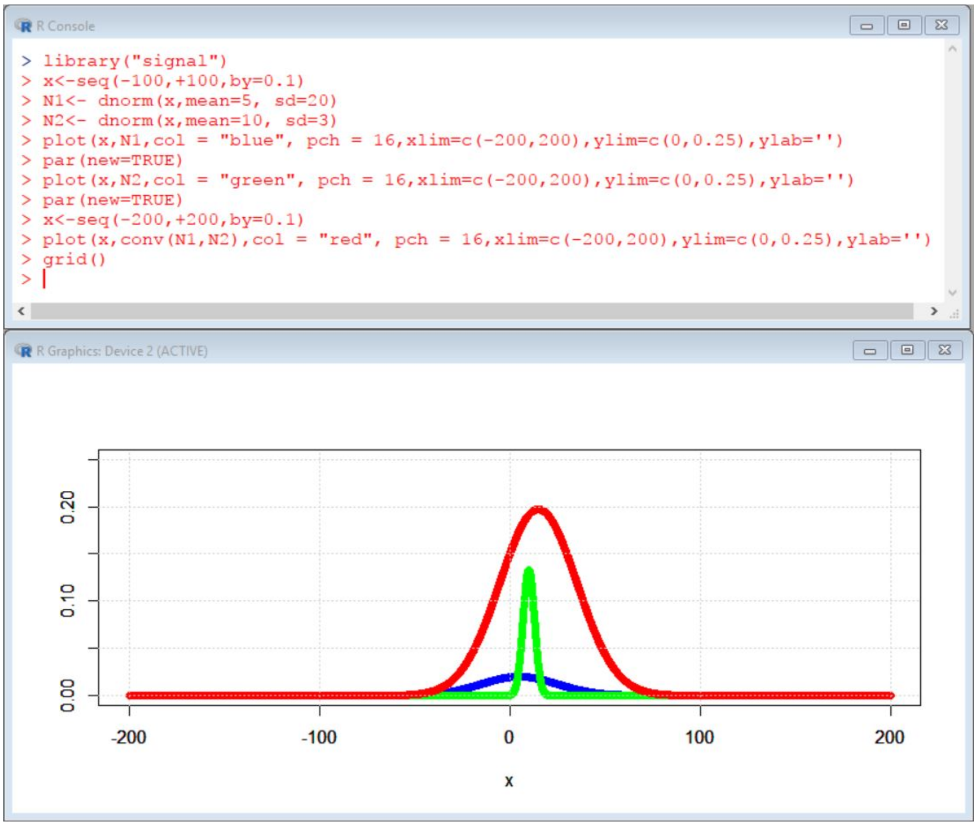
The $c[n>4]$ being all equal to zero.

The final result will then be written:

$$c[n]=6\delta(n)+7\delta(n-1)+13\delta(n-2)+7\delta(n-3)+3\delta(n-4) \tag{7.1.480}$$

Such a calculations can obviously be done very simply with softwares like R for example (see the corresponding companion book).

E2. A well known example is the convolution of two gaussians (see section Statistics page 709) that also result in a... gaussian. And obviously the discrete version of the convolution also give a gaussian as illustrated below with R (see the corresponding companion book for more details):



```
> library("signal")
> x<-seq(-100,+100,by=0.1)
> N1<- dnorm(x,mean=5, sd=20)
> N2<- dnorm(x,mean=10, sd=3)
> plot(x,N1,col = "blue", pch = 16,xlim=c(-200,200),ylim=c(0,0.25),ylab='')
> par(new=TRUE)
> plot(x,N2,col = "green", pch = 16,xlim=c(-200,200),ylim=c(0,0.25),ylab='')
> par(new=TRUE)
> x<-seq(-200,+200,by=0.1)
> plot(x,conv(N1,N2),col = "red", pch = 16,xlim=c(-200,200),ylim=c(0,0.25),ylab='')
> grid()
> |
```

where we have as the reader have noticed above, the following convolution $\mathcal{N}(5,20)*\mathcal{N}(10,3)$ and we see obviously that the result gives a gaussian of $\mathcal{N}(5+10=15,\sqrt{20^2+3^2}=15)$.

Let us now prove that the discrete convolution is also associative for the discrete convolution (as promised earlier!):

Proof 7.3.7.

$$\begin{aligned}((f*g)\star h)(n)&=\sum_{k=0}^{n}(f\star g)[k]h[n-k]\\&=\sum_{k=0}^{n}\sum_{l=0}^{k}f[l]g[k-l]h[n-k]\\&=\sum_{l=0}^{n}\sum_{k=l}^{n}f[l]g[k-l]h[n-k]\\&=\sum_{l=0}^{n}\sum_{k=0}^{n-l}f[l]g[k]h[n-k-l]\\&=\sum_{l=0}^{n}f[l](g*h)[n-l]\\&=(f*(g*h))(n)\end{aligned} \tag{7.1.481}$$

■ Q.E.D.

However, in image processing, we often consider convolution in spatial domain where causality does not apply.

If $h[m]=h[-m]$ is symmetric (almost always true in image processing), then replacing m by $-m$ we get:

$$y[n]=\sum_{m=-k}^{k}x[n+m]\,h[-m]=\sum_{m=-k}^{k}x[n+m]\,h[m] \tag{7.1.482}$$

We see now that the convolution is similar to the *correlation* of two functions!

If the input $x[m]$ is finite (always true in reality), i.e.:

$$x[m]=\begin{cases}x[m] & 0\leq m<N\\ 0 & \text{otherwise}\end{cases} \tag{7.1.483}$$

its index $n+m$ in the convolution has to satisfy the following for x to be in the valid non-zero range:

$$0\leq n+m\leq N-1 \tag{7.1.484}$$

or correspondingly, the index n of the output $y[n]$ has to satisfy:

$$-m\leq n\leq N-m-1 \tag{7.1.485}$$

When the variable index m in the convolution is equal to k, the index of output $y[n]$ reaches its lower bound $n=-k$; when $m=-k$, the index of $y[n]$ reaches its upper bound $n=N+k-1$. In other words, there are $N+2k$ valid (non-zero) elements in the output:

$$y[n],\quad(-k\leq n\leq N+k-1) \tag{7.1.486}$$

Assume the size of the input signal $x[n]$ is N $(n=0,\cdots,n=N-1)$ and the size of h is $M=2k+1$ (usually an odd number), then the size of the resulting convolution $y=x*h$ is $N+M-1=N+2k$. However, as it is usually desirable for the output y to have the same size as the input x, we can drop k components at each end of y. When the size of h is even,we can drop k components at one end and $k-1$ from the other of y.

7.1.4.2 Matrix Convolution

In image processing, all of the discussions above for one-dimensional convolution are generalized into two dimensions!

Definition(#7.1.283): "Matrix convolution" is the treatment of a matrix by another one which is named the "kernel". Most of the time, the convolution matrix filter uses a first matrix which is the image to be treated. The image is a bi-dimensional collection of pixels in rectangular coordinates. The used kernel h of always odd $k\times k$ dimensions depends on the effect you want!

By construction, we have:

$$\boxed{C[m,n]=\text{Image}*\text{Kernel}=f*h[m,n]=\sum_{i=-\frac{k-1}{2}}^{+\frac{k-1}{2}}\sum_{j=-\frac{k-1}{2}}^{+\frac{k-1}{2}} h\left(i+\frac{k+1}{2},j+\frac{k+1}{2}\right)f(m+i,n+j)} \tag{7.1.487}$$

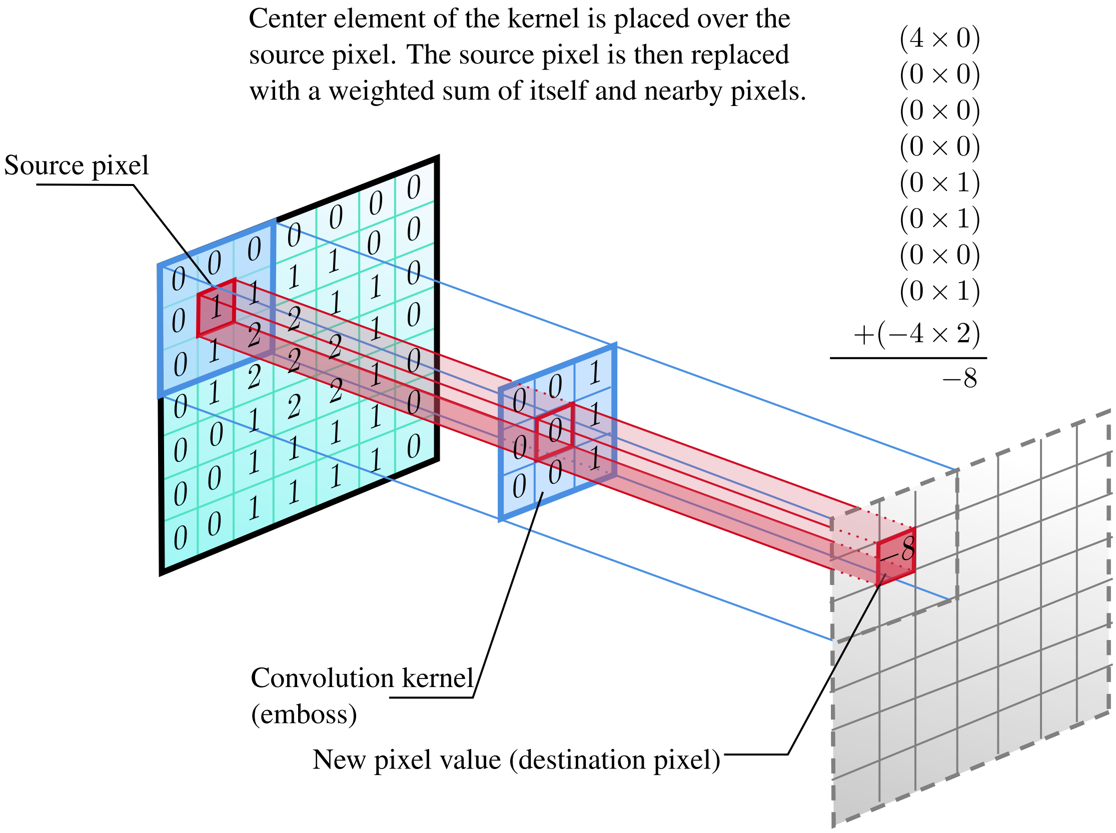
For example, if we have the following matrices:

$$f=\begin{bmatrix}0&0&0&0&0&0&0\\0&1&1&1&1&0&0\\0&1&2&2&1&1&0\\0&1&2&2&2&1&0\\0&0&1&2&2&1&0\\0&0&1&1&1&1&0\\0&0&1&1&1&1&0\end{bmatrix}\qquad h=\begin{bmatrix}-4&0&0\\0&0&0\\0&0&4\end{bmatrix} \tag{7.1.488}$$

Then (don't forget that the columns and row of the kernel matrix are flipped!):

$$\begin{aligned}C[2,2]=&\begin{bmatrix}0&0&0&0&0&0&0\\0&1&1&1&1&0&0\\0&1&2&2&1&1&0\\0&1&2&2&2&1&0\\0&0&1&2&2&1&0\\0&0&1&1&1&1&0\\0&0&1&1&1&1&0\end{bmatrix}*h[2,2]=\sum_{i=-1}^{+1}\sum_{j=-1}^{+1}h(i+2,j+2)f(2+i,2+j)\\&=f(1,1)h(1,1)+f(1,2)h(1,2)+f(1,3)h(1,3)+f(2,1)h(2,1)+f(2,2)h(2,2)\\&+f(2,3)h(2,3)+f(1,3)h(1,3)+f(2,3)h(2,3)+f(3,3)h(3,3)\\&=0\cdot 4+0\cdot 0+0\cdot 0+0\cdot 0+1\cdot 0+1\cdot 0+0\cdot 0+1\cdot 0+2\cdot(-4)=-8\end{aligned} \tag{7.1.489}$$

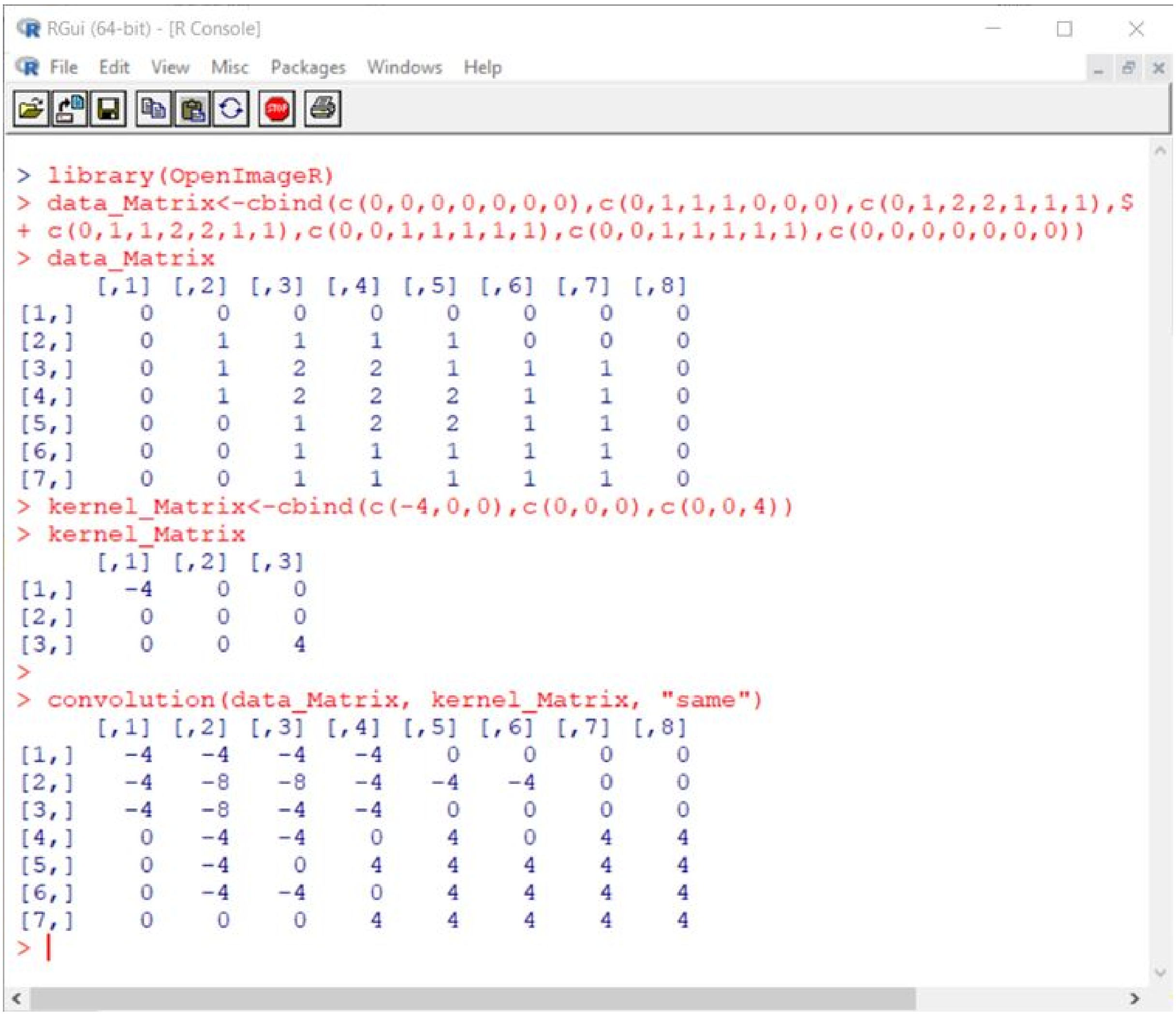
That can be illustrated as following (h has already been double-flipped in the figure below):


Center element of the kernel is placed over the source pixel. The source pixel is then replaced with a weighted sum of itself and nearby pixels.

(4 × 0)
(0 × 0)
(0 × 0)
(0 × 0)
(0 × 1)
(0 × 1)
(0 × 0)
(0 × 1)
+(−4 × 2)
−8
Source pixel
Convolution kernel (emboss)
New pixel value (destination pixel)

Figure 7.65: **Matrix Convolution**

The full example can be run and reproduced with a free software like `R`:


RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help

```
> library(OpenImageR)
> data_Matrix<-cbind(c(0,0,0,0,0,0,0),c(0,1,1,1,0,0,0),c(0,1,2,2,1,1,1),$
+ c(0,1,1,2,2,1,1),c(0,0,1,1,1,1,1),c(0,0,1,1,1,1,1),c(0,0,0,0,0,0,0))
> data_Matrix
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
[1,]    0    0    0    0    0    0    0    0
[2,]    0    1    1    1    1    0    0    0
[3,]    0    1    2    2    1    1    1    0
[4,]    0    1    2    2    2    1    1    0
[5,]    0    0    1    2    2    1    1    0
[6,]    0    0    1    1    1    1    1    0
[7,]    0    0    1    1    1    1    1    0
> kernel_Matrix<-cbind(c(-4,0,0),c(0,0,0),c(0,0,4))
> kernel_Matrix
     [,1] [,2] [,3]
[1,]   -4    0    0
[2,]    0    0    0
[3,]    0    0    4
>
> convolution(data_Matrix, kernel_Matrix, "same")
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
[1,]   -4   -4   -4   -4    0    0    0    0
[2,]   -4   -8   -8   -4   -4   -4    0    0
[3,]   -4   -8   -4   -4    0    0    0    0
[4,]    0   -4   -4    0    4    0    4    4
[5,]    0   -4    0    4    4    4    4    4
[6,]    0   -4   -4    0    4    4    4    4
[7,]    0    0    0    4    4    4    4    4
> 
```

Or for people who may prefer a slow implementation in C++:

Listing 7.1: **C++ for matrix convolution**

```cpp
// find center position of kernel (half of kernel size)
kCenterX = kCols / 2;
kCenterY = kRows / 2;

for(i=0; i < rows; ++i)              // rows
{
    for(j=0; j < cols; ++j)          // columns
    {
        for(m=0; m < kRows; ++m)     // kernel rows
        {
            mm = kRows - 1 - m;      // row index of flipped kernel

            for(n=0; n < kCols; ++n) // kernel columns
            {
                nn = kCols - 1 - n;  // column index of flipped kernel

                // index of input signal, used for checking
```

```
                   boundary
                ii = i + (kCenterY - mm);
                jj = j + (kCenterX - nn);

                // ignore input samples which are out of bound
                if( ii >= 0 && ii < rows && jj >= 0 && jj < cols )
                    out[i][j] += in[ii][jj] * kernel[mm][nn];
            }
        }
    }
}
```

7.1.5 Integral Transforms

An "integral transform" is an operator that maps functions from one space to another. Formally:

$$\boxed{\mathcal{T}(f(x))=\int\limits_{-\infty}^{+\infty}K(x,y)f(y)\mathrm{d}y} \tag{7.1.490}$$

Now the practical motivation for an integral transform is to reduce the complexity of the problem i.e the mathematical operations will be much easier to handle in the image space (typically the resolution of differential equations!).

However, as much as it is fun to do work in the image space, one has to be able to interpret the results in the original space. To do so requires the study of the operator K. Usually one knows a priori the nature of the function f by the nature of the problem one is dealing. Hence the study of integral transforms is the study of the operator $\mathcal{T}$. Two properties come very easily:

$$\mathcal{T}(f+g)=\mathcal{T}(f)+\mathcal{T}(g) \tag{7.1.491}$$

$$\mathcal{T}(cf)=c\mathcal{T}(f) \tag{7.1.492}$$

To ensure invertibility, one has to show that the kernel space only contains the null function.

The Fourier and Laplace transforms are for example continuous (integral) transforms of continuous functions (even if there exist a discrete version of the Fourier transform!).

The Laplace transform maps a function $f(t)$ to a function $\mathcal{F}(s)$ of the complex variable s, where:

$$s=\sigma+\mathrm{i}\omega \tag{7.1.493}$$

Since the derivative:

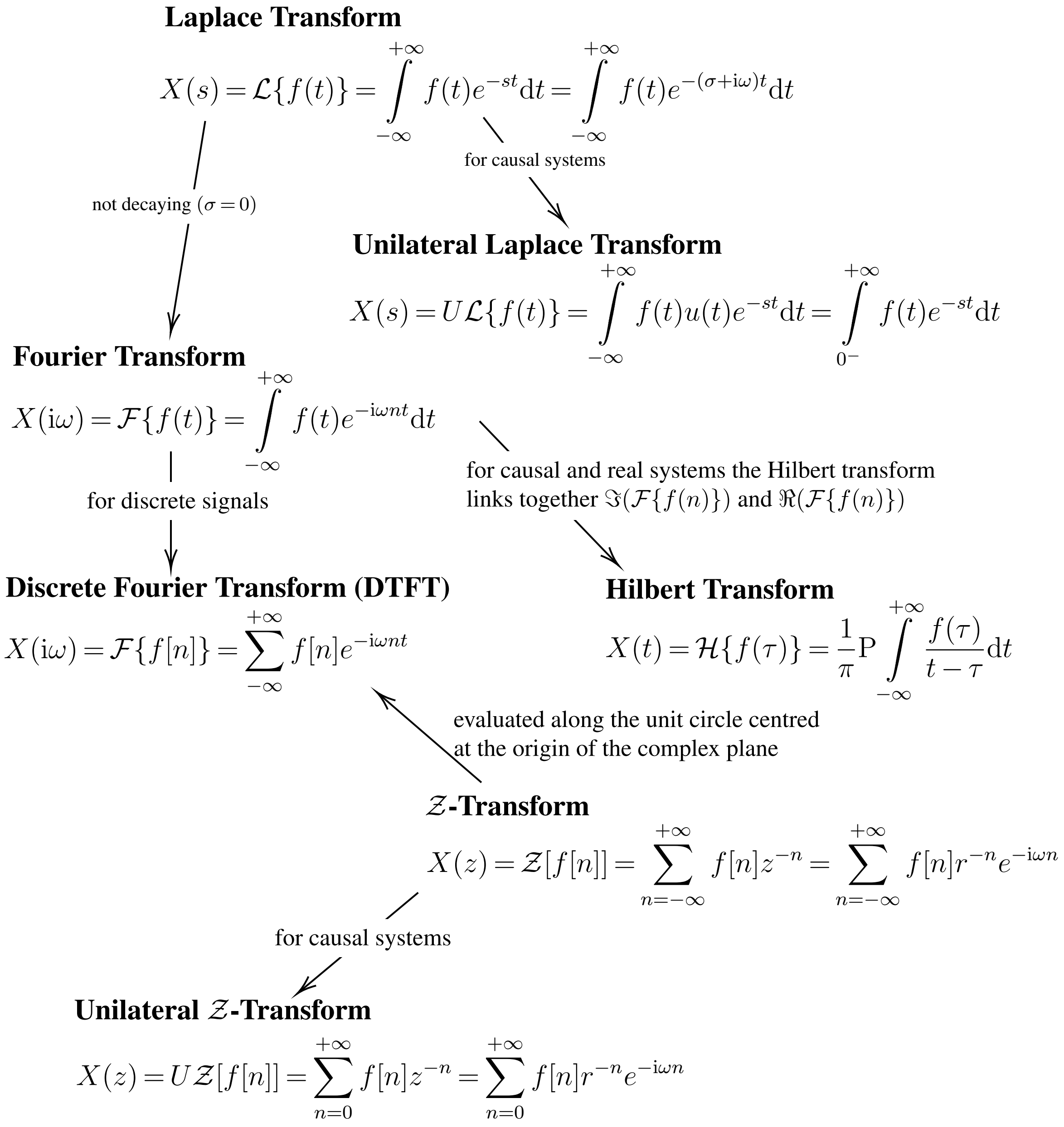
$$f'(t)=\frac{\mathrm{d}f(t)}{\mathrm{d}t} \tag{7.1.494}$$

maps to $s\mathcal{F}(s)$, the Laplace transform $\mathcal{L}$ of a linear differential equation is an algebraic equation. Thus, the Laplace transform is useful for, among other things, solving linear differential equations.

If we set the real part of the complex variable s to zero, $\sigma=0$, the result is the Fourier transform $\mathcal{F}(\mathrm{i}\omega)$ which is essentially the frequency domain representation of $f(t)$ (note that this is true only if for that value of σ the formula to obtain the Laplace transform of $f(t)$ exists, i.e., it does not go to infinity).

The $\mathcal{Z}$-transform is essentially a discrete version of the Laplace transform and, thus, can be useful in solving difference equations, the discrete version of differential equations. The $\mathcal{Z}$-transform maps a sequence $f[n]$ to a continuous function $F(z)$ of the complex variable $z=re^{\mathrm{i}\Omega}$.

If we set the magnitude of z to unity, $r=1$, the result is the Discrete Time Fourier Transform (DTFT) $\mathcal{F}(\mathrm{i}\Omega)$ which is essentially the frequency domain representation of $f[n]$.



Laplace Transform

$$X(s)=\mathcal{L}\{f(t)\}=\int\limits_{-\infty}^{+\infty} f(t)e^{-st}\mathrm{d}t=\int\limits_{-\infty}^{+\infty} f(t)e^{-(\sigma+\mathrm{i}\omega)t}\mathrm{d}t$$

Unilateral Laplace Transform

$$X(s)=U\mathcal{L}\{f(t)\}=\int\limits_{-\infty}^{+\infty} f(t)u(t)e^{-st}\mathrm{d}t=\int\limits_{0^-}^{+\infty} f(t)e^{-st}\mathrm{d}t$$

Fourier Transform

$$X(\mathrm{i}\omega)=\mathcal{F}\{f(t)\}=\int\limits_{-\infty}^{+\infty} f(t)e^{-\mathrm{i}\omega nt}\mathrm{d}t$$

Discrete Fourier Transform (DTFT)

$$X(\mathrm{i}\omega)=\mathcal{F}\{f[n]\}=\sum_{-\infty}^{+\infty} f[n]e^{-\mathrm{i}\omega nt}$$

Hilbert Transform

$$X(t)=\mathcal{H}\{f(\tau)\}=\frac{1}{\pi}\mathrm{P}\int\limits_{-\infty}^{+\infty}\frac{f(\tau)}{t-\tau}\mathrm{d}t$$

$\mathcal{Z}$-Transform

$$X(z)=\mathcal{Z}[f[n]]=\sum_{n=-\infty}^{+\infty} f[n]z^{-n}=\sum_{n=-\infty}^{+\infty} f[n]r^{-n}e^{-\mathrm{i}\omega n}$$

Unilateral $\mathcal{Z}$-Transform

$$X(z)=U\mathcal{Z}[f[n]]=\sum_{n=0}^{+\infty} f[n]z^{-n}=\sum_{n=0}^{+\infty} f[n]r^{-n}e^{-\mathrm{i}\omega n}$$

for causal systems
not decaying ($\sigma=0$)
for discrete signals
for causal and real systems the Hilbert transform links together $\Im(\mathcal{F}\{f(n)\})$ and $\Re(\mathcal{F}\{f(n)\})$
evaluated along the unit circle centred at the origin of the complex plane
for causal systems

Figure 7.66: **Most common types of integral transforms with harmonized notation**

Remarks

R1. The three integral transformations mentioned above (Fourier, Laplace, $\mathcal{Z}$) are only a sample of what exists in practice. Let us also mention the Hartley transform, the Mellin transform, the Weierstrass transform, the Hankel transform (Fourier-Bessel), the Abel transform, the Hilbert transform, the Gauss-Weierstrass transform, etc.

R2. For information we will calculate cases and prove properties of the Fourier, Laplace, $\mathcal{Z}$ and Hilbert transforms only used in concrete applications in the industry and mainly useful in other chapters of this textbook! Indeed a rather general presentation would require a few hundred pages and it is not the objective of this textbook, as you already know it, to prove mathematical properties not associated with concrete cases.

7.1.5.1 Fourier Transform

So we have already introduced Fourier Transforms in the section of Sequences and Series (page 1758), we will come back here more in details on this topic in (we wish) are more structured way...

7.1.5.1.1 Continuous Time Fourier Transform

The Fourier expansion coefficient $X[k]$ of a continuous periodic signal $x_T(t)=x_T(t+T)$ is:

$$X[k]=\frac{1}{T}\int_T x_T(t)e^{-\mathrm{i}k\omega_0 t}\mathrm{d}t \quad (k=0,\pm 1,\pm 2,\cdots) \tag{7.1.495}$$

and the Fourier expansion of the signal is:

$$x_T(t)=\sum_{k=-\infty}^{+\infty} X[k]e^{\mathrm{i}k\omega_0 t} \tag{7.1.496}$$

which can also be written as:

$$x_T(t)=\frac{1}{T}\sum_{k=-\infty}^{+\infty}(TX[k])e^{\mathrm{i}k\omega_0 t}=\frac{\omega_0}{2\pi}\sum_{k=-\infty}^{+\infty}X(k\omega_0)e^{\mathrm{i}k\omega_0 t} \quad (a) \tag{7.1.497}$$

where $X(k\omega_0)$ is defined as:

$$X(k\omega_0)\triangleq T\,X[k]=\int_T x_T(t)e^{-\mathrm{i}k\omega_0 t}\mathrm{d}t \quad (b) \tag{7.1.498}$$

When the period of $x_T(t)$ approaches infinity $T\rightarrow+\infty$, the periodic signal $x_T(t)$ becomes a non-periodic signal $x(t)$ and the following will result:

- Interval between two neighbouring frequency components becomes zero:

$$T\rightarrow+\infty\Longrightarrow\omega_0=\frac{2\pi}{T}\rightarrow 0 \tag{7.1.499}$$

- Discrete frequency becomes continuous frequency:

$$k\omega_0|_{\omega_0\rightarrow 0}\Longrightarrow\omega \tag{7.1.500}$$

- Summation of the Fourier expansion becomes an integral:

$$x(t)\triangleq\lim_{T\rightarrow+\infty}x_T(t)=\lim_{\omega_0\rightarrow 0}\frac{1}{2\pi}\sum_{k=-\infty}^{+\infty}X(k\omega_0)e^{\mathrm{i}k\omega_0 t}\omega_0=\frac{1}{2\pi}\int_{-\infty}^{+\infty}X(\omega)e^{\mathrm{i}\omega t}\mathrm{d}\omega \tag{7.1.501}$$

  the second equal sign is due to the general fact:

$$\lim_{\Delta x\rightarrow 0}\sum_{k=-\infty}^{+\infty}f(k\Delta x)\Delta x=\int_{-\infty}^{+\infty}f(x)\mathrm{d}x \tag{7.1.502}$$

- Time integral over T becomes over the entire time axis:

$$X(\omega)\stackrel{\triangle}{=}\lim_{T\to+\infty}X(k\omega_0)=\lim_{T\to+\infty}\int\limits_T x_T(t)e^{-\mathrm{i}k\omega_0 t}\mathrm{d}t=\int\limits_{-\infty}^{+\infty}x(t)e^{-\mathrm{i}\omega t}\mathrm{d}t \qquad (7.1.503)$$

In summary, when the signal is non-periodic $x(t)=\lim_{T\to+\infty}x_T(t)$, the Fourier expansion becomes Fourier transform. The forward transform (analysis) is:

$$X(\omega)=\int\limits_{-\infty}^{+\infty}x(t)e^{-\mathrm{i}\omega t}\mathrm{d}t \quad \text{or} \quad X(f)=\int\limits_{-\infty}^{+\infty}x(t)e^{-\mathrm{i}2\pi f t}\mathrm{d}t \qquad (7.1.504)$$

and the inverse transform (synthesis) is:

$$x(t)=\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty}X(\omega)e^{\mathrm{i}\omega t}\mathrm{d}\omega=\int\limits_{-\infty}^{+\infty}X(f)e^{\mathrm{i}2\pi f t}\mathrm{d}f \qquad (7.1.505)$$

Comparing Fourier coefficient of a periodic signal $x_T(t)$ with Fourier spectrum of a non-periodic signal $x(t)$:

$$X[k]=\frac{1}{T}\int\limits_T x_T(t)e^{-\mathrm{i}k\omega_0 t}\mathrm{d}t,\quad X(\omega)=\int\limits_{-\infty}^{+\infty}x(t)e^{-\mathrm{i}\omega t}\mathrm{d}t \qquad (7.1.506)$$

The spectrum of a time signal can be denoted by $X(\omega)$ or $X(f)$ to emphasize the fact that the spectrum represents how the energy contained in the signal is distributed as a function of frequency ω or f. Moreover, if $X(f)$ is used, the factor $1/2\pi$ in front of the inverse transform is dropped so that the transform pair takes a more symmetric form. On the other hand, as Fourier transform can be considered as a special case of Laplace transform when the real part σ of the complex argument $s=\sigma+\mathrm{i}\omega=\mathrm{i}\omega$ is zero:

$$X(s)|_{s=\mathrm{i}\omega}=\int\limits_{-\infty}^{+\infty}x(t)e^{-st}\mathrm{d}t|_{s=\mathrm{i}\omega}=\int\limits_{-\infty}^{+\infty}x(t)e^{-\mathrm{i}\omega t}\mathrm{d}t=X(\mathrm{i}\omega) \qquad (7.1.507)$$

It is also natural to denote the spectrum of $x(t)$ by $X(\mathrm{i}\omega)$.

Ok this done let us see now some important example for physics (especially quantum physics) and signal processing!

✍**Examples:**

E1. Consider the unit impulse function (Dirac function):

$$x(t-a)=\delta(t-a) \qquad (7.1.508)$$

Therefore:

$$X(\mathrm{i}\omega)=\int_{-\infty}^{+\infty}\delta(t-a)e^{-\mathrm{i}\omega t}\mathrm{d}t=e^{-\mathrm{i}\omega a} \tag{7.1.509}$$

and if $a=0$ we have then:

$$X(\mathrm{i}\omega)=\int_{-\infty}^{+\infty}\delta(t)e^{-\mathrm{i}\omega t}\mathrm{d}t=1 \tag{7.1.510}$$

E2. If the spectrum of a signal $x(t)$ is a delta function in frequency domain $X(\mathrm{i}\omega)=2\pi\,\delta(\omega)$, the signal can be found to be:

$$x(t)=\mathcal{F}^{-1}(X(\mathrm{i}\omega))=\frac{1}{2\pi}\int_{-\infty}^{+\infty}2\pi\,\delta(\omega)e^{\mathrm{i}\omega t}\mathrm{d}\omega=e^{0}=1 \tag{7.1.511}$$

i.e.:

$$\mathcal{F}(x(t))=\int_{-\infty}^{+\infty}e^{-\mathrm{i}\omega t}\mathrm{d}t=2\pi\,\delta(\omega) \tag{7.1.512}$$

E3. We consider:

$$x(t)=\begin{cases}1 & |t|<a\\ 0 & \text{else}\end{cases} \tag{7.1.513}$$

The spectrum is:

$$X(\mathrm{i}\omega)=\int_{-a}^{a}e^{-\mathrm{i}\omega t}\mathrm{d}t=\frac{1}{-\mathrm{i}\omega}e^{-\mathrm{i}\omega t}|_{-a}^{a}=\frac{2}{\omega}\sin(a\omega) \tag{7.1.514}$$

This is the sinc function with a parameter a.

Note that the height of the main peak is $2a$ and it gets taller and narrower as a gets larger.

Also note:

$$\int_{-\infty}^{+\infty}X(\mathrm{i}\omega)\mathrm{d}\omega=2\int_{-\infty}^{+\infty}\frac{\sin(a\omega)}{\omega}\mathrm{d}\omega=2\pi \tag{7.1.515}$$

When a approaches infinity, $x(t)=1$ for all t, and the spectrum becomes:

$$X(\mathrm{i}\omega)=\int_{-\infty}^{+\infty}e^{-\mathrm{i}\omega t}\mathrm{d}t=\lim_{a\to+\infty}\left[\frac{2}{\omega}\sin(a\omega)\right]=2\pi\delta(\omega)=\delta(f) \tag{7.1.516}$$

Recall that the Fourier coefficient of $x(t)=1$ is:

$$X[k]=\delta[k]=\begin{cases}1 & k=0\\ 0 & \text{else}\end{cases} \tag{7.1.517}$$

which represents the energy contained in the signal at $k=0$ (DC component at zero frequency), and the spectrum $X(\mathrm{i}\omega)=X[k]/\omega$ is the energy density or distribution which is infinity at zero frequency.

The integral in the above transform is an important formula to be used frequently later:

$$\int\limits_{-\infty}^{+\infty} e^{-\mathrm{i}\omega t}\mathrm{d}t=2\pi\delta(\omega)\quad\text{or}\quad\int\limits_{-\infty}^{+\infty} e^{-\mathrm{i}2\pi ft}\mathrm{d}t=\delta(f) \tag{7.1.518}$$

which can also be written as:

$$\int\limits_{-\infty}^{+\infty} e^{-\mathrm{i}\omega t}\mathrm{d}t=\int\limits_{-\infty}^{+\infty}[\cos(\omega t)+\mathrm{i}\,\sin(-\mathrm{i}\omega t)]\mathrm{d}t=\int\limits_{-\infty}^{+\infty}\cos(\omega t)\mathrm{d}t=\delta(f) \tag{7.1.519}$$

Switching t and f in the equation above, we also have:

$$\int\limits_{-\infty}^{+\infty} e^{-\mathrm{i}2\pi ft}\mathrm{d}f=\int\limits_{-\infty}^{+\infty}\cos(2\pi ft)\mathrm{d}f=\delta(t) \tag{7.1.520}$$

representing a superposition of an infinite number of cosine functions of all frequencies, which cancel each other any where along the time axis except at $t=0$ where they add up to infinity, an impulse.

E3. Let us now consider:

$$x(t)=\cos(\omega_0 t)=\frac{1}{2}[e^{\mathrm{i}\omega_0 t}+e^{-\mathrm{i}\omega_0 t}] \tag{7.1.521}$$

The spectrum of the cosine function is:

$$\begin{aligned}X(\mathrm{i}\omega)&=\int\limits_{-\infty}^{+\infty}x(t)e^{-\mathrm{i}\omega t}\mathrm{d}t=\frac{1}{2}\int\limits_{-\infty}^{+\infty}[e^{\mathrm{i}\omega_0 t}+e^{-\mathrm{i}\omega_0 t}]e^{-\mathrm{i}\omega t}\mathrm{d}t\\&=\pi[\delta(\omega-\omega_0)+\delta(\omega+\omega_0)]\end{aligned} \tag{7.1.522}$$

The spectrum of the sine function:

$$x(t)=\sin(\omega_0 t)=\frac{1}{2\mathrm{i}}[e^{\mathrm{i}\omega_0 t}-e^{-\mathrm{i}\omega_0 t}] \tag{7.1.523}$$

can be similarly obtained to be:

$$X(\mathrm{i}\omega)=\frac{\pi}{\mathrm{i}}[\delta(\omega-\omega_0)-\delta(\omega+\omega_0)]=-\mathrm{i}\pi[\delta(\omega-\omega_0)-\delta(\omega+\omega_0)] \tag{7.1.524}$$

Again, these spectra represent the energy density distribution of the sinusoids, while the corresponding Fourier coefficients:

$$X[k]=\mathcal{F}(\cos(\omega_0 t))=\frac{1}{2}[\delta[k-1]+\delta[k+1]] \tag{7.1.525}$$

and:

$$X[k]=\mathcal{F}(\sin(\omega_0 t))=\frac{1}{2\mathrm{i}}[\delta[k-1]-\delta[k+1]] \tag{7.1.526}$$

represent the energy contained at frequency $\omega=\omega_0$.

7.1.5.1.1.1 Properties of Fourier Transform

The properties of the Fourier transform are summarized below. For some of them already proved in the section of Sequences and Series we will not give the proof. But for others we will do it again!

In the following, we assume $\mathcal{F}(x(t))=X(\mathrm{i}\omega)$ and $\mathcal{F}(y(t))=Y(\mathrm{i}\omega)$.

P1. Linearity:

$$\mathcal{F}(ax(t)+by(t))=a\mathcal{F}(x(t))+b\mathcal{F}(y(t)) \tag{7.1.527}$$

P2. Time shift:

$$\mathcal{F}(x(t\pm t_0))=X(\mathrm{i}\omega)e^{\pm\mathrm{i}\omega t_0} \tag{7.1.528}$$

Proof 7.3.8. Let $t'=t\pm t_0$, i.e., $t=t'\mp t_0$, we have:

$$\begin{aligned}\mathcal{F}(x(t\pm t_0)) &= \int_{-\infty}^{+\infty} x(t\pm t_0)e^{-\mathrm{i}\omega t}\mathrm{d}t=\int_{-\infty}^{+\infty} x(t')e^{-\mathrm{i}\omega(t'\mp t_0)}\mathrm{d}t'\\ &= e^{\pm\mathrm{i}\omega t_0}\int_{-\infty}^{+\infty} x(t')e^{-\mathrm{i}\omega t'}\mathrm{d}t'=X(\mathrm{i}\omega)e^{\pm\mathrm{i}\omega t_0}\end{aligned}$$

■ Q.E.D.

P3. Frequency shift:

$$\mathcal{F}^{-1}\left(X(\mathrm{i}\omega\pm\omega_0)\right)=x(t)e^{\mp\mathrm{i}\omega_0 t} \tag{7.1.529}$$

Proof 7.3.9. Let $\omega'=\omega\pm\omega_0$, i.e., $\omega=\omega'\mp\omega_0$, we have:

$$\begin{aligned}\mathcal{F}^{-1}[X(\mathrm{i}(\omega\pm\omega_0))]&=\frac{1}{2\pi}\int_{-\infty}^{+\infty} X(\mathrm{i}(\omega\pm\omega_0))e^{\mathrm{i}\omega t}\mathrm{d}\omega=\frac{1}{2\pi}\int_{-\infty}^{+\infty} X(\mathrm{i}\omega')e^{\mathrm{i}\omega(\omega'\mp\omega_0)}\mathrm{d}\omega'\\ &=e^{\mp\mathrm{i}\omega_0 t}\frac{1}{2\pi}\int_{-\infty}^{+\infty} X(\mathrm{i}\omega')e^{\mathrm{i}\omega'}\mathrm{d}\omega'=x(t)e^{\mp\mathrm{i}\omega_0 t}\end{aligned} \tag{7.1.530}$$

■ Q.E.D.

P4. Time reversal:

$$\mathcal{F}(x(-t))=X(-\omega) \tag{7.1.531}$$

Proof 7.3.10.

$$\mathcal{F}(x(-t))=\int\limits_{-\infty}^{+\infty} x(-t)e^{-\mathrm{i}\omega t}\mathrm{d}t \tag{7.1.532}$$

Replacing t by $-t'$, we get:

$$\mathcal{F}(x(-t))=-\int\limits_{+\infty}^{-\infty} x(t')e^{\mathrm{i}\omega t'}\mathrm{d}t'=\int\limits_{-\infty}^{+\infty} x(t')e^{\mathrm{i}\omega t'}\mathrm{d}t'=X(-\omega) \tag{7.1.533}$$

■ Q.E.D.

P5. Even and Odd Signals and Spectra:

If the signal $x(t)$ is an even (or odd) function of time, its spectrum $X(\mathrm{i}\omega)$ is an even (or odd) function of frequency:

$$\text{if } x(t)=x(-t) \text{ then } X(\mathrm{i}\omega)=X(-\mathrm{i}\omega) \tag{7.1.534}$$

and:

$$\text{if } x(t)=-x(-t) \text{ then } X(\mathrm{i}\omega)=-X(-\mathrm{i}\omega) \tag{7.1.535}$$

Proof 7.3.11. If $x(t)=x(-t)$ is even, then according to the time reversal property, we have:

$$X(\mathrm{i}\omega)=\mathcal{F}(x(t))=\mathcal{F}(x(-t))=X(-\omega) \tag{7.1.536}$$

i.e., the spectrum $X(\mathrm{i}\omega)=X(-\omega)$ is also even. Similarly, if $x(t)=-x(-t)$ is odd, we have:

$$X(\mathrm{i}\omega)=\mathcal{F}(x(t))=\mathcal{F}(-x(-t))=-X(-\omega) \tag{7.1.537}$$

i.e., the spectrum $X(\mathrm{i}\omega)=-X(-\omega)$ is also odd.

■ Q.E.D.

P6. Time and frequency scaling:

$$\mathcal{F}(x(at))=\frac{1}{a}X\left(\frac{\omega}{a}\right) \quad \text{or} \quad \mathcal{F}(ax(at))=X\left(\frac{\omega}{a}\right) \tag{7.1.538}$$

Proof 7.3.12. Let $u = at$, i.e., $t = u/a$, where $a > 0$ is a scaling factor, we have:

$$\mathcal{F}(x(at)) = \int_{-\infty}^{+\infty} x(at)e^{-\mathrm{i}\omega t}\mathrm{d}t = \int_{-\infty}^{+\infty} x(u)e^{-\mathrm{i}\omega u/a}\mathrm{d}(u/a) = \frac{1}{a}X\left(\frac{\omega}{a}\right) \tag{7.1.539}$$

Note that when $a < 1$, time function $x(at)$ is stretched, and $X(\mathrm{i}\omega/a)$ is compressed; when $a > 1$, $x(at)$ is compressed and $X(\mathrm{i}\omega/a)$ is stretched. This is a general feature of Fourier transform, i.e., compressing one of the $x(t)$ and $X(\mathrm{i}\omega)$ will stretch the other and vice versa. In particular, when $a \to 0$, $x(at)$ is stretched to approach a constant, and $X(\mathrm{i}\omega/a)/a$ is compressed with its value increased to approach an impulse; on the other hand, when $a \to +\infty$, $ax(at)$ is compressed with its value increased to approach an impulse and $X(\mathrm{i}\omega/a)$ is stretched to approach a constant.

■ Q.E.D.

P7. Complex Conjugation:

$$\text{if } \mathcal{F}(x(t)) = X(\mathrm{i}\omega), \quad \text{then } \mathcal{F}(x^*(t)) = X^*(-\mathrm{i}\omega) \tag{7.1.540}$$

Proof 7.3.13. Taking the complex conjugate of the inverse Fourier transform, we get:

$$x^*(t) = \left[\frac{1}{2\pi}\int_{-\infty}^{+\infty} X(\mathrm{i}\omega)e^{\mathrm{i}\omega t}\mathrm{d}\omega\right]^* = \frac{1}{2\pi}\int_{-\infty}^{+\infty} X^*(\mathrm{i}\omega)e^{-\mathrm{i}\omega t}\mathrm{d}\omega \tag{7.1.541}$$

Replacing ω by $-\omega'$ we get the desired result:

$$x^*(t) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} X^*(-\omega')e^{\mathrm{i}\omega' t}\mathrm{d}\omega' = \mathcal{F}^{-1}\left(X^*(-\omega)\right) \tag{7.1.542}$$

■ Q.E.D.

We further consider two special cases:

- If $x(t) = x^*(t)$ is real, then:

  $$\begin{aligned}\mathcal{F}(x(t)) &= X(\mathrm{i}\omega) = X_r(\mathrm{i}\omega) + \mathrm{i}X_i(\mathrm{i}\omega)\\ &= \mathcal{F}(x^*(t)) = X^*(-\omega) = X_r(-\omega) - \mathrm{i}X_i(-\omega)\end{aligned} \tag{7.1.543}$$

  i.e., the real part of the spectrum is even (with respect to frequency ω), and the imaginary part is odd:

  $$\begin{cases} X_r(\mathrm{i}\omega) = X_r(-\mathrm{i}\omega)\\ X_i(\mathrm{i}\omega) = -X_i(-\mathrm{i}\omega)\end{cases} \tag{7.1.544}$$

- If $x(t) = -x^*(t)$ is imaginary, then:

  $$\begin{aligned}\mathcal{F}(x(t)) &= X(\mathrm{i}\omega) = X_r(\mathrm{i}\omega) + \mathrm{i}X_i(\mathrm{i}\omega)\\ &= \mathcal{F}(-x^*(t)) = -X^*(-\mathrm{i}\omega) = -X_r(-\mathrm{i}\omega) + \mathrm{i}X_i(-\mathrm{i}\omega)\end{aligned} \tag{7.1.545}$$

  i.e., the real part of the spectrum is odd, and the imaginary part is even:

  $$\begin{cases} X_r(\mathrm{i}\omega) = -X_r(-\mathrm{i}\omega)\\ X_i(\mathrm{i}\omega) = X_i(-\mathrm{i}\omega)\end{cases} \tag{7.1.546}$$

If the time signal $x(t)$ is one of the four combinations shown in the table (real even, real odd, imaginary even, and imaginary odd), then its spectrum $X(\mathrm{i}\omega)$ is given in the corresponding table entry:

	if $x(t)$ is real X_r even, X_i odd	if $x(t)$ is imaginary X_r odd, X_i even
if $x(t)$ is Even X_r and X_i even	$X_i = 0$, $X = X_r$ even	$X_r = 0$, $X = X_i$ even
if $x(t)$ is Odd X_r and X_i odd	$X_r = 0$, $X = X_i$ odd	$X_i = 0$, $X = X_r$ odd

Note that if a real or imaginary part in the table is required to be both even and odd at the same time, it has to be zero.

These properties are summarized below:

	$x(t) = x_r(t) + \mathrm{i}x_i(t)$	$X(\mathrm{i}\omega) = X_r(\mathrm{i}\omega) + \mathrm{i}X_i(\mathrm{i}\omega)$
1	real $x(t) = x_r(t)$	even $X_r(\mathrm{i}\omega)$, odd $X_i(\mathrm{i}\omega)$
2	real and even $x(-t) = x_r(t)$	real and even $X_r(\mathrm{i}\omega)$
3	real and odd $x(-t) = -x_r(t)$	imaginary and odd $X_i(\mathrm{i}\omega)$
4	imaginary $x(t) = x_i(t)$	odd $X_r(\mathrm{i}\omega)$, even $X_i(\mathrm{i}\omega)$
5	imaginary and even $x(-t) = x_i(t)$	imaginary and even $X_i(\mathrm{i}\omega)$
6	imaginary and odd $x(-t) = -x_i(t)$	real and odd $X_r(\mathrm{i}\omega)$

As any signal can be expressed as the sum of its even and odd components, the first three items above indicate that the spectrum of the even part of a real signal is real and even, and the spectrum of the odd part of the signal is imaginary and odd.

P8. Symmetry (or Duality):

$$\text{if } \mathcal{F}(x(t)) = X(\mathrm{i}\omega), \text{ then } \mathcal{F}(X(t)) = 2\pi\, x(-\mathrm{i}\omega) \tag{7.1.547}$$

Or in a more symmetric form:

$$\text{if } \mathcal{F}(x(t)) = X(f), \text{ then } \mathcal{F}(X(t)) = x(-f) \tag{7.1.548}$$

Proof 7.3.14. As $\mathcal{F}(x(t)) = X(\mathrm{i}\omega)$, we have:

$$x(t) = \mathcal{F}^{-1}\left(X(\mathrm{i}\omega)\right) = \frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} X(\mathrm{i}\omega)e^{\mathrm{i}\omega t}\mathrm{d}\omega \tag{7.1.549}$$

Letting $t' = -t$, we get:

$$x(-t') = \frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} X(\mathrm{i}\omega)e^{-\mathrm{i}\omega t'}\mathrm{d}\omega \tag{7.1.550}$$

Interchanging t' and ω we get:

$$2\pi x(-\omega)=\int\limits_{-\infty}^{+\infty}X(t')e^{-\mathrm{i}\omega t'}\mathrm{d}t'=\mathcal{F}(X(t)) \tag{7.1.551}$$

or:

$$x(-f)=\int\limits_{-\infty}^{+\infty}X(t')e^{-\mathrm{i}2\pi f t'}\mathrm{d}t'=\mathcal{F}(X(t)) \tag{7.1.552}$$

In particular, if the signal is even:

$$x(t)=x(-t) \tag{7.1.553}$$

then we have:

$$\text{if }\ \mathcal{F}(x(t))=X(f)\ \text{ then }\mathcal{F}(X(t))=x(f) \tag{7.1.554}$$

■ Q.E.D.

For example, the spectrum of an even square wave is a sinc function, and the spectrum of a sinc function is an even square wave.

P9. Multiplication theorem:

$$\int\limits_{-\infty}^{+\infty}x(t)y^*(t)\mathrm{d}t=\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty}X(\mathrm{i}\omega)Y^*(\mathrm{i}\omega)\mathrm{d}\omega \tag{7.1.555}$$

Proof 7.3.15.

$$\begin{aligned}\int\limits_{-\infty}^{+\infty}x(t)y^*(t)\mathrm{d}t&=\int\limits_{-\infty}^{+\infty}x(t)\left[\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty}Y^*(\mathrm{i}\omega)e^{-\mathrm{i}\omega t}\mathrm{d}\omega\right]\mathrm{d}t\\&=\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty}Y^*(\mathrm{i}\omega)\left[\int\limits_{-\infty}^{+\infty}x(t)e^{-\mathrm{i}\omega t}\mathrm{d}t\right]\mathrm{d}\omega=\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty}X(\mathrm{i}\omega)Y^*(\mathrm{i}\omega)\mathrm{d}\omega\end{aligned} \tag{7.1.556}$$

■ Q.E.D.

P10. Parseval's equation for the Fourier transform (for more details see page 1756):

In the special case when $y(t)=x(t)$, the above becomes the Parseval's equation:

$$\int\limits_{-\infty}^{+\infty}|x(t)|^2\mathrm{d}t=\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty}|X(\mathrm{i}\omega)|^2\mathrm{d}\omega=\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty}S_X(\mathrm{i}\omega)\mathrm{d}\omega=\text{Total energy in }x(t) \tag{7.1.557}$$

where:

$$S_X(\mathrm{i}\omega)\triangleq|X(\mathrm{i}\omega)|^2 \tag{7.1.558}$$

is the energy density function, commonly named "power spectrum", representing how the signal's energy is distributed along the frequency axes. The total energy contained in the signal is obtained by integrating $S(\mathrm{i}\omega)$ over the entire frequency axes.

The Parseval's equation also indicates that the energy or information contained in the signal is reserved, i.e., the signal is represented equivalently in either the time or frequency domain with no energy gained or lost!

The latter relation is more commonly written:

$$\boxed{P_f = |\mathcal{F}(x(t))(\omega)|^2} \tag{7.1.559}$$

or more simply:

$$P_f = |\mathcal{F}(x(t))|^2 \tag{7.1.560}$$

✍**Example:**

We have already proved in the section of Sequences and Series that the Fourier transform of the square pulse was given by (see page 1763):

$$\mathcal{F}(\text{rect}(t)) = A\Delta t\text{sinc}(\pi f\Delta t) \tag{7.1.561}$$

Hence:

$$P_f = (A\Delta t\text{sinc}(\pi f\Delta t))^2 \tag{7.1.562}$$

The energy spectrum of the pulse we have just calculated shows a great similarity with the Fraunhofer diffraction pattern due to a narrow slit (see section Wave Optics page 4492). In reality, it is more than a similarity because it is possible to prove in physics that any diffraction pattern is the Fourier transform of the object that is the cause!

P11. Correlation:

The "cross-correlation" of two real signals $x(t)$ and $y(t)$ is defined as (notice that it is strictly equivalent to the definition of the continuous convolution seen just earlier above at page 2293):

$$R_{xy}(t) \triangleq \int\limits_{-\infty}^{+\infty} x(\tau)y(\tau - t)\mathrm{d}\tau = \int\limits_{-\infty}^{+\infty} x(t+\tau)y(\tau)\mathrm{d}\tau \tag{7.1.563}$$

Specially, when $x(t) = y(t)$, the above becomes the "auto-correlation" of signal $x(t)$:

$$R_x(t) \triangleq \int\limits_{-\infty}^{+\infty} x(\tau)x(\tau - t)\mathrm{d}\tau \tag{7.1.564}$$

Assuming $\mathcal{F}(x(t)) = X(\mathrm{i}\omega)$, we have $\mathcal{F}(x(t-\tau)) = X(\mathrm{i}\omega)e^{-\mathrm{i}\omega\tau}$ and according to multiplication theorem, $R_x(\tau)$ can be written as:

$$\begin{aligned} R_x(\tau) &= \int\limits_{-\infty}^{+\infty} x(t)x(t-\tau)\mathrm{d}t = \frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} X(\mathrm{i}\omega)X^*(\mathrm{i}\omega)e^{\mathrm{i}\omega\tau}\mathrm{d}\omega \\ &= \frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} |X(\mathrm{i}\omega)|^2 e^{\mathrm{i}\omega\tau}\mathrm{d}\omega = \frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} S_X(\mathrm{i}\omega)e^{\mathrm{i}\omega\tau}\mathrm{d}\omega = \mathcal{F}^{-1}\left(S_X(\mathrm{i}\omega)\right) \end{aligned} \tag{7.1.565}$$

i.e.:

$$\mathcal{F}(R_x(t)) = S_X(\mathrm{i}\omega) \tag{7.1.566}$$

that is, the auto-correlation and the energy density function of a signal $x(t)$ are a Fourier transform pair.

P12. Convolution Theorems:

The "convolution theorem" states that convolution in time domain corresponds to multiplication in frequency domain and vice versa:

$$\mathcal{F}(x(t) * y(t)) = X(\mathrm{i}\omega)\,Y(\mathrm{i}\omega) \quad (a) \tag{7.1.567}$$

$$\mathcal{F}(x(t)\,y(t)) = X(\mathrm{i}\omega) * Y(\mathrm{i}\omega) \quad (b) \tag{7.1.568}$$

Let us start with the proof of (a)!

Proof 7.3.16.

$$\begin{aligned} \mathcal{F}(x(t)*y(t)) &= \int\limits_{-\infty}^{+\infty}\left[\int\limits_{-\infty}^{+\infty} x(\tau)y(t-\tau)\mathrm{d}\tau\right]e^{-\mathrm{i}\omega t}\mathrm{d}t \\ &= \int\limits_{-\infty}^{+\infty} x(\tau)\left[\int\limits_{-\infty}^{+\infty} y(t-\tau)e^{-\mathrm{i}\omega t}\mathrm{d}t\right]\mathrm{d}\tau \\ &= \int\limits_{-\infty}^{+\infty} x(\tau)e^{-\mathrm{i}\omega\tau}\left[\int\limits_{-\infty}^{+\infty} y(t-\tau)e^{-\mathrm{i}\omega(t-\tau)}\mathrm{d}(t-\tau)\right]\mathrm{d}\tau \\ &= X(\mathrm{i}\omega)\,Y(\mathrm{i}\omega) \end{aligned} \tag{7.1.569}$$

■ Q.E.D.

And now let us continue with the proof of (b)!

Proof 7.3.17.

$$\begin{aligned}\mathcal{F}(x(t)\,y(t)) &= \int\limits_{-\infty}^{+\infty} x(t)\,y(t)e^{-\mathrm{i}\omega t}\mathrm{d}t\\ &= \int\limits_{-\infty}^{+\infty}\left[\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} X(\mathrm{i}\omega')e^{\mathrm{i}\omega' t}\mathrm{d}\omega'\right] y(t)e^{-\mathrm{i}\omega t}\mathrm{d}t\\ &= \frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} X(\mathrm{i}\omega')\left[\int\limits_{-\infty}^{+\infty} y(t)e^{\mathrm{i}\omega' t}e^{-\mathrm{i}\omega t}\mathrm{d}t\right]\mathrm{d}\omega'\\ &= \frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} X(\mathrm{i}\omega')\left[\int\limits_{-\infty}^{+\infty} y(t)e^{-\mathrm{i}(\omega-\omega')t}\mathrm{d}t\right]\mathrm{d}\omega'\\ &= \frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} X(\mathrm{i}\omega')Y(\mathrm{i}(\omega-\omega'))\mathrm{d}\omega' = X(\mathrm{i}\omega)*Y(\mathrm{i}\omega)\end{aligned} \tag{7.1.570}$$

■ Q.E.D.

P13. Time Derivative:

$$\mathcal{F}\left(\frac{\mathrm{d}}{\mathrm{d}t}x(t)\right) = \mathrm{i}\omega\, X(\mathrm{i}\omega) \tag{7.1.571}$$

Proof 7.3.18. Differentiating the inverse Fourier transform $X(\mathrm{i}\omega)$ with respect to t we get:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}x(t) = \frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} X(\mathrm{i}\omega)e^{\mathrm{i}\omega t}\mathrm{d}\omega\right] &= \frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} X(\mathrm{i}\omega)\frac{\mathrm{d}}{\mathrm{d}t}e^{\mathrm{i}\omega t}\mathrm{d}\omega\\ = \frac{1}{2\pi}\int\limits_{-\infty}^{+\infty}[\mathrm{i}\omega X(\mathrm{i}\omega)]e^{\mathrm{i}\omega t}\mathrm{d}\omega = \mathcal{F}^{-1}(\mathrm{i}\omega X(\mathrm{i}\omega))\end{aligned} \tag{7.1.572}$$

Repeating this process we get:

$$\mathcal{F}\left(\frac{\mathrm{d}^n}{\mathrm{d}t^n}x(t)\right) = (\mathrm{i}\omega)^n X(\mathrm{i}\omega) \tag{7.1.573}$$

■ Q.E.D.

P14. Time Integration:

First consider the Fourier transform of the following two signals:

$$u(t) = \begin{cases} 0 & t<0\\ 1 & t>0\end{cases}, \qquad \mathrm{sgn}(t) = \begin{cases} -1/2 & t<0\\ +1/2 & t>0\end{cases} = u(t) - \frac{1}{2} \tag{7.1.574}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}u(t)=\delta(t),\qquad \frac{\mathrm{d}}{\mathrm{d}t}\mathrm{sgn}(t)=\frac{\mathrm{d}}{\mathrm{d}t}\left[u(t)-\frac{1}{2}\right]=\delta(t) \tag{7.1.575}$$

According to the time derivative property above:

$$X(\mathrm{i}\omega)=\mathcal{F}(x(t))=\frac{1}{\mathrm{i}\omega}\mathcal{F}\left(\frac{\mathrm{d}}{\mathrm{d}t}x(t)\right) \tag{7.1.576}$$

we get:

$$\mathcal{F}(u(t))=\frac{1}{\mathrm{i}\omega}\mathcal{F}\left(\frac{\mathrm{d}}{\mathrm{d}t}x(t)\right)=\frac{1}{\mathrm{i}\omega}\mathcal{F}(\delta(t))=\frac{1}{\mathrm{i}\omega} \tag{7.1.577}$$

and:

$$\mathcal{F}(\mathrm{sgn}(t))=\frac{1}{\mathrm{i}\omega}\mathcal{F}\left(\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{sgn}(t)\right)=\frac{1}{\mathrm{i}\omega}\mathcal{F}(\delta(t))=\frac{1}{\mathrm{i}\omega} \tag{7.1.578}$$

Why do the two different functions have the same transform?

In general, any two function $f(t)$ and $g(t)=f(t)+c^{te}$ with a constant difference c^{te} have the same derivative $\mathrm{d}\,f(t)/\mathrm{d}t$, and therefore they have the same transform according the above method. This problem is obviously caused by the fact that the constant difference c^{te} is lost in the derivative operation.

To recover this constant difference in time domain, a delta function needs to be added in frequency domain. Specifically, as function $\mathrm{sgn}(t)$ does not have DC component, its transform does not contain a delta:

$$\mathcal{F}(\mathrm{sgn}(t))=\frac{1}{\mathrm{i}\omega} \tag{7.1.579}$$

To find the transform of $u(t)$, consider:

$$u(t)=\mathrm{sgn}(t)+\frac{1}{2} \tag{7.1.580}$$

and:

$$\mathcal{F}(u(t))=\mathcal{F}(\mathrm{sgn}(t))+\mathcal{F}\left(\frac{1}{2}\right)=\frac{1}{\mathrm{i}\omega}+\pi\delta(\omega) \tag{7.1.581}$$

The added impulse term $\pi\delta(\omega)$ directly reflects the constant $c=1/2$ in time domain.

Now we show that the Fourier transform of a time integration is:

$$\mathcal{F}\left(\int\limits_{-\infty}^{t}x(\tau)\mathrm{d}\tau\right)=\frac{1}{\mathrm{i}\omega}X(\mathrm{i}\omega)+\pi X(0)\delta(\omega) \tag{7.1.582}$$

Proof 7.3.19. First consider the convolution of $x(t)$ and $u(t)$:

$$x(t)*u(t)=\int\limits_{-\infty}^{+\infty}x(\tau)u(t-\tau)\mathrm{d}\tau=\int\limits_{-\infty}^{t}x(\tau)\mathrm{d}\tau \tag{7.1.583}$$

Due to the convolution theorem, we have:

$$\mathcal{F}\left(\int\limits_{-\infty}^{t}x(\tau)\mathrm{d}\tau\right)=\mathcal{F}(x(t)*u(t))=X(\mathrm{i}\omega)\left[\frac{1}{\mathrm{i}\omega}+\pi\delta(\omega)\right]=\frac{1}{\mathrm{i}\omega}X(\mathrm{i}\omega)+\pi X(0)\delta(\omega) \tag{7.1.584}$$

■ Q.E.D.

P15. Frequency Derivative:

$$\mathcal{F}(tx(t)) = \mathrm{i}\frac{\mathrm{d}}{\mathrm{d}\omega}X(\mathrm{i}\omega) \tag{7.1.585}$$

Proof 7.3.20. We differentiate the Fourier transform of $x(t)$ with respect to ω to get:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\omega}X(\mathrm{i}\omega) &= \frac{\mathrm{d}}{\mathrm{d}\omega}\left[\int\limits_{-\infty}^{+\infty} x(t)e^{-\mathrm{i}\omega t}\mathrm{d}t\right] = \int\limits_{-\infty}^{+\infty} x(t)\frac{\mathrm{d}}{\mathrm{d}\omega}e^{-\mathrm{i}\omega t}\mathrm{d}t\\ &= \int\limits_{-\infty}^{+\infty} x(t)(-\mathrm{i}t)e^{-\mathrm{i}\omega t}\mathrm{d}t\end{aligned} \tag{7.1.586}$$

i.e.:

$$\mathcal{F}(-\mathrm{i}tx(t)) = \frac{\mathrm{d}}{\mathrm{d}\omega}X(\mathrm{i}\omega) \tag{7.1.587}$$

Multiplying both sides by i, we get:

$$\mathrm{i}\frac{\mathrm{d}}{\mathrm{d}\omega}X(\mathrm{i}\omega) = \int\limits_{-\infty}^{+\infty} tx(t)e^{-\mathrm{i}\omega t}\mathrm{d}t = \mathcal{F}(tx(t)) \tag{7.1.588}$$

Repeating this process we get:

$$\mathcal{F}(t^n x(t)) = \mathrm{i}^n\frac{\mathrm{d}^n}{\mathrm{d}\omega^n}X(\mathrm{i}\omega) \tag{7.1.589}$$

■ Q.E.D.

Property name	**Time domainfunction** $f(t)$	**Fourier Transform** $F(\omega)$
Linearity	$af_1(t)+bf_2(t)$	$aF_1(\omega)+bF_2(\omega)$
Time Scaling	$f(at)$	$\frac{1}{\lvert a \rvert}F\left(\frac{\omega}{a}\right)$
Time Shifting	$f(t-t_0)$	$e^{-\mathrm{i}\omega t_0}F(\omega)$
Frequency Shifting	$e^{\mathrm{i}\omega_0 t}f(t)$	$F(\omega-\omega_0)$
Time Reversal	$f(-t)$	$F(-\omega)$
Time Multiplication	$t^n f(t)$	$\mathrm{i}^n\frac{\mathrm{d}}{\mathrm{d}\omega^n}X(\omega)$
Differential	$\frac{\mathrm{d}f(t)}{\mathrm{d}t}$	$\mathrm{i}\omega F(\omega)$
n^{th} Differential	$\frac{\mathrm{d}^n f(t)}{\mathrm{d}t^n}$	$(\mathrm{i}\omega)^n F(\omega)$
Integration	$\int\limits_{-\infty}^{t} f(t)\mathrm{d}t$	$\frac{F(\omega)}{\mathrm{i}\omega}+\pi F(0)\delta(\omega)$
Multiplication	$f_1(t)f_2(t)$	$\frac{1}{2\pi}F_1(\omega)\star F_2(\omega)$
Convolution	$f_1(t)*f_2(t)$	$F_1(\omega)\cdot F_2(\omega)$

Table 7.7: **Summary of most common Fourier transform properties**

7.1.5.1.1.2 Usual Fourier transforms

There are in math and physics many Fourier transforms of signals we see quite frequently (but not exclusively). Furthermore, all Fourier transforms proved below will be used in the various chapters on Physics, Engineering, Atomistic, Social Mathematics, etc of this textbook. So, as in any formula booklet, we propose you the most well known Fourier transforms but with... the proofs!

1. Impulse:

   As shown above:

   $$\mathcal{F}(\delta(t-a))=\int\limits_{-\infty}^{+\infty}\delta(t-a)e^{-\mathrm{i}\omega t}\mathrm{d}t=e^{\mathrm{i}\omega a} \qquad (7.1.590)$$

   or:

   $$\mathcal{F}(\delta(t))=1 \qquad (7.1.591)$$

   It is therefore immediate that the inverse Fourier transform of the complex exponential is the Dirac delta:

   $$\mathcal{F}^{-1}(e^{\mathrm{i}\omega a})=\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty}e^{\mathrm{i}\omega a}e^{-\mathrm{i}\omega t}\mathrm{d}t=\delta(t-a) \qquad (7.1.592)$$

2. Unit Step:

   As shown above:

   $$\mathcal{F}(u(t))=\frac{1}{\mathrm{i}\omega}+\pi\delta(\omega) \qquad (7.1.593)$$

3. Constant:

   As shown above:

   $$\mathcal{F}(x(t))=\int\limits_{-\infty}^{+\infty}e^{-\mathrm{i}\omega t}\mathrm{d}t=2\pi\,\delta(\omega) \qquad (7.1.594)$$

   This is a useful formula.

4. Complex exponential:

   The spectrum of a complex exponential can be found from the above due to the frequency shift property:

   $$\mathcal{F}(e^{\mathrm{i}\omega_0 t})=2\pi\,\delta(\omega-\omega_0) \qquad (7.1.595)$$

   Let us write this more explicitly:

   $$\mathcal{F}(e^{\mathrm{i}\omega_0 t})=\int\limits_{-\infty}^{+\infty}e^{\mathrm{i}\omega_0 t}e^{-\mathrm{i}\omega t}\mathrm{d}t=\int\limits_{-\infty}^{+\infty}e^{\mathrm{i}(\omega_0-\omega)t}\mathrm{d}t=\left.\frac{e^{\mathrm{i}(\omega_0-\omega)t}}{\mathrm{i}(\omega_0-\omega)}\right|_{-\infty}^{+\infty}=\text{hummm....} \qquad (7.1.596)$$

Now let us try to solve this using the physicist way... Using what we have just seen before (the inverse Fourier transform of the Dirac pulse), we have:

$$\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} e^{\mathrm{i}\omega a}e^{-\mathrm{i}\omega t}\mathrm{d}t=\delta(t-a) \tag{7.1.597}$$

it comes obviously after rearranging (multiplying both sides by 2π):

$$\int\limits_{-\infty}^{+\infty} e^{\mathrm{i}\omega a}e^{-\mathrm{i}\omega t}\mathrm{d}t=2\pi\delta(t-a) \tag{7.1.598}$$

5. Sinusoids:

$$x(t)=\cos(\omega_0 t)=\frac{1}{2}(e^{\mathrm{i}\omega_0 t}+e^{-\mathrm{i}\omega_0 t}) \tag{7.1.599}$$

Therefore:

$$X(\mathrm{i}\omega)=\mathcal{F}(\cos(\omega_0 t))=\frac{1}{2}\left(\mathcal{F}(e^{\mathrm{i}\omega_0 t})+\mathcal{F}(e^{-\mathrm{i}\omega_0 t})\right)=\pi[\delta(\omega-\omega_0)+\delta(\omega+\omega_0)] \tag{7.1.600}$$

Similarly, we have:

$$X(\mathrm{i}\omega)=\mathcal{F}(\sin(\omega_0 t))=\frac{\pi}{\mathrm{i}}[\delta(\omega-\omega_0)-\delta(\omega+\omega_0)] \tag{7.1.601}$$

6. Exponential decay (right-sided):

$$x(t)=e^{-at}u(t)\quad (a>0) \tag{7.1.602}$$

Therefore:

$$\begin{aligned}\mathcal{F}(e^{-at}u(t))&=\int\limits_{-\infty}^{+\infty} x(t)e^{-\mathrm{i}\omega t}\mathrm{d}t=\int\limits_{0}^{+\infty} e^{-at}e^{-\mathrm{i}\omega t}\mathrm{d}t\\&=\frac{-1}{a+\mathrm{i}\omega}e^{-(a+\mathrm{i}\omega)t}|_0^{+\infty}=\frac{1}{a+\mathrm{i}\omega}\end{aligned} \tag{7.1.603}$$

7. Exponential decay (left-sided):

Due to the time reversal property, we also have (for $a>0$):

$$\mathcal{F}(e^{at}u(-t))=\frac{1}{a-\mathrm{i}\omega} \tag{7.1.604}$$

or:

$$\mathcal{F}(-e^{at}u(-t))=\frac{1}{\mathrm{i}\omega-a} \tag{7.1.605}$$

8. Exponential decay (two-sided):

$$x(t)=e^{-a|t|}\quad(a>0) \tag{7.1.606}$$

As the two-sided exponential decay is the sum of the right and left-sided exponential decays, its spectrum of $x(t)$ is the sum of their spectra due to linearity:

$$\mathcal{F}(e^{-a|t|})=\frac{1}{a+\mathrm{i}\omega}+\frac{1}{a-\mathrm{i}\omega}=\frac{2a}{a^2+\omega^2} \tag{7.1.607}$$

9. Comb function:

The comb function is defined as:

$$\text{comb}(t)=\sum_{n=-\infty}^{+\infty}\delta(t-nT) \tag{7.1.608}$$

Its Fourier series coefficient is:

$$\text{COMB}[k]=\frac{1}{T}\int\limits_{-T/2}^{T/2}\text{comb}(t)e^{-\mathrm{i}\omega_0 t}\mathrm{d}t=\frac{1}{T}\int\limits_{-T/2}^{T/2}\delta(t)e^{-\mathrm{i}\omega_0 t}\mathrm{d}t=\frac{1}{T} \tag{7.1.609}$$

and its spectrum is:

$$\text{COMB}(\mathrm{i}\omega)=\frac{2\pi}{T}\sum_{k=-\infty}^{+\infty}\delta\left(\omega-k\omega_0\right)=\frac{1}{T}\sum_{k=-\infty}^{+\infty}\delta\left(f-\frac{k}{T}\right) \tag{7.1.610}$$

We see that the spectrum of an impulse train with time interval T is also an impulse train with frequency interval $\omega_0=2\pi/T$. Also, according to the definition of the Fourier transform, we have:

$$\begin{aligned}\text{COMB}(\mathrm{i}\omega)&=\int\limits_{-\infty}^{+\infty}\text{comb}(t)e^{-\mathrm{i}\omega t}\mathrm{d}t=\int\limits_{-\infty}^{+\infty}\left[\sum_{n=-\infty}^{+\infty}\delta(t-nT)\right]e^{-\mathrm{i}\omega t}\mathrm{d}t\\&=\sum_{n=-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty}\delta(t-nT)e^{-\mathrm{i}\omega t}\mathrm{d}t=\sum_{n=-\infty}^{+\infty}e^{-\mathrm{i}\omega nT}\end{aligned} \tag{7.1.611}$$

Therefore we have this equation:

$$\sum_{n=-\infty}^{+\infty}e^{-\mathrm{i}\omega nT}=\frac{2\pi}{T}\sum_{k=-\infty}^{+\infty}\delta(\omega-k\omega_0)=\frac{1}{T}\sum_{k=-\infty}^{+\infty}\delta\left(f-\frac{k}{T}\right) \tag{7.1.612}$$

which can be compared with the equation in continuous case:

$$\int\limits_{-\infty}^{+\infty}e^{-\mathrm{i}\omega t}\mathrm{d}t=2\pi\delta(\omega)=\delta(f) \tag{7.1.613}$$

10. Square wave (for another detailed derivation see page 1763):

    A square wave or rectangular function of width a can be considered as the difference between two unit step functions:

$$\text{rect}_a(t)=u(t+a)-u(t-a)=\begin{cases}1 & -a\leq t<a\\ 0 & \text{else}\end{cases}\tag{7.1.614}$$

    and due to linearity, its Fourier spectrum is the difference between the two corresponding spectra:

$$\begin{aligned}\mathcal{F}(\text{rect}_a(t))&=\mathcal{F}(u(t+a))-\mathcal{F}(u(t-a))=\frac{1}{\mathrm{i}\omega}e^{\mathrm{i}a\omega}-\frac{1}{\mathrm{i}\omega}e^{-\mathrm{i}a\omega}\\&=\frac{2}{\omega}\,\frac{e^{\mathrm{i}a\omega}-e^{-\mathrm{i}a\omega}}{2\mathrm{i}}=\frac{2}{\omega}\sin(a\omega)=\frac{2a}{\omega}\text{sinc}(a\omega)\end{aligned}\tag{7.1.615}$$

11. Sinc function:

    The spectrum of an ideal low-pass filter is:

$$X(\mathrm{i}\omega)=\begin{cases}1 & -a\leq\omega<a\\ 0 & \text{else}\end{cases}\tag{7.1.616}$$

    and its impulse response can be found by inverse Fourier transform:

$$x(t)=\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty}X(\mathrm{i}\omega)e^{\mathrm{i}\omega t}\mathrm{d}\omega=\frac{1}{2\pi\mathrm{i}t}(e^{\mathrm{i}at}-e^{-\mathrm{i}at})=\frac{\sin(at)}{\pi t}\tag{7.1.617}$$

12. Triangle function:

$$x(t)=\text{triangle}(t)=\begin{cases}1-|t| & |t|<1\\ 0 & |t|\geq 1\end{cases}\tag{7.1.618}$$

    As $x(t)$ is an even function, its Fourier transform is:

$$\begin{aligned}\mathcal{F}(x(t))&=\int\limits_{-\infty}^{+\infty}x(t)e^{-\mathrm{i}\omega t}\mathrm{d}t=2\int\limits_{0}^{1}(1-t)\cos(\omega t)\mathrm{d}t\\&=2\left[\int\limits_{0}^{1}\cos(\omega t)\mathrm{d}t-\int\limits_{0}^{1}t\,\cos(\omega t)\mathrm{d}t\right]=\frac{2}{\omega}\left[\sin(\omega)-\int\limits_{0}^{1}t\mathrm{d}\sin(\omega t)\right]\\&=\frac{2}{\omega}\left[\sin(\omega)-(t\,\sin(\omega t)|_0^1-\int\limits_{0}^{1}\sin(\omega t)\mathrm{d}t)\right]=\frac{2}{\omega^2}\cos(\omega t)|_1^0\\&=\frac{2}{\omega^2}(1-\cos(\omega))=\left(\frac{2}{\omega}\right)^2\sin^2\left(\frac{\omega}{2}\right)=\frac{\sin^2(\pi f)}{(\pi f)^2}=\text{sinc}^2(\omega)\end{aligned}\tag{7.1.619}$$

    Alternatively, as the triangle function is the convolution of two square functions ($a=1/2$), its Fourier transform can be more conveniently obtained according to the convolution theorem as:

$$\mathcal{F}(\text{triangle}(t))=\mathcal{F}(\text{rect}(t)*\text{rect}(t))=\text{sinc}(\omega)\,\text{sinc}(\omega)=\left(\frac{2}{\omega}\right)^2\sin^2\left(\frac{\omega}{2}\right)\tag{7.1.620}$$

13. Gaussian function:

    The Fourier transform of a Gaussian or bell-shaped function $x(t)=e^{-\pi t^2}$ is:

$$\begin{aligned}X(\mathrm{i}\omega)=\mathcal{F}(x(t))&=\int\limits_{-\infty}^{+\infty}[e^{-\pi t^2}]e^{-\mathrm{i}\omega t}\mathrm{d}t\\&=\int\limits_{-\infty}^{+\infty}e^{-\pi(t^2+\mathrm{i}2ft)}\mathrm{d}t=e^{\pi(\mathrm{i}f)^2}\int\limits_{-\infty}^{+\infty}e^{-\pi[t^2+\mathrm{i}2ft+(\mathrm{i}f)^2]}\mathrm{d}t\\&=e^{-\pi f^2}\int\limits_{-\infty}^{+\infty}e^{-\pi(t+\mathrm{i}f)^2}\mathrm{d}(t+\mathrm{i}f)=e^{-\pi f^2}\end{aligned}\tag{7.1.621}$$

    Here we have used the identity:

$$\int\limits_{-\infty}^{+\infty}e^{-\pi x^2}\mathrm{d}x=1\tag{7.1.622}$$

    We see that the Fourier transform of a bell-shaped function is also a bell-shaped function:

$$\mathcal{F}(e^{-\pi t^2})=e^{-\pi f^2}\tag{7.1.623}$$

    Note that the area underneath either $x(t)$ or $X(\mathrm{i}\omega)$ is unity. Moreover, due to the property of time and frequency scaling, we have:

$$\mathcal{F}(a\,e^{-\pi(at)^2})=e^{-\pi(f/a)^2}\tag{7.1.624}$$

    Note that if $a=1/\sqrt{2\pi\sigma^2}$, then $a\,x(at)$ above is a normal distribution with variance σ^2 and mean $\mu=0$. If we let $a\rightarrow\infty$, $x(t)$ becomes narrower and taller and approaches $\delta(t)$, and its spectrum $e^{-\pi(f/a)^2}$ becomes wider and approaches constant 1. On the other hand, if we rewrite the above as:

$$\mathcal{F}(e^{-\pi(at)^2})=\frac{1}{a}e^{-\pi(f/a)^2}\tag{7.1.625}$$

    and let $a\rightarrow 0$, $x(t)$ approaches 1 and $X(\mathrm{i}\omega)$ approaches $\delta(\omega)$.

To summarize here is a table with the most important Fourier Transforms:

Function $f(t)$	**Fourier Transform** $\mathcal{F}(f(t))$
$f(t-t_0)$	$\mathcal{F}(\omega)e^{-\mathrm{i}\omega t_0}$
$f(t)e^{\mathrm{i}\omega_0 t}$	$\mathcal{F}(\omega-\omega_0)$
$f(\alpha t)$	$\dfrac{1}{\lvert\alpha\rvert}\mathcal{F}\left(\dfrac{\omega}{\alpha}\right)$
$\mathcal{F}(t)$	$2\pi f(-\omega)$
$\dfrac{\mathrm{d}^n f(t)}{\mathrm{d}t^n}$	$(\mathrm{i}\omega)^n\mathcal{F}(\omega)$
$(-\mathrm{i}t)^n f(t)$	$\dfrac{\mathrm{d}^n\mathcal{F}(\omega)}{\mathrm{d}\omega^n}$
$\displaystyle\int_{-\infty}^{t} f(\tau)d\tau$	$\dfrac{\mathcal{F}(\omega)}{\mathrm{i}\omega}+\pi\mathcal{F}(0)\delta(\omega)$
$h\cdot\text{rect}\left(\dfrac{t}{a}\right)$	$h\cdot a\cdot\text{sinc}\left(\dfrac{\omega a}{2}\right)$
$\text{rect}(t)$	$\frac{2a}{\omega}\text{sinc}(a\omega)$
e^{-at^2}	$\dfrac{1}{\sqrt{2a}}e^{-\frac{\omega^2}{4a}}$
$\delta(t-a)$	1
c^{te}	$2\pi\delta(\omega)$
$u(t)$	$\dfrac{1}{\mathrm{i}\omega}+\pi\delta(\omega)$
$e^{i\omega_0 t}$	$2\pi\delta(\omega-\omega_0)$
$\cos(\omega_0 t)$	$\pi\left[\delta(\omega-\omega_0)+\delta(\omega+\omega_0)\right]$
$\sin(\omega_0 t)$	$-\mathrm{i}\pi\left[\delta(\omega-\omega_0)-\delta(\omega+\omega_0)\right]$
$e^{-at}u(t)$	$\dfrac{1}{a+\mathrm{i}\omega}$
$e^{at}u(t)$	$\dfrac{1}{a-\mathrm{i}\omega}$
$-e^{at}u(-t)$	$\dfrac{1}{\mathrm{i}\omega-a}$
$e^{-a\lvert t\rvert}$	$\dfrac{2a}{a^2+\omega^2}$
$\text{comb}(t)=\displaystyle\sum_{n=-\infty}^{+\infty}\delta(t-nT)$	$\displaystyle\sum_{n=-\infty}^{+\infty}e^{-i\omega nT}$
$\text{triangle}(t)$	$\text{sinc}^2(\omega)$

Table 7.8: **Most common Fourier transforms**

Now that we have seen quite a number of properties and usual Fourier Transform, let us go back to our heat equation determined in the section of Thermodynamics (see page 3684) in the form:

$$\frac{\partial u}{\partial t}-D\frac{\partial^2 u}{\partial x^2}=0 \tag{7.1.626}$$

And let us also be interested in the case where:

$$\begin{aligned} u(x,0)&=f(x)\\ f(\pm\infty)&\rightarrow 0 \end{aligned} \tag{7.1.627}$$

To simplify the writing, we will write the differential equation in the following form:

$$\partial_t u=D\partial_x^2 u \tag{7.1.628}$$

Let us take the Fourier transform relatively to x of this equality. Let us recall that for this purpose we have proved in the section of Sequences and Series and just earlier at page 2316 that (time derivative property):

$$\mathcal{F}(\partial_x^2 u(x))=(\mathrm{i}\omega)^2\mathcal{F}(u(x))=-\omega^2\mathcal{F}(u(x)) \qquad (7.1.629)$$

Let us put to simplify the notations:

$$\mathcal{F}(u(x))=\hat{u} \qquad (7.1.630)$$

As we take the Fourier transform with respect to x, we can take out the partial derivative of the integral of the Fourier transform such as:

$$\mathcal{F}(\partial_t u(x))=\partial_t\mathcal{F}(u(x))=\partial_t\hat{u} \qquad (7.1.631)$$

Our differential equation is then reduced to:

$$\partial_t\hat{u}=-\omega^2 D\hat{u} \qquad (7.1.632)$$

Thus explicitly:

$$\partial_t\left(\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}ue^{-\mathrm{i}\omega x}\mathrm{d}x\right)=-\omega^2 D\left(\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}ue^{-\mathrm{i}\omega x}\mathrm{d}x\right) \qquad (7.1.633)$$

We see then immediately with the simplified version that a particular solution is:

$$\hat{u}(\omega,t)=C(\omega)e^{-\omega^2 Dt} \qquad (7.1.634)$$

The constant has to be determined by the initial condition:

$$u(x,0)=f(x)\Rightarrow\hat{u}(\omega,0)=\hat{f}(x) \qquad (7.1.635)$$

Therefore:

$$\hat{u}(\omega,t)=C(\omega)e^{-\omega^2 Dt}=\hat{u}(\omega,0)e^{-\omega^2 Dt}=\hat{f}(x)e^{-\omega^2 Dt} \qquad (7.1.636)$$

Then it comes by doing the inverse Fourier transform in x:

$$u(x,t)=\mathcal{F}^{-1}(\hat{u}(\omega,t))=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}\hat{f}(x)e^{-\omega^2 Dt}e^{\mathrm{i}\omega x}\mathrm{d}x=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}\hat{f}(x)e^{-\omega^2 Dt+\mathrm{i}\omega x}\mathrm{d}\omega \qquad (7.1.637)$$

Now let's make a small change of notation by putting:

$$\omega=\lambda \qquad \hat{f}(x)=F(\omega)=F(\lambda) \qquad (7.1.638)$$

Then we have:

$$u(x,t)=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}F(\omega)e^{-\lambda^2 Dt+\mathrm{i}\lambda x}\mathrm{d}\omega \qquad (7.1.639)$$

We then fall much more quickly on the same integral that we obtained in the section of Thermodynamics in our study of the heat equation, the finishing work being the same, the reader can refer to it!

We had also proved in the section Thermodynamics that we obtained:

$$u(x,t)=\frac{1}{2\sqrt{\pi Dt}}\int\limits_{-\infty}^{+\infty}e^{-\frac{(v-x)^2}{4Dt}}\mathrm{d}v \tag{7.1.640}$$

We then notice that the Fourier transform is not only a tool for analysing a frequency domain signal but also for solving some differential equations more quickly.

But as many times in mathematics, we must be careful using such a tool. They may be some trap and subtilities. Let us see on famous example!

We know that the set of solutions of the differential equation:

$$y'=0 \tag{7.1.641}$$

where y is a continuous function from $\mathbb{R}\mapsto\mathbb{C}$, is made of the set of constant functions!

Let us try to solve this equation using the Fourier transform. By taking the Fourier transform on the left and on the right we get:

$$\omega\mathcal{F}(y)=0 \tag{7.1.642}$$

An error would be to think that we can divide left and right by ω and thus get:

$$\mathcal{F}(y)=0 \tag{7.1.643}$$

In this case, taking the inverse transform would get $y=0$ as the only solution to the equation, then we would lose all other solutions.

In fact, what we must remember is that the Fourier transform is defined on the space of "temperate distributions" and that therefore, as long as we decide to use this integral transform to solve the differential equation above, we also decide, implicitly, to look for solutions in this space. Now in the space of temperate distributions, the equation $\omega\mathcal{F}(y)=0$ possesses an infinity of solutions that are given by (without proof) $c\cdot\delta$ where c is a complex number and δ is the Dirac distribution. As a result:

$$\hat{y}=c\cdot\delta \tag{7.1.644}$$

and taking the inverse transform we get indeed:

$$y=c^{te} \tag{7.1.645}$$

We therefore conclude that when we solve differential equations with the Fourier transform it must be remembered that in the space of temperate distributions the usual algebraic calculation rules are to be handled with care. This remark is obviously valid only for people who know the Fourier transform but have only a vague idea of what is a temperate distribution (which is normally the case for engineers....

7.1.5.1.2 Discrete Time Fourier Transform

A discrete-time signal can be considered as a continuous signal $x(t)$ sampled at a rate $F=1/t_0$ or $\Omega=2\pi/t_0$, where t_0 is the sampling period (time interval between two consecutive samples). The corresponding sampling function (comb function) is:

$$\text{comb}(t)=\sum_{m=-\infty}^{+\infty}\delta(t-mt_0) \tag{7.1.646}$$

The sampling process can be represented by:

$$x_s(t)=x(t)\,\text{comb}(t)=x(t)\sum_{m=-\infty}^{+\infty}\delta(t-mt_0)=\sum_{m=-\infty}^{+\infty}x[m]\delta(t-mt_0) \tag{7.1.647}$$

where $x[m]=x(mt_0)$ is the value of $x(t)$ at $t=mt_0$. The Fourier transform of this discrete signal (treated as a special case of continuous signal) is:

$$\begin{aligned}X(\mathrm{i}\omega)&=\int_{\infty}^{+\infty}x_s(t)e^{-\mathrm{i}\omega t}\mathrm{d}t=\int_{\infty}^{+\infty}\left[\sum_{m=-\infty}^{+\infty}x[m]\delta(t-mt_0)\right]e^{-\mathrm{i}\omega t}\mathrm{d}t\\&=\sum_{m=-\infty}^{+\infty}x[m]\int_{\infty}^{+\infty}\delta(t-mt_0)\,e^{-\mathrm{i}\omega t}\mathrm{d}t=\sum_{m=-\infty}^{+\infty}x[m]\,e^{-\mathrm{i}\omega mt_0}\end{aligned} \tag{7.1.648}$$

This is the forward Fourier transform (analysis) of a discrete signal $x_s(t)$. The spectrum $X(\mathrm{i}\omega)$ is periodic with period $\Omega=2\pi F=2\pi/t_0$:

$$X(\mathrm{i}(\omega+\Omega))=\sum_{m=-\infty}^{+\infty}x[m]e^{-\mathrm{i}(\omega+\Omega))mt_0}=\sum_{m=-\infty}^{+\infty}x[m]e^{-\mathrm{i}\omega mt_0}e^{-\mathrm{i}\Omega mt_0}=X(\mathrm{i}\omega) \tag{7.1.649}$$

as :

$$e^{-\mathrm{i}\Omega mt_0}=e^{-\mathrm{i}2m\pi}=1 \tag{7.1.650}$$

To get back the time signal $x[m]$ from its spectrum:

$$X(\mathrm{i}\omega)=\sum_{m=-\infty}^{+\infty}x[m]\,e^{-\mathrm{i}\omega mt_0} \tag{7.1.651}$$

we multiply the equation by $e^{\mathrm{i}\omega nt_0}/\Omega$ and integrate both sides with respect to ω over the period $\Omega=2\pi F=2\pi/t_0$ to obtain the inverse Fourier transform (synthesis):

$$\begin{aligned}\frac{1}{\Omega}\int_{\Omega}X(\mathrm{i}\omega)e^{\mathrm{i}\omega nt_0}\mathrm{d}\omega&=\frac{1}{\Omega}\int_{\Omega}[\sum_{m=-\infty}^{+\infty}x[m]e^{-\mathrm{i}\omega mt_0}]e^{\mathrm{i}\omega nt_0}\mathrm{d}\omega\\&=\sum_{m=-\infty}^{+\infty}x[m]\frac{1}{\Omega}\int_{\Omega}e^{-\mathrm{i}\omega(m-n)t_0}\mathrm{d}\omega=\sum_{m=-\infty}^{+\infty}x[m]\delta[m-n]=x[n]\end{aligned} \tag{7.1.652}$$

Note that here we used:

$$\frac{1}{\Omega}\int\limits_{\Omega} e^{-\mathrm{i}\omega(m-n)t_0}\mathrm{d}\omega=\frac{1}{\Omega}\int\limits_{\Omega} e^{-\mathrm{i}(m-n)2\pi\omega/\Omega}\mathrm{d}\omega=\delta[m-n]=\begin{cases}1 & m=n\\ 0 & m\neq n\end{cases} \tag{7.1.653}$$

which can be compared with:

$$\frac{1}{T}\int\limits_{T} e^{\mathrm{i}(m-n)\omega_0 t}\mathrm{d}t=\frac{1}{T}\int\limits_{T} e^{\mathrm{i}(m-n)2\pi t/T}\mathrm{d}t=\delta[m-n]=\begin{cases}1 & m=n\\ 0 & m\neq n\end{cases} \tag{7.1.654}$$

To summarize, the spectrum of a given discrete signal:

$$\boxed{x_s(t)=\sum_{m=-\infty}^{+\infty} x[m]\delta(t-mt_0)} \tag{7.1.655}$$

can be found by the "forward discrete Fourier transform" to be:

$$\boxed{X_\Omega(\mathrm{i}\omega)=\mathcal{F}[x[m]]=\sum_{m=-\infty}^{+\infty} x[m]e^{-\mathrm{i}\omega m t_0}=\sum_{m=-\infty}^{+\infty} x[m]e^{-\mathrm{i}2\pi m t_0}} \tag{7.1.656}$$

and the signal can be expressed by inverse Fourier transform:

$$\boxed{x[m]=\mathcal{F}^{-1}[X_\Omega(\mathrm{i}\omega)]=\frac{1}{\Omega}\int\limits_{\Omega} X_\Omega(\mathrm{i}\omega)e^{\mathrm{i}\omega m t_0}\mathrm{d}\omega=\int\limits_{F} X_F(f)e^{\mathrm{i}2\pi m t_0}\mathrm{d}f} \tag{7.1.657}$$

It is interesting to compare this discrete time Fourier transform pair with the Fourier series expansion (the Fourier transform of a periodic signal):

$$x_T(t)=\mathcal{F}^{-1}[X[n]]=\sum_{n=-\infty}^{+\infty} X[n]e^{\mathrm{i}n\omega_0 t}=\sum_{n=-\infty}^{+\infty} X[n]e^{\mathrm{i}2\pi n f_0 t} \tag{7.1.658}$$

$$X[n]=\mathcal{F}[x_T(t)]=\frac{1}{T}\int\limits_{T} x_T(t)e^{-\mathrm{i}n\omega_0 t}\mathrm{d}t=\frac{1}{T}\int\limits_{T} x_T(t)e^{-\mathrm{i}2\pi n f_0 t}\mathrm{d}t \tag{7.1.659}$$

with discrete spectrum:

$$X(\mathrm{i}\omega)=2\pi\sum_{n=-\infty}^{+\infty} X[n]\delta(\omega-n\omega_0)\quad\text{or}\quad X(f)=\sum_{n=-\infty}^{+\infty} X[n]\delta(f-nf_0) \tag{7.1.660}$$

We see symmetry between these two different forms of Fourier transform. If the signal $x(t)=x(t+T)$ is periodic, its spectrum $X(\mathrm{i}\omega)$ is discrete, the coefficients of the Fourier series with interval $\omega_0=2\pi/T$. On the other hand, if $x(t)$ is discrete with interval $t_0=2\pi/\Omega$, its spectrum $X(\mathrm{i}\omega)=X(\mathrm{i}\omega+\Omega)$ is periodic.

In particular, if the unit of time is so chosen that the sampling period is $t_0=1$, then $\Omega=2\pi/t_0=2\pi$, and the forward Fourier transform of a discrete signal becomes:

$$X(\mathrm{i}\omega)=\sum_{m=-\infty}^{+\infty} x[m]e^{-\mathrm{i}m\omega}=\sum_{m=-\infty}^{+\infty} x[m]e^{-\mathrm{i}m2\pi f} \tag{7.1.661}$$

The inverse transform becomes:

$$x_s[m] = \frac{1}{2\pi}\int\limits_0^{2\pi} X(\mathrm{i}\omega)e^{\mathrm{i}m\omega}\mathrm{d}\omega = \int\limits_0^{1} X(\mathrm{i}\omega)e^{\mathrm{i}m2\pi f}\mathrm{d}f \tag{7.1.662}$$

The spectrum $X(\mathrm{i}\omega) = X(\mathrm{i}\omega + 2\pi)$ is periodic.

Remark

The spectrum of a time signal (continuous or discrete) can be denoted by $X(\mathrm{i}\omega)$ or $X(f)$ to emphasize the fact that the spectrum represents how the energy contained in the signal is distributed as a function of frequency ω or f. Moreover, if $X(f)$ is used, the factor $1/2\pi$ in front of the inverse transform is dropped so that the transform pair takes a more symmetric form. On the other hand, as Fourier transform of discrete signal can be considered as a special case of $\mathcal{Z}$ transform when the real part of $s = \sigma + \mathrm{i}\omega$ is zero, i.e., $z = e^s = e^{\mathrm{i}\omega}$:

$$X(z)|_{z=e^{\mathrm{i}\omega}} = \sum_{n=-\infty}^{+\infty} x[n]z^{-n}|_{z=e^{\mathrm{i}\omega}} = \sum_{n=-\infty}^{+\infty} x[n]e^{-\mathrm{i}\omega n} = X(e^{\mathrm{i}\omega}) \tag{7.1.663}$$

it is also natural to denote the spectrum of $x[n]$ by $X(e^{\mathrm{i}\omega})$.

7.1.5.1.3 Properties of Discrete Fourier Transform

As a special case of general Fourier transform, the discrete time transform shares all properties (and their proofs) of the Fourier transform discussed above, except now some of these properties may take different forms. In the following, we always assume $\mathcal{F}[x[m]] = X(e^{\mathrm{i}\omega})$ and $\mathcal{F}[y[m]] = Y(e^{\mathrm{i}\omega})$.

P1. Linearity:

$$\mathcal{F}[ax[m] + by[m]] = aX(e^{\mathrm{i}\omega}) + bY(e^{\mathrm{i}\omega}) \tag{7.1.664}$$

P2. Time Shifting:

$$\mathcal{F}[x[m - m_0]] = e^{-\mathrm{i}m_0\omega}X(e^{\mathrm{i}\omega}) \tag{7.1.665}$$

Proof 7.3.21.

$$\mathcal{F}[x[m - m_0]] = \sum_{m=-\infty}^{+\infty} x[m - m_0]e^{-\mathrm{i}\omega m} \tag{7.1.666}$$

If we let $m' = m - m_0$, the above becomes:

$$\mathcal{F}[x[m - m_0]] = \sum_{m=-\infty}^{+\infty} x[m']e^{-\mathrm{i}\omega(m'+m_0)} = e^{-\mathrm{i}\omega m_0}X(e^{\mathrm{i}\omega}) \tag{7.1.667}$$

■ Q.E.D.

P3. Time Reversal:

$$\mathcal{F}[x[-m]] = X(e^{-\mathrm{i}\omega}) \tag{7.1.668}$$

P4. Frequency Shifting:

$$\mathcal{F}[x[m]e^{\mathrm{i}\omega_0 m}] = X(e^{\mathrm{i}(\omega-\omega_0)}) \tag{7.1.669}$$

P5. Differencing:

Differencing is the discrete-time counterpart of differentiation.

$$\mathcal{F}[x[m] - x[m-1]] = (1-e^{-\mathrm{i}\omega})X(e^{\mathrm{i}\omega}) \tag{7.1.670}$$

Proof 7.3.22.

$$\begin{aligned}\mathcal{F}[x[m]-x[m-1]] &= \mathcal{F}[x[m]] - \mathcal{F}[x[m-1]]\\ &= X(e^{\mathrm{i}\omega}) - X(e^{\mathrm{i}\omega})e^{-\mathrm{i}\omega} = (1-e^{-\mathrm{i}\omega})X(e^{\mathrm{i}\omega})\end{aligned} \tag{7.1.671}$$

■ Q.E.D.

P6. Differentiation in frequency:

$$\mathcal{F}^{-1}\left[\mathrm{i}\frac{\mathrm{d}}{\mathrm{d}\omega}X(e^{\mathrm{i}\omega})\right] = m\,x[m] \tag{7.1.672}$$

Proof 7.3.23. Differentiating the definition of discrete Fourier transform with respect to ω, we get:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\omega}X(e^{\mathrm{i}\omega}) &= \frac{\mathrm{d}}{\mathrm{d}\omega}\sum_{m=-\infty}^{+\infty} x[m]e^{-\mathrm{i}\omega m} = \sum_{m=-\infty}^{+\infty} x[m]\frac{\mathrm{d}}{\mathrm{d}\omega}e^{-\mathrm{i}\omega m}\\ &= \sum_{m=-\infty}^{+\infty} -\mathrm{i}mx[m]e^{-\mathrm{i}\omega m}\end{aligned} \tag{7.1.673}$$

■ Q.E.D.

P7. Convolution Theorems:

The convolution theorem states that convolution in time domain corresponds to multiplication in frequency domain and vice versa:

$$\mathcal{F}[x[n]*y[n]] = X(e^{\mathrm{i}\omega})\,Y(e^{\mathrm{i}\omega}) \quad (a) \tag{7.1.674}$$

$$\mathcal{F}[x[n]\,y[n]] = X(e^{\mathrm{i}\omega}) * Y(e^{\mathrm{i}\omega}) \quad (b) \tag{7.1.675}$$

Recall that the convolution of periodic signals $x_T(t)$ and $y_T(t)$ is:

$$x_T(t)*y_T(t) \triangleq \frac{1}{T}\int\limits_T x_T(\tau)y_T(t-\tau)\mathrm{d}\tau \tag{7.1.676}$$

Here the convolution of periodic spectra $X(f)$ and $Y(f)$ is similarly defined as:

$$X(e^{\mathrm{i}\omega})*Y(e^{\mathrm{i}\omega})=\frac{1}{\Omega}\int\limits_{\Omega}X(e^{\mathrm{i}\omega'})Y(e^{\mathrm{i}(\omega-\omega')})\mathrm{d}\omega'=\frac{1}{2\pi}\int\limits_{0}^{2\pi}X(e^{\mathrm{i}\omega'})Y(e^{\mathrm{i}(\omega-\omega')})\mathrm{d}\omega' \quad (7.1.677)$$

Proof of (a):

$$\begin{aligned}\mathcal{F}[x[n]*y[n]]&=\sum_{n=-\infty}^{+\infty}\left[\sum_{m=-\infty}^{+\infty}x[m]y[n-m]\right]e^{-\mathrm{i}n\omega}\\&=\sum_{m=-\infty}^{+\infty}x[m]\left[\sum_{n=-\infty}^{+\infty}y[n-m]e^{-\mathrm{i}(n-m)\omega}\right]e^{-\mathrm{i}m\omega}\\&=X(\mathrm{i}\omega)\,Y(\mathrm{i}\omega)\end{aligned} \quad (7.1.678)$$

Proof of (b):

$$\begin{aligned}\mathcal{F}[x[n]y[n]]&=\sum_{n=-\infty}^{+\infty}x[n]y[n]e^{-\mathrm{i}n\omega}=\sum_{n=-\infty}^{+\infty}\left[\frac{1}{2\pi}\int\limits_{0}^{2\pi}X(\mathrm{i}\omega')e^{\mathrm{i}n\omega'}\mathrm{d}\omega'\right]y[n]e^{-\mathrm{i}n\omega}\\&=\frac{1}{2\pi}\int\limits_{0}^{2\pi}X(\mathrm{i}\omega')\left[\sum_{n=-\infty}^{+\infty}e^{\mathrm{i}n\omega'}y[n]e^{-\mathrm{i}n\omega}\right]\mathrm{d}\omega'\\&=\frac{1}{2\pi}\int\limits_{0}^{2\pi}X(\mathrm{i}\omega')\sum_{n=-\infty}^{+\infty}y[n]e^{-\mathrm{i}n(\omega-\omega')}\mathrm{d}\omega'\\&=\frac{1}{2\pi}\int\limits_{0}^{2\pi}X(\mathrm{i}\omega')Y(\mathrm{i}(\omega-\omega'))\mathrm{d}\omega'=X(\mathrm{i}\omega)*Y(\mathrm{i}\omega)\end{aligned} \quad (7.1.679)$$

■ Q.E.D.

P8. Parseval's relation for the discrete Fourier transform:

$$\sum_{n=-\infty}^{+\infty}|x[n]|^2=\frac{1}{2\pi}\int\limits_{0}^{2\pi}|X(e^{\mathrm{i}\omega})|^2\mathrm{d}\omega \quad (7.1.680)$$

✍**Examples:**

E1. The discrete Fourier transform of the Dirac delta:

$$\mathcal{F}[\delta[n]]=\sum_{n=-\infty}^{+\infty}\delta[n]e^{-\mathrm{i}\omega n}=1 \quad (7.1.681)$$

E2. The spectrum of:

$$x[n]=a^n u[n]\quad(|a|<1) \quad (7.1.682)$$

is:

$$X(e^{\mathrm{i}\omega})=\mathcal{F}[x[n]]=\sum_{n=-\infty}^{+\infty}a^n u[n]e^{-\mathrm{i}n\omega}=\sum_{n=0}^{+\infty}(ae^{-\mathrm{i}\omega})^n=\frac{1}{1-ae^{-\mathrm{i}\omega}}\tag{7.1.683}$$

If the signal is two-sided:

$$x[n]=a^{|n|}=a^n u[n]+a^{-n}u[-n]-\delta[n],\quad (|a|<1)\tag{7.1.684}$$

Due to the time reversal property, its spectrum is:

$$X(e^{\mathrm{i}\omega})=\frac{1}{1-ae^{-\mathrm{i}\omega}}+\frac{1}{1-ae^{\mathrm{i}\omega}}-1=\frac{1-a^2}{1-2a\,\cos(\omega)+a^2}\tag{7.1.685}$$

E3. Consider a LTI system (Linear Time-Invariant) with impulse response:

$$h[n]=a^n u[n],\quad (|a|<1)\tag{7.1.686}$$

and input:

$$x[n]=b^n u[n],\quad (|b|<1)\tag{7.1.687}$$

The output $y[n]$ can be found in either time domain by convolution or in frequency domain by multiplication. In time domain, we have:

$$\begin{aligned}y[n]&=h[n]*x[n]=\sum_{m=-\infty}^{+\infty}a^{n-m}u[n-m]b^m u[m]=a^n\sum_{m=0}^{n}a^{-m}b^m\\&=a^n\frac{1-(b/a)^{n+1}}{1-(b/a)}u[n]=\frac{1}{a-b}(a^{n+1}-b^{n+1})u[n]\end{aligned}\tag{7.1.688}$$

When $a=b$, we have:

$$y[n]=a^n\sum_{m=0}^{n}a^{-m}b^m=(n+1)a^n u[n]\tag{7.1.689}$$

In frequency domain, we first find the spectra of both $x[n]$ and $h[n]$ to be:

$$X(e^{\mathrm{i}\omega})=\frac{1}{1-ae^{-\mathrm{i}\omega}};\quad H(e^{\mathrm{i}\omega})=\frac{1}{1-be^{-\mathrm{i}\omega}}\tag{7.1.690}$$

and the spectrum of the output is:

$$Y(e^{\mathrm{i}\omega})=H(e^{\mathrm{i}\omega})\,X(e^{\mathrm{i}\omega})=\frac{1}{(1-be^{-\mathrm{i}\omega})(1-ae^{-\mathrm{i}\omega})}\tag{7.1.691}$$

To find $y(n)$ in time domain by inverse transform of $Y(e^{\mathrm{i}\omega})$, we use partial fraction expansion to rewrite the above as:

$$Y(e^{\mathrm{i}\omega})=\frac{A}{1-ae^{-\mathrm{i}\omega}}+\frac{B}{1-be^{-\mathrm{i}\omega}}=\frac{A-Abe^{-\mathrm{i}\omega}+B-aBe^{-\mathrm{i}\omega}}{(1-be^{-\mathrm{i}\omega})(1-ae^{-\mathrm{i}\omega})}\tag{7.1.692}$$

By equating the coefficients of $e^{-\mathrm{i}\omega}$ and the constants, we get:

$$A+B=1,\quad aB+bA=0 \tag{7.1.693}$$

which can be solved to get:

$$A=\frac{a}{a-b},\qquad B=\frac{-b}{a-b} \tag{7.1.694}$$

In this form, $Y(\mathrm{i}\omega)$ can be easily inverse transformed to yield:

$$h[n]=\left[\frac{a}{a-b}a^n-\frac{b}{a-b}b^n\right]u[n]=\frac{1}{a-b}(a^{n+1}-b^{n+1})u[n] \tag{7.1.695}$$

same as the result from convolution. Again when $a=b$, we have:

$$Y(e^{\mathrm{i}\omega})=\frac{1}{(1-ae^{-\mathrm{i}\omega})^2}=\frac{e^{\mathrm{i}\omega}}{a}\mathrm{i}\frac{\mathrm{d}}{\mathrm{d}\omega}\left(\frac{1}{1-ae^{-\mathrm{i}\omega}}\right) \tag{7.1.696}$$

But since:

$$\mathcal{F}^{-1}\left[\frac{1}{1-ae^{-\mathrm{i}\omega}}\right]=a^n u[n] \tag{7.1.697}$$

by the frequency differentiation property, we have:

$$\mathcal{F}^{-1}\left[\mathrm{i}\frac{\mathrm{d}}{\mathrm{d}\omega}\left(\frac{1}{1-ae^{-\mathrm{i}\omega}}\right)\right]=na^n u[n] \tag{7.1.698}$$

and the output in time domain is obtained as:

$$\begin{aligned}y[n]&=\mathcal{F}^{-1}[Y(e^{\mathrm{i}\omega})]=\mathcal{F}^{-1}\left[\frac{e^{\mathrm{i}\omega}}{a}\mathrm{i}\frac{\mathrm{d}}{\mathrm{d}\omega}\left(\frac{1}{1-ae^{-\mathrm{i}\omega}}\right)\right]\\&=\frac{1}{a}(n+1)a^{n+1}u[n+1]=(n+1)a^n u[n+1]=(n+1)a^n u[n]\end{aligned} \tag{7.1.699}$$

Note that the time-shifting property is used due to the factor $e^{\mathrm{i}\omega}$. Also note that $u[n+1]$ (starting at $n=-1$) is replaced by $u[n]$ (starting at $n=0$) as $n+1=0$ when $n=-1$.

E4. The impulse response of a discrete LTI system is:

$$h[m]=a^m u[m] \tag{7.1.700}$$

where $|a|<1$ so that the system is stable. The output $y[m]$ of the system with an input:

$$x[m]=\cos\left(\frac{2\pi m}{N}\right)=\frac{e^{\mathrm{i}2\pi m/N}+e^{-\mathrm{i}2\pi m/N}}{2} \tag{7.1.701}$$

can be found in three different ways:

- Time domain convolution:

  The output is the convolution of $x[m]$ and $h[m]$:

$$\begin{aligned} y[m] = h[m] * x[m] &= \sum_{k=0}^{+\infty} a^k \frac{e^{i2\pi(m-k)/N} + e^{-i2\pi(m-k)/N}}{2} \\ &= \frac{1}{2} e^{i2\pi m/N} \sum_{k=0}^{+\infty} a^k e^{-i2\pi k/N} + \frac{1}{2} e^{-i2\pi m/N} \sum_{k=0}^{+\infty} a^k e^{i2\pi k/N} \\ &= \frac{1}{2} e^{i2\pi m/N} \frac{1}{1-ae^{-i2\pi/N}} + \frac{1}{2} e^{-i2\pi m/N} \frac{1}{1-ae^{i2\pi/N}} \end{aligned} \tag{7.1.702}$$

- The eigenequation method:

  We first get the frequency response function from $h[m]$:

$$H(e^{i\omega}) = \sum_{k=-\infty}^{+\infty} h[k] e^{-i\omega k} = \sum_{k=0}^{+\infty} (ae^{-i\omega})^k = \frac{1}{1-ae^{-i\omega}} \tag{7.1.703}$$

  which is the eigenvalue of the system when the input is a complex exponential $e^{in\omega}$. Now the system's response to:

$$x[m] = \cos\left(\frac{2\pi m}{N}\right) = \frac{e^{i2\pi m/N} + e^{-i2\pi m/N}}{2} \tag{7.1.704}$$

  can be found to be:

$$\begin{aligned} y[m] &= \frac{1}{2}\left[H(e^{i2\pi/N}) e^{i2\pi m/N} + H(e^{-i2\pi/N}) e^{-i2\pi m/N}\right] \\ &= \frac{1}{2} e^{i2\pi m/N} \frac{1}{1-ae^{-i2\pi/N}} + \frac{1}{2} e^{-i2\pi m/N} \frac{1}{1-ae^{i2\pi/N}} \end{aligned} \tag{7.1.705}$$

- Frequency domain multiplication:

  If we find the spectra of both $h[m]$ and $x[m]$ in the frequency domain, the spectrum of $y[m]$ can be found by multiplication. We already know:

$$H(e^{i\omega}) = \mathcal{F}[h[m]] = \frac{1}{1-ae^{-i\omega}} \tag{7.1.706}$$

  We next find the spectrum of $x[m]$:

$$\begin{aligned} X(e^{i\omega}) = \mathcal{F}[x[m]] &= \sum_{m=-\infty}^{+\infty} \frac{e^{i2\pi m/N} + e^{-i2\pi m/N}}{2} e^{-im\omega} \\ &= \pi \sum_{k=-\infty}^{+\infty} [\delta(\omega - 2k\pi - 2\pi/N) + \delta(\omega - 2k\pi - 2\pi/N)] \end{aligned} \tag{7.1.707}$$

  Now the spectrum of the output $y[m]$ can be found:

$$\begin{aligned} Y(e^{i\omega}) &= H(e^{i\omega}) X(e^{i\omega}) \\ &= \frac{\pi}{1-ae^{-i\omega}} \sum_{k=-\infty}^{+\infty} [\delta(\omega - 2k\pi - 2\pi/N) + \delta(\omega - 2k\pi - 2\pi/N)] \end{aligned} \tag{7.1.708}$$

and the output $y[m]$ is obtained by inverse Fourier transform:

$$\begin{aligned} y[m] &= \frac{1}{2\pi}\int\limits_0^{2\pi}\left[\frac{\pi}{1-ae^{-\mathrm{i}\omega}}\sum_{k=-\infty}^{+\infty}\left[\delta\left(\omega-2k\pi-\frac{2\pi}{N}\right)+\delta\left(\omega-2k\pi-\frac{2\pi}{N}\right)\right]\right]e^{\mathrm{i}m\omega}\mathrm{d}\omega \\ &= \frac{1}{2}e^{\mathrm{i}2\pi m/N}\frac{1}{1-ae^{-\mathrm{i}2\pi/N}}+\frac{1}{2}e^{-\mathrm{i}2\pi m/N}\frac{1}{1-ae^{\mathrm{i}2\pi/N}} \end{aligned}$$

The physical meaning of this result will be clear if we write $H(2\pi/N)$ in polar form:

$$H(e^{\mathrm{i}2\pi/N})=\frac{1}{1-ae^{\mathrm{i}2\pi/N}}=re^{\mathrm{i}\theta} \tag{7.1.709}$$

and the output becomes:

$$y[m]=r\cos\left(\frac{2\pi}{N}m+\theta\right) \tag{7.1.710}$$

That is, the output of the system is also a sinusoidal signal of the same frequency as the input, but with different magnitude r and a phase angle θ. For example, if $N=4$, we have:

$$H(e^{\mathrm{i}\pi/2})=\frac{1}{1+\mathrm{i}a}=\frac{1}{\sqrt{1+a^2}}e^{-\mathrm{i}\tan^{-1}(a)} \tag{7.1.711}$$

and the output is:

$$y[m]=\frac{1}{\sqrt{1+a^2}}\cos\left(\frac{\pi m}{2}-\tan^{-1}(a)\right) \tag{7.1.712}$$

To finish depending on donations

7.1.5.2 Laplace Transform

The Laplace transform (LT), as we have already mention it, is a generalization of the Fourier transform (TF), however, although it is so called in his honour because he used it in his work on the theory of probability, seems to have been originally discovered by Leonhard Euler. The Laplace transform also appears in all branches of mathematical physics (mechanical engineering, electronics, quantitative finance, etc.) and is used extensively in order to solve differential equations that arise in many modelling situations of real life.

As for the Fourier transform, the Laplace transformation allows us to get rid of differentiations. The transform can do this because it has the wonderful property of converting the operation of differentiation into the far simpler one of multiplication. That is, it transforms a differential equation into an algebraic equation. This process is analogous to how logarithms transform multiplication into the simpler operation of addition (because $\log(xy)=\log(x)+\log(y)$).

A power series about an origin is any series that can be written in the following form:

$$\sum_{n=0}^{+\infty} a_n x^n \tag{7.1.713}$$

where a_n are numbers and n is a non-negative integer. One can think of $a_n=a(n)$ as a function of n for each non-negative integer $n=0,1,2,\ldots$. In order to give birth to Laplace transformation technique, we make some associations. The discrete variable n is converted into a real variable t. The coefficient term a_n is written as $f(t)$. The term x^n can equivalently be written as $e^{(\ln(x^t))}$. Finally, summation notation can be replaced by its continuous analogue, that is, integration. By doing so, we have following:

$$\sum_{n=0}^{+\infty} a_n x^n \rightsquigarrow \int\limits_0^{+\infty} f(t)e^{\ln(x^t)}\,\mathrm{d}t \tag{7.1.714}$$

For convergence, it is obviously important to have following condition for the above integral (yes think for this about the original sum!):

$$\sum_{n=0}^{+\infty} a_n x^n \rightsquigarrow \int\limits_0^{+\infty} f(t)e^{\ln(x^t)}\mathrm{d}t; \qquad 0<x<1 \tag{7.1.715}$$

Therefore:

$$\sum_{n=0}^{+\infty} a_n x^n \rightsquigarrow \int\limits_0^{+\infty} f(t)e^{t\ln(x)}\mathrm{d}t \tag{7.1.716}$$

Since $0<x<1$ so it implies that:

$$\ln(0)<\ln(x)<\ln(1) \quad \text{or} \quad -\infty<\ln(x)<0 \tag{7.1.717}$$

Thus $\ln(x)$ has to be negative for the integral to converge, in this regard, we suppose $\ln(x)=-s$ where $s>0$. Thus, the final integral takes the form:

$$\sum_{n=0}^{+\infty} a_n x^n \rightsquigarrow \int\limits_0^{+\infty} f(t)e^{-st}\mathrm{d}t=F(s) \quad s>0 \tag{7.1.718}$$

In this way, we can say that Laplace Transform is simply stretching a discrete (infinite series) into a continuous (integration) analogue.

Let us recall that if $s=\mathrm{i}\omega$ (the real part of s is purely imaginary), Laplace transform becomes Fourier transform! In general, any continuous time signal $x(t)$ can be Laplace transformed to get:

$$\boxed{X(s)=\mathcal{L}(x(t))\triangleq\int\limits_{-\infty}^{+\infty}x(t)e^{-st}\mathrm{d}t} \tag{7.1.719}$$

provided the integral converges, i.e., the function $X(s)$ exists. This general form of "bilateral Laplace transform" is related to the Fourier transform:

$$\mathcal{L}(x(t))=\int\limits_{-\infty}^{+\infty}x(t)e^{-(\sigma+\mathrm{i}\omega)t}\mathrm{d}t=\int\limits_{-\infty}^{+\infty}[x(t)e^{-\sigma t}]e^{-\mathrm{i}\omega t}\mathrm{d}t=\mathcal{F}[x(t)e^{-\sigma t}] \tag{7.1.720}$$

i.e., Laplace transform of a generic function is Fourier transform of the same function multiplied by $e^{-\sigma t}$. This exponential factor has the effect of forcing the signals to converge. This is why the Laplace transform can be applied to a broader class of signals than the Fourier transform, including exponentially growing signals shown in the following two cases:

- Right sided:

  $$x(t)=e^{at}u(t),\quad (a>0) \tag{7.1.721}$$

  The Fourier transform does not exist as the signal grows exponentially when $t\rightarrow+\infty$, i.e., the transform integral does not converge (not integrable). However, its Laplace transform exists if $\Re[s]=\sigma>a$ (i.e., $\sigma-a>0$), as the modified signal $x(t)e^{-\sigma t}=e^{-(\sigma-a)t}u(t)$ will converge.

- Left sided:

  $$x(t)=e^{at}u(-t),\quad (a<0) \tag{7.1.722}$$

  Again the Fourier transform does not exist as the signal grows exponentially when $t\rightarrow-\infty$, i.e., the transform integral does not converge. But if $\Re[s]=\sigma<a$ (i.e., $\sigma-a<0$), the modified signal $x(t)e^{-\sigma t}=e^{-(\sigma-a)t}u(-t)$ will converge and its Laplace transform exists.

In Fourier transform, both the signal $x(t)$ in time domain and its spectrum $X(\mathrm{i}\omega)$ in frequency domain are a one-dimensional (1D) complex function. However, the Laplace transform $X(s)$ of the 1D signal $x(t)$ is a complex function defined over a two-dimensional *complex plane*, named the "s-plane", spanned by the two variables σ (for the horizontal real axis) and ω (for the vertical imaginary axis).

In particular, if this 2D function $X(s)=X(\sigma+\mathrm{i}\omega)$ is evaluated along the imaginary axis $\Re[s]=\sigma=0$, it becomes a 1D function $X(\mathrm{i}\omega)$, the Fourier transform of $x(t)$. Graphically, the Fourier transform, the spectrum of the signal, can be found as the cross section of the 2D function $X(s)=X(\sigma+\mathrm{i}\omega)$ along the line $\Re[s]=\sigma=0$.

Before we go further another way to introduce the Laplace transform may be useful and help the reader. But if you have understand the above introduction you can go over this alternate presentation.

The Fourier transform works quite well only when the argument of the integral does not diverge, where the kernel (the multiplication by $e^{-\mathrm{i}\omega t}$) is purely complex, and its integral sweeps the set of reals (bilateral infinity $]-\infty,+\infty[$).

Typically, mathematicians, that like to generalize stuff...., have generalized the Fourier transform to functions kernel that are not purely imaginary and whose integral could be one-sided $[0,+\infty[$. Thus, the Laplace transform converges for a larger set of functions than the Fourier transform.

Indeed, let us see a simple example with a diverging function:

$$\mathcal{F}(e^t)(\omega)=\int\limits_{-\infty}^{+\infty} e^t e^{-\mathrm{i}\omega t}\mathrm{d}t=\int\limits_{-\infty}^{+\infty} e^{t-\mathrm{i}\omega t}\mathrm{d}t=\int\limits_{-\infty}^{+\infty} e^{(1-\mathrm{i}\omega)t}\mathrm{d}t=\left.\frac{e^{\alpha t}}{\alpha}\right|_{-\infty}^{+\infty}=+\infty \tag{7.1.723}$$

This last integral does not converge (at least as far as I know ...) whatever the value of α different from zero!

We can define a generalized version of the Fourier transform and denote it (we already know that it is the bilateral Laplace transform):

$$\mathcal{L}(f(t))(s)=\int\limits_{-\infty}^{+\infty} f(t)e^{-st}\mathrm{d}t \tag{7.1.724}$$

with (we already know this!):

$$s=\sigma+\mathrm{i}\omega t \tag{7.1.725}$$

where, in general, the convergence of the integral is not guaranteed for all s. We then name "abscissa of absolute convergence" of the Laplace transform the set of values that σ must take for the integral to converge.

Let's revisit the non-convergent Fourier transform seen just above, but where the function is defined only for positive times (one-sided Laplace transform):

$$\mathcal{L}(e^t)(s)=\int\limits_{-\infty}^{+\infty} e^t e^{-st}\mathrm{d}t=\int\limits_{0}^{+\infty} e^t e^{-st}\mathrm{d}t=\int\limits_{0}^{+\infty} e^t e^{-(\sigma+\mathrm{i}\omega)t}\mathrm{d}t=\int\limits_{0}^{+\infty} e^{-((\sigma-1)+\mathrm{i}\omega)t}\mathrm{d}t=\left.\frac{e^{-t((\sigma-1)+\mathrm{i}\omega)}}{-((\sigma-1)+\mathrm{i}\omega)}\right|_{0}^{+\infty} \tag{7.1.726}$$

The result converges if $\omega\geq 0$ (which is always the case in physics) for $t>0$ and if and only if

$\sigma > 1$. Then in this case we have:

$$\begin{aligned}\mathcal{L}(e^t)(s) &= \left.\frac{e^{-t((\sigma-1)+\mathrm{i}\omega)}}{-((\sigma-1)+\mathrm{i}\omega)}\right|_0^{+\infty} = \left.\frac{e^{(1-\sigma)t}e^{-\mathrm{i}\omega t}}{-((\sigma-1)+\mathrm{i}\omega)}\right|_0^{+\infty} = \frac{\underbrace{e^{(1-\sigma)\cdot+\infty}}_{\to 0}\underbrace{e^{-\mathrm{i}\omega\cdot+\infty}}_{\to 0}}{-((\sigma-1)+\mathrm{i}\omega)} - \frac{e^{(1-\sigma)\cdot 0}e^{-\mathrm{i}\omega\cdot 0}}{-((\sigma-1)+\mathrm{i}\omega)}\\ &= 0 - \frac{1}{-((\sigma-1)+\mathrm{i}\omega)} = \frac{1}{(\sigma+\mathrm{i}\omega)-1} = \frac{1}{s-1}\end{aligned} \tag{7.1.727}$$

7.1.5.2.1 Inverse Laplace Transform

The inverse Laplace transform can be obtained from the corresponding Fourier transform:

$$X(s) = X(\sigma+\mathrm{i}\omega) = \mathcal{L}(x(t)) = \int\limits_{-\infty}^{+\infty} x(t)e^{-(\sigma+\mathrm{i}\omega)t}\mathrm{d}t = \mathcal{F}(x(t)e^{-\sigma t}) \tag{7.1.728}$$

The inverse Fourier transform of the above is:

$$x(t)e^{-\sigma t} = \mathcal{F}^{-1}(X(\sigma+\mathrm{i}\omega)) = \frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} X(\sigma+\mathrm{i}\omega)e^{\mathrm{i}\omega t}\mathrm{d}\omega \tag{7.1.729}$$

Multiplying both sides by $e^{\sigma t}$, we get the inverse Laplace transform:

$$x(t) = \mathcal{L}^{-1}[X(s)] = \frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} X(\sigma+\mathrm{i}\omega)e^{(\sigma+\mathrm{i}\omega)t}\mathrm{d}\omega \tag{7.1.730}$$

As we want to represent the inverse transform in terms of s (instead of ω), we realize that:

$$\mathrm{d}s = \mathrm{d}(\sigma+\mathrm{i}\omega) = \mathrm{i}\,\mathrm{d}\omega, \quad \text{ie,} \quad \mathrm{d}\omega = \frac{\mathrm{d}s}{\mathrm{i}} \tag{7.1.731}$$

We also realize that the integral in the inverse transform is along a vertical line in the s-plane from $\mathrm{i}\omega = -\mathrm{i}\infty$ to $\mathrm{i}\omega = +\mathrm{i}\infty$ while the other variable σ is fixed. The "inverse Laplace transform" above can therefore be written as:

$$\boxed{x(t) = \mathcal{L}^{-1}(X(s)) = \frac{1}{\mathrm{i}2\pi}\int\limits_{\sigma-\mathrm{i}\infty}^{\sigma+\mathrm{i}\infty} X(s)e^{st}\mathrm{d}s} \tag{7.1.732}$$

Now we have the Laplace transform pair:

$$X(s) = \mathcal{L}(x(t)) = \int\limits_{-\infty}^{+\infty} x(t)e^{-st}\mathrm{d}t \tag{7.1.733}$$

$$x(t) = \mathcal{L}^{-1}(X(s)) = \frac{1}{\mathrm{i}2\pi}\int\limits_{\sigma-\mathrm{i}\infty}^{\sigma+\mathrm{i}\infty} X(s)e^{st}\mathrm{d}s \tag{7.1.734}$$

The forward (bilateral) and inverse Laplace transform pair can also be represented as:

$$x(t) \xleftrightarrow{\mathcal{L}} X(s) \tag{7.1.735}$$

7.1.5.2.2 Region of Convergence

An essential issue of Laplace transform of $x(t)$ is whether the transform $X(s)$ even exists, and under what condition it exists. To see this, consider the following examples.

✍**Examples:**

E1. The Fourier transform of a signal $x(t)=e^{-at}u(t)$ is:

$$X(\mathrm{i}\omega)=\int\limits_{-\infty}^{+\infty} e^{-at}u(t)e^{-\mathrm{i}\omega t}\mathrm{d}t=\int\limits_{0}^{+\infty} e^{-(a+\mathrm{i}\omega)t}\mathrm{d}t=-\frac{1}{a+\mathrm{i}\omega}e^{-(a+\mathrm{i}\omega)}|_0^{+\infty} \tag{7.1.736}$$

This integral does not converge unless $a>0$. In other words, only when the signal $x(t)$ decays (instead of grows) exponentially, will its Fourier transform $X(\mathrm{i}\omega)$ exist:

$$X(\mathrm{i}\omega)=\frac{1}{a+\mathrm{i}\omega}\quad a<0 \tag{7.1.737}$$

Now consider the Laplace transform of the same signal:

$$X(s)=\int\limits_{0}^{+\infty} e^{-at}e^{-st}\mathrm{d}t=\int\limits_{0}^{+\infty} e^{-at}e^{-(\sigma+\mathrm{i}\omega)t}\mathrm{d}t=\int\limits_{0}^{+\infty} e^{-(a+\sigma)t}e^{-\mathrm{i}\omega t}\mathrm{d}t \tag{7.1.738}$$

Similar to the Fourier transform, for this integral to converge, i.e., for Laplace transform $X(s)$ to exist, it is necessary for $\sigma=\Re[s]$ to satisfy:

$$a+\sigma>0\quad\text{or}\quad\Re[s]=\sigma>-a \tag{7.1.739}$$

in which case the Laplace transform is:

$$X(s)=\frac{1}{(\sigma+a)+\mathrm{i}\omega}=\frac{1}{s+a}\quad\Re[s]>-a \tag{7.1.740}$$

As a special case where $a=0$, $x(t)=u(t)$ we have:

$$\mathcal{L}(u(t))=\frac{1}{s}\quad\Re[s]>0 \tag{7.1.741}$$

When $\Re[s]=\sigma=0$, the Laplace transform $X(s)$ becomes Fourier transform $X(\mathrm{i}\omega)$.

E2. The non-causal version of the signal above is $x(t)=-e^{-at}u(-t)$ and its Laplace transform is:

$$X(s)=-\int\limits_{-\infty}^{0} e^{-at}e^{-st}\mathrm{d}t=-\int\limits_{-\infty}^{0} e^{-((a+\sigma)+\mathrm{i}\omega)t}\mathrm{d}t=\frac{1}{a+\sigma+\mathrm{i}\omega}e^{-((a+\sigma)+\mathrm{i}\omega)t}|_{-\infty}^{0} \tag{7.1.742}$$

Only when:

$$a+\sigma<0\quad\text{or}\quad\Re[s]=\sigma<-a \tag{7.1.743}$$

will this integral converge and Laplace transform $X(s)$ exists:

$$X(s)=\frac{1}{a+\sigma+\mathrm{i}\omega}=\frac{1}{a+s} \quad \Re[s]<-a \tag{7.1.744}$$

Again as a special case when $a=0$, $x(t)=-u(-t)$ we have:

$$\mathcal{L}[-u(-t)]=\frac{1}{s} \quad \Re[s]<0 \tag{7.1.745}$$

Comparing the two examples above we see that two different signals can have identical Laplace transform $X(s)$ but s may have to satisfy different conditions for $X(s)$ to exist. In general, the set of all s values satisfying the conditions for the integral of Laplace transform to converge is called the "region of convergence" (ROC) in the complex s-plane. In the first case above, the ROC is $\Re[s]>0$, and in the second case, the ROC is $\Re[s]<0$.

E3. Consider the signal:

$$x(t)=\left[e^{-2t}+e^{t}\cos(3t)\right]u(t)=\left[e^{-2t}+\frac{1}{2}e^{-(1-\mathrm{i}3)t}+\frac{1}{2}e^{-(1+\mathrm{i}3)t}\right]u(0) \tag{7.1.746}$$

The Laplace transform of this signal is:

$$\begin{aligned}X(s)&=\int\limits_0^{+\infty}\left[e^{-2t}+\frac{1}{2}e^{-(1-\mathrm{i}3)t}+\frac{1}{2}e^{-(1+\mathrm{i}3)t}\right]e^{-st}\mathrm{d}t\\&=\int\limits_0^{+\infty}e^{-2t}e^{-st}\mathrm{d}t+\frac{1}{2}\int\limits_0^{+\infty}e^{-(1-\mathrm{i}3)t}e^{-st}\mathrm{d}t+\frac{1}{2}\int\limits_0^{+\infty}e^{-(1+\mathrm{i}3)t}e^{-st}\mathrm{d}t\\&=\frac{1}{s+2}+\frac{1/2}{s+(1-\mathrm{i}3)}+\frac{1/2}{s+(1+\mathrm{i}3)}\end{aligned} \tag{7.1.747}$$

Following example 1, we see that the conditions for the three integrals to converge are, respectively:

$$\Re[s]>-2,\quad \Re[s]>-1,\quad \Re[s]>-1 \tag{7.1.748}$$

i.e., when $\Re[s]>-1$, $X(s)$ exists and can be written as:

$$X(s)=\frac{1}{s+2}+\frac{1/2}{s+(1-\mathrm{i}3)}+\frac{1/2}{s+(1+\mathrm{i}3)}=\frac{2s^2+5s+12}{(s^2+2s+10)(s+2)} \quad \Re[s]>-1 \tag{7.1.749}$$

E4. Let us consider:

$$\mathcal{L}(\delta(t-T))=\int\limits_{-\infty}^{+\infty}\delta(t)e^{-st}\mathrm{d}t=e^{-sT} \tag{7.1.750}$$

As the Laplace integration converges independently of s, the ROC is the entire s-plane. In particular, when $T=0$, we have:

$$\mathcal{L}(\delta(t))=1 \tag{7.1.751}$$

7.1.5.2.3 Zeros and Poles of Laplace Transform

All Laplace transforms in the above examples are rational, i.e., they can be written as a ratio of polynomials of variable s in the general form:

$$X(s) = \frac{N(s)}{D(s)} = \frac{\sum_{k=0}^{M} b_k s^k}{\sum_{k=0}^{N} a_k s^k} = \frac{\prod_{k=1}^{M} (s - s_{z_k})}{\prod_{k=1}^{N} (s - s_{p_k})} \tag{7.1.752}$$

where $N(s)$ is the numerator polynomial of order M with roots $s_{z_k}, (k = 1, 2, \cdots, M)$, and $D(s)$ is the denominator polynomial of order N with roots $s_{p_k}, (k = 1, 2, \cdots, N)$. In general, we assume the order of the numerator polynomial is lower than that of the denominator polynomial, i.e., $M < N$. If this is not the case, we can always expand $X(s)$ into multiple terms so that $M < N$ is true for each of terms.

✍**Example:**

Consider:

$$X(s) = \frac{s^2 - 3}{s + 2} \tag{7.1.753}$$

As the order of the numerator $M = 2$ is higher than that of the denominator $N = 1$, we expand it into the following terms:

$$X(s) = \frac{s^2 - 3}{s + 2} = A + Bs + \frac{C}{s + 2} \tag{7.1.754}$$

and get:

$$s^2 - 3 = (A + Bs)(s + 2) + C = Bs^2 + (A + 2B)s + (2A + C) \tag{7.1.755}$$

Equating the coefficients for terms s^k $(k = 0, 1, \ldots, M)$ on both sides, we get:

$$B = 1, \quad A + 2B = 0, \quad 2A + C = -3 \tag{7.1.756}$$

Solving this system of equations, we get the following coefficients:

$$A = -2; \quad B = 1, \quad C = 1 \tag{7.1.757}$$

and:

$$X(s) = s - 2 + \frac{1}{s + 2} \tag{7.1.758}$$

Alternatively, the same result can be obtained by carrying out a long-hand division:

$$(s^2 - 3) \div (s + 2) \tag{7.1.759}$$

Definitions (#7.1.284):

D1. Any complex value s_z of s for which $H(s)|_{s=s_z}=H(s_z)=0$ is a "zero" of $H(s)$.

D2. Any complex value s_p of s for which $H(s)|_{s=s_p}=H(s_p)=\infty$ is a "pole" of $H(s)$.

Obviously, all roots of the numerator polynomial $N(s)$ are zeros of $H(s)$ and all roots of the denominator polynomial $D(s)$ are poles of $H(s)$. Moreover, if the order of $D(s)$ exceeds the order of $N(s)$ (i.e., $N>M$), then $H(\infty)=0$, i.e., there is a zero at infinity. On the other hand, if the order of $N(s)$ exceeds that of $D(s)$ (i.e., $M>N$), then $H(\infty)=\infty$, i.e, there is a pole at infinity. On the s-plane zeros and poles can be indicated by o and x respectively. Most essential behaviour properties of a LTI system can be obtained graphically from the ROC and the zeros and poles of its transfer function $H(s)$ on the s-plane.

7.1.5.2.4 Properties of region of convergence

Whether the Laplace transform $X(s)$ of a function $x(t)$ exists depends on whether or not the transform integral converges:

$$X(s)=\int\limits_{-\infty}^{+\infty}x(t)e^{-st}\mathrm{d}t=\int\limits_{-\infty}^{+\infty}x(t)e^{-\sigma t}e^{-\mathrm{i}\omega t}\mathrm{d}t<\infty \tag{7.1.760}$$

which in turn depends on the duration and magnitude of $x(t)$ as well as the real part of s $\Re[s]=\sigma$. When $x(t)$ is right sided (i.e., $x(t)=0$ for $t<t_0$), it may have infinite duration for $t>0$, and the larger σ the more quickly $x(t)e^{-\sigma t}$ decays as $t\rightarrow+\infty$. On the other hand, if $x(t)$ is left sided (i.e., $x(t)$ for $t<t_0$), it may have infinite duration for $t<0$, and the smaller σ the more quickly $x(t)e^{-\sigma t}$ decays as $t\rightarrow-\infty$. The imaginary part of s $\Im[s]=\mathrm{i}\omega$ determines the frequency of a sinusoid which is bounded and has no effect on the convergence of the integral. Based on these observations, we can get the following properties for the ROC:

- If $x(t)$ is absolutely integrable and of finite duration, then the ROC is the entire s-plane (the Laplace transform integral is finite, i.e., $X(s)$ exists, for any s).
- The ROC of $X(s)$ consists of strips parallel to the $\mathrm{i}\omega$-axis in the s-plane.
- If $x(t)$ is right sided and $\Re[s]=\sigma_0$ is in the ROC, then any s to the right of σ_0 (i.e., $\Re[s]>\sigma_0$) is also in the ROC, i.e., ROC is a right sided half plane.
- If $x(t)$ is left sided and $\Re[s]=\sigma_0$ is in the ROC, then any s to the left of σ_0 (i.e., $\Re[s]<\sigma_0$) is also in the ROC, i.e., ROC is a left sided half plane.
- If $x(t)$ is two-sided, then the ROC is the intersection of the two one-sided ROCs corresponding to the two one-sided parts of $x(t)$. This intersection can be either a vertical strip or an empty set.
- If $X(s)$ is rational, then its ROC does not contain any poles (by definition $X(s)|_{s=s_p}=\infty$ does not exist). The ROC is bounded by the poles or extends to infinity.
- If $X(s)$ is a rational Laplace transform of a right sided function $x(t)$, then the ROC is the half plane to the right of the rightmost pole, and if $X(s)$ is a rational Laplace transform of a left sided function $x(t)$, then the ROC is the half plane to the left of the leftmost pole.

- A signal $x(t)$ is absolutely integrable, i.e., its Fourier transform $X(\mathrm{i}\omega)$ exists, if and only if the ROC of the corresponding Laplace transform $X(s)$ contains the imaginary axis $\Re[s]=0$ or $s=\mathrm{i}\omega$.

✍**Examples:**

E1. Consider the Laplace transform of a two-sided signal $x(t)=e^{-b|t|}$:

$$X(s)=\mathcal{L}(x(t))=\mathcal{L}(e^{-b|t|})=\mathcal{L}(e^{-bt}u(t))+\mathcal{L}(e^{bt}u(-t)) \qquad (7.1.761)$$

The Laplace transform of the two components can be obtained from the two examples discussed earlier above. Then we get:

$$\mathcal{L}(e^{-bt}u(t))=\frac{1}{s+b} \quad \Re[s]>-b \qquad (7.1.762)$$

and let $b=-a$, we then have:

$$\mathcal{L}(e^{bt}u(-t))=\mathcal{L}(e^{-at}u(-t))=-\frac{1}{s+a}=-\frac{1}{s-b} \quad \Re[s]<-a=b \qquad (7.1.763)$$

Combining the two components, we have:

$$\mathcal{L}(e^{-b|t|})=\frac{1}{s+b}-\frac{1}{s-b}=\frac{-2b}{s^2-b^2} \quad -b<\Re[s]<b \qquad (7.1.764)$$

Whether $X(s)$ exists or not depends on b. If $b>0$, i.e., $x(t)$ decays exponentially as $|t|\to+\infty$, then the ROC is the strip between $-b$ and b and $X(s)$ exists. But if $b<0$, i.e., $x(t)$ grows exponentially as $|t|\to+\infty$, then the ROC is an empty set and $X(s)$ does not exist.

E2. Given the following Laplace transform, find the corresponding signal:

$$X(s)=\frac{1}{(s+1)(s+2)}=\frac{1}{s+1}-\frac{1}{s+2} \qquad (7.1.765)$$

Given the two poles $s_{p_1}=-1$ and s_{p_2} of the expression, there are three associated ROCs:

- The half plane to the right of the rightmost pole $s_{p_2}=-1$, with the corresponding right sided time function:

  $$x(t)=e^{-t}u(t)-e^{-2t}u(t) \qquad (7.1.766)$$

- The half plane to the left of the leftmost pole $s_{p_1}=-2$, with the corresponding left sided time function:

  $$x(t)=-e^{-t}u(-t)+e^{-2t}u(-t) \qquad (7.1.767)$$

- The vertical strip between the two poles $-1<\Re[s]<-2$, with the corresponding two sided time function:

  $$x(t)=-e^{-t}u(-t)-e^{-2t}u(t) \qquad (7.1.768)$$

In particular, note that only the first ROC includes the $\mathrm{i}\omega$-axis and the corresponding time function has a Fourier transform. Fourier transform of the other two functions do not exist.

7.1.5.2.5 Properties of Laplace Transform

The Laplace transform has a set of properties in parallel with that of the Fourier transform. The difference is that we need to pay special attention to the ROCs. In the following, we always assume:

$$\mathcal{L}[x(t)]=X(s) \qquad \text{ROC}=R_x \tag{7.1.769}$$

and:

$$\mathcal{L}[y(t)]=Y(s) \qquad \text{ROC}=R_y \tag{7.1.770}$$

P1. Linearity:

$$\mathcal{L}[ax(t)+by(t)]=aX(s)+bY(s) \quad \text{ROC}\supseteq(R_x\cap R_y) \tag{7.1.771}$$

While it is obvious that the ROC of the linear combination of $x(t)$ and $y(t)$ should be the intersection of the their individual ROCs $R_x\cap R_y$ in which both $X(s)$ and $Y(s)$ exist, we note that in some cases when zero-pole cancellation occurs, the ROC of the linear combination could be larger than $R_x\cap R_y$, as shown in the example below.

✍**Example:**

Assume:

$$X(s)=\mathcal{L}[x(t)]=\frac{1}{s+1} \qquad \Re[s]>-1 \tag{7.1.772}$$

and:

$$Y(s)=\mathcal{L}[y(t)]=\frac{1}{(s+1)(s+2)} \qquad \Re[s]>-1 \tag{7.1.773}$$

then:

$$\mathcal{L}[x(t)-y(t)]=\frac{1}{s+1}-\frac{1}{(s+1)(s+2)}=\frac{s+1}{(s+1)(s+2)}=\frac{1}{s+2} \qquad \Re[s]>-2 \tag{7.1.774}$$

P2. Time Shifting:

$$\mathcal{L}[x(t-t_0)]=e^{-t_0 s}X(s) \qquad \text{ROC}=R_x \tag{7.1.775}$$

P3. Shifting in s-Domain:

$$\mathcal{L}[e^{-s_0 t}x(t)]=X(s+s_0) \qquad \text{ROC}=R_x+\Re[s_0] \tag{7.1.776}$$

Note that the ROC is shifted by s_0, i.e., it is shifted vertically by $Im[s_0]$ (with no effect to ROC) and horizontally by $\Re[s_0]$. If R_x is $\Re[s]>0$, the new ROC is $\Re[s+s_0]>0$, i.e., $\Re[s]>-\Re[s_0]$.

P4. Time Scaling:

$$\mathcal{L}[x(at)] = \frac{1}{|a|}X\left(\frac{s}{a}\right) \qquad \text{ROC} = \frac{R_x}{a} \tag{7.1.777}$$

Note that the ROC is horizontally scaled by $1/a$, which could be either positive ($a > 0$) or negative ($a < 0$) in which case both the signal $x(t)$ and the ROC of its Laplace transform are horizontally flipped.

P5. Conjugation:

$$\mathcal{L}[x^*(t)] = X^*(s^*) \qquad \text{ROC} = R_x \tag{7.1.778}$$

Proof 7.3.24.

$$X^*(s^*) = \left[\int\limits_{-\infty}^{+\infty} x(t)e^{-s^*t}\right]^* \mathrm{d}t = \int\limits_{-\infty}^{+\infty} x^*(t)e^{-st}\mathrm{d}t = \mathcal{L}[x^*(t)] \tag{7.1.779}$$

■ Q.E.D.

P6. Convolution:

$$\mathcal{L}[x(t) * y(t)] = X(s)Y(s) \qquad \text{ROC} \supseteq (R_x \cap R_y) \tag{7.1.780}$$

Note that the ROC of the convolution could be larger than the intersection of R_x and R_y, due to the possible pole-zero cancellation caused by the convolution, similar to the linearity property.

✍**Example:**

Assume:

$$X(s) = \mathcal{L}[x(t)] = \frac{s+1}{s+2} \;\; \Re[s] > -2 \;\; \text{and} \quad Y(s) = \mathcal{L}[y(t)] = \frac{s+2}{s+1} \;\; \Re[s] > -1 \tag{7.1.781}$$

then:

$$\mathcal{L}[x(t) * y(t)] = X(s)Y(s) = 1 \;\; \text{ROC} = \text{the entire s-plane} \tag{7.1.782}$$

P7. Differentiation in Time Domain:

$$\mathcal{L}\left[\frac{\mathrm{d}}{\mathrm{d}t}x(t)\right] = sX(s) \quad \text{ROC} \supseteq R_x \tag{7.1.783}$$

This can be proven by differentiating the inverse Laplace transform:

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = \frac{1}{\mathrm{i}2\pi}\int\limits_{\sigma-\mathrm{i}\infty}^{\sigma+\mathrm{i}\infty} X(s)\frac{\mathrm{d}}{\mathrm{d}t}e^{st}\mathrm{d}s = \frac{1}{\mathrm{i}2\pi}\int\limits_{\sigma-\mathrm{i}\infty}^{\sigma+\mathrm{i}\infty} sX(s)e^{st}\mathrm{d}s \tag{7.1.784}$$

Again, multiplying $X(s)$ by s may cause pole-zero cancellation and therefore the resulting ROC may be larger than R_x. For example, when $x(t)=u(t)$ and $X(s)=1/s$ with $\Re[s]>0$ as the ROC, $\mathrm{d}x(t)/\mathrm{d}t=\delta(t)$ and with $sX(s)=1$ whose ROC is the entire s-plane. In general, we have:

$$\mathcal{L}\left[\frac{\mathrm{d}^n}{\mathrm{d}t^n}x(t)\right]=s^nX(s) \tag{7.1.785}$$

✍**Example:**

The ROC of $\mathcal{L}[\delta(t)]=1$ is the entire s-plane, and we have:

$$\mathcal{L}\left[\frac{d}{\mathrm{d}t}\delta(t)\right]=s \qquad \text{ROC}=\text{entire s-plane} \tag{7.1.786}$$

and more generally:

$$\mathcal{L}\left[\frac{d^n}{\mathrm{d}t^n}\delta(t)\right]=s^n \qquad \text{ROC}=\text{entire s-plane} \tag{7.1.787}$$

P8. Differentiation in s-Domain:

$$\mathcal{L}[tx(t)]=-\frac{\mathrm{d}}{\mathrm{d}s}X(s) \qquad \text{ROC}=R_x \tag{7.1.788}$$

This can be proven by differentiating the Laplace transform:

$$\frac{\mathrm{d}}{\mathrm{d}s}X(s)=\int\limits_{-\infty}^{+\infty}x(t)\frac{\mathrm{d}}{\mathrm{d}s}e^{-st}\mathrm{d}t=\int\limits_{-\infty}^{+\infty}(-t)x(t)e^{-st}\mathrm{d}t \tag{7.1.789}$$

Repeat this process we get:

$$\mathcal{L}[t^nx(t)]=(-1)^n\frac{\mathrm{d}^n}{\mathrm{d}s^n}X(s) \qquad \text{ROC}=R_x \tag{7.1.790}$$

P9. Integration in Time Domain:

$$\mathcal{L}\left[\int\limits_{-\infty}^{t}x(\tau)\mathrm{d}\tau\right]=\frac{X(s)}{s} \qquad \text{ROC}\supseteq(R_x\cap\{\Re[s]>0\}) \tag{7.1.791}$$

This can be proven by realizing that:

$$x(t)*u(t)=\int\limits_{-\infty}^{+\infty}x(\tau)u(t-\tau)\mathrm{d}\tau=\int\limits_{-\infty}^{t}x(\tau)\mathrm{d}\tau \tag{7.1.792}$$

and therefore by convolution property we have:

$$\mathcal{L}[x(t)*u(t)]=X(s)\frac{1}{s} \tag{7.1.793}$$

Also note that as the ROC of $\mathcal{L}[u(t)]=1/s$ is the right half plane $\Re[s]>0$, the ROC of $X(s)/s$ is the intersection of the two individual ROCs $R_x\cap\{\Re[s]>0\}$, except if pole-zero cancellation occurs (when $x(t)=\mathrm{d}\delta(t)/\mathrm{d}t$ with $X(s)=s$) in which case the ROC is the entire s-plane.

Property name	**Time domain function $f(t)$**	**Laplace Transform $F(s)$**
Linearity	$af(t)+bg(t)$	$aF(s)+bG(s)$
Scale change	$f(at)$	$\dfrac{1}{a}F\left(\dfrac{s}{a}\right)$
Shift	e^{-at}	$F(s+a)$
Delay	$f(t-a)$	$e^{-as}F(s)$
Derivation	$\dfrac{\mathrm{d}f(t)}{\mathrm{d}t}$	$sF(s)-f(0)$
n^{th} Derivation	$\dfrac{\mathrm{d}^nf(t)}{\mathrm{d}t^n}$	$s^nf(s)-s^{n-1}f(0)-s^{n-2}f^1(0)-\ldots-f^{n-1}(0)$
Power	$t^n\dfrac{\mathrm{d}f(t)}{\mathrm{d}t}$	$(-1)^n\dfrac{\mathrm{d}^nF(s)}{\mathrm{d}s^n}$
Integration	$\displaystyle\int\limits_0^t f(x)\mathrm{d}x$	$\dfrac{1}{s}F(s)$
Reciprocal	$\dfrac{1}{t}f(t)$	$\displaystyle\int\limits_s^{+\infty} F(x)\mathrm{d}x$
Convolution	$f(t)*g(t)$	$F(s)\cdot G(s)$
Periodic function	$f(t)=f(t+T)$	$\dfrac{1}{1-e^{-sT}}\displaystyle\int\limits_0^T e^{-sx}f(x)\mathrm{d}x$

Table 7.9: **Summary of most common Laplace transform properties**

7.1.5.2.6 Usual Laplace transforms

As we always do in this textbook, let us see some common usual Laplace Transforms that are useful in some physics and finance well known problems (and also used as example in the MATLAB™ companion book):

1. $\delta(t)$, $\delta(t-\tau)$

$$\mathcal{L}[\delta(t)]=\int\limits_{-\infty}^{+\infty}\delta(t)e^{-st}\mathrm{d}t=e^0=1,\forall s \tag{7.1.794}$$

   Moreover, due to time shifting property, we have:

$$\mathcal{L}[\delta(t-\tau)]=e^{-s\tau},\forall s \tag{7.1.795}$$

2. $u(t)$ (Heaviside function), $t\,u(t)$, $t^n\,u(t)$

   Due to the property of time domain integration, we have:

$$\mathcal{L}[u(t)]=\mathcal{L}\left[\int\limits_{-\infty}^{t}\delta(\tau)d\tau\right]=\frac{1}{s},\quad \Re[s]>0 \tag{7.1.796}$$

   Let us have another more explicit approach as the previous one may be a bit hard.... The function $u(t)$ (for recall it is the Heaviside function defined as $H(t)=0,\forall t<0$) for the usage of the unilateral Laplace transform as it non-null only for $t\geq 0$. Then we have:

$$\mathcal{L}(H(t))(s)=\int\limits_0^{+\infty} H(t)e^{-st}\mathrm{d}t=\int e^{-st}\mathrm{d}t=\left.\frac{e^{-st}}{-s}\right|_0^{+\infty} \tag{7.1.797}$$

This integral converge only if $\sigma\geq 0$. Hence:

$$\mathcal{L}(H(t))(s)=\left.\frac{e^{-st}}{-s}\right|_0^{+\infty}=\frac{\overbrace{e^{-\sigma\cdot+\infty}}^{}_{\to 0}\overbrace{e^{-\mathrm{i}\omega\cdot+\infty}}^{}_{\to 0}}{-(\sigma+\mathrm{i}\omega)}-\frac{e^{\sigma\cdot 0}e^{-\mathrm{i}\omega\cdot 0}}{-(\sigma+\mathrm{i}\omega)}=0-\frac{1}{-s}=\frac{1}{s} \tag{7.1.798}$$

Applying the s-domain differentiation property to the above, we have:

$$\mathcal{L}[tu(t)]=-\frac{\mathrm{d}}{\mathrm{d}s}\left[\frac{1}{s}\right]=\frac{1}{s^2},\quad \Re[s]>0 \tag{7.1.799}$$

Or more explicitly (it sometimes good to see different ways...!) using integration by part:

$$\mathcal{L}(t)(s)=\int\limits_0^{+\infty} te^{-st}\mathrm{d}t=t\left.\frac{e^{-st}}{-s}\right|_0^{+\infty}-\int\limits_0^{+\infty}\frac{e^{-st}}{-s}\mathrm{d}t=t\left.\frac{e^{-st}}{-s}\right|_0^{+\infty}-\frac{1}{s}\left.\frac{e^{-st}}{-s}\right|_0^{+\infty} \tag{7.1.800}$$

By the first term we also see that we must have $\sigma>0$. Which gives us:

$$\mathcal{L}(t)(s)=t\left.\frac{e^{-st}}{-s}\right|_0^{+\infty}-\frac{1}{s}\left.\frac{e^{-st}}{-s}\right|_0^{+\infty}=[0-0]-\left[0-\frac{1}{s^2}\right]=\frac{1}{s^2} \tag{7.1.801}$$

And in general (useful especially in Mechanical Engineering even if so far we were able to avoid its usage in this textbook):

$$\mathcal{L}[t^n u(t)]=\frac{n!}{s^{n+1}},\quad \Re[s]>0 \tag{7.1.802}$$

Indeed, using integration by parts:

$$\mathcal{L}(t^n)s=\int\limits_0^{+\infty} t^n e^{-st}\mathrm{d}t=t^n\left.\frac{e^{-st}}{-s}\right|_0^{+\infty}-n\int\limits_0^{+\infty} t^{n-1}\frac{e^{-st}}{-s}\mathrm{d}t \tag{7.1.803}$$

From the first term we also see that we must have $\sigma>0$. Hence it remains:

$$\begin{aligned}\mathcal{L}(t^n)(s)=-n\int\limits_0^{+\infty} t^{n-1}\frac{e^{-st}}{-s}\mathrm{d}t=\frac{n}{s}\int\limits_0^{+\infty} t^{n-1}e^{-st}\mathrm{d}t=\frac{n}{s}\mathcal{L}(t^{n-1})(s)=\frac{n}{s}\frac{n-1}{s}\mathcal{L}(t^{n-2})(s)\\ =\ldots=\frac{n!}{s^n}\mathcal{L}(1)(s)=\frac{n!}{s^{n+1}}\end{aligned} \tag{7.1.804}$$

It then comes under the condition $\sigma>0$:

$$\begin{aligned}\mathcal{L}(t^0)(s)&=\mathcal{L}(1)(s)=\frac{0!}{s}=\frac{1}{s}\\ \mathcal{L}(t^1)(s)&=\mathcal{L}(t)(s)=\frac{1!}{s^2}=\frac{1}{s^2}\\ \mathcal{L}(t^2)(s)&=\mathcal{L}(t^2)(s)=\frac{2!}{s^3}=\frac{2}{s^3}\\ &\ldots\end{aligned} \tag{7.1.805}$$

3. $e^{-at}u(t)$, $te^{-at}u(t)$

Applying the s-domain shifting property to:

$$\mathcal{L}[u(t)]=\frac{1}{s},\quad \Re[s]>0 \tag{7.1.806}$$

we have:

$$\mathcal{L}[e^{-at}u(t)]=\frac{1}{s+a},\quad \Re[s]>-a \tag{7.1.807}$$

If you don't like using the shift property for that proof, here is a more explicit and long way to get the same result (we change the notation a bit to avoid any confusion):

$$\begin{aligned}\mathcal{L}(e^{xt}H(t))(x)&=\int\limits_{-\infty}^{+\infty} e^{xt}H(t)e^{-st}\mathrm{d}t=\int\limits_{0}^{+\infty} e^{xt}e^{-st}\mathrm{d}t=\int\limits_{0}^{+\infty} e^{(a+\mathrm{i}b)t}e^{-(\sigma+\mathrm{i}\omega)t}\mathrm{d}t\\&=\int\limits_{0}^{+\infty} e^{-((\sigma-a)+\mathrm{i}(\omega-b))t}\mathrm{d}t=\left.\frac{e^{-((\sigma-a)+\mathrm{i}(\omega-b))t}}{-((\sigma-a)+\mathrm{i}(\omega-b))}\right|_0^{+\infty}=\left.\frac{e^{-((\sigma-a)+\mathrm{i}(\omega-b))t}}{x-s}\right|_0^{+\infty}\end{aligned} \tag{7.1.808}$$

We see that the integral will converge if and only if:

$$\sigma>a\qquad \omega>b \tag{7.1.809}$$

Therefore:

$$\mathcal{L}(e^{xt}H(t))(x)=\left.\frac{e^{-((\sigma-a)+\mathrm{i}(\omega-b))t}}{x-s}\right|_0^{+\infty}=0-\frac{1}{x-s}=\frac{1}{s-x} \tag{7.1.810}$$

Applying the same property to:

$$\mathcal{L}[t^n u(t)]=\frac{n!}{s^{n+1}},\quad \Re[s]>0 \tag{7.1.811}$$

we have:

$$\mathcal{L}[t^n e^{-at}u(t)]=\frac{n!}{(s+a)^{n+1}},\quad \Re[s]>-a \tag{7.1.812}$$

4. $e^{-\mathrm{i}\omega_0 t}u(t)$, $\sin(\omega_0 t)u(t)$, $\cos(\omega_0 t)u(t)$

We could make it by integration by parts but it's quite long. The trick is letting $a=\pm\mathrm{i}\omega_0$ in:

$$\mathcal{L}[e^{-at}u(t)]=\frac{1}{s+a},\quad \Re[s]>-\Re[a] \tag{7.1.813}$$

we get:

$$\mathcal{L}[e^{-\mathrm{i}\omega_0 t}u(t)]=\frac{1}{s+\mathrm{i}\omega_0}\ \text{ and }\ \mathcal{L}[e^{\mathrm{i}\omega_0 t}u(t)]=\frac{1}{s-\mathrm{i}\omega_0}\quad \Re[s]>0 \tag{7.1.814}$$

and therefore:

$$\mathcal{L}[\cos(\omega_0 t)u(t)]=\frac{1}{2}\mathcal{L}\left[e^{\mathrm{i}\omega_0 t}+e^{-\mathrm{i}\omega_0 t}\right]=\frac{1}{2}\left[\frac{1}{s-\mathrm{i}\omega_0}+\frac{1}{s+\mathrm{i}\omega_0}\right]=\frac{s}{s^2+\omega_0^2} \tag{7.1.815}$$

and:

$$\mathcal{L}[\sin(\omega_0 t)u(t)]=\frac{1}{2\mathrm{i}}\mathcal{L}\left[e^{\mathrm{i}\omega_0 t}-e^{-\mathrm{i}\omega_0 t}\right]=\frac{1}{2\mathrm{i}}\left[\frac{1}{s-\mathrm{i}\omega_0}-\frac{1}{s+\mathrm{i}\omega_0}\right]=\frac{\omega_0}{s^2+\omega_0^2} \tag{7.1.816}$$

5. $t\ \cos(\omega_0 t)u(t)$, $t\ \sin(\omega_0 t)u(t)$

   Letting $a = \pm \mathrm{i}\omega_0$ in:

$$\mathcal{L}[te^{-at}u(t)] = \frac{1}{(s+a)^2}, \quad \Re[s] > -a \tag{7.1.817}$$

   we get:

$$\mathcal{L}[te^{-\mathrm{i}\omega_0 t}u(t)] = \frac{1}{(s+\mathrm{i}\omega_0)^2} \quad \text{and} \quad \mathcal{L}[te^{\mathrm{i}\omega_0 t}u(t)] = \frac{1}{(s-\mathrm{i}\omega_0)^2}, \quad \Re[s] > -a \tag{7.1.818}$$

   Furthermore we have:

$$\mathcal{L}[t\ \cos(\omega_0 t)u(t)] = \frac{1}{2}\mathcal{L}[t\,(e^{\mathrm{i}\omega_0 t} + e^{-\mathrm{i}\omega_0 t})] = \frac{1}{2}\left[\frac{1}{(s-\mathrm{i}\omega_0)^2} + \frac{1}{(s+\mathrm{i}\omega_0)^2}\right] = \frac{s^2-\omega_0^2}{(s^2+\omega_0^2)^2} \tag{7.1.819}$$

   and:

$$\mathcal{L}[t\ \sin(\omega_0 t)u(t)] = \frac{1}{2\mathrm{i}}\mathcal{L}\left[t\,(e^{\mathrm{i}\omega_0 t} - e^{-\mathrm{i}\omega_0 t})\right] = \frac{1}{2\mathrm{i}}\left[\frac{1}{(s-\mathrm{i}\omega_0)^2} - \frac{1}{(s+\mathrm{i}\omega_0)^2}\right] = \frac{2s\omega_0}{(s^2+\omega_0^2)^2} \tag{7.1.820}$$

6. $e^{-at}\cos(\omega_0 t)u(t)$, $e^{-at}\sin(\omega_0 t)u(t)$

   Applying s-domain shifting property to:

$$\mathcal{L}[\cos(\omega_0 t)u(t)] = \frac{s}{s^2+\omega_0^2} \tag{7.1.821}$$

   and:

$$\mathcal{L}[\sin(\omega_0 t)u(t)] = \frac{\omega_0}{s^2+\omega_0^2} \tag{7.1.822}$$

   we get, respectively:

$$\mathcal{L}[e^{-at}\cos(\omega_0 t)u(t)] = \frac{s+a}{(s+a)^2+\omega_0^2} \tag{7.1.823}$$

   and:

$$\mathcal{L}[e^{-at}\sin(\omega_0 t)u(t)] = \frac{\omega_0}{(s+a)^2+\omega_0^2} \tag{7.1.824}$$

7. Let us see an important transform pair that we will find invaluable when we get to transients in transmission lines and in advanced quantitative finance[7]. Specifically, if we define the "Gauss error function":

$$\text{erf}(t) = \frac{2}{\sqrt{\pi}}\int_0^t e^{-u^2}\mathrm{d}u \tag{7.1.825}$$

   then we will derive here the pair of unilateral Laplace transform:

$$1 - \text{erf}\left(\frac{a}{2\sqrt{t}}\right) \leftrightarrow \frac{1}{s}e^{-a\sqrt{s}} \tag{7.1.826}$$

---

[7] Notice that the entire development that will follow is inspired from the excellent book: *Transients for Electrical Engineers* **??**.

Proof 7.3.25. We will start by computing the Laplace transform of:

$$x(t)=\frac{e^{-a^2/4t}}{\sqrt{\pi t^3}} \tag{7.1.827}$$

How do we know we have to start from this? Because this function appears in the book *Analytical Theory of Heat* (11822 according to holocene calendar) of Joseph Fourier where he already did the job before us but in a different painful way[8]!

In any case what we have is:

$$X(s)=\mathcal{UL}(x(t))=\int\limits_0^{+\infty}x(t)e^{-st}\mathrm{d}t=\int\limits_0^{+\infty}\frac{e^{-a^2/4t}}{\sqrt{\pi t^3}}\mathrm{d}t \tag{7.1.828}$$

Next, let us make the change of variable:

$$\tau^2=\frac{a^2}{4t} \tag{7.1.829}$$

and so:

$$\tau=\frac{a}{2t^{1/2}} \tag{7.1.830}$$

Therefore:

$$\mathrm{d}t=-\sqrt{t^3}\frac{4}{a}\mathrm{d}\tau \tag{7.1.831}$$

With this our unilateral Laplace transforms becomes:

$$\begin{aligned}X(s)&=\int\limits_{+\infty}^{0}\frac{e^{-\tau^2}}{\sqrt{\pi t^3}}e^{-s\frac{a^2}{4\tau^2}}\left(-\sqrt{t^3}\frac{4}{a}\mathrm{d}\tau\right)=-\frac{4}{a\sqrt{\pi}}\int\limits_{+\infty}^{0}e^{-\left[\tau^2+s\frac{a^2}{4\tau^2}\right]}\mathrm{d}\tau\\&=\frac{4}{a\sqrt{\pi}}\int\limits_0^{+\infty}e^{-\left[\tau^2+s\dfrac{a^2}{4\tau^2}+a\sqrt{s}-a\sqrt{s}\right]}\mathrm{d}\tau=\frac{4}{a\sqrt{\pi}}e^{-a\sqrt{s}}\int\limits_0^{+\infty}e^{-\left[\tau^2-a\sqrt{s}+\left(s\frac{a\sqrt{s}}{2\tau}\right)^2\right]}\mathrm{d}\tau\\&=\frac{4}{a\sqrt{\pi}}e^{-a\sqrt{s}}\int\limits_0^{+\infty}e^{-\left(\tau-\frac{a\sqrt{s}}{2\tau}\right)^2}\mathrm{d}\tau\end{aligned} \tag{7.1.832}$$

Or:

$$X(s)=\frac{4}{a\sqrt{\pi}}e^{-a\sqrt{s}}\int\limits_0^{+\infty}e^{-\left(\tau-\frac{b}{\tau}\right)^2}\mathrm{d}\tau \tag{7.1.833}$$

where:

$$b=\frac{a\sqrt{s}}{2} \tag{7.1.834}$$

---

[8] You can find it in the English translation of year 11878 according to holocene calendar (see [188]) at the page 384 for $\mathbb{R}^3$.

Now, concentrate on the integral, alone. Multiplying out the exponent of the integrand, we have:

$$\int\limits_0^{+\infty} e^{-\left(\tau^2-2b+\frac{b^2}{\tau^2}\right)^2}\mathrm{d}\tau=e^{2b}\int\limits_0^{+\infty} e^{-\left(\tau^2+\frac{b^2}{\tau^2}\right)^2}\mathrm{d}\tau=e^{2b}I(b) \tag{7.1.835}$$

where:

$$I(b)=\int\limits_0^{+\infty} e^{-\left(\tau^2+\frac{b^2}{\tau^2}\right)^2}\mathrm{d}\tau \tag{7.1.836}$$

The integral $I(b)$, despite its perhaps complicated appearance, can be evaluated as follows.

Differentiating with respect to b we get:

$$\frac{\mathrm{d}I}{\mathrm{d}b}=\int\limits_0^{+\infty}\left(-\frac{2b}{\tau^2}\right)e^{-\left(\tau^2+\frac{b^2}{\tau^2}\right)^2}\mathrm{d}\tau \tag{7.1.837}$$

Now, make the change of variable:

$$u=\frac{b}{\tau}\Leftrightarrow\tau=\frac{b}{u} \tag{7.1.838}$$

which leads to:

$$\mathrm{d}\tau=-\frac{b}{u^2}\mathrm{d}u \tag{7.1.839}$$

and so:

$$\frac{\mathrm{d}I}{\mathrm{d}b}=\int\limits_{+\infty}^{0}\left(-\frac{2u^2}{b}\right)e^{-\left(\frac{b^2}{u^2}+u^2\right)}\left(-\frac{b}{u^2}\mathrm{d}u\right)=-2\int\limits_0^{+\infty}e^{-\left(\frac{b^2}{u^2}+u^2\right)}\mathrm{d}u=-2I(b) \tag{7.1.840}$$

Thus:

$$\frac{\mathrm{d}I}{I}=-2\mathrm{d}b \tag{7.1.841}$$

which the primitive gives:

$$I(b)=c^{te}e^{-2b} \tag{7.1.842}$$

where c^{te} is some constant. We can determine it easily with (see section Statistics page 695):

$$I(0)=c^{te}=\int\limits_0^{+\infty}e^{-\tau^2}\mathrm{d}\tau=\frac{\sqrt{\pi}}{2} \tag{7.1.843}$$

the reader may have indeed recognized the Gauss integral here.

So:

$$I(b)=\frac{\sqrt{\pi}}{2}e^{-2b} \tag{7.1.844}$$

And putting into:

$$\int\limits_0^{+\infty} e^{-\left(\tau-\frac{b}{\tau}\right)^2}\mathrm{d}\tau=\int\limits_0^{+\infty} e^{-\left(\tau^2-2b+\frac{b^2}{\tau^2}\right)^2}\mathrm{d}\tau=e^{2b}\int\limits_0^{+\infty} e^{-\left(\tau^2+\frac{b^2}{\tau^2}\right)^2}\mathrm{d}\tau=e^{2b}I(b) \tag{7.1.845}$$

We get:

$$\int\limits_0^{+\infty} e^{-\left(\tau-\frac{b}{\tau}\right)^2}\mathrm{d}\tau=\frac{\sqrt{\pi}}{2} \tag{7.1.846}$$

And putting also back into:

$$X(s)=\frac{4}{a\sqrt{\pi}}e^{-a\sqrt{s}}\int\limits_0^{+\infty} e^{-\left(\tau-\frac{b}{\tau}\right)^2}\mathrm{d}\tau=\frac{4}{a\sqrt{\pi}}e^{-a\sqrt{s}}\frac{\sqrt{\pi}}{2}=\left(\frac{2}{a}e^{-a\sqrt{s}}\right) \tag{7.1.847}$$

Which gives us indeed the pair (that will be useful to us for the study of the Telegrapher equation):

$$\frac{e^{-a^2/4t}}{\sqrt{\pi t^3}}\leftrightarrow\frac{2}{a}e^{-a\sqrt{s}} \tag{7.1.848}$$

Which is equivalent:

$$\frac{ae^{-a^2/4t}}{2\sqrt{\pi t^3}}\leftrightarrow e^{-a\sqrt{s}} \tag{7.1.849}$$

Finally, recall that we have proved during our studies of main Laplace transform properties the (see section Analysis page 2345) following pair (integration in time domain):

$$\int\limits_0^{+\infty} x(t)\mathrm{d}t\leftrightarrow\frac{X(s)}{s} \tag{7.1.850}$$

Therefore:

$$\int\limits_0^{+\infty}\frac{ae^{-a^2/4t}}{2\sqrt{\pi t^3}}\mathrm{d}t\leftrightarrow\frac{e^{-a\sqrt{s}}}{s} \tag{7.1.851}$$

In the integral above let us make the following change of variable:

$$u=\frac{a}{2\sqrt{t}} \tag{7.1.852}$$

Therefore:

$$\mathrm{d}t=-\frac{4\sqrt{t^3}}{a}\mathrm{d}u \tag{7.1.853}$$

and so:

$$\frac{a}{2\sqrt{\pi}}\int\limits_{+\infty}^{\frac{a}{2\sqrt{t}}}\frac{e^{-u^2}}{\sqrt{x^3}}\left(-\frac{4\sqrt{x^3}}{a}\mathrm{d}u\right)=\frac{2}{\sqrt{\pi}}\int\limits_{\frac{a}{2\sqrt{t}}}^{+\infty}e^{-u^2}\mathrm{d}u=\frac{2}{\sqrt{\pi}}\left[\int\limits_0^{+\infty}e^{-u^2}\mathrm{d}u-\int\limits_0^{\frac{a}{2\sqrt{t}}}e^{-u^2}\mathrm{d}u\right] \tag{7.1.854}$$

In the first integral in the square brackets we recognize the half of the Gauss integral. Therefore:

$$\frac{2}{\sqrt{\pi}}\left[\int\limits_0^{+\infty}e^{-u^2}\mathrm{d}u-\int\limits_0^{\frac{a}{2\sqrt{t}}}e^{-u^2}\mathrm{d}u\right]=\frac{2}{\sqrt{\pi}}\left[\frac{\sqrt{\pi}}{2}-\int\limits_0^{\frac{a}{2\sqrt{t}}}e^{-u^2}\mathrm{d}u\right] \tag{7.1.855}$$

In the second one we recognize the Gauss error function (see page 3697), therefore:

$$\frac{2}{\sqrt{\pi}}\left[\frac{\sqrt{\pi}}{2}-\int\limits_0^{\frac{a}{2\sqrt{t}}}e^{-u^2}\mathrm{d}u\right]=\frac{2}{\sqrt{\pi}}\left[\frac{\sqrt{\pi}}{2}-\frac{\sqrt{\pi}}{2}\mathrm{erf}\left(\frac{a}{2\sqrt{t}}\right)\right]=1-\mathrm{erf}\left(\frac{a}{2\sqrt{t}}\right) \tag{7.1.856}$$

Therefore our pair:

$$\int\limits_0^{+\infty}\frac{ae^{-a^2/4t}}{2\sqrt{\pi t^3}}\mathrm{d}t\leftrightarrow\frac{e^{-a\sqrt{s}}}{s} \tag{7.1.857}$$

can be rewritten:

$$1-\mathrm{erf}\left(\frac{a}{2\sqrt{t}}\right)\leftrightarrow\frac{e^{-a\sqrt{s}}}{s} \tag{7.1.858}$$

and so, at last, we have the claimed pair of the beginning!

■ Q.E.D.

To summarize:

$f(t)$	$F(s)=\mathcal{L}[f(t)]$	**ROC**
$f(t)=1$	$F(s)=\frac{1}{s}$	$s>0$
$f(t)=\delta(t)$	$F(s)=1$	$s\forall s0$
$f(t)=u(t)$ (heaviside)	$F(s)=\frac{1}{s}$	$\Re[s]>0$
$f(t)=t^b u(t)$	$F(s)=\frac{n!}{s^{n+1}}$	$\Re[s]>0$
$f(t)=tu(t)$	$F(s)=\frac{1}{s^2}$	$\Re[s]>0$
$f(t)=e^{at}$	$F(s)=\frac{1}{(s-a)}$	$s>a$
$f(t)=e^{-at}$	$F(s)=\frac{1}{s+a}$	$\Re[s]>-a$
$f(t)=e^{-a\|t\|}$	$F(s)=\frac{-2a}{s^2-a^2}$	$-b<\Re[s]<b$
$f(t)=t^n e^{-at}$	$F(s)=\frac{n!}{(s+a)^{n+1}}$	$\Re[s]>-a$
$f(t)=e^{-\mathrm{i}at}$	$F(s)=\frac{1}{s+\mathrm{i}a}$	$\Re[s]>-a$
$f(t)=t^n$	$F(s)=\frac{n!}{s^{(n+1)}}$	$s>0$
$f(t)=\sin(at)$	$F(s)=\frac{a}{s^2+a^2}$	$s>0$
$f(t)=\cos(at)$	$F(s)=\frac{s}{s^2+a^2}$	$s>0$
$f(t)=t\sin(at)$	$F(s)=\frac{2sa}{(s^2+a^2)^2}$	$s>0$
$f(t)=t\cos(at)$	$F(s)=\frac{s^2-a^2}{(s^2+a^2)^2}$	$s>0$
$f(t)=\sinh(at)$	$F(s)=\frac{a}{s^2-a^2}$	$s>\|a\|$
$f(t)=\cosh(at)$	$F(s)=\frac{s}{s^2-a^2}$	$s>\|a\|$
$f(t)=t^n e^{at}$	$F(s)=\frac{n!}{(s-a)^{(n+1)}}$	$s>a$
$f(t)=e^{at}\sin(bt)$	$F(s)=\frac{b}{(s-a)^2+b^2}$	$s>a$
$f(t)=e^{at}\cos(bt)$	$F(s)=\frac{(s-a)}{(s-a)^2+b^2}$	$s>a$
$f(t)=e^{at}\sinh(bt)$	$F(s)=\frac{b}{(s-a)^2-b^2}$	$s-a>\|b\|$
$f(t)=e^{at}\cosh(bt)$	$F(s)=\frac{(s-a)}{(s-a)^2-b^2}$	$s-a>\|b\|$
$f(t)1-\operatorname{erf}\left(\frac{a}{2\sqrt{t}}\right)=1-\frac{2}{\sqrt{\pi}}\int_0^{a/(2\sqrt{t})}e^{-u^2}\mathrm{d}u$	$F(s)=\frac{1}{s}e^{-a\sqrt{s}}$	$s>0$

Table 7.10: **Most common Laplace transform**

7.1.5.2.7 Solving differential systems

Let us see now a well known simple academic companion example of the application of the Laplace transform to deal with differential equations.

$$\text{origin space}\left\{\begin{tikzcd} \text{ODE in } t \arrow[r, "\mathcal{L}"] \arrow[d, dashed] & \text{eqn. in } s \arrow[d] \\ y(t) & Y(s) \arrow[l, "\mathcal{L}^{-1}"'] \end{tikzcd}\right\}\text{transform space}$$

Let us consider that a voltage $x(t)$ is applied as the input to a resistor R, a capacitor C and an inductor L connected in series (the case without the use of the Laplace transform is introduced in details in the section of Electrical Engineering page 8475). The output $y(t)$ is the voltage across one of the three elements. The system can be described by a differential equation in time domain:

$$x(t)=v_L(t)+v_R(t)+v_C(t)=L\frac{\mathrm{d}}{\mathrm{d}t}\,i(t)+R\,i(t)+\frac{1}{C}\int_{-\infty}^{t}i(\tau)\mathrm{d}\tau \tag{7.1.859}$$

or by an algebraic equation in s-domain:

$$X(s)=V_L(s)+V_R(s)+V_C(s)=\left[sL+R+\frac{1}{sC}\right]I(s) \tag{7.1.860}$$

and the overall impedance of the circuit is defined as the ratio between voltage $V(s)$ and the current $I(s)$:

$$Z(s)\overset{\triangle}{=}\frac{V(s)}{I(s)}=sL+R+\frac{1}{sC} \tag{7.1.861}$$

which is composed of the individual impedance of the three elements:

$$Z_L(s)=sL,\quad Z_R(s)=R,\quad Z_C(s)=\frac{1}{sC} \tag{7.1.862}$$

	resistor R	capacitor C	inductor L
time domain	$i=\frac{v}{R}$	$i=\frac{1}{C}\frac{\mathrm{d}v}{\mathrm{d}t}$	$v=\frac{1}{L}\frac{\mathrm{d}i}{\mathrm{d}t}$
s-domain	$V_R=IR$	$V_C=I/Cs$	$V_L=IsL$
impedance $Z=V/I$	R	$1/sC$	sL

If the output is the voltage across one of the three elements (V_L, V_R, or V_C), the transfer function $H(s)$ can be easily obtained by treating the series circuit as a voltage divider:

- Output is voltage across the capacitor $v_C(t)$

$$H_C(s)=\frac{1/sC}{Ls+R+1/sC}=\frac{1/LC}{s^2+(R/L)s+(1/LC)} \tag{7.1.863}$$

- Output is voltage across the resistor $v_R(t)$

$$H_R(s)=\frac{R}{Ls+R+1/sC}=\frac{(R/L)s}{s^2+(R/L)s+(1/LC)} \tag{7.1.864}$$

- Output is voltage across the inductor $v_L(t)$

$$H_L(s)=\frac{sL}{Ls+R+1/sC}=\frac{s^2}{s^2+(R/L)s+(1/LC)} \tag{7.1.865}$$

If we define:

$$\zeta\overset{\triangle}{=}\frac{R}{2}\sqrt{\frac{C}{L}},\quad \omega_n\overset{\triangle}{=}\frac{1}{\sqrt{LC}} \tag{7.1.866}$$

the common denominator of the transfer functions can be written in standard (canonical) form:

$$s^2+(R/L)s+(1/LC)=s^2+2\zeta\omega_n s+\omega_n^2 \tag{7.1.867}$$

with two roots:

$$p_{1,2}=(-\zeta\pm\sqrt{\zeta^2-1})\omega_n \tag{7.1.868}$$

and the transfer functions above can be written in standard forms:

- The first one:

  $$H_C(s)=\frac{\omega_n^2}{s^2+2\zeta\omega_n s+\omega_n^2} \tag{7.1.869}$$

  with two poles p_1,p_2 and no zeros.

- The second one:

  $$H_R(s)=\frac{2\zeta\omega_n s}{s^2+2\zeta\omega_n s+\omega_n^2} \tag{7.1.870}$$

  with two poles p_1,p_2 and one zero at the origin.

- The third one:

  $$H_L(s)=\frac{s^2}{s^2+2\zeta\omega_n s+\omega_n^2} \tag{7.1.871}$$

  with two poles p_1,p_2 and two repeated zeros at the origin.

The three transfer functions behave like low-pass, band-pass and high-pass filter, respectively. Moreover, when the common real part $-\zeta\omega_n$ of the two complex conjugate poles is small (i.e., $0<\zeta<0.5$), there will be a narrow pass-band around $\omega=\omega_n$ for all three transfer functions. The magnitude and phase angle of the corresponding frequency response function $|H(\mathrm{i}\omega)|$ can be qualitatively determined in the s-plane, as to be discussed later.

7.1.5.2.8 Unilateral Laplace Transform

The Laplace transform so far discussed is the "bilateral Laplace transform" as it can be applied to left sided signals as well as right sided ones. Now we will consider the "unilateral Laplace transform" of an arbitrary signal $x(t)$ defined as:

$$\boxed{\mathcal{UL}(x(t))=X(s)\triangleq\int_{-\infty}^{+\infty}x(t)u(t)e^{-st}\mathrm{d}t=\int_{0^-}^{+\infty}x(t)e^{-st}\mathrm{d}t} \tag{7.1.872}$$

where for recall $\mathcal{U}$ and $u(t)$ are another typical notation of the Heaviside function.

This definition always assumes $x(t)=0$ for $t<0$. When the unilateral Laplace transform is applied to find the transfer function $H(s)=\mathcal{UL}[h(t)]$ of a LTI system, it is always assumed to be causal. And the ROC is always right sided in s-plane.

By definition, the unilateral Laplace transform of any signal $x(t)=x(t)u(t)$ is identical to its bilateral Laplace transform. However, when $x(t)\neq x(t)u(t)$, the two Laplace transforms are different.

Unilateral Laplace Transform shares all the properties of bilateral Laplace transform, except some of the properties are expressed in different forms. Here we only consider the differentiation in time domain:

$$\mathcal{UL}\left[\frac{\mathrm{d}}{\mathrm{d}t}x(t)\right]=\mathcal{UL}[x'(t)]=sX(s)-x(0^-) \tag{7.1.873}$$

Proof 7.3.26.

$$\begin{aligned}\mathcal{UL}\left[\frac{\mathrm{d}}{\mathrm{d}t}x(t)\right]&=\int\limits_{0^-}^{+\infty}x'(t)e^{-st}\mathrm{d}t=\int\limits_{0^-}^{+\infty}e^{-st}\mathrm{d}[x(t)]=x(t)e^{-st}|_{0^-}^{+\infty}-\int\limits_{0^-}^{+\infty}x(t)\mathrm{d}(e^{-st})\\&=-x(0)+s\int\limits_{0^-}^{+\infty}x(t)e^{-st}\mathrm{d}t=sX(s)-x(0^-)\end{aligned} \tag{7.1.874}$$

■ Q.E.D.

We can further get Laplace transform of higher order derivatives:

$$\mathcal{UL}[x''(t)]=s\mathcal{UL}[x'(t)]-x'(0^-)=s^2X(s)-sx(0^-)-x'(0^-) \tag{7.1.875}$$

and in general:

$$\mathcal{UL}[x^{(n)}(t)]=s^nX(s)-\sum_{k=0}^{n-1}s^kx^{(n-1-k)}(0^-) \tag{7.1.876}$$

To finish depending on donations

7.1.5.3 $\mathcal{Z}$-Transform

Laplace transforms (whose Fourier transform is for recall a special case) is applicable only for functions say ... "continuous" to make it simple.... But since the advent of computing the vast majority of functions (signals) are samples at a time interval T_s such that the functions are actually discrete.

From then on it becomes useful to define a transformation similar to the Laplace transform but in the discrete case! Thus, the Laplace transform becomes a special case of the $\mathcal{Z}$-transform when the sampling period T_s tends to 0.

The discrete Fourier transform of a discrete signal $x[n]$ is defined for recall as:

$$X(e^{\mathrm{i}\omega})=\mathcal{F}[x[n]]=\sum_{n=-\infty}^{+\infty}x[n]e^{-\mathrm{i}2\pi fn}=\sum_{n=-\infty}^{+\infty}x[n]e^{-\mathrm{i}\omega n} \tag{7.1.877}$$

provided $x[n]$ is absolutely summable:

$$\sum_{n=-\infty}^{+\infty}|x[n]|<+\infty \tag{7.1.878}$$

Obviously some signals may not satisfy this condition and their Fourier transform do not exist. To overcome this difficulty, we can multiply the given $x[n]$ by an exponential function $e^{-\sigma n}$ so that $x[n]$ may be forced to be summable for certain values of the real parameter σ. Now the discrete time Fourier transform becomes:

$$\begin{aligned}\mathcal{F}[x[n]e^{-\sigma n}]&=\sum_{n=-\infty}^{+\infty}\left[x[n]e^{-\sigma n}\right]e^{-\mathrm{i}\omega n}=\sum_{n=-\infty}^{+\infty}x[n]e^{-(\sigma+\mathrm{i}\omega)n}\\&=\sum_{n=-\infty}^{+\infty}x[n]e^{-sn}=\sum_{n=-\infty}^{+\infty}x[n]z^{-n}=X(z)\end{aligned} \tag{7.1.879}$$

The result of this summation is a function of a complex variable defined as:

$$z=e^s=e^{\sigma+\mathrm{i}\omega} \tag{7.1.880}$$

This is the "forward (bilateral) $\mathcal{Z}$-transform" of the discrete signal $x[n]$:

$$\boxed{X(z)=\mathcal{Z}[x[n]]=\sum_{n=-\infty}^{+\infty}x[n]z^{-n}} \tag{7.1.881}$$

There is another way to introduce the $\mathcal{Z}$-transform that result to the same result but with different notation and that is more explicit. As it may help the reader to better understand, let us show that but by limiting ourselves to positive times only (the case of interest in engineering is the unilateral $\mathcal{Z}$-transform!).

So we know that a continuous function $f(t)t$ can be discretized (sampled) in a way to take at each time step nT_s with $n\in\mathbb{N}$ the value of the function of interest, multiplied by a Dirac pulse on this same point such that:

$$f_s(t) \underset{\lim T_s \to 0}{=} \sum_{0}^{+\infty} f(nT_s)\delta(t-nT_s) \qquad (7.1.882)$$

Now let us calculate the Laplace transform of this discretized function at the limit and using the fact that the $f(nT_e)$ are henceforth independent from the time and the linearity property of the Laplace transform:

$$\mathcal{L}(f_s(t))(s) = \int_0^{+\infty} \sum_{n=0}^{+\infty} f(nT_s)\delta(t-nT_s)\mathrm{d}t = \sum_{n=0}^{+\infty} \left[f(nT_s) \int_0^{+\infty} \delta(t-nT_s)\mathrm{d}t \right] \qquad (7.1.883)$$

As we have shown above in our study of the usual Laplace transforms, we have:

$$\int_0^{+\infty} \delta(t-nT_s) = e^{-nsT} \qquad (7.1.884)$$

Then we have:

$$\mathcal{L}(f_s(t))(s) = \sum_{n=0}^{+\infty} f(nT_s)e^{-nst} \qquad (7.1.885)$$

That is usage as we have just seen before to write as following and define as the "unilateral $\mathcal{Z}$-transform" (on which we will come back later below):

$$\mathcal{Z}(f_s(t))(z) = \sum_{n=0}^{+\infty} f_n z^{-n} \qquad (7.1.886)$$

with:

$$f_n = f(nT_s) \qquad z = e^{st} \qquad (7.1.887)$$

That's it for the second approach!

Now, given the $\mathcal{Z}$-transform $X(z)$, the original time signal can be obtained by the inverse $\mathcal{Z}$-transform, which can be derived from the corresponding discrete Fourier transform. As shown above, we have:

$$X(z) = X(e^{\sigma+\mathrm{i}\omega}) = \sum_{n=-\infty}^{+\infty} [x[n]e^{-\sigma n}]e^{-\mathrm{i}\omega n} = \mathcal{F}[x[n]e^{-\sigma n}] \qquad (7.1.888)$$

Now $x[n]e^{-\sigma n}$ can be obtained by the inverse Fourier transform:

$$x[n]e^{-\sigma n} = \mathcal{F}^{-1}[X(e^{\sigma+\mathrm{i}\omega})] = \frac{1}{2\pi}\int_0^{2\pi} X(e^{\sigma+\mathrm{i}\omega})e^{\mathrm{i}\omega n}\mathrm{d}\omega \qquad (7.1.889)$$

Multiplying both sides by $e^{\sigma n}$, we get:

$$x[n] = \frac{1}{2\pi}\int_0^{2\pi} X(e^{\sigma+\mathrm{i}\omega})e^{(\sigma+\mathrm{i}\omega)n}\mathrm{d}\omega = \frac{1}{2\pi}\int_0^{2\pi} X(z)z^n\mathrm{d}\omega \qquad (7.1.890)$$

To represent the inverse transform in terms of z (instead of ω), we note:

$$\mathrm{d}z = \mathrm{d}(e^{\sigma+\mathrm{i}\omega}) = e^{\sigma}\mathrm{i}e^{\mathrm{i}\omega}\mathrm{d}\omega = \mathrm{i}z\,\mathrm{d}\omega \tag{7.1.891}$$

i.e.:

$$\mathrm{d}\omega = \frac{\mathrm{d}z}{\mathrm{i}z} \tag{7.1.892}$$

and the "inverse $\mathcal{Z}$-transform" can be obtained as:

$$\boxed{x[n] = \mathcal{Z}^{-1}[X(z)] = \frac{1}{\mathrm{i}2\pi}\oint X(z)z^{n-1}\mathrm{d}z} \tag{7.1.893}$$

Note that the integral with respect to ω from 0 to 2π becomes an integral with respect to $z = e^{\sigma+\mathrm{i}\omega}$ in the complex z-plane, along a circle with a fixed radius e^{σ} and a varying angle ω from 0 to 2π. Now we have the $\mathcal{Z}$-transform pair:

$$X(z) = \mathcal{Z}[x[n]] = \sum_{n=-\infty}^{+\infty} x[n]z^{-n} \tag{7.1.894}$$

$$x[n] = \mathcal{Z}^{-1}[X(z)] = \frac{1}{2\pi\mathrm{i}}\oint X(z)z^{n-1}\mathrm{d}z \tag{7.1.895}$$

The forward and inverse $\mathcal{Z}$-transform pair can also be represented as:

$$x[n] \overset{\mathcal{Z}}{\longleftrightarrow} X(z) \tag{7.1.896}$$

In particular, if we let $\sigma = 0$, i.e., $z = e^{\mathrm{i}\omega}$, then the $\mathcal{Z}$-transform becomes the discrete-time Fourier transform:

$$X(z)\Big|_{z=e^{\mathrm{i}\omega}} = \sum_{n=-\infty}^{+\infty} x[n]e^{-\mathrm{i}\omega n} = X(e^{\mathrm{i}}\omega) \tag{7.1.897}$$

This is the reason why sometimes the discrete Fourier spectrum is expressed as a function of $e^{\mathrm{i}\omega}$.

Different from the discrete-time Fourier transform which converts a 1-D signal $x[n]$ in time domain to a 1-D complex spectrum $X(e^{\mathrm{i}\omega})$ in frequency domain, the $\mathcal{Z}$-transform $X(s)$ converts the 1-D signal $x[n]$ to a complex function defined over a 2-D complex plane, named the "z-plane", represented in polar form by radius $|z| = |e^{\sigma+\mathrm{i}\omega}| = e^{\sigma}$ and angle $\angle z = \angle(e^{\sigma+\mathrm{i}\omega}) = \omega$.

In particular, when this 2-D function $X(z) = X(e^{\sigma+\mathrm{i}\omega})$ is evaluated along the unit circle $|z| = e^{0} = 1$ corresponding to $\sigma = 0$, it becomes a 1-D periodic function $X(e^{\mathrm{i}\omega})$, the discrete Fourier transform of $x[n]$.

7.1.5.3.1 Transfer function of LTI system

The output $y[n]$ of a discrete LTI (Linear Time-Invariant) system with input $x[n]$ can be found by convolution (see page 2293):

$$y[n]=\mathcal{O}[x[n]]=h[n]*x[n]=\sum_{k=-\infty}^{+\infty} h[k]x[n-k] \tag{7.1.898}$$

where $h[n]$ is the impulse response function of the system. In particular, if the input is a complex exponential:

$$x[n]=z^n=(e^s)^n=e^{sn} \tag{7.1.899}$$

then the output $y[n]$ can be found to be:

$$y[n]=\mathcal{O}[z^n]=\sum_{k=-\infty}^{+\infty} h[k]z^{n-k}=z^n\sum_{k=-\infty}^{+\infty} h[k]z^{-k}=H(z)z^n \tag{7.1.900}$$

This is the eigenequation with the complex exponential $x[n]=z^n=e^{sn}$ being the eigenfunction of any discrete LTI system, corresponding to its eigenvalue defined as:

$$H(z)\triangleq\sum_{k=-\infty}^{+\infty} h[k]z^{-k} \tag{7.1.901}$$

which is the $\mathcal{Z}$-transform of its impulse response $h[n]$, called the "transfer function" of the LTI system. In particular, when $\sigma=0$, i.e., $z=e^s=e^{\mathrm{i}\omega}$, the transfer function $H(z)$ becomes the frequency response function, the Fourier transform of the impulse response:

$$H(e^{\mathrm{i}\omega})\triangleq\sum_{n=-\infty}^{+\infty} h[n]e^{-\mathrm{i}\omega n} \tag{7.1.902}$$

7.1.5.3.2 Conformal mapping between s-plane to z-plane

The s-plane and the z-plane are related by a conformal mapping (i.e. conserves the angles) specified by the analytic complex function:

$$z=e^s=e^{\sigma+\mathrm{i}\omega}=e^{\sigma}e^{\mathrm{i}\omega}=re^{\mathrm{i}\omega} \tag{7.1.903}$$

where :

$$\begin{cases}\Re[s]=\sigma\\ \Im[s]=\mathrm{i}\omega\end{cases}\quad\text{and}\quad\begin{cases}|z|=r=e^{\sigma}\\ \angle z=\omega\end{cases} \tag{7.1.904}$$

Even if, as far as we know, there are no important practical application of this mapping, it is more a mathematical curiosity the may help some students to understand how the s-plane and z-plane are related together.

The mapping is continuous, i.e., neighbouring points in s-plane are mapped to neighbouring points in z-plane and vice versa. Consider the mapping of these specific features:

- The origin $s=0$ of s-plane is mapped to $z=e^0=1$ on the real axis in z-plane.
- Each vertical line $\Re[s]=\sigma_0$ in s-plane is mapped to a circle $|z|=e^{\sigma_0}$ centered about the origin in z-plane. In particular:
  - Leftmost vertical line $\Re[s]=\sigma=-\infty$ is mapped as the origin $|z|=e^{-\infty}=0$
  - Imaginary axis $\Re[s]=0$ is mapped as the unit circle $|z|=e^0=1$
  - Rightmost vertical line $\Re[s]=\sigma=+\infty$ is mapped as a circle of infinite radius $|z|=e^{+\infty}=+\infty$.
- Each horizontal line $\Im[s]=\mathrm{i}\omega_0$ in s-plane is mapped to $\angle z=\omega_0$, a ray from the origin in z-plane of angle ω_0 with respect to the positive horizontal direction.
- A right angle formed by a pair vertical and horizontal lines in s-plane is conserved by the mapping, as the corresponding circle and ray in z-plane also form a right angle (in fact any angle is conserved, an important property of the conformal mapping).

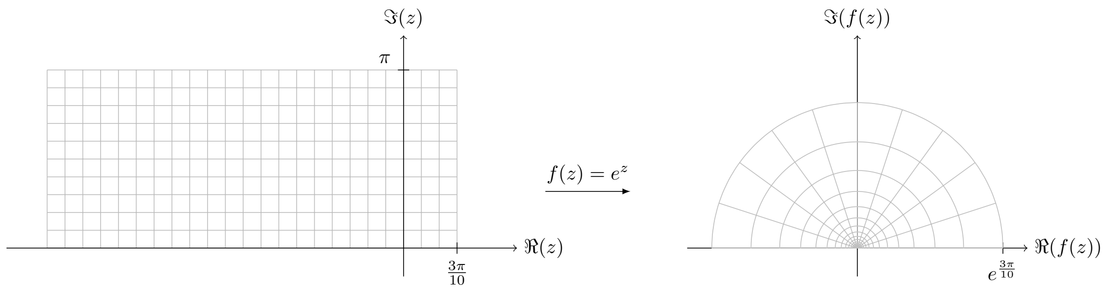


Figure 7.67: Example of a simple complex conformal mapping

The infinite range $-\infty<\omega<+\infty$ for frequency ω along a vertical line $\Re[s]=\sigma_0$ in s-plane is mapped repeatedly to a finite range $0\leq\omega<2\pi$ around a circle $|z|=e^{\sigma_0}$ in z-plane, corresponding to the conversion of a continuous signal $x(t)$ with non-periodic spectrum $X(\mathrm{i}\omega)$ for $-\infty<\omega<+\infty$ to a discrete signal $x[n]$ with periodic spectrum $X(e^{\mathrm{i}\omega})$ for $0\leq\omega<2\pi$.

7.1.5.3.3 Region of Convergence

Whether the $\mathcal{Z}$-transform $X(z)$ of a signal $x[n]$ exists depends on the complex variable $z=e^s$ as well as the signal itself. $X(z)$ exists if and only if the argument z is inside the region of convergence (ROC) in the z-plane, which is composed of all z values for the summation of the $\mathcal{Z}$-transform to converge. The ROC of the $\mathcal{Z}$-transform is determined by $|z|=|e^s|=e^\sigma$ (a circle), the magnitude of variable z, while the ROC for the Laplace transform is determined by $\sigma=\Re[s]$, (a vertical line), the real part of s.

✍**Examples:**

E1. Let us see our first $\mathcal{Z}$-transform with at the same time its corresponding region of convergence. For this purpose, let us first recall the following geometric series (see page 2419):

$$\sum_{n=0}^{+\infty}=\frac{1}{1-x} \tag{7.1.905}$$

for $|x|<1$. The $\mathcal{Z}$-transform of the following right sided signal $x[n]=a^n u[n]$ is:

$$X(z)=\sum_{n=-\infty}^{+\infty} a^n u[n]z^{-n}=\sum_{n=0}^{+\infty}(az^{-1})^n=\frac{1}{1-az^{-1}}=\frac{z}{z-a} \tag{7.1.906}$$

If we put $a=1$, we simply get the $\mathcal{Z}$-transform of the Heaviside function (as a simple geometric series: $1+z^{-1}+z^{-2}+\ldots$)

For this summation to converge, i.e., for $X(z)$ to exist, it is necessary to have $|az^{-1}|<1$, i.e., the ROC is $|z|>|a|$. As a special case when $a=1$, $x[n]=u[n]$ and we have:

$$\mathcal{Z}[u[n]]=\frac{1}{1-z^{-1}},\quad |z|>1 \tag{7.1.907}$$

E2. Let us see now a simple trivial example of an inverse of the following given $\mathcal{Z}$-transform:

$$X(z)=4z^2+2+3z^{-1} \tag{7.1.908}$$

Comparing this with the definition of $\mathcal{Z}$-transform:

$$X(z)=\sum_{n=-\infty}^{+\infty} x[n]z^{-n}=x[-2]z^2+x[-1]z^1+x[0]+x[1]z^{-1}+x[2]z^{-2} \tag{7.1.909}$$

we get:

$$x[n]=4\delta[n+2]+2\delta[n]+3\delta[n-1] \tag{7.1.910}$$

In general, we can use the time shifting property:

$$\mathcal{Z}[\delta[n+n_0]]=z^{n_0} \tag{7.1.911}$$

to inverse transform the $X(z)$ given above to $x[n]$ directly.

E3. Let us determine the obvious inverse transform of:

$$X(z)=\frac{1}{1-az^{-1}} \tag{7.1.912}$$

As we know that:

$$\frac{1}{1-az^{-1}}=1+az^{-1}+a^2z^{-2}+\cdots \tag{7.1.913}$$

which converges if the ROC is $|z|>|a|$, i.e., $|az^{-1}|<1$ then we get:

$$x[n]=a^n u[n] \tag{7.1.914}$$

7.1.5.3.4 Zeros and Poles of $\mathcal{Z}$-Transform

All $\mathcal{Z}$-transforms in the above examples are rational, i.e., they can be written as a ratio of polynomials of variable z in the general form:

$$X(z)=\frac{N(z)}{D(z)}=\frac{\displaystyle\sum_{k=0}^{M} b_k z^k}{\displaystyle\sum_{k=0}^{N} a_k z^k}=\frac{b_M}{a_N}\frac{\displaystyle\prod_{k=1}^{M}(z-z_{z_k})}{\displaystyle\prod_{k=1}^{N}(z-z_{p_k})} \tag{7.1.915}$$

where $N(z)$ is the numerator polynomial of order M with roots $z_{z_k}, (k=1,2,\cdots,M)$, and $D(z)$ is the denominator polynomial of order N with roots $z_{p_k}, (k=1,2,\cdots,N)$. In general, we assume the order of the numerator polynomial is lower than that of the denominator polynomial, i.e., $M<N$. If this is not the case, we can always expand $X(z)$ into multiple terms so that $M<N$ is true for each of terms.

The "zeros" and "poles" of a rational $X(z)=N(z)/D(z)$ are defined as:

Definitions (#7.1.285):

D1. Each of the roots of the numerator polynomial z_z for which:

$$X(z)\Big|_{z=z_z}=X(z_z)=0 \tag{7.1.916}$$

is a "zero" of $X(z)$.

If the order of $D(z)$ exceeds that of $N(z)$ (i.e., $N>M$), then $X(+\infty)=0$, i.e., there is a zero at infinity:

$$\left.\frac{b_1z+b_0}{a_2z^2+a_1z+a_0}\right|_{z\to+\infty}=0 \tag{7.1.917}$$

D2. Each of the roots of the denominator polynomial z_p for which :

$$X(z)\Big|_{z=z_p}=X(z_p)=+\infty \tag{7.1.918}$$

is a "pole" of $X(z)$.

If the order of $N(z)$ exceeds that of $D(z)$ (i.e., $M>N$), then $X(+\infty)=+\infty$, i.e, there is a pole at infinity:

$$\left.\frac{b_2z^2+b_1z+b_0}{a_1z+a_0}\right|_{z\to+\infty}\to+\infty \tag{7.1.919}$$

Most essential behaviour properties of an LTI system can be obtained graphically from the ROC and the zeros and poles of its transfer function $H(z)$ on the z-plane.

7.1.5.3.5 Properties of $\mathcal{Z}$-Transform

The $\mathcal{Z}$-transform has a set of properties in parallel with that of the Fourier transform (and Laplace transform). The difference is that we need to pay special attention to the ROCs. In the following, we always assume:

$$\mathcal{Z}[x[n]]=X(z)\quad \text{ROC}=R_x \tag{7.1.920}$$

and:

$$\mathcal{Z}[y[n]]=Y(z)\quad \text{ROC}=R_y \tag{7.1.921}$$

- Linearity:

$$\mathcal{Z}[ax[n]+by[n]]=aX(z)+bY(z),\quad \text{ROC}\supseteq(R_x\cap R_y) \tag{7.1.922}$$

While it is obvious that the ROC of the linear combination of $x[n]$ and $y[n]$ should be the intersection of the their individual ROCs $R_x\cap R_y$ in which both $X(z)$ and $Y(z)$ exist, note that in some cases the ROC of the linear combination could be larger than $R_x\cap R_y$. For example, for both $x[n]=a^nu[n]$ and $y[n]=a^nu[n-1]$, the ROC is $|z|>|a|$, but the ROC of their difference $a^nu[n]-a^nu[n-1]=\delta[n]$ is the entire z-plane.

- Time shifting:

$$\mathcal{Z}[x[n-n_0]]=z^{-n_0}X(z),\quad \text{ROC}=R_x \tag{7.1.923}$$

Proof 7.3.27.

$$\mathcal{Z}[x[n-n_0]]=\sum_{n=-\infty}^{+\infty}x[n-n_0]z^{-n} \tag{7.1.924}$$

Define $m=n-n_0$, we have $n=m+n_0$ and:

$$\sum_{m=-\infty}^{+\infty}x[m]z^{-m}z^{-n_0}=z^{-n_0}X(z) \tag{7.1.925}$$

The new ROC is the same as the old one except the possible addition/deletion of the origin or infinity as the shift may change the duration of the signal.

■ Q.E.D.

A slightly different approach lead us to a different result that is more explicit and by the way more useful especially in business applications (we change the notation a bit to be in conformity with the tradition in the financial field). First let us consider the special case (shift of one unit only in the positive direction[9]):

$$\begin{aligned}\mathcal{Z}(f_{n+1})(z)&=\sum_{n=0}^{+\infty}f_{n+1}z^{-n}=z\sum_{n=0}^{+\infty}f_{n+1}z^{-(n+1)}=z\sum_{k=1}^{+\infty}f_kz^{-k}\\&=z\sum_{k=1}^{+\infty}f_kz^{-k}+(zf_0-zf_0)=z\left(\sum_{k=1}^{+\infty}f_kz^{-k}+f_0\right)-zf_0\\&=z\sum_{k=0}^{+\infty}f_kz^{-k}-zf_0=z\mathcal{Z}(f_n)(z)-zf_0\end{aligned} \tag{7.1.926}$$

---

[9] This is then a "unilateral $\mathcal{Z}$-transform" and as we will see it on page 2373, the properties then differ slightly from the bilateral version.

Repeating exactly the same type of procedure we get quite simply in the general way for $m>0$:

$$\begin{aligned}\mathcal{Z}(f_{n+1})(z)&=z^m\mathcal{Z}(f_n)(z)-z^mf_0-z^{m-1}f_1-\ldots-zf_{m-1}\\&=z^m\left[\mathcal{Z}(f_n)(z)-f_0-z^{-1}f_1-\ldots-z^{-m+1}f_{m-1}\right]\end{aligned}\tag{7.1.927}$$

- Time Expansion (Scaling):

$$\mathcal{Z}[x[n/k]]=X(z^k),\quad \text{ROC}=R_x^{1/k}\tag{7.1.928}$$

The discrete signal $x[n]$ cannot be continuously scaled in time as n has to be an integer (for a non-integer n $x[n]$ is zero). Therefore $x[n/k]$ is defined as:

$$x[n/k]\triangleq\begin{cases}x[n/k] & \text{if } n \text{ is a multiple of } k\\ 0 & \text{else}\end{cases}\tag{7.1.929}$$

✍**Example:**

If $x[n]$ is ramp:

n	1	2	3	4	5	6
$x[n]$	1	2	3	4	5	6

then the expanded version $x[n/2]$ is :

n	1	2	3	4	5	6
$n/2$	0.5	1	1.5	2	2.5	3
m		1		2		3
$x[n/2]$	0	1	0	2	0	3

where m is the integer part of n/k.

Proof 7.3.28. The $\mathcal{Z}$-transform of such an expanded signal is:

$$\mathcal{Z}[x[n/k]]=\sum_{n=-\infty}^{+\infty}x[n/k]z^{-n}=\sum_{m=-\infty}^{+\infty}x[m]z^{-km}=X(z^k)\tag{7.1.930}$$

Note that the change of the summation index from n to m has no effect as the terms skipped are all zeros.

■ Q.E.D.

- Convolution:

$$\mathcal{Z}[x[n]*y[n]]=X(z)Y(z),\quad \text{ROC}\supseteq(R_x\cap R_y)\tag{7.1.931}$$

The ROC of the convolution could be larger than the intersection of R_x and R_y, due to the possible pole-zero cancellation caused by the convolution.

- Time Difference:

$$\mathcal{Z}[x[n]-x[n-1]]=(1-z^{-1})X(z), \quad \text{ROC}=R_x \tag{7.1.932}$$

Proof 7.3.29.

$$\mathcal{Z}[x[n]-x[n-1]]=X(z)-z^{-1}X(z)=(1-z^{-1})X(z)=\frac{z-1}{z}X(z) \tag{7.1.933}$$

Note that due to the additional zero $z=1$ and pole $z=0$, the resulting ROC is the same as R_x except the possible deletion of $z=0$ caused by the added pole and/or addition of $z=1$ caused by the added zero which may cancel an existing pole.

■ Q.E.D.

- Time Accumulation:

$$\mathcal{Z}[\sum_{k=-\infty}^{n} x[k]]=\frac{1}{1-z^{-1}}X(z), \quad \text{ROC}\supseteq[R_x\cap(|z|>1)] \tag{7.1.934}$$

Proof 7.3.30. The accumulation of $x[n]$ can be written as its convolution with $u[n]$:

$$u[n]*x[n]=\sum_{k=-\infty}^{+\infty} u[n-k]x[k]=\sum_{k=-\infty}^{n} x[k] \tag{7.1.935}$$

Applying the convolution property, we get:

$$\mathcal{Z}[\sum_{k=-\infty}^{n} x[k]]=\mathcal{Z}[u[n]*x[n]]=\frac{1}{1-z^{-1}}X(z) \tag{7.1.936}$$

as $\mathcal{Z}[u[n]]=1/(1-z^{-1})$.

■ Q.E.D.

- Time Reversal:

$$\mathcal{Z}[x[-n]]=X(1/z) \quad \text{ROC}=1/R_x \tag{7.1.937}$$

Proof 7.3.31.

$$\mathcal{Z}[x[-n]]=\sum_{n=-\infty}^{+\infty} x[-n]z^{-n}=\sum_{m=-\infty}^{+\infty} x[m]\left(\frac{1}{z}\right)^{-m}=X(1/z) \tag{7.1.938}$$

where $m=-n$.

■ Q.E.D.

- Scaling in z-domain:

$$\mathcal{Z}[a^n x[n]]=X\left(\frac{z}{a}\right), \quad \text{ROC}=|a|R_x \tag{7.1.939}$$

Proof 7.3.32.

$$\mathcal{Z}[a^n x[n]] = \sum_{n=-\infty}^{+\infty} x[n] \left(\frac{z}{a}\right)^{-n} = X\left(\frac{z}{a}\right) \tag{7.1.940}$$

In particular, if $a = e^{\mathrm{i}\omega_0}$, the above becomes:

$$\mathcal{Z}[e^{jn\omega_0} x[n]] = X(e^{-\mathrm{i}\omega_0} z) \quad \text{ROC} = R_x \tag{7.1.941}$$

The multiplication by $e^{-\mathrm{i}\omega_0}$ to z corresponds to a rotation by angle ω_0 in the z-plane, i.e., a frequency shift by ω_0. The rotation is either clockwise ($\omega_0 > 0$) or counter clockwise ($\omega_0 < 0$) corresponding to, respectively, either a left-shift or a right shift in frequency domain. The property is essentially the same as the frequency shifting property of discrete Fourier transform.

■ Q.E.D.

- Conjugation:

$$\mathcal{Z}[x^*[n]] = X^*(z^*), \quad \text{ROC} = R_x \tag{7.1.942}$$

Proof 7.3.33. Complex conjugate of the $\mathcal{Z}$-transform of $x[n]$ is:

$$X^*(z) = [\sum_{n=-\infty}^{+\infty} x[n] z^{-n}]^* = \sum_{n=-\infty}^{+\infty} x^*[n](z^*)^{-n} \tag{7.1.943}$$

Replacing z by z^*, we get the desired result.

■ Q.E.D.

- Differentiation in z-domain:

$$\mathcal{Z}[nx[n]] = -z\frac{\mathrm{d}}{\mathrm{d}z} X(z), \quad \text{ROC} = R_x \tag{7.1.944}$$

Proof 7.3.34.

$$\frac{\mathrm{d}}{\mathrm{d}z} X(z) = \sum_{n=-\infty}^{+\infty} x[n] \frac{\mathrm{d}}{\mathrm{d}z}(z^{-n}) = \sum_{n=-\infty}^{+\infty} (-n) x[n] z^{-n-1} = \frac{-1}{z} \sum_{n=-\infty}^{+\infty} n x[n] z^{-n} \tag{7.1.945}$$

i.e.:

$$\mathcal{Z}[nx[n]] = -z\frac{\mathrm{d}}{\mathrm{d}z} X(z) \tag{7.1.946}$$

■ Q.E.D.

✍**Example:**

Taking derivative with respect to z of the right side of:

$$\mathcal{Z}[a^n u[n]] = \frac{1}{1-az^{-1}} \quad |z|>|a| \tag{7.1.947}$$

we get:

$$\frac{\mathrm{d}}{\mathrm{d}z}\left[\frac{1}{1-az^{-1}}\right] = \frac{-az^{-2}}{(1-az^{-1})^2} \tag{7.1.948}$$

Due to the property of differentiation in z-domain, we have:

$$\mathcal{Z}[na^n u[n]] = \frac{az^{-1}}{(1-az^{-1})^2} \quad |z|>|a| \tag{7.1.949}$$

Note that for a different ROC $|z|<|a|$, we have:

$$\mathcal{Z}[-na^n u[-n-1]] = \frac{az^{-1}}{(1-az^{-1})^2} \quad |z|<|a| \tag{7.1.950}$$

7.1.5.3.6 Usual $\mathcal{Z}$-Transforms

As we always do in this textbook, let us see some common usual $\mathcal{Z}$-Transforms that are useful in engineering in some well known problems:

1. $\delta[n]$, $\delta[n-m]$

$$\mathcal{Z}[\delta[n]] = \sum_{n=-\infty}^{+\infty} \delta[n]z^{-n} = \delta[0]z^{-0} = \delta[0] = 1 \qquad \forall z \tag{7.1.951}$$

Due to the time shifting property, we also have:

$$\mathcal{Z}[\delta[n-m]] = z^{-m} \qquad \forall z \tag{7.1.952}$$

2. $u[n]$, $a^n u[n]$, $na^n u[n]$

$$\mathcal{Z}[u[n]] = \sum_{n=0}^{+\infty} z^{-n} = \frac{1}{1-z^{-1}} = \frac{z}{z-1} \qquad |z|>1 \tag{7.1.953}$$

Due to the scaling in z-domain property, we have:

$$\mathcal{Z}[a^n u[n]] = \frac{1}{1-(z/a)^{-1}} = \frac{1}{1-az^{-1}} \qquad S|z|>|a| \tag{7.1.954}$$

Or for those that don't like using this property, we have simply using the previous result:

$$\mathcal{Z}[a^n u[n]] = \sum_{n=0}^{+\infty} a^n z^{-n} = \sum_{n=0}^{+\infty}(a^{-1}z)^{-n} = \frac{1}{1-(z/a)^{-1}} = \frac{\frac{z}{a}}{\frac{z}{a}-1} = \frac{z}{z-a} \tag{7.1.955}$$

Applying the property of differentiation in z-domain to the above, we have:

$$\mathcal{Z}[na^n u[n]]=-z\frac{\mathrm{d}}{\mathrm{d}z}\left[\frac{1}{1-az^{-1}}\right]=-z\frac{-az^{-2}}{(1-az^{-1})^2}=\frac{az^{-1}}{(1-az^{-1})^2}\qquad |z|>|a| \tag{7.1.956}$$

3. $u[n-k]$

   Using the time shifting property that is for recall given by:

$$\mathcal{Z}(x[n-n_0])=z^{-n_0}X(z) \tag{7.1.957}$$

   we have:

$$\mathcal{Z}(u[n-k])=z^{-k}X(u[n])=z^{-k}\frac{z}{z-1}=\frac{z^{1-k}}{z-1} \tag{7.1.958}$$

✍**Example:**

A useful case for a simple business application that we will use later is:

$$\mathcal{Z}(u[n-k])=\mathcal{Z}(u[n-l-1])=\mathcal{Z}(u[n-(l+1)])=\frac{z^{2-l}}{z-1} \tag{7.1.959}$$

4. $e^{\pm in\omega_0}u[n]$, $\cos[n\omega_0]u[n]$, $\sin[n\omega_0]u[n]$

   Applying the scaling in z-domain property to $\mathcal{Z}[u[n]]=1/(1-z^{-1})$, we have:

$$\mathcal{Z}[e^{in\omega_0}u[n]]=\frac{1}{1-(e^{\mathrm{i}\omega_0}z)^{-1}}=\frac{1}{1-e^{-\mathrm{i}\omega_0}z^{-1}}\qquad |z|>1 \tag{7.1.960}$$

   and similarly, we have:

$$\mathcal{Z}[e^{-in\omega_0}u[n]]=\frac{1}{1-e^{\mathrm{i}\omega_0}z^{-1}}\qquad |z|>1 \tag{7.1.961}$$

   Moreover, we have:

$$\begin{aligned}\mathcal{Z}[\cos(n\omega_0)u[n]]=\mathcal{Z}\left[\frac{e^{in\omega_0}+e^{-in\omega_0}}{2}u[n]\right]&=\frac{1}{2}\left[\frac{1}{1-e^{\mathrm{i}\omega_0}z^{-1}}+\frac{1}{1-e^{-\mathrm{i}\omega_0}z^{-1}}\right]\\&=\frac{2-(e^{\mathrm{i}\omega_0}+e^{-\mathrm{i}\omega_0})z^{-1}}{2[1-(e^{\mathrm{i}\omega_0}+e^{-\mathrm{i}\omega_0})z^{-1}+z^{-2}]}\\&=\frac{1-\cos(\omega_0)z^{-1}}{1-2\cos(\omega_0)z^{-1}+z^{-2}}\qquad |z|>1\end{aligned} \tag{7.1.962}$$

   Similarly we have:

$$\mathcal{Z}[\sin(n\omega_0)u[n]]=\frac{\sin\left(\omega_0 z^{-1}\right)}{1-2\cos\left(\omega_0 z^{-1}\right)+z^{-2}}\qquad |z|>1 \tag{7.1.963}$$

5. $r^n\cos[n\omega_0]u[n]$, $r^n\sin[n\omega_0]u[n]$

Applying the z-domain scaling property to the above, we have:

$$\mathcal{Z}[r^n \cos(n\omega_0)u[n]] = \frac{1 - r\ \cos\left(\omega_0 z^{-1}\right)}{1 - 2r\ \cos\left(\omega_0 z^{-1}\right) + r^2 z^{-2}} \qquad |z| > r \tag{7.1.964}$$

and:

$$\mathcal{Z}[r^n \sin(n\omega_0)u[n]] = \frac{r\ \sin\left(\omega_0 z^{-1}\right)}{1 - 2r\ \cos\left(\omega_0 z^{-1}\right) + r^2 z^{-2}} \qquad |z| > r \tag{7.1.965}$$

To summarize:

$x[n]$	$X(z) = \mathcal{Z}(x[n])$	**ROC**
$A\delta[n]$	A	Everywhere
$A\delta\left[n - n_0\right]$	Az^{-n_0}	Everywhere
$A \cdot u[n]$	$\frac{A}{1-z^{-1}} = \frac{Az}{z-1}$	$\lvert z\rvert > 1$
$u[n-k]$	$\frac{z^{1-k}}{z-1}$	$\lvert z\rvert > 1$
$Aa^n \cdot u[n]$	$\frac{A}{1-az^{-1}} = \frac{Az}{z-a}$	$\lvert z\rvert > \lvert a\rvert$
$nA^n a^n \cdot u[n]$	$\frac{Az^{-1}}{(1-Az)^2}$	$\lvert z\rvert > \lvert a\rvert$
$Aa^n e^{\mathrm{i}\omega n} \cdot u[n]$	$\frac{A}{1-ae^{j\omega}z^{-1}} = \frac{Az}{z-ae^{j\omega}}$	$\lvert z\rvert > \lvert a\rvert$
$Aa^n \cos(\omega n) \cdot u[n]$	$\frac{Az(z-a\cos\omega)}{z^2-(2a\cos\omega)z+a^2}$	$\lvert z\rvert > \lvert a\rvert$
$Aa^n \cos(\omega n + \alpha) \cdot u[n]$	$\frac{\frac{A}{2}e^{j\alpha}}{1-ae^{\mathrm{i}\omega}z^{-1}} + \frac{\frac{A}{2}e^{-j\alpha}}{1-ae^{-j\omega}z^{-1}}$	$\lvert z\rvert > \lvert a\rvert$
$A\cos(\omega n) \cdot u[n]$	$\frac{Az(z-\cos\omega)}{z^2-(2\cos\omega)z+1}$	$\lvert z\rvert > 1$
$A\sin(\omega n) \cdot u[n]$	$\frac{Az\sin\omega}{z^2-(2\cos\omega)z+1}$	$\lvert z\rvert > 1$
$An \cdot u[n]$	$\frac{Az^{-1}}{(1-z^{-1})^2} = \frac{Az}{(z-1)^2}$	$\lvert z\rvert > 1$
$An^2 \cdot u[n]$	$\frac{Az^{-1}\left(1+z^{-1}\right)}{(1-z^{-1})^3} = \frac{Az(z+1)}{(z-1)^2}$	$\lvert z\rvert > 1$
$Ana^n \cdot u[n]$	$\frac{Aaz^{-1}}{(1-az^{-1})^2} = \frac{Aaz}{(z-a)^2}$	$\lvert z\rvert > \lvert a\rvert$
$Aa^n \cdot u[-n-1]$	$\frac{-A}{1-az^{-1}} = \frac{-Az}{z-a}$	$\lvert z\rvert < \lvert a\rvert$

Table 7.11: **Most common $\mathcal{Z}$-Transforms**

7.1.5.3.7 Unilateral $\mathcal{Z}$-Transform

The "unilateral $\mathcal{Z}$-transform" of an arbitrary signal $x[n]$ is defined as:

$$\boxed{\mathcal{UZ}[x[n]] = X(z) \triangleq \sum_{n=-\infty}^{+\infty} x[n]u[n]z^{-n} = \sum_{n=0}^{+\infty} x[n]z^{-n}} \tag{7.1.966}$$

where for recall $\mathcal{U}$ and $u[n]$ are another typical notation of the Heaviside function.

Some of the properties of the unilateral $\mathcal{Z}$-transform that differ slightly from the bilateral $\mathcal{Z}$-transform are listed below:

P1. Time Advance:

$$\begin{aligned}\mathcal{UZ}\left[x[n+1]\right] &= \sum_{n=0}^{+\infty} x[n+1]z^{-n} = z\sum_{m=1}^{+\infty} x[m]z^{-m} \\ &= z\left[\sum_{m=0}^{+\infty} x[m]z^{-m} - x[0]\right] = zX(z) - zx[0]\end{aligned} \tag{7.1.967}$$

where we have assumed $m = n+1$.

P2. Time Delay:

$$\begin{aligned}\mathcal{UZ}\left[x[n-1]\right] &= \sum_{n=0}^{+\infty} x[n-1]z^{-n} = z^{-1}\sum_{m=-1}^{+\infty} x[m]z^{-m} \\ &= z^{-1}\left[\sum_{m=0}^{+\infty} x[m]z^{-m} + zx[-1]\right] = z^{-1}X(z) + x[-1]\end{aligned} \tag{7.1.968}$$

where $m = n-1$. Similarly, we have:

$$\begin{aligned}\mathcal{UZ}[x[n-2]] &= \sum_{n=0}^{+\infty} x[n-2]z^{-n} = z^{-2}\sum_{m=-2}^{+\infty} x[m]z^{-m} \\ &= z^{-2}\left[\sum_{m=0}^{+\infty} x[m]z^{-m} + zx[-1] + z^2x[-2]\right] \\ &= z^{-2}X(z) + x[-1]z^{-1} + x[-2]\end{aligned} \tag{7.1.969}$$

where $m = n-2$. In general, we have:

$$\mathcal{UZ}[x[n-n_0]] = z^{-n_0}X(z) + \sum_{k=0}^{n_0-1} z^{-k}x[k-n_0] \tag{7.1.970}$$

P3. Convolution:

$$\mathcal{UZ}[x[n]*y[n]] = X(z)Y(z) \tag{7.1.971}$$

If both $x[n]$ and $y[n]$ are causal, i.e., $x[n] = y[n] = 0$ for $n < 0$, the unilateral and bilateral $\mathcal{Z}$ transforms are identical.

P4. Time Difference:

$$\mathcal{UZ}[x[n] - x[n-1]] = (1-z^{-1})X(z) \quad \text{ROC} = R_x \tag{7.1.972}$$

Proof 7.3.35.

$$\mathcal{UZ}[x[n] - x[n-1]] = X(z) - z^{-1}X(z) - x[-1] = (1-z^{-1})X(z) - x[-1] \tag{7.1.973}$$

■ Q.E.D.

P5. Time Accumulation:

$$\mathcal{UZ}\left[\sum_{k=0}^{n} x[k]\right] = \frac{1}{1-z^{-1}}X(z) = \frac{z}{z-1}X(z) \tag{7.1.974}$$

P6. Initial Value Theorem:

If $x[n]=x[n]u[n]$, i.e., $x[n]=0$ for $n<0$, then:

$$x[0]=\lim_{z\rightarrow+\infty}X(z) \tag{7.1.975}$$

Proof 7.3.36.

$$\lim_{z\rightarrow+\infty}X(z)=\lim_{z\rightarrow+\infty}\left[\sum_{n=0}^{+\infty}x[n]z^{-n}\right]=x[0] \tag{7.1.976}$$

All terms with $n>0$ become zero as $z^{-n}=1/z^{n}\rightarrow 0$ as $z\rightarrow+\infty$, except the first one which is always $x[0]$.

■ Q.E.D.

P7. Final Value Theorem:

If $x[n]=x[n]u[n]$, i.e., $x[n]=0$ for $n<0$, then:

$$\lim_{n\rightarrow+\infty}x[n]=x[+\infty]=\lim_{z\rightarrow 1}(1-z^{-1})X(z) \tag{7.1.977}$$

Proof 7.3.37.

$$\mathcal{Z}\left[x[n]-x[n-1]\right]=X(z)-X(z)z^{-1}=\sum_{n=0}^{+\infty}\left[x[n]-x[n-1]\right]z^{-n} \tag{7.1.978}$$

i.e.:

$$(1-z^{-1})X(z)=\lim_{N\rightarrow+\infty}\sum_{n=0}^{N}\left[x[n]-x[n-1]\right]z^{-n} \tag{7.1.979}$$

Letting $z\rightarrow 1$ in the above, we get:

$$\begin{aligned}\lim_{z\rightarrow 1}(1-z^{-1})X(z)&=\lim_{N\rightarrow+\infty}\left[\sum_{n=0}^{N}[x[n]-x[n-1]\right]\lim_{N\rightarrow+\infty}[x[N]-x[-1]]\\&=\lim_{N\rightarrow+\infty}x[N]=x[+\infty]\end{aligned} \tag{7.1.980}$$

where $x[-1]=0$.

■ Q.E.D.

To finish depending on donations

7.1.5.4 Hilbert Transform

The information about the Hilbert transform is often scattered in most textbooks about signal processing. Their authors frequently use mathematical formulas without explaining them thoroughly to the reader at the opposite of the Fourier, Laplace or $\mathcal{Z}$-transform. Our purpose is to make a more stringent presentation of the Hilbert transform but still with the signal processing application in mind.

Definition (#7.1.286): The "Hilbert transform" of a function $f(t)$ is defined for all t by:

$$\boxed{X(t)=\frac{1}{\pi}\mathrm{P}\int\limits_{-\infty}^{+\infty}\frac{f(\tau)}{t-\tau}\mathrm{d}\tau} \qquad (7.1.981)$$

when the integral exists.

It is normally not possible to calculated the Hilbert transform as an ordinary improper integral because of the pole at $\tau=t$. However, the P in front of the integral denotes the "Cauchy principal value" which expanding the class of function for which the integral definition above exist. It can be defined as following:

Definition (#7.1.287): The "Cauchy principal value" is a method for assigning values to certain improper integrals (see section Differential and Integral Calculus page 1595) which would otherwise be undefined. Depending on the type of singularity in the integrand f, the Cauchy principal value is defined according to the following rules:

- For a singularity at the finite number b:

$$\lim_{\varepsilon\to 0^+}\left[\int\limits_a^{b-\varepsilon} f(x)\mathrm{d}x+\int\limits_{b+\varepsilon}^{c} f(x)\mathrm{d}x\right] \qquad (7.1.982)$$

  where b is a point at which the behaviour of the function f is such that:

$$\begin{gathered}\int\limits_a^b f(x)\mathrm{d}x=\pm\infty \text{ for any } a<b \text{ and}\\ \int_b^c f(x)\mathrm{d}x=\mp\infty \text{ for any } c>b\end{gathered} \qquad (7.1.983)$$

- For a singularity at infinity:

$$\lim_{a\to+\infty}\int\limits_{-a}^{a} f(x)\mathrm{d}x \qquad (7.1.984)$$

  where:

$$\int\limits_{-\infty}^{0} f(x)\mathrm{d}x=\pm\infty \qquad \text{and} \qquad \int\limits_{0}^{+\infty} f(x)\mathrm{d}x=\mp\infty \qquad (7.1.985)$$

In some cases it is necessary to deal simultaneously with singularities both at a finite number b and at infinity. This is usually done by a limit of the form:

$$\lim_{\varepsilon\to 0+}\left[\int\limits_{b-\frac{1}{\varepsilon}}^{b-\varepsilon} f(x)\mathrm{d}x+\int\limits_{b+\varepsilon}^{b+\frac{1}{\varepsilon}} f(x)\mathrm{d}x\right] \tag{7.1.986}$$

The Cauchy principal value can also be defined in terms of contour integrals of a complex-valued function $f(z);\, z=x+\mathrm{i}y$, with a pole on a contour C. Define $C(\varepsilon)$ to be the same contour where the portion inside the disk of radius ε around the pole has been removed. Provided the function $f(z)$ is integrable over $C(\varepsilon)$ no matter how small ε becomes, then the Cauchy principal value is the limit:

$$\mathrm{P}\int\limits_{C} f(z)\mathrm{d}z=\lim_{\varepsilon\to 0+}\int\limits_{C(\varepsilon)} f(z)\mathrm{d}z \tag{7.1.987}$$

✍**Example:**

Consider the ill-defined expression:

$$\int\limits_{-1}^{+1}\frac{\mathrm{d}x}{x} \tag{7.1.988}$$

As the primitive of $1/x$ is equal to $\ln(x)$ there is obviously a problem... So the idea is to write instead:

$$\lim_{a\to 0+}\left(\int\limits_{-1}^{-a}\frac{\mathrm{d}x}{x}+\int\limits_{a}^{+1}\frac{\mathrm{d}x}{x}\right) \tag{7.1.989}$$

but as the $1/x$ function is odd, we can write this:

$$\lim_{a\to 0+}\left(-\int\limits_{a}^{+1}\frac{\mathrm{d}x}{x}+\int\limits_{a}^{+1}\frac{\mathrm{d}x}{x}\right)=\lim_{a\to 0+}(0) \tag{7.1.990}$$

that is obviously equal to 0! So 0 is the principal value of the initial integral.

Remark

Sadly different authors use different notations for the Cauchy principal value of a function f, among others:

$$P\int f(x)\mathrm{d}x,\, PV\int f(x)\mathrm{d}x,\, \text{p.v.}\int f(x)\mathrm{d}x,\, \text{P.V.}\int f(x)\mathrm{d}x,\, \mathcal{P}\int f(x)\mathrm{d}x,\, P_v\int f(x)\mathrm{d}x,\, \int_L^{*} f(z)\mathrm{d}z,\, ⨍ f(x)\mathrm{d}x,\ldots$$

To finish depending on donations

7.1.6 Functional dot product (inner product)

The "functional dot product" (very strong analogy with the dot product in seen in the section Vector Calculus) may seem unnecessary when examined for the first time outside of an application context or only as generalization purpose, but in fact it has many practical applications. We will make such direct use in the section of Wave Quantum Physics and Quantum Chemistry, or even more important in the context of trigonometric polynomials through the Fourier series and transforms (see section Sequences and Series page 1727) that we find everywhere in contemporary physics and computer science.

However, if the reader has not travelled the section of Vector Calculus and the part treating the vector dot product, we would highly recommend reading it otherwise what follows may be a little incomprehensible.

We put ourselves in the space $\mathcal{C}([a,b],\mathbb{R})$ of continuous functions in the interval $[a,b]$ into $\mathbb{R}$ with the inner product defined by (we find here again the specific notation of the dot product in its functional - i.e. integral - version as we had mentioned during our definition of the vector dot product in the section of Vector Calculus):

$$\boxed{\langle f|g\rangle=\int\limits_a^b f(x)\cdot g(x)\mathrm{d}x} \tag{7.1.991}$$

A family of orthogonal polynomials, as we can make the analogy with the dot product in the section Vector Calculus, is therefore a polynomial family $(p_0,\ldots,p_n,\ldots)$ such as:

$$\langle p_j|p_k\rangle=0 \tag{7.1.992}$$

if $j\neq k$. We recall that an orthogonal family is a free family. We also saw in the section of Vector Calculus that in the space $\mathcal{C}([a,b],\mathbb{C})$ the only possible coherent choice was:

$$\boxed{\langle f|g\rangle=\int\limits_a^b f(x)\cdot \overline{g(x)}\mathrm{d}x} \tag{7.1.993}$$

We name the two previous relations "L^2-dot product".

Remember now that if we want to show something is an inner product, we need to show three things for all $f,g\in\mathbb{R}^2$ and $\alpha\in\mathbb{R}$:

1. Symmetry: $\langle f|g\rangle=\langle g|f\rangle$ (or, if the field is the complex numbers $\mathbb{C}^n$, $\langle f|g\rangle=\overline{\langle g|f\rangle}$, i.e. "conjugate symmetry)
2. Linearity: $\langle \alpha f|g\rangle=\alpha\langle f|g\rangle$. Notice this also implies $\langle f|\alpha g\rangle=\alpha\langle f|g\rangle$ ($\bar{\alpha}$ in the complex case) by symmetry
3. Positive-definite: $\langle f,f\rangle\geq 0$ with equality if and only if $f=0$, the zero function

The first two properties follow directly from the definition of an integral. For the third property, it's quite obvious (the integral of f^2 can be zero only if f is zero).

Remark

If such a dot product exist, we can then the Cauchy-Schwarz inequality also applies to it. We name it the "Cauchy-Schwarz inequality for integrals".

Therefore we can build the "L^2-norm":

$$\|f\|_{L^2}=\sqrt{\langle f|g\rangle}=\left(\int_a^b f(x)\cdot\overline{g(x)}\mathrm{d}x\right)^{1/2} \tag{7.1.994}$$

Remark

We will see further below that the definition above is not the most general one as especially physicists and engineers say that functions are orthogonal under a more constraint situation!

The development that will follow will remind us the Gram-Schmidt procedure (see section Vector Calculus page 2062) to build an orthogonal family.

Theorem 7.4. Given $(p_0,\ldots,p_n,\ldots)$ a family of linearly independent polynomial defined on $[a,b]$ and V the vector space defined by this family. The family $(y_0,\ldots,y_n,\ldots)$ defined by recurrence in the following way:

$$y_n=p_n-\sum_{j=0}^{n-1}\langle p_n|y_j\rangle\frac{y_j}{\|y_j\|^2}\text{ for }n>1 \tag{7.1.995}$$

and $y_0=p_0$ is orthogonal and generates V.

Proof 7.4.1. Let us show by induction on n that $(y_0,y\ldots,y_n,\ldots)$ is an orthogonal family which generates the same space as $(p_0,\ldots,p_n,\ldots)$. The assertion holds for $n=0$. Let us suppose that the assertion holds for $n\geq 0$, for $0\leq k\leq n$ we have:

$$\begin{aligned}\langle y_{n+1}|y_k\rangle&=\left\langle p_{n+1}-\sum_{j=0}^{n}\langle p_{n+1}|y_j\rangle\cdot\frac{y_j}{\|y_j\|^2}\middle|y_k\right\rangle=\langle p_{n+1}|y_k\rangle-\sum_{j=0}^{n}\langle p_{n+1}|y_j\rangle\cdot\langle y_j|y_k\rangle\cdot\frac{1}{\|y_j\|^2}\\&=\langle p_{n+1}|y_k\rangle-\langle p_{n+1}|y_k\rangle\cdot\langle y_k|y_k\rangle\cdot\frac{1}{\|y_k\|^2}=\langle p_{n+1}|y_k\rangle-\langle p_{n+1}|y_k\rangle=0\end{aligned} \tag{7.1.996}$$

$(y_0,\ldots,y_n,\ldots)$ is therefore orthogonal. Finally, we get the equality:

$$y_{n+1}=p_{n+1}-\sum_{j=0}^{n}\langle p_{n+1}|y_j\rangle\cdot\frac{y_j}{\|y_j\|^2} \tag{7.1.997}$$

■ Q.E.D.

Let us see a first example very important is signal processing and statistics that is relatively to frequency analysis:

✍**Example:**

Let us consider the very important example in modern physics that is the set of continuous 2π-periodic function denoted $P_{2\pi}$ that forms a vector space (see section Set Theory page 1805).

We define the dot product of two functions of this set by:

$$\forall f,g\in P_{2\pi}\quad \langle f|g\rangle=\int\limits_{-\pi}^{\pi} f(x)g(x)\mathrm{d}x=\int\limits_{0}^{-\pi} f(x)g(x)\mathrm{d}x \tag{7.1.998}$$

The aim of this definition is to build an abstract functional basis $P_{2\pi}$ on which we can break down any 2π-periodic function!!!

The simplest idea is then to use the trigonometric functions sine and cosine:

$$\begin{gathered}\cos(x),\cos(2x),\cos(3x),\ldots,\cos(nx)\\ \sin(x),\sin(2x),\sin(3x),\ldots,\sin(nx)\end{gathered} \tag{7.1.999}$$

The relations below show that the basis chosen above are orthogonal and therefore form a free family, plus it's a generating family of the vector space $P_{2\pi}$ because as we have seen in our study of Fourier series (see section Sequences and Series page 1727) we have the following values:

$$\begin{gathered}\int\limits_{0}^{2\pi}\cos(kx)\cos(mx)\mathrm{d}x=\delta_{km}\pi\\ \int\limits_{0}^{2\pi}\sin(kx)\sin(mx)\mathrm{d}x=\delta_{km}\pi\\ \int\limits_{0}^{2\pi}\cos(kx)\sin(mx)\mathrm{d}x=0\end{gathered} \tag{7.1.1000}$$

where δ_{km} is the kronecker symbol (see section Tensor Calculus page 2044).

Therefore it is also an orthogonal basis but not orthonormal. If we want to normalized the vectors of the basis we just need obviously to take:

$$\begin{gathered}\frac{1}{\sqrt{\pi}}\cos(x),\frac{1}{\sqrt{\pi}}\cos(2x),\frac{1}{\sqrt{\pi}}\cos(3x),\ldots,\frac{1}{\sqrt{\pi}}\cos(nx)\\ \frac{1}{\sqrt{\pi}}\sin(x),\frac{1}{\sqrt{\pi}}\sin(2x),\frac{1}{\sqrt{\pi}}\sin(3x),\ldots,\frac{1}{\sqrt{\pi}}\sin(nx)\end{gathered} \tag{7.1.1001}$$

Remark

If the reader remembers that for a random variable X defined on $\mathbb{R}$, the mean was calculated as (see section Statistics page 634):

$$\mathrm{E}(X)=\int\limits_{-\infty}^{+\infty} xf(x)\mathrm{d}x \qquad (7.1.1002)$$

Then we can assimilate:

$$\mathrm{E}(g(x))=\int\limits_{-\infty}^{+\infty} g(x)f(x)\mathrm{d}x \qquad (7.1.1003)$$

where:

$$\int\limits_{-\infty}^{+\infty} f(x)\mathrm{d}x=1 \qquad (7.1.1004)$$

to the expected mean of the function $g(x)$! Analogy sometimes very useful in practice!

Let us see now another example that is an extension of the previous one and that has also a great importance in signal processing but also in quantum physics and quantum chemistry:

✍**Example:**

Let us consider a basis of complex functions of the form $re^{\mathrm{i}n\varphi}$ with $n\in\mathbb{Z}$. We therefore can write:

$$\langle ae^{\mathrm{i}m\varphi}|be^{\mathrm{i}m\varphi}\rangle=\int\limits_{-\pi}^{+\pi} ae^{\mathrm{i}m\varphi}\cdot\overline{be^{\mathrm{i}n\varphi}}\mathrm{d}\varphi=ab\int\limits_{-\pi}^{+\pi} e^{\mathrm{i}(m-n)\varphi}\mathrm{d}\varphi \qquad (7.1.1005)$$

We get:

$$ab\int\limits_{-\pi}^{+\pi} e^{\mathrm{i}(m-n)\varphi}\mathrm{d}\varphi=\begin{cases} ab\displaystyle\int\limits_{-\pi}^{+\pi} 1\mathrm{d}x=ab2\pi & m=n \\ ab\left[\dfrac{-\mathrm{i}}{m-n}e^{\mathrm{i}(m-n)\varphi}\right]_{-\pi}^{+\pi}=0 & m\neq n \end{cases} \qquad (7.1.1006)$$

It is obvious that if we take for basis functions of the type:

$$\varphi_n=\frac{1}{2\pi}e^{\mathrm{i}n\varphi} \qquad (7.1.1007)$$

then we have an orthonormal basis (and not just an orthogonal one).

Let us see again another example that will be useful for us in the section of Wave Mechanics for our study of the ideal circular membrane of a drum.

✍**Example:**

We have proved in the section Sequences and Series that the Bessel function (see page 1768) $J_p(x)$ satisfies the following differential equation (Bessel's equation):

$$x^2\frac{\mathrm{d}^2 J_p}{\mathrm{d}x^2}+x\frac{\mathrm{d}J_p}{\mathrm{d}x}+(x^2-p^2)J_p=0 \tag{7.1.1008}$$

which can be written as:

$$x\frac{\mathrm{d}}{\mathrm{d}x}\left(x\frac{\mathrm{d}J_p}{\mathrm{d}x}\right)+(x^2-p^2)J_p=0 \tag{7.1.1009}$$

The variable p need not be an integer as we will see it in the section of Mechanical Engineering with the study case of the self-buckling column.

It turns out to be useful to define a new variable t by $x=at$, where a is a constant which we will take to be a zero of J_p, i.e. $J_p(a)=0$. Let us define:

$$u(t)=J_p(at) \tag{7.1.1010}$$

which implies:

$$u(1)=0 \tag{7.1.1011}$$

and substituting into (7.1.1009) gives:

$$t\frac{\mathrm{d}}{\mathrm{d}t}\left(t\frac{\mathrm{d}u}{\mathrm{d}t}\right)+(a^2t^2-p^2)u=0 \tag{7.1.1012}$$

since $x\mathrm{d}/\mathrm{d}x$ is equivalent to $t\mathrm{d}/\mathrm{d}t$.

We can also write down the equation obtained by picking another zero, b. Defining:

$$v(t)=J_p(bt) \tag{7.1.1013}$$

which implies:

$$v(1)=0 \tag{7.1.1014}$$

we have then:

$$t\frac{\mathrm{d}}{\mathrm{d}t}\left(t\frac{\mathrm{d}v}{\mathrm{d}t}\right)+(b^2t^2-p^2)v=0 \tag{7.1.1015}$$

To derive the orthogonality relation, we multiply (7.1.1012) by v, and (7.1.1015) by u. Subtracting and dividing by t gives:

$$v\frac{\mathrm{d}}{\mathrm{d}t}\left(t\frac{\mathrm{d}u}{\mathrm{d}t}\right)-u\frac{\mathrm{d}}{\mathrm{d}t}\left(t\frac{\mathrm{d}v}{\mathrm{d}t}\right)+(a^2-b^2)tuv=0 \tag{7.1.1016}$$

The first two terms in (7.1.1016) can be combined as:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(vt\frac{\mathrm{d}u}{\mathrm{d}t}-ut\frac{\mathrm{d}v}{\mathrm{d}t}\right) \tag{7.1.1017}$$

since the extra terms present in (7.1.1017), but not in (7.1.1016), when the derivatives are expanded out are equal and opposite and so cancel. Hence we have:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(vt\frac{\mathrm{d}u}{\mathrm{d}t}-ut\frac{\mathrm{d}v}{\mathrm{d}t}\right)+(a^2-b^2)tuv=0 \tag{7.1.1018}$$

We next integrate this over the range of t from 0 to 1 (0 since the Bessel function is not defined for $t<0$ and to 1 since it's the place where there is a zero by construction for recall!), which gives:

$$\left[vt\frac{\mathrm{d}u}{\mathrm{d}t}-ut\frac{\mathrm{d}v}{\mathrm{d}t}\right]_0^1+(a^2-b^2)\int_0^1 tu(t)v(t)\mathrm{d}t=0 \tag{7.1.1019}$$

The integrated term vanishes at the lower limit because $t=0$, and it also vanishes at the upper limit because $u(1)=v(1)=0$, see (7.1.1011) and (7.1.1013). Hence, if $a\neq b$, (7.1.1019) gives:

$$\int_0^1 tu(t)v(t)\mathrm{d}t=0 \tag{7.1.1020}$$

which, using (7.1.1010) and (7.1.1013), can be written:

$$\boxed{\int_0^1 tJ_p(at)J_p(bt)\mathrm{d}t=0} \tag{7.1.1021}$$

This is the desired orthogonality equation! Remember we require that a and b are distinct zeroes of J_p, so both Bessel functions in (7.1.1021) vanish at the upper limit.

Let us see again another another example that will be useful for us in the section of Wave Quantum Physics during our study of the harmonic oscillator (see page 4817):

✍**Example:**

We will introduce in the section of Wave Quantum Physics the following "physicist Hermite polynomial":

$$H_n(x)=(-1)^n e^{x^2}\frac{\mathrm{d}^n}{\mathrm{d}x^n}e^{-x^2} \tag{7.1.1022}$$

Therefore (see the plot in the section of Wave Quantum Physics at page 4817):

$$\begin{aligned}
&H_0=1\\
&H_1=2x\\
&H_2=4x^2-2\\
&H_3=8x^3-12x\\
&H_4=16x^4-48x+12\\
&H_5=32x^5-160x^3+120x\\
&H_6=64x^6-480x^4+720x^2-120\\
&\ldots
\end{aligned} \tag{7.1.1023}$$

where we notice almost immediately that (useful for further below):

$$\frac{\mathrm{d}^n}{\mathrm{d}x^n}H_n=2^n n! \tag{7.1.1024}$$

And we need to prove that they are orthogonal (or even better: orthonormal!).

For this purpose we introduce the weight function $w(x)=e^{-x^2}$, therefore:

$$\begin{aligned}
\int\limits_{-\infty}^{+\infty}H_n(x)H_m(x)\,\mathrm{e}^{-x^2}\,\mathrm{d}x&=(-1)^m\int\limits_{-\infty}^{+\infty}\left(\frac{\mathrm{d}^m}{\mathrm{d}x^m}\mathrm{e}^{-x^2}\right)H_n(x)\mathrm{d}x\\
&=(-1)^{m-n}\int\limits_{-\infty}^{+\infty}\left(\frac{\mathrm{d}^{m-n}}{\mathrm{d}x^{m-n}}\mathrm{e}^{-x^2}\right)H_n^{(n)}(x)\mathrm{d}x\\
&=(-1)^{m-n}2^n n!\int\limits_{-\infty}^{+\infty}\frac{\mathrm{d}^{m-n}}{\mathrm{d}x^{m-n}}\mathrm{e}^{-x^2}\mathrm{d}x
\end{aligned} \tag{7.1.1025}$$

So we get (we use the Gauss integral as proved in the section Statistics page 695):

$$\int\limits_{-\infty}^{+\infty}H_n(x)H_m(x)\,\mathrm{e}^{-x^2}\,\mathrm{d}x=\begin{cases}\text{if } m>n: & 0\\ \text{if } m=n: & n!2^n\displaystyle\int\limits_{-\infty}^{+\infty}\mathrm{e}^{-x^2}\mathrm{d}x=n!2^n\sqrt{\pi}\end{cases} \tag{7.1.1026}$$

Finally, using Kronecker symbol (see section Tensor Calculus page 2044):

$$\boxed{\int\limits_{-\infty}^{+\infty} H_n(x)H_m(x)\,\mathrm{e}^{-x^2}\mathrm{d}x=2^n n!\sqrt{\pi}\delta_{nm}} \tag{7.1.1027}$$

Now let us see a last example useful for our study of the non-rigid rotator in Quantum Chemistry (radial part solution of the hydrogenoid Schrödinger equation):

✍**Example:**

If $L_m(x)$ and $L_n(x)$ are Laguerre's polynomials (see section Calculus page 1404) m,n being positive integers, then:

$$\int\limits_{0}^{+\infty} e^{-x}L_n(x)L_m(x)\mathrm{d}x=\delta_{mn} \tag{7.1.1028}$$

where:

$$\delta_{mn}=\begin{cases}0 & m\neq n\\ 1 & m=n\end{cases} \tag{7.1.1029}$$

The generating function for Laguerre's polynomial gives:

$$\begin{gathered}\sum_{n=0}^{+\infty}L_n(x)t^n=\frac{e^{-\frac{xt}{1-t}}}{1-t}\\ \sum_{m=0}^{+\infty}L_m(x)s^m=\frac{e^{-\frac{xs}{1-s}}}{1-s}\\ \Rightarrow\quad\sum_{n=0}^{+\infty}\sum_{m=0}^{+\infty}L_n(x)L_m(x)t^ns^m=\frac{e^{-x\{t/(1-t)+s/(1-s)\}}}{(1-t)(1-s)}\end{gathered} \tag{7.1.1030}$$

We now multiply both sides of the above by e^{-x} and integrate both sides from 0 to $+\infty$ with respect to x, which gives:

$$\begin{aligned}&\sum_{n=0}^{+\infty}\sum_{m=0}^{+\infty}\left\{\int\limits_{0}^{+\infty}e^{-x}L_n(x)L_m(x)\mathrm{d}x\right\}t^ns^m\\ &=\frac{1}{(1-t)(1-s)}\int\limits_{0}^{+\infty}e^{-x\{1+t/(1-t)+s/(1-s)\}}\mathrm{d}x\\ &=\frac{1}{(1-t)(1-s)}\left|\frac{e^{-x\{1+t/(1-t)+s/(1-s)\}}}{-\left\{1+\frac{t}{1-t}+\frac{s}{1-s}\right\}}\right|_{0}^{+\infty}\\ &=\frac{1}{(1-t)(1-s)}\cdot\frac{1}{1+\frac{t}{1-t}+\frac{s}{1-s}}=\frac{1}{1-st}\\ &=(1-st)^{-1}=1+st+(st)^2+(st)^3+\ldots=\sum_{n=0}^{+\infty}s^nt^n\end{aligned} \tag{7.1.1031}$$

Therefore:

$$\sum_{n=0}^{+\infty}\sum_{m=0}^{+\infty}\left\{\int_0^{+\infty} e^{-x}L_n(x)L_m(x)\mathrm{d}x\right\}t^n s^m=\sum_{n=0}^{+\infty}s^n t^n \tag{7.1.1032}$$

When $m\neq n$, equating coefficients of $t^n s^m$ on both sides of the above relation gives:

$$\int_0^{+\infty} e^{-x}L_n(x)L_m(x)\mathrm{d}x=0 \tag{7.1.1033}$$

When $m=n$, equating coefficients of $t^n s^n$ from both sides of the prior previous relation gives:

$$\int_0^{+\infty} e^{-x}\left[L_n(x)\right]^2\mathrm{d}x=1 \tag{7.1.1034}$$

Combining we get:

$$\int_0^{+\infty} e^{-x}L_n(x)L_m(x)\mathrm{d}x=\delta_{mn} \tag{7.1.1035}$$

where:

$$\delta_{mn}=\begin{cases}0 & \mathrm{m}\neq\mathrm{n}\\ 1 & \mathrm{m}=\mathrm{n}\end{cases} \tag{7.1.1036}$$

Remark

We have already proved in the section of Calculus at page 1400 that the Legendre polynomials (useful for our study of Quantum Chemistry) are orthogonal.

From what we have seen above we deduce that:

$$\langle f|g\rangle=\int_a^b f(x)\cdot\overline{g(x)}\mathrm{d}x \tag{7.1.1037}$$

is in fact generalized by:

$$\boxed{\langle f|g\rangle=\int_a^b f(x)\cdot\overline{g(x)}w(x)\mathrm{d}x} \tag{7.1.1038}$$

where $w(x)$ is the "weight function".

So the engineer, physicist, mathematician must always be careful when he see in a textbook a sentence of the type: «*these functions are orthogonal*». Indeed the author/redactor should instead read: «*these functions are orthogonal with a given weight*».

7.1.6.1 Cauchy-Schwarz inequality for integrals

Let us now provide a detailed proof of what we have mentioned earlier above: the "Cauchy-Schwarz inequality for integrals". As this will be an important result with quite important application in Statistics, Topology and Quantum Physics.

Theorem 7.5. Let f and g be real functions which are continuous on the closed interval $[a,b]$. Then:

$$\boxed{\left(\int_a^b f(t)\,g(t)\,\mathrm{d}t\right)^2 \leq \int_a^b (f(t))^2\,\mathrm{d}t \int_a^b (g(t))^2\,\mathrm{d}t} \tag{7.1.1039}$$

Proof 7.5.1. For the proof we use the following clever trick! We put $\forall x \in \mathbb{R}$:

$$(xf(t)+g(t))^2 \tag{7.1.1040}$$

Hence using relative sizes of definite integrals:

$$0 \leq \int_a^b (xf(t)+g(t))^2\,\mathrm{d}t \tag{7.1.1041}$$

Using the linear combination of integrals property:

$$0 \leq x^2 \int_a^b (f(t))^2\,\mathrm{d}t + 2x \int_a^b f(t)\,g(t)\;\mathrm{d}t + \int_a^b (g(t))^2\,\mathrm{d}t \tag{7.1.1042}$$

Let us put:

$$0 \leq Ax^2+2Bx+C \tag{7.1.1043}$$

where:

$$A=\int_a^b (f(t))^2\,\mathrm{d}t \qquad B=\int_a^b f(t)\,g(t)\,\mathrm{d}t \qquad C=\int_a^b (g(t))^2\,\mathrm{d}t \tag{7.1.1044}$$

The quadratic equation $Ax^2+2Bx+C$ is non-negative for all x. It follows (using the same reasoning as in Cauchy's Inequality) that the discriminant $(2B)^2-4AC$ of this polynomial must be non-positive.

Thus:

$$B^2 \leq AC \tag{7.1.1045}$$

and hence the result!

■ Q.E.D.

That can be interpreted as the functional version of the Cauchy-Schwarz equality for real vectors that we have proven in the section of Vector Calculus (see page 1826):

$$|\vec{x}\circ\vec{y}|\leq\|\vec{x}\|\|\vec{y}\| \tag{7.1.1046}$$

Remember that we have proved the complex version of the latter also (see page 2388):

$$|\langle\vec{x},\vec{y}\rangle|^2\leq\langle\vec{x},\vec{y}\rangle\langle\vec{y},\vec{y}\rangle \tag{7.1.1047}$$

That naturally leads us to build the hermitian Cauchy-Shwarz inequality for integral where $f,g\in\mathbb{C}$:

$$\boxed{\left|\int_a^b \langle f(t),g(t)\rangle|\mathrm{d}t\right|^2\leq\int_a^b |f(t)|^2\,\mathrm{d}t\int_a^b |g(t)|^2\,\mathrm{d}t} \tag{7.1.1048}$$

Also sometimes denoted:

$$\left|\int_a^b f(t)\bar{g}(t)\mathrm{d}t\right|^2\leq\int_a^b |f(t)|^2\,\mathrm{d}t\int_a^b |g(t)|^2\,\mathrm{d}t \tag{7.1.1049}$$

among many other notations with small variations...

~100 %

★★★★☆

47 votes, 71.49%

7.2. Complex Analysis

BEFORE STARTING this section on the study of differential and integral calculus in the generalized case of complex numbers, I should point out that I used many illustrations inspired by the PDF of E. Hairer (with his permission). This text also contains many sentences and developments taken, homogenized and simplified from the same PDF (at the risk to make some purist readers climbing to the curtains...) according to the notations and educational objectives of the rest of this textbook.

The subject of Complex Analysis is the study of functions $\mathbb{C} \mapsto \mathbb{C}$ and their differentiability (which is different from that in $\mathbb{R}^2$). The "holomorphic functions" (that is to say differentiable in a subset of $\mathbb{C}$) have as we will see it later surprising and elegant properties that can be reused in the situation of the special case of functions in $\mathbb{R}^2$ (remember that $\mathbb{C}$ is a generalization of $\mathbb{R}^2$) that have important applications in advanced physics (we will use the results of this section for our study of quantum physics and also for some applications of fluid mechanics and also for advanced models in financial options pricing).

Before we start let us first explain the interest of Complex Analysis in a simplified way!

We studied in the chapter Algebra a part of the Differential and Integral Calculus with some useful and important theorems in physics and engineering. However, staying in $\mathbb{R}$ or $\mathbb{R}^2$ the list of theorems runs out somehow and we end up finding much relevant tools in practice that allows to simplify the integration calculation that we can sometimes found in industrial applications. So, when we remember that $\mathbb{R} \subset \mathbb{C}$ (thus the set of complex numbers generalizes the set of real numbers) and that we can also build a correspondence $\mathbb{R}^2 \mapsto \mathbb{C}$ as we shall see, then new theorems appear with very interesting results that can be exploit for the integrals in $\mathbb{R}$ or $\mathbb{R}^2$!! It is because of this reason that the engineer needs to know Complex Analysis!

After studying this particular field of mathematics, it is common to say that the shortest path between two truths of the real domain often requires to pass trough the complex domain...

7.2.1 Linear Applications

A good introduction to complex analysis and its representation is to look at first (for educational purposes mainly) the special case of complex linear applications. Let us see this!

Let $U \subset \mathbb{C}$ be a set and $V \subset \mathbb{C}$ another set. A function that associates to each $z \in U$ an $w \in V$ such that:

$$w = f(z) = f(x + \mathrm{i}y) \in V \qquad (7.2.1)$$

is a "complex function":

$$f : U \mapsto V \qquad (7.2.2)$$

What is important is to remember (see section Numbers page 243) that we can identify:

$$z = x + \mathrm{i}y \in \mathbb{C} \mapsto (x,y) \in \mathbb{R}^2 \tag{7.2.3}$$

and:

$$\boxed{w = f(z) = u(x,y) + \mathrm{i}v(x,y) \mapsto (u,v) \in \mathbb{R}^2} \tag{7.2.4}$$

We have thus two functions of two real variables x, y:

$$u(x,y), v(x,y) \tag{7.2.5}$$

which are the coordinates of the point w.

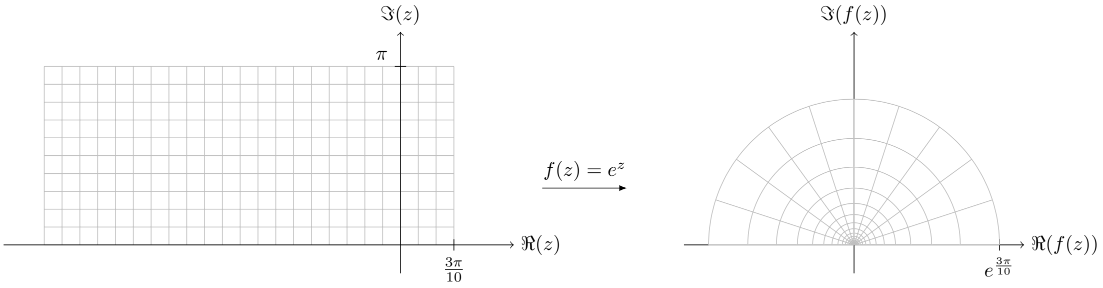


Figure 7.68: **Example of a complex function (conformal mapping)**

Definition (#7.2.288): An application is named "$\mathbb{C}$-linear" if for example a function of the type:

$$w = f(z) = c \cdot z \tag{7.2.6}$$

where c is a fixed complex number and z an arbitrary complex number, satisfies:

$$f(cz_1 + cz_2) = cf(z_1) + cf(z_2) \tag{7.2.7}$$

That is to say that $f(z)$ must me additive and homogeneous or just briefly when this two properties are satisfied we say that $f(z)$ is a "linear map".

We have seen and proved in the section on Numbers during our study of complex numbers, that the multiplication of two complex numbers could be equivalent to an orthogonal rotation followed by a scaling and that this same multiplication could be represented in matrix form! Or the transcription into a matrix form involves as we saw in the section on Linear Algebra automatically the property of linearity!

So the reader can easily check that a matrix of rotation/scaling is an example of an $\mathbb{C}$-linear application (on request we can detail) that we will now write:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = s \cdot \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \tag{7.2.8}$$

Which can be typically represented as follows (we can clearly see a rotation and a scaling which conserve the angles and proportions):

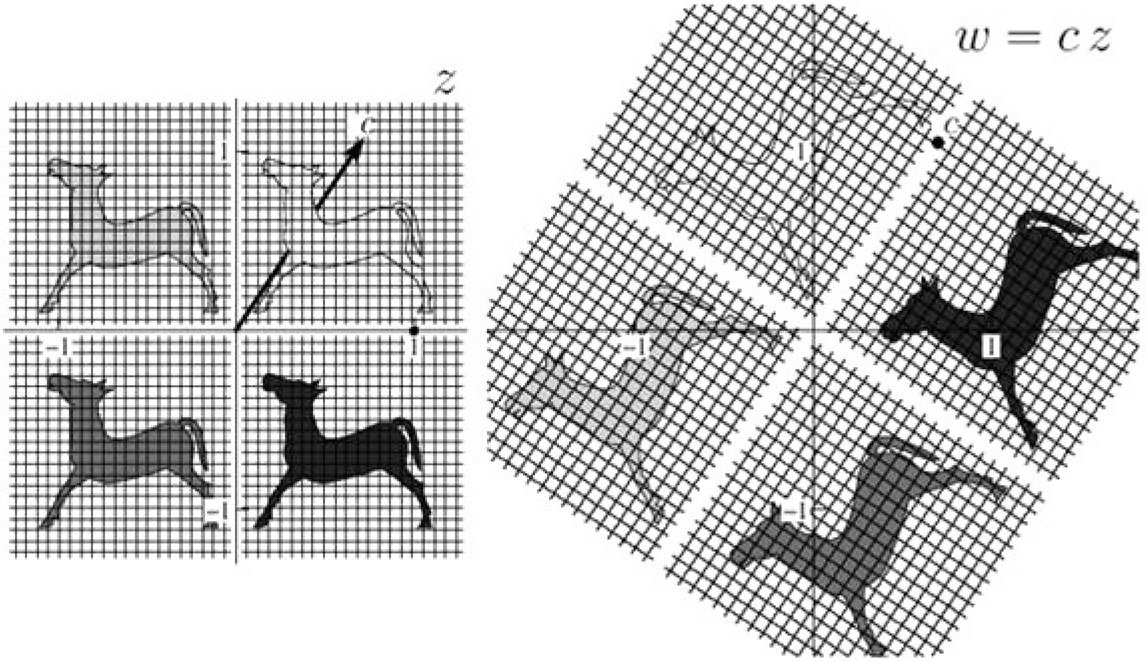


Figure 7.69: $\mathbb{C}$-linear application example

It is the fact that the proportions and the angles are kept that makes a complex function $\mathbb{C}$-linear. Otherwise, we would say that the function is $\mathbb{R}$-linear.

So a matrix equation is $\mathbb{C}$-linear if and only if it is of the form:

$$A = \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \tag{7.2.9}$$

Let us see some examples of quite remarkable $\mathbb{C}$ non-linear functions.

✍**Examples:**

E1. Consider the function:

$$f(z) = z^2 \tag{7.2.10}$$

In real coordinates this gives:

$$f(z) = u + \mathrm{i}v = (x + \mathrm{i}y)^2 = x^2 - y^2 + 2\mathrm{i}xy \Rightarrow \begin{cases} x_1 & = \dfrac{c_1 - b_1 x_2}{a_1} \\ x_2 & = \dfrac{c_2 - a_2 x_1}{b_2} \end{cases} \tag{7.2.11}$$

So let's look what this function do with the points of the complex plane which are coincident with the vertical lines of this same plane (which takes us to write $x = a$). Then we have:

$$u = a^2 - y^2 \qquad v = 2ay \tag{7.2.12}$$

and eliminating y, we find the equation of a parabola or rather a family of parabolas (for several values of b) which are open to the left of the pictured complex plane.

Here is a picture representation of the complex plane on which we have drawn a cat head:

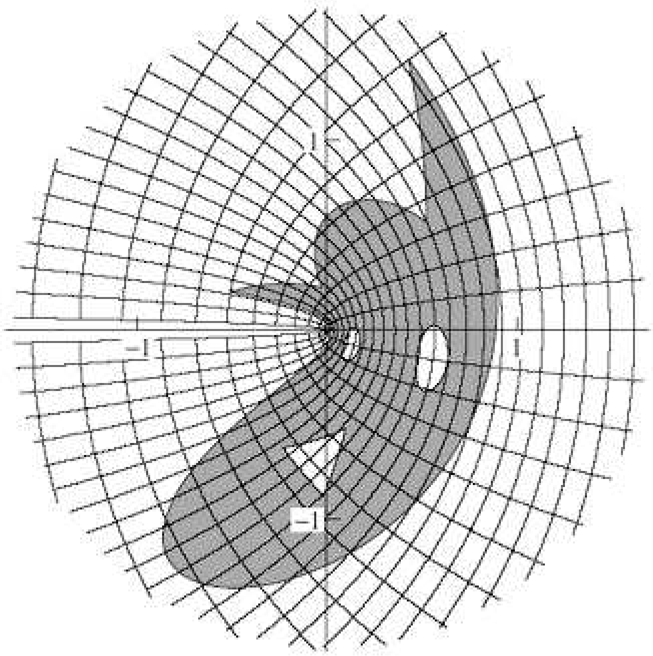


Figure 7.70: **Complex representation of the image of the example function**

and if we look at the corresponding pre-image complex plane then we have two heads of cats that appear:

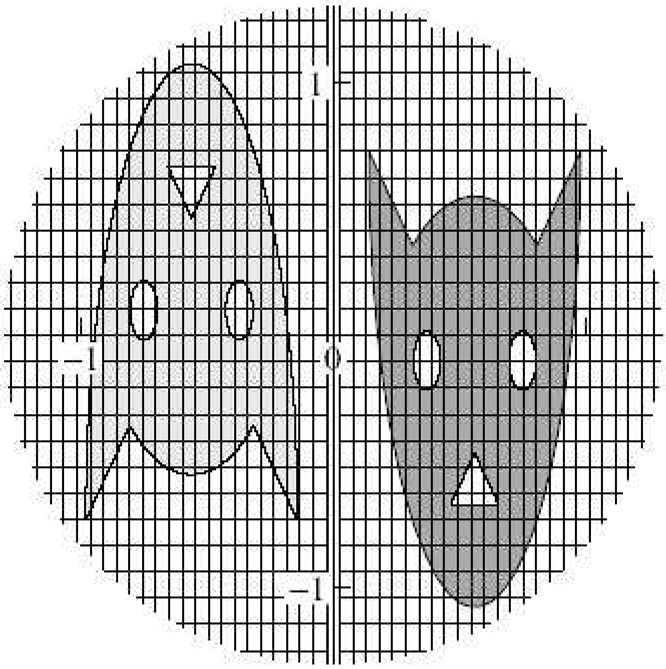


Figure 7.71: **Pre-image representation of the example function**

The appearance of these two heads of cats is that this function has 2 possible pre-images for each image point (so it is a surjective function - see section Set Theory page 406).

Here is a nice Maple™ 17.00 script by Carl Ebehart to check this (shame that this can not be done in an easier way in Maple™):

```
> complextools[gridimage] := proc(p)
local llhc, width, height, xres, yres, clrs, V, H,
i,j,k,l,pz,x,y,z,f,g,xtcs,ytcs,opts,margs;
llhc := [-1, -1];
width := 2;height := 2;
xres := .25;yres := .25;
xtcs := 1; ytcs := 1;
clrs := [red, black];
opts := NULL;
opts := op(select(type,[args],‘=‘));
margs:= remove(type, [args] ,‘=‘) ;
if nops(margs) >1 and margs[2] <> “ then llhc := margs[2] fi:
if nops(margs) >2 and margs[3] <> “ then width := margs[3] fi:
if nops(margs) >3 and margs[4] <> “ then height := margs[4] fi:
if nops(margs) >4 and margs[5] <> “ then xres := margs[5] fi:
if nops(margs) >5 and margs[6] <> “ then yres := margs[6] fi:
if nops(margs) >6 and margs[7] <> “ then xtcs := margs[7] fi:
if nops(margs) >7 and margs[8] <> “ then ytcs := margs[8] fi:      if
nops(margs) >8 and margs[9] <> “ then clrs := margs[9] fi:
z:= x + I*y;
pz := evalc(p(z));
f := unapply(evalc( Re(pz)),x,y); g := unapply(evalc(Im(pz)),x,y);
V:= plot( [ seq([seq(op([[f(llhc[1] + i*xres
,llhc[2]+(j-1)*yres/ytcs),g(llhc[1] + i*xres
,llhc[2]+(j-1)*yres/ytcs)], [f(llhc[1] + i*xres , llhc[2]
+j*yres/ytcs),g(llhc[1] + i*xres , llhc[2] +j*yres/ytcs)]]),
j=1..ytcs*height/yres)], i = 0 ..  width/xres) ],color=clrs[1]);
H := plot( [ seq([seq(op([[f(llhc[1]+(j-1)*xres/xtcs,llhc[2] +
i*yres), g(llhc[1]+(j-1)*xres/xtcs,llhc[2] + i*yres)], [f(llhc[1]
+j*xres/xtcs, llhc[2] + i*yres), g(llhc[1] +j*xres/xtcs, llhc[2]
+ i*yres)]]), j=1..xtcs*width/xres)], i = 0 ..  height/yres)
],color=clrs[2]);
plots[display]([V,H],scaling=constrained,opts); end:
with(complextools);

>plots[display]([seq(plots[display]([gridimage(z->z),
gridimage(z->z^2)]),i=10)],insequence=true);
```

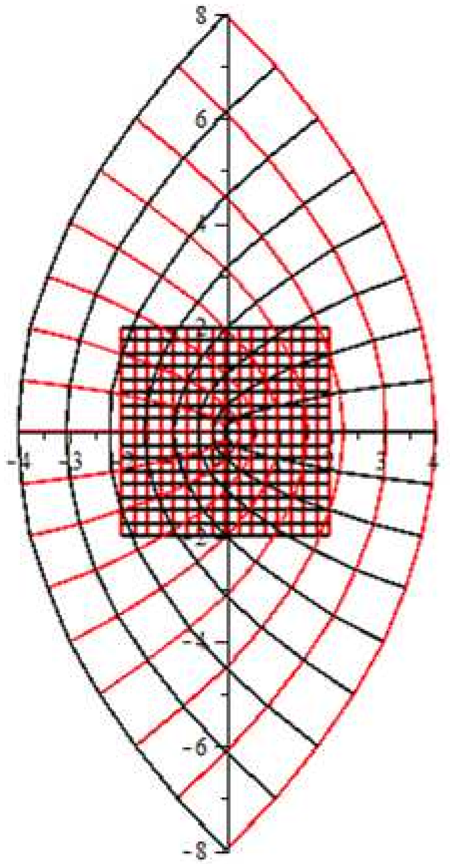


Figure 7.72: **Practical Maple™ 17.00 example of simple C-linear application**

E2. Another interesting feature is the "Cayley transformation" used in some areas of physics and defined as:

$$f(z)=\frac{z+1}{z-1} \tag{7.2.13}$$

having as domain definition: $\mathbb{C}/\{1\}$.

We notice that this is an involutive function since:

$$f(f(z))=f\left(\frac{z+1}{z-1}\right)=\frac{\dfrac{z+1}{z-1}+1}{\dfrac{z+1}{z-1}-1}=\frac{\dfrac{z+1+(z-1)}{z-1}}{\dfrac{z+1-(z-1)}{z-1}}=\frac{z+1+(z-1)}{z+1-(z-1)}=\frac{2z}{2}=z \tag{7.2.14}$$

and as we have proved in the section of Proof Theory (see page 181) that any involution function is both injective and surjective, then the Cayley transform is a bijective function.

This function transforms the imaginary axis $\mathrm{i}y$ in unit circle (and vice versa as it is involutive). Let us see that:

$$\begin{aligned}f(z)=f(\mathrm{i}y)=\frac{\mathrm{i}y+1}{\mathrm{i}y-1}=\frac{\mathrm{i}y+1}{\mathrm{i}y-1}\frac{-\mathrm{i}y-1}{-\mathrm{i}y-1}=\frac{y^2-\mathrm{i}y-\mathrm{i}y-1}{y^2-\mathrm{i}y+\mathrm{i}y+1}=\frac{y^2-2\mathrm{i}y-1}{1+y^2}\\=\frac{y^2-1}{1+y^2}-\mathrm{i}\frac{2y}{1+y^2}=u+\mathrm{i}v\end{aligned} \tag{7.2.15}$$

where:

$$u=\frac{y^2-1}{1+y^2} \qquad v=-\frac{2y}{1+y^2} \tag{7.2.16}$$

satisfies:

$$u^2+v^2=\frac{(y^2-1)^2+(-2y)^2}{(1+y^2)^2}=\frac{(y^4-2y^2+1)+(4y^2)}{1+2y^2+y^4}=1 \tag{7.2.17}$$

That is to say:

$$u^2+v^2=1 \tag{7.2.18}$$

This is the equation of a circle as proven in the section of Analytical Geometric.

E3. As another example of function, consider the "Joukovski transformation" defined by:

$$f(z)=\frac{1}{z}\left(z+\frac{1}{z}\right) \tag{7.2.19}$$

If the definition domain is built in polar coordinates look at how a circle or ellipse transforms with this function:

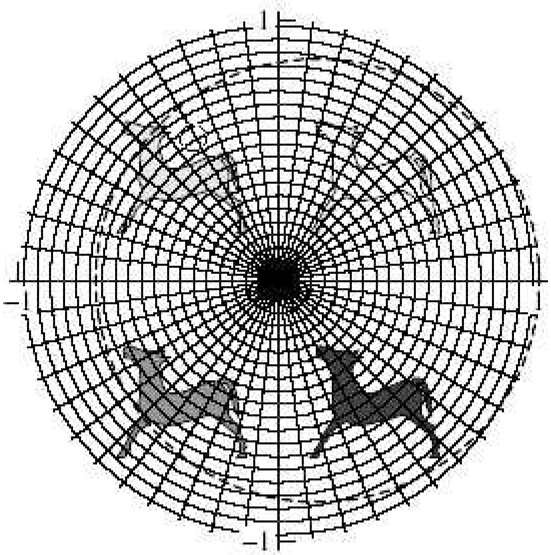


Figure 7.73: Transformation into polar coordinates of an ellipse with the example function

Then the image plane will be:

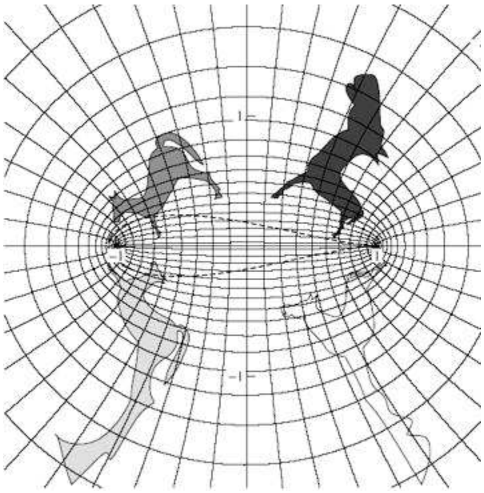


Figure 7.74: **Result of the Joukovski transformation in polar coordinates**

It thus transforms the circles respectively centered at 0 and the rays passing through 0 into a family of cofocal ellipses and hyperbolas. To prove this fact, we use the complex polar coordinates (Euler formula) seen in the section on Numbers (see page 251):

$$z=r(\cos(\varphi)+\mathrm{i}\sin(\phi)) \tag{7.2.20}$$

and:

$$\begin{aligned}z^{-1}&=\frac{1}{r(\cos(\varphi)+\mathrm{i}\sin(\phi))}=r^{-1}\frac{1}{\cos(\varphi)+\mathrm{i}\sin(\phi)}\frac{\cos(\varphi)-\mathrm{i}\sin(\phi)}{\cos(\varphi)-\mathrm{i}\sin(\phi)}\\&=r^{-1}\frac{\cos(\varphi)-\mathrm{i}\sin(\phi)}{\cos^2(\varphi)+\sin^2(\phi)}=r^{-1}(\cos(\varphi)-\mathrm{i}\sin(\phi))\end{aligned} \tag{7.2.21}$$

Then we have:

$$\begin{aligned}f(z)&=\frac{1}{2}\left(r(\cos(\varphi)+\mathrm{i}\sin(\phi))+\frac{1}{r}(\cos(\varphi)-\mathrm{i}\sin(\phi))\right)\\&=\left(\frac{r}{2}+\frac{1}{2r}\right)\cos(\varphi)+\mathrm{i}\left(\frac{r}{2}-\frac{1}{2r}\right)\sin(\phi)\end{aligned} \tag{7.2.22}$$

therefore:

$$u=\left(\frac{r}{2}+\frac{1}{2r}\right)\cos(\varphi)\qquad v=\left(\frac{r}{2}-\frac{1}{2r}\right)\sin(\phi) \tag{7.2.23}$$

and we immediately see that (see section Trigonometry page 2519):

$$\cos^2(\varphi)+\sin^2(\phi)=\frac{u^2}{\left(\dfrac{r}{2}+\dfrac{1}{2r}\right)^2}+\frac{v^2}{\left(\dfrac{r}{2}-\dfrac{1}{2r}\right)^2}=1 \tag{7.2.24}$$

has the form of the equation of an ellipse (see section Analytical Geometry page 2768) and we also have:

$$\begin{aligned}\frac{u^2}{\cos^2(\varphi)}-\frac{v^2}{\sin^2(\phi)}&=\left(\frac{r}{2}+\frac{1}{2r}\right)^2-\left(\frac{r}{2}-\frac{1}{2r}\right)^2\\&=\left(\frac{r^2}{4}+2\frac{r}{2}\frac{1}{2r}+\frac{1}{4r^2}\right)-\left(\frac{r^2}{4}-2\frac{r}{2}\frac{1}{2r}+\frac{1}{4r^2}\right)\\&=2\frac{r}{2}\frac{1}{2r}+2\frac{r}{2}\frac{1}{2r}=\frac{1}{2}+\frac{1}{2}=1\end{aligned}\tag{7.2.25}$$

which is the equation of a hyperbola (see section Analytical Geometry page 2768).

This function is useful in case we cleverly place a circle through the point $z=1$ (as in the case of the first figure) the plan represented in polar coordinates with a dotted line might looks like an airplane wing. This allowed a time ago in aerodynamics (but the technique is obsolete today) to transpose the study of a vector field of an airplane wing profile to the study of a circle profile and to do after the Joukovski transformation.

Indeed, let us see a part of this still with Maple™ 4.00:

```
> assume(x,real,y,real);
> z:=x+I*y;
> F:=1/2*(z+1/z);
> u:=Re(F);
> u:=evalc(u);
> v:=Im(F); > v:=evalc(v);
> with(plots):with(plottools):
> p1:=disk([0,0],1,color=black):
> p2:=implicitplot(seq(v=b8,b=-10..10),x=-4..4,y=-2..2,color=black):
> display([p2,p1],scaling=constrained);
```

We thus get:

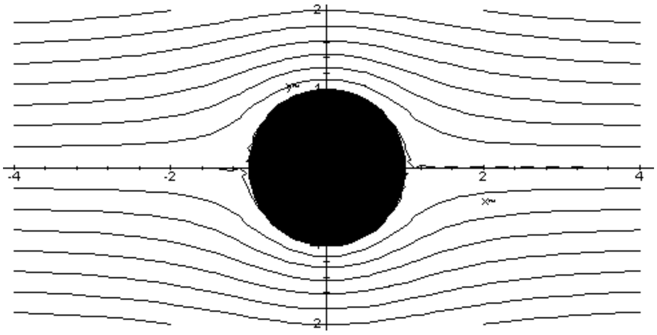


Figure 7.75: **Important application example of the Joukovski function**

Let us see a last example that shows an electric dipole with its electric field and potential lines (see section Electrostatics page 3980) can bee seen as the emergence of the $\mathbb{C}$-linear function $1/z$:

✍**Examples:**

E4. Always with Maple™ 4.00b we write:

```
>assume(x,real,y,real);
> z:=x+I*y;
> F:=1/z;
> u:=Re(F);u:=evalc(u);
> v:=Im(F);v:=evalc(v);
> with(plots):
> p1:=implicitplot(seq(u=a,a=-5..5),x=-1..1,y=-1..1,numpoints=1000):
> p2:=implicitplot(seq(v=b,b=-5..5),x=-1..1,y=-1..1,numpoints=1000,
color=green):
> display([p1,p2],scaling=constrained);
```

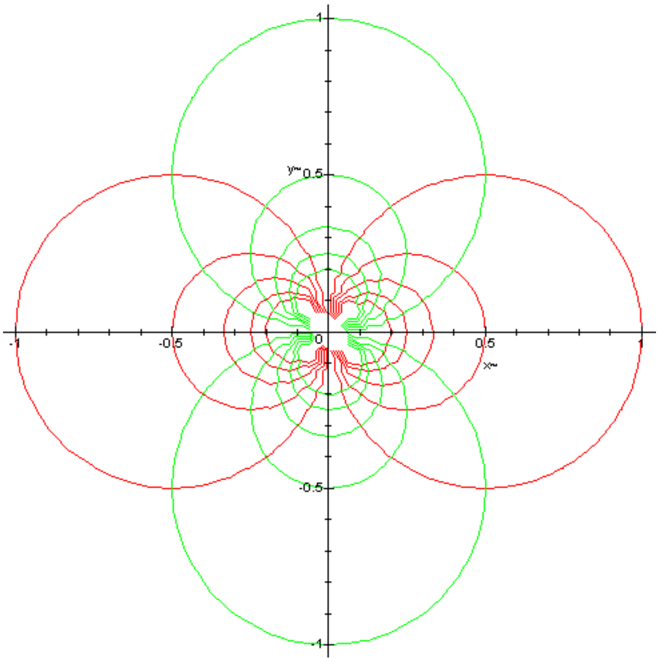


Figure 7.76: **Another important application of a complex function**

7.2.2 Holomorphic Functions

The definition of the derivative with respect to a complex variable is naturally formally identical to the derivative with respect to a real variable.

We then have, if the function $f(z)$ is differentiable in z_0:

$$\frac{\mathrm{d}f_h}{\mathrm{d}z}=\lim_{z\to z_0}\frac{f(z)-f(z_0)}{z-z_0}=\lim_{h\to 0}\frac{f(z_0+h)-f(z_0)}{(z_0+h)-z_0}=\lim_{h\to 0}\frac{f(z_0+h)-f(z_0)}{h}=f'(z) \quad (7.2.26)$$

and we say (abusively in this textbook) that function is "holomorphic" (while in $\mathbb{R}$ we say "differentiable") or "analytical" in its domain or in a subset of it if it is differentiable at any point.

In other words a holomorphic functions is a complex-valued function of one or more complex variables that is complex differentiable in a neighbourhood of every point in its domain. The existence of a complex derivative in a neighbourhood is a very strong condition, for it implies that any holomorphic function is actually infinitely differentiable and equal to its own Taylor series.

Remarks

R1. A complex function is derived like a real function, we just have to put z as being x... at the condition of what we will see in what follows is respected!

R2. In fact if the function is holomorphic in a subset of the complex plane, we will see a little further below in our study of the convergence of power series that this is than always an open subset.

Equivalently, we say that the function f is $\mathbb{C}$-differentiable if the following limit exists in $\mathbb{C}$:

$$\lim_{h\rightarrow 0}\frac{f(z_0+h)-f(z_0)}{h}=f'(z_0) \tag{7.2.27}$$

Let us now present and prove a central theorem for complex analysis named "Cauchy-Riemann theorem"!

If the function:

$$f(z)=f(x,y)=f(x+\mathrm{i}y)=u(x,y)+\mathrm{i}v(x,y) \tag{7.2.28}$$

is $\mathbb{C}$-differentiable on $z_0=x_0+\mathrm{i}y_0$, then we have:

$$\boxed{\frac{\partial u}{\partial x}=\frac{\partial v}{\partial y} \qquad \frac{\partial v}{\partial x}=-\frac{\partial u}{\partial y}} \tag{7.2.29}$$

which is somewhat the equivalent to the Schwarz theorem (limited to $\mathbb{R}$) proved in the section of Differential and Integral Calculus (see page 1469). The above two relations are named "Cauchy conditions". So these are the two conditions that must verify a complex function to be differentiable on z_0. Thus, it is possible to use these relations to examine the points where the function is not analytic.

Theorem 7.6. If these conditions are satisfied (what we will prove right below!), then we deduce that u and v must both be harmonic functions of x and y.

Proof 7.6.1. As:

$$\lim_{h\rightarrow 0}\frac{f(z_0+h)-f(z_0)}{h}=f'(z_0) \tag{7.2.30}$$

by choosing:

$$z=z_0+x \tag{7.2.31}$$

with $x \in \mathbb{R}$, we get:

$$f'(z_0)=\lim_{z\rightarrow z_0}\frac{f(x_0+x,y_0)-f(x_0,y_0)}{z_0+x-z_0}=\lim_{z\rightarrow z_0}\frac{f(x_0+x,y_0)-f(x_0,y_0)}{x} \tag{7.2.32}$$

and as x approaches a small value $\mathrm{d}x$, we have (see section Differential and Integral Calculus page 1450):

$$\begin{aligned}f'(z_0)&=\lim_{\mathrm{d}x\rightarrow 0}\frac{f(x_0+\mathrm{d}x,y_0)-f(x_0,y_0)}{\mathrm{d}x}=\frac{\partial f(z_0)}{\partial x}=\frac{\partial(u(x_0,y_0)+\mathrm{i}v(x_0,y_0))}{\partial x}\\&=\frac{\partial}{\partial x}u(x_0,y_0)+\mathrm{i}\frac{\partial}{\partial x}v(x_0,y_0)\end{aligned} \tag{7.2.33}$$

by choosing:

$$z=z_0+\mathrm{i}y \tag{7.2.34}$$

with $y \in \mathrm{R}$, we get:

$$f'(z_0)=\lim_{z\rightarrow z_0}\frac{f(x_0,y_0+\mathrm{i}y)-f(x_0,y_0)}{z_0+\mathrm{i}y-z_0}=\lim_{z\rightarrow z_0}\frac{f(x_0,y_0+\mathrm{i}y)-f(x_0,y_0)}{\mathrm{i}y} \tag{7.2.35}$$

and when y tends to a small value we have (see section Differential and Integral Calculus page 1450):

$$\begin{aligned}f'(z_0)&=\lim_{\mathrm{d}y\rightarrow 0}\frac{f(x_0,y_0+\mathrm{i}\mathrm{d}y)-f(x_0,y_0)}{\mathrm{i}\mathrm{d}y}=\frac{1}{\mathrm{i}}\frac{\partial f(z_0)}{\partial y}=-\frac{\partial f(z_0)}{\partial y}\\&=-\mathrm{i}\frac{\partial(u(x_0,y_0)+\mathrm{i}v(x_0,y_0))}{\partial y}=-\mathrm{i}\left(\frac{\partial u(x_0,y_0)}{\partial y}+\mathrm{i}\frac{\partial v(x_0,y_0)}{\partial y}\right)\\&=\frac{\partial}{\partial y}v(x_0,y_0)-\mathrm{i}\frac{\partial}{\partial y}u(x_0,y_0)\end{aligned} \tag{7.2.36}$$

So now we have:

$$\begin{aligned}\left(\frac{\partial f}{\partial x}\right)_y&=\frac{\partial f}{\partial y}=\frac{\partial}{\partial x}u(x_0,y_0)+\mathrm{i}\frac{\partial}{\partial x}v(x_0,y_0)\\\left(\frac{\partial f}{\partial y}\right)_x&=\frac{\partial f}{\partial x}=\frac{\partial}{\partial y}v(x_0,y_0)-\mathrm{i}\frac{\partial}{\partial y}u(x_0,y_0)\end{aligned} \tag{7.2.37}$$

But remember we proved in the section of Integral and Differential Calculus the following theorem (see page 1469):

$$\frac{\partial^2 f}{\partial x\partial y}=\frac{\partial^2 f}{\partial y\partial x} \tag{7.2.38}$$

Hence:

$$\begin{aligned}\frac{\partial^2 f}{\partial y\partial x}&=\frac{\partial}{\partial y}\left(\frac{\partial}{\partial x}u(x_0,y_0)+\mathrm{i}\frac{\partial}{\partial x}v(x_0,y_0)\right)\\\frac{\partial^2 f}{\partial x\partial y}&=\frac{\partial}{\partial x}\left(\frac{\partial}{\partial y}v(x_0,y_0)-\mathrm{i}\frac{\partial}{\partial y}u(x_0,y_0)\right)\end{aligned} \tag{7.2.39}$$

Therefore using directly Schwartz theorem:

$$\frac{\partial}{\partial y}\left(\frac{\partial}{\partial x}u(x_0,y_0)+\mathrm{i}\frac{\partial}{\partial x}v(x_0,y_0)\right)=\frac{\partial}{\partial x}\left(\frac{\partial}{\partial y}v(x_0,y_0)-\mathrm{i}\frac{\partial}{\partial y}u(x_0,y_0)\right) \qquad (7.2.40)$$

Which can be written:

$$\frac{\partial}{\partial y}\left(\frac{\partial}{\partial x}u(x_0,y_0)+\mathrm{i}\frac{\partial}{\partial x}v(x_0,y_0)\right)-\frac{\partial}{\partial x}\left(\frac{\partial}{\partial y}v(x_0,y_0)-\mathrm{i}\frac{\partial}{\partial y}u(x_0,y_0)\right)=0 \qquad (7.2.41)$$

A trivial solution is obviously to have:

$$\begin{aligned}\frac{\partial}{\partial x}u(x_0,y_0)+\mathrm{i}\frac{\partial}{\partial x}v(x_0,y_0)=0)\\ \frac{\partial}{\partial y}v(x_0,y_0)-\mathrm{i}\frac{\partial}{\partial y}u(x_0,y_0)=0\end{aligned} \qquad (7.2.42)$$

Therefore we are authorized to write:

$$\frac{\partial}{\partial x}u(x_0,y_0)+\mathrm{i}\frac{\partial}{\partial x}v(x_0,y_0)=\frac{\partial}{\partial y}v(x_0,y_0)-\mathrm{i}\frac{\partial}{\partial y}u(x_0,y_0)=0 \qquad (7.2.43)$$

By identifying real and imaginary parts, we finish the proof!

■ Q.E.D.

So for f to be differentiable in the complex domain $\mathbb{C}$ (holomorphic) at a point, it is sufficient that it be differentiable as a function of two real variables ($\mathbb{R}^2$-differentiable on (x_0,y_0)) and that its partial first derivatives at this point satisfy the Cauchy-Riemann equations.

But, for it to be $\mathbb{C}$-differentiable, Cauchy-Riemann's equations must valid at all points of the complex plane (we sometimes speaks about "complete functions") and not only in a subdomain thereof! Otherwise, it contains therefore "singularities" and we then speak of "meromorphic function" (which is therefore a holomorphic function excepted on singularities points).

The Gamma function studied in the section of Differential and Integral Calculus (see page 1574) and plotted with Maple™ 4.00b (code provided!) is such a well-known meromorphic function.

Remark

Geometrically, we will prove later that a holomorphic function has a possible interpretation in the sense that it is a conformal transformation (angles conservation).

Notice therefore that if $f(z)$ is $\mathbb{C}$-differentiable it can be developed as Taylor series (see section Sequences and Series page 1696):

$$f(z)=\sum_{i=0}^{+\infty}\frac{1}{n!}\frac{\mathrm{d}^n f(z_0)}{\mathrm{d}z^n}(z-z_0)^n \qquad (7.2.44)$$

Note an important thing too. If we rewrite:

$$\frac{\partial u}{\partial x}=\frac{\partial v}{\partial y} \qquad \frac{\partial v}{\partial x}=-\frac{\partial u}{\partial y} \qquad (7.2.45)$$

as following:

$$\begin{aligned}\frac{\partial u}{\partial x}-\frac{\partial v}{\partial y}=0\\ \frac{\partial v}{\partial x}+\frac{\partial u}{\partial y}=0\end{aligned} \tag{7.2.46}$$

Then we say that f is "irrotational" (see section Vector Calculus page 1889) since the first relation can be seen as:

$$\frac{\partial u}{\partial x}-\frac{\partial v}{\partial y}=\text{rot}(\vec{f})_z=0 \tag{7.2.47}$$

which is an important analogy! Finally, the second equation:

$$\frac{\partial v}{\partial x}+\frac{\partial u}{\partial y}=0 \tag{7.2.48}$$

also let us say by analogy (but it stops at a simple analogy!) that the bivariate function f is non-divergent (see section Vector Calculus page 1879) what is good mnemonic way to remember this equation.

Let's also show something else in evidence. If we take the two Cauchy-Riemann equations:

$$\begin{aligned}\frac{\partial u}{\partial x}=\frac{\partial v}{\partial y}\\ \frac{\partial u}{\partial y}=-\frac{\partial v}{\partial x}\end{aligned} \tag{7.2.49}$$

and that we derivate them once again, we get:

$$\begin{aligned}\frac{\partial^2 u}{\partial x^2}=\frac{\partial^2 v}{\partial x\partial y}\\ \frac{\partial^2 u}{\partial y^2}=-\frac{\partial^2 v}{\partial y\partial x}\end{aligned} \tag{7.2.50}$$

and that we sum these two relations, we get then:

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}=0 \tag{7.2.51}$$

It is the same with v. Then we have:

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}=0 \qquad \frac{\partial^2 v}{\partial x^2}+\frac{\partial^2 v}{\partial y^2}=0 \tag{7.2.52}$$

And we know very well this form of equations (Maxwell-Poisson equation in the section of Electrodynamics and Newton-Poisson equation in the section of Astronomy...). This is a wave equation also named "Laplace equation" (nothing to do with that of the same name seen in our study of the hydrostatic!) and given by the scalar Laplacian (see section Vector Calculus page 1899):

$$\boxed{\Delta\cdot u=\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}=0 \qquad \Delta\cdot v=\frac{\partial^2 v}{\partial x^2}+\frac{\partial^2 v}{\partial y^2}=0} \tag{7.2.53}$$

Then it is traditional to say that u is harmonic and of course we can get the same result with v! Well obviously ... we knew it, since we have already studied in the section Numbers that the real and imaginary parts of a complex number could be put in trigonometric form.

Thanks to this discovery, Riemann opened the application of holomorphic functions in many problems of physics, since these equations are satisfied by the gravitational potential (Newton-Poisson equation in the section of Astronomy page 5267) by electric and magnetic fields (Maxwell-Poisson equation in the section of Electrodynamics page 4099) by heat balance (no examples yet in this textbook) and by movements without rotational of certain fluids (no examples either in this textbook yet).

✍Example:

The potential of a dipole can be described by the following holomorphic function:

$$f(z)=\frac{1}{z+1}+\frac{1}{z-1} \tag{7.2.54}$$

The figure below:

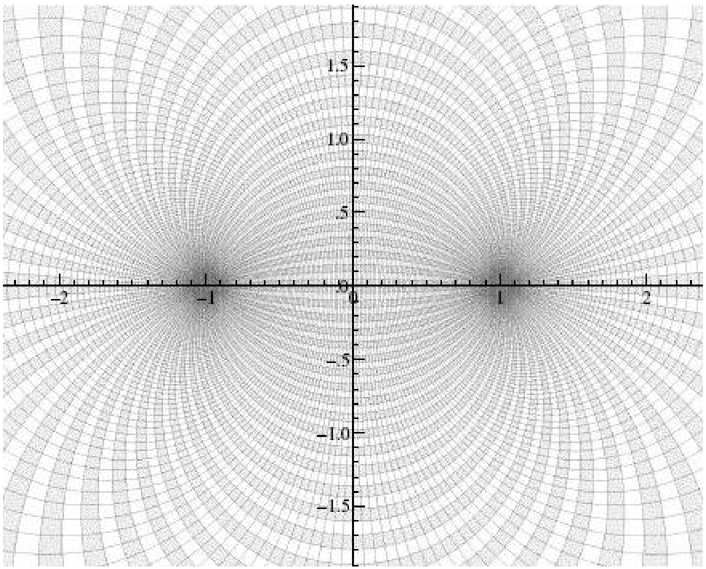


Figure 7.77: **Plane representation of a well known holomorphic function...**

shows level-curves (iso-curves) of the given harmonic functions $u(x,y)$ and $v(x,y)$ as real and complex parts of the function $f(z)$ of this example.

7.2.2.1 Orthogonality of real and imaginary iso-curves

We will now prove a friendly property that have the functions that satisfy the Cauchy conditions (i.e. that are analytic functions!). Indeed, remember that we have already seen above the function:

$$f(z)=z^2 \tag{7.2.55}$$

which gave the following diagram:

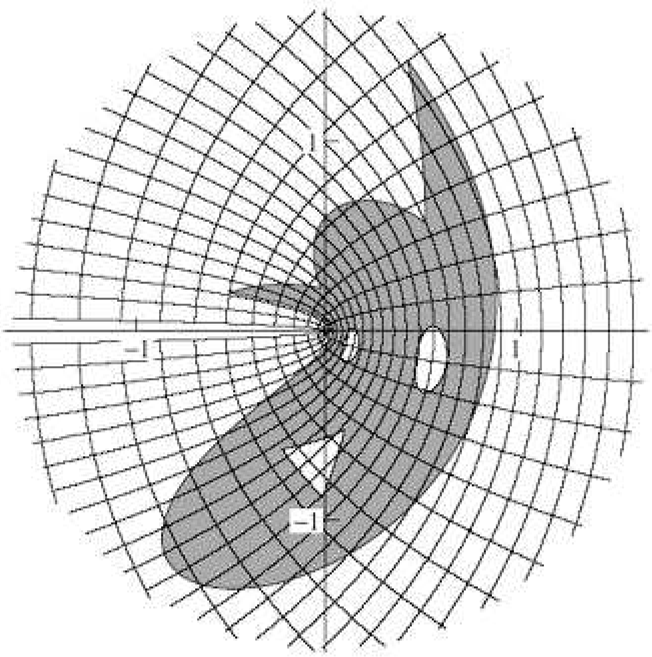


Figure 7.78: **Reminder of plane representation of a complex function seen earlier**

Theorem 7.7. The functions satisfying the Cauchy conditions have the simple following geometrical property: the lines whose real part of the function is constant $\mathcal{R}(f(z))=c^{te}$ and the lines whose imaginary part is constant $\mathcal{I}(f(z))=c^{te}$ are orthogonal to each other (think to the trigonometric form of complex numbers it helps to better visualize!).

In other words, the analytical complex functions are transformation functions of an area of the plane into a new plane where the angles are preserved. Then we say that the function is a "complete transformation".

Proof 7.7.1. For the proof remember that we have proved in the section of Vector Calculus that gradient of a function f of $\mathbb{R}^2$ is given by:

$$\vec{\nabla}(f)=\left(\frac{\partial f}{\partial x},\frac{\partial f}{\partial y}\right) \tag{7.2.56}$$

and as part of our study of isolines in the section of Differential Geometry that the tangent vector to isolines of the function f will always be parallel to the vector of the plane:

$$\vec{t}=\left(\frac{\partial f}{\partial y},-\frac{\partial f}{\partial x}\right) \tag{7.2.57}$$

and that the latter two vectors are perpendicular, such that:

$$\vec{\nabla}(f)\circ\vec{t}=0 \tag{7.2.58}$$

Now assimilate the tangent (parallel) vector $\vec{t}_u$ to the real isolines:

$$\Re(f(z))=u(x,y)=c^{te} \tag{7.2.59}$$

with:

$$\vec{t}_u=\left(\frac{\partial u}{\partial y},-\frac{\partial u}{\partial x}\right) \tag{7.2.60}$$

and the normal vector to the imaginary isolines:

$$\Im(f(z))=v(x,y)=c^{te} \tag{7.2.61}$$

with the gradient v of components:

$$\vec{\nabla}(v)=\left(\frac{\partial v}{\partial x},\frac{\partial v}{\partial y}\right) \tag{7.2.62}$$

Using the Cauchy conditions proved above, we have for this last relation:

$$\vec{\nabla}(v)=\left(\frac{\partial v}{\partial x},\frac{\partial v}{\partial y}\right)=\left(-\frac{\partial u}{\partial y},\frac{\partial u}{\partial x}\right) \tag{7.2.63}$$

By comparing:

$$\vec{t}_u=\left(\frac{\partial u}{\partial y},-\frac{\partial u}{\partial x}\right)\quad\text{and}\quad\vec{\nabla}(v)=\left(-\frac{\partial u}{\partial y},\frac{\partial u}{\partial x}\right) \tag{7.2.64}$$

we therefore see that $\vec{t}_u$ and $\vec{\nabla}(v)$ are parallel (collinear). And since $\vec{t}_u$ is colinear the real isolines and that $\vec{\nabla}(v)$ is perpendicular to the imaginary isolines we finished our proof.

■ Q.E.D.

The reader may take as an example the function:

$$f(z)=z^2 \tag{7.2.65}$$

mathematically and schematically detailed earlier above! But to change a little bit, consider an example that will accompany us throughout the rest of this section and that is the following holomorphic function:

$$f(z)=\frac{1}{1+z^2} \tag{7.2.66}$$

That gives us with Maple™ 4.00b:

```
 >assume(x,real,y,real);
> z:=1/(1+(x+I*y)^2);
> F:=1/z;
> u:=Re(F);
> u:=evalc(u);
> v:=Im(F);
> v:=evalc(v);
> with(plots):
> p1:=implicitplot(seq(u=a,a=-5..5),x=-5..5,y=-5..5,numpoints=1000):
> p2:=implicitplot(seq(v=b,b=-5..5),x=-5..5,y=-5..5,numpoints=1000,color=green):
> display([p1,p2]);
```

which gives:

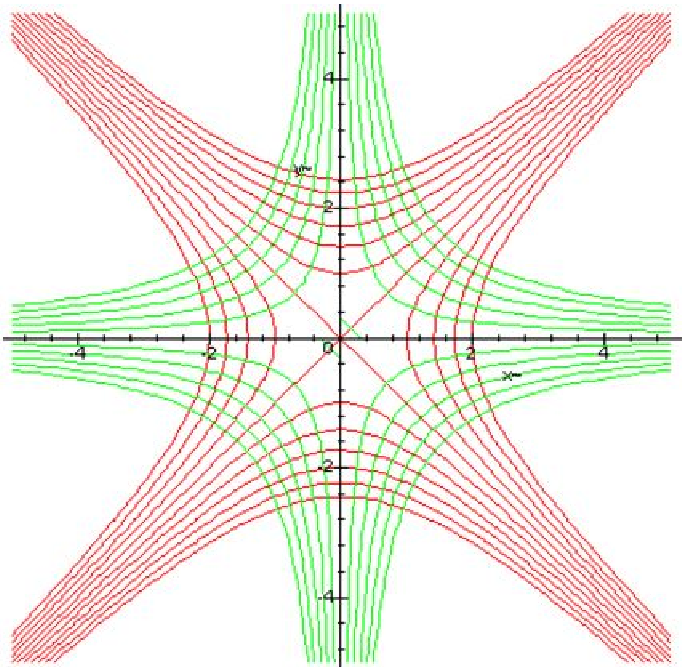


Figure 7.79: **Representation of an important holomorphic function with its isolines**

7.2.3 Complex Logarithm

We need for all functions built into $\mathbb{R}$ found their equivalent in $\mathbb{C}$ while knowing that if we reduce the case of $\mathbb{R}$ to $\mathbb{C}$ we must get back on our feet!

To do this, let us start with the most classical and academic function which is the logarithm and also the only one function for which we will need the complex version in other sections of this textbook. As always we will focus only on the properties that we will need later for practical applications and nothing more!

In the same way that we built the logarithm as being by definition by the inverse function of the natural exponential e^x in the section of Functional Analysis, we first start from:

$$f(z)=e^z \tag{7.2.67}$$

where z is a complex number and we will define the complex logarithm that must be reduced to the natural logarithm if z has no imaginary part!

So by definition the complex logarithm will be:

$$\boxed{w=e^z \Leftrightarrow z=\text{Log}(w)} \tag{7.2.68}$$

and in this entire book, the complex logarithm will be differentiated by the real logarithm by a capital L for the first letter!

Let us write z and w in the Euler form as viewed in the section Numbers:

$$\begin{aligned} w&=x+\mathrm{i}y=r(\cos(\theta)+\mathrm{i}\sin(\theta))=|w|e^{\mathrm{i}\text{arg}(w)}=|w|(\cos(\text{arg}(w))+\sin(\text{arg}(w)))\\ z&=u+\mathrm{i}v \end{aligned} \tag{7.2.69}$$

Then we have:

$$w=e^z \Rightarrow r(\cos(\theta)+\mathrm{i}\sin(\theta))=e^{u+\mathrm{i}v}=e^u e^{\mathrm{i}v}=e^u \tag{7.2.70}$$

By correspondence, we find immediately:

$$r=|w|=e^u \Leftrightarrow u=\ln(r) \quad \text{and} \quad v=\text{arg}(w)=\theta+2\pi k \tag{7.2.71}$$

with $k\in\mathbb{Z}$. Therefore we get:

$$z=u+\mathrm{i}v=\ln(r)+\mathrm{i}\text{arg}(w)=\ln(r)+\mathrm{i}(\theta+2\pi k)=\text{Log}(w) \tag{7.2.72}$$

Therefore:

$$\boxed{z=\text{Log}(w)=\ln(r)+\mathrm{i}(\theta+2\pi k)} \tag{7.2.73}$$

or more explicitly:

$$\boxed{\text{Log}(w)=\ln(|w|)+\mathrm{i}\text{arg}(w)} \tag{7.2.74}$$

So if w has no imaginary part, we fall back on our feet since arg(w) becomes zero.

A big difference is highlighted between the logarithm of the complex and real numbers: the complex numbers logarithms can take several values because of the argument!!

We can also check that:

$$e^{\text{Log}(w)}=e^{\ln|w|+\mathrm{i}\,\text{arg}(w)}=e^{\ln|w|}e^{\mathrm{i}\,\text{arg}(w)}=|w|e^{\mathrm{i}\,\text{arg}(w)}=w \tag{7.2.75}$$

For a function to have an inverse, it must map distinct values to distinct values, i.e., be injective. But the complex exponential function is not injective, because $e^{z+2\pi\mathrm{i}}=e^z$ for any z, since adding $\mathrm{i}\theta$ to w has the effect of rotating e^z counter-clockwise of θ radians. So all the points of the form $z+\mathrm{i}k\theta$ are all mapped to the same number by the exponential function. So the exponential function does not have an inverse function in the standard sense.

There are two solutions to this problem:

1. One is to restrict the domain of the exponential function to a region that does not contain any two numbers differing by an integer multiple of $2\pi\mathrm{i}$: this leads naturally to the definition of "branches" of Log(w), which are certain functions that single out one logarithm of each number in their domains.

2. Another way to resolve the indeterminacy is to view the logarithm as a function whose domain is not a region in the complex plane, but a "Riemann surface" that covers the punctured complex plane in an infinite-to-1 way.

Branches have the advantage that they can be evaluated at complex numbers. On the other hand, the function on the Riemann surface is elegant in that it packages together all branches of $\text{Log}(w)$ and does not require an arbitrary choice as part of its definition.

We can see this with Maple™ 4.00b easily:

```
>plot3d([r*cos(f),r*sin(f),f],r=0..1,f=-2*Pi..2*Pi,axes=boxed,style=patch,
shading=ZHUE);
```

which gives:

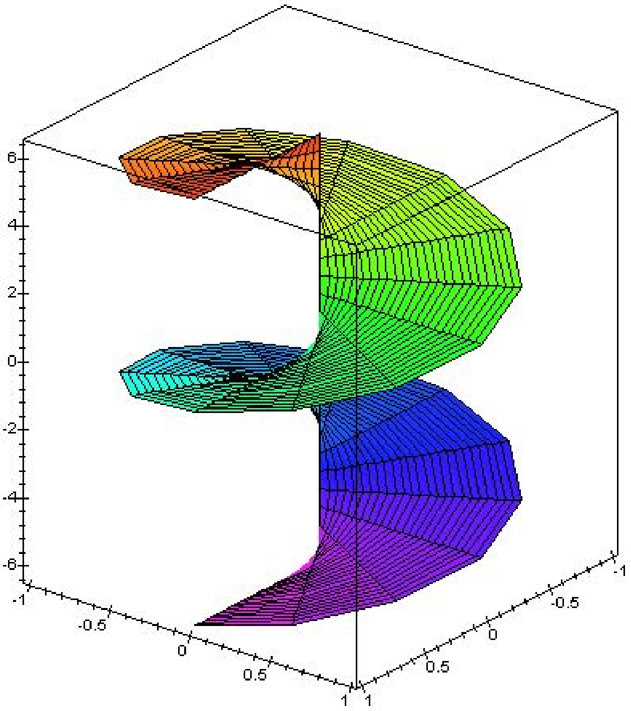


Figure 7.80: **Complex Logarithm plot in Maple™ 4.00b**

Or in high resolution using TikZ:

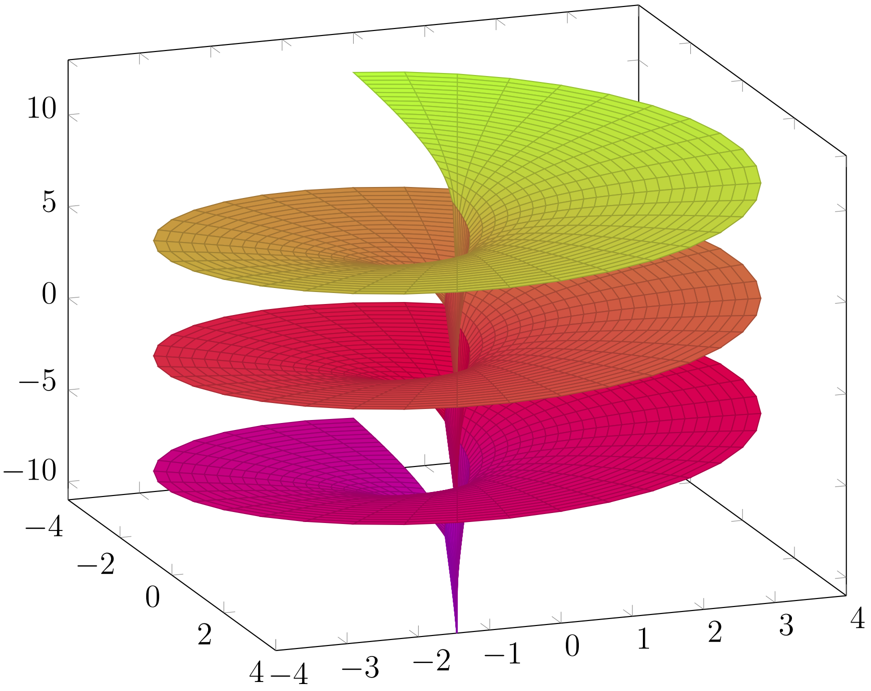


Figure 7.81: **Complex Logarithm with TikZ**

For this reason, one cannot always apply Log to both sides of an identity $e^{z_1}=e^{z_2}$ to deduce $z_1=z_2$. Also, the identity $\text{Log}(z_1z_2)=\text{Log}(z_1)+\text{Log}(z_2)$ can fail: the two sides can differ by an integer multiple of $2\pi\text{i}$.

For each non-zero complex number $w=x+\text{i}y$, the principal value $\text{Log}(w)$ is the logarithm whose imaginary part lies in the interval $[-\pi,+\pi]$. The expression $\text{Log}(0)$ is left undefined since there is no complex number z satisfying $e^z=0$.

Then the principal value of the complex logarithm can be defined by (see section Trigonometry page 2503):

$$\boxed{\text{Log}(w)=\ln(|w|)+\text{i}\arg(w)=\ln(\sqrt{x^2+y^2})+\text{i}\arctan2(y,x)} \tag{7.2.76}$$

We see also obviously that the function $\text{Log}(w)$ is discontinuous at each negative real number (we can see it on the figure above), but continuous everywhere else in $\mathbb{C}^*$.

Riemann surfaces can be thought of as deformed versions of the complex plane: locally near every point they look like patches of the complex plane, but the global topology can be quite different. For example, they can look like a sphere or a torus or a couple of sheets glued together.

The main point of Riemann surfaces is that holomorphic functions may be defined between them. Riemann surfaces are nowadays considered the natural setting for studying the global behaviour of these functions, especially multi-valued functions such as the square root and other algebraic functions, or the logarithm.

Basically, a Riemann surface is simply just a surface, as far as the shape is concerned. Any normal surface you can think of (for example, plane, sphere, torus, ...) are all Riemann surfaces. We say it's a "Riemann surface", is due to the context, is that we define the surface using complex functions, and for use in studying complex functions.

7.2.4 Complex Integral Calculus

We have seen just above how to check if a complex function $f(z)$ was differentiable (it must at least respect the Cauchy-Riemann equations) at any point.

Now let us see the opposite case... the integration that is absolutely fascinating in complex plane!

We have obviously taking again the notations of the section of Differential and Integral Calculus:

$$F(z)=\int f(z)\mathrm{d}z \tag{7.2.77}$$

either in explicit form:

$$\begin{aligned}F(z)&=\int f(x+\mathrm{i}y)\mathrm{d}(x+\mathrm{i}y)=\int(u+\mathrm{i}v)(\mathrm{d}x+\mathrm{i}\mathrm{d}y)\\&=\left[\int(u\mathrm{d}x-v\mathrm{d}y)\right]+\mathrm{i}\left[\int(v\mathrm{d}x+u\mathrm{d}y)\right]\\&=\int u\mathrm{d}x-\int v\mathrm{d}y+\mathrm{i}\int v\mathrm{d}x+\mathrm{i}\int u\mathrm{d}y\end{aligned} \tag{7.2.78}$$

Well once this expression established, let us give a little explanation about how to read it:

1. We know that u and v are dependent both in the general case of x and y.
2. We know that u and v represent (see examples at the beginning of this section) closed or open curves and also straight lines when x (or respectively y) is fixed and that the other associated variable varies!

So each term have an integral in the above expression is in fact a line integral on a family of open or closed curves (including a specific case that is straight lines...)!

This integral can be evaluated using the Green's theorem in the plane (see section Vector Calculus page 1890) if we consider the particular case of a closed curvilinear path such as:

$$F(z)=\oint(u\mathrm{d}x-v\mathrm{d}y)+\mathrm{i}\oint(v\mathrm{d}x+u\mathrm{d}y) \tag{7.2.79}$$

Let us first study the real part:

Indeed we proved (it is strongly advised to read again this Green's theorem) in the section of Vector Calculus that:

$$\oint(f_x\mathrm{d}x+f_y\mathrm{d}y)=\int\limits_S\left(\frac{\partial f_y}{\partial x}-\frac{\partial f_x}{\partial y}\right)\mathrm{d}S_z=\int\limits_S\left(\frac{\partial f_y}{\partial x}-\frac{\partial f_x}{\partial y}\right)\mathrm{d}x\mathrm{d}y \tag{7.2.80}$$

What will be written in our situation:

$$\oint(u\mathrm{d}x+(-v)\mathrm{d}y)=\int\limits_S\left(\frac{\partial(-v)}{\partial x}-\frac{\partial u}{\partial y}\right)\mathrm{d}x\mathrm{d}y=-\int\limits_S\left(\frac{\partial v}{\partial x}+\frac{\partial u}{\partial y}\right)\mathrm{d}x\mathrm{d}y \tag{7.2.81}$$

However, if the function is holomorphic and thus satisfies the Cauchy-Riemann equations we get immediately:

$$\oint (u\mathrm{d}x + (-v)\mathrm{d}y) = -\int_S \underbrace{\left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right)}_{=0} \mathrm{d}x\mathrm{d}y = 0 \qquad (7.2.82)$$

Thus our integral is reduced in the particular case of a closed path:

$$F(z) = \mathrm{i}\oint (v\mathrm{d}x + u\mathrm{d}y) \qquad (7.2.83)$$

and... reusing Green's theorem for this imaginary part:

$$\oint (v\mathrm{d}x + u\mathrm{d}y) = \int_S \left(\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y}\right) \mathrm{d}x\mathrm{d}y \qquad (7.2.84)$$

However, if the function is holomorphic (for reminder that is to say differentiable at every point of the complex plane or an open subset of it) and thus satisfies the Cauchy-Riemann equations we get immediately:

$$\int_S \underbrace{\left(\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y}\right)}_{=0} \mathrm{d}x\mathrm{d}y \qquad (7.2.85)$$

and we thus obtain the "Cauchy theorem", or "Cauchy-Goursat theorem" for its generalized version for non continuous functions, which says that if a function is holomorphic (thus satisfying the Cauchy-Riemann equations) and integrated on a closed contour then:

$$\boxed{\oint f(z)\mathrm{d}z = \oint (u + \mathrm{i}v)\mathrm{d}z = 0} \qquad (7.2.86)$$

As a corollary (without proof), any function that satisfies the above relation is holomorphic (in the whole complex plane or an open subset of it).

This result gives the possibility in certain fields like quantum physics fields (we think of the Yukawa potential that is not yet treated in this textbook in detailed) to calculate complicated real definite integrals using the above property. The idea is when choosing the closed contour of the path integral to play to make the real definite integral only as a part only of the path (by generalizing to the complex case) and by equality with zero we deduce its value thanks to the other parts of the integrals of the chosen path (parts that are obviously simple to calculate).

In other words, the idea is to calculate by difference! The difficulty residing in practice in finding the function $f(z)$ and the closed contour that permits to make appear the function $f(x)$ of the researched definite integral...

Using this result, let us make a very important academic example which will be useful later (but who has no connection with the case of calculating a real definite integral).

✍**Example:**

Let us calculate:

$$\oint \frac{\mathrm{d}z}{z} \tag{7.2.87}$$

For this purpose, we will use the simplification that consist to remember (see section Numbers page 251) that:

$$z = re^{\mathrm{i}\theta} \tag{7.2.88}$$

Therefore:

$$\mathrm{d}z = r\mathrm{i}e^{\mathrm{i}\theta}\mathrm{d}\theta \tag{7.2.89}$$

We can then write the path integral as:

$$\oint \frac{\mathrm{d}z}{z} = \int \frac{r\mathrm{i}e^{\mathrm{i}\theta}\mathrm{d}\theta}{re^{\mathrm{i}\theta}} = \mathrm{i}\mathrm{d}\theta \tag{7.2.90}$$

Or as on a closed path differentiable at any point (without nodes) the angle to make a full turn will necessarily be between 0 and $+\pi$. It comes then:

$$\oint \frac{\mathrm{d}z}{z} = \int\limits_0^{2\pi} \mathrm{d}\theta = 2\pi\mathrm{i} \tag{7.2.91}$$

Before we continue by noticing a very interesting and important fact that we will detail later formally: An integral (we do not speak of primitive but of integral!) of the type $1/x$ in $\mathbb{R}$ would not be calculable. But now if we generalize the concept of $\mathbb{C}$, we see that we get go around the singularity via a path integral that enclose the singularity. And ... and ... in our previous calculation z might have only the real value and not the imaginary one (so z reduce to x). So the integral of $1/x$ becomes calculable and has a result in the set of complex numbers which is remarkable!

Some mathematicians interpret this by figuring that $1/x$ is a flat projection of a three-dimensional space in which the imaginary axis is perpendicular to the plane $\mathbb{R}^2$ (see figures below). Hence the fact that $1/x$ can integrated in the set $\mathbb{C}$.

Finally, let us indicate that $1/z$ is holomorphic on the whole complex plane except on 0 (the derivative being the same as $1/x$). Then the function $1/z$ is thus not $\mathbb{C}$-differentiable!

This being done, let us do an important and similar case with the following path integral:

$$\oint \frac{1}{z - z_0}\mathrm{d}z \tag{7.2.92}$$

where z_0 is a constant complex number. Let us write:

$$z - z_0 = \rho e^{\mathrm{i}\varphi} \Rightarrow \frac{\mathrm{d}z}{\mathrm{d}\varphi} = \rho\mathrm{i}e^{\mathrm{i}\varphi} \Rightarrow \mathrm{d}z = \rho\mathrm{i}e^{\mathrm{i}\varphi}\mathrm{d}\varphi \tag{7.2.93}$$

We can then write if we make one turn counter clockwise:

$$\oint \frac{\mathrm{d}z}{z-z_0}=\int \frac{\rho \mathrm{i}e^{\mathrm{i}\varphi}\mathrm{d}\varphi}{\rho e^{\mathrm{i}\varphi}}=\mathrm{i}\int \mathrm{d}\varphi=2\pi\mathrm{i} \tag{7.2.94}$$

which is valid only if our integration path avoids z_0 what otherwise there is a singularity. This latter integral is a little simplistic generalization of the previous one.

Now let us show the important theorem that interests us since the beginning of this section using many proven results so far!

We know that if a function $f(z)$ satisfies the Cauchy-Riemann equations, then if we carefully avoid the value z_0 (as in the above calculations), the expression:

$$\frac{f(z)-f(z_0)}{z-z_0} \tag{7.2.95}$$

is differentiable at all points except in on z_0 (where the expression is no longer holomorph) is name a "singularity".

Indeed, take a holomorphic function $f(z)$ satisfying Cauchy-Riemann equation and subtract a constant $(f(z_0))$ does not change the fact that the expression (in this case the numerator in previous relationship) remains holomorphic. Finally, multiply it by a fraction (the denominator of the above equation) which is also holomorph gives a holomorphic function. But singularities can then appear, we then speak of "meromorphic functions" (this is the ratio of two holomorphic functions).

Remark

A meromorphic function is a function holomorphic in the whole complex plane, except possibly on a set of isolated points each of which is a pole (singularity) for the function (see further below for the concept of pole/singularity). The gamma function (see the plot in the Differential and Integral Calculus section at page **??**) is a famous example of meromorphic function!

So if we take the path integral on a closed path avoiding z_0, the Cauchy theorem gives us immediately (remember the proof above):

$$\oint \frac{f(z)-f(z_0)}{z-z_0}\mathrm{d}z=0 \tag{7.2.96}$$

However, this can also be written after rearrangement of terms:

$$\oint \frac{f(z)}{z-z_0}\mathrm{d}z=f(z_0)\oint \frac{1}{z-z_0}\mathrm{d}z \tag{7.2.97}$$

Therefore:

$$f(z_0)=\frac{1}{\displaystyle\oint \frac{1}{z-z_0}\mathrm{d}z}\oint \frac{f(z)}{z-z_0}\mathrm{d}z \tag{7.2.98}$$

But we have proved above that:

$$\oint \frac{\mathrm{d}z}{z-z_0}=2\pi\mathrm{i} \tag{7.2.99}$$

Then we get the result named "Cauchy's integral theorem", or more rarely "Cauchy formula" (of which there is a generalized result we will prove later below):

$$\boxed{f(z_0)=\frac{1}{2\pi\mathrm{i}}\oint\frac{f(z)}{z-z_0}\mathrm{d}z} \tag{7.2.100}$$

In fact, in practice all the subtlety is to be able to take back a given holomorphic function $g(z)$ (which therefore satisfies the Cauchy-Riemann equations) by manipulating it in a form of the type:

$$g(z)=\frac{f(z)}{z-z_0} \tag{7.2.101}$$

when its possible... then the calculation of its path integral (closed path) becomes extremely simple since it will be equal to:

$$2\pi\mathrm{i}f(z_0) \tag{7.2.102}$$

by the Cauchy's integral theorem!

Remark

So we know how to calculate the value of a path integral of an expression that is not holomorph but for which the numerator is holomorph!

Caution! The sign of the value of a path integral will depend on the direction in which its integration path will be done. If the direction is straightforward (that is to say "counter-clockwise") its sign will be positive; if on the contrary the direction is clockwise his sign will be negative. You probably think that this information is irrelevant since this value is usually zero. Yes... it is, but we will see later the importance of this information when referring to the calculation of what we name the "residuals".

✍**Example:**

An important application example is named the "Gauss' mean value theorem" that states given $f(z)$ an analytic function on and inside a disk $C_r : |z - z_0| = r$, then:

$$f(z_0) = \frac{1}{2\pi} \int_0^{2\pi} f(z_0 + re^{\mathrm{i}\theta}) \mathrm{d}\theta \tag{7.2.103}$$

Indeed, from the Cauchy's integral relation we have (we simply make the change of variable $z = z_0 + re^{\mathrm{i}\theta}$):

$$\begin{aligned} f(z_0) &= \frac{1}{2\pi\mathrm{i}} \oint_{C_r} \frac{f(z)}{z - z_0} \mathrm{d}z) = \frac{1}{2\pi\mathrm{i}} \int_0^{2\pi} \frac{f(z_0 + re^{\mathrm{i}\theta})\mathrm{i}re^{\mathrm{i}\theta}}{(z_0 + re^{\mathrm{i}\theta}) - z_0} \mathrm{d}\theta = \frac{1}{2\pi\mathrm{i}} \int_0^{2\pi} \frac{f(z_0 + re^{\mathrm{i}\theta})\mathrm{i}re^{\mathrm{i}\theta}}{re^{\mathrm{i}\theta}} \mathrm{d}\theta \\ &= \frac{1}{2\pi} \int_0^{2\pi} f(z_0 + re^{\mathrm{i}\theta}) \mathrm{d}\theta \end{aligned} \tag{7.2.104}$$

There is a similar relation for the derivative $f'(z_0)$ to that given by the Cauchy's integral theorem. Let us see this:

$$\begin{aligned} f'(z_0) &= \lim_{h \to 0} \frac{f(z_0 + h) - f(z_0)}{h} \\ &= \lim_{h \to 0} \left(\frac{f(z_0 + h)}{h} - \frac{f(z_0)}{h} \right) = \lim_{h \to 0} \frac{1}{h} (f(z_0 + h) - f(z_0)) \\ &= \lim_{h \to 0} \frac{1}{h} \left(\frac{1}{2\pi\mathrm{i}} \oint \frac{f(z)}{z - (z_0 + h)} \mathrm{d}z - \frac{1}{2\pi\mathrm{i}} \oint \frac{f(z)}{z - z_0} \mathrm{d}z \right) \\ &= \lim_{h \to 0} \frac{1}{2\pi\mathrm{i}h} \left(\oint \frac{f(z)\mathrm{d}z}{z - z_0 - h} - \oint \frac{f(z)\mathrm{d}z}{z - z_0} \right) \\ &= \lim_{h \to 0} \frac{1}{2\pi\mathrm{i}h} \left(\oint \frac{f(z)\mathrm{d}z}{z - z_0 - h} \frac{(z - z_0)}{(z - z_0)} - \oint \frac{f(z)\mathrm{d}z}{z - z_0} \frac{(z - z_0 - h)}{(z - z_0 - h)} \right) \\ &= \lim_{h \to 0} \oint \frac{1}{2\pi\mathrm{i}h} \frac{f(z)\left[(z - z_0) - (z - z_0 - h)\right]\mathrm{d}z}{(z - z_0 - h)(z - z_0)} = \lim_{h \to 0} \oint \frac{1}{2\pi\mathrm{i}h} \frac{hf(z)\mathrm{d}z}{(z - z_0 - h)(z - z_0)} \end{aligned} \tag{7.2.105}$$

Therefore:

$$f'(z_0) = \frac{1}{2\pi\mathrm{i}} \oint \frac{f(z)\mathrm{d}z}{(z - z_0)^2} \tag{7.2.106}$$

thereby continuing, we have:

$$\begin{aligned} f''(z_0) &= \frac{2}{2\pi\mathrm{i}} \oint \frac{f(z)\mathrm{d}z}{(z - z_0)^3} \\ f'''(z_0) &= \frac{6}{2\pi\mathrm{i}} \oint \frac{f(z)\mathrm{d}z}{(z - z_0)^4} \end{aligned} \tag{7.2.107}$$

In short, we therefore note that:

$$\boxed{f^{(n)}(z_0)=\frac{\mathrm{d}^n f(z_0)}{\mathrm{d}z^n}=\frac{n!}{2\pi\mathrm{i}}\oint\frac{f(z)}{(z-z_0)^{n+1}}\mathrm{d}z} \tag{7.2.108}$$

which is the "Generalized Cauchy's integral theorem".

This result is very powerful because it shows that holomorphic functions are infinitely differentiable (because of the denominator), that is to say analytical, and it is much more difficult to find an equivalent theorem with such simple conditions for real functions.

If we now return to our Taylor expansion of a complex function:

$$f(z)=\sum_{i=0}^{+\infty}\frac{1}{n!}\frac{\mathrm{d}^n f(z_0)}{\mathrm{d}z^n}(z-z_0)^n \tag{7.2.109}$$

um ... and what do we see here? Well this !:

$$\begin{aligned}f(z)&=\sum_{i=0}^{+\infty}\frac{1}{n!}\frac{\mathrm{d}^n f(z_0)}{\mathrm{d}z^n}(z-z_0)^n=\sum_{i=0}^{+\infty}\frac{1}{n!}\frac{n!}{2\pi\mathrm{i}}\oint\frac{f(z)\mathrm{d}z}{(z-z_0)^{n+1}}(z-z_0)^n\\&=\sum_{i=0}^{+\infty}\frac{1}{2\pi\mathrm{i}}\oint\frac{f(z)}{(z-z_0)^{n+1}}\mathrm{d}z(z-z_0)^n=\sum_{i=0}^{+\infty}c_n(z-z_0)^n\end{aligned} \tag{7.2.110}$$

It follows the following relation named "Laurent series in positive powers" (there is a more generalized version will be prove later below):

$$\boxed{f(z)=\sum_{i=0}^{+\infty}c_n(z-z_0)^n} \tag{7.2.111}$$

that gives us the formal expression of a complex function in the form of infinite series of integer powers near a point z_0 of the complex plane with therefore:

$$\boxed{c_n=\frac{1}{2\pi\mathrm{i}}\oint\frac{f(z)}{(z-z_0)^{n+1}}\mathrm{d}z=\frac{f^n(z_0)}{n!}} \tag{7.2.112}$$

Remembering that $\mathrm{d}^n f(z_0)/\mathrm{d}z^n$ can be written equivalently $f^n(z_0)$, we see that all the two previous relations gives us the Taylor series expansion that we had obtained in real analysis (see section Sequences and Series page 1696) and that was:

$$\boxed{f(z)=\sum_{i=0}^{+\infty}\frac{1}{n!}\frac{\mathrm{d}^n f(z_0)}{\mathrm{d}z^n}(z-z_0)^n} \tag{7.2.113}$$

Thus, the Taylor series in $\mathbb{R}$ are a special case of Laurent series that are in $\mathbb{C}$!!!

This result is quite remarkable because it also shows that we can use the path integral in the complex plane for calculating the coefficients c_n of the Laurent series instead of calculating the derivatives of order n of the function f if these latter are too complicated to determine. Or vice versa... calculate a simple derivation instead of calculating a headache type path integral (typically the case in physics) using the fact that:

$$\oint \frac{f(z)}{(z-z_0)^{n+1}}\mathrm{d}z = 2\pi\mathrm{i}\frac{f^{(n)}(z_0)}{n!} \tag{7.2.114}$$

The only unfortunate point being that the latter relation is calculable only if we can put the function in path line integral in the form:

$$\frac{f(z)}{(z-z_0)^{n+1}} \tag{7.2.115}$$

where n is a positive or null integer. This is honestly far from to be easy in most cases! The idea would be to find a general path for line integral, valid for any function $f(z)$ such that the denominator (which additionally contains a singularity on z_0) disappears. That would be ideal ... but we need a track ... and it will come from the study of the convergence of series of complex powers. Let's see what it is with a qualitative approach!

7.2.4.1 Convergence of a complex series

We saw in the section of Sequences and Series that many real functions could be expressed in Maclaurin series (special case of the Taylor series on $x_0=0$) in the form:

$$f(x)=\sum_{k=0}^{+\infty} a_k x^k \tag{7.2.116}$$

We also showed, by example the only, that this series expansion of infinite powers was valid for some real functions only in a certain domain of definition named "radius of convergence".

Even if this radius of convergence can be determined more or less easily in each case, there are some baffling examples that could not in the early 119th century (holocene calendar) be understood without complex analysis.

Let's see a simple example to understand what kind of problem it is. Consider for this the two functions:

$$g(x)=\frac{1}{1-x^2} \quad \text{and} \quad h(x)=\frac{1}{1+x^2} \tag{7.2.117}$$

and before continuing our example, recall that we have proved in the section of Sequences and Series the relation:

$$S_n=u_0+u_1+u_2+\ldots+u_n=\frac{u_0-u_{n+1}}{1-q}=u_0\frac{1-q^{n+1}}{1-q} \tag{7.2.118}$$

relative to a geometric series, that is to say a series whose terms are of the type:

$$u_n=u_0q^n \tag{7.2.119}$$

Therefore it comes immediately if $n\rightarrow+\infty$ and $q\in]-1,+1[$:

$$S_\infty=u_0+u_1+u_2+\ldots+u_{+\infty}=\sum_{k=0}^{+\infty}u_k=u_0\frac{1-q^{+\infty+1}}{1-q}=u_0\frac{1-0}{1-q}=\frac{u_0}{1-q} \tag{7.2.120}$$

if $u_0=1$, we get:

$$\sum_{k=0}^{+\infty} u_k=\sum_{k=0}^{+\infty} u_0 q^k=\sum_{k=0}^{+\infty} u_0=\frac{1}{1-q} \tag{7.2.121}$$

So if we change the notation, we have:

$$\frac{1}{1-x}=\sum_{n=0}^{+\infty} x^n \tag{7.2.122}$$

Then it comes immediately:

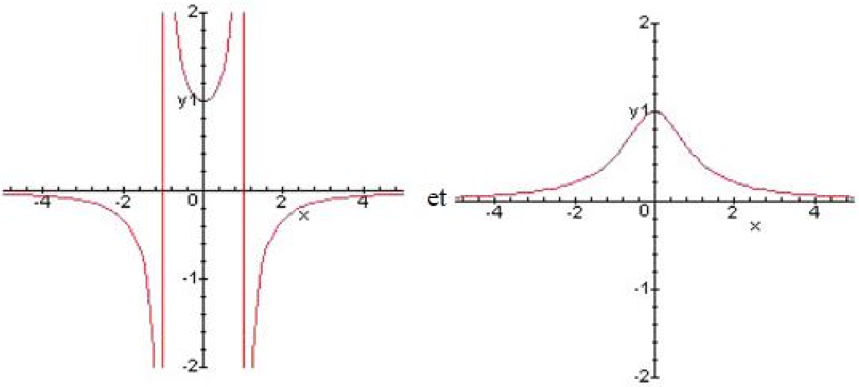
$$g(x)=\sum_{n=0}^{+\infty}(x^2)^n=\sum_{n=0}^{+\infty}(x^2)^n \quad \text{and} \quad h(x)=\sum_{n=0}^{+\infty}(-x^2)^n=\sum_{n=0}^{+\infty}(-1)^n x^{2n} \tag{7.2.123}$$

Therefore the two previous functions $g(x)$ and $h(x)$ are defined for a infinite series expansion in powers only in radius of convergence $x\in]-1,+1[$.

We would get the same result by making a Maclaurin series expansion!

We see that there is trivially for $g(x)$ two singularities that are $x=\{-1,+1\}$ by cons, basically we do not see trivial singularities for $h(x)$ if we reason only in $\mathbb{F}$ so it can be hard for the latter function to understand the origin of the radius of convergence!

Indeed, if we plot these two functions in $\mathbb{R}$ with Maple™ 4.00b we get respectively:



hence the problem of why there is still implicitly a radius of convergence $x\in]-1,+1[$ for $h(x)$???

An even more blatantly way to highlight the problem, is to show the approach of these two functions by a Maclaurin series expansion with ten terms.

For $g(x)$ we get for example:

```
>with(plots):
>xplot:= plot(1/(1-x^2),x=-5..5,thickness=2,color=red):
>tays:= plots[display](xplot):
```

```
>for i from 1 by 2 to 10 do
tpl:= convert(taylor(1/(1-x^2), x=0,i),polynom):
tays:= tays,plots[display]([xplot,plot(tpl,x=-5..5,y=-2..2,
color=black,title=convert(tpl,string))])
od:
>plots[display]([tays],view=[-5..5,-2..2]);
```

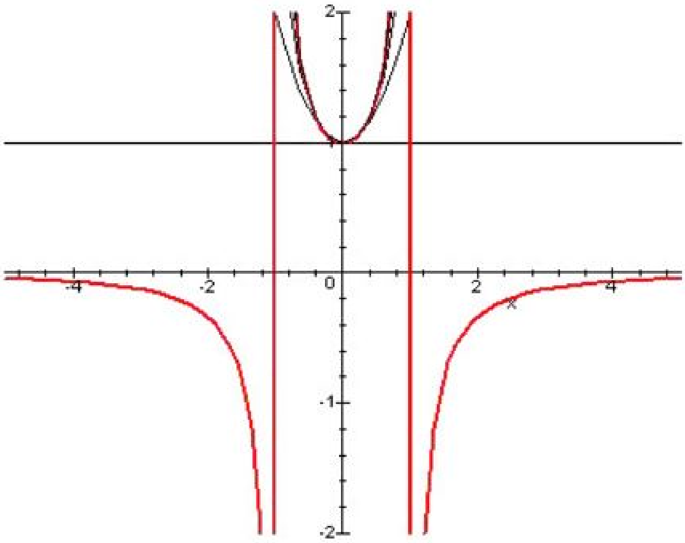


Figure 7.82: **Plane representation of the function g to visualize the problem**

where we see well that the Maclaurin series (or expression in power series) does not converge outside $x \in]-1,+1[$ which can be intuitive because of both singularities.

For $h(x)$ we have by cons:

```
 > with(plots):
> xplot:= plot(1/(1+x^2),x=-5..5,thickness=2,color=red):
> tays:= plots[display](xplot):
> for i from 1 by 2 to 10 do
tpl:= convert(taylor(1/(1+x^2), x=0,i),polynom):
tays:= tays,plots[display]([xplot,plot(tpl,x=-5..5,y=-2..2,
color=black,title=convert(tpl,string))])
od:
> plots[display]([tays],view=[-5..5,-2..2]);
```

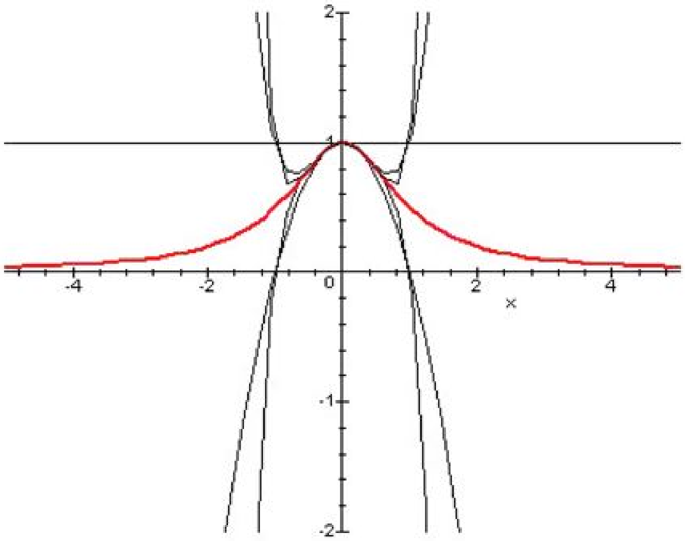


Figure 7.83: **Surprisingly, here the Maclaurin series (in black) does not converge**

where we see well that the Maclaurin series (or the expression in power series) does not converge either outside $x \in]-1,+1[$ which was unsettling and against-intuitive at the beginning of the history of real analysis.

Today even a high school student knows that he can also think in $\mathbb{C}$ and that $\mathbb{R} \subset \mathbb{C}$. So the real analysis is just a special case and restricted of the field of complex analysis. The fact to extend the domain of a given analytic function is named "analytic continuation". As we will see just now analytic continuation often succeeds in defining further values of a function, for example in a new region where an infinite series representation in terms of which it is initially defined becomes divergent!

The singularity for $h(x)$ in $\mathbb{C}$ comes that latter is then written:

$$h(z) = \frac{1}{1+z^2} \tag{7.2.124}$$

and there are therefore two singularities $z = \{-\mathrm{i}, +\mathrm{i}\}$ that we see well if we represent:

$$|h(z)| = \left|\frac{1}{1+z^2}\right| \tag{7.2.125}$$

with Maple™ 4.00b (fortunately we now have the equivalent of a microscope in mathematics with Maple™...):

```
>plot3d(abs(1/(1+(re+I*im)^2)),re=-3..3,im=-3..3,view=[-2..2,-2..2,-2..2]
,orientation=[-130,70],contours=50,style=PATCHCONTOUR,axes=frame,
grid=[100,100],numpoints=10000);
```

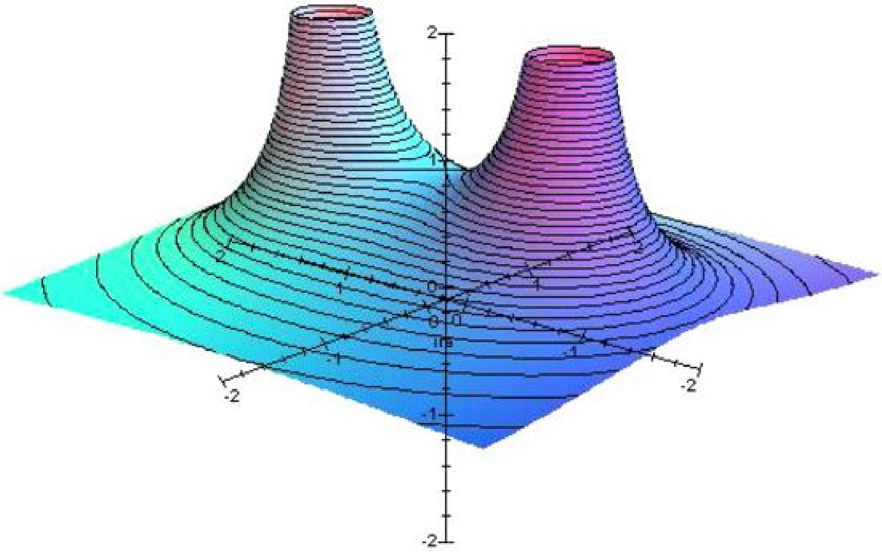


Figure 7.84: **Complex representation of the function h to highlight the reason for the divergence**

where we can see the two singularities on the imaginary axis and the function $h(x)$ on the real axis (between the two peaks). So when we develop a function in power series, we conclude that the radius of convergence is defined by the whole complex plane and not by the traditional axis of the real analysis.

This makes it more natural to understand why we were talking in the section of Sequences and Series of "radius" as seen from above, we have in the complex plane:

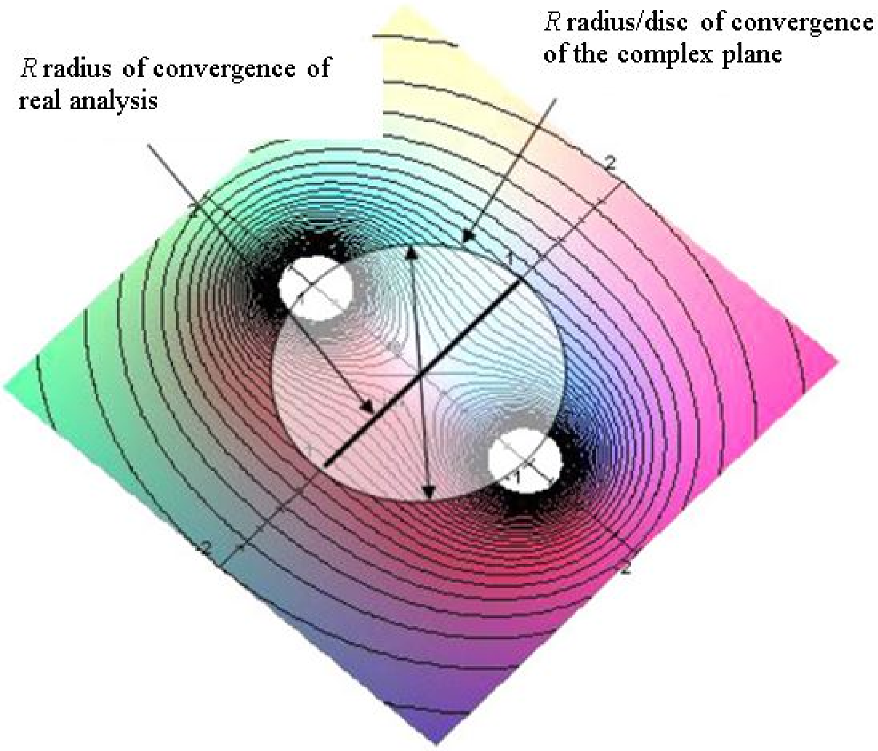


Figure 7.85: **Representation of the various convergence of radius of $h(z)$**

hence the fact that we are talking sometimes about (open) convergence disk and sometimes of (open) convergence radius. Moreover, we notice on the chart that the domain of convergence is convex (any couple of points of the domain can be connected by a straight line that is in the area of convergence).

Remark

Let us Recall that a subset, interval or "open" disc means that we do not take its border as we have seen in the section Topology.

Then we understand better why the Taylor series does not converge trivially for $h(x)$: it must converge on the whole disc of the complex plane and not just converge on the real axis!

From all this we deduce that our Laurent series in positive powers proved above:

$$f(z)=\sum_{k=0}^{+\infty}c_n(z-z_0)^n \tag{7.2.126}$$

not necessarily converge, unsurprisingly... on the whole complex plane (just like the Taylor series on the real line as this is the equivalent!) but sometimes only in a opened subdomain (convex?) of this plane around z_0 (which in the particular example taken above was obviously: 0).

With our function $h(x)$ expressed using a development of Maclaurin with 5 terms, we see immediately with Maple™ 4.00b that on the borders of the square inscribed in the disc of convergence, the series does not converge and we're guessing the start of the two singularities:

```
>plot3d(abs(1-(re+I*im)^2+(re+I*im)^4-(re+I*im)^6+(re+I*im)^8),
re=-0.7..0.7,im=-0.7..0.7,view=[-1.5..1.5,-1.5..1.5,0..1.5]
,orientation=[-130,70],contours=50,style=PATCHCONTOUR,axes=frame,
grid=[100,100],numpoints=10000);
```

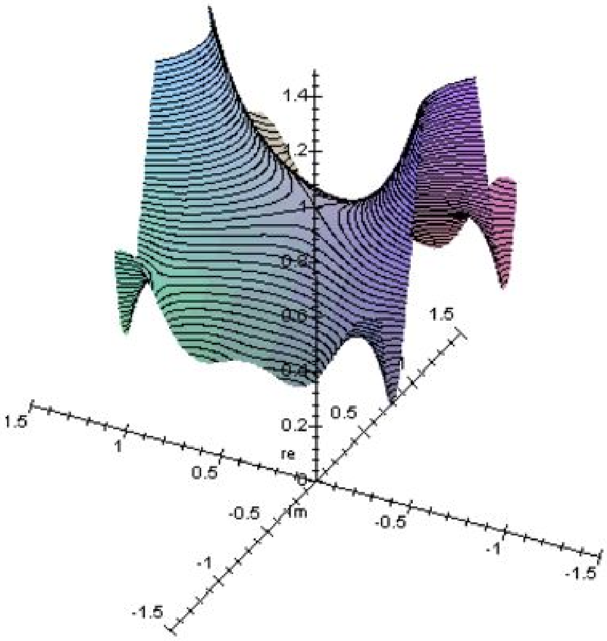


Figure 7.86: Focus on the complex representation to understand the reason for the divergence

a little outside the disc of convergence, we obviously have a little bit nonsense:

```
>plot3d(abs(1-(re+I*im)^2+(re+I*im)^4-(re+I*im)^6+(re+I*im)^8),re=-3..3,
im=-3..3,view=[-1.5..1.5,-1.5..1.5,0..1.5],orientation=[-130,70]
,contours=50,style=PATCHCONTOUR,axes=frame,grid=[100,100],numpoints=10000);
```

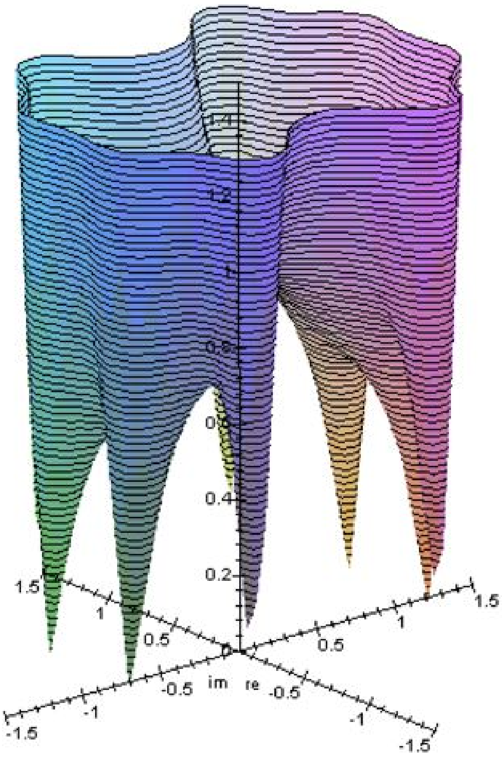


Figure 7.87: This diverges ... (stalactites ???)

There is still something interesting to try ... since we are now on a plane, not a straight line right (axis), it is possible for us to make the Taylor expansion around a singularity z_0 by deforming the disk in a convex crown/ring simply connected as shown below (the crown/ring being the simplest simply convex geometry arising from the deformation of a disk):

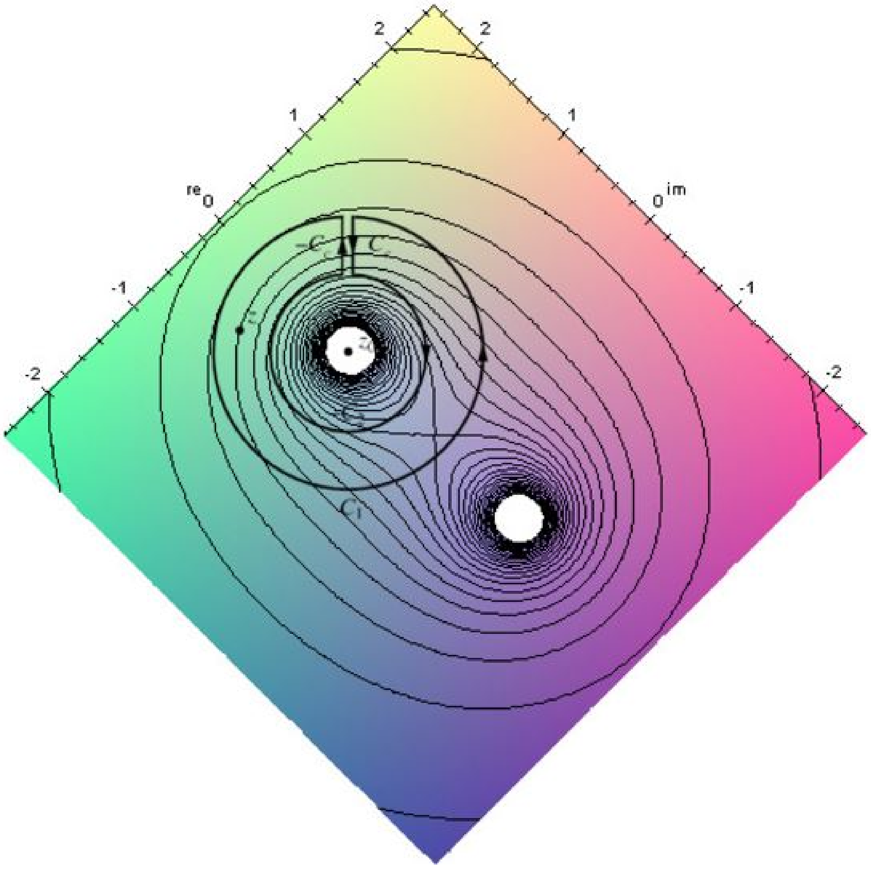


Figure 7.88: **Representation of the deformation of a disc in a crown/ring**

The advantage of this is to deform the area of convergence on the whole complex plane by avoiding (bypassing) all the singularities. Thus, unlike the Taylor series that are only valid on an interval of the x-axis, we would have a new type of series describing a function absolutely everywhere, that is to say before AND after (so around...) singularities!

So obviously we will require that in the deformed crown above the function is always holomorph and analytical (as in the initial convex disc). Before determining what we are going down (generalized Laurent series!), we must first do a study of the decomposition of path integral:

7.2.5 Path Decomposition

The path integrals as given previously can also be written in another form almost classical and used many times in the literature.

Let us see this. First, remember that we have just proved in the special case of a holomorphic function that:

$$\oint f(z)\mathrm{d}z=0 \tag{7.2.127}$$

But a closed path can be seen as a path having a round trip:

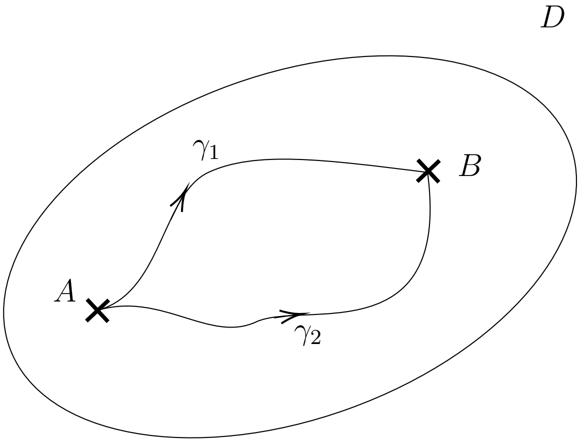


Figure 7.89: **Representation of a closed path with round trip**

Therefore we can write:

$$\int\limits_{\gamma_1} f(z)\mathrm{d}z-\int\limits_{\gamma_2} f(z)\mathrm{d}z=0\Leftrightarrow\int\limits_{\gamma_1} f(z)\mathrm{d}z=\int\limits_{\gamma_2} f(z)\mathrm{d}z \tag{7.2.128}$$

And now comes what interest us... for this purpose let us focus one of the path integral of the type:

$$\int\limits_{\gamma} f(z)\mathrm{d}z \tag{7.2.129}$$

We already well known (1st form of notation) that any complex number z of the type:

$$z=x+\mathrm{i}y \tag{7.2.130}$$

can be (2nd form of notation) written as (Euler form):

$$z(x,y)=x+\mathrm{i}y=r\cos(\theta)+\mathrm{i}\sin(\theta)=re^{\mathrm{i}\theta}=z(r,\theta) \tag{7.2.131}$$

and to integrate on a path, nothing prevent us to choose a path where r (the module) would be fixed and θ variable (we could not have the possibility to do this with the 1st form because by modifying the imaginary or real part, we can't get be guarantee to get a nice smooth curve but this is possible with the Euler form of a complex number)!

Therefore we have:

$$z(\theta)=re^{\mathrm{i}\theta} \tag{7.2.132}$$

We write then naturally:

$$\int\limits_{\gamma} f(z)\mathrm{d}z=\int\limits_{\gamma} f(z(\theta))\mathrm{d}z(\theta) \tag{7.2.133}$$

and as:

$$z'(\theta)=\dfrac{\mathrm{d}z}{\theta}\Leftrightarrow \mathrm{d}z=z'(\theta)\mathrm{d}\theta \tag{7.2.134}$$

Therefore:

$$\int\limits_{\gamma} f(z)\mathrm{d}z=\int\limits_{\gamma} f(z(\theta))z'(\theta)\mathrm{d}\theta \tag{7.2.135}$$

That we often find in the following form in the literature:

$$\boxed{\int\limits_{\gamma} f(z)\mathrm{d}z=\int\limits_{\gamma} f(z(t))\dot{z}(t)\mathrm{d}t} \tag{7.2.136}$$

7.2.5.1 Inverse Path

If C is a curve going from a point P to a point Q, then we denote by C^{-} the same curve but travelled from Q to P.

Now we will parametrize C^{-}.

If $C(t)$ it is the curve defined on $[a,b]$, then we define the curve $C^{-}(t)$ on $[a,b]$ by:

$$C^{-}(t)=C(a+b-t) \tag{7.2.137}$$

Indeed we have with this parametrization:

$$C^{-}(a)=C(a+b-a)=C(b)\quad \text{and}\quad C^{-}(b)=C(a+b-b)=C(a) \tag{7.2.138}$$

and when t increases from a to b, $a+b-t$ decreases from b to a. C^{-} is therefore only C but travelled in the opposite direction.

We then have using the last proof:

$$\int\limits_{C^{-}} f(C^{-})\mathrm{d}C^{-}=\int\limits_{a}^{b} f(C^{-}(t))C'^{-}(t)\mathrm{d}t=\int\limits_{a}^{b} f(C(a+b-t))C'(a+b-t)\mathrm{d}t \tag{7.2.139}$$

Let us put:

$$u=a+b-t \tag{7.2.140}$$

Therefore:

$$\mathrm{d}u = -\mathrm{d}t \tag{7.2.141}$$

Then we have:

$$\begin{aligned}\int\limits_{C^-} f(C^-)\mathrm{d}C^- &= \int\limits_a^b f(C(a+b-t))C'(a+b-t)\mathrm{d}t = \int\limits_a^b f(C(u))C'(u)(-\mathrm{d}u)\\ &= -\int\limits_a^b f(C(u))C'(u)\mathrm{d}u = -\int\limits_a^b f(C)\mathrm{d}C\end{aligned} \tag{7.2.142}$$

Therefore if C^- and C are the paths of the same function but travel in the opposite direction, we have by taking our conventional notation (caution! In the second term it is implicit that the parametrization is different from the first one!):

$$\int\limits_{\gamma} f(z)\mathrm{d}z = -\int\limits_{\gamma^-} f(z)\mathrm{d}z \tag{7.2.143}$$

Therefore:

$$\boxed{\int\limits_{\gamma} f(z)\mathrm{d}z + \int\limits_{\gamma^-} f(z)\mathrm{d}z = 0} \tag{7.2.144}$$

this is why we often say that the sign of the value of a line integral will depend on the direction in which its integration path is travel. If the direction is straightforward (that is to say "counter-clockwise") its sign will be positive; if on the contrary the direction is clockwise its sign will be negative (see section Differential and Integral Calculus page 1586).

7.2.6 Laurent Series

This last relation obtained, we can return to the deformation of our disc of convergence in a crown. We recall that initially the idea is to have the analytical expression of a function as an infinite series of powers in a limited area around a singularity point and all this... in the purpose to be able to calculate for physicists complex path integrals through a method using the properties of complex series!

Let's start with the point (2) that is to say have an infinite power series for a path integral, which will take us more easily to point (1) that is to say get the an analytical expression of a function around a singularity point, by zooming on our crown:

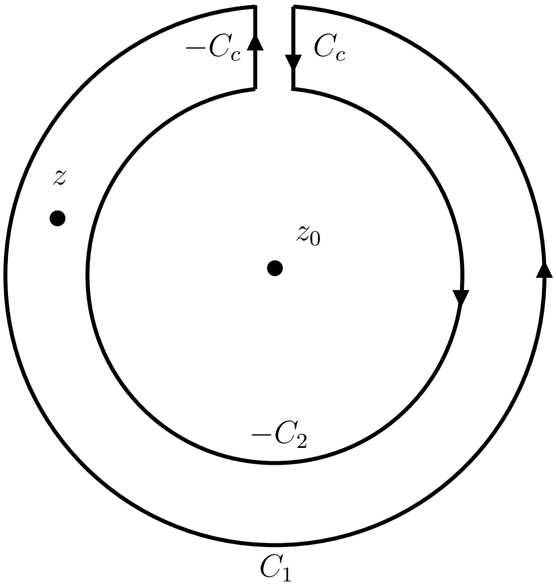


Figure 7.90: **Zoom on our crown from our starting example**

We therefore have if the function f is analytic and holomorphic in the crown of outer radius R and inside radius r, the following path curvilinear integral in the crown as we proved above (we change notation: $z = z'$ and $z_0 = z$):

$$f(z) = \frac{1}{2\pi\mathrm{i}} \oint \frac{f(z')}{z'-z}\mathrm{d}z' \tag{7.2.145}$$

therefore we denote now by z the point where we want to know the function and z' variable of which f depends. This notation change will be justified later for a purely practical reason.

The crown can be broken down into four paths:

$$f(z) = \frac{1}{2\pi\mathrm{i}} \int\limits_{C_1} \frac{f(z')}{z'-z}\mathrm{d}z' + \frac{1}{2\pi\mathrm{i}} \int\limits_{C_c} \frac{f(z')}{z'-z}\mathrm{d}z' + \frac{1}{2\pi\mathrm{i}} \int\limits_{-C_c} \frac{f(z')}{z'-z}\mathrm{d}z' + \frac{1}{2\pi\mathrm{i}} \int\limits_{-C_2} \frac{f(z')}{z'-z}\mathrm{d}z' \tag{7.2.146}$$

If both segments C_c and $-C_c$ are infinitely close, they then correspond to the same path travelled once in a positive direction and once in the negative direction. As we have proved just above that:

$$\int\limits_{\gamma} f(z)\mathrm{d}z=-\int\limits_{\gamma^-} f(z)\mathrm{d}z \tag{7.2.147}$$

It therefore follows that:

$$\int\limits_{C_c} \frac{f(z')}{z'-z}\mathrm{d}z'=-\int\limits_{-C_c} \frac{f(z')}{z'-z}\mathrm{d}z' \tag{7.2.148}$$

Which brings us to write:

$$\begin{aligned} f(z)&=\frac{1}{2\pi\mathrm{i}}\int\limits_{C_1}\frac{f(z')}{z'-z}\mathrm{d}z'+\frac{1}{2\pi\mathrm{i}}\int\limits_{C_c}\frac{f(z')}{z'-z}\mathrm{d}z'-\frac{1}{2\pi\mathrm{i}}\int\limits_{C_c}\frac{f(z')}{z'-z}\mathrm{d}z'+\frac{1}{2\pi\mathrm{i}}\int\limits_{-C_2}\frac{f(z')}{z'-z}\mathrm{d}z'\\ &=\frac{1}{2\pi\mathrm{i}}\int\limits_{C_1}\frac{f(z')}{z'-z}\mathrm{d}z'-\frac{1}{2\pi\mathrm{i}}\int\limits_{C_2}\frac{f(z')}{z'-z}\mathrm{d}z'\\ &=\frac{1}{2\pi\mathrm{i}}\int\limits_{\gamma_R}\frac{f(z')}{z'-z}\mathrm{d}z'-\frac{1}{2\pi\mathrm{i}}\int\limits_{\gamma_r}\frac{f(z')}{z'-z}\mathrm{d}z'=f_1(z)+f_2(z)\end{aligned} \tag{7.2.149}$$

where we have put a "+" between the last two terms, because as we shall see immediately, the convergence criterion associated with the traditional notation in this field of study, makes automatically emerge the sign "-".

For the two integral f_1,f_2, we know that the fraction can be written as a geometric series as already seen above. Effectively, starting from (now you will understand why we changed the notation):

$$\frac{1}{z'-z}=\frac{1}{(z'-z_0)+(z_0-z)}=\frac{1}{z'-z_0}\frac{1}{1-\dfrac{z-z_0}{z'-z_0}} \tag{7.2.150}$$

by assimilating:

$$x=\frac{z-z_0}{z'-z_0} \tag{7.2.151}$$

where as we have seen, the convergence requires that:

$$|z-z_0|\leq|z'-z_0| \tag{7.2.152}$$

so that x is lower in absolute value to 1.

We then see the infinite geometric series appearing:

$$\frac{1}{z'-z_0}=\frac{1}{z'-z_0}\left[1+\left(\frac{z-z_0}{z'-z_0}\right)+\left(\frac{z-z_0}{z'-z_0}\right)^2+\ldots+\left(\frac{z-z_0}{z'-z_0}\right)^n+\ldots\right] \tag{7.2.153}$$

Therefore:

$$\frac{1}{z'-z_0}=\frac{1}{z'-z_0}+\frac{z-z_0}{(z'-z_0)^2}+\frac{(z-z_0)^2}{(z'-z_0)^3}+\ldots+\frac{(z-z_0)^n}{(z'-z_0)^{n+1}}+\ldots \tag{7.2.154}$$

To come back to:

$$f_1(z)=\frac{1}{2\pi \mathrm{i}}\int\limits_{\gamma_R}\frac{f(z')}{z'-z}\mathrm{d}z' \tag{7.2.155}$$

we have in any point z inside the circle of radius R whose border is described by the variable z' and of center z_0 the convergence that is assured because:

$$|z-z_0|\leq|z'-z_0| \tag{7.2.156}$$

Then we can write:

$$f_1(z)=\frac{1}{2\pi \mathrm{i}}\int\limits_{\gamma_R}f(z')\left[\frac{1}{z'-z_0}+\frac{z-z_0}{(z'-z_0)^2}+\frac{(z-z_0)^2}{(z'-z_0)^3}+\ldots+\frac{(z-z_0)^n}{(z'-z_0)^{n+1}}+\ldots\right]\mathrm{d}z' \tag{7.2.157}$$

Integrating term by term integration, we highlight the development (already known):

$$f_1(z)=\sum_{n=0}^{+\infty}c_n(z-z_0)^n \tag{7.2.158}$$

with the definition of coefficients c_n, where n is a positive or null integer:

$$c_n=\frac{1}{2\pi \mathrm{i}}\int\limits_{\gamma_R}\frac{f(z')}{(z'-z_0)^{n+1}}\mathrm{d}z' \tag{7.2.159}$$

This development may do think to the development of Taylor in the sense that only positive (or zero) powers of $(z'-z_0)$ appear, but this is not Taylor development in the case of the crown! Indeed, c_n can not be written this time as:

$$c_n=\frac{f^n(z_0)}{n!} \tag{7.2.160}$$

since, by assumption, $f(z)$ is assumed analytic in the crown only and may therefore very well not be inside the small circle of radius r, in particular on z_0, in which case $f^n(z_0)$ may simply not exist (let us repeat that z is strictly constrained to be in the crown, therefore $r<|z|<R$). We will see later what happens when $f(z)$ is holomorphic in this disk and that, in particular, z_0 is not a singular point.

We still need to treat f_2. We then do the same type of development as for f_1, with the difference that now:

$$|z-z_0|>|z'-z_0| \tag{7.2.161}$$

when z' browse the small circle of radius r. To make a geometric series appear, we must write this time:

$$\frac{1}{z-z'}=-\frac{1}{z-z'}=\frac{1}{(z-z_0)-(z'-z_0)}=-\frac{1}{z-z_0}\frac{1}{1-\dfrac{z'-z_0}{z-z_0}} \tag{7.2.162}$$

Therefore:

$$\frac{1}{z-z'}=-\frac{1}{z-z_0}\frac{1}{1-\dfrac{z'-z_0}{z-z_0}}=-\frac{1}{z-z_0}\left[1+\left(\frac{z'-z_0}{z-z_0}\right)+\left(\frac{z'-z_0}{z-z_0}\right)^2+\ldots+\left(\frac{z'-z_0}{z-z_0}\right)^n+\ldots\right] \tag{7.2.163}$$

So we have:

$$\frac{1}{z-z'}=-\frac{1}{z-z_0}-\frac{z'-z_0}{(z-z_0)^2}-\frac{(z'-z_0)^2}{(z-z_0)^3}-\ldots-\frac{(z'-z_0)^n}{(z-z_0)^{n+1}}-\ldots \tag{7.2.164}$$

Integrating term by term, we highlight the (new) development:

$$f_2(z)=\sum_{n=1}^{+\infty}c_{-n}(z-z_0)^{-n} \tag{7.2.165}$$

with:

$$c_{-n}=\frac{1}{2\pi\mathrm{i}}\int\limits_{\gamma_r}f(z')(z'-z_0)^{-n-1}\mathrm{d}z' \tag{7.2.166}$$

By changing n in $-n$ in the summation for f_2, we have for the sum $f_1(z)+f_2(z)$:

$$f(z)=\sum_{n=-\infty}^{+\infty}c_n(z-z_0)^n \tag{7.2.167}$$

with at this time two distinct c_n:

$$c_{n\geq 0}=\frac{1}{2\pi\mathrm{i}}\int\limits_{\gamma_R}\frac{f(z')}{(z'-z_0)^{n+1}}\mathrm{d}z' \quad \text{and} \quad c_{n<0}=\frac{1}{2\pi\mathrm{i}}\int\limits_{\gamma_r}f(z')(z'-z_0)^{-n-1}\mathrm{d}z' \tag{7.2.168}$$

We will now see that these two relations can be combined into one!

For this purpose if we observe well the last two relations, we find that they do not depend at all of z (!) and this is normal since the c_n are the coefficients of the series expansion of $f(z)$ and these are the same at any point of the domain of definition of the function where it is analytic!

So the two contours (circles) can be merged into only one circle since it is located in the crown and has for center z_0:

$$c_{n\geq 0}=\frac{1}{2\pi\mathrm{i}}\int\limits_{\gamma_R}\frac{f(z')}{(z'-z_0)^{n+1}}\mathrm{d}z' \qquad (n=0,\pm 1,\pm 2,\ldots) \tag{7.2.169}$$

Furthermore, the attentive reader will have noticed that this contour does not even need to be a circle finally! It may be any geometry as long as it is closed and is located in an analytical area!

Thus, we get the two relations:

$$\boxed{\begin{aligned} f_{(z)}&=\sum_{n=-\infty}^{+\infty}c_n(z-z_0)^n\\ c_n&=\frac{1}{2\pi\mathrm{i}}\int\limits_{\gamma_R}\frac{f(z')}{(z'-z_0)^{n+1}}\mathrm{d}z' \qquad (n=0,\pm 1,\pm 2,\ldots)\end{aligned}} \tag{7.2.170}$$

The two previous relation define the "general Laurent series". It is remarkable and differs from a Taylor series in the sense that it contains all the positive and negative integer powers and the coefficients c_n can a priori not be expressed with the derivatives of f.

The power series of $n \geq 0$ is named "regular part", the negative powers is commonly named "main part".

The series of negative powers converges uniformly everywhere outside γ_r, that of positive powers within γ_R. In total the development of Laurent converges uniformly in the common area, which is the crown and therefore also on the unique path γ.

Let us now show a point that we have mentioned above. If the circle contains no singularity, then all the coefficients:

$$c_{n<0} = \frac{1}{2\pi \mathrm{i}} \int\limits_{\gamma} f(z')(z'-z_0)^{-n-1} \mathrm{d}z' \qquad (7.2.171)$$

are zero. First note that $-n-1$ is a positive or zero integer, which we will denote by p such as:

$$c_{n<0} = \frac{1}{2\pi \mathrm{i}} \int\limits_{\gamma} f(z')(z'-z_0)^{p} \mathrm{d}z' \quad p \in \mathbb{N} \qquad (7.2.172)$$

We then have the following integrand along a closed path:

$$f(z')(z'-z_0)^p \qquad (7.2.173)$$

But, if we remove the singularity that requires $f(z')$ is holomorphic (and in anyway this is required by all initial developments of the Laurent series).

As $(z'-z_0)^p$ is polynomial with positive and not null integer powers and that as we know any polynomial satisfying these conditions is differentiable at least once without showing singularity. Thus this term is also holomorphic.

Assuming that the product of two holomorphic functions is holomorphic and that contour γ is closed, then we have using the following result proved above (for a holomorphic function):

$$\oint f(z) \mathrm{d}z = 0 \qquad (7.2.174)$$

the following immediate consequence:

$$\int\limits_{\gamma} f(z')(z'-z_0)^p \mathrm{d}z' = 0 \Rightarrow c_{n<0} = \frac{1}{2\pi \mathrm{i}} \int\limits_{\gamma} f(z')(z'-z_0)^p \mathrm{d}z' = 0 \qquad (7.2.175)$$

if there is no singularity in the small circle of the crown. We then fall back again on a development with only positive powers, the $c_{n\geq 0}$ this time being equal to:

$$c_n = \frac{f^n(z_0)}{n!} \qquad (7.2.176)$$

according to the generalized Cauchy's integral theorem proved earlier above. Conversely, we see well that this is the main part (when it exists!) which contains the information on the fact that f is not a priori holomorphic in the small disk. The existence of negative powers shows that f is clearly not bounded on z_0.

The classification of singularities of a function will be precisely based on the consideration of the characteristics of the main part of the Laurent development centered on a singular point of this function.

✍**Example:**

Let us see to what looks like the Laurent series of our famous example function:

$$f(z)=\frac{1}{1+z^2} \tag{7.2.177}$$

on a simply convex domain that would be the crown rounding the singularity i for example (we could have taken the second singularity $-\mathrm{i}$ but we had to choose one to not repeat twice the explanations below...). This is equivalent therefore to search the power series development of $z-\mathrm{i}$.

We will proceed as following:

$$f(z)=\frac{1}{1+z^2}=\frac{1}{(z+\mathrm{i})(z-\mathrm{i})} \tag{7.2.178}$$

For what will follow we will use:

$$f(z)=\frac{1}{(z+\mathrm{i})}=\frac{1}{2\mathrm{i}+(z-\mathrm{i})}=-\frac{\mathrm{i}}{2}\frac{1}{1-\dfrac{\mathrm{i}}{2}(z-\mathrm{i})} \tag{7.2.179}$$

The second fraction can be expressed as a geometric series if as we have already seen:

$$\left|\frac{\mathrm{i}}{2}\frac{1}{1-\dfrac{\mathrm{i}}{2}(z-\mathrm{i})}\right|<1 \tag{7.2.180}$$

Therefore it comes:

$$\frac{1}{1-\dfrac{\mathrm{i}}{2}(z-\mathrm{i})}=1+\frac{\mathrm{i}}{2}(z-\mathrm{i})+\left(\frac{\mathrm{i}}{2}(z-\mathrm{i})\right)^2+\left(\frac{\mathrm{i}}{2}(z-\mathrm{i})\right)^3+\ldots \tag{7.2.181}$$

Let us multiply both sides of this equality by $-\mathrm{i}/2$ and then divide them by $z-\mathrm{i}$ (the second term in the denominator of the original fraction) for obtain the left term:

$$\begin{aligned}\frac{1}{1-\dfrac{\mathrm{i}}{2}(z-\mathrm{i})}\frac{-\mathrm{i}}{2}\frac{1}{z-\mathrm{i}}&=\frac{-\mathrm{i}}{2-\mathrm{i}(z-\mathrm{i})}\frac{\mathrm{i}}{\mathrm{i}}\frac{1}{z-\mathrm{i}}=\frac{1}{2\mathrm{i}+(z-\mathrm{i})}\frac{1}{(z-\mathrm{i})}\\&=\frac{1}{2\mathrm{i}z+2+z^2-2z\mathrm{i}-1}=\frac{1}{z^2+1}\end{aligned} \tag{7.2.182}$$

and for the right term:

$$-\frac{\mathrm{i}}{2}\frac{1}{z-\mathrm{i}}-\left(\frac{\mathrm{i}}{2}\right)^2-\left(\frac{\mathrm{i}}{2}\right)^3(z-\mathrm{i})-\left(\frac{\mathrm{i}}{2}\right)^4(z-\mathrm{i})^2-\ldots \tag{7.2.183}$$

Finally we have the following geometric series:

$$\frac{1}{z^2+1} = -\frac{\mathrm{i}}{2}\frac{1}{z-\mathrm{i}} - \left(\frac{\mathrm{i}}{2}\right)^2 - \left(\frac{\mathrm{i}}{2}\right)^3 (z-\mathrm{i}) - \left(\frac{\mathrm{i}}{2}\right)^4 (z-\mathrm{i})^2 - \ldots \qquad (7.2.184)$$

We see then on this Laurent series around i of the holomorphic function $f(z)$ that the following coefficient appears:

$$c_{-1} = -\frac{\mathrm{i}}{2}, c_0 = -\left(\frac{\mathrm{i}}{2}\right)^2, c_1 = -\left(\frac{\mathrm{i}}{2}\right)^3 \qquad (7.2.185)$$

and then we have with Maple™ 4.00b:

```
>plot3d(abs(-I/2*1/((re+I*im)-I)-(I/2)^2-(I/2)^3*(re-I*im)-
(I/2)^4*(re-I*im)^2-(I/2)^5*(re-I*im)^3),
re=-1.5..1.5,im=-1.5..1.5,view=[-2..2,-2..2,-1..2],
orientation=[-130,70],contours=50,style=PATCHCONTOUR,axes=frame,
grid=[100,100],numpoints=10000);
```

This gives the following figure:

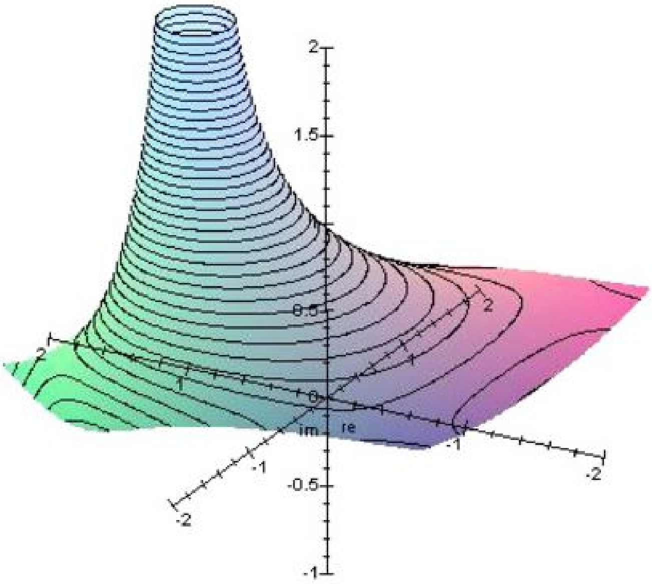


Figure 7.91: **Laurent series representation of $f(z)$ in Maple™ 4.00b**

where we see that the Laurent series allows us to express $f(z)$ in a neighbourhood close to the singularity i by taking five terms.

Ditto if we make the sum of the two Laurent series for the two singularities with seven terms:

```
>plot3d(abs(-I/2*1/((re+I*im)-I)-(I/2)^2-(I/2)^3*(re-I*im)-
(I/2)^4*(re-I*im)^2-(I/2)^5*(re-I*im)^3 -(I/2)^6*(re-I*im)^4
-(I/2)^7*(re-I*im)^5+I/2*1/((re+I*im)+I)+(I/2)^2+
(I/2)^3*(re+I*im)+(I/2)^4*(re+I*im)^2+(I/2)^5*(re+I*im)^3
+(I/2)^6*(re+I*im)^4 +(I/2)^7*(re+I*im)^5),re=-1.5..1.5,
im=-1.5..1.5, view=[-2..2,-2..2,-1..2],orientation=[130,70],
contours=50, style=PATCHCONTOUR, axes=frame,grid=[100,100],
numpoints=10000);
```

This gives the following graph:

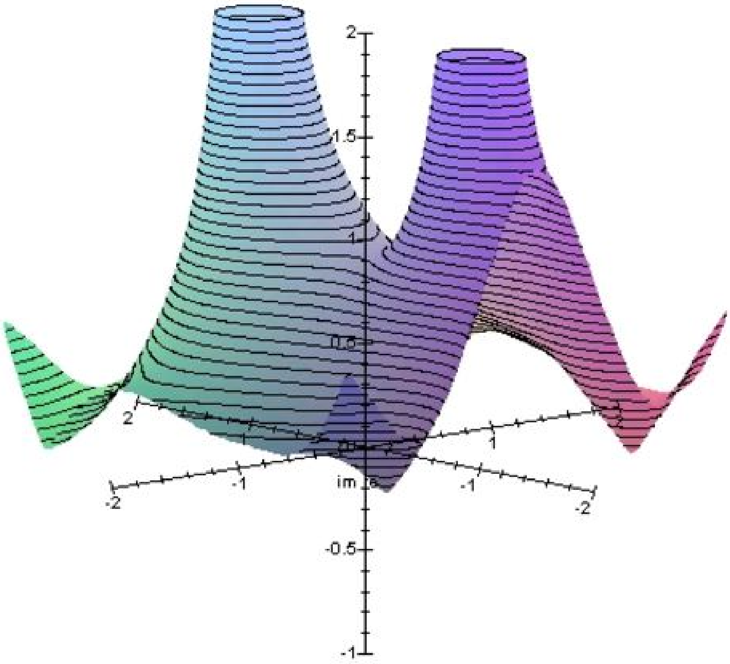


Figure 7.92: **Sum of the two Laurent series of $f(z)$ for both singularities in Maple™ 4.00b**

7.2.7 Singularities

We have seen just before that it was possible to calculate the path integral of a function, on condition of analyticity, on the outline of a singularity. Our goal will now be to enhance this approach.

We have already mentioned and highlighted in our previous proofs that the integrant in the "Cauchy's integral theorem" was of the form:

$$\frac{f(z)}{z-z_0} \tag{7.2.186}$$

where $f(z)$ is well defined in z_0.

The point $z-z_0$ is of course a singularity of $f(z)$ and it is not defined there.

As we saw during our proof of Laurent series, $f(z)$ can be expressed as a Laurent series in the form a positive power Laurent series in a convergence disk (or what remains the same, as a series of Laurent in a crown not centered on a singularity...) in the form:

$$\begin{aligned} f(z)=\frac{1}{z-z_0}&=\left(f(z_0)+\frac{1}{1!}f'(z_0)(z-z_0)^1+\frac{1}{2!}f''(z_0)(z-z_0)^2+\ldots\right)\\ &=\frac{f(z_0)}{z-z_0}+f'(z_0)+\frac{1}{2!}f'(z_0)(z-z_0)+\ldots \end{aligned} \tag{7.2.187}$$

Before continuing, it is customary in mathematics to define a small conventional vocabulary regarding this time the possible singularities of $f(z)$!

Let us first recall that we know, and that we have proved, that all information on the singularities of $f(z)$ are contained in the main part of the Laurent series (negative powers) defined on the crown surrounding z_0:

$$\sum_{n=1}^{+\infty} c_{-n}(z-z_0)^{-n} \tag{7.2.188}$$

The following classification focus on "isolated singularities", that is to say, a singular point where $f(z)$ is analytic everywhere in the neighbourhood of z_0 excepted on z_0 itself. This classification, as we will see permits us to distinguish three types of singular points, will be useful when developing the theory of residues further.

Definitions (#7.2.289):

D1. When the limit of the function $|f(z)|$ exists on z_0, we say that the singularity is a "removable singular point" or "apparent singularity".

For example:

$$f(z)=\frac{\sin(z)}{z} \tag{7.2.189}$$

does not seem to be defined on $z=z_0=0$ but we have a numerator having a Laurent series without negative powers (therefore a simple Taylor series). It then comes into by doing the Maclaurin series (that is to say the Taylor series on $z=z_0=0$...):

$$f(z)=\frac{\sin(z)}{z}=\frac{1}{z}\left(z-\frac{z^3}{3}+\ldots\right)=1-\frac{z^2}{3!}+\ldots \tag{7.2.190}$$

We then see that $f(z)$ finally has no term with negative power and therefore we have eliminated the singularity (or that it contains simply no singularities... which can easily be check with Maple™ 4.00b).

D2. When on z_0 the limit $|f(z)|$ does not exist we speak about "essential singularity".

For example, $z_0=0$ is an essential singularity for the function:

$$f(z)=e^{\frac{1}{z}} \tag{7.2.191}$$

Indeed, if z approaches zero coming from the positive real axis $\mathbb{R}_+$, the function diverges, more precisely, it tends to $+\infty$. If z comes from $\mathbb{R}_-$, the function tends to zero as illustrated by the following Maple™ 4.00b plot:

```
>plot3d(abs(exp(1/(re+I*im))),re=-5..5,im=-5..5,
view=[-3..3,-3..3,-0.5..3],orientation=[-130,70],contours=50,
style=PATCHCONTOUR,axes=frame,grid=[100,100],numpoints=10000)
>plot3d(abs(exp(1/(re+I*im))),re=-5..5,im=-5..5,
view=[-3..3,-3..3,-0.5..3],orientation=[-130,70],contours=50,
style=PATCHCONTOUR,axes=frame,grid=[100,100],numpoints=10000)
```

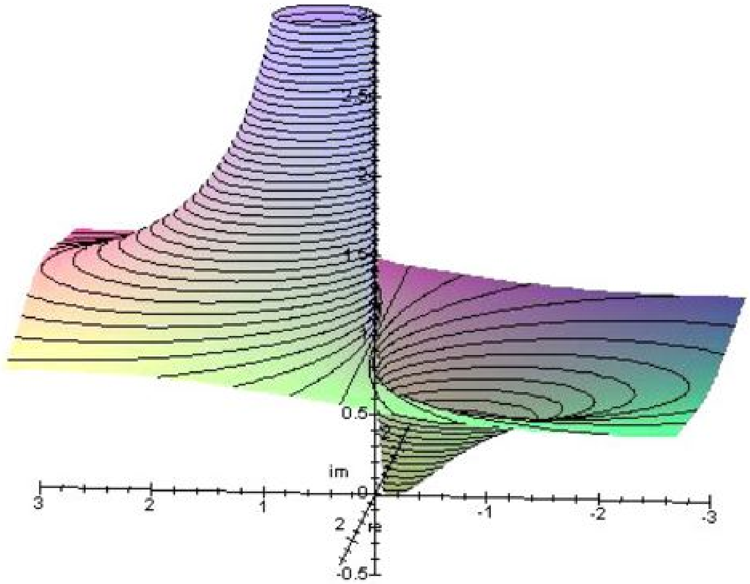


Figure 7.93: **Essential singularity example with $e^{1/z}$ in Maple™ 4.00b**

Indeed:

$$f(z)=e^{1/z}=\sum_{k=0}^{+\infty}\frac{\left(\dfrac{1}{z}\right)^2}{n!}=\sum_{k=0}^{+\infty}\frac{z^{-n}}{n!} \tag{7.2.192}$$

So an equivalent way of defining an essential singularity, is to say that there are an infinite number of terms with negative powers in the main part of the Laurent series.

D3. When on z_0 the limit of $|f(z)|$ is $+\infty$, we speak about a "pole".

This is the last category (as far as we know...) in which we can store a function that is not classifiable neither in the first nor in the second definition above.

So another equivalent way of defining a "pole", is to say that there is a finite number of terms with negative powers in the main part of the Laurent series. If the number of terms is k, then we speak of "pole of order k".

Remarks

R1. We sometimes say that an essential singularity is a " pole of order $+\infty$".

R2. A pole of order 1 is named a "simple pole". One of order 2 is named a "double pole" and so on...

If we come back on our example:

$$f(z)=\frac{1}{1+z^2} \tag{7.2.193}$$

We have proved previously that the Laurent series of this function in $z_0=\mathrm{i}$ was:

$$\frac{1}{z^2+1}=-\frac{\mathrm{i}}{2}\frac{1}{z-\mathrm{i}}-\left(\frac{\mathrm{i}}{2}\right)^2-\left(\frac{\mathrm{i}}{2}\right)^3(z-\mathrm{i})-\left(\frac{\mathrm{i}}{2}\right)^4(z-\mathrm{i})^2-\ldots \tag{7.2.194}$$

This function has therefore a trivial pole of order 1 on $z_0=\mathrm{i}$ and also on $z_0=-1$ because in this latter case this infinite series diverge to $+\infty$ and we can easily check this with the following Maple™ 17.00 command:

```
>sum(-(I*(1/2))^n*(-2*I)^(n-2), n = 0 .. infinity)
```

7.2.8 Residue Theorem

Consider a function $f(z)$ whose pole is of order less or equal to k.

Let us make it analytic:

$$\phi(z)=(z-z_0)^k f(z) \tag{7.2.195}$$

that is to say that we have take a function $f(z)$ that we have made analytic after elimination of the poles supposed in finite number - order - less than or equal to k on z_0.

This function $\phi(z)$ has therefore a Laurent series development in a disc center on z_0.

As we have prove it previously, we can therefore by using the following relation:

$$\phi^{(k)}(z_0)=\frac{\mathrm{d}^k\phi(z_0)}{\mathrm{d}z^k}=\frac{k!}{2\pi\mathrm{i}}\oint\frac{\phi(z)}{(z-z_0)^{k+1}}\mathrm{d}z \tag{7.2.196}$$

write:

$$\phi^{(k-1)}(z_0)=\frac{(k-1)!}{2\pi\mathrm{i}}\oint\frac{\phi(z)}{(z-z_0)^{k}}\mathrm{d}z \tag{7.2.197}$$

Using $f(z)$ under the integral it comes:

$$\phi^{(k-1)}(z_0)=\frac{(k-1)!}{2\pi\mathrm{i}}\oint f(z)\mathrm{d}z \tag{7.2.198}$$

You must deeply analyse this relation and understand that it link together the integral of a function having singularities with the value on one point of an analytical function having no more singularities!!!

This latter relation can be rewritten by rearranging terms:

$$\oint f(z)\mathrm{d}z=\frac{2\pi\mathrm{i}}{(k-1)!}\phi^{(k-1)}(z_0) \tag{7.2.199}$$

And by expressing $\phi^{(k)}(z_0)$ by using (this is authorized because this latter function is analytical) the fact that by definition:

$$\phi^{(k)}(z_0)=\frac{\mathrm{d}^k\phi(z_0)}{\mathrm{d}z^k} \tag{7.2.200}$$

We get obviously:

$$\oint f(z)\mathrm{d}z=\frac{2\pi\mathrm{i}}{(k-1)!}\frac{\mathrm{d}^{(k-1)}\phi(z_0)}{\mathrm{d}z^{(k-1)}} \tag{7.2.201}$$

Therefore by making $\phi(z)$ explicit again:

$$\boxed{\oint f(z)\mathrm{d}z=\frac{2\pi\mathrm{i}}{(k-1)!}\frac{\mathrm{d}^{(k-1)}}{\mathrm{d}z^{(k-1)}}\left[(z-z_0)^k f(z)\right]_{z=z_0}} \tag{7.2.202}$$

This latter relation is valid only for ONE isolated singularity (in case you forget!) and where k is equal at least to 1!

Mathematicians therefore define:

$$\boxed{\mathrm{Res}[f(z);z_0]=\lim_{z\rightarrow z_0}\frac{1}{(k-1)!}\frac{\mathrm{d}^{k-1}}{\mathrm{d}z^{k-1}}\left[(z-z_0)^k f(z)\right]_{z=z_0}} \tag{7.2.203}$$

as being the residue of the function $f(z)$ at the point being an isolated singularity of order k. Or respectively:

$$\mathrm{Res}[f(z);z_0]=\frac{1}{2\pi\mathrm{i}}\oint f(z)\mathrm{d}z \tag{7.2.204}$$

where the path integral is centered on z_0.

Now notice that the term on the right of the equality in the previous relation correspond to the coefficient c_{-1} of the Laurent series. Indeed:

$$c_n=\frac{1}{2\pi\mathrm{i}}\int\limits_{\gamma}\frac{f(z')}{(z'-z_0)^{n+1}}\mathrm{d}z'\Rightarrow c_{-1}=\frac{1}{2\pi\mathrm{i}}\int\limits_{\gamma}\frac{f(z')}{(z'-z_0)^{-1+1}}\mathrm{d}z'=\frac{1}{2\pi\mathrm{i}}\int\limits_{\gamma}f(z')\mathrm{d}z' \tag{7.2.205}$$

Therefore:

$$\boxed{\mathrm{Res}[f(z);z_0]=c_{-1}} \tag{7.2.206}$$

Remark

Therefore it comes that on an isolated singularity that can be eliminated, the residue is null because as we saw it before, the path integral rounding a domain without singularity is equal to zero!

To resume, the relation:

$$\oint f(z)\mathrm{d}z=2\pi\mathrm{i}\mathrm{Res}[f(z);z_0]=2\pi\mathrm{i}c_{-1} \tag{7.2.207}$$

is very interesting for the physicist... because this is a very elegant way for him to calculate the path integral of a non analytic function $f(z)$ having a unique isolated singularity and this just by knowing the order of its poles!

For example if a function $f(z)$ has only a pole of order 1, we have therefore:

$$\oint f(z)\mathrm{d}z=\frac{2\pi\mathrm{i}}{(1-1!)}\frac{\mathrm{d}^{1-1}}{\mathrm{d}z^{1-1}}\left[(z-z_0)^1f(z)\right]_{z=z_0}=2\pi\mathrm{i}\lim_{z\to z_0}\left[(z-z_0)^1f(z)\right] \tag{7.2.208}$$

and we replace therefore z_0 by the desired value in the parenthesis $(z-z_0)$ and after we calculate the limit between brackets!

Now to go further, let us remind that the contour of the path integral:

$$\oint f(z)\mathrm{d}z \tag{7.2.209}$$

and the curvilinear path of the integral:

$$c_n=\frac{1}{2\pi\mathrm{i}}\int\limits_{\gamma}\frac{f(z')}{(z'-z_0)^{n+1}}\mathrm{d}z' \tag{7.2.210}$$

are in fact combined (identical) and the coefficients c_n do not depend on z! The only constraint on the path is that is closed and in an analytical domain centered on one point.

So if we have several isolated singularities, surrounded by connected curvilinear paths as shown below on the complex plane of a function having a pole of order 3 (i.e. three non-removable singularities z_0, z_1, z_2):

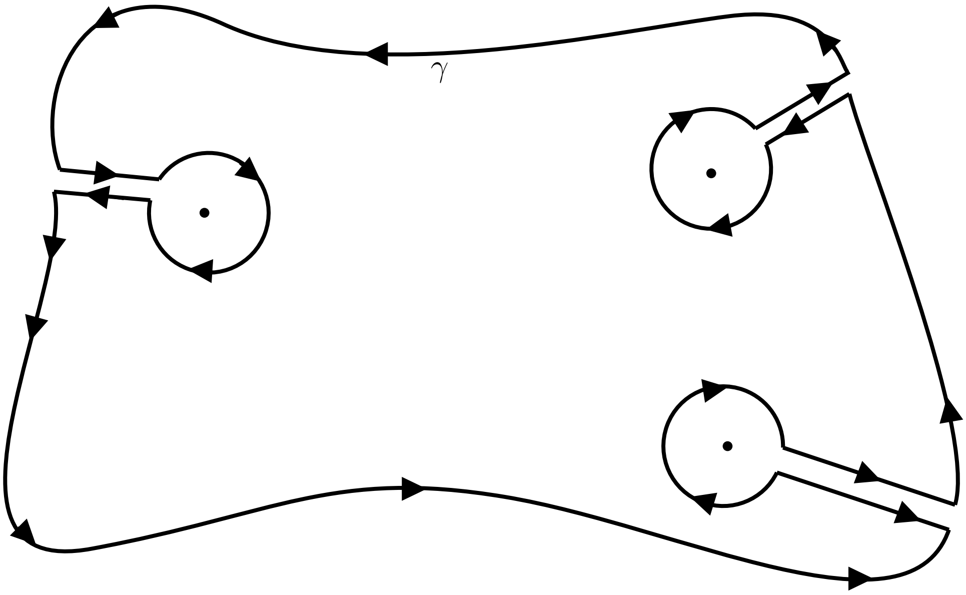


Figure 7.94: **Multiple isolated singularities surrounded by curvilinear paths**

then we have still only one closed curvilinear path but whose different isolated singularities are connected by cross which as we know: the paths that are opposed cancel themselves! And let us remind that the coefficients c_n are the same throughout on all the path since it is on an analytical domain.

We then have the generalized version of the residue theorem for a function f with n isolated singularities:

$$\boxed{\begin{gathered}\oint f(z)\mathrm{d}z=2\pi\mathrm{i}\sum_{i=1}^{n}\text{Res}\left[f(z);z_i\right]\\ \text{Res}\left[f(z);z_i\right]=\lim_{z\to z_0}\frac{1}{(k_i-1)!}\frac{\mathrm{d}^{k_i-1}}{\mathrm{d}z^{k_i-1}}\left[(z-z_i)^{k_i}f(z)\right]\end{gathered}} \tag{7.2.211}$$

with a rigorous approach that is specific to engineers ... who sometimes write this latter relation as following:

$$\oint f(z)\mathrm{d}z=2\pi\mathrm{i}\sum r_i \tag{7.2.212}$$

where r is therefore a residue. This is an important result in the field of solving differential equations associated with some inverse Laplace transforms (see section Functional Analysis page 2336). This intermediate result will give us the possibility to get an another one a little further of major importance for the section of Corpuscular Quantum Physics.

✍**Example:**

Let us take again our famous function:

$$f(z)=\frac{1}{1+z^2} \tag{7.2.213}$$

We know it has a pole of order 1 on $z_0=\mathrm{i}$ and a pole of order 1 on $z_0=-\mathrm{i}$. So if we take this time Laurent series with a path that surrounds the two singularities (and not only one) then we have a function with a pole of order 2.

It comes then for this particular case:

$$\oint f(z)\mathrm{d}z=2\pi\mathrm{i}\sum_{i=1}^{n}\text{Res}\left[f(z);z_i\right] \tag{7.2.214}$$

with n being equal to 2.

Then we have:

$$\begin{aligned}\text{Res}\left[f(z);z_1\right]=\text{Res}\left[f(z);\mathrm{i}\right]&=\lim_{z\to z_1}\frac{1}{(k_1-1)!}\frac{\mathrm{d}^{k_1-1}}{\mathrm{d}z^{k_1-1}}\left[(z-z_1)^{k_1}f(z)\right]\\&=\lim_{z\to\mathrm{i}}\frac{1}{(1-1)!}\frac{\mathrm{d}^{1-1}}{\mathrm{d}z^{1-1}}\left[(z-\mathrm{i})^1\frac{1}{z^2+1}\right]=\lim_{z\to\mathrm{i}}\left[(z-\mathrm{i})\frac{1}{(z+\mathrm{i})(z-\mathrm{i})}\right]\\&=\lim_{z\to\mathrm{i}}\frac{1}{z+\mathrm{i}}=-\frac{1}{2}\mathrm{i}\end{aligned} \tag{7.2.215}$$

and:

$$\begin{aligned}\text{Res}\left[f(z);z_2\right]=\text{Res}\left[f(z);-\mathrm{i}\right]&=\lim_{z\to z_2}\frac{1}{(k_2-1)!}\frac{\mathrm{d}^{k_2-1}}{\mathrm{d}z^{k_2-1}}\left[(z-z_2)^{k_2}f(z)\right]\\&=\lim_{z\to-\mathrm{i}}\frac{1}{(1-1)!}\frac{\mathrm{d}^{1-1}}{\mathrm{d}z^{1-1}}\left[(z-(-\mathrm{i}))^1\frac{1}{z^2+1}\right]\\&=\lim_{z\to-\mathrm{i}}\left[(z+\mathrm{i})\frac{1}{(z+\mathrm{i})(z-\mathrm{i})}\right]\\&=\lim_{z\to-\mathrm{i}}\frac{1}{z-\mathrm{i}}=\frac{1}{2}\mathrm{i}\end{aligned}\tag{7.2.216}$$

We can easily check this with Maple™ 4.00b:

```
>readlib(singular):
>singular(1/(1+z^2),z);
>readlib(residue):
>residue(1/(1+z^2),z=-I);
>residue(1/(1+z^2),z=I);
```

and therefore:

$$\oint\frac{1}{1+z^2}\mathrm{d}z=2\pi\mathrm{i}\sum_{i=1}^{2}\text{Res}\left[f(z);z_i\right]=0\tag{7.2.217}$$

In fact in this case, the residue theorem gives zero because the function has no poles to infinity which is true since in our example:

$$\lim_{z\to\pm\infty}|f(z)|\neq\infty\tag{7.2.218}$$

Physicists meanwhile say that... the force does make any work on this path ...!

7.2.8.1 Pole at infinity

We have say before that any function that did not have poles at infinity had therefore the sum of the residues of all the poles that are equal to zeros. This result is very important in physics and merits to be study!

It is almost trivial to recognize the number of poles... but to recognize the poles that are at infinity there are many times traps we can easily fall in.

Let us consider the expression $f(z)\mathrm{d}z$. If z is at the neighbourhood of the infinity then $1/z$ is near 0. Let us write:

$$t=\frac{1}{z}\tag{7.2.219}$$

Then we have:

$$f(z)\mathrm{d}z=-\frac{1}{t^2}f\left(\frac{1}{t}\right)\mathrm{d}t \tag{7.2.220}$$

Then the residue at infinity is such that:

$$\text{Res}\left[f(z);\infty\right]=\text{R}\left[g(t);0\right] \tag{7.2.221}$$

with:

$$g(t)=-\frac{1}{t^2}f\left(\frac{1}{t}\right) \tag{7.2.222}$$

Therefore with:

$$\boxed{\text{Res}\left[g(t);0\right]=\lim_{t\to 0}\frac{1}{(k-1)!}\frac{\mathrm{d}^{k-1}}{\mathrm{d}t^{k-1}}\left[(t-0)^k g(t)\right]=-\lim_{t\to 0}\frac{1}{(k-1)!}\frac{\mathrm{d}^{k-1}}{\mathrm{d}t^{k-1}}\left[t^k\frac{1}{t^2}f\left(\frac{1}{t}\right)\right]} \tag{7.2.223}$$

The latter relation we will be indispensable to us in the section of Corpuscular Quantum Physics Corpuscular to build the relativistic Sommerfeld model of hydrogenous atom because we will need to calculate a path integral with a pole.

Let's see an example with the function that accompanies us since the beginning of this section. That is to say:

$$f(z)=\frac{1}{z^2+1} \tag{7.2.224}$$

Therefore it comes:

$$g(t)=-\frac{1}{t^2}f\left(\frac{1}{t}\right)=-\frac{1}{t^2}\frac{t^2}{1+t^2}=-\frac{1}{1+t^2} \tag{7.2.225}$$

And we recognize immediately the initial function, in absolute value, and that has therefore no pole on 0. Therefore $f(z)$ has no pole at infinity.

~100 %

★★★☆☆

14 votes, 61.43%

7.3. Topology

TOPOLOGY is an extremely broad field of mathematics for which it is difficult to define precisely the object so the areas where it applies are varied (real line topology, graphs topology, differential topology, complex topology, symplectic topology, etc.).

We mainly make a distinction with:

- "General topology" that establishes the foundational aspects of topology and investigates properties of topological spaces and investigates concepts inherent to topological spaces. It includes point-set topology, which is the foundational topology used in all other branches (including topics like compactness and connectedness).
- "Algebraic topology" tries to measure degrees of connectivity using algebraic constructs such as homology and homotopy groups.
- "Differential topology" is the field dealing with differentiable functions on differentiable manifolds. It is closely related to differential geometry (see section of the same name in the chapter about Geometry page 2850) and together they make up the geometric theory of differentiable manifolds.
- "Geometric topology" primarily studies manifolds and their embeddings (placements) in other manifolds. A particularly active area is low dimensional topology, which studies manifolds of four or fewer dimensions. This includes knot theory, the study of mathematical knots.

Remark

A "manifold" is a higher dimensional analogue of a curve or a surface. A manifold of dimension n is a space, you can think of a collection of points, that locally looks like $\mathbb{R}^n$ for some integer n. All curves in the two dimensional plane that do not intersect locally look like part of a line. All smooth surfaces in $\mathbb{R}^3$, that is surfaces that do not have sharp kinks, edges (boundaries), or points all locally look like the two dimensional plane. Thus such surfaces are two dimensional manifolds. One should imagine being able to tear any small piece off the smooth surface, then being able to stretch it, push down any "hills" and push up any "valleys" to end up with a flat piece of the 2-plane. Therefore a manifold is a space that is locally Euclidean.

The picture to have in your mind when thinking about manifolds is the relation between a globe and a map. Small pieces of the globe can always be described by maps (pieces of the 2-plane). Moreover, the entire globe can be covered by a collection of maps: an atlas.

What we can say at first is that in its foundations Topology is very closely related to the set theory, the study of convergence of sequences and series, functional analysis, analysis complex, the differential and integral calculus, vector calculus and the geometry to mention only the most important cases that the reader can already found in this textbook.

The origin of Topology comes from the problems that laid to the progress of functional analysis in the rigorous study of continuous functions, their differentiability, their limits at a point (finite or note), the existence of extremums, etc. in higher-dimensional spaces (in fact, implicitly, the goal for topology is to create tools that easily allow to study the properties of functions in all dimensions). All these concepts, needed a rigorous mathematical definition of the intuitive idea of proximity, especially when doing operations on such functions.

We will try in this section to identify the basis of the structures that allow us to speak about limits and continuity and this only for curiosity as consultants in R&D and financial engineering we never saw a business application where the subjects below are absolutely necessary to develop a new business or solve a problem.

The majority of examples that we will take in this section will be in $\mathbb{R}$ (the $\mathbb{R}$ straight line to be more exact...) because it is the most used one by the engineers (and most of times the only one!) and the one we will have to use for the sections on Graph Theory, Statistics, Differential and Integral Calculus and also on Fractals. When we restrict our study of Topology on $\mathbb{R}$ we then speak of "Real Analysis".

7.3.1 General Topology

General topology is the branch of topology dealing with the basic set-theoretic definitions and constructions used in topology. It is the foundation of most other branches of topology, including differential topology, geometric topology, and algebraic topology.

The fundamental concepts in point-set topology are "continuity", "compactness", and "connectedness".

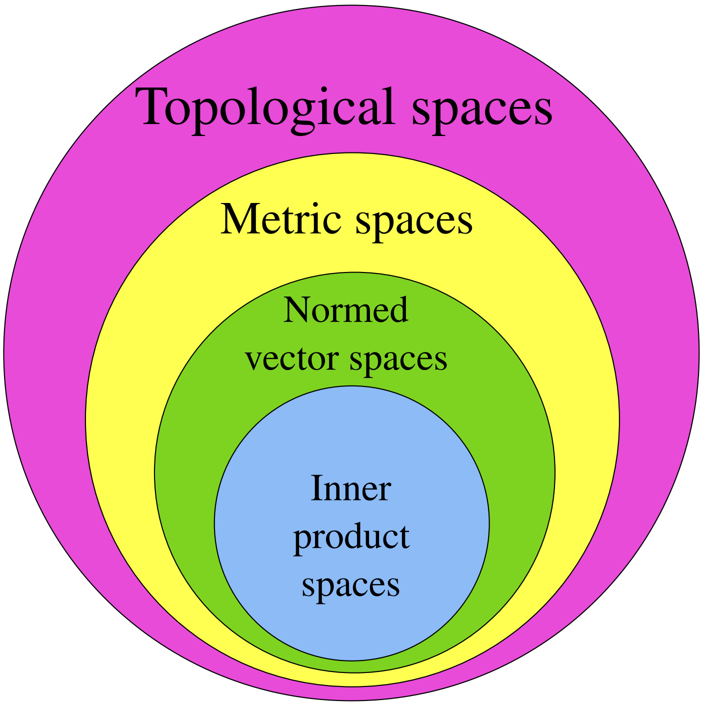
- Continuous functions take nearby points to nearby points.
- Compact sets are those that can be covered by finitely many sets of arbitrarily small size.
- Connected sets are sets that cannot be divided into two pieces that are far apart.

The words nearby, arbitrarily small, and far apart can all be made precise by using open sets. If we change the definition of open set, we change what continuous functions, compact sets, and connected sets are. Each choice of definition for open set is named a "topology". A set with a topology is named a "topological space".

"Metric spaces" are an important class of topological spaces where distances can be assigned a number named a "metric". Having a metric simplifies many proofs, and many of the most common topological spaces are metric spaces.

An "inner product" (see section Vector Calculus page 1829) induces a "norm" (see section Vector Calculus page 1818) and the norm induces a metric space.

Therefore we understand better what we will study in this section that can be summarized by the following figure:



7.3.1.1 Topological Spaces

Topological spaces form the conceptual foundation on which the concepts of limit, continuity or equivalence are defined.

The framework is general enough to be applied in many different situations: finite sets, discrete sets, geometry spaces, n dimensional numerical spaces and most complex functional areas. These concepts appear in almost all branches of mathematics, they are therefore central to the modern view of mathematics.

Definition (#7.3.290): Consider a non-empty set X (the length of a plastic ruler for example). A "topology $\mathcal{T}$" or "topological space $(x,\mathcal{T})$" on X is a family $\mathcal{T}$ of parts of X (of length of our rule...) named "open V" (as the open intervals seen in the section of Functional Analysis) such that the following axioms are true:

A1. The empty set $\varnothing$ and X are considered as open V and must belong to the family of the topology $\mathcal{T}$ (these only both open sets define what we name the "trivial topology" that is the most minimal one satisfying all the axioms):

$$\varnothing \in \mathcal{T} \quad \text{and} \quad X \in \mathcal{T} \tag{7.3.1}$$

In other words, if we imagine our plastic ruler, the measure zero (strictly speaking: the empty set) must belong to the topology defined on the ruler and the ruler itself (seen as a subset).

A2. Any finite intersection of open of $\mathcal{T}$ will be an open of $\mathcal{T}$:

$$V_1, V_2, \ldots, V_n \in \mathcal{T} \quad \text{implies} \quad \bigcap_{i=1}^{n} V_i \in \mathcal{T} \tag{7.3.2}$$

A3. Any union of open of $\mathcal{T}$ will be an open of $\mathcal{T}$:

$$V_1, V_2, \ldots, V_n \in \mathcal{T} \quad \text{implies} \quad \bigcup_{i=1}^{n} V_i \in \mathcal{T} \tag{7.3.3}$$

Remarks

R1. Mathematicians frequently note by O the family of open sets and by F the family of closed sets. Convention we will not follow in this textbook.

R2. The close sets of a topology are complementary of open sets. Therefore, the family of close sets contains among other X and the empty set $\varnothing$...

R3. There is no difference between part and subset of a set.

A4. The couple $(X, \mathcal{T})$ is a "Hausdorff space" or "separate space" if moreover the property named "Hausdorff axiom" is verified:

$$\forall p, q \in X \quad \text{with} \quad p \neq q, \exists V_p, V_q \in F \quad \text{such that} \quad p \in V_p, q \in V_q \quad \text{and} \quad V_p \cap V_q = 0 \tag{7.3.4}$$

Remarks

R1. A well known example of topological space is $\mathbb{R}$ provided with the set F generated by the open intervals (by the union law), that is to say the intervals of the type $]a, b[$.

R2. We will see a very concrete and beautiful + nice application of Hausdorff spaces in our study of fractals in the chapter Theoretical Computing.

Definitions (#7.3.291):

D1. If we denote by (X, V) a topological space, V designating the open sets of X, a "base", in the topological sense, of (X, V) is a part B of V such that any open set of V is an union of open sets of B (this is the same idea as vector spaces but in fact applied to sets ... nothing bad and difficult! If you want an example see the section of Measure Theory).

D2. In topology a subset A of a topological space X is named "dense" (in X) if for every point x in X either belongs to A or is a limit point of A. Informally, for every point in X, the point is either in A or arbitrarily "close" to a member of A. For instance, every real number is either a rational number or has one arbitrarily close to it. Therefore $\mathbb{Q}$ is dense in $\mathbb{R}$.

7.3.2 Metric Space and Distance

Definition (#7.3.292): A "metric space" denoted by (X,d) or sometimes X_d (or even sometimes just X if the type of distance d cannot not be confused) is by definition a set X with provided with an application:

$$d : X \times X \to \mathbb{R} \tag{7.3.5}$$

named "distance". If the distance satisfies the following "metric axioms":

A1. Positivity:

$$\forall x \neq y \in X \quad d(x,y) \geq 0 \tag{7.3.6}$$

A2. Separation:

$$\forall x,y \in X \quad d(x,y) = 0 \Leftrightarrow x = y \tag{7.3.7}$$

A3. Triangular inequality:

$$\forall x,y,z \in X \quad d(x,z) \leq d(x,y) + d(y,z) \tag{7.3.8}$$

A4. Symmetry:

$$d(x,y) = d(y,x) \tag{7.3.9}$$

we say then that the distance is a "metric". If a distance respect all axioms above excepted the triangular inequality we say then that the distance is a "dissimilarity measure".

The triangle inequality is necessary to define a metric because it ensures that the distance between two points is never shorter than the distance between them via a third point. This property is essential for many of the important concepts in metric spaces, such as convergence and continuity.

Remarks

R1. Some readers will probably see immediately that some of these properties have already been seen in other sections of this textbook during our study of distances between functional points and during our study of norms (like the riangle inequality proved in the section of Vector Calculus at page 1828).

R2. Some authors omit the axiom A1 which is strictly correct as it trivially follows from A3.

The "distance function" of $\forall x,y \in X$ is thus usually denoted in the more possible sense in mathematics (at least as far as we know):

$$d_{\text{name/symbol of distance}}(x,y) \tag{7.3.10}$$

we will see three examples much more further below with a schema.

Definition (#7.3.293): If we do not impose the axiom A2, we say that d is a "semi-distance" on X and if we allow a semi-distance d to take the value $+\infty$, we prefer to say that d is a "gap".

Remarks

R1. If a distance d satisfies the property (named "strong triangle inequality" or "ultrametric inequality"):

$$\forall x,y,z \in X \quad d(x,y) \leq \max\{d(x,y),d(y,z)\} \tag{7.3.11}$$

that is more restrictive than the triangle inequality in some spaces, we say that d is "ultrametric".

An example of ultrametric distance is the family tree (...):

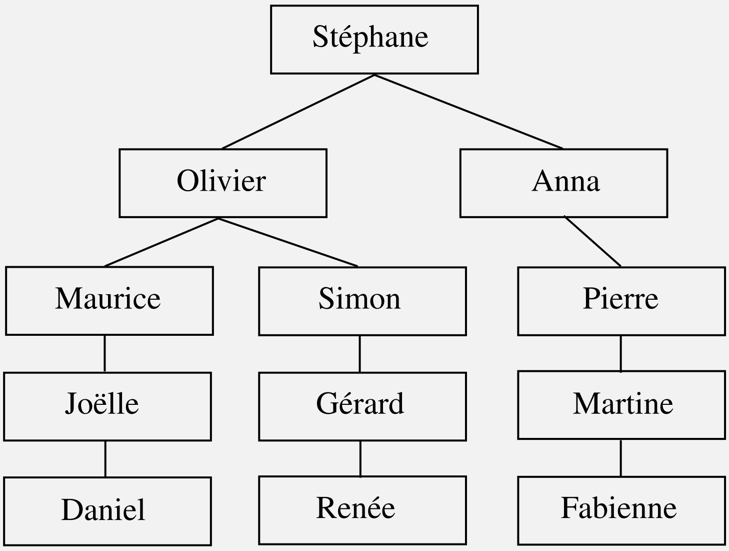


Figure 7.95: **Example of ultrametric distance with an orgchart**

We have the following distances:

$$\begin{aligned} d(\text{Daniel},\text{Stéphane}) &= 4\\ d(\text{Renée},\text{Stéphane}) &= 4\\ d(\text{Renée},\text{Fabienne}) &= 8\\ d(\text{Daniel},\text{Fabienne}) &= 8\\ d(\text{Daniel},\text{Renée}) &= 6 \end{aligned} \tag{7.3.12}$$

We note that the distances above do not add up, but we have by cons:

$$d(x,z) \leq \max\{d(x,y),d(y,z)\} \tag{7.3.13}$$

Therefore:

$$\begin{aligned} d(\text{Daniel},\text{Fabienne}) &\leq \max\{d(\text{Daniel},\text{Renée}),d(\text{Renée},\text{Fabienne})\}\\ 8 &\leq \max\{6,8\} \end{aligned} \tag{7.3.14}$$

Notice that the strong triangle inequality implies the triangular inequality:

$$d(x,z) = \max(d(x,y),d(y,z)) \leq d(x,y)+d(x,z) \tag{7.3.15}$$

because of the fact that the maximum of two numbers is less than or equal to their sum if both are greater or equal to zero.

R2. Let (X,d) be a metric space and consider $F=\varnothing$ a part of the set E. The metric space (F,δ) where δ denotes the restriction $d_{F\times F}$ of d is named "metric subspace" of (X,d) (we should check that the distance d is equivalent to the distance δ). In this case, we also say that F is provided with the distance induced by this of X. We therefore simply note d the induced distance.

Let us give now some examples:

✍**Examples:**

E1. If we take for X the plane, or the three-dimensional space of Euclidean geometry and a unit of length, the "distance" in the usual sense is a distance within the meaning of the 4 axioms mentioned above. In these spaces, the three points A,B,C satisfy as we have proved it in the section Vector Calculus:

$$d(A,B)\leq d(A,C)+d(C,B) \quad d(A,B)\geq |d(A,C)-d(C,B)| \tag{7.3.16}$$

with other inequality obtained by circular permutation of A,B,C. These inequalities are well known, for example between the side lengths of a triangle.

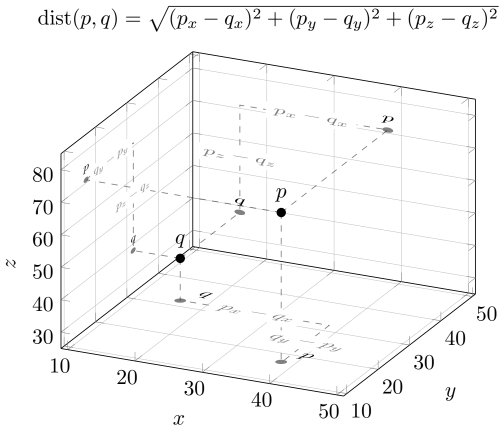
E2. If we take $X=\mathbb{R}^n$, $n\in\mathbb{N}\geq 1$ and that we equip $\mathbb{R}^n$ of an Euclidean vector space structure (and not non-Euclidean!) and we take two points:

$$x=(x_1,x_2,\ldots,x_n) \qquad y=(y_1,y_2,\ldots,y_n) \tag{7.3.17}$$

in $\mathbb{R}^n$, the distance is then given by (we have already proved this in the sections of Vector Calculus at page 1818):

$$d(x,y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\ldots+(x_n-y_n)^2}=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{7.3.18}$$

This distance is often represented in three dimensions as following:



This latter distance satisfies the four axioms of distance and we name it the "Euclidean distance" and is often denoted L_2. We can take (it is an interesting property for the general culture), any relation of the form:

$$d(x,y)=\sqrt[p]{(x_i-y_i)^p} \tag{7.3.19}$$

is also a distance in $\mathbb{R}^n$ (without proof) named "p-norm". In the particular case with $n=1$, we have of course:

$$d(x,y)=|x-y|=\max(x-y,y-x) \tag{7.3.20}$$

This is the usual distance on $\mathbb{R}$ and is often denoted L_2.

Mathematicians are even stronger by generalizing ever more (the proof has little interest for now in this textbook) the prior-previous relation (taking into account the definition of the distance) in the form:

$$\forall p\in[1,+\infty[\subset\mathbb{R}: \quad \delta_p(x,y)=\sqrt[p]{\sum_{i=1}^{n}(d(x,y))^n} \tag{7.3.21}$$

which is named "Hölder distance". Also sometimes denoted for $L_p:\mathbb{R}^n\mapsto\mathbb{R}$:

$$\boxed{\forall p\in[1,+\infty[\subset\mathbb{R}: \quad L_p(x)=||x||_p=\left(\sum_{i=1}^{n}|x_i|^p\right)^{1/p}} \tag{7.3.22}$$

Remark

Following the intervention of a reader we would like to point out that strictly speaking the above inclusion should be noted $[1,+\infty[\subset\overline{\mathbb{R}}$ where $\overline{\mathbb{R}}$ is the achieved line (also valid for precision for the Minkowski inequality below).

As for the triangle inequality given for recall by (see section Vector Calculus page 1828):

$$\vec{x},\vec{y}\in\mathbb{R}^3: \quad \|\vec{x}+\vec{y}\|\leq\|\vec{x}\|+\|\vec{y}\|\Leftrightarrow\sqrt[2]{\sum_{i=1}^{3}(x_i+y_i)^2}\leq\sqrt[2]{\sum_{i=1}^{3}x_i^2}+\sqrt[2]{\sum_{i=1}^{3}y_i^2} \tag{7.3.23}$$

the generalization, by the verification of the existence of the Hölder distance, gives us the true "Minkowski inequality":

$$\forall x,y\in\mathbb{R},p\in[1,+\infty[\subset\mathbb{R}: \quad \sqrt[p]{\sum_{i=1}^{n}(x_i+y_i)^p}\leq\sqrt[p]{\sum_{i=1}^{n}x_i^p}+\sqrt[p]{\sum_{i=1}^{n}y_i^p} \tag{7.3.24}$$

Let us continue with our examples:

✍**Examples:**

E3. If we take $X=\mathbb{C}$ we will consider the distance:

$$d(z,z')=|z-z'| \tag{7.3.25}$$

Therefore if $z=a+\mathrm{i}b=(a,b)$ and $z'=a'+\mathrm{i}b'=(a',b')$ we have the module that the same manner as norm in $\mathbb{R}^2$, forms a distance:

$$d(z,z')=\sqrt{(a-a')^2+(b-b')^2} \tag{7.3.26}$$

E4. Let us consider $E=\varnothing$ an arbitrary set. Let us write:

$$\delta(x,y)=1 \quad \text{if} \quad x\neq y \quad \text{and} \quad \delta(x,y)=0 \quad \text{if} \quad x=y \tag{7.3.27}$$

It is quite check that this distance satisfies the four axioms and that is furthermore an ultrametric distance. This distance is named "discreet distance" and the reader will notice that, by analogy, we choosed to express this distance by the Dirac symbol δ (this is not innocent !!) rather than the traditional d.

There are plenty of other distances used in practice that satisfies the four axioms of a distance. The technique used to measure distances depends on a particular situation you are working on. For instance, in some areas, the euclidean distance can be optimal and useful for computing distances. Other applications require a more sophisticated approach for calculating distances between points or observations like the cosine distance. As an image is many times worth one thousand words here is an illustration of the most well knows distances that respect the four axioms:

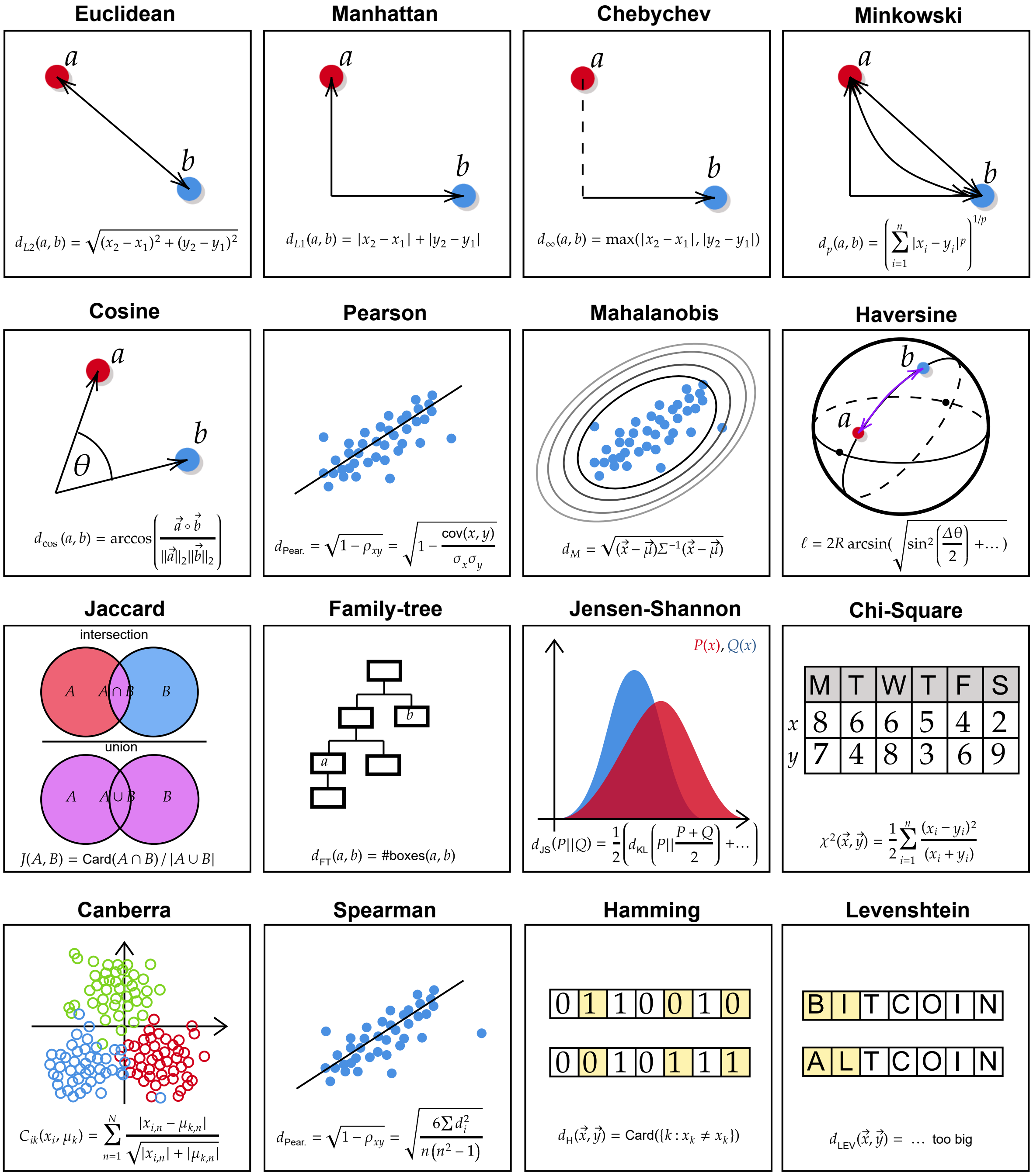

Euclidean
a
b
$d_{L2}(a,b)=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$
Manhattan
a
b
$d_{L1}(a,b)=|x_2-x_1|+|y_2-y_1|$
Chebychev
a
b
$d_{\infty}(a,b)=\max(|x_2-x_1|,|y_2-y_1|)$
Minkowski
a
b
$d_p(a,b)=\left(\sum_{i=1}^{n}|x_i-y_i|^p\right)^{1/p}$
Cosine
a
b
θ
$d_{\cos}(a,b)=\arccos\left(\frac{\vec{a}\circ\vec{b}}{\|\vec{a}\|_2\|\vec{b}\|_2}\right)$
Pearson
$d_{\text{Pear.}}=\sqrt{1-\rho_{xy}}=\sqrt{1-\frac{\text{cov}(x,y)}{\sigma_x\sigma_y}}$
Mahalanobis
$d_M=\sqrt{(\vec{x}-\vec{\mu})\Sigma^{-1}(\vec{x}-\vec{\mu})}$
Haversine
b
a
$\ell=2R\arcsin(\sqrt{\sin^2\left(\frac{\Delta\theta}{2}\right)+\ldots})$
Jaccard
intersection
A
$A\cap B$
B
union
A
$A\cup B$
B
$J(A,B)=\text{Card}(A\cap B)/|A\cup B|$
Family-tree
b
a
$d_{\text{FT}}(a,b)=\#\text{boxes}(a,b)$
Jensen-Shannon
$P(x)$, $Q(x)$
$d_{\text{JS}}(P\|Q)=\frac{1}{2}\left(d_{\text{KL}}\left(P\|\frac{P+Q}{2}\right)+\ldots\right.$
Chi-Square
M T W T F S
x 8 6 6 5 4 2
y 7 4 8 3 6 9
$\chi^2(\vec{x},\vec{y})=\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-y_i)^2}{(x_i+y_i)}$
Canberra
$C_{ik}(x_i,\mu_k)=\sum_{n=1}^{N}\frac{|x_{i,n}-\mu_{k,n}|}{\sqrt{|x_{i,n}|+|\mu_{k,n}|}}$
Spearman
$d_{\text{Pear.}}=\sqrt{1-\rho_{xy}}=\sqrt{\frac{6\sum d_i^2}{n(n^2-1)}}$
Hamming
0 1 1 0 0 1 0
0 0 1 0 1 1 1
$d_{\text{H}}(\vec{x},\vec{y})=\text{Card}(\{k:x_k\neq x_k\})$
Levenshtein
B I T C O I N
A L T C O I N
$d_{\text{LEV}}(\vec{x},\vec{y})=\ldots$ too big

Figure 7.96: Common types of distances

We have already proven that the euclidean distance satisfies the four axioms at page 1818. We haven't done the proof that the four axioms are respected for all other distances visible in the figure above year in this textbook (the proof of the triangle inequality is quite boring to write most of the time).

💣Caution!!! Some computer scientists claim that the Sørensen–Dice distance and the Squared Euclidean distance are distances... but the latter doesn't respect the triangle equality as far as we know. Moreover there are also various definitions of the cosine distance and Pearson's distance and some of them doesn't respect the triangular equality. So be careful!

7.3.2.1 Equivalent Distances

Sometimes two different distances d and δ on the same set E are quite similar so that the related metric spaces $(E,d),(E,\delta)$ have the same properties for certain mathematical objects defined by d on one hand, and by δ on the hand. There are several concepts of equivalences for example first (before the others that require mathematical tools that we have not yet defined):

Definition (#7.3.294): Let d and δ be two distances on the same set E, d and δ are named "equivalent distances" if there are two real constants $c>0,C>0$ such that:

$$\forall x,y\in E:\quad c\cdot d(x,y)\leq\delta(x,y)\leq C\cdot d(x,y)\tag{7.3.28}$$

Therefore:

$$c\leq\frac{\delta(x,y)}{d(x,y)}\leq C\tag{7.3.29}$$

with $c\leq C$. We note this equivalence by:

$$d\sim\delta\tag{7.3.30}$$

The advantage of this definition is the following: if we have convergence for one of the metric, then we have convergence for the other too. More clearly:

$$d(x_n,x)\rightarrow 0\Leftrightarrow\delta(x_n,x)\rightarrow 0\tag{7.3.31}$$

verbatim:

$$x_n\overset{d}{\rightarrow}x\Leftrightarrow x_n\overset{\delta}{\rightarrow}x\tag{7.3.32}$$

7.3.2.2 Lipschitz Functions

With respect to the above definitions, we can now assign some additional properties to functions such as we had define in the section of Set Theory or Functional Analysis and analysed (in part...) in the section of Differential and Integral Calculus. The idea is also mainly to build a set of tools enabling the study of differential properties of non differentiable functions.

Let (E,d) and (F,δ) be metric spaces, and $f:E\rightarrow F$ a function. We define the following properties:

P1. We say that f is an "isometry" if (it is rather intuitive ...!):

$$\forall x,y\in E:\quad\delta[f(x),f(y)]=d(x,y)\tag{7.3.33}$$

✍**Exemple:**

On $]0,+\infty[$, let $\delta(x,y)=|\log(x)-\log(y)|$ and on $\mathbb{R}$ let $d(x,y)=|x-y|$. Then:

$$\begin{array}{ccc}(]0,+\infty[,\delta) & \longrightarrow & (\mathbb{R},d)\\ x & \mapsto & \log(x)\end{array}$$

is an isometry!

P2. If we take the usual distance, the L-Lipschitz or "Lipschits function of order L" is then defined on a given interval by:

$$\forall x,y\in E: \quad |f(x)-f(y)|\leq L|x-y| \tag{7.3.34}$$

that we can also write:

$$\forall x,y\in E: \quad \frac{|f(x)-f(y)|}{|x-y|}\leq L \tag{7.3.35}$$

or what remains the same: all line drawn between two arbitrary points of the graph must have a bounded and finite slope coefficient (derivative) between L and $-L$.

Any such L is referred to as a "Lipschitz constant" for the function f. As L can be define on intervals (not necessarily the whole domain of definition) smallest constant is sometimes named the "best Lipschitz constant".

Otherwise, one can equivalently define a function to be Lipschitz continuous if and only if there exists a constant L such that, for all $x\neq y$:

$$\frac{\mathrm{d}(f(x),f(y))}{\mathrm{d}(x,y)}\leq L \tag{7.3.36}$$

For real-valued functions of several real variables, this holds if and only if the absolute value of the slopes of all secant lines are bounded by k.

Therefore all Lipschitz must be continuous and any function f that has a bounded L value is more restrictive than just simply being continuous!

✍**Examples:**

E1. The function $f(x)=\sin(x)$ is 1-Lipschitz as the derivative of the cosine is between -1 and 1.

E2. The function $f(x)=x^2$ is locally Lipschitz as for any interval closed and finite interval we can find a bound L but is not globally Lipschitz as when $x\rightarrow\pm\infty$ then the derivative has also no bounds.

E3. The function $f(x)=|x|$ has no derivatives on $x=0$ in the ordinary sense. But its derivative in the Lipschitz sense on $x=0$ is given by the a closed interval denoted $\partial_L f(0)=[-1,1]$ given by the bound of L as the ordinary derivatives is less than or equal to L in absolute value!

This last example show us that the notion of a local minimum of a function $f(x)$ in the ordinary sense is generalized with Lipschitz condition. As we can now simply define the condition of local minimum of a non-smooth function $f(x)$ as:

$$0\in\partial_L f(x) \tag{7.3.37}$$

rather than in the ordinary sense (more restrictive):

$$f'=0 \tag{7.3.38}$$

Schematically for a Lipschitz continuous function, there is a double cone (shown in white) whose vertex can be translated along the graph, so that the graph always remains entirely outside the cone.

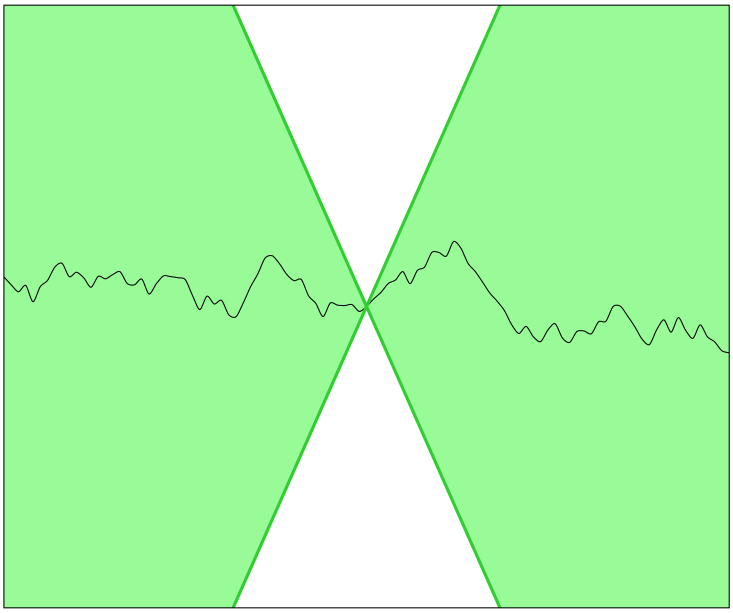


Figure 7.97: **Example of Lipschitz function (source: Wikipedia, author: Robert Zhang)**

Intuitively, a Lipschitz continuous function is therefore limited in how fast it can change: there exists a definite real number such that, for every pair of points on the graph of this function, the absolute value of the slope of the line connecting them is not greater than this real number; this bound is named a "Lipschitz constant" of the function (or "modulus of uniform continuity"). For instance, every function that has bounded first derivatives is Lipschitz.

P3. If $L=1$, we say that the function $f(x)$ is a "short map". If $-1<L<1$, we say that $f(x)$ is "strictly contracting".

P4. We say that two metric spaces are "isometric spaces" if there is a surjective isometry of one over the other (which is quite natural in geometry ...).

Remarks

R1. An isometry is always injective as by one of the axioms of distance:

$$f(x)=f(y)\Rightarrow \delta[f(x),f(y)]=0=d(u,v)\Rightarrow u=v \tag{7.3.39}$$

but in general it is not necessarily surjective (for a reminder of these concepts see page 406).

R2. If (E,d) and (F,δ) are isometric, of the point of view of the theory of metric spaces they are not discernible, as all their properties are the same, but their elements can be of very different nature (sequences in one and functions in the other).

7.3.2.3 Continuity and Uniform Continuity

As we already see it in the section of Functional Analysis (see page 2226), a continuous function is, roughly speaking, a function for which small changes in the input result in small changes in the output and that permits the analysis of limits. Otherwise, a function is said to be a discontinuous function. Formally it was defined by:

$$\forall x_0 \in E \forall \varepsilon > 0 \exists \delta > 0 \forall x \in E: \quad |x-x_0| < \delta \Rightarrow |f(x)-f(x_0)| < \varepsilon \tag{7.3.40}$$

In other words remember that this mean that a function is continuous if for every point x_0 in the domain E, we can make the images of that point ($f(x_0)$) and another point ($f(x)$) arbitrarily close (of a distance ε) if we move the other point (x) close enough (distance δ) to our given point.

Hence it is not continuous if:

$$\exists x_0 \in E \exists \varepsilon > 0 \forall \delta > 0 \exists x \in E: \quad |x-x_0| < \delta \Rightarrow |f(x)-f(x_0)| \geq \varepsilon \tag{7.3.41}$$

The previous definition is not so good for mathematicians as:

1. It makes usage of a special case of distance (the absolute value)
2. It defines continuity only for a given point x_0

It is therefore more common to generalize by writing:

$$\boxed{\forall x_0 \in E \forall \varepsilon > 0 \exists \lambda > 0 \forall x \in E: \quad d(x,x_0) < \lambda \Rightarrow \delta(f(x),f(x_0)) < \varepsilon} \tag{7.3.42}$$

Now let us state a more restrictive definition!

Definition (#7.3.295): A function $f(x)$ is "uniformly continuous" if it satisfies:

$$\boxed{\forall \varepsilon > 0 \exists \lambda > 0 \forall x_0, x \in E: \quad d(x_0,x) < \lambda \Rightarrow \delta[f(x),f(x_0)] \leq \varepsilon} \tag{7.3.43}$$

with $\lambda = \varepsilon / L$ and $L \neq 0$. In other words, if we can bring two points as close as we want in a space, so can we in the other way (which ensures somehow the derivation!).

Hence it is not uniformly continuous if:

$$\exists \varepsilon > 0 \forall \lambda > 0 \exists x_0, x \in E: \quad d(x_0,x) < \lambda \text{ and } \delta[f(x),f(x_0)] \geq \varepsilon \tag{7.3.44}$$

If we compare the two relations:

$$\forall x_0 \in E \forall \varepsilon > 0 \exists \lambda > 0 \forall x \in E: \quad d(x,x_0) < \lambda \Rightarrow \delta(f(x),f(x_0)) < \varepsilon \tag{7.3.45}$$

$$\forall \varepsilon > 0, \exists \lambda > 0, \forall x_0, x \in E: \quad d(x_0,x) < \lambda \Rightarrow \delta[f(x),f(x_0)] \geq \varepsilon \tag{7.3.46}$$

the only difference is the order of the quantifiers. Indeed, for something to be continuous, you can check "one x at a time", so for each x, you pick a ε and then find some $\varepsilon > 0$ that depends on both x and ε so that $|f(x)-f(x_0)| < \varepsilon$ if $|x-x_0| < \delta$. If we want uniform continuity, we need to pick first a ε, then find a δ which is good for ALL the x values we might have.

As the previous definition are not quite easy for everybody let us see the engineer version of these two definitions:

Definitions (#7.3.296):

D1. A function $f(x): E \rightarrow \mathbb{R}$ is "continuous" at a point $x_0 \in A$ if, for all $\varepsilon > 0$, there exists a $\delta > 0$ such that whenever $|x - x_0| < \delta$ (and $x \in E$) it follows that $|f(x) - f(x_0)| < \varepsilon$.

D2. A function $f(x): E \rightarrow \mathbb{R}$ is "uniformly continuous" on A if, for all $\varepsilon > 0$, there exists a $\delta > 0$ such that whenever $|x - x'| < \delta$ (and $(x, x') \in E$) it follows that $|f(x) - f(x')| < \varepsilon$.

Therefore we see better the difference: continuity is defined at a point x_0, whereas uniform continuity is defined on a set E. Roughly speaking, uniform continuity requires the existence of a single $\delta > 0$ that works for the whole set E, and not near the single point x_0.

The concept is illustrated below:

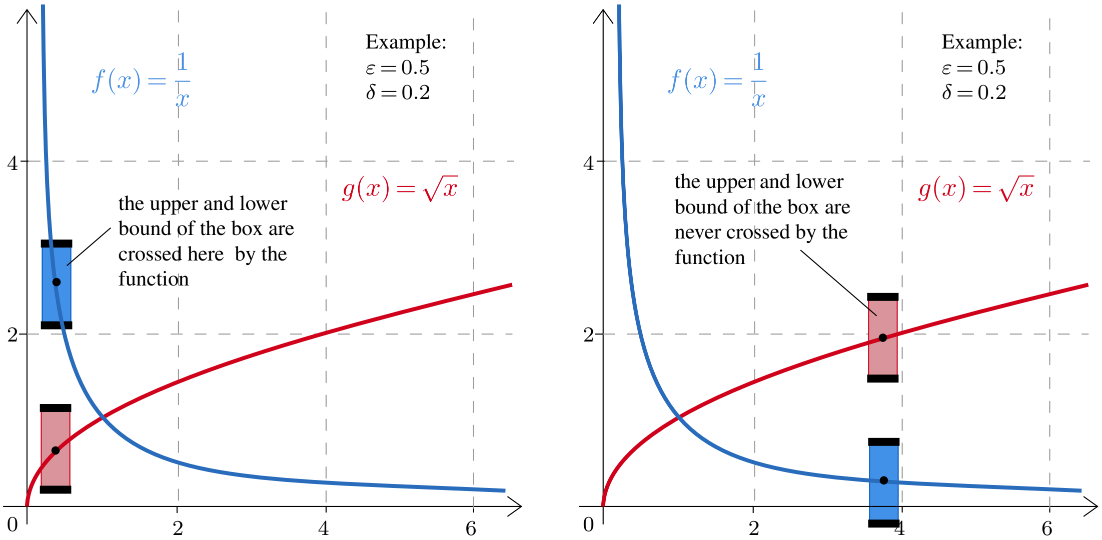
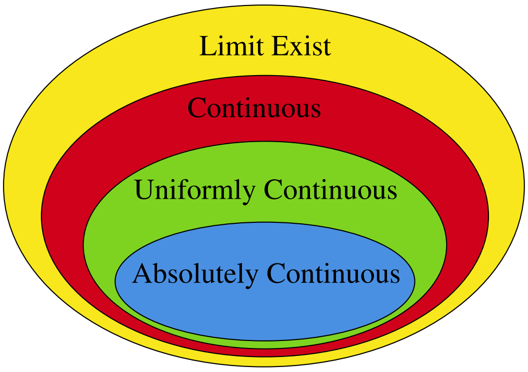


Figure 7.98: **Uniform continuity illustrated (source Wikpedia, author: Eskil Simon Kanne Wadsholt)**

At some point the graph of $f(x) = 1/x$ will penetrate the top and/or bottom of a (here blue) window with height $2\varepsilon \in \mathbb{R}_{>0}$ and width $2\delta \in \mathbb{R}_{>0}$ while (f) ranging over an x-interval smaller than δ, as that window's center moves over the graph in the direction of $x = 0$. If there existed a window whereof top and/or bottom is never penetrated by this graph as the window moves over the graph in the direction of $x = 0$, then its width would be infinitesimally small, meaning that $f(x)$ is not uniformly continuous. The function $g(x) = \sqrt{x}$, on the other hand, is uniformly continuous.

From this definition we see that any uniformly continuous function is continuous but the reciprocity is not true (any uniformly continuous function is not necessarily continuous):



If the chosen distance is known (for example the absolute value for scalar functions such that $d = |\cdot|$ and $\delta = |\cdot|$) the previous definition notation changes obviously a little bit:

$$\forall \varepsilon > 0, \exists \delta > 0, \forall x_0, x \in E : \quad |x_0 - x| \leq \delta \Rightarrow |f(x_0) - f(x)| \leq \varepsilon \tag{7.3.47}$$

✍**Example:**

The function $f(x) = x^2$ is continuous but not uniformly continuous on the interval $E = [0, +\infty[$.

We prove first that our function $f(x)$ is continuous on E. Remember first that:

$$\forall x_0 \in E \forall \varepsilon > 0 \exists \lambda > 0 \forall x \in E : \quad d(x, x_0) < \lambda \Rightarrow \delta(f(x), f(x_0)) < \varepsilon \tag{7.3.48}$$

In our case we can therefore write and check that:

$$\forall \varepsilon > 0 \exists \delta > 0 \forall x_0, x \in E : \quad |x - x_0| < \delta \Rightarrow |x^2 - x_0^2| \leq \varepsilon \tag{7.3.49}$$

Let us choose $x_0 = a - 1$ with $a > 1$ and $\delta = \min(1, \varepsilon/2a)$ (note that δ depends on x_0 since a does). Choose $x \in S$. Assume $|x - x_0| < \delta$. Then $|x - x_0| < 1$ so $x < x_0 + 1$ so $x, x_0 < a$ so:

$$|x^2 - x_0^2| = (x + x_0)|x_0| \leq 2a|x - x_0| < 2a\delta \leq 2a\frac{\varepsilon}{2a} = \varepsilon \tag{7.3.50}$$

We prove now that $f(x)$ is not uniformly continuous on E, i.e.:

$$\exists \varepsilon > 0 \forall \delta > 0 \exists x_0 \in E \exists x \in S : \quad |x - x_0| < \delta \text{ and } |x^2 - x_0^2| \geq \varepsilon \tag{7.3.51}$$

Let $\varepsilon = 1$. Choose $\delta > 0$. Let $x_0 = 1/\delta$ and $x = x_0 + \delta/2$. Then $|x - x_0| = \delta/2 < \delta$ but:

$$|x^2 - x_0^2| = \left| \left(\frac{1}{\delta} + \frac{\delta}{2}\right)^2 - \left(\frac{1}{\delta}\right)^2 \right| = 1 + \frac{\delta^2}{4} > 1 = \varepsilon \tag{7.3.52}$$

as required.

7.3.3 Opened and Closed Set

Definition (#7.3.297): Consider a set E with a distance d. A subset U of E is named "open subset" if, for each element of U, there is a non-null distance r for which all the elements of E whose distance from this element is less than or equal to r, belong to U, which gives in mathematical language:

$$U \text{ open subset of } E \Leftrightarrow \forall x \in U, \exists r \in \mathbb{R}^{+} | \forall y \in E: \quad d(x,y) \leq r \Rightarrow y \in U \tag{7.3.53}$$

In topology, an open subset is then only an abstract concept generalizing the idea of an open interval in the real line.

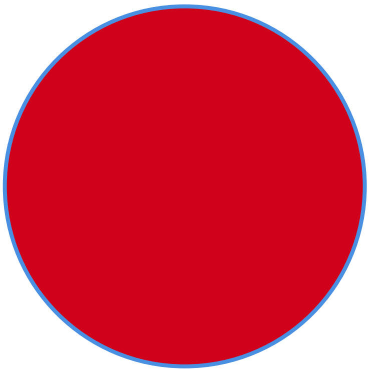
Remark

For recall, the symbol "|" means in this context: satisfies the property...

In practice, however, open sets are usually chosen to be similar to the open intervals of the real line. The notion of an open set provides a fundamental way to speak of nearness of points in a topological space, without explicitly having a concept of distance defined.

✍**Example:**

The points (x,y) satisfying $x^2+y^2=r^2$ are colored blue. The points (x,y) satisfying $x^2+y^2<r^2$ are colored red. The red points form an open subset of the plane $\mathbb{R}^2$. The blue points form a boundary set. The union of the red and blue points is a closed set.



This definition may perhaps seem complicated but in fact, its real meaning is simpler than it seems. In fact, according to this definition, an open set in a topological space is nothing more than a set of contiguous points and without borders.

The lack of border comes from the condition $r \neq 0$. Indeed, by reductio ad absurdum, if an open set U had an edge, then for each point on it (the edge) it would still be possible to find a point not belonging to U as close as we want from it. It follows that the distance r becomes necessary therefore zero.

Definitions (#7.3.298):

D1. A "closed subset" is an "open with edge".

D2. A "neighbourhood" of a point E is a subset of E containing an open subset containing this point.

The definition of an open set can be simplified by introducing an additional concept, that of "open ball":

7.3.3.1 Balls

Given x an element of E:

Definition (#7.3.299): An "open ball of center x and radius $r>0$" or "metric ball of radius r centered at x without border" is the subset of all the points of E whose the distance x is less than r, that we write in general:

$$_oB_x^r(E)=\{y\in E|d(x,y)<r\} \tag{7.3.54}$$

An open set can also be defined as a set for which it is possible to define an open ball at each point.

Typically in the real plane where d is the euclidean distance:

$$_oB_x^r(\mathbb{R}^2)=\left\{y\in\mathbb{R}^2|d(x,y)<r\right\} \tag{7.3.55}$$

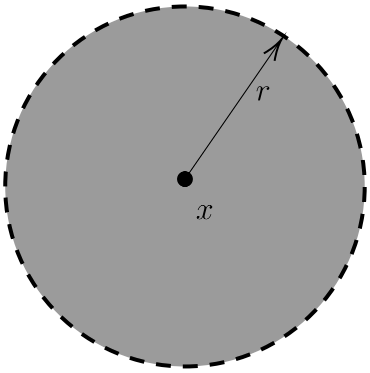


Figure 7.99: An open ball of radius r, centered at the point x

An open set can also be defined as a set for which it is possible to define an open ball at each point.

Remarks

R1. The open such defined, form what we name an "induced topology" by the distance d or also a "metric topology".

R2. We name an "open cover" U of E, a set of open of E whose union is equal to E. In other words: A collection of open sets that collectively cover another set.

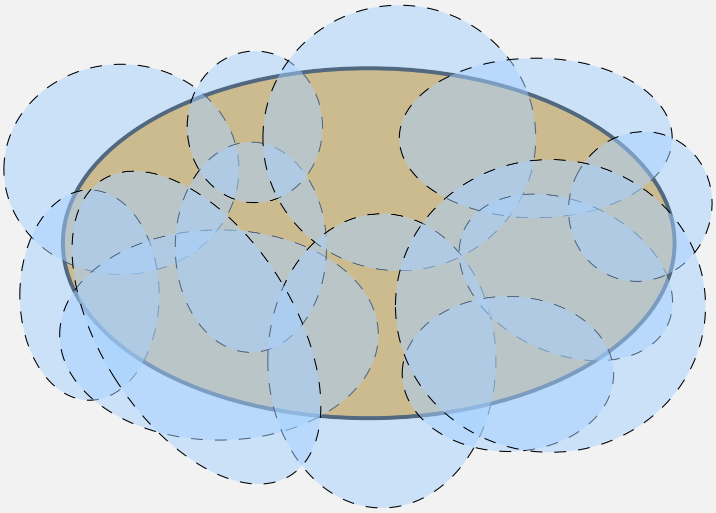
Formally, if:

$$C=\{U_\alpha : \alpha\in A\} \tag{7.3.56}$$

is an indexed family of open sets U_α, then C is a cover of X if:

$$E\subseteq\bigcup_{\alpha\in A}U_\alpha \tag{7.3.57}$$

Visually in a naive way this gives:



Open Cover

Definition (#7.3.300): A "closed ball" is similar to an open ball but differs in the sense that we include the elements located at a distance r from the center:

$$_fB_x^r(E)=\{y\in E|d(x,y)<r\} \tag{7.3.58}$$

Remark

For $0<r<r'$ the inclusions $_oB_x^r\subset B_x^r\subset B(x,r')$ are direct consequences of the definition of the open and closed ball.

✍**Example:**

The usual distance in $\mathbb{R}$ is given by $d(x,y)=|x-y|$. The balls are there simple intervals. For $x\in\mathbb{R}$ and $r\in\mathbb{R}_+^*$, we have:

$$_oB_x^r=]x-r,x+r[\quad\text{and}\quad {}_fB_x^r=[x-r,x+r] \tag{7.3.59}$$

Definition (#7.3.301): A "sphere" is given by:

$$S_x^r(E)=\{y\in E|d(x,y)=r\} \tag{7.3.60}$$

Remark

Since by definition $r>0$, open and closed balls are not empty because they contain at least their center. By cons, a sphere may be empty!

Example:

With $\mathbb{R}^n,\mathbb{C}^n$ we have seen in the previous examples we could set different distances. To distinguish them, we denote then by:

$$\begin{aligned} d_1(x,y)&=\sum_{i=1}^{n}|y_i-x_i|\\ d_2(x,y)&=\sqrt{\sum_{i=1}^{n}(y_i-x_i)^2}\\ d_\infty&=\max\{y_i-x_i,x_i-y_i\}\end{aligned}\tag{7.3.61}$$

So in $\mathbb{R}^2$ the closed balls with center O and of radius unit equivalent to the previous three formulations, have the following shapes (remember that $0<r\leq 1$ in this example):

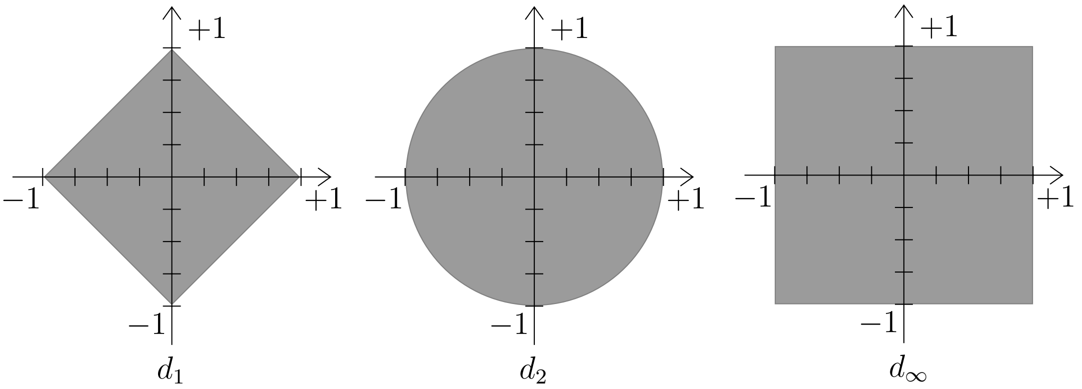


Figure 7.100: **Examples of closed balls of unit radius with different distances**

For example in statistics (see the section of the same name) we also use (among a lot of others) the Chi-2 distance given by:

$$d_{\chi^2}(x,y)=\sqrt{\sum_{i=1}^{n}\frac{(y_i-x_i)^2}{x_i}}\tag{7.3.62}$$

Or always in (multivariate) statistics (have a look to the Statistics section but also to the Industrial Engineering one) the "Mahalanobis distance":

$$d_M(x)=\sqrt{(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})}\tag{7.3.63}$$

Remarks

The proof that the Mahalonobis distance is indeed a distance is straightforward. Indeed, without loss of generality, let us rewrite $x-\mu=\vec{x}$. Then:

$$d_M(x)=\sqrt{\vec{x}^T\Sigma^{-1}\vec{x}} \qquad (7.3.64)$$

But as we have proved it in the section Statistics at page 607, Σ is positive semi-definite, hence Σ^{-1} is most of time invertible (see section Statistics page 608). When it's invertible, all its eigenvalues are positive, then Σ^{-1} is also positive semi-definite (see section Linear Algebra page 1957). This implies that:

$$\vec{x}^T\Sigma^{-1}\vec{x}\geq 0 \qquad (7.3.65)$$

Therefore the Mahalonobis distance is indeed a distance (we should also prove that it respects the triangular inequality to show it is also a metric!). Another way to see it as engineer is just to consider that as $\Sigma^{-1}\vec{x}$ simply gives just another vector $\vec{y}$ such that:

$$d_M(x)=\sqrt{\vec{x}^T\Sigma^{-1}\vec{x}}=\sqrt{\vec{x}^T\vec{y}}=\sqrt{\vec{x}\circ\vec{y}} \qquad (7.3.66)$$

This it's just the euclidean distance between two different vectors!

For any other semi-positive definite matrix A other than the variance-covariance matrix we also define the "elliptic metric":

$$d_e(x)=\sqrt{\vec{x}^TA^{-1}\vec{x}} \qquad (7.3.67)$$

Or in the Error Correcting Codes section (see page 7167) we use the Hamming distance given by:

$$\forall x,y\in Q^n: \quad d_H(x,y)=\text{Card}\left\{i\in\mathbb{N}|x_i\neq y_i\right\}=\#\left\{i\in\mathbb{N}|x_i\neq y_i\right\} \qquad (7.3.68)$$

and so on...

7.3.3.2 Partitions

Now that we have defined the concepts of balls, we can finally (almost) rigorously define the concepts of open and closed intervals (which in a space of more than one dimension are named "partitions") that we have so often used in the section of Functional Analysis and Differential and Integral Calculus.

Definition (#7.3.302): Let (X,d) of a metric space. We say that a subset A of X is "bounded" if there is a closed ball ${}_fB^r_{r_0}(X)$ such that $A\subseteq_f B^r_{r_0}(X)$:

$$\forall x\in A, d(x_0,x)\leq r \qquad (7.3.69)$$

Given the previous note on balls inclusions, it is clear that we can replace the word "closed" by "open". Moreover the triangle inequality implies that the bounded character of A does not depend on the choice of x_0 (with a x_0' we simply need to replace r by $r'=r+d\left(x_0,x_0'\right)$).

✍**Example:**

The odd–even topology is the topology where $X = \mathbb{N}$ and:

$$P = \{\{2k-1, 2k\}, k \in \mathbb{N}\} \tag{7.3.70}$$

the unbounded partitions of radius $r < 1$.

Therefore we see that unless P is trivial, at least one set in P contains more than one point, and the elements of this set are topologically indistinguishable: the topology does not separate points!

Definitions (#7.3.303):

D1. Let X be a set and (Y, d) a metric space. If X is a set, we say that a function $f : X \mapsto Y$ is "bounded" if its image $f(X)$ is bounded (the case of the sine or cosine function, for example).

D2. Given (E, d) a metric space, and given A a non-empty subset of E. For any $u \in E$ we note $d(u, A)$ and name "distance u to A", the positive real number:

$$d(u, A) = \inf_{a \in A} d(u, a) = \inf \{d(u, a) | a \in A\} \tag{7.3.71}$$

We extend the concept by writing:

$$\forall u \in E: \quad d(u, \varnothing) = +\infty \tag{7.3.72}$$

If A and B are two parts (subsets) of E we have respectively (perhaps this is more understandable in this way for some readers...):

$$d(A, B) = \inf \{d(x, y) | x \in A, y \in B\} \tag{7.3.73}$$

The reader must take care here to interpret $d(A, B)$ as the infinimum of the distance between the sets A and B, because the distance between the parties does not always define a distance in the usual way on the part for example of $\mathcal{P}(\mathbb{R})$.

Indeed, if we take again our famous example:

$$A = \{0\} \quad \text{and} \quad B = \left\{\frac{1}{n+1}, n \in \mathbb{N}\right\} \tag{7.3.74}$$

we have $d(A, B) = 0$ when $n \to 0$ while $A \neq B$.

Remarks

R1. If the reader has well understood the definition of "parts" (and especially the previous example) he has probably noticed that it does not necessarily always exist a $a \in A$ such that $d(u, A) = d(u, a)$. Accordingly, we write:

$$\forall \varepsilon > 0, \exists \alpha = \alpha(\varepsilon) \in A: \quad d(u, A) \leq d(u, \alpha) < d(u, A) + \varepsilon \tag{7.3.75}$$

Moreover, if such an α exists, it is obviously not necessarily unique.

R2. It should be remembered that this distance also meets the four axioms of distances (we can give the proof on request)!

D3. Given (E,d) a metric space, and let A be a part (subset) of E. We name "adhesion" of A and denote by adh(A) the subset of E defined by:

$$\text{adh}(A)=\{u\in E|d(u,A)=0\} \tag{7.3.76}$$

For example, the adhesion of the part (subset) of rational numbers $\mathbb{Q}$ (part A) of $\mathbb{R}$ (the metric space E) is a subset of $\mathbb{R}$ itself since any real number is the limit of a rational.

Especially, since $\forall u\in A: \quad =+\infty$, we have $\text{adh}(\varnothing)=\varnothing$, and since $\forall u\in E: \quad d(u,E)=0$, we have $\text{adh}(E)=E$.

Remarks

R1. Any element of the set adh(A) is named "adherent point" of A.

R2. We say that a part A of E is a "closed part" if it is equal to its adherence.

R3. We say that a part A of E is an "open part" if its complementary relatively to E:

$$\complement_E A=A\setminus E \tag{7.3.77}$$

is closed.

It follows from the definitions that (without proof):

$$u\in\text{adh}(A)\Leftrightarrow\forall r>0: \quad {}_fB_u^r\cap A=\varnothing \tag{7.3.78}$$

and:

$$A\text{ is open }\Leftrightarrow\forall u\in A,\exists\rho>0: \quad {}_fB_u^r\subset A \tag{7.3.79}$$

with some properties (supposed as very obvious but we can give the proof on request):

P1. If $A\subset E$ and $B\subset E$ satisfies $A\subset B$, then we have:

$$\forall u\in E: \quad d(u,B)\leq d(u,A) \tag{7.3.80}$$

P2. For all $A\subset E$, any $u\in E$ we have:

$$d(u,A)=d(u,\text{adh}(A)) \tag{7.3.81}$$

The latter property has for corollary (obvious and therefore without proof excepted on request): If for any $u\in E$, we have $d(u,A)=d(u,B)$ and $A,B\neq\varnothing$, we then have:

$$\text{adh}(A)=\text{adh}(B) \tag{7.3.82}$$

7.3.3.3 Formal Ball

The concept of distance from a point to a set gives the possibility to extend the notions of ball and sphere see previously. We will see now the basis concepts of a "formal ball" also named "generalized ball".

D1. Given $A \neq \varnothing$ and given a $r > 0$. We name "generalized open ball" of center A of radius r, the following set:

$$_{o}B_{A}^{r} = \{y \in E | d(y, A) < r\} \tag{7.3.83}$$

and respectively "generalized closed ball":

$$_{f}B_{A}^{r} = \{y \in E | d(y, A) \leq r\} \tag{7.3.84}$$

and respectively a "generalized sphere":

$$S_{A}^{r} = \{y \in E | d(y, A) = r\} \tag{7.3.85}$$

D2. Given (E, d) a metric space and let A, B be two non-empty parts (subsets) of E. We denote by $g(A, B)$ and name "gap" of A to B, the real number greater than or equal to zero such that:

$$\begin{aligned} g(A,B) &= \inf_{a \in A} g(a,B) = \inf_{b \in B} g(b,A) = \inf_{a \in A} \inf_{b \in B} g(a,b) = \inf_{b \in A} \inf_{a \in B} g(a,b) \\ &= \inf\{d(a,b) | (a,b) \in A \times B\} \end{aligned} \tag{7.3.86}$$

Remark

The triangle inequality $g(A, B) \leq g(A, C) + g(C, B)$ is not valid in the context of gaps. To prove it, a single example that contradicts this inequality is sufficient.

✍**Example:**

In $\mathbb{R}$ let us take $A = \{0, 1\}, B = \{2, 3\}, C = \{1, 3\}$ then we have:

$$(g(A, B) = 1) > g(A, C) + g(C, B) = 0 + 0 = 0 \tag{7.3.87}$$

There is therefore a contradiction.

7.3.3.4 Diameter

Definition (#7.3.304): Given (E, d) a metric space and A a non-empty part (subset) of E. We denote $\text{diam}(A)$ and name "diameter" of A, the positive non-zero real number:

$$\text{diam}(A) = \sup\{d(a, b) | (a, b) \in A \times A\} \tag{7.3.88}$$

Every non-empty part (subset) A of a metric space satisfying $\text{diam} < +\infty$ will also be say "bounded".

Remark

We consider the empty set $\varnothing$ as bound set of diameter A.

If the whole metric space (E,d) is bounded, we say that the distance d is bounded. For example, the discrete distance is limited, the usual distance on $\mathbb{R}$ is not.

We also have the following properties (the first two are usually trivial, the third one comes from the definition of the diameter itself):

P1. $\text{diam}(A)=0\Leftrightarrow A=\{a\}$ or $A=\varnothing$

P2. $A\subset B\Rightarrow \text{diam}(A)\leq \text{diam}(B)$

P3. $\text{diam}({}_fB_x^r)\leq 2r, \text{diam}({}_oB_x^r)\leq 2r, \text{diam}(S_x^r)\leq 2r$

> Caution!!! Concerning the latter property, the reader must take the habit of thinking with the Euclidean distance. The first common pitfall is to think that the second diameter (that of the open ball) should be strictly less but that would be forgetting that the board has no thickness strictly speaking!

There is also often an understanding problem with $\text{diam}(S_x^r)\leq 2r$. To be convinced just take the discrete distance (that for two points that are not confused is equal to 1, otherwise 0). Thus, in a metric space where we take S_x^r with $r=1$, we have indeed $\text{diam}(S_x^1)\leq 1$ (that is an interesting case because almost completely counter-intuitive).

P4. $\text{diam}(A\cup B)\leq \text{diam}(A)+g(A,B)+\text{diam}(B)$

To be convinced, in $\mathbb{R}$ take $A=B$, then we have (trivial strict inferiority):

$$(\text{diam}(A\cup B)=\text{diam}(A))\leq \text{diam}(A)+g(A,B)+\text{diam}B) \tag{7.3.89}$$

P5. A is bounded if and only if: $\exists r>0,\exists x\in E:\quad A\subset_o B_x^r$

Definition (#7.3.305): We name "Hausdorff excess" or "Hausdorff distance" from X to Y:

$$e(X,Y)=\sup\{d(x,Y)|x\in X\}=\max\{d(x,Y)|x\in X\} \tag{7.3.90}$$

that we found often in the literature with the more condensed notation:

$$e(X,Y)=d_H(X,Y) \tag{7.3.91}$$

or much more explicitly:

$$d_{\text{H}}(X,Y)=\max\left\{\sup_{x\in X}\inf_{y\in Y}d(x,y),\ \sup_{y\in Y}\inf_{x\in X}d(x,y)\right\} \tag{7.3.92}$$

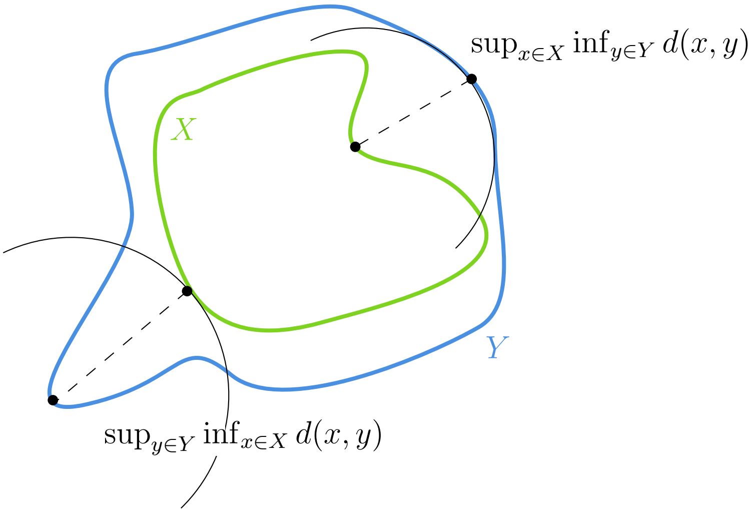

$\sup_{x\in X}\inf_{y\in Y} d(x,y)$
X
Y
$\sup_{y\in Y}\inf_{x\in X} d(x,y)$

Figure 7.101: Components of the calculation of d_H between X and Y (source: Wikipedia)

✍**Example:**

Let us take $X\subset\mathbb{R}^2$ as the unit radius circle centered at the origin and Y to the square circumscribing it. Elementary geometry concepts obviously leads to finding that the Hausdorff distance between the circle and the square is therefore:

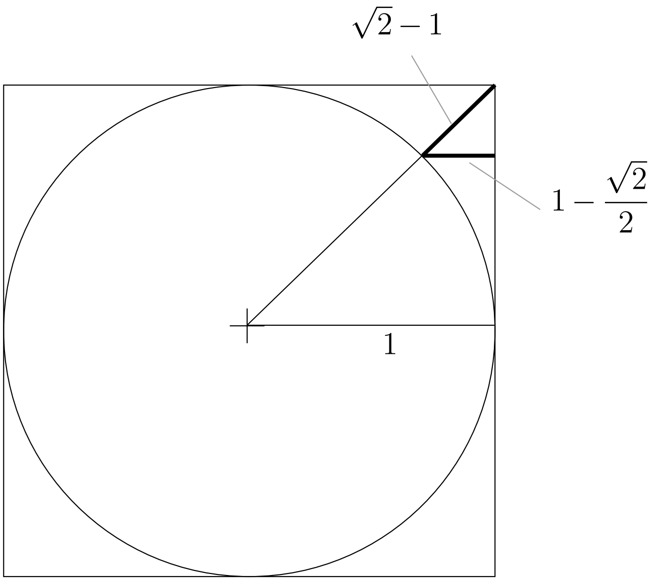

$\sqrt{2}-1$
$1-\dfrac{\sqrt{2}}{2}$
1

Figure 7.102: High-school example of a Hausdorff distance in the plane

technically:

$$e(X,Y)=d_H(X,Y)=\sqrt{2}-1 \tag{7.3.93}$$

Remark

We have generally $e(X,Y)\neq e(Y,X)$ and these quantities may not be finite.

7.3.4 Varieties

We now introduce the "varieties". These are topological spaces that are "locally as $\mathbb{R}^2$" (our space for example ...), that is locally euclidean.

Definitions (#7.3.306):

D1. A "topological variety of dimension n" is a Hausdorff space M such that for every $p \in M$ there exists an open neighbourhood $U \subset M$ with $p \in U$, an open neighbourhood $U' \subset \mathbb{R}^n$ and a homeomorphism such that:

$$\varphi : U \mapsto U' \quad (7.3.94)$$

D2. A "homeomorphism" between two spaces is a continuous bijection whose inverse is also continuous.

D3. The pairs (U,φ) are named "maps", U being the "domain of the map" and φ the "coordinate application". Instead of "map" sometimes we say also "coordinate system".

Remark

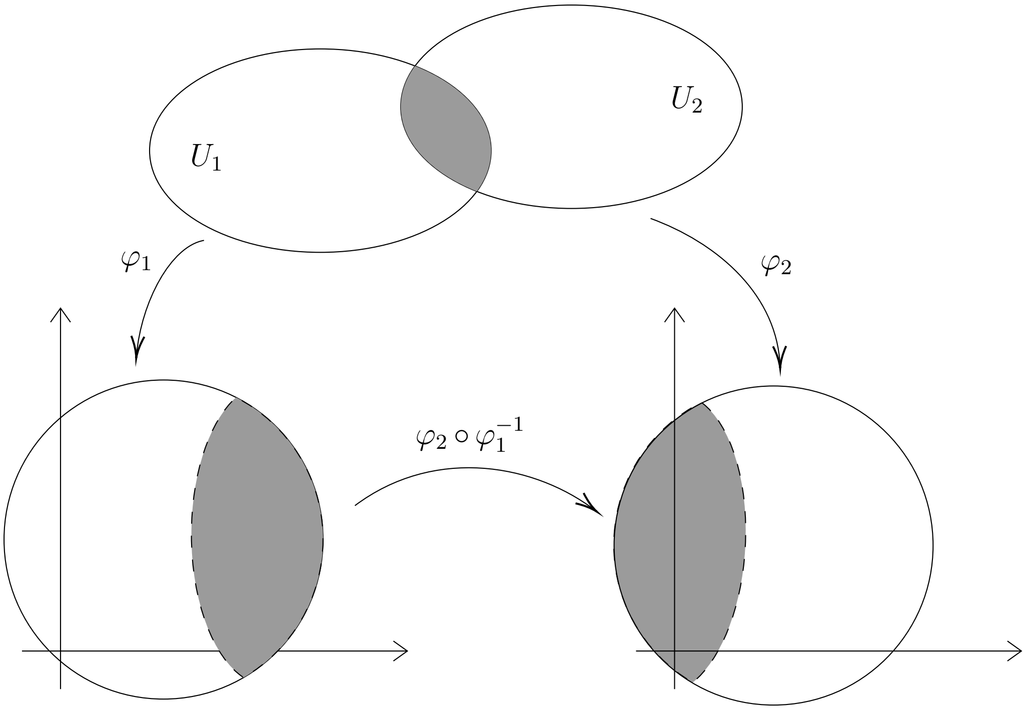
We will denote by $\dim(M)$ the dimension of a topological variety. Therefore:

$$\dim(M^n) = n \quad (7.3.95)$$

D4. Given M be a topological variety of dimension n. A family A of maps of M is named a "atlas" if for each $x \in M$, there exists a map $(U,\varphi) \in A$ such as $x \in U$.

If $(U_1,\varphi_1),(U_2,\varphi_2)$ are two maps of M such as $U_1 \cap U_2 \neq \varnothing$, then the application of map changes:

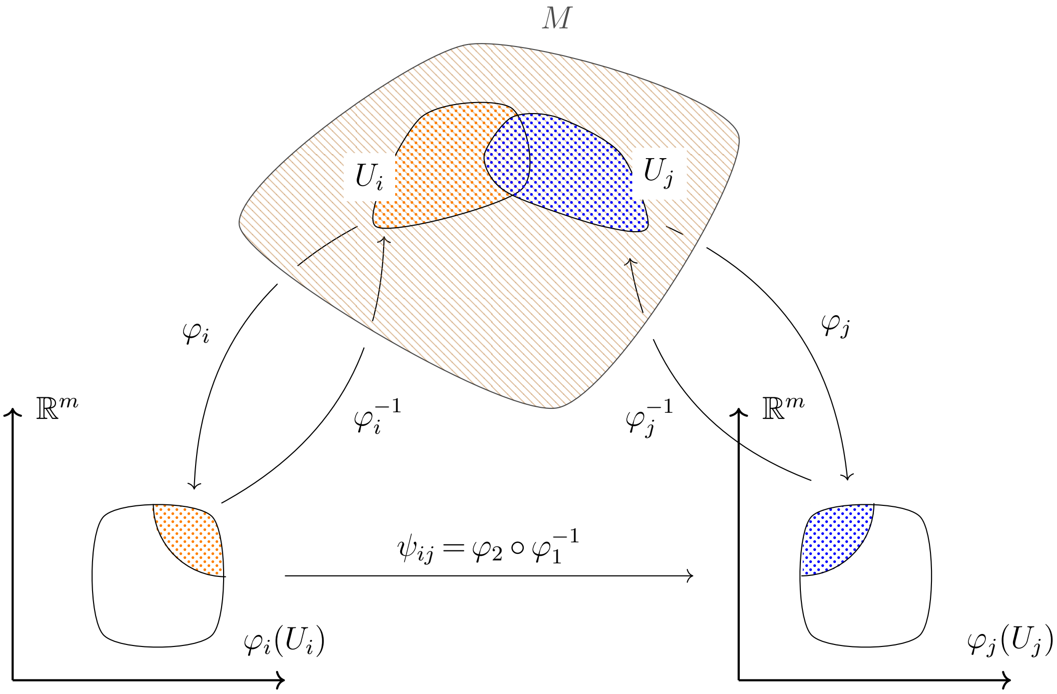
$$\varphi_2 \circ \varphi_1^{-1} : \varphi_1(U_1 \cap U_2) \rightarrow \varphi_2(U_1 \cap U_2) \quad (7.3.96)$$



U_1
U_2
φ_1
φ_2
$\varphi_2 \circ \varphi_1^{-1}$

Figure 7.103: A naive maps homeomorphisms

is obviously also a homeomorphism. We can find also the above figure sometimes in textbooks in a more explicit shape:



M
U_i
U_j
φ_i
φ_j
φ_i^{-1}
φ_j^{-1}
$\mathbb{R}^m$
$\mathbb{R}^m$
$\psi_{ij} = \varphi_2 \circ \varphi_1^{-1}$
$\varphi_i(U_i)$
$\varphi_j(U_j)$

Figure 7.104: A more complete maps homeomorphisms

More "geometrically" it looks like this:

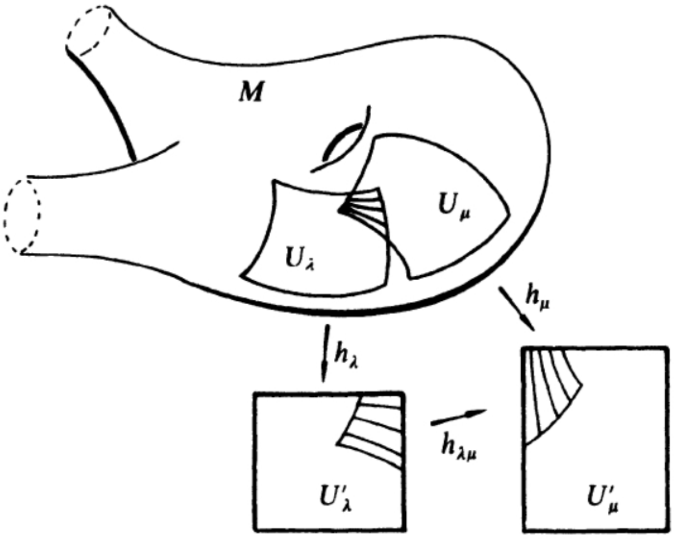


Figure 7.105: **Intuitive maps homeomorphism (source: ?)**

7.3.4.1 Subvariety

Definition (#7.3.307): A subset of a variety is itself a variety named a "subvariety".

✍**Example:**

A sphere of the three-dimensional Euclidean space $\mathbb{R}^3$ is an (algebraic) variety since it is defined by a polynomial equation. For example and is obviously smooth and locally euclidean:

$$x^2+y^2+z^2-1=0 \tag{7.3.97}$$

defines the sphere of radius 1 centered at the origin. Its intersection with the xy-plane is a circle given by the system of polynomial equations:

$$\begin{aligned} x^2+y^2+z^2-1&=0\\ z&=0 \end{aligned} \tag{7.3.98}$$

Hence the circle is itself an algebraic variety, and a subvariety of the sphere, and of the plane as well.

7.3.4.2 Surfaces Homeomorphism

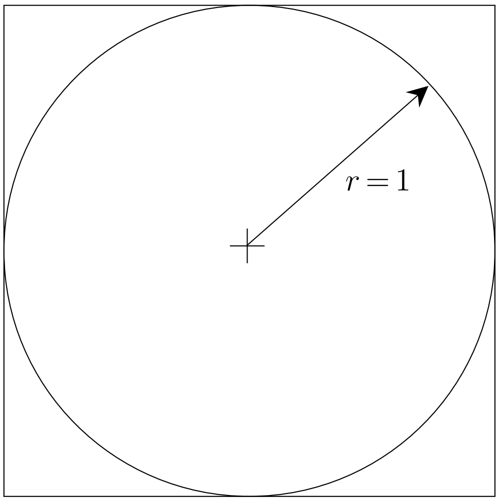
Definition (#7.3.308): In the mathematical field of topology, a "homeomorphism or "topological isomorphism" is a continuous function between topological spaces that has a continuous inverse function. Roughly speaking, a topological space is a geometric object, and the homeomorphism is a continuous stretching and bending of the object into a new shape. Thus, a square and a circle are homeomorphic to each other, but a sphere and a torus are not.

More formally, remember that an application $\varphi : X \mapsto Y$ between two topological spaces is named a homeomorphism if it has the following properties:

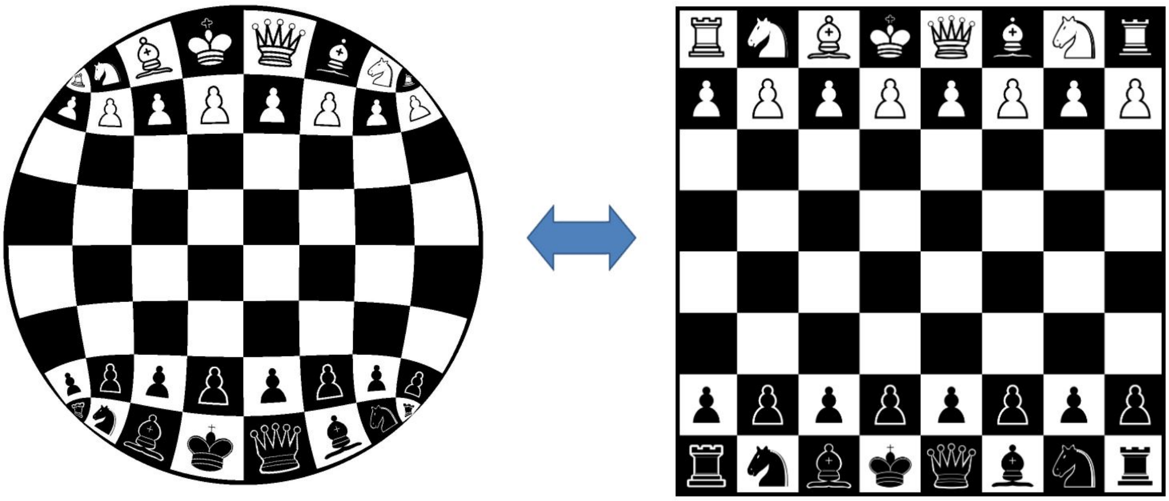
1. φ is a bijection (one-to-one and onto)
2. φ is continuous
3. The reciprocal function φ^{-1} is continuous

Can we say that a square (being a special map in $\mathbb{R}^2$) is homeomorph to a circle (being another special map in $\mathbb{R}^2$), or a torus to a cup of tea... If this is possible we must be able to find a closed form bijective expression between the two surfaces.

As the pure theoretical concepts are very not friendly in our point of view let us begin with a two dimensional special case. Let us show (prove) first that we can transform all interior points of square of side 1 into all interior points circle of radius 1. This is represented by the well known following figure:



Such mappings have particular interest in industrial design or just simply for communication purposes (Photoshop effects or Statistics charts deformation as we do many times in the `R` Software):



or for defishing fish eyes captors, picture or security mirrors:

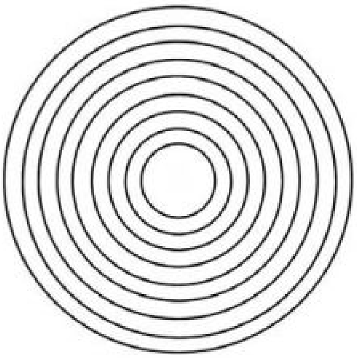
Recall that we defined unit disc as the set:

$$\mathcal{D}=\left\{(u,v)\in\mathbb{R}^2|u^2+v^2\leq 1\right\} \tag{7.3.99}$$

If we think of the unit disc as a continuum of concentric circles with radii growing from zero to one, we can parametrize the unit disc as the set:

$$\mathcal{D}=\left\{(u,v)\in\mathbb{R}^2|u^2+v^2=t^2,0\leq t\leq 1\right\} \tag{7.3.100}$$

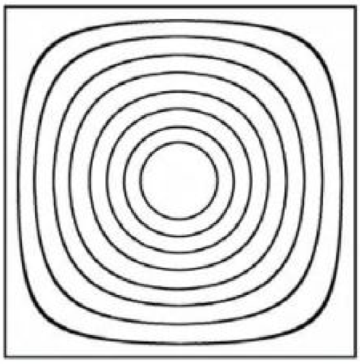
In doing so, we introduced a parameter t this is the distance of point (u,v) to the origin.



In analogy to the circular continuum of the unit disc, one can write the square region $[-1,1]\times[-1,1]$ as the set:

$$\mathcal{S}=\left\{(x,y)\in\mathbb{R}^2|x^2+y^2=t^2, 0\leq t\leq 1\right\} \tag{7.3.101}$$

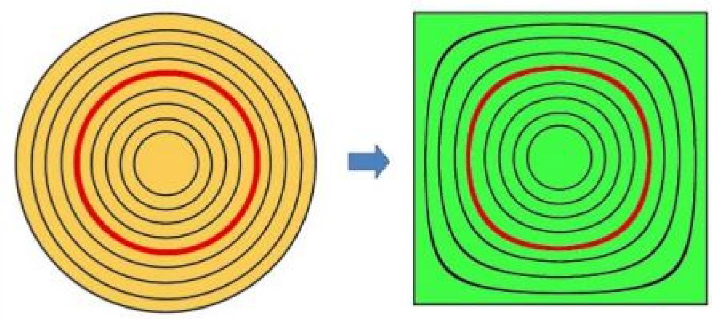
In other words, the square can be considered as a continuum of concentric shrunken FG-squircles (see section Analytical Geometry page 2812).



Topologists denote the proof that the interior points of two geometries are homeomorph in this special case using the following notation:

$$\mathring{\mathcal{D}}\sim\mathring{\mathcal{S}} \tag{7.3.102}$$

We will now show that $\mathring{\mathcal{D}}\mapsto\mathring{\mathcal{D}}$ as it is the most common case in practice in our point of view. That is to say:



We can establish a correspondence between the unit disc and the square region by mapping every circular contour in the interior of the disc to a squircular contour in the interior of the square. In other words, we map contour curves in the circular continuum of the disc to those in the squircular continuum of the square. This can be done by equating the parameter t of both sets to get the equation:

$$u^2+v^2=x^2+y^2-x^2y^2 \tag{7.3.103}$$

We name this equation the "squircularity condition" for mapping a circular disc to a square region.

It is easy to derive the FG-Squircular mapping by combining the squircularity condition:

$$t^2=u^2+v^2=x^2+y^2-x^2y^2 \tag{7.3.104}$$

That we can also write:

$$t^2=(t\cos(\theta))^2+(t\cos(\theta))^2=x^2+y^2-x^2y^2 \tag{7.3.105}$$

Using radial to cartesian coordinates (see section Vector Calculus page 1863):

$$t^2=\left(t\frac{x}{\sqrt{x^2+y^2}}\right)^2+\left(t\frac{y}{\sqrt{x^2+y^2}}\right)^2=x^2+y^2-x^2y^2 \tag{7.3.106}$$

Therefore by equivalence we get:

$$u=t\frac{x}{\sqrt{x^2+y^2}} \qquad v=t\frac{x}{\sqrt{x^2+y^2}} \tag{7.3.107}$$

After substitution of parameter t, we get:

$$\boxed{\begin{aligned}u&=\sqrt{x^2+y^2-x^2y^2}\frac{x}{\sqrt{x^2+y^2}}\\ v&=\sqrt{x^2+y^2-x^2y^2}\frac{x}{\sqrt{x^2+y^2}}\end{aligned}} \tag{7.3.108}$$

In other words, this is a radial mapping that converts circular contours on the disc to squircular contours on the square.

We shall now derive the inverse equations for the FG-Squircular mapping. But as it is boring to write in LaTeX and it is not used to much in practice we will omit the latter for the moment.

7.3.4.3 Differential Varieties

"Differential varieties" or "differentiable manifolds" are the basic objects of differential topology and differential geometry. These are varieties, "curved spaces" locally modelled on the Euclidean space of dimension n, on which it is possible to generalize a good part of the operations of differential and integral calculus.

A differential variety is therefore first defined by the data of a topological variety, a topological space locally homeomorphic to the space $\mathbb{R}^n$. Local homeomorphisms are named "maps" and define local coordinate systems. The differential structure is defined by requiring certain regularity properties of the transition maps between maps. This structure makes it possible, for example, to give a global definition of the notion of differentiable application, or of vector field with its integral curves.

On the other hand, unless the variety is provided with additional structures (such as a Riemannian metric), the calculations of derivatives of order 2, the notion of measurement of a part, do not admit of natural generalization.

Definitions (#7.3.309):

D1. A "differentiable variety" is a topological space M where the applications φ are of class $\mathcal{C}^{+\infty}$.

D2. A "diffeomorphism" is an application $\varphi : U \mapsto U'$ where U, U' are open domains of $\mathbb{R}^n$ and if φ is a homeomorphism and furthermore U, U' are differentiable!

> **Remark**
>
> Remember that "Differentiable" in this context will always mean of class $\mathcal{C}^{+\infty}$.

D3. Given a topological variety $M := M^n$ (to simplify the notations), two maps $(U_1, \varphi_1), (U_2, \varphi_2)$ of M are named compatible maps (more precisely: compatible of class $\mathcal{C}^{+\infty}$) if one of these two properties is satisfied:

P1. $U_1 \cap U_2 \neq \varnothing$ and the application $\varphi_2 \circ \varphi_1^{-1}$ of map changes is a diffeomorphism.

P2. $U_1 \cap U_2 = \varnothing$

An atlas A of M is differentiable if all maps of A are compatible between them.

> **Remarque**
>
> Given a differentiable atlas, it is sometimes necessary to complete it: we say that a map of M is compatible with a differentiable atlas if it is compatible with every map of A. An atlas of A is a "maximal atlas" if every compatible map of A belongs already to A. A maximal atlas is named a "differentiable structure".

~70 %

★★★★☆

12 votes, 71.67%

7.4. Measure Theory

Caution! The level of abstraction and of motivation required for reading and understanding this section is quite high for engineers (target audience of this textbook for recall). The reader should be comfortable with the concepts seen in the section Set Theory as well as the one of Topology. We also apologize for the actual lack of figures.

THE measure, in the topological sense, will allow us to generalize the elementary notion of measure of a segment or area (in the Riemann sense, for example) and is inseparable from the new theory of integration that will build Lebesgue from the years to 11901-11902 (holocene calendar) and we will address here to build mathematical tools much more powerful than the simple Riemann integral (see section Differential and Integral Calculus page 1499) with practical and numerical example in MATLAB™.

The philosophers of science who developed measurement theory were largely concerned with epistemic questions like: we can't observe correlations between physical objects and real numbers, so how can the use of real numbers be justified in terms of things we can observe? Indeed, privileges a single unit of mass, involves real numbers in the facts of mass. Why is the latter bad?:

1. Real numbers are abstract and therefore causally inert
2. Real numbers don't fundamentally exist
3. Real numbers are constructed entities, and constructed entities can't be involved in fundamental facts

Measure Theory will also allow us to rigorously define the concept of "measurement" (no matter what is the measure) and so to return to important results of the study of probabilities (see section Probabilities page 450)! Indeed, we will see (we will define the vocabulary that follows just now further below) why (U, A, P) is[10] a "probability space" where A is in fact a "tribe" (also named "algebra") on U and P a measure on the measurable space (U, A).

---

[10] The notation of that tuple varies depending if we study topology, statistics or integral calculus.

7.4.1 Measurable Spaces

When in mathematics we calculate derivatives, primitives or simply count stuff, we carry implicitly a measure of an object or set of objects. Rigorously, mathematicians want to define how the measured thing can be structured, how to make a measurement of it and the properties resulting!

Measure theory is all about abstracting the idea of "size". What do we mean by size? Size is a number that we attribute to an object that obeys a specific, intuitive property: if we break the object apart, the sizes of the pieces should add up to the size of the whole.

There are many phenomena that follow this basic "size"-like idea including such phenomena as mass, length, and volume (and as we will see later, probability). For example, if we have a ball of clay and we split the ball of clay into two, then the masses of the two smaller balls of clay will add up to the mass of the original ball of clay.

Before we can discuss the "sizes" of objects and pieces of objects, we need a mathematical way to describe an object. Measure theory uses a set to denote the object that we are considering. For example, a set may represent a ball of clay. Given our object/set, we then need a way to describe how a set can be broken into pieces. This notion of "breaking an object into pieces" is described by the definition below.

Definitions (#7.4.310):

D1. Let E be a set, a "tribe" on E is a family $\mathcal{A}$ (this notation comes from the fact that many people speak of "**A**lgebra sets" instead of "tribe") of subsets of E satisfying the following axioms:

A1. $E \in \mathcal{A}$ (see examples below - E being one of the possible elements of $\mathcal{A}$).

A2. If A is a member of a tribe then:

$$A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A} \tag{7.4.1}$$

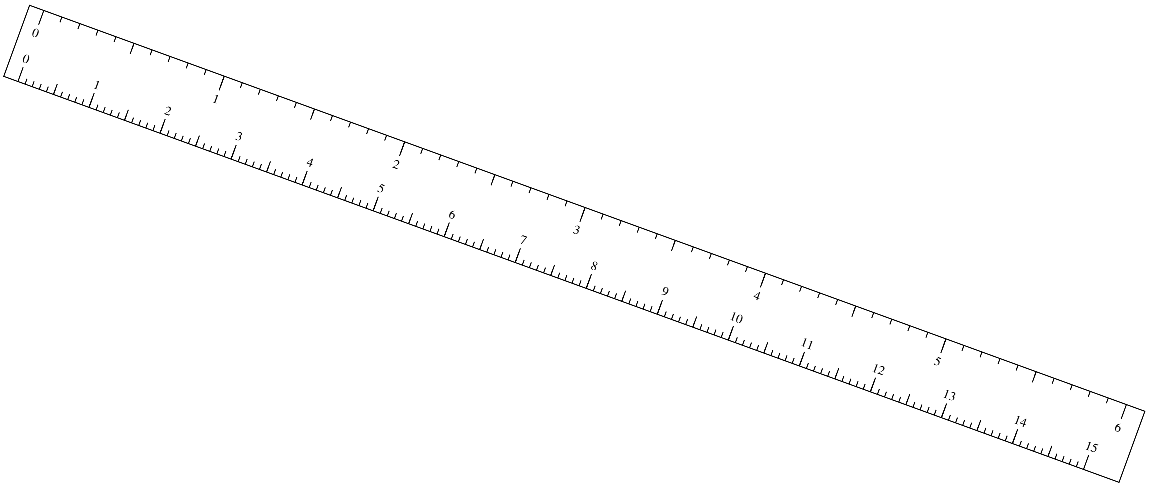
This means that $\mathcal{A}$ is "stable by transition to complementary". This axiom implies that the empty set is always an element of a tribe!

A3. For any sequence (A_n) of elements of $\mathcal{A}$ we have:

$$\bigcup_{i=1}^{n} A_n \in \mathcal{A} \tag{7.4.2}$$

We then say that $\mathcal{A}$ is then "stable by countable union" or also said to be "closed under set unions".

For example, the graduating from a simple ruler of measurement ... satisfies these three axioms!



Remarks

R1. We write $E \in \mathcal{A}$ because we consider with this notation E not anymore as a subset of $\mathcal{A}$ but as an element of $\mathcal{A}$!

R2. The uncountable cases are typical of topology, statistics or integral calculus!

D2. If in the third axiom we require that $\mathcal{A}$ is stable under finite (uncountable) union, i.e.:

$$\bigcup_{i=1}^{+\infty} A_n \in \mathcal{A} \tag{7.4.3}$$

then we impose the more general notion of "σ-algebra". Thus, a tribe is necessarily contained in an σ-algebra (but the opposite is not true just because the axiom is stronger) such that we can write:

$$\mathcal{A} \subseteq \sigma\text{-algebra}(\mathcal{A}) \tag{7.4.4}$$

With these properties in mind, another way to imagine this is to think of E as a ceramic plate that is cracked. The σ-algebra describes all the ways that the plate can fall apart along its "cracks" for a given type of crack.

Remark

In the field of probabilities, E is assimilated to the Universe of events and $\mathcal{A}$ to a family of events and we speak then of "probabilistic space" or simply of... "measurable space".

Okay, we have spent a while thinking about representing objects and pieces of an object. Now how do we assign each piece a "size"? This is where a measure comes into play. A measure is simply a function that assigns each piece of an object a size. As we discussed before, the sizes of two pieces of an object must add up to the size of the whole, which is exactly how a measure is defined!

D3. Given a set E and a σ-algebra on E, denoted $\mathcal{A}$, the function:

$$\mu:\mathcal{A}\rightarrow\mathbb{R} \tag{7.4.5}$$

is a "Lebesgue measure" if the following properties hold for μ (also denoted sometimes μ_L):

P1. Nonnegative: $\forall A\in\mathcal{A},\mu(A)\geq 0$

P2. Empty set has zero Lebesgue measure: $\mu(\emptyset)=0$

P3. The Lebesgue measure is invariant, i.e. $\mu(A)+\mu(A+c^{te})$

P4. Countable additivity: Given $A,B\in\mathcal{F}$ where $A\cap B=\emptyset$, then $\mu(A\cup B)=\mu(A)+\mu(B)$

The first part of the definition requires that a "size" (i.e. measure) cannot be negative. That's intuitive. The second part says that the "null piece" (i.e. nothing) has a size of zero. Finally, the third part says that the sum of the sizes of two pieces of an object must equal the size of the two pieces glued together.

D4. The pair $(E,\mathcal{A})$ is named "measurable space" and we say that the elements of $\mathcal{A}$ are "measurable sets".

D5. The tuple $(E,\mathcal{A},\mu)$ is named "measure space".

Therefore a measure space is simply the combination of a set (i.e. object), a σ-algebra (i.e. pieces of the object), and a measure μ (i.e. ways to describe the sizes of the pieces).

Intuitively, each element of the σ-algebra, represents a "piece" of the object. The first criteria in the definition above establishes a "null piece" (i.e. a piece of zero size) can be considered a piece of the object. The second criteria establishes the fact that if we break off a piece, A, of the object, then the remaining object A^c (the complement of A) is also a valid piece. Finally, the third criteria establishes the fact that if we glue a countable set of pieces together, the resultant piece is also a piece of the object.

✍**Examples:**

E1. Given $E=\{1,2\}$ a set of cardinal 2... The only two tribe $\mathcal{A}$ that satisfy the three axioms are:

$$\mathcal{A}_1=\{\varnothing,\{1\},\{2\},\{1,2\}\} \qquad \mathcal{A}_2=\{\varnothing,\{1,2\}\} \tag{7.4.6}$$

There are no other tribes for the set E as these two (the normal one, and the maximum one), because we must not forget that the union of each elements of the tribe must also be in the tribe (axiom A3), and also the complement of a member (axiom A2).

We also see from this example that if E is a set then $\{E,\varnothing\}$ is indeed a tribe!

E2. The set of parts of E, denoted $\mathcal{P}(E)$ is also a tribe (dixit previous example).

A tribe $\mathcal{A}$ is also "stable by the union of the finite complementaries". Indeed, if (A_n) is a sequence of elements of $\mathcal{A}$ we have (trivial when taking for example the previous first example):

$$\bigcup_n A_i^c \in \mathcal{A} \quad \text{and} \quad \left(\bigcup_n A_i^c\right)^c \in \mathcal{A} \tag{7.4.7}$$

A tribe is also "stable by finite intersection", that is to say (trivial also by taking the previous first example):

$$\bigcap_{i=1}^{+\infty} A_i \in \mathcal{A} \tag{7.4.8}$$

which brings to the property that a tribe is stable by finite unions and intersections. Especially, if we take two elements of a tribe $A, B \in \mathcal{A}$, then $A \setminus B \in \mathcal{A}$ with for recall (see section Set Theory page 398):

$$\boxed{A \setminus B = A \cap B^c} \tag{7.4.9}$$

Obviously it can be proven with a general theorem that tribes (i.e. algebras) are closed under set intersections, differences, and symmetric differences but we consider that latter as quite intuitive and not necessary in a book dedicated to engineers.

Remark

Most readers should probably easily see with the previous first example that if $(\mathcal{A}_i)_I$ is a family of tribes on E then $\bigcap_I \mathcal{A}_i$ is also a tribe (the verification is almost immediate).

Well it is nice to play with potatoes and sub-potatoes... and their complementary but let us continue...

Definition (#7.4.311): Given E a set and $\mathcal{B}$ a family of subsets of $\mathcal{P}(E)$ such that $\mathcal{B} \subset \mathcal{P}(E)$. We denote by definition:

$$\boxed{\sigma(\mathcal{B}) = \bigcap_{\mathcal{A} \in \mathcal{S}} \mathcal{A}} \tag{7.4.10}$$

the "generated tribe" by $\mathcal{B}$. Therefore $\sigma(\mathcal{B})$ is by definition the smallest tribe containing $\mathcal{B}$ (and by extension the smallest tribe of E).

Below are three small examples that gives the opportunity to check if what precedes has been well understood and that also gives the possibility to highlight important results for what will follow:

✍**Examples:**

E1. Given a set E and $A \subset E, A \neq E$ and also $\mathcal{B} = \{A\}$ then (when A is seen as a subset of E as given by the statement of a family of subsets!):

$$\sigma(\mathcal{B}) = \{\varnothing, E, A, A^c\} \tag{7.4.11}$$

E2. If $\mathcal{A}$ is a tribe on E then:

$$\sigma(\mathcal{A})=\mathcal{A} \tag{7.4.12}$$

E3. Given $E=\{1,2,3,4\}$ and $A=\{\{1,2\},3\}$ we then have (take care because now A is a family of parts (subsets) and not only a unique subset!) the following generated tribe:

$$\sigma(\mathcal{A})=\{\{\varnothing\},\{3\},\{4\},\{1,2\},\{3,4\},\{1,2,3\},\{1,2,4\},\{1,2,3,4\}\} \tag{7.4.13}$$

Rather than determining this tribe by seeking the smallest tribe $\mathcal{P}(E)$ containing A (which would be laborious) we play with the axioms defining a tribe to easily find it.

So therefore we find well in $\sigma(A)$ at least the obligatory empty set $\{\varnothing\}$ and also:

$$E=\{1,2,3,4\} \tag{7.4.14}$$

following the axiom A1 and:

$$E=\{\{1,2\},\{3\}\} \tag{7.4.15}$$

itself by the definition of $\sigma(A)$ and the complementaries of:

$$\{\{3\}^c=\{1,2,4\},\{1,2\}^c=\{3,4\},\{\{1,2\}\cup\{3\}\}^c=\{4\}\} \tag{7.4.16}$$

following the axiom A2 and also the unions:

$$\bigcup_i A_n=\{\{1,2\}\cup\{3\}=\{1,2,3\},\{1,2\}\cup\{1,2,3\}=\{1,2,3\},\ldots\} \tag{7.4.17}$$

following the axiom A3.

Definition (#7.4.312): Let E be a topological space (see section Topology page 2447). We denote by $\mathcal{B}(E)$ the tribe generated by the open sets of E. $\mathcal{B}(E)$ is named the "borelian tribe" of E. The elements of $\mathcal{B}(E)$ are named the "borelians" of E.

Remarks

R1. The notion of borelian tribe is especially interesting because it is necessary for the definition of "Lebesgue tribe" and afterwards to the "Lebesgue measure" that will lead us to define the famous "Lebesgue integral"!

R2. The tribe $\mathcal{B}(E)$ being stable by going to the complementary, it also contains all closed subsets.

R3. If E is a topological space with a finite basis, $\mathcal{B}(A)$ in generated by the opens of the basis.

✍**Example:**

If $\mathbb{R}$ designates the space provided of real numbers with the Euclidean topology (see section Topology page 2453), the family of open intervals with rational bounds is a "countable base" (given the bounds...) of $\mathbb{R}$ and therefore generates $\mathcal{B}(\mathbb{R})$. Same thing for $\mathbb{R}^d$ with for countable basis the family of open spaces with rational bounds.

Theorem 7.8. Let us now consider $\mathcal{S}$ a dense set (see section Topology page 2449) in $\mathbb{R}$. The following families generate $\mathcal{B}(\mathbb{R})$:

$$\{]-\infty,a[\,,a\in\mathcal{S}\}\quad\{]-\infty,a]\,,a\in\mathcal{S}\}\quad\{]a,+\infty[\,,a\in\mathcal{S}\}\quad\{[a,+\infty[\,,a\in\mathcal{S}\}\tag{7.4.18}$$

Proof 7.8.1. Given (the family of open subsets):

$$\mathcal{F}=\{]-\infty,a]\,,a\in\mathcal{S}\}\tag{7.4.19}$$

We have obviously:

$$\sigma(\mathcal{F})\subseteq\mathcal{B}(\mathbb{R})\tag{7.4.20}$$

Furthermore:

$$([a,+\infty[\cap]-\infty,b[)=[a,b[\text{ if }a<b\tag{7.4.21}$$

Therefore the intervals of the type $[a,b[$ with a and b in $\mathcal{S}$ also belongs to $\sigma(\mathcal{F})$. Therefore, if we generalize, with $x<y$, it exists a sequence (a_n) of elements of $\mathcal{S}$ decreasing to x and a sequence (b_n) of elements of $\mathcal{S}$ increasing to y such that:

$$]x,y[=\bigcup_n[a_b,b_n[\in\sigma(\mathcal{F})\tag{7.4.22}$$

which bring in the same way as $E\in\mathcal{A}$ that $\mathcal{B}(\mathbb{R})\subseteq\sigma(\mathcal{F})$. Other cases can be treated analogously.

■ Q.E.D.

Theorem 7.9. Given $(E,\mathcal{A})$ a measurable space and $A\subseteq E$ (and $A\in\mathcal{A}$) (where A is therefore considerate as a subset and non as an element!). The family $\{A\cap B|B\in\mathcal{A}\}$ is a tribe on A named "trace tribe" of $\mathcal{A}$ on A, that we will denote by $A\cap\mathcal{A}$. Furthermore, if $A\in\mathcal{A}$, the trace tribe is formed by the measurable elements contained in A.

Proof 7.9.1. We will do a proof by the example (... yes it is not a real proof...). For this we check the three points that define a tribe:

1. $A=(E\cap A)\Rightarrow A\in(A\cap\mathcal{A})$
2. Given $B\in\mathcal{A}$, $A\setminus(A\cap B)=A\setminus B=A\cap B^c=A\cap B^c$ and therefore $A\setminus(A\cap B)\in(A\cap\mathcal{A})$

✍**Example:**

Given $E=\{1,2,3\}$ then (a tribe among others - do not forget the stability by union!):

$$\mathcal{A}=\{\varnothing,\{1\},\{2\},\{3\},\{1,2\},\{1,3\},\{2,3\},\{1,2,3\}\} \tag{7.4.23}$$

Let us choose $A=\{1,2\},B=\{2,3\}$ (it is obvious that $\{A\cap B|B\in\mathcal{A}\}$ is a tribe on A). Then:

$$\begin{aligned}A\setminus(A\cap B)=\{1,2\}-(\{1,2\}\cap\{2,3\})=\{1,2\}-\{2\}&=\{1\}\\ A\setminus B=\{1,2\}-\{2,3\}&=\{1\}\\ A\cap B^c=\{1,2\}\cap\{2,3\}^c=\{1,2\}\cap\{1\}&=\{1\}\end{aligned} \tag{7.4.24}$$

and we have well $\{1\}\in A$ and also $A\in(A\cap\mathcal{A})$.

3. Given $(A\cap B_n)$ a sequence of elements of $A\cap\mathcal{A}\,(B_n\in\mathcal{A})$ then:

$$\bigcup_n(A\cap B_n)=A\cap\underbrace{\left(\bigcup_n B_n\right)}_{\in\mathcal{A}}\in A\cap\mathcal{A} \tag{7.4.25}$$

The last statement of the proposition will be supposed as obvious (if not, let us know!).

■ Q.E.D.

Given now E a set, $\mathcal{C}$ a family of subsets of E and $A\subseteq E$ non empty. We denote by $A\cap\mathcal{C}$, the trace $\mathcal{C}$ on A and $\sigma_A(A\cap\mathcal{C})$ the tribe generated by $A\cap\mathcal{C}$ on A. Therefore:

$$A\cap\sigma(\mathcal{C})=\sigma_A(A\cap\mathcal{C}) \tag{7.4.26}$$

✍**Example:**

Given the set $E=\{1,2,3,4\},\mathcal{C}=\{\{1,2\},\{3\}\},A=\{3,4\}$ then:

$$A\cap\mathcal{C}=\{\varnothing,\{3\}\}\quad\sigma_A(A\cap\mathcal{C})=\sigma_A\big(\{\varnothing,\{3\}\}\big)=\big\{\varnothing,\{3\},\{4\},\{3,4\}\big\} \tag{7.4.27}$$

and let us check that $A\cap\sigma(\mathcal{C})=\sigma_A(A\cap\mathcal{C})$:

$$\begin{aligned}A\cap\sigma(\mathcal{C})&=\{3,4\}\cap\sigma\big\{\{1,2\},\{3\}\big\}\\&=\{3,4\}\cap\big\{\varnothing,\{3\},\{4\},\{1,2\},\{3,4\},\{1,2,3\},\{1,2,4\},\{1,2,3,4\}\big\}\\&=\big\{\varnothing,\{3\},\{4\},\{3,4\}\big\}\end{aligned} \tag{7.4.28}$$

So the equality is satisfied!

A trivial corollary of this equality is that if we consider a topological space E and $A \subseteq E$ with the induced topology, then:

$$A \cap \mathcal{B}(E) = \mathcal{B}(A) \qquad (7.4.29)$$

We will study more in details later σ-algebras but first let us recall that a tribe (sometimes named "algebra of sets") on E must satisfy the following properties:

P1. Has to contain E

P2. Must be stable by the complementary

P3. Must be stable by countable union or intersection

Let us recall (see section Set Theory page 398) that if we have E that is a set, then for every $A, B \subseteq E$ we define the symmetric difference $A\Delta B$ between A and B by:

$$A\Delta B = (A \setminus B) \cup (B \setminus A) \qquad (7.4.30)$$

Trivial properties are as follows:

P1. A σ-algebra is stable by symmetric difference ($A, B \in \mathcal{A}$ we have $A\Delta B \in \mathcal{A}$)

P2. $A\Delta B = B\Delta A$

P3. $A^c \Delta B^c = A\Delta B$

P4. $A\Delta B = (A \cup B) \setminus (A \cap B)$

Theorem 7.10. If $\mathcal{B}$ is a σ-algebra over E, then $(\mathcal{B}, \Delta, \cap)$ is a "Boolean ring" (or "Boolean algebra" but be careful with the term "algebra" here which can cause confusion with the corresponding structure in Set theory) with $\varnothing$ and E as neutral "additive" element (Δ) and respectively "multiplicative" ($\cap$).

Remark

For reminders on the items listed in the preceding paragraph, the reader can refer to the section of Set Theory at page 378 and the subsection of Boolean Algebra (see section Formal Logic Systems page 7105).

Proof 7.10.1. The "addition" Δ is associative because developing we get (this can verified by an arrow diagram if needed - the "potatoes"):

$$A\Delta(B\Delta C) = (A \cap B \cap C) \cup (A \cap B^c \cap C^c) \cup (A^c \cap B \cap C^c) \cup (A^c \cap B^c \cap C) \qquad (7.4.31)$$

and the latter expression is stable by permutation (commutation) of A and C (same method of verification). Therefore:

$$A\Delta(B\Delta C) = C\Delta(B\Delta A) = (B\Delta A)\Delta C = (A\Delta B)\Delta C \qquad (7.4.32)$$

We check that $\varnothing$ is neutral with respect to the symmetric difference (the proof that E is neutral with respect to inclusion is obvious). It is trivial that:

$$A\Delta\varnothing=\varnothing\Delta A \quad \text{and that} \quad A\Delta A=\varnothing \tag{7.4.33}$$

$(\mathcal{B},\Delta,\varnothing)$ is therefore well an Abelian group with respect to the law Δ (symmetric difference).

Finally $\cap$ is distributive with respect to Δ. Indeed:

$$\begin{aligned} A\cap(B\Delta C)&=A\cap(B\setminus C\cup C\setminus B)\\ &=(A\cap B\setminus C)\cup(A\cap C\setminus B)\\ &=\big(A\cap B\setminus(A\cap C)\big)\cup\big(A\cap C\setminus(A\cap B)\big)\\ &=(A\cap B)\Delta(A\cap C)\end{aligned} \tag{7.4.34}$$

What makes $(\mathcal{B},\Delta,\varnothing)$ is indeed a ring (furthermore of a commutative ring!).

■ Q.E.D.

7.4.1.1 Monotone Classes

Definition (#7.4.313): Let E be a set. A "monotone class" on E is a family $\mathcal{C}$ of subsets of E satisfying the following axioms:

A1. $E \in \mathcal{C}$

A2. $A, B \in \mathcal{C}$ and $A \subseteq B \Rightarrow \setminus A \in \mathcal{C}$

A3. If (A_n) is an increasing sequence (take care to the word "increasing"!) of elements of $\mathcal{C}$ then $\bigcup\limits_{i=1}^{+\infty} A_i \in \mathcal{C}$ (stable by countable increasing union).

Remarks

R1. An increasing sequence of sets is: $A_1 \subseteq A_2 \subseteq A_3 \ldots$

R2. The first two axioms imply that $\mathcal{C}$ is by complementary.

R3. The three axioms together leads to that the monotonous class is stable by decreasing intersection. A way to check this to take the complement of each element of the increasing sequence to fall back on the decreasing sequence and vice versa.

Every σ-algebra is a monotone class, because σ-algebras are closed under arbitrary countable unions and intersections.

Therefore:

$$\mathcal{A} \subseteq \sigma\text{-algebra}(\mathcal{A}) \subseteq \mathcal{M}(\mathcal{A}) \tag{7.4.35}$$

In the same way as for the tribes, if we consider a family $(\mathcal{C}_i)_I$ of monotone classes on E. Then $\bigcap_I \mathcal{C}_i$ is a monotone class (the proof is verified immediately by the three previous axioms).

✍**Example:**

Given E a set, $\mathcal{P}(E)$ is a monotone class on E. More generally, a tribe is a monotone class.

Equivalently to tribes, let us consider a set E and $\mathcal{C} \subseteq \mathcal{P}(E)$. Given $\mathcal{S}$ the family of all monotone class containing $\mathcal{C}$, $\mathcal{S}$ is not empty because $\mathcal{P}(E) \in \mathcal{S}$. We denote by:

$$\mathcal{M}(\mathcal{C}) = \bigcap_{\mathcal{T} \in \mathcal{S}} \mathcal{T} \tag{7.4.36}$$

the monotone class generated by $\mathcal{C}$. Therefore $\mathcal{M}(\mathcal{C})$ is the smallest monotone class containing $\mathcal{C}$ (and satisfying obviously the previous axioms).

Remark

If E is a set and $\mathcal{C} \subseteq \mathcal{P}(E)$ then $\mathcal{C} \subseteq \sigma(\mathcal{C})$, as $\sigma(\mathcal{C})$ is a monotone class (and also a tribe) containing $\mathcal{C}$ and therefore contains also $\mathcal{M}(\mathcal{C})$ (see the examples with tribes).

The theorem (of the monotonic class) is stated as follows:

Theorem 7.11. Given E as set. If $\mathcal{C}$ is a family of parts of E that we impose as stable by finite intersection then $\mathcal{C} = \sigma\mathcal{C}$ (we then have to prove that the smallest tribe of $\mathcal{C}$ is equal to the smallest monotone class of $\mathcal{C}$). If we do not impose that $\mathcal{C}$ is stable by finite intersection we would not have necessarily the equality!

Proof 7.11.1. As already said: $\mathcal{M}(\mathcal{C}) \subseteq \sigma(\mathcal{C})$ (which is trivial). We will prove first that $\mathcal{M}(\mathcal{C})$ is a tribe on E. For this it is sufficient to show that $\mathcal{M}(\mathcal{C})$ is (also) stable by countable union (and not necessarily by an increasing sequence of elements!).

Let us considerate following families for the proof:

$$\begin{aligned}\mathcal{M}_1 &= \left\{A \in \mathcal{M}(\mathcal{C}) | A \cap B \in \mathcal{M}(\mathcal{C}), \forall B \in \mathcal{C}\right\} \\ \mathcal{M}_2 &= \left\{A \in \mathcal{M}(\mathcal{C}) | A \cap B \in \mathcal{M}(\mathcal{C}), \forall B \in \mathcal{M}(\mathcal{C})\right\}\end{aligned} \tag{7.4.37}$$

By the previous definitions $\mathcal{M}_1 \subseteq \mathcal{C}$ but $\mathcal{C}$ being (imposed) stable by finite intersections implies that $\mathcal{C} \subseteq \mathcal{M}_1$ and therefore (it is the same reasoning as for the tribes):

$$\mathcal{C} \subseteq \mathcal{M}_1 \subseteq \mathcal{M}(\mathcal{C}) \tag{7.4.38}$$

$\mathcal{M}_1$ is a monotone class, indeed $E \in \mathcal{M}_1$, if $A_1, A_2 \in \mathcal{M}_1$ and that $A_1 \subseteq A_2$ (second axiom) then:

$$\forall B \in \mathcal{C}, \underbrace{A_1 \cap B}_{\in \mathcal{M}(\mathcal{C})} \subseteq \underbrace{A_2 \cap B}_{\in \mathcal{M}(\mathcal{C})} \tag{7.4.39}$$

and therefore (which supports the fact that the other elements (A_n) satisfy the previous relation):

$$\forall B \in \mathcal{C}, (A_2 \cap B) \setminus (A_1 \cap B) \in \mathcal{M}(\mathcal{C}) \Rightarrow (A_2 \setminus A_1) \cap B \in \mathcal{M}(\mathcal{C}) \tag{7.4.40}$$

If (A_n) is an increasing sequence of elements of $\mathcal{M}_1$ then:

$$\forall B \in C, \left(\bigcup A_n\right) \cap B = \bigcup \underbrace{(A_n \cap B)}_{\in \mathcal{M}(\mathcal{C})} \in \mathcal{M}(\mathcal{C}) \tag{7.4.41}$$

as $(A_n \cap B)$ is an increasing sequence.

Therefore $\mathcal{M}_1$ is indeed a monotone class and by $\mathcal{C} \subseteq \mathcal{M}_1 \subseteq \mathcal{M}(\mathcal{C})$, we therefore have:

$$\mathcal{M}_1 = \mathcal{M}(\mathcal{C}) \tag{7.4.42}$$

The latter equality implies $\mathcal{C} \subseteq \mathcal{M}_2$. As for $\mathcal{M}_1$, we show that $\mathcal{M}_2$ is a monotone class and therefore $\mathcal{C} = \mathcal{M}_2$, which means by extension that $\mathcal{M}(\mathcal{C})$ is therefore stable by finite intersections.

$\mathcal{M}(\mathcal{C})$ being stable by complementary this takes us to that $\mathcal{M}(\mathcal{C})$ is, we just proved it, stable by finite unions (but we want to prove that it is stable by countable union!).

Given now a sequence (A_n) of elements of $\mathcal{M}(\mathcal{C})$. We consider the sequence:

$$B_n = \bigcup_{i=1}^{n} A_i \tag{7.4.43}$$

(B_n) is an increasing sequence of elements of $\mathcal{M}(\mathcal{C})$, therefore:

$$\bigcup_{n=1}^{+\infty} B_n \in \mathcal{M}(\mathcal{C}) \tag{7.4.44}$$

but:

$$\bigcup_{n=1}^{+\infty} B_n = \bigcup_{n=1}^{+\infty} A_n \tag{7.4.45}$$

Therefore:

$$\bigcup_{n=1}^{+\infty} A_n \in \mathcal{M}(\mathcal{C}) \tag{7.4.46}$$

Therefore $\mathcal{M}(\mathcal{C})$ is stable by countable union and finally $\mathcal{M}(\mathcal{C})$ is a tribe. But as $\mathcal{C} \subseteq \mathcal{M}(\mathcal{C})$ this brings us to $\mathcal{M}(\mathcal{C}) = \sigma(\mathcal{C})$.

■ Q.E.D.

We will later see some important applications of this theorem (but first we want to improve the above text with figure and more simple and practical examples!).

7.4.2 Lebesgue measure in probabilities

Probability Theory has become increasingly important in multiple parts of science (and therefore engineering too!) as we have already mentioned it during our study of the topic. Getting deeply into Probability Theory requires a full book, not just a chapter as we also already know. For readers who intend to pursue further studies in the roots of Probability Theory, this subsection will give you a good head start. For readers not intending to delve further into the roots of Probability Theory, this subsection will give you a taste of the subject.

Modern Probability Theory makes major use of Measure Theory. As we will see, a probability measure is simply a measure such that... the measure of the whole space equals 1. Thus a thorough understanding of the paragraphs that will follow dealing with Measure Theory and integration provides a solid foundation for Probability Theory!

Studying modern Probability Theory based on Measure Theory helps through rigorous proofs and definitions to explain to engineers or scientifically illiterate people typically why:

- A punctual estimation has a probability of zero in all problems involving variables (or combination of variables) having for domain of definition $\mathbb{R}$ and therefore such an estimation doesn't make sense rigorously speaking

- The previous point explain why we should always in such cases compute ranges of estimation that are plausible rather than punctual estimation that have zero probability
- A event having a probability of zero doesn't mean it can't happen
- That in an infinite universe with an infinite time a sequence of events has a probability of 1 to occur (see page 455)

Now that we have some basic measure theory, let's go back to the definition of "probability". A probability space is simply a measure space where the measure assigned to the entire object is one!

Definition (#314): A "probability space" W is a unique triple $W=\{\Omega,\mathcal{F},P\}$ where :

- Ω is its sample space
- $\mathcal{F}$ its σ-algebra of events
- P its probability measure

For reminder:

- The object Ω, named the "sample space", represents all the conceivable futures of the universe. Let's consider a simple coin flip. Before we flip the coin, there are a bunch of conceivable ways in which the coin spins through the air and settles on the table. The set of outcomes is a theoretical representation of all possible conceivable futures in which this process will unfold.
- The σ-algebra, is simply a way of partitioning all of these conceivable futures. Each of these partitions is named an "event". For example in some set of futures, the coin lands on heads - in which case, we would say that the event "heads" occurred. In other futures, the coin lands as "tails". If one of these futures occurs, then we say the event "tails" occurred.
- The probability measure assigns a degree of certainty towards each partition of the conceivable futures. Specifically, the probability measure assigns each event $\omega\in\Omega$ a value between 0 and 1 , which represents our degree of certainty that our future will be contained in ω. Specifically, 1 represents full certainty that our future will be contained in the event ω and 0 represents full certainty that our future will not be contained in ω.

Definition (#315): A "random variable" is a mapping:

$$X:\Omega\rightarrow\mathbb{R} \tag{7.4.47}$$

that assigns a real number $X(\omega)$ to each outcome ω.

All of this begs the question, why do we need Measure Theory for all of this? First we see that Measure Theory provides a natural description of uncertainty in the following ways:

- A σ-algebra gives us a way to ensure that if we are discussing our certainty towards an event occurring, we can also discuss our certainty towards the event not occurring. That is, the complement of any element of the σ-algebra is also in the σ-algebra.
- The σ-algebra ensures that we can discuss the probability of unions and intersections of events. The fact that $P(\varnothing):=0$ expresses the impossibility that no outcome in the sample space occurs. That is, we are certain to see some conceivable future.
- The fact that $P(\Omega):=1$ expresses that some outcome of the sample space will certainly occur. That is, 1 is the value that equates to maximal certainty.
- Uncertainty has a "size" in the following sense: if we have two disjoint events, then the probability of either of these events occurring should be the sum of the probabilities of the individual events. By using a measure space to model uncertainty, we ensure that given $A,B\subset\Omega$ where $A\cap B=\varnothing$, it follows that $P(A\cup B)=P(A)+P(B)$.

Remarks

R1. The sample "Universe" space Ω is the set of all possible samples or elementary events $\omega:\Omega=\{\omega\,|\,\omega\in\Omega\}$.

R2. The σ-algebra $\mathcal{F}$ is the set of all of the considered events A, i.e., subsets of $\Omega:\mathcal{F}=\{A\,|\,A\subseteq\Omega,A\in\mathcal{F}\}$.

R3. The probability measure P assigns a probability $P(A)$ to every event $A\in\mathcal{F}:P:\mathcal{F}\rightarrow[0,1]$. With $P(\Omega)=1$ and $P(\omega_i)=p_i$.

Suppose $n\in\mathbb{Z}^+$ and Ω is a set containing exactly n elements. Let $\mathcal{F}$ denote the collection of all subsets of Ω. Then:

$$\frac{\text{counting measure on }\Omega}{n}$$

is a probability measure P on $(\Omega,\mathcal{F})$. Obviously on uncountable sets like $\mathbb{R}$ or $\mathbb{R}^+$ that are frequent in engineering, we have $P(\omega)=0$.

✍**Example:**

As a more specific example of the previous item, suppose that $\Omega=\{40,41,\ldots,49\}$ and $P=(\text{counting measure on }\Omega)/10$. Let $A=\{\omega\in\Omega:\omega\text{ is even }\}$ and $B=\{\omega\in\Omega:\omega\text{ is prime }\}$. Then $P(A)$ [which is the probability that an element of this sample space Ω is even] is $\frac{1}{2}$ and $P(B)$ [which is the probability that an element of this sample space Ω is prime] is $\frac{3}{10}$.

Definition (#316): The "Borel σ-algebra" is the smallest σ-algebra (see page 2483) containing open interval in $\mathbb{R}$. The sets in $\mathcal{B}(\mathbb{R})$ are named "Borel sets". The extension to the multi-dimension case $\mathcal{B}(\mathbb{R}^n)$, is straightforward. The Borel σ-algebra therefore includes all subsets of $\mathbb{R}$ which are of interest in practical applications (scientific or engineering).

Let λ denote Lebesgue measure on $\mathbb{R}$, and let $\mathcal{B}$ denote the σ-algebra of Borel subsets of $\mathbb{R}$. Define a density distribution function $h:\mathbb{R}\rightarrow[0,+\infty[$. Then $hd\lambda$ is a probability measure on $(\mathbb{R},\mathcal{B})$.

Definition (#317): Suppose $(\Omega, \mathcal{F}, P)$ is a probability space. An event A is said to happen "almost surely" if the probability of A is 1, or equivalently if $P(\Omega \backslash A) = 0$.

7.4.2.1 Lebesgue Measure of the rational and irrational numbers

We can now begin to calculate the Lebesgue measure of the rational numbers! The calculation uses the first following lemma:

Lemma 7.11.1. If $X \subseteq Y$ are both measurable sets, $\mu(X) \leq \mu(Y)$. While this statement is quite intuitive (a subset should not be larger than it's superset), we give the short proof:

Proof 7.11.2. We may decompose Y as $Y = X \cup (Y \backslash X)$ (here $Y \backslash X$ is the relative complement of X in Y). Since X and $Y \backslash X$ are disjoint, by the third property (countable additivity) of the Lebesgue measure μ introduced earlier above (see page 2484):

$$\mu(Y) = \mu(X) + \mu(Y \backslash X) \geq \mu(X) \tag{7.4.48}$$

since $\mu(Y \backslash X) \geq 0$.

■ Q.E.D.

We now prove the next lemma:

Lemma 7.11.2. The Lebesgue measure of any singleton set, $\{x\}$, is zero.

Proof 7.11.3. This fact is intuitive, because a singleton is the smallest possible nonempty set; if we are to have any nonempty sets of measure zero, surely they must include the singletons.

Let x be any real number. The set $\{x\}$ is measurable, and for any $\varepsilon > 0$, $\{x\} \subset \left(x - \frac{\varepsilon}{2}, x + \frac{\varepsilon}{2}\right)$. Therefore:

$$\begin{aligned} 0 &\leq \mu(\{x\}) \\ &\leq \mu\left(\left(x - \frac{\varepsilon}{2}, x + \frac{\varepsilon}{2}\right)\right) \\ &= x + \frac{\varepsilon}{2} - \left(x - \frac{\varepsilon}{2}\right) \\ &= \varepsilon \end{aligned} \tag{7.4.49}$$

so $0 \leq \mu(\{x\}) \leq \varepsilon$. Since $\varepsilon > 0$ was arbitrary, we may let $\varepsilon \rightarrow 0$, which shows that $\mu(\{x\}) = 0$.

■ Q.E.D.

Finally, we are in a position to calculate the Lebesgue measure of the rational numbers:

Proof 7.11.4. Since the rationals are countable, we may enumerate them as $\mathbb{Q} = \cup_{i=1}^{+\infty} \{q_i\}$, so:

$$\mu(\mathbb{Q}) = \sum_{i=1}^{+\infty} \mu\left(\{q_i\}\right) = \sum_{i=1}^{+\infty} 0 = 0 \tag{7.4.50}$$

We have arrived at the remarkable (and counter-intuitive) fact that the Lebesgue measure of the rational numbers is zero. In a very precise sense, almost all real numbers are not rational.

■ Q.E.D.

The reader must keep in mind that we have this result because the set of rational numbers is countable. One of the properties of Lebesgue measure (as opposed to Peano-Jordan measure, for example) is that the measure of a countable union of disjoint sets is the sum of the measures of the individual sets. A single point has Lebesgue measure zero, so a countable set, being a countable union of singletons, also has Lebesgue measure zero!

Returning to the problem of comparing the size of the rational to the unit interval $[0,1]$, which motivated our introduction of measure theory above. The measure of the unit interval is $\mu([0,1])=1$, so the interval is larger than the rationals.

Remark

Diligent readers may have noticed that our proof that the rational numbers have measure zero may be applied to any countable subset of the real numbers. While this may seem to validate our initial idea that the small sets of the real are those that are countable, it is not true. The Cantor set (see section Fractals page 7049), a well-known example in analysis and topology, is uncountable and has measure zero. Therefore the class of small subsets of the real numbers is much larger than just the countable subsets.

Let P denote Lebesgue measure on the interval $[0,1]$. If $\omega\in[0,1]$, then ω is almost surely an irrational number (because the set of rational numbers has Lebesgue measure 0).

This example shows that an event having probability 1 (equivalent to happening almost surely) does not mean that the event definitely happens. Conversely, an event having probability 0 does not mean that the event is impossible. Specifically, if a real number is chosen at random from $[0,1]$ using Lebesgue measure as the probability, then the probability that the number is rational is 0, but that event can still happen.

In other words, zero probability does not mean that event is impossible. Every time you pick a random number from $[0,1]$, you have an infinite set of possible outcomes, so probability of picking a number from any finite subset of $[0,1]$ is equal to zero. However, impossible event is defined otherwise: It's an event which can't happen due to experiment conditions. In this case it's picking any number outside of $[0,1]$.

Thus, impossible and zero-probable events have the same probability of 0, however impossible events never happen, while zero-probable events are possible! This is sometimes named the "miracle theorem". That's why in science a "miracle" is associated to an impossible event but most people who don't have a graduate level in science education associate miracles (wrongly!) to zero-probable events... It follows from the fact that a length of segment $[1,1]$ (single point, i.e. "singleton") is 0, however point 1 (as well as any other point) exists.

At this point the reader can now understand what we stated earlier:

- A punctual estimation has a probability of zero in all problems involving variables (or combination of variables) having for domain of definition $\mathbb{R}$ and therefore such an estimation doesn't make sense rigorously speaking

- The previous point explain why we should always in such cases compute ranges of estimation that are plausible rather than punctual estimation that have zero probability

- An event having a probability of zero doesn't mean it can't happen

Caution! On the internet, and in particular on social networks (unsurprisingly...), propagates a disinformation which consists of asserting (obviously without proof!), that if a probability is sufficiently small, then the related event can be considered as having a probability strictly equal to zero and therefore be an impossible event. This is obviously false and we have therefore proven it above! It's also very easy to refute this dumb claim experimentally just by using a few dozens of hexagonal dice (among many other possibilities!).

> *The theory of probability [...] becomes the happiest supplement to the ignorance and weakness of the human mind. If we consider the analytical methods to which this theory gave birth, the truth of the principles which serve as its basis, the fine and delicate logic which their use requires in the solution of problems, the establishments of public utility which rely on it, and the extension which it has received and which it can still receive, by its application to the most important questions of natural philosophy and moral sciences; if we then observe that in the very things which can be subjected to calculation, it gives the surest insights which can guide us in our judgements, and that it learns to guard against the illusions which often lead us astray, we will see that there is no science more worthy of our meditations and that it is more useful to introduce into the system of public education.*

quote 7.4:2498.#59

Pierre-Simon de Laplace

~50 %

★★★★☆

17 votes, 76.47%

8

Geometry

Geometry is the mathematical discipline that focuses on the rigorous study of spaces and forms. (Larousse)

Contents

8.1 Trigonometry **2503**

- 8.1.1 Radian 2503
- 8.1.2 Circle Trigonometry 2508
  - 8.1.2.1 Remarkable trigonometric triangle identities 2519
- 8.1.3 Hyperbolic Trigonometry 2532
  - 8.1.3.1 Remarkable hyperbolic identities 2539
- 8.1.4 Spherical Trigonometry 2542
  - 8.1.4.1 Girard's theorem 2546
  - 8.1.4.2 Four-parts formula 2548
- 8.1.5 Solid Angle 2549

8.2 Euclidean Geometry **2553**

- 8.2.1 Objects of Euclidean Geometry 2553
  - 8.2.1.1 Dimensions 2555
- 8.2.2 Euclid's Constructions 2563
  - 8.2.2.1 Segments and Lines 2566
- 8.2.3 Plane Geometry 2573
  - 8.2.3.1 Displacements and Turnarounds 2574
  - 8.2.3.2 Plane angles 2575
  - 8.2.3.3 Triangles 2593
  - 8.2.3.4 Parallelism 2618
  - 8.2.3.5 Circle 2620
- 8.2.4 Hilbert's Axioms 2631

8.2.4.1 Incidence Axioms (axioms of association) 2632
8.2.4.2 Order Axioms 2632
8.2.4.3 Congruence Axioms 2633
8.2.4.4 Continuity Axioms 2633
8.2.4.5 Parallels Axioms 2634
8.2.5 Barycenter (centroid) 2635
8.2.6 Geometric Transformations 2640
8.2.6.1 Translation 2642
8.2.6.2 Homothety (scaling) 2643
8.2.6.3 Shear (skew) transformation 2645
8.2.6.4 Rotation 2647
8.2.6.5 Reflection 2659
8.3 Non-Euclidean Geometry 2663
8.3.1 Curvature(s) 2665
8.3.2 Axioms of non-euclidean geometry 2667
8.3.3 Generalized angle sum theorems 2668
8.3.4 Geodesic and Metric Equation 2669
8.3.5 Riemann Spaces 2673
8.4 Projective Geometry 2679
8.4.1 Conical Perspective (Central Perspective) 2680
8.4.1.1 Images of Points 2684
8.4.1.2 Images of Straight Lines 2701
8.4.2 Affine projections 2712
8.4.2.1 Isometric perspective 2714
8.4.2.2 Oblique perspective 2720
8.4.2.3 Orthogonal projection 2722
8.4.3 Spherical projections 2727
8.4.3.1 Stereographic projection 2729
8.4.3.2 Cylindrical projection 2738
8.4.3.3 Mercator projection 2739
8.4.3.4 Lambert's equivalent projection (Peters projection) 2743
8.4.3.5 Mollweide projection 2746
8.4.4 Other perspectives and Laws of linear perspective 2751
8.4.5 Homogeneous Coordinates (projection coordinates) 2759
8.4.5.1 $\mathcal{P}^2$ Projective Space 2759
8.4.5.2 $\mathcal{P}^3$ Projective Space 2761
8.5 Analytical Geometry 2764

8.5.1 Conics 2764
8.5.1.1 Algebraic approach 2765
8.5.1.2 Geometric Approach 2775
8.5.1.3 Dandelin Theorem (Dandelin spheres) 2793
8.5.1.4 Classification of conical by the determinant 2794
8.5.2 Parametrizations 2800
8.5.2.1 Equation of the Plane 2800
8.5.2.2 Equation of the Straight line 2804
8.5.2.3 Equation of a Circle 2810
8.5.2.4 Equation of a Square 2812
8.5.2.5 Equation of an Hexagon 2813
8.5.2.6 Equation of a Cycloid 2814
8.5.2.7 Equation of an Epicycloid 2816
8.5.2.8 Equation of a Spiral 2818
8.5.2.9 Equation of a Hypocycloid 2821
8.5.2.10 Surface of revolution 2823
8.6 Differential Geometry 2850
8.6.1 Parametric Curves 2850
8.6.2 Isolines 2857
8.6.3 Frenet Frame 2863
8.6.4 Surface Patches 2878
8.6.4.1 Metric of a Surface Patch 2879
8.7 Geometric Shapes 2886
8.7.1 Usual Surfaces (areas) 2888
8.7.1.1 Polygons 2888
8.7.1.2 Rectangle 2894
8.7.1.3 Square 2895
8.7.1.4 Unspecified Triangle 2896
8.7.1.5 Isosceles Triangle 2900
8.7.1.6 Equilateral Triangle 2902
8.7.1.7 Right Triangle 2905
8.7.1.8 Trapezoid 2906
8.7.1.9 Parallelogram 2908
8.7.1.10 Hexagon 2909
8.7.1.11 Rhombus 2914
8.7.1.12 Circle 2915
8.7.1.13 Ellipse 2917

8.7.2 Usual Volumes . . . 2923
8.7.2.1 Polyhedron . . . 2924
8.7.2.2 Usual Solids of Revolution . . . 2956
8.7.3 Isoperimetric inequality (2D) . . . 2993
8.8 Graph Theory . . . 2999
8.8.1 Type of Graphs and Structures . . . 3000
8.8.2 Graph Adjacency Matrix . . . 3026
8.8.3 Categories . . . 3031
8.9 Knot Theory . . . 3035
8.9.1 Braids Representation . . . 3035
8.9.1.1 Braids Group . . . 3037
8.9.2 Knot Representation . . . 3040
8.9.2.1 Knots Group . . . 3044
8.9.3 Tait's Knot . . . 3048
8.9.4 Mathematical Formalisation . . . 3052
8.9.4.1 Planar Representation . . . 3061

8.1. Trigonometry

TRIGONOMETRY is part of the science of geometry. Geometry whose etymological root is "measure of the Earth", trigonometry has etymological root "measuring of body with three angles (trine)". Thus the trigonometry is a branch of mathematics that studies relations involving lengths and angles of triangles.

There are currently four known trigonometries (defined) commonly used in mathematics: the "circle trigonometry" (assimilated with the study of "circular functions"), the "hyperbolic trigonometry" and "spherical trigonometry" and the more exotic "elliptical trigonometry" (related to the elliptic integrals and their related elliptic function). We propose in the text that follows an attempt to a fairly rigorous approach to all the most famous relations of the three first type of trigonometries (the last one - the elliptical trigonometry - will be reserved to another section of this textbook).

Remarks

R1. We will not deal here on the quadratic and rhombic trigonometries that are used by the electricians and have little or no interest in theoretical physics. The same remark is valid for the lemniscatique trigonometry which is related to pure mathematics and in particular to the Riemann zeta function.

R2. The reader who is looking for the proofs of derivatives and integrals of trigonometric functions defined below should read the Differential and Integral Calculus section (see chapter Algebra page 1450) where the derivatives and integrals of common trigonometric functions are all proven.

The purpose of this section will be to determine the most common relations in trigonometry and that are used extensively in all sections of this textbook (Mechanics, Astronomy, Statistics, Atomistic, etc.). Note that the majority of the relations (but not all!) that we will define and prove were determined in the 116th century (holocene calendar) by algebraists like Viète.

8.1.1 Radian

When we speak about trigonometry, the first thing that should come to mind and emerge as standard for plan angles measurements (see the section of Euclidean Geometry chapter for the definition of the concept of angle) is the notion of "radians ".

Definition (#8.1.318): 1 "radian" (denoted sometimes [rad]) is the angle described by a plane secant to a circle passing through its center as the arc thus defined by the horizontal axis passing through the center of the circle and the secant is equal in length to the radius of this circle.

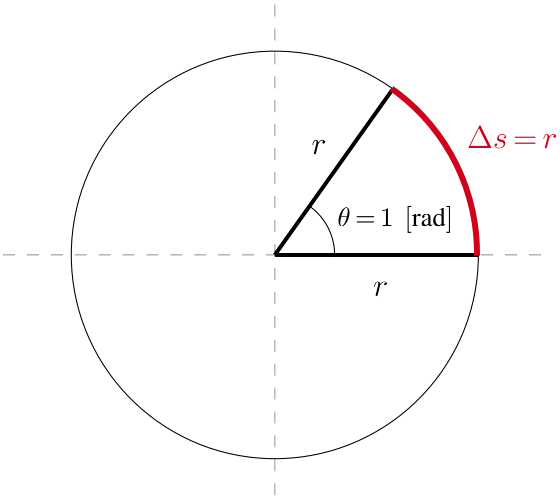


Figure 8.1: **Schematic definition of the radian**

Remark

The radian is widely used in physics, as we will see it in the corresponding chapters of this textbook, when angular measurements are required. For example, angular velocity is typically measured in radians per second ([rad/s]). One revolution per second is equal to 2π radians per second.

For example, for a circle of radius $R=1$ therefore of circumference (or "perimeter") $P=2\pi R = 2\pi$ the length of the arc defined by a secant having an angle of 1 radian with respect to the horizontal passing through the center of the circle will be equal 1.

Hence it follows that the angle for a "round" of the circle will be:

$$2\pi \text{ [rad]} \tag{8.1.1}$$

This can be summarized in the following drawing:

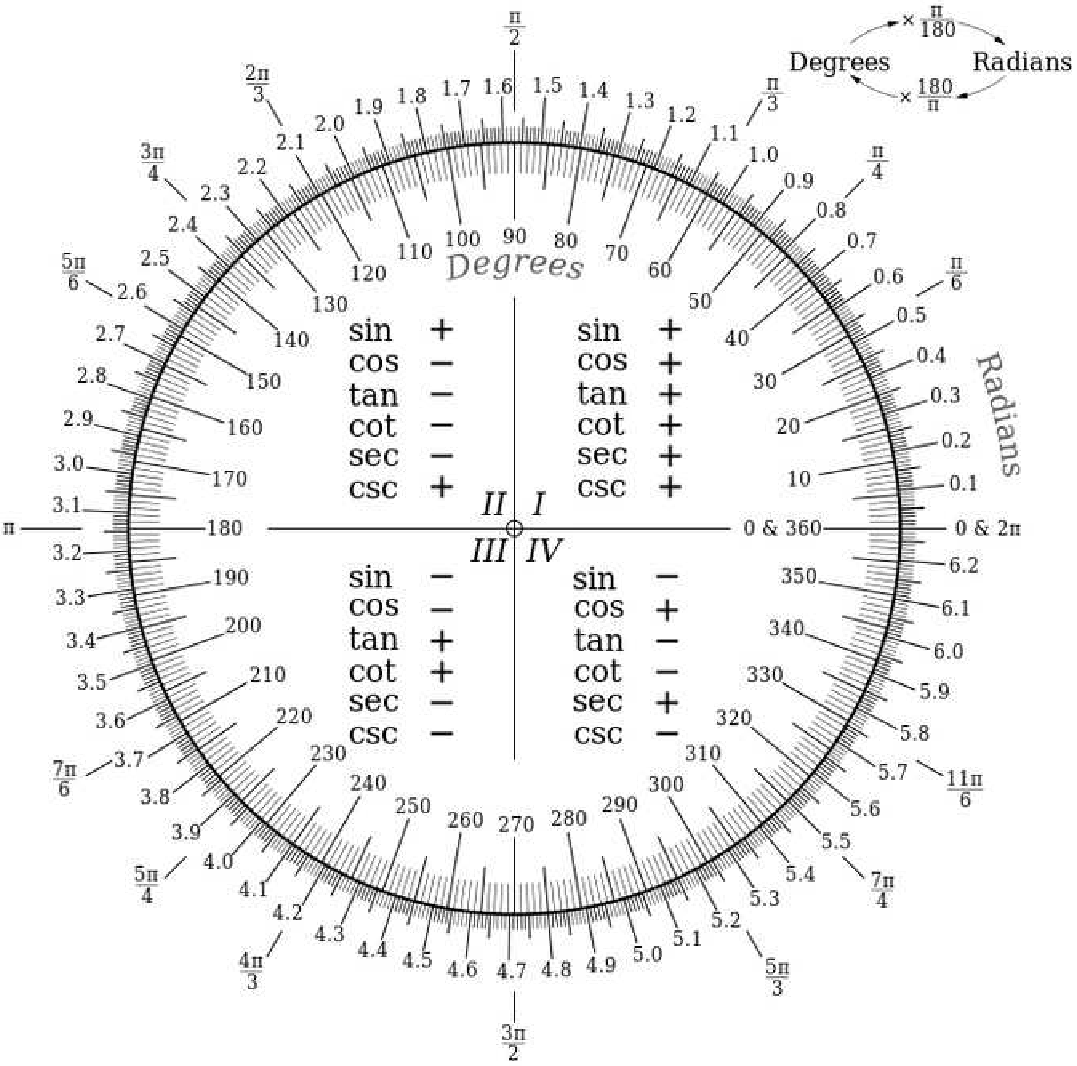

Degrees
Radians
× π/180
× 180/π
sin +
cos −
tan −
cot −
sec −
csc +
sin +
cos +
tan +
cot +
sec +
csc +
sin −
cos −
tan +
cot +
sec −
csc −
sin −
cos +
tan −
cot −
sec +
csc −
II
I
III
IV
0 & 360
0 & 2π
π/2
π
3π/2

Figure 8.2: **A chart to convert between degrees and radians (source: Wikipedia)**

The previous example can be generalized to any circle of radius R as the angle for a complete tour will always be for 2π[rad] a whole turn-around, π[rad] for a half turn-around and $\dfrac{\pi}{2}$[rad] for fourth turn-around...

Unfortunately in high-schools most teachers still teach children to measure angles in degrees. Fortunately conversion to do is not too difficult ... (it's a simple rule of three):

$$\boxed{\text{angle in radians} = \text{angle in degrees} \cdot \frac{\pi}{180}} \tag{8.1.2}$$

Let r be the measurement of an angle in radians, d the measurement of the same angle in degrees, g and the measurement of the same angle in gradients (old unit) we have by definition:

$$\boxed{\frac{r}{2\pi} = \frac{d}{360} = \frac{g}{400}} \tag{8.1.3}$$

Astronomers and astrophysicists like to talk in minutes or seconds of arc such as:

$$1^\circ = 60' \text{ minutes arc} \qquad 1' = 60'' \text{ seconds arc} \tag{8.1.4}$$

For example if the Moon has a diameter angle (i.e. angular diameter) of approximately $29'$ ($0.5°$) in the sky and if our eyes would have a deep exposure capability we would be able to see that Andromeda (M31) has a diameter angle of $178'$ ($3°$), therefore about 6 times as wide as the Moon!

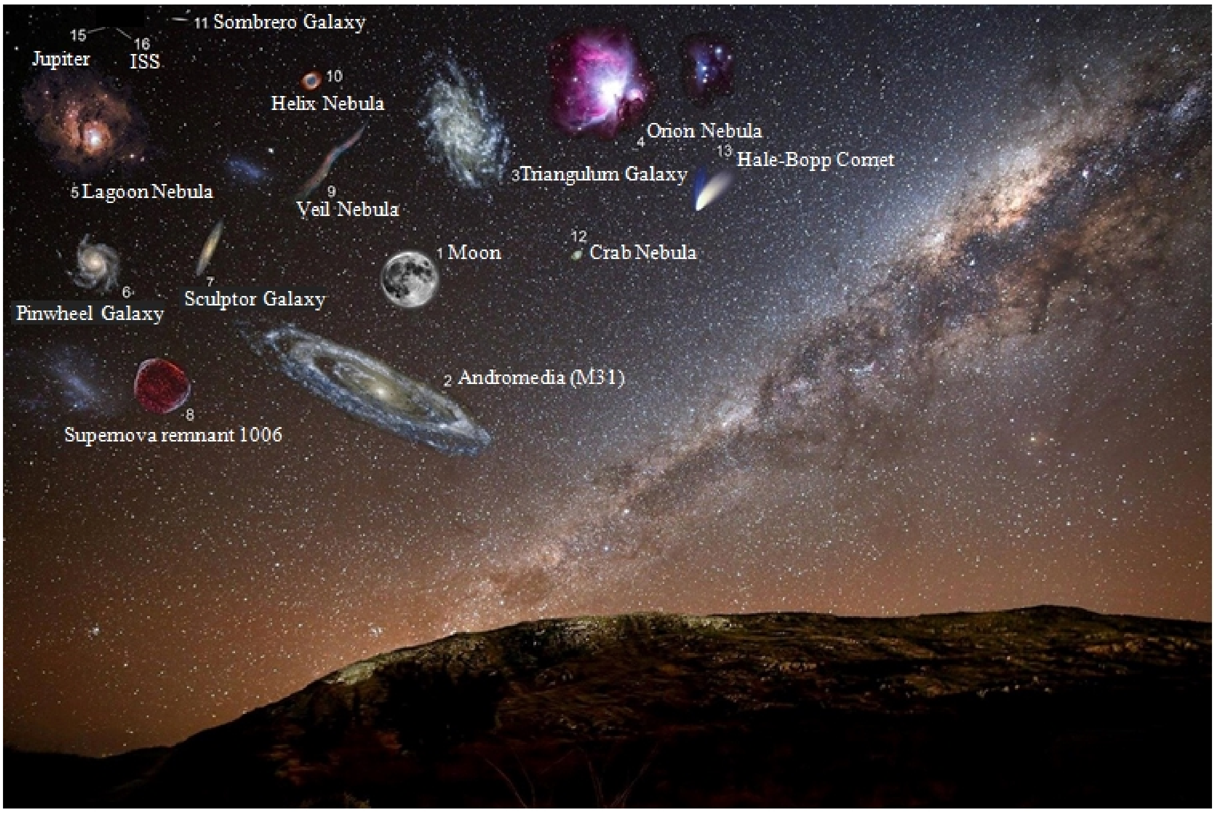

15 Jupiter
16 ISS
11 Sombrero Galaxy
10 Helix Nebula
4 Orion Nebula
13 Hale-Bopp Comet
3 Triangulum Galaxy
5 Lagoon Nebula
9 Veil Nebula
1 Moon
12 Crab Nebula
6 Pinwheel Galaxy
7 Sculptor Galaxy
2 Andromeda (M31)
8 Supernova remnant 1006

Figure 8.3: Earth's sky as seen if human eyes had deep exposure capabilities

Here is a nice visual summarizing this concept of sub-angles units:

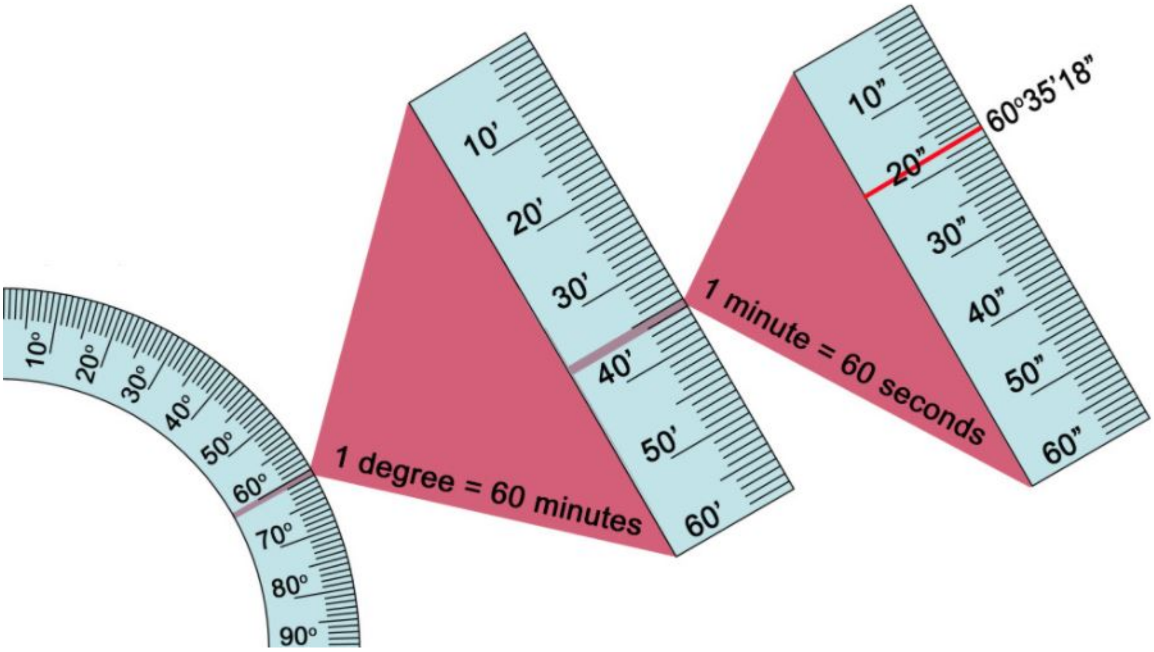

10°
20°
30°
40°
50°
60°
70°
80°
90°
1 degree = 60 minutes
10'
20'
30'
40'
50'
60'
1 minute = 60 seconds
10"
20"
30"
40"
50"
60"
60°35'18"

Figure 8.4: Degrees, minutes of arc, seconds of arc (author: ?)

✍**Example:**

Imagine you are observing at the zenith (at your perpendicular from the ground) a huge sphere of 51 kilometers in radius and at distance of $5,000$ kilometres.

The angular size of this object when it is at your zenith is given by:

$$\alpha = \tan\left(\frac{51}{5000}\right) \cong 0.01 \cong 0.573^\circ = 0^\circ 34' 22.8" \tag{8.1.5}$$

Now consider this same object at the ground level but further away from you at a distance of $20,000$ kilometres. Now its angular size is:

$$\alpha = \tan\left(\frac{51}{5000}\right) \cong 0.0025 \cong 0.143^\circ = 0^\circ 8' 34.8" \tag{8.1.6}$$

Hence its size between the two positions change of approximately a factor 4.

8.1.2 Circle Trigonometry

Consider the figure below showing any circle centered at the origin in a direct basis:

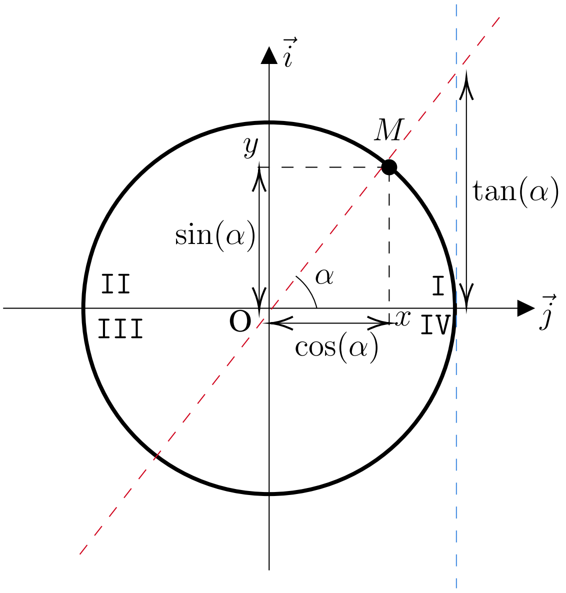


Figure 8.5: **Construction idea of elementary (circle) trigonometric functions**

First, by applying the Pythagorean theorem (see section Euclidean Geometry page 2609) in the visible right triangle above, we get:

$$\|\overrightarrow{\mathrm{O}M}\|=R=\sqrt{x^2+y^2} \tag{8.1.7}$$

where R is the radius of the circle.

To define the trigonometric functions for the angle α, start with any right triangle that contains the angle α. The three sides of the triangle are named as follows:

1. The "hypotenuse" is the side opposite the right angle, in this case the side is the segment $\overline{\mathrm{O}M}$. The hypotenuse is always the longest side of a right-angled triangle.
2. The "opposite side" is the side opposite to the angle we are interested in.
3. The "adjacent side" is the side adjacent to the angle we are interested in.

From this representation we can define mathematical entities named "trigonometric functions of the circle", also sometimes named by the ancient (...) "cyclometrics functions", such as (for the most known one):

$$\boxed{\begin{array}{lll}\cos(\alpha)=\dfrac{x}{R}=\dfrac{\text{adjacent}}{\text{hypotenuse}} & \arccos\left(\dfrac{x}{R}\right)=\alpha & \sec(\alpha)=\dfrac{1}{\cos(\alpha)}=\cos^{-1}(\alpha)\\ \sin(\alpha)=\dfrac{y}{R}=\dfrac{\text{opposite}}{\text{hypotenuse}} & \arcsin\left(\dfrac{y}{R}\right)=\alpha & \csc(\alpha)=\dfrac{1}{\sin(\alpha)}=\sin^{-1}(\alpha)\\ \tan(\alpha)=\dfrac{y}{x}=\dfrac{\text{opposite}}{\text{adjacent}} & \arctan\left(\dfrac{y}{x}\right)=\alpha & \cot(\alpha)=\dfrac{1}{\tan(\alpha)}=\tan^{-1}(\alpha)\end{array}} \tag{8.1.8}$$

Be careful because depending on the authors and on the context (and that's our case!) $\arccos, \arcsin, \arctan$, are respectively denoted by $\cos^{-1}, \sin^{-1}, \tan^{-1}$.

Read "cosine" for "cos", "sine" for "sin", "tangent" for "tan", "cotangent" for "cot", "secant" for "sec", "cosecant" for "csc".

"SOHCAHTOA" is a mnemonic acronym used to remember the three main trigonometric ratios (sine, cosine, tangent) in a right-angled triangle. It breaks down as Sine = Opposite/Hypotenuse (SOH) ; Cosine = Adjacent/Hypotenuse (CAH), and Tangent = Opposite/Adjacent (TOA).

Remarks

R1. When the context allows it, and that there can not be any ambiguity, the parentheses after the name of the trigonometric function may be omitted (this is often the case in physics).

R2. The arc... functions are obviously the reciprocal functions of the trigonometric functions!

In practice (school or high level engineering) you have to take care to two important properties:

1. Trigonometric functions are far from being bijective. Indeed, they tend to repeat themselves periodically as we will see further below. Nonetheless with a little restriction on their definition domain, they may acquire this quality. In this context, the perspective of a reciprocal becomes possible.

2. If the sign of the bracket of the arc tangent is negative, then we do not know exactly in which quadrant (I, II, III or IV) we are so we do not know the sign of the numerator or denominator! It is for this reason that most calculators and softwares have two arc tangent functions: `arctan` and `arctan2`. The first gives an angle between $\left[-\dfrac{\pi}{2},+\dfrac{\pi}{2}\right]$ and therefore we can not know precisely in which quadrant we are. The second gives us an angle between $[-\pi,\pi]$ so we can know in which quadrant we are but we must provide the sign of the numerator or denominator.

From these functions, we can make some combinations and define simple but remarkable relations but whom usefulness is debatable (and which are very rarely used) such as:

$$\begin{aligned}\text{covers}(\alpha)&=1-\sin(\alpha)\\ \text{vers}(\alpha)&=1-\cos(\alpha)\\ \text{cexc}(\alpha)&=\csc(\alpha)-1\\ \text{exsec}(\alpha)&=\sec(\alpha)-1\end{aligned} \tag{8.1.9}$$

but you will never meet these functions in this textbook because we personally never use such notations (it is rather customary in some U.S. books).

Here is a nice figure... that resume all that stuff:

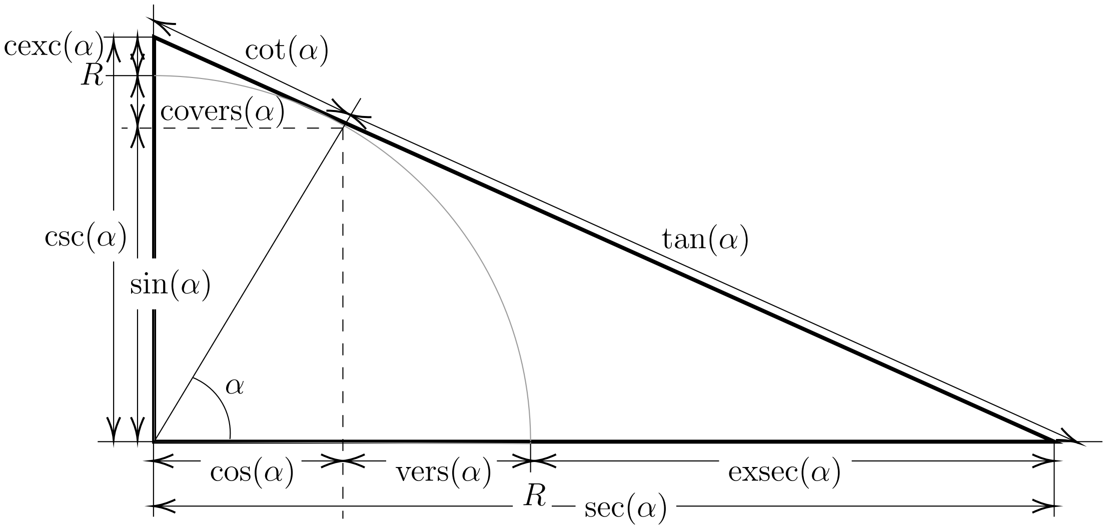
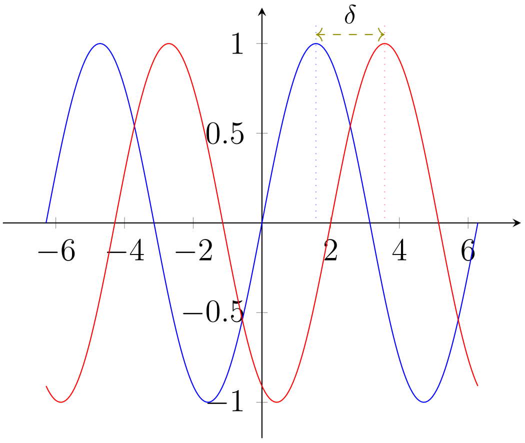

cexc(α)
R
cot(α)
covers(α)
csc(α)
sin(α)
tan(α)
α
cos(α)
vers(α)
R
exsec(α)
sec(α)

Figure 8.6: **Design principle of common circle trigonometric functions**

Definition (#8.1.319): We name "phase-shift" the angle δ that shift any trigonometric function argument of a given value. For example as illustrated below:

$$\sin(x+\delta) \tag{8.1.10}$$



OK for people reading this textbook on a computer with Adobe Flash player installed and activated, here is an animated version (otherwise see here: `https://vimeo.com/575748635`):

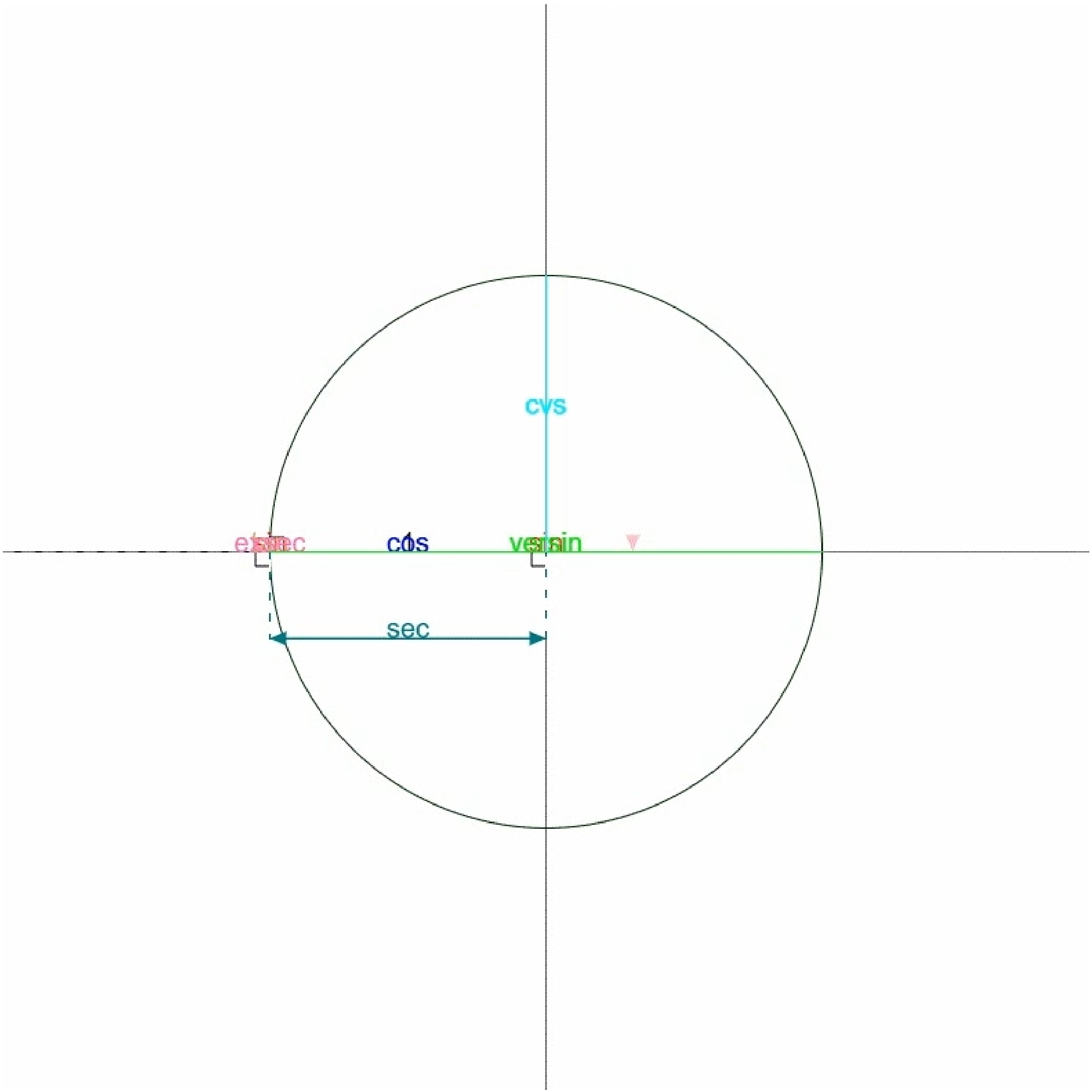


Figure 8.7: **Animation of the trigonometric circle**

Let us see now some very simple angular properties of these common functions:

P1. If we remain in the study of the circle, hence named "trigonometric circle", we must put for the above definitions: $R=1$. Thus, it appears more clearly the physical meaning of these definitions and it will result in a number of properties and applications directly usable in theoretical physics and pure mathematics.

Indeed, if $R=1$ we trivially have (see section Vector Calculus page 1863):

$$\overrightarrow{\text{O}M}=\cos(\alpha)\vec{i}+\sin(\alpha)\vec{j} \tag{8.1.11}$$

and by applying the Pythagorean theorem (see section Euclidean Geometry page 2609):

$$\|\overrightarrow{\text{O}M}\|=1=\sqrt{\cos^2(\alpha)+\sin^2(\alpha)} \tag{8.1.12}$$

Therefore:

$$\boxed{\cos^2(\alpha)+\sin^2(\alpha)=1} \tag{8.1.13}$$

P2. If α is a real number, and for $\forall k \in \mathbb{N}$, the real numbers α and $\alpha+2k\pi$ are associated with the same point M by the periodicity of the unit circle. Indeed, α and $\alpha+2k\pi$ are two measures of the same oriented angle. So:

$$\forall\alpha\in\mathbb{R},\forall k\in\mathbb{N}\quad \cos(\alpha\pm 2k\pi)=\cos(\alpha)\text{ and }\sin(\alpha\pm 2k\pi)=\sin(\alpha) \tag{8.1.14}$$

Ditto for all trigonometric functions arising from the definition of sine and cosine functions.

Remark

In the measure of "oriented angles", we say that two measures are congruent modulo $2k\pi$ [rad] if and only if their difference is a multiple of $2k\pi$ [rad]. This characterize two measurements of the same angle.

By definition, the sine and cosine of any real number part of the interval $[-1,+1]$. Specifically, the position of the point M allows us to learn more about the cosine and sine of α. So:

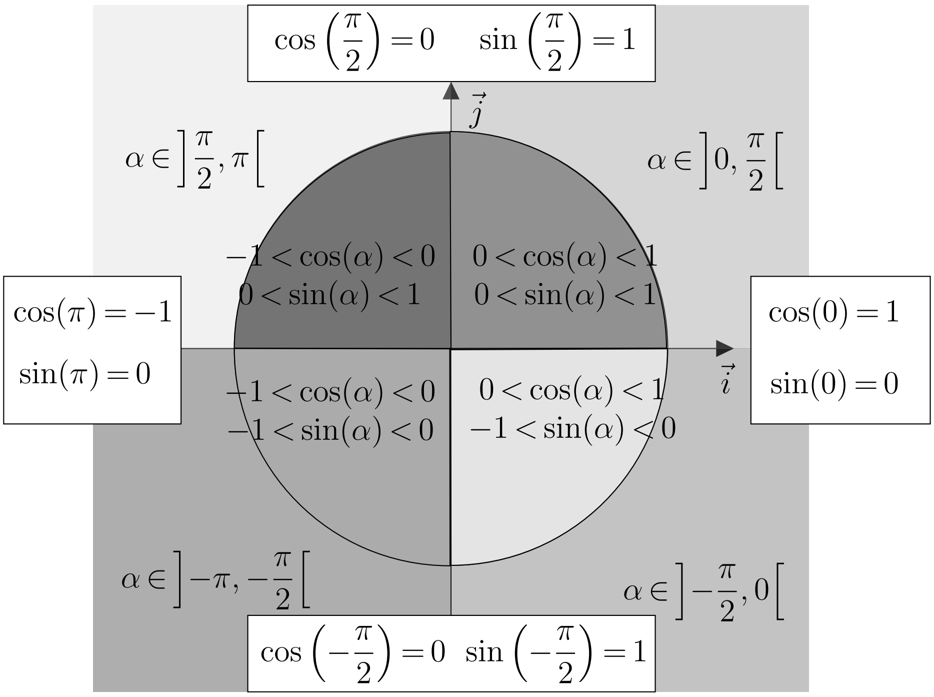

$\cos\left(\frac{\pi}{2}\right)=0 \quad \sin\left(\frac{\pi}{2}\right)=1$
$\vec{j}$
$\alpha\in\left]\frac{\pi}{2},\pi\right[$
$\alpha\in\left]0,\frac{\pi}{2}\right[$
$-1<\cos(\alpha)<0$
$0<\sin(\alpha)<1$
$0<\cos(\alpha)<1$
$0<\sin(\alpha)<1$
$\cos(\pi)=-1$
$\sin(\pi)=0$
$\cos(0)=1$
$\sin(0)=0$
$\vec{i}$
$-1<\cos(\alpha)<0$
$-1<\sin(\alpha)<0$
$0<\cos(\alpha)<1$
$-1<\sin(\alpha)<0$
$\alpha\in\left]-\pi,-\frac{\pi}{2}\right[$
$\alpha\in\left]-\frac{\pi}{2},0\right[$
$\cos\left(-\frac{\pi}{2}\right)=0 \quad \sin\left(-\frac{\pi}{2}\right)=1$

Figure 8.8: **Remarkable angles of common circle trigonometric functions**

There is also another representation of the trigonometric functions of the circle, a bit more technical but almost important to understand well later, at least visually, for the study of wave mechanics (see section of the same name page 3475):

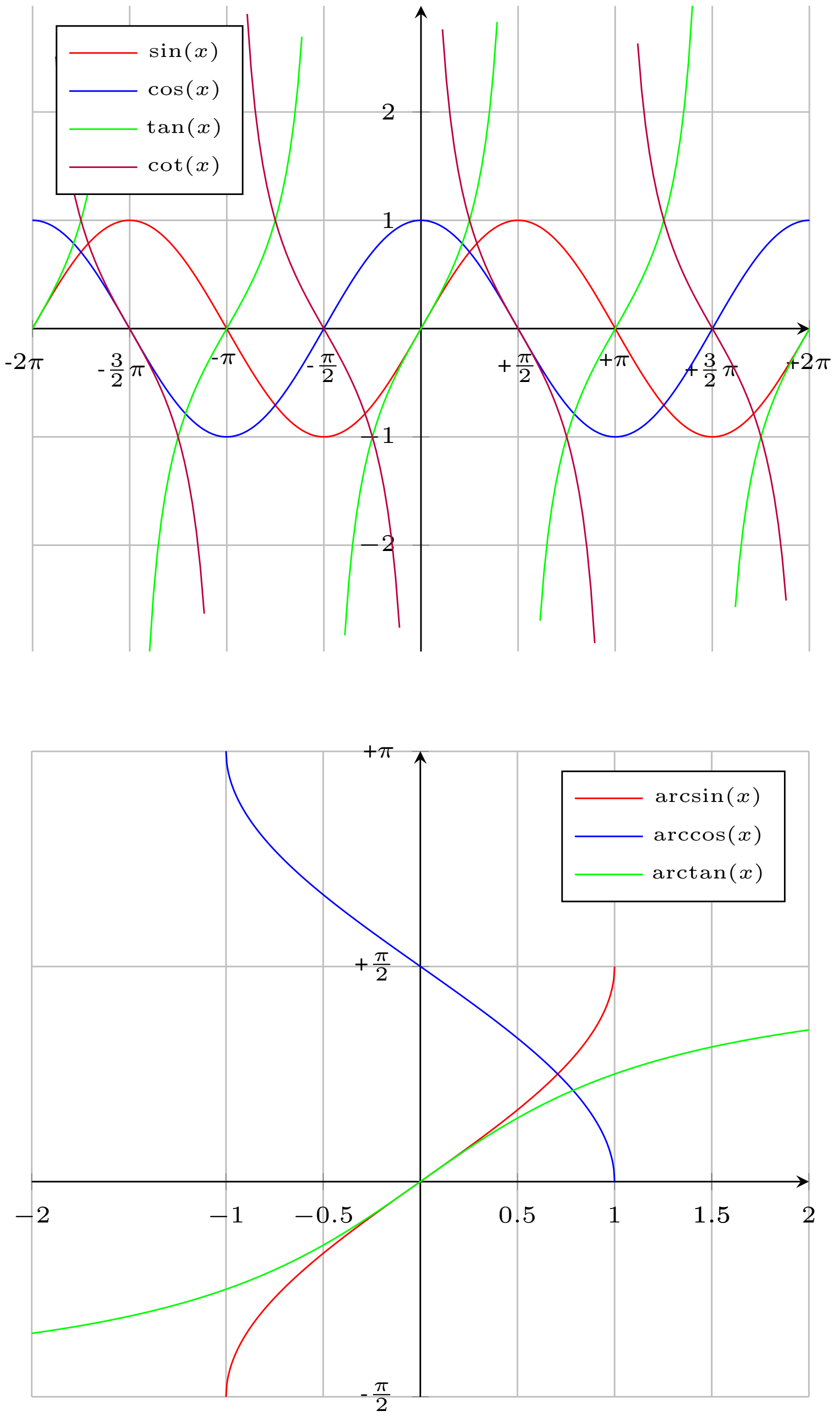


Figure 8.9: Plots of some common circle trigonometric functions

The reader should notice at this point without too much difficulties the following properties (often used in physics!) with $n \in \mathbb{Z}$:

$$\begin{aligned} \sin(n\pi) &= 0 \\ \cos(n\pi) &= (-1)^n \\ \sin\left((2n+1)\frac{\pi}{2}\right) &= (-1)^n \\ \cos\left((2n+1)\frac{\pi}{2}\right) &= 0 \end{aligned} \tag{8.1.15}$$

and easily recognize that the sine is an odd function and the cosine an even function (observation that we often will be useful in various mathematical developments on trigonometric series).

As we saw at the beginning of this section, following the definition of the trigonometric functions we obviously have:

$$\tan(\alpha)=\frac{\sin(\alpha)}{\cos(\alpha)}\qquad \cot(\alpha)=\frac{1}{\tan(\alpha)}=\frac{\cos(\alpha)}{\sin(\alpha)}\tag{8.1.16}$$

and also:

$$\cos^2(\alpha)+\sin^2(\alpha)=1\Leftrightarrow\frac{\cos^2(\alpha)}{\sin^2(\alpha)}+1=\frac{1}{\sin^2(\alpha)}\Leftrightarrow\cot^2(\alpha)+1=\frac{1}{\sin^2(\alpha)}\tag{8.1.17}$$

In exactly the same way we prove that:

$$\tan^2(\alpha)+1=\frac{1}{\cos^2(\alpha)}\tag{8.1.18}$$

From previous equalities we found without too much difficulty (elementary algebra):

$$\sin(\alpha)=\sqrt{1-\cos^2(\alpha)}=\sqrt{1-\frac{1}{1+\tan^2(\alpha)}}=\frac{\tan(\alpha)}{\sqrt{1+\tan^2(\alpha)}}=\frac{1}{\sqrt{1+\cot^2(\alpha)}}\tag{8.1.19}$$

We have identically:

$$\cos(\alpha)=\sqrt{1-\sin^2(\alpha)}=\sqrt{1-\frac{1}{1+\cot^2(\alpha)}}=\frac{\cot(\alpha)}{\sqrt{1+\cot^2(\alpha)}}=\frac{1}{\sqrt{1+\tan^2(\alpha)}}\tag{8.1.20}$$

By reverse reasoning we get just as easily:

$$\begin{aligned}\tan(\alpha)&=\frac{\sin(\alpha)}{\sqrt{1-\sin^2(\alpha)}}=\frac{\sqrt{1-\cos^2(\alpha)}}{\cos(\alpha)}\\\cot(\alpha)&=\frac{\cos(\alpha)}{\sqrt{1-\cos^2(\alpha)}}=\frac{\sqrt{1-\sin^2(\alpha)}}{\sin(\alpha)}\end{aligned}\tag{8.1.21}$$

It could comes without difficulty by observing the unit circle with its remarkable angles that:

$$\begin{aligned}&\sin(-\alpha)=-\sin(\alpha)\\&\sin(\pi\pm\alpha)=\mp\sin(\alpha)\\&\sin\left(\frac{\pi}{2}\pm\alpha\right)=\cos(\alpha)\\[2ex]&\cos(-\alpha)=\cos(\alpha)\\&\cos(\pi\pm\alpha)=-\cos(\alpha)\\&\cos\left(\frac{\pi}{2}\pm\alpha\right)=\mp\sin(\alpha)\\[2ex]&\tan(-\alpha)=-\tan(\alpha)\\&\tan(\pi\pm\alpha)=\pm\tan(\alpha)\\&\tan\left(\frac{\pi}{2}\pm\alpha\right)=\mp\cot(\alpha)\end{aligned}\tag{8.1.22}$$

Or summarized as a table for only the simplest one:

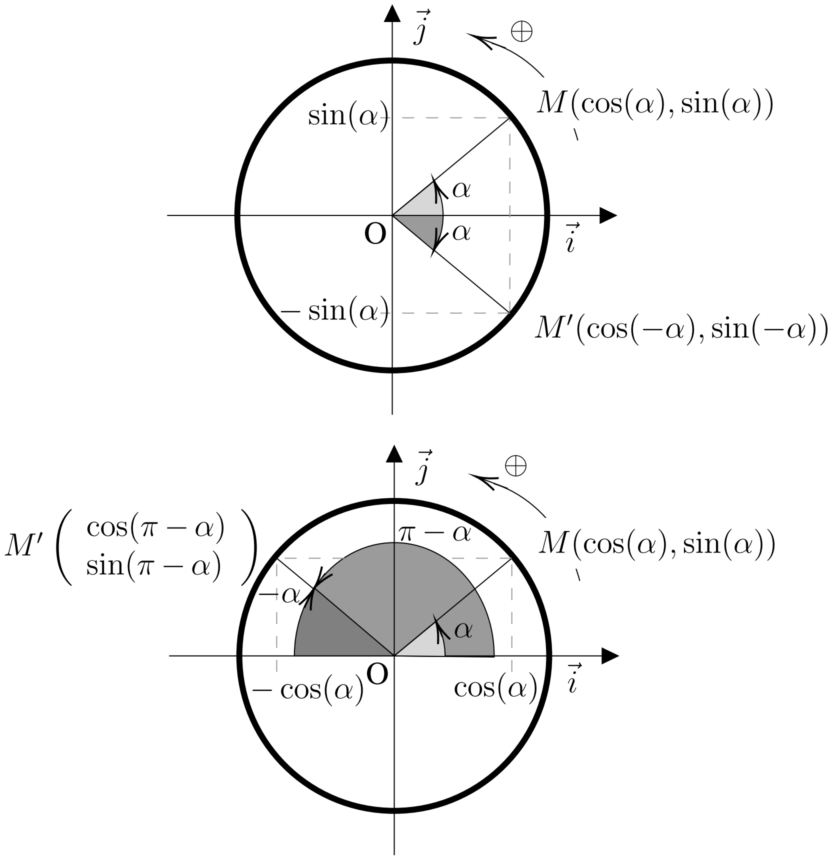
Trigonometric Functions	
Even Functions $f(-x)=f(x)$	**Odd Functions** $f(-x)=-f(x)$
$\cos(-x)=\cos(x)$ $\sec(-x)=\sec(x)$	$\sin(-x)=-\sin x$ $\csc(-x)=-\csc(x)$ $\tan(-x)=-\tan(x)$ $\cot(-x)=-\cot(x)$

Table 8.1: **Odd and even trigonometric functions**

Remark

The minus-plus sign "$\mp$" is generally used in conjunction with the "$\pm$" sign in a same relation. So when the first is at the "$+$" level, then the other one is at "$-$" level and conversely.

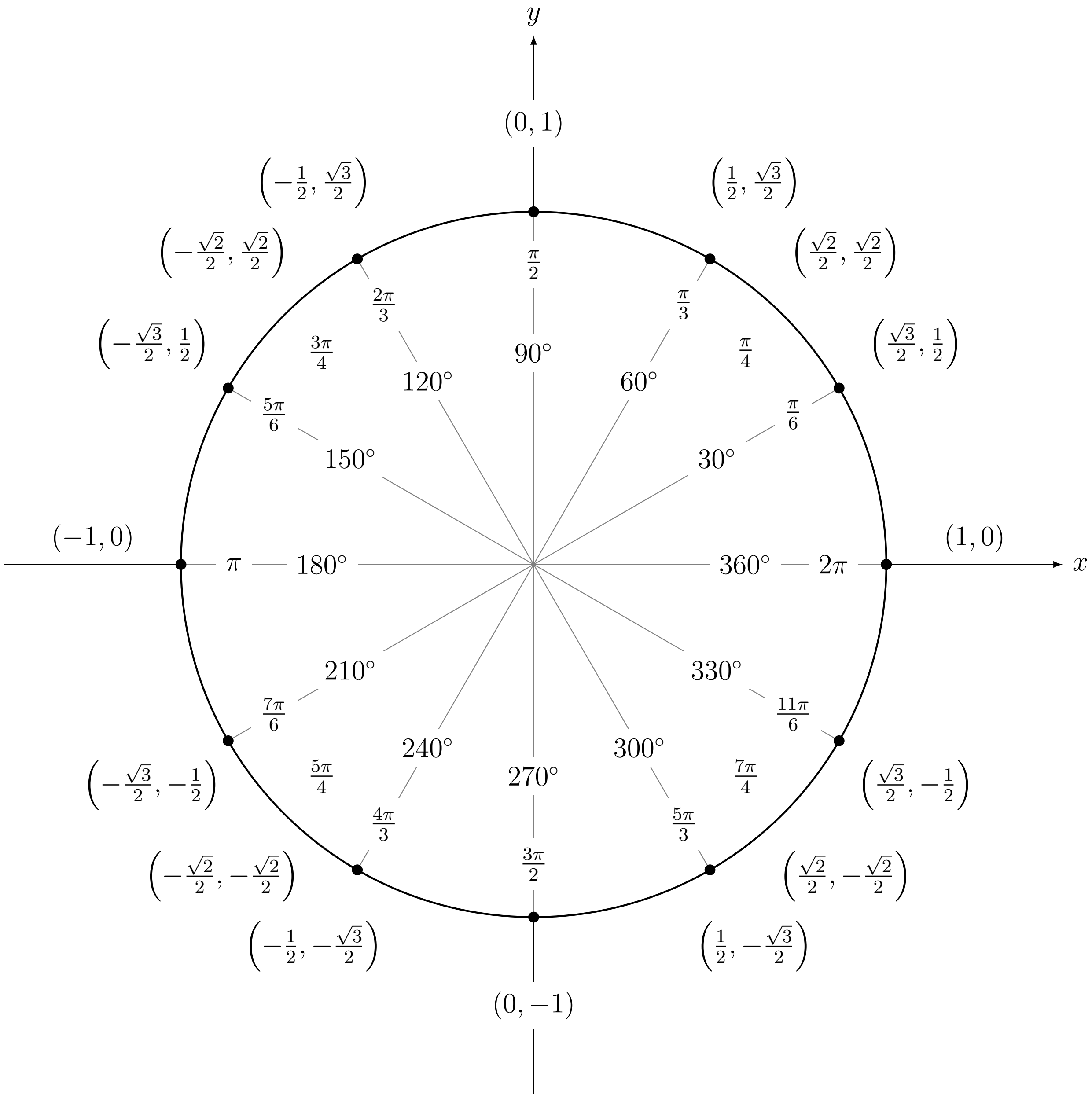
Here are the figures that summarize how to analyse some of these properties (for other relations, the method is the same):



$\vec{j}$
$\oplus$
$M(\cos(\alpha),\sin(\alpha))$
$\sin(\alpha)$
α
O
α
$\vec{i}$
$-\sin(\alpha)$
$M'(\cos(-\alpha),\sin(-\alpha))$
$\vec{j}$
$\oplus$
$M'\begin{pmatrix}\cos(\pi-\alpha)\\\sin(\pi-\alpha)\end{pmatrix}$
$\pi-\alpha$
$M(\cos(\alpha),\sin(\alpha))$
$-\alpha$
α
O
$-\cos(\alpha)$
$\cos(\alpha)$
$\vec{i}$

Figure 8.10: **Visual equivalences of some trigonometric relations**

Here are some remarkable angles associated with the values of cosine and sine functions that many students must memorize during their studies:



y
x
$(0,1)$
$(0,-1)$
$(1,0)$
$(-1,0)$
$\left(\frac{1}{2},\frac{\sqrt{3}}{2}\right)$
$\left(\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2}\right)$
$\left(\frac{\sqrt{3}}{2},\frac{1}{2}\right)$
$\left(-\frac{1}{2},\frac{\sqrt{3}}{2}\right)$
$\left(-\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2}\right)$
$\left(-\frac{\sqrt{3}}{2},\frac{1}{2}\right)$
$\left(-\frac{\sqrt{3}}{2},-\frac{1}{2}\right)$
$\left(-\frac{\sqrt{2}}{2},-\frac{\sqrt{2}}{2}\right)$
$\left(-\frac{1}{2},-\frac{\sqrt{3}}{2}\right)$
$\left(\frac{1}{2},-\frac{\sqrt{3}}{2}\right)$
$\left(\frac{\sqrt{2}}{2},-\frac{\sqrt{2}}{2}\right)$
$\left(\frac{\sqrt{3}}{2},-\frac{1}{2}\right)$
$\frac{\pi}{2}$ 90°
$\frac{\pi}{3}$ 60°
$\frac{\pi}{4}$
$\frac{\pi}{6}$ 30°
$\frac{2\pi}{3}$ 120°
$\frac{3\pi}{4}$
$\frac{5\pi}{6}$ 150°
π 180°
360° 2π
$\frac{7\pi}{6}$ 210°
$\frac{5\pi}{4}$
$\frac{4\pi}{3}$ 240°
$\frac{3\pi}{2}$ 270°
$\frac{5\pi}{3}$ 300°
$\frac{7\pi}{4}$
$\frac{11\pi}{6}$ 330°

Figure 8.11: Some identity angles and conventional trigonometric associated values for sine and cosine

Trigono-metric functions	**The angles of** $\mathrm{arcx}(x)$																
	0° 0	30° $\frac{\pi}{6}$	45° $\frac{\pi}{4}$	60° $\frac{\pi}{3}$	90° $\frac{\pi}{2}$	120° $\frac{2\pi}{3}$	135° $\frac{3\pi}{4}$	150° $\frac{5\pi}{6}$	180° π	210° $\frac{7\pi}{6}$	225° $\frac{5\pi}{4}$	240° $\frac{4\pi}{3}$	270° $\frac{3\pi}{2}$	300° $\frac{5\pi}{3}$	315° $\frac{7\pi}{4}$	330° $\frac{11\pi}{6}$	360° 2π
$\sin(x)$	0	$\frac{1}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{\sqrt{3}}{2}$	1	$\frac{\sqrt{3}}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$	$-\frac{\sqrt{2}}{2}$	$-\frac{\sqrt{3}}{2}$	-1	$-\frac{\sqrt{3}}{2}$	$-\frac{\sqrt{2}}{2}$	$-\frac{1}{2}$	0
$\cos(x)$	1	$\frac{\sqrt{3}}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$	$-\frac{\sqrt{2}}{2}$	$-\frac{\sqrt{3}}{2}$	-1	$-\frac{\sqrt{3}}{2}$	$-\frac{\sqrt{2}}{2}$	$-\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{\sqrt{3}}{2}$	1
$\tan(x)$	0	$\frac{\sqrt{3}}{3}$	1	$\sqrt{3}$	$\pm\infty$	$-\sqrt{3}$	-1	$-\frac{\sqrt{3}}{3}$	0	$\frac{\sqrt{3}}{3}$	1	$\sqrt{3}$	$\pm\infty$	$-\sqrt{3}$	-1	$-\frac{\sqrt{3}}{3}$	0
$\cot(x)$	$\mp\infty$	$\sqrt{3}$	1	$\frac{\sqrt{3}}{3}$	0	$-\frac{\sqrt{3}}{3}$	-1	$-\sqrt{3}$	$\mp\infty$	$\sqrt{3}$	1	$\frac{\sqrt{3}}{3}$	0	$-\frac{\sqrt{3}}{3}$	-1	$-\sqrt{3}$	$\mp\infty$

Table 8.2: Some identity angles and conventional trigonometric associated values table

Let us now introduce one last definition concerning circle trigonometric function relation that we will meet again in the sections of Wave Optics or during our study of Fourier transforms in the section of Sequences and Series which is the "sine cardinal":

$$\text{sinc}(x)=\frac{\sin(x)}{x} \tag{8.1.23}$$

represented by (with Maple™ 4.00b):

```
>with(plots):
>sinc:=sin(x)/x;
>plot([sinc(x)], x=-5*Pi..5*Pi, Sinc);
```

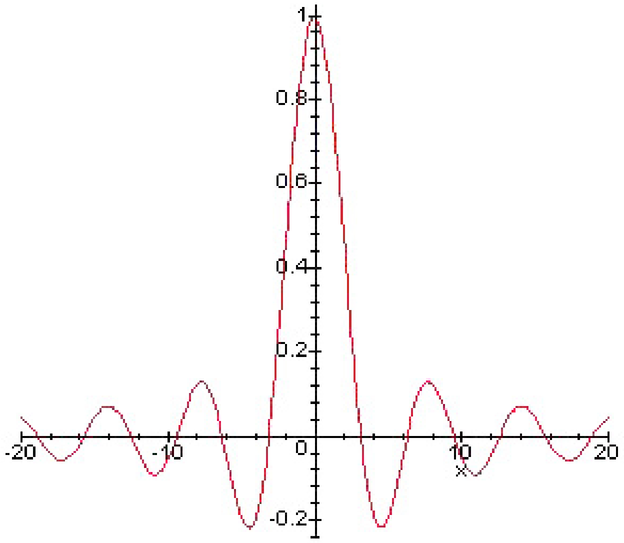


Figure 8.12: **2D plot of the sinc function in Maple™ 4.00b**

Or in a much better quality using TikZ:

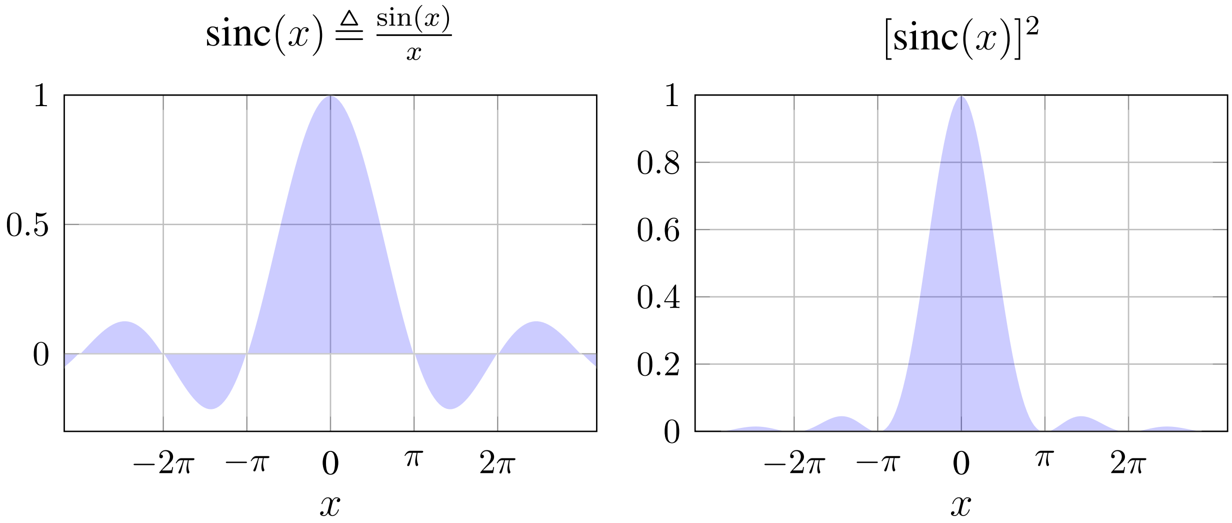


Figure 8.13: **Sine cardinal and its famous square equivalent (author: Jimmy Roussel)**

Even if engineers know very well the previous figures it is especially its pseudo-3D representation which is known by common peoples because often used for marketing reasons as reminiscent of a drop of water falling into water (figure made with Maple™ 4.00b) and it is always pretty to look at ...:

```
>plot3d(sin(sqrt(x^2+y^2))/(sqrt(x^2+y^2)),x=-20..20,y=-20..20);
```

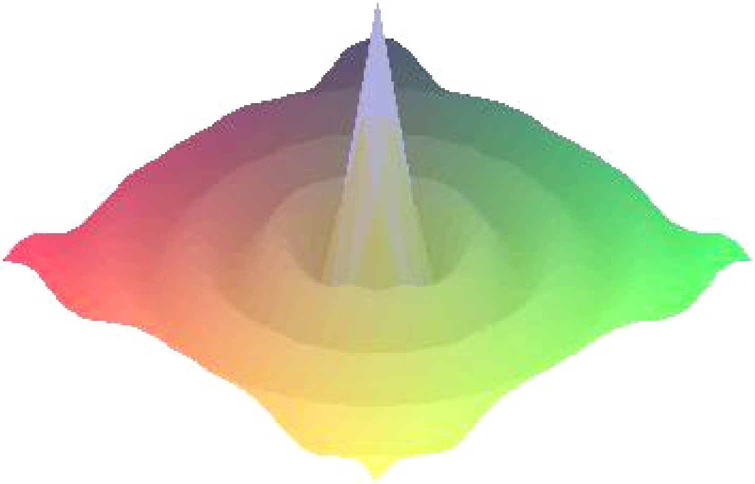


Figure 8.14: **Pseudo 3D plot of the sinc function in Maple™ 4.00b**

Or in a much better quality using TikZ:

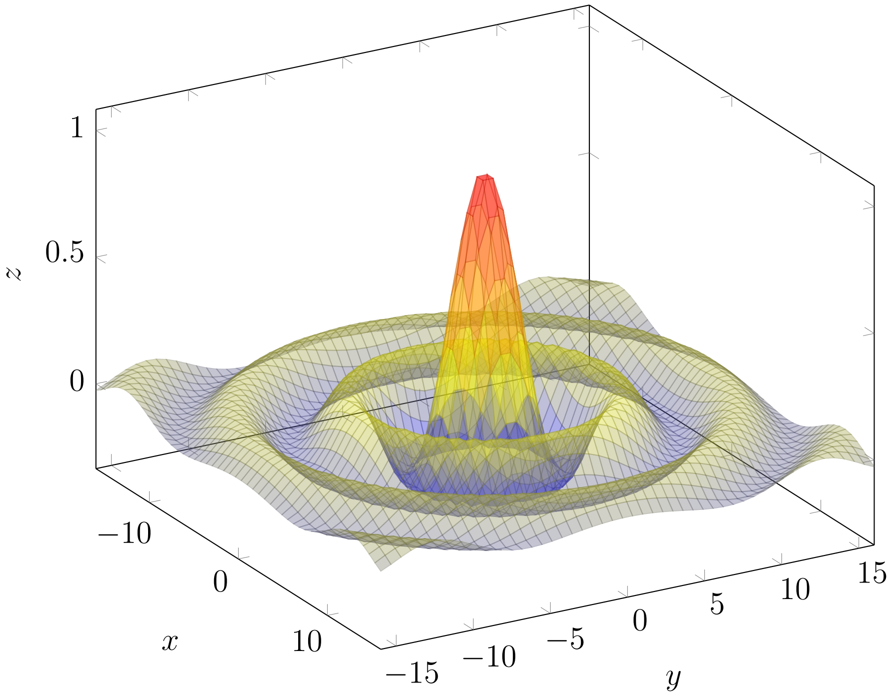


Figure 8.15: **Sine cardinal function with TikZ**

In either case, the value at $x = 0$ is defined to be the limiting value $\mathrm{sinc}(0) = 1$ that is really easy to prove using l'Hôpital's rule (see section Differential and Integral Calculus page 1461).

Keep in mind that the "sinc" function is especially important in signal processing and is the Fourier transform of a rectangular pulse and also in some Optics Phenomenon.

8.1.2.1 Remarkable trigonometric triangle identities

The drawing below will allow us to build relations that will help to solve equations involving trigonometric functions (all these relations are of primary importance in physics to simplify problem solving!).

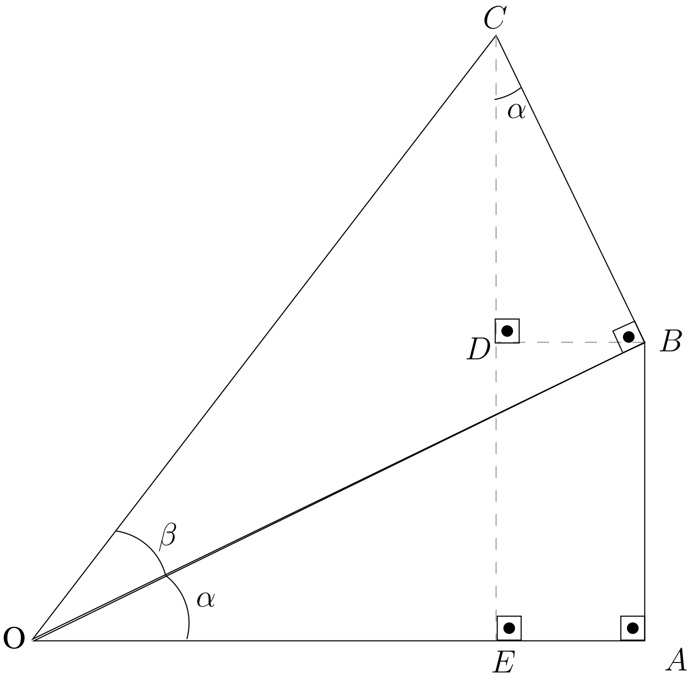


Figure 8.16: Basic construction for determining remarkable circle trigonometric identities for triangles

First we will note on the above figure the following relation:

$$\overline{DE}=\overline{AB} \tag{8.1.24}$$

Therefore:

$$\begin{aligned}\sin(\alpha+\beta)&=\frac{\overline{CD}+\overline{DE}}{\overline{\text{O}C}}=\frac{\overline{CD}}{\overline{\text{O}C}}+\frac{\overline{DE}}{\overline{\text{O}C}}=\frac{\overline{CD}}{\overline{\text{O}C}}\frac{\overline{CB}}{\overline{CB}}+\frac{\overline{DE}}{\overline{\text{O}C}}\frac{\overline{\text{O}B}}{\overline{\text{O}B}}\\&=\frac{\overline{CB}}{\overline{\text{O}C}}\frac{\overline{CD}}{\overline{CB}}+\frac{\overline{\text{O}B}}{\overline{\text{O}C}}\frac{\overline{AB}}{\overline{\text{O}B}}=\sin(\beta)\cos(\alpha)+\cos(\beta)\sin(\alpha)\end{aligned} \tag{8.1.25}$$

To resume:

$$\sin(\alpha+\beta)=\sin(\beta)\cos(\alpha)+\cos(\beta)\sin(\alpha) \tag{8.1.26}$$

Remark

We hesitated to put the main triangle identities in boxes (frames) but in fact all the results above and below are important and therefore everything will have been into boxes... So we apology to the reader if the absence of boxes (frames) disturbs.

This implies trivially if $\alpha=\beta$:

$$\sin(2\alpha)=2\sin(\alpha)\cos(\alpha) \tag{8.1.27}$$

and:

$$\sin(\alpha-\beta)=\sin(\alpha+(-\beta))=\sin(\alpha)\cos(-\beta)+\cos(\alpha)\sin(-\beta) \tag{8.1.28}$$

Therefore:

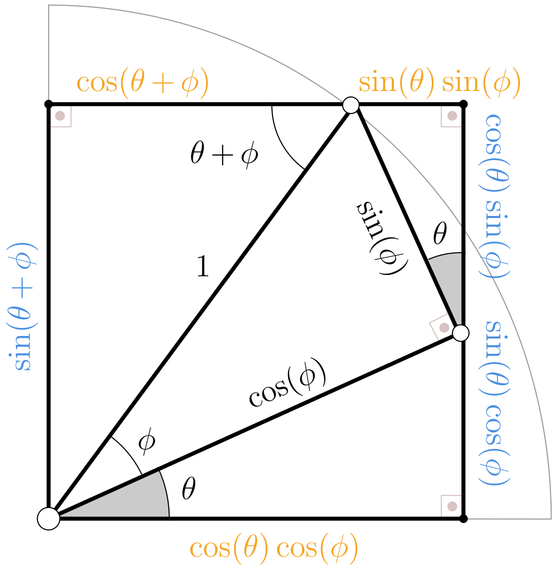
$$\sin(\alpha-\beta)=\sin(\alpha)\cos(\beta)-\cos(\alpha)\sin(\beta) \tag{8.1.29}$$

We also have:

$$\begin{aligned}\sin\left([\alpha+\beta]+\frac{\pi}{2}\right)&=\cos(\alpha+\beta)=\sin(\alpha)\cos\left(\beta+\frac{\pi}{2}\right)+\cos(\alpha)\sin\left(\beta+\frac{\pi}{2}\right)\\&=\sin(\alpha)(-\sin(\beta))+\cos(\alpha)\cos(\beta)\end{aligned} \tag{8.1.30}$$

Therefore:

$$\cos(\alpha+\beta)=\cos(\alpha)\cos(\beta)-\sin(\alpha)\sin(\beta) \tag{8.1.31}$$


$\cos(\theta+\phi)$
$\sin(\theta)\sin(\phi)$
$\theta+\phi$
$\cos(\theta)\sin(\phi)$
$\sin(\phi)$
θ
1
$\sin(\theta+\phi)$
$\sin(\theta)\cos(\phi)$
$\cos(\phi)$
ϕ
θ
$\cos(\theta)\cos(\phi)$

$$\begin{gathered}\sin(\theta+\phi)=\sin(\theta)\cos(\phi)+\cos(\theta)\sin(\phi)\\\cos(\theta+\phi)=\cos(\theta)\cos(\phi)-\sin(\theta)\sin(\phi)\end{gathered}$$

This implies trivially if $\alpha=\beta$:

$$\cos(2\alpha)=\cos^2(\alpha)-\sin^2(\alpha) \tag{8.1.32}$$

With the already proven identity $\cos^2(\alpha)+\sin^2(\alpha)=1$ we also get the very important following relations:

$$\cos(2\alpha)=\cos^2(\alpha)-\sin^2(\alpha)=1-2\sin^2(\alpha)=2\cos^2(\alpha)-1 \tag{8.1.33}$$

Relations with which we get very easily relations that seems to be named "Carnot's Formula":

$$\sin^2(\alpha)=\frac{1-\cos(2\alpha)}{2}\qquad\cos^2(\alpha)=\frac{1+\cos(2\alpha)}{2} \tag{8.1.34}$$

and:

$$\cos(\alpha-\beta)=\cos(\alpha+(-\beta))=\cos(\alpha)\cos(-\beta)-\sin(\alpha)\sin(\beta) \tag{8.1.35}$$

We also have:

$$\tan(\alpha+\beta)=\frac{\sin(\alpha+\beta)}{\cos(\alpha+\beta)}=\frac{\dfrac{\sin(\alpha)\cos(\beta)}{\cos(\alpha)\cos(\beta)}+\dfrac{\cos(\alpha)\sin(\beta)}{\cos(\alpha)\cos(\beta)}}{\dfrac{\cos(\alpha)\cos(\beta)}{\cos(\alpha)\cos(\beta)}-\dfrac{\sin(\alpha)\sin(\beta)}{\cos(\alpha)\cos(\beta)}}=\frac{\tan(\alpha)+\tan(\beta)}{1-\tan(\alpha)\tan(\beta)} \tag{8.1.36}$$

This, to get the relation:

$$\tan(\alpha+\beta)=\frac{\tan(\alpha)+\tan(\beta)}{1-\tan(\alpha)\tan(\beta)} \tag{8.1.37}$$

which implies when $\alpha=\beta$:

$$\tan(2\alpha)=\frac{2\tan(\alpha)}{1-\tan^2(\alpha)} \tag{8.1.38}$$

and obviously:

$$\tan(\alpha-\beta)=\tan(\alpha+(-\beta))=\frac{\tan(\alpha)+\tan(-\beta)}{1-\tan(\alpha)\tan(-\beta)}=\frac{\tan(\alpha)-\tan(\beta)}{1+\tan(\alpha)\tan(\beta)} \tag{8.1.39}$$

Therefore to summarize:

$$\tan(\alpha\pm\beta)=\frac{\tan(\alpha)\pm\tan(\beta)}{1\pm\tan(\alpha)\tan(\beta)} \tag{8.1.40}$$

We also get trivially (ask us if you have problems!) from previous relations (we do a little mixing and we shake...):

$$\begin{aligned}\tan^2(\alpha)&=1-\frac{2\tan(\alpha)}{\tan(2\alpha)}\\ \tan^2(\alpha)&=\frac{1-\cos(2\alpha)}{1+\cos(2\alpha)}\end{aligned} \tag{8.1.41}$$

We also have using above results:

$$\sin(2\alpha)=2\sin(\alpha)\frac{\cos(\alpha)}{\cos(\alpha)}\cos(\alpha)=2\tan(\alpha)\cos^2(\alpha) \tag{8.1.42}$$

using the relation proved far above :

$$\frac{1}{\cos^2(\alpha)}=\tan^2(\alpha)+1 \tag{8.1.43}$$

Therefore:

$$\sin(2\alpha)=2\tan(\alpha)\cos^2(\alpha)=\frac{2\tan(\alpha)}{\tan^2(\alpha)+1}\Leftrightarrow\sin(\alpha)=\frac{2\tan\left(\dfrac{\alpha}{2}\right)}{\tan^2\left(\dfrac{\alpha}{2}\right)+1} \tag{8.1.44}$$

Exactly similarly we get (ask us if you need the details):

$$\cos(2\alpha)=\cos^2(\alpha)-\sin^2(\alpha)=\cos^2(\alpha)(1-\tan^2(\alpha)) \tag{8.1.45}$$

Using always:

$$\frac{1}{\cos^2(\alpha)}=\tan^2(\alpha)+1 \tag{8.1.46}$$

Therefore:

$$\cos(2\alpha)=\cos^2(\alpha)(1-\tan^2(\alpha))=\frac{1-\tan^2(\alpha)}{\tan^2(\alpha)+1}\Leftrightarrow\cos(\alpha)=\frac{1-\tan^2\left(\dfrac{\alpha}{2}\right)}{\tan^2\left(\dfrac{\alpha}{2}\right)+1} \tag{8.1.47}$$

Now let us determine additional trigonometric formulas that seems to be name "Simpson's Formula" or "trigonometric addition formulas" which express the sum of sine and / or cosine in product of sine and / or cosine.

Given the identities already previously proved:

$$\begin{aligned}&(1)\quad \cos(\alpha+\beta)=\cos(\alpha)\cos(\beta)-\sin(\alpha)\sin(\beta)\\&(2)\quad \cos(\alpha-\beta)=\cos(\alpha)\cos(\beta)+\sin(\alpha)\sin(\beta)\end{aligned} \tag{8.1.48}$$

That can also be summarized as:

$$\cos(\alpha\pm\beta)=\cos(\alpha)\cos(\beta)\mp\sin(\alpha)\sin(\beta) \tag{8.1.49}$$

Let us put $\alpha+\beta=p$ and $\alpha-\beta=q$ and this also give us:

$$\alpha=\frac{p+q}{2}\qquad\beta=\frac{p-q}{2} \tag{8.1.50}$$

We get by summing (1) and (2):

$$\cos(p)+\cos(q)=2\cos\left(\frac{p+q}{2}\right)\cos\left(\frac{p-q}{2}\right) \tag{8.1.51}$$

and by subtracting:

$$\cos(p)-\cos(q)=2\sin\left(\frac{p+q}{2}\right)\sin\left(\frac{p-q}{2}\right) \tag{8.1.52}$$

In the same way we but starting from:

$$\begin{aligned}&(1)\quad \sin(\alpha+\beta)=\sin(\alpha)\cos(\beta)+\cos(\alpha)\sin(\beta)\\&(2)\quad \sin(\alpha-\beta)=\sin(\alpha)\cos(\beta)-\cos(\alpha)\sin(\beta)\end{aligned} \tag{8.1.53}$$

That can also be summarized as:

$$\sin(\alpha\pm\beta)=\sin(\alpha)\cos(\beta)\pm\cos(\alpha)\sin(\beta) \tag{8.1.54}$$

We get first by summing:

$$\sin(p)+\sin(q)=2\sin\left(\frac{p+q}{2}\right)\cos\left(\frac{p-q}{2}\right) \tag{8.1.55}$$

and by subtracting:

$$\sin(p)-\sin(q)=2\sin\left(\frac{p-q}{2}\right)\cos\left(\frac{p+q}{2}\right) \tag{8.1.56}$$

and re-injecting the angles we easily (normally but if you have difficulties tell us!) fall back on the following "Simpson's inverse formulas":

$$\begin{aligned}\cos(\alpha)\cos(\beta)&=\frac{1}{2}\left(\cos(\alpha+\beta)+\cos(\alpha-\beta)\right)\\ \sin(\alpha)\sin(\beta)&=\frac{1}{2}\left(\cos(\alpha-\beta)-\cos(\alpha+\beta)\right)\\ \sin(\alpha)\cos(\beta)&=\frac{1}{2}\left(\sin(\alpha+\beta)+\sin(\alpha-\beta)\right)\end{aligned} \tag{8.1.57}$$

We also have starting from:

$$\sin(\alpha+\beta)=\sin(\alpha)\cos(\beta)+\cos(\alpha)\sin(\beta) \tag{8.1.58}$$

and we divide by $\cos(\alpha)\cos(\beta)$, to get:

$$\frac{\sin(\alpha+\beta)}{\cos(\alpha)\cos(\beta)}=\frac{\sin(\alpha)\cos(\beta)+\cos(\alpha)\sin(\beta)}{\cos(\alpha)\cos(\beta)}=\frac{\sin(\alpha)}{\cos(\alpha)}+\frac{\sin(\beta)}{\cos(\beta)}=\tan(\alpha)+\tan(\beta) \tag{8.1.59}$$

And exactly in the same way we get:

$$\frac{\sin(\alpha-\beta)}{\cos(\alpha)\cos(\beta)}=\frac{\sin(\alpha)\cos(\beta)-\cos(\alpha)\sin(\beta)}{\cos(\alpha)\cos(\beta)}=\frac{\sin(\alpha)}{\cos(\alpha)}-\frac{\sin(\beta)}{\cos(\beta)}=\tan(\alpha)-\tan(\beta) \tag{8.1.60}$$

All these relations will help us in our study of general physics (especially in the section of Wave Mechanics and therefore in the sections of Electrodynamics and Acoustics) and particularly in the case of integral calculations or superposition of waves.

Remarque

To summarize (this may help...), the following equalities demonstrated previously:

$$\begin{aligned}\cos(p)+\cos(q)&=2\cos\left(\frac{p+q}{2}\right)\cos\left(\frac{p-q}{2}\right)\\ \cos(p)-\cos(q)&=2\sin\left(\frac{p+q}{2}\right)\sin\left(\frac{p-q}{2}\right)\\ \sin(p)+\sin(q)&=2\sin\left(\frac{p+q}{2}\right)\cos\left(\frac{p-q}{2}\right)\\ \sin(p)-\sin(q)&=2\sin\left(\frac{p-q}{2}\right)\cos\left(\frac{p+q}{2}\right)\\ \tan(\alpha)+\tan(\beta)&=\frac{\sin(\alpha+\beta)}{\cos(\alpha)\cos(\beta)}\\ \tan(\alpha)-\tan(\beta)&=\frac{\sin(\alpha-\beta)}{\cos(\alpha)\cos(\beta)}\end{aligned} \tag{8.1.61}$$

are therefore named "Simpson's formulas".

Let us also derive the following identity that will be useful to us in the section of Differential and Integral Calculus:

$$\begin{aligned}\tan\left(x+\frac{\pi}{2}\right)&=\frac{\sin\left(x+\frac{\pi}{2}\right)}{\cos\left(x+\frac{\pi}{2}\right)}=\frac{\sin x\cos\left(\frac{\pi}{2}\right)+\cos x\sin\left(\frac{\pi}{2}\right)}{\cos x\cos\left(\frac{\pi}{2}\right)-\sin x\sin\left(\frac{\pi}{2}\right)}\\&=\frac{0+\cos(x)}{0-\sin(x)}=-\frac{\cos(x)}{\sin(x)}=-\cot(x)\end{aligned}\tag{8.1.62}$$

and respectively:

$$\cot(x)=-\tan\left(x-\frac{\pi}{2}\right)\tag{8.1.63}$$

We also have this identity that will be useful to us in the section of Astronomy:

$$\begin{aligned}\tan^2(x)-\cot^2(x)&=\frac{\sin^2(x)}{\cos^2(x)}-\frac{\cos^2(x)}{\sin^2(x)}\\&=\frac{\sin^4(x)-\cos^4(x)}{\cos^2(x)\sin^2(x)}\\&=\frac{\left(\sin^2(x)+\cos^2(x)\right)\left(\sin^2(x)-\cos^2(x)\right)}{(\cos(x)\sin(x))^2}\\&=\frac{(1)(-\cos(2x))}{\frac{1}{4}\sin^2(2x)}\\&=-4\frac{\cos(2x)}{\sin^2(2x)}\end{aligned}\tag{8.1.64}$$

8.1.2.1.1 Laws of Cosines (cosine rule)

Let us prove now the cosine theorem that will be very useful to us in the sections of Geometry, Cosmology, Electrodynamics and Mechanics.

Theorem 8.1. In any triangle, the square of one of the three sides is equal to the sum of the squares of the two others reduced by the product of these two sides by the cosine of the angle between them:

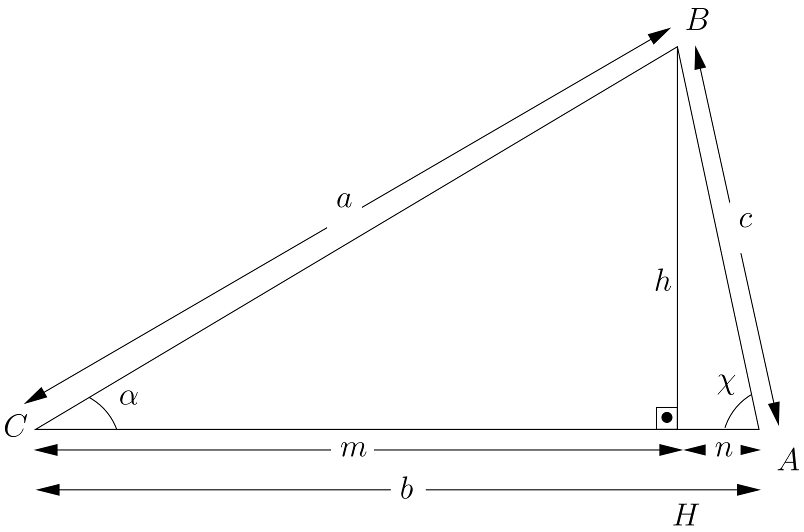


Figure 8.17: **Construction for the proof of the law of cosines**

Proof 8.1.1. One possible proof has the following path by using the Pythagorean theorem:

$$\begin{array}{r}a^2=h^2+m^2\\ -c^2=h^2+n^2\\ \hline a^2-c^2=m^2-n^2\\ a^2-c^2=(m+n)(m-n)\\ a^2-c^2=b(b-2n)=b^2-2nb\end{array} \tag{8.1.65}$$

But in the ABH triangle, rectangle in H, we have the relation $n/c=\cos(\chi)$ seen above therefore:

$$a^2-c^2=b(b-2n)=b^2-2nb=b^2-2bc\cos(\chi) \tag{8.1.66}$$

So we get one of the relations of the "cosine theorem":

$$\boxed{a^2=b^2+c^2-2bc\cos(\chi)} \tag{8.1.67}$$

By circular permutation, we get the other two known relations of the cosine theorem.

■ Q.E.D.

Remark

The cosine theorem is sometimes named "Al-Kashi's formula". And you can note that if a is the hypotenuse and its opposite angle a right angle such that $\cos(\chi)$ is zero, so we fall back on the Pythagorean theorem. Here's why we sometimes name the Al-Kashi's formula: "generalized Pythagorean formula".

An interesting case for applying the cosine theorem is to determine the distance l between a point b on the circumference of the circle and a point a situated inside the circle depending on the direction (angle) in which the eye is directed:

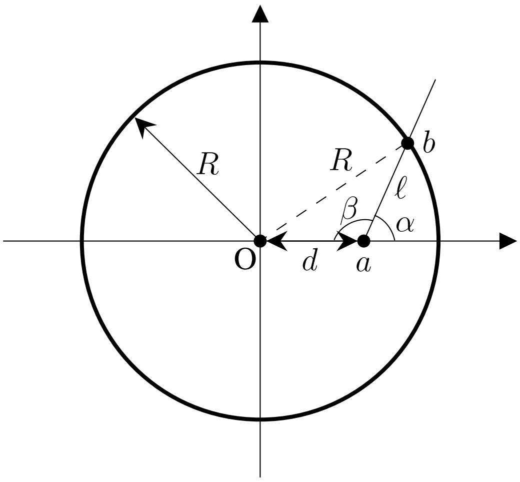


Figure 8.18: **Distance from a point inside a circle at its circumference**

We can therefore apply the cosine theorem that gives us (knowing d, R and β):

$$R^2=d^2+l^2-2dl\cos(\beta) \tag{8.1.68}$$

Therefore:

$$l^2-(2d\cos(\beta))l-(R^2-d^2)=0 \tag{8.1.69}$$

It is therefore an equation of the second degree whose discriminant is (see section Calculus page 1383):

$$\Delta^2=(2d\cos(\beta))^2+4(R^2-d^2) \tag{8.1.70}$$

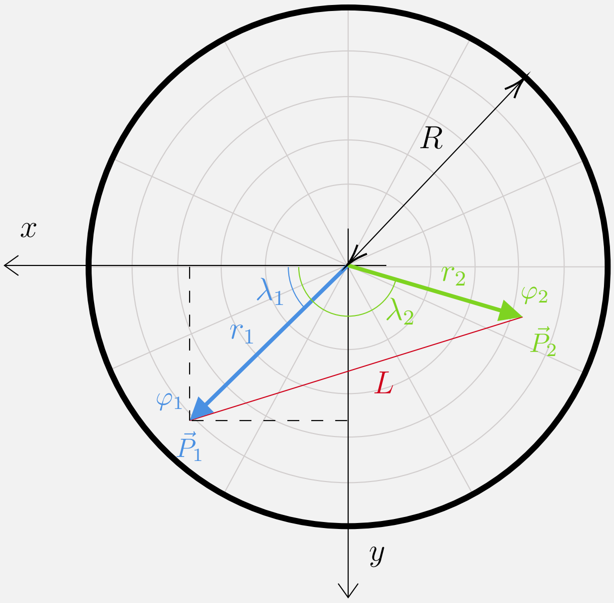
and whose two solutions are (see section Calculus page 1383):

$$l_{1,2}=\frac{2d\cos(\beta)\pm\sqrt{(2d\cos(\beta))^2+4(R^2-d^2)}}{2}=d\cos(\beta)\pm\sqrt{(d\cos(\beta))^2+(R^2-d^2)} \tag{8.1.71}$$

We will also see an important application in the section of Geometric Shapes to calculate the area of any triangle (Heron's formula) page 2898.

Remark

A funny application of the law of cosines is also the flat Earth joke. Indeed, in this joke the Earth is represented by a disc of radius R as seen from the top as:


R
x
y
λ_1
r_1
φ_1
$\vec{P}_1$
r_2
λ_2
φ_2
$\vec{P}_2$
L

where $\vec{P}_1$ and $\vec{P}_2$ are two localization of this flat Earth. Then from the law of cosines, we get:

$$L = R \cdot \sqrt{r_1^2 + {r_2}^2 - 2 \cdot r_1 \cdot r_2 \cdot \cos(\Delta\lambda)} \tag{8.1.72}$$

with:

$$r_1 = \left(1 - \frac{\varphi_1}{90^\circ}\right) \quad r_2 = \left(1 - \frac{\varphi_2}{90^\circ}\right) \quad \Delta\lambda = \lambda_2 - \lambda_1 \tag{8.1.73}$$

8.1.2.1.2 Laws of Sines (sine rule)

Consider the below triangle for whom we represent the two heights:

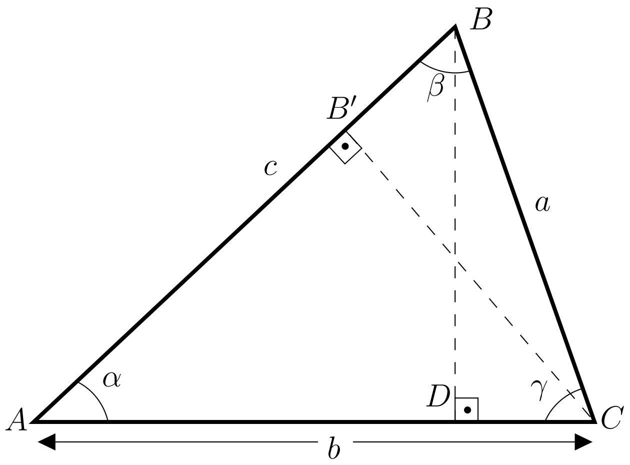


Figure 8.19: **Construction for the proof of the law of sines**

In the triangle above we have the relations:

$$\sin(\alpha)=\frac{\overline{DB}}{c}\qquad \sin(\gamma)=\frac{\overline{DB}}{a}\tag{8.1.74}$$

which leads us to the expression:

$$\overline{DB}=c\sin(\alpha)=a\sin(\gamma)\tag{8.1.75}$$

Therefore:

$$\frac{a}{\sin(\alpha)}=\frac{c}{\sin(\gamma)}\tag{8.1.76}$$

By similar reasoning we have:

$$\sin(\alpha)=\frac{\overline{CB'}}{b}\qquad \sin(\beta)=\frac{\overline{c}}{a}\tag{8.1.77}$$

That gives in the same way:

$$\frac{a}{\sin(\alpha)}=\frac{b}{\sin(\beta)}\tag{8.1.78}$$

All combined provides us the "sine theorem", the most beautiful example of application in this textbook is certainly the determination of the Lagrange points L4 and L5 in the section Astronomy:

$$\boxed{\frac{a}{\sin(\alpha)}=\frac{b}{\sin(\beta)}=\frac{c}{\sin(\gamma)}}\tag{8.1.79}$$

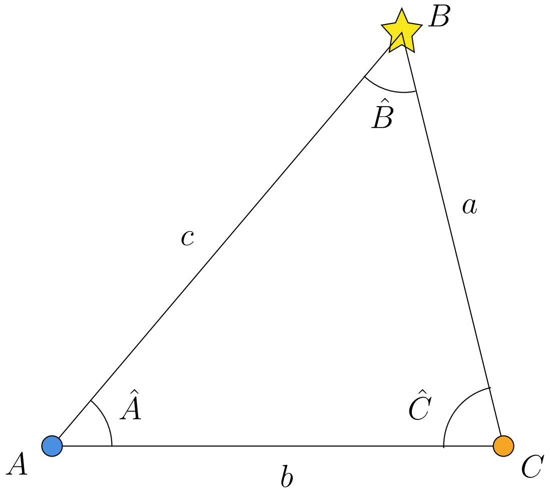
Obviously, there is not here all the existing trigonometric circle relations as we have already said, but at least the most important you need to know to find we you study physical systems.

The law of sines in famously know for its application to "triangulation"! In trigonometry and geometry, triangulation is the process of determining the location of a point by forming triangles to the point from known points. The properties used for triangulation are:

- The sum of the angles of a triangle, equal to π radians (180°)
- The law of sines

If we denote a the side opposite the angle $\hat{A}$ and c the side opposite the angle $\hat{C}$, we express the law of sines as follows:

$$\frac{a}{\sin(\hat{A})}=\frac{b}{\sin(\hat{B})}=\frac{c}{\sin(\hat{C})} \tag{8.1.80}$$



Hence:

$$\begin{aligned}AB=c=\frac{b\sin(\hat{C})}{\sin(\hat{B})}\\ CB=a=\frac{b\sin(\hat{A})}{\sin(\hat{B})}\end{aligned} \tag{8.1.81}$$

We obviously have:

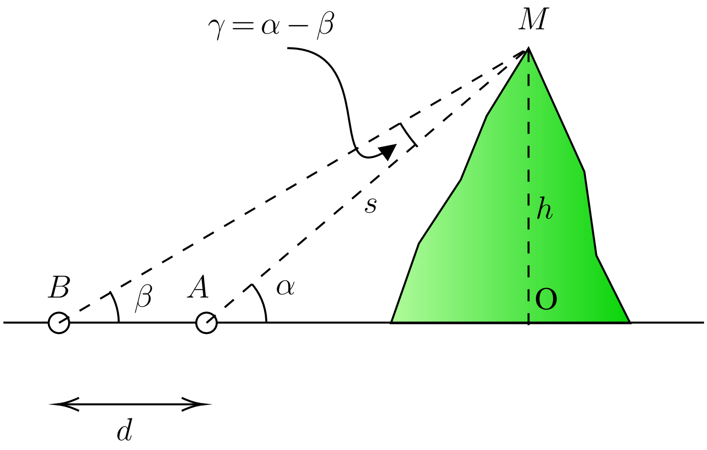
$$\hat{C}=180^\circ-\hat{A}-\hat{B} \tag{8.1.82}$$

✍**Example:**

Abū Rayḥān Al-Bīrūnī, a pioneering Muslim scientist, figured out a truly remarkable and ingenious method to calculate the radius of the Earth (and subsequently its circumference) in the 111th century (holocene calendar) different from that of Eratosthenes (that used another method 13 centuries before). This method was very simple yet accurate, requiring just four measurements in all to be taken and then applying a trigonometric relation to arrive at the solution.

To carry out his method Biruni only needed an astrolabe, a suitable mountain with a flat horizon in front of it so that the angle of depression of the horizon could be accurately measured and basic knowledge of trigonometry.

Let us consider the figure below depicting the scenario:



Considering the right-angled triangle OMA, the mountain height h and the (unknown) slope height s are related obviously by:

$$h=s\cdot\sin(\alpha) \tag{8.1.83}$$

Considering angles in triangle AMB, we easily see that $\gamma=\alpha-\beta$. Then, applying the sine rule to AMB, we get:

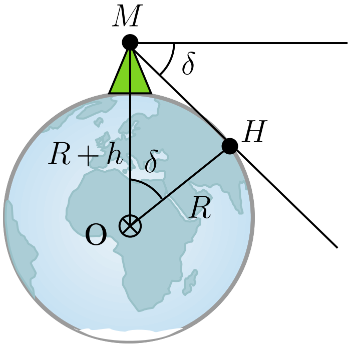
$$\frac{s}{\sin(\beta)}=\frac{d}{\sin(\alpha-\beta)} \tag{8.1.84}$$

Combining the two relations above and using the trigonometric identities proved earlier we easily get using the Simpson's relation also proved earlier above:

$$h=\left(\frac{\sin(\alpha)\sin(\beta)}{\sin(\alpha-\beta)}\right)d=\frac{\tan(\alpha)\cos(\alpha)\tan(\beta)\cos(\beta)}{\sin(\alpha-\beta)}d=\left(\frac{\tan(\alpha)\tan(\beta)}{\tan(\alpha)-\tan(\beta)}\right)d \tag{8.1.85}$$

This is the relationship that Al Biruni used to calculate the height of the mountain. The input values $\{\alpha,\beta,d\}$ can all be obtained with good precision so that, provided the tangents of angles can be evaluated accurately, an excellent estimate of the height h can be made. Indeed, the relation has been used in surveying up to the present day!

We now describe how al-Biruni used the mountain height to estimate the size of the Earth. The single additional measurement he required was the dip angle δ of the horizon as viewed from the mountain top (M in the figure below):



It is clear that the angle MOH is equal to δ. It is also clear from the figure that $\cos(\delta) = R/(R+h)$. This implies:

$$R = \frac{\cos(\delta)}{1-\cos(\delta)}h \tag{8.1.86}$$

which gives us a formula for Earth radius R. Provided that he could evaluate the cosine of a small angle δ, al-Biruni could use this formula to obtain R from h and δ.

8.1.3 Hyperbolic Trigonometry

We have shown in the section of Functional Analysis that any function $f(x)$ can be divided into an odd and even function as:

$$f(x)=\frac{f(x)-f(-x)}{2}+\frac{f(x)+f(-x)}{2} \tag{8.1.87}$$

Thus, for the function $f(x)=e^x$, we get:

$$e^x=\frac{e^x-e^{-x}}{2}+\frac{e^x+e^{-x}}{2} \tag{8.1.88}$$

Remember that in our study of complex numbers (see section Numbers page 252) we proved the following Euler formulas:

$$\cos(\phi)=\frac{e^{\mathrm{i}\phi}+e^{-\mathrm{i}\phi}}{2} \qquad \sin(x)=\frac{e^{\mathrm{i}\phi}-e^{-\mathrm{i}\phi}}{2\mathrm{i}} \tag{8.1.89}$$

and we talked about the fact that even if ϕ should be most of time be in $\mathbb{R}$, it can also be in $\mathbb{C}$ and therefore we generalize the definition of the sine and cosine by analogy with the hyperbolic sine and cosine (we will demonstrate the origin of this term later below) by:

$$\boxed{\cosh(z)=\frac{e^z+e^{-z}}{2} \qquad \sinh(z)=\frac{e^z-e^{-z}}{2}} \tag{8.1.90}$$

where $z\in\mathbb{C}$. Therefore we can also obviously define:

$$\tanh(z)=\frac{\sinh(z)}{\cosh(z)} \tag{8.1.91}$$

And so on... like for circle trigonometry.

By the way, the reader can notice that:

$$\cos(z)=\cosh(\mathrm{i}z) \quad \cosh(z)=\cos(\mathrm{i}z) \tag{8.1.92}$$

We can therefore write:

$$e^z=\cosh(z)+\sinh(z) \tag{8.1.93}$$

relation that we can again put in analogy with:

$$e^{\mathrm{i}x}=\cos(x)+\mathrm{i}\sin(x) \tag{8.1.94}$$

Interestingly, we can now work with complex angles in trigonometry!!! Indeed, if we put $a+ib$ then we have:

$$e^{a+ib}=\cosh(a+ib)+\sinh(a+ib) \tag{8.1.95}$$

But as we know (see section Numbers page 1353):

$$e^{a+\mathrm{i}b}=e^a e^{\mathrm{i}b}=re^{\mathrm{i}b}=r(\cos(b)+\mathrm{i}\sin(b)) \tag{8.1.96}$$

Therefore we can write the following equality:

$$e^a\cos(b)+e^a\mathrm{i}\sin(b)=\cosh(a+\mathrm{i}b)+\sinh(a+\mathrm{i}b) \tag{8.1.97}$$

Therefore the hyperbolic trigonometric function of a complex angle exists and the image is also a complex number. We can see abusively hyperbolic trigonometry as a kind of generalization of trigonometry circle with real and complex angles.

As opposed to the circle trigonometry where we have proved that:

$$\sin^2(\alpha)+\cos^2(\alpha)=1 \tag{8.1.98}$$

In hyperbolic trigonometry we have:

$$\sinh^2(z)-\cosh^2(z)=1 \tag{8.1.99}$$

The proof is easy:

$$\begin{aligned}\cosh^2(z)&=\left(\frac{e^z-e^{-z}}{2}\right)^2=\frac{e^{2z}-2e^ze^{-z}+e^{-2z}}{4}=\frac{1}{2}+\frac{e^{2z}+e^{-2z}}{4}\\&=1-\frac{1}{2}+\frac{e^{2z}+e^{-2z}}{4}=1+\frac{e^{2z}-2+e^{-2z}}{4}=1+\frac{e^{2z}-2e^ze^{-z}+e^{-2z}}{4}\\&=1+\left(\frac{e^z+e^{-z}}{2}\right)^2=1+\sinh^2(z)\end{aligned} \tag{8.1.100}$$

By rearranging we get:

$$\sinh^2(z)-\cosh^2(z)=1 \tag{8.1.101}$$

Before we prove why this is named "hyperbolic trigonometric" let us see some important identities that we will use in various sections of this textbook.

First we will look for a closed form of the reciprocal hyperbolic sine and cosine functions. For this purpose remember that:

$$y:=\cosh(z)=\frac{e^z+e^{-z}}{2} \tag{8.1.102}$$

and that the search for the inverse function is always to isolate the variable (in this case: z).

Therefore:

$$2y-e^{-z}-e^z=2\cosh(z)-e^{-z}-e^z=0 \tag{8.1.103}$$

That is to say (multiplying by $-e^z$):

$$(e^z)^2-2ye^z+1=0 \tag{8.1.104}$$

by solving the polynomial this second degree polynomial on e^z we get:

$$e^z_{1,2}=\frac{2y\pm\sqrt{4y^2-4}}{2}=y\pm\sqrt{y^2-1} \tag{8.1.105}$$

As $e^z > 0$ we must reject the solution with the sign "$-$" (yes! otherwise with a $z \in \mathbb{R}$ you could have a negative e^z). It comes then:

$$e^z = y + \sqrt{y^2 - 1} \tag{8.1.106}$$

Therefore:

$$z = \ln\left(y + \sqrt{y^2 - 1}\right) \tag{8.1.107}$$

Using the same for:

$$y := \sinh(z) = \frac{e^z + e^{-z}}{2} \tag{8.1.108}$$

Therefore:

$$2y + e^{-z} - e^z = 2\sinh(z) + e^{-z} - e^z = 0 \tag{8.1.109}$$

That is to say (multiplying by $-e^z$):

$$(e^z)^2 - 2ye^z - 1 = 0 \tag{8.1.110}$$

by solving the polynomial this second degree polynomial on e^z we get:

$$e^z_{1,2} = \frac{2y \pm \sqrt{4y^2 + 4}}{2} = y \pm \sqrt{y^2 + 1} \tag{8.1.111}$$

As $e^z > 0$ we must reject the solution with the sign "-" (yes! otherwise with a $z \in \mathbb{R}$ you could have a negative e^z). It comes then:

$$e^z = y + \sqrt{y^2 + 1} \tag{8.1.112}$$

Therefore:

$$z = \ln\left(y + \sqrt{y^2 + 1}\right) \tag{8.1.113}$$

Finally :

$$\begin{aligned} z = \operatorname{arccosh}(y) \quad &= \ln\left(y + \sqrt{y^2 - 1}\right) \\ z = \operatorname{arcsinh}(y) \quad &= \ln\left(y + \sqrt{y^2 + 1}\right) \end{aligned} \tag{8.1.114}$$

Caution! Following the authors arccosh and arcsinh are denoted by argcosh, argsinh or $\cosh^{-1}, \sinh^{-1}$ (the two last notations are confusing with the corresponding inverse hyperbolic function!).

To study now a simple important geometric representation of that stuff let us now write:

$$\begin{aligned} \cos(\alpha) = X \qquad & \sin(\alpha) = Y \\ \cosh(z) = X' \qquad & \sinh(z) = Y' \end{aligned} \tag{8.1.115}$$

With a restriction $\alpha, z \in \mathbb{R}$ we have:

$$\begin{aligned} X &= [-1,1] \qquad Y = [-1,1] \\ X' &= [1,+\infty] \qquad Y' =]-\infty,+\infty[\end{aligned} \tag{8.1.116}$$

This can be represented as follow:

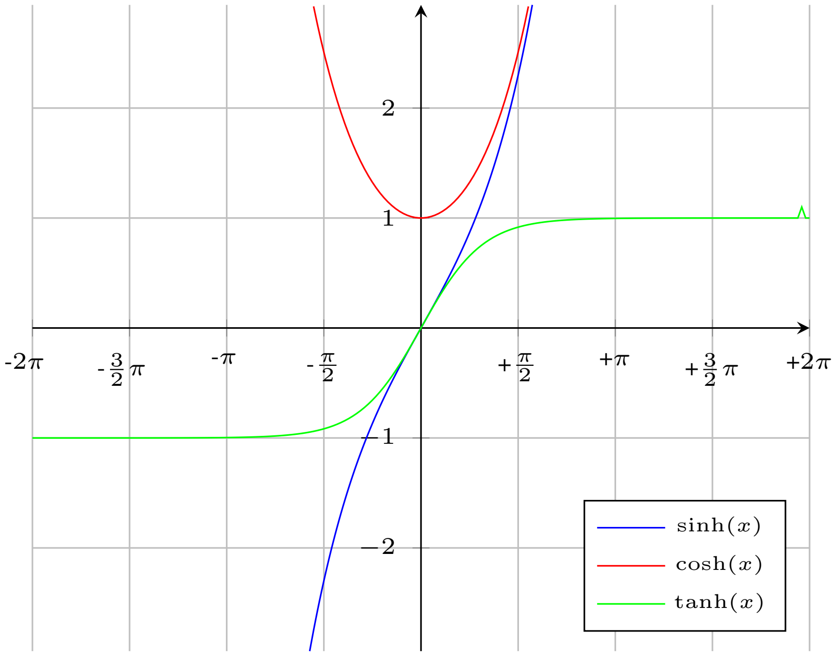


Figure 8.20: **Hyperbolic functions domain**

With this new notation, we have:

$$\begin{aligned} \cos^2(\alpha)+\sin^2(\alpha) &= X^2+Y^2 = 1 \\ \cosh^2(z)+\sinh^2(z) &= X'^2-Y'^2 = 1 \end{aligned} \tag{8.1.117}$$

But as we shall see in our study on conical in the section of Analytical Geometry:

1. The first of these two relations is for the entire given domain of definition, a circle of unit radius centered on the origin. The reader will note that it is curious enough for circle trigonometry to get a circle ☺
2. The second of these two relations is for the entire domain of a definition, an equilateral hyperbola oriented along the X axis whose apex is $S(1,0)$ and of focus $F(\sqrt{2},0)$. The reader will note again that he is curious enough for hyperbolic trigonometry to get a hyperbola ☺

For the second case this takes us to have the possibility to represent the two main hyperbolic trigonometric functions as:

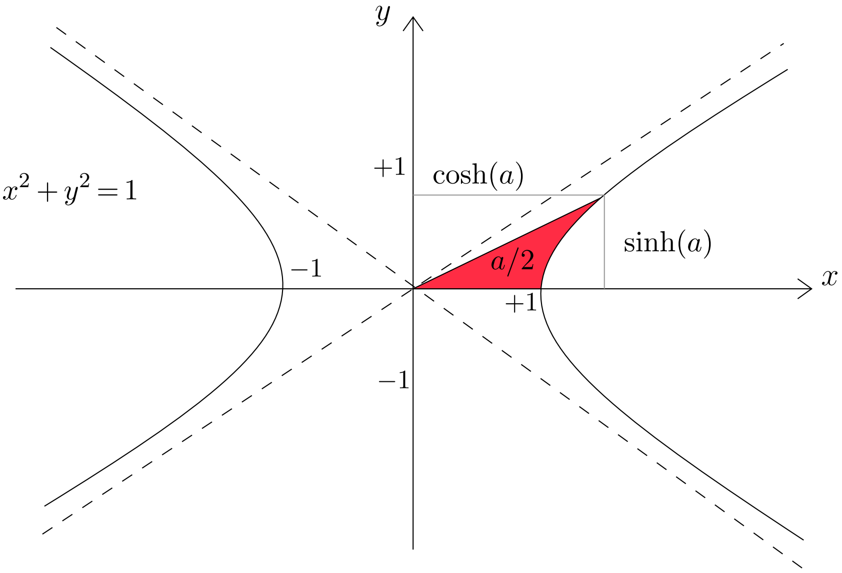


Figure 8.21: Equivalence of hyperbolic functions and hyperbola (source: Wikipedia)

This figure is very interesting because as we will prove it now, from the calculation of the red area $a/2$, where a is assimilated to a "hyperbolic angle", we can also define the hyperbolic trigonometric functions.

Let us denote by x and y the abscissa and ordinate of the red triangle (by imagining this is full a triangle). We therefore have (see section Geometric Shapes page 2896) therefore for its area (base multiplied by high divided by 2):

$$A_{\text{tr}}=\frac{xy}{2} \tag{8.1.118}$$

But we see that we have to subtract a part of an area under the hyperbola of equation $X^2-Y^2=1$. That is to say of equation:

$$Y=\sqrt{X^2-1}:=\sqrt{t^2-1} \tag{8.1.119}$$

And the area to subtract is therefore given by:

$$A_{\text{hyp}}=\int\limits_1^x\sqrt{t^2-1}\mathrm{d}t \tag{8.1.120}$$

Therefore the red area is given by:

$$A=A_{\text{tr}}-A_{\text{hyp}}=\frac{xy}{2}-\int\limits_1^x\sqrt{t^2-1}\mathrm{d}t \tag{8.1.121}$$

Using the primitive proved in the section of Differential and Integral Calculus (see page 1534), we get:

$$A=\frac{xy}{2}-\left(\frac{t}{2}\sqrt{t^2-1}-\frac{1}{2}\ln\left(t+\sqrt{t^2-1}\right)\right)\Bigg|_1^x \tag{8.1.122}$$

Therefore:

$$A=\frac{xy}{2}-\frac{x}{2}\sqrt{x^2-1}+\frac{1}{2}\ln\left(x+\sqrt{x^2-1}\right) \tag{8.1.123}$$

But we can rewrite this:

$$A=\frac{x\sqrt{x^2-1}}{2}-\frac{x}{2}\sqrt{x^2-1}-\frac{1}{2}\ln\left(x+\sqrt{x^2-1}\right):=\frac{a}{2} \tag{8.1.124}$$

Therefore:

$$a=x\sqrt{x^2-1}-x\sqrt{x^2-1}+\ln\left(x+\sqrt{x^2-1}\right) \tag{8.1.125}$$

Finally:

$$a=\ln\left(x+\sqrt{x^2-1}\right)=2A \tag{8.1.126}$$

We recognize here the first relation of the both that we proved a little sooner above if we put $z=a,y=x$:

$$\begin{aligned} z=\text{arccosh}(y) &=\ln\left(y+\sqrt{y^2-1}\right)\\ z=\text{arcsinh}(y) &=\ln\left(y+\sqrt{y^2+1}\right)\end{aligned} \tag{8.1.127}$$

So the interpretation is very interesting: the hyperbolic angle can be seen as the double of the red triangle surface (or if you prefer: the half of the hyperbolic angle is equal to the red surface)!

These two relations:

$$\begin{aligned}&\cos^2(\alpha)+\sin^2(\alpha)=X^2+Y^2=1\\ &\cosh^2(z)+\sinh^2(z)=X'^2-Y'^2=1\end{aligned} \tag{8.1.128}$$

should help, we hope to understand the origin of the name of the hyperbolic trigonometry and note that the study of hyperbolic trigonometry, is analogous to the study of circle trigonometry on the circle.

If we represent the trigonometric circle and the trigonometric hyperbole and we add some additional information, here is what we get:

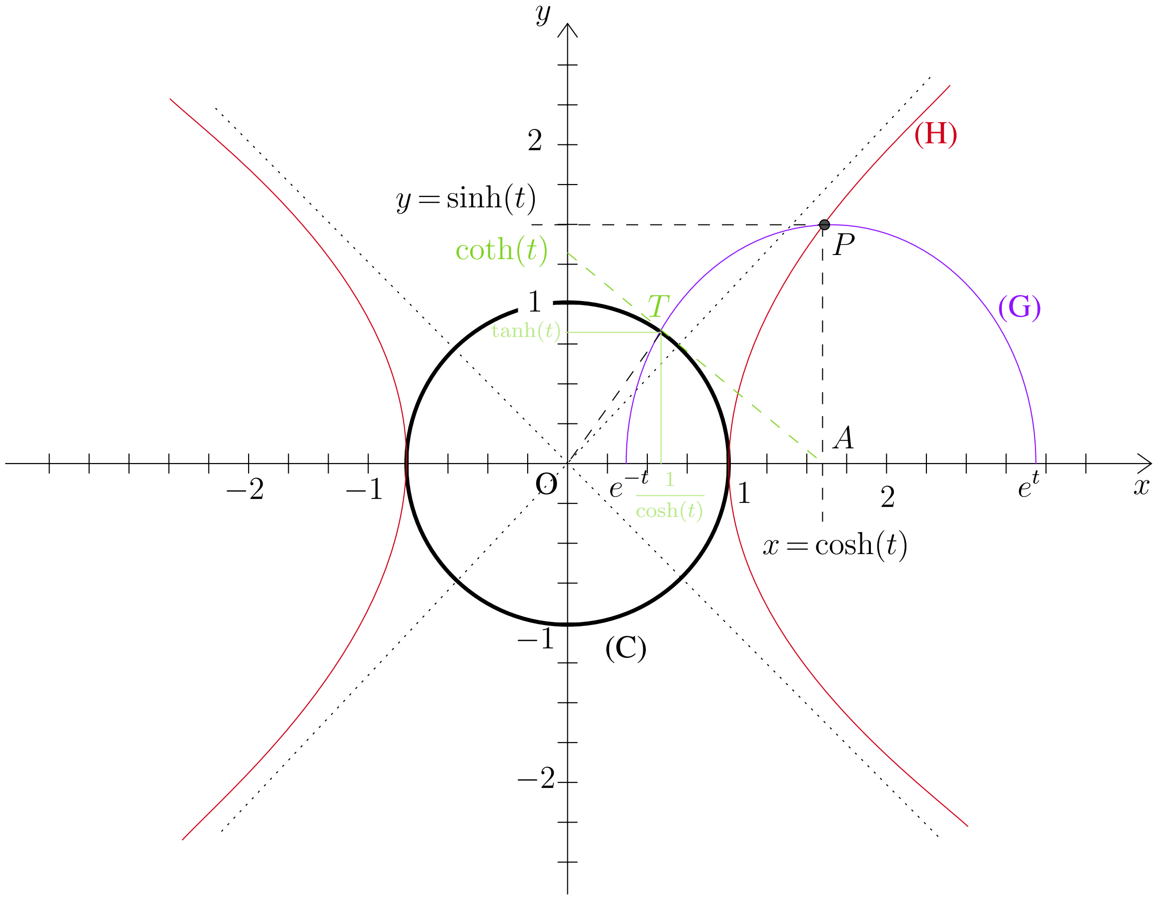


Figure 8.22: Visual definition of circle and hyperbolic trigonometric functions

Explanations:

To trace with the ruler and compass the point $P(x,y)$ of the hyperbole, we give ourself a x, that is to say the point $A(x,0)$. We draw the tangent to the circle (C) through $A(x,0)$ which gives us the tangent point T. We trace the circle (G) of center $A(x,0)$ and passing through T. This circle intersects the hyperbola at the point $P(x,y)$ perpendicular to the $A(x,0)$ to Ox.

We see appear on the figure several values of the hyperbolic functions corresponding to $x = \cosh(t), y = \sinh(t)$ but also $\tanh(t), \coth(t)$, etc. Among others, the circle (G) intersects the axis Ox in two points whose abscissas are e^t and e^{-t}.

If the reader wants to check thanks to the figure, it can control that in all points of the hyperbole, the relations (among others):

$$\cosh(t) = \frac{e^t + e^{-t}}{t} \qquad \sinh(t) = \frac{e^t - e^{-t}}{t} \qquad \cosh^2(t) - \sinh^2(t) = 1 \tag{8.1.129}$$

are always verified.

We get for the readers who wants to play with software using Maple™ 4.00b:

```
>plot([sinh(x),cosh(x),tanh(x)],x=-2*Pi..2*Pi,color=[red,black,blue]);
```

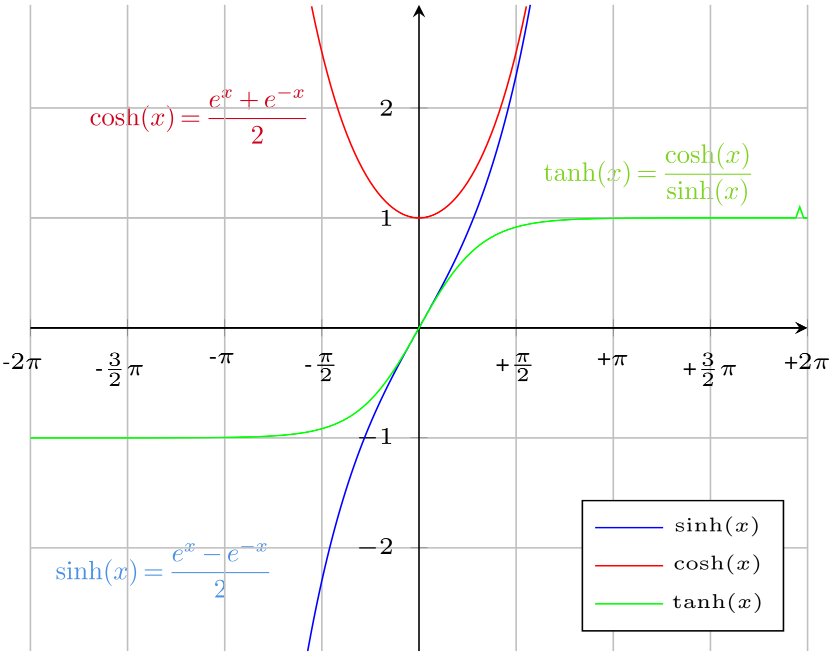


Figure 8.23: Visual definition of hyperbolic trigonometric

We will see again the $\cosh(x)$ in the section of Civil Engineering for example in the context of suspended cables. We also see again $\sinh(x)$ and $\tanh(x)$ as part of the study of gravity waves in the section of Marine and Weather Engineering.

8.1.3.1 Remarkable hyperbolic identities

The purpose now is as for the circle trigonometry to prove some remarkable identities but for hyperbolic trigonometry.

So we start with the definition:

$$\cosh(z)=\frac{e^z+e^{-z}}{2}\qquad \sinh(z)=\frac{e^z-e^{-z}}{2}\tag{8.1.130}$$

and:

$$\tanh(z)=\frac{\cosh(z)}{\sinh(z)}=\frac{e^z-e^{-z}}{e^z+e^{-z}}=\frac{e^{2z}-1}{e^{2z}+1}\qquad \coth(z)=\frac{1}{\tanh(z)}=\frac{\sinh(z)}{\cosh(z)}=\frac{e^{2z}+1}{e^{2z}-1}\tag{8.1.131}$$

From these definitions and using the basic algebra operations we can determine the following outstanding relations (this is much easier than determining the remarkable relations of circle trigonometry, so unless requested we give most of these relations without proof).

First let us begin with the very easiest one:

$$\begin{aligned}&\cosh(z)+\sinh(z)=e^z &&\cosh(z)-\sinh(z)=e^{-z}\\&\cosh(-z)=\cosh(z)&&\sinh(-z)=-\sinh(z)\\&\tanh(-z)=-\tanh(z)&&\coth(-z)=-\coth(z)\end{aligned}\tag{8.1.132}$$

We also have:

$$\cosh^2(z) = \frac{1}{1-\tanh^2(z)} = \frac{\cosh(2z)+1}{2} \qquad \sinh^2(z) = \frac{\tanh^2(z)}{1-\tanh^2(z)} = \frac{\cosh(2z)-1}{2} \tag{8.1.133}$$

And we have the addition relations:

$$\begin{aligned}\cosh(a+b) &= \cosh(a)\cosh(b)+\sinh(a)\sinh(b)\\ \cosh(a-b) &= \cosh(a)\cosh(b)-\sinh(a)\sinh(b)\\ \sinh(a+b) &= \sinh(a)\cosh(b)+\cosh(a)\sinh(b)\\ \sinh(a-b) &= \sinh(a)\cosh(b)-\cosh(a)\sinh(b)\end{aligned} \tag{8.1.134}$$

Following the request of a reader, let us prove the first and third above relations.

For the first:

$$\begin{aligned}&\cosh(a)\cosh(b)+\sinh(a)\sinh(b) = \frac{e^a+e^{-a}}{2}\frac{e^b+e^{-b}}{2}+\frac{e^a-e^{-a}}{2}\frac{e^b-e^{-b}}{2}\\ &= \frac{1}{4}\left[\left(e^{a+b}+e^{a-b}+e^{-a+b}+e^{-a-b}\right)+\left(e^{a+b}-e^{a-b}-e^{-a+b}+e^{-a-b}\right)\right]\\ &= \frac{1}{4}\left[\left(2e^{a+b}+2e^{-a-b}\right)\right] = \frac{1}{2}\left[\left(e^{a+b}+e^{-(a+b)}\right)\right] = \cosh(a+b)\end{aligned} \tag{8.1.135}$$

and for the third:

$$\begin{aligned}&\sinh(a)\cosh(b)+\cosh(a)\sinh(b) = \frac{e^a-e^{-a}}{2}\frac{e^b+e^{-b}}{2}+\frac{e^a+e^{-a}}{2}\frac{e^b-e^{-b}}{2}\\ &= \frac{1}{4}\left[\left(e^{a+b}+e^{a-b}-e^{-a+b}-e^{-a-b}\right)+\left(e^{a+b}-e^{a-b}+e^{-a+b}-e^{-a-b}\right)\right]\\ &= \frac{1}{4}\left[\left(2e^{a+b}-2e^{-a-b}\right)\right] = \frac{1}{2}\left[\left(e^{a+b}-e^{-(a+b)}\right)\right] = \sinh(a+b)\end{aligned} \tag{8.1.136}$$

Also let us mention other remarkable identities (once again don't hesitate to request us the detail proof if you need help):

$$\begin{aligned}\cosh(a)+\cosh(b) &= 2\cosh\left(\frac{a+b}{2}\right)\cosh\left(\frac{a-b}{2}\right)\\ \cosh(a)-\cosh(b) &= 2\sinh\left(\frac{a+b}{2}\right)\sinh\left(\frac{a-b}{2}\right)\\ \sinh(a)+\sinh(b) &= 2\sinh\left(\frac{a+b}{2}\right)\cosh\left(\frac{a-b}{2}\right)\\ \sinh(a)-\sinh(b) &= 2\cosh\left(\frac{a+b}{2}\right)\sinh\left(\frac{a-b}{2}\right)\end{aligned} \tag{8.1.137}$$

and finally (we stop here because we don't need anything else to study the other chapters and sections of this textbook):

$$\begin{aligned}
\tanh(a+b) &= \frac{\tanh(a)+\tanh(b)}{1+\tanh(a)\tanh(b)}\\
\tanh(a-b) &= \frac{\tanh(a)-\tanh(b)}{1-\tanh(a)\tanh(b)}\\
\tanh(a)+\tanh(b) &= \frac{\sinh(a+b)}{\cosh(a)\cosh(b)}\\
\tanh(a)-\tanh(b) &= \frac{\sinh(a-b)}{\cosh(a)\cosh(b)}
\end{aligned} \tag{8.1.138}$$

8.1.4 Spherical Trigonometry

The aim of spherical trigonometry is to determine remarkable relations between the angles and sides of projected forms (also named "geodetic forms" because following the curvature of space) on the surface of a sphere. To determine these relations, we will look at the special case of a sphere of radius unity and to the relations between the sides of a spherical triangle (planar elementary surface) and the various existing angles (because this is used a lot in spherical triangulation!). We will see that the results are completely independent of the radius of the sphere.

Consider the figure in which is located a geodesic triangle with vertices A,B,C of respective opening angles $\hat{A},\hat{B},\hat{C}$ and opposite sides a,b,c and three unit vectors $\vec{i},\vec{j},\vec{k}$ such as $\vec{i}\perp\vec{k},\vec{j}\perp\vec{k}$ and that the end of $\vec{k}$ is merged with the vertices A:

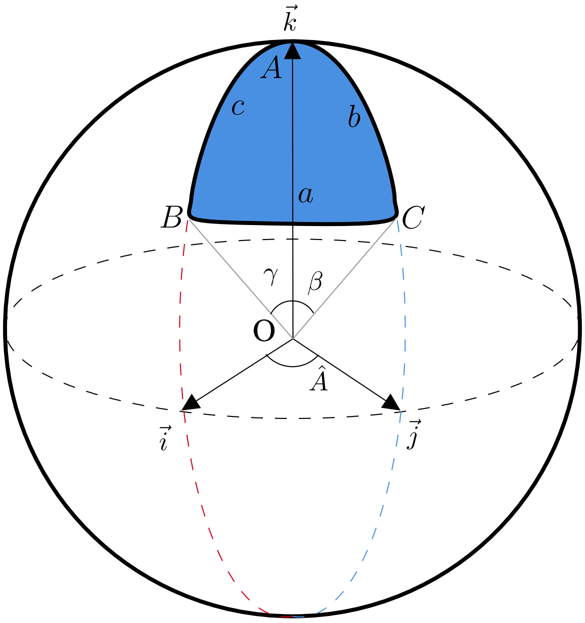
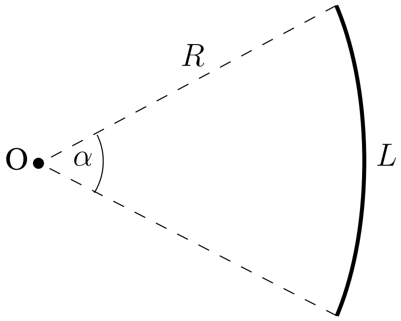


Figure 8.24: Construction to introduce the spherical geometry

The angle between points B and C, denoted by α, could not be shown in the diagram above because of lack of space.

Let us recall that the perimeter of a unit circle on the sphere of radius unity is obviously $P=2\pi R$ (see section Euclidean Geometry page 2620). The perimeter of the circle as a function of the opening angle of the latter being given by (relation very often used in physics !!!):

$$\boxed{L=2\pi R\frac{\alpha}{2\pi}=R\alpha} \tag{8.1.139}$$



If the circle is of radius $R=1$ (as is the case for our sphere), the calculation of the arc length is simplified and becomes:

$$L=\alpha \tag{8.1.140}$$

We will keep this constraint of a unitary radius subsequently to simplify the expressions we'll get later on.

Consequently regarding the points on our sphere; the sides of the triangle are given by:

$$a=\alpha, b=\beta, c=\gamma \tag{8.1.141}$$

Now consider the scalar product (see section Vector Calculus page 1821):

$$\overrightarrow{\text{O}B}\circ\overrightarrow{\text{O}C}=\|\overrightarrow{\text{O}B}\|\|\overrightarrow{\text{O}C}\|\cos(\alpha) \tag{8.1.142}$$

and as $\overrightarrow{\text{O}B}=\overrightarrow{\text{O}C}$ (unit radius) we have:

$$\|\overrightarrow{\text{O}B}\|\|\overrightarrow{\text{O}C}\|\cos(\alpha)=\cos(\alpha)=\cos(a)=\overrightarrow{\text{O}B}\circ\overrightarrow{\text{O}C} \tag{8.1.143}$$

If we decompose the two vectors $\overrightarrow{\text{O}B}$ and $\overrightarrow{\text{O}C}$ vectors on the orthonormal vectors we have:

$$\begin{aligned}\overrightarrow{\text{O}B}&=\cos(\gamma)\vec{k}+\sin(\gamma)\vec{i}\\ \overrightarrow{\text{O}C}&=\cos(\beta)\vec{k}+\sin(\beta)\vec{j}\end{aligned} \tag{8.1.144}$$

This gives us:

$$\cos(a)=\overrightarrow{\text{O}B}\circ\overrightarrow{\text{O}C}=\left[\cos(\gamma)\vec{k}+\sin(\gamma)\vec{i}\right]\circ\left[\cos(\beta)\vec{k}+\sin(\beta)\vec{j}\right] \tag{8.1.145}$$

giving (distributive property of scalar product):

$$\cos(a)=\cos(\gamma)\cos(\beta)\vec{k}\circ\vec{k}+\cos(\gamma)\sin(\beta)\vec{k}\circ\vec{j}+\sin(\gamma)\cos(\beta)\vec{i}\circ\vec{k}+\sin(\gamma)\sin(\beta)\vec{i}\circ\vec{j} \tag{8.1.146}$$

As $\vec{k}\circ\vec{k}=1$ and $\vec{k}\circ\vec{i}=\vec{k}\circ\vec{j}=0$ the above relation reduces to:

$$\cos(a)=\cos(\gamma)\cos(\beta)+\sin(\gamma)\sin(\beta)\vec{i}\circ\vec{j} \tag{8.1.147}$$

And as:

$$\vec{i}\circ\vec{j}=\cos(\hat{A}) \tag{8.1.148}$$

We get:

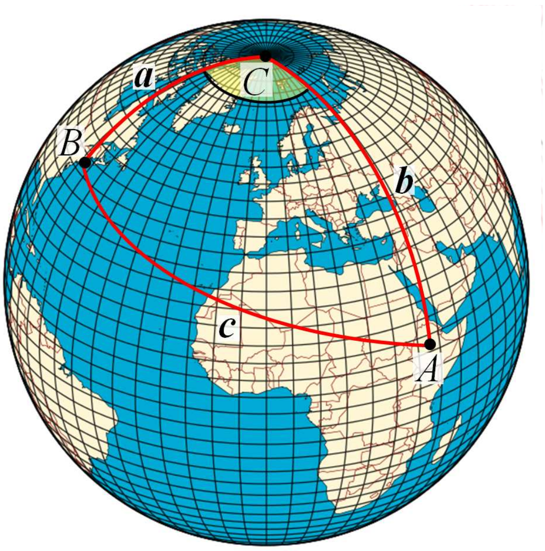
$$\cos(a)=\cos(\gamma)\cos(\beta)+\sin(\gamma)\sin(\beta)\cos(\hat{A}) \tag{8.1.149}$$

relation named "fundamental relation" or "cosine formula" that we can therefore as well write (because of the unitary radius assumption):

$$\boxed{\cos(a)=\cos(c)\cos(b)+\sin(c)\sin(b)\cos(\hat{A})} \tag{8.1.150}$$

✍**Example:**

On the globe below, Boston (B) and Addis Ababa (A) are plotted as the vertices of a spherical triangle with apex at the North Pole (C). Boston is at latitude $42°19'$ North which means edge a is subtended by a central angle of $90°-42°19'\approx 47.68°$.



Similarly Addis Ababa, latitude $9°03'$ North, gives the angle subtending edge b as about $80.95°$. And angle C is the sum of the longitudes: $71°05'\,\mathrm{W}+38°42'\,\mathrm{E}\approx 109.78°$. So the great circle distance c from B to A is given approximately by:

$$\cos^{-1}\left(\cos\left(48°\right)\cos\left(81°\right)+\sin\left(48°\right)\sin\left(81°\right)\cos\left(110°\right)\right) \tag{8.1.151}$$

which is about $98°=98\cdot 60$ nautical miles (1 nautical mile $=1$ minute of great circle arc), or about $10,900$ [km].

Also often rearranged as following (especially in statistics to introduce the partial correlation coefficient):

$$\cos(\hat{A})=\frac{\cos(a)-\cos(c)\cos(b)}{\sin(c)\sin(b)} \tag{8.1.152}$$

The latter relation is invariant by circular permutation of the variables $a,b,c,\hat{A}$. It is also interesting to note before we continue that if the spherical triangle has right angle at A, the previous relation simplifies to:

$$\cos(a)=\cos(c)\cos(b) \tag{8.1.153}$$

and if the triangle is sufficiently small relative to the radius and we make a second order Taylor expansion (see section Sequences and Series page 1696) near 0 for each of the terms we get:

$$\left(1-\frac{a^2}{2}\right)=\left(1-\frac{c^2}{2}\right)\left(1-\frac{b^2}{2}\right) \tag{8.1.154}$$

Therefore:

$$\left(1-\frac{a^2}{2}\right)=1-\frac{c^2}{2}-\frac{b^2}{2}+\underbrace{\frac{c^2b^2}{2}}_{\rightarrow 0} \tag{8.1.155}$$

After simplification:

$$a^2=b^2+c^2 \tag{8.1.156}$$

we find back the Pythagorean theorem (see section Euclidean Geometry page 2609). Therefore:

$$\cos(a)=\cos(c)\cos(b) \tag{8.1.157}$$

is the spherical geometry (non-Euclidean geometry) equivalent of the Pythagorean theorem in plane geometry (Euclidean geometry)!!!

This parenthesis now closed, let's us go back on our main business. The sine of all angles being positive (as all angles are $[0,\pi]$), we can write:

$$\begin{aligned}\frac{\sin(\hat{A})}{\sin(a)}=\frac{\sqrt{1-\cos^2(\hat{A})}}{\sin(\hat{a})}&=\frac{\sqrt{1-\left(\dfrac{\cos(a)-\cos(b)\cos(c)}{\sin(b)\sin(c)}\right)^2}}{\sin(\hat{a})}\\&=\frac{\sqrt{(1-\cos^2(b))(1-\cos^2(c))-\cos^2(a)-\cos^2(b)\cos^2(c)+2\cos(a)\cos(b)\cos(c)}}{\sin(a)\sin(b)\sin(c)}\\&=\frac{1-\cos^2(a)-\cos^2(b)-\cos^2(c)+2\cos(a)\cos(b)\cos(c)}{\sin(a)\sin(b)\sin(c)}\end{aligned} \tag{8.1.158}$$

The latter relation is of course also invariant by circular permutation of the variables $a,b,c,\hat{A}$. So we get a remarkable relation of the spherical triangle, named "sines relations" or "sines formula" and given by:

$$\boxed{\frac{\sin(\hat{A})}{\sin(a)}=\frac{\sin(\hat{B})}{\sin(b)}=\frac{\sin(\hat{C})}{\sin(c)}} \tag{8.1.159}$$

As spherical trigonometry is often used for land trails, often with two very specific and orthogonal circles: the Earth's equator and a meridian or any parallel, this case is of particular interest! Therefore in the case of a right triangle in A we obviously get:

$$\frac{1}{\sin(a)}=\frac{\sin(\hat{B})}{\sin(b)}=\frac{\sin(\hat{C})}{\sin(c)} \tag{8.1.160}$$

All the relationships that we have determined so far allow us, in the case where $\hat{A}=\pi/2$ and $R=1$, to draw very interesting relationships for geophysics:

$$\begin{aligned}\sin(b)&=\sin(a)\sin(\hat{B})\\ \sin(c)&=\sin(a)\sin(\hat{C})\\ \cos(a)&=\cos(b)\cos(\hat{c})\\ \cos(\hat{C})&=\sin(\hat{B})\cos(c)\\ \cos(\hat{C})&=\sin(\hat{B})\cos(c)\\ \cos(\hat{B})&=\sin(c)\cos(b)\end{aligned}\tag{8.1.161}$$

For sure, we have not presented here all the relations or remarkable identities existing about spherical trigonometry, but at least the most important one that should be know to read the other sections of this textbook.

Remark

We define the "surplus" or "spherical excess" by the number:

$$E=\hat{A}+\hat{B}+\hat{C}-\pi\tag{8.1.162}$$

8.1.4.1 Girard's theorem

While we are at it, let us calculate a classic problem which is that of the area of a triangle on a sphere. Consider the following figure:

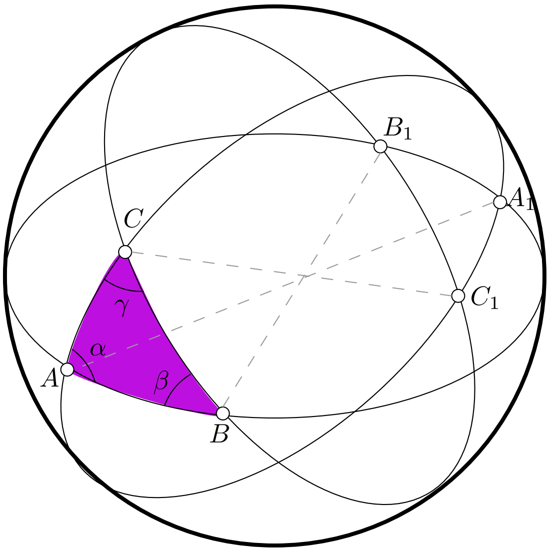


Figure 8.25: **Construction for the study of the surface of a spherical triangle**

If we extend the geodesic arcs $\overset{\frown}{AB}$ and $\overset{\frown}{AC}$ until A_1 we get a slice of a sphere whose surface is proportional to the angle α in A. If this angle was equal to 2π, we would have the whole sphere and the surface would be equal to $4\pi R^2$ (see section Geometric Shapes page 2964). As the angle is equal to α, the proportionality told us that S_1 is equal to:

$$S_1=\frac{\alpha}{2\pi}4\pi R^2=\alpha 2R^2 \tag{8.1.163}$$

Similarly, if we extend the arcs $\overset{\frown}{BC}$ and $\overset{\frown}{BA}$ up to B_1 and if we extend the arcs $\overset{\frown}{CA}$ and $\overset{\frown}{CB}$ until C_1, we get two slices with surfaces S_2 and S_3 are equal to:

$$S_2=\beta 2R^2 \quad S_3=\lambda 2R^2 \tag{8.1.164}$$

Now suppose we add these three areas:

$$S_1+S_2+S_3=2R^2(\alpha+\beta+\gamma) \tag{8.1.165}$$

we then get the half of the sphere $S_\odot$ (look to the figure to represent it yourself mentally) more two times the geodesic triangle area S in pink on the figure (as counted two times in excess).

We must subtract twice the blue triangle area to get the surface of the hemisphere:

$$\frac{S_\odot}{2}=2R^2(\alpha+\beta+\gamma)-2S \tag{8.1.166}$$

Therefore:

$$S_\odot=4R^2(\alpha+\beta+\gamma)-4S \tag{8.1.167}$$

as $S_\odot=4\pi R^2$, we get:

$$S=\frac{4R^2(\alpha+\beta+\gamma)-S_\odot}{4}=\frac{4R^2(\alpha+\beta+\gamma)-4\pi R^2}{4} \tag{8.1.168}$$

After simplification we deduce that the area S of the triangle ABC is:

$$S=R^2(\alpha+\beta+\gamma-\pi)=R^2\varepsilon \tag{8.1.169}$$

where ε is a solid angle (see below definition). That latter relation is sometimes named "Girard's theorem". More commonly written:

$$\boxed{\alpha+\beta+\gamma=\pi+\frac{S}{R^2}} \tag{8.1.170}$$

and $K=1/R^2$ is defined as the "curvature" (it indicates how much the geometry of the sphere deviates from the Euclidean plane geometry where $K=0$). On the basis of its definition K seems to depend on the radius of the sphere considered as a two dimensional manifold embedded in the ambient Euclidean space $\mathbb{R}^3$. The boxed relation above however tells us that the curvature K can be evaluated by just performing measures of angles and areas on the two-dimensional surface of the sphere. It follows that K has an intrinsic geometric meaning, that is independent of its embedding in the flat 3-dimensional space, since it can be computed by only using measures on the spherical surface.

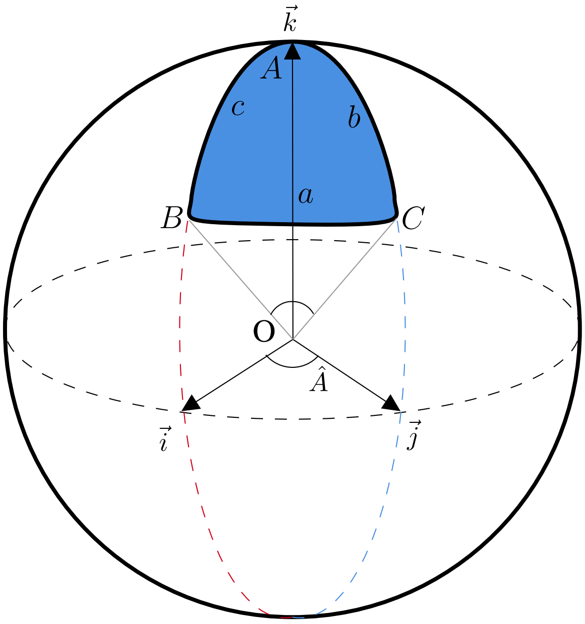
It is fairly simple to generalize this concept to other forms following the same philosophy (especially those composed of triangles...).

8.1.4.2 Four-parts formula

The "four-parts formulafour-parts formula" is a mathematical theorem in spherical trigonometry. It describes the relationship between four consecutive parts of a spherical triangle.

This formula is particularly useful in astronomy, geodesy, and navigation, where spherical triangles are commonly used. In this textbook we will use it to prove the property of a particular type of sundial!

Theorem 8.2. Let $\triangle ABC$ be a spherical triangle on the surface of a sphere whose center is O:



Let the sides a, b, c of $\triangle ABC$ be measured by the angles subtended at O, where a, b, c are opposite A, B, C respectively.

We have:

$$\cos(a)\cos(\hat{C}) = \sin(a)\cot(b) - \sin(\hat{C})\cot(\hat{B}) \tag{8.1.171}$$

That is:

$$\cos(\text{ inner side }) \cdot \cos(\text{ inner angle }) = \sin(\text{ inner side }) \cdot \cot(\text{ other side }) - \sin(\text{ inner angle }) \cdot \cot(\text{ other angle }) \tag{8.1.172}$$

This is known as the "four-parts formula", as it defines the relationship between each of four consecutive parts of $\triangle ABC$.

Proof 8.2.1. First we use the spherical cosine formula proved earlier above where we have for reminder:

$$\begin{aligned}\cos(b) &= \cos(a)\cos(c) + \sin(a)\sin(c)\cos(\hat{B})\\ \cos(c) &= \cos(b)\cos(a) + \sin(b)\sin(a)\cos(\hat{B})\end{aligned} \tag{8.1.173}$$

substituting for $\cos(c)$ we get:

$$\cos(b)=\cos(a)[\cos(b)\cos(a)+\sin(b)\sin(a)\cos(\hat{C})]+\sin(a)\sin(c)\cos(\hat{B}) \tag{8.1.174}$$

We distribute:

$$\cos(b)=\cos^2(a)\cos(b)+\cos(a)\sin(b)\sin(a)\cos(\hat{C})+\sin(a)\sin(c)\cos(\hat{B}) \tag{8.1.175}$$

We rearrange:

$$\cos(b)\left(1-\cos^2(a)\right)=\cos(a)\sin(b)\sin(a)\cos(\hat{C})+\sin(a)\sin(c)\cos(\hat{C}) \tag{8.1.176}$$

Hence using a simple trigonometric identity (sum of squares of sine and cosine):

$$\cos(b)\sin^2(a)=\cos(a)\sin(b)\sin(a)\cos(\hat{C})+\sin(a)\sin(c)\cos(\hat{B}) \tag{8.1.177}$$

dividing both sides by $\sin(a)\sin(b)$:

$$\cot(b)\sin(a)=\cos(a)\cos(\hat{C})+\frac{\sin(c)}{\sin(b)}\cos(\hat{B}) \tag{8.1.178}$$

Using the spherical law of since proved earlier above and given for reminder by:

$$\frac{\sin(\hat{A})}{\sin(a)}=\frac{\sin(\hat{B})}{\sin(b)}=\frac{\sin(\hat{C})}{\sin(c)} \tag{8.1.179}$$

we have:

$$\cot(b)\sin(a)=\cos(a)\cos(\hat{C})+\frac{\sin(\hat{C})}{\sin(\hat{B})}\cos(\hat{B}) \tag{8.1.180}$$

Rearranging and simplifying we get finally the "four-parts formula":

$$\boxed{\cos(a)\cos(\hat{C})=\sin(a)\cot(b)-\sin(\hat{C})\cot(\hat{B})} \tag{8.1.181}$$

■ Q.E.D.

8.1.5 Solid Angle

In spatial geometry, arise the problem of opening angle of a portion of the space (in extension to the so named "plane angle" that applies to $\mathbb{R}^2$). We define the "solid angle" by measuring the portion of space delimited by a conical surface of apex O and we express it in "steradian", given by the ratio (logical extension of the definition of the plane angle):

$$\Omega=\frac{S}{r^2} \tag{8.1.182}$$

S being the area of the cap cutted by a cone on a sphere of radius r.

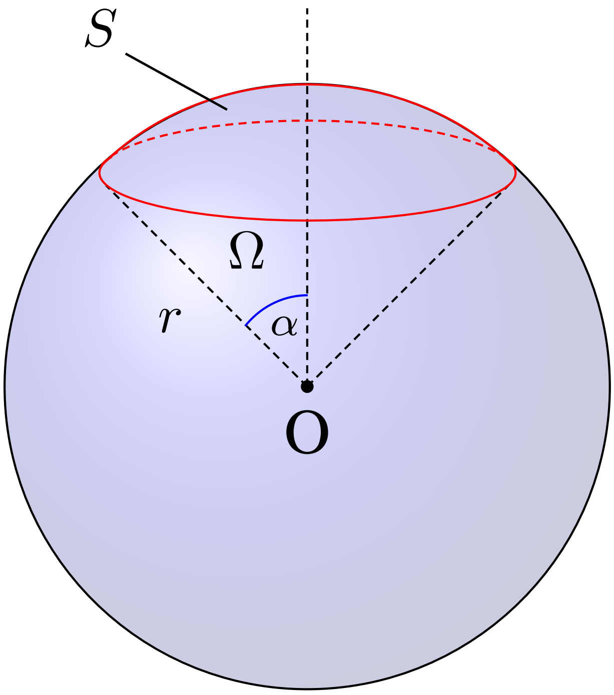


Figure 8.26: **Configuration for the definition of the solid angle**

If α is the half plane-angle of the cone, we get for this ratio (for calculation of the cap of a spherical surface see the section Geometric Shapes at page 2886):

$$\Omega = 2\pi(1-\cos(\alpha)) = 4\pi\sin^2\left(\frac{\alpha}{2}\right) \tag{8.1.183}$$

Hence we conclude that the total solid angle is by definition on the whole space:

$$\Omega_{\text{tot}} = 4\pi \tag{8.1.184}$$

We can also calculate the "elementary solid angle" as shown below:

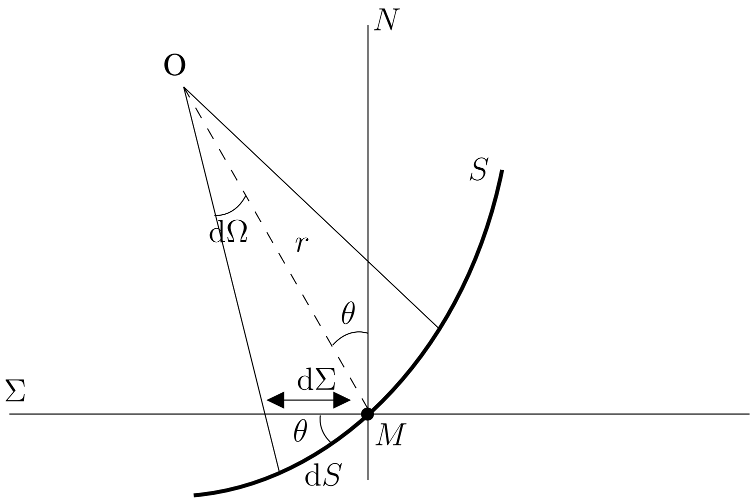


Figure 8.27: **Configuration for the definition of the elementary solid angle**

Given an elementary solid angle $\mathrm{d}\Omega$ and $\overline{\mathrm{O}M}$ the cone axis. We put:

$$\overline{\mathrm{O}M} = r \tag{8.1.185}$$

We consider any surface Σ passing through the point M. $\mathrm{d}\Omega$ cut on this surface a portion $\mathrm{d}\Sigma$.

If we draw the sphere $\mathcal{S}^2$ of center O and radius r, this solid angle cut on the sphere a cap of area $\mathrm{d}S$:

$$\mathrm{d}S = r^2\mathrm{d}\Omega \tag{8.1.186}$$

Given MN the normal to $\mathrm{d}\Sigma$ which make an angle θ with OM. We have assimilating $\mathrm{d}S$ and $\mathrm{d}\Sigma$ to portions of a plane:

$$\mathrm{d}S = \cos(\theta)\mathrm{d}\Sigma \tag{8.1.187}$$

Therefore:

$$\mathrm{d}\Omega = \frac{\mathrm{d}\Sigma\cos(\theta)}{2} \tag{8.1.188}$$

This concept of solid angle will be very useful especially in the field of theoretical physics that deals with the thermal radiation (see sections on Optics page 4339 and Thermodynamics page 3709 and Atomic Physics page 5100).

We can still calculate from the previous concepts, the elementary solid angle of revolution as presented in the figure below:

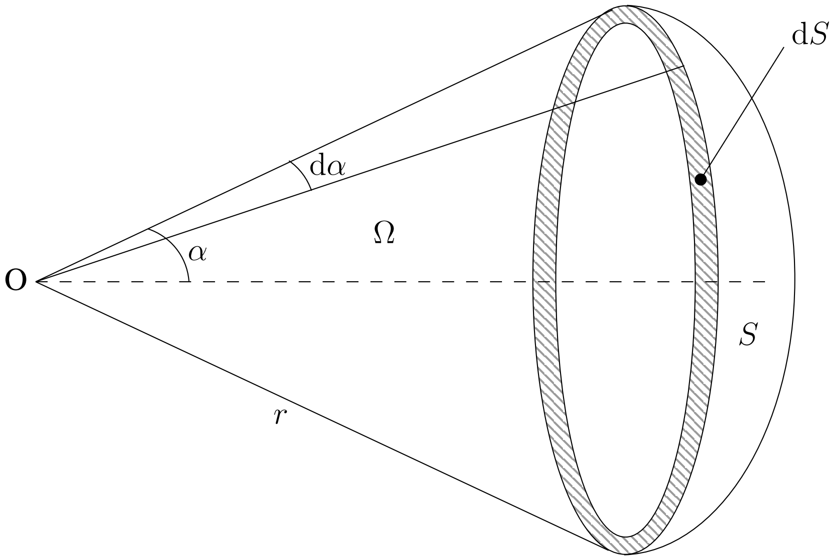


Figure 8.28: **Configuration for the definition of the elementary solid angle of revolution**

It is between two solid angles of revolution whose half-angles at the apex differ of $\mathrm{d}\alpha$:

$$\mathrm{d}\Omega = \mathrm{d}(2\pi(1-\cos(\alpha))) = 2\pi\sin(\alpha)\mathrm{d}\alpha \tag{8.1.189}$$

Therefore:

$$\mathrm{d}\Omega = \frac{\mathrm{d}S}{r^2} \tag{8.1.190}$$

Proof 8.2.2. In the section dealing with Geometric Shapes (see section Geometric Shapes page 2964) we proved the different ways to calculate the area of a sphere. From these calculations it was concluded that the elementary surface with constant radius R was:

$$\mathrm{d}S = r^2\sin(\theta)\mathrm{d}\theta\mathrm{d}\phi \tag{8.1.191}$$

and since:

$$\mathrm{d}\Omega=\frac{\mathrm{d}S}{r^2} \tag{8.1.192}$$

the elementary solid angle is then:

$$\mathrm{d}\Omega=\sin(\theta)\mathrm{d}\theta\mathrm{d}\phi \tag{8.1.193}$$

Thus, the solid angle defined by a cone of revolution, with an apex plane angle equal to α is given by:

$$\begin{gathered}\Omega=\int\mathrm{d}\Omega=\int\limits_0^{2\pi}\mathrm{d}\phi\int\limits_0^{\alpha}\sin(\theta)\mathrm{d}\theta=\phi|_0^{2\pi}\cdot(-\cos(\theta))|_0^{\alpha}\\=2\pi(-\cos(\alpha)-(-\cos(0)))=2\pi(-\cos(\alpha)+1)=2\pi(1-\cos(\alpha))\end{gathered} \tag{8.1.194}$$

■ Q.E.D.

~50 %

★★★☆☆

186 votes, 80.54%

8.2. Euclidean Geometry

The purpose of the "Euclidean geometry" (more commonly named "plane geometry") is, in principle, the study of forms and properties of natural bodies. The geometry is however not an experimental science, since its purpose is not to study certain aspects of nature, but a necessarily arbitrary reproduction of it.

We will introduce in this section implicitly, at first, the five postulates of the Euclidean geometry (the first four are nowadays regarded as axioms) and then develop around these basic geometry that will be necessary to the reader need to study the rest of this textbook. Once this is done, we will summarize our study by presenting explicitly the five postulates of Euclid's and then Hilbert's axioms.

Remark

We tried to preserve Euclid's notations as best as possible but although with a little more modern approach to certain concepts and present only those that are useful to Applied Mathematics.

8.2.1 Objects of Euclidean Geometry

Before stating the five Euclid's postulates, it seems good to set some intuitive concepts first (we will introduce the twenty-three Euclid's definition later just before his five postulates!):

Definitions (#8.2.320):

D1. The simplest experimental concept is that of "volume". We say that a body occupies a given volume when it occupies in the three-dimensional space a certain place (for spaces with higher dimensions, we talk about "hyper-volumes").

D2. We assume as obvious that a volume is limited by a "surface"; but if the existence of the volume is physically controllable and measurable, the surface is a creation of the mind; it is something similar to a balloon, for example, wrapping any volume, but only similar. It is a two-dimensional geometric without thickness.

D3. When a surface is limited, this limit is a "line". Again, the line is a creation of the mind, a line has no experimental existence; it is something similar to the figure formed by a wire. Still geometric entity but without width (otherwise it is a surface)... only a length.

D4. A "straight-line" is defined as the unbounded line of shortest distance joining two points on a surface. In other words, A straight line is a line with no beginning and no end, and therefore infinitely long.

D5. When a line is limited (bounded), its limit is a "point": the point is something analogous to the intersection of two stretched wires. This is still a creation of the mind, a geometric being (because it is supposed to have no width, no length, no height... just no dimensions).

Remark

It is customary in geometry to represent a point by a letter $A, B, \ldots$; a line or a surface by a letter in brackets (but this is rarely respected because we often assume the reader knows what we're talking about). Then we write, for example: the line (L), the surface (S).

D6. The term the "segment $\overline{AB}$" generally designates a line bounded by the points A and B. We say that a point M is on the segment $\overline{AB}$ is to translate the following fact: any segment $\overline{AB}$ can be split in an infinity of ways into two pieces limited by A and M on one hand, by M and B on the other - fact actually inspired by the experimental possibility to cut a piece of wire in half, and that in an infinite of ways (we will come back later about this).

Remark

The sentence: «The line (L) is drawn on a surface (S)» means that the surface (S) could be divided into several pieces, so that the line (L) is the boundary or part of a boundary of one of these pieces. This definition is based on the fact that it is possible to cut a tissue, for example by following with scissors any line on this tissue.

When a line (L) is drawn on a surface (S), any point M which is located on the line (L) is, by definition, also located on the surface (S). Then we say that it is a "point of this surface."

D7. A "semi-straight line" is a line which has a boundary (or vertex) on one side and is infinite on the other side.

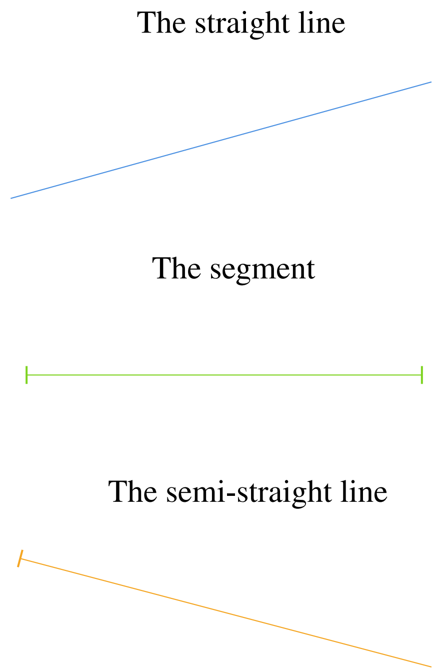


Figure 8.29: **Difference between line, segment and semi-straight line**

D8. We name "angle" (or "plane angle") or more strictly "straight angle" the plane portion limited by two half-lines (see further below the definition of a half-line).

8.2.1.1 Dimensions

We talked before about volume, surface and line to whom we can associate dimensions. But what is a dimension?

We will try to try to define it as best as possible but first, it is important to know that there are several types of dimensions in geometry. The best known and common one is that we name "topological dimension".

For example, the point (mathematical and geometrical abstraction) has a topological dimension of 0, the curve (continuous line of zero thickness) a dimension of 1, a surface a dimension of 2, a volume of a dimensions 3 and a hyper- volume a dimensions of 4 (to represent a hyper-volume take a volume drawn on a paper (...) make it a translation and connect the vertices). These are all integer values by definition:

Dimension: 0
Dimension: 1

(a) A point (b) A line (straight or curved)

Dimension: 2
Dimension: 3

(c) Plane figure/Surface (polygon, ellipse, etc.) (d) A solid or body (sphere, parallelepiped, etc.)

Table 8.3: **Objects, representations and corresponding dimensions**

In physics and mathematics, the dimension of a mathematical space (or object) is informally defined as the minimum number of coordinates needed to specify any point within it. Thus a line has a dimension of one because only one coordinate is needed to specify a point on it.

In mathematics, the dimension of an object is an intrinsic property independent of the space in which the object is embedded. For example, a point on the unit circle in the plane can be specified by two Cartesian coordinates, but a single polar coordinate (the angle) would be sufficient (for more details on coordinate systems see the section of Vector Calculus), so the circle is 1-dimensional even though it exists in the 2-dimensional plane. This intrinsic notion of dimension is one of the chief ways the mathematical notion of dimension differs from its common usages.

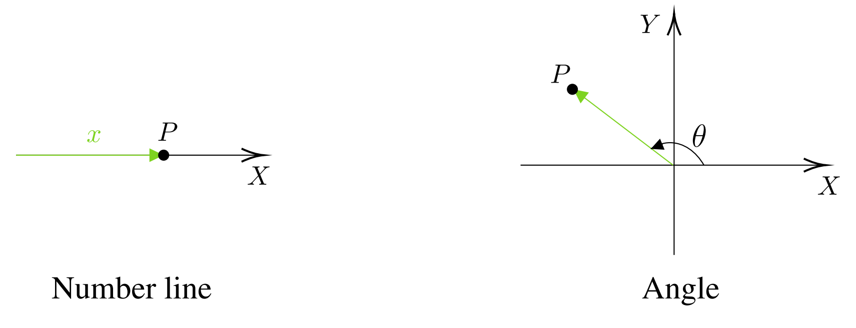
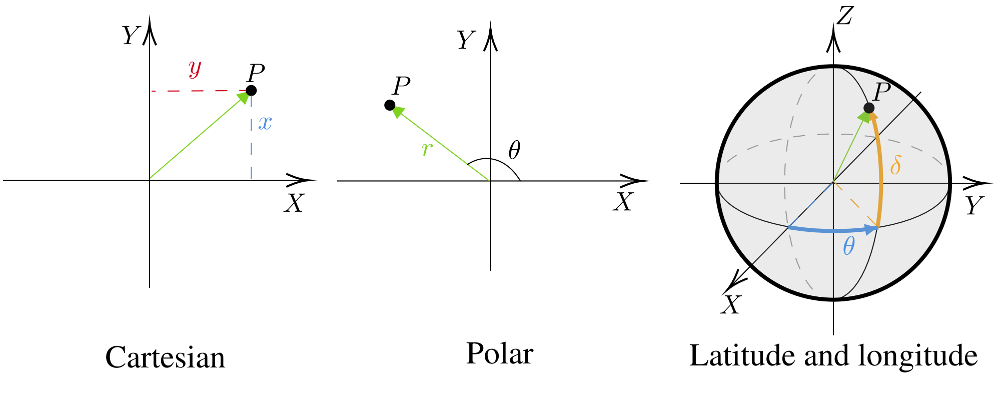
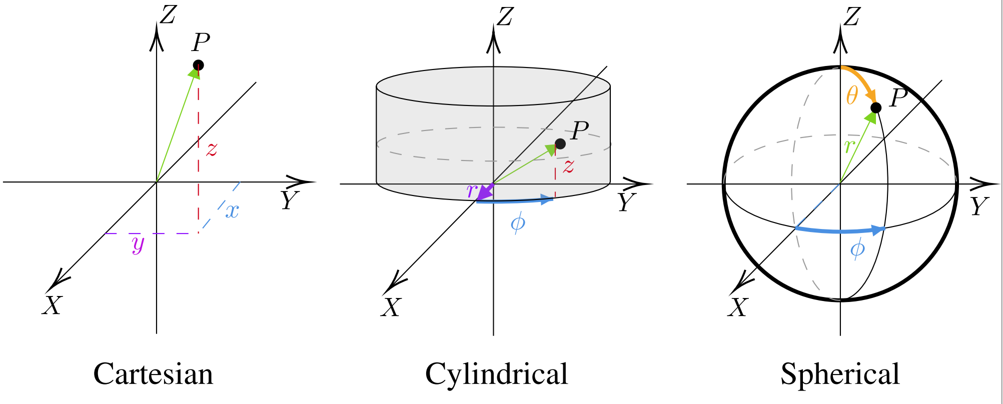
Dimensions	Example coordinate systems
1	 Number line Angle
2	 Cartesian Polar Latitude and longitude
3	 Cartesian Cylindrical Spherical

Figure 8.30: **Example of some coordinate systems**

To calculate the size of certain objects, we will use the method of the plane metric geometry by taking a standard measure of this object, that is to say the object itself but smaller, and to carry it over our object a given number of times:

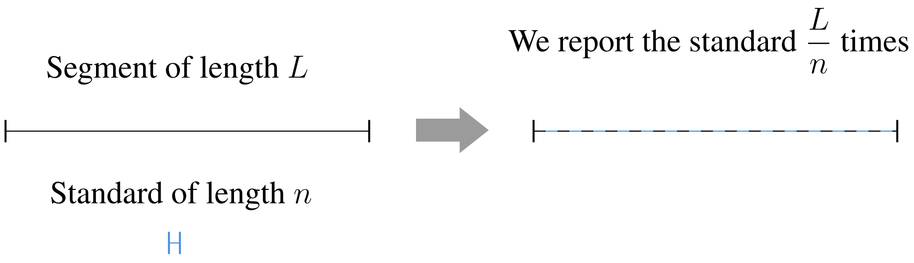

Segment of length L
We report the standard $\dfrac{L}{n}$ times
Standard of length n

Figure 8.31: **Concept of one-dimensional standard**

Let L be the total length of the segment. We will take a standard segment of length n that we will report on the big segment. This standard will be reported L/n times. We will write this:

$$\frac{L}{n}=\left(\frac{L}{n}\right)^{1} \tag{8.2.1}$$

We can apply the same reasoning to a surface:

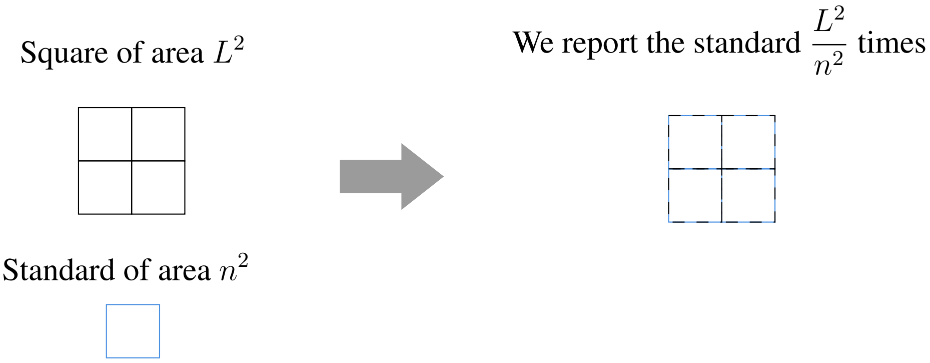

Square of area L^2
We report the standard $\frac{L^2}{n^2}$ times
Standard of area n^2

Figure 8.32: **Concept of two-dimensional standard**

Let L^2 be the total area of the square. We will take a standard square of area n^2 that we will report into the big square. This standard will be reported L^2/n^2 times. We note that:

$$\frac{L^2}{n^2}=\left(\frac{L}{n}\right)^{2} \tag{8.2.2}$$

We can apply the same reasoning to a volume:

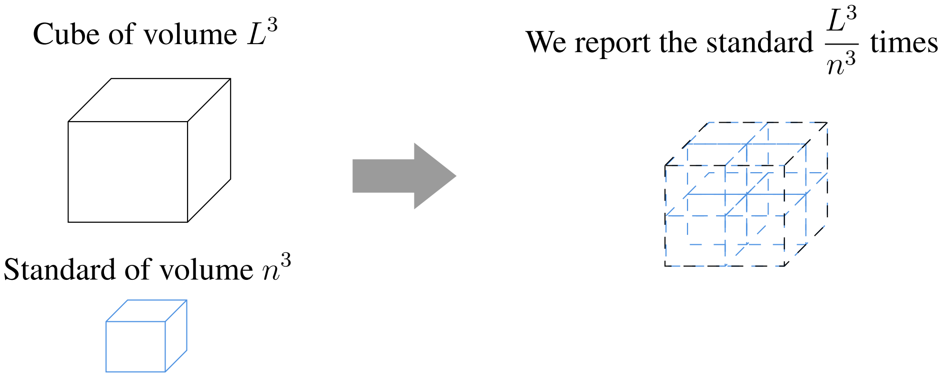

Cube of volume L^3
We report the standard $\frac{L^3}{n^3}$ times
Standard of volume n^3

Figure 8.33: **Concept of three-dimensional standard**

Let L^3 be the total volume of the cube. We will take a standard cube of volume n^3 that we will report into the big cube. This standard will be reported L^3/n^3 times. We note that:

$$\frac{L^3}{n^3}=\left(\frac{L}{n}\right)^{3} \tag{8.2.3}$$

and so on...

Through these three examples, we make appear the number 1 for the segment, the number 2 for the square, the number 3 for the volume. These numbers are the "dimension" of the object (same applies for a sphere for example).

So, to summarize:

1. The lines are of dimension 1, because to measure a length with a finer division by a factor n of a standard segment, the number of subdivisions will be multiplied by the same factor. So there is a unity power relation between the subdivision and the measurement.
2. Surfaces are two-dimensional, as for measuring a surface with a finer division by a factor n of a standard square, the number of subdivisions will be given by the square power of n. Therefore there is a power of two relation between the subdivision and the measurement (if we take the square two times smaller to cover a surface, we will need four times the standard).
3. Volumes are 3-dimensional, as for measuring a volume with a finer division by a factor n of a standard cube, the number of subdivisions will be given by the cubic power of n. So there is a power of three relation between the subdivision and the measurement (if we take cubes two time smaller to fill in a volume, we will need eight times the standard).

Let us generalize: let N be the number of times we report the standard of length n on our object of length L, and given d the dimension of the object, we have:

$$\boxed{N=\left(\frac{L}{n}\right)^2 \Leftrightarrow d=\frac{\ln(N)}{\ln(n)}} \tag{8.2.4}$$

In the case of fractals (see section Fractals page 7038) the dimensions are variable and fractional. Consider the Von Koch curve (for example) after one iteration of the sequence defining it:

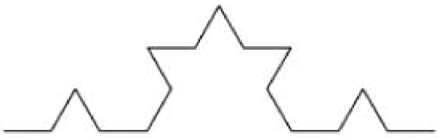


Figure 8.34: **von Koch curve**

Let L be its size such that $L=1$. To calculate its dimension we choose the fundamental element of the curve in red below (yes... we could take the segment of line for sure...):

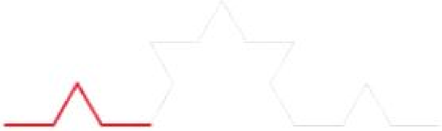


Figure 8.35: **Standard choice for Von Koch curve**

Let n be the size of this standard such as $n=1/3$. We see very well that we can report it 4 times on the curve. Therefore:

$$d=\frac{\ln(4)}{\ln\left(\dfrac{1}{1/3}\right)}=\frac{\ln(4)}{\ln(3)}\cong 1.26 \tag{8.2.5}$$

The dimension of the Von Koch fractal therefore has a fractional value and is more a curve (because close to 1), rather than a surface (which is of dimension 2). Cauliflowers are for example fractals with dimension 2.33...

So we can calculate the size of any fractal objects on condition of knowing their standard measurement.

We don't venture fetching complex objects in some galaxies, while the most famous fractal is in your plate (the second being your lungs ...). Eh yes! Cauliflower is a fractal (as well as your lungs)! You've probably noticed that when we cut cauliflower (not something recommended to try with your lungs ...), we are breaking it instead of cutting it, and it gives lots of small cauliflower, which themselves may give other smaller cauliflower. This characteristic of self-similarity at different scales is makes the cauliflower a fractal.

Let us calculate now the fractal dimension of cauliflower. When we break cauliflower, we get between 12 and 14 branches that resemble to the whole cauliflower close to a given scale... This expansion is, if we calculate it, of a factor 3. So according to the formula above, the fractal dimension of the cauliflower is approximately:

$$d=\frac{\ln(13)}{\ln(3)}\cong 2.33 \tag{8.2.6}$$

So the cauliflower is closer of a surface than a volume.

Let us do one last example with Sierpinski triangle fractal (see section Fractals page 7062):

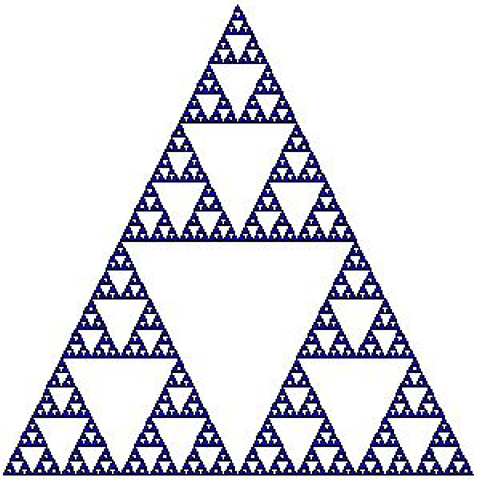


Figure 8.36: Sierpinski triangle fractal

First, it is clear that we need a square to cover it (round it) completely. With squares two times smaller, we will need three squares:

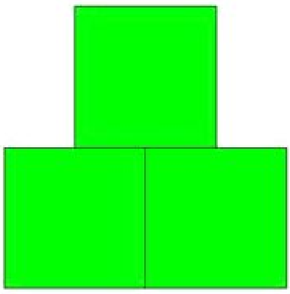


Figure 8.37: Three squares to cover the Sierpinski fractal

If we divide again once the size of the squares by two, we will need 9 squares to cover the triangle:

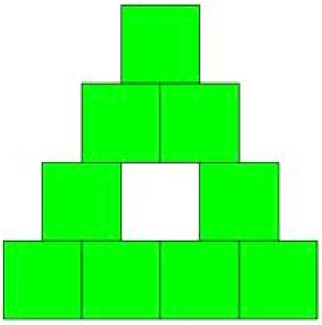


Figure 8.38: Nine squares to cover the Sierpinski fractal

If we divide once again the size of the square by two we will need 27 of them:

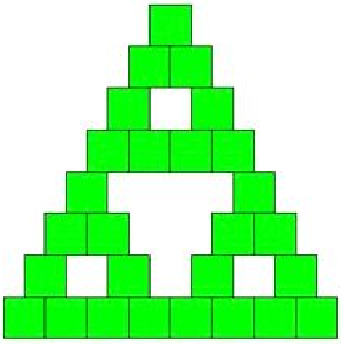


Figure 8.39: Twenty seven squares to cover the Sierpinski fractal

And there we see that:

$$3=2^{\frac{\ln(3)}{\ln(2)}},9=4^{\frac{\ln(3)}{\ln(2)}},27=8^{\frac{\ln(3)}{\ln(2)}} \tag{8.2.7}$$

And therefore:

$$\frac{\ln(3)}{\ln(2)}\cong 1.585 \tag{8.2.8}$$

So the Sierpinski fractal is closer to a surface than a curve.

There exist also other dimensions. Take for example the "homothetic dimensions" Here are some simple examples (see further below the strict definition of "homothetic transformation"):

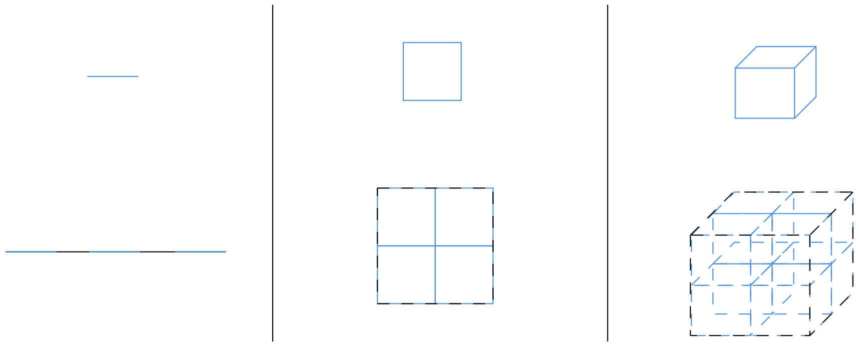


Figure 8.40: **Representation of homothetic dimensions**

The segment (far left), of 1-dimension, has by homothetic transformation seen its length doubled, we note that:

$$d=1, H=2, N=2^1=2 \tag{8.2.9}$$

The square (at the center), of 2-dimension has by homothetic transformation seen its surface doubled and we note that:

$$d=2, H=2, N=2^2=4 \tag{8.2.10}$$

The cube (far right), of 3-dimension has by homothetic transformation seen its volume has quadruple and we note that:

$$d=3, H=2, N=2^3=8 \tag{8.2.11}$$

The duplication of scale factor (homothetic) is therefore equal to:

$$2^{\text{Topological dimension}} \tag{8.2.12}$$

As you can see, this is always an integer value but of a different kind of dimension.

The concept of dimensions having been introduced let us look now to the Euclid's postulates which may seem vague at first but will be detailed as we go far in our reading of this section.

8.2.2 Euclid's Constructions

The construction of the planar Euclidean geometry is based on twenty-three definitions and five postulates (the first four are today regarded as axioms as we have already mentioned it).

Let us start with the traditional definitions:

Definitions (#8.2.321):

D1. A "point" is that which has no part.

D2. A line is a breadthless "length".

D3. The "ends" of a line are points.

D4. A "straight line" is a line which lies evenly with the points on itself.

D5. A "surface" is that which has length and breadth only.

D6. The "edges" of a surface are lines.

D7. A "plane surface" is a surface which lies evenly with the straight lines on itself.

D8. A "plane angle" is the inclination to one another of two lines in a plane which meet one another and do not lie in a straight line.

D9. And when the lines containing the angle are straight, the angle is named a "rectilinear angle".

D10. When a straight line standing on a straight line makes the adjacent angles equal to one another, each of the equal angles is a "right angle", and the straight line standing on the other is named a "perpendicular line" to that on which it stands.

Remark

As we will see in the section of Analytical Geometry (see page 2811), perpendicularity doesn't only exists with straight lines on a plane but also with curves like the famous case of a circle and a straight line passing through its origin!

D11. An "obtuse" angle is an angle greater than a right angle.

D12. An "acute" angle is an angle less than a right angle.

D13. A "boundary" is that which is an extremity of anything.

D14. A "figure" is that which is contained by any boundary or boundaries.

D15. A "circle" is a plane figure contained by one line such that all the straight lines falling upon it from one point among those lying within the figure equal one another.

D16. And the point is named the "center" of the circle.

D17. A "diameter" of the circle is any straight line drawn through the center and terminated in both directions by the circumference of the circle, and such a straight line also bisects the circle.

D18. A "semicircle" is the figure contained by the diameter and the circumference cut off by it. And the center of the semicircle is the same as that of the circle.

D19. "Rectilinear figures" are those which are contained by straight lines, trilateral figures being those contained by three, quadrilateral those contained by four, and multilateral those contained by more than four straight lines.

D20. Of "trilateral figures", an equilateral triangle is that which has its three sides equal, an isosceles triangle that which has two of its sides alone equal, and a scalene triangle that which has its three sides unequal.

D21. Further, of trilateral figures, a "right-angled triangle" is that which has a right angle, an "obtuse-angled triangle" that which has an obtuse angle, and an "acute-angled triangle" that which has its three angles acute.

D22. Of quadrilateral figures, a "square" is that which is both equilateral and right-angled; an "oblong" that which is right-angled but not equilateral; a "rhombus" that which is equilateral but not right-angled; and a "rhomboid" that which has its opposite sides and angles equal to one another but is neither equilateral nor right-angled. And let quadrilaterals other than these be named "trapezia".

D23. "Parallel straight lines" are straight lines which, being in the same plane and being produced indefinitely in both directions, do not meet one another in either direction.

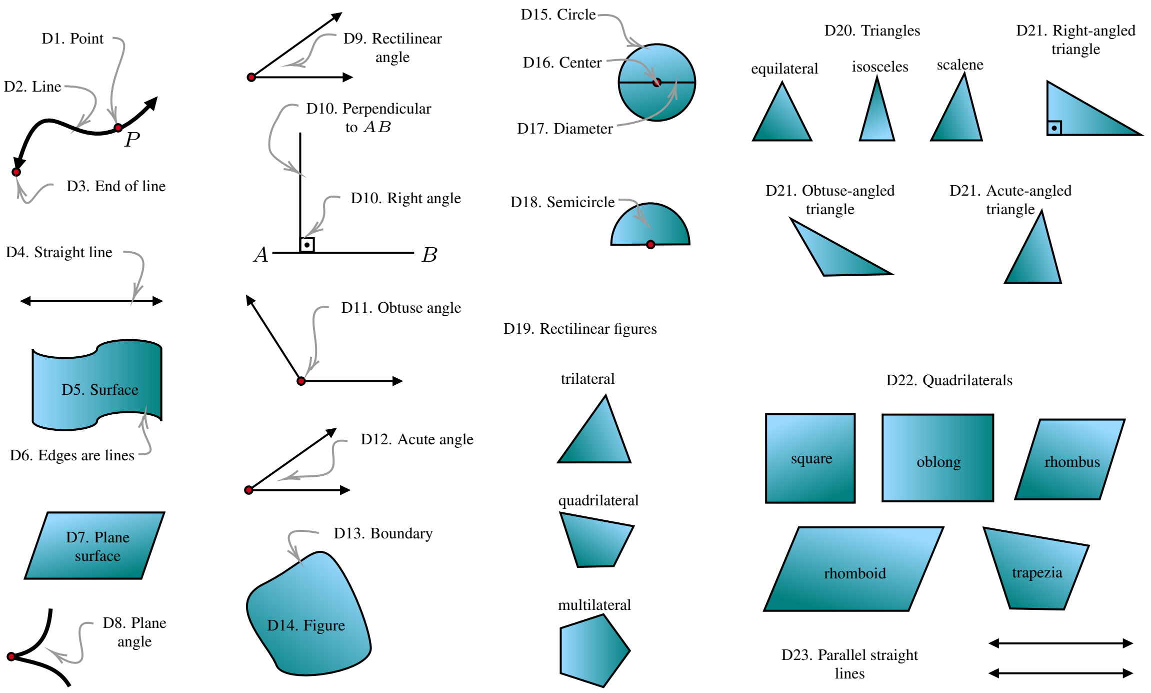

D1. Point
D2. Line
P
D3. End of line
D4. Straight line
D5. Surface
D6. Edges are lines
D7. Plane surface
D8. Plane angle
D9. Rectilinear angle
D10. Perpendicular to AB
D10. Right angle
A
B
D11. Obtuse angle
D12. Acute angle
D13. Boundary
D14. Figure
D15. Circle
D16. Center
D17. Diameter
D18. Semicircle
D19. Rectilinear figures
trilateral
quadrilateral
multilateral
D20. Triangles
equilateral
isosceles
scalene
D21. Right-angled triangle
D21. Obtuse-angled triangle
D21. Acute-angled triangle
D22. Quadrilaterals
square
oblong
rhombus
rhomboid
trapezia
D23. Parallel straight lines

Figure 8.41: **The 23 definitions of Euclid illustrated**

And now let us introduce Euclid's five postulates:

P1. Drawing a straight line of any point to any point.

In a modern way we would say that through two distinct points A and B, it passes a straight line and it goes in only one.

In other words, two straight lines (D) and (D') which have two common points are coincident, any point of one is a point of the other and vice versa.

It follows from this postulate that two lines (D) and (D') have no common point or have a single common point named "intersection" and are then "intersecting" and "distinct" or have more than one common point and are then "combined".

P2. Extend indefinitely, in its direction, a finite straight line.

In a modern way we will say that any finite segment $\overline{AB}$ can be extended in a straight line passing through A and B (given the first axiom, it is unique in Euclidean geometry).

P3. From any point and with any interval, describe a circle circumference.

In a modern way we will say that for any point A and any separate point B of A, we can draw a circle with center A passing through B.

P4. All right angles are equal.

Under modern form we say that for every angle $\widehat{xyz}$ of the plane corresponds its measurement θ (also denoted in high-school by $\measuredangle AOB$), performed with a unit chosen once for all where θ is a positive number, less than 2π. Conversely, let θ be any positive number between 0 and 2π, we shall assume that there is an infinite number of angles $\widehat{xyz}$ equal to one another whose measurement with the chosen angle unit is θ.

P5. If a straight line, falling on two straight lines, makes the interior angles on the same side smaller than two right angles, these lines, extended to infinity, will meet at the side where the angles are smaller than the two rights.

In a modern way we will say that given a straight line and a point, there is a unique line through this point and not intersecting the initial right.

Notice that these postulates were constructed in an old way as independent constructions. In modern mathematics we always try to define axioms (or postulates) based on relationships between concepts and not as independent constructions.

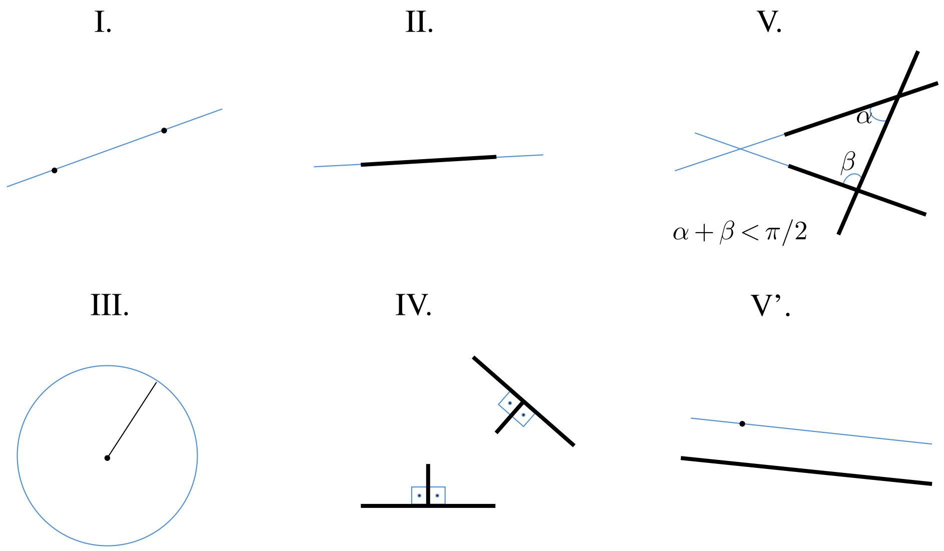


Figure 8.42: **Illustrated summary of five Euclid's postulates**

The Euclid's postulates will allow us the development of the concept of measurement of a length, area, volume, angle, as we shall see below.

Euclid's fifth postulate cannot be proven as a theorem from the four first postulates, although this was attempted by many people. Euclid himself used only the first four postulates ("absolute geometry") for the first 28 propositions of the *Elements*, but was forced to invoke the parallel postulate on the 29th. In 11823 (holocene calendar), Janos Bolyai and Nicolai Lobachevsky independently realized that entirely self-consistent "non-Euclidean geometries" could be created in which the parallel postulate did not hold (Gauss had also discovered but suppressed the existence of non-Euclidean geometries).

The two most fundamental theorems of Euclidean geometry are the Pythagorean theorem and the Thales theorem that we will prove later! A bit of analysis permits us to go further with Trigonometry that we have already developed in the previous section.

8.2.2.1 Segments and Lines

First, the simplest geometric figure (excepted the point ...) in Euclidean geometry is the "straight line" and it is directly concerned by the first two Euclid's postulates.

Definitions (#8.2.322):

D1. The "straight line" is the image given by a tensioned wire of zero thickness and infinite length.

> **Remark**
>
> We can also define the "straight line" as an infinity of points placed next to each other in a same direction on a plane.

D2. We name "half-line" the portion of a line limited to a point O named "origin".

Remark

The expression the "half-line $\overline{\mathrm{O}A}$" means the half-line of origin O, point named the "first", which contains the point A.

D3. We say that two half-lines $\overline{\mathrm{O}A}$, $\overline{\mathrm{O}B}$ are "opposing half-lines" when they constitute the whole line $\overline{AB}$.

D4. We name "segment $\overline{AB}$" the portion of a right straight line limited by two points A and B. These points are named the "extremities" of the segment.

8.2.2.1.1 Quantities of the same type

We say that geometric figures (meaning straight lines) are of the "same type quantities" when it is possible to define:

1. In which case a figure (A) is said to be equal to a figure (B) and, if they are unequal, which is the smallest.
2. What we need to understand when we sum a figure (A) and a figure (B).

The definitions should be chosen such that if (A) is smaller than said (B) and (B) smaller than (C), then (A) must be reported to be small than (C).

It is necessary, moreover, that the figure resulting of the sum of (A) and (B) is equal to that which is named: sum of (B) and (A).

Finally, a substitution in a comparison, or an equal amount, of a figure by an identical figure should not affect the result of operations.

To understand what are the quantities of the same type, take the example of line segments:

- We admit that it is possible to determine the equality of two segments $\overline{AB}$ and $\overline{A'B'}$ when we can make them coincide.
- We also admit that it is possible to replace the segment $\overline{A'B'} \in AB$ by an equal segment $\overline{AD}$ and putt on the line AB.
- Finally, we will assume that it is possible to distinguish between three points A, B, C, taken at random on a line or segment and to know which is between the other two.

We then agree to say that the segment $\overline{A'B'}$ is smaller than the segment $\overline{AB}$, which is written formally (see section Operators page 288):

$$\overline{A'B'} < \overline{AB} \tag{8.2.13}$$

when the point C, obtained by putting it on the segment $\overline{AB}$ a segment $\overline{AC}$ equal to $\overline{A'B'}$, falls between A and B.

If the point C was on B, the segments $\overline{AB}$ and $\overline{A'B'}$ would be equal and then we would write:

$$\overline{A'B'} = \overline{AB} \tag{8.2.14}$$

We agree to named "sum of two segments" $\overline{AB}$, $\overline{A'B'}$, the segment $\overline{AC}$ obtained by translating on the half-line opposed to the segment $\overline{BA}$ a segment $\overline{BC}$ equal to $\overline{A'B'}$. We translate this by writing:

$$\overline{AC} = \overline{AB} + \overline{A'B'} \tag{8.2.15}$$

Let us still consider quantities of the same species ... Add therebetween several of these quantities is to add one of them to another, the sum determined to another third, etc. For example, add the segments $\overline{AB},\overline{BC},\overline{CD}$, is add $\overline{AB}$ and $\overline{BC}$ which gives $\overline{AC}$, then $\overline{AC}$ and $\overline{CD}$ which gives $\overline{AD}$. We summarize the operation by writing:

$$\overline{AD}=\overline{AB}+\overline{BC}+\overline{CD} \tag{8.2.16}$$

Multiply by a quantity by an integer n, it is n quantities to itself. For example, if we have $\overline{AB}=\overline{BC}=\overline{CD}$, the above relation would be written:

$$\overline{AD}=3\overline{AB}=3\overline{BC}=3\overline{CD} \tag{8.2.17}$$

We will now define what we name "compare two quantities (A) and (B) of the same kind". For this let us arbitrarily select a quantity (C) of the same kind as (A) and (B) and smaller than each of them. Let us build a sequence of quantities such that:

$$(C_1)=(C),(C_2)=2(C),\ldots,(C_p)=p(C),\ldots,(C_n)=n(C) \tag{8.2.18}$$

We find that the quantity (A) is between two quantities (C_p) and (C_{p+1}) and that (B) is located between two others (C_q) and $(C_{q+1}$ by construction.

We then say by definition that the ratio of the quantities (A) and (B) is a positive number $(A)/(B)$ between:

$$\left[\frac{(C_p)}{(C_{q+1})},\frac{(C_{p+1})}{(C_q)}\right] \tag{8.2.19}$$

✍**Example:**

Consider then, for example, any two segments $\overline{AB},\overline{A'B'}$ but different (e.g. $1.2\,[\text{cm}]$ and $3.5\,[\text{cm}]$):

To perform the previous operation, let us use a graduated rule which unit will be the quantity (C), an arbitrary segment (for example $1\,[\text{cm}]$).

We apply the zero of the rule on A, and B will by design intercalate between two graduations of the rule numbered p and $p+1$ unless the size (C) is equal to $\overline{AB}$ ... (either with the choice taken as an example, B intercalate between the 1st and 2nd graduation).

For $\overline{A'B'}$, we apply the zero of the rule also on A', and B' also intercalate between two values of the rule numbered q and $q+1$ unless the variable C is equal to $\overline{A'B'}$ ... (either with the choice taken as an example, B' will intercalate between the 3rd and the 4th one).

We will express the results of these measurements and their ratio by writing:

$$\frac{p}{q+1}<\frac{\overline{AB}}{\overline{A'B'}}<\frac{p+1}{q} \tag{8.2.20}$$

where the term on the left extremity is named a "default measurement" and this at the opposite a "measurement by excesses".

So with the measures taken as an example we have:

$$\frac{1}{3+1}<\frac{\overline{1.2}}{\overline{3.5}}<\frac{1+1}{3}\Leftrightarrow 0.25<0.342<0.\bar{6} \tag{8.2.21}$$

Definition (#8.2.323): And in physics, we name "measure of a quantity" (A) the positive number which measures the ratio of this size and a quantity (U) arbitrarily chosen as reference and which we name the "unit", the measuring unit being "1", by definition.

We can show that if a is a measure of (A), b that of (B) evaluated both with the same unit (U), the number $(A)/(B)$ is equal to the ratio a/b. This ratio is (and this should be obvious for the reader at this level) independent of the chosen unit of measurement.

Remark

We say that (B) is an "aliquot part" of (A) if the ratio $(A)/(B)$ is an integer.

We will agree once and for all, that in geometry all quantities of the same species that occur in a given figure are measured with the same unit.

Given $(A),(B),(C),\ldots,(S)$ quantities of the same species and $(A'),(B'),(C'),\ldots,(S')$, quantities of the same species, but that are not necessarily of the same kind as the previous. We say that these quantities are "homologous" if we can group them in pairs, (A') homologous to (A), (B') homologous to (B), ..., etc., so that following conditions are met:

- If (A) is equal to (B), (A') is equal to (B').
- If (A) is smaller than (B), (A') is smaller than (B').
- If (S) is the sum of (A) and (B), (S') is the sum of (A') and (B').

To calculate the ratio $(A)/(B)$, let us build the following sequence $(C_1),(C_2),\ldots,(C_p)$ and to calculate the ratio $(A')/(B')$, let us build the sequence $(C'_1),(C'_2),\ldots,(C'_p)$ following equation, obtained as the previous one, but from the quantity (C') homolog of (C).

It is obvious that if (A) can be intercalate between (C_p) and (C_{p+1}), (A') could therefore also be intercalate between (C'_p) and (C'_{p+1}); Similarly, if (B) can be intercalated between (C_q) and (C_{q+1}). The ratios $(A)/(B)$ and $(A')/(B')$ will be bounded by the same numbers:

$$\frac{p}{q+1} \qquad \frac{p+1}{q} \tag{8.2.22}$$

Consequently, the ratio between two quantities (A) and (B) is equal to the ratio of the homolog quantities (A') and (B').

If, in particular, the quantities $(A),(B),\ldots$ are measured with a unity (U), and if the quantities $(A'),(B'),\ldots$ are measured with the sames unities (U'), homolog of (U), the equal ratios:

$$\frac{(A)}{(U)} \qquad \frac{(A')}{(U')} \tag{8.2.23}$$

are only the measurement of (A) and of (A'). Consequently:

The measurements of two homolog quantities (A) and (A') are equal at the condition that the chosen units to measure them are also homolog quantities.

Les us consider now on a half-line along Ox-axis a point M. Given x the measurement (following previous definition) of $\overline{\text{O}M}$. For every point M of the half-line $\overline{\text{O}M}$ corresponds a unique positive x. We will admit that to any positive value x randomly chosen correspond a unique point M on the half-line.

A consequence of this assumption is that there is a point, and only one, C, which divides the segment $\overline{\text{O}M}$ in equal parts. This point is the point of the half-line $\overline{\text{O}M}$, such that:

$$\overline{\text{O}C} = \frac{\overline{\text{O}M}}{2} \tag{8.2.24}$$

We name this point the "middle of the segment" $\overline{\text{O}M}$.

Theorem 8.3. There exists a point M and only one located on the segment $\overline{AB}$ such that the relation $\overline{MA}/\overline{MB}$ is equal to a given positive number.

Remark

Obviously if $\lambda = 1$, this point is the middle of the segment.

Proof 8.3.1. Given M any point of the segment $\overline{AB}$ on a plane; given x, the measurement of $\overline{AM}$ and a the measurement of $\overline{AB}$: the measurement of $\overline{MB}$ will be equal to $a-x$ since M is placed between A and B. We will have:

$$\frac{\overline{MA}}{\overline{MB}} = \frac{x}{a-x} \tag{8.2.25}$$

For this ratio to be equal to λ, we have to, and it is sufficient to, that x is solution of the equation:

$$\frac{x}{a-x} = \lambda \tag{8.2.26}$$

This equation has for unique solution:

$$x = \frac{a\lambda}{1+\lambda} \tag{8.2.27}$$

To this positive x value that is less than a corresponds a point M and only one of the half-line $\overline{AB}$ such as $\overline{MA} = x$. This point M satisfies, and satisfied alone, to the imposed conditions.

■ Q.E.D.

Theorem 8.4. There is a point M and only one of the line $\overline{AB}$, located outside the segment $\overline{AB}$, such as the ratio $\overline{MA}/\overline{MB}$ equals a given definite number λ and from 1.

Proof 8.4.1. Let us prove this uniqueness in two steps:

1. Let us suppose $\lambda > 1$. Given M any point of the line $\overline{AB}$ located outside the segment $\overline{AB}$: even A is on the $\overline{MB}$ segment, or B is on the $\overline{MA}$ segment. If A is the $\overline{MB}$ segment, we have necessarily $\overline{MB} > \overline{MA}$, thus:

   $$\frac{\overline{MA}}{\overline{MB}} < 1 < \lambda \tag{8.2.28}$$

   The point M thus does not respond to the question of uniqueness.

   If B is on the segment $\overline{MA}$ and we denote $\overline{MA} = x$, $\overline{AB} = a$, $\overline{MB} = x - a$. So we have:

   $$\frac{\overline{MA}}{\overline{MB}} = \frac{x}{x-a} \tag{8.2.29}$$

   For this ratio to be equal to λ, it is necessary, and sufficient that x is a solution of the equation:

   $$\frac{x}{x-a} = \lambda \tag{8.2.30}$$

   This equation admits as unique solution:

   $$x = \frac{a\lambda}{\lambda - 1} \tag{8.2.31}$$

   that gives a value of x that will always be positive and greater than λ. To this positive x value and greater than a corresponds a point M and only one of the half-line $\overline{AB}$. This point M satisfies, and only it, to the conditions of uniqueness imposed.

2. Let us suppose that $\lambda < 1$. We will seek the point M for which (we simply reversed the ratio):

   $$\frac{\overline{MB}}{\overline{MA}} = \frac{1}{\lambda} \tag{8.2.32}$$

   is a number greater than 1. There is one and only one from the point (1). It is the only one that satisfies the imposed conditions.

Remark

There is no point M located outside the segment $\overline{AB}$ for which $\overline{MA}/\overline{MB} = 1$. Indeed, if A is on the $\overline{MB}$ segment, we have regardless of M, $\overline{MA}/\overline{MB} < 1$, and if B is on the segment $\overline{MA}$, $\overline{MA}/\overline{MB} > 1$....

■ Q.E.D.

8.2.3 Plane Geometry

Let us now study a geometrical object of greater dimension than that of the straight line that is the Plane and the Surface.

A plane is a flat, two-dimensional surface that extends infinitely far. A plane is the two-dimensional analogue of a point (zero dimensions), a line (one dimension) and three-dimensional space. Planes can arise as subspaces of some higher-dimensional space, as with a room's walls extended infinitely far, or they may enjoy an independent existence in their own right, as in the setting of Euclidean geometry.

When working exclusively in two-dimensional Euclidean space, the definite article is used, so, "the plane" refers to the whole space. Many fundamental tasks in mathematics, geometry, trigonometry, graph theory and graphing are performed in a two-dimensional space, or in other words, in: "the plane".

Let us now consider a finite surface (S) and two points A and B of this surface. At least two simple cases are possible:

1. There are points of the straight segment $\overline{AB}$ that are not on the surface (S). We say in this case that: "the segment $\overline{AB}$ intersects the surface"; the common point to the segment $\overline{AB}$ and the surface (S) are the intersection points of the surface (S) and of the segment $\overline{AB}$. Among these common points there are especially the points A and B.

2. All points on the segment $\overline{AB}$ are points of the surface (S). We say then that: "the segment $\overline{AB}$ is on the surface (S)".

Definition (#8.2.324): We name "plane" the surface such that any segment $\overline{AB}$ joining two points arbitrarily chosen on the surface, is on the surface.

We will assume that such a surface exists and that in an Euclidean space of any number of dimensions, a plane is uniquely determined by any of the following:

- Three non-collinear points (points not on a single line).
- A line and a point not on that line.
- Two distinct but intersecting lines.
- Two parallel lines.

The following properties hold in three-dimensional Euclidean space but not in higher dimensions, though they have higher-dimensional analogues:

P1. Two distinct planes are either parallel or they intersect in a line.

P2. A line is either parallel to a plane, intersects it at a single point, or is contained in the plane.

P3. Two distinct lines perpendicular to the same plane must be parallel to each other.

P4. Two distinct planes perpendicular to the same line must be parallel to each other.

The study of planes will be made later in this section (but the mathematical vectorial analysis of a plane is done the in the section of Analytical Geometry). We will first focus on the study of geometric figures drawn in a given plane, figures named "plane figures". Their study is related to the field of "plane geometry".

All the shapes exist in a flat plane. A plane can be thought of an a flat sheet with no thickness, and which goes on for ever in both directions. It is absolutely flat and infinitely large, which makes it hard to draw.

Remarks

R1. The study of geometry can be broken into two broad types: plane geometry, which deals with only two dimensions, and solid geometry which allows all three. The world around us is obviously three-dimensional, having width, depth and height, Solid geometry deals with objects in that space such as cubes and spheres.

R2. In practice, the figures are drawn either on a sheet of paper or on the surface of the blackboard (or on a computer screen).

8.2.3.1 Displacements and Turnarounds

Given (F) a drawing made on a plane board. Imagine we do on a transparent layer, which front is applied on the board, a copy (C) of the drawing (F). Let us perform afterwards, by moving this transparent layer on another point of the plane board, a new drawing a new drawing (F') identical to (F).

Two major cases have to be considered mathematically speaking:

1. If the recto side of the transparent layer is remained applied to the plane board, the drawing (F') is obtained from (F) by an operation named "displacement" or more commonly "translation".
2. If, however, the transparent layer was horizontally or vertically returned, and if it is the verso that is applied to the plane board, the operation is named "turnaround" or more commonly a "vertical or horizontal symmetry".

At each point A of the drawing (F), any one of these two operations corresponds to a point A' of the drawing (F'), which we name "the homologue" of A. A segment $\overline{AB}$ of (F) coincides with a homologous segment $\overline{A'B'}$ of (F'); these two segments are by definition equal regardless of A and B.

In the figure below the red triangle is a translation of the blue one. The green triangle is a horizontal symmetry of the blue one plus a translation:

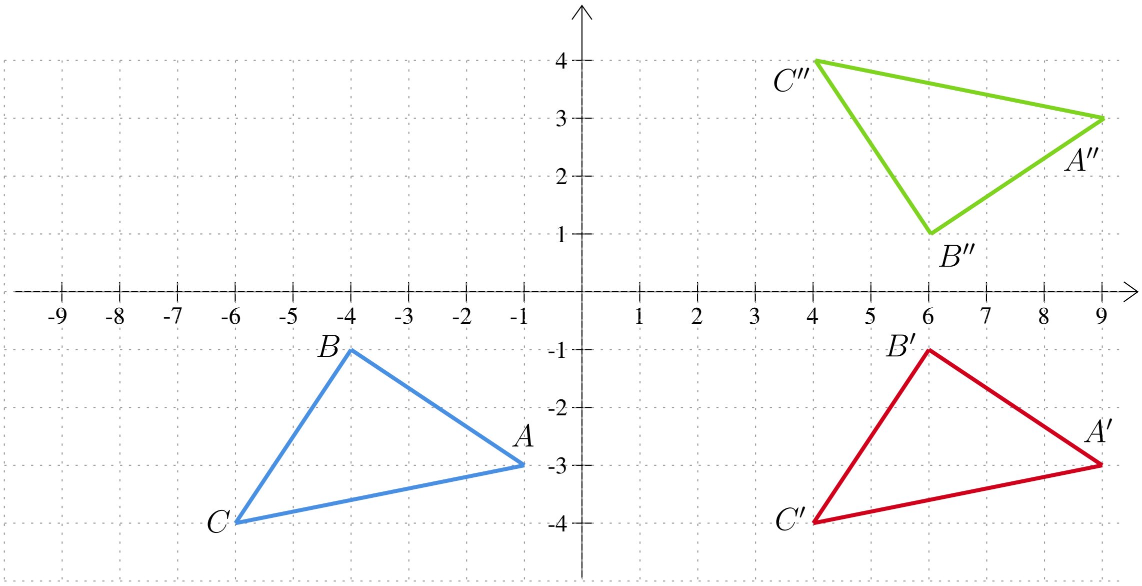


Figure 8.43: **Illustrated example of a translation and a symmetry**

Definition (#8.2.325): We commonly say that the drawing (F) is "stackable" to drawing (F') and that these drawings represent equal figures.

8.2.3.2 Plane angles

We have already defined the concept of "angle" in the previous section on Trigonometry (circle angle and spherical angle). We now have more the fourth Euclid's postulate at our disposal regarding the concept of plane angle.

We will now come back more in detail about the concepts of plane angle and see the underlying concepts that will enable us to address further a particularly useful object which is the bisecting line!

Definitions (#8.2.326):

D1. We name "angle" (or "plane angle") or more strictly "linear angle" the plane portion bounded by two half-lines OA, OB, for example. The point O is named the "top" of the angle, the half-lines OA, OB are named the "sides" of the angle.

In highs-school, angles are measured with an instrument named a "protractor":

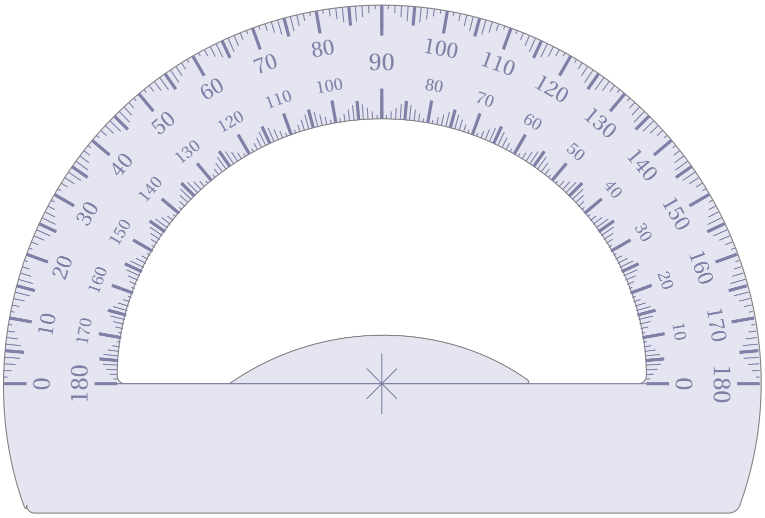


Figure 8.44: **A half circle protractor marked in degrees** (180°)

D2. We name "angle between two half-lines AB, AC", the angle of top A which sides are the half-lines AB, AC.

D3. The half-lines OA, OB divide the plane into two regions: they therefore define two angles:

(a) One consists of the covered hatch area (see figure on the far left below) is named "salient angle".

(b) The other consists of the covered hatch covered (see figure on the center below) is named "reflex angle".

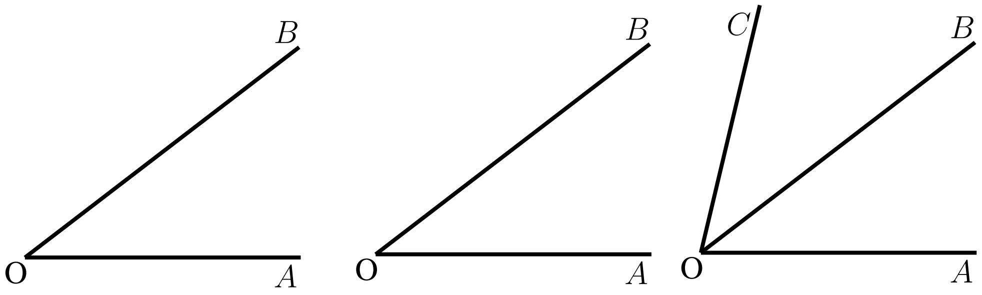


Figure 8.45: **Salient angle, reflex angle, adjacent angle**

The notation $\widehat{BOA}$ or $\widehat{AOB}$ denotes one of the two angles: the letter that designate the top should be (usually) written in the middle (often we do not mention the top if the context is clear). Where no precision accompanies this notation, it is by definition the salient angle!!

D4. We name "adjacent angles" two angles which have the top and one side that are common and which are placed on either side of the common side. In the figure above the far right, the saillant angles $\widehat{AOB}$, $\widehat{BOC}$ are adjacent.

Given $\widehat{AOB}$ and $\widehat{A'OB'}$ two angles of a sample plane. We have assumed previously that there is a displacement that brings O$'$ on O and A' on a point A of the line OA. This movement causes $\text{O}'B'$ to move either on $\text{O}B_1$ so that the two angles $(\widehat{A\text{O}B},\widehat{A\text{O}B_1})$ are not adjacent, or on $\text{O}B_2$ so that $(\widehat{A\text{O}B},\widehat{A\text{O}B_2})$ are adjacent.

In the latter case, an additional half turn around $\text{O}A$ will bring $\widehat{A\text{O}B_2}$ into the position $\widehat{A\text{O}B_1}$. This movement and this rotation, if it occurs, replace $\widehat{A'\text{O}'B'}$ by an angle $\widehat{A\text{O}B_1}$ that is equal by definition.

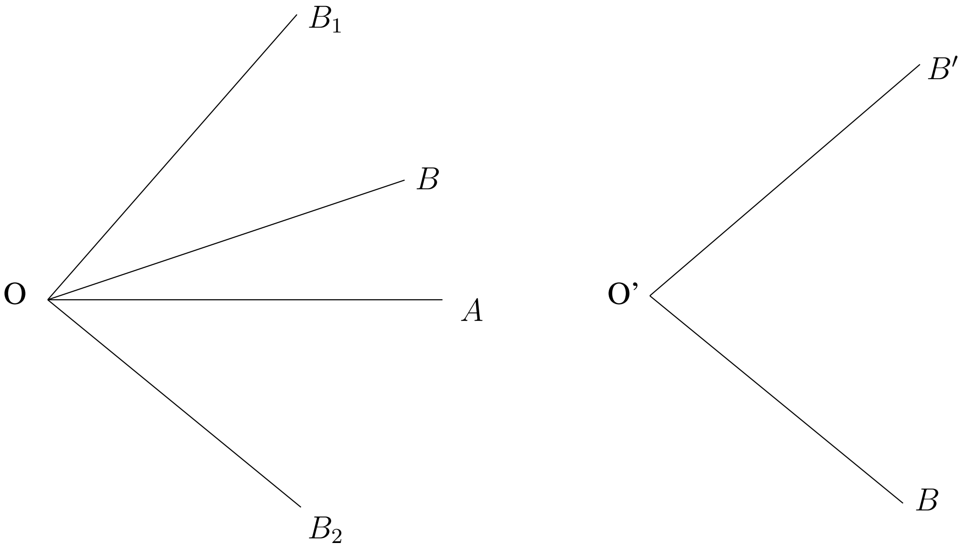


Figure 8.46: **Representation of the movement of two angles**

If $\text{O}B_1$ is merged with $\text{O}B$, there are points of one of the two angles $\widehat{A\text{O}B_1}$ and $\widehat{A\text{O}B}$ that are equal, any point of one being a point of the other, we say, in this case that: the angles $\widehat{A\text{O}B}$, $\widehat{A'\text{O}'B'}$ are "equal angles", which is expressed by the expression:

$$\widehat{A\text{O}B}=\widehat{A'\text{O}'B'} \tag{8.2.33}$$

If $\text{O}B_1$ is not merged with $\text{O}B$, there are points of one of the two angles $\widehat{A\text{O}B_1}$, for example, which are not points of $A\text{O}B$. On the figure above, the angle $\widehat{A\text{O}B}$ is covered by hatch, the angle $\widehat{A\text{O}B_1}$ also. The points we are speaking about are those of the angle $\widehat{B\text{O}B_1}$ covered only once by hatch. We will agree to say that the angle $\widehat{A\text{O}B_1}$ and, therefore, the angle equal to $\widehat{A'\text{O}'B'}$ are, in this case, greater than the angle $\widehat{A\text{O}B}$, which is expressed by the inequality:

$$\widehat{A'\text{O}'B'}>\widehat{A\text{O}B} \tag{8.2.34}$$

Now that we are able to compare angles, let us study how we can sum the up (and thus subtract respectively).

Let us study first the case of the sum of two adjacent angles $\widehat{A\text{O}B}$, $\widehat{B\text{O}C}$. Two cases can occur following that the angles are salient angles or reflex angles:

(a) Given $\widehat{A\text{O}B}$ and $\widehat{B\text{O}C}$ the two angles to sum (see left figure below). By definition, the sum of these angles is the angle $\widehat{A\text{O}C}$, what we express by the equality:

$$\widehat{A\text{O}C}=\widehat{A\text{O}B}+\widehat{B\text{O}C} \tag{8.2.35}$$

(b) Given $\widehat{AOB}$ a salient angle to sum up to the reflex angle $\widehat{BOC}$. If we cover of hatch successively the both angles (see right figure below), the salient angle $\widehat{AOC}$ is covered twice:

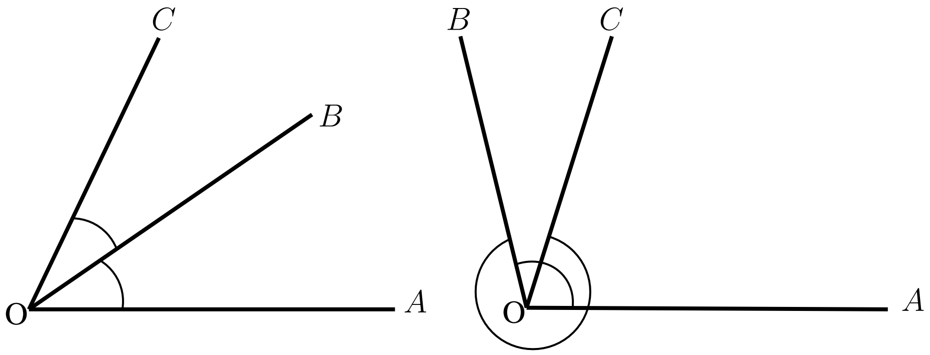


Figure 8.47: **Sum of two adjacent angles**

In this case, the sum of the angles $\widehat{AOB}$ (saillant) and $\widehat{BOC}$ (reflex) is therefore equal to $\widehat{AOC}$ plus two "flat angles" (see below for the definition), which is expressed by writing:

$$\widehat{AOB}+\widehat{BOC}=\widehat{AOC}+2\,\text{flat angles} \tag{8.2.36}$$

Remark

This may seem confusing to some but those who have already read the section of Trigonometry know already that the angles of the trigonometric circle are equal to themselves modulo 2π [rad].

Let us now study the case of the sum of any two angles:

The sum of two angles $\widehat{AOB}$, $\widehat{A'O'B'}$ is, by definition, equal to the sum of the angle $\widehat{AOB}$ and of an angle $\widehat{BOC}$ equal to the angle $\widehat{A'O'B'}$ and adjacent to the angle $\widehat{AOB}$.

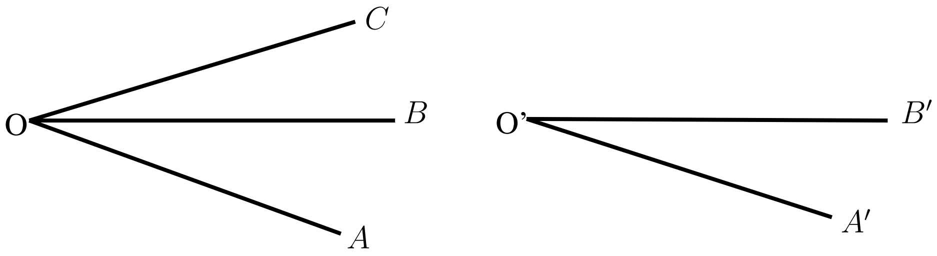


Figure 8.48: **Sum of any two angles**

Such an angle is obtained by a displacement which brings O on O$'$ and A at a point OA, followed or not by a rotation around OA.

Let us study for last case the sum of more than two angles:

The sum of several angles $\widehat{AOB}$, $\widehat{A'O'B'}$, etc., is by definition equal to the sum obtained by adding the first to the second, this sum to the third, and so on.

Given $\widehat{AOB}$ the first angle, $\widehat{BOC}$ the second angle, $\widehat{KOL}$ an angle equal to the last angle to add and adjacent to the previous one. The result of the sum will be $\widehat{AOL}$ augmented as many times of two flat angles that the plane was recovered during this operation. We easily see that this result does not depend on the order of angles $\widehat{A\text{O}B}$, $\widehat{A'\text{O}'B'}$ to add, etc.

D5. Two angles formed by two lines cut by a secant are named "alternate angles" if:

  (a) They are located on either side of the secant
  (b) They are located between the two lines
  (c) They are not adjacent angles

In the example below, the straight lines (X) and (Y) are cut respectively in A and B by the secant (S):

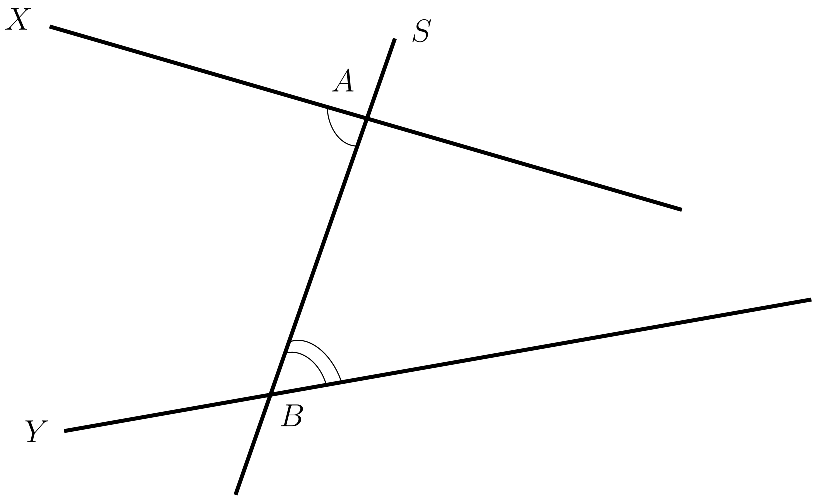


Figure 8.49: **Example of alternate angle**

and the two shown angles are alternate angles. In the case where (X) and (Y) are parallel, the alternate angles are equal.

It is also this which seems to have allowed Eratosthenes to deduce the circumference of the Earth from a purely geometrical way (thus it is not useless!)... even if there is quite good evidence that someone else did it long time before and that in reality Eratosthenes didn't do the measurements himself but just took the results reported in an old papyrus stored in the library of Alexandria[1].

✍**Example:**

So, according to the legend Eratosthenes used the observation he made on the shadow of two objects in two different places, Syene (now Aswan) and Alexandria, considered on the same meridian, the June 21 (summer solstice at his time) at local solar noon. It is at this precise time of the year that in the northern hemisphere the Sun (actually a G2V-type star, aka yellow dwarf) holds the highest position above the horizon. In a previous observation, Eratosthenes noticed that there was no shadow in a water well at Syene. Thus, at that moment, the Sun was vertical and its light shone directly the bottom of the water well. Eratosthenes noticed however that the same day at the same hour, an obelisk located in Alexandria formed a shadow. The Sun was no longer vertical and the obelisk had an off centered shadow. Eratosthenes considered as parallel the light rays of the Sun at any point on the Earth (considered as spherical). He deduced that the angle between the sunlight and the vertical was of 7.2 degrees.

---

[1] Indeed, the description of the experience - especially the position of the Sun at noon in a given city- leads to the conclusion that the actual experiment and record of this in Egypt occurred very likely 3,500 years before Eratosthenes time!!!

Eratosthenes then estimated the distance between Syene and Alexandria by using a bématiste which extrapolate upon the time of camel-day journey between the two cities: the distance achieved was estimated as 787.5 kilometers (measure very close to reality).

By the geometric theory of alternate congruent angles Eratosthenes proposed a simple figure: it consisted of a single circle having a central angle of 7.2 degrees which intercepts an arc (connecting Syene to Alexandria) of 800 kilometers.

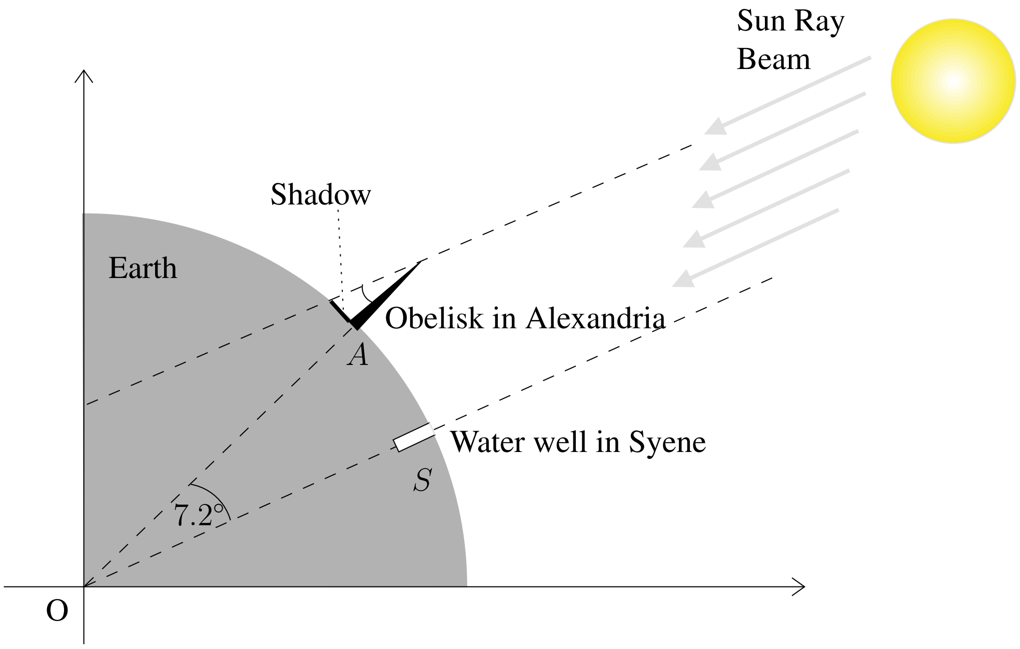
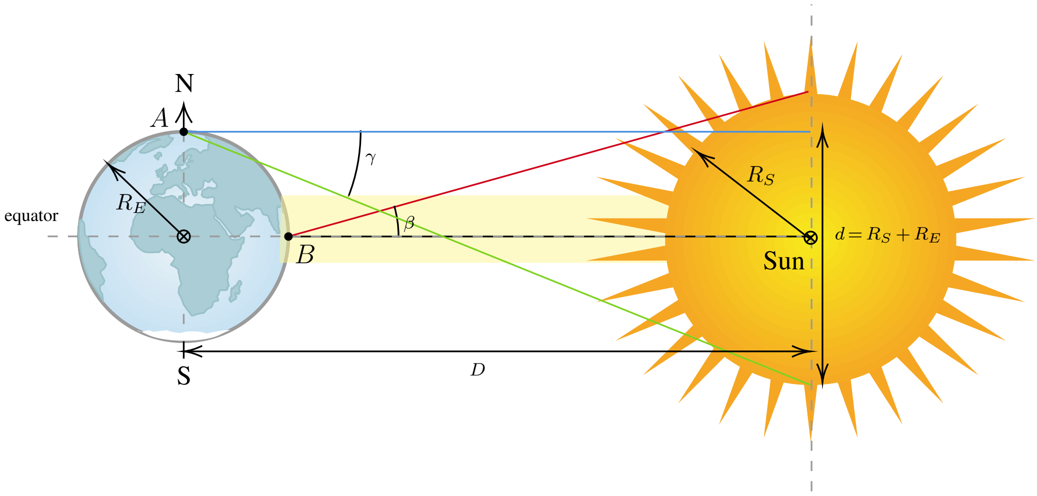


Figure 8.50: Alternate angle principle as used by Eratosthenes

By the ratio in the circles (already known at that time), he calculated the circumference of the Earth was by a simple rule of three equal to $39,375$ kilometers. What was a remarkably accurate measurement for that time (the actual measures give an average of $40,075.02$ kilometers).

However...is the parallel rays assumption good enough? To answer let us consider the following figure:



Where we take for the radius of Earth $R_E \cong 6,371\,[\text{km}]$, for the radius of the Sun $R_S \cong 696,340\,[\text{km}]$ and for the average distance center-to-center of $D \cong 149,597,871\,[\text{km}]$ (see derivation in the section of Astronomy page 5218).

Let us compute first β (in degrees) that is half the angular size (angular diameter) of the Sun and corresponds to the most oblique angle a ray of light can be seen from the equator at the summer solstice (point B in the figure above):

$$\beta = \arcsin\left(\frac{R_S}{D - R_E}\right) \cong 0.26670959° \tag{8.2.37}$$

in most tables we can found $2 \cdot \beta \cong 0.5332°$.

Now let us take the most extreme case in the figure above doing the same kind of calculation for the point A. For the latter we have:

$$\gamma = \arcsin\left(\frac{d}{D}\right) = \arcsin\left(\frac{R_S + R_E}{D}\right) \cong 0.269138339° \tag{8.2.38}$$

Hence in the worst case we have a difference of $0.00242875°$ between both most extreme angles in people look on the same point of the Sun. Or if we add up (people on point A and B look in opposite direction) we have $\beta + \gamma = 0.5358°$. So whatever the scenario we take $\beta - \gamma = 0.00242875°$ or $\beta + \gamma = 0.5358°$ the claim - or approximation - of parallel rays is quite acceptable.

Caution!!! Eratosthenes calculation is not an experimental evidence that the Earth is a geoid (spherical shaped like!). Indeed, the calculations would lead to the same result if the Earth was flat, toroidal or many other non-spherical shapes! Better direct experimental evidence are: 1) Satellites photos, 2) Astronauts direct observations 3) The observation of low-pressure weather systems that rotate anti-clockwise in the northern hemisphere due to the Coriolis effect 4) Foucault's pendulum shifts with time 5) The observation that stars seen at southern hemisphere are not the same as those seen at the northern hemisphere etc.

Let us do a summary of most type on angles that the reader mus remember:

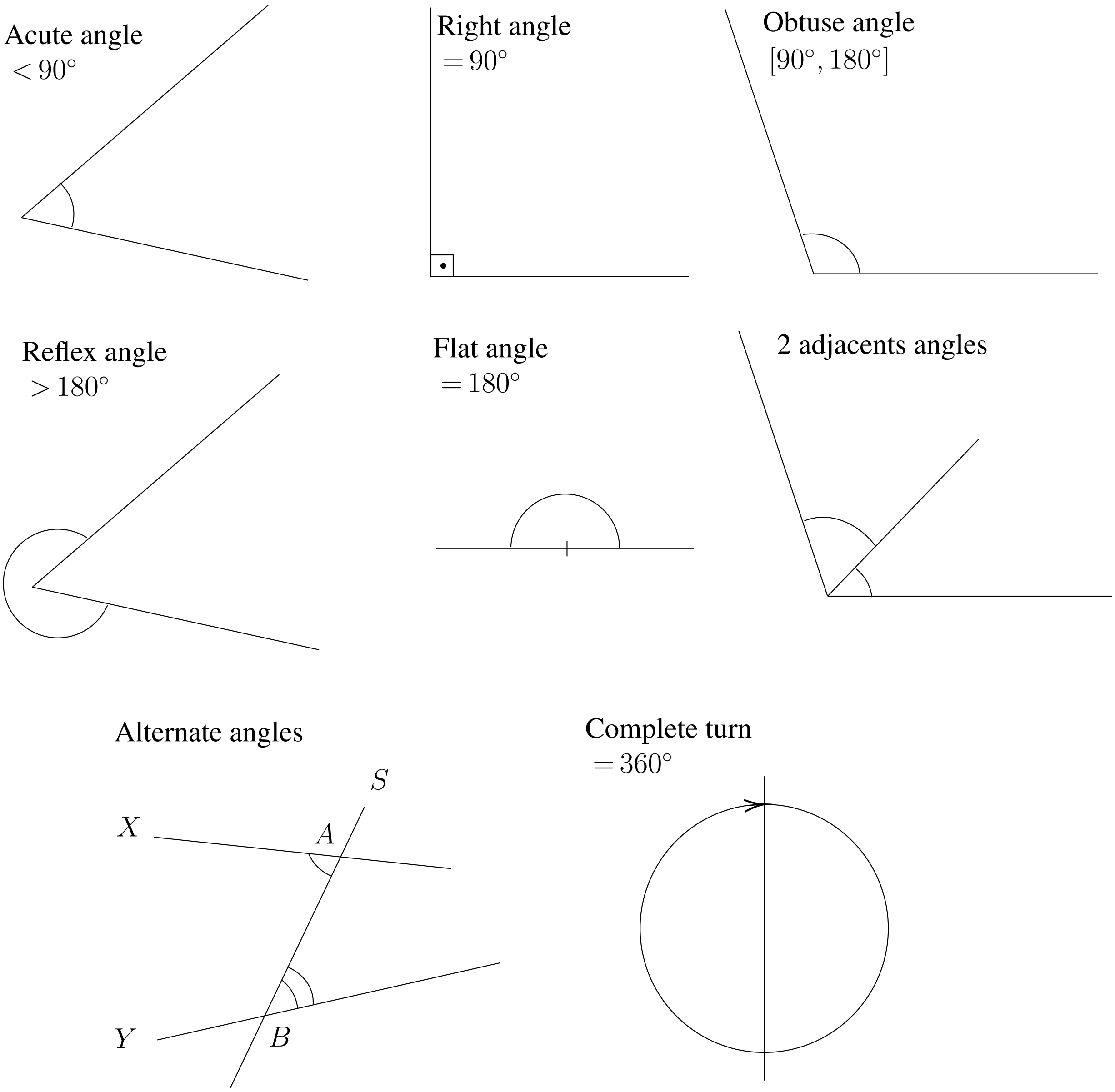

Acute angle
$< 90°$
Right angle
$= 90°$
Obtuse angle
$[90°, 180°]$
Reflex angle
$> 180°$
Flat angle
$= 180°$
2 adjacents angles
Alternate angles
S
X
A
Y
B
Complete turn
$= 360°$

Figure 8.51: **Angle types summary**

8.2.3.2.1 Angle Measurements

We have defined the equality and the sum of two or more angles. These definitions meet the conditions of the same quantities variables we have already seen previously.

So let us choose arbitrarily an angle of the plane $\widehat{aob}$, which will bet the unit angle for the plane. The measurement of the ratio:

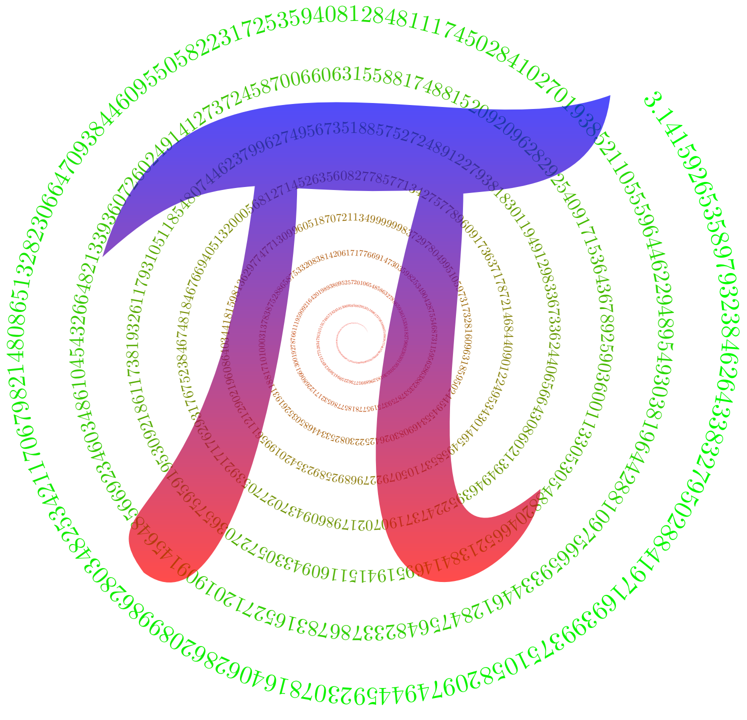
$$\frac{\widehat{AOB}}{\widehat{aob}} \tag{8.2.39}$$

proceed as it has been explained previously during our study of the quantities of the same species, will be a positive number denoted most of time by the greek letters α or θ named by definition the "angle measurement $\widehat{AOB}$, with the selected unit $\widehat{aob}$".

We denote by the lowercase Greek letter "pi" the irrational number:

$$\boxed{\pi = 3.14159265358979323846264338327950288419716939\ldots} \tag{8.2.40}$$

which is by definition the measurement value of a "flat angle" or also named "plane angle" (we have already see where this value comes from in the section of Trigonometry).



The reader can refer to the section of Numerical Methods to see another definition of π based on the circumference of the circle and based on the Archimedes algorithm (see page 6298).

Remark

As all angles of the plane are smaller than two flat angles, the number θ (which is the measurement of $\widehat{AOB}$) must be less than 2π.

Having defined the flat angle, we can now define other types of commonly used angles:

Definitions (#8.2.327):

D1. We name "right angle" any angle equal to half of a flat angle.

D2. Two angles are "perpendicular angles", denoted by the symbol $\perp$, when they are both rights and adjacent.

D3. We name "acute angle" any angle less than a right angle and "obtuse angle" any angle greater than a right angle.

D4 We name "oriented angle" or "vector angle", the angle between two vectors or lines (see section Vector Calculus page 1797) having same origin and which value measured counter-clockwise will be taken as positive and negative if taken in the clockwise direction.

The best known example of oriented angle is the trigonometric unit circle:

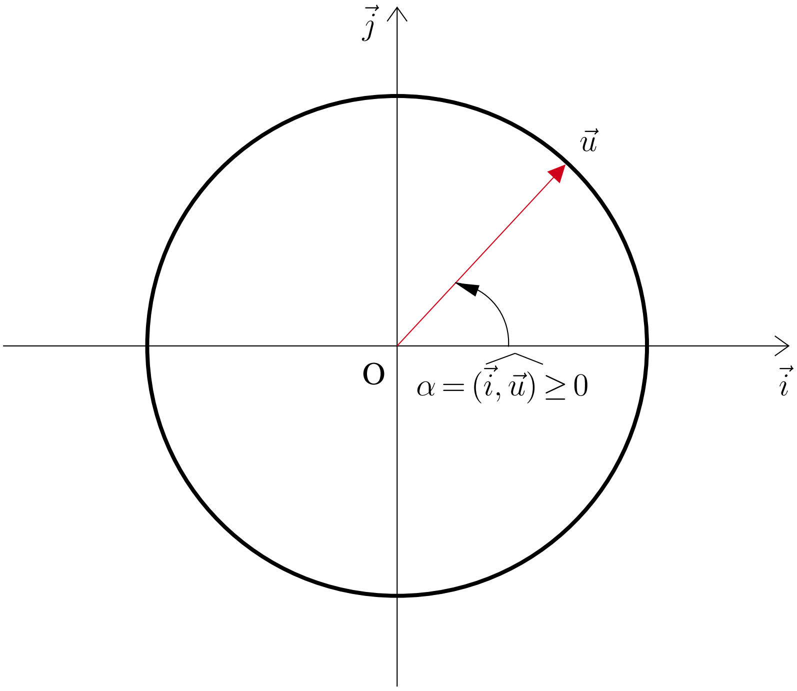


Figure 8.52: **Positive or null oriented angle**

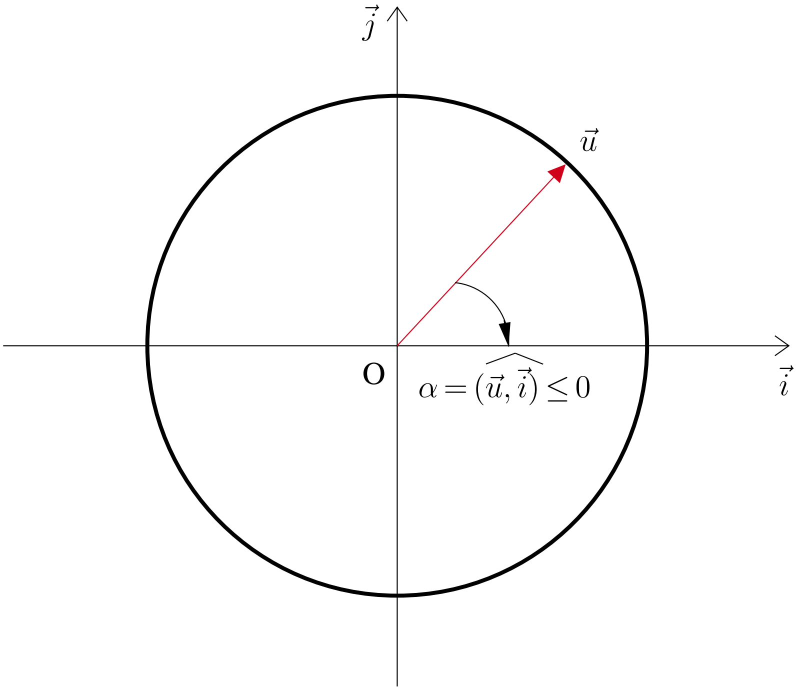

$\vec{j}$
$\vec{u}$
O
$\alpha=\widehat{(\vec{u},\vec{i})}\leq 0$
$\vec{i}$

Figure 8.53: **Negative or null oriented angle**

D5. We say that two angles are "additional angles" or "supplementary angles" when their sum is equal to two right angles (i.e. a flat angle).

D6. We say that two angles are "complementary angles" when their sum is equal to a right angle (thus two complementary angles are always right...).

The "internal angle" and "external angle" of a regular polygons are complementary angles as by construction:

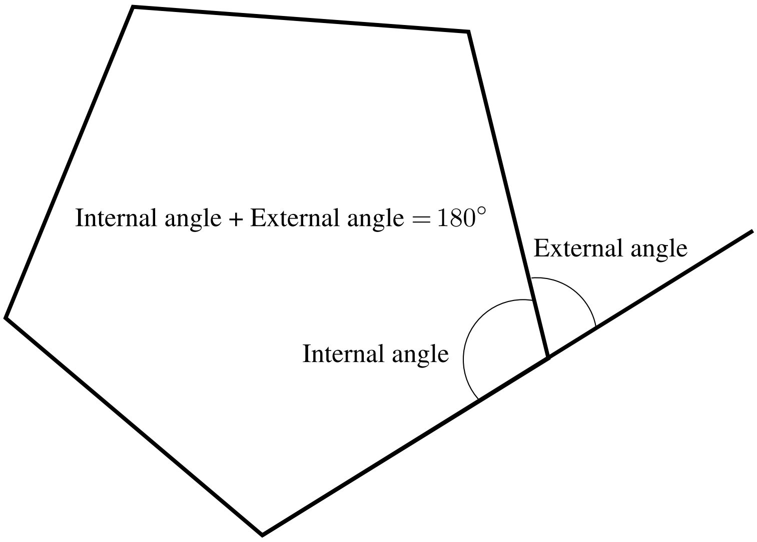

Internal angle + External angle $=180^\circ$
External angle
Internal angle

Figure 8.54: **Internal + External angle (source: Wikipedia)**

Let us now consider the symbols α,β,γ as the measurements of several angles $\widehat{AOB}$, $\widehat{BOC}$, $\widehat{COD}$.

We will not insist on the obvious fact that the equalities $\widehat{AOB} = \widehat{BOC}$ or $\alpha = \beta$ are equivalent and thus the inequalities $\widehat{AOB} < \widehat{BOC}$, or $\alpha < \beta$. These common sense remarks are applied whenever the same species of quantities will have been measured, of course with the same unit!

However, we insist that, according to the definition of the sum of several angles, $\alpha + \beta + \gamma$ is the measure of the angle $\widehat{AOD}$ increased as many times of two flat angles π that the plane was covered during the operations of addition:

$$\alpha + \beta + \gamma = \widehat{AOD} + 2n\pi \qquad (8.2.41)$$

The relative integer n which is introduce in this calculation has a value that can be specified, but that has no importance for the mathematician or physicist, as we will see later. We thus do not decide not to write the $2n\pi$ in geometry (as it will be always implicit). Also, we decide by convention to write:

$$\widehat{AOD} = \theta \qquad (8.2.42)$$

So we have:

$$\alpha + \beta + \gamma = \theta \qquad (8.2.43)$$

the notation convention means that θ is the measurement of the angle $\widehat{AOD}$ if we have $0 \leq \theta < 2\pi$.

If θ is greater than 2π, this equality means that the measurement of the angle $\widehat{AOD}$ is $\theta - 2n\pi$, n being a relative integer so that we have $0 \leq \theta - 2n\pi < 2\pi$.

Remarks

R1. If $\theta > 2\pi$ there is only one unique positive integer n and only one, as we have $0 \leq \theta - 2n\pi < 2\pi$, that is to say:

$$\frac{\theta}{2\pi} - 1 < n \leq \frac{\theta}{2\pi} \qquad (8.2.44)$$

because the two numbers $(\theta/2\pi) - 1$ and $\theta/2\pi$ are positive and different from 1.

R2. Saying that θ is the measurement of the angle $\widehat{AOD}$ suppose that we have $0 \leq \theta < 2\pi$ and leads to the equality $\widehat{AOD} = \alpha$. But write $\widehat{AOD} = \alpha$ does not necessarily mean that θ is the measurement of $\widehat{AOD}$; it is necessary for this that we have $0 \leq \theta < 2\pi$.

8.2.3.2.2 Units of Angle Measurements

We have define the flat angle to be equal to π without specifying the unit. This is what we will now apply to do. There are (still and sadly) several units of angle measurements which most important and common one are listed below:

Definitions (#8.2.328):

D1. We name "degree" the 180th part of a flat angle.

Remark

The original motivation for choosing the degree as a unit of rotations and angles is unknown.

Many old former are still made in degrees (and still today in some areas of physics/astronomy). The sub-multiple of the degree is the: "Sexagesimal minute" equal to the 60th of 1 degree, and the "Sexagesimal second" is equal to the 60th of the sexagesimal minute.

✍**Example:**

The value:

$$\theta = 236^\circ 20' 40''$$

will be read as two hundred thirty six degrees, twenty minutes and forty two seconds.

The conversion from degrees to fractional decimal angles is quite easy:

$$\begin{aligned} 236.345^\circ &= 236^\circ + 0.345^\circ = 236^\circ + 0.345^\circ \frac{60'}{1} \\ &= 236^\circ + 20.7' = 236^\circ + 20' + 0.7' \frac{60''}{1'} \\ &= 236^\circ 20' 42'' \end{aligned}$$

Most countries around the world sadly still use today degrees in Kindergarten and Highschools but without making use of the notation of minutes and seconds (not very convenient for pedagogical reasons).

D2. As we have already see it in the section of Trigonometry, we name "radian" (denoted [rad]) the plane angle described by a secant to a circle, passing through its center, such that the arc as defined by the horizontal axis through the center of the circle and the secant be of equal length of the radius of the circle:

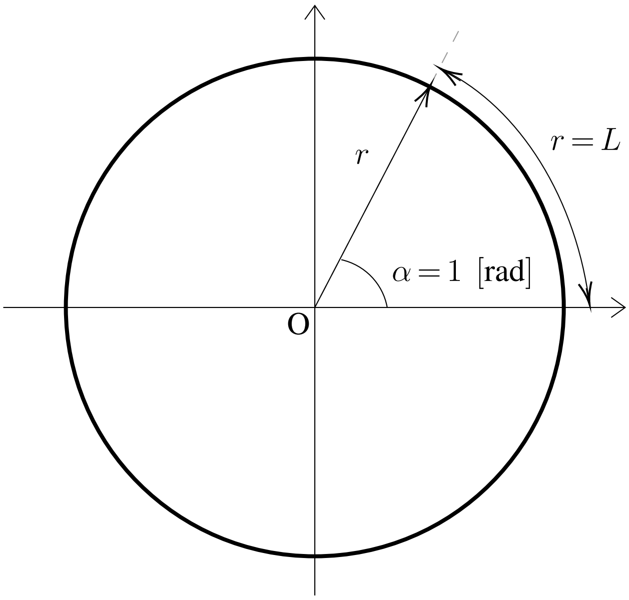


Figure 8.55: **Geometric definition of the radian**

In most mathematical work beyond practical geometry, angles are typically measured in radians rather than degrees. This is for a variety of reasons; for example, the trigonometric functions have simpler and more "natural" properties when their arguments are expressed in radians (see section Trigonometry page 2503).

One complete turn ($360°$) is equal to 2π radians, so a half turn of $180°$ is equal to π radians, or equivalently, the degree is a mathematical constant given by:

$$\boxed{1° = \frac{2\pi}{260} = \frac{\pi}{180}} \qquad (8.2.45)$$

Thus, in radians, a flat angle is equal to π and all other angles are real multiples of this constant.

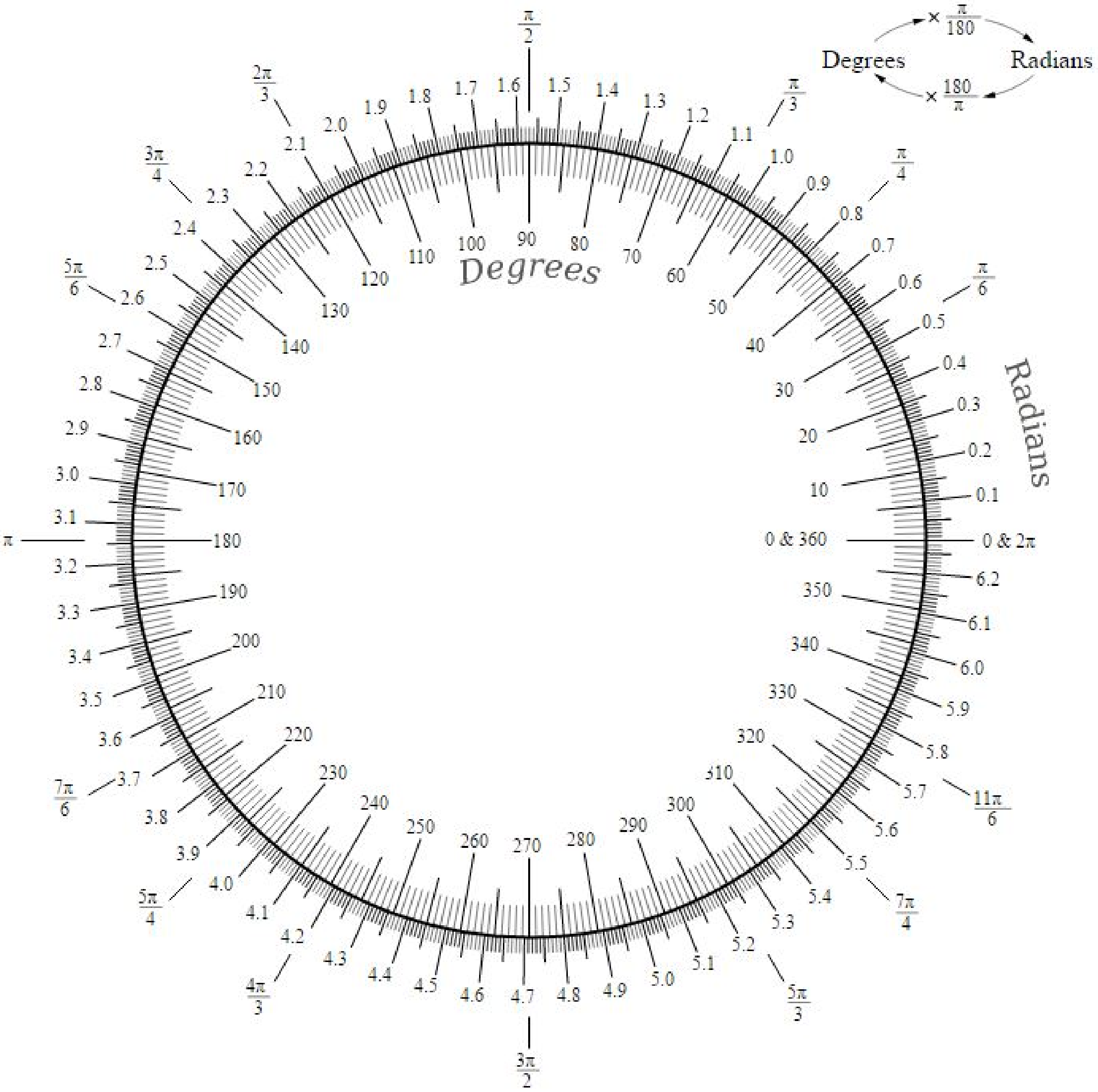

Degrees
Radians
$\times\frac{\pi}{180}$
$\times\frac{180}{\pi}$
Degrees
Radians
0 & 360
0 & 2π
$\frac{\pi}{2}$
π
$\frac{3\pi}{2}$

Figure 8.56: **Degrees-Radians conversions figure (source: Wikipedia)**

D3. We name "grad" the 200th part of the flat angle.

The grade is also an old angle unit. Its sub-multiples are: the "centesimal minute," equal to 100th of one grad, and the "centesimal second", equal to one hundredth of the centesimal minute.

The value:

$$\theta = 236.345\,[\text{gr}]$$

will be read as two hundred thirty six grads, eighteen minutes and five second.

Here are some typical angles values:

Value [degrees]	**Description**
0°	A flat surface
0.0000305°	Average angular diameter of Pluto
0.000632°	Average angular diameter of Neptune

0.00096°	Average angular diameter of Uranus
0.0023°	Average angular diameter of Mercury
0.0043°	Angular diameter of Hubble deep field
0.0047°	Average angular diameter of Saturn
0.00645°	Average angular diameter of Mars
0.012°	Average angular diameter of Jupiter
0.017°	Average resolvable diameter by the human eye
0.529°	Actual average angular diameter of the Sun
0.53°	Actual average angular diameter of the Moon
1.4°	Earth precession angle by century
3.16°	Actual angular diameter of Andromeda Galaxy
5.33°	Actual angular diameter of the small Magellanic cloud
5.14°	Moon-Earth orbital angle
10°	Actual angular diameter of large Magellanic cloud
15°	Typical airplane take-off angle
19.3°	Steepwest street in the World: Baldwin Street - Dunedin, New Zealand
20°	Width of spread out hand with arm stretched out
23.44°	Earth's obliquity
30°	Incline of many laptop stands
38°	Actual angular diameter of Gum Nebula
45°	Staircase handrail
48°	Steepest funiclar in the World: Stoosbahn - Switzerland
60°	The angle between spokes in a 6-spoke wheel
90°	A perpendicular
$\gtrsim 100°$	Magellanic stream angular diameter
125°	Human binocular vision range
360°	Milky way angular diameter

Table 8.4: **Some common angles values**

8.2.3.2.3 Bisector

Now that we know how to compare, add and measure angles we will be able to look at an important concept in geometry that is the one of "bisection" and some relative properties that we will reuse later for quite important theorems.

Definitions (#8.2.329):

D1. We name "bisector" the straight line that divides an angle into two equal parts.

D2. We name "half-bisector" the segment that divides an angle into two equal parts and start at the origin of the angle.

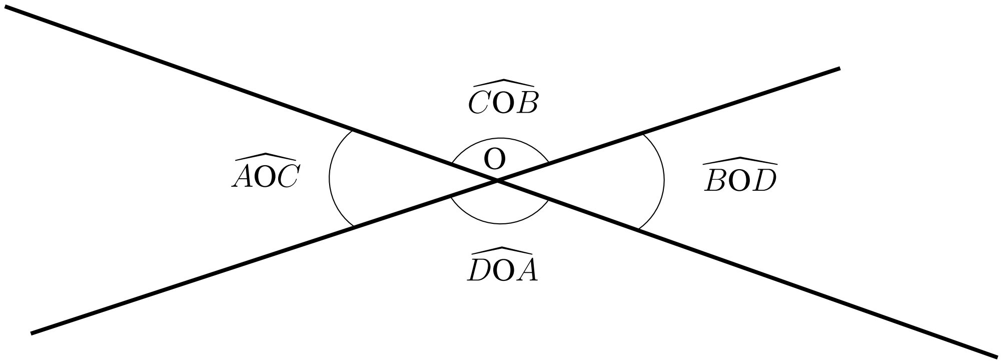
D3. We name "line segment bisector" the (infinite) straight line that passes through the mid-point of a segment.

Particularly important is the "perpendicular bisector" of a segment, which, according to its name, meets the segment at right angles.

Two straight lines AB and CD intersecting on a point O form as we already know intuitively, four angles:

$$\widehat{AOC},\widehat{COB},\widehat{BOD},\widehat{DOA} \tag{8.2.46}$$

The angles $\widehat{AOC}$, $\widehat{BOD}$, as well as the angles $\widehat{COB}$, $\widehat{DOA}$ whose sides are opposite, are known as "opposing angles by the top".



Obviously, if θ is the measurement of the angle $\widehat{AOC}$, the measurement of the adjacent angle $\widehat{COB}$ is $\pi-\theta$. That of the angle $\widehat{DOB}$ adjacent to the previous is $\pi-(\pi-\theta)$, that of $\widehat{DOA}$ is $\pi-\theta$.

Given OE the half-bisector of the angle $\widehat{AOC}$, OG that of the angle $\widehat{COB}$, OF that of the angle $\widehat{BOD}$, OH that of the angle $\widehat{DOA}$, we have:

$$\widehat{COE}=\frac{\theta}{2} \qquad \widehat{COG}=\frac{\pi-\theta}{2} \tag{8.2.47}$$

and consequently:

$$\widehat{GOE}=\frac{\pi}{2} \tag{8.2.48}$$

We would have also same:

$$\widehat{GOF}=\widehat{FOH}=\widehat{HOE}=\frac{\pi}{2} \tag{8.2.49}$$

We summarize as following all these results as well as measurement properties:

P1. Two opposite angles that share the same top are equals.

P2. Two opposite angles that share the same top have the same bisector.

P3. The bisectors of two additional adjacent angles are rectangular bisector.

P4. The bisectors of angles formed by two secants are two perpendicular lines (right angles).

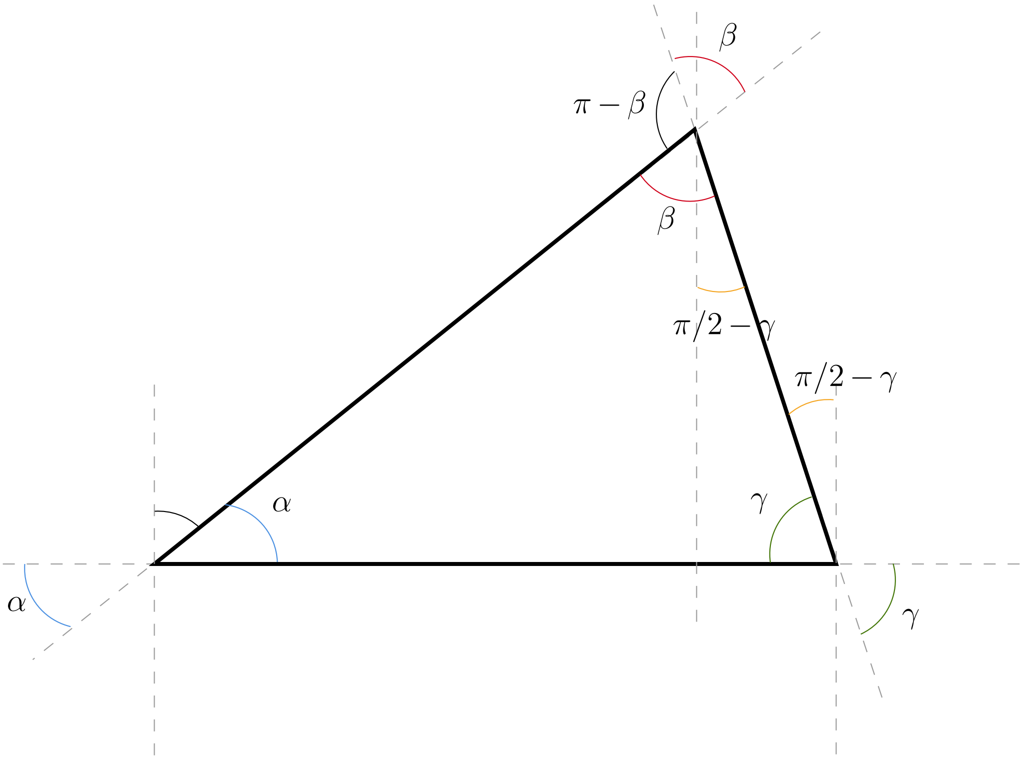


Figure 8.57: **Typical useful triangle angles in the study of Statics**

8.2.3.3 Triangles

We have study so far until now the concept of dimensions, point, straight segment, straight line, angle and of open (infinite) plan. However, a plane may be defined by several lines to thereby obtain geometric plane shapes the simplest of which may be regarded as triangles (for many other shapes see the section Geometrical Shapes at page 2886).

Definition (#8.2.330): We name "triangle" a figure (see below) formed by three segments (or lines) AB, BC, CA, the points A, B, C being not aligned. The segments AB, BC, CA, are the "sides" of the triangle. The points A, B, C are the "vertices" of the triangle. The salient angle $\widehat{BAC}$, which contains all the points of BC, is named angle $\hat{A}$ of the triangle and BC is then named the "opposite side".

Remark

We use the notation $\hat{A}$ when no confusion is possible. If necessary we use the notation $\widehat{BAC}$ with the same meaning.

There are 6 elements in a triangle, namely: three angles $\hat{A}$, $\hat{B}$, $\hat{C}$ and three finite sides $\overline{AB}$, $\overline{BC}$, $\overline{CA}$ that are most of time simply denoted by AB, BC, CA.

We will designate by $a=BC$, $b=AC$, $c=AB$ the measured lengths of the sides with the same unit of measurement, and obviously by $\hat{A}$, $\hat{B}$, $\hat{C}$ the measurements of the angles.

Theorem 8.5. The sum of the angles of a plane triangle is always equal to π radians (or 180°). The proof is quite simple and is named the "angle sum theorem".

Proof 8.5.1. In the figure below, ABC is any triangle, and D is the parallel to $\overline{BC}$ passing through A. We observe:

1. The blue angles have the same amplitude (value) because they are alternate angles (the segment $\overline{BC}$ and the line D being parallels).
2. Also, the green angles have the same amplitude (value) because they are alternate angles[2].
3. We notice that the sum of the angles blue + red + green are equal to a flat angle on A.

From the equalities seen in (1) and (2) previously, we deduce that:

$$\vec{A}+\vec{B}+\vec{C}=\pi=180^\circ \tag{8.2.50}$$

---

[2] If two lines are parallel then their alternate interior angles are equal. If the alternate interior angles of two lines are equal then the lines must be parallel. This is for reminder the "Z theorem".

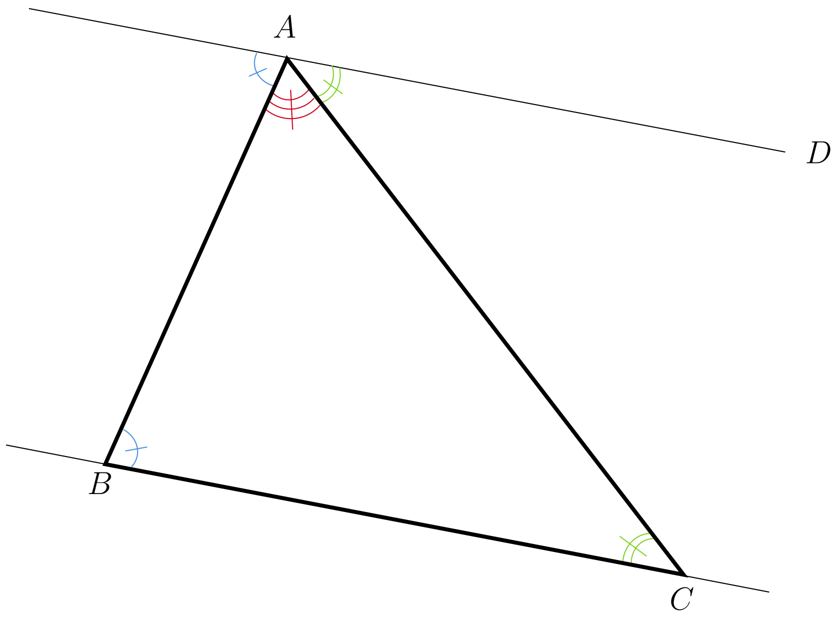


Figure 8.58: **Sum of angles of a plane triangle**

This proof is valid regardless the type of triangle drawn in the plane.

■ Q.E.D.

As for all other shapes that we will see in this section, the calculations of the perimeter and area (surface) are given in the section of Geometric Shapes page 2888.

8.2.3.3.1 Equal Triangles (congruent triangles)

Definition (#8.2.331): We say that two triangles are "equal triangles" when we can by movement or by a reversal or both combined, superimpose all vertices of the first triangle with that of the second. Then we say also that the triangles are "homologous triangles" or "congruent triangles".

If a triangle ABC is congruent to a triangle DEF, the relation can be written mathematically as:

$$\boxed{\triangle \text{ABC} \cong \triangle \text{DEF}} \tag{8.2.51}$$

From this definition, it comes that two triangles are congruent when either the following two properties are satisfied:

P1. They have one equal side and two equal angles.

Proof 8.5.2. So we state that two triangles have equal side $\overline{BC} = \overline{B'C'}$ between two equal angles $\hat{B} = \hat{B}'$,$\hat{C} = \hat{C}'$ are equal.

In other words, if 2 triangles have 2 equal angles 2 by 2, then the 3rd angles are equal too. That said we can now take as equal side the one that is located between the 2 equal angles without loss of generality.

Since $\overline{BC}=\overline{B'C'}$ (see figure below), there is a movement which brings B' to B and C' in C. This movement brings A' on A_1 located on the same side relatively to the segment $\overline{BC}$, or on A_2 symmetric of A_1 relatively to this same segment.

The two segments $\overline{BA}$ and $\overline{BA_1}$ do by hypothesis, with $\overline{BC}$ the same angle $\hat{B}=\hat{B}'$. As they are by construction on the same side of $\overline{BC}$, they are merged. The both segments $\overline{CA}$ and $\overline{CA_1}$ are merged for the same reason and because $\hat{C}=\hat{C}'$. A_1 is then merged with A. The two triangles ABC, $A'B'C'$ are then indeed congruent.

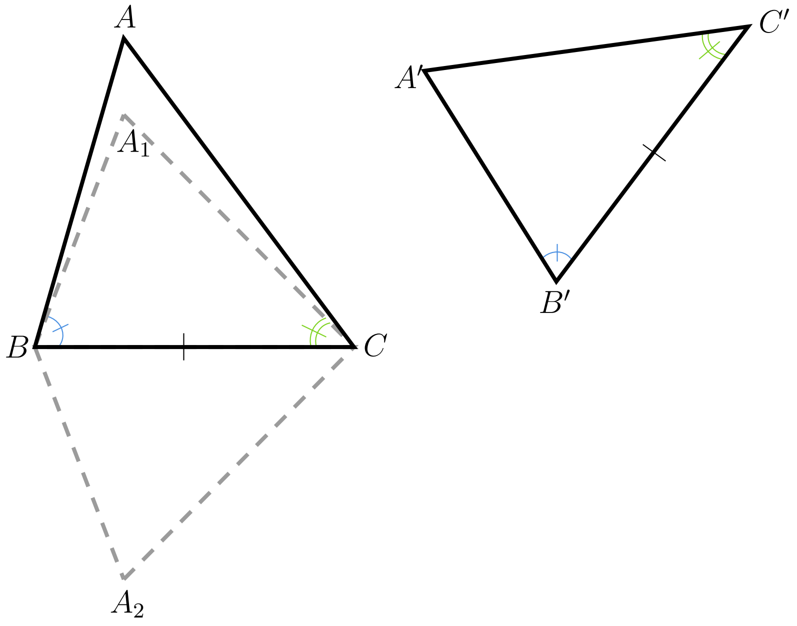


Figure 8.59: **Two triangles having equal side and two equal angles**

■ Q.E.D.

This result is named sometimes the "angle-side-angle theorem".

P2. They have a same angle between two sides of equal length.

Proof 8.5.3. So we state that two triangles with an equal angle $\hat{A}=\hat{A}'$ between two equal sides $\overline{AB}=\overline{A'B'}$ and $\overline{AC}=\overline{A'C'}$ are equal.

Since $AB=A'B'$ (see figure below) there exist a shift located relatively to $\overline{AB}$ at the same side as the point. If it brought it on C_2 symmetric of C_1 with respect to $\overline{AB}$, a half turn around the $\overline{AB}$ would take it on C_1. The segments $\overline{AC}$ and $\overline{AC_2}$ located on the same side of $\overline{AB}$ do, by hypothesis the same angle with $\overline{AB}$, since $\hat{A}=\hat{A}'$. They are then merged. The hypothesis $\overline{AC}=\overline{AC_1}$ then makes us conclude that C_1 and C coincide. The two triangles ABC, $A'B'C'$ are then congruent.

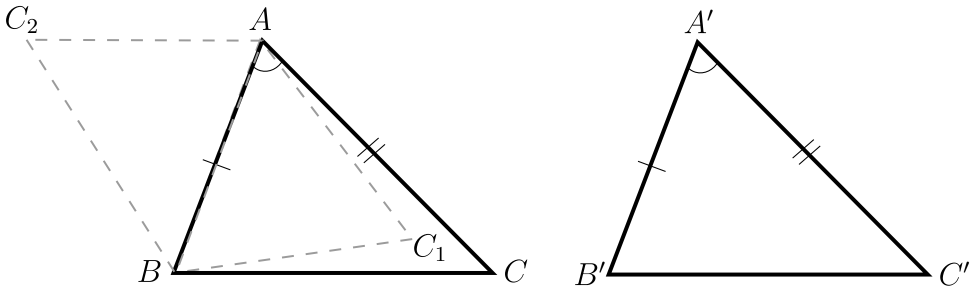


Figure 8.60: **Displacement showing the equality of the 2 triangles**

■ Q.E.D.

Corollary 8.5.1. As corollary we have that two triangles are congruent if their three sides are equal. This is named sometimes the "SSS (side-side-side) theorem".

Now about drawing symbols, it is common to represent various equal angles or equal sides in triangle as following :

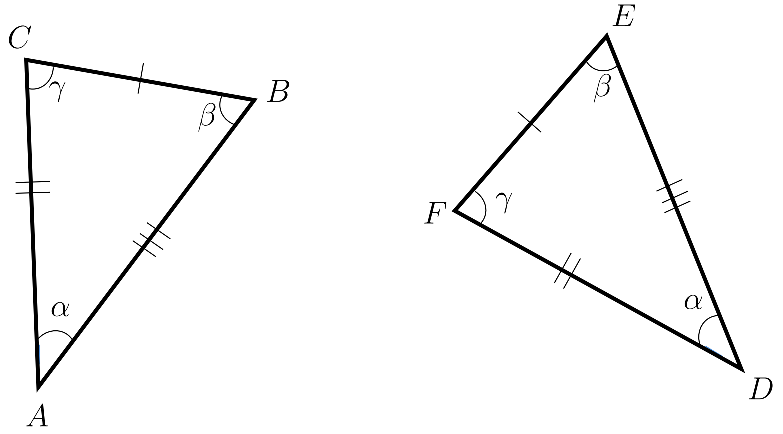


Figure 8.61: **First typical notation for congruent triangles**

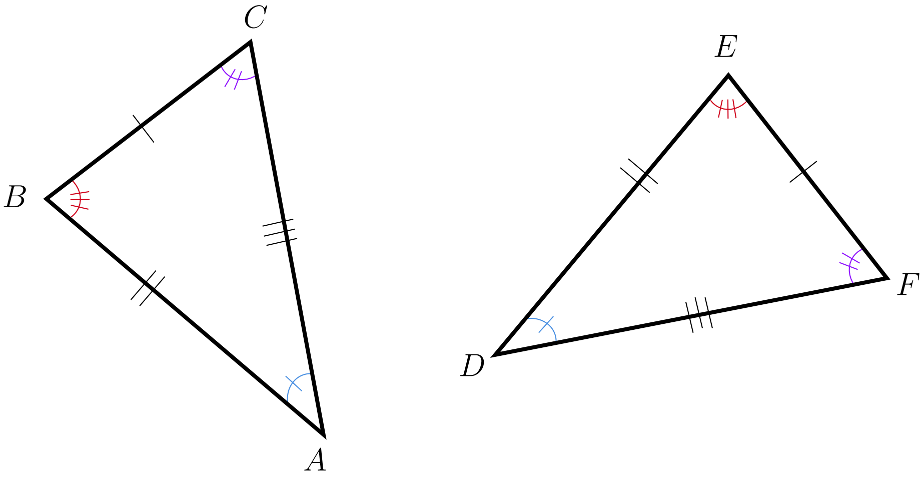


Figure 8.62: **Second typical notation for congruent triangles**

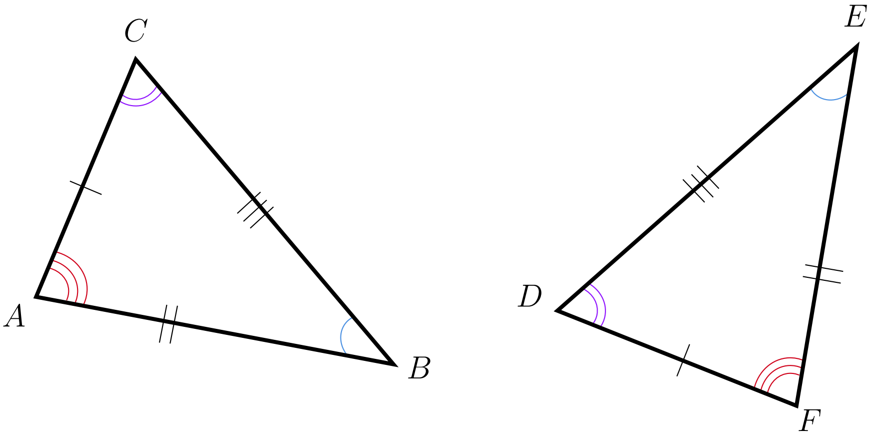


Figure 8.63: **Third typical notation for congruent triangles**

8.2.3.3.2 Isosceles Triangles

Definition (#8.2.332): We say that ABC is an "isosceles triangle" when two of its sides AB and AC are equal ("iso" meaning "same"). The third side BC is then named the "base" of the triangle.

Remark

We say that a triangle is "scalene" when it has three unequal sides.

Definition (#8.2.333): We name "mediator" of a segment $\overline{BC}$, the perpendicular to $\overline{BC}$ at point H of this segment, middle of $\overline{BC}$.

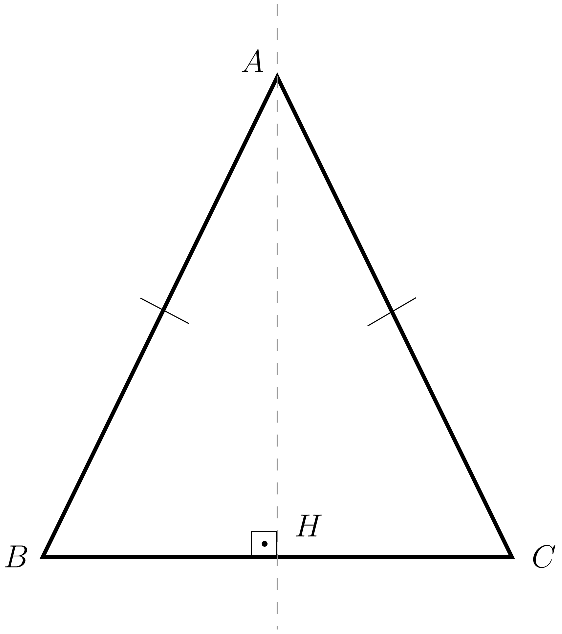


Figure 8.64: Representation of the mediator in isosceles triangle

Theorem 8.6. In an isosceles triangle ABC as shown above, the angles $\hat{B}$ and $\hat{C}$ opposed to the equal sides are equal.

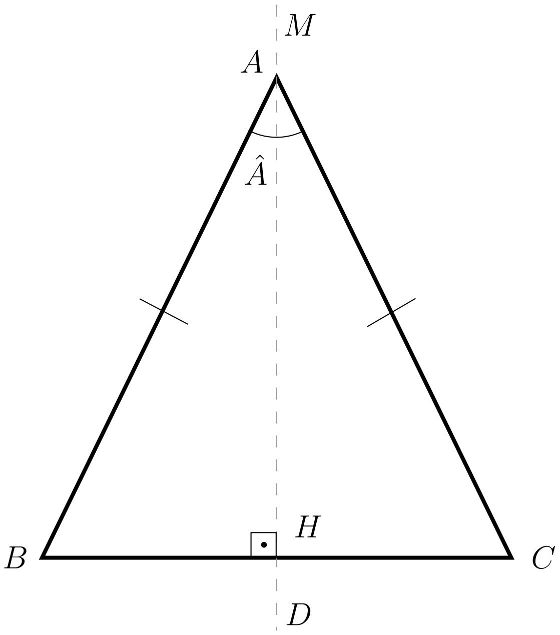
Proof 8.6.1. The both (right) triangles BAH and CAH defined by the bisector of $\widehat{A}$ have an equal angle $\widehat{BAH}=\widehat{CAH}$ by construction between two equal sides: $\overline{AH}$ that is the shared (common) side and $\overline{AB}=\overline{AC}$ by hypothesis. As the angles $\widehat{BHA}=\widehat{CHA}$ and rights and equals and that the sum of angles of a triangle is equal to a straight angle (180° or π radians), then the angles $\widehat{HBA}$ and $\widehat{ACH}$ are equal.

■ Q.E.D.

Remark

This latest proof is one to which refers the comic Logicomix on page 55 and it is this type of proof by logical induction which would have brought the young Bertrand Russell to became interested in logic to later become one of the most famous logicians in modern history.

Theorem 8.7. In an isosceles triangle ABC as shown above the mediator of $\overline{BC}$ and the bisector of the angle $\hat{A}$ coincide (figure below):



In other words, the locus of equidistant points M (at same distance) between of two given points B and C is given by the bisector (D) of segment $\overline{BC}$.

Remark

In geometry, a "locus" is a set of points (commonly, a line, a line segment, a curve or a surface), whose location satisfies or is determined by one or more specified conditions. More formally we name "locus" of a point M, subject to some constraints (conditions), all positions occupied by this point M that still satisfy the constraints (conditions).

Proof 8.7.1. We will prove first that any point of the locus is on the line (D) and afterwards that any point of (D) is a point of the locus.

1. Any point of the locus is on the line (D):

   In other words, the assumption, in the figure above, that $\overline{MB}=\overline{MC}$ implies that M is the mediator of $\overline{BC}$. Indeed, if $\overline{MB}=\overline{MC}$, the triangle MBC is isosceles and the top M is on the mediator of $\overline{BC}$ (that coincides as we know with the bisector).

2. Any point of (D) is a point of the locus:

   This means that if M is on the mediator of $\overline{BC}$, we have $\overline{MB}=\overline{MC}$. Indeed, if M is on the line (D) that meets in H, midpoint of $\overline{BC}$, the line BC, the triangles MHB, MHC are then equal (second case of equality: $\widehat{BHM}=\widehat{CHM}$ because these angles are straight, $\overline{HM}$ common; $\overline{HB}=\overline{HC}$ because H is the midpoint of $\overline{BC}$): the sides $\overline{MB}$, $\overline{MC}$ are then also equal and we have indeed $\overline{MB}=\overline{MC}$. Therefore point M is indeed a point of the locus.

■ Q.E.D.

Using this theorem we can state another one:

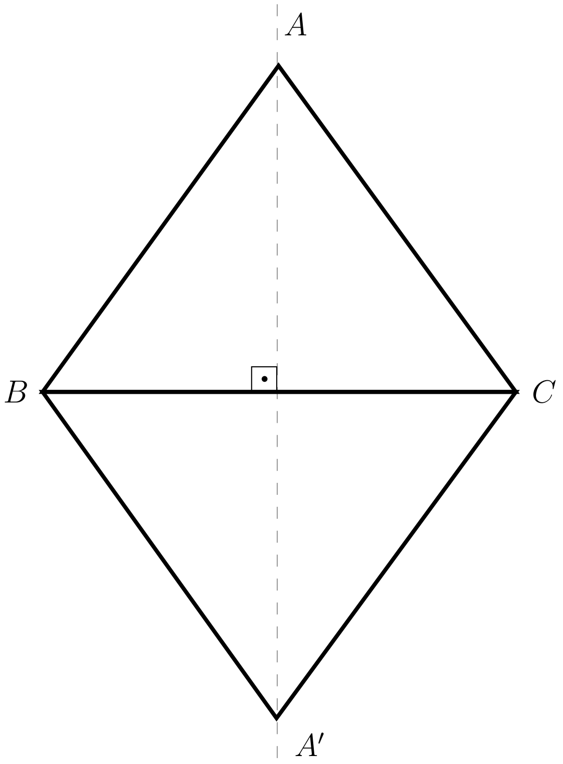
Theorem 8.8. By taking a point A outside a line $\overline{BC}$, we can draw to this line a single perpendicular (or in other words: to a given point outside of a given line we can draw one and only one perpendicular line).

This is an equivalent to the Euclid's fifth postulate and a way to build an isoceles triangle as the apex is on the perpendicular of the base.

Proof 8.8.1. The proof is done in two steps:

1. There exist a perpendicular line:

   Indeed, given a triangle ABC, let us subject this triangle to half a turn around BC (horizontal symmetry) as in the figure below:

   The apex A comes in A' symmetrical by definition, relative to $\overline{BC}$. Since the figures, ABC and $A'BC$ are equal, $\overline{AB}=\overline{A'B}$ and $\overline{AC}=\overline{A'C}$.

   $\overline{BC}$ is then the mediator of $\overline{AA'}$ and the lines $\overline{BC}$ and $\overline{AA'}$ are perpendicular. $\overline{AA'}$ is therefore well perpendicular to $\overline{BC}$ passing through A.

2. The perpendicular line is unique:

   Given $\overline{AH}$ a perpendicular drawn from A to $\overline{BC}$, it meets the line $\overline{BC}$ on a point H that is different from either B or C.

   Let us Suppose that H is different from B. The angles $\widehat{AHB}$, $\widehat{A'HB}$ which are deduced from each other by rotation, are equal, and as each of them is a right angle, the angle $\widehat{AHA'}$ is a flat angle. The line $\overline{AH}$ thus coincides with the straight line $\overline{AA'}$ and is therefore the only perpendicular to $\overline{BC}$ which passes through the apex A.

■ Q.E.D.

Definitions (#8.2.334):

D1. We name "orthogonal projection" from point A on a line $\overline{BC}$ the point H point where the perpendicular drawn from A at this line meet H. The point H is also named the "foot" of the perpendicular.

We have already see the vector version of this projection in the section of Vector Calculus and that it was denoted by $\text{proj}_{\overline{BC}A}$.

D2. We name "geometrical distance" from A to $\overline{BC}$ the length of the segment $\overline{AH}$.

Since in the previous figure $\overline{BC}$ is the mediator of $\overline{AA'}$, H is the midpoint of $\overline{AA'}$. So: A point A and its symmetrical point A' are with respect to a straight line (D) are characterized by the following two properties that we will not prove as we consider them sufficiently intuitive:

P1. $\overline{AA'}$ is perpendicular to (D)

P2. The middle of $\overline{AA'}$ is on (D)

The line $\overline{AB}$, which connects the apex A to a point of the straight line $\overline{BC}$, other than the foot H of the perpendicular drawn from A to this line, is named an "oblique line". The point B being named the "oblique foot".

8.2.3.3.3 Equilateral Triangles

Definition (#8.2.335): A triangle ABC is say to be an "equilateral triangle" when all its three sides are equal. In the familiar Euclidean geometry, equilateral triangles are also equiangular; that is, all three internal angles $\hat{A},\hat{B},\hat{C}$ are also congruent to each other and are each equal to $\pi/3$ (60°). They are regular polygons (see definition further below), and can therefore also be referred to as "regular triangles":

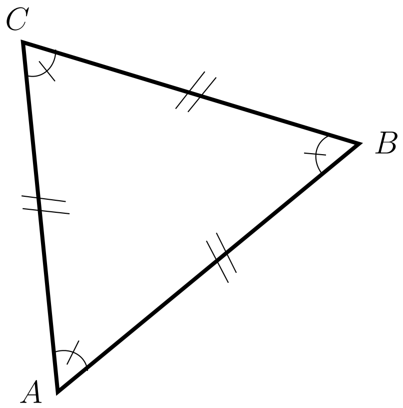


Figure 8.65: **Equilateral triangle**

As for all other shapes that we will see in this section the calculation of the perimeter and area (surface) are given in the section of Geometric Shapes at the page 2886.

8.2.3.3.4 Right Triangle

Definition (#8.2.336): A "right triangle" (American English) or "right-angled triangle" (British English) is a triangle in which one angle is a right angle (that is, a 90° angle or $\pi/2$ [rad]). The relation between the sides and angles of a right triangle is the basis for trigonometry!

The side opposite the right angle is named the "hypotenuse" (side c on the figure). The sides adjacent to the right angle are named "legs". Side a may be identified as "the side adjacent to angle $\hat{B}$" and "opposed to (or opposite) angle $\hat{C}$", while side b is the side adjacent to angle $\hat{C}$ and opposed to angle $\hat{B}$.

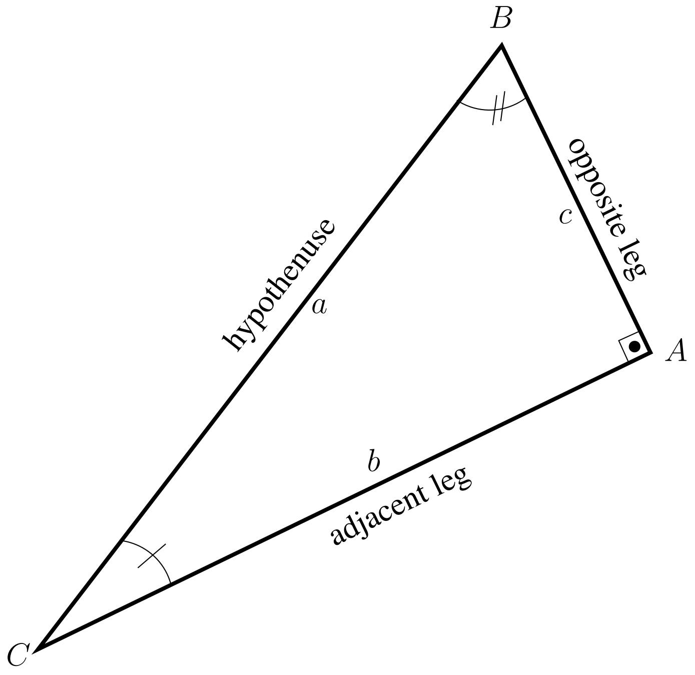

B
hypothenuse
a
opposite leg
c
A
b
adjacent leg
C

Figure 8.66: **Right triangle**

To say that the triangle is right at A means it is in $\hat{A}$ that is the right angle. Another quite common representation of a right triangle in college level textbook is:

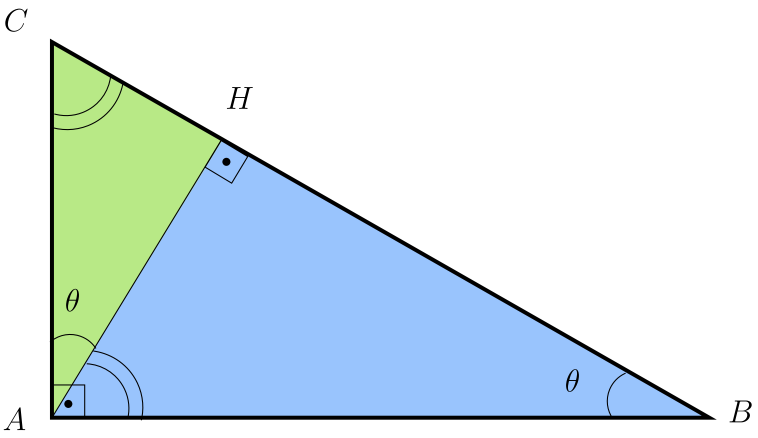

C
H
θ
θ
A
B

Figure 8.67: **Another well known representation of right triangle**

Remark

In a right triangle, the longest side is always the side opposite the right angle. We will prove this property with the Pythagorean theorem (see further below).

8.2.3.3.5 Right Isosceles Triangle

Definition (#8.2.337): A "right isosceles triangle" ABC is both right and isosceles, meaning that it has both a right angle and two sides of equal length.

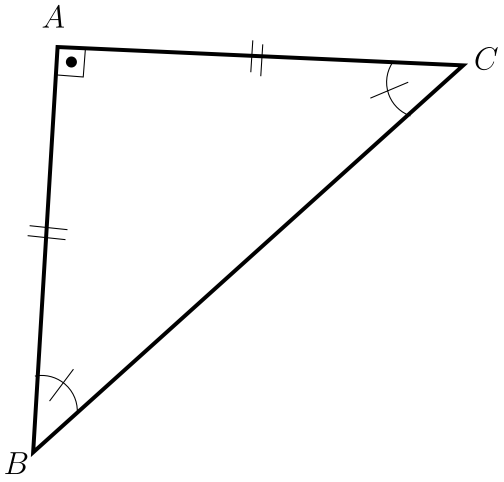


Figure 8.68: **Right isosceles triangle**

The main vertex (apex) corresponds to the right angle (A). Indeed, as $\overline{BC}$, the hypotenuse must be the greatest side, the sides $\overline{AB}$ and $\overline{AC}$ have the same length (smaller)!

So far we have for summary:

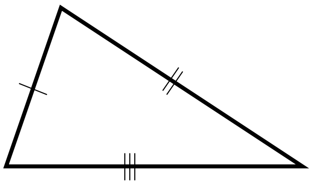
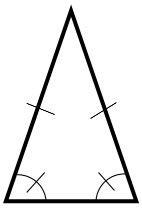
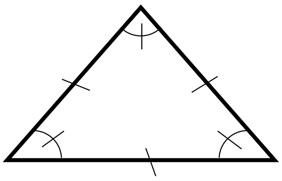
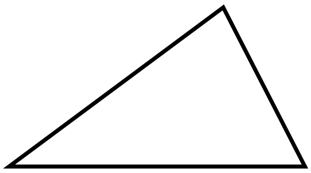
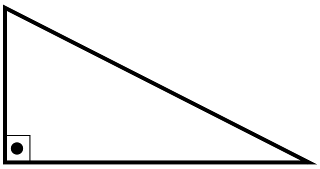
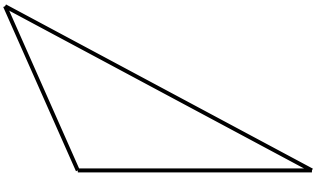
Based on sides	
Scalene triangle All three sides have different lengths. Its angles are also all different.	
Isosceles triangle Two sides have equal lengths. Two of its angles are also equals	
Equilateral (equiangular) triangle All three sides are of same length. All three angles are equal (60°)	
Based on angles	
Acute triangle All angles are less than (90°)	
Right triangle Has one right angle (90°)	
Obtuse triangle Has one angle more than 90°	

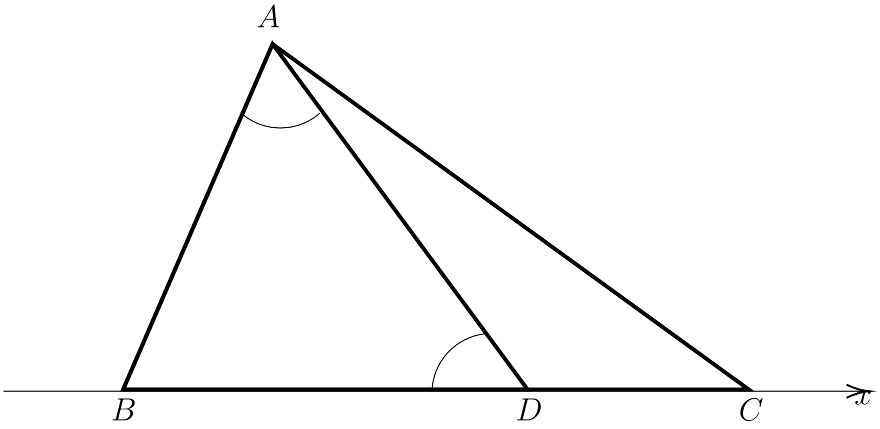
Figure 8.69: **Classification of plane triangles**

8.2.3.3.6 Inequalities in the triangles

Let us now see some interesting inequalities (properties) in the triangle.

P1. First let us prove that in any triangle, a side opposite to a right or obtuse angle (greater than 90° or $\pi/2$ [rad]) is greater than each of the other two sides of the triangle.

Proof 8.8.2. Consider the triangle ABC below where $\hat{A}\geq\pi/2$ and $\overline{Cx}$ the half-line extension of $\overline{BC}$. Let us draw, on the half-line Bx, a length $\overline{BD}=\overline{BA}$ to construct an isosceles triangle in the original triangle:



Therefore the triangle BAD is an isosceles triangle whose angle at the base AD is obviously acute (less than $\pi/2$ or less than 90°).

So by construction $\widehat{BAD}<\widehat{BAC}$. The straight line $\overline{AD}$ is interior by construction the angle $\widehat{BAC}$ and then it follows:

$$\overline{BC}>\overline{BD} \tag{8.2.52}$$

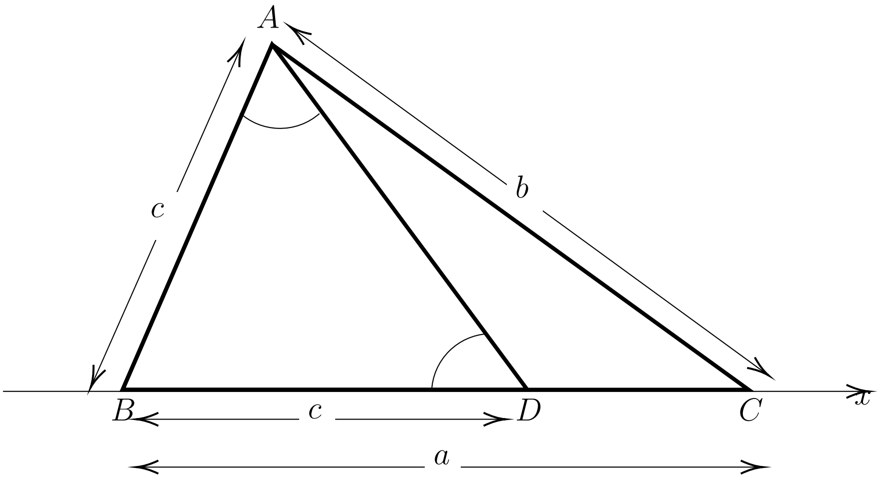
and as $\overline{BD}=\overline{BA}$ by construction we have:

$$\overline{BC}>\overline{BA} \tag{8.2.53}$$

Which finites our proof. As the reasoning is the same to prove that $\overline{BC}>\overline{CA}$.

■ Q.E.D.

P2. In any triangle whose sides have strictly increasing lengths, one side is always less than the sum of the other two.



Proof 8.8.3. Let D be the point of the side $\overline{BC}$ such that $\overline{BD}=\overline{BA}=c$ are the sides of the isosceles triangle ABD. We get:

$$\overline{DC}=\overline{BC}-\overline{BD}=a-c \tag{8.2.54}$$

The triangle ABD being isosceles, the angle at the base $\widehat{ADB}$ is acute and its complement $\widehat{ADC}$ is obtuse. In the ADC triangle, we get from the preceding equation P1 property $\overline{DC}<\overline{CA}$, that is to say:

$$a-c<b \tag{8.2.55}$$

where:

$$a<b+c \tag{8.2.56}$$

this is the famous "triangle inequality" in it geometrical version. We will see it again in many other sections of this textbook in spaces and more abstract mathematical concepts.

The property is the immediate for other sides b and c by permutation of the method:

$$\begin{gathered}a<b+c\\ b<a+c\\ c<a+b\end{gathered} \tag{8.2.57}$$

■ Q.E.D.

P3. In any triangle any side is greater than the difference of the other two.

Proof 8.8.4. Suppose we have $a>b>c$. Therefore we also have:

$$a+c>b \tag{8.2.58}$$

Subtracting c to both sides of the inequality gives:

$$a>b-c \tag{8.2.59}$$

The property is the immediate for the other sides b and c by permutation of the method:

$$a>b-c \tag{8.2.60}$$

$$b > a - c \tag{8.2.61}$$
$$c > a - b \tag{8.2.62}$$

And finally as:

$$a > b - c \quad \text{and} \quad a < b + c \tag{8.2.63}$$

for any triangle with increasing sides we have:

$$\boxed{b - c < a < b + c} \tag{8.2.64}$$

■ Q.E.D.

8.2.3.3.7 Triangles remarkable interior lines

A little summary about triangles remarkable interior lines is probably necessary at this point (before continuing on other triangles theorems) to avoid if possible to much confusion as we have given a lot of definitions so far:

- The "perpendicular bisector" of a segment is the line which is perpendicular to the segment and which passes through its center.
- The "height" or "altitude" of a triangle is a line which passes through a vertex of the triangle and which intercept the opposite edge with right angle.
- The "angle bisector" of an angle is the semi-straight line that divides the angle into two equal angles.
- The "median" of triangle passes through a vertex and intercepts the opposite edge at its center.
- The "mediator" is the perpendicular line to the crossing point of the median.

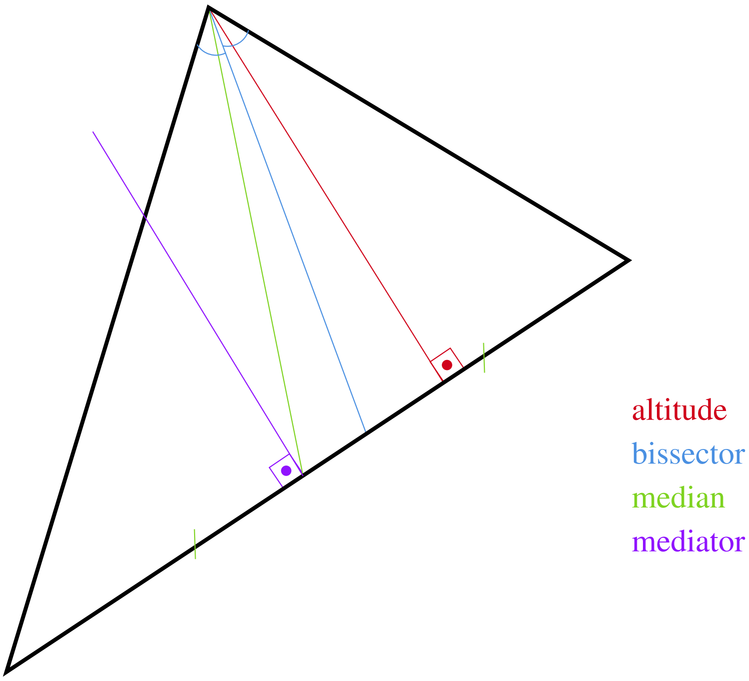


Figure 8.70: **Bisectors, Median, Mediator and Height in a triangle**

Obviously in some geometries all a part of all the lines above are not discernible.

In a triangle:

- The intersection of 3 altitudes is named the "orthocenter".
- The intersection of 3 medians is named the "centroid".
- The intersection of 3 angle bisectors is the "center of the inscribed circle".
- The intersection of 3 perpendicular bisectors is the center of the circumscribed circle.

8.2.3.3.8 Pythagorean theorem

Now that we have seen what was a triangle and reviewed some of its properties as well as the concept angle, we can prove the famous "Pythagorean theorem" that gives the relation that must satisfy three numbers that represent the sides of a right triangle in Euclidean Geometry and also permits to make circle trigonometry (see section Trigonometry page 2508).

Theorem 8.9. The square of the hypotenuse (the side opposite the right angle) is equal to the sum of the squares of the other two sides.

The theorem has been given numerous proofs - possibly the most for any mathematical theorem. They are very diverse, including both geometric proofs and algebraic proofs, with some dating back thousands of years. The theorem can be generalized in various ways, including higher-dimensional spaces, to spaces that are not Euclidean, to objects that are not right triangles, and indeed, to objects that are not triangles at all, but n-dimensional solids.

Among all possible proof we chose to present the next one as it seemed to most easy one to us for readers (student mainly in this case):

Proof 8.9.1. Given a square (four right angles polygon) inside which is inscribed another square of smaller sides. We can calculate the surface of the inscribed square thanks to the right triangles resulting from the empty spaces between the two squares such as presented below:

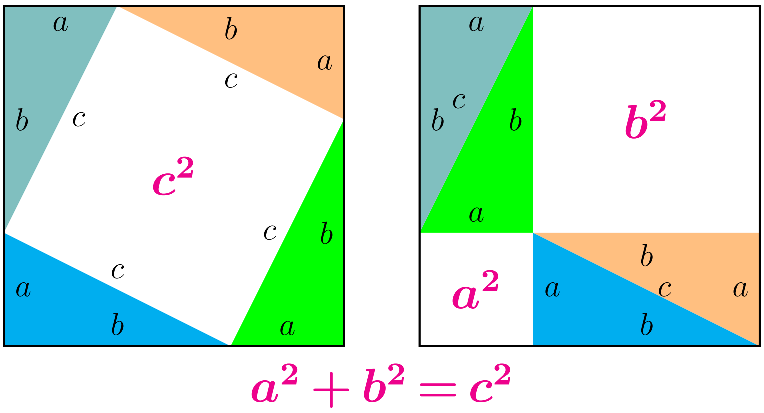


Figure 8.71: Pythagorean theorem illustrated

The surface of the whole square is obviously:

$$S=(a+b)^2=a^2+2ab+b^2 \tag{8.2.65}$$

and that of the white square:

$$S=c^2 \tag{8.2.66}$$

To get the surface of the white square we can subtract from the whole square the surface of the four right triangles having each for surface (see section Geometric Shapes page 2896):

$$S_\Delta=\frac{ab}{2} \tag{8.2.67}$$

The surface of the white square is finally:

$$(a+b)^2=a^2+2ab+b^2-4\frac{ab}{2}=a^2+b^2=c^2 \tag{8.2.68}$$

We then get the final result of the famous "Pythagorean theorem":

$$\boxed{a^2+b^2=c^2} \tag{8.2.69}$$

In a right triangle, the square of the hypotenuse (side opposite the right angle) is equal to the sum of the squares of sides of the right angle[3].

■ Q.E.D.

In the particular case we have three integers a, b and c that satisfy the Pythagorean theorem (there are infinite combinations of integers satisfying the Pythagorean theorem), then we speak of "Pythagorean triple".

Remark

We own this possible proof to the Chinese Chao King in the 102th century (holocene calendar).

It is often mentioned in smaller class of the reciprocal of Pythagorean theorem which states: In a triangle, if the square of one side is equal to the sum of the squares of the other two sides, then the triangle is rectangle and the hypotenuse will be the longest side of the triangle.

8.2.3.3.9 Thales' Theorem (intercept theorem)

The "intercept theorem", also known as "Thales' theorem" (not to be confused with another theorem with that name), is an important theorem in elementary geometry about the ratios of various line segments that are created if two intersecting lines are intercepted by a pair of parallels.

Having just proved the Pythagorean theorem and that now the concepts of parallels, segments, angles and others are quite well known to us, we can finally prove the Thales' theorem which is given here a possible proof that first requires the development of two lemmas:

Lemma 8.9.1. Triangles with same surfaces: If two triangles have a common side and if the third vertices are on a parallel to this common side, then they have the same surface!

Proof 8.9.2. Given the figure:

---

[3] Modern scholars seems to agree that Pythagoras of Samos never existed for real (as Homere, Adam, Eve, Moise, Noah, Jesus, etc.), that's why you won't found an entry about him in the section of Biographies. The Pythagorean theorem was also known and used by the Babylonians and Indians centuries before Pythagoras, but he may have been the first to introduce it to the Greeks

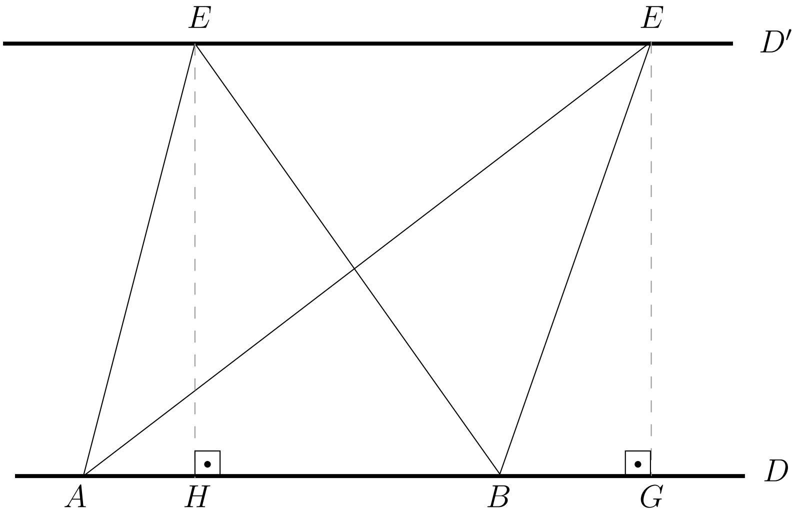


Figure 8.72: **First construction for Thales' theorem proof**

We have:

$$D//D' \quad (8.2.70)$$

$$\overline{EH} \perp D \quad (8.2.71)$$

$$\overline{FG} \perp D \quad (8.2.72)$$

$EFGH$ is a rectangle because its sides are parallel in pairs and have at least two right angles. So its opposite sides have the same length: $\overline{EH} = \overline{FG}$.

$\overline{EH}$ is the relative height to $\overline{AB}$ in the triangle EAB and $\overline{FG}$ is the relative height $\overline{AB}$ in the triangle FAB.

The surface of the triangle depends only on the length of the side and the length of the height relative this side (see section Geometric Shapes page 2896). For the both triangles EAB and FAB, these lengths are equal, so they have the same surface!

So we have indeed that if two triangles have a common side and if the third vertices are on a parallel to this common side, then they have the same surface!

■ Q.E.D.

Lemma 8.9.2. We have:

$$\frac{a+c}{b+d} = \frac{c}{d} \quad (8.2.73)$$

Proof 8.9.3. Given the proportions ratio ("proportional calculus" or "product in cross"):

$$\frac{a+c}{b+d} = \frac{c}{d} \quad (8.2.74)$$

then:

$$ad = bc \quad (8.2.75)$$

If $ad=bc$, then:

$$ad+cd=bc+cd \tag{8.2.76}$$

where we added a same positive or negative number to the two members of the equality. Thus after factorization:

$$d(a+c)=c(b+d) \tag{8.2.77}$$

and by applying contreversly the product in cross:

$$\boxed{\frac{a+c}{b+d}=\frac{c}{d}} \tag{8.2.78}$$

■ Q.E.D.

Let us now expose what is the Thales theorem.

Theorem 8.10. If two intersecting lines are cut by parallel lines, the line segments cut by the parallel lines from one of the lines are proportional to the corresponding line segments cut by them from the other line and from this we can also deduce also other remarkable identities.

Proof 8.10.1. Given the figure:

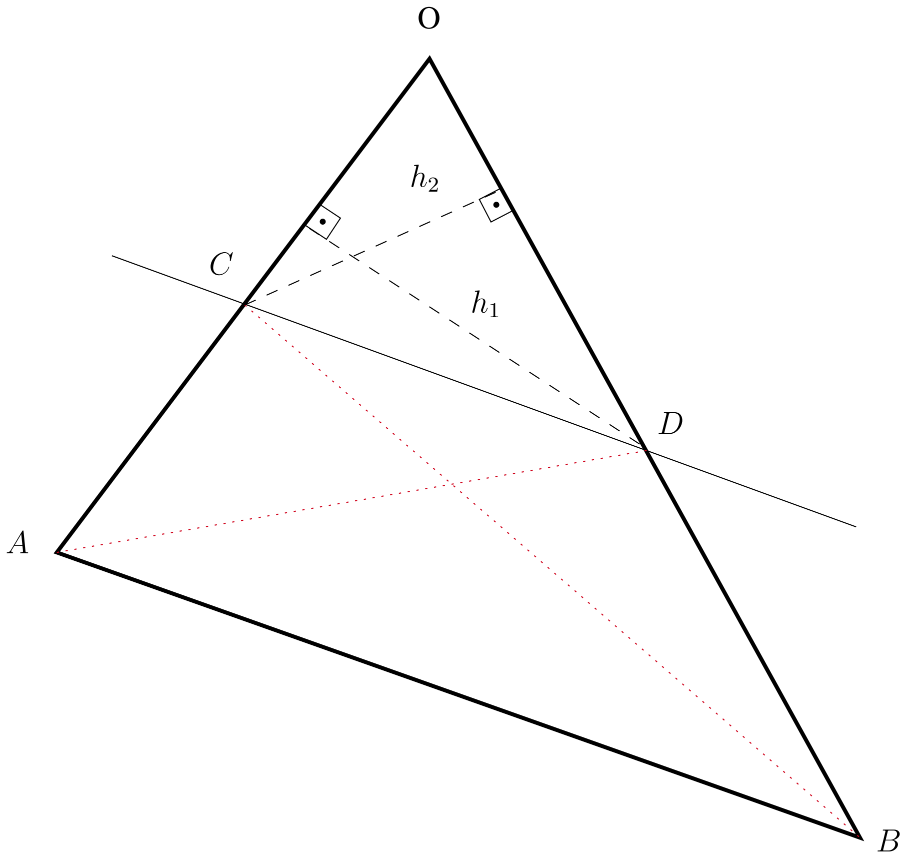


Figure 8.73: First approach of Thales' theorem

With:

$$\overline{AB}//CD \tag{8.2.79}$$

$$h_1 \perp \overline{\text{O}A} \tag{8.2.80}$$

$$h_2 \perp \overline{\text{O}B} \tag{8.2.81}$$

We have proven previously that if two triangles have a common side and if the third vertices are on a parallel, then they have the same surface. So the triangles ΔACD and ΔBCD have the same surface!

Adding to each of these two surfaces that of the triangle $\Delta \text{O}CD$, we get that the triangles $\Delta \text{O}DA$ and $\Delta \text{O}CB$ have the same surface!

We deduce then that using the product in cross that:

$$\frac{S_{\Delta \text{O}CD}}{S_{\Delta \text{O}AD}} = \frac{S_{\Delta \text{O}CD}}{S_{\Delta \text{O}CB}} \tag{8.2.82}$$

Given h_1 the height issue from D in the triangle $\Delta \text{O}CD$ and h_2 the height issue from C in the triangle $\Delta \text{O}CD$:

$$\frac{S_{\Delta \text{O}CD}}{S_{\Delta \text{O}AD}} = \frac{\overline{\text{O}C}\dfrac{h_1}{2}}{\overline{\text{O}A}\dfrac{h_1}{2}} = \frac{\overline{\text{O}C}}{\overline{\text{O}A}} \quad \text{and} \quad \frac{S_{\Delta \text{O}CD}}{\Delta S_{\text{O}CB}} = \frac{\overline{\text{O}D}\dfrac{h_2}{2}}{\overline{\text{O}B}\dfrac{h_2}{2}} = \frac{\overline{\text{O}D}}{\overline{\text{O}B}} \tag{8.2.83}$$

Conclusion:

$$\boxed{\frac{\overline{\text{O}C}}{\overline{\text{O}A}} = \frac{\overline{\text{O}D}}{\overline{\text{O}B}}} \tag{8.2.84}$$

Given now the figure:

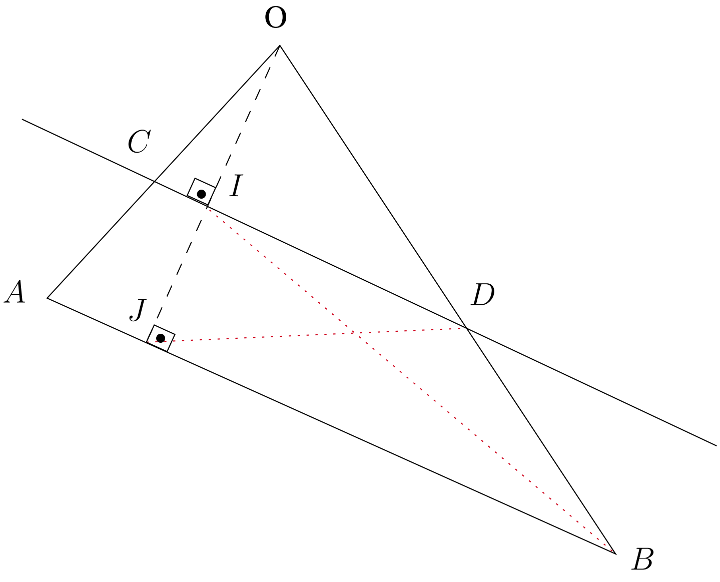


Figure 8.74: **Second approach of Thales' theorem**

The triangles ΔIJD and ΔIDB have the same surface from our first lemma, also same for the triangles $\Delta \text{O}JD$ and $\Delta \text{O}IB$ therefore:

$$S_{\Delta \text{O}JD} = \frac{1}{2}\overline{\text{O}J} \cdot \overline{DI} \qquad S_{\Delta \text{O}IB} = \frac{\overline{\text{O}J} \cdot \overline{JB}}{2} - \frac{\overline{JI} \cdot \overline{JB}}{2} = \frac{1}{2}\overline{\text{O}I} \cdot \overline{BJ} \tag{8.2.85}$$

hence:

$$\overline{\mathrm{O}J}\cdot\overline{DI}=\overline{\mathrm{O}I}\cdot\overline{BJ} \tag{8.2.86}$$

therefore:

$$\boxed{\frac{\overline{\mathrm{O}I}}{\overline{\mathrm{O}J}}=\frac{\overline{DI}}{\overline{BJ}}} \tag{8.2.87}$$

In the same way, in the triangles $\Delta \mathrm{O}IA$ and $\Delta \mathrm{O}CJ$, we get:

$$\frac{\overline{\mathrm{O}I}}{\overline{\mathrm{O}J}}=\frac{\overline{CI}}{\overline{AJ}} \tag{8.2.88}$$

From the second lemma as:

$$\boxed{\frac{\overline{\mathrm{O}I}}{\overline{\mathrm{O}J}}=\frac{\overline{DI}}{\overline{BJ}}=\frac{\overline{CI}}{\overline{AJ}}} \tag{8.2.89}$$

Then:

$$\frac{\overline{\mathrm{O}I}}{\overline{\mathrm{O}J}}=\frac{\overline{DI}}{\overline{BJ}}=\frac{\overline{CI}}{\overline{AJ}}=\frac{\overline{DI}+\overline{CI}}{\overline{BJ}+\overline{AJ}}=\frac{\overline{D}}{\overline{BA}} \tag{8.2.90}$$

By applying to the first triangle $\Delta \mathrm{O}JB$ what we did during the first approach, we also get:

$$\frac{\overline{\mathrm{O}I}}{\overline{\mathrm{O}J}}=\frac{\overline{\mathrm{O}D}}{\overline{\mathrm{O}B}} \tag{8.2.91}$$

So finally by taking all previous results we get:

$$\boxed{\frac{\overline{\mathrm{O}I}}{\overline{\mathrm{O}J}}=\frac{\overline{DC}}{\overline{BA}}=\frac{\overline{\mathrm{O}D}}{\overline{\mathrm{O}B}}} \tag{8.2.92}$$

which constitutes the "Thales' theorem" of ratios.

■ Q.E.D.

Here is an excellent summary of what we have seen so far about Thales' theorem:

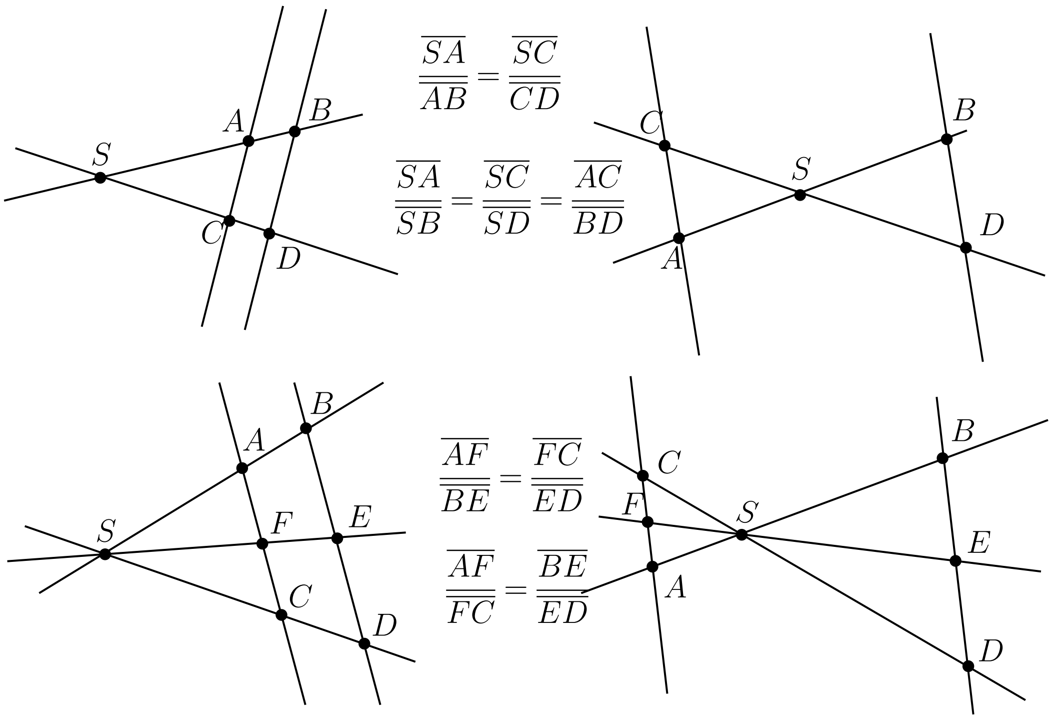
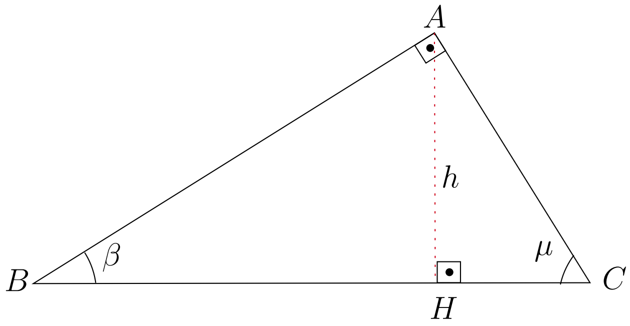

$\frac{\overline{SA}}{\overline{AB}}=\frac{\overline{SC}}{\overline{CD}}$
$\frac{\overline{SA}}{\overline{SB}}=\frac{\overline{SC}}{\overline{SD}}=\frac{\overline{AC}}{\overline{BD}}$
$\frac{\overline{AF}}{\overline{BE}}=\frac{\overline{FC}}{\overline{ED}}$
$\frac{\overline{AF}}{\overline{FC}}=\frac{\overline{BE}}{\overline{ED}}$
S
A
B
C
D
E
F

Figure 8.75: Thales' theorem (source: Wikipedia)

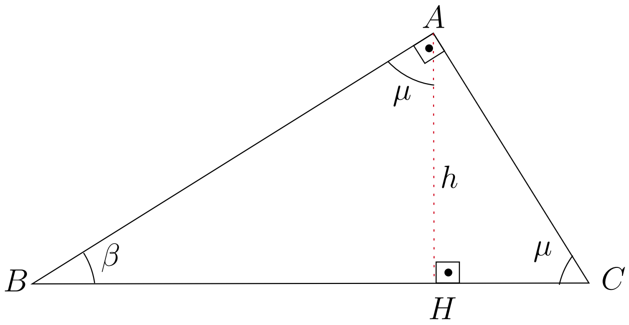
8.2.3.3.9.1 Euclid's (triangle) theorem

The Euclid's theorem uses Thales' theorem to give a relation in the below right angle triangle:

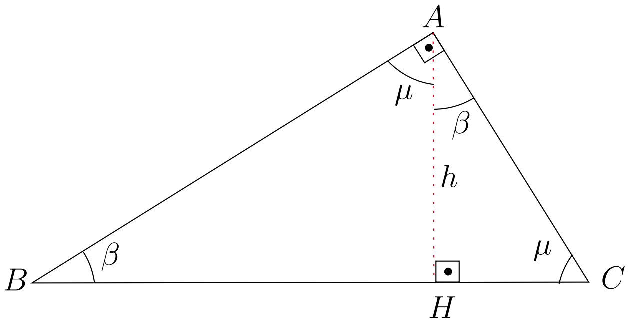


between h (i.e. $\overline{AH}$) and $\overline{BH}$ and $\overline{HC}$.

From the triangle above we see that $\beta+\mu=\pi/2$ as the sum of all angles of a triangle in the plane are equal to π. Therefore we can deduce that the angle μ is also here as the sum of the angles of ΔABH must also be equal to π:



Applying the same reasoning in the triangle ΔAHC we see also that the missing angle is by deduction obviously β itself:



Therefore the both triangles ΔABH and ΔAHC are similar and we can apply the Thales' theorem such that:

$$\frac{\overline{BH}}{\overline{AH}}=\frac{\overline{AH}}{\overline{HC}} \tag{8.2.93}$$

Rearranging we get:

$$\overline{AH}^2=\overline{BH}\cdot\overline{HC} \tag{8.2.94}$$

Or written differently:

$$\boxed{h^2=\overline{BH}\cdot\overline{HC}} \tag{8.2.95}$$

And that's the "Euclid's theorem" for right angle triangles in plane.

However, always in the same triangles we can write:

$$\frac{\overline{HC}}{\overline{AC}}=\frac{\overline{AC}}{\overline{BC}} \tag{8.2.96}$$

Therefore we get the second equality of Euclid's theorem for right triangles:

$$\overline{AC}^2=\overline{HC}\cdot\overline{BC} \tag{8.2.97}$$

And similarly we can get the third equality of Euclid's theorem:

$$\overline{AB}^2=\overline{HB}\cdot\overline{BC} \tag{8.2.98}$$

Remark

Sadly there are three theorems in mathematics named "Euclid's theorem" therefore that name is often quite confusing! Indeed, one of them is the theorem just seen above, the second one is a theorem of Euclid about the congruence of triangles (not presented in this textbook) and the third one is about prime numbers (also not presented in this textbook) but also named "Euclid's lemma" (see section Number Theory page 351).

8.2.3.4 Parallelism

Definition (#8.2.338): We name "parallels" two line (non-necessarily straight one!) equally distant from one another over their entire length.

This concept is directly connected to Euclid's fifth postulate (parallel postulate) and is often considered as the most important being been proved that it can not be considered as an axiom as it is not respected in non-Euclidean geometries (see section Non-Euclidean Geometry page 2663).

The consequences of this postulate are as we know the following in a Euclidean geometry:

1. If two lines AB and CD are parallels, any straight line $E'F'$ that cut one, cut the other one.

   Proof 8.10.2. Given F the common point to the straight line CD and the straight line $E'F'$: if the straight line $E'F'$ did not cut the straight line AB, it would be parallel to it, and through the point F would pass two straight lines CD and $E'F'$ parallel to a third one AB, which is not the case. So the straight line $E'F'$, cuts the line AB.

   ■ Q.E.D.

2. Two straight lines AB and CD parallels to a third line $E'F'$ are parallel between them.

   Proof 8.10.3. If the straight line CD was not parallel to the straight line AB it would cut (cross) it. It would also cut (cross) the straight line $E'F'$ parallel to the straight line AB, it would therefore not be parallel to $E'F'$.

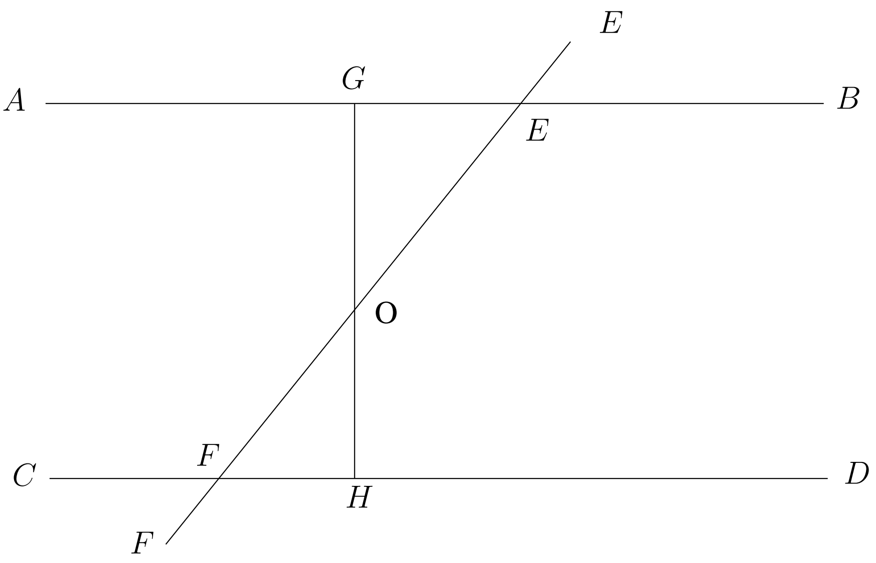
   ■ Q.E.D.

Theorem 8.11. If two parallels lines are cut by a secant then:

1. The intern-alternate angles are equal;
2. The external-alternate angles are equal;
3. The corresponding angles are equal.

If two lines are parallel then their alternate interior angles are equal. If the alternate interior angles of two lines are equal then the lines must be parallel. This is commonly named the "Z theorem".

Proof 8.11.1. Consider two (coplanar) parallel AB and CD and a secant EF:



1. By the middle O of EF let us conduct the perpendicular $\overline{GH}$ to AB, which is also perpendicular to CD. The right triangles $E\mathrm{O}G$ and $F\mathrm{O}H$ an acute angle equal to $\widehat{G\mathrm{O}E} = \widehat{F\mathrm{O}H}$ and an equal hypotenuse: $\mathrm{O}F = \mathrm{O}E$. They are equal, and the angles $\widehat{\mathrm{O}FH}$ and $\widehat{GE\mathrm{O}}$ are equal.

2. The alternate exterior angles $\widehat{E'EB}$ and $\widehat{F'FC}$ are equal, because $\widehat{E'EB}$ is opposed by the apex at the angle $\widehat{GE\mathrm{O}}$, which an intern-alternate with the angle $\widehat{\mathrm{O}FH}$.

■ Q.E.D.

... and vice versa if two straight lines are cut by a secant forming with these straight lines:

1. Two intern-alternate equal angles;
2. Two external-alternate equal angles;
3. Two corresponding equal angles.

then these two lines are parallel!

Remark

To prove the parallelism of two straight lines, it is necessary and sufficient that the intern-alternate angles, extern-alternate angles or corresponding angles, formed by these two straight lines with secant transversal, are all equal. That is to say equal to $\pi/2$.

8.2.3.5 Circle

Definition (#8.2.339): We name "circle" the locus of points M of the plane that are a given distance R, named "radius" of the circle, of a fixed point O, named "center of the circle". Or using the concept of "locus" defined earlier: a circle is defined as the locus of a point that is at a given distance of a fixed point, the center of the circle:

$$\overline{\mathrm{O}M}=R \tag{8.2.99}$$

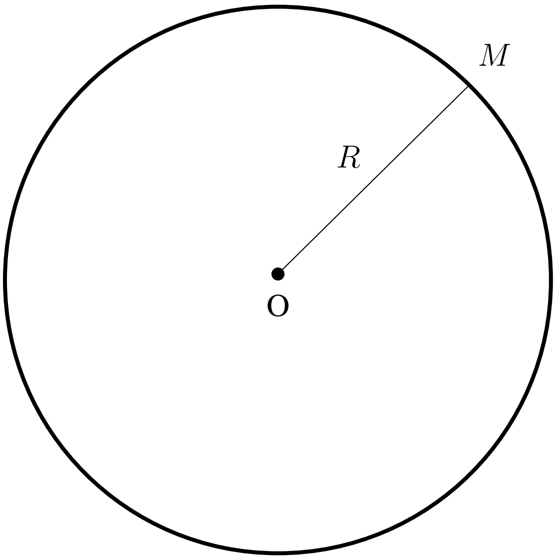


Figure 8.76: **Circle and Radius**

Remark

The word "radius" means either the segment $\overline{\mathrm{O}M}$, or its measurement R.

Circles are directly concerned by the third Euclid that we stated earlier above.

We name "diameter", denoted $\varnothing$, of a circle every line passing through the center O of the circle. Any diameter meets the circle at two points A and B, such that by construction:

$$\varnothing :=\overline{\mathrm{O}A}=\overline{\mathrm{O}B}=R \tag{8.2.100}$$

which we name "extremities of the diameter". We reserve the expression "diameter $\overline{AB}$" for the diameter of extremities A and B. We say that two points of a circle are "diametrically opposed" when they are the two ends of the same diameter:

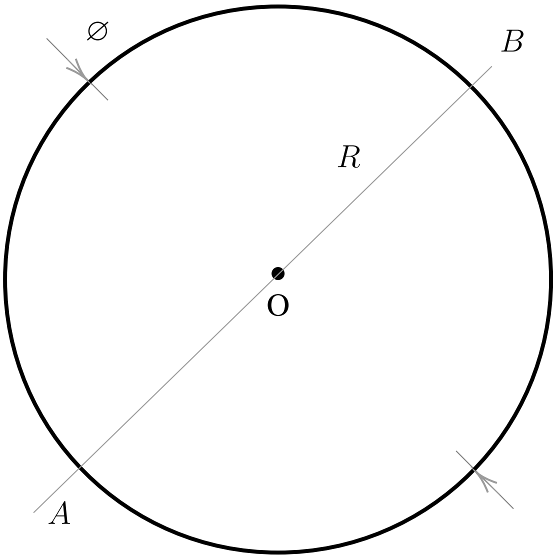
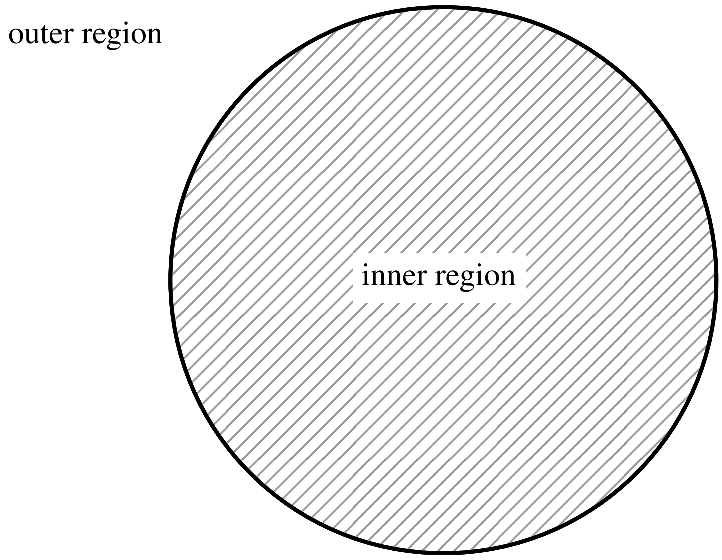


Figure 8.77: **Circle and Diameter**

A circle divides the plane into two parts: one which contains the center, which we name "inner region" and one that does not contain it, which we name "outer region":



Theorem 8.12. The necessary and sufficient condition for a point P to be strictly inside a circle (C), of center O and radius R, is $\overline{OP} < R$.

Proof 8.12.1. There are two case to consider for the proof:

1. The condition is necessary: If, by hypothesis, P is inside the circle (C), it is located either on O, or between the ends A and B of the diameter defined by the locus of the points M. If it is on O the proposition is obvious, if it is not on O, then it is between O and A for example, and we have $\overline{\text{O}P} < \overline{\text{O}A}$, that is to say $\overline{OP} < R$.

2. The condition is sufficient: If, by hypothesis $\overline{OP} < R$, P is between the extremities A and B of the loci defined by the points M, so inside the circle (C).

■ Q.E.D.

Corollary 8.12.1. The necessary and sufficient condition for a point P to be outside a circle (C) is $\overline{\text{O}P} > R$.

We name "string of a circle" (C) the segment $\overline{CD}$ whose extremities C and D are on the circle (on the loci of M) as visible in the figure below.

Theorem 8.13. The mediator of a string $\overline{CD}$ is a diameter and obviously $C\text{O}D$ is an isosceles triangle.

Proof 8.13.1. The mediator Δ of $\overline{CD}$ (see figure below), string of the circle (C) of center O and radius R pass through the point O because as proved it during our study of the triangles we have $\overline{\text{O}R} = \overline{\text{O}D} = R$.

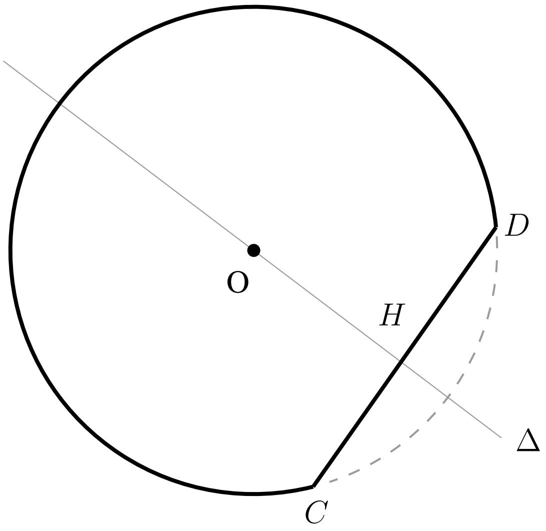


Figure 8.78: Mediator of circle string

■ Q.E.D.

Corollary 8.13.1. The perpendicular drawn from the center O of a circle of any circle string $\overline{CD}$ pass by construction through the middle H of this string.

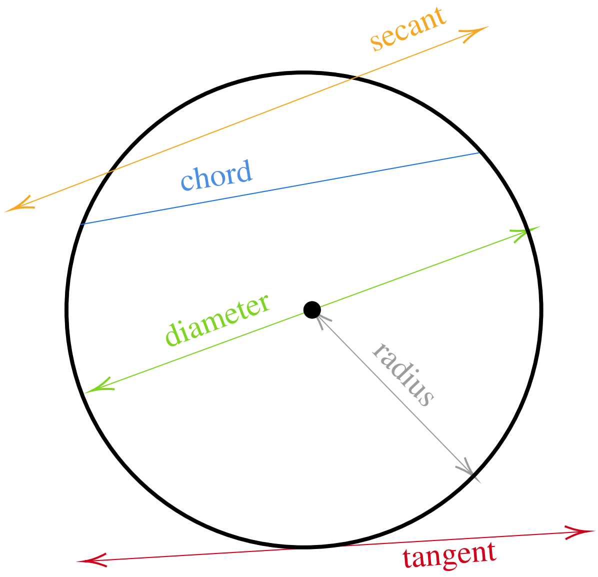


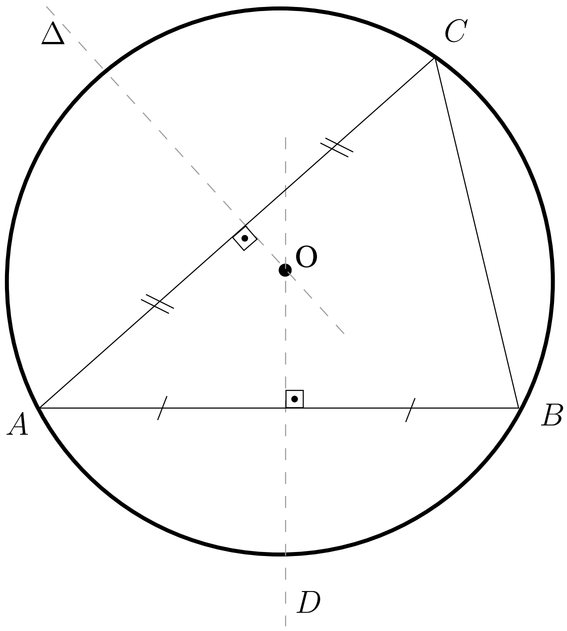
Figure 8.79: **Some common vocabulary for a circle**

8.2.3.5.1 Circumscribed circle theorem

Theorem 8.14. By three points non-aligned A, B, C we can draw a circle and only one. This is the "circumscribed circle theorem".

Or in other words: All triangles are cyclic, i.e. every triangle has a circumscribed circle or "circumcircle".

Proof 8.14.1. First, we draw the mediator D of the segment $\overline{AB}$ and the mediator Δ of the segment $\overline{AC}$. If D and Δ were parallel, the perpendicular $\overline{AB}$ to D would be perpendicular also to Δ so coinciding with $\overline{AC}$. The points A, B, C would then be aligned. Therefore D and Δ are non-parallel and intersect at a point O:



1. There is one circle going through A, B, C:

   The point O being on D, mediator of $\overline{AB}$, $\overline{\mathrm{O}A}=\overline{\mathrm{O}B}$ by construction: the point O being on D, mediator of $\overline{AC}$, $\overline{\mathrm{O}A}=\overline{\mathrm{O}C}$. The circle (C), of center O and of radius $\overline{\mathrm{O}A}$ passes through B (since $\overline{\mathrm{O}A}=\overline{\mathrm{O}B}$) and also by C (since $\overline{\mathrm{O}A}=\overline{\mathrm{O}C}$). It therefore passes through A, B, C.

2. It goes through the points A, B, C only one circle:

   If it was going through A, B, C a different circle than that (C) of center O and radius $\overline{\mathrm{O}A}$, its center O' would be on the mediator of $\overline{AB}$ and $\overline{AC}$ that are two strings of the circle; it is then confused with O!

■ Q.E.D.

Remark

The mediator of $\overline{BC}$, string of the circle (C) above, also passes through the point O. We can say (important result!) that the three mediators of the sides of a triangle ABC are concurrent in a same point named the "orthocenter".

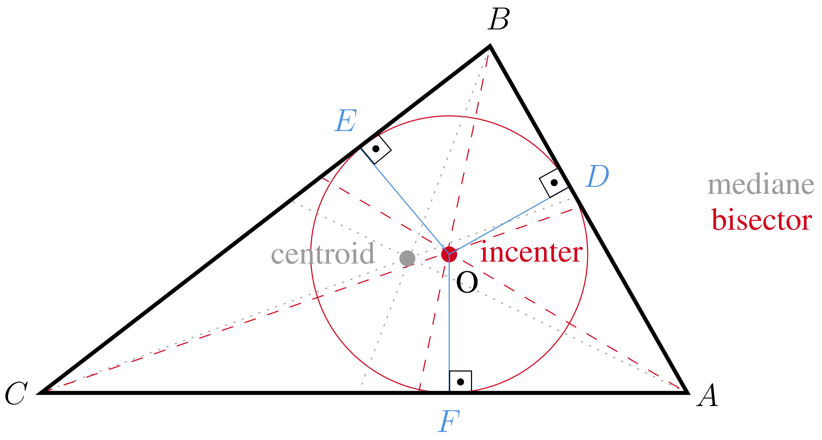
We will see further below that using the law of sines it's quite easy to get the diameter of the circumscribed circle and therefore its surface if we know its intern-angles. But if we don't know its intern-angles we need the center angles theorem that we will prove further below to calculate its surface.

8.2.3.5.2 Inscribed circle theorem

Theorem 8.15. It is the opposite of the previous theorem and states therefore a unique circle can be inscribed in any triangle, i.e. every triangle has an incircle tangent to it.

Proof 8.15.1. Given ABC, bisect the angles at the vertices A and B.

These angles bisectors must interact at a point O:



We locate afterwards the point D, E and F on sides $\overline{AB}$, $\overline{BC}$ and $\overline{CA}$ respectively so that:

$$\overline{\mathrm{O}D}\perp\overline{AB} \qquad \overline{\mathrm{O}E}\perp\overline{BC} \qquad \overline{\mathrm{O}F}\perp\overline{CA}$$

Observe that the triangles $\Delta A\mathrm{O}D$ and $\Delta A\mathrm{O}F$ are congruent and also the triangles $\Delta B\mathrm{O}D$ and $\Delta B\mathrm{O}E$ by the angle-side-angle theorem.

Since corresponding sides of congruent triangles are equal, we also know that:

$$\overline{\mathrm{O}D}=\overline{\mathrm{O}F} \quad \text{and } \overline{\mathrm{O}D}=\overline{\mathrm{O}E} \tag{8.2.101}$$

Hence:

$$R=\overline{\mathrm{O}D}=\overline{\mathrm{O}F}=\overline{\mathrm{O}E} \tag{8.2.102}$$

So the circle with center O and radius R is an incircle for the triangle.

■ Q.E.D.

We can be interested to calculate the surface of the incircle of radius R of the triangle ΔABC knowing only the distances between the points A, B and C.

So following the above figure, we consider the triangle ΔABC with incircle having center at O, we notice that:

$$S_{\Delta ABC}=S_{\Delta A\mathrm{O}B}+S_{\Delta B\mathrm{O}C}+S_{\Delta C\mathrm{O}A} \tag{8.2.103}$$

Thus:

$$S_{\Delta ABC}=\frac{1}{2}\overline{AB}R+\frac{1}{2}\overline{BC}R+\frac{1}{2}\overline{CA}R \tag{8.2.104}$$

That is:

$$S_{\Delta ABC}=\frac{R}{2}\left(\overline{AB}r+\overline{BC}r+\overline{CA}\right) \tag{8.2.105}$$

So:

$$S_{\Delta ABC}=\frac{R}{2}P_{\Delta ABC} \tag{8.2.106}$$

Therefore:

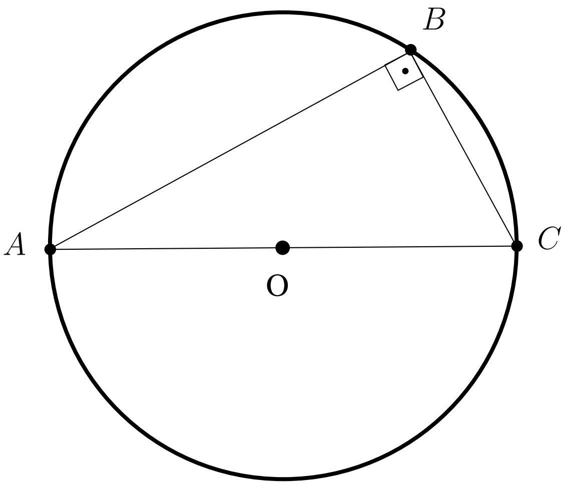
$$\boxed{R=\frac{2S_{\Delta ABC}}{P_{\Delta ABC}}} \tag{8.2.107}$$

where P is the perimeter of the triangle.

8.2.3.5.3 Thales' theorem of the circle

Theorem 8.16. If A, B and C are points on a circle where the segment $\overline{AC}$ is a diameter of the circle, then the angle $\widehat{ABC}$ is a right angle. This is the "Thales' theorem for circle".

Proof 8.16.1. Given the figure:



Since $\overline{\mathrm{O}A} = \overline{\mathrm{O}B} = \overline{\mathrm{O}C}$, the triangles $\Delta \mathrm{O}BA$ and $\Delta \mathrm{O}BC$ are therefore isosceles triangles, and by the equality of the base angles of an isosceles triangle, $\widehat{\mathrm{O}BC} = \widehat{\mathrm{O}CB}$ and $\widehat{BA\mathrm{O}} = \widehat{AB\mathrm{O}}$.

Let $\alpha = \widehat{BA\mathrm{O}}$ and $\beta = \widehat{\mathrm{O}BC}$. The three internal angles of the triangle ABC are α, $(\alpha+\beta)$, and β. Since the sum of the angles of a triangle is equal to π [rad] (180°), we have:

$$\begin{aligned} &\alpha + (\alpha+\beta) + \beta = \pi \\ &\Rightarrow 2\alpha + 2\beta = \pi \\ &\Rightarrow 2(\alpha+\beta) = \pi \\ &\therefore \alpha+\beta = \frac{\pi}{2} \end{aligned} \tag{8.2.108}$$

■ Q.E.D.

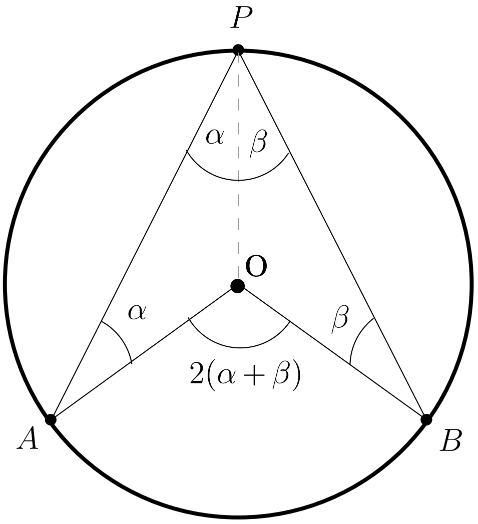
Corollary 8.16.1. If and only if the inscribed triangle is a right triangle, the circumcenter lies at the center of the hypotenuse.

8.2.3.5.4 Central angle theorem

The "central angle theorem" is very useful in solving questions that deals with angles within circles as we will see just after. The use of this theorem often simplifies a complicated situation into a rather simple one.

Theorem 8.17. The central angle theorem states that the central angle from two chosen points A and B on the circle is always twice the inscribed angle from those two points. The inscribed angle can be defined by any point P along the outer arc AB and the two points A and B.

Proof 8.17.1. Consider the following figure:



If we denote $\widehat{AP\text{O}}=\alpha$ and $\widehat{BP\text{O}}=\beta$ then it follows as $\Delta P\text{O}A$ and $\Delta P\text{O}B$ are isosceles triangle that:

$$\widehat{PA\text{O}}=\alpha \quad \text{and} \quad \widehat{PB\text{O}}=\beta \tag{8.2.109}$$

Using the fact that the sum of the angles of a triangle is equal to π (or $180°$) then:

$$\widehat{P\text{O}A}=\pi-2\alpha \tag{8.2.110}$$

$$\widehat{P\text{O}B}=\pi-2\beta \tag{8.2.111}$$

Since all angles at center O must add up to 2π (or $360°$), we have:

$$\widehat{A\text{O}B}=2\pi-(\pi-2\alpha)-(\pi-2\beta)=2\alpha+2\beta=2(\alpha+\beta) \tag{8.2.112}$$

Therefore:

$$\boxed{\widehat{A\text{O}B}=2\widehat{APB}} \tag{8.2.113}$$

Don't forget that this result is valid whatever the position of P on the circle!!!!!!

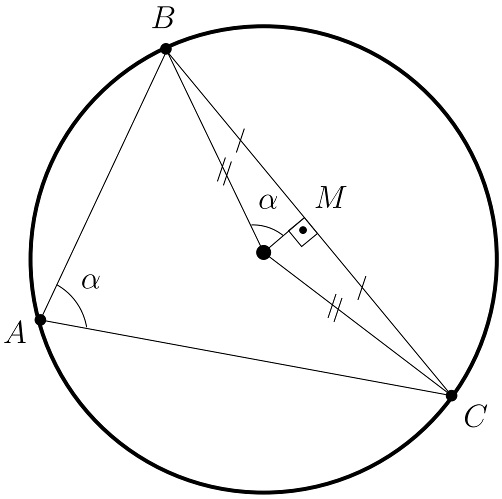
■ Q.E.D.

Using this theorem we can now prove that the relation that appears in the law of sines (see section Trigonometry page 2528) is the diameter of the circumcircle of the triangle ABC.

Theorem 8.18.

$$\varnothing=\frac{a}{\sin(\alpha)}=\frac{b}{\sin(\beta)}=\frac{c}{\sin(\gamma)} \tag{8.2.114}$$

Proof 8.18.1. For the proof consider the following figure:



Given the triangle ABC, let O denoted the center of its circumcircle.

Observe that $\widehat{BOC}=2\widehat{BAC}$ due to the central angle theorem.

That is:

$$\widehat{BOC}=2\alpha \tag{8.2.115}$$

Let M be the midpoint of $\overline{BC}$.

Then the triangle ΔBOM is congruent to the triangle ΔCOM by the side-side-side theorem.

So we have that:

$$\widehat{BOM}=\widehat{COM}=\frac{1}{2}\widehat{BOC} \tag{8.2.116}$$

since congruent angles in congruent triangles are equal.

By construction we have:

$$\widehat{OMB}=\widehat{OMC}=\frac{\pi}{2} \tag{8.2.117}$$

Let us now write:

$$\widehat{BOM}=\alpha \tag{8.2.118}$$

Now in the right triangle BOM we see that:

$$\sin(\alpha)=\frac{\overline{BM}}{\overline{BO}}=\frac{2\overline{BM}}{2\overline{BO}}=\frac{\overline{BM}+\overline{MC}}{2R}=\frac{\overline{BC}}{2R}=\frac{a}{\varnothing} \qquad (8.2.119)$$

Therefore:

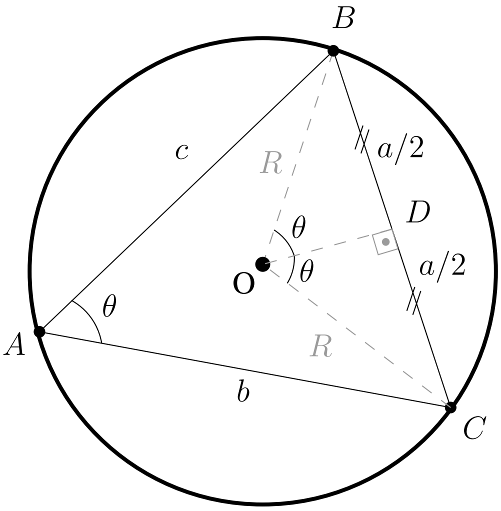
$$\varnothing=\frac{a}{\sin(\alpha)} \qquad (8.2.120)$$

Hence, by applying the law of sines, we conclude that:

$$\boxed{\varnothing=\frac{a}{\sin(\alpha)}=\frac{b}{\sin(\beta)}=\frac{c}{\sin(\gamma)}} \qquad (8.2.121)$$

■ Q.E.D.

So we can easily calculate the surface knowing the intern angles. But consider now that we don't know the intern-angles. So we also use the central angle theorem and the following figure:



From triangle $\Delta BD\text{O}$ we have:

$$\sin(\theta)=\frac{a/2}{R}=\frac{a}{2R} \qquad (8.2.122)$$

As we will prove in the section of Geometric Shapes, the surface of the internal triangle is:

$$S=\frac{1}{2}\text{basis}\cdot\text{height}=\frac{1}{2}b\cdot h \qquad (8.2.123)$$

So if we choose for basis of the triangle b and knowing that therefore:

$$\sin(\theta)=\frac{h}{c} \qquad (8.2.124)$$

we get:

$$S=\frac{1}{2}bc\sin(\theta)=\frac{1}{2}bc\left(\frac{a}{2R}\right)=\frac{abc}{4R} \qquad (8.2.125)$$

Therefore:

$$\boxed{R=\frac{abc}{4S_{\Delta ABC}}} \tag{8.2.126}$$

and we can then easily get the surface of the circumscribed circle from its radius.

8.2.4 Hilbert's Axioms

Euclid gathered all the geometric knowledge of his time in the form of the 5 postulates. He left his name to the Euclidean geometry that uses its fifth postulate, to non-Euclidean geometry that does not use it, and to Euclidean spaces.

This postulated base, however, is imperfect, to rigorously prove theorems associated with this geometry, it is necessary to accept as true additional implicit assumptions. David Hilbert built a corresponding axiomatic to the idea that made Euclid of the geometry by adding ad hoc assumptions (mathematicians make definitions with coffee...).

The Hilbert's axioms are grouped into five categories: the association, the order, the congruence, the continuity and the parallels.

Hilbert's axiom system is constructed with 6 primitive notions: three primitive terms and three primitive relations.

The three primitive terms are:

1. Point
2. Line
3. Plane

Remark

Points, lines and planes are considered a distinct by default.

The three primitive relations are:

1. That of association defines the word "contains" in geometry! It corresponds to the ideas "is part of" and "is included in" of Set Theory.
2. That of "order" is a binary relation between a couple of points and one point, it appears in the terms "between" and to define the segments.
3. That of "congruence", which corresponds to the three "equivalence relations" for the couples of points, the triangles and the angles.

Here are the "Hilbert's axioms":

8.2.4.1 Incidence Axioms (axioms of association)

A.A1. Given two points, there is a line going through these two points (contains them both).

A.A2. Given two points, there is one line and only one passing through these two points (verbatim the line described in A.A1).

A.A3. A line contains at least two points, and for a given line, there exists at least one point not contained into that line.

A.A4. Given three points that are not contained in a line, there is a plane containing the three points. Any plane contains at least one point.

A.A5. Given three points that are not contained in a line, there is only one plane containing the three points.

A.A6. Given two points contained in a line D and in a plane A, then A contains all the points of D.

A.A7. If two planes A and B both contain a point C, then the intersection of A and B contains at least one other point.

A.A8. There are at least four non coplanar points.

8.2.4.2 Order Axioms

A.O1. If point B is between the points A and C, B is then also between the points C and A, and there is a line containing the three points A, B, C.

A.O2. Given two points A and C, there is a point B element of the line $\overline{AC}$ such that C is between A and B.

A.O3. Given three points contained in a line, then a one and only is located between the other two.

A.O4. ("Pasch's theorem") Given three points A, B, C non-collinear, and given a line D in the plane ABC but containing none of the points A, B, C: If D contains a point of the segment $\overline{AB}$, then D contains either a point of the segment $\overline{AC}$ or a point of the segment $\overline{BC}$.

8.2.4.3 Congruence Axioms

First let us recall that intuitively "congruent" means in geometry "stackable" (exactly the same size and shape) at the opposite of "similar" (only one of the both properties are satisfied).

A.G1. Given two points A, B and a point A' element of a line d, there are two and two only unique points C and D, such that A' is between C and D, and $\overline{AB}$ is congruent to $\overline{A'C}$ and $\overline{AB}$ is congruent to $\overline{A'D}$.

A.G2. The congruence relation is transitive, that is to say, if $\overline{AB}$ is congruent to $\overline{CD}$ and if $\overline{CD}$ is congruent to $\overline{EF}$, then $\overline{AB}$ is congruent to $overlineEF$.

A.G3. Given a straight line d containing the adjacent segments $\overline{AB}$ and $\overline{BC}$, and given a straight line containing the adjacent segments $\overline{A'B'}$ and $\overline{B'C'}$. If $\overline{AB}$ is congruent to $\overline{A'B'}$ and $\overline{BC}$ is congruent to $\overline{B'C'}$ then $\overline{AC}$ is congruent to $\overline{A'C'}$.

A.G4. Given an angle $\widehat{ABC}$ and a half-line $\overline{B'C'}$, there are two and only two half-lines, $\overline{B'D}$ and $\overline{B'E}$, such that the angle $\widehat{DB'C'}$ is congruent to the angle $\widehat{ABC}$ and the angle $\widehat{EB'C'}$ is congruent to the angle $\widehat{ABC}$.

A.G5. Given two triangles ABC and $A'B'C'$ such that $\overline{AB}$ is congruent to $\overline{A'B'}$, $\overline{AC}$ is congruent to $\overline{A'C'}$, and the angle $\widehat{BAC}$ is congruent to the angle $\widehat{B'A'C'}$, then the triangle ABC is congruent to the triangle $A'B'C'$.

Therefore we see that these axioms are used to compare segments, and also angles, to define the center of a segment, the orthogonal lines, to talk about equilateral triangles, isosceles, etc. They also enable to strictly define the translations which was so often used by Euclid without having been defined.

8.2.4.4 Continuity Axioms

A.C1. Also named "Archimedes' axiom", this axiom assume that given two segments $\overline{AB}$ and $\overline{CD}$ then there is always a finite sequence of points $A_1,\ldots,A_n$ belonging to the line containing the segment $\overline{AB}$ and such that $A<A_1<\ldots<A_{n-1}<B<A_n$ that can satisfy $\overline{AA_1}\equiv\overline{AA_2}\equiv\ldots\overline{A_{n-1}A_n}\equiv\overline{CD}$.

A.C2. Also named "Cantor's Axiom" it assumes if (A_n) and (B_n) are two infinite sequence of points and that we build overlapping line segments $\overline{A_iB_i}$ such as $\overline{A_kB_k}\subset\overline{A_iB_i}$ if $i<k$ (also sometimes written $]A_{i+1},B_{i+1}[\subset]A_i,B_i[$) located on a line L and such that $\forall\overline{CD},\exists i:\overline{A_iB_i}\equiv\overline{CD}$, then there exists a point X belonging to all segments $\overline{A_iB_i}$. In other words: let there be a sequence of nested segments whose length tends towards 0 then there is a point common to all the segments.

It is not obvious but the Archimedean axiom is the result of the experience in a large number of measurements and guarantees the measurability of line segments. An arbitrary line segment can be attached a real number named the "line segment length". The axiom does not guarantee the inverse relation, that an arbitrary real number can be related to some line segment as the length of this line segment. This fact is guaranteed by the Cantor axiom.

8.2.4.5 Parallels Axioms

A.P1. Also known as the "Euclidean axiom" it assumes that given a straight line L and a point P not belonging to L, it passes one and only one straight line L' through P which is parallel to d.

Other equivalent formulation is:

A.P1'. Given a straight line L, a point P not included on L, then there exists a plane containing d and P. This plane contains one and only one straight line containing P and containing no point of d.

Any non-empty set of points (in the plane or in the space) consistent to the axiomatic system with the Euclidean postulate forms the Euclidean geometry (Euclidean plane $\mathcal{E}^2$, or Euclidean space $\mathcal{E}^3$) which is named also "parabolic geometry". One of the most well known theorems following directly from the Euclidean axiom of parallelism is that the sum of the interior angles in an arbitrary triangle equal π.

We can't really prove until now the non-contradictory logic of all these axioms. However, we know two things if we make an analogy with what we studied in the chapters Arithmetic and Algebra of this textbook (especially the section on Sets Theory, Functional Analysis and Sequences and Series):

1. If these axioms are contradictory, then the theory of real numbers is also contradictory.
2. If the system of axioms obtained by removing the axiom of Cantor is contradictory, then the theory of rational numbers is contradictory.

Thus, the confidence we have in the strength of these axioms lies on the one we have in the theory of real numbers, which is already also quite big.

8.2.5 Barycenter (centroid)

Now that we have discussed the minimum of the construction of Euclid and Hilbert geometry, we can move to a higher level to the analysis of properties of geometric shapes. We will begin by studying the concept of "center of gravity", also referred to but rarely to "centroid" or "barycenter" and that is very important not only in geometry but also in classical mechanics, astronomy and many other domains where multiple masses must be reduced to a single virtual point.

Remark

The definition of the barycenter requires some of the mathematical tools defined in the Vector Calculus section (see page 1794). Reading this section is therefore recommended if the reader wishes to understand what follows.

Definition (#8.2.340): We name "center of gravity" or "centroid" or "geometric center" of the points A_i of the plane or of the space affected respectively by the coefficients $\alpha_1, \alpha_2, \ldots, \alpha_n$ (where the α_i are real such that $\sum \alpha_i \neq 0$) the single point G (or also denoted sometimes CM for "Center of Mass") such that:

$$\boxed{\sum_{i=1}^{n} \alpha_i \overrightarrow{GA_i} = \vec{0}} \tag{8.2.127}$$

The couple denoted (A_i, α_i) is named "ponderated point" or "massive point" in the context of study of physics and when $\alpha_i > 0$ it represents the most often a mass. If all the α_i have the same value then the centroid is assimilated obviously to the mean position of all the points in all of the coordinate directions. Therefore a physical object has uniform density, then its center of mass is the same as the centroid of its shape!

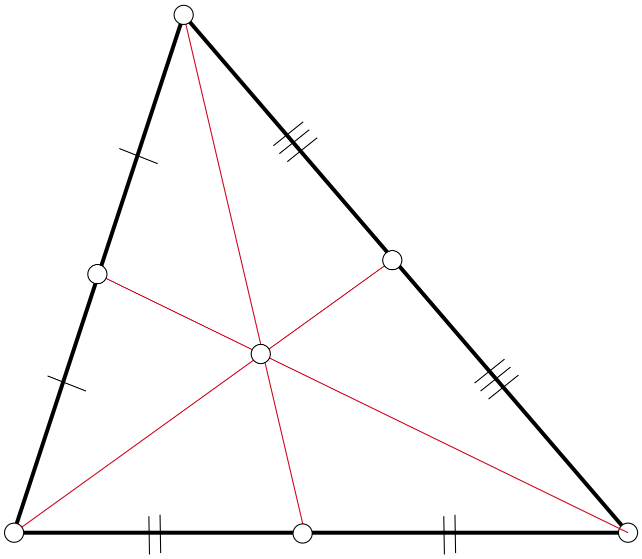


Figure 8.80: **Example of planar centroid (source: Wikpedia)**

Remarks

R1. In physics, the "center of inertia" of a body corresponds to the centroid of the particles that make up the body in question. Each particle being weighted by its own mass. So this is the point from which the mass is evenly distributed. If the density is constant, the center of mass coincides with the center of gravity as already mention.

R2. The "center of gravity" of a body corresponds to the centroid of the particles that make up the body in question, each particle being weighted by its own weight! Very often in mechanics, the size of the body is small compared with the roundness of the Earth, we consider a uniform gravity field. Under this assumption, the center of gravity and the center of mass coincide.

R3. When for any massive point (A_i,α_i) we have $\alpha_i=\alpha_j\;\forall(i,j)\in\mathbb{N}^*$, then we speak of "isobarycenter".

For an arbitrary point O, we obviously have by simple vector addition:

$$\sum_{i=1}^{n}\alpha_i\overrightarrow{GA_i}=\sum_{i=1}^{n}\alpha_i\left(\overrightarrow{\mathrm{O}A_i}-\overrightarrow{\mathrm{O}G}\right)=\sum_{i=1}^{n}\alpha_i\overrightarrow{\mathrm{O}A_i}-\left(\sum_{i=1}^{n}\alpha_i\right)\overrightarrow{\mathrm{O}G}=\vec{0} \tag{8.2.128}$$

hence the major result:

$$\boxed{\overrightarrow{\mathrm{O}G}=\frac{\displaystyle\sum_{i=1}^{n}\alpha_i\overrightarrow{\mathrm{O}A_i}}{\displaystyle\sum_{i=1}^{n}\alpha_i}} \tag{8.2.129}$$

Passing to the limit, if the domain is continuous (or can be considered as), we have:

$$\boxed{\overrightarrow{\mathrm{O}G}=\frac{\displaystyle\int\overrightarrow{\mathrm{O}M}\mathrm{d}l}{\displaystyle\int\mathrm{d}l}=\frac{\displaystyle\int\overrightarrow{\mathrm{O}M}\mathrm{d}l}{L}} \tag{8.2.130}$$

We may well also work with the surface or volumes elements (to mention only the more trivial case) to determine the centroid:

$$\overrightarrow{\mathrm{O}G}=\frac{\displaystyle\iint\overrightarrow{\mathrm{O}M}\mathrm{d}S}{\displaystyle\iint\mathrm{d}S}=\frac{\displaystyle\iint\overrightarrow{\mathrm{O}M}\mathrm{d}S}{S}\qquad\text{and}\qquad\overrightarrow{\mathrm{O}G}=\frac{\displaystyle\iiint\overrightarrow{\mathrm{O}M}\mathrm{d}V}{\displaystyle\iiint\mathrm{d}V}=\frac{\displaystyle\iiint\overrightarrow{\mathrm{O}M}\mathrm{d}V}{V} \tag{8.2.131}$$

✍**Example:**

To calculate for example the centroid of the area between the $0x$ axis and the parabola function $y=kx^2$ for $0\leq x\leq a$ we have (do not forget that when we calculate the centroid along the y axis that we have to take it the right side of the curve!):

$$\overrightarrow{\mathrm{OG}}=\begin{pmatrix}G_x\\G_y\end{pmatrix}=\begin{pmatrix}\dfrac{\iint x\mathrm{d}S}{\iint \mathrm{d}S}\\\dfrac{\iint xy\mathrm{d}S}{\iint \mathrm{d}S}\end{pmatrix}=\begin{pmatrix}\dfrac{\iint x\mathrm{d}S}{\iint \mathrm{d}S}\\\dfrac{\iint xkx^2\mathrm{d}S}{\iint \mathrm{d}S}\end{pmatrix} \tag{8.2.132}$$

in the first component we have $\mathrm{d}S=y\mathrm{d}x$ so the double integral becomes a simple integral!

For the second component we can't use $\mathrm{d}S=x\mathrm{d}y$ as it represent an element of area between the y-axis and the parabola curve. Indeed, we want the outside area (above the parabola curve when see of the point of view of the y-axis!). Therefore the idea is to translate the parabola of a so that the parabola surface relatively to y will be the expected one. Therefore we have $\mathrm{d}S=(x-a)\mathrm{d}y$.

And it comes:

$$\begin{aligned}\overrightarrow{\mathrm{OG}}&=\begin{pmatrix}\dfrac{\iint x\mathrm{d}S}{\iint \mathrm{d}S}\\\dfrac{\iint xkx^2\mathrm{d}S}{\iint \mathrm{d}S}\end{pmatrix}=\begin{pmatrix}\dfrac{\int\limits_0^a xy\mathrm{d}x}{\int\limits_0^a y\mathrm{d}x}\\\dfrac{\int\limits_0^b kx^2(x-a)\mathrm{d}y}{\int\limits_0^b (x-a)\mathrm{d}y}\end{pmatrix}=\begin{pmatrix}\dfrac{\int\limits_0^a xkx^2\mathrm{d}x}{\int\limits_0^a kx^2\mathrm{d}x}\\\dfrac{\int\limits_0^a kx^2(x-a)2kx\mathrm{d}x}{\int\limits_0^a (x-a)2kx\mathrm{d}x}\end{pmatrix}\\&=\begin{pmatrix}\dfrac{\int\limits_0^a x^3\mathrm{d}x}{\int\limits_0^a x^2\mathrm{d}x}\\k\dfrac{\int\limits_0^a x^3(x-a)\mathrm{d}x}{\int\limits_0^a x(x-a)\mathrm{d}x}\end{pmatrix}=\begin{pmatrix}\dfrac{\left.\frac{1}{4}x^4\right|_0^a}{\left.\frac{1}{3}x^3\right|_0^a}\\\dfrac{\left.\frac{1}{5}x^5-\frac{1}{4}ax^4\right|_0^a}{\left.\frac{1}{3}x^3-\frac{1}{2}ax^2\right|_0^a}\end{pmatrix}=\begin{pmatrix}\dfrac{3}{4}\dfrac{a^4}{a^3}\\k\dfrac{\frac{1}{5}a^5-\frac{1}{4}a^5}{\frac{1}{3}a^3-\frac{1}{2}a^3}\end{pmatrix}=\begin{pmatrix}\dfrac{3}{4}a\\k\dfrac{-\frac{1}{20}a^5}{-\frac{1}{6}a^3}\end{pmatrix}\\&=\begin{pmatrix}\dfrac{3}{4}a\\k\dfrac{3}{10}a^2\end{pmatrix}\end{aligned} \tag{8.2.133}$$

Remark

When we calculate the center of gravity of a function as in the example above, then we speak of "weighted curve".

In space with a reference frame $(\mathrm{O},\vec{i},\vec{j},\vec{k})$ by writing (x_i,y_i,z_i) the coordinates of the weighted points (A_i,α_i) and (x_G,y_G,z_G) those of G, then we have (in the discrete case):

$$\boxed{x_G=\frac{\displaystyle\sum_{i=1}^{n}\alpha_i x_i}{\displaystyle\sum_{i=1}^{n}\alpha_i}\qquad y_G=\frac{\displaystyle\sum_{i=1}^{n}\alpha_i y_i}{\displaystyle\sum_{i=1}^{n}\alpha_i}\qquad z_G=\frac{\displaystyle\sum_{i=1}^{n}\alpha_i z_i}{\displaystyle\sum_{i=1}^{n}\alpha_i}}\tag{8.2.134}$$

Let us now give and prove some properties of the centroid (center of mass/gravity):

P1. Given $(A_1,\alpha_1),(A_2,\alpha_2),\ldots,(A_n,\alpha_n)$, n weighted points. If $\sum\alpha_i\neq 0$, we have then for any point M:

$$\sum_{i=1}^{n}\alpha_i\overrightarrow{MA_i}=\left(\sum_{i=1}^{n}\alpha_i\right)\overrightarrow{MA_i}\tag{8.2.135}$$

Proof 8.18.2.

$$\begin{aligned}\sum_{i=1}^{n}\alpha_i\overrightarrow{MA_i}&=\alpha_1\overrightarrow{MA_1}+\alpha_2\overrightarrow{MA_2}+\ldots+\alpha_n\overrightarrow{MA_n}\\&=\alpha_1\left(\overrightarrow{MG}+\overrightarrow{GA_1}\right)+\alpha_2\left(\overrightarrow{MG}+\overrightarrow{GA_2}\right)+\ldots+\alpha_n\left(\overrightarrow{MG}+\overrightarrow{GA_n}\right)\\&=(\alpha_1+\alpha_2+\ldots+\alpha_n)\overrightarrow{MG}+\alpha_1\overrightarrow{GA_1}+\alpha_2\overrightarrow{GA_2}+\ldots+\alpha_n\overrightarrow{GA_n}\\&=\left(\sum_{i=1}^{n}\alpha_i\right)\overrightarrow{MG}+\sum_{i=1}^{n}\alpha\overrightarrow{GA_i}\end{aligned}\tag{8.2.136}$$

As by definition of the barycenter we have:

$$\sum_{i=1}^{n}\alpha_i\overrightarrow{GA_i}=\vec{0}\tag{8.2.137}$$

Then we have indeed:

$$\sum_{i=1}^{n}\alpha_i\overrightarrow{MA_i}=\left(\sum_{i=1}^{n}\alpha_i\right)\overrightarrow{MG}+\sum_{i=1}^{n}\alpha_i\overrightarrow{GA_i}=\left(\sum_{i=1}^{n}\alpha_i\right)\overrightarrow{MG}\tag{8.2.138}$$

■ Q.E.D.

P2. For $\forall k\in\mathbb{R}^*$, the weighted points $(A_1,\alpha_1),(A_2,\alpha_2),\ldots,(A_n,\alpha_n)$ and $(A_1,k\alpha_1),(A_2,k\alpha_2),\ldots,(A_n,k\alpha_n)$ have same barycenter because (invariance of the barycenter):

$$\sum_{i=1}^{n}\alpha_i\overrightarrow{GA_i}=\vec{0}\Leftrightarrow\sum_{i=1}^{n}(k\alpha_i)\overrightarrow{GA_i}=\vec{0}\tag{8.2.139}$$

Proof 8.18.3. The proof is in our point of view immediate (but if you don't see it don't hesitate as always to contact us and we will write it).

■ Q.E.D.

P3. The barycenter G of n weighted points is invariant when we replace p of them, by their barycenter G', assigned to the condition that their coefficients are such that $\sum_{i=1}^{n}\alpha_i \neq 0$, then G is the barycenter of:

$$\left(G'_i, \sum_{i=1}^{p}\alpha_i\right), (A_{p+1}, \alpha_{p+1}), \ldots, (A_n, \alpha_n) \tag{8.2.140}$$

Proof 8.18.4. If G' is the barycenter of the weighted points $(A_1, \alpha_1), (A_2, \alpha_2), \ldots, (A_p, \alpha_p)$ then:

$$\sum_{i=1}^{p}\alpha_i\overrightarrow{MA_i} = \left(\sum_{i=1}^{p}\alpha_i\right)\overrightarrow{MG'} \tag{8.2.141}$$

For the particular case where $M = G$:

$$\sum_{i=1}^{p}\alpha_i\overrightarrow{GA_i} = \left(\sum_{i=1}^{p}\alpha_i\right)\overrightarrow{GG'} \tag{8.2.142}$$

But G being the barycenter of the n weighted points $(A_1, \alpha_1), (A_2, \alpha_2), \ldots, (A_n, \alpha_n)$ we have then:

$$\sum_{i=1}^{n}\alpha_i\overrightarrow{GA_i} = \vec{0} \Leftrightarrow \sum_{i=1}^{p}\alpha_i\overrightarrow{GA_i} + \sum_{i=p+1}^{n}\alpha_i\overrightarrow{GA_i} = \left(\sum_{i=1}^{p}\alpha_i\right)\overrightarrow{GG'} + \sum_{i=p+1}^{n}\alpha_i\overrightarrow{GA_i} = \vec{0} \tag{8.2.143}$$

As:

$$\sum_{i=1}^{p}\alpha_i + \sum_{i=p+1}^{n}\alpha_i \neq 0 \tag{8.2.144}$$

the previous equality proves that G is the barycenter of weighted points:

$$\left(G'_i, \sum_{i=1}^{p}\alpha_i\right), (A_{p+1}, \alpha_{p+1}), \ldots, (A_n, \alpha_n) \tag{8.2.145}$$

■ Q.E.D.

P4. If $\sum\alpha_i = 0$, for any point M, N:

$$\sum_{i=1}^{n}\alpha_i\overrightarrow{MA_i} = \sum_{i=1}^{n}\alpha_i\overrightarrow{NA_i} = c^{te} \tag{8.2.146}$$

Proof 8.18.5. For $M \neq N$ let us calculate:

$$\begin{aligned}
&\sum_{i=1}^{n}\alpha_i\overrightarrow{MA_i}-\sum_{i=1}^{n}\alpha_i\overrightarrow{NA_i}\\
&=\alpha_1\overrightarrow{MA_1}+\alpha_2\overrightarrow{MA_2}+\ldots+\alpha_n\overrightarrow{MA_n}-\alpha_1\overrightarrow{NA_1}-\alpha_2\overrightarrow{NA_2}-\alpha_2\overrightarrow{NA_2}-\ldots-\alpha_n\overrightarrow{NA_n}\\
&=\alpha_1\left(\overrightarrow{MA_1}-\overrightarrow{NA_1}\right)+\alpha_2\left(\overrightarrow{MA_2}-\overrightarrow{NA_2}\right)+\ldots+\alpha_n\left(\overrightarrow{MA_n}-\overrightarrow{NA_n}\right)\\
&=\alpha_1\overrightarrow{MN}+\alpha_2\overrightarrow{MN}+\ldots+\alpha_n\overrightarrow{MN}\\
&=(\alpha_1+\alpha_2+\ldots+\alpha)\overrightarrow{MN}=\left(\sum_{i=1}^{n}\alpha_i\right)\overrightarrow{MN}=\vec{0}
\end{aligned} \tag{8.2.147}$$

as $\sum_{i=1}^{n}\alpha_i=0$, we then have:

$$\sum_{i=1}^{n}\alpha_i\overrightarrow{MA_i}=\sum_{i=1}^{n}\alpha_i\overrightarrow{NA_i}=c^{te} \tag{8.2.148}$$

■ Q.E.D.

Remark

When a body has a certain symmetry, the calculations are simplified because the center of gravity (barycenter) should coincide with the element of symmetry. If a body such as a sphere, parallelepiped, etc., has a center of symmetry, the centroid is coincides with it. If the body only has one axis of symmetry, the centroid is then on this axis.

8.2.6 Geometric Transformations

The transformations in the plane (and spaces of more dimensions) are usually strictly defined using group theory (see section Set Algebra page 1417). But in the context of Euclidean geometry, this Set Theory approach does not interest us. So we will do in this subsection only a simple formal approach (and therefore more intuitive) of elementary transformations in the plane that are: translation, scaling, rotation and skew (shear). For the translation we will see also the equivalent in space as it is very important for our study of Physics (especially for the study of the gyroscope).

Remark

By definition an "isometry" is a transformation that preserves distances and areas (distance-preserving injective map between metric spaces). As we will see below, the translation, rotation and reflection are isometries. The scaling being obviously not an isometry in the plane or the space.

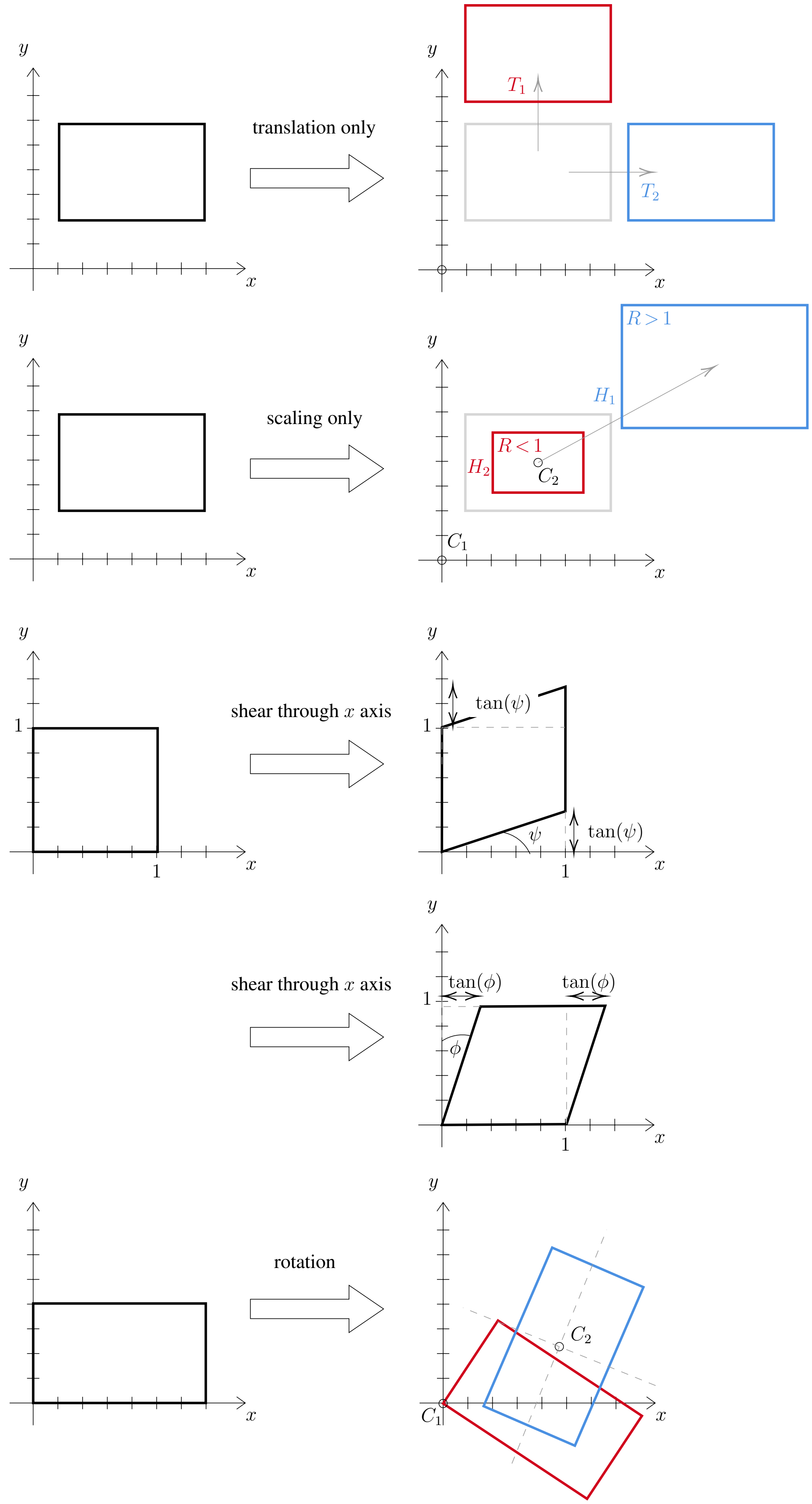

y
x
translation only
y
x
T_1
T_2
y
x
scaling only
y
x
$R>1$
H_1
$R<1$
H_2
C_2
C_1
y
1
1
x
shear through x axis
y
1
$\tan(\psi)$
ψ
$\tan(\psi)$
1
x
shear through x axis
y
1
$\tan(\phi)$
$\tan(\phi)$
ϕ
1
x
y
x
rotation
y
C_2
C_1
x

Figure 8.81: **Most common type of geometric transformations**

8.2.6.1 Translation

Given a straight line in a plane P on which two points A and B define a segment of line denoted $\overline{AB}$.

Definition (#8.2.341): A "translation T" ("displacement in a given direction" as said Euclid) of this line segment associates with each point A and B new points A' and B' such that $\overline{AB}=\overline{A'B'}$. So we can restrict the notion of translation to a point only if we want such that mathematically we can write:

$$T:P\rightarrow P\Rightarrow T(A)=A' \tag{8.2.149}$$

In other words, a transformation function (application) of the type transformation T of the of the entire plane in itself associates with each pre-image not more than a single image. Translation is therefore a bijective function. So we can define a unique reciprocal transformation application denoted as T^{-1} (recall of what was seen in the chapter Arithmetics):

$$T^{-1}(A')=A \tag{8.2.150}$$

We say by definition a point is "translationally invariant" if and only if:

$$T(A)=A'=A \tag{8.2.151}$$

In another type of formalism, the displacement of the point A to point B following the vector $\overrightarrow{AB}$ is named "translation vector $\overrightarrow{AB}$" equation (see section Vector Calculus page 1815). It is expressed mathematically by the sum of the point coordinates and the matrix of the coordinates of the vector (see just below the details).

For example in a three-dimensional space:

$$(x,y,z)+(x',y',z')=(x+x',y+y',z+z') \tag{8.2.152}$$

The translation is not a linear transformation (see section Linear Algebra page 1934), we can not allow ourselves to be represented by the multiplication of a square matrix as we will see it for following other transformations (scaling and transformation).

For this purpose (be able to express this transformation as square matrix linear application) we must use a workaround consisting in using a system named "homogeneous coordinates" (see section Projective Geometry page 2759) where the points of the plane are represented by a vector having 3 components (and respectively those of space by a vector with 4 components):

$$\begin{pmatrix}x\\y\end{pmatrix}\Rightarrow\begin{pmatrix}x_h\\y_h\\w\end{pmatrix}\qquad\text{with:}\qquad\begin{aligned}x_h&=x\cdot w\\y_h&=x\cdot w\end{aligned} \tag{8.2.153}$$

In the context of the study of translation we put $w=1$ because in this case:

$$\boxed{\begin{pmatrix}x'\\y'\\1\end{pmatrix}=\begin{bmatrix}1&0&T_x\\0&1&T_y\\0&0&1\end{bmatrix}\begin{pmatrix}x\\y\\1\end{pmatrix}=\begin{pmatrix}1\cdot x+0\cdot y+T_x\cdot 1\\0\cdot x+1\cdot y+T_y\cdot 1\\0+0+1\cdot 1\end{pmatrix}=\begin{pmatrix}x+T_x\\y+T_y\\1\end{pmatrix}} \tag{8.2.154}$$

This system of homogeneous coordinates is applicable to all other transformation we will see later by adding each time a coordinate (see section Projective Geometry page 2759).

Remark

Translation sends a line on a parallel line to the original line of course!

8.2.6.2 Homothety (scaling)

Given any shape in the plane or space (point, line, oval, polygon, etc.), a number R, and a point placed C at a predefined location.

Definition (#8.2.342): A "homothetic transformation H" (also named "scaling" or rarely "enlargement") of ratio R and center C is the application that at every point M of the shape associates to the segment $\overline{CM}$ a new point collinear to $\overline{CM}$ but arranged at a greater or smaller distance of ratio R from the center C such that $\overline{CM'} = R \cdot \overline{CM}$.

We can restrict the concept of scaling to a line segment such that we can write mathematically:

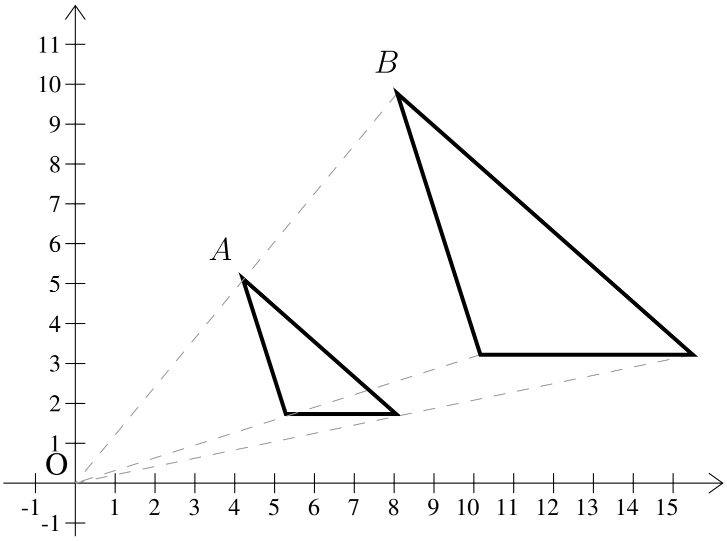
$$H : P \to P \Rightarrow H(\overline{CM}) = R \cdot \overline{CM} = \overline{CM'} \qquad (8.2.155)$$

In other words, a transformation of the type scaling in the entire plan in itself associates with each pre-image not more than a single image. Dilation is therefore as for the translation a bijective function. We can the define a reciprocal transformation application denoted H^{-1} such that as:

$$H^{-1}(R \cdot \overline{CM}) = \overline{CM} \qquad (8.2.156)$$

If $R \neq 1$, then C is the only invariant point. If $R = 1$, then all points are invariant and the scaling is say to be of the type "scaling identity". If $R = -1$, then we say that we have a "central symmetry". A central symmetry is then a rotation of $180°$ (or π [rad]).

In another type of formalism, a scaling of center O and ratio k, associate to the point A a point B such that $\overline{\text{O}B} = k\overline{\text{O}A}$. The point B being located on the straight line OA and at a distance $\overline{\text{O}B} = k\overline{\text{O}A}$. The sign of k determines the position of B with respect to O:



$$\overline{\text{O}B} = k\overline{\text{O}B} \Leftrightarrow (x', y', z') = k(x, y, z)$$

Figure 8.82: **Example of scaling of center O**

We now make a jump in the spatial geometry (the jump is not very large and require just the knowledge of elementary vector calculus and linear algebra).

We can also replace the scalar k by a square matrix such that:

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} ax+by+cz \\ dx+ey+fz \\ gx+hy+iz \end{pmatrix} \begin{pmatrix} kx \\ ky \\ kz \end{pmatrix} \tag{8.2.157}$$

A trivial solution for a scaling is to put that $b=c=d=f=g=h=0$ from which the matrix form obvious diagonal k:

$$k = \begin{bmatrix} k & 0 & 0 \\ 0 & k & 0 \\ 0 & 0 & k \end{bmatrix} \tag{8.2.158}$$

This matrix is named "transformation matrix by scaling of center O (in this case the origin of the reference frame) and ratio k" and therefore the scaling being a diagonal matrix commutes with any linear application.

Obviously for the two dimensional case we have immediately:

$$\boxed{\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{bmatrix} k & 0 \\ 0 & k \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{bmatrix} W & 0 \\ 0 & H \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} kx \\ ky \end{pmatrix}} \tag{8.2.159}$$

where W is for "width" and H is for "height".

Remark

It should be noticed that as many planar shapes have their surface that depends on the multiplication of two of their geometrical parameters, as scaling of factor k has a squared effect of factor k on the surface. The same principle applied for a volume but it's cubic amplitude rather than a squared one.

In the case presented above, the scaling conserves the forms in all directions (its geometry is invariant under this transformation) if we use indeed the same factor k for all axes (we speak then of "uniform scaling"). But we may also use the following matrix:

$$k = \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{bmatrix} \tag{8.2.160}$$

that would deforms the shape according to a different factor for each axis.

The inverse transformation of the scaling is obviously the scaling of center O and ratio k^{-1} thus in the form of a matrix:

$$k^{-1} = \begin{bmatrix} 1/k & 0 & 0 \\ 0 & 1/k & 0 \\ 0 & 0 & 1/k \end{bmatrix} \tag{8.2.161}$$

When the scaling center does not coincide with the origin of the selected coordinate system (which almost happens almost all the time), the procedure for calculating the coordinates of the image point is very simple. It is necessary to:

1. Carry out a translation to match the center of scaling with the origin of the frame and apply this translation to all points involved.
2. Make the scaling as described above (the center being the origin of the reference frame).
3. Make the reverse translation to take back the of center and the image in its original place.

To conclude, notice that the successive operations of translation and scaling are not commutative in the general case. In practice, we often performs first scaling and then translation as described in the three steps above.

8.2.6.3 Shear (skew) transformation

Definition (#8.2.343): In plane geometry, a "shear mapping" is a linear application that displaces each point in fixed direction, by an amount proportional to its signed distance from a line that is parallel to that direction.This type of mapping is also named "shear transformation", "transvection", or just "shearing".

Before formalize the general case of shear let us consider the following special case of a shearing factor k along the x axis:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} \begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x+2y \\ y \end{pmatrix} \tag{8.2.162}$$

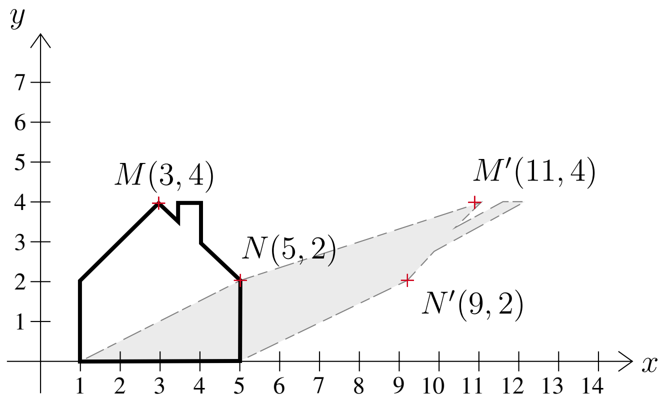


Figure 8.83: **Scaling with shear**

OK let us generalize that stuff! In fact for the shearing we must consider the case of shearing through the x-axis whose situation can be summarized by:

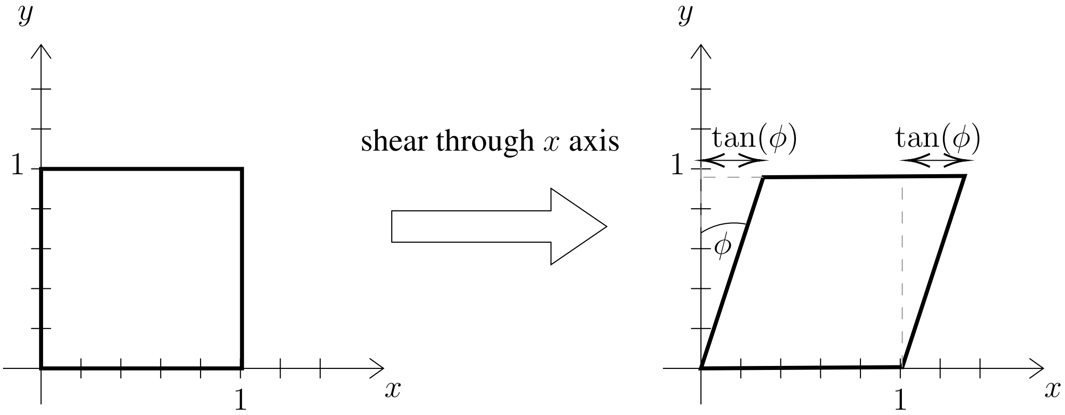

y
1
1
x
shear through x axis
y
$\tan(\phi)$
$\tan(\phi)$
1
ϕ
1
x

Figure 8.84: **Shearing trough x-axis**

So it is immediate that we have:

$$\boxed{\begin{pmatrix} x' \\ y' \end{pmatrix} \begin{bmatrix} 1 & \tan(\phi) \\ 0 & 1 \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x + y\tan(\phi) \\ y \end{pmatrix}} \tag{8.2.163}$$

and the case of shearing through the y-axis whose situation can be summarized by:

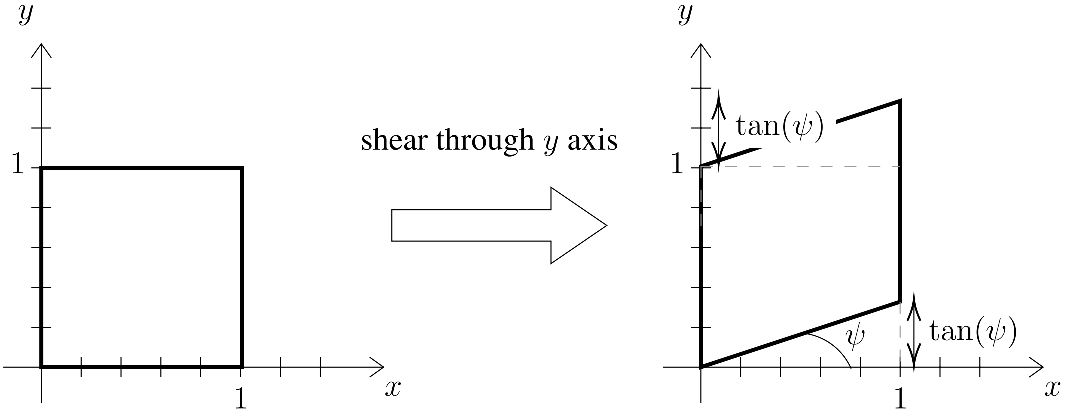

y
1
1
x
shear through y axis
y
$\tan(\psi)$
1
ψ
$\tan(\psi)$
1
x

Figure 8.85: **Shearing trough y-axis**

So it is immediate that we have:

$$\boxed{\begin{pmatrix} x' \\ y' \end{pmatrix} \begin{bmatrix} 1 & 0 \\ \tan(\psi) & 1 \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x(\psi) \\ x\tan(\psi) + y \end{pmatrix}} \tag{8.2.164}$$

8.2.6.4 Rotation

Given any some form in the plane (point, line, oval, polygon, etc.), a real number α, and a point C located at a predefined location.

Definition (#8.2.344): A "rotation R" of angle α and of center C is the application that to every point M of the shape associates to the segment $\overline{CM}$ a new point, but having had a positive or negative angle of rotation α of center C such that $\overline{CM}$ and $\overline{CM'}$ have the same length but not same direction.

From this definition, it is apparent that the axis of rotation of an object is the locus of points of the object which remain stationary!

Remark

The rotation is also, in a more academic way, a bijective mapping in the plane, we can therefore also define a reciprocal transformation application denoted R^{-1}. But the rotation in the space is not bijective so we can not define a unique reciprocal transformation.

If $\alpha \neq 0$, then C is the only invariant point. If $\alpha = 2k\pi$ (with $k \in \mathbb{Z}$), then all points are invariant and the rotation is say to be of the type "identity rotation". If we choose an appropriate system of perpendicular axes such that their intersection coincides with C and that $\pm\pi$ then R is named a "rotation of central symmetry".

In another type of formalism, the rotation is expressed much more rigorously. We will help us with the drawing of a unit circle (in the plane) to study this type of transformation. We will consider the first case the origin of the reference and translation coincides:

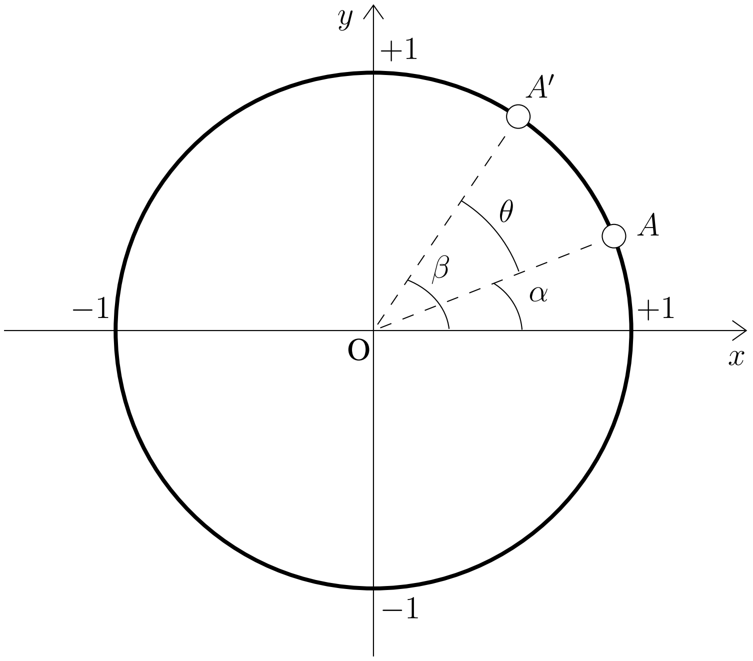


Figure 8.86: **Example of rotation in the plane**

where A' is the image of A by the rotation of center O and angle θ.

We have in the plane for the point A (see section Trigonometry page 2508):

$$x=r\cos(\alpha) \qquad y=r\sin(\alpha) \tag{8.2.165}$$

and identically for the point A':

$$x'=r\cos(\beta) \qquad y'=r\sin(\beta) \tag{8.2.166}$$

with $\beta=\alpha+\theta$.

Which brings us to write:

$$\dfrac{x}{\cos(\alpha)}=\dfrac{x'}{\cos(\alpha+\theta)}=r \tag{8.2.167}$$

Therefore using the Trigonometric identities (see section Trigonometry page 2519):

$$\begin{aligned}x'&=x\dfrac{\cos(\alpha+\theta)}{\cos(\alpha)}=x\dfrac{\cos(\alpha)\cos(\theta)-\sin(\alpha)\sin(\theta)}{\cos(\alpha)}=x\cos(\theta)-x\dfrac{\sin(\alpha)\sin(\theta)}{\cos(\alpha)}\\&=x\cos(\theta)-x\dfrac{y\sin(\theta)}{r\cos(\alpha)}=x\cos(\theta)-r\cos(\alpha)\dfrac{y\sin(\theta)}{r\cos(\alpha)}=x\cos(\theta)-y\sin(\theta)\end{aligned} \tag{8.2.168}$$

Identically (based on the fact that the basic trigonometric identities presented in the section Trigonometry of this textbook are known), we find:

$$y'=x\sin(\theta)+y\cos(\theta) \tag{8.2.169}$$

This allows us to write the rotation matrix in the plane (imagining that the z-axis out of the sheet/screen):

$$\begin{pmatrix}x'\\y'\end{pmatrix}=\begin{bmatrix}\cos(\theta) & -\sin(\theta)\\ \sin(\theta) & \cos(\theta)\end{bmatrix}\begin{pmatrix}x\\y\end{pmatrix}=\begin{bmatrix}x\cos(\theta) & -y\sin(\theta)\\ x\sin(\theta) & y\cos(\theta)\end{bmatrix} \tag{8.2.170}$$

Therefore the rotation matrix in the plane is given by:

$$\boxed{R=\begin{bmatrix}\cos(\theta) & -\sin(\theta)\\ \sin(\theta) & \cos(\theta)\end{bmatrix}} \tag{8.2.171}$$

The inverse transformation is simply the rotation of center O (same as before) but of angle $-\theta$ that is (we use obviously again the trigonometric relations of opposite angles):

$$R^{-1}=\begin{bmatrix}\cos(\theta) & \sin(\theta)\\ -\sin(\theta) & \cos(\theta)\end{bmatrix} \tag{8.2.172}$$

An easy way to see this is to notice that it is immediate that:

$$RR^{-1}=\begin{bmatrix}\cos(\theta) & -\sin(\theta)\\ \sin(\theta) & \cos(\theta)\end{bmatrix}\begin{bmatrix}\cos(\theta) & \sin(\theta)\\ -\sin(\theta) & \cos(\theta)\end{bmatrix}=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} \tag{8.2.173}$$

When we want to make a rotation of an object around its barycenter (centroid) we should proceed as for the scaling, that is to say: bring the barycenter of the shape back to origin O of the reference frame with a translation, apply the rotation on the shape and only after take back the shape to its original position with a translation.

During the rotation of an object in space (we take the opportunity to introduce this now... because we need it in several on physics concepts that will follow in this textbook), the transformation is quite similar to the previous one except that the product of two rotations does not give a rotation and that it is not commutative when the number of dimensions is strictly greater than 2 (we can make the detailed calculation on request).

Indeed, during a rotation of angle ϕ around the z-axis coordinate, the coordinate z does not change. Which brings us to write the rotation matrix in three-dimensional space relatively to the plane $x-y$ as:

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{bmatrix} \cos(\phi) & \sin(\phi) & 0 \\ -\sin(\phi) & \cos(\phi) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = R_Z(\phi) \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x\cos(\phi)+y\sin(\phi) \\ -x\sin(\phi)+y\cos(\phi)z \end{pmatrix} \tag{8.2.174}$$

For people reading this textbook on a computer with Adobe Flash player installed and activated, here is an animated version (otherwise see here: https://vimeo.com/575753732):

So therefore the rotation matrix application around the z axis of angle ϕ is:

$$\boxed{R_Z(\phi) = \begin{bmatrix} \cos(\phi) & \sin(\phi) & 0 \\ -\sin(\phi) & \cos(\phi) & 0 \\ 0 & 0 & 1 \end{bmatrix}} \tag{8.2.175}$$

So we can see that $R_Z(\phi)$ contains the rotation matrix in the plane $x-y$ that we have proved earlier above but with one more component according to the idea of homogeneous coordinates!!!

The philosophy is then always the same relative to other axes:

The three-dimensional rotation around the x-axis (relatively to the plane $z-y$) of angle θ:

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta) & \sin(\theta) \\ 0 & -\sin(\theta) & \cos(\theta) \end{bmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = R_X(\theta) \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x \\ y\cos(\theta)+z\sin(\theta) \\ -y\sin(\theta)+z\cos(\theta) \end{pmatrix} \tag{8.2.176}$$

For people reading this textbook on a computer with Adobe Flash player installed and activated, here is an animated version (otherwise see here: https://vimeo.com/575753223):

So therefore the rotation matrix application around the x axis of angle θ is :

$$\boxed{R_X(\theta)=\begin{bmatrix}1 & 0 & 0\\ 0 & \cos(\theta) & \sin(\theta)\\ 0 & -\sin(\theta) & \cos(\theta)\end{bmatrix}} \tag{8.2.177}$$

And finally the three-dimensional rotation around the y-axis (relatively to the plane $z-x$) of angle γ:

$$\begin{pmatrix}x'\\ y'\\ z'\end{pmatrix}=\begin{bmatrix}\cos(\gamma) & 0 & -\sin(\gamma)\\ 0 & 1 & 0\\ \sin(\gamma) & 0 & \cos(\gamma)\end{bmatrix}\begin{pmatrix}x\\ y\\ z\end{pmatrix}=R_Y(\gamma)\begin{pmatrix}x\\ y\\ z\end{pmatrix}=\begin{pmatrix}x\cos(\gamma)-z\sin(\gamma)\\ y\\ x\sin(\gamma)+z\cos(\gamma)\end{pmatrix} \tag{8.2.178}$$

OK for people reading this textbook on a computer with Adobe Flash player installed and activated, here is an animated version (otherwise see here: `https://vimeo.com/575753505`):

So therefore the rotation matrix application around the y axis of angle γ is:

$$\boxed{R_Y(\gamma)=\begin{bmatrix}\cos(\gamma) & 0 & -\sin(\gamma)\\ 0 & 1 & 0\\ \sin(\gamma) & 0 & \cos(\gamma)\end{bmatrix}} \tag{8.2.179}$$

We finally have three rotation matrices $R_X(\theta)$, $R_Y(\gamma)$, $R_Z(\phi)$ each corresponding to one of the planes (xy, yz, xz) of three-dimensional space. The three angles referred above are named "Euler angles" and sadly there is no convention in the notation of the angles, they depends on the authors and teachers.

These three matrices are part of the group of matrices of order three, denoted "SO (3)" and named by physicists and mathematicians "group of spatial rotations SO(3)" as seen in the section of Set Algebra at page 1425. Any rotation can be represented by the product matrix resulting from the product of these three (orthogonal) matrices (whose determinant is equal to 1 as proved in the section of Linear Algebra).

Remark

There is not only one eigenvector for one rotation matrix. There can be one but also a maximum of three. Indeed, consider a rotation of an object around the z-axis of π [rad] then obviously the z eigenvector $(0,0,1)$ with eigenvalues $-1,+1$ is obviously an eigenvector but also in this special case $(1,0,0)$ is an eigenvector but with an eigenvalue of -1.

Any rotation is a composition of these three rotations but it is important that the reader remembers the in the section of Linear Algebra where we saw that matrix multiplication is not commutative. Thus, rotation about the x-axis of $90°$ ($\pi/2$ [rad]) and the about the z-axis of $90°$ ($\pi/2$ [rad]) is not equivalent to rotate first around the z-axis and then around the x-axis by the same angle as shown in the figure below:

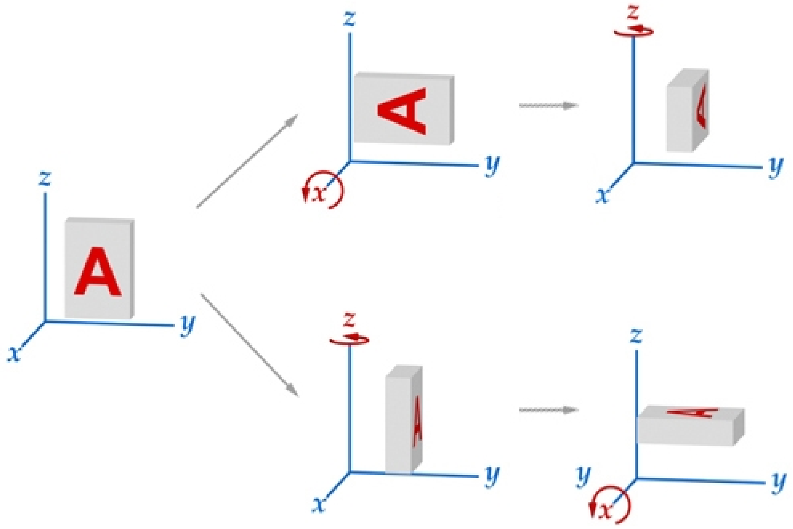


Figure 8.87: **Example of non-commutativity of the rotation matrix**

Finally, to get the matrix that makes up a particular case of the three rotations here for example are the commands to provide in Maple™ 4.00b:

```
>X:=array([[1,0,0],[0,cos(theta),sin(theta)],[0,-sin(theta),cos(theta)]]);
>Y:=array([[cos(lambda),0,-sin(lambda)],[0,1,0],[sin(lambda),0,cos(lambda)]]);
>Z:=array([[cos(phi),sin(phi),0],[-sin(phi),cos(phi),0],[0,0,1]]);
>evalm(X&*Y&*Z);
```

It is not very interesting to provide to our reader the expression of $R_{XYZ}=R_X R_Y R_Z$ as the rotations are not commutative and that therefore (excepted for some special cases) we have $R_{XYZ}\neq R_{XZY}\neq R_{YZX}\neq R_{YXZ}\neq R_{ZXY}\neq R_{ZYX}$. This is why most book don't give them (excepted some PDFs on Internet).

But as a reader requested it for R_{ZYX} let us give it:

$$\begin{aligned}R_{ZYX}&=\begin{bmatrix}\cos(\phi) & \sin(\phi) & 0\\ -\sin(\phi) & \cos(\phi) & 0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}\cos(\gamma) & 0 & -\sin(\gamma)\\ 0 & 1 & 0\\ \sin(\gamma) & 0 & \cos(\gamma)\end{bmatrix}\begin{bmatrix}1 & 0 & 0\\ 0 & \cos(\theta) & \sin(\theta)\\ 0 & -\sin(\theta) & \cos(\theta)\end{bmatrix}\\ &=\begin{bmatrix}\cos(\phi) & \sin(\phi) & 0\\ -\sin(\phi) & \cos(\phi) & 0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}\cos(\gamma) & \sin(\gamma)\sin(\theta) & -\sin(\gamma)\sin(\theta)\\ 0 & \cos(\theta) & \sin(\theta)\\ \sin(\gamma) & -\cos(\gamma)\sin(\gamma) & \cos(\gamma)\cos(\theta)\end{bmatrix}\\ &=\begin{bmatrix}\cos(\phi)\cos(\gamma) & \cos(\phi)\sin(\gamma)\sin(\theta)+\cos(\phi)\cos(\theta) & -\cos(\phi)\sin(\gamma)\sin(\theta)+\sin(\phi)\sin(\theta)\\ -\sin(\phi)\cos(\gamma) & -\sin(\phi)\sin(\gamma)\sin(\theta)+\cos(\phi)\cos(\theta) & \sin(\phi)\sin(\gamma)\sin(\theta)+\cos(\phi)\sin(\theta)\\ \sin(\gamma) & -\cos(\gamma)\sin(\gamma) & \cos(\gamma)\cos(\theta)\end{bmatrix}\end{aligned} \tag{8.2.180}$$

If we are looking to achieve the composition of a translation T and a rotation R and a scale scaling H the transformation matrix will be typically:

$$[R][H][T] \tag{8.2.181}$$

Remarks

R1. Let us recall (see section Linear Algebra page 1947) that the multiplication of two matrices is generally not commutative.

R2. The direct similarity of center C, of ratio R and angle α is the composite of the scaling of center C with ratio R and a rotation of center C and angle α. We refer the reader to the section Numbers to review that complex numbers (see page 243) allow to operate formally to similarities with the operations of addition and multiplication (direct or retrograde).

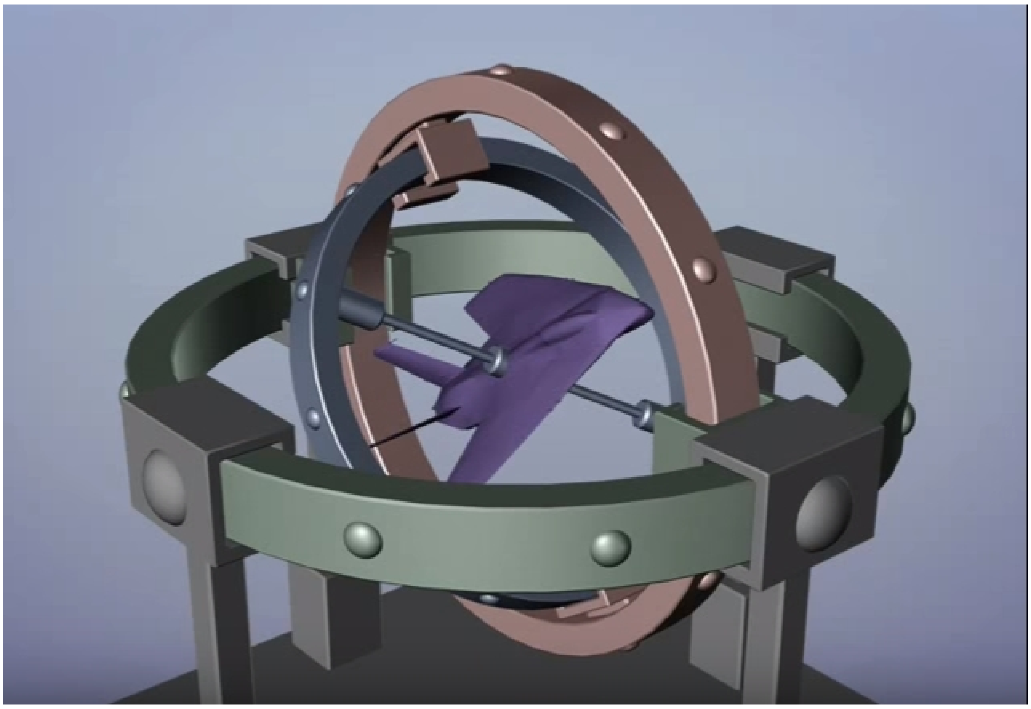
R3. We can make much more powerful and variable rotations using quaternions numbers (or "hypercomplex numbers"). For more information about this type of numbers the reader should refer to the section Numbers where the quaternions (see page 261) are introduced with a lot of details that makes more clear why they are preferred sometimes in 3D computing for rotations instead of the rotation matrices.

8.2.6.4.1 Gimbal lock

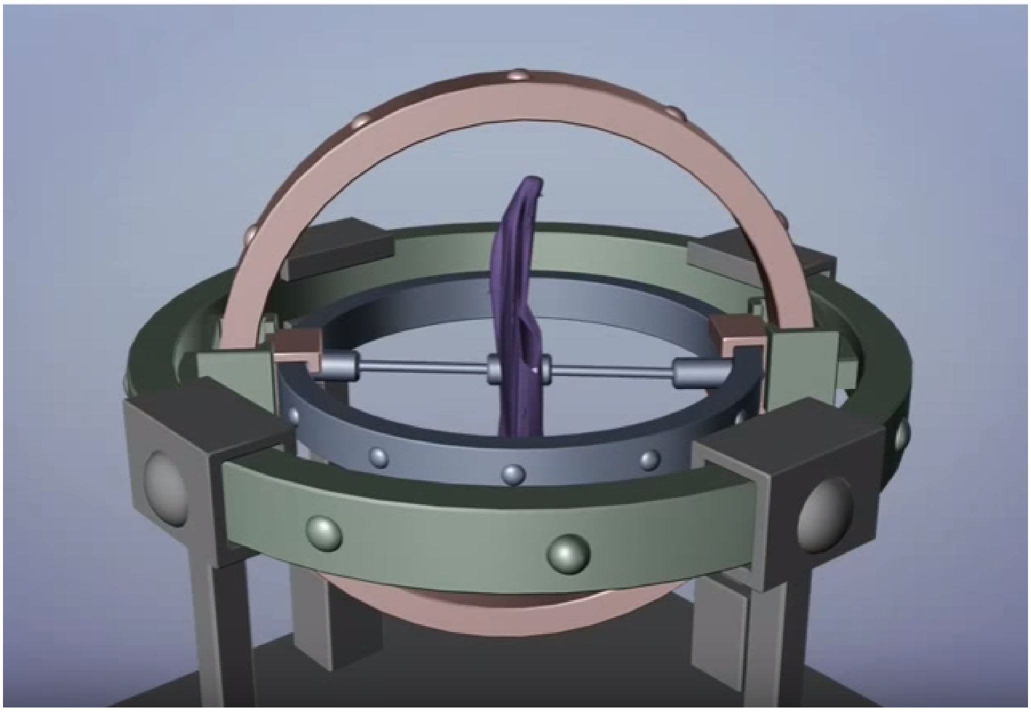
Gimbal lock is the loss of one degree of freedom in a three-dimensional, three-gimbal mechanism that occurs when the axes of two of the three gimbals are driven into a parallel configuration, "locking" the system into rotation in a degenerate two-dimensional space.

The problem of gimbal lock appears when one uses Euler angles (see further below the definition) in applied mathematics. Developers of 3D computer programs, such as 3D modelling, embedded navigation systems, robotics and video games must take care to avoid it.

Before going in the math stuff let us show an example of gimbal lock. Consider the following gyroscope like structure:



In the above configuration we can rotate the Jet in any angle (that means that a jet will know its position wherever it go) Indeed, the inner disc can rotate uniquely, same for the inner and outer disc. But now if the following configuration happens to occur:



The reader can then see that the inner disc then can only rotate around the same axis as the outer disc. So... yes!!! We lost one degree of freedom.

A lot of confusion results sometimes from the term "lock" in this issue. Indeed, we see from the above figure the gimbals don't actually "lock". They're all still free to spin about like they normally would, no ring is being held up or forcibly locked into position. It's only that if two of the three axis are aligned we can't say anymore the difference between rotating about one axis or another one. We can't because they're effectively the same axis in this special position. Both gimbals would rotate in the same direction at the same time.

So in certain orientations a way, pitch or roll axis aligns with one of the others. A that point, a change in the first can't be distinguished from a change in the other. That's a big problem if we need to keep track of our orientation accurately.

Remark

It is said that well-known gimbal lock incident happened in the Apollo 11 Moon mission.

In formal language, gimbal lock occurs because the map from Euler angles to rotations. Indeed, Euler angles provide a means for giving a numerical description of any rotation in three-dimensional space using three numbers, but not only is this description not unique, but there are some points where not every change in the target space (rotations) can be realized by a change in the source space (Euler angles). This is a topological constraint - there is no covering map from the 3-torus to the 3-dimensional real projective space; the only (non-trivial) covering map is from the 3-sphere, as in the use of quaternions.

A rotation in 3D space can be represented numerically with matrices in several ways. One of these representations is as we have just seen it:

$$R=R_ZR_XR_Y=\begin{bmatrix}\cos(\phi) & \sin(\phi) & 0\\ -\sin(\phi) & \cos(\phi) & 0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}1 & 0 & 0\\ 0 & \cos(\theta) & \sin(\theta)\\ 0 & -\sin(\theta) & \cos(\theta)\end{bmatrix}\begin{bmatrix}\cos(\gamma) & 0 & -\sin(\gamma)\\ 0 & 1 & 0\\ \sin(\gamma) & 0 & \cos(\gamma)\end{bmatrix}\tag{8.2.182}$$

Let us observe what happens if $\theta=0$ (that is to say we block the rotation around the x-axis):

$$R=R_ZR_XR_Y=\begin{bmatrix}\cos(\phi) & \sin(\phi) & 0\\ -\sin(\phi) & \cos(\phi) & 0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}\cos(\gamma) & 0 & -\sin(\gamma)\\ 0 & 1 & 0\\ \sin(\gamma) & 0 & \cos(\gamma)\end{bmatrix}\tag{8.2.183}$$

Carrying out matrix multiplication:

$$\begin{aligned}R=R_ZR_XR_Y&=\begin{bmatrix}0 & 0 & -1\\ \sin(\phi) & \cos(\phi) & 0\\ \cos(\phi) & -\sin(\phi) & 0\end{bmatrix}\begin{bmatrix}\cos(\gamma) & \sin(\gamma) & 0\\ -\sin(\gamma) & \cos(\gamma) & 0\\ 0 & 0 & 1\end{bmatrix}\\ &=\begin{bmatrix}0 & 0 & 1\\ \sin(\phi)\cos(\gamma)-\cos(\phi)\sin(\gamma) & \sin(\phi)\sin(\gamma)+\cos(\phi)\cos(\gamma) & 0\\ \cos(\phi)\cos(\gamma)+\sin(\phi)\sin(\gamma) & \cos(\phi)\sin(\gamma)-\sin(\phi)\cos(\gamma) & 0\end{bmatrix}\end{aligned}\tag{8.2.184}$$

And finally using some of our trigonometry identities (see section Trigonometry page 2519):

$$R=\begin{bmatrix}0 & 0 & 1\\ \sin(\phi-\gamma) & \cos(\phi-\gamma) & 0\\ \cos(\phi-\gamma) & \sin(\gamma-\phi) & 0\end{bmatrix}=\begin{bmatrix}0 & 0 & 1\\ \sin(\phi-\gamma) & \cos(\phi-\gamma) & 0\\ \cos(\phi-\gamma) & -\sin(\phi-\gamma) & 0\end{bmatrix}\tag{8.2.185}$$

Changing the values of ϕ and γ in the above matrix has the same effects: the rotation angle $\phi-\gamma$ changes, but the rotation axis remains in the z direction! Indeed, we know that in a rotation matrix that when the first row and first column are fixed then we have a rotation around the x-axis, when the second row and second column are fixed then we have a rotation around the y-axis, when we the third a row and third column are fixed then we have a rotation around the z axis. So when the third column is fixed but instead of the third row it is the first row that is fixed than the rotation the last column and the first row in the matrix won't change. The only solution for ϕ and γ to recover different roles is to change θ.

In this specific situation, the gimbal lock behaviour can NOT be avoided. Quaternions (see page 261) can uniquely identify a rotation, but when they are converted into Euler rotations, they lose also one degree of freedom.

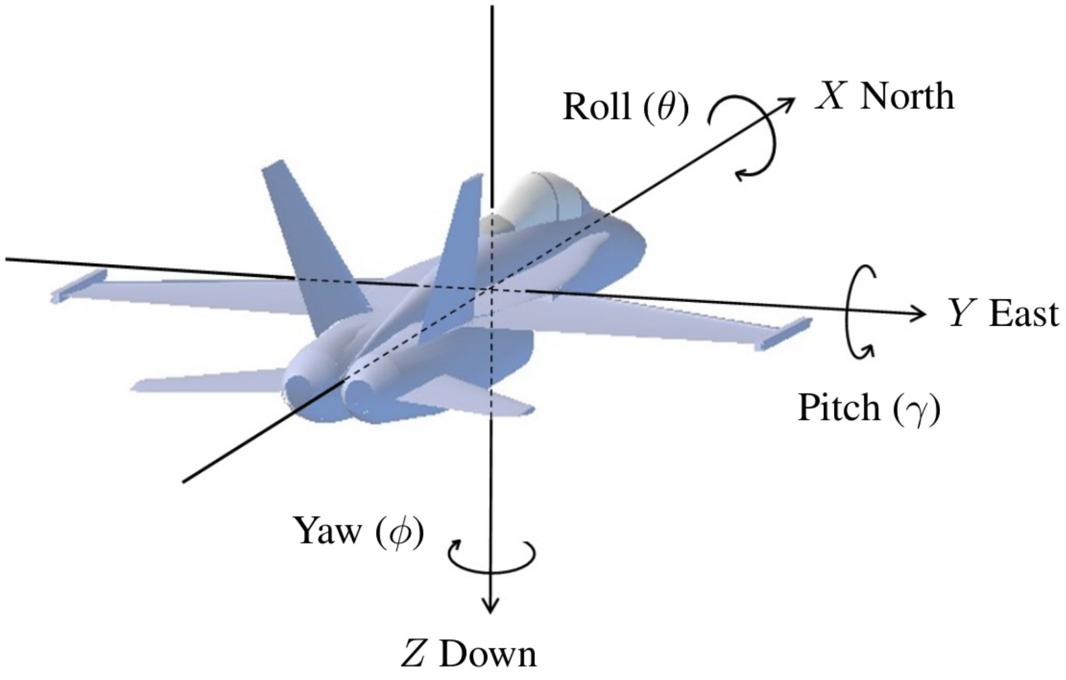


Figure 8.88: Vocabulary of rotation in aeronautical science

8.2.6.4.2 Euler angles

Let us come back on the Euler angles for an important result for our study late of the gyroscope effect in Classical Mechanics (especially the "nutation" effect)!

So we characterize a general orientation of a body with respect to the inertial system xyz (see figure below) in term of the following 3 rotations respectively by tradition the following order:

1. Rotation by angle ϕ around the z-axis
2. Rotation by angle θ around the new x_1' axis, which we name the "line of nodes".
3. Rotation by angle γ about the new x_3-axis

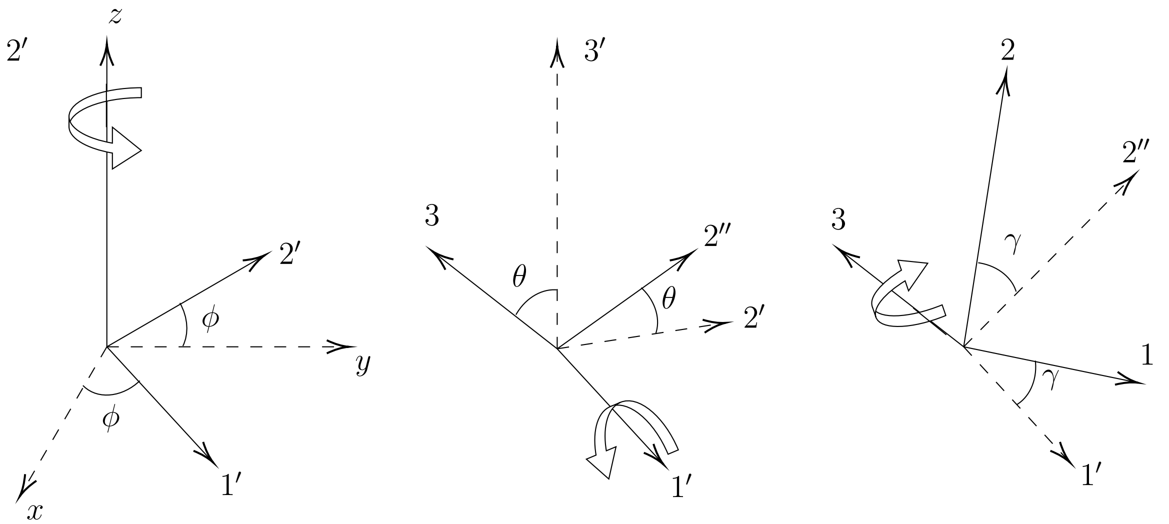


Figure 8.89: **Euler angles rotation convention**

Formally as we know this is given by:

$$R_{YXZ}=\begin{bmatrix}\cos(\gamma) & 0 & -\sin(\gamma)\\ 0 & 1 & 0\\ \sin(\gamma) & 0 & \cos(\gamma)\end{bmatrix}\begin{bmatrix}1 & 0 & 0\\ 0 & \cos(\theta) & \sin(\theta)\\ 0 & -\sin(\theta) & \cos(\theta)\end{bmatrix}\begin{bmatrix}\cos(\phi) & \sin(\phi) & 0\\ -\sin(\phi) & \cos(\phi) & 0\\ 0 & 0 & 1\end{bmatrix} \tag{8.2.186}$$

Fortunately, we most of time not have to use the resulting matrix of the three matrix multiplication above. In most kinematic problems what we really need is to be able to write the components of the angular velocity $\vec{\Omega}$ in both systems of coordinates. Since $\vec{\Omega}$ describes precisely how fast the angles vary in time, we have:

$$\boxed{\vec{\Omega}=\dot{\vec{\phi}}+\dot{\vec{\theta}}+\dot{\vec{\gamma}}=\frac{\mathrm{d}\vec{\phi}}{\mathrm{d}t}+\frac{\mathrm{d}\vec{\theta}}{\mathrm{d}t}+\frac{\mathrm{d}\vec{\gamma}}{\mathrm{d}t}=\dot{\phi}\vec{e}_z+\dot{\theta}\vec{e}_{1'}+\dot{\gamma}\vec{e}_3} \tag{8.2.187}$$

since the three rotations are about these particular axes.

Let us analyze each contribution to $\vec{\Omega}$:

1. $\dot{\vec{\phi}}=\vec{e}_z\dot{\phi}$ (with respect to xyz system). Following the rotations, we find that with respect to 123 system we have first (the reader can take the extreme situations where $\theta=0$ and $\theta=\pi/2$ to see that this is accurate):

   $$\vec{e}_z=\cos(\theta)\vec{e}_3+\sin(\theta)\vec{e}_{2''} \tag{8.2.188}$$

   but we notice that we also have with the same idea:

   $$\vec{e}_{2''}=\sin(\gamma)\vec{e}_1+\cos(\gamma)\vec{e}_2 \tag{8.2.189}$$

   Therefore we have:

   $$\vec{e}_z=\cos(\theta)\vec{e}_3+\sin(\theta)\vec{e}_{2''}=\cos(\theta)\vec{e}_3+\sin(\theta)\left(\sin(\gamma)\vec{e}_1+\cos(\gamma)\vec{e}_2\right) \tag{8.2.190}$$

   Hence:

   $$\dot{\vec{\phi}}=\dot{\phi}\vec{e}_z=\sin(\theta)\sin(\gamma)\dot{\phi}\vec{e}_1+\sin(\theta)\cos(\gamma)\dot{\phi}\vec{e}_2+\cos(\theta)\dot{\phi}\vec{e}_3 \tag{8.2.191}$$

   with respect to the 123 system.

2. $\dot{\vec{\theta}}=\vec{e}_{1'}\dot{\theta}$ (with respect to $1'2'3'$ system). Following the rotations, we find that with respect to 123 system we have obviously (the reader can take the extreme situations where $\gamma=0$ and $\gamma=\pi/2$ to see that this is accurate):

   $$\vec{e}_{1'}=\cos(\gamma)\vec{e}_1-\sin(\gamma)\vec{e}_2 \tag{8.2.192}$$

   Therefore:

   $$\dot{\vec{\theta}}=\vec{e}_{1'}\dot{\theta}=\left(\cos(\gamma)\vec{e}_1-\sin(\gamma)\vec{e}_2\right)\dot{\theta}=\cos(\gamma)\dot{\theta}\vec{e}_1-\sin(\gamma)\dot{\theta}\vec{e}_2 \tag{8.2.193}$$

   with respect to 123 system.

3. $\dot{\vec{\gamma}}=\vec{e}_3\dot{\gamma}$. Following the rotations, we have then immediately with respect to 123:

   $$\dot{\vec{\gamma}}=\vec{e}_3\dot{\gamma} \tag{8.2.194}$$

Therefore summing the relations:

$$\begin{aligned}\dot{\vec{\phi}}&=\sin(\theta)\sin(\gamma)\dot{\phi}\vec{e}_1+\sin(\theta)\cos(\gamma)\dot{\phi}\vec{e}_2+\cos(\theta)\dot{\phi}\vec{e}_3\\ \dot{\vec{\theta}}&=\cos(\gamma)\dot{\theta}\vec{e}_1-\sin(\gamma)\dot{\theta}\vec{e}_2\\ \dot{\vec{\gamma}}&=\vec{e}_3\dot{\gamma}\end{aligned} \tag{8.2.195}$$

Therefore the sum finely regrouped gives:

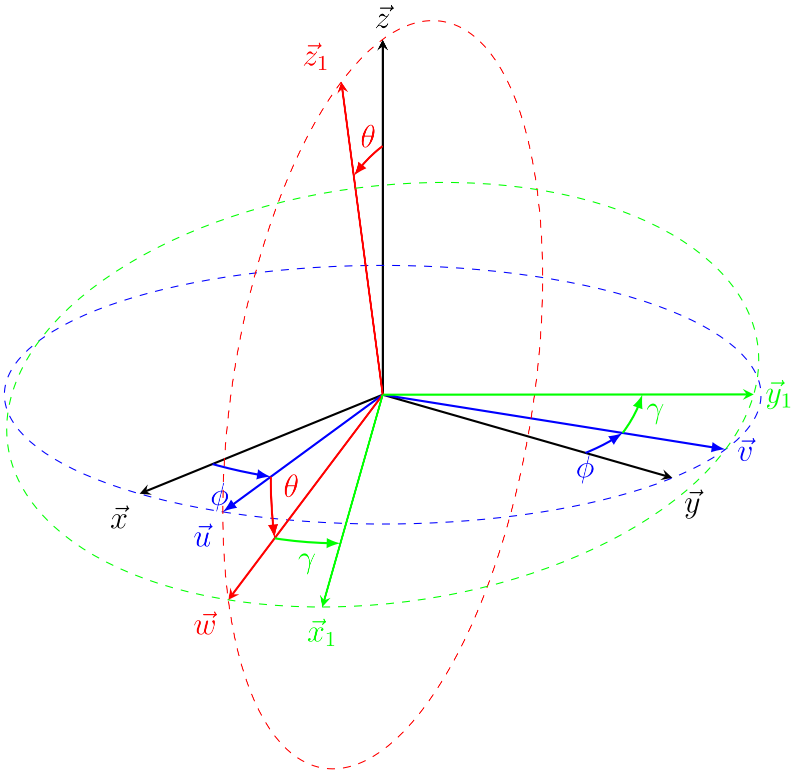
$$\vec{\Omega}=\left(\sin(\theta)\sin(\gamma)\dot{\phi}+\cos(\gamma)\dot{\theta}\right)\vec{e}_1+\left(\sin(\theta)\cos(\gamma)\dot{\phi}-\sin(\gamma)\dot{\theta}\right)\vec{e}_2+\left(\cos(\theta)\dot{\phi}+\dot{\gamma}\right)\vec{e}_3 \tag{8.2.196}$$

with respect to 123 system. Many times denoted:

The system:

$$\boxed{\begin{aligned}\Omega_1 &= \sin(\theta)\sin(\gamma)\dot{\phi}+\cos(\gamma)\dot{\theta}\\ \Omega_2 &= \sin(\theta)\cos(\gamma)\dot{\phi}-\sin(\gamma)\dot{\theta}\\ \Omega_3 &= \cos(\theta)\dot{\phi}+\dot{\gamma}\end{aligned}} \tag{8.2.197}$$

is named "kinematic Euler equations" and is very useful to express the rotation speed of a body in its own reference frame (most of time its center of mass).



$\vec{z}$
$\vec{z}_1$
θ
$\vec{y}_1$
γ
$\vec{v}$
ϕ
$\vec{y}$
$\vec{x}$
ϕ
θ
$\vec{u}$
γ
$\vec{w}$
$\vec{x}_1$

Figure 8.90: Another common representation of Euler angles (author: Dorian Depriester)

8.2.6.5 Reflection

Definition (#8.2.345): The "reflection", also named "axial symmetry", denoted S_Δ (in geometry) relatively to the straight line Δ (in space the reflection is done relatively to a plane) is the application that associates to each point M outside Δ the point M' such that Δ is the mediator of $\overline{MM'}$. If M belongs to Δ, then $M = M'$.

Mathematically it is written:

$$S_\Delta : P \rightarrow P \Rightarrow S_\Delta(M) = M' \tag{8.2.198}$$

In other words, a transformation function of the type reflection of the entire plan in itself associates with each pre-image not more than a single image. Reflection is then a bijective function. So we can define a reciprocal transformation denoted S_Δ^{-1} such that:

$$S_\Delta^{-1}(M') = M \tag{8.2.199}$$

Remark

All points of Δ are trivially invariant under reflection in the plane.

The most obvious and well known reflection axes are those included into the shape itself as depicted with the examples visible in the figure below:

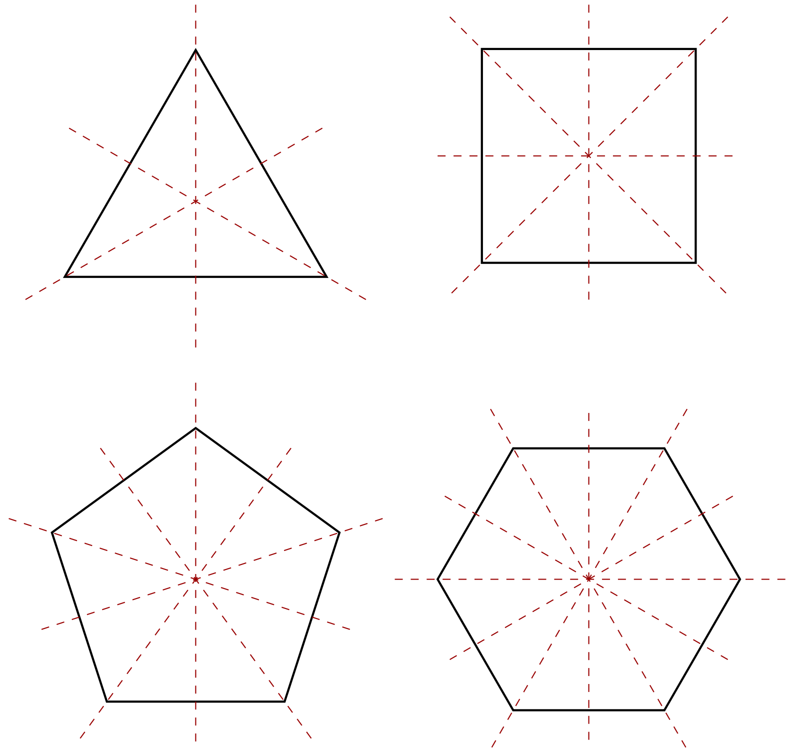


Figure 8.91: **Reflection axes (author: Lauren K. Williams)**

In matrix form the reflections in the plane are extremely simple to formalize using linear algebra (see section of the same name page 1924) as shown in the situations below:

- Reflection relatively to the y-axis:

$$S_{\Delta_y}=\begin{pmatrix}x'\\y'\end{pmatrix}=\begin{bmatrix}-1&0\\0&1\end{bmatrix}\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}-x\\y\end{pmatrix}\tag{8.2.200}$$

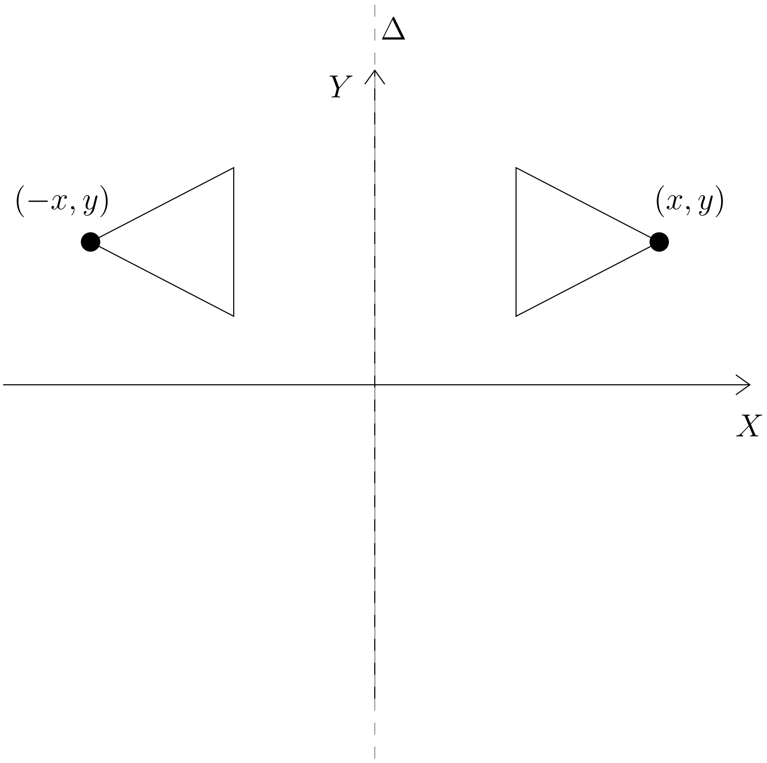


Figure 8.92: **Reflection around y-axis**

- Reflection relatively to the x-axis:

$$S_{\Delta_x}=\begin{pmatrix}x'\\y'\end{pmatrix}=\begin{bmatrix}1&0\\0&-1\end{bmatrix}\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}x\\-y\end{pmatrix}\tag{8.2.201}$$

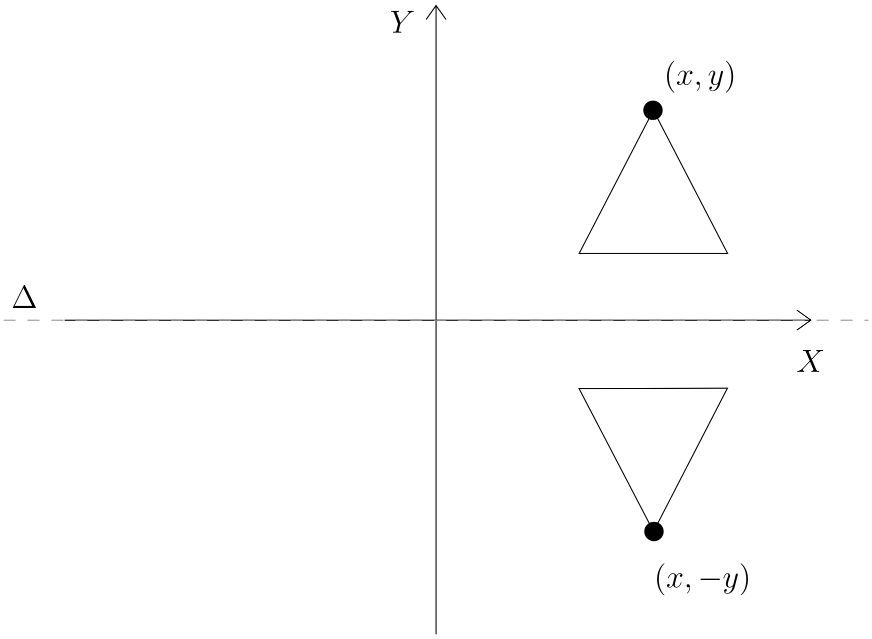


Figure 8.93: **Reflection around x-axis**

- Reflection relatively to the origin O:

$$S_{\Delta}=\begin{pmatrix}x'\\y'\end{pmatrix}=\begin{bmatrix}-1&0\\0&-1\end{bmatrix}\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}-x\\-y\end{pmatrix}\tag{8.2.202}$$

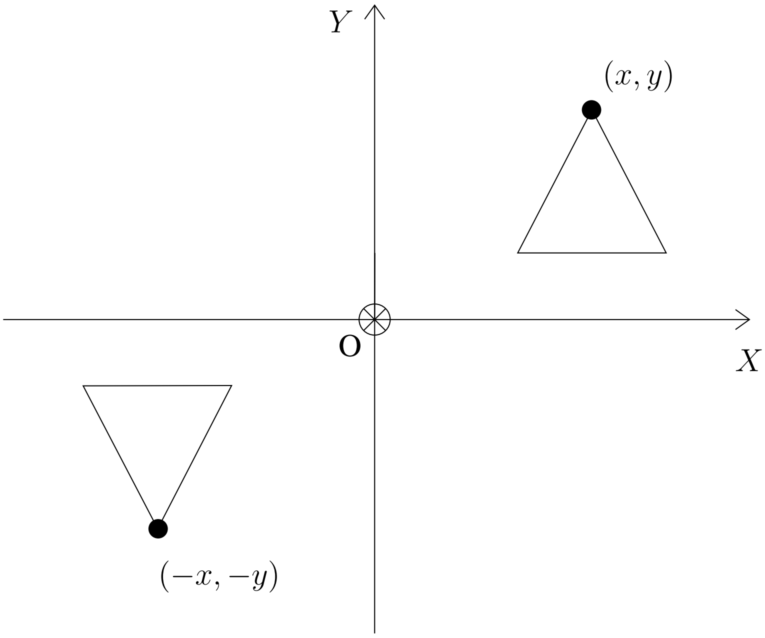
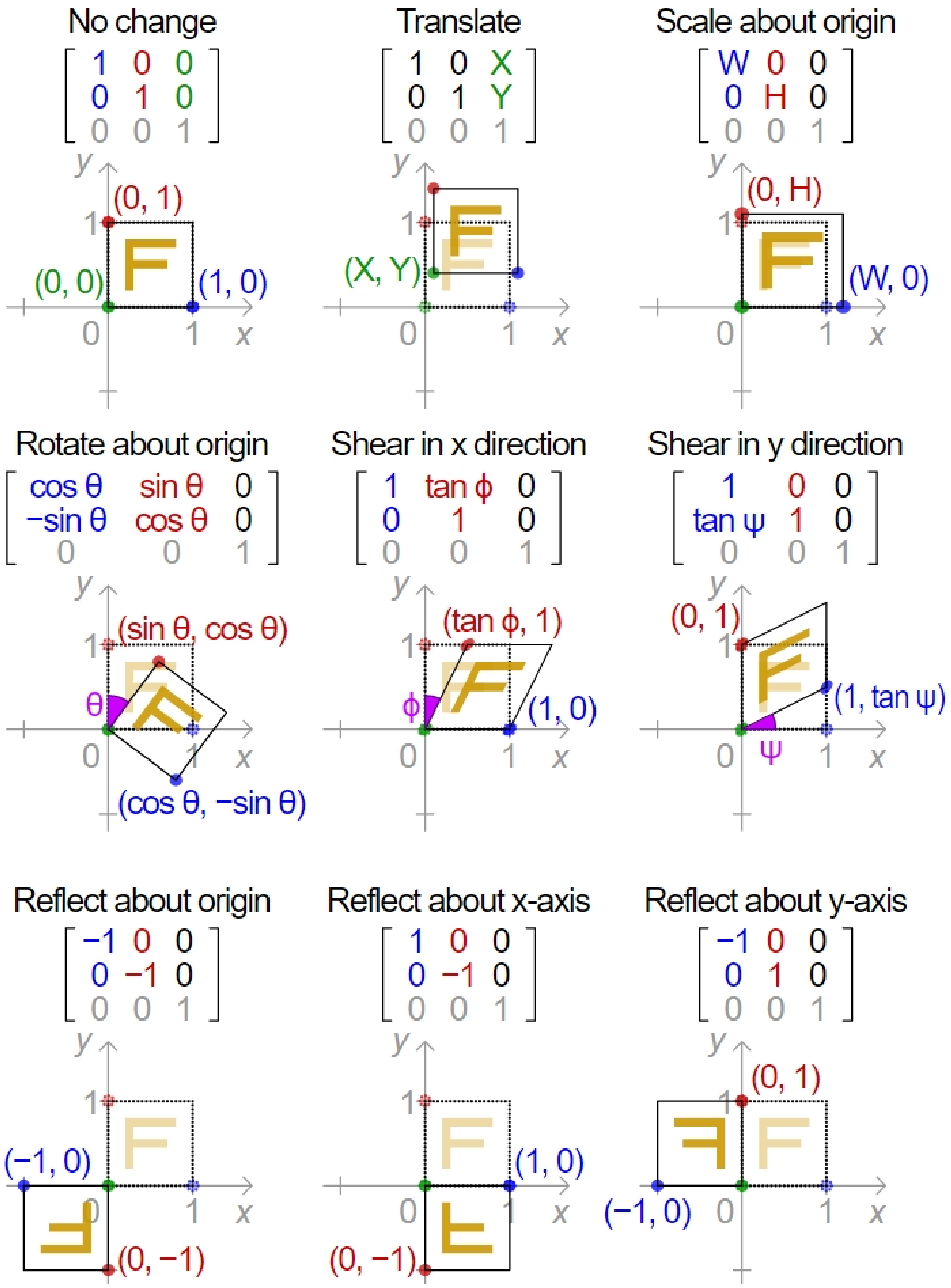


Figure 8.94: **Reflection around origin O**

The reader has very likely noticed that every reflection matrix is in fact only the special case of the plane rotation matrix and this was not difficulty to guess as we can see on every figure that a reflection is only a rotation around a special center of rotation.

Here is a quite good summary of what we have seen so far:



No change

[1 0 0; 0 1 0; 0 0 1]
y
(0, 1)
1
(0, 0)
(1, 0)
0
1
x

Translate

[1 0 X; 0 1 Y; 0 0 1]
y
1
(X, Y)
0
1
x

Scale about origin

[W 0 0; 0 H 0; 0 0 1]
y
(0, H)
1
(W, 0)
0
1
x

Rotate about origin

[cos θ sin θ 0; −sin θ cos θ 0; 0 0 1]
y
(sin θ, cos θ)
1
θ
0
1
x
(cos θ, −sin θ)

Shear in x direction

[1 tan ϕ 0; 0 1 0; 0 0 1]
y
(tan ϕ, 1)
1
ϕ
(1, 0)
0
1
x

Shear in y direction

[1 0 0; tan ψ 1 0; 0 0 1]
y
(0, 1)
1
(1, tan ψ)
0
ψ
1
x

Reflect about origin

[−1 0 0; 0 −1 0; 0 0 1]
y
1
(−1, 0)
0
1
x
(0, −1)

Reflect about x-axis

[1 0 0; 0 −1 0; 0 0 1]
y
1
(1, 0)
0
1
x
(0, −1)

Reflect about y-axis

[−1 0 0; 0 1 0; 0 0 1]
y
(0, 1)
1
(−1, 0)
0
1
x

Figure 8.95: **Summary of affine transformations (source: Wikipedia)**

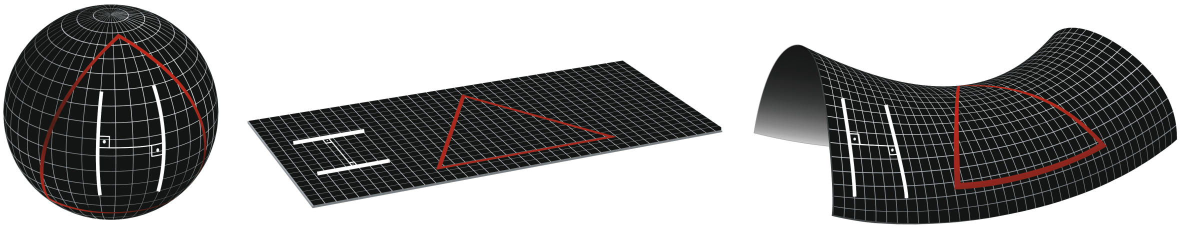
~70 %

★★★★★

31 votes, 71.61%

8.3. Non-Euclidean Geometry

ONE may recall from their lecture the section of Euclidean Geometry (see page 2563) that the sum of the angles in a triangle is 180 degrees. This comes from the Euclidean or "flat" geometry, which includes something named the "parallel postulate" (i.e. the 5th postulate) which states that if you were to draw two points next to one another, then extend from those points two lines that are parallel to one another, those lines could be extended to infinity without ever becoming closer together or further apart! This makes perfect intuitive sense; Euclidean geometry seems to be the structure of our world. However, our senses often deceive us. It turns out that gravity literally warps the geometry of space-time (for a naive and technical definition of "space-time" see the section of Special Relativity at page 5708) itself as we will see it in the section of General Relativity (see page 5720). That's right, because of the gravitational field, space-time is non-Euclidean (and there is some amount of gravity everywhere, since it is an interaction with infinite range). If not Euclidean, what can this geometry even be? As illustrated below, geometry on curved surfaces is a little different from geometry on flat (Euclidean) surfaces:



Positive curvature
(spherical like/elliptic)

Flat (zero) curvature
(plane like/euclidian)

Negative curvature
(horse's saddle like/hyperbolic)

Figure 8.96: **Positive (elliptic), negative (hyperbolic) and flat (euclidean) curves (source: NASA / WMAP Science Team)**

In a negatively curved ("hyperbolic") geometry, the sum of the angles in a triangle add up to less than 180 degrees (π radian), and parallel lines diverge from one another. In a positively curved geometry (also known as an "elliptic" geometry), the sum of the angles in a triangle add up to greater than 180 degrees (π radian) as proved in the section of Trigonometry, and parallel lines converge towards one another. A "flat" geometry, the Euclidean case, has zero curvature. It is a special case that is the exact point in between hyperbolic and elliptic geometries and in such an euclidean geometry the sum of the angles of a triangle is always equal to 180 degrees (π radian) as proven at page 2593. Another illustration below shows the failure of the parallel postulate in non-Euclidean spaces.

Therefore we guess obviously that non-Euclidean geometries are all geometries that not necessarily satisfy all the Hilbert's axioms (see section Euclidean Geometry page 2631) but without any contradiction between them (unlike the old axioms of Euclid, particularly that on parallels).

A particular representation of this type of geometry consists to define the points as being distributed over the surface of a sphere (that are the intersection of the diameters of the sphere with the surface), and the lines, to generalize the concept of straight line (now we say "geodesic"), as the intersections of the surface of the sphere with the planes containing the center of the sphere.

Two points then define uniquely a line and a point is always given by the intersection of two lines. However, in this geometry, if we give ourselves a line AB and a external point P to the line, there is no line passing through P and not intersecting AB. So Euclid's fifth postulate is not satisfied because in P we can not draw any parallel to AB.

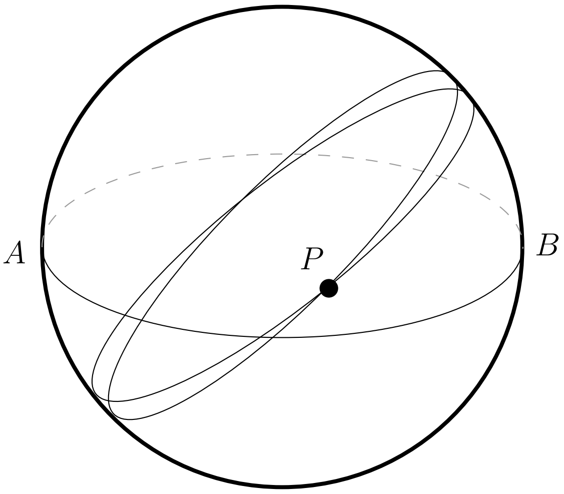


Figure 8.97: **Illustrated example of the violation of the fifth Euclide postulate**

Remark

Before starting this section, we strongly recommend the reader to consult and if possible to understand the content of the section on Trigonometry, Euclidean Geometry and Tensor Calculus as we will use many results, not necessarily trivial, we have proved in them.

In section on Euclidean Geometry, we studied a number of theorems related to plans. Let us insist on the fact that the "plane" is a two-dimensional figure whose curvature is null and immersed in a 3-dimensional space (therefore the plane can be moved in it). This said, it should perhaps now necessary to define a little bit more rigorously the intuitive concept of "curvature(s)".

8.3.1 Curvature(s)

Definition (#8.3.346): A figure is said to be "curved" if there is at least one point lying on the straight or straights defining the bounds of the figure (surfaces, volumes, etc.) defining a tangent on the bound and not crossing the bound elsewhere.

This is Friedrich Gauss who in 11824 (holocene calendar) had formulated the possibility that there are alternatives to Euclidean geometries. We distinguish between geometries with "negative curvature", like the one of the Russian Nikolai Lobachevsky (11829 according to holocene calendar) and Bolyai (11832 according to holocene calendar) (sum of angles of a triangle below π, infinite number of possible parallel to a line through a point), from geometries with "positive curvature" as the Riemann (11867 according to holocene calendar) one (sum of the angles of a triangle greater than π, parallel joining to the poles).

We will see in this section various non-Euclidean geometries, the best known are the "Riemannian geometries" (constant curvature) and "Lobachevsky geometries" (hyperbolic type with non-constant curvature).

Remark

The geometry commonly named "Riemann geometry" is a three-dimensional spherical space, finite space but boundless, with regular curvature, alternative to the Euclidean postulate on parallels.

The interest in the study of these geometries is that we can't determine whether the Universe we live in is made of a type of geometry rather than another because given our (physical) size, immersed we are in any geometry whatsoever with low curvature, any surface space seems locally Euclidean to us (two parallel lines do not intersect). However, in General Relativity, which makes an excessive us of tensor calculus (generalization of any geometry as seen in the section on Tensor Calculus at page 2038) shows that there are region of the space where the geometry is strongly curved and therefore locally non-Euclidean and only the study of this kind of geometries allows us to draw theories explaining observations that are not exploitable only with human intuition.

A distinction has also to be made between "intrinsic curvature" and "extrinsic curvature[4]". The easy to visualize curvature is extrinsic. The physically meaningful curvature is intrinsic! Intrinsic curvature is what General Relativity deals with! Imagine a sheet of paper laid flat on a table. On this paper is a bug. The bug walks along the shortest path on the paper from point A to point B. We can all agree that the paper is flat and that the line is straight. Now roll up the paper into a tube. Again the bug walks along the shortest path from point A to point B. The bug will have walked across the exact same path on the paper (assuming that the shortest path does not cross the seam). From the point of view of the paper, the path is straight. From our external point of view, the paper is curved and the path follows the paper. The above is an example of extrinsic curvature. We have a two dimensional space (the paper) and a larger three dimensional space (our ordinary three dimensional geometry) in which it is embedded. The curvature of the path depends on how we roll up the paper, i.e. how we choose to do the embedding. For an inhabitant living on the surface of the paper using measurements made only on that surface, extrinsic curvature is not detectable.

---

[4] As a common question on Internet is that a space can only be curved in relation to an absolute "straight space"? Otherwise we would never know that it is curved. There must be an absolute straight space geometry underlying the curved space?

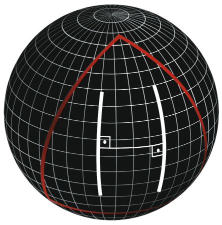
Extrinsic curvature is a property of the embedding. The bug traces out the same path on the paper, whether it is rolled or unrolled. It is only from our external viewpoint that we can see any "curvature". In the case of a sphere, things are not so simple. We cannot roll up a flat sheet of paper into a sphere without stretching or wrinkling it. If the bug on a sphere follows shortest paths and keeps careful track of distances, he can notice that the surface he lives on is not Euclidean. The interior angles of a triangle will sum to more than 180 degrees as we have already proved in the section of Trigonometry. The above is an example of intrinsic curvature. For an inhabitant living on the surface of the paper using measurements made only on that surface, intrinsic curvature is detectable. Here comes the hard part... In the examples above, we talked about a two dimensional surface embedded within a three dimensional Euclidean space. That was just an aid to visualization. We can talk about the geometry of a two (or three or four or more) dimensional space without requiring that it be embedded in a higher dimensional space at all. A common way to do that is to imagine that the inhabitants of the space are able to measure distances. From any point in the space to any other point in the space, they can measure the distance between them. The distance measurement is the mathematical notion of a "metric". A space for which a metric exists is named a "metric space" as we know it. Given a metric space, one can define intrinsic curvature in terms of the metric. A standard way of handling a metric space in physics is to divide it up (if needed) into pieces and equip each piece with a Cartesian coordinate system. There are some rules about the boundaries and how to handle the seams between the pieces that we need not concern ourselves with. The result is a "manifold". A surface of a sphere can be modelled as a two dimensional manifold. There is no embedding in three dimensional space and no meaningful notion of extrinsic curvature. There is still non-zero intrinsic curvature for this space.

Definition (#8.3.347): "Intrinsic curvature" comes from the parallel translation of a vector tangent to the path of translation. If a vector is translated around a loop and it fails to come back onto itself that is intrinsic curvature (see an example in the section of Tensor Calculus page 2114). "Extrinsic curvature" is computed by the parallel translation of a vector normal to a surface or space. If the translated normal deviates from the normal vector at a point the difference in the two normal vectors defines the extrinsic curvature for a unit of translation along the space.

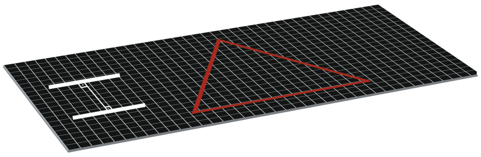
Before we tackle formally and with an abstract manner to certain non-Euclidean geometries we will first make a pragmatic and specific introduction of certain concepts that we are not totally foreign for because already covered in other section theoretically. Once made this introduction, which will be very useful pedagogically, we will address these concepts more rigorously.

8.3.2 Axioms of non-euclidean geometry

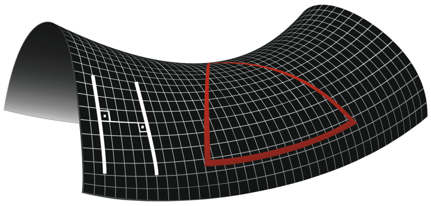
To obtain a non-Euclidean geometry, the parallel postulate A.P1 (or its equivalent) that we have seen during our study of Euclidean Geometry must be replaced by its negation (A.NP1). Everything that follows is related to that figure again:



Positive curvature
(spherical like/elliptic)



Flat (zero) curvature
(plane like/euclidian)



Negative curvature
(horse's saddle like/hyperbolic)

Then the axiom of parallels becomes:

A.NP1. The also named "Lobachevsky axiom" or "Hyperbolic Parallel Postulate (HPP)" assume that there exist at least two lines passing through the given point P not located on a given line and parallel to this line (we speak then of "two diverging ultra-parallel lines").

An alternative and more easy formulation is:

A.NP1'. There exist a line L and a point P not on L such that at least two distinct lines parallel to l pass through P. In other words: Through a point not on a line there is more than one line parallel to the given line!!!

The above form of the axiom of parallelism ensures the existence of an infinite number of parallel lines satisfying the given conditions. Geometric space defined by the axiomatic system including Lobachevsky axiom is named "hyperbolic non-Euclidean geometry". Geometry of the curved surface hyperboloid can be described as a hyperbolic geometry. In the hyperbolic space, the interior angles in an arbitrary triangle is less than π.

A second possible form of the negation of axiom of parallelism is as follows:

A.NP1. There exist no line passing through the given point not located on a given line and parallel to this line.

The geometric space defined by this last axiomatic system is named "elliptic non-Euclidean geometry". Geometry of the sphere (spherical geometry) can be regarded as elliptic geometry. In the elliptic space, the interior angles in an arbitrary triangle is greater than π (see the proof just below).

8.3.3 Generalized angle sum theorems

We have already proven the angle sum theorem for the euclidean geometry earlier in the section Trigonometry at page 2593. Let us see now a very important practical case for non-euclidean geometries: the case of the sphere!

To prove that for the spherical case the sum of the angles of a triangle is greater than 180 degrees (π radian), we first need to introduce the concept of Lune on a sphere.

Definition (#8.3.348): In spherical geometry, a "spherical lune" (or "biangle") is an area on a sphere bounded by two half great circles which meet at antipodal points.

Given the figure below that will be of main interest for us later:

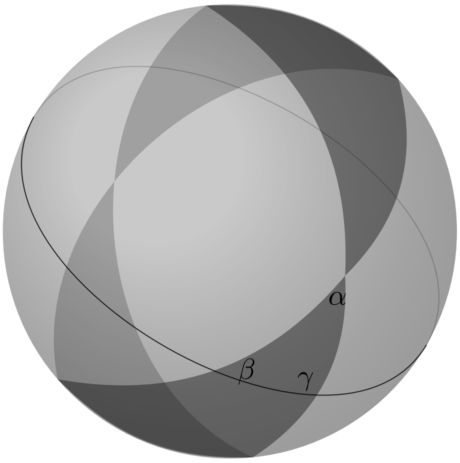


Figure 8.98: **Lunes in a sphere**

where we have 6 lunes denoted $L(\alpha)$, $L(\bar{\alpha}$, $L(\beta)$, $L(\bar{\beta})$, $L(\gamma)$ and $L(\bar{\gamma})$.

We know that the surface of a sphere is $4\pi R^2$ (see page 2964), the surface of a Lune of angle θ is therefore straightforward:

$$S_{\text{Lune}}(\theta)=4\pi R^2\frac{\theta}{2\pi}=2\theta R^2 \tag{8.3.1}$$

Now that we have introduce what a Lune is and how to compute its surface we can prove the angle sum theorem for a spherical non-euclidean space!

Consider the three angles α, β, and γ of the triangle visible on the sphere in the figure above.

The six lunes defined by the angles of the triangle encompass the entire sphere as well as overlapping the triangle area in the front and rear of the sphere 4 extra times. Thus, we begin with the radius of a sphere plus 4 times the area of the triangle is equal to the 6 lunes.

Therefore:

$$4\pi R^2+4S_{\alpha\beta\gamma}=2L(\alpha)+2L(\beta)+2L(\gamma) \tag{8.3.2}$$

Simplifying we get:

$$2\pi R^2 + 2S_{\alpha\beta\gamma} = L(\alpha) + L(\beta) + L(\gamma) \tag{8.3.3}$$

At this point we need to use a the previous result that states that a lune whose corner angle is θ radians has area $2\theta R^2$:

$$2\pi R^2 + 2S_{\alpha\beta\gamma} = 2\alpha R^2 + 2\beta R^2 + 2\gamma R^2 \tag{8.3.4}$$

Hence:

$$2\pi R^2 + 2S_{\alpha\beta\gamma} = 2R^2(\alpha + \beta + \gamma) \tag{8.3.5}$$

Therefore:

$$\pi + \frac{S_{\alpha\beta\gamma}}{R^2} = \alpha + \beta + \gamma \tag{8.3.6}$$

So as ${}^{S_{\alpha\beta\gamma}}/_{R^2} > 0$, we have finally:

$$\alpha + \beta + \gamma > \pi \tag{8.3.7}$$

This result is sometimes named the "spherical triangle theorem".

As we have proved for the flat euclidean geometry that $\alpha + \beta + \gamma = \pi$ and just now for the spherical non-euclidean geometry that $\alpha + \beta + \gamma > \pi$, therefore by deduction for the hypergeometric non-euclidean geometry we have $\alpha + \beta + \gamma < \pi$.

8.3.4 Geodesic and Metric Equation

Let us come back on the concepts of geodesic and curvature which we have often mentioned in the section of Tensor Calculus (the fact of not having read the section of Tensor Calculus is normally not a problem in understanding what follows).

Consider, as an introduction example, the two-dimensional surface of a sphere of radius R. Given two points B and C diametrically opposed, we seek the shortest distance s measured on the sphere between B and C. The curve we get is as we know a "geodesic", a concept that generalizes therefore, for an arbitrary surface, the notion of the straight lines for planes:

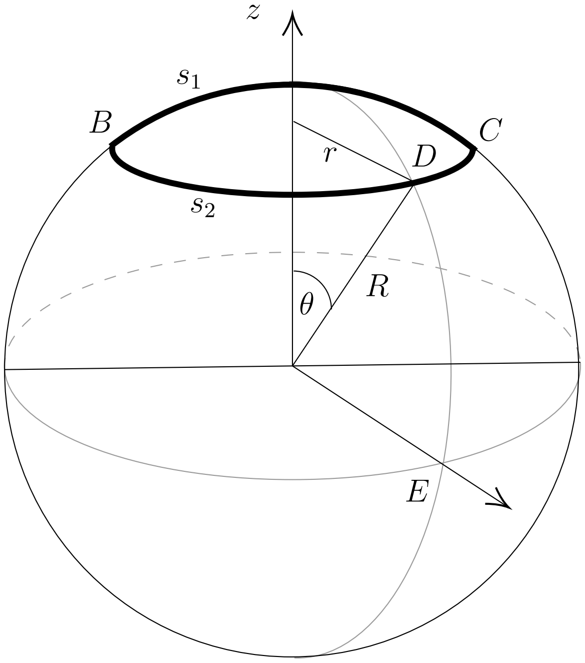


Figure 8.99: **Illustration of the problem to research the geodesic on a sphere**

Remark

We assume as intuitive that the length of a curve in the three-dimensional Euclidean space is always greater than or equal to the length of any planar projection of this curve. The geodesic curve is therefore necessarily a plane curve.

The radius between the axis Oz and one of the points B and C is trivially given by some elementary trigonometry (see section Trigonometry page 2519):

$$r = R\sin(\theta) \tag{8.3.8}$$

And therefore the half circumference of the circle at the height of B and C is given by:

$$s_2 = \frac{P_2}{2} = \frac{2\pi R\sin(\theta)}{2} = \pi R\sin(\theta) \tag{8.3.9}$$

And we have proved it in the section of Trigonometry that the perimeter of a circle depending on the opening angle of the latter was given by (α must be in radians!):

$$L = R\alpha \tag{8.3.10}$$

It therefore comes automatically:

$$s_1 = R2\theta \tag{8.3.11}$$

Finally:

$$\begin{aligned} s_1 &= R2\theta \\ s_2 &= R\pi\sin(\theta) \end{aligned} \tag{8.3.12}$$

A plot or Taylor approximation shows easily that $\pi\sin(\theta) \geq 2\theta$ on the interval $[0,\pi/2]$ then on the same interval $s_2 \geq s_1$ (there is equality in $\theta = 0$ and $\theta = \pi/2$).

The geodesics of the sphere are therefore the big circles arcs s_1, paths taken by aircraft for inter-continental flight (in empty space, without wind and perfect spherical planet), and correspond to the lines obtained between the surface of the sphere and a plane passing through the center thereof!

The geometrical properties of figures drawn on the surface of a sphere are obviously no longer those of Euclidean Geometry. Thus, the shortest path from point B to point C on the spherical surface, consists of a great circle passing through the points B and C plotted on a plane passing through the center of the sphere. Great circle arcs play the same role for the sphere that lines in the plane. These are the "geodesic" of the sphere.

Now consider two dimensional surfaces: the surface of the sphere and that of the cylinder. Given two points B and C, we trace the geodesic curve between these points:

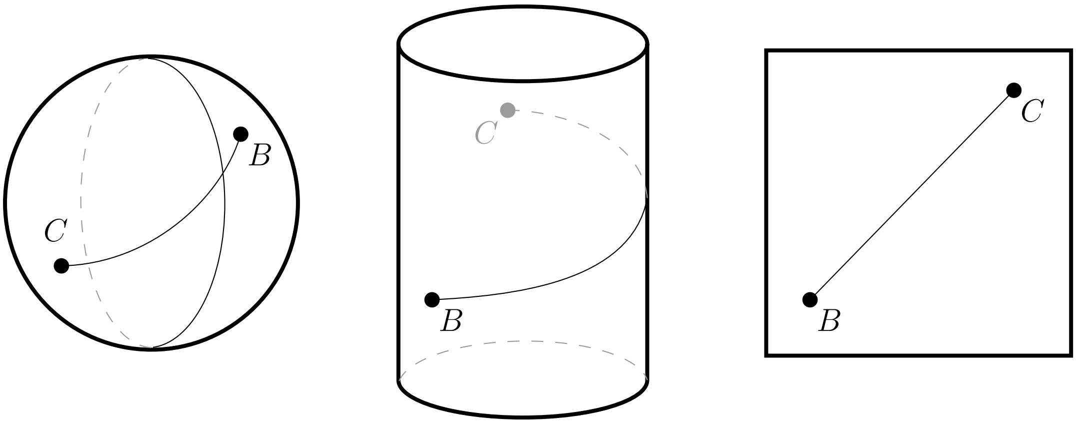


Figure 8.100: **Some geodesics of trivial surface on trivial volumes**

The cylinder may be cut parallel to its axis and unfolded flat. The geodesic thus appears as a straight line of the plane! We say then that the cylinder is "inherently flat" (even if its topology is different from the plane, we must especially avoid that the cut crosses the geodesic). This is intuitively obviously not the case of the surface of the sphere.

In the case of the cylindrical surface, we can define the Cartesian coordinates of the plane $B(y_1,z_1)$ and $C(y_2,z_2)$ that permits to write the length s of the curve (straight line) BC with the Pythagorean theorem:

$$s^2=(y_2-y_1)^2+(z_2-z_1)^2 \tag{8.3.13}$$

The metric of the plane is Euclidean and under infinitesimal form we get the "Euclidean metric equation":

$$\mathrm{d}s^2=\mathrm{d}y^2+\mathrm{d}z^2 \tag{8.3.14}$$

On the cylinder, the change of variable gives $y=r\theta$ (with the angle in radians!):

$$\mathrm{d}s^2=(r\theta_2-r\theta_1)^2+(z_2-z_1)^2=r^2(\theta_2-\theta_1)^2+(z_2-z_1)^2 \tag{8.3.15}$$

Or under local form:

$$\mathrm{d}s^2=r^2\mathrm{d}\theta^2+\mathrm{d}z^2 \tag{8.3.16}$$

The surface of the cylinder can thus be represented by Cartesian coordinates similar to those of the plane, the metric of the cylinder surface being Euclidean under infinitesimal form and under global form.

Remark

The previous relation is what we got in the section of Tensor Calculus for the metric equation in polar coordinates.

We can interest us now to the problem to write the analogue of Pythagorean theorem to a spherical surface. The impossibility to cut the ball and flatten it to marry a plane suggests some difficulties...

That is why the equation of the metric can't be written in general form as the Pythagorean theorem. Indeed, we proved in the section of Tensor Calculus (page 2099) that this latter was given for a spherical surface by:

$$\mathrm{d}s^2=\mathrm{d}r^2+r^2\mathrm{d}\theta^2+r^2\sin^2(\theta)\mathrm{d}\phi^2 \tag{8.3.17}$$

However, locally (that is to say in a small region of small dimension relatively to the radius of the sphere), the properties of the sphere can be described by Cartesian coordinates of a plane tangent to the surface (the essential property of Riemann spaces) as the metric equation is locally Euclidean!:

$$\mathrm{d}s^2\cong r^2\mathrm{d}\theta^2+r^2\sin^2(\theta)\phi z^2 \tag{8.3.18}$$

By writing:

$$\mathrm{d}\xi=\sqrt{g_{\theta\theta}}\mathrm{d}\theta \qquad \mathrm{d}\eta=\sqrt{g_{\phi\phi}}\mathrm{d}\phi \tag{8.3.19}$$

We therefore have:

$$\mathrm{d}s^2\cong g_{\theta\theta}\mathrm{d}\theta^2+g_{\phi\phi}\mathrm{d}\phi^2=\mathrm{d}\xi^2+\mathrm{d}\eta^2 \tag{8.3.20}$$

with:

$$g_{\theta\theta}=r^2 \qquad g_{\phi\phi}=r^2\sin^2(\theta) \tag{8.3.21}$$

While (θ,ϕ) are the "Gauss coordinates", (ξ,η) are the "Riemann coordinates" of the locally tangent plane. Thus, we notice that Euclidean space is a special case of Gauss coordinates for which we have (see section Special Relativity page 5704):

$$\begin{aligned}\mathrm{d}s^2&=g_{ij}\mathrm{d}x^i\mathrm{d}x^j=g_{11}\mathrm{d}u\mathrm{d}u+g_{12}\mathrm{d}u\mathrm{d}v+g_{21}\mathrm{d}v\mathrm{d}u+g_{22}\mathrm{d}v\mathrm{d}v\\&=1\mathrm{d}u\mathrm{d}u+0\mathrm{d}u\mathrm{d}v+0\mathrm{d}v\mathrm{d}u+1\mathrm{d}v\mathrm{d}v=\mathrm{d}u^2+\mathrm{d}v^2\end{aligned} \tag{8.3.22}$$

and thus the metric tensor is a identity matrix (the diagonal elements are equal to 1 the rest being zero):

$$g_{ij}=\begin{bmatrix}1&0\\0&1\end{bmatrix} \tag{8.3.23}$$

The reader will have perhaps noticed that if we take the vector basis of the polar coordinates as seen in the section of Vector Calculus but without normalizing them to then unit (that is to say without dividing the first vector by r and the second by $\sin(\theta)$) we have:

$$g_{ij}=\vec{e}_i\circ\vec{e}_j \tag{8.3.24}$$

with $i,j=\{\theta,\phi\}$ and so on for every other coordinate system.

8.3.5 Riemann Spaces

To better understand what is a Riemann space, we will now go through a small example of a two-dimensional surface (classic example):

Consider a sphere of radius r, of surface S, located in the ordinary three-dimensional space. The Cartesian coordinates x,y,z of a point M on the surface S can be expressed, for example, depending on the spherical coordinate (r,θ,ϕ) (see section Vector Calculus page 1849). The sphere is fully described for a given radius r and $0 \leq \phi < 2\phi$ and $0 \leq \theta < \pi$.

Three such parameters, giving the possibility to determine a point on the surface of a sphere, are as we know (see section Tensor Calculus page 2092) curvilinear coordinates of the surface or also named "Gaussian coordinates" (Gauss being one of the first mathematicians interested in the study of bodies immersed in non-Euclidean spaces). Other arbitrary parameters u,v,w may of course be chosen as curvilinear coordinates on the surface.

The linear element $\mathrm{d}s^2$ of the surface, square distance between two infinitely close points M,M', is written based on spherical coordinates, as we have seen in the section of Tensor Calculus and just above:

$$\mathrm{d}s^2 \cong r^2\mathrm{d}\theta^2 + r^2\sin^2(\theta)\mathrm{d}\phi^2 \tag{8.3.25}$$

We thus obtain an expression of the linear element based on only the three Gauss coordinates (r,θ,ϕ). We could of course impose a local study (tangent plane) so that the linear element is only function of (θ,ϕ) as we have seen above:

$$\mathrm{d}s^2 \cong r^2\sin^2(\theta)\mathrm{d}\phi^2 \tag{8.3.26}$$

Written using the three parameters, the surface of the sphere (considered as a two-dimensional space) is an example of two-dimensional Riemann space whose linear element is in the general form well known (see the chapter of tensor calculus):

$$\mathrm{d}s^2 = g_{ij}\mathrm{d}u^i\mathrm{d}u^j \tag{8.3.27}$$

where the $\mathrm{d}u^i, \mathrm{d}u^j$ are the contravariant components of the vector $\mathrm{d}\vec{M} = \overrightarrow{MM'}$ relatively to the referential denoted by $(M,\vec{e}_j)$.

Remark

The study of figures on Riemannian surfaces is part of differential geometry to which we devote a whole section in this chapter.

Now consider any surface of coordinates u^1,u^2. The Cartesian coordinates x,y,z of ordinary space where this surface is dived are written in general terms with the Gauss coordinates:

$$x = x(u^1,u^2) \qquad y = y(u^1,u^2) \qquad z = z(u^1,u^2) \tag{8.3.28}$$

Remember also that the metric equation can be written in tensorial notation:

$$\mathrm{d}s^2 = g_{ij}\mathrm{d}u^i\mathrm{d}u^j \tag{8.3.29}$$

can be written in an expanded form as follows (this is proved in details with a geometric approach in the section of Differential Geometry at page 2880):

$$\boxed{\mathrm{d}s^2=E\mathrm{d}u^2+2F\mathrm{d}u\mathrm{d}v+G\mathrm{d}v^2} \tag{8.3.30}$$

with:

$$\begin{aligned}
E&=\left(\frac{\partial}{\partial u}x\right)^2+\left(\frac{\partial}{\partial u}y\right)^2+\left(\frac{\partial}{\partial u}z\right)^2=(\partial_1 x)^2+(\partial_1 y)^2+(\partial_1 z)^2\\
G&=\left(\frac{\partial}{\partial v}x\right)^2+\left(\frac{\partial}{\partial v}y\right)^2+\left(\frac{\partial}{\partial v}z\right)^2=(\partial_2 x)^2+(\partial_2 y)^2+(\partial_2 z)^2\\
F&=\frac{\partial}{\partial u}x\frac{\partial}{\partial v}x+\frac{\partial}{\partial u}y\frac{\partial}{\partial v}y+\frac{\partial}{\partial u}z\frac{\partial}{\partial v}z=\partial_1 x\partial_2 x+\partial_1 y\partial_2 y+\partial_1 z\partial_2 z
\end{aligned} \tag{8.3.31}$$

Remarks

R1. The expression given above for the linear elements $\mathrm{d}s$ is named "quadratic fundamental form" of the surface considered. The coefficients E,F,G are functions of curvilinear coordinates. Generally this surface, considered a two-dimensional space, will be an example of Riemann space, for arbitrary curvilinear coordinates.

R2. The different Riemann spaces are what we name in general form (because there is not only Riemannian spaces of constant curvature) a "variety" equipped with a Riemannian metric. A variety may be defined (not formally), for example, by a set of points into an existing space. Generally a surface gives the idea of a two-dimensional variety. The sphere and the torus are two-dimensional varieties without borders. A cylinder of revolution, a hyperbolic paraboloid are two-dimensional open varieties with borders to infinity. But we can also consider abstract varieties. This is the case for example of a configuration space. This is then an n-dimensional space represented by a set q^i (or denoted by u^i) of generalized coordinates (see the introduction to the Lagrangian formalism in the section of Analytical Mechanics page 3155), the latter can have values in a finite domain or not.

We can now define in a little better way what is a Riemann space.

Definition (#8.3.349): A "Riemann space" is a variety to which we attached a metric. This means that in every part of the variety analytically represented by a coordinate system u^i, we gave ourselves a quadratic differential form:

$$\mathrm{d}s^2=g_{ij}\mathrm{d}u^i\mathrm{d}u^j \tag{8.3.32}$$

which is the metric of space.

The coefficients g_{ij} are not entirely arbitrary and should verify, as we have shown in the section of Tensor Calculus, the following conditions:

C1. The components are symmetric as $g_{ij}=g_{ji}$

C2. The determinant of the matrix of g_{ij} is not zero

C3. The differential form of the linear element $\mathrm{d}s^2$, and therefore the concept of distance defined by the g_{ij}, is invariant vis-a-vis any change of coordinates.

C4. All partial differential of second order of the g_{ij} exist and are continuous and therefore of class $\mathcal{C}^2$.

A Riemann space is a space of points, each being identified by a system of n coordinates u^i, with a metric such as the differential form of the linear element that satisfies the above conditions. This metric is therefore named "Riemannian metric".

Remarks

R1. If the metric is positive definite, that is to say, if $g_{ij}u^iu^j$ for any non-zero vector $\vec{v}$, we say that the space is "strictly Riemannian". In this case, the determinant of the matrix g_{ij} is strictly positive and all the eigenvalues of the matrix are strictly positive (see section Analytical Geometry page 2794).

R2. By definition, we say that a metric space is Euclidean when any fundamental tensor of this space may be reduced, by an appropriate change of coordinates, to a form such as (see section Vector Calculus page 1810) the canonical orthonormal basis. That is to say $g_{ij}=\delta_{ij}$.

R3. The definition of Riemannian spaces shows that Euclidean space is a very special case of these spaces. So there exists only one Euclidean space but we can create an infinite number of Riemannian spaces.

It is important to see that:

$$\mathrm{d}s^2=E\mathrm{d}u^2+2F\mathrm{d}u\mathrm{d}v+G\mathrm{d}v^2 \tag{8.3.33}$$

Can be rewritten as:

$$\frac{\mathrm{d}s^2}{\mathrm{d}t^2}=E\frac{\mathrm{d}u^2}{\mathrm{d}t^2}+2F\frac{\mathrm{d}u\mathrm{d}v}{\mathrm{d}t^2}+G\frac{\mathrm{d}v^2}{\mathrm{d}t^2}=E\left(\frac{\mathrm{d}u}{\mathrm{d}t}\right)^2+2F\frac{\mathrm{d}u}{\mathrm{d}t}\frac{\mathrm{d}v}{\mathrm{d}t}+G\left(\frac{\mathrm{d}v}{\mathrm{d}t}\right)^2 \tag{8.3.34}$$

that we denote by tradition:

$$\frac{\mathrm{d}s^2}{\mathrm{d}t^2}=E\dot{\mathrm{u}}^2+2F\dot{\mathrm{u}}\dot{\mathrm{v}}+G\dot{\mathrm{v}}^2 \tag{8.3.35}$$

Therefore:

$$\mathrm{d}s=\sqrt{E\dot{\mathrm{u}}^2+2F\dot{\mathrm{u}}\dot{\mathrm{v}}+G\dot{\mathrm{v}}^2}\mathrm{d}t \tag{8.3.36}$$

Hence the length of a curve on any differential surface:

$$\boxed{s(t_0,t_1)=\int\limits_{t_0}^{t_1}\sqrt{E\dot{\mathrm{u}}^2+2F\dot{\mathrm{u}}\dot{\mathrm{v}}+G\dot{\mathrm{v}}^2}\mathrm{d}t} \tag{8.3.37}$$

✍**Example:**

Let us consider the parametric equation of a special hyperbolic surface:

$$\Sigma(u,v)=\begin{pmatrix}u\\v\\uv\end{pmatrix},\ u,v\in\mathbb{R}^2 \tag{8.3.38}$$

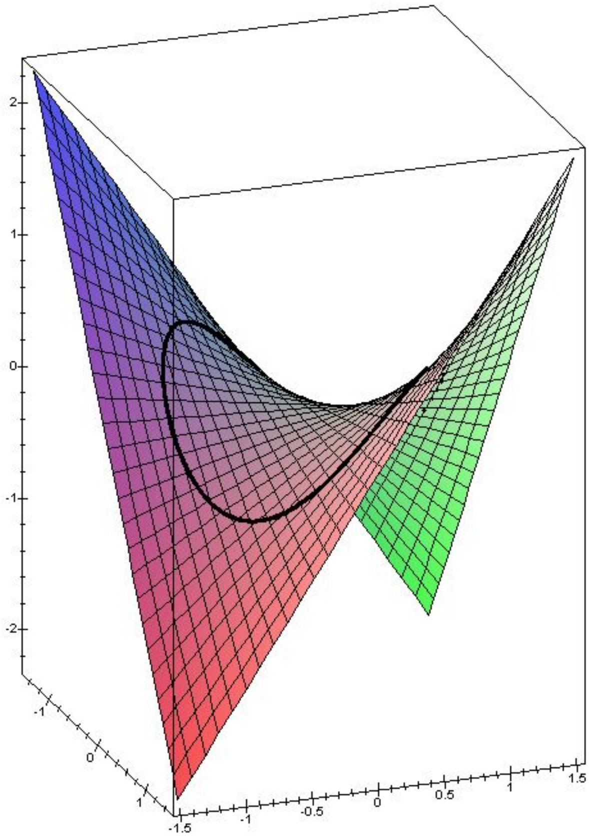
We take now the parametric equation of a circle of radius 1 on the plane:

$$\Gamma(u(t),v(t))=\begin{pmatrix}\cos(t)\\\sin(t)\end{pmatrix} \tag{8.3.39}$$

That is to say visually with Maple™ 4.00b:

```
>x:=u;y:=v;z:=u*v;
>with(plots):
>surface:=plot3d([x,y,z],u=-1.5..1.5,v=-1.5..1.5,grid=[30,30],
scaling=constrained,axes=boxed,style=PATCH):
>curve:=spacecurve([cos(t),sin(t),cos(t)*sin(t)],t=-4*Pi..4*Pi,
numpoints=1000,scaling=CONSTRAINED,orientation=[50,60],style=PATCH,
axes=NORMAL,color=black,thickness=3):
>display(surface,curve,orientation=[-20,70]);
```

Which gives:



Therefore:

$$\begin{aligned}
E&=(\partial_1 x)^2+(\partial_1 y)^2+(\partial_1 z)^2=1^2+0^2+v^2=1+v^2\\
G&=(\partial_2 x)^2+(\partial_2 y)^2+(\partial_2 z)^2=0^2+1^2+u^2=1+u^2\\
F&=\partial_1 x\partial_2 x+\partial_1 y\partial_2 y+\partial_1 z\partial_2 z=0^2+0^2+uv=uv
\end{aligned}\tag{8.3.40}$$

Then we have:

$$s(t_0,t_1)=\int\limits_{t_0}^{t^1}\sqrt{(1+v^2)\dot{\mathrm{u}}^2+2uv\dot{\mathrm{u}}\dot{\mathrm{v}}+(1+u^2)\dot{\mathrm{v}}^2}\mathrm{d}t\tag{8.3.41}$$

And we want to calculate the perimeter of the circle on the surface. Therefore:

$$\begin{aligned}
&s(0,2\pi)=\int\limits_0^{2\pi}\sqrt{(1+v^2)\dot{\mathrm{u}}^2+2uv\dot{\mathrm{u}}\dot{\mathrm{v}}+(1+u^2)\dot{\mathrm{v}}^2}\mathrm{d}t\\
&=\int\limits_0^{2\pi}\sqrt{(1+\cos^2(t))\dot{\mathrm{u}}^2+2\cos(t)\sin(t)\dot{\mathrm{u}}\dot{\mathrm{v}}+(1+\sin^2(t))\dot{\mathrm{v}}^2}\mathrm{d}t\\
&=\int\limits_0^{2\pi}\sqrt{(1+\sin^2(t))(-\sin(t))^2+2\cos(t)\sin(t)(-\sin(t))(\cos(t))+(1+\cos^2(t))(\cos(t))^2}\mathrm{d}t\\
&=\int\limits_0^{2\pi}\sqrt{(1+\sin^2(t))(-\sin(t))^2-2\cos^2(t)\sin^2(t)+(1+\cos^2(t))(\cos(t))^2}\mathrm{d}t\\
&=\int\limits_0^{2\pi}\sqrt{1+2\cos^2(t)\sin^2(t)+\sin^4(t)+\cos^4(t)}\mathrm{d}t\\
&=\int\limits_0^{2\pi}\sqrt{1+(\cos^2(t)-\sin^2(t))^2}\mathrm{d}t\\
&=\int\limits_0^{2\pi}\sqrt{1+\cos^2(2t)}\mathrm{d}t
\end{aligned}\tag{8.3.42}$$

We make a change of variable $\tau=2t$, and therefore:

$$\mathrm{d}\tau=2\mathrm{d}t\qquad 2\pi\rightarrow 2\pi\qquad 0\rightarrow 0\tag{8.3.43}$$

and therefore:

$$s(0,2\pi)=\frac{1}{2}\int\limits_0^{4\pi}\sqrt{1+\cos^2(\tau)}\mathrm{d}\tau\tag{8.3.44}$$

As the integrand is even:

$$s(0,2\pi)=\int\limits_0^{2\pi}\sqrt{1+\cos^2(\tau)}\mathrm{d}\tau \tag{8.3.45}$$

We have then:

$$\int\limits_0^{2\pi}\sqrt{1+\cos^2(\tau)}\mathrm{d}\tau=4\int_0^{\pi/2}\sqrt{1+\cos^2(\tau)}\mathrm{d}\tau=4\int\limits_0^{\pi/2}\sqrt{2-\sin^2(\tau)}\mathrm{d}x \tag{8.3.46}$$

$$=4\sqrt{2}\int\limits_0^{\pi/2}\sqrt{1-\frac{\sin^2(\tau)}{2}}\mathrm{d}\tau=4\sqrt{2}E\left(\frac{1}{\sqrt{2}}\right) \tag{8.3.47}$$

With:

$$\boxed{E(k)=\int\limits_0^{\pi/2}\sqrt{1-k^2\sin^2(\tau)}\mathrm{d}x} \tag{8.3.48}$$

and is referred to as the "complete elliptic integral of second kind" (see section Differential and Integral Calculus page 1599). This integral cannot be formally calculated. But a numerical integration gives:

$$s(0,2\pi)=\int\limits_0^{2\pi}\sqrt{1+\cos^2(\tau)}\mathrm{d}\tau\cong 7.6404 \tag{8.3.49}$$

~30 %

★★★★☆

26 votes, 80.77%

8.4. Projective Geometry

SINCE Chasles and this until the years 1930s (holocene calendar), Projective Geometry was often synonymous of "higher geometry." It was opposed to Euclidean Geometry: that was said to be elementary and analytical. At the time of Monge, Carnot and von Staudt, we also spoke of "geometry of position" or "geometry of situation". These geometries study figures at the point of view of their respective positions and of their invariant properties which bind them in a geometric transformation (rotation, symmetry, scaling, etc.), homographic in particular. Besides the harmonic division, basic concept, it uses the famous anharmonic ratio (cross ratio), the inverse, the involution the transformation by reciprocal polar, stereographic projection[5], the correlation, the homology, the duality, the conical, etc. (see later in this text for the definitions).

Definition (#8.4.350): "Projective geometry" is a topic of mathematics. It is the study of geometric properties that are invariant with respect to projective transformations. This means that, compared to elementary geometry, projective geometry has a different setting, "projective space", and a selective set of basic geometric concepts.

Project Geometry is quite abstract apart from some general and basic principles outlined below (undergraduate level). It is therefore relatively difficult to understand and requires before study it: a good knowledge of elementary geometry even in space, a perfect mastery of three-dimensional space in the observation-representation-interpretation cycle, to master also complex analysis for advanced topics. This will led us to introduce some concepts that do not normally have their place in this section, but that we think, can greatly help the reader to understand this branch of mathematics.

In a first place, we will discuss the basic concepts of perspective with a particular focus on the concept of "projective" representation (there are other empirical perspective methods: cavalier, cabinet, isometric, military, ... but these latter don't have mathematical or real meaning even if they represent quite properly volumetric objects). Then we will study the mathematical representations of some three-dimensional objects with some computer application to finally go study "hard" and "pure" projective geometry.

Definition (#8.4.351): "perspective" refers to a way of seeing, analyzing, or interpreting something, often influenced by cultural, historical, or personal contexts. In visual arts, it is a technique for representing depth and three-dimensionality on a flat surface. In social science or philosophy, it can refer to a point of view or conceptual position on a particular subject.

Remark

The "descriptive geometry" is a rigorous artistic form of projective geometry but not a formal on (in the sense that it is not mathematically formalized... at least as far as we know...).

---

[5] ie: projection on a plane

8.4.1 Conical Perspective (Central Perspective)

One problem with the study of three-dimensional volumes and their representation is the concept of "perspective". Indeed, human beings can not see the 3 dimensions of an object, it is the brain that interprets the shadows and reflections of an object so that we can interpret it as having a volume (there are optical illusions that go in this direction ...: the "trompe l'oeil").

Here is an animated example of one of the most famous optical illusion, named the Shepard tabletop illusion, that shows that we should, statistically speaking, not trust our intuitions too much...:

The animation above will run for people having a PDF reader with Adobe Flash player installed and activated (otherwise see here: `https://vimeo.com/575741013`).

We will focus in the following paragraphs on the "conical perspective", also named "central perspective" or "linear perspective". Linear perspective is thought to have been devised about 11415 (holocene calendar) by Italian Renaissance architect Filippo Brunelleschi and later documented by architect and writer Leon Battista Alberti in 11435 (holocene calendar).

Remark

In the field of projective geometry we are not talking about "conical perspective" but "conical projection".

Definition (#8.4.352): The "conical perspective" or "perspective projection" (to not be confused with the "conical projection"!!!!!) is by construction the closest representation of our human visual perceptions, it allows especially to see a sphere as a circle as the human vision is based on a visual cone that when looking to a sphere is equivalent to see a circle as illustrated below:

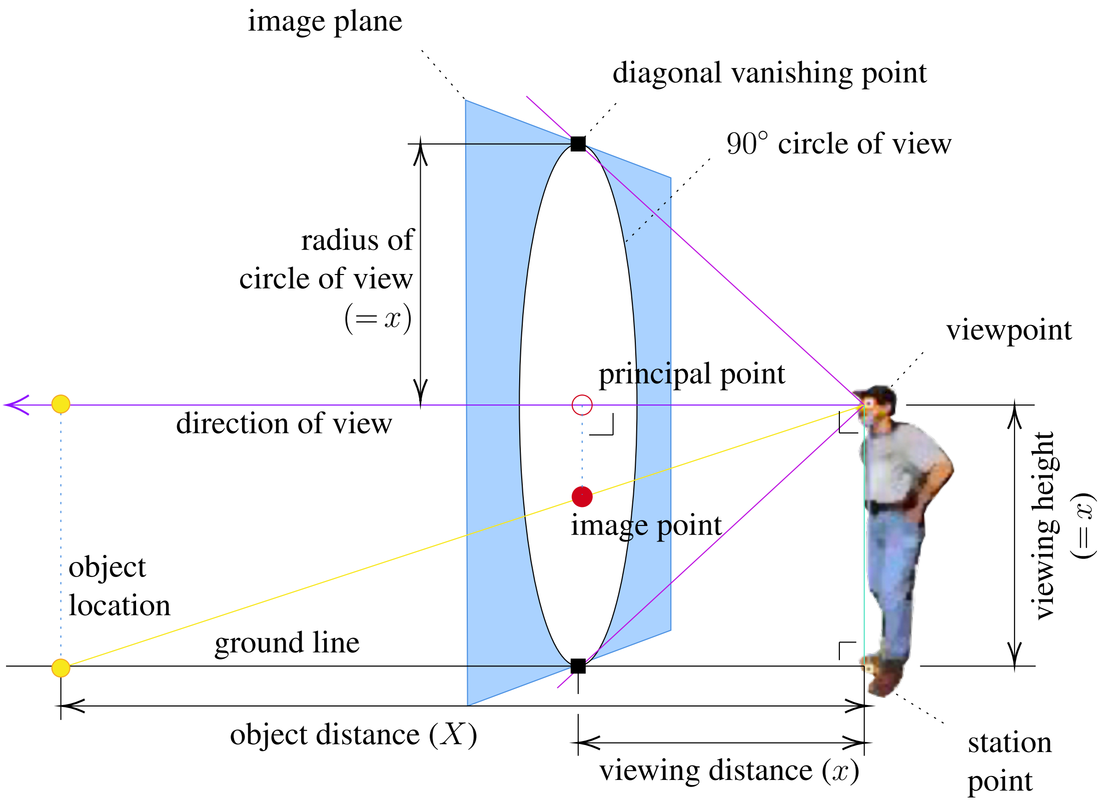


image plane
diagonal vanishing point
90° circle of view
radius of circle of view $(= x)$
viewpoint
principal point
direction of view
image point
object location
ground line
viewing height $(= x)$
object distance (X)
viewing distance (x)
station point

Figure 8.101: **Visual cone and corresponding projection circle**

Remark

The conical perspective is that of Renaissance painters. It is also the one that appears on photos.

The difficulty of the representation in perspective is to translate it in a plane (for example that of the paper or the computer screen as our eyes makes a projection of a 3D world in a 2D image in our brain) a construction (object) that is defined - fairly simply, by the way - in space using mathematical tools.

A conical perspective of the conventional 3-dimensional space is a projective transformation that sends all the points of space on the same plane of this space. It requires the information of a point O (equivalent to the position of the observer's eye) and a projection plane named the "table" or "Dürer glass" (the equivalent of our retina) as referring to a quite old copy painting method:

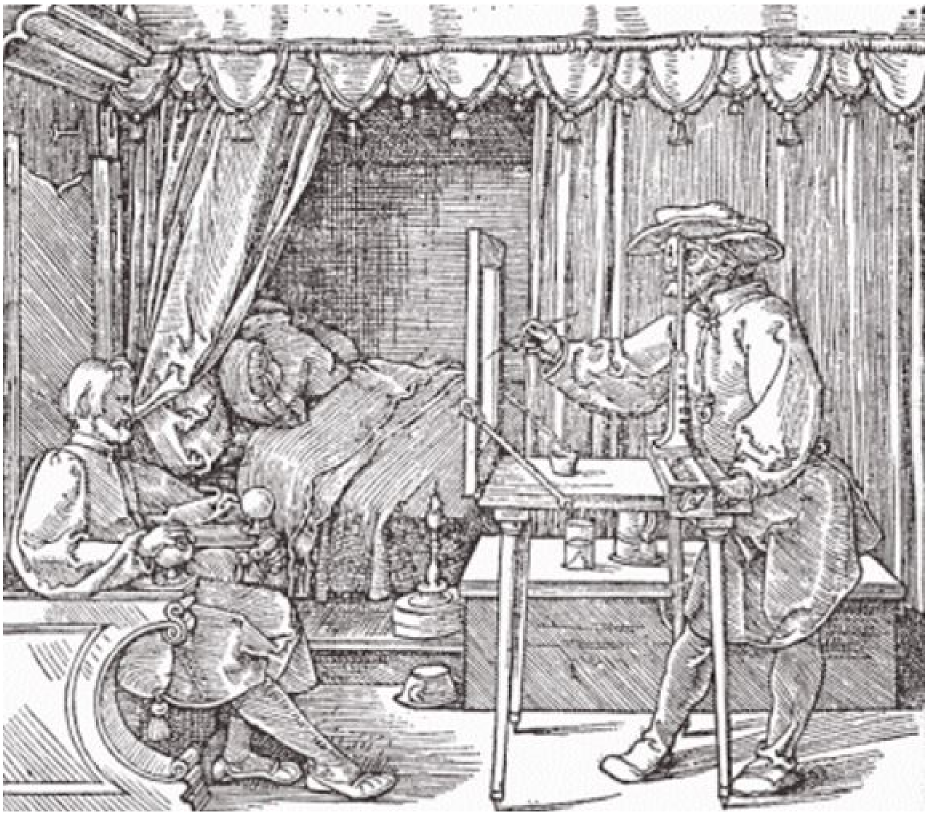


Figure 8.102: **Origin of Durer glass method illustration**

We will see during the theoretical mathematical developments that unlike affine projections (see the following subsection), the conical projection does not retain the barycenter (therefore the ratios of lengths on a given straight line) but it conserve the alignment and cross ratio.

When we speak of conical perspective, we use a few special planes and straight lines in three-dimensional space (see figure below).

Definitions (#8.4.353):

D1. The "picture plane" or "Dürer glass", denoted T, is the plane on which we are drawing (projection plane).

D2. The "ground plane", denoted S, is a fixed plane, perpendicular to the picture plane T.

D3. The "point of view" (or "center of projection"), denoted O, is a point out of T and of S: this is the point where we will have to place the eye for the drawing on the picture plane T coincides with the real image.

D4. The "horizon plane" denoted H, is the plane parallel to the ground plane S passing through the point of view point O.

D5. The "horizon line", denoted h, is the intersection of the horizon plane H and the picture plane T.

D6. The "Earth line", denoted EL, is the intersection of the ground plane S plan and of the picture plane T.

D7. A plane or a straight line parallel to the ground plane S that are named "horizontal".

D8. A plane or a straight line perpendicular to the ground plane S that are named "vertical".

D9. A plane or a straight line parallel to the picture plane T that are named "front".

D10. A plane or a straight line perpendicular to the picture plane T that are named "end".

Here is a diagram showing these different concepts:

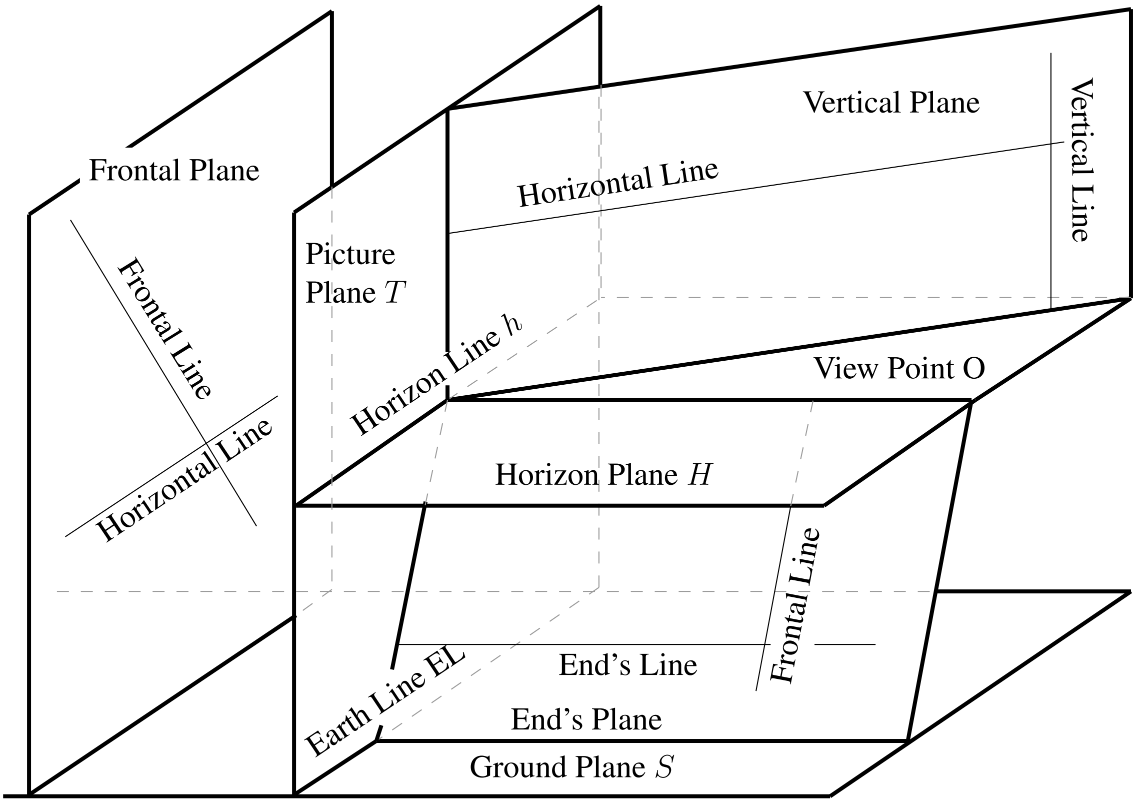

Frontal Plane
Frontal Line
Horizontal Line
Picture
Plane T
Vertical Plane
Horizontal Line
Vertical Line
Horizon Line h
View Point O
Horizon Plane H
Frontal Line
End's Line
Earth Line EL
End's Plane
Ground Plane S

Figure 8.103: **Definition of the planes and lines of projective geometry**

8.4.1.1 Images of Points

Any volumetric object (which can only be determined by touching or at the level of mathematical abstraction) composed by a set of points M, is for our brain, the image of a plane projection m whose support is a surface in the space between the observed object and our eye.

In mathematics, this surface, we know, named the "table" is defined in its central view by the horizon line (where are located the vanishing points) and a physical repository so named the Earth line (see figure below where the observer is at a higher point than the observed object):

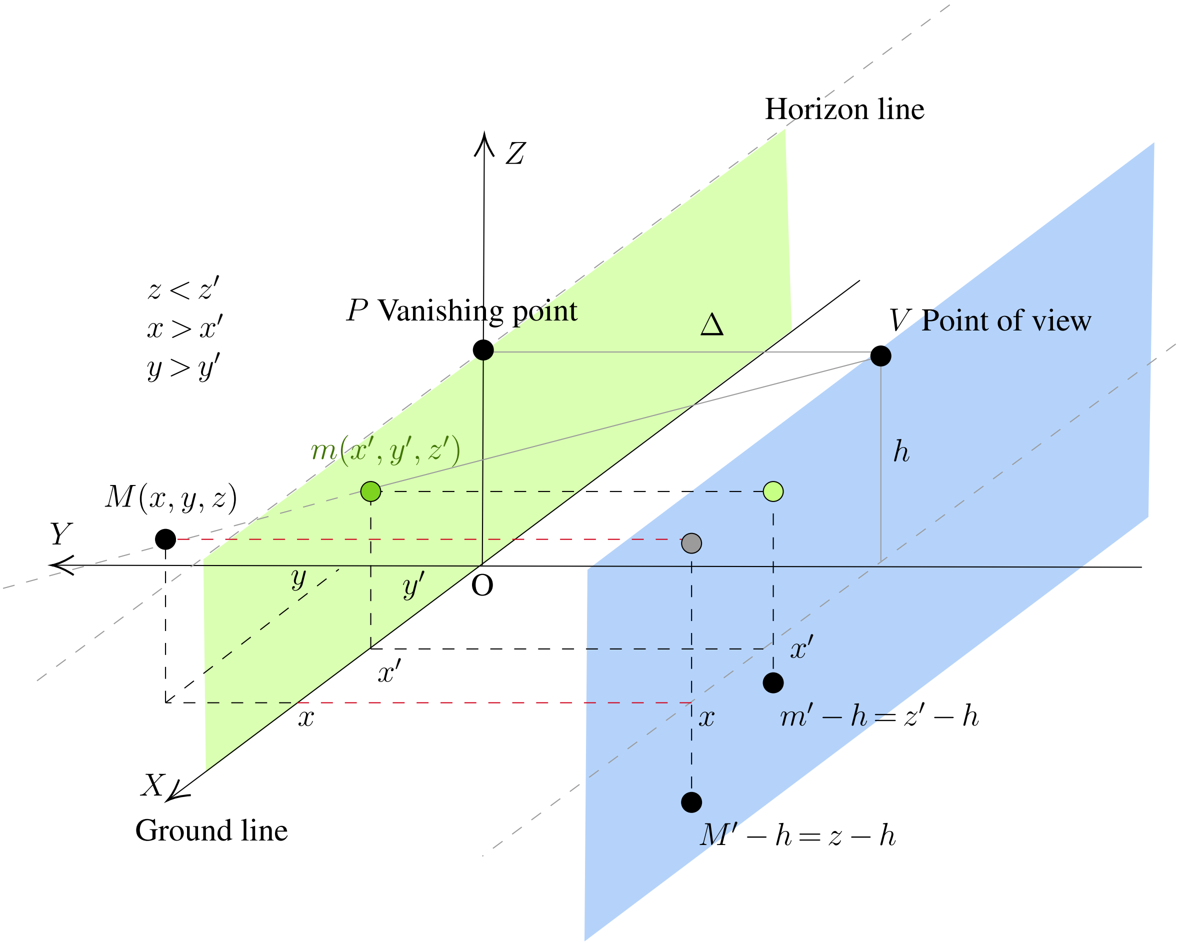


Figure 8.104: Example of what is a "table" in projective geometry

The height between the ground line and the point view is named the "horizon height" and is denoted by h.

The objective from this figure is to mathematically determine a representation of a solid object on a surface (on a plane in a simple case, otherwise any) by knowing the equation of the line between point M (point of the observed object) and the point V (point of view) in the purpose to determine the coordinates of intersection between this line and the table.

We can from the above diagram draw the following relations using the Thales theorem (see section Euclidean Geometry page 2610) in the different triangles:

$$\frac{x}{x'}=\frac{z-h}{z'-h}=\frac{y+\Delta}{y'+\Delta} \tag{8.4.1}$$

with $\Delta \geq 0$.

If we put $y'=0$, according to what shows the figure above, the coordinates of m become:

$$x'=x\frac{\Delta}{y+\Delta} \tag{8.4.2}$$

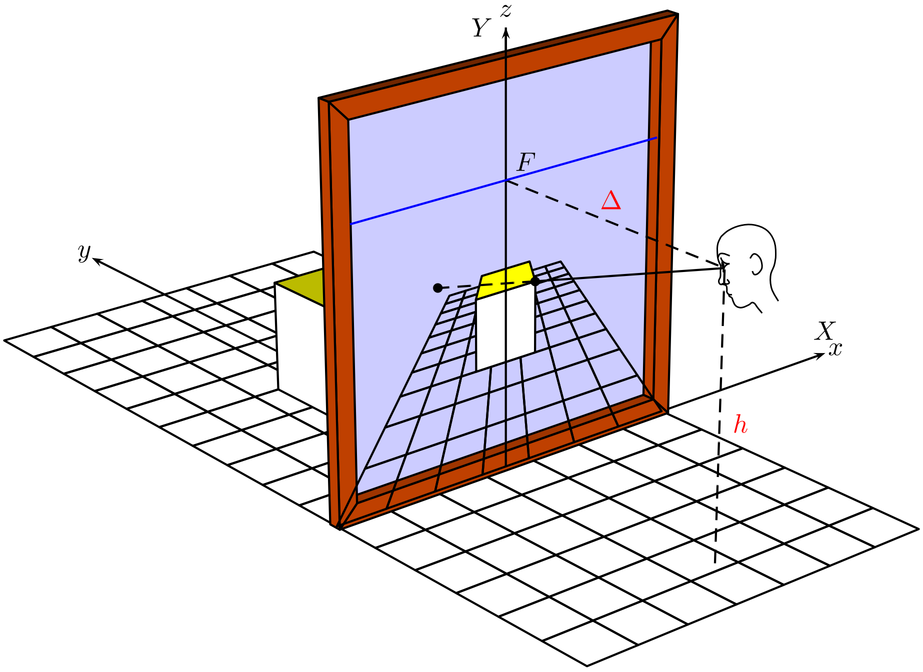
and:

$$z'=h+(z-h)\frac{\Delta}{y+\Delta}=\frac{hy+\Delta z}{y+\Delta} \tag{8.4.3}$$

Often summarized as the "homographic transformations" (we will come back on the latter further below and prove some of their properties):

$$\boxed{\begin{aligned} X&=\frac{x\Delta}{y+\Delta}\\ Y&=\frac{hy+\Delta z}{y+\Delta}\end{aligned}} \tag{8.4.4}$$

and illustrated by:



Notice that whatever the values of $h\geq 0$, $z\geq 0$ or $\Delta\geq 0$ we always have $z'\geq 0$. Hence in linear perspective an object can never disappear by the bottom!

From these relations, the problem of a plane representation of a volumetric shape completely solved, since we can always project a point (or the distance between two points) on a table from the coordinates of the original.

The term Δ is commonly named the "focal length" from the point of view to the table and optical specialists usually denote it with the letter f.

To understand this result, we can put ourselves in the context of a two-dimensional study where the observer is at the height of the ground line ($h=0$) and arranged to represent a person looking at the table (comparable to a TV screen, computer screen or other screen) in which we ask the conventional x and y axes in the screen plane and the z axes perpendicular to it (thus relative to the last figure, Y becomes Z and vice versa).

Thus, the above relation of the ratios:

$$\frac{x}{x'}=\frac{z-h}{z'-h}=\frac{y+\Delta}{y'+\Delta} \quad (8.4.5)$$

becomes with this change of axis:

$$\frac{x}{x'}=\frac{y-h}{y'-h}=\frac{z+\Delta}{z'+\Delta} \quad (8.4.6)$$

and as $h=0$ (which is often the case in front of screens):

$$y'=\frac{y(z'+\Delta)}{z+\Delta} \quad (8.4.7)$$

If the table is placed on the axis of reference (the projection table is assimilated to the screen), then we have $z'=0$ which gives us:

$$y'=\frac{y\Delta}{z+\Delta} \quad (8.4.8)$$

and by proceeding in the same way:

$$x'=\frac{x\Delta}{z+\Delta} \quad (8.4.9)$$

Remark

The last two relations are those that we use to make 3D animations programmed in the Macromedia Flash 6.0 (see example further below) or in Javascript (see example also further below) on in any other programming language.

We see with the last two relations an identical term:

$$\boxed{P=\frac{\Delta}{z+\Delta}} \quad (8.4.10)$$

This term corresponds to the "depth" of the perspective.

In some textbooks, this depth is denoted by (simply factorization):

$$P=\frac{1}{1+\dfrac{z}{\Delta}} \quad (8.4.11)$$

If we consider two points (x_1,x_2 or y_1,y_2) visible from the surface of a volume seen by an observer and their respective distance Δx or Δy, these quantities are preserved if the two points coincide in the plane of the table because we have then:

$$P=1 \tag{8.4.12}$$

as $z=0,\forall\Delta$.

It is interesting to study what should be the value of the focal length for $z\neq 0$ to have $\Delta x'=\Delta x$ or $\Delta y'=\Delta y$. Thus, if we take the limit:

$$\lim_{\Delta\rightarrow+\infty} P=\frac{\Delta}{z+\Delta} \tag{8.4.13}$$

by applying l'Hôpital's rule (derivative of the numerator and denominator as seen in the section of Differential and Integral Calculus page 1461) and remembering that z is fixed, then:

$$\lim_{\Delta\rightarrow+\infty} P=\frac{\Delta}{z+\Delta}=\frac{\dfrac{\mathrm{d}}{\mathrm{d}\Delta}\Delta}{\dfrac{\mathrm{d}}{\mathrm{d}\Delta}(1+\Delta)}=\frac{1}{0+1}=1 \tag{8.4.14}$$

From this result we can conclude the following:

For any real distance between two points that do not coincide in the table but located in a same plane have an equal projection distance, it is necessary that the equations of the two lines that determine their intersection with table are parallel. This implies, since the observer is a convergent point, that we have to take away the observer to an infinite distance from the plane to keep the quantities projected on the table: this is the "orthogonal parallel projection" also sometimes named "orthographic parallel projection".

A very good example to see these results is to program in pseudo-3D on a computer.

✍**Example:**

There are many ways to do pseudo 3D with computers. The best known are technically OpenGL or DirectX and C++ but are not very easy to introduce ... so we'll see how to turn a pseudo-sphere in the projective space with Macromedia Flash 6.0 in order to show how apply the different theoretical elements presented above, but also to show that these are not the only tools available.

For this purpose, open the Macromedia Flash 6.0 software and save the new animation with the name *Circle.fla*:

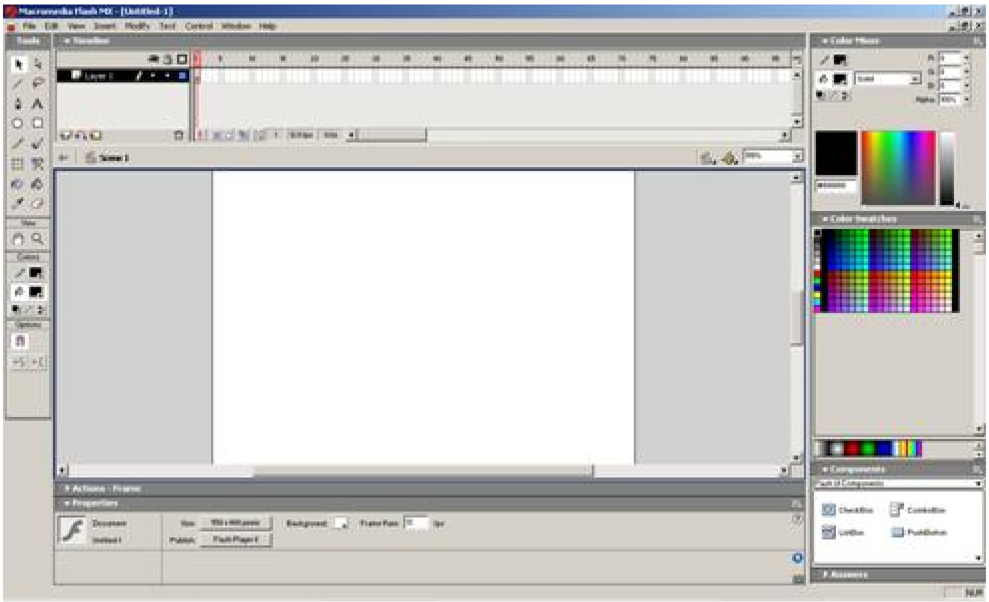


Figure 8.105: **Macromedia Flash 6.0 GUI**

With the **Circle** tool in the **Drawing** toolbar, draw a respectable size disc in the animation area:

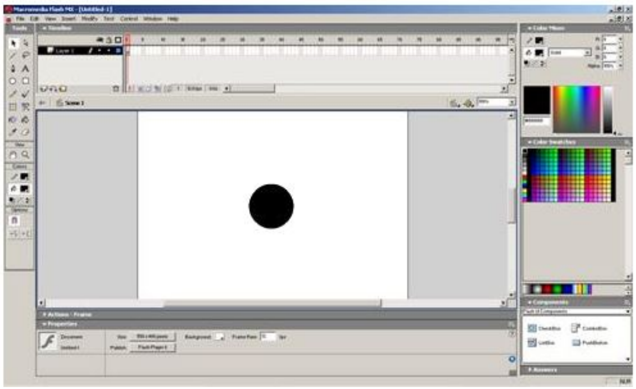
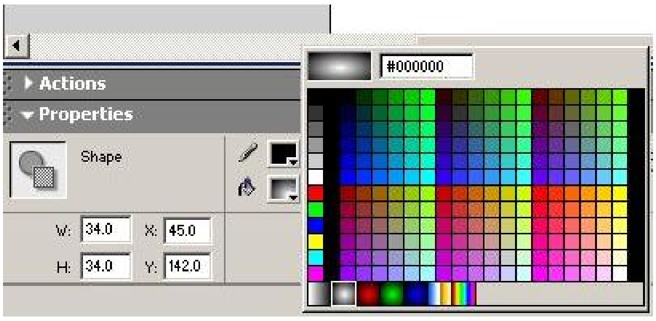


Figure 8.106: **Disc in the animation area**

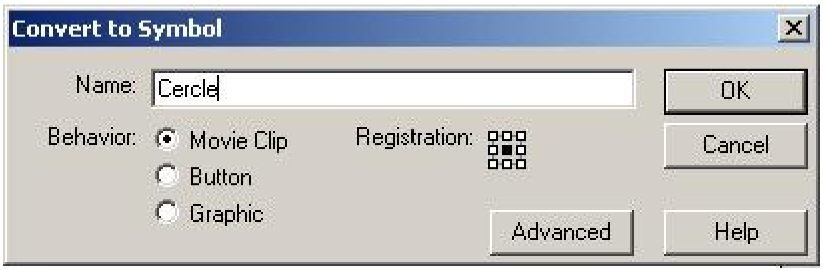
After having selected the circle with the **Fill** tool choose a *Radial Gradient*:



W: 34.0 X: 45.0
H: 34.0 Y: 142.0
#000000

Figure 8.107: **Color palette to fill-in the circle**

Right-click on the circle and choose the option **Convert to Symbol** and enter the information as presented below:

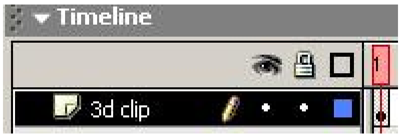

Convert to Symbol
OK
Cancel
Advanced
Help
Registration:
Button
Graphic

Name: Cercle

Behavior: ● Movie Clip

Figure 8.108: **Conversion dialogue box from of Object to Symbol**

Rename the layer where lies your animation clip with the name *3d clip*:

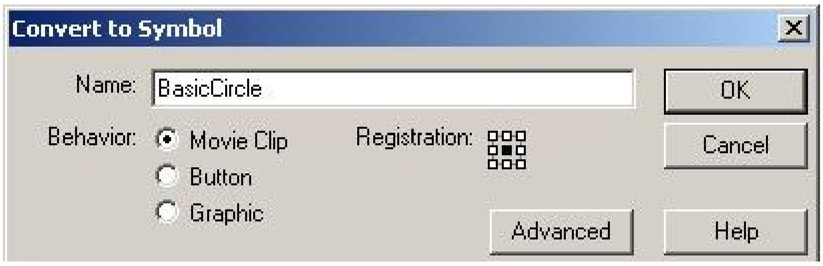


3d clip

Figure 8.109: **Renaming the animation layer**

Afterwards double-click on your circle to enter your animation Clip.

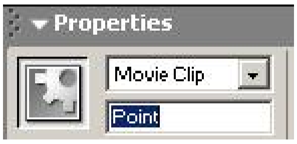
There select the circle again, right-click it and select **Convert to Symbol**:


Convert to Symbol
OK
Cancel
Advanced
Help
Registration:
Button
Graphic

Name: BasicCircle

Behavior: ● Movie Clip

Then, in the properties of the circle enter the name *Point*:

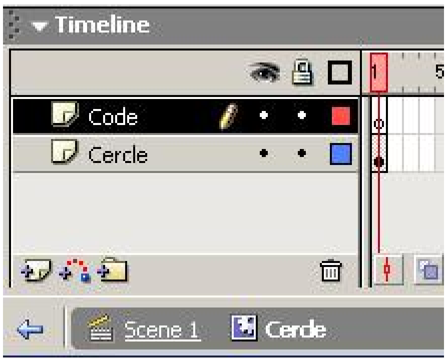

Properties

Movie Clip

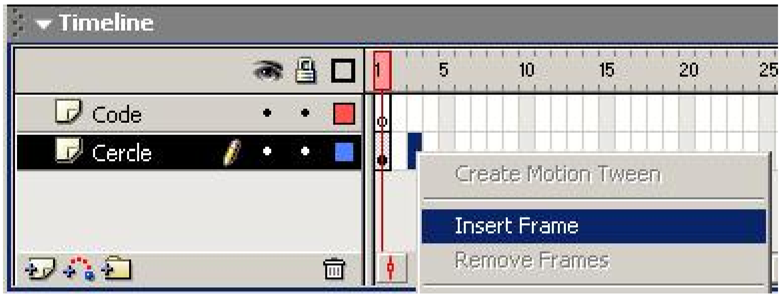
Point

Now, in the animation Clip named *Cercle*, we will insert three frames: the first to define the mathematical functions necessary for recalculating the variables, the second calling functions, the third that allows us to loop and go the second frame indefinitely.

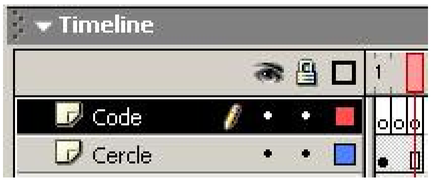
To make things in an almost clean way, we will create a second layer (by renaming the first one that contains our circle with the name *Cercle*) that we will rename *Code*:


Timeline
Code
Cercle
Scene 1
Cercle

Do a right-click on the third image of the layer containing our circle:


Timeline
Code
Cercle
Create Motion Tween
Insert Frame
Remove Frames

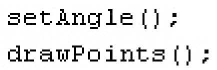
and choose the option **Insert Frame** and for the layer *Code* made almost the same, but by choosing **Insert Key Frame**. You should then get the following display:


Timeline
Code
Cercle

Then by selecting the first image of the layer *Code* activates the display of the Action to insert the following code:

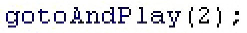
```
Radius = 100;
degrees = 0;
Xcenter=0;
Ycenter=0;
speed=2;

function setAngle () {
     Angle = degrees * (Math.PI / 180);
     degrees = degrees + speed;
}

function drawPoints () {
     Point._x = Radius * Math.cos(Angle)+Xcenter;
     Point._y = Radius * Math.sin(Angle)+Ycenter;
}
```

Repeat the same with the second image of the layer *Code*, but putting this time inside it:



```
setAngle();
drawPoints();
```

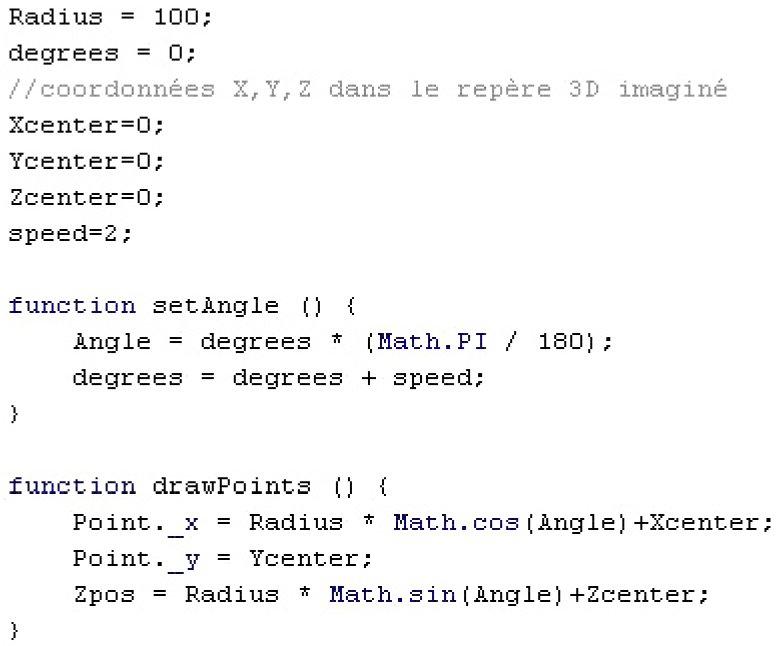
and finally doing the same with the third image of the same layer, but by putting inside it:



```
gotoAndPlay(2);
```

We then get the following animated result (the animation is only visible on a PDF reader with Adobe Flash installed and activated otherwise go see here https://vimeo.com/578782166):

We will now make appear the z-axis by fixing y (and therefore by moving z). Obviously, we will not see anything happen with regard to z until we define the homographic projection and this because a computer is unable to show basically the concept of depth... The code is then written:



```
Radius = 100;
degrees = 0;
//coordonnées X,Y,Z dans le repère 3D imaginé
Xcenter=0;
Ycenter=0;
Zcenter=0;
speed=2;

function setAngle () {
    Angle = degrees * (Math.PI / 180);
    degrees = degrees + speed;
}

function drawPoints () {
    Point._x = Radius * Math.cos(Angle)+Xcenter;
    Point._y = Ycenter;
    Zpos = Radius * Math.sin(Angle)+Zcenter;
}
```

The calculation of `Zpos` will enable us further below to calculate the depth of the movement of the object along the z-axis and this is where the homographic projection will play a role.

We then get a pseudo-sphere which rotates around an axis in a plane perpendicular to the screen. This is why we see the pseudo-sphere make left / right trips (the concept of distance is not yet present because of a lack of presence of the depth factor):

The animation above will run for people having a PDF reader with Adobe Flash player installed and activated (otherwise see here: https://vimeo.com/578782447).

Now we will use the relations:

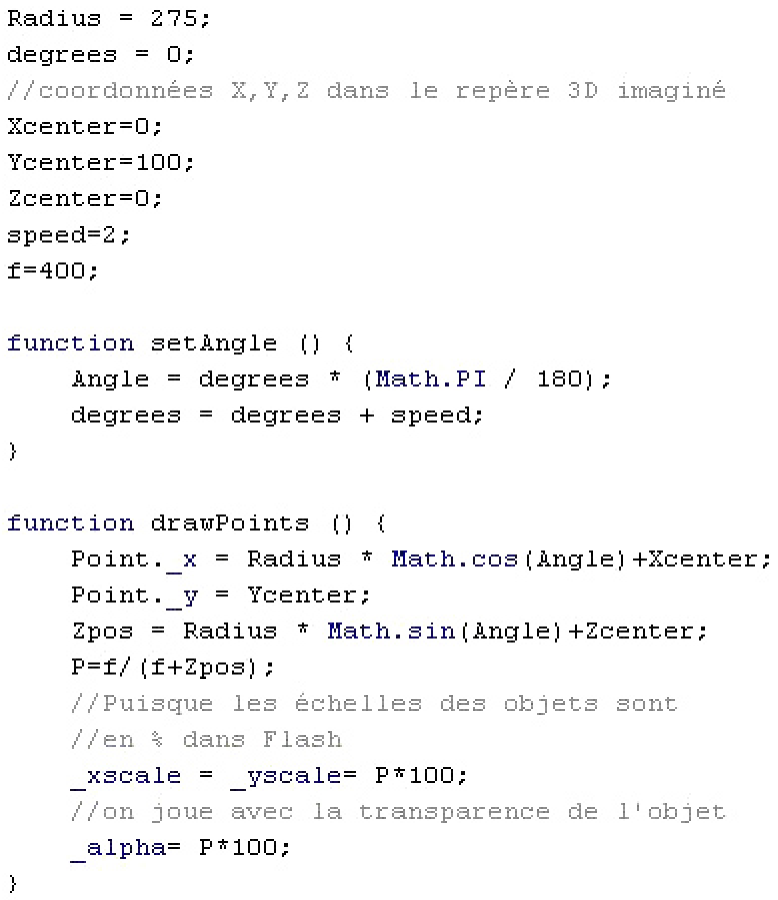
$$\begin{aligned}x'(z)&=xP=x\frac{f}{z+f}\\ y'(z)&=yP=y\frac{f}{z+f}\end{aligned}\tag{8.4.15}$$

proved earlier. If we seek to represent the depth of any point of the projection table, it is the distance ratio between two points in the table that will interest us to determine the scaling factor:

$$\begin{aligned}\frac{x'_2(z)-x'_1(z)}{x_2-x_1}\%&=\frac{f}{z+f}\cdot 100=P\cdot 100\\ \frac{y'_2(z)-y'_1(z)}{y_2-y_1}\%&=\frac{f}{z+f}\cdot 100=P\cdot 100\end{aligned}\tag{8.4.16}$$

We will have therefore to apply this result as defining the scale of the projection table.

We then have the following code where the depth P plays on the height and width of the animated surface of the instance *Point* of our pseudo-sphere:



```
Radius = 275;
degrees = 0;
//coordonnées X,Y,Z dans le repère 3D imaginé
Xcenter=0;
Ycenter=100;
Zcenter=0;
speed=2;
f=400;

function setAngle () {
    Angle = degrees * (Math.PI / 180);
    degrees = degrees + speed;
}

function drawPoints () {
    Point._x = Radius * Math.cos(Angle)+Xcenter;
    Point._y = Ycenter;
    Zpos = Radius * Math.sin(Angle)+Zcenter;
    P=f/(f+Zpos);
    //Puisque les échelles des objets sont
    //en % dans Flash
    _xscale = _yscale= P*100;
    //on joue avec la transparence de l'objet
    _alpha= P*100;
}
```

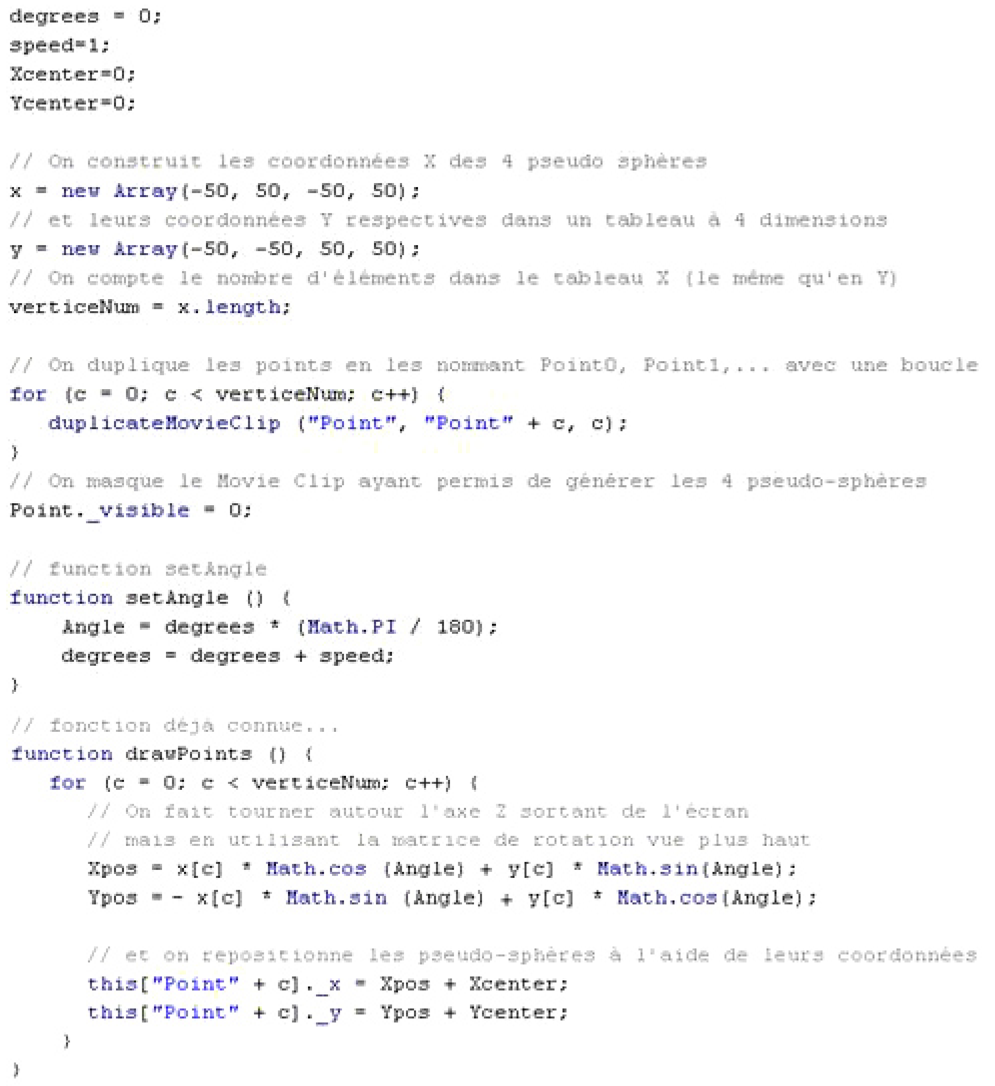
Which gives us:

and obviously it works regardless of the parametric equation of the path followed! We can then copy this instance animation, change the height, the starting angle for the get the 4 vertices of a cube rotating in space.

The animation above will run for people having a PDF reader with Adobe Flash player installed and activated (otherwise see here: `https://vimeo.com/578783079`).

This is the next step:

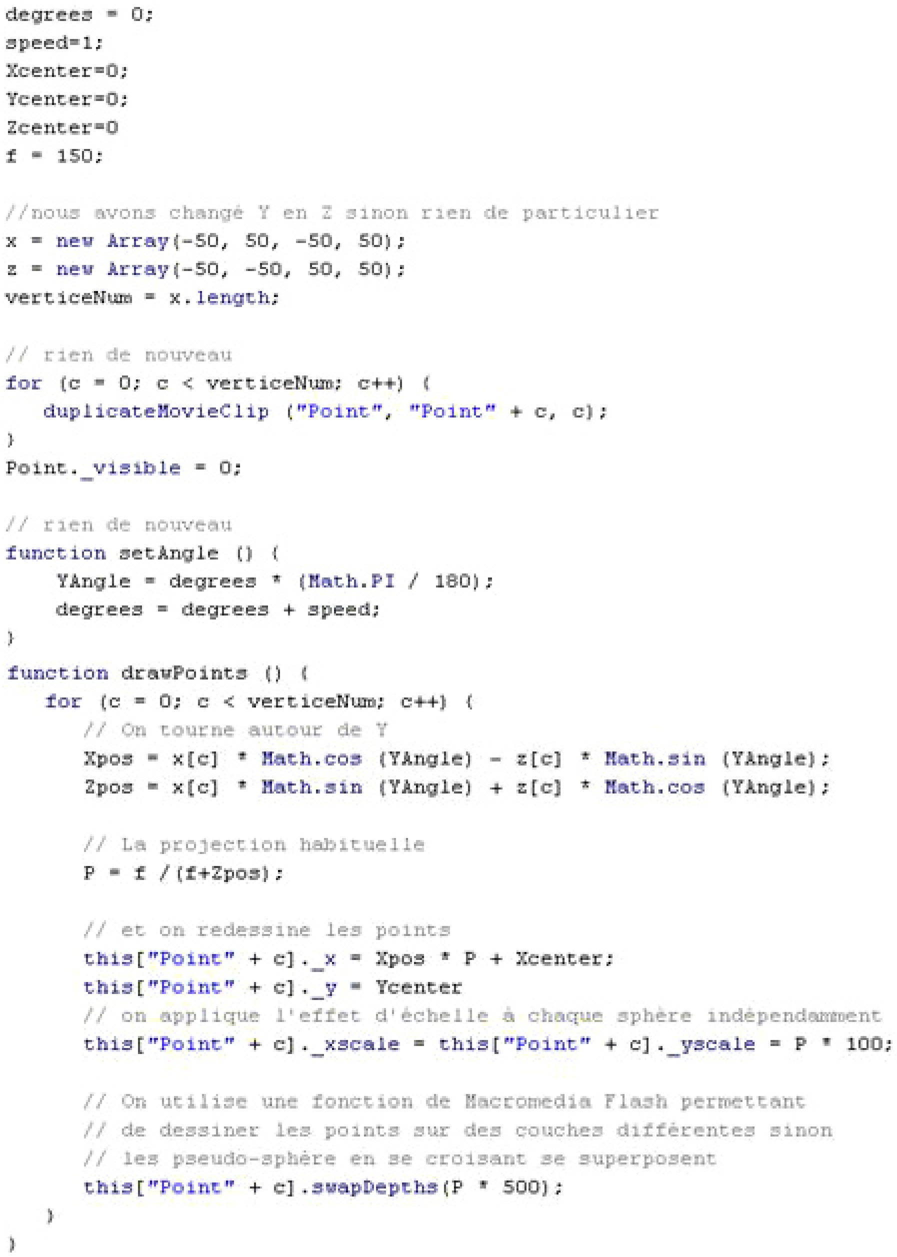
Indeed, let us change our code as shown below for 4 pseudo-spheres rotating around an imaginary z-axis going out of the screen (we use the rotation matrices proved in the section of Euclidean Geometry) and sorry for the comments in French as I did first the code for french readers:



```
degrees = 0;
speed=1;
Xcenter=0;
Ycenter=0;

// On construit les coordonnées X des 4 pseudo sphères
x = new Array(-50, 50, -50, 50);
// et leurs coordonnées Y respectives dans un tableau à 4 dimensions
y = new Array(-50, -50, 50, 50);
// On compte le nombre d'éléments dans le tableau X (le même qu'en Y)
verticeNum = x.length;

// On duplique les points en les nommant Point0, Point1,... avec une boucle
for (c = 0; c < verticeNum; c++) {
   duplicateMovieClip ("Point", "Point" + c, c);
}
// On masque le Movie Clip ayant permis de générer les 4 pseudo-sphères
Point._visible = 0;

// function setAngle
function setAngle () {
    Angle = degrees * (Math.PI / 180);
    degrees = degrees + speed;
}
// fonction déjà connue...
function drawPoints () {
   for (c = 0; c < verticeNum; c++) {
      // On fait tourner autour l'axe Z sortant de l'écran
      // mais en utilisant la matrice de rotation vue plus haut
      Xpos = x[c] * Math.cos (Angle) + y[c] * Math.sin(Angle);
      Ypos = - x[c] * Math.sin (Angle) + y[c] * Math.cos(Angle);

      // et on repositionne les pseudo-sphères à l'aide de leurs coordonnées
      this["Point" + c]._x = Xpos + Xcenter;
      this["Point" + c]._y = Ypos + Ycenter;
    }
}
```

This gives us ... four charming pseudo-spheres turning around a common center (the animation below will run for people having a PDF reader with Adobe Flash player installed and activated otherwise see here: https://vimeo.com/578783905):

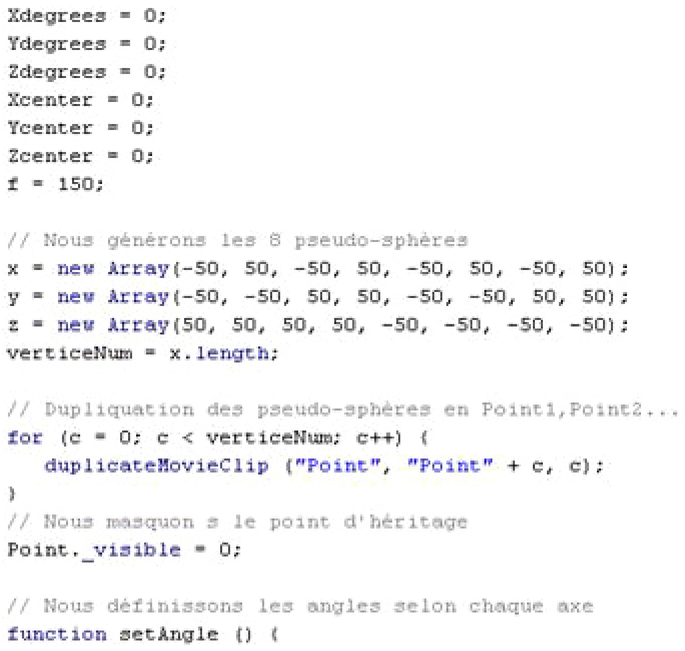
Now we reverse y and z again and apply the homographic projection:



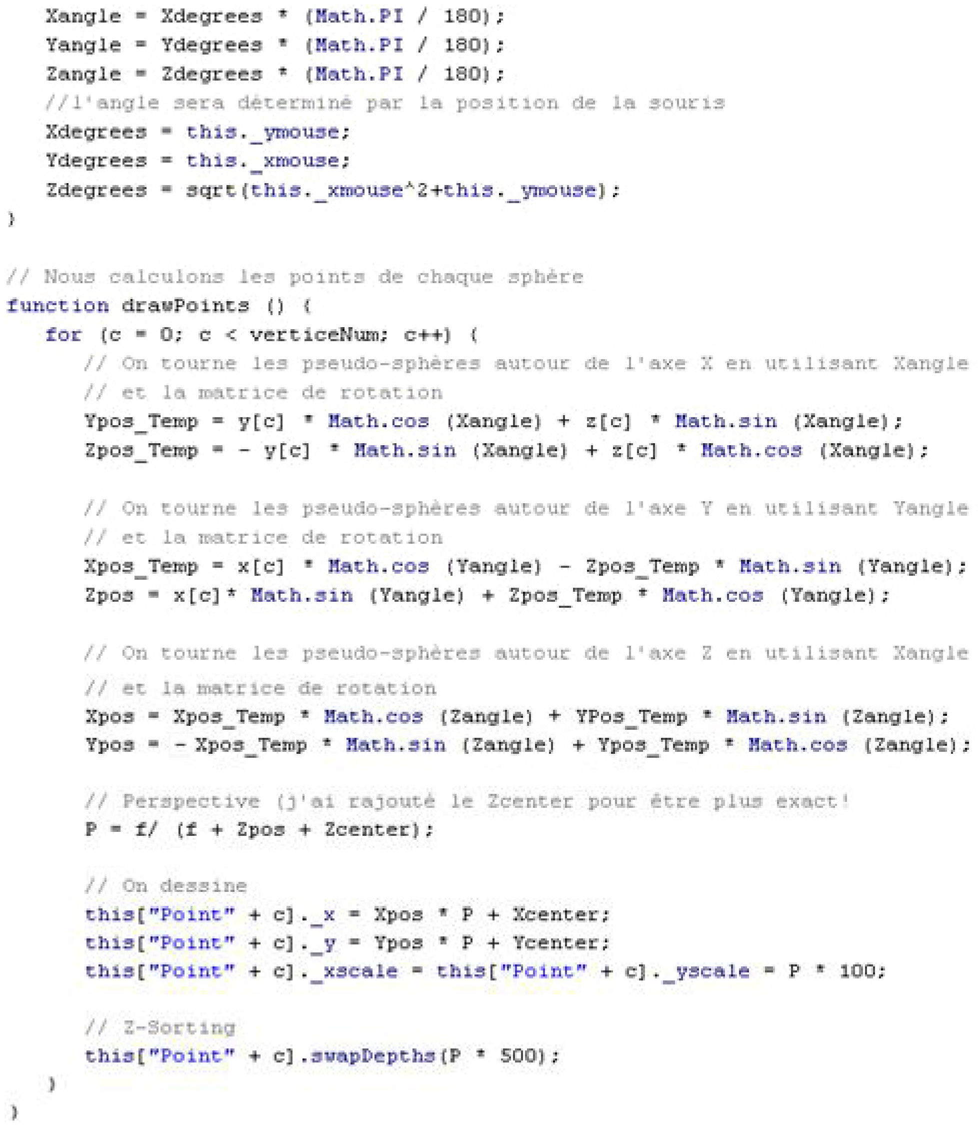
```
degrees = 0;
speed=1;
Xcenter=0;
Ycenter=0;
Zcenter=0
f = 150;

//nous avons changé Y en Z sinon rien de particulier
x = new Array(-50, 50, -50, 50);
z = new Array(-50, -50, 50, 50);
verticeNum = x.length;

// rien de nouveau
for (c = 0; c < verticeNum; c++) {
   duplicateMovieClip ("Point", "Point" + c, c);
}
Point._visible = 0;

// rien de nouveau
function setAngle () {
    YAngle = degrees * (Math.PI / 180);
    degrees = degrees + speed;
}
function drawPoints () {
   for (c = 0; c < verticeNum; c++) {
      // On tourne autour de Y
      Xpos = x[c] * Math.cos (YAngle) - z[c] * Math.sin (YAngle);
      Zpos = x[c] * Math.sin (YAngle) + z[c] * Math.cos (YAngle);

      // La projection habituelle
      P = f /(f+Zpos);

      // et on redessine les points
      this["Point" + c]._x = Xpos * P + Xcenter;
      this["Point" + c]._y = Ycenter
      // on applique l'effet d'échelle à chaque sphère indépendamment
      this["Point" + c]._xscale = this["Point" + c]._yscale = P * 100;

      // On utilise une fonction de Macromedia Flash permettant
      // de dessiner les points sur des couches différentes sinon
      // les pseudo-sphère en se croisant se superposent
      this["Point" + c].swapDepths(P * 500);
   }
}
```

and we get (the animation below will run for people having a PDF reader with Adobe Flash player installed and activated otherwise see here: https://vimeo.com/578784812):

For the remaining part, we will generate 8 pseudo-sphere and instead to turn them always about the same axis, we will rotate around three axes x, y or z using the variables `XAngle`, `YAngle` or `ZAngle` and rotation matrices about each of these respective axes:



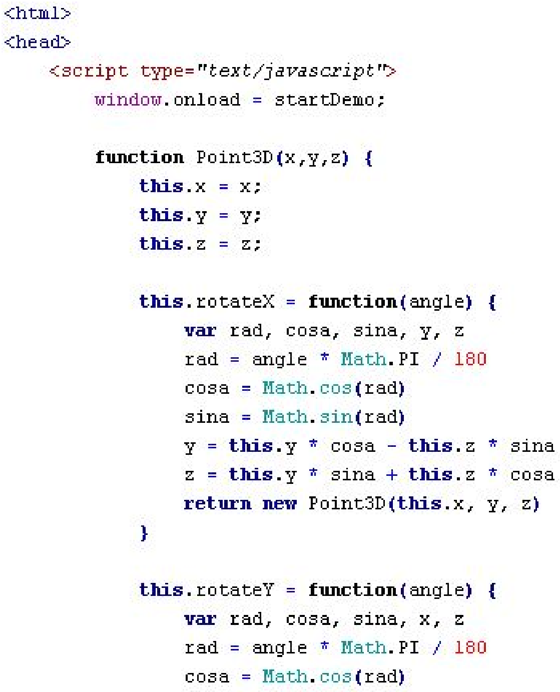
```
Xdegrees = 0;
Ydegrees = 0;
Zdegrees = 0;
Xcenter = 0;
Ycenter = 0;
Zcenter = 0;
f = 150;

// Nous générons les 8 pseudo-sphères
x = new Array(-50, 50, -50, 50, -50, 50, -50, 50);
y = new Array(-50, -50, 50, 50, -50, -50, 50, 50);
z = new Array(50, 50, 50, 50, -50, -50, -50, -50);
verticeNum = x.length;

// Dupliquation des pseudo-sphères en Point1,Point2...
for (c = 0; c < verticeNum; c++) {
   duplicateMovieClip ("Point", "Point" + c, c);
}
// Nous masquon s le point d'héritage
Point._visible = 0;

// Nous définissons les angles selon chaque axe
function setAngle () {
```



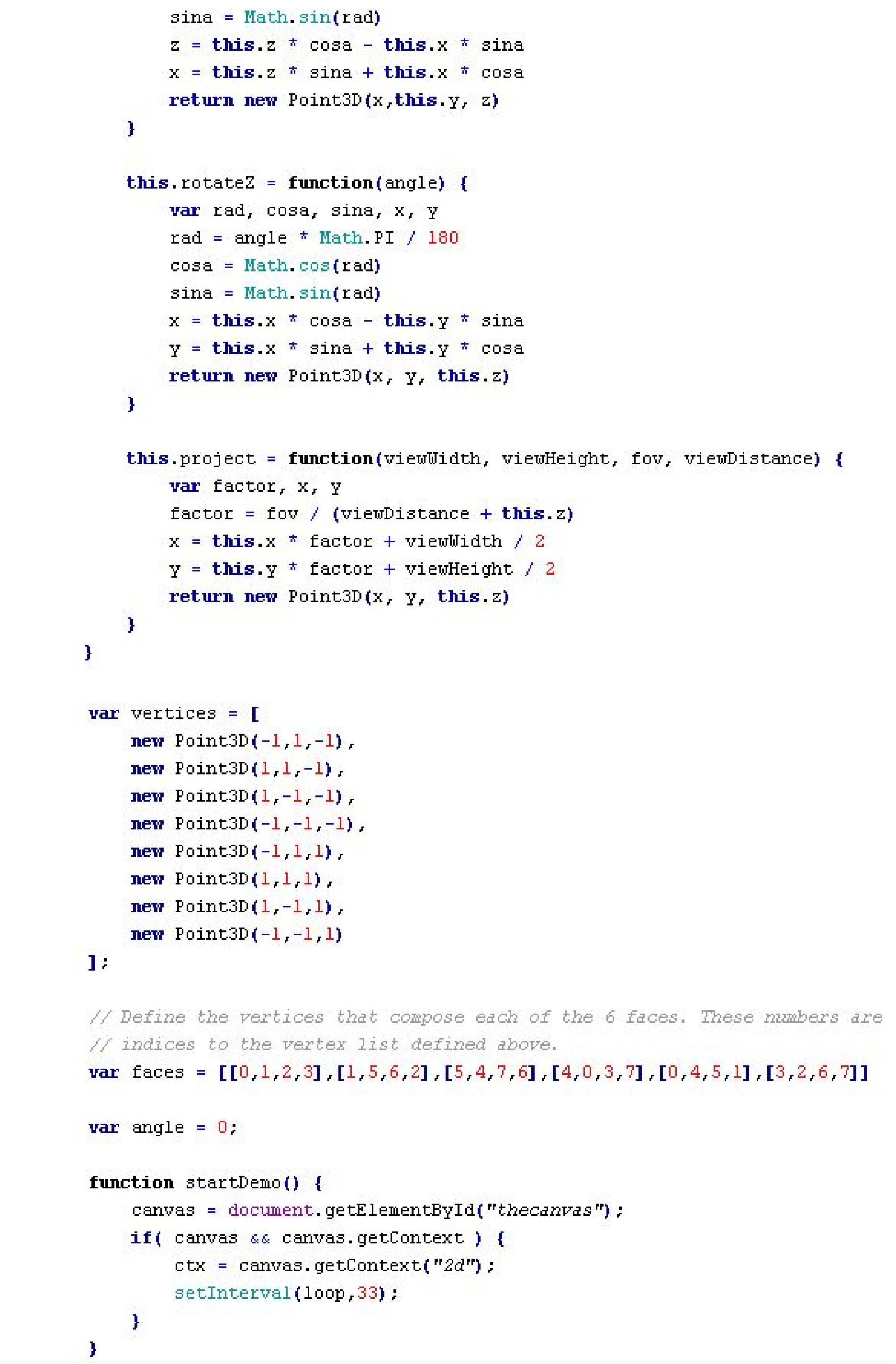
```
    Xangle = Xdegrees * (Math.PI / 180);
    Yangle = Ydegrees * (Math.PI / 180);
    Zangle = Zdegrees * (Math.PI / 180);
    //l'angle sera déterminé par la position de la souris
    Xdegrees = this._ymouse;
    Ydegrees = this._xmouse;
    Zdegrees = sqrt(this._xmouse^2+this._ymouse);
}

// Nous calculons les points de chaque sphère
function drawPoints () {
    for (c = 0; c < verticeNum; c++) {
        // On tourne les pseudo-sphères autour de l'axe X en utilisant Xangle
        // et la matrice de rotation
        Ypos_Temp = y[c] * Math.cos (Xangle) + z[c] * Math.sin (Xangle);
        Zpos_Temp = - y[c] * Math.sin (Xangle) + z[c] * Math.cos (Xangle);

        // On tourne les pseudo-sphères autour de l'axe Y en utilisant Yangle
        // et la matrice de rotation
        Xpos_Temp = x[c] * Math.cos (Yangle) - Zpos_Temp * Math.sin (Yangle);
        Zpos = x[c]* Math.sin (Yangle) + Zpos_Temp * Math.cos (Yangle);

        // On tourne les pseudo-sphères autour de l'axe Z en utilisant Xangle
        // et la matrice de rotation
        Xpos = Xpos_Temp * Math.cos (Zangle) + YPos_Temp * Math.sin (Zangle);
        Ypos = - Xpos_Temp * Math.sin (Zangle) + Ypos_Temp * Math.cos (Zangle);

        // Perspective (j'ai rajouté le Zcenter pour être plus exact!
        P = f/ (f + Zpos + Zcenter);

        // On dessine
        this["Point" + c]._x = Xpos * P + Xcenter;
        this["Point" + c]._y = Ypos * P + Ycenter;
        this["Point" + c]._xscale = this["Point" + c]._yscale = P * 100;

        // Z-Sorting
        this["Point" + c].swapDepths(P * 500);
    }
}
```

and here is the final result (the animation below will run for people having a PDF reader with Adobe Flash player installed and activated otherwise see here: `https://vimeo.com/578785768`):

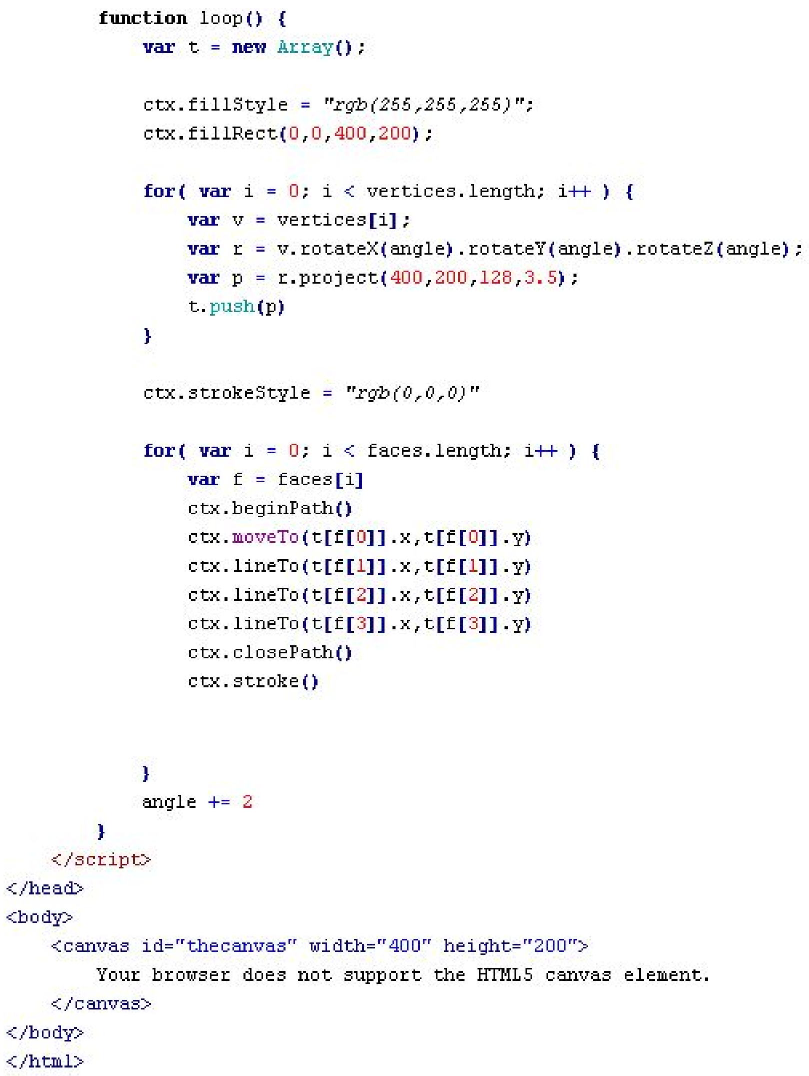
Since the Adobe Flash technology is virtually dead since the time I wrote the code above... an anonymous reader rewrote the above code again into WebGL and here is the corresponding code (if you want to give the same example in other languages don't hesitate):



```
<html>
<head>
    <script type="text/javascript">
        window.onload = startDemo;

        function Point3D(x,y,z) {
            this.x = x;
            this.y = y;
            this.z = z;

            this.rotateX = function(angle) {
                var rad, cosa, sina, y, z
                rad = angle * Math.PI / 180
                cosa = Math.cos(rad)
                sina = Math.sin(rad)
                y = this.y * cosa - this.z * sina
                z = this.y * sina + this.z * cosa
                return new Point3D(this.x, y, z)
            }

            this.rotateY = function(angle) {
                var rad, cosa, sina, x, z
                rad = angle * Math.PI / 180
                cosa = Math.cos(rad)
```



```
        sina = Math.sin(rad)
        z = this.z * cosa - this.x * sina
        x = this.z * sina + this.x * cosa
        return new Point3D(x,this.y, z)
    }

    this.rotateZ = function(angle) {
        var rad, cosa, sina, x, y
        rad = angle * Math.PI / 180
        cosa = Math.cos(rad)
        sina = Math.sin(rad)
        x = this.x * cosa - this.y * sina
        y = this.x * sina + this.y * cosa
        return new Point3D(x, y, this.z)
    }

    this.project = function(viewWidth, viewHeight, fov, viewDistance) {
        var factor, x, y
        factor = fov / (viewDistance + this.z)
        x = this.x * factor + viewWidth / 2
        y = this.y * factor + viewHeight / 2
        return new Point3D(x, y, this.z)
    }
}

var vertices = [
    new Point3D(-1,1,-1),
    new Point3D(1,1,-1),
    new Point3D(1,-1,-1),
    new Point3D(-1,-1,-1),
    new Point3D(-1,1,1),
    new Point3D(1,1,1),
    new Point3D(1,-1,1),
    new Point3D(-1,-1,1)
];

// Define the vertices that compose each of the 6 faces. These numbers are
// indices to the vertex list defined above.
var faces = [[0,1,2,3],[1,5,6,2],[5,4,7,6],[4,0,3,7],[0,4,5,1],[3,2,6,7]]

var angle = 0;

function startDemo() {
    canvas = document.getElementById("thecanvas");
    if( canvas && canvas.getContext ) {
        ctx = canvas.getContext("2d");
        setInterval(loop,33);
    }
}
```



```
        function loop() {
            var t = new Array();

            ctx.fillStyle = "rgb(255,255,255)";
            ctx.fillRect(0,0,400,200);

            for( var i = 0; i < vertices.length; i++ ) {
                var v = vertices[i];
                var r = v.rotateX(angle).rotateY(angle).rotateZ(angle);
                var p = r.project(400,200,128,3.5);
                t.push(p)
            }

            ctx.strokeStyle = "rgb(0,0,0)"

            for( var i = 0; i < faces.length; i++ ) {
                var f = faces[i]
                ctx.beginPath()
                ctx.moveTo(t[f[0]].x,t[f[0]].y)
                ctx.lineTo(t[f[1]].x,t[f[1]].y)
                ctx.lineTo(t[f[2]].x,t[f[2]].y)
                ctx.lineTo(t[f[3]].x,t[f[3]].y)
                ctx.closePath()
                ctx.stroke()


            }
            angle += 2
        }
    </script>
</head>
<body>
    <canvas id="thecanvas" width="400" height="200">
        Your browser does not support the HTML5 canvas element.
    </canvas>
</body>
</html>
```

8.4.1.2 Images of Straight Lines

Let us determine from the previous results, the image of a straight line parallel to the ground line (hence the x-axis) of the table XY. In this case, we have:

$$z=a \qquad y=b \tag{8.4.17}$$

where a and b are constants, and for any value of h. This gives us from the relations obtained previously:

$$\begin{aligned} x' &= \frac{\Delta}{\Delta+a}x \\ y' &= h+(b-h)\frac{\Delta}{\Delta+a}=\frac{ha+\Delta b}{\Delta a}=c^{te} \end{aligned} \tag{8.4.18}$$

So as we could expect it, a straight line parallel to the ground line becomes in perspective still a straight line but at a height y' in the XY plane parallel to our screen (it was quite intuitive to guess...).

For any line parallel to the z-axis of the screen (and therefore in its "depth"), we have:

$$x = a \qquad y = b \tag{8.4.19}$$

where a and b are constants for any value of h. Which gives us:

$$\left.\begin{aligned} x' &= \frac{\Delta}{\Delta+z}a \\ y' &= h+(b-h)\frac{\Delta}{\Delta+z} \end{aligned}\right\} : \frac{\Delta}{\Delta+z}=\frac{x'}{a} \Rightarrow y' = h+(b-h)\frac{x'}{a} \tag{8.4.20}$$

The straight lines of equation:

$$y' = h+(b-h)\frac{x'}{a} \tag{8.4.21}$$

all pass through the point $P(x'=0, y'=h)$ when $z \to +\infty$ which is the "main vanishing point" and through the point $P(x'=a, y'=b)$ when $z \to 0$ as shown below in the projection made in Adobe Photoshop (the horizontal lines corresponding to the ground live have been added to give the effect of perspective):

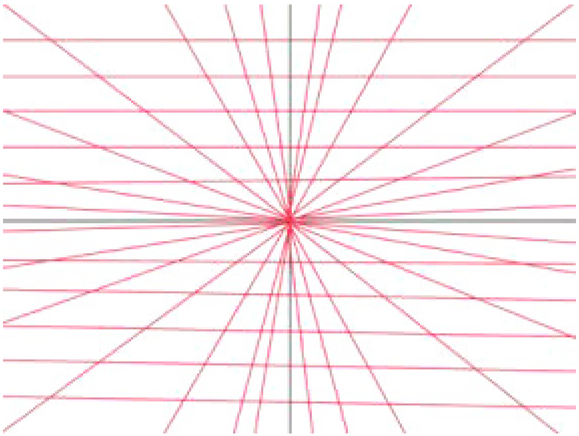
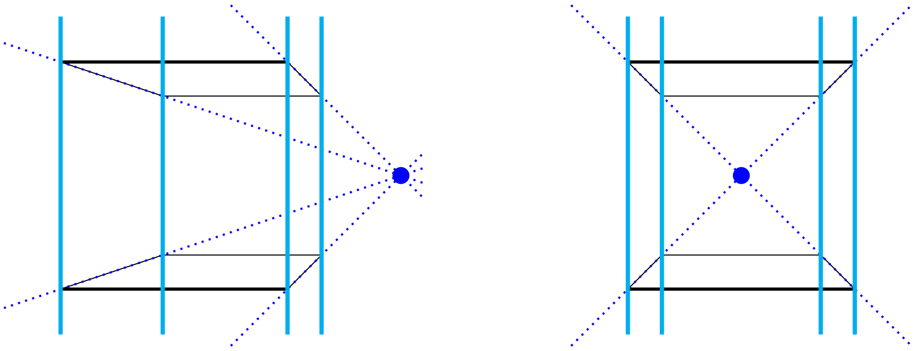


Figure 8.110: **Example of a single vanishing point**

Or in a more academic way:



This is why the shadows from objects seen from the ground seem to diverge (and hence some scientifically illiterate people will claim that therefore the Sun is local...), when in reality seen from the top (in the photo below thanks to a drone) they are indeed parallel because the Sun is quite far (approximately 150 millions kilometres):

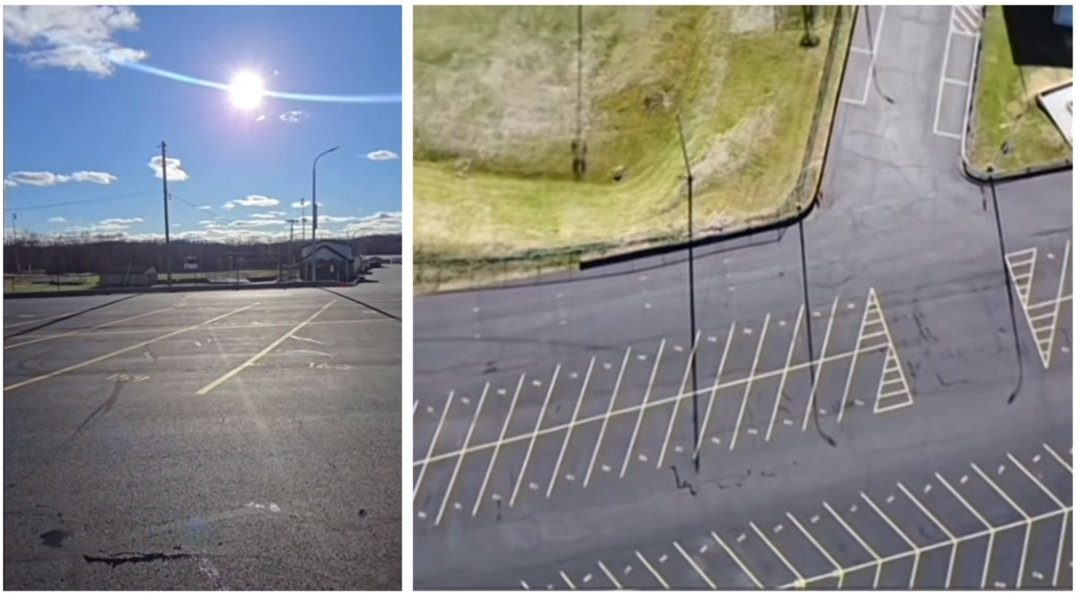


Figure 8.111: **Divergent shadows due to perspective**

From the figure above, we can define the concept of "departure angle" given by the figure below:

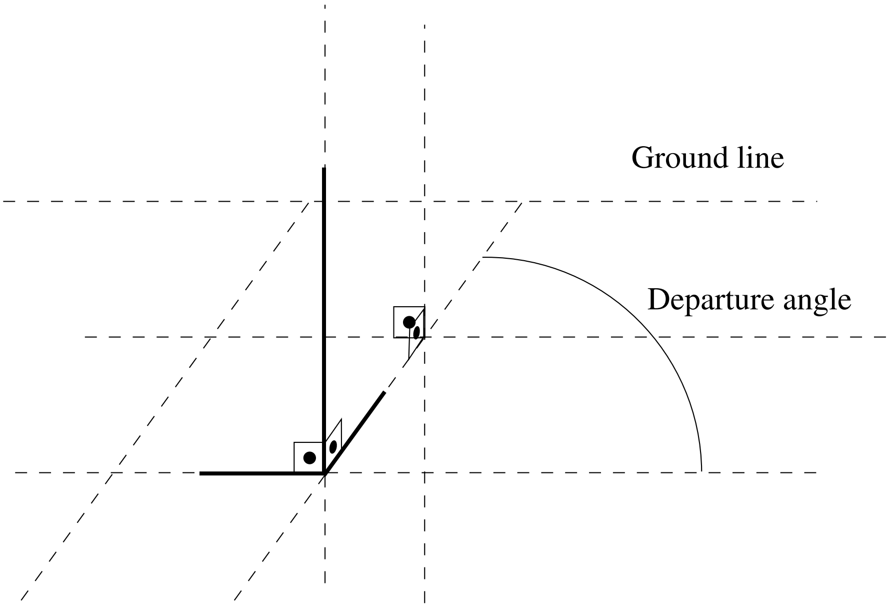
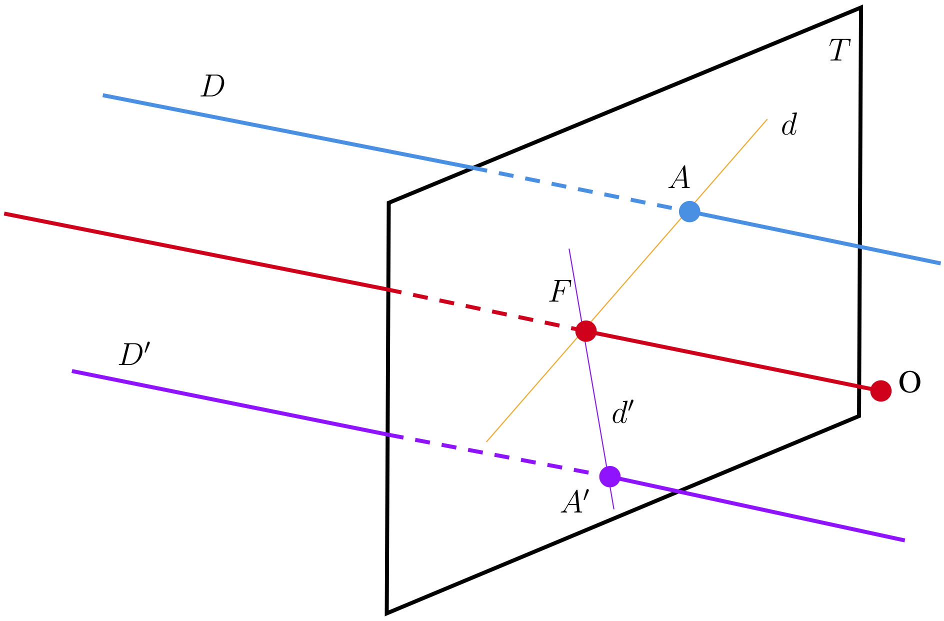


Figure 8.112: **Departure angle**

Another geometric representation can perhaps help to better understand the result. Let us recall that the vanishing point of a line (D) is the intersection point F of the table plane T with the line parallel to D passing through O. Two parallel lines (D) and (D') therefore have the same vanishing point (in the perspective point of view obviously!). As shown below:



If we denote by A the intersection point of (D) with T, the perspective drawing of D is the straight line $\overline{AF}$, intersection of T with the plane containing O and D. Since two parallel lines in a conical perspective point of view have the same vanishing point F, they are therefore represented by two intersecting straight lines on F.

For any straight line being in the XZ plane of the screen (i.e. in its "depth"), we have:

$$z=mx+p \tag{8.4.22}$$

and for $y=0$. Which gives us:

$$\begin{gathered}x'=\frac{\Delta}{\Delta+z}x\Rightarrow x=\frac{\Delta+z}{\Delta}x'\\ y'=h+(y-h)\frac{\Delta}{\Delta+z}=h-h\frac{\Delta}{\Delta+z}=\frac{hz}{\Delta+z}\end{gathered} \tag{8.4.23}$$

From the latter equation we deduce:

$$z=-\frac{\Delta}{y'-h}y' \tag{8.4.24}$$

by putting z in the expression of x:

$$x=\left(1+\frac{z}{\Delta}\right)x'=\left(1-\frac{y'}{y'-h}\right)x'=-\frac{hx'}{y'-h} \tag{8.4.25}$$

we replace x and y in the equation of the straight line $z=mx+p$, and once calculations and simplifications made, we find:

$$y'=\frac{mh}{\Delta+p}x'+\frac{ph}{\Delta+p} \tag{8.4.26}$$

Let us consider the particular case of straight lines passing through the opposite vertices of a tile, that is to say inclined at $\pm\pi/4$ therefore with a director coefficient (slope) of ± 1.

The image of these straight lines are then given by:

$$y'=\pm\frac{h}{\Delta+p}x'+\frac{ph}{\Delta+p} \tag{8.4.27}$$

If $y'=h$, then we have following the previous relation:

$$x'=\pm\Delta \tag{8.4.28}$$

this means that any projection of lines of director coefficient ±1 lying in the xy-plane for all x' constitutes secondary vanishing points situated on the same remote horizon line at an equal distance of the main vanishing point as shown below (in the context of a on Adobe Photoshop training, we added a cube is in this perspective):

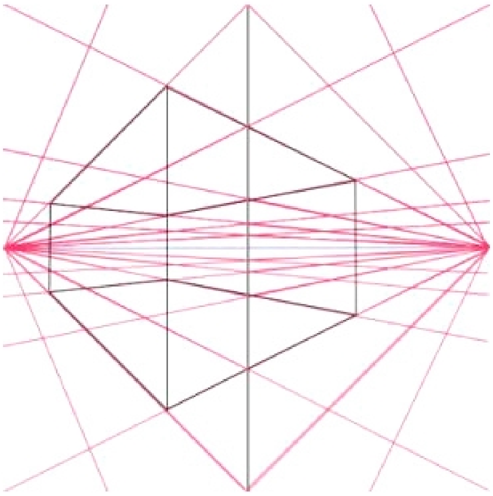
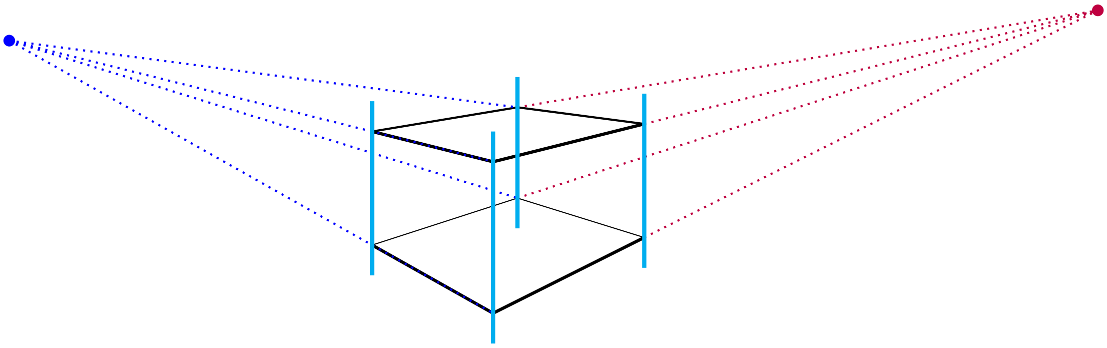


Figure 8.113: **Example of secondary vanishing point**

Or in a more academic way:



We see well in the figure above the symmetry about the vertical axis and the two vanishing point that are causing the tiles.

Now let us consider the straight lines parallel to the y-axis of the table (display) and their projection equation if $x=a$ and $z=b$:

$$\begin{aligned}x'&=\frac{\Delta}{\Delta+b}a=c^{te}\\ y'&=h+(y-h)\frac{\Delta}{\Delta+b}\end{aligned} \tag{8.4.29}$$

These right images thus remain straight lines parallel to the y axis. In other words, the right images remain parallel to straight "object" as shown below (for different positions of the observation point):

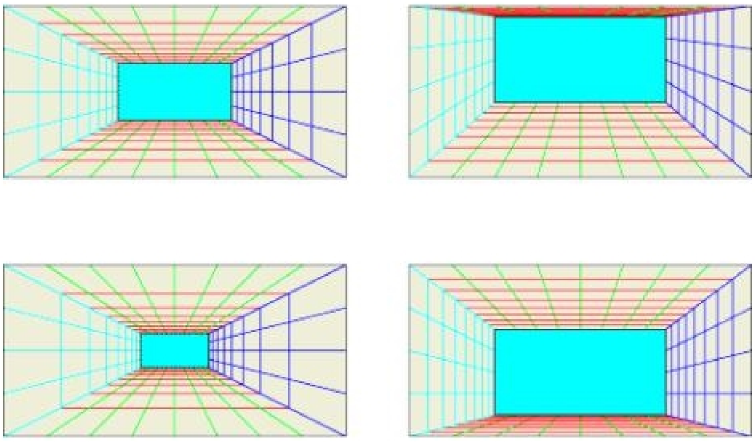


Figure 8.114: Different perspective based on the values of the constants

Let us take for example of this last result, segment of height H whose footer is on the horizontal plane (ground):

$$y = H \qquad x = a \tag{8.4.30}$$

The height of the segment is given by:

$$y' = H' = h + (H - h)\frac{\Delta}{\Delta + b} \tag{8.4.31}$$

Let us consider now $x = a$, that is to say the vertical columns of height H are aligned on the line $x = a$:

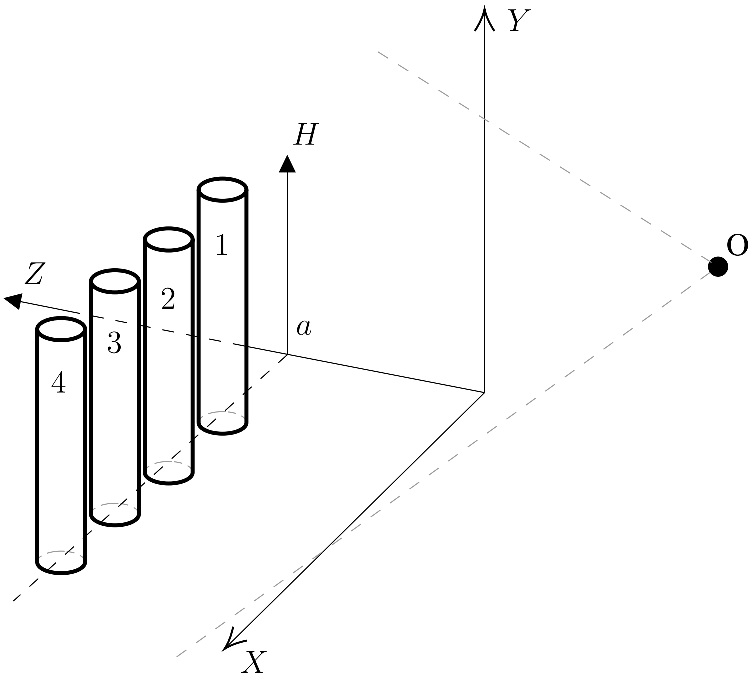


Figure 8.115: Aligned columns perspective problem

Let us calculate the coordinates of the vertices of these images:

$$\begin{aligned} x' &= \frac{\Delta}{\Delta+b}a \Rightarrow \frac{\Delta}{\Delta+b}=\frac{x'}{a} \\ y' &= H' = h+(H-h)\frac{\Delta}{\Delta+b}=h+\frac{H-h}{a}x \end{aligned} \tag{8.4.32}$$

This is obviously still the equation of a straight line and let us notice that all these line pass trough the point of coordinates $(x'=0, y'=h)$ that is just the main vanishing point!

So as we experiment in real life, if we consider the above figure corresponding to the previous developments when $H<h$ the columns seems to have their height that decrease as the are more far from the observer and if $h>H$ the columns height seems to increase.

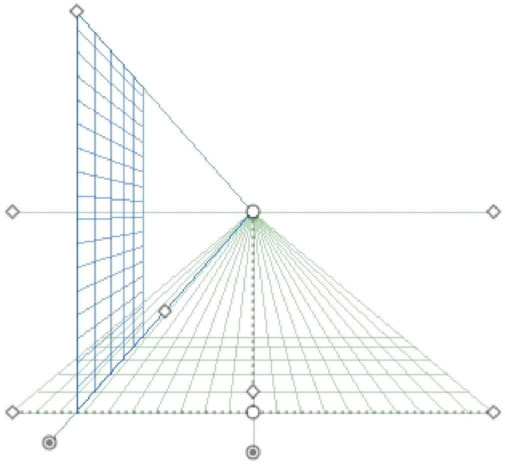
From the previous discussion, we have therefore also obviously a new method for rotating object in three dimensional space. Instead of rotating the object around different axes, we can imagine using the above equations to rotate the viewer around the object (that is a point of view...).

We have considered so far only the projective perspective on a plane. By the way, to work on any projection methods (sphere on a plane, plane on a sphere, sphere on sphere, anything about anything) it is simply enough to extend the analysis that we made above in a coordinate system suitable for the studied system (polar coordinates, cylindrical coordinates, spherical coordinates, ...). This is probably that way that some 3D simulation software use to project an image on a reflective surface such as a semi-transparent undulated glass at the difference that on computer a surface don't have an infinite number of points but a limited number of pixels and therefore the projection need some smoothing trough special interpolation techniques.

Remark

From the results we have obtained above, we can extract an interesting and intuitive conclusion For an observer of a picture or a table seethe image as it was originally, it must be placed at specific coordinates of the table (of the picture or of the table).

Software such as Adobe Illustrator offers since the early 12010s (holocene calendar) a tool to create perspectives for one vanishing point, for two vanishing points and three vanishing points:



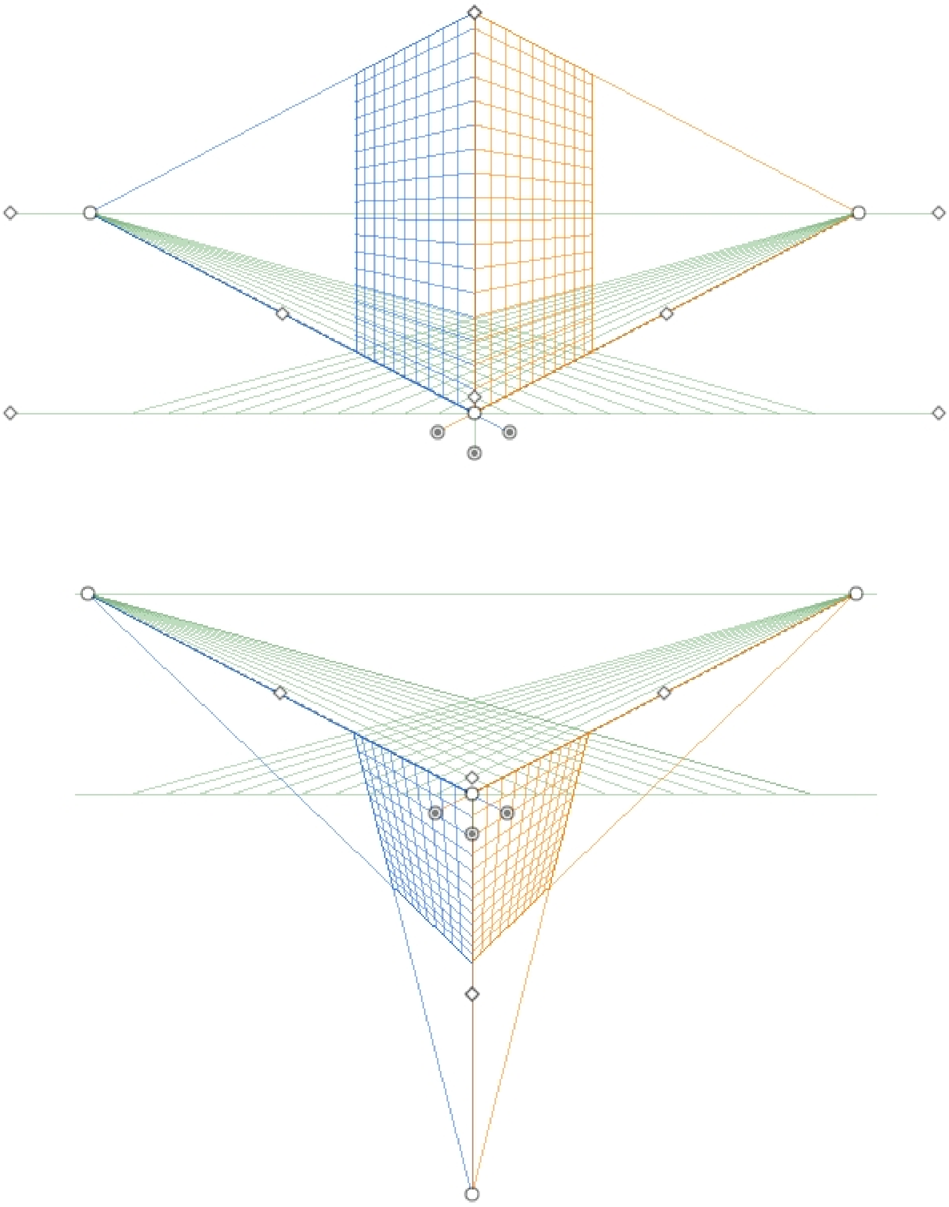
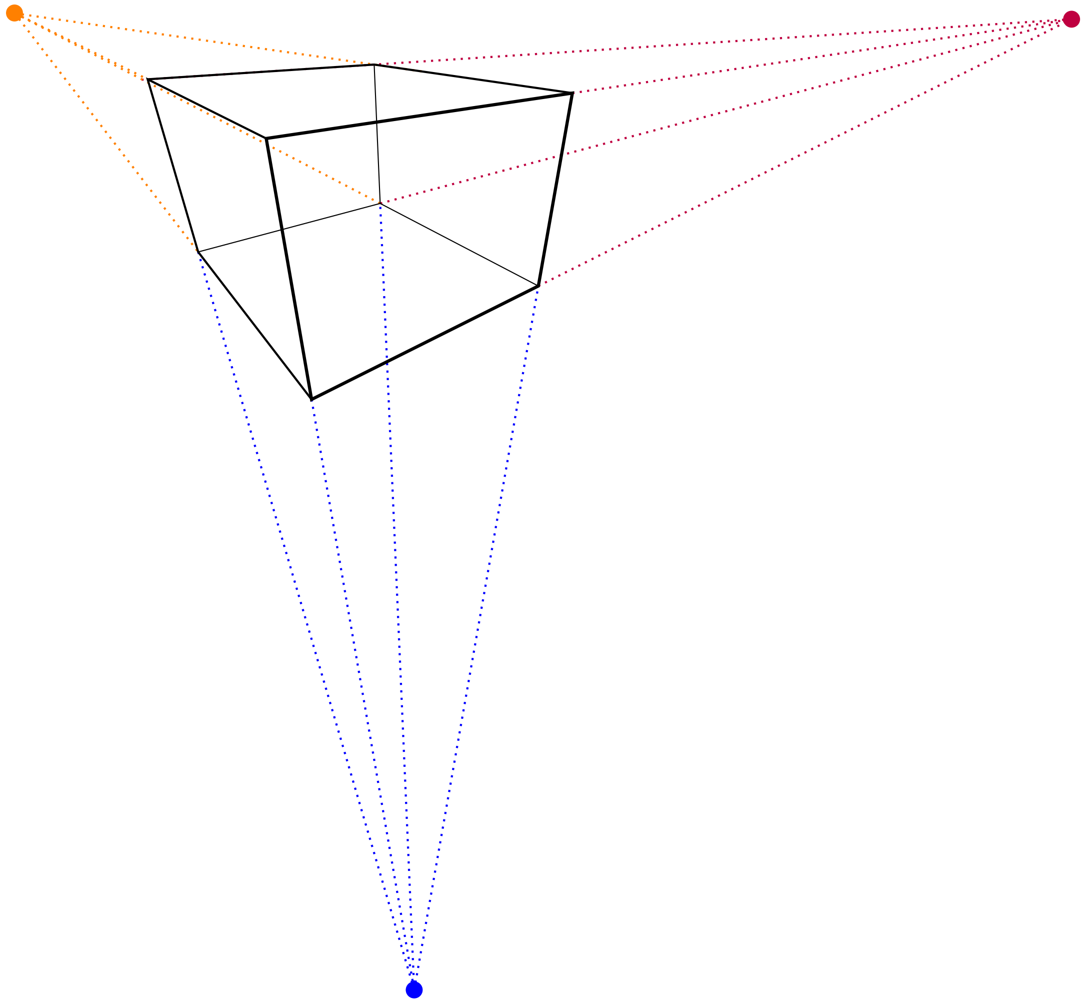
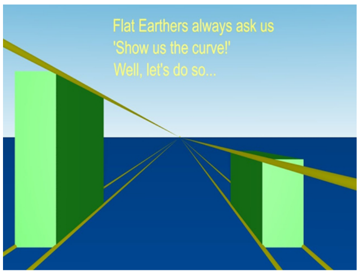


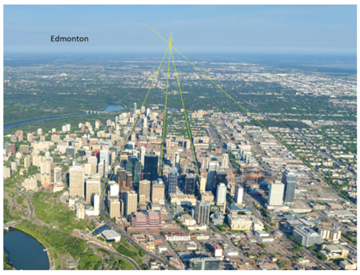
Figure 8.116: **One, two and three vanishing points in Adobe Illustrator**

Or in a more academic way:



And to finish a wink for all our favourite flat earthers (...):


Flat Earthers always ask us
'Show us the curve!'
Well, let's do so...



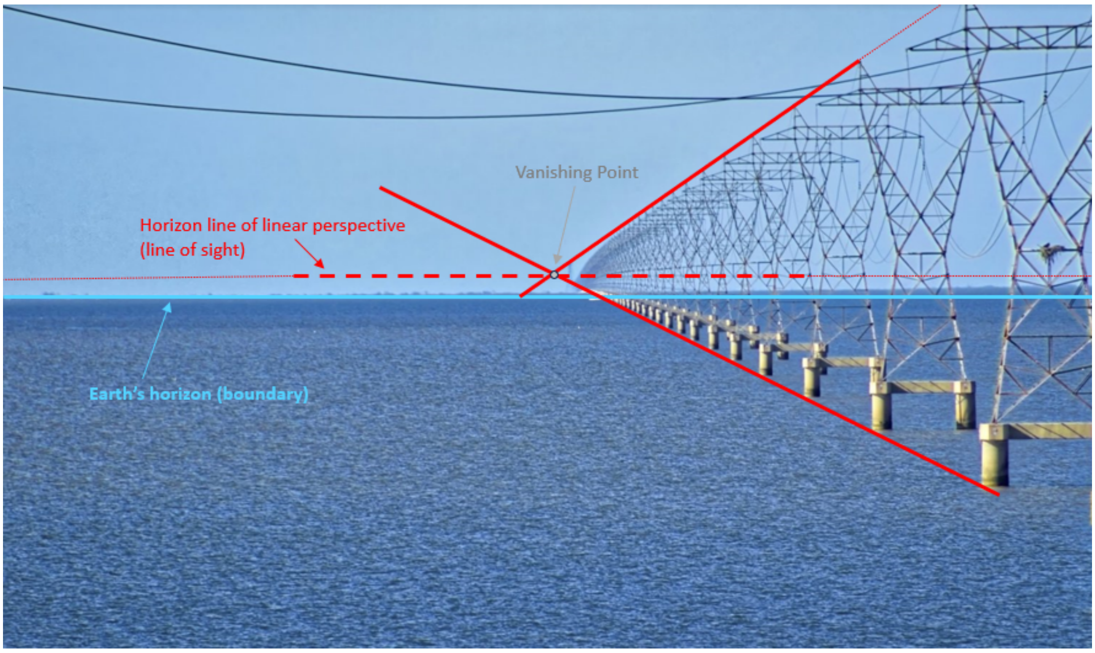
Medicine Head Alberta

Antwerp docks

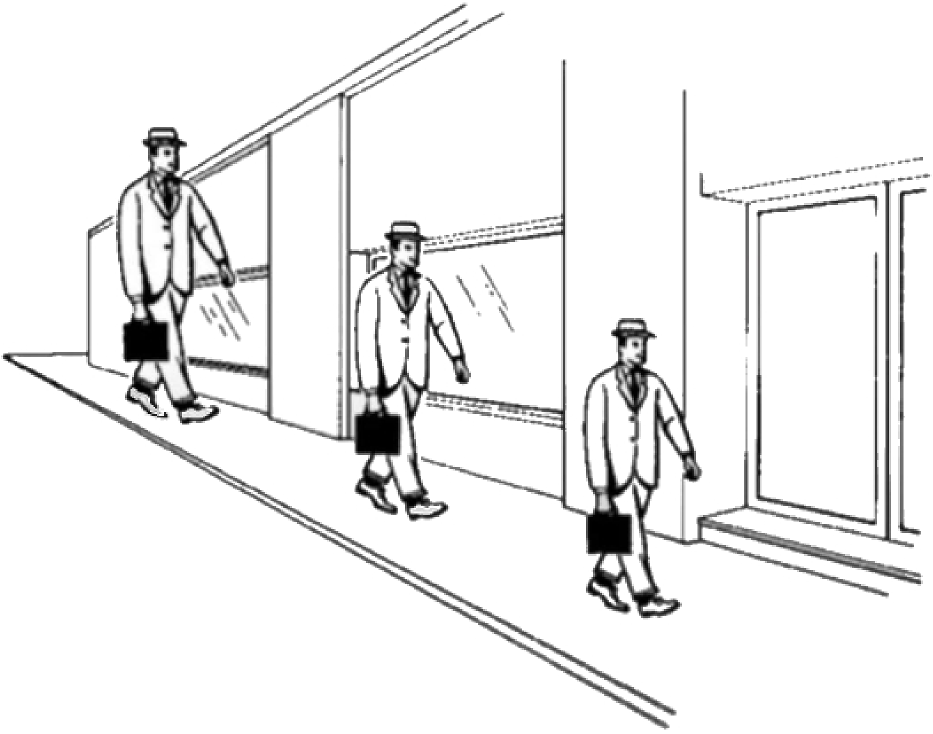
Antwerp Kieldrecht lock

Washington D.C.

And a last good one with a reminder to not confuse the boundary of an object on a projection and the horizon line (line of sight):

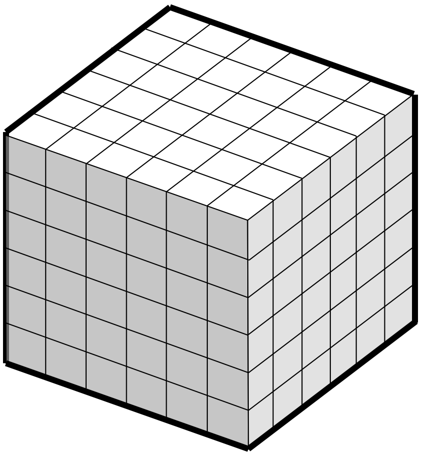

Vanishing Point
Horizon line of linear perspective
(line of sight)
Earth's horizon (boundary)

The reader must keep in mind that the human eye and brain are not very precise nor accurate and unbiased tools ton investigate Nature. A basic well known example is the illustration below involving linear perspective where the individual seems to have different heights when in reality with a ruler we can check that their height is always the same...:

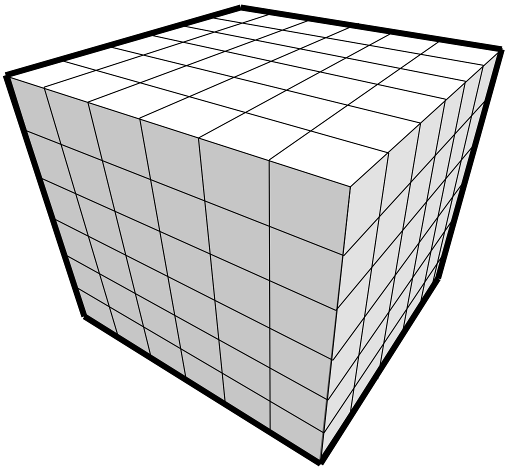


8.4.2 Affine projections

Although the best method of perspective representation is the conical perspective method, we can not always tolerate to do a large quantity of computation to represent a volume. Thus, it is possible to define projection techniques that result from an approximation of the mathematical results obtained previously to get two new techniques (there are more but these two perspectives we will see further below are by far the most used) which we see daily on many technical papers or artistic works. These two techniques are respectively the "isometric projection" and "orthogonal projection" that are part of the family of "affine projections" (in fact the use of the word "projection" must be understood as "projection on the Table of the observer" and therefore it is also a perspective).



isometric projection



perspective projection

An example of projective transformation that is not affine is photography through a Camera:

Definition (#8.4.354): An "affine projection" of the usual 3-dimensional space is an affine transformation that sends all the points of this space on a same plane of this space. If the point $M(x, y, z)$ is not on the projection plane, he and its image $m(x', y', z')$ form a line whose direction is constant: we name it the "projection direction". The corresponding perspective is named "parallel perspective" or "cylindrical perspective".

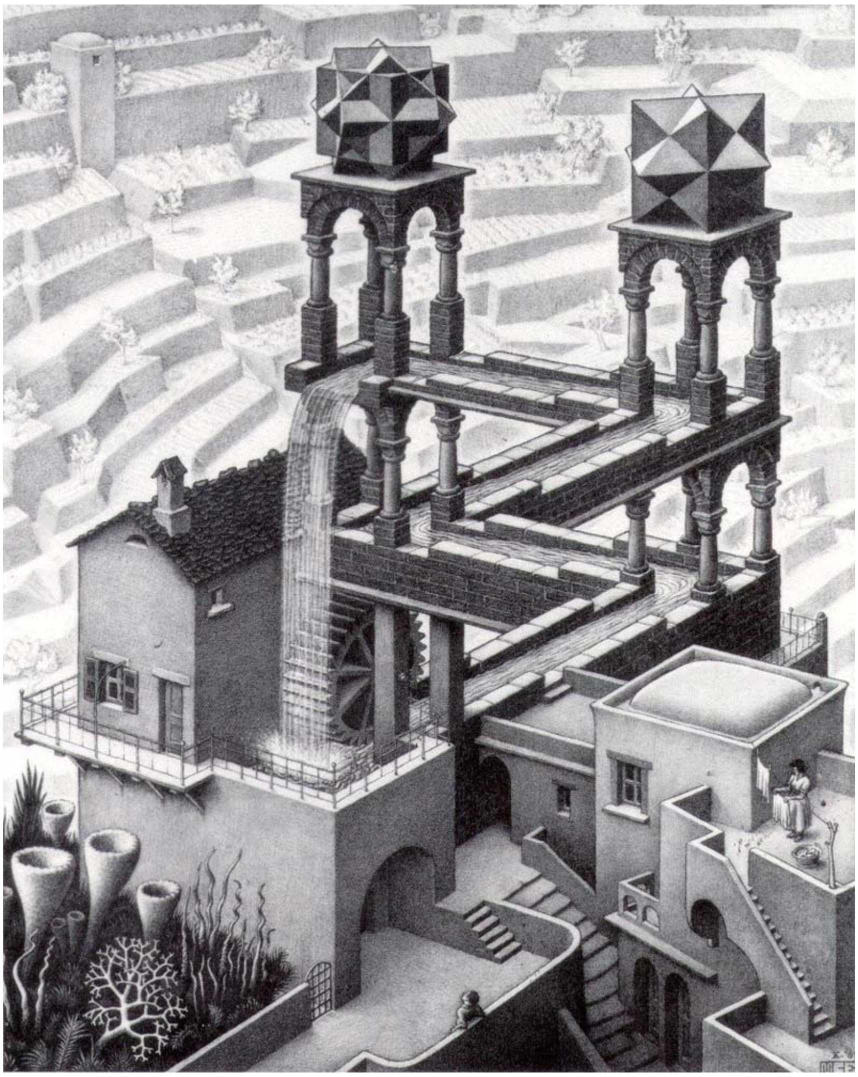
Like all affine transformations of space, an affine projection preserves:

- The parallelism between the lines
- The barycenter, that is all the proportions existing on a given line

Only the lengths and angles within a parallel plane to the projection plane are kept.

Let us present briefly these two techniques because they must be part of the general culture of the engineer.

Like all projections and all perspective, the loss of the third dimension and the non-respect of real mathematical transformation induces possible errors of interpretation. This has been extensively used by the artist M. C. Escher to create impossible situations:



8.4.2.1 Isometric perspective

Isometric projection is a method for visually representing three-dimensional objects in two dimensions in technical and engineering drawings. It is a parallel projection (object is rotated along one or more of its axes relative to the plane of projection) in which the three coordinate axes appear equally foreshortened (hence the "iso") and the angle between any two of them is approximated to 120 degrees ($2\pi/3$ [rad]):

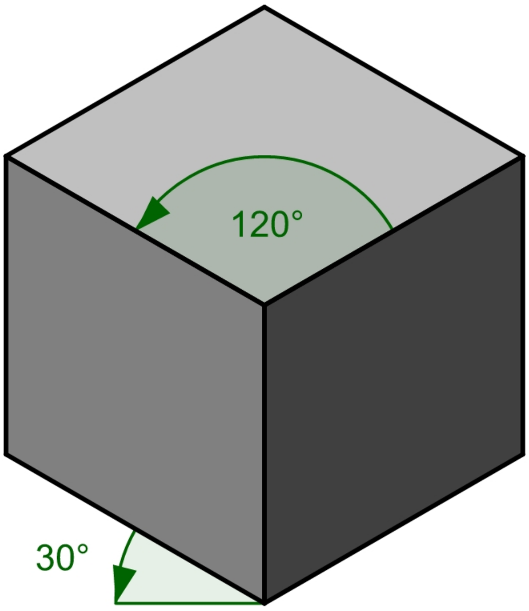


Figure 8.117: **Isometric angles (source: Wikipedia)**

Indeed, as we will see during our study of Analytical Geometry, the angle of 30° above is only an approximation of the angle between the xy-plane and the (isometric) plane vertical to the vector $\vec{n}=(1,1,1)$ passing through the corner of a cube of equal sides of length 1. In fact the choice of the angle is empirical and for example in computer games it varies between 30° and 60° depending on the game editor.

Let us consider some 3D shapes using the isometric drawing method. In the figure below coming from Wikipedia (year 12016 according to holocene calendar) the black dimensions are supposed true lengths as found in an orthographic projection (see further below). The red dimensions are used when drawing with the isometric drawing method. The same 3D shapes drawn in isometric projection would appear smaller; an isometric projection will show the object's sides foreshortened, by approximately 80%.

In fact we will see that the figure below coming from Wikipedia (year 12016 according to holocene calendar) is in fact quite wrong or at least can bring the reader to some misunderstanding!!!!

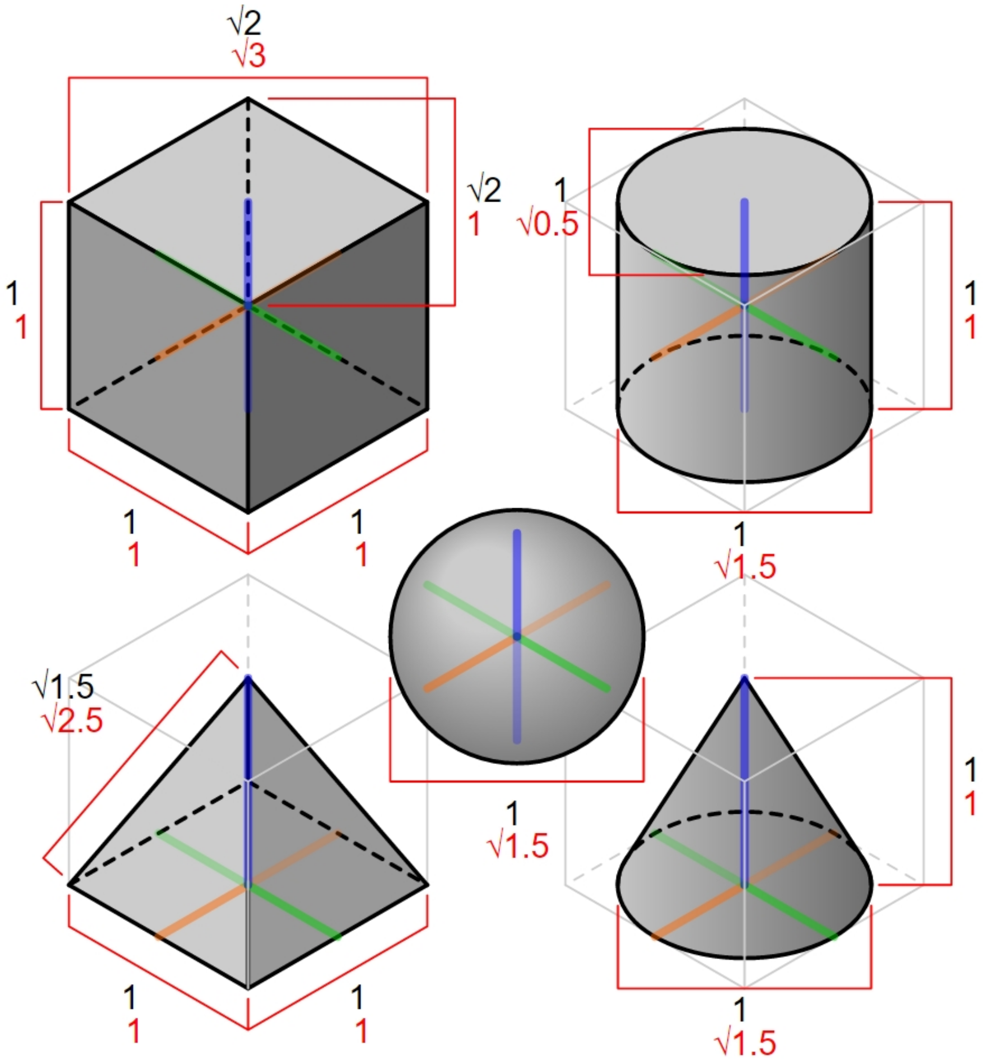


Figure 8.118: **Isometric volumes (source: Wikipedia)**

Ok let us see closer to the figure above. To prove where this values comes from we need to use the change of basis mathematics tools (see section Linear Algebra page 1992). This will be a veeery nice application of Linear Algebra and Vector Calculus!!!

So let us consider the following situation:

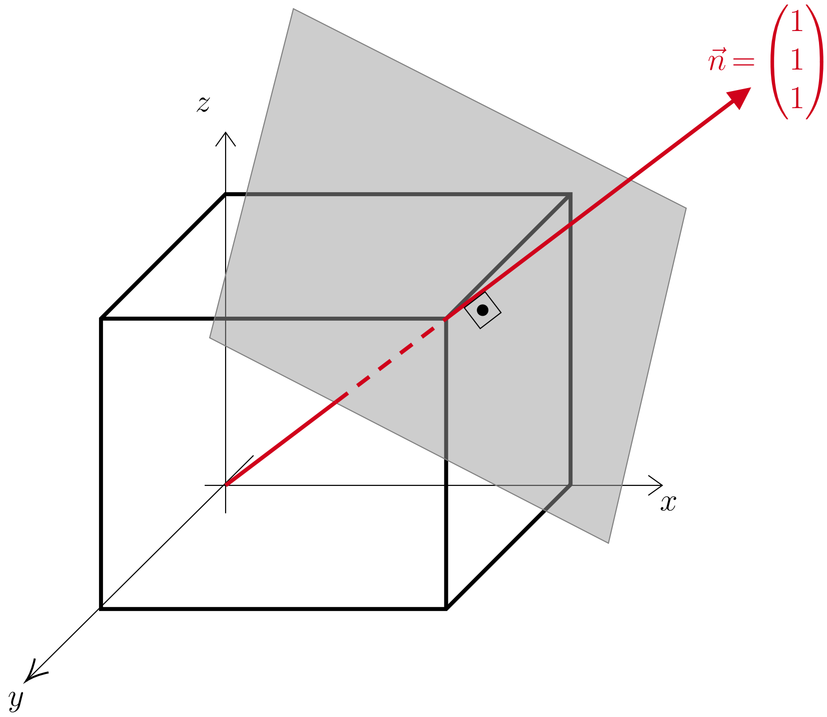


Figure 8.119: **Isometric plane to find the cube dimensions**

To simplify the determination of the new basis in the plane we well make pass our plane trough the origin (isometric affine plane) such that its equation becomes therefore:

$$x+y+z=0 \tag{8.4.33}$$

Obviously as our plane is perpendicular to $\vec{n}=(1,1,1)$ it comes immediately that a first basis vector of our orthonormal 2D basis will be the vector:

$$\vec{u}_1=\begin{pmatrix}-1\\-1\\2\end{pmatrix} \tag{8.4.34}$$

We can easily control that this point belongs well to our isometric affine plane of equation:

$$x+y+z=0 \tag{8.4.35}$$

by just replacing the values and see that the equality holds.

Now we are looking for the second orthonormal vector of our plane basis. So we first simply use the cross product:

$$\vec{u}_2=\vec{n}\times\vec{u}_1=\begin{pmatrix}1\\1\\1\end{pmatrix}\times\begin{pmatrix}-1\\-1\\2\end{pmatrix}=\begin{pmatrix}3\\-3\\0\end{pmatrix} \tag{8.4.36}$$

We can also easily control that this point belongs well to our isometric affine plane.

And we must normalize to the unit so that finally:

$$\vec{u}_1=\frac{1}{\sqrt{6}}\begin{pmatrix}-1\\-1\\2\end{pmatrix} \qquad \vec{u}_2=\frac{1}{\sqrt{18}}\begin{pmatrix}3\\-3\\0\end{pmatrix} \tag{8.4.37}$$

Now we calculate the images of the vectors of the canonical basis $(\vec{e}_1,\vec{e}_2,\vec{e}_3)$ by the orthogonal projection on our isometric affine plane P:

$$\begin{gathered}\text{proj}_P\vec{e}_1=(\vec{e}_1\circ\vec{u}_1)\vec{u}_1+(\vec{e}_1\circ\vec{u}_2)\vec{u}_2=-\frac{1}{6}\begin{pmatrix}-1\\-1\\2\end{pmatrix}+\frac{1}{6}\begin{pmatrix}3\\-3\\0\end{pmatrix}=\frac{1}{3}\begin{pmatrix}2\\-1\\-1\end{pmatrix}\\ \text{proj}_P\vec{e}_2=(\vec{e}_2\circ\vec{u}_1)\vec{u}_1+(\vec{e}_2\circ\vec{u}_2)\vec{u}_2=-\frac{1}{6}\begin{pmatrix}-1\\-1\\2\end{pmatrix}-\frac{1}{6}\begin{pmatrix}3\\-3\\0\end{pmatrix}=\frac{1}{6}\begin{pmatrix}-2\\4\\-2\end{pmatrix}=\frac{1}{3}\begin{pmatrix}-1\\2\\-1\end{pmatrix}\\ \text{proj}_P\vec{e}_3=(\vec{e}_3\circ\vec{u}_1)\vec{u}_1+(\vec{e}_3\circ\vec{u}_2)\vec{u}_2=\frac{4}{6}\begin{pmatrix}-1\\-1\\2\end{pmatrix}-0=\frac{2}{3}\begin{pmatrix}-1\\-1\\2\end{pmatrix}=\frac{1}{3}\begin{pmatrix}-2\\-2\\4\end{pmatrix}\end{gathered} \tag{8.4.38}$$

Finally the projection matrix in the canonical basis is:

$$\boxed{Z=\frac{1}{3}\begin{bmatrix}2&-1&-2\\-1&2&-2\\-1&-1&4\end{bmatrix}} \tag{8.4.39}$$

Ok now let us consider the projection of the origin $(0,0,0)$ we then have obviously:

$$\frac{1}{3}\begin{bmatrix}2&-1&-2\\-1&2&-2\\-1&-1&4\end{bmatrix}\begin{pmatrix}0\\0\\0\end{pmatrix}=\begin{pmatrix}0\\0\\0\end{pmatrix} \tag{8.4.40}$$

and now one of its X or Y unit vector :

$$\frac{1}{3}\begin{bmatrix}2&-1&-2\\-1&2&-2\\-1&-1&4\end{bmatrix}\begin{pmatrix}1\\0\\0\end{pmatrix}=\frac{1}{3}\begin{pmatrix}2\\-1\\-1\end{pmatrix} \tag{8.4.41}$$

So we get for the norm of the distance of a unit vector of the original base of the XY-plane:

$$\sqrt{\frac{2^2}{3^2}+\frac{(-1)^2}{3^2}+\frac{(-1)^2}{3^2}}=\sqrt{\frac{6}{9}}=\sqrt{\frac{2}{3}} \tag{8.4.42}$$

And finally the unit vector along the z-axis:

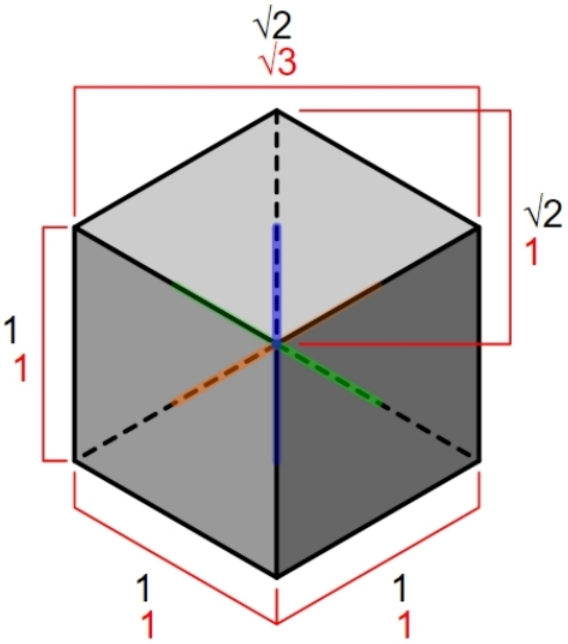
$$\frac{1}{3}\begin{bmatrix}2&-1&-2\\-1&2&-2\\-1&-1&4\end{bmatrix}\begin{pmatrix}0\\0\\1\end{pmatrix}=\frac{1}{3}\begin{pmatrix}-2\\-2\\4\end{pmatrix} \tag{8.4.43}$$

So we get for the norm of the distance of a unit vector of the original base along the Z-axis:

$$\sqrt{\frac{2^2}{3^2}+\frac{(-2)^2}{3^2}+\frac{(4)^2}{3^2}}=\sqrt{\frac{24}{9}}=\sqrt{\frac{8}{3}}\cong 2.828 \tag{8.4.44}$$

So now let us come back on Wikipedia's figure:

- First the cube...



So the real sides of a cube of side length equal to 1 transform same as the basis vector as seen just before. That is to say an isometric transformation through Z gives for the sides in or parallel to the XY plane:

$$Z:1\rightarrow\sqrt{\frac{2}{3}} \tag{8.4.45}$$

and along the Z-axis:

$$Z:1\rightarrow\sqrt{\frac{8}{3}} \tag{8.4.46}$$

So when Wikipedia says that the distances are shortened this is true... but they omit to say that this is only in two space directions. Also when on the figure they show that 1 in real dimensions is equal to 1 for the sides in isometric perspective is completely false.

The distance between the two horizontal top vertices is given by the calculation of the norm of the projection:

$$\begin{pmatrix}1\\0\\1\end{pmatrix}-\begin{pmatrix}0\\1\\1\end{pmatrix}=\begin{pmatrix}1\\-1\\0\end{pmatrix} \tag{8.4.47}$$

So:

$$\frac{1}{3}\begin{bmatrix}2&-1&-2\\-1&2&-2\\-1&-1&4\end{bmatrix}\begin{pmatrix}1\\-1\\0\end{pmatrix}=\frac{1}{3}\begin{pmatrix}3\\-3\\0\end{pmatrix} \tag{8.4.48}$$

So the norm is:

$$\sqrt{\frac{3^2}{3^2}+\frac{(-3)^2}{3^2}+\frac{0^2}{3^2}}=\sqrt{2} \tag{8.4.49}$$

And therefore the original diagonal transforms to:

$$Z:\sqrt{2}\rightarrow\sqrt{2} \tag{8.4.50}$$

So once again Wikipedia is wrong. In fact on Wikipedia they take for original sides size the inverse of $\sqrt{2/3}$, that is to say $\sqrt{3/2}$. Indeed we have in this case:

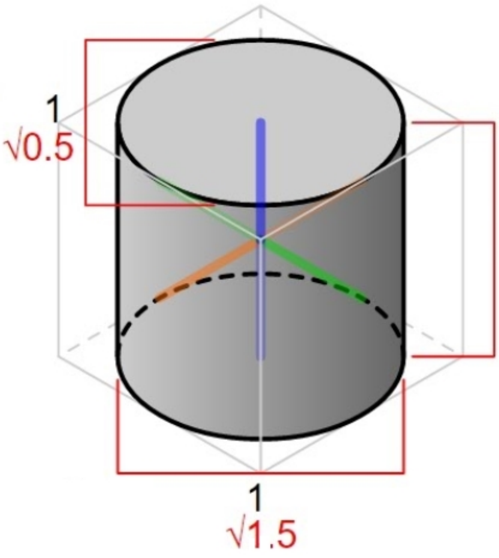
$$\frac{1}{3}\begin{bmatrix}2 & -1 & -2\\ -1 & 2 & -2\\ -1 & -1 & 4\end{bmatrix}\sqrt{\frac{3}{2}}\begin{pmatrix}1\\ -1\\ 0\end{pmatrix}=\frac{1}{\sqrt{3}\sqrt{2}}\begin{pmatrix}1\\ -5\\ 4\end{pmatrix} \tag{8.4.51}$$

So the norm is:

$$\sqrt{\frac{3^2}{(\sqrt{3}\sqrt{2})^2}+\frac{(-3)^2}{(\sqrt{3}\sqrt{2})^2}+\frac{0^2}{(\sqrt{3}\sqrt{2})^2}}=\frac{1}{\sqrt{3}\sqrt{2}}\sqrt{2\cdot 3^2}=\frac{3}{\sqrt{3}}=\sqrt{3} \tag{8.4.52}$$

So we fall back on the value given by Wikipedia...

- Second the cylinder...



The height of the cylinder follows exactly the same procedure as for the cube and the remark relatively to Wikipedia's figure remains the same!

So what will interest us here is especially the "isometric ellipse" (also named "isocircle") visible in the figure above and we will focus on the minor and major axes of that latter.

First, to find these values we need to project the circle on the isometric plane. The figure (where we added a circle on the cube) is given now by:

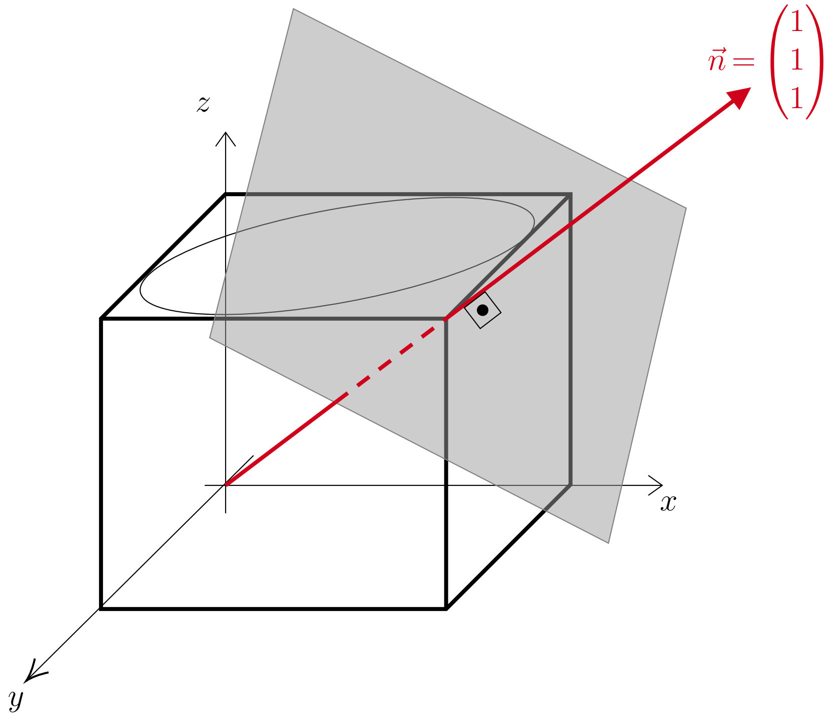


Figure 8.120: Isometric plane to find the isocircle dimensions

To finish depending on donations

8.4.2.2 Oblique perspective

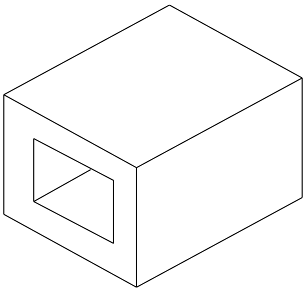
Let us take a view (eg the front view), and let us name the axes "x" (horizontal) and y (vertical) as usual. The z-axis being the axis perpendicular to the view (also as usual).

In isometric projection, we plot the z-axis with an angle relative to the x axis (for example $\pi/6$ ($30°$) or sometimes $\pi/4$ ($45°$)), and we report then the distances by multiplying by a coefficient less than 1 that is empirical on by using trigonometric basic rules.

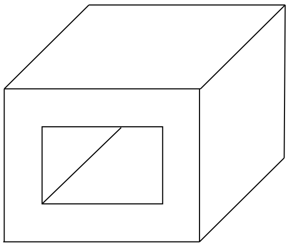
Remark

The oblique projections for all angles other than $\pi/4$ are named "Cavalier projection". When the angle is equal to $\pi/4$ we then speak of "Cabinet projection".

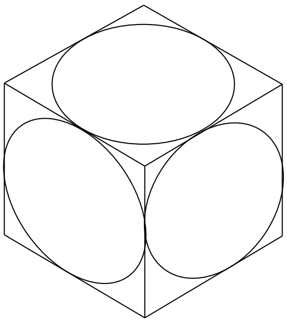
Oblique projection is a simple type of technical drawing of graphical projection used for producing two-dimensional images of three-dimensional objects. The objects are not in perspective, so they do not correspond to any view of an object that can be obtained in practice, but the technique does yield somewhat convincing and useful images:



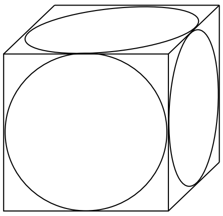
Isometric Pictorial



Oblique Pictorial



Isometric Pictorial



Oblique Pictorial

In the oblique pictorials coordinate system only one axes is at an angle. The angle may range between 0 and 90 degrees ($\pi/2$ [rad]); however, the most commonly used angle is 45 degrees.

It is thus often used when a figure must be drawn by hand, e.g. on a black board (lesson, oral examination).

8.4.2.3 Orthogonal projection

If the projection direction is orthogonal to the plane of projection, then the perspective becomes transforms a sphere into a simple circle. This is a type of perspective drawing used as an alternative to the conical perspective (with which it will also coincide when the eye of the observer is placed infinitely far away from the "table").

The simplest orthogonal projection to express is obviously the one that sends space on a plane parallel to the table, this is the "parallel orthogonal projection" or "orthographic projection", of distance d. In other words, such that for example for the "above view" we have for any point of the volume:

$$M(x,y,z)\rightarrow m(x,y,d) \tag{8.4.53}$$

So we trivially obtain all the coordinates in an orthonormal 2D frame proper to this plane itself where $x=x'$ and $y=y'$:

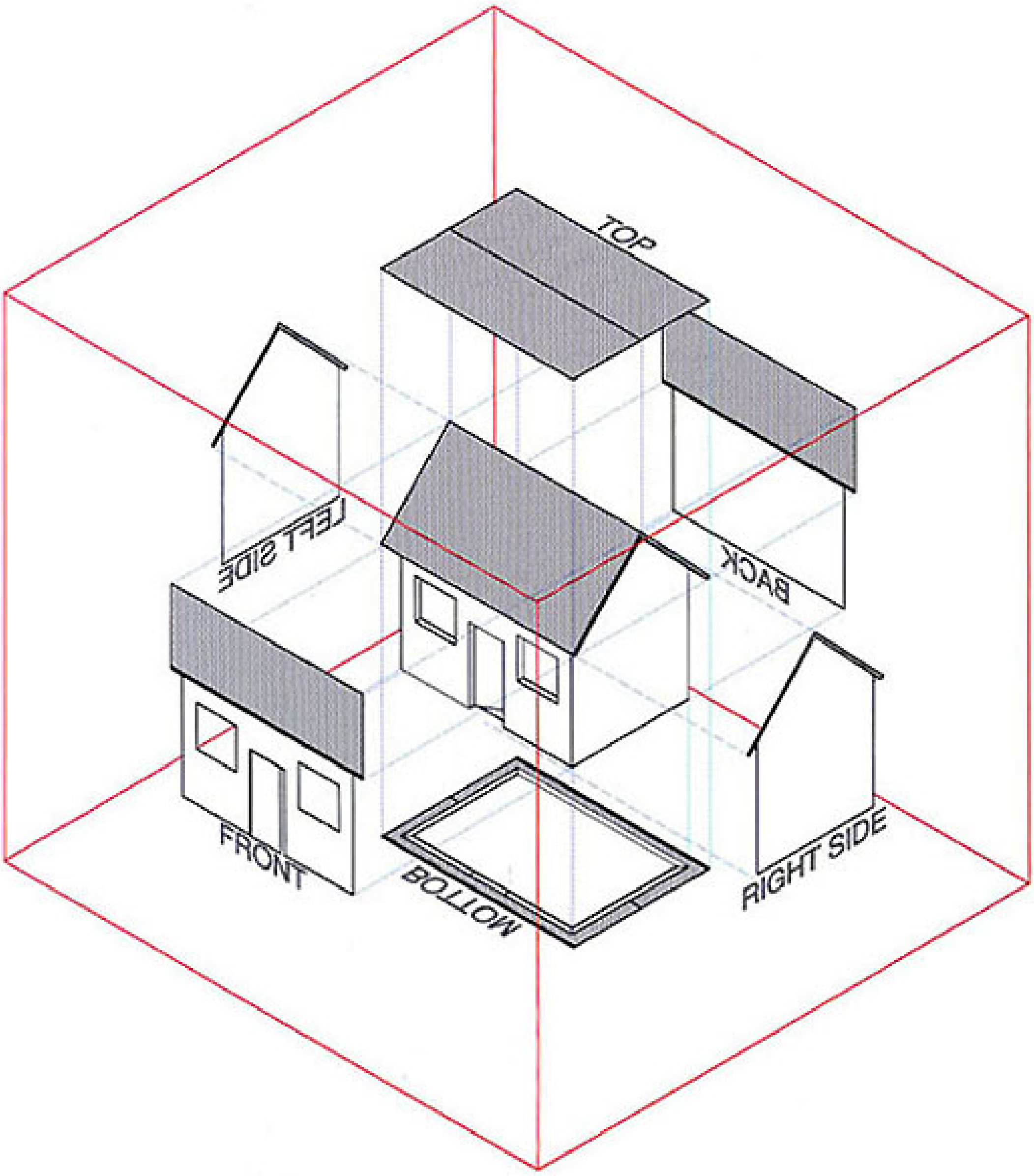


Figure 8.121: **Orthogonal projection in real life...**

Nowadays, softwares like AutoCAD, Inventor, Cathia and so on... generate automatically orthogonal projections from an isometric view of a drawing.

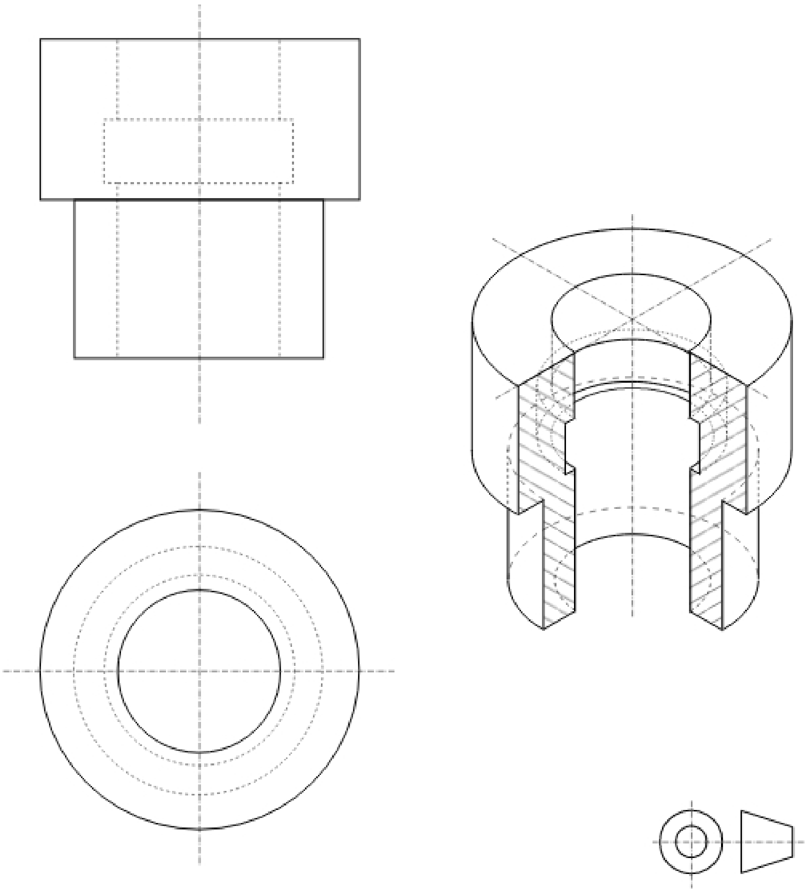


Figure 8.122: **Orthogonal Projection with a software**

In computer graphics, one of the most common matrices used for orthographic projection can be defined by a 6-tuple, (left, right, bottom, top, near, far), which defines the clipping planes (that is, the region of eye space that contains all the geometry you want to display). These planes form a box with the minimum corner at (left, bottom, -near) and the maximum corner at (right, top, -far). The box is then translated so that its center is at the origin, then it is scaled to the unit cube which is defined by having a minimum corner at $(-1,-1,-1)$ and a maximum corner at $(1,1,1)$.

The orthographic transform to get the canonical can be given by the following matrix:

$$\boxed{C_V=\begin{bmatrix}\frac{2}{\text{right}-\text{left}} & 0 & 0 & -\frac{\text{right}+\text{left}}{\text{right}-\text{left}}\\ 0 & \frac{2}{\text{top}-\text{bottom}} & 0 & -\frac{\text{top}+\text{bottom}}{\text{top}-\text{bottom}}\\ 0 & 0 & \frac{-2}{\text{far}-\text{near}} & -\frac{\text{far}+\text{near}}{\text{far}-\text{near}}\\ 0 & 0 & 0 & 1\end{bmatrix}} \tag{8.4.54}$$

which is in fact obviously a scaling followed by a translation of the form (the positions are abbreviated now with their first letter):

$$C_V=ST=\begin{bmatrix}\frac{2}{r-l} & 0 & 0 & 0\\ 0 & \frac{2}{t-b} & 0 & 0\\ 0 & 0 & \frac{2}{f-n} & 0\\ 0 & 0 & 0 & 1\end{bmatrix}\begin{bmatrix}1 & 0 & 0 & -\frac{l+r}{2}\\ 0 & 1 & 0 & -\frac{t+b}{2}\\ 0 & 0 & -1 & -\frac{f+n}{2}\\ 0 & 0 & 0 & 1\end{bmatrix} \tag{8.4.55}$$

As illustrated in the figure below:

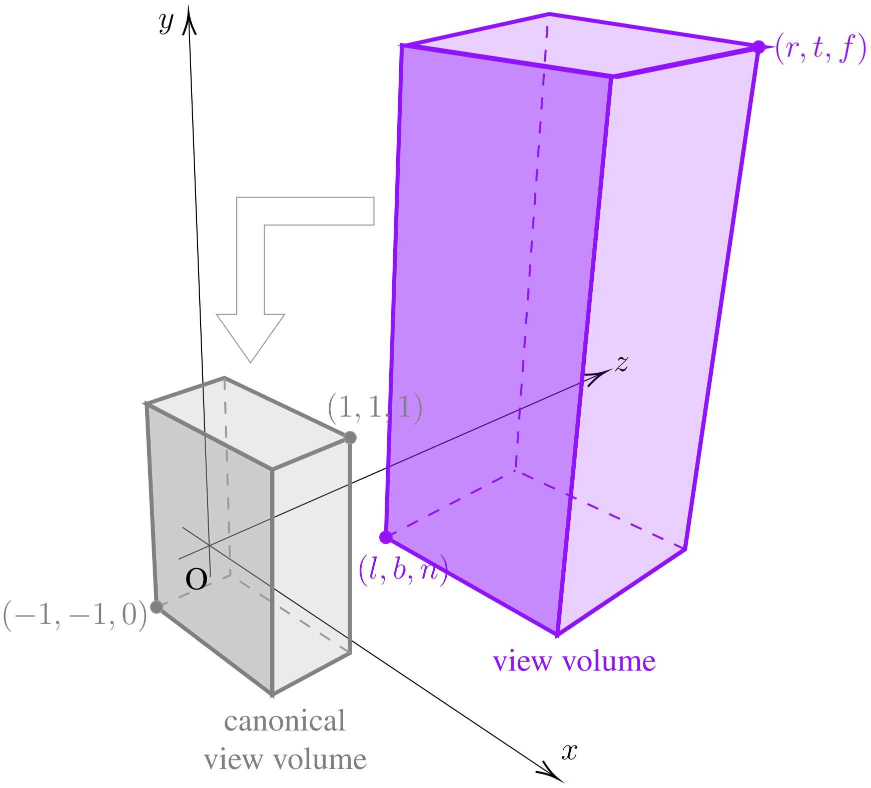


Figure 8.123: **Canonical transformation for orthogonal projection**

If the origin of the matrix C_V is not obvious here are the details to get it:

The easiest approach may be to consider each of the three axes separately, and compute how to map points along that axis from the original view volume into the canonical view volume. We begin with the x-coordinate. A point within our view volume will have an x-coordinate on the range $[l,r]$, and we want to transform it to the range $[-1,+1]$!

So first we see that we have:

$$l\leq x\leq r \tag{8.4.56}$$

Now, in preparation to scale the range down to the size we want, we subtract l from all terms to produce a zero on the left-hand side. Another approach we could take here would be to translate the range so that it centers on zero, rather than having one of its endpoints at zero, but the algebra is a bit neater this way, so we will do it like this for the sake of readability:

$$0\leq x-l\leq r-l \tag{8.4.57}$$

Now that one end of our range is positioned at zero, we can scale it down to the size we want. We want the range of x-values to be two units wide, from $+1$ to -1, so we multiply through by $2/(r-l)$. Notice that $r-l$ is the width of our view volume and is thus always a positive number, so we don't have to worry about the inequalities changing directions:

$$0 \leq \frac{2x-2l}{r-l} \leq 2 \tag{8.4.58}$$

Next, we subtract one from all terms to produce our desired range of $[-1,+1]$:

$$-1 \leq \frac{2x-2l}{r-l} - 1 \leq +1 \tag{8.4.59}$$

A bit of basic algebra allows us to write the center term as a single fraction:

$$-1 \leq \frac{2x-r-l}{r-l} - 1 \leq +1 \tag{8.4.60}$$

Finally, we split the center term into two fractions so that it takes the form $px+q$. For this we need to group our terms this way so that the equations we derive can be easily translated into matrix form:

$$-1 \leq \frac{2x}{r-l} - \frac{r+l}{r-} \leq +1 \tag{8.4.61}$$

The center term of this inequality now gives us the equation we need to transform x into the canonical view volume:

$$x' = \frac{2x}{r-l} - \frac{r+l}{r-l} \tag{8.4.62}$$

The steps required to obtain a formula for y are exactly the same. We just have to substitute y for x, t for r, and b for l. So rather than repeat them here, we will just show the result:

$$y' = \frac{2y}{t-b} - \frac{t+b}{t-b} \tag{8.4.63}$$

Finally, we need to derive a relation for z. It's a little different in this case because we are mapping z to the range $[0,+1]$ rather than $[-1,+1]$, but this should look very familiar. Here our starting condition, a z-coordinate on the range $[n,f]$:

$$n \leq z \leq f \tag{8.4.64}$$

We subtract n from all terms so the lower end of the range is positioned at zero:

$$0 \leq z-n \leq f-n \tag{8.4.65}$$

And now, all that's left is to divide through by $f-n$ to produce a final range of $[0,1]$. As before, we notice that $f-n$ indicates the depth of our viewing volume and thus will never be negative:

$$0 \leq \frac{z-n}{f-n} \leq 1 \tag{8.4.66}$$

Finally, we split this into two fractions so it takes the form $pz+q$:

$$0 \leq \frac{z}{f-n} - \frac{n}{f-n} \leq 1 \tag{8.4.67}$$

This gives us our formula for transforming z:

$$z' = \frac{z}{f-n} - \frac{n}{f-n} \tag{8.4.68}$$

Now, we are ready to write our centering orthographic projection matrix. To recap our work thus far, here are the three projection equations we have derived:

Once the canonical volume obtained we apply for example the simple orthographic projection onto the plane $z=0$ that is intuitively defined by the following matrix:

$$P_{xy} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{8.4.69}$$

For each point (x,y,z), the transformed point would be:

$$P_{xy} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x \\ y \\ 0 \end{pmatrix} \tag{8.4.70}$$

Often, it is more useful to use homogeneous coordinates. The transformation above can be represented for homogeneous coordinates as:

$$P_{xy} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{8.4.71}$$

For each homogeneous vector (x,y,z), the transformed vector will be obviously:

$$P_{xy} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} = \begin{pmatrix} x \\ y \\ 0 \end{pmatrix} \tag{8.4.72}$$

Although this is the most used method in the industry on paper (or screen), however, it has a drawback: the real effect of depth is lost entirely and this is why orthogonal projection is often accompanied of an isometric projection view of the object of interest.

8.4.3 Spherical projections

Spherical projections is an important field in engineering topography and geography (and therefore navigation). As far as we know there is not far from more than 200 various techniques of spherical projection.

Indeed, it is impossible to render paths on a sphere onto a flat surface in such a way that all distances remain the same. In drawing a map of a sphere, therefore, some compromises must be made. Most maps adopt one of two possible strategies that areas are preserved or that angles are preserved. Stereographic projection (i.e. projection on a plane) is one way of making maps, and it preserves angles. It has been used since ancient times for this purpose, and its basic geometrical properties were known even then.

We will focus in this textbook only on the most known spherical projections and also without detailing the problems of projecting images from a sphere to a plane when the number of points (pixels) is not infinite...

For example here is a summary of some well known spherical projections[6]:

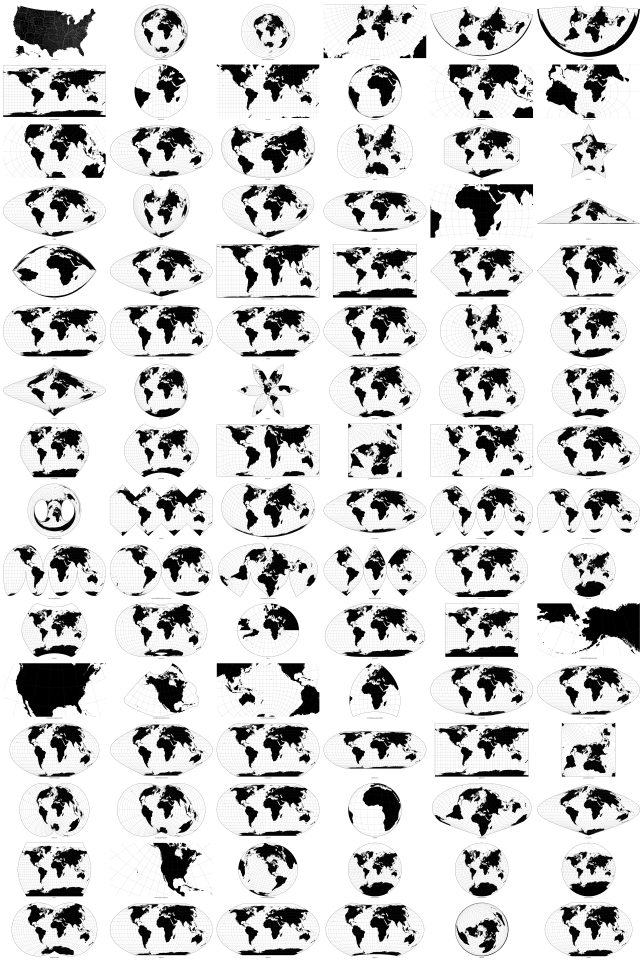


Figure 8.124: Various spherical and cylindrical projections like Mercator, Gall-Peters, Mollweide, Robinson, etc. (author: Mike Bostock)

---

[6] The choice of the projection used in school depends mainly on the country government as some projections give more importance to some countries than others... But anyway any student can now see an Earth sphere (globe) in its class and this is the only totally conform existing projection!

Or with a more detailed idea of each technique and its corresponding name:

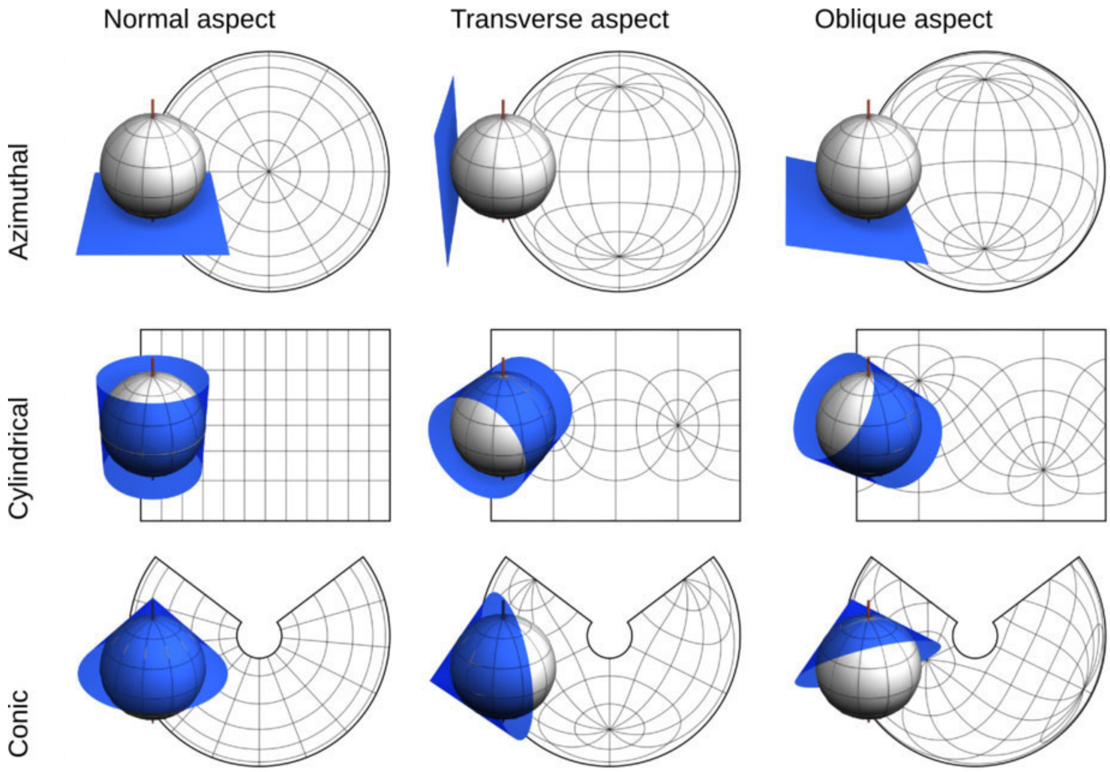

Normal aspect
Transverse aspect
Oblique aspect
Azimuthal
Cylindrical
Conic

Figure 8.125: Technical view of various spherical projections techniques

Remark

An exhaustive list of spherical projections is given on the projection list web page of the Geocart software available here:

`https://www.mapthematics.com/ProjectionsList.php`

Obviously we could avoid all the issues of spherical projections by using in schools and on logo of corporations or governmental agencies (like some already do!) a "hemispherical map":

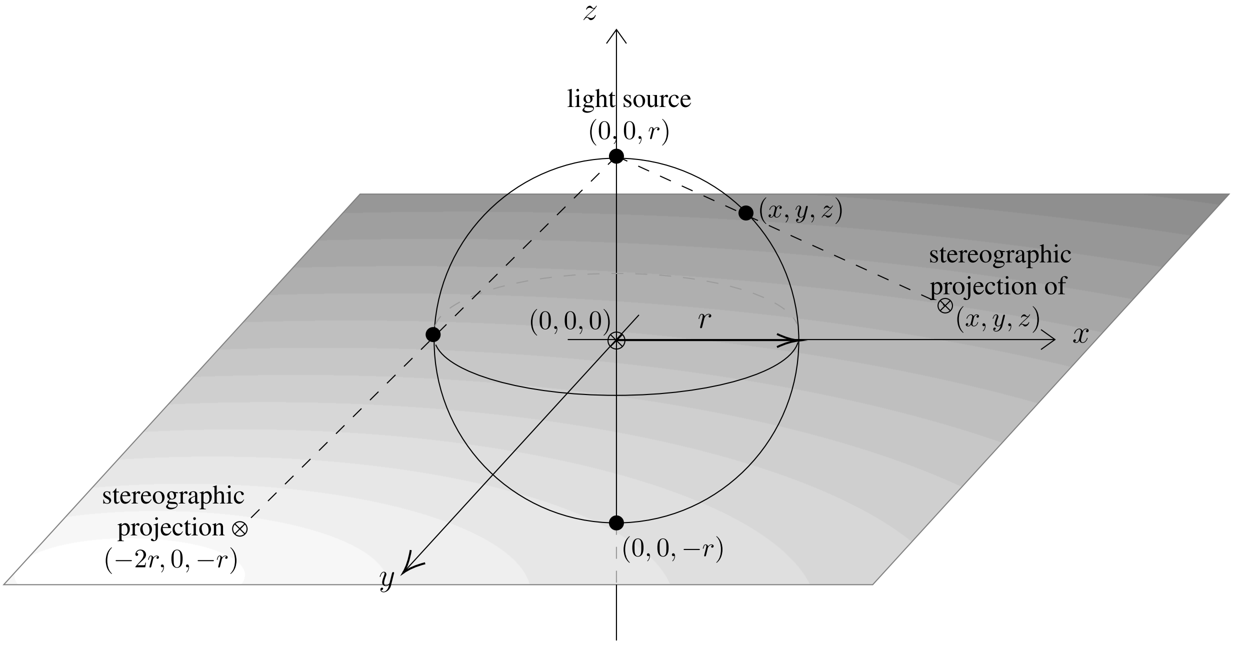
TABLEAU SPHÉROÏDAL DE LA TERRE
N° 13.
OU MAPPEMONDE EN DEUX HÉMISPHÈRES,
Substituant à la Projection qui donnait à la Terre une forme Concave, celle, plus rationnelle, figurant sa Convexité;
et résumant, à l'aide de Lignes Isothermes, le Système de l'Égalité des Températures à différentes Latitudes du Globe.
HÉMISPHÈRE OCCIDENTAL.
HÉMISPHÈRE ORIENTAL.
AMÉRIQUE SEPTENTRIONALE
AMÉRIQUE MÉRIDIONALE
AFRIQUE
OCÉAN INDIEN
MER DES INDES
LÉGENDE EXPLICATIVE... PROJECTION ORTHOGRAPHIQUE.
TABLEAU DES LIGNES ISOTHERMES et CARTE D'ÉGALES TEMPÉRATURES.
TEMPÉRATURES... LIGNES ISOTHERMES... PÔLES DU FROID.

8.4.3.1 Stereographic projection

The "stereographic projection" (in the classical sense of the term), or more precisely the "azimuthal normal aspect projection", is one way of projecting the points that lie on a spherical surface onto a plane. Such projections are commonly used in Earth and space mapping where the geometry is often inherently spherical and needs to be displayed on a flat surface such as paper or a computer display. Any attempt to map a sphere onto a plane requires distortion, stereographic projections are no exception and indeed it is not an ideal approach if minimal distortion is desired.

A physical model of stereographic projections is to imagine a transparent sphere sitting on a plane. If we name the point at which the sphere touches the plane the "south pole" then we place a light source at the "north pole". Each ray from the light passes through a point on the and then strikes the plane, this is the stereographic projection of the point on the sphere.

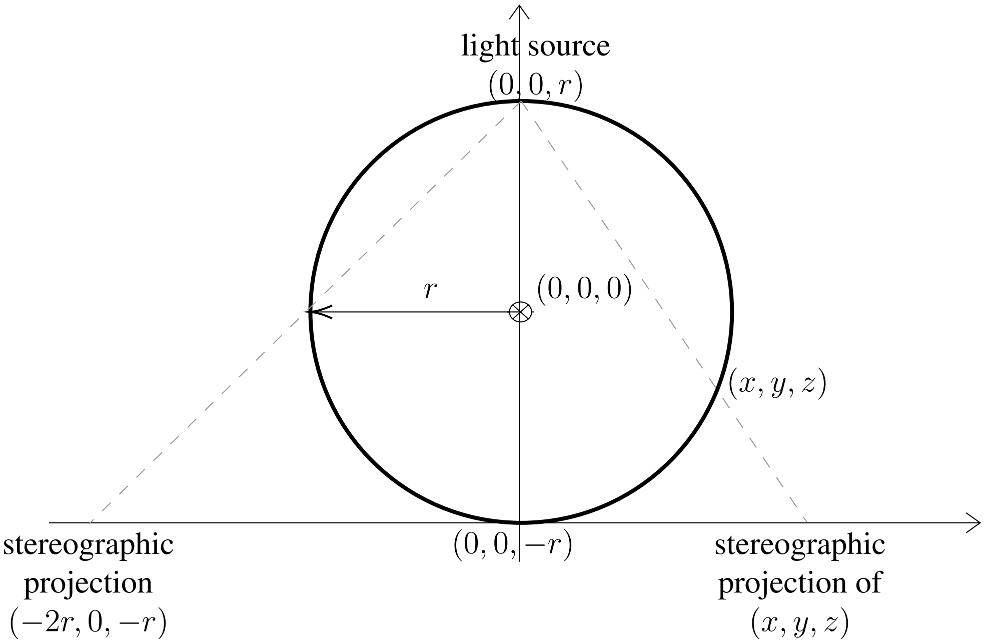
In order to derive the formula for the projection of a point (x,y,z) lying on the sphere assume the sphere is centered at the origin and is of radius r. The plane is all the points $z=-r$, and the light source is at point $(0,0,r)$. The cross section of this arrangement is shown below in what is commonly named a "Schlegel diagram" that looks like below in pseudo-perspective:



z
light source
$(0, 0, r)$
(x, y, z)
stereographic
projection of
(x, y, z)
$(0, 0, 0)$
r
x
stereographic
projection ⊗
$(-2r, 0, -r)$
y
$(0, 0, -r)$

Figure 8.126: **Schlegel diagram in pseudo-perspective (stereographic projection)**

and in front view:



light source
$(0, 0, r)$
r
$(0, 0, 0)$
(x, y, z)
stereographic
projection
$(-2r, 0, -r)$
$(0, 0, -r)$
stereographic
projection of
(x, y, z)

Figure 8.127: **Schlegel diagram in front view**

As we can see above the projection of $(0,0,r)$ gives obviously $(0,0,-r)$ and using Thales's theorem (or "intercept theorem") as proved in the section of Euclidean Geometry (see page 2610) we have immediately that the projection of $(-r,0,0)$ is $(-2r,0,r)$.

Let us consider now the equation of the line from $P_1=(0,0,r)$ through a point $P_2=(x,y,z)$

on the sphere. Therefore we have for the projected point P:

$$\vec{P} = \vec{P}_1 + a(\vec{P}_1 - \vec{P}_2) \tag{8.4.73}$$

Solving this for a for the z component yields:

$$-r = r + a(z - r) \tag{8.4.74}$$

or:

$$a = \frac{2r}{r - z} \tag{8.4.75}$$

Finally:

$$\boxed{\begin{pmatrix} x_p \\ y_p \\ z_p \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ r \end{pmatrix} + \frac{2r}{r-z}\begin{pmatrix} x \\ y \\ z - r \end{pmatrix} = \begin{pmatrix} \dfrac{2rx}{r-z} \\ \dfrac{2ry}{r-z} \\ -r \end{pmatrix}} \tag{8.4.76}$$

Notice that:

- The South pole is at the center of the projected points
- Lines of latitude project to concentric circles about $(0, 0, -r)$
- Lines of longitude project to rays from the point $(0, 0, -r)$
- There is little distortion near the South pole
- The equation project to a circle of radius $2r$
- The distortion increases the closer one gets to the north pole finally becoming infinite at the north pole
- As we can see in the image below, angles are preserved (since latitude and longitude are still perpendicular on the projected plane)

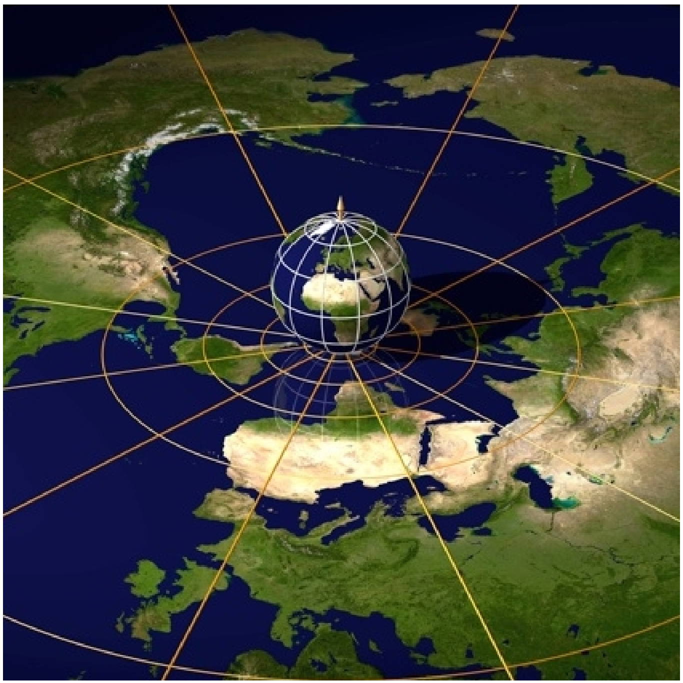


Figure 8.128: **Earth's stereographic projection (source: ?)**

Some author put the origin of the frame at the center of the sphere. Therefore the previous relations becomes:

$$\vec{P}=\begin{pmatrix}\dfrac{2rx}{r-z}\\\dfrac{2ry}{r-z}\\0\end{pmatrix}\tag{8.4.77}$$

and take for radius $r=1$. Then we have:

$$\vec{P}=\begin{pmatrix}\dfrac{x}{1-z}\\\dfrac{y}{1-z}\\0\end{pmatrix}=\left(\frac{x}{1-z},\frac{y}{1-z},0\right)\tag{8.4.78}$$

And as we denote the components of the projected points by a,b,c instead of x,y,z, we get:

$$\vec{P}=\left(\frac{a}{1-c},\frac{b}{1-c},0\right)\tag{8.4.79}$$

And as the last coordinate is always 0 we only write:

$$\vec{P}=\left(\frac{a}{1-c},\frac{b}{1-c}\right)\tag{8.4.80}$$

And when the projection plane is assimilated to the (complex) projection plane, then we get the relation that we can find most of time in books:

$$\boxed{P=\frac{a}{1-c}+\mathrm{i}\frac{b}{1-c}=\frac{a+\mathrm{i}b}{1-c}}\tag{8.4.81}$$

Theorem 8.19. The stereographic projection is a conformal projection (i.e. conserves the angles). That is to say:

If we consider the stereographic projection $\pi: S^2 \setminus \{\vec{e}_3\} \in \mathbb{R}^3$ as we have proved it, defined by:

$$\pi: (x,y,z)^T \Rightarrow \left(\frac{x}{1-z}, \frac{y}{1-z}, 0\right)^T \tag{8.4.82}$$

with $x^2+y^2+z^2=1$ and $\vec{e}_3=(0,0,1)^T$.

Given two paths φ and γ of type $\mathcal{C}^1$ on the sphere (that is to say two differentiable applications in which the derivative is continuous for recall...) such as (empirical choice):

$$\varphi(0)=\gamma(0) \tag{8.4.83}$$

and such that $\varphi(0)=\gamma(0)\neq\vec{e}_3$, we want to check if (we use the bracket notation for the dot product to not confuse it with the round symbol of function composition):

$$\frac{\langle(\pi\circ\varphi)'(0),(\pi\circ\gamma)'(0)\rangle}{||(\pi\circ\varphi)'(0)||\cdot||(\pi\circ\gamma)'(0)||}=\frac{\langle\varphi'(0),\gamma'(0)\rangle}{||\varphi'(0)||\cdot||\gamma'(0)||} \tag{8.4.84}$$

where $\varphi'(0)$ and $\gamma'(0)$ are the tangent vectors of φ and γ on point 0 and $(\pi\circ\varphi)'(0)$ and $(\pi\circ\gamma)'(0)$ are there corresponding tangent vectors on the projection plane.

Proof 8.19.1. To do the proof let start by developing the expressions $(\pi\circ\varphi)'(0)$ and $(\pi\circ\gamma)'(0)$. Let us write $\varphi=(\varphi_x,\varphi_y,\varphi_z)^T$. We have :

$$\pi\circ\varphi=\left(\frac{\varphi_x}{1-\varphi_z},\frac{\varphi_y}{1-\varphi_z}\right)^T \tag{8.4.85}$$

Let us simplify the denominator by putting:

$$\vec{u}=\vec{\varphi}-\vec{e} \tag{8.4.86}$$

Then we get:

$$\pi\circ\varphi=\left(-\frac{u_x}{u_z},-\frac{u_y}{u_z},0\right)^T \tag{8.4.87}$$

Hence:

$$\begin{aligned}(\pi\circ\varphi)'(0)&=\left(\frac{-u_x{}'(0)\cdot u_z(0)+u_x(0)\cdot u_z{}'(0)}{u_z(0)^2},\frac{-u_y{}'(0)\cdot u_z(0)+u_y(0)\cdot u_z{}'(0)}{u_z(0)^2},0\right)^T\\&=\frac{1}{u_z(0)}(-u_x{}'(0),-u_y{}'(0),0)^T+\frac{u_z{}'(0)}{u_z(0)^2}(u_x(0),u_y(0),0)^T\\&=-\frac{1}{u_z(0)}(u_x{}'(0),u_y{}'(0),u_z{}'(0))^T+\frac{u_z{}'(0)}{u_z(0)^2}(u_x(0),u_y(0),u_z(0))^T\\&=-\frac{1}{u_z(0)}\vec{u}'(0)+\frac{u_z{}'(0)}{u_z(0)^2}\vec{u}(0)\end{aligned} \tag{8.4.88}$$

And same, by putting:

$$\vec{v}=\vec{\gamma}-\vec{e}_3 \tag{8.4.89}$$

we get:

$$(\pi\circ\gamma)'(0)=-\frac{1}{v_z(0)}\vec{v}\,'(0)+\frac{v'_z(0)}{v_z(0)^2}\vec{v}(0) \tag{8.4.90}$$

Let us now develop $\langle(\phi\circ\varphi)'(0),(\phi\circ\gamma)'(0)\rangle$:

$$\begin{aligned}&\left\langle(\pi\circ\varphi)'(0),(\pi\circ\gamma)'(0)\right\rangle=\left\langle-\frac{1}{u_z\,(0)}u'\,(0)+\frac{u_z{}'\,(0)}{u_z(0)^2}u\,(0)\,,-\frac{1}{v_z\,(0)}v'\,(0)+\frac{v_z{}'\,(0)}{v_z(0)^2}v\,(0)\right\rangle\\&=\frac{1}{u_z\,(0)\,v_z\,(0)}\left\langle u'\,(0)\,,v'\,(0)\right\rangle\\&-\frac{v_z{}'\,(0)}{u_z\,(0)\,v_z(0)^2}\left\langle u'\,(0)\,,v\,(0)\right\rangle-\frac{u_z{}'\,(0)}{u_z(0)^2v_z\,(0)}\left\langle u\,(0)\,,v'\,(0)\right\rangle+\frac{u_z{}'\,(0)\,v_z{}'\,(0)}{u_z(0)^2\,v_z(0)^2}\left\langle u\,(0)\,,v\,(0)\right\rangle\end{aligned} \tag{8.4.91}$$

The change of variables takes us immediately to:

$$\varphi(0)=\gamma(0)\Leftrightarrow u(0)=v(0) \tag{8.4.92}$$

Therefore:

$$\begin{aligned}&\left\langle(\pi\circ\varphi)'(0),(\pi\circ\gamma)'(0)\right\rangle=\\&=\frac{1}{u_z(0)^2}\left\langle u'\,(0)\,,v'\,(0)\right\rangle-\frac{v_z{}'\,(0)}{u_z(0)^3}\left\langle u'\,(0)\,,u\,(0)\right\rangle-\frac{u_z{}'\,(0)}{u_z(0)^3}\left\langle v\,(0)\,,v'\,(0)\right\rangle+\frac{u_z{}'\,(0)\,v_z{}'\,(0)}{u_z(0)^4}\left\langle u\,(0)\,,u\,(0)\right\rangle\end{aligned} \tag{8.4.93}$$

As by construction:

$$\langle\varphi,\varphi'\rangle=0 \tag{8.4.94}$$

and that we also have:

$$\langle u,u'\rangle=\langle\varphi-\vec{e}_3,\varphi'\rangle=-\langle\vec{e}_3,\varphi'\rangle=-\varphi'_z=-u'_z \tag{8.4.95}$$

and identically:

$$\langle v,v'\rangle=-v'_z \tag{8.4.96}$$

then the above dot product becomes:

$$\begin{aligned}&\left\langle(\pi\circ\varphi)'(0),(\pi\circ\gamma)'(0)\right\rangle\\&=\frac{1}{u_z(0)^2}\left\langle u'\,(0)\,,v'\,(0)\right\rangle+2\frac{v_z{}'\,(0)\,u_z{}'\,(0)}{u_z(0)^3}+\frac{u_z{}'\,(0)\,v_z{}'\,(0)}{u_z(0)^4}\left\langle u\,(0)\,,u\,(0)\right\rangle\\&=\frac{1}{u_z(0)^2}\left\langle u'\,(0)\,,v'\,(0)\right\rangle+\frac{u_z{}'\,(0)\,v_z{}'\,(0)}{u_z(0)^4}\left(\left\langle u\,(0)\,,u\,(0)\right\rangle+2u_z\,(0)\right)\end{aligned} \tag{8.4.97}$$

But:

$$\begin{aligned}\langle u(0),u(0)\rangle+2u_z(0)&=\langle u(0),u(0)\rangle+2\langle u(0),e_3\rangle=\langle u(0),u(0)+2e_3\rangle\\&=\langle\varphi(0)-e_3,\varphi(0)+e_3\rangle=\varphi_x^2+\varphi_y^2+(\varphi_z-1)(\varphi_z+1)\\&=\varphi_x^2+\varphi_y^2+\varphi_z^2-1^2=1-1=0\end{aligned}\tag{8.4.98}$$

since by construction $\varphi_x^2+\varphi_y^2+\varphi_z^2=1$.

So finally our numerator is given by:

$$\left\langle(\pi\circ\varphi)'(0),(\pi\circ\gamma)'(0)\right\rangle=\frac{1}{u_z(0)^2}\langle u'(0),v'(0)\rangle=\frac{1}{(\varphi_z(0)-1)^2}\langle\varphi'(0),\gamma'(0)\rangle\tag{8.4.99}$$

Let us now calculate the first term of the denominator $||(\pi\circ\varphi)'(0)||$:

$$\begin{aligned}\left\|(\pi\circ\varphi)'(0)\right\|^2&=\left\langle-\frac{1}{u_z(0)}u'(0)+\frac{u_z{}'(0)}{u_z(0)^2}u(0),-\frac{1}{u_z(0)}u'(0)+\frac{u_z{}'(0)}{u_z(0)^2}u(0)\right\rangle\\&=\frac{1}{u_z(0)^2}\langle u'(0),u'(0)\rangle+2\frac{u_z{}'(0)^2}{u_z(0)^3}+\frac{u_z{}'(0)^2}{u_z(0)^4}\langle u(0),u(0)\rangle\end{aligned}\tag{8.4.100}$$

as for recall (see proof just before) $\langle u,u'\rangle=-u'_z$.

But we also have:

$$\begin{aligned}\langle u,u\rangle&=\langle\varphi-\vec{e}_3,\varphi-\vec{e}_3\rangle=\varphi_x^2+\varphi_y^2+(\varphi_z-1)^2\\&=\varphi_x^2+\varphi_y^2+\varphi_z^2-2\varphi_z+1=1-2\varphi_z+1=2(1-\varphi_z)=-2u_z\end{aligned}\tag{8.4.101}$$

Therefore the previous expressions becomes:

$$\left\|(\pi\circ\varphi)'(0)\right\|^2=\frac{1}{u_z(0)^2}\langle u'(0),u'(0)\rangle=\frac{1}{(\varphi_z(0)-1)^2}\langle\varphi'(0),\varphi'(0)\rangle\tag{8.4.102}$$

Similarly:

$$\left\|(\pi\circ\gamma)'(0)\right\|^2=\frac{1}{(\gamma_z(0)-1)^2}\langle\gamma'(0),\gamma'(0)\rangle\tag{8.4.103}$$

To finish, we put:

$$\begin{aligned}\left\langle(\pi\circ\varphi)'(0),(\pi\circ\gamma)'(0)\right\rangle&=\frac{1}{(\varphi_z(0)-1)^2}\langle\varphi'(0),\gamma'(0)\rangle\\\left\|(\pi\circ\varphi)'(0)\right\|^2&=\frac{1}{(\varphi_z(0)-1)^2}\langle\varphi'(0),\varphi'(0)\rangle\\\left\|(\pi\circ\gamma)'(0)\right\|^2&=\frac{1}{(\gamma_z(0)-1)^2}\langle\gamma'(0),\gamma'(0)\rangle\end{aligned}\tag{8.4.104}$$

inside:

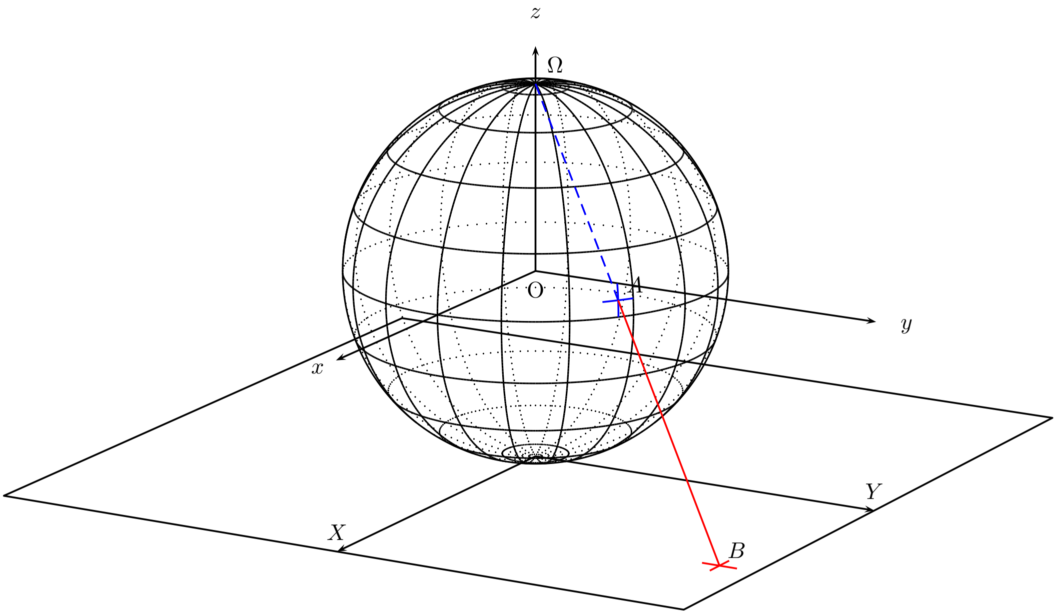
$$\frac{\langle(\pi\circ\varphi)'(0),(\pi\circ\gamma)'(0)\rangle}{||(\pi\circ\varphi)'(0)||\cdot||(\pi\circ\gamma)'(0)||}\tag{8.4.105}$$

and we get:

$$\frac{\left\langle(\pi\circ\varphi)'(0),(\pi\circ\gamma)'(0)\right\rangle}{\left\|(\pi\circ\varphi)'(0)\right\|\cdot\left\|(\pi\circ\gamma)'(0)\right\|}=\frac{\langle\varphi'(0),\gamma'(0)\rangle}{\sqrt{\langle\varphi'(0),\varphi'(0)\rangle}\sqrt{\langle\gamma'(0),\gamma'(0)\rangle}}=\frac{\langle\varphi'(0),\gamma'(0)\rangle}{\|\varphi'(0)\|\cdot\|\gamma'(0)\|}\tag{8.4.106}$$

■ Q.E.D.

Many people need in 3D softwares to map an image (a flat one) on a sphere. How do software do the inverse of a stereographic projection? To see that, let us consider first the following illustration made by Manuel Luque:



In the figure above we have first:

$$\text{O}=\begin{pmatrix}0\\0\\0\end{pmatrix}\quad \Omega=\begin{pmatrix}0\\0\\1\end{pmatrix}\quad A=\begin{pmatrix}x\\y\\z\end{pmatrix} \tag{8.4.107}$$

and we also have:

$$x^2+y^2+z^2=1 \tag{8.4.108}$$

and we also have:

$$B=\begin{pmatrix}X\\Y\\-1\end{pmatrix}\quad \overrightarrow{\Omega A}=\begin{pmatrix}x\\y\\z-1\end{pmatrix}\quad \overrightarrow{\Omega B}=\begin{pmatrix}X\\Y\\-2\end{pmatrix} \tag{8.4.109}$$

and trivially:

$$\overrightarrow{\Omega A}=k\overrightarrow{\Omega B}\Longleftrightarrow\begin{pmatrix}x\\y\\z-1\end{pmatrix}=k\begin{pmatrix}X\\Y\\-2\end{pmatrix} \tag{8.4.110}$$

The problem posed is the following: we give ourself the coordinates of a point B of the plane $z=-1$, we must deduce the coordinates of the point A of the sphere $x=f(X,Y);y=g(X,Y);z=h(X,Y)$.

First we have easily:

$$\begin{cases} x = kX \\ y = kY \\ 1 - z = 2k \end{cases} \Longrightarrow \begin{array}{l} x^2 + y^2 = k^2\left(X^2 + Y^2\right) \\ k = \frac{1-z}{2} \end{array} \tag{8.4.111}$$

Knowing that $x^2 + y^2 = 1 - z^2$ we get by replacing:

$$1 - z^2 = \frac{(1-z)^2}{4}\left(X^2 + Y^2\right) \Longrightarrow 4(1+z) = (1-z)\left(X^2 + Y^2\right) \tag{8.4.112}$$

Therefore:

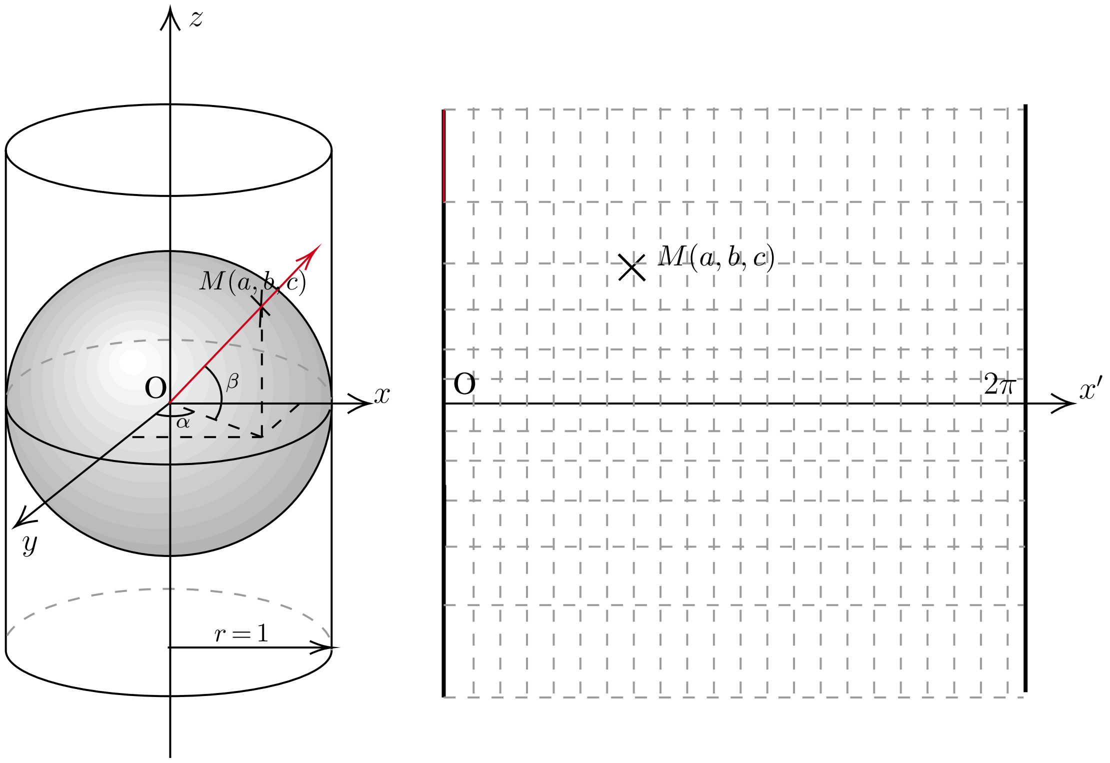
$$z = \frac{X^2 + Y^2 - 4}{X^2 + Y^2 + 4} \tag{8.4.113}$$

Doing the same for the other two coordinates, we get:

$$\begin{cases} x = \frac{4X}{X^2+Y^2+4} \\ y = \frac{4Y}{X^2+Y^2+4} \\ z = \frac{X^2+Y^2-4}{X^2+Y^2+4} \end{cases} \tag{8.4.114}$$

8.4.3.2 Cylindrical projection

The "cylindrical projection" (in the classical sense of the term) or more precisely named the "cylindrical normal aspect projection", is one where lines of latitude are projected to equally spaced parallel lines and lines of longitude are projected onto not necessarily equally spaced parallel lines. The figure below illustrates the basic projection, a line is projected from the center of the sphere through each point on the sphere until it intersects:



As we can see from the above figure it is quite obvious first that:

$$\tan(\beta)=\frac{z'}{r}=\frac{z'}{1}\Rightarrow z'=\tan(\beta)=\tan\left(\arctan\left(\frac{c}{\sqrt{a^2+b^2}}\right)\right)=\frac{c}{\sqrt{a^2+b^2}} \tag{8.4.115}$$

and for x' we have obviously:

$$x'=\alpha=\operatorname{sgn}(x)\arccos\left(\frac{b}{\sqrt{a^2+b^2}}\right) \tag{8.4.116}$$

with for recall:

$$\operatorname{sgn}(x)=\begin{cases}-1 & x<0\\ 0 & x=0\\ +1 & x>0\end{cases} \tag{8.4.117}$$

So finally:

$$\boxed{\begin{aligned}&x'=\alpha=\operatorname{sgn}(x)\arccos\left(\frac{b}{\sqrt{a^2+b^2}}\right)\\&z'=\tan(\beta)=\frac{c}{\sqrt{a^2+b^2}}\end{aligned}} \tag{8.4.118}$$

Schematically the steps of construction are the following:

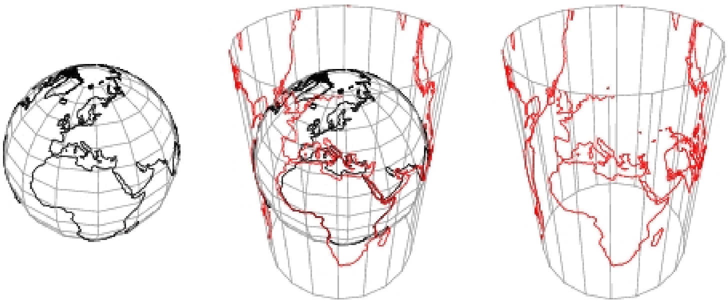
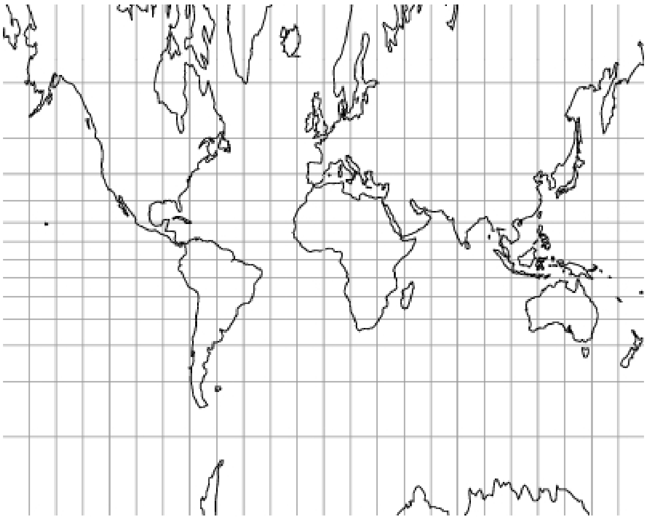


Figure 8.129: **Earth's cylindrical projection**

which results in:



It should be quite obvious (without proof) that this projection does not conserve the areas, but conserve the angles!

8.4.3.3 Mercator projection

A "Mercator projection" is similar in appearance to a cylindrical projection but has a different distortion in the spacing of the lines of longitude (the same idea applies for the so named "Gall stereographic projection" or "Gall–Peters projection" that we will not introduce in this textbook). The Mercator projection is still in the beginning of the 121st century the most common projection used in mapping the Earth onto a flat surface (it advantages, using human bias, the countries habited by white color christian people...).

Like the cylindrical projection North and South are always vertical, we cannot therefore represent the poles because the mathematics have an infinity singularity there.

The result of such a projection is:

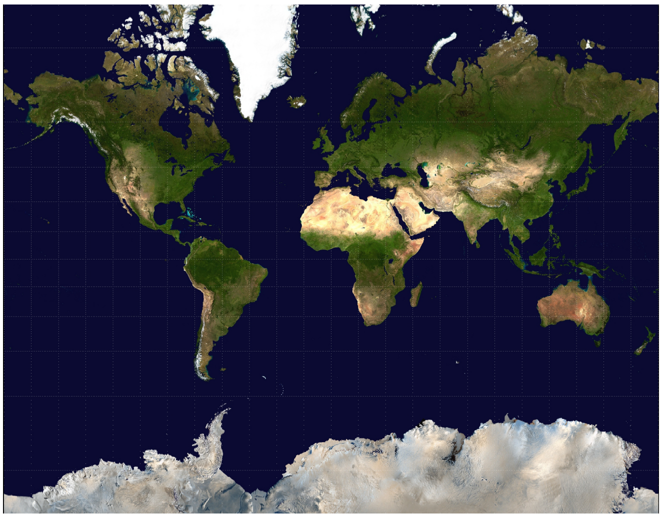


Figure 8.130: **Earth's Mercator projection**

The Mercator map projection with the true sizes and shapes of the countries overlaid:

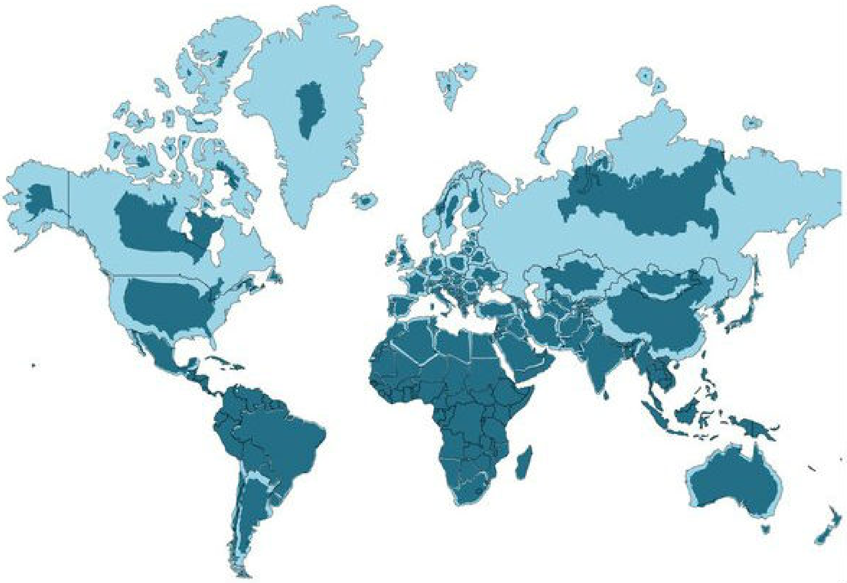


Figure 8.131: **Mercator map projection with the true sizes and shapes of the countries overlaid (author: Neil Kaye)**

As we have seen previously, the cylindrical map projection is specified by formula linking the geographic coordinates of latitude β and longitude α to Cartesian coordinates on the map with origin on the equator and x-axis along the equator. By construction, all points on the same meridian lie on the same generator of the cylinder at a constant value of x, but the distance y along the generator (measured from the equator) is an arbitrary function of latitude, $y(\beta)$. In general this function does not describe the geometrical projection (as of light rays onto a screen) from the center of the globe to the cylinder, which is only one of an unlimited number of ways to conceptually project a cylindrical map.

Since the cylinder is tangential to the globe of radius R at the equator, the scale factor between globe and cylinder is obviously unity on the equator but nowhere else. In particular since the radius of a parallel, or circle of latitude, is $R\cos(\beta)$, the corresponding parallel on the map must have been stretched by a factor of:

$$\frac{R}{R\cos(\beta)}=\frac{1}{\cos(\beta)}=\sec(\beta) \tag{8.4.119}$$

This scale factor on the parallel is conventionally denoted by k and the corresponding scale factor on the meridian is denoted by h.

The relations between $y(\beta)$ and properties of the projection, such as the transformation of angles and the variation in scale, follow from the geometry of corresponding small elements on the globe and map. The figure below shows a point P at latitude β and longitude α on the globe and a nearby point Q at latitude $\phi+\delta\phi$ and longitude $\alpha+\delta\alpha$. The vertical segments $\overset{\frown}{PK}$ and $\overset{\frown}{MQ}$ are arcs of meridians of length $R\delta\beta$ (see the proof in the section Trigonometry). The horizontal segment $\overset{\frown}{PM}$ and $\overset{\frown}{KQ}$ are arcs of parallels of length $R\cos(\beta)\delta\alpha$ (still following from the same proof of the section Trigonometry). The corresponding points on the projection define a rectangle of width δx and height δy:

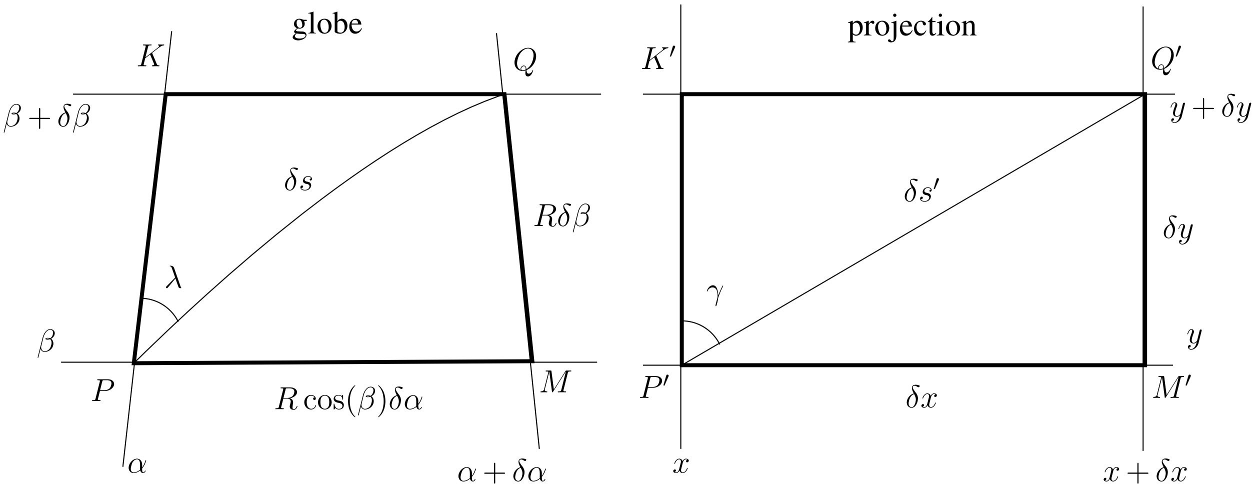


Figure 8.132: **Mercator projection function construction**

For small elements, the angle $\widehat{PKQ}$ is approximately a right angle and therefore:

$$\tan(\lambda)\cong\frac{R\cos(\beta)\alpha}{R\delta\beta}=\frac{\cos(\beta)\delta\alpha}{\delta\beta} \qquad \tan(\gamma)\cong\frac{\delta x}{\delta y} \tag{8.4.120}$$

The previously mentioned scaling factors from globe to cylinder are given by:

- Parallel scale factor:

$$k(\alpha)=\frac{\overline{P'M'}}{\overline{PM}}=\frac{\delta x}{R\cos(\beta)\delta\alpha} \tag{8.4.121}$$

- Meridian scale factor:

$$k(\alpha)=\frac{\overline{P'K'}}{\overline{PK}}=\frac{\delta y}{R\delta\alpha} \tag{8.4.122}$$

Consider now that the projection plane we take has for x-coordinates the range $[-\pi R, +\pi R]$ (the whole being then equal to $2\pi R$). Then for every meridian:

$$x = R(\alpha - \alpha_0) \tag{8.4.123}$$

Indeed, it is easy to see that if $(\alpha - \alpha_0) = 2\pi$ we have well $x = 2\pi r$ when we take time $\alpha_0 = 0$. But in the general case we have obviously:

$$\delta x = R\delta\alpha \tag{8.4.124}$$

Therefore we can write:

$$\tan(\gamma) \cong \frac{\delta x}{\delta y} = \frac{R\delta\alpha}{\delta y} = \frac{R\dfrac{\tan(\lambda)\delta\beta}{\cos(\beta)}}{\delta y} = \frac{R\sec(\beta)}{\delta y/\delta\beta}\tan(\lambda) \cong \frac{R\sec(\beta)}{y'(\beta)}\tan(\lambda) \tag{8.4.125}$$

Where we have considered that:

$$\frac{\delta y}{\delta \beta} \cong \frac{\mathrm{d}y}{\mathrm{d}\beta} = y'(\beta) \tag{8.4.126}$$

The choice of the function $y(\beta)$ for the Mercator projection is determined by the demand that the projection to be conformal[7] (!), a condition which can be defined as the equality of angles as proved earlier above. The condition that a sailing course of constant azimuth λ on the globe is mapped into a constant grid bearing γ on the map. Setting $\lambda = \gamma$ in:

$$\tan(\gamma) \cong \frac{\sec(\beta)}{y'(\beta)}\tan(\lambda) \tag{8.4.127}$$

gives immediately after rearranging:

$$y'(\beta) = R\sec(\beta) \tag{8.4.128}$$

with $y(0) = 0$, by using the primitives proved in the section of Differential and Integral Calculus (see page 1534) we have immediately:

$$y(\beta) = \ln\left(\left|\tan\left(\frac{\beta}{2} + \frac{\pi}{4}\right)\right|\right) \tag{8.4.129}$$

So finally:

$$\boxed{\begin{aligned} &x = (\alpha - \alpha_0) \\ &y = \ln\left(\left|\tan\left(\frac{\beta}{2} + \frac{\pi}{4}\right)\right|\right) \end{aligned}} \tag{8.4.130}$$

In the first equation α_0 is the longitude of an arbitrary central meridian usually, but not always, that of Greenwich (i.e., zero).

---

[7] Very useful for navigators as a few centuries ago they were using mainly magnetic compass (and also star position techniques for double check...) to drive their boat. With a conformal map, you just draw a straight line between two points of the map (departure and arrival points), you measure the angle relatively to the vertical grid-lines of the map and after you just have to keep in mind while driving the boat to conserve this angle with the compass relatively to the North direction to reach your target.

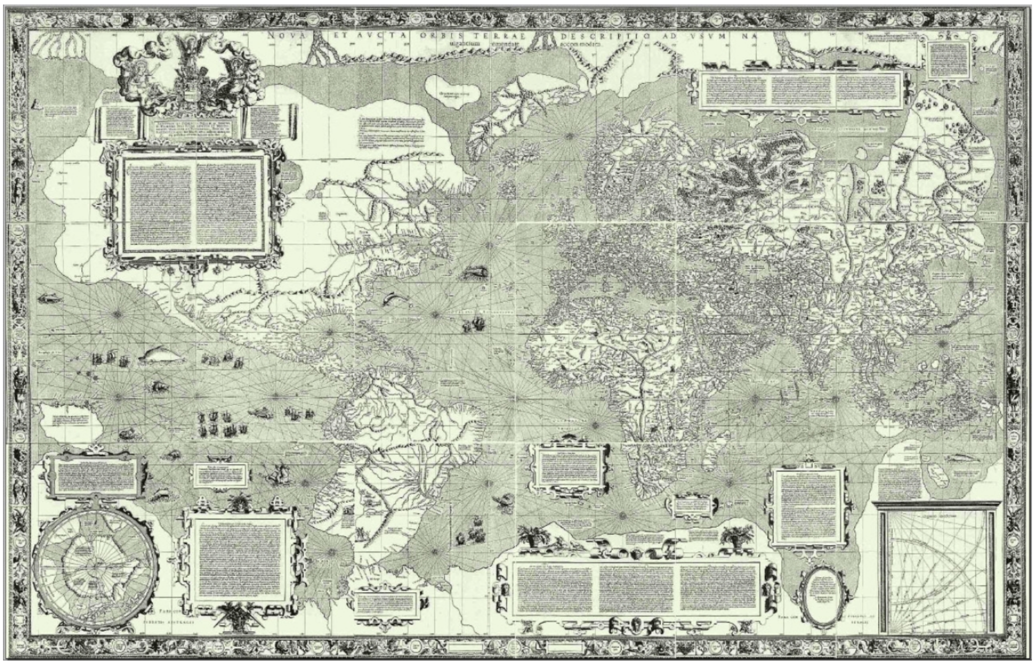


Figure 8.133: **Mercator world map as seen in 11569 (holocene calendar)... (source: Wikipedia)**

8.4.3.4 Lambert's equivalent projection (Peters projection)

The idea behind the "Lambert's equivalent projection", also named "Lambert's cylindrical equal-area projection" or sometimes "Peter's projection" (must not be confused with "Gall's-Peter projection"!), is quite easy as illustrated in the figure below:

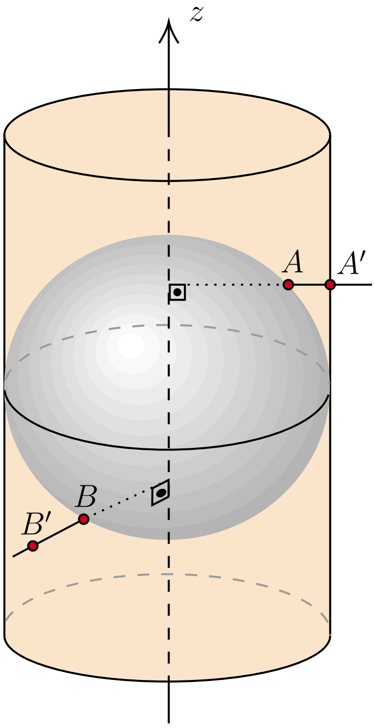
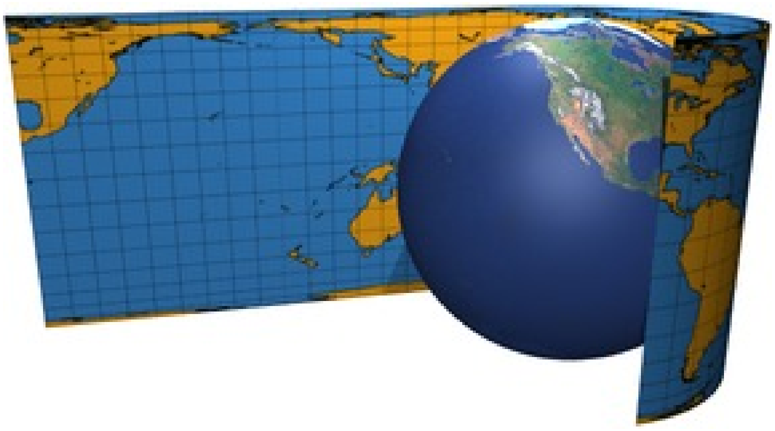


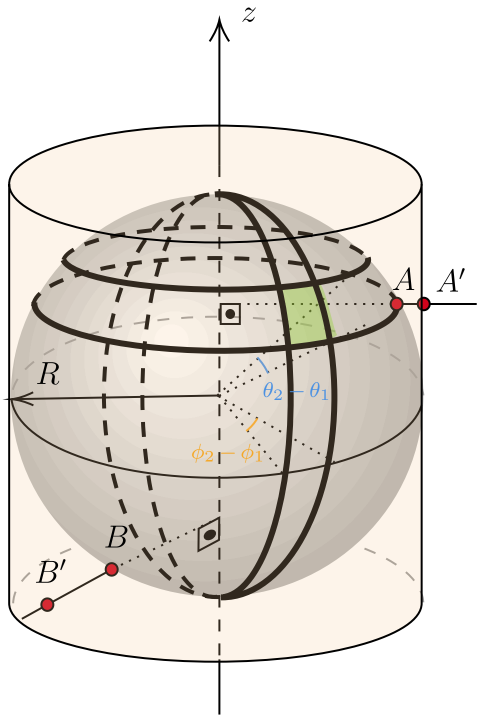
Figure 8.134: **Lambert's equivalent projection (or Lambert's equal area projection)**

Suppose we have a terrestrial globe as visible above. It is enough to wrap a sheet of paper around the globe as in the figure above and to project horizontally each point of the sphere on the cylinder. For that, if A is a point of the sphere, we take the line passing through A and perpendicular to the axis of the poles This line intersects the cylinder at A': the point A' represents the point A on the map.

Or more explicitly:



Now using the following figure, let us prove that the Lambert's cylindrical projection conserve the areas (that's its huge advantage in comparison of the Mercator map that biases the areas of countries), in other words it is a projection named "authalic projection:



For this proof, we will look to a small area of the sphere delimited by the parallels of latitude θ_1 and θ_2 and of the meridians of longitude ϕ_1 and ϕ_2. We will denote by R the radius of the sphere and we will assume that $\theta_2-\theta_1$ and $\phi_2-\phi_1$ are positive and small.

According to this assumption, the area of this region is that of a small rectangle (see figure above). At the latitude θ_1, the radius of the parallel is obviously $R\cos(\theta_1)$. Then by a simple rule of three, the length of the arc from angle $\phi_2 - \phi_1$ is equal to $R(\phi_2 - \phi_1)\cos(\theta_1)$, as the length of the area is the product or the radius by the angle in radian (see section Trigonometry page 2542).

What is now the height of this rectangle? As the meridian is a circle of radius R and that the height of the rectangle is given by angle $\theta_2 - \theta_1$, then this height is given approximately by $R(\theta_2 - \theta_1)$. Then the area of our small region is approximately equal to:

$$R^2(\theta_2 - \theta_1)(\phi_2 - \phi_1)\cos(\theta_1) \tag{8.4.131}$$

Let us calculate now the area of its projection on the cylinder. This projection is now a real rectangle. The parallel has been projected on a circle of radius R. As the corresponding angle has remained the same, that is $\phi_2 - \phi_1$, the width of the rectangle is $R(\phi_2 - \phi_1)$. For the height, the reader must notice that the tangent to the meridian at the latitude θ_1 make an angle of:

$$\frac{\pi}{2} - \theta_1 \tag{8.4.132}$$

with the horizontal. The projection on the vertical of a segment of length $R(\theta_2 - \theta_1)$ is then equal to:

$$R(\theta_2 - \theta_1)\sin\left(\frac{\pi}{2} - \theta_1\right) = R(\theta_2 - \theta_1)\cos(\theta_1) \tag{8.4.133}$$

The area on the map is then equal to:

$$R^2(\theta_2 - \theta_1)(\phi_2 - \phi_1)\cos(\theta_1) \tag{8.4.134}$$

Thus, the same area than that of the small area on the sphere. This finish the proof that Lambert's cylindrical projection conserves the area.

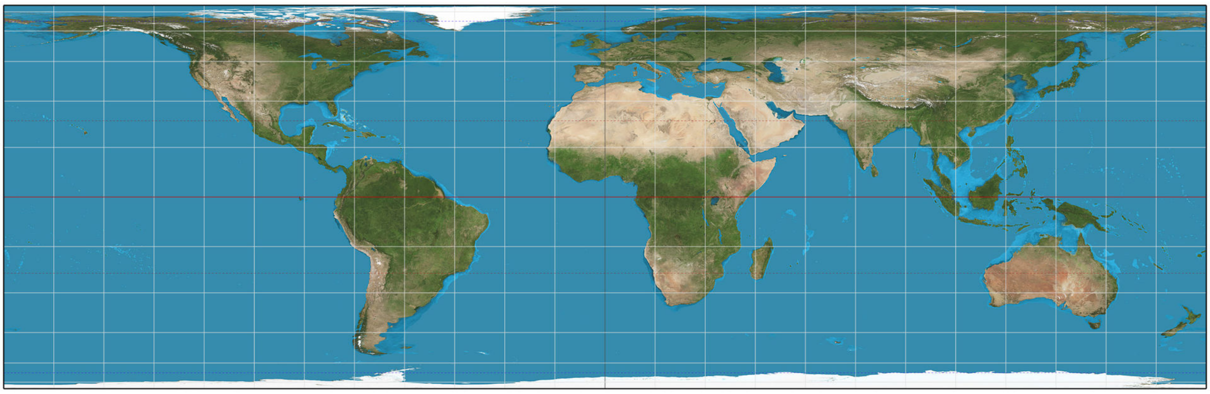


Figure 8.135: **Earth's Lambert projection (source: Wikipedia)**

8.4.3.5 Mollweide projection

The "Mollweide projection" is a very famous equal-area, pseudocylindrical map projection generally used for global maps of the world or night sky (it is also known as the "Babinet projection", "homalographic projection", "homolographic projection" and "elliptical projection"). The projection trades accuracy of angle and shape for accuracy of proportions in area, and as such is used where that property is needed, such as maps depicting global distributions.

The text that follows is entirely inspired from the paper [335] where the method is derived in details using three different proofs. Here we will focus on what seems to us the simplest of the three derivations that make sense.

Pseudocylindrical map projections have in common straight parallel lines of latitude and curved meridians. Until the 119th century (holocene calendar) the only pseudocylindrical projection with important properties was the sinusoidal or Sanson-Flamsteed. The sinusoidal has equally spaced parallels of latitude, true scale along parallels, and equivalency or equal-area. As a world map, it has disadvantage of high distortion at latitudes near the poles, especially those far from the central meridian:

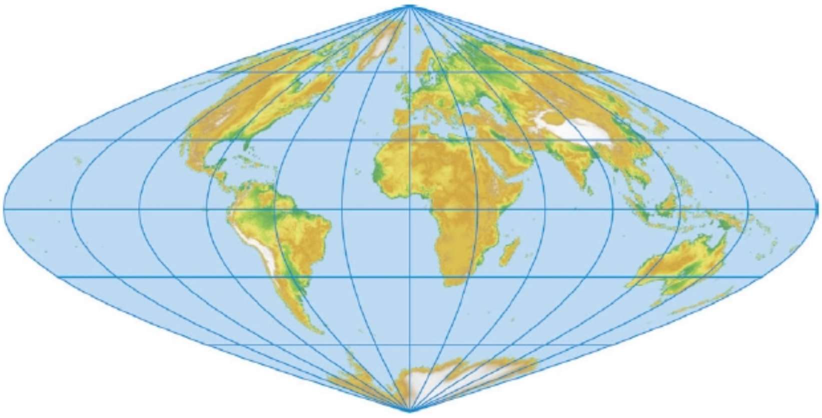


Figure 8.136: **Sansons or Sanson-Flamsteed or Sinuoidal projection**

In 11805 (holocene calendar), Mollweide announced an equal-area world map projection that is aesthetically more pleasing than the sinusoidal because the world is placed in an ellipse with axes in a ratio and all the meridians are equally spaced semi-ellipses. The Mollweide projection was the only new pseudocylindrical projection of the 119th century (holocene calendar) to receive much more than academic interest:

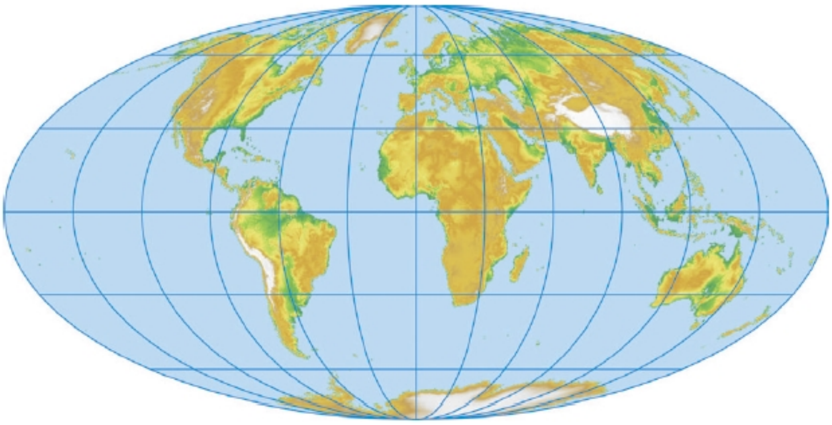


Figure 8.137: **Mollweide projection**

O'Connor and Robertson stated that Mollweide produced the map projection to correct the distortions in the Mercator projection, first used by Gerardus Mercator in 11569 (holocene calendar). While the Mercator projection is well adapted for sea charts, its very great exaggeration of land areas in high latitudes makes it unsuitable for most other purposes. In the Mercator projection the angles of intersection between the parallels and meridians, and the general configuration of the land, are preserved but as a consequence areas and distances are increasingly exaggerated as one moves away from the equator. To correct these defects, Mollweide drew his elliptical projection; but in preserving the correct relation between the areas he was compelled to sacrifice configuration and angular measurement. The Mollweide projection lay relatively dormant until J. Babinet reintroduced it in 11857 (holocene calendar) under the name "homalographic projection".

The well known equations of the Mollweide projections that we will prove further below read as follows:

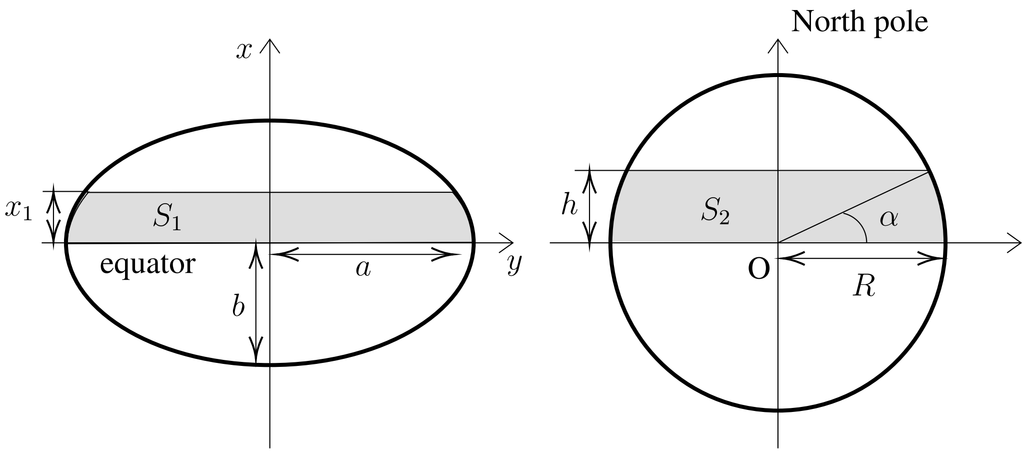
$$\boxed{x=\sqrt{2}R\sin(\beta)\qquad y=\frac{2\sqrt{2}}{\pi}R\lambda\cos(\beta)\qquad 2\beta+\sin(2\beta)=\pi\sin(\varphi)}\tag{8.4.135}$$

In these formulas x and y are rectangular coordinates in the plane of projection, φ (latitude angle) and λ (longitude angle) are geographic coordinates of the points on the sphere and R is the radius of the sphere to be mapped. The angle β is an auxiliary angle that is connected with the latitude φ by the relation $2\beta+\sin(2\beta)=\pi\sin(\varphi)$. For given latitude φ, the equation $2\beta+\sin(2\beta)=\pi\sin(\varphi)$ is a transcendental equation in β. In the past, it was solved by using tables and interpolation method. In our days, it is usually solved by using some iterative numerical method, like bisection or Newton-Raphson method.

Proof 8.19.2. Given the Earth's radius R, suppose the equatorial aspect of an equal-area projection with the following properties:

- A world map is bounded by an ellipse twice broader than tall
- Parallels map into parallel straight lines with uniform scale
- The central meridian is a part of straight line; all other ones are semi-elliptical arcs

Suppose an Earth-sized map; let us define two regions, S_1 on the map and S_2 on the Earth, both bounded by the equator and a parallel (the figure depicts a 2D representation but the left geometry is an ellipse - not an ellipsoid ! - and the right one a sphere):



The equal-area property can be used to calculate x for given φ. Given x and λ (longitude), y can be calculated immediately from the ellipse equation, since horizontal scale is constant. Equation of ellipse centred in origin, with major axis on y-axis is as we already know:

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1 \Rightarrow y^2=b^2\left(1-\frac{x^2}{a^2}\right) \qquad (8.4.136)$$

For $0 \leq x \leq a$:

$$y=\frac{b}{a}\sqrt{a^2-x^2} \qquad (8.4.137)$$

The area between y-axis and parallel mapped into $x=x_1$ is:

$$S_1=2\int_0^{x_1} y\mathrm{d}x=2\frac{b}{a}\int_0^{x_1}\sqrt{a^2-x^2}\mathrm{d}x \qquad (8.4.138)$$

Let us introduce the auxiliary angle:

$$x=a\sin(\beta) \qquad 0\leq x\leq a, 0\leq\beta\leq\frac{\pi}{2} \qquad (8.4.139)$$

and:

$$\mathrm{d}x=a\cos(\beta)\mathrm{d}\beta \qquad (8.4.140)$$

then:

$$\int\sqrt{a^2-x^2}\mathrm{d}x=\int\sqrt{a^2\left(1-\sin^2(\beta)\right)}a\cos(\beta)\mathrm{d}\beta=a^2\int\cos^2(\beta)\mathrm{d}\beta \qquad (8.4.141)$$

Since:

$$\cos^2(\alpha)=\frac{1+\cos(2\alpha)}{2} \qquad (8.4.142)$$

Then:

$$\begin{aligned}a^2\int\cos^2(\beta)\mathrm{d}\beta&=a^2\int\frac{1+\cos(2\beta)}{2}\mathrm{d}\beta=\frac{a^2}{2}\left(\int\mathrm{d}\beta+\int\cos(2\beta)\mathrm{d}\beta\right)\\&=\frac{a^2}{2}\left(\beta+\frac{\sin(2\beta)}{2}\right)+c^{te}\end{aligned}\tag{8.4.143}$$

Therefore:

$$S_1=2\frac{b}{a}\left(\frac{a^2}{2}\left(\beta+\frac{\sin(2\beta)}{2}\right)+c^{te}\right)_0^\beta=\frac{ab}{2}(2\beta+\sin(2\beta))=2R^2(2\beta+\sin(2\beta))\tag{8.4.144}$$

for some $0\leq\beta\leq\pi/2$, corresponding to $x_1=a\sin(\beta)$ and because of $ab\pi=4R^2\pi$.

Now let us calculate the "spherical zone":

$$\begin{aligned}S_2&=\int_{\theta=\mathrm{lat}_1}^{\mathrm{lat}_2}2\pi r'\mathrm{d}\theta=\int_{\theta=\mathrm{lat}_1}^{\mathrm{lat}_2}2\pi r\cos(\theta)\mathrm{d}\theta=2\pi r\int_{\theta=\mathrm{lat}_1}^{\mathrm{lat}_2}\cos(\theta)\mathrm{d}\theta\\&=2\pi r(-\sin(\theta))\Big|_{\theta=\mathrm{lat}_1}^{\mathrm{lat}_2}=2\pi r(\sin(\mathrm{lat}_1)-\sin(\mathrm{lat}_2))\end{aligned}\tag{8.4.145}$$

Therefore on a sphere, the area between the equator and parallel φ (i.e. $\mathrm{lat}_2=0$) is:

$$S_2=2\pi Rh=2\pi R^2\sin(\varphi)\tag{8.4.146}$$

But as $S_1=S_2$, then:

$$2R^2(2\beta+\sin(2\beta))=2\pi R^2\sin(\varphi)\tag{8.4.147}$$

Hence:

$$S_1=S_2\Leftrightarrow 2R^2(2\beta+\sin(2\beta))=2\pi R^2\sin(\varphi)\tag{8.4.148}$$

Simplifying we fall back the third relation of the Mollweide projection:

$$2\beta+\sin(2\beta)=\pi\sin(\varphi)\tag{8.4.149}$$

The auxiliary angle β, as already mentioned, must be found by interpolation or successive approximation.

Finally, since horizontal scale is uniform, from the two equalities (remember that the left $ab\pi$ is the surface of the elliptical projection 2D and the right $4R^2\pi$ is the surface of the projected sphere!):

$$ab\pi=4R^2\pi\qquad b=2a\tag{8.4.150}$$

We get:

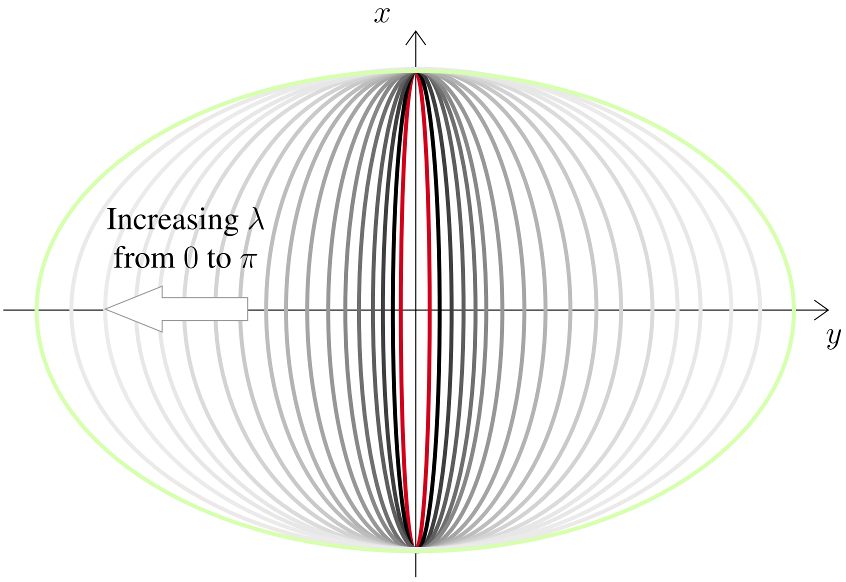
$$a=\sqrt{2}R\tag{8.4.151}$$

and therefore we can fall back on the first relation of the Mollweide projection:

$$x=a\sin(\beta)=\sqrt{2}R\sin(\beta)\tag{8.4.152}$$

Now, remember that λ is the longitude that goes starting from 0 [rad], to $-\pi$ on the West to $+\pi$ on the East. As it is symmetrical, from the meridian 0 [rad] just having one of the value of the angle going to West or East in enough information to make the mapping on the ellipse (as it is symmetrical!). The clever idea now is to shrink the elliptical projection so that when the latitude is near zero we get a very small projection ellipse as visible in red below and when it is equal to $+\pi$ we get the whole projection ellipse in green below (remember we do that in the horizontal direction of y). So multiplying the elliptic function y by the ratio λ/π will give the expected shrinkage:

$$y=\left(\frac{b}{a}\sqrt{a^2-x^2}\right)\frac{\lambda}{\pi} \tag{8.4.153}$$



Therefore:

$$y=\frac{2\lambda}{\pi}\sqrt{2R^2-x^2}=2\sqrt{2R^2-2R^2\sin^2(\beta)}\frac{\lambda}{\pi}=\frac{2\sqrt{2}\lambda}{\pi}R\sqrt{1-\sin^2(\beta)} \tag{8.4.154}$$

Then we fall back on the second relation of the Mollweide projection:

$$y=\frac{2\sqrt{2}}{\pi}R\lambda\cos(\beta) \tag{8.4.155}$$

We have all of the three relations and this finish the proofs!

■ Q.E.D.

Although the applied condition was that a half of the sphere has to be mapped onto a disc, the final projection equations hold for the whole sphere and give its image situated into an ellipse.

Let us consider the shape of the Mollweide projection of the whole sphere. From:

$$x=\sqrt{2}R\sin(\beta) \qquad y=\frac{2\sqrt{2}}{\pi}R\lambda\cos(\beta) \tag{8.4.156}$$

it is easy to get the equation of a meridian in the projection:

$$\left(\frac{x}{\sqrt{2}R}\right)^2+\left(\frac{\pi y}{2\sqrt{2}R\lambda}\right)^2=1 \tag{8.4.157}$$

It is obvious that for a given λ, the relation above is the equation of an ellipse. It follows that the semi-axis a is constant, while b depends on the longitude λ. If we take $\lambda=\pi$, than $b=2\sqrt{2}R$, and $a:b=1:2$ and that is the ratio of semi-axes in the Mollweide projection. The question arises: is it possible to find out a pseudocylindrical equal-area projection that will give the whole world in an arbitrary ellipse satisfying any given ratio $a:b$ or $b:a$? The answer is yes (!) and this leads to "Generalized Mollweide projection". But as the proof is of no interest in this textbook, the reader can refer to [335] for a detailed proof of it.

So the Mollweide is a pseudocylindrical projection in which the equator is represented as a straight horizontal line perpendicular to a central meridian one-half its length. The other parallels compress near the poles, while the other meridians are equally spaced at the equator. The meridians at 90 degrees East and West form a perfect circle, and the whole Earth is depicted in a proportional $2:1$ ellipse.

Notice that the Mollweide and Hammer projections are occasionally confused, since they are both equal-area and share the elliptical boundary; however, the latter design has curved parallels and is not pseudocylindrical:

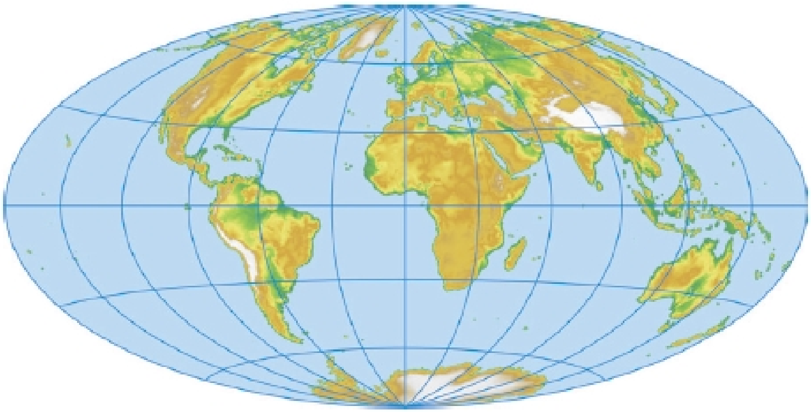


Figure 8.138: **Hammer projection**

8.4.4 Other perspectives and Laws of linear perspective

We can also choose any other angle of perspective and that not keep anything as equal but that are just well adapted to the technology or to human perception. For example the game Diablo II or Age of Empire use an angle in degrees of $53.130102°$ (rounded to 53) as for old $4:3$[8] screens it gives as result that an object have a vanishing line on the upper left corner of the $4:3$ screen will also pass trough the bottom right corner of the screen. It is therefore just choice of comfort!

---

[8] The origin of the $4:3$ is interesting and useful as the ratio forms a Pythagorean triangle, meaning that the length of the diagonal (hypotenuse) is an exact integer value (of pixels...).

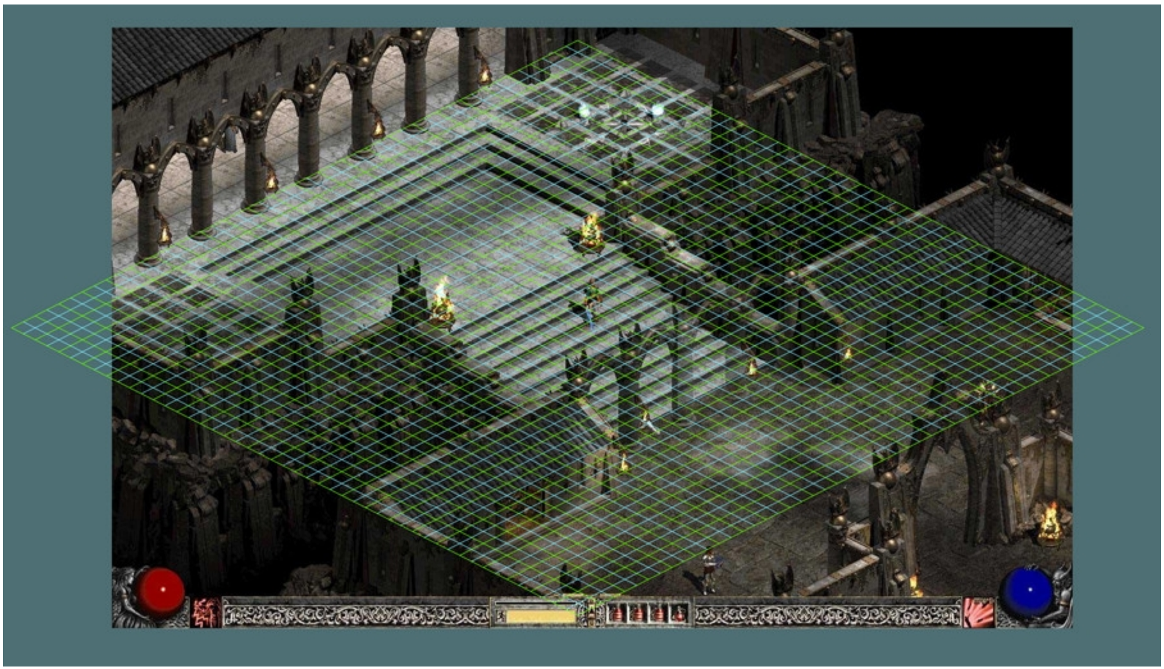


Figure 8.139: **Diablo II perspective grid**

Alternatively there are also a few games which use a trimetric projection where one axis is $1:2$ and the other is $2:1$ as below:

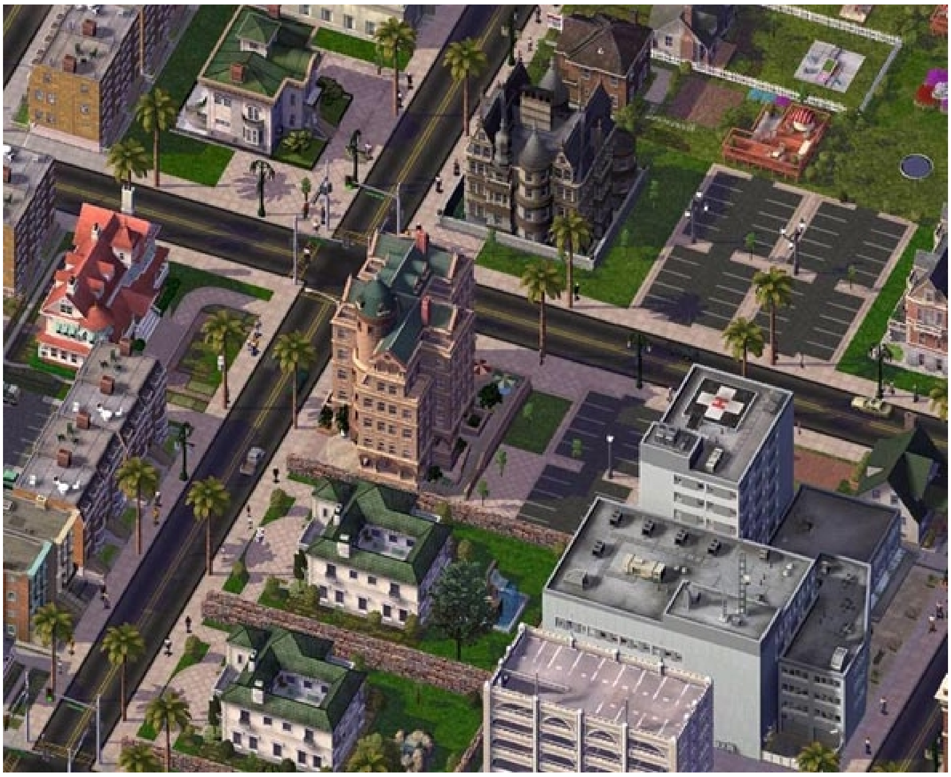


Figure 8.140: **Example of trimetric projection**

To resume we have seen so far the following projections:

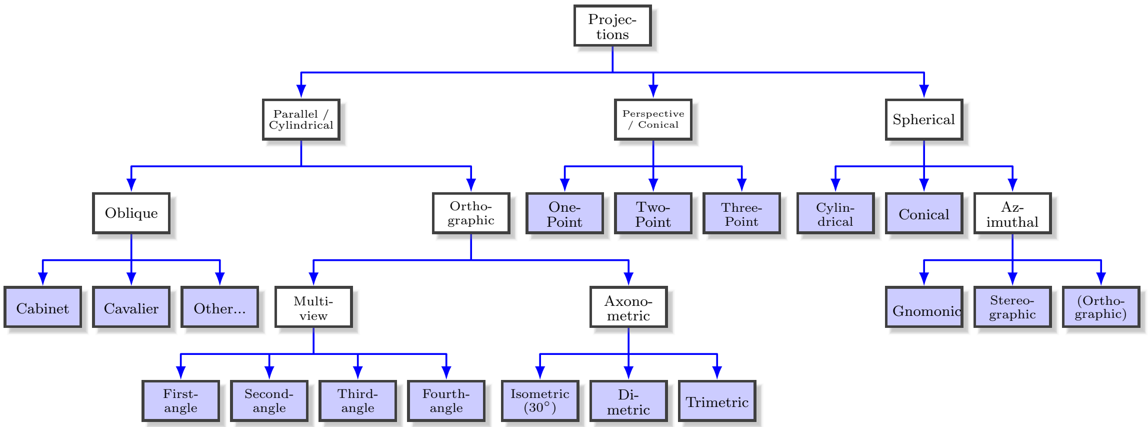

Projec-
tions
Parallel /
Cylindrical
Perspective
/ Conical
Spherical
Oblique
Ortho-
graphic
One-
Point
Two-
Point
Three-
Point
Cylin-
drical
Conical
Az-
imuthal
Cabinet
Cavalier
Other...
Multi-
view
Axono-
metric
Gnomonic
Stereo-
graphic
(Ortho-
graphic)
First-
angle
Second-
angle
Third-
angle
Fourth-
angle
Isometric
(30°)
Di-
metric
Trimetric

Figure 8.141: Non-exhaustive orgchart of various projections methods

And a non-exhaustive visual summary:

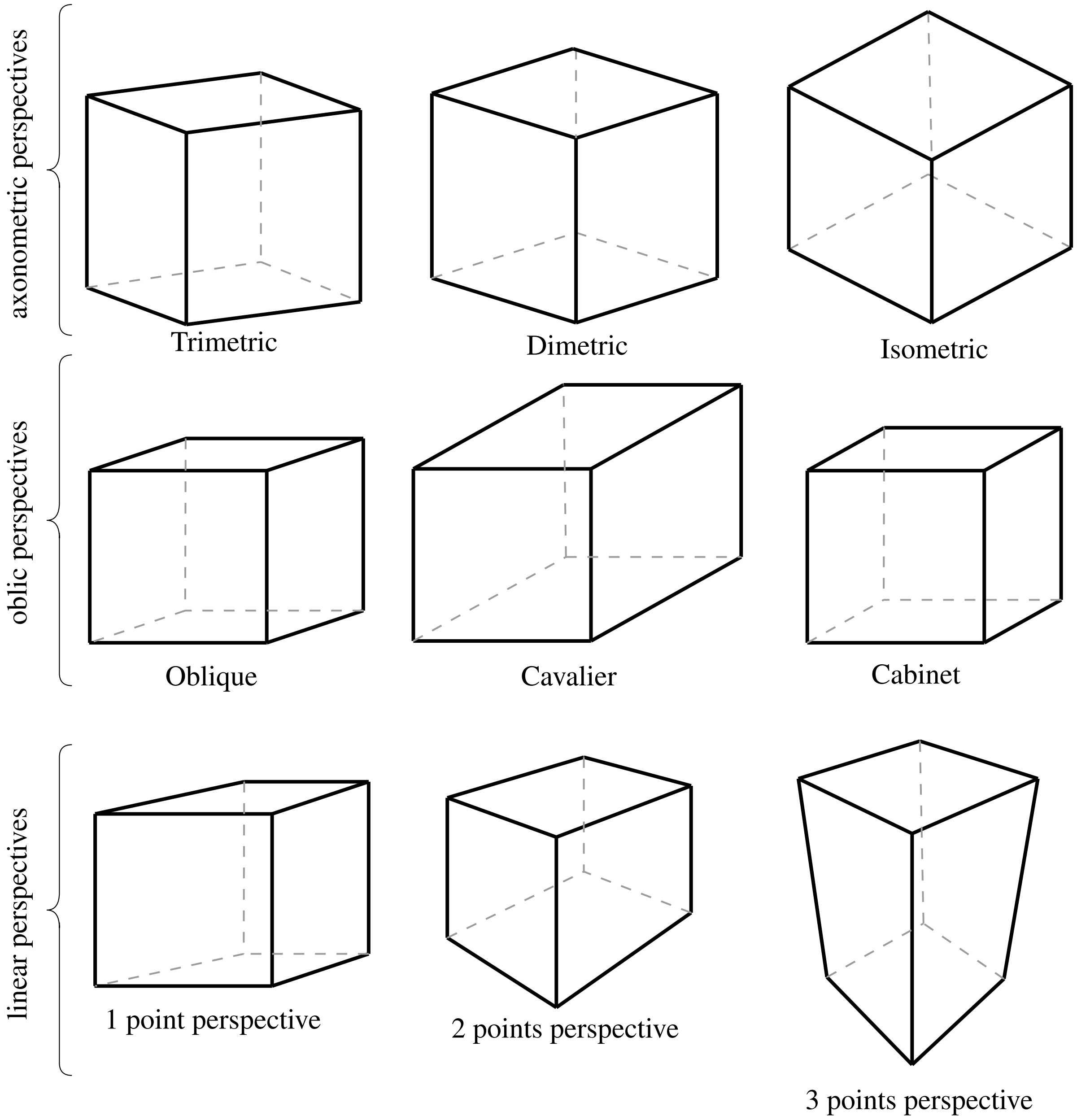

axonometric perspectives
Trimetric
Dimetric
Isometric
oblic perspectives
Oblique
Cavalier
Cabinet
linear perspectives
1 point perspective
2 points perspective
3 points perspective

Figure 8.142: **Some common types of projections**

Now the reader has enough knowledge to fully understand the laws of linear perspective:

L1. **The law of horizon line:** This law states that there is only one horizon line (but there can be no horizon line like in the intergalactic outer space).

L2. **The law of eye height:** The horizon line, when it exists, is the line of the eye level of the observer (and obviously the eye height is independent of gaze direction!).

L3. **The law of vanishing points:** This law states that all straight lines that defines the contour of a shape or object, if extended, arrive on the horizon line at a vanishing point when we consider conical projection/perspective (and therefore all parallel straight lines defining the contour of a shape or object converge on the horizon line at the same vanishing points[9]).

Notice that vanishing points don't have to be aligned on the horizon line even if it is most common case!

When the vanishing point is placed at the zenith (above the horizon), we look up to see the object. We speak then of "worm's eye view", since we are in the position of a worm looking towards the zenith. All vanishing lines above the horizon line are descending lines.

When the vanishing point is placed at the nadir (below the horizon), we look down. This is named the "bird's eye view". All vanishing lines below the horizon line are ascending lines.

Remark

In two-point perspective, we draw vertical lines, vertical! In three-point perspective, our vertical lines are at an angle, converging on the third vanishing point above or below. Two-point perspective is generally sufficient for most basic scenes, while three-point perspective corresponds to reality (because two-point perspective does not account for vertical distortion).

L4. **The law of diminution:** This law states that as objects get farther away from us they appear to get small and can even seems to vanish if their size is smaller than the resolving power of the human eye.

In perfect conditions, the human eye has a resolving power of 60 arc seconds (i.e. angular resolution of 0.008 degrees). In an older person, however, the ability to resolve certain objects is somewhat reduced. This is mainly due to the reduced size of the pupil, but also to the constraints of anatomical and physiological vision.

L5. **The law of Overlap:** This law states that when one object is in front of another, it will cover it and describe their spatial relationships. If the projection plane is curved (positive curvature like a sphere for example) the landscape may be in front of objects that are further than the horizon.

L6. **The law of Convergence:** This law states that as parallel lines go away from us, they appear to converge, or meet, at a point when we consider conical projection/perspective. We can also say this in reverse: When parallel lines come toward us, they diverge, or get farther apart.

L7. **The law of Atmospheric perspective:** This law describes the properties that affect an objects appearance as it moves away in an environment having an isotropic and homogeneous atmosphere and standard pressure and temperature:

- Light and dark variation of a same color get closer creating less contrast (the light is diffused in an atmosphere)
- Colors get less saturated (for the same reasons as the previous point)
- Colors get cooler, shifting towards blues and purples in a typical Earth like atmosphere during a sunny day.

---

[9] All vanishing straight lines below the horizon line are said to be "ascending lines" and all vanishing straight lines below the horizon line are said to be "descending lines".

- Edges gets softer (again because of the resolving power of the human eye or any other intermediate apparatus)
- We see less detail until the object cannot be distinguished from its background (resolving power again) or disappear below the horizon when the projective plan is a geometry with positive curvature.

Below we have an illustrated summary of these seven laws:

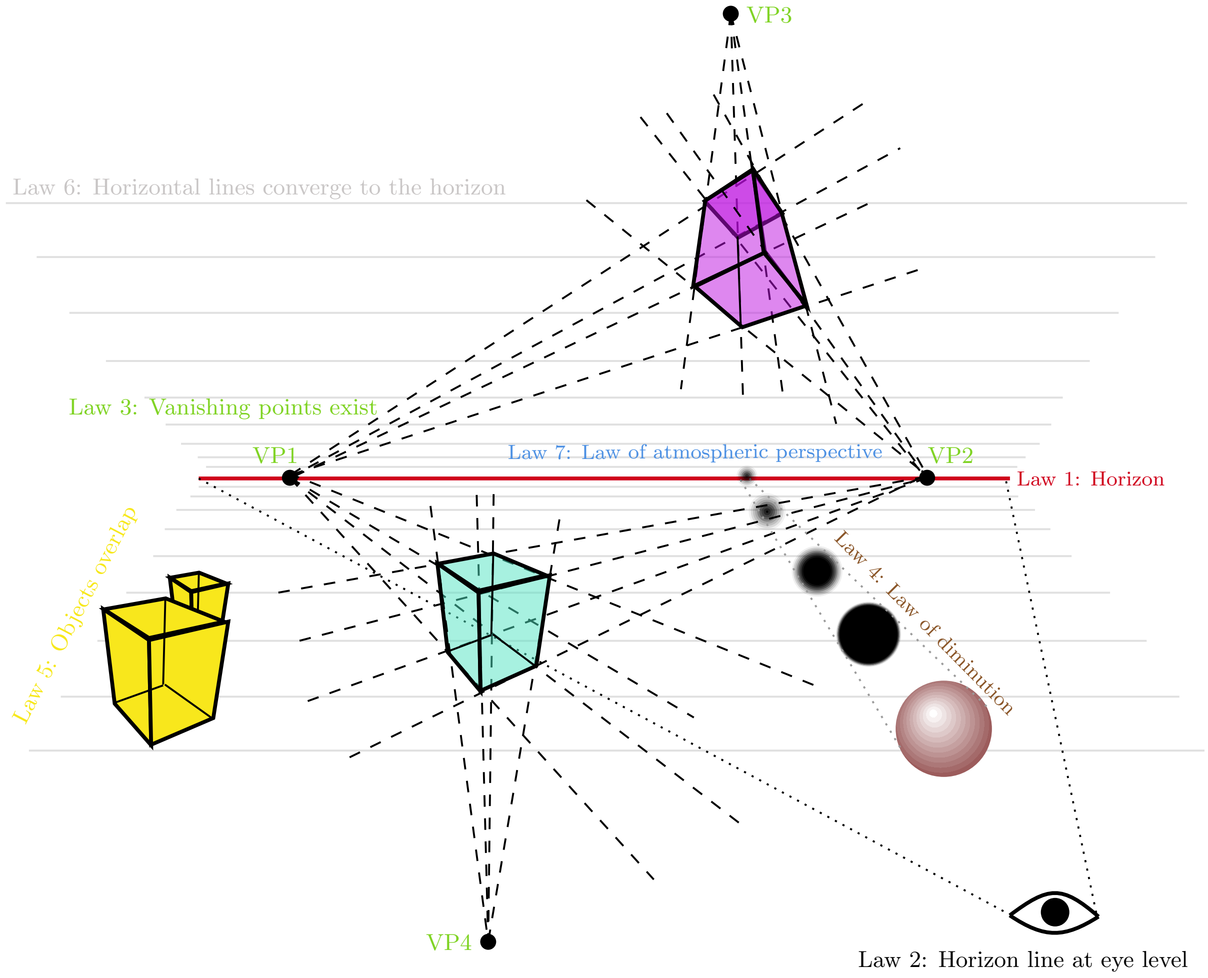

VP3
Law 6: Horizontal lines converge to the horizon
Law 3: Vanishing points exist
VP1
Law 7: Law of atmospheric perspective
VP2
Law 1: Horizon
Law 5: Objects overlap
Law 4: Law of diminution
VP4
Law 2: Horizon line at eye level

Figure 8.143: Summary of the seven laws of linear perspective

Two parallels lines with the laws of perspective seem to converge when in fact they don't:

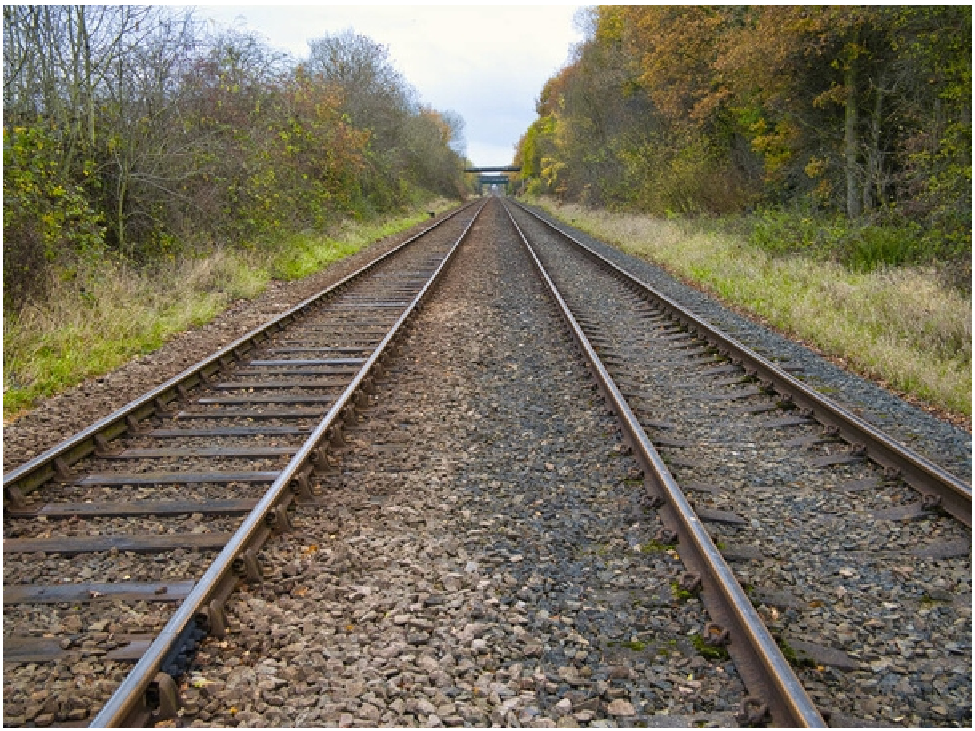


Figure 8.144: **Railway convergence perspective effect**

The same effect of perspective occurs with crepuscular rays that seem to converge locally when in fact they are almost parallel:

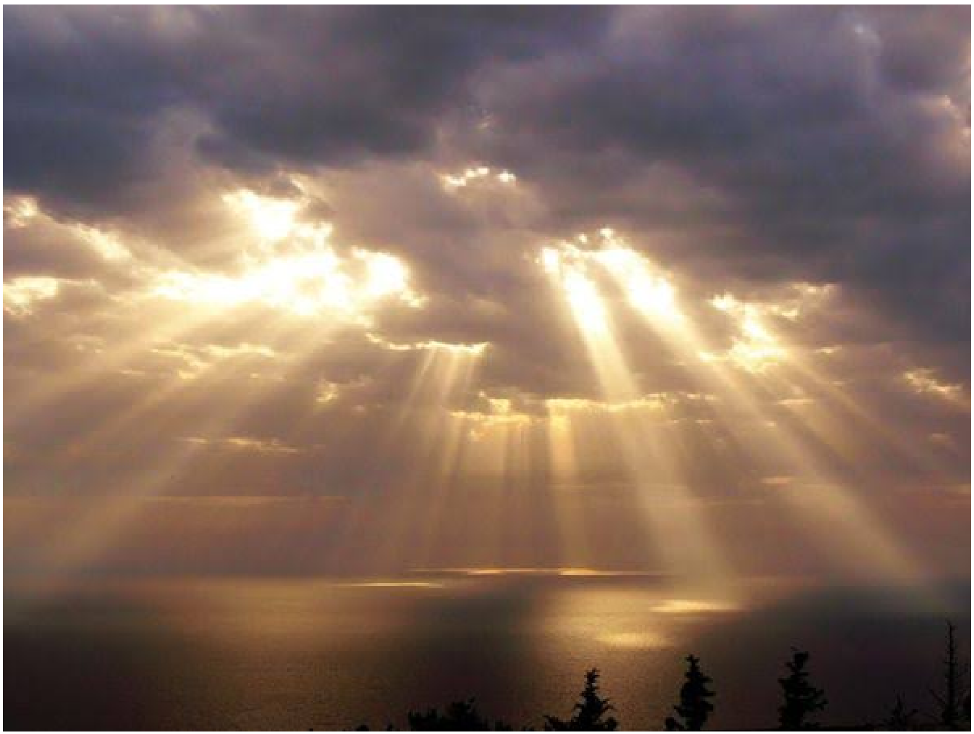


Figure 8.145: **Crepuscular rays**

And the same goes for shadows that seems to diverge (when in fact they converge towards the source) to the left for objects that are on the left and diverge to the right for objects that are on the right:

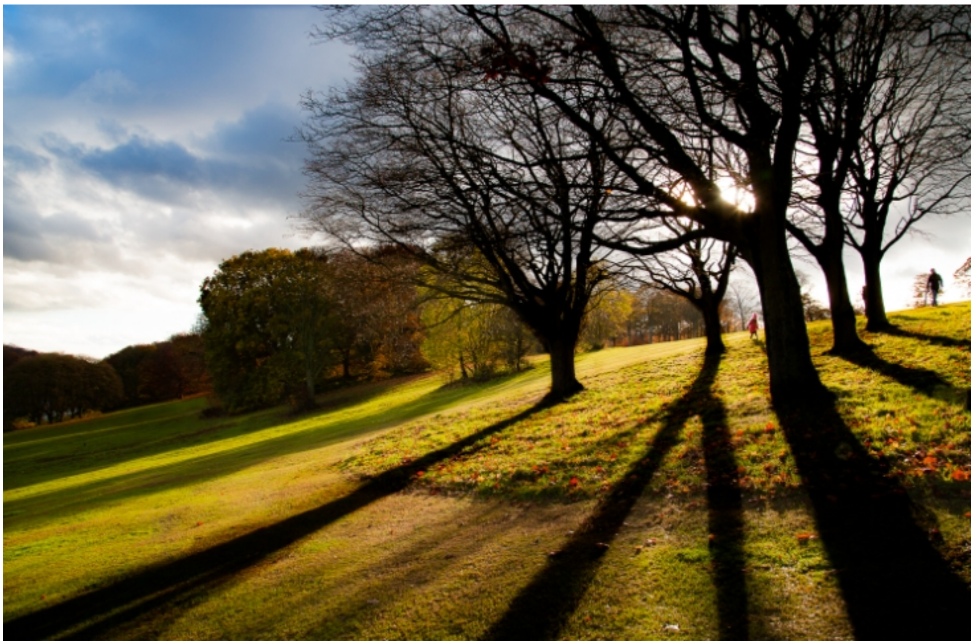
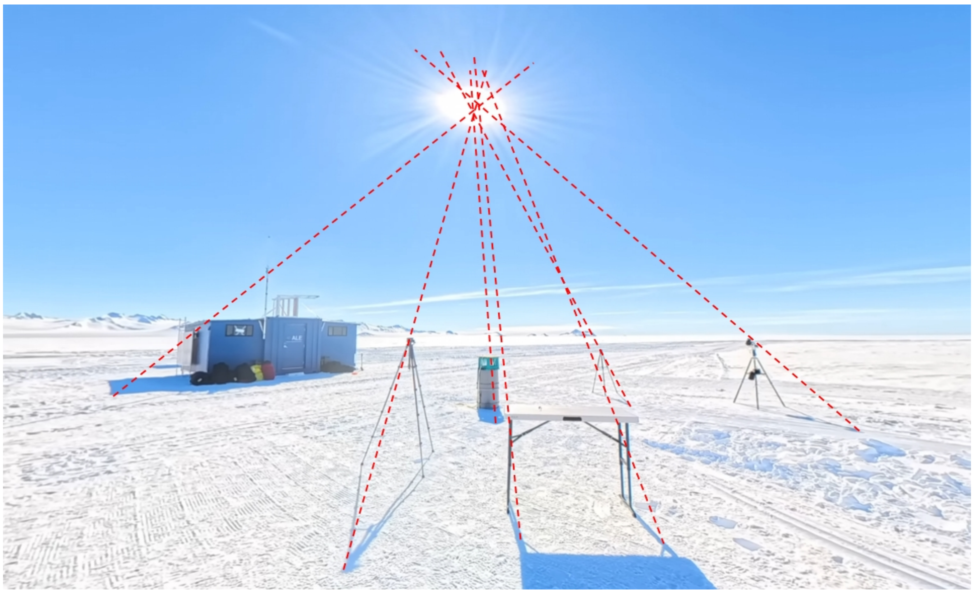


Figure 8.146: Shadows that seem to diverge

or the shadows in Antarctica during the Midnight Sun (here with a wide-angle camera that distorts the objects a little) if we trace the vanishing lines of the shadows we see that although they give the impression of converging towards a local Sun in accordance with the laws of linear perspective (while we will prove in the section of Astronomy that the Sun is about 150 million kilometers away) and that the shadows seem to diverge on the ground:



Definition (#8.4.355): The "midnight Sun" is the period when the entire Sun disc (determined thanks to a solar filter) remains continuously visible for 24 hours or more above the horizon, occurring in summer at latitudes north of the Arctic Circle (66°33′ N) and south of the Antarctic Circle (66°33′ S), due to the tilt of Earth's axis (sadly some scientists use a slightly different non-academic definition hence a variation of a few days in period value).

8.4.5 Homogeneous Coordinates (projection coordinates)

In mathematics, "homogeneous coordinates" also named "projection coordinates", introduced by August Ferdinand Möbius in the 119th century (holocene calendar), make calculations possible in the projective space as do the Cartesian coordinates in Euclidean space. Homogeneous coordinates are used extensively in computer graphics and more particularly for the representation of three dimensional (3D) scenes because they are adapted to projective geometry and allow to characterize the transformations of space under optimal algorithmic form. The matrix notation form is particularly used in 3D graphics programming libraries such as OpenGL and Direct3D.

We proved in our study of transformations in the plane and space (see section Euclidean Geometry page 2640) that the translation, the scaling, rotation or reflection were not possible to be translate represented in (real) matrix form without a trick that was to add a dummy extra dimension to the vector of coordinates and to the associated matrix transformation but we did it only for the translation. So let us see it now for all other classical transformation.

8.4.5.1 $\mathcal{P}^2$ Projective Space

Thus, we saw that the translation in $\mathbb{R}^2$ could be written in matrix form:

$$\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = T \cdot \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{bmatrix} 1 & 0 & T_x \\ 0 & 1 & T_y \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} x+T_x \\ y+T_y \\ 1 \end{pmatrix} \tag{8.4.158}$$

Therefore, always in $\mathbb{R}^2$, a scaling of factor k who was written:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{bmatrix} k & 0 \\ 0 & k \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} k \cdot x \\ k \cdot y \end{pmatrix} \tag{8.4.159}$$

becomes in homogeneous coordinates:

$$\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = S \cdot \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{bmatrix} S_x & 0 & 0 \\ 0 & S_y & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \begin{pmatrix} S_x \cdot x \\ S_y \cdot y \\ 1 \end{pmatrix} \tag{8.4.160}$$

Hence, always in $\mathbb{R}^2$, a rotation of angle θ in the $x-y$ plane:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x\cos(\theta) - y\sin(\theta) \\ x\sin(\theta) + y\cos(\theta) \\ 1 \end{pmatrix} \tag{8.4.161}$$

becomes (and as we proved it, this is indeed a rotation around the z-axis):

$$\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = R \cdot \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} x\cos(\theta) - y\sin(\theta) \\ x\sin(\theta) + y\cos(\theta) \\ 1 \end{pmatrix} \tag{8.4.162}$$

So always in $\mathbb{R}^2$, the reflection that was written along the x-axis by:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \begin{pmatrix} -x \\ y \end{pmatrix} \tag{8.4.163}$$

becomes:

$$\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = \Delta \cdot \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} -1 & 0 & \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \begin{pmatrix} -x \\ y \\ 1 \end{pmatrix} \qquad (8.4.164)$$

and so on for other transformations that we have proved and that were resumed in the following figure we have already see and where the homogeneous coordinates are in gray:

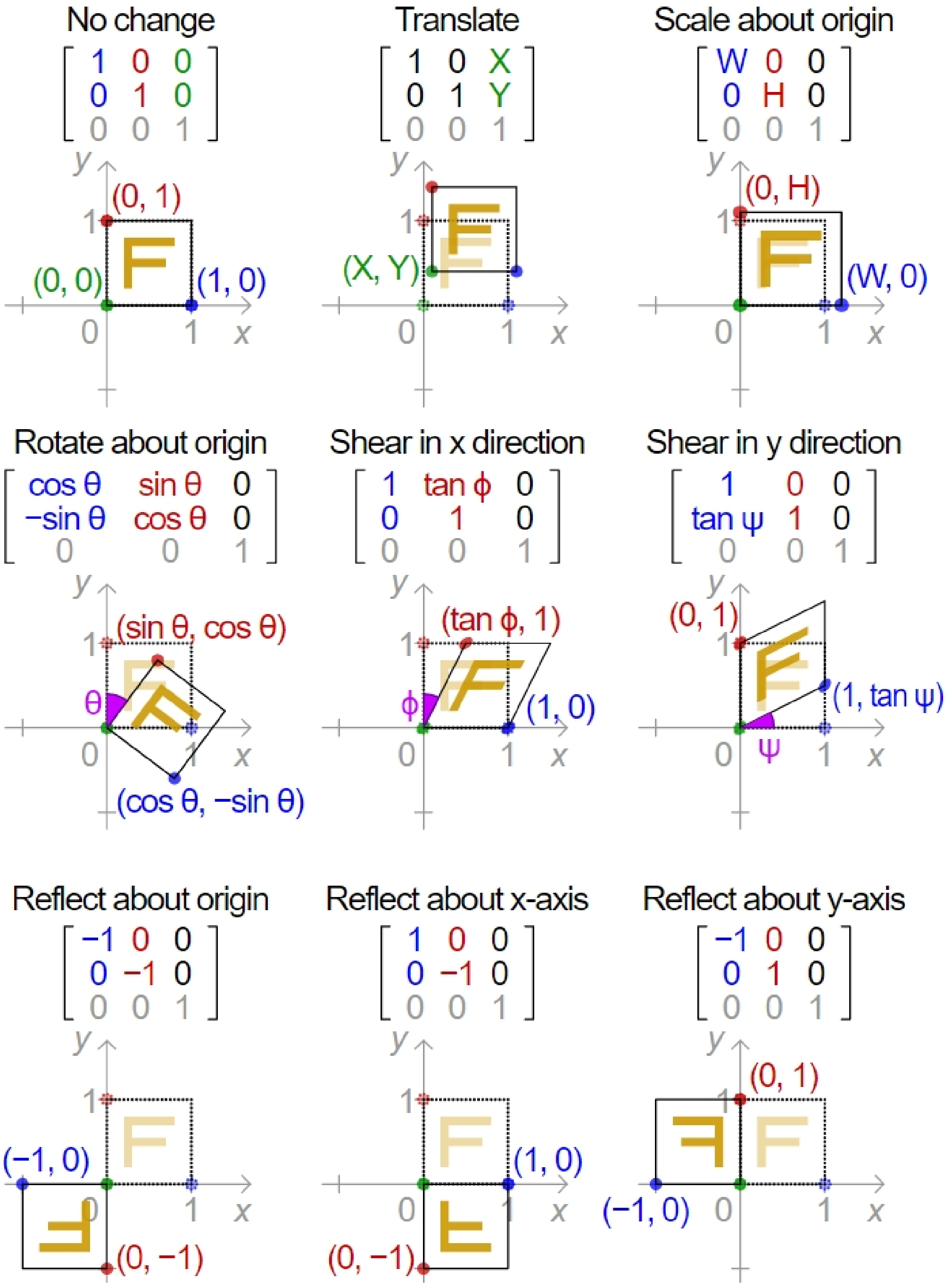

No change
1 0 0
0 1 0
0 0 1
(0, 1)
(0, 0)
(1, 0)
Translate
1 0 X
0 1 Y
0 0 1
(X, Y)
Scale about origin
W 0 0
0 H 0
0 0 1
(0, H)
(W, 0)
Rotate about origin
cos θ sin θ 0
−sin θ cos θ 0
0 0 1
(sin θ, cos θ)
θ
(cos θ, −sin θ)
Shear in x direction
1 tan ϕ 0
0 1 0
0 0 1
(tan ϕ, 1)
ϕ
(1, 0)
Shear in y direction
1 0 0
tan ψ 1 0
0 0 1
(0, 1)
(1, tan ψ)
ψ
Reflect about origin
−1 0 0
0 −1 0
0 0 1
(−1, 0)
(0, −1)
Reflect about x-axis
1 0 0
0 −1 0
0 0 1
(1, 0)
(0, −1)
Reflect about y-axis
−1 0 0
0 1 0
0 0 1
(0, 1)
(−1, 0)
x
y
0
1

Figure 8.147: **Summary of homogeneous coordinates transformations**

Mathematicians say then that we when we use the above transformations or the composition of some of them, we put ourselves in the projective space $\mathcal{P}^2$.

8.4.5.2 $\mathcal{P}^3$ Projective Space

Verbatim, we can do the same with a point P of $\mathbb{R}^3$ which will then in $\mathcal{P}^3$ be represented by a 4×1 coordinate vector:

$$\begin{pmatrix} x\\ y\\ z\\ w \end{pmatrix} \tag{8.4.165}$$

Thus we have in space the following transformation matrices:

- For translations:

$$T=\begin{bmatrix} 1 & 0 & 0 & T_x\\ 0 & 1 & 0 & T_y\\ 0 & 0 & 1 & T_z\\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{8.4.166}$$

- For rotations it is enough for sure in 3D to have as we proved it:

$$R_{ZYX}=R_ZR_YR_X=\begin{bmatrix} \cos(\phi) & \sin(\phi) & 0\\ -\sin(\phi) & \cos(\phi) & 0\\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} \cos(\gamma) & 0 & -\sin(\gamma)\\ 0 & 1 & 0\\ \sin(\gamma) & 0 & \cos(\gamma) \end{bmatrix}\begin{bmatrix} 1 & 0 & 0\\ 0 & \cos(\theta) & \sin(\theta)\\ 0 & -\sin(\theta) & \cos(\theta) \end{bmatrix} \tag{8.4.167}$$

  But if we want to work with projection coordinates, this becomes obviously:

$$R_{ZYX}=\begin{bmatrix} \cos(\phi) & \sin(\phi) & 0 & 0\\ -\sin(\phi) & \cos(\phi) & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} \cos(\gamma) & 0 & -\sin(\gamma) & 0\\ 0 & 1 & 0 & 0\\ \sin(\gamma) & 0 & \cos(\gamma) & 0\\ 0 & 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 & 0\\ 0 & \cos(\theta) & \sin(\theta) & 0\\ 0 & -\sin(\theta) & \cos(\theta) & 0\\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{8.4.168}$$

- For homoteties (scaling) obviously:

$$S=\begin{bmatrix} S_x & 0 & 0 & 0\\ 0 & S_y & 0 & 0\\ 0 & 0 & S_z & 0\\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{8.4.169}$$

- In space, one can push in two coordinate axis directions and keep the third one fixed. The following is the shear transformation in both $x-$ and $y-$directions with shearing factors a and b, respectively, keeping the $z-$ coordinate the same (so its only one of the possible special cases):

$$\begin{pmatrix} x'\\ y'\\ z'\\ 1 \end{pmatrix}=\begin{bmatrix} 1 & 0 & a & 0\\ 0 & 1 & b & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1 \end{bmatrix}\begin{pmatrix} x\\ y\\ z\\ 1 \end{pmatrix}=\begin{pmatrix} x+az\\ y+bz\\ z\\ 1 \end{pmatrix} \tag{8.4.170}$$

  Thus, a point (x,y,z) in space is transformed to $(x+az,y+bz,z)$. Therefore, the z-coordinate does not change, while (x,y) is "pushed” in the direction of $(a,b,0)$ with a factor z.

Mathematicians say then that we when we use the above transformations or the composition of some of them, we put ourselves in the projective space $\mathcal{P}^3$.

We have proved above that in the case of the conical project, we had the following homographic transformations:

$$\begin{aligned} x' &= \frac{x\Delta}{z+\Delta} \\ y' &= \frac{y\Delta}{z+\Delta} \\ z' &= \Delta \end{aligned} \tag{8.4.171}$$

where x', y', z' are the projection of the relation points x, y, z on the table with a focal distance Δ for recall...

What is traditionally written by putting $z := z+\Delta$ and $f := \Delta$:

$$\begin{aligned} x' &= \frac{x\cdot f}{z} \\ y' &= \frac{y\cdot f}{z} \\ z' &= f \end{aligned} \tag{8.4.172}$$

where we see that if the focal distance f is infinite, the subject coincides with the XY plane (or if the object is infinitely far away as it is equivalent...).

Let us put:

$$\frac{z}{f} = s \tag{8.4.173}$$

We can the write:

$$\begin{aligned} sx' &= x \\ sy' &= y \\ z' &= f \\ s &= \frac{z}{f} \end{aligned} \tag{8.4.174}$$

Or better:

$$\begin{aligned} sx' &= x \\ sy' &= y \\ sz' &= z \\ s &= \frac{z}{f} \end{aligned} \tag{8.4.175}$$

We then use the following matrix for the conical projection:

$$\boxed{P_C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1/f & 0 \end{bmatrix}} \tag{8.4.176}$$

and then we normalize the coordinates by the ratio z/f. Indeed (a reader requested the details):

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \frac{1}{\frac{z}{f}} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1/f & 0 \end{bmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} = \frac{f}{z} \begin{pmatrix} x \\ y \\ z \\ z/f \end{pmatrix} = \begin{pmatrix} \frac{xf}{z} \\ \frac{yf}{z} \\ z \\ 1 \end{pmatrix} \tag{8.4.177}$$

~80 %

★★★★☆

35 votes, 82.29%

8.5. Analytical Geometry

THE "analytic geometry" is the branch of geometry that deals with the study of geometric shapes and their properties using the advanced tools of Calculus such as functional analysis, vector calculus and linear algebra. Its border lies in the tools used and originated by the work of René Descartes in the early 117th century (holocene calendar). The "vector geometry" is a subset of analytical geometry and we will also us it in this section as in this of Differential Geometry.

Remark

When we use for these same subjects that will follow the differential and integral calculus we then say that we do "differential geometry" (see section of the same name page 2850).

Analytic geometry is a very broad area (like everything else in this textbook) then... we discussed here that the elements indispensable for the study of physics (especially astronomy and quantum physics particle) and engineering (spacial engineering). These elements are also often studied in small classes and are (cited in order of study at high-school): the conics, the equations of the line, of the plan, of the sphere, etc. their intersections, their tangent planes and many others.

The reader will notice that we begin this section with conics and not straight lines and this is simply because we will use much more conics properties than straight lines in this textbook. For analytical geometric properties of straight lines see the subsection after the conics.

8.5.1 Conics

It has been very difficult for us to choose whether to put the study of the conics in the chapter on Algebra or on Geometry. We finally decided to put this study in this chapter (hence Geometry...) which suggests that the reader that has made a linear reading of this textbook has already covered all chapters and sections presenting the mathematical tools necessary to study of conics. We hope that our choice will be the best one for the reader.

The study of conic will be very useful in the section of Astronomy (we own to Kepler is many results from on the study on conics) and also in the sections of Geometrical Optics, Statistics and Industrial Engineering. It is therefore appropriate to go in the details of this subject.

8.5.1.1 Algebraic approach

Consider $(\text{O},\vec{i},\vec{j})$ an orthonormal plan. The simplest algebraic curves that we can find after the lines whose general form equations are (reminder):

$$ax+by+c=0 \tag{8.5.1}$$

are the curves of the second degree, named "analytical conics", that is to say by extension:

$$\boxed{\underbrace{\alpha x^2+\beta y^2+\gamma xy}_{\text{quadratic term}}+\delta x+\varepsilon y+\zeta=0} \tag{8.5.2}$$

with α,β,γ all non-zero.

The second degree corresponding curves are named "conics" (also named "quadrics" because of the presence of a quadratic term). The conic will be named "proper conic" if it is an ellipse, a parabola or a hyperbole as we shall see further below.

This last relation can also be written in matrix form (very important in the practice of Numerical Methods as we will see it in the section of the same name and in the classification of differential equations as discussed in the section of Differential and Integral Calculus):

$$[x \quad y]=\begin{bmatrix}\alpha & \dfrac{\beta}{2}\\ \dfrac{\beta}{2} & \gamma\end{bmatrix}\begin{bmatrix}x\\ y\end{bmatrix}+[\delta \quad \varepsilon]\begin{bmatrix}x\\ y\end{bmatrix}+\zeta=\vec{x}^TA\vec{x}+B^T\vec{x}+C \tag{8.5.3}$$

notation for which there exists several variants...

For example, the function:

$$f(x,y)=x^2+y^2+2bxy-5x-4y+g \tag{8.5.4}$$

gives for different values of g the following paths:

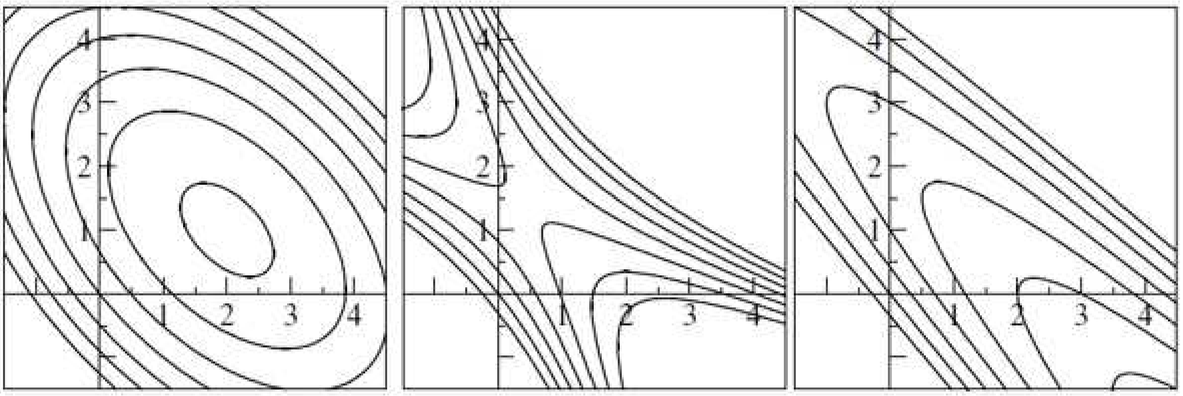


Figure 8.148: **Conics with respectively $b=0.5,1.5$ and 1 and more values of g**

Our first task will be to obtain, by translation and rotation of the coordinate system in which this relation is expressed, a reduced equation much simpler by eliminating the xy term. Indeed, let us choose a new referential being deduced by the previous one just by a rotation of angle θ. Let x' and y' be the coordinates of the new points. We have (see section Euclidean Geometric page 2648):

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \tag{8.5.5}$$

Thus:

$$\begin{aligned} x &= \cos(\theta)x' - \sin(\theta)y' \\ y &= \sin(\theta)x' + \cos(\theta)y' \end{aligned} \tag{8.5.6}$$

The equation becomes:

$$\alpha(\cos(\theta)x' - \sin(\theta)y')^2 + \beta(\sin(\theta)x' + \cos(\theta)y')^2 \tag{8.5.7}$$

$$+\gamma(\cos(\theta)x' - \sin(\theta)y')(\sin(\theta)x' + \cos(\theta)y') + \ldots = 0 \tag{8.5.8}$$

So we want all the terms $x'y'$ to be grouped as:

$$-2\alpha\cos(\theta)\sin(\theta) + 2\beta\cos(\theta)\sin(\theta) + \gamma(\cos^2(\theta) - \sin^2(\theta)) \tag{8.5.9}$$

Since (see section Trigonometry page 2519):

$$\cos^2(\theta) - \sin^2(\theta) = \cos(2\theta) \quad \text{and} \quad \sin(\alpha+\beta) = \sin(\beta)\cos(\alpha) + \cos(\beta)\sin(\alpha) \tag{8.5.10}$$

substituting, we get:

$$(\beta - \alpha)\sin(2\theta) + \gamma\cos(2\theta) = 0 \tag{8.5.11}$$

To have only the terms in $x'y'$ that simplifies, we just have to choose the rotation angle θ such as:

$$\frac{\sin(2\theta)}{\cos(2\theta)} = \frac{\gamma}{\alpha - \beta} = \tan(2\theta) \tag{8.5.12}$$

Therefore we will consider now the equation:

$$\boxed{\alpha x^2 + \beta y^2 + \delta x + \varepsilon y + \zeta = 0} \tag{8.5.13}$$

1. If we write that $\alpha = 0$ and $\beta = 0$. And we divide by β, we can take us back to an equation of the following type:

   $$y^2 + \gamma x + \delta y + \varepsilon = 0 \tag{8.5.14}$$

   Where:

   - If $\gamma \neq 0$, we end up with an equation describing the figure of a "parabola" of axis parallel to OX.
   - If $\gamma = 0$, it is a degenerate case (a degenerate conic is a conic that is reduced to be two lines that may or may not be parallel or a single line).

2. If we write $\alpha \neq 0$ and $\beta = 0$ thus the situation is treated exactly as before.

3. If $\alpha \neq 0$ and $\beta \neq 0$, we can remove the terms δx and εy terms as follows:

$$\alpha\left(x+\frac{\delta}{2\alpha}\right)^2+\beta\left(y+\frac{\varepsilon}{2\beta}\right)+\gamma=0 \tag{8.5.15}$$

And by a simple change of referential via translations, we arrive at an equation that looks like following:

$$\alpha X^2+\beta Y^2+\gamma=0 \tag{8.5.16}$$

- If $\gamma=0$ the previous relations simplifies to:

$$\alpha X^2=-\beta Y^2 \tag{8.5.17}$$

And therefore it is obvious that if $\text{sgn}(x)=\text{sgn}(y)$ there is only the trivial solution $X=Y=0$ if we stay in $\mathbb{R}$. Instead if $\text{sgn}(x)\neq\text{sgn}(y)$ thus we have an infinity of solutions represented by a straight line.

- If $\gamma \neq 0$ let us put that:

$$\alpha=\text{sgn}(\alpha)\frac{1}{\alpha'^2} \quad \text{and} \quad \beta=\text{sgn}(\beta)\frac{1}{\beta'^2} \tag{8.5.18}$$

and let us divide by $|\gamma|$ such that:

$$\text{sgn}(\alpha)\frac{X^2}{|\gamma|\alpha'^2}+\text{sgn}(\beta)\frac{Y^2}{|\gamma|\beta'^2}+\text{sgn}(\gamma)=0 \tag{8.5.19}$$

If we put now:

$$\frac{X^2}{|\gamma|\alpha'^2}=\frac{1}{a^2} \quad \text{and} \quad \frac{Y^2}{|\gamma|\beta'^2}=\frac{1}{b^2} \tag{8.5.20}$$

We get:

$$\text{sgn}(\alpha)\frac{X^2}{a^2}+\text{sgn}(\beta)\frac{Y^2}{b^2}=-\text{sgn}(\gamma) \tag{8.5.21}$$

So we therefore have several possible situations:

$$\boxed{\begin{array}{ll} +\dfrac{X^2}{a^2}+\dfrac{Y^2}{b^2}=+1 & -\dfrac{X^2}{a^2}-\dfrac{Y^2}{b^2}=-1 \\ +\dfrac{X^2}{a^2}-\dfrac{Y^2}{b^2}=+1 & -\dfrac{X^2}{a^2}+\dfrac{Y^2}{b^2}=-1 \\ -\dfrac{X^2}{a^2}+\dfrac{Y^2}{b^2}=+1 & +\dfrac{X^2}{a^2}-\dfrac{Y^2}{b^2}=-1 \\ \xcancel{-\dfrac{X^2}{a^2}-\dfrac{Y^2}{b^2}=+1} & \xcancel{+\dfrac{X^2}{a^2}+\dfrac{Y^2}{b^2}=-1} \end{array}} \tag{8.5.22}$$

Two terms above are impossible in $\mathbb{R}^2$, this is why we crossed them (the sum of two positive numbers can not be negative and vice versa).
There are several cases of interesting figures:

(a) For:

$$+\frac{X^2}{a^2}+\frac{Y^2}{b^2}=+1 \qquad -\frac{X^2}{a^2}-\frac{Y^2}{b^2}=-1 \tag{8.5.23}$$

and $a=b=1$, we have a unit circle. The reader can have fun testing this with the following commands in Maple™ 4.00b:

```
>with(plots):
>a:=1;b:=1;
>implicitplot(x^2/a^2+y^2/b^2=1,x=-10..10,y=-10..10);
```

(b) For:

$$+\frac{X^2}{a^2}+\frac{Y^2}{b^2}=+1 \qquad -\frac{X^2}{a^2}-\frac{Y^2}{b^2}=-1 \tag{8.5.24}$$

and $\forall a,b\in\mathbb{R}^*$ we have an ellipse (which therefore contains the specific case of the circle) with half-axes a and b (for the visual representation of an ellipse see further below or otherwise see the section of Geometric Forms).

The reader can have fun testing this with the following command in Maple™ 4.00b:

```
>with(plots):
>a:=4;b:=8;
>implicitplot(x^2/a^2+y^2/b^2=1,x=-10..10,y=-10..10);
```

(c) For:

$$\begin{array}{cc} \dfrac{X^2}{a^2}+\dfrac{Y^2}{b^2}=+1 & -\dfrac{X^2}{a^2}+\dfrac{Y^2}{b^2}=-1 \\ -\dfrac{X^2}{a^2}+\dfrac{Y^2}{b^2}=+1 & \dfrac{X^2}{a^2}-\dfrac{Y^2}{b^2}=-1 \end{array} \tag{8.5.25}$$

and $\forall a,b\in\mathbb{R}^*$, we get hyperbolas whose symmetry axis is parallel to OX or OY (for visual representation see further below). We say that the hyperbole is an "equilateral hyperbola" when $a=b$.

The reader can have fun testing this with the following commands in Maple™ 4.00b:

```
>with(plots):
>a:=4;b:=4;
>implicitplot(x^2/a^2-y^2/b^2=1,x=-10..10,y=-10..10);
```

The term "conical" comes from the fact that one of the first definitions of conicals consisted of the intersection of a cone and a plane.

Indeed, given:

$$z^2=x^2+y^2 \tag{8.5.26}$$

the equation of a cone having an angle at the top of $\pi/4$ (or an angle of $\pi/2$ relatively to its surface and the x-axis). This is simple the application of Pythagorean theorem...

Plus also given an oriented plane with an angle θ relatively to the Z axis:

$$x=f(y,z)=\cos(\theta)y+\sin(\theta)z:=h \tag{8.5.27}$$

where the reader can see that if $\theta=0$ we get $z=h$ for $\forall x,y$ the plane is therefore parallel to $x\text{O}y$, if $\theta=\pi/2$ we get $y=h$ for $\forall z,y$ the plane is parallel to $x\text{O}z$. In both situation and for every θ the plane never intersect with the axis X and we have a normal vector that is always in the plane $x\text{O}y$.

This normal vector has almost obviously the following coordinates (we use the cosine-director notation):

$$\vec{n}=(0,\cos(\theta),\sin(\theta))=\left(0,\cos(\theta),\cos\left(\frac{\pi}{2}-\theta\right)\right) \tag{8.5.28}$$

Consider now the rotation matrix in the space $\mathbb{R}^3$ relative to the axis Oz (see section Euclidean Geometry page 2648):

$$\begin{pmatrix}x'\\y'\\z'\end{pmatrix}=\begin{bmatrix}1&0&0\\0&\cos(\theta)&-\sin(\theta)\\0&\sin(\theta)&\cos(\theta)\end{bmatrix}\begin{pmatrix}x\\y\\z\end{pmatrix}=\begin{pmatrix}x\\y\cos(\theta)-z\sin(\theta)\\y\sin(\theta)+z\cos(\theta)\end{pmatrix} \tag{8.5.29}$$

So we have for expression of rotation of our plane:

$$\cos(\theta)y+\sin(\theta)z=\cos(\theta)(\cos(\theta)-z\sin(\theta))+\sin(\theta)(\cos(\theta)-z\sin(\theta)):=h \tag{8.5.30}$$

After simplification, $\forall\theta$:

$$y=h \tag{8.5.31}$$

This just mean that when we apply the rotation matrix to our plane with the same angle θ we take it to is "initial" position.

Identically, for the cone, a rotation according to the Z axis :

$$(y\sin(\theta)+z\cos(\theta))^2=x^2+(y\cos(\theta)-z\sin(\theta))^2 \tag{8.5.32}$$

After development and simplification:

$$x^2-(y^2-z^2)\cos(2\theta)+yz2\sin(2\theta)=0 \tag{8.5.33}$$

Equation that gives a horizontal cone for $\theta=0$ and a vertical one for $\theta=\frac{\pi}{2}$.

Thus, we have one possible general system:

$$\left\{\begin{array}{l}y=h\\x^2-(y^2-z^2)\cos(2\theta)+yz2\sin(2\theta)=0\end{array}\right. \tag{8.5.34}$$

We see then that for:

- $\theta\in\left]\dfrac{\pi}{4},\dfrac{\pi}{2}\right]$ we get an intersection between the plane and the cone giving an ellipse
- $\theta=\dfrac{\pi}{4}$ we get a parabola
- $\theta\in\left[0,\dfrac{\pi}{4}\right[$ we get a hyperbola

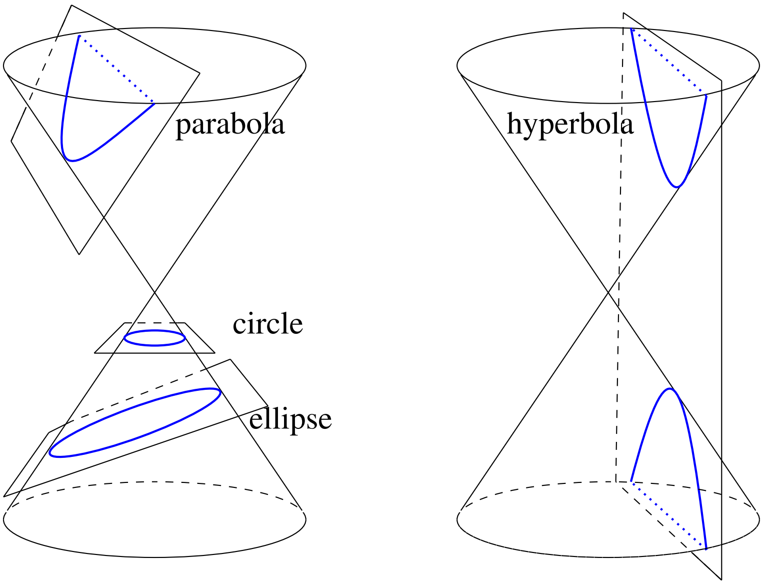


Figure 8.149: **Various intersections between a cone and a plane (author: Ridlo W. Wibowo)**

and possible real world neighbourhood noise impact analysis:

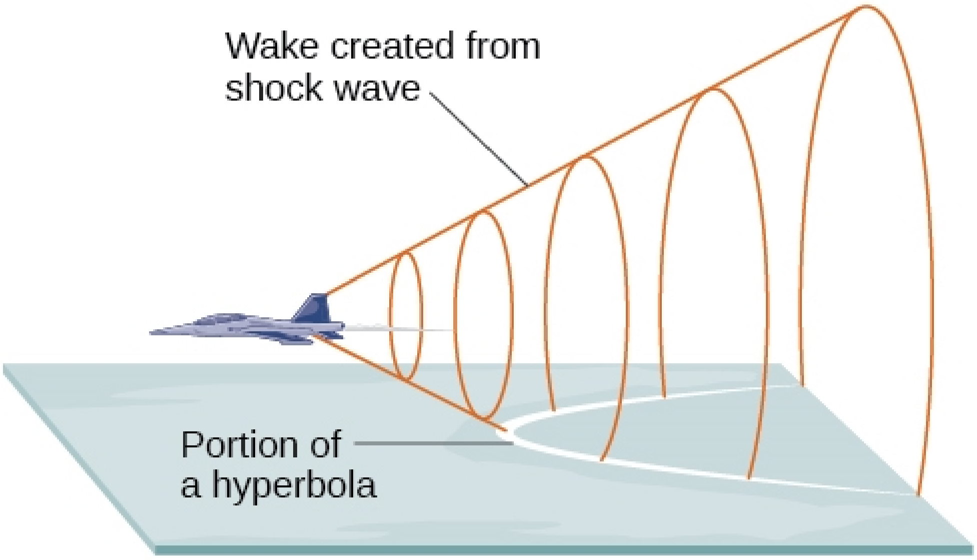

Wake created from shock wave
Portion of a hyperbola

Figure 8.150: **Shock wave intersecting the ground (source: OpenStax)**

The hyperbola of a sonic boom is described as a cone-shaped shock wave that intersects the ground in a hyperbolic curve. Every point on this curve is hit by the boom at the same time, allowing people in different locations to hear it simultaneously. This is because the hyperbolic curve moves forward as the airplane does, eventually allowing everyone in its path to hear the boom.

Remark

We also give the curve of equation $xy=1$ the name hyperbole because, with a variable change:

$$\begin{aligned} x &= X+Y \\ y &= X-Y \end{aligned} \tag{8.5.35}$$

Which brings us to:

$$X^2-Y^2=1 \tag{8.5.36}$$

which as we have seen, is the equation of a hyperbole.

For those who have Maple™ 4.00b or later here are some non-trivial commands to have fun to do a parabola:

```
>restart:  with(plots):
>c:=sqrt(x^2+y^2):
>p:=y+3:
>Y:=solve(c=p,y);
>intsect:=subs(y=Y,c);
>intsect := (x +(1/6*x-3/2))
>P1:=plot3d(c,x=-5..5,y=-5..5,axes=normal,color=red,numpoints=2000
,view=[-5..5,-5..5,0..5],style=wireframe):
>P2:=plot3d(p,x=-5..5,y=-5..5,axes=normal,color=yellow,
numpoints=2000,view=[-5..5,-5..5,0..5],style=patchnogrid):
>display(P1,P2,scaling=constrained, orientation=[-10,75]);
```

that gives:

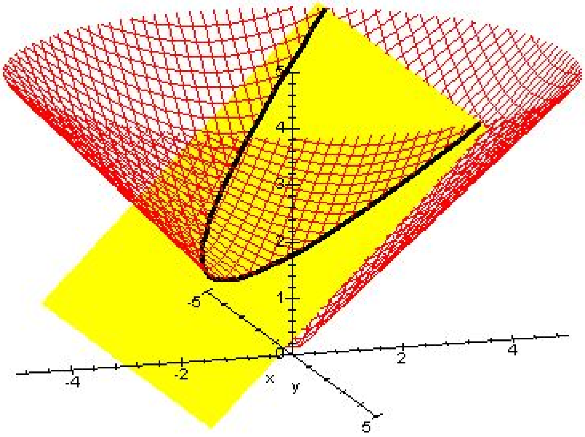


Figure 8.151: Maple™ 4.00b plot to obtain a parabola by the intersection of a cone and a plane

or to obtain a hyperbola:

```
>restart:  with(plots):
>c:=sqrt(x^2+y^2):
>p:=4*y+5:
>Y:=solve(c=p,y);
>intsect:=subs(y=Y[1],c);
>P1:=plot3d(c,x=-5..5,y=-5..5,axes=normal,color=red,numpoints=2000
,view=[-5..5,-5..5,0..5],style=wireframe):
>P2:=plot3d(p,x=-5..5,y=-5..5,axes=normal,color=yellow,numpoints=2000
,view=[-5..5,-5..5,0..5],style=patchnogrid):
>P3:=spacecurve([x,Y[1],intsect],x=-5..5,color=black,thickness=3):
```

```
P3:=spacecurve([x,Y[1],intsect],x=-5..5,color=black,thickness=3):
>display(P1,P2,P3,scaling=constrained);
```

that gives:

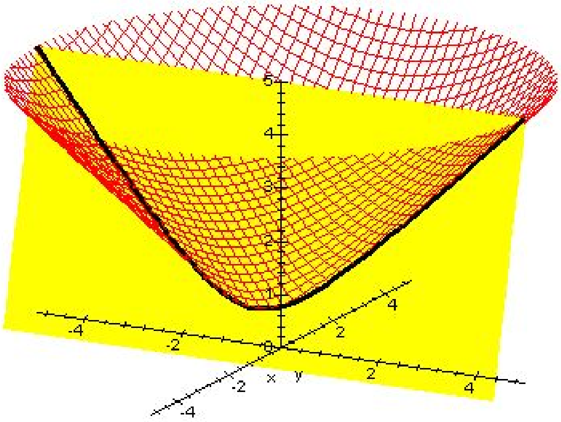


Figure 8.152: **Maple™ 4.00b plot to obtain a hyperparabola by the intersection of a cone and a plane**

or to obtain an ellipse:

```
>restart:  with(plots):
>c:=sqrt(x^2+y^2):
>p:=y/3+3:
>Y:=solve(c=p,y);
>E1:=subs(y=Y[1],c); E2:=subs(y=Y[2],c);
>P1:=plot3d(c,x=-5..5,y=-5..5,axes=normal,color=red,numpoints=2000,
view=[-5..5,-5..5,0..5],style=wireframe):
>P2:=plot3d(p,x=-5..5,y=-5..5,axes=normal,color=yellow,numpoints=2000,
view=[-5..5,-5..5,0..5],style=patchnogrid):
>P3:=spacecurve([x,Y[1],E1],[x,Y[2],E2],x=-5..5,color=black,thickness=3,
numpoints=2000):
>display(P1,P2,P3,scaling=constrained);
```

that gives:

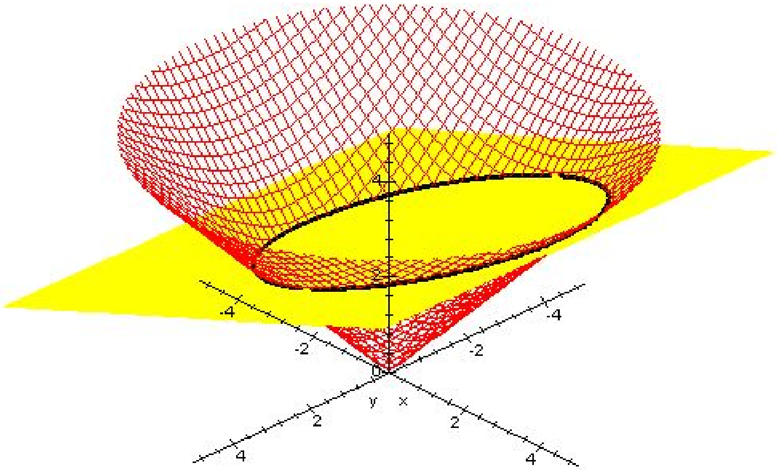


Figure 8.153: Maple™ 4.00b plot to obtain an ellipse by the intersection of a cone and a plane

or finally to obtain a circle:

```
>restart: with(plots):
>c:=sqrt(x^2+y^2):
>p:=3:
>Y:=solve(c=p,y);
>circ1:=subs(y=Y[1],c); circ2:=subs(y=Y[2],c);
>P1:=plot3d(c,x=-5..5,y=-5..5,axes=normal,color=red,numpoints=2000,
view=[-5..5,-5..5,0..5],style=wireframe):
>P2:=plot3d(p,x=-5..5,y=-5..5,axes=normal,color=yellow,numpoints=2000,
view=[-5..5,-5..5,0..5],style=patchnogrid):
>P3:=spacecurve([x,Y[1],circ1],[x,Y[2],circ2],
x=-5..5,color=black,thickness=3,numpoints=2000):
>display(P1,P2,P3);
```

that gives:

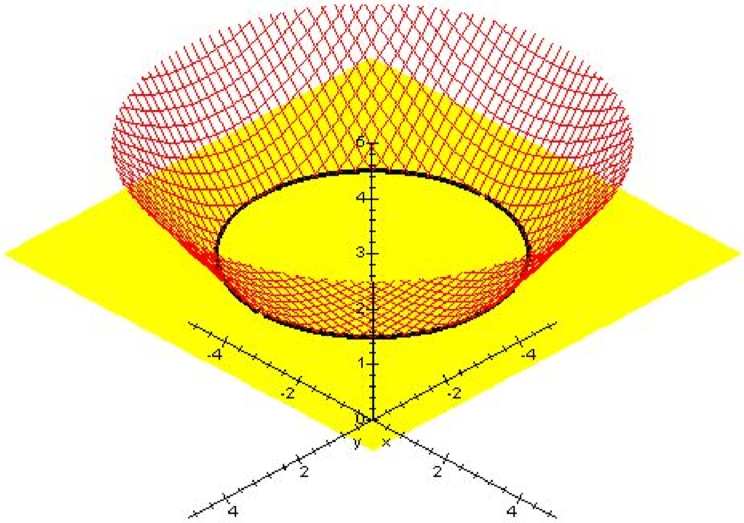


Figure 8.154: Maple™ 4.00b plot to obtain a circle by the intersection of a cone and a plane

However, the conicals also have a geometrical definition:

8.5.1.2 Geometric Approach

Let F be a point in the plane, D a line but not containing F (but to a non-zero distance of it) and e a positive real number. We are interested in the set of points M defined by:

$$\boxed{\{M|e\cdot d(M,D)=d(M,F)\}} \tag{8.5.37}$$

where F is named the "focus", D the "direction of the conical" and e the "eccentricity[10]":

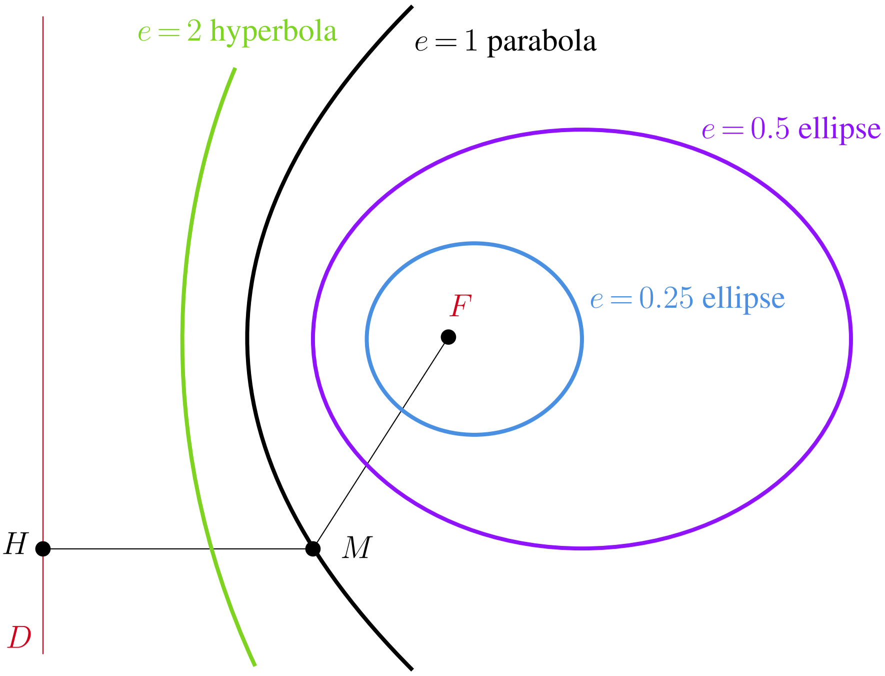


Figure 8.155: **Definition of the focus (F), direction (D) of the conical and eccentricity (e)**

We will arrange subsequently to have always F as the origin of the basis for the conical, therefore D will have by definition for equation:

$$x=-h \tag{8.5.38}$$

with $h>0$ so we therefore denote by $d(M,D)$ the distance between the point M and D.

Therefore:

$$\begin{aligned}&e\cdot d(M,D)=d(M,F)\\&\Leftrightarrow e|x+h|=\sqrt{x^2+y^2}\\&\Leftrightarrow e^2(x+h)^2=x^2+y^2\\&\Leftrightarrow (1-e^2)x^2+y^2-2e^2hx-e^2h^2=0\end{aligned} \tag{8.5.39}$$

We well fall back on the equation of a conic since the latter is a special case of:

$$\alpha x^2+\beta y^2+\delta x+\varepsilon y+\xi=0 \tag{8.5.40}$$

We can now consider several specific cases:

---

[10] The eccentricity of the Earth's orbit varies over time, due to gravitational interaction with other celestial bodies. The eccentricity value can get as low as 0.0034 and up to 0.058. Because of the eccentricity the orbit of the Earth from the farthest (aphelion) to nearest (perihelion) point of the Sun its distance has a differential of almost 3 millions kilometers actually.

1. Case where $e=1$:

   The equation:

   $$(1-e^2)x^2+y^2-2e^2hx-e^2h^2=0 \tag{8.5.41}$$

   then reduces trivially to:

   $$y^2-h^2=2hx \tag{8.5.42}$$

   It is therefore a parabola of axis orthogonal to D, whose top Ω is the mid-segment $\overline{FK}$, where K is the projection of F on D (see figure below).

   If we rewrite the last relation in the form:

   $$y^2=2hx+h^2=2h\left(x+\frac{h}{2}\right) \tag{8.5.43}$$

   and redefining the origin with respect to Ω by a translation of $h/2$, the generating focus of the parabola will be in $F=(h/2,0)$ and the latter equation is then reduced to:

   $$y^2=2hx \tag{8.5.44}$$

   where h is named "parameter of the parabola" and relatively to Ω, the focus will be given by the coordinates $F=(h/2,0)$ and the direction (directrix) by the equation $D:x=-h/2$. Written in a different way we fall back on the famous relation of the parabola in functional analysis (the reader can refer to a practical calculation example for an antenna in the section about Optics):

   $$x=\frac{1}{2h}y^2 \tag{8.5.45}$$

   That is:

   $$\boxed{x\propto y^2} \tag{8.5.46}$$

   As shown in the figure below, the distance to the direction to Ω is imposed by the conditions of the model:

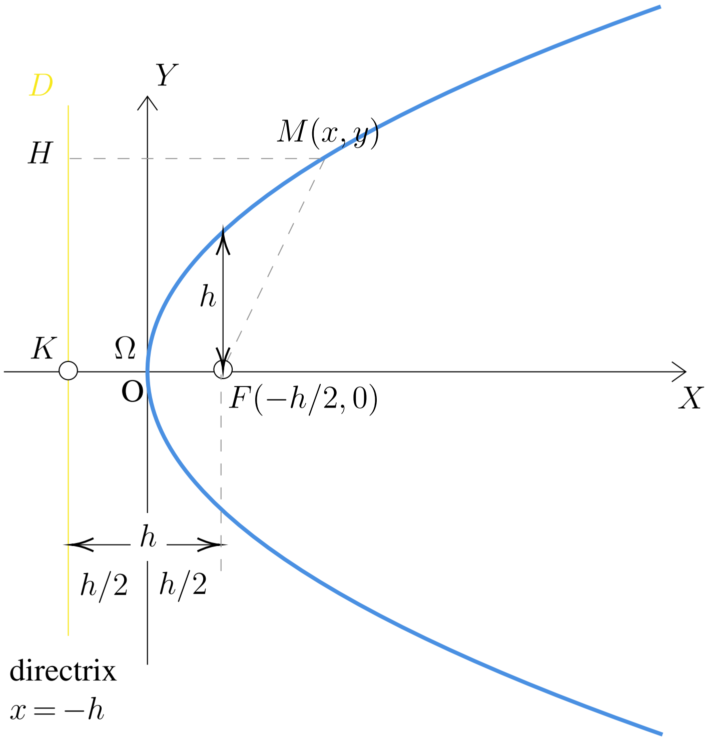


Figure 8.156: **Representation of the parameters of the conical for the parabola**

Based on the feedback of a reader let us introduce the parabola in another way that seems for some people more intuitive and more pedagogical (take from [309]).

We want to introduce the equation:

$$y = ax^2 + bx + c \tag{8.5.47}$$

which is a U-shaped curve named a "vertical parabola". We begin with the general definition of a parabola in the plane.

If we are given a line L and a point F not on L, the set of all points equidistant from L and F will form a U-shaped curve that passes midway between L and F and named a "parabola":

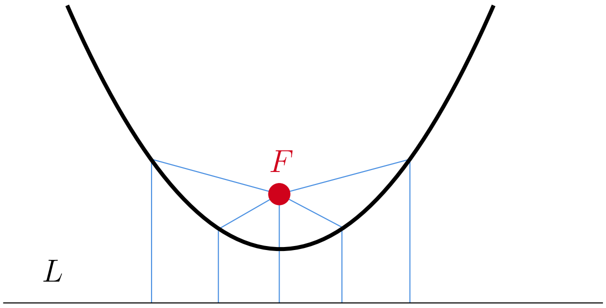


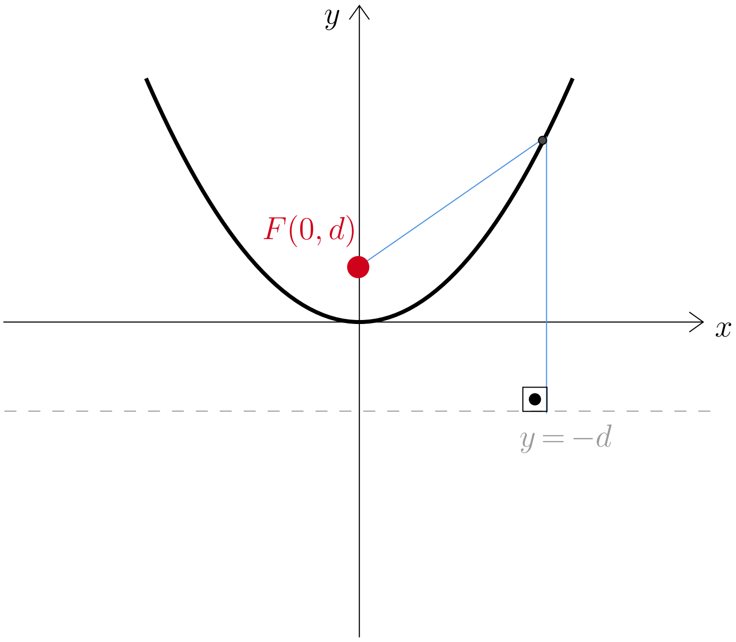
Figure 8.157: **Parabola = set of points equidistant from L and F**

Theorem 8.20. The graph of the equation:

$$y = ax^2 \tag{8.5.48}$$

(where $a \neq 0$) is the parabola with focus $F\left(0, \frac{1}{4}a\right)$ and directrix $y = -\left(\frac{1}{4}a\right)$. Its vertex is (0.0) and it axis is the y-axis.

Proof 8.20.1. Let us find the equation of the parabola with focus $F(0,d)$ and directrix $y=-d$, as shown below where obviously $d={}^1\!/\!_{4a}$:



Our plan is to show that the equation is $y=ax^2$ where $d={}^1\!/\!_{4a}$. Given a point $P(x,y)$, the perpendicular from P to the directrix is a vertical line of length $y+d$. Thus:

$$\begin{aligned}\text{distance from } P \text{ to directrix} &= y+d\\ \text{distance from } P \text{ to focus} &= \sqrt{x^2+(y-d)^2}\end{aligned} \tag{8.5.49}$$

The point P lies on the parabola exactly when these distances are equal:

$$y+d=\sqrt{x^2+(y-d)^2} \tag{8.5.50}$$

Simplifying we get:

$$\begin{aligned}(y+d)^2&=x^2+(y-d)^2\\ y^2+2yd+d^2&=x^2+y^2-2yd+d^2\end{aligned} \tag{8.5.51}$$

Therefore:

$$y=\frac{1}{4d}x^2=ax^2 \tag{8.5.52}$$

■ Q.E.D.

2. Case where $e<1$:

   It is an ellipse. Indeed it is not easy to see it but by rearranging the terms of equation:

$$(1-e^2)x^2+y^2-2e^2hx-e^2h^2=0 \tag{8.5.53}$$

we have (and this independently of whether $e<1$ or not):

$$(1-e^2)\left(x-\frac{e^2h}{1-e^2}\right)^2+y^2-\frac{e^2h^2}{1-e^2}=0\Leftrightarrow\frac{\left(x-\dfrac{e^2h}{1-e^2}\right)}{\dfrac{e^2h^2}{(1-e^2)^2}}+\frac{y^2}{\dfrac{e^2h^2}{1-e^2}}=1 \tag{8.5.54}$$

The last term of the ante-previous relation can recovered as following after development:

$$\frac{e^4h^2}{1-e^2}-\frac{e^2h^2}{1-e^2}=e^2h^2\left(\frac{e^2-1}{1-e^2}\right)=-e^2h^2 \tag{8.5.55}$$

Let us define:

$$F=\frac{e^2h}{1-e^2} \tag{8.5.56}$$

is the origin of the ellipse on X (then the x and y must now be taken relative to this new origin!!!). The previous equation can therefore be simplified and becomes (this always regardless of whether $e<1$ or not):

$$\boxed{\frac{x^2}{\dfrac{e^2h^2}{(1-e^2)^2}}+\frac{y^2}{\dfrac{e^2h^2}{1-e^2}}=\frac{x^2}{a^2}+\frac{y^2}{b^2}=1} \tag{8.5.57}$$

Therefore if $e<1$, the denominators are both positive and we fall back on the reduced equation of an ellipse since $a\neq b$.

Remark

You must notice that this definition can not include the circle among the ellipses otherwise there is a singularity because the denominators would be zero (the eccentricity being equal to zero for the circle).

To get the semi-major axis of the ellipse we just have to put $y=0$. Thus, we have:

$$\frac{x^2}{\dfrac{e^2h^2}{(1-e^2)^2}}=\frac{x^2}{a^2}=1 \tag{8.5.58}$$

then the semi-major axis is equal to:

$$a=\pm\frac{eh}{1-e^2}=\pm\frac{F}{e} \tag{8.5.59}$$

in the same way, we get for the semi-minor axis:

$$b=\pm\frac{eh}{\sqrt{1-e^2}} \tag{8.5.60}$$

putting $p=eh$ being the "parameter of the ellipse" or "focal parameter of the ellipse", we get:

$$\boxed{|a|=\frac{p}{1-e^2}\qquad |b|=\frac{p}{\sqrt{1-e^2}}} \tag{8.5.61}$$

whose first relation will be very useful in the section of Astronomy and also in the section of General Relativity.

Notice also by the way that we have then:

$$\frac{b^2}{a}=\frac{\dfrac{p^2}{(\sqrt{1-e^2})^2}}{\dfrac{p}{1-e^2}}=p \tag{8.5.62}$$

It is customary to denote the distance from the center Ω of the ellipse to the foci point F with the letter c such that:

$$c=\overline{\Omega F}=F \tag{8.5.63}$$

We then have:

$$c=F=\pm ae\Leftrightarrow e=\pm\frac{c}{a} \tag{8.5.64}$$

There are therefore two foci to the ellipse at a distance equivalent but opposite of the center Ω. We define therefore the eccentricity of an ellipse by the ratio:

$$e=\frac{c}{a}=\frac{\overline{\Omega F}}{a} \tag{8.5.65}$$

The reader will perhaps also noticed that we for $x=0$ in:

$$(1-e^2)x^2+y^2-2e^2hx-e^2h^2=0 \tag{8.5.66}$$

the ordinate that is therefore given by the product eh and it can be expressed by the ellipse parameters under the form (even if this may seem odd at units level this is implicitly correct):

$$p=eh=b \tag{8.5.67}$$

We can then also prove a relation that is very common in formula compilation hanbooks about ellipses:

$$a^2=b^2+c^2 \tag{8.5.68}$$

That is to say:

$$\frac{e^2h^2}{(1-e^2)^2}=\frac{e^2h^2}{1-e^2}+e^2a^2=\frac{e^2h^2}{1-e^2}+e^2\frac{e^2h^2}{(1-e^2)^2}=\frac{e^2h^2(1-e^2)+e^4h^2}{(1-e^2)^2}=\frac{e^2h^2}{(1-e^2)^2} \tag{8.5.69}$$

This property is then proved and that brings us to be able to write the eccentricity only thanks to the conventional parameters of the ellipse:

$$\boxed{e=\frac{c}{a}=\frac{\sqrt{a^2-b^2}}{a}} \tag{8.5.70}$$

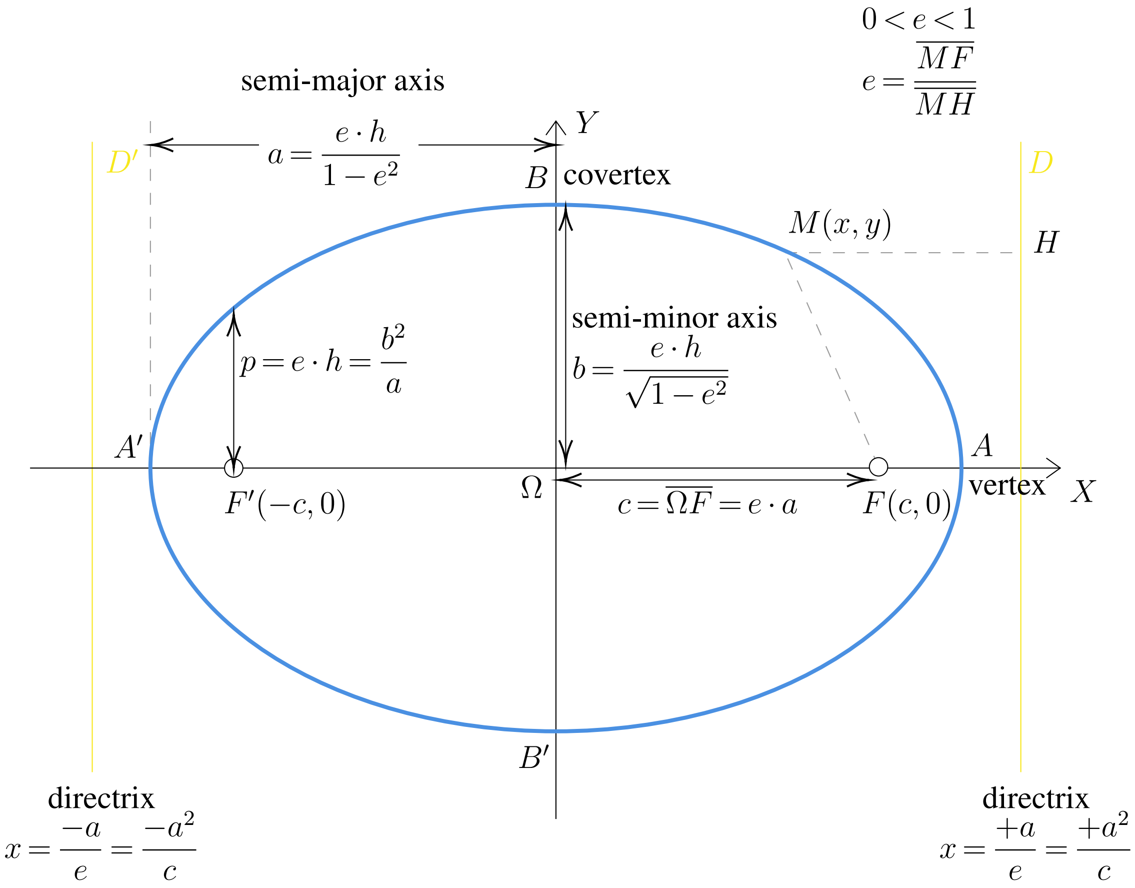

$0<e<1$
$e=\frac{\overline{MF}}{\overline{MH}}$
semi-major axis
$a=\frac{e\cdot h}{1-e^2}$
Y
D'
D
B covertex
$M(x,y)$
H
semi-minor axis
$b=\frac{e\cdot h}{\sqrt{1-e^2}}$
$p=e\cdot h=\frac{b^2}{a}$
A'
A
$F'(-c,0)$
Ω
$c=\overline{\Omega F}=e\cdot a$
$F(c,0)$
vertex
X
B'
directrix
$x=\frac{-a}{e}=\frac{-a^2}{c}$
directrix
$x=\frac{+a}{e}=\frac{+a^2}{c}$

Figure 8.158: Representation of the parameters of the conical for the ellipse

The distance from the center of the ellipse to each focus is $c=ae$. The distance from the center to a vertex is a. Therefore, the distance from a focus to the closer vertex is the difference between these distances, $a-c$, which can be expressed in terms of a and e as:

$$f=a-c=a-ae=a(1-e) \tag{8.5.71}$$

The distance from a focus to the farther vertex is $a+c$ or $a(1+e)$.

Here are some well known elliptic eccentricities (the reader must keep in mind that these are not constants!):

Celestial object orbit	Orbit eccentricity e
Mercury	0.206
Venus	0.007
Earth	0.017
Mars	0.093
Jupiter	0.048
Saturn	0.056
Uranus	0.047
Neptun	0.09
Pluton	0.250
Hubble Telescope	0.0002842
International Space Station	0.0005309
Halley's comet	0.970

Table 8.5: Some elliptic eccentricity values

We can also find out where are the director lines D of the ellipse with respect to the edge of this latter by using the definition of eccentricity when $y=0$. Then we have:

$$\begin{aligned} &e\cdot d(M,D)=d(M,F)\\ &\Leftrightarrow d(M,D)=\frac{d(M,F)}{e}=\frac{a-ea}{e}=a\frac{1-e}{e} \end{aligned} \tag{8.5.72}$$

therefore the factor of a is between 0 and $+\infty$ and this has for meaning that the director lines will be situated (at least in the special case presented here where the eccentricity is positive and strictly less than the unit) always outside of the ellipse border and in the nearest case tangential to this same border (but in all cases they will be outside of the ellipse).

A useful and obvious parametric representation of the ellipse is an extension of the trigonometric circle (see section Trigonometry page 2511):

$$\begin{aligned} x&=a\cos(t)\\ y&=b\sin(t) \end{aligned} \tag{8.5.73}$$

Indeed, if we consider the Cartesian equation of the ellipse proved above:

$$\frac{X^2}{a^2}+\frac{Y^2}{b^2}=1 \tag{8.5.74}$$

and putting $R=X/a$ and $S=Y/b$ we get:

$$R^2+S^2=1 \tag{8.5.75}$$

If we remember the unit circle in trigonometry, this equation admits as solutions $R=\cos(t)$ and $S=\sin(t)$. It comes then:

$$\begin{aligned} R&=\frac{X}{a}=\cos(t)\Rightarrow X=a\cos(t)\\ S&=\frac{Y}{b}=\sin(t)\Rightarrow Y=b\sin(t) \end{aligned} \tag{8.5.76}$$

That's it! And obviously if $a=b\neq 0$ we fall back on the parametric equation of a circle!

Now let us prove an important property of ellipses: The sum of the distance of the two foci to a point M on the border of the ellipse is constant and equal to the major axis.

Proof 8.20.2. We consider the center is the midpoint of the foci of coordinate $\Omega(0,0)$. The foci lie on the major axis $y=0$ and same for the equation of the minor axis for which $x=0$.

Now, if the length of the major, minor axes are $2a, 2b$ respectively with eccentricity e the equation of the ellipse is:

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1 \tag{8.5.77}$$

Any point M on the ellipse can be described as we have just seen before by:

$$M(a\cos(\theta),b\sin(\theta)) \tag{8.5.78}$$

So, the distance between a point $M(a\cos(\theta), b\sin(\theta))$ and one of the foci of coordinate $(ae, 0)$ is :

$$\begin{aligned}\overline{MF} &= \sqrt{(ae-a\cos(\theta)^2+(b\sin(\theta)-0)^2} = \sqrt{a^2e^2+a^2\cos^2(\theta)-2a^2e\cos(\theta)+b^2\sin^2(\theta)}\\ &= \sqrt{a^2e^2+a^2\cos^2(\theta)-2a^2e\cos(\theta)+b^2(1-\cos^2(\theta))}\\ &= \sqrt{a^2e^2+a^2\cos^2(\theta)-2a^2e\cos(\theta)+a^2(1-e^2)(1-\cos^2(\theta))}\\ &= \sqrt{a^2e^2+a^2\cos^2(\theta)-2a^2e\cos(\theta)+(a^2-a^2e^2)(1-\cos^2(\theta))}\\ &= \sqrt{a^2e^2+a^2\cos^2(\theta)-2a^2e\cos(\theta)+a^2-a^2e^2-a^2\cos^2(\theta)+a^2e^2\cos^2(\theta)}\\ &= \sqrt{-2a^2e\cos(\theta)+a^2+a^2e^2\cos^2(\theta)}\\ &= a\sqrt{-2e\cos(\theta)+1+e^2\cos^2(\theta)}\\ &= a\sqrt{(1-e\cos(\theta))^2} = a(1-e\cos(\theta))\end{aligned} \tag{8.5.79}$$

Similarly (as always you can request the details if necessary), the distance between $(a\cos(\theta), b\sin(\theta)), (-ae, 0)$ is equal to $a(1+e\cos(\theta))$.

Therefore the sum gives:

$$a(1-e\cos(\theta))+a(1+e\cos(\theta)) = 2a \tag{8.5.80}$$

■ Q.E.D.

It is this result that authorizes us to say that an ellipse can be constructed in the famous following way:

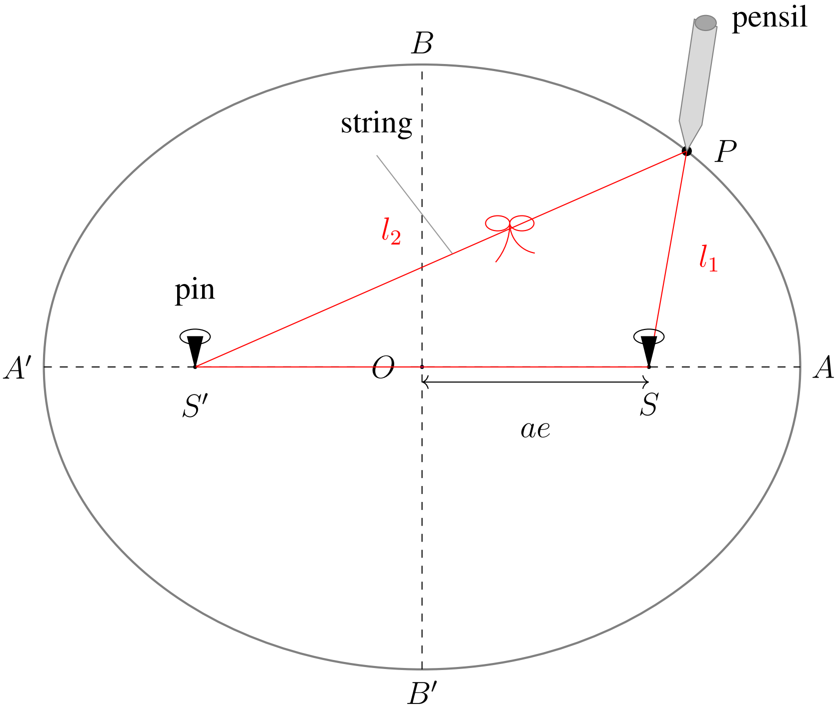


Figure 8.159: **Hand-made construction of an ellipse (author: Ridlo W. Wibowo)**

It is interesting also to see the opposite reasoning! We suppose the construction above as a definition and we build from it the equation of the ellipse centered at the origin!

For this we begin with the foci $(-c,0)$ and $(c,0)$. The ellipse is then by definition the set of all points (x,y) such that the sum of the distances from (x,y) to the foci is constant, as shown in the figure above.

If $(a,0)$ is a vertex of the ellipse, the distance from $(-c,0)$ to $(a,0)$ is:

$$a-(-c)=a+c \tag{8.5.81}$$

The distance from $(c,0)$ to $(a,0)$ is $a-c$. The sum of the distances from the foci to the vertex is:

$$(a+c)+(a-c)=2a \tag{8.5.82}$$

If (x,y) is a point on the ellipse, then we can define the following variables:

$$\begin{aligned}d_1 &= \text{the distance from } (-c,0) \text{ to } (x,y)\\ d_2 &= \text{the distance from } (c,0) \text{ to } (x,y)\end{aligned} \tag{8.5.83}$$

By the definition of an ellipse, the sum of distances d_1, d_2 of the foci to any point (x,y) of the ellipse is equal to and is constant! We know that the sum of these distances is $2a$ for the vertex $(a,0)$. It follows that:

$$d_1+d_2=2a \tag{8.5.84}$$

for any point on the ellipse. We will begin the derivation by applying the distance formula. The rest of the derivation is algebraic:

$$\begin{aligned}
& d_1+d_2=\sqrt{(x-(-c))^2+(y-0)^2}+\sqrt{(x-c)^2+(y-0)^2}=2a && \text{Distance formula}\\
& \sqrt{(x+c)^2+y^2}+\sqrt{(x-c)^2+y^2}=2a && \text{Simplify expressions}\\
& \sqrt{(x+c)^2+y^2}=2a-\sqrt{(x-c)^2+y^2} && \text{Move radical to opposite side}\\
& (x+c)^2+y^2=\left[2a-\sqrt{(x-c)^2+y^2}\right]^2 && \text{Square both sides}\\
& x^2+2cx+c^2+y^2=4a^2-4a\sqrt{(x-c)^2+y^2}+(x-c)^2+y^2 && \text{Expand the squares}\\
& x^2+2cx+c^2+y^2=4a^2-4a\sqrt{(x-c)^2+y^2}+x^2-2cx+c^2+y^2 && \text{Expand remaining squares}\\
& 2cx=4a^2-4a\sqrt{(x-c)^2+y^2}-2cx && \text{Combine like terms}\\
& 4cx-4a^2=-4a\sqrt{(x-c)^2+y^2} && \text{Isolate the radical}\\
& cx-a^2=-a\sqrt{(x-c)^2+y^2} && \text{Divide by 4}\\
& [cx-a^2]^2=a^2\left[\sqrt{(x-c)^2+y^2}\right]^2 && \text{Square both sides}\\
& c^2x^2-2a^2cx+a^4=a^2\left(x^2-2cx+c^2+y^2\right) && \text{Expand the squares}\\
& c^2x^2-2a^2cx+a^4=a^2x^2-2a^2cx+a^2c^2+a^2y^2 && \text{Distribute } a^2\\
& a^2x^2-c^2x^2+a^2y^2=a^4-a^2c^2 && \text{Rewrite}\\
& x^2\left(a^2-c^2\right)+a^2y^2=a^2\left(a^2-c^2\right) && \text{Factor common terms}\\
& x^2b^2+a^2y^2=a^2b^2 && \text{Set } b^2=a^2-c^2\\
& \tfrac{x^2b^2}{a^2b^2}+\tfrac{a^2y^2}{a^2b^2}=\tfrac{a^2b^2}{a^2b^2} && \text{Divide both sides by } a^2b^2\\
& \tfrac{x^2}{a^2}+\tfrac{y^2}{b^2}=1 && \text{Simplify}
\end{aligned} \tag{8.5.85}$$

Thus, the standard equation of an ellipse is:

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1 \tag{8.5.86}$$

The equation defines an ellipse centered at the origin, with for recall:

- If $a > b$, the ellipse is stretched further in the horizontal direction
- If $b > a$, the ellipse is stretched further in the vertical direction
- The length of the major axis is $2a$
- The coordinates of the vertices are $(0, \pm a)$
- The length of the minor axis is $2b$
- The coordinates of the co-vertices are $(\pm b, 0)$
- The coordinates of the foci are $(0, \pm c)$
- We have the identity $c^2 = a^2 - b^2$

The standard form of the equation of an ellipse with center (h, k) is then obviously:

$$\frac{(x-h)^2}{a^2} + \frac{(y-k)^2}{b^2} = 1 \tag{8.5.87}$$

However, there is another form of the equation of the ellipse, much more important, which is found in the sections of mechanics, astrophysics and quantum physics of this textbook.

First let us recall that:

$$e \cdot d(M, D) = d(M, F) \Leftrightarrow e|x+h| = \sqrt{x^2 + y^2} \tag{8.5.88}$$

In polar coordinates, this gives:

$$e \cdot d(M, D) = d(M, F) \Leftrightarrow e|r\cos(\theta) + h| = r \tag{8.5.89}$$

Therefore:

$$r = e(r\cos(\theta) - h) \quad \text{or} \quad r = -e(r\cos(\theta) - h) \tag{8.5.90}$$

after identification:

$$r_1 = \frac{-eh}{1 - e\cos(\theta)} \qquad r_2 = \frac{eh}{1 + e\cos(\theta)} \tag{8.5.91}$$

We obtain two different equations, but it is actually the same curve that describes the radius of the ellipse from one of its two foci. And we can see that:

$$r_1(\theta + \pi) = -r_2(\theta) \tag{8.5.92}$$

Since the $p = eh$ is defined as the parameter of the conical, the polar equation of the ellipse is given by:

$$r = \frac{p}{1 + e\cos(\theta)} \tag{8.5.93}$$

where θ is named the "true anomaly".

Notice the three special values:

$$\begin{aligned} &\theta = \frac{\pi}{2} \Rightarrow r = p = eh : \text{focal parameter} \\ &\theta = 0 \Rightarrow r_p = \frac{p}{1+e} : \text{pericenter (perihelion)} \\ &\theta = \pi \Rightarrow r_a = \frac{p}{1-e} : \text{apocenter (aphelion)} \end{aligned} \tag{8.5.94}$$

and we get obviously:

$$a=\frac{r_a+r_p}{2} \qquad \text{and} \qquad c=\frac{r_a-r_p}{2} \tag{8.5.95}$$

and since $e=\pm c/a$ we get also another famous relation for eccentricity:

$$e=\pm\frac{\dfrac{r_a-r_p}{2}}{\dfrac{r_a+r_p}{2}}=\pm\frac{r_a-r_p}{r_a+r_p} \tag{8.5.96}$$

The pericenter is better known as the "perigee" in astronomy as well as the apocenter which is better known as the "apogee" and are used a lot in astronomy and astrodynamics.

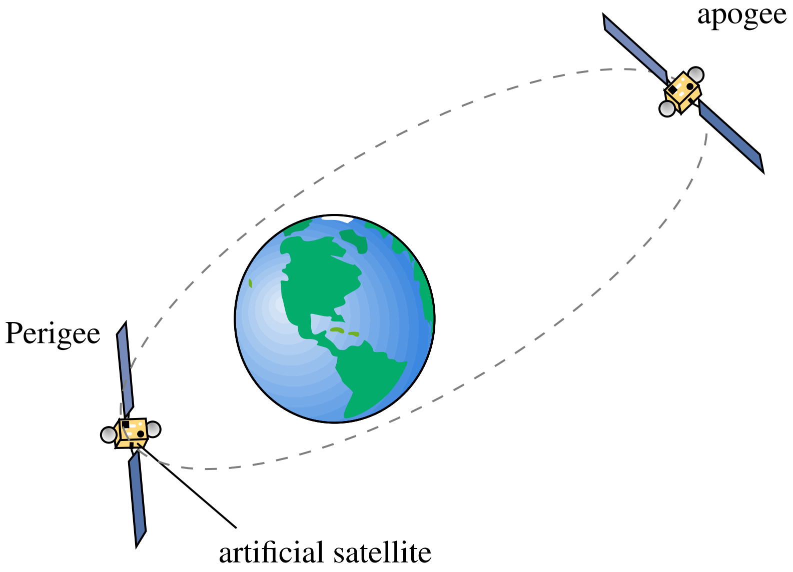


Figure 8.160: **Apogee and Perigee**

Remark

In facts, "apogee" and "perigee" are reserved words when the Earth is on the foci. If it is the Sun (actually a G2V-type star, aka yellow dwarf) that is the foci we then speak of "aphelion" (corresponding approximately to the beginning of Winter) and "perihelion" (corresponding approximately to the beginning of Summer).

In the general case, D may do any angle with the axis of polar angles, and the general equation is then (very important relation in astronomy and aerospace engineering!):

$$\boxed{r=\frac{p}{1+e\cos(\theta-\theta_0)}=\frac{a(1-e^2)}{1+e\cos(\theta-\theta_0)}=\frac{a\left(1-\left(\dfrac{c}{a}\right)^2\right)}{1+\left(\dfrac{c}{a}\right)\cos(\theta-\theta_0)}} \tag{8.5.97}$$

Let us prove now the interesting result that a special kind of parabola is an ellipse, but with one focal point at infinity!

Let us choose the equation for an ellipse with a focus at $(0,0)$ and the other at $(0,2ae)$ keeping $a(1-e)=f$ (where for reminder f is distance from the vertex to the focus of the ellipse, which ends up being the focal length of the parabola) such that:

$$\frac{x^2}{a^2\left(1-e^2\right)}+\frac{(y-ae)^2}{a^2}=1 \tag{8.5.98}$$

which is equivalent to:

$$\frac{x^2}{f(1+e)}+\frac{y^2-2aey}{a}=f(1+e) \tag{8.5.99}$$

If we let $a\rightarrow+\infty$ (and therefore $e=1-\frac{f}{a}\rightarrow 1$), we get:

$$\frac{x^2}{2f}-2y=2f \tag{8.5.100}$$

Hence:

$$y=\frac{x^2}{4f}-f \tag{8.5.101}$$

which is a parabola! This results explains why in ballistics when we take into account Earth's curvature (see section Classical Mechanics page 3278) the trajectory in the low speed regime is almost a parabola because Earth's radius is huge!

3. Case where $e>1$:

   This is a hyperbola (same reasoning as the ellipse considering by starting with any other value than $e>1$ or not):

$$\left(1-e^2\right)\left(x-\frac{e^2h}{1-e^2}\right)^2+y^2-\frac{e^2h^2}{1-e^2}=0\Leftrightarrow\frac{\left(x-\dfrac{e^2h}{1-e^2}\right)}{\dfrac{e^2h^2}{(1-e^2)^2}}+\frac{y^2}{\dfrac{e^2h^2}{1-e^2}}=1 \tag{8.5.102}$$

   Where we put again:

$$F=\frac{e^2h}{1-e^2} \tag{8.5.103}$$

   is the origin of hyperbole. The above equation is simplified and becomes (so far, we end up with exactly the same expression as for the ellipse):

$$\frac{x^2}{\dfrac{e^2h^2}{(1-e^2)^2}}+\frac{y^2}{\dfrac{e^2h^2}{1-e^2}}=\frac{x^2}{a^2}+\frac{y^2}{b^2}=1 \tag{8.5.104}$$

   But as $e>1$ the denominator of the second term will be negative and therefore we fall back on the reduced equation of a hyperbole!

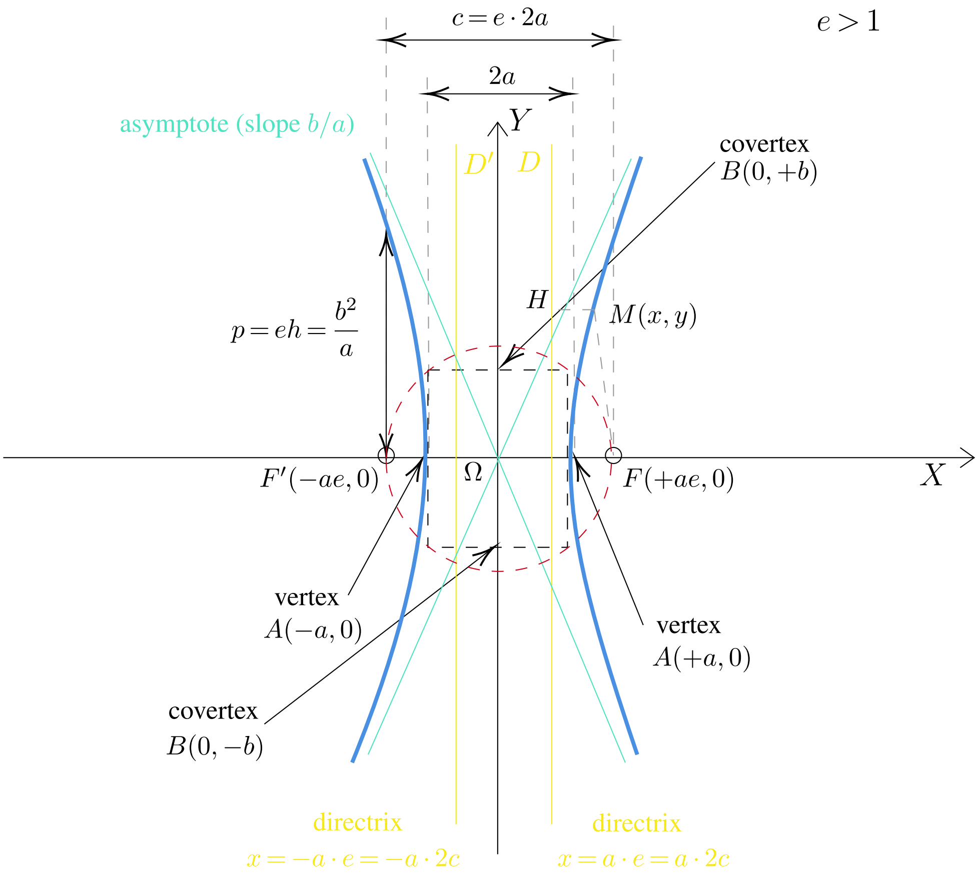
   We have therefore for semi-major axis and semi-minor axis (same reasoning as for the ellipse):

$$|a|=\frac{p}{1-e^2}\qquad |b|=\frac{p}{\sqrt{1-e^2}} \tag{8.5.105}$$

and:

$$e=\frac{c}{a}=\frac{\sqrt{a^2+b^2}}{a} \tag{8.5.106}$$

with the corresponding figure:



$c=e\cdot 2a$
$e>1$
$2a$
asymptote (slope b/a)
Y
D'
D
covertex
$B(0,+b)$
H
$M(x,y)$
$p=eh=\frac{b^2}{a}$
Ω
$F'(-ae,0)$
$F(+ae,0)$
X
vertex
$A(-a,0)$
vertex
$A(+a,0)$
covertex
$B(0,-b)$
directrix
$x=-a\cdot e=-a\cdot 2c$
directrix
$x=a\cdot e=a\cdot 2c$

Figure 8.161: **Representation of the parameters of the conical for the hyperbola**

where we presented the two asymptotes using a simple technique of passing to the limit:

$$\frac{x^2}{a^2}-\frac{y^2}{b^2}=1\Leftrightarrow y=\sqrt{\frac{b^2}{a^2}x^2-b^2}\underset{x\to+\infty}{\cong}\sqrt{\frac{b^2}{a^2}x^2}=\pm\frac{b}{a}x=\triangle \tag{8.5.107}$$

If the hyperbole is equilateral, then it is obvious that the two asymptotes are perpendiculars since their slope is then respectively $+1$ and -1. In the equilateral case, we also have the eccentricity which is immediately given by:

$$e=\sqrt{\frac{a^2+b^2}{a}}=\frac{\sqrt{a^2+a^2}}{a}=\sqrt{2} \tag{8.5.108}$$

We can also determine where are the director lines D of the hyperbolas with respect to their border by using the definition of the eccentricity when $y=0$ by performing exactly in the same as for the ellipse. Then we have:

$$ed(M,D)=d(M,F)\quad\Leftrightarrow d(M,D)=\frac{d(M,F)}{e}=\frac{ea-e}{e}=a\frac{e-1}{e} \tag{8.5.109}$$

therefore the factor of a is between 0 and tend to 1 as e is very large meaning that the director lines will be located (at least in the case here where this eccentricity is positive and strictly smaller than unity) is tangential to the hyperbola that is to say the closest to the y-axis (but in all cases they will therefore be between the two hyperbolas).

There is another nice way to build the equation of the hyperbola by assuming that he hyperbola is the set of all points (x,y) such that the difference of the distances d_1, d_2 from (x,y) to the foci is constant (so keep in mind that the ellipse is defined by the sum of distance, whereas the hyperboloid is defined by the difference).

For this proof let us consider $(-c,0)$ and $(c,0)$ be the foci of a hyperbola centered at the origin.

If $(a,0)$ is a vertex of the hyperbola, the distance from $(-c,0)$ to $(a,0)$ is:

$$a-(-c)=a+c \tag{8.5.110}$$

The distance from $(c,0)$ to $(a,0)$ is $c-a$. The sum of the distances from the foci to the vertex is:

$$(a+c)-(c-a)=2a \tag{8.5.111}$$

If (x,y) is a point on the hyperbola, we can define the following variables:

$$\begin{aligned}d_2&=\text{the distance from } (-c,0) \text{ to } (x,y)\\ d_1&=\text{the distance from } (c,0) \text{ to } (x,y)\end{aligned} \tag{8.5.112}$$

By assuming that for a hyperbola, d_2-d_1 is constant for any point (x,y) on the hyperbola it follows for the vertex $(a,0)$ that:

$$d_2-d_1=2a \tag{8.5.113}$$

and therefore the same value should be consistent for the whole hyperbola. As with the derivation of the equation of an ellipse seen just previously, we will begin by applying the distance formula. The rest of the derivation is algebraic. Compare this derivation with the one from the previous section for ellipses:

$$\begin{aligned}
& d_2-d_1=\sqrt{(x-(-c))^2+(y-0)^2}-\sqrt{(x-c)^2+(y-0)^2}=2a && \text{Distance Formula}\\
& \sqrt{(x+c)^2+y^2}-\sqrt{(x-c)^2+y^2}=2a && \text{Simplify expressions}\\
& \sqrt{(x+c)^2+y^2}=2a+\sqrt{(x-c)^2+y^2} && \text{Move radical to opposite side}\\
& (x+c)^2+y^2=\left(2a+\sqrt{(x-c)^2+y^2}\right)^2 && \text{Square both sides}\\
& x^2+2cx+c^2+y^2=4a^2+4a\sqrt{(x-c)^2+y^2}+(x-c)^2+y^2 && \text{Expand the squares}\\
& x^2+2cx+c^2+y^2=4a^2+4a\sqrt{(x-c)^2+y^2}+x^2-2cx+c^2+y^2 && \text{Expand remaining square}\\
& 2cx=4a^2+4a\sqrt{(x-c)^2+y^2}-2cx && \text{Combine like terms}\\
& 4cx-4a^2=4a\sqrt{(x-c)^2+y^2} && \text{Isolate the radical}\\
& cx-a^2=a\sqrt{(x-c)^2+y^2} && \text{Divide by 4}\\
& (cx-a^2)^2=a^2\left[\sqrt{(x-c)^2+y^2}\right]^2 && \text{Square both sides}\\
& c^2x^2-2a^2cx+a^4=a^2\left(x^2-2cx+c^2+y^2\right) && \text{Expand the squares}\\
& c^2x^2-2a^2cx+a^4=a^2x^2-2a^2cx+a^2c^2+a^2y^2 && \text{Distribute } a^2\\
& a^4+c^2x^2=a^2x^2+a^2c^2+a^2y^2 && \text{Combine like terms}\\
& c^2x^2-a^2x^2-a^2y^2=a^2c^2-a^4 && \text{Rearrange terms}\\
& x^2\left(c^2-a^2\right)-a^2y^2=a^2\left(c^2-a^2\right) && \text{Factor common terms}\\
& x^2b^2-a^2y^2=a^2b^2 && \text{Set } b^2=c^2-a^2.\\
& \tfrac{x^2b^2}{a^2b^2}-\tfrac{a^2y^2}{a^2b^2}=\tfrac{a^2b^2}{a^2b^2} && \text{Divide both sides by } a^2b^2\\
& \tfrac{x^2}{a^2}-\tfrac{y^2}{b^2}=1
\end{aligned} \tag{8.5.114}$$

This equation defines a hyperbola centered at the origin with vertices $(\pm a,0)$ and co-vertices $(0,\pm b)$.

Therefore the standard form of the equation of a hyperbola with center $(0,0)$ and transverse axis on the x-axis is:

$$\frac{x^2}{a^2}-\frac{y^2}{b^2}=1 \tag{8.5.115}$$

where:

- The length of the transverse axis is $2a$
- The coordinates of the vertices are $(\pm a,0)$
- The length of the conjugate axis is $2b$
- The coordinates of the co-vertices are $(0,\pm b)$
- The distances between the foci is $2c$
- We have the identity $c^2=a^2+b^2$
- The coordinates of the foci are $(\pm c,0)$

- The equations of the asymptotes are $y=\pm{}^{b}/_{a}$

And for the standard form equation of a hyperbola with center (h,k) and transverse axis parallel to the x-axis we have:

$$\frac{(x-h)^2}{a^2}-\frac{(y-k)^2}{b^2}=1 \tag{8.5.116}$$

Remember now that in the figure above we have represented the hyperbola:

$$\frac{x^2}{a^2}-\frac{y^2}{b^2}=1 \tag{8.5.117}$$

for some $0<a<b$. The asymptotes $y/b=\pm x/a$ have also been shown, and it is easy to see algebraically that the asymptotes indeed must be these equations.

The rectangle visible in the figure between the both hyperbolas is drawn such that the horizontal width is the distance $2a$, thus opposite sides are tangent to the vertices of the hyperbola at $(\pm a,0)$. The vertical height of the rectangle is chosen such that the rectangle's vertices are on the asymptotes, thus the height is $2b$ and the rectangle's vertices are $(\pm a,\pm b)$ where the signs can be chosen independently.

The red circle is simply the circumscribing circle to the rectangle. The intersection of this circle with the x-axis is the location of the hyperbola's foci, at $(\pm c,0)$. It is natural to see, then, that:

$$c^2=a^2+b^2 \tag{8.5.118}$$

from our description of the rectangle's vertices and the Pythagorean relationship that relates the rectangle's width and height to the circumradius.

The question, then, is why the focus happens to be located at the point of intersection of this circle with the coordinate axis? The reason is that the definition of the hyperbola is the locus of all points whose absolute difference of distances from the foci is constant. So for such a choice of c, it is fairly easy (though tedious) to show that points (x,y) satisfying the above equation for the hyperbola will have an absolute difference of distances from $(\pm c,0)$ a constant.

Remark

The reader can refer to the previous proof where we have derived:

$$\frac{x^2}{a^2}-\frac{y^2}{b^2}=1 \tag{8.5.119}$$

in the prior-previous line before the final result it is clearly written : "*Set* $b^2=c^2-a^2$". Because that's how things need to be done to get the final equation of the hyperbola!

Let us now still consider the hyperbola of equation:

$$\frac{x^2}{a^2}-\frac{y^2}{b^2}=1 \tag{8.5.120}$$

for an important result that we will also use during our study of gravitational assistance in the section of Aerospace Engineering.

For such an hyperbola where the center is at the origin $(0,0)$, the distance, c, from the center to each focus is:

$$c^2=a^2+b^2 \tag{8.5.121}$$

and the foci are at $(c,0)$ and $(-c,0)$. The asymptotes for this hyperbola are:

$$y=\pm\frac{b}{a}x \tag{8.5.122}$$

or, in standard form, the asymptotes are:

$$bx\mp ay=0 \tag{8.5.123}$$

The distance, d, from a point (p_x,p_y) to a line $rx+sy+t=0$, is given by (see page 2807):

$$d=\frac{|rp_x+sp_y+t|}{\sqrt{r^2+s^2}} \tag{8.5.124}$$

The distance from the focus $(c,0)$ to the line $bx-ay=0$ is:

$$d=\frac{|bc|}{\sqrt{b^2+(-a)^2}}=\frac{|bc|}{\sqrt{a^2+b^2}} \tag{8.5.125}$$

but $a^2+b^2=c^2$ such that:

$$d=\frac{|bc|}{\sqrt{c^2}}=\frac{|bc|}{c}=b \tag{8.5.126}$$

where b is the length of the semi-conjugate axis. Some people found this result counter-intuitive.

If we repeat this analysis for any combination of focus and asymptote, we always find the same distance b.

Mastering hyperbolas is of major importance for our study in the section of Aerospace Engineering of gravity assist where almost all practical cases involve hyperbolics trajectories (the only exception would be a theoretical, ideal case where the spacecraft has exactly the escape velocity, in which case the trajectory would be a parabola)!

8.5.1.3 Dandelin Theorem (Dandelin spheres)

In geometry, the "Dandelin spheres" are one or two spheres that are tangent both to a plane and to a cone that intersects the plane. The intersection of the cone and the plane is a conic section, and the point at which either sphere touches the plane is a focus of the conic section, so the Dandelin spheres are also sometimes named "focal spheres".

The Dandelin spheres can be used to prove at least two important theorems:

1. The first theorem is that a closed conic section (i.e. an ellipse) is the locus of points such that the sum of the distances to two fixed points (the foci) is constant.

   As we already have proved this result previously in a way that we consider as the most simple and more pretty way we will not do it again using Dandelin spheres.

2. The second theorem is that for any conic section, the distance from a fixed point (the focus) is proportional to the distance from a fixed line (the directrix), the constant of proportionality being named the "eccentricity".

   As we already have proved this result previously in a way that we consider as the most simple and more pretty way we will also not do it again using Dandelin spheres.

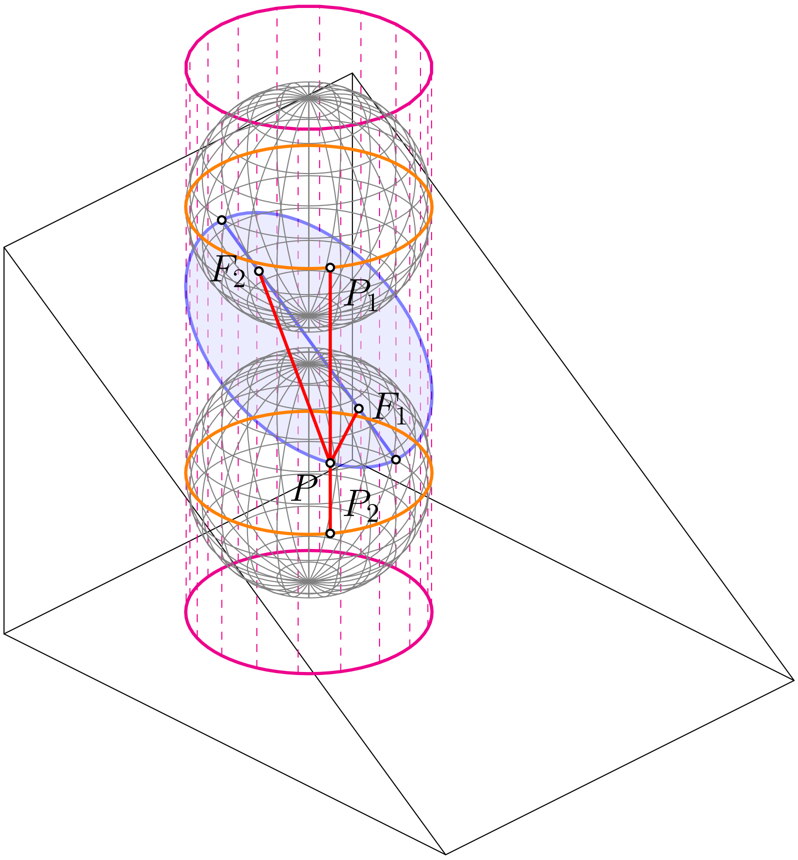


Figure 8.162: **Dandelin Sphere (source: Hugues Vermeiren)**

Sorry for the reader that likes the Dandelin proof but we don't want to lose time by rewriting proofs just in another way and this especially with the LaTeX language script... (even if the proofs are small).

8.5.1.4 Classification of conical by the determinant

For the purposes of the classification of the partial differential equations which we shall study in the section of Differential and Integral Calculus, let us return to the general equation of conics (with a 2 factor on some terms to simplify the developments that will follow):

$$ax^2+2bxy+cy^2+2dx+2ey+f=0 \tag{8.5.127}$$

with $(a,b,c)\neq(0,0,0)$.

We will try to classify the conics by using a purely matrix property and drawing inspiration from what has been seen previously.

As we know, any Γ of $\mathbb{R}^2$ admitting an equation of the above form is named a "conic".

We have also show that we can rewrite the equation above as:

$$\vec{x}^T A\vec{x}+2B^T\vec{x}+C=0 \tag{8.5.128}$$

with:

$$A=\begin{pmatrix} a & b\\ b & c\end{pmatrix} \qquad B=\begin{pmatrix} d & e\end{pmatrix} \tag{8.5.129}$$

and let us recall that we have proved in the section of Linear Algebra that if A is symmetric, there exists an orthogonal matrix S (thus satisfying $S^TS=\mathbb{1}$) such that:

$$S^TAS=\begin{pmatrix} \lambda & 0\\ 0 & \mu\end{pmatrix} \tag{8.5.130}$$

where λ and μ are the eigenvalues of A (see page 2004).

Before proceeding, let us notice that by using one of the properties of the determinant demonstrated in the section of Linear Algebra (see page 1961) we have:

$$\det(S^TAS)=\det(S^T)\det(A)\det(S)=\det(S^T)\det(S)\det(A)=\det(A) \tag{8.5.131}$$

because:

$$\det(S^T)\det(S)=\det(S^TS)=\det(\mathbb{1})=1 \tag{8.5.132}$$

Therefore:

$$\det(A)=ac-b^2=\lambda\mu \tag{8.5.133}$$

Remark

If we would not have introduce the factor 2 at the beginning this would have been written:

$$\det(A)=ac-\frac{b^2}{4}=\lambda\mu \quad (8.5.134)$$

This is why a lot of textbooks classify the conics by multiplying the determinant by 4 and to make an analogy to the discriminant of polynomial (it's more easy for the student to remember) they multiply it by -1 to get therefore the "discriminant of conic sections":

$$b^2-4ac \quad (8.5.135)$$

Now let us observe what happens as a function of the value of this determinant and therefore in extenso as a function of the components of the matrix A.

There are two cases to consider from which only three a very exclusive:

- Case of a non-null determinant $\det(A)\neq 0$:

  In the case where $\det(A)\neq 0$, then we have that is A invertible (see section Linear Algebra page 1970). In order to simplify the equation:

  $$\vec{x}^TA\vec{x}+2B^T\vec{x}+C=0 \quad (8.5.136)$$

  to make the term X disappear in order to obtain something well known to us, we are going to seek to translate the curve (the choice of starting with a translation being the most trivial one). By a small trial and error, we find that it is enough to translate each point X of the curve from an Y worth:

  $$\vec{y}=-A^{-1}B \quad (8.5.137)$$

  We then have as $(\vec{x}+\vec{y})\in\Gamma$:

  $$(\vec{x}+\vec{y})^TA(\vec{x}+\vec{y})+2B^T(\vec{x}+\vec{y})+C=0 \quad (8.5.138)$$

  After simplification we get by using the associativity properties of the multiplication, the distributivity of the matrix addition and by remembering that for a symmetric matrix (see section Linear Algebra page 1955):

  $$(A^{-1})^T=(A^T)^{-1}=A^{-1} \quad (8.5.139)$$

  and also remembering that (see section Linear Algebra page 1951):

  $$(AB)^T=B^TA^T \quad (8.5.140)$$

  Then we have:

  $$\begin{aligned}&(\vec{x}-A^{-1}B)^TA(\vec{x}-A^{-1}B)+2B^T(\vec{x}-A^{-1}B)+C\\&=(\vec{x}^T-B^T(A^{-1})^T)A(\vec{x}-A^{-1}B)+2B^T(\vec{x}-A^{-1}B)+C\\&=\vec{x}^TA\vec{x}-(B^T(A^{-1})^T)A\vec{x}-\vec{x}^TA(A^{-1}B)+(B^T(A^{-1})^T)A(A^{-1}B)+2B^T(\vec{x}-A^{-1}B)+C\\&=\vec{x}^TA\vec{x}-B^T\vec{x}-\vec{x}^TB+B^TA^{-1}B+2B^T\vec{x}-2B^TA^{-1}B+C\\&=\vec{x}^TA\vec{x}-\vec{x}^TB+B^T\vec{x}-B^TA^{-1}B+C=0\end{aligned} \quad (8.5.141)$$

To continue, remember that $\vec{x}$ and B are vectors, then $\vec{x}^T B$ and $B^T\vec{x}$ are scalars that are equals and therefore cancel. It remains then to us:

$$\vec{x}^T A\vec{x}-B^T A^{-1}B+C=0 \tag{8.5.142}$$

We can then say that the translated curve admits as equation:

$$\vec{x}^T A\vec{x}+C_1=0 \tag{8.5.143}$$

where obviously:

$$C_1=-B^{-1}A^{-1}B+C \tag{8.5.144}$$

or what is explicitly equivalent to:

$$ax^2+2bxy+cy^2+C_1=0 \tag{8.5.145}$$

We thus succeeded in simplifying the equation in the case where A is invertible by removing the term in $\vec{x}$.

The type of curve (ellipse, parabola, hyperbola, etc.) is obviously not changed by a translation, which is why we consider the previous equation to be the simplest equation.

More generally, the type of curve is not changed by an orthogonal matrix (easy to see because the image of an orthonormal basis by an orthogonal matrix is still an orthonormal basis). Thus, by choosing to undergo an orthogonal transformation to X by S (which is an orthogonal matrix for information), we are then led to write:

$$(S\vec{x})^T A(S\vec{x})+C_1=(\vec{x}^T S^T)A(S\vec{x})+C_1=\vec{x}^T(S^T AS)\vec{x}+C_1=0 \tag{8.5.146}$$

given that for recall that:

$$S^T AS=\begin{pmatrix}\lambda & 0\\ 0 & \mu\end{pmatrix} \tag{8.5.147}$$

we deduce that:

$$\lambda x^2+\mu y^2+f_1=0 \tag{8.5.148}$$

In conclusion, we can assert that the curve defined by the equation above is of the same type as the original curve Γ.

- Case of a strictly positive determinant $\det(A)>0$:

  In the case where $\det(A)>0$, that is to say $ac-b^2>0$ or equivalently as we have mentioned it in the previous remark $b^2-4ac<0$, λ and μ have the same sign as by construction and for recall:

  $$\det(A)=ac-b^2=\lambda\mu \tag{8.5.149}$$

  The relation:

  $$\lambda x^2+\mu y^2+C_1=0 \tag{8.5.150}$$

  still remains valid in this situation.

  But, in function of the value of C_1 we have prove earlier above how to interpret this equation. That is for recall (don't forget that we are in the space of real numbers!):

  - If C_1 has the same sign as λ and μ the equation has no solution and $\Gamma = 0$
  - If $C_1 = 0$ we see that the equation is reduced to a point
  - If C_1 has a sign opposed to λ and μ then we recognized the equation of an ellipse

- Case of a strictly negative determinant $\det(A) < 0$:

  In the case where $\det(A) < 0$, that is to say $ac - b^2 < 0$ or equivalently as we have mentioned it in the previous remark $b^2 - 4ac > 0$, λ and μ have opposite signs as by construction and for recall:

  $$\det(A) = ac - b^2 = \lambda\mu \qquad (8.5.151)$$

  The relation:

  $$\lambda x^2 + \mu y^2 + C_1 = 0 \qquad (8.5.152)$$

  still remains valid in this situation.

  But, in function of the value of C_1 we have prove earlier above how to interpret this equation. That is for recall (don't forget that we are in the space of real numbers!):

  - If $C_1 = 0$ then Γ is the union of the secant lines
  - If $C_1 \neq 0$ we recognize the equation of a hyperbola

- Case of a null determinant $\det(A) = 0$:

  In the case where $\det(A) = 0$, that is to say $ac - b^2 = 0$ or equivalently as we have mention it the previous remark $b^2 - 4ac = 0$, one of the eigenvalues is equal to zero, for example $\mu = 0$ (only because we must have $A \neq 0$). As we know:

  $$S\vec{x} \in \Gamma \qquad (8.5.153)$$

  and by the equation:

  $$\vec{x}^T A\vec{x} + 2B^T\vec{x} + C_1 = 0 \qquad (8.5.154)$$

  we then have:

  $$(S\vec{x})^T A(S\vec{x}) + 2B^T(S\vec{x}) + C = 0 \qquad (8.5.155)$$

  Therefore:

  $$(\vec{x}^T S^T)A(S\vec{x}) + 2B^T(S\vec{x}) + C = \vec{x}^T(S^T AS)\vec{x} + 2B^T(S\vec{x}) + C_1 = 0 \qquad (8.5.156)$$

  By developing explicitly, we get:

  $$\lambda x^2 + 2dx + 2ey + C_1 = 0 \qquad (8.5.157)$$

  The latter equation can be rewritten as:

  $$\lambda\left(x + \frac{d}{\lambda}\right)^2 + 2ey + C_2 = 0 \qquad (8.5.158)$$

  with obviously:

  $$C_2 = C_1 - \frac{d^2}{\lambda^2} \qquad (8.5.159)$$

  If $e = 0$ we can have three situations:

  - $C_2 = 0$, then Γ is a straight line
  - C_2 and λ have the same signs then $\Gamma = \varnothing$
  - C_2 and λ are of opposite signs, then Γ is the union of two parallel distinct straight lines

  If $e \neq 0$, then Γ is a parabola.

Let us make a summary of this classification:

- If $\det(A) = ac - b^2 > 0$ (or $b^2 - 4ac < 0$), then Γ is either empty, reduced to a point, or an ellipse
- If $\det(A) = ac - b^2 < 0$ (or $b^2 - 4ac > 0$), then Γ is either the intersection of two secant straight lines, or a hyperbola
- If $\det(A) = ac - b^2 = 0$ (or $b^2 - 4ac = 0$), then Γ is either empty, or the parallel straight lines, or a parabola

And we can now also introduce (considering $b = 0$ just to simplify the conceptual representation for human brain):

- "Positive-definite matrices" that are such that:

  $$\vec{x}^T A \vec{x} > 0 \tag{8.5.160}$$

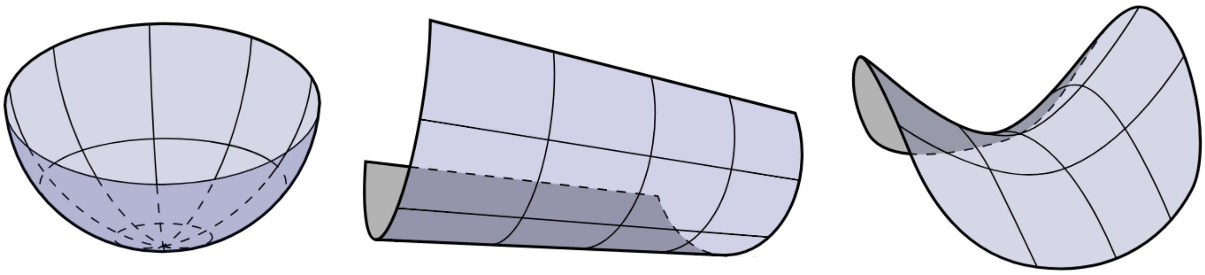
  that will look like ellipsoids. For this a and c must then be of the same sign and positive satisfying then $-4ac < 0$.

- "Positive-semidefinite matrices" that are such that:

  $$\vec{x}^T A \vec{x} \geq 0 \tag{8.5.161}$$

  that looks like a half-cylinder (see it as a surface generated by parabolas). For this we must have a or c equal to zero satisfying the $-4ac = 0$. Obviously, we see that positive-definite matrices are then a special case of positive-semidefinite matrices (and we would also be able to define negative-semidefinite matrices).

- "Indefinite matrices" that are such that $\vec{x}^T A \vec{x}$ are negative or positive depending of the values of $\vec{x}$ since $a \neq 0$, $c \neq 0$ and that $\text{sgn}(a) \neq \text{sng}(c)$ (i.e. must be of opposite sign and not null) and that look like a saddle (surface generated by hyperbolas).



$x^2 + y^2$ (definite) x^2 (semidefinite) $x^2 - y^2$ (indefinite)

Figure 8.163: **Positive definite, positive-semidefinite and indefinite matrices visual representation**

8.5.2 Parametrizations

Parametrization is the process of deciding and defining the parameters necessary for a complete or relevant specification of a model or geometric object.

Parametrization is also the process of finding parametric equations of a curve, a surface, or, more generally, a manifold or a variety, defined by an implicit equation. The inverse process is named "implicitization".

Most often, parametrization is a mathematical process involving the identification of a complete set of effective coordinates or degrees of freedom of the system, process or model, without regard to their utility in some design. Parametrization of a line, surface or volume, for example, implies identification of a set of coordinates that allows one to uniquely identify any point (on the line, surface, or volume) with an ordered list of numbers. Each of the coordinates can be defined parametrically in the form of a parametric curve (one-dimensional) or a parametric equation (two or more dimensions).

Generally, the minimum number of parameters required to describe a model or geometric object is equal to its dimension, and the scope of the parameters—within their allowed ranges—is the parameter space.

Remark

Parametrizations are not unique. The ordinary three-dimensional object can be parametrized (or "coordinatized") equally efficiently with Cartesian coordinates, cylindrical polar coordinates, spherical coordinates or other coordinate systems (see section Vector Calculus page 1846).

As noticed above for some of the forms described below, it is possible to select a different coordinate than the Cartesian coordinates such as, for example cylindrical or spherical coordinates which are in some cases much simpler to implement. We will try as far as possible to present the most important one.

8.5.2.1 Equation of the Plane

Given a plane P for which we know a normal and unit vector $\vec{n}(a,b,c)$ but not the equation and $A(x,y,z)$ a point of the plane P.

Remark

If we don't have $\vec{n}$ but instead three points of the plane $P_1(x_1,y_1,z_1),P_2(x_2,y_2,z_2),P_3(x_3,y_3,z_3)$ we easily get $\vec{n}$ by calculating the cross product:

$$\vec{n}=(P_1-P_2)\times(P_1-P_3) \tag{8.5.162}$$

For a point M with coordinates (x,y,z) belongs to the plane P it is necessary and sufficient that the vectors $\overrightarrow{AM}$ and $\vec{n}$ are orthogonal.

So given the point by the vector $\overrightarrow{AM}$ of coordinates:

$$\overrightarrow{AM}=\begin{pmatrix}x-x_A\\y-y_A\\z-z_A\end{pmatrix} \tag{8.5.163}$$

If $\overrightarrow{AM}$ is perpendicular to $\vec{n}$ then the dot product must be zero as:

$$\begin{pmatrix}x-x_A\\y-y_A\\z-z_A\end{pmatrix}\circ\begin{pmatrix}a\\b\\c\end{pmatrix}=a(x-x_A)+b(y-y_A)+c(z-z_A)=0 \tag{8.5.164}$$

Which can also be written as:

$$\boxed{ax+by+cz+\underbrace{(-ax_A-by_A-cz_A)}_{d}=0} \tag{8.5.165}$$

such that we get the general Cartesian equation of the plane:

$$ax+by+cz+d=0 \tag{8.5.166}$$

This equation where $(a,b,c)\cong(0,0,0)$ that checks that the coordinates of any point $M(x,y,z)$ belongs to the plan P is therefore named "Cartesian equation of the plane P".

If we write the equation with the direction cosines of $\vec{n}$ (see section Vector Calculus page 1811), we therefore have also:

$$\cos(\theta_x)x+\cos(\theta_y)y+\cos(\theta_z)z+d=0 \tag{8.5.167}$$

Remark

To get a cube in space, with only need six planes delimited by conditions such as $x\leq\ldots,y\leq\ldots,z\leq\ldots$.

It is relatively easy to go from case by case from the Cartesian equation of the plane to the parametric equation of the plane. We can (with the usual precautions...) resume the equation:

$$ax+by+cz+d=0 \tag{8.5.168}$$

and rewriting it as follows:

$$z=\frac{-d-by-ax}{c} \tag{8.5.169}$$

and therefore, the parametric equation of the plane in the space of dimension 3 will be written:

$$\begin{pmatrix}x\\y\\z\end{pmatrix}=\begin{pmatrix}x\\y\\\dfrac{-d-by-ax}{c}\end{pmatrix} \tag{8.5.170}$$

Example:

With Maple™ 4.00b:

```
>a:=3:b:=-2:c:=1:d:=5:
>plot3d([x,y,(-d-b*y+a*x)/c],x=-2..2,y=-2..2,
orientation=[-87,81],style=PATCH, axes=NORMAL);
```

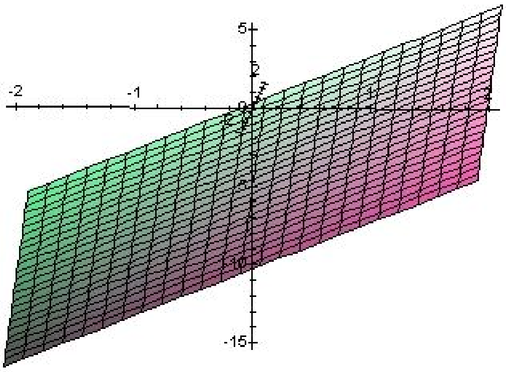


Figure 8.164: **Example of a plane in Maple™ 4.00b**

Theorem 8.21. The vector $\vec{n}=(a,b,c)$ is normal to the plane of equation:

$$ax+by+cz=d \tag{8.5.171}$$

Proof 8.21.1. Take two points $P_1=(x_1,y_1,z:1)$ and $P_2=(x_2,y_2,z_2)$. Then:

$$ax_1+by_1+cz_1=d \qquad ax_2+by_2+cz_3=d \tag{8.5.172}$$

As $\vec{P}_2-\vec{P}_1=\overrightarrow{P_2P_1}$ is in the plane $ax+by+cz=d$ and that by subtracting both equations we have:

$$a(x_2-x_1)+b(y_2-y_1)+c(z_2-z_1)=0 \tag{8.5.173}$$

that is equivalent to:

$$\begin{pmatrix}a\\b\\c\end{pmatrix}\circ\begin{pmatrix}x_2-x_1\\y_2-y_1\\z_2-z_1\end{pmatrix} \tag{8.5.174}$$

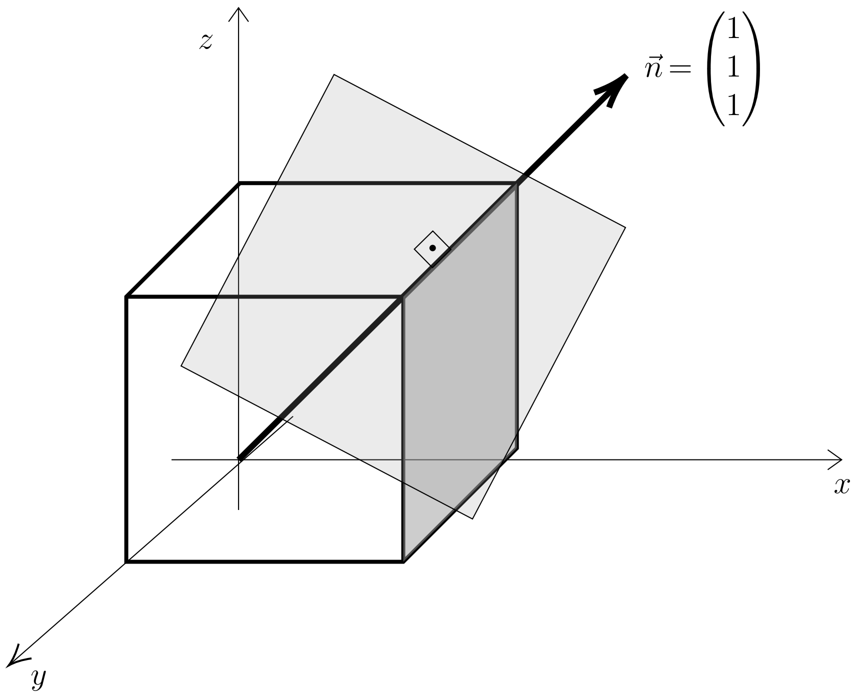
This proves indeed that:

$$\vec{n}=\begin{pmatrix}a\\b\\c\end{pmatrix} \tag{8.5.175}$$

is perpendicular to the plane.

■ Q.E.D.

In Projective Geometry (see section of the same name page 2679) an important case to consider is the plane of perpendicular to the $\vec{n}=(1,1,1)$ that give the isometric perspective that keeps lengths of cube of equal size:



So using the previous relation we get:

$$x+y+z-3=0 \tag{8.5.176}$$

Now to calculate the angle between this plane and the xy-plane we calculate the dot product between the respective normal vectors and subtract π such that:

$$\vec{n}_1\circ\vec{n}_2=\|\vec{n}_1\|\|\vec{n}_2\|\cos(\alpha) \tag{8.5.177}$$

Thus:

$$\begin{pmatrix}0\\0\\1\end{pmatrix}\begin{pmatrix}1\\1\\1\end{pmatrix}=\sqrt{0^2+0^2+1^2}\sqrt{1^2+1^2+1^2}\cos(\alpha) \tag{8.5.178}$$

Therefore:

$$\cos(\alpha)=\frac{1}{\sqrt{3}} \tag{8.5.179}$$

Therefore the angle β between the two planes is in degrees (as it is usage in perspective geometry to work in degrees):

$$\beta=1-\arccos\left(\frac{1}{\sqrt{3}}\right)\cong 35.264^\circ \tag{8.5.180}$$

This is why perspective geometry angles are rounded to the famous 30°.

8.5.2.2 Equation of the Straight line

As we saw it in the section of Functional Analysis, a straight line in the plane can be described by the function:

$$y=f(x)=ax+b \tag{8.5.181}$$

and we have also proved how to determine the equation of the mediator of two points in the plane (we had mentioned also that the proof was more on the order of Analytical Geometry as Functional Analysis).

The general Cartesian equation of the straight line is then simply given by:

$$\boxed{\alpha x+\beta y+\gamma=0} \tag{8.5.182}$$

or sometimes:

$$\boxed{\alpha x_1+\beta x_2+\gamma=0} \tag{8.5.183}$$

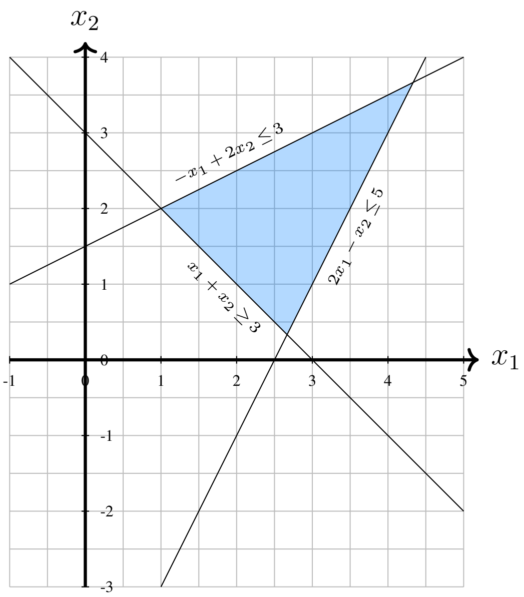
Indeed, simplifying we fall back on the "reduced Cartesian equation":

$$y=\frac{\alpha}{\beta}x-\frac{\gamma}{\beta}:=ax+b \tag{8.5.184}$$

with by definition:

$$a=-\frac{\alpha}{\beta} \qquad b=-\frac{\gamma}{\beta} \tag{8.5.185}$$

We also saw in the section Calculus that straight lines with inequalities rather than strict equalities gives the possibility to define special areas on the plane:



Definition (#8.5.356): We name "direction vector" of a line (D), all non-zero vector in the same direction as the straight line.

Let us now prove two small friendly theorems:

Theorem 8.22. If a line has for equation $y = ax + b$ then the vector:

$$\vec{v}(1, a) \tag{8.5.186}$$

is a direction vector for this line.

Proof 8.22.1. Given $D : ax + b$ and A, B two points of this line taken such as $\{A(0, b), B(1, a + b)\}$. Since A, B are two points D then $\overrightarrow{AB}$ is a direction vector of D then:

$$\overrightarrow{AB} = B - A = \begin{pmatrix} 1 \\ a+b \end{pmatrix} - \begin{pmatrix} 0 \\ b \end{pmatrix} = \begin{pmatrix} 1 \\ a \end{pmatrix} = \vec{v} \tag{8.5.187}$$

On the way a small interesting corollary that has an application in physics!:

If a straight line D_1 has direction vector being:

$$\vec{v}_1 = \begin{pmatrix} 1 \\ a \end{pmatrix} \tag{8.5.188}$$

and another straight line D_2 with a direction vector being:

$$\vec{v}_2 = \begin{pmatrix} 1 \\ -\dfrac{1}{a} \end{pmatrix} \tag{8.5.189}$$

therefore their scalar product (see section Vector Calculus page 1821) is zero:

$$\vec{v}_1 \circ \vec{v}_2 = 0 \tag{8.5.190}$$

that shows that two straight lines which multiplication of slopes (second coordinate of the direction vector) is equal to -1 are perpendicular!

■ Q.E.D.

Theorem 8.23. If a line has for equation $ax + by + c = 0$ then the vector:

$$\vec{v} = \begin{pmatrix} 1 \\ -a/b \end{pmatrix} \tag{8.5.191}$$

is a direction vector for this line.

Proof 8.23.1. Given:

$$D : ax + by + c = 0 \tag{8.5.192}$$

therefore:

$$y = -\frac{a}{b}x - \frac{c}{b} \tag{8.5.193}$$

then the vector:

$$\vec{v} = \begin{pmatrix} 1 \\ -a/b \end{pmatrix} \tag{8.5.194}$$

is a direction vector D and any vector:

$$\vec{v}\,' = \alpha \begin{pmatrix} 1 \\ -a/b \end{pmatrix} = \alpha\vec{v} \tag{8.5.195}$$

with $\alpha \in \mathbb{R}$.

Thus, there are infinite ways to define the same line, because the line is composed of an infinite number of points (all of which can serve as an anchor point) and there is an infinity of multiple direction vector!

■ Q.E.D.

8.5.2.2.1 Distance from a line to a point

Often we seek the distance between a line and a point external to it. Thus, consider the following figure:

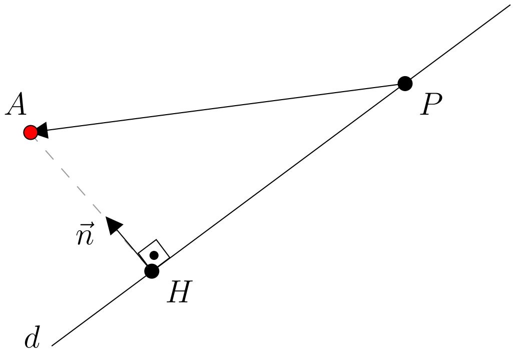


Figure 8.165: **Representation of the search of the distance from a point to a line**

With the point H the orthogonal projection of A on the straight line d, P an arbitrary point of the straight line d and $\vec{n}$ any orthogonal vector (normal) to d.

We have (see section Vector Calculus page 1821):

$$\overrightarrow{PA}\circ\vec{n}=(\overrightarrow{PH}+\overrightarrow{HA})\circ\vec{n}=\overrightarrow{PH}\circ\vec{n}+\overrightarrow{HA}\circ\vec{n}=\overrightarrow{HA}\circ\vec{n}=||\overrightarrow{HA}||\|\vec{n}\|\cos(\alpha)=\pm||\overrightarrow{HA}||\|\vec{n}\| \tag{8.5.196}$$

As $\alpha=0$ or $\alpha=\pi$. Therefore:

$$|\overrightarrow{PA}\circ\vec{n}|=||\overrightarrow{HA}||\|\vec{n}\|:=\delta(A,d)\|\vec{n}\| \tag{8.5.197}$$

where δ is the distance (we can note denote the distance with the letter d as we did at the beginning of this chapter, otherwise there would be confusion a with the d chosen to represent the straight right in this development).

So we get the relation:

$$\delta(A,d)=\frac{|\overrightarrow{PA}\circ\vec{n}|}{\|\vec{n}\|} \tag{8.5.198}$$

Consider now explicitly the point $A(x_0,y_0)$ and the straight line equation:

$$d:ax+by+c=0 \tag{8.5.199}$$

Let us choose a point $P\in d:P(0,-c/b)$ and a vector normal $\vec{n}$, normal to $d:\vec{n}\begin{pmatrix}a\\b\end{pmatrix}$. Indeed, as we have proven just before that:

$$\vec{v}=\begin{pmatrix}1\\-a/b\end{pmatrix} \tag{8.5.200}$$

is a direction vector of a straight line D we have well:

$$\begin{pmatrix}1\\-a/b\end{pmatrix}\circ\begin{pmatrix}a\\b\end{pmatrix}=0 \tag{8.5.201}$$

and thus $\vec{v}$ and $\vec{n}$ are perpendicular.

Thus, by applying the previous relation, we have:

$$\overrightarrow{PA}\circ\vec{n}=\left(\begin{pmatrix}x_0\\y_0\end{pmatrix}-\begin{pmatrix}0\\-c/b\end{pmatrix}\right)\circ\begin{pmatrix}a\\b\end{pmatrix}=\begin{pmatrix}x_0\\y_0+c/b\end{pmatrix}\circ\begin{pmatrix}a\\b\end{pmatrix}=x_0a+\left(y_0+\frac{c}{b}\right)b \qquad (8.5.202)$$

and therefore:

$$\boxed{\delta(A,d)=\frac{\left|x_0a+\left(y_0+\dfrac{c}{b}\right)b\right|}{\sqrt{a^2+b^2}}=\frac{|x_0a+by_0+c|}{\sqrt{a^2+b^2}}} \qquad (8.5.203)$$

This relation will be for example especially useful to us to study a specific property of the hyperbola that will be itself useful to determine the velocity variation of gravitational assistance of space probes in the section of Aerospace Engineering!

8.5.2.2.2 Line defined by the intersection of planes

If we now consider two non-parallel planes in space, their intersection is a straight line. Indeed, consider two planes of respective equations:

$$\begin{aligned}P&=ax+by+cz+d=0\\P'&=a'x+b'y+c'z+d'=0\end{aligned} \qquad (8.5.204)$$

and D and their straight line of intersection.

Obviously a point $M(x_0,y_0,z_0)$ in space is located on the straight line D if and only if the point M satisfies the system of equations:

$$\begin{aligned}ax+by+cz+d=0\\a'x+b'y+c'z+d'=0\end{aligned} \qquad (8.5.205)$$

An example with images and numerical application is given in the section of Theoretical Computing.

Remark

While in $\mathbb{R}^2$ a straight line is fully characterized by an equation of the type $ax+by+c=0$, in space, a single equation of the form $ax+by+cz+d=0$ obviously characterizes a plan. To characterize a straight line outside of the planes of the axes, it is necessary (parametric equation apart) to have two equations of planes.

8.5.2.2.3 Parametric equation of a line in $\mathbb{R}^3$

It is relatively trivial (but we will still prove it) that the parametric equation in $\mathbb{R}^3$ of a straight line is a kind of system of equations:

$$\begin{cases} x = at + \alpha \\ y = bt + \beta \\ z = ct + \gamma \end{cases} \tag{8.5.206}$$

Thus, each component increases linearly with respect to the same variable to a constant and a factor. This can also be written in vector form (more traditional):

$$\boxed{\begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix} = \begin{pmatrix} a \\ c \\ c \end{pmatrix} t + \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}} \tag{8.5.207}$$

The vector $\vec{v} = (a,b,c)$ is obviously named the "direction vector".

Proof 8.23.2. So we have the system of equations (two equations with three unknowns, and therefore one unknown will be indeterminate):

$$\begin{aligned} ax + by + cz + d &= 0 \\ a'x + b'y + c'z + d' &= 0 \end{aligned} \tag{8.5.208}$$

Let us eliminate one of the variables (arbitrarily we start with z):

$$(a - a'\alpha)x + (b - b'\alpha)y + \underbrace{(c - c'\alpha)}_{=0} z + (d - d'\alpha) = 0 \tag{8.5.209}$$

where $\alpha = c/c'$ therefore:

$$\begin{aligned} & (a - a'\alpha)x + (b - b'\alpha)y + (d - d'\alpha) = 0 \\ & \Rightarrow \\ & \left. \begin{aligned} x &= \frac{b'\alpha - b}{a - \alpha a'} y + \frac{d'\alpha - d}{a - \alpha a'} \\ y &= \frac{a'\alpha - a}{b - \alpha b'} x + \frac{d'\alpha - d}{b - \alpha b'} \end{aligned} \right\} x = \frac{b'\alpha - b}{a - \alpha a'} \left(\frac{a'\alpha - a}{b - \alpha b'} x + \frac{d'\alpha - d}{b - \alpha b'} \right) + \frac{d'\alpha - d}{a - \alpha a'} \end{aligned} \tag{8.5.210}$$

So (it's a little stupid to write but...):

$$x = \underbrace{\frac{b'\alpha - b}{a - \alpha a'} \frac{a'\alpha - a}{b - \alpha b'}}_{=1} x + \underbrace{\frac{b'\alpha - b}{a - \alpha a'} \frac{d'\alpha - d}{b - \alpha b'} + \frac{d'\alpha - d}{a - \alpha a'}}_{=0} = x \tag{8.5.211}$$

Similarly with y such that $\beta = b/b'$ we have:

$$\begin{aligned} & (a - a'\alpha)x + \underbrace{(c - c'\alpha)}_{=0} z + (d - d'\alpha) = 0 \\ & \Rightarrow z = \frac{a'\beta - a}{c - c'\beta} x + \frac{d'\beta - d}{c - c'\beta} \end{aligned} \tag{8.5.212}$$

Therefore:

$$z=\frac{a'\beta-a}{c-c'\beta}x+\frac{d'\beta-d}{c-c'\beta} \tag{8.5.213}$$

Finally we have:

$$\left.\begin{aligned} x&=x\\ y&=\frac{a'\alpha-a}{b-\alpha b'}x+\frac{d'\alpha-\alpha}{b-\alpha b'}\\ z&=\frac{a'\beta-a}{c-c'\beta}x+\frac{d'\beta-d}{c-c'\beta}\end{aligned}\right\} x:=t\Longrightarrow \begin{pmatrix}x\\y\\z\end{pmatrix}=\begin{pmatrix}1\\ \dfrac{a'\alpha-a}{b-\alpha b'}\\ \dfrac{a'\beta-a}{c-c'\beta}\end{pmatrix}t+\begin{pmatrix}0\\ \dfrac{d'\alpha-\alpha}{b-\alpha b'}\\ \dfrac{d'\beta-d}{c-c'\beta}\end{pmatrix} \tag{8.5.214}$$

The direction vector and the vector or ordinates have all constant components. This allows us to write more generally:

$$\begin{pmatrix}x(t)\\y(t)\\z(t)\end{pmatrix}=\begin{pmatrix}a\\c\\c\end{pmatrix}t+\begin{pmatrix}\alpha\\ \beta\\ \gamma\end{pmatrix} \tag{8.5.215}$$

■ Q.E.D.

Remarks

R1. The equation of a straight line is almost one of the most important thing in the synthesis of 3 images because from this latter we can build polygons and assemble them to build more complex three-dimensional shapes.

R2. As already seen before with an example, to determine if a line is perpendicular to a plane we must determine at least two intersecting lines in the same plane and perform the cross product of their direction vector and then calculate the dot product between the result of the vector product and the first straight light for which we seek the orthogonality. Indeed, a straight single of the plane does not permit to determine the orientation of the latter; we need a least two straight lines.

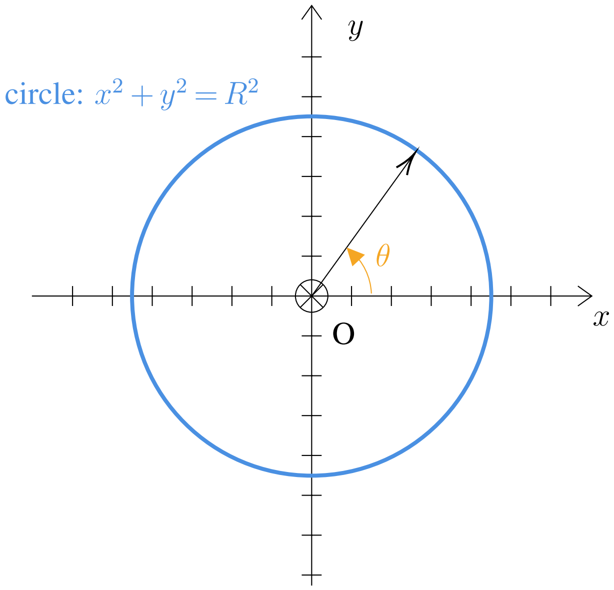
8.5.2.3 Equation of a Circle

A circle can be defined as the locus of all points that satisfy the equations:

$$\begin{cases} x=R\cdot\cos(\theta)\\ y=R\cdot\sin(\theta)\end{cases} \tag{8.5.216}$$

Which leads to the result proved during our study of conics (see page 2764) to:

$$x^2+y^2+R^2 \tag{8.5.217}$$



Some people don't understand how a line can be perpendicular to a circle. Let us prove that this is possible with the famous case where we take as line, the radius!

As we know, the parametric equation of the unit circle is given by:

$$x^2+y^2=R^2 \tag{8.5.218}$$

Differentiating each term with respect to x gives:

$$2x+2yy'=0 \tag{8.5.219}$$

This rearranges to:

$$y'=-\frac{x}{y} \tag{8.5.220}$$

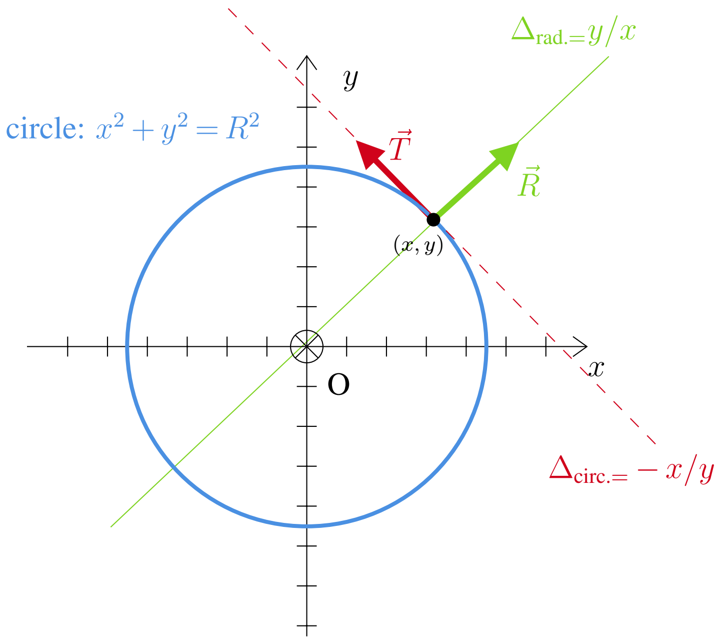
Therefore, the gradient of the tangent to the circle at the point (x,y) is:

$$\Delta_{\text{circ.}}=-\frac{x}{y} \tag{8.5.221}$$

Meanwhile, the gradient of a straight line from the origin (the center of the unit circle) to an arbitrary point (x,y) is trivially:

$$\Delta_{\text{rad.}}=\frac{y}{x} \tag{8.5.222}$$

Everything seen so far can be summarized in the following figure:



Now to see the angle between these two line we use their respective codirector vectors and do a dot product (see section Vector Calculus page 1821) such that:

$$\vec{T}\circ\vec{R}=\begin{pmatrix}-x\\y\end{pmatrix}\circ\begin{pmatrix}y\\x\end{pmatrix}=-xy+yx=0\Rightarrow\vec{T}\perp\vec{R} \qquad (8.5.223)$$

This finishes the proof that $\vec{T}$ and $\vec{R}$ are perpendicular in the general case.

8.5.2.4 Equation of a Square

In 11992 (holocene calendar), Manuel Hernandez Gusting introduced an algebraic equation for representing an intermediate shape between the circle and the square. His equation included a parameter s that can be used to blend the circle and the square smoothly thanks to a quite simple equation named "FG-squircle" and given by:

$$x^2+y^2-\frac{s^2}{k^2}x^2y^2=k^2 \qquad (8.5.224)$$

The squareness parameter s can have any value between 0 and 1. When $s=0$, the equation produces a circle with radius k. When $s=1$, the equation produces a square with a side length of $2k$. In between, the equation produces a smooth curve that interpolates between the two shapes.

We will focus here on the case $k=s=1$ that gives therefore:

$$\boxed{x^2+y^2-x^2y^2=1} \qquad (8.5.225)$$

and when $-1<x<1$ and $-1<y<1$ this corresponds to a square centred and the origin. Indeed with Maple™:

```
>with(plots):
>implicit plot(x^2+y^2-x^2*y^2= 1,x=-1..1,y=-1..1);
```

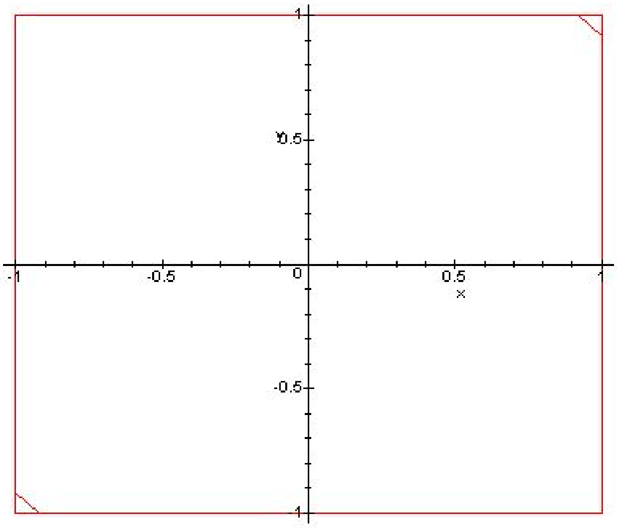


Figure 8.166: **FG-squircle representation in Maple™ 4.00**

This implicit equation of the square will be useful to us to study the isomorphism of a circle and a square in the section of Topology.

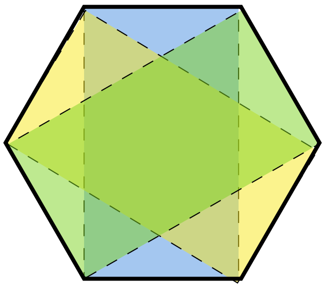
For the fun we can also define a square of side $2a$, having four straight sides lying along the lines $x=\pm a$ and $y=\pm a$ by a piecewise function using trigonometry as:

$$f(\theta)=\begin{cases}(a,a\tan(\theta)), & -\frac{\pi}{4}\leq\theta\leq\frac{\pi}{4}\\(a\cot(\theta),a), & \frac{\pi}{4}\leq\theta\leq\frac{3\pi}{4}\\(a,-a\tan(\theta)), & \frac{3\pi}{4}\leq\theta\leq\frac{5\pi}{4}\\(-a\cot(\theta),a), & \frac{5\pi}{4}\leq\theta\leq\frac{7\pi}{4}\end{cases}\qquad(8.5.226)$$

8.5.2.5 Equation of an Hexagon

A regular hexagon centered at the origin $(0,0)$ with a distance from center to side (apothem) of $\sqrt{3}/2$ (which corresponds to a side length $s=1$) can be defined by the intersection of three rectangular regions:

- $|y|\leq\frac{\sqrt{3}}{2}$ (top and bottom flat edges)
- $\left|\frac{\sqrt{3}}{2}x+\frac{1}{2}y\right|\leq\frac{\sqrt{3}}{2}$ (slanted edges)
- $\left|\frac{\sqrt{3}}{2}x-\frac{1}{2}y\right|\leq\frac{\sqrt{3}}{2}$ (slanted edges)



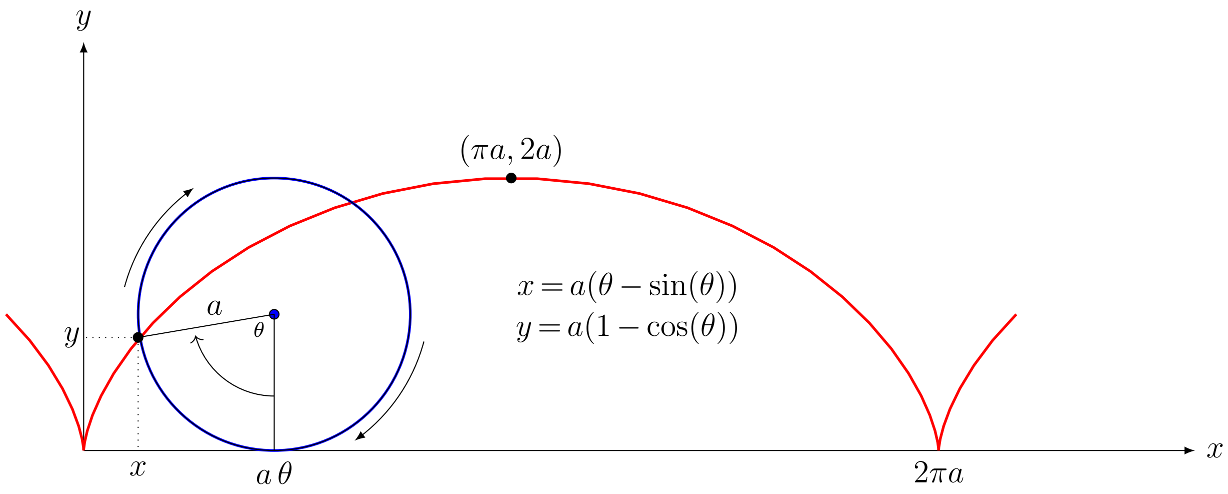
Combined into a single inequality using absolute values, the interior of the hexagon is given by the following piecewise equation:

$$\boxed{\max\left(|y|,\left|\frac{\sqrt{3}}{2}x+\frac{1}{2}y\right|,\left|\frac{\sqrt{3}}{2}x-\frac{1}{2}y\right|\right)\leq\frac{\sqrt{3}}{2}} \qquad (8.5.227)$$

8.5.2.6 Equation of a Cycloid

A cycloid is the curve traced by a point on the rim of a circular wheel as the wheel rolls along a straight line without slippage. It is an example of a roulette, a curve generated by a curve rolling on another curve.

The cycloid is an important curve in physics! Indeed as we will prove it in the section of Classical Mechanics, the cycloid, with the cusps pointing upward, is the curve of fastest descent under constant gravity, and is also the form of a curve for which the period of an object in descent on the curve does not depend on the object's starting position. It is also a mathematical curve on which we will fall back during our detailed study of Friedmann-Lemaître-Robertson-Walker Universe Models in the section of Cosmology (based on the mathematics of General Relativity)!



We see on the figure above that as an arc θ of the circle of radius a has a length:

$$2\pi a\frac{\theta}{2\pi}=a\theta \qquad (8.5.228)$$

This corresponds to the abscissa distance in the range $[0,2\pi a]$.

But to get the x value of the tracked point we must subtract the horizontal projection a' of the radius a on the abscissa and as:

$$\sin(\theta)=\frac{a'}{a}\Leftrightarrow a'=a\sin(\theta) \qquad (8.5.229)$$

Therefore:

$$x=a\theta-a\sin(\theta)=a(\theta-\sin(\theta)) \tag{8.5.230}$$

And the same procedure applies for y.

Remark

Notice that it means if θ do a whole round ($\theta=2\pi$), then the horizontal travelled distance will be $2\pi a$.

So we finally get:

$$\boxed{\begin{aligned}x&=a\left(\theta-\sin(\theta)\right)\\ y&=a\left(1-\cos(\theta)\right)\end{aligned}} \tag{8.5.231}$$

Notice that the complete arc length L of the cycloid is given by:

$$L=\int\limits_0^{2\pi}\sqrt{\left(\frac{\mathrm{d}x}{\mathrm{d}\theta}\right)^2+\left(\frac{\mathrm{d}y}{\mathrm{d}\theta}\right)^2}\mathrm{d}\theta=\int\limits_0^{2\pi}a\sqrt{2-2\cos(\theta)}\mathrm{d}\theta=2a\int\limits_0^{2\pi}\sin\left(\frac{\theta}{2}\right)\mathrm{d}\theta=8a \tag{8.5.232}$$

This analysis explains by the why the quite famous "Aristotle's wheel paradox": A wheel is depicted in two-dimensional space as two circles. Its larger, outer circle is tangential to a horizontal surface (e.g. a road that it rolls on), while the smaller, inner one has the same center and is rigidly affixed to the larger. (The smaller circle could be the bead of a tire, the rim it is mounted upon, or the axle.) Assuming the larger circle rolls without slipping (or skidding) for one full revolution, the distances moved by both circles' circumferences are the same. The distance travelled by the larger circle is equal to its circumference, but for the smaller it is greater than its circumference, thereby creating a paradox. But in reality the paradox is easily explained by the fact that the real distance travelled is not the horizontal one but the one described by the cycloid and the inner point has a shorter cycloid path than the external point!

8.5.2.7 Equation of an Epicycloid

In geometry, an epicycloid is a plane curve produced by tracing the path of a chosen point on the circumference of a circle - named an "epicycle" - which rolls without slipping around a fixed circle. It is a particular kind of roulette.

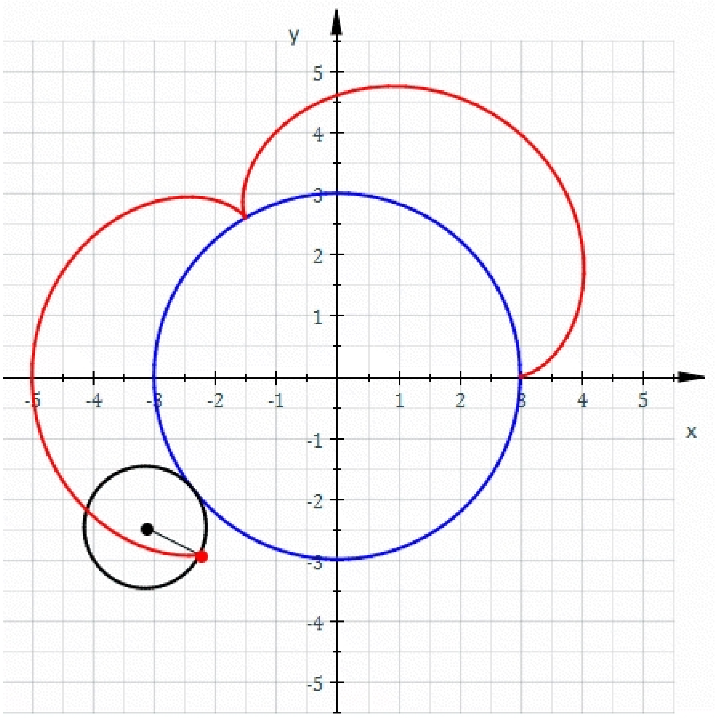


Figure 8.167: Epicycloid path (source: Wikipedia, author: Sam Derbyshire)

As we will see in the section of Mechanical Engineering page 8376, the epicycloid is used to build efficient energy gearing geometry shapes!

For the need of the section of Mechanical Engineering we need to determine the parametric equation of the epicycloid. For this consider the following figure:

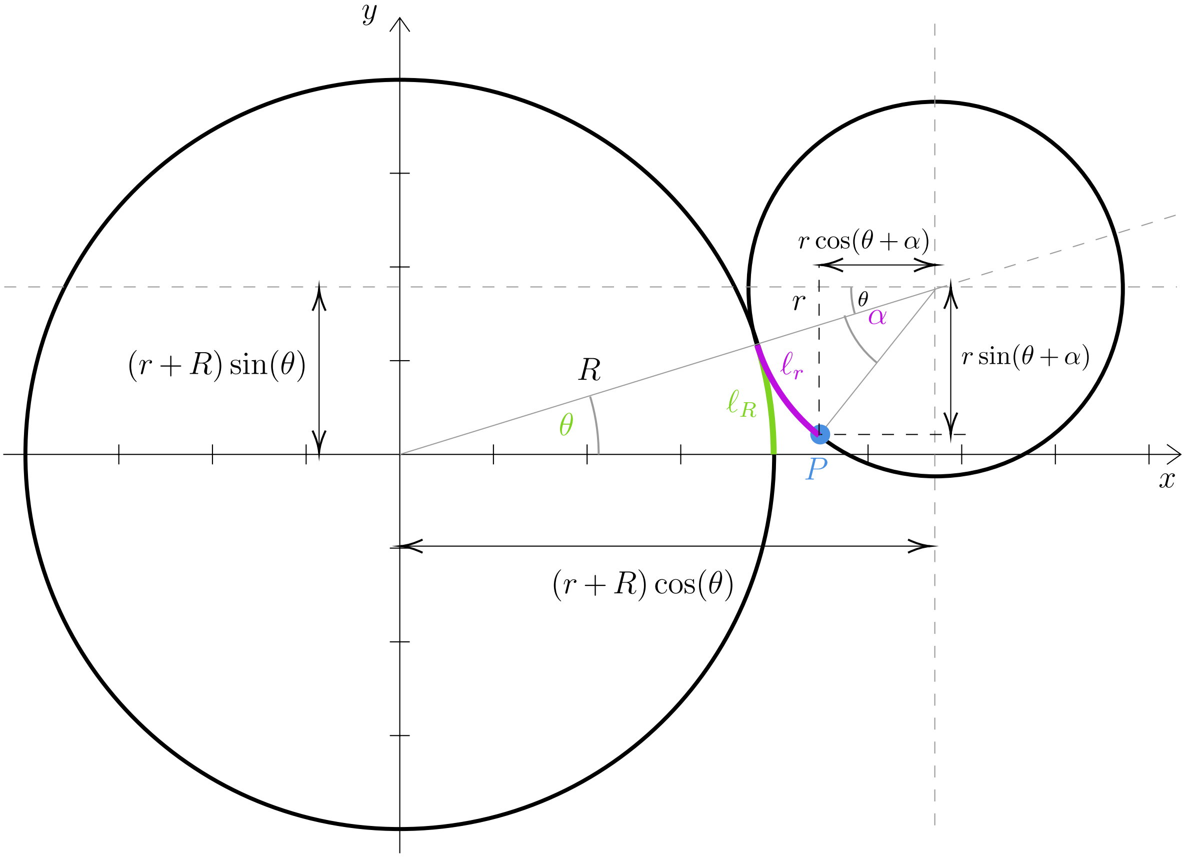


Figure 8.168: **Epicycloid sketch for proof (source: Wikipedia, Sam Derbyshire)**

We assume that the position of P is what we want to solve, α is the radian from the tangential point to the moving point P, and θ is the radian from the starting point to the tangential point.

Since there is no sliding between the two cycles, then we have that:

$$\ell_R=\ell_r \tag{8.5.233}$$

By the definition of radian (which is the rate arc over radius), then we have that:

$$\ell_R=\theta R \qquad \ell_r=\alpha r \tag{8.5.234}$$

From these two conditions, we get the identity:

$$\theta R=\alpha r \tag{8.5.235}$$

By rearranging, we get the relation between α and θ, which is:

$$\alpha=\frac{R}{r}\theta \tag{8.5.236}$$

From the figure, we see the position of the point P is clearly given by:

$$\boxed{\begin{aligned}x&=(R+r)\cos\theta-r\cos(\theta+\alpha)=(R+r)\cos\theta-r\cos\left(\frac{R+r}{r}\theta\right)\\y&=(R+r)\sin\theta-r\sin(\theta+\alpha)=(R+r)\sin\theta-r\sin\left(\frac{R+r}{r}\theta\right)\end{aligned}} \tag{8.5.237}$$

8.5.2.8 Equation of a Spiral

A "spiral" is a curve in the plane or in the space, which runs around a center in a special way. It sometimes a useful shape in some industrial fields (robotics, watchmaking, texts designs, etc.). We can see Archimedean Spirals in the spring mechanism of clocks and luxury watches but they study is also useful for some calculations related to galaxies or to atmospheric re-entry of passive space vehicles.

We can make a spiral by two motions of a point: There is a uniform motion in a fixed direction and a motion in a circle with constant speed. Both motions start at the same point. This construction is named the "Archimedean Spiral" (also known as the "arithmetic spiral").

We get the Archimedean Spiral by analogy to the circle parametric equations. Let us recall that the parametric equation of a circle is given by:

$$\begin{aligned} x &= R\cos(\theta) \\ y &= R\sin(\theta) \end{aligned} \tag{8.5.238}$$

Let us rewrite it as following:

$$\begin{aligned} x &= R\cos(t) \\ y &= R\sin(t) \end{aligned} \tag{8.5.239}$$

where a is a chosen constant.

Now for the Archimedean Spiral, the idea is to make R also proportional of t. Therefore the parametric equation for an Archimedean spiral centered at the origin is:

$$\boxed{\begin{aligned} x &= R(t)\cos(t) = at\cos(t) \\ y &= R(t)\sin(t) = at\sin(t) \end{aligned}} \tag{8.5.240}$$

Below is an example of an Archimedean Spiral with in polar coordinate $r = a\theta$:

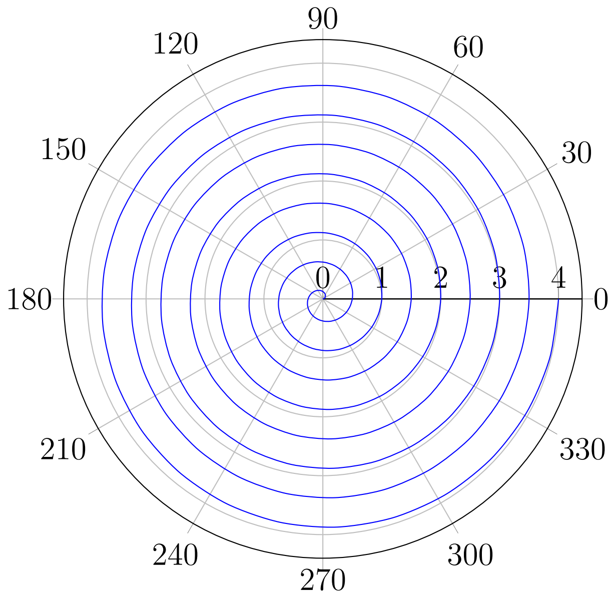


Figure 8.169: **Archimedean Spiral**

The distances between the spiral branches are the same. More exactly: The distances of intersection points along a line through the origin are the same.

Obviously there are plenty of other spirals like the exponential one illustrated here below:

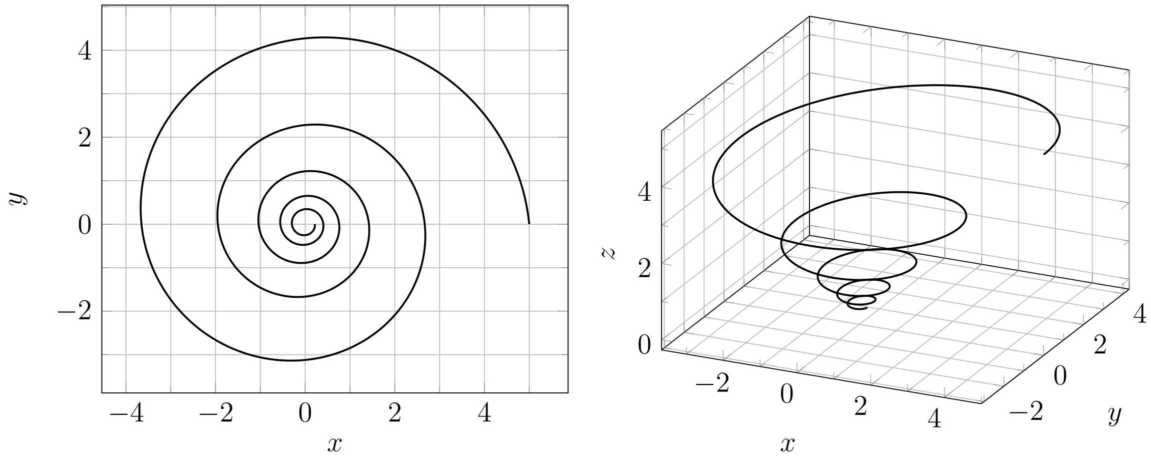


Figure 8.170: **Exponential spiral in 2D and 3D of the type $r_0 e^{bt}$**

Remark

Contrary to popular belief (and belief is not science for reminder!), a space probe does not complete the final stage of its atmospheric entry vertically over a distance of ~ 120 kilometers, nor in a spiral (although it does spiral before entering the upper layers of the atmosphere!); instead, the process typically takes an average of ~ 10 minutes and covers a distance of $\sim 4,000$ kilometers, corresponding to an average deceleration of approximately $2g$. Here is the typical profile of an unassisted atmospheric entry:

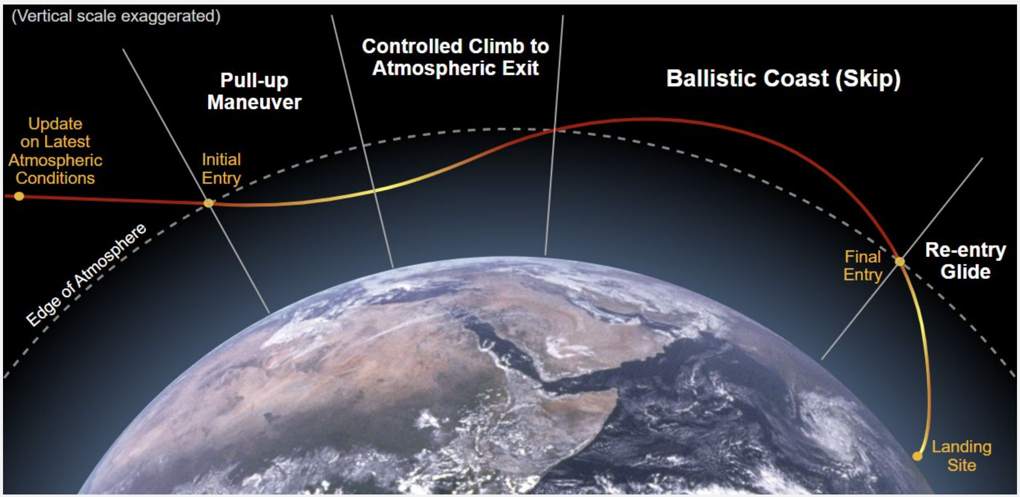
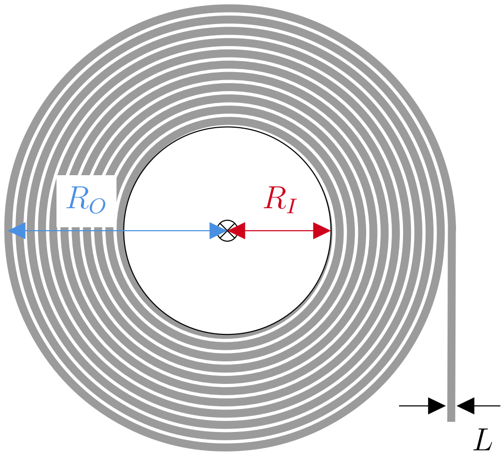

(Vertical scale exaggerated)
Update on Latest Atmospheric Conditions
Pull-up Maneuver
Initial Entry
Controlled Climb to Atmospheric Exit
Ballistic Coast (Skip)
Edge of Atmosphere
Final Entry
Re-entry Glide
Landing Site

Figure 8.171: **Boost-glide atmospheric re-entry (source: Wikipedia, author: Clem Tillier)**

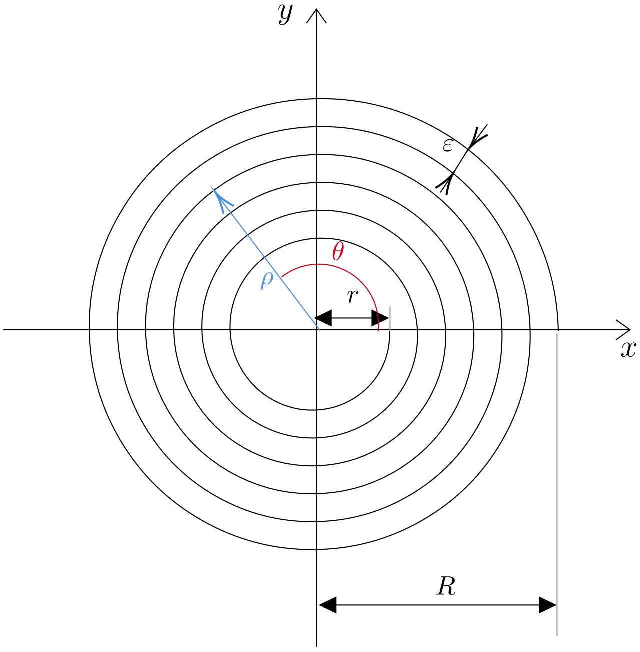
Regarding the well-known case of the length of a coil (very useful in electromagnetism):



We simply have:

$$L \cong \frac{\pi\left(R_O^2 - R_I^2\right)}{T} \tag{8.5.241}$$

And for the more realistic case of a spiral:



We obviously have:

$$\begin{gathered}\rho = r + \varepsilon\frac{\theta}{2\pi}\\ R = r + n\varepsilon\end{gathered} \tag{8.5.242}$$

and:

$$\begin{cases} x = \rho\cos(\theta)\\ y = \rho\sin(\theta)\end{cases} \tag{8.5.243}$$

Hence:

$$\left\{\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}\theta}&=\frac{\varepsilon}{2\pi}\cos(\theta)-\left(r+\varepsilon\frac{\theta}{2\pi}\right)\sin(\theta)\\\frac{\mathrm{d}y}{\mathrm{d}\theta}&=\frac{\varepsilon}{2\pi}\sin(\theta)+\left(r+\varepsilon\frac{\theta}{2\pi}\right)\cos(\theta)\end{aligned}\right.\tag{8.5.244}$$

And this gives us:

$$\frac{\mathrm{d}L}{\mathrm{d}\theta}=\sqrt{\left(\frac{\mathrm{d}x}{\mathrm{d}\theta}\right)^2+\left(\frac{\mathrm{d}y}{\mathrm{d}\theta}\right)^2}=\sqrt{\left(\frac{\varepsilon}{2\pi}\right)^2+\left(r+\varepsilon\frac{\theta}{2\pi}\right)^2}=\frac{\varepsilon}{2\pi}\sqrt{1+\left(2\pi\frac{r}{\varepsilon}+\theta\right)^2}\tag{8.5.245}$$

Therefore:

$$\begin{aligned}L&=\int\limits_{\theta=0}^{\theta=2\pi n}\left(\frac{\mathrm{d}L}{\mathrm{d}\theta}\right)\mathrm{d}\theta=\frac{\varepsilon}{2\pi}\int\limits_{0}^{2\pi n}\sqrt{1+\left(2\pi\frac{r}{\varepsilon}+\theta\right)^2}\mathrm{d}\theta\\&=\frac{\varepsilon}{4\pi}\left(\left(2\pi\frac{r}{\varepsilon}+\theta\right)\sqrt{1+\left(2\pi\frac{r}{\varepsilon}+\theta\right)^2}+\ln\left(2\pi\frac{r}{\varepsilon}+\theta+\sqrt{1+\left(2\pi\frac{r}{\varepsilon}+\theta\right)^2}\right)\right)_{\theta=0}^{\theta=2\pi n}\\&=\frac{\varepsilon}{4\pi}\left(\left(2\pi\frac{r}{\varepsilon}+2\pi n\right)\sqrt{1+\left(2\pi\frac{r}{\varepsilon}+2\pi n\right)^2}-\left(2\pi\frac{r}{\varepsilon}\right)\sqrt{1+\left(2\pi\frac{r}{\varepsilon}\right)^2}+\ln\left(\frac{2\pi\frac{r}{\varepsilon}+2\pi n+\sqrt{1+\left(2\pi\frac{r}{\varepsilon}+2\pi n\right)^2}}{2\pi\frac{r}{\varepsilon}+\sqrt{1+\left(2\pi\frac{r}{\varepsilon}\right)^2}}\right)\right)\end{aligned}\tag{8.5.246}$$

and as $R=r+n\varepsilon$, this last relation simplifies to:

$$\begin{aligned}L&=\frac{\varepsilon}{4\pi}\left(2\pi\frac{R}{\varepsilon}\sqrt{1+\left(2\pi\frac{R}{\varepsilon}\right)^2}-\left(2\pi\frac{r}{\varepsilon}\right)\sqrt{1+\left(2\pi\frac{r}{\varepsilon}\right)^2}+\ln\left(\frac{2\pi\frac{R}{\varepsilon}+\sqrt{1+\left(2\pi\frac{R}{\varepsilon}\right)^2}}{2\pi\frac{r}{\varepsilon}+\sqrt{1+\left(2\pi\frac{r}{\varepsilon}\right)^2}}\right)\right)\\&=\pi\frac{R^2}{\varepsilon}\sqrt{1+\left(\frac{\varepsilon}{2\pi R}\right)^2}-\pi\frac{r^2}{\varepsilon}\sqrt{1+\left(\frac{\varepsilon}{2\pi r}\right)^2}+\frac{\varepsilon}{2\pi}\ln\left(\frac{R}{r}\right)+\frac{\varepsilon}{4\pi}\ln\left(\frac{1+\left(\frac{\varepsilon}{2\pi R}\right)^2\sqrt{1+\left(\frac{\varepsilon}{2\pi R}\right)^2}}{1+\left(\frac{\varepsilon}{2\pi r}\right)^2\sqrt{1+\left(\frac{\varepsilon}{2\pi r}\right)^2}}\right)\end{aligned}\tag{8.5.247}$$

If ε is small compared to r and R, let us see that we fall back on:

$$L\cong\pi\frac{R^2}{\varepsilon}-\pi\frac{r^2}{\varepsilon}=\frac{\pi(R^2-r^2)}{\varepsilon}\tag{8.5.248}$$

8.5.2.9 Equation of a Hypocycloid

In geometry, a hypocycloid is a special plane curve generated by the trace of a fixed point on a small circle that rolls within a larger circle. It is comparable to the cycloid or epicycloid but instead of the circle rolling along a line, it rolls within a circle.

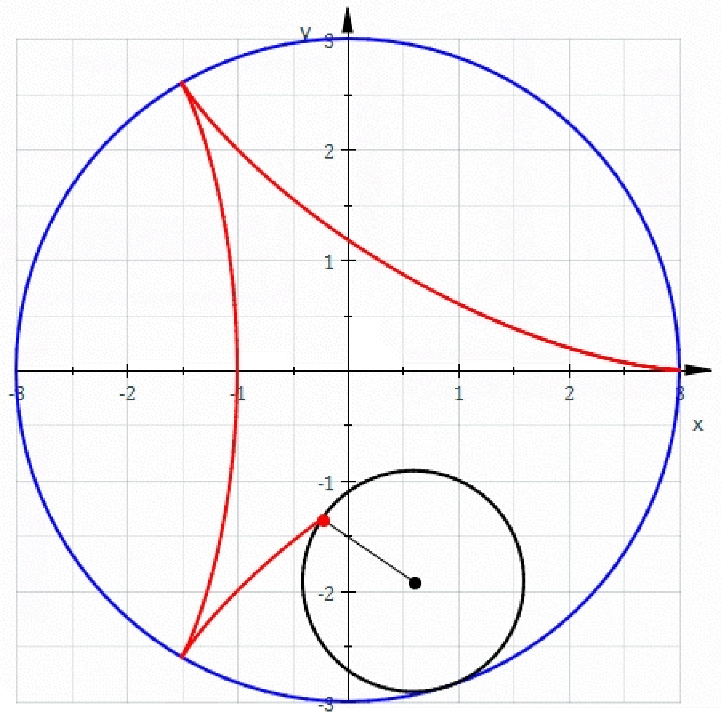
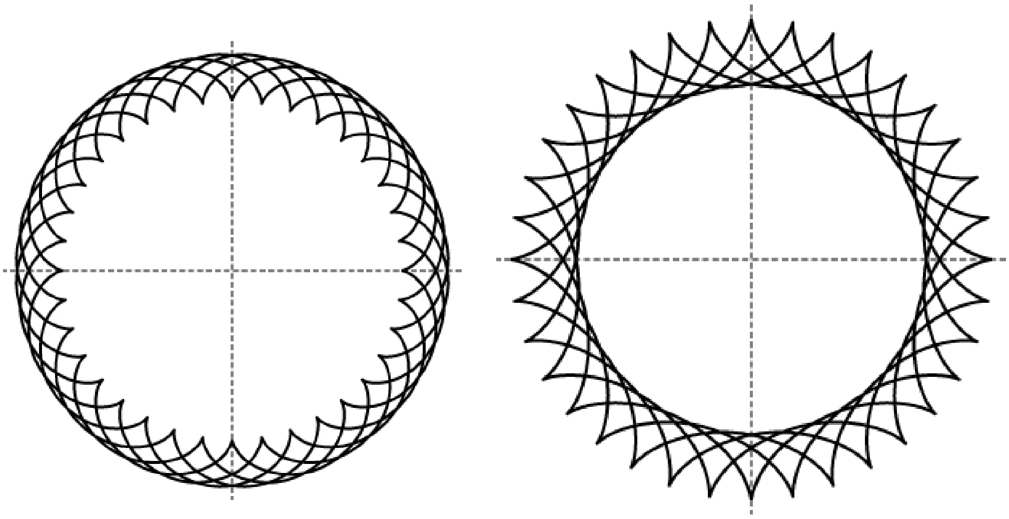


Figure 8.172: **Hypocycloid path (source: Wikipedia, Sam Derbyshire)**

When we know the parametric equation of the epicycloid, that of the hypocycloid is immediate:

$$\boxed{\begin{aligned} x &= (R-r)\cos\theta + r\cos\left(\frac{R-r}{r}\theta\right) \\ y &= (R-r)\sin\theta + r\sin\left(\frac{R-r}{r}\theta\right) \end{aligned}} \tag{8.5.249}$$

Now for the section of Mechanical Engineering it will help if the reader compare side by side and epicycloid and a hypocycloid one:



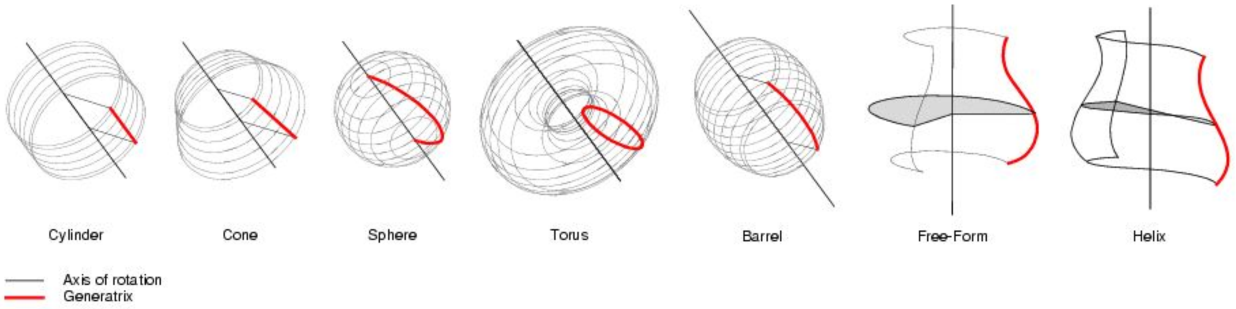
(a) **Epicycloid (source: Wikipedia)**

(b) **Hipocycloid**

8.5.2.10 Surface of revolution

More generally many surfaces (and also some of which we've seen before as the sphere, tore and cylinder) can be described by revolving a primary form of smaller size and then by rotation.

Definition (#8.5.357): A "surface of revolution" is a surface obtained by rotating a plane curve (e.g. $z=f(x)$), named the "generatrix", around the z-axis (for example!). So we pass from a plane of $\mathbb{R}^2$ to a basis in $\mathbb{R}^3$, the x-axis then generates a plane that has become the yOz plane.


Cylinder
Cone
Sphere
Torus
Barrel
Free-Form
Helix
Axis of rotation
Generatrix

Let us see seven (cone, sphere, ellipsoid, cylinder, paraboloid, hyperboloid, torus) famous cases:

8.5.2.10.1 Cone

Consider a circular cone of vertex O (origin) with a circle of center $(0,0,h)$ (where h is a positive real number corresponding to the height of the cone) and radius R:

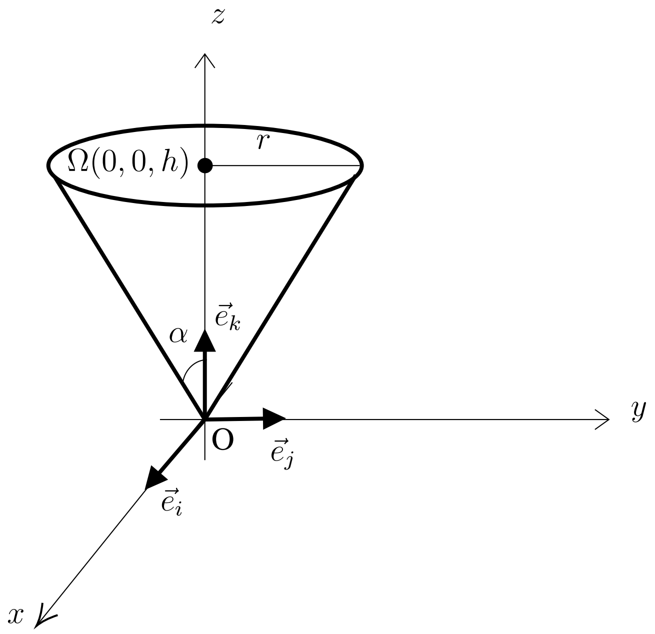

z
r
$\Omega(0,0,h)$
α
$\vec{e}_k$
y
O
$\vec{e}_j$
$\vec{e}_i$
x

Figure 8.174: **Representation of a cone**

A parametric representation of the circle at the height h is given as we already proved earlier by:

$$\begin{cases} x(t)=r\cos(t) \\ y(t)=r\sin(t) \\ z(t)=h \end{cases} \tag{8.5.250}$$

where t belongs to the interval $[-\pi,\pi]$.

By extension, the parametric representation of a cone is:

$$\begin{cases} x(t,k)=kr\cos(t) \\ y(t,k)=kr\sin(t) \\ z(t,k)=kh \end{cases} \tag{8.5.251}$$

this can also be written in vector form (more traditional):

$$\boxed{\begin{pmatrix} x(t,k) \\ y(t,k) \\ z(t,k) \end{pmatrix} = \begin{pmatrix} kr\cos(t) \\ kr\sin(t) \\ kh \end{pmatrix}} \tag{8.5.252}$$

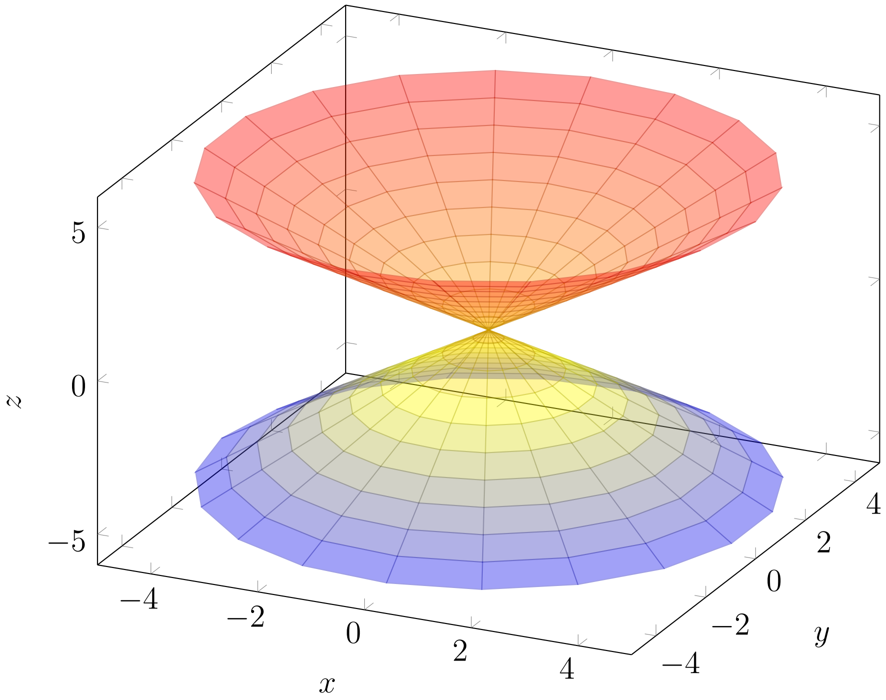


Figure 8.175: **High definition Cone made with LaTeX/Tikz**

In other words, the circle linearly propagates in all directions.

✍**Example:**

With Maple™ 4.00b:

```
>r:=1:h:=4:
>plot3d([k*r*cos(t),k*r*sin(t),k*h],k=0..10,t=0..2*Pi,
orientation=[50,60],style=PATCH,axes=NORMAL);
```

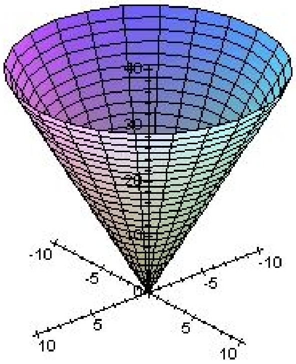


Figure 8.176: **Parametric representation of a cone in Maple™ 4.00b**

This suggests also the following cartesian equation of this cone:

$$x^2+y^2=k^2r^2=\left(\frac{z}{h}\right)^2r^2 \qquad (8.5.253)$$

As we have $\tan(\alpha)=r/h$ we can write:

$$x^2+y^2=\left(\frac{z}{h}\right)^2r^2=z^2\tan^2(\alpha)=0 \qquad (8.5.254)$$

Therefore:

$$\boxed{x^2+y^2-z^2\tan^2(\alpha)=0} \qquad (8.5.255)$$

which is the Cartesian equation of a cone in space that we will see again in the section of Special Relativity in our study of light cones.

8.5.2.10.2 Sphere

Consider the orthonormal basis $\{\vec{i},\vec{j},\vec{k}\}$. Given the sphere $\mathcal{S}^2$ of center $\Omega(a,b,c)$ and of radius r:

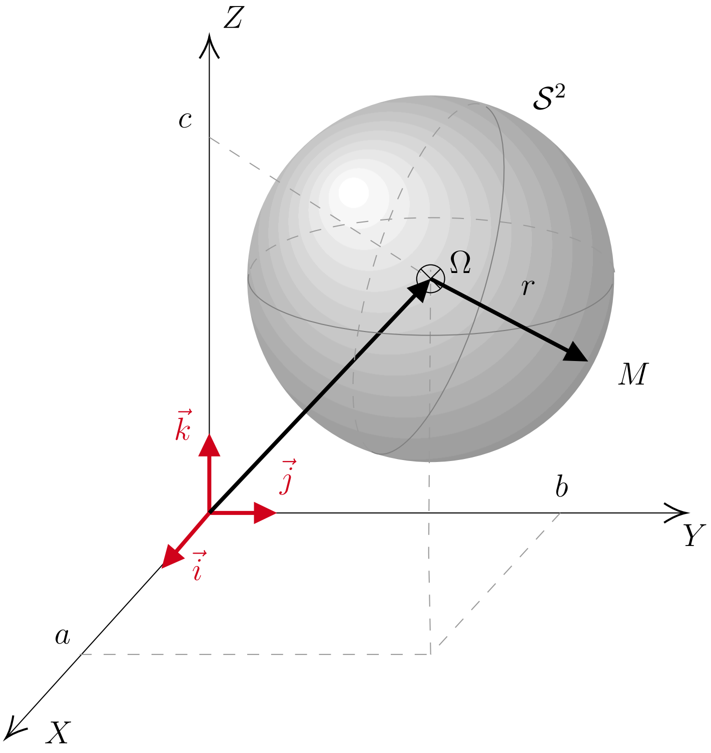


Figure 8.177: Representation of a $\mathcal{S}^2$ sphere

$M(x,y,z)$ belongs to the sphere S^2 of radius r if and only if:

$$\overline{\Omega M} = r \tag{8.5.256}$$

that is to say by applying Pythagoras theorem:

$$\overline{\Omega M} = r^2 = (x-a)^2 + (y-b)^2 + (z-c)^2 \tag{8.5.257}$$

Hence the "Cartesian equation of the sphere" in the basis $\{\vec{i},\vec{j},\vec{k}\}$:

$$\boxed{r^2 = (x-a)^2 + (y-b)^2 + (z-c)^2 = r^2} \tag{8.5.258}$$

There is another way to describe the sphere using the parametric equation. Indeed, we saw in the section of Vector Calculus that the transformation of the Cartesian coordinates to spherical coordinates is given by the following curvilinear coordinates:

$$x = r\sin(\theta)\cos(\phi),\; y = r\sin(\theta)\sin(\phi),\; z = r\cos(\theta) \tag{8.5.259}$$

Therefore we have well (by a change of variable this also represent a sphere not centered at the origin):

$$\begin{aligned} x^2+y^2+z^2 &= r^2\sin^2(\theta)\cos^2(\phi) + r^2\sin^2(\theta)\sin^2(\phi) + r^2\cos^2(\theta) \\ &= r^2\sin^2(\theta)(\cos^2(\phi)+\sin^2(\phi)) + r^2\cos^2(\theta) = r^2\sin^2(\theta) + r^2\cos^2(\theta) \\ &= r^2(\sin^2(\theta)+\cos^2(\theta)) = r^2 \end{aligned} \tag{8.5.260}$$

So the parametric equation of the sphere is indeed:

$$\boxed{\begin{pmatrix} x(\theta,\phi) \\ y(\theta,\phi) \\ z(\theta,\phi) \end{pmatrix} = \begin{pmatrix} r\sin(\theta)\cos(\phi) \\ r\sin(\theta)\sin(\phi) \\ r\cos(\theta) \end{pmatrix}} \tag{8.5.261}$$

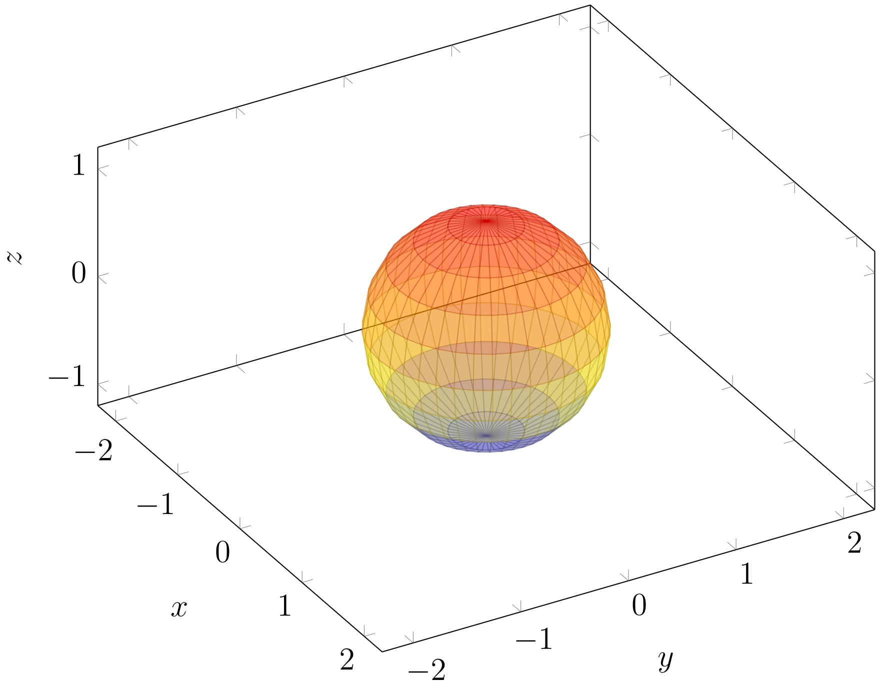


Figure 8.178: **High definition Sphere made with LaTeX/Tikz**

We fall back therefore on the Cartesian equation of a sphere to a given translation constant.

Example:

With Maple™ 4.00b for a sphere of unit radius:

```
>r:=1:h:=4:
>plot3d([sin(theta)*cos(phi),sin(theta)*sin(phi),cos(theta)],
theta=0..Pi,phi=-Pi..Pi,scaling=CONSTRAINED,orientation=[50,60],
style=PATCH,axes=NORMAL);
```

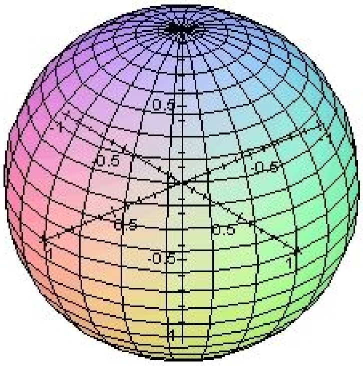


Figure 8.179: **Parametric representation of a sphere in Maple™ 4.00b**

8.5.2.10.2.1 Great circle of sphere

We will introduce here a concept that will be useful to us when we will study the geodesic of a sphere in the section of Analytical Mechanics.

Definition (#8.5.358): A "great circle", also known as an "orthodrome" or "Riemannian circle", of a sphere is the intersection of the sphere and a plane that passes through the center point of the sphere. This partial case of a circle of a sphere is opposed to a "small circle", the intersection of the sphere and a plane that does not pass through the center. Any diameter of any great circle coincides with a diameter of the sphere, and therefore all great circles have the same circumference as each other, and have the same center as the sphere. A great circle is the largest circle that can be drawn on any given sphere. Every circle in Euclidean 3-space is a great circle of exactly one sphere.

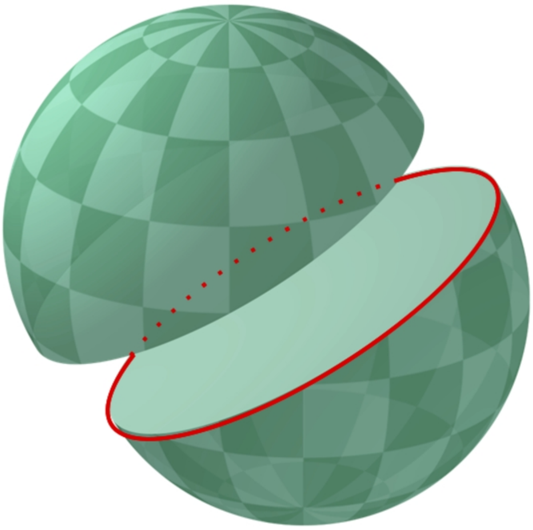


Figure 8.180: **A great circle divides the sphere in two equal hemispheres**

Let use choose now a unit normal vector to a plane passing through the origin using Euler angles be:

$$\vec{n}=\begin{pmatrix}-\sin(\gamma)\\0\\\cos(\gamma)\end{pmatrix} \tag{8.5.262}$$

where for rotated our such that the component n_2 is equal to 0 (as it is always possible to do so whatever the plan we choose that pass through the origin!).

If we choose a sphere of radius $r=1$ such that it's parametric equation is written as:

$$\vec{\mathcal{S}}^2=\begin{pmatrix}r\sin(\theta)\cos(\phi)\\r\sin(\theta)\sin(\phi)\\r\cos(\theta)\end{pmatrix}=\begin{pmatrix}\sin(\theta)\cos(\phi)\\\sin(\theta)\sin(\phi)\\\cos(\theta)\end{pmatrix} \tag{8.5.263}$$

And as the set of points that are at the intersection of the plane and the sphere are by construction such that:

$$\vec{\mathcal{S}}^2\circ\vec{n}=0 \tag{8.5.264}$$

Therefore we get:

$$-\sin(\gamma)\sin(\theta)\cos(\phi)+0\cdot\sin(\theta)\sin(\phi)+\cos(\gamma)\cos(\theta)=0 \tag{8.5.265}$$

After rearranging:

$$\frac{\sin(\gamma)}{\cos(\gamma)}\cos(\phi)=\frac{\cos(\theta)}{\sin(\theta)} \tag{8.5.266}$$

Let us put:

$$a=\frac{\sin(\gamma)}{\cos(\gamma)} \tag{8.5.267}$$

and using the definition of the cotangent:

$$\boxed{a\cos(\phi)=\cot(\theta)} \tag{8.5.268}$$

This is therefore the "equation of a great circle".

8.5.2.10.2.2 n-Sphere volume

Before we deal with the subject, we want to thanks professor Howard Haber that has provided us the major LaTeXsource code of the text below that will be useful to us in the sections of Statistical Mechanics (especially to derive the Sackur-Tetrode equation) and of General Relativity!

An n-dimensional hypersphere of radius R consists of the locus of points such that the distance from the origin is less than or equal to R.

A point in an n-dimensional Euclidean space is designated by $(x_1\,,\,x_2\,,\,\ldots\,,\,x_n)$.

In equation form, the hypersphere corresponds then by definition to the set of points such that:

$$x_1^2+x_2^2+\cdots+x_n^2\leq R^2 \tag{8.5.269}$$

To compute the volume of this hypersphere, we simply integrate the infinitesimal volume element $\mathrm{d}V=\mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_n$ over the region of n-dimensional space indicated by 8.5.269. We wish to compute this volume $V_n(R)$. Explicitly:

$$V_n(R)=\int\ldots\int_{x_1^2+x_2^2+\cdots+x_n^2\leq R^2}\mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_n=C_nR^n \tag{8.5.270}$$

The factor of R^n is a consequence of dimensional analysis.

The surface-area of the n-dimensional hypersphere will be denoted in what follows by $S_{n-1}(R)$.

The surface of the hypersphere corresponds to the locus of points such that:

$$x_1^2+x_2^2+\cdots+x_n^2=R^2 \tag{8.5.271}$$

We can construct the volume $V_n(R)$ by adding infinitely thin spherical shells of radius $0 \leq r \leq R$. In equation form, this reads:

$$V_n(R) = \int_0^R S_{n-1}(r)\mathrm{d}r \tag{8.5.272}$$

And as we know:

$$S_{n-1}(R) = \frac{\mathrm{d}V_n(R)}{\mathrm{d}R} = nC_nR^{n-1} \tag{8.5.273}$$

where we have used 8.5.270. Thus, the only remaining task is to compute C_n.

In order to obtain a better intuition on the meaning of C_n, let us equate the two expressions we have for $V_n(R)$, namely 8.5.270 and 8.5.271.

In the latter, $S_{n-1}(r)$ is determined by 8.5.273. Thus:

$$V_n(R) = \int\!\cdots\!\int_{x_1^2+x_2^2+\cdots+x_n^2 \leq R^2} \mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_n = nC_n\int_0^R r^{n-1}\,\mathrm{d}r \tag{8.5.274}$$

This last relation is simply the evaluation of an n-dimensional integral either in rectangular co-ordinates or hyper-spherical co-ordinates. In n-dimensions, we would write:

$$\mathrm{d}x_1\,\mathrm{d}x_2\cdots\mathrm{d}x_n = r^{n-1}\,\mathrm{d}r\,\mathrm{d}\Omega_{n-1} \tag{8.5.275}$$

where $\mathrm{d}\Omega_{n-1}$ contains all of the angular factors. For example, for $n=2$, $\mathrm{d}\Omega_1 = \mathrm{d}\theta$; for $n=3$, $\mathrm{d}\Omega_2 = \sin(\theta)\,\mathrm{d}\theta\,\mathrm{d}\phi$, etc.

One could explicitly define the $n-1$ angular variables in n-dimensions.

However, if we are integrating over a spherically symmetric function, then we will simply integrate over $\mathrm{d}\Omega_{n-1}$. Comparing 8.5.274 and 8.5.275, we see that:

$$\int\!\cdots\!\int \mathrm{d}\Omega_{n-1} = nC_n \tag{8.5.276}$$

We will present a trick for computing C_n without ever explicitly parametrizing the angles of the hyper-spherical coordinate system. For this purpose, consider the function:

$$f(x_1\,,\,x_2\,,\,\ldots\,,\,x_n) = e^{-(x_1^2+x_2^2+\cdots+x_n^2)} = e^{-r^2} \tag{8.5.277}$$

If we integrate this function over the full n-dimensional space in both rectangular and hyper-spherical coordinates, we obtain:

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} e^{-(x_1^2+x_2^2+\cdots+x_n^2)}\mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_n = \int_0^{+\infty} r^{n-1}\,\mathrm{d}r\int \mathrm{d}\Omega_{n-1}e^{-r^2} \tag{8.5.278}$$

Since the integrand on the right hand side depends only on r (there is no angular dependence), we can immediately perform the integral over $\mathrm{d}\Omega_{n-1}$. Using 8.5.276:

$$\int_{-\infty}^{+\infty} e^{-x_1^2}\,\mathrm{d}x_1\int_{-\infty}^{+\infty} e^{-x_2^2}\mathrm{d}x_2\ldots\int_{-\infty}^{+\infty} e^{-x_n^2}\mathrm{d}x_n = nC_n\int_0^{+\infty} r^{n-1}\,e^{-r^2}\mathrm{d}r \tag{8.5.279}$$

All the integrals above can be evaluated exactly using Gauss integral (see section Statistics page 695) and Euler-Gamma integral (see section Differential and Integral Calculus page 1574):

$$\begin{gathered}\int\limits_{-\infty}^{+\infty} e^{-x^2}\,\mathrm{d}x=\sqrt{\pi}\\ \int\limits_{0}^{+\infty} r^{n-1}\,e^{-r^2}\,\mathrm{d}r=\int\limits_{0}^{+\infty} e^{-t}t^{\frac{n-1}{2}}\left(\frac{1}{2}t^{-1/2}\mathrm{d}t\right)=\frac{1}{2}\Gamma\left(\frac{n}{2}\right)\end{gathered}\tag{8.5.280}$$

Inserting these results into 8.5.279, we get :

$$\pi^{n/2}=C_n\,\frac{n}{2}\,\Gamma\left(\frac{n}{2}\right)=C_n\,\Gamma\left(1+\frac{n}{2}\right)\tag{8.5.281}$$

where we used the property of the Gamma function, $x\Gamma(x)=\Gamma(x+1)$, at the final step. Solving for C_n, we get:

$$C_n=\frac{\pi^{n/2}}{\Gamma\left(1+\dfrac{n}{2}\right)}\tag{8.5.282}$$

Although we chose a particular function $f(x_1\,,\,x_2\,,\,\ldots\,,\,x_n)$ to get the final result for C_n, it is clear that the value of C_n is independent of this function. After all, the defining equation for C_n as in 8.5.276 makes no reference to any function. As a check, let us evaluate $n\,C_n$ for $n=2$ and $n=3$:

$$\begin{gathered}2\,C_2=\frac{2\pi}{\Gamma(2)}=2\pi=\int\limits_0^{2\pi}\mathrm{d}\theta\\ 3\,C_3=\frac{3\pi^{3/2}}{\Gamma\left(\dfrac{5}{2}\right)}=4\pi=\int\limits_0^{2\pi}\mathrm{d}\phi\int\limits_0^{\pi}\sin\theta\mathrm{d}\theta\end{gathered}\tag{8.5.283}$$

In the last computation, we used $\Gamma\left(\frac{5}{2}\right)=\frac{3}{2}\Gamma\left(\frac{3}{2}\right)=\left(\frac{3}{2}\right)\left(\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right)=\frac{3}{4}\sqrt{\pi}$. Indeed, we have produced the correct values for the integration over the angular factors in two and three dimensions.

We are finally ready to compute the volume and surface-area of an n-dimensional hypersphere. Inserting 8.5.282 into 8.5.270 and 8.5.273, we find:

$$\boxed{\begin{aligned}V_n(R)&=\frac{\pi^{n/2}\,R^n}{\Gamma\left(1+\dfrac{n}{2}\right)}\\ S_{n-1}(R)&=\frac{n\pi^{n/2}\,R^{n-1}}{\Gamma\left(1+\dfrac{n}{2}\right)}\end{aligned}}\tag{8.5.284}$$

The expression for $S_n(R)$ can be slightly simplified by writing $\Gamma\left(1+\dfrac{n}{2}\right)=\dfrac{n}{2}\,\Gamma\left(\dfrac{n}{2}\right)$. This yields:

$$\boxed{S_{n-1}(R)=\frac{2\pi^{n/2}\,R^{n-1}}{\Gamma\left(\dfrac{n}{2}\right)}}\tag{8.5.285}$$

You should check that you get the expected results for $n = 2$ and $n = 3$. For example,$V_3(R) = \frac{4}{3}\pi R^3$ and $S_2(R) = 4\pi R^2$.

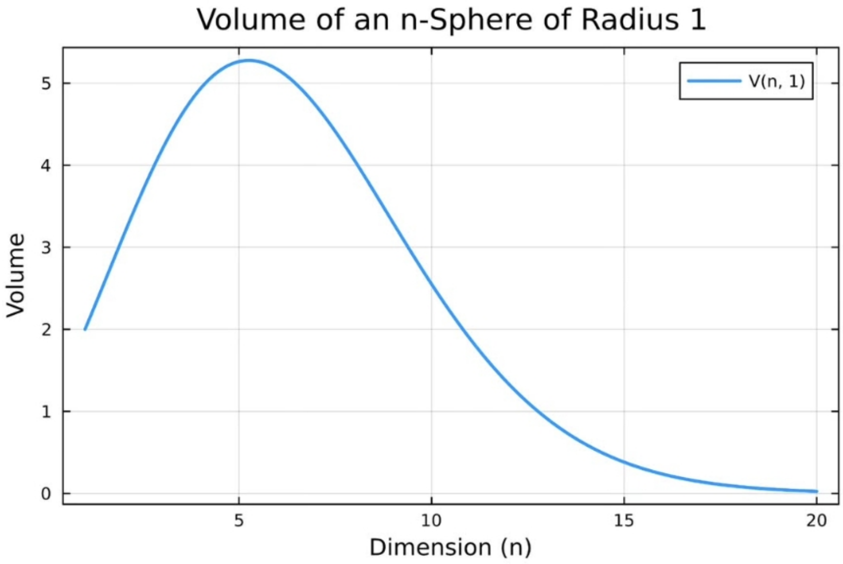
Finally, it is amusing to note that $\lim_{n\rightarrow+\infty} V_n(R) = 0$ and $\lim_{n\rightarrow+\infty} S_n(R) = 0$.

To understand this peculiar behaviour, consider a hypercube in n-dimensional space measuring one unit on each side. The total volume of this hypercube is 1. We can fit a hypersphere of diameter 1 (or radius $\frac{1}{2}$) inside the hypercube such that the surface of the hypersphere just touches each of the walls of the hypercube. Then $1 - V_n(\frac{1}{2})$ is the volume inside the cube but outside the hypersphere.

In particular, as n becomes large, $1 - V_n(\frac{1}{2})$ rapidly approaches 1, which is consistent with the assertion that $\lim_{n\rightarrow+\infty} V_n(R) = 0$. This simply means that as the number of dimensions becomes larger and larger, the amount of space outside the hypersphere (but inside the cube) is become relatively more and more important. This is already happening as you go from 2 to 3 dimensions. So you can check your intuition by inscribing a circle in a unit square and a sphere in a unit cube and computing the total volume in three dimensions (area in two dimensions) outside the sphere (circle) but inside the cube (square). If you take the ratio of volumes (areas) of the sphere (circle) to that of the cube (square), this ratio actually *decreases* as you go from 2 dimensions to 3 dimensions!

This shows that the integer dimension with maximum integer coefficient is $n = 5$.

Dimension	**Volume**	**Volume at $r = 1$**
2	πr^2	3.14159
3	$4/3\pi r^3$	4.18879
4	$1/2\pi^2 r^4$	4.93480
5	$8/15\pi^2 r^5$	5.26379
6	$1/6\pi^3 r^6$	5.16771
7	$16/105\pi^3 r^7$	4.72477
8	$1/24\pi^4 r^8$	4.05871
9	$32/945\pi^4 r^9$	3.29851
10	$1/120\pi^5 r^{10}$	2.55016


Volume of an n-Sphere of Radius 1
V(n, 1)
Volume
0
1
2
3
4
5
5
10
15
20
Dimension (n)

8.5.2.10.3 Ellipsoid (spheroid)

We saw in our study of conic that the cartesian equation of an ellipse in the plane was given by:

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1 \tag{8.5.286}$$

with a, b the two axes of the ellipse (small and large one).

Thus, without rigorous proof, we can verify by hand or using computers that the Cartesian equation:

$$\boxed{\frac{x^2}{a^2}+\frac{y^2}{b^2}+\frac{z^2}{b^2}=1} \tag{8.5.287}$$

is an ellipsoid:

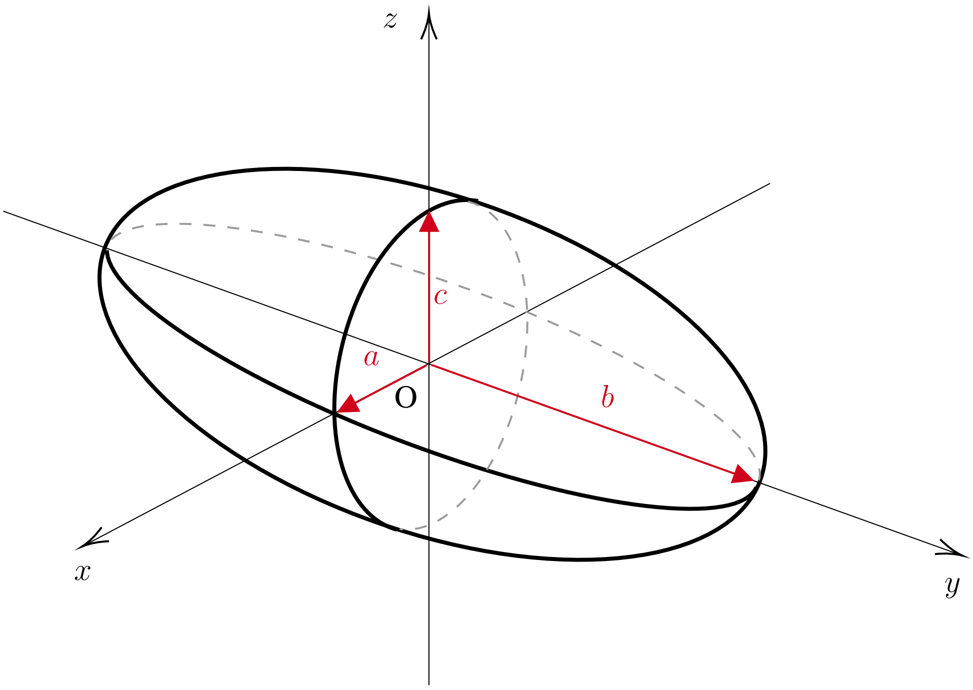


Figure 8.181: **Representation of an ellipsoid**

However, there is another way of describing an ellipsoid using also curvilinear coordinates:

$$x=a\cos(\theta)\cos(\phi), y=b\cos(\theta)\sin(\phi), z=c\sin(\theta) \tag{8.5.288}$$

So the parametric equation of an ellipsoid will be:

$$\boxed{\begin{pmatrix}x(\theta,\phi)\\y(\theta,\phi)\\z(\theta,\phi)\end{pmatrix}=\begin{pmatrix}a\cos(\theta)\cos(\phi)\\b\cos(\theta)\sin(\phi)\\c\sin(\theta)\end{pmatrix}} \tag{8.5.289}$$

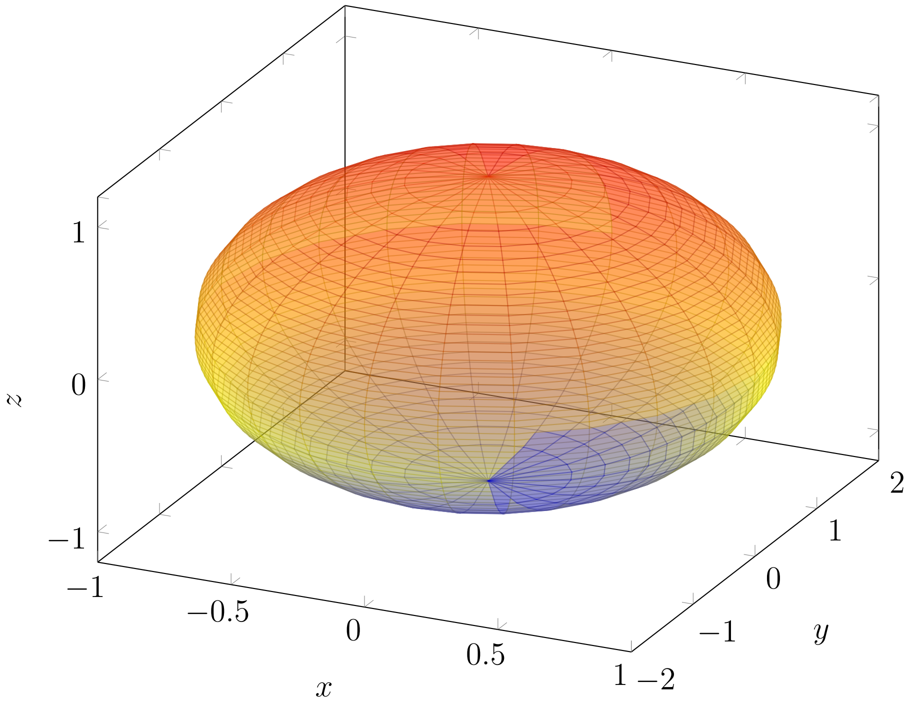


Figure 8.182: **High definition Ellipsoid made with LaTeX/Tikz**

We can see that this is simply the parametric equation of a sphere but with different radii along the axes of the chosen basis.

Example:

With Maple™ 4.00b:

```
>a:=100:b:=2:c:=20:
>plot3d([a*cos(theta)*cos(lambda),b*cos(theta)*sin(lambda),
c*sin(theta)],lambda=0..Pi,theta=-Pi..Pi,orientation=[50,60],
style=PATCH,axes=NORMAL);
```

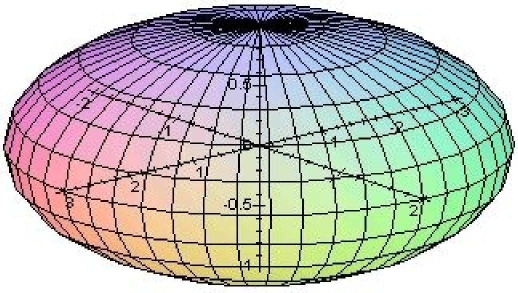


Figure 8.183: **Parametric representation of an ellipsoid in Maple™ 4.00b**

So we have:

$$x^2=a^2\cos^2(\theta)\cos^2(\phi),y^2=b^2\cos^2(\theta)\sin^2(\phi),z^2=c^2\sin^2(\theta) \tag{8.5.290}$$

hence:

$$\begin{aligned}\frac{x^2}{a^2}+\frac{y^2}{b^2}+\frac{z^2}{c^2}&=\cos^2(\theta)\cos^2(\phi)+\cos^2(\theta)\sin^2(\phi)+\sin^2(\theta)=\cos^2(\theta)(\cos^2(\phi)+\sin^2(\phi))+\sin^2(\theta)\\&=\cos^2(\theta)+\sin^2(\theta)=1\end{aligned} \tag{8.5.291}$$

Finally:

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}+\frac{z^2}{c^2}=1 \tag{8.5.292}$$

In astronomy (see section Astronomy page 5217) the ellipsoid is mainly named "oblate spheroid".

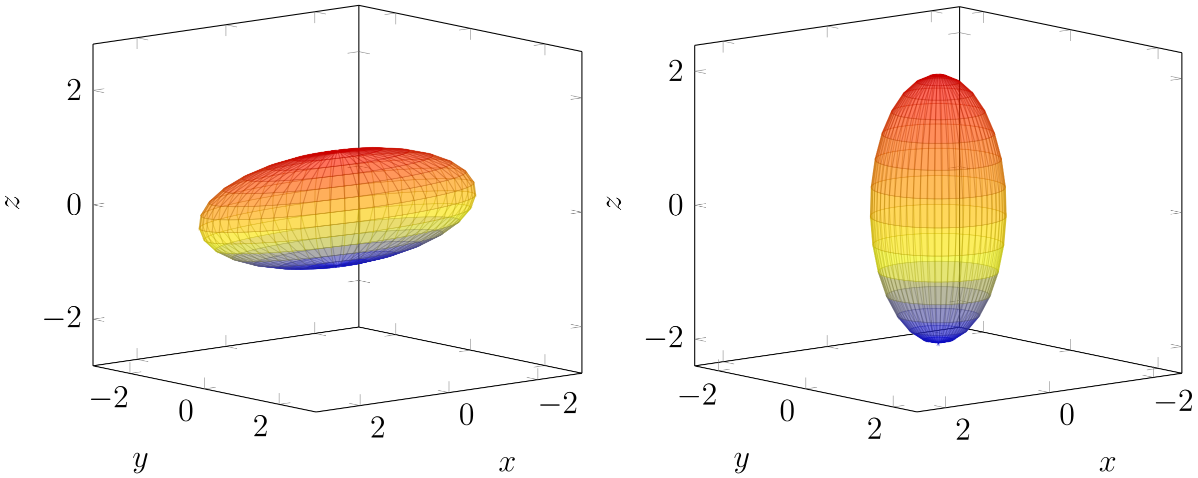


Figure 8.184: Oblate spheroid (on the left), Prolate spheroid (on the right)

and we have technically a oblate spheroid when $c<a$ and prolate one when $c>a$. Obviously the case $a=c$ reduces to a sphere.

Remark

Because of the combined effects of gravity and rotation, the Earth's shape is not quite a sphere but instead is slightly flattened in the direction of its axis of rotation. For that reason, in cartography the Earth is often approximated by an oblate spheroid instead of a sphere. The actual World Geodetic System model uses a spheroid whose radius is $6,378.137$ [km] at the equator and $6,356.752$ [km] at the poles.

The aspect ratio of an oblate spheroid/ellipse, $b:a$, is the ratio of the polar to equatorial lengths, while the flattening (also named "oblateness") f, is the ratio of the equatorial-polar length difference to the equatorial length:

$$f=\frac{a-b}{a}=1-\frac{b}{a} \tag{8.5.293}$$

8.5.2.10.4 Cylinder

It goes without saying that the parametric equation of a cylinder of radius r is given by:

$$\boxed{\begin{pmatrix} x(\phi) \\ y(\phi) \\ z \end{pmatrix} = \begin{pmatrix} r\cos(\phi) \\ r\sin(\phi) \\ z \end{pmatrix}} \qquad (8.5.294)$$

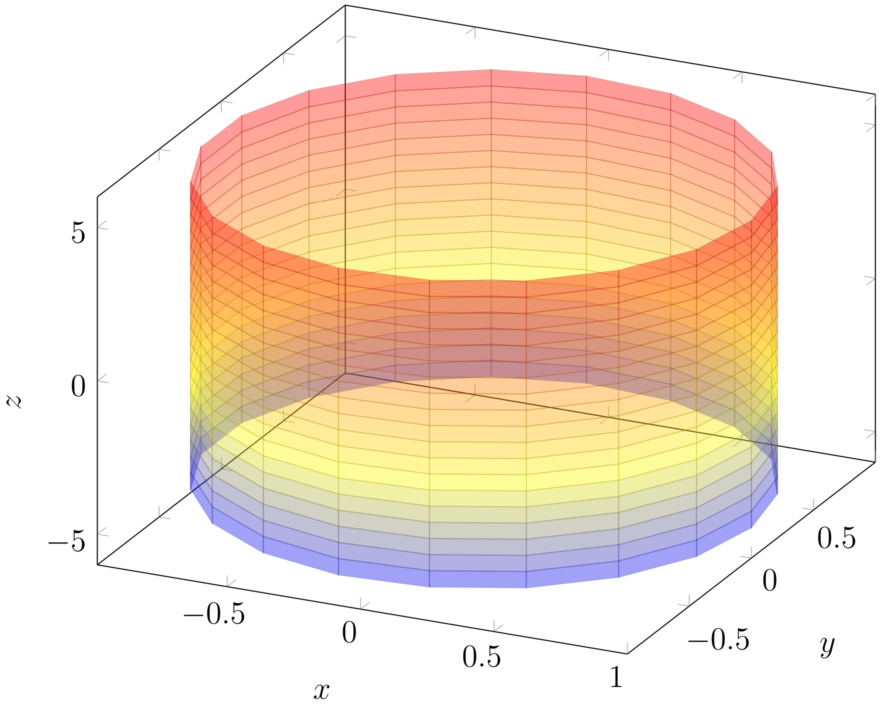


Figure 8.185: **High definition Cylinder made with LaTeX/Tikz**

We see well that the components x, y satisfy the Cartesian equation of a circle for every z since:

$$x^2 + y^2 = r^2 \qquad (8.5.295)$$

✍**Example:**

With Maple™ 4.00b a cylinder of radius $r = 1$:

```
>plot3d([cos(phi),sin(phi),z],phi=-Pi..Pi,z=0..2,orientation=[50,60],
style=PATCH,axes=NORMAL);
```

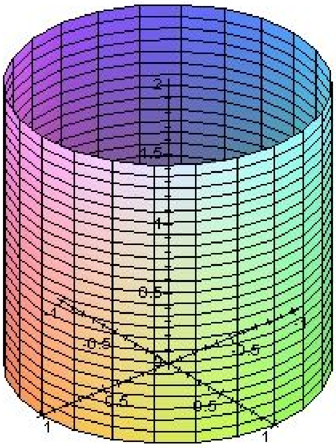


Figure 8.186: **Parametric representation of a cylinder in Maple™ 4.00b**

It is obvious that the parametric equation of an elliptical base cylinder is given by:

$$\boxed{\begin{pmatrix} x(\phi) \\ y(\phi) \\ z \end{pmatrix} = \begin{pmatrix} a\cos(\phi) \\ b\sin(\phi) \\ z \end{pmatrix}} \tag{8.5.296}$$

which also satisfies the parametric equation of an ellipse in the plane:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \tag{8.5.297}$$

8.5.2.10.5 Paraboloid

Given the equation of the parabola around the z-axis:

$$z=\frac{x^2}{2p} \tag{8.5.298}$$

with for reminder:

$$p=eh \tag{8.5.299}$$

We have obviously (by cutting the paraboloid by a horizontal plane which gives therefore a circle of radius r) the relation:

$$z=\frac{r^2}{2p} \tag{8.5.300}$$

named "cylinder equation". But we also have by applying the Pythagorean theorem in the circle above:

$$r^2=x^2+y^2 \tag{8.5.301}$$

Therefore:

$$x^2+y^2=2pz \tag{8.5.302}$$

Hence we get the "Cartesian equation of the paraboloid (of revolution)":

$$z=\frac{1}{2}\left(\frac{1}{p}x^2+\frac{1}{p}y^2\right) \tag{8.5.303}$$

We build the parametric equation of the paraboloid exactly in the same way as for the cone, at the difference that the evolution along the z-axis will not be linear relatively to a parameter k but to the square of the latter. Therefore we have:

$$\boxed{\begin{pmatrix} x(t,k) \\ y(t,k) \\ z(k) \end{pmatrix}=\begin{pmatrix} kr\cos(t) \\ kr\sin(t) \\ kh^2 \end{pmatrix}=\begin{pmatrix} k\dfrac{1}{2p}\cos(t) \\ k\dfrac{1}{2p}\sin(t) \\ kh^2 \end{pmatrix}} \tag{8.5.304}$$

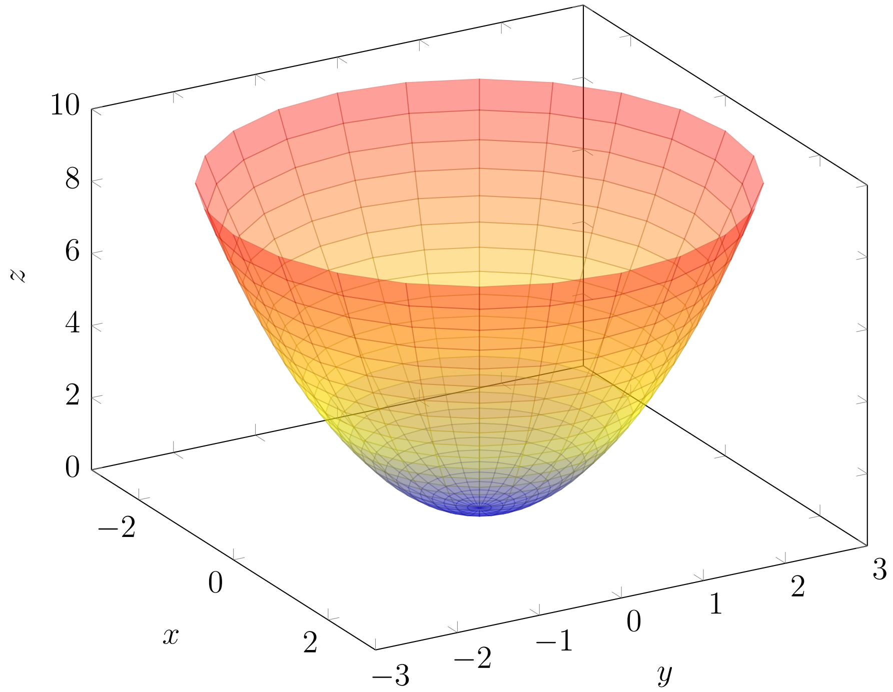


Figure 8.187: **High definition Paraboloid made with LaTeX/Tikz**

✍**Example:**

With Maple™ 4.00b a paraboloid with $h = 1, p = 1$:

```
>p:=1:h:=1:
>plot3d([k*1/(2*p)*cos(t),k*1/(2*p)*sin(t),k^2*h],k=0..10,
t=0..2*Pi,orientation=[50,60],style=PATCH,axes=NORMAL);
```

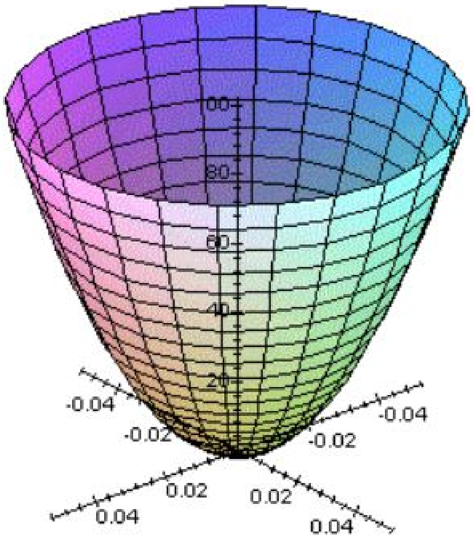


Figure 8.188: **Parametric representation of a paraboloid in Maple™ 4.00b**

8.5.2.10.6 Hyperboloid

Given the linear function:

$$z=x \tag{8.5.305}$$

that we turn around the z-axis. As the line passes trough the origin, the revolution around the z-axis is similar to two cones facing each other with their apex at the origin O. Therefore:

$$z^2=x^2+y^2 \tag{8.5.306}$$

which gives us obviously:

$$x^2+y^2-z^2=0 \tag{8.5.307}$$

Definition (#8.5.359): Any surface generated by a line is a "ruled surface". More explicitly a surface S is ruled if through every point of S there is a straight line that lies on S. A ruled surface S is a curved surface which can be generated by the continuous motion of a straight line in space along a space curve named the "directrix".

Let us take an important example (nuclear central chimney shape, gears, etc.) that is the "hyperboloid of one sheet" of equation:

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}-\frac{z^2}{c^2}=1 \tag{8.5.308}$$

More often denoted as:

$$-\frac{x^2}{a^2}-\frac{y^2}{b^2}+\frac{z^2}{c^2}=1 \tag{8.5.309}$$

Or as:

$$\frac{z^2}{c^2}-\left(\frac{x^2}{a^2}+\frac{y^2}{b^2}\right)=1 \tag{8.5.310}$$

If we put $a=b=c=1$ we have therefore (rotated line around the z axis but not going through the origin O):

$$x^2+y^2-z^2=1\Leftrightarrow x^2-z^2=1-y^2 \tag{8.5.311}$$

which can also be written as the product of the equation of two lines such that:

$$(x-z)(x+z)=(1-y)(1+y) \tag{8.5.312}$$

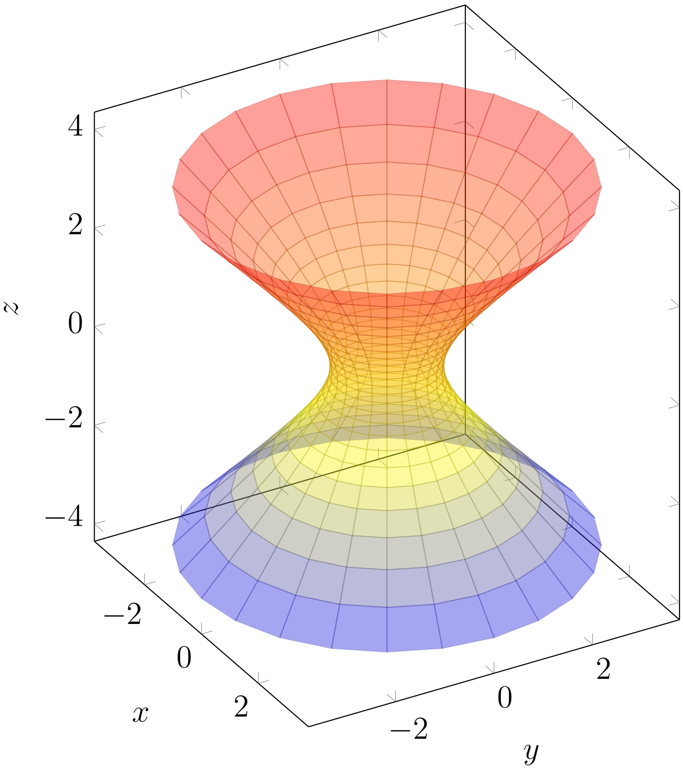


Figure 8.189: **High definition Hyperboloid of one sheet made with LaTeX/Tikz**

That we can also rewrite with $k \in \mathbb{R}$:

$$(x-z)k(x+z)=k(1-y)(1+y) \tag{8.5.313}$$

by identification:

$$\text{D1}: x-z=k(1-y) \tag{8.5.314}$$
$$\text{D2}: 1+y=k(x+z) \tag{8.5.315}$$

It is not trivial but these to two lines belong to the same surface and any point belonging to one of these two lines is contained therein. The figures below show well that fact, every point belongs to these two lines:

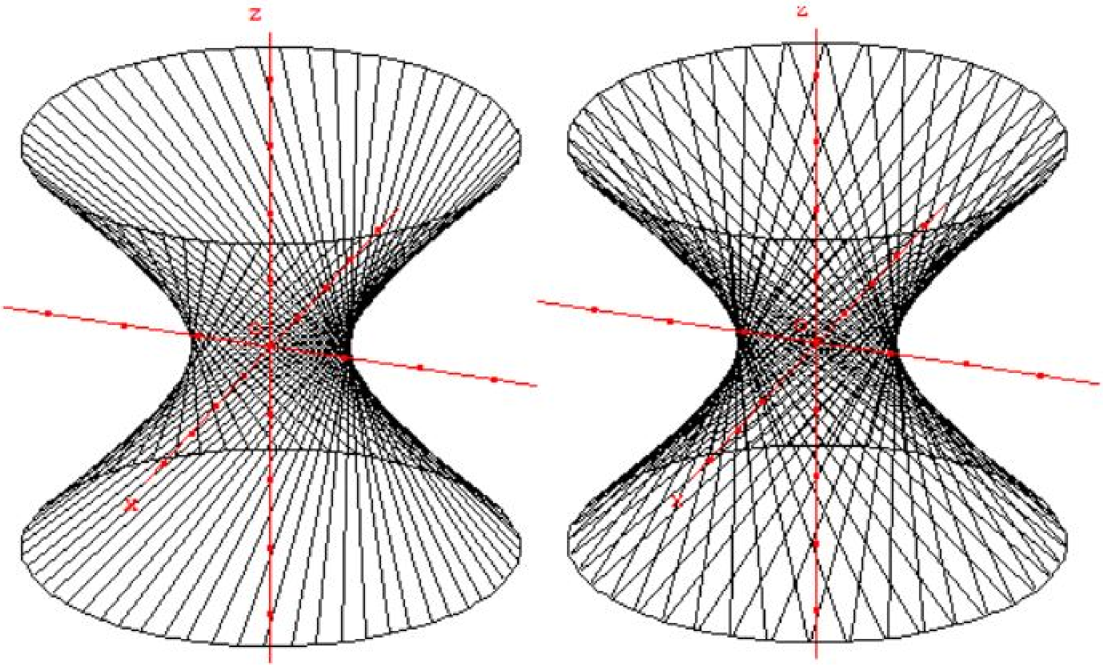


Figure 8.190: **Hyperbolic Rules Surface**

We could also describe this by circles such that (this is the traditional way on computers):

$$x^2+y^2-z^2=1 \leftrightarrow x^2+y^2=(1+z^2)=r^2 \tag{8.5.316}$$

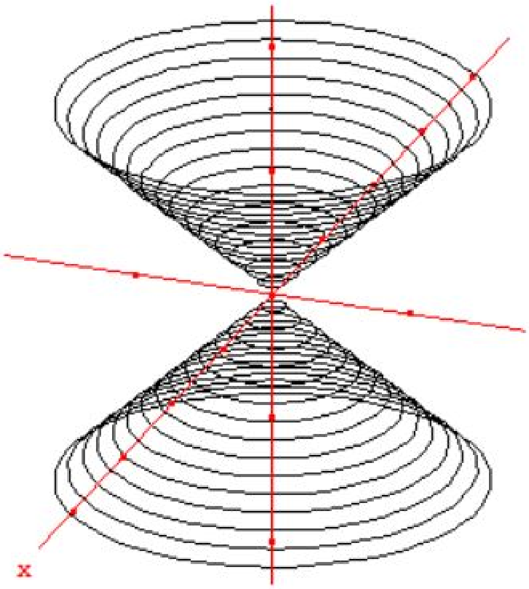


Figure 8.191: **Hyperbolic Circles Surface**

✍**Example:**

With Maple™ 4.00b a hyperboloid with $a=2, b=2, c=3$:

```
>with(plots):
>a:=2,b:=2,c:=3:
>plot3d(>implicitplot3d(x^2/a^2+y^2/b^2-z^2/c^2-1,x=-4..4,y=-4..4,
z=-4..4,orientation=[50,60],style=PATCH,axes=NORMAL);
```

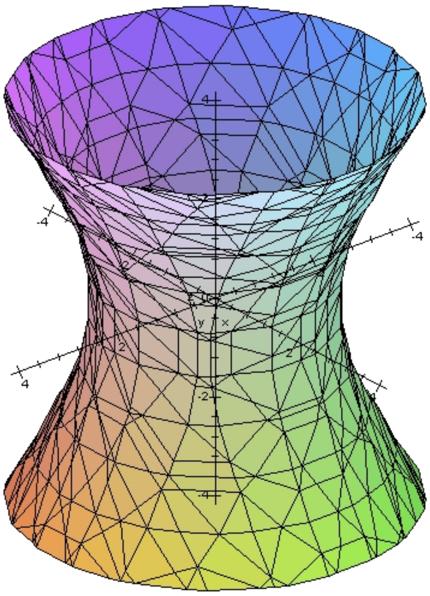


Figure 8.192: Parametric representation of a hyperboloid of one sheet in Maple™ 4.00b

Consider the case the equation $Ax^2+By^2+Cz^2=D$. Assume that $D>0$. Therefore:

- If two coefficients are positive and one is negative we speak of : "hyperboloid of one sheet".

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}-\frac{z^2}{c^2}=1 \tag{8.5.317}$$

- If two coefficients are negative and one is positive we speak of: "hyperboloid of two sheets

$$-\frac{x^2}{a^2}-\frac{y^2}{b^2}+\frac{z^2}{c^2}=1 \tag{8.5.318}$$

W will meet the latter case during our study of the Michelson interferometer.

Here is also a very interesting figure that highlights the difference between a one sheet and two sheets hyperboloids:

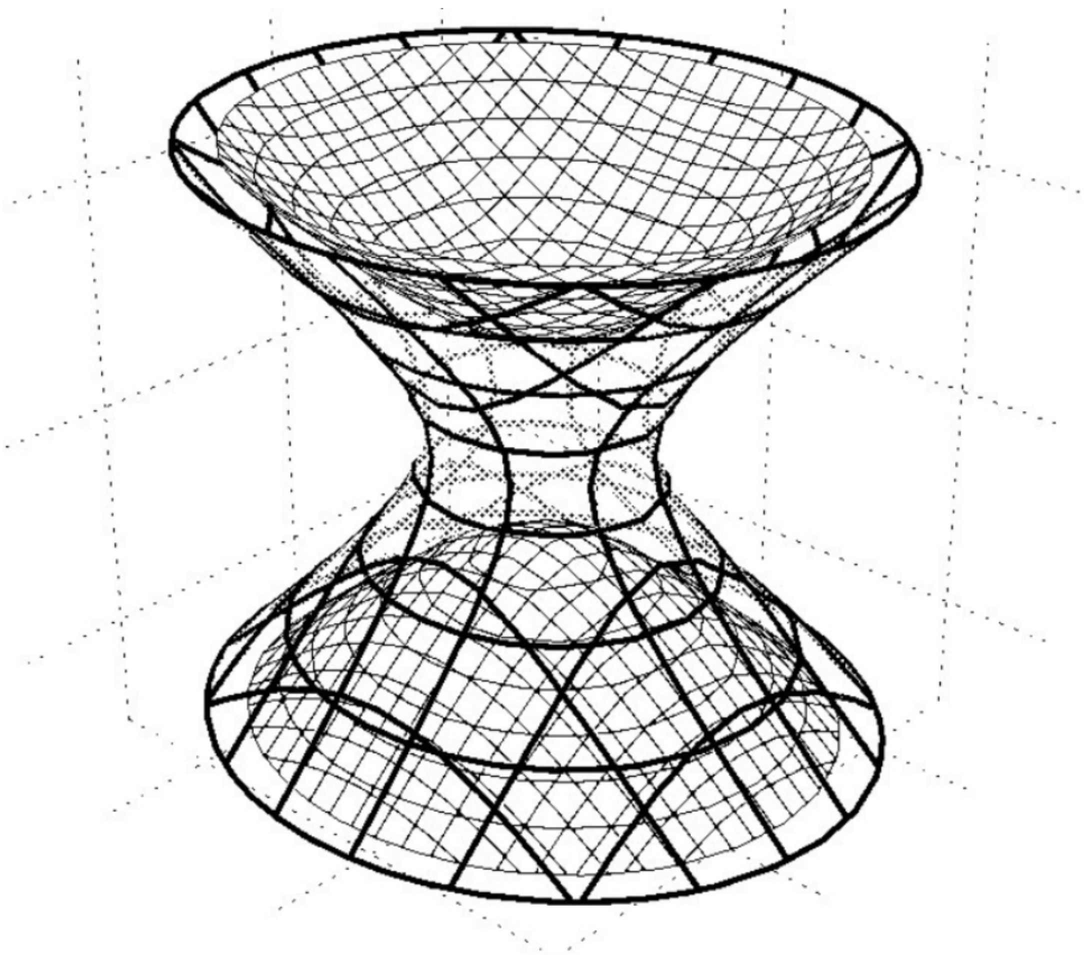


Figure 8.193: Two hyperboloids (of one sheet and two sheets) which are asymptotic to an identical elliptic cone

There are many other parametrizations, of the hyperboloid and so no such thing as the parametrization.

We have $x^2+y^2+1=z^2$. If we write $x^2+y^2=r^2$ (so we can take $x=r\cos(v)$, $y=r\sin(v)$), then $r^2+1=z^2$. Now since $\cosh^2(u)-\sinh^2(u)=1$, we could take $z=\cosh(u), r=\sinh(u)$, and thus:

$$\begin{aligned} x &= \sinh(u)\cos(v) \\ y &= \sinh(u)\sin(v) \\ z &= \cosh(u) \end{aligned} \tag{8.5.319}$$

However, this is not a good parametrization at $u=0$ (corresponding to the single point $(x,y,z)=(0,0,1)$.

Now let's try a different way. Write the equation as $t^2-y^2=1$ where $t^2=z^2-x^2$. We can take $y=\sinh(v)$ (note that this ranges over $\mathbb{R}$ as v does), and then $t^2=\cosh^2(v)$. If we choose the positive t (and we can), we have $t=\cosh(v)$. Now $(z/t)^2-(x/t)^2=1$. For the sheet with $z>0$ we can take $z/t=\cosh(u)$ and $x/t=\sinh(u)$. As u ranges over $\mathbb{R}$, we get all possible (x,z) pairs with $z>0$. So we have:

$$\begin{aligned} x &= \sinh(u)\cosh(v) \\ y &= \sinh(v) \\ z &= \cosh(u)\cosh(v) \end{aligned} \tag{8.5.320}$$

This parametrization is better because it maps $\mathbb{R}^2$ one-to-one onto one sheet.

We would like to share now a very nice and interesting (instructive) summary of most well known conical surfaces that we have studied so far and provided by OpenStax:

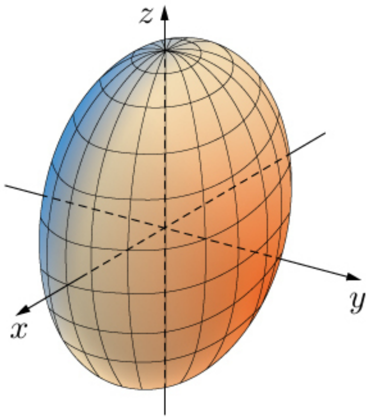
Ellipsoid

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}+\frac{z^2}{c^2}=1$$

Traces:

- In plane $z=p$: an ellipse
- In plane $y=q$: an ellipse
- In plane $x=r$: an ellipse

If $a=b=c$ then this surface is sphere.



Hyperboloid of One Sheet

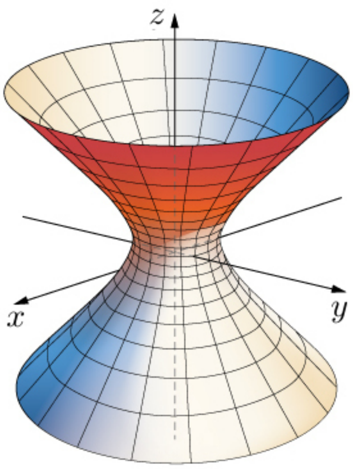
$$\frac{x^2}{a^2}+\frac{y^2}{b^2}-\frac{z^2}{c^2}=1$$

Traces:

- In plane $z=p$: an ellipse
- In plane $y=q$: a hyperbola
- In plane $x=r$: a hyperbola

In the equation for this surface, two of the variables have positive coefficients and one has negative coefficient.

The axis of the surface corresponds to the variable with the negative coefficient.



Hyperboloid of Two Sheets

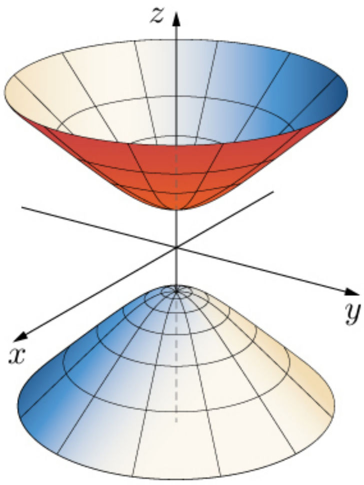
$$-\frac{x^2}{a^2}-\frac{y^2}{b^2}+\frac{z^2}{c^2}=1$$

Traces:

- In plane $z=p$: an ellipse or the empty set (no trace)
- In plane $y=q$: a hyperbola
- In plane $x=r$: a hyperbola

In the equation for this surface, two of the variables have negative coefficients and one has positive coefficient.

The axis of the surface corresponds to the variable with a positive coefficient. The surface does not intersect the coordinate plane perpendicular to the axis.



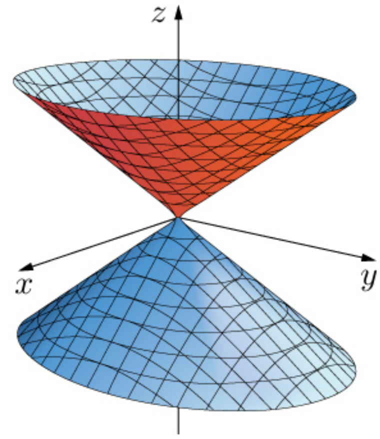
Elliptic Cone

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}-\frac{z^2}{c^2}=0$$

Traces:

- In plane $z=p$: an ellipse
- In plane $y=q$: a hyperbola
- In plane $x=r$: a hyperbola
- In the xy-plane: a pair of lines that intersect at the origin
- In the yz-plane: a paire of lines that intersect at the origin

The axis of the surface corresponds to the variable with a negative coefficient. The traces in the coordinate planes parallel to the axis are intersecting lines.



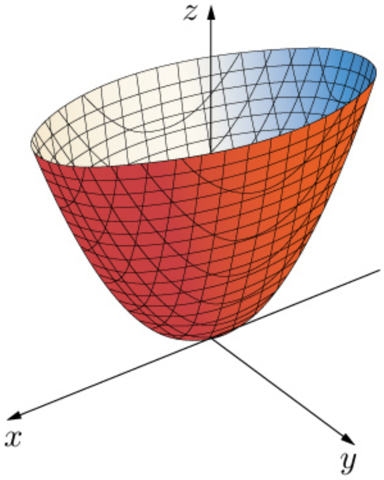
Elliptic Paraboloid

$$z=\frac{x^2}{a^2}+\frac{y^2}{b^2}$$

Traces:

- In plane $z=p$: an ellipse
- In plane $y=q$: a parabola
- In plane $x=r$: a parabola

The axis of the surface corresponds to the linear variable.



Hyperbolic Paraboloid

$$z=\frac{x^2}{a^2}-\frac{y^2}{b^2}$$

Traces:

- In plane $z=p$: a hyperbola
- In plane $y=q$: a parabola
- In plane $x=r$: a parabola

The axis of the surface corresponds to the linear variable.

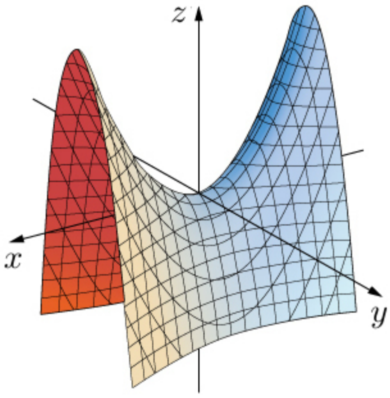


Figure 8.194: Common conical surfaces (source: OpenStax)

8.5.2.10.7 Torus

Definition (#8.5.360): In geometry, a "torus" is a surface of revolution generated by revolving a circle in three-dimensional space about an axis coplanar with the circle. If the axis of revolution does not touch the circle, the surface has a ring shape and is named a "torus of revolution".

We know that to generate a circle in the plane $x\text{O}y$, a possible parametric equation is:

$$x = r\cos(\phi) \qquad (8.5.321)$$

$$y = r\sin(\phi) \qquad (8.5.322)$$

To offset this circle to the right (in the direction of the positive x), we'll just add a strictly positive constant term x:

$$\begin{aligned} x &= a + r\cos(\phi) \\ y &= r\sin(\phi) \end{aligned} \qquad (8.5.323)$$

In space to draw a shifted circle in the $x\text{O}z$-plane we have therefore:

$$\begin{aligned} x &= a + r\cos(\phi) \\ z &= r\sin(\phi) \end{aligned} \qquad (8.5.324)$$

Which traditional notation is in the context of the study of torus:

$$\begin{aligned} x &= R + r\cos(\phi) \\ z &= r\sin(\phi) \end{aligned} \qquad (8.5.325)$$

If we want to generate a torus, we will rotate the circle around the u-axis by making it following a circle in the plane $x\text{O}y$. We then have the "parametric equation of the torus of revolution":

$$\boxed{\begin{pmatrix} x(\phi,\theta) \\ y(\phi,\theta) \\ z(\phi) \end{pmatrix} = \begin{pmatrix} (R + r\cos(\phi))\cos(\theta) \\ (R + r\cos(\phi)\sin(\theta) \\ r\sin(\phi) \end{pmatrix}} \qquad (8.5.326)$$

✍**Example:**

With Maple™ 4.00b a torus with $r = 1, R = 4$:

```
>r:=1:R:=4
>plot3d([(R+r*cos(phi))*cos(theta),(R+r*cos(phi))*sin(theta),
1*sin(phi)],
theta=-Pi..Pi,phi=-Pi..Pi,scaling=CONSTRAINED,orientation=[50,60],
style=PATCH,axes=NORMAL);
```

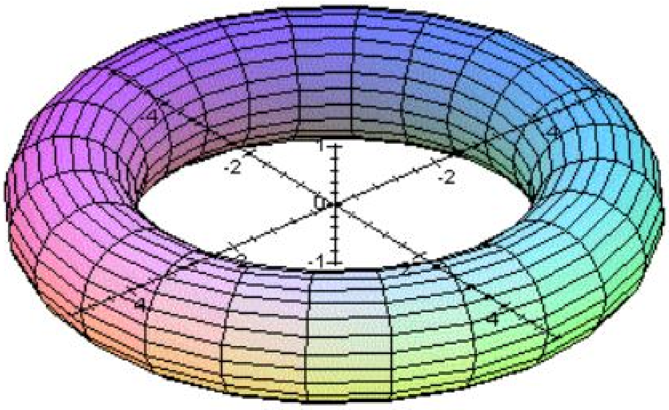


Figure 8.195: **Parametric representation of a torus in Maple™ 4.00b**

~80 %

★★★★☆

14 votes, 67.14%

8.6. Differential Geometry

As we have already mentioned it in the section of Non-Euclidean geometry, differential geometry is the branch of geometry that aims to study the local and intrinsic properties (in the neighbourhood of a point) of curves and non-Euclidean surfaces (as a generalization of Euclidean surfaces!).

Differential geometry takes its name from the fact that it was born from the possibility of a kinematic interpretation that infinitesimal calculus brings to the study of curves. The points that we will discuss here will also serve well in the study of classical mechanics as complex analysis applied to many areas of study physical fields.

Remark

Before we tackle the very formal and abstract manner to address differential geometry with the tools of Topology (usual method used by mathematicians) we have chose in a first time to present the essential elements in the most simple and enjoyable as this is done in some engineering schools. Purists will therefore perhaps excuse us... or wait for something better...

8.6.1 Parametric Curves

Definition (#8.6.361): We will assimilate the "physical space" to $\mathbb{R}^3$ and will suppose that it includes a referential $R=(\mathrm{O},\vec{i},\vec{j},\vec{k})$ and we will denote by B the base $(\vec{i},\vec{j},\vec{k})$.

Let us consider a set $I\subset\mathbb{R}$ and a function $\Gamma=f(I)$ such that:

$$f\begin{pmatrix} I\rightarrow\mathbb{R}^3 \\ t\rightarrow f(t)=M \end{pmatrix},M(x,y,z)_R \tag{8.6.1}$$

that is a parametric equations that defines a group of quantities as functions of one or more independent "parameters".

Parametric curves are commonly used in kinematics, where the trajectory of an object is represented by equations depending on time as the parameter. Because of this application, a single parameter is often labelled t. However, parameters can represent other physical quantities (such as geometric variables) or can be selected arbitrarily for convenience.

Remarks

R1. If f is continuous, then Γ is a curve in space named "curve in one piece".

R2. A parabola, a sinusoid are curves named "flat curves". An ellipse, a circle are named "closed planar curves". For these examples, all points of the considered curves are located in a same plane. Conversely, a curve is named "left curve" if it is not so.

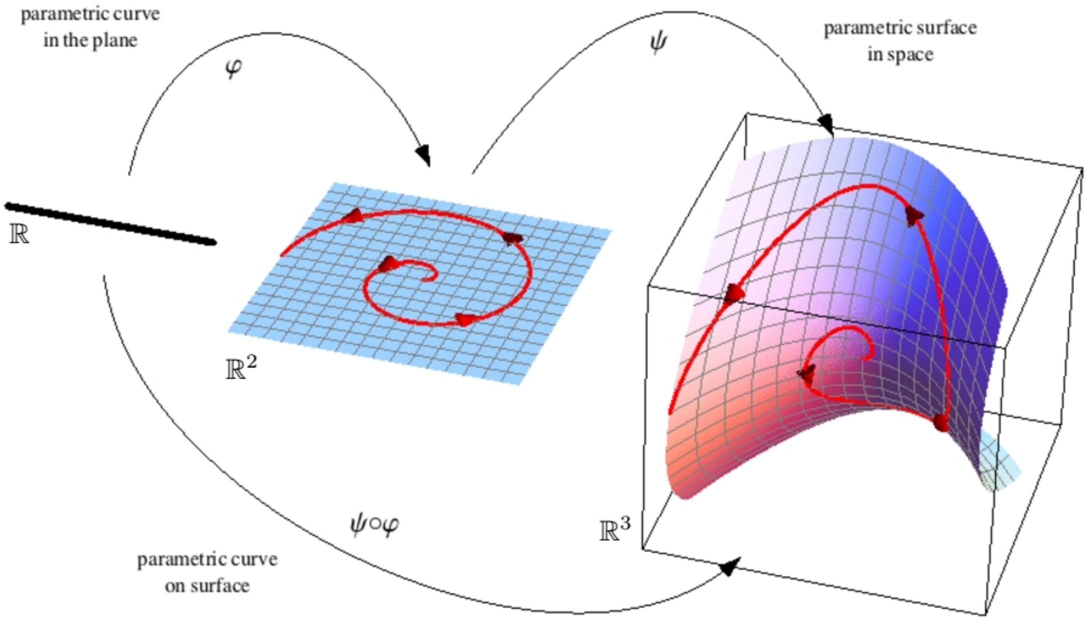


Figure 8.196: **Parametric curve and parametric surface**

Let us choose $t_0 \in I$ and put $M_0 = f(t_0)$ which we will denote by $M(t_0)$ then we can state the following definition: the pair (f, I) where f is a continuous function is named "parametric arc". Γ is named the "support" of (f, I) and t_0 is an "origin" of (f, I).

Note that parametric representations are generally non-unique, so the same quantities may be expressed by a number of different parametrizations.

Remarks

R1. Abusively, and more frequently, we also say that (f, I) is a "parametrization" of Γ.

R2. It is easy to define other parametrized arcs admitting also Γ as support. To do this, it is sufficient to give us a bijective function φ of I to $J \subset \mathbb{R}$ and such that $f(t) = f(\varphi^{-1}(x))$.

Before continuing, remember that in differential geometry, the "curvilinear abscissa" or "line-element" is a kind of algebraic variation of the length of an arc (this is therefore the analogous, on a curve, of the abscissa on a straight oriented line).

Let us consider now the following curvilinear abscissa (see section Special Relativity page 5704):

$$\mathrm{d}s^2 = g_{ij}\mathrm{d}u^i\mathrm{d}u^j \tag{8.6.2}$$

we have already seen that in a canonical Euclidean space (in $\mathbb{R}^n$) the curvilinear abscissa is then written:

$$\mathrm{d}s^2 = \delta_{ij}\mathrm{d}u^i\mathrm{d}u^j \tag{8.6.3}$$

with $i,j=1,2,3$ and as we have $\delta_{ij}=0, i\neq j$, it remains in $\mathbb{R}^3$:

$$\mathrm{d}s^2=\mathrm{d}u^1\mathrm{d}u^1+\mathrm{d}u^2\mathrm{d}u^2+\mathrm{d}u^3\mathrm{d}u^3=(\mathrm{d}u^1)^2+(\mathrm{d}u^2)^2+(\mathrm{d}u^3)^2 \tag{8.6.4}$$

In the Cartesian system:

$$u^1=x, u^2=y, u^3=z \tag{8.6.5}$$

so it comes:

$$\boxed{\mathrm{d}s^2=\mathrm{d}x^2+\mathrm{d}y^2+\mathrm{d}z^2} \tag{8.6.6}$$

which is therefore the linear differential element of an Euclidean space (the shortest path or the "geodesic" or "differential curvilinear abscissa") that we have already met many times in various section of this textbook. So this is nothing new or surprising!

If we restrict ourselves to the plane, the differential curvilinear abscissa of a plane curve is then obviously:

$$\boxed{\mathrm{d}s^2=\mathrm{d}x^2+\mathrm{d}y^2} \tag{8.6.7}$$

We already know how to use this equation (we used it in the section of Analytical Mechanics). But as a reminder never hurts, let us do examples with a straight line, a parabola and a half-circle (the choice is not innocent!).

✍**Examples:**

E1. Consider the general equation of a line in the plane (it's not a plane curve for reminder but a straight flat line!):

$$y=ax+b \tag{8.6.8}$$

It then comes immediately:

$$\mathrm{d}y=a\mathrm{d}x \tag{8.6.9}$$

Therefore:

$$\mathrm{d}s^2=\mathrm{d}x^2+\mathrm{d}y^2=\mathrm{d}x^2+(a\mathrm{d}x)^2=(1+a^2)\mathrm{d}x^2 \tag{8.6.10}$$

E2. Consider now the following equation of a parabola in the plane:

$$y=ax^2+b \tag{8.6.11}$$

It then comes immediately:

$$\mathrm{d}y=2ax\mathrm{d}x \tag{8.6.12}$$

Therefore:

$$\mathrm{d}s^2=\mathrm{d}x^2+\mathrm{d}y^2=\mathrm{d}x^2+(2ax\mathrm{d}x)^2=(1+4a^2x^2)\mathrm{d}x^2 \tag{8.6.13}$$

E3. Consider the equation of an ellipse (see section Analytical Geometry page 2768) written as:

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=R^2 \tag{8.6.14}$$

after rearrangement:

$$y=\sqrt{b^2\left(R-\frac{x^2}{a^2}\right)}=\sqrt{b^2R-b^2\frac{x^2}{a^2}} \tag{8.6.15}$$

It then comes immediately:

$$\mathrm{d}y=-\frac{b^2}{a^2\sqrt{b^2R-b^2\dfrac{x^2}{a^2}}}x\mathrm{d}x \tag{8.6.16}$$

Hence:

$$\mathrm{d}s^2=\mathrm{d}x^2+\mathrm{d}y^2=\mathrm{d}x^2+\left(-\frac{b^2}{a^2\sqrt{b^2R-b^2\dfrac{x^2}{a^2}}}x\mathrm{d}x\right)^2=\left(1+\frac{b^2}{a^4R-a^2x^2}x^2\right)\mathrm{d}x^2 \tag{8.6.17}$$

Notice that by making a Maclaurin approximation (when x is therefore zero, which corresponds to the study of the pole of the ellipse) we get (see section Sequences and Series page 1703):

$$y=\sqrt{b^2\left(R^2-\frac{x^2}{a^2}\right)}\cong\sqrt{b^2R}-\frac{1}{2}\frac{\sqrt{b^2R}}{a^2R}x^2 \tag{8.6.18}$$

Following the request of a reader here are the details of the development of the previous result. First recall the Taylor series (see section Sequences and Series page 1696):

$$f(x)\cong f(x_0)+\frac{1}{1!}f'(x_0)(x-x_0)+\frac{1}{2!}f''(x_0)(x-x_0)^2+\ldots \tag{8.6.19}$$

If we put $x_0=0$, we get the Maclaurin series:

$$f(x)\cong f(0)+\frac{1}{1!}f'(0)x+\frac{1}{2!}f''(0)x^2+\ldots \tag{8.6.20}$$

Therefore we get proceeding in two stages:

$$\begin{aligned}
y \cong\ & \frac{\mathrm{d}}{\mathrm{d}x^0}\sqrt{b^2\left(R^2-\frac{x^2}{a^2}\right)}\Bigg|_{x=x_0}(x-x_0)^0+\frac{1}{1!}\frac{\mathrm{d}}{\mathrm{d}x^1}\sqrt{b^2\left(R^2-\frac{x^2}{a^2}\right)}\Bigg|_{x=x_0}(x-x_0)^1\\
&+\frac{1}{2!}\frac{\mathrm{d}}{\mathrm{d}x^2}\sqrt{b^2\left(R^2-\frac{x^2}{a^2}\right)}\Bigg|_{x=x_0}(x-x_0)^2\\
&=\sqrt{b^2\left(R^2-\frac{x_0^2}{a^2}\right)}+\frac{1}{1}\frac{\mathrm{d}}{\mathrm{d}x}\sqrt{b^2\left(R^2-\frac{x^2}{a^2}\right)}\Bigg|_{x=x_0}(x-x_0)\\
&+\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}x^2}\sqrt{b^2\left(R^2-\frac{x^2}{a^2}\right)}\Bigg|_{x=x_0}(x-x_0)^2\\
&=\sqrt{b^2\left(R^2-\frac{x_0^2}{a^2}\right)}+\frac{-\dfrac{b^2x}{a^2}}{\sqrt{b^2\left(R^2-\dfrac{x_0^2}{a^2}\right)}}\Bigg|_{x=x_0}(x-x_0)\\
&+\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}x}\frac{-\dfrac{b^2x}{a^2}}{\sqrt{b^2\left(R^2-\dfrac{x^2}{a^2}\right)}}\Bigg|_{x=x_0}(x-x_0)^2
\end{aligned}\tag{8.6.21}$$

$$\begin{aligned}
&=\sqrt{b^2\left(R^2-\frac{x_0^2}{a^2}\right)}+\frac{-\dfrac{b^2x_0}{a^2}}{\sqrt{b^2\left(R^2-\dfrac{x_0^2}{a^2}\right)}}(x-x_0)\\
&+\frac{1}{2}\left(\frac{-\dfrac{b^2}{a^2}}{\sqrt{b^2\left(R^2-\dfrac{x^2}{a^2}\right)}}+-\frac{2b^2}{a^2}x\cdot\ldots\right)\Bigg|_{x=x_0}(x-x_0)^2\\
&=\sqrt{b^2\left(R^2-\frac{x_0^2}{a^2}\right)}+\frac{-\dfrac{b^2x_0}{a^2}}{\sqrt{b^2\left(R^2-\dfrac{x_0^2}{a^2}\right)}}(x-x_0)\\
&+\frac{1}{2}\left(\frac{-\dfrac{b^2}{a^2}}{\sqrt{b^2\left(R^2-\dfrac{x_0^2}{a^2}\right)}}+-\frac{2b^2}{a^2}x_0\cdot\ldots\right)(x-x_0)^2
\end{aligned}\tag{8.6.22}$$

by putting $x_0=0$ we get:

$$\begin{aligned}
y &\cong \sqrt{b^2R}+0\cdot x+\frac{1}{2}\left(\frac{-\dfrac{b^2}{a^2}}{\sqrt{b^2\left(R^2-\dfrac{0^2}{a^2}\right)}}+-\frac{2b^2}{a^2}0\cdot\ldots\right)x^2\\
&=\sqrt{b^2R}+\frac{1}{2}\frac{-\dfrac{b^2}{a^2}}{\sqrt{b^2R}}x^2=\sqrt{b^2R}-\frac{1}{2}\frac{b^2}{a^2\sqrt{b^2R}}x^2=\sqrt{b^2R}-\frac{1}{2}\frac{\sqrt{b^2R}}{\sqrt{b^2R}}\frac{b^2}{a^2\sqrt{b^2R}}x^2\\
&=\sqrt{b^2R}-\frac{1}{2}\frac{\sqrt{b^2R}}{a^2R}x^2
\end{aligned}\tag{8.6.23}$$

It then comes immediately:

$$\mathrm{d}y=-\frac{\sqrt{b^2R}}{a^2R}x\mathrm{d}x\tag{8.6.24}$$

Therefore:

$$\mathrm{d}s^2=\mathrm{d}x^2+\mathrm{d}y^2=\mathrm{d}x^2+\left(-\frac{\sqrt{b^2R}}{a^2R}x\mathrm{d}x\right)^2=\left(1+\frac{b^2}{a^4R}x^2\right)\mathrm{d}x^2\tag{8.6.25}$$

We see then that the curvilinear abscissa of an ellipse in the plane becomes that of a parabola when we make a Maclaurin series expansion of the equation of the ellipse at the pole.

We could do the same with an hyperbole and end up with the same form of differential curvilinear abscissa, usually denoted by tradition:

$$\boxed{\mathrm{d}s^2=\mathrm{d}x^2+\mathrm{d}y^2=\left(1+k_x^2x^2\right)\mathrm{d}x^2}\tag{8.6.26}$$

where k_x is named the "osculator parameter of the parabola".

What is veeery important to notice with the developments above (we can do it also identically for the hyperbola that is just the elliptic case with a different sign on readers request) is that we finally have (remember that the ellipse is a general case of the sphere!):

$$\begin{aligned}
\mathrm{d}s_{\text{plane}}^2&=(1+a^2)\mathrm{d}x^2\\
\mathrm{d}s_{\text{parabole}}^2&=(1+4a^2x^2)\mathrm{d}x^2\\
\mathrm{d}s_{\text{ellipse}}^2&=(1+k_x^2x^2)\mathrm{d}x^2\\
\mathrm{d}s_{\text{hyperbola}}^2&=(1-k_x^2x^2)\mathrm{d}x^2
\end{aligned}\tag{8.6.27}$$

Where the last three cases are generally denoted:

$$\boxed{\mathrm{d}s=(1+K^2x^2)\mathrm{d}x^2}\tag{8.6.28}$$

We then understand better why many books on General Relativity or Differential Geometry say that:

- When $K=0$ all case reduce to a flat space
- When $K=+1$ (or more generally just positive) we are in a spherical space (in fact it's ellipsoidal)
- When $K=-1$ (or more generally just negative) we are in a hyperbolic space

These examples being closed, let us continue with the theory. We can obviously rewrite our differential curvilinear abscissa by dividing both sides of equality by $\mathrm{d}t$ as:

$$\boxed{\left(\frac{\mathrm{d}s}{\mathrm{d}t}\right)^2=\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2+\left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2+\left(\frac{\mathrm{d}z}{\mathrm{d}t}\right)^2} \tag{8.6.29}$$

Hence we get the famous relation:

$$\mathrm{d}s=\sqrt{\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2+\left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2+\left(\frac{\mathrm{d}z}{\mathrm{d}t}\right)^2}\,\mathrm{d}t \tag{8.6.30}$$

✍**Example:**

Let us see an application of the parametrized differential curvilinear abscissa with an helix (the examples are pretty in differential geometry and therefore nice to see...) which is a typical left curve:

Let $t,r,h\in\mathbb{R}^3,t>0,r>,h>0$ and the function:

$$f(t):\begin{cases}[0,t]\mapsto\mathbb{R}^3\\ t\rightarrow f(t)=M\end{cases} \tag{8.6.31}$$

with $M(x,y,z)$ and parametric coordinates:

$$\begin{pmatrix}x\\y\\z\end{pmatrix}=\begin{pmatrix}r\cos(t)\\r\sin(t)\\ht\end{pmatrix} \tag{8.6.32}$$

We then have with Maple™ 4.00b by taking r and h as equal to 1:

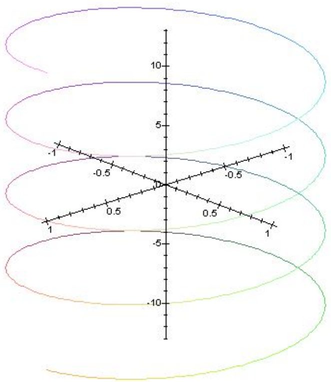


Figure 8.197: **Parametric representation of an helix in Maple™ 4.00b**

The function f is a parametric arc whose support is named an "helix", r is the radius and h is the step. Taking $t_0 = 0$ as the origin, the curvilinear abscissa of this helix (one piece) is given by:

$$\left(\frac{\mathrm{d}s}{\mathrm{d}t}\right)^2 = \left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2 + \left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2 + \left(\frac{\mathrm{d}z}{\mathrm{d}t}\right)^2 = r^2\sin^2(t) + r^2\cos^2(t) + h^2 = r^2 + h^2 \tag{8.6.33}$$

Therefore:

$$\frac{\mathrm{d}s}{\mathrm{d}t} = \sqrt{r^2 + h^2} \tag{8.6.34}$$

and hence by integration:

$$s = \sqrt{r^2 + h^2}t \tag{8.6.35}$$

8.6.2 Isolines

Let us see now a very important point in mathematics but also in medical engineering, astrophysics, meteorology (among many other areas) that are isolines.

Before addressing the subject in a purely mathematical form, we suggest the reader to open MATLAB® 5.0.0.473 (we've done also pretty much the same example with Maple™ 4.00b in the section of Functional Analysis) and write in it:

```
»[xx,yy,z]=peaks;
»figure(1);mesh(xx,yy,z);title('peaks')
```

We get:

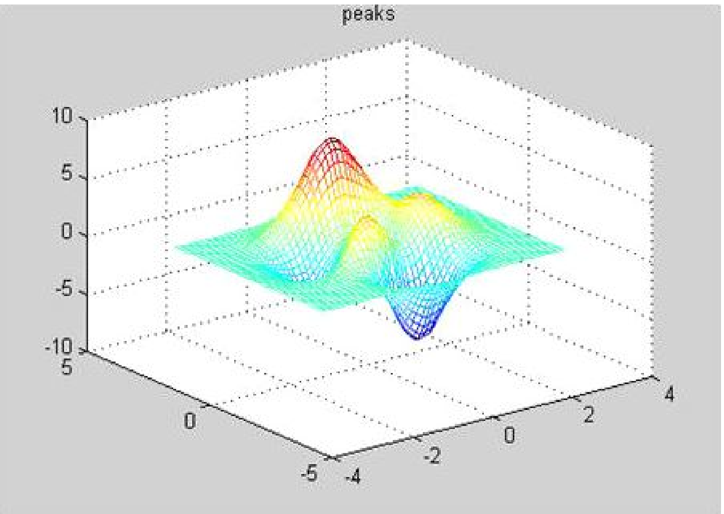


Figure 8.198: **Initial plot in MATLAB® 5.0.0.473**

then for aesthetic reasons, to write:

```
»figure(2);surf(xx,yy,z);title('surf')
```

We get:

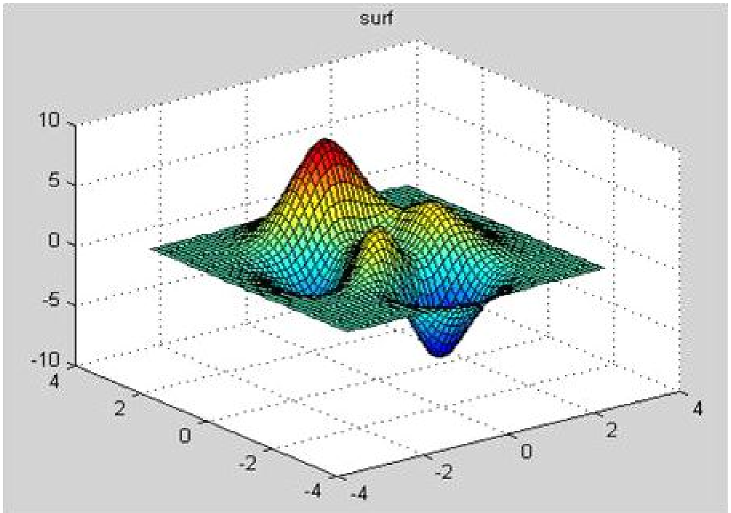


Figure 8.199: Plot rendering modification in MATLAB® 5.0.0.473

Then we would like MATLAB® to plot for us some level curves (the points where the value of the function $f(x, y)$ is constant), named by the mathematicians "isolines" or "iso-level curves". For this we must write:

```
»figure(3);contour3(xx,yy,z);title('contour')
```

We get:

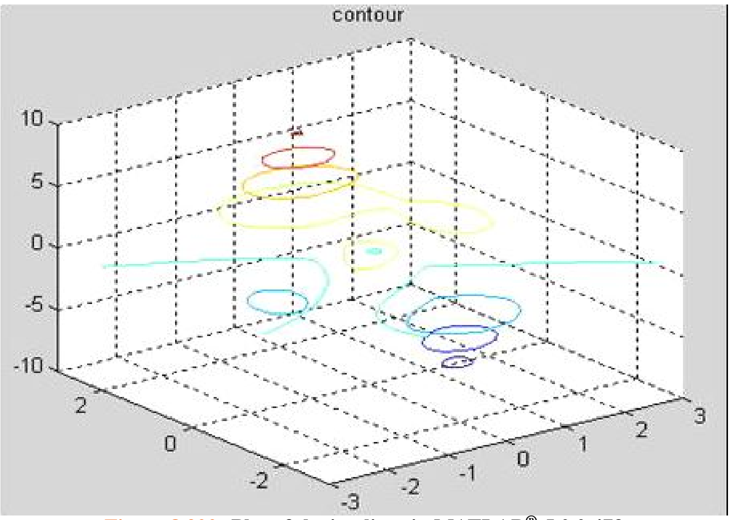


Figure 8.200: Plot of the isoclines in MATLAB® 5.0.0.473

We will then ask the software to project the isoclines on the X, Y plane. This is done with:

```
»figure(3);contour(xx,yy,z);title('contour')
```

We get:

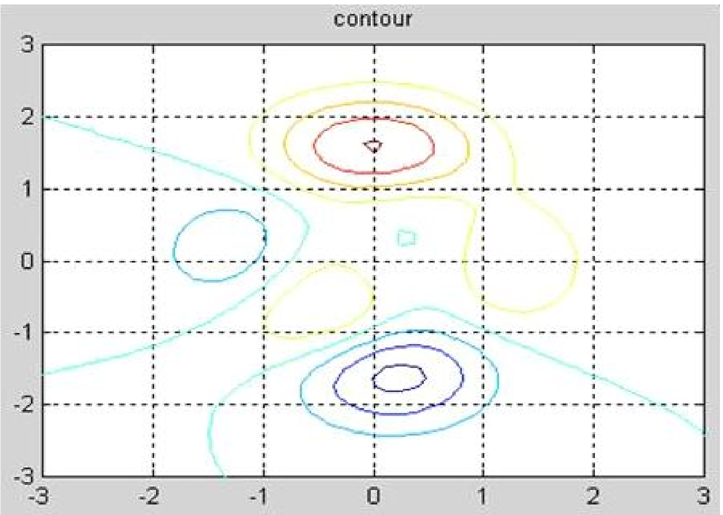


Figure 8.201: **Projection of the isoclines in MATLAB® 5.0.0.473**

And it is these curves that will interest us. We would like to determine the algebraic expression in the plane of the latter. But first let's have fun with MATLAB® by writing again:

```
»figure(3);contour(xx,yy,z);title('contour')
```

We get:

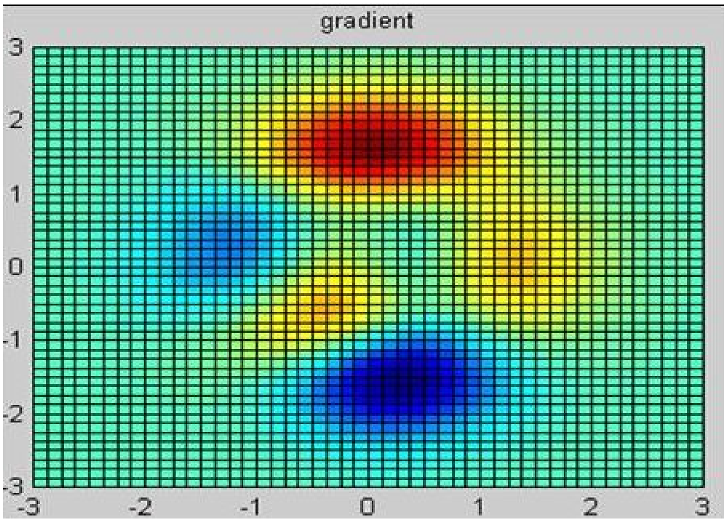


Figure 8.202: **Plane representation of isoclines with coloured gradients in MATLAB® 5.0.0.473**

but we can do even better by removing the grid with the command:

```
»shading interp
```

We get:

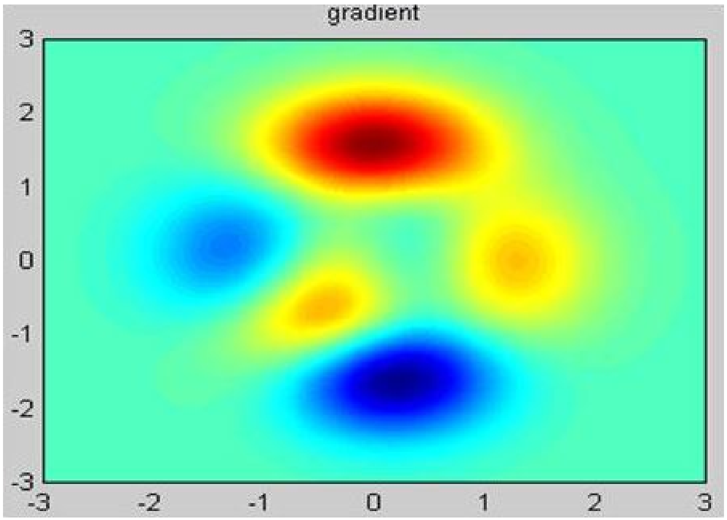


Figure 8.203: **Plane representation of isoclines with coloured gradients and without grid in MATLAB® 5.0.0.473**

Then, without closing the above now add the line:

```
»hold on; contour(xx,yy,z,'k')
```

We get:

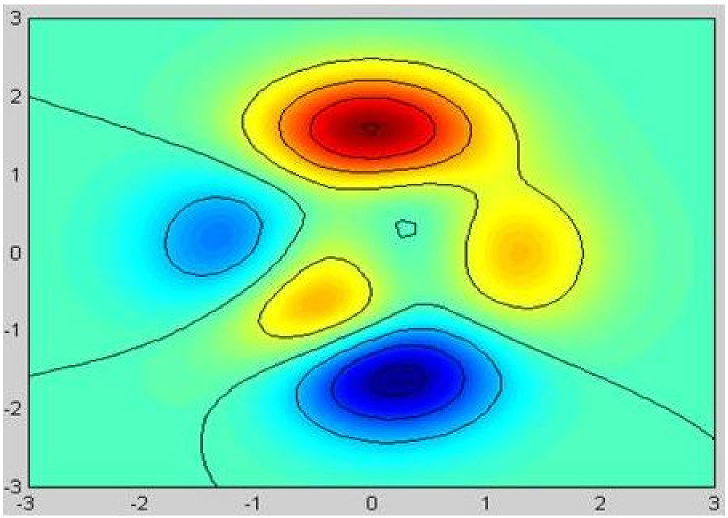


Figure 8.204: **Isoclines association projected with color gradients in MATLAB® 5.0.0.473**

Definition (#8.6.362): In mathematics, a level set of a real-valued function f of n real variables is a set of the form:

$$L_c(f)=\{(x_1,\cdots,x_n)\mid f(x_1,\cdots,x_n)=c^{te}\} \tag{8.6.36}$$

that is, a set where the function takes on a given constant value c^{te}.

When the number of variables is two, a level set is generically a curve, named a "level curve", "contour line", or "isoline". So a level curve is the set of all real-valued solutions of an equation in two variables x_1 and x_2. When $n=3$, a level set is named a "level surface" or "isosurface", and for higher values of n the level set is a named a "hypersurface". So a "level surface" is the set of all real-valued roots of an equation in three variables x_1, x_2 and x_3, and a "level hypersurface" is the set of all real-valued roots of an equation in $n\ (n>3)$ variables.

Let us now consider to determine the equation of the isolines the bivariate function $f(x,y)$ function and that we will impose as being $\mathbb{R}^2$-differentiable.

The relation:

$$f(x,y)=c^{te} \tag{8.6.37}$$

defines a plan curve Γ therefore named "isocline". It is a curve such that when x varies, then y will thus not vary such that f remains constant.

We saw in the section of Differential and Integral Calculus that the differential of f, for any infinitesimal variations of x and y is:

$$\mathrm{d}f=\frac{\partial f}{\partial x}\mathrm{d}x+\frac{\partial f}{\partial y}\mathrm{d}y \tag{8.6.38}$$

Now, if we want when x varies of $\mathrm{d}x$, the value of the function f does not change, $\mathrm{d}y$ must not be anything but such that the variation of $\mathrm{d}f$ is zero. In other words:

$$\mathrm{d}f=\frac{\partial f}{\partial x}\mathrm{d}x+\frac{\partial f}{\partial y}\mathrm{d}y=0 \tag{8.6.39}$$

along Γ. But this equation is useless as such, but it fixed us the ratio of the derivative of the isoline in the plane such as:

$$\frac{\mathrm{d}y}{\mathrm{d}x}=-\frac{\dfrac{\partial f}{\partial x}}{\dfrac{\partial f}{\partial y}} \tag{8.6.40}$$

which gives us the slope of the tangent to Γ and thus after by integration, the desired function itself!!!

It goes without saying that the tangent vector to the curve Γ is a vector parallel to the one whose having for components (by correspondence with the above relation):

$$(\mathrm{d}x,\mathrm{d}y)=\left(\frac{\partial f}{\partial y},-\frac{\partial f}{\partial x}\right) \tag{8.6.41}$$

that we will denote by:

$$\vec{t}=\left(\frac{\partial f}{\partial y},-\frac{\partial f}{\partial x}\right) \tag{8.6.42}$$

Also recall that the gradient is given by (see section Vector Calculus page 1869):

$$\vec{\nabla}(f)=\left(\frac{\partial f}{\partial x},\frac{\partial f}{\partial y}\right) \tag{8.6.43}$$

We note that the latter two vectors are perpendicular (result which will be very useful in the section of Complex Analysis). Indeed, calculating the dot product (see section Vector Calculus page 1821):

$$\vec{t}\circ\vec{\nabla}(f)=\left(\frac{\partial f}{\partial y},-\frac{\partial f}{\partial x}\right)\circ\left(\frac{\partial f}{\partial x},\frac{\partial f}{\partial y}\right)=\frac{\partial f}{\partial x}\frac{\partial f}{\partial y}+\left(-\frac{\partial f}{\partial y}\right)\frac{\partial f}{\partial x}=0 \quad (8.6.44)$$

In other words, the vector $\vec{\nabla}(f)$ defines the orthogonal lines to the curve Γ.

✍**Example:**

Consider the equation of a special parabola in $\mathbb{R}^3$:

$$f(z)=\sqrt{x^2+y^2} \quad (8.6.45)$$

So we have the isoclines which are given by:

$$f(z)=\sqrt{x^2+y^2}=c^{te} \quad (8.6.46)$$

hence their equation in the plane:

$$y=\sqrt{c^{te}-x^2} \quad (8.6.47)$$

That is to say circles in the plane whose radius is equal to the square root of the corresponding selected constant at the height z height of f!

Let us now calculate the slope of the tangent Γ to these circles:

$$\frac{\mathrm{d}y}{\mathrm{d}x}=-\frac{\dfrac{\partial f}{\partial x}}{\dfrac{\partial f}{\partial y}}=-\frac{\dfrac{x}{\sqrt{x^2+y^2}}}{\dfrac{y}{\sqrt{x^2+y^2}}}=-\frac{x}{y}=-\frac{x}{\sqrt{c^{te}-x^2}} \quad (8.6.48)$$

This is consistent with the simple derivative of:

$$\frac{\mathrm{d}y}{\mathrm{d}x}=\frac{\mathrm{d}}{\mathrm{d}x}\sqrt{c^{te}-x^2}=-\frac{x}{\sqrt{c^{te}-x^2}} \quad (8.6.49)$$

We also have:

$$\vec{t}=\left(\frac{y}{\sqrt{c^{te}-x^2}},-\frac{x}{\sqrt{c^{te}-x^2}}\right) \quad (8.6.50)$$

We see that on $x=0$ this vector is equal to:

$$\vec{t}=(1,0) \quad (8.6.51)$$

which is intuitively consistent with the vector tangent to the circle that we have at this point of the x-axis.

8.6.3 Frenet Frame

One of the most important tools used to analyse a curve is the Frenet frame, a moving frame that provides a coordinate system at each point of the curve that is "best adapted" to the curve near that point.

The Frenet frame is a tool for the study of the local behaviour of the curves. More precisely, it is a local coordinate system associated with a point describing a curve Γ. Its construction method is different depending on the ambient space is 2-dimensional (plane curve) or 3-dimensional (left curve).

The Frenet frame and the Frenet formulas (giving the derivatives vectors of this frame) give the opportunity to conduct in a systematic way calculations of curvature and twisting of left curves and to introduce interesting geometric concepts associated with curves.

Let us first consider a curve Γ with its curvilinear abscissa $s(t)$ and $M_0=s(t_0)$ its origin. We denote by definition:

$$\vec{T}=\frac{\mathrm{d}\vec{f}(t)}{\mathrm{d}s}=\frac{\vec{f}(t+\mathrm{d}t)-\vec{f}(t)}{\mathrm{d}s}=\frac{\mathrm{d}\overrightarrow{\mathrm{O}M}}{\mathrm{d}s} \tag{8.6.52}$$

the tangent to the parametrized curve Γ of parameter t near a point M with respect to a frame put on O with $\mathrm{d}s$ that is calculated as we have shown previously.

It is interesting to note that if t is interpreted as begin the time (don't forget that this textbook is mainly target for engineering application), then we get a tangential speed:

$$\vec{v}=\vec{T}\frac{\mathrm{d}s}{\mathrm{d}t}=\vec{T}\|\vec{v}\|=\vec{T}v=\frac{\mathrm{d}\vec{f}(t)}{\vec{d}t} \tag{8.6.53}$$

and thus the $\vec{T}$ vector is obviously directed in the direction of movement.

Moreover, by construction and definition of the curvilinear abscissa, we always have:

$$\|\vec{T}\|=\frac{\mathrm{d}\overrightarrow{\mathrm{O}M}}{|\mathrm{d}s|} \tag{8.6.54}$$

and therefore the tangent vector $\vec{T}$ to the point M is unitary (and not zero!).

Now, without knowing exactly for what it will be useful for us now, let us look closer to the vector:

$$\frac{\vec{T}}{\mathrm{d}s} \tag{8.6.55}$$

Knowing trivially from the foregoing that:

$$\|\vec{T}\|=\vec{T}\circ\vec{T}=1 \tag{8.6.56}$$

Then we have:

$$\frac{\mathrm{d}\left(\vec{T}\circ\vec{T}\right)}{\mathrm{d}s}=\frac{\mathrm{d}\vec{T}}{\mathrm{d}s}\circ\vec{T}+\vec{T}\circ\frac{\mathrm{d}\vec{T}}{\mathrm{d}s}=2\vec{T}\circ\frac{\mathrm{d}\vec{T}}{\mathrm{d}s}=0 \tag{8.6.57}$$

therefore first $\mathrm{d}\vec{T}/\mathrm{d}s$ is a posteriori not unitary and $\vec{T}$ is perpendicular to it (results that will serve us several times afterwards so it must be remember!).

Therefore:

$$\vec{T}\circ\frac{\mathrm{d}\vec{T}}{\mathrm{d}s}=0 \tag{8.6.58}$$

Let us put:

$$\vec{N}=\frac{\dfrac{\mathrm{d}\vec{T}}{\mathrm{d}s}}{\left\|\dfrac{\mathrm{d}\vec{T}}{\mathrm{d}s}\right\|} \tag{8.6.59}$$

Given the previous result, $\vec{N}$ is the perpendicular vector, named unitary "curvature vector" at $\vec{T}$ on M. We say that this couple of vector $\left(\vec{T},\mathrm{d}\vec{T}/\mathrm{d}s\right)$ is "direct orthonormal" and C is by definition the "curvature".

We can approach the "curvature", denoted C, in a geometrically way. Let us see how!

We know at this step of our study that on a point M_0 of a curve Γ (differentiable at least once on every point - that is to say of class $\mathcal{C}^1$), there is a non-zero vector $\vec{T}$ which is tangent.

In any neighbouring point M (of curvilinear abscissa s), the tangent vector $\vec{T}$ can be written in approximation:

$$\vec{T}(s)\cong\vec{T}+\frac{\mathrm{d}\vec{T}}{\mathrm{d}s}\mathrm{d}s \tag{8.6.60}$$

if the curve is locally in the same plane (for now we are here studying the curvature and not the twisting of a curve)!

Two normal at M and M_0 thus intersecting in a point Ω, the following figure:

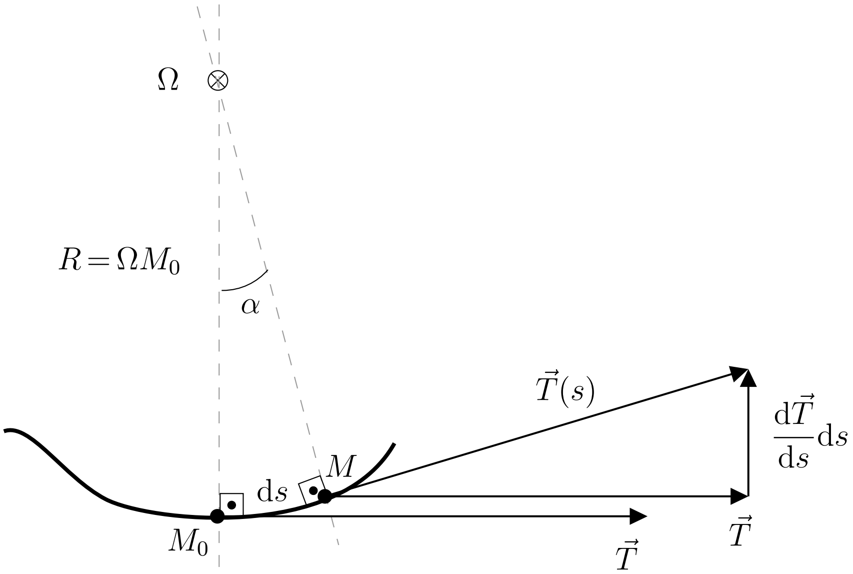


Figure 8.205: **Decomposition of the tangent vector of the curved path**

shows that at the first order in $\mathrm{d}s$, the point M can be considered locally as deduced from the point M_0 by a rotation of center Ω.

The circle thus defined of radius:

$$\overline{\Omega M_0} = R \tag{8.6.61}$$

is the one that best tangent the curve locally at the point M_0. Its radius is derived from the figure (two similar triangles at the limit):

$$\frac{\mathrm{d}s}{R} = \frac{\left\|\dfrac{\mathrm{d}\vec{T}}{\mathrm{d}s}\right\|\mathrm{d}s}{\|\vec{T}\|} \tag{8.6.62}$$

hence, since $\vec{T}$ is unitary, the definition of the value of the curvature for a plane curve:

$$\boxed{C = \frac{1}{R} = \left\|\frac{\mathrm{d}\vec{T}}{\mathrm{d}s}\right\|} \tag{8.6.63}$$

and that's it!

It is possible to interpret the concept of curvature as the speed of rotation of the Frenet frame relatively to a fixed direction.

The pair of vectors $(\vec{T},\vec{N})$ is named a "Frenet frame" and the basis vectors the "Frenet vectors".

The Frenet frame is a moving frame and the elements of this frame change depending on the point in question. In physics, we must not confuse this with the repository: as Frenet vectors move with the point!

Remark

The definition of C as above is true in the context of a choice of positive curvature. It's taken in a mechanical point of view but not necessary in pure mathematics.

If $C \neq 0$, then as seen above we can now write:

$$\vec{N} = \frac{\dfrac{\mathrm{d}\vec{T}}{\mathrm{d}s}}{\left\|\dfrac{\mathrm{d}\vec{T}}{\mathrm{d}s}\right\|} = \frac{\dfrac{\mathrm{d}\vec{T}}{\mathrm{d}s}}{C} = \frac{1}{C}\frac{\mathrm{d}\vec{T}}{\mathrm{d}s} = R\frac{\mathrm{d}\vec{T}}{\mathrm{d}s} \tag{8.6.64}$$

where R is named the "curvature radius". Depending on the situation (i.e. position on the curve) $\|\vec{T}\|$ can be bigger or smaller than the curvature radius R!

About the relation:

$$\boxed{\frac{\vec{N}}{R} = \frac{\mathrm{d}\vec{T}}{\mathrm{d}s}} \tag{8.6.65}$$

is named the "first Frenet formula" and shows obviously that $\vec{N}$ and $\mathrm{d}\vec{T}/\mathrm{d}s$ are collinear and hence their cross product is zero (result that will be used later).

These relations are justified by the analogy with mechanics. Indeed, remember that we have shown above that:

$$\vec{v}=\vec{T}\frac{\mathrm{d}s}{\mathrm{d}t}=\vec{T}\|\vec{v}\|=\vec{T}v=\frac{\mathrm{d}\vec{f}(t)}{\vec{d}t} \qquad (8.6.66)$$

Let us now calculate the acceleration:

$$\begin{aligned}\vec{a}&=\frac{\mathrm{d}\vec{v}}{\mathrm{d}t}=\frac{\vec{T}v}{\mathrm{d}t}=v\frac{\mathrm{d}\vec{T}}{\mathrm{d}t}+\vec{T}\frac{\mathrm{d}v}{\mathrm{d}t}=v\frac{\mathrm{d}s}{\mathrm{d}t}\frac{\mathrm{d}\vec{T}}{\mathrm{d}s}+\vec{T}\frac{\mathrm{d}v}{\mathrm{d}t}=v^2\frac{\mathrm{d}\vec{T}}{\mathrm{d}s}+\vec{T}\frac{\mathrm{d}v}{\mathrm{d}t}\\&=v^2C\vec{N}+\vec{T}\frac{\mathrm{d}v}{\mathrm{d}t}=\vec{T}\frac{\mathrm{d}v}{\mathrm{d}t}+\frac{v^2}{R}\vec{N}=\vec{a}_{\text{tg}}+\vec{a}_n\end{aligned} \qquad (8.6.67)$$

we fall back on the result obtained in the section of Classical Mechanics in our study of the osculating plane (see page 3237).

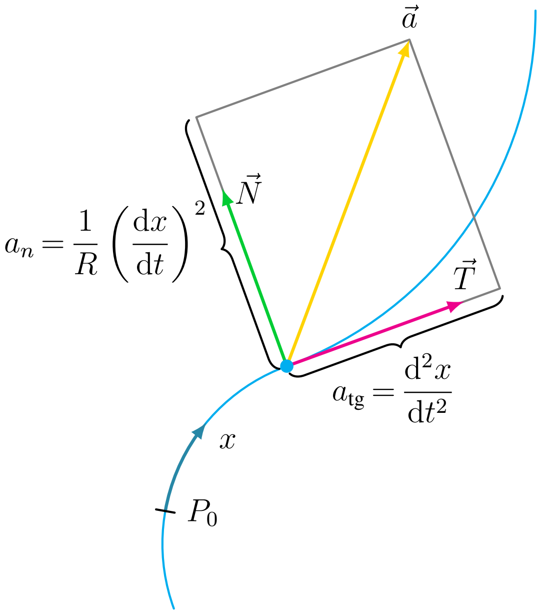

$\vec{a}$
$\vec{N}$
$a_n=\frac{1}{R}\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2$
$\vec{T}$
$a_{\text{tg}}=\frac{\mathrm{d}^2x}{\mathrm{d}t^2}$
x
P_0

Figure 8.206: **Tangential and Normal acceleration**

To give a more accurate geometric interpretation of the curvature, we define first through Ω the "center of curvature" of the "osculating circle" (placed in the osculating plane) or "curvature circle" of radius R that tangent locally the best the curve Γ in the Frenet frame:

$$\vec{\Omega}=\vec{M}+R\vec{N} \qquad (8.6.68)$$

To clarify geometrically what is the osculating circle, take a curve and a point M on that curve. Then draw the normal at this point of the locally plane curve equation and take a point Ω on the normal. Then, the center Ω of the circle passing through the point M is tangent to the curve. But all the circles tangent to the curve are not tangent in the same way! Indeed, if Ω is far from M, the circle will be located rather outside of the curve (blue and red circles in the figure below). If Ω is close to M, the circle will be located rather inside of the curve (yellow circle in the figure below). The limit radius between being "inside the curve" and being "outside the curve" is conventionally the "radius of curvature" we defined earlier above. The circle of that radius is then the famous "osculating circle"!

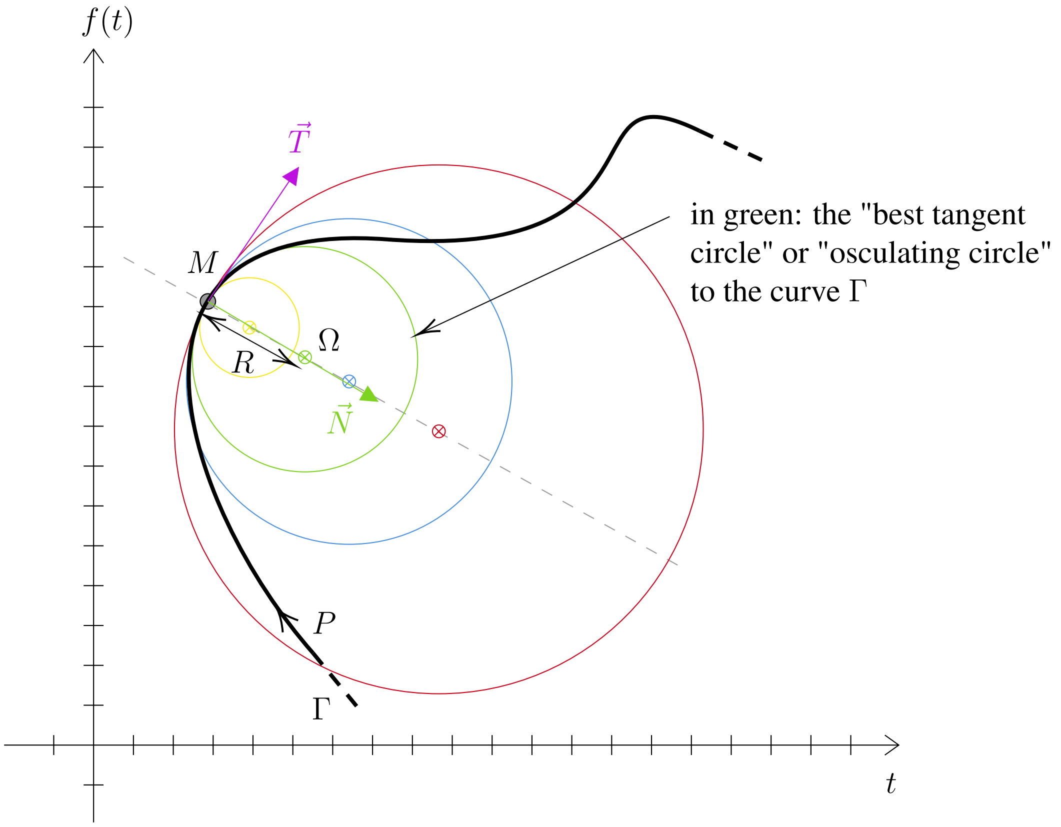

$f(t)$
$\vec{T}$
M
Ω
R
$\vec{N}$
P
Γ
t
in green: the "best tangent circle" or "osculating circle" to the curve Γ

Figure 8.207: Representation of an osculating circle

In the particular case where $\vec{T}$ is a constant vector we have obviously:

$$\frac{\mathrm{d}\vec{T}}{\mathrm{d}s}=\vec{0} \tag{8.6.69}$$

and therefore $C=0$ implying that R is no longer define. We sometimes say that in this case that the radius of curvature of Γ is infinite (a straight line then has a zero curvature at any point).

Let us now consider the vector perpendicular to the osculating plane defined and denoted by:

$$\vec{B}=\vec{T}\times\vec{N} \tag{8.6.70}$$

We can already say, since $\vec{T}$ and $\vec{N}$ are of unit norm that $\vec{B}$ is also of unit norm (which will serve us later)!

We can see that $\mathrm{d}\vec{B}/\mathrm{d}s$ is obviously orthogonal to $\vec{T}$. Indeed:

$$\frac{\mathrm{d}\vec{B}}{\mathrm{d}s}=\frac{\mathrm{d}\vec{T}}{\mathrm{d}s}\times\vec{N}+\vec{T}\frac{\mathrm{d}\vec{N}}{\mathrm{d}s}=\vec{0}\times\vec{N}+\vec{T}\frac{\mathrm{d}\vec{N}}{\mathrm{d}s}=\vec{T}\frac{\mathrm{d}\vec{N}}{\mathrm{d}s} \tag{8.6.71}$$

where we took the special case $\mathrm{d}\vec{T}/\mathrm{d}s=\vec{0}$ (but anyway in general $\mathrm{d}\vec{T}/\mathrm{d}s$ and $\vec{N}$ are collinear as we have prove it earlier and therefore the cross product of these two vectors is always zero).

As $\mathrm{d}\vec{B}/\mathrm{d}s$ is perpendicular to $\vec{T}$ it is therefore collinear to $\vec{N}$.

Therefore, as $\mathrm{d}\vec{T}/\mathrm{d}s$ is perpendicular to $\vec{T}$, $\mathrm{d}\vec{B}/\mathrm{d}s$ is perpendicular $\vec{B}$. This can be proven using exactly the same trick as when we have proved that $\mathrm{d}\vec{T}/\mathrm{d}s$ is perpendicular to $\vec{T}$.

Therefore:

$$\vec{B}\circ\frac{\mathrm{d}\vec{B}}{\mathrm{d}s}=0 \tag{8.6.72}$$

Let us put:

$$\boxed{\frac{\mathrm{d}\vec{B}}{\mathrm{d}s}=\frac{\vec{N}}{R'}=\tau\vec{N}} \tag{8.6.73}$$

This relation is the "second Frenet formula" where by definition, $\vec{B}$ is the "bi-normal vector" of Γ at the point M and τ the "twist" and R' the "torsion radius".

We can now establish the third Frenet formula. For this we start from:

$$\vec{B}=\vec{T}\times\vec{N}\Leftrightarrow\vec{N}=\vec{B}\times\vec{T} \tag{8.6.74}$$

from which we get:

$$\frac{\mathrm{d}\vec{N}}{\mathrm{d}s}=\frac{\mathrm{d}\vec{B}}{\mathrm{d}s}\times\vec{T}+\vec{B}\times\frac{\mathrm{d}\vec{T}}{\mathrm{d}s}=\frac{\vec{N}}{R'}\times\vec{T}+\vec{B}\times\frac{\vec{N}}{R} \tag{8.6.75}$$

But, by the properties of the cross product (see section Vector Calculus page 1830):

$$\vec{N}\times\vec{T}=-\vec{B}\qquad\vec{B}\times\vec{N}=-\vec{T} \tag{8.6.76}$$

hence the "third Frenet formula":

$$\boxed{\frac{\mathrm{d}\vec{N}}{\mathrm{d}s}=-\frac{\vec{B}}{R'}-\frac{\vec{T}}{R'}} \tag{8.6.77}$$

We name "Frenet trihedron" or "Frenet-Serret frame" associated to the curve Γ at point M, the natural orthonormal space frame $(M,\vec{T},\vec{N},\vec{B})$:

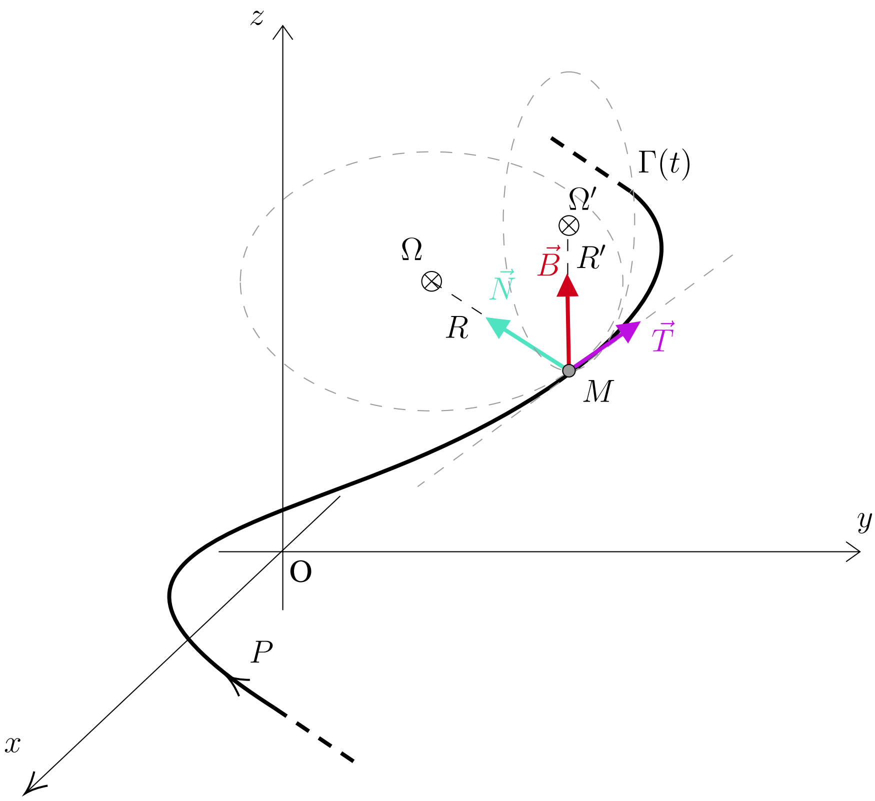


Figure 8.208: **Representation of the trihedron**

where, in mechanics, the vector $\vec{T}$ is collinear to the velocity and to tangential acceleration and $\vec{N}$ is collinear with the normal acceleration.

Remark

The radius of curvature R is in the osculating plane (the plane formed by the tangent and normal vector to the curve) which is the best plan in which is contained the curve. So, the radius of curvature gives in a point (locally) the best ("truest") radius of the curve. The twist gives us against the tendency of the curve of going out of the osculating plane (verbatim if the curve is contained in a plane, torsion is zero).

✍**Examples:**

E1. We consider the plane parametric equation of a parabola:

$$\Gamma(t)=\begin{pmatrix} t \\ t^2 \end{pmatrix} \tag{8.6.78}$$

Therefore we have:

$$\left(\frac{\mathrm{d}s}{\mathrm{d}t}\right)^2=\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2+\left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2=1+4t^2 \tag{8.6.79}$$

Hence:

$$\frac{\mathrm{d}s}{\mathrm{d}t}=\sqrt{1+4t^2} \tag{8.6.80}$$

Therefore we have:

$$\vec{T}=\frac{\mathrm{d}\overrightarrow{\mathrm{O}M}}{\mathrm{d}s}\frac{\mathrm{d}t}{\mathrm{d}s}\Rightarrow\vec{T}=\begin{pmatrix}\dfrac{1}{\sqrt{1+4t^2}}\\ \dfrac{1}{\sqrt{1+4t^2}}2t\end{pmatrix}\tag{8.6.81}$$

By the way you will notice that we have well:

$$\|\vec{T}\|=1\tag{8.6.82}$$

Thus, the curvature (inverse of the radius of curvature) is given by:

$$\begin{aligned}\frac{\mathrm{d}\vec{T}}{\mathrm{d}s}&=\frac{\mathrm{d}\overrightarrow{\mathrm{O}M}}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}s}=\begin{pmatrix}\left(-4\dfrac{t}{(1+4t^2)^{3/2}}\right)\dfrac{1}{\sqrt{1+4t^2}}\\ \left(\dfrac{2}{\sqrt{1+4t^2}}-8\dfrac{t^2}{(1+4t^2)^{3/2}}\right)\dfrac{1}{\sqrt{1+4t^2}}\end{pmatrix}\\ &=\begin{pmatrix}\dfrac{-4t}{(1+4t^2)^2}\\ \dfrac{2}{1+4t^2}-\dfrac{8t^2}{(1+4t^2)^2}\end{pmatrix}\frac{1}{R(t)}\Rightarrow\left\|\frac{\mathrm{d}\vec{T}}{\mathrm{d}s}\right\|=\sqrt{\frac{4}{(1+4t^2)^3}}=\frac{2}{(1+4t^2)^{3/2}}\end{aligned}\tag{8.6.83}$$

And therefore:

$$R(t)=\frac{(1+4t^2)^{3/2}}{2}\tag{8.6.84}$$

At $t=0$ the parabola has therefore a curvature $R(0)=0.5$. We have the vectors at $t=0$:

$$\vec{T}=\begin{pmatrix}1\\0\end{pmatrix}\qquad\frac{\mathrm{d}\vec{T}}{\mathrm{d}s}=\begin{pmatrix}0\\2\end{pmatrix}\qquad\vec{N}=R\frac{\mathrm{d}\vec{T}}{\mathrm{d}s}=\begin{pmatrix}0\\1\end{pmatrix}\tag{8.6.85}$$

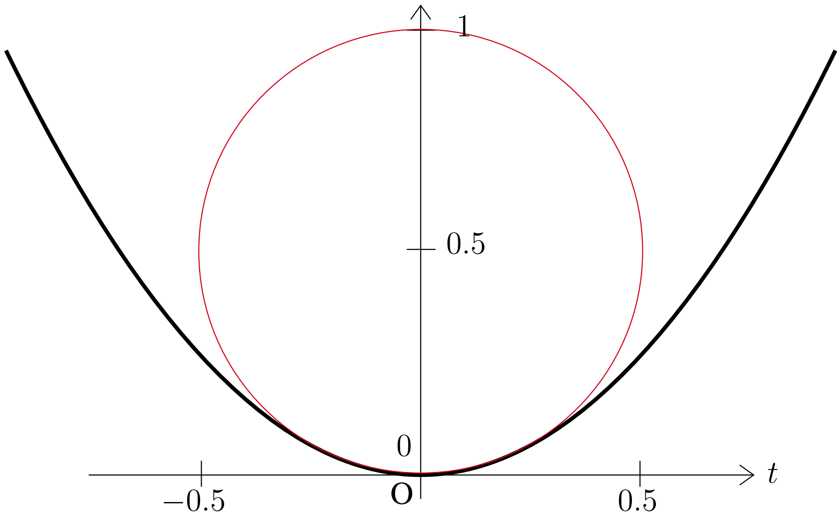


Figure 8.209: Osculating circle to parabola at point $t=0$

E2. Let us seek the radius and the center of curvature for any M of our helix defined above as practical example. Recall that its parametric function is given by:

$$\overrightarrow{\text{O}M}=\begin{pmatrix}x\\y\\z\end{pmatrix}=\begin{pmatrix}r\cos(t)\\r\sin(t)\\ht\end{pmatrix} \tag{8.6.86}$$

and that:

$$\dfrac{\mathrm{d}s}{\mathrm{d}t}=\sqrt{r^2+h^2} \tag{8.6.87}$$

Therefore we have:

$$\vec{T}=\dfrac{\mathrm{d}\overrightarrow{\text{O}M}}{\mathrm{d}s}=\dfrac{\mathrm{d}\overrightarrow{\text{O}M}}{\mathrm{d}t}\dfrac{\mathrm{d}t}{\mathrm{d}s}\Rightarrow\vec{T}=\begin{pmatrix}-\dfrac{r}{\sqrt{r^2+h^2}}\sin(t)\\\dfrac{r}{\sqrt{r^2+h^2}}\cos(t)\\\dfrac{h}{\sqrt{r^2+h^2}}\end{pmatrix} \tag{8.6.88}$$

On the way, the reader will have probably noticed that we have:

$$\|\vec{T}\|=1 \tag{8.6.89}$$

Thus, the curvature (the inverse of the radius of curvature) is given by:

$$\dfrac{\mathrm{d}\vec{T}}{\mathrm{d}t}=\dfrac{\mathrm{d}\vec{T}}{\mathrm{d}t}\dfrac{\mathrm{d}t}{\mathrm{d}s}=\begin{pmatrix}-\dfrac{r}{r^2+h^2}\cos(t)\\-\dfrac{r}{r^2+h^2}\sin(t)\\0\end{pmatrix}=\left\|\dfrac{\mathrm{d}\vec{T}}{\mathrm{d}s}\right\|=\dfrac{r}{r^2+h^2} \tag{8.6.90}$$

Therefore the radius of curvature is equal to:

$$R=\dfrac{r^2+h^2}{r} \tag{8.6.91}$$

This is consistent with intuition because when the step h of the helix is equal to zero, the radius or curvature is equal to r (the radius of the corresponding circle) and when h tends to infinity the radius of curvature tends to infinity and the curvature tends to zero. This example is a famous case in engineering applied to the exhaust chimney of smoke evacuation which are surrounded by a spiral of Archimedes and whose objective is to raise the air flows upward (the difficulty being to identify the radius R of the metal plate to cut that will follow the desired curvature at best... at least locally by knowing the radius of the chimney and the height h of the step of the helix):

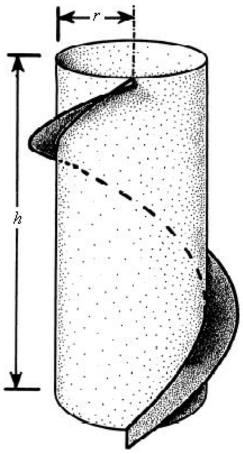


Figure 8.210: **Basic principle of an industrial chimney with a spiral (source: Frank Morgan, Riemmanian Geometry)**

Or the real corresponding version:

Figure 8.211: **Industrial evacuation chimney with spiral**

But in this engineering case, the height h should be obtained by a complete rotation. Therefore the radius of curvature will be written:

$$R = \frac{r^2 + \left(\dfrac{h}{2\pi}\right)^2}{r} \tag{8.6.92}$$

Anyway, to get back to our example and finish it, it comes with the first Frenet formula the following normal vector:

$$\vec{N} = \begin{pmatrix} -\cos(t) \\ -\sin(t) \\ 0 \end{pmatrix} \tag{8.6.93}$$

and for which the point (extremity of the vector) are indistinctness of the z-axis (merged with it) of our helix regardless the height h! The coordinate of the component z of this vector is zero since the normal is taken with respect to a point M of the curve already at a given implicit height h.

By the third Frenet formula with the binormal vector:

$$\begin{aligned}\vec{B}=\vec{T}\times\vec{N}&=\begin{pmatrix}-\dfrac{r}{\sqrt{r^2+h^2}}\sin(t)\\ \dfrac{r}{\sqrt{r^2+h^2}}\cos(t)\\ \dfrac{h}{\sqrt{r^2+h^2}}\sin(t)\end{pmatrix}\times\begin{pmatrix}-\cos(t)\\-\sin(t)\\0\end{pmatrix}\\ &=\frac{1}{\sqrt{r^2+h^2}}\begin{pmatrix}r\cos(t)\cdot 0-h(-\sin(t))\\ h(-\cos(t))-(-r\sin(t))\cdot 0\\ (-r\sin(t))(-\sin(t))-r\cos(t)(-\cos(t))\end{pmatrix}\end{aligned} \tag{8.6.94}$$

and the torsion radius is given by the relation:

$$\frac{\mathrm{d}\vec{B}}{\mathrm{d}s}=\frac{\vec{N}}{R'} \tag{8.6.95}$$

we therefore have:

$$\vec{N}=R'\frac{\mathrm{d}\vec{B}}{\mathrm{d}s}=R'\frac{\mathrm{d}\vec{B}}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}s} \tag{8.6.96}$$

And as we got the three following relations:

$$\vec{N}=\begin{pmatrix}-\cos(t)\\-\sin(t)\\0\end{pmatrix}\qquad\frac{\mathrm{d}\vec{B}}{\mathrm{d}t}=\begin{pmatrix}\dfrac{h}{\sqrt{r^2+h^2}}\cos(t)\\ \dfrac{h}{\sqrt{r^2+h^2}}\sin(t)\\0\end{pmatrix}\qquad\frac{\mathrm{d}t}{\mathrm{d}s}=\frac{1}{\sqrt{r^2+h^2}} \tag{8.6.97}$$

We deduce the torsion radius:

$$R'=-\frac{r^2+h^2}{h} \tag{8.6.98}$$

Let us now determine the important case of the explicit expression of the radius of curvature in Cartesian coordinates (result used in the section of Civil Engineering and useful in many other areas of physics!). Consider for this purpose the following figure:

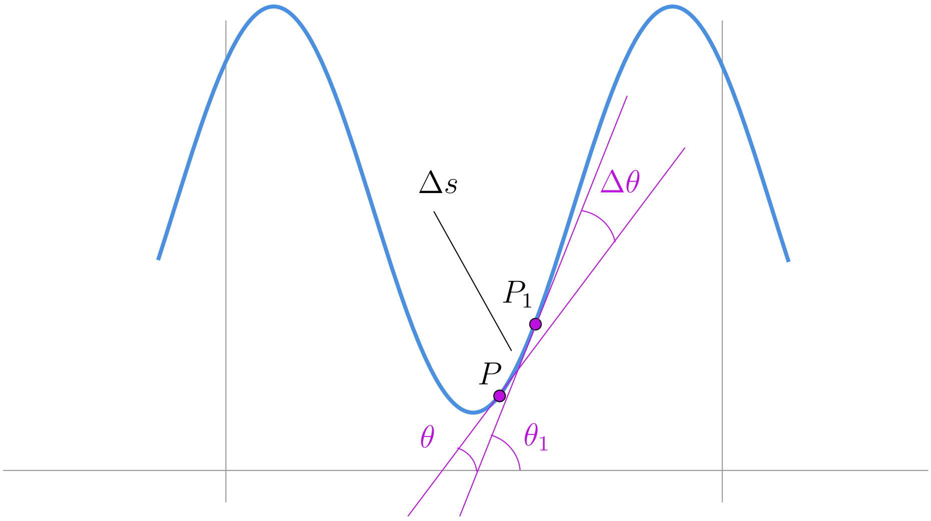


Figure 8.212: **Illustrated approach of the "1D" curvature expression**

So we have the radius of curvature which is intuitively given by the following relation if we do not make use of vector analysis:

$$C=\lim_{\Delta\to\theta}\frac{\Delta\theta}{\Delta s}=\frac{\mathrm{d}\theta}{\mathrm{d}\theta} \tag{8.6.99}$$

We also have:

$$\frac{\mathrm{d}\theta}{\mathrm{d}s}=\frac{\mathrm{d}\theta}{\mathrm{d}x}\frac{\mathrm{d}x}{\mathrm{d}s} \tag{8.6.100}$$

and as:

$$\tan(\theta)=\frac{\mathrm{d}y}{\mathrm{d}x} \tag{8.6.101}$$

therefore we have:

$$\theta=\arctan\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right) \tag{8.6.102}$$

and therefore:

$$\frac{\mathrm{d}\theta}{\mathrm{d}x}=\frac{\mathrm{d}}{\mathrm{d}x}\left(\arctan\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)\right) \tag{8.6.103}$$

We have proved in the section of Differential and Integral Calculus the following derivative (see page 1475):

$$\frac{\mathrm{d}}{\mathrm{d}x}\arctan(x)=\frac{1}{1+x^2} \tag{8.6.104}$$

We then have immediately by the composed derivatives:

$$\frac{\mathrm{d}\theta}{\mathrm{d}x}=\frac{\mathrm{d}}{\mathrm{d}x}\left(\arctan\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)\right)=\frac{1}{1+\left(\dfrac{\mathrm{d}y}{\mathrm{d}x}\right)^2}\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)=\frac{\dfrac{\mathrm{d}^2y}{\mathrm{d}x^2}}{1+\left(\dfrac{\mathrm{d}y}{\mathrm{d}x}\right)^2} \tag{8.6.105}$$

Having done this we also need $\mathrm{d}x/\mathrm{d}s$. But, we have already proved with different approaches in the section of Analytical Mechanics and Geometric Shapes (among others) by simply using Pythagoras theorem that:

$$\frac{\mathrm{d}x}{\mathrm{d}s}=\frac{1}{\dfrac{\mathrm{d}s}{\mathrm{d}x}}=\frac{1}{\sqrt{1+\left(\dfrac{\mathrm{d}y}{\mathrm{d}x}\right)^2}} \tag{8.6.106}$$

By grouping all we finally have:

$$C=\frac{\mathrm{d}\theta}{\mathrm{d}s}=\frac{\mathrm{d}\theta}{\mathrm{d}x}\frac{\mathrm{d}x}{\mathrm{d}s}=\frac{\dfrac{\mathrm{d}^2y}{\mathrm{d}x^2}}{1+\left(\dfrac{\mathrm{d}y}{\mathrm{d}x}\right)^2}\frac{1}{\sqrt{1+\left(\dfrac{\mathrm{d}y}{\mathrm{d}x}\right)^2}}=\frac{\dfrac{\mathrm{d}^2y}{\mathrm{d}x^2}}{\left(1+\left(\dfrac{\mathrm{d}y}{\mathrm{d}x}\right)^2\right)^{3/2}}=\frac{y''}{(1+y'^2)^{3/2}} \tag{8.6.107}$$

So it comes that the radius of the local osculating circle of a Cartesian function in the plane (by taking the absolute value of the second derivative to avoid having a negative radius...) is given by:

$$\boxed{R=\frac{1}{C}=\frac{\left(1+\left(\dfrac{\mathrm{d}y}{\mathrm{d}x}\right)^2\right)^{3/2}}{\left|\dfrac{\mathrm{d}^2y}{\mathrm{d}x^2}\right|}=\frac{(1+y'^2)^{3/2}}{|y''|}} \tag{8.6.108}$$

The last relation is often denoted in other textbooks as:

$$\kappa=\frac{|y''|}{(1+y'^2)^{3/2}} \tag{8.6.109}$$

Let us now derive another famous expression for the curvature based on the method and notation given in [323].

We start again from:

$$\begin{aligned}\kappa:=\frac{\mathrm{d}\phi}{\mathrm{d}s}=\frac{\dfrac{\mathrm{d}\phi}{\mathrm{d}t}}{\dfrac{\mathrm{d}s}{\mathrm{d}t}}&=\frac{\dfrac{\mathrm{d}\phi}{\mathrm{d}t}}{\sqrt{\left(\dfrac{\mathrm{d}x}{\mathrm{d}t}\right)^2+\left(\dfrac{\mathrm{d}y}{\mathrm{d}t}\right)^2}}\\&=\frac{\dfrac{\mathrm{d}\phi}{\mathrm{d}t}}{\sqrt{x'^2+y'^2}}\end{aligned} \tag{8.6.110}$$

where ϕ is the tangential angle and s is the arc length. The $\mathrm{d}\phi/\mathrm{d}t$ derivative in the above equation can be found using the identity:

$$\tan(\phi)=\frac{\mathrm{d}y}{\mathrm{d}x}=\frac{\mathrm{d}y/\mathrm{d}t}{\mathrm{d}x/\mathrm{d}t}=\frac{y'}{x'} \tag{8.6.111}$$

So:

$$\frac{\mathrm{d}}{\mathrm{d}t}\tan(\phi)=\sec^2(\phi)\frac{\mathrm{d}\phi}{\mathrm{d}t}=\frac{x'y''-y'x''}{x'^2} \tag{8.6.112}$$

Therefore:

$$\begin{aligned}\frac{\mathrm{d}\phi}{\mathrm{d}t}&=\frac{1}{\sec^2(\phi)}\frac{\mathrm{d}}{\mathrm{d}t}(\tan\phi)=\frac{1}{1+\tan^2(\phi)}\frac{x'y''-y'x''}{x'^2}\\&=\frac{1}{1+\frac{y'^2}{x'^2}}\frac{x'y''-y'x''}{x'^2}=\frac{x'y''-y'x''}{x'^2+y'^2}\end{aligned} \tag{8.6.113}$$

Hence:

$$\boxed{\kappa=\frac{x'y''-y'x''}{(x'^2+y'^2)^{3/2}}} \tag{8.6.114}$$

Let us derive now a third quite famous expression for the curvature based on the method and notation given in [2].

We start again from:

$$\hat{\vec{T}}:=\frac{\frac{\mathrm{d}\vec{r}}{\mathrm{d}t}}{\left\|\frac{\mathrm{d}\vec{r}}{\mathrm{d}t}\right\|}=\frac{\frac{\mathrm{d}\vec{r}}{\mathrm{d}t}}{\frac{\mathrm{d}s}{\mathrm{d}t}} \tag{8.6.115}$$

Therefore:

$$\begin{aligned}\frac{\mathrm{d}\vec{r}}{\mathrm{d}t}&=\frac{\mathrm{d}s}{\mathrm{d}t}\hat{\vec{T}}\\\frac{\mathrm{d}^2\vec{r}}{\mathrm{d}t^2}&=\frac{\mathrm{d}^2s}{\mathrm{d}t^2}\hat{\vec{T}}+\frac{\mathrm{d}s}{\mathrm{d}t}\frac{\mathrm{d}\hat{\vec{T}}}{\mathrm{d}t}\\&=\frac{\mathrm{d}^2s}{\mathrm{d}t^2}\hat{\vec{T}}+\kappa\hat{\vec{N}}\left(\frac{\mathrm{d}s}{\mathrm{d}t}\right)^2\end{aligned} \tag{8.6.116}$$

where $\hat{\vec{N}}$ is the normal vector. But:

$$\begin{aligned}\frac{\mathrm{d}\vec{r}}{\mathrm{d}t}\times\frac{\mathrm{d}^2\vec{r}}{\mathrm{d}t^2}&=\frac{\mathrm{d}s}{\mathrm{d}t}\frac{\mathrm{d}^2s}{\mathrm{d}t^2}(\hat{\vec{T}}\times\hat{\vec{T}})+\kappa\left(\frac{\mathrm{d}s}{\mathrm{d}t}\right)^3(\hat{\vec{T}}\times\hat{\vec{N}})=\frac{\mathrm{d}s}{\mathrm{d}t}\frac{\mathrm{d}^2s}{\mathrm{d}t^2}\vec{0}+\kappa\left(\frac{\mathrm{d}s}{\mathrm{d}t}\right)^3(\hat{\vec{T}}\times\hat{\vec{N}})\\&=\kappa\left(\frac{\mathrm{d}s}{\mathrm{d}t}\right)^3(\hat{\vec{T}}\times\hat{\vec{N}})\end{aligned} \tag{8.6.117}$$

so taking norms of both sides gives:

$$\left\|\frac{\mathrm{d}\vec{r}}{\mathrm{d}t}\times\frac{\mathrm{d}^2\vec{r}}{\mathrm{d}t^2}\right\|=\kappa\left(\frac{\mathrm{d}s}{\mathrm{d}t}\right)^3=\kappa\left\|\frac{\mathrm{d}\vec{r}}{\mathrm{d}t}\right\|^3 \tag{8.6.118}$$

Solving for κ then gives:

$$\kappa=\left\|\frac{\mathrm{d}\hat{\vec{T}}}{\mathrm{d}s}\right\|=\frac{\left\|\frac{\mathrm{d}\vec{r}}{\mathrm{d}t}\times\frac{\mathrm{d}^2\vec{r}}{\mathrm{d}t^2}\right\|}{\left\|\frac{\mathrm{d}\vec{r}}{\mathrm{d}t}\right\|^3}=\frac{\left\|\dot{\vec{r}}\times\ddot{\vec{r}}\right\|^3}{\left\|\dot{\vec{r}}\right\|^3} \tag{8.6.119}$$

Hence:

$$\boxed{\kappa = \frac{\left\| \dot{\vec{r}} \times \ddot{\vec{r}} \right\|^3}{\left\| \dot{\vec{r}} \right\|^3}} \tag{8.6.120}$$

8.6.4 Surface Patches

The study of surface patches is strongly related to the properties of parametric surfaces expression as seen at the end of the section of Analytical Geometry. We are therefore interested to calculate the tangent plane and the curvature of a surface on a given point. It is therefore an extension of the previous subjects. For example, the study of isolines of a surface that we have seen earlier can also be considered as belonging the field of Surface Patches study.

Now let us consider to begin $D \subset \mathbb{R}^2$:

$$g\begin{pmatrix} D \to \mathbb{R}^3 \\ (u,v) \to g(u,v) = M \end{pmatrix} \tag{8.6.121}$$

with:

$$\overrightarrow{\text{O}M} = \begin{pmatrix} x(u,v) \\ y(u,v) \\ z(u,v) \end{pmatrix} \tag{8.6.122}$$

such that $h(I) \subset D$.

We can define $g \circ h$:

$$g \circ h\begin{pmatrix} I \to \mathbb{R}^3 \\ t \to (g \circ h)(t) = g(u(t), v(t)) \end{pmatrix} \tag{8.6.123}$$

If we assume h continuous, it is clear that $g \circ h$ is a parametric arc. Let us denote by Γ its support, we have $\Sigma \subset \Sigma$ and we say that Σ a "plotted curve" or "inscribed curve" on Σ defined by the "Gauss Coordinates" u and v (already seen in the section of Non-Euclidean Geometry).

Remark

We always assume now that $D = I \times J$.

Given $M_0 \in \Sigma$, $M_0 = g(u_0, v_0)$. Let us look at the two curves drawn on Σ defined by the following parametrized arcs:

$$(g_{v_0}, I) \text{ with } g_{v_0}\begin{pmatrix} I \to \mathbb{R}^3 \\ u \to g_{v_0} = g(u, v_0) \end{pmatrix} \tag{8.6.124}$$

$$(g_{u_0}, J) \text{ with } g_{u_0}\begin{pmatrix} J \to \mathbb{R}^3 \\ v \to g_{u_0} = g(u_0, v) \end{pmatrix} \tag{8.6.125}$$

g_{u_0} and g_{v_0} are the two functions named "partial functions" of g on (u_0, v_0).

The support of (g_{u_0}, J) and (g_{v_0}, I) are named "coordinate-curves" of Σ on M_0 relatively to the parametrization (g, D). We denote them respectively Γ_{u_0} and Γ_{v_0} (see figure below). We also name Γ_{u_0} "first coordinate-curve" and Γ_{v_0} "second coordinate-curve".

It is of course quite obvious that (see section Differential and Integral Calculus page 1465) that:

$$\frac{\mathrm{d}g_{u_0}}{\mathrm{d}v} = \frac{\partial g}{\partial v} = \frac{\partial \overrightarrow{\text{O}M}}{\partial v} \tag{8.6.126}$$

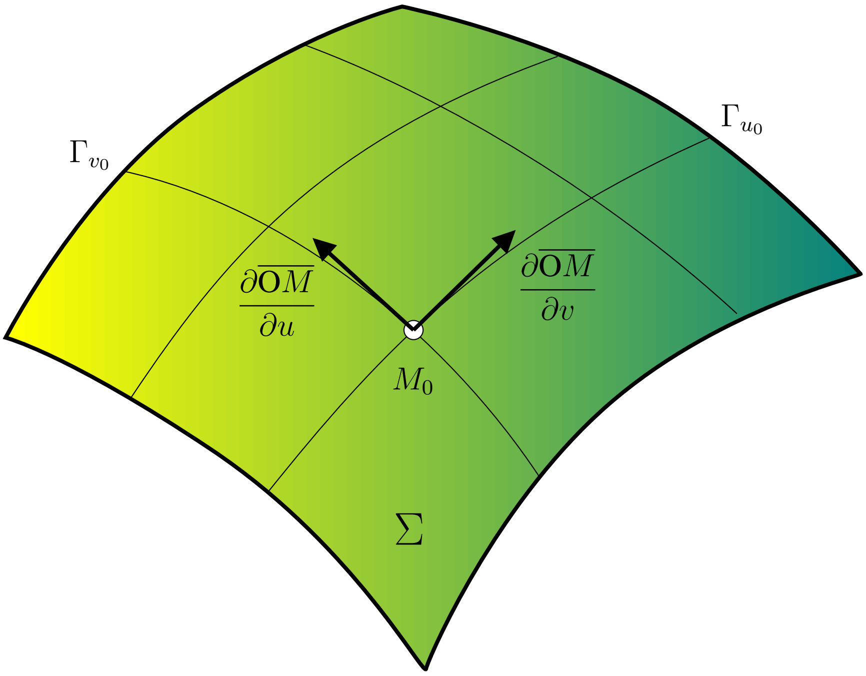
is tangent to Γ_{u_0} on M_0 and that $\partial\overrightarrow{\text{O}M}/\partial u$ is tangent to Γ_{v_0} on M_0.

As we have already prove it in the section of Tensor Calculus, this expression is independent of the surface patch as the infinitesimal element length $\mathrm{d}s$ is independent of the parametrization of Σ. This quadratic form is therefore an invariant that represents the metric of Σ. It is also denoted as follows by tradition:

$$\boxed{\mathrm{d}s^2 = g_{uu}\mathrm{d}u^2 + 2g_{uv}\mathrm{d}u\mathrm{d}v + g_{vv}\mathrm{d}2} \tag{8.6.127}$$

or even in more condensed form using tensor notation:

$$\mathrm{d}s^2 = g_{\mu\nu}\mathrm{d}x^\mu\mathrm{d}x^\nu \tag{8.6.128}$$



Γ_{v_0}
Γ_{u_0}
$\dfrac{\partial\overline{\text{O}M}}{\partial u}$
$\dfrac{\partial\overline{\text{O}M}}{\partial v}$
M_0
Σ

Figure 8.213: Representation of a surface patch

8.6.4.1 Metric of a Surface Patch

Given again:

$$g\begin{pmatrix} D \rightarrow \mathbb{R}^3 \\ (u,v) \rightarrow g(u,v) = M \end{pmatrix} \tag{8.6.129}$$

with:

$$\overrightarrow{\text{O}M} = \begin{pmatrix} x(u,v) \\ y(u,v) \\ z(u,v) \end{pmatrix} \tag{8.6.130}$$

Let us write $\mathrm{d}g = (\mathrm{d}x, \mathrm{d}y, \mathrm{d}z)$ or in other words:

$$\mathrm{d}\overrightarrow{\text{O}M} = \begin{pmatrix} \mathrm{d}x \\ \mathrm{d}y \\ \mathrm{d}z \end{pmatrix} \tag{8.6.131}$$

We also have (see section Differential and Integral Calculus page 1466):

$$\begin{aligned}\mathrm{d}x&=\frac{\partial x}{\partial u}\mathrm{d}u+\frac{\partial x}{\partial v}\mathrm{d}v\\ \mathrm{d}y&=\frac{\partial y}{\partial u}\mathrm{d}u+\frac{\partial y}{\partial v}\mathrm{d}v\\ \mathrm{d}z&=\frac{\partial z}{\partial u}\mathrm{d}u+\frac{\partial z}{\partial v}\mathrm{d}v\end{aligned} \tag{8.6.132}$$

and we have proved at the beginning of this section that the curvilinear abscissa in a Cartesian space (in Riemann coordinates Riemann) was given by:

$$\mathrm{d}s^2=\mathrm{d}x^2+\mathrm{d}y^2+\mathrm{d}z^2 \tag{8.6.133}$$

We therefore have after substitution in Gaussian coordinates and grouping terms of the same type:

$$\begin{aligned}\mathrm{d}s^2=&\left[\left(\frac{\partial x}{\partial u}\right)^2+\left(\frac{\partial y}{\partial u}\right)^2+\left(\frac{\partial z}{\partial u}\right)^2\right]^2\mathrm{d}u^2+2\left[\frac{\partial x}{\partial u}\frac{\partial x}{\partial v}+\frac{\partial y}{\partial u}\frac{\partial y}{\partial v}+\frac{\partial z}{\partial u}\frac{\partial z}{\partial v}\right]\mathrm{d}u\mathrm{d}v\\ &+\left[\left(\frac{\partial x}{\partial v}\right)^2+\left(\frac{\partial y}{\partial v}\right)^2+\left(\frac{\partial z}{\partial v}\right)^2\right]^2\mathrm{d}v^2\end{aligned} \tag{8.6.134}$$

Which is equivalent to write (take caution not to read there the square of a vector but it is the dot product of the vector with itself!):

$$\mathrm{d}s^2=\left(\frac{\partial\overrightarrow{\mathrm{O}M}}{\partial u}\right)^2\mathrm{d}u^2+2\frac{\partial\overrightarrow{\mathrm{O}M}}{\partial u}\frac{\partial\overrightarrow{\mathrm{O}M}}{\partial v}\mathrm{d}u\mathrm{d}v+\left(\frac{\partial\overrightarrow{\mathrm{O}M}}{\partial v}\right)^2\mathrm{d}v^2 \tag{8.6.135}$$

In a more traditional way with the notation:

$$E=\left(\frac{\partial\overrightarrow{\mathrm{O}M}}{\partial u}\right)^2\qquad F=\frac{\partial\overrightarrow{\mathrm{O}M}}{\partial u}\frac{\partial\overrightarrow{\mathrm{O}M}}{\partial v}\qquad G=\left(\frac{\partial\overrightarrow{\mathrm{O}M}}{\partial v}\right)^2 \tag{8.6.136}$$

we get a relation named the "first fundamental quadratic form" (we will note prove in this textbook the second one):

$$\mathrm{d}s^2=E\mathrm{d}u^2+2F\mathrm{d}u\mathrm{d}v+G\mathrm{d}v^2 \tag{8.6.137}$$

also named "first Gauss differential form". It is interesting to write this last relation in the form:

$$\mathrm{d}s^2=\frac{1}{E}(E\mathrm{d}u+F\mathrm{d}v)^2+\frac{EG-F^2}{E}\mathrm{d}v \tag{8.6.138}$$

and we see for $\mathrm{d}s^2$ to be positive, E and also $EG-F^2$ must be positive!

The first fundamental form may be also represented as a symmetric matrix:

$$\mathrm{d}s^2=\vec{u}^T\begin{bmatrix}E & G\\ F & G\end{bmatrix}\vec{v} \tag{8.6.139}$$

The first fundamental form is often written in the modern notation of the metric tensor. The coefficients may then be written as g_{ij}:

$$g_{ij}=\begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix}=\begin{bmatrix} E & F \\ F & G \end{bmatrix} \tag{8.6.140}$$

and we will prove in the section of Tensor Calculus during our study of Gram's determinant (see page 2054) that this relation can be used to calculate the surface of any (regular) surface patch!!

8.6.4.1.1 Regularity of a Surface

Definition (#8.6.363): A point M belonging to the surface Σ is said (this is quite relatively logical...) a "regular point" if and only if:

$$\frac{\partial\overrightarrow{\mathrm{O}M}}{\partial u}\times\frac{\partial\overrightarrow{\mathrm{O}M}}{\partial v}\neq\vec{0} \tag{8.6.141}$$

A surface patch is logically named "smooth surface" if and only if all its points are regular (if the cross product is zero then there is somewhere a "fold" at $\pi/2$... and this is quite annoying for continuity purposes).

Let us notice that:

$$\begin{aligned}
&\left\|\frac{\partial\overrightarrow{\mathrm{O}M}}{\partial u}\times\frac{\partial\overrightarrow{\mathrm{O}M}}{\partial v}\right\|^2=\left\|\begin{pmatrix}\dfrac{\partial x(u,v)}{\partial u}\\ \dfrac{\partial y(u,v)}{\partial u}\\ \dfrac{\partial u(u,v)}{\partial u}\end{pmatrix}\times\begin{pmatrix}\dfrac{\partial x(u,v)}{\partial v}\\ \dfrac{\partial y(u,v)}{\partial v}\\ \dfrac{\partial u(u,v)}{\partial v}\end{pmatrix}\right\|^2=\left\|\begin{pmatrix}\partial_u y\partial_v z-\partial_u z\partial_v y\\ \partial_u z\partial_v x-\partial_u x\partial_v z\\ \partial_u x\partial_v y-\partial_u y\partial_v x\end{pmatrix}\right\|^2\\
&=\left\|\sqrt{(\partial_u y\partial_v z-\partial_u z\partial_v y)^2+(\partial_u z\partial_v x-\partial_u x\partial_v z)^2+(\partial_u x\partial_v y-\partial_u y\partial_v x)^2}\right\|\\
&=(\partial_u y\partial_v z-\partial_u z\partial_v y)^2+(\partial_u z\partial_v x-\partial_u x\partial_v z)^2+(\partial_u x\partial_v y-\partial_u y\partial_v x)^2\\
&=(\partial_u y\partial_v z)^2-2\partial_u y\partial_v z\partial_u z\partial_v y+(\partial_u z\partial_v y)^2\\
&+(\partial_u z\partial_v x)^2-2\partial_u z\partial_v x\partial_u x\partial_v z+(\partial_u x\partial_v z)^2\\
&+(\partial_u x\partial_v y)^2-2\partial_u x\partial_v y\partial_u y\partial_v x+(\partial_u y\partial_v x)^2\\
&=\ldots\text{boring to do algebra}\ldots=EG-F^2
\end{aligned} \tag{8.6.142}$$

The angle α between the two coordinate-curves Γ_{u_0} and Γ_{v_0} to Σ on M_0 is given by the product definition of the dot product:

$$\frac{\partial\overrightarrow{\mathrm{O}M}}{\partial u}\circ\frac{\partial\overrightarrow{\mathrm{O}M}}{\partial v}=\left\|\frac{\partial\overrightarrow{\mathrm{O}M}}{\partial u}\right\|\left\|\frac{\partial\overrightarrow{\mathrm{O}M}}{\partial v}\right\|\cos(\alpha)=F=0 \tag{8.6.143}$$

Hence the expression:

$$\cos(\alpha)=\frac{\dfrac{\partial\overrightarrow{\mathrm{O}M}}{\partial u}\circ\dfrac{\partial\overrightarrow{\mathrm{O}M}}{\partial v}}{\left\|\dfrac{\partial\overrightarrow{\mathrm{O}M}}{\partial u}\right\|\left\|\dfrac{\partial\overrightarrow{\mathrm{O}M}}{\partial v}\right\|}=\frac{\dfrac{\partial\overrightarrow{\mathrm{O}M}}{\partial u}\circ\dfrac{\partial\overrightarrow{\mathrm{O}M}}{\partial v}}{\sqrt{\left(\dfrac{\partial\overrightarrow{\mathrm{O}M}}{\partial u}\right)^2\left(\dfrac{\partial\overrightarrow{\mathrm{O}M}}{\partial v}\right)^2}}=\frac{F}{\sqrt{EG}} \tag{8.6.144}$$

Therefore a necessary and sufficient condition for the coordinates-curves Γ_{u_0} and Γ_{v_0} to be perpendicular to Σ on M_0 is that F is equal to zero. In this particular case, we say that the curvilinear coordinates u, v on the surface are orthogonal coordinates!!!

✍**Examples:**

E1. Let us consider the parametrization of the cartesian plane. Then we have:

$$\overrightarrow{\text{O}M}=\begin{pmatrix}x\\y\\0\end{pmatrix}\Rightarrow\begin{cases}\dfrac{\partial\overrightarrow{\text{O}M}}{\partial x}=\begin{pmatrix}1\\0\\0\end{pmatrix}\\[2ex]\dfrac{\partial\overrightarrow{\text{O}M}}{\partial y}=\begin{pmatrix}0\\1\\0\end{pmatrix}\end{cases}\tag{8.6.145}$$

hence:

$$E=G=1,F=0\tag{8.6.146}$$

Therefore:

$$\mathrm{d}s^2=\mathrm{d}x^2+\mathrm{d}y^2\tag{8.6.147}$$

We fall back on the same differential curvilinear abscissa than that seen in the section of Tensor Calculus and General Relativity with the diagonal metric of the flat space.

We also have:

$$EG-F^2=1\neq 0\tag{8.6.148}$$

Then the surface is indeed regular.

We also have:

$$\cos(\alpha)=\frac{F}{\sqrt{EG}}=\frac{0}{\sqrt{1}}=0\Rightarrow\alpha=\pm\frac{\pi}{2}\tag{8.6.149}$$

So the both coordinate-curves are perpendicular!

E2. Let us consider the parametrization of the cylinder surface. We then have (see section Vector Calculus page 1861):

$$\overrightarrow{\text{O}M}=\begin{pmatrix}r\cos(\phi)\\r\sin(\phi)\\z\end{pmatrix}\Rightarrow\begin{cases}\dfrac{\partial\overrightarrow{\text{O}M}}{\partial r}=\begin{pmatrix}\cos(\phi)\\\sin(\phi)\\0\end{pmatrix}\\[2ex]\dfrac{\partial\overrightarrow{\text{O}M}}{\partial \phi}=\begin{pmatrix}-r\sin(\phi)\\r\cos(\phi)\\0\end{pmatrix}\end{cases}\tag{8.6.150}$$

hence:

$$E=1,G=r^2,F=0\tag{8.6.151}$$

Therefore:

$$\mathrm{d}s^2=\mathrm{d}r^2+r^2\mathrm{d}\phi^2\tag{8.6.152}$$

We fall back on the same differential curvilinear abscissa than that seen in the section of Tensor Calculus and General Relativity with the diagonal metric in polar coordinates.

We also have:

$$EG-F^2=r^2\neq 0 \tag{8.6.153}$$

Therefore the surface is regular as long as r is not zero. We also have:

$$\cos(\alpha)=\frac{F}{\sqrt{EG}}=\frac{0}{\sqrt{r^2}}\Rightarrow\alpha=\pm\frac{\pi}{2} \tag{8.6.154}$$

So the both coordinates-curves are perpendicular on the cylinder.

E3. Let us consider the parametrization of the sphere of origin O and radius r . Then we have (see section Vector Calculus page 1849):

$$\overrightarrow{\text{O}M}=\begin{pmatrix}r\sin(\theta)\cos(\phi)\\ r\sin(\theta)\sin(\phi)\\ r\cos(\theta)\end{pmatrix}\Rightarrow\begin{cases}\dfrac{\partial\overrightarrow{\text{O}M}}{\partial\theta}=\begin{pmatrix}r\cos(\theta)\cos(\phi)\\ r\cos(\theta)\sin(\phi)\\ -r\sin(\theta)\end{pmatrix}\\ \dfrac{\partial\overrightarrow{\text{O}M}}{\partial\phi}=\begin{pmatrix}-r\sin(\theta)\sin(\phi)\\ r\sin(\theta)\cos(\phi)\\ 0\end{pmatrix}\end{cases} \tag{8.6.155}$$

hence:

$$E=r^2,G=r^2\sin^2(\theta),F=0 \tag{8.6.156}$$

Therefore:

$$\mathrm{d}s^2=r^2\mathrm{d}\theta^2+r^2\sin^2(\theta)\mathrm{d}\phi^2 \tag{8.6.157}$$

We fall back on the same differential curvilinear abscissa than that seen in the section of Tensor Calculus and General Relativity with the diagonal metric in spherical coordinates.

We also have:

$$EG-F^2=r^4\sin^2(\theta)\neq 0 \tag{8.6.158}$$

Therefore the surface is regular as long as r and θ are not zero. We also have:

$$\cos(\alpha)=\frac{F}{\sqrt{EG}}=\frac{0}{\sqrt{r^4\sin^2(\theta)}}\Rightarrow\alpha=\pm\frac{\pi}{2} \tag{8.6.159}$$

So the both coordinates-curves are perpendicular on the sphere.

E4. Let us consider the parametrization of the hyperboloid. Then we have (see section Analytical Geometry page 2841):

$$\overrightarrow{\mathrm{O}M}=\begin{pmatrix}a\sqrt{1+t^2}\cos(u)\\ a\sqrt{1+t^2}\sin(u)\\ bu\end{pmatrix}\Rightarrow\begin{cases}\dfrac{\partial\overrightarrow{\mathrm{O}M}}{\partial t}=\begin{pmatrix}a\dfrac{a\cos(u)t}{\sqrt{1+t^2}}\\ a\dfrac{a\sin(u)t}{\sqrt{1+t^2}}\\ 0\end{pmatrix}\\ \dfrac{\partial\overrightarrow{\mathrm{O}M}}{\partial u}=\begin{pmatrix}-a\sqrt{1+t^2}\sin(u)\\ a\sqrt{1+t^2}\cos(u)\\ b\end{pmatrix}\end{cases}\tag{8.6.160}$$

hence:

$$E=a^2\frac{t^2}{1+t^2},G=a^2(1+t^2)+b^2,F=0\tag{8.6.161}$$

Therefore:

$$\mathrm{d}s^2=a^2\frac{t^2}{1+t^2}\mathrm{d}t^2+(a^2(1+t^2)+b^2)\mathrm{d}u^2\tag{8.6.162}$$

Let us take b as being equal to zero. Then we have:

$$EG-F^2=a^4t^2\tag{8.6.163}$$

Therefore the surface is regular as long as a not equal to zero. We also have:

$$\cos(\alpha)=\frac{F}{\sqrt{EG}}=\frac{0}{\sqrt{a^4t^2}}\Rightarrow\alpha=\pm\frac{\pi}{2}\tag{8.6.164}$$

So the both coordinates-curves are perpendicular on the hyperboloid.

~80 %

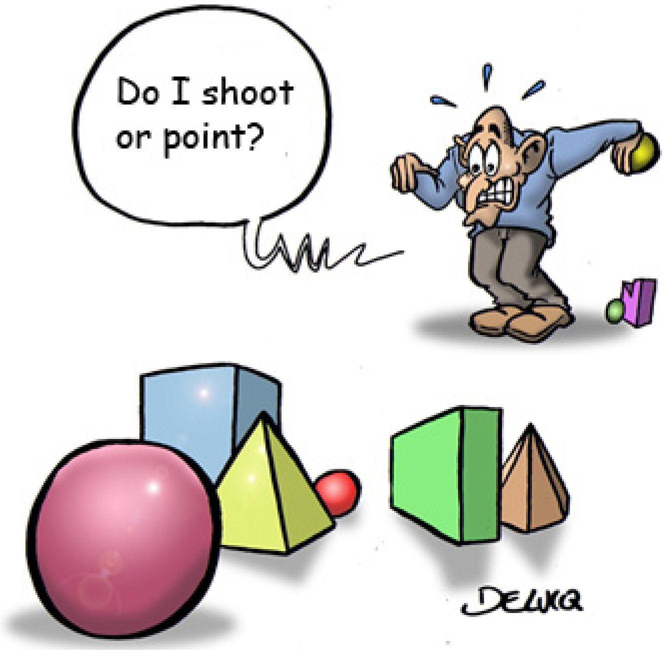
★★★★☆

35 votes, 74.86%

8.7. Geometric Shapes

We have already defined at the beginning of the section on Euclidean Geometry the concepts of topological dimensions, what were a zero-dimensional point and a unit dimensional curve. We will not come back on these concepts we will focus here mostly on shapes of higher dimensions that we will use a lot in the chapters related to theoretical Physics.

The purpose of section is to categorize with proofs some remarkable mathematical properties of well-known geometric shapes and bodies (area, volume, center of mass, moment of inertia). Indeed, there are many mathematical books listing those properties without proofs but few or simply no books at all prove them all (at least we have never seen such a book at this day...). The list below is at this date far away to exhaustive (since there are an infinite number of geometric shapes), but it will be completed in time.


Do I shoot or point?
DELUKA

The few shapes that we wanted to present in this section allow easily to find the remarkable properties of a large number of shapes not listed in this section by assembly or decomposition.

Remarks

R1. The trigonometric relations and remarkable integrals that will be used for the study of the geometric shapes below are not proven in this section. They all have already been proved in the sections dealing specifically with Trigonometry (see page 2503) and Differential Calculus and Integral (see page 1450).

R2. We understand by "center of gravity", the "barycenter" as discussed in the chapter of Euclidean Geometry (see page 2635).

Here are some typical areas scales (the years are given relatively to the holocene calendar):

Magnitude $[\text{m}^2]$	**Description**
$2.6\cdot 10^{-70}$	Planck area $\frac{G\hbar}{c^3}$
$66.52\cdot 10^{-29}$	Thomson cross-section of the electron
$1\cdot 10^{-28}$	1 barn, roughly the cross-sectional area of a uranium nucleus
$7.17\cdot 10^{-9}$	A dot printed using 300 dots per inch resolution
$1\cdot 10^{-8}$	A red blood cell
$1.57\cdot 10^{-8}$	Cross-sectional area of a straight human hair
$2.4\cdot 10^{-8}$	Common LCD screen pixel sizes
$5.65\cdot 10^{-7}$	Cross-sectional area of a mechanical pencil lead
$1.5\cdot 10^{-7}$	Typical for one of the 20 standard amino acids
$1\cdot 10^{-6}$	Head of a pin
$6\cdot 10^{-6}$	E. coli bacterium
$4\cdot 10^{-5}$	Typical hole punched in a piece of paper by a hole punch
$2.9\cdot 10^{-4}$	One side of a U.S. penny
$5\cdot 10^{-4}$	Typical postage stamp
$1.1\cdot 10^{-3}$	Average human retina
$4.6\cdot 10^{-3}$	Face of a typical credit card
$6.237\cdot 10^{-2}$	International A4 paper (210×297 [mm])
0.125	International A3 paper (297×420 [mm])
0.25	International A2 paper (420×594 [mm])
0.5	International A1 paper (594×841 [mm])
1	International A0 paper (594×1189 [mm])
1-4	Top of an office desk
10-20	A typical parking space for tourists cars
100	One are
261	Size of a tennis court
845	Wing area of Airbus A380, the largest commercial airliner
978	Primary mirror of the Extremely Large Telescope, the largest optical telescope in the world (under construction)
$1,250$	Water in an Olympic-size swimming pool
$4,047$	1 acre
$7,140$	Typical football (soccer) field
$10,000$	1 hectare
$53,000$	Base of the Great Pyramid of Giza
$1.05\cdot 10^{8}$	Paris (inner city only)
$7.74\cdot 10^{12}$	Australia
$1.687\cdot 10^{14}$	Pacific Ocean
$5.1\cdot 10^{14}$	Earth
$6.1\cdot 10^{16}$	Jupiter
$6.1\cdot 10^{42}$	Approximate actual area of Milky Way's galactic disk
10^{54}	Approximate actual area the observable Universe

Table 8.6: **Some common areas scales**

8.7.1 Usual Surfaces (areas)

There are several definitions of the concept of surface (area) due to Euclid and another modern due to the topology (see sections of the corresponding names).

Definitions (#8.7.364):

D1. A "plane surface" is a closed area that can be identify almost approximatively by a height and a width.

D2. A "surface" is a topological variety of dimension 2.

Depending on the authors a surface or area is denoted by the symbols: $A, S, \mathcal{A}, \mathcal{S}, \Sigma$.

Remark

We will focus initially only on the properties (perimeter, area, center of gravity, etc.) of surfaces dived in Euclidean plane and space.

8.7.1.1 Polygons

Definition (#8.7.365): A "polygon" is a plane figure bounded by straight line segments (i.e.: with a closed polyline).

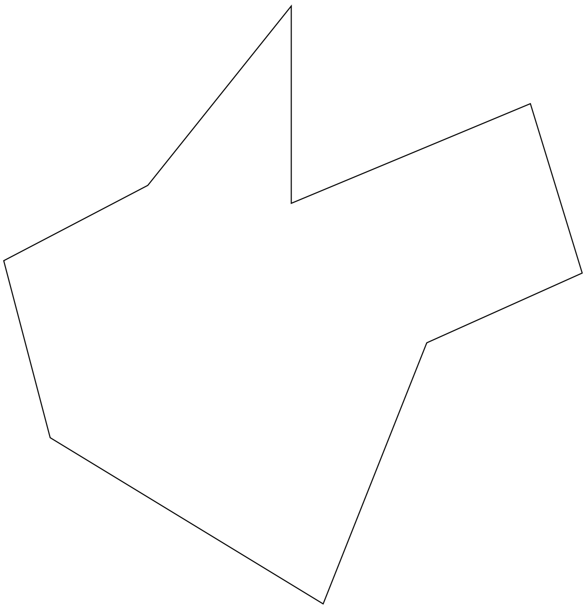


Figure 8.214: **Example of an arbitrary plane polygon**

Definition (#8.7.366): A "quadrilateral", "pentagon", "hexagon", "heptagon" are polygons with respectively four, five, six, seven... sides.

We distinguish three main families (but they are not the only ones!) of polygons:

1. Crossed polygons
2. Concave polygons (non-convex)
3. Convex polygons

We will find these two last families in different sections of this textbook (but for details see below).

Before going any further we would like to clarify to the reader that there is, at least as far as we know, only mathematical relations to calculate the surface for simple polygons. Although in practice we almost always meet non simple polygons and we considered as useless to dwell on the determination of a relation that would reduce the calculation of the surface of any polygon to that of simple polygons or to use Monte Carlo methods (see section of Numerical Methods page 6708).

Definition (#8.7.367): A polygon is named "crossed polygon" if at least two of its sides intersect, that is to say, if at least two of its sides cross each other.

This is the case of the polygone $ABCDE$ below:

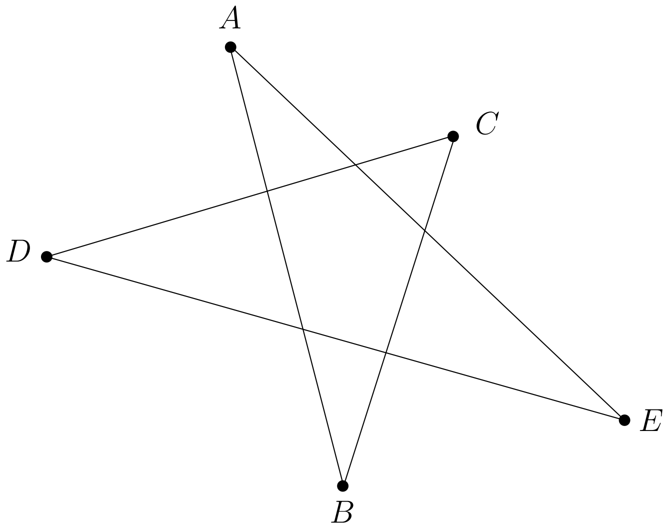


Figure 8.215: **Example of a crossed polygon**

Remark

The "envelope" of a polygon is the polygon obtained by following its outer contour. For example, the envelope of the previous polygon is a decagon whose vertices are the five summits of the pentagon and the five intersections of its sides.

Definition (#8.7.368): A polygon is named "concave polygon" if it is not crossed and if one or more of its diagonals are not entirely within the area bounded by the polygon.

This is the case of the pentagon $ABCDE$ below:

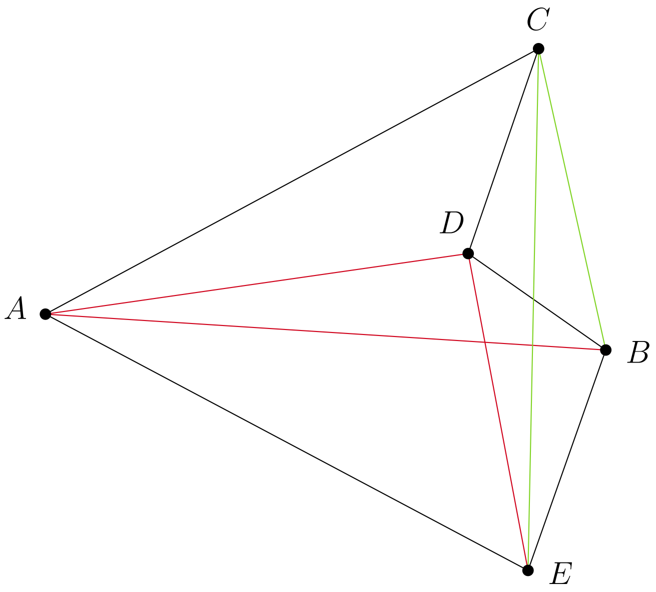


Figure 8.216: **Example of a concave polygon**

Definition (#8.7.369): A polygon is named "convex polygon" if it is not crossed and if all its diagonals are entirely within the area bounded by the polygon.

The hexagon $MNOPQR$ below is an example of convex polygon:

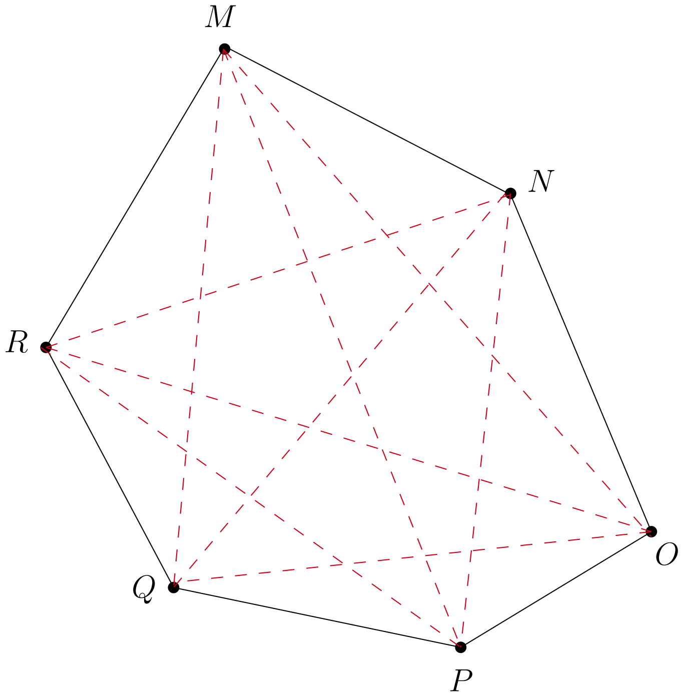


Figure 8.217: **Example of a convex polygon**

With respect to the definitions given above where the diagonals were identified, let us see if there is a relation permitting to know the number of diagonals relatively to the number of edges of a polygon.

For this let us start from a polygon of n sides (also note that it has n vertices):

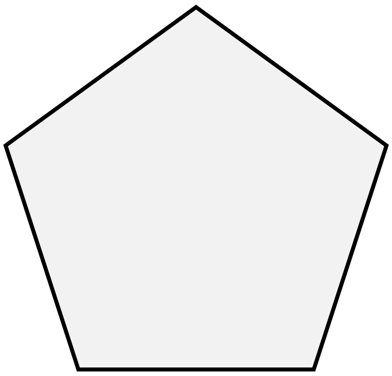


Figure 8.218: **Example of a polygon (pentagon with 5 edges and 5 vertices)**

We define now the total number segments S as equal to the quantity of the sides (edges) E more the quantity of diagonals D such that:

$$S = E + D \tag{8.7.1}$$

With our example we can see that $S = 10$:

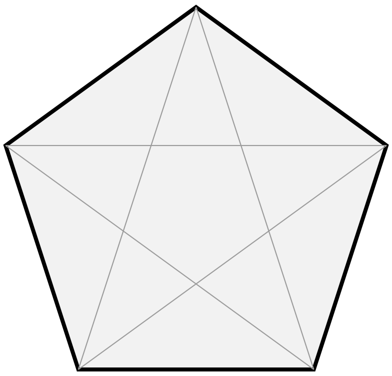


Figure 8.219: **Pentagon with all its 10 segments**

Now take one first point of our pentagon. We see that we can reach all n points (vertices), excepted the chosen one (-1) therefore we can draw $n-1$ segments as shown in the figure below:

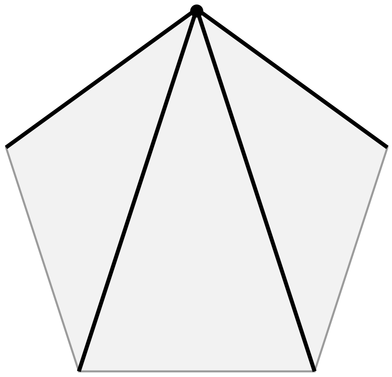


Figure 8.220: Example of segments starting from a chosen point

With the second point, we can also reach all the points n, except the chosen point itself (-1) and we must remove the first point already taken in the previous step (-1). We therefore draw $n-2$ segments:

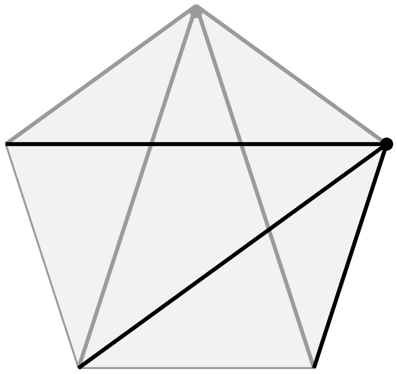


Figure 8.221: Approach with the 2nd point

With the third one, we can also reach all the n points, excepted the considerated point (-1) and less the two points already seen (-2) that is to say the formation of $n-3$ segments:

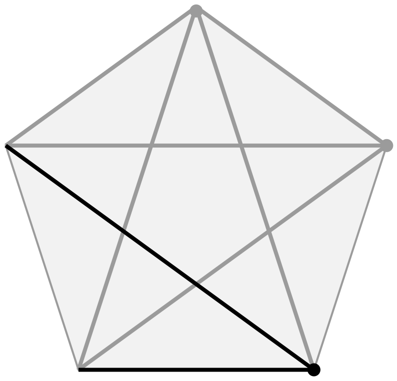


Figure 8.222: Approach with the 3rd point

We continue with the other points: the 4th that gives $n-4$ segments, the fifth that gives $n-5$ segments ... Therefore we see that the $(n-2)$th item thus gives $n-(n-2)$ segments, etc.

So we finally have for:

$$\begin{array}{r@{}l} s &= 1+2+\ldots+(n-2)+(n-1) \\ +s &= (n-1)+(n-2)+\ldots+2+1 \\ \hline 2s &= n+n+\ldots+n+n \end{array} \qquad (8.7.2)$$

By simplifying we find therefore:

$$s=\frac{n(n-1)}{2} \qquad (8.7.3)$$

We find ourselves with two relations:

$$s=\frac{n(n-1)}{2} \quad \text{and} \quad s=n+d \qquad (8.7.4)$$

Therefore it follows that:

$$\boxed{d=s-n=\frac{n(n-1)}{2}-n=\frac{n(n-3)}{2}} \qquad (8.7.5)$$

Definition (#8.7.370): A "regular polygon" is a convex-polygon in which every side has the same length and every angle is the same. For any natural number $n \geq 3$, we can draw a regular polygon with n sides; sometimes we'll name this polygon a "regular n-gon" to emphasize that it has n sides. Below are the regular n-gons for $n=3,4,\ldots,12$. As n gets larger, the regular n-gons look more like a perfect circle:

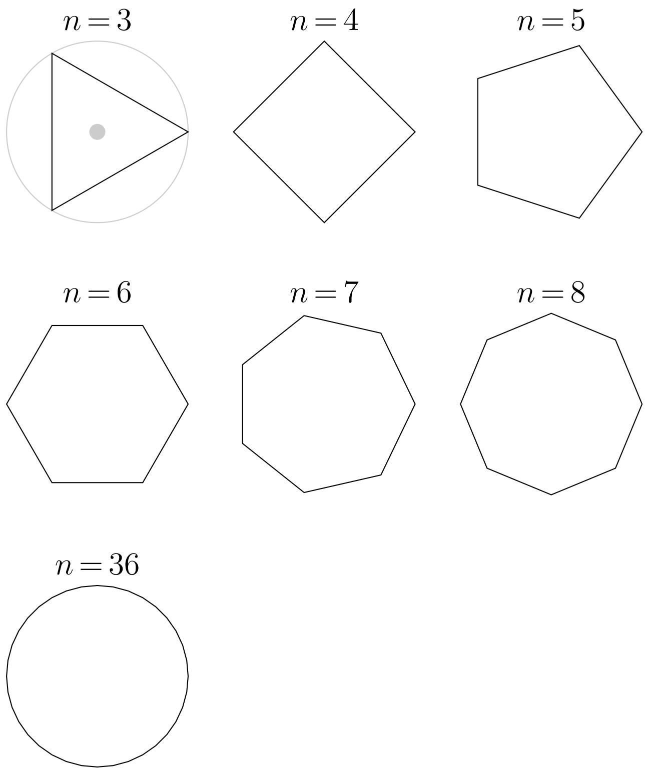


Figure 8.223: **Some regular polygons (subfamily of convex polygons)**

8.7.1.2 Rectangle

Definition (#8.7.371): The "rectangle" is a special case of the quadrangle in the sense that his sides L and H (notation according to length and height as we can see in figure below) are equal in pairs and with right angle (in other words, L is not necessarily equal to H).

Other possible definitions include that a rectangle is a parallelogram with a right angle or a quadrilateral with four right angles...

> **Remark**
>
> The rectangle can be seen as the composition of two (or more) triangles (see below for definition). To build a rectangle, it would be sufficient to have a single rectangle triangle and make him undergo a double reflection and rotation relative to a well-chosen axis (see section Euclidean Geometry page 2640).

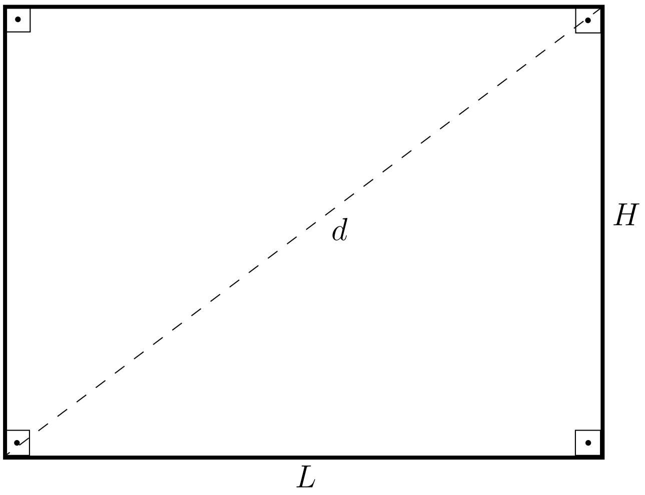


Figure 8.224: **Example of rectangle**

By the Euclid's axioms (see section Euclidean Geometry page 2563), the perimeter of a rectangle is given by:

$$\boxed{P=2(L+H)} \tag{8.7.6}$$

And by definition its surface by:

$$\boxed{S=L\cdot H} \tag{8.7.7}$$

and the length of its diagonal (application of the Pythagorean theorem):

$$d=\sqrt{L^2+H^2} \tag{8.7.8}$$

The position of the rectangle's center of gravity, if we put the Cartesian coordinate system in the lower left corner of this geometry, is trivially given by:

$$G(x,y)=\left(\frac{L}{2},\frac{H}{2}\right) \tag{8.7.9}$$

Finally, indicate that if we were living beings in a two-dimensional space, the rectangle is what we should perceive if a cuboid crossed our universe parallel to its faces.

Finally, let us notice, and without proof, that the sum of the angles or a plane rectangle is equal to $4\cdot\pi/2=2\pi$.

8.7.1.3 Square

Definition (#8.7.372): The "square" is a special case of the rectangle in the sense that its four sides are equal such that $L=H$.

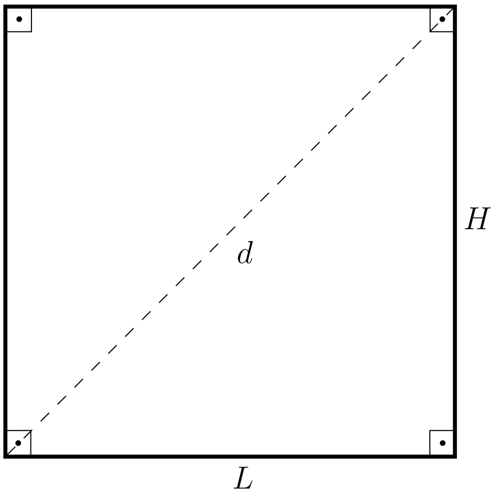


Figure 8.225: **Example of square**

By the Euclid's axioms (see section Euclidean Geometry page 2563), the perimeter of the square is given by:

$$\boxed{P=4L} \tag{8.7.10}$$

So it comes for the surface:

$$\boxed{S=L^2} \tag{8.7.11}$$

and for its diagonal:

$$d=\sqrt{L^2+H^2}=\sqrt{L^2+L^2}=\sqrt{2L^2}=\sqrt{2}L \tag{8.7.12}$$

The position of the square's center of gravity, if put the Cartesian coordinate system in the lower left corner of the form, is trivially given by:

$$G(x,y)=\left(\frac{L}{2},\frac{L}{2}\right) \tag{8.7.13}$$

Finally, let us indicate that if we were living beings in a two-dimensional space, the square is what we should perceive if a cube crossed our universe parallel to its faces.

Also notice, and without proof, that the sum of the angles or a plane square is equal to $4\cdot\pi/2=2\pi$.

8.7.1.4 Unspecified Triangle

Definition (#8.7.373): An "unspecified triangle" is a polygon with three sides and includes in particular cases the: isosceles, equilateral and rectangles triangles. A triangle with vertices A, B, and C is denoted $\triangle ABC$.

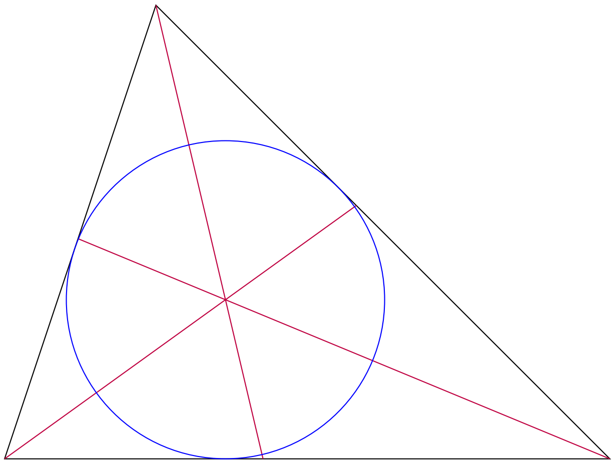


Figure 8.226: **Example of an unspecified triangle with inscribed circle and bisectors**

Remark

For the details about the Circumcircles and Incircles of triangles we send the reader to the section of Euclidean Geometry (see page 2620).

By the Euclid's axioms (see section Euclidean Geometry page 2563), the perimeter of an unspecified triangle is given by:

$$\boxed{P=a+b+c} \tag{8.7.14}$$

An unspecified plane triangle can always be decomposed into two right angle triangles. Thus, the one of the above figure can be divided into two right angle triangles:

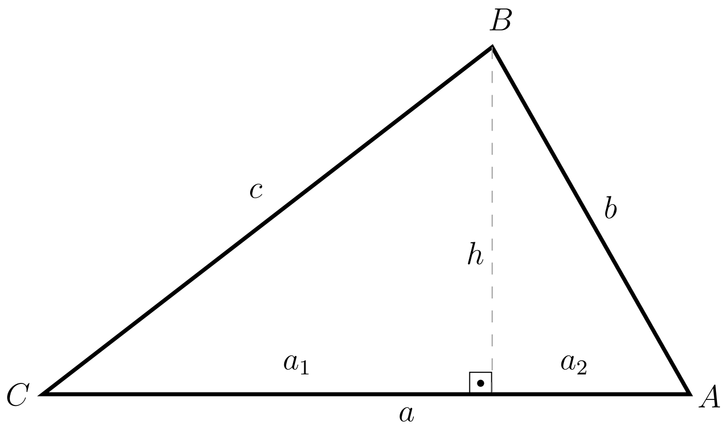


Figure 8.227: **Unspecified triangle divided into two right triangles**

of respective basis a_1 and a_2 (defined by the orthogonal projection of the top on opposite to the segment a) such as:

$$a = a_1 + a_2 \tag{8.7.15}$$

The surface of each of these two right angle triangles is as we have said already implicitly in our study of the rectangle, half the area of a rectangle of the same length and same height. Therefore:

$$S_1 = \frac{a_1 h}{2} \qquad S_2 = \frac{a_2 h}{2} \tag{8.7.16}$$

Therefore, the sum of these surfaces, gives us the surface of unspecified triangle:

$$\boxed{S_{\triangle} = S_1 + S_2 = \frac{a_1 h}{2} + \frac{a_2 h}{2} = \frac{a_1 + a_2}{2} h = \frac{ah}{2}} \tag{8.7.17}$$

We can say from this latter equality that the surface of any one triangle is similar to the half of the area of a rectangle of length $L = a$ and height $H = h$.

Remark

Whatever the basis a, b, c and the respective height h_a, h_b, h_c, the previous reasoning remains of course completely correct.

That said ... it is nice and pretty but there are very many practical cases where we do not know the height but only the length of the three sides. So, what do we do? Well, we will use the cosine theorem (Al-Kashi's Theorem) proved in the section Trigonometry (see page 2544) which gives us a reminder for any unspecified triangle of the type:

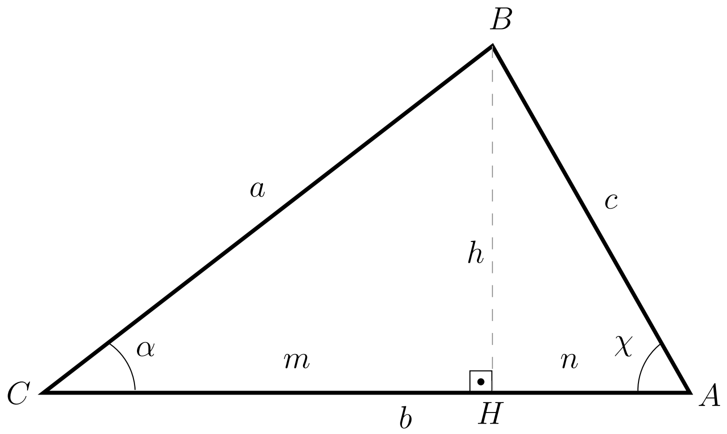


Figure 8.228: **Reminder of the construction for the cosine's theorem proof**

$$\cos(\chi) = \frac{b^2 + c^2 - a^2}{2bc} \tag{8.7.18}$$

and also the relation resulting in the triangle relative to the surface:

$$S = \frac{bh}{2} = \frac{bc\cos\left(\pi - \frac{\pi}{2} - \chi\right)}{2} = \frac{bc\cos\left(\frac{\pi}{2} - \chi\right)}{2} = \frac{bc\sin(\chi)}{2} \tag{8.7.19}$$

Therefore we have:

$$\begin{aligned}S&=\frac{bc\sqrt{1-\cos^2(\chi)}}{2}=\frac{bc\sqrt{(1-\cos(\chi))(1+\cos(\chi))}}{2}\\&=\frac{bc}{2}\sqrt{\left(1-\frac{b^2+c^2-a^2}{2bc}\right)\left(1+\frac{b^2+c^2-a^2}{2bc}\right)}\\&=\frac{bc}{2}\sqrt{\frac{1}{4b^2c^2}(2bc-b^2-c^2+a^2)(2bc+b^2+c^2-a^2}\\&=\frac{1}{4}\sqrt{(-(b-c)^2+a^2)\left((b+c)^2-a^2)\right)}=\frac{1}{4}\sqrt{(a^2-(b-c)^2)\left((b+c)^2-a^2\right)}\\&=\frac{1}{4}\sqrt{((a+b-c)(a-b+c))\left((b+c+a)(b+c-a)\right)}\end{aligned}\tag{8.7.20}$$

Therefore after rearrangement we get finally:

$$\boxed{S=\frac{1}{4}\sqrt{(a+b+c)(b+c-a)(a+b-c)(a-b+c)}}\tag{8.7.21}$$

Relation that is commonly named "Heron's formula" also sometimes named "Hero's formula" most know in the following form when we put $\Sigma:=\dfrac{1}{2}(a+b+c)$:

$$\begin{aligned}S&=\sqrt{\frac{1}{16}(a+b+c)(b+c-a)(a-b+c)(a+b-c)}\\&=\sqrt{\left(\frac{a+b+c}{2}\right)\left(\frac{b+c-a}{2}\right)\left(\frac{a-b+c}{2}\right)\left(\frac{a+b-c}{2}\right)}\\&=\sqrt{\Sigma\left(\Sigma-a\right)\left(\Sigma-b\right)\left(\Sigma-c\right)}\end{aligned}\tag{8.7.22}$$

The determination of the center of gravity (or centroid) G (see section Euclidean Geometry page 2635) is a little less intuitive than in the case of the rectangle...

Theorem 8.24. Given a triangle $\triangle ABC$. We name A' the "middle of the segment" $\overline{BC}$, B' that of $\overline{AC}$ and C' that of $\overline{AB}$:

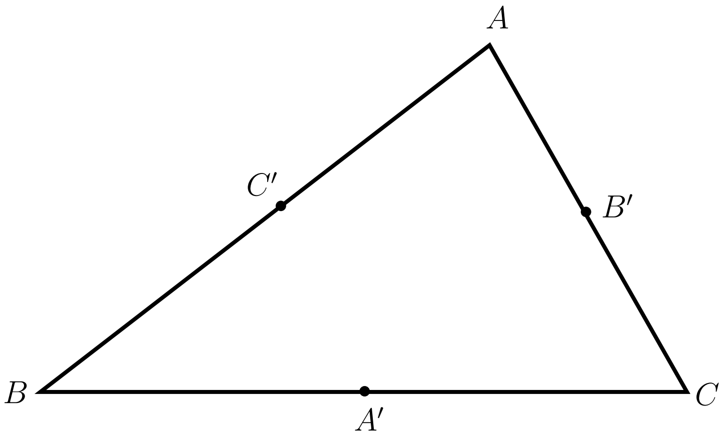


Figure 8.229: **Starting point for determining the center of gravity**

We will prove that only the point G satisfying (see section Euclidean Geometry):

$$\overrightarrow{GA}+\overrightarrow{GB}+\overrightarrow{GC}=\vec{0}\tag{8.7.23}$$

is the point of intersection of the three medians of triangle $\triangle ABC$. This proof will be made two stages: two proposals (at the end, we can conclude by the theorem):

P1. If $\triangle ABC$ is a triangle then there exists one and only one center of gravity G such that $\overrightarrow{GA}+\overrightarrow{GB}+\overrightarrow{GC}=\vec{0}$.

Proof 8.24.1. Let G be a point in the plane such as $\overrightarrow{GA}+\overrightarrow{GB}+\overrightarrow{GC}=\vec{0}$. Then we can write that:

$$\overrightarrow{GA}+\underbrace{\overrightarrow{GA}+\overrightarrow{AB}}_{\overrightarrow{GB}}+\underbrace{\overrightarrow{GA}+\overrightarrow{AC}}_{\overrightarrow{GC}}=\vec{0} \tag{8.7.24}$$

Therefore:

$$\overrightarrow{AG}=\frac{1}{3}\left(\overrightarrow{AB}+\overrightarrow{AC}\right) \tag{8.7.25}$$

This vector equality ensures that the G point is unique and we can even move it by moving the edges of the triangle!

■ Q.E.D.

P2. The three medians of a triangle are concurrent. Their intersection is the point G.

Proof 8.24.2. To prove that the three medians are concurrent, we will prove that G belongs to each of the three medians.

Just before we proved G satisfies the equality:

$$\overrightarrow{AG}=\frac{1}{3}\left(\overrightarrow{AB}+\overrightarrow{AC}\right) \tag{8.7.26}$$

As A' is the midpoint of $\overline{BC}$, then we can write that:

$$\overline{AB}+\overline{AC}=2\overline{AA'} \tag{8.7.27}$$

Therefore we get:

$$\overrightarrow{AG}=\frac{1}{3}\left(\overrightarrow{AB}+\overrightarrow{AC}\right)=\frac{1}{3}\left(2\overrightarrow{AA'}\right)=\frac{2}{3}\overrightarrow{AA'} \tag{8.7.28}$$

The vectors $\overrightarrow{AG}$ and $\overrightarrow{AA'}$ are collinear! Therefore the points A, G, A' are aligned. Writing in an other way, the point G is part of the median $\overline{AA'}$ of the triangle $\triangle ABC$. We can even say that is placed at the two thirds of the segment $\overline{AA'}$ from the top A.

What we have just shown with the median AA' is of course also true for the other two medians. Therefore:

$$\begin{gathered}\overrightarrow{AG}=\frac{2}{3}\overrightarrow{AA'}\\ \overrightarrow{BG}=\frac{2}{3}\overrightarrow{BB'}\\ \overrightarrow{CG}=\frac{2}{3}\overrightarrow{CC'}\end{gathered} \tag{8.7.29}$$

To resume, the point G is therefore part of the three medians $\overline{AA'}$, $\overline{BB'}$ and $\overline{CC'}$. These three lines are concurrent and the point G is the intersection point. This result will be useful later in our study of polyhedra.

■ Q.E.D.

Finally, let us indicate that if we were living beings in a two-dimensional space, the triangle is what we should perceive if geometric shapes composed of at least three joined sides would cross our universe through one of its vertex.

Also let us recall that the sum of the angles of any plane triangle is equal to π (180°) as proved in the section of Euclidean Geometry page 2593.

We will stop here this analogy with a two-dimensional space that can be generalized to each geometric shape that we present thereafter (circle and sphere, ellipse and ellipsoid, etc.). The idea was mainly to present the conception that volumes we experience in our daily lives can also be seen as 4-dimensional space shapes crossing our 3-dimensional space.

8.7.1.5 Isosceles Triangle

Definition (#8.7.374): An "isosceles triangle" is a triangle that has two sides of equal length. Sometimes it is specified as having two and only two sides of equal length, and sometimes as having at least two sides of equal length, the latter version thus including the equilateral triangle as a special case.

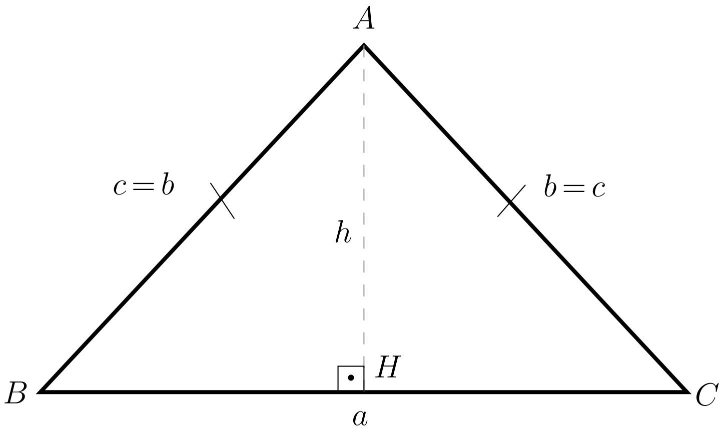


Figure 8.230: **Example of isosceles triangle**

In an isosceles triangle, the two equal sides are named "legs", and the remaining side is named the "base". The opposite angle to the base is named the "vertex angle", and the point associated with that angle is named the "apex". The two equal angles are named the "isosceles angles".

The perimeter of such a triangle remains:

$$\boxed{P = a + b + c} \quad (8.7.30)$$

but as it has two equal sides, then we can always simplify as follows:

$$P = a + b + b = a + 2b = a + 2c \quad (8.7.31)$$

The surface as we have proved above in the general case remains:

$$\boxed{S=\frac{ah}{2}} \tag{8.7.32}$$

And if we don't know the height but we know the internal angles we can by apply elementary trigonometry if needed (see section Trigonometry page 2508).

And the center of gravity remains, as we have proved in the general case (unspecified triangle), at the position:

$$\begin{aligned}\overrightarrow{AG}&=\frac{2}{3}\overrightarrow{AA'}\\ \overrightarrow{BG}&=\frac{2}{3}\overrightarrow{BB'}\\ \overrightarrow{CG}&=\frac{2}{3}\overrightarrow{CC'}\end{aligned} \tag{8.7.33}$$

A remarkable property of an isosceles triangle is that the bisector and the median of the third side, that is of not equal size than the two others, are not-distinguishable and therefore obviously of equal length (see section Euclidean Geometry page 2607).

A famous isosceles triangle is the right angle isosceles triangle because its angles are canonical and we get immediately $\theta=\pi/4$ and also we have immediately the length of the hypotenuse by applying the Pythagorean theorem:

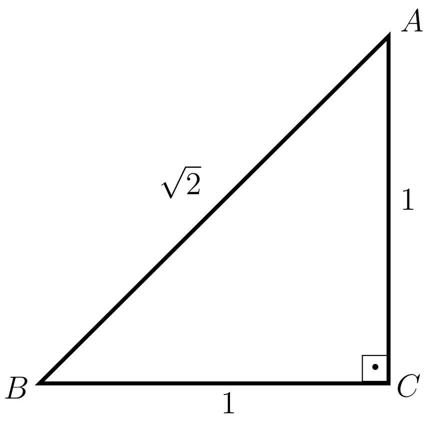


Figure 8.231: **Right angle isosceles triangle**

Also let us recall that the sum of the angles of any plane triangle is equal to π (180°) as proved in the section of Euclidean Geometry page 2593.

8.7.1.6 Equilateral Triangle

Definition (#8.7.375): An "equilateral triangle" is a triangle in which all three sides are equal. In the familiar Euclidean geometry, equilateral triangles are also "equiangular"; that is, all three internal angles are also congruent to each other and are each equal to $\pi/3$. They are regular polygons, and can therefore also be referred to as regular triangles.

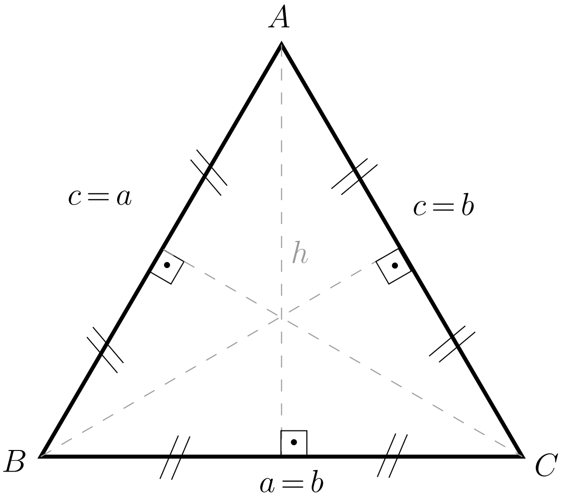


Figure 8.232: **Equilateral triangle**

Theorem 8.25. Each angle of an equilateral triangle is $\pi/3$ [rad] (60°).

Proof 8.25.1. First let us recall that by the angle sum theorem (page 2593) we have:

$$\hat{A}+\hat{B}+\hat{C}=\pi \tag{8.7.34}$$

We know also that Δabc is equilateral, then:

$$a=b=c \tag{8.7.35}$$

Then as $\angle\overline{ab}$ and $\angle\overline{ac}$ are congruent, we have:

$$\hat{A}=\hat{B} \tag{8.7.36}$$

And proceeding identically:

$$\hat{B}=\hat{C} \tag{8.7.37}$$

and:

$$\hat{C}=\hat{A} \tag{8.7.38}$$

Hence we can write that:

$$\hat{\theta}:=\hat{A}=\hat{B}=\hat{C} \tag{8.7.39}$$

And using the angle sum theorem:

$$\hat{A}+\hat{B}+\hat{C}=\pi=3\theta \tag{8.7.40}$$

We get:

$$\theta=\frac{\pi}{3} \tag{8.7.41}$$

■ Q.E.D.

The perimeter of such a triangle remains:

$$P=a+b+c \tag{8.7.42}$$

but as it has three equal sides, then we can always simplify as follows:

$$\boxed{P=3a=3b=3c} \tag{8.7.43}$$

The surface as we have proved above in the general case remains:

$$\boxed{S=\frac{ah}{2}} \tag{8.7.44}$$

Applying Pythagorean theorem we get if we don't know h:

$$S=\frac{a}{2}h=\frac{a}{2}\sqrt{a^2-\left(\frac{1}{2}a\right)^2}=\frac{a^2}{2}\sqrt{1-\frac{1}{4}}=\frac{a^2}{2}\sqrt{\frac{3}{4}}=\frac{\sqrt{3}}{4}a^2 \tag{8.7.45}$$

It can be shown that a triangle with given perimeter has the maximum possible area, if it is equilateral.

Proof 8.25.2. Remember the Heron's formula proved above ($\Sigma=\dfrac{1}{2}(a+b+c)$):

$$S=\sqrt{\Sigma\left(\Sigma-a\right)\left(\Sigma-b\right)\left(\Sigma-c\right)} \tag{8.7.46}$$

To find the conditions for the maximum area, we want to compute derivatives with respect a, b and c but since they are not independent variables, we first substitute the semi-perimeter Σ into Heron's equation to get rid of one degree of freedom:

$$c=2\Sigma-a-b \tag{8.7.47}$$

Therefore:

$$S=\Sigma(\Sigma-a)(\Sigma-b)(a+b-\Sigma) \tag{8.7.48}$$

Setting the derivative with respect to a to zero:

$$\begin{aligned}\frac{\mathrm{d}S}{\mathrm{d}a}&=\frac{1}{2}\frac{-\Sigma(\Sigma-b)(a+b-\Sigma)+\Sigma(\Sigma-a)(\Sigma-b)}{\sqrt{\Sigma\left(\Sigma-a\right)\left(\Sigma-b\right)\left(\Sigma-c\right)}}\\&=\frac{-\Sigma(\Sigma-b)(a+b-s)+\Sigma(\Sigma-a)(\Sigma-b)}{2S}\\&=\frac{\Sigma(\Sigma-b)(2\Sigma-2a-b)}{2S}=0\\&\Rightarrow 2\Sigma-2a-b=0\end{aligned} \tag{8.7.49}$$

Since $b=\Sigma$ is not a solution $A\neq 0\Rightarrow b\neq\Sigma$.

Similarly, setting the derivative with respect to b to zero yields:

$$2\Sigma-2b-a=0 \tag{8.7.50}$$

Solving simultaneously gives:

$$a=b=\frac{2\Sigma}{3} \tag{8.7.51}$$

and substituting back gives:

$$c=\frac{2\Sigma}{3} \tag{8.7.52}$$

So $a=b=c$ and the triangle is equilateral.

■ Q.E.D.

And the center of gravity remains, as we have proved in the general case (unspecified triangle), at the position:

$$\begin{gathered}\overrightarrow{AG}=\frac{2}{3}\overrightarrow{AA'}\\ \overrightarrow{BG}=\frac{2}{3}\overrightarrow{BB'}\\ \overrightarrow{CG}=\frac{2}{3}\overrightarrow{CC'}\end{gathered} \tag{8.7.53}$$

A remarkable property of an equilateral triangle is that all its bisectors and medians are non-distinguishable and obviously of the same length (see section Euclidean Geometry page 2607)!

Let us now see the important little "Viviani's theorem" regularly used for graphical representation in the field of materials engineering (mixture), or statistics (mixture designs or distribution of three frequencies of data whose sum is always equal).

Consider the following figure:

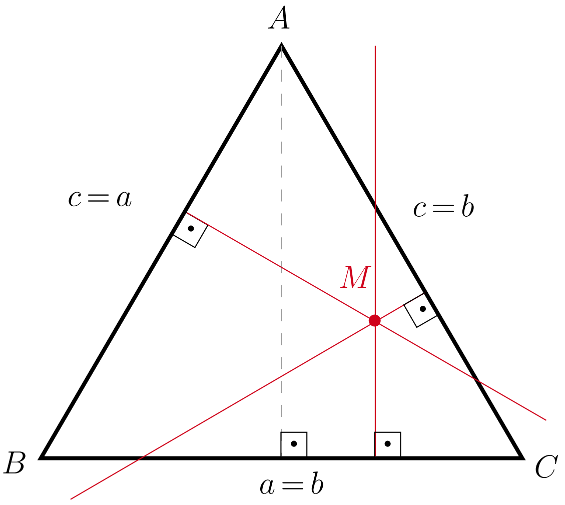


Figure 8.233: **Illustrative figure for Viviani's theorem**

Theorem 8.26. If we place a point in an equilateral triangle and that from this point we draw a line in the direction of each side, so that the lines are perpendicular to each side. No matter where we place the point, the sum of the perpendicular distances between the point and the sides is equal to the height of the triangle!

Proof 8.26.1. For the proof, let us denote by d_a, d_b, d_c the distances M to the sides, the length of one side and the height h. Then we have:

$$\frac{lh}{2} = S_{\Delta ABC} = S_{MBC} + S_{MCA} + S_{MAC} = \frac{ld_a}{2} + \frac{ld_b}{2} + \frac{ld_c}{2} = \frac{l(d_a + d_b + d_c)}{2} \tag{8.7.54}$$

and we have therefore well:

$$h = d_a + d_b + d_c \tag{8.7.55}$$

■ Q.E.D.

Also let us recall that the sum of the angles of any plane triangle is equal to π (180°) as proved in the section of Euclidean Geometry page 2593.

8.7.1.7 Right Triangle

Definition (#8.7.376): A "right triangle" (American English) or "right-angled triangle" (British English) is a triangle in which one angle is a right angle (that is, a $\pi/2$-radian angle). The relation between the sides and angles of a right triangle is the basis for trigonometry (see corresponding section).

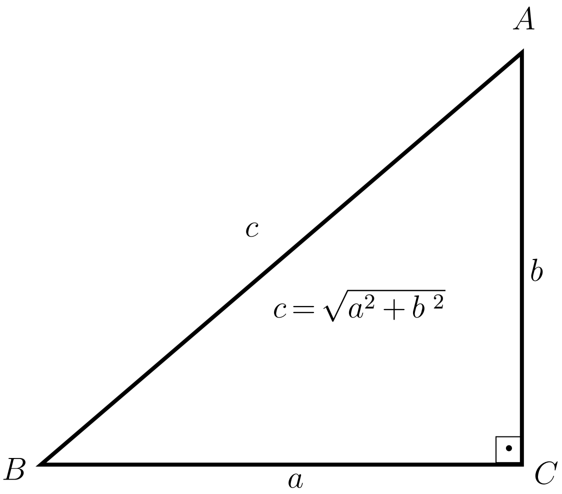


Figure 8.234: **Example of a right angle triangle**

The side opposite to the right angle is named the "hypotenuse".

The perimeter of such a triangle remains:

$$\boxed{P = a + b + c = a + b + \sqrt{a^2 + b^2}} \tag{8.7.56}$$

The surface as we have proved above in the general case remains:

$$\boxed{S = \frac{ah}{2} = \frac{ab}{2}} \tag{8.7.57}$$

And the center of gravity remains, as we have proved in the general case (unspecified triangle), at the position:

$$\begin{aligned}\overrightarrow{AG} &= \frac{2}{3}\overrightarrow{AA'}\\ \overrightarrow{BG} &= \frac{2}{3}\overrightarrow{BB'}\\ \overrightarrow{CG} &= \frac{2}{3}\overrightarrow{CC'}\end{aligned} \tag{8.7.58}$$

Remarkable property of a right triangle: this triangle has this unique property that we can directly apply to it the Pythagorean theorem (see section Euclidean Geometry page 2609).

Also let us recall that the sum of the angles of any plane triangle is equal to π (180°) as proved in the section of Euclidean Geometry page 2593.

8.7.1.8 Trapezoid

Definition (#8.7.377): A convex quadrilateral with at least one pair of parallel sides is referred to as a "trapezoid" in American and Canadian English but as a "trapezium" in English outside North America. The parallel sides are named the "bases" of the trapezoid and the other two sides are named the "legs" or simply the "lateral sides" (if they are not parallel; otherwise there are two pairs of bases). A "scalene trapezoid" is a trapezoid with no sides of equal length.

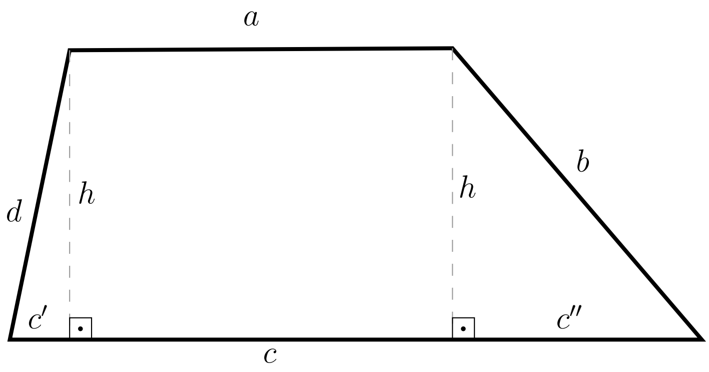


Figure 8.235: **Example of trapezoid**

Calculating the perimeter of the trapezoid is obvious:

$$\boxed{P = a + b + c + d} \tag{8.7.59}$$

Its surface is calculated using the decomposition:

$$\begin{aligned}S &= \frac{c'h}{2} + ah + \frac{c''h}{2} = \frac{(c'+c''+2a)h}{2} = \frac{(c'+c''+a+a)h}{2}\\ &= \frac{(c'+c''+a+(c-c'-c''))h}{2} = \frac{(a+c)h}{2}\end{aligned} \tag{8.7.60}$$

Finally:

$$\boxed{S = \frac{(a+c)h}{2}} \tag{8.7.61}$$

When both sides have the same length, we get the special cases of the square, the rectangle, of the rhombus or of the parallelogram (here we put the order from more specific to the more general, we could put the diamond rhombus as n°2):

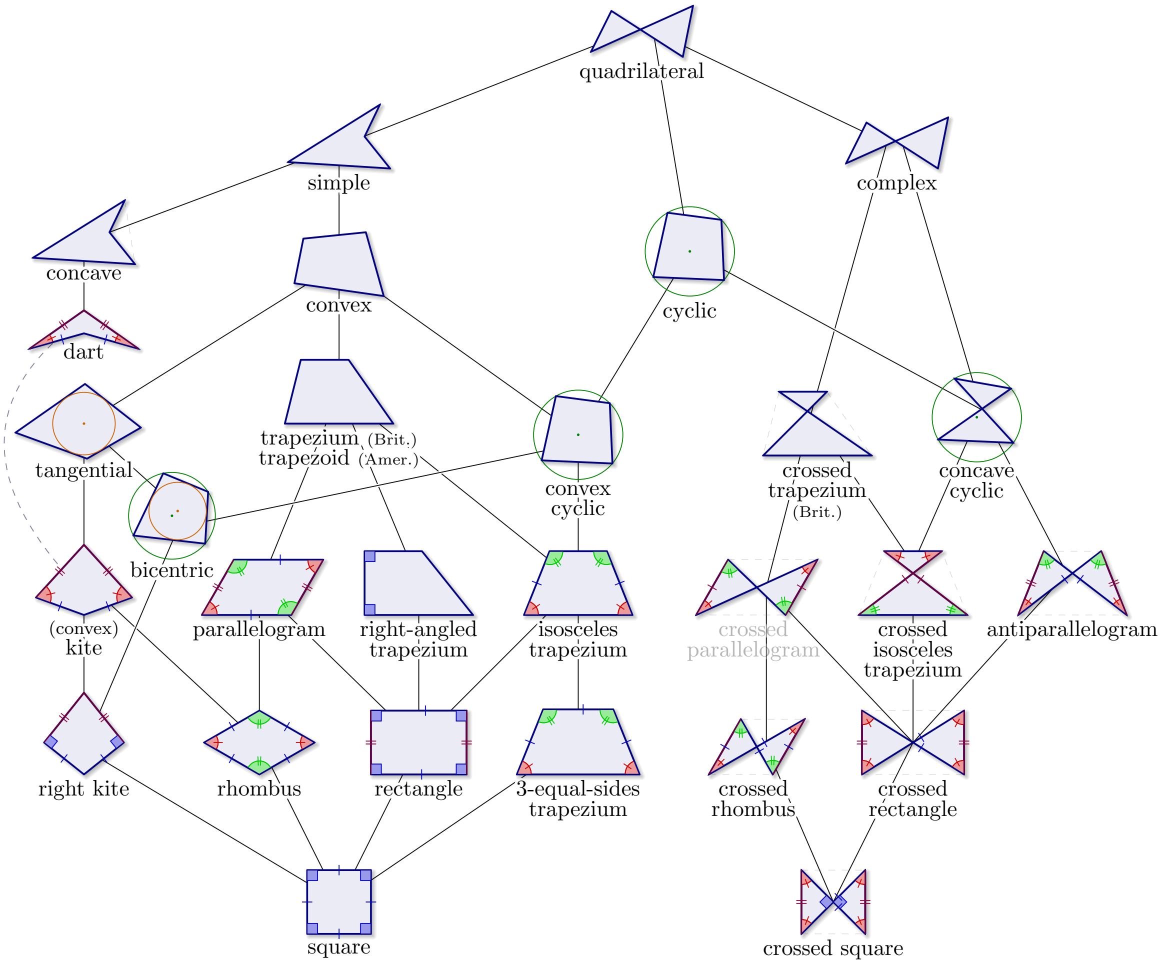

quadrilateral
simple
complex
concave
convex
cyclic
dart
tangential
trapezium (Brit.)
trapezoid (Amer.)
convex
cyclic
crossed
trapezium
(Brit.)
concave
cyclic
bicentric
(convex)
kite
parallelogram
right-angled
trapezium
isosceles
trapezium
crossed
parallelogram
crossed
isosceles
trapezium
antiparallelogram
right kite
rhombus
rectangle
3-equal-sides
trapezium
crossed
rhombus
crossed
rectangle
square
crossed square

Figure 8.236: **Quadrilateral hierarchy including Trapezoid classification (author: Izaak Neutelings)**

Also a common use is to retain a more restrictive definition, in order to not take into account these particular figures. We add in this case that the lengths of two parallel sides are not equal (this allows students of high-school classes to avoid confusion arising from the existence of two names for the same object, such as rhombus and trapezius).

Remark

There is a particular case of the trapezoid, the "isosceles trapezoid" as we can see on the previous figure, whose two non-parallel sides are the same length (we can add: as both sides are not parallel, it is not a parallelogram).

8.7.1.9 Parallelogram

Definition (#8.7.378): The "parallelogram" is a special case of the quadrilateral and very important (in the context of the analysis of shapes in physics), where the sides are parallel in pairs and of opposite equal angles (opposite angles are congruent):

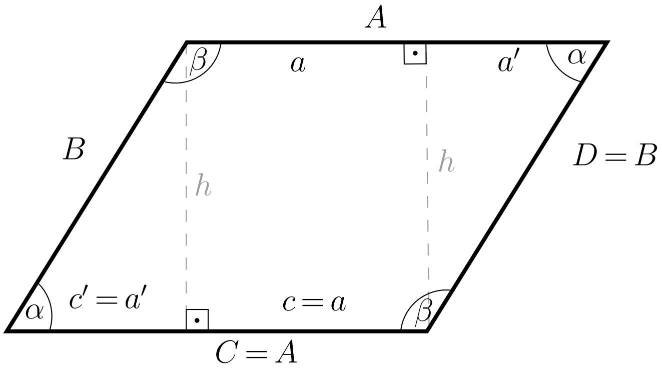


Figure 8.237: **Parallelogram example**

The rhombus being a special case of the parallelogram as we already know (the rhombus is a parallelogram with all sides equal)!

The perimeter of the parallelogram is obviously given by:

$$\boxed{P=a+a+b+b=2a+2b=2(a+b)} \tag{8.7.62}$$

For the surface, as this is a special case of the trapezoid, it comes immediately (as we will use only this relation in the book we will not prove other method to calculate the surface):

$$\boxed{S=\frac{(A+A)h}{2}=Ah} \tag{8.7.63}$$

Given the above figure, it is clear that if we cut the parallelogram into two as below, we see that this is twice the same triangle:

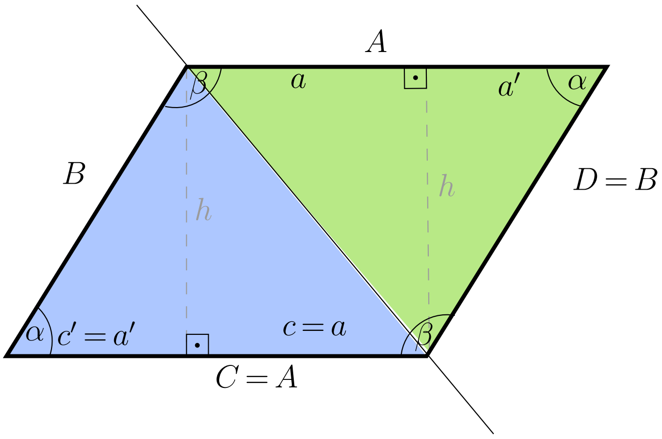


Figure 8.238: **Cutted parallelogram**

Thanks to this we will have a very useful property for the physical analysis of the static forces or phasors in the studies of the wave superposition (see section Wave Mechanics page ??).

Remark

As we already know parallelograms are therefore trapezoids like (see classification figure above for a refresh).

Also let us recall that as the sum of the angles of any plane triangle is equal to π (180°) as proved in the section of Euclidean Geometry page 2593, and that a parallelogram is made of two identical triangle, the sum of the angles of a parallelogram is then equal to 2π.

8.7.1.10 Hexagon

The hexagon is a quite interesting case. Indeed, is used for satellite signal Earth tiling and also because the "Honeycomb conjecture" states that the hexagonal tiling is the best way to divide a surface into regions of equal area with the least total perimeter.

This structure exists naturally in the form of graphite, where each sheet of graphene resembles chicken wire, with strong covalent carbon bonds. Tubular graphene sheets have been synthesised. These are known as "carbon nanotubes". They have many potential applications, due to their high tensile strength and electrical properties. Silicene is similar.

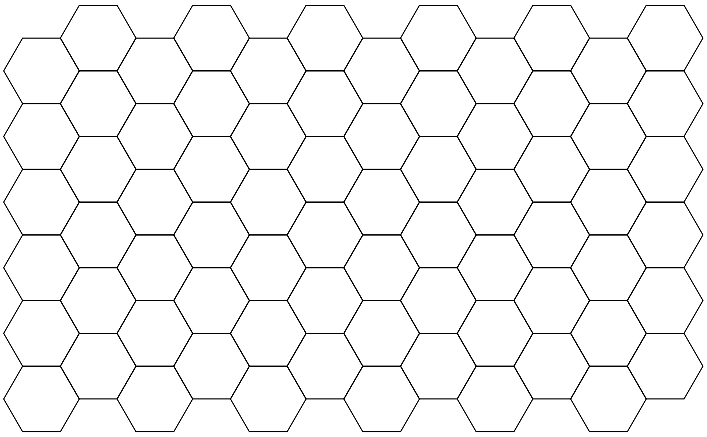


Figure 8.239: **A regular hexagonal grid**

Some applications (the reader can also take a look about the European Extra Large Telescope):

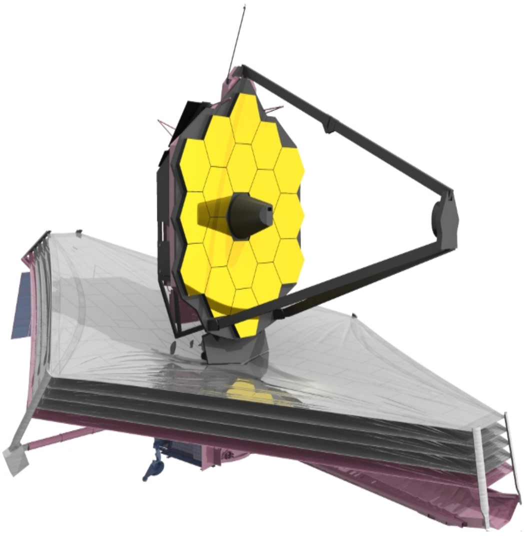


Figure 8.240: James Webb Space Telescope

Figure 8.241: TV broadcast reflector antenna

The study of the geometrical properties of the hexagon is also useful to study the "22° Sun hallo effect" and also the "Sun dog effect" (see section Geometric Optics page 4385).

Definition (#8.7.379): In geometry, a hexagon is a 6-sided polygon or 6-gon. The total of the internal angles of any hexagon is 4π. A "regular hexagon" is defined as a hexagon that is both equilateral and equiangular.

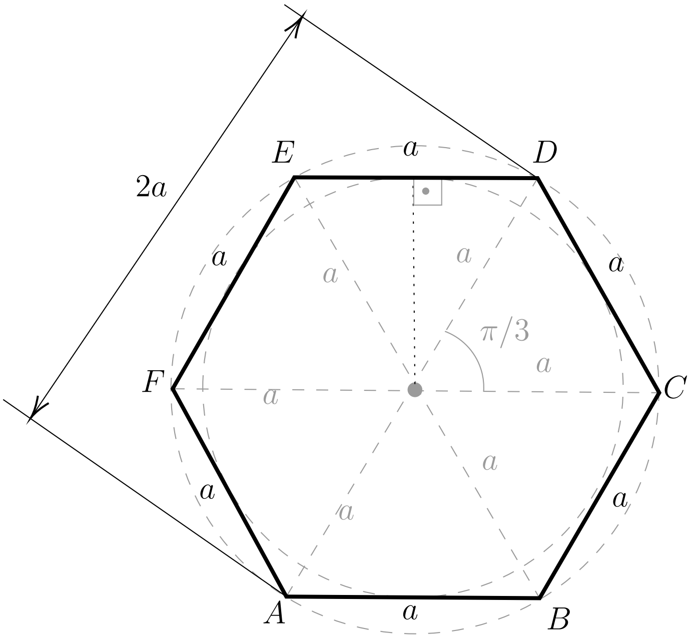


Figure 8.242: Regular hexagon details (source: Wikipedia)

First the perimeter is obviously:

$$\boxed{P=6a} \tag{8.7.64}$$

The surface of a regular hexagon of side length a is obviously given by:

$$S=6\left(\frac{1}{2}a(a\tan\left(\frac{\pi}{6}\right))\right)=3a^2\frac{\sqrt{3}}{2}=\frac{\sqrt{27}}{2}a^2 \tag{8.7.65}$$

Therefore:

$$\boxed{S=\frac{\sqrt{27}}{2}a} \tag{8.7.66}$$

Now let us come back to our honeycomb conjecture!

To begin our search, we had three ideas: the structure of the honeycomb cells should be strong, stable and economical.

The cells are solid if they can be nested and leave no holes. We are therefore interested in tiling.

It can be considered as intuitive that all cells must have the same form and that they will be stable if their form is regular. So we will study the regular polygons.

Let us recall that regular polygon is a geometric figure that has all sides of the same length and all angles of equal measure.

Of all the regular polygons, we have sought those who realize a tiling.

This problem is like looking how it is possible to assemble identical regular polygons around its edges.

For our study, we need to know two things:

1. The sum of the measures of the three angles of a planar triangle is equal to π
2. A round is equal to 2π

Let us try with different shapes:

1. Equilateral triangles ($\pi/3$)

   The three corners of an equilateral triangle are equal, and their measurement is $\pi/3=$. And as $6\cdot(\pi/2)=2\pi$, we can assembled six equilateral triangles around the vertices and thus realize a tiling:

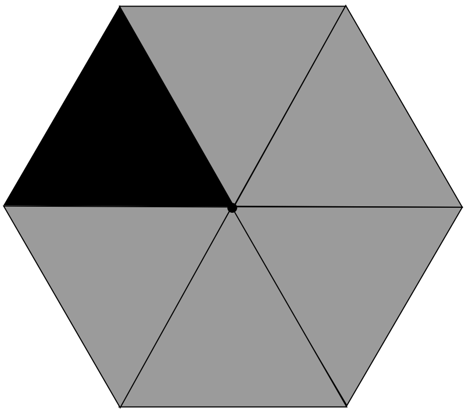


Figure 8.243: **Equilateral Triangle Tiling**

2. Squares $(2\pi/4)$

   The corners angles of a square measure $\pi/2$. And as $4\cdot(\pi/2)=2\pi$, we can assembled four square around the vertices and so achieve a tiling:

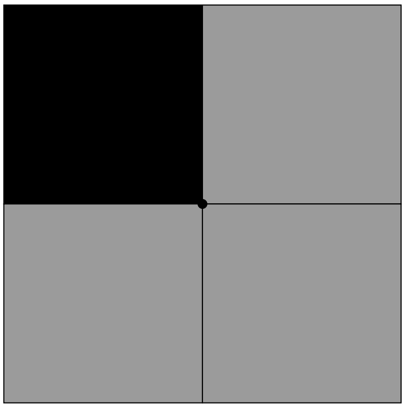


Figure 8.244: **Square Tiling**

3. Regular pentagons $(2\pi/5)$

   The internal angles of a regular pentagon is equal to $2\pi/5$ therefore the internal vertices angles are equal to $\pi-2\pi/5=3\pi/5$.

   If we assembled three of theme around a node this gives holes as $3\cdot(3\pi/5)<2\pi$, and obviously add one more pentagon will not solve the problem!

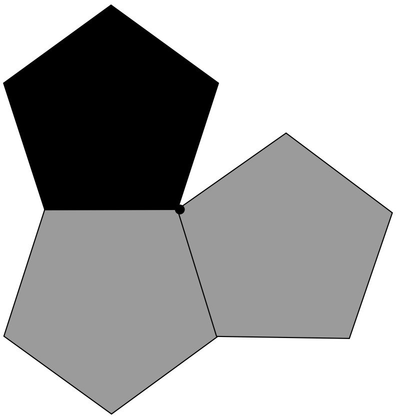


Figure 8.245: **Pentagon Tiling**

We therefore can not provide a tiling with pentagons.

4. Regular hexagons ($2\pi/6$)

   The internal angles of a regular pentagon is equal to $2\pi/6$ therefore the internal vertices angles are equal to $\pi-2\pi/6=2\pi/3$.

   And as $3\cdot(2\pi/3)=2\pi$, we can assemble three regular hexagons around the vertices and so achieve a tiling:

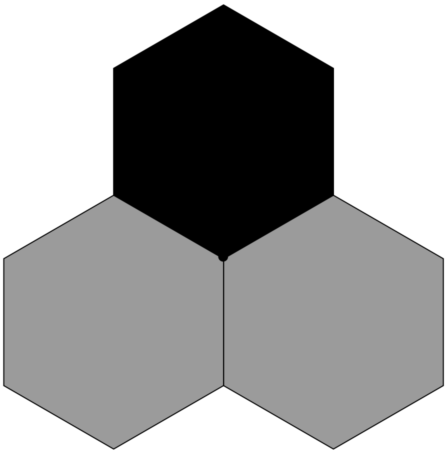


Figure 8.246: **Hexagon Tiling**

5. Regular polygons with seven sides and more

   From polygon with seven sides and more, tiling is impossible because the internal angles are not integer dividers from 2π.

The general relation to get the measurement of the angles of a regular polygon with n sides is:

$$\pi-\left(\frac{2\pi}{n}\right) \tag{8.7.67}$$

Therefore, the only regular polygons that realize a tiling are equilateral triangles, squares and regular hexagons!!

Finally, we thought that bees must use the least amount of wax to make their cells, while having the maximum space inside.

Therefore the question is: Among the equilateral triangle, the square and the regular hexagon, which he has the greatest area for the smallest perimeter?

In other word we seek to maximize the ratio A/P (or the smallest ratio P/A).

So we have for each shape using our previous results:

1. Equilateral triangle:

$$\frac{A}{P}=\frac{\dfrac{1}{2}bh}{3b}=\frac{1}{6}h=\frac{1}{6}b\tan(\pi/6)=\frac{1}{6}b\frac{\sqrt{3}}{3}b\cong\frac{\sqrt{3}}{18}b\cong 0.0962b \tag{8.7.68}$$

2. Square:

$$\frac{A}{P}=\frac{b^2}{4b}=\frac{b}{4}=0.25b \tag{8.7.69}$$

3. Regular hexagon:

$$\frac{\frac{\sqrt{27}}{2}b^2}{6b}=\frac{\sqrt{27}}{12}b=\frac{\sqrt{3}}{3}b\cong 0.5773b \tag{8.7.70}$$

Among the equilateral triangle, the square and the regular hexagon it is therefore the regular hexagon which gives for the maximum ratio "area on perimeter".

So that the cell structure is solid, stable and economical, bees are incentive to build basic regular hexagons. We can see that this is what they do naturally (after a lot of trial and errors during their evolution)!

8.7.1.11 Rhombus

Definition (#8.7.380): A "rhombus" is a simple (non-self-intersecting) quadrilateral all of whose four sides have the same length. Another name is "equilateral quadrilateral", since equilateral means that all of its sides are equal in length. The rhombus is often named a "diamond", after the diamonds suit in playing cards which resembles the projection of an octahedral diamond, though the former sometimes refers specifically to a rhombus with a $\pi/3$ angle and the latter sometimes refers specifically to a rhombus with a $\pi/4$ angle.

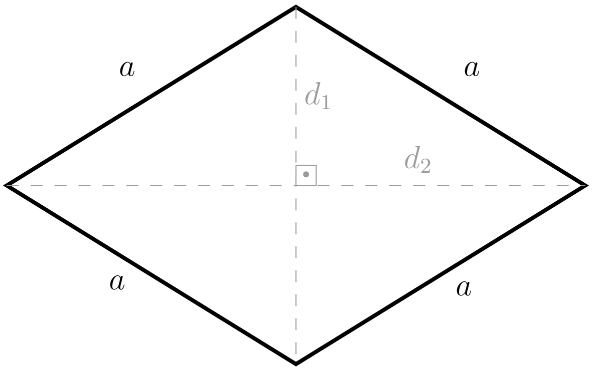


Figure 8.247: **Example of a rhombus**

Calculating the perimeter and the surface of the rhombus immediately follow from those of the trapezium.

8.7.1.12 Circle

There are several possible definitions of the circle. Let's look at least two.

Definitions (#8.7.381):

- D1. A "circle" is a special case of a polygon with an infinite number of sides.
- D2. A "circle" is a flat curve all of whose points are equidistant from a fixed point named "center".

Remark

A circle may also be defined as a special ellipse in which the two foci are coincident and the eccentricity is 0.

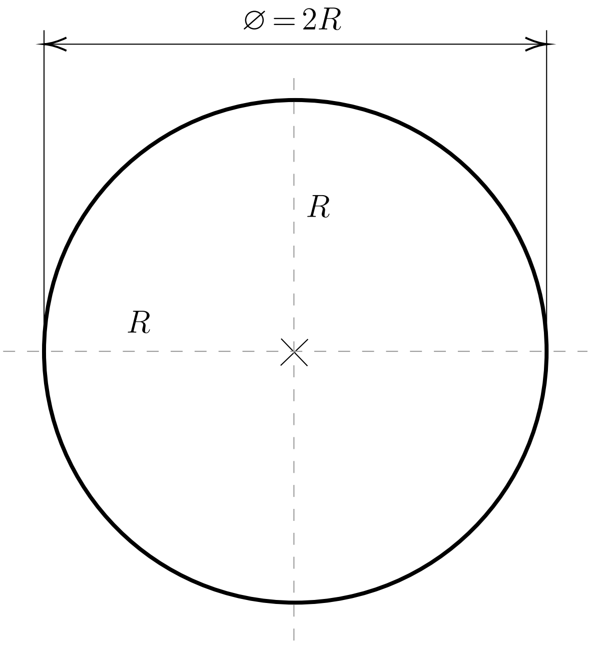


Figure 8.248: **Example of a circle**

We will prove in the section of Theoretical Computing at page 6298 that the perimeter of a circle of radius R and therefore of diameter $\varnothing=2R$ is given by:

$$\boxed{P=2\pi R=(2R)\pi=\varnothing\pi} \quad (8.7.71)$$

The relation for the surface (of the corresponding "disc") can be obtained in two common ways:

1. By the search of the primitive of perimeter P which gives us:

$$S=\int P\mathrm{d}R=2\pi\int R\mathrm{d}R=\pi R^2 \quad (8.7.72)$$

2. The second method is more aesthetic and uses the parametric equation of the circle, trivially given by the orthogonal projections of the Cartesian coordinates (see section Vector Calculus page 1863):

$$\begin{aligned} x &= R\cos(\theta) \\ y &= R\sin(\theta) \end{aligned} \tag{8.7.73}$$

We know that the area described by a function $f(x)$ is given by (see section Differential and Integral Calculus page 1497):

$$\boxed{S=\int\limits_a^b f(x)\mathrm{d}x=\int\limits_a^b y\mathrm{d}x} \tag{8.7.74}$$

We just have to substitute in this integral the parametrized variables:

$$y=R\sin(\theta) \qquad \mathrm{d}x=\mathrm{d}(R\cos(\theta))=R\mathrm{d}\cos(\theta)=-R\sin(\theta)\mathrm{d}\theta \tag{8.7.75}$$

Therefore:

$$S=\int\limits_a^b y\mathrm{d}x=\int\limits_a^b (R\sin(\theta))(-R\sin(\theta)\mathrm{d}\theta)=-R^2\int\limits_a^b \sin^2(\theta)\mathrm{d}\theta \tag{8.7.76}$$

The limits of integration are obviously $\theta=[0,2\pi]$ then we have:

$$\begin{aligned} S&=-R^2\int\limits_0^{2\pi}\sin^2(\theta)\mathrm{d}\theta=-2R^2\int\limits_0^{\pi}\sin^2(\theta)\mathrm{d}\theta=-2R^2\int\limits_0^{\pi}\frac{1-\cos(2\theta)}{2}\mathrm{d}\theta \\ &=2R^2\left(\frac{\sin(2\theta)}{4}-\frac{\theta}{2}\right)\Bigg|_0^{-\pi}=R^2\theta\Big|_0^{-\pi}-R^2\frac{\sin(2\theta)}{2}\Bigg|_0^{-\pi}=R^2\pi \end{aligned} \tag{8.7.77}$$

We therefore also by this method:

$$S=\pi R^2=\pi\left(\frac{\varnothing}{2}\right)^2=\pi\frac{\varnothing^2}{4} \tag{8.7.78}$$

The length l of an opening angle range α of a circle of radius R is obviously given by:

$$l=\frac{2\pi R}{2\pi}\alpha=R\alpha \tag{8.7.79}$$

and the surface S_t of an opening angle α of a circle of radius R identically by:

$$S_t=\frac{\pi R^2}{2\pi}\alpha=\frac{R^2}{2}\alpha \tag{8.7.80}$$

Let us now determine the surface S_d of cutted a part of disc that is to say:

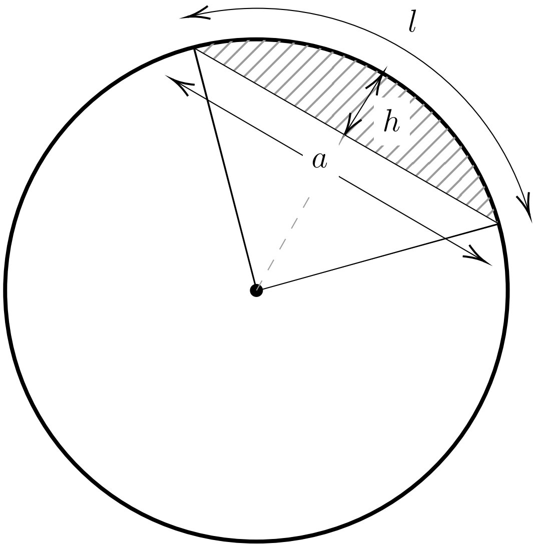


Figure 8.249: **Surface S_d of a disc cutted part**

We have:

$$S_d = \frac{R^2}{2}\alpha - \frac{a(R-h)}{2} = \frac{R^2}{2}\frac{l}{R} - \frac{a(R-h)}{2} = \frac{Rl - a(R-h)}{2} \tag{8.7.81}$$

Remark

A circle may also be defined as a special ellipse in which the two foci are coincident and the eccentricity is 0.

8.7.1.13 Ellipse

Definition (#8.7.382): An "ellipse" is a closed curve where every point P is such that the sum of its distances from two fixed points named "foci $F1, F2$" is constant (as we see it in details in the section of Analytical Geometry the ellipse can also be seen as an affine transformation of the circle).

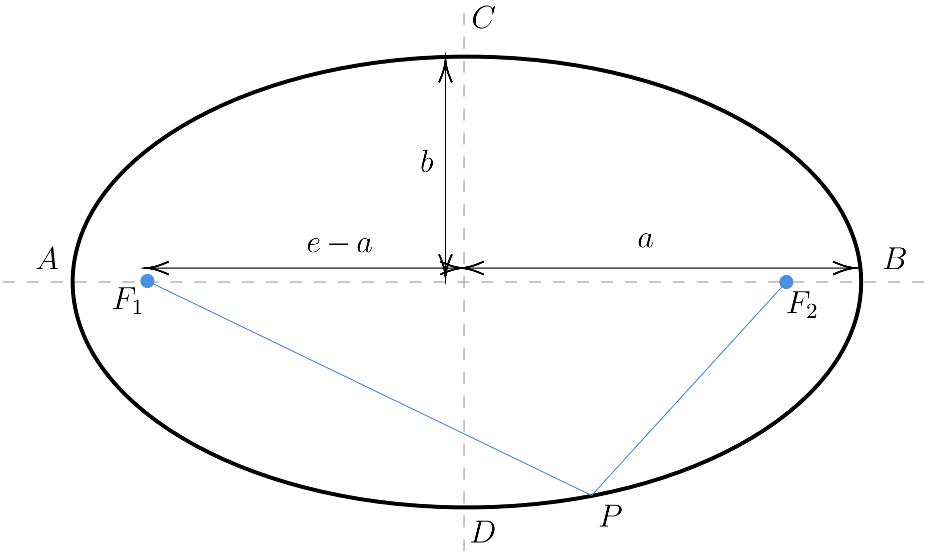


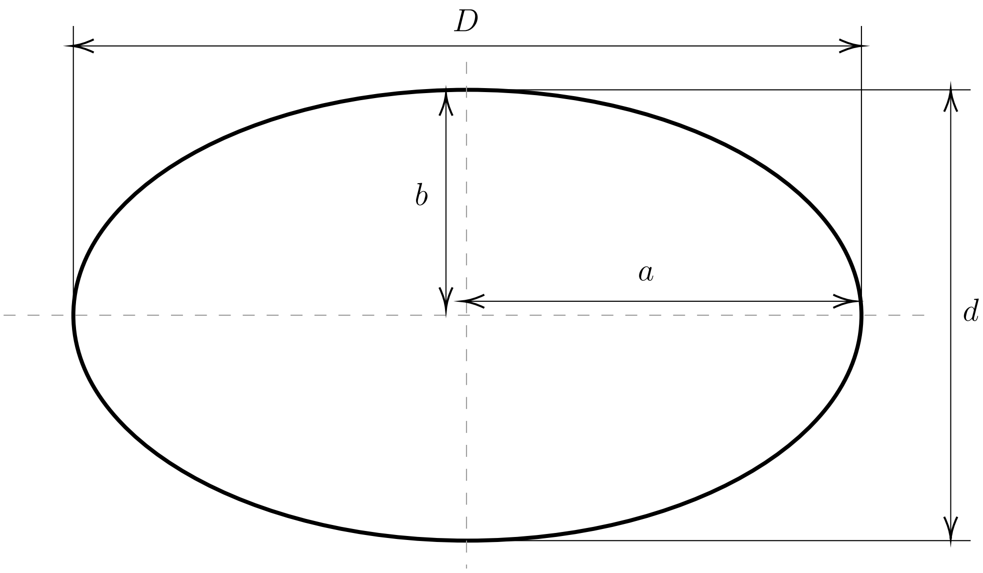
Figure 8.250: **Example of an ellipse with its visible foci and eccentricity e**

Remark

When am "ellipse" is a closed curve where every point P is such that the sum of its distances from two fixed points named "foci $F1, F2$" is constant an "oval" is a closed curve in a plane which resembles the outline of an egg (i.e. an oval is a curve resembling a squashed circle but, unlike the ellipse, without a precise mathematical definition).

The shape of an ellipse (how 'elongated' it is) is represented by its "eccentricity e", which for an ellipse can be any number from 0 (the limiting case of a circle) to arbitrarily close to but less than 1.

For the mathematical developments that will follow we will rather use the notation of the following figure:



Remark

The reader must therefore not forget that most important mathematical developments about the ellipse are in the section of Analytical Geometry (see page 2764).

Let us introduce to start a small text relative to the calculation of the perimeter of the ellipse!

Given the parametric equation in Cartesian coordinates of an ellipse (see section Vector Calculus page 1863):

$$\begin{aligned} x &= a\cos(\theta) \\ y &= b\sin(\theta) \end{aligned} \tag{8.7.82}$$

The distance between the center of the ellipse and its perimeter is given by the Pythagorean Theorem:

$$R(\theta)=\sqrt{a^2\cos^2(\theta)+b^2\sin^2(\theta)} \tag{8.7.83}$$

An arc element is then given by:

$$R(\theta)\mathrm{d}\theta=\sqrt{a^2\cos^2(\theta)+b^2\sin^2(\theta)}\mathrm{d}\theta \tag{8.7.84}$$

The perimeter of the ellipse is the given by the integral:

$$P=\int\limits_0^{2\pi} R(\theta)\mathrm{d}\theta=\int\limits_0^{2\pi}\sqrt{a^2\cos^2(\theta)+b^2\sin^2(\theta)}\mathrm{d}\theta=4\int\limits_0^{\pi/2}\sqrt{a^2\cos^2(\theta)+b^2\sin^2(\theta)}\mathrm{d}\theta \tag{8.7.85}$$

and then it starts to get tougher... This type integral can not easily be calculated using the usual primitives, integration by parts or change of variables. This is what we name a "second-order elliptic integral J" for $0<b<a$ (see section Differential and Integral Calculus page 1599):

$$\boxed{J(a,b)=\int\limits_0^{\pi/2}\sqrt{a^2\cos^2(\theta)+b^2\sin^2(\theta)}\mathrm{d}\theta} \tag{8.7.86}$$

Long developments which we present in a few years in the section of Differential and Integral Calculus give for calculating the perimeter after a limited series:

$$\boxed{P\cong\sqrt{2(a^2+b^2)}=2\pi a\sqrt{1-e^2}} \tag{8.7.87}$$

The surface of the ellipse can be obtained in a very similar manner to that of the circle and the calculations are surprisingly much simpler than those of its perimeter. Remember that the parametric equation the ellipse is:

$$\begin{aligned} x &= a\cos(\theta) \\ y &= b\sin(\theta) \end{aligned} \tag{8.7.88}$$

We know that the area described by a function $f(x)$ is given by:

$$S=\int\limits_a^b f(x)\mathrm{d}x=\int\limits_a^b y\mathrm{d}x \tag{8.7.89}$$

We just have to substitute in this integral the parametrized variables:

$$y=b\sin(\theta) \qquad \mathrm{d}x=\mathrm{d}(a\cos(\theta))=a\mathrm{d}\cos(\theta)=-a\sin(\theta)\mathrm{d}\theta \tag{8.7.90}$$

Therefore:

$$S=\int\limits_a^b y\mathrm{d}x=\int\limits_a^b (b\sin(\theta))(-a\sin(\theta)\mathrm{d}\theta)=-ab\int\limits_a^b \sin^2(\theta)\mathrm{d}\theta \tag{8.7.91}$$

The limits of integration being obviously $\theta=[0,-2\pi]$ we have:

$$\begin{aligned}S&=-ab\int\limits_0^{-2\pi}\sin^2(\theta)\mathrm{d}\theta=-2ab\int\limits_0^{-\pi}\frac{1-2\cos(2\theta)}{2}\mathrm{d}\theta=-2ab\left(\frac{1}{2}\theta-\frac{1}{4}\sin(2\theta)\right)\Bigg|_0^{-\pi}\\&=-2ab\left[\left(\frac{1}{2}(-\pi)-\frac{1}{4}\sin(2(-\pi))\right)-\left(\frac{1}{2}\cdot 0-\frac{1}{4}\sin(2\cdot 0)\right)\right]\\&=-2ab\frac{1}{2}(-\pi)=ab\pi\end{aligned} \tag{8.7.92}$$

Remark

Be careful with this kind of calculations with the order of integration bounds. Indeed, if we had taken the bounds from $[0,\pi]$ (instead of $[0,-\pi]$) one must imagine that the integrated function browse the perimeter in the negative direction of the x-axis. So the integral would be necessarily be negative.

We therefore also have for this method:

$$\boxed{S=\pi ab=\frac{\pi}{4}D\cdot d} \tag{8.7.93}$$

Remarks

R1. We assume as obvious (and therefore without proof) that the ellipse's center of gravity coincides with the center of it.

R2. We send the reader to the study of conical (see section Analytical Geometry page 2764) for the calculation of the area of an ellipse from its "ellipse parameter" and its "eccentricity" (everything is prove there).

8.7.1.13.1 Ellipse section surface

Now let us focus on a common question on Internet Forums: What is the area of a section of an ellipse. To answer that question we start again from:

$$\begin{aligned}x&=a\cos(\theta)\\y&=b\sin(\theta)\end{aligned} \tag{8.7.94}$$

but that we will write a in slightly different way:

$$\begin{aligned} x &= r(\theta)\cos(\theta) \\ y &= r(\theta)\sin(\theta) \end{aligned} \tag{8.7.95}$$

An element of surface in polar coordinates is given by:

$$\mathrm{d}S = r(\theta)\mathrm{d}r\mathrm{d}\theta \tag{8.7.96}$$

Therefore the area of a section has for expression:

$$S = \int\limits_{\theta_1}^{\theta_2}\int\limits_{0}^{r(\theta)} r\mathrm{d}r\mathrm{d}\theta \tag{8.7.97}$$

Thus:

$$S = \frac{1}{2}\int\limits_{\theta_1}^{\theta_2} r^2(\theta)\mathrm{d}\theta \tag{8.7.98}$$

But using what we have seen in the section of Analytical Geometry we can write that latter:

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = 1 = r(\theta)\left(\frac{\cos^2(\theta)}{a^2} + \frac{\sin^2(\theta)}{b^2}\right) \tag{8.7.99}$$

Therefore:

$$r^2(\theta) = \frac{1}{\dfrac{\cos^2(\theta)}{a^2} + \dfrac{\sin^2(\theta)}{b^2}} \tag{8.7.100}$$

Thus:

$$S = \frac{1}{2}\int\limits_{\theta_1}^{\theta_2} r^2(\theta)\mathrm{d}\theta = \frac{1}{2}\int\limits_{\theta_1}^{\theta_2} \frac{\mathrm{d}\theta}{\dfrac{\cos^2(\theta)}{a^2} + \dfrac{\sin^2(\theta)}{b^2}} \tag{8.7.101}$$

To continue let us make a simple change of variables:

$$\int\limits_{\theta_1}^{\theta_2} \frac{\mathrm{d}\theta}{a'^2\cos^2(\theta) + b'^2\sin^2(\theta)} \tag{8.7.102}$$

Pulling out $a'^2\cos^2(\theta)$ we get:

$$\int \frac{\mathrm{d}\theta}{a'^2\cos^2(\theta)\left(1 + \dfrac{b'^2}{a'^2}\tan^2(\theta)\right)} \tag{8.7.103}$$

Now we make a serious change of variable:

$$u = \frac{b'}{a'}\tan(\theta) \tag{8.7.104}$$

Thus according to the usual derivatives (see section Differential and Integral Calculus page 1475):

$$\mathrm{d}u=\frac{b'}{a'}\frac{1}{\cos^2(\theta)}\mathrm{d}\theta \tag{8.7.105}$$

Then the integral becomes:

$$\frac{1}{a'^2}\int\frac{\mathrm{d}\theta}{\cos^2(\theta)\left(1+\dfrac{b^2}{a^2}\tan^2(\theta)\right)}=\frac{1}{a'^2}\int\frac{\dfrac{a'}{b'}\mathrm{d}u}{1+u^2}=\frac{1}{a'b'}\int\frac{\mathrm{d}u}{1+u^2} \tag{8.7.106}$$

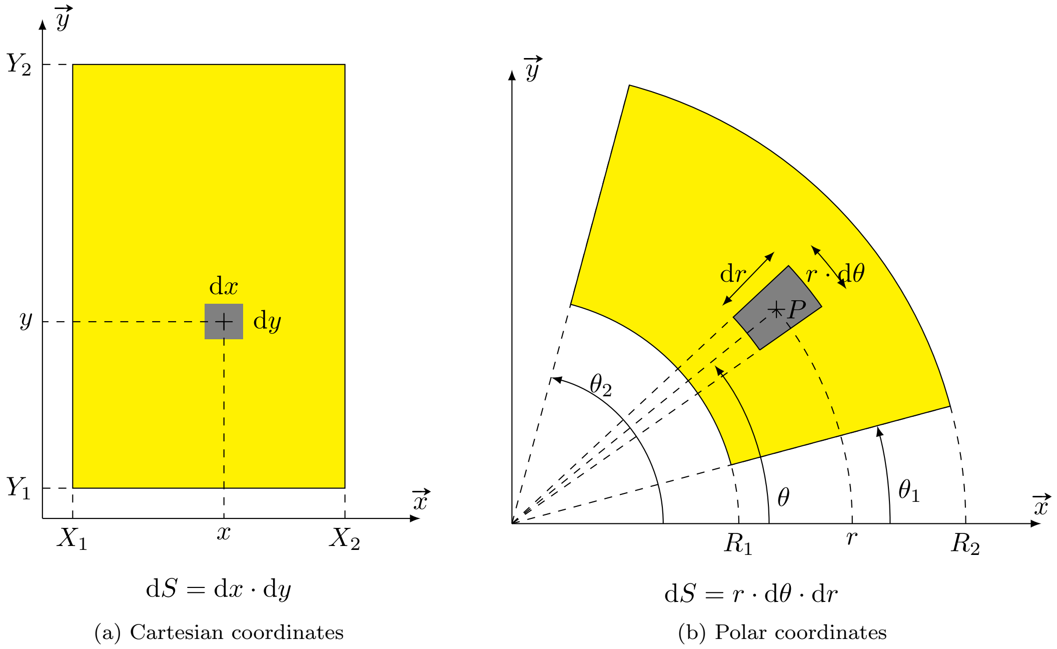
We know according to usual integrals that (see section Differential and Integral Calculus page 1534):

$$\frac{1}{a'b'}\int\frac{\mathrm{d}u}{1+u^2}=\frac{1}{a'b'}\arctan(u)=\frac{1}{a'b'}\arctan\left(\frac{b'}{a'}\tan(\theta)\right) \tag{8.7.107}$$

So finally:

$$\boxed{S=\frac{ab}{2}\arctan\left(\frac{a}{b}\tan(\theta)\right)\Bigg|_{\theta_1}^{\theta_2}} \tag{8.7.108}$$

A small reminder of some principles used so far:



$\vec{y}$
Y_2
y
Y_1
X_1
x
X_2
$\vec{x}$
$\mathrm{d}x$
$\mathrm{d}y$
$\mathrm{d}S=\mathrm{d}x\cdot\mathrm{d}y$
$\vec{y}$
$\mathrm{d}r$
$r\cdot\mathrm{d}\theta$
P
θ_2
θ
θ_1
R_1
r
R_2
$\vec{x}$
$\mathrm{d}S=r\cdot\mathrm{d}\theta\cdot\mathrm{d}r$

(a) Cartesian coordinates (b) Polar coordinates

Figure 8.251: Surface elements $\mathrm{d}S$ in 2D

8.7.2 Usual Volumes

There are several definitions of the concept of volume (surface that limits a body). A definition due to Euclid and another due to the field of Topology (see the section of the same name).

Definitions (#8.7.383):

D1. A "volume" is an entity that has a length, a width and a height.

D2. A "volume" is a 3-manifold topology.

The surfaces that surround a body may be planar or curved:

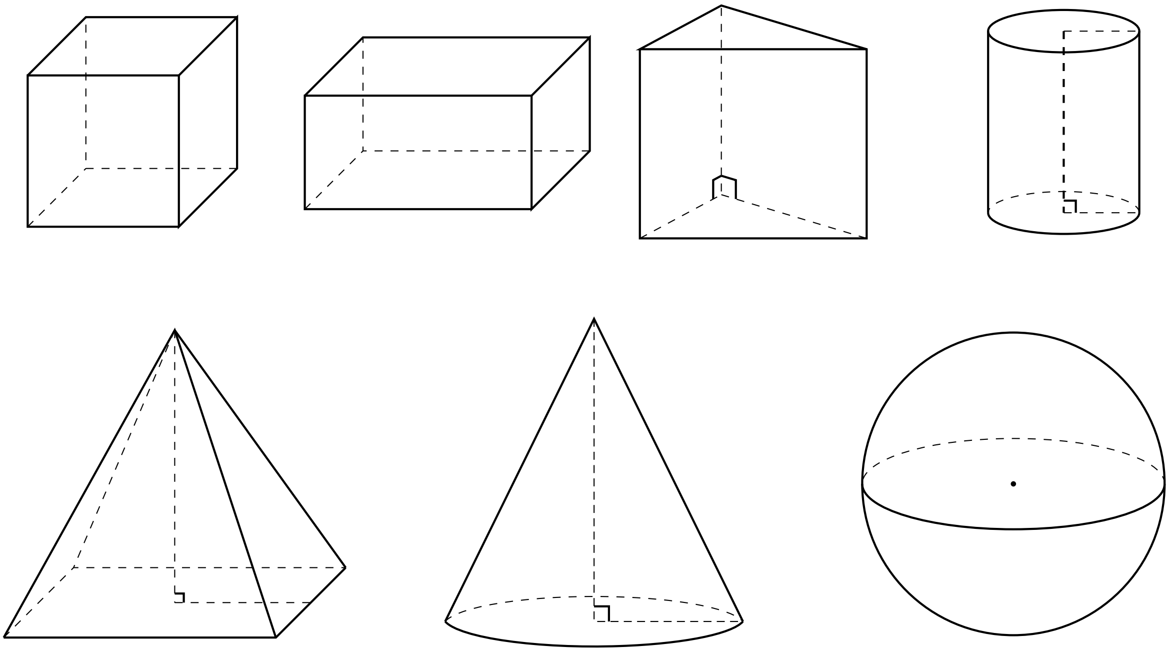


Figure 8.252: **Simple volumes bounded by a surface**

On the left, the body is limited only by planar surfaces, in the middle by one and only one single curved surface, and on the right by a curved surface and two planar surfaces.

Remark

We will focus initially only to the properties (area, volume, center of gravity, moment of inertia ...) of volumes plunged into Euclidean geometries.

8.7.2.1 Polyhedron

The study of polyhedra (especially Platonic polyhedra) is very important in physics (e.g. crystallography) and also in mathematics because it gives a nice application of finite groups (see section Set Algebra page 1420). It is therefore appropriate to carefully read what will follow.

Furthermore, the study of polyhedra is also a very educational and aesthetic way to see the implementation of several geometric, trigonometric and vector algebra theorems.

It should be noted before all that the various polyhedra will deliberately not presented as equal in their study. Thus, we will focus only on given properties for some and not for others.

Definitions (#8.7.384):

D1. A "polyhedron" is a solid whose border is formed of plans or plan portions. The plan the portions, which include between them the polyhedron, are the faces; each face being limited by intersections (edges) with neighbouring faces, is a polygon. The sides of this polygon are the edges of the polyhedron. We name "vertices" of a polyhedron any vertex of any of its faces.

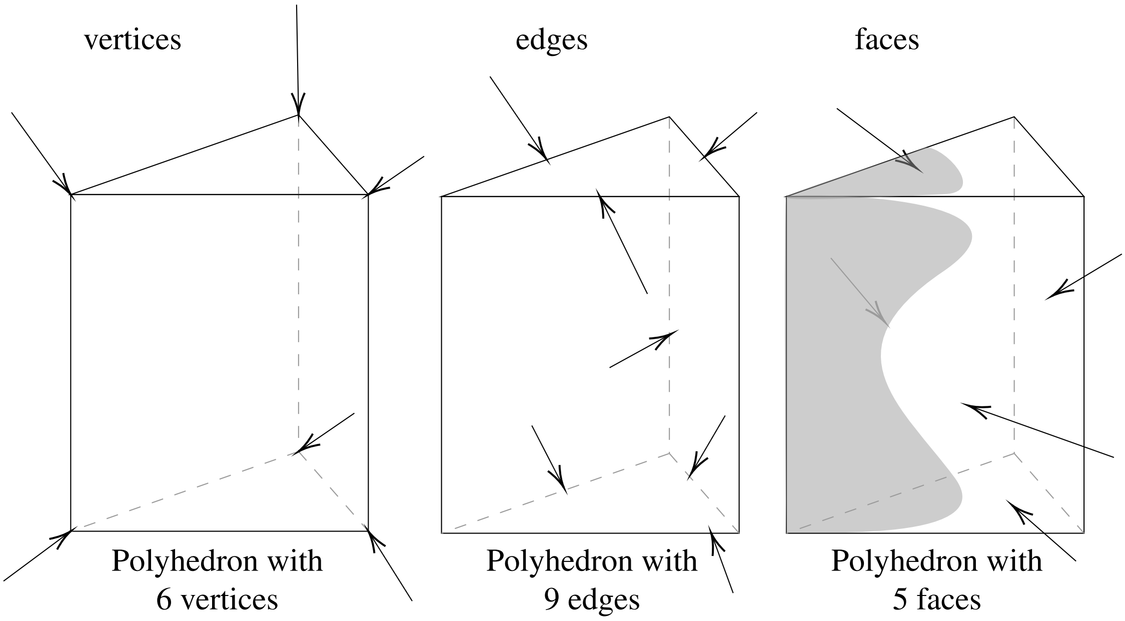


Figure 8.253: **Polyhedron vocabulary**

D2. A "regular polygon" is a polygon which sides and all angles are equal (this definition will be useful for regular polyhedra further below).

8.7.2.1.1 Parallelepiped

Definition (#8.7.385): In geometry, a "parallelepiped" is a three-dimensional figure formed by six parallelograms (the term rhomboid is also sometimes used with this meaning). By analogy, it relates to a parallelogram just as a cube relates to a square or as a cuboid to a rectangle.

Three equivalent definitions of parallelepiped are:

- A polyhedron with six faces (hexahedron), each of which is a parallelogram.
- A hexahedron with three pairs of parallel faces
- A prism of which the base is a parallelogram.

The rectangular cuboid (six rectangular faces), cube (six square faces), and the rhombohedron (six rhombus faces) are all specific cases of parallelepiped (see further below for details about these polyhedrons).

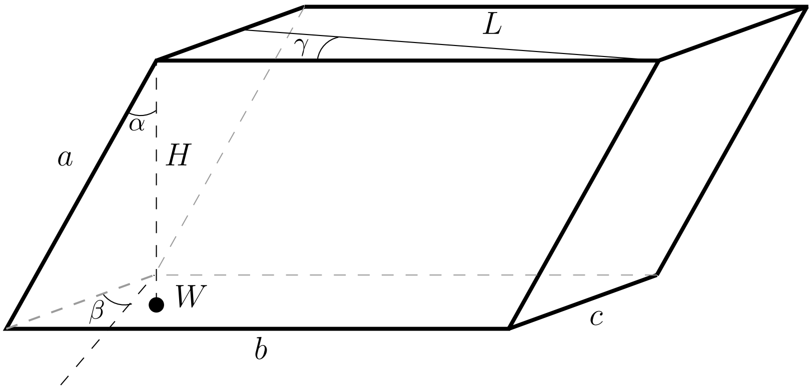


Figure 8.254: **General parallelepiped**

Definitions (#8.7.386):

D1. The parallelepiped is a "cube" if and only if:

$$a=b=c \qquad \text{and} \qquad \alpha=\beta=\gamma=0 \tag{8.7.109}$$

D2. The parallelepiped is a "cuboid" (rectangular parallelepiped) if and only if:

$$\alpha=\beta=\gamma=0 \tag{8.7.110}$$

D3. The parallelepiped is a "rhombohedron" if and only if:

$$a=b=c \tag{8.7.111}$$

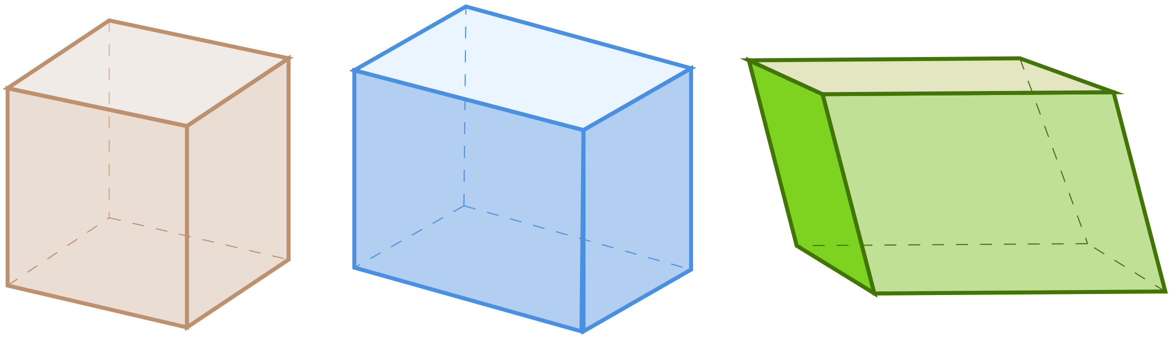


Figure 8.255: Cube, Cuboid and Rhombohedron

The volume of the parallelepiped is simply obtained by using trigonometric properties. We can consider mainly two situations:

- We know H, b, c, α and the volume is obviously given by:

$$\boxed{V = Hbc = \frac{\cos(\alpha)}{a}bc} \tag{8.7.112}$$

Obviously we get therefore for a cuboid where $a = b = c$:

$$\boxed{V = a^3} \tag{8.7.113}$$

- If we know only α, β, γ and H, W, L:

$$V = \frac{J}{\cos(\alpha)}\frac{L}{\cos(\gamma)}\frac{W}{\cos(\beta)} = abc \tag{8.7.114}$$

About its surface, it is simply the sum of the surfaces of each of its parallelogram without anything special (we have already prove earlier how to calculate the surface of a parallelogram) and therefore we need to know all deformation angles of each parallelogram.

8.7.2.1.1.1 Moment of Inertia of a rectangular plate

Now let us calculate the moment of inertia of a plate (rectangular parallelepiped) of cross-sectional area S which axis of rotation is the y-axis:

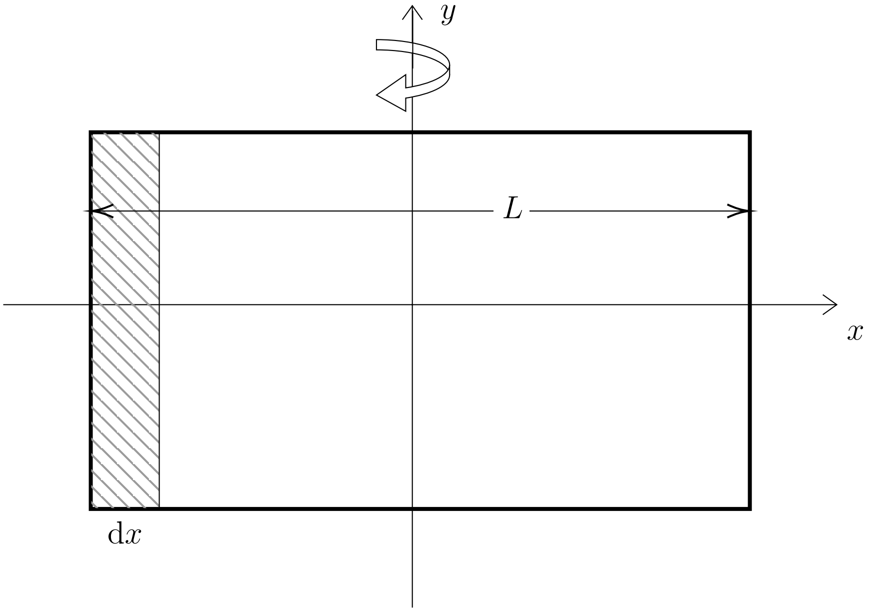


Figure 8.256: Search for the moment of inertia of a plate along the section

A volume element of the rectangle (in gray) is given by:

$$\mathrm{d}V = S\mathrm{d}x \Longrightarrow \mathrm{d}m = \rho S\mathrm{d}x \qquad (8.7.115)$$

and (see section Classical Mechanics page 3297):

$$J_y = \int x^2\mathrm{d}m = \int_{-L/2}^{+L/2} x^2\rho\mathrm{d}m = \rho S\left\{\frac{\left(\frac{L}{2}\right)^3}{3} - \frac{\left(\frac{L}{2}\right)^3}{3}\right\} = \rho\frac{SL^3}{12} = \rho SL\frac{L^2}{12} = M\frac{L^2}{12} \qquad (8.7.116)$$

Using Steiner's theorem (see section Classical Mechanics page 3310) we can easily get the moment of inertia of this plate when the rotation axis is collinear to the borders of the latter:

$$J_{y'} = M\frac{L^2}{36}\left(\frac{L}{e}\right)^2 M = M\frac{L^2}{12} + \frac{L^2}{4}M = \frac{4L^2M}{12} = \frac{L^2M}{3} \qquad (8.7.117)$$

and let us now focus on the moment of inertia of the plate relative to the z-axis (perpendicular to x and y therefore) and let us put axes so we have:

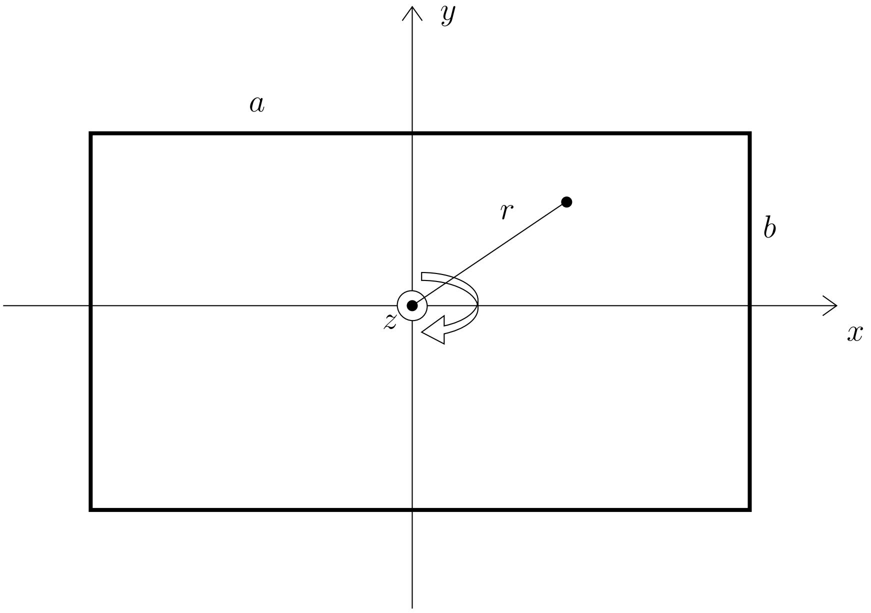


Figure 8.257: **Search for the moment of inertia of a plate along perpendicular axis**

We have:

$$J_z = \int r^2 \mathrm{d}m = \int (x^2+y^2)\mathrm{d}m = \int x^2 \mathrm{d}m + \int y^2 \mathrm{d}m = J_x + J_y \tag{8.7.118}$$

where r is in the x and y plane.

With:

$$\begin{aligned} J_x &= \int y^2 \mathrm{d}m = \int_{-b/2}^{+b/2} y^2 \rho e a \mathrm{d}y = \rho e a \int_{-b/2}^{+b/2} y^2 \mathrm{d}y = \frac{\rho e a}{3}\left\{\left(\frac{b}{2}\right)^3 - \left(-\frac{b}{2}\right)^3\right\} = \frac{M}{12}b^2 \\ J_y &= \int x^2 \mathrm{d}m = \int_{-a/2}^{+a/2} x^2 \rho e a \mathrm{d}x = \rho e a \int_{-a/2}^{+a/2} x^2 \mathrm{d}x = \frac{\rho e a}{3}\left\{\left(\frac{a}{2}\right)^3 - \left(-\frac{a}{2}\right)^3\right\} = \frac{M}{12}a^2 \end{aligned} \tag{8.7.119}$$

Therefore:

$$\boxed{J_{\square,z} = J_x + J_y = \frac{M}{12}\left(a^2+b^2\right)} \tag{8.7.120}$$

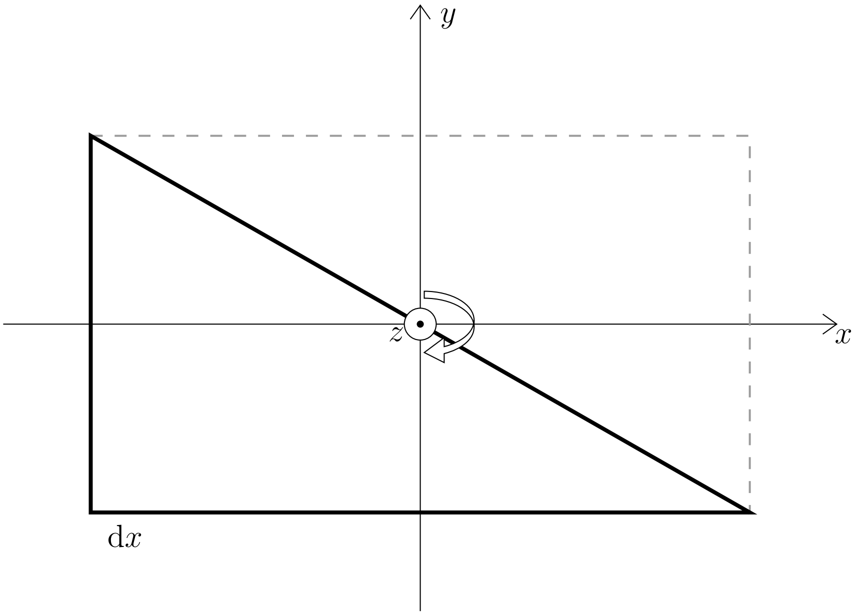
If the plate is a square of side L:

$$J_z = \frac{M}{12}\left(a^2+b^2\right) = \frac{M}{12}\left(L^2+L^2\right) = \frac{ML^2}{6} \tag{8.7.121}$$

8.7.2.1.1.2 Moment of Inertia of a triangular plate

We will now prove that it possible to calculate the moment of inertia of an equilateral triangle plate and rectangle triangle plate for an axis passing still through the centroid from the latter.

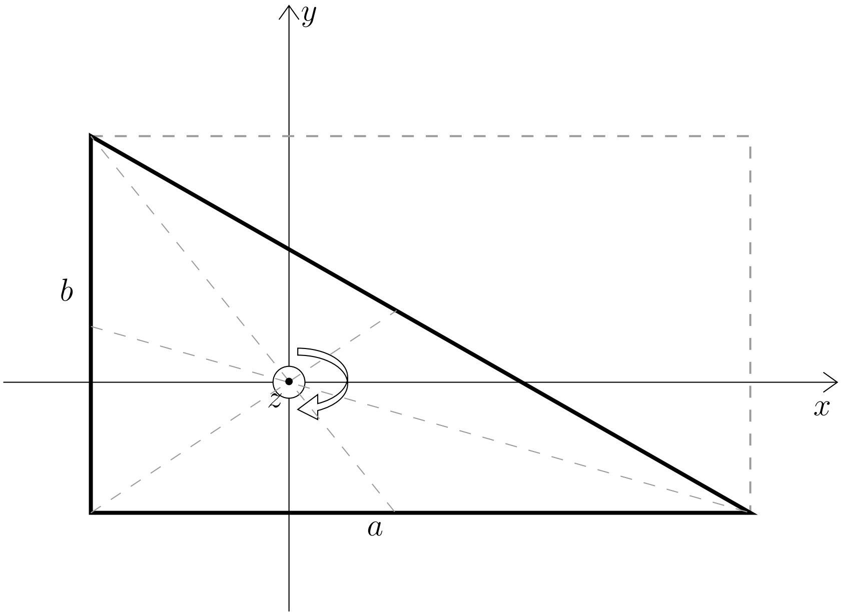
The method we propose here is direct and almost... brutal.... Indeed, we consider the moment of inertia always from the same axis, but for half the rectangle:



First we have obviously now:

$$J_z = \frac{1}{2}\frac{M}{12}(a^2+b^2) = \frac{M}{24}(a^2+b^2) \tag{8.7.122}$$

But we want the moment inertia of this equilateral triangle trough its centroid! That is to say:



Therefore as the center of gravity is placed now on the third of the median starting at the square's center (see proof previously) of gravity and that we make use of Steiner's theorem (see section Classical Mechanics page 3310) we get:

$$J_z=\frac{M}{24}(a^2+b^2)-\left(\frac{1}{3}\left(\sqrt{(0.5a)^2+(0.5b)^2}\right)\right)^2M=\frac{M}{24}(a^2+b^2)-\left(\frac{1}{6}\left(\sqrt{a^2+b^2}\right)\right)^2M \tag{8.7.123}$$

If $a=b=L$ we have the famous special case:

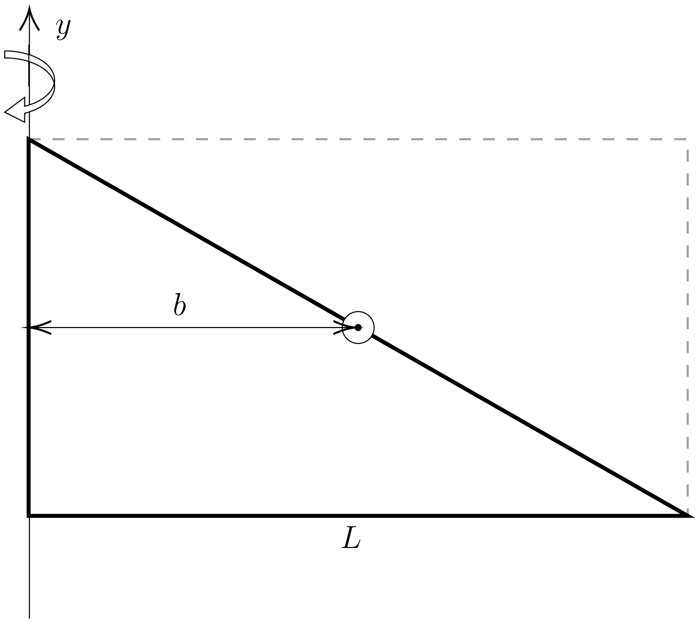
$$\begin{aligned}J_z&=\frac{M}{24}(L^2+L^2)-\left(\frac{1}{6}\left(\sqrt{L^2+L^2}\right)\right)^2M=\frac{L^2M}{12}-\left(\frac{1}{6}\left(\sqrt{2}L\right)\right)^2M\\&=\frac{L^2M}{12}-\frac{2L^2M}{36}=\frac{3L^2M}{36}-\frac{2L^2M}{36}=\frac{ML^2}{36}\end{aligned} \tag{8.7.124}$$

If we have $a=b=L$ then the rectangle has a height $h=L$ this is why we can found most of time the previous relation in textbooks written as:

$$\boxed{J_{\triangle,z}=\frac{Mh^2}{36}} \tag{8.7.125}$$

which is the moment of inertia of an equilateral triangle plate for the special case of an axis passing through the center of mass (gravity).

For sure as for other geometries we can choose any other axis and the possibilities are limitless... But to end we will calculate the moment of inertia of a rectangular triangle plate (right angle triangle) of dimensions a,b whose rotation axis is along the y-axis but translated of $b/2$ (that is to say: the rotation axis is collinear to the basis of length a of the rectangle triangle plate):



Remember for this that we have first proved just above that for a rectangular plate in rotation along the a border was:

$$J_{y'}=\frac{L^2M}{3} \tag{8.7.126}$$

Applying once again Steiner's theorem AND for half a plate (therefore for a rectangle triangle plate) we get immediately:

$$J_y = \frac{L^2M}{3} - \left(\frac{L}{2}\right)^2 M = \frac{L^2M}{3} - \frac{L^2M}{4} = \frac{4L^2M - 3L^2M}{12} = \frac{L^2M}{12} \tag{8.7.127}$$

And if we put as usual $L := h$ then we get the another famous result:

$$\boxed{J_y = \frac{h^2M}{12}} \tag{8.7.128}$$

As we can see the inertia momentum in this configuration is the same as that one with a rectangular plate having the rotation axis trough the centroid following the y-axis.

8.7.2.1.2 Pyramid

Definition (#8.7.387): The "pyramid" is a polyhedron whose base is a polygon and for side faces triangles connected on a single point made the "apex". The pyramid is not in the general case a regular polyhedron! A "right pyramid" has its apex directly above the centroid of its base. Non-right pyramids are named "oblique pyramids". A "regular pyramid" has a regular polygon base and is usually implied to be a right pyramid. When unspecified, a pyramid is usually assumed to be a "regular square pyramid".

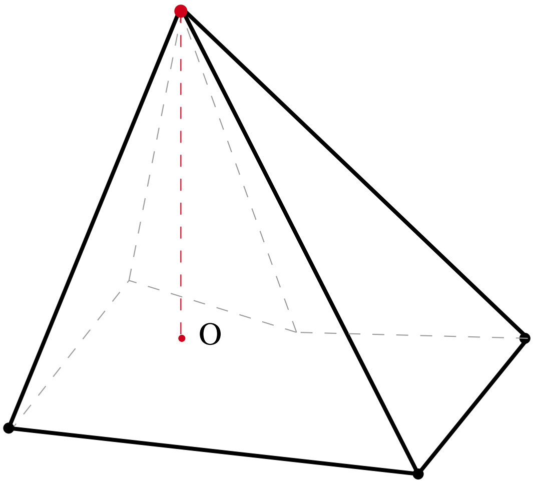


Figure 8.258: **Example of pyramid (irregular basis) but non-oblique**

Consider a surface $S(t)$ of the section of the pyramid with the plane of $z = t$ (that is to say for $t = 0$ we have the surface of the basis of the pyramid), then the volume V we are looking for will be equal to:

$$V = \int_0^h S(t)\mathrm{d}t \tag{8.7.129}$$

We speak of plane equation when there is no defined basis for now. In fact, in the integral, t varies between 0 and h. This implies that we take a basis centered on H (the projection of the apex of the pyramid on the basis), of axis therefore oriented towards $\overline{HO}$. The other two axes are chosen arbitrary in the plane of the basis of the pyramid.

We must now clarify what is $S(t)$ depending on t:

Let $S(0)=S$ be the area of the base of the pyramid. The section of the pyramid by the plane of equation $z=t$ is deduced by the scaling of center O and ratio t/h. So the integral is written:

$$V=\int_0^h S(t)\mathrm{d}t=\int_0^h S\frac{t^2}{h^2}\mathrm{d}t \tag{8.7.130}$$

where the fact that having taken the square of t/h is due to the fact that each inner term of S is the product of two terms (by the definition of a surface that is the multiplication of the length) each of homothetic ratio t/h.

Thus we have:

$$\boxed{V=\frac{S}{h^2}\int_0^h t^2\mathrm{d}t=\frac{Sh}{3}} \tag{8.7.131}$$

8.7.2.1.2.1 Moment of Inertia of a regular square pyramid

We have along the axis of symmetry of the regular square pyramid of height h and side a:

$$J_z=\int r^2\mathrm{d}m=\int(\sqrt{x^2+y^2})^2\mathrm{d}m=\int(x^2+y^2)\mathrm{d}m \tag{8.7.132}$$

As:

$$V=\frac{Sh}{3} \tag{8.7.133}$$

We have for the mass of the pyramid:

$$M=\rho\frac{Sh}{3}=\rho\frac{a^2h}{3} \tag{8.7.134}$$

So:

$$\mathrm{d}m=\rho\frac{\mathrm{d}x\mathrm{d}y\mathrm{d}z}{3} \tag{8.7.135}$$

Therefore:

$$\begin{aligned}
J_z=J_z&=\rho\int\limits_0^h \mathrm{d}z\int\limits_{-\frac{az}{2h}}^{\frac{az}{2h}}\mathrm{d}y\int\limits_{-\frac{az}{2h}}^{\frac{az}{2h}}(x^2+y^2)\mathrm{d}x=\rho\int\limits_0^h \mathrm{d}z\int\limits_{-\frac{az}{2h}}^{\frac{az}{2h}}\mathrm{d}y\left(\frac{1}{3}x^3+y^2x\right)\Bigg|_{-\frac{az}{2h}}^{\frac{az}{2h}}\\
&=2\rho\int\limits_0^h \mathrm{d}z\int\limits_{-\frac{az}{2h}}^{\frac{az}{2h}}\mathrm{d}y\left(\frac{1}{3}\frac{a^3z^3}{8h^3}+\frac{az}{2h}y^2\right)=2\rho\int\limits_0^h \mathrm{d}z\left(\frac{1}{3}\frac{a^3z^3}{8h^3}y+\frac{1}{3}\frac{az}{2h}y^3\right)\Bigg|_{-\frac{az}{2h}}^{\frac{az}{2h}}\\
&=4\rho\int\limits_0^h \mathrm{d}z\left(\frac{1}{3}\frac{a^4z^4}{16h^4}+\frac{1}{3}\frac{a^4z^4}{16h^4}\right)=4\rho\int\limits_0^h\frac{2}{3}\frac{a^4z^4}{16h^4}\mathrm{d}z=\frac{a^4\rho}{6h^4}\int_0^h z^4\mathrm{d}z=\frac{a^4\rho}{6h^4}\frac{1}{5}h^5\\
&=\frac{a^4\rho}{6}\frac{1}{5}h=\frac{1}{10}\frac{1}{3}a^2h\rho a^2=\frac{1}{10}ma^2
\end{aligned}\tag{8.7.136}$$

8.7.2.1.3 Right Prism

Definition (#8.7.388): The "(right) prism" is a polyhedron whose bases are two equals polygons with parallel sides (they have the same surface!), The side faces are parallelograms. Therefore, the prism is not a regular polyhedron! The two parallel sides and shapes are named bases of the (right) prism.

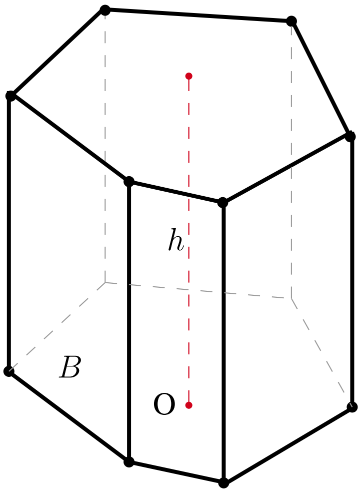


Figure 8.259: Example of right prism

To calculate the volume V of a right prism, we simply multiply the area of its base B by its height h:

$$V=Bh\tag{8.7.137}$$

As its base is a polygon, that is to say, it can be a triangle, a quadrilateral or a pentagon ... The reader must know how to calculate these areas to calculate the volume of the right prism.

A right prism with a triangular basis is named a "triangular prism", a right prism with a rectangular basis is named a... "cuboid", a right prism with a square base is named a... "cube" and a right prism with a circular basis is named a... "cylinder"...

8.7.2.1.4 Regular Polyhedron

Definitions (#8.7.389):

D1. A "regular polyhedron" is made of all identical and regular surfaces. A common equivalent definition is that a regular polyhedron is such that faces are congruent regular polygons which are assembled in the same way around each vertex.

D2. A "convex polyhedron" is such that each point of a line segment joining any two points belongs to the (inside of) polyhedron.

The regular polyhedra are at the number of nine, where five of them are convex and are known as the Platonic Solids. We sometimes named regular polyhedra only the Platonic solids and it is these that will interest us here.

Theorem 8.27. We will prove that there are only five convex regular polyhedra, which are therefore named the "five Platonic solids" (the other columns of the table below will be proved and explained a little further), a regular polyhedron is identified by its "Schläfli symbol" of the form (m, n), where m is the number of sides of each face and n the number of faces meeting at each vertex.

For the proof we will denote by A the number of edges, S the number of nodes (vertices) and F the number of faces.

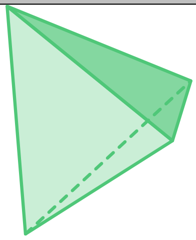
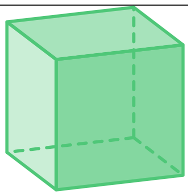
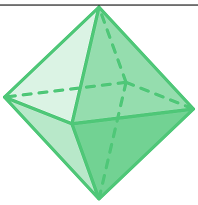
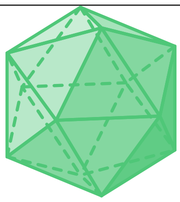
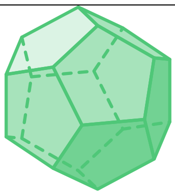
Name	**Schläfli** (m,n)	**Image**	S	A	F	$F-A+S$
Tetrahedron	$(3,3)$		4	6	4	2
Cube	$(4,3)$		8	12	6	2
Octahedron	$(3,4)$		6	12	8	2
Dodecahedron	$(5,3)$		20	30	12	2
Icosahedron	$(3,5)$		12	30	20	2

Table 8.7: **Five regular polyhedra (Platonic Solids)**

Remark

A "Platonic hydrocarbon" is a hydrocarbon (molecule) whose structure matches one of the five Platonic solids, with carbon atoms replacing its vertices, carbon–carbon bonds replacing its edges, and hydrogen atoms as needed. As far as we know, not all Platonic solids have molecular hydrocarbon counterparts.

Proof 8.27.1. Let m be the number of sides of each surface of a regular polyhedron, n the number of edges that meet at each vertex. We have therefore that each angle of any face is given by:

$$2\beta=\frac{\pi}{2}\left(2-\frac{4}{m}\right) \tag{8.7.138}$$

Caution!!! It is the angle 2β that defines the angle of a face and not just β!

Which is deduced from the following figure:

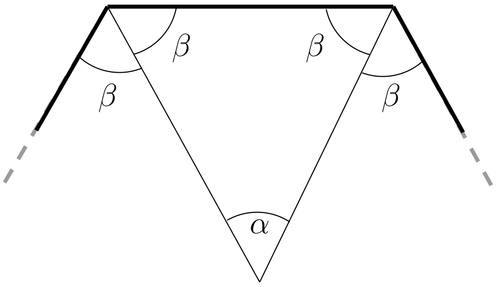


Figure 8.260: Angles between the edges of a polyhedron

where we have:

$$\alpha=\frac{2\pi}{m} \tag{8.7.139}$$

and:

$$2\beta=\pi-\alpha=\pi-\frac{2\pi}{m}=\frac{\pi}{2}\left(2-\frac{4}{m}\right) \tag{8.7.140}$$

But the sum of n angles grouped around a vertex is smaller than the n angles which intersect a plane into equal parts (we assume that intuitive by cutting)! Each of them is less than:

$$\frac{2\pi}{n} \tag{8.7.141}$$

therefore:

$$\frac{\pi}{2}\left(2-\frac{4}{m}\right)<\frac{2\pi}{n} \tag{8.7.142}$$

hence:

$$\frac{1}{m}+\frac{1}{n}>\frac{1}{2} \tag{8.7.143}$$

The numbers m and n are both at least equal to 3 (the smallest polygon being a triangle). It follows that the only possible cases are:

$$(m,n)=\{(3,3),(3,4),(4,3),(3,5),(5,3)\} \tag{8.7.144}$$

■ Q.E.D.

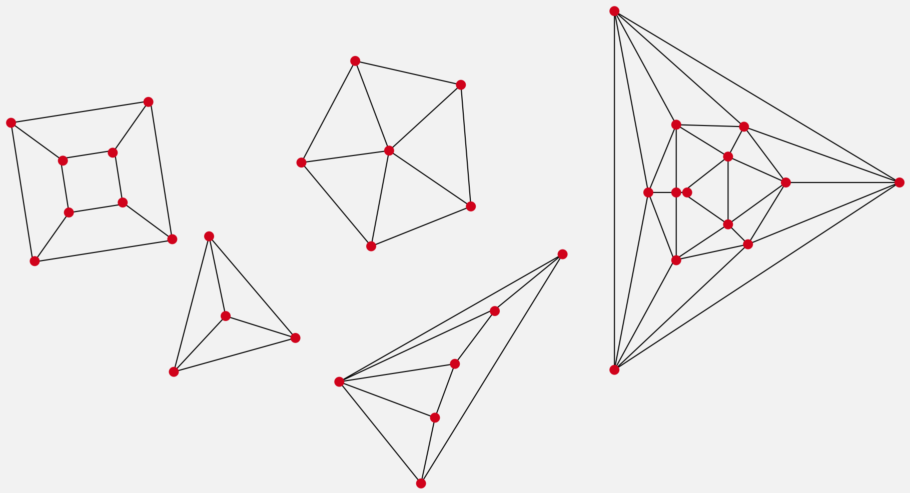
Let us denote now by F the number of faces, A the number of edges and S the number of vertices (nodes). Let recall that we have proved in the section of Graph Theory the "Euler's formula" at page 3002 (also named "Descartes-Euler's theorem" or "polyhedral formula") such that:

$$\boxed{F-A+S=2} \tag{8.7.145}$$

and it is of course also valid for the flattening of a polyhedron in the plane (and therefore of a polyhedron in the space).

Remark

The representation as a flattening graph of a polyhedron is named a "Schlegel diagram" and is used in some fields of organic chemistry.



In the case of regular polyhedra, each side (face) has m edges so that $m \cdot F$ is the whole number of edges of all sides and as each edge as encounter exactly two faces, we have the equality (take an example to be convinced if necessary!):

$$mF = 2A \tag{8.7.146}$$

and as n is the number of edges that meet at each vertex, and that each edge connects two vertices, we also have:

$$mF = nS \tag{8.7.147}$$

Therefore:

$$mF = nS = 2A \tag{8.7.148}$$

By injecting into Euler's formula, we have then:

$$\frac{2A}{m} - A + \frac{2A}{n} = 2 \Rightarrow \frac{1}{m} + \frac{1}{n} = \frac{1}{A} + \frac{1}{2} > 2 \tag{8.7.149}$$

and we fall back on the inequality of the previous theorem. Let us take again our calculation:

$$F - A + S = 2 = F - \frac{mF}{2} + \frac{mF}{n} = F\left(1 - \frac{m}{2} + \frac{m}{n}\right) \tag{8.7.150}$$

from which we get:

$$F = \frac{2}{1 - \dfrac{m}{2} + \dfrac{m}{n}} = \frac{2}{\dfrac{2n}{2n} - \dfrac{mn}{2n} + \dfrac{2m}{2n}} = \frac{4n}{2n - mn + 2m} \tag{8.7.151}$$

We can now undertake the classification of regular polyhedra:

1. The tetrahedron $(m,n) = (3,3)$:

$$F = \frac{4n}{2n - mn + 2m} = \frac{12}{6 - 9 + 6} = 4 \quad A = \frac{mF}{2} = \frac{12}{2} = 6 \quad S = \frac{mF}{n} = \frac{12}{3} = 4 \tag{8.7.152}$$

2. The octahedron $(m,n)=(3,4)$:

$$F=\frac{4n}{2n-mn+2m}=\frac{16}{8-12+6}=8 \quad A=\frac{mF}{2}=\frac{24}{2}=12 \quad S=\frac{mF}{n}=\frac{24}{2}=6 \tag{8.7.153}$$

3. The hexahedron or cube $(m,n)=(4,3)$:

$$F=\frac{4n}{2n-mn+2m}=\frac{16}{6-12+8}=6 \quad A=\frac{mF}{2}=\frac{24}{2}=12 \quad S=\frac{mF}{n}=\frac{24}{3}=6 \tag{8.7.154}$$

4. The icosahedron $(m,n)=(3,5)$:

$$F=\frac{4n}{2n-mn+2m}=\frac{20}{10-15+6}=20 \quad A=\frac{mF}{2}=\frac{60}{2}=30 \quad S=\frac{mF}{n}=\frac{60}{5}=12 \tag{8.7.155}$$

5. The dodecahedron $(m,n)=(5,3)$:

$$F=\frac{4n}{2n-mn+2m}=\frac{12}{6-15+10}=12 \quad A=\frac{mF}{2}=\frac{60}{2}=30 \quad S=\frac{mF}{n}=\frac{60}{3}=20 \tag{8.7.156}$$

this completes our classification.

Let us prove now that any polyhedron composed only of pentagons and hexagons has exactly mandatory twelve pentagons (and therefore there is no constraint on the number of hexagons). It is not a priori not very intuitive!

This proof will explain us for example why the soccer ball and the fullerene molecule have twelve pentagons:

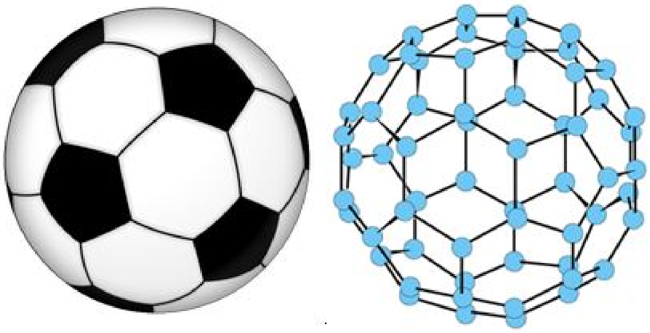


Figure 8.261: **Known polyhedra made of 12 pentagons**

Proof 8.27.2. For the proof, we start from the Euler's formula:

$$F-A+S=2 \tag{8.7.157}$$

Given M, the number of pentagons and N the number of hexagons, then we have the number of faces that is equal to:

$$F=M+N \tag{8.7.158}$$

In addition, each vertex is shared by three polygons in the case of the soccer ball and of the fullerene. Therefore, the number of vertices is equal to:

$$S=\frac{5M+6N}{3} \tag{8.7.159}$$

Each edge is shared by two polygons. Therefore, the number of edges is:

$$A=\frac{5M+6N}{2} \tag{8.7.160}$$

Then injecting Euler's formula:

$$\begin{aligned}F-A+S=2=(M+N)-\frac{5M+6N}{2}+\frac{5M+6N}{3}\\=\frac{6(M+N)-3(5M+6N)+2(5M+6N)}{6}=\frac{M}{6}\end{aligned} \tag{8.7.161}$$

Therefore, to satisfy:

$$\frac{M}{6}=2 \tag{8.7.162}$$

M, the number of pentagons, must be equal to 12.

■ Q.E.D.

8.7.2.1.5 Regular Tetrahedron

A tetrahedron is a polyhedron composed of 4 triangular faces, 6 straight edges, and 4 vertex corners. The tetrahedron is the simplest of all the ordinary convex polyhedra and the only one that has fewer than 5 faces.

The tetrahedron is one kind of pyramid, which is a polyhedron with a flat polygon base and triangular faces connecting the base to a common point. In the case of a tetrahedron the base is a triangle (any of the four faces can be considered the base), so a tetrahedron is also known as a "triangular pyramid".

Definition (#8.7.390): A "regular tetrahedron" is one in which all 4 faces are equilateral triangles. It is one of the five regular Platonic solids, which have been known since antiquity.

We have shown that for the tetrahedron $F=4,m=3$ and it is relatively easy to guess that such a polyhedron is formed by 3 identical equilateral triangles as shown in the figure below:

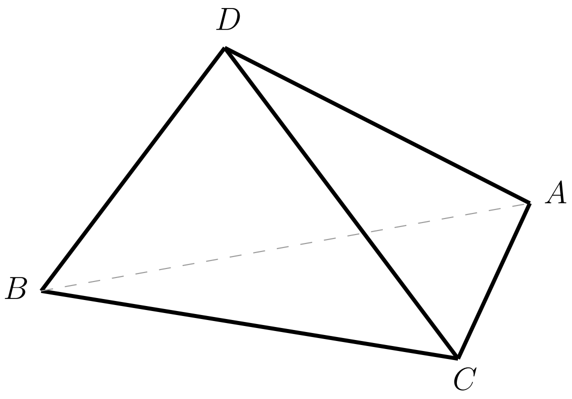


Figure 8.262: **Example of regular tetrahedron**

For this purpose let us start by studying the following equilateral triangle:

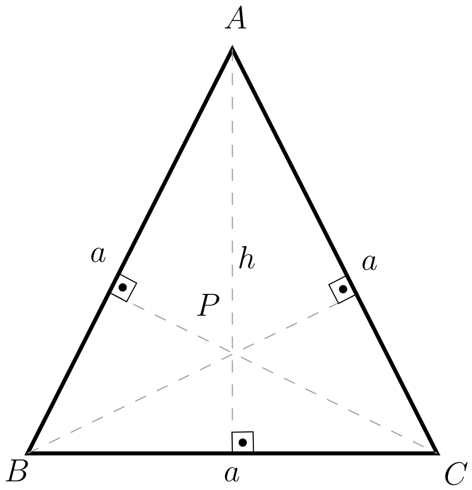


Figure 8.263: **Equilateral triangle to start with to calculate the volume of the tetrahedron**

In this equilateral triangle, a is the side, h is the height. The perpendicular bisectors are h, h', h'' of the respective sides $\overline{BC}$, $\overline{AB}$, $\overline{AC}$.

The segments h and h' intersect at a point P (barycenter). By construction of the equilateral triangle, we have $\overline{PC}=\overline{PB}=\overline{PA}$ (just apply Pythagoras theorem to prove it if you want).

Furthermore, we have already proved in our study of the triangle, that the barycenter of the latter is always at $2/3$ of the median height. As medians and bisectors are merged in the case of the equilateral triangle, then we have $\overline{AP}=2/3\cdot h$.

Now, if we draw a straight line passing through the point P and perpendicular to the plane in which the triangle is located. Given D a point on this line, as we have $\overline{PB}=\overline{PC}=\overline{PA}=2/3\cdot h$ we will have of course $\overline{BD}=\overline{AD}=\overline{CD}$ (again just apply Pythagoras theorem!).

We only have to find a way such that $\overline{BD}=a$ and we will have the regular tetrahedron we wanted. We then calculate:

$$\overline{BP}^2+\overline{PD}^2=\overline{BD}^2=a^2 \tag{8.7.163}$$

and therefore:

$$\overline{PD}^2=a^2-\left(\frac{2}{3}h\right)^2=a^2-\frac{4}{9}h^2=a^2-\frac{4}{9}\left(a^2-\left(\frac{1}{2}a\right)^2\right)=a^2-\frac{4}{9}\left(\frac{3}{4}a^2\right)=\frac{2}{3}a^2 \quad (8.7.164)$$

therefore:

$$H=\sqrt{\frac{2}{3}}a \quad (8.7.165)$$

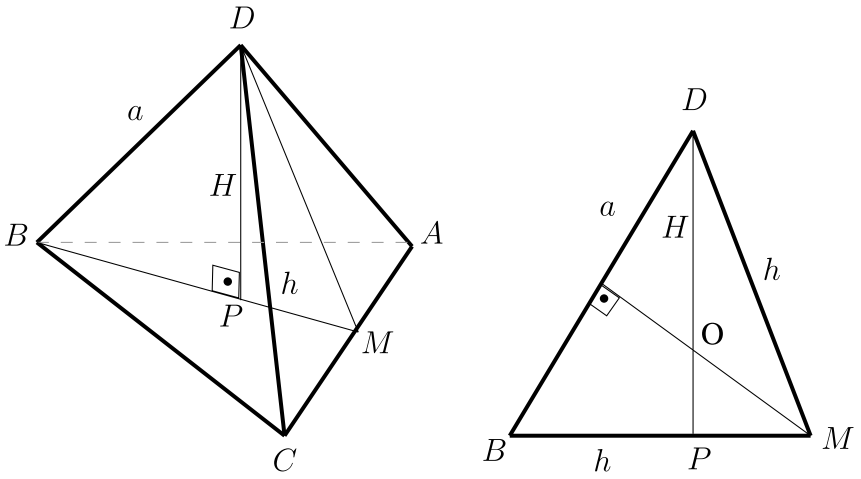


Figure 8.264: **Tetrahedron variables summary**

The bisector of $\overline{BD}$ (on the right figure) passing through M intersects H at a point O, which is nothing more than the center of the sphere circumscribed to the tetrahedron. Indeed, by construction, we have $\overline{\text{O}B}=\overline{\text{O}C}=\overline{\text{O}A}$ and the bisector gives us $\overline{\text{O}B}=\overline{\text{O}D}$.

The Thales theorem (see section Plane Geometry page 2610) also gives us:

$$\frac{\overline{\text{O}D}}{\frac{1}{2}a}=\frac{a}{\overline{PD}}\Rightarrow\overline{\text{O}D}=\frac{1}{2}a^2\frac{1}{\overline{PD}}=\frac{1}{2}a^2\frac{1}{\frac{\sqrt{2}}{\sqrt{2}}a}=\frac{\sqrt{3}}{\sqrt{2}}\frac{a}{2} \quad (8.7.166)$$

and for the developments that will follows we will put $H=\overline{PD}$, $\overline{\text{O}D}=R$.

Now let us calculate the total area. It will necessarily be given by the surface of one side multiplied by the number of faces, and as we have proved how to calculate the area of a triangle above it comes immediately:

$$S=n\left(\frac{1}{2}\cdot\text{basis}\cdot\text{height}\right)=4\frac{1}{2}ah=4\frac{1}{2}a\frac{\sqrt{3}a}{2}=\sqrt{3}a^2 \quad (8.7.167)$$

For the volume, it's also just as simple as we have proved earlier above that is was equal to that of a pyramid. It then comes immediately:

$$V=\frac{1}{3}\cdot\text{basis}\cdot\text{height}=\frac{1}{3}\frac{S}{4}\cdot H=\frac{1}{3}\frac{\sqrt{3}a^2}{4}\frac{\sqrt{2}}{\sqrt{3}}a=\frac{\sqrt{2}}{12}a^3 \quad (8.7.168)$$

Therefore:

$$\boxed{S=\sqrt{3}a^2} \tag{8.7.169}$$

and:

$$\boxed{V=\frac{\sqrt{2}}{12}a^3} \tag{8.7.170}$$

8.7.2.1.6 Regular hexahedron (cube)

The cube is a regular polyhedron that is probably the most familiar to us, it has 6 faces and its construction does not require to be presented.

The cube is also a square parallelepiped, an equilateral cuboid and a right rhombohedron. It is a regular square prism in three orientations, and a trigonal trapezohedron in four orientations.

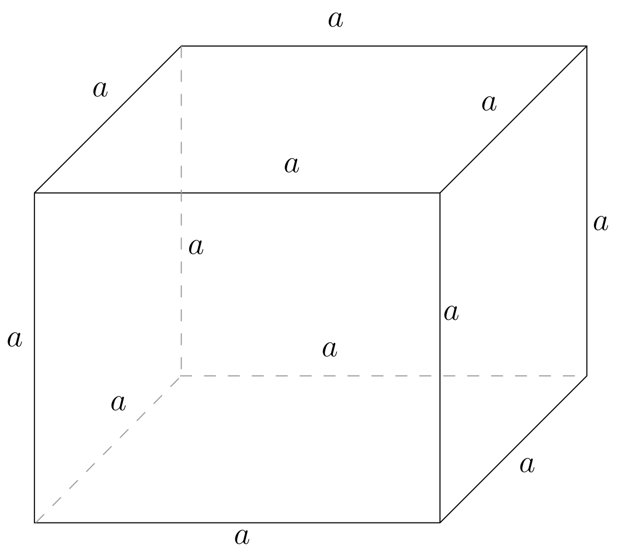


Figure 8.265: Example of regular hexahedron (cube)

Since all sides are of length a, the surface is simply given by the multiplication of the 6 faces surfaces. So:

$$\boxed{S=6a^2} \tag{8.7.171}$$

and the volume:

$$\boxed{V=a^3} \tag{8.7.172}$$

8.7.2.1.7 Regular octahedron

We have shown that for the octahedron $F=8, m=4$ and it is relatively easy to guess that the regular octahedron is formed (by definition!!!) of 8 identical equilateral triangles.

A regular octahedron is a square bi-pyramid in any of three orthogonal orientations. It is also a triangular anti-prism in any of four orientations.

To build, and demonstrate that it is possible to build such a polyhedron, we put as before that his side is equal to a:

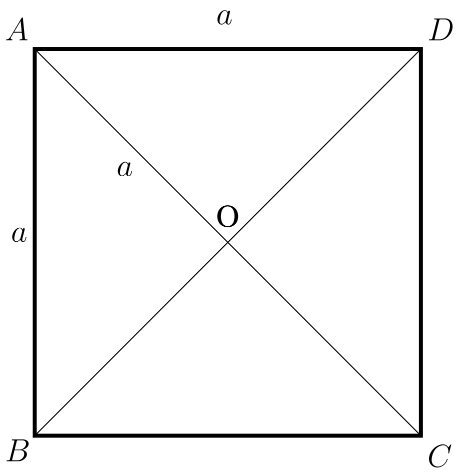


Figure 8.266: Example of regular octahedron seen from above

Then we denote by O the point of intersection of the two diagonals. We have therefore by applying the Pythagorean theorem:

$$\overline{AC}^2=a^2+a^2=2a^2 \Rightarrow \overline{AC}=\sqrt{2}a \tag{8.7.173}$$

and:

$$\overline{\text{O}A}^2=\frac{1}{2}\overline{AC}=\frac{a}{\sqrt{2}} \tag{8.7.174}$$

On the perpendicular line to the plane containing our square, and passing through O, we add two vertices E, F at a distance that we calculate as follows:

$$\overline{\text{O}A}^2+\overline{\text{O}E}^2=\overline{AE}^2=a^2 \tag{8.7.175}$$

from which we get:

$$\overline{\text{O}E}^2=\overline{AE}^2-\overline{\text{O}A}^2=a^2-\frac{1}{2}a^2=\frac{1}{2}a^2 \tag{8.7.176}$$

Therefore:

$$\overline{\text{O}E}^2=\frac{a}{\sqrt{2}}=\overline{\text{O}A}=-\overline{\text{O}F} \tag{8.7.177}$$

Our polyhedron is then well formed by 8 equilateral identical triangles. Each vertex has 4 corners and 4 faces, which allows us to affirm that it is regular and so ends our construction.

The surface of the regular octahedron is:

$$S=n\frac{1}{2}ah=8\frac{1}{2}a\frac{\sqrt{3}a}{2}=2\sqrt{3}a^2 \tag{8.7.178}$$

where h is the height of the equilateral triangle of side a that we have already calculated above.

For the volume, this is still based on that of the pyramid. So:

$$V=2\frac{1}{3}\text{basis}\cdot\text{height}=2\frac{1}{3}a^2\frac{a}{\sqrt{2}}=\frac{\sqrt{2}}{3}a^2 \tag{8.7.179}$$

And we will assume that it is obvious to the reader that our octahedron is inscribed in a sphere of radius R whose center is the point O. For R, we have:

$$R=\overline{\text{O}A}=\overline{\text{O}E}=\frac{a}{\sqrt{2}} \tag{8.7.180}$$

Let us show already that now we can build the regular icosahedron from the octahedron and this latter exists and is buildable.

For this purpose, we will first consider a reference frame of the octahedron placed at the origin O corresponding to its barycenter as visible in the figure below:

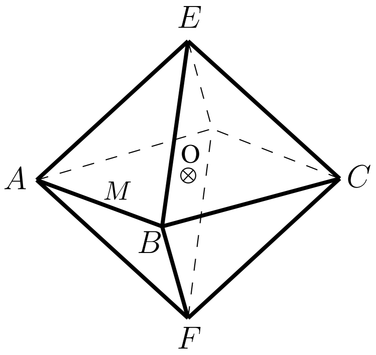


Figure 8.267: **Example of the octahedron**

We then have:

$$\begin{aligned}\vec{A}&=(R,0,0)\\ \vec{B}&=(0,R,0)\\ \vec{C}&=(-R,0,0)\\ \vec{D}&=(0,-R,0)\\ \vec{E}&=(0,0,R)\\ \vec{F}&=(0,0,-R)\end{aligned} \tag{8.7.181}$$

After this, let us consider the following figure:

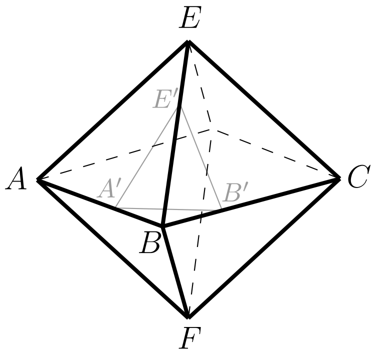


Figure 8.268: Searches of the position of the equilateral triangle

In the figure above, A' is a point that start from A and that arrives in B, and given B' a point which start from C and arrives at B, and finally E' a point that start from B and arrives on E. These three points go together and moving at the same speed. If we follow these three points that form a triangle $A'B'E'$, we feel intuitively that there is a place such that $A'B'E'$ is an equilateral triangle.

Let us determine this place:

$$\begin{aligned}\overrightarrow{AB}&=\overrightarrow{\mathrm{O}B}-\overrightarrow{\mathrm{O}A}=(-R,R,0) & \overrightarrow{\mathrm{O}A'}&=\overrightarrow{\mathrm{O}A}+\mu\overrightarrow{AB}=R(1-\mu,\mu,0)\\ \overrightarrow{CB}&=\overrightarrow{\mathrm{O}B}-\overrightarrow{\mathrm{O}C}=(R,R,0) & \overrightarrow{\mathrm{O}B'}&=\overrightarrow{\mathrm{O}C}+\mu\overrightarrow{CB}=R(\mu-1,\mu,0)\\ \overrightarrow{BE}&=\overrightarrow{\mathrm{O}E}-\overrightarrow{\mathrm{O}B}=(0,-R,R) & \overrightarrow{\mathrm{O}E'}&=\overrightarrow{\mathrm{O}B}+\mu\overrightarrow{BE}=R(0,1-\mu,\mu)\end{aligned} \tag{8.7.182}$$

and therefore:

$$\begin{aligned}\overrightarrow{A'B'}&=\overrightarrow{\mathrm{O}B'}-\overrightarrow{\mathrm{O}A'}=R(2\mu-2,0,0)\\ \overrightarrow{A'E'}&=\overrightarrow{\mathrm{O}E'}-\overrightarrow{\mathrm{O}A'}=R(\mu-1,1-2\mu,\mu)\end{aligned} \tag{8.7.183}$$

and we want:

$$||\overrightarrow{A'B'}||=||\overrightarrow{A'E'}||=||\overrightarrow{B'E'}|| \tag{8.7.184}$$

Therefore:

$$||\overrightarrow{A'B'}||^2=R^2(2\mu-2)^2=||\overrightarrow{A'E'}||^2=R^2\left[(\mu-1)^2+(1-2\mu)^2+\mu^2\right] \tag{8.7.185}$$

That is to say:

$$4\mu^2-8\mu+4=6\mu^2-6\mu+2 \tag{8.7.186}$$

Which simplifies in:

$$\mu^2-\mu-1=0 \tag{8.7.187}$$

and as $0\leq\mu\leq 1$, we get for the resolution of this polynomial of the second degree (see section Calculus page 1382) the only acceptable solution:

$$\mu=\frac{1}{2}(-1+\sqrt{5}) \tag{8.7.188}$$

where the reader may have perhaps noticed that this is the inverse of the golden ratio...

According to the figure below, if we put:

$$\overrightarrow{\text{O}I'}=\overrightarrow{\text{O}E}+\mu\overrightarrow{EA} \tag{8.7.189}$$

then we fall back on the same value for μ (the reader can check this himself, or we can on request add the detailed calculations here) and ditto for all other points:

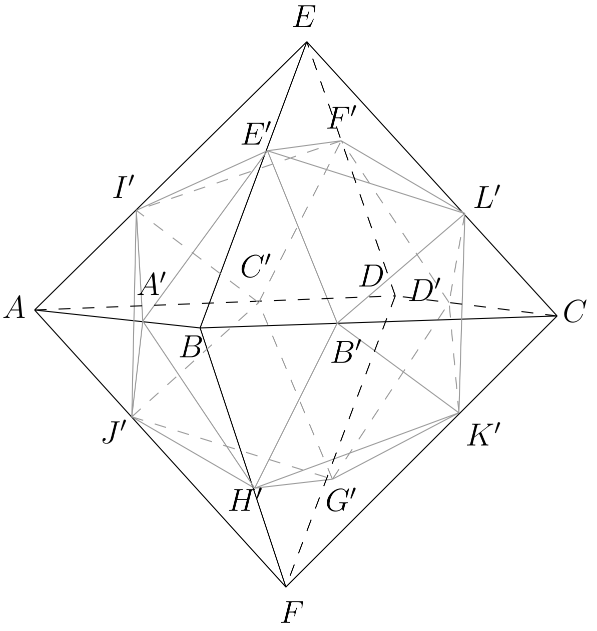


Figure 8.269: **Construction of a regular icosahedron**

Our new polyhedron therefore has one face by face of the octahedron and one face by edge of the octahedron. We have therefore 20 faces composed of identical equilateral triangles. In addition, 5 edges and 5 faces meet at each vertices. We then get a regular icosahedron!!!

We then have for the coordinates of each vertex (the reader must be observed that the vertices are opposed in pairs one component on the figure):

$$\begin{array}{ll}
\vec{A}'=R(1-\mu,\mu,0) & \vec{D}'=R(1-\mu,-\mu,0)\\
\vec{B}'=R(\mu-1,\mu,0) & \vec{C}'=R(1-\mu,-\mu,0)\\
\vec{E}'=R(0,1-\mu,\mu) & \vec{H}'=R(0,1-\mu,-\mu)\\
\vec{F}'=R(0,\mu-1,\mu) & \vec{G}'=R(0,\mu-1,-\mu)\\
\vec{I}'=R(\mu,0,1-\mu) & \vec{L}'=R(-\mu,0,1-\mu)\\
\vec{J}'=R(\mu,0,\mu-1) & \vec{K}'=R(-\mu,0,1-\mu)
\end{array} \tag{8.7.190}$$

8.7.2.1.8 Regular Icosahedron

We saw earlier how to build the regular icosahedron. There it exists!

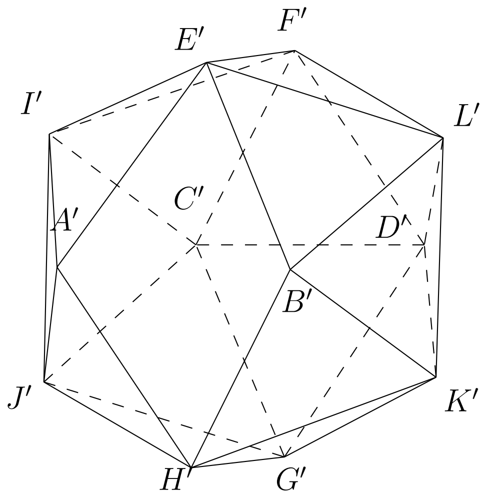


Figure 8.270: **Representation of the regular icosahedron**

Knowing the coordinates of the different edges, let us now calculate the surface and volume of the regular icosahedron.

The calculation of the surface is simple since it is 20 equilateral triangles. We have then:

$$\overrightarrow{A'B'} = R(2\mu - 2, 0, 0) \tag{8.7.191}$$

therefore:

$$||\overrightarrow{A'B'}|| = 2R(\mu - 1) = a' \tag{8.7.192}$$

Hence:

$$S = 20\frac{1}{2}a'h = 20\frac{1}{2}a'\sqrt{a'^2 - \left(\frac{1}{2}a'\right)^2} = 20\frac{1}{2}a'2\sqrt{\frac{3}{4}a'^2} = 20\frac{1}{2}a'\frac{\sqrt{3}}{2}a' \tag{8.7.193}$$

Therefore:

$$\boxed{S = 5\sqrt{3}a'^2} \tag{8.7.194}$$

The calculation of the volume being quite more subtle! Let's go for it...

The icosahedron is built around the pentagon and the golden section as we were able to notice in our previous study of the octahedron.

If ever the reader is not convinced here is an additional figure where we see that each edge of the icosahedron is an edge of a pentagon ($AFECB$, $LGHJK$, $DAJKC$, $DEGHA$, $BJILC$, $FELIH$,...):

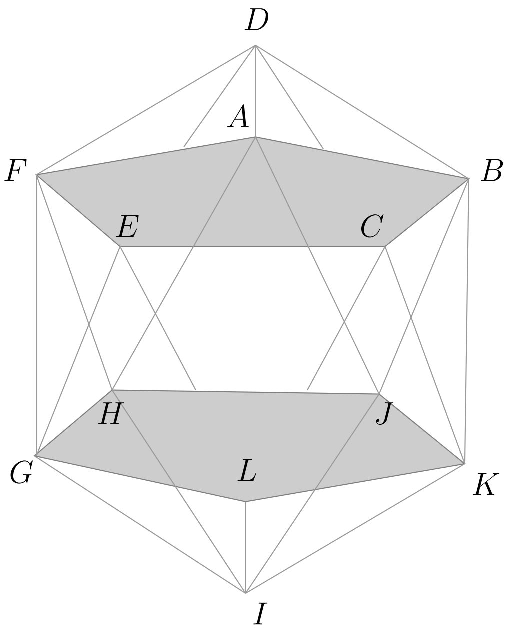


Figure 8.271: **Pentagons in the regular icosahedron**

Using the method of the pyramids, we have 20 equilateral triangles that form the basis of a pyramid whose height goes up to the origin O of the icosahedron (or the origin confused the with circumscribed or inscribed corresponding sphere).

Take for example the base ABD with the intersection of mediators located at the point M as shown below:

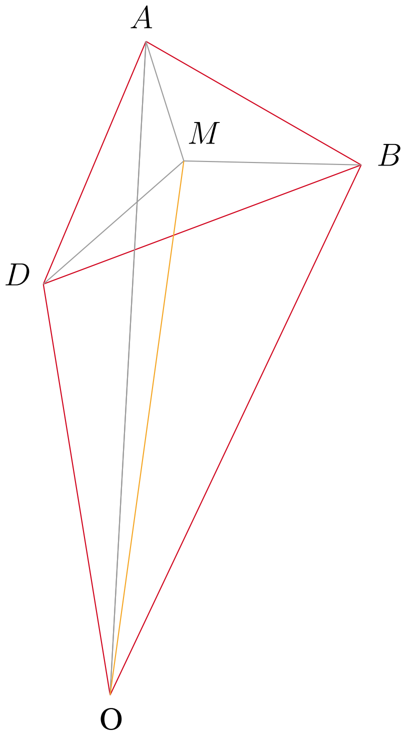


Figure 8.272: **Representation of one of the pyramids of the regular icosahedron**

As we know, we proved that the volume of a pyramid is:

$$V_p = \frac{1}{3}bh \tag{8.7.195}$$

The surface b in our situation that of the equilateral triangle ΔADB and the height h is the segment $\overline{\text{O}M}$.

If we denote by a the triangle side, then the surface is given by:

$$S = \frac{ah}{2} = \frac{a\sqrt{a^2 - \left(\frac{1}{2}a\right)^2}}{a} = \frac{\sqrt{3}}{4}a^2 \tag{8.7.196}$$

To find h, we know by construction of the point M that the triangles $\text{O}MA$, $\text{O}MB$ and $\text{O}MD$ are right triangles.

Let us work arbitrarily with the triangle $\text{O}MD$. First, let determine the length $\overline{DM}$. We have proved in our study of the bisector of length H of the equilateral triangle (see section Euclidean Geometry page 2601) that $\overline{DM}$ is then equal to:

$$\overline{DM} = \frac{2}{3}H \tag{8.7.197}$$

But:

$$H = \sqrt{a^2 - \left(\frac{1}{2}a\right)^2} = \frac{\sqrt{3}}{2}a \tag{8.7.198}$$

So finally:

$$\overline{DM} = \frac{2}{3}\frac{\sqrt{3}}{2}a = \frac{\sqrt{3}}{3}a = \frac{1}{\sqrt{3}}a \tag{8.7.199}$$

To find h we must find the length $\overline{\text{O}D} = r$ in terms of length of the edges a of icosahedron. For this, we must recognize one of the fundamental properties of the icosahedron.

Before going further, let us show a property of the pentagon below with its diagonals d and sides c:

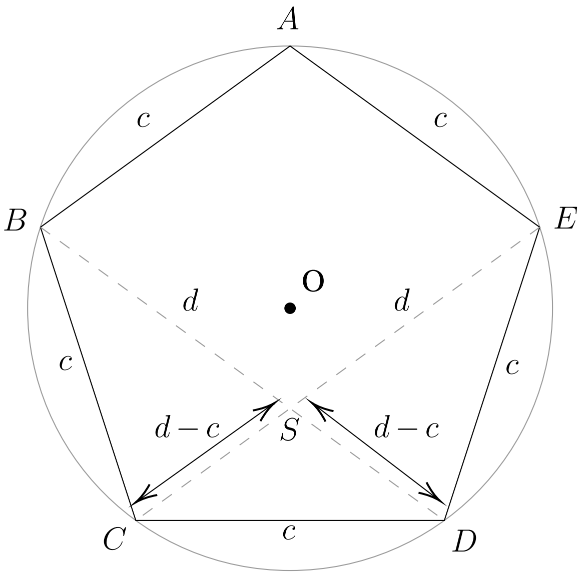


Figure 8.273: **Parallelogram in a pentagon**

$BSEA$ is a parallelogram. Indeed, the diagonal $\overline{BD}$ is parallel to the side $\overline{AE}$ (for example, because both are perpendicular to the axis of symmetry passing through $\overline{\text{O}C}$). Since S is on $\overline{BD}$, it proves that $\overline{BS}$ and $\overline{AE}$ are parallel. We show in the same manner that $\overline{AB}$ and $\overline{SE}$ are parallel.

We deduce that:

$$\overline{SD}=\overline{BD}-\overline{BS}=\overline{BD}-\overline{AE}=d-c \qquad (8.7.200)$$

and same for $\overline{CS}$:

$$\overline{CS}=\overline{EC}-\overline{ES}=\overline{EC}-\overline{AB}=d-c \qquad (8.7.201)$$

Let us continue ... we have the equality $\widehat{BAE}=\widehat{CSD}$. Furthermore, as CD and BE are parallel, the triangles SCD and ABE are similar. Therefore, the distances between their sides are kept (Thales):

$$\frac{\overline{AE}}{\overline{BE}}=\frac{\overline{SD}}{\overline{CD}} \qquad (8.7.202)$$

hence the relation:

$$\frac{c}{d}=\frac{d-c}{c} \qquad (8.7.203)$$

After some modifications:

$$\frac{d}{c}-\frac{c}{d-c}=0\Leftrightarrow\frac{d(d-c)-c\cdot c}{c(d-c)}=0\Leftrightarrow d^2-dc-c^2=0\Leftrightarrow\frac{d^2}{c^2}-\frac{d}{c}-1=0 \qquad (8.7.204)$$

If μ now designates the ratio d/c, the above relation becomes:

$$\mu^2-\mu-1=0 \qquad (8.7.205)$$

and μ being strictly positive, we have already seen during our study of the octahedron that the only positive root is the golden ratio:

$$\mu=\frac{d}{c}=\frac{1+\sqrt{5}}{2} \qquad (8.7.206)$$

We have just finish to prove that a diagonal of a pentagon is equal to the golden ratio multiplied by the length of an edge of this same pentagon.

Thus we have in the pentagons $AFECB$ and $LGHJK$ of our icosahedron:

$$\overline{FB}=\overline{GK}=\mu a \qquad (8.7.207)$$

Let us notice also the rectangle $FBGK$ whose barycenter (centroid) coincides with that of the icosahedron. Moreover, $\overline{FK}$ and $\overline{BG}$ represent by construction the diameter of the sphere circumscribed to the icosahedron and therefore $\overline{OF}=\overline{OK}$ is the radius r that we are looking for.

We have:

$$\overline{FK}^2=\overline{FB}^2+\overline{BK}^2=\mu^2a^2+a^2=(\mu^2+1)a^2 \qquad (8.7.208)$$

Therefore:

$$\overline{FK} = a\sqrt{\mu^2+1} = 2r \tag{8.7.209}$$

hence:

$$r = \frac{a\sqrt{\mu^2+1}}{2} \tag{8.7.210}$$

Now, we can calculate h:

$$h^2 = \overline{OM}^2 = r^2 - \overline{DM}^2 = \frac{a^2(\mu^2+1)}{4} - \frac{1}{3}a^2 = \frac{3(\mu^2+1)-4}{12}a^2 = \frac{3\mu^2-1}{12}a^2 \tag{8.7.211}$$

But:

$$\mu^2 = \left(\frac{1+\sqrt{5}}{2}\right)^2 = \frac{6+2\sqrt{5}}{4} = \mu+1 \tag{8.7.212}$$

as the golden ratio is a root of $x^2 = x+1$.

Therefore:

$$(\mu^2)^2 = (\mu+1)^2 = \mu^2+2\mu+1 = \mu^2+2(\mu^2-1)+1 = 3\mu^2-1 \tag{8.7.213}$$

Finally:

$$h^2 = \frac{3\mu^2-1}{12}a^2 = \frac{\mu^4}{12}a^2 \tag{8.7.214}$$

and:

$$h = \frac{\mu^2}{2\sqrt{3}}a \tag{8.7.215}$$

Therefore the volume of one pyramid of the icosahedron is:

$$V_{\text{pyr}} = \frac{1}{3}Sh = \frac{1}{3}\frac{\sqrt{3}}{4}a^2\frac{\mu^2}{2\sqrt{3}}a = \frac{\mu^2}{24}a^3 \tag{8.7.216}$$

As the are 20 pyramids:

$$\boxed{V = 20\frac{\mu^2}{24}a^3 = \frac{5}{6}\mu^2a^3} \tag{8.7.217}$$

8.7.2.1.9 Regular Dodecahedron

After failing to find in the literature an aesthetic and simple way the of the proof of the construction feasibility of the dodecahedron, we will move in for now (it is possible to live without it...).

Let us simply notice that the dodecahedron consists of 12 pentagons and its volume is comparable to that of a cube on which we have placed on each face a sort of little roof which ultimately will give the pentagons:

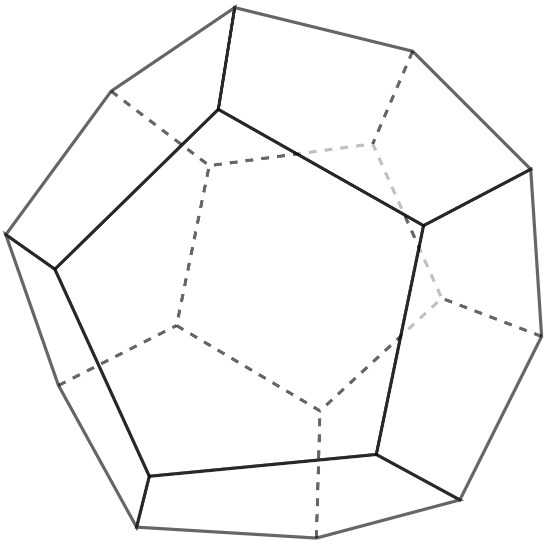


Figure 8.274: **Regular dodecahedron**

For our study of the dodecahedron, we will only focus to determine its surface and volume.

For this, let us consider first the regular pentagon below:

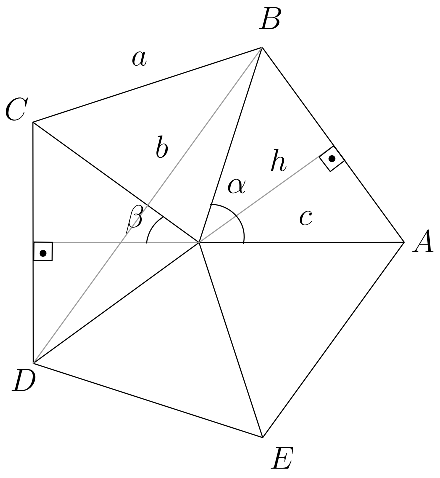


Figure 8.275: **Regular pentagon for the study of the volume of the regular dodecahedron**

We will first have to determine the length of h and b.

Let us first recall that we have already during our study of the icosahedron proved that the diagonal of a pentagon is connected to the length of its sides by the relation:

$$b = a\mu = a\frac{1+\sqrt{5}}{2} \tag{8.7.218}$$

where μ is the golden ratio. It remains to then to determine h.

It is obvious first that $\beta = 2\pi/10 = \pi/5 = \alpha/2$ and that:

$$c \cdot \cos(\beta) = h \tag{8.7.219}$$

But we have two missing information here: the angle and c. Begin by determining how much is the cosine without using the calculator (you will understand why ...).

We have first by the identity (see section Trigonometry page 2519):

$$\cos(2x)=2\cos^2(x)-1 \tag{8.7.220}$$

that:

$$\cos\left(4\frac{\pi}{5}\right)=2\cos^2\left(2\frac{\pi}{5}\right)-1 \tag{8.7.221}$$

Which can also be written:

$$\cos\left(\pi-\frac{\pi}{5}\right)=2\cos^2\left(2\frac{\pi}{5}\right)-1 \tag{8.7.222}$$

But this can also be written always using the same trigonometric identity:

$$\cos\left(\pi-\frac{\pi}{5}\right)=2\left[2\cos^2\left(\frac{\pi}{5}\right)-1\right]-1 \tag{8.7.223}$$

Thus after simplification:

$$-\cos\left(\frac{\pi}{5}\right)=8\cos^4\left(\frac{\pi}{5}\right)-8\cos^2\left(\frac{\pi}{5}\right)+1 \tag{8.7.224}$$

By a change of variable and rearranging the different terms:

$$8X^4-8X^2+X+1=0 \tag{8.7.225}$$

We have -1 and $1/2$ that are two obvious roots so we get (see section Calculus page 1383):

$$(X+1)\left(X-\frac{1}{2}\right)\left(X^2-\frac{1}{2}X-\frac{1}{4}\right)=0 \tag{8.7.226}$$

We just have to solve a simple equation of the second degree, the solution is trivial using the methods seen in the section Calculus and we get:

$$X=\frac{1\pm\sqrt{5}}{4} \tag{8.7.227}$$

Either by taking the only feasible solution, then we have:

$$X=\frac{\mu}{2}=\frac{1+\sqrt{5}}{4} \tag{8.7.228}$$

So we fall back on the golden ratio again! and this leads us directly to write:

$$\boxed{h=\frac{1}{2}\mu c=\frac{1}{2}\frac{1+\sqrt{5}}{2}c} \tag{8.7.229}$$

It remains to us to determine c. We have:

$$\left(\frac{1}{2}a\right)^2+h^2=c^2\frac{1}{4}a^2+\frac{1}{4}\mu^2c^2\Rightarrow c^2(4-\mu^2)=a^2 \tag{8.7.230}$$

and as $\mu^2-\mu-1=0$ we have:

$$\begin{aligned}&(4-\mu^2)(2+\mu)=8+4\mu-2\mu^2-\mu^3=8+4\mu-2\mu^2-\mu(\mu+1)=8+3\mu-3\mu^2\\&=8+3(\mu-\mu^2)=8+3(-1)=5\end{aligned}\tag{8.7.231}$$

and therefore:

$$c^2=\frac{1}{4-\mu^2}a^2=\frac{2+\mu}{5}a^2=\frac{\mu^2+1}{5}a^2\tag{8.7.232}$$

hence:

$$c=\frac{\sqrt{\mu^2+1}}{\sqrt{5}}a\tag{8.7.233}$$

We have therefore for the calculation of the area of the dodecahedron, a surface composed of 12 pentagons, each of which is composed of by 5 triangle of base a and height h:

$$\begin{aligned}S&=12\cdot 5\frac{1}{2}ah=30ah=30a\frac{1}{2}\mu c=15a\mu c=15a\mu\frac{\sqrt{\mu^2+1}}{\sqrt{5}}a\\&=15\mu\frac{\sqrt{\mu^2+1}}{\sqrt{5}}a^2\end{aligned}\tag{8.7.234}$$

Therefore:

$$\boxed{S=3\mu a^2\sqrt{5}\sqrt{\mu^2+1}}\tag{8.7.235}$$

To calculate the volume, we're going to use the trick mentioned at the beginning. That is to say to cut in a first time the dodecahedron in a parallelepiped of side:

$$c=s\mu=s\frac{1+\sqrt{5}}{2}\tag{8.7.236}$$

since the side of the parallelepiped is a diagonal of the pentagon of side s and of 6 small roofs (which are clearly visible on the figure previously given of the dodecahedron).

Every little roof following two different views have the following lengths (where we fall back obviously for some edges those of pentagons s or the diagonals c thereof):

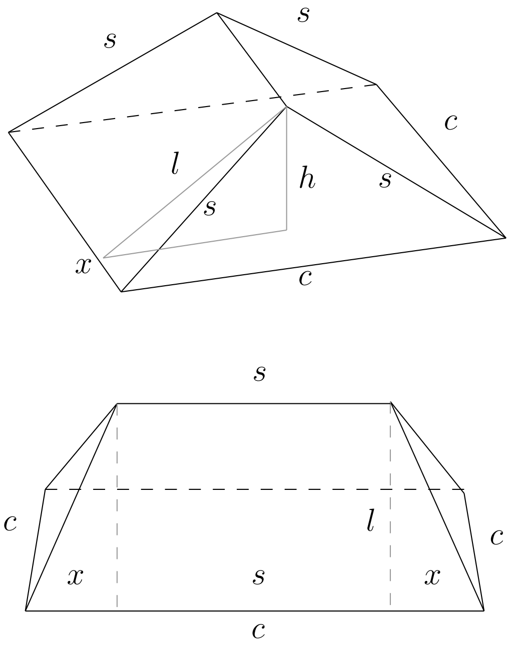


Figure 8.276: **Dodecahedron roofs**

Each small roof, we treat separately their ends by separating (cutting) them and merging them together. Finally, we have two pieces to study: the major visible part of the roof left in the figure below and the secondary part of the roof on the right in the figure below, which is nothing else other than the merging of the extremities of one small roof:

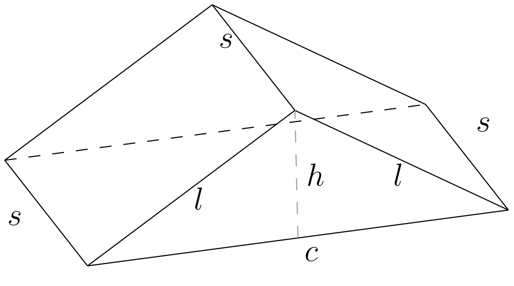


Figure 8.277: **Dodecahedron roofs decomposition for volume calculation**

We must therefore determine x and l and h since c and s are already known to us.

First, we see trivially that:

$$x=\frac{c-s}{2} \tag{8.7.237}$$

From the Pythagorean theorem, we then have:

$$\begin{aligned}&s^2=l^2+x^2=l^2+\left(\frac{c-s}{2}\right)^2\\&l^2=h^2+\left(\frac{c}{2}\right)^2\end{aligned} \tag{8.7.238}$$

Combining these two relationships, we get:

$$s^2-\left(\frac{c-s}{2}\right)^2=h^2+\left(\frac{c}{2}\right)^2 \tag{8.7.239}$$

Therefore it comes:

$$\begin{aligned}h^2&=s^2-\left(\frac{c-s}{2}\right)^2-\left(\frac{c}{2}\right)^2=s^2-\frac{1}{4}(c^2-2cs+s^2)-\frac{c^2}{4}=\frac{3}{4}s^2-\frac{c^2}{2}+\frac{cs}{2}\\&=\frac{3}{4}s^2-\frac{\left(\dfrac{1+\sqrt{5}}{2}s\right)^2}{2}+\frac{\left(\dfrac{1+\sqrt{5}}{2}s\right)s}{2}=\frac{3}{4}s^2-\frac{(1+\sqrt{5})^2s^2}{8}+\frac{(1+\sqrt{5})s^2}{4}\\&=s^2\left(\frac{3}{4}-\frac{(1+\sqrt{5})^2}{8}+\frac{(1+\sqrt{5})}{4}\right)=s^2\left(\frac{6-1-2\sqrt{5}-5+2+2\sqrt{5}}{8}\right)=s^2\frac{2}{8}=\frac{1}{4}s^2\end{aligned} \tag{8.7.240}$$

Hence:

$$h=\frac{s}{2} \tag{8.7.241}$$

We can now calculate the volume V_p of the 6 small roofs:

$$\begin{aligned}V_p&=6\left(\frac{1}{3}2xch+\frac{csh}{2}\right)=6\left(\frac{1}{3}2\frac{c-s}{2}c\frac{s}{2}+\frac{cs}{2}\frac{s}{2}\right)\\&=6\left(\frac{1}{3}\left(s\frac{1+\sqrt{5}}{2}-s\right)s\frac{1+\sqrt{5}}{2}\frac{s}{2}+s\frac{1+\sqrt{5}}{2}\frac{s^2}{4}\right)\\&=s^3\left(\left(\frac{1+\sqrt{5}}{2}-1\right)\frac{1+\sqrt{5}}{2}+3\frac{1+\sqrt{5}}{2}\right)=s^3\left(\frac{(1+\sqrt{5}-2)(1+\sqrt{5})}{4}+\frac{3+3\sqrt{5}}{3}\right)\\&=s^3\left(\frac{(1+\sqrt{5}-2)(1+\sqrt{5})+3+3\sqrt{5}}{4}\right)=s^3\left(\frac{(1+\sqrt{5}-2+\sqrt{5}+5-2\sqrt{5})+3+3\sqrt{5}}{4}\right)\\&=s^3\left(\frac{7+3\sqrt{5}}{4}\right)\end{aligned} \tag{8.7.242}$$

Therefore, the total volume of the dodecahedron is finally the volume of 6 small roofs summed with the volume of the central parallelepiped:

$$\begin{aligned}V&=c^3+s^3\left(\frac{7+3\sqrt{5}}{4}\right)=\left(\frac{1+\sqrt{5}}{2}\right)^3+s^3\left(\frac{7+3\sqrt{5}}{4}\right)=s^3\left[\left(\frac{7+3\sqrt{5}}{4}\right)+\left(\frac{1+\sqrt{5}}{2}\right)^3\right]\\&=s^3\left[\left(\frac{7+3\sqrt{5}}{4}\right)+\frac{1}{8}(1+\sqrt{5})^3\right]=s^3\left[\frac{7+3\sqrt{5}}{4}+\frac{1}{8}(1+3\sqrt{5}+3\cdot 5+5\sqrt{5})\right]\\&=s^3\left[\frac{7+3\sqrt{5}}{4}+\frac{1}{8}(16+8\sqrt{5})\right]=s^3\left[\frac{7+3\sqrt{5}}{4}+\frac{1}{4}(8+4\sqrt{5})\right]=s^3\frac{15+7\sqrt{5}}{4}\end{aligned} \tag{8.7.243}$$

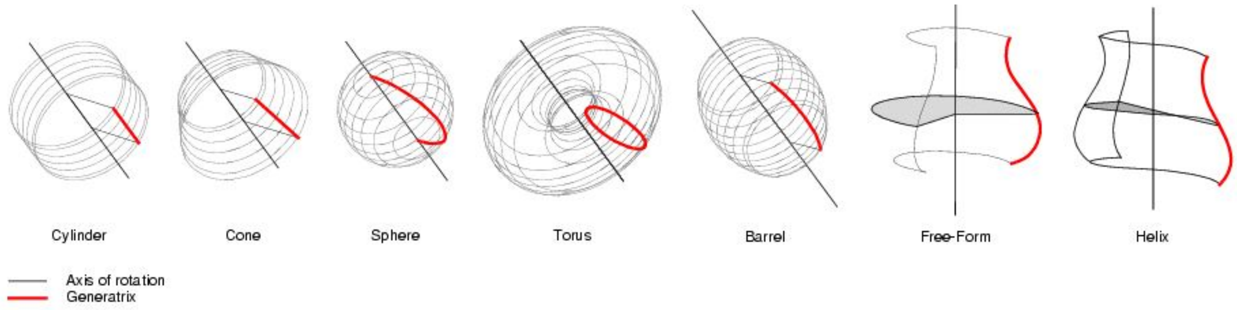
Finally:

$$\boxed{V=s^3\frac{15+7\sqrt{5}}{4}} \tag{8.7.244}$$

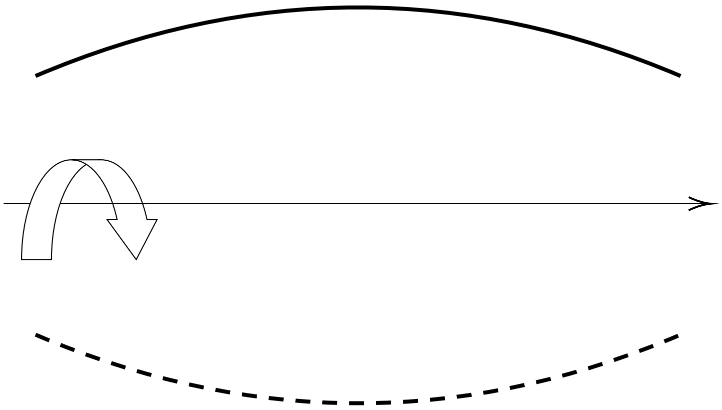
8.7.2.2 Usual Solids of Revolution

Let us recall that as seen in the section of Analytical Geometry a "solid of revolution" is a solid figure obtained by rotating a plane curve around some straight line (the axis) that lies on the same plane. and the more formally a surface of revolution is a surface obtained by rotating a plane curve (e.g. $y=f(x)$), named the "generatrix", around the y-axis (for example!). So we pass from a plane of $\mathbb{R}^2$ to a basis in $\mathbb{R}^3$.

Indeed many surfaces (and also some of which we've seen in the section of Analytical Geometry as the sphere, tore and cylinder) can be described by revolving a primary form of smaller size and then by rotation.


Cylinder
Cone
Sphere
Torus
Barrel
Free-Form
Helix
Axis of rotation
Generatrix

Let us see before going further a general method for determining the area of a body of revolution. That is to say the surface of the body generated by the rotation of a finite length curve about an axis:

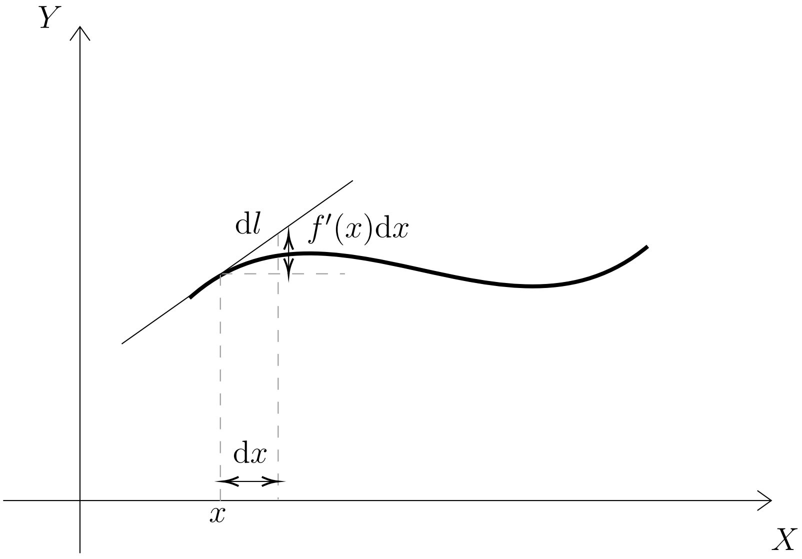


For this, we notice that when the curve is given a function $y=f(x)$ and we notice by Pythagoras (see figure below) that the element of length $\mathrm{d}l$ satisfies (relation that we have already meet in other sections of this textbook):

$$\mathrm{d}l^2=\mathrm{d}x^2+\mathrm{d}y^2=\mathrm{d}x^2+(f'(x))^2\mathrm{d}x^2 \tag{8.7.245}$$

Therefore:

$$\mathrm{d}l=\sqrt{1+(f'(x))^2}\mathrm{d}x^2 \tag{8.7.246}$$



Thus, the surface element generated by the rotation of the element of length $\mathrm{d}l$ is given by:

$$\mathrm{d}S = 2\pi|f(x)|\mathrm{d}l = 2\pi|f(x)|\sqrt{1+(f'(x))^2}\mathrm{d}x^2 \qquad (8.7.247)$$

The area of the surface of revolution generated by a function $f:[a,b]\mapsto\mathbb{R}$ continuously differentiable is then given by the relation:

$$\boxed{S = 2\pi\int\limits_a^b |f(x)|\sqrt{1+(f'(x))^2}\mathrm{d}x^2} \qquad (8.7.248)$$

Remark

In the section Mechanics we will use another approach that can be used to calculate the surface (and volume) of bodies of revolution and that is named "Pappus's second centroid Theorem". In fact as we will see, it is more used in practice to determine the centroid position rather than the surface (or volume) of a body of revolution.

8.7.2.2.1 Cylinder

Definition (#8.7.391): A "cylinder" is a surface generated by a line moving parallel to a fixed direction by meeting a fixed plane curve (circle), which plane cuts the given direction.

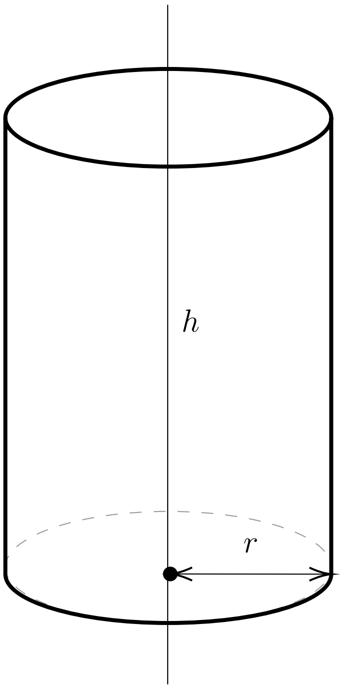


Figure 8.278: **Example of a cylinder**

The volume of a cylinder of revolution of height $x = r$ and height h is calculated by the method of discs knowing that the surface of a circle (disc) is πr^2:

$$V = \pi r^2 \int_0^h \mathrm{d}y = \pi r^2 y \Big|_0^h = \pi r^2 h \tag{8.7.249}$$

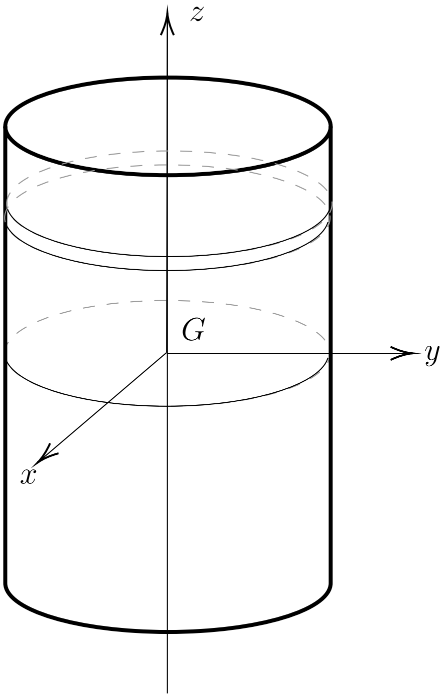
Therefore:

$$\boxed{V = \pi r^2 h} \tag{8.7.250}$$

The surface of a cylinder is simply the sum of the surface of the both discs and of the surface of the folded rectangle of height h and length of $2\pi r$:

$$\boxed{S = 2(\pi r^2) + 2\pi r h} \tag{8.7.251}$$

Now let us calculate the moment of inertia of a solid cylinder with respect to its vertical symmetry axis (revolution axis):



We have:

$$\mathrm{d}m=\rho\mathrm{d}V=\rho 2\pi rh\mathrm{d}r \tag{8.7.252}$$

Therefore:

$$J_z=\int r^2\mathrm{d}m=\int \rho r^2 2\pi rh\mathrm{d}r=2\pi\rho h\int\limits_0^R\int\limits_0^R r^3\mathrm{d}r=2\pi\rho h\frac{R^4}{4}=\rho\pi R^2 h\frac{R^2}{2}=M\frac{R^2}{2} \tag{8.7.253}$$

Given now G the cylinder's center of gravity, G_z coincides with the cylinder axis of the revolution. The axes G_y and G_x play identical roles. The moments of inertia J_{G_x} and J_{G_y} with respect to these axes are equal and are written:

$$J_{G_y}=\int(x^2+z^2)\mathrm{d}m=J_{G_x}=\int(y^2+z^2)\mathrm{d}m=J \tag{8.7.254}$$

hence:

$$J_{G_y}+J_{G_x}=2J_{G_x}=2J_{G_y}=\int(x^2+z^2+y^2+z^2)\mathrm{d}m=\int(x^2+y^2+2z^2)\mathrm{d}m \tag{8.7.255}$$

hence:

$$J_{G_y}=J_{G_x}=\frac{1}{2}(x^2+y^2+2z^2)\mathrm{d}m=\frac{1}{2}(x^2+y^2)\mathrm{d}m+\int z^2\mathrm{d}m \tag{8.7.256}$$

The first integral is in fact the moment of inertia of the cylinder relatively to the axis G_z and we know it's equal to:

$$J_{G_z}=J_z=M\frac{R^2}{2} \tag{8.7.257}$$

The second integral is easily calculated by cutting the cylinder into slices of thickness $\mathrm{d}z$ perpendicular to the axis G_z. The mass of the elementary slice is $\mathrm{d}m=\pi r^2\rho\mathrm{d}z$ thus:

$$\int z^2\mathrm{d}m=2\int_0^{h/2}\pi r^2\rho z^2\mathrm{d}z=2\pi r^2\rho\frac{h^3}{24}=M\frac{h^2}{12} \qquad (8.7.258)$$

The moment of inertia of a cylinder relative to an axis perpendicular to its axis of revolution is thus written:

$$J_{G_y}=J_y=J_{G_x}=J_x=M\left(\frac{R^2}{4}+\frac{h^2}{12}\right) \qquad (8.7.259)$$

In many high-school problems $R\ll h$ then we can find the latter relation under the form (named a "slender rod"):

$$J_{G_y}=J_{G_x}\cong M\frac{h^2}{12} \qquad (8.7.260)$$

Now let us found the moment of inertia of a cylinder rotation around an axis perpendicular to its rotation axis and at a distance h of one of it ends. To calculate it, we consider:

- The cylinder is cut into a infinitely many pieces of infinitesimally thin slices
- Each of these slices have a mass of $\mathrm{d}m$ and length of $\mathrm{d}x$
- A variable to sum. E.g. in this problem, we are summing from left of the axis to right of axis. The variable is x.

Now, we show our formula for the calculation for moment of inertia first:

$$\mathrm{d}J=x^2\mathrm{d}m \qquad (8.7.261)$$

Recall that we are using x to sum. Hence, we have to force a $\mathrm{d}x$ into the equation for moment of inertia. Now, lets find an expression for $\mathrm{d}m$. Since the rod is uniform, the mass varies linearly with distance such that:

$$\mathrm{d}m=\frac{M}{L}\mathrm{d}x \qquad (8.7.262)$$

Using the relation for $\mathrm{d}m$, we substitute it into the first relation. Hence, we have:

$$\mathrm{d}J=x^2\mathrm{d}m=\frac{M}{L}x^2\mathrm{d}x \qquad (8.7.263)$$

Now we know obviously that:

$$J=\int\mathrm{d}J \qquad (8.7.264)$$

Substituting $\mathrm{d}J$ (with the appropriate limits):

$$J_h=\frac{M}{L}\int\limits_{-h}^{L-h}x^2\mathrm{d}x \qquad (8.7.265)$$

where the lower limit is $-h$ because the left side of the cylinder is $-h$ units away from the axis of rotation (we take right as positive). This part is in fact the tricky part...

Solving the integration, we have the moment of inertia for a uniform rigid cylinder for any perpendicular rotation axis:

$$J_h=\frac{1}{3}M(L^2-3Lh+3h^2) \tag{8.7.266}$$

Now in the special case where $h=0$ (used in many high-school to study a falling chimney) or very small in comparison to L we found the famous relation that can been found in many textbooks (again under the name "slender rod"):

$$J_0=\frac{1}{3}mL^2 \tag{8.7.267}$$

If the rotation axis is through the center of mass of the rod (remember rod is uniform, hence $h=L/2$) we fall back on:

$$J_{h/2}=\frac{1}{12}mL^2 \tag{8.7.268}$$

Now let's also calculate the moment of inertia of an empty tube or cylinder of non-zero thickness (always provided in formula summary textbooks).

The moment of inertia of a tube with respect to its axis of revolution is a classical notion of the moment of inertia of a cylinder. Let us therefore consider a tube of outer radius r_e and inner radius r_i. As (see section Classical Mechanics page 3306):

$$\sum_{i=1}^{n} J_i=J \tag{8.7.269}$$

Therefore, the moment of inertia of a tube can be considered as the moment of inertia of a cylinder of radius equal to the outer radius of the tube, minus the moment of inertia of a cylinder of radius equal to the inner radius of the tube. Therefore:

$$\begin{aligned}J_{\text{tube}}&=J_{r_e}-J_{r_i}=\frac{M_e r_e^2}{2}-\frac{M_i r_i^2}{2}=\frac{1}{2}(\rho\pi r_e^2 h r_e^2-\rho\pi r_i^2 h r_i^2)=\frac{\rho\pi h}{2}(r_e^4-r_i^4)\\&=\frac{\rho\pi h}{2}(r_e^2-r_i^2)(r_e^2+r_i^2)=\frac{\rho\pi h(r_e^2-r_i^2)}{2}(r_e^2+r_i^2)=\frac{M}{2}(r_e^2+r_i^2)\end{aligned} \tag{8.7.270}$$

and if $\bar{r}=r_e^2\cong r_i^2$, then we get the classical relation available in many physics textbooks:

$$J_{\text{tube}}=\frac{M2\bar{r}^2}{2}=M\bar{r}^2 \tag{8.7.271}$$

8.7.2.2.2 Cone

Definition (#8.7.392): A "cone" is a three-dimensional geometric shape that tapers smoothly from a flat base (frequently, though not necessarily, circular) to a point named the "apex" or "vertex".

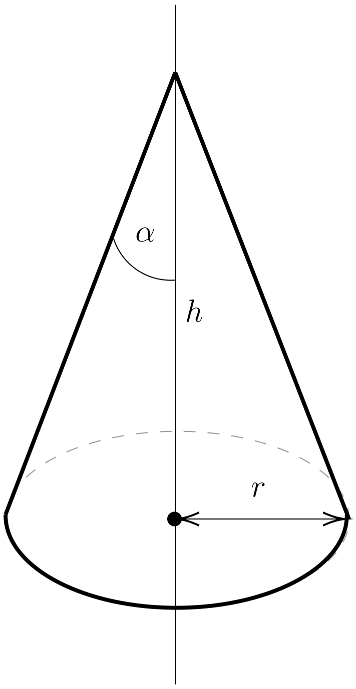


Figure 8.279: **Cone**

The line passing through the points $(r,0)$ (end of the base of the cone/center of the disc) and $(0,h)$ (apex of the cone) is given obviously by:

$$y=ax+b=\frac{-h}{r}x+1h \tag{8.7.272}$$

Indeed, when $x=0$, we have $y=h$ and when $x=r$ we have $y=0$.

The rotation of this line relative to the y-axis gives the volume of the cone:

$$V=\pi\int_a^b x^2\mathrm{d}y=\pi\int_0^h r^2\left(1-\frac{y}{h}\right)^2\mathrm{d}y=\pi r^2\frac{1}{3}\left[\left(1-\frac{y}{h}\right)\right]_0^h(-h)=\frac{-\pi r^2 h}{3}(0-1)=\frac{\pi r^2 h}{3} \tag{8.7.273}$$

Finally:

$$\boxed{V=\frac{\pi r^2 h}{3}} \tag{8.7.274}$$

To calculate the lateral surface of a cone, we will parametrized the straight line that go from the summit of the cone (0.0) to (r,h) so this is a different parametrization than for the volume (this gives the possibility to simplify some calculations). Then we have:

$$f(x)=\frac{r}{h}x \tag{8.7.275}$$

and therefore:

$$\begin{aligned}S_l&=2\pi\int\limits_0^h|f(x)|\sqrt{1+(f'(x))^2}\mathrm{d}x=2\pi\int\limits_0^h\left|\frac{r}{h}x\right|\sqrt{1+\frac{r^2}{h^2}}\mathrm{d}x=2\pi\frac{r}{h}\sqrt{1+\frac{r^2}{h^2}}\int\limits_0^h x\mathrm{d}x\\&=2\pi\frac{r}{h}\sqrt{1+\frac{r^2}{h^2}}\frac{1}{2}x^2\bigg|_0^h=\pi\frac{r}{h}\sqrt{1+\frac{r^2}{h^2}}h^2=\pi rh\sqrt{1+\frac{r^2}{h^2}}=\pi r\sqrt{h^2+r^2}\\&=\pi r^2\sqrt{1+\frac{h^2}{r^2}}\end{aligned}\tag{8.7.276}$$

Therefore, the total surface of the cone (base + side surface) is:

$$S=S_b+S_l=\pi r^2+\pi r^2\sqrt{1+\frac{h^2}{r^2}}=\pi r^2\left(1+\sqrt{1+\frac{h^2}{r^2}}\right)\tag{8.7.277}$$

Finally:

$$\boxed{S=\pi r^2\left(1+\sqrt{1+\frac{h^2}{r^2}}\right)}\tag{8.7.278}$$

Let us now calculate the moment of inertia of a cone with respect to its axis of revolution:

For these calculations we will use the moment of inertia of the cylinder $J_{z,\text{cyl}}$ and consider the cone as a stack of infinitesimal cylinders.

Therefore:

$$\begin{aligned}J_z&=\int\mathrm{d}J_{z,\text{cyl}}=\int\frac{r^2}{2}\mathrm{d}m=\int\frac{r^2}{2}(\rho\pi r^2\mathrm{d}h)=\frac{\pi\rho}{2}\int r^2\mathrm{d}h=\frac{\pi\rho}{2}\frac{H}{R}\int\limits_0^R r^4\mathrm{d}r=\frac{\pi\rho HR^5}{10R}\\&=\left(\rho\frac{\pi R^2H}{3}\right)\frac{3R^2}{10}=\frac{3}{10}MR^2\end{aligned}\tag{8.7.279}$$

8.7.2.2.3 Sphere

Definition (#8.7.393): A "sphere" is a perfectly round geometrical object in three-dimensional space that is the surface of a completely round ball. Like a circle, which geometrically is a two-dimensional object, a sphere is defined mathematically as the set of points that are all at the same distance r (radius) from a given point, but in three-dimensional space. The longest straight line through the ball, connecting two points of the sphere, passes through the center and its length is thus twice the radius (as for the circle); it is the diameter $\varnothing$ of the sphere.

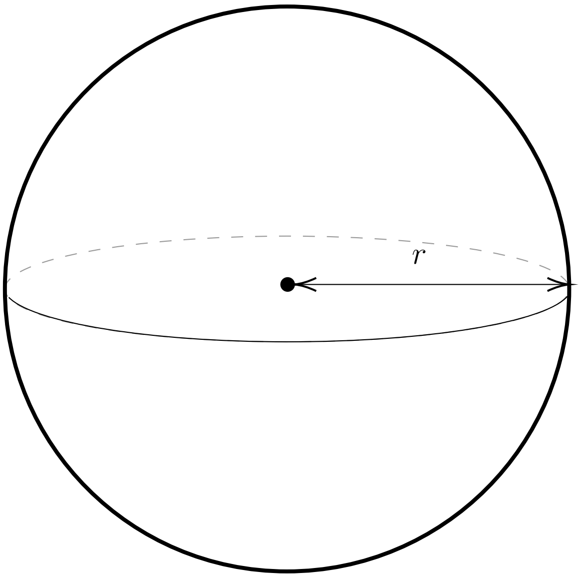


Figure 8.280: **Example of a Sphere**

As the figure above is quite academic, here is a more friendly one if it can help:

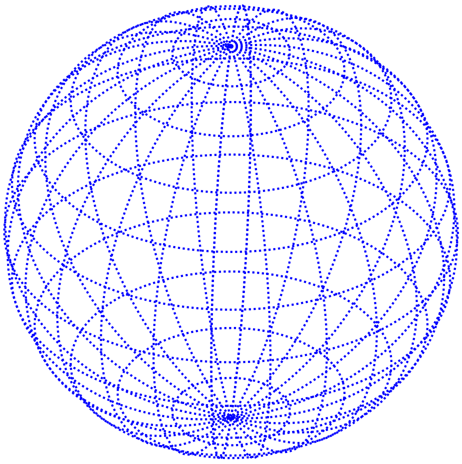


Figure 8.281: **Sphere grid (author: Jeff Hein)**

Obviously we can see a sphere of radius R, as a surface formed by rotating a semicircle around its major axis. The function describing a semicircle being (see section Analytical Geometry page 2768):

$$y=f(x)=\sqrt{R^2+x^2} \qquad (8.7.280)$$

The sphere can be dissected as a sum of disk of thickness Δx. The half-discs being perpendicular to the x-axis and of width Δx at the position x_k (see figure below).

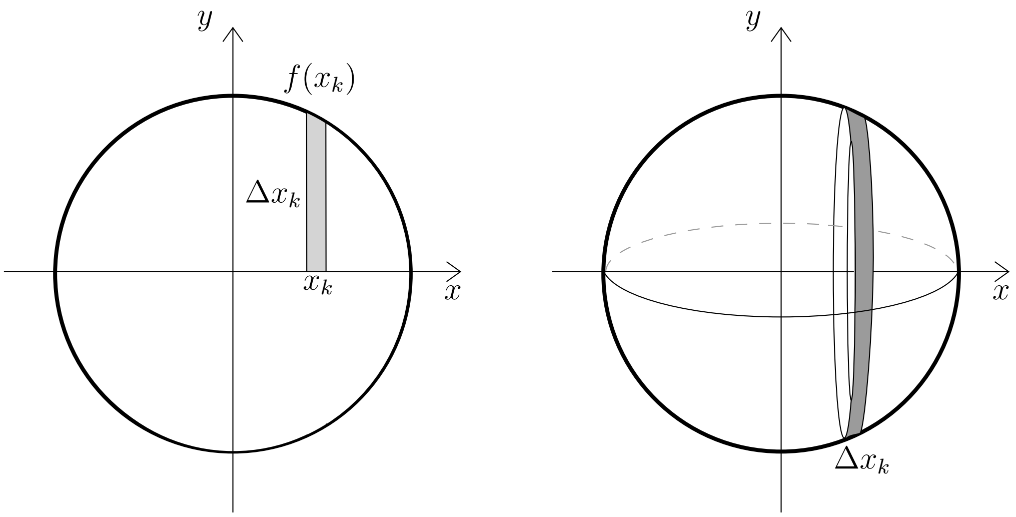


Figure 8.282: **Calculation of the volume of a sphere by a stack of discs**

Then we have:

$$x_k = k\Delta x \tag{8.7.281}$$

The volume of a disk (cylinder) being given by (passing to the limit):

$$\mathrm{d}V_k = \pi r_k^2 \mathrm{d}x \tag{8.7.282}$$

and the radius r_k being given by the function:

$$r_k = f(x_k) = \sqrt{R^2 - x_k^2} \tag{8.7.283}$$

then we have:

$$\mathrm{d}V_k = \pi \left(\sqrt{R^2 - x_k^2}\right)^2 \mathrm{d}x = \pi(R^2 - x_k^2)\mathrm{d}x \tag{8.7.284}$$

By integrating between $x = [-R, R]$, we then have:

$$V = \int\limits_{-R}^{R} \pi(R^2 - r_k^2)\mathrm{d}x = \pi \int\limits_{-R}^{R} \mathrm{d}x = \pi \left(R^2 x - \frac{x^3}{3}\right)\Bigg|_{-R}^{R} \tag{8.7.285}$$

We can also make the terminals as $x = [0, R]$ it gives the same to a factor 2:

$$V = 2\pi \left(R^2 x - \frac{x^3}{3}\right)\Bigg|_{0}^{R} = 2\pi \left(R^3 - \frac{R^3}{3}\right) = 2\pi \frac{3R^3 - R^3}{3} = 2\pi \frac{2R^3}{3} = \frac{4}{3}\pi R^3 \tag{8.7.286}$$

Finally:

$$\boxed{V = \frac{4}{3}\pi R^3} \tag{8.7.287}$$

The expression of the surface being given by differentiation with respect to the element generating the surface, we get (it's a bit limit to say this...):

$$\boxed{S = 4\pi R^2} \tag{8.7.288}$$

There is another way, more rigorous, to approach these both calculations (especially the last one...). Indeed, in the section of Differential and Integral Calculus we have introduced the concept of Jacobian that allows changing the coordinate system based on integration variables we are working on (for a detailed definition the reader should refer to section Differential and Integral Calculus page 1524):

$$\iiint\limits_D f(x,y)\mathrm{d}x\mathrm{d}y=\iiint\limits_D F(u,v,w)|\det(J)|\mathrm{d}u\mathrm{d}v\mathrm{d}w \tag{8.7.289}$$

and we have proved (always in the section of Differential and Integral Calculus page 1528) that the Jacobian in spherical coordinates was:

$$|\det(J)|=r^2\sin(\theta) \tag{8.7.290}$$

So just as $\mathrm{d}x\mathrm{d}y\mathrm{d}z$ is a differential element of volume, we can convert this element into spherical coordinates and make verbatim appear a differential element of volume of the sphere of radius r. We then have just to integrate correctly for the size of the whole sphere.

Therefore we have:

$$\begin{aligned}V&=\iiint\limits_D r^2\sin(\theta)\mathrm{d}r\mathrm{d}\phi\mathrm{d}\theta=\int\limits_0^R r^2\mathrm{d}r\int\limits_0^\pi\sin(\theta)\mathrm{d}\theta\int\limits_0^{2\pi}\mathrm{d}\pi=\left.\frac{r^3}{3}\right|_0^R\cdot\left.(-\cos(\theta))\right|_0^\pi\left.\phi\right|_0^{2\pi}\\&=\frac{R^3}{3}(-\cos(\pi)-(-\cos(0)))2\pi=\frac{2}{3}\pi R^3(-(-1)-(-1))=\frac{4}{3}\pi R^3\end{aligned} \tag{8.7.291}$$

and for the surface (for which the radius is constant):

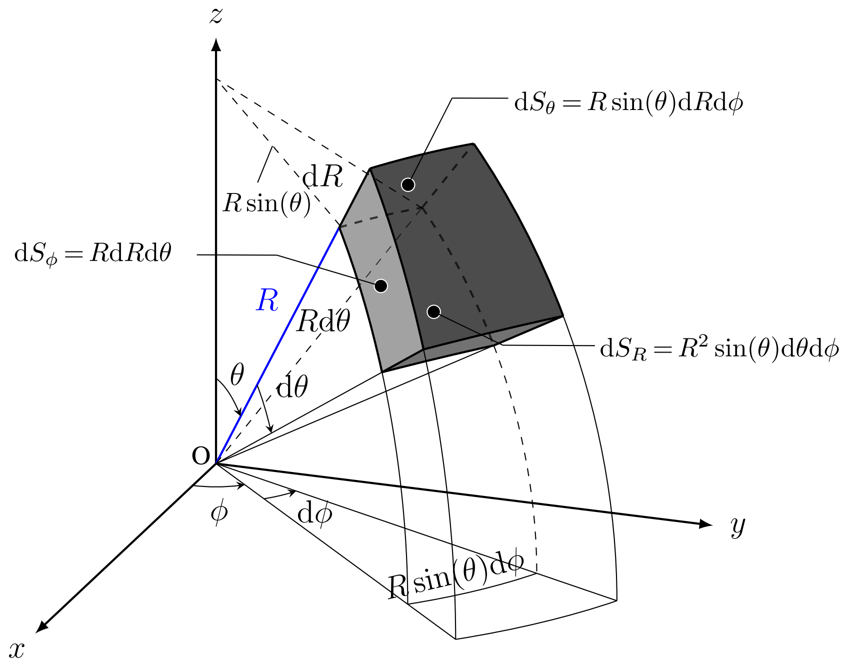
$$\begin{aligned}S&=R^2\iint\sin(\theta)\mathrm{d}\phi\mathrm{d}\theta=R^2\int_0^{2\pi}\sin(\theta)\mathrm{d}\theta\int_0^{2\pi}\mathrm{d}\phi=R^2\left.(-\cos(\theta))\right|_0^\pi\left.\phi\right|_0^{2\pi}\\&=R^2(-\cos(\pi)-(-\cos(0)))2\pi=2\pi R^2(-(-1)-(-1))=4\pi R^2\end{aligned} \tag{8.7.292}$$

If necessary, we can find the element of surface geometrically rather than through the Jacobian because the latter is not very educational in small classes...

Then, remembering that in the section of Trigonometry (see page 2542), we proved that the length of an element circular arc is given by:

$$\mathrm{d}L=R\mathrm{d}\alpha \tag{8.7.293}$$

So, it becomes quite easy to complete the following figure:



z
$\mathrm{d}S_{\theta} = R\sin(\theta)\mathrm{d}R\mathrm{d}\phi$
$\mathrm{d}R$
$R\sin(\theta)$
$\mathrm{d}S_{\phi} = R\mathrm{d}R\mathrm{d}\theta$
R
$R\mathrm{d}\theta$
$\mathrm{d}S_R = R^2\sin(\theta)\mathrm{d}\theta\mathrm{d}\phi$
θ
$\mathrm{d}\theta$
O
ϕ
$\mathrm{d}\phi$
y
$R\sin(\theta)\mathrm{d}\phi$
x

Figure 8.283: **Representation of a surface element in spherical coordinates**

and then we see immediately that (most important surface element for us in this textbook so far for the sections treating of Physics and Engineering!):

$$\mathrm{d}S = R^2\sin(\theta)\mathrm{d}\theta\mathrm{d}\phi \tag{8.7.294}$$

what is obviously more funny...

Let us now calculate the moment of inertia of a homogeneous solid ball of mass M and density ρ. For this, the ball having a maximal symmetry, it is more convenient to first calculate the polar moment of inertia (see section Classical Mechanics page 3306), and then determine the axial moment of inertia using the first result:

$$J_0 = \int r^2\mathrm{d}m = \int r^2\rho\mathrm{d}V = \int r^2\rho(4\pi r^2\mathrm{d}r) = 4\pi\rho\int_0^R r^4\mathrm{d}r = 4\pi\rho\frac{1}{5}R^5 = \frac{4}{3}\pi\rho\frac{3}{5}R^3R^2 = \frac{3}{5}MR^2 \tag{8.7.295}$$

where we used the fact that:

$$\rho = \frac{M}{\dfrac{4}{3}\pi R^3} \tag{8.7.296}$$

As J_x, J_y, J_z are equal by the symmetry of the ball, it comes:

$$J_0 = \frac{1}{2}(3J_x) \Rightarrow J_x = J_y = J_z = \frac{2}{3}J_0 = \frac{2}{5}MR^2 \tag{8.7.297}$$

We give also another derivation if the reader don't like the previous one. For that purpose remember that:

$$J = \sum_i m_i r^2 = \int_{\text{body}} r^2\mathrm{d}m \tag{8.7.298}$$

For a body of uniform composition $\mathrm{d}m = \rho \mathrm{d}V$, where ρ is the density and $\mathrm{d}V$ is the change in volume. For a sphere we have just seen that $\mathrm{d}V = 4\pi/3r^2\mathrm{d}r$. Substitution gives:

$$J = \int_{\text{body}} r^2 \mathrm{d}m = \int_{\text{body}} r^2 \rho \mathrm{d}V = \int_{\text{body}} r^2 \rho \frac{4}{3}\pi r^2 \mathrm{d}r \qquad (8.7.299)$$

So:

$$J = \frac{4\pi\rho}{3}\int\limits_{-R}^{+R} r^4 \mathrm{d}r = \frac{4\pi\rho}{3}\frac{r^5}{5}\bigg|_{-R}^{+R} = \frac{4\pi\rho}{3}\frac{2R^5}{5} \qquad (8.7.300)$$

The mass of the object is given by $M = 4\pi\rho/3R^3$. So:

$$J = \frac{2}{5}MR^2 \qquad (8.7.301)$$

8.7.2.2.3.1 Spherical cap surface

In the field of astronomy we need sometimes to compute the surface of a spherical cap of high h! Let us see how to do this based on the following figure:

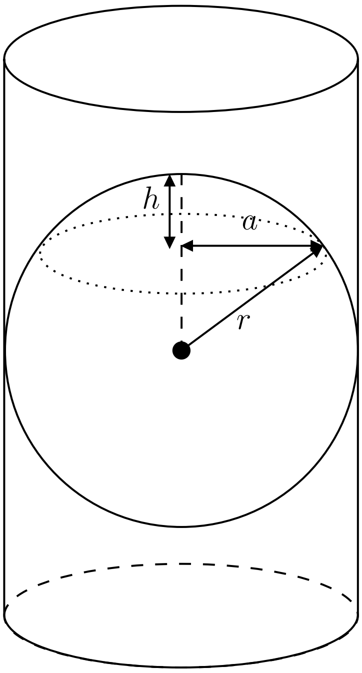


Figure 8.284: **Spherical cap surface derivation trick**

The idea to determiner the spherical cap area we enclose the sphere inside a cylinder of radius r and height $2r$ just touching at a great circle. The projection of the sphere onto the cylinder preserves area.

That is the way it is assumed Archimedes derived that the area of the sphere is same as lateral surface area of the cylinder which is for reminder for our special case:

$$S_{\text{cylinder}} = (2\pi r)(2r) = 4\pi r^2 \qquad (8.7.302)$$

The projection of the cap on the cylinder has area:

$$S_{\text{sph. cap}} = (2\pi r)\cdot h \tag{8.7.303}$$

And since:

$$a^2 = r^2 - (r-h)^2 = 2rh - h^2 \Rightarrow 2rh = h^2 + a^2 \tag{8.7.304}$$

the area of the cap can therefore be written:

$$S_{\text{sph. cap}} = (2\pi r)\cdot h = \pi(2rh) = \pi\left(h^2 + a^2\right) \tag{8.7.305}$$

In astronomy we have most of the time as starting information not h (nor a) but the percentage length $\%\ell$ of a half great circle (that has obviously for length ${}^1\!/\!{}_2\mathcal{C} = 2\pi r/2 = \pi r$) whose corresponding projection in the sphere axis is $h < r$ (in the case the projection is greater than r we will sum et subtract multiple sphere caps).

The trick to get the area of spherical cap in percentage $\%\ell$ of a half great circle is to think in its equivalent in angle π (half of the corresponding angle of the entire great circle). Therefore:

$$h = r - r\cos\left(\frac{\%\ell\pi}{2}\right) = r\left[1-\cos\left(\frac{\%\ell\pi}{2}\right)\right] \tag{8.7.306}$$

and identically:

$$a = r\sin\left(\frac{\%\ell\pi}{2}\right) \tag{8.7.307}$$

Hence the corresponding spherical cap area as used in astronomy:

$$\begin{aligned} S_{\text{sph. cap}} &= \pi\left(h^2+a^2\right) = \pi\left\{r^2\left[1-\cos\left(\frac{\%\ell\pi}{2}\right)\right]^2 + r^2\sin^2\left(\frac{\%\ell\pi}{2}\right)\right\} \\ &= 2r^2 - 2r^2\cos^2\left(\frac{\%\ell\pi}{2}\right) = 2r^2\sin^2\left(\frac{\%\ell\pi}{2}\right) \end{aligned} \tag{8.7.308}$$

8.7.2.2.3.2 Sphere packing

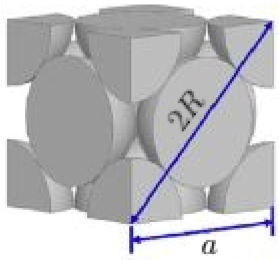
In geometry, a sphere packing is an arrangement of non-overlapping spheres within a containing space. The spheres considered are usually all of identical size, and the space is usually three-dimensional Euclidean space. A typical sphere packing problem is to find an arrangement in which the spheres fill as much of the space as possible. The proportion of space filled by the spheres is named the "density of the arrangement".

Many problems in the chemical and physical sciences can be related to sphere packing problems where more than one size of sphere is available. Here there is a choice between separating the spheres into regions of close-packed equal spheres, or combining the multiple sizes of spheres into a compound or interstitial packing (when many sizes of spheres - or a distribution - are available, the problem quickly becomes intractable).

For the very special case of equal spheres in three dimensions, the densest packing uses approximately 74% of the volume.

Proof 8.27.3. Let us see for the two famous cases:

- FCC (face cubic center) structure:



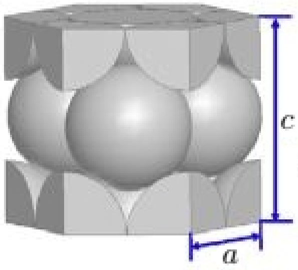
First we see obviously as the diagonal of the cube is equal to four times a sphere radius ($4R$) that using the Pythagoras theorem:

$$(4R)^2=a^2+a^2\Rightarrow 4R=\sqrt{2}a\Rightarrow R=\frac{\sqrt{2}}{4}a \tag{8.7.309}$$

Therefore, denoting V_s the volume of a sphere and V_e the volume of the cube:

$$\text{Packing factor}=\frac{V_s}{V_e}=\frac{4\cdot\frac{4\pi}{3}R^3}{a^3}=\frac{4\cdot\frac{4\pi}{3}\cdot\left(\frac{\sqrt{2}}{4}a\right)^3}{a^3}=\frac{\sqrt{2}\pi}{6}=74.05\% \tag{8.7.310}$$

- HCP (hexagonal close packed):



First we see obviously that:

$$c=\sqrt{\frac{8}{3}}a \qquad \text{and} \qquad R=\frac{a}{2} \tag{8.7.311}$$

Therefore:

$$\text{Packing factor}=\frac{V_s}{V_3}=\frac{6\cdot\frac{4\pi}{3}R^3}{6\cdot\frac{1}{2}\cdot\frac{\sqrt{3}}{2}a^2\cdot c}=\frac{6\cdot\frac{4\pi}{3}\cdot\left(\frac{a}{2}\right)^2}{6\cdot\frac{1}{2}\cdot\frac{\sqrt{3}}{2}a^2\cdot\frac{\sqrt{8}}{3}a}=\frac{\sqrt{2}\pi}{6}=74.05\% \tag{8.7.312}$$

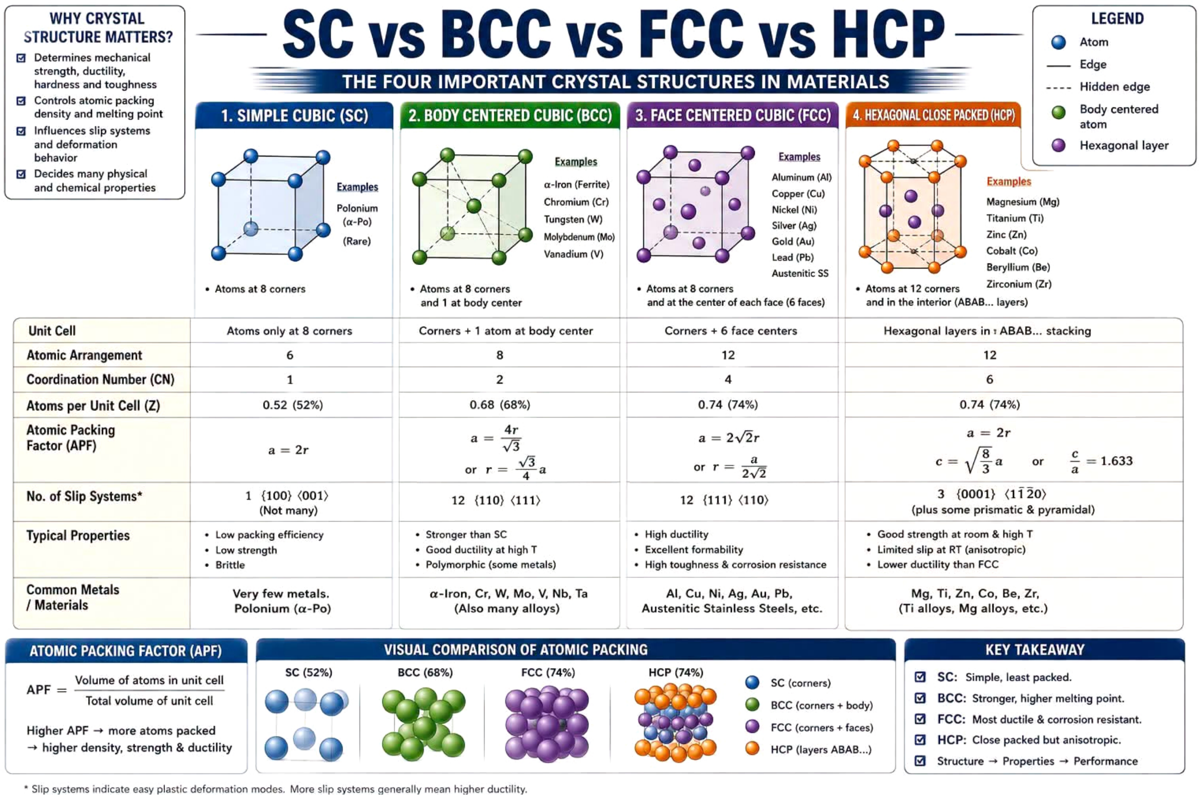
■ Q.E.D.

Carl Friedrich Gauss proved in 11831 (holocene calendar) in a non-trivial and non-formal way that these packings have the highest density amongst all possible lattice packings (we won't provide the proof here as it is quite non-elegant in our personal and subjective point of view).

In 11611 (holocene calendar) Johannes Kepler conjectured that this is the maximum possible density amongst both regular and irregular arrangements—this became known as the "Kepler conjecture". In 11998 (holocene calendar), Thomas Callister Hales, following the approach suggested by László Fejes Tóth in 11953 (holocene calendar), announced a proof of the Kepler conjecture (for a good presentation of the proof see [222]). Hales' proof is a proof by exhaustion involving checking of many individual cases using complex computer calculations. Referees said that they were "99% certain" of the correctness of Hales' proof. On 10 August 12014 (holocene calendar) Hales announced the completion of a formal proof using automated proof checking, removing any doubt.

Remark

Some other lattice packings are often found in physical systems. These include the cubic lattice with a density of $\frac{\pi}{6}\cong 0.5236$, the hexagonal lattice with a density of $\frac{\pi}{3\sqrt{3}}\cong 0.6046$ and the tetrahedral lattice with a density of $\frac{\pi\sqrt{3}}{16}\cong 0.3401$, and loosest possible at a density of 0.0555.



WHY CRYSTAL STRUCTURE MATTERS?

- ☑ Determines mechanical strength, ductility, hardness and toughness
- ☑ Controls atomic packing density and melting point
- ☑ Influences slip systems and deformation behavior
- ☑ Decides many physical and chemical properties

SC vs BCC vs FCC vs HCP

THE FOUR IMPORTANT CRYSTAL STRUCTURES IN MATERIALS

LEGEND

- ● Atom
- — Edge
- ---- Hidden edge
- ● Body centered atom
- ● Hexagonal layer

1. SIMPLE CUBIC (SC)

Examples: Polonium (α-Po) (Rare)

- Atoms at 8 corners

2. BODY CENTERED CUBIC (BCC)

Examples: α-Iron (Ferrite), Chromium (Cr), Tungsten (W), Molybdenum (Mo), Vanadium (V)

- Atoms at 8 corners and 1 at body center

3. FACE CENTERED CUBIC (FCC)

Examples: Aluminum (Al), Copper (Cu), Nickel (Ni), Silver (Ag), Gold (Au), Lead (Pb), Austenitic SS

- Atoms at 8 corners and at the center of each face (6 faces)

4. HEXAGONAL CLOSE PACKED (HCP)

Examples: Magnesium (Mg), Titanium (Ti), Zinc (Zn), Cobalt (Co), Beryllium (Be), Zirconium (Zr)

- Atoms at 12 corners and in the interior (ABAB... layers)

	1. SIMPLE CUBIC (SC)	2. BODY CENTERED CUBIC (BCC)	3. FACE CENTERED CUBIC (FCC)	4. HEXAGONAL CLOSE PACKED (HCP)
Unit Cell	Atoms only at 8 corners	Corners + 1 atom at body center	Corners + 6 face centers	Hexagonal layers in ı ABAB... stacking
Atomic Arrangement	6	8	12	12
Coordination Number (CN)	1	2	4	6
Atoms per Unit Cell (Z)	0.52 (52%)	0.68 (68%)	0.74 (74%)	0.74 (74%)
Atomic Packing Factor (APF)	$a = 2r$	$a = \frac{4r}{\sqrt{3}}$ or $r = \frac{\sqrt{3}}{4}a$	$a = 2\sqrt{2}r$ or $r = \frac{a}{2\sqrt{2}}$	$a = 2r$ $c = \sqrt{\frac{8}{3}}a$ or $\frac{c}{a} = 1.633$
No. of Slip Systems*	1 {100} ⟨001⟩ (Not many)	12 {110} ⟨111⟩	12 {111} ⟨110⟩	3 {0001} $\langle 1\bar{1}\bar{2}0\rangle$ (plus some prismatic & pyramidal)
Typical Properties	• Low packing efficiency • Low strength • Brittle	• Stronger than SC • Good ductility at high T • Polymorphic (some metals)	• High ductility • Excellent formability • High toughness & corrosion resistance	• Good strength at room & high T • Limited slip at RT (anisotropic) • Lower ductility than FCC
Common Metals / Materials	Very few metals. Polonium (α-Po)	α-Iron, Cr, W, Mo, V, Nb, Ta (Also many alloys)	Al, Cu, Ni, Ag, Au, Pb, Austenitic Stainless Steels, etc.	Mg, Ti, Zn, Co, Be, Zr, (Ti alloys, Mg alloys, etc.)

ATOMIC PACKING FACTOR (APF)

$$\text{APF} = \frac{\text{Volume of atoms in unit cell}}{\text{Total volume of unit cell}}$$

Higher APF → more atoms packed → higher density, strength & ductility

VISUAL COMPARISON OF ATOMIC PACKING

- ● SC (corners)
- ● BCC (corners + body)
- ● FCC (corners + faces)
- ● HCP (layers ABAB...)

KEY TAKEAWAY

- ☑ **SC:** Simple, least packed.
- ☑ **BCC:** Stronger, higher melting point.
- ☑ **FCC:** Most ductile & corrosion resistant.
- ☑ **HCP:** Close packed but anisotropic.
- ☑ Structure → Properties → Performance

* Slip systems indicate easy plastic deformation modes. More slip systems generally mean higher ductility.

Finally notice that many people like to compute how many times the Earth can be contained inside Jupiter or inside the Sun. Dividing the volume for example of the Sun (actually a G2V-type star, aka yellow dwarf) by that of Earth obviously gives a value ($\sim 1,300,000$) that assumes that the Earth can be cut into infinitesimal volumes. But if we consider that we can't cut the Earth into infinitesimal volumes (sphere packing problem), then the number is approximately $\sim 932,884$. Value to compare with $1,300,000\cdot 74.05\% = 962,624$.

8.7.2.2.4 Torus

Definition (#8.7.394): A "torus" is the surface generated by rotating a circle c of "minor radius" r around a straight line at a distance R, named "major radius" from its center located in its plane, but not passing through its center. In other words, it is a surface of revolution generated by revolving a circle in three-dimensional space about an axis coplanar with the circle and that does not touch the circle:

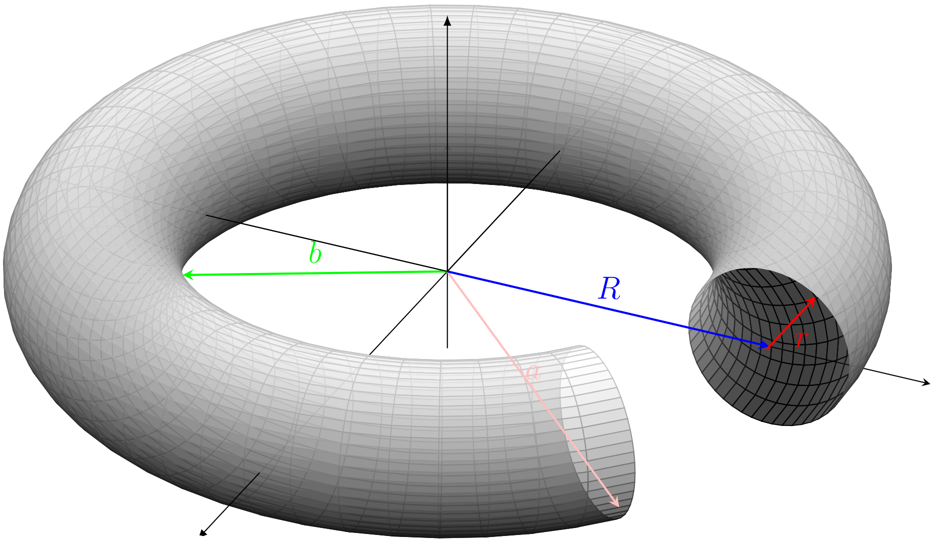


Figure 8.285: Torus

We see obviously in the figure above that: $a-b=2R$.

To calculate the volume we could obviously use Pappus's centroid theorem as proved in the section of Classical Mechanic. But the latter method is not very elegant and quite "physics" oriented. We will look therefore for a more pure mathematical approach.

Given the equation of a semicircle of center $(0,c)$ (see section Analytical Geometry page 2785):

$$x^2+(y-c)^2=r^2 \tag{8.7.313}$$

In order to write y as a function of x, let us isolate y in this equation:

$$x^2+(y-c)^2=r^2 \Leftrightarrow (y-c)^2=r^2-x^2 \Leftrightarrow y-c=\pm\sqrt{r^2-x^2} \Leftrightarrow y=c\pm\sqrt{r^2-x^2} \tag{8.7.314}$$

The circle is then made of the graphs of the following functions:

- Top semicircle:

$$f_1: x \mapsto c+\sqrt{r^2-x^2} \tag{8.7.315}$$

- Bottom semicircle:

$$f_2: x \mapsto c-\sqrt{r^2-x^2} \tag{8.7.316}$$

Let us denote now by usage with the study of the torus $R := c$.

The requested volume is the difference between the volumes generated by the rotation of the surfaces (surfaces defined by the area between the function of the circle in question and the axis of abscissa $x = r, x = -r$) in the space around the x-axis.

By applying the integration relation of solids of revolution we get:

$$\begin{aligned}
V=V_1-V_2&=\pi\int\limits_{-r}^{r}(R+\sqrt{r^2-x^2})^2\mathrm{d}x-\pi\int\limits_{-r}^{r}(R-\sqrt{r^2-x^2})^2\mathrm{d}x\\
&=\pi\int\limits_{-r}^{r}\left[(R+\sqrt{r^2-x^2})^2-(R-\sqrt{r^2-x^2})^2\right]\mathrm{d}x\\
&=\pi\int\limits_{-r}^{r}\left[(R^2+2R\sqrt{r^2-x^2}+r^2-x^2)-(R^2-2R\sqrt{r^2-x^2}+r^2-x^2\right]\mathrm{d}x\\
&=\pi\int\limits_{-r}^{r}(R^2+2R\sqrt{r^2-x^2}+r^2-x^2-R^2+2R\sqrt{r^2-x^2}-r^2+x^2)\mathrm{d}x\\
&=\pi\int\limits_{-r}^{r}4R\sqrt{r^2-x^2}\mathrm{d}x=4\pi c\int\limits_{-r}^{r}\sqrt{r^2-x^2}\mathrm{d}x
\end{aligned}\tag{8.7.317}$$

Let us calculate the latter integral by the classical substitution $x = r\sin(t)$ thus:

$$\mathrm{d}x=(r\sin(t))'\mathrm{d}t=r\cos(t)\mathrm{d}t\tag{8.7.318}$$

if $x = -r$:

$$-r=r\sin(t)\Rightarrow\sin(t)=-1\Rightarrow -t=\frac{\pi}{2}\tag{8.7.319}$$

if $x = r$:

$$r=r\sin(t)\Rightarrow\sin(t)=1\Rightarrow t=\frac{\pi}{2}\tag{8.7.320}$$

Therefore:

$$\begin{aligned}
V&=4\pi R\int\limits_{-\pi/2}^{\pi/2}\sqrt{r^2-r^2\sin^2(t)}r\cos(t)\mathrm{d}t=4\pi R\int\limits_{-\pi/2}^{\pi/2}\sqrt{r^2(1-\sin^2(t)}r\cos(t)\mathrm{d}t\\
&=4\pi R\int\limits_{-\pi/2}^{\pi/2}r\sqrt{\cos^2(t)}r\cos(t)\mathrm{d}t=4\pi Rr^2\int\limits_{-\pi/2}^{\pi/2}|\cos(t)|\cos(t)\mathrm{d}t=4\pi Rr^2\int\limits_{-\pi/2}^{\pi/2}\cos^2(t)\mathrm{d}t
\end{aligned}\tag{8.7.321}$$

Let us linearise that expression again using the trigonometric identities proved in the section

Trigonometry (Carnot's formula at page 2520):

$$\begin{aligned}V&=4\pi Rr^2\int\limits_{-\pi/2}^{\pi/2}\frac{1}{2}(1+\cos(2t))\mathrm{d}t=2\pi Rr^2\left(\int\limits_{-\pi/2}^{\pi/2}\mathrm{d}t+\int\limits_{-\pi/2}^{\pi/2}\cos(2t)\mathrm{d}t\right)\\&=2\pi Rr^2\left(t\Big|_{-\pi/2}^{\pi/2}+\frac{1}{2}\int\limits_{-\pi/2}^{\pi/2}2\cos(2t)\mathrm{d}t\right)=2\pi Rr^2\left(\frac{\pi}{2}-\left(-\frac{\pi}{2}\right)+\frac{1}{2}\left[\sin(2t)\right]\Big|_{-\pi/2}^{\pi/2}\right)\\&=2\pi Rr^2\left(\pi+\frac{1}{2}(\sin(\pi)-\sin(-\pi))\right)=2\pi Rr^2\left(\pi+\frac{1}{2}(0-0)\right)=2R\pi^2r^2\end{aligned}\tag{8.7.322}$$

So the volume of a torus of a minor radius r and major radius R is given by:

$$\boxed{V=2R\pi^2r^2}\tag{8.7.323}$$

and the surface (by derivation of the generating surface element... but there are many other elementary way to get this result):

$$\boxed{S=\frac{\mathrm{d}V}{\mathrm{d}r}=4Rr\pi^2}\tag{8.7.324}$$

The moment of inertia of the torus in relation to its axis of revolution is calculated as follows:

First we start using the torus volume by change the notations a little bit:

$$V=2\pi^2a^2b\tag{8.7.325}$$

The volumetric density of the torus is given by (weight to volume):

$$\rho=\frac{M}{V}=\frac{\mathrm{d}m}{\mathrm{d}V}\Leftrightarrow\mathrm{d}m=\frac{M}{2\pi^2a^2b}\mathrm{d}V\tag{8.7.326}$$

In cylindrical coordinates, we know that we have:

$$\mathrm{d}V=r\mathrm{d}r\mathrm{d}\theta\mathrm{d}z\tag{8.7.327}$$

Therefore:

$$\mathrm{d}m=\frac{M}{2\pi^2a^2b}\mathrm{d}V=\frac{M}{2\pi^2a^2b}r\mathrm{d}r\mathrm{d}\theta\mathrm{d}z\tag{8.7.328}$$

The moment of inertia is given by:

$$\begin{aligned}J&=\int r^2\mathrm{d}m=2\left(\frac{M}{2\pi^2a^2b}\int\limits_0^{2\pi}\mathrm{d}\theta\int\limits_{b-a}^{b+a}r^2\mathrm{d}r\int\limits_0^{\sqrt{a^2-(r-b)^2}}\mathrm{d}z\right)=\frac{M}{\pi^2a^2b}\int_0^{2\pi}\mathrm{d}\theta\int\limits_{b-a}^{b+a}r^3\mathrm{d}r\sqrt{a^2-(r-b)^2}\\&=\frac{M}{\pi^2a^2b}\int_0^{2\pi}\mathrm{d}\theta\int\limits_{b-a}^{b+a}r^3\sqrt{a^2-(r-b)^2}\mathrm{d}r=\frac{2\pi M}{\pi^2a^2b}\int\limits_{b-a}^{b+a}\sqrt{a^2-(r-b)^2}\mathrm{d}r\\&=\frac{2M}{\pi a^2b}\int\limits_{b-a}^{b+a}\sqrt{a^2-(r-b)^2}\mathrm{d}r\end{aligned}\tag{8.7.329}$$

Let us put $s=r-b$, then:

1. The limits of integration then become $-a$, $+a$, as we bring all integration points at the origin by putting $s=r-b$.
2. Trivially, since $r=s+b$ so we have $\mathrm{d}s=\mathrm{d}r$.

Which gives:

$$\begin{aligned}J_z&=\frac{2M}{\pi a^2 b}\int\limits_{b-a}^{b+a} r^2\sqrt{a^2-(r-b)^2}\mathrm{d}r=\frac{2M}{\pi a^2 b}\int\limits_{-a}^{a}(s+b)^3\sqrt{a^2-s^2}\mathrm{d}s\\&=\frac{2M}{\pi a^2 b}\int_{-a}^{+a}(s^3+3s^2b+3sb^2+b^3)\sqrt{a^2-s^2}\mathrm{d}s\end{aligned}\tag{8.7.330}$$

As we have proved in the section of Differential and Integral Calculus that the integral with two symmetrical terminals of an odd function (product of an even and odd function) is zero (see page 1516), the integrals of:

$$s^3\sqrt{a^2-s^2}\mathrm{d}s\qquad 3sb^2\sqrt{a^2-s^2}\mathrm{d}s\tag{8.7.331}$$

are equal to zero.

Then we have to calculate:

$$J_z=\frac{2M}{\pi a^2 b}\int\limits_{-a}^{+a}(3s^2b+b^3)\sqrt{a^2-s^2}\mathrm{d}s\tag{8.7.332}$$

Now let us put $s=a\sin(t)$ and therefore $\mathrm{d}s=a\cos(t)\mathrm{d}t$. Therefore it comes:

$$J_z=\frac{2M}{\pi a^2 b}\int\limits_{-\pi/2}^{+\pi/2}(3a^2\sin^2(t)b+b^3)\sqrt{a^2-a^2\sin^2(t)}a\cos(t)\mathrm{d}t\tag{8.7.333}$$

But as:

$$\cos^2(t)+\sin^2(t)=1\Rightarrow a^2\cos^2(t)+a^2\sin^2(t)=a^2\Rightarrow\sqrt{a^2-a^2\sin^2(t)}=|a\cos(t)|\tag{8.7.334}$$

Therefore:

$$\begin{aligned}J_z&=\frac{2M}{\pi a^2}\int\limits_{-\pi/2}^{+\pi/2}(3a^2\sin^2(t)+b^2)|a\cos(t)|a\cos(t)\mathrm{d}t=\frac{2M}{\pi a^2}\int\limits_{-\pi/2}^{+\pi/2}(3a^2\sin^2(t)+b^2)a^2\cos^2(t)\mathrm{d}t\\&=\frac{2M}{\pi}\int\limits_{-\pi/2}^{+\pi/2}(3a^2\sin^2(t)+b^2)\cos^2(t)\mathrm{d}t=\frac{2M}{\pi}\int\limits_{-\pi/2}^{+\pi/2}(3a^2\sin^2(t)\cos^2(t)+b^2\cos^2(t))\mathrm{d}t\\&=\frac{2M}{\pi}\int\limits_{-\pi/2}^{+\pi/2}(3a^2(1-\cos^2(t))\cos^2(t)+b^2\cos^2(t))\mathrm{d}t\\&=\frac{2M}{\pi}\int\limits_{-\pi/2}^{+\pi/2}(3a^2\cos^2(t)-3a^2\cos^4(t)+b^2\cos^2(t))\mathrm{d}t\end{aligned}\tag{8.7.335}$$

Therefore (see section Differential and Integral Calculus page 1534):

$$\int\limits_{-\pi/2}^{+\pi/2} 3a^2\cos^2(t)\mathrm{d}t = 3a^2\int\limits_{-\pi/2}^{+\pi/2}\cos^2(t)\mathrm{d}t = \frac{1}{2}3a^2(\cos(t)\sin(t)+t)\Big|_{-\pi/2}^{+\pi/2} = 3a^2\frac{\pi}{2}$$
$$\int\limits_{-\pi/2}^{+\pi/2} b^2\cos^2(t)\mathrm{d}t = b^2\int\limits_{-\pi/2}^{+\pi/2}\cos^2(t)\mathrm{d}t = \frac{b^2}{2}(\cos(t)\sin(t)+t)\Big|_{-\pi/2}^{+\pi/2} = b^2\frac{\pi}{2}$$
$$\int\limits_{-\pi/2}^{+\pi/2} 3a^2\cos^4(t)\mathrm{d}t = 3a^2\int\limits_{-\pi/2}^{+\pi/2}\cos^4(t)\mathrm{d}t = \left(\frac{1}{4}\cos^3(t)\sin(t)+\frac{3}{8}\cos(t)\sin(t)+\frac{3}{8}t\right)\Big|_{-\pi/2}^{+\pi/2} = 3a^2\frac{3\pi}{8} \tag{8.7.336}$$

So finally:

$$J_z = \frac{2M}{\pi}\left(-3a^2\frac{3\pi}{8}+b^2\frac{\pi}{2}+3a^2\frac{\pi}{2}\right) = M\left(\frac{3}{4}a^2+b^2\right) \tag{8.7.337}$$

8.7.2.2.5 Ellipsoid (spheroid)

Definitions (#8.7.395):

D1. An "ellipsoid" is a surface of the second degree of a three dimensional Euclidean space. It is therefore part of quadrics (see section Analytical Geometry page 2765) given by the equation:

$$\boxed{\frac{x^2}{a^2}+\frac{y^2}{b^2}+\frac{z^2}{c^2}=1} \tag{8.7.338}$$

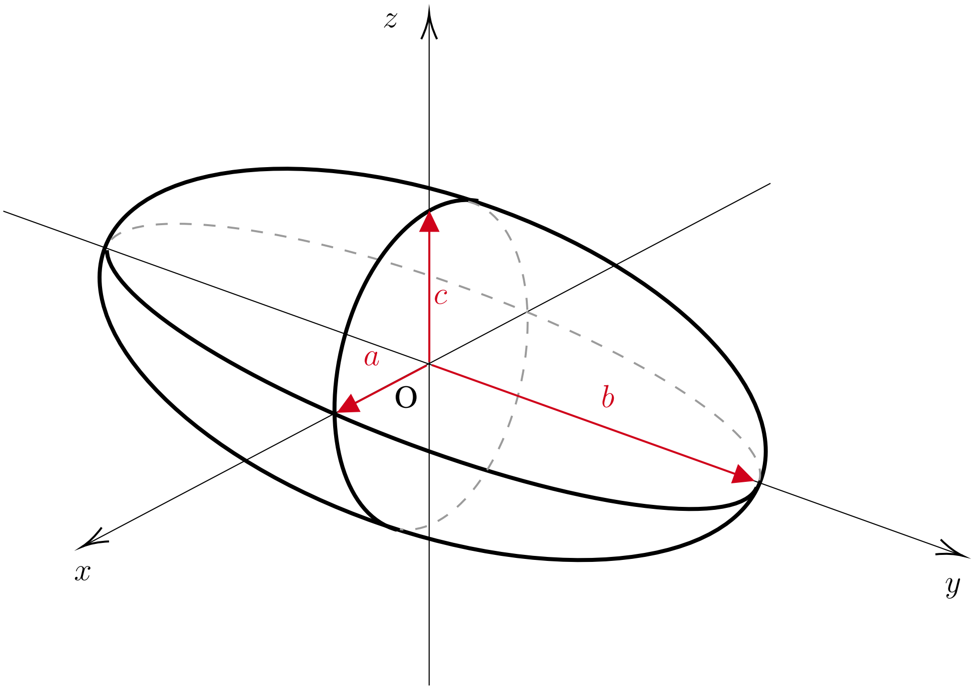


Figure 8.286: **Ellipsoid**

D2. A "spheroid", or "ellipsoid of revolution, is a quadric surface obtained by rotating an ellipse about one of its principal axes; in other words, an ellipsoid with two equal semi-diameters.

The equation of a spheroid with z as the symmetry axis is given by setting $a=b$:

$$\boxed{\frac{x^2+y^2}{a^2}+\frac{z^2}{c^2}=1} \tag{8.7.339}$$

If the ellipse is rotated about its major axis, the result is a "prolate (elongated) spheroid", like an American football or rugby ball. If the ellipse is rotated about its minor axis, the result is an "oblate (flattened) spheroid", like a lentil:

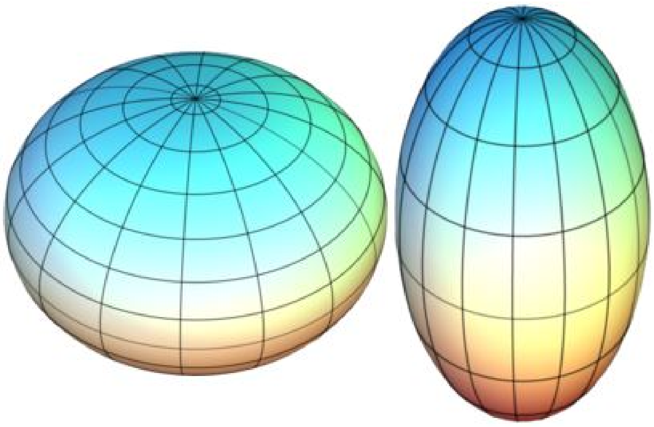


Figure 8.287: **Oblate spheroid on the left, prolate one on the right (source: Wikipedia)**

To calculate the volume delimited by an ellipsoid, we take the equation that we have determined during our study of conicals:

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}+\frac{z^2}{c^2}=1 \tag{8.7.340}$$

The section by a plane parallel to the plane YZ and lying at a distance x of the latter, gives the ellipse:

$$\frac{y^2}{b^2}+\frac{z^2}{c^2}=1-\frac{x^2}{a^2} \tag{8.7.341}$$

or:

$$\frac{y^2}{b^2\left(1-\dfrac{x^2}{a^2}\right)}+\frac{z^2}{c^2\left(1-\dfrac{x^2}{a^2}\right)}=\frac{y^2}{\left(b\sqrt{1-\dfrac{x^2}{a^2}}\right)}+\frac{z^2}{\left(c\sqrt{1-\dfrac{x^2}{a^2}}\right)}=1 \tag{8.7.342}$$

with for semi-axes:

$$b_1=b\sqrt{1-\frac{x^2}{a^2}} \qquad c_1=c\sqrt{1-\frac{x^2}{a^2}} \tag{8.7.343}$$

But as we proved it earlier, the surface of an ellipse is equal to $\pi b_1 c_2$. Therefore:

$$\boxed{S=\pi bc\left(1-\frac{x^2}{a^2}\right)} \tag{8.7.344}$$

The volume of the ellipsoid is therefore equal to:

$$\begin{aligned}V&=\pi bc\int\limits_{-a}^{+a}\left(1-\frac{x^2}{a^2}\right)\mathrm{d}x=\pi bc\left(x-\frac{x^3}{3a^2}\right)\Bigg|_{-a}^{+a}=\pi bc\left[\left(a-\frac{a^3}{3a^2}\right)-\left((-a)-\frac{(-a)^3}{3a^2}\right)\right]\\&=\pi bc\left(2a-2\frac{a^3}{3a^2}\right)=\pi bc\left(2a-\frac{2}{3}a\right)\end{aligned} \tag{8.7.345}$$

Therefore:

$$\boxed{V=\frac{4}{3}\pi abc} \tag{8.7.346}$$

If $a=b=c=R$ we fall back on the volume of a sphere:

$$V=\frac{4}{3}\pi R^3 \tag{8.7.347}$$

Let us now calculate the surface and volume of an oblate spheroid. Remember that its Cartesian equation is:

$$\frac{x^2+y^2}{a^2}+\frac{z^2}{c^2}=1 \tag{8.7.348}$$

The ellipticity of an oblate spheroid is defined by:

$$e=\sqrt{1-\frac{c^2}{a^2}} \tag{8.7.349}$$

The surface area of an oblate spheroid can be computed as a surface of revolution about the z-axis:

$$S=2\pi\int r(z)\sqrt{1+(r'(z))^2}\mathrm{d}z \tag{8.7.350}$$

with radius function of z given by:

$$r(z)=a\sqrt{1-\left(\frac{z}{c}\right)^2} \tag{8.7.351}$$

Therefore:

$$(r'(z))^2=\left(-\frac{az}{c^2\sqrt{1-\frac{z^2}{c^2}}}\right)^2=\frac{a^2z^z}{c^4\left(1-\left(\frac{z}{c}\right)^2\right)} \tag{8.7.352}$$

Therefore:

$$\begin{aligned}S&=2\pi\int a\sqrt{1-\left(\frac{z}{c}\right)^2}\sqrt{1+\frac{a^2z^z}{c^4\left(1-\left(\frac{z}{c}\right)^2\right)}}\mathrm{d}z=2\pi a\int\sqrt{\left(1-\left(\frac{z}{c}\right)^2\right)+\frac{a^2z^z}{c^4}}\mathrm{d}z\\&=2\pi a\int\sqrt{1-\left(\frac{z}{c}\right)^2+\frac{a^2z^z}{c^4}}\mathrm{d}z=2\pi a\int\sqrt{1-\frac{a^2z^2-z^2c^4}{c^2}}\mathrm{d}z=2\pi a\int\sqrt{1-\frac{(a-c)(a+c)}{c^2}z^2}\mathrm{d}z\end{aligned} \tag{8.7.353}$$

If we put:

$$\alpha^{-1}:=\frac{(a-c)(a+c)}{c^2} \tag{8.7.354}$$

We can then write the integral as following:

$$F(z)=\int\sqrt{\alpha-z^2}\mathrm{d}z \tag{8.7.355}$$

And as we have seen in the section of Differential and Integral Calculus the usual primitives (see page 1534) and the following identity $-\arcsin(\alpha)=\arcsin(-\alpha)$ in the section of Trigonometry (see 2519):

$$\begin{aligned}F(z)&=\left(\frac{\alpha^2}{2}\arcsin\left(\frac{z}{\alpha}\right)+\frac{z}{2}\sqrt{\alpha^2-z^2}\right)\Bigg|_{-c}^{+c}\\&=\left(\frac{\alpha^2}{2}\arcsin\left(\frac{-c}{\alpha}\right)+\frac{-c}{2}\sqrt{\alpha^2-(-c)^2}\right)-\left(\frac{\alpha^2}{2}\arcsin\left(\frac{c}{\alpha}\right)+\frac{c}{2}\sqrt{\alpha^2-(c)^2}\right)\\&=-c\sqrt{\alpha^2-c^2}+\alpha^2\arcsin\left(\frac{c}{\alpha}\right)\end{aligned} \tag{8.7.356}$$

Therefore:

$$S=2\pi a\left[-c\sqrt{\alpha^2-c^2}+\alpha^2\arcsin\left(\frac{c}{\alpha}\right)\right]=2\pi a\left[-c\sqrt{\left(\frac{c^2}{a^2-c^2}\right)^2-c^2}+\alpha^2\arcsin\left(\frac{c}{\alpha}\right)\right] \tag{8.7.357}$$

The calculation of the moment of inertia of an ellipsoid is very important in astrophysics since a large number of stars or planets in rotation on themselves by their deformation at the equator because of the centrifugal (fictional) force (see section Classical Mechanics page 3388) are being distorted in a first approximation in this volume.

For an ellipsoid, define C as being the moment of inertia along the c-axis, A is the moment of inertia along the a-axis, and B the moment of inertia along the b-axis.

To begin, let us consider the moment of inertia along the c-axis that we will assimilate to the z-axis. Thus, in Cartesian coordinates, we have:

$$J_z=C=\int r^2\mathrm{d}m=\int\rho r^2\mathrm{d}V=\rho\int r^2\mathrm{d}v=\rho\int(x^2+y^2)\mathrm{d}x\mathrm{d}y\mathrm{d}z \tag{8.7.358}$$

By making the following substitution, we implicitly do that the previous integral is a normalization of an ellipsoid:

$$x'=\frac{x}{a},y'=\frac{y}{b},z'=\frac{z}{c} \tag{8.7.359}$$

what gives us for our integral (so we transform the volume V of the ellipsoid in the volume V' of a sphere):

$$C=\rho\int\left((ax')^2+(by')2\right)abc \tag{8.7.360}$$

We can now move from Cartesian coordinates to spherical coordinates (see section Vector Calculus page 1849) without forgetting to use the Jacobian (see section Differential and Integral Calculus page 1524) that we had proved as being in spherical coordinates given by:

$$|\det(J)|=r^2\sin(\theta) \tag{8.7.361}$$

Therefore (we use again the common primitives proved in the section of Differential and Integral

Calculus):

$$\begin{aligned}
C &= abc\rho \int\limits_0^1 \int\limits_0^{2\pi} \int\limits_0^{\pi} (a^2 r^2 \cos^2(\phi)\sin^2(\theta) + b^2 r^2 \sin^2(\phi)\sin^2(\theta) r^2 \sin(\theta)\mathrm{d}\theta\mathrm{d}\phi\mathrm{d}r \\
&= abc\rho \int\limits_0^1 \int\limits_0^{2\pi} \int\limits_0^{\pi} (a^2\cos^2(\phi) + b^2\sin^2(\phi))r^4\sin^3(\theta)\mathrm{d}\theta\mathrm{d}\phi\mathrm{d}r \\
&= abc\rho \int\limits_0^1 \int\limits_0^{2\pi} \int\limits_0^{\pi} (a^2\cos^2(\phi) + b^2\sin^2(\phi))r^2 \left(-\frac{1}{3}\cos(\theta)\sin^2(\theta) - \frac{2}{3}\cos(\theta)\right)\Bigg|_0^{\pi} \mathrm{d}\phi\mathrm{d}r \\
&= \frac{4}{3}abc\rho \int\limits_0^1 \int\limits_0^{2\pi} (a^2\cos^2(\phi) + b^2\sin^2(\phi) r^2 \mathrm{d}\phi\mathrm{d}r \\
&= \frac{4}{3}abc\rho \int\limits_0^1 \left(a^2\pi + b^2\pi\right) r^4 \mathrm{d}r = \frac{4}{3}abc\pi\rho(a^2+b^2) \int\limits_0^1 r^4\mathrm{d}r \\
&= \frac{4}{15}abc\pi\rho(a^2+b^2)r^5\Bigg|_0^1 = \frac{4}{15}\pi\rho abc(a^2+b^2)
\end{aligned} \tag{8.7.362}$$

By injecting for the ellipsoid:

$$\rho = \frac{M}{V} = \frac{M}{\dfrac{4}{3}\pi abc} \tag{8.7.363}$$

Then we get:

$$C = \frac{1}{5}M(a^2+b^2) \tag{8.7.364}$$

and by symmetry, we get the following obvious results:

$$\begin{aligned}
A &= J_x = \frac{1}{5}M(b^2+c^2) \\
B &= J_y = \frac{1}{5}M(a^2+c^2) \\
C &= J_z = \frac{1}{5}M(a^2+b^2)
\end{aligned} \tag{8.7.365}$$

The inertia matrix (see section Classical Mechanics page 3310) is therefore:

$$(J_0)_{(\vec{i},\vec{j},\vec{k})} = \begin{pmatrix} \frac{1}{5}M(b^2+c^2) & 0 & 0 \\ 0 & \frac{1}{5}M(a^2+c^2) & 0 \\ 0 & 0 & \frac{1}{5}M(a^2+b^2) \end{pmatrix} \tag{8.7.366}$$

8.7.2.2.6 Paraboloid

Definition (#8.7.396): A "paraboloid" is as we already know a quadric surface of special kind. There are two kinds of paraboloids: elliptic and hyperbolic.

The "elliptic paraboloid" is shaped like an oval cup and can have a maximum or minimum point. In a suitable coordinate system with three axes x, y, and z, it can be represented by the equation:

$$\frac{z}{c}=\frac{x^2}{a^2}+\frac{y^2}{b^2} \tag{8.7.367}$$

where a and b are constants that dictate the level of curvature in the xz and yz planes respectively. This is an elliptic paraboloid which opens upward for $c>0$ and downward for $c<0$.

The latter relations is more commonly written as:

$$z=c^2\frac{x^2}{a^2}+c^2\frac{y^2}{b^2}=a'x^2+b'y^2 \tag{8.7.368}$$

The "hyperbolic paraboloid" - not to be confused with a hyperboloid (see section Analytical Geometry page 2841) - is a doubly ruled surface shaped like a saddle. In a suitable coordinate system, a hyperbolic paraboloid can be represented by the equation:

$$\frac{z}{c}=\frac{y^2}{b^2}-\frac{x^2}{a^2} \tag{8.7.369}$$

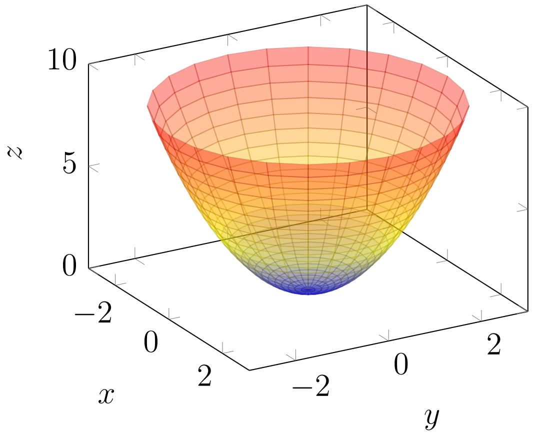


Figure 8.288: Paraboloid of revolution

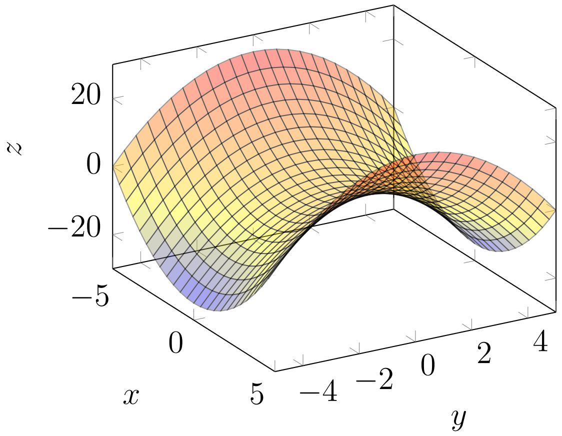


Figure 8.289: Hyperbolic paraboloid

We will only calculate the volume of the paraboloid of revolution as it is the only result that we need in the other chapters of this textbook at this date. For this purpose let us consider the following figure:

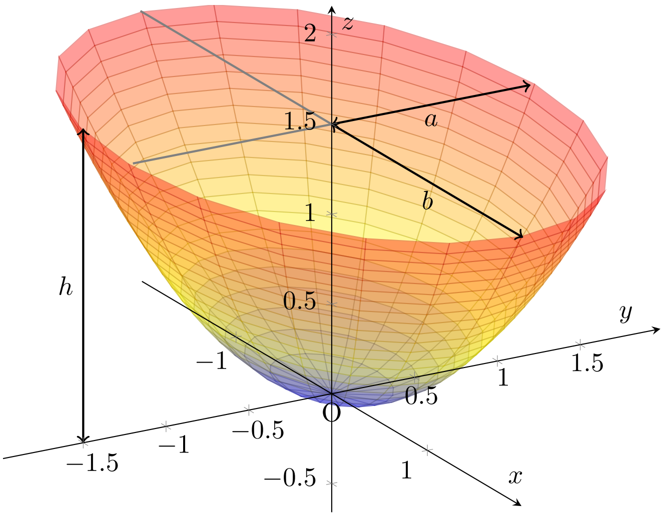


Figure 8.290: **Elliptic paraboloid**

Given by:

$$z=\frac{x^2}{a^2}+\frac{y^2}{b^2} \tag{8.7.370}$$

Looking at the figure above we see that the parameters a and b depends on the value of z. So as the surface of an ellipse is given by $\mathrm{d}S=\pi ab$, we have for every height of the paraboloid:

$$S(z)=\pi a(z)b(z) \tag{8.7.371}$$

So in the xz plane as we have $y=0$ then:

$$z=\frac{x^2}{a^2}\Leftrightarrow a(z)=|x|\sqrt{\frac{1}{z}} \tag{8.7.372}$$

and also same for the plane yz:

$$z=\frac{y^2}{b^2}\Leftrightarrow b(z)=|y|\sqrt{\frac{1}{z}} \tag{8.7.373}$$

And we know that when $z=h$ we must have $a(h)=a$, and when $z=0$ we must have $a(h)=0$ (same with b). So we have to write:

$$a(z)=a\sqrt{\frac{z}{h}} \qquad b(z)=b\sqrt{\frac{z}{h}} \tag{8.7.374}$$

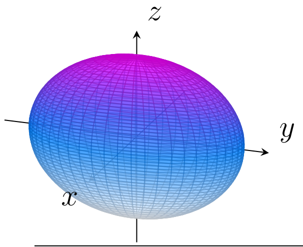
Therefore:

$$V_h=\int\limits_0^h S(z)\mathrm{d}z=\int\limits_0^h \pi a(z)b(z)\mathrm{d}z=\pi\int\limits_0^h a\sqrt{\frac{z}{h}}b\sqrt{\frac{z}{h}}\mathrm{d}z=\pi ab\frac{1}{h}\int\limits_0^h |z|\mathrm{d}z=\pi ab\frac{1}{h}\frac{1}{2}h^2 \tag{8.7.375}$$

Hence:

$$\boxed{V_h=\frac{1}{2}\pi abh} \tag{8.7.376}$$

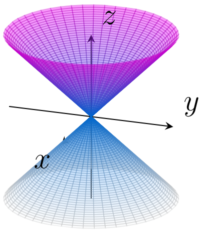
Quadric Surfaces at a Glance



Ellipsoid

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$$

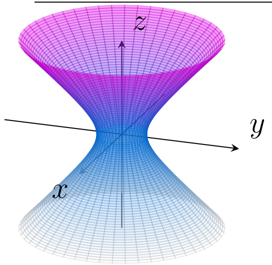
All traces are ellipses. If $a = b = c$, the ellipsoid is a sphere.



Cone

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = \frac{z^2}{c^2}$$

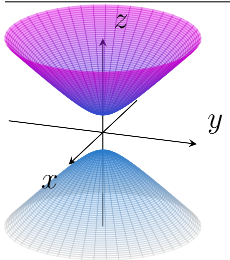
Horizontal traces are ellipses. Vertical traces are hyperbolas (or a pair of lines). In general, the surface opens along the axis corresponding to the squared term that is on a side of the equation by itself.



Hyperboloid of One Sheet

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1$$

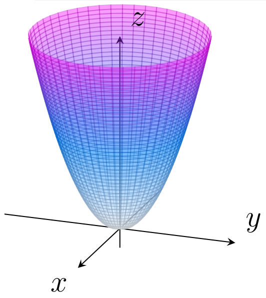
Horizontal traces are ellipses. Vertical traces are hyperbolas (or pairs of lines). In general, the surface opens along the axis corresponding to the variable whose coefficient is negative.



Hyperboloid of Two Sheets

$$\frac{z^2}{c^2} - \frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$$

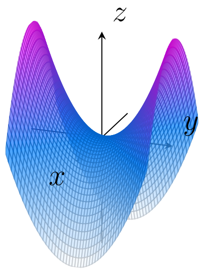
Horizontal traces are ellipses (or a single point or nothing). Vertical traces are hyperbolas. The surface opens along the axis corresponding to the variable whose coefficient is positive.



Elliptic Paraboloid

$$\frac{z}{c} = \frac{x^2}{a^2} + \frac{y^2}{b^2}$$

Horizontal traces are ellipses (or a point or nothing). Vertical traces are parabolas. The paraboloid opens upward when $c > 0$ and downward when $c < 0$. In general, the surface opens along the axis corresponding to the variable raised to the first power.



Hyperbolic Paraboloid

$$\frac{z}{c} = \frac{x^2}{a^2} - \frac{y^2}{b^2}$$

Horizontal traces are hyperbolas (or a pair of lines). Vertical traces are parabolas. In general, the parabolas open along the axis corresponding to the variable raised to the first power.

Figure 8.291: Quadric surfaces at a glance (author: Dave Richeson)

8.7.2.2.7 Gabriel's horn

"Gabriel's horn" (also named "Torricelli's trumpet") is a geometric figure which has infinite surface area but finite volume. The name refers to the Abrahamic tradition identifying the Archangel Gabriel as the angel who blows the horn to announce Judgement Day, associating the divine, or infinite, with the finite. The properties of this figure were first studied by Italian physicist and mathematician Evangelista Torricelli in the 117th century (holocene calendar).

Gabriel's horn is formed by taking the graph of:

$$f(x) \mapsto \frac{1}{x} \tag{8.7.377}$$

with the domain $x \geq 1$ and rotating it in three dimensions about the x-axis. With Maple™ 4.00b we can plot it using (notice that is has the same shape as a gravitational potential or electric potential):

```
>plot3d(1/x, theta=0..2*Pi, x = 1..20, coords=cylindrical
```

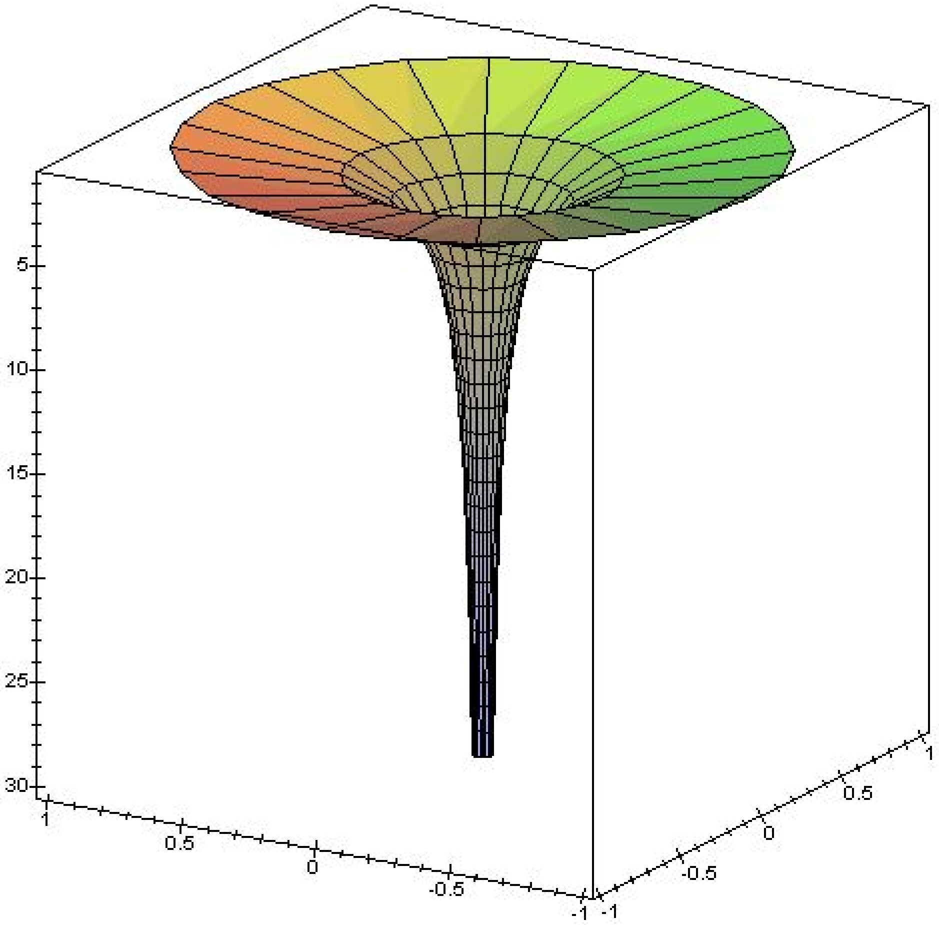


Figure 8.292: **Gabriel's horn with Maple™ 4.00b**

The discovery was made using Cavalieri's principle before the invention of calculus, but today calculus can be used to calculate the volume and surface area of the horn between $x = 1$ and $x = a$, where $a > 1$. Using integration, it is possible to find the volume V (using the method of discs again for the volume) and the surface area S:

$$\begin{gathered}
V=\pi\int\limits_1^a (f(x))^2\,\mathrm{d}x=\pi\int\limits_1^a\left(\frac{1}{x}\right)^2\mathrm{d}x=\pi\left(1-\frac{1}{a}\right)\\
S=2\pi\int\limits_a^b f(x)\sqrt{1+(f'(x))^2}\mathrm{d}x=2\pi\int\limits_1^a\frac{1}{x}\sqrt{1+\left(-\frac{1}{x^2}\right)^2}\mathrm{d}x>2\pi\int\limits_1^a\frac{\mathrm{d}x}{x}=2\pi\ln(a)
\end{gathered}\tag{8.7.378}$$

The parameter a can be as large as required, but it can be seen from the equation that the volume of the part of the horn between $x=1$ and $x=a$ will never exceed π; however, it will get closer and closer to π as a becomes larger. Mathematically, the volume approaches π as a approaches infinity. Using the limit notation of calculus:

$$\lim_{a\rightarrow+\infty}V=\lim_{a\rightarrow+\infty}\pi\left(1-\frac{1}{a}\right)=\pi\cdot\lim_{a\rightarrow+\infty}\left(1-\frac{1}{a}\right)=\pi\tag{8.7.379}$$

The surface area formula above gives a lower bound for the area as 2π times the natural logarithm of a. There is no upper bound for the natural logarithm of a as a approaches infinity. That means, in this case, that the horn has an infinite surface area. That is to say:

$$\lim_{a\rightarrow+\infty}S\geq\lim_{a\rightarrow+\infty}2\pi\ln(a)=+\infty\tag{8.7.380}$$

The apparent paradox formed part of a dispute over the nature of infinity involving many of the key thinkers of the time including Thomas Hobbes, John Wallis and Galileo Galilei.

Remark

Since the horn has finite volume but infinite surface area, there is an apparent paradox that the horn could be filled with a finite quantity of paint, and yet that paint would not be sufficient to coat its inner surface. The "paradox" is resolved by realizing that a finite amount of paint can in fact coat an infinite surface area - it simply needs to get thinner at a fast enough rate.

The converse phenomenon of Gabriel's horn – a surface of revolution that has a finite surface area but an infinite volume – cannot occur (there is a proof but it is not a very aesthetic one so we will not present it here).

8.7.2.2.8 Wine Barrel with Circular Section

Now let us look for fun to a well known volume for winemakers (and not only!):

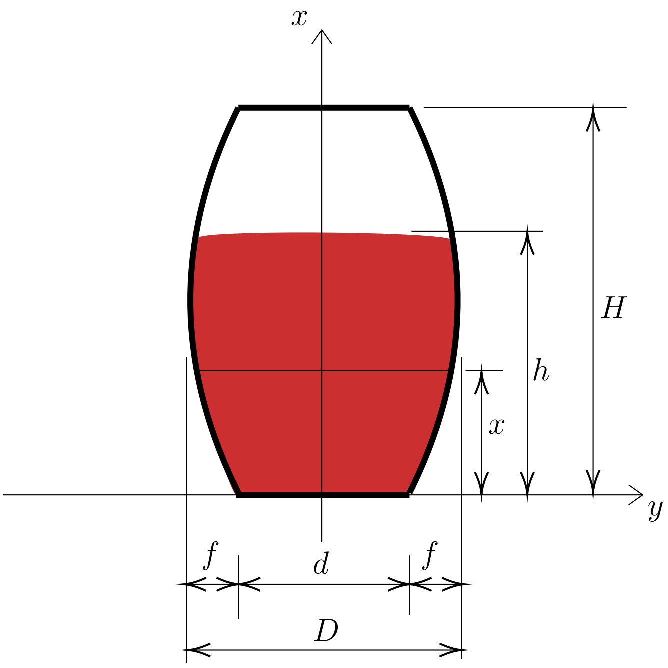


Figure 8.293: **Wine Barrel with circular section**

Let us consider that the lateral curvature of the barrel is a parabola:

$$y=ax^2+bx+c \tag{8.7.381}$$

Let us put:

$$f=\frac{D-d}{2} \quad \text{and} \quad L=\frac{H}{2} \tag{8.7.382}$$

given the way we have put the x,y-axis it is relatively easy to determine the coefficients of the polynomial. The coefficient c is the simplest:

$$c=y(x=0)=\frac{d}{2} \tag{8.7.383}$$

We also have:

$$y'=0=2ax+b\Rightarrow 2a\frac{H}{2}+b=2aL+b=0\Rightarrow a=\frac{-b}{2L} \tag{8.7.384}$$

Also:

$$\begin{gathered}y\left(x=\frac{H}{2}\right)=a\left(x=\frac{H}{2}\right)^2+b\frac{H}{2}+c=-\frac{b}{2L}L^2+bL+\frac{d}{2}=\frac{D}{2}\\ \Rightarrow b\left(-\frac{L}{2}+L\right)=\frac{D-d}{2}\Rightarrow b\frac{L}{2}=f\Rightarrow b=\frac{2f}{L}\end{gathered} \tag{8.7.385}$$

Therefore we have:

$$a=-\frac{f}{L^2} \quad b=\frac{2f}{L} \quad c=\frac{d}{2} \tag{8.7.386}$$

The radius of a horizontal section of ordinate x is $r=y$ and its surface is therefore:

$$S=\pi r^2 \tag{8.7.387}$$

or:

$$S=\pi(ax^2+bx+c)^2 \tag{8.7.388}$$

Let us develop this:

$$\begin{aligned}S&=\pi\left(a^2x^4+b^2x^2+c^2+2ax^2(bx+c)+bxc\right)\\&=\pi\left(a^2x^4+b^2x^2+c^2+2abx^3+2acx^2+2bxc\right)\end{aligned} \tag{8.7.389}$$

The volume of liquid for a height h will then be:

$$\begin{aligned}V&=\int\limits_0^h(a^2x^4+b^2x^2+c^2+2abx^3+2acx^2+2bcx)\mathrm{d}x\\&=\pi\left[a^2\frac{x^5}{5}+b^2\frac{x^3}{3}+c^2x+2ab\frac{x^4}{4}+2ac\frac{x^3}{3}+2bc\frac{x^2}{2}\right]\Bigg|_0^h\\&=\pi\left[a^2\frac{h^5}{5}+2ab\frac{h^4}{4}+b^2\frac{h^3}{3}+2ac\frac{h^3}{3}+2bc\frac{h^2}{2}+c^2h\right]\\&=\pi\left[a^2\frac{h^5}{5}+ab\frac{h^4}{2}+(2ac+b^2)\frac{h^3}{3}+bch^2+c^2h\right]\end{aligned} \tag{8.7.390}$$

To calculate the inner surface of the wine barrel, we consider the outer curve given by a parabolic arc as shown below:

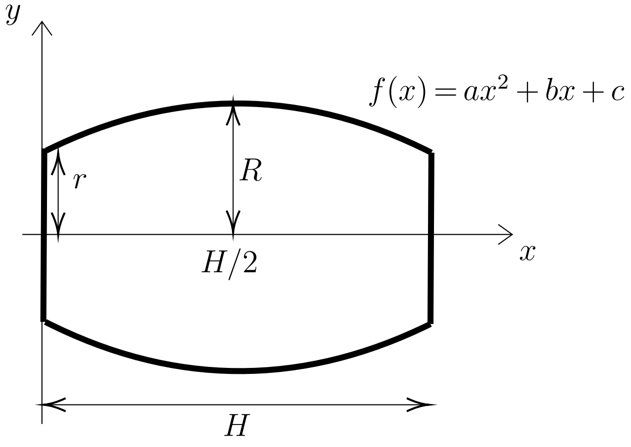


Figure 8.294: **Vertical section of the wine barell**

To calculate the lateral area of the barrel, we must first determine the expression of the above parabola.

Looking at the figure, we get:

$$\begin{aligned}f(0)&=r=c\\f(H)&=r=aH^2+bH+c\\f\left(\frac{H}{2}\right)&=a\frac{H^2}{4}+b\frac{H}{2}+c=R\end{aligned} \tag{8.7.391}$$

which is a system of three equations with the unknowns a, b, c.

After resolution, we get:

$$\begin{aligned} a &= \frac{4(r-R)}{H^2} \\ b &= \frac{4(R-r)}{H} \\ c &= r \end{aligned} \tag{8.7.392}$$

The side surface of the barrel including the surface of the two disks at the ends is then given by:

$$S = 2\pi r^2 + 2\pi \int\limits_0^H |f(x)|\sqrt{1+(f'(x))^2}\mathrm{d}x = 2\pi r^2 + 2\pi \int\limits_0^H (ax^2+bx+c)\sqrt{1+(2ax+b)^2}\mathrm{d}x \tag{8.7.393}$$

By doing the change of variable $u = 2ax+b$ we get:

$$S = 2\pi r^2 + \frac{\pi}{4a^2}\int\limits_b^{2aH+b}(u^2-b^2+4ac)\sqrt{u^2+1}\mathrm{d}u \tag{8.7.394}$$

The latter integral can be calculated using the following relations (the second has been proved in the section of Differential and Integral Calculus the first one need to be detailed when we will have the time...):

$$\begin{aligned} \int u^2\sqrt{u^2+1}\mathrm{d}u &= \frac{1}{4}u\left(u^2+\frac{1}{2}\right)\sqrt{u^2+1}-\frac{1}{8}\text{arcsinh}(u) \\ \int \sqrt{u^2+1}\mathrm{d}u &= \frac{1}{2}u\sqrt{u^2+1}+\frac{1}{2}\text{arcsinh}(u) \end{aligned} \tag{8.7.395}$$

Where for recall (see section Trigonometry page 2532):

$$\text{arcsinh}(z) = \ln(u+\sqrt{u^2+1}) \tag{8.7.396}$$

We will not go further because the resulting formula would be huge and not very interesting in our point of view (but on request we will do it).

Nevertheless, here is a numerical application. Suppose $r = 20$ [cm], $R = 30$ [cm], $H = 60$ [cm].

We calculate:

$$a = -0.0111, b = 0.667, c = 20 \tag{8.7.397}$$

Therefore:

$$S = 2.513\cdot 10^3 + 6.362\cdot 10^3\int\limits_{0.667}^{-0.667}u^2\sqrt{u^2+1}\mathrm{d}u - 8.483\cdot 10^3\int\limits_{0.667}^{-0.667}\sqrt{u^2+1}\mathrm{d}u \tag{8.7.398}$$

The first integral is equal to:

$$\int\limits_{0.667}^{-0.667}u^2\sqrt{u^2+1}\mathrm{d}u = -0.222 \tag{8.7.399}$$

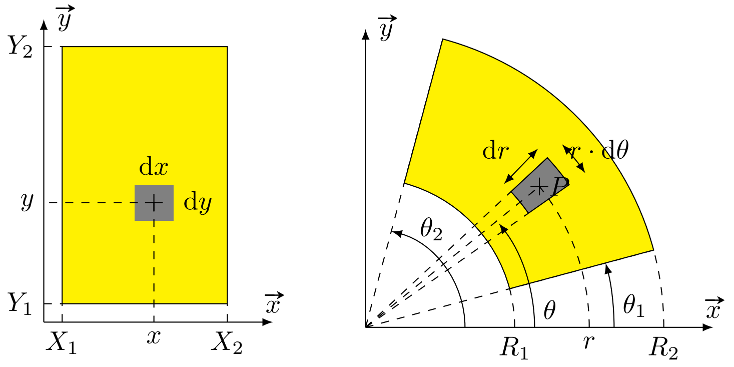
The second integral is equal to:

$$\int\limits_{0.667}^{-0.667}\sqrt{u^2+1}\mathrm{d}u=-1.427 \tag{8.7.400}$$

So finally we get:

$$S\cong 1.321\,[\mathrm{m}^2] \tag{8.7.401}$$

To summarize everything seen so far about surfaces and volumes, let us show all surfaces elements in 2D and also all surfaces elements in 3D and volumes techniques used so far:

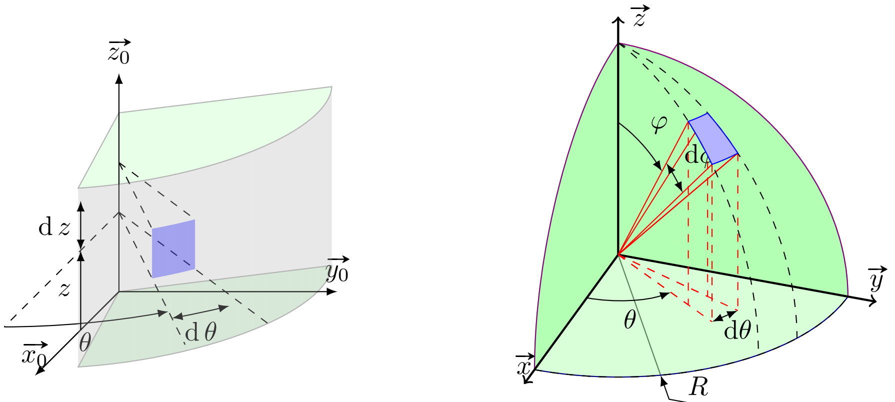

$\vec{y}$
Y_2
y
Y_1
X_1
x
X_2
$\vec{x}$
$\mathrm{d}x$
$\mathrm{d}y$
dr
$r\cdot\mathrm{d}\theta$
P
θ_2
θ
θ_1
R_1
r
R_2

$\mathrm{d}S = \mathrm{d}x \cdot \mathrm{d}y$

(a) Cartesian coordinates

$\mathrm{d}S = r \cdot \mathrm{d}\theta \cdot \mathrm{d}r$

(b) Polar coordinates

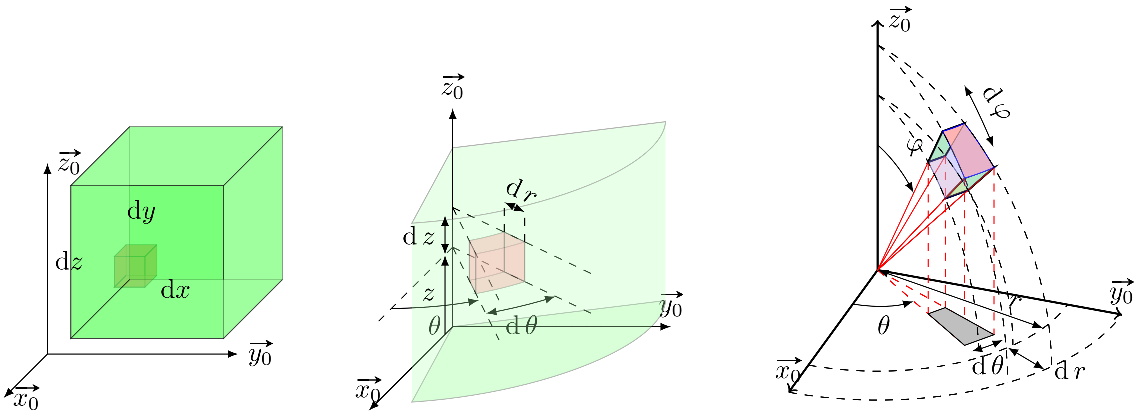

$\vec{z_0}$
d z
z
$\vec{y_0}$
$\vec{x_0}$
θ
d θ
$\vec{z}$
φ
dφ
$\vec{y}$
$\vec{x}$
θ
dθ
R

$\mathrm{d}S = r \cdot \mathrm{d}\theta \cdot \mathrm{d}z$

(c) Cylindrical coordinates

$\mathrm{d}S = R \cdot \sin\varphi \cdot \mathrm{d}\theta \cdot R \cdot \mathrm{d}\varphi$

(d) Spherical coordinates


$\vec{z_0}$
dy
dz
dx
$\vec{y_0}$
$\vec{x_0}$
d r
d z
z
θ
d θ
φ
d φ
r
d r

$\mathrm{d}V = \mathrm{d}x \cdot \mathrm{d}y \cdot \mathrm{d}z$

(e) Cartesian coordinates

$\mathrm{d}V = r \cdot \mathrm{d}\theta \cdot \mathrm{d}r \cdot \mathrm{d}z$

(f) Cylindrical coordinates

$\mathrm{d}V = r \cdot \sin\varphi \cdot \mathrm{d}\theta \cdot r \cdot \mathrm{d}\phi \cdot \mathrm{d}r.$

(g) Spherical coordinates

Figure 8.295: 2D and 3D surface and volumes elements techniques in different coordinates systems (author: Robert Papanicola)

8.7.3 Isoperimetric inequality (2D)

The text that follows is almost just a copy/paste of a text written by Jeremy Booher where we added some missing details and cross-references! The reader will see that the method follows a clever process that is very interesting to learn how to make some kind of abstract proofs on geometric objects.

Of all figures in the plane with a given perimeter, which has the greatest area? The Greeks knew the answer to this question, although they could not prove it. Part of the problem was that the notion of perimeter and area are tricky notions, requiring analysis to deal with properly. A rigorous proof had to wait until the 119th century (holocene calendar). There are now many proofs of the isoperimetric inequality. We will focus on the case where the boundary of the figure is a piecewise $\mathcal{C}^1$ curve in order to sweep the analytic issues about area and perimeter under the rug. Remember that being piecewise $\mathcal{C}^1$ means that the closed curve is continuously differentiable except at a finite number of points.

Theorem 8.28. (Isoperimetric Inequality) Let γ be a closed curve in the plane that is piecewise $\mathcal{C}^1$. Let $A=\mathcal{A}(\text{Int}(\gamma))$ be the area enclosed by the curve and $L=\mathcal{L}(\gamma)$ the perimeter. We have:

$$4\pi A \leq L^2 \tag{8.7.402}$$

with equality holding if and only if the curve is a circle (that's what lead us to claim that the circle is for a shape of a given perimeter, the shape having the biggest area). Indeed in the case where the curve is a circle, we have:

$$4\pi(\pi R^2)\leq(2\pi R)^2 \Rightarrow 1=1 \tag{8.7.403}$$

Hence the area is maximized when the shape is a circle (therefore the equality holds!). In other words, the isoperimetric inequality constrains how large the surface can be. You'll note that this inequality is maximized when the bounding perimeter is a circle!

In other words, any 2D (i.e. planer) surface that satisfies:

$$\frac{L^2}{A}=4\pi \tag{8.7.404}$$

maximizes the surface for a given perimeter!

Remark

The boundary curve need not be piecewise $\mathcal{C}^1$: with more analysis, less needs to be assumed about γ. However, this is more than sufficient to encompass common figures likes polygons and circles.

The proof we will introduce below is named the "Hurwitz's proof via Actions". For this to make sense, the curve must be piecewise $\mathcal{C}^1$ as we will assume throughout. Several useful facts are:

- If $\gamma(t)=(x(t),y(t))$ is a curve from $[a,b]\rightarrow\mathbb{R}^2$, its length can be calculated as we know by the relation (see section Differential Geometry 2856):

$$\mathcal{L}(\gamma)=\int_a^b\sqrt{\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2+\left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2}\,\mathrm{d}t \tag{8.7.405}$$

- If $\gamma(t)$ is a curve from $[0,\pi]\rightarrow\mathbb{R}^2$ (you can apply the same reasoning with the same result by choosing $[0,2\pi]\rightarrow\mathbb{R}^2$) with the polar coordinates of $\gamma(t)$ given by $r(t)$ and $\theta(t)$, the length of the curve is:

$$\mathcal{L}(\gamma)=\int\limits_0^{\pi}\sqrt{\left(\dfrac{\mathrm{d}r}{\mathrm{d}t}\right)^2+r^2\left(\dfrac{\mathrm{d}\theta}{\mathrm{d}t}\right)^2}\mathrm{d}t \tag{8.7.406}$$

This follows by making the change of variable $x=r\cos(\theta)$ and $y=r\sin(\theta)$ in the previous relation. Indeed, in polar coordinates we define the curve by the equation $r=f(t)$. In order to adapt the arc length formula for a polar curve, we use the equations:

$$x=r\cos(\theta)=f(t)\cos(\theta(t)) \tag{8.7.407}$$

$$y=r\sin(\theta)=f(t)\sin(\theta(t)) \tag{8.7.408}$$

Therefore:

$$\begin{aligned}\dfrac{\mathrm{d}x}{\mathrm{d}t}&=f'(t)\cos(\theta(t))-f(t)\sin(\theta(t))\dfrac{\mathrm{d}\theta}{\mathrm{d}t}\\ \dfrac{\mathrm{d}y}{\mathrm{d}t}&=f'(t)\sin(\theta(t))+f(t)\cos(\theta(t))\dfrac{\mathrm{d}\theta}{\mathrm{d}t}\end{aligned} \tag{8.7.409}$$

Hence:

$$\begin{aligned}L&=\int\limits_a^b\sqrt{\left(\dfrac{\mathrm{d}x}{\mathrm{d}t}\right)^2+\left(\dfrac{\mathrm{d}y}{\mathrm{d}t}\right)^2}\mathrm{d}t\\ &=\int\limits_a^b\sqrt{\left(f'(t)\cos(\theta(t))-f(t)\sin(\theta(t))\dfrac{\mathrm{d}\theta}{\mathrm{d}t}\right)^2+\left(f'(t)\sin(\theta(t))+f(t)\cos(\theta(t))\dfrac{\mathrm{d}\theta}{\mathrm{d}t}\right)^2}\mathrm{d}t\\ &=\int\limits_a^b\sqrt{(f'(t))^2\left(\cos^2(\theta(t))+\sin^2(\theta(t))\right)+f(t)^2\left(\dfrac{\mathrm{d}\theta}{\mathrm{d}t}\right)^2\left(\cos^2(\theta(t))+\sin^2(\theta(t))\right)}\mathrm{d}t\\ &=\int\limits_a^b\sqrt{(f'(t))^2+f(t)^2\left(\dfrac{\mathrm{d}\theta}{\mathrm{d}t}\right)^2}\mathrm{d}t=\int\limits_{\alpha}^{\beta}\sqrt{r^2+\left(\dfrac{\mathrm{d}r}{\mathrm{d}\theta}\right)^2}\mathrm{d}t=\int\limits_a^b\sqrt{\left(\dfrac{\mathrm{d}r}{\mathrm{d}t}\right)^2+r^2\left(\dfrac{\mathrm{d}\theta}{\mathrm{d}t}\right)^2}\mathrm{d}t\end{aligned} \tag{8.7.410}$$

- With the notation above, the area inside the curve is given by:

$$\mathcal{A}(\mathrm{Int}(\gamma))=\dfrac{1}{2}\int\limits_0^{\pi}r(t)^2\dfrac{\mathrm{d}\theta}{\mathrm{d}t}\mathrm{d}t \tag{8.7.411}$$

Indeed, similar to Cartesian equations, we can find the exact area of the polar region using an integral. The only exception is that we are using sectors to approximate the area, not rectangles, therefore:

$$A=\dfrac{\theta}{2\pi}\pi r^2 \tag{8.7.412}$$

Therefore:

$$\mathrm{d}A=\frac{\mathrm{d}\theta}{2\pi}\pi r(\theta)^2=\frac{1}{2}r(t)^2\mathrm{d}\theta=\frac{1}{2}r(t)^2\frac{\mathrm{d}\theta}{\mathrm{d}t}\mathrm{d}t \tag{8.7.413}$$

Hence as $\gamma(t)$ is a curve from $[0,\pi]\rightarrow\mathbb{R}^2$, we have:

$$A=\int\limits_0^\pi \frac{1}{2}r(t)^2\frac{\mathrm{d}\theta}{\mathrm{d}t}\mathrm{d}t \tag{8.7.414}$$

- The Cauchy-Schwarz inequality for integrals says that for square integrable functions f and g from $[0,\pi]\rightarrow\mathbb{R}$ we have (see section Functional Analysis 2388):

  $$\left(\int\limits_0^\pi f(t)g(t)\mathrm{d}t\right)^2\leq\int\limits_0^\pi f(t)^2\mathrm{d}t\int\limits_0^\pi g(t)^2\mathrm{d}t \tag{8.7.415}$$

  Equality is achieved if and only if g is a multiple of f.

The concept of action appears in physics and allows a reformulation of classical mechanics in terms of action and Lagrangians. For our purposes, all we need is the following definition of the action of a curve, which should be thought of as a physical property of a particle following that trajectory.

So let us remind that given $\gamma:[a,b]\rightarrow\mathbb{R}^2$ a plane curve. The action of the path γ is defined to be (see section Classical Mechanics page 3364):

$$\mathcal{E}(\gamma):=\frac{1}{2}\int_a^b\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2+\left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2\mathrm{d}t=\frac{1}{2}\int_a^b\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2+r^2\left(\frac{\mathrm{d}\theta}{\mathrm{d}t}\right)^2\mathrm{d}t \tag{8.7.416}$$

Unlike the length of the curve, the action of the curve depends on the parametrization of the curve. Therefore, we need to standardize how to parametrize curves.

The "standard parametrization" of a curve γ is the one for which the parameter's interval is $[0,\pi]$.

Since the definitions of action and length are similar, it is no surprise that they are related!

Lemma 8.28.1. Let γ be a curve with standard parametrization. Then:

$$\mathcal{L}(\gamma)^2\leq 2\pi\mathcal{E}(\gamma) \tag{8.7.417}$$

with equality achieved if and only if the parameter is a multiple of arc length along γ.

Proof 8.28.1. By Cauchy-Schwarz:

$$\left(\int_0^\pi f(t)g(t)\mathrm{d}t\right)^2\leq\int_0^\pi f(t)^2\mathrm{d}t\int_0^\pi g(t)^2\mathrm{d}t \tag{8.7.418}$$

Taking $f(t)=1$ and $g(t)=\sqrt{\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2+\left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2}$ gives the desired inequality. Equality is achieved if $g(t)$ is a constant, which says that the point $(x(t),y(t))$ moves along the curve with constant speed. Thus the parameter is a multiple of arc length.

The relevance to the isoperimetric inequality is that the action is also related to the area the curve encloses.

■ Q.E.D.

Proposition: Let γ be a simple closed curve with standard parametrization. Then:

$$2\mathcal{A}(\text{Int}(\gamma)) \leq \mathcal{E}(\gamma) \tag{8.7.419}$$

with equality achieved if and only if γ is a circle.

The proof uses the following analytic lemma:

Lemma 8.28.2. (Wirtinger Inequality) Let $f:[0,\pi]\rightarrow\mathbb{R}$ be a piecewise $\mathcal{C}^1$ function with $f(0)=f(\pi)=0$. Then we have:

$$\int\limits_0^\pi f'(t)^2\text{d}t \geq \int\limits_0^\pi f(t)^2\text{d}t \tag{8.7.420}$$

Equality is achieved if and only if $f(t)$ is a multiple of $\sin(t)$.

Proof 8.28.2. Consider the function $g(t)=\frac{f(t)}{\sin(t)}$. It is differentiable at the endpoints by l'Hopital's rule and the fact that $f(0)=f(\pi)=0$. Then:

$$f'(t)=g(t)\cos(t)+g'(t)\sin(t) \tag{8.7.421}$$

so:

$$\int\limits_0^\pi f'(t)^2\text{d}t=\int\limits_0^\pi \left(g(t)^2\cos(t)^2+2g(t)g'(t)\cos(t)\sin(t)+g'(t)^2\sin(t)^2\right)\text{d}t \tag{8.7.422}$$

Integrating the middle term by parts (see section Differential and Integral Calculus page 1532) given for recall by:

$$\int\limits_a^b x(t)y'(t)\text{d}t=x(t)y(t)\Big|_a^b-\int\limits_a^b x'(t)y(t)\text{d}t \tag{8.7.423}$$

(where we choose $y'(t)=g(t)g'(t)$ and $x(t)=\cos(t)\sin(t)$) and using the boundary conditions gives:

$$2\int\limits_0^\pi g(t)g'(t)\cos(t)\sin(t)\text{d}t=-\int\limits_0^\pi g(t)^2\left(\cos(t)^2-\sin(t)^2\right)\text{d}t \tag{8.7.424}$$

Indeed we have for the first term of the integration by part (we substitute $u=g(t)\longrightarrow \text{d}u=g'(t)\text{d}t$):

$$y(t)=\int g(t)g'(t)\text{d}t=\int u\text{d}u=\frac{u^2}{2}=\frac{f(t)^2}{2} \tag{8.7.425}$$

And for the second term of the integration by part:

$$\begin{aligned}x'(t)&=\frac{\text{d}(\cos(t)\sin(t))}{\text{d}t}=\frac{\text{d}\cos(t)}{\text{d}t}\sin(t)+\cos(t)\frac{\text{d}\sin(t)}{\text{d}t}\\&=(-\sin(t))\sin(t)+\cos(t)\cos(t)=\cos^2(t)-\sin^2(t)\end{aligned} \tag{8.7.426}$$

So coming back to the whole result, we get:

$$\int_0^\pi f'(t)^2 \mathrm{d}t = \int_0^\pi \left(g(t)^2 + g'(t)^2\right)\sin(t)^2 \mathrm{d}t = \int_0^\pi f(t)^2 \mathrm{d}t + \int_0^\pi g'(t)^2 \sin(t)^2 \mathrm{d}t \geq \int_0^\pi f(t)^2 \mathrm{d}t \tag{8.7.427}$$

Equality is achieved if and only if $g'(t) = 0$ on the interval, which implies $g(t)$ is constant and $f(t)$ is a multiple of $\sin(t)$.

■ Q.E.D.

We can now prove the proposition stated earlier above:

$$2\mathcal{A}(\mathrm{Int}(\gamma)) \leq \mathcal{E}(\gamma) \tag{8.7.428}$$

with equality achieved if and only if γ is a circle.

Proof 8.28.3. Remember that:

$$\mathcal{A}(\mathrm{Int}(\gamma)) = \frac{1}{2}\int_0^\pi r(t)^2 \frac{\mathrm{d}\theta}{\mathrm{d}t}\mathrm{d}t \tag{8.7.429}$$

and:

$$\mathcal{E}(\gamma) := \frac{1}{2}\int_a^b \left(\frac{dr}{dt}\right)^2 + r^2\left(\frac{d\theta}{dt}\right)^2 dt \tag{8.7.430}$$

Therefore:

$$\mathcal{E}(\gamma) - 2\cdot\mathcal{A}(\mathrm{Int}(\gamma)) = \frac{1}{2}\int_0^\pi \left(\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2 + r^2\left(\frac{\mathrm{d}\theta}{\mathrm{d}t}\right)^2 - 2r^2\frac{\mathrm{d}\theta}{\mathrm{d}t}\right)\mathrm{d}t \tag{8.7.431}$$

Completing the square turns the right side into:

$$\frac{1}{2}\int_0^\pi r^2\left(\frac{\mathrm{d}\theta}{\mathrm{d}t} - 1\right)^2 \mathrm{d}t + \frac{1}{2}\int_0^\pi \left(\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2 - r^2\right)\mathrm{d}t \tag{8.7.432}$$

The first term is non-negative since it is a square, and the second is non-negative by Wirtinger's inequality as this term can be rewritten:

$$\int_0^\pi f'(t)^2 dt \geq \int_0^\pi f(t)^2 \mathrm{d}t \Rightarrow \int_0^\pi \left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2 - \int_0^\pi r^2 \mathrm{d}t \geq 0 \tag{8.7.433}$$

So we have indeed:

$$\mathcal{E}(\gamma) - 2\cdot\mathcal{A}(\mathrm{Int}(\gamma)) = \frac{1}{2}\int_0^\pi \left(\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2 + r^2\left(\frac{\mathrm{d}\theta}{\mathrm{d}t}\right)^2 - 2r^2\frac{\mathrm{d}\theta}{\mathrm{d}t}\right)\mathrm{d}t \geq 0 \tag{8.7.434}$$

So the proposition:

$$2\mathcal{A}(\mathrm{Int}(\gamma)) \leq \mathcal{E}(\gamma) \tag{8.7.435}$$

is proven!

The curve is closed, so we may pick the origin of the polar coordinate system to be at $\gamma(0)=\gamma(\pi)$.

To have equality, i.e.:

$$2\mathcal{A}(\mathrm{Int}(\gamma))=\mathcal{E}(\gamma) \Rightarrow \frac{1}{2}\int_0^{\pi} r^2\left(\frac{\mathrm{d}\theta}{\mathrm{d}t}-1\right)^2 \mathrm{d}t+\frac{1}{2}\int_0^{\pi}\left(\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2-r^2\right)\mathrm{d}t=0 \tag{8.7.436}$$

So we see that if we pick up:

$$\frac{\mathrm{d}\theta}{\mathrm{d}t}=1 \qquad \text{and} \qquad r=a\sin(\theta+\varphi) \tag{8.7.437}$$

We get the equality and $r=a\sin(\theta+\varphi)$ is the polar equation of a circle!

■ Q.E.D.

The isoperimetric equality now follows easily.

Proof 8.28.4. Pick a standard parametrization for γ. By the two propositions, we have therefore:

$$\left.\begin{array}{l} 2\mathcal{A}(\mathrm{Int}(\gamma)) \leq \mathcal{E}(\gamma) \\ \mathcal{L}(\gamma)^2 \leq 2\pi\mathcal{E}(\gamma) \end{array}\right\} \Rightarrow 4\pi\mathcal{A}(\mathrm{Int}(\gamma))=\mathcal{L}(\gamma)^2 \tag{8.7.438}$$

What is often denoted in a vulgar way $4\pi A=L^2$. The latter result can be written explicitly:

$$\boxed{\begin{gathered} 4\pi\mathcal{A}(\mathrm{Int}(\gamma)) \leq \mathcal{L}(\gamma)^2 \\ \Downarrow \\ 4\pi\frac{1}{2}\int_0^{\pi} r(t)^2\frac{\mathrm{d}\theta}{\mathrm{d}t}\,\mathrm{d}t \leq \left(\frac{1}{2}\int\limits_a^b \left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2+r^2\left(\frac{\mathrm{d}\theta}{\mathrm{d}t}\right)^2\mathrm{d}t\right)^2 \end{gathered}} \tag{8.7.439}$$

with equality achieved if and only if the curve is a circle!

■ Q.E.D.

Perhaps the most familiar physical manifestation of the three-dimensional isoperimetric inequality is the shape of a drop of water. Namely, a drop will typically assume a symmetric round shape. Since the amount of water in a drop is fixed, surface tension forces the drop into a shape which minimizes the surface area of the drop, namely a round sphere (this type of experiment has been reproduced many times in the International Space Station).

~80 % ★★★★☆

171 votes, 74.86%

8.8. Graph Theory

THE history of graph theory (or also named "complex cellulars") may have start with the work of Euler in the 118th century (holocene calendar) and has its origins in the study of certain problems, such as the bridges of Königsberg (the inhabitants of Königsberg wondered if it was possible, starting from any part of the city, to cross all the bridges without passing two times by the same bridge and return to their starting point), the walk of the rider on the chessboard cards or the colouring problem and the shortest path between two points.

Graph theory was then developed in various disciplines such as chemistry (isomers), biology, social sciences (transport networks, social networks), project management (critical path method), IT (network topology, computational complexity, protocols transfers), quantum physics, spectral clustering, etc. Since the early 120th century (holocene calendar), it is a full branch of mathematics, thanks to the works of König, Menger, Cayley and then Berge and Erdös. This branch of mathematics has get a great resurgence of interest following the emergence of Internet social networks whose connections between "friends" and "followers" are graphs whose topological and statistical analysis can teach us many things.

In general, a graph is used to represent the structure, the connections of a complex set by expressing the links between its components: communication network, road networks, interaction of various animal species, electrical networks, etc. The graphs are therefore a thought method to model a wide variety of problems by reducing them to the study of vertices and edges.

Recent work in graph theory are often done by computer, because of the importance of the algorithmic aspect in their field (see the beginning of the section of Theoretical Computing some small examples).

Indeed, the purpose is essentially to modelize problems. We express the problem in terms of graphs so that it reports to a problem of graph theory that we know usually how to resolve because falling within a class of known problems.

The solutions to graph problems can be easy and effective (the time required to process computationally being reasonable because they depend polynomially on the number of vertices of the graph) or difficult (the processing time is then exponential) in which case we use a heuristic, that is to say a search process for a solution (not necessarily the best).

If graph theory knows a quite big interests since the years 11980 (holocene calendar), maybe is it because it does not require in its elementary concepts considerable mathematical background. Actually, just going through the sections of Probability, Set Theory, Linear Algebra and Topology presented in this textbook to already feel comfortable with the different definitions.

We will introduce the basic vocabulary of the graph theory. The terms used are those of the common language of Euclidean geometry (and unfortunately they are also in large numbers...).

8.8.1 Type of Graphs and Structures

Definitions (#8.8.397):

D1. A "graph" (or "polygraph") G is a pair $G(X, E)$ consisting of a non-empty and finite set X (the vertices/nodes), and a set E (the edges/links) of X of ordered pairs elements of X connected by a line segment:

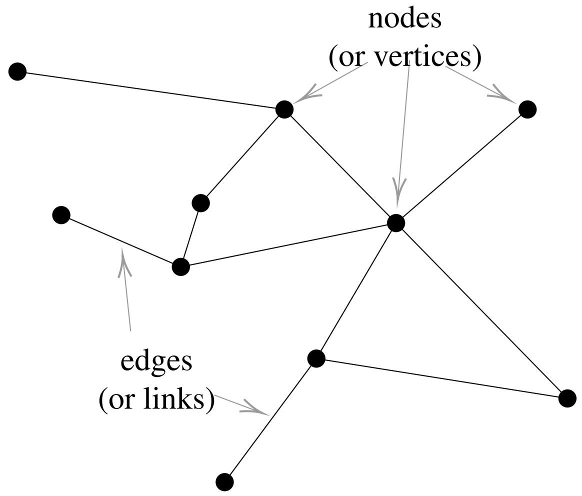

nodes
(or vertices)
edges
(or links)

Figure 8.296: **Graphs basic vocabulary**

or said in another way (...) a graph is a part of the Cartesian product X^2 (see section Set Theory page 391):

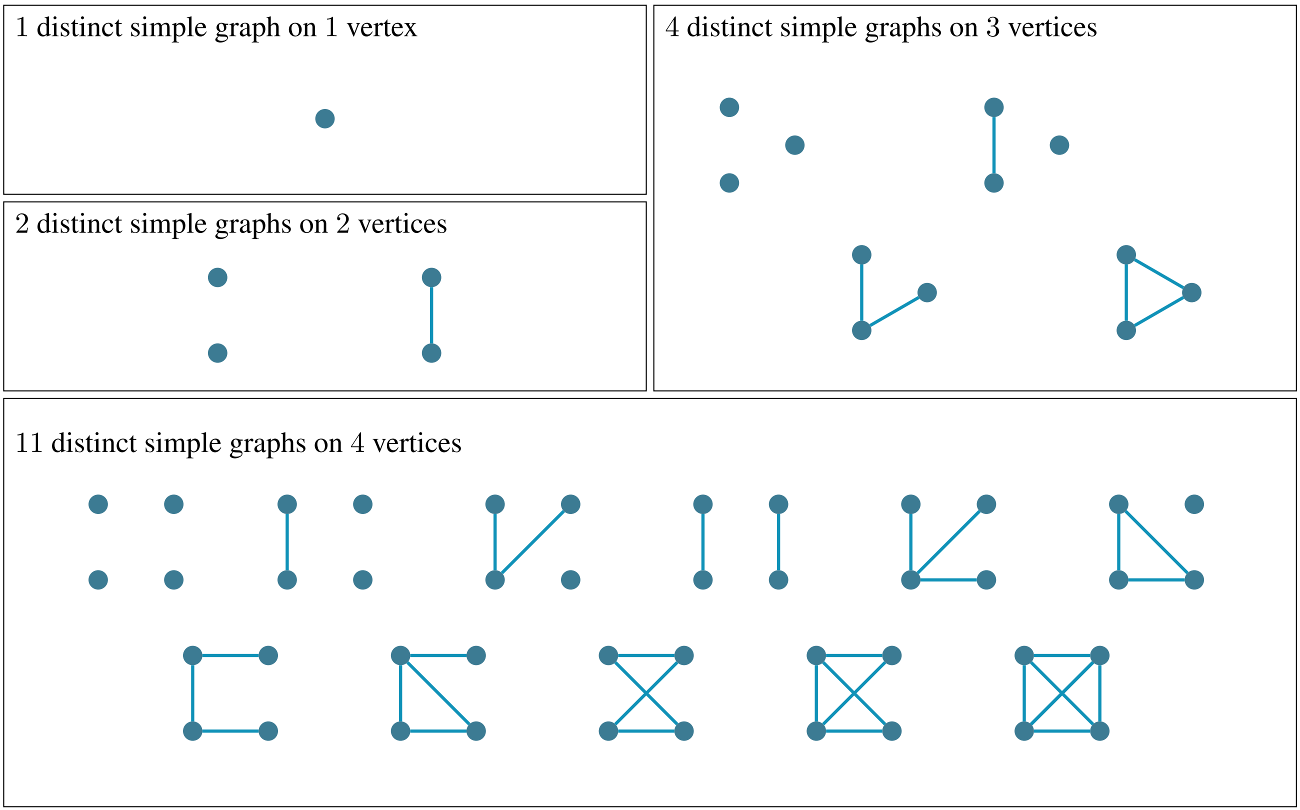

1 distinct simple graph on 1 vertex
4 distinct simple graphs on 3 vertices
2 distinct simple graphs on 2 vertices
11 distinct simple graphs on 4 vertices

Figure 8.297: **Simple Graphs**

Remark

A graph is often denoted in English $G=(E,V)$ where E is the first letter of the word "edges" and V the first letter of the word "vertices".

D2. Let (a_1,b_1) and (a_2,b_2) be pairs. Then the characteristic (or defining) property of the "ordered pair" is:

$$(a_1,b_1)=(a_2,b_2) \quad \text{if and only if} \quad a_1=a_2 \wedge b_1=b_2 \tag{8.8.1}$$

The order in which the objects appear in the pair is significant: the ordered pair (a,b) is different from the ordered pair (b,a) unless $a=b$.

In graph theory it is useful to precise if a graph G is made of ordered pairs in this way we can work if necessary with edges orientation to distinguish edges sharing the same pairs of vertices but having not the same orientation path.

D3. A "multigraph" is a graph G which is permitted to have edges sharing the same vertices and having edges looping on the same vertices.

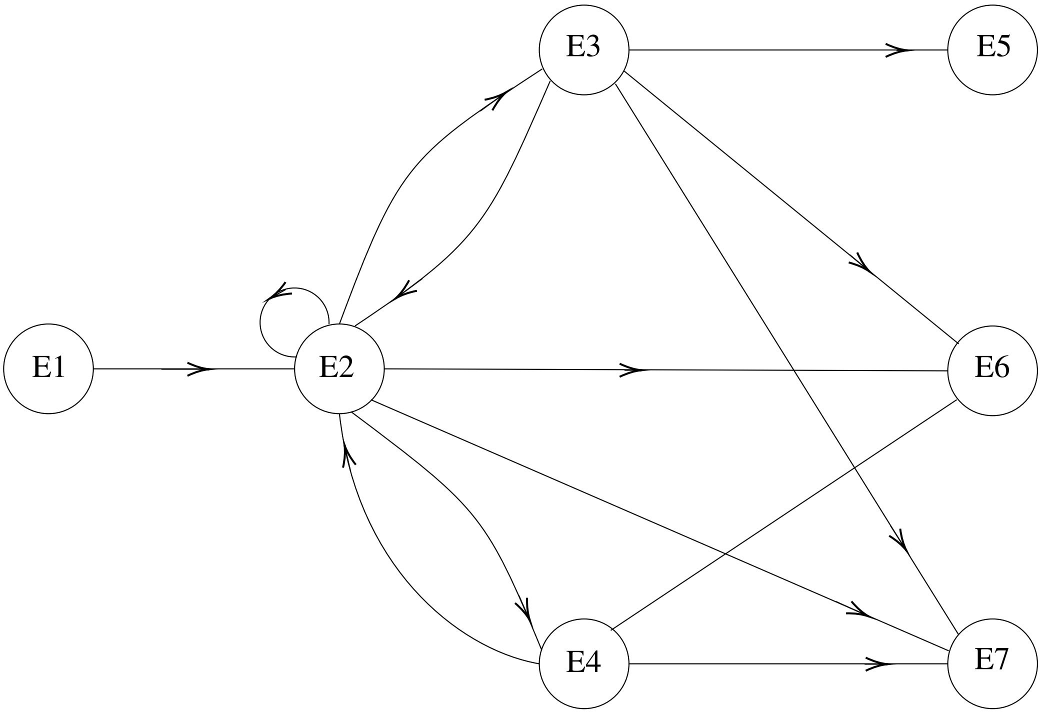


Figure 8.298: Example of a multigraph that we will use much later

Remark

This a very common case in Markov chains analysis (see sections on Probabilities page 537 or Games and Decision Theory page 7456) and as we will see further below the Königsberg bridge problem is a multigraph!

D4. The elements of X are the "vertices" of the graph G, those of E are the "edges" of the graph G (indeed an edge - link - is composed of two vertices joined by a line segment, hence the reference to pairs of elements in the previous definition). The set of vertices of a graph G is denoted by $G(X)$ and the set of edges $G(E)$.

D5. The plane where a graph is immersed is a "face F" and each close surface (area) on this plane defined by a close path of edges is also a face F.

D6. A graph is named "planar graph" when we can represent it in a plane without intersecting edges.

The following are some examples of (connected) planar graphs:

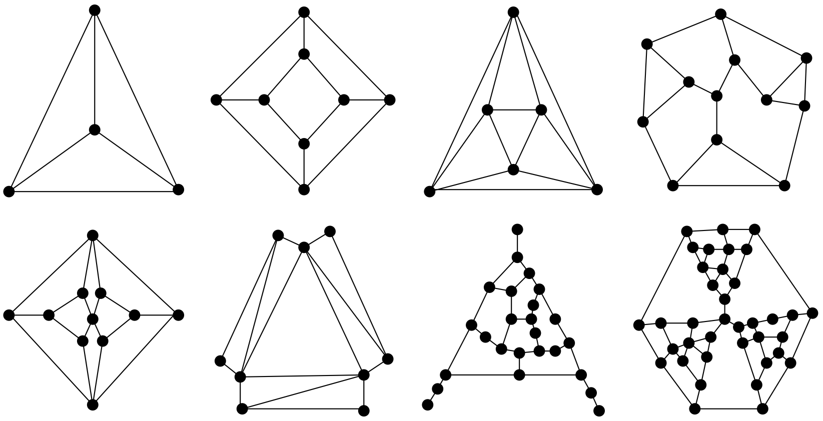


Figure 8.299: **Example of planar graphs**

Theorem 8.29. Now let us show that if F is the number of faces of a planar graph, A the number of edges and S the number of nodes, then we have using a old traditional notation:

$$\boxed{F-A+S=2} \tag{8.8.2}$$

which is the relation known under the name "Euler's formula" or "Descartes-Euler's theorem" (proof after the example below) and which will be useful many times in this textbook (in this section and during our study of polyhedra in the section of Geometrical Shapes). So keep in mind that any graph that doesn't respect this relation cannot be a planar graph!

✍**Example:**

A graph with 2 faces (the light gray side is an infinite outside face), 4 vertices and 4 nodes:

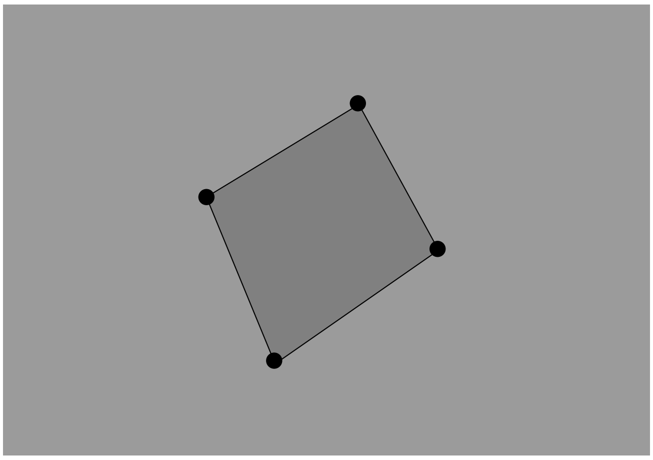


Figure 8.300: **Graph with 2 faces, 4 vertices and 4 nodes**

Proof 8.29.1. We prove this theorem by performing a recurrence (see section Proof Theory page 171) on the difference $A-S$ (this is the trick!).

First it is obvious that the theorem is true for:

$$A-S=-1 \tag{8.8.3}$$

because in this case the graph is a tree so it only has one face (the outer infinite face), thus $F=1$.

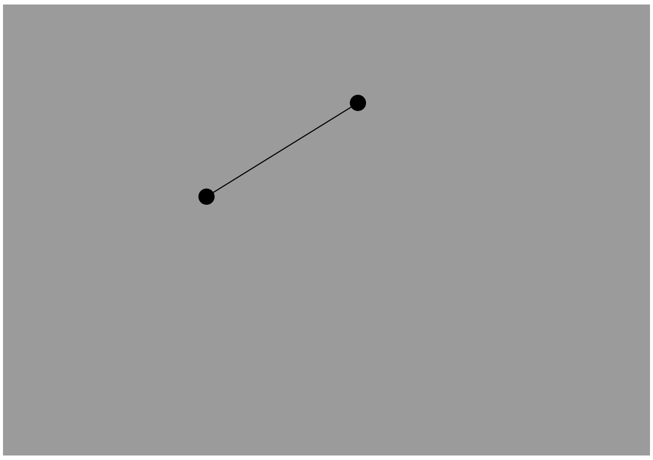


Figure 8.301: **Tree with one edge and two vertices**

Therefore:

$$F-A+S=F-(A-S)=1-(-1)=2 \tag{8.8.4}$$

Then let us take a connected graph (see definition further below) containing at least one cycle G (figure below is an example of a graph with 3 cycles that is to say the two small insiders and the global one):

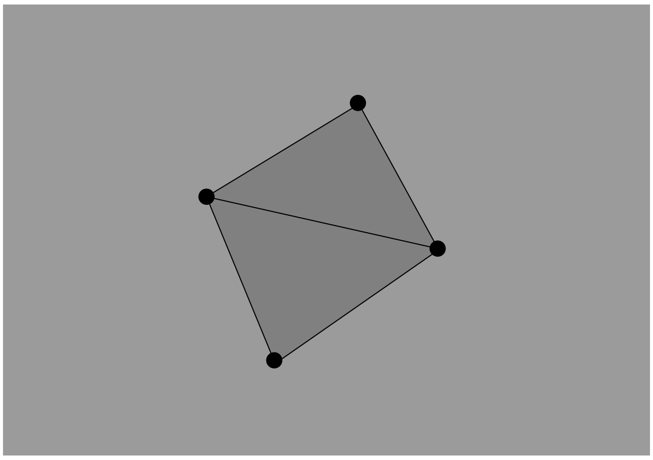


Figure 8.302: **Graph with at least one cycle**

If we remove one edge e to this cycle, then we should be able recursively to apply the same relation to the graph $G-\{e\}$ if it is right. Indeed, the graph without the edge e will have F faces, S vertices and A edges and thus we can use the relation proved above with the simple tree (that had no cycles):

$$F - A + S = 2 \tag{8.8.5}$$

if we give put the edge back on place then we write:

$$F - A + S = 2 \tag{8.8.6}$$

$$(F+1) - (A+1) + S = F - A + S = 2 \tag{8.8.7}$$

■ Q.E.D.

Regular polygons (see section Geometric Shapes page 2893) with their diagonals drawn in can be viewed as graphs (with a vertex at each intersection), more precisely as "complete graphs" (see the definition at page 3010), as illustrated below:

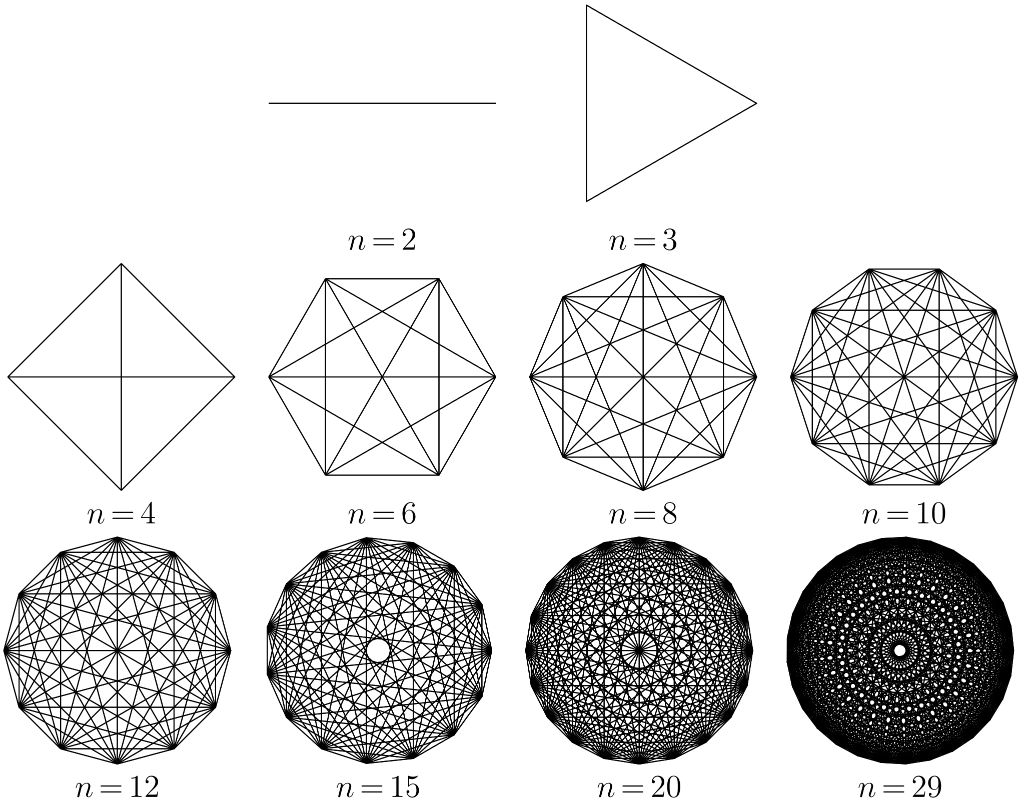
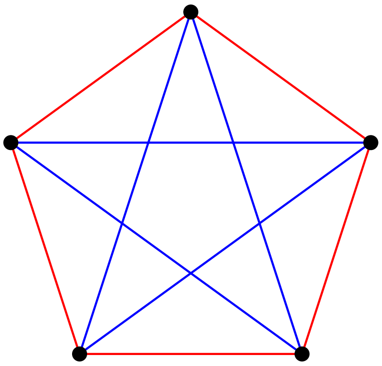


Figure 8.303: Regular polygons seen as graphs

If we make every intersection point a vertex, then these pictures automatically become connected, planar graphs! If we know the number of intersection points and the number of segments, we can then use Euler's formula to find the number of regions (for example, drawing the two diagonals of a square results in 5 intersections, 4 regions, and 8 segments. The 5 diagonals of a regular pentagon make 10 intersections, 11 regions, and 20 segments).

✍**Example:**

Let us consider K_5:



We take the Euler's formula:

$$F-A+S=2 \tag{8.8.8}$$

K_5 has 5 vertices and 10 edges, so we get:

$$5-10+S=2 \tag{8.8.9}$$

which says that if the graph is drawn without any edges crossing, there would be $S=7$ faces. And there is no way to find a way to do that with K_5.

D7. Given $e=x,y$ an edge of the graph G, we say that the vertices x,y which are the "ends" of the edge of G are "adjacent" or "neighbours" in the graph G and that the edge e is "incident" to the vertices x,y.

D8. If two edges e and e' have one end in common, we say they are "incidental edges" otherwise that they are "independent edges".

Remark

If e is an edge of a graph G, we denote by $G-e$ the subgraph (see definition just below) $G'=(X,E-\{e\})$. If X' is a subset of X, we denote by $G-X'$ the graph G without the vertices X'.

D9. What we name "order of a graph $\mathcal{O}$" is the number of its vertices (nodes).

D10. What we name "size of a graph" of the graph is the number of its edges.

Let G be a complete graph (and not a multigraph!) of order n, the set E of edges must be by definition chosen as a subset of all pairs of elements of the set X, therefore it's a set of cardinal:

$$\boxed{\text{Card}(E)=C_2^n=\frac{n(n-1)}{2}} \tag{8.8.10}$$

This is a relatively trivial result since each vertices is linked to all the other vertices except himself (hence the numerator) and we divide by two simply to not count the neighbours vertices twice (and they all are neighbours when we go through them all in a complete graph!).

More explicitly (if needed) $n(n-1)/2$ comes from simple counting argument. Label the vertices $1,2,\ldots,n$. The first vertex is now joined to $n-1$ other vertices. The second vertex has already been joined to vertex 1 and hence has to be joined to the remaining $n-2$ vertices and in general the kth vertex has already been joined to the previous $k-1$ vertices and hence has to be joined to the remaining $n-k$ vertices. So the total number of edges is given by:

$$(n-1)+(n-2)+\ldots+2+1=\frac{n(n-1)}{2} \tag{8.8.11}$$

This result is illustrated by some examples further below.

Consequently, there are (see section Probabilities on the arrangements of n indistinguishable elements by pairs of two 523):

$$2^{\frac{n(n-1)}{2}} \tag{8.8.12}$$

possible choices for E and so many graphs admitting X as set of vertices.

Remark

The value obtained with $n(n-1)/2$ must not be confused with the quite famous "six degrees of separation" theory popularized by Frigyes Karinthy in 11929 and later tested in experiments like Stanley Milgram's in the 11960s (modern studies using digital communication networks like Facebook, Twitter, LinkedIn found the average path length is often between 4 and 6 degrees). This concepts suggesting that 6 links of friends are enough to cover the whole actual World population of 8.142 billions individuals. Indeed, if you know 45 friends and each of these 45 friends know 45 mutually exclusive friends and so on six times, you get $45^6=8,303,765,625$.

✍**Example:**

We take a triangular graph - with three points - as example where we have $n(n-1)/2=3$ and therefore $2^{\frac{n(n-1)}{2}}=8$:

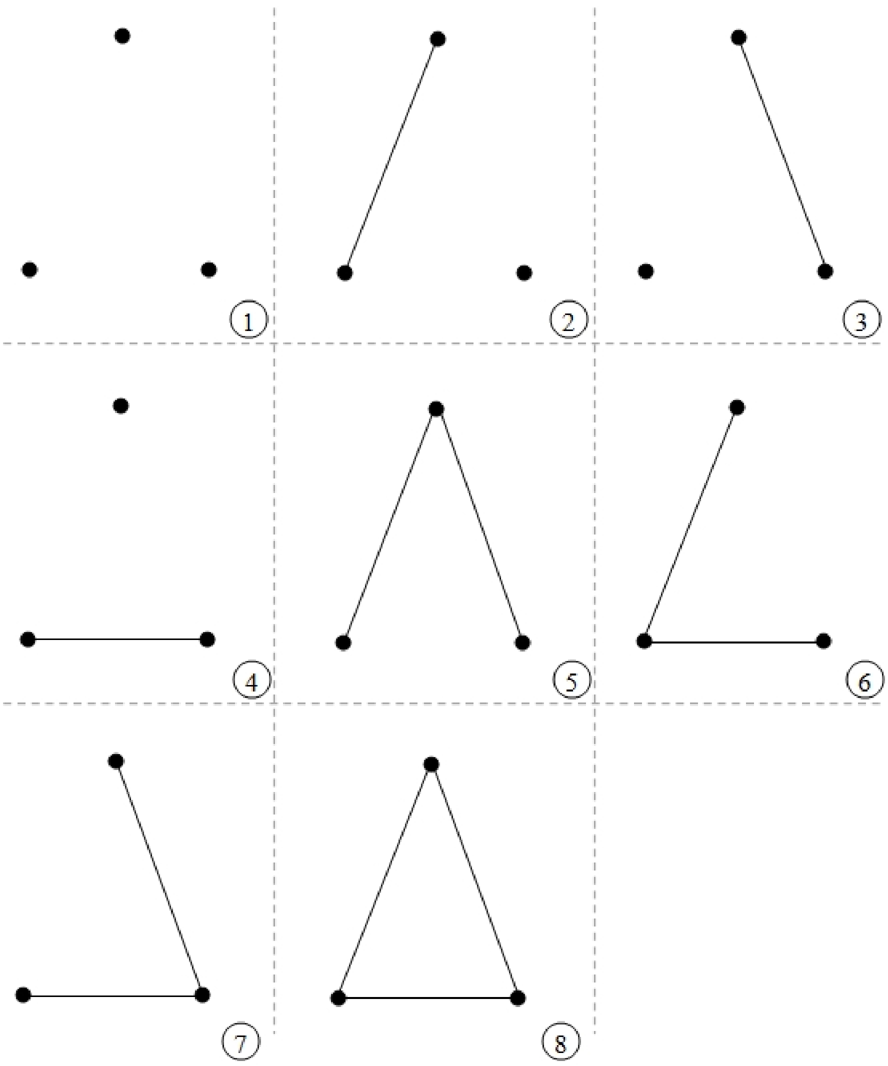


Figure 8.304: **Example of graphs subsets**

Some of these graphs are, as we consider their vertices as indistinguishable, "automorphic" (see definition of this term a little bit late below in this section).

This result means that there are about 2 million of possible graphs for 7 vertices and almost $4\cdot10^{105}$ possible graphs for 27 vertices - a number to be compared with the fact that we estimate at 10^{100} the number of atoms in the observable universe...

D11. The "neighbourhood" of a vertices is the set V of all its neighbours.

D12. We name "degree" of a vertices and we denote by D the number of its neighbours (the cardinal of V), which is also the number of edges that are incident to it (zero degrees vertices being named "isolated vertices" and degree 1 vertices being named "pending vertices").

Properties (without proof):

P1. The sum of the degrees of the vertices is equal to twice the number of edges:

$$\sum_{v\in V}\deg(v)=2\text{Card}(E) \tag{8.8.13}$$

P2. In a graph such that $\mathcal{O}>3$, the number of vertices with odd degrees is always even (handshaking lemma).

Indeed, if the sum of all the degrees is equal to twice the number of edges. Since the sum of the degrees is even and the sum of the degrees of vertices with even degree is even, the sum of the degrees of vertices with odd degree must be even. If the sum of the degrees of vertices with odd degree is even, there must be an even number of those vertices.

Remark

A "regular graph" is a graph in which all vertices have same degree k. We say then that the graph is "k-regular".

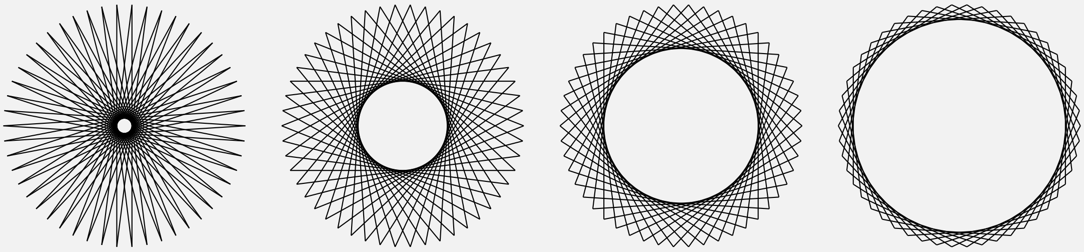


Figure 8.305: A few possible 2-regular graphs

D13. We say that a graph $G'=(X',E')$ is a "subgraph" of a graph $G=(X,E)$ when $X'\subseteq X$ and $E'\subseteq E$. We say that $G'=(X',E')$ is an "induced subgraph" of a graph $G=(X,E)$ provided $X'\subseteq X$ and E' contains all edges of E which are subsets of X'.

If G' is a subgraph of G and u and w are vertices of G', then by the definition of a subgraph, u and w are also vertices of G. However, if u and w are adjacent in G (i.e., there is an edge of G joining them), the definition of subgraph does not require that the edge joining them in G is also an edge of G'. If the subgraph g' has the property that whenever two of its vertices are joined by an edge in G, this edge is also in G', then we say that G' is an "induced subgraph".

Notice that every induced subgraph is also an ordinary subgraph, but not conversely. Think of a subgraph as the result of deleting some vertices and edges from the larger graph. For the subgraph to be an induced subgraph, we can still delete vertices, but now we only delete those edges that included the deleted vertices.

✍**Example:**

Consider the graphs:

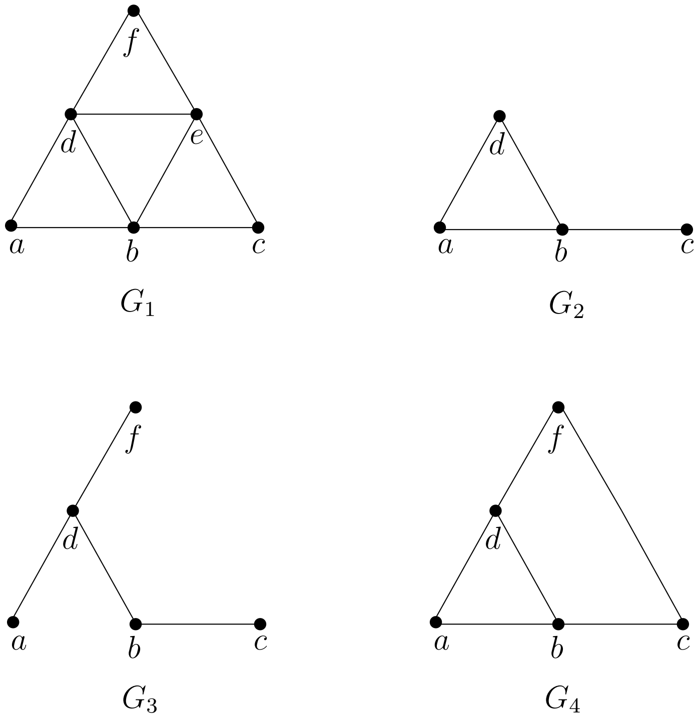


Figure 8.306: **Examples of subgraphs and induced graphs**

Here both $G_2(X_2, E_2)$ and $G_3(X_2, E_3)$ are subgraphs of $G_1(X_1, E_1)$. But only G_2 is an induced subgraph. Every edge in G_1 that connects vertices in G_2 is also an edge in G_2. In G_3, the edge $\{a, b\}$ is in E_1 but not E_3, even though vertices a and b are in X_3.

The graph $G_4(X_4, E_4)$ is not a subgraph of G_1, even though it looks like all we did is remove vertex e. The reason is that in E_4 we have the edge $\{c, f\}$ but this is not an element of E_1, so we don't have the required $E_4 \subseteq E_1$.

D14. A "covering subgraph" of a graph $G = (X, E)$ is a subgraph $G' = (X, E')$, that is to say a subgraph whose vertices are all vertices of G and whose edges are in E'.

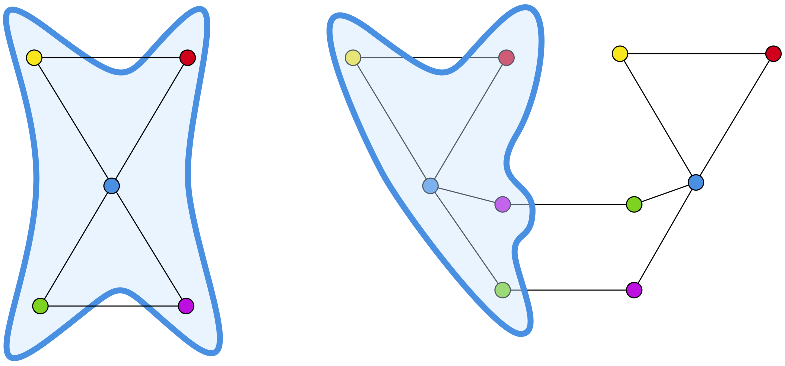


Figure 8.307: In the above figure, the graph on the right is a covering graph of the graph on the left

D15. For a graph of order n, there are two extreme cases for all its edges: either the graph has no edges or all possible edges connecting the vertices by pairs are present. In the latter case the graph is said named a "complete graph" or "strongly connex graph".

✍**Example:**

Here are some complete graphs for which we have well:

$$\frac{n(n-1)}{2} \tag{8.8.14}$$

edges. We notice that the first four graphs are planar (indeed you can see how it is possible, by projecting a vertices in the plane, to transform the fourth K_4 in K_4' in a way so that there are no more intersections). The fifth graph K_5 is non-planar (we can not find movements avoiding crossings):

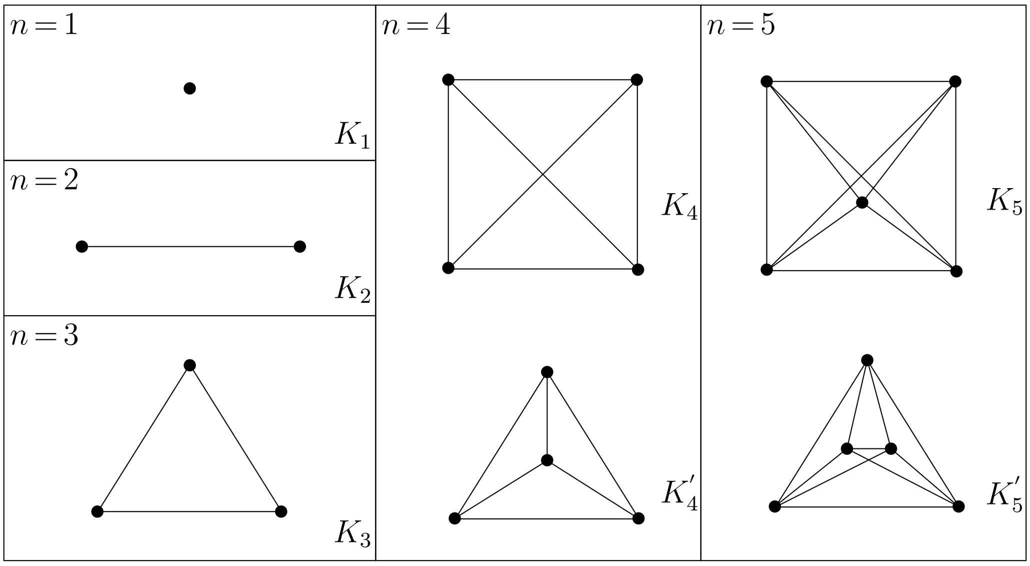


Figure 8.308: Example of complete graphs

A complete graph is therefore a graph where each vertex is connected to every other. The complete graph of order n is denoted by K_n. In this type of graph each vertex is of degree $n-1$.

A nice case to be processed is the "Star of David":

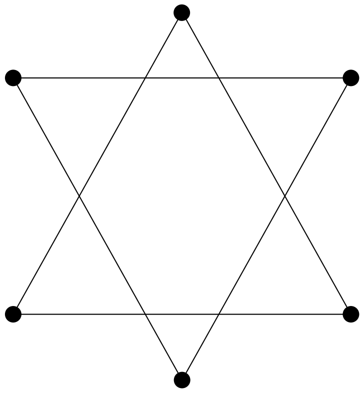


Figure 8.309: **Star of David**

which is obviously a complete graph by definition if and only if we join all vertices between them (and we lose the geometry of the star but we get a graph K_6):

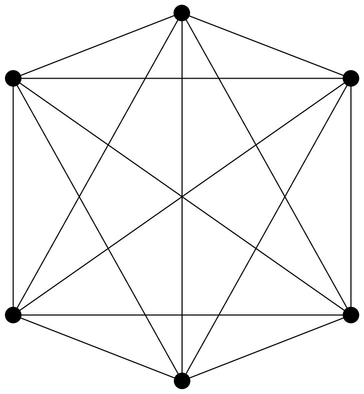


Figure 8.310: **Star of David completed**

Remark

This result is interesting in the field of management of the communication in business projects and IT security (see for the latter the section Cryptography at page 7216). For example, if you manage a project with 10 stakeholders (corresponding to n), so there are $n(n-1)/2$ possible communication channels (email or phone) possible, that is to says 45 (and in the field of IT security there would be 45 encryption symmetric keys system to generate). Hence the importance in project management to establish clear communication rules (in the form of a graph) if one does not want to be flooded by emails or phones unnecessarily (and in the field of security to implement asymmetric systems). We will also encounter this results in the liquid drop model of the nucleus in the Nuclear Physics section of the chapter Atomistic (see page 5141).

D16. A "stable graph" or "independent set" is subgraph without edges and a "clique" a complete subgraph.

In other words an independent vertex set of a graph G is a subset of the vertices such that no two vertices in the subset represent an edge of G.

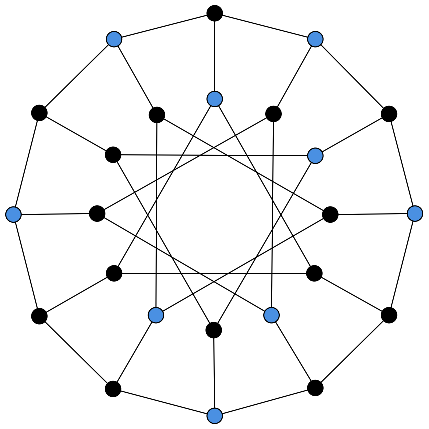


Figure 8.311: **Stable subgraph example (in blue)**

Remark

A "maximum independent" set is an independent set of largest possible size for a given graph G. The problem of finding such a set is a NP-hard optimization problem (see section Numerical Methods page 6291). As such, it is unlikely that there exists an efficient algorithm for finding a maximum independent set of a graph.

D17. In a graph, it is natural to want to move from summit to summit following the edges. Such a walk through n vertices is named a "string P_n" or "path":

A path is a list $P_k=(x_1,\ldots,x_k)$ of vertices such that there exists an edge in the graph between every pair of successive vertices: $\forall i=1\ldots k-1\;(x_i,x_{i+1})\in E$. The path length corresponds to the number of edges traversed: $k-1$.

A path is named "simple path" if each edge of the path is taken only once. Here for example a simple path with 5 vertices:

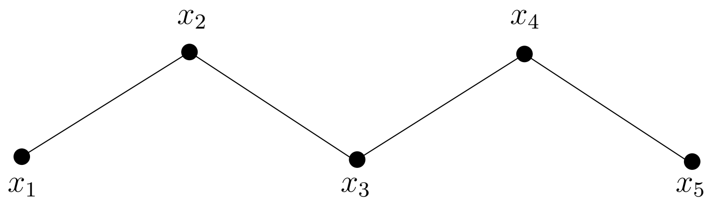


Figure 8.312: **Simple path example**

Thus, we also define a "cycle" as:

$$C_n=C_{k+1}=(x_1,\ldots,x_k,x_{k+1}) \tag{8.8.15}$$

as being a simple path ending at the starting point as $x_1=x_{k+1}$.

D18. A "simple cycle" is a cycle in which all the edges are different.

D19. A "directed graph" or "oriented graph" is a graph whose edges have a direction and re therefore named "arcs" (thus opposite to an undirected graph).

Remarks

R1. The terms "path" and "circuit" are normally reserved to directed graphs. For undirected graphs we use mainly the words "string" or "cycle". However, the formal definition is exactly the same in both cases, only changes the structure (directed graph or not) on which they are defined.

R2. An undirected graph is a symmetrical directed graph. Indeed, if an arc connects the vertice a to the vertice b top and another arc the vertice b to the vertice a, then it is of usage to trace only a simple line between a and b which we name ... an "edge".

D20. A path $P_k=(x_1,\ldots,x_k)$ is named "elementary path" if each of the vertices of the path is visited only once: $\forall i,j=1\ldots k \quad i\neq j, x_i\neq x_j$. An elementary path is therefore a simple path without cycle.

In a graph of order G we have the following properties:

P1. All elementary path is of length at most $n-1$. Indeed, an elementary path visiting at most 1 time each vertex of the graph, its length (number of edges) may actually not exceed $n-1$.

P2. The number of elementary paths in a graph is finite. Indeed, the number of paths of length $k(k=0,1,\ldots,n-1)$ is at most the combinatorial of a sequence of $k+1$ peaks distinguishable from n. There are therefore (see section Probabilities page 527):

$$A_n^p=\frac{n!}{(n-p)!} \tag{8.8.16}$$

possible paths.

The elementary paths are the natural restrictions we are looking at the concept of path. The question is whether we lose something by considering only the basic paths in a graph: can we always replace a path in a graph by an elementary path?

Lemma 8.29.1. The "König's lemma" answers this question affirmatively: from any path we can extract an elementary sub-path.

In other words: If there is a path between two vertices x and y, then there exists an elementary path between x and y.

Proof 8.29.2. The idea of the proof is to choose a particular path between x and y and show that it is elementary. Which path to choose? If a path has a circuit, this circuit is a detour on the road of x and y. A good candidate to be an elementary path appears to be a shortest path.

Among all the paths connecting x to y, let us choose a path:

$$P_k=(x_1=x,\ldots,x_k=y) \tag{8.8.17}$$

with the fewest possible edges. Suppose by contradiction that P_k is not elementary. There exists in this path a vertices z appearing at least two times along the path P_k.

Given i,j the first two indices such as $x_i=z$ and $x_j=z$:

$$P_k=(x_1=x,\ldots,x_i=z,\ldots,x_j=z,\ldots,x_k=y) \quad (8.8.18)$$

To get a contradiction, we just delete the cycle between $x_i=z$ and $x_j=z$. So we have a new path:

$$P'_k=(x_1=x,\ldots,x_{i_1},\ldots,x_{j+1},\ldots,x_k=y) \quad (8.8.19)$$

which is a path connecting x to y. Its length strictly less than P_k:

$$\text{Card}(P'_k)<\text{Card}(P'_k) \quad (8.8.20)$$

which contradicts our initial choice as a shortest path.

■ Q.E.D.

D21. A graph G is named "connected graph" or "connex graph" if and only if there is at least one simple path between each pair of vertices (the path being therefore not necessarily direct- can pass through one or more intermediate vertices).

Therefore, we have that a "connected component" $G'=(X',E')$ is a subgraph of G in which any two vertices are connected to each other by paths, and which is connected to no additional vertices in the supergraph.

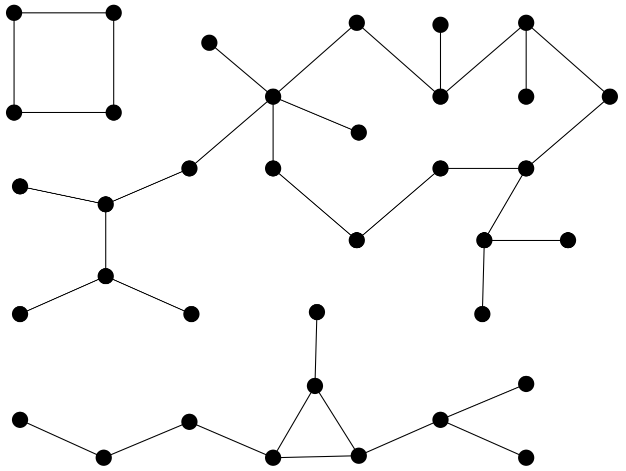


Figure 8.313: **Various connected graphs**

Remarks

R1. A graph with only one connected component is simply a... connected graph.

R2. An isolated node (of degree 0) always constitutes a connected component alone.

R3. The relation on the nodes "there is a path between ..." is an equivalence relation (reflexive, symmetric and transitive). The connected components of a graph correspond to equivalence classes of this relation.

We will suppose intuitive that a graph G of order n is always associated with at least $n-1$ edges.

D22. A "tree" or "spanning tree" is a connected graph (undirected) with no cycles (acyclic). In a tree the number of edges is equal to the number of vertices -1:

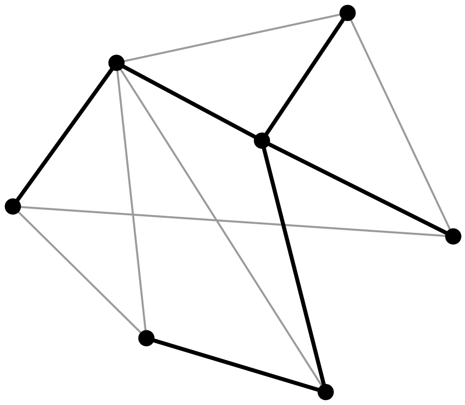


Figure 8.314: Example of a spanning tree

A minimum spanning tree with the minimum weight values of a valuated graph where all nodes are browsed is named a "minimum spanning tree":

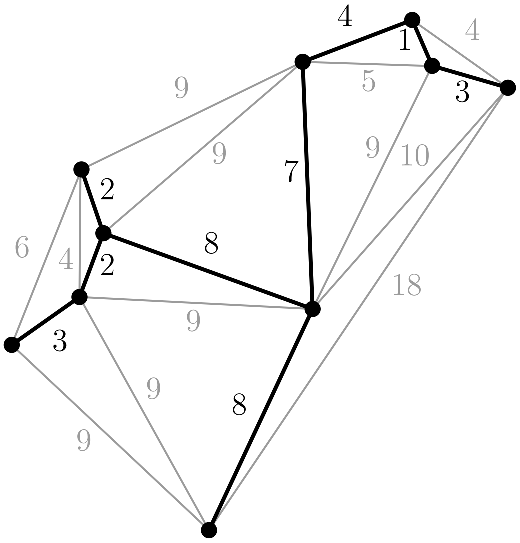


Figure 8.315: Example of a minimum spanning tree (source: Wikipedia)

If you imagine that every point is a city, and that a country has not enough money for now for the maintenance of all existing routes between several cities, the minimum spanning tree gives mathematically (without any other considerations) the road to clean and repair to minimize costs while connecting all the towns together.

D23. A "valuated tree" or "weighted graph" is a tree (respectively a graph) where the edges have positive values (weights). The sum of all the values that are on the browsed edges is then named the "cost tree Value" (respectively "cost graph value").

Remark

The valued trees are used in many fields. These include computer networks in which we seek to optimize the number of interconnections between machines to avoid duplication of items of data packets or project management (see example below).

✍**Example:**

An excellent practical example of a connected weighted and directed graph and directed (abbreviated to as "digraph") is the one used in project management to calculate the critical path (of a project without constraints). This is a graph representing the dependencies between n intermediate tasks required to complete a project, commonly named "Gantt chart" or even "scheduling graph". The length (weight) of each task is the value of the externally incident weights to the correspondent node arcs. Edges represent the sequences of tasks. We are always adding an initial and final node (in the world of project management people rather talking about "milestone"...). The first node is connected by a zero value arc to all nodes without predecessors, and all nodes without successors are connected to the final node. The resulting graph should obviously be acyclic.

A "critical path" is a path of maximum length between the two extreme nodes (milestones). It may have possibly be more than one of the same length. Any task located on a critical path can not be delayed without affecting the total duration of a project. In other words, it has zero "total slack" (we then also say so that its early start finish date is strictly equal to late finish date). Furthermore, we also define in project management the "free slack" that indicates the time over which a task can slide without moving/impacting the successor task. The free slack is calculated as the difference between the date of earliest start of a task summed of its duration and the late start of the successor task. By construction the free slack is always small or equal tot the total slack.

Let us take for example a project that consists of the following tasks:

Tasks	Predecessor Tasks	Duration
A	E	3
B	K,C	4
C	-	3
E	E,J	2
E	-	2
F	G,L	3
G	-	4
H	A, M, R	2
J	E	2
K	C	2
L	G	5
M	C	4
N	G	3
R	J	2

Table 8.8: **Data for critical path example**

The valued associated directed graph corresponding to this table once the definition of the critical path applied is the following using the start dates (this method of calculation is named "Potential Metra Method"):

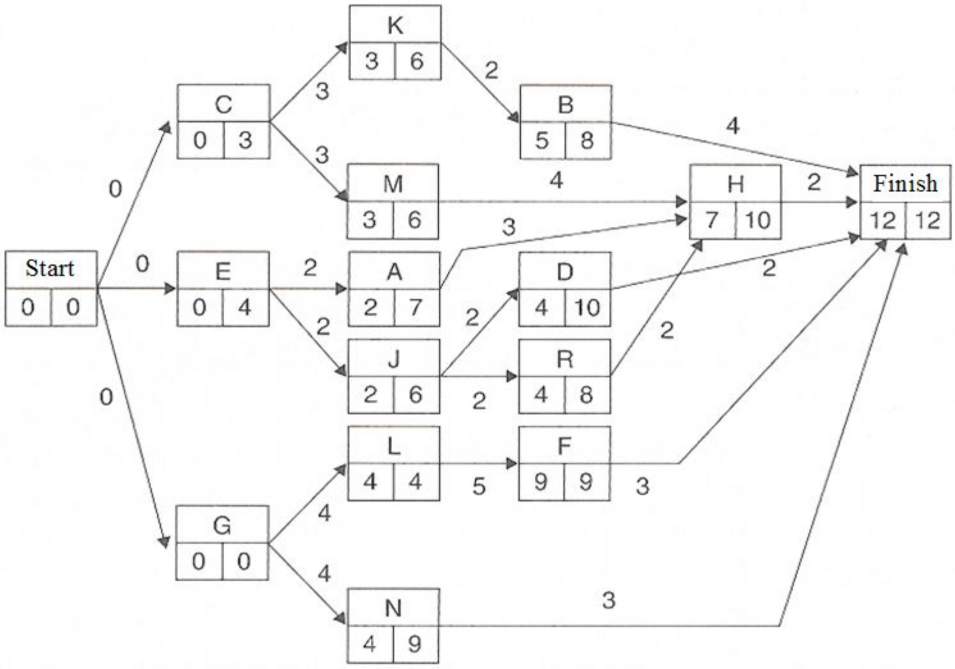
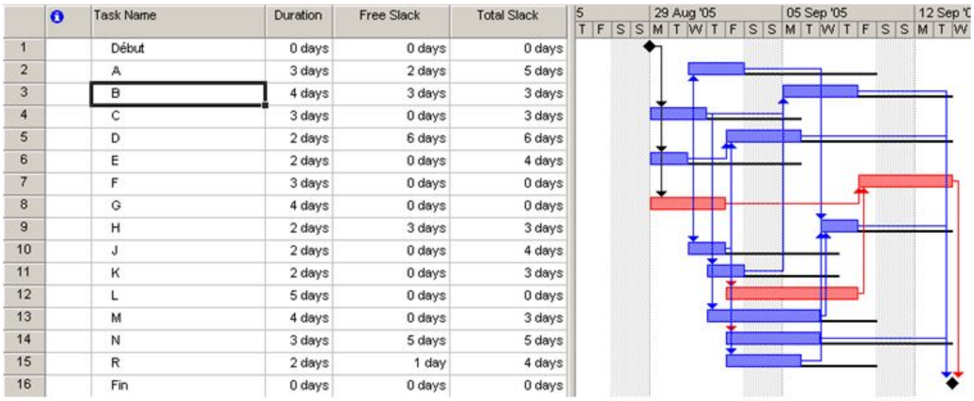

Start
0 0
C
0 3
E
0 4
G
0 0
K
3 6
M
3 6
A
2 7
J
2 6
L
4 4
N
4 9
B
5 8
D
4 10
R
4 8
F
9 9
H
7 10
Finish
12 12

Figure 8.316: **PERT (Program Evaluation Review Technique) diagram**

We see well in this graph that the finite set of tasks $\{\text{Start}, G, L, \text{Finish}\}$ are critical.

A great tool for using such graphs is (among others) Microsoft Project whose the corresponding Gantt diagram for the example above is:



		Task Name	Duration	Free Slack	Total Slack
1		Début	0 days	0 days	0 days
2		A	3 days	2 days	5 days
3		B	4 days	3 days	3 days
4		C	3 days	0 days	3 days
5		D	2 days	6 days	6 days
6		E	2 days	0 days	4 days
7		F	3 days	0 days	0 days
8		G	4 days	0 days	0 days
9		H	2 days	3 days	3 days
10		J	2 days	0 days	4 days
11		K	2 days	0 days	3 days
12		L	5 days	0 days	0 days
13		M	4 days	0 days	3 days
14		N	3 days	5 days	5 days
15		R	2 days	1 day	4 days
16		Fin	0 days	0 days	0 days

5
29 Aug '05
05 Sep '05
12 Sep '0
T F S S M T W T F S S M T W T F S S M T W

Figure 8.317: **Same graph but seen in Microsoft Project 97**

D24. A "cycle" is a simple path looping on itself. A graph in which there is no cycles is known as "acyclic".

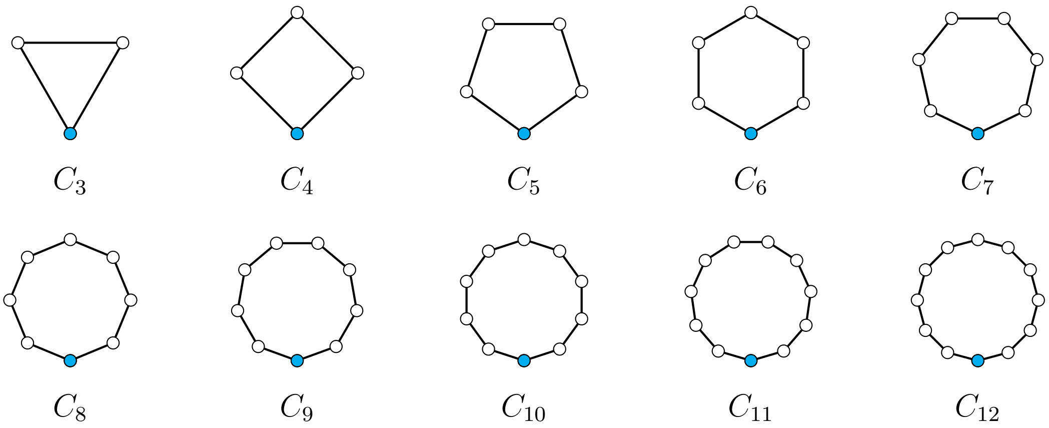


Figure 8.318: **Examples of cyclic graphs**

Non-connex acyclic graphs made of trees are an interesting class of graphs with remarkable properties and a name: "forests" (a term often used in computer networks).

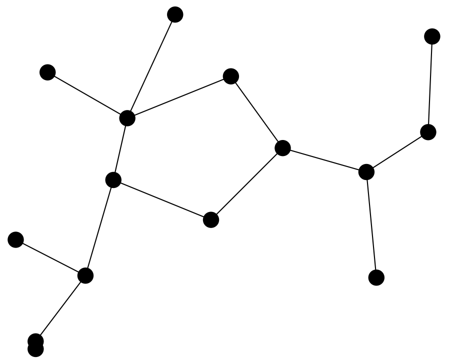


Figure 8.319: **Example of a connex graph containing a cycle**

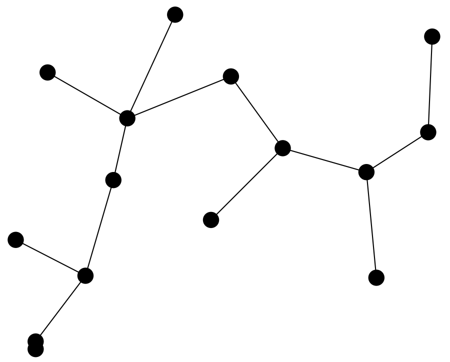


Figure 8.320: **Example of a connex graph containing no cycles**

Below an example of forest (composed of trees, each tree being connex but the whole forming an acyclic and non-connex graph):

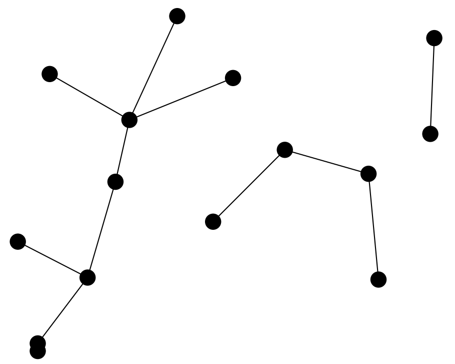


Figure 8.321: **Example of a forest (3 connex graphs)**

Thus we see that a forest is a graph whose components are trees. The vertices of degree 1 are named "leaves" of the tree.

Properties:

P1. If in a graph G is any vertex is of degree greater than or equal to 2, then G has at least one cycle (trivial).

> **Remark**
>
> This simple property implies that a graph without cycle without graph has at least one vertex of degree 0 or 1.

P2. An acyclic graph G with n vertices has at most $n-1$ edges as we already know.

D25. An "Eulerian cycle" is a cycle passing once and only once by each edge of the graph and returning to the starting point (we will see further properties required by a graph for such a cycle exists on it) and a graph having an eulerian cycle is named "Eulerian graph".

D26. A "Hamiltonian cycle" is a simple cycle through all the vertices of the graph once and only once and a graph having an eulerian cycle is named "Hamiltonian graph". For a Hamiltonian cycle, the graph must be connex and must not have any single node.

To better understand the difference between the two previous definitions the reader can see in the figure below some graphs that are Eulerian and Hamiltonian, Eulerian but not Hamiltonian and so on... (the four possibilities):

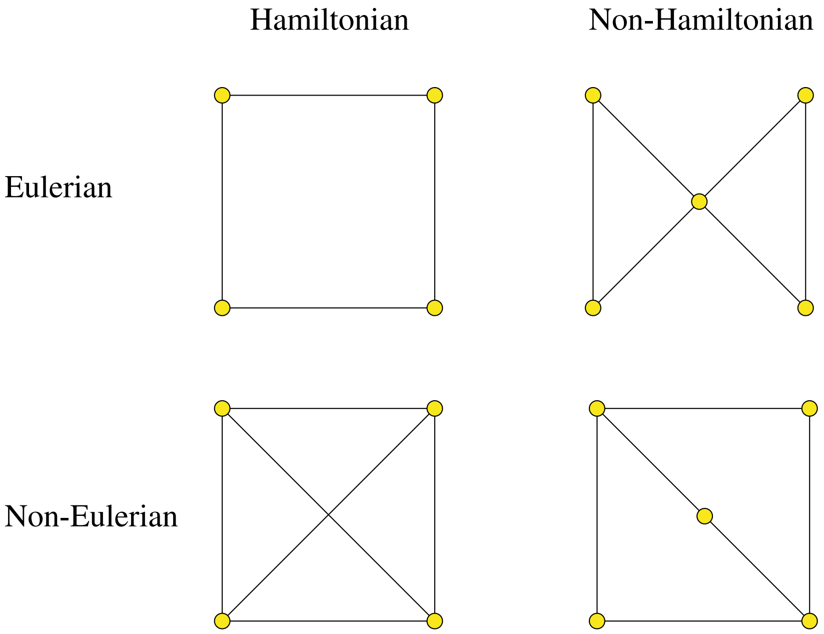


Figure 8.322: **Example of Hamiltonian or Non-Hamiltonian graphs**

It is appropriate now to open a parenthesis (for straws ...) on one of the most famous problem in graph theory: Königsberg bridges.

Leonhard Euler, liked to take a walk in his city of Königsberg. Following the legend he liked especially to walk through seven bridges that cross the river. Age coming (mathematical knowledge too ...), he wondered if without sacrificing his walk, he could shorten the length of his walk by crossing each bridge only once. This problem is probably one of the oldest known one in graph theory: that the existence of a chain passing once and only once by each edge (sadly today only four bridges remain standing).

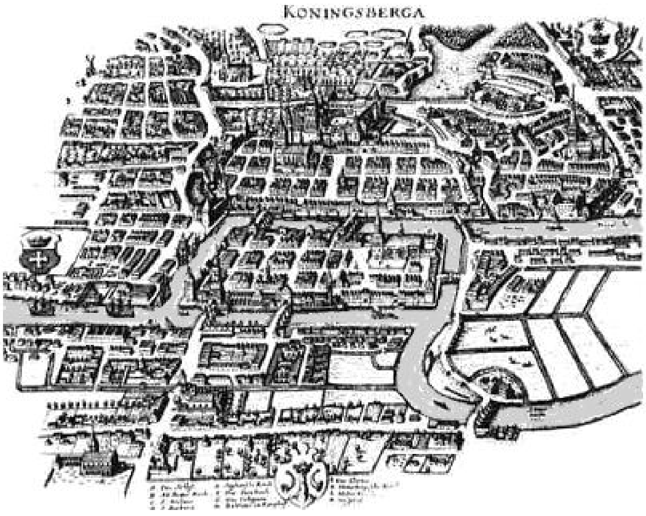


Figure 8.323: **Antique map of the city of Königsberg**

The river divides the city of Königsberg in four parts: a, b, c, d. Each bridge connects two such parties. Then we can represent our problem by a graph with four vertices, wherein each edge represents one of the seven bridges of Königsberg. In this example the graph is not a simple graph:

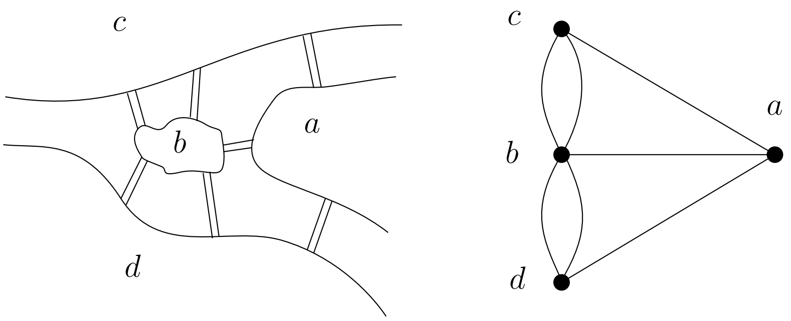


Figure 8.324: **Königsberg equivalent graph**

The question here is whether the graph is Eulerian or not? If our problem the obtained graph is Eulerian, we need to exhibit an Euler cycle, which does not seem easy. But if it is not? Euler gave a strong characterization of Eulerian graphs given by the following statement:

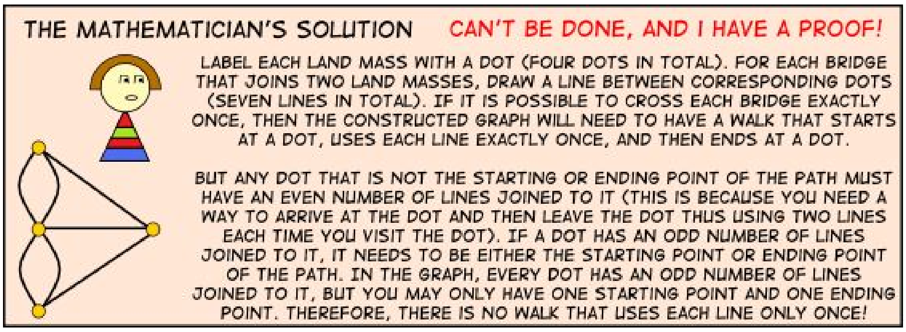
Theorem 8.30 (Euler Theorem)**.** A graph is Eulerian if and only if it is connex and all vertices have even degree (so there is an even number of edges that arrive on each vertex of which half of them used to arrive at the vertices, the others to leave it) except for maximum two (these two exceptions being the vertice of departure and this of arrival).

Specifically for a connex graph:

- The graph has no odd vertices, then it is Eulerian (and the chain is cyclic).
- The graph can have only one odd vertices from the property (already stated above) that in a graph, the vertices having odd order is always in even.
- If the graph has two odd vertices, then these vertices are the ends of the Eulerian cycle.

As corollary we have that a graph have more than two odd vertices never have an Eulerian cycle.

With this characterization (as we will prove it just after), the vertices a, b, c, d being of odd order, we know immediately that it is impossible to walk cross all the Königsberg bridges in only one time during a walk.



THE MATHEMATICIAN'S SOLUTION CAN'T BE DONE, AND I HAVE A PROOF!

LABEL EACH LAND MASS WITH A DOT (FOUR DOTS IN TOTAL). FOR EACH BRIDGE THAT JOINS TWO LAND MASSES, DRAW A LINE BETWEEN CORRESPONDING DOTS (SEVEN LINES IN TOTAL). IF IT IS POSSIBLE TO CROSS EACH BRIDGE EXACTLY ONCE, THEN THE CONSTRUCTED GRAPH WILL NEED TO HAVE A WALK THAT STARTS AT A DOT, USES EACH LINE EXACTLY ONCE, AND THEN ENDS AT A DOT.

BUT ANY DOT THAT IS NOT THE STARTING OR ENDING POINT OF THE PATH MUST HAVE AN EVEN NUMBER OF LINES JOINED TO IT (THIS IS BECAUSE YOU NEED A WAY TO ARRIVE AT THE DOT AND THEN LEAVE THE DOT THUS USING TWO LINES EACH TIME YOU VISIT THE DOT). IF A DOT HAS AN ODD NUMBER OF LINES JOINED TO IT, IT NEEDS TO BE EITHER THE STARTING POINT OR ENDING POINT OF THE PATH. IN THE GRAPH, EVERY DOT HAS AN ODD NUMBER OF LINES JOINED TO IT, BUT YOU MAY ONLY HAVE ONE STARTING POINT AND ONE ENDING POINT. THEREFORE, THERE IS NO WALK THAT USES EACH LINE ONLY ONCE!

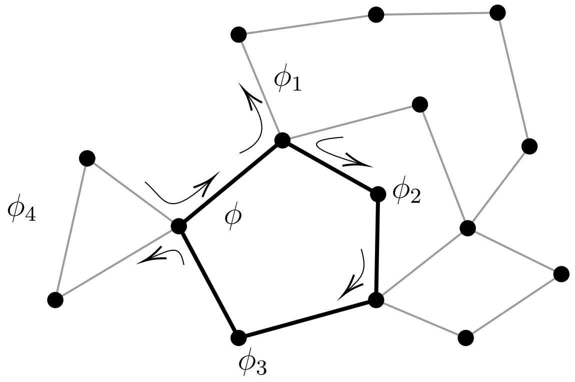
Proof 8.30.1. The proof will be done if three steps:

(a) Let us suppose that a graph G is Eulerian. Then there exists a chain c browsing once and only once each edge (until then it's easy). Of course, in the case of a chain, the peaks being located at the ends of the chain are of odd order and are only two.

(b) Consider now a vertex x and let us suppose this time not a Eulerian graph but an Euler cycle. During the course of this cycle, each time we go through the vertex x, we find ourselves at the starting point and to perform a new round two edged are available to us (indeed the path can be walked in two direction since the graph is undirected!). The vertex x is the of even order and can be defined anywhere in the cycle, hence the fact that all the vertices are of even degree.

(c) Conversely, consider a connected graph G whose vertices are all of even degree. We will show by induction on the number of edges that G is then Eulerian:

- If G is reduced to a single isolated vertex, it is obviously Eulerian. Otherwise all the vertices of G are of degree greater or equal to 2. This implies that there is a cycle ϕ on G:
- Consider the partial graph H consisting of the edges outside the cycle ϕ. The vertices of H are also of even degree, the cycle containing an even number of edges incident to each vertex. By induction each connex component H_i of H is an Eulerian graph, and so has an Eulerian cycle ϕ_i. To rebuild an Euler cycle on G, we only need to merge the cycle ϕ the with different cycles ϕ_i. For this, we travel the cycle ϕ from an arbitrary vertices; when we meet for the first time a vertices x from H_i, we substitute it the cycle ϕ_i. The resulting cycle is an Euler cycle in G, the cycle ϕ and the cycles ϕ_i form a partition of the edges.



■ Q.E.D.

Remark

This principle of decomposing a graph into connex graphs and summing them permits to build a recursive algorithm for determining if a graph is Eulerian or not.

D27. Two graphs G and G' are "complementary graphs" when they satisfy the following conditions:

(a) They have the same set of vertices.

(b) Two vertices x, y that are neighbours in G are not neighbours in G'.

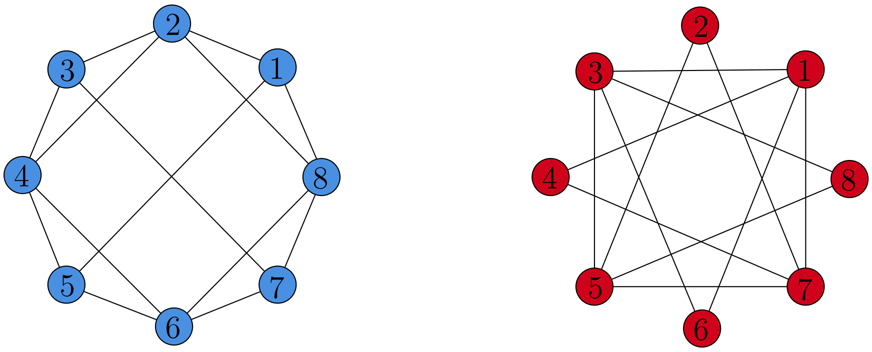


Figure 8.325: **Example of complementary graph**

A "self-complementary" graph is a graph which is isomorphic to its complement. The simplest non-trivial self-complementary graphs are the 4-vertex path graph and the 5-vertex cycle graph.

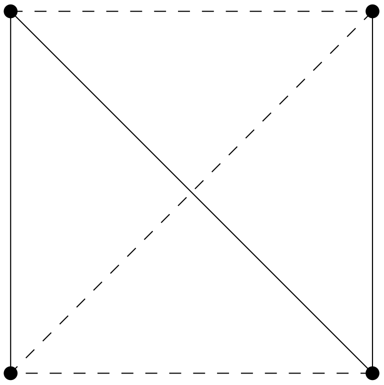


Figure 8.326: **Example of self-complementary graph**

D28. A "bipartite graph" is a graph $K_{p,q}$ such that we can partition the set of all its vertices into two classes cardinal of p and q respectively so that each edge has one end in each of the two classes.

✍Example:

Here is a representation of a classical bipartite graph $K_{3,3}$. It is the famous problem of the supply of three houses from three supplier (water, electricity, gas) without the right of alignment of services (using same "pipelines"). We immediately see that this graph is non-planar.

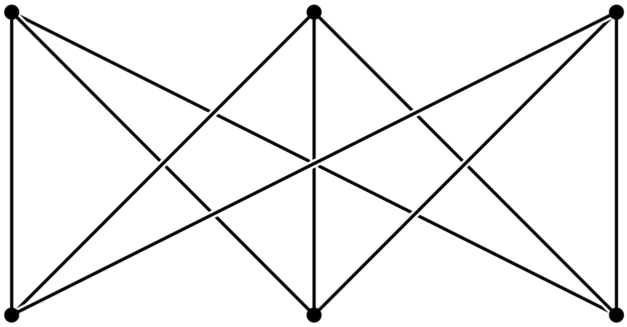


Figure 8.327: **Example of a $K_{3,3}$ bipartite graph**

Remark

The complete bipartite graph $K_{p,q}$ is a graph with $n=p+1$ vertices and configured in such a way that each vertex of a class is adjacent to all the vertices of the other and only thereto.

D29. Two graphs are "isomorphic graphs" if there exists a bijection of f from X into X' such that for any vertices x and y of G:

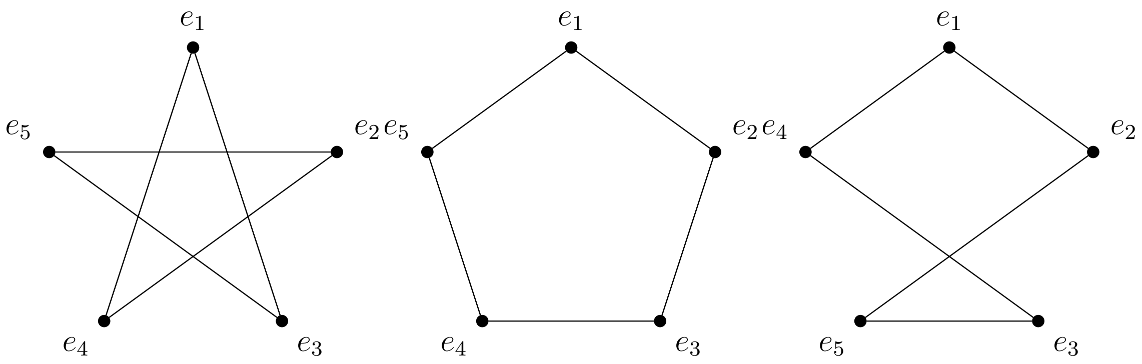
$$x \text{ adjacent to } y \text{ in } G \leftrightarrow f(x) \text{ adjacent to } f(y) \text{ in } G' \tag{8.8.21}$$

We also say that f is an "isomorphism of G on G'".

More formally, an isomorphism of graphs G_1 and G_2 is a bijection $f:V(G_1)\leftrightarrow V(G_2)$ that preserves adjacency. That is to say:

$$u \sim_{G_1} v \leftrightarrow f(u) \sim_{G_2} f(v) \tag{8.8.22}$$

For example on the figure below we have three isomorphic graphs:



	1	**2**	**3**	**4**	**5**
1			1	1	
2				1	1
3	1				1
4	1	1			
5		1	1		

	1	**2**	**3**	**4**	**5**
1		1			1
2	1		1		
3		1		1	
4			1		1
5	1			1	

	1	**2**	**3**	**4**	**5**
1		1		1	
2	1				1
3				1	1
4	1		1		
5		1	1		

Figure 8.328: **Isomorphic graphs with their adjacency matrix**

Caution!!! Sometimes we talk about graphs equivalence "to a given isomorphism". This means more clearly, that at the exception of a single violation among the set of all edges, that the graphs are isomorphic.

If there exists a bijection of f of X into itself such that:

$$x \text{ adjacent to } y \text{ in } G \leftrightarrow f(x) \text{ adjacent to } f(y) \text{ in } G \tag{8.8.23}$$

then we say that f is an "automorphism of graph" in G (by permutation of the vertices there are many possible examples ...).

As the isomorphism in the case of graphs go from one set to another of the same size n (n to n mapping), the number of different possible bijections is (see section Probabilities page 525):

$$n! \tag{8.8.24}$$

This means there is a maximum of $n!$ graphs which can be grouped in a same equivalence class. Consequently, there exists at least (lower bound) the ratio of the total n of vertices on the cardinal majorated by the greatest possible equivalence class (but not necessarily existing!):

$$N=\frac{2^{\frac{n(n-1)}{2}}}{n!} \tag{8.8.25}$$

Using the gross majoration $n!<n^n$, we have:

$$N>\frac{2^{\frac{n(n-1)}{2}}}{n^n} \tag{8.8.26}$$

Therefore:

$$\log_2(N)>\frac{n(n-1)}{2}-n\log_2(n) \tag{8.8.27}$$

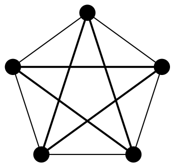
Given:

$$\log_2(N)>\frac{n^2}{2}\left(1-\frac{1}{n}-\frac{2\log_2(n)}{(n)}\right) \tag{8.8.28}$$

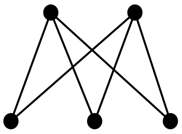
Thus, when n approaches the infinity, $\log_2(N)$ admits a lower bound of order n^2.

The reader should also keep in mind that some graphs are used more than other, and get special names:

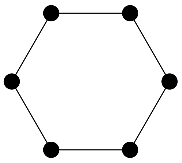
- K_n: The complete graph on n vertices
- $K_{m,n}$: The complete bipartite graph with sets of m and n vertices
- C_n: The cycle on n vertices, just one big loop
- P_n: The path on n vertices, just one long path



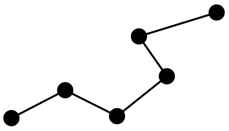
K_5



$K_{2,3}$



C_6



P_6

8.8.2 Graph Adjacency Matrix

Formally, a graph is also a set on which we have defined a binary, anti-reflexive (no element is related to itself) and symmetric (if x is related to y, then y is related to x) relation. The graph structure may then seem particularly poor.

But we can also associate to a graph G vertices $x_1,\ldots,x_n$ a square matrix M of dimensions $m\times n$ named the "adjacency matrix" of the graph and whose elements are 0 or 1. Denoting by m_{ij} the term located at the intersection of row number i (representing the vertices x_i) and column j (representing the vertices x_j), M is defined as:

$$m_{ij}=1 \quad \text{if and only if}\, x_i,x_j \text{ are adjacents} \tag{8.8.29}$$

Remark

We also know that graphs can be represented by matrices in the context of the study of Markov chains in the field of probabilities (see section Probabilities page 537).

Let us see an example both abstract but also easily applicable to many fields of industry, sociology and biology (there is also application for the adjacency matrix in Text Mining as you can see in the `R` companion book). Consider the following directed graph and observe that it is not symmetrical nor antireflexive:

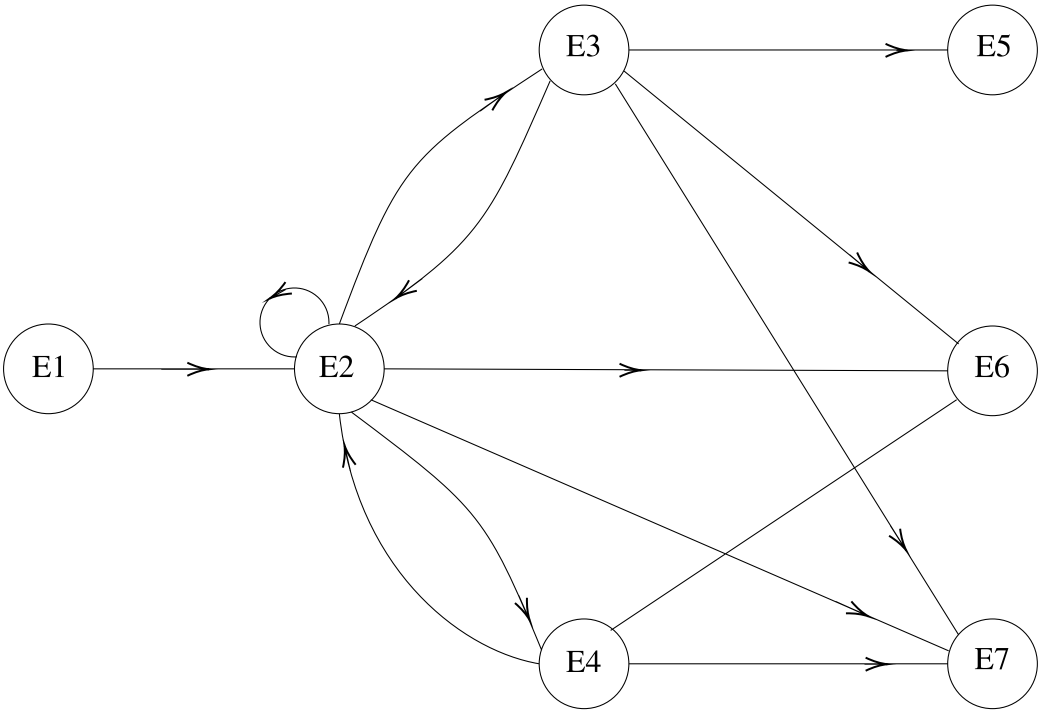


Figure 8.329: Example of Markov chain (nor antireflexive or symmetric)

And as we just say we can represent this diagram (connected, and oriented graph) in the form of a table in which we denote by "1" a possible transition from the state mentioned at the top of the column to the state mentioned at the beginning of the row and "0" otherwise:

↗	E1	E4	E2	E3	E5	E7	E6
E1.	0	0	0	0	0	0	0
E4.	0	0	1	0	0	0	0
E2.	1	1	1	1	0	0	0
E3.	0	0	1	0	0	0	0
E5.	0	0	0	1	0	0	0
E7.	0	1	1	1	0	0	0
E6.	0	1	1	1	0	0	0

Table 8.9: **Adjacency matrix of the connected graph**

It is essential to fully understand the meaning of the values in this table! But at this level of the book that should not pose major difficulties to the reader.

Now let us evaluate the number of ways to go:

1. From E2 to E2 in two stages
2. From E3 to E4 in two stages
3. Form E2 to E7 in two stages

It is easy in the particular case above to count these possibilities. But in the case of a more complex graph it becomes difficult or even impossible for a human being within a reasonable time (and also for some computer at this time).

We have to use then the following theorem:

Theorem 8.31. Given an orient graph with $x_1, x_2, \ldots, x_n$ vertices and of adjacency matrix:

$$M = (m_{ij}) \tag{8.8.30}$$

For any integer k, then the number of paths of length k of vertices x_1 to a vertices x_j is given by:

$$M^k = m_{ij}^{(k)} \tag{8.8.31}$$

where the exponent of M, denotes the power k of the adjacency matrix.

Proof 8.31.1. Let us perform an induction on k:

$$m_{ij}^{(1)} = m_{ij} \tag{8.8.32}$$

designates the number of paths going from x_i to x_j (where i and j can be equal). Let us suppose the result true for the integer $k-1$, with $k \geq 1$ as:

$$M^k = M^{k-1} M \tag{8.8.33}$$

we then have (by construction of matrix multiplication):

$$m_{ij}^{(k)} = \sum_{I=1}^{n} m_{ij}^{(k-1)} m_{lj} \tag{8.8.34}$$

By induction hypothesis, $m_{il}^{(k1)}$ is the number of paths of length $k-1$ from x_i to x_l and m_{lj} is, we know it (by construction!), the number of paths of length 1 from x_l to x_j and in particular it is equal to 1 if (x_i, x_j) is an edge of the graph and equal to 0 otherwise!

So the product:

$$m_{il}^{(k-1)}m_{lj} \tag{8.8.35}$$

gives for a given value of l the number of paths of length k from x_i to x_j whose last edge is (x_l, x_j).

The sum:

$$\sum_{l=1}^{n} m_{il}^{(k-1)}m_{lj} \tag{8.8.36}$$

therefore gives all the possibilities (paths) of length k from x_i to x_j regardless of the starting point of the last edge!

■ Q.E.D.

Thus, in our example, the adjacency matrix M is given by:

$$M=\begin{bmatrix}0&0&0&0&0&0&0\\0&0&1&0&0&0&0\\1&1&1&1&0&0&0\\0&0&1&0&0&0&0\\0&0&0&1&0&0&0\\0&1&1&1&0&0&0\\0&1&1&1&0&0&0\end{bmatrix} \tag{8.8.37}$$

and is neither symmetric or antireflexive as we have already mentioned.

So if we carry this matrix at the power of k, each component $m_{ij}^{(k)}$ will give all possibilities (paths) of length k going from x_i to x_j Thus we have for example with $k=2$:

$$M^2=\begin{bmatrix}0&0&0&0&0&0&0\\1&1&1&1&0&0&0\\1&1&3&1&0&0&0\\1&1&1&1&0&0&0\\0&0&1&0&0&0&0\\1&1&3&1&0&0&0\\1&1&3&1&0&0&0\end{bmatrix} \tag{8.8.38}$$

using for example the function `MMULT( )` available in most spreadsheet softwares.

We then have the answer to all our three initial questions by reading the above matrix:

1. From E2 to E2 in two stages there are 3 possibilities

2. From E3 to E4 in two stages there is 1 possibility
3. From E2 to E7 in two stages there are 3 possibilities

It is possible to gain some time in such calculations. If we denote C_i, the i-th column of the matrix M, then:

$$MC_i = M_i^2 \tag{8.8.39}$$

So we always get the number of paths of length k from a given starting point corresponding to the column i.

This example has a very interesting approach in some areas studying the behaviour of individuals in different situations (shopping, tourisms, accidents, etc.).

If instead of writing in the matrix M the number of possible paths from one vertices to another, we write the probability (part) of the total number of individuals who go from one vertices to reach another one we have for example (in practice the values are imposed by experimental observation!) the following matrix:

$$M = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.094 & 0 & 0 & 0 & 0 \\ 1 & 0.954 & 0.662 & 0.75 & 0 & 0 & 0 \\ 0 & 0 & 0.218 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.175 & 1 & 0 & 0 \\ 0 & 0.034 & 0.017 & 0.035 & 0 & 1 & 0 \\ 0 & 0.012 & 0.009 & 0.04 & 0 & 0 & 1 \end{bmatrix} \tag{8.8.40}$$

which is already the stochastic transpose matrix of the visible graph below (according to the theory of Markov chains we saw in the section of Probabilities that we need to take the transpose of the stochastic matrix to calculate the probabilities of states).

Considering that these probabilities do not change over time, then we have a homogeneous Markov chain (without cycles). We then see that:

1. The sum of the transition probabilities going from each vertices (state) must always logically be equal to 1 (which we already mentioned in the section of Probabilities)
2. Everyone start from the first vertices E1
3. Some stagnate (stop) at some stage
4. Those arriving at an end stage E5, E6 or E7 stay there and do not retrace their steps (absorbing states).

The equivalent graph is therefore:

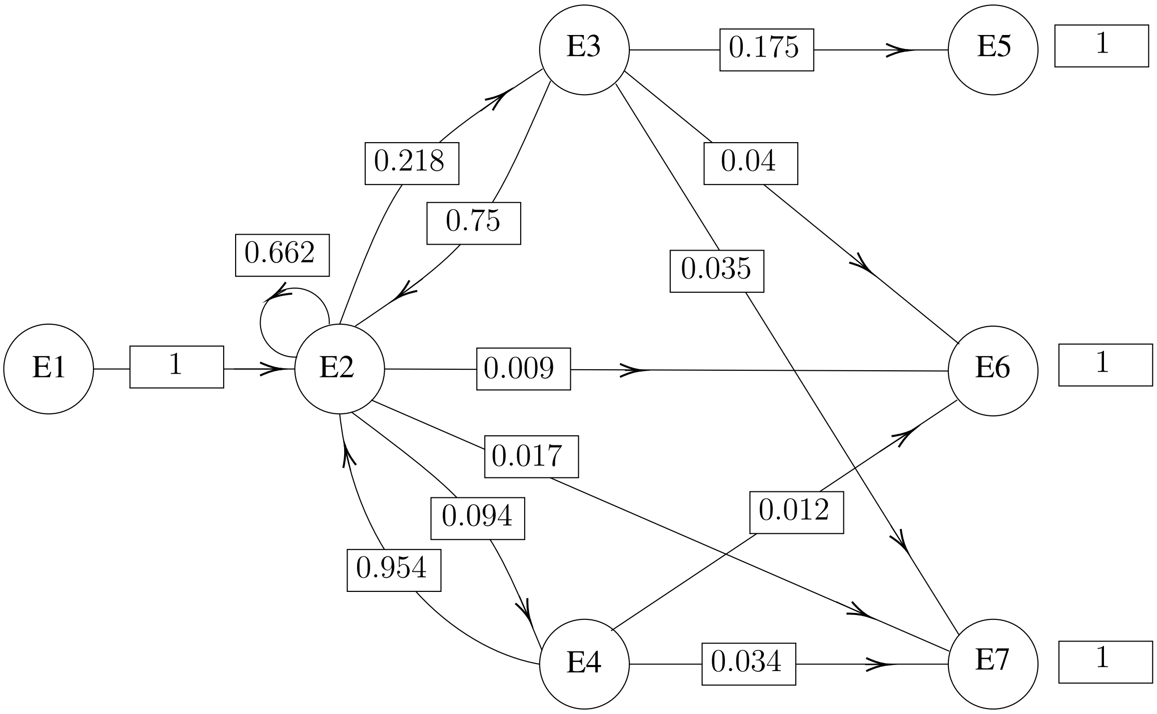


Figure 8.330: **Markov probabilistic transition graph**

By denoting N (instead of M to avoid confusion) the matrix constructed from the graph above we see that C_i is if one of the columns of the matrix M then for example:

$$NC_1 = T_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.094 & 0 & 0 & 0 & 0 \\ 1 & 0.954 & 0.662 & 0.75 & 0 & 0 & 0 \\ 0 & 0 & 0.218 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.175 & 1 & 0 & 0 \\ 0 & 0.034 & 0.017 & 0.035 & 0 & 1 & 0 \\ 0 & 0.012 & 0.009 & 0.04 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0.094 \\ 0.662 \\ 0.218 \\ 0 \\ 0.017 \\ 0.009 \end{bmatrix} \tag{8.8.41}$$

gives the sum of the transition probabilities that what goes from E1 arrives 2 in respectively E1(0), E2(0.662), E3(0.218) ... (the distribution of the initial vector may be any as long as the sum of column of values is equal to 1)!

If we then multiply again:

$$NT_1 = T_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.094 & 0 & 0 & 0 & 0 \\ 1 & 0.954 & 0.662 & 0.75 & 0 & 0 & 0 \\ 0 & 0 & 0.218 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.175 & 1 & 0 & 0 \\ 0 & 0.034 & 0.017 & 0.035 & 0 & 1 & 0 \\ 0 & 0.012 & 0.009 & 0.04 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0.094 \\ 0.662 \\ 0.218 \\ 0 \\ 0.017 \\ 0.009 \end{bmatrix} = \begin{bmatrix} 0 \\ 0.062 \\ 0.691 \\ 0.144 \\ 0.038 \\ 0.039 \\ 0.025 \end{bmatrix} \tag{8.8.42}$$

gives the sum of the transition probabilities that what goes from E1 arrives 3 in respectively E1(0), E2(0.691), ... and so on. We can therefore know what is the probability that an individual arriving on E2 can arrive to one of the terminal nodes (E5, E6 or E7) on a number of finite step.

Remark

Remember that the sum of the probabilities of the resulting columns T is always equal to 1 for the transpose of the stochastic matrix.

Continuing a long time as well ... we find that the equilibrium measurement π that satisfies (see section Probabilities page 540):

$$P^T\pi = \pi \tag{8.8.43}$$

is:

$$\pi = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0.45 \\ 0.32 \\ 0.23 \end{bmatrix} \tag{8.8.44}$$

regardless of the distribution of the starting vector. Property named "ergodic" in the field of Markov chains.

Which means 45% probability of being in E5, 32% probability of being in E7 and 23% probability of being in E6. Another way to look at it is to say that if a cohort of 100 individuals start at node E1 with constant probability in time between the various states transitions, the equilibrium we will have is 45 people in E5, 32 people in E7 and 23 people in E6.

8.8.3 Categories

The introduction of categories through the "category theory" by Eilenberg and MacLane in 11942 (holocene calendar) was to transform difficult problems of topology in more affordable algebra problems. Later, the theory of categories has grown significantly, both for itself and for its applications in the most diverse fields of mathematics (e.g. differential geometry). Although some of its autonomous development has sometimes been criticized, categories are now recognized as a powerful language to develop a universal semantics of mathematical structures. They are also used in logic and more recently in physics, and a fruitful collaboration seems to develop between IT and categoricians.

Definitions (#8.8.398):

D1. Intuitively a "category" is just a directed graph on which we have given a law to compose consecutive arrows, satisfying certain axioms.

D2. An "oriented graph" is formed by a set of nodes of graph, with links between them, represented by arrows from a node A to a node B, what we denote by $f: A \mapsto B$. We say then that A is the "source" of the arrow, and B and its "goal". There may be several arrows having same source and the same goal (we say then that they are "parallel") and there can be "closed" arrows, which source and target coincide.

D3. Two arrows f, g are say to be "consecutive arrow" if the target of the first is also the source of the second:

$$f : A \to B \qquad g : B \to C \tag{8.8.45}$$

then we say that they form a path of length 2 from A to C.

A category is then a graph in which we define a composition of arrows, associating to any path (f, g) of length 2 from A to C an arrow of the graph from A to A, named "composed of the path", and denoted fg:

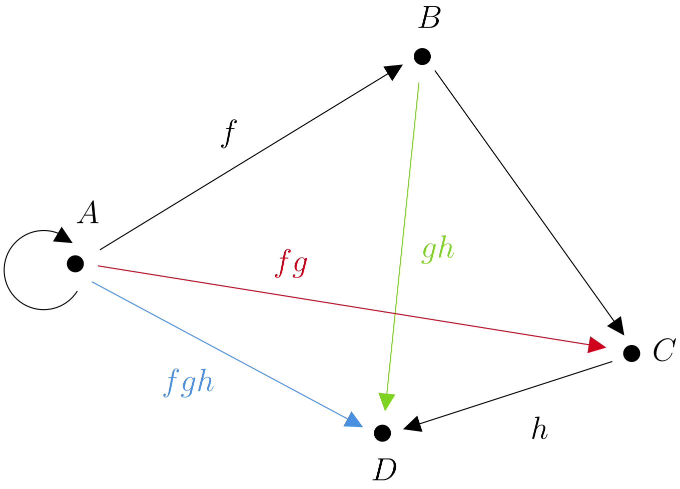


Figure 8.331: Example of graph of categories

This composition satisfies the following axioms:

A1. Associativity: If fgh is a path of length 3, the both compositions $f(gh)$ and $(fg)h$ that we deduce are associative. It follows that to any path of length n is also associated only a single composition of vertices (invariance of the path).

A2. Identities: At any node A is associated a closed arrow from A to A, obviously named "identity" of A and denoted Id_A, from which the composition with an arrow of source or target A is equal to this other arrow.

More generally, a path (of length n) from A to A_n is a sequence of n consecutive arrows:

$$f_1 : A \to A_1, f_2 : A_1 \to A_2, \ldots, f_n : A_{n-1} \to A_n \tag{8.8.46}$$

Remarks

R1. The graph nodes are named "objects" of category and the arrows the "morphisms" (or simply "links") within the framework of category theory.

R2. An arrow f is an isomorphism (see section Set Theory page 443) if there exists an arrow g (named "inverse") such that the compositions fg and gf are identities (this inverse is therefore unique).

Thus, a category is formed by objects (the nodes of the graph) and by the links between them (the arrows or morphisms), but the essential idea is to emphasize the links relatively to the objects. In fact, the success of the categories in the most varied fields is due to the quantity of information on objects that can be inferred from the mere consideration of links and operations on them, regardless of the nature and anatomy of these objects.

In the next few lines, we will explain how to read the directed graphs we can sometimes encounter in math books. This will be a good example of category theory, because we have already encountered such graphs without describing them in the section Numbers and Cryptography for example.

For simplicity we will explain these graphs when the basic objects are sets (which is the most common case in this whole book anyway actually).

Let us consider three sets A, B, and C and three applications such that:

$$f:A\mapsto B, g:A\mapsto C, h:B\mapsto C \tag{8.8.47}$$

and let us rewrite it as following:

$$f:A\rightarrow B, g:A\rightarrow C, h:B\rightarrow C \tag{8.8.48}$$

We can therefore consider the applications f, g and h as arrows that connect the objects (sets) A, B, and C to form a triangle:

$$\begin{tikzcd}
A \arrow[r, "g"] \arrow[d, "f"'] & C \\
B \arrow[ur, "h"'] &
\end{tikzcd}$$

Where obviously $g=h\circ f$ or more explicitly for those that have forgotten the signification of this notation that for any $a\in A$ we have $h(f(a))=g(a)$.

Definition (naive version #399): We say that an oriented graph is a "commutative diagram" if all the paths we take to get from one object (set) to another one represent the same application.

Or let us consider:

$$f:A\rightarrow B, g:B\rightarrow C, h:A\rightarrow C \tag{8.8.49}$$

and such that each application f,g,h has its Kernel that is equal to the set of neutral elements of the starting set. Therefore:

$$\begin{tikzcd} A \arrow[loop left, "1_A"] \arrow[r, "f"] \arrow[dr, "g \circ f"'] & B \arrow[loop right, "1_B"] \arrow[d, "g"] \\ & C \arrow[loop right, "1_C"] \end{tikzcd}$$

We can complicate almost without limits such diagrams considering more sets and arrows (applications) connecting them. For example:

$$\begin{tikzcd} A \arrow[r, "g"] \arrow[d, "f"'] & C \arrow[d, "r"] \\ B \arrow[r, "h"'] & D \end{tikzcd}$$

This diagram is commutative if and only if $h \circ f = r \circ g$.

Remarks

R1. Usually in the mathematical literature such diagrams are implicitly commutative.

R2. As already mentioned, the objects of these diagrams can more generally be groups, rings, fields, topological space, etc. In these cases, the arrows are no longer any applications but group homomorphisms, or ring homomorphisms, continuous applications, etc.

~90 %

★★★★☆

28 votes, 70.00%

8.9. Knot Theory

René Descartes had imagined a system of the World, when the universe was animated by vortices. At the end of 118th century (holocene calendar), Peter Guthrie Tait et William Thomson Kelvin have resurrected this theory, and interpreted the chemical bonds imagining knotted molecules that interlace.

Remarks

R1. The name "Knots theory" that has emerged in the scientific community is quite unfortunate. Some (French) mathematicians rightly use instead the name of "theory of intertwining and braids" which is more correct and general (since a node is interlaced with several components).

R2. Most of the text and figures below is a reproduction, with agreement, of Professor Michael Eisermann course from the University of Stuttgart (`http://www.igt.uni-stuttgart.de/eiserm`).

The mathematical knot theory was pioneered by the work of Charles Newton Little and Thomas Kirkman who tried to give a basis for the physical ideas of Peter Guthrie Tait et William Thomson Kelvin. The first nodes classifications using a wire image of the nodes, without thickness, and a planar projection, as the shadow cast on a screen. One immediately noticeable character concerns crossings where we distinguish the strand above and strand below. The cannages, or braiding strips are based on the consideration of the crossings, and the number of such intersections will be the first classifying element.

8.9.1 Braids Representation

An example of braid (set of strands) and the corresponding string (a string is a braid that was closed):

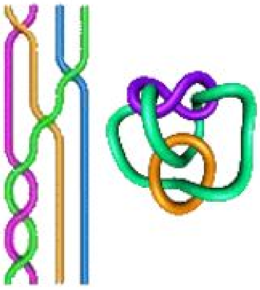


Figure 8.332: **Braid and corresponding string (node diagram)**

Let us focus on the braid to the left of the figure above to begin and let us ask ourselves what would be the best way of putting things in order to compare braids? Let us first attempt with the following representation of a particular braid:

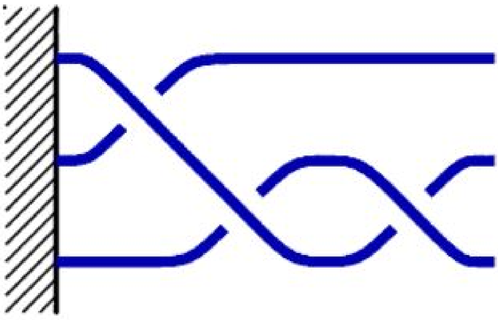


Figure 8.333: **Particular presentation of a braid**

As the strands of a braid are flexible and can move, we notice immediately that this representation has a problem: all the braids are equal. Effectively:

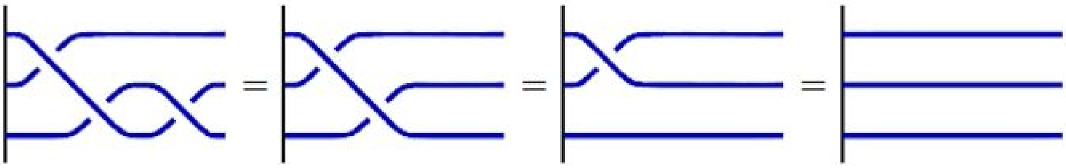


Figure 8.334: **Example of the problem of the braid representation**

A better model then is perhaps to secure the ends of both sides:

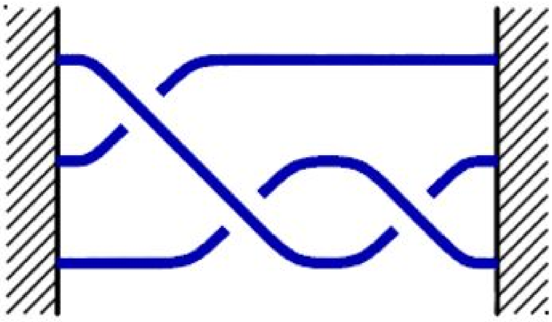


Figure 8.335: **Other choice of representation**

Of course, with this model the middle strands may still move as below and the braid remains invariant:

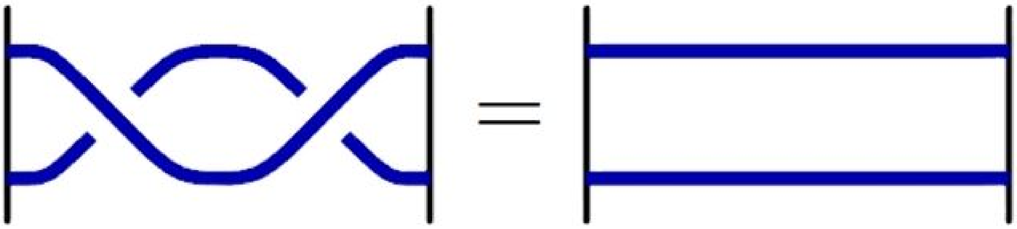


Figure 8.336: **First example of invariant braid**

or otherwise as this:

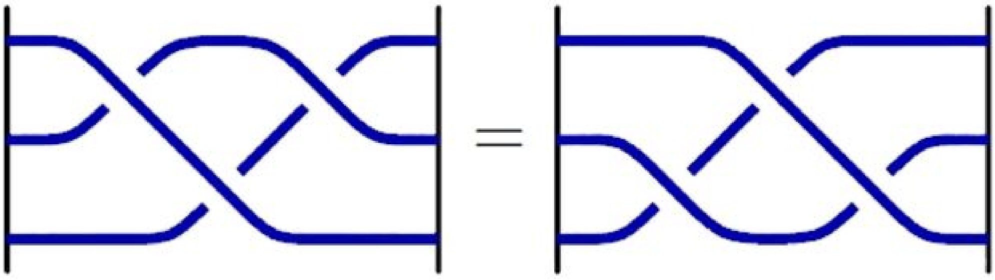


Figure 8.337: **Second example of invariant braid**

We can also translate the crossings and the braid remains invariant:

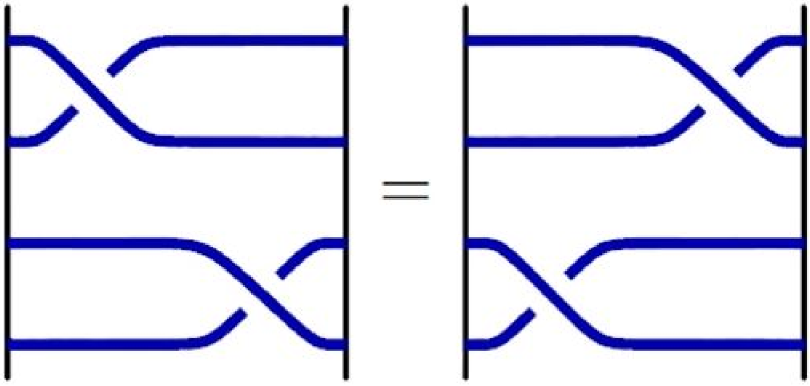


Figure 8.338: **Translation of crossing and preserving the trivial invariance**

The length is not a variable in this model:

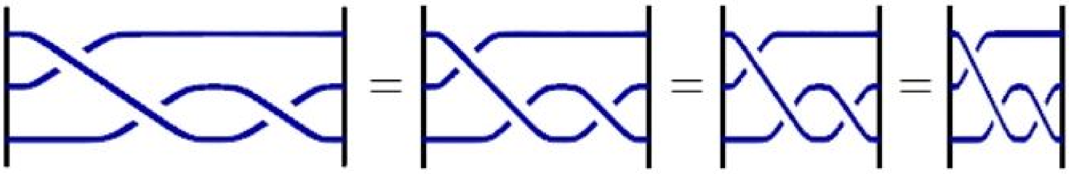


Figure 8.339: **Length invariance**

8.9.1.1 Braids Group

If we define the multiplication of two braids as the operation that consists schematically to:

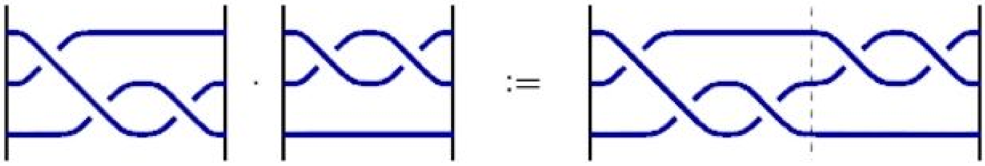


Figure 8.340: **Example of multiplication of two braids**

We can ask ourselves if this is a type of commutative group?

Remember before that we have defined the group structure in the section of Set Theory as follows: We refer to a set by the term "group" if its components meets the three conditions of what we name "internal law group ", defined below:

$$(G,\star)\text{ is a group if}=\begin{cases}\star\text{ is associative}\\ \text{it exists for }G\text{ a neutral element for }\star\\ \text{any elelement of }G\text{ has a symetric for }\star\end{cases}\tag{8.9.1}$$

If furthermore, the internal law is also commutative, then we say that the group is an "abelian group" or simply "commutative group".

1. Let us start with the first check. Is this representation is associative?:

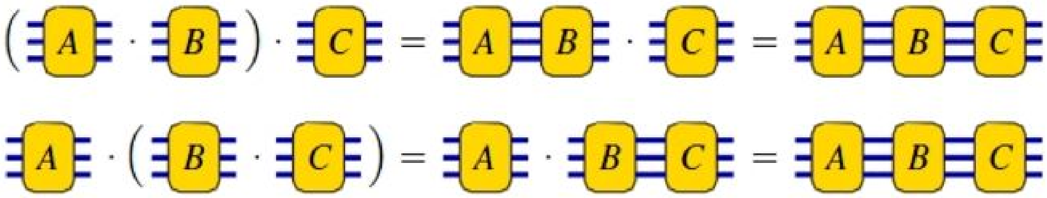


Figure 8.341: **Check of associativity of braids (by the multiplication)**

So the answer is YES regardless of the number of strands!

2. Is this representation commutative?:

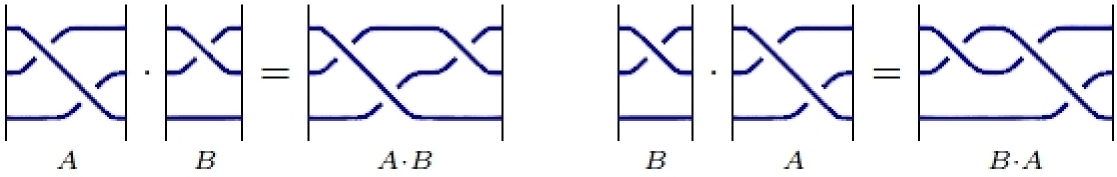


Figure 8.342: **Control of the commutativity of braids (by multiplication)**

The answer is therefore NO beyond 2 strands (because with 2 strands it is commutative)!

3. Does it admits a neutral element?:

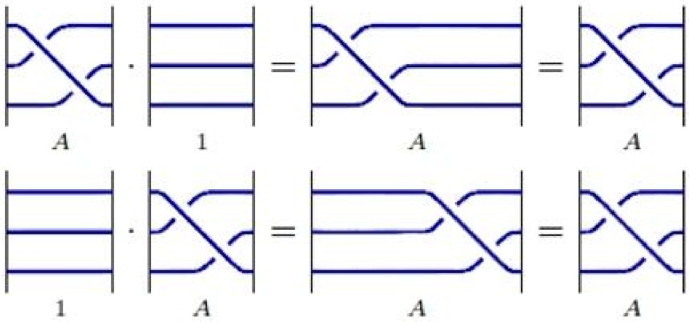


Figure 8.343: **Check of the existence of a neutral element for braids (by the multiplication)**

The answer is therefore YES whatever the number of strands!

4. Are there inverse elements (symmetrical)?:

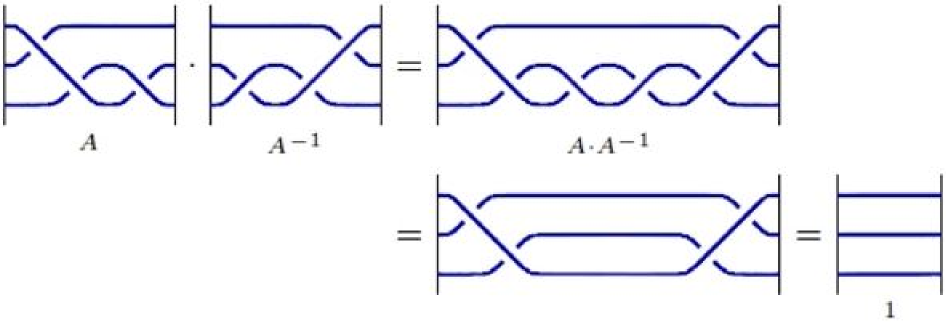


Figure 8.344: Check of the existence of an inverse element for braids (by the multiplication)

The answer is therefore YES whatever the number of strands!

Therefore the braids with n strands form a non-commutative group denoted by $(\mathcal{B}_n, \cdot)$ where the stylized B represent the word "braids".

Let us observe something interesting about the braids with two double strands. If we index their tendrils by integer numbers as below (imagine the braid from the middle of which you rotate the ends, it gives all other left and right tendrils):

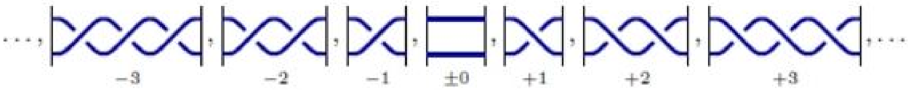


Figure 8.345: Indexing braids tendrils

with the convention of positive tendril v and negative tendril v:

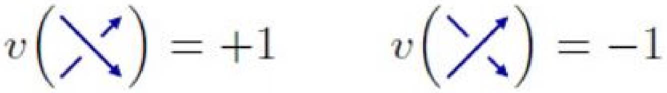


Figure 8.346: Indexation convention

Beware to the following trap! (try to find it ... it is not always easy at first glance):

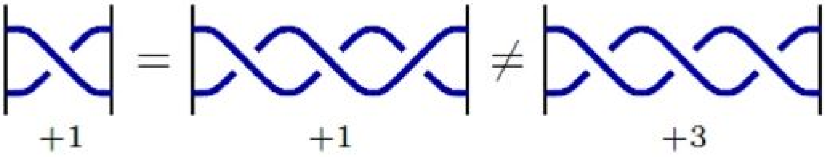


Figure 8.347: Indexation tendril trap

Indeed, the right crossing above of the second braid is precisely not .... a cross! Then we say that the number of crossing is not an invariant. It is for this reason that what matters is the "tendril" because it does not change in the movements of the braid at the opposite of a simply cross!!!

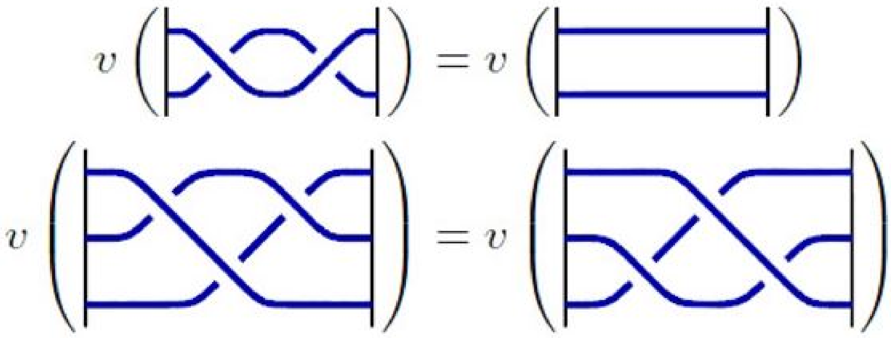


Figure 8.348: **Difference between tendril and cross**

We then notice that regarding to the multiplication of braids (see section Set Theory page 441) and their indexing under the form of their number of tindrels, we have the following properties:

$$f(A \cdot B) = f(A) + f(B) \quad \forall A, B \in \mathcal{B}_2 \qquad (8.9.2)$$
$$f(1_A) = 1_{\mathbb{Z}} \qquad (8.9.3)$$
$$f(A^{-1}) = (f(A))^{-1} \quad \forall A \in \mathcal{B}_2 \qquad (8.9.4)$$

So the multiplicative group of braid with 2 strands is a homomorphism of group with the addition law of the group of relative integers. Moreover, we see that the application f is trivially bijective. We have therefore a isomorphism of group! Beyond 3 strands, we saw that the group was not more commutative!

8.9.2 Knot Representation

Let us return to the first figure presented above:

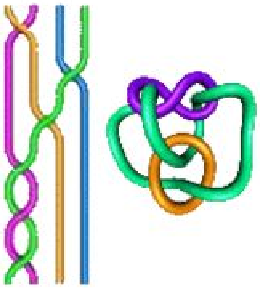


Figure 8.349: **Braid and corresponding string (node diagram)**

A first difficulty then appears for the image on the right (which is a "knot diagram"): a knot (a chain with a single strand) is a three-dimensional geometric object, and we are using a plane representation obtained by projecting the knot onto a plane. This could therefore complicate things... we will see!

Remark

Any node or any chain can be obtained from a braid (without proof).

There exists many possible views, and different appearances of the same node. Two nodes of different projections are distinct (not isotopes)? In the first published table of node classification (that of Tait and Little) the minimum number of crossings is used as a classification principle (up to ten crossings) and it took almost a century to detect one duplication (see table further below): two identical nodes were taken for different.

In practice a node rather looks like this:

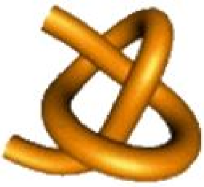


Figure 8.350: **Example of a node in real life**

But mathematicians have the usage to connect both ends of the string to get this representation of a node:

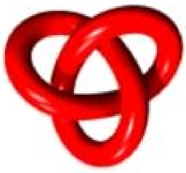


Figure 8.351: **Knot representation by mathematicians**

therefore a node is always a loop that closes on itself. In other words, an open node is always equivalent to a closed node:

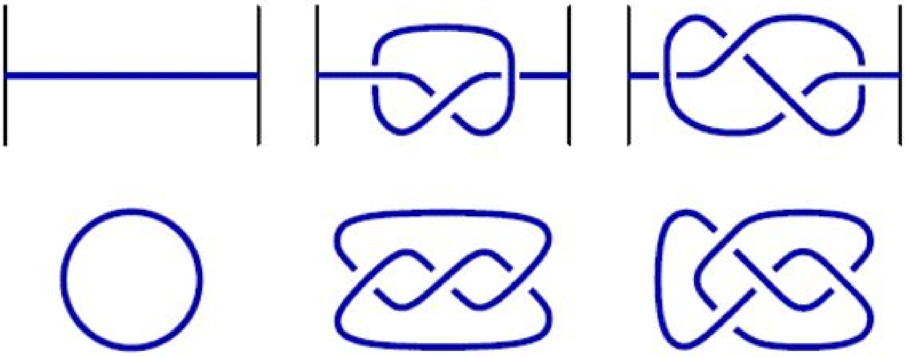


Figure 8.352: **Equivalence between an open and a closed knot**

A closed knot may look different depending on the angle from which we look at it. Thus, the two nodes below are two representations of the same type of knot named the "trefoil knot":

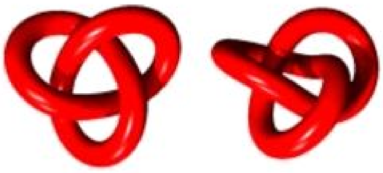


Figure 8.353: **2 different perspectives of the same trefoil knot**

Remark

The trefoil knot has a simple parametric equation given by:

$$\begin{aligned} x &= -\cos(t) + 2\cos(2t) \\ y &= \sin(t) + 2\sin(2t) \end{aligned} \tag{8.9.5}$$

That leads to:

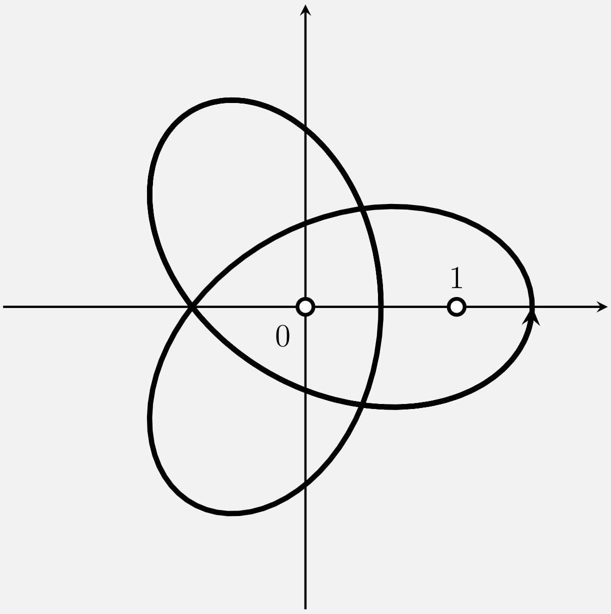


Figure 8.354: **Trefoil plot**

We can also ask ourselves if the left clover node and its mirror on the right are the same?:

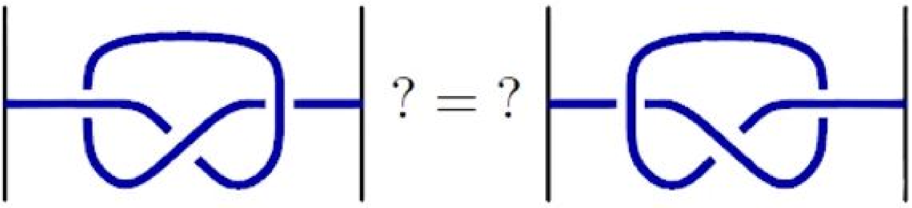


Figure 8.355: **Searching for equality between mirror symmetry of the clover node**

or the same question but when the two knots are closed:

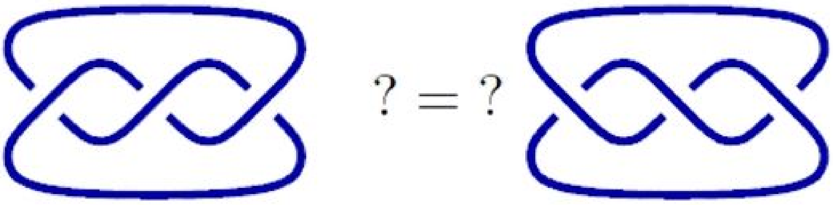


Figure 8.356: **Same situation but by closing the knot**

More difficult, K. A. Perko has showed that the two knots below, named "Perko pair", are in fact the same knot:

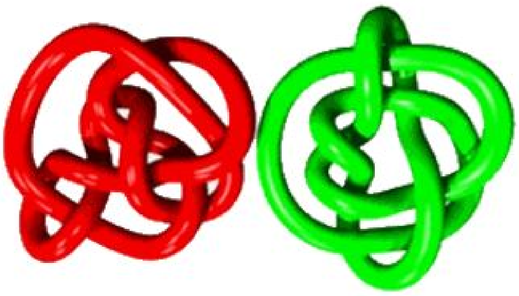


Figure 8.357: **Example of a Perko pair**

Another extreme (and funny) example are the Ochiai, Goeritz and Thistlethwaite knots below are in reality a simple unknot (!!!):

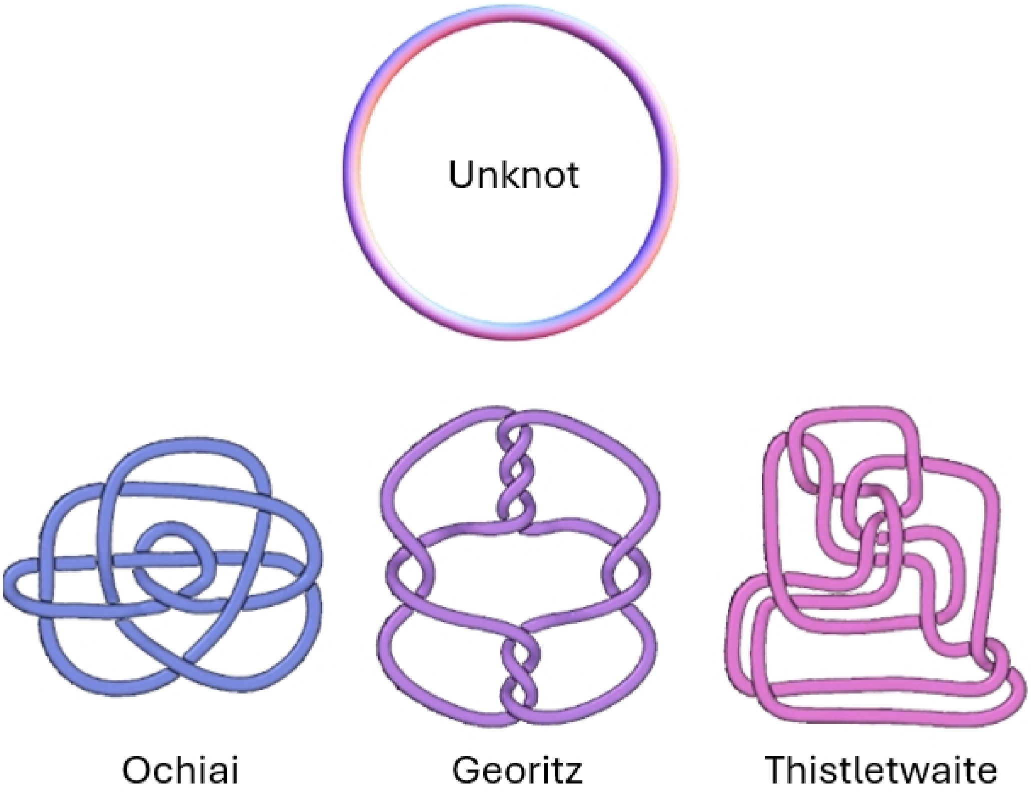


Figure 8.358: **Knots that are simply an unknot**

It is therefore not trivial to know two physical objects represents basically the same knot. But, that's what interests mathematicians! Specifically, they want to classify the knots, that is to say determining all types of nodes that are fundamentally different, and not just in appearance (it's the same idea than in topology where mathematicians successfully proved that every volume could be reduced to three primary volumes).

8.9.2.1 Knots Group

Just as we did for braids, let us look if knots form a group?

So we define the multiplication of two knots as the operation that is schematically:

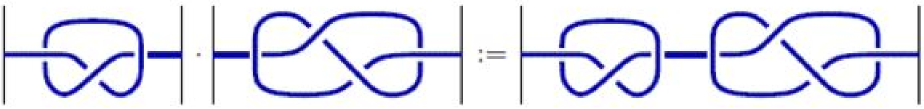


Figure 8.359: Refresh of the definition of the multiplication of two knots

We can ask ourselves if this is a commutative type group?

1. Let us start with the first check. Is this representation associative ?:

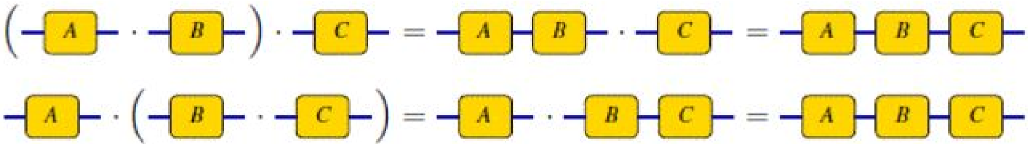


Figure 8.360: Check of the commutativity of knots (by the multiplication)

   The answer is YES whatever the number of knots!

2. Second control: does it have a neutral element?:

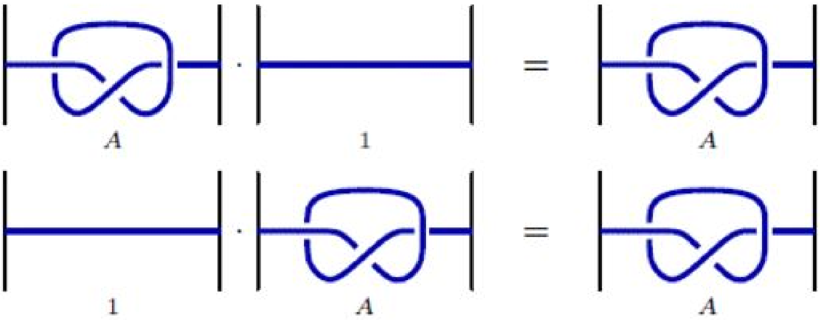


Figure 8.361: Check of the existence of a neutral element for the knots (by the multiplication)

   The answer is YES whatever the number of knots!

3. Third check: is it commutative?:

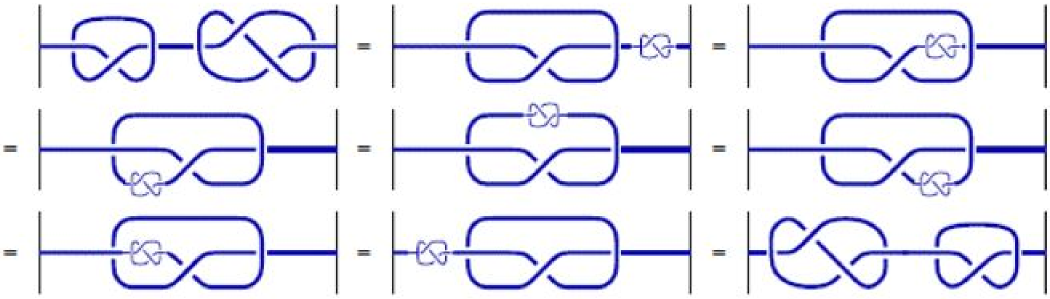


Figure 8.362: **Check of the commutativity of knots (by the multiplication)**

So unlike braids the answer is YES!

4. Fourth and last check: Are there inverse elements (symmetrical)?:

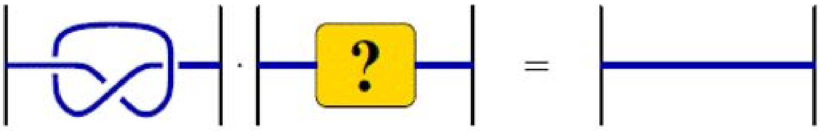


Figure 8.363: **Check of the existence of an inverse element for the knots (by the multiplication)**

and that's the one million dollar question!

With time and the efforts of mathematicians, three essential mathematical perspectives for studying knots emerged to solve different problems:

1. Topological, where the node is conceived as the union of a finite number of closed curves, define close to a given deformation.
2. Algebraic, where there count the intersections for example, and where we associate the groups to the nodes.
3. Geometrical, where we take into account the shape of the knot, measuring lengths or angles. In particular, the idea of the number of rotations appears as twisting and number of interlaces, and these concepts are fundamental to the study of DNA in molecular biology.

The link between these views is tricky. There are evidences hard to prove rigorously. We think especially to the Jordan theorem asserting that a closed flat curve without crossing defines an interior and an exterior! It took almost two centuries to properly define the concept of curve and there are wild knots to eliminate before any classification. We must take care to properly define the deformation of a knot, that mathematician named "isotopies".

The most effective mathematical idea for the study of the knots is probably that of "knot invariant". An invariant of a node is a characteristic (integer number, real number, polynomial, group, etc.) which remains unchanged during deformation. If we have/found an invariant, we can say that two knots are really different when the invariant does not take the same value for both knots. But if two knots have the same invariant, we can not say they are the same type (deformable into each other). A typical example of two knots having the same invariant and for which it is not trivial to say whether they are the isotopes are the right and the left trefoil knots when the chosen invariant is the number of crossings denoted by $c(D)$. We should for this purpose have a complete system of invariants.

Here, invariants are indispensable, and mathematicians have come up with dozens that distils different features of knots. But these invariants tend to have blind spots.

Take, for example, an invariant named "tricolorability". A knot diagram is tricolorable if there's a way to color its strands red, blue and green so that at every crossing, the three strands that meet are either all the same color or all different colors. Mathematicians have shown that even when you move the strands of a knot around, its tricolorability (or lack thereof) is unchanged. In other words, tricolorability is an innate feature of a knot.

The trefoil is tricolorable:

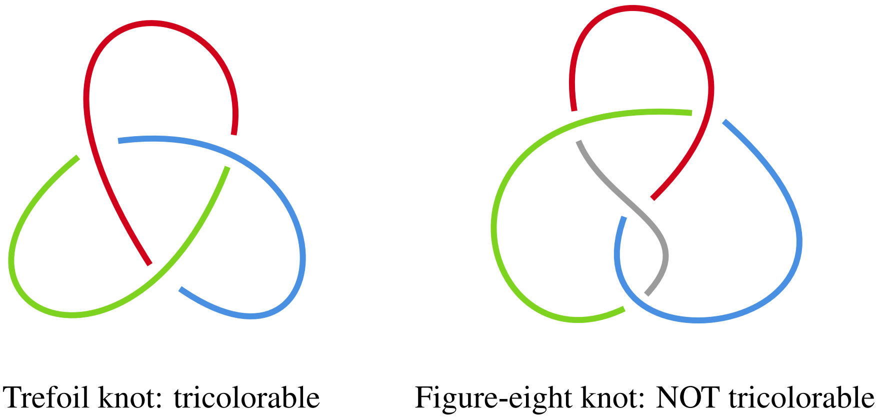

Trefoil knot: tricolorable
Figure-eight knot: NOT tricolorable

Figure 8.364: **Tricolorability of trefoil knot (source: Quanta Magazine)**

But the "unknot" (a loop that has no actual knots, even if it appears tangled) is not tricolorable, providing an instant proof that the trefoil is not just the unknot in disguise. But while tricolorability enables us to distinguish some knots from the unknot, it's not a perfect tool for this purpose: Knots that are tricolorable are definitely knotted, but knots that aren't tricolorable aren't definitely unknotted. For instance, the figure-eight knot is not tricolorable, but it is genuinely knotted. This knot falls into tricolorability's blind spot — it's as if the invariant is saying: "The figure-eight knot is unknotted as far as I can tell".

The "Conway knot", an ill-crossing knot discovered by John Horton Conway more than fifty years ago, is extraordinarily skilled at fooling knot invariants!

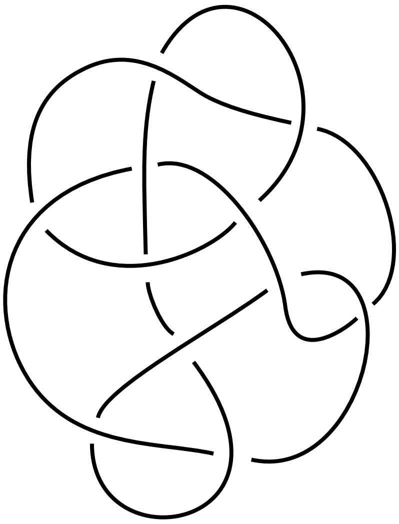


Figure 8.365: **Conway knot**

The Russian mathematician Victor Vasiliev introduced in 11990 (holocene calendar) a new class of invariants. It remains to make them explicitly computable and prove that they form a complete system.

Henri Poincaré introduced in 11900 (holocene calendar) the concept of fundamental group of a space which describes the possibilities of paths with return to starting point. Applied to the outer space to a node, this provides the "knot group" and the "Alexander polynomial" linked to it (see below the mathematical formalization).

The same construction applied to what is named the "configuration space" provides an effective definition of "braid groups". These groups were introduced in an intuitive form, circa 11920 (holocene calendar), by the Viennese mathematician Emil Artin, one of the fathers of modern algebra. Since 11937 (holocene calendar), the Russian mathematician Markov connected knots and braids, and gave a theoretical method to define, with braids, knots invariants.

At the beginning of years 11990 (holocene calendar), braid groups were a curiosity and their complexity repulsive. Then, suddenly, they have become a central theme of scientific research. Let us give an idea of the diversity of views that lead to braid groups:

1. In geometry, the Russian mathematician Vladimir Arnold classified under the name of "disasters" singularities of geometric configurations, which leads to the configuration spaces.
2. In algebra, specifically in group theory, two avatars of braid groups play a central role: the Coxeter's groups and the Hecke's algebras.
3. In statistical mechanics, the study of exactly soluble models is done through the use of Yang-Baxter relations and quantum groups. The link with the braid groups is profound.
4. The two-dimensional physics has recently taken great importance and we expect from it

a high temperature model of superconductivity. The usual classification of fermions and bosons particles is complicated in two dimensions. The new concept is that of anyone whose mathematical model is linked to braid groups.

5. Quantum field theory is the mathematical model of elementary particles. Edward Witten has made the bridge between this theory and braid groups, via the Jones polynomial.

The Knots Theory is therefore an active interface between physics and mathematics. Knots and braids now provide an effective modernization tool for the physic of polymers to the liquid crystal, through molecular biology. In the opposite direction, imported ideas of physics have sparked a revolution in mathematics: to somewhat marginal issue at the beginning of the years 11990 (holocene calendar), the Knots Theory at the beginning of the years 12000 (holocene calendar) one of the greatest mathematical subject.

8.9.3 Tait's Knot

The main aim of the knot theory is to clarify all knots and has its origin in physics in the early 119th century (holocene calendar) as we have already mentioned. Why?

Recall that at that time atoms were a mystery: why they seemed indestructible and why they exist in so many varieties and can be combined to give countless other components?

At that time the most beautiful equations of physics (which are often good candidates to explain things we do not understand...) were the Maxwell equations, then it was natural (or attempting) for physicists to try to explain the atomic mechanics in terms of electromagnetism, although we now know today that this path was destined to fail. Later in the mid 119th century (holocene calendar), electromagnetic waves were largely conceptualized as vibrations of a medium named at the time "phosphor ether". The repository of this ether was then defined (or at least suspected) as an absolute reference... But later, the experimenters Michelson and Morley (see [390]) showed that the relative motion of our planet in this ether was undetectable (and worse... they measured the absolute constancy of the speed of light as did De Sitter later with another type of experiment as described in his paper [133]!). Their experiences also led Albert Einstein and Henri Poincaré to develop their famous theory of Special Relativity (see section Special Relativity page 5627):

$$\vec{\nabla}\circ\vec{E}=0 \qquad \vec{\nabla}\times\vec{E}=\vec{0} \tag{8.9.6}$$

which is, we must admit, similar to fluid mechanics (see section of the same name page 3770) when we have an (assumed...) incompressible fluid without viscosity and without vortices:

$$\vec{\nabla}\circ\vec{v}=0 \qquad \vec{\nabla}\times\vec{v}=\vec{0} \tag{8.9.7}$$

More generally, if the rotational $\vec{\nabla}\times\vec{v}$ is not zero, Helmholtz proved in 11858 (holocene calendar) that the vortex field lines - defined by the lines of $\vec{\nabla}\times\vec{v}$ - move in the direction of $\vec{v}$ as if they had an independent existence (and here this is the open breach ... of course!). These lines could have no end but could form loops.

In 11867 (holocene calendar) the physicist Peter Guthrie Tait (assistant of Hamilton and a champion of quaternions) founded an ingenious way to prove this effect by cutting a circular hole in a box, filling it with smoke, and then expelling smoke afterwards through compression of the air in the box thereby forming smoke circles. It also showed this to his friend Kelvin that noted the analogy with electromagnetism and proposed a theory in which the atoms were vortex (knots) in ether! He hypothesized that the different types of atoms correspond to different types of knotted vortices (yes it's a bit farfetched but...)!

Tait subsequently tried to classify (see table below) tied lines in agreement with:

1. The number of interlaces when they are projected on a plane
2. By representing only the "main nodes"

Definition (#8.9.400): A node can be complicated because it is the succession of simple knots:

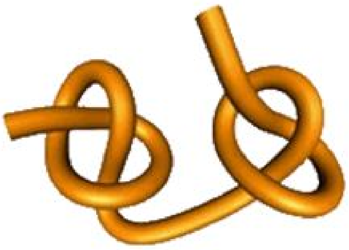


Figure 8.366: **Example of a succession of knots on an open knot**

These simple knots (with one strand) can be separated by cutting the string:

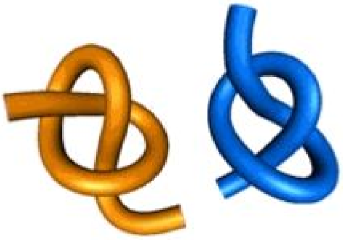


Figure 8.367: **Separation of the knot**

A main knot is a knot that can not be separated into simpler knots: cut the string unties the knot.

Classify knots this is trying to determine the building blocks: all main knots. The list Tait got initially was the following nodes (in the hope of obtaining the periodic table of elements but "ether" version...) and if necessary we recommend you to bring a string to ensure they are main one (sometimes it is mentally difficult):

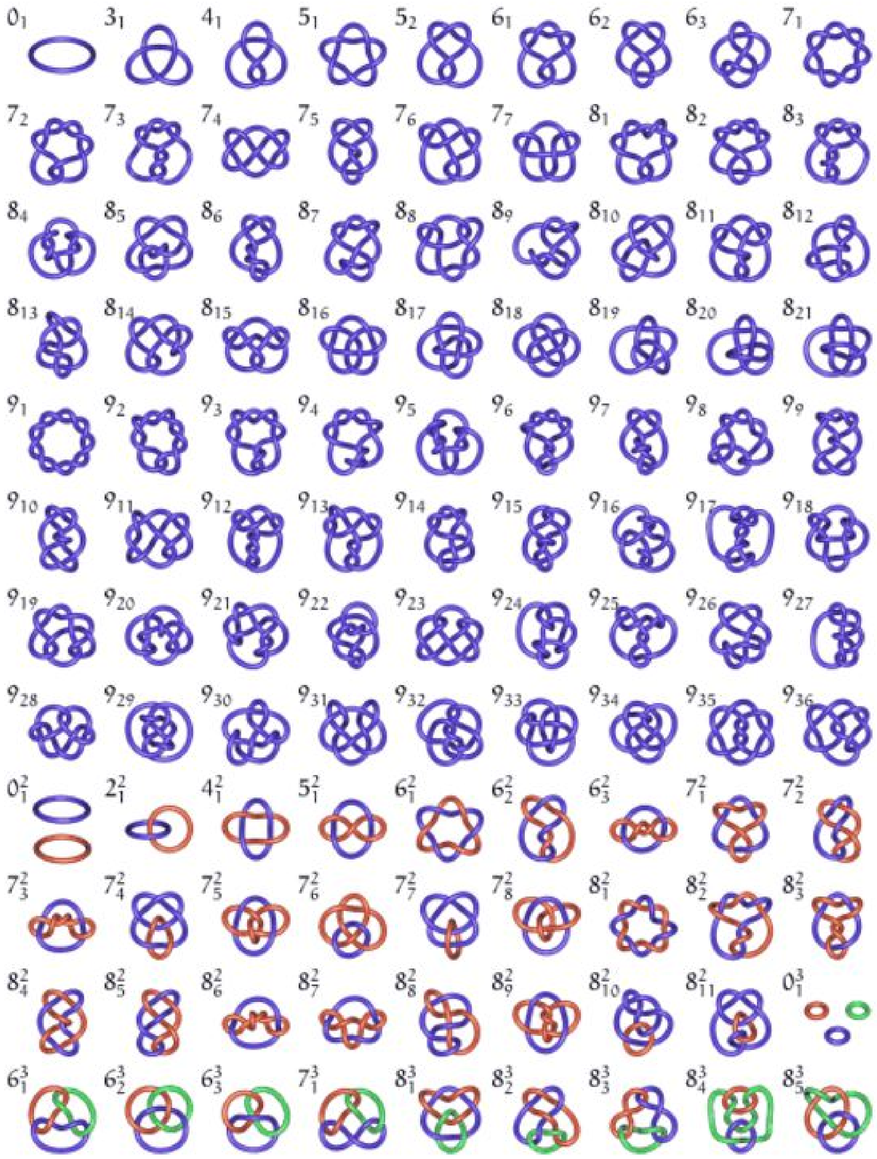

0_1
3_1
4_1
5_1
5_2
6_1
6_2
6_3
7_1
7_2
7_3
7_4
7_5
7_6
7_7
8_1
8_2
8_3
8_4
8_5
8_6
8_7
8_8
8_9
8_{10}
8_{11}
8_{12}
8_{13}
8_{14}
8_{15}
8_{16}
8_{17}
8_{18}
8_{19}
8_{20}
8_{21}
9_1
9_2
9_3
9_4
9_5
9_6
9_7
9_8
9_9
9_{10}
9_{11}
9_{12}
9_{13}
9_{14}
9_{15}
9_{16}
9_{17}
9_{18}
9_{19}
9_{20}
9_{21}
9_{22}
9_{23}
9_{24}
9_{25}
9_{26}
9_{27}
9_{28}
9_{29}
9_{30}
9_{31}
9_{32}
9_{33}
9_{34}
9_{35}
9_{36}
0_1^2
2_1^2
4_1^2
5_1^2
6_1^2
6_2^2
6_3^2
7_1^2
7_2^2
7_3^2
7_4^2
7_5^2
7_6^2
7_7^2
7_8^2
8_1^2
8_2^2
8_3^2
8_4^2
8_5^2
8_6^2
8_7^2
8_8^2
8_9^2
8_{10}^2
8_{11}^2
0_1^3
6_1^3
6_2^3
6_3^3
7_1^3
8_1^3
8_2^3
8_3^3
8_4^3
8_5^3

Figure 8.368: Tait's classification

Remarks

R1. This table is very important and we will very often make reference to it for application examples using the proposed nomenclature.

R2. The corresponding values in the above table, named "knot complexity measures", are noted $c(D)_{\text{Id}}^{c(B)}$ in generality where $c(D)$ is the number of crossings, $c(B)$ the number of strands, and Id the number id of the node in the class $c(D)$.

The beauty of this model of "vortex atoms" was the fact that it was in agreement with the continuous appearance of the wonderful world of fluids, or to the Maxwell's equations by extension, to discretize the different types of atoms. A difficulty with this theory, however, was the remarkable stability of atoms. Kelvin admitted in 11905 (holocene calendar) after many years of failure in trying to prove that the movement Helmholtz circles was stable, the conclusion was that these circles were essentially unstable and should therefore dissipate. Curiously it is the extreme stability of the atoms that was one of the puzzle pieces to build the corpuscular quantum physics.

Moreover, with the theory of Special Relativity, the concept of ether that was precious to Maxwell and his contemporaries, especially regarding the Knots Theory, became synonymous of a useless concept and with the addition of the theoretical quantum theory vortex atoms was completely forgotten. Knot theory resurfaced because of some conjecture that Tait was not able to proved (we will come back about this later) that were proved only in the 11980s (holocene calendar) in a hazardous turn in theoretical physics.

What physicists abandoned intrigue mathematicians. The basic question remains the same: how can we say that two nodes are "isotopies" of each other (we will define further below what it is exactly). This question is closely related to the famous conjectures of Tait. To attack these conjectures and the basic problem of knots resemblance, the topologists developed invariants knots. An example of well known knot invariant and that have had a great success are the Alexander polynomial discovered by J. W. Alexander in 11927 according to holocene calendar (see further below). Thus, if the polynomials of two nodes are different, these are then not isotopies. Unfortunately, there are some knots having polynomials Alexander that are equivalents which are not isotopies of each other...

The mathematical knot theory then developed during fifty years was a bit forgotten until the thunderclap of Jones in 11984 (holocene calendar) of a new knot invariant (the Jones polynomial). The discovery of Jones is pretty exemplar of the scientific point of view and can lead to meditation on the current organization of science. Jones was not at all an expert on knots. He was interested in the classification of factors in von Neumann algebra (functional analysis). He obtained matrices algebras whose commutation relations (Yang-Baxter equations) were close relations of the braid group. From braids to knots, there is only one step that Jones has taken with the help of Joan Birman which is a specialist in knots.

Then we are see an explosion of discoveries: purely combinatorial version, new polynomial, etc. In 11989 (holocene calendar), Witten proved that the Jones polynomial can be obtained from quantum field theory by thanks to a Feynman integral, giving the first definition that does not use the flat projections of a knot. Somehow the Jones-Witten theory is a non-commutative extension of the work of Gauss. The Lie group (see section Set Algebra page 1417) which acts in magnetism is $\mathrm{U}(1)$, while the Witten invariant is a Feynman integral over a space of $\mathrm{SU}(2)$-connections.

8.9.4 Mathematical Formalisation

A node is mathematically modelled by an injective application, differentiable and whose derivative does not vanish, of the circle in oriented three-dimensional space (trivial knot).

The two central problems of the theory of knots is to decide in a calculable way if the knot is trivial (can be discard without cutting the string) or not (this problem is in this early 121st century (holocene calendar) not resolved as far as we know) and if two nodes are really equivalents.

The first type of problem can be well represented on the fact whether the knot below is knotted or not...?:

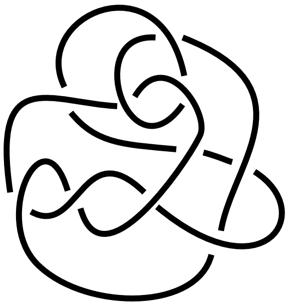


Figure 8.369: **Knotted knot?**

The answer is No! As shown in the figure below (read from left to right and top to bottom):

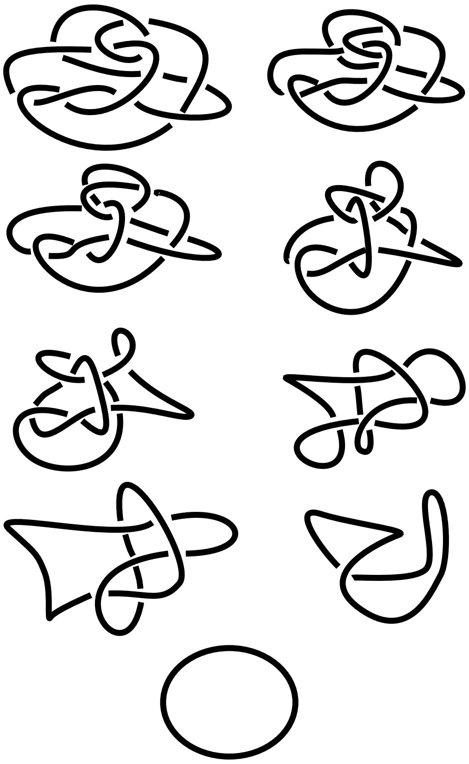


Figure 8.370: **It is not...!**

The goal for the second problem is to combine the knots computable mathematical objects (polynomials, numbers) named "knot invariants" and that are insensitive to the deformation of a knot. If the invariant is not equal to that the trivial knot that is:

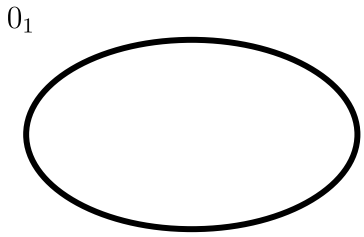


Figure 8.371: **Trivial Knot**

we are therefore sure that the knot is not trivial. And for example, the simplest non-trivial knot is the trefoil knot:

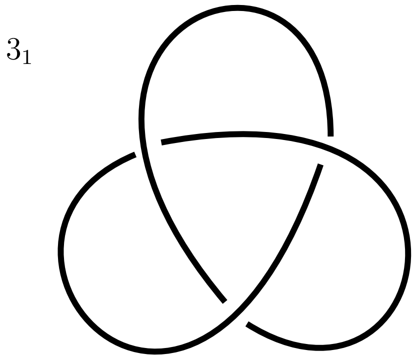


Figure 8.372: **Trefoil Knot**

The problem is finding enough fine tuning invariants. We will describe two of them:

1. The number of interlacing (of two knots), idea of Gauss, and which occurs in electromagnetism
2. The Jones polynomial (introduced in 11985 by Vaughan Jones, 11990 Fields Medal according to holocene calendar), which is subtle enough to distinguish for example right of the left trefoil knot.

We will also describe a general class of invariants: finite type invariants or Vassiliev invariants. These invariants defined in a rather unconstructive way are maybe complete invariants, but we do not know for sure at this beginning of the 121st century (holocene calendar). We will describe the number of interlacing of 2 knots as a combinatorial invariant calculable from a node diagram. We will then write the classical integral formula related to magnetism (Gauss) to calculate it. We will then do a little detour by the global differential geometry of curves in three-dimensional space to show the "White formula" that connects three geometric invariants associated with a ribbon. We will then describe the new polynomial invariant in a combinatorial point of view. The point of view of Feynman integrals will be discussed.

Victor Vasiliev has introduced a general family of invariants that contains most known invariants and we can say that they are of finite type. We will describe the principle.

In biology the RNA and DNA strand and also amino acid filaments are wound in complex three-dimensional shapes (these are closed braids, as we know a more general case of knots). However, often, through the microscope, we see only a two-dimensional projection. Invariants give the possibility to trace back to three-dimensional information from 2D views we have:

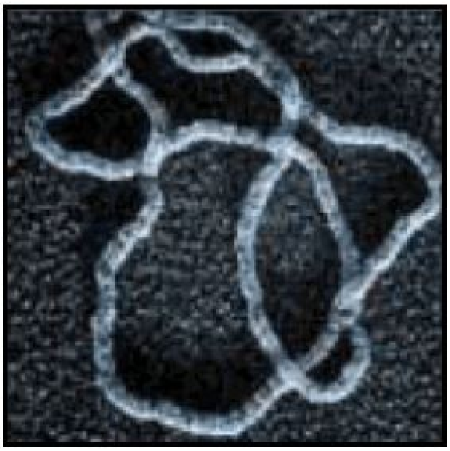


Figure 8.373: **Example of RNA or DNA strand**

On the other hand, biologists have observed knotted DNA molecules and found that the topological nature of the DNA molecule, that is to say the knot type formed by the molecule, affects its operational mechanism in the cells by conditioning some of its chemical properties. Some viruses attack cells by changing the long DNA molecules by interlacing them in different ways. Indeed, by the means of enzymes named topoisamérases, viruses cut and reattach different strands of the DNA molecule such that they take the form of a knot which can be very complex. It turns out that the type of knot obtained is in a way the virus signature. To fight effectively against the virus, it is imperative to recognize their signature by their action on DNA. Therefore to identify different viruses we must be able to recognize the different types of knots and this is where the theory of knots can help the biologist.

Remark

The identification of some knot type has been transformed by research centers in an online gaming which aims it to give a funny way to the people to contribute to find a solution. This seems to work since late 12011 (holocene calendar) some players seems to have made relevant discovery regarding the analysis of proteins.

Let us start now with some definitions. We have deliberately chosen to not to list a given number of definitions (as many in Graph Theory...) that the reader can easily find in literature or on the

Internet. We allow ourselves to omit them in the sense that they will not be helpful in the application of knot theory in quantum field physics (and except for those who like make small drawings... they are useless in our point of view).

Definitions (#8.9.401):

D1. A knot may be defined (as there are several possible definitions and some have small enough bothersome uncertainties...) by the image of the circle denoted by S^1 by a continuous application (the string is assumed as such), injective (this avoiding the string enters in itself):

$$f : S^1 \mapsto \mathbb{R}^3 \tag{8.9.8}$$

in other words, it is a curve without double points, drawn in Euclidean space of three dimensions.

A knot is therefore represented by an injective application $f \in \mathcal{C}^1([0,1],\mathbb{R}^3)$ (impose $\mathcal{C}^1$ class knots avoids having too wild curves ...) verifying that $f(0)=f(1)$ (closed knot). The image of f is sometimes named "support of the knot f": this is the "physical" realization of the knot in space.

Remark

All knots will subsequently denoted by the letter N.

To summarize crudely..., a knot is a string we for which we welded the ends.

D2. We say that a knot is a "trivial knot" if the application f that defines it extends in a continuous application of the disc $D^2 \mapsto \mathbb{R}^3$ and always injective: a trivial knot is therefore a knot that border an immersed disc of $\mathbb{R}^3$.

D3. Two knots γ_1,γ_2 are equivalent if there exists a continuous application:

$$F : [1,2] \mapsto N \tag{8.9.9}$$

such as $F(1)=\gamma_1$ and $F(2)=\gamma_2$. It remains to find F that is a curve in the set of curves.

D4. The $c(K)$ of a knot K is the natural integer representing the minimum number of crossings for every diagram of a knot type (that is a natural measure of complexity).

✍**Example:**

The knot 0_1 has a $c(K)=c(0_1)=0$ therefore null. It seems there exist no knots with $c(K)$ equal to the unit. A proof consist is to list all possible diagrams with one or two cross and see that they are in fact equivalent knots of the type 0_1 or just interlaces. The trefoil knot (knot 3_1) has a $c(K)=c(3_1)=3$.

D5. The mirror image $\bar{K}$ of a knot K is obtained by reflection on a plane in $\mathbb{R}^3$. The knot K can be constructed by inversion of crossing of the knot diagram:

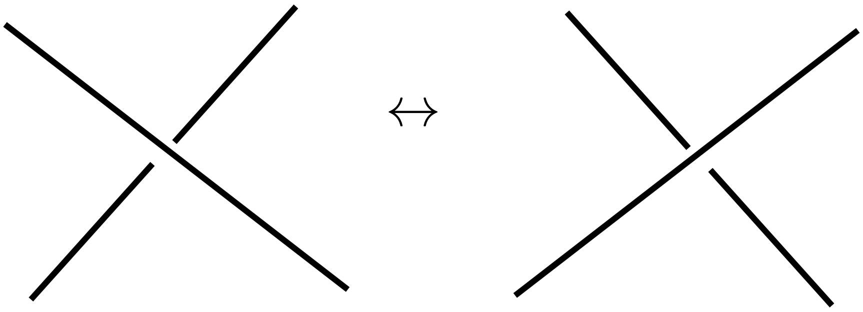


Figure 8.374: **Mirror image by crosses inversion**

✍**Example:**

With the trefoil knot ("left trefoil" and "right trefoil"):

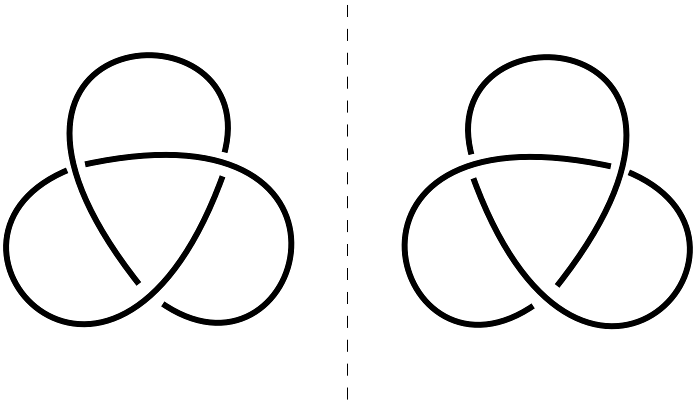


Figure 8.375: **Mirror of trefoil knot**

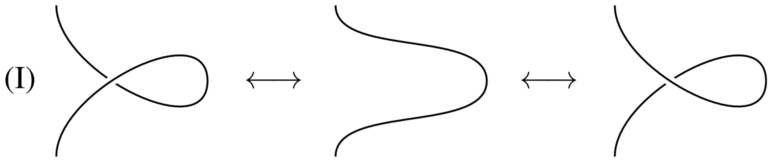
Remark

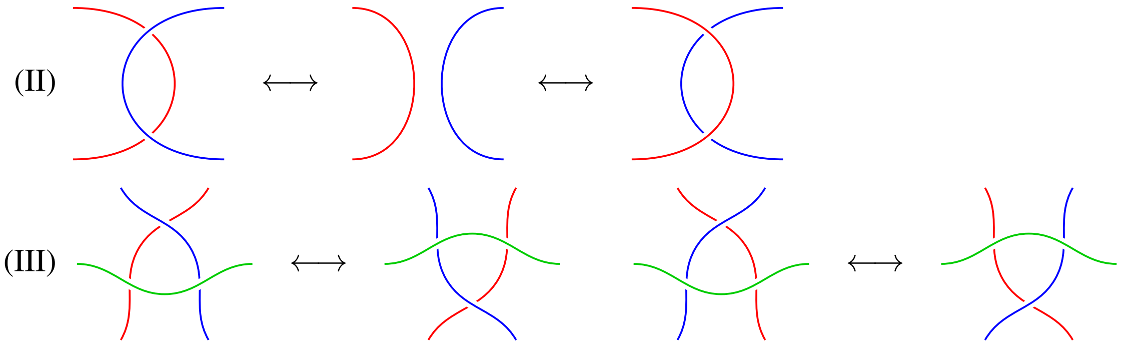
We also notice that in the Tait classification table, the mirrors knots are not represented!

D6. Two knots are say to be "isotopes knots" if we can switch from one to the other by continuous manipulation named "isotopies" (the left trefoil knot and right trefoil knot are for example not isotopes!) and as we know this is the problem number one in the knot theory: detect isotopes knots.

In other words, an isotopie is a movement which does not deforms/change the knot.

The isotopies give in the plane three particular types of movements named "Reidemeister movements":





It is therefore well three simple operations for changing a part of a knot without changing the nature of the knot itself.

So two diagrams or knots are isotopes, if we can pass from one to the other by a finite sequence of Reidemeister movements. Thus, with the example below, we show that the initial knot is equivalent to the trivial knot:

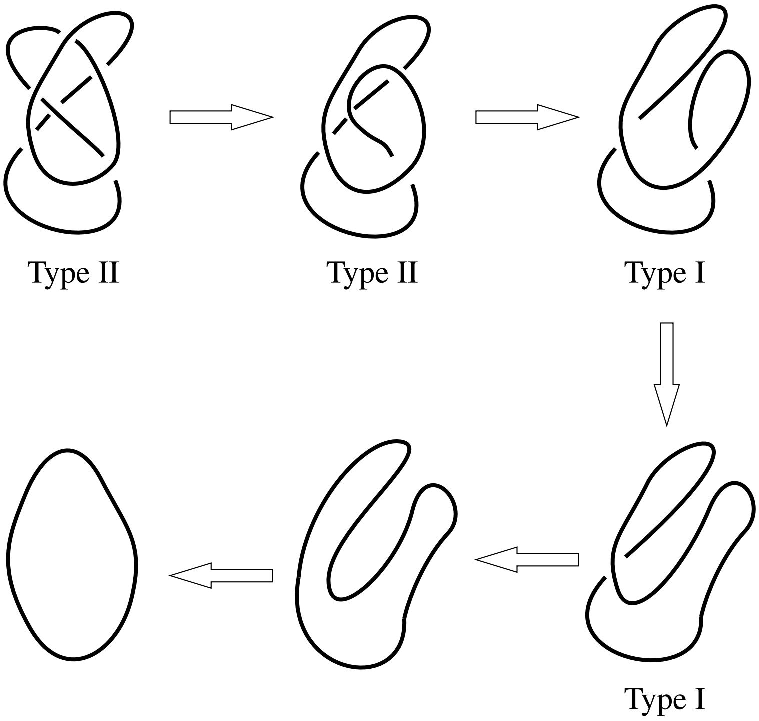


Figure 8.376: **Simplistic application example of the Reidemeister movements**

D7. Two knots are say to be "equivalent knot" or "amphicheiral". if they are isotopes OR if one is isotope to the image of the other in a mirror. According to the above definitions, each knot is necessarily equivalent to its own mirror image, but only reflective knots are isotopes to their mirror image. The eight knot 4_1 is a good example of this kind of knots, which are almost rare:

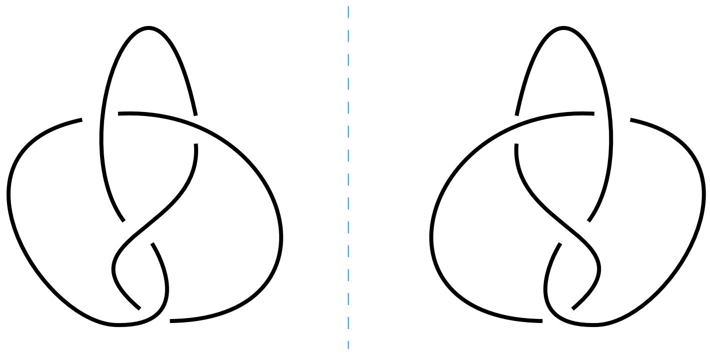
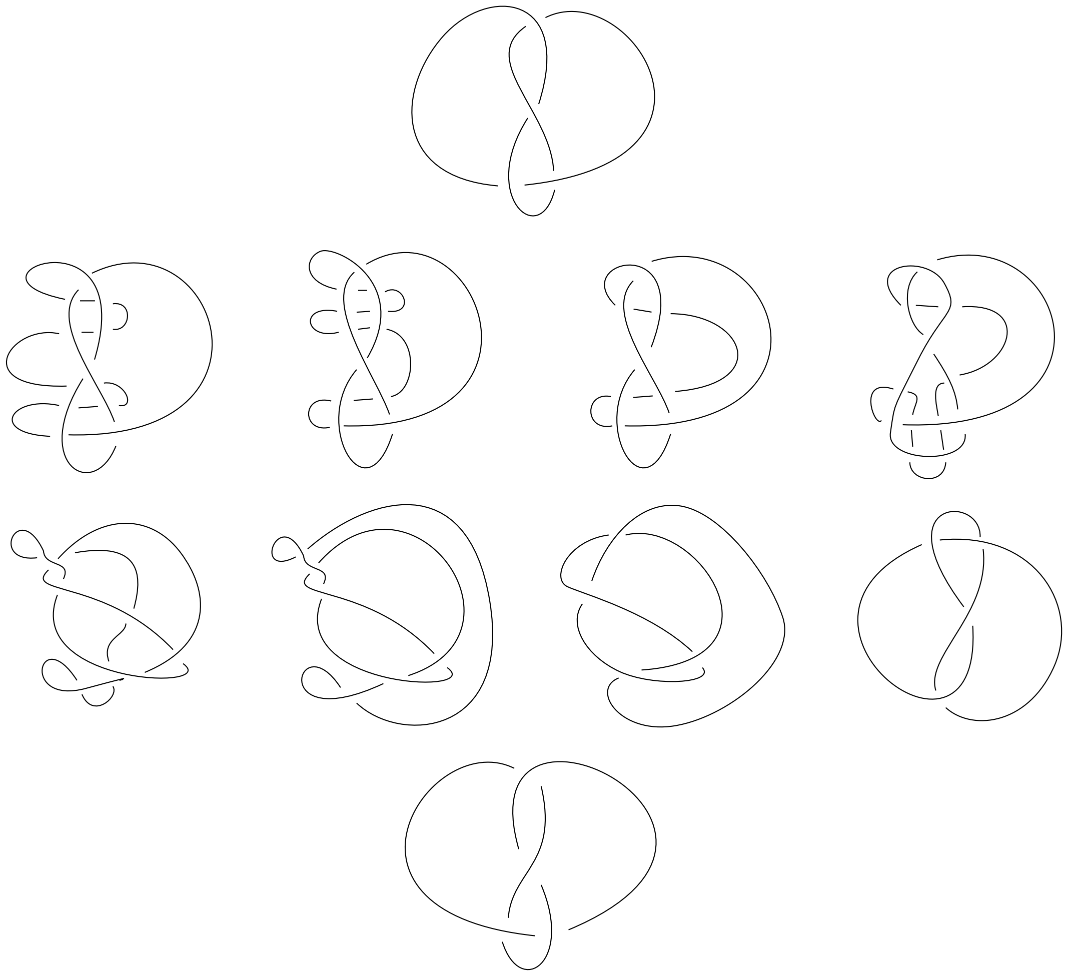


Figure 8.377: **Knot isotopie**

The corresponding Reidemeister movements are not always obvious to guess. Let us show you the details:



D8. An "interlaced" is a subvariety (see section Topology 2475) compact (see section Topology page 2447) of class $\mathcal{C}^{\infty}$ and of dimension 1.

D9. The "number of convex components" is denoted by $c(E)$. If $c(E)=1$ we say that E is a knot.

Most of time, interlaces are oriented (see section Graph Theory page 3012) and we will identify the isotopes interlacing. So we represent the interlacing in the plane by projecting and specifying the type of crossing points.

✍**Example:**

Interlaces with three convex components (left) and trefoil knot (right):

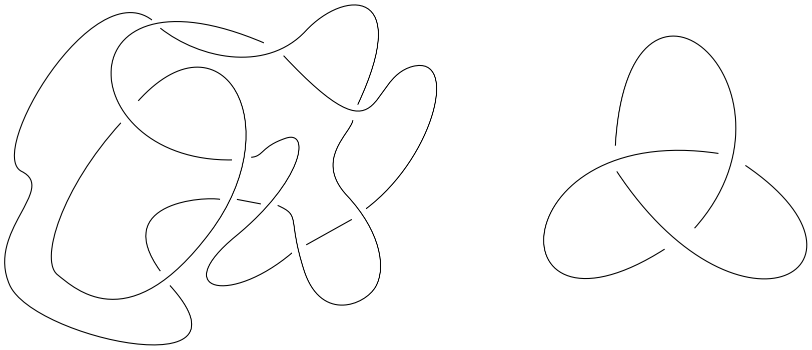


Figure 8.378: **Interlaces with three convex components**

For sure two interlaces are isotopes if and only if we can move from one to the other by a finite sequence of Reidemeister movements.

D10. Three interlaces E_+,E_-,E_0 are named "associates interlaces" if they differ only in a crossing point and that at this point they are in one of the following configurations:

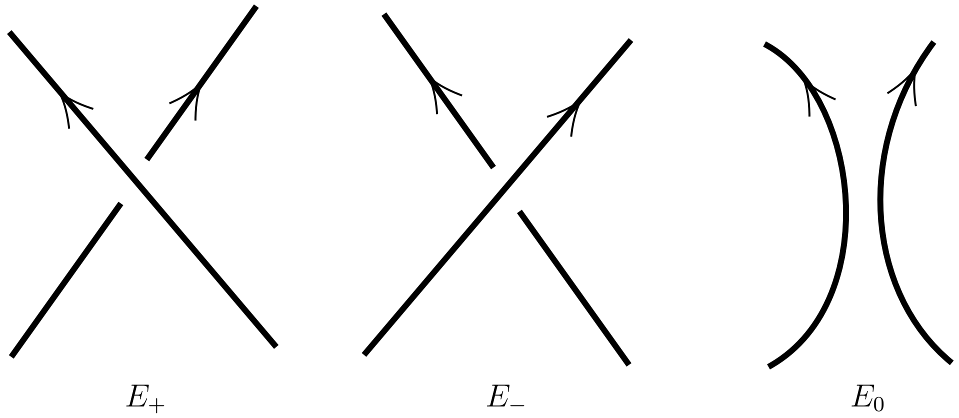


Figure 8.379: **Example of associated interlaces**

We denote by $\mathcal{E}$ the set of all interlacing classes and we will now focus in functions of the type $P:\mathcal{E}\rightarrow A$ where A is a commutative ring (the reader will must forget that we have seen that the coefficients of a polynomial are elements of a ring).

Remark

We must also remember that a knot is a curve and that any curve can be represented by a polynomial, hence the idea...!

D11. P is say to be "invariant by association" (by association interlaces...) if:

$$P(\bigcirc)=1 \tag{8.9.10}$$

and if there exist invertible $a_+,a_-,a_0\in A$ such that for any associated interlacing triplet (E_+,E_-,E_0) we have (!):

$$a_+P(E_+)+a_-P(E_-)+a_0P(E_0)=0 \tag{8.9.11}$$

We can already prove in a fairly basic way that if such a polynomial exists, then it is uniquely determined by the coefficients of the above equality. We summarize this in the following theorem:

Theorem 8.32. If P is invariant by association, then it is uniquely determined by the coefficients a_+,a_-,a_0.

Proof 8.32.1. Let us notice first that:

$$\forall r\in\mathbb{N}\qquad P(\bigcirc^r)=\left(-\frac{a_+a_-}{a_0}\right)^{r-1} \tag{8.9.12}$$

where $\bigcirc^r$ describes the knot consisting of r non-interlaced circles. Indeed, the relation $P(\bigcirc)=1$ and the invariance property of P applied to the following interlaces:

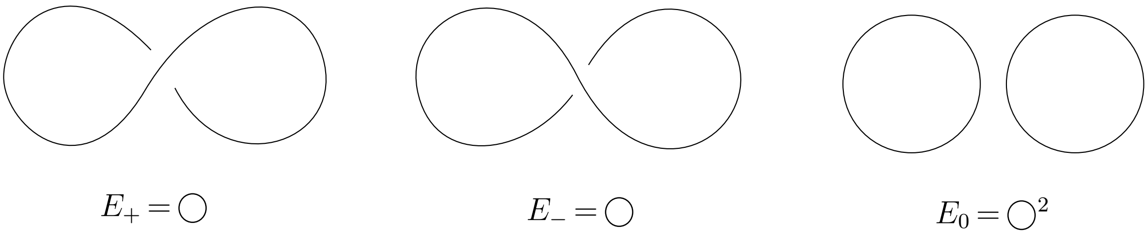


Figure 8.380: **Example of non-interlaces notation**

gives:

$$a_+P(\bigcirc)+a_-P(\bigcirc)+a_0P(\bigcirc^2)=0 \tag{8.9.13}$$

Therefore we get:

$$P(\bigcirc^2)=-\frac{a_+P(\bigcirc)+a_-P(\bigcirc)}{a_0}=-\frac{a_+\cdot 1+a_-\cdot 1}{a_0}=\left(-\frac{a_+a_-}{a_0}\right)^{2-1} \tag{8.9.14}$$

Thus, by induction on r we get:

$$P(\bigcirc^r)=\left(-\frac{a_+a_-}{a_0}\right)^{r-1} \tag{8.9.15}$$

■ Q.E.D.

In other words, the function P can be expressed only by its coefficients!!! So we can now try to see if a knot full of interlaces may always be reduced to $P(\bigcirc)$ recursively.

8.9.4.1 Planar Representation

We assume that the knot is in the oriented three dimensional Euclidean space. We look at projections of a knot on the plane $z = 0$. It is intuitively clear that we can assume that node has no vertical tangent and that the crossing points are only double and transversal. Such screening will be named a "good projection".

We then indicate at each crossing point which is the strand which passes over the other. Such a drawing represents a knit in a non-unambiguous way: we say that we have a "knot graph". Of course, two equivalent knots have good projections that give different knots diagrams in general (herein lies also one of the problems as well). It is therefore important to read the equivalence of two knots directly on their charts.

Here is an example with the trefoil knot:

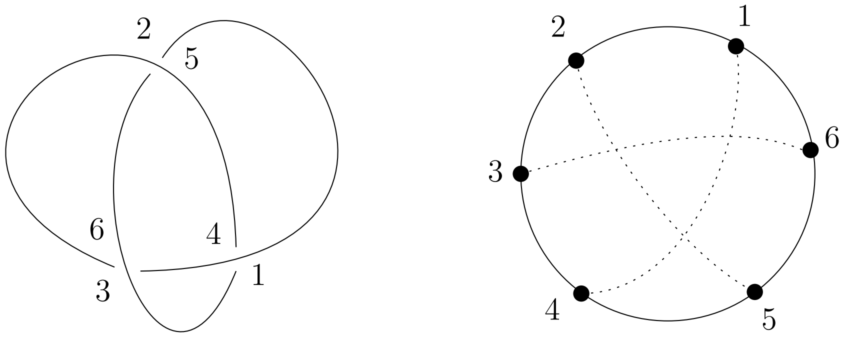


Figure 8.381: **Trefoil knot and its diagram**

> **Remark**
>
> The trefoil knot is the knot having the minimum number of intersections, namely 3; it actually exists in two amphiceiral (images of one another by reflection for refresh...). This is in fact in a simple knot the ends have been welded. The trefoil knot is the border of a Möbius with three half-twists and also the torus knot of order $(3,2)$ (3 windings around the core, on two revolutions), as well as the on of order $(2,3)$.

As a drawing is not necessarily easy to transmit, or to draw on a computer, we can also give a coding diagram of the knot by a matrix with integer coefficients at three lines and whose number of columns equals the number of double points.

We number the $2n$ crossing points on the closed curve (circle) in the order they arrive. We note that the pairs are all formed of an even number and an odd number. We build then the first two rows of the matrix by putting in each column both numbers giving the same projection: the odd on the first row, the odds on the second one. We then add to each column, an ± 1 indicating the orientation of the two strands (oriented) ($+1$ if the bottom one passes through from right to left when we traverse the top one, -1 otherwise). For example, the matrix associated with the trefoil knot of the previous figure is:

$$3_1=\begin{bmatrix}1 & 3 & 5\\ 4 & 6 & 2\\ + & + & +\end{bmatrix} \tag{8.9.16}$$

~15 %

★★★★☆

20 votes, 77.00%

9

Mechanics

Mechanics is the branch of physics that relates to the study of forces and their actions in abstract form. (Larousse)

Contents

9.1 Principia . . . 3068

9.1.1 System of Units . . . 3072

9.1.1.1 Dimensional Analysis . . . 3081

9.1.1.2 Scientific Notation and Metric Prefixes . . . 3098

9.1.1.3 Scales of Measurements . . . 3101

9.1.2 Distributions . . . 3112

9.1.3 Constants . . . 3113

9.1.3.1 Mathematical Constants . . . 3114

9.1.3.2 Universal Constants (fundamental constants) . . . 3114

9.1.3.3 Astronomical/Astrophysical parameters and constants . . . 3119

9.1.3.4 Chemical parameters . . . 3119

9.1.3.5 Material parameters . . . 3121

9.1.3.6 Planck's constants . . . 3122

9.1.4 Principles of Physics . . . 3127

9.1.4.1 Principle of Causality . . . 3128

9.1.4.2 Principle of Conservation of Energy . . . 3129

9.1.4.3 Principle of Least Action . . . 3131

9.1.4.4 Noether's Principle (Noether's theorem) . . . 3132

9.1.4.5 Curie's Principle . . . 3144

9.1.5 Point Spaces . . . 3145

9.2 Analytical Mechanics . . . 3152

9.2.1 Lagrangian formalism . . . 3155

9.2.1.1 Generalized coordinates and frames . . . 3156
9.2.1.2 Variational Principle . . . 3160
9.2.1.3 Euler-Lagrange Equation . . . 3162
9.2.2 Canonical Formalism . . . 3180
9.2.2.1 Legendre Transform . . . 3180
9.2.2.2 Hamiltonian . . . 3181
9.2.2.3 Poisson bracket . . . 3189
9.2.2.4 Canonical transformations . . . 3192
9.3 Classical Mechanics . . . 3194
9.3.1 Newton's laws . . . 3199
9.3.1.1 Newton's first law (Inertia law) . . . 3199
9.3.1.2 Newton's second law (Fundamental Principle of Dynamics) . . . 3202
9.3.1.3 Newton's third law (Law of Action and Reaction) . . . 3209
9.3.2 Center of Mass and Reduced Mass . . . 3213
9.3.2.1 Center of Mass Theorem . . . 3216
9.3.2.2 Guldin's Theorem . . . 3221
9.3.3 Kinematics of Rectilinear Motion . . . 3224
9.3.3.1 Position . . . 3225
9.3.3.2 Velocity and speed . . . 3225
9.3.3.3 Acceleration . . . 3231
9.3.3.4 Galilean Relativity Principle . . . 3239
9.3.4 Angular Momentum . . . 3244
9.3.4.1 Moments . . . 3251
9.3.5 Static Forces . . . 3256
9.3.5.1 Lami's theorem . . . 3261
9.3.6 Ballistics . . . 3263
9.3.6.1 Ballistic equation without friction . . . 3265
9.3.6.2 Ballistic equation with friction . . . 3269
9.3.6.3 Ballistic accounting Earth's curvature . . . 3271
9.3.7 Kinematics of Circular Motion (rotational kinematics) . . . 3280
9.3.8 Energy, Work and Power . . . 3293
9.3.8.1 Conservative vector field . . . 3294
9.3.8.2 Kinetic Energy . . . 3297
9.3.8.3 Gravitational Potential Energy . . . 3351
9.3.8.4 Conservation of Total Mechanical Energy . . . 3360
9.3.8.5 Conservation of Linear Momentum . . . 3367
9.3.8.6 Power . . . 3379

9.3.9 Relative Movements and Inertial Forces 3383
9.3.9.1 Coriolis force and deflection magnitude 3391
9.3.9.2 Eötvös effect 3398
9.3.10 Oscillating Movements 3403
9.3.10.1 Newton's cradle 3403
9.3.10.2 Simple Pendulum (mass-wire pendulum) 3405
9.3.10.3 Physical Pendulum 3413
9.3.10.4 Elastic Pendulum (mass-spring pendulum) 3416
9.3.10.5 Conical Pendulum 3423
9.3.10.6 Torsion Pendulum 3425
9.3.10.7 Foucault's Pendulum 3426
9.3.10.8 Huygens' Pendulum (and brachistochrone curve) 3433
9.3.10.9 Double Pendulum 3440
9.3.10.10 Triple linear coupled pendulum 3446
9.3.10.11 Inverted Pendulum 3448
9.3.11 Tribology 3452
9.3.11.1 Exponential Friction 3462
9.3.11.2 Horizontal Viscous Friction 3465
9.3.11.3 Vertical Viscous Friction 3466
9.3.11.4 Stokes' Vertical Viscous Friction 3467
9.3.11.5 Stokes' Horizontal Viscous Friction 3470
9.3.11.6 Friction's Heat Factor 3473
9.4 Wave Mechanics 3475
9.4.1 Wave Function 3475
9.4.2 Wave Equation 3476
9.4.3 Type of Waves 3478
9.4.3.1 Periodic Waves 3478
9.4.3.2 Harmonic Waves 3479
9.4.3.3 Stationary Waves 3483
9.4.3.4 Vibration Modes in a Stretched String 3486
9.4.4 Non-relativistic Lagrangian of a String 3496
9.4.5 Vibrational modes of a circular membrane 3500
9.4.6 Phasors 3510
9.4.7 Superposition of periodic waves 3517
9.4.7.1 AM/FM Broadcasting 3523
9.5 Statistical Mechanics 3525
9.5.1 Statistical Information Theory 3525

9.5.1.1 High Entropy Structures . . . 3536
9.5.1.2 Usual Shannon entropies . . . 3538
9.5.1.3 Perplexity . . . 3541
9.5.1.4 Kullback–Leibler divergence . . . 3542
9.5.2 Boltzmann law . . . 3549
9.5.3 Statistical Physics Distributions . . . 3554
9.5.3.1 Maxwell Distribution (velocity distribution) . . . 3554
9.5.3.2 Maxwell-Boltzmann Distribution . . . 3561
9.5.3.3 Boltzmann Distribution . . . 3568
9.5.3.4 Fermi-Dirac Distribution . . . 3571
9.5.3.5 Bose-Einstein Distribution . . . 3576
9.5.4 Brownian Motion (random walk) . . . 3582
9.6 Thermodynamics . . . 3597
9.6.1 Thermodynamic Variables . . . 3599
9.6.2 Thermodynamics Systems . . . 3602
9.6.3 Thermodynamic Transformations . . . 3603
9.6.4 State Variables . . . 3607
9.6.4.1 Phases . . . 3609
9.6.5 Equation of State . . . 3611
9.6.5.1 Ideal Gas law . . . 3612
9.6.5.2 State equation of a Liquid . . . 3613
9.6.5.3 State equation of Solids . . . 3615
9.6.6 Laws of Thermodynamics . . . 3618
9.6.7 Calorific Capacities (heat capacity) . . . 3622
9.6.7.1 Newton’s Law of Cooling . . . 3635
9.6.8 Internal Energy . . . 3637
9.6.8.1 Work (energy) of Mechanical Forces . . . 3641
9.6.8.2 Enthalpy . . . 3643
9.6.8.3 Laplace’s law . . . 3646
9.6.8.4 Saint-Venant Thermodynamic Equation . . . 3651
9.6.8.5 Thermoelastics coefficients . . . 3653
9.6.9 Heat . . . 3656
9.6.9.1 Entropy . . . 3658
9.6.9.2 Carnot Cycle . . . 3668
9.6.10 Maxwell relations . . . 3672
9.6.11 Continuity Equation . . . 3679
9.6.11.1 Heat Equation . . . 3684

9.6.12 Thermal radiation . . . 3703
9.6.12.1 Black Body radiation . . . 3704
9.6.12.2 Summing up all types of heat transfer . . . 3731
9.7 Continuum Mechanics . . . 3734
9.7.1 Rigid Bodies . . . 3735
9.7.1.1 Pressures . . . 3735
9.7.1.2 Elasticity of Solids . . . 3738
9.7.2 Liquids (fluid mechanics) . . . 3770
9.7.2.1 Pascal's Theorem . . . 3773
9.7.2.2 Viscosity . . . 3775
9.7.2.3 Bernoulli's Theorem . . . 3783
9.7.2.4 Navier-Stokes Equations . . . 3811
9.7.2.5 Levelling law . . . 3869
9.7.2.6 Hydrostatic Pressure . . . 3872
9.7.2.7 Archimedes' principle . . . 3875
9.7.2.8 Surface tension . . . 3882
9.7.2.9 Speed of sound in a liquid . . . 3885
9.7.2.10 Summary of fluid dynamics laws . . . 3886
9.7.3 Gas . . . 3888
9.7.3.1 Types of Gas . . . 3888
9.7.3.2 Virial Theorem . . . 3897
9.7.3.3 Kinetic Pressures (kinetic theory of gases) . . . 3918
9.7.3.4 Kinetic Temperature . . . 3921
9.7.3.5 Amagat and Dalton's law . . . 3924
9.7.3.6 Mean free path (in kinetic theory) . . . 3925
9.7.3.7 Sackur-Tetrode equation . . . 3931
9.7.3.8 Generalized Bernoulli equation (for gases) . . . 3936
9.7.3.9 Magnus (gravitational) Effect . . . 3938
9.7.3.10 Summary of gas laws . . . 3943
9.7.4 Plasmas . . . 3945
9.7.4.1 Plasma Frequency . . . 3947
9.7.4.2 Lawson Criterion for D-T . . . 3951
9.7.4.3 Saha ionization equation . . . 3954
9.7.4.4 Ionospheric propagation . . . 3958

9.1. Principia

BY introducing the Mechanics through the Principia of Physics we finally do, after a veeeery long and necessary travel into the beautiful world of mathematics, our first steps in the field of theoretical physics (in a simplified version) who does not seek only the precision but also most of times the understanding of phenomena (exact numerical simulations and calculations and real life experiments give anyway very similar results of the theories as we will see). Understanding a theory in physics is primarily equivalent understanding the underlying mathematics behind. The rest is just some imagination and if possible, observation.

As we have already said, physics is a fundamental science that has a profound influence on all other sciences and also on human society and that aims to explain the *how*, not the *why* (see the Introduction of this textbook for details). Future physicists and engineers are thus not the only ones who should have understood its basic ideas, but anyone considering a career in science (including students majoring in biology, chemistry, mathematics and finance) should acquire the same comprehension.

quote
9.1:3068.#60

> *All science is either physics or stamp collecting!*
>
> Ernest Rutherford

The primary purpose of this textbook, we insist..., is to give the reader a unified view of physics by presenting what we think are the fundamental ideas that constitute the body of knowledge (absolute minimum) of contemporary physics.

Often, physics is taught as if it were a juxtaposition of several sciences, more or less connected, but without any real concern for unity and detail (which is why the majority of books and websites are without proofs). We rejected this mode of teaching (to have suffered of this during our studies) and opted for a unified and carefully detailed presentation by making when necessary reference to a section of the books that contains the proofs of mathematical tools used or another underlying physical theory. Obviously the law, relations and equations of physics can be compared to a computer program that possess, in addition to their immediate logical sense, most of time incomplete, a more global sense that can be revealed only sequentially according to the progress of mathematics and of our understanding of our Universe.

We also insist that every reader should know the basics of logic, arithmetic, algebra, vector calculus, linear algebra, tensor calculus, differential and integral calculus and analytical geometry prior to the study of physical phenomena to work with all rigour and all the necessary understanding to the mathematical reasoning that will be introduced from now (mathematics is the foundation of the huge edifice of theoretical physics!). We will see in what follows that all tools or mathematical results presented so far will be used !!

Let us recall that "Theoretical Physics" is the "exact/deductive science" that takes care of the best mathematical modelling natural phenomena, artificial, observable or non-observable. In shorter terms, we could talk about description of "reality" as to know if it is sensible or true reality...

When we want to predict or describe a specific physical phenomenon, we can usually go through an analytical model where different values given by unspecified (abstract values: variables) and the laws of physics by functions, when they are known (if so, we can make a hypothesis and test it). By equating a physical phenomenon, we translate reality (measurements) into a mathematical experience, virtual, according to certain rules. We conduct a simulation of reality on expressed quantities.

The different "laws" are often historically developed on empirical facts first and are subsequently verified experimentally (see the hypothetical-deductive method in the section of Proof Theory at page 141). Assuming that these laws are valid in the context (time frame, basis frame, etc.), we can therefore expect that the mathematics are in line with the experimental facts (or vice versa). Of course, a virtual experience is not real and can not express the reality in all its subtlety. This is only a model! It is therefore clear that the prediction of a physical phenomenon may differ from actual experimental facts.

Remarks

R1. We would like to recall (because of an abuse or misunderstanding that we find too often on various Internet forums), that a "scientific experiment" is a practical work of the study of a phenomenon which is reproducible (by independent research groups) in similar conditions and which number of reproductions is high enough to ensure that the errors (standard deviations) on measures become negligible between the independent groups (there are also international ISO norms for this purpose!!!).

R2. It should also be noted that most theoretical models that we will study now in this textbook making use of algebra or vector analysis and can be rewritten with the tools tensor analysis and based on the methods of Lagrangian formalism (see section Analytical/Lagrangian Mechanics page 3155 for know what it is...). However, these latter methods can not easily be used for a simple introduction to physics because they require additional efforts from reader and much more paper and time to get the same results (often at least...). However, and we will come back about this, these methods are now essential and of primary importance in the various fields of high-level modern physics such as Fluid Mechanics, General Relativity, Quantum Field Physics, analysis of chaotic systems and some others.

Before we begin our study of physical phenomena, we must define the main concepts on which theoretical physics is based. Thus, we will see in the order:

1. Human beings have created a system of measurement units which are arbitrary standard quantities to a given coefficient (this is why still some countries in the world don't have the same methods to count time or to weight objects even if there are ISO norms), specific to identify simply and in a reproducible way every physical phenomenon.
2. Some concepts interrelated to our vision of our environment lead us to make Assumptions/Hypothesis and Principles (to postulate something...) which relate to our sensory

reality while being transposed to any other reality of this type.

3. Theoretical physics leads us to consider the fundamentals of nature as abstract mathematical concepts. Thus our common observation gives us a concrete view of the Universe while theoretical physics gives us an abstract view of it. But theoretical physics let us sometimes inspects and understand (at least we suppose it) phenomenons that cannot be observed experimentally (or some dozens or hundreds years later only after technologies developments).

Then we can be led to ask however the following question: do the facts determine which theory/model is true?

By observing nature, we can see facts: this is data that we do not create. Astronomers, for example, note the position of celestial objects. We understand a fact since it appears as the consequence of the order of things described by a theory. But theories still retain the status of assumptions: even when a theory is consistent with all observed facts, it does not prove it to be true (a famous example is the Theory of Evolution considered by many people as being absolutely true even if it does not respect the scientific protocol). Indeed, there is still an infinite number of possible theories that are all compatible with all the observed facts. Then we say that the facts "sub-determine" theories: facts impose constraints on theories in the sense that only the theories consistent with the observed facts, are acceptable. But these constraints will always be low enough to leave the choice among an infinite number of theories.

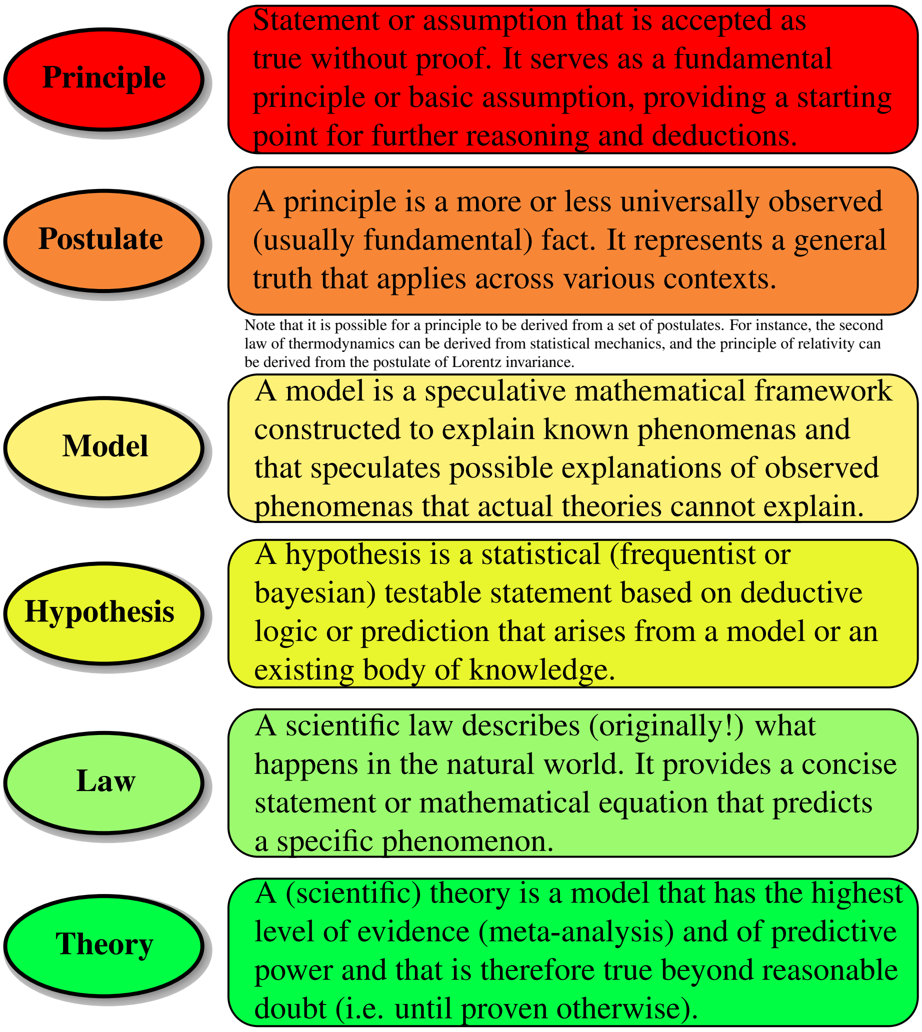
Theoretical advances may consist in setting aside old, incorrect paradigms (e.g., aether theory of light propagation, caloric theory of heat, burning consisting of evolving phlogiston, or astronomical bodies revolving around the Earth) or may be an alternative model that provides answers that are more accurate or that can be more widely applied.

Theoretical physics consists of several different approaches. In this regard, theoretical particle physics forms a good example. For instance: "phenomenologists" might employ (semi-) empirical formulas to agree with experimental results, often without deep physical understanding. "Modelers" (also named "model-builders") often appear much like phenomenologists, but try to model speculative theories that have certain desirable features (rather than on experimental data), or apply the techniques of mathematical modelling to physics problems. Some attempt to create approximate theories, named "effective theories", because fully developed theories may be regarded as unsolvable or too complicated. Other theorists may try to unify, formalise, reinterpret or generalise extant theories, or create completely new ones altogether. Sometimes the vision provided by pure mathematical systems can provide clues to how a physical system might be modelled. Theoretical problems that need computational investigation are often the concern of computational physics.

Of course, scientists are actually considering a finite number of theories, depending on what seems easiest in a given conceptual framework. At the time of Kepler, for example, there were three main theories to explain the movements of the planets, all consistent with the observed facts. According to the Ptolemaic theory, planets orbits are circular or located on a epicycle around the Earth motionless at the center of the universe. In the Copernican theory[1], the Sun (actually a G2V-type star, aka yellow dwarf) occupies the center, the orbits of the planets and the Earth being located on circles and epicycles. In the Keplerian theory, planets orbits are ellipses with the Sun at one focus and finally comes General Relativity that we will study later... (see page 5720).

---

[1] In reality the first one to make the assumption of heliocentric system seems to be the greek Aristarchus of Samos almost 1600 years before Copernicus (see [563])!



Principle

Statement or assumption that is accepted as true without proof. It serves as a fundamental principle or basic assumption, providing a starting point for further reasoning and deductions.

Postulate

A principle is a more or less universally observed (usually fundamental) fact. It represents a general truth that applies across various contexts.

Note that it is possible for a principle to be derived from a set of postulates. For instance, the second law of thermodynamics can be derived from statistical mechanics, and the principle of relativity can be derived from the postulate of Lorentz invariance.

Model

A model is a speculative mathematical framework constructed to explain known phenomenas and that speculates possible explanations of observed phenomenas that actual theories cannot explain.

Hypothesis

A hypothesis is a statistical (frequentist or bayesian) testable statement based on deductive logic or prediction that arises from a model or an existing body of knowledge.

Law

A scientific law describes (originally!) what happens in the natural world. It provides a concise statement or mathematical equation that predicts a specific phenomenon.

Theory

A (scientific) theory is a model that has the highest level of evidence (meta-analysis) and of predictive power and that is therefore true beyond reasonable doubt (i.e. until proven otherwise).

Figure 9.1: **Reminder of hierarchies of concepts in modern experimental science**

9.1.1 System of Units

The range of objects and phenomena studied in physics is immense. From the incredibly short lifetime of an atomic nucleus to the age of Earth, from the tiny sizes of sub-nuclear particles to the vast distance to the edges of the known universe, from the force exerted by a jumping flea to the force between Earth and the Sun, there are enough factors of 10 to challenge the imagination of even the most experienced scientist. Giving numerical values for physical quantities and equations for physical principles allows us to understand nature much more deeply than qualitative descriptions alone. To comprehend these vast ranges, we must also have accepted units in which to express them. We shall find that even in the potentially mundane discussion of meters, kilograms, and seconds, a profound simplicity of nature appears: all physical quantities can be expressed as combinations of only a few base physical quantities.

Definition (#9.1.402): A "measurement" or "physical quantity" is the quantitative nomological expression of a property, of an effect or an abstract quantity defined by a model that represent the object or phenomenon under study. A fundamental measurement does not explain itself, it described itself with respect to a definition.

Definition (#9.1.403): A "unit of measurement" is a definite magnitude of a physical quantity, defined and adopted by convention or by law, that is used as a standard for measurement of the same physical quantity. Any other value of the physical quantity can be expressed as a simple multiple of the unit of measurement.

In fact it's not enough to simply agree on units of measurement! These have to be consistent, reliable and standardized in order to enable international collaboration and interoperability. There is a science to measurement, and it's named "metrology".

✍**Example:**

For example, length is a physical quantity. The "meter" is a unit of length that represents a definite predetermined length. When we say 10 meters (or 10 [m]), we actually mean 10 times the definite predetermined length named "meter".

We recognize two types of measurements:

1. The "constants": they have a concrete value expressible numerically and are supposed not to change during the studied phenomenon. They are "passive measurements" (we come back later on this and will enumerate some of them).
2. The "variables": they have a concrete value only in a particular state, but not when we observe the physical phenomenon as a whole (during all a range of time). These are "active variables". The variables describing a physical phenomenon are often correlated to each other by means of functions. Then by definition we say that these variables have a "functional relation" between them.

Remark

One measurement only makes sense if it is "observable", measurement to which we associate a number, the result of the measurement performed by using a device.

Measure a physical quantity is equivalent to compare it to a known quantity of the same kind (we also say "of the same dimension"), taken as an arbitrary standard (the rigorous approach of this related the abstract modern Measurement Theory that we have already study in the section of the same name). The measurement result is expressed thanks to two elements:

1. A number which is the ratio of the measured quantity to the standard of the measurement quantity.
2. A name or acronym that identifies clearly the chosen standard.

The notation of a measurement A is therefore in abstract form:

$$A=\{A\}\,[A]=n\{S_A\}\,[\mathrm{A}] \tag{9.1.1}$$

The "number" $\{A\}$ is in fact the measured value (magnitude) of the quantity A as multiple n of the corresponding standardized unit $\{S_A\}$ and $[A]$ is its "name" that we commonly name "physical unit" or simply "unit" (the quantitative expression of a variable or a function). These two elements are inseparable, the measured value is meaningful only if we show the same time the selected unit. It changes if we change the unit!!!

Remark

The transition from one unit to another to express the same quantity is named "unit conversion". Conversion of units involves obviously the comparison of different standard physical values, either of a single physical quantity or of a physical quantity and a combination of other physical quantities.

Starting with:

$$A=n\{S_A\}\,[\mathrm{A}]$$

just replace the original standard reference unit $\{S_A\}$ with its meaning in terms of the desired unit $\{S_B\}$, if:

$$\{S_A\}=c_{A\to B}\,\{S_B\}$$

then:

$$A=n\{S_A\}\,[\mathrm{A}]=nc_{A\to B}\{S_A\}\,[\mathrm{B}]=n\{S_B\}\,[\mathrm{B}]$$

Now n and $c_{A\to B}$ are just the numerical values corresponding respectively to the multiplicand of the standard and the conversion factor, we just have to multiply them to finish the job.

✍**Examples:**

E1. One example of the importance of agreed units is the failure of the NASA Mars Climate Orbiter, which was accidentally destroyed on a mission to Mars in September 11999 (holocene calendar) instead of entering orbit due to miscommunications about the value of forces: different computer programs used different units of measurement (newton versus pound force). Considerable amounts of effort, time, and money were wasted.

E2. On 15 April 11999 (holocene calendar), Korean Air cargo flight 6316 from Shanghai to Seoul was lost due to the crew confusing tower instructions (in meters) and altimeter readings (in feet). Three crew and five people on the ground were killed. Thirty-seven were injured.

E3. In 11983 (holocene calendar), a Boeing 767 ran out of fuel in mid-flight because of two mistakes in figuring the fuel supply of Air Canada's first aircraft to use metric measurements. This accident was the result of both confusion due to the simultaneous use of metric and Imperial measures and confusion of mass and volume measures.

Definition (#9.1.404): Some variables can, for the sake of simplification of writing, be expressed from other quantities. We then say that the new quantity "derives" from base units. We also say that two physical quantities are "homogeneous quantities", or "dimensionally consistent", if they are of the same physical nature or if we can express them both in the same basis unit(s).

This also means that the equations in physics must obey the following rules:

- Every term in an expression that is added or subtracted must have the same dimensions; it does not make sense to add or subtract quantities of differing dimension. In particular, the expressions on each side of the equality in an equation must have the same dimensions.
- The arguments of any of the standard mathematical functions such as trigonometric functions (such as sine and cosine), logarithms, or exponential functions that appear in the equation must be dimensionless. These functions require pure numbers as inputs and give pure numbers as outputs.

So, after a long period of reflection the physical world seems to be reducible to the concepts of space, time and energy.

Thus appears another possible definition of Physics:

Definition (#9.1.405): "Physics" is the science of the properties and mutual relations in time of matter and energy to a given factor.

Our role is to give a description of these properties and relations under the form of physical equations or laws applied to observed phenomena in the context of a theory providing predictable facts.

Physical quantities are not all independent of each other but linked by certain laws or relations. It would be unreasonable, but possible, to choose a particular unit for each of the physical quantities regardless of their mutual relations.

Build a coherent system of units consists therefore to determine a minimum number of units that establish the rules of construction of these mutual relations. Indeed, historically, a lot of units has been set for the measurement of a single quantity to meet specific needs of areas of practical life or science and technology. Their choices and their definitions were often empirical (or even based on religious beliefs sometimes!), and hence the conversions between units not always easy or clear.

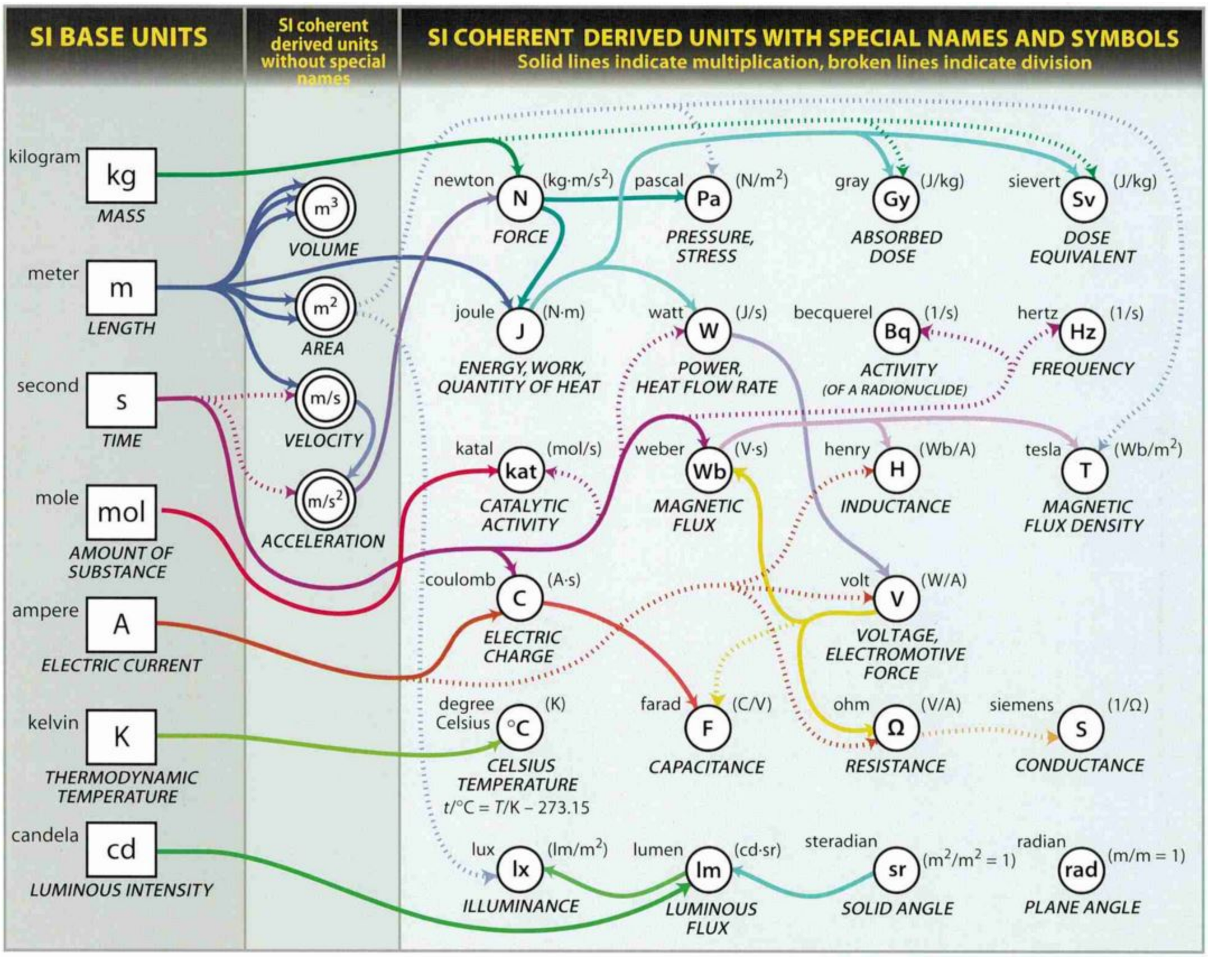

SI BASE UNITS
SI coherent derived units without special names
SI COHERENT DERIVED UNITS WITH SPECIAL NAMES AND SYMBOLS
Solid lines indicate multiplication, broken lines indicate division
kilogram kg MASS
meter m LENGTH
second s TIME
mole mol AMOUNT OF SUBSTANCE
ampere A ELECTRIC CURRENT
kelvin K THERMODYNAMIC TEMPERATURE
candela cd LUMINOUS INTENSITY
m3 VOLUME
m2 AREA
m/s VELOCITY
m/s2 ACCELERATION
newton N (kg·m/s2) FORCE
pascal Pa (N/m2) PRESSURE, STRESS
gray Gy (J/kg) ABSORBED DOSE
sievert Sv (J/kg) DOSE EQUIVALENT
joule J (N·m) ENERGY, WORK, QUANTITY OF HEAT
watt W (J/s) POWER, HEAT FLOW RATE
becquerel Bq (1/s) ACTIVITY (OF A RADIONUCLIDE)
hertz Hz (1/s) FREQUENCY
katal kat (mol/s) CATALYTIC ACTIVITY
weber Wb (V·s) MAGNETIC FLUX
henry H (Wb/A) INDUCTANCE
tesla T (Wb/m2) MAGNETIC FLUX DENSITY
coulomb C (A·s) ELECTRIC CHARGE
volt V (W/A) VOLTAGE, ELECTROMOTIVE FORCE
degree Celsius °C (K) CELSIUS TEMPERATURE t/°C = T/K – 273.15
farad F (C/V) CAPACITANCE
ohm Ω (V/A) RESISTANCE
siemens S (1/Ω) CONDUCTANCE
lux lx (lm/m2) ILLUMINANCE
lumen lm (cd·sr) LUMINOUS FLUX
steradian sr (m2/m2 = 1) SOLID ANGLE
radian rad (m/m = 1) PLANE ANGLE

Figure 9.2: **Based ISQ units and derived units**

In regards to the evolution of sciences and technologies, at the increased exchange of information, the need to unify the system appeared obvious, the aim being to keep a small as possible number of units. These are the "fundamental units". From the physical laws and relations between the different fundamental units, we deduce the units of other values which then become then by sake of simplification of writing "derived units".

The choice and definition of the fundamental units set is a complex issue. Indeed, these units must be highly accurate and available in measurement laboratories of the entire planet.

Remark

The International System of Quantities (ISQ) is a system based on seven base quantities: length, mass, time, electric current, thermodynamic temperature (see definition at page 3935), amount of substance, and luminous intensity. Other quantities such as area, pressure, and electrical resistance are derived from these base quantities by clear non-contradictory equations. The ISQ defines the quantities that are measured with the SI units. The ISQ is defined in the international standard ISO/IEC 80000, and was finalized in 12009 (holocene calendar) with the publication of ISO 80000-1 *Quantities and units*. A small summary is given the figure above.

The basic units are at the number of 4 (we will justify this later): the length (meters), mass (kg), time (seconds), the electric charge (coulombs). This system then constitutes the "M.K.S.C. system" (the main redactor of this textbook assumes the choice of adding the Coulomb as fundamental unit).

The units of the M.K.S.C system are within this textbook (according to the early 121st century standards in most countries around planet Earth):

1. The meter [m], for the length L (we have already defined the concept of length in the chapter Geometry but we will come back again on it later). Since 12018 (holocene calendar) it is defined by taking the fixed numerical value of the speed of light in vacuum (as we will see during our study of Quantum Field Theory, vacuum isn't really empty) c to be $299,792,458$ when expressed in the unit $\mathrm{m}\cdot\mathrm{s}^{-1}$, where the second [s] is defined in terms of the caesium frequency $\Delta\nu_{\mathrm{Cs}}$. Hence:

$$1\;[\mathrm{m}]=\frac{c}{299,792,458}\cdot 1\;[\mathrm{s}]=\frac{9,192,631,770}{299,792,458}\frac{c}{\Delta\nu_{\mathrm{Cs}}}\cong 30.663319\frac{c}{\Delta\nu_{\mathrm{Cs}}}\tag{9.1.2}$$

   The effect of this definition is that one metre is the length of the path travelled by light in vacuum during a time interval with duration of $1/299,792,458$ of a second.

2. The kilogram [kg] for the mass M (we will return back later to the definition of the concept of mass). Since 12018 (holocene calendar) it is defined by taking the fixed numerical value of the Planck constant h to be $6.62607015\cdot 10^{-34}$ when expressed in the unit $[\mathrm{J}\cdot\mathrm{s}^{-1}]$, which is equal to $[\mathrm{kg}\cdot\mathrm{m}^2\cdot\mathrm{s}^{-1}]$, where the metre and the second are defined in terms of c and $\Delta\nu_{\mathrm{Cs}}$. Hence:

$$1\;[\mathrm{kg}]=\frac{h}{6.62607015\cdot 10^{-34}}\cdot 1\;[\mathrm{m}^{-2}\cdot\mathrm{s}]=\frac{299,792,458^2}{9,192,631,770\cdot 6.62607015\times 10^{-34}}\frac{h\Delta\nu_{\mathrm{Cs}}}{c^2}\cong 1.4755214\cdot 10^{40}\frac{h\Delta\nu_{\mathrm{Cs}}}{c^2}\tag{9.1.3}$$

   The effect of this definition is to introduce the unit $\mathrm{kg}\cdot\mathrm{m}^2\cdot\mathrm{s}^{-1}$ (the unit of both the physical quantities of action and angular momentum). Together with the definitions of the second and the metre this leads to a definition of the unit of mass expressed in terms of the Planck constant h.

3. The second [s], for time T (time is not measurable in himself as it cannot be touched, but the time interval Δt is an arbitrary concept quite valid - we also will also return back on the definition of this concept further below). Since of 12018 (holocene calendar) it is defined by taking the fixed numerical value of the caesium frequency $\Delta\nu_{\mathrm{Cs}}$, the

unperturbed ground-state hyperfine transition frequency of the Caesium-133 atom, to be $9,192,631,770$ when expressed in the unit $[\text{Hz}]$, which is equal to $[\text{s}^{-1}]$. Hence:

$$1\,[\text{s}] = \frac{9,192,631,770}{\Delta\nu_{\text{Cs}}} \tag{9.1.4}$$

The effect of this definition is that the second is equal to the duration of $9,192,631,770$ periods of the radiation corresponding to the transition between the two hyperfine levels of the unperturbed ground state of the ^{133}Cs atom.

4. The Coulomb [C] used as a fundamental unit of electric charge $q_e = 1.602176634 \cdot 10^{-19}$ (not derived from any known at this date - we will also return to the definition of this concept further below).

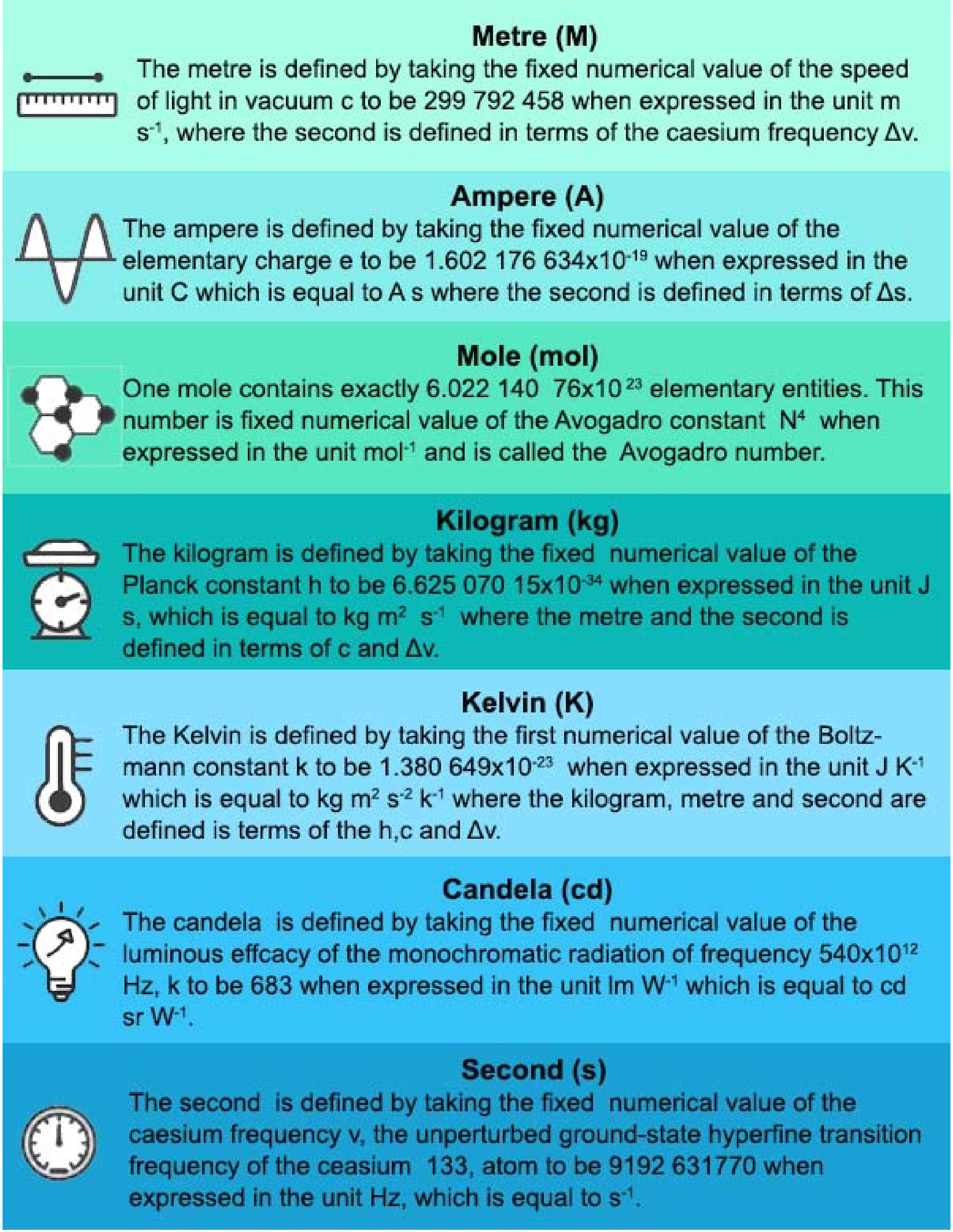
Remark

The electric charge isn't listed among the 19 free parameters of the Standard Model of particles because it's not put in by hand-it, it emerges naturally from deeper symmetries in the theory and therefore isn't something fundamental in reality... The electric charge arises from a combination of two of the gauge coupling constants g_1 and g_2. From there we can calculate a fine-structure constant, as per:

$$\alpha = \frac{1}{4\pi}\frac{(g_1 g_2)^2}{g_1^2 + g_2^2} \tag{9.1.5}$$

which then is proportional to the elementary charge squared, i.e. $\alpha \propto e^2$.

The 12019 (holocene calendar) redefinition of the SI base units came into force on 20 May 12019. It can be summarized by the following figure (see [206] for more details):



Metre (M)

The metre is defined by taking the fixed numerical value of the speed of light in vacuum c to be 299 792 458 when expressed in the unit m s^{-1}, where the second is defined in terms of the caesium frequency Δv.

Ampere (A)

The ampere is defined by taking the fixed numerical value of the elementary charge e to be 1.602 176 634x10^{-19} when expressed in the unit C which is equal to A s where the second is defined in terms of Δs.

Mole (mol)

One mole contains exactly 6.022 140 76x10^{23} elementary entities. This number is fixed numerical value of the Avogadro constant N^4 when expressed in the unit mol^{-1} and is called the Avogadro number.

Kilogram (kg)

The kilogram is defined by taking the fixed numerical value of the Planck constant h to be 6.625 070 15x10^{-34} when expressed in the unit J s, which is equal to kg m^2 s^{-1} where the metre and the second is defined in terms of c and Δv.

Kelvin (K)

The Kelvin is defined by taking the first numerical value of the Boltzmann constant k to be 1.380 649x10^{-23} when expressed in the unit J K^{-1} which is equal to kg m^2 s^{-2} k^{-1} where the kilogram, metre and second are defined is terms of the h,c and Δv.

Candela (cd)

The candela is defined by taking the fixed numerical value of the luminous effcacy of the monochromatic radiation of frequency 540x10^{12} Hz, k to be 683 when expressed in the unit lm W^{-1} which is equal to cd sr W^{-1}.

Second (s)

The second is defined by taking the fixed numerical value of the caesium frequency v, the unperturbed ground-state hyperfine transition frequency of the ceasium 133, atom to be 9192 631770 when expressed in the unit Hz, which is equal to s^{-1}.

Figure 9.3: Based 12019 SI

The reader has to known that since its adoption into the International System of Units in 11971 (holocene calendar), criticisms of the concept of the mole as a unit like the metre or the second have arisen as the number of molecules, etc. in a given amount of material is a fixed dimensionless quantity that can be expressed simply as a number, not requiring a distinct base unit. Therefore the mole is not a true metric (i.e. measuring) unit, rather it is a parametric unit, and amount of substance is a parametric base quantity. We can only hope that in the future this unit will be removed.

Remarks

R1. The concept of angle θ (in radians, degrees or steradians - see the texts on plane trigonometry, spherical trigonometry and plane geometry in the chapter Geometry starting at page 2503) since it has no unit because by definition its a ratio of lengths (for the radian or degree of angle) or surfaces (for the steradian). It is therefore appropriate to assimilate it to a derived unit and not as fundamental unit. However, in physics, we have sometimes the habit to indicate its presence in the dimensional equations to help re-reading some of them and to know that their result is given with respect to an angle unit (otherwise it could risky generate misinterpretations for those using the equations without have read the proof...).

R2. The development of science has led the General Conference of Weights and Measures to introduce some additional practices (but not necessary) units such as (among others as you can see it in the previous figure) the temperature in "Kelvin" [K] (derived from the average energy - Brownian motion), the amount of expressed in terms Moles, the current intensity in Ampere [A] and the light intensity expressed in Candelas [cd]. Thus, the actual International System (SI) with its seven fundamental base units (kilogram-meter-second, kelvin, candela, mole and ampere) and its seventeen-derived units does it suggests that seven units are needed to describe all physics? Actually no! As the Gauss analysis suggests it, among the seven fundamental units, four of them - Kelvin, Candela, the Mole and the Ampere - can be derived from the other three. The introduction of seven base units represents a pragmatic balance between experimentalists who require units adapted to their measures, and idealism theorists, whose goal is to reduce arbitrary redundancy to a minimum.

R3. This is a huge chance to have a homogeneous system such as we have in the 121st century (holocene calendar). Indeed, for the record, in 11522 (holocene calendar) alone in the Baden-Baden area (Germany) there were 112 units of measures for the concept of length and 92 units of measures for the concept of area.... that is to say the mess... a nightmare! In the early 121st century (holocene calendar) there are ISO 80000 norms which aim to harmonize the notations, definitions and values of units in all fields of sciences.

These clarifications done, any physical quantity known at this date can be expressed using a unit which is expressed as the product at a given power (positive or negative power!) of the five fundamental dimensional factors and an arbitrary scale factor K (dimensionless number, which acts as a proportionality factor for taking from a scale, that of the fundamental unit, to that of the studied variable):

$$K\cdot[M]^m\cdot[L]^l\cdot[T]^t\cdot[\theta]^\theta\cdot[C]^c \tag{9.1.6}$$

where the numbers m,l,t,θ,c named respectively "mass order", "length order", "time order", "angle order" and "charge order" are positive, negative or zero integers (for the huge majority of cases!).

The above expression will be written in the "canonical form" defined by the standard units:

$$K\cdot[\text{kg}]^m\cdot[\text{m}]^l\cdot[\text{s}]^t\cdot[\text{C}]^c \tag{9.1.7}$$

the angle having no units, we don't write it anymore (but can be still here implicitly depending on the context!).

Any physical quantity X is therefore expressed as:

$$x = x \cdot K \cdot [\text{kg}]^m \cdot [\text{m}]^l \cdot [\text{s}]^t \cdot [\text{C}]^c \tag{9.1.8}$$

where x is the magnitude of the physical quantity in the unit system associated with the scaling factor K. There are several possible pairs (x, K), but we will always have:

$$x \cdot K = x_0 = c^{te} \tag{9.1.9}$$

where the constant x_0 is the value of the physical magnitude when we choose to express it in the system M.K.S.C.

So two physical quantities are homogeneous if and only if the quadruplets:

$$(m_1, l_1, t_1, c_1), (m_2, l_2, t_2, c_2) \tag{9.1.10}$$

that are linked to them are equals:

$$m_1 = m_2 \quad l_1 = l_2 \quad t_1 = t_2 \quad c_1 = c_2 \tag{9.1.11}$$

It follows from what we have said, that:

- The sum or difference of any number of quantities is meaningful only if these quantities are homogeneous and the result will have the same units as the operands.
- The product or division of multiple quantities has for units respectively the product or division of the operands units.

Remark

The units of the different quantities have a practical side but not infallible in theoretical physics: they nevertheless allow the physicist to verify whether a proven relation between two quantities is at least correct at the unit level. We name this kind of approach a "dimensional analysis" (we advise you to go see the proof of the Stokes law in the section of Continuum Mechanics at page 3853 for a very good example of application).

For a little summary:

Basic units

Quantity	Unit	Sym.
Length	meter	m
Mass	kilogram	kg
Time	second	s
Therm. temp.	kelvin	K
Electr. current	ampere	A
Luminous intens.	candela	cd
Amount of subst.	mol	mol

Extra units

Plane angle	radian	rad
Solid angle	sterradian	sr

Derived units with special names

Quantity	Unit	Sym.	Derivation
Frequency	hertz	Hz	s^{-1}
Force	newton	N	$kg \cdot m \cdot s^{-2}$
Pressure	pascal	Pa	$N \cdot m^{-2}$
Energy	joule	J	$N \cdot m$
Power	watt	W	$J \cdot s^{-1}$
Charge	coulomb	C	$A \cdot s$
El. Potential	volt	V	$W \cdot A^{-1}$
El. Capacitance	farad	F	$C \cdot V^{-1}$
El. Resistance	ohm	Ω	$V \cdot A^{-1}$
El. Conductance	siemens	S	$A \cdot V^{-1}$
Mag. flux	weber	Wb	$V \cdot s$
Mag. flux density	tesla	T	$Wb \cdot m^{-2}$
Inductance	henry	H	$Wb \cdot A^{-1}$
Luminous flux	lumen	lm	$cd \cdot sr$
Illuminance	lux	lx	$lm \cdot m^{-2}$
Activity	bequerel	Bq	s^{-1}
Absorbed dose	gray	Gy	$J \cdot kg^{-1}$
Dose equivalent	sievert	Sv	$J \cdot kg^{-1}$

9.1.1.1 Dimensional Analysis

Dimensional analysis is an area of physics with that concerns the to units of the magnitudes. In particular, the fact that the units are relatively arbitrary makes that any valid equation of physics must be homogeneous: something that is measured in meters per second can not be equal to something that is measured in kilograms per meter. It is a very popular way and very effective way to check its own calculations (and that of others ...).

This statement, which can seem paradoxical, effectively means: Do not start (if possible ...) in a complicated calculation without finding beforehand the qualitative form of the result with dimensional analysis.

This qualitative form is traditionally named the "dimensional equation" and thus represents the formula that gives the possibility to determines the unit in which the result must be expressed in a research. This is an equation of variables (quantities/magnitudes), that is to say, in which the measured phenomena are represented by a unit symbol those we have seen in the previous paragraphs.

✍**Example:**

Let's see an example of a legend (?) often quoted in magazines and popular books:

Dimensional analysis should have enabled G. I. Taylor in 11950 (holocene calendar) to estimate the energy released by the explosion of an atomic bomb (first detonation of a nuclear weapon in July 16, 11945 (holocene calendar): code name "Trinity", tested in the Jornada del Muerto desert), while this information was classified top secret, it would have been enough for him to watch a movie of the explosion, unwisely made public by the US military.

The physicist G. I. Taylor suppose to arrive at this result that the process of expansion of the gas sphere depends at the minimum of the parameters of time t, the generated energy E by the explosion and the density of air ρ.

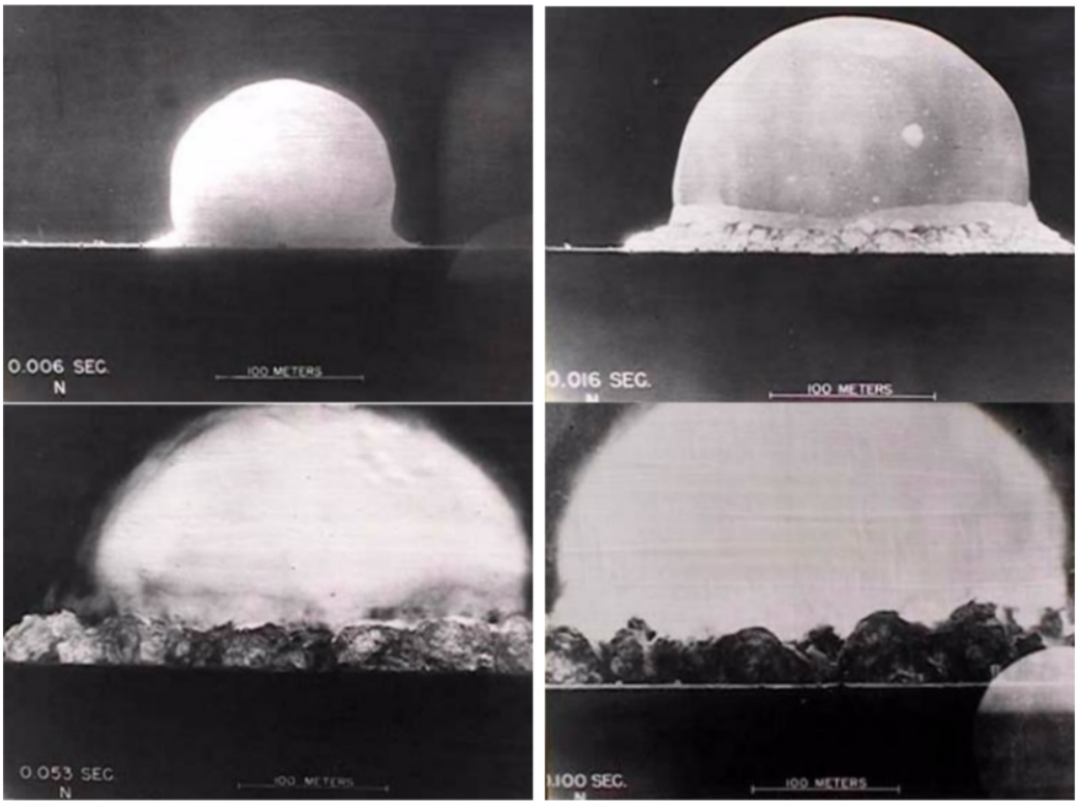


Figure 9.4: **Supposed picture used by G. I. Taylor for the estimation (source: ?)**

Dimensional analysis then conducted him for the radius of the sphere of gas at time t to:

$$r\,[\mathrm{m}] = k \cdot \left(E\,[\mathrm{kg}\cdot\mathrm{m}^2\cdot\mathrm{s}^{-2}]\right)^x \cdot \left(\rho\,[\mathrm{kg}\cdot\mathrm{m}^3]\right)^y \cdot \left(t\,[\mathrm{s}]\right)^z \tag{9.1.12}$$

where k is a dimensionless constant.

And by trial and error we find relatively quickly $x = 1/5, y = -1/5, z = 2/5$ such as:

$$r = kE^{1/5}\rho^{-1/5}t^{2/5} \tag{9.1.13}$$

Indeed:

$$m = \left(\mathrm{kg}^{1/5}\cdot\mathrm{m}^{2/5}\cdot\mathrm{s}^{-2/5}\right)\cdot\left(\mathrm{kg}^{-1/5}\cdot\mathrm{m}^{3/5}\right)\cdot\left(\mathrm{s}^{2/5}\right) \tag{9.1.14}$$

Taylor then finds the time dilation law of the radius of the atomic mushroom is proportional to (there is no need to indicate other units, since only the temporal part interests us here!):

$$r \propto t^{2/5} \tag{9.1.15}$$

If we know r and t from a film, and k being supposed to be of the order of the unit and ρ being known, we finally get:

$$E \cong \frac{\rho r^5}{t^2} \tag{9.1.16}$$

what remains a rough approximation. But getting such a result (of magnitude) with the heavy artillery of theoretical physics would require much more time and paper.

As we can see in the example above, the power of units aren't always an integer! The most well-known case of non-integer dimension units are first the one in financial engineering (Value at Risk) involving de square root of time (i.e. $[\text{s}^{1/2}]$) and the second most well known one is in fiber optics technology (polarization mode dispersion coefficient) involving an inverse of square root distance (i.e. $[\text{s}^{-1/2}]$).

9.1.1.1.1 Time

Definition (#9.1.406): The "time" is a state variable (not a "measurable") and therefore a notion impalpable but however rigorously defined. It also is a mathematical tool to equate the observation of physical phenomena (observable) and thus draw some information. This tool exists because there are living beings to observe (and measure) the Nature and its changes (Socratic principle) and matter and movement for there to be changes. Time is therefore the indefinite continued progression of existence and events that occur in apparently irreversible succession from the past through the present to the future. The SI derived unit of time is the second [s]. Time is a component quantity of various measurements used to sequence events, to compare the duration of events or the intervals between them, and to quantify rates of change of quantities in material reality or in the conscious experience. Time is often referred to in General Relativity as one of the four dimensions, along with the three spatial dimensions (see section Special Relativity page 5627), when multiplied by the speed of light, of a four-dimensional pseudo-Riemmanian manifold.

Caution! As we will prove in the section of Quantum Gravity (see page 6087), the Wheeler-DeWitt model suggests that the universe as a whole may be timeless, and what we perceive as time is a local, emergent phenomenon that is not fundamental because as we know from General Relativity (see page 5720) there's no universal clock!

Remarks

R1. Time (and time intervals) being an arbitrary concept, it is symmetrical, that is to say that every phenomenon observed can be in the imagination of a possible inverted time finding his initial conditions. We speak then of "symmetry of time" (for now it has never been proven as far as we know that time may suffer of a "breaking symmetry"). In fact as we have seen in the section of Statistical Mechanics (see page 3658), the direction of time is defined by the concept of entropy and as the entropy can only increase (as far as we know), the question to go back in time is a non-sense (however in thermodynamic systems that are not closed, remember that entropy can decrease with time). We speak then of "thermodynamic time".

R2. Certain subatomic interactions involving the weak nuclear force violates the conservation of parity, but only very rarely. According to the CPT theorem, this means they should also be time irreversible, and so establish an arrow of time. This, however, is neither linked to the thermodynamic arrow of time, nor has anything to do with our daily experience of time irreversibility.

R3. If an intensive variable is one that is unchanged with the thermodynamic operation of scaling of a system (see section Thermodynamics page 3608) the time is not an intensive variable if we think to General Relativity (mass influence on time).

We represent very often in physics time (within a interval) by a horizontal arrow (axis). Since time is a relative concept, we can then define each instant of time as being zero time denoted t_0. This concept is very often used in physics because the only thing that interests physicists is the time difference denoted Δt (by the use of differential and integral calculus) or $\mathrm{d}t$ when the interval is very small.

Let us prove now that the time reference is independent of the choice for an observer at rest. Given a time denoted by the letter t, then we have:

$$\Delta t=(t_1+t')-(t_0+t')=t_1-t_0 \tag{9.1.17}$$

where t' is an arbitrary basis (not necessary) when we compare a temporal difference.

The time interval is given by a standard measure that can only be at best a perfectly periodic movement (repeated over time). So, the first means of measuring time were day and night..., the positions of the Sun (actually a G2V-type star, aka yellow dwarf) and Moon in the sky..., the pendulum, the relaxation of springs, the Cesium 137 degeneration period or even binary systems binaries of massive stars. In short, an observable system that produces a stable periodic phenomenon and small enough for any physical action that can be reduced to it can be used as a standard time interval!!!

Definitions (#9.1.407):

D1. An "event" consists to give meaning to a point in space-time (for a naive and technical definition of "space-time" see the section of Special Relativity at page 5708).

D2. Two events are say to be "simultaneous events" if they have the same value of the time coordinate.

D3. We name "coincidence" the simultaneity of two events at the same point in space. The coincidence is an absolute fact, independent of the choice of the repository. This is actually a particular case of the causal conservation principle. Two events coinciding in a referential frame may be the cause and effect of each other (and vice versa), and that possibility is retained in the new marker.

After Albert Einstein's Special Relativity (see section on the subject page 5627), time can no longer be thought of being absolute in all reference frames. Thus a watch that indicate one minute in one frame of reference will not be one minute in a reference frame that is travelling at a velocity relative to the first frame. This means that the time of an event is variable depending on the observer!

In high level physics, time (human creation) is always multiplied by the constant of speed of light c to get a distance and deal with it like a standard fourth dimension. As also in high level physics practitioners use natural units where among other $c=1$, then time is strictly and simply equal to length such that $t=\ell$.

9.1.1.1.2 Length

Definition (#9.1.408): The concept of "length", denoted most of the times x or sometimes ℓ, is given by the information that gives the path taken by an object in a given time interval or simply be the distance between two points. The SI derived unit of length is the meter [m].

Length may be distinguished from "height", which is vertical extent, and "width" or "breadth", which are the distance from side to side, measuring across the object at right angles to the length. Length is a measure of one dimension, whereas "area" is a measure of two dimensions (length squared) and "volume" is a measure of three dimensions (length cubed) as we will study this in details in the section of Euclidean Geometry (see page 2553). In most systems of measurement, the unit of length (meter) is a base unit, from which other units dimensions are defined.

Remarks

R1. If there was no matter in the Universe there would be no concept of movement and thus no path length and as we have already pointed it out, also no time (and again ... it is without considering some results of quantum physics that we will prove in the sections dedicated to it).

R2. Length is obviously an extensive quantity (additive). Indeed, the total length of an item is the sum of the length that composed it.

As for time, there is no absolute origin of length measurement (there is no "zero point" in the Universe as postulated, among others, by the theory of relativity) and physicists are interested anyway especially only to differences of travelled path Δx relative to an origin as they do for the time.

Thus, identically as for the time, we have for an observer at rest observing a moving material point:

$$\Delta x=(x_1+x')-(x_0+x')=x_1-x_0 \tag{9.1.18}$$

where x' is a mathematically arbitrary basis that is useless when comparing a difference of position in a time difference of a material point (as it cancels as you can see in the previous relation).

If a material point is located in a three dimensional space (the most frequent case in Classical

Mechanics, Astronomy, Astrophysics and Particle Physics) we arbitrarily chose an origin O, we note the position by of this body by $\vec{r}$, its distance in length by x, in width by y and in height by y (named "Cartesian coordinates" as we already know) by an imaginary arrow named "vector" (see section Vector Calculus page 1797) connecting the arbitrary origin of the reference frame to the interested point as follows:

$$\vec{r}=\begin{pmatrix}x\\y\\z\end{pmatrix}=\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix} \tag{9.1.19}$$

Remark

The arrow above the $\vec{r}$ means of course that it is a vector!

The notation:

$$(x,y,z):=(x_1,x_2,x_3) \tag{9.1.20}$$

is a simplified notation used frequently in physics and who should prevail in small classes (take care! the fact that the numbers are in indices here, does not mean that they are covariant components - see section of Linear Algebra page 1956 - this is just a notation convention). However in this textbook we move from one to the other notations based on the needs and usage (it will be up to you to be careful to not confuse).

The matrix:

$$\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix} \tag{9.1.21}$$

is the metric tensor of a canonical pre-Euclidean space with positive signature (see section Vector Calculus page 1810 and General Relativity page 5731). This is a particular case in theoretical physics but still a very common case study in Classical Mechanics (you must start with simple spaces before going further...).

We will return in more detail on these concepts during our study punctual spaces further below.

After Albert Einstein's Special Relativity (see section on the subject page 5627), length can no longer be thought of being constant in all reference frames. Thus a ruler that is one meter long in one frame of reference will not be one meter long in a reference frame that is travelling at a velocity relative to the first frame. This means length of an object is variable depending on the observer!

Remark

All measurement apparatus that Humans use to measure other physical quantities are if you think about it related in one way or another to lengths (distance) comparisons. The fact that the measurement of a physical property can be reduced to a measurement of distance does not imply the non-existence of the physical phenomenon which produces the measured displacement. The measurement of the distance constitutes only our mode of access to the physical phenomenon in question. And this phenomenon can be explained independently of the measurement one makes of it!

Definition (#9.1.409): The "litre", SI symbols L, is a metric unit of volume. It is equal to 1 cubic decimetre (dm^3), $1,000$ cubic centimetres (cm^3) or 0.001 cubic metres (m^3). A cubic decimetre (or litre) occupies a volume of $10\,[\text{cm}]\cdot 10\,[\text{cm}]\cdot 10\,[\text{cm}]$. One litre of liquid water has a mass of almost exactly one kilogram, because the kilogram was originally defined in 11795 (holocene calendar) as the mass of one cubic decimetre of water at the temperature of melting ice ($0\,[^\circ\text{C}]$) at standard pressure conditions. Subsequent redefinitions of the metre and kilogram mean that this relationship is no longer completely exact.

9.1.1.1.3 Mass

In general the "mass" m of a body is in a closed system an amount which is preserved and that characterizes the amplitude with which a body interacts with other body through different (attractive) forces. The SI derived unit of the mass is the kilogram [kg].

> **Remark**
>
> In an isolated system and where there is no expansion of space-time, there can not be spontaneous creation or destruction of mass (however this is not the case for energy as we will see it during our study of the Casimir Effect in the section of Wave Quantum Physics at page 4825!).

But in most textbooks and classes we consider five type of masses.

Definitions (#9.1.410):

D1. The "invariant mass" or "proper mass" (also sometimes named "effective mass") denoted by m_0 is scalar quantity, an intrinsic property of a body (preserved by Poincaré transformations, and with some clauses, in General Relativity). It is constant for elementary subatomic particles (each species has its value of effective mass), as well as for atoms (and similar objects) on a specified energy level. It does not have a conservation law. It is not strictly constant for large bodies (for example, increases with heat because of the relation $E=mc^2$ that we will introduce later during our study of Special Relativity at page 5679).

D2. The "rest mass" m_0 is the same as invariant mass if it is positive. For particles with zero invariant mass, strictly speaking, rest mass is undefined (although both terms might be synonymous in colloquial speech)! A separate term is motivated by the fact that massive bodies, and only them, have their rest frames (may be also named "comoving frame").

D3. The "relativistic mass" is the implicit mass of the photon in its linear momentum p given by the Einstein's relation (see later our study of Special Relativity at page 5689). The "relativistic mass" must not be confused with the "apparent relativistic mass" given by γm_0 where γ is the Lorentz-Fitzgerald coefficient that we will introduce in our study of Special Relativity later (see page 5739).

Caution! We don't like the use of the word "photon", which dates from 11926 (holocene calendar) and was first used by the chemist Gilbert N. Lewis. In our view, there is no such thing as a photon that is useful when Quantum Electrodynamics is mastered at its finest! Only a comedy of errors and historical accidents led to the popularity of the word "photon" among physicists and optical scientists. However we admit that the word is short and convenient. Its use is also habit forming (similar objections are possible to use of the word "phonon", which dates from 11932).

D4. The "inertial mass" is the "m_i" thing in Newtonian mechanic (i.e. in the $F=m_ia$ relation that we will introduce later). Thought to be a conserved scalar quantity. For low speeds approximately equals to the rest mass m_0, but generally an obsolete concept since relativistic dynamics. Depending on context, is superseded in relativity either with the rest mass or (along momentum vectors) with 4-momenta. One can't consistently define inertial mass of a photon, for example.

D5. The "gravitational mass" m_g supposed as gravitational charge (for Newton's gravity law or some another theory). The interesting thing is that, physically, no difference has been found so far between gravitational and inertial mass. Many experiments have been performed to check the values and the experiments always agree to within the margin of error for the experiment. Albert Einstein used the fact that gravitational and inertial mass were equal to begin his theory of General Relativity (see page 5720) in which he postulated that gravitational mass was the same as inertial mass.

We also differentiate the "active gravitational mass" denoted M, from the "passive gravitational mass" denoted m in the Newton's gravitational law that we will introduce later (see page 5255):

$$F=m\frac{GM}{r^2} \tag{9.1.22}$$

But by the Newton's third law (see page 3209) lead us to assume (and this is verified experimentally!) that both mass (active and passive) are equivalent such that we can also write:

$$F=M\frac{Gm}{r^2} \tag{9.1.23}$$

Hence the fact that we speak only and simply about the "gravitational mass" (without specifying if it is active or passive since they seem to be equivalent).

Remarks

R1. Experiments have shown that the inertial and gravitational masses were proportional to the ten-billionth. This experimental identity named "Galilean principle of equivalence" is the basis of the postulates of General Relativity (see section of the same name page 5720) such that we assume:

$$m_i=m_g \tag{9.1.24}$$

that we denote simply by the letter m.

R2. Unlike electric charges (see further below the definition of "electric charge"), which characterize the amplitude of interaction by the electric force, it seems that they are only positive masses. Indeed, the electric charges can push or attract between them.

There are other type of masses like the "Bare Mass" and the "Self Energy Mass" in Quantum Physics but they are actually out of the scope of this textbook!

In addition, the mass being an additive property (so "extensive" as we have already said) of matter, for a system of n masses m_i $(i=1\ldots n)$, the total mass is:

$$M_{\text{tot}}=\sum_{i=1}^{n} m_i \tag{9.1.25}$$

Similarly, for a continuous volume distribution, the total mass of a system of volume V is given by:

$$M_{\text{tot}}=\iiint\limits_{V}\rho_V(A)\mathrm{d}V=\iiint\limits_{V=\{A|\vec{r}=\overrightarrow{\mathrm{O}A}\}}\rho(\vec{r})\circ\mathrm{d}^3\vec{r} \tag{9.1.26}$$

where $\rho_V(A)$ is the "mass density" or "volume density" of the system at point A and where $\rho_V(\vec{r})$ is the density of the system at the point indicated by $\vec{r}=\overrightarrow{\mathrm{O}A}$ (this is what is meant in the braces below the second triple integral).

Remarks

R1. The triple integral (on the three dimensions of space), can be reduced to a simple integral when exploiting the possible symmetries of the system.

R2. The calculation of the mass in a non-continuous distribution (discrete) material should be done with the vector components calculated separately. Once done, then it is appropriate to take the norm.

Therefore $\rho_V(A)$ is the mass density of a material elements, centered around A, of characteristic dimensions of those of the system, but large compared to the inter-atomic distances in the system defined by:

$$\rho_V=\frac{\mathrm{d}m}{\mathrm{d}V} \tag{9.1.27}$$

If the system is homogeneous we have obviously:

$$\rho_V=\frac{M_{\text{tot}}}{V} \tag{9.1.28}$$

Remark

The mass density is an intensive quantity. Indeed, the density of a physical system is not equal to the sum of these densities (this is common sense!). The reader will notice that this intensive quantity that is the density is the ratio of two extensive quantities.

Definitions (#9.1.411):

D1. We say that a system is a "massic homogeneous system" if its volumetric, surfacic, linear densities (see definition below) are constant on any small element of volume that constitute it.

D2. We say that a system is an "isotropic system", if its physical properties are identical in all points and directions.

We also define sometimes the "surfacic mass density" (or simply "surfacic density") for systems with virtually no thickness and a "linear mass density" (or simply "linear density") for systems with also negligible width relatively to their length. We then have (S being a surface and s curvilinear abscissa):

$$M_{\text{tot}}=\iint\limits_S \rho_S(A)\mathrm{d}S=\iint_S \rho_S(\vec{r})\mathrm{d}^2\vec{r} \quad \text{and} \quad M_{\text{tot}}=\int\limits_L \rho_L(s)\mathrm{d}s=\int\limits_L \rho_L(\vec{r})\mathrm{d}\vec{r} \qquad (9.1.29)$$

with in the general case:

$$\rho_S=\frac{\mathrm{d}m}{\mathrm{d}S} \qquad \rho_L=\frac{\mathrm{d}m}{\mathrm{d}L} \qquad (9.1.30)$$

If the system is homogeneous we have again obviously with the traditional notation:

$$\sigma:=\rho_S=\frac{M}{S} \qquad \gamma:=\rho_L=\frac{M}{L} \qquad (9.1.31)$$

Remark

Notice the interesting following relationship with the delta Dirac function (see section Differential and Integral Calculus page 1569) if we take the special case of linear density (but the same can be generalized to the volumic case!):

$$\int\limits_{-\infty}^{+\infty}\rho(x)\mathrm{d}x=\int\limits_{-\infty}^{+\infty}m\delta(x)\mathrm{d}x=m \qquad (9.1.32)$$

Hence:

$$\rho(x)=m\delta(x) \qquad (9.1.33)$$

This relationship is also used a lot in quantum field theory with the energy density rather than the mass density.

Definition (#9.1.412): With the above, we can define the "density" as the number of identical elements and all countable per unit volume, or unit surface or unit linear.

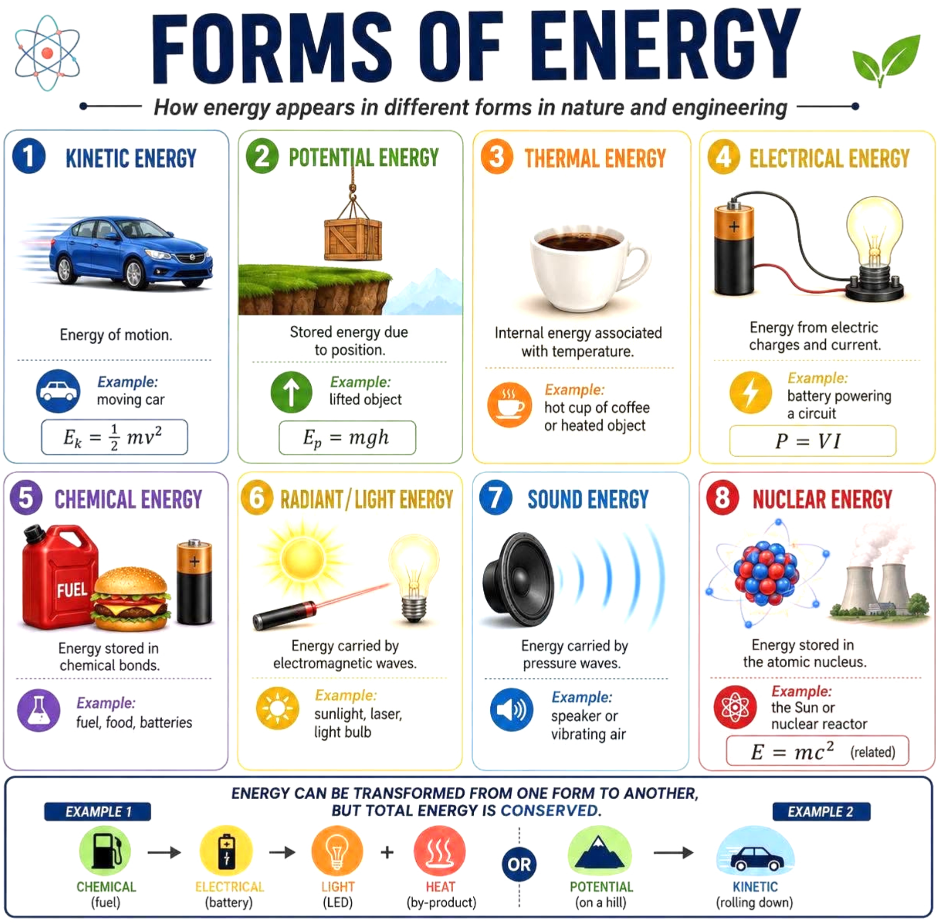
The density of a material varies with temperature and pressure. This variation is typically small for solids and liquids but much greater for gases. Increasing the pressure on an object decreases the volume of the object and thus increases its density. Increasing the temperature of a substance (with a few exceptions) decreases its density by increasing its volume. In most materials, heating the bottom of a fluid results in convection of the heat from the bottom to the top, due to the decrease in the density of the heated fluid. This causes it to rise relative to more dense unheated material.

Caution! The opinion that everything in our Universe is made of mass, opinion named in philosophy "materialism", has been definitely proven wrong with the confirmation of the discovery of the Higgs boson (and its corresponding field) the 4th July 12022 (holocene calendar) at CERN in Geneva (Switzerland). We know today that the mass is a property acquired by the excitation of the various existing quantum fields through another quantum field named the "Higgs field". The Brout-Englert-Higgs mechanism (11964 according to holocene calendar) introduced the Higgs field, whose quantum manifestation is the Higgs boson. Only particles that interact with the Higgs field acquire mass! The Higgs mechanism therefore makes it misleading to speak of the mass of such and such a particle, since mass is no longer an intrinsic property of elementary particles, but a measure of their interactions with the Higgs field.

9.1.1.1.4 Energy

We still do not know what exactly is energy (denoted in its general form with the letter E) but we know its effects. What we know however is that the energy can be classified can be transferred partially to other objects or converted into different forms. The main common forms being:

- "Work Energy" which is the energy generated by the application of a force on a body giving it a certain kinetic energy (thermal[2], electric kinetic energy) or potential energy (gravitational, electrostatic, magnetic potential energy) as studied in the sections of Classical Mechanics, Astronomy, Electrostatics, Electrodynamics, Thermodynamics and Continuum Mechanics.
- "Heat Energy" which is a form of energy determined by the number of micro-states of a system as seen in the sections of Thermodynamics and Statistical Mechanics.
- "Mass Energy" is the energy contained in a certain amount of mass (nuclear energy, chemical energy) as seen in the sections of Special Relativity and Statistical Mechanics.
- "Radiation energy" is the energy held in light waves (electromagnetic wave of a particle without rest mass), allowing them to travel across space.


FORMS OF ENERGY
How energy appears in different forms in nature and engineering
1 KINETIC ENERGY
Energy of motion.
Example: moving car
$E_k = \frac{1}{2}mv^2$
2 POTENTIAL ENERGY
Stored energy due to position.
Example: lifted object
$E_p = mgh$
3 THERMAL ENERGY
Internal energy associated with temperature.
Example: hot cup of coffee or heated object
4 ELECTRICAL ENERGY
Energy from electric charges and current.
Example: battery powering a circuit
$P = VI$
5 CHEMICAL ENERGY
FUEL
Energy stored in chemical bonds.
Example: fuel, food, batteries
6 RADIANT/LIGHT ENERGY
Energy carried by electromagnetic waves.
Example: sunlight, laser, light bulb
7 SOUND ENERGY
Energy carried by pressure waves.
Example: speaker or vibrating air
8 NUCLEAR ENERGY
Energy stored in the atomic nucleus.
Example: the Sun or nuclear reactor
$E = mc^2$ (related)
ENERGY CAN BE TRANSFORMED FROM ONE FORM TO ANOTHER, BUT TOTAL ENERGY IS CONSERVED.
EXAMPLE 1
CHEMICAL (fuel)
ELECTRICAL (battery)
LIGHT (LED)
HEAT (by-product)
OR
EXAMPLE 2
POTENTIAL (on a hill)
KINETIC (rolling down)

---

[2] The virial theorem, as proved in the section of Continuum Mechanics page 3897, indicates the absolute temperature as proportional to the average kinetic energy of the random microscopic motions of those of their constituent microscopic particles, such as electrons, atoms, and molecules, that move freely within the material.

The total energy of a system can be subdivided and classified in various ways. For example, Classical Mechanics distinguishes between kinetic energy, which is determined by an object's movement through space, and potential energy, which is a function of the position of an object within a field. It may also be convenient to distinguish gravitational energy, thermal energy, several types of nuclear energy (which utilize potentials from the nuclear force and the weak force), electric energy (from the electric field), and magnetic energy (from the magnetic field), among others. Many of these classifications overlap; for instance, thermal energy usually consists partly of kinetic and partly of potential energy.

Some types of energy are a varying mix of both potential and kinetic energy. An example is mechanical energy which is the sum of (usually macroscopic) kinetic and potential energy in a system. Elastic energy in materials is also dependent upon electrical potential energy (among atoms and molecules), as is chemical energy, which is stored and released from a reservoir of electrical potential energy between electrons, and the molecules or atomic nuclei that attract them..The list is also not necessarily complete. Whenever physical scientists discover that a certain phenomenon appears to violate the law of energy conservation, new forms are typically added that account for the discrepancy.

Potential energies are often measured as positive or negative depending on whether they are greater or less than the energy of a specified base state or configuration such as two interacting bodies being infinitely far apart. Wave energies (such as radiant or sound energy), kinetic energy, and rest energy are each greater than or equal to zero because they are measured in comparison to a base state of zero energy: "no wave", "no motion", and "no inertia", respectively.

We can still try to ask ourselves what is exactly energy?

Definition (#9.1.413): The "energy" is the observable of the result of dynamic patterns of modifications or a conservation property of a system (many textbooks define energy as the variation of work but this is circular definition). These dynamic patterns may not be necessarily deterministic and of zero mean and conservative in a closed system! The SI derived unit of energy is the Joule [J]. It is also common in Nuclear Physics, Plasma Physics and Particle Physics to use the electron-volt that is by definition the energy gain by a charge of a single electron moving across an electric potential difference of one volt and equal to $1.602176634\cdot 10^{-19}$ [J] (see section Nuclear Physics page 5027).

It is essential to exercise caution when encountering assertions from individuals such as charlatans, enthusiasts, popular science publications, and some engineers, who claim that «*everything is energy!*». Contemporary scientific evidence disproves this assertion, as numerous fundamental properties of nature cannot be solely described in terms of energy. These include angular momentum and momentum (which are two of the three conserved quantities explained by Noether's theorem, the third being energy), spin, quark colors, the chirality of photons, the polarity of light, and neutrino flavors among many others. Consequently, you will not find any peer-reviewed papers in prestigious scientific journals like Physical Review Letters stating «*everything is energy!*», as such a statement is simply incorrect. An ancient Greek proverb states «*Mēden agan*», which can be translated to «*Excess is wisdom's enemy*». As always, excessive simplification indeed makes things simpler—but at the risk of drawing conclusions that verge on absurdity.

As we will see in the section of Analytical Mechanics during our study of Lagrangian formalism, the energy can be defined as the conserved quantity associated with time translation symmetry of the Lagrangian (see page 3132).

Remarks

R1. The mass and energy are equivalent as we will see during our study of Special Relativity (see section Special Relativity page 5627), if we define a system of units such that the speed of light system is $c = 1$ (widely used convention by physicists in advanced research) the mass is then just equal to the energy when at rest (since $E = mc^2$ in natural units becomes $E = m$)!

R2. The energy, in the same way that mass, is an extensive quantity.

R3. In an isolated system and in a static space-time, they cannot be spontaneous creation or destruction of energy by definition. Spontaneous apparition of energy can only be due to an external action in such a system. Another way of saying the same thing is that the total energy in the Universe is constant if the latter is considered as isolated and not expanding.

R4. In certain study fields, such as plasma physics and particle physics, it is sometimes common to convert the electron-volt (energy) as a unit of temperature as it helps for comparison analysis with an intuitive everyday life quantity (we will have such examples in this textbook). The conversion to the Kelvin scale is defined by using the Boltzmann constant k (see section Continuum Mechanics page 3921):

$$\frac{1\,[\text{eV}]}{k} = \frac{1.602176634 \cdot 10^{-19}\,[\text{J}]}{1.380649 \cdot 10^{-23}\,[\text{J} \cdot \text{K}^{-1}]} \cong 11604.518121\,[\text{K}] \tag{9.1.34}$$

For more detail and a criticism of the conversion, the reader must refer to the section of Nuclear Physics.

To finish, let us notice that as the mass and energy are equivalent to a given factor (c), we can obviously also define an energy volumetric density, energy surfacic density and energy linear density with also the same three remarks that apply.

You will notice in all calculations in this textbook that the Boltzmann constant $k = 1.380649 \cdot 10^{-23}\,[\text{m}^2 \cdot \text{kg} \cdot \text{s}^{-2} \cdot \text{K}^{-1}]$ and T almost always shows up together; it's no accident, k is just a dummy factor which converts energy to temperature. That's why in high level physics, practitioners use natural units where $k = 1$ such that $E = T$.

9.1.1.1.5 Electric Charge

It is difficult to say something about the electric charge (you can search a definition on Internet you'll see ...). However, if we refer to the approach of theoretical physicist Hideki Yukawa we can try to give the following definition:

Definition (#9.1.414): The "electric charge" q seems to be a conservative physical property of matter that causes it to experience a force when placed in an electric or magnetic field. There are two types of electric charges: positive and negative. Like charges repel and unlike attract. An object (one not made of antimatter) is negatively charged if it has an excess of electrons, and is otherwise positively charged or uncharged. The SI derived unit of electric charge is the coulomb [C]. In electrical engineering, it is also common to use the ampere-hour [Ah], and, in chemistry, it is common to use the elementary charge (e) as a unit. The symbol Q often denotes multiple electric charges. Early knowledge of how electrically charged substances interact is now named "Classical Electrodynamics" (see section Electrodynamics page 4094), and is still accurate for problems that don't require consideration of quantum effects (see section Quantum Field Theory page 5192).

More technically... (we refer to what we will see in the section of Quantum Field Theory at page 5158) an another but still not accurate definition is to say that: An "electric charge" q is a conservative property that has a particle lying in a spherically symmetric potential field to interact with the source of that field in the context of the exchange of a quantum of interaction (the photon is the quantum of this interaction generated by the quantum fluctuations of vacuum in the presence of a mass) defining a vector field of the Coulomb type.

Or another definition similar to that of the mass: The "electric charge" q of a body in a closed system is a conservative quantity and that characterizes algebraically the amplitude with which the body interacts with other body through electrostatic and magnetic forces.

The electric charge is a fundamental conserved property of some subatomic particles, which determines their electromagnetic interaction. Electrically charged matter is influenced by, and produces, electromagnetic fields. The interaction between a moving charge and an electromagnetic field is the source of the electromagnetic force, which is one of the four fundamental forces.

Remarks

R1. In an isolated system, globally they cannot be spontaneous creation or removal of electric charges. The emergence of electric charges can only be due to an external action as far as we know. Another way of saying the same thing - as for the mass - is that the total charge contained an isolated system seems to be constant.

R2. The charge is an extensive quantity. Indeed, the total electric charge of a physical system is equal to the algebraic sum of the charges which constitutes it.

120th century (holocene calendar) experiments demonstrated that electric charge is quantized; that is, it comes in integer multiples of individual small units named the "elementary charge", e, approximately equal to $1.602176634 \cdot 10^{-19}$ [C] (except for particles named quarks, which have charges that are integer multiples of $e/3$). The proton has a charge of $+e$, and the electron has a charge of $-e$. The study of charged particles, and how their interactions are mediated by photons, is named "Quantum Electrodynamics" (see section Elementary Particle Physics page 5194).

As already mentioned, contrary to the mass, there are positive and negative electric charges this is why we use the term "algebraic". The electric charge remains an additive property (extensive). Thus, for a system of q_i electric charges, the total electric charge is therefore:

$$Q := Q_{\text{tot}} = \sum_{i=1}^{n} \pm q_i \tag{9.1.35}$$

The electric charge of a macroscopic object is the algebraic sum of the electric charges of the particles that make it up. This charge is often small, because matter is made of atoms, and atoms typically have equal numbers of protons and electrons, in which case - if the distribution of charges can be considered as homogeneous[3] - their charges cancel out, yielding a net charge of zero, thus making the atom neutral.

Similarly as for the mass and the energy, for a total continuous volumetric distribution of the electric charge of a system of volume V (we denote the electric charge density in an identical way as that of the mass if the ambiguity is not possible):

$$Q_{\text{tot}} = \iiint_V \rho_V(A)\mathrm{d}V = \iiint_{v=\{A|\vec{r}=\overrightarrow{\mathrm{O}A}\}} \rho_V(\vec{r}) \circ \mathrm{d}^3\vec{r} \tag{9.1.36}$$

where $\rho_V(A)$ is the "volumetric electric charge density" of the system at point A, that is to say the electric charge of a material element, centered around A, of characteristic dimensions in comparison to that of the system, but large in comparison of the inter-atomic distances in this system ($\rho_V(\vec{r})$ is the volumetric density of electric charges of the system to the point indicated by $\vec{r} = \overrightarrow{\mathrm{O}A}$) defined by:

$$\rho_V = \frac{\mathrm{d}q}{\mathrm{d}V} \tag{9.1.37}$$

with also the same three remarks that apply.

As for the mass, we can give the following definitions:

Definitions (#9.1.415):

D1. We say that a system is an "electric homogeneous system" if its volumetric, surfacic, linear charge densities (see definition below) are constant on any small element of volume that constitute it.

D2. We say that a system is an "electric isotropic system", if its electric properties are identical in all points and directions.

---

[3] For example the disposition of the quarks in the Neutron is not homogeneous this is why even if the Neutron seems... neutral under a strong magnetic field the neutron will have a precession movement and this is why even if the neutron has an electric charge that is to weak to be deviated by a magnetic field it is not small enough to hide the Larmor effect.

We also define sometimes the "electric charge surfacic density" (or simply "surfacic density") for systems with virtually no thickness and a "electric charge linear density" (or simply "linear density") for systems with also negligible width relatively to their length. We then have (S being a surface and s curvilinear abscissa):

$$Q_{\text{tot}}=\iint\limits_S \rho_S(A)\mathrm{d}S=\iint_S \rho_S(\vec{r})\mathrm{d}^2\vec{r} \quad \text{and} \quad Q_{\text{tot}}=\int\limits_L \rho_L(s)\mathrm{d}s=\int\limits_L \rho_L(\vec{r})\mathrm{d}\vec{r} \tag{9.1.38}$$

with in the general case:

$$\rho_S=\frac{\mathrm{d}q}{\mathrm{d}S} \qquad \rho_L=\frac{\mathrm{d}q}{\mathrm{d}L} \tag{9.1.39}$$

If the system is electrically homogeneous we have again obviously with the traditional notation:

$$\sigma:=\rho_S=\frac{Q}{S} \qquad \gamma:=\rho_L=\frac{Q}{L} \tag{9.1.40}$$

Definition (#9.1.416): With the above, we can define the "electric density" as the number of identical elements and all countable per unit volume, or unit surface or unit linear.

The electric charge density of a material varies, as for the mass, with temperature and pressure. This variation is typically small for solids and liquids but much greater for gases. Increasing the pressure on an object decreases the volume of the object and thus increases its density. Increasing the temperature of a substance (with a few exceptions) decreases its density by increasing its volume. In most materials, heating the bottom of a fluid results in convection of the heat from the bottom to the top, due to the decrease in the density of the heated fluid. This causes it to rise relative to more dense unheated material.

9.1.1.2 Scientific Notation and Metric Prefixes

Ok just now we have defined the concept of quantity/magnitude and associated main units as used by the physicists and engineers. But what about the scientific notation for big numbers (magnitudes) and associated metric prefixes??? Let us see this now!

Scientific notation (also referred to as standard form or standard index form) is a way of expressing numbers that are too big or too small to be conveniently written in decimal form. It is commonly used by scientists, mathematicians and engineers. On scientific calculators it is known as "SCI" display mode.

Indeed, it is common in physics that the manipulated variables are very large and heavy to write. For example, it is always a hassle to have magnitudes as $8,000,000,000$ or $0.000,000,000,1$.

In scientific notation all numbers are written in the form:

$$m\cdot 10^n \tag{9.1.41}$$

(m times ten raised to the power of n), where the exponent n is an integer, and the coefficient m is any real number, named the "significand" or "mantissa".

Definition (#9.1.417): We say that $m\cdot 10^n$ is "scientific notation" of a given number if $-10<m<10$ (that is to say that m is written with a single digit other than zero before the decimal point) and if n is a signed integer.

Remark

"Engineering notation" (often named "ENG" display mode on scientific calculators) differs from normalized scientific notation in that the exponent n is restricted to multiples of 3.

✍**Examples:**

E1. The number $8,000,000,000$ will be written in scientific and engineer notation as:

$$8\cdot 10^9 \tag{9.1.42}$$

That is to say 9 zeros after the eight.

E2. The number $0.000,000,000,1$ will be written in scientific notation:

$$1\cdot 10^{-10} \tag{9.1.43}$$

That is to say 10 zeros before the one. Or in engineering notation:

$$100\cdot 10^{-12} \tag{9.1.44}$$

The advantage of the scientific notation is to give an order of magnitude of a number between two consecutive powers of 10 as:

$$10^8<8\cdot 10^9<10^{10} \qquad 10^{-11}<1\cdot 10^{-10}<10^{-9} \tag{9.1.45}$$

Remarks

R1. If we have a number like $154,434,347,786$ depending on the context we allow ourselves to truncate it and then write it with a precision to three decimal places and the latter number becomes $154.434 \cdot 10^9$ in engineering notation which is easier to write but dangerous to handle because of the error induced by the truncation. We refer the reader to the subject of Errors Propagation in the section Statistics at page 1307 for more informations on the topic.

R2. For the mathematicians the scientific notation is only one writing of a number among many and the choice of this writing is related to the context of the problem. Obviously these "results numbers " obtained may be simple numbers coming from abstract but also practical problems from experiences, measurement operations, etc. and here we will agree with the way of physicists.

A "metric prefix" is a unit prefix that precedes a basic unit of measure to indicate a multiple or fraction of the unit.

Decimal multiplicative prefixes have been a feature of all forms of the metric system with six dating back to the system's introduction in the 11790s (holocene calendar).

Each prefix has a unique symbol that is pretended to the unit symbol. The prefix kilo-, for example, may be added to gram to indicate multiplication by one thousand: one kilogram is equal to one thousand grams. The prefix milli-, likewise, may be added to meter to indicate division by one thousand; one millimetre is equal to one thousandth of a meter.

The prefix and corresponding magnitudes are given in the table below:

Metric Prefix				Scientific Notation	
Prefix	**Symbol**	**Multiplier**	**Exponential**	**Normalized**	**E Notation**
yotta	Y	$1,000,000,000,000,000,000,000,000,000$	10^{24}	$1.0 \cdot 10^{24}$	1.0E24
zetta	Z	$1,000,000,000,000,000,000,000$	10^{21}	$1.0 \cdot 10^{21}$	1.0E21
exa	E	$1,000,000,000,000,000,000$	10^{15}	$1.0 \cdot 10^{18}$	1.0E18
peta	P	$1,000,000,000,000,000$	10^{15}	$1.0 \cdot 10^{15}$	1.0E15
tera	T	$1,000,000,000,000$	10^{12}	$1.0 \cdot 10^{12}$	1.0E12
giga	G	$1,000,000,000$	10^{9}	$1.0 \cdot 10^{\circ}$	1.0E9
mega	M	$1,000,000$	10^{6}	$1.0 \cdot 10^{6}$	1.0E6
kilo	k	$1,000$	10^{3}	$1.0 \cdot 10^{3}$	1.0E3
hecto	h	100	10^{2}	$1.0 \cdot 10^{2}$	1.0E2
deca	da	10	10^{1}	$1.0 \cdot 10^{1}$	1.0E1
		1	10°	$1.0 \cdot 10^{\circ}$	1.0E0
deci	d	0.1	10^{-1}	$1.0 \cdot 10^{-1}$	1.0E-1
centi	c	0.01	10^{-2}	$1.0 \cdot 10^{-2}$	1.0E-2
milli	m	0.001	10^{-3}	$1.0 \cdot 10^{-3}$	1.0E-3
micro	μ	0.000001	10^{-6}	$1.0 \cdot 10^{-8}$	1.0E-6
nano	n	0.000000001	10^{-9}	$1.0 \cdot 10^{-9}$	1.0E-9
pico	p	0.000000000001	10^{-12}	$1.0 \cdot 10^{-12}$	1.0E-12
femto	f	0.000000000000001	10^{-15}	$1.0 \cdot 10^{-15}$	1.0E-15
atto	^{a}	0.000000000000000001	10^{-18}	$1.0 \cdot 10^{-18}$	1.0E-18
zepto	z	0.0000000000000000000001	10^{-21}	$1.0 \cdot 10^{-21}$	1.0E-21
yocto	y	$0.000000000000000000000001$	10^{-24}	$1.0 \cdot 10^{-24}$	1.0E-24

Table 9.1: **Metric prefix**

For example, $10,000,000$ grams conventionally denoted by:

$$10,000,000\,[\text{g}] \tag{9.1.46}$$

will be written in scientific notation:

$$1 \cdot 10^7\,[\text{g}] \tag{9.1.47}$$

but with a physicist notation (according to the previous table):

$$1 \cdot 10^4\,[\text{kg}] \quad \text{or} \quad 10\,[\text{Mg}] \tag{9.1.48}$$

9.1.1.3 Scales of Measurements

As the main part of physics is understanding our surrounding it seemed important to us to give to the reader some scales of measurements of typical quantities. We own the tables below to William Nielsen Brandt.

Some pressures values first:

Place	**Average pressure**	**Mean free path**	**Molecules by [cm^3]**
Rough neutron star pressure	$1\cdot 10^{34}$ [Pa]		
Rough White Dwarf pressure	$1\cdot 10^{23}$ [Pa]		
Rough Sun central pressure	$27,000$ [GPa]		
Central pressure of the Earth	100 [GPa]		
Pressure needed for natural diamonds crystallization	10 [GPa]		
Conventional high pressure laboratory press	6 [GPa]		
Pressure at the bottom of the Marianas trench	110 [MPa]		
Karate fist peak pressure	90 [MPa]		
Typical scuba tank pressure	20 [MPa]		
High pressure bicycle tire	1.5 [MPa]		
Standard atmosphere	~ 100 [kPa]	~ 66 [nm]	$\sim 2.5\cdot 10^{19}$
Vacuum cleaner	~ 88 [kPa]	~ 70 [nm]	$\sim 10^{19}$
Steam turbine	~ 9 [kPa]		
Standing person	15 [Pa]		
Human arterial blood	10 [Pa]		
Mars	~ 0.6 [kPa]		
Freeze Dryer	~ 50 [Pa]	~ 500 [μm]	$\sim 10^{16}$
Pressure of a sound at the human threshold of pain	30 [Pa]		
Radiation pressure at the surface of the Sun	0.2 [Pa]		
Thermos bottle	1 to 0.01 [Pa]	~ 1 [m]	$\sim 10^{13}$
Vacuum tube	0.1 [μPa]	$\sim 1,000$ [km]	$\sim 10^{7}$
Human threshold of hearing	0.01 [μPa]		
Cryopump (LHC/LIGO)	0.01 [μPa]	$\sim 10,000$ [km]	$\sim 10^{6}$
Solar radiation pressure	0.005 [μPa]		
Moon	0.1 [nPa]	$\sim 10,000$ [km]	$\sim 10^{5}$
Interplanetary space	$0.7\cdot 10^{-12}$ [Pa]		~ 11
Best vacuum performed on Earth (Tokyo University)	10^{-16} [Pa]		
Interstellar space	10^{-17} [Pa]		~ 1
Intergalactic space			~ 0.000001

Table 9.2: **Typical vacuum values and mean free paths**

Remark

It is common to define the unit of 1 [bar] as the atmospheric pressure at sea level on Earth equal by definition to $100,000$ [Pa], but this unit is not approved as part of the International System of Units.

And some time scales:

Magnitude [s]	**Description**
$5 \cdot 10^{-44}$	Planck-Wheeler time $= \left(\frac{G\hbar}{c^5}\right)^{\frac{1}{2}}$
$4 \cdot 10^{-24}$	Typical lifetime of strong interaction resonance $= \frac{h}{m_{\mathrm{p}}c^2}$
$274 \cdot 10^{-18}$	Shortest span of time actually measured by physicists
$1 \cdot 10^{-13}$	Typical period of vibration of an atom in a solid
$1 \cdot 10^{-13}$	Typical X-ray line electric dipole radiative transition time
$1.6 \cdot 10^{-9}$	Typical hydrogen 2p→1s radiative transition time (electric dipole one photon process)
$8 \cdot 10^{-4}$	Mass shedding minimum spin period for a neutron star
$1.6 \cdot 10^{-3}$	Spin period of PSR 1957+20
0.12	Typical hydrogen 2s→1s radiative transition time
10	Median duration of a classical γ-ray burst
887	Mean life of a neutron in free space
$2,000$	Sun dynamic time scale
$8.6 \cdot 10^{4}$	Earth's rotation time
$3.2 \cdot 10^{7}$	Earth orbit time around the Sun
$1.6 \cdot 10^{9}$	Typical time between Milky Way supernovae
$1.9 \cdot 10^{11}$	Carbon-14 half-life
$3 \cdot 10^{13}$	Rough time for evolution of a biological species
$3 \cdot 10^{14}$	Rough Lyapunov time of the Solar System
$6.3 \cdot 10^{14}$	Sun thermal time scale
$2 \cdot 10^{15}$	Timescale for Los Angeles to pass San Francisco via continental drift
$7.3 \cdot 10^{15}$	Orbit time for Sun around galaxy center
$2 \cdot 10^{16}$	Rough supernova biological extinction time
$6 \cdot 10^{16}$	Time for galaxy to cross a cluster
$1.1 \cdot 10^{17}$	Primeval slime to man time
$1.5 \cdot 10^{17}$	Age of Earth (see [599]) and Sun
$1.5 \cdot 10^{17}$	Uranium-238 half-life
$2.7 \cdot 10^{17}$	Look back time to $z = 1$
$3 \cdot 10^{17}$	Main sequence lifetime for a 1 $\mathrm{M}_{\odot}$ star
$3.3 \cdot 10^{17}$	Look back time to $z = 2$
$3.3 \cdot 10^{17}$	Sun nuclear time scale
$3.7 \cdot 10^{17}$	Look back time to $z = 4$
$3.8 \cdot 10^{17}$	Rough age of the Milky Way
$4.1 \cdot 10^{17}$	Age of the universe $= \frac{2}{3H_0}$
$1 \cdot 10^{39}$	Lower limit on the proton lifetime
$4.7 \cdot 10^{73}$	1 $\mathrm{M}_{\odot}$ Hawking's Black Hole

Table 9.3: **Typical time scales**

And some length scales:

Magnitude [m]	Description
$1.6 \cdot 10^{-35}$	Planck length
$2 \cdot 10^{-35}$	Rough postulated superstring size
$1.5 \cdot 10^{-18}$	Classical proton radius
$1.6 \cdot 10^{-17}$	Rough weak force length
$4 \cdot 10^{-17}$	LIGO 4 km gravity-wave detector needed sensitivity
$1.44 \cdot 10^{-15} A^{\frac{1}{3}}$	Nuclear radius
$2.8 \cdot 10^{-15}$	Classical electron radius
$8.8 \cdot 10^{-15}$	Attractive strong force length
$5.3 \cdot 10^{-11}$	Bohr radius (electron-proton distance)
$2.6 \cdot 10^{-10}$	Copper atom spacing in solid copper
$3.5 \cdot 10^{-10}$	H_2O molecular diameter
$4 \cdot 10^{-10}$	ROSAT X-ray satellite mirror rms surface error
$\sim 3 \cdot 10^{-9}$	Mean nucleon spacing on primordial nucleosynthesis, DNA turn length
$1 \cdot 10^{-7}$	Typical size of a virus, Interstellar dust size
$4 \cdot 10^{-6}$	Typical size of a cell
$2 \cdot 10^{-4}$	Small dust particle size
$0.5 - 2.5$	Human Adult tall range
$3,700$	Mean ocean depth
$5,500$	Rough radius of Halley's comet
$8,847$	Height of Mount Everest
$10,000$	Typical Neutron star, asteroid or comet radius
$11,032$	Mariana Trench depth's, Troposphere height, Airliner cruising altitude
$30,000$	Typical thickness of the Earth's crust
$3.2 \cdot 10^{6}$	Length of the Great Wall of China
$6.3 \cdot 10^{6}$	Radius of the Earth
$4.2 \cdot 10^{7}$	Geostationary satellite orbit height
$7.1 \cdot 10^{7}$	Radius of Jupiter
$9 \cdot 10^{7}$	Distance to the Earth's solar wind bow shock
$3.8 \cdot 10^{8}$	Distance to the Moon
$7.0 \cdot 10^{8}$	Radius of the Sun
$1.50 \cdot 10^{11}$	Earth/Sun mean distance
$5 \cdot 10^{11}$	Radius of the 20 $M_{\odot}$ red supergiant Betelgeuse at maximum light
$5.91 \cdot 10^{12}$	Pluton/Sun mean distance
$4 \cdot 10^{14}$	Rough stellar separation in a globular cluster
$5 \cdot 10^{17}$	Typical interstellar medium cloud size
$2 \cdot 10^{18}$	Rough radius of the local interstellar hot gas bubble
$5 \cdot 10^{18}$	Rough height of Milky Way, Distance to Betelgeuse supergiant
$2 \cdot 10^{19}$	Characteristic height of the Milky Way main disk
$2.4 \cdot 10^{20}$	Distance from Sun to galactic center
10^{22}	Typical active galaxy jet length and rough local group radius
$7 \cdot 10^{23}$	Distance to the center of the Virgo cluster of galaxies
$3 \cdot 10^{25}$	Schwarzschild radius of a singularity with Universe critical density

Table 9.4: **Typical distances scales**

And some velocities scales:

Magnitude [m·s^{-1}]	**Description**
$1 \cdot 10^{-9}$	Sea floor spreading rate
$1.6 \cdot 10^{-9}$	Average slip rate of the San Andreas fault
$1 \cdot 10^{-8}$	Typical rainfall rate in a semi-arid climate
$2 \cdot 10^{-8}$	Grass growth rate
$3 \cdot 10^{-6}$	Typical glacial advance rate
$1 \cdot 10^{-3}$	Equivalent radial velocity resolution of pulsar pulse arrival time analysis
1.3	Human walking speed
3	Radial velocity accuracy of high precision Doppler spectroscopy
13	Speed of the reflex motion induced on the Sun by Jupiter
25	Car speed
100	Typical speed of an electric pulse in the nervous system
330	Sound speed in air
480	Earth's atmosphere molecular rms velocity
600	Fighter jet speed
$2,380$	Escape velocity from Moon's surface
$10,000$	Typical longitudinal seismic wave velocity in the Earth's mantle
$11,000$	Escape velocity from the Earth's surface
$20,000$	Globular cluster stellar velocity dispersion
$29,000$	Earth's motion around the Sun
$40,000$	Globular cluster stellar escape velocity
$1 \cdot 10^{5}$	Typical Galactic pulsar vertical velocity component
$1 \cdot 10^{5}$	Average speed of the initial stroke of a lightning flash
$2.2 \cdot 10^{5}$	Rotational velocity of the Sun around the Milky Way's center
$3 \cdot 10^{5}$	Orbital speed of PSR 1913+16
$3 \cdot 10^{5}$	Rough velocity of Geminga's proper motion
$3.65 \cdot 10^{5}$	Motion of the Solar System barycentre relative to the cosmic microwave background
$6.2 \cdot 10^{5}$	Escape velocity from the Sun's surface
$6.2 \cdot 10^{5}$	Escape velocity from the Milky Way for objects in the solar neighbourhood (ARAA 29, 429)
$6.22 \cdot 10^{5}$	Motion of the Local Group relative to the cosmic microwave background
$8 \cdot 10^{5}$	Typical galaxy cluster galaxy velocity dispersion
$2 \cdot 10^{6}$	Speed of $n=1$ hydrogen electron = $Z\alpha c$
$5 \cdot 10^{6}$	Young (months old) supernova ejecta
$1.2 \cdot 10^{7}$	Velocity of the wind streaming out from H1413+117 "Cloverleaf" (*Nature* 371, 559)
$7.8 \cdot 10^{7}$	SS433 jet speed
$1.4 \cdot 10^{8}$	Keplerian orbital velocity at the surface of a neutron star
$2 \cdot 10^{8}$	Escape velocity from neutron star surface
$2.998 \cdot 10^{8}$	Light in a vacuum
$2.7 \cdot 10^{10}$	Apparent superluminal motion of the jet from the $z=0.940$ BL Lac AO 0235+164 (may be lensed)

Table 9.5: **Typical velocities scales**

And some masses scales:

Magnitude [kg]	Description
$4.2 \cdot 10^{-36}$	Mass equivalent of a green light photon
$9.1 \cdot 10^{-36}$	Electron anti-neutrino upper mass limit
$4.8 \cdot 10^{-31}$	Muon neutrino upper mass limit
$9.11 \cdot 10^{-31}$	Electron mass
$5.5 \cdot 10^{-29}$	Tau neutrino upper mass limit
$1.67 \cdot 10^{-27}$	Proton mass
$9 \cdot 10^{-27}$	Bottom quark mass
$4.8 \cdot 10^{-26}$	Mean mass of atmosphere molecule
$1.4 \cdot 10^{-25}$	$W^{\pm}$ mass (80.22 ± 0.22 GeV/c^2)
$1.6 \cdot 10^{-25}$	Z^0 mass (91.187 ± 0.007 GeV/c^2)
$2 \cdot 10^{-25}$	Favoured Higgs boson mass
$3.1 \cdot 10^{-25}$	Top quark mass (176 ± 13 GeV/c^2)
$4 \cdot 10^{-25}$	DNA nucleotide
$1 \cdot 10^{-22}$	Typical protein molecule mass
$5 \cdot 10^{-21}$	*E. Coli* ribosome
$1 \cdot 10^{-16}$	Interstellar dust grain mass
$8 \cdot 10^{-15}$	Rough mass of a human DNA molecule
$7 \cdot 10^{-13}$	Typical mass of a cell
$2.2 \cdot 10^{-8}$	Planck mass
0.02	Typical goldfish mass
$2 - 150$	Typical human mass
$500 - 2,500$	Typical car mass
$10,000$	Tyrannosaurus Rex
$1 \cdot 10^{13}$	Typical comet/mountain mass
$5.3 \cdot 10^{18}$	Total mass of Earth's atmosphere
$3 \cdot 10^{19}$	Typical asteroid mass
$1.4 \cdot 10^{21}$	Total mass of Earth's oceans
$7.3 \cdot 10^{22}$	Mass of the Moon
$5.98 \cdot 10^{24}$	Mass of the Earth
$1.9 \cdot 10^{27}$	Mass of Jupiter
$1.6 \cdot 10^{29}$	Minimum mass to fusion burn hydrogen
$1.99 \cdot 10^{30}$	Mass of the Sun
$2.8 \cdot 10^{30}$	Chandrasekhar mass (maximum mass for a white dwarf)
$6.0 \cdot 10^{30}$	Oppenheimer-Volkoff mass (maximum mass for a neutron star)
$4 \cdot 10^{31}$	Rough stellar mass above which the endpoint is a Black Hole
$1.2 \cdot 10^{32}$	Rough mass at which a star becomes unstable to pulsations
$2 \cdot 10^{33}$	Typical interstellar cloud mass
$1 \cdot 10^{36}$	Typical mass of a globular cluster
$5 \cdot 10^{36}$	Approximate mass of the Milky Way central Black Hole
$1.4 \cdot 10^{49}$	Rough total mass in spiral galaxies
$3 \cdot 10^{49}$	Rough total mass in elliptical and spheroidal galaxies
$8 \cdot 10^{49}$	Rough total mass of visible matter in the universe
$2 \cdot 10^{52}$	Rough total mass of a critical density universe

Table 9.6: **Typical masses scales**

And some densities scales:

Magnitude [kg· m^{-3}]	**Description**
$2 \cdot 10^{-38}$	Effective density of the 100-300 MHz radio background
$1 \cdot 10^{-35}$	Effective density of the 1-10 MeV γ-ray background
$8 \cdot 10^{-35}$	Effective density of the 2-100 keV X-ray background
$1.1 \cdot 10^{-33}$	Upper limit to the effective density of the gravitational wave background
$1 \cdot 10^{-32}$	Effective density of the starlight released in a Hubble time
$4.6 \cdot 10^{-31}$	Effective density of the cosmic microwave background radiation
$2 \cdot 10^{-29}$	Smoothed density of visible galactic material throughout universe
$2 \cdot 10^{-28}$	Smoothed baryon density predicted by primordial nucleosynthesis
$4.7 \cdot 10^{-27}$	Critical density of the universe $= \frac{3H_0^2}{8\pi G}$
$2 \cdot 10^{-24}$	Typical gas in a cluster of galaxies
$3 \cdot 10^{-21}$	Typical gas in the interstellar medium of the Milky Way
$7 \cdot 10^{-21}$	Dynamically inferred Milky Way disk density
$5 \cdot 10^{-20}$	Typical density of the gas in the central kiloparsec of an interacting or starburst galaxy
$1 \cdot 10^{-9}$	Best room temperature vacuum achieved on Earth
$1.7 \cdot 10^{-4}$	Mean density of Antares (19 $\text{M}_\odot$)
1.3	Density of air
700	Mean density of Saturn
$1,000$	Density of water
$1,300$	Mean density of Jupiter
$1,400$	Mean density of the Sun
$3,300$	Mean density of the Moon
$5,500$	Mean density of the Earth
$7,860$	Density of iron
$19,300$	Density of gold
$5 \cdot 10^{7}$	Typical White Dwarf mean density
$3 \cdot 10^{10}$	Typical White Dwarf central density
$1.1 \cdot 10^{12}$	Inverse β decay threshold
$4.3 \cdot 10^{14}$	Neutron drip density
$6 \cdot 10^{17}$	Nuclear density
$1 \cdot 10^{18}$	Typical neutron star central density
$5 \cdot 10^{96}$	Planck-Wheeler density, at which quantum gravitational effects become important $= \frac{c^5}{G^2\hbar}$

Table 9.7: **Typical densities scales**

And some energies scales:

Magnitude [J]	**Description**
$4 \cdot 10^{-21}$	kT_{room} = translational kinetic energy of atmosphere gas molecule
$7 \cdot 10^{-21}$	Donor level/conduction band gap in a doped semiconductor
$3 \cdot 10^{-20}$	Molecular vibration transition
$1.6 \cdot 10^{-19}$	Valence band/conduction band gap in a semiconductor
$3.8 \cdot 10^{-19}$	Green light photon
$9 \cdot 10^{-19}$	Valence band/conduction band gap in an insulator
$1.1 \cdot 10^{-18}$	Fermi energy in copper = depth of Fermi sea
$2.2 \cdot 10^{-18}$	Hydrogen $n = 1$ binding energy
$1.6 \cdot 10^{-16}$	1 keV X-ray
$1.5 \cdot 10^{-15}$	Hydrogenic iron $n = 1$ binding energy
$8.18 \cdot 10^{-14}$	Electron rest mass
$1.6 \cdot 10^{-13}$	1 MeV γ-ray
$1.6 \cdot 10^{-13}\ (Z_1 \cdot Z_2)$	Coulomb barrier height
$1.3 \cdot 10^{-12}$	Nucleon binding energy
$4.3 \cdot 10^{-12}$	Energy from $4 \cdot (^1\mathrm{H}) \rightarrow\ ^4\mathrm{He}$
$1.50 \cdot 10^{-10}$	Proton rest mass
$1.6 \cdot 10^{-7}$	Particle kinetic energy in a 1 TeV accelerator
10	Well hit tennis ball
$2 \cdot 10^{5}$	Energy from a light bulb burning for 1 hour
$2 \cdot 10^{7}$	Rough energy from a 1 kilogram meal
$4.2 \cdot 10^{9}$	Explosion energy of 1 ton of TNT
$2 \cdot 10^{11}$	Rough human total energy output in a lifetime
$1.5 \cdot 10^{14}$	Typical atomic bomb explosion energy
$6 \cdot 10^{16}$	Tunguska 50 m diameter meteorite impact energy
$1 \cdot 10^{17}$	Powerful H-bomb explosion energy
$3 \cdot 10^{17}$	Elastic wave energy release from a large $(M = 8.5)$ earthquake
$9 \cdot 10^{18}$	USA electricity usage in 11986 (holocene calendar)
$5 \cdot 10^{19}$	Rough explosion energy of the Lake Toba eruption in Sumatra
$2.5 \cdot 10^{22}$	Shoemaker-Levy fragment G impact on Jupiter
$4 \cdot 10^{23}$	K-T 10 km diameter meteorite impact energy
$1 \cdot 10^{24}$	Power released by an evaporating Black Hole the last second of its life
$2.1 \cdot 10^{29}$	The Earth's rotational energy
$3 \cdot 10^{31}$	The Earth's total heat content
$7 \cdot 10^{34}$	Total rotational energy of the planets
$2.5 \cdot 10^{35}$	Rotational energy of the Sun
$3 \cdot 10^{36}$	Gravitational internal binding energy of Jupiter
$3 \cdot 10^{41}$	A 450 km s^{-1} neutron star kick
$1.3 \cdot 10^{44}$	Total radiant energy from the Sun — $(\frac{1}{10})(0.007)\mathrm{M}_\odot c^2$
$3 \cdot 10^{44}$	Energy in photons from a type II supernova explosion
$1 \cdot 10^{45}$	Rough total energy from a cosmological γ-ray burst
$3 \cdot 10^{46}$	Energy in neutrinos from a type II supernova explosion
$2 \cdot 10^{53}$	Typical gravitational binding energy of a galaxy
$2 \cdot 10^{54}$	Typical magnetic and kinetic energy in a large radio lobe (Shu 311)
$5.2 \cdot 10^{54}$	Rotational energy of a 10^8 M$_\odot$ maximal Kerr Black Hole $= 0.29\mathrm{M}_\bullet\ c^2$
$5 \cdot 10^{57}$	Typical gravitational binding energy of a cluster of galaxies

Table 9.8: **Typical energy scales**

And some powers scales:

Magnitude [W]	Description
$4\cdot 10^{-6}$	Average energy output from 1 kilogram of the Milky Way
$1\cdot 10^{-3}$	Optical disc player LASER
6	Amateur short wave radio transmitter
60	Light bulb
100	Gravitational wave power from the Earth-Sun system
150	Human being under normal conditions
750	Maximum long duration horse output
$1,500$	Typical fireplace fire
$20,000$	Car
$1\cdot 10^{5}$	Running Tyrannosaurus Rex
$3\cdot 10^{8}$	Nuclear power reactor
$3\cdot 10^{8}$	Moderate thunderstorm electrical power generation rate
$1.3\cdot 10^{9}$	Hoover dam
$3\cdot 10^{11}$	USA average electricity usage rate in 11986 (holocene calendar)
$1\cdot 10^{13}$	Solar radio luminosity
$8\cdot 10^{13}$	Powerful nanosecond pulse LASER
$1.7\cdot 10^{17}$	Insolation of Earth
$1\cdot 10^{20}$	X-ray luminosity of the quiet Sun
$1\cdot 10^{23}$	Typical white dwarf luminosity
$8\cdot 10^{24}$	Gravitational wave radiation from PSR 1913+16
$3.9\cdot 10^{26}$	Solar luminosity
$1\cdot 10^{28}$	X-ray luminosity associated with Sgr A^*
$3.5\cdot 10^{28}$	Rotational energy loss rate from the Geminga pulsar
$4.3\cdot 10^{28}$	Megamaser in NGC 4258
$2.4\cdot 10^{29}$	5 $M_{\odot}$ star on main sequence
$4\cdot 10^{30}$	Cygnus X-1 X-ray luminosity
$5.5\cdot 10^{30}$	Luminosity of the 20 $M_{\odot}$ red supergiant Betelgeuse at maximum light
$1\cdot 10^{31}$	Crab Nebula energy output
$2\cdot 10^{31}$	Eddington limit for a 1.4 $M_{\odot}$ neutron star
$3\cdot 10^{31}$	Typical luminosity of GRS 1915+105 in outburst (*Nature* 371, 46)
$1\cdot 10^{33}$	Rough luminosity of Eta Carinae in April 11843 (holocene calendar)
$5\cdot 10^{35}$	Type II supernova peak photon luminosity
$3\cdot 10^{36}$	Milky Way
$1.5\cdot 10^{38}$	Coma cluster X-ray gas luminosity
$1\cdot 10^{39}$	10^8 $M_{\odot}$ Black Hole accreting at $\frac{1}{10}$ of the Eddington limit
$1\cdot 10^{39}$	Typical quasar luminosity
$2\cdot 10^{41}$	Rough luminosity of the $z=2.286$ ultraluminous IRAS galaxy IRAS F10214+4724
$6\cdot 10^{43}$	Rough luminosity of a cosmological γ-ray burst
$2\cdot 10^{46}$	Rough luminosity of all the stars in the universe
$1\cdot 10^{48}$	Core collapse neutrino luminosity of a type II supernova (QJRAS 30, 423)

Table 9.9: **Typical powers scales**

And some temperature scales:

Magnitude [K]	Description
$7 \cdot 10^{-7}$	Laser cooling of cesium atoms
$1.3 \cdot 10^{-5}$	Cosmic microwave background quadrupole anisotropy
$3.3 \cdot 10^{-3}$	Cosmic microwave background dipole anisotropy
0.01	Typical limit of liquid helium dilution cooling
0.3	Typical limit of liquid helium evaporation cooling
2.17	Liquid ^{4}He superfluid transition temperature
2.726	Cosmic microwave background temperature today
3.20	Liquid ^{3}He boiling point
4	Typical limit of Joule-Thomson effect cooling
4.18	Liquid ^{4}He boiling point
6	Typical noise temperature of a HEMT receiver at 30 GHz
12	Lanthanum (under pressure) superconductivity critical temperature (highest for a pure element)
20	Liquid H_2 boiling temperature
77	Liquid N_2 boiling temperature
133	Mercury-barium-calcium-copper oxide compound superconductivity critical temperature
273	Water freezing temperature
311	Human surface temperature
373	Water boiling temperature
388	Brimstone melting temperature - upper limit to the temperature of Hell
506	Paper burning temperature
740	Typical surface temperature of Venus
1811	Melting temperature of iron
3000	Cosmic microwave background temperature at decoupling
5770	Solar effective temperature
$5 \cdot 10^5$	Surface temperature of the Geminga pulsar
$3 \cdot 10^6$	Polar cap temperature of the Geminga pulsar
$3 \cdot 10^6$	Typical fusion experiment
$1.4 \cdot 10^7$	Center of the Sun
$1.5 \cdot 10^7$	Changeover temperature from the proton-proton chain to the CNO cycle
$2.7 \cdot 10^7$	Center of a 5 $M_\odot$ star
$5 \cdot 10^7$	Typical gas temperature in a cluster of galaxies
$4 \cdot 10^8$	Characteristic temperature for electron-positron pair production
$4 \cdot 10^8$	Minimum primordial nucleosynthesis temperature
$5 \cdot 10^8$	Inner accretion disc temperature of Cyg X-1
$7 \cdot 10^8$	Thermal electrons become relativistic $(v_e = \frac{c}{2})$
$1 \cdot 10^9$	Maximum primordial nucleosynthesis temperature
$1 \cdot 10^9$	Rough superconductivity critical temperature in a neutron star
$1 \cdot 10^{10}$	Rough plasma pair catastrophe temperature
$3 \cdot 10^{10}$	Core collapse temperature of a supernova (QJRAS 30, 424)
$3 \cdot 10^{15}$	Rough electroweak unification temperature

Table 9.10: **Typical temperature scales**

And some magnetic intensities scales:

Magnitude [T]	Description
$1\cdot10^{-15}$	Smallest measured magnetic field (Schumann resonances)
$1\cdot10^{-13}$	Spontaneous human brain activity
$1\cdot10^{-12}$	Evoked human brain activity
$1\cdot10^{-12}$	Typical magnetic field needed for good radio reception
$1\cdot10^{-11}$	Typical magnetic field from a human heart
$1\cdot10^{-10}$	Typical 50/60 Hz magnetic field inside a building
$5\cdot10^{-10}$	Typical magnetic field strength in the local interstellar medium
$1\cdot10^{-9}$	Typical magnetic field strength in a radio lobe
$1\cdot10^{-8}$	Typical magnetic field strength in the central 300 parsecs of the Milky Way
$5\cdot10^{-8}$	Magnetic field in the Crab Nebula
$1\cdot10^{-7}$	Magnetic field in The Arc at the Milky Way center
$1\cdot10^{-6}$	Typical magnetic field from a hand held cordless phone
$3\cdot10^{-5}$	Magnetic field at Earth's surface
$1\cdot10^{-4}$	Magnetic field near Sun's pole
$4\cdot10^{-4}$	Magnetic field at Jupiter's cloud tops
$1\cdot10^{-4}$	Magnetic field near mobile phone
0.1	Ap star magnetic field
0.2	Sunspot magnetic field
1	Typical medical NMR magnetic field or solar spots
2	Magnetic field felt by the electron in an $n=1$ hydrogen atom
2	RS CVn star spot magnetic field
12	Typical magnetic field used in high resolution NMR spectroscopy
25	Powerful superconducting/normal hybrid magnet
45	World record magnetic field in laboratory
$1\cdot10^{4}$	Field of white dwarf
$3\cdot10^{4}$	Fields in petawatt LASER pulses
$1\cdot10^{8}$	Typical single pulsar dipole magnetic field strength
$4.4\cdot10^{9}$	Magnetic field strength when the energy of the first electron Landau level is comparable to the electron rest mass (MNRAS 275, 257)
$1\cdot10^{11}$	Highest field ever measured, on magnetar and soft gamma repeater SGR-1806-20
$1\cdot10^{12}$	Estimated magnetic field near atomic nucleus

Table 9.11: **Typical magnetic scales**

Without measurements, there are also no emotions! Indeed, our emotions are triggered by our senses. And all the impressions and all the information that our senses provide us are - among others - measurements!

Because of all these concepts and definitions and the complexity of data analysis, engineers and scientists define experts committees in a given field that publish international coherent standards that almost all specialists follow to allow science to progress faster and therefore improve people's conditions and save every year the life of million of people. An example of such standards is given in the table below:

Basic standards for Physics ISO/IEC 80000 Quantities and Units	ISO basics standards for measurement analysis in science
ISO 80000-1:2022 General	ISO 3534:2006 Statistics Vocabulary and symbols
ISO 80000-2:2019 Mathematical signs & symbols	ISO 2602:1980 Statistical interpretation of test results Estimation of the mean
ISO 80000-3:2019 Space and time	ISO 3301:1975 Statistical interpretation of data Comparison of two means in the case of paired data
ISO 80000-4:2019 Mechanics	ISO 16269-6:2014 Statistical interpretation of data Part 6: Determination of statistical tolerance intervals
ISO 80000-5:2022 Thermodynamics	ISO 16269-8:2004 Statistical interpretation of data Part 8: Determination of prediction intervals
ISO 80000-6:2022 Electromagnetism	ISO 16269-7:2001 Statistical interpretation of data Part 7: Median and confidence intervals
ISO 80000-7:2019 Light and radiation	ISO 5779:1997 Statistical interpretation of data Tests for departure from the normal distribution
ISO 80000-8:2020 Acoustics	ISO 3494:1976 Statistical interpretation of data Power of tests relating to means and variances
ISO 80000-9:2019 Physical chemistry and Molecular Physics	ISO 16269-4:2010 Statistical interpretation of data Detection and treatment of outliers
ISO 80000-10:2019 Atomic and Nuclear Physics	ISO 2854:1976 Statistical interpretation of data Techniques of estimation and tests relating to means and variances
ISO 80000-11:2019 Characteristic (dimensionless) numbers	ISO 3534-3:2013 Vocabulary and symbols Part 3: Design of experiments
ISO 80000-12:2019 Condensed matter physics	ISO/IEC 17025:2017 General requirements for the competence of testing and calibration laboratories
...	ISO 5725 Accuracy (trueness and precision) of measurement methods and results

Table 9.12: Some famous international standards that make authority in the field of experimental science

9.1.2 Distributions

In physics and mechanics, mass distribution is the spatial distribution of mass within a solid body. In principle, it is relevant also for gases or liquids, but on Earth their mass distribution is almost homogeneous.

Definitions (#9.1.418):

D1. A mass or an electric charge are named "punctual" if they occupy a volume whose dimensions are much less than the distances of observations.

Remark

The elementary electric charge is a good approximation of a punctual charge given its small size which the classical radius of the order of the femtometer, which is of course very small compared to conventional observation dimensions (at least until the time of writing...).

D2. We consider N bodies of finite mass or electric charge magnitude in a finite volume V. If this volume is assumed large enough so that the average distance between the bodies is much greater than the size of these, we are dealing with a "discontinuous distribution" or "discrete distributions" of these bodies (we sometimes talk about non-uniform distribution).

The calculations are most of time very difficult to do by starting with a discrete distribution because, in general, the number of bodies to be considered is very high when the volume has macroscopic dimensions. In this case, many theoretical models (but not all), try to introduce another type of distribution using some approximation tricks.

D3. We consider N bodies of finite mass or electric charge magnitude in a finite volume V. If the distribution of elements such that there is no "holes" between all of them (in other words, each element is clamped against a other) then we are dealing with a "continuous distribution". A continuous distribution can then be described by a function that shows how the elements are distributed in a volume, surface or line.

Remarks

R1. We can specify sometimes, as we have already mentioned it during the definitions of mass or electric charge, that the distributions can be of volume, surface or linear type. If this is not stated is that the information is implicitly trivial in the context of the mathematical developments.

R2. The term "continuous" in "continuous distribution" come from the fact that many times we need to integrate functions, hence the need of a continuous distribution.

9.1.3 Constants

As we already know, physics, unlike mathematics, is an "exact science" in the sense that its verification and validity are based on and are constantly tested for refutation by experimental facts.

As the human being had to arbitrarily choose a system of measures, some laws theoretically derived thanks to the properties of matter are often accurate only to a given multiplicative constant factor of normalization with respect to the measurement system. Then appear in the equations of physics constants whose existence is only due to this measurement system (but however this is not always the case), some constants although adapting the measurement system will equal (at least it seems) never the unit.

There are many constants in physics (very likely an infinity) but some have a special status in that they can not be deduce from other constants.

A constant can be seen also as out level of ignorance. Every constant that cannot be explained theoretically is the sign that perhaps a theory must be in-deep (until a given level). Actually the Standard Model and General Relativity need 20 fundamentals constants[4].

We propose below a list and values (not exact values) that we will frequently met again in our future mathematical developments in this textbook.

Remarks

R1. The constants below are given for some for the moment at the reader read them (...) because according to some theories and models, values are for some not fixed at all (time dependent) or just their accuracy increase in function of our technologies.

R2. The series of international norms ISO 80000:2009 *Quantities and Units* define the values of numerous scientific constants.

---

[4] The universal constant G, the Yukawa coefficients determining the masses of the six quark (u, d, c, s, t, b) and three lepton (e, μ, τ) flavours, the Higgs mass and vacuum expectation value, three angles and a phase of the Cabibbo-Kobayashi-Maskawa matrix, a phase for the QCD vacuum and three coupling constants g_s, g_W, g_1 for the gauge group SU$(3)\times$SU$(2)\times$SU(1) respectively.

9.1.3.1 Mathematical Constants

We give here only common mathematical constants which origin (definition) was proven and usage is very common in this textbook like in the sections of Functional Analysis, Euclidean Geometry, Differential and Integral Calculus. The list will be expanded in function of the evolution of the book itself!

Name	Symbol	Value	Unit
Archimede's constant (pi)	π	$\sim 3.14159265358979323846$	-
Euler's number	e	$\sim 2.71828182845904523536$	-
Golden ratio	φ	$(1+\sqrt{5})/2 \cong 1.61803398879894$	-
Euler-Mascheroni constant	γ	~ 0.5772156629	-
Apéry's constant	$\zeta(3)$	~ 1.202	-

Even if these are mathematical constants they appear a lot in physics in so many fields that the best for the reader is just to read the whole remaining part of this textbook to see where...

A mathematical constant is therefore a special number, usually a real number, that is "significantly interesting in some way". What it means for a constant to arise "naturally", and what makes a constant "interesting", is ultimately a matter of taste, and some mathematical constants are notable more for historical reasons than for their intrinsic mathematical interest. The more popular constants have been studied throughout the ages and computed to many decimal places.

Remark

The reader interested in the properties of chemical elements can download the periodic table of elements proposed in the download section of the companion website.

9.1.3.2 Universal Constants (fundamental constants)

A fundamental physical constant is a physical quantity that is generally assumed (for simplification purposes at undergraduate level) to be both universal in nature and constant in time. It can be contrasted with a mathematical constant, which is a fixed numerical value, but does not directly involve any physical measurement.

Before discussing some constants of Nature of our observable Universe, we must have an idea of which constants to consider. First, all constants of physics do not play the same role, and some have a much deeper one than others. Following Levy-Leblond, we can define three classes of fundamental constants, class A being the class of the constants characteristic of particular objects, class B being the class of constants characteristic of a class of physical phenomena, and class C being the class of universal constants. Indeed, the status of a constant can change with time. For instance, the velocity of light was a initially a type A constant (describing a property of light) which then became a type B constant when it was realized that it was related to electromagnetic phenomena and, to finish, it ended as a type C constant (it enters many laws of physics from electromagnetism to relativity including the notion of causality...). It has even become a much more fundamental constant since it has been chosen as the definition of the meter. A more conservative definition of a fundamental constant would thus be to state that it is any parameter that can not be calculated with our present knowledge of physics, i.e. a free parameter of our theory at hand. Each free parameter of a theory is in fact a challenge for future theories to explain its value as we have already mentioned it.

The values listed below are values which scientists noticed that they seemed (...) constant and independent of all parameters used at least.... at normal energy levels (indeed some of the constants below like the fine-structure constant are observed as being variable at high energy levels like in the Large Hadron Collider collisions at CERN!):

Name	**Symbol**	**Value**	**Unit**
Elementary charge	e	$1.602176634\cdot 10^{-19}$	C
Gravitational constant	G,κ	$6.67430(15)\cdot 10^{-11}$	$\text{m}^3\cdot\text{kg}^{-1}\cdot\text{s}^{-2}$
Fine-structure constant	$\alpha=e^2/2hc\varepsilon_0$	$\approx 1/137=7.7923525643(11)\cdot 10^{-3}$	
Speed of light in vacuum	$c=(\varepsilon_0\mu_0)^{-1}$	$2.99792458\cdot 10^{8}$	$\text{m}\cdot\text{s}^{-1}$ (def)
Permittivity of the vacuum	$\varepsilon_0=1/\mu_0c^2$	$8.8541878188\cdot 10^{-12}$	$\text{F}\cdot\text{m}^{-1}$
Permeability of the vacuum	$\mu_0=4\pi\alpha\hbar/e^2c$	$4\pi\cdot 10^{-7}$	$\text{H}\cdot\text{m}^{-1}$
Impedance of the vacuum	Z_0	376	$\text{C}\cdot\text{V}\cdot\text{s}^{-1}$
Coulomb's constant	$k_e=(4\pi\varepsilon_0)^{-1}$	$8.9876\cdot 10^{9}$	$\text{N}\cdot\text{m}^2\cdot\text{C}^{-2}$
Planck's constant	h	$6.62607015\cdot 10^{-34}$	$\text{J}\cdot\text{s}$
Dirac's constant	$\hbar=h/2\pi$	$1.054571817...\cdot 10^{-34}$	$\text{J}\cdot\text{s}$
Bohr magneton	$\mu_{\text{B}}=e\hbar/2m_{\text{e}}$	$9.2740100657(29)\cdot 10^{-24}$	$\text{A}\cdot\text{m}^2$
Electron magnetic moment	μ_e	$-9.2847646917(29)\times 10^{-24}$	$\text{J}\cdot\text{T}^{-1}$
Electron g-factor	$g=-2\lvert\mu_e\rvert/\mu_B$	$-2.00231930436092(36)$	
Bohr radius	$a_0=4\pi\epsilon_0\hbar^2/m_{\text{e}}e^2$	$5.29177210544(82)$	Å
Rydberg's constant	R_y	$10973731.568157(12)$	m^{-1}
Electron Compton wavelength	$\lambda_{\text{Ce}}=h/m_ec$	$2.242631023538(76)\cdot 10^{-12}$	m
Proton Compton wavelength	$\lambda_{\text{Cp}}=h/m_{\text{p}}c$	$1.3214\cdot 10^{-15}$	m
Reduced mass of the H-atom	μ_{H}	$9.1045755\cdot 10^{-31}$	kg
Classical electron radius	$r_e=\alpha^2a_0$	$2.8179403205(13)\cdot 10^{-15}$	m
Thomson cross section	$\sigma_e=(8\pi/3)r_{\text{e}}^2$	$6.6524587051(62)\cdot 10^{-29}$	m^2
Electron-Volt	1[eV]	$1.602176634\cdot 10^{-19}$	J
Stefan constant	$\sigma=\pi^2k^4/60\hbar^3c^2$	$5.670374419...\cdot 10^{-8}$	$\text{W}\cdot\text{m}^{-2}\text{K}^{-4}$
Stefan-Boltzmann's constant	$\sigma_{\text{S.B.}}$	$7.56\cdot 10^{-16}$	$\text{W}\cdot\text{m}^{-2}\cdot\text{K}^{-4}$
Wien's constant	k_{W}	$2.897771995\ldots\cdot 10^{-3}$	mK
Absolute Temperature	T_0	273.15	K
Molar gas constant	R	$8.314462618\ldots$	$\text{J}\cdot\ \text{mol}^{-1}\cdot\text{K}^{-1}$
Faraday's constant	F	$96485.33212\ldots$	$\text{C}\cdot\text{mol}^{-1}$
Avogadro's constant	N_{A}	$6.02214076\cdot 10^{23}$	mol^{-1}
Boltzmann's constant	$k=R/N_{\text{A}}$	$1.380649\cdot 10^{-23}$	$\text{J}\cdot\text{K}^{-1}$
Electron mass	m_{e}	$9.1093837139(28)\cdot 10^{-31}$	kg
Proton mass	m_{p}	$1.67262192595(52)\cdot 10^{-27}$	kg
Neutron mass	m_{n}	$1.67492750056(85)\times 10^{-27}$	kg
Atomic mass constant	u	$1.66053906892(52)\cdot 10^{-27}$	kg
Muon mass	m_μ	$1.883531627(42)\cdot 10^{-28}$	kg
Elementary mass unit	$m_{\text{u}}=\frac{1}{12}m(^{12}_{6}\text{C})$	$1.6605656\cdot 10^{-27}$	kg
Nuclear magneton	μ_{N}	$5.0508\cdot 10^{-27}$	$\text{J}\cdot\text{T}^{-1}$
Muon magnetic moment	μ_μ	$-4.49044830(10)\times 10^{-26}$	$\text{J}\cdot\text{T}^{-1}$
Muon magnetic moment anomaly	a_μ	$1.16592062(41)\cdot 10^{-3}$	
Muon g-factor	$g_\mu=-2(1+a_\mu)$	$-2.00233184123(82)$	
...	...	...	...

Table 9.13: **CODATA recommended values of the fundamental physical constants (12022)**

Most of the constants above can be obtained by theoretical models (see next chapters of this textbook for the proofs). As far as we know today (early 121st century according to holocene calendar), it seems that only e, μ_0, ε_0, h, m_e, k cannot be obtained from other constants (yes! there are theoretical models to get approximate values for c and G by the study of the quantum vacuum properties!).

Notice that you should never see a value of the speed of light with an uncertainty, because it's based on a definition (that we already provided earlier) because in reality we cannot easily measure it. Indeed that's because of effect on relativity on clocks that should be synchronised to measure the distance travelled during a given time between two points... and the calculations depends on the speed of light itself... So a solution may be to measure a two way round trip of light instead that a direct trip as in that the two way round trip case you may be able to use the same clock! But then another problem appears: speed of light might not be symmetrical on a two way round trip... how can we be sure that it is really the case? So actually it's a convention (sometimes named the "two way speed of light 11905[5] Einstein's convention") and not a fact that light on a two way trip travel at the same speed for go and come back (Albert Einstein wrote that about that difficulty in one of it's paper[6]).

In fact, the number of fundamental physical constants depends on the physical theory accepted as "fundamental". Currently, this is the theory of General Relativity for gravitation and the Standard Model for electromagnetic, weak and strong nuclear interactions and the matter fields. Between them, these theories seems to account for a total of 26 independent fundamental constants already mentioned earlier above.

Remark

The best papers that we have read so far about the details on the quite huge possible variation ranges of the Universal constants - and that debunk the fine-tuning fallacy - are *The degree of fine-tuning in our universe - and others* (see [6]) by Fred C. Adams and *The fundamental constants and their variation: observational and theoretical status* (see [552]) by Jean-Philippe Uzan and also *The Fallacy of Fine-Tuning: Why the Universe Is Not Designed for Us* (see [552]) by V. J. Stenger. We therefore recommend any reader interested in exploring the subject in greater depth (and understand the fine-tuning fallacies and lies of people have a religious agenda) to read them!

In the table below we can see ranges of parameters for viable universes according to Fred C. Adams in *The degree of fine-tuning in our universe - and others* (see [6]). The results are expressed in decades, and calculated by taking $\log_{10}(X_{\max}/X_{\min})$, where $X_{\max}$ is the maximum value a parameter could take while still being compatible with a habitable universe. $X_{\min}$ is the minimum such value. For ranges marked with an asterisk* (the weak coupling constant and the vacuum energy scale) the quoted range only applies to values above that measured in our universe; any values below that measured in our universe are allowed, without limitation:

---

[5] According to holocene calendar.

[6] In it's paper Einstein define speed of light as $c = 2AB/(t'_B - t_A)$ and still in the early 121st century (holocene calendar) there are researches who try to find a solution to measure the one-way speed of light but that are systemically debunked.

Quantity	**Observed Value**	**Range (decades)**	$X_{\max}/X_{\min}$
Up quark mass	2.3 [MeV]	>3	$>10^3$
Down quark mass	4.8 [MeV]	0.85	7.08
Electron-proton mass ratio	$1/1836$	5	10^5
Up-down quark mass difference	2.5 [MeV]	1	>10
Gravitational constant	$6\cdot10^{-39}$	>10	$>10^{10}$
Weak coupling constant	10^{-5}	6*	10^6
Fine structure constant	$1/137$	4	10^4
Strong coupling constant	15	3	10^3
Fluctuation amplitude	10^{-5}	4	10^4
Baryon to photon ratio	$6\cdot10^{-10}$	6	10^6
Dark matter abundance	$3\cdot10^{-9}$	6	10^6
Vacuum energy scale	0.0003 [eV]	10*	10^{10}

In a survey made on 79 professional physicists (see [96]) in 12024 (holocene calendar) we get the following result for the opinion of professional scientist about the values of the constants:

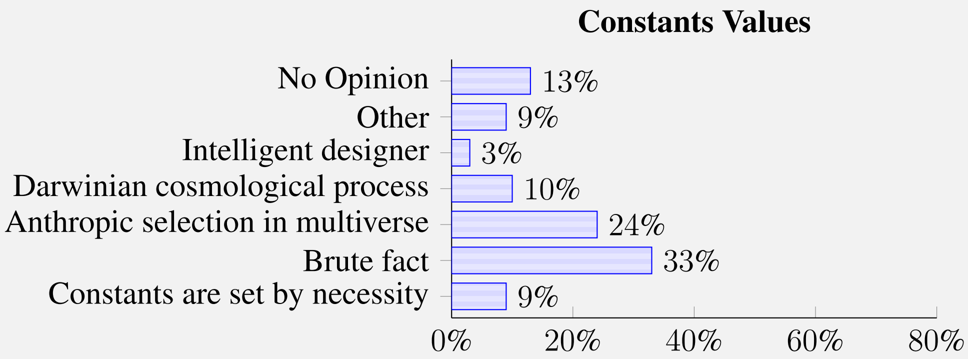


Figure 9.5: Copenhagen survey on how the "Big Bang" should be understood

The developments of high energy physics theories such as multi-dimensional (Kaluza-Klein) and string theories provide new motivations to consider the time variation of the fundamental constants. The observation of the variability of these constants constitutes one of the very few hope to test directly the existence of extra-dimensions and to test these high energy-physics models.

> *The Universe is not fine-tuned to us: We are fine-tuned to our particular Universe!*
>
> Victor. J. Stenger

quote 9.1:3117.#61

Caution! The readers must be careful about the fallacy of "fine tuning argument" from people having very likely a precise agenda and using only rhetorical arguments to manipulate their audience (we have already speak about the danger of rhetoric in this textbook introduction). As we will prove it during all coming chapters and sections, most constants can vary (more or less in a given range) from their actual value without forbidding universe, stars, planets and life to emerge (and even if the constants can vary of only $\pm 0.1\%$ as we have proved it in the section of Set Theory - Cantor theorem at page 232 - there are as many infinite numbers in a small range of $\mathbb{R}$ than in the whole set of $\mathbb{R}$, i.e. the infinite choices of values are of same cardinal... therefore speaking of "fine tuning" for an infinite possibility of choices is quite a non-sense).

So claiming that only a very small subset of constant values permits the existence of intelligent life is mathematically a non-sense as all sets have the same cardinal! Furthermore the proponents of fine-tuning make serious undergraduate level errors in physics, cosmology, probability theory, data analysis and even worst in the basics understanding of the scientific method (keep in mind by the way that $99.9999999999999999999999999999999999999999999999999999999999999999999\%$ at least of the observable universe cannot sustain life and only 15% of Earth can naturally support humans and that almost $6,000$ genetic disease are known so far... and that the human body has almost weekly cells that turn out into small cancers...). While not every proponent makes every error, there is a remarkable similarity to the arguments you find in the theist literature, which leads you to suspect that they do not do much research to support their claims beyond simply reading each other's books. With so many errors and misjudgements, and with such a gross lack of understanding of the basic science we have seen exhibited by the supporters of supernatural fine tuning, we can safely say that their motivation is more wishful thinking than truthful scientific inference.

A proper analysis finds there is no evidence that the Universe is fine-tuned. In their 12006 (holocene calendar) paper Roni Harnik, Graham Kribs, and Gilad Perez (see [231]) put forward a universe that seems capable of creating preconditions for life and yet has fundamental particles entirely different from our own. In 12013 (holocene calendar), Abraham Loeb, of Harvard, argued (see [356]) that a primitive form of life might have been possible in the early universe. Fred Adams and Evan Grohs showed that if we vary several parameters in our theories at once, there are ways for stars to produce carbon other than the mechanism Hoyle predicted it's just that those other options conflict with observations that Hoyle already knew of. The reader can also read [511] if interested in a undergraduate treatment of this fallacy.

These three examples show that a chemistry complex enough to support life can arise under circumstances that are not anything like the ones we experience, and therefore our universe isn't all that special!

9.1.3.3 Astronomical/Astrophysical parameters and constants

Obviously an astronomical or astrophysics parameter is a physical "constant" used in astronomy theoretical models that in fact vary with time and can be obtained only by measurement.

There are billions of astronomical/astrophysics parameters but here are the really basic ones that every civilized person should have read at least once in their life:

Name	**Symbol**	**Value**	**Unit**
Diameter of the Sun	$D_\odot$	$1392 \cdot 10^6$	m
Earth's average gravitation	g	9.81	m·s^{-1}
Mass of the Sun	$M_\odot$	$1.989 \cdot 10^{30}$	kg
Rotational period of the Sun	$T_\odot$	25.38	days
Radius of Earth	R_A	$6.378 \cdot 10^6$	m
Mass of Earth	M_A	$5.976 \cdot 10^{24}$	kg
Rotational period of Earth	T_A	23.96	hours
Earth orbital period	Tropical year	365.24219879	days
Astronomical unit	AU	$1.4959787066 \cdot 10^{11}$	m
Light year	lj	$9.4605 \cdot 10^{15}$	m
Parsec (Sun-Earth distance)	pc	$3.0857 \cdot 10^{16}$	m
Hubble constant (non-SI)	H_0	73.52 ± 1.62	km·s^{-1}·Mpc^{-1}
Hubble constant (SI units)	H_0	$2.4305(\pm 0.0525) \cdot 10^{-18}$	s^{-1}
Universe critical density	ρ_c	$5 \cdot 10^{-27}$	kg·m^{-3}

Remark

The astronomical unit (AU) was initially the Earth-Sun distance but as the latter varies obviously with time it has become a standardized value at $149,597,870,700$ [m] by the XXVIII General Assembly of International Astronomical Union. However notice that in modern times, Earth-Sun distance is measured using telemetry (radar echo) with an accuracy of a dozen of meters only (see [286])!

9.1.3.4 Chemical parameters

The physicochemical parameters are physical constants that are especially found in all areas relating to chemistry and thermodynamics.

There are also billions of physicochemical parameters, but here are the really basic ones that every civilized person should have read at least once in their life:

Name	**Symbol**	**Value**	**Unit**
Standard Pressure	P_0	$101,325$	Pa
Standard Temperature	T	293.15	K
Molar Volume	V_M	22.4141	l mol^3

Remark

The reader interested in the properties of chemical elements can download the periodic table of elements proposed in the download section of the companion website.

9.1.3.5 Material parameters

The material parameters are especially found in all areas relating to mechanics, electrodynamics, thermodynamics, architecture, civil engineering, etc.

Name	Symbol	Value	Unit
Refractive index of vacuum	n	1	
Refractive Vacuum/Air ($\lambda = 589.29$ [nm])	n	1.000293	
Refractive Vacuum/Water ($\lambda = 589.29$ [nm], 293 [K])	n	1.330	
Refractive Vacuum/Diamond ($\lambda = 589.29$ [nm])	n	2.419	
Refractive Vacuum/Ethanol (ethyl alcohol) ($\lambda = 589.29$ [nm])	n	1.361	
Resistivity copper (293 [K])	ρ	0.017	$\Omega\cdot\text{mm}^2\cdot\text{m}^{-1}$
Resistivity aluminium (293 [K])	ρ	0.0278	$\Omega\cdot\text{mm}^2\cdot\text{m}^{-1}$
Resistivity sast steel (293 [K])	ρ	0.13	$\Omega\cdot\text{mm}^2\cdot\text{m}^{-1}$
Resistivity silver (293 [K])	ρ	0.016	$\Omega\cdot\text{mm}^2\cdot\text{m}^{-1}$
Resistivity carbon (293 [K])	ρ	40	$\Omega\cdot\text{mm}^2\cdot\text{m}^{-1}$
Resistivity pure steel (293 [K])	ρ	0.1	$\Omega\cdot\text{mm}^2\cdot\text{m}^{-1}$
Resistivity quicksilver (293 [K])	ρ	0.941	$\Omega\cdot\text{mm}^2\cdot\text{m}^{-1}$
Resistivity gold (293 [K])	ρ	0.0222	$\Omega\cdot\text{mm}^2\cdot\text{m}^{-1}$
Resistivity lead (293 [K])	ρ	0.208	$\Omega\cdot\text{mm}^2\cdot\text{m}^{-1}$
Resistivity bakelite (293 [K])	ρ	10^{18}	$\Omega\cdot\text{cm}$
Resistivity pure water (293 [K])	ρ	10^{7}	$\Omega\cdot\text{cm}$
Resistivity marble (293 [K])	ρ	10^{10}	$\Omega\cdot\text{cm}$
Resistivity glas (293 [K])	ρ	10^{15}	$\Omega\cdot\text{cm}$

9.1.3.6 Planck's constants

The Planck's constants are mainly physical curiosities that arise from a specific system of units and whose values in the S.I. system are given in the table below (for their derivation see page 4905):

Name	Symbol	Value	Unit
Planck length	l_P	$1.616199(97)\cdot 10^{-35}$	m
Planck area	l_P^2	$2.6121\cdot 10^{-70}$	m^2
Planck volume	l_P^3	$4.2217\cdot 10^{-105}$	m^3
Planck mass	m_P	$2.17651(13)\cdot 10^{-8}$	kg
Planck time	t_P	$5.39106(32)\cdot 10^{-44}$	s
Planck electric charge	q_P	$1.875545956(41)\cdot 10^{-18}$	C
Planck temperature	T_P	$1.416833(85)\cdot 10^{32}$	K
Planck linear momentum	$p_P = m_p \cdot c$	6.52485	kg m s^{-1}
Planck energy	$E_P = m_P c^2$	$1.9561\cdot 10^{9}$	J
Planck force	$f_P = E_P \cdot l_P^{-1}$	$1.21027\cdot 10^{44}$	N
Planck power	$P_P = E_P \cdot t_P^{-1}$	$3.62831\cdot 10^{52}$	W
Planck density	$\rho_P = m_P \cdot l_P^{-3}$	$5.15500\cdot 10^{96}$	kg m^{-3}
Planck energy density	$u_P = E_P \cdot l_P^{-3}$	$4.63298\cdot 10^{113}$	J m^{-3}
Planck energy intensity	$I_P = u_P \cdot c$	$1.38893\cdot 10^{122}$	W m^{-2}
Planck angular frequency	$\omega_P = t_P^{-1}$	$1.85487\cdot 10^{43}$	s^{-1}
Planck pressure	$p_P = f_p \cdot l_P^2$	$4.63309\cdot 10^{113}$	N m^{-2}
Planck electric current	I_P	$3.4789\cdot 10^{25}$	A
Planck voltage	V_P	$1.04295\cdot 10^{27}$	V
Planck impedance	$Z_P = V_P \cdot I_P^{-1}$	29.9792458	Ohm
Planck acceleration	$a_P = c \cdot t_P^{-1}$	$5.560815\cdot 10^{51}$	m s^{-2}

We see above that some Planck units are suitable for measuring quantities that are familiar from daily experience. For example:

- 1 Planck mass is only about 22 micrograms
- 1 Planck momentum is only about 6.5 $[\text{kg}\cdot\text{m}\cdot\text{s}^{-1}]$
- 1 Planck energy[7] is only about 500 $[\text{kW}\cdot\text{h}]$
- 1 Planck charge is close to only 11.7 elementary charges
- 1 Planck impedance is close to only 30 $[\Omega]$

Planck constants are based on the fact that as we will see further below that their definition comes only from properties of Nature and not from any human construct. Planck units are only one system of several systems of natural units, but Planck units are not based on properties of any prototype object or particle (that would be arbitrarily chosen), but rather on only the properties of free space.

---

[7] Much smaller than the energy produced by the Tsar Bomba $\sim 2.1\cdot 10^{17}$ [J].

In this system of units the fundamentals constants G, c, $\hbar$, $(4\pi\varepsilon_0)^{-1}$ and k (Boltzmann constant) are such that:

$$\boxed{G=c=\hbar=\frac{1}{4\pi\varepsilon_0}=k=1} \tag{9.1.49}$$

Some physicists argue that communication with extraterrestrial intelligence would have to employ such a system of units in order to be understood as it is once again the Natural choice!!!

The Planck constant are mainly physical curiosities and the reader must not see any special interpretation about them (at least as far as we know today).

Remark

For reminder, the reader interested in the exact provenance of the various Planck constants (Planck length, Planck mass, etc.) should go read the section of Wave Quantum Physics where these constants are determined with the necessary details at page 4905.

The table below shows how the use of Planck units simplifies many fundamental equations of physics that we will prove in details in the next chapters, because this gives each of the five fundamental constants, and products of them, a simple numeric value of 1. In the SI form, the units should be accounted for. In the non-dimensionalized form, the units, which are now Planck units, need not to be written if their use is understood.

Model	SI form	Non-dimensionalized form
Newton's gravitational law	$F=-G\frac{m_1m_2}{r^2}$	$F=-\frac{m_1m_2}{r^2}$
Coulomb's law	$F=\frac{1}{4\pi\varepsilon_0}\frac{q_1q_2}{r^2}$	$F=\frac{q_1q_2}{r^2}$
Einstein field equations	$G_{\mu\nu}=8\pi\frac{G}{c^4}T_{\mu\nu}$	$G_{\mu\nu}=8\pi T_{\mu\nu}$
Mass–energy equivalence	$E=mc^2$	$E=m$
Energy–momentum relation	$E^2=m^2c^4+p^2c^2$	$E^2=m^2+p^2$
Thermal energy	$E=\frac{1}{2}kT$	$E=\frac{1}{2}T$
Planck-Einstein relation	$E=h\nu=\hbar\omega$	$E=\omega$
Planck's law	$I(\omega,T)=\frac{\hbar\omega^3}{4\pi^3c^2}\frac{1}{e^{\frac{\hbar\omega}{k_BT}}-1}$	$I(\omega,T)=\frac{\omega^3}{4\pi^3}\frac{1}{e^{\omega/T}-1}$
Stefan–Boltzmann constant	$\sigma=\frac{\pi^2k_B^4}{60\hbar^3c^2}$	$\sigma=\frac{\pi^2}{60}$
Schrödinger's equation	$-\frac{\hbar^2}{2m}\nabla^2\left(1+V(\vec{r})\right)\Psi(\vec{r},t)=\mathrm{i}\hbar\dot{\Psi}(\vec{r},t)$	$-\frac{1}{2m}\nabla^2\left(1+V(\vec{r})\right)\Psi(\vec{r},t)=\mathrm{i}\dot{\Psi}(\vec{r},t)$
Dirac covariant equation	$(\mathrm{i}\hbar\gamma^\mu\partial_\mu-mc)\psi=0$	$(\mathrm{i}\gamma^\mu\partial_\mu-m)\psi=0$
Maxwell Equations	$\vec{\nabla}\circ\vec{E}=\dfrac{1}{\varepsilon_0}\rho$ $\vec{\nabla}\circ\vec{B}=0$ $\vec{\nabla}\times\vec{E}=\dfrac{\partial\vec{B}}{\partial t}$ $\vec{\nabla}\times\vec{B}=$ $\dfrac{1}{c^2}\left(\dfrac{1}{\varepsilon_0}\vec{j}+\dfrac{\partial\vec{E}}{\partial t}\right)$	$\vec{\nabla}\circ\vec{E}=4\pi\rho$ $\vec{\nabla}\circ\vec{B}=0$ $\vec{\nabla}\times\vec{E}=\dfrac{\partial\vec{B}}{\partial t}$ $\vec{\nabla}\times\vec{B}=$ $4\pi\vec{j}+\dfrac{\partial\vec{E}}{\partial t}$

Table 9.14: **Simplification of physics equations in Planck units**

Despite the examples given how many constant are there? Why play they a "central role" in physical theories? Do they all have the same importance or are some of them more fundamental than the others? Following what criteria? Can we then test whether they are really constant?

To try to answer some of these questions, first of all we must notice that at each step of our theoretical constructions there remain some constant parameters that are not and can not be explained in terms of more fundamental quantities, simply because these latter do not exist in the state of our knowledge. When the theories are refined, it is possible that a constant is explained in terms of new parameters, more fundamental. Thus, the mass of the proton is a fundamental constant of nuclear physics, but must in principle be able to calculate, within the framework of quantum chromodynamics, according to the quark masses and energies of the electromagnetic and strong interactions. This change of status is associated with that of the proton, which is of elementary particle become a composite body.

We will define constants modestly as all non-determined parameters in a given theoretical framework. This definition is equivalent to accept that we are unable to write an evolution equation for these constants and that thus revealed the limit of what the theories in which they appear are able to explain. However the constancy hypothesis is implicitly controlled by experimental validation of these theories. The results of experiments must be reproducible at various times and in various laboratories. If this is the case, within the limits of experimental details, then it is legitimate to consider that the constancy hypothesis is valid. This definition implies that there is no absolute constants list, because such a list depends on the theoretical frameworks chosen to describe the Nature and may change with the progress of knowledge.

Now arises the question whether it is possible to characterize more precisely the concept of constant and determine whether, among all the constants, some are more fundamental than others. For this, we must first prove a relation between constants and units.

Thus, Planck discovered in 11900 (holocene calendar) that it was possible to use the three fundamental physical constants:

$$\boxed{\begin{aligned} G &\cong 6.67408(31)\cdot 10^{-11} && [\mathrm{m}^3\cdot\mathrm{s}^2\cdot\mathrm{kg}^{-1}]\\ c &\cong 299,729,458 && [\mathrm{m}\cdot\mathrm{s}^{-1}]\\ \hbar &= \frac{h}{2\pi}\cong 1.05\cdot 10^{-34} && [\mathrm{J}\cdot\mathrm{s}] \end{aligned}} \tag{9.1.50}$$

to define the three units of mass, time and length from the Planck mass, the Planck length and the Planck time (see again the chapter of Atomistic for the mathematical approach for determining their values).

Planck named those units "Natural System Units" (NSU) because they are independent of a body or a material and [...] necessarily retain their significance for all times and all civilizations, even for aliens or non-human one. The significance of these units take long to emerge in the theoretical physicists community. They point the scales where quantum mechanics and gravitation become of the same amplitude. They are therefore very suited to the study of primordial cosmology and to that of Black Holes and relativistic quantum mechanics.

The choice of Planck units as natural units is linked to the considerations justifying that G, c, $\hbar$ are the three most fundamental constants (known at this date). In these units, the numerical value of these three fundamental constants is 1 as we have already noticed earlier above.

The role of the constants in the structure of physic theories can be quite beautifully illustrated by the magic cube named the "Okun cube" of physical theories visible in the figure below (whose validity remains to be check of course). The idea is to "turn on" or "turn off" one by one the three fundamental constants to see how physical theories are articulated with each other.

Remark

The reader will certainly better understand the explanations that follow when it will have studied and read the sections of General Relativity, Quantum Field Theory and therefore the Wave Quantum Physics, so if necessary he can skip the following explanations.

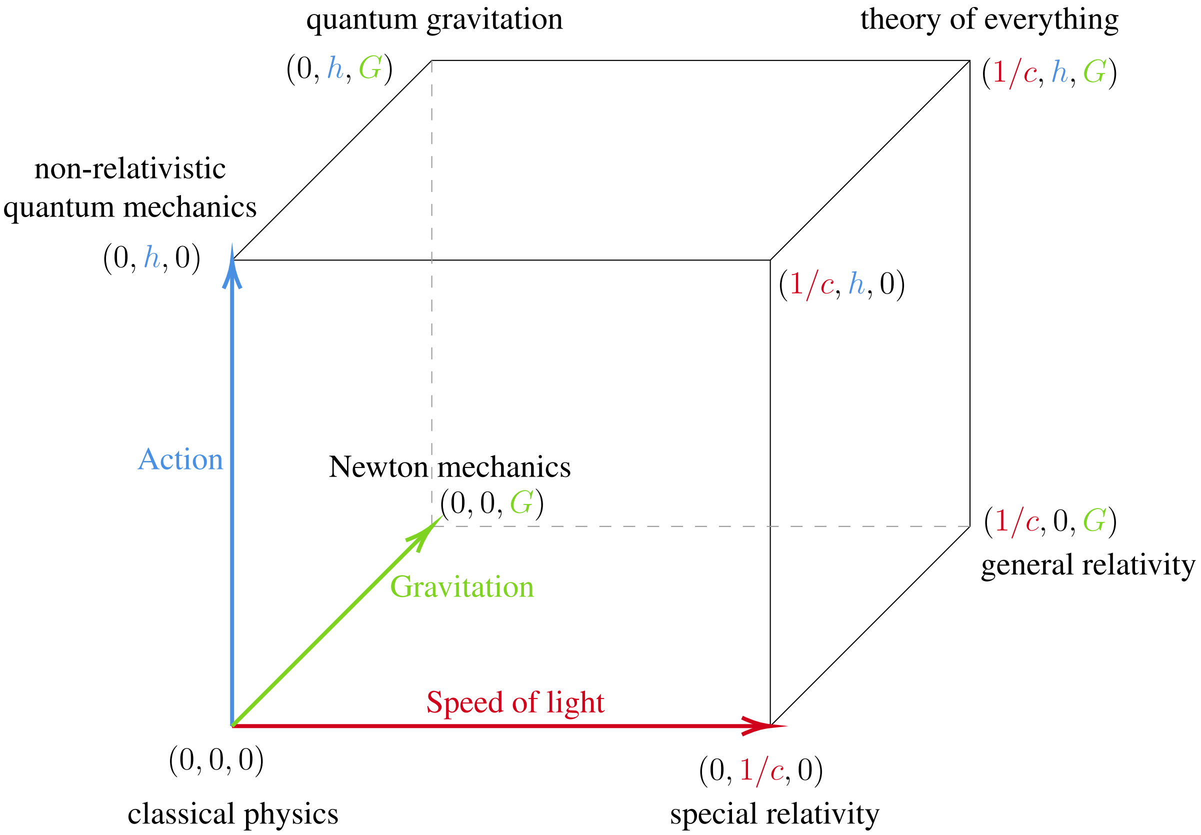

quantum gravitation
theory of everything
$(0, h, G)$
$(1/c, h, G)$
non-relativistic quantum mechanics
$(0, h, 0)$
$(1/c, h, 0)$
Action
Newton mechanics
$(0, 0, G)$
$(1/c, 0, G)$
general relativity
Gravitation
Speed of light
$(0, 0, 0)$
classical physics
$(0, 1/c, 0)$
special relativity

Figure 9.6: Okun cube (author: Juergen Giesen)

When G is set to 0, this means removing all the gravitational forces and decoupling the matter of space and time. When $\hbar$ is set to 0, we delete the quantum character of nature and we decouple the corpuscular and wave nature (consequence of De Broglie relation), when $1/c$ is set to 0, the speed of light is infinite and the time and space are uncoupled from each other (consequence of Lorentz transformation). To visualize this, we consider the cube above introduced by the Soviet physicist Mikhail Bronshtein, who gets an idea originally developed by Lev Landau, Dmitri Ivanenko and Georgi Gamow.

Naturally, at the lowest level $(0, 0, 0)$, we find the Newtonian mechanics, which does not take into account relativistic, quantum or gravitational effects. At the top level where we consider the effect of one constant, we find the three of Special Relativity $(1, 0, 0)$, quantum mechanics $(0, 1, 0)$ and Newtonian gravitation $(0, 0, 1)$, three theories tested with great precision in their field of validity and therefore having the status of "scientific theory".

At an even higher level, the Quantum Field Theory in $(1,1,0)$ takes into account the quantum and relativistic effects; the General Relativity in $(1,0,1)$ takes into account the gravitational and relativistic effects, and Newtonian quantum gravity in $(0,1,1)$ is supposed to provide a quantum and non-relativistic description of gravity. Only the first two theories are currently tested experimentally and have therefore the status of "scientific theory".

The ultimate actual level is at $(1,1,1)$ the "theory of everything" (name very pretentious and too commercial), supposed to give a quantum and relativistic description of gravity. Its formulation is not yet known, even if String Theories (see section of the same name page 6099), intensively studied nowadays seem to be a serious candidates. These theories appear as limiting cases of a broader and deeper theory but not yet formulated: the M theory (M states for "Mother").

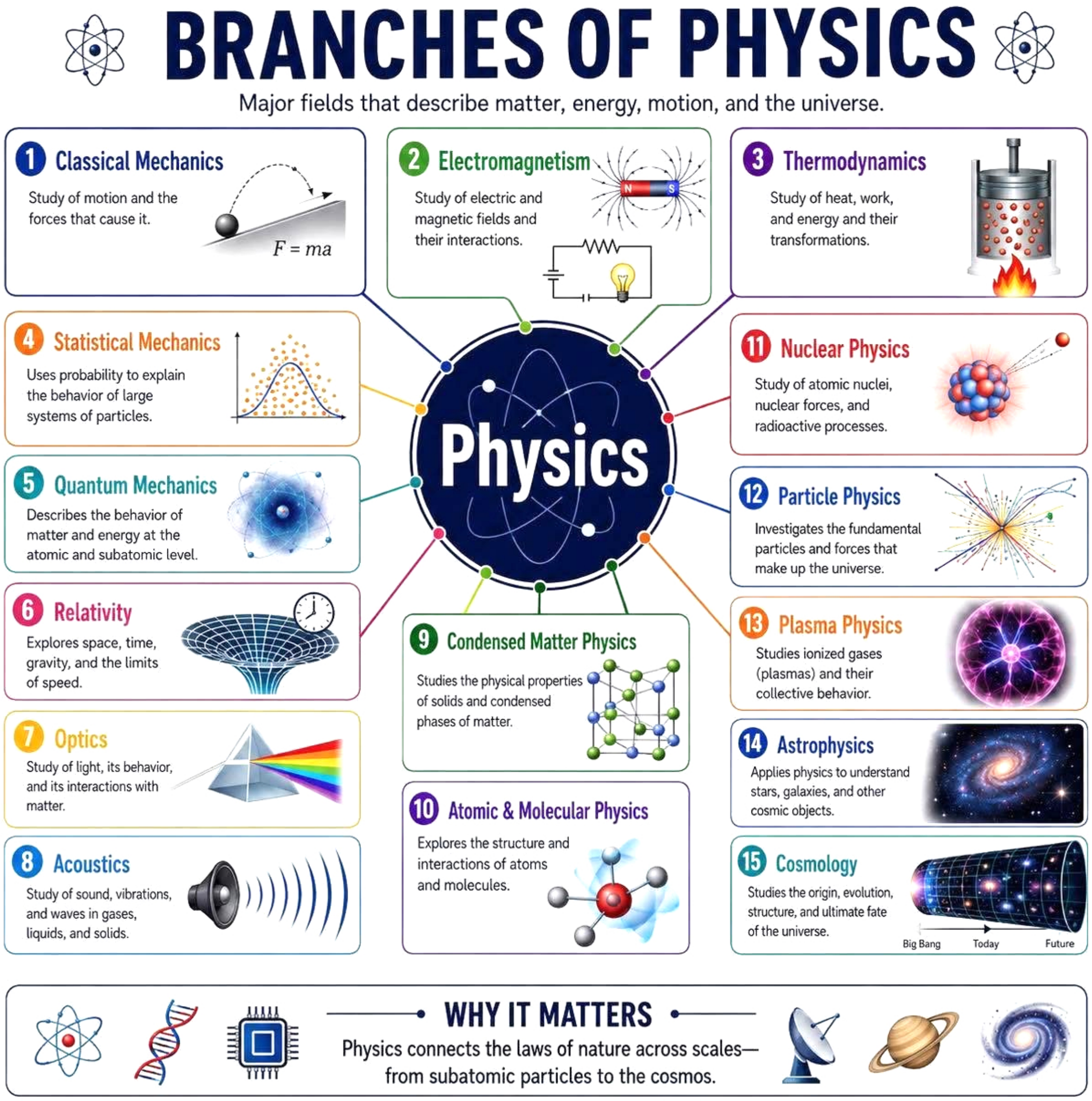
Caution!!! There is no principle of physics yet (in the beginning in the 121st century according to holocene calendar) that says physicals laws or constants have to be the same everywhere and always. Indeed, the reader must remember that the "laws" of physics are **descriptive**, not **imperative** and therefore they are some observed regularity in what we can observe experimentally with sufficiently wide occurrence or relevance! Asking whether the laws can change is no more and no less than asking whether regularities which have been observed up till now might not be quite as universal as was thought. The answer to that is: YES! This could drive the reader a bit mad. Indeed, how can we possibly do Science if the rules keep changing? But if you are a professional active scientist, you have to face up to that possibility - and you have a responsibility to check out if there is evidence beyond reasonable doubt about it! We will repeat this warning multiple times in this textbook as its seems that this is something necessary!

9.1.4 Principles of Physics

The advances of science in general and physics in particular were based two centuries before mainly on experimentation, that is to say that we reproduced in the laboratories given phenomena to analyse them systematically (the reproducibility with a given accuracy validating a hypothesis). This was equivalent as asking systematically specific questions to Nature and to describe and study the reactions caused thereby. The repetition at will of a phenomenon in an experiment would not be guaranteed if there was not a general principle of causality.

> *Physics doesn't captivate everyone - but if it draws you in naturally, take pride in being among the rare few whom the Universe seems to have chosen to explore its mysteries up close.*

quote 9.1:3127.#62



BRANCHES OF PHYSICS

Major fields that describe matter, energy, motion, and the universe.

Physics

1. **Classical Mechanics** Study of motion and the forces that cause it.

2. **Electromagnetism** Study of electric and magnetic fields and their interactions.

3. **Thermodynamics** Study of heat, work, and energy and their transformations.

4. **Statistical Mechanics** Uses probability to explain the behavior of large systems of particles.

5. **Quantum Mechanics** Describes the behavior of matter and energy at the atomic and subatomic level.

6. **Relativity** Explores space, time, gravity, and the limits of speed.

7. **Optics** Study of light, its behavior, and its interactions with matter.

8. **Acoustics** Study of sound, vibrations, and waves in gases, liquids, and solids.

9. **Condensed Matter Physics** Studies the physical properties of solids and condensed phases of matter.

10. **Atomic & Molecular Physics** Explores the structure and interactions of atoms and molecules.

11. **Nuclear Physics** Study of atomic nuclei, nuclear forces, and radioactive processes.

12. **Particle Physics** Investigates the fundamental particles and forces that make up the universe.

13. **Plasma Physics** Studies ionized gases (plasmas) and their collective behavior.

14. **Astrophysics** Applies physics to understand stars, galaxies, and other cosmic objects.

15. **Cosmology** Studies the origin, evolution, structure, and ultimate fate of the universe.

WHY IT MATTERS

Physics connects the laws of nature across scales—from subatomic particles to the cosmos.

9.1.4.1 Principle of Causality

Definition (#9.1.419): We define the "principle of causality" by the fact that in exactly the same conditions, the same causes always lead to the same effects. In other words, if certain initial conditions are well known, the phenomenon will take place in a determined manner, always the same (at least in Classical Physics, but not in Quantum Physics!!!).

In fact, experience is not always necessary if we consider the special case of "first order principles" which are by definition “*basic proposition or assumption that cannot be deduced from any other proposition or assumption*”.

This method involves breaking a problem down to its most basic, established laws and building up a solution from there. A key example is deriving an empirical law, like the Stefan-Boltzmann law, directly from fundamental physics principles, rather than using experimental data.

However, our modern society often give only a little time left to the great people of science to think about these first-order principles by imaginary experiences (very popular method used by Albert Einstein for the parenthesis...).

It is in a trilemma proposed by the scepticist of antiquity named Agrippa, according to an argument reported by Sextus Empiricus, that the question of justification of knowledge (based on reasoning and not on instrumentalism!) has been put most explicitly:

H1. Knowledge is ultimately based on first principles but arbitrary one (axiomatic argument, which rests on accepted precepts).

H2. Or we do not find such principles and we have an infinite regress (regressive argument, in which each proof requires a further proof, ad infinitum).

H3. Or justification is circular (circular argument, in which theory and proof support each other).

The last two methods of reasoning are fundamentally weak, and because the Greek sceptics advocated deep questioning of all accepted values, they refused to accept proofs of the first sort. The trilemma, then, is the decision among the three equally unsatisfying options.

This trilemma has also often, in contemporary philosophy, especially by Karl Popper, be nameed "Fries trilemma" or "Münchhausen trilemma" or "Agrippa trilemma" and we currently do not know in what scenario (H1, H2 or H3) we are.

On multiple occasions in history, it has been asked how we can know for certain what is true and what is not. For example, it is possible that rather than being many billions of years old as the scientific evidence strongly suggests, the Universe was actually made last Thursday. It is also possible that the entire world is the creation of your own mind, like a dream it takes you several decades to wake up from. Again, the is why we never speak about "proofs" in experimental science but about "strong statistical evidence", or "statistical evidence above any reasonable doubts" (see section Statistics page 948), and "consensus".

Attempts to attach any objective importance to speculations on this latter topic may alternatively be called "bullshit"...

Let us now introduce these fundamental principles (or assumptions):

9.1.4.2 Principle of Conservation of Energy

The first principle of conservation of energy is stated as following (basically at least... see below ... for some remarks ...): The total energy, denoted E or E_{tot} of any isolated and inertial system does not vary with time relatively to a fixed reference frame if there is no transfer of energy (or mass/radiation) or heat from outside the system.

This principle can be expressed by the formula:

$$\boxed{\Delta E_{\text{tot}} = \Delta U + \Delta E_c + \Delta E_p} \qquad (9.1.51)$$

where ΔE_{tot} is the total energy variation of the system, ΔU the variation of internal energy of the system, that is to says its own energy corresponding to the kinetic and potential microscopic energies of particles that constitute it.

The term ΔE_c, also sometimes denoted ΔE_{cin}, is the change in kinetic energy at a macroscopic scale (motion of the system in a given reference frame) and the term ΔE_p, also sometimes denoted ΔE_p, the variation of potential energy at a macroscopic scale, for the system interacting with gravitational or electromagnetic fields.

In physics, a conservation law (nothing is lost, nothing is created) expresses the idea that a particular measurable property of a local isolated physical (closed) system remains constant during the evolution of this system. For example for energy this means in the case of a conservative system:

$$\Delta U + \Delta E_c + \Delta E_p = 0 \qquad (9.1.52)$$

The list below lists some conservation laws useful to the engineer and have never been put in default as far as we know until this day and resulting for most of them of the conservation energy principle:

- Conservation of inertia
- Conservation of momentum
- Conservation of angular momentum
- Conservation of electric charge
- Conservation of the magnetic flux
- Conservation of mass

The Noether's theorem that we will see further below expresses the equivalence between conservation laws and invariance of physical laws regarding certain transformations (typically named "symmetries")!!!!!!

This theorem applies only to system describable by a Lagrangian (see next section of Analytical Mechanics). For example, the invariance by translation over time implies that energy is conserved (!!!), the translation invariance in space implies that the amount of movement is maintained (!!!), and the rotation invariance in space implies that the angular momentum is conserved (!!!).

This equivalence can be proven and follows from the invariance in time of the laws of physics. This is the first Noether's principle (theorem) that we'll prove a little further below.

Remarks

R1. Energy can be found as we know in several forms (this does not to mean that there are several types of energies!!!) as heat, kinetic energy, potential, electrical, magnetic, etc.

For example, in consumer applications, including in the field of nutrition, we often express the energy in "calories". The calorie being strictly speaking the energy we must provide to heat up one gram of pure water of one degree Kelvin (or one degree Celsius) in standard conditions of pressure as we will see more in details in the section of Thermodynamics (we can convert calories to joules by multiplying by approximately 4.186). Even worst... nutritionists name for communication purposes "calorie" what physicists name (properly) "kilocalorie".

R2. The violation of this principle of conservation of energy in an isolated system has never been observed but its validity can not be proved as far as we know (hence the fact that it is a "first principle").

R3. Some physicists are debating the fact that this first principle follows from Noether's theorem we shall see. But it is quite debatable as Noether's theorem considers the potential energy as constant over time...

9.1.4.3 Principle of Least Action

The "principle of least action" (also named "variational principle") is stated as following:

All natural phenomena are consistent with the fact that Nature, in producing its effects, always acts by the simplest and most direct way.

This single rule as we will see it through this entire book underpins all of physics.

With this statement and the principle of conservation of energy, we can then establish a great powerful mathematical tools for the study of theoretical physics. But we can not introduce this tool at this level of our discussion since it requires mathematical tools that we will introduce later in details, in the section of Analytical Mechanics (during the study of the Lagrangian formalism to be more precise at page 3155).

Meanwhile here are the two main relations that summarize this principle:

$$\boxed{\begin{gathered}S(q,\dot{q})=\int\limits_{t_A}^{t_B}L(q,\dot{q})\mathrm{d}t\\ \frac{\partial L}{\partial q_i}-\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}_i}=0\end{gathered}} \tag{9.1.53}$$

We will see through this textbook that all 121st century (holocene calendar) science can actually be derived from the latter, like Quantum Field Theory and General Relativity for example (and we will prove that in details in the corresponding sections of our book!).

Remark

The violation of this principle in an isolated system has never been observed as far as we know but its validity can also not be proved as far as we know (hence the fact that it is also a "first principle").

If evolution by natural selection can be connected to the Principle of Least Action, which has been theorised on for many years, then it corresponds to the direction of Ultimate Causation. As an organism evolves and follows a path of Proximate Causation, if the vector is closely parallel to that of the Ultimate Causation then the changes will confer desirable heritable attributes which will lead to further development. If however the variations do not occur in a direction close to the of the Ultimate Causation vector the evolved organism will quickly die out. This may be viewed as similar to Feynman's sum over all histories. Therefore, the Principle of Least Action gives a direction to Evolution.

9.1.4.4 Noether's Principle (Noether's theorem)

The Noether's principle, traditionally named "Noether's theorem" as it can be proved, combines elegantly conserved physical quantities to the symmetries of the laws of nature!!!

That is:

- Translational symmetry in the time (time invariant phenomenon) corresponds to the energy conservation
- Translation in the space corresponds to the conservation of momentum (quantity of motion)
- Rotation in space corresponds to the conservation of angular momentum
- Gauge transformations corresponds conservation of electric charge (see section Quantum Field Theory page 5180), weak and color charge.

The reader must keep in mind that energy and momentum are not the same because they are fundamentally different quantities: momentum is a vector (it has direction) and is defined by mass times velocity, while energy is a scalar (it has no direction) and is related to the capacity to do work!

In other words, Noether's theorem states that physics is:

- Symmetrical (invariant) by translation in time (this having the consequence that there is no origin in time)
- Symmetrical (invariant) by translation in space (this having the consequence that there is no origin in the space)
- Symmetrical (invariant) by rotation (this having the consequence that there is no preferred direction in space)

Noether's theorem is important, both because of the insight it gives into conservation laws, and also as a practical calculational tool. It allows investigators to determine the conserved quantities (invariants) from the observed symmetries of a physical system. Conversely, it allows researchers to consider whole classes of hypothetical Lagrangians with given invariants, to describe a physical system. As an illustration, suppose that a physical theory is proposed which conserves a quantity X. A researcher can calculate the types of Lagrangians that conserve X through a continuous symmetry. Due to Noether's theorem, the properties of these Lagrangians provide further criteria to understand the implications and judge the fitness of the new theory.

Remarks

R1. This theorem implies that a Galilean reference frame (see section Classical Mechanics page 3157) is homogeneous (no privileged time or space origin) and isotropic (no preferred direction).

R2. We must not confuse between the invariance of the laws and the non-invariance of the theoretical solutions which lead to these laws! For example, the discharge of a capacitor (see section Electrical Engineering page 8467) is translation invariant in time but not its solution.

This result established in 11915 (holocene calendar) by Emmy Noether just after arriving in Göttingen, was described by Albert Einstein as "*monument of mathematical thinking*". It is now one of the pillars of theoretical physics.

Today it is still often presented only during course on Quantum Field Theory. This makes it seem more complicated and mysterious than it is, and this is forget that it also applies to Classical Mechanics!!!

Remark

We recommend to most readers to read the proof of Noether's theorem in parallel of the section Analytical Mechanics (see page 3155).

Thus, symmetries play a major role in physics. They allow in one hand to simplify the problems and on the other hand to draw new laws. To illustrate the first application of symmetries let us simply evoke the mathematical form of the gravitational potential generated by a punctual mass located at the origin of reference frame (see section Classical Mechanics page 5266). In Cartesian coordinates, the expression of the gravitational potential is relatively complex:

$$\phi=-\frac{G}{\sqrt{x^2+y^2+z^2}} \tag{9.1.54}$$

while in spherical coordinates (coordinate system that takes advantage of the spherical symmetry of the potential) it takes a very simple form:

$$\phi=-\frac{G}{r}\;[\text{J}\cdot\text{kg}^{-1}] \tag{9.1.55}$$

The properties of symmetry of a problem are exploited here in order to simplify the mathematical treatment of physical laws. Although these mathematical considerations tell us about the physical properties of the system considered, they conserve however a purely technical nature.

Symmetries also found another application whose physical significance is much deeper. The fact, not accidental, that a system possesses symmetries should certainly have physical implications. Intuitively, we understand that the presence of symmetries in a physical system translates the invariance of some of its physical properties under the application of space-time transformations or, more generally: under geometric transformations. The invariance of physical properties must necessarily induce relations of a new kind between the variables of the system. Such relations must in turn reveal deeper laws that combine the geometry of the system to the laws of Nature. This reasoning, even if it is intuitive, invites us to delve deeper into the relations that may exist between the physics laws and the geometrical properties of space-time.

Let us consider a mechanical experience more or less complex observed simultaneously by two physicists O and O' at rest or in uniform translation located in different places as each of them chooses a reference frame which he is the origin.

They undertake to measure various physical quantities and obtain numerically results that generally differ. However, the physical laws they draw (same level of knowledge) are identical!! This conclusion is trivial because we all know that the laws of Nature must not depend not on the location of the observers in uniform translation or a rest.

Mathematically, the difference between repositories O and O' following the chosen reference frame of the studied experiment is the change of basis from one to the other by a rotation $R(\phi)$ and / or translation $T(l)$ (see section Euclidean Geometry page 2640).

The fact that the physical laws are independent of the position of the observer means that they do not vary after have applied to them a rotation and / or translation. We then say that they are "invariant under rotation and translation" or they are "symmetrical in rotation and translation".

Remark

In physics, the definition of symmetry is similar to that of mathematicians but applies to the laws of nature and not the geometric figures. Thus a physical symmetry is a transformation of system variables - which may be geometrical variables or more abstract variables - that does not change the formulation of physical laws.

Let us give a more or less rigorous definition of a physical symmetry.

Before let us consider a system S whose state evolves over time. Let us denote the state of S at time t by $S(t)$. Then at the initial moment t_0, S is therefore in the state $S(t_0)$. Let us consider a geometrical transformation T (rotation or translation) that acts at each point of space and optionally of time. In a time t, the action of T on the system S has for effect of transforming it into a system $S'(t)=T[S(t)]$ such that at the moment t_0, the transformation by T of $S(t_0)$ is $S'(t_0)$.

Definition (#9.1.420): A transformation T named a "physical symmetry" if the transformation by T of a system S (which gives S') evolves in the same way as S, that is to say, if we apply the laws physics on $S'(t_0)$ to know its states S'' in a later time t then $S''(t)=S'(t)=T[S(t)]$.

9.1.4.4.1 Invariance by translation in space

Let us consider an isolated system consisting of n interacting particles identified by the position vectors $\vec{r}_i$. The interaction of two particles i,j derives from a potential $\varphi(\vec{r}_i,\vec{r}_j,t)$ (see section Classical Mechanics page 3351). Each particle is subjected to forces resulting from the interaction with other particles. For a given particle i for example, the resultant of these forces is expressed according to Newton's second law (see section Classical Mechanics page 3202):

$$\vec{F}_i(\vec{r},t)=m\ddot{\vec{r}}(\vec{r}_i,t)=\sum \vec{F}_{i,j}(\vec{r}_i,\vec{r}_j,t)=\sum -m\vec{\nabla}\cdot\varphi(\vec{r}_i,\vec{r}_j,t) \tag{9.1.56}$$

Let us apply to the system the following translation in space:

$$\vec{r}_i \rightarrow \vec{r}_i^{\,\prime}=\vec{r}_i+\vec{l} \tag{9.1.57}$$

where $\vec{l}$ is any space direction vector.

Say that the translation in space of a system is a symmetry means that the acceleration and the force acting on each particle are unchanged after the transformation:

$$\vec{F}_i(\vec{r},t)=\vec{F}_i(\vec{r}_i+\vec{l},t) \tag{9.1.58}$$

Which implies:

$$\sum -m\vec{\nabla}\cdot\varphi(\vec{r}_i+\vec{l},\vec{r}_j+\vec{l},t)=\sum -m\vec{\nabla}\cdot\varphi(\vec{r}_i,\vec{r}_j,t) \tag{9.1.59}$$

This equality must be true regardless of the position of the particles, so whatever are the $\vec{r}_i$ and the $\vec{r}_j$. It is clear that the only way to check the last equality in these conditions is to equalize then by pairs the potential between each particle j with the particle i, that is to say:

$$\varphi(\vec{r}_i+\vec{l},\vec{r}_j+\vec{l},t)=\varphi(\vec{r}_i,\vec{r}_j,t) \tag{9.1.60}$$

The potentials are then necessarily (and this is the power of Noether's theorem!) functions of the $\vec{r}_i-\vec{r}_j$ such that:

$$\varphi(\vec{r}_i,\vec{r}_j,t)=\varphi(\vec{r}_i-\vec{r}_j,t) \tag{9.1.61}$$

Therefore, we deduce that:

$$\vec{F}_{i,j}(\vec{r}_i,\vec{r}_j,t)=-m\vec{\nabla}\cdot\varphi(\vec{r}_i-\vec{r}_j,t)=-\left(-m\vec{\nabla}\cdot\varphi(\vec{r}_j-\vec{r}_i,t)\right)=-\vec{F}_{j,i}(\vec{r}_j,\vec{r}_i,t) \tag{9.1.62}$$

This immediately implies that the resultant of all the forces applied to the particles in the system is zero and therefore the total linear momentum is conserved:

$$\sum\vec{F}_{i,j}(\vec{r}_i,\vec{r}_j,t)+\sum\vec{F}_{j,i}(\vec{r}_j,\vec{r}_i,t)=\sum\vec{F}_i=\vec{0}\rightarrow\frac{\mathrm{d}}{\mathrm{d}t}\left(\sum m\,\dot{\vec{r}}_i\right)=\vec{0}\Leftrightarrow\sum m\,\dot{\vec{r}}_i=c^{te} \tag{9.1.63}$$

The translation invariance in space of Newton's law thus leads necessarily to:

1. The potential between the particles of an isolated system is a function of their relative distance (this will be confirmed in the section Astronomy during our study of the gravitational potential field, as well as in the section of Electrostatics regarding the electrostatic potential and also the Yukawa potential with spherical symmetry during our study of Quantum Field Theory).
2. The law of equality between action and reaction.
3. The conservation of total linear momentum of a system!!!

As consequence of the point (3) the origin of the space is unobservable (since conservation of linear momentum is equivalent to translation invariance in space)!

9.1.4.4.2 Invariance by rotation in space

Let us request now that rotations about a fixed point are symmetrics. This property must be true whatever the considered fixed point, in particular, if the fixed point is precisely the position of one of the particles in the system. It follows that the potential necessarily follows a spherical symmetry. The forces acting between the particles are then collinear to vectors that link them.

The angular momentum of the system is then expressed as follows (see section Classical Mechanics page 3244):

$$\vec{b}=\sum\vec{b}_i=\sum m_i\vec{r}_i\times\dot{\vec{r}}_i \tag{9.1.64}$$

The derivative with respect to the time of the total angular momentum gives:

$$\frac{\mathrm{d}\vec{b}}{\mathrm{d}t}=\sum m_i\frac{\mathrm{d}}{\mathrm{d}t}(\vec{r}_i\times\dot{\vec{r}}_i)=\sum m_i\frac{\mathrm{d}\vec{r}_i}{\mathrm{d}t}\times\dot{\vec{r}}_i+\sum m_i\vec{r}_i\times\ddot{\vec{r}}_i=\sum m_i\vec{r}_i\times\ddot{\vec{r}}_i=\sum\vec{r}_i\times m_i\ddot{\vec{r}}_i \tag{9.1.65}$$

But the last term of the vector product can be written:

$$m_i\ddot{\vec{r}}=\vec{F}_{ij} \tag{9.1.66}$$

where the $\vec{F}_{ij}$ are the internal system forces of the particle j acting on the particle i.

The prior-previous expression the becomes:

$$\frac{\mathrm{d}\vec{b}}{\mathrm{d}t}=\sum\vec{r}_i\times\vec{F}_{ij} \tag{9.1.67}$$

We can group the terms $\vec{r}_j\times\vec{F}_{ji}$ and $\vec{r}_i\times\vec{F}_{ij}$ terms by pairs and by the property of the cross product we have necessarily:

$$\frac{\mathrm{d}\vec{b}}{\mathrm{d}t}=\sum\vec{r}_i\times\vec{F}_{ij}=\vec{0} \tag{9.1.68}$$

So we conclude that the angular momentum is conserved and that the conservation of angular momentum is thus equivalent to the rotational invariance.

Consequence: there is no preferred direction in space!

9.1.4.4.3 Invariance by translation in time

The total energy of a system is the sum of the kinetic energy of its particles and the potential energy resulting from the mutual interaction of this same particles, either in the form of Classical Mechanics:

$$E_{\text{tot}}=E_c+E_p=\sum\frac{1}{2}m_i\dot{\vec{r}}_i^{\,2}+V_{i,j}(\vec{r}_i,\vec{r}_j) \tag{9.1.69}$$

We will assume that the potential $V_{i,j}(\vec{r}_i,\vec{r}_j)$ does not vary with time. This assumption is justified empirically by the fact that potential observed in nature are independent of time in closed systems at equilibrium.

Let us calculate the derivative of total energy with respect to time:

$$\begin{aligned}\frac{\mathrm{d}E_{\text{tot}}}{\mathrm{d}t}&=\frac{\mathrm{d}}{\mathrm{d}t}\sum\frac{1}{2}m_i\dot{\vec{r}}_i^2+\frac{\mathrm{d}}{\mathrm{d}t}V_{i,j}(\vec{r}_i,\vec{r}_j)=\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\sum m_i\dot{\vec{r}}_i^2+0=\frac{1}{2}\sum\frac{\mathrm{d}m_i}{\mathrm{d}t}\dot{r}_i^2+\frac{1}{2}\sum\frac{\mathrm{d}\dot{\vec{r}}_i^2}{\mathrm{d}t}m_i\\&=\frac{1}{2}\sum\frac{\mathrm{d}m_i}{\mathrm{d}t}\dot{\vec{r}}_i^2+\sum\dot{\vec{r}}_i\frac{\mathrm{d}\dot{\vec{r}}_i}{\mathrm{d}t}m_i=\frac{1}{2}\sum\frac{\mathrm{d}m_i}{\mathrm{d}t}\vec{v}_i^2+\sum\vec{v}_i\frac{\mathrm{d}\dot{\vec{r}}_i}{\mathrm{d}t}m_i\end{aligned} \tag{9.1.70}$$

But if the system is closed (no external mass input or external energy input), the term $\mathrm{d}m_i/\mathrm{d}t$ is zero (no relativistic mass change either because the speed of each particle or the entire system is constant or its variation is in average equal to zero or just because we are in a comoving reference frame). It is the same for the term $\mathrm{d}\dot{\vec{r}}_i/\mathrm{d}t$ where if the system is closed (no external mass input or external energy input) the average acceleration of each particle or the entire system relatively the center of gravity will be zero. So:

$$\frac{\mathrm{d}E_{\text{tot}}}{\mathrm{d}t}=\frac{1}{2}\sum\frac{\mathrm{d}m_i}{\mathrm{d}t}\vec{v}_i^2+\sum\vec{v}_i\frac{\mathrm{d}\dot{\vec{r}}_i}{\mathrm{d}t}m_i=\frac{1}{2}\sum 0\cdot\vec{v}_i^2+\sum\vec{v}_i\cdot\vec{0}\cdot m_i=0 \tag{9.1.71}$$

So we conclude that the total energy of such a system is a constant! The above result is often more elegantly rewritten using the Lagrangian formalism that we already know as:

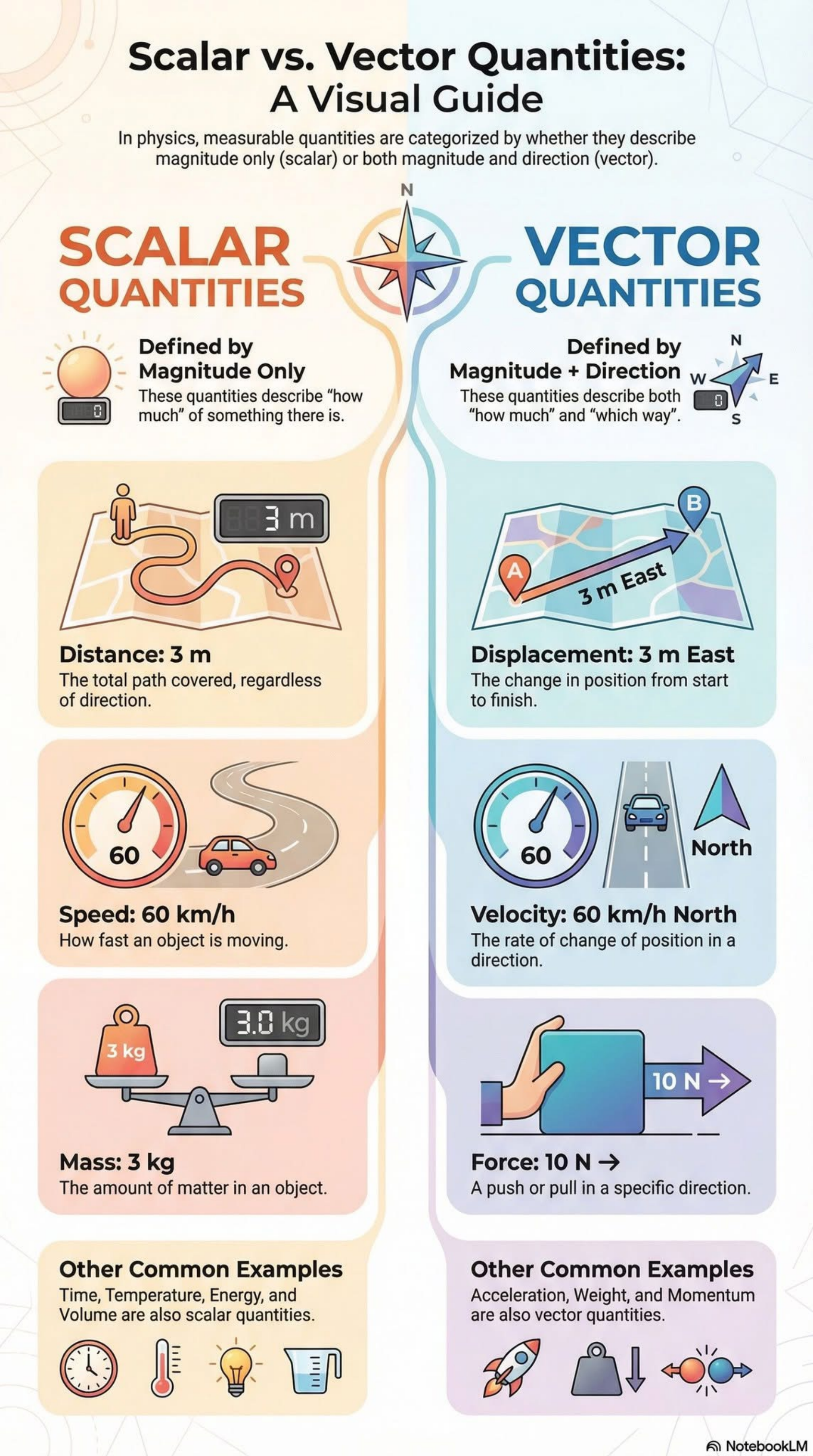
$$\boxed{\frac{\partial L}{\partial q}-\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial\dot{q}}\right)=0\Leftrightarrow\frac{\mathrm{d}L}{\mathrm{d}t}=0} \tag{9.1.72}$$

Caution! The Universe at its beginning was only made of radiation (photons). Therefore there was no mass, hence no time! So the law of conservation of energy doesn't apply to the beginning of the Universe as we observe it with experimental data of the early 121st century (holocene calendar)! Even the actual Universe isn't invariant by time translation because of the space frame that is expanding (see [22] and [129]). So the law of conservation applies only to small (classical) local frames typically like physics laboratories made by humans at their scale! On a very small scale, the law of conservation of energy is respected only statistically, however it is not systematically respected at the individual quantum level (see [11]).

Ask the question "*what is the mechanical quantity invariant by translation*" is equivalent to ask "*what are the mathematical quantities which are unchanged when we apply to them a translation*". There are two of them: the scalars and vectors.

Intuitively, a scalar is treated as a real number (see section Numbers page 209). But, in mechanics, the real numbers that we can construct are made from vector quantities as the vectors position, velocity, etc. For such a real number to have the status of scalar it must be independent of space. Thus, a position vector can obviously not be regarded as a scalar. The energy of the particle is a real number, but is not a scalar because as it depends explicitly, in its formulation, of the position vector.

Similarly, a vector is not only a mathematical entity having its components in a base. To have the status of vector, a mathematical entity must be transformed in the same way as the basis vectors of the vector space. By this definition, the angular momentum is not a true vector as, being the composition by the cross product of two vectors, it does not change as the basis vectors. From a mathematical point of view it is a pseudo-vector (see section Vector Calculus page 1799).



Scalar vs. Vector Quantities: A Visual Guide

In physics, measurable quantities are categorized by whether they describe magnitude only (scalar) or both magnitude and direction (vector).

SCALAR QUANTITIES

Defined by Magnitude Only

These quantities describe "how much" of something there is.

Distance: 3 m

The total path covered, regardless of direction.

Speed: 60 km/h

How fast an object is moving.

Mass: 3 kg

The amount of matter in an object.

Other Common Examples

Time, Temperature, Energy, and Volume are also scalar quantities.

VECTOR QUANTITIES

Defined by Magnitude + Direction

These quantities describe both "how much" and "which way".

Displacement: 3 m East

The change in position from start to finish.

Velocity: 60 km/h North

The rate of change of position in a direction.

Force: 10 N →

A push or pull in a specific direction.

Other Common Examples

Acceleration, Weight, and Momentum are also vector quantities.

Figure 9.7: **Reminder of scalar vs vector quantities in physics**

The only real vector that remains is the linear momentum as it is built using the derivative of the position vector which is, of course, a real vector. We conclude that the only quantity likely to be retained by translation is the total linear momentum of the system.

By similar reasoning as the above one, it is possible to assume what quantity could be invariant under rotation. Knowing that only scalar and some pseudo-vectors are indeed invariant under rotation, we conclude that the only quantity that can be retained when rotations occurs is the total angular momentum of the system.

Finally, still by the same type reasoning, the invariance of the laws of mechanics by time translation is equivalent to search the quantities conserved by a translation in time. These quantities are the real scalar and vector on the time-line. No mechanical quantity can be treated as a vector on the time-line. However, energy is a scalar indeed translationally invariant in time, since the potential energy is independent by hypothesis of the time. The invariance of the laws of mechanics by translation in time therefore suggests intuitively the conservation of energy.

These reasoning can obviously not be used as a proof but they show a close relation between geometry and invariance properties of a system.

9.1.4.4.4 Noether's theorem

We will now prove that Noether's (first) theorem states that every differentiable symmetry of the action of a physical system has a corresponding conservation law.

For this purpose let us consider L the Lagrangian (see section Analytical Mechanics page 3155) of a system represented by the $2n$ generalized coordinates $q_i, \dot{q}_i$ (in a space of dimension n) in the configuration space. Let us suppose that this system is invariant under the following infinitesimal transformation denoted h_s :

$$h_s(q): \quad q_i \rightarrow {q'}_i \tag{9.1.73}$$

Where s is a real number (rotation angle, translation distance, or translation time value) and continuous parameter and for which we have:

$$h_{s=0}(q) = q \tag{9.1.74}$$

The function $h_s(q)$ acts continuously on the variational path following the intellectual approach that will be stated in the section of Analytical Mechanics.

Let us suppose that the functions $q_i, \dot{q}_i$ are solutions of the Lagrange equations (which we will prove in the section of Analytical Mechanics). According to our assumptions the functions (defined):

$$\varphi(s,t) = h_s \qquad \dot{\varphi}(s,t) = \frac{\mathrm{d}}{\mathrm{d}t} h_s \tag{9.1.75}$$

are therefore necessarily also solutions of the Lagrange equations, which is translated explicitly by (we omit the indication of the sum after to make the notation more light...!):

$$\sum_i \left(\frac{\partial L(\varphi(s,t),\dot{\varphi}(s,t))}{\partial q_i} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L(\varphi(s,t),\dot{\varphi}(s,t))}{\partial \dot{q}_i} \right) = 0 \tag{9.1.76}$$

Moreover, by hypothesis, the Lagrangian is invariant for the transformations of the type described by $h_s : q_i \rightarrow {q'}_i$. It follows that its derivative with respect to the parameter s is necessarily zero:

$$\frac{\mathrm{d}}{\mathrm{d}s} L(\varphi(s,t), \dot{\varphi}(s,t)) = 0 \tag{9.1.77}$$

And we will prove in the section of Analytical Mechanics (under the form of an integral being equal to zero) the relation:

$$\frac{\partial L}{\partial q_i}\delta q_i + \frac{\partial L}{\partial \dot{q}_i}\delta \dot{q}_i = 0 \tag{9.1.78}$$

which can finally be written:

$$\frac{\partial L}{\partial q_i}\delta \phi_s + \frac{\partial L}{\partial \dot{q}_i}\delta \dot{\phi}_{si} = 0 \tag{9.1.79}$$

by dividing the both sides by $\mathrm{d}s$:

$$\frac{\partial L}{\partial q_i}\frac{\mathrm{d}\varphi_i}{\mathrm{d}s} + \frac{\partial L}{\partial \dot{q}_i}\frac{\mathrm{d}\dot{\varphi}_i}{\mathrm{d}s} = 0 \tag{9.1.80}$$

but we also have from the Euler-Lagrange equation (see section Analytical Mechanics page 3162):

$$\frac{\partial L}{\partial q_i} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}_i} = 0 \Leftrightarrow \frac{\partial L}{\partial q_i} = \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}_i} \tag{9.1.81}$$

We then get:

$$\left(\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}_i}\right)\frac{\mathrm{d}\varphi_i}{\mathrm{d}s} + \frac{\partial L}{\partial \dot{q}_i}\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\mathrm{d}\varphi_i}{\mathrm{d}s}\right) = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{q}_i}\frac{\mathrm{d}\varphi_i}{\mathrm{d}s}\right) = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{q}_i}\frac{\mathrm{d}h_s}{\mathrm{d}s}\right) = 0 \tag{9.1.82}$$

So the quantity:

$$\frac{\partial L}{\partial \dot{q}_i}\frac{\mathrm{d}h_s}{\mathrm{d}s} \tag{9.1.83}$$

is a constant of the movement!

The Noether's theorem then is stated as follows:

Given a system with a Lagrangian $L(q_i, \dot{q}_i, t)$ on which we apply an infinitesimal transformation $h_s : q_i \rightarrow {q'}_s$, where s is a real and continuous parameter (angle rotation, distance translation, time value translation). Then there exist a constant of motion denoted $C({q'}_i, \dot{q'}_i, t)$ whose expression is given by:

$$\boxed{C({q'}_i, \dot{q'}_i, t) = \left(\frac{\partial L}{\partial \dot{q}_i}\frac{\mathrm{d}h_s}{\mathrm{d}s}\right)_{s\rightarrow 0}} \tag{9.1.84}$$

in other words: every differentiable symmetry of the action of a physical system has a corresponding conservation law.

Let us apply now the Noether's theorem to the cases previously studied. Let us fix an arbitrary Cartesian reference frame of origin O. Let us write $(\vec{e}_1,\vec{e}_2,\vec{e}_3)$ the orthonormal basis of this reference frame. Let us consider a system of n particles identified in O by their position vectors $\vec{r}_i$. The Lagrangian of this system is then $L(\vec{r}_{j,i},\dot{\vec{r}}_{j,i},t)$ where $j=x,y,z$ distinguishes the spatial components of the vectors $\vec{r}_i,\dot{\vec{r}}_i$.

Let us suppose now that the system is invariant by translation of distance s along the x-axis only. The translation along that axis will be written as follows:

$$r_{x,i}(t)\rightarrow h_s(r_{x,i}(t))=r_{x,i}(t)+s(\vec{e}_x\circ\vec{e}_x) \tag{9.1.85}$$

The constant of the motion given by application of Noether theorem is then (still following the x axis):

$$C({r'}_{x,i},\dot{r'}_{x,i},t)=\left(\frac{\partial L}{\partial\dot{r}_{x,i}}\frac{\mathrm{d}h_s}{\mathrm{d}s}\right)_{s\rightarrow 0}=\frac{\partial L}{\partial\dot{r}_{x,i}}\frac{\mathrm{d}}{\mathrm{d}s}\left(r_{x,i}(t)+s(\vec{e}_x\circ\vec{e}_x)\right)=\frac{\partial L}{\partial\dot{r}_{x,i}}\underbrace{(\vec{e}_x\circ\vec{e}_x)}_{=1}=\frac{\partial L}{\partial\dot{r}_{x,i}} \tag{9.1.86}$$

We will see in the section of Analytical Mechanics analytical that $\partial L/\partial\dot{r}_{x,j}$ is the conjugate momentum $p_{x,i}$. We conclude therefore that the conserved quantity following Noether's theorem is: $p_{x,i}$ that is to say, the total linear momentum of the system along the x axis !!!

If we did the same with the other axes, we would easily prove also the conservation of total momentum along the axes, allowing us to conclude that in the general case of an infinitesimal translation:

$$\mathrm{d}\vec{r}=s(\vec{e}_x+\vec{e}_y+\vec{e}_z) \tag{9.1.87}$$

the conserved quantity is then the total momentum of the system.

Let us suppose now that the system is invariant under rotation of an infinitesimal angle s around the z-axis. This rotation will be written (for the cross product imagine in your head what happen in the space... if you don't see let us know and we will add a figure):

$$r_i(t)\rightarrow h_s(r_i(t))=r_{x,i}(t)+s(\vec{e}_z\times\vec{r}_i(t)) \tag{9.1.88}$$

and it is therefore a vector.

And we also get:

$$C({r'}_i,\dot{r'}_i,t)=\left(\frac{\partial L}{\partial\dot{\vec{r}}_i}\frac{\mathrm{d}h_s}{\mathrm{d}s}\right)_{s\rightarrow 0}=\frac{\partial L}{\partial\dot{\vec{r}}_i}\circ\frac{\mathrm{d}}{\mathrm{d}s}\left(\vec{r}_i+s(\vec{e}_z\times\vec{r}_i)\right)=\frac{\partial L}{\partial\dot{\vec{r}}_i}\circ(\vec{e}_z\times\vec{r}_i) \tag{9.1.89}$$

Once again, knowing that:

$$\vec{p}_i=\frac{\partial L}{\partial\dot{\vec{r}}_i} \tag{9.1.90}$$

the conserved quantity obtained by application of Noether's theorem is therefore:

$$C({r'}_i,\dot{r'}_i,t)=\left(\frac{\partial L}{\partial\dot{\vec{r}}_i}\frac{\mathrm{d}h_s}{\mathrm{d}s}\right)_{s\rightarrow 0}=\vec{p}_i\circ(\vec{e}_z\times\vec{r}_i) \tag{9.1.91}$$

and we have shown in the section of Vector Calculus page 1911 that:

$$\vec{x}\circ(\vec{y}\times\vec{z})=-\vec{z}\circ(\vec{y}\times\vec{x})=-(-\vec{y}\circ(\vec{z}\times\vec{x}))=\vec{y}\circ(\vec{z}\times\vec{x}) \tag{9.1.92}$$

which brings us to write:

$$C(r'_i,\dot{r'}_i,t)=\vec{e}_z\circ(\vec{p}_i\times\vec{r}_i(t))=\vec{e}_z\circ\vec{b}_i=b_{z,i} \tag{9.1.93}$$

We prove in the same way the invariance of the Lagrangian under rotations in the other axes which leads to the conservation of the components along these axes of the total angular momentum of the system.

In conclusion, we have identified three well known different readings of the laws of physics (we omit the gauge transformations):

$$C(q'_i,\dot{q'}_i,t)=\left(\frac{\partial L}{\partial\dot{q}_i}\frac{\mathrm{d}h_s}{\mathrm{d}s}\right)_{s\to 0}$$

Observation	Conservation Law	Physical interpretation
Invariance of physics laws by translation	Conservation of momentum	Homogeneity of space: space has the same properties in all its points. $C\left(r'_{x,i},\dot{r}'_{x,i},t\right)=p_{x,i}$
Invariance of physics laws by rotation	Conservation of angular momentum	Isotropy of space: space has the same properties in all directions. $C\left(r'_{x,i},\dot{r}'_{x,i},t\right)=b_{z,i}$
Invariance of physics laws by translation in time	Conservation of energy	Homogeneity of the time: the laws of nature do not change over time. $\frac{\partial L}{\partial q}-\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial\dot{q}}\right)=0\Leftrightarrow\frac{\mathrm{d}L}{\mathrm{d}t}=0$

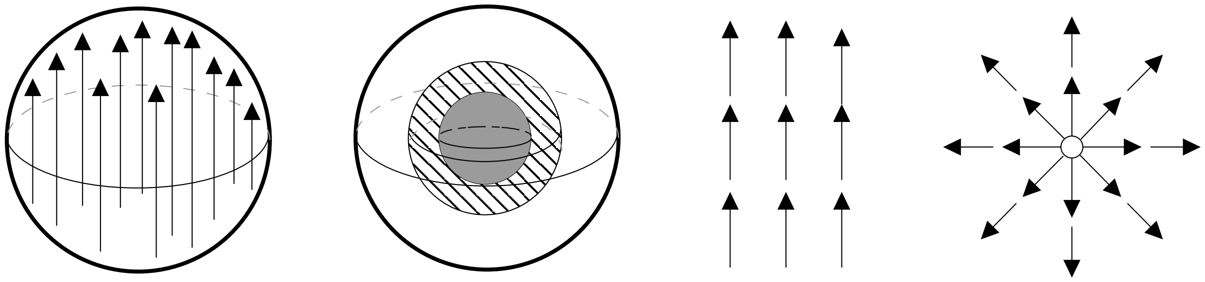
Table 9.15: **Conservation Laws**

By analogy we can see by the proofs above that momentum is to position what energy is to time (the reader must keep this information in mind for when we will discuss the Heisenberg uncertainty principle in the section of Quantum Wave Physics)!

In other words, the Universe seems to be:

P1. Homogeneous (no origin of time, or space, that is privileged) that is to say invariant under translation

P2. Isotropic (no preferred direction) that is to say invariant under rotations



homogeneous
not isotropic

isotropic
not homogeneous

homogeneous
not isotropic

isotropic
not homogeneous

Figure 9.8: Isotropic vs Homogeneous

9.1.4.5 Curie's Principle

The Curie's principle (which we owe to Pierre Curie) arises somewhat intuitively from Noether's theorem and is stated as following: If a cause has a certain symmetry or invariance, then its effect will be the same symmetry (or the same invariance), or a higher symmetry, provided that the problem of the solution is unique.

> **Remark**
>
> Note that the elements of symmetries act on the directions of the vectorial quantities, while invariance act on variables that depend on these quantities.

Thus, in a homogeneous and isotropic space, if we apply a geometric transformation to a physical system that is able to generate certain effects (forces, fields), then these effects undergo the same transformation.

In other words, if a physical system $\mathcal{S}$ has a certain degree of symmetry, then we can infer the effects generated by the system at a point from effects at another point.

Here are the 6 properties of symmetries derived from the Curie's principle:

P1. **Invariance by translation:** if $\mathcal{S}$ is invariant throughout translation parallel to an axis, the effects are independent of the coordinate axis (interest is then work in Cartesian coordinates).

P2. **Axial symmetry:** if $\mathcal{S}$ is invariant in rotation about a given axis, then its effects do not depend on the angle that defines the rotation (the interest being then to work in cylindrical coordinates).

P3. **Cylindrical symmetry:** if $\mathcal{S}$ is invariant by translation AND rotation, then its effects depend only on the distance from the axis of rotation (the interest being then to work in cylindrical coordinates).

P4. **Spherical symmetry:** if $\mathcal{S}$ invariant in rotation around a fixed point, then its effects depend only on the distance to the fixed point (the interest being then to work in spherical coordinates).

P5. **Plane of symmetry:** if $\mathcal{S}$ has a plane of symmetry, then at any point of this plane:

- an effect that has a vectorial origin is contained in the plane
- an effect that has pseudo-vector character (see the section of Vector Calculus for the definition of a pseudo-vector at page 1799) is perpendicular to the plane

P6. **Plane of antisymmetry:** if, by symmetry with respect to a plane, $\mathcal{S}$ is converted to $-\mathcal{S}$ then at any point of this plane:

- an effect that has a vectorial origin is perpendicular to the plane
- an effect that has pseudo-vector character is contained in the plane

9.1.5 Point Spaces

We have hesitated quite long to write the text below as it already partially in the section of Vector Calculus and respectively of Tensor Calculus. But finally do a summary of the main concepts of these both fields only for Classical Mechanics seems to be perhaps a good choice or at least a good reminder!

The study of physical phenomena in Classical Mechanics use initially their representation in the space of classical Euclidean geometry with one time dimension and any number of spatial dimensions.

The vectors we studied in the section of Vector Calculus (tensor of order 1) and the tensors (of any order) we studied also in the section of Tensor Calculus can as we have already mentioned, be used to connect each point in space-time to the origin of a given repository or reference frame, thereby forming vector and/or tensor fields. This mathematical fact, requires the mathematical definition of spaces formed by points also named "point spaces" or "punctual spaces".

The precise definition of punctual vector space-time that we will give will be built from the concept of vector space that we saw in the section of Vector Calculus (see page 1805).

Let us see first the particular example of the point space formed by triplets of numbers that is directly derived from the conventional three-dimensional euclidean geometrical space $\mathcal{E}^3$.

So, let us take two triplets of numbers denoted by:

$$\vec{A}=(a_1,a_2,a_3) \qquad \vec{B}=(b_1,b_2,b_3) \tag{9.1.94}$$

etc.

Let us denote by $\mathcal{E}'^3$ the set of all the elements $\{\vec{A},\vec{B},\ldots\}$ formed by triplets of numbers. At any pair $(\vec{A},\vec{B})$ of two elements of two elements of $\mathcal{E}'^3$, taken in this order (because in physics the time has a direction!!! the vectors have a given order), we can make correspond a vector $\vec{x}$ belonging to a vector space $\mathcal{E}^3$, and denoted geometrically by $\overrightarrow{AB}$ and defining this latter space by a triplet of numbers such that (we use the index notation as studied in the section of Tensor Calculus):

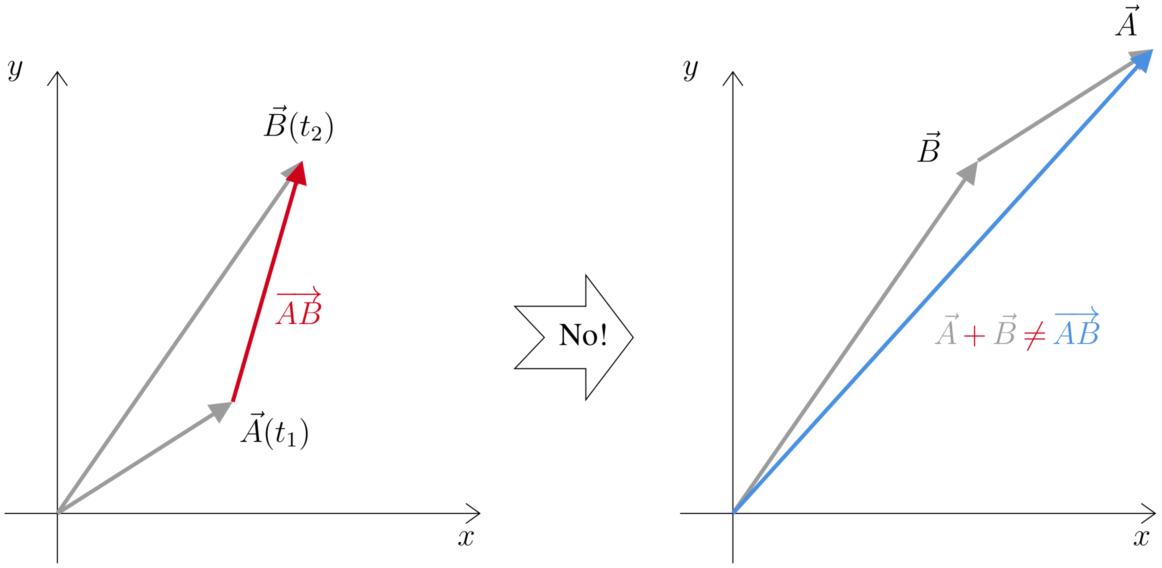
$$\vec{x}=\overrightarrow{AB}=\vec{B}-\vec{A}\Leftrightarrow x_i=b_i-a_i \tag{9.1.95}$$

with $i=1,2,3$.

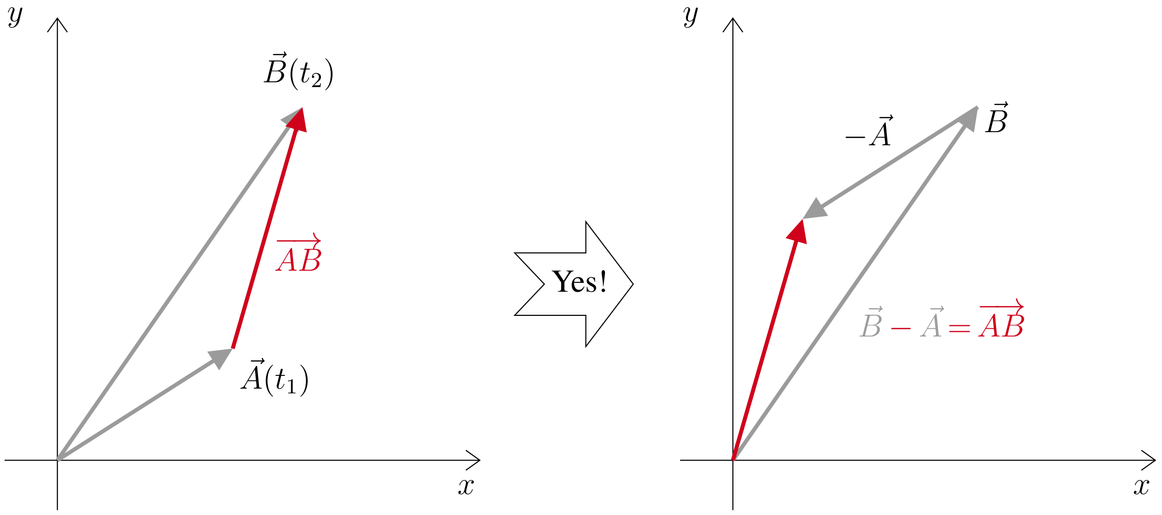
We then have:

$$\vec{x}=\overrightarrow{AB}=\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix} \tag{9.1.96}$$

Indeed, the choice of this expression comes once again from the fact that for example position vectors in physics are ordered by the time t:


y
$\vec{B}(t_2)$
$\overrightarrow{AB}$
$\vec{A}(t_1)$
x
No!
y
$\vec{A}$
$\vec{B}$
$\vec{A}+\vec{B}\neq\overrightarrow{AB}$
x

But:


y
$\vec{B}(t_2)$
$\overrightarrow{AB}$
$\vec{A}(t_1)$
x
Yes!
y
$-\vec{A}$
$\vec{B}$
$\vec{B}-\vec{A}=\overrightarrow{AB}$
x

If we define relatively to this element the addition and the multiplication by a scalar, we fall back on as we have already see it in the section of Set Theory with a vector space structure.

The correspondence we thus build between any ordered pair of elements $(\vec{A},\vec{B})$ of two elements of $\mathcal{E}'^3$ and a vector of a vector space of $\mathcal{E}^3$, obviously satisfies the following properties:

P1. Antisymmetry: $\overrightarrow{AB}=-\overrightarrow{BA}$

P2. Associativity with respect to addition: $\overrightarrow{AC}=\overrightarrow{AB}+\overrightarrow{BC}$

P3. If O (an origin) is an arbitrary point chosen in $\mathcal{E}'^3$, to any vector $\vec{x}$ of $\mathcal{E}^3$, there is a point M and a unique one, such that $\overrightarrow{\mathrm{O}M}=\vec{x}$.

Mathematicians say that a space can be say to be a "point space" if and only if the associated vector space respect the previous three relations (which seems quite obvious in fact...). The elements of $\mathcal{E}^3$ are then obviously named "points".

When $\mathcal{E}^3$ is associated with a dot product (see section Vector Calculus page 1821) then we know that $\mathcal{E}^3$ will be say to be a "pre-Euclidean point space".

Let us consider any point O of a pre-Euclidean point space and base $(\vec{e}_i)$ associated with it.

Definitions (#9.1.421):

D1. We name "reference of the space" $\mathcal{E}^3$ the set consisting of the elements O (origin) and of the base $(\vec{e}_i)$. This kind of reference frame is denoted:

$$(\vec{e}_i) := (\mathrm{O}, \vec{e}_i) = (\vec{e}_1, \vec{e}_2, \vec{e}_3) \tag{9.1.97}$$

D2. The "components" or "coordinates" of a point M of a pre-Euclidean space point $\mathcal{E}^3$, relative to the reference $(\vec{e}_i)$, are the (when not specified: contravariant[8]) components x^i of the vector $\vec{x} = \overrightarrow{\mathrm{O}M}$ relatively to the base $(\vec{e}_i)$.

Given two points M and M' of say $\mathcal{E}^3$ (as in fact we know that we can choose any number n of dimensions) defined by respective (contravariant) coordinates x^i and x'^i, we then have as we know from the section Tensor Calculus:

$$\overrightarrow{\mathrm{O}M} = x^i \vec{e}_i \qquad \overrightarrow{\mathrm{O}M'} = x'^i \vec{e}_i \tag{9.1.98}$$

Using the properties P1 and P2 given above:

$$\overrightarrow{MM'} = \overrightarrow{M\mathrm{O}} + \overrightarrow{\mathrm{O}M'} = -\overrightarrow{\mathrm{O}M} + \overrightarrow{\mathrm{O}M'} = (-x^i + x'^i)\vec{e}_i \tag{9.1.99}$$

We conclude that the components of the vector $\overrightarrow{MM'}$, in the base $(\vec{e}_i)$ are the 3 quantities $(x'^i - x^i)$ (n quantities in the case of a n dimension space), differences of the contravariant coordinates of the points M and M'.

Given $(\vec{e}_i)$ and $(\vec{e'}_j)$ any two basis of $\mathcal{E}^n$ and linked by the general relations of change of basis (see sections of Tensor Calculus page 2066 and Linear Algebra page 1992 for the details):

$$\vec{e}_i = A'^k_i \vec{e'}_k \qquad \text{and} \qquad \vec{e'}_k = A^i_k \vec{e}_i \tag{9.1.100}$$

Let us seek the relations between the coordinates of a point M of $\mathcal{E}^n$ relatively to these two bases. For this, let us express the vector $\overrightarrow{\mathrm{OO'}}$ and $\overrightarrow{\mathrm{O'O}}$ on each of these bases:

$$\overrightarrow{\mathrm{OO'}} = \alpha^i \vec{e}_i \qquad \overrightarrow{\mathrm{OO'}} = \alpha'^j \vec{e'}_j \tag{9.1.101}$$

and same for the vectors $\overrightarrow{\mathrm{O}M}$ and $\overrightarrow{\mathrm{O'}M}$, thus:

$$\overrightarrow{\mathrm{O}M} = x^i \vec{e}_i \qquad \overrightarrow{\mathrm{O'}M} = x'^j \vec{e'}_j \tag{9.1.102}$$

On another part we have:

$$\overrightarrow{\mathrm{O}M} = \overrightarrow{\mathrm{OO'}} + \overrightarrow{\mathrm{O'}M} = \alpha^i \vec{e}_i + x'^j \vec{e'}_j = \alpha^i \vec{e}_i + x'^j A^i_j \vec{e}_i = \left(\alpha^i + A^i_j x'^j\right) \vec{e}_i \tag{9.1.103}$$

---

[8] As we have seen in the section of Tensor Calculus the contravariant and covariant components are the same in an orthonormal base.

Identifying the result relatively to the vector $\vec{e}_i$ in the expression of $\overrightarrow{\mathrm{O}M}$, we have:

$$\boxed{x^i=\alpha^i+A^i_jx'^j} \tag{9.1.104}$$

And in an analogue way:

$$\boxed{x'^j=\alpha'^j+A'^j_ix'^i} \tag{9.1.105}$$

These two relations are more than useful in physics where we often have to consider a base in a given reference frame (so we can express the position of a point from either one or the other using these relations).

Let us now consider a pre-Euclidean point space as well as M and M' two points of this space. We have proved in our study of Topology (see section Topology page 2450) that the norm of the vector $\overrightarrow{MM'}$ is a possible measure of the distance between M and M' respecting some interesting and useful axioms. So we have if we denote by y_i the covariant components of $\overrightarrow{MM'}$:

$$\overline{MM'}=||\overrightarrow{MM'}||=\sqrt{\sum_i y_i^2} \tag{9.1.106}$$

If the both points M and M' have respectively for contravariant coordinates x^i and x'^i with respect to a base $(\vec{e}_i)$, we know following our definition above in the case of ordered vectors that:

$$\overrightarrow{MM'}=(-x^i+x'^i)\vec{e}_i \tag{9.1.107}$$

The squared norm is given as we have seen in our study of Tensor Calculus (see section Tensor Calculus page 2051) by the more general explicit relation:

$$||\overrightarrow{MM'}||^2=g_{ij}(x'^i-x^i)(x'^j-x^j)=g_{ij}\Delta x^i\Delta x^j \tag{9.1.108}$$

If we denote by $\mathrm{d}s$ the distance between the points M and M'. The above relation provides the expression of the square of the distance between these two points in the form:

$$\boxed{\mathrm{d}s^2=g_{ij}\mathrm{d}x^i\mathrm{d}x^j} \tag{9.1.109}$$

Let us also recall (see section Tensor Calculus page 2051) that for a pre-Euclidean point space where the basis vectors are orthonormal, we have:

$$g_{ij}=\delta_{ij} \tag{9.1.110}$$

and the expression of the distance becomes:

$$\mathrm{d}s^2=\mathrm{d}x^i\mathrm{d}x^j=\mathrm{d}x_i\mathrm{d}x_j \tag{9.1.111}$$

We thus get wan expression that generalizes to n dimensions, the square of the elementary distance, relatively to an orthogonal Cartesian coordinate system in the space of classical geometry (Euclidean).

The vectors in physics are generally functions of one or more variables, they may be variables of space or time. When to a point M of a point space $\mathcal{E}^n$, we attach a tensor, defined by its components relatively to the base $(\vec{e}_i)$, we say that we have given ourselves a "tensor field" (the tensor fields of order 1 being the vector fields as we have seen it in the section of Tensor Calculus).

For n-dimensional vectors, the concept of derivative from a three-dimensional vector is generalized and we get all the classic relations relatively to derivatives[9].

Let us consider a vector $\vec{x}$ belonging to a pre-Euclidean space $\mathcal{E}^2$ whose components on a base $(\vec{e}_i)$ are functions of some parameter α (typically the time). We will denote this vector $\vec{x}(\alpha)$ and we will have:

$$\vec{x}(\alpha)=x^i(\alpha)\vec{e}_i \tag{9.1.112}$$

By definition, the derivative from the vector $\vec{x}$ relatively to the component α is a vector denoted:

$$\vec{x}\,'(\alpha) \tag{9.1.113}$$

following the notation of most mathematicians. Or:

$$\dot{\vec{x}}(\alpha) \tag{9.1.114}$$

following the tradition of some physicists. Or even:

$$\mathrm{d}_\alpha\vec{x}(\alpha) \tag{9.1.115}$$

that we see most of time with engineers. Or more explicitly if we respect the pure mathematical notations:

$$\dfrac{\mathrm{d}\vec{x}}{\mathrm{d}\alpha} \tag{9.1.116}$$

In this textbook, we switch to a notation to another without prior notice depending on the desire and obviously to simplify the notation (the reader will have to do with it).

Since we are doing more mathematics than physics, let us write just now explicitly:

$$\vec{x}\,'(\alpha)=\dfrac{\mathrm{d}x^i(\alpha)}{\mathrm{d}\alpha}\vec{e}_i \tag{9.1.117}$$

Recalling (see section Differential and Integral Calculus page 1464) that the differential is given by:

$$\mathrm{d}\vec{x}=\vec{x}\,'(\alpha)\mathrm{d}\alpha \tag{9.1.118}$$

The different expressions of the derivatives of three-dimensional vectors relatively to the sum of vectors,dot product of two vectors are easily transferable to the case of n-dimensional vectors.

---

[9] A vector derivative - like the gradient, divergence, laplacian, etc. - is a derivative taken with respect to a vector field. Vector derivatives are extremely important in physics, where they arise throughout fluid mechanics, electricity and magnetism, elasticity, and many other areas of theoretical and applied physics

If a vector $\vec{x}$ of $\mathcal{E}^n$ depends on several independent parameters, α, β, γ, the partial derivative of $\vec{x}(\alpha,\beta,\gamma)$ vector with respect to the variable α, for example, is a vector denoted:

$$\frac{\partial\vec{x}}{\partial\alpha} \qquad \text{or} \qquad \partial_\alpha\vec{x} \tag{9.1.119}$$

whose components are the partial derivatives of the components of $\vec{x}$, namely:

$$\frac{\partial\vec{x}}{\partial\alpha}=\frac{\partial x^i}{\partial\alpha}\vec{e}_i \tag{9.1.120}$$

The total differential of the vector $\vec{x}(\alpha,\beta,\gamma)$ being written (see section of Differential and Integral Calculus page 1466):

$$\mathrm{d}\vec{x}=\frac{\partial\vec{x}}{\partial\alpha}\mathrm{d}\alpha+\frac{\partial\vec{x}}{\partial\beta}\mathrm{d}\beta+\frac{\partial\vec{x}}{\partial\gamma}\mathrm{d}\gamma \tag{9.1.121}$$

In a bases $(\vec{e}_i)$ we will consider as obvious that if a vector $\vec{x}$ is dependent of a parameter α, that its associated point M such that $\vec{x}=\overrightarrow{\mathrm{O}M}$ is also dependent of α. So if $\vec{x}$ admits a derivative, it is also same of $\overrightarrow{\mathrm{O}M}$!

Theorem 9.1. The derivative $\vec{x}\,'(\alpha)$ of a vector $\vec{x}$ does not depend on the (static) origin point O of the base but only of the considered point M.

Proof 9.1.1. Indeed, if O′ is another point of origin, we have:

$$\overrightarrow{\mathrm{O}M}=\overrightarrow{\mathrm{OO}'}+\overrightarrow{\mathrm{O}'M} \tag{9.1.122}$$

and since the vector $\overrightarrow{\mathrm{OO}'}$ is fixed and does not depend on α, we have:

$$\frac{\mathrm{d}\overrightarrow{\mathrm{OO}'}}{\mathrm{d}\alpha}=\vec{0} \tag{9.1.123}$$

hence:

$$\boxed{\frac{\mathrm{d}\overrightarrow{\mathrm{O}M}}{\mathrm{d}\alpha}=\frac{\mathrm{d}\overrightarrow{\mathrm{O}'M}}{\mathrm{d}\alpha}=\vec{x}'(\alpha)} \tag{9.1.124}$$

■ Q.E.D.

We can write the derivative of the vector $\overrightarrow{\mathrm{O}M}$ by only mentioning the point M as it is the usage in physics and therefore write:

$$\vec{x}\,'(\alpha)=\mathrm{d}_\alpha\vec{x}=\frac{\overrightarrow{\mathrm{O}'M}}{\mathrm{d}\alpha}=\frac{\vec{M}}{\mathrm{d}\alpha}=\vec{M}'(\alpha)=\mathrm{d}_\alpha\vec{M} \tag{9.1.125}$$

The differential of $\overrightarrow{\mathrm{O}M}$ is then written:

$$\mathrm{d}\vec{M}=\vec{M}\,'(\alpha)\mathrm{d}\alpha \tag{9.1.126}$$

If a point M of $\mathcal{E}^n$ is associated, relatively to base $(\vec{e}_i)$, to a vector $\vec{x}(\alpha,\beta,\gamma)=\overrightarrow{\mathrm{O}M}$, the partial derivatives of $\overrightarrow{\mathrm{O}M}$ will depend only of the point M and we will write, for example:

$$\frac{\partial \vec{x}}{\partial \alpha}=\frac{\partial \vec{M}}{\partial \alpha} \tag{9.1.127}$$

To simplify the expressions of total partial derivatives of the functions dependent of n independent variables we use, when the context permits, the index-based notations as seen in the section of Tensor Calculus. Therefore:

$$f(x,y,z,\ldots,t)=f(x^1,x^2,\ldots,x^n) \tag{9.1.128}$$

for a function f dependent of n independent variables x^n (this includes the time in physics). There partial derivatives are written into the form:

$$\partial_i f=\frac{\partial f}{\partial y^i} \tag{9.1.129}$$

The second partial derivatives with respect to the variables y^i and y^k will be written:

$$\partial_{ik} f=\frac{\partial f}{\partial y^i \partial y^k} \tag{9.1.130}$$

When $\vec{x}$ is a vector such that $\vec{x}=x^i\vec{e}_i$ whose components are dependent of n independent variables y^k such that:

$$x^i=x^i(y^1,y^2,\ldots,y^n) \tag{9.1.131}$$

the partial derivatives of the vector will then be written, using the summation convention:

$$\frac{\partial \vec{x}}{\partial y^k}=\partial_k\vec{x}=\partial_k\left(x^i\vec{e}_i\right)=(\partial_k x^i)\vec{e}_i \tag{9.1.132}$$

The concept of space-time being now introduced, we can now proceed to study the Lagrangian formalism and determining the mathematical formulation of the principle of least action (see next section).

~95 %

★★★☆☆

71 votes, 67.32%

9.2. Analytical Mechanics

ANALYTICAL MECHANICS is a reformulation of Classical Mechanics that we due the present form, also named "Lagrangian mechanics", from the work of Bernoulli brothers and especially Euler and Lagrange. It is indeed in 11696 (holocene calendar) that began the story of the true theoretical physics.

In fact, the analytical mechanics starts from the event from the following observation (in the 117th century according to holocene calendar): Any conservative system seems to move from one state to another still using the simplest means and maintaining a constant magnitude between the two states.

Remarks

R1. The above means can be: the shortest path, the fastest path (that is to say the spatio-temporal paths with lower amplitudes...).

R2. According to the first fundamental principle of physics, the constant quantity is selected as being the total energy.

This statement is named in the context of the Mechanics the "principle of least action (of Maupertuis)" or as part of general physics "variational principle" or sometimes as part of optics "principle of economy" or "Fermat's principle". In the purely mathematical abstraction framework we speak of "Hamilton's principle".

More technically, it is also formulated as follows: A system moves from one configuration to another so that the variation of the action (see below for details) between the natural path actually followed and any infinitely close and virtual trajectory having the same starts and ends in space and time is zero.

In fact, although this statement may seem as consistent, it raises doubts ... but we will see that:

1. In Classical Mechanics we can prove the Newton's first law by admitting this principle as true and by superimposing the principle of conservation of energy[10] and we can therefore also explain the wobbling motion of almost any simple solid.
2. In Electromagnetism, we will find back all the Maxwell equations (verbatim the Biot-Savart law, the Faraday law, the Lorentz force, the Laplace force, etc.) from the properties of the principle of least action and energy conservation.
3. In Optics, we will prove that the path followed by the light is always the shortest and this will allow us to prove Fermat's principle underlying any geometrical optics. Also in

---

[10] It is strongly advised that the reader remembers the situation where this principle doesn't apply as seen during our study of Noether's theorem (see page 3136)!

General Relativity this same assumption will take us to prove the deflection of light near massive bodies.

4. In Atomic Physics, the properties of the principle of least action will allow us to determine some mathematical properties of atoms and other particles (fermions and bosons in quantum field theory).

5. The principle of least action will also enable us to prove that all bodies, with or without mass, is deflected by a field of acceleration and ... so will give us the opportunity to determine the Einstein Field Equations that are the basis of the section on General Relativity (see page 5720).

6. This principle also applies to get powerful results in geometry as we shall see a little further. Thus, the analytical mechanics techniques are very intimately related to pure mathematics.

7. We can merge all of our actual knowledge of electromagnetism, quantum physics and gravitation in a single formulation (obviously long ...) using the Lagrangian formalism.

It needless to say that through this small examples the huge applications of this principle!!!

Historically, it is interesting to know that this is Pierre-Louis Moreau de Maupertuis who first articulated the principle of least action in shortly scientific form. The intervention of Euler and Lagrange in this area has been to put this principle into a mathematical form and to prove that it follows from a simple mathematical property of optima of continuous functions. It is needless to say that this latter fact that allowed to re-prove all the laws of classical physics from only pure mathematical concepts has upset at this time some people... (and still do today!).

This principle was (and always has) a significant impact and the problem was to apply the mathematical expression of the latter to all physical phenomena that had already been experimentally demonstrated and empirically at that time. Perform this proof was similar to explain why such phenomenon or such law was thus rather than otherwise. Imagine!

For example given locality, Lorentz invariance (see section Special Relativity page 5639) and know physical data since 11860 (holocene calendar) it can be shown that the Lagrangian describing all observed physical process (without gravity) can be written:

$$
\begin{gathered}
-\tfrac{1}{2}\partial_\nu g^a_\mu \partial_\nu g^a_\mu - g_s f^{abc}\partial_\mu g^a_\nu g^b_\mu g^c_\nu - \tfrac{1}{4}g_s^2 f^{abc}f^{ade}g^b_\mu g^c_\nu g^d_\mu g^e_\nu + \\
\tfrac{1}{2}ig_s^2(\bar{q}^\sigma_i\gamma^\mu q^\sigma_j)g^a_\mu + \bar{G}^a\partial^2 G^a + g_s f^{abc}\partial_\mu\bar{G}^a G^b g^c_\mu - \partial_\nu W^+_\mu\partial_\nu W^-_\mu - \\
M^2 W^+_\mu W^-_\mu - \tfrac12\partial_\nu Z^0_\mu\partial_\nu Z^0_\mu - \tfrac{1}{2c_w^2}M^2 Z^0_\mu Z^0_\mu - \tfrac12\partial_\mu A_\nu\partial_\mu A_\nu - \tfrac12\partial_\mu H\partial_\mu H - \\
\tfrac12 m_h^2 H^2 - \partial_\mu\phi^+\partial_\mu\phi^- - M^2\phi^+\phi^- - \tfrac12\partial_\mu\phi^0\partial_\mu\phi^0 - \tfrac{1}{2c_w^2}M\phi^0\phi^0 - \beta_h[\tfrac{2M^2}{g^2} + \\
\tfrac{2M}{g}H + \tfrac12(H^2+\phi^0\phi^0+2\phi^+\phi^-)] + \tfrac{2M^4}{g^2}\alpha_h - igc_w[\partial_\nu Z^0_\mu(W^+_\mu W^-_\nu - \\
W^+_\nu W^-_\mu) - Z^0_\nu(W^+_\mu\partial_\nu W^-_\mu - W^-_\mu\partial_\nu W^+_\mu) + Z^0_\mu(W^+_\nu\partial_\nu W^-_\mu - \\
W^-_\nu\partial_\nu W^+_\mu)] - igs_w[\partial_\nu A_\mu(W^+_\mu W^-_\nu - W^+_\nu W^-_\mu) - A_\nu(W^+_\mu\partial_\nu W^-_\mu - \\
W^-_\mu\partial_\nu W^+_\mu) + A_\mu(W^+_\nu\partial_\nu W^-_\mu - W^-_\nu\partial_\nu W^+_\mu)] - \tfrac12 g^2 W^+_\mu W^-_\mu W^+_\nu W^-_\nu + \\
\tfrac12 g^2 W^+_\mu W^-_\nu W^+_\mu W^-_\nu + g^2c_w^2(Z^0_\mu W^+_\mu Z^0_\nu W^-_\nu - Z^0_\mu Z^0_\mu W^+_\nu W^-_\nu) + \\
g^2 s_w^2(A_\mu W^+_\mu A_\nu W^-_\nu - A_\mu A_\mu W^+_\nu W^-_\nu) + g^2 s_w c_w[A_\mu Z^0_\nu(W^+_\mu W^-_\nu - \\
W^+_\nu W^-_\mu) - 2A_\mu Z^0_\mu W^+_\nu W^-_\nu] - g\alpha[H^3 + H\phi^0\phi^0 + 2H\phi^+\phi^-] - \\
\tfrac18 g^2\alpha_h[H^4+(\phi^0)^4+4(\phi^+\phi^-)^2+4(\phi^0)^2\phi^+\phi^-+4H^2\phi^+\phi^-+2(\phi^0)^2H^2] - \\
gMW^+_\mu W^-_\mu H - \tfrac12 g\tfrac{M}{c_w^2}Z^0_\mu Z^0_\mu H - \tfrac12 ig[W^+_\mu(\phi^0\partial_\mu\phi^- - \phi^-\partial_\mu\phi^0) - \\
W^-_\mu(\phi^0\partial_\mu\phi^+ - \phi^+\partial_\mu\phi^0)] + \tfrac12 g[W^+_\mu(H\partial_\mu\phi^- - \phi^-\partial_\mu H) - W^-_\mu(H\partial_\mu\phi^+ - \\
\phi^+\partial_\mu H)] + \tfrac12 g\tfrac{1}{c_w}(Z^0_\mu(H\partial_\mu\phi^0 - \phi^0\partial_\mu H) - ig\tfrac{s_w^2}{c_w}MZ^0_\mu(W^+_\mu\phi^- - W^-_\mu\phi^+) + \\
igs_wMA_\mu(W^+_\mu\phi^- - W^-_\mu\phi^+) - ig\tfrac{1-2c_w^2}{2c_w}Z^0_\mu(\phi^+\partial_\mu\phi^- - \phi^-\partial_\mu\phi^+) + \\
igs_wA_\mu(\phi^+\partial_\mu\phi^- - \phi^-\partial_\mu\phi^+) - \tfrac14 g^2W^+_\mu W^-_\mu[H^2+(\phi^0)^2+2\phi^+\phi^-] - \\
\tfrac14 g^2\tfrac{1}{c_w^2}Z^0_\mu Z^0_\mu[H^2+(\phi^0)^2+2(2s_w^2-1)^2\phi^+\phi^-] - \tfrac12 g^2\tfrac{s_w^2}{c_w}Z^0_\mu\phi^0(W^+_\mu\phi^- + \\
W^-_\mu\phi^+) - \tfrac12 ig^2\tfrac{s_w^2}{c_w}Z^0_\mu H(W^+_\mu\phi^- - W^-_\mu\phi^+) + \tfrac12 g^2 s_w A_\mu\phi^0(W^+_\mu\phi^- + \\
W^-_\mu\phi^+) + \tfrac12 ig^2 s_w A_\mu H(W^+_\mu\phi^- - W^-_\mu\phi^+) - g^2\tfrac{s_w}{c_w}(2c_w^2-1)Z^0_\mu A_\mu\phi^+\phi^- - \\
g^1 s_w^2 A_\mu A_\mu\phi^+\phi^- - \bar{e}^\lambda(\gamma\partial+m_e^\lambda)e^\lambda - \bar\nu^\lambda\gamma\partial\nu^\lambda - \bar{u}^\lambda_j(\gamma\partial+m_u^\lambda)u^\lambda_j - \\
\bar{d}^\lambda_j(\gamma\partial+m_d^\lambda)d^\lambda_j + igs_wA_\mu[-(\bar{e}^\lambda\gamma^\mu e^\lambda) + \tfrac23(\bar u^\lambda_j\gamma^\mu u^\lambda_j) - \tfrac13(\bar d^\lambda_j\gamma^\mu d^\lambda_j)] + \\
\tfrac{ig}{4c_w}Z^0_\mu[(\bar\nu^\lambda\gamma^\mu(1+\gamma^5)\nu^\lambda) + (\bar e^\lambda\gamma^\mu(4s_w^2-1-\gamma^5)e^\lambda) + (\bar u^\lambda_j\gamma^\mu(\tfrac43 s_w^2 - \\
1-\gamma^5)u^\lambda_j) + (\bar d^\lambda_j\gamma^\mu(1-\tfrac83 s_w^2-\gamma^5)d^\lambda_j)] + \tfrac{ig}{2\sqrt2}W^+_\mu[(\bar\nu^\lambda\gamma^\mu(1+\gamma^5)e^\lambda) + \\
(\bar u^\lambda_j\gamma^\mu(1+\gamma^5)C_{\lambda\kappa}d^\kappa_j)] + \tfrac{ig}{2\sqrt2}W^-_\mu[(\bar e^\lambda\gamma^\mu(1+\gamma^5)\nu^\lambda) + (\bar d^\kappa_j C^\dagger_{\lambda\kappa}\gamma^\mu(1+ \\
\gamma^5)u^\lambda_j)] + \tfrac{ig}{2\sqrt2}\tfrac{m_e^\lambda}{M}[-\phi^+(\bar\nu^\lambda(1-\gamma^5)e^\lambda) + \phi^-(\bar e^\lambda(1+\gamma^5)\nu^\lambda)] - \\
\tfrac{g}{2}\tfrac{m_e^\lambda}{M}[H(\bar e^\lambda e^\lambda) + i\phi^0(\bar e^\lambda\gamma^5 e^\lambda)] + \tfrac{ig}{2M\sqrt2}\phi^+[-m_d^\kappa(\bar u^\lambda_j C_{\lambda\kappa}(1-\gamma^5)d^\kappa_j) + \\
m_u^\lambda(\bar u^\lambda_j C_{\lambda\kappa}(1+\gamma^5)d^\kappa_j] + \tfrac{ig}{2M\sqrt2}\phi^-[m_d^\lambda(\bar d^\lambda_j C^\dagger_{\lambda\kappa}(1+\gamma^5)u^\kappa_j) - m_u^\kappa(\bar d^\lambda_j C^\dagger_{\lambda\kappa}(1- \\
\gamma^5)u^\kappa_j] - \tfrac g2\tfrac{m_u^\lambda}{M}H(\bar u^\lambda_j u^\lambda_j) - \tfrac g2\tfrac{m_d^\lambda}{M}H(\bar d^\lambda_j d^\lambda_j) + \tfrac{ig}{2}\tfrac{m_u^\lambda}{M}\phi^0(\bar u^\lambda_j\gamma^5 u^\lambda_j) - \\
\tfrac{ig}{2}\tfrac{m_d^\lambda}{M}\phi^0(\bar d^\lambda_j\gamma^5 d^\lambda_j) + \bar X^+(\partial^2-M^2)X^+ + \bar X^-(\partial^2-M^2)X^- + \bar X^0(\partial^2 - \\
\tfrac{M^2}{c_w^2})X^0 + \bar Y\partial^2 Y + igc_wW^+_\mu(\partial_\mu\bar X^0X^- - \partial_\mu\bar X^+X^0) + igs_wW^+_\mu(\partial_\mu\bar YX^- - \\
\partial_\mu\bar X^+Y) + igc_wW^-_\mu(\partial_\mu\bar X^-X^0 - \partial_\mu\bar X^0X^+) + igs_wW^-_\mu(\partial_\mu\bar X^-Y - \\
\partial_\mu\bar YX^+) + igc_wZ^0_\mu(\partial_\mu\bar X^+X^+ - \partial_\mu\bar X^-X^-) + igs_wA_\mu(\partial_\mu\bar X^+X^+ - \\
\partial_\mu\bar X^-X^-) - \tfrac12 gM[\bar X^+X^+H + \bar X^-X^-H + \tfrac{1}{c_w^2}\bar X^0X^0H] + \\
\tfrac{1-2c_w^2}{2c_w}igM[\bar X^+X^0\phi^+ - \bar X^-X^0\phi^-] + \tfrac{1}{2c_w}igM[\bar X^0X^-\phi^+ - \bar X^0X^+\phi^-] + \\
igMs_w[\bar X^0X^-\phi^+ - \bar X^0X^+\phi^-] + \tfrac12 igM[\bar X^+X^+\phi^0 - \bar X^-X^-\phi^0]
\end{gathered}
$$

Thus, as far as we know, it seems that the first to tackle the problem was the Bâlois (Switzerland) Leonhard Euler. But we have also kept the name of Lagrange (hence the expression "Lagrangian formalism") to define the entire method and the mathematical formalism built around the principle of least action.

In fact we will see in the section the differences between:

1. Newtonian Mechanics (or Newtonian Formalism)
2. Lagrangian Mechanics (or Lagrangian Formalism)
3. Hamiltonian Mechanics (or Hamiltonian Formalism)

and we will see that 1 is a special case of 3 which is a special case of 2. We will also prove with in some context (as quantum physics) the Hamiltonian is sometimes preferred to the Lagrangian and vice versa.

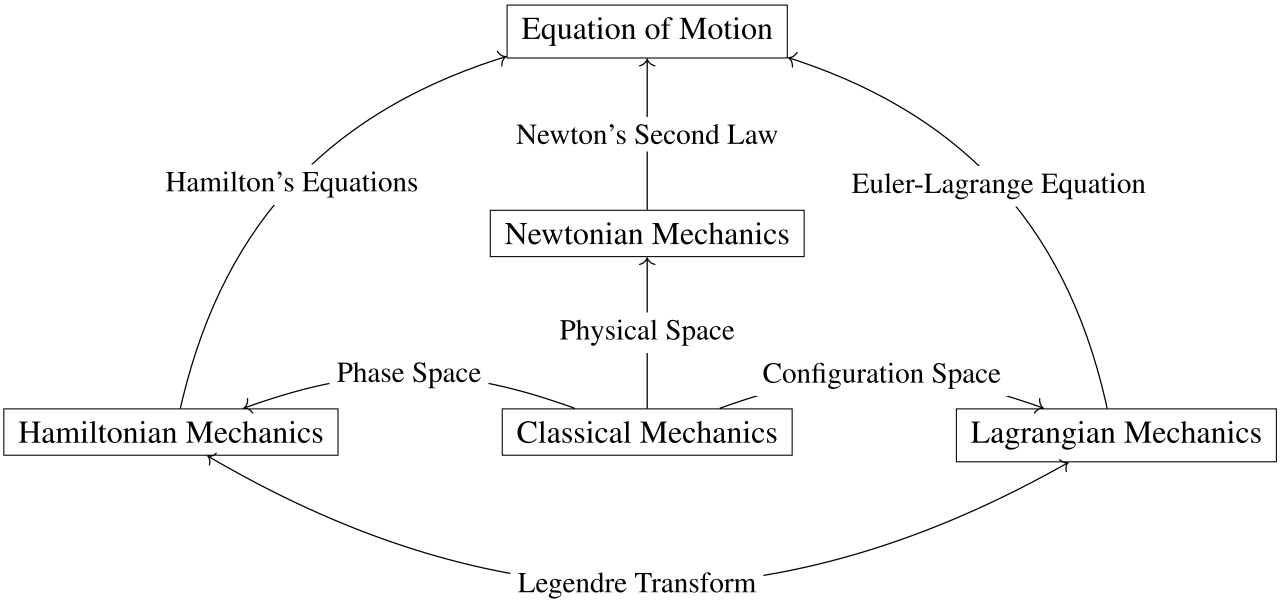

Equation of Motion
Newton's Second Law
Hamilton's Equations
Euler-Lagrange Equation
Newtonian Mechanics
Physical Space
Phase Space
Configuration Space
Hamiltonian Mechanics
Classical Mechanics
Lagrangian Mechanics
Legendre Transform

Figure 9.9: Landscape of Analytical and Classical Mechanics

9.2.1 Lagrangian formalism

Classical Mechanics can be formulated in different ways. The most common is the formulation of Newton, which uses the concept of force (see section Classical Mechanics page 3199). It is by far the easiest when it comes to consider a specific issue and that's why this is the one that is taught in small classes. But to treat more complex problems or more precisely and to be able to do rigorous proofs the Newton formulation is not the most practical one!

Analytical Mechanics, initiated in the 118th century (holocene calendar), regroups therefore very different mathematized formulations of Classical Mechanics, including Hamiltonian and Lagrangian mechanics (all these formulations are equivalent!).

This formulation is taught only a little bit in small schools because we must admit the Lagrangian and Hamiltonian formalism (therefore containing the principle of least action in mathematical form) uses an abstraction level a bit higher than the normal methods and despite being often of great help in the development of theories (basic physics, Quantum Physics, General Relativity, Quantum Field Theory, Superstring Theory), it rarely follows new discovers from it (but rather useful and powerful reduction and validation method).

Let's begin our work:

9.2.1.1 Generalized coordinates and frames

A natural reflex usually leads to refer to the position of a point in space with the only knowledge of its three Cartesian coordinates x, y, z (or polar, cylindrical, spherical, etc.). This attitude is also often justified by the simplicity of a lot of situations encountered in practice where it is not necessary to search for more elaborate methods or to go in other coordinate systems (see section Vector Calculus page 1846).

To locate the position of a mobile (or a material point) in physics it is necessary at first to associate a frame to the repository. Thus, a "frame" is a system (concrete physical) of location (or "reference system") in the space associated with the repository.

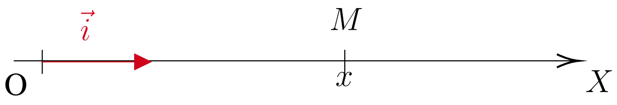
✍**Examples:**

E1. In astronomy, for many problems, the Earth is taken as repository, but on the center of it we put the reference frame (most of time in spherical coordinates).

E2. In mechanics a body is taken as a repository but it is on its inertial center is that we put the reference frame.

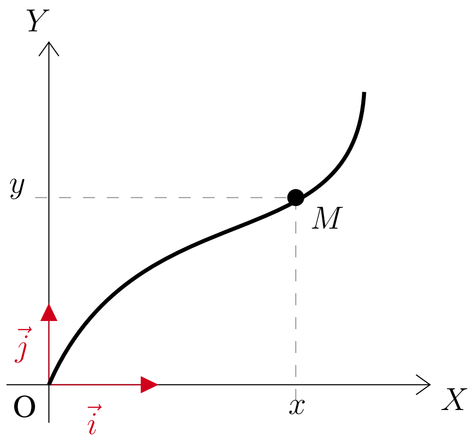
Conventional frames in Classical Mechanics are mostly canonical oriented pre-Euclidean spaces basis (see section Vector Calculus page 2051) and where each point or each vector of the space, can be represented algebraically by its affixes values (the value in the ordinate corresponding to the projection on the vertical axis) and its abscissas values (corresponding to the projection on the horizontal axis).

Here are some trivial examples:



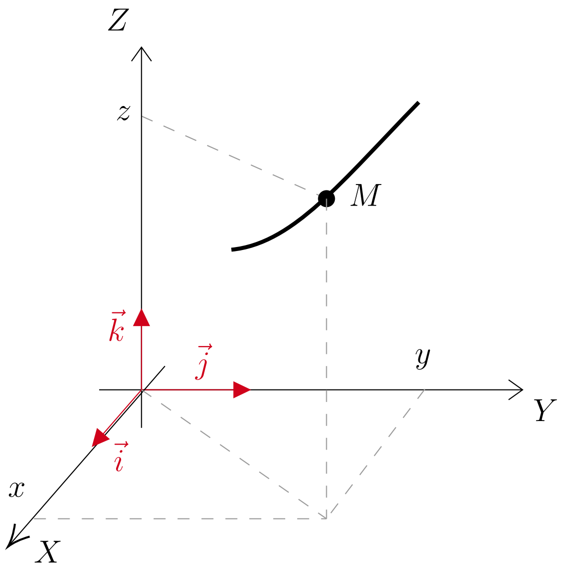
One dimensional reference frame

The motion of M is done along a straight line



Two dimensional reference frame
(Argand-Cauchy plane)

The motion of M is done in the plane or in a subset of it



Three dimensional reference frame

The motion of M is done in the space or in a subset of it

Figure 9.10: **Movements in referential frames from to 1, 2 or 3 dimensions**

Definitions (#9.2.422):

D1. A repository, assimilated to a reference frame, is said to be a "Galilean reference frame" (it's rare that we make explicit mention of in physics for lack of rigour) if:

- We can consider it as motionless during the study of the movement of the system or to be in an uniform rectilinear translation relative to another repository itself motionless (this is also named an "inertial reference frame" because not accelerated).

  So if we neglect the rotational movement of the Sun (actually a G2V-type star, aka yellow dwarf) around the center of the galaxy, then the heliocentric reference frame may be regarded as a Galilean reference frame. If we neglect the rotational movement of the Earth around the Sun (actually a G2V-type star, aka yellow dwarf), then the geocentric reference may be regarded as a Galilean reference frame. If we neglect the rotational motion of the Earth itself, then the terrestrial reference frame can be considered as Galilean. In many mechanical experiments on the surface of the Earth, we find that the terrestrial reference frame can be considered as Galilean with very good accuracy for short time experiments. Fortunately there are still a lot of phenomena which must take into account the rotation of the Earth (deviation to the East, Foucault pendulum, etc.)

- We can consider it as a system in which Newton's laws are satisfied (see section Classical Mechanics page 3199).

D2. A repository, assimilated to a reference frame, is named "barycentric reference frame" (see section Euclidean Geometry page 2635) if it has for origin the center of mass (see section Classical Mechanics page 3213) of the body being studied.

Therefore, the "Copernicus reference frame" is assimilated to the center of gravity (inertia) of our Solar System, the "heliocentric reference frame" also named "Kepler's reference frame" to the Sun center of mass.

D3. A repository, assimilated to a reference frame, is named "geocentric reference frame" when we make reference to a system of axes placed at the center of mass of the Earth. The axes, parallel to those of a Copernicus reference frame, point in the direction of to three fixed stars. In this reference the Earth rotates on itself in 24h.

D4. A repository, assimilated to a reference frame, is named "Terrestrial reference frame" when we make reference to a system of axes placed at the center of mass of the Earth that has a uniform rotational movement corresponding to the speed of rotation of the Earth. Traditionally one of the axes is directed to the pole star. This is the reference frame that we refer to as in real life (it is not strictly speaking a Galilean reference frame)! This will induce specific effects on the movements in the atmosphere such as we experience them.

Remarks

R1. In the 2-dimensions case, say that an orthonormal basis $(\mathrm{O},\vec{i},\vec{j})$ is a "direct reference frame" means that the oriented angle of $(\vec{i},\vec{j})$ has for main value $\pi/2$ (clockwise). Say that an orthonormal basis $(\mathrm{O},\vec{i},\vec{j})$ is an "indirect reference frame" means that the angle has for main value $-\pi/2$. In all what will follow, if we do not specify the orientation, this suggests that $(\mathrm{O},\vec{i},\vec{j})$ is a direct reference frame.

R2. Obviously, there are a number of arbitrary choices to make when we set up a reference frame: for example, where to place the origin of the spatial coordinates, how to orient the axes, and where to put the origin of time. But this freedom is not a real issue, indeed if we make a choice, there are mathematical techniques (see section Linear Algebra page 1992) that can be used to move from an old system (basis) to a new one.

It is true that the three parameters x, y, z are perfectly sufficient to identify a material point in ordinary space as we have already mentioned it in our study punctual spaces (see section Principles page 3145), but nevertheless it is sometimes inevitable, or just more advantageous to use a larger number of parameters than three. We can obviously consider all sorts of settings to achieve the coordinates of a point in space, so that in a more generalized manner we shall have to take into account relation of the type (we do not keep the same writing that the one we had in our study punctual spaces for consistency with the many existing references on the subject):

$$\begin{aligned} x&=x^1=x^1(q^1,q^2,\ldots,q^n)\\ y&=x^2=x^2(q^1,q^2,\ldots,q^n)\\ z&=x^3=x^3(q^1,q^2,\ldots,q^n) \end{aligned} \tag{9.2.1}$$

The parameters $q^1, q^2, \ldots, q^n$ are named "generalized coordinates" and are the parameters to which a problem will most often referred. Knowing their expression over time is a fundamental problem of dynamics (think for example a satellite which coordinates x, y, z are dependant on spherical components φ, θ, r that are themselves depending of the time t). This means that we will have reach a solution when we will have the independent relation:

$$\begin{aligned} q^1 &= q^1(t) \\ q^2 &= q^2(t) \\ &\ldots \\ q^n &= q^n(t) \end{aligned} \tag{9.2.2}$$

Geometrically these components can be lengths along straight lines, or arc lengths along curves, or angles; not necessarily Cartesian coordinates or other standard orthogonal coordinates. There is one for each degree of freedom, so the number of generalized coordinates equals the number of "degrees of freedom", n. In physics a "degree of freedom" corresponds to a quantity that changes the configuration of the system, for example the angle of a pendulum, or the arc length traversed by a bead along a wire, etc.

It is therefore important to notice that the number of parameters q^i defining locating a point in space is at least equal to three in most common cases. It is ultimately the nature of the envisaged situations that suggest the choice of the number of parameters to be used (cartesian coordinates, cylindrical, spherical, ...).

The vector:

$$\vec{q}(t) = (q^1(t), q^2(t), \ldots, q^n(t)) \tag{9.2.3}$$

is a point in the "configuration space" of the system.

Indeed, it is quite natural mathematically speaking to associate the manipulation of the n generalized coordinates q^i to a n-dimensional space, in which the q^i would appear as the coordinates of a point P representative of the configuration of any system. This is why we name this space the "configuration space" and is denoted $\mathcal{C}^n$.

But the rigour of mathematical-physical, brings us to provide a more accurate description of the phenomena by adding the important variable that is the time, often seen as independent variable in the q^i. We thus arrive inevitably to use another configuration space $\mathcal{C}^{n+1}$ that we name "event space".

This configuration space is of great interest to many problems of modern science and is particularly well suited to the relativistic problems. The independent variables constituting the spatial and temporal coordinates then form what we name the "Euler variables" (as we will see later, they have this name because we can find them in the Euler-Lagrange equation).

The components $x^1, x^2, \ldots, x^n$, that can be seen as functions as we have just see, will be assumed now defined, continuous, of class $\mathcal{C}^2$ (to work with the acceleration) relatively to the parameters $q^1, q^2, \ldots, q^n$ and will lead to a Jacobian different from zero (see section Differential and Integral Calculus page 1524).

Definitions (#9.2.423):

D1. A "holonomic constraint" is a constraint equation of the following shape for a particle:

$$f(x^1,x^2,x^3,t)=f(\vec{x},t)=0 \tag{9.2.4}$$

which connects all the 3 spatial coordinates of that particle together, so they are not independent and founding such a function can help to solve the study of a dynamic system.

D2. A mechanical system can involve constraints on both the generalized coordinates and their derivatives. Constraints of this type are known as "non-holonomic constraint" and are of the form:

$$f(q^1,\ldots,q^n,\partial_t q^1,\ldots,\partial_t q^n,t)=f(\vec{q},\dot{\vec{q}},t)=0 \tag{9.2.5}$$

To study continuum medium (radically different concept from a material point), we will have however two different approaches:

1. The Lagrange method or Hilbert Method: we seek to characterize the movement of the medium described by a Lagrangian formulation consisting in characterizing it by a system of equations in the Newtonian sense. For derivatives, we the have the speed and acceleration of the medium.
2. The Euler method: Instead of following the path of a point, we focus on the evolution of a physical characteristics at a given point as the speed, temperature acceleration, pressure or other. We speak so often of "Eulerian system".

9.2.1.2 Variational Principle

A "variational principle" is a scientific principle used within the calculus of variations, which develops general methods for finding functions which extremize the value of quantities that depend upon those functions. For example, to answer this question: « *What is the shape of a chain suspended at both ends?* » we can use the variational principle that the shape must minimize the gravitational potential energy.

The "variational principle" is thus only the contemporary mathematical form of the principle of least action, which, as we have already mentioned, is at the base of the Lagrangian formalism and of all physics!

Let us recall that according to the statement of the variational principle we must find in any physical phenomenon, an amount that is naturally optimized (minimized or maximized) and describing all the variables of the system studied and thus its outcome.

Here are the steps that we will follow and once this approach presented, we will address its mathematical formalization.

The proposals are the following:

P1. We assume the variational principle and the principle of conservation of energy as correct.

P2. The total energy of a closed system is constant and consists only of the summation of the kinetic energy and potential energy.

If we consider only kinetic energy, then the system is named "free system"; if both energies are considered, then we say that the system is a "generalized system".

P3. We define a mathematical function (whose variables are the generalized coordinates) named "Lagrangian" which is given by the difference of the two previous mentioned energies.

P4. On the evolution of a system between two states, we seek for the properties of the function (Lagrangian) that minimize the variation of the difference of the two energies on temporal or metric evolution of the system.

So to put it in a mathematical form, we first ask that there is a real function of $2n$ variables (the n general coordinates and the corresponding n derivatives to the implicit variable that will typically be in physics the time t):

$$(q_1,\ldots,q_n,\dot{q}_1,\ldots,\dot{q}_n)\mapsto L(q_1,\ldots,q_n,\dot{q}_1,\ldots,\dot{q}_n) \tag{9.2.6}$$

that we will name "generalized Lagrangian" of the system, whose integral satisfies the following statement:

In a natural movement starting from a point $A(q_1^A,\ldots,q_n^A)$ at the time t_A, arriving at the point $B(q_1^A,\ldots,q_n^A)$ at the time t_B, the following integral (that is a pillar of modern physics!!!!!) is named "action integral" or simply "action":

$$S(q,\dot{q})=\int_{t_A}^{t_B} L[q_1(t),\ldots,q_n(t),\dot{q}_1(t),\ldots,\dot{q}_n(t)]\mathrm{d}t \tag{9.2.7}$$

which may also be denoted more shortly:

$$S(q,\dot{q})=\int_{t_A}^{t_B} L(q,\dot{q})\mathrm{d}t \tag{9.2.8}$$

must be an extremal value (in fact, a "minimum" or "maximum"), since we could as well take $-L$ instead of $+L$ in the choice of the definition of the generalized Lagrangian).

This "action" or "Lagrangian's action" S is what we commonly name in physics a "functional" and has units of energy multiplied by time since L is energy, or said otherwise: mass × velocity × distance, or formally $S_0=m\cdot v\cdot\ell$.

It was Maupertuis that proposed this new metric based on its comprehensions of shortest time problem in Mechanics (leading to the brachistrone) and in Optics (leading to Snell's law).

The faster it goes, the greater the action. And if it's a particle, then the more massive it is, the greater the action. If there are multiple segments to the journey, then the total action is just the sum of the mass times velocity times distance for each segment:

$$S_0=\Sigma m\cdot v\cdot\ell \tag{9.2.9}$$

In 11744 (holocene calendar), Maupertuis wrote: «*This action is the true expense of Nature, which she manages to make as small as possible*».

Maupertuis was attacked and ridiculed by some people with this idea coming from nowhere.

But there was one man who vehemently defended it, and that man was Leonhard Euler. The first thing Euler did was he replaced the sum with an integral:

$$\sum mv\ell \rightarrow \int mv\mathrm{d}\ell \tag{9.2.10}$$

so we could calculate the action while speed or direction changed continuously. And he used this to find the path of a particle around a central mass, like the orbit of a planet around a star.

Solving this meant that out of every possible path between two points, Euler would have to find the one for which the action was the smallest. This is similar to the problem Fermat tried to solve, only now, instead of changing one variable, he would have to vary every possible point along the path, which is infinitely many. Needless to say, this was an arduous task.

Math had not yet developed the tools required to handle such problems. Fortunately, Euler himself invented a new method. It was clunky and time consuming, but it worked.

Through this process, he realized that the principle of least action only works if the total energy is conserved, and it is the same for all paths considered. These were two conditions that Maupertuis hadn't realized were necessary. So Euler improved the mathematical rigor of the principle.

But Euler was still far from a general proof. That would have to wait for another legendary mathematician, Joseph-Louis Lagrange.

Joseph-Louis Lagrange was a shy 19-year-old, mostly self-taught. But despite his age, he was already working at the forefront of mathematics, including with Euler's new method. In 11754 (holocene calendar), he shared his results with Euler, who replied that Lagrange had « *extolled the theory to the highest summit of perfection*». But besides being world class mathematicians, the two had another thing in common. They were both huge proponents of the principle of least action. And around five years later, just a year after Maupertuis' death, Lagrange succeeded in providing a general proof!

9.2.1.3 Euler-Lagrange Equation

The principle of least action therefore states that the (integral) S is extremal if:

$$t \mapsto [q_1, \ldots, q_n, \dot{q}_1, \ldots, \dot{q}_n] \tag{9.2.11}$$

is the natural trajectory actually followed by the physical system.

Let us consider then a very similar trajectory to the previous, that we will denote by:

$$(q_1 + \delta q_1, \ldots, q_n + \delta q_n, \dot{q}_1 + \delta \dot{q}_1, \ldots, \dot{q}_n + \delta \dot{q}_n) \tag{9.2.12}$$

That is to say that for every i we put:

$$\begin{aligned} q_i(t) &\rightarrow q_i(t)+\delta q_i(t) \\ \dot{q}_i(t) &\rightarrow \dot{q}_i(t)+\delta \dot{q}_i(t) \end{aligned} \tag{9.2.13}$$

with:

$$\delta q_i(t_A)=\delta q_i(t_B)=0 \tag{9.2.14}$$

to ensure that we always start from the same point A to reach the same point B. The perfect determination of the extreme points (characterized by variations of a path to another one identically zero) is a major assumption of the principle of least action!

Remark

Now we will omit writing the time arguments to facilitate the notations.

If $S(q,\dot{q})$ is indeed the evolution of a system evolving following the principle of least action, then the action given by the variation:

$$\delta S := S(q+\delta q,\dot{q}+\delta\dot{q})-S(q,\dot{q}) \tag{9.2.15}$$

is zero for δq and $\delta\dot{q}$ going toward zero (meaning that any physical system returns to its original state without external intervention).

Which brings us to write:

$$\delta S=\int\limits_{t_A}^{t_B} L(q+\delta q,\dot{q}+\delta\dot{q})\mathrm{d}t-\int\limits_{t_A}^{t_B} L(q,\dot{q})\mathrm{d}t=\int\limits_{t_A}^{t_B}[L(q+\delta q,\dot{q}+\delta\dot{q})-L(q,\dot{q})]\mathrm{d}t=0 \tag{9.2.16}$$

This allows us to justify the name of "variational principle" (also sometimes named the "principle of stationarity of action"):

$$\delta S=\delta\int\limits_{t_A}^{t_B} L\mathrm{d}t=\int\limits_{t_A}^{t_B}\delta L\mathrm{d}t \tag{9.2.17}$$

This principle states that the path of a particle (or a more general system) is obtained by requiring a certain functional S named "action" is stationary relative to a infinitesimal variation of the trajectory. In other words, if we make an infinitesimal variation of the trajectory, the change must be almost zero:

$$\boxed{\delta S=\delta\int\limits_{t_A}^{t_B} L\mathrm{d}t=\int\limits_{t_A}^{t_B}\delta L\mathrm{d}t=0} \tag{9.2.18}$$

A single, simple rule that underpins all of physics, encompassing every principle from Classical Mechanics to Electromagnetism, Quantum Theory, and General Relativity, as well as the fundamental particles that make up matter!

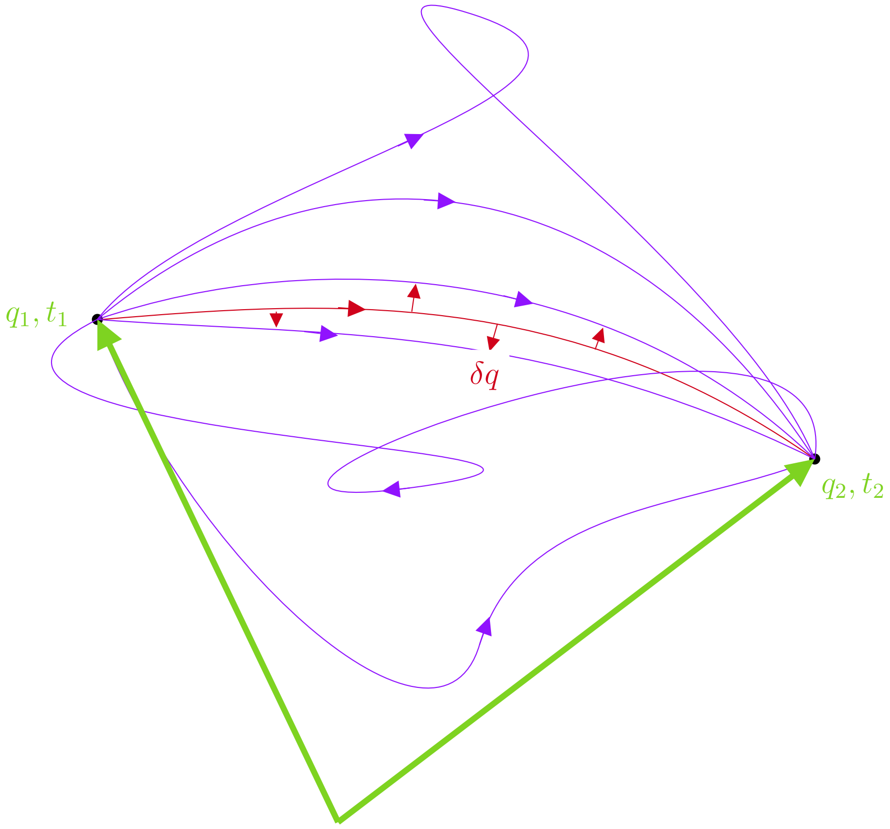


Figure 9.11: **Variation idea of Least Action Principle**

For a simple mechanical system, the action is therefore by evidence, by the principle of conservation of energy equal to the integral on the trajectory of (by definition of the Lagrangian) the difference between the kinetic energy and potential energy.

If you want to slide a mass from point A to point B, what shape of ramp will get it there the fastest? This is known as the "problem of fastest descent". Most people not educated to science will answer that fastest descent is a straight line from A to B. But if you bend the ramp down a bit at the beginning, well, the mass accelerates to a higher speed earlier. So even though it travels slightly farther, it travels faster and beats the straight ramp.

So the question is, what shape provides the perfect balance of acceleration and path length to minimize travel time? According to Galileo, it was the arc of a circle. He showed that this is faster than any polygon.

But is it the fastest? Nearly sixty years later, in June of 11696 (holocene calendar), Johann Bernoulli set this problem as a challenge to the best mathematicians in the world (mainly because he's a big show off and he wants to show that he's better than all of them). He gave everyone six months to come up with solutions, but none were submitted. So Gottfried Leibniz, a friend of Bernoulli's, persuaded him to extend the deadline to give foreigners a chance. We may think Isaac Newton was probably the intended target because everybody thought of him as the best, and so Johann would have wanted to show that he was better than Newton.

Isaac Newton was no longer a really active mathematician or physicist. He was working as the warden of the Mint, like a big high government position and on the 29th of January, 11697 (holocene calendar), Newton returned home after a long day at work to find Bernoulli's challenge in the mail.

Irritated, he wrote, «*I do not love to be dunned and teased by foreigners about mathematical things*». But the problem was too enticing, so Newton spent the rest of the day and night on it, and by 4:00 a.m. He had come up with a solution, something that took Bernoulli two weeks. Newton submitted his solution to the journal *Philosophical Transactions of the Royal Society of London* (see [414]). He sent his solution there but didn't sign it.

And Johann Bernoulli, after seeing the solution, is alleged to have said, «*I recognize the lion by his claw*». Who else could give such a solution (see page 3433).

And while overall Newton dominated Bernoulli, in this instance, Bernoulli's solution actually was much more elegant and interesting than Newton's.

How does light travel from one place to another? This was contemplated by Hero of Alexandria in the 101st century (according to holocene calendar). He realized that in a single medium, like air, light always follows the shortest path.

A consequence is that when light reflects, say, off a lake, the angle of incidence is always equal to the angle of reflection. Any other path between the start and end points would be longer. But when light goes from one medium into another, like from air into water, it bends in a peculiar way. So what is the guiding principle here?

Well, over the next 1,600 years people slowly figured out that the sine of the angle of incidence divided by the sine of the angle of refraction is equal to a constant, n, which depends on the nature of the two media. This came to be known as "Snell's Law" (see section Optics page pagerefSnell law). But no one knew why it worked.

That is, until 11657 (holocene calendar). So another great mathematician enters our story at this point. This is Pierre Fermat!

By day he was a judge, but at night he would come home and hang out with his wife and kids and then do what he really loved the most, which was work on math for fun. And he mostly worked in pure math but at one point he got interested in the question of, why does light obey this principle of refraction?

And he thought maybe Hero of Alexandria was on the right track, but it's not distance that is being minimized, but rather time.

But to see if this was true for refraction would be difficult. He would have to work out every possible path light could take by varying the point where it intersects the boundary and compute the time for each, and then show that light takes the path for which the total travel time is the shortest. He doesn't know how to solve it and he worries that it is going to be overcomplicated even if he could solve it, so he will not do it...

Years go by, Fermat gets interested in it five years later and tries to solve it, and then he does solve it and he shows that Snell's Law actually pops out as the minimizing path for light, you know, under these conditions of moving from one medium with one speed of light to another with another (see section Optics page pagerefSnell law).

And Fermat said that this is «*the most extraordinary, the most unforeseen, and the happiest calculation" of his life*».

If you use that principle of least time, you can explain everything that was known about light at the time of Fermat. It's the first time that anyone showed that nature obeys an optimization principle, that nature does the best possible thing. In this case, that light takes the shortest possible time!

Upon finding this solution, Bernoulli wrote, «*In this way I have solved at one stroke two important problems, an optical and a mechanical one, and have achieved more than I demanded from others. I have shown that the two problems taken from entirely separate fields of mathematics have the same character*».

Little did Bernoulli know he was onto something much bigger. Around 40 years later, one of his students, Pierre Louis de Maupertuis, also studied the behavior of light and particles, and he noticed that there are cases where the two behave very similarly. This made him think, what if Fermat's principle of least time wasn't the most fundamental principle?

Why should Nature care about minimizing time? Maybe there is a more foundational quantity being minimized, one that doesn't only govern light, but also particles. So, in the 11740s (holocene calendar), he proposed a new quantity, which he named the "action".

Therefore, in a theory in which the forces derives from a potential V, we are naturally led to define the "classic Lagrangian" by (the reader must absolutely try to remember it!):

$$\boxed{L(q,\dot{q})=T(q,\dot{q})-V(q,\dot{q})} \tag{9.2.19}$$

where T and V are the traditional notation in the Lagrangian formalism for the kinetic energy and respectively potential energy given by (for a unit mass):

$$T=\frac{1}{2}\dot{q}^2 \qquad V=V(q) \tag{9.2.20}$$

The variational principle is then written in the conservative case (which facilitates the interpretation of the integral as an infinite sum of infinitesimal differences of the kinetic and potential energy between two instants and that must obviously be zero in a conservative system):

$$\delta S=\int_{t_A}^{t_B}\delta L\mathrm{d}t=\int_{t_A}^{t_B}\delta(T-V)\mathrm{d}t=\int_{t_A}^{t_B}(\delta T-\delta V)\mathrm{d}t=0 \tag{9.2.21}$$

Which implies that well known fact that in any conservative system any change of kinetic energy requires a contrary variation of potential energy (hence the fact that it is implicitly a negative sign):

$$\delta T=\delta V \tag{9.2.22}$$

Strictly speaking, we do not always have the equality between the variational of the integral of the Lagrangian and the integral of the Lagrangian variational. Sometimes it can happen that:

$$\left(\delta\int_{t_A}^{t_B}L\mathrm{d}t\right)\neq\left(\int_{t_A}^{t_B}\delta L\mathrm{d}t=0\right) \tag{9.2.23}$$

Indeed, while the integral on the right of inequality has a quite relatively easy interpretation since it is in physics since the infinite sum of infinitesimal differences between potential and kinetic energy in a conservative system is zero, the integral on the left is the inexact differential (see section Differential and Integral Calculus page 1469) who will be zero if the Lagrangian does not depend on the way and therefore by extension does not explicitly depend on the variable t.

Remark

For the study of General Relativity (see page 5720), we are not looking for the variation of the difference of energy to be minimal as it is the case for mechanical systems, but the variation in the length $\mathrm{d}s$ of an arc (not time dependent unlike the previous example) in any space during a trajectory of a free system. Which bring us to simply write (remember it also because it will be very important!) the action:

$$\delta S=\int \delta(\mathrm{d}S)=0 \tag{9.2.24}$$

for a unit mass and taking the natural units.

Maybe let us do that in another way for pedagogical reasons (some people will like the first approach and some others the second approach). Remember that using Mapertuis notation:

$$\delta S_0=0 \tag{9.2.25}$$

And using Euler generalisation:

$$\delta\int mv\mathrm{d}\ell=0 \tag{9.2.26}$$

But we have:

$$v=\frac{\mathrm{d}\ell}{\mathrm{d}t}\rightarrow \mathrm{d}\ell=v\mathrm{d}t \tag{9.2.27}$$

Therefore:

$$\delta\int mv^2\mathrm{d}t=0 \tag{9.2.28}$$

But wait, that's just twice the kinetic energy such that:

$$\delta\int 2T\mathrm{d}t=\delta\int (T+T)\mathrm{d}t=0 \tag{9.2.29}$$

And as Euler pointed out, the total energy must be conserved. Total energy is just kinetic plus potential:

$$E=T+V\rightarrow T=E-V \tag{9.2.30}$$

So we can rewrite as:

$$\delta\int (T+E-V)\mathrm{d}t=\delta\int (T-V)\mathrm{d}t+\delta\int E\mathrm{d}t=0 \tag{9.2.31}$$

And since energy is constant:

$$\delta\int(T-V)\mathrm{d}t+\delta(Et)=\delta\int(T-V)\mathrm{d}t+E\delta t+t\delta E=0 \tag{9.2.32}$$

But remember, as Euler found, the energy of different paths has to be the same, so the variation between them is 0 and the term $t\delta E$ drops out. So it remains:

$$\delta\int(T-V)\mathrm{d}t+E\delta t=0 \tag{9.2.33}$$

If we rearrange:

$$\delta\int(T-V)\mathrm{d}t=-E\delta t \tag{9.2.34}$$

This looks a lot like some other minimization principle, if only $E\delta t$ was equal to 0. Well, we can make this term 0 by only considering paths that have the same travel time. So if we do that, then there's no longer any variation in time and this term drops out and it remains:

$$\delta\int(T-V)\mathrm{d}t=\delta\int L\mathrm{d}t=0 \tag{9.2.35}$$

The first person to write the principle of least action like this was William Rowan Hamilton in 11834 (holocene calendar), and by doing so, he got the principle also named after him.

Hamilton's Principle is the modern way of writing the principle of least action and the way you would find it in almost every physics textbook.

In part that's because Hamilton's Principle tells you how objects move from one place to another rather than just giving you the shape of the path. Two other important differences between both principles are that the action is now an integral over time instead of space. And a consequence is that with Hamilton's Principle, you now need a start and end point, and also a start and end time.

After Euler found out about Lagrange's proof, he wrote to him the following: «*How satisfied would not Mr Maupertuis be, were he still alive, if he could see his principle of least action applied to the highest degree of dignity to which it is susceptible*». With Lagrange's proof, we now have two ways to solve any mechanics problem. You can either use forces and vectors or you can use energies and scalars.

To return to our application of the variational principle in the case of the generalized Lagrangian, we can write the following exact total differential (see section Differential and Integral Calculus page 1469) of δL and then we get the relation:

$$\delta S=\int\limits_{t_A}^{t_B}\delta L\mathrm{d}t=\int\limits_{t_A}^{t_B}\sum_{i=1}^{n}\left(\frac{\partial L}{\partial q_i}\delta q_i+\frac{\partial L}{\partial \dot{q}_i}\delta\dot{q}_i\right)\mathrm{d}t=0 \tag{9.2.36}$$

So the idea here is that if we have the path of least action, then if we were to change it a little bit by, say, adding a tiny bump here or flattening it out there to our path of least action, well, then the action basically shouldn't change because we are in this really special configuration: the path of least action. We add a little bit to the minimal path but the action is still the same!

Let us integrate by parts (see section Differential and Integral Calculus page 1532) the second term of the sum of the previous integral:

$$\int_{t_A}^{t_B} \frac{\partial L}{\partial \dot{q}_i}\delta \dot{q}_i \mathrm{d}t = \int_{t_A}^{t_B} \frac{\partial L}{\partial \dot{q}_i}\delta \frac{\mathrm{d}}{\mathrm{d}t}q_i \mathrm{d}t = \frac{\partial L}{\partial \dot{q}_i}\delta q_i \bigg|_{t_A}^{t_B} - \int_{t_A}^{t_B} \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}_i}\delta q_i \mathrm{d}t \qquad (9.2.37)$$

The first term of the last equality is equal to zero:

$$\frac{\partial L}{\partial \dot{q}_i}\delta q_i \bigg|_{t_A}^{t_B} = \frac{\partial L}{\partial \dot{q}_i}(t_B)\delta q_i(t_B) - \frac{\partial L}{\partial \dot{q}_i}(t_A)\delta q_i(t_A) = \frac{\partial L}{\partial \dot{q}_i}(t_B)\cdot 0 - \frac{\partial L}{\partial \dot{q}_i}(t_A)\cdot 0 = 0 \qquad (9.2.38)$$

since we have already mentioned above that by construction we must have:

$$\delta q_i(t_A) = \delta q_i(t_B) = 0 \qquad (9.2.39)$$

The expression of the integral of least action can finally be written:

$$\delta S = \int_{t_A}^{t_B} \delta L \mathrm{d}t = \int_{t_A}^{t_B} \sum_{i=1}^{n} \left(\frac{\partial L}{\partial q_i} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}_i}\right)\delta q_i \mathrm{d}t = 0 \qquad (9.2.40)$$

But as the δq_i and $\dot{q}_i$ tend to 0 in an infinite number of ways and we must have however $\delta S = 0$. This means that each term in the sum of the integral can be taken independently and has to satisfy:

$$\boxed{\forall i \in \mathbb{N}^* : \delta S = \int_{t_A}^{t_B} \delta L \mathrm{d}t = \int_{t_A}^{t_B} \left(\frac{\partial L}{\partial q_i} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}_i}\right)\delta q_i \mathrm{d}t = 0} \qquad (9.2.41)$$

But as the functions δq_i and $\delta \dot{q}_i$ functions can always approach zero in many ways, and that this integral should be anyway equal to zero , we deduce that it is the integrands that are zero:

$$\boxed{\forall i \in \mathbb{N}^* : \delta S = \frac{\partial L}{\partial q_i} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}_i} = 0} \qquad (9.2.42)$$

These n equations, satisfied by the generalized Lagrangian of the system to effectively track movement, are named "Euler-Lagrange equations", or more briefly (but rarely) "Lagrange equations". These are, as we shall see, the equations of motion of the system: resolved, they give the effective evolution of the system over time:

$$\frac{\partial L}{\partial q_i} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}_i} = 0 \qquad (9.2.43)$$

or more explicitly we have the following second order differential equation:

$$\boxed{\frac{\partial L(q_i,\dot{q}_i,t)}{\partial q_i} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L(q_i,\dot{q}_i,t)}{\partial \dot{q}_i} = 0} \qquad (9.2.44)$$

And the reader must not forget that this in the case of a conservative system!!!!

If the system is not conservative, by the analysis of units we have:

$$\boxed{\frac{\partial L(q_i,\dot{q}_i,t)}{\partial q_i}-\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L(q_i,\dot{q}_i,t)}{\partial \dot{q}_i}=F_i} \tag{9.2.45}$$

where F_i represents the generalized forces not arising from a potential (even if sometimes we consider a force arising from a potential as not arising from a potential as it is more easy to manipulate for students as it brings anyway to the same result when the force depends explicitly on the acceleration).

The Lagrange equations should describe a particle motion. Given $L=T-V$ we should expect to recover Newton's second law for a particle moving under the influence of a force. We have (see section Classical Mechanics page 3297):

$$T=\frac{1}{2}m(\dot{x}^2+\dot{y}^2+\dot{z}^2) \tag{9.2.46}$$

It follows that in absence of a potential:

$$L=T-V=T-0=T \tag{9.2.47}$$

Therefore:

$$\frac{\partial T}{\partial x}=\frac{\partial T}{\partial y}=\frac{\partial T}{\partial z}=0 \tag{9.2.48}$$

and:

$$\frac{\partial T}{\partial \dot{x}}=m\dot{x},\frac{\partial T}{\partial \dot{y}}=m\dot{y},\frac{\partial T}{\partial \dot{z}}=m\dot{z} \tag{9.2.49}$$

The non-conservative equation of motion is:

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}_i}-\frac{\partial L}{\partial q_i}=F_i \tag{9.2.50}$$

where $q_1=x,q_2=y,q_3=z$. The equation of motion then becomes:

$$\frac{\mathrm{d}}{\mathrm{d}t}(m\dot{x})=F_x \tag{9.2.51}$$

and identically for y and z, meaning that as expected we recovered Newton's 2nd law!

✍**Example:**

A common model of friction is:

$$F_f=-Kv \tag{9.2.52}$$

(see the study of tribology in the section of Classical Mechanics at page 3452) where v is the speed of the body in translational motion. Consider the corresponding "Rayleigh dissipation function" (energy dissipation due to friction):

$$E_F=\frac{1}{2}Kv^2 \tag{9.2.53}$$

If we use a Cartesian coordinate system, the generalized force due to friction, in x-direction for instance is given by:

$$F_{f_x}=-Kv_x=-\frac{\partial E_f}{\partial v_x} \tag{9.2.54}$$

that we will also denote:

$$F_{f_1}=-Kv_1=-\frac{\partial E_f}{\partial v_1} \tag{9.2.55}$$

or as in Analytical Mechanics some physicists denote E_f by $\mathcal{F}$ we have also sometimes:

$$F_{f_1}=-Kv_1=-\frac{\partial \mathcal{F}}{\partial v_1} \tag{9.2.56}$$

Therefore through this example we see that we define the "generalized force" as:

$$\boxed{F_i=-\frac{\partial \mathcal{F}}{\partial \dot{q}_i}} \tag{9.2.57}$$

So, the equations of motion in the case of a non-conservative force is written:

$$\boxed{\frac{\partial L(q_i,\dot{q}_i,t)}{\partial q_i}-\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L(q_i,\dot{q}_i,t)}{\partial \dot{q}_i}-\frac{\partial \mathcal{F}}{\partial \dot{q}_i}=0} \tag{9.2.58}$$

If instead, the friction was proportional to the acceleration for some reason (as do a gravitational force for example but this latter is in the V term as it arise from a potential), we could go one step further and define a new Lagrangian:

$$L=T-V+\mathcal{F} \tag{9.2.59}$$

and using Lagrange's equations on L would give us the correct answer.

Most of time when we use the Lagrangian formalism in quantum physics the studied system is in vacuum so the first boxed relation above applies but when the system if for example in a gravitational field then an external force F is applied and then second boxed relation has to be applied (see the study of the gyroscope nutation in the section of Classical Mechanics at page 3324 for a detailed example).

Remark

This is while studying physics (the following sections of this textbook) that we can better understand the applications of this equation (obtained almost by purely mathematical developments !!!) and it becomes afterwards possible to understand its deep meaning. At our level of discourse, it is pointless to say anything for the moment. We must do physics, and physics again and again... to understand it well (learning by doing as we say!).

So in the Lagrangian approach, we learn to reason from concepts of potential and kinetic energy instead of force concepts. The two approaches are obviously physically equivalent, but the energies are not vector quantities, they are conceptually easier to use in a wide range of situations. In quantum physics, for example, the notion of force has no meaning but energy concepts remains valid and same in General Relativity!!! This is one more reason to become familiar with their use. In addition, the force in the sense of Newton is instantaneous distance action. In relativity, such a thing is impossible. The concept of force is therefore a purely classical macroscopic creation and contrary to our intuition, its interest is limited.

Let us see a very first simple example of application of Euler-Lagrange's equation (other examples will be seen during our study of Newton's laws, of Electrodynamics, Special Relativity, General Relativity, Quantum Physics Fields, etc.):

✍**Examples:**

E1. First, let us write in a conventional mathematical form the Euler-Lagrange equation (the notation of the generalized coordinates is not the same in mathematics than in physics...):

$$\frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial Z(x,y,y')}{\partial y'}-\frac{\partial Z(x,y,y')}{\partial y}=0 \tag{9.2.60}$$

And now we take a simple practical mathematical example internationally known and very important (we will reuse the developments made below four our study of the Huygens pendulum later).

The problem statement is: determine the shortest path between two points of a plane (we're guessing it's the straight line but must prove it!). In other words: to find the geodesic of a plane!

This problem consist to find the shortest parametrized curve $x(t), t\in[t_0,t_1]$that connects two points (caution with the variable t has nothing to do with the time in this example!) such that:

$$x(t_0)=x_0 \qquad x(t_1)=x_1 \tag{9.2.61}$$

Thus the infinitesimal length $\mathrm{d}L$ by applying Pythagoras is (differential curvilinear abscissa):

$$\begin{aligned}&\mathrm{d}L^2=\mathrm{d}t^2+\mathrm{d}x^2\\&\mathrm{d}L=\sqrt{\mathrm{d}t^2+\mathrm{d}x^2}=\sqrt{1+\frac{\mathrm{d}x^2}{\mathrm{d}t^2}}\mathrm{d}t=\sqrt{1+\dot{x}^2(t)}\mathrm{d}t\end{aligned} \tag{9.2.62}$$

Thus, the length L of the parametric curve is given by:

$$\boxed{\int\limits_{t_0}^{t_1}\sqrt{1+\dot{x}^2(t)}\mathrm{d}t} \tag{9.2.63}$$

It is a relation we will often meet again in physics and math !!

Thus, this problem, which the geometrical solution is very simple, is formulated as variational calculation problem as follows:

$$\min L(x,\dot{x},t)=\sqrt{1+\dot{x}^2} \tag{9.2.64}$$

Let us write the Euler-Lagrange of the solution to this problem, solution that if it exists, it has to satisfy.

Therefore we have:

$$\begin{gathered}
Z(x,y,y')=\sqrt{1+y'^2}\\
\frac{\partial Z}{\partial y}=0\\
\frac{\partial Z}{\partial y'}=\frac{y'}{\sqrt{1+y'^2}}
\end{gathered} \tag{9.2.65}$$

The Euler-Lagrange equation therefore becomes:

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial Z(x,y,y')}{\partial y'}\right)-\frac{\partial Z(x,y,y')}{\partial y}=\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{y}{\sqrt{1+y'^2}}\right)=0 \tag{9.2.66}$$

Therefore:

$$\frac{y}{\sqrt{1+y'^2}}=C \tag{9.2.67}$$

where $C\in\mathbb{R}$ is a constant of integration (see section Differential and Integral Calculus page 1512). The latter equality implies that (we add an index for the constant that will avoid us to confused it with two other constants that will appear little further below):

$$y'=\frac{\mathrm{d}y}{\mathrm{d}x}=C\sqrt{1-y'^2} \tag{9.2.68}$$

Returning to the notation used in the beginning of our problem statement:

$$x'=\frac{\mathrm{d}x}{\mathrm{d}t}=\dot{x}=C_1\sqrt{1-\dot{x}^2} \tag{9.2.69}$$

Therefore:

$$\dot{x}^2=C_1^2-C_1^2\dot{x}^2\Leftrightarrow(C_1^2+1)\dot{x}^2=C_1^2\Leftrightarrow\dot{x}=\sqrt{\frac{C_1^2}{C_1^2+1}} \tag{9.2.70}$$

Therefore:

$$\mathrm{d}x=C_2\mathrm{d}t \tag{9.2.71}$$

and by integration it comes:

$$x(t)=C_2t+C_2 \tag{9.2.72}$$

This is the indeed the of a line named therefore the "geodesic equation in cartesian coordinates". This can be seen obviously by rewriting the constants as:

$$x(t)=at+b \tag{9.2.73}$$

Identical to the famous line equation:

$$y(x)=ax+b \tag{9.2.74}$$

E2. We will now focus on an example frequently asked on Internet forums: what is the geodesic of an $\mathcal{S}^2$ sphere?

For this, and without using Tensor Calculus (!), let us recall that in spherical coordinates (see section Vector Calculus page 1849):

$$\mathrm{d}\vec{r}=\vec{e}_r\mathrm{d}r+\vec{e}_\theta r\mathrm{d}\theta+\vec{e}_\phi\sin(\theta)\mathrm{d}\phi \tag{9.2.75}$$

Without loss of generality, we may take the sphere to be of unit constant radius. Therefore the previous relations reduce to:

$$\mathrm{d}\vec{r}=\vec{e}_\theta\mathrm{d}\theta+\vec{e}_\phi\sin(\theta)\mathrm{d}\phi \tag{9.2.76}$$

The length of a path from a point A to B on the sphere is the given by:

$$\ell=\int\limits_A^B||\mathrm{d}\vec{r}||=\int\limits_A^B\sqrt{\mathrm{d}\theta^2+\sin^2(\theta)\mathrm{d}\phi^2}=\int\limits_{\theta_A}^{\theta_B}\sqrt{1+\sin^2(\theta)\phi'^2}\mathrm{d}\theta \tag{9.2.77}$$

So as before the variational problem is equivalent to put that:

$$\min L(\theta,\phi,\phi')=\sqrt{1+\sin^2(\theta)\phi'^2} \tag{9.2.78}$$

Now applying the Euler-Lagrange equation:

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial Z(x,y,y')}{\partial y'}\right)-\frac{\partial Z(x,y,y')}{\partial y}=0 \tag{9.2.79}$$

in spherical coordinates:

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\frac{\partial L(\theta,\phi,\phi')}{\partial \phi'}\right)-\frac{\partial L(\theta,\phi,\phi')}{\partial \phi}=0 \tag{9.2.80}$$

Explicitly:

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\frac{\partial}{\partial\phi'}\sqrt{1+\sin^2(\theta)\phi'^2}\right)-\frac{\partial}{\partial\phi}\sqrt{1+\sin^2(\theta)\phi'^2}=0 \tag{9.2.81}$$

As:

$$\frac{\partial}{\partial\phi}\sqrt{1+\sin^2(\theta)\phi'^2}=0 \tag{9.2.82}$$

It remains:

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\frac{\partial}{\partial\phi'}\sqrt{1+\sin^2(\theta)\phi'^2}\right)=0 \tag{9.2.83}$$

Therefore:

$$\frac{\partial}{\partial\phi'}\sqrt{1+\sin^2(\theta)\phi'^2}=c^{te} \tag{9.2.84}$$

Hence:

$$\frac{1}{2}\frac{2\sin^2(\theta)\phi'}{\sqrt{1+\sin^2(\theta)\phi'^2}}=\frac{\sin^2(\theta)\phi'}{\sqrt{1+\sin^2(\theta)\phi'^2}}=c^{te} \tag{9.2.85}$$

Therefore after squaring and rearranging we get:

$$\sin^4(\theta)\phi'^2=(c^{te})^2(1+\sin^2(\theta)\phi'^2) \tag{9.2.86}$$

Hence:

$$\phi'^2\left(\sin^4(\theta)-(c^{te})^2\sin^2(\theta)\right)=(c^{te})^2 \tag{9.2.87}$$

Finally:

$$\phi'=\frac{c^{te}}{\sqrt{\sin^4(\theta)-(c^{te})^2\sin^2(\theta)}}=\frac{c^{te}}{\sin(\theta)\sqrt{\sin^2(\theta)-(c^{te})^2}} \tag{9.2.88}$$

and the problem reduces to integrating this with respect to θ as $\phi'=\mathrm{d}\phi/\mathrm{d}\theta$:

$$\phi=\int\frac{c^{te}}{\sin(\theta)\sqrt{\sin^2(\theta)-(c^{te})^2}}\mathrm{d}\theta \tag{9.2.89}$$

We substitute (see section Differential and Integral Calculus page 1475):

$$u=\frac{\cos(\theta)}{\sin(\theta)}=\frac{1}{\tan(\theta)}\overset{\text{def}}{=}\cot(\theta)\Rightarrow \mathrm{d}u=-\frac{1}{\sin^2(\theta)}\mathrm{d}\theta\overset{\text{def}}{=}-\text{cosec}^2(\theta)\mathrm{d}\theta \tag{9.2.90}$$

Then:

$$\begin{aligned}\phi&=\int\frac{c^{te}}{\sin(\theta)\sqrt{\sin^2(\theta)-(c^{te})^2}}\mathrm{d}\theta=-\int\frac{c^{te}}{\text{cosec}^2(\theta)\sin(\theta)\sqrt{\sin^2(\theta)-(c^{te})^2}}\\&=-\int\frac{c^{te}}{\sqrt{1-(c^{te})^2\text{cosec}^2(\theta)}}\mathrm{d}u\end{aligned} \tag{9.2.91}$$

But as:

$$\cos^2(\theta)+\sin^2(\theta)=1\Leftrightarrow\frac{\cos^2(\theta)}{\sin^2(\theta)}+1=\frac{1}{\sin^2(\theta)}\Leftrightarrow u^2+1=\text{cosec}^2(\theta) \tag{9.2.92}$$

Therefore:

$$\phi=-\int\frac{c^{te}}{\sqrt{1-(c^{te})^2(1+u^2)}}\mathrm{d}u \tag{9.2.93}$$

Now we substitute:

$$a=\frac{\sqrt{1-(c^{te})^2}}{c^{te}} \tag{9.2.94}$$

Hence (see section Differential and Integral Calculus page 1534):

$$\phi=-\int\frac{1}{\sqrt{a^2-u^2}}\mathrm{d}u=-\arcsin\left(\frac{u}{a}\right)+C^{te} \tag{9.2.95}$$

where C^{te} is a constant of integration that we will denote by ϕ_0. Therefore:

$$\phi-\phi_0=-\arcsin\left(\frac{u}{a}\right)\Rightarrow\sin(\phi-\phi_0)=-\frac{u}{a}\Rightarrow a\sin(\phi-\phi_0)=-u \tag{9.2.96}$$

Since, for recall, $u=\cot(\theta)$, we have finally:

$$a\sin(\phi-\phi_0)=-\cot(\theta) \tag{9.2.97}$$

Let us choose $\phi_0=\pi/2$. Then:

$$-a\cos(\phi)=-\cot(\theta) \tag{9.2.98}$$

Therefore:

$$\boxed{a\cos(\phi)=\cot(\theta)} \tag{9.2.99}$$

So this is the "parametric equation of a geodesic on a 2-sphere" and we see that this is identical to relation of the "great circle" that we get in the section of Analytical Geometry during our study of the parametrization of a sphere in a three dimensional space.

It seems to be personal preference, and all academic, whether you use the Lagrangian method $L=T-V$ or the $F=ma$ method. The two methods produce the same equations. However, in problems involving more than one variable (see typically the study of oscillating motions in the section of Classical Mechanics), it usually turns out to be much easier to write down T and V, as opposed to writing down all the forces. This is because T and V are nice and simple scalars. The forces, on the other hand, are vectors, and it is easy to get confused if they point in various directions. The Lagrangian method has the advantage that once you've written down $L=T-V$, you don't have to think anymore. All you have to do is blindly take some derivatives.

Therefore:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial\dot{x}}\right)=\frac{\partial L}{\partial x}\Rightarrow m\ddot{x}=m(l+x)\dot{\theta}^2+mg\cos(\theta)-kx \tag{9.2.100}$$

and:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial\dot{\theta}}\right)\Rightarrow\frac{\partial L}{\partial\theta}&\Rightarrow\frac{\mathrm{d}}{\mathrm{d}t}(m(l+x)^2\dot{\theta})=-mg(l+x)\sin(\theta)\\&\Rightarrow m(l+x)^2\ddot{\theta}+2m(l+x)\dot{x}\dot{\theta}=-mg(l+x)\sin(\theta)\\&\Rightarrow m(l+x)\ddot{\theta}+2m\dot{x}\dot{\theta}=-mg\sin(\theta)\end{aligned} \tag{9.2.101}$$

9.2.1.3.1 Beltrami Identity

We will prove here a useful relation named "Beltrami identity" which simplifies the application of the Euler-Lagrange equation in some particular situations!

Let us first recall the Euler-Lagrange equation:

$$\frac{\partial L(q_i,\dot{q}_i,t)}{\partial q_i}-\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L(q_i,\dot{q}_i,t)}{\partial \dot{q}_i} \tag{9.2.102}$$

Let us write the exact total differential (see section Differential and Integral Calculus page 1466):

$$\mathrm{d}L=\frac{\partial L}{\partial t}\mathrm{d}t+\frac{\partial L}{\partial q_i}\mathrm{d}q_i+\frac{\partial L}{\partial \dot{q}_i}\mathrm{d}\dot{q}_i \tag{9.2.103}$$

What we will write under the form:

$$\frac{\mathrm{d}L}{\mathrm{d}t}=\frac{\partial L}{\partial t}+\frac{\partial L}{\partial q_i}\frac{\mathrm{d}q_i}{\mathrm{d}t}+\frac{\partial L}{\partial \dot{q}_i}\frac{\mathrm{d}\dot{q}_i}{\mathrm{d}t}=\frac{\partial L}{\partial t}+\frac{\partial L}{\partial q_i}\dot{q}_i+\frac{\partial L}{\partial \dot{q}_i}\ddot{q}_i \tag{9.2.104}$$

Rearranging:

$$\frac{\partial L}{\partial q_i}\dot{q}_i=\frac{\mathrm{d}L}{\mathrm{d}t}-\frac{\partial L}{\partial \dot{q}_i}\ddot{q}_i-\frac{\partial L}{\partial t} \tag{9.2.105}$$

Let us multiply before continuing the Euler-Lagrange by $\dot{q}_i$:

$$\dot{q}_i\frac{\partial L}{\partial q_i}-\dot{q}_i\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}_i}=0 \tag{9.2.106}$$

and let us inject the prior previous relation in the previous one:

$$\dot{q}_i\frac{\partial L}{\partial q_i}-\dot{q}_i\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}_i}=\left(\frac{\mathrm{d}L}{\mathrm{d}t}-\frac{\partial L}{\partial \dot{q}_i}\ddot{q}_i-\frac{\partial L}{\partial t}\right)-\dot{q}_i\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}_i}=0 \tag{9.2.107}$$

After a small factorization we get:

$$-\frac{\partial L}{\partial t}+\frac{\mathrm{d}}{\mathrm{d}t}\left(L-\dot{q}_i\frac{\partial L}{\partial \dot{q}_i}\right)=0 \tag{9.2.108}$$

In the particular conditions (but relatively frequent in physics), which we will name in this textbook the "Beltrami condition", where:

$$\frac{\partial L}{\partial t}=0 \tag{9.2.109}$$

We get the "Beltrami identity":

$$\boxed{\frac{\mathrm{d}}{\mathrm{d}t}\left(L-\dot{q}_i\frac{\partial L}{\partial \dot{q}_i}\right)=0} \tag{9.2.110}$$

which bring us to have:

$$L-\dot{q}_i\frac{\partial L}{\partial \dot{q}_i}=c^{te} \tag{9.2.111}$$

that we will be very useful for example in the section of Classical Mechanics for our study of the brachistochrone page 3433.

9.2.1.3.2 Theorem of Variational Calculus

The theorem of variational calculus consist to prove that by considering a continuous function f on the interval $[a,b]$ on $\mathbb{R}$ and H the set of continuous functions on $[a,b]$ infinitely differentiable on $]a,b[$ and which vanish on a and b, then for any function $h\in H$:

$$\int\limits_a^b f(x)h(x)\mathrm{d}x=0 \qquad (9.2.112)$$

f is zero on $[a,b]$.

Why focus on this theorem? Because we will meet it frequently during the application of the variational principle in similar configuration. Indeed, let us recall that the variational principle leads to have:

$$\delta S=\int\delta(\mathrm{d}s)=0 \qquad (9.2.113)$$

and the integrated expression is rarely a simple function as the reader will notice in reading the different following chapters of this textbook. It is therefore important to know a property that sometimes simplifies the analysis of the problem.

Remark

Some people may think that the case $f(x)=1$ with $h(x)=\sin(x)$ and $x\in[0,2\pi]$ contradicts the statement of the theorem! In fact it's not really that ... this theorem must be valid for $\forall h\in H$ and not only just for the mentioned example. Hence the fact that f will have to be equal zero as we will prove it just now.

Proof 9.1.2. To simplify we will take the case $a=-1,b=1$. With a little bit more technical details the proof by contradiction below can be adapted to the case of any a, b.

Let us suppose that f is not on $[-1,1]$. Then there exists $x_0\in]-1,1[$ such that $f(x_0)\neq 0$. We can the suppose $f(x_0)>0$ (same reasoning if $f(x_0)<0$).

By the initial assumption of continuity and of not nullity of f there exists a small interval around x_0 on which f is strictly positive. That is to say, there exists $\varepsilon>0$ such that $[x_0-\varepsilon,x_0+\varepsilon]\subset[-1,1]$ and $\forall y\in[x_0-\varepsilon,x_0+\varepsilon], f(y)>0$.

Let us consider now the function $\pi:[-1,1]\mapsto\mathbb{R}$:

$$\phi(x)=\begin{cases} e^{-\frac{1}{1-x^2}} & \text{if } |x|<1\\ 0 & \text{otherwise}\end{cases} \qquad (9.2.114)$$

We check easily that ϕ is continuous (positive) on $[-1,1]$ and infinitely differentiable on $[-1,1[$ (see section Differential and Integral Calculus page 1496).

In addition, $\phi(-1)=\phi(1)=0$. And therefore, $\phi\in H$. Here is a graphical representation of ϕ:

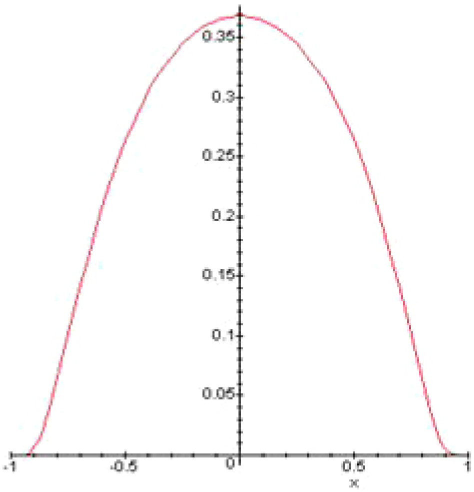


Figure 9.12: **Sample Function to prove the variational calculus**

From ϕ we want get a continuous function on $[-1,1]$, infinitely differentiable on $]-1,1[$, positive on $[x_0-\varepsilon, x_0+\varepsilon]$ and zero outside $[x_0-\varepsilon, x_0+\varepsilon]$ to show the absurdity of the assumption of non nullity of f so that the theorem is verified (remember that we are doing a proof by contradiction!).

For this, it is enough to center ϕ on x_0 and contract it.

The function $h:[-1,1]\mapsto\mathbb{R}$ defined by:

$$h(x)=\begin{cases}\phi\left(\dfrac{x-x_0}{\varepsilon}\right) & \text{if } x\in[x_0-\varepsilon,x_0+\varepsilon]\\ 0 & \text{otherwise}\end{cases} \tag{9.2.115}$$

satisfies the required criteria. Moreover, $h(-1)=h(1)=0$ and therefore $h\in H$.

Therefore, the function $f\cdot h$ will be continuous on $[-1,1]$ and positive on $[x_0-\varepsilon,x_0+\varepsilon]$ and zero everywhere else.

We have:

$$\int_{-1}^{1} f(x)h(x)\mathrm{d}x=0 \tag{9.2.116}$$

But, if a function $g:[a,b]\mapsto\mathbb{R}$ is continuous and positive and:

$$\int_{a}^{b} g(x)\mathrm{d}x=0 \tag{9.2.117}$$

this leads inevitably (we would suppose that as too intuitive to be proved) to $g=0$ on $[a,b]$.

Therefore $f(x)h(x)=0$ on $[x_0-\varepsilon,x_0+\varepsilon]$ or $f(x_0)h(x_0)\neq 0$ according to our initial absurd assumption, which is contradictory.

The starting assumption is therefore false and f must be zero on $[-1,1]$.

■ Q.E.D.

9.2.2 Canonical Formalism

The canonical formalism does not introduce a new type of physics but offers a new range of tools to study physical phenomena. Its masterpiece, the "Hamiltonian", plays an important role in quantum physics.

As in the Lagrange formalism we work with quantities (scalars) such as energy, T and V rather than with vector quantities such as the Newton force.

In the formalism of Lagrange, the description of an n degrees of freedom mechanical system is as we know described by the general (unconstrained) independent coordinates q_i that leads to n Euler-Lagrange equations of the type:

$$\frac{\partial L}{\partial q_i}-\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}_i}=0 \Leftrightarrow \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}_i}-\frac{\partial L}{\partial q_i}=0 \tag{9.2.118}$$

which are second order differential equations (see section Differential and Integral Calculus page 1610).

With Lagrange we compare mainly trajectories and therefore the q_i and the $\dot{q}_i$ are independent. By Hamilton we must first learn to define the "general momentum" denoted $\vec{p}_i$, replacing the generalized coordinates $\dot{q}_i$ and q_i which are also independent.

Remark

The origin of the conjugated momentum will be trivial when we will see further below a first concrete example.

9.2.2.1 Legendre Transform

The "Legendre transformation" is often used in thermodynamics (see section of the same name page 3677) because it allows to interconnect the various thermodynamic potentials. In mechanics or in geometry it gives the possibility to define the Hamiltonian from the Lagrangian and vice versa. We will give here a simplified and sufficient description of it in the univariate case (that is to say: in one dimension).

Given a function $f(u,v)$ where u,v are the both independent variables of which f depends.

Let us define:

$$w:=\frac{\partial f(u,v)}{\partial}=w(u,v) \tag{9.2.119}$$

The Legendre transformation gives the possibility to define a function $g(u,w)$ that can replace $f(u,v)$ in the sense that it rid of one of the explanatory variables (which is very useful in thermodynamics!):

$$f(u,v)\mapsto g(u,w)=vw-f \tag{9.2.120}$$

Given now the total differential of f (see section Differential and Integral Calculus page 1466):

$$\mathrm{d}f=\frac{\partial f}{\partial u}\mathrm{d}u+\frac{\partial f}{\partial v}\mathrm{d}v=\frac{\partial f}{\partial u}\mathrm{u}+w\mathrm{d}v \tag{9.2.121}$$

From the definition of g we get:

$$\begin{aligned}\mathrm{d}g&=w\mathrm{d}v+v\mathrm{d}w-\mathrm{d}f\\&=w\mathrm{d}v+v\mathrm{d}w-\frac{\partial f}{\partial u}\mathrm{d}u-w\mathrm{d}v\\&=v\mathrm{d}w-\frac{\partial f}{\partial u}\mathrm{d}u=\frac{\partial g}{\partial w}\mathrm{d}w+\frac{\partial g}{\partial u}\mathrm{d}u\end{aligned} \tag{9.2.122}$$

and then we have:

$$v=\frac{\partial g}{\partial w}\qquad\frac{\partial g}{\partial u}=-\frac{\partial f}{\partial u} \tag{9.2.123}$$

Obviously we end up just with a sign difference to the same result if we put since the beginning:

$$f(u,v)\mapsto g(u,w)=f-vw \tag{9.2.124}$$

As this is usage in Thermodynamics.

9.2.2.2 Hamiltonian

Given a $L(q_i,\dot{q}_i)$ that we will assimilate the function f introduce just above with the q_i serving playing the role of the u and the $\dot{q}_i$ the role of the v. For w, we define the generalized moments also named "canonical moments".

Then for each generalized velocity, there is one corresponding conjugate momentum, defined as:

$$\boxed{p_i:=\frac{\partial L}{\partial\dot{q}_i}=p_i(q_j,\dot{q}_j)} \tag{9.2.125}$$

with $i,j\in\mathbb{N}^*$.

In Cartesian coordinates, the generalized momenta are precisely the physical linear momenta. In circular polar coordinates, the generalized momentum corresponding to the angular velocity is the physical angular momentum. For an arbitrary choice of generalized coordinates, it may not be possible to obtain an intuitive interpretation of the conjugate momenta.

Before continuing let's see what allows us to do this definition:

So we define, in analogy with g, a function of the q_i and the p_i which we will denote by $H(q_i,p_i)$:

$$\begin{array}{ccccc} g(u,w) & = & vw & - & f \\ \Downarrow & & \Downarrow & & \Downarrow \\ H(q_i,p_i) & = & \displaystyle\sum_{i=1}^{n}\dot{q}_i p_i & - & L(q_i,\dot{q}_i) \end{array}$$

The relation obtained just above (that shows clearly that the Hamiltonian is the Legendre transform of the Lagrangian!):

$$\boxed{H(q_i,p_i)=\sum_{i=1}^{n}\dot{q}_i p_i - L(q_i,\dot{q}_i)} \tag{9.2.126}$$

named "Hamilton function" or "Hamiltonian" is more than important (like everything else anyway). We will see this relation again, among others, in the section of Relativistic Quantum Physics or Quantum Field Physics. Moreover, a beautiful example of what we have seen now is given in the section of Special Relativity where we calculate the Lagrangian and Hamiltonian of a free particle. The results are quite relevant and useful and their accuracy in electrodynamics are more than amazing!

If we rewrite the relation above for one degree of freedom, we have obviously:

$$H(q_i,p_i)=\dot{q}p-L(q,\dot{q})=\dot{q}\frac{\partial L}{\partial \dot{q}}-L \tag{9.2.127}$$

Now consider the special case where:

$$H(q_i,p_i)=\dot{q}\frac{\partial L}{\partial \dot{q}}-L=0 \tag{9.2.128}$$

This means obviously that:

$$\dot{q}\frac{\partial L}{\partial \dot{q}}=L \tag{9.2.129}$$

We recognize here a special case of Euler's homogeneous function theorem (see section Functional Analysis page 2282). So $H=0$ for every Lagrangian that is homogeneous of degree 1 (i.e. re-parametrization invariant).

✍**Example:**

Another important and well known application analytical mechanics is the calculation of minimal surfaces (for physics and architecture). If we focus on the determination of such a surface by requiring it to be a surface of revolution, we will see that we find a catenoid (the form taken by a soap film between two rings).

We give us ourselves the radius R_1 and R_2 of two circles of distance between them of l. We seek a function y of class $\mathcal{C}^1$ such that:

$$y(0)=R_1 \qquad \text{and} \qquad y(l)=R_2 \tag{9.2.130}$$

and that the parametric surface of revolution of the type:

$$\{(x,y(x)\cos(\phi),y(x)\sin(\phi)),x\in[0,l],\phi[0,2\pi]\} \tag{9.2.131}$$

has a minimal surface.

We know that the surface of a volume of revolution can be written (see section Geometric Shapes page 2956):

$$S=\int 2\pi f(x)\sqrt{1+f'(x)^2}\mathrm{d}x=2\pi\int y\sqrt{1+y'^2}\mathrm{d}x=2\pi\int L(y,y')\mathrm{d}x \tag{9.2.132}$$

Thus by making the function vary:

$$\delta S=\int \delta L(y,y')\mathrm{d}x=\int\left(\frac{\partial L}{\partial y}\delta y+\frac{\partial L}{\partial y'}\delta y'\right)\mathrm{d}x \tag{9.2.133}$$

As $\delta(y')=(\delta y)'$ the integration by parts of the second term gives:

$$\int\frac{\partial L}{\partial y'}\delta y'\mathrm{d}x=\frac{\partial L}{\partial y'}\delta y-\int\left(\frac{\partial L}{\partial y'}\right)'\delta y\mathrm{d}x \tag{9.2.134}$$

As the limits of integration are fixed, the first term is zero. It then remains:

$$\int\frac{\partial L}{\partial y'}\delta y'\mathrm{d}x=-\int\left(\frac{\partial L}{\partial y'}\right)'\delta y\mathrm{d}x \tag{9.2.135}$$

and therefore:

$$\delta S=\int \delta L(y,y')\mathrm{d}x=\int\left(\frac{\partial L}{\partial y}\delta y'-\left(\frac{\partial L}{\partial y'}\right)'\delta y\right)\mathrm{d}x=\int\left(\frac{\partial L}{\partial y}-\left(\frac{\partial L}{\partial y'}\right)'\right)\delta y\mathrm{d}x \tag{9.2.136}$$

The minimum we sought corresponds to $\delta S=0$ whatever δy, which imposes the condition:

$$\left(\frac{\partial L}{\partial y'}\right)'-\frac{\partial L}{\partial y}=0 \tag{9.2.137}$$

we fall back on the Euler-Lagrange equation.

This equation can also be written in another form. By introducing the canonical momentum to simplify:

$$p=\frac{\partial L}{\partial y'} \tag{9.2.138}$$

Then we get immediately:

$$\left(\frac{\partial L}{\partial y'}\right)'-\frac{\partial L}{\partial y}=0\Rightarrow p'-\frac{\partial L}{\partial y}=0\Rightarrow p'=\frac{\partial L}{\partial y} \tag{9.2.139}$$

We then get:

$$\mathrm{d}L=\frac{\partial L}{\partial y}\mathrm{d}y+\frac{\partial L}{\partial y'}\mathrm{d}y'=p'\mathrm{d}y+p\mathrm{d}y' \tag{9.2.140}$$

Thus, putting the analogy seen earlier above (Hamilton's method):

$$H(y,p)=py'-L \tag{9.2.141}$$

We arrive at:

$$\mathrm{d}H=\mathrm{d}(y'p-L)=\mathrm{d}\left(y'\frac{\partial L}{\partial y'}-L\right)=0 \tag{9.2.142}$$

Therefore, remembering that at the beginning we had:

$$S=2\pi\int\sqrt{1+y'^2}\mathrm{d}x=2\pi\int L(y,y')\mathrm{d}x \tag{9.2.143}$$

we arrive at:

$$y'\frac{\partial L}{\partial y'}-L=\frac{yy'^2}{\sqrt{1+y''2}}-y\sqrt{1+y'^2}=\frac{-y}{\sqrt{1+y'^2}}=c^{te} \tag{9.2.144}$$

What we can also note (because the constant has an undetermined sign):

$$\frac{-y}{\sqrt{1+y'^2}}=-c^{te} \tag{9.2.145}$$

Then we have:

$$\frac{y}{\sqrt{1+y'^2}}=c^{te} \tag{9.2.146}$$

We have already integrated this type of differential equation in detail in the section of Civil Engineering during the study of the chain shape (see page 8558). The result is:

$$y=\frac{1}{c^{te}}\cosh\left(\frac{x}{c^te}+c_1^{te}\right)+c_2^{te} \tag{9.2.147}$$

the surface of revolution of this curve being a catenoid:

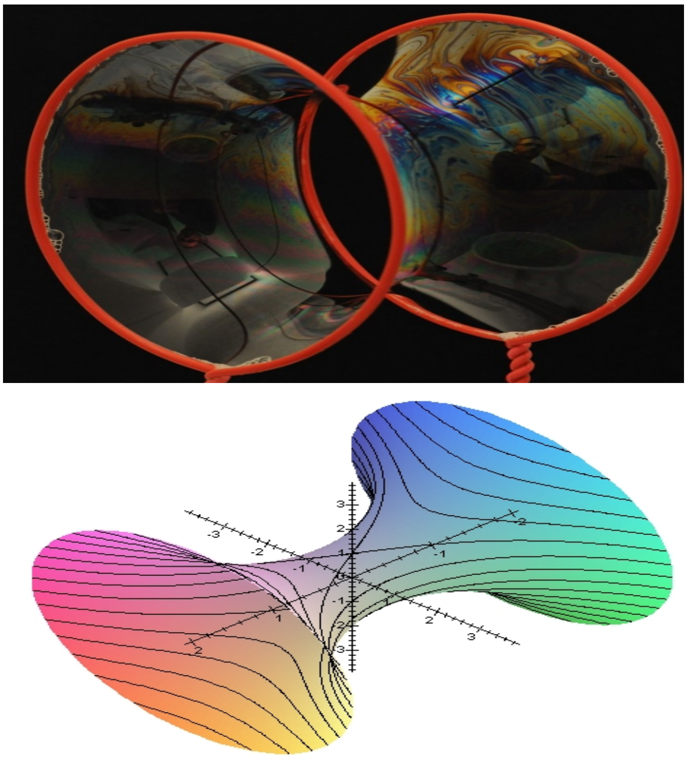


Figure 9.13: **Catenoid made of soap between two circular rings and with Maple™ 4.00b**

This figure can be obtained with Maple™ 4.00b as follows:

```
>y:=cosh(x);
>y:=cosh(x);plot3d([x,y*cos(phi),y*sin(phi)],x=-2..2,phi=-2*Pi..2*Pi,
numpoints = 50000,axes=normal,contours=10,style=patchcontour);
```

Now if L depends on the time (which is still often the case ...) we have as total differential:

$$\mathrm{d}L=\sum_{i=1}^{n}\frac{\partial L}{\partial q_i}\mathrm{d}q_i+\sum_{i=1}^{n}\frac{\partial L}{\partial \dot{q}_i}\mathrm{d}\dot{q}_i+\frac{\partial L}{\partial t}\mathrm{d}t=\sum_{i=1}^{n}\frac{\partial L}{\partial q_i}\mathrm{d}q_i+\sum_{i=1}^{n}p_i\mathrm{d}\dot{q}_i+\frac{\partial L}{\partial t}\mathrm{d}t \qquad (9.2.148)$$

we also calculate the total differential of $H(q_i,p_i)$ and substitute in it results previously obtained:

$$\mathrm{d}H=\sum_{i=1}^{n}q_i\mathrm{d}p_i+\sum_{i=1}^{n}p_i\mathrm{d}\dot{q}_i-\left[\sum_{i=1}^{n}\frac{\partial L}{\partial q_i}\mathrm{d}q_i+\sum_{i=1}^{n}p_i\mathrm{d}\dot{q}_i+\frac{\partial L}{\partial t}\mathrm{d}t\right]=\sum_{i=1}^{n}q_i\mathrm{d}p_i-\sum_{i=1}^{n}\frac{\partial L}{\partial q_i}\mathrm{d}q_i-\frac{\partial L}{\partial t}\mathrm{d}t \qquad (9.2.149)$$

which shows well that $H(q_i,p_i)$ is function of the q_i and of the time t.

We can also write for its total differential:

$$\mathrm{d}H=\sum_{i=1}^{n}\frac{\partial H}{\partial p_i}\mathrm{d}p_i+\sum_{i=1}^{n}\frac{\partial H}{\partial q_i}\mathrm{d}q_i+\frac{\partial H}{\partial t}\mathrm{d}t \qquad (9.2.150)$$

and as the q_i and p_i are independent we identify, by comparing the two expressions, that:

$$\boxed{\begin{aligned}\frac{\partial H}{\partial q_i}&=-\frac{\partial L}{\partial q_i}\\ \frac{\partial H}{\partial p_i}&=\dot{q}_i\\ \frac{\partial H}{\partial t}&=-\frac{\partial L}{\partial t}\end{aligned}} \tag{9.2.151}$$

These relations are extremely important because we will see and use them again in the section of Magnetostatics in Relativistic Quantum Physics and Quantum Field Theory in a somewhat more barbaric form (but beautiful too...).

The prior previous relation can be written obviously:

$$\frac{\mathrm{d}H}{\mathrm{d}t}=\sum_{i=1}^{n}\left[\frac{\partial H}{\partial p_i}\frac{\mathrm{d}p_i}{\mathrm{d}t}+\frac{\partial H}{\partial q_i}\frac{\mathrm{d}q_i}{\mathrm{d}t}\right]+\frac{\partial H}{\partial t} \tag{9.2.152}$$

Let us now consider the second term of the first member of the Euler-Lagrange equation. We have:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial}{\partial \dot{q}_i}\right)=\frac{\mathrm{d}}{\mathrm{d}t}(p_i)=\frac{\partial L}{\partial q_i}=\dot{p}_i \tag{9.2.153}$$

and therefore we get the $2n$ equations below (the first line describes velocity as a function of how energy changes with momentum, and the second line describes how force varies with the negative of the energy gradient with respect to position):

$$\left.\begin{aligned}\dot{q}_i&=\frac{\partial H}{\partial p_i} \qquad \text{i=1,2,}\ldots\text{,n}\\ \dot{p}_i&=-\frac{\partial H}{\partial q_i} \qquad \text{i=1,2,}\ldots\text{,n}\end{aligned}\right\}2n$$

These $2n$ equations are named "canonical equations of motion" and are differential equations of the first order.

So now we can write the prior previous relation after rearrangement as:

$$\boxed{\frac{\mathrm{d}H}{\mathrm{d}t}=\sum_{i=1}^{n}\left[-\frac{\partial H}{\partial q_i}\frac{\partial H}{\partial p_i}-\frac{\partial H}{\partial p_i}\right]+\frac{\partial H}{\partial t}} \tag{9.2.154}$$

Remark

The appearance of the minus sign "$-$" between the equations for the q_i and those for their conjugate moments, is named a "symplectic symmetry".

From:

$$\mathrm{d}H=\sum_{i=1}^{n}\frac{\partial H}{\partial p_i}\mathrm{d}p_i+\sum_{i=1}^{n}\frac{\partial H}{\partial q_i}\mathrm{d}q_i+\frac{\partial H}{\partial t}\mathrm{d}t \tag{9.2.155}$$

we can, on a trajectory that obeys the canonical equations, compute:

$$\frac{\mathrm{d}H}{\mathrm{d}t}=\sum_{i=1}^{n}\frac{\partial H}{\partial p_i}\mathrm{d}\dot{p}_i+\sum_{i=1}^{n}\frac{\partial H}{\partial q_i}\mathrm{d}\dot{q}_i+\frac{\partial H}{\partial t}\quad=-\sum_{i=1}^{n}\dot{p}_i\dot{q}_i+\sum_{i=1}^{n}\dot{q}_i\dot{p}_i+\frac{\partial H}{\partial t}=\frac{\partial H}{\partial t}=-\frac{\partial L}{\partial t} \tag{9.2.156}$$

If H does not depend on time then we have $\partial H/\partial t=0$, i.e.:

$$\frac{\partial}{\partial t}\left(\sum_{i=1}^{n}\dot{q}_i\frac{\partial L}{\partial \dot{q}_i}-L(q_i,\dot{q}_i)\right)=0 \tag{9.2.157}$$

then H (and L) are a "constant of movement", i.e.:

$$\sum_{i=1}^{n}\dot{q}_i\frac{\partial L}{\partial \dot{q}_i}-L(q_i,\dot{q}_i)=c^{te} \tag{9.2.158}$$

What must be understood so far is that in Newtonian mechanics, the time evolution is obtained by computing the total force being exerted on each particle of the system, and from Newton's second law, the time-evolutions of both position and velocity are computed. In contrast, in Hamiltonian mechanics, the time evolution is obtained by computing the Hamiltonian of the system in the generalized coordinates and inserting it in the Hamiltonian equations. This approach is equivalent to the one used in Lagrangian mechanics. In fact, the Hamiltonian is as we just seen the Legendre transform of the Lagrangian when holding q and t fixed and denoting p as the dual variable, and thus both approaches give the same equations for the same generalized momentum. The main motivation to use Hamiltonian mechanics instead of Lagrangian mechanics comes from the symplectic structure of Hamiltonian systems.

An example seems indispensable to us at this level of advancement of the study of the Lagrangian formalism. We will restrict ourselves to a particular case of a particle subjected to a force in one dimension. But even if this example and the related developments are simple we will find the results here in many other following sections of this textbook. It is therefore important for the reader to study and understand well what will follow (which unfortunately also requires that the contents of the section of Classical Mechanics is known by the reader).

✍**Example:**

Given a particle of mass m travelling in one dimension (say x) and subjected to a force derived from a potential such that:

$$F_x=-\frac{\partial V}{\partial x}\Leftrightarrow \ddot{x}=-\frac{1}{m}\frac{\partial V}{\partial x}=-(\vec{\nabla}V)_x \tag{9.2.159}$$

We know that its Lagrangian is:

$$L=\frac{1}{2}m\dot{x}^2-V(x) \tag{9.2.160}$$

We will have only one momentum (the linear momentum), denoted p as always, conjugated to x and defined as we also know by:

$$p=\frac{\partial L}{\partial \dot{x}} \tag{9.2.161}$$

equation that we can (that we must!) reverse (from the definition of the linear momentum):

$$\dot{x}=\frac{p}{m} \tag{9.2.162}$$

We may notice at this point that the momentum p corresponds (hazard...!!) to the component x of the elementary definition $p=mv$ (which is not always the case trivially).

As defined by the Hamiltonian then it comes:

$$H=\dot{x}(p)p-L(x,\dot{x}(p))=\frac{p}{m}p-\frac{m}{2}\left(\frac{p}{m}\right)^2+V(x)=\frac{p^2}{2m}+V(x) \tag{9.2.163}$$

That we often write in the form:

$$\boxed{H=T+V} \tag{9.2.164}$$

where T is therefore the kinetic energy expressed in function of the momentum.

Caution! A typical error when comparing side by side the hamiltonian and the lagrangian:

$$\begin{aligned}\mathrm{d}L&=\left(\frac{\partial L}{\partial q}\right)\mathrm{d}q+\left(\frac{\partial L}{\partial \dot{q}}\right)\mathrm{d}\dot{q}\\ \mathrm{d}H&=\left(\frac{\partial H}{\partial q}\right)\mathrm{d}q+\left(\frac{\partial H}{\partial p}\right)\mathrm{d}p\end{aligned} \tag{9.2.165}$$

is to forget that finding $\frac{\partial L}{\partial q}$ corresponds to wiggling q while holding $\dot{q}$ fixed. On the other hand, finding $\frac{\partial H}{\partial q}$ corresponds to wiggling q while holding p fixed. If p can be expressed as a function of $\dot{q}$ only (!!!), then these two situations coincide - however, if it also depends on q, then they do not, and the two partial derivatives are referring to two different things!

Explicitly, having $p=p(q,\dot{q})$. Then using the chain rule, we find that:

$$\mathrm{d}H=\left(\frac{\partial H}{\partial q}\right)\mathrm{d}q+\left(\frac{\partial H}{\partial p}\right)\left[\frac{\partial p}{\partial q}\mathrm{d}q+\frac{\partial p}{\partial \dot{q}}\mathrm{d}\dot{q}\right] \tag{9.2.166}$$

So, if we shift q but hold $\dot{q}$ fixed (as a special example), we find that:

$$\mathrm{d}L=\left(\frac{\partial L}{\partial q}\right)\mathrm{d}q \tag{9.2.167}$$

while:

$$\mathrm{d}H=\left[\left(\frac{\partial H}{\partial q}\right)+\left(\frac{\partial H}{\partial p}\right)\left(\frac{\partial p}{\partial q}\right)\right]\mathrm{d}q \tag{9.2.168}$$

If $L(q,\dot{q})=H(q,p(q,\dot{q}))$ as in the case of a free particle, then we would find that:

$$\mathrm{d}L=\mathrm{d}H \tag{9.2.169}$$

and that therefore:

$$\left(\frac{\partial L}{\partial q}\right)=\left(\frac{\partial H}{\partial q}\right)+\left(\frac{\partial H}{\partial p}\right)\left(\frac{\partial p}{\partial q}\right) \tag{9.2.170}$$

9.2.2.3 Poisson bracket

The Poisson bracket is an important binary operation in Hamiltonian mechanics, playing a central role in Hamilton's equations of motion, which govern the time evolution of a Hamiltonian dynamical system. This formalism and definition will be quite useful to us in the section of Wave Quantum Mechanics.

The Poisson bracket $\{A,B\}_{q,p}$ is the standard way of writing a certain operation that involves the quantities $A(q_i,p_i)$ and $B(q_i,p_i)$ and also all canonical variables (q_i,p_i) defined by:

$$\boxed{\{A,B\}_{q,p} := \sum_{i=1}^{n}\left[\frac{\partial A}{\partial q_i}\frac{\partial B}{\partial p_i}-\frac{\partial A}{\partial p_i}\frac{\partial B}{\partial q_i}\right]} \tag{9.2.171}$$

that expresses the way to travel in a field (the bracket content is zero if the two paths are equal).

We recognize here (!!!) the expression in brackets we introduced earlier above:

$$\frac{\mathrm{d}H}{\mathrm{d}t}=\sum_{i=1}^{n}\left[\frac{\partial H}{\partial p_i}\frac{\mathrm{d}p_i}{\mathrm{d}t}+\frac{\partial H}{\partial q_i}\frac{\mathrm{d}q_i}{\mathrm{d}t}\right]+\frac{\partial H}{\partial t} \tag{9.2.172}$$

Where as:

$$\boxed{\frac{\mathrm{d}H}{\mathrm{d}t}=\sum_{i=1}^{n}\left[\frac{\partial H}{\partial q_i}\frac{\partial H}{\partial p_i}-\frac{\partial H}{\partial p_i}\frac{\partial H}{\partial q_i}\right]+\frac{\partial H}{\partial t}} \tag{9.2.173}$$

From the definition of $\{A,B\}_{q,p}$ we can deduce some relatively trivial properties that will be useful to us in the sections of Wave Quantum Physics and Relativistic Quantum Physics:

P1. $\{A,B\}=-\{B,A\}$

Proof 9.1.3.

$$\begin{aligned}\{A,B\}&=\sum_{i=1}^{n}\left[\frac{\partial A}{\partial q_i}\frac{\partial B}{\partial p_i}-\frac{\partial A}{\partial p_i}\frac{\partial B}{\partial q_i}\right]=\sum_{i=1}^{n}\left[-\frac{\partial B}{\partial q_i}\frac{\partial A}{\partial p_i}+\frac{\partial B}{\partial p_i}\frac{\partial A}{\partial q_i}\right]\\&=-\sum_{i=1}^{n}\left[\frac{\partial B}{\partial p_i}\frac{\partial A}{\partial q_i}-\frac{\partial B}{\partial q_i}\frac{\partial A}{\partial p_i}\right]\\&=-\{B,A\}\end{aligned} \tag{9.2.174}$$

■ Q.E.D.

P2. $\{A,B+C\}=\{A,B\}+\{A,C\}$

Proof 9.1.4.

$$\begin{aligned}\{A,B+C\}&=\sum_{i=1}^{n}\left[\frac{\partial A}{\partial q_i}\frac{\partial(B+C)}{\partial p_i}-\frac{\partial A}{\partial p_i}\frac{\partial(B+C)}{\partial q_i}\right]\\&=\sum_{i=1}^{n}\left[\frac{\partial A}{\partial q_i}\left(\frac{\partial B}{\partial p_i}+\frac{\partial C}{\partial p_i}\right)-\frac{\partial A}{\partial p_i}\left(\frac{\partial B}{\partial q_i}+\frac{\partial C}{\partial q_i}\right)\right]\\&=\sum_{i=1}^{n}\left[\left(\frac{\partial A}{\partial q_i}\frac{\partial B}{\partial p_i}-\frac{\partial A}{\partial p_i}\frac{\partial B}{\partial q_i}\right)+\left(\frac{\partial A}{\partial q_i}\frac{\partial C}{\partial p_i}-\frac{\partial A}{\partial p_i}\frac{\partial C}{\partial q_i}\right)\right]\\&=\sum_{i=1}^{n}\left[\frac{\partial A}{\partial q_i}\frac{\partial B}{\partial p_i}-\frac{\partial A}{\partial p_i}\frac{\partial B}{\partial q_i}\right]+\sum_{i=1}^{n}\left[\frac{\partial A}{\partial q_i}\frac{\partial C}{\partial p_i}-\frac{\partial A}{\partial p_i}\frac{\partial C}{\partial q_i}\right]\\&=\{A,B\}+\{A,C\}\end{aligned} \tag{9.2.175}$$

■ Q.E.D.

P3. $\{A,BC\}=B\{A,C\}+C\{A,B\}$

Proof 9.1.5.

$$\begin{aligned}\{A,BC\}&=\sum_{i=1}^{n}\left[\frac{\partial A}{\partial q_i}\frac{\partial(BC)}{\partial p_i}-\frac{\partial A}{\partial p_i}\frac{\partial(BC)}{\partial q_i}\right]\\&=\sum_{i=1}^{n}\left[\frac{\partial A}{\partial q_i}\left(B\frac{\partial C}{\partial p_i}+C\frac{\partial B}{\partial p_i}\right)-\frac{\partial A}{\partial p_i}\left(B\frac{\partial C}{\partial q_i}+C\frac{\partial B}{\partial q_i}\right)\right]\\&=\sum_{i=1}^{n}\left[B\left(\frac{\partial A}{\partial q_i}\frac{\partial C}{\partial p_i}-\frac{\partial A}{\partial p_i}\frac{\partial C}{\partial q_i}\right)+C\left(\frac{\partial A}{\partial q_i}\frac{\partial B}{\partial p_i}-\frac{\partial A}{\partial p_i}\frac{\partial B}{\partial q_i}\right)\right]\\&=B\sum_{i=1}^{n}\left[\frac{\partial A}{\partial q_i}\frac{\partial C}{\partial p_i}-\frac{\partial A}{\partial p_i}\frac{\partial C}{\partial q_i}\right]+C\sum_{i=1}^{n}\left[\frac{\partial A}{\partial q_i}\frac{\partial B}{\partial p_i}-\frac{\partial A}{\partial p_i}\frac{\partial B}{\partial q_i}\right]\\&=B\{A,C\}+C\{A,B\}\end{aligned} \tag{9.2.176}$$

■ Q.E.D.

P4. $\{A,\{B,C\}\}+\{C,\{A,B\}\}+\{B,\{C,A\}\}=0$

Proof 9.1.6.

$$\begin{aligned}&\{A,\{B,C\}\}+\{C,\{A,B\}\}+\{B,\{C,A\}\}\\&=\sum_{i=1}^{n}\left[\partial_q A\partial_p\left(\sum_{i=1}^{n}[\partial_q B\partial_p C-\partial_p B\partial_q C]\right)-\partial_p A\partial_q\left(\sum_{i=1}^{n}[\partial_q B\partial_p C-\partial_p B\partial_q C]\right)\right]\\&+\sum_{i=1}^{n}\left[\partial_q C\partial_p\left(\sum_{i=1}^{n}[\partial_q A\partial_p B-\partial_p A\partial_q B]\right)-\partial_p C\partial_q\left(\sum_{i=1}^{n}[\partial_q A\partial_p B-\partial_p A\partial_q B]\right)\right]\\&+\sum_{i=1}^{n}\left[\partial_q B\partial_p\left(\sum_{i=1}^{n}[\partial_q C\partial_p A-\partial_p C\partial_q A]\right)-\partial_p B\partial_q\left(\sum_{i=1}^{n}[\partial_q C\partial_p A-\partial_p C\partial_q A]\right)\right]\end{aligned} \tag{9.2.177}$$

Well and here, just to not have an unreadable long and boring stuff we will prove the property for $n=1$ and assume (of course) that it is valid for all n:

$$\begin{aligned}
&\{A,\{B,C\}\}+\{C,\{A,B\}\}+\{B,\{C,A\}\}\\
&=\partial_q A\partial_p\left[\partial_q B\partial_p C-\partial_p B\partial_q C\right]-\partial_p A\partial_q\left[\partial_q B\partial_p C-\partial_p B\partial_q C\right]\\
&+\partial_q C\partial_p\left[\partial_q A\partial_p B-\partial_p A\partial_q B\right]-\partial_p C\partial_q\left[\partial_q A\partial_p B-\partial_p A\partial_q B\right]\\
&+\partial_q B\partial_p\left[\partial_q C\partial_p A-\partial_p C\partial_q A\right]-\partial_q B\partial_q\left[\partial_q C\partial_p A-\partial_p C\partial_q A\right]\\
&=\partial_q A\partial_{p,q}B\partial_p C+\partial_q A\partial_q B\partial_{p,p}C-\partial_q A\partial_{p,p}B\partial_q C-\partial_q A\partial_p B\partial_{p,q}C\\
&-\partial_p A\partial_{q,q}B\partial_p C-\partial_p A\partial_q B\partial_{q,p}C+\partial_p A\partial_{q,p}B\partial_q C+\partial_p A\partial_p B\partial_{q,q}C\\
&+\partial_q C\partial_{p,q}A\partial_p B+\partial_q C\partial_q A\partial_{p,p}B-\partial_q C\partial_{p,p}A\partial_q B-\partial_q C\partial_p A\partial_{p,q}B\\
&-\partial_p C\partial_{q,q}A\partial_p B-\partial_p C\partial_q A\partial_{q,p}B+\partial_p C\partial_{q,p}A\partial_q B+\partial_p C\partial_p A\partial_{q,q}B\\
&+\partial_q B\partial_{p,q}C\partial_p A+\partial_q B\partial_q C\partial_{p,p}A-\partial_q B\partial_{p,p}\partial_q A-\partial_q B\partial_p C\partial_{p,q}A\\
&-\partial_p B\partial_{q,q}C\partial_p A-\partial_p B\partial_q C\partial_{q,p}A+\partial_p B\partial_{q,p}C\partial_q A+\partial_p B\partial_p C\partial_{q,q}A
\end{aligned}\tag{9.2.178}$$

We have also (see section Differential and Integral Calculus page 1469) under certain conditions the property $\partial_{p,q}=\partial_{q,p}$. Therefore all the terms cancel (this is elementary algebra normally...) for finally have:

$$\{A,\{B,C\}\}+\{C,\{A,B\}\}+\{B,\{C,A\}\}\tag{9.2.179}$$

$$\tag{9.2.180}$$

expression named "Jacobi identity".

That is to say written in another way (perhaps more obvious) for binary operation × on a set S possessing a binary operation $+$ with an additive identity denoted by 0 satisfies the Jacobi identity if:

$$a\times(b\times c)+c\times(a\times b)+b\times(c\times a)=0\quad\forall a,b,c\in S\tag{9.2.181}$$

That is, the sum of all even permutations of $(a,(b,c))$ must be zero.

■ Q.E.D.

Now let us use the Poisson brackets with physical quantities! As by definition the coordinates and the linear momentum are not directly dependent, we have when we put $A=q_j$ and $B=q_k$:

$$\frac{\partial q_j}{\partial p_i}=0,\ \frac{\partial q_k}{\partial p_i}=0\tag{9.2.182}$$

Therefore:

$$\{q_k,q_j\}=\sum_{i=1}^{n}\left[\frac{\partial q_k}{\partial q_i}\frac{\partial q_j}{\partial p_i}-\frac{\partial q_k}{\partial p_i}\frac{\partial q_j}{\partial q_i}\right]=\sum_{i=1}^{n}\left[\frac{\partial q_k}{\partial q_i}\cdot 0-0\cdot\frac{\partial q_j}{\partial q_i}\right]=0\tag{9.2.183}$$

hence:

$$\{q_k,q_j\}=\{q_j,q_k\}=0\tag{9.2.184}$$

and identically:

$$\{p_k,p_j\}=\{p_j,p_k\}=0\tag{9.2.185}$$

But:

$$\{q_k,p_j\}=\sum_{i=1}^{n}\left[\frac{\partial q_k}{\partial q_i}\frac{\partial p_j}{\partial p_i}-\frac{\partial q_k}{\partial p_i}\frac{\partial p_j}{\partial q_i}\right]=\sum_{i=1}^{n}\frac{\partial q_k}{\partial q_i}\frac{\partial p_j}{\partial p_i}=\sum_{i=1}^{n}\delta_{ki}\delta_{ji}=\delta_{kj} \tag{9.2.186}$$

where for recall δ_{kj} is the Kronecker symbol defined par:

$$\delta_{kj}=\delta^{jk}=\begin{cases}1 & \text{if } k=j\\ 0 & \text{if } k\neq j\end{cases} \tag{9.2.187}$$

We have also:

$$\{p_j,q_k\}=\sum_{i=1}^{n}\left[\frac{\partial p_j}{\partial q_i}\frac{\partial q_k}{\partial p_i}-\frac{\partial p_j}{\partial p_i}\frac{\partial q_k}{\partial q_i}\right]=-\sum_{i=1}^{n}\frac{\partial p_j}{\partial p_i}\frac{\partial q_k}{\partial q_i}=-\sum_{i=1}^{n}\delta_{ji}\delta_{ki}=-\delta_{kj} \tag{9.2.188}$$

It comes therefore:

$$\boxed{\{q_k,p_j\}\neq\{p_j,q_k\}} \tag{9.2.189}$$

This implies a fairly general result that we see again in the section of Wave Quantum Physics:

$$\{q_k,p_j\}-\{p_j,q_k\}\neq 0 \tag{9.2.190}$$

when $k\neq j$.

9.2.2.4 Canonical transformations

We say that of the q_i,p_i that they are "generalized canonical variables". It is not an understatement as there is virtually no limit to what they can represent physically.

That being the case, there must be transformations between these different choices. We will denote by Q_i,P_i the new canonical variables obtained after such a transformation.

We are not surprised to see that these changes are subject to fairly strict conditions. Indeed, the q_i,p_i are generalized and obey as we have just prove above:

$$\{q_k,q_j\}=0 \qquad \{p_k,p_j\}=0 \qquad \{q_k,p_j\}=\delta_{kj} \tag{9.2.191}$$

and the canonical equations:

$$\dot{q}_i=\frac{\partial H}{\partial p_i} \qquad \dot{p}_i=-\frac{\partial H}{\partial q_i} \tag{9.2.192}$$

are form invariant.

Thus, after a transformation of the q_i,p_i to the Q_i,P_i and defining a new hamiltonian that we will denote by $K(Q_i,P_i)$ we should have:

$$\{Q_k,Q_j\}=0 \qquad \{P_k,P_j\}=0 \qquad \{Q_k,P_j\}=\delta_{kj} \tag{9.2.193}$$

and the canonical equations:

$$\dot{Q}_i = \frac{\partial K}{\partial p_i} \qquad \dot{P}_i = -\frac{\partial K}{\partial q_i} \tag{9.2.194}$$

Strictly speaking, the equations of transformation can be written:

$$\begin{aligned} Q_i &= Q_i(q_j, p_j, t) \\ P_i &= P_i(q_j, p_j, t) \end{aligned} \tag{9.2.195}$$

with $i, j = 1, 2, \ldots, n$ and must be able to reverse as the physics remains independent of the variables we use to describe it, so we can write the inverse transformations:

$$\begin{aligned} q_i &= q_i(Q_j, P_j, t) \\ p_i &= p_i(Q_j, P_j, t) \end{aligned} \tag{9.2.196}$$

with $i, j = 1, 2, \ldots, n$.

The q_i, p_i, Q_i, P_i form $4n$ variables but it is obvious that only $2n$ of them are independent.

~95 %

★★★☆☆

39 votes, 68.21%

9.3. Classical Mechanics

BEFORE introducing the study of moving solid bodies as part of Classical Mechanics (at the opposite of relativistic mechanics) also named "rational mechanics" or "Newtonian mechanics", it may seem in the natural order of things to define and study first the properties relative to their static state.

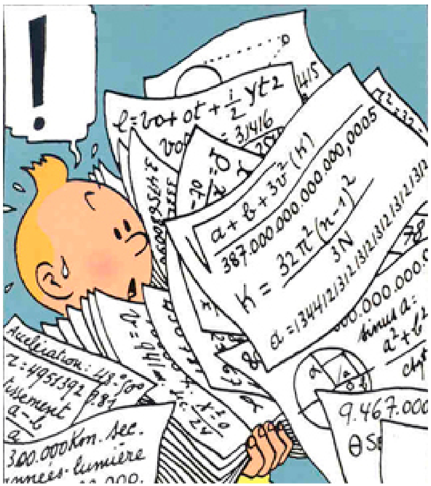


Figure 9.14: **All rights reserved ©Moulinsart**

Definitions (#9.3.424):

D1. A phenomenon is said to "static" or in "equilibrium" when it suffers of no dynamics (no acceleration or verbatim: "no unbalanced forces"), at least in appearance. We can consider an equilibrium as a static state, although even if it is apparent because it can be the result of two opposing dynamics that cancel themselves! Thus, the variables which describe a static phenomenon are constants, the concrete values of these quantities can be calculated. Static is a major case study in Mechanical engineering and Civil Engineering (see sections of the same name page 8359 and page 8539).

In a more technical way this definition is raised to the rank of principle named the "fundamental principle of static" which states that for a system to be in equilibrium, it is necessary that the general resultant and the resultant moment of the external forces is equivalent to zero relative to its center of mass or gravity (the condition is sufficient for mechanical problems that deal with non-deformable solids).

D2. The "Static" is the study of the conditions of equilibrium of a material point or body subjected to forces that are in equilibrium (cancelling themselves).

D3. Any cause that can accelerate (concept defined in detail further below) or distort a body is named a "force" (a concept introduced by Newton and rigorously to which we return in detail later in the statement of three Newton's laws).

An in-deep observation showed that a force is a macroscopic result of complex microscopic phenomena, namely interactions distance between particles. These interactions are in the number of 4 and we do not wish to talk about these 4 interactions now because they make use of mathematical tools that are out of context in this section of the book.

Remark

In Classical Mechanics we naturally do not ask ourself on the question of a transformation of time. The changes we focus on a related only to the measures "position" and its derivatives (velocity and acceleration). Indeed, in Classical Mechanics, we postulate the "Newton's time": time flows in the same way from one referential to another.

D4. If the lines of action of all the forces acting on a body are in the same plane, the system of forces named "coplanar system":

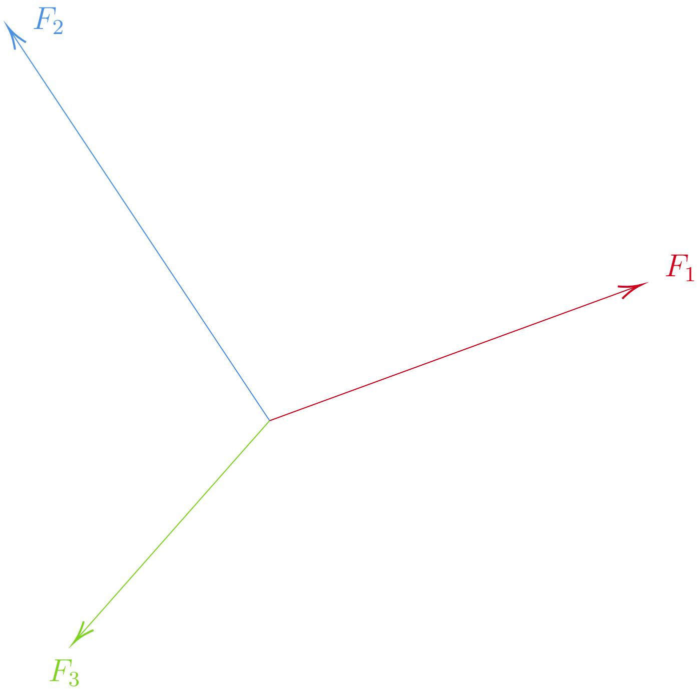


Figure 9.15: **Particular case of coplanar system**

where the intersection of forces is named a "forces node".

If we consider for example a particular set of coplanar forces whose intensities (norms) were measured using a dynamometer and angles relatively to an appropriately selected reference with a protractor. We will have a diagram of the following type named "dyname" (abbreviation of "dynamic diagram"):

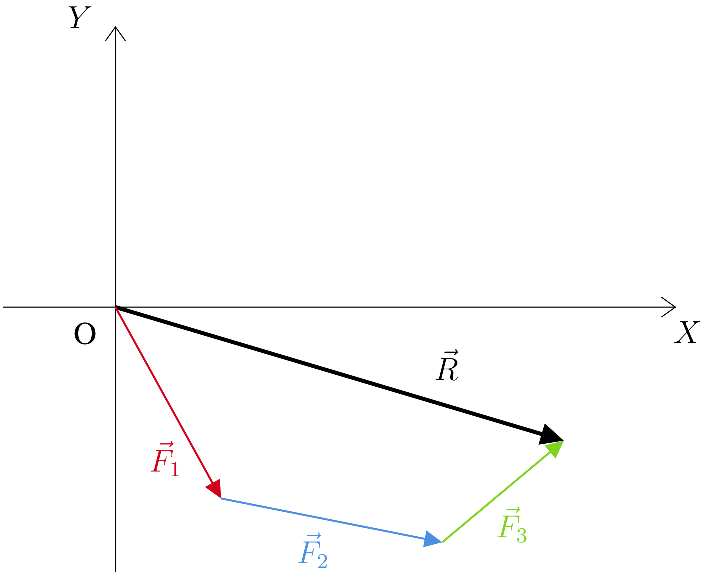


Figure 9.16: **Particular example of a measured coplanar system (dyname)**

To calculate the components of the resulting (and thus its norm), it may be easier sometimes to represent the three forces in the following form (after translation to a common point):

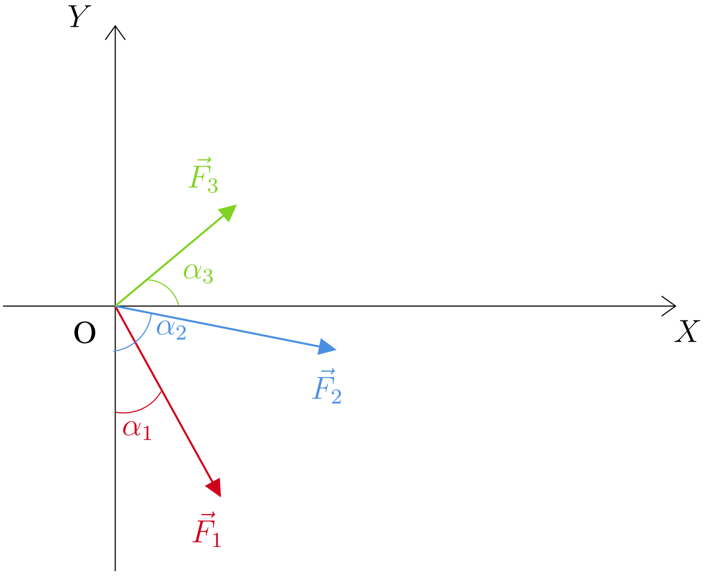


Figure 9.17: **Simplified representation of forces for treatment (translated dyname)**

Then, using elementary trigonometry, knowing the angles and the intensity of each of the three forces, it is possible to determine their respective components along the X and Y and their algebraic sum along each of the axes give the components of the resultant (today this type of reasoning sounds simple but it took still until the late 116th century (holocene calendar) that these reasonings on static forces emerges).

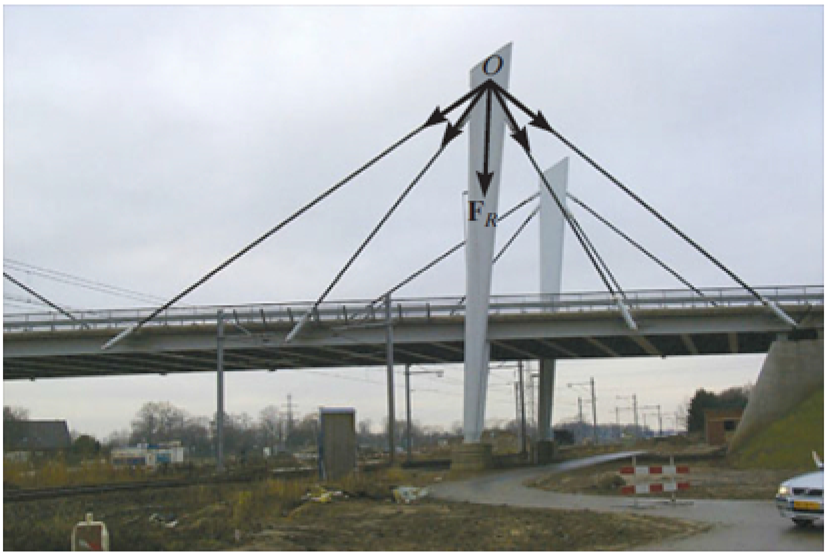


Figure 9.18: **Representation of the resultant of forces of a suspended bridge**

A small basic application case is the porting of a load:

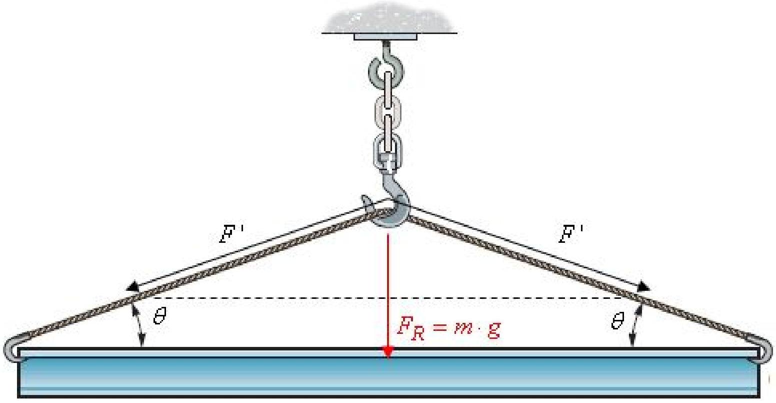


Figure 9.19: **Coplanar symmetrical load porting**

By symmetry each of the left/right part of the cable is under the same tension (force). Applying elementary trigonometry (see section Trigonometry page 2503), we have:

$$\frac{\frac{F_R}{2}}{F'} = \cos\left(\pi - \frac{\pi}{2} - \theta\right) = \cos\left(\frac{\pi}{2} - \theta\right) = \sin(\theta) \tag{9.3.1}$$

Therefore:

$$F' = \frac{\frac{F_R}{2}}{\sin(\theta)} \tag{9.3.2}$$

We notice that we have an interest in having:

$$\sin(\theta) > 0.5 \tag{9.3.3}$$

(therefore an angle greater than $\sim 30°$) otherwise we would have:

$$F' > F_R \tag{9.3.4}$$

As we have already mentioned it, a system of concurrent coplanar forces is in static equilibrium when the resultant of all forces (and moments) is zero. Graphically, the end of the last force on the dyname representation coincides with the origin of the first force. Mathematically this can be written:

$$\left.\begin{aligned}\sum_{i=1}^{n} F_{x_i}=R_x=0\\ \sum_{i=1}^{n} F_{y_i}=R_y=0\end{aligned}\right\}\Rightarrow \vec{R}=\vec{0} \qquad (9.3.5)$$

A diagram representing such forces is named a "free body diagram" and is well known by college students as looking as following:

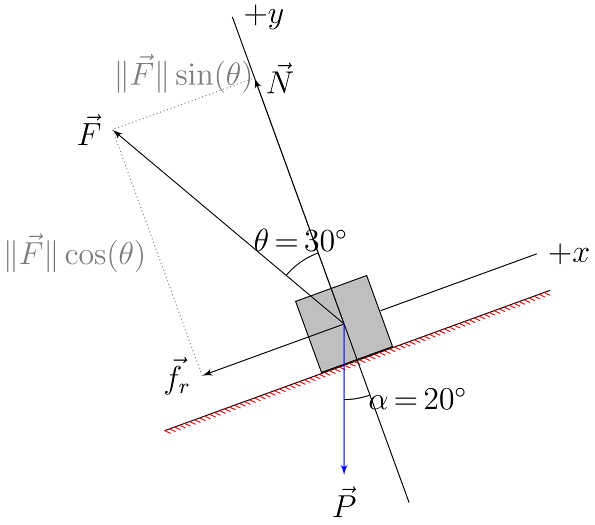
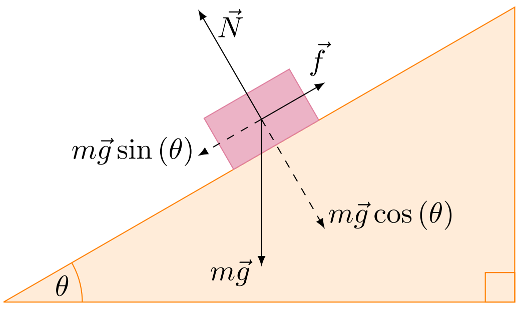


Figure 9.20: **Typical example of a free body diagram**

Or also as following:



We will meet and analyse quite a lot of such free body diagrams during the study of various subjects through this textbook (so don't be impatient to practice them...!).

D5. A material system $\mathcal{S}$ (set of material points (P_i, m_i)) is named "rigid solid" or "rigid body", or simply "solid", if the mutual distances of the constituent material points do not vary over time. What we note technically as follows:

$$\forall t, \forall (i,j): \quad \left\|\overline{P_iP_j}\right\| = c^{te} \qquad (9.3.6)$$

9.3.1 Newton's laws

Newton's three laws are the foundation for Classical Mechanics. They describe the relation between a body and the forces acting upon it, and its motion in response to those forces. These laws are retrospectively unprovable and can not be formalized (proved) because they result from observations and thus derive from our daily experience.

However, developments in modern physics and that are based on the consequences of these three laws are in such agreement with theoretical conditions that impose the principle of least action and relative experiences, that their validity could make any doubt (...).

9.3.1.1 Newton's first law (Inertia law)

Definition (#9.3.425): Any point or wide body mass perseveres in its shape (geometry) or its state of rest or of uniform motion (described by the center of mass), unless (in a static space-time[11]!) if "applied forces" force bring the body to change its shape or its movement.

$$\boxed{\vec{\Sigma} F=\vec{0} \Leftrightarrow \vec{v}=\vec{c^{te}}} \tag{9.3.7}$$

The punctual body is obviously dimensionless. It is a creation of the mind, a model representing a physical object that is animated as a translational motion (no rotation on itself because it is dimensionless!). We assume here that our physical space is three dimensional to which we add time that is not a spatial dimension in the context of Classical Mechanics but an immutable and independent parameter!

In other words: Every body at rest or in uniform motion is under pressure of a given number of zero forces or of a sum of forces whose resultant is zero (that is the fundamental principle of statics also named "principle of inertia")!

Corollary: When the trajectory of a body is not a straight line or when the speed of this body is not constant, we can conclude from the Principle of Inertia that the forces acting on this body do not cancel each other out.

Remarks

R1. We have proved this corollary in our study of Noether's theorem in the section of Analytical Mechanics at page 3132.

R2. Some people criticize the principle of inertia by arguing that there are not such possible experimental configuration in the Universe where a body has zero forces (or a zero resultant) on it as there is gravitation everywhere and that latter is a force! But these people forget that Albert Einstein has proved that gravity is NOT a force (as we will prove it during our study of General Relativity on page 5720.

In this corollary (and also before), comes to us in full light the word "force $\vec{F}$". Let us questioning this word for which everyday language is full of different meanings: wrist force, soul's force, think force... Also, the force can be visible or not, measurable or not depending on the case... Anyway, it has the "power" to change the path (the movement) and shape (geometry) of things. Without ignoring this halo that surrounds the word that has puzzled more than a physicist before him, Newton gives the force a very precise meaning, that stands out from the intuitive idea of a physical effort.

---

[11] Energy isn't conserved in an expanding or contracting space as we will prove it during our study of Noether theorem (see page 3136).

Let us give some intuitive properties of a force:

P1. The force is a vector quantity (see section Vector Calculus page 1797) with corresponding mathematical properties.

P2. The effect of a force, does not change if we slide the force on his line of action.

P3. The force is proportional to the mass.

Caution! Strictly speaking, force is not a vector. Like momentum, it is a covector or one-form, and covariant. You can see this in several ways:

- From the principle of virtual work: force is a linear function mapping infinitesimal displacements $\delta\vec{x}$ (a vector) to infinitesimal changes in energy $F\delta\vec{x}$ (a scalar) and hence a covector by definition.
- From Newton's second law $\vec{F}=m\vec{a}$ we see that the acceleration is a vector, which is "index-lowered" by the mass to give the covector that is the force.
- Conservative forces arise from the differential of potential energy $F=-\mathrm{d}U$, and the differential of a function is a one-form (covariant).

The difference between a vector and covector may not make sense if you're just starting to learn about physics, and for now, knowing that forces can be "added tip to tail" like vectors may be enough for practical calculations. But it is something you should start to pay attention to as your understanding matures: like dimensional analysis, carefully keeping track of what your physical objects are, mathematically, is helpful both for building deeper understanding, and catching errors.

A force is therefore a physical quantity manifested by its effects:

E1. "Dynamic effect": a force is capable of producing a cause or alter the movement of a body.

E2. "Static effect: a force is a cause capable of producing a deformation of the shape (geometry) of a body.

Any force can be represented by a vector whose four properties are at least:

P1. "Direction": the line in that the action takes place (this include the angle).

P2. "Orientation": The way the line in which the action takes place is followed.

P3. "Application point": the point where the action is applied on the body.

P4. "Intensity": the amplitude (norm) of the force

The photo below is an example of Newton's first law (an object in motion or at rest tends to stay in motion or at rest). The object in this picture is the man, and even though his bike stopped, he kept moving forward (...):

It is possible to classify most of the forces by families such as:

F1. The "forces of reaction": each body applies a force on another body that is in contact with himself. For example, if an object is putted on a table, this table applied (most of time...) an equal and opposite force on the object (so that it does not sink into the table - that are quantum mechanisms that are responsible for this reaction force). This force is always vertical to the point of contact.

F2. The "frictional forces" the frictional force exists when two bodies are in contact. It always opposes to the movement. The friction force which opposes to the movement has not only has a negative effect, it is also essential to ensure contact between two surfaces (e.g. contact of tires with the road, brake, etc.).

F3. The "tension forces": it is a force that pulls on an element of a body such as the tension exerted by a thread or a spring (see section Mechanical Engineering page 3416).

F4. The "action-at-a-distance forces": these are the forces that act through vector fields such as electric field, magnetic field or gravitational field (types of forces that result even when the two interacting objects are not in physical contact with each other, yet are able to exert a push or pull despite their physical separation). This last type of common force has the particularity, if the field is isotropic, to be reduced to the study of the body's center of gravity (we will prove it further below during our study of static forces).

Remark

As far as we know the are no official conventions for a finite number of properties for forces and for categories of forces. In some books the authors will tell you that there are only two type of forces: Contact Forces (frictional forces, tension forces, normal forces, air resistance forces, applied forces, spring forces, etc.) and Action-at-a-Distance Forces (gravitational force, electric force, magnetic force).

9.3.1.2 Newton's second law (Fundamental Principle of Dynamics)

Definition (#9.3.426): The change in motion is proportional to the "hard driving force", and is carried along the line through which this force is printed. Or more precisely: The acceleration of an object as produced by a net force is directly proportional to the magnitude of the net force, in the same direction as the net force, and inversely proportional to the mass of the object.

A force, we know it, is in the language of Newton what causes the "change of the movement" and not something else...! But, supplement to the program, the words "change of motion" of the law hide a mathematical meaning, different from the intuition "variation in speed". For Newton, we saw that a body at rest was characterized by its quantity of matter, its mass. Inspired by some predecessors, Newton put that a body in motion "carries a certain amount", simply named at his time: the "linear momentum". This is actually this amount which, under the simple word "movement", is contained in the statement of the second law. The amount of movement is the measure that we derive from both speed (concept which we will define later in our study of the kinematics) and quantity of matter, that is, by definition, the product of its mass by its speed:

$$\boxed{\vec{p}=m\vec{v}} \tag{9.3.8}$$

The reader may have noticed that linear momentum is as for the force, a convector and not a vector!

Caution! The linear momentum relation $p=mv$ doesn't apply to particles without mass. However such particles can transfer momentum and we will explain and prove that during our study of Special Relativity (see page 5627).

Using modern mathematical symbols, the first part of this second law can then be reformulated:

The force is equal to the variation with time of the amount of movement or in a non-relativistic framework:

$$\vec{F}=\frac{\mathrm{d}\vec{p}}{\mathrm{d}t}=\frac{\mathrm{d}(m\vec{v})}{\mathrm{d}t}\cong m\frac{\mathrm{d}(\vec{v})}{\mathrm{d}t}=m\vec{a} \tag{9.3.9}$$

This latter relation is valid until the speed is much lower than that of light as we will see in our study of Special Relativistic much later, as Newton supposed that the mass did not vary (or do not seem to vary...) as a function of speed. Thus, the "fundamental relation of dynamics (FRD)" is also given by:

$$\boxed{\vec{F}=m\vec{a}} \tag{9.3.10}$$

and can be stated as follows: Given a constant mass m of a body, the acceleration experienced by a body in an inertial frame is proportional to the net force it undergoes, and inversely proportional to its mass m in a non relativistic framework. The law addresses the cause and effect relationship between force and motion and the classical sense (as far as we know, in this early 121st century - holocene calendar -, in many Quantum Physics effects, causality doesn't exist as we will prove it during our study of the subject later). So it may be useful that we define now these both concepts in the context of our modern physical comprehension of the observable Universe.

Remark

Notice the interesting following relationship with the delta Dirac function (see section Differential and Integral Calculus page 1569):

$$\Delta\int\limits_{-\infty}^{+\infty} F(t)\mathrm{d}t=\int_{-\infty}^{+\infty} mv\delta(t)\mathrm{d}t \tag{9.3.11}$$

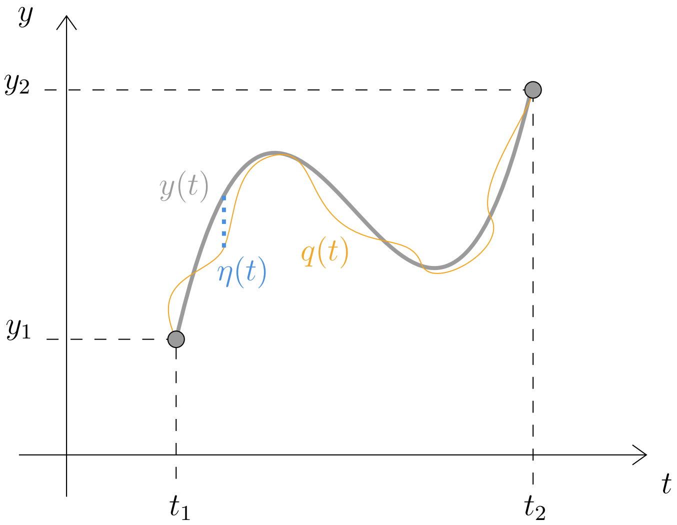
Typical apparatus (PASCO® scientific) used in high schools to verify Newton's second law with a dynamic system and a smart cart:

PASCO wireless smart cart

Let us see now that we can prove $F=ma$ using the Maupertuis principle (i.e. principle of least action) as introduced in the section of Analytical Mechanics at page 3152.

For this we imagine a body travelling a path that can be describe by a function of time $y(t)$. We add at every moment t a small perturbation $\eta(t)$, let us note this new perturbated path:

$$q(t)=y(t)+\eta(t) \tag{9.3.12}$$



And since the variations are small, the difference in action between these two paths is 0, i.e.:

$$\delta S = S[q(t)] - S[y(t)] = 0 \tag{9.3.13}$$

Our next task is to solve this equation. So we compute the action for each path. For this, we need the kinetic and potential energy for each, and we write them as a function of y and η.

We have for that path $q(t)$:

$$\begin{aligned} T(q) &= \frac{1}{2}mv^2 & V &= V(q) \\ &= \frac{1}{2}m\left(\frac{\mathrm{d}q}{\mathrm{d}t}\right)^2 & V(q) &= V(y+\eta) \\ &= \frac{1}{2}m\left(\frac{\mathrm{d}}{\mathrm{d}t}(y+\eta)\right)^2 & V(q) &= V(y)+\eta V'(y)+\eta^2V''(y)+\ldots \\ &= \frac{1}{2}m\left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2+m\frac{\mathrm{d}y}{\mathrm{d}t}\frac{\mathrm{d}\eta}{\mathrm{d}t}+\frac{1}{2}m\left(\frac{\mathrm{d}\eta}{\mathrm{d}t}\right)^2 & V(q) &\cong V(y)+\eta V'(y) \\ &\cong \frac{1}{2}m\left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2+m\frac{\mathrm{d}y}{\mathrm{d}t}\frac{\mathrm{d}\eta}{\mathrm{d}t} \end{aligned} \tag{9.3.14}$$

and for the path $y(t)$:

$$\begin{aligned} T(y) &= \frac{1}{2}mv^2 & V = V(y) \\ &= \frac{1}{2}m\left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2 \end{aligned} \tag{9.3.15}$$

Therefore for the path $q(t)$ we have:

$$\begin{aligned}S[q(t)]&=\int\limits_{t_1}^{t_2}(T(q)-V(q))\mathrm{d}t\\&=\int\limits_{t_1}^{t_2}\left(\frac{1}{2}m\left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2+m\frac{\mathrm{d}y}{\mathrm{d}t}\frac{\mathrm{d}\eta}{\mathrm{d}t}-V(y)-\eta V'(y)\right)\mathrm{d}t\\&=\int\limits_{t_1}^{t_2}\left(\frac{1}{2}m\left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2-V(y)\right)\mathrm{d}t+\int\limits_{t_1}^{t_2}\left(m\frac{\mathrm{d}y}{\mathrm{d}t}\frac{\mathrm{d}\eta}{\mathrm{d}t}-\eta V'(y)\right)\mathrm{d}t\end{aligned} \tag{9.3.16}$$

and for the path $y(t)$:

$$S[y(t)]=\int\limits_{t_1}^{t_2}(T-V)\mathrm{d}t=\int\limits_{t_1}^{t_2}\left(\frac{1}{2}m\left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2-V(y)\right)\mathrm{d}t \tag{9.3.17}$$

Hence:

$$\begin{aligned}&\delta S=S[q(t)]-S[y(t)]\\&=\int\limits_{t_1}^{t_2}\left(\frac{1}{2}m\left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2-V(y)\right)\mathrm{d}t+\int\limits_{t_1}^{t_2}\left(m\frac{\mathrm{d}y}{\mathrm{d}t}\frac{\mathrm{d}\eta}{\mathrm{d}t}-\eta V'(y)\right)\mathrm{d}t-\int\limits_{t_1}^{t_2}\left(\frac{1}{2}m\left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2-V(y)\right)\mathrm{d}t\\&\int\limits_{t_1}^{t_2}\left(m\frac{\mathrm{d}y}{\mathrm{d}t}\frac{\mathrm{d}\eta}{\mathrm{d}t}-\eta V'(y)\right)\mathrm{d}t=0\end{aligned} \tag{9.3.18}$$

We can rewrite the first term in the parenthesis this further by using integration by parts:

$$\int\limits_{t_1}^{t_2}m\frac{\mathrm{d}y}{\mathrm{d}t}\frac{\mathrm{d}\eta}{\mathrm{d}t}\mathrm{d}t=m\frac{\mathrm{d}y}{\mathrm{d}t}\eta\Bigg|_{t_1}^{t_2}-\int\limits_{t_1}^{t_2}m\frac{\mathrm{d}^2y}{\mathrm{d}t^2}\eta\mathrm{d}t \tag{9.3.19}$$

The first right hand term cancel as at the starting point t_1 we have η that is zero and same for the final point t_2 where η is zero. Therefore:

$$\int\limits_{t_1}^{t_2}m\frac{\mathrm{d}y}{\mathrm{d}t}\frac{\mathrm{d}\eta}{\mathrm{d}t}\mathrm{d}t=-\int\limits_{t_1}^{t_2}m\frac{\mathrm{d}^2y}{\mathrm{d}t^2}\eta\mathrm{d}t \tag{9.3.20}$$

Therefore:

$$\delta S=S[q(t)]-S[y(t)]=\int\limits_{t_1}^{t_2}\left(\left[-m\frac{\mathrm{d}^2y}{\mathrm{d}t^2}-V'(y)\right]\eta\right)\mathrm{d}t=0 \tag{9.3.21}$$

But since η can be just about anything, this can only be true if the parenthesis is zero:

$$-m\frac{\mathrm{d}^2y}{\mathrm{d}t^2}-V'(y)=0 \tag{9.3.22}$$

So what we found is that the action is minimal for the path that satisfies this curious differential equation. Now it might look complicated, but it's not.

Now it might look complicated, but it's not! Minus the derivative of the potential is just the force. And the second derivative of height? Well, that's just acceleration. Such that:

$$-ma+F(y)=0 \tag{9.3.23}$$

So if we rearrange this, we find that the path that satisfies the principle of least action is the one that obeys:

$$F=ma \tag{9.3.24}$$

In other words, the principle of least action is equivalent to Newton's second law!

Definitions (#9.3.427):

D1. A "cause" (also referred to as "causation", "causality", or "cause and effect[12]") is the influence by which one event, process or state (a cause) contributes to the production of another event, process or state (an effect) where the cause is totally or partly responsible for the effect, and the effect is totally or partly dependent on the cause.

D2. An "effect" is a change which is a result or consequence of an action, a physical cause or of an immaterial probability[13].

Caution! Newton's second law of motion is at the heart of Classical Mechanics. Like other stuffs, it is insubstantial. Yet clearly Newton's second law is full of meaning, by the highest standard: It proves itself useful in demanding situations. Splendid, unlikely looking bridges, like the Erasmus Bridge (known as the Swan of Rotterdam), do bear their loads; spacecraft do reach Saturn. The paradox deepens when we consider force from the perspective of modern physics. In fact, the concept of force is conspicuously absent from our most advanced formulations of the basic laws as we will prove it in the next chapters. It doesn't appear in the Schrödinger's equation, or in any reasonable formulation of Quantum Field Theory, or in the foundations of general relativity...

Remember that the "mass" is a measure for the quantity of material contained in a body (see section Principia page 3087). The mass is a quantity independent of where it is located (S.I.

---

[12] *Post hoc ergo propter hoc* (latin of "after this, therefore because of this:) is an informal fallacy that states: "Since event Y followed event X, event Y must have been caused by event X." It is often shortened simply to *post hoc fallacy*. A logical fallacy of the questionable cause variety, it is subtly different from the fallacy *cum hoc ergo propter hoc* ("with this, therefore because of this"), in which two events occur simultaneously or the chronological ordering is insignificant or unknown. Post hoc is a logical fallacy in which an event says to be the cause of a later event because it occurred earlier!

[13] All speculative Quantum Causal Models created so far until year 12017 (holocene calendar) have failed to find a deterministic causality to some Quantum effects that have well explained probabilistic behaviours! Thus it follows that the Kalâm cosmological argument fails because its first premise fails (but it also fails for the reason as we have already seen during our study of Logic that if in case there is a premise, either implicit or explicit, that is logically equivalent to the conclusion then the reasoning is said to be "circular". In many cases, circularity is a problem and we name such circularity a "vicious circularity".

unit kilogram: [kg]). The "weight" (the oldest trace differentiating mass and weight seems to date from the early 117st century by Newton), corresponds to the force (S.I. unit Newton: [N]) that an object applied on another one through an action (or field). The latter depends on where we are located (see below the equation of the gravitational force of Newton).

We will see (prove) that in the context of a body falling in a gravitational field with a spherical symmetry, we have see section Astronomy page 5255):

$$\vec{F}=m\vec{a}=m\left(\frac{GM}{R^2}\frac{\vec{R}}{\|\vec{R}\|}\right) \tag{9.3.25}$$

as part of our good old Earth, we have the habit to write:

$$\|\vec{a}\|=g \tag{9.3.26}$$

In an Eulerian system in Cartesian coordinates, a given quantity φ of a continuous medium will have a distribution based on four independent variables x, y, z, t. For small changes $\mathrm{d}x$, $\mathrm{d}y$, $\mathrm{d}z$ and $\mathrm{d}t$, the total variation φ being expressed by (see section Differential and Integral Calculus page 1466):

$$\mathrm{d}\varphi=\frac{\partial\varphi}{\partial x}\mathrm{d}x+\frac{\partial\varphi}{\partial y}\mathrm{d}y+\frac{\partial\varphi}{\partial z}\mathrm{d}z+\frac{\partial\varphi}{\partial t}\mathrm{d}t \tag{9.3.27}$$

By following a particle in its motion, we observe during a time $\mathrm{d}t$ displacements $\mathrm{d}x$, $\mathrm{d}y$, $\mathrm{d}z$. We can therefore express from the previous relation of the total variation of φ for the time $\mathrm{d}t$. We obtain therefore the expression of a very important derivative in theoretical physics named the "particular derivative":

$$\frac{\mathrm{d}\varphi}{\mathrm{d}t}=\frac{\partial\varphi}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t}+\frac{\partial\varphi}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t}+\frac{\partial\varphi}{\partial z}\frac{\mathrm{d}z}{\mathrm{d}t}+\frac{\partial\varphi}{\partial t} \tag{9.3.28}$$

In Classical Mechanics we will particularly work with the Newtonian gravitational field. Therefore, the relation linking the force to the acceleration takes a more general form. Let's see how!

Given the particular derivative of the velocity (for the three spatial coordinates):

$$\frac{\mathrm{d}\vec{v}}{\mathrm{d}t}=\frac{\partial\vec{v}}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t}+\frac{\partial\vec{v}}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t}+\frac{\partial\vec{v}}{\partial z}\frac{\mathrm{d}z}{\mathrm{d}t}+\frac{\partial\vec{v}}{\partial t} \tag{9.3.29}$$

Which can also be written:

$$\frac{\mathrm{d}\vec{v}}{\mathrm{d}t}=\frac{\partial\vec{v}}{\partial t}+v_x\frac{\partial\vec{v}}{\partial x}+v_y\frac{\partial\vec{v}}{\partial y}+v_z\frac{\partial\vec{v}}{\partial z} \tag{9.3.30}$$

Which can also be written in condensed form:

$$\frac{\mathrm{d}\vec{v}}{\mathrm{d}t}=\frac{\partial\vec{v}}{\partial t}+(\vec{v}\circ\vec{\nabla})\vec{v}=\left(\frac{\partial}{\partial t}+\vec{v}\circ\vec{\nabla}\right)\vec{v} \tag{9.3.31}$$

The Newton's second law is written then in generalized form:

$$\boxed{\vec{F}=m\left(\frac{\partial}{\partial t}+\vec{v}\circ\vec{\nabla}\right)\vec{v}} \tag{9.3.32}$$

where as we will see it during our study of fluid dynamics (see section Continuous Mechanics page 3830):

$$\vec{v} \circ \vec{\nabla} \tag{9.3.33}$$

is the "advection term".

This formulation of Newton's second law is of utmost importance in physics. It makes explicit account of the force experienced by a material point in a vector field depending on the speed and not only of the position. We will use this formulation in the section of Continuum Mechanics in our study of fluids and plasmas (see page 3945), in the section of Electromagnetism and in that of General Relativity (see page 5720)!

We can see that if $\vec{\nabla}(\vec{v}) = \vec{0}$, that is to say if the gradient of the velocity field is zero (i.e. the velocity field is constant or equal to zero - the latter case being typical for any experience outside a force field like a gravitational one for example!), then we fall back on the special case of the simple second Newton law as:

$$\vec{F} = m\left(\frac{\partial}{\partial t} + \vec{v} \circ \vec{\nabla}\right)\vec{v} = m\frac{\partial}{\partial t} + \underbrace{\vec{v} \circ \vec{\nabla}(\vec{v})}_{=0} = m\frac{\partial \vec{v}}{\partial t} = m\vec{a} \tag{9.3.34}$$

As we will prove during our study of Fluid Mechanics, the generalized Newton's second law can be written:

$$\vec{F} = m\frac{\partial \vec{v}}{\partial t} + m\vec{v} \circ (\vec{\nabla}\vec{v}) = m\frac{\partial \vec{v}}{\partial t} + m\frac{1}{2}\vec{\nabla} \cdot v^2 + (\vec{\nabla} \times \vec{v}) \times \vec{v} \tag{9.3.35}$$

In the case of a normal uniform gravitation field the last term vanish ("vorticity" as we will see it again during our study of Fluid Mechanics at page 3837), then it remains:

$$\vec{F} = m\frac{\partial \vec{v}}{\partial t} + m\frac{1}{2}\vec{\nabla} \cdot v^2 \tag{9.3.36}$$

That is to say in the one dimensional case this reduces to:

$$F = ma + \frac{\partial}{\partial x}\left(\frac{1}{2}mv^2\right) \tag{9.3.37}$$

And as we will see later the last term between parenthesis is the Kinetic Energy. Therefore:

$$F = ma + \frac{\partial}{\partial x}E_c \tag{9.3.38}$$

But as variation of the kinetic energy is in absolute value equal to that of the variation of potential energy, denoted E_p and sometimes U, on the same distance as we will prove later then we have:

$$\boxed{F = ma + \frac{\partial}{\partial x}U} \tag{9.3.39}$$

That means once again that if we are far away from any force field (gravitational one for example in Classical Mechanics), the latter relation reduce to the simple Newton's second law again as $U = 0$:

$$F = ma + \frac{\partial}{\partial x}\underbrace{U}_{=0} = ma \tag{9.3.40}$$

so if we see a body that accelerates it can only be due to an impulse (a direct applied force on it).

If there is nobody or nothing entering in collision on with our body or applying an impulse on it but we see that our body accelerates, then it mean that we are in the presence of a force field and the generalized Newton's second law reduced to:

$$F=m\underbrace{a}_{=0}+\frac{\partial}{\partial x}U=\frac{\partial}{\partial x}U \tag{9.3.41}$$

That is implicitly... a force due to a gravitational field (we still omit the case of electric field at this level).

But of course we can have the both terms that are non-zero! That is when a body accelerates because of an impulse plus because of a force field like the gravitation field (typically a rocket entering into orbit in a far away planet).

9.3.1.3 Newton's third law (Law of Action and Reaction)

Definition (#9.3.428): In every two-body static interaction, there is a pair of forces acting on the two interacting objects. The size of the forces on the first object equals the size of the force on the second object. The direction of the force on the first object is opposite to the direction of the force on the second object. Forces always come in pairs - equal and opposite action-reaction force pairs! This third law is most known under the name of "action / reaction principle" and follows from the first Newton law by mathematical through the reasoning that we have already seen during our study of Noether's theorem in the section Principles.

This law can be simply expressed in the following form:

$$\vec{F}_{12}=\vec{F}_{21} \tag{9.3.42}$$

Also sometimes written as:

$$\vec{F}_1=-\vec{F}_2 \qquad \text{or} \qquad \vec{F}_2=-\vec{F}_1 \tag{9.3.43}$$

and in a little bit more general and explicit way as following:

$$\boxed{\sum\vec{F}_{1\to 2}=\sum\vec{F}_{2\to 1}} \tag{9.3.44}$$

Thus, in nature, according to this principle, there is no isolated force, each force has its "opposite", they work in pairs!

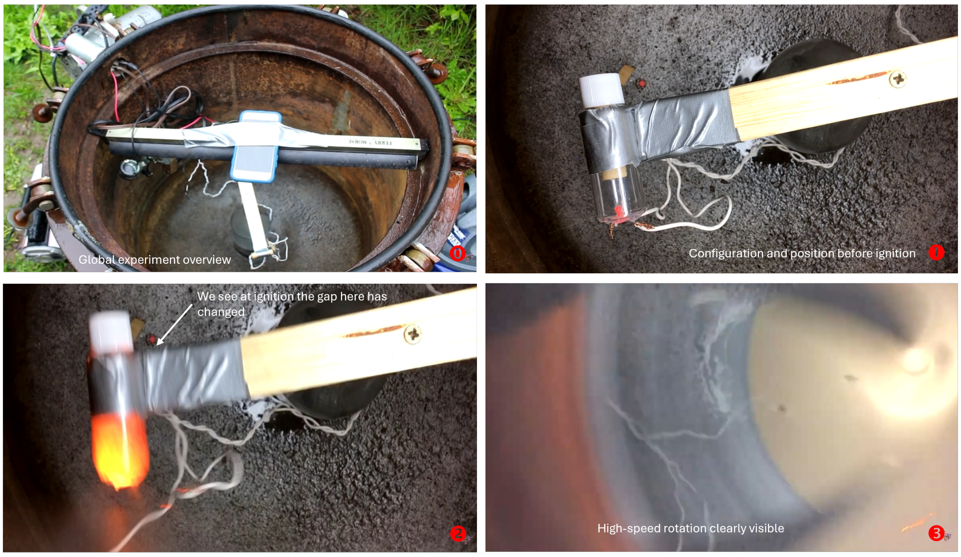
There is also a nice easy experiment (among many others) done by Cody Reeder that profs beyond reasonable doubt that a thruster also works in vacuum. Using his *Leader Iron Works Inc.* 189 [L] (50 gallons) vacuum/pressure vessel:

LEADER IRON WORKS INC.

that once connected to a vacuum pump looks like:

A coarse vacuum is created in the vessel of $-600\,[\text{mmHg}]$ (i.e. of $-0.789474\,[\text{atm}]$) so we have only 22% of the ambient normal pressure (corresponding approximately to the atmosphere pressure at an altitude of $11,000\,[\text{m}]$):

Now attaching an Estes 1/2A3-4T model rocket motor ($13\,[\text{mm}]$ engine) to a wooden board an igniting it we get:


Global experiment overview
⓪
Configuration and position before ignition
①
We see at ignition the gap here has changed
②
High-speed rotation clearly visible
③

and after ignition the coarse vacuum has obviously changed a little bit:

The same kind of experiment is visible in the video here https://www.youtube.com/watch?v=T8MOoUuLnug with a soda can hanged on a small wire instead of a miniature rocket thruster with still a coarse vacuum but however a better one of $-0.9\,[\text{atm}]$ corresponding approximately to an altitude of $16,000\,[\text{m}]$ (and we can see the action/reaction even before the can liquid hits the walls).

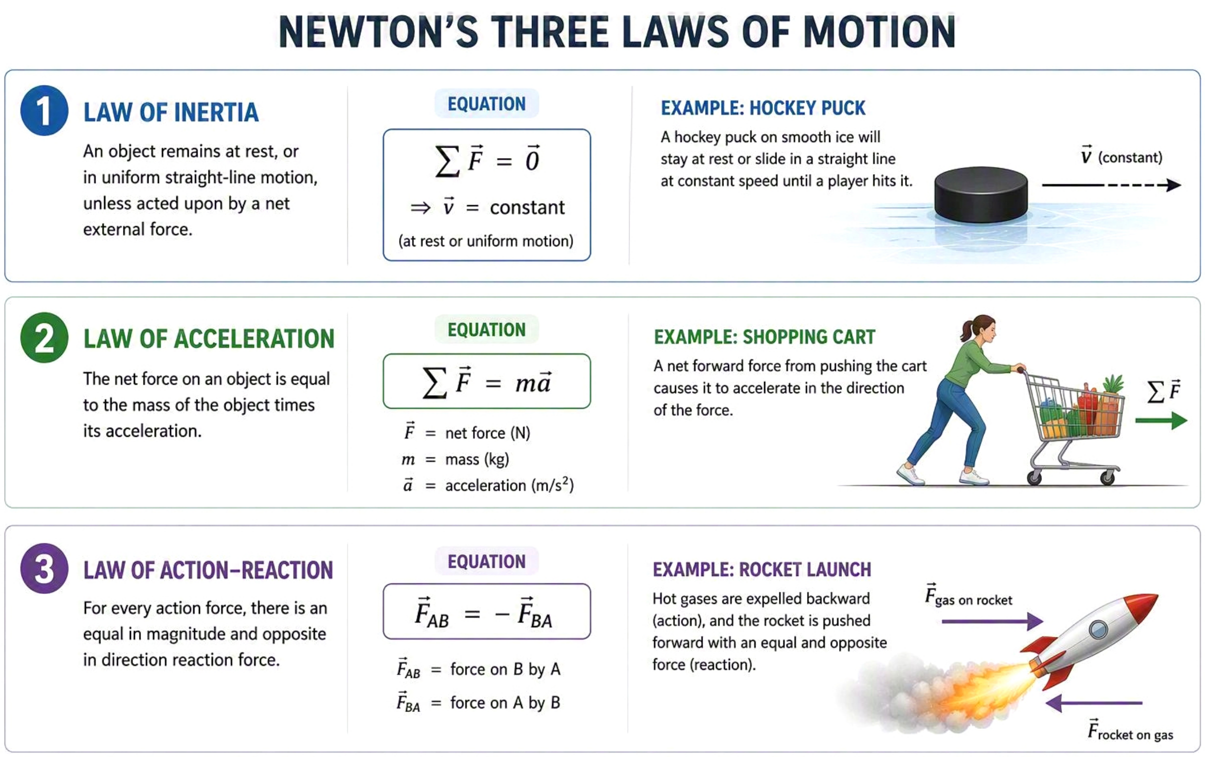
We can also say that two solid points or extended bodies in static contact exert on each other always opposing forces in direction but equal in intensity and direction.

It follows an important property of the third Newton law: For a material point, subject to forces $\vec{F}_1,\vec{F}_2,\ldots,\vec{F}_n$ to be in static equilibrium the resultant of the forces must be equal to zero. That is the material point:

$$\sum_{i=1}^{n}\vec{F}_i=\vec{0} \tag{9.3.45}$$

satisfies the "equilibrium condition".

The previous relation, which thus defines any body in equilibrium, opens the study of huge field of practical applications names "static forces" and that we will develop further below after having introduced the concept of force momentum.



NEWTON'S THREE LAWS OF MOTION

① **LAW OF INERTIA**

An object remains at rest, or in uniform straight-line motion, unless acted upon by a net external force.

EQUATION

$$\sum \vec{F} = \vec{0}$$

$$\Rightarrow \vec{v} = \text{constant}$$

(at rest or uniform motion)

EXAMPLE: HOCKEY PUCK

A hockey puck on smooth ice will stay at rest or slide in a straight line at constant speed until a player hits it.

② **LAW OF ACCELERATION**

The net force on an object is equal to the mass of the object times its acceleration.

EQUATION

$$\sum \vec{F} = m\vec{a}$$

$\vec{F}$ = net force (N)

m = mass (kg)

$\vec{a}$ = acceleration (m/s^2)

EXAMPLE: SHOPPING CART

A net forward force from pushing the cart causes it to accelerate in the direction of the force.

③ **LAW OF ACTION–REACTION**

For every action force, there is an equal in magnitude and opposite in direction reaction force.

EQUATION

$$\vec{F}_{AB} = -\vec{F}_{BA}$$

$\vec{F}_{AB}$ = force on B by A

$\vec{F}_{BA}$ = force on A by B

EXAMPLE: ROCKET LAUNCH

Hot gases are expelled backward (action), and the rocket is pushed forward with an equal and opposite force (reaction).

$\vec{F}_{\text{gas on rocket}}$
$\vec{F}_{\text{rocket on gas}}$

9.3.2 Center of Mass and Reduced Mass

The center of mass is a special case of the barycenter with all its properties that we have already largely developed in the section of Euclidean Geometry (therefore we strongly recommend to reader to refer to it at page 2635) but applied to physics.

We can make no distinction between "center of mass" and "center of gravity" (also named "centroid") if and only if the mass of the body is considered homogeneous.

For those who want a simple but impressive example of the importance of understanding the concept of center of gravity in the context of engineering and business, I recommend searching the Internet for video on the keyword "Steadicam". This is a technique of moving the center of gravity of cameras for image stabilization that has revolutionized the way of filming in the movie industry (and not only!). Or playing with center of mass can also be an art:

Definition (#9.3.429): Let us consider a solid formed of n mass points $m_1, m_2, \ldots, m_n$ and identified by their respective position vectors $\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_n$. We name "center of mass" (or "center of inertia" if there is strict equality between gravitational mass and inertial mass) a point G where we can connect all the mass of the system (and therefore its analysis !!) and such that, the origin being arbitrarily chosen it is given by:

$$\vec{r}_G = \frac{\sum\limits_{i=1}^{n} \vec{r}_i m_i}{\sum\limits_{i=1}^{n} m_i} \tag{9.3.46}$$

Relation that has to be compared by the reader with the one of the barycenter that we get in the section of Euclidean Geometry:

$$\overrightarrow{\text{O}G} = \frac{\sum\limits_{i=1}^{n} \alpha_i \overrightarrow{\text{O}A_i}}{\sum\limits_{i=1}^{n} \alpha_i} \tag{9.3.47}$$

In an identical way, we define the "reduced mass" of the system by:

$$\boxed{\mu = \frac{\prod\limits_{i=1}^{n} m_i}{\sum\limits_{i=1}^{n} m_i}} \tag{9.3.48}$$

The reduced mass is the "effective" inertial mass appearing in the two-body problem of Newtonian Mechanics. It is a quantity which allows the two-body problem to be solved as if it were a one-body problem.

Theorem 9.2. Given two bodies, one with mass m_1 and the other with mass m_2, the equivalent one-body problem, with the position of one body with respect to the other as the unknown, is that of a single body of mass:

$$\mu=\frac{1}{\dfrac{1}{m_1}+\dfrac{1}{m_2}}=\frac{m_1m_2}{m_1m_2} \tag{9.3.49}$$

Proof 9.2.1. This relation can be derived using Newton's second law, i.e. the force exerted by body 2 on body 1:

$$\vec{F}_{12}=m_1\vec{a}_1 \tag{9.3.50}$$

and the force exerted by body 2 on body 1 is:

$$\vec{F}_{21}=m_2\vec{a}_2 \tag{9.3.51}$$

According to Newton's third law, the force that body 2 exerts on body 1 is equal and opposite to the force that body 1 exerts on body 2:

$$\vec{F}_{12}=\vec{F}_{21} \tag{9.3.52}$$

Therefore:

$$m_1\vec{a}_1=-m_2\vec{a}_2 \tag{9.3.53}$$

and:

$$\vec{a}_2=-\frac{m_1}{m_2}\vec{a}_1 \tag{9.3.54}$$

The relative acceleration $\vec{a}_{\text{rel}}$ between the two bodies is given by:

$$\vec{a}_{\text{rel}}=\vec{a}_1-\vec{a}_2=\left(1+\frac{m_1}{m_2}\right)\vec{a}_1=\frac{m_2+m_1}{m_1m_2}m_1\vec{a}_1=\frac{\vec{F}_{12}}{m_{\text{red}}} \tag{9.3.55}$$

So we conclude that body 1 moves with respect to the position of body 2 as a body of mass equal to the reduced mass.

■ Q.E.D.

If we consider the solid as continuous (true only at the microscopic scale in a first approximation) therefore it comes:

$$\vec{r}_G=\frac{\displaystyle\int_V \vec{r}\mathrm{d}m}{\displaystyle\int_V \mathrm{d}m} \tag{9.3.56}$$

Where the integrals are extended to whole solid volume.

Furthermore, if the solid is homogeneous (special case), of volumetric mass ρ, then $\mathrm{d}m = \rho \mathrm{d}V$, $\mathrm{d}V$ being the volume element. The equation can then be written (the notation of the triple integral is reduced to only one integral by for clarity purpose):

$$\boxed{\vec{r}_G = \frac{\displaystyle\int_V \vec{r}\mathrm{d}V}{\displaystyle\int_V \mathrm{d}V}} \tag{9.3.57}$$

That is to say in components:

$$x_G = \frac{\displaystyle\int x\mathrm{d}V}{V} \qquad y_G = \frac{\displaystyle\int y\mathrm{d}V}{V} \qquad z_G = \frac{\displaystyle\int z\mathrm{d}V}{V} \tag{9.3.58}$$

Properties:

P1. If the solid has an axis of symmetry, then G is on that axis.

P2. If the solid has a plane of symmetry, then G is on this plane.

P3. If the solid has several axes of symmetry, then G is at their intersection.

Remark

The center of mass M may be out of the solid (eg: a stool, a boomerang, etc.).

It is not always obvious to calculate the center of mass of a relatively simple given body. It's not that the mathematical tools to handle are complex (simple integral, Pythagoras theorem and some multiplication and integration by parts) but we must approach the problem in an elegant way and if we do not immediately the right approach we will fail very quick. We advise the teachers that address this topic and the related exercises to do it with their students (i.e. in class) but by leaving them to discuss how the teacher must attack the problem on the blackboard (it works very well).

9.3.2.1 Center of Mass Theorem

Under the action of external forces $\vec{F}_1, \ldots, \vec{F}_i, \ldots, \vec{F}_n$, acting at each point of a solid, each of these points takes an acceleration equal to the applied force $\vec{F}_i$. Using Newton's second law for each point and summing the effects we have (in a non-relativistic case):

$$\vec{F}_{\text{ext}} = \sum_{i=1}^{n} F_i = \sum_{i=1}^{n} F_i = \sum_{i=1}^{n} m_i \frac{\mathrm{d}^2\vec{r}_i}{\mathrm{d}^2 t} = \frac{\mathrm{d}^2}{\mathrm{d}t^2} \sum_{i=1}^{n} m_i \vec{r}_i \tag{9.3.59}$$

by the position of the center of mass given by the relation:

$$\vec{r}_G = \frac{\displaystyle\sum_{i=1}^{n} \vec{r}_i m_i}{\displaystyle\sum_{i=1}^{n} m_i} \tag{9.3.60}$$

In the special case of a binary system (used a lot in Astronomy and Corpuscular Quantum physics as we will see it) we have:

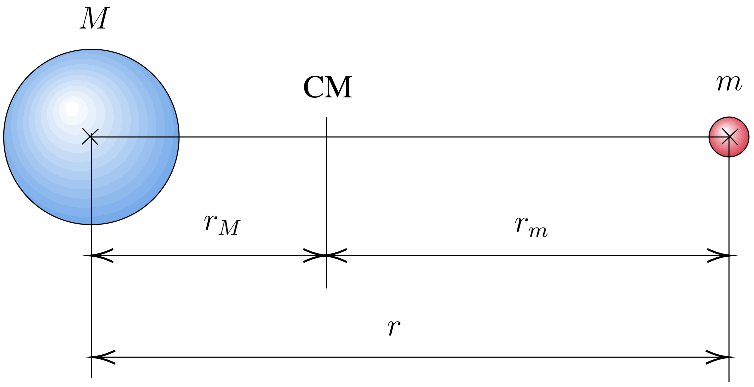


Figure 9.21: **Binary system center of mass (profile view)**

Where:

$$\vec{r}_G=\frac{M\vec{r}_M+m\vec{r}_m}{M+m} \tag{9.3.61}$$

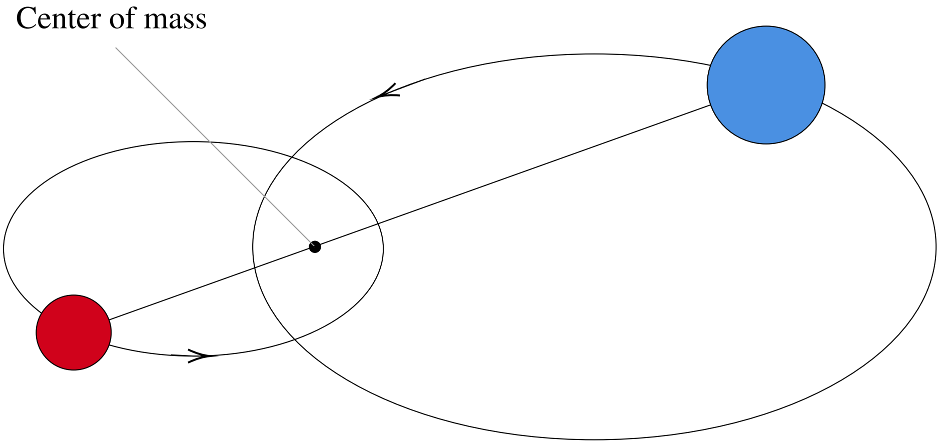


Figure 9.22: **Binary system center of mass when in motion**

In the special case where $M=m$ we have obviously:

$$\vec{r}_G=\frac{m\vec{r}_M+m\vec{r}_m}{2m}=\frac{\vec{r}_M+\vec{r}_m}{2} \tag{9.3.62}$$

Therefore the center of mass (CM) is indeed at the middle point!

If we set the position of the center of mass at the origin ($\vec{r}_G=\vec{0}$). Then we have in the case where $M\neq m$:

$$M\vec{r}_M+m\vec{r}_m=\vec{0} \tag{9.3.63}$$

and then we have taking the norm:

$$\frac{M}{m}=-\frac{r_m}{r_M} \tag{9.3.64}$$

In other words, the distances r_M and r_M can change, but the ratio r_m/r_M is fixed. Also, the center of mass always lies on the line connecting the two stars.

To come back to the general case, it comes if the repository is placed on the center of mass:

$$\vec{F}_{\text{ext}}=\frac{\mathrm{d}^2\vec{r}_G}{\mathrm{d}t^2}\sum_{i=1}^{n}m_i \tag{9.3.65}$$

where:

$$\sum_{i=1}^{n}m_i=M \tag{9.3.66}$$

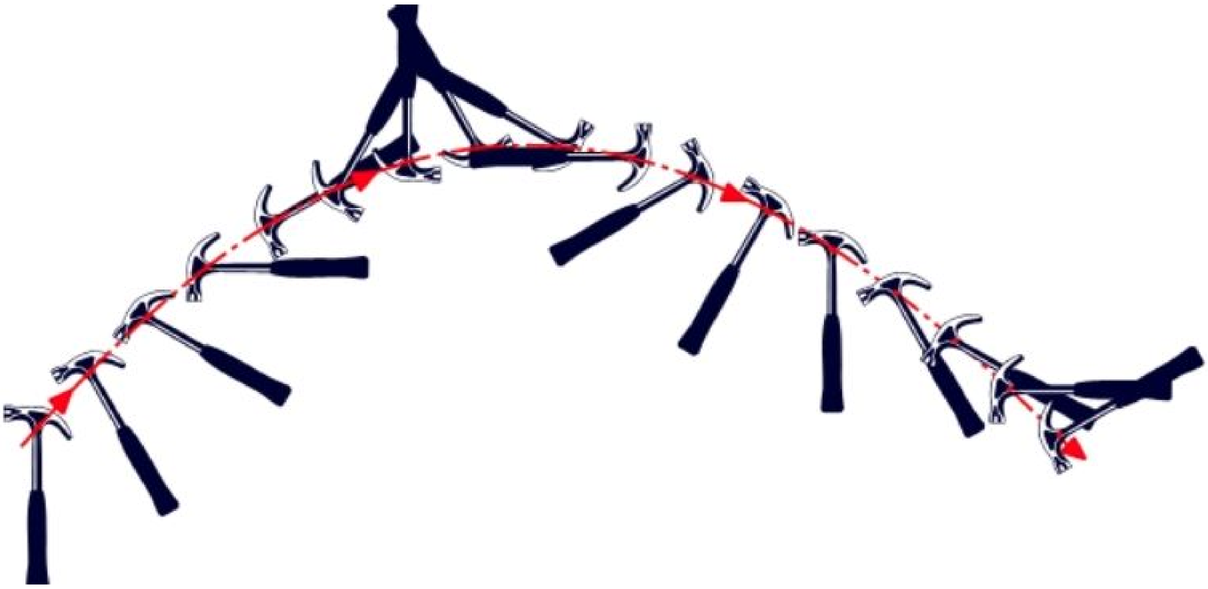
therefore:

$$\boxed{\vec{F}_{\text{ext}}=M\frac{\mathrm{d}\vec{r}_G}{\mathrm{d}t}=M\vec{a}_G} \tag{9.3.67}$$

This is the theorem of the center of mass that we can state as follows:

The center of mass of a solid moves as a material point of mass equal to that of the solid and on which would be applied the sum of external forces. A simple example is that of an explosive projectile in absence of gravity describing a curved path. If the projectile explodes and breaks, the shards of the center of mass continues to describe the curved path he had started.

Or another famous example:



Remark

In the particular case of a solid (set of points) subjected to the field of gravity, $\vec{F}_{\text{ext}}$ is the weight of the solid and then G is named the "center of gravity" (hence the origin of this name).

Let us come back on the relation:

$$\vec{r}_G=\frac{\sum\limits_{i=1}^{n}\vec{r}_i m_i}{\sum\limits_{i=1}^{n} m_i} \tag{9.3.68}$$

giving the position of center of mass. Its speed is:

$$\vec{v}_G=\frac{\mathrm{d}\vec{r}_G}{\mathrm{d}t}=\frac{1}{M}\sum_i\frac{\mathrm{d}\vec{r}_i}{\mathrm{d}t}=\frac{1}{M}\sum_i m_i\vec{v}_i=\frac{1}{M}\sum_i\vec{p}_i \tag{9.3.69}$$

By denoting:

$$\vec{P}=\sum_i\vec{p}_i \tag{9.3.70}$$

where $\vec{P}$ is the linear momentum of the system, it comes:

$$\vec{P}=M\vec{v}_G \tag{9.3.71}$$

This relation shows that if the sum of the external forces is zero then:

$$\vec{F}_{\text{ext}}=\frac{\mathrm{d}\vec{P}}{\mathrm{d}t}=\vec{0}\Rightarrow\vec{P}=\vec{c}^{te}\Rightarrow\vec{v_G}=\vec{c}^{te} \tag{9.3.72}$$

Thus the amount of movement of the entire system is conserved and movement of the system center of mass is unaltered. This justifies the remarks made during the study of the conservation of momentum.

In the study of interactions between particles, it is often convenient to use a reference frame linked to the center of mass of all particles. This center of mass is at rest in that frame, its speed is zero and also the total linear momentum, as shown by the above equations. This property is the powerful benefit of this description.

Remark

In mechanics, the use of the center of mass (material point) is particularly easy because the system of forces is only governed by the Newton's law. With electric charged particles, it is quite different. Electromagnetic effects are dominant in their acceleration, which induces interactive wave phenomena significantly more complex. That is why we will never see a study in this textbook of "charge center" when we will get in the section of Electrostatic or Electrodynamics...

✍**Example:**

We want to determine how far the center of mass of the Earth-Moon system is from the center of Earth by ignoring the other objects in our Solar System.

We define the origin of the coordinate system as the center of Earth. Then, with just two objects, then the center of mass theorem will be written:

$$r_G=\frac{\sum\limits_{i=1}^{2} r_i m_i}{\sum\limits_{i=1}^{2} m_i}=\frac{M_{\text{Earth}} r_{\text{Earth}}+M_{\text{Moon}} r_{\text{Moon}}}{r_{\text{Earth}}+r_{\text{Moon}}} \tag{9.3.73}$$

But as we choose the center of Earth as the origin, we then have $r_{\text{Earth}}=0$ [m]. Inserting this into the equation for r_G gives:

$$r_G=\frac{M_{\text{Moon}} r_{\text{Moon}}}{r_{\text{Earth}}+r_{\text{Moon}}} \cong \frac{7.36 \cdot 10^{22} \cdot 3.82 \cdot 10^5}{5.97 \cdot 10^{24}+7.36 \cdot 10^{22}} \cong 4.64 \cdot 10^6 \text{ [m]} \tag{9.3.74}$$

The radius of Earth is $6.37 \cdot 10^6$ [m] , so the center of mass of the Earth-Moon system is approximately $1,730$ [km] below the surface of Earth.

9.3.2.2 Guldin's Theorem

The "Guldin's Theorems" or also named "Pappus's Centroid Theorem" gives the possibility in some cases to simplify the calculation of the center of mass of some bodies (or reciprocally the volume or surface when the coordinates of the center of mass is known).

1. First Theorem:

   Theorem 9.3. The volume V of a solid of revolution generated by rotating a plane figure F about an external axis is equal to the product of the area S of F and the distance d travelled by its geometric centroid.

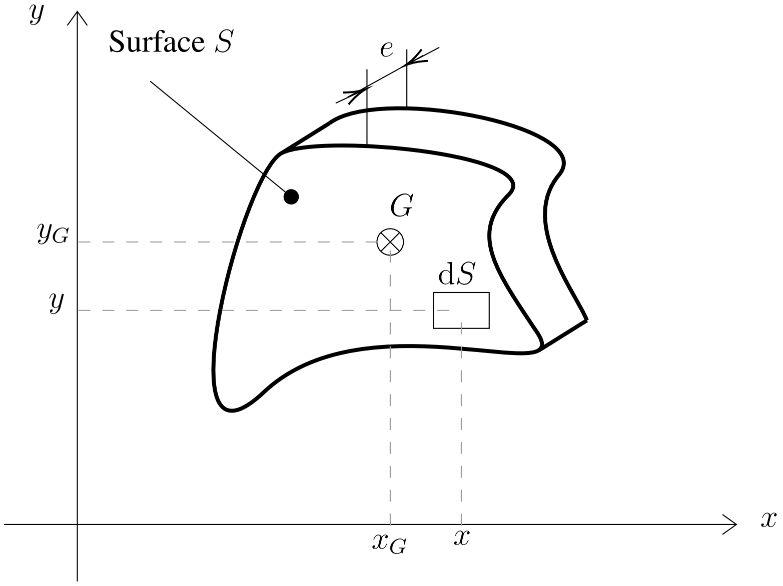
   Proof 9.3.1. Given a flat plate, homogeneous, of constant thickness e, of volumetric density ρ placed in a Cartesian plane $x\text{O}y$. We know that we have compared to the y-axis:

$$y_G=\frac{\displaystyle\int \mathrm{d}m}{\displaystyle\int m}=\frac{\displaystyle\int y\mathrm{d}m}{M}=\frac{e\rho\displaystyle\int y\mathrm{d}S}{e\rho S}=\frac{\displaystyle\int \mathrm{d}S}{S} \tag{9.3.75}$$

   Therefore:

$$y_G\cdot S=\int \mathrm{d}S \tag{9.3.76}$$

   Let us consider now the following figure:



   Let us consider a rotation of $2\pi y$ around the x axis of small surface located on y following the vertical axis. The element of volume described by a surface element $\mathrm{d}S$ during this rotation is obviously equal to:

$$\mathrm{d}V=2\pi y\mathrm{d}S \tag{9.3.77}$$

and therefore, the total volume described by the complete surface S is using the previous relation:

$$V=\int \mathrm{d}V=2\pi\int \mathrm{d}S=2\pi y_G S \tag{9.3.78}$$

Thus, by doing the same for x_G, we finally get:

$$y_G=\frac{V}{2\pi S} \qquad x_G=\frac{V}{2\pi S} \tag{9.3.79}$$

■ Q.E.D.

✍**Example:**

For example, the volume of the torus (see section Geometric Shapes page 2848) with minor radius r and major radius R is:

$$V=(\pi r^2)(2\pi R)=2\pi^2 Rr^2 \tag{9.3.80}$$

2. Second Theorem (or first depends on the point of view...):

Theorem 9.4. The surface area S of a surface of revolution generated by rotating a plane curve C about an axis external to C and on the same plane is equal to the product of the arc length L of C and the distance d travelled by its geometric centroid.

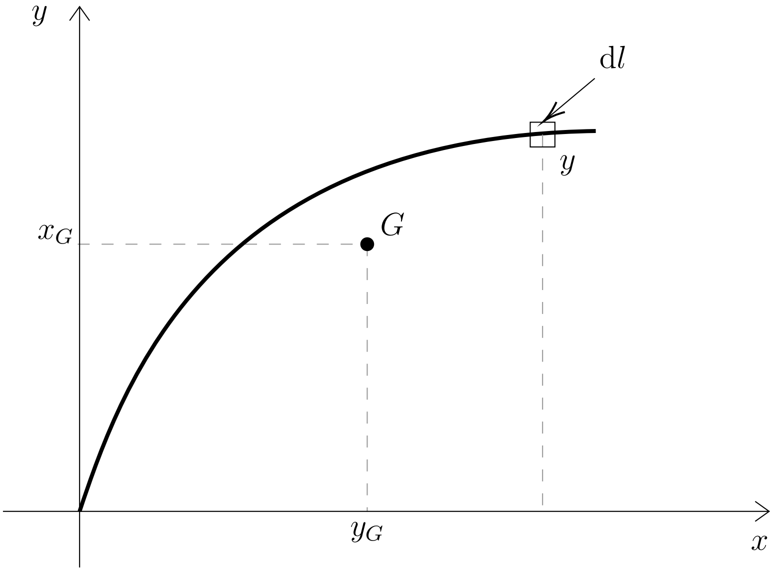
Proof 9.4.1. Let us consider a curved path or string, homogeneous, of length L, of constant section and of linear mass density ρ_l. We know that we have compared to the y-axis:

$$y_G=\frac{\displaystyle\int y\mathrm{d}m}{\displaystyle\int \mathrm{d}m}=\frac{\rho_l\displaystyle\int y\mathrm{d}l}{\rho_l L}=\frac{\displaystyle\int y\mathrm{d}l}{L} \tag{9.3.81}$$

and therefore:

$$y_G L=\int y\mathrm{d}l \tag{9.3.82}$$

Let us consider now the following figure:



Let us consider a rotation around the x axis. The surface described by an element of length $\mathrm{d}l$ during this rotation is equal to:

$$\mathrm{d}V=2\pi\mathrm{d}l \tag{9.3.83}$$

and therefore the total area described by the string of length L is:

$$S=\int\mathrm{d}S=2\pi\int y\mathrm{d}l=2\pi y_G L \tag{9.3.84}$$

Thus, by doing the same for x_G, we finally get:

$$y_G=\frac{S}{2\pi L}\qquad x_G=\frac{S}{2\pi L} \tag{9.3.85}$$

■ Q.E.D.

✍**Example:**

For example, the surface area of the torus (see section Geometric Shapes page 2848) with minor radius r and major radius R is:

$$A=(2\pi r)(2\pi R)=4\pi^2 Rr \tag{9.3.86}$$

9.3.3 Kinematics of Rectilinear Motion

If a phenomenon is evolving, by observing it, we see a shift from the concrete value of one or more characteristic variables. These values are obviously not constant but variable. An evolution implies that there is a beginning, an infinite number of intermediate states and an end.

Definition (#9.3.430): A "state" is the description of a snapshot of an evolving phenomenon (not necessarily in the temporal sense).

The functional relation between variables for a given state can be described by an equation. For an evolutionary phenomenon, there may be an infinite number of states that we can describe by as many equations. In this form, it has no interest. We are looking for a single equation that relates the different quantities verifying all the states that the evolutionary phenomenon may admit. By this equation, we can then calculate any state of the evolutionary phenomenon studied: this is the "equation of state" (a concept taken from the section of Thermodynamics).

Therefore kinematics is the branch of Classical Mechanics which describes the motion of points (alternatively "particles"), bodies (objects), and systems of bodies without consideration of the masses of those objects nor the forces that may have caused the motion.

Kinematics begins with a description of the geometry of the system and the initial conditions of known values of the position, velocity and or acceleration of various points that are a part of the system, then from geometrical arguments it can determine the position, the velocity and the acceleration of any part of the system. The study of the influence of forces acting on masses falls within the purview of kinetics.

Kinematics is obviously used in wide range of fields like astrophysics to describe the motion of celestial bodies and collections of such bodies. In mechanical engineering, robotics, and biomechanics kinematics is used to describe the motion of systems composed of joined parts (multi-link systems) such as an engine, a robotic arm or the skeleton of the human body.

We consider mainly three sub-domains of kinematics:

1. "Rectilinear kinematics" or "Translation kinematics" investigates laws of motion of objects along straight line without any reference to forces that cause the motion to change.
2. "Curvilinear kinematics" investigates lows of motion of objects in space in two and three directions without any reference to forces that cause the motion to change.
3. "Rotational kinematics" investigates lows of motion of objects along circular path without any reference to forces that cause the motion to change.

9.3.3.1 Position

Definition (#9.3.431): The "position" of an object is defined by its position vector which in the particular case of a three-dimensional space is given by:

$$\vec{r}=(x,y,z) \tag{9.3.87}$$

Obviously each coordinate of a moving object may vary in function of time as:

$$\vec{r}=(x(t),y(t),z(t)) \tag{9.3.88}$$

Rather than this notation rather heavy with parentheses ... physicists frequently note the position vector (or vector space) in the form of a vector of 4 dimensions (three spatial components and one time component) and then we write:

$$\vec{r}=(x,y,z,t) \tag{9.3.89}$$

or simply denoted by:

$$\vec{r}(x,y,z,t) \tag{9.3.90}$$

and we then name it a "space-time four-vector" whose components are the generalized coordinates of the system.

9.3.3.2 Velocity and speed

Definition (#9.3.432): The "scalar velocity", also named "speed", denoted v, is defined as the distance travelled by an object during a given amount of time:

$$v=\frac{d}{t} \tag{9.3.91}$$

The SI unit for speed is obviously meter per second: $[\text{m}\cdot\text{s}^{-1}]$.

The latter must not be confused with the "velocity" that is the vectorial version of the speed defined as the displacement travelled by an object during a given amount of time:

$$\vec{v}=\frac{\vec{r}}{t} \tag{9.3.92}$$

✍**Example:**

A vehicle start its journey from a city A to go to a city B at a speed of $60\,[\text{km}\cdot\text{h}^{-1}]$ during one hour and comes back following the same way during the same duration of one hour. Therefore its average speed is:

$$\vec{v}=\frac{\Delta x}{\Delta t}=\frac{\text{total distance}}{\text{elapsed time}}=\frac{120}{2}=60\,[\text{km}\cdot\text{h}^{-1}] \tag{9.3.93}$$

and the average velocity is:

$$v=\frac{\vec{x}_{\text{final}}-\vec{x}_{\text{initial}}}{\Delta t}=\frac{\vec{0}-\vec{0}}{2}=0\,[\text{km}\cdot\text{h}^{-1}] \tag{9.3.94}$$

When a body is in uniform motion, that is to say, it travels a given distance following a dimension x_i with $i=\{1,2,3\}$ in a time always equal, the previous ratio is constant in time:

$$v_i=\frac{x_i}{t}=c^{te} \tag{9.3.95}$$

The "arithmetic average speed" or "average linear speed" is defined as the ratio of the distance between a given starting point $x_{i,0}$ at instant t_0 and an end point $x_{i,f}$ at a time t_f:

$$\bar{v}_i=\frac{x_{i,f}-x_{i,0}}{t_f-t_0}=\frac{x_i(t_f)-x_i(t_0)}{t_f-t_0}=\frac{\Delta x}{\Delta t}=c^{te} \tag{9.3.96}$$

Remark

You must take care when calculation the average speeds of as there are several types of averages in mathematics... (see section Statistics page 546)! For example, it is common to use the harmonic mean speed as we prove it in the section Statistics.

This represents an average (because we are not interested to know how the path between $x_{i,f}$ and $x_{i,0}$ has been travelled) but not the instantaneous speed of the vehicle at any given time.

If we want to know the speed named "instantaneous speed" of the vehicle at a point of its trajectory must take the limit of the delta time Δt to a differential element $\mathrm{d}t$ (see section Differential and Integral Calculus page 1464) as:

$$\mathrm{d}t=t_f-t_0=(t_0+\varepsilon)-t_0 \tag{9.3.97}$$

with $\varepsilon\to 0$.

Mathematically, we note this correctly as follows:

$$\mathrm{d}t=\lim_{\varepsilon\to 0}(t_0+\varepsilon)-t_0 \tag{9.3.98}$$

Thus, during an infinitely small time difference, the distance difference will also be infinitesimal. So we have:

$$\mathrm{d}x_i=\lim_{\varepsilon\to 0}(x_{i,0}+\varepsilon)-x_{i,0} \tag{9.3.99}$$

and finally:

$$\boxed{v_i=\frac{\mathrm{d}x_i}{\mathrm{d}t}=\dot{x}} \tag{9.3.100}$$

The junior engineer or scientist must be careful when calculating the average velocity. Indeed as we have prove it in the section of Statistics, the harmonic average has to be used and not the arithmetic average!!!

If the studied body is not in linear motion in three-dimensional Cartesian coordinate then its position will be given by the vector (x,y,z) and we will denote its velocity by:

$$\vec{v}=\left(\frac{\mathrm{d}x}{\mathrm{d}t},\frac{\mathrm{d}y}{\mathrm{d}t},\frac{\mathrm{d}z}{\mathrm{d}t}\right)=(v_x,v_y,v_z)=(\dot{x},\dot{y},\dot{z}) \tag{9.3.101}$$

Remark

If all the parts of a body moving at the same speed and in the same direction, then we have a "translation movement". By cons, in a "rotational movement", the speeds of the various body parts are not the same, in amplitude and in direction (we will see it further below in details), and may vary over time.

Caution! A movement can only be described scientifically with respect to a documented fixed reference point: the absolute motion does not exist. Galileo had already understood that: "*The movement is as nothing.*" The movement does not exist in itself, but in relation to something else.

Here is a non-exhaustive list of some typical speed magnitudes (the years are given relatively to the holocene calendar):

Magnitude $[\mathrm{m}\cdot\mathrm{s}^{-1}]$	**Magnitude** $[\mathrm{km}\cdot\mathrm{h}^{-1}]$	**Description**
$2.2\cdot 10^{-18}$	$7.8\cdot 10^{-18}$	Expansion rate between two points in free space 1 meter apart under Hubble's law
$1.4\cdot 10^{-14}$	$5\cdot 10^{-14}$	Cave of the Crystals gypsum crystals' growth rate
$1\cdot 10^{-13}$	$3\cdot 10^{-13}$	Rate of erosion of bedrock
$4.12\cdot 10^{-12}$	$1.48\cdot 10^{-11}$	Average growth rate of a limestone stalactite
$9.8\cdot 10^{-11}$	$3.5\cdot 10^{-10}$	Rate of global sea level rise in between 11993–12003 (approx. 3.1 $[\mathrm{mm}\cdot\mathrm{year}^{-1}]$)
$8\cdot 10^{-9}$	$5\cdot 10^{-8}$	Typical magnitude speed of continental drift
$1.3\cdot 10^{-9}$	$4.7\cdot 10^{-8}$	Average rate of the Moon receding from the Earth (approx. 38 $[\mathrm{mm}\cdot\mathrm{year}^{-1}]$)
$4.8\cdot 10^{-9}$	$1.7\cdot 10^{-8}$	Human hair growth (approx. 1.24 $[\mathrm{cm}\cdot\mathrm{month}^{-1}]$)
$1.4\cdot 10^{-5}$	$5\cdot 10^{-5}$	Average growth rate of bamboo (approx. 1.2 $[\mathrm{cm}\cdot\mathrm{day}^{-1}]$)
$6\cdot 10^{-4}$	$2.2\cdot 10^{-3}$	Typical speed of Thiovulum majus, the fastest-swimming known bacterium
1-5	3.6-5.4	Average walking human speed
2.39	8.53	World record time 50m freestyle swim
5.78	20.81	World record time marathon
6-7	20-25	Comfortable bicycling speed
10.438	37.578	Average speed of Jamaican athlete Usain Bolt while setting the 100m world record in Berlin on 16 August 12009
30	120	Typical speed of car (freeway)
33.67	121.21	Fastest sail boat Vestas Sailrocket 2
40	140	Typical peak speed of a local service train (or intercity on lower standard tracks)

54	195	Maximum speed an average human can attain during a face-down free-fall in standard conditions
67	240	The top speed of the world's fastest roller coaster, Formula Rossa
90	320	Diving peregrine falcon-fastest bird
91	328	Fastest recorded ball (a golf ball) in sports
113.88	410	Maximum speed of a sport car Bugatti Veyron Super Sport
119.44	430	Maximum speed of a street reglemented motorcycle Suzuki Hayabusa
126-143	452-517	The fastest wind speed ever recorded on Earth, caused by the 11999 Bridge Creek-Moore tornado
167.5	603	Top speed of Japanese L0 Series Maglev train in 12015
250	900	Typical cruising speed of a modern jet airliner, e.g. an Airbus A380
320	$1,200$	The speed of a typical .22 LR bullet
340.3	$1,225$	Speed of sound in standard atmosphere (15 [°C] and 1 [atm])
344.66	$1,240.77$	Max speed reached by the jet-propelled car ThrustSSC in 11997 - Land speed record
373	$1,342.8$	Highest speed recorded during a free fall set by Felix Baumgartner
464	$1,670$	Speed of Earth's rotation at the equator
603	$2,170.8$	Speed of the Concorde airliner
981	$3,532$	SR-71 Blackbird, the fastest aircraft driven by a mechanical jet engine
$1,022$	$3,679$	Average orbital velocity of the Moon around Earth
$1,400$	$5,040$	Speed of the Space Shuttle when the solid rocket boosters separate
$1,500$	$5,400$	Speed of sound in water
$1,789$	$6,443$	Speed of BrahMos II hypersonic cruise missile
$2,000$	$7,200$	Estimated speed of a thermal neutron
$2,375$	$8,550$	Escape velocity from Moon
$3,373$	$12,144$	Speed of the unmanned X-43 rocket/scramjet plane
$7,700$	$27,700$	Speed of International Space Station
$11,107$	$39,985.2$	Speed of Apollo 10 – high speed record for human-crewed vehicle
$11,200$	$40,320$	Average approx. escape velocity from Earth

$17,000$	$61,000$	The approximate speed of the Voyager 1 probe relative to the Sun, when it exited our Solar System
$29,800$	$107,280$	Actual average speed of the Earth in orbit around the Sun
$47,800$	$172,100$	Atmospheric entry speed of the Galileo atmospheric probe inside Jupiter
$192,000$	$690,000$	Top speed of the Parker Solar Probe
$200,000$	$700,000$	Average approx. orbital speed of our Solar System in the Milky Way galaxy
$308,571$	$1,080,000$	Actual approaching velocity of Andromeda Galaxy to our Galaxy
$450,000$	$1,600,000$	Typical speed of a particle of the solar wind, relative to the Sun
$617,700$	$2,224,000$	Escape velocity from the surface of the Sun
$1,610,000$	$5,800,000$	Speed of hypervelocity star PSR B2224+65, which currently seems to be leaving the Milky Way
$14,000,000$	$50,000,000$	Typical speed of a fast neutron
$30,000,000$	$100,000,000$	Typical speed of an electron in a cathode-ray tube
$100,000,000$	$360,000,000$	Typical speed of the return stroke of lightning
$150,000,000$	$540,000,000$	The theoretical escape velocity of a neutron star
$200,000,000$	$720,000,000$	Speed of a signal in an optical fiber
$299,792,455$	$1,079,252,838$	Maximal speed of a proton in LHC accelerator at CERN
$299,792,458$	$1,079,252,848.8$	Speed of light or other electromagnetic radiation in a vacuum or massless particles

Table 9.16: **Some common speed scales**

9.3.3.2.1 Motion parallax

Definition (#9.3.433): "Motion parallax" is a monocular depth cue arising from the relative velocities of objects moving across the retinae of a moving person. The term parallax refers to a change in position. Thus, motion parallax is a change in position caused by the movement of the viewer. Motion parallax arises from the motion of the observer in the environment.

It is perhaps easier to think of what motion parallax is by imagining yourself as a passenger in a car looking out the side window. The car is moving very fast down the highway. The objects very close to the window, such as the small trees planted by the highway, seem to rush by. Beyond the small trees, you can see a distant farmhouse. The farmhouse appears to move more slowly relative to you in the car. You know that the trees and the farmhouse are standing still; you are the object that is moving!

There are some famous videos on Internet highlighting this effect. One such clip of is a moving train with different apparent speeds at the same time has left netizens bamboozled. In this viral clip, a train appears to be moving at very different speeds when looking outside from different windows. In the first scene, the train can be seen moving at a high speed when looking outside from the left window where there is a wall, however, in the next part, the passing visuals on the other side of the window where there is also a wall show the speed of the train to be significantly slower. This is simply due to the fact that the two walls are at very different distance outside from the windows.

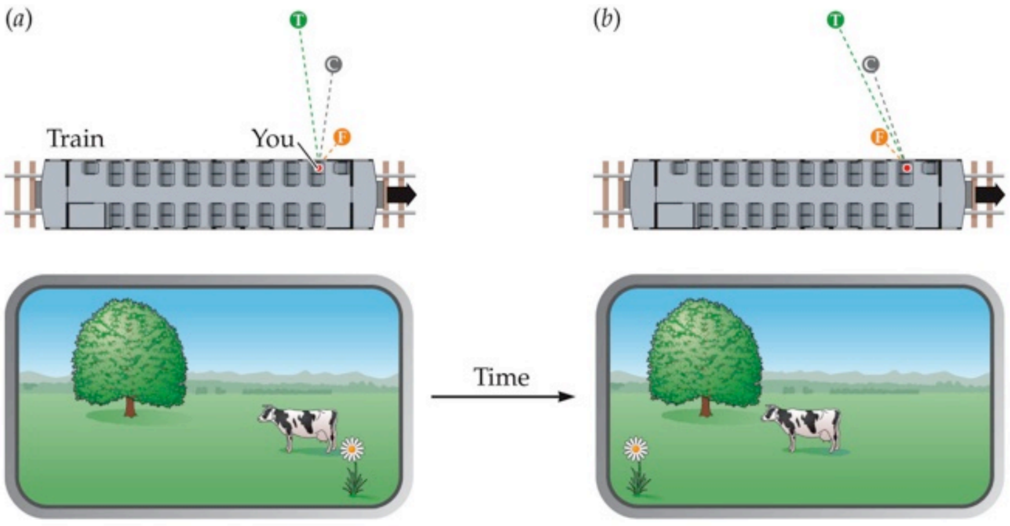


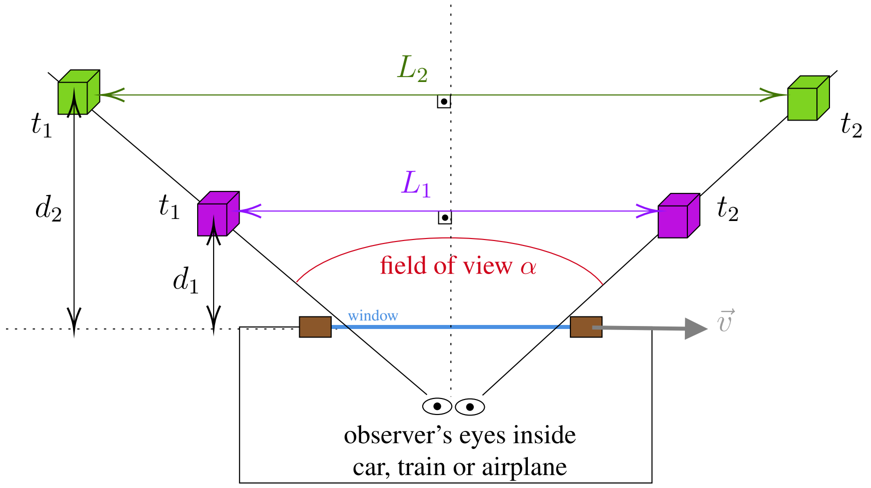
Figure 9.23: **Motion parallax**

Another famous video is the one of the Professor of psychology at Japan's Ritsumeikan University, Dr. Akiyoshi Kitaoka, who filmed from the front of a train and zoomed in on the tracks ahead. The bizarre clip tricks the brain into thinking the train is travelling slower when the footage is zoomed in.

This is completely a result of vector mathematics and relativity. Normally, velocity is not a scalar quantity! Speed is vectorial. In order to perceive velocity perfectly, the velocity vector must be perpendicular to our viewing angle. That is, the reason why we see a vehicle coming towards us from a distance as if it is moving slower. Because we perceive as speed perfectly only the perpendicular component of the speed vector to our viewing angle. This component is smaller than the absolute speed. As the vehicle approaches, we shift our perspective towards the vehicle. The velocity vector begins to be perpendicular to our field of view. When the vehicle is right next to us, this vector is almost perpendicular to our field of view. But this event is momentary. So normally it is quite difficult for us to predict the absolute velocity. For this reason, when the vehicle is right next to us, we notice a speed close to the actual speed of the vehicle. If we narrow our field of view by zooming the camera, the velocity vector of vehicle entering our field of view will be less perpendicular to our view angle, due to the narrowing of our field of view. For this reason, it will be more difficult to perceive the absolute speed of objects and we will see objects as moving more slowly.

This is one of the reasons why it is not recommended to walk or cross a motorway or railway tracks! Because the perception of speed of vehicles moving parallel to the point of view is highly biased and very likely leads to deadly decisions if not all necessary precautions are taken!

Let us formalize the concept mathematically now! Consider the figure below:



We see that for two different objects being in the same line of sight at time t_1 the distance they need to travel to disappear from the line of sight at time t_2 is obviously not the same as $L_2 > L_1$!

We see that the time while an object is in the field view is obviously given by (we neglect the distance between the observer's eyes and the window):

$$\Delta t_i = \frac{L_i}{v} = \frac{2d_i \tan\left(\dfrac{\alpha}{2}\right)}{v} \tag{9.3.102}$$

So we see that the time motion parallax is directly proportional do the distance from the observer. In other words, if an object is two times further away from a reference object observer it will take twice the time to disappear from the field of view! That's why the further is an object from the observer, the slower it looks like as it takes more time to disappear from the field of view!

9.3.3.3 Acceleration

Definition (#9.3.434): The "scalar acceleration", also known simply "acceleration", denoted "a", is by definition, the variation of the scalar speed during a given amount of time as (we go straight to the limit otherwise we speak of "average linear acceleration"):

$$\boxed{a_i = \frac{\mathrm{d}v_i}{\mathrm{d}t} = \frac{\mathrm{d}^2 x_i}{\mathrm{d}t^2} = \dot{v}} \tag{9.3.103}$$

or in other words: the speed at which the speed change...

The SI unit for acceleration is obviously meter per second squared: $[\mathrm{m}\cdot\mathrm{s}^{-2}]$.

For example, when a car starts from a standstill (zero relative velocity) and travels in a straight line at increasing speeds, it is accelerating in the direction of travel. If the car turns, there is an acceleration toward the new direction. In this example, we can name the forward acceleration of the car a "linear acceleration", which passengers in the car might experience as a force pushing them back into their seats. When changing direction, we might name this "non-linear acceleration", which passengers might experience as a sideways force. If the speed of the car decreases, this is an acceleration in the opposite direction from the direction of the vehicle, sometimes named "deceleration". Passengers may experience deceleration as a force lifting them forwards. Mathematically, there is no separate formula for deceleration: both are changes in velocity. Each of these accelerations (linear, non-linear, deceleration) might be felt by passengers until their velocity (speed and direction) matches that of the car.

Again, if the body is not in uniform motion we have:

$$\vec{a}=\left(\frac{\mathrm{d}v_x}{\mathrm{d}t},\frac{\mathrm{d}v_y}{\mathrm{d}t},\frac{\mathrm{d}v_z}{\mathrm{d}t}\right)=(a_x,a_y,a_z)=(\dot{v}_x,\dot{v}_y,\dot{v}_y) \tag{9.3.104}$$

If the body is in uniform rectilinear motion (we can always generalize to non-rectilinear movement as we will see further below) then we have:

$$\begin{gathered}v_i=\frac{\mathrm{d}x_i}{\mathrm{d}t}\Leftrightarrow v_i\mathrm{d}t=\mathrm{d}x_i\\ \Rightarrow\int\limits_{t_0}^{t}v_i\mathrm{d}t=\int\limits_{t_0}^{t}\mathrm{d}x_i\Rightarrow v_it|_{t_0}^{t}+c_1^{te}=x_i|_{t_0}^{t}+c_2^{te}\\ \Rightarrow x_i(t)-x_i(t_0)=c^{te}+v_it-v_it_0\Rightarrow\Delta x_i=c^{te}+v_i\Delta t\end{gathered} \tag{9.3.105}$$

The constant is to be determined depending on the initial conditions. If the initial position at time zero is zero constant will be zero. If not we write:

$$\boxed{\Delta x_i=x_{i,0}+v_i\Delta t} \tag{9.3.106}$$

which gives us the distance travelled by a body during a given period of time.

If the body is in rectilinear motion and constantly accelerates then we have:

$$\begin{gathered}a_i=\frac{\mathrm{d}v_i}{\mathrm{d}t}\Leftrightarrow a_i\mathrm{d}t=\mathrm{d}v_i\\ \Rightarrow\int\limits_{t_0}^{t}a_i\mathrm{d}t=\int\limits_{t_0}^{t}\mathrm{d}v_i\Rightarrow a_it|_{t_0}^{t}+c_1^{te}=v_i|_{t_0}^{t}+c_2^{te}\\ \Rightarrow v_i(t)-v_i(t_0)=c^{te}+a_it-a_it_0\Rightarrow\Delta v_i=c^{te}+a_i\Delta t\end{gathered} \tag{9.3.107}$$

The constant is to be determined depending on the initial conditions. If the initial velocity at time zero is zero, the constant will be zero. Otherwise we write:

$$\Delta v_i=v_{i,0}+a_i\Delta t \tag{9.3.108}$$

We see this relation more frequently in the form:

$$v_i(t)=v_{i,0}+a_it \tag{9.3.109}$$

or predominantly in the form:

$$\boxed{v=v_0+at} \tag{9.3.110}$$

but we have:

$$v=\frac{\mathrm{d}x}{\mathrm{d}t}=v_0+at\Rightarrow \mathrm{d}x=v_0\mathrm{d}t+at\mathrm{d}t \tag{9.3.111}$$

and if we integrate this relation, we get:

$$\int\limits_{t_0}^{t}\mathrm{d}x=\int\limits_{t_0}^{t}v_0\mathrm{d}t+\int\limits_{t_0}^{t}\mathrm{d}at\mathrm{d}t\Rightarrow \Delta x=x_0+v_0\Delta t+\frac{1}{2}a\Delta t^2 \tag{9.3.112}$$

that we find in the schools most frequently in the form:

$$\boxed{x=x_0+v_0t+\frac{1}{2}at^2} \tag{9.3.113}$$

where the acceleration a can be obviously negative when the object decelerate.

This relation gives the position of a mobile in rectilinear and uniformly accelerated motion (in the majority of physical problems we consider the acceleration as constant). Of the latter, we deduce a large quantity of other useful mathematical relations that are very interesting in physics as well considering ideals cases (frictionless) or real cases. Indeed, rearranging it (elementary algebra), it is possible to obtain acceleration when the time to reach a given speed is known. We can also calculate the distance it takes for a mobile (always in rectilinear movement) with a given acceleration to reach a certain speed.

The first case that we will consider and the best known, is the constant speed acceleration of a falling body in an environment free from any friction (case further treated in our study of tribology).

As we have already proved above, we have when the initial velocity is zero:

$$v=at\Leftrightarrow t=\frac{v}{a}\quad\text{and}\quad x=\frac{1}{2}at^2 \tag{9.3.114}$$

The two combined relations give (according to the tradition of use in schools, we replaced x by h to indicate that the position is often assimilated in practice at a height):

$$x=\frac{1}{2}a\left(\frac{v}{a}\right)^2=\frac{1}{2}\frac{v^2}{a}\Rightarrow v=\sqrt{2ah} \tag{9.3.115}$$

We can take out from this relation the escape velocity of body (useful relation when we will study Astrophysics - see corresponding section - and interesting for comparison when for when we will study General Relativity).

Suppose we already know that two bodies attract each other with an acceleration following classical Newton's model (that we will prove later at page 5255):

$$a=g=\frac{Gm}{R^2} \tag{9.3.116}$$

Placed in the relation $v=\sqrt{2ah}$ of a falling body, we get:

$$v=\sqrt{2\frac{Gm}{R^2}h} \tag{9.3.117}$$

So at the surface of the main attractor body we have the "escape velocity":

$$v_e=\sqrt{\frac{2Gm}{R^2}R}=\sqrt{\frac{2Gm}{R}} \tag{9.3.118}$$

We can answer from this relation, the question of why some solar system planets have atmospheres and others do not (obviously we must take normally into account the molecular agitation ...) as we will discussed ii more in details in the section Astronomy.

What is also interesting in this relation is that we can calculate what should be the radius R of a body of mass m so that to its escape velocity is equal to that of light (referring to Black Holes as we will study them in the section of General Relativity page 5720).

Therefore we have:

$$c=\sqrt{\frac{2Gm}{R}}\Rightarrow R=\frac{2G}{c^2}m \tag{9.3.119}$$

We will see in the section of General Relativity after relatively long calculations in an isotropic gravitational field (Schwarzschild metric) that we will fall back on that relation.

However...! This relation is not directly associated to Black Holes but to a hypothetical "dark star" that is a theoretical object compatible with Newtonian Mechanics that, due to its large mass, has a surface escape velocity that equals or exceeds the speed of light. Whether light is affected by gravity under Newtonian Mechanics is unclear but if it were accelerated the same way as projectiles, any light emitted at the surface of a dark star would be trapped by the star's gravity, rendering it dark, hence the name. Dark stars are only analogous to Black Holes in General Relativity.

Remark

The limits of acceleration that can be perceived by humans were studied by Nesti et al (see [412]) who summarized the results of several experiments. The conclusion is that the lowest acceleration we can perceive is $0.02\,[\mathrm{m\cdot s^{-2}}]$. The variation of acceleration of the Earth on its orbit is as we will prove it in the section of Astronomy (see page 5237) is of the order of $10^{-6}\,[\mathrm{m\cdot s^{-2}}]$ and the variation of acceleration on Earth's due to its rotation when we change of 1° of latitude is of the order of $10^{-4}\,[\mathrm{m\cdot s^{-2}}]$ as we will prove it during our study of centrifugal force later. This explains why the humans cannot detect such small variations. This is why scientists use instruments like accelerometers because as always ... all human senses are very poor to investigate the surrounding world with precision.

Here are some typical acceleration scales (the years are given relatively to the holocene calendar):

Magnitude $[\mathrm{m\cdot s^{-2}}]$	**Magnitude g**	**Description**

$5 \cdot 10^{-14}$	$5 \cdot 10^{-15}g$	Smallest acceleration in a scientific experiment
$1 \cdot 10^{-10}$	$5 \cdot 10^{-11}g$	Typical gravitational acceleration of stars in the Milky Way
$5.93 \cdot 10^{-3}$	$6.04 \cdot 10^{-4}g$	Acceleration of Earth toward the Sun due to Sun's gravitational attraction
1.62	$0.1654g$	Standing on the Moon at its equator
4.3	$0.44g$	Car acceleration 0-100 $[\text{km} \cdot \text{h}^{-1}]$ in 6.4 $[\text{s}]$ with a Saab 9-5 Hirsch
9.80665	$1g$	Standard gravity, the gravity acceleration on Earth at sea level standard
11.2	$1.14g$	Saturn V Moon rocket just after launch
15.2	$1.55g$	Bugatti Veyron from 0 to 100 $[\text{km} \cdot \text{h}^{-1}]$ in 2.4 s
29	$3g$	Space Shuttle, maximum during launch and reentry
29	$3g$	Sustainable for >25 seconds, for an untrained average human
34-39	3.5-$5g$	High-G roller coasters
41	$4.2g$	Top Fuel drag racing world record of 4.4 $[\text{s}]$
49	$5g$	Causes disorientation, dizziness and fainting for most humans (formula One car, maximum under heavy braking)
59	$6g$	Parachutist peak during normal opening of parachute
70.6	$7.19g$	Apollo 16 on reentry
79	$8g$	F-16 aircraft pulling out of dive
88	$9g$	Maximum for a fit, trained person with G-suit to keep consciousness, avoiding G-LOC
88-118	9-$12g$	Typical maximum turn acceleration in an aerobatic plane or fighter jet
147	$15g$	Explosive seat ejection from aircraft
177	$18g$	Physical damage in humans like broken capillaries
209	$21.3g$	Peak acceleration experienced by cosmonauts during the Soyuz 18a abort
333	$34g$	Peak deceleration of the Stardust Sample Return Capsule on reentry to Earth
454	$46.2g$	Maximum acceleration a human has survived on a rocket sled
>491	$>50g$	Death or serious injury likely
982	$100g$	Sprint missile, automobile crash (100 $[\text{km} \cdot \text{h}^{-1}]$ into wall)
$1,540$	$157g$	Peak acceleration of fastest rocket sled run
$2,256$	$230g$	Peak acceleration experience by the Galileo probe during descent into Jupiter's atmosphere

$2,490$	$254g$	Peak deceleration experienced by Jules Bianchi in crash of Marussia MR03, 2014 Japanese Grand Prix
$3,200$	$320g$	A jumping human flea
$3,800$	$380g$	A jumping click beetle
$4,944$	$504g$	Clothes on washing machine, during dry spinning (46 [cm] drum / $1,400$ rpm)
$29,460$	$3,000g$	Baseball struck by bat
$33,000$	$3,400g$	Standard requirement for decelerative crashworthiness in certified flight recorders (such as a Boeing 737 'black box')
$>49,100$	$>5,000g$	Shock capability of mechanical wrist watches
$84,450$	$8,600g$	Current Formula One engines, maximum piston acceleration (up to $10,000g$ before rev limits)
$102,000$	$10,400g$	A mantis shrimp punch
$196,400$	$20,000g$	Spore acceleration of the Pilobolus fungi
$1,000,000$	$100,000g$	Closing jaws of a trap-jaw ant
$1,865,000$	$190,000g$	9×19 [mm] Parabellum handgun bullet, peak
$3,800,000$	$390,000g$	Theoretical surface gravity of white dwarf Sirius B
$3,900,000$	$400,000g$	Ultracentrifuge
$1\cdot 10^9$	$100,000,000g$	The record peak acceleration of a projectile in a coilgun, a 2 gram projectile accelerated in 1 [cm] from rest to 5 $[\text{km}\cdot\text{s}^{-1}]$
$1\cdot 10^{12}$ to $1\cdot 10^{13}$	$1\cdot 10^{11}$ to $1\cdot 10^{12}$	Theoretical urface gravity of a neutron star
$2.1\cdot 10^{13}$	$2.1\cdot 10^{12}$ to $1\cdot 10^{12}$	Protons in the LHC at CERN
$9.149\cdot 10^{21}$	$9.33\cdot 10^{20}$	Classical (Bohr model) acceleration of an electron around a 1H nucleus

Table 9.17: **Some common acceleration scales**

9.3.3.3.1 Osculator Plane

For this subject the reader should first absolutely refer to the section Differential Geometry of the chapter Geometry and especially the part concerning the Frenet frame (see page 2863) that is a prerequisite to understand what will follow.

The vectors speed $\vec{v}$ and acceleration $\vec{a}$ associated with a point P moving form, at every moment t a plane named "osculating plane" of the path (usually curvilinear otherwise the plane is reduced to a straight line in the case of a rectilinear motion).

It is often useful to split the acceleration vector in the osculating plane respectively along the tangent and the normal to the trajectory:

$$\boxed{\vec{a}=\vec{a}_{\text{tg}}+\vec{a}_{\text{n}}} \tag{9.3.120}$$

where the first term of the right part is a vector parallel to the velocity and the second term a vector perpendicular to the velocity and located on the concave side of the path.

Let us express these two vectors explicitly (a more general example is given in the section of Differential Geometry as already mentioned!).

We can write:

$$\vec{v}=\frac{\mathrm{d}\vec{r}}{\mathrm{d}t}=\frac{\mathrm{d}\vec{r}}{\mathrm{d}s}\frac{\mathrm{d}s}{\mathrm{d}t}=\vec{T}v \tag{9.3.121}$$

where $\mathrm{d}s$ is an element of the curve (curvilinear abscissa) from the path Λ and $\vec{T}$ a unit vector tangent to the path and linked to the point P.

The acceleration is then written:

$$\vec{a}=\frac{\mathrm{d}\vec{v}}{\mathrm{d}t}=\frac{\mathrm{d}(\vec{T}v)}{\mathrm{d}t}=\vec{T}\frac{\mathrm{d}v}{\mathrm{d}t}+v\frac{\mathrm{d}\vec{T}}{\mathrm{d}t} \tag{9.3.122}$$

The first term on the right of the equality is the tangential acceleration and for the second term, even if the speed is constant it appears in the expression of the acceleration to express the change of direction of the speed vector.

Let us decompose the vector $\vec{T}$ in the orthonormal Euclidean base $\mathbb{R}^2$ generated by the family of vector $\vec{i},\vec{j}$:

$$\vec{T}=T_x\vec{i}+T_y\vec{j}=\|\vec{T}\|\cos(\alpha)\vec{i}+\|\vec{T}\|\sin(\alpha)\vec{j}=\cos(\alpha)\vec{i}+\sin(\alpha)\vec{j} \tag{9.3.123}$$

After by derivating in respect to time:

$$\begin{aligned}\frac{\mathrm{d}\vec{T}}{\mathrm{d}t}&=-\sin(\alpha)\frac{\mathrm{d}\alpha}{\mathrm{d}t}\vec{i}+\cos(\alpha)\frac{\mathrm{d}\alpha}{\mathrm{d}t}=(-\sin(\alpha)\vec{i}+\cos(\alpha)\vec{j})\frac{\mathrm{d}\alpha}{\mathrm{d}t}\\&=\left[\cos\left(\alpha+\frac{\pi}{2}\right)\vec{e}_i+\sin\left(\alpha+\frac{\pi}{2}\right)\vec{e}_j\right]\frac{\mathrm{d}\alpha}{\mathrm{d}t}\end{aligned} \tag{9.3.124}$$

By comparing with the initial expression of the vector $\vec{T}$, we see that the terms in brackets above are the components of a new unit vector $\vec{n}$ perpendicular to the vector $\vec{T}$, thus perpendicular to the path and directed toward the center of curvature (see section Differential Geometry page 2863).

Moreover by the definition of the radian (see section Trigonometry page 2503), we have:

$$\frac{\mathrm{d}\alpha}{\mathrm{d}t}=\frac{\mathrm{d}s}{R}\frac{1}{\mathrm{d}t}=\frac{1}{R}\frac{\mathrm{d}s}{\mathrm{d}t}=\frac{v}{R} \tag{9.3.125}$$

where R is the instantaneous (local) radius of curvature of the trajectory.

The expression $\mathrm{d}\vec{T}/\mathrm{d}T$ then becomes:

$$\frac{\mathrm{d}\vec{T}}{\mathrm{d}t}=\frac{v}{R}\vec{n} \tag{9.3.126}$$

and the second term of the general expression of the acceleration then becomes:

$$v\frac{\mathrm{d}\vec{T}}{\mathrm{d}t}=\frac{v^2}{R}\vec{n} \tag{9.3.127}$$

So we have finally (relation proved with another approach in the section of Differential Geometry):

$$\boxed{\vec{a}=\frac{\mathrm{d}v}{\mathrm{d}t}\vec{T}+\frac{v^2}{R}\vec{n}=\vec{a}_{\text{tg}}+\vec{a}_{\text{n}}} \tag{9.3.128}$$

where once again $\vec{a}_{\text{tg}}$ is the tangential acceleration that express the variation of speed on the trajectory of the body (or point) P and where $\vec{a}_{\text{n}}$ is a term that express the change of direction of the body (point) P without that the latter see its speed change! Commonly the normal acceleration is assimilated, when multiplied by the body (point) mass, to the "centrifugal (fictional) force" (centrifugal meaning: that go away from the center). We will come back more in the details on this force further below.

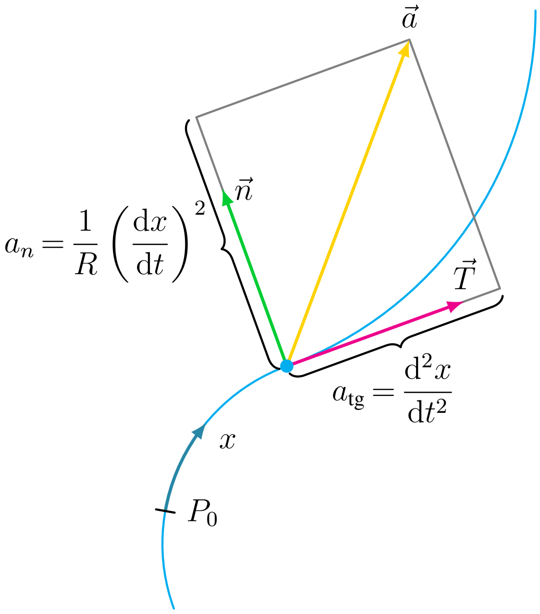

$\vec{a}$
$\vec{n}$
$a_n=\frac{1}{R}\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2$
$\vec{T}$
$a_{\text{tg}}=\frac{\mathrm{d}^2x}{\mathrm{d}t^2}$
x
P_0

Figure 9.24: Tangential and Normal acceleration

Caution! The centrifugal force is considered in physics as a fictitious force because in fact this is not a force that tends to move the body (particle) away from a center of rotation but it is just that there is a force which is not sufficient (the friction force in the case of carousel or the gravity force in case of planets or satellites) to prevent us from following a straight course by simple inertia. This is why when we are ejected from a carousel we leave tangentially to the rotation and not perpendicularly thereto.

We immediately see that if $\vec{a}_\text{n}=\vec{0}$ the movement can be at least necessarily linear but surely uniform, accelerated or not, while if $\vec{a}_\text{n}\neq\vec{0}$ the trajectory is necessarily curved.

Indeed, as we just say the acceleration is basically rate of change of velocity. In a rotational motion, there are two components of the net acceleration: one Normal(along the radius) and one Tangential (along the circumference). The Normal acceleration serves the purpose of causing the motion to be circular, whereas the tangential one can make the rotation faster or slower. Since a uniform motion itself defines to be a constant velocity motion (i.e. the rotation is going on at a constant speed), the tangential component of acceleration is zero. If the rotation is non-uniform, that means the speed of rotation is changing, thus inevitably leading to the conclusion that the tangential component of the net acceleration is non-zero.

9.3.3.4 Galilean Relativity Principle

Definition (#9.3.435): As far as we know it is impossible for a lively observer in uniform motion to know if he moves relatively to its environment or conversely if the environment is moving relative to him (we can not distinguish the rest and motion at constant speed and direction). Therefore, there can be no absolute reference frame (or privileged reference frame) that can be considered fixed vis-à-vis all other Galileans referential frame which clearly means that all Galilean reference frame must enjoy the same status in Mechanics as they can not be distinguished from each other. This principle is named the "principle of Galilean relativity" or "Galilean invariance".

Specifically, the term "Galilean invariance" today usually refers to this principle as applied to Newtonian Mechanics, that is, Newton's laws hold in all frames related to one another by a Galilean transformation. In other words, all frames related to one another by such a transformation is inertial (meaning, Newton's equation of motion is valid in this frame). In this context it is sometimes named "Newtonian relativity".

So for example, the two observers depicted below, the one on the ground in the reference frame S and the one in unform motion in a vehicle in the reference frame S' will observe exactly the same trajectory of the ball they throw from their point of view and will conclude to the same laws of physics (!):

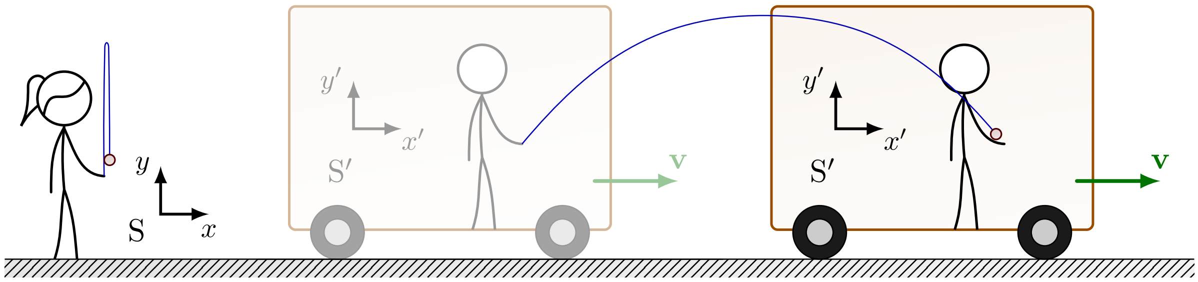


Figure 9.25: **The principle of Galilean relativity illustrated (author: Izaak Neutelings)**

Hence uniform motion (constant velocity) in a vacuum is undetectable! This is a core principle of relativity, stating that no physical measurement can distinguish between inertial frames of reference in uniform motion.

This principle (not to be confused with the principle of Special Relativity because the assumptions differ a little bit ...) follows directly from the study of what we name the "Galilean transformation" that we will study below for Kinematics and for which we will give a more in-deep example in the section of Special Relativity with the Galilean Transformation of a wave equation.

Definition (#9.3.436): A "Galilean transformation" is a sequence of mathematical operations on a physical law that determines the properties of one or more "observable" (speed, strength, momentum, etc.) when we move during the study of a physical phenomenon from one repository to another repository: one supposed to rest (relatively to the second), and the other in uniform motion (or vice versa).

The question at the historical origin was to answer if it was more legitimate to study a phenomenon in a repository or another one. More exactly, we want to determine whether the form of physical laws keep the same algebraic form regardless of the repository in which we study them.

Let us have a more closer look at this:

Given two repositories in movement relative to each other at a constant speed $\vec{u}(u_x, u_y, u_z)$. For a given Cartesian reference frame $(\vec{x}, \vec{y}, \vec{z})$ at rest (or supposed such) we will put for the second repository $(\vec{x}', \vec{y}', \vec{z}')$ so that it is aligned with the x-axis of the first repository to simplify calculations with $\vec{u}(u_x, 0, 0)$:

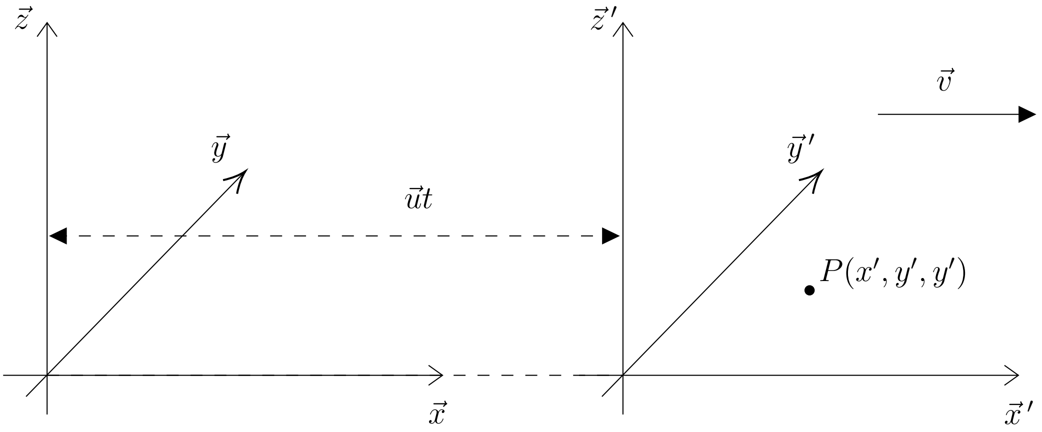


Figure 9.26: **Example repositories in uniform relative motion with respect to each other**

We will also put in the second frame of reference, a material point P of coordinates $\vec{P}(x',y',z')$.

Remark

We assume known the concept of "linear momentum" p defined further below rigorously. Let us just recall here that the linear momentum of the animated point P of mass m', with a speed v (norm) in $(\vec{x}',\vec{y}',\vec{z}')$ is given by:

$$p' = m'v' \tag{9.3.129}$$

We then have by applying the conventional relations of rectilinear kinematics:

$$x' = x - u_x t \qquad y' = y \qquad z' = z \tag{9.3.130}$$

$$t' = t \qquad m' = m \tag{9.3.131}$$

Remark

Note that the last equation expresses the assumption of a universal time and mass independent of the relative motion of different observers. This has to be compared with our future study of Special Relativity (see the section with the corresponding name page 5627).

We get:

$$\begin{aligned} p'_x &= mv'_x = m\frac{\mathrm{d}x'}{\mathrm{d}t'} = m\frac{\mathrm{d}(x-u_x t)}{\mathrm{d}t} = m\left(\frac{\mathrm{d}x}{\mathrm{d}t} - \frac{\mathrm{d}u_x}{\mathrm{d}t}t - \frac{\mathrm{d}t}{\mathrm{d}t}u_x\right) = m\left(\frac{\mathrm{d}x}{\mathrm{d}t} - 0\cdot t - 1\cdot u_x\right) \\ &= m\left(\frac{\mathrm{d}x}{\mathrm{d}t} - u_x\right) = m(v_x - u_x) \end{aligned} \tag{9.3.132}$$

hence $v'_v = v_x - u_x$ and therefore we get for the force (we will study further below what links linear momentum and force together):

$$F_x = \frac{\mathrm{d}p_x}{\mathrm{d}t} = \frac{\mathrm{d}m(v_x-u)}{\mathrm{d}t} = 0 = F'_x = \frac{\mathrm{d}p'_x}{\mathrm{d}t'} = 0 \tag{9.3.133}$$

The result is interesting since Newton's second law keeps exactly the same shape and the same value in the two repositories (we better understand the name of "Newton relativity"). The fact that we were moving or not has therefore no impact on our vision of the world that remains exactly the same (in this model because when we will study Special Relativity we will see that this is not satisfied anymore!!!).

Consequence: Since the forces are the same, no mechanical experience can determine whether a Galilean is the absolute coordinate system (i.e. two observers in two different Galilean frames, cannot with mechanical experience determine which moves relative to the other one).

So in Classical Mechanics, there is no absolute Galilean reference frame!

Notice, however, that this is achieved by assuming that:

$$\frac{\mathrm{d}mu}{\mathrm{d}t}=\frac{\mathrm{d}mv_x}{\mathrm{d}t}=\frac{\mathrm{d}mv'_x}{\mathrm{d}t'}=0 \tag{9.3.134}$$

that is to say, we require that the relative velocity is uniform (constant), the mass is also constant and especially that $t=t'$.

But in fact, this transformation is fundamentally wrong as discussed in more detail in our study of Special Relativity (see the section with the corresponding name page 5627). Indeed, the easiest way to see it that is more experimental than theoretical (we keep the theoretical one for the section of Special Relativity) is to consider an object moving along the axis with a speed v measured in the apostrophed reference frame:

$$x'=vt' \tag{9.3.135}$$

then what will be its velocity w in the apostrophed reference frame? If the Galilean transformation is fundamentally true, it would be enough to replace in the above relation x' and t' by their expressions in function of t:

$$x-ut=vt \qquad \text{or} \qquad x=(u+v)t \tag{9.3.136}$$

thus (velocity addition law):

$$w=\frac{\mathrm{d}x}{\mathrm{d}t}=u+v \tag{9.3.137}$$

But there was a problem ... an experiment involving light rays (electromagnetic waves) was conducted at the beginning of the 120th century (holocene calendar), and showed that this addition law was false. This experiment named "Michelson-Morley experiment" (see [390]) and later the "de Sitter double star experiment" (see [133]) changed forever our vision of the world ... and brought Albert Einstein to develop the theory of Special Relativity by requiring that the speed of light to be constant regardless of the movement of the reference frame (see section Special Relativity page 5627):

$$\boxed{c:=v_{\text{light}}=c^{te}} \tag{9.3.138}$$

If we measure the speed and other vector quantities, we find that the results of measurements of the components x', y', z', t' are not identical to those obtained on x, y, z, t. In physics, the "principle of covariance" emphasizes the formulation of physical laws using only certain physical quantities such that their measurements in different frames of reference can be unambiguously correlated. Mathematically, the physical quantities must transform "covariantly", that is, under a certain representation of the group of coordinate transformations between admissible frames of reference of the physical theory. The corresponding group is the referred to as the "covariance group". The covariance principle states that the laws of physics should transform from one frame to another covariantly, that is, according to a representation of the covariance group (Lorentz matrices in Special Relativity and in General Relativity the covariance group consists of all arbitrary invertible and differentiable coordinate transformations).

As we know, laws (formulas) are relations between observable (terms), relations deduced from many observations. The research of laws is governed by what we might name the "principle of simplicity": laws in the smallest possible number, expression the simple as possible between terms and in minimum quantity. But the characteristic of a good law is also the covariance during a change of reference frame. The Physics (in the sense of the theory that describes the sensitive reality) will no longer be linked to the observer nor his associated Galilean space-time.

However, let us give an important counterexample in Classical Mechanics: the force between two stationary electric charges in a referential frame use in this reference frame only electrostatic theory. If the same system is observed from a moving reference frame relative to the first, it will have to be described by the electromagnetism theory (as we will see in the section of Electrodynamics an electric charge in movement produces a magnetic field). This can be explained only thanks to the covariance group of Special Relativity[14].

By construction Classical Mechanics is being covariant by Galilean transformation (change of Galilean reference frame): the dynamic postulate (force) takes the same form in the different Galilean frames as we have just seen!

---

[14] See our study of the Joules-Bernoulli equations in the section of Special Relativity at page 5696

9.3.4 Angular Momentum

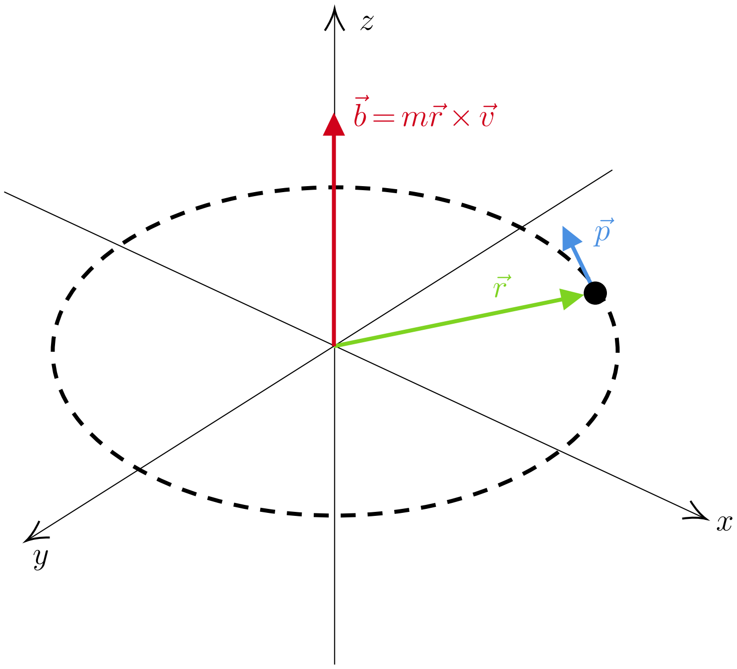
Definition (#9.3.437): The "angular momentum" (rarely "moment of momentum" or "rotational momentum" or commonly named "orbital angular momentum" in astronomy) $\vec{b}$ with respect to a point O of a particle of mass m moving at the speed $\vec{v}$ on $\vec{r}$ is defined by:

$$\boxed{\vec{b}=\vec{r}\times\vec{p}} \tag{9.3.139}$$

with $\vec{p}$ being the linear momentum (see definition during the Newton's second law statement) given by:

$$\vec{p}=m\vec{v} \tag{9.3.140}$$

By definition, the angular momentum is a vector perpendicular to the plane containing the vectors $\vec{r}$ and $\vec{p}$ and if the particle moves in a plane, the direction of $\vec{b}$ is constant but not necessarily in the same direction depending or the rotating sense of the particle.



A special but important case in Astronomy and Mechanics is to calculate the angular momentum of a circular plane motion of radius r. In this situation, the "radius vector" $\vec{r}$ is then always perpendicular to the direction of the velocity vector $\vec{v}$ and therefore:

$$\boxed{\vec{b}=m\vec{r}\times\vec{v}\Rightarrow\|\vec{b}\|=mvr=m\omega r^2\Rightarrow mr^2\vec{\omega}} \tag{9.3.141}$$

We see here appear the definition of the "rotation vector $\vec{\omega}$" also sometimes denoted by $\vec{\Omega}$.

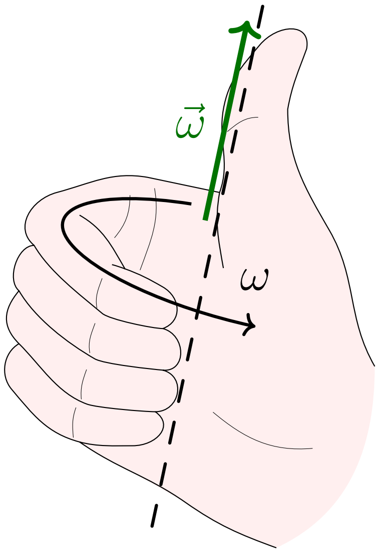


Figure 9.27: **Angular momentum right-handed rule (author: Izaak Neutelings)**

For a non rectilinear motion path (a conical for example), we introduce the normal and tangential velocity components (we will in-deep this further below):

$$\vec{v}=\vec{v}_n+\vec{v}_{\text{tg}} \tag{9.3.142}$$

But as the velocity is the change in position over time the velocity is (by definition) tangent to the curve. So we get:

$$\vec{b}=m\vec{r}\times(\vec{v}_n+\vec{v}_{\text{tg}})=m\vec{r}\times\vec{v}_{\text{tg}} \tag{9.3.143}$$

If we take a differentiable vector function $\vec{v}(t)$ (a velocity vector). If its speed is such that $\vec{v}(t)=c^{te}$ then at any point which $\mathrm{d}\vec{v}/\mathrm{d}t$ is non zero, $\mathrm{d}\vec{v}/\mathrm{d}t$ is perpendicular to $\vec{v}(t)$.

The shortest way to prove this assertion is to remember that:

$$\|\vec{v}\|^2=v^2=\vec{v}\circ\vec{v} \tag{9.3.144}$$

After differentiating:

$$2v\frac{\mathrm{d}v}{\mathrm{d}t}=\frac{\mathrm{d}\vec{v}}{\mathrm{d}t}\circ\vec{v}+\vec{v}\circ\frac{\mathrm{d}\vec{v}}{\mathrm{d}t}=2\vec{v}\circ\frac{\mathrm{d}\vec{v}}{\mathrm{d}t} \tag{9.3.145}$$

and since its speed is constant:

$$2v\cdot 0=2\vec{v}\circ\frac{\mathrm{d}\vec{v}}{\mathrm{d}t} \tag{9.3.146}$$

Since $\vec{v}\neq 0$ and $\mathrm{d}\vec{v}/\mathrm{d}t\neq 0$ then they have to be perpendicular!

So, coming back to:

$$\vec{b}=\vec{r}\times\vec{v}_{\text{tg}} \tag{9.3.147}$$

in scalar form this gives:

$$b=mrv_{\text{tg}}=mr^2\frac{\mathrm{d}\theta}{\mathrm{d}t} \tag{9.3.148}$$

where r is therefore named the "radius of curvature" of the trajectory as we already know (see section Differential Geometry page 2865).

Let us now study the derivative of the angular momentum:

$$\frac{\mathrm{d}\vec{b}}{\mathrm{d}t}=\frac{\mathrm{d}(\vec{r}\times\vec{p})}{\mathrm{d}t}=\frac{\mathrm{d}\vec{r}}{\mathrm{d}t}\times\vec{p}+\vec{r}\times\frac{\mathrm{d}\vec{p}}{\mathrm{d}t} \tag{9.3.149}$$

In the right-hand side, we have by the definition of the vector product:

$$\frac{\mathrm{d}\vec{r}}{\mathrm{d}t}\times\vec{p}=\vec{v}\times m\vec{v}=\vec{0} \tag{9.3.150}$$

and also:

$$\vec{r}\times\frac{\mathrm{d}\vec{p}}{\mathrm{d}t}=\vec{r}\times\vec{F}=\vec{M} \tag{9.3.151}$$

Which gives finally:

$$\boxed{\frac{\mathrm{d}\vec{b}}{\mathrm{d}t}=\vec{r}\times\vec{R}=\vec{M}} \tag{9.3.152}$$

The derivative with respect to time of the angular momentum of a moving point is equal to what we define as the "moment of force $\vec{M}$" which will also be discussed further below and whose units are that of energy and is a vector perpendicular to the plane defined by $\vec{r}$ and by $\vec{F}$ (by construction of the cross product as seen in the section of Vector Calculus!).

Remark

This latter relation makes that we sometimes name the angular momentum as "moment of linear momentum"...

What is highly impressive in this result (instantaneous change of angular momentum), is that any body having a non-zero angular momentum and subject to no moment of force, maintains the direction and norm of $\vec{b}$ in space and time.

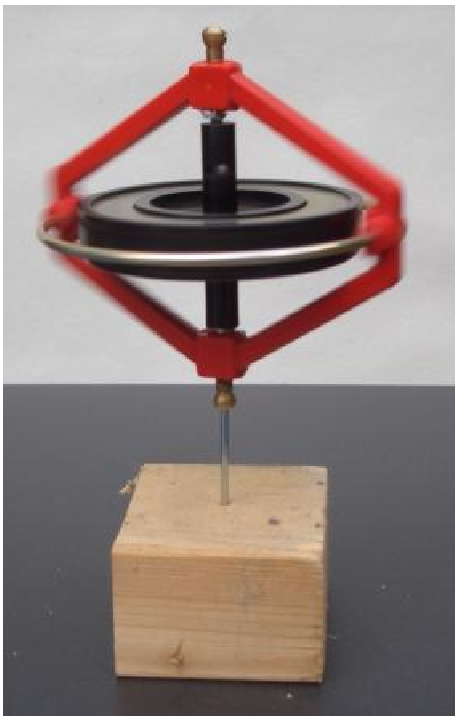


Figure 9.28: Gyroscope remaining upright while spinning owing to the conservation of its angular momentum (source: Wikipedia)

This result will allow us to study the dynamic of the gyroscope and all other bodies having similar properties (such as the Earth rotating on itself and that points approximately to the North Star, i.e. Polaris nowadays, which is a major cause of origin of the seasons!). We will study further below the gyroscope and its properties, because their behaviour is fascinating and the theoretical results arising find applications in astrophysics, atomic physics and even in philosophy. Indeed, the conservation of the orientation of the angular momentum vector leads to the conclusion that even if the space was completely empty of its contents in the whole Universe and contained an object with an angular momentum, the empty space has however a property which allows the object in question where it oriented it angular momentum... which is quite disturbing! So the empty space is not nothing otherwise relatively to what turns the object into space? It rotates relative to the space itself!

We also have:

$$\vec{b}=\int_{t_1}^{t_2}\vec{M}\mathrm{d}t \qquad (9.3.153)$$

where the integral is named "angular impulsion" and the previous relation sometimes is named "theorem of angular momentum" (we'll see a generalization of this theorem in the proof of König's theorem later). It is stated as following: The angular impulsion provided by a moment of force between the moments t_1 and t_2 is equal to the variation of the angular momentum during this time interval.

In the study of solid dynamic this theorem plays a fundamental role, similar to the dynamic Newton equation $\vec{F}=m\vec{a}$.

The use of angular momentum makes it easy to prove the Kepler law of areas (Kepler's second law), which plays an important role in the understanding of planetary motion (see section

Astronomy page 5233) or to prove that in an isolated Earth-Moon system, the total angular momentum is conserved, so if the Earth slows its rotation and the Moon keep it constant, it forces the Moon to increase its distance from Earth!!!

Let us see this last statement:

Imagine a particle moving under the action of a force $\vec{F}$ constant and parallel to $\vec{r}$. We say that this force is a "central force" if its direction constantly goes through the same fixed point, named the "center of force". The magnitude of the force only depends on the distance r of the object from the origin and is directed along the line joining them (in the case of a force field):

$$\boxed{\vec{F}=\vec{F}(\vec{r})=F(\|\vec{r}\|)\vec{r}} \tag{9.3.154}$$

Therefore:

$$\frac{\mathrm{d}\vec{b}}{\mathrm{d}t}=\vec{r}\times\vec{R}=\vec{0} \tag{9.3.155}$$

as in central force problems $\vec{F}$ is parallel to $\vec{r}$.

Therefore the angular momentum relative to the center of force is constant if the force is central! The converse is also true: if the angular momentum is constant, its derivative with respect to time is zero and the direction of the force is always collinear to $\vec{r}$ so the force is central.

For example, in the motion of a planet around the Sun (actually a G2V-type star, aka yellow dwarf) or an electron around the nucleus of the atom (in the Bohr model...) the momentum of this force relatively to the center because is obviously zero as no outside element acts on the system, that is to say based on the figure below:

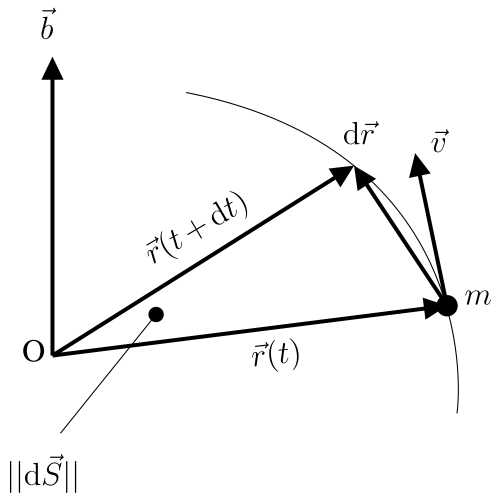


Figure 9.29: **Illustration of momentum**

Then we have:

$$\vec{M}=\vec{r}\times\vec{F}=\vec{0}\Leftrightarrow\vec{M}=\frac{\mathrm{d}\vec{b}}{\mathrm{d}t}=\vec{0} \tag{9.3.156}$$

Therefore:

$$\vec{b}=\vec{r}\times m\vec{v}=\vec{c}^{te} \tag{9.3.157}$$

On the other hand, the element surface $\mathrm{d}\vec{S}$ described by the motion of the radius $\vec{r}$ is (according to the figure above and the properties of the vector product): therefore:

$$\dfrac{\mathrm{d}\vec{S}}{\mathrm{d}t}=\dfrac{1}{2}\left(\vec{r}\times\dfrac{\mathrm{d}\vec{r}}{\mathrm{d}t}\right)=\dfrac{1}{2}(\vec{r}\times\vec{v}) \tag{9.3.158}$$

Using the relation $\vec{b}=\vec{r}\times m\vec{v}=\vec{c}^{te}$ we get:

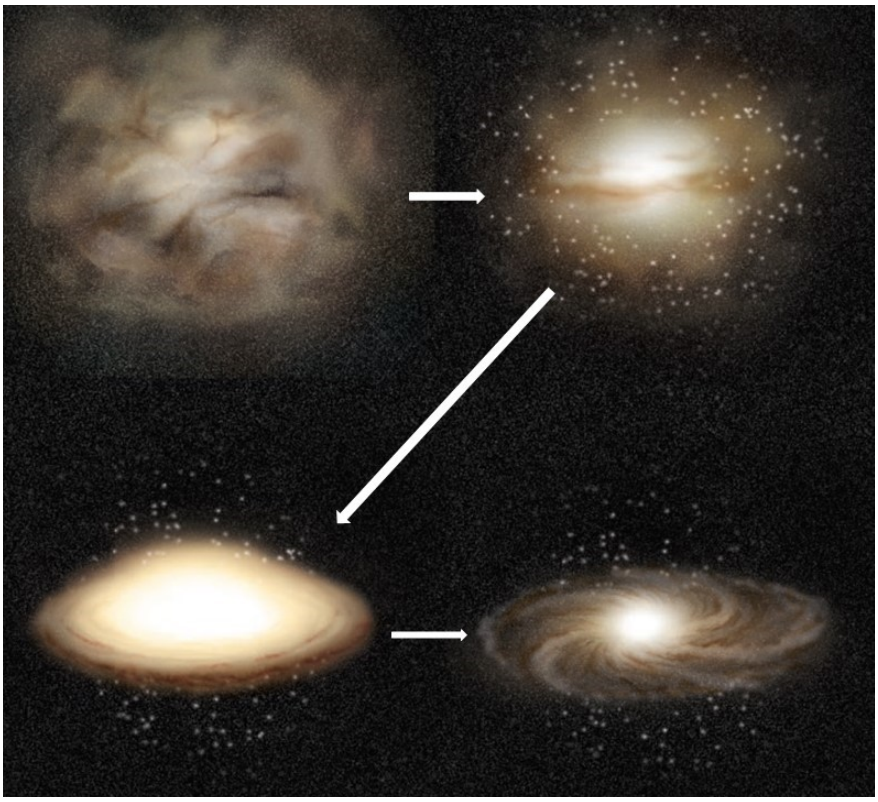
$$\boxed{\dfrac{\mathrm{d}\vec{S}}{\mathrm{d}t}=\dfrac{1}{2m}(\vec{r}\times m\vec{v})=\vec{c}^{te}} \tag{9.3.159}$$

Consequences:

1. The areal velocity is constant in a central force problem, that is to say that the areas scanned in equal times are equal. It is the Kepler areas law (see section Astronomy page 5234)!
2. The plane $\vec{r},\vec{v}$ is fixed because $\vec{r}\times m\vec{v}=\vec{c}^{te}$. Thus the trajectory of a planet in a ideal case is viewed as moving in a plane relatively to the reference frame of the force center.

We'll obviously come back on this relation in the section Astronomy to write it in a slightly more traditional form.

We (almost) all know that the Earth, the other planets and their moons all revolve around the Sun in elliptical orbits (from the point of view of the Sun referential otherwise the trajectory is helicoidal...!). However, have you ever stopped to wonder why these celestial objects move around the Sun in such a way that our Solar System appears to be lying on a plane, rather than going every which way and the same observation can also be done for Saturn's rings, galaxies or Black Holes disk accretion?



We can trace the origin of our Solar System to a massive shapeless 'blob' of matter floating through space about 4.6 billion years ago (see [69]).

The particles in this blob gradually began to move closer due to gravity, and whenever there are multiple particles and powerful gravitational forces, there are also collisions and therefore heating and emissions of radiations outside of the cloud. These collisions and the subsequent trajectories of these particles are obviously random, making them almost impossible to predict. Although these objects move randomly, the one thing that remains constant as we just saw is the angular Momentum as the latter does not irradiate anything outside the cloud and to be changed must be influence by an external force!!! The solar system is an isolated system in itself (in first approximation!). The galaxy is also an isolated system (in first approximation), since the gravitational effect of other cosmic objects is negligible.

So an isolated system and therefore without interaction with external objects, the total Angular Momentum has to be conserved!

Remark

A common question is: what is the origin of angular momentum in the early time of our universe? First a thing to know is that for the same reason energy and momentum is not conserved in an expanding universe (violation of space translation), conservation of angular momentum (result of rotation invariance) is also not conserved in curved space-time unless the metric has rotational symmetry. And keep in mind the spin of an object can be compensated by an opposite change in spin in one or more other objects (as angular momentum is a vectorial quantity).

This physical quantity is constant around a fixed axis. This axis is a point in 2-dimensional space, and in our world, which exists in 3 dimensions, this axis turns out to be a line. The system rotates along a plane that is perpendicular to the axis. Therefore, whenever particles collide, they may move in any direction, but all the up and down motion cancels out, always following the rule that the total spin in that plane must be constant. Over time and after countless collisions, these particles lose their freedom in everything except 2-dimensional space, thereby aligning themselves in a plane.

9.3.4.1 Moments

A "moment" (not to be confused with Momentum!) or "torque" is an expression involving the product of a distance and a physical quantity, and in this way they account for how the physical quantity is located or arranged. Moments are usually defined with respect to a fixed reference point; they deal with physical quantities as measured at some distance from that reference point. For example, the moment of force acting on an object, often named "torque", is the product of the force and the distance from a reference point. In principle, any physical quantity can be multiplied by distance to produce a moment; commonly used quantities include forces, masses, and electric charge distributions.

OK let's go! In fact we have seen just above that the "moment of force" was defined by (temporal variation of the angular momentum):

$$\boxed{\vec{M}=\vec{r}\times\vec{F}} \tag{9.3.160}$$

where $\vec{M}$ is the moment (or "torque" also denoted sometimes by the letter τ) of the force $\vec{F}$ relatively to the origin of the vector $\vec{r}$. It is important to note that the force momentum has for units "energy" ([J]) and is also perpendicular to $\vec{r}$ and $\vec{F}$ by construction!

It must be also notice that increasing the radius of force application thereby giving the opportunity to decrease the force necessary to keep a constant force momentum in a mechanical system certainly reduces the effort (force) but in the end not the energy expended since the traveled distance is bigger.

If we express the norm of the force moment $\vec{M}$, by the definition of the cross product, we get:

$$\|\vec{M}\|=\|\vec{F}\|\cdot\|\vec{r}\|\sin(\alpha)=\|\vec{F}\|b \tag{9.3.161}$$

Therefore appears a quantity:

$$b=\|\vec{r}\|\sin(\alpha) \tag{9.3.162}$$

which by definition is the "lever arm" of the force $\vec{F}$ and whose location is given by the axis of rotation of the body due to the resulting moment of force (do not confuse this with the notation $\vec{b}$ of the angular momentum!).

Warning! The principle of the lever arm is thus a great force multiplier but in no case it multiplies the work (energy)!

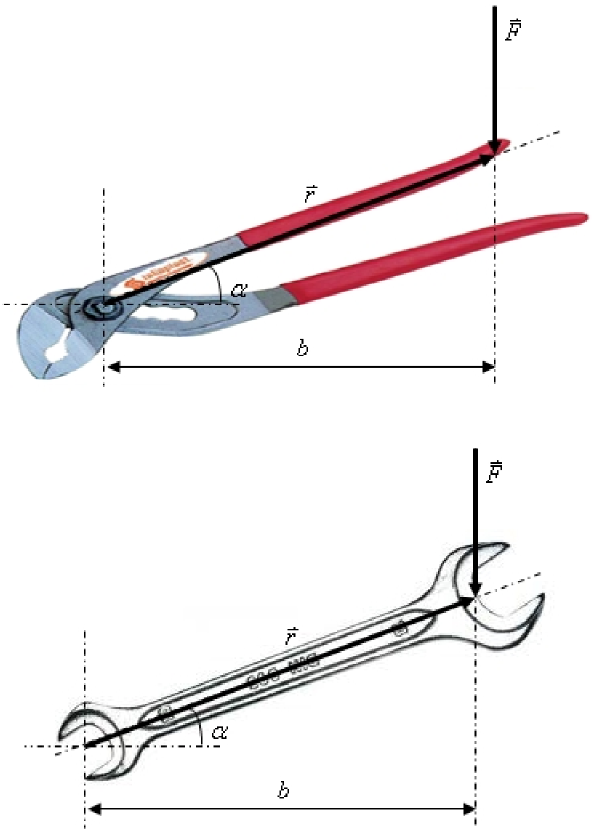


Figure 9.30: **Example of some everyday lever arm**

Remark

The case of application of changes of winter/summer of tourism cars is well known by motorists as it is recommended by most manufacturers to apply a torque of 120 [Nm] for the tightening of the bolts.

For an extended body, subject to forces $\vec{F}_1, \vec{F}_2, \ldots, \vec{F}_n$ to be in total equilibrium, it is not only necessary that the resultant of these forces is zero (no translation) but also the resulting force moments is zero too (no rotation). That is to say:

$$\sum_{i=1}^{n} \vec{F}_i = \vec{0} \quad \text{and} \quad \sum_{i=1}^{n} \vec{M}_i = \vec{0} \tag{9.3.163}$$

Thus more explicitly:

$$\boxed{\begin{aligned} &\sum_{i=1}^{n} F_{i,x} = \sum_{i=1}^{n} F_i \cos(\alpha_i) = 0 \\ &\sum_{i=1}^{n} F_{i,y} = \sum_{i=1}^{n} F_i \sin(\alpha_i) = 0 \\ &\sum_{i=1}^{n} M_i = \sum_{i=1}^{n} (x_i F_{i,y} - y_i F_{i,x}) = \sum_{i=1}^{n} F_i (x_i \sin(\alpha_i) - y_i \cos(\alpha_i)) = 0 \end{aligned}} \tag{9.3.164}$$

When the components of a system that meets the two above relations are known, then we speak of sometimes of "isostatic system".

By definition, a "couple of force" or simply "couple" is defined as a set of two forces of equal magnitude but opposite direction, acting along two parallel lines on the same extended body. The resultant of course no forces indicates that the couple produces no translation effect. But the sum of the moments is not zero, the body undergoes a rotation such that:

$$\vec{M}=\vec{r}_1\times\vec{F}_1+\vec{r}_2\times\vec{F}_2=\vec{r}_1\times\vec{F}_1-\vec{r}_2\times\vec{F}_1=(\vec{r}_1-\vec{r}_2)\times\vec{F}_1=\vec{b}\times\vec{F} \tag{9.3.165}$$

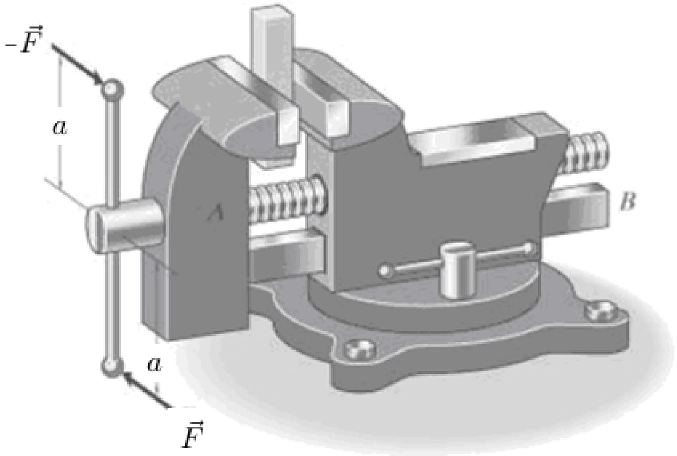


Figure 9.31: **Example of couple of force**

To avoid any confusion here is a small refresh:

- A Moment is equivalent to a force multiplied by the length of the line passing through the point of reaction and that is perpendicular to the force.
- Torque is a moment that is applied in such a way that it tends to rotate a body around its axis. The usual example being a torsional moment on a shaft.
- A Couple is the moment that is the resultant of two forces of the same magnitude, acting in opposite direction at the same distance from the reaction point. Therefore, this force system is statically equivalent to a pure moment (no resultant force).

Let us also mention the composition of force moments with the following ultra-classic case of balanced moment:

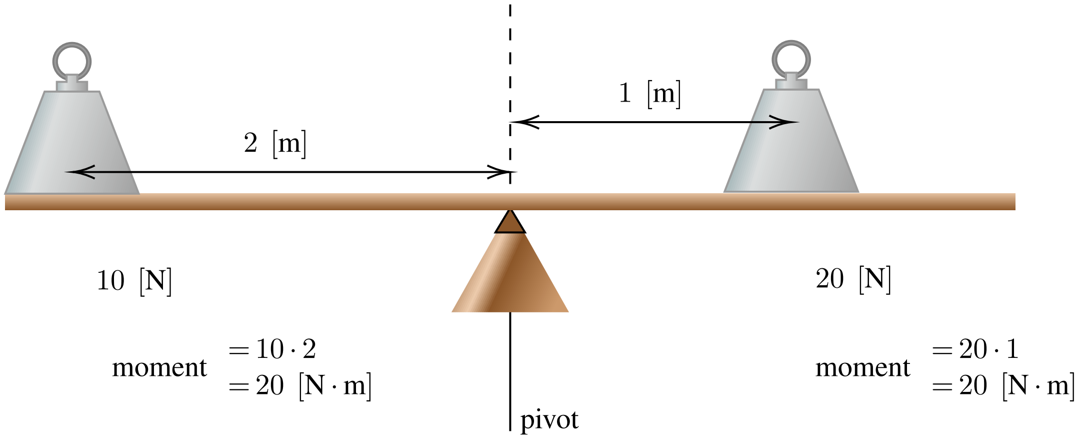


Figure 9.32: **Balanced moment example**

And also the analysis of the typical well known situation involving two moments of force:

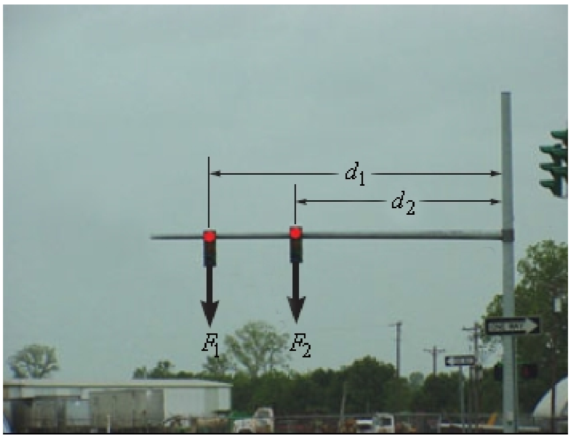


Figure 9.33: **Composition of force moments**

for which it is quite obvious that the force of the resulting moment of force has magnitude the sum of the two elementary forces:

$$F_R = F_1 + F_2 \tag{9.3.166}$$

To calculate the corresponding resultant distance, we will use the theorem of the center of mass, which as we have prove above and in the section of Euclidean Geometry is given in the general case by:

$$\overrightarrow{OG} = \frac{\sum\limits_{i=1}^{n} \alpha \overrightarrow{OA_i}}{\sum\limits_{i=1}^{n} \alpha_i} \tag{9.3.167}$$

and that in the present case is reduced to:

$$d = \frac{\sum\limits_{i=1}^{n} F_i d_i}{\sum\limits_{i=1}^{n} F_i} = \frac{F_1 d_1 + F_2 d_2}{F_1 + F_2} \tag{9.3.168}$$

We thus have the resulting final moment which is:

$$M_R = F_r d = (F_1+F_2)\frac{F_1 d_1 + F_2 d_2}{F_1+F_2} = F_1 d_1 + F_2 d_2 = M_1 + M_2 \tag{9.3.169}$$

Thus, the resultant moment of force is the simple sum of the elementary moment of forces of forces.

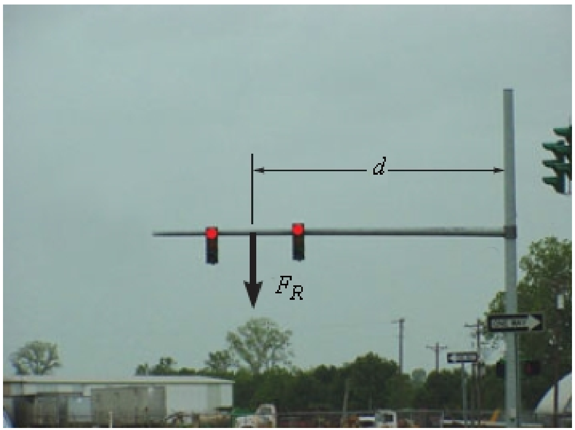


Figure 9.34: **Resulting composition of force moments**

Now that we have quite properly defined what was a force and moment of force, we can immediately begin the study of static forces:

9.3.5 Static Forces

The study field of static forces is a quite difficulty domain to generalize when the target students discover the concept of forces and momentum. Most books use many examples (like pulley systems, levers, balances, friction, truss, etc.) to get the reader to assimilate the method of analysis needed to resolve problems in this field of Classical Mechanics. Far from being against this method, we do not want to restrict ourselves or dwell (following viewpoints) to specific examples, but we have preferred to propose a method of analysis that would work without fail in any case.

Definitions (#9.3.438):

D1. The "static forces" is the field of physics that studies the effect of the resulting forces (or moments of force) that are constant over time, applied on a punctual or extended body.

D2. When the sum of all vector forces and moments of force is zero, there is no movement for a body initially not moving. We then speak of "static equilibrium" (but the forces still exist within the system) such that the forces and moments of forces cancel each other:

$$\sum_{i=1}^{n} \vec{F}_i = \vec{0} \quad \text{or} \quad \sum_{i=1}^{n} \vec{M}_i = \vec{0} \tag{9.3.170}$$

More generally an object is in equilibrium if the linear momentum of its center of mass is constant and if its angular momentum about its center of mass is constant.

Remark

The previous relations show us that it is not because a system is in static equilibrium that is not subject to any force (the vector sum of forces can cancel but the forces are non-zero)!

We have then the following corollaries:

C1. When analysing a static forces system, we must always (!!!) work with vector components of forces and moments of forces (by the first Newton's law!).

C2. We must therefore impose a reference frame in which all forces components will be expressed:

- In the case of punctual body (particle point) on which forces are applied, we must assimilate the origin of the coordinate to the position of that point.
- If the extension lines of all forces applied on an extended body are all converging at a given point, the system can be considered as an isolated body reduced at this point of convergence.
- If the extended body is plunged into an isotropic, coplanar and constant over time field of forces (gravitational, electrostatic, magnetic ...) then all applied forces can be reported to the center of gravity.

Proof 9.4.2. We have seen during our study of the section of Vector Calculus that the sum of vectors of a same set, put extremity by extremity (at the pictorial representation level) or algebraically added constitute what we name the "resultant" of the system of forces or momentum forces (torques):

$$\vec{R}_{\vec{F}}=\sum_{i=1} n\vec{F}_i \quad \text{and/or} \quad \vec{R}_{\vec{M}}=\sum_{i=1}^{n}\vec{M}_i \tag{9.3.171}$$

It is clear that a material point is therefore defined in a static state if the resultant of the competing forces is equal to zero. Thus, a punctual body is at rest (constant speed equal to zero) if the quantity $\vec{R}_{\vec{F}}$ is zero.

This condition is however not sufficient for an extended body (non-punctual): the latter can not move (no translation movement), but can turn on itself by the application forces outside its barycenter (the forces are then momentum forces- torques - acting on points of the body in question).

Now let us imagine a set of forces $\vec{F}_i$, each of them applied at a point of position-vector $\vec{r}_i$ of an extended mobile and all parallel to a given common direction, indicated by a unit vector $\vec{u}$. The resultant of these forces is then:

$$\vec{R}_{\vec{F}}=\sum_{i=1}^{n}\vec{F}_i=\sum_{i=1}^{n}F_i\vec{u}=\left(\sum_{i=1}^{n}F_i\right)\vec{u} \tag{9.3.172}$$

The resulting norm is then obviously:

$$\|\vec{R}_{\vec{F}}\|=\sum_{i=1}^{n}F_i \tag{9.3.173}$$

Similarly, the vector sum of the parallel corresponding moment of forces (torques) are written:

$$\vec{R}_{\vec{M}}=\sum_{i=1}^{n}\vec{r}_i\times\vec{F}_i=\sum_{i=1}^{n}\vec{r}_i\times(F_i\vec{u})=\left(\sum_{i=1}^{n}\vec{r}_iF_i\right)\times\vec{u} \tag{9.3.174}$$

Let us now search the vector position $\vec{r}_C$ of a fictive point C, named the "center of force" such as the moment of the resultant $\vec{R}_{\vec{F}}$ applied at point C is equal to the total torque $\vec{R}_{\vec{M}}$. In other words, $\vec{r}_C$ must be the solution of the vector equation:

$$\vec{R}_{\vec{M}}=\vec{r}_C\times\vec{R}_{\vec{F}} \tag{9.3.175}$$

If it is possible to find such a point C, so we do not longer need, in principle, to calculate the individual force momentum of each force and make the vector sum. Rather it is enough to determine the resultant $\vec{R}_{\vec{F}}$ and evaluate its resultant momentum force (torque) applied to the fictive point C.

Combining the above relations, we have:

$$\vec{R}_{\vec{M}}=\vec{r}_C\times\vec{R}_{\vec{F}}=\vec{r}_C\times\left(\sum_{i=1}^{n}F_i\vec{u}\right)=\left(\sum_{i=1}^{n}F_i\right)\vec{r}_C\times\vec{u}=\left(\sum_{i=1}^{n}F_i\vec{r}_C\right)\times\vec{u} \tag{9.3.176}$$

But we showed just before that:

$$\vec{R}_{\vec{M}}=\left(\sum_{i=1}^{n}\vec{r}_iF_i\right)\times\vec{u} \tag{9.3.177}$$

thus by equating:

$$\left(\sum_{i=1}^{n}F_i\vec{r}_C\right)\times\vec{u}=\left(\sum_{i=1}^{n}\vec{r}_iF_i\right)\times\vec{u} \tag{9.3.178}$$

It follows:

$$\sum_{i=1}^{n}F_i\vec{r}_C=\sum_{i=1}^{n}\vec{r}_iF_i \tag{9.3.179}$$

From which we finally draw:

$$\boxed{\vec{r}_C=\frac{\displaystyle\sum_{i=1}^{n}\vec{r}_iF_i}{\displaystyle\sum_{i=1}^{n}F_i}} \tag{9.3.180}$$

as $F_i=m_ia_i$ (Newton's second law) let us now assume (special case but well known in astronomy when we deal with gravitational force at almost constant distance) that $\forall i,j:a_i=a_j$ we can then write this very important result:

$$\vec{r}_C=\frac{\cancel{a_i}\displaystyle\sum_{i=1}^{n}\vec{r}_im_i}{\cancel{a_i}\displaystyle\sum_{i=1}^{n}m_i}=\frac{\displaystyle\sum_{i=1}^{n}\vec{r}_im_i}{\displaystyle\sum_{i=1}^{n}m_i} \tag{9.3.181}$$

We recognize in the last expression the center of mass that we have studied earlier above:

$$\boxed{\vec{r}_G=\frac{\displaystyle\sum_{i=1}^{n}\vec{r}_im_i}{\displaystyle\sum_{i=1}^{n}m_i}} \tag{9.3.182}$$

■ Q.E.D.

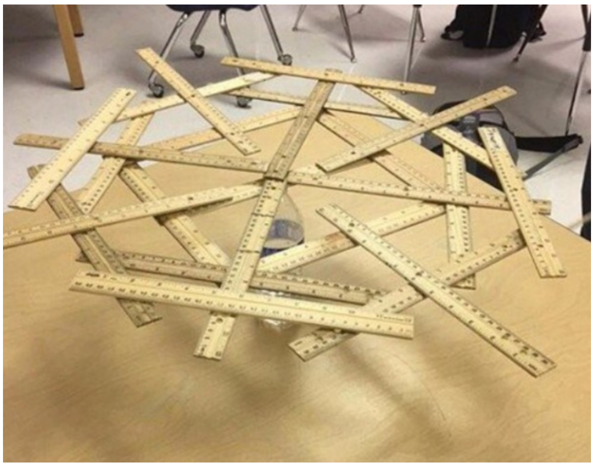
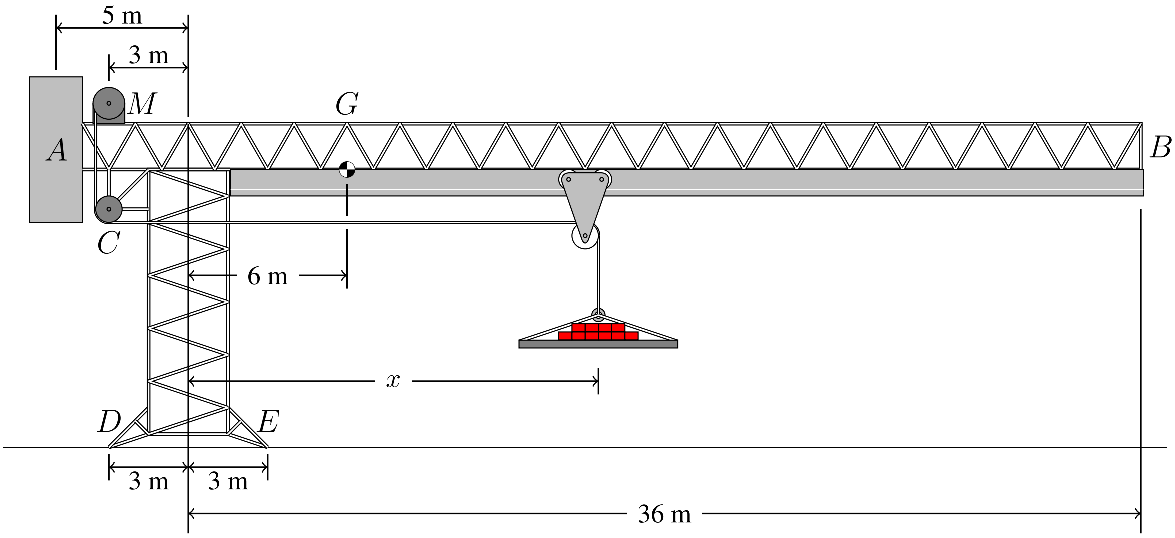


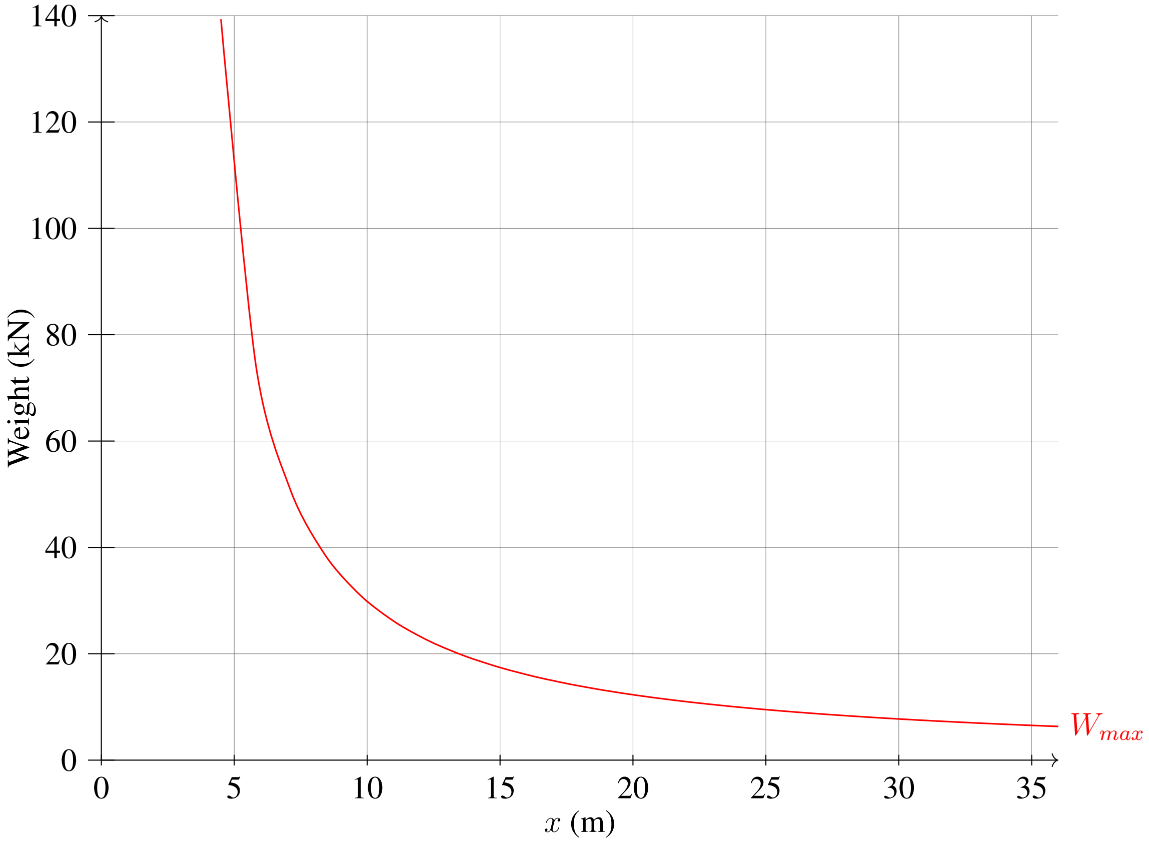
Figure 9.35: **Funny application example of what we have just proved**

C3. By the third Newton's law, any rigid solid body in stable equilibrium in contact with a set of rigid solids also in stable equilibrium, suffer all of the same equal force at each point of contact (identically distributed) but opposed by these latest (and assimilated through their center of gravity when it is an isotropic and constant vector field that is causing the contact). Therefore:

- The application points of action/reaction forces must be placed on the various points of contact when it is a countable amount of forces that are behind and that are not parallel.
- The application points of action/reaction forces can be places at the center of mass or gravity if the forces causing the contact (verbatim: acceleration) are the source of a gravitational vector field, respectively electrostatic / magnetic (meaning the forces are in a first approximation all parallels).

For example a typical problem in the analysis of the static of forces is to find the maximum weight that a crane could carry as a function of distance before it would tip over (the supports at D and E aren't bolted to the ground).

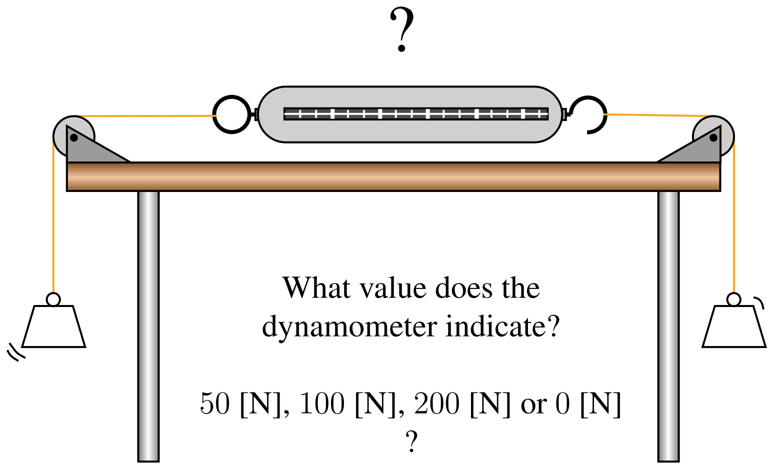




Below, the reader can see a photo of the biggest crane in the world in the early 121st century (Sarens SGC-250, $5,000$ Tonne lifting capacity, 250 [m] tall):

Figure 9.36: **Sarens SGC-250**

A simple example to check if someone masters the basics of static is to show the figure below and ask for an answer:


?
What value does the dynamometer indicate?
50 [N], 100 [N], 200 [N] or 0 [N] ?

For information the correct answer is 100 [N].

9.3.5.1 Lami's theorem

"Lami's theorem" states that when three coplanar, concurrent, non-collinear forces keep a body in static equilibrium, each force is proportional to the sine of the angle between the other two forces. In formula form, this is expressed as :

$$\frac{F_A}{\sin(\alpha)}=\frac{F_B}{\sin(\beta)}=\frac{F_C}{\sin(\gamma)} \tag{9.3.183}$$

This relationship arises because the three force vectors must form a closed triangle when placed tip-to-tail, and the law of sines (see section Trigonometry page 2528) then applies to that triangle.

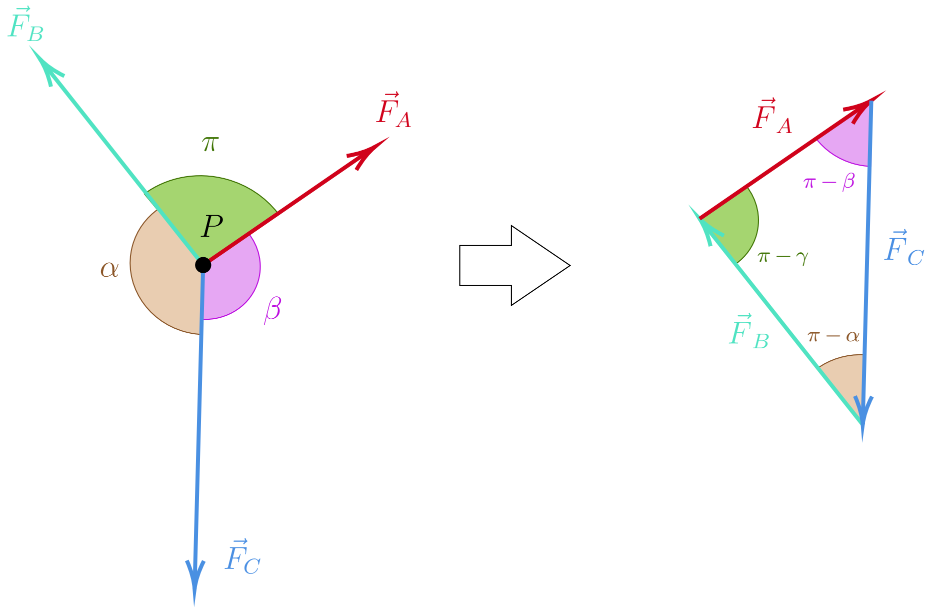

$\vec{F}_B$
$\vec{F}_A$
π
P
α
β
$\vec{F}_C$
$\vec{F}_A$
$\pi-\beta$
$\vec{F}_C$
$\pi-\gamma$
$\vec{F}_B$
$\pi-\alpha$

Figure 9.37: Lami's theorem illustrated

Proof 9.4.3. As the vectors must balance $\vec{F}_A+\vec{F}_B+\vec{F}_C=\vec{0}$, hence by making all the vectors touch its tip and tail the result is a triangle with sides F_A, F_B, F_C and angles $180°-\alpha, 180°-\beta, 180°-\gamma$ where α,β,γ are the exterior angles.

By the law of sines then:

$$\frac{F_A}{\sin(180°-\alpha)}=\frac{F_B}{\sin(180°-\beta)}=\frac{F_C}{\sin(180°-\gamma)} \tag{9.3.184}$$

Then by applying that for any angle $\theta, \sin(180°-\theta)=\sin\theta$ (supplementary angles have the same sine), and the result is:

$$\frac{F_A}{\sin(\alpha)}=\frac{F_B}{\sin(\beta)}=\frac{F_C}{\sin(\gamma)} \tag{9.3.185}$$

■ Q.E.D.

Lami's theorem is widely used in statics, mechanical engineering, and structural engineering, especially for analyzing systems where exactly three forces act at a point. Typical applications include trusses, frames, bridges, cranes, and other structures where loads meet at joints or nodes. It is particularly useful for solving equilibrium problems quickly when only three forces are involved and the geometry is known.

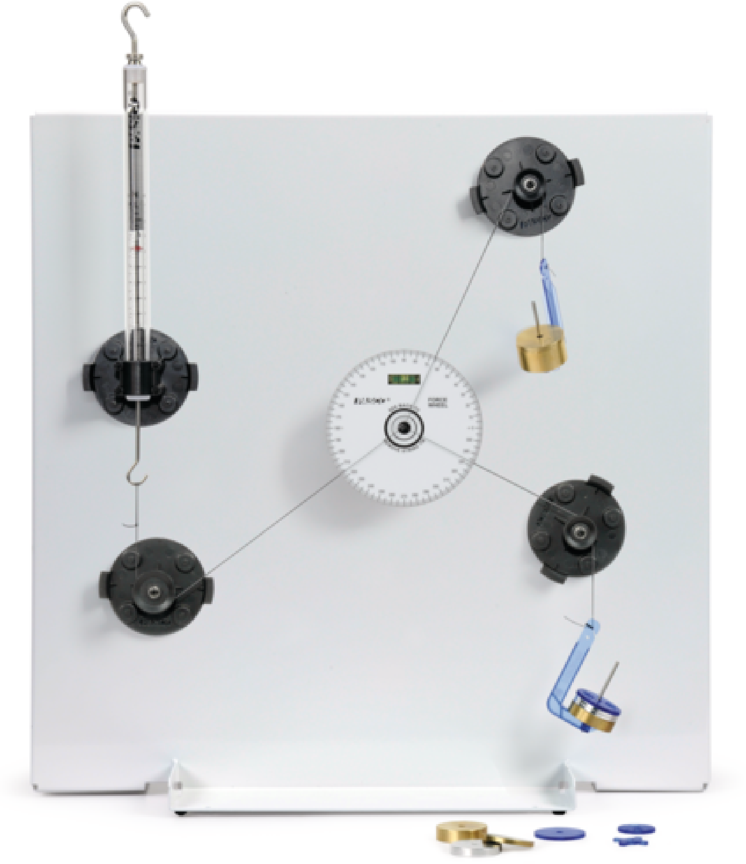


Figure 9.38: **PASCO® scientific static system to check Lami's theorem in schools**

9.3.6 Ballistics

Definition (#9.3.439): "Ballistics" is the science or study of the motion path of a body having an initial velocity through a given medium in space immersed in a given (most of time: constant) field (gravitational, magnetic or electrostatic).

> Remark
>
> We have hesitate, and still do, to write this subject in the section of Spatial Engineering. So it is not impossible that in a near of or far future this subject moves in another section.

We will prove now thanks to our knowledge in kinematics that the parabolic movement is the natural (approximative) movement of a mobile body in the field of gravity in empty space animated (without friction or where friction can be neglected) by a travel speed $\vec{v}_0$ not necessarily parallel to the acceleration of gravity $\vec{g}$ (in general this is therefore a curvilinear motion). For example, a projectile having an initial velocity $\vec{v}_0$ inclined from an angle β from the horizontal:

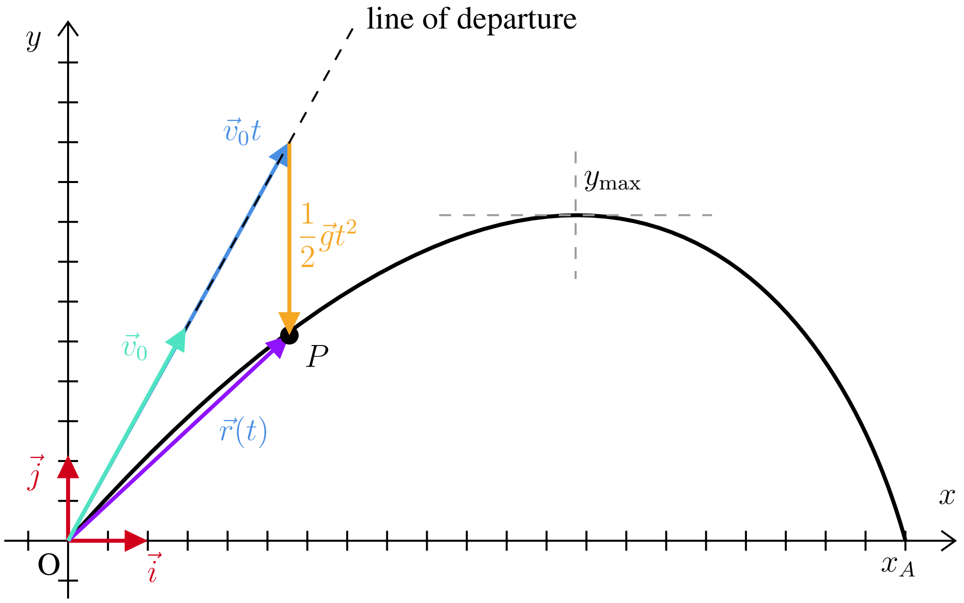


Figure 9.39: **Ballistics configuration study**

And with some vocabulary stuff associated:

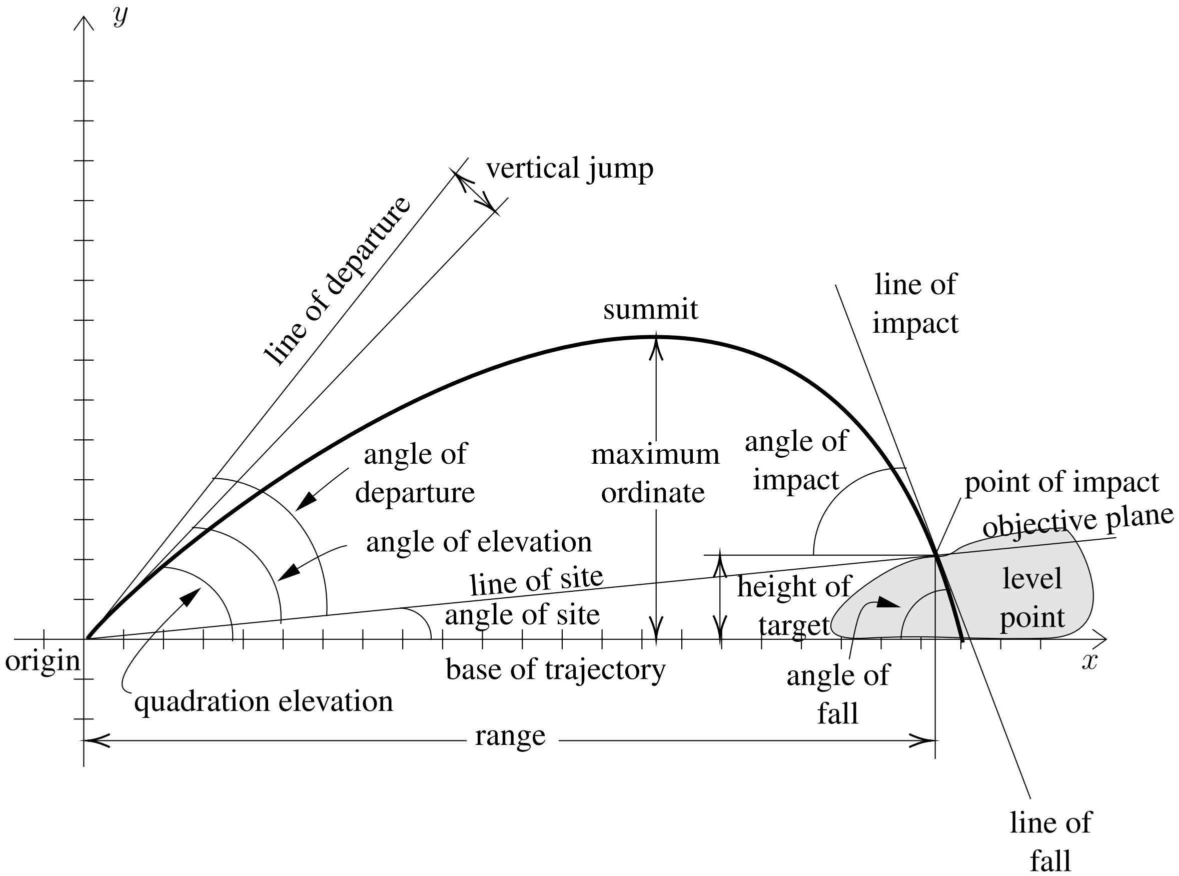

y
vertical jump
line of departure
summit
line of
impact
angle of
departure
maximum
ordinate
angle of
impact
point of impact
objective plane
angle of elevation
line of site
angle of site
height of
target
level
point
origin
base of trajectory
quadration elevation
angle of
fall
x
range
line of
fall

Figure 9.40: **Typical Artillery vocabulary in ballistics**

Obviously the reader can ask himself if the study of ballistics is useful outside artillery?! In fact even if does not apply to rocket launch as their velocity is not constant (boosters) and the application to golf is not quite interesting for engineers we have some practical situation where such calculations are useful. A contemporary one is the "ice ballistic problem" for the security radius around wind turbine to avoid people or houses to receive blocks of ice (in winter) that detach from the pales in rotation at a given rotation speed ω and for sure the trajectories if these ice blocks are ballistic trajectories!!!

For pedagogical reasons we will first do the developments assuming wrongly a flat Earth...

Before we start let us give a definition!

Definition (#9.3.440): An "elevation angle" is the angular measurement between the reference plane tangent to the observer (instrument) position and a line of sight directed toward an object located above that tangent plane.

9.3.6.1 Ballistic equation without friction

In the absence of gravity and friction the mobile P would follow the straight departure line indefinitely. The action of the gravity is to take it down, at time t, of the well known value to us $\dfrac{1}{2}gt^2$.

We put the projection on the axes:

$$\vec{r}(t)=v_0\cos(\beta)t\vec{i}+\left(v_0\sin(\beta)t-\frac{1}{2}gt^2\right)\vec{j} \tag{9.3.186}$$

combination of smooth movement along the x-axes and a falling motion with initial velocity v_{0y} along the y-axes. Which corresponds to the following equations:

$$\begin{aligned}x(t)&=v_0\cos(\beta)t\\ y(t)&=v_0\sin(\beta)t-\frac{1}{2}gt^2\end{aligned} \tag{9.3.187}$$

eliminating the time between these two equations we get the "parabolic trajectory equation (without friction)":

$$\boxed{y=x\tan(\beta)-\frac{g}{2v_0^2\cos^2(\beta)}x^2} \tag{9.3.188}$$

which had already been obtained by Galileo in the early 117th century (holocene calendar).

Remark

Notice that the same relation applies wit the Electric force (in non-relativistic speed) where the acceleration g must simply be replaced by qE/m (see section Electrostatics page 3968).

We calculate also the target range $x_{\max}$ of the projectile by putting $y=0$ in the above equation and we get easily :

$$\boxed{x_{\max}=\frac{v_0^2}{g}\sin(2\beta)} \tag{9.3.189}$$

the solution $x=0$ have no interest.

The maximum height $y_{\max}$ can be calculated by cancelling the derivative of the equation of the trajectory. So we get easily:

$$\boxed{y_{\max}=\frac{v_0^2}{2g}\sin^2(\beta)} \tag{9.3.190}$$

We notice that for the maximum range for a given initial speed we have two typical cases to consider:

1. We give ourselves and unreachable target range $x_{\max}$ as it is not possible that $\sin(\beta)>1$.

2. As we have $\sin(2\beta)$ in the expression of $x_{\max}$ it is obvious that it is $\beta = \pi/4$ ($45°$) that gives the maximum range for a given initial speed.

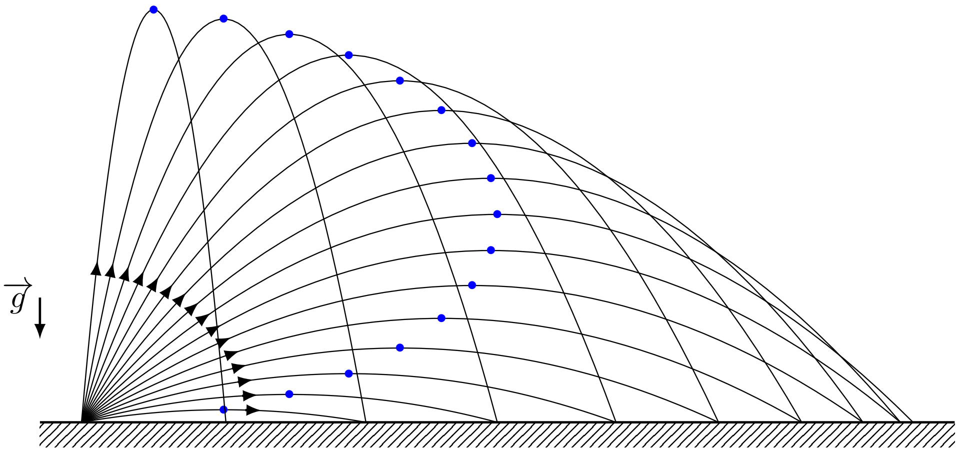


Figure 9.41: **Influence of the angle β on the trajectory (source: Femtophysique)**

The curve enveloping all parables, plotted for a given initial speed v_0 in all possible directions, is still a parable named "safety parabola". Its rotation about the y-axis creates a paraboloid which circumscribes (contains) the region of the space accessible only to projectiles.

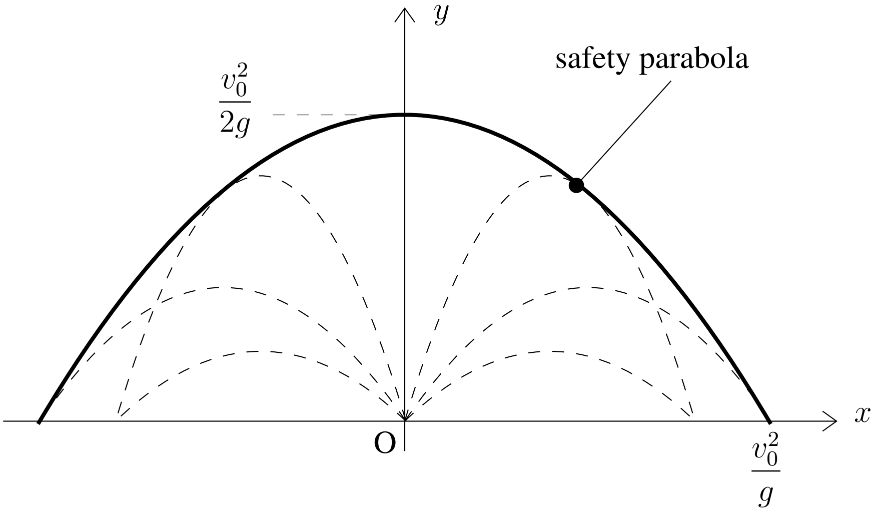


Figure 9.42: **Ballistic safety parabola**

Thus, it is not too difficult to find the equation of this safety parabola !

The vertically shot ($\beta = \pi/4$) is known to us and is given by:

$$v_0 = \sqrt{2gy_{\max}} \Leftrightarrow y_{\max} = \frac{v_0^2}{2g} \tag{9.3.191}$$

The maximal range is as we know given by:

$$x_{\max} = \frac{v_0^2}{g} \tag{9.3.192}$$

So when $x = x_{\max} \Rightarrow y = 0$ such that:

$$y_{\text{safety}} = \frac{v_0^2}{2g} - \frac{g}{2v_0^2}x^2 \tag{9.3.193}$$

which is the equation of the safety parabola.

The time of flight is the time it takes for the projectile to finish its trajectory. As we have for the vertical component:

$$y(t) = v_0 \sin(\beta)t - \frac{1}{2}gt^2 \tag{9.3.194}$$

and the "time of flight" is given when $y = 0$, it comes therefore:

$$v_0 \sin(\beta)t = \frac{1}{2}gt^2 \tag{9.3.195}$$

That is to say:

$$\boxed{t_{\text{flight}} = 2v_0\frac{\sin(\theta)}{g}} \tag{9.3.196}$$

The "angle of reach" is the angle β at which a projectile must be launched in order to go a distance $x_{\max}$, given the initial velocity v_0. So from:

$$x_{\max} = \frac{v_0^2}{g}\sin(2\beta) \tag{9.3.197}$$

We get:

$$\boxed{\beta_{\text{reach}} = \frac{1}{2}\arcsin\left(\frac{x_{\max}g}{v_0^2}\right)} \tag{9.3.198}$$

Ok now much more interesting... To hit a target at range x_T and altitude y_T when fired from $(0,0)$ and with initial speed v_0 the required angle(s) of launch θ are:

$$\theta = \arctan\left(\frac{v_0^2 \pm \sqrt{v^4 - g(gx_T^2 + 2y_Tv_0^2)}}{gx_T}\right) \tag{9.3.199}$$

The two roots of the equation correspond to the two possible launch angles, so long as they aren't imaginary, in which case the initial speed is not great enough to reach the point (x,y) selected.

Proof 9.4.4. First, let us recall that:

$$\begin{aligned} x_T(t) &= v_0\cos(\beta)t \\ y_T(t) &= v_0\sin(\beta)t - \frac{1}{2}gt^2 \end{aligned} \tag{9.3.200}$$

Solving the first for t and substituting this expression in the second gives (using the trigonometric identities proved in the section of Trigonometry):

$$y_T = x_T\tan(\beta) - \frac{gx_T^2}{2v_0^2\cos^2(\beta)} = x\tan(\beta) - \frac{gx_T^2}{2v_0^2}(1+\tan^2(\beta)) \tag{9.3.201}$$

After a small rearrangement:

$$\frac{-gx_T^2}{2v_0^2}\tan^2(\beta)+x\tan(\beta)-\frac{gx_T^2}{2v_T^2}-y_T=0 \tag{9.3.202}$$

Let us put $\Gamma=\tan(\beta)$:

$$\frac{-gx_T^2}{2v_T^2}\Gamma^2+x\Gamma-\frac{gx_T^2}{2v_0^2}-y_T=0 \tag{9.3.203}$$

Solving this second order polynomial as we learned to do it in the section Calculus we have for discriminant (see page 1383) :

$$\Gamma=\frac{-x_T\pm\sqrt{x_T^2-4\left(\frac{-gx_T^2}{2v_0^2}\right)\left(\frac{-gx_T^2}{2v_0^2}-y_T\right)}}{2\left(\frac{-gx_T^2}{2v_0^2}\right)} \tag{9.3.204}$$

So we get after simplifications:

$$\tan(\beta)=\frac{v_0^2\pm\sqrt{v_0^4-g(gx_T^2+2y_Tv_0^2)}}{gx_T} \tag{9.3.205}$$

Hence:

$$\boxed{\beta=\tan^{-1}\left(\frac{v_0^2\pm\sqrt{v_0^4-g(gx_T^2+2y_Tv_0=^2)}}{gx_T}\right)} \tag{9.3.206}$$

If instead of a coordinate (x,y) it is required to hit a target at distance r and angle of elevation θ (polar coordinates), we use the relations $x_T=r_T\cos(\theta_T)$, $y_T=r_T\sin(\theta_T)$ and substitute to get:

$$\beta=\tan^{-1}\left(\frac{v_0^2\pm\sqrt{v_0^4-g(gr_T^2\cos^2(\theta_T)+2v_0^2r_T\sin(\theta_T))}}{gr_T\cos(\theta_T)}\right) \tag{9.3.207}$$

■ Q.E.D.

Remark

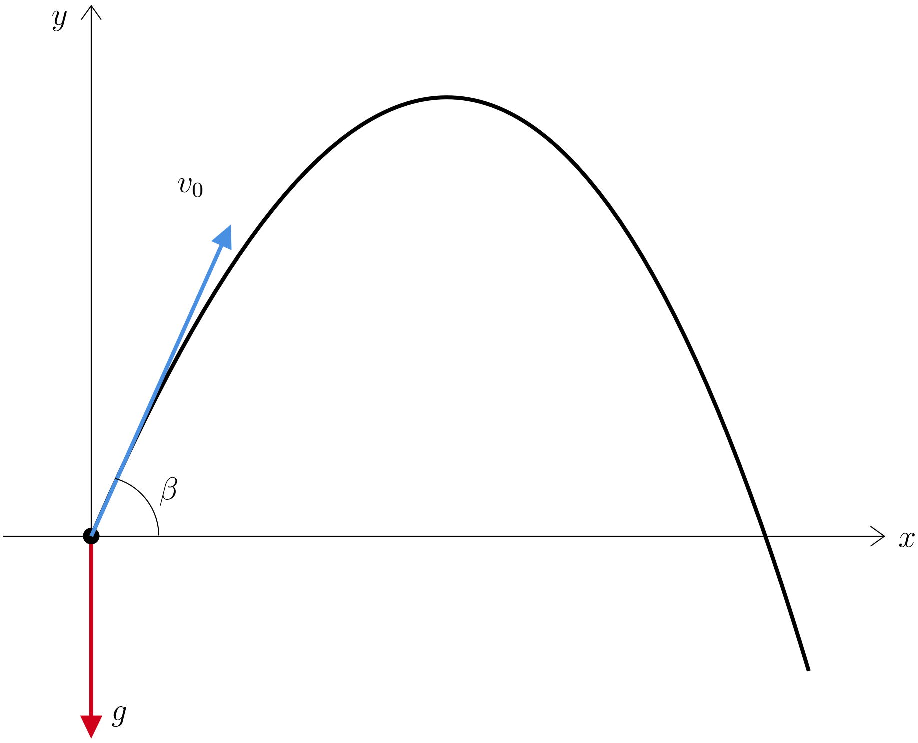
Keep in mind that if what interest you is the horizontal distance traveled by an item with an initial angle $\beta=0$ and initial speed v_0 under gravity g and without friction (in vacuum) from a height h, then we have:

$$x_{\max}=x_0+v_0t+\frac{1}{2}at^2=v_0t=v_0\sqrt{\frac{2h}{g}} \tag{9.3.208}$$

Therefore from a height of $h=1.70$ [m], we have $x_{\max}\cong v_0\cdot 0.58$. So as a breath is estimated at $v_0\cong 1\,[\text{m}\cdot\text{s}^{-1}]$, and sneeze germs at $v_0\cong 40\,[\text{m}\cdot\text{s}^{-1}]$, and a cough at $v_0\cong 10\,[\text{m}\cdot\text{s}^{-1}]$ just do the maths...

9.3.6.2 Ballistic equation with friction

Suppose that a projectile of mass m is launched, at $t=0$, from ground level (in a flat plain), making an angle β to the horizontal. Suppose, further, that, in addition to the force of gravity, the projectile is subject to an air resistance force which acts in the opposite direction to its instantaneous direction of motion, and whose magnitude is directly proportional to its instantaneous speed. This is not a particularly accurate model of the drag force (see section Fluid Mechanics page 3856) due to air resistance (the magnitude of the drag force is typically proportion to the square of the speed), but it does lead to tractable equations of motion...



Consider now the motion of the point mass in the presence of a resistant force proportional to the point velocity, i.e., of the form $-bv$, with $b>0$ denoting a suitable parameter. In this case Newton's law can be cast in the following form:

$$a=g-\frac{1}{\tau}v \tag{9.3.209}$$

Let us denote that:

$$\vec{a}=\vec{g}-\frac{1}{\tau}\vec{v} \tag{9.3.210}$$

where $\vec{v}=(v_x,v_y)$ is the projectile velocity, $\vec{g}=(0,-g)$ the acceleration due to gravity. In component form the above equation becomes:

$$\begin{aligned}\frac{\mathrm{d}v_x}{\mathrm{d}t}&=-g\frac{v_x}{v_T}\\ \frac{\mathrm{d}v_y}{\mathrm{d}t}&=-g\left(1+\frac{v_y}{v_T}\right)\end{aligned} \tag{9.3.211}$$

where $v_T = mg/b$ is named the "terminal velocity" i.e., the velocity at which the drag force balances the gravitational force (for a projectile falling vertically downward).

In component form, the above equation becomes:

$$\begin{aligned}\frac{\mathrm{d}v_x}{\mathrm{d}t} &= -g\frac{v_x}{v_T}\\ \frac{\mathrm{d}v_y}{\mathrm{d}t} &= -g\left(1+\frac{v_y}{v_T}\right)\end{aligned} \tag{9.3.212}$$

Integrating the first equation above we get:

$$\int\limits_{v_{x0}}^{v} \frac{dv_x}{v_x} = -\frac{g}{v_T}t \tag{9.3.213}$$

Here $v_{x0} = v_0\cos(\beta)$ is the initial x-component of the launch velocity. Hence:

$$v_x = v_0\cos(\beta)e^{-gt/v_T} \tag{9.3.214}$$

And integration yields to (this takes into account that the constant of integration must satisfy $x(0) = 0$:

$$x = \frac{v_0 v_T\cos(\beta)}{g}\left(1-e^{-gt/v_T}\right) \tag{9.3.215}$$

In the limit $t \ll v_T/g$, the above relation reduces to:

$$x = v_0\cos(\beta t) \tag{9.3.216}$$

Which is the standard result in the absence of air drag. In the opposite limit, $t \gg v_T/g$, we get:

$$x = \frac{v_0 v_T\cos(\beta)}{g} \tag{9.3.217}$$

The above expression clearly sets an effective upper limit on how far the projectile can travel in the horizontal direction!

Now let us come back to the y component! Remember that:

$$\begin{aligned}\frac{\mathrm{d}v_x}{\mathrm{d}t} &= -g\frac{v_x}{v_T}\\ \frac{\mathrm{d}v_y}{\mathrm{d}t} &= -g\left(1+\frac{v_y}{v_T}\right)\end{aligned} \tag{9.3.218}$$

Integrating the second equation will be written:

$$\int\limits_{v_{y0}}^{v_z} \frac{\mathrm{d}v_y}{v_T+v_y} = -\frac{g}{v_T}t \tag{9.3.219}$$

where $v_{y0} = v_0\sin(\beta)$ is the y-component of the launch velocity. Hence:

$$\ln\left(\frac{v_T+v_y}{v_T+v_{y0}}\right) = -\frac{g}{v_T} \tag{9.3.220}$$

or:

$$v_y = v_0 \sin(\beta) e^{-gt/v_T} - v_T \left(1 - e^{-gt/v_T}\right) \tag{9.3.221}$$

Integration gives:

$$y = \frac{v_T}{g}\left(v_0 \sin(\beta) + v_T\right)\left(1 - e^{-gt/v_T}\right) - v_T t \tag{9.3.222}$$

In the limit $t \ll v_T/g$, this equation reduces to:

$$y = v_0 \sin(\beta) t - \frac{g}{2} t^2 \tag{9.3.223}$$

which is similar to the standard result in the absence of air drag. In the opposite limit, $t \gg v_T/g$, we get:

$$y = \frac{v_T}{g}\left(v_0 \sin(\beta) + v_T\right) - v_T t \tag{9.3.224}$$

So finally:

$$\boxed{\begin{aligned} x &= \frac{v_0 v_T \cos(\beta)}{g}\left(1 - e^{-gt/v_T}\right) \\ y &= \frac{v_T}{g}\left(v_0 \sin(\beta) + v_T\right) - v_T t \end{aligned}} \tag{9.3.225}$$

We know that in the absence of air resistance, the maximum horizontal range is achieved when the launch angle β takes the value $45°$ ($\pi/4$ [rad]).

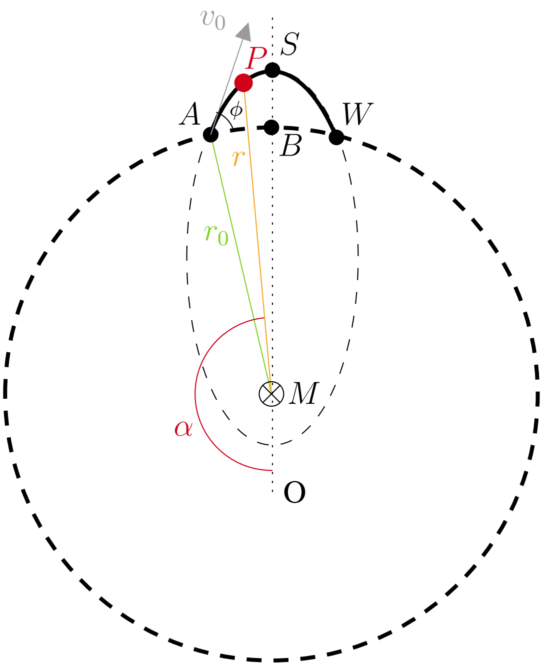
On the other hand in the presence of air resistance, the maximum horizontal range is achieved when β is made as small as possible!

9.3.6.3 Ballistic accounting Earth's curvature

We will focus now on the trajectory in a vacuum, taking into account the decrease of gravity with the height and the convergence of vertical lines from the curvature of the Earth.

The text that follows is almost a simple copy/paste with some improvements of the very good book *Handbook of Ballistics, Volume I.* of Dr C. Cranz, Cpt. K. Becker (11921 according to holocene calendar) from page 22 to 33 (see [156]).

Take the center M of the Earth (as in the figure below) as the pole in a system of polar coordinates; and let any arbitrary point P have the polar coordinates $\overline{MP} =$ radius vector r, and $\angle \text{O}MP =$ polar angle α; the direction OM of the polar axis from which the polar angle α is measured may be left undetermined at first:



At A, the point of departure of the shell, let $r=r_0=$ radius of the Earth, $6,370,300$ [m]; let the initial velocity be v_0, and the angle of departure ϕ.

According to Newton's law of gravitation (see section Astronomy page 5255), the acceleration of gravity is given by:

$$a=g\frac{r_0^2}{r^2}:=\frac{\mu}{r^2} \tag{9.3.226}$$

Moreover by second Kepler's law/area law (see section Astronomy page 5234):

$$r^2\frac{\mathrm{d}\alpha}{\mathrm{d}t} \tag{9.3.227}$$

is a constant, that we will denote here C, along the whole path of flight ; the value of (rearrangement of the previous relation):

$$r\frac{\mathrm{d}\alpha\cdot r}{\mathrm{d}t} \tag{9.3.228}$$

at the point A is given by (tangential velocity):

$$C=r_0v_0\cos(\phi) \tag{9.3.229}$$

since here:

$$r\mathrm{d}\alpha=\cos(\phi)\mathrm{d}s \tag{9.3.230}$$

where $\mathrm{d}s$ is the element of arc, and:

$$\frac{\mathrm{d}s}{\mathrm{d}t}=v_0 \tag{9.3.231}$$

we have:

$$r^2\frac{\mathrm{d}\alpha}{\mathrm{d}t}=C=r_0v_0\cos(\phi) \tag{9.3.232}$$

Further, in the motion of the shell along its path:

$$\frac{\mathrm{d}v}{\mathrm{d}t}=-\frac{\mu}{r^2}\frac{\mathrm{d}r}{\mathrm{d}s} \quad (9.3.233)$$

Or after rearranging:

$$v\frac{\mathrm{d}v}{\mathrm{d}r}=-\frac{\mu}{r^2} \quad (9.3.234)$$

Rearranging:

$$v\mathrm{d}v=-\frac{\mu}{r^2}\mathrm{d}r \quad (9.3.235)$$

and integrating from A to P and rearranging:

$$v^2-v_0^2=-2\mu\int\limits_{r_0}^{r} r^{-2}\mathrm{d}r=2\mu\left(\frac{1}{r}-\frac{1}{r_0}\right) \quad (9.3.236)$$

That we can rewrite:

$$v^2=q+\frac{2\mu}{r} \quad (9.3.237)$$

where:

$$q=v_0^2-\frac{2\mu}{r_0} \quad (9.3.238)$$

Since (see section Vector Calculus page 1865):

$$v^2=\left(\frac{\mathrm{d}s}{\mathrm{d}t}\right)^2=\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2+r^2\left(\frac{\mathrm{d}\alpha}{\mathrm{d}t}\right)^2 \quad (9.3.239)$$

and from:

$$r^2\frac{\mathrm{d}\alpha}{\mathrm{d}t}=C=r_0v_0\cos(\phi) \quad (9.3.240)$$

We have:

$$\frac{\mathrm{d}r}{\mathrm{d}t}=\frac{\mathrm{d}r}{\mathrm{d}\alpha}\frac{\mathrm{d}\alpha}{\mathrm{d}t}=\frac{\mathrm{d}r}{\mathrm{d}\alpha}\frac{C}{r^2} \quad (9.3.241)$$

Therefore:

$$v^2=\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2+r^2\left(\frac{\mathrm{d}\alpha}{\mathrm{d}t}\right)^2 \quad (9.3.242)$$

may be written:

$$v^2=q+\frac{2\mu}{r}=\left(\frac{\mathrm{d}r}{\mathrm{d}\alpha}\right)^2\frac{C^2}{r^4}+\frac{C^2}{r^2} \quad (9.3.243)$$

Or after rearranging:

$$\mathrm{d}\alpha=\frac{\dfrac{C}{r^2}\mathrm{d}r}{\sqrt{q+\dfrac{2\mu}{r}-\dfrac{C^2}{r^2}}}=\frac{-\mathrm{d}\dfrac{\dfrac{C}{r}-\dfrac{\mu}{C}}{\sqrt{q+\frac{\mu^2}{C^2}}}}{\sqrt{1-\left[\dfrac{\dfrac{C}{r}-\dfrac{\mu}{C}}{\sqrt{\left(q+\dfrac{\mu^2}{C^2}\right)}}\right]^2}} \tag{9.3.244}$$

This is the differential equation of the path of the shell, with r and α as the two variables. Integration gives (see section Differential and Integral Calculus page 1534):

$$\alpha-\gamma=\cos^{-1}\left(\frac{\dfrac{C}{r}-\dfrac{\mu}{C}}{\sqrt{q+\dfrac{\mu^2}{C^2}}}\right) \tag{9.3.245}$$

Hence after rearranging:

$$r=\frac{C}{\dfrac{\mu}{C}+\sqrt{q+\dfrac{\mu^2}{C^2}}\cos(\alpha-\gamma)} \tag{9.3.246}$$

That can be rewritten as:

$$r=\frac{p}{1+e\cos(\alpha-\gamma)} \tag{9.3.247}$$

if we put:

$$p=\frac{C^2}{\mu},\qquad e=\sqrt{1+\frac{qC^2}{\mu^2}} \tag{9.3.248}$$

We recognize that:

$$r=\frac{p}{1+e\cos(\alpha-\gamma)} \tag{9.3.249}$$

is the conic section equation (see section Analytical Geometry page 2782) in which we have parameter:

$$p=\frac{b^2}{a}=\frac{a^2-d^2}{a}=a-e\cdot d=a-e^2a=a\left(1-e^2\right) \tag{9.3.250}$$

where a and b are the two semi-axes, a the one that contains the focus, d the distance between center and focus of the conic section, e is the eccentricity.

One focus M is here the pole of the system of polar coordinates, and the polar angle α is measured from that vertex O of the major axis that lies nearest to the focus M; with $e<1$ we have an ellipse; with $e=0$, a circle; $e=1$ a parabola; $e>1$ an hyperbola.

When the integration constant γ is zero, the polar axis OM of the polar coordinate system is the line joining the perihelion O to the center of the Earth. We have then:

$$\begin{gathered}r=\frac{p}{1+e\cos(\alpha)}\\ v^2=v_0^2-\frac{2\mu}{r_0}+\frac{2\mu}{r}\end{gathered} \tag{9.3.251}$$

where:

$$p=\frac{C^2}{\mu},\qquad e=\sqrt{1+\frac{qC^2}{\mu^2}} \tag{9.3.252}$$

and:

$$q=v_0^2-\frac{2\mu}{r_0},\qquad \mu=gr_0^2,\qquad C=r_0v_0\cos(\phi),\qquad r_0=6,370,300\,[\text{m}] \tag{9.3.253}$$

Since r_0,ϕ,v_0, are known, and also C,μ,q,e,p, we are in a position to determine for any value of α the corresponding distance r of the shell from the center of the Earth from:

$$r=\frac{p}{1+e\cos(\alpha)} \tag{9.3.254}$$

and from:

$$v^2=v_0^2-\frac{2\mu}{r_0}+\frac{2\mu}{r} \tag{9.3.255}$$

the corresponding velocity v in the trajectory; the time of flight is obtained then from the integration of:

$$\mathrm{d}t=\frac{r^2\mathrm{d}\alpha}{C} \tag{9.3.256}$$

The trajectory is an ellipse when $e<1$, that is when:

$$\sqrt{1+\frac{C^2}{\mu^2}\left(v_0^2-\frac{2\mu}{r_0}\right)}<1,\ \text{or } v_0<\sqrt{\frac{2\mu}{r_0}} \tag{9.3.257}$$

Now:

$$\sqrt{\frac{2\mu}{r_0}}=\sqrt{2\cdot(9.81)\cdot 6,370,300}=11,050\,[\text{m}\cdot\text{s}^{-1}] \tag{9.3.258}$$

so that the path is an ellipse so long as v_0 remains $<11,050\,[\text{m}\cdot\text{s}^{-1}]$.

This elliptic orbit is a circle in the special case when $e=0$, or:

$$\sqrt{1+\frac{r_0^2v_0^2\cos^2(\phi)}{\mu^2}\left(v_0^2-\frac{2\mu}{r_0}\right)}=0 \tag{9.3.259}$$

Let us put:

$$\frac{r_0v_0^2}{\mu}=z \tag{9.3.260}$$

This equation becomes a second order polynomial:

$$z^2-2z+\frac{1}{\cos^2(\phi)}=0 \qquad (9.3.261)$$

Whose roots are (see section Calculus page 1382):

$$z=\frac{r_0{v_0}^2}{\mu}=1\pm\sqrt{1-\frac{1}{\cos^2(\phi)}} \qquad (9.3.262)$$

and this expression is real only when $\cos(\phi)=\pm1, \phi=0$ or π; in that case:

$$\frac{r_0v_0^2}{\mu}=1 \qquad (9.3.263)$$

Therefore:

$$v_0=\sqrt{\frac{\mu}{r_0}}=7,900\,[\mathrm{m}\cdot\mathrm{s}^{-1}] \qquad (9.3.264)$$

Thus under the specified assumptions the trajectory with any ordinary initial velocity v_0 is an ellipse; it is a parabola when $v_0=11,050\,[\mathrm{m}\cdot\mathrm{s}^{-1}]$; with still greater initial velocity it would be an hyperbola.

The trajectory can be a circle only in the special case when the shell is fired horizontally, with an initial velocity of about $7,900\,[\mathrm{m}\cdot\mathrm{s}^{-1}]$.

To calculate the range $\overline{AW}$ and the vertex ordinate $\overline{BS}$ of an elliptic trajectory from A, we should proceed as follows: first calculate the polar angle $\alpha=\angle\mathrm{O}MA$, which corresponds to the point A, from the relation:

$$r_0=\frac{p}{1+e\cos(\alpha_0)} \qquad (9.3.265)$$

which holds if A is to lie on the trajectory.

Twice the supplement of this angle is the polar angle $\angle AMW$; from this and r_0, the range $\overline{AW}$ can be calculated.

Further, the vertex ordinate $\overline{BS}$ of the trajectory $=\overline{MS}-r_0$, in which $\overline{MS}$ denotes the maximum value of r; this is when the denominator in:

$$r=\frac{p}{1+e\cos(\alpha)} \qquad (9.3.266)$$

assumes its least value, that is when $\cos(\alpha)=-1$. C'est-à-dire:

$$r_{\max}=\frac{p}{1-e} \qquad (9.3.267)$$

and:

$$r_{\min}=\frac{p}{1+e}=\overline{M\mathrm{O}} \qquad (9.3.268)$$

and thus the major axis of the ellipse is given by $r_{\max}+r_{\min}$. Thus:

$$\overline{BS}=\frac{p}{1-e}-r_0 \qquad (9.3.269)$$

✍**Example:**

Let us consider:

$$v_0=820\,[\mathrm{m}\cdot\mathrm{s}^{-1}]=2,952\,[\mathrm{km}\cdot\mathrm{h}^{-1}],\qquad \phi=44^\circ,\qquad r_0=6,370,300\,[\mathrm{m}] \tag{9.3.270}$$

Then:

$$e=\sqrt{1+\frac{qC^2}{\mu^2}}=\sqrt{1+\frac{v_0^2\cos^2(\phi)}{g^2r_0^2}\left(v_0^2-2gr_0\right)}=0.99445 \tag{9.3.271}$$

and:

$$\begin{aligned}\alpha_0&=\cos^{-1}\left(\frac{\frac{C}{r_0}-\frac{\mu}{C}}{\sqrt{q+\frac{\mu^2}{C^2}}}\right)=\cos^{-1}\left(\frac{v_0\cos(\phi)-\frac{gr_0}{v_0\cos(\phi)}}{\sqrt{v_0^2-2gr_0+\frac{g^2r_0^2}{v_0^2\cos(\phi)}}}\right)\\&=179^\circ41'23''26\cong 179.68972^\circ\end{aligned} \tag{9.3.272}$$

Hence:

$$\overline{AW}=\frac{2\cdot(180-\alpha_0)}{360^\circ}2\pi r_0\cong 68,995\,[\mathrm{m}] \tag{9.3.273}$$

with vertex height:

$$\overline{BS}=\frac{p}{1-e}=\frac{C^2/\mu}{1-e}\cong 16,620\,[\mathrm{m}] \tag{9.3.274}$$

On the other hand, in the parabolic path with the same v_0 and ϕ, the range will be $68,500\,[\mathrm{m}]$, and the vertex height $16,538\,[\mathrm{m}]$.

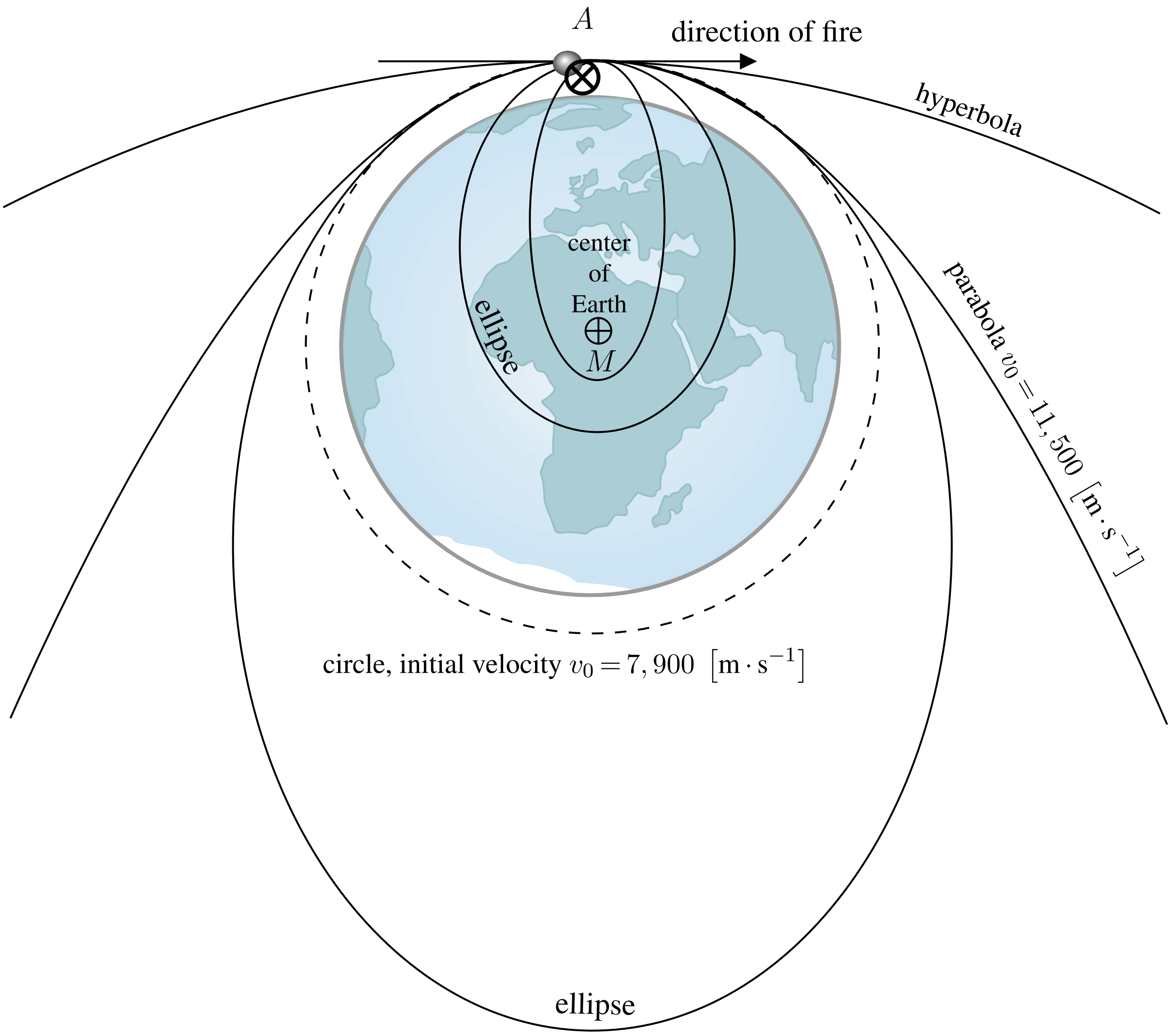
Hence the curvature of the Earth makes an alteration range of:

$$68,995-68,500=495\,[\mathrm{m}] \tag{9.3.275}$$

We therefore understand better why military have to take into account the curvature for long distance missiles!

It is interesting to discuss the trajectories if a shell is fired from the same point always in the same direction, but with increasing initial velocities v_0.

It is assumed in the figure below that shells are fired horizontally from an elevation A in the neighbourhood of the surface of the Earth.


A
direction of fire
hyperbola
parabola $v_0 = 11,500\ \left[\mathrm{m}\cdot\mathrm{s}^{-1}\right]$
center of Earth
M
ellipse
circle, initial velocity $v_0 = 7,900\ \left[\mathrm{m}\cdot\mathrm{s}^{-1}\right]$
ellipse

This is the famous "Newton's cannonball" that was a thought experiment Isaac Newton used to hypothesize that the force of gravity was universal, and it was the key force for planetary motion. It appeared in his posthumously published 11728 (holocene calendar) work *De mundi systemate* (also published in English as *A Treatise of the System of the World*).

As the initial velocity v_0 increases, so the ellipse broadens; the focus moves from A towards M; with $v_0 = 7,900\,\left[\mathrm{m}\cdot\mathrm{s}^{-1}\right]$ the shell describes a circular path round the Earth, in perpetual motion, and the movable focus comes into coincidence with the fixed focus M. In this case the shell flies round always at the same distance from the ground.

When the initial velocity increases still more, the shell at first goes further from the Earth's surface, in an ellipse, but returns again to A the point of departure. At the opposite side of the Earth the shell recedes more and more from the Earth's surface; the movable focus goes back beyond M, on the prolongation of the line AM.

With the velocity $v_0 = 11,050\,[\text{m}\cdot\text{s}^{-1}]$ the shell does not come back to A; the ellipse is changed into a parabola; the movable focus has receded to infinity.

As soon as this initial velocity $11,050\,[\text{m}\cdot\text{s}^{-1}]$ is exceeded, the trajectory is an hyperbola, the branch of which through A approximates more and more to the horizontal direction; the movable focus then approaches A on the line AM produced backward.

This result is of major importance for our study in the section of Aerospace Engineering of gravity assist where all practical cases involve hyperbolics trajectories (the only exception would be a theoretical, ideal case where the spacecraft has exactly the escape velocity, in which case the trajectory would be a parabola)!

9.3.7 Kinematics of Circular Motion (rotational kinematics)

The motion of objects as they translate – move bodily from one place to another – follows a simple set of rules. It turns out that a very similar set of rules describes the motion of objects as the rotate – spin around in place.

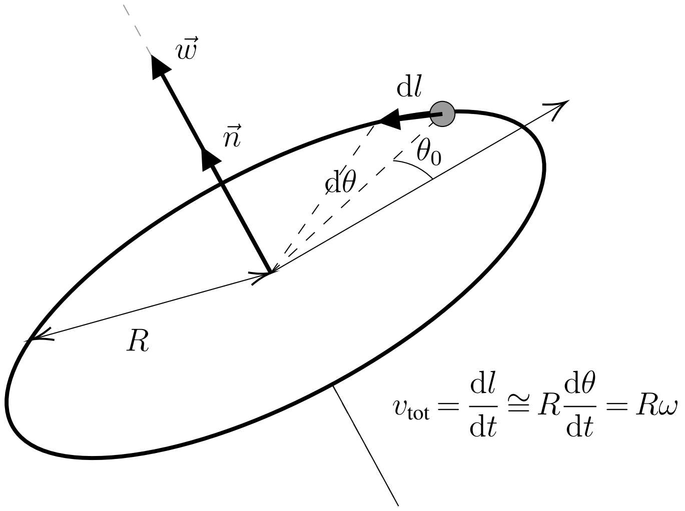
Most situations we have considered so far involve the type of motion named "translation" (pick it up and put it down somewhere else).

But centrifuges, planets, space stations, satellites, molecules, motor engines exhibit a different sort of motion: rotation (spit in place)!

The circular movement, also named "rotation movements", thus describe the rotation of an object around an axis (or a point to make simpler). The usage is to define it with the following information:

- The direction of the axis of rotation plane in space
- The direction of rotation on the constant radius circle around this axis
- The speed v on the circular path.

All this can be summarized in the following figure:


$\vec{w}$
$\mathrm{d}l$
$\vec{n}$
θ_0
$\mathrm{d}\theta$
R
$v_{\text{tot}}=\dfrac{\mathrm{d}l}{\mathrm{d}t}\cong R\dfrac{\mathrm{d}\theta}{\mathrm{d}t}=R\omega$

where for:

$$\boxed{v_{\text{rot}}=\frac{\mathrm{d}l}{\mathrm{d}t}\cong R\frac{\mathrm{d}\theta}{\mathrm{d}t}=R\omega} \tag{9.3.276}$$

we have used the relation proved in the section of Trigonometry (it is just logical by definition if the angle is measured in radians):

$$\mathrm{d}l=R\mathrm{d}\theta \tag{9.3.277}$$

It is therefore absolutely nonsense to express a rotation speed in "distance-per-unit-time" if the radius is not mentioned. A rotation speed of an object must be expressed in radians (or degrees) per unit-time if the dimension of the rotation radius is not explicitly given!

We summarize these three directions by the data given by a vector "instantaneous angular velocity":

$$\vec{\omega}=\omega\vec{n} \tag{9.3.278}$$

The direction of rotation is said to be "positive" when the thumb drawn in the direction of the unit vector $\vec{n}$, we take the axis of the right hand and one sees the object rotate in the direction of the other four fingers ("right hand rule"):

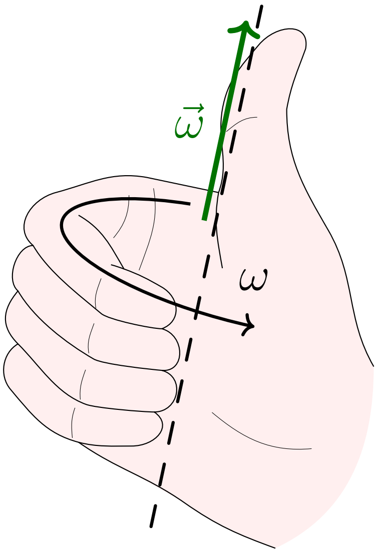


Figure 9.43: **Angular momentum right-handed rule (author: Izaak Neutelings)**

The norm of the instantaneous angular speed, represents obviously the angle traveled per unit of time,by the object that moves in the plane perpendicular to the unit vector $\vec{n}$:

$$\boxed{\omega=\frac{\mathrm{d}\theta}{\mathrm{d}t}} \tag{9.3.279}$$

Of course, it goes without saying that the angular velocity is given in radians per second and not in revolutions per minute or degrees per second!!! We must therefore be careful to always do the right conversion!

Remarks

R1. In the general case of the circular motion, angular velocity of the object studied varies over time: $\omega=\omega(t)$.

R2. When the direction of the rotation axis changes, the components of unit vector $\vec{n}$ are functions of time. This is the case of a motorcycle wheel in a turn for example.

If $\mathrm{d}t$ is the time required for this movement, the curvilinear velocity of the point is therefore:

$$v=\frac{\mathrm{d}l}{\mathrm{d}t}=R\frac{\mathrm{d}\theta}{\mathrm{d}t}=\omega R:=v_{\text{rot}} \tag{9.3.280}$$

then we fall back on the result already given previously.

Now let us do the same that during our study of the rectilinear uniform movement and let us determine the angular position over time. Then we have:

$$\omega=\frac{\mathrm{d}\theta}{\mathrm{d}t}\Rightarrow\mathrm{d}\theta=\omega\mathrm{d}t \tag{9.3.281}$$

Therefore:

$$\int\limits_{\theta_0}^{\theta}\mathrm{d}\theta=\int\limits_{\theta_0}^{t}\omega\mathrm{d}t\Rightarrow\theta-\theta_0=\omega t-\omega t_0 \tag{9.3.282}$$

and if $t_0=0$, then we have:

$$\boxed{\theta=\theta_0+\omega t} \tag{9.3.283}$$

to be compared with the equivalent relation between position and speed obtained during our study of rectilinear kinematics.

Now let us consider the definition of "angular acceleration" (the traditional notation is a bit unfortunate ...):

$$\alpha=\dot{\omega}=\frac{\mathrm{d}\omega}{\mathrm{d}t}=\frac{\mathrm{d}^2\theta}{\mathrm{d}^2}=\ddot{\theta} \tag{9.3.284}$$

Then we have:

$$\alpha=\frac{\mathrm{d}\omega}{\mathrm{d}t}\Rightarrow\mathrm{d}\omega=\alpha\mathrm{d}t \tag{9.3.285}$$

and therefore:

$$\int\limits_{\omega_0}^{\omega}\mathrm{d}\omega=\int\limits_{t_0}^{t}\alpha\mathrm{d}t\Rightarrow\omega-\omega_0=\alpha t-\alpha t_0 \tag{9.3.286}$$

and if $t_0=0$, then we have:

$$\omega=\omega_0+\alpha t \tag{9.3.287}$$

What we can write:

$$\dot{\theta}=\frac{\mathrm{d}\theta}{\mathrm{d}t}=\omega_0+\alpha t \tag{9.3.288}$$

Therefore it comes:

$$\int_{t_0}^{t} \mathrm{d}\theta = \int_{t_0}^{t} (\omega_0 + \alpha t)\mathrm{d}t = \int_{t_0}^{t} \omega_0 \mathrm{d}t + \int_{t_0}^{t} \alpha t \mathrm{d}t \tag{9.3.289}$$

Hence:

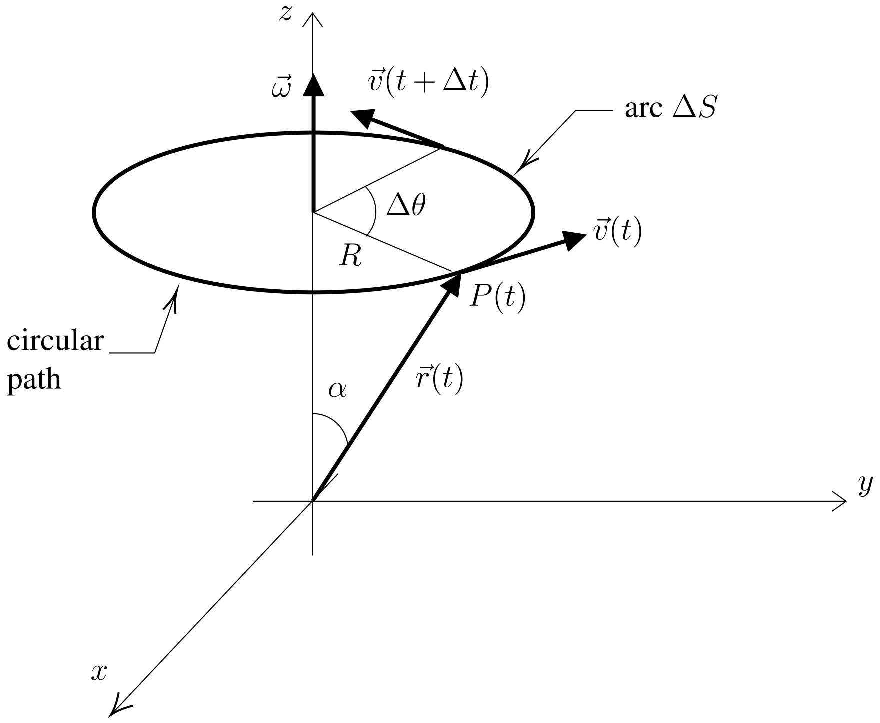
$$\theta(t) - \theta_0 = \omega_0(t - t_0) + \frac{1}{2}\alpha(t^2 - t_0^2) \tag{9.3.290}$$

and if $t_0 = 0$, then we have:

$$\boxed{\theta(t) = \theta_0 + \omega_0 t + \frac{1}{2}\alpha t^2} \tag{9.3.291}$$

compared with the equivalent relation between position, velocity and acceleration obtained in our study of the kinematics of rectilinear motions.

Let us now look to the vector of circular motion which will be extremely important one step further and also in many other chapters. So given an Euclidean orthonormal basis such as:



We see well from the above figure that:

$$\|\vec{r}(t)\| \sin(\alpha) = R \tag{9.3.292}$$

So finally we have:

$$v = \omega R = \omega \|\vec{r}(t)\| \sin(\alpha) = \|\vec{\omega}\| \cdot \|\vec{r}(t)\| \sin(\alpha) \tag{9.3.293}$$

We see then that we are dealing with a cross product (see section Vector Calculus page 1830) such that:

$$\vec{v} = \vec{\omega} \times \vec{r}(t) \tag{9.3.294}$$

So we have:

$$\vec{a}=\frac{\mathrm{d}\vec{v}}{\mathrm{d}t}=\frac{\mathrm{d}(\vec{\omega}\times\vec{r})}{\mathrm{d}t}=\frac{\mathrm{d}\vec{\omega}}{\mathrm{d}t}\times\vec{r}+\vec{\omega}\times\frac{\mathrm{d}\vec{r}}{\mathrm{d}t} \tag{9.3.295}$$

that we also write:

$$\boxed{\vec{a}=\frac{\mathrm{d}\vec{\omega}}{\mathrm{d}t}\times\vec{r}+\vec{\omega}\times\frac{\mathrm{d}\vec{r}}{\mathrm{d}t}=\vec{a}_{\text{tg}}+\vec{a}_n} \tag{9.3.296}$$

The acceleration of the circular motion is therefore is formed in the general case, of two terms, the first being the "tangential acceleration" always expressing the variation of the speed along the path and the second acceleration perpendicular along the named "normal acceleration" or "centripetal acceleration" (centripetal meaning: "that tends to bring the center"). This is why the normal acceleration is also sometimes denoted by $\vec{a}_c$.

Remark

Following Newton's second law we can associate a force to each of this both acceleration. However there must be no confusion between: "centrifugal force" (Latin for "center fleeing") that describes the tendency of an object following a curved path to fly outwards, away from the center of the curve. It's not really a force as we have already mentioned it plenty of times; it results from inertia — the tendency of an object to resist any change in its state of rest or motion, and the "centripetal force" that is a real force that counteracts the centrifugal force and prevents the object from "flying out," keeping it moving instead with a uniform speed along a circular path.

✍**Examples:**

E1. The centrifugal "force" on Earth's equator is equal to:

$$F_c=m\frac{v^2}{R}=m\frac{(\omega R)^2}{R}=m\omega^2 R=m\cdot(7.292115\cdot 10^{-5})\cdot 6,378,137\cong m\cdot 0.034 \tag{9.3.297}$$

and obviously the tangential force is zero as the tangential acceleration is zero since $\mathrm{d}\omega/\mathrm{d}t=0$! For information the $0.034\,[\mathrm{m}\cdot\mathrm{s}^{-2}]$ is equal to $440.64\,[\mathrm{km}\cdot\mathrm{h}^{-2}]$ (there is a conversion factor of $3,600^2/1,000=12,600$).

To compare at the same place to Earth's gravitational acceleration that is of $a_g\cong 9.800\,[\mathrm{m}\cdot\mathrm{s}^{-2}]$. The resultant is then:

$$a_R=a_g-a_c\cong 9.8-0.034=9.766\,[\mathrm{m}\cdot\mathrm{s}^{-2}] \tag{9.3.298}$$

Interesting result for flat earthers who "believe" that if the Earth was really rotating we should be put in orbit because of the centripetal force....

This is why scientific illiteracy leads ignorant people to not be able to explain the following meme (that is misleading by the way because the face deformation is due to the air speed and not to the the centripetal force) using college level physics:

10 km/h
IG: space.check
100 km/h
1675 km/h

when instead the correct scientific corresponding photomontage is:

20 RPM
60 RPM
0.00069 RPM

E2. We have the angular velocity ω and the orbital radius R for the Earth orbit, we can calculate its centripetal acceleration a using the relation derived earlier above:

$$a_n = \omega^2 r = \left(1.991 \cdot 10^{-7}\right)^2 \cdot \left(1.5 \cdot 10^{11}\right) \cong \left(3.9686321 \cdot 10^{-14}\right) \cdot \left(1.5 \cdot 10^{11}\right) \cong 5.9529 \cdot 10^{-3}\,[\mathrm{m} \cdot \mathrm{s}^{-2}] \tag{9.3.299}$$

For information the $5.9529 \cdot 10^{-3}\,[\mathrm{m} \cdot \mathrm{s}^{-2}]$ is equal to $77.15\,[\mathrm{km} \cdot \mathrm{h}^{-2}]$.

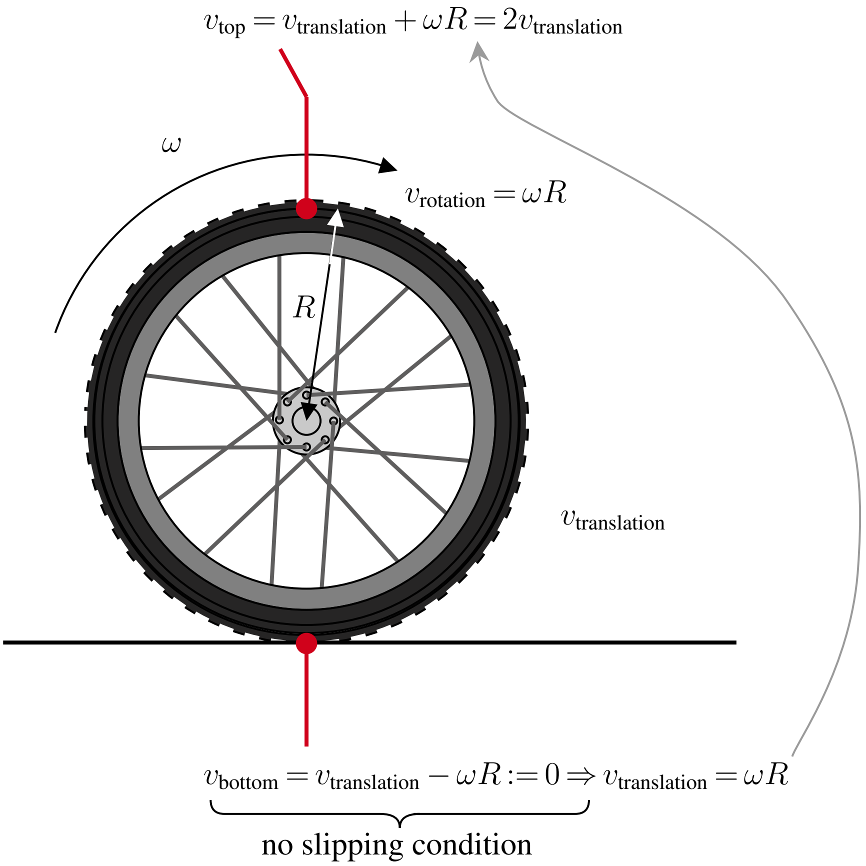
Again the tangential acceleration of the Earth assuming it's on a circular orbit is equal to zero $a_{\text{tg}} = \mathrm{d}\omega/\mathrm{d}t = 0$.

Some sciences prove that everything you know is wrong. Physics doesn't. It just proves that everything you know is relative. A good example of this is the motion of the wheel. Although it was one of humankind's earliest machines, it still has some tricks up its sleeve.

When we measure speed, we generally measure the distance an object goes over time. For an overall car, that's a fine measurement. But the wheels are in different kinds of motion. When a wheel skids, it's moving over the road without any traction. The rubber of the wheel is sliding with respect to the surface of the road. Tires don't usually, or optimally, skid. They aren't usually moving over the ground – at least the part that touches the ground isn't. It's gripping, and staying still.

The car is moving forward at a certain speed, $v_{\text{translational}}$, with respect to the ground. The bottom of the wheel is assumed motionless in respect to the ground. In order for the bottom of the wheel to have a speed of zero when the car is racing forward, it has to be going backwards just as much as the car is going forwards. It's like someone on a moving sidewalk walking backwards as fast as the sidewalk is pushing them forwards, so they don't appear to move when observed from someone off the sidewalk.

Here's where things get weird. The exact center of the wheel moves at the same speed of the car, of course. So looked at from the road, it is moving at speed $v_{\text{translational}}$. But the top of the wheel is moving with the speed of the car, and it is also moving forward at exactly the speed that the bottom of the wheel is moving backward. Since the bottom of the wheel is spinning backwards at negative $v_{\text{translational}}$, so the top of the wheel is spinning forwards at speed positive v_{top}, and also moving with the car at speed positive $v_{\text{translational}}$, making for a combined speed of $2v_{\text{translational}}$.


$v_{\text{top}} = v_{\text{translation}} + \omega R = 2v_{\text{translation}}$
ω
$v_{\text{rotation}} = \omega R$
R
$v_{\text{translation}}$
$\underbrace{v_{\text{bottom}} = v_{\text{translation}} - \omega R := 0}_{\text{no slipping condition}} \Rightarrow v_{\text{translation}} = \omega R$

And that is how a single wheel, all in one piece and attached to a car going at a steady pace can have a measured speed of 0, $v_{\text{translational}}$, and $2v_{\text{translational}}$, all at the same time.

The relation between the velocity at the top, center and bottom of the wheel is geometrical and is not affected by friction per-se. If the car wheel spins with angular frequency ω, has a radius R and velocity at the axle of $v_{\text{translational}}$ then the velocity of the wheel at the top and bottom is:

$$v_{\text{top}} = v_{\text{translational}} + \omega R, v_{\text{bottom}} = v_{\text{translational}} - \omega R \tag{9.3.300}$$

If the no slip condition holds then $v_{\text{bottom}} = 0$, so $v_{\text{translational}} = \omega R$. Using this in the top equation gives $v_{\text{top}} = v_{\text{translational}} + v_{\text{translational}} = 2v_{\text{translational}}$, independent of R! This is because both v and v_{top} increase in the same proportion as R increases.

The reader interested to vector expressions of the speed and acceleration in cartesian, polar, cylindrical and spherical coordinates should refer to the section of Vector Calculus (see page 1846) where all the details are given where we get for recall:

- In cartesian coordinates:
  - Velocity:

$$\vec{v} = \left(\dot{x}(t), \dot{y}(t), \dot{z}(t),\right) \tag{9.3.301}$$

  - Acceleration:

$$\vec{a} = \left(\ddot{x}(t), \ddot{y}(t), \ddot{z}(t), t\right) \tag{9.3.302}$$

- In polar coordinates:
  - Velocity:

$$\vec{v}=\dot{r}\vec{e}_r+\vec{e}_\phi r\dot{\phi} \tag{9.3.303}$$

  - Acceleration:

$$\vec{a}=\vec{e}_r(\ddot{r}-\dot{\phi}^2 r)+\vec{e}_\phi(2\dot{r}\dot{\phi}+r\ddot{\phi}) \tag{9.3.304}$$

- In cylindrical coordinates:
  - Velocity:

$$\vec{v}=\dot{r}\vec{e}_r+\vec{e}_\phi r\dot{\phi}+\vec{e}_z\dot{z} \tag{9.3.305}$$

  - Acceleration:

$$\vec{a}=\vec{e}_r\left(\ddot{r}-r\dot{\phi}^2\right)+\vec{e}_\phi\left(r\ddot{\phi}+2\dot{r}\dot{\phi}\right)+\vec{e}_z\ddot{z} \tag{9.3.306}$$

- In spherical coordinates:
  - Velocity:

$$\vec{v}=\dot{r}\vec{e}_r+\vec{e}_\theta r\dot{\theta}+\vec{e}_\phi r\dot{\phi}\sin(\theta) \tag{9.3.307}$$

  - Acceleration:

$$\begin{aligned}\vec{a}=&\vec{e}_r(\ddot{r}-r\dot{\theta}^2-r\dot{\phi}^2\sin^2(\theta))+\vec{e}_\theta(r\ddot{\theta}+2\dot{r}\dot{\theta}-r\dot{\phi}^2\sin\theta\cos(\theta))\\&+\vec{e}_\phi(r\ddot{\phi}\sin(\theta)+2r\dot{\theta}\dot{\phi}\cos(\theta)+2\dot{r}\dot{\phi}\sin(\theta))\end{aligned} \tag{9.3.308}$$

For the details of the analysis of the velocity and acceleration of a elliptical motion the reader should refer to the section of Astronomy.

Also from the vis-viva equation proved in the section of Aerospace Engineering (see page 8619) we get the speed of a body in any point of an ellipse:

$$v=\sqrt{GM\left(\frac{2}{r(\theta)}-\frac{1}{a}\right)} \tag{9.3.309}$$

where as proved in the section of Analytical Geometry (see page 2785):

$$r(\theta)=\frac{a(1-e^2)}{1+e\cos(\theta)} \tag{9.3.310}$$

where a is the semi-major axis of the ellipse.

For non-planar and not rectilinear motions (that is to say: the generalization of what we have seen so far). The reader should refer to the section of Differential Geometry (see page 2850).

If we express the circular movement of a point P from a system of axes lying in the plane of the trajectory, for simplicity, then the position of the point P is given parametrically by (see section Vector Calculus page 1864):

$$\vec{r}(t)=r\cos(\omega t)\vec{i}+r\sin(\omega t)\vec{j}=r\sin\left(\omega t+\frac{\pi}{2}\right)\vec{i}+r\sin(\omega t)\vec{j} \tag{9.3.311}$$

This shows that the circular movement can be considered as the superposition of two sinusoidal movements phase-shifted of $\pi/2$. But if we write now more generally we could also have:

$$\vec{r}(t)=r_1\sin\left(\omega t+\frac{\pi}{2}\right)\vec{i}+r_2\sin(\omega t)\vec{j} \tag{9.3.312}$$

Or even more general:

$$\vec{r}(t)=r_1\sin\left(\omega t+\varphi\right)\vec{i}+r_2\sin(\omega t)\vec{j} \tag{9.3.313}$$

or more general again:

$$\vec{r}(t)=r_1\sin\left(\omega_x t+\varphi\right)\vec{i}+r_2\sin(\omega_y t)\vec{j} \tag{9.3.314}$$

We get by focusing on the coordinates:

$$\begin{aligned}x&=A\sin(\omega_x t)\\ y&=B\sin(\omega_y t+\varphi)\end{aligned} \tag{9.3.315}$$

by varying the phase difference φ and the ratio ω_x/ω_y we get curves that we name "Lissajous figures":

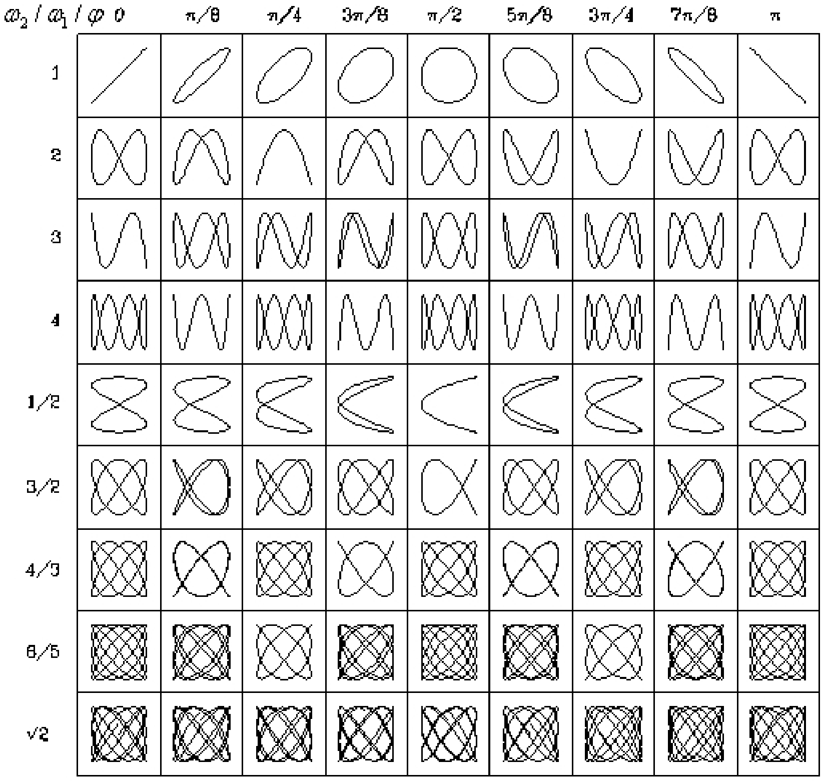


Figure 9.44: Lissajous figures

A much more beautiful summary is given by the following illustration:

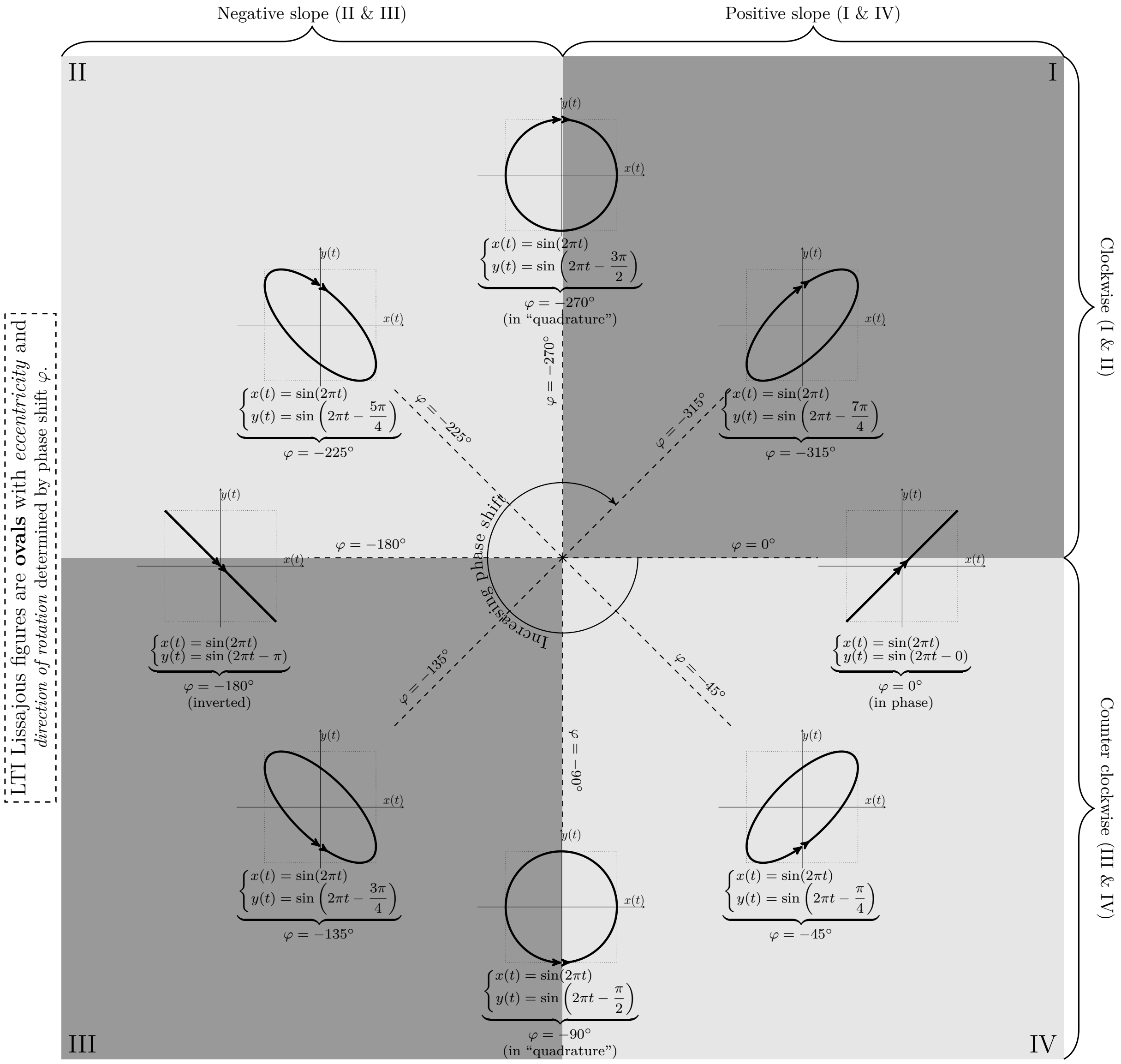

Negative slope (II & III)
Positive slope (I & IV)
II
I
III
IV
Clockwise (I & II)
Counter clockwise (III & IV)
LTI Lissajous figures are **ovals** with *eccentricity* and *direction of rotation* determined by phase shift φ.
Increasing phase shift
$y(t)$
$x(t)$
$x(t)=\sin(2\pi t)$, $y(t)=\sin\left(2\pi t-\frac{3\pi}{2}\right)$
$\varphi=-270°$ (in "quadrature")
$x(t)=\sin(2\pi t)$, $y(t)=\sin\left(2\pi t-\frac{5\pi}{4}\right)$
$\varphi=-225°$
$x(t)=\sin(2\pi t)$, $y(t)=\sin\left(2\pi t-\frac{7\pi}{4}\right)$
$\varphi=-315°$
$x(t)=\sin(2\pi t)$, $y(t)=\sin(2\pi t-\pi)$
$\varphi=-180°$ (inverted)
$x(t)=\sin(2\pi t)$, $y(t)=\sin(2\pi t-0)$
$\varphi=0°$ (in phase)
$x(t)=\sin(2\pi t)$, $y(t)=\sin\left(2\pi t-\frac{3\pi}{4}\right)$
$\varphi=-135°$
$x(t)=\sin(2\pi t)$, $y(t)=\sin\left(2\pi t-\frac{\pi}{4}\right)$
$\varphi=-45°$
$x(t)=\sin(2\pi t)$, $y(t)=\sin\left(2\pi t-\frac{\pi}{2}\right)$
$\varphi=-90°$ (in "quadrature")
$\varphi=-270°$
$\varphi=-225°$
$\varphi=-315°$
$\varphi=-180°$
$\varphi=0°$
$\varphi=-135°$
$\varphi=-45°$
$\varphi=-90°$

Figure 9.45: Several Lissajous figures for different phase delays (author: Krishnatej Vedala, source: Wikipedia)

An important case to consider for the study of circular particles accelerators is the acceleration vector of a uniform circular movement. For such a movement, the position vector in time t is obviously given by:

$$\vec{r}=R\sin(\omega t)\vec{e}_x+R\cos(\omega t)\vec{e}_y \tag{9.3.316}$$

where R is the radius, ω the (constant) angular velocity and $\vec{e}_x$,$\vec{e}_y$ the orthonormal unit vectors for x and y.

The acceleration vector being given by:

$$\vec{a}=\frac{\mathrm{d}^2\vec{r}}{\mathrm{d}t^2} \tag{9.3.317}$$

Carrying out the two derivations we find:

$$\vec{a}=-\omega^2\left[R\sin(\omega t)\vec{e}_x+R\cos(\omega t)\vec{e}_y\right] \tag{9.3.318}$$

Therefore we get for the "centrifugal acceleration":

$$\boxed{\vec{a} = -\omega^2 \vec{r}} \tag{9.3.319}$$

So in uniform circular motion the only acceleration vector always points opposite to the position vector, to the center of the trajectory.

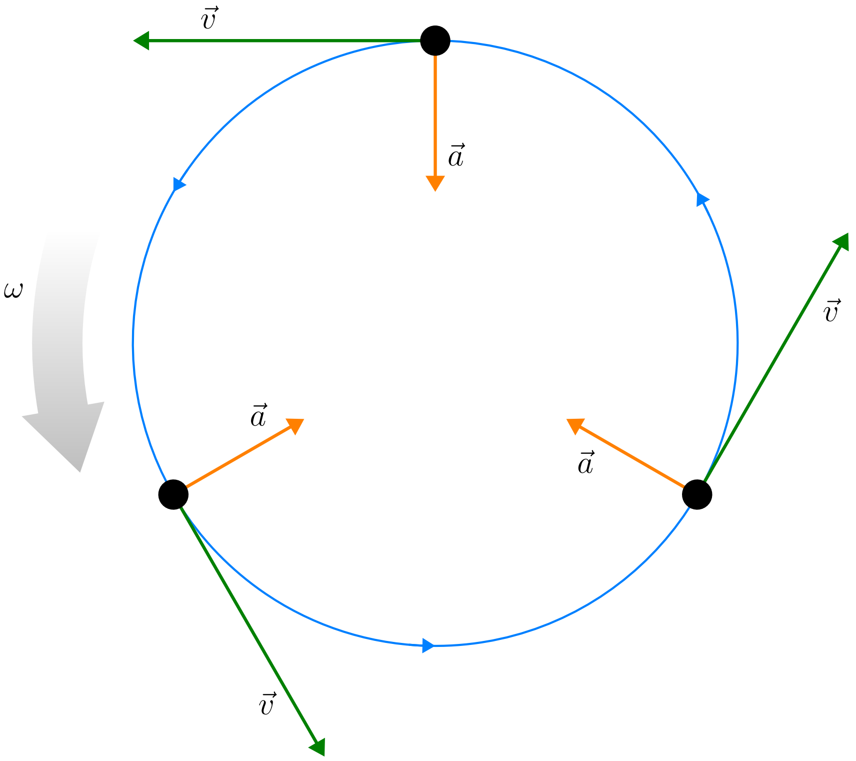
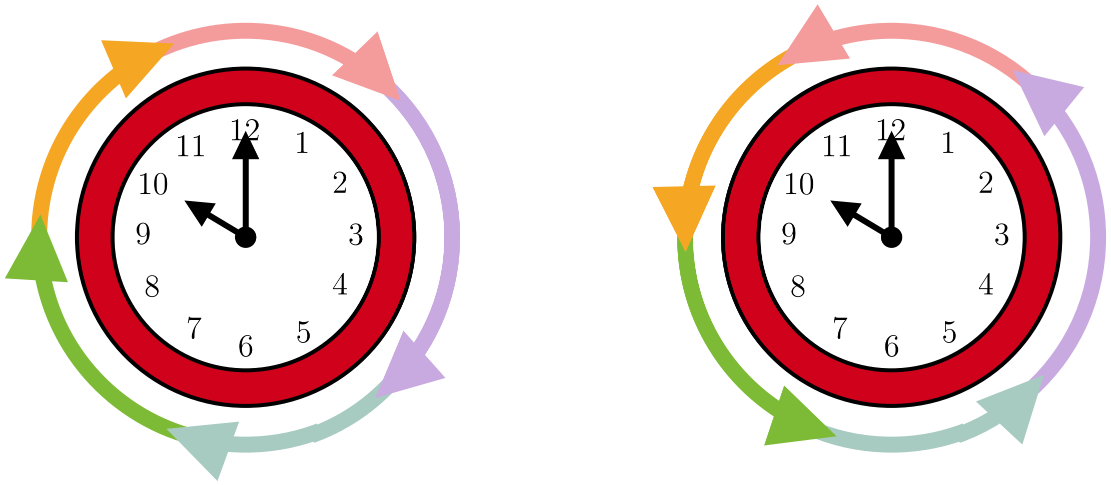


Figure 9.46: Centrifugal acceleration

The reader will find very important practical applications of the kinematics of the circular motion to for the industry relatively to mechanics in the section of Mechanical Engineering (see chapter Engineering page 8359) and also to Astronomy (see chapter Cosmology page 5212) and finally to Corpuscular Quantum Physics (see chapter Atomistic page 4616).



These arrows show
clockwise direction

These arrows show
anticlockwise direction

Clockwise is the same direction the
hands of a clock move

(for an observer inside the watch the hands go from left to right)

(for an observer outside the watch the hands go from right to left)

Anticlockwise (or "counter-clockwise"/"trigonometric")
is the opposite direction to the hands of a clock move

(for an observer inside the watch the hands go from right to left)

(for an observer outside the watch the hands go from left to right)

The rotation is observer dependant!
(that's why the stars appear to rotate anticlockwise in northern hemisphere
and clockwise in the southern hemisphere)

Figure 9.47: **Reminder of typical rotation descriptions**

9.3.8 Energy, Work and Power

When a force acts upon an object to cause a displacement of the object, it is said that work was done upon the object. There are three key ingredients to work - force, displacement, and cause. In order for a force to qualify as having done work on an object, there must be a displacement and the force must cause the displacement. There are several good examples of work that can be observed in everyday life - a horse pulling a plow through the field, a father pushing a grocery cart down the aisle of a grocery store, a freshman lifting a backpack full of books upon her shoulder, a weightlifter lifting a barbell above his head, an Olympian launching the shot-put, etc. In each case described here there is a force exerted upon an object to cause that object to be displaced.

If a mass point m undergoes an elementary displacement $\mathrm{d}\vec{r}$ under the effect of a force $\vec{F}$, this force performs an elementary work W given by definition by the "work equation":

$$\boxed{\mathrm{d}W=\vec{F}\circ\mathrm{d}\vec{r}=\|\vec{F}\|\cdot||\mathrm{d}\vec{r}\|\cos(\alpha)} \qquad (9.3.320)$$

However the reader must be careful. It's now why a body accelerates (begin to move) that there is necessarily a force!!! Indeed, Einstein has proved that what we have consider for centuries as the "gravitational force" was in fact only a body following the space-time curvature. So never go too quick in conclusion when making science...

On occasion, a force acts upon a moving object to hinder a displacement. Examples might include a car skidding to a stop on a roadway surface or a baseball runner sliding to a stop on the infield dirt. In such instances, the force acts in the direction opposite the objects motion in order to slow it down. The force doesn't cause the displacement but rather hinders it. The negative work refers obviously to the numerical value that results when values of F, d and θ are substituted into the work equation. Since the force vector is directly opposite the displacement vector, theta is π degrees. The $\cos(\pi)$ is -1 and so a negative value results for the amount of work done upon the object.

Definition (#9.3.441): If the work W is positive the work is named "engine work". If negative we speak then of "resistive work".

If the mass m moves from a point A to place B, the total work is:

$$\Delta W=W_{A,B}=\int\limits_A^B\vec{F}\circ\mathrm{d}\vec{r}=\int\limits_A^B\|\vec{F}\|\cos(\alpha)||\mathrm{d}\vec{r}\| \qquad (9.3.321)$$

For the units, we have the option of expressing them in "Joules [J]" or in Newton-meters:

$$[\mathrm{N}\cdot\mathrm{m}]=[\mathrm{J}] \qquad (9.3.322)$$

From the definition above we can also quickly deduce the work of a moment of force (or in other words, the work of a force in a rotational movement) because of an infinitesimal element of movement, we have:

$$\mathrm{d}W=(\vec{F}\times\vec{r})\circ\mathrm{d}\vec{\theta}=(\vec{F}\times\vec{r})\circ\vec{n}\mathrm{d}\theta=\vec{M}\circ\vec{n}\mathrm{d}\theta \qquad (9.3.323)$$

So in the case of circular motion the work of a moment of force will be given by:

$$\boxed{\Delta W = W_{A,B} = \int\limits_{\theta_A}^{\theta_B} = \vec{M} \circ \vec{n} \mathrm{d}\theta} \tag{9.3.324}$$

9.3.8.1 Conservative vector field

If the force $\vec{F}$ is constant in magnitude and in direction (gravity near the Earth's surface for example), the integral calculation of W takes a simpler form:

$$\Delta W = \int\limits_{A}^{B} \vec{F} \circ \mathrm{d}\vec{r} = \vec{F} \circ \int\limits_{A}^{B} \mathrm{d}\vec{r} = \vec{F} \circ (\vec{r}_B - \vec{r}_A) \tag{9.3.325}$$

This result shows that the work depends only the initial and final positions and not of the path travelled in this special condition.

But in fact there is a more general result that we have already proved in the section of Differential and Integral Calculus that is the "fundamental theorem for line integrals" or "gradient theorem for line integrals" that gives us:

$$\boxed{\int\limits_{\Gamma} \vec{F} \circ \mathrm{d}\vec{r} \Leftrightarrow \int\limits_{\Gamma} \vec{\nabla}(U) \circ \mathrm{d}\vec{r} : \quad \int\limits_{\Gamma} \vec{\nabla}(U) \circ \mathrm{d}\vec{r} = U(b) - U(a)} \tag{9.3.326}$$

where $\vec{\nabla}()$ is the gradient operator (see section Vector Calculus page 1869).

The force of gravity, of electrostatic force is a special case of this type but in fact in all situations as it derives from a potential (see details further below)!

A "conservative vector field" (also named a "path-independent vector field") is a vector field $\vec{F}$ whose line integral $\oint_{\Gamma} \vec{F} \circ \mathrm{d}\vec{r}$ over any curve Γ depends only on the endpoints of Γ. Therefore a vector field for which we have:

$$\oint_{\Gamma} \vec{F} \circ \mathrm{d}\vec{r} = 0 \tag{9.3.327}$$

is by definition automatically a conservative vector field and then we say that there is "no circulation around a closed curve".

We have also proved in the section of Vector Calculus the Stokes theorem (see page 1889) that gives for recall a simply connected domain, a continuously differentiable vector field $\vec{F}$:

$$\boxed{\oint \vec{F} \circ \mathrm{d}\vec{r} = \int\limits_{S} \left(\vec{\nabla} \times \vec{F} \right) \circ \mathrm{d}\vec{S}} \tag{9.3.328}$$

Therefore for a conservative vector field:

$$\oint_{\Gamma} \vec{F} \circ \mathrm{d}\vec{r} = \int\limits_{S} \left(\vec{\nabla} \times \vec{F} \right) \circ \mathrm{d}\vec{S} = 0 \Leftrightarrow \vec{\nabla} \times \vec{F} = \vec{0} \tag{9.3.329}$$

So a vector field that is continuously differentiable on a simply connected domain such that:

$$\vec{\nabla}\times\vec{F}=\vec{0} \tag{9.3.330}$$

is also a possible criteria of identification of a conservative vector field.

✍**Example:**

Our goal is to determine if the vector field:

$$\vec{F}=\begin{pmatrix}-\dfrac{y}{x^2+y^2}\\ \dfrac{x}{x^2+y^2}\\ 0\end{pmatrix} \tag{9.3.331}$$

is conservative or not.

One condition for path independence is as we know the following: for a simply connected domain, a continuously differentiable vector field $\vec{f}$ is path-independent if and only if its curl is zero.

Then we have:

$$\vec{\nabla}\times\vec{F}=\begin{pmatrix}\dfrac{\partial}{\partial x}\\ \dfrac{\partial}{\partial y}\\ \dfrac{\partial}{\partial z}\end{pmatrix}\begin{pmatrix}-\dfrac{y}{x^2+y^2}\\ \dfrac{x}{x^2+y^2}\\ 0\end{pmatrix}=\begin{pmatrix}0\\ 0\\ \dfrac{\partial}{\partial x}\left(\dfrac{x}{x^2+y^2}\right)-\dfrac{\partial}{\partial y}\left(-\dfrac{y}{x^2+y^2}\right)\end{pmatrix} \tag{9.3.332}$$

We calculate:

$$\begin{aligned}\frac{\partial F_2}{\partial x}&=\frac{1}{x^2+y^2}-\frac{x(2x)}{(x^2+y^2)^2}=\frac{y^2-x^2}{(x^2+y^2)^2}\\ \frac{\partial F_1}{\partial y}&=-\frac{1}{x^2+y^2}+\frac{y(2y)}{(x^2+y^2)^2}=\frac{y^2-x^2}{(x^2+y^2)^2}\end{aligned} \tag{9.3.333}$$

Since these partial derivatives are equal, the curl is zero.

Can we conclude $\vec{F}$ is conservative? The problem is that $\vec{F}$ is not defined at the origin $(0,0)$. Its domain of definition has a hole in it, which for two-dimensional regions, is enough to prevent it from being simply connected. The test does not apply, and we still don't know whether or not $\vec{F}$ is conservative.

Let's try another test, this time a test for path-dependence. If we can find a closed curve along which the integral of $\vec{F}$ is non-zero, then we can conclude that $\vec{F}$ is path-dependent. If the curve does not go around the origin, then we can use Green's theorem (see section Vector Calculus page 1890) as we have only the third component of the curl that was non-null above:

$$\oint\limits_{\Gamma}(f_x\mathrm{d}x+f_y\mathrm{d}y)=\int_S\left(\frac{\partial f_y}{\partial x}-\frac{\partial f_x}{\partial y}\right)\mathrm{d}S_z \tag{9.3.334}$$

to show the integral of $\vec{F}$ is zero or not using path integral rather than directly the curl.

We choose for closed path a circle and then we can use a re-parametrization of the field using $x=\cos(t)$ and $y=\sin(t)$ with $0\leq t\leq 2\pi$. Then on the unit circle, $\vec{F}$ takes a simple form:

$$\vec{F}=\left(\frac{-\sin(t)}{\cos^2(t)+\sin^2(t)},\frac{\cos(t)}{\cos^2(t)+\sin^2(t)}\right)=(-\sin(t),\cos(t)) \tag{9.3.335}$$

and we have obviously:

$$\begin{aligned}x=\cos(t)&\Rightarrow \mathrm{d}x=-\sin(t)\mathrm{d}t\\ y=\sin(t)&\Rightarrow \mathrm{d}y-\cos(t)\mathrm{d}t\end{aligned} \tag{9.3.336}$$

Therefore:

$$\begin{aligned}\oint_\Gamma \vec{F}\circ\mathrm{d}\vec{r}&=\int_a^b\begin{pmatrix}F_1\\F_2\end{pmatrix}\circ\begin{pmatrix}\mathrm{d}x\\ \mathrm{d}y\end{pmatrix}\mathrm{d}t=\int_0^{2\pi}\begin{pmatrix}-\sin(t)\\ \cos(t)\end{pmatrix}\circ\begin{pmatrix}-\sin(t)\\ \cos(t)\end{pmatrix}\mathrm{d}t\\ &=\int_0^{2\pi}\left(\sin^2(t)+\cos^2(t)\right)\mathrm{d}t=\int_0^{2\pi}1\mathrm{d}t=2\pi\end{aligned} \tag{9.3.337}$$

The hole in the domain at the origin did end up causing trouble. We found a curve Γ where the circulation around Γ is not zero. The vector field $\vec{F}$ is therefore path-dependent.

The relation:

$$\Delta W=E_c(B)-E_c(A) \tag{9.3.338}$$

is sometimes named "theorem of kinetic energy" or "work-energy theorem".

Note that in the case of a rectilinear movement, to which a mobile initially at rest travels a distance d distance being subjected to a certain force, we have (omitting of course the change in potential energy and energy loss by friction):

$$W=Fd=\frac{1}{2}mv^2\Leftrightarrow F=\frac{mv^2}{2d} \tag{9.3.339}$$

that we name "working force".

9.3.8.2 Kinetic Energy

Newton's second law $\vec{F}=m\vec{a}$ applies along a path A-B. By using the expression of the work we have:

$$\Delta W=\int\limits_A^B \vec{F}\circ\mathrm{d}\vec{r}=\int\limits_A^B m\frac{\mathrm{d}\vec{v}}{\mathrm{d}t}\circ\mathrm{d}\vec{r}=\int\limits_A^B m\frac{\mathrm{d}\vec{v}}{\mathrm{d}t}\circ\vec{v}\mathrm{d}t=m\int\limits_A^B \vec{v}\circ\mathrm{d}\vec{v} \tag{9.3.340}$$

and by developing the dot product using the components we have:

$$\Delta W=m\int\limits_A^B (v_x\mathrm{d}v_x+v_y\mathrm{d}v_y+v_z\mathrm{d}v_z)=\frac{1}{2}m(v_x^2+v_y^2+v_z^2)\Big|_A^B=\frac{1}{2}mv_B^2-\frac{1}{2}mv_A^2 \tag{9.3.341}$$

By definition we say that each term in the last equality is the (non-relativistic) "kinetic energy" denoted by:

$$\boxed{E_{\text{cin}}=E_c=\frac{1}{2}mv^2} \tag{9.3.342}$$

and it is measured obviously also in Joules [J] (or other exotic derivate units whose theoretical physicists sometimes abuse a little bit too much...) and is always positive in mechanics or any other field of physics.

The previous development can then be rewritten as:

$$\Delta W=m\int\limits_A^B (v_x\mathrm{d}v_x+v_y\mathrm{d}v_y+v_z\mathrm{d}v_z)=\frac{1}{2}mv_B^2-\frac{1}{2}v_A^2=E_c(B)-E_c(A) \tag{9.3.343}$$

Therefore when a body moves (without rotation on itself) under the action of any force $\vec{F}$, the work of this acceleration force on any path A, B is equal to the change in kinetic energy of the body.

9.3.8.2.1 Moment of inertia

The "moment of inertia", otherwise known as the "angular mass" or "rotational inertia", of a rigid body determines the torque needed for a desired angular acceleration about a rotational axis. It depends on the body's mass distribution and the axis chosen, with larger moments requiring more torque to change the body's rotation.

For bodies constrained to rotate in a plane, we will see that is sufficient to consider their moment of inertia about an axis perpendicular to the plane. For bodies free to rotate in three dimensions, their moments must be described by a symmetric 3×3 matrix; each body has a set of mutually perpendicular principal axes for which this matrix is diagonal and torques around the axes act independently of each other.

To a rigid solid rotating about an axis at the angular velocity $\vec{\omega}$, the elementary kinetic energy of any point of mass $\mathrm{d}m$, located off-axis, is:

$$\mathrm{d}E_c=\frac{1}{2}v^2\mathrm{d}m=\frac{1}{2}\mathrm{d}m=\frac{1}{2}\left(\vec{\omega}\times\vec{r}\right)^2\mathrm{d}m=\frac{1}{2}(\omega r\sin(\alpha))^2\mathrm{d}m=\frac{1}{2}(\omega r)^2\mathrm{d}m \tag{9.3.344}$$

since $\vec{\omega}$ and $\vec{r}$ are perpendicular. The total kinetic energy is then:

$$E_c = \int \mathrm{d}E_c = \frac{1}{2}\omega^2 \int r^2 \mathrm{d}m \tag{9.3.345}$$

We have become accustomed in physics to write this last relation:

$$\boxed{E_c = \frac{1}{2}\omega^2 \int r^2 \mathrm{d}m = \frac{1}{2}\omega^2 J} \tag{9.3.346}$$

with the "moment of inertia" defined by:

$$J = \int r^2 \mathrm{d}m \tag{9.3.347}$$

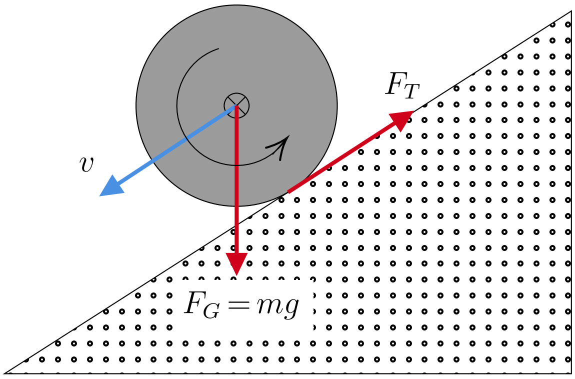
Examples:

E1. Let us calculate the final velocity of a rigid ball falling on an inclined rigid plane with non-zero friction (therefore the ball is going to rotate) in a gravitational potential field where the friction heat can be neglected and in vacuum.

The answer of the beginner will often be obtained by considering only the kinetic energy but not the rotating speed of the ball. Now we must take this into account through its moment of inertia.

So we have the total kinetic energy that the translation kinetic energy of the center of mass plus the rotational energy around the same center of mass:

$$E_c = E_{\text{transl.}} + E_{\text{rot.}} = \frac{1}{2}mv_G^2 + \frac{1}{2}J_G\omega^2 = \frac{1}{2}v_G^2\left(m + \frac{J}{r^2}\right) \tag{9.3.348}$$



equating that value with that of the gravitational potential energy (see further below), assuming a zero initial velocity of fall, we have:

$$\Delta E_c = \Delta E_p \Leftrightarrow mgh = \frac{1}{2}v_G^2\left(m + \frac{J_G}{r^2}\right) \tag{9.3.349}$$

Thus we get the velocity acquired at the bottom of the plane (rolling friction not included...):

$$v_G=\sqrt{\frac{2gh}{1+\dfrac{J_G}{mr^2}}} \tag{9.3.350}$$

and we have proved in the section of Geometric Shapes (see page 2886) that the moment of inertia of a solid ball around its center of mass is:

$$J_G=\frac{2}{5}mr^2 \tag{9.3.351}$$

Therefore it comes:

$$v_G=\sqrt{\frac{2gh}{1+\dfrac{2}{5}}}=\sqrt{\frac{5}{7}2gh}=\sqrt{\frac{10}{7}gh} \tag{9.3.352}$$

We see in the special case of the ball, that final speed of drop is (without air friction and without energy loss due to friction) independent of its mass and radius and of then angle of the plane (whether the ball is empty or full) which is relatively counter-intuitive.

E2. A second famous example is the calculation of the rotational kinetic energy of a perfectly spherical planet of homogeneous mass and constant rotational period. Then we have:

$$\begin{aligned}E_c&=\frac{1}{2}\omega^2J=\frac{1}{2}\omega^2\left(\frac{2}{5}MR^2\right)=\frac{2}{10}\omega^2MR^2=\frac{1}{5}\left(\frac{v}{R}\right)^2MR^2=\frac{1}{5}v^2M\\&=\frac{1}{5}v^2(\rho V)=\frac{1}{5}v^2\left(\rho\frac{4}{3}\pi R^3\right)=\frac{4\pi}{15}\rho R^3v^2=\frac{4\pi}{15}\rho R^2\left(\frac{2\pi R}{T}\right)^2=\frac{16\pi^3}{15}\rho\frac{R^5}{T^2}\end{aligned} \tag{9.3.353}$$

In a solid, the distribution of the material about an axis will obviously be different depending on the chosen axis. The moment of inertia corresponding will also be obviously different. It is therefore essential to specify the axis with respect to which we wish to determine the moment of inertia. We see that in practice that engineers often place the axis so that it passes through the center of mass. In the tables, we frequently find expressions of inertia moments of common forms (along a given axis) such as cylinder, cone, sphere, bar, tube.

Here is a summary of some moment of inertia of plane figures among all those proved in detail in the section of Geometric Shapes (see page 2886):

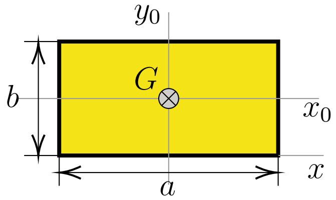
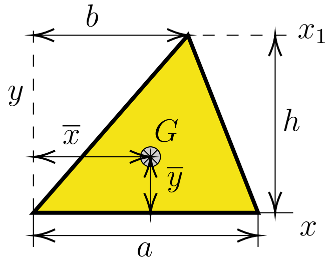
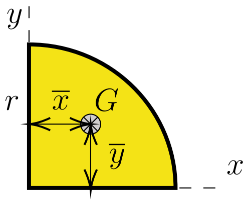
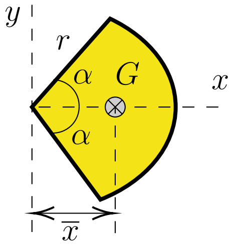
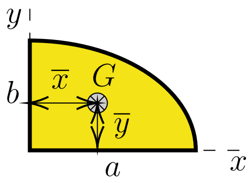
Shape	Figure	Area moment of inertia
Rectangular area		$I_x = \frac{ab^3}{3}$ $I_g = \frac{ab^3}{12}$ $J = \frac{ab}{12}(a^2+b^2)$
Triangular area	 b x_1 y h $\overline{x}$ G $\overline{y}$ x a	$I_x = \frac{ah^3}{12}$ $I_g = \frac{ah^3}{36}$ $I_g = \frac{ah^3}{4}$
Quarter circular area	 y r $\overline{x}$ G $\overline{y}$ x	$I_x = I_y = \frac{\pi r^4}{16}$ $J = \frac{\pi r^4}{8}$
Area of circular sector	 y r α G x α $\overline{x}$	$I_x = \frac{r^4}{4}\left(\alpha - \frac{1}{2}\sin(2\alpha)\right)$ $I_y = \frac{r^4}{4}\left(\alpha + \frac{1}{2}\sin(2\alpha)\right)$ $J = \frac{1}{2}r^4\alpha$
Area of elliptical quadrant	 y $\overline{x}$ G b $\overline{y}$ x a	$I_x = \frac{\pi ab^3}{16}$ $I_y = \frac{\pi a^3 b}{16}$ $J = \frac{\pi ab}{16}(a^2+b^2)$

Figure 9.48: **Summary Moment of inertia of some planar shapes**

and same for some body (volumetric) shapes (proofs are also available in the same section of Geometric Shapes at page 2886):

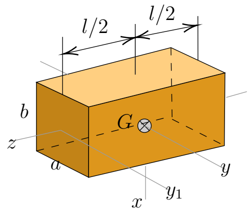
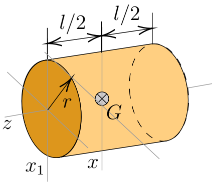
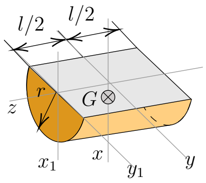
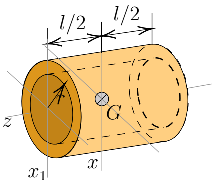
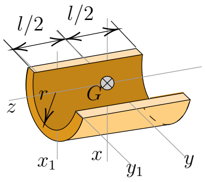
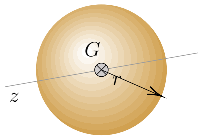
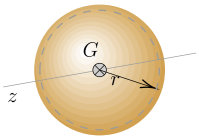
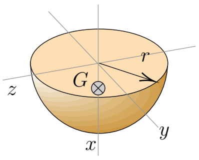
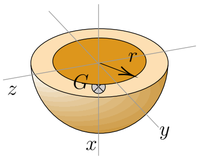
Body	Figure	Moments of inertia
Rectangular Parallelepiped		$I_{xx}=\frac{1}{12}m\left(a^2+l^2\right)$ $I_{yy}=\frac{1}{12}m\left(b^2+l^2\right)$ $I_{zz}=\frac{1}{12}m\left(a^2+b^2\right)$ $I_{y_1y_1}=\frac{1}{12}mb^2+\frac{1}{3}ml^2$
Circular Cylinder		$I_{xx}=\frac{1}{4}mr^2+\frac{1}{12}ml^2$ $I_{x_1x_1}=\frac{1}{4}mr^2+\frac{1}{3}ml^2$ $I_{zz}=\frac{1}{2}mr^2$
Semicylinder		$I_{xx}=I_{yy}=\frac{1}{4}mr^2+\frac{1}{12}ml^2$ $I_{x_1x_1}=I_{y_1y_1}=\frac{1}{4}mr^2+\frac{1}{3}ml^2$ $I_{zz}=\frac{1}{2}mr^2$
Circular Cylindrical shell		$I_{xx}=\frac{1}{2}mr^2+\frac{1}{12}ml^2$ $I_{x_1x_1}=\frac{1}{2}mr^2+\frac{1}{3}ml^2$ $I_{zz}=mr^2$
Half Cylindrical shell		$I_{xx}=I_{yy}=\frac{1}{2}mr^2+\frac{1}{12}ml^2$ $I_{x_1x_1}=I_{y_1y_1}=\frac{1}{2}mr^2+\frac{1}{3}ml^2$ $I_{zz}=mr^2$
Sphere		$I_{zz}=\frac{2}{5}mr^2$
Spherical shell		$I_{zz}=\frac{2}{3}mr^2$
Hemisphere		$I_{xx}=I_{yy}=I_{zz}=\frac{2}{5}mr^2$
Hemispherical Shell		$I_{xx}=I_{yy}=I_{zz}=\frac{2}{3}mr^2$

Figure 9.49: Summary of very common bodies moments of inertia

We saw in our study of angular momentum that:

$$\vec{b}=mr^2\vec{\omega} \tag{9.3.354}$$

and just before that the moment of inertia is given by:

$$J=\int r^2\mathrm{d}m \tag{9.3.355}$$

So we have:

$$\mathrm{d}\vec{b}=r^2\vec{\omega}\mathrm{d}m \tag{9.3.356}$$

hence:

$$\vec{b}=\vec{\omega}\int r^2\mathrm{d}m \tag{9.3.357}$$

We finally get:

$$\boxed{\vec{b}=J\vec{\omega}} \tag{9.3.358}$$

That it is the expression of the angular momentum of a rigid body rotating on itself (on one of its possible axes of rotation) commonly named "spin angular moment".

As we showed during our study of the angular momentum:

$$\frac{\mathrm{d}\vec{b}}{\mathrm{d}t}=\vec{r}\times\vec{F}=\vec{M} \tag{9.3.359}$$

then it comes under the assumption that the mass and geometry of the solid remain constant ... that the moment of force is then:

$$\boxed{\vec{M}=\frac{\mathrm{d}}{\mathrm{d}t}(J\vec{\omega})=J\frac{\mathrm{d}\vec{\omega}}{\mathrm{d}t}} \tag{9.3.360}$$

And obviously, if we study a system in which the angular momentum is conservative, it is obvious that:

$$\boxed{\vec{M}=\vec{0}\Rightarrow\frac{\mathrm{d}\vec{b}}{\mathrm{d}t}=\vec{0}\Rightarrow\vec{b}=c^{te}=J\vec{\omega}} \tag{9.3.361}$$

This conservation of angular momentum is applicable in a multitude of experiences as the well known which is to turn on a chair and spread out hands or legs which will decrease the speed (and vice versa) as illustrated in figure below:

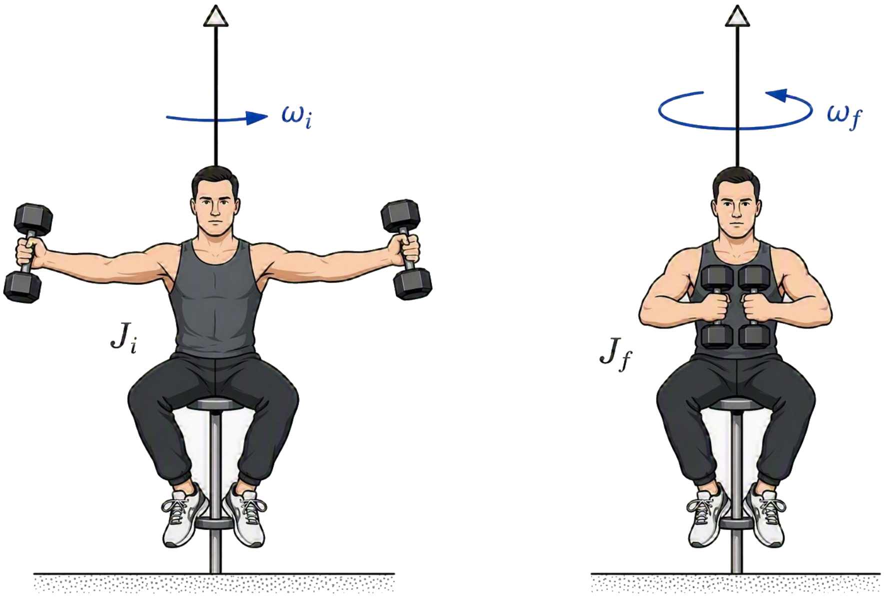
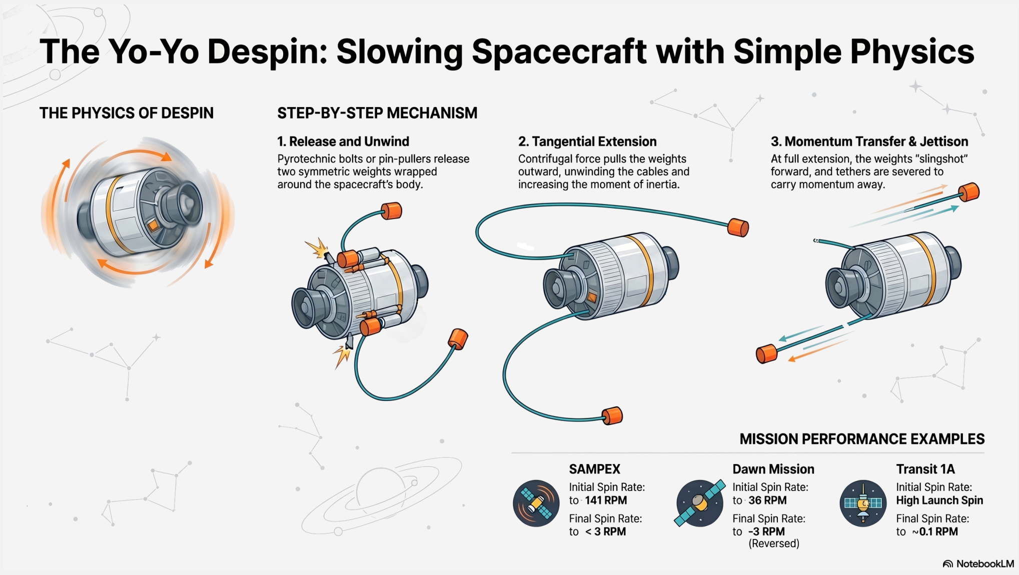


Figure 9.50: **Conservation of the angular momentum on variation of inertia momentum**

In figure skating and pole dance, the concept of angular momentum is extremely important in order to complete a successful figure, without it there would be no nice spinning effect at all!

Remark

A technique named the "yo-yo despin effect" reduces an spacecraft's rotation by exploiting moment of inertia following the idea just seen above. Two small masses attached to cables are released and swing outward as the spacecraft spins; when the masses move farther from the rotation axis, the system's moment of inertia increases, and—because angular momentum must stay constant—the rotation rate decreases. Once the spin slows to the desired level, the cables can be jettisoned, leaving the spacecraft stabilized without needing continuous propulsion.



The Yo-Yo Despin: Slowing Spacecraft with Simple Physics

THE PHYSICS OF DESPIN

STEP-BY-STEP MECHANISM

1. Release and Unwind
Pyrotechnic bolts or pin-pullers release two symmetric weights wrapped around the spacecraft's body.

2. Tangential Extension
Contrifugal force pulls the weights outward, unwinding the cables and increasing the moment of inertia.

3. Momentum Transfer & Jettison
At full extension, the weights "slingshot" forward, and tethers are severed to carry momentum away.

MISSION PERFORMANCE EXAMPLES

SAMPEX
Initial Spin Rate:
to **141 RPM**
Final Spin Rate:
to **< 3 RPM**

Dawn Mission
Initial Spin Rate:
to **36 RPM**
Final Spin Rate:
to **−3 RPM**
(Reversed)

Transit 1A
Initial Spin Rate:
High Launch Spin
Final Spin Rate:
to **~0.1 RPM**

Apart for the Earth that rotates on itself, the Sun (actually a G2V-type star, aka yellow dwarf) rotates on itself also (as most stars do!). But when a star collapse into a neutron star or even into a Black Hole, most of the angular momentum is conserved and therefore the spin rotation increase drastically:

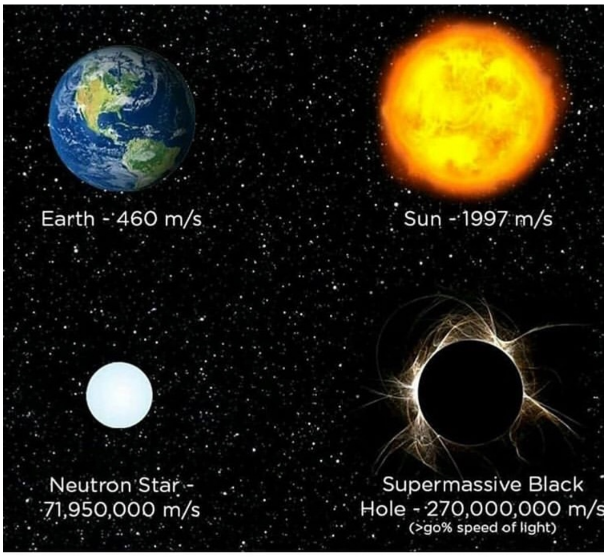

Earth - 460 m/s
Sun - 1997 m/s
Neutron Star - 71,950,000 m/s
Supermassive Black Hole - 270,000,000 m/s
(>90% speed of light)

Figure 9.51: How fast does it spin at equator?

Another curious experience (but mathematically correct) to take place ourselves on a turntable with a wheel in rotation held horizontally (vertical angular momentum is zero) and put it then vertically (same type of experiment can be done in the opposite direction). As the vertical momentum must remain zero to counteract this (or must remain constant in direction and in amplitude as it is a vectorial quantity), the plateau on which is put the experimenter will rotate in the opposite direction of rotation of the wheel.

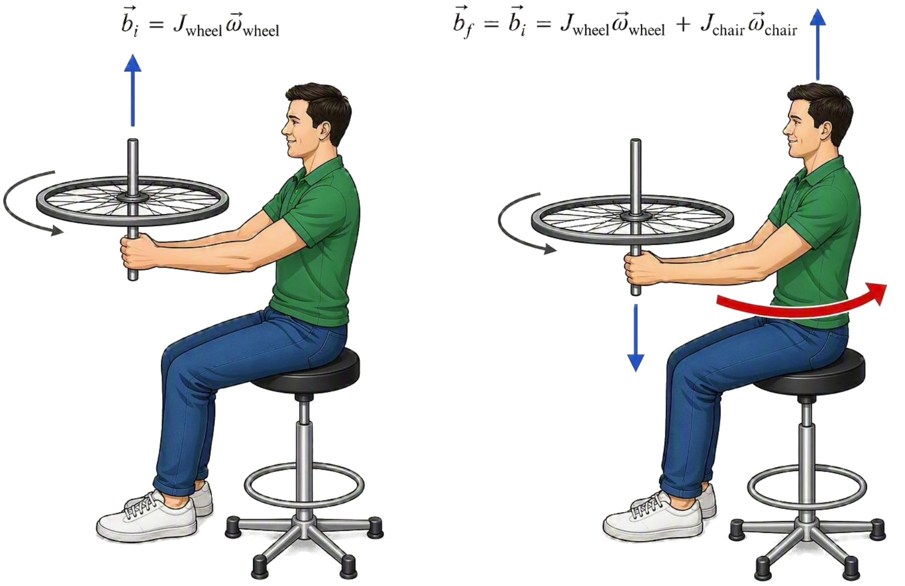

$\vec{b}_i = J_{\text{wheel}}\vec{\omega}_{\text{wheel}}$
$\vec{b}_f = \vec{b}_i = J_{\text{wheel}}\vec{\omega}_{\text{wheel}} + J_{\text{chair}}\vec{\omega}_{\text{chair}}$

Figure 9.52: Conservation of total angular momentum by variation of intern angular momentum

Remarks

R1. The movement of large masses on the surface of the Earth (icebergs, river flooding, tectonic plates, etc.) cause changes in moment of inertia of the Earth. It follows therefore fluctuations of the angular velocity and therefore imperfection of astronomical time standard (a few thousandths per day).

R2. A "reaction wheel" (or "momentum wheel") is a type of flywheel used primarily by spacecraft for attitude control without using fuel for rockets or other reaction devices. They are particularly useful when the spacecraft must be rotated by very small amounts, such as keeping a telescope pointed at a star. They may also reduce the mass fraction needed for fuel. This is accomplished by equipping the spacecraft with an electric motor attached to a flywheel which, when its rotation speed is changed, causes the spacecraft to begin to counter-rotate proportionately through conservation of angular momentum. Reaction wheels can rotate a spacecraft only around its center of mass; they are not capable of moving the spacecraft from one place to another.

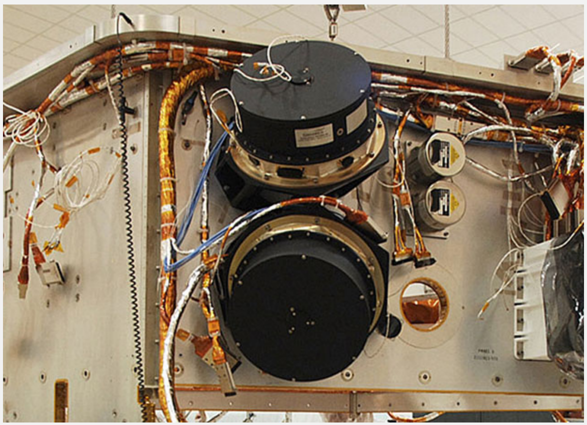


Figure 9.53: **Kepler satellite reaction wheels**

R3. A "control moment gyroscope" is a related but different type of attitude actuator. When mounted to a rigid spacecraft, applying a constant torque to the wheel using one of the gimbal motors causes the spacecraft to develop a constant angular velocity about a perpendicular axis, thus allowing control of the spacecraft's pointing direction. CMGs are generally able to produce larger sustained torques than reaction wheel with less motor heating, and are preferentially used in larger and/or more-agile spacecraft, including Skylab, Mir, and the International Space Station (ISS).

ISS011E11356

Figure 9.54: One of the four, 122 [cm] diameter, ISS control moment gyroscopes (source: NASA)

Now let us come back to the methods of calculation of the moments of inertia. The kinetic energy of a body is the sum of kinetic energy of each element of this body at the macroscopic level, then we have:

$$E_c=\sum_{i=1}^{n} E_{c,i} \tag{9.3.362}$$

Therefore in the context of a solid rigid body rotating around an axis, we have:

$$E_c=\sum_{i=1}^{n} E_{c,i}=\sum_{i=1}^{n}\frac{1}{2}J_i\omega^2=\frac{1}{2}\omega^2\sum_{i=1}^{n} J_i=\frac{1}{2}\omega^2 J \tag{9.3.363}$$

Thus, for a body composed of a plurality of different geometries, the total moment of inertia is the sum of the moments of inertia with respect to the axis of rotation of each of the sub-geometries such that:

$$\sum_{i=1}^{n} J_i=J \tag{9.3.364}$$

When we calculate the moment of a body of inertia relative to a given axis Δ of rotation, it may be interesting to know what is the distance from the axis where we can place fictitiously the whole mass of this body concentrated on a given point for the same moment of inertia. By definition, this distance denoted k and named the "radius of gyration" is trivially given by:

$$\boxed{J_\Delta=Mk^2 \Rightarrow k=\sqrt{\frac{J_\Delta}{M}}} \tag{9.3.365}$$

where J_Δ is the known moment of inertia of the body of mass M relatively to an axis Δ.

By definition, the "polar moment of inertia" (or also "squared moment of inertia") is the moment of inertia defined with respect to a point (a "pole") and not relatively to an axis and written:

$$J_0=\int r^2\mathrm{d}m \tag{9.3.366}$$

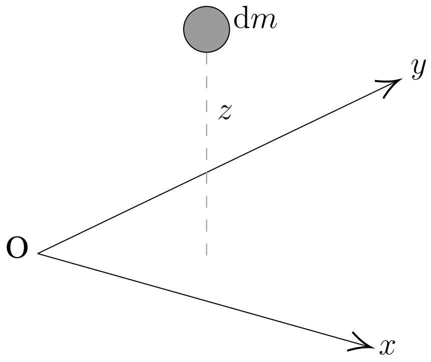
This quantity intervene only for free spins and is of interest, for rotations around a fixed axis, only because it sometimes facilitates the calculation of the axial moments of inertia under the follows relation (in Cartesian coordinates):

$$J_0=\frac{1}{2}(J_x+J_y+J_z) \tag{9.3.367}$$

Proof 9.4.5.

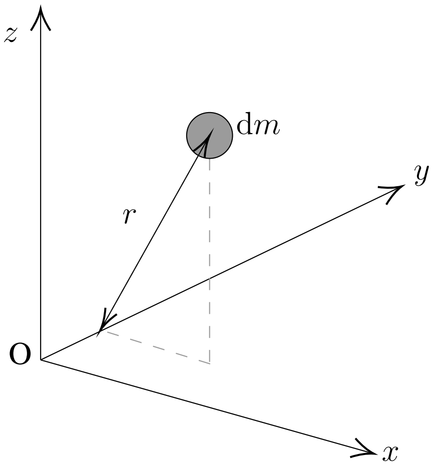
Lemma 9.4.1. The moment of inertia compared to the $x\mathrm{O}y$ plane is trivially given by:

$$J_{x,y}=\int z^2\mathrm{d}m \tag{9.3.368}$$



Lemma 9.4.2. The moment of inertia with respect to an axis is given by:

$$\begin{aligned}J_x&=\int r^2\mathrm{d}m=\int(y^2+z^2)\mathrm{d}m\\J_y&=\int(x^2+z^2)\mathrm{d}m\\J_z&=\int(x^2+y^2)\mathrm{d}m\end{aligned} \tag{9.3.369}$$

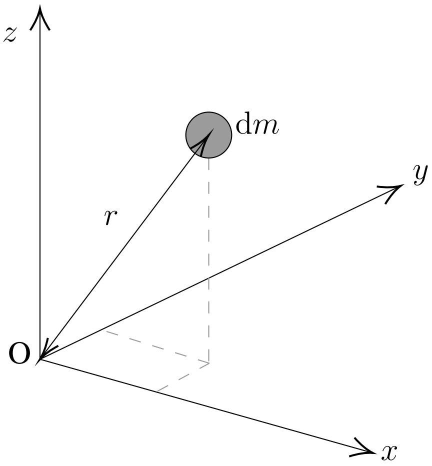


Summing these relations, we deduce:

$$J_x+J_y+J_z=2\int(x^2+y^2+z^2)\mathrm{d}m \tag{9.3.370}$$

The polar moment of inertia is then given by:

$$J_0=\int r^2\mathrm{d}m=\int(x^2+y^2+z^2)\mathrm{d}m \tag{9.3.371}$$



Comparing with the second lemma above we get:

$$\boxed{J_0=\frac{1}{2}(J_x+J_y+J_z)} \tag{9.3.372}$$

■ Q.E.D.

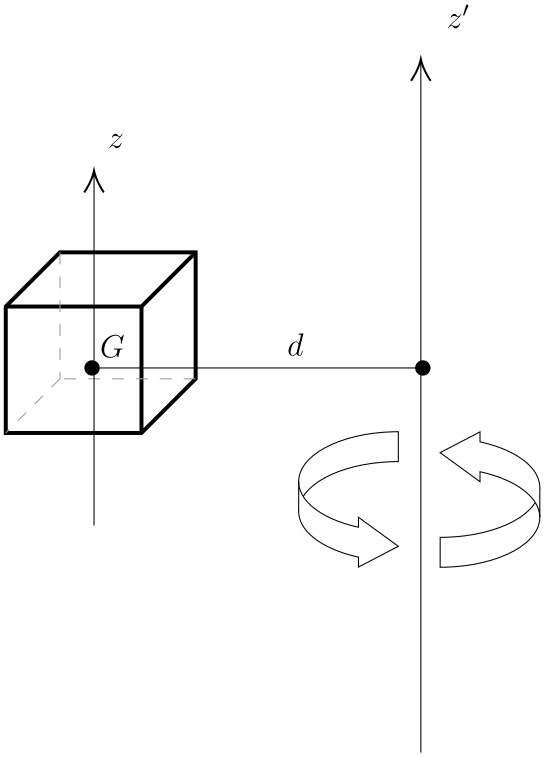
If the body in question has a spherical symmetry, it comes immediately as $J_x=J_y=J_z:=J$ that:

$$J_0=\frac{1}{2}(3J)\Rightarrow J=\frac{3}{2}J_0 \tag{9.3.373}$$

An example is given with the ball (full sphere) in the section on Geometric Shapes in the chapter Geometry (see page 2964).

Suppose now that we know the moment of inertia J_z of a rigid solid body compared to any other perpendicular axis (the chosen axis is not necessarily only assimilated to the common z axis) passing through the center of mass G. Then let us calculate the moment of inertia J_z, relatively to another axis z', parallel to z and remote from distance d, and let us make up appear the existing link between these two different moments of inertia.

So consider the following figure representing our subject of interest:



In a planar Cartesian reference frame, we have for any points (x,y):

$$r^2=x^2+y^2 \tag{9.3.374}$$

and:

$$r'^2=x^2+(y+d)^2 \tag{9.3.375}$$

Then we have:

$$J_z=\int r^2\mathrm{d}m=\int(x^2+y^2)\mathrm{d}m \tag{9.3.376}$$

and:

$$\begin{aligned}J_{z'}&=\int r'^2\mathrm{d}m=\int(x^2+(y+d)^2)\mathrm{d}m=\int(x^2+y^2+2dy+d^2)\mathrm{d}m\\&=\int(x^2+y^2)\mathrm{d}m+\int(2dy+d^2)\mathrm{d}m=J_z+2d\int y\mathrm{d}m+d^2\int\mathrm{d}m=J_z+2d\int y\mathrm{d}m+d^2M\end{aligned} \tag{9.3.377}$$

The term:

$$2d\int y\mathrm{d}m \tag{9.3.378}$$

is zero because if the moment of inertia is calculated relatively to the center mass G as we have assumed it from the beginning, then:

$$y_G=\frac{\displaystyle\int y\mathrm{d}V}{\displaystyle\int \mathrm{d}V}=0 \tag{9.3.379}$$

Ultimately, we finally get the "Huygens-Steiner theorem" also named "parallel axis theorem":

$$\boxed{J_{z'}=J_z+d^2M} \tag{9.3.380}$$

As we will see it in the section on Geometric Shapes in the chapter Geometry (see page 2886), it then becomes easy to calculate the moment of inertia of an equilateral triangle knowing that of a square plate and moving the axis inertia to the point where the triangle is the center of gravity (i.e. at the third of the median between the center of the rectangle and of its vertices).

Since there are as many moments of inertia as axes of rotation and that they are often in practical study cases related to the principal axes of inertia (axis related to the axes of revolution or symmetry planes - see below), it may be useful to introduce a mathematical being useful in the context of representation of moments of inertia which is none other than the "inertia matrix" or also known as (modern formulation) "inertia tensor".

The approach to rigorously determine the expression of this tensor is the following:

Given A_i a given point of a solid, for which we seek to calculate the moment of inertia and Δ the axis of origin O and vector of unit vector $\vec{u}$ relative to which we wish to calculate the moment of inertia. Any point A_i of the solid can be projected (orthogonal projection) on a point $H\in\Delta$ from the knowledge of the angle θ between Δ and $\overrightarrow{\mathrm{O}A_i}$ such that:

$$\overline{HA}=\overline{\mathrm{O}A}\sin(\theta)\Rightarrow\overline{HA}^2=\overline{\mathrm{O}A_i}^2\sin^2(\theta)=||\overrightarrow{\mathrm{O}A_i}\times\vec{u}||^2 \tag{9.3.381}$$

Therefore:

$$J_\Delta=\int r^2\mathrm{d}m=\int||\overrightarrow{\mathrm{O}A_i}\times\vec{u}||^2\mathrm{d}m=\int||\overrightarrow{\mathrm{O}A_i}\times\vec{u}||\cdot||\overrightarrow{\mathrm{O}A_i}\times\vec{u}||\mathrm{d}m=\int(\overrightarrow{\mathrm{O}A_i}\times\vec{u})\circ(\overrightarrow{\mathrm{O}A_i}\times\vec{u})\mathrm{d}m \tag{9.3.382}$$

Based on the properties of the mixed product (see section Vector Calculus page 1838) and of the scalar product:

$$(\vec{x}\times\vec{y})\circ\vec{z}=\vec{x}\circ(\vec{y}\times\vec{z})\qquad\text{and}\qquad(\vec{x}\times\vec{y})\circ\vec{z}=\vec{z}\circ(\vec{x}\times\vec{y}) \tag{9.3.383}$$

we have:

$$(\overrightarrow{\mathrm{O}A_i}\times\vec{u})\circ(\overrightarrow{\mathrm{O}A_i}\times\vec{u})=\vec{w}\circ(\overrightarrow{\mathrm{O}A_i}\times\vec{u})=(\vec{w}\times\overrightarrow{\mathrm{O}A_i})\circ\vec{u}=\vec{u}\circ(\vec{w}\times\overrightarrow{\mathrm{O}A_i})=\vec{u}\circ\left(\left(\overrightarrow{\mathrm{O}A_i}\times\vec{u}\right)\times\overrightarrow{\mathrm{O}A_i}\right) \tag{9.3.384}$$

and therefore:

$$J_\Delta=\int\vec{u}\circ\left(\left(\overrightarrow{\mathrm{O}A_i}\times\vec{u}\right)\times\overrightarrow{\mathrm{O}A_i}\right)\mathrm{d}m \tag{9.3.385}$$

As $\vec{u}$ is a vector of constant direction whatever the point of integration, we can take it out of the integral such that:

$$J_\Delta=\vec{u}\circ\int\left(\left(\overrightarrow{\mathrm{O}A_i}\times\vec{u}\right)\times\overrightarrow{\mathrm{O}A_i}\right)\mathrm{d}m=\vec{u}\circ\vec{v} \tag{9.3.386}$$

We can check that if we replace $\vec{u}$ by $\vec{u}_1+\vec{u}_2$, we will get a result that also sum (of the type $\vec{v}=\vec{v_1}+\vec{v}_2$ that is the sum of two integrals) up by the property of linearity of the cross product (see section Vector Calculus page 1833).

Therefore, the application that to $\vec{u}$ associated a given $\vec{v}$ is therefore a linear application that can be represented in a given basis $\mathcal{B}$ by a matrix:

$$\vec{u}\rightarrow\vec{v}=[J_0]_{\mathcal{B}}\vec{u} \tag{9.3.387}$$

The matrix $[J_0]_{\mathcal{B}}$ is named "inertia tensor" of the system relatively to the point O, in the base $\mathcal{B}$.

The moment of inertia of a system with respect to any axis Δ of unit vector $\vec{u}$ is given by:

$$\boxed{\forall \text{O}\in\Delta:\quad J_{\Delta}=\vec{u}\circ([J_0]_{\mathcal{B}}\vec{u})} \tag{9.3.388}$$

The problem now is to calculate the elements of the tensor $[J_0]$ for a given basis $\mathcal{B}$. Given a reference frame $(\text{O},\vec{i},\vec{j},\vec{k})$ such that $\text{O}\in\Delta$ we put:

$$\vec{u}=\alpha\vec{i}+\beta\vec{j}+\gamma\vec{k} \tag{9.3.389}$$

and:

$$\overrightarrow{\text{O}A}=x\vec{i}+y\vec{j}+z\vec{k} \tag{9.3.390}$$

Using the fact that we know that a cross product can be represented by as an asymmetric matrix operation (see section Vector Calculus page 1831):

$$\overrightarrow{\text{O}A_i}\times\vec{u}=\begin{pmatrix}0 & -z & y\\ z & 0 & -x\\ -y & x & 0\end{pmatrix}\begin{pmatrix}\alpha\\ \beta\\ \gamma\end{pmatrix} \tag{9.3.391}$$

and therefore:

$$\begin{aligned}[J_0]_{\mathcal{B}}\vec{u}&=\int\left(\left(\overrightarrow{\text{O}A_i}\times\vec{u}\right)\times\overrightarrow{\text{O}A_i}\right)=-\int\left(\overrightarrow{\text{O}A_i}\times\left(\overrightarrow{\text{O}A_i}\times\vec{u}\right)\right)\text{d}m\\ &=-\int\begin{pmatrix}-z^2-y^2 & xy & xz\\ xy & -z^2-x^2 & yz\\ xz & yz & -y^2-x^2\end{pmatrix}\begin{pmatrix}\alpha\\ \beta\\ \gamma\end{pmatrix}\text{d}m\\ &=\begin{pmatrix}\int(z^2+y^2)\text{d}m & -\int xy\text{d}m & -\int xz\text{d}m\\ -\int xy\text{d}m & \int(z^2+x^2)\text{d}m & -\int yz\text{d}m\\ -\int xz\text{d}m & -\int yz\text{d}m & \int(y^2+x^2)\text{d}m\end{pmatrix}\begin{pmatrix}\alpha\\ \beta\\ \gamma\end{pmatrix}\end{aligned} \tag{9.3.392}$$

In the above expression of the inertia matrix, we recognize the diagonal elements: it is simply the moments of inertia of the system relatively to the different axes of the base. We name "product of inertia" the non-diagonal elements of the matrix and we write them:

$$J_{xy}=-\int xy\text{d}m\qquad J_{xz}=-\int xz\text{d}m\qquad J_{yz}=-\int yz\text{d}m \tag{9.3.393}$$

Therefore we have:

$$\boxed{J=\begin{pmatrix}J_{\text{O}x} & J_{xy} & J_{xz}\\ J_{xy} & J_{\text{O}y} & J_{yz}\\ J_{xz} & J_{yz} & J_{\text{O}z}\end{pmatrix}=\begin{pmatrix}\int(z^2+y^2)\text{d}m & -\int xy\text{d}m & -\int xz\text{d}m\\ -\int xy\text{d}m & \int(z^2+x^2)\text{d}m & -\int yz\text{d}m\\ -\int xz\text{d}m & -\int yz\text{d}m & \int(y^2+x^2)\text{d}m\end{pmatrix}} \tag{9.3.394}$$

where obviously:

- $J_{\mathrm{O}x}$ is the inertial momentum relatively to $(\mathrm{O},\vec{x})$
- $J_{\mathrm{O}y}$ is the inertial momentum relatively to $(\mathrm{O},\vec{y})$
- $J_{\mathrm{O}z}$ is the inertial momentum relatively to $(\mathrm{O},\vec{z})$
- J_{yz} is the inertial momentum relatively to $(\vec{y},\vec{z})$
- J_{xz} is the inertial momentum relatively to $(\vec{x},\vec{z})$
- J_{xy} is the inertial momentum relatively to $(\vec{x},\vec{y})$

Also sometimes written as (USA notation and mechanical engineers notation):

$$J=\begin{pmatrix} I_{\mathrm{O}x} & I_{xy} & I_{xz} \\ I_{xy} & I_{\mathrm{O}y} & I_{yz} \\ I_{xz} & I_{yz} & I_{\mathrm{O}z} \end{pmatrix}=\begin{pmatrix} I_{xx} & I_{xy} & I_{xz} \\ I_{xy} & I_{yy} & I_{yz} \\ I_{xz} & I_{yz} & I_{zz} \end{pmatrix}=\begin{pmatrix} A & -F & -E \\ -F & B & -D \\ -E & -D & C \end{pmatrix} \tag{9.3.395}$$

If O is assimilated to the center of mass of the solid, we simply write:

$$(J_{\mathrm{O}x},J_{\mathrm{O}y},J_{\mathrm{O}z})=(J_x,J_y,J_z) \tag{9.3.396}$$

It follows if thickness of the study body is negligible compared to the other dimensions then $z=0$ and then we have immediately:

$$J_{xy}=\begin{pmatrix} \int y^2\mathrm{d}m & -\int xy\mathrm{d}m & 0 \\ -\int xy\mathrm{d}m & \int x^2\mathrm{d}m & 0 \\ 0 & 0 & \int(y^2+x^2)\mathrm{d}m \end{pmatrix}=\begin{pmatrix} J_{\mathrm{O}x} & J_{xy} & 0 \\ J_{xy} & J_{\mathrm{O}y} & 0 \\ 0 & 0 & J_{\mathrm{O}z} \end{pmatrix}=\begin{pmatrix} J_{\mathrm{O}x} & J_{xy} & 0 \\ J_{xy} & J_{\mathrm{O}y} & 0 \\ 0 & 0 & J_{\mathrm{O}x}+J_{\mathrm{O}y} \end{pmatrix} \tag{9.3.397}$$

We can also generalize Huygens theorem by making use of this matrix. To do so, let us write (x',y',z') the coordinates of any point A in R' and (x,y,z) its coordinates in R. We write (a,b,c) the coordinates of the O' of R' in R:

$$R\Rightarrow\begin{cases} x'=x-a \\ y'=y-b \\ z'=z-c \end{cases} \tag{9.3.398}$$

as:

$$\overrightarrow{\mathrm{OO}'}+\overrightarrow{\mathrm{O}'A}=\overrightarrow{\mathrm{O}A}\Leftrightarrow\overrightarrow{\mathrm{O}'A}=\overrightarrow{\mathrm{O}A}-\overrightarrow{\mathrm{OO}'} \tag{9.3.399}$$

We then have:

$$\begin{aligned} J_{\mathrm{O}z}&=\int(x^2+y^2)\mathrm{d}m=\int\left((x'+a)^2+(y'+b)^2\right)\mathrm{d}m \\ &=\int(x'^2+y'^2)\mathrm{d}m+2a\int x'\mathrm{d}m+a^2\int\mathrm{d}m+2b\int y'\mathrm{d}m+b^2\int\mathrm{d}m \\ &=J_{\mathrm{O}z'}+M(a^2+b^2)+2a\int x'\mathrm{d}m+2b\int y'\mathrm{d}m \end{aligned} \tag{9.3.400}$$

Now if O' coincides with the center of mass G, then according to the definition of the center of mass:

$$\int \overrightarrow{GA}\mathrm{d}m=\vec{0}\Rightarrow\begin{cases}\int x'\mathrm{d}m=0\\ \int y'\mathrm{d}m=0\\ \int z'\mathrm{d}m=0\end{cases}\tag{9.3.401}$$

We deduce then:

$$J_{\mathrm{O}z}=J_{\mathrm{G}z'}+M(a^2+b^2)\tag{9.3.402}$$

and also:

$$J_{\mathrm{O}x}=J_{\mathrm{G}x'}+M(b^2+c^2)\qquad J_{\mathrm{O}y}=J_{\mathrm{G}y'}+M(a^2+c^2)\tag{9.3.403}$$

with:

$$\overrightarrow{\mathrm{O}G}=a\vec{i}+b\vec{j}+c\vec{k}=x_G\vec{i}+y_G\vec{j}+z_G\vec{k}\tag{9.3.404}$$

We fall back on the classic Huygens theorem since a^2+b^2 is none other than the squared distance between the axis Oz and Gz and same for b^2+c^2 that is the squared distance between Ox and Gx and a^2+c^2 which is the distance between Oy and Gy.

If we now focus on the products of inertia, we get:

$$J_{xy}=-\int xy\mathrm{d}m=-\int(x'+a)(y'+b)\mathrm{d}m=-\int x'y'\mathrm{d}m-a\int y'\mathrm{d}m-b\int x'\mathrm{d}m-ab\int\mathrm{d}m\tag{9.3.405}$$

Therefore, if O' coincide with G:

$$\begin{aligned}J_{xy}&=J_{x'y'}-Mab\\ J_{xz}&=J_{x'z'}-Mac\\ J_{yz}&=J_{y'z'}-Mbc\end{aligned}\tag{9.3.406}$$

In summary, the "Huygens generalized theorem" is written:

$$\boxed{[J_0]_{(\vec{i},\vec{j},\vec{k})}=[J_G]_{(\vec{i},\vec{j},\vec{k})}+\begin{pmatrix}M(y_G^2+z_G^2) & -Mx_Gy_G & -Mx_Gz_G\\ -Mx_Gy_G & M(x_G^2+z_G^2) & -My_Gz_G\\ -Mx_Gz_G & -My_Gz_G & M(x_G^2+y_G^2)\end{pmatrix}}\tag{9.3.407}$$

The inertia tensor being real and symmetric, we have proved in the section of Linear Algebra (spectral theorem at page) that it is always possible to find three perpendicular directions of vectors $\vec{e}_1,\vec{e}_2,\vec{e}_3$ such as the symmetrical tensor (matrix) is diagonalisable:

$$[J_0]_{\vec{e}_1,\vec{e}_2,\vec{e}_3}=\begin{pmatrix}J_1 & 0 & 0\\ 0 & J_2 & 0\\ 0 & 0 & J_3\end{pmatrix}\tag{9.3.408}$$

The trihedron formed by the vector $\vec{e}_1,\vec{e}_2,\vec{e}_3$ is named "main trihedron of inertia" and its axes are named "principal axes of inertia". In this reference frame $[J_0]$ takes the name of "principal inertia tensor". If furthermore O is assimilated to G, we speak of "central tensor of inertia".

In fact, to find the moments of inertia relatively to the principal axes it is almost never necessary to diagonalise the inertia tensor because it is often enough to be guided by the symmetry of the system. We'll see the following theorems that if there are axes or planes of symmetry for the mass distribution, the axes of inertia are easy to find. In addition, the system is generally simple enough (or sufficiently broken down into simple elements ...) for these axes to be obvious.

Theorem 9.5. If the system has a material plane of symmetry (i.e.: $\rho(A)=\rho(A')$ if A is the symmetric of A' relatively to the plane) then any axis perpendicular to this plane is a principal axis of inertia and therefore the inertia matrix reduce to:

$$J_{xy}=\begin{pmatrix} J_{\mathrm{O}x} & J_{xy} & 0 \\ J_{xy} & J_{\mathrm{O}y} & J_{yz} \\ 0 & 0 & J_{\mathrm{O}z} \end{pmatrix} \tag{9.3.409}$$

Proof 9.5.1. Let us choose a reference frame $x\mathrm{O}y$ in the plane relatively to which the system has a mass distribution that is symmetric and $\mathrm{O}z$ an axis perpendicular to this plane. To calculate $\int xz\mathrm{d}m$ or $\int yz\mathrm{d}m$ let us group the points by pairs, symmetrical to $x\mathrm{O}y$. that is to say such that $z_A=-z_{A'}$. We then have:

$$x_Az_A+x_{A'}z_{A'}=0 \tag{9.3.410}$$

and even:

$$y_Az_A+y_{A'}z_{A'}=0 \tag{9.3.411}$$

that is to say, as (material symmetry!):

$$\mathrm{d}m_A=\mathrm{d}m_{A'} \tag{9.3.412}$$

that all contributions of symmetrical pairs of points are zero, which implies:

$$J_{xz}=J_{yz}=0 \tag{9.3.413}$$

that is to say, the z axis is a main direction of inertia.

■ Q.E.D.

Theorem 9.6. Let us Choose $\mathrm{O}z$ as axis of symmetry. As above, we have:

$$\int xz\mathrm{d}m=\int yz\mathrm{d}m=0 \tag{9.3.414}$$

Proof 9.6.1. Indeed, because if we group the points by pairs A and A' symmetrical to $\mathrm{O}z$, we have:

$$x_A=-x_A \qquad y_A=-y_{A'} \tag{9.3.415}$$

but as $z_A=z_{A'}$ we have always:

$$x_Az_A+x_{A'}z_{A'}=0 \tag{9.3.416}$$

and even:

$$y_Az_A+y_{A'}z_{A'}=0 \tag{9.3.417}$$

■ Q.E.D.

Remark

When we have identified two principal axes of inertia through previous symmetries, the third is simply the necessary one to complete an orthogonal (trihedron) axis system.

Theorem 9.7. If a system admits an axis of revolution for its mass distribution, then all orthogonal trihedron including the axis of revolution, is the principal trihedron of inertia. The material system is then said to be a "cylindrical system" and inside the principal inertial trihedron the inertial tensor takes the form (assuming that the axis of rotation is the third axis of the coordinate system):

$$[J_0]_{\vec{e}_1,\vec{e}_2,\vec{e}_3}=\begin{pmatrix} J & 0 & 0 \\ 0 & J & 0 \\ 0 & 0 & J' \end{pmatrix} \tag{9.3.418}$$

Proof 9.7.1. If Oz is an axis of revolution, any plane including Oz is a plane of symmetry and line perpendicular to Oz is principal axis of inertia (first theorem). In addition, all these lines perpendicular to Oz are equivalent.

■ Q.E.D.

Definition (#9.3.442): If the matrix of inertia O of a material system is of the type:

$$[J_0]_{\vec{e}_1,\vec{e}_2,\vec{e}_3}=\begin{pmatrix} J & 0 & 0 \\ 0 & J & 0 \\ 0 & 0 & J \end{pmatrix} \tag{9.3.419}$$

then we say that the system is a "spherical system" (or a "spherically symmetric system").

Remark

The systematic choice of a main trihedron of inertia gives the possibility to reduce the inertia tensor of 6 to 3 components only, calculated once and for all. However, this choice involves the use of a base that will most often be in motion relative to the repository used, which may pose problems derivations with respect to time of the basis vectors. We can then, if it's easier, obtain the components of the tensor of symmetry in any base with a matrix of bases change between the main base and the base used for the calculation of the main trihedron of inertia.

Theorem 9.8. When the moments of inertia of a solid are known in the directions of all principal axes of inertia, we can easily determine the moment of inertia J with respect to any other axis through the center of gravity using what we name an "ellipsoid of inertia" (not to be confused with the moment of inertia of an ellipsoid - proved in the section Geometric Shapes at page 2981).

Proof 9.8.1. Given three axes, centered on G, parallel to the main axes. In their directions, let us choose lengths proportional to:

$$A=\frac{1}{\sqrt{J_x}} \qquad B=\frac{1}{\sqrt{J_y}} \qquad C=\frac{1}{\sqrt{J_z}} \tag{9.3.420}$$

In this phase space of moments of inertia, any point $P=(x,y,z)$ denotes a moment of inertia J such that:

$$||\overrightarrow{GP}||=x^2+y^2+z^2=\frac{1}{J} \tag{9.3.421}$$

To determine J based on the J_x,J_y,J_z, without having to calculate x,y,z, we identify the cosine directions (see section Vector Calculus page 1811) of the axis of rotation to those of the straight line $\overline{GP}$:

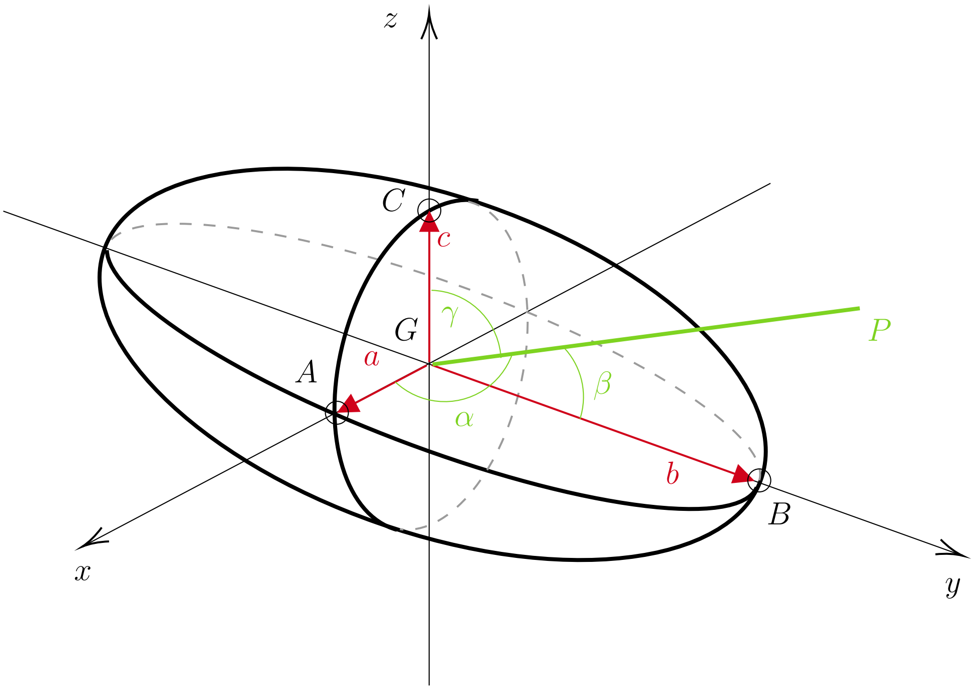


Figure 9.55: **Illustration of the ellipsoid of inertia**

Thus we have:

$$(\cos(\alpha),\cos(\beta),\cos(\gamma))=\left(\frac{x}{||\overrightarrow{GP}||},\frac{y}{||\overrightarrow{GP}||},\frac{z}{||\overrightarrow{GP}||}\right)=\left(\frac{x}{\dfrac{1}{\sqrt{J}}},\frac{y}{\dfrac{1}{\sqrt{J}}},\frac{z}{\dfrac{1}{\sqrt{J}}}\right)=\sqrt{J}(x,y,z) \tag{9.3.422}$$

Therefore:

$$x=\frac{\cos(\alpha)}{\sqrt{J}}\qquad y=\frac{\cos(\beta)}{\sqrt{J}}\qquad z=\frac{\cos(\gamma)}{\sqrt{J}} \tag{9.3.423}$$

Therefore:

$$J||\overrightarrow{GP}||^2=Jx^2+Jy^2+Jz^2=1 \tag{9.3.424}$$

We can now calculate the conditions of normalization of this relation. Thus, if $y=z=0$ and $x=a$, we have:

$$Jx^2+Jy^2+Jz^2=JA^2=1\Rightarrow A^2=\frac{1}{J}=\frac{1}{J_x} \tag{9.3.425}$$

Respectively we will have:

$$B^2=\frac{1}{J}=\frac{1}{J_y} \qquad C^2=\frac{1}{J}=\frac{1}{J_z} \tag{9.3.426}$$

As:

$$A=\frac{1}{\sqrt{J_x}}\Rightarrow A^2=\frac{1}{J_x} \qquad B=\frac{1}{\sqrt{J_y}}\Rightarrow B^2=\frac{1}{J_y} \qquad C=\frac{1}{\sqrt{J_z}}\Rightarrow C^2=\frac{1}{J_z} \tag{9.3.427}$$

■ Q.E.D.

Which brings us to write:

$$J_x x^2+J_y y^2+J_z z^2=\frac{x^2}{A^2}+\frac{y^2}{B^2}+\frac{z^2}{C^2}=1 \tag{9.3.428}$$

By substitution, we get:

$$\frac{x^2}{A^2}+\frac{y^2}{B^2}+\frac{z^2}{C^2}=\frac{1}{J}\left(\frac{\cos^2(\alpha)}{A^2}+\frac{\cos^2(\beta)}{B^2}+\frac{\cos^2(\gamma)}{C^2}\right)=\frac{1}{J}\left(J_x\cos^2(\alpha)+J_y\cos^2(\beta)+J_z\cos^2(\gamma)\right) \tag{9.3.429}$$

So finally:

$$\boxed{J_x\cos^2(\alpha)+J_y\cos^2(\beta)+J_z\cos^2(\gamma)=J} \tag{9.3.430}$$

Thus, knowing the moments of inertia of a body with respect to its principal axis J_x, J_y, J_z, we can know its moment of inertia relatively to any axis having an angle α,β,γ relatively to the main axes.

In other words, the moment of inertia J of a symmetric body like a sphere, ellipsoid, etc. around any general axis (here $\overrightarrow{GP}$) can be expressed as a linear combination of the moments of inertia around the principal axes (i.e. thanks to the squares of the cosines of the angles of the line OP with the x-, y- and z-axis. This is as powerful result used quite a lot in Astronomy and Astrophysics.

9.3.8.2.2 Euler's rotation equation

In what follow we will present a derivation in modern fashion of "Euler rotation equations", or also sometimes named "Euler equations", that is almost a simple copy/paste of Vedat Tanriverdi paper *Euler's derivation of rigid body equations*. This derivation can help understand rigid body rotations and inertia products. We will start by explaining the general structure and continue by giving derivations of Euler equations in the stationary reference frame and in the body reference frame.

Leonhard Euler first derived equations in the stationary reference frame, and then he realized that moments of inertia are time-dependent. Then, he derived equations in the body reference frame. We will also follow the same path!

In the figure below one can see a rigid body, stationary reference frame (x', y', z'), body reference frame (x, y, z), applied force $\vec{F}_a$ and position vector $\vec{l}$ showing acting position of the force. In the figure, the rigid body is represented by a stone to emphasize that it can have any shape, and it should be solid. Euler defines a rigid body with the usage of solid instead of rigid by following statement "*A body is called solid whose interior is not subject to any change, or which is such that all its parts constantly maintain the same relative position, whatever motion the whole body may have.*".

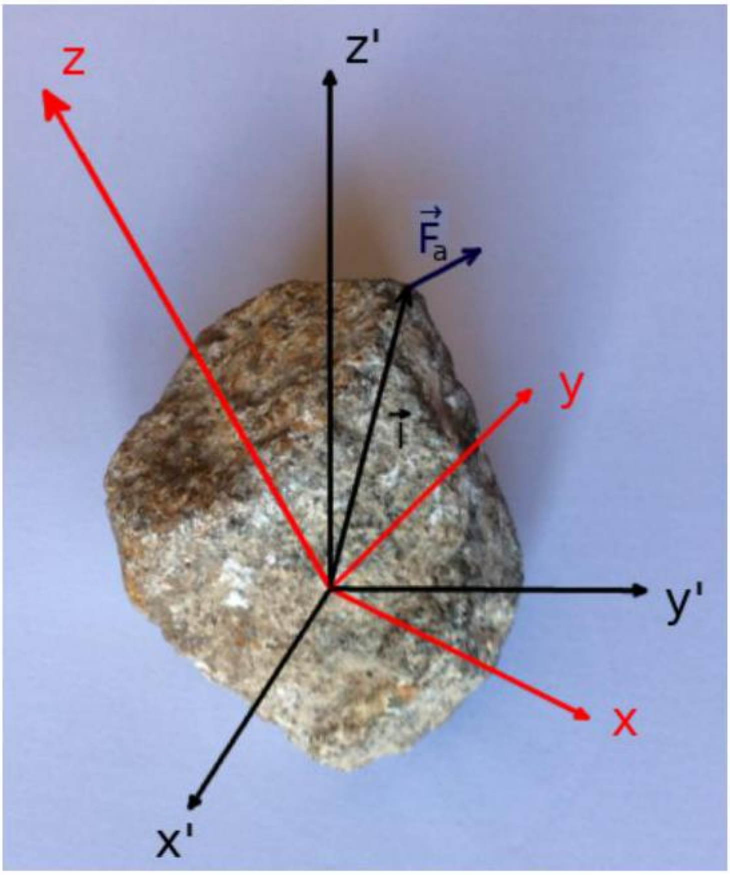


Figure 9.56: A rigid body (stone), stationary reference frame (x', y', z'), body reference frame (x, y, z), applied force $\vec{F}_a$ and position vector $\vec{l}$.

In general, a rigid body can make both translational and rotational motion. In this work, we will follow Euler and will consider only the rotational part of the motion. In the fibure above the origins of reference frames are fixed to the (approximate) center of mass.

We assume that the axes of the body reference frame lie on the principal axes of the rigid body. This assumption will provide simplicities in the calculations in the body reference frame by letting inertia products equal to zero.

An applied force $\vec{F}_a$ acting on the rigid body can cause changes in its angular momentum. We will follow Euler and consider an infinitesimal part of the rigid body to find how applied force changes the rigid body's angular momentum. This infinitesimal part should not lie on the rotation axis for the sake of generality.

9.3.8.2.2.1 Derivation in the stationary reference frame

In this section, we will follow the steps of Euler's derivation for the rotation of a rigid body in a stationary or inertial reference frame.

Like Euler, we will consider an infinitesimal part of the rotating rigid body. We can write its position in a stationary reference frame as:

$$\vec{r}=x'\hat{x}'+y'\hat{y}'+z'\hat{z}' \tag{9.3.431}$$

We can write the angular velocity of this infinitesimal part as:

$$\vec{w}=w_{x'}\hat{x}'+w_{y'}\hat{y}'+w_{z'}\hat{z}' \tag{9.3.432}$$

Then, we can obtain its velocity from the cross product $\vec{v}=\vec{w}\times\vec{r}$ as:

$$\vec{v}=\left(z'w_{y'}-y'w_{z'}\right)\hat{x}'-\left(z'w_{x'}-x'w_{z'}\right)\hat{y}'+\left(y'w_{x'}-x'w_{y'}\right)\hat{z}' \tag{9.3.433}$$

As a result, we can write the components of velocity as:

$$\begin{aligned}\frac{\mathrm{d}x'}{\mathrm{d}t}&=z'w_{y'}-y'w_{z'}\\ \frac{\mathrm{d}y'}{\mathrm{d}t}&=x'w_{z'}-z'w_{x'}\\ \frac{\mathrm{d}z'}{\mathrm{d}t}&=y'w_{x'}-x'w_{y'}\end{aligned} \tag{9.3.434}$$

If we take the derivative of these with respect to time and use them in derivatives, we can get components of acceleration as:

$$\begin{aligned}\frac{\mathrm{d}^2x'}{\mathrm{d}t^2}&=z'\dot{w}_{y'}-y'\dot{w}_{z'}+y'w_{x'}w_{y'}-x'w_{y'}^2-x'w_{z'}^2+z'w_{z'}w_{x'}\\ \frac{\mathrm{d}^2y'}{\mathrm{d}t^2}&=x'\dot{w}_{z'}-z'\dot{w}_{x'}+z'w_{y'}w_{z'}-y'w_{z'}^2-y'w_{x'}^2+x'w_{x'}w_{y'}\\ \frac{\mathrm{d}^2z'}{\mathrm{d}t^2}&=y'\dot{w}_{x'}-x'\dot{w}_{y'}+x'w_{z'}w_{x'}-z'w_{x'}^2-z'w_{y'}^2+y'w_{y'}w_{z'}\end{aligned} \tag{9.3.435}$$

By multiplying components of acceleration with the mass of the infinitesimal part, m, we can get components of force $\vec{F}=F_{x'}\hat{x}'+F_{y'}\hat{y}'+F_{z'}\hat{z}'$ according to Newton's second law $\vec{F}=m\vec{a}$. Then by using the force $\vec{F}$ and the equations above, we can obtain moments of the force (torque) $M_{x'}=y'F_{z'}-z'F_{y'}$, $M_{y'}=z'F_{x'}-x'F_{z'}$ and $M_{z'}=x'F_{y'}-y'F_{x'}$ as:

$$\begin{aligned}\mathrm{d}M_{x'}=&\mathrm{d}m\left[\left(y'^2+z'^2\right)\dot{w}_{x'}-x'y'\dot{w}_{y'}-x'z'\dot{w}_{z'}+x'y'w_{x'}w_{z'}-x'z'w_{x'}w_{y'}\right.\\ &\left.-y'z'\left(w_{y'}^2-w_{z'}^2\right)+w_{y'}w_{z'}\left(y'^2-z'^2\right)\right]\\ \mathrm{d}M_{y'}=&\mathrm{d}m\left[\left(x'^2+z'^2\right)\dot{w}_{y'}-x'y'\dot{w}_{x'}-y'z'\dot{w}_{z'}+y'z'w_{x'}w_{y'}-y'x'w_{y'}w_{z'}\right.\\ &\left.-x'z'\left(w_{z'}^2-w_{x'}^2\right)+w_{x'}w_{z'}\left(z'^2-x'^2\right)\right]\\ \mathrm{d}M_{z'}=&\mathrm{d}m\left[\left(x'^2+y'^2\right)\dot{w}_{z'}-x'z'\dot{w}_{x'}-y'z'\dot{w}_{y'}+x'z'w_{y'}w_{z'}-y'z'w_{x'}w_{z'}\right.\\ &\left.-x'y'\left(w_{x'}^2-w_{y'}^2\right)+w_{x'}w_{y'}\left(x'^2-y'^2\right)\right]\end{aligned} \tag{9.3.436}$$

We will integrate these equations to find the effect of applied force on the totality of the rigid body. Integration of the left-hand side gives the moment of the applied force relevant to that axis, e.g., $M_{x'}=\int \mathrm{d}M'_x$. By using integrals at the right-hand side, we can define components of moments of inertia tensor as:

$$\begin{array}{ll} J_{x'x'}=\int\left(y'^2+z'^2\right)\mathrm{d}m, & J_{x'y'}=J_{y'x'}=-\int \rho x'y'\mathrm{d}m \\ J_{y'y'}=\int\left(x'^2+z'^2\right)\mathrm{d}m, & J_{x'z'}=J_{z'x'}=-\int \rho x'z'\mathrm{d}m \\ J_{z'z'}=\int\left(x'^2+y'^2\right)\mathrm{d}m, & J_{y'z'}=J_{z'y'}=-\int y'z'\mathrm{d}m \end{array} \quad (9.3.437)$$

Therefore:

$$\boxed{\begin{aligned} M_{x'}=&J_{x'x'}\dot{w}_{x'}+J_{x'y'}\dot{w}_{y'}+J_{x'z'}\dot{w}_{z'}-J_{x'y'}w_{x'}w_{z'}+J_{x'z'}w_{x'}w_{y'}\\ &+J_{y'z'}\left(w_{y'}^2-w_{z'}^2\right)+w_{y'}w_{z'}\left(J_{z'z'}-J_{y'y'}\right)\\ M_{y'}=&J_{y'y'}\dot{w}_{y'}+J_{x'y'}\dot{w}_{x'}+J_{y'z'}\dot{w}_{z'}-J_{y'z'}w_{x'}w_{y'}+J_{x'y'}w_{y'}w_{z'}\\ &+J_{x'z'}\left(w_{z'}^2-w_{x'}^2\right)+w_{x'}w_{z'}\left(J_{x'x'}-J_{z'z'}\right)\\ M_{z'}=&J_{z'z'}\dot{w}_{z'}+J_{x'z'}\dot{w}_{x'}+J_{y'z'}\dot{w}_{y'}-J_{x'z'}w_{y'}w_{z'}+J_{y'z'}w_{x'}w_{z'}\\ &+J_{x'y'}\left(w_{x'}^2-w_{y'}^2\right)+w_{x'}w_{y'}\left(J_{y'y'}-J_{x'x'}\right) \end{aligned}} \quad (9.3.438)$$

These equations describe the rotation of a rigid body in the stationary reference frame by the effect of the moment of force, $\vec{M}=\vec{F}_a\times\vec{l}$.

In Classical Mechanics, Euler's rotation equations are a vectorial quasilinear first-order ordinary differential equation describing the rotation of a rigid body, using a rotating reference frame with angular velocity ω whose axes are fixed to the body.

One should be careful while using these equations since moments of inertia are (may be) time dependent in the stationary reference frame!

9.3.8.2.2.2 Derivation in the body reference frame

Now we will follow Euler's method in the stationary reference frame to get equations of rigid body rotations in the body reference frame under the influence of torque. We will consider an infinitesimal part of the rigid body again to see the effect of torque on the rotation of that infinitesimal part. The axes of the body reference frame are chosen as the principal axes of the body.

In the body reference frame, we can show the position of an infinitesimal part of the rigid body by:

$$\vec{r}=x\hat{x}+y\hat{y}+z\hat{z} \quad (9.3.439)$$

We can write the components of angular velocity as:

$$\vec{w}=w_x\hat{x}+w_y\hat{y}+w_z\hat{z} \quad (9.3.440)$$

Then, we can get the velocity of the infinitesimal part from cross-product $\vec{v}=\vec{w}\times\vec{r}$ as:

$$\vec{v}=\left(zw_y-yw_z\right)\hat{x}-\left(zw_x-xw_z\right)\hat{y}+\left(yw_x-xw_y\right)\hat{z} \quad (9.3.441)$$

As a result, we can write the components of the velocity as:

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}t}&=zw_y-yw_z\\ \frac{\mathrm{d}y}{\mathrm{d}t}&=xw_z-zw_x\\ \frac{\mathrm{d}z}{\mathrm{d}t}&=yw_x-xw_y\end{aligned} \tag{9.3.442}$$

By taking the time derivative of the equations above and using them, we can find the components of acceleration as:

$$\begin{aligned}\frac{\mathrm{d}^2x}{\mathrm{d}t^2}&=z\dot{w}_y-y\dot{w}_z+yw_xw_y-xw_y^2-xw_z^2+zw_zw_x\\ \frac{\mathrm{d}^2y}{\mathrm{d}t^2}&=x\dot{w}_z-z\dot{w}_x+zw_yw_z-yw_z^2-yw_x^2+xw_xw_y\\ \frac{\mathrm{d}^2z}{\mathrm{d}t^2}&=y\dot{w}_x-x\dot{w}_y+xw_zw_x-zw_x^2-zw_y^2+yw_yw_z\end{aligned} \tag{9.3.443}$$

where dots over w_i's represent again time derivatives. According to Newton's second law, $\vec{F}=m\vec{a}$, by multiplying the equations above with mass of the infinitesimal part, m, we can get components of the force $\vec{F}=m\mathrm{d}^2x/\mathrm{d}t^2\hat{x}+m\mathrm{d}^2y/\mathrm{d}t^2\hat{y}+m\mathrm{d}^2z/\mathrm{d}t^2\hat{z}$. This force is responsible for the acceleration of the rigid body's infinitesimal part and different than the applied force. It results from the combination of applied force and rigidity of the object.

Then by using the force $\vec{F}$ as we did before we can obtain the effects of moments of the force on angular accelerations and angular velocities as:

$$\begin{aligned}\mathrm{d}M_x=&\mathrm{d}m\left[\left(y^2+z^2\right)\dot{w}_x-xy\dot{w}_y-xz\dot{w}_z+xyw_xw_z-xzw_xw_y\right.\\ &\left.-yz\left(w_y^2-w_z^2\right)+w_yw_z\left(y^2-z^2\right)\right]\\ \mathrm{d}M_y=&\mathrm{d}m\left[\left(x^2+z^2\right)\dot{w}_y-xy\dot{w}_x-yz\dot{w}_z+yzw_xw_y-yxw_yw_z\right.\\ &\left.-xz\left(w_z^2-w_x^2\right)+w_xw_z\left(z^2-x^2\right)\right]\\ \mathrm{d}M_z=&\mathrm{d}m\left[\left(x^2+y^2\right)\dot{w}_z-xz\dot{w}_x-yz\dot{w}_y+xzw_yw_z-yzw_xw_z\right.\\ &\left.-xy\left(w_x^2-w_y^2\right)+w_xw_y\left(x^2-y^2\right)\right]\end{aligned} \tag{9.3.444}$$

We can integrate these equations over the volume of the rigid body. If we consider the torque on different infinitesimal parts, it changes as the distance from the rotation axis changes, and the moments of force are distributed over the rigid body to satisfy the rotation of the rigid body as a whole with the help of atomic interactions. This is the result of the object's rigidity. Then, the integration of the left-hand side gives torque relevant to that axis, e.g., $M_x=\int \mathrm{d}M_x$. Integrals at the right-hand side are related to moments of inertia and inertia products, and they are defined as:

$$\begin{aligned}J_{xx}&=\int\left(y^2+z^2\right)\mathrm{d}m, & J_{xy}&=J_{yx}=-\int xy\mathrm{d}m=0\\ J_{yy}&=\int\left(x^2+z^2\right)\mathrm{d}m, & J_{xz}&=J_{zx}=-\int xz\mathrm{d}m=0\\ J_{zz}&=\int\left(x^2+y^2\right)\mathrm{d}m, & J_{yz}&=J_{zy}=-\int yz\mathrm{d}m=0\end{aligned} \tag{9.3.445}$$

We choose the body axes as principal axes, then terms in the second column, inertia products, are equal to zero since the definition of principal axes is based on this.

Then the integration of these equations over the volume of the rigid body gives:

$$\boxed{\begin{aligned} M_x &= J_{xx}\dot{w}_x - w_y w_z \left(J_{yy} - J_{zz}\right) \\ M_y &= J_{yy}\dot{w}_y - w_z w_x \left(J_{zz} - J_{xx}\right) \\ M_z &= J_{zz}\dot{w}_z - w_x w_y \left(J_{xx} - J_{yy}\right) \end{aligned}} \qquad (9.3.446)$$

The above relation is most famously know under the following shape:

$$\boxed{\vec{M} = \vec{J}\dot{\vec{\omega}} + \vec{\omega} \times (\vec{J}\vec{\omega})} \qquad (9.3.447)$$

or sometimes:

$$\vec{M} = \vec{L} + \vec{\omega} \times \vec{L} \qquad (9.3.448)$$

or sometimes:

$$\tau_\alpha = J_\alpha \dot{\omega}_\alpha + \varepsilon_{\alpha l k}\omega_l \omega_k J_k \qquad (9.3.449)$$

with α fixed but a sum implied over the repeated l and k indices. Here ε_{abc} is the fully antisymmetric Levi-Civita symbol.

These are Euler equations for rigid bodies describing the change of the angular momentum under the influence of torque, $\vec{M} = \vec{F}_a \times \vec{l}$, in the body reference frame. In these equations, moments of inertia and inertia products are independent of time.

In Classical Mechanics, Euler's rotation equations are a vectorial quasilinear first-order ordinary differential equation describing the rotation of a rigid body, using a rotating reference frame with angular velocity ω whose axes are fixed to the body.

By following Euler's ideas, we have obtained equations describing the rotation of a rigid body. We have used Euler's method to obtain equations in the stationary reference frame. While obtaining equations in the body reference frame, we have followed Euler's method in the stationary reference frame and have not use direction cosines though Euler used them. Our derivation gives the same equations as equations in the body reference frame, known as "Euler equations". The equations in the stationary reference frame are more complicated than the ones in the body reference frame due to non-zero inertia products. In addition to this, it should be remembered that the moments of inertia change in the stationary reference frame during the rotation.

If we rotate a rigid body with an initial effect, it will rotate around one of its principal axes provided that there is not any disturbing effect. We have seen during the derivation that studying rotation of a rigid body by using principal axes provides simplifications since inertia products become zero. In formal definition, principal axes are defined by using inertia products: if we choose an axis passing through the center of mass and if inertia products related to that axis are zero then that axis is principal axes. As we mentioned this axis is physically exist and the rigid body naturally rotates around it, if an initial rotation is given. If we try to define the rotation of a rigid body by using axes other than principal axes, we need to consider inertia products. Moments of inertia and inertia products play a similar role to mass in translational motion, in which mass is related to inertia against the change in linear momentum. Moment of inertia corresponds to the inertia of the rigid body against the change in angular momentum around rotation axis, and inertia product is the inertia of the rigid body against the change in angular momentum around rotation axis due to effects of other rotations and asymmetries in mass distribution.

During the derivation, we considered that the fixed point is the center of mass of the rigid body. However, this derivation is also valid if the fixed point of the rigid body is different than the center of mass. In such cases, the origins should be set to that fixed point, and moments of inertia in the body reference frame should be calculated by considering rotations around the fixed point. This calculation can be done with the help of Steiner's parallel-axis theorem.

During the derivation, we considered an applied force. Any force that can change atomic interactions will deform the object, and it can not be considered as a rigid body anymore. Hence, the force that will be considered in Euler equations should not change the atomic structure. If it changes, other physical interactions should be considered, and such cases are not subject to rigid body rotations.

✍**Example:**

The Earth as an oblate ellipsoid is an object with $J_x = J_y \neq J_z$. Therefore, Euler's rotation equation reduces to (as their at no externe torque on the Earth in first approximation) after rearranging:

$$\begin{aligned}
J_x\dot{\omega}_x &= \omega_y\omega_z\left(J_x - J_z\right)\\
J_y\dot{\omega}_y &= -\omega_x\omega_z\left(J_x - J_z\right)\\
J_z\dot{\omega}_z &= 0
\end{aligned}\tag{9.3.450}$$

So, in this case, we see that ω_z, which is the spin about the symmetric axis, is a constant of motion. And we have using the first two relations by rearranging:

$$\dot{\omega}_1 = \Omega\omega_2 \quad \text{and} \quad \dot{\omega}_2 = -\Omega\omega_1 \tag{9.3.451}$$

where:

$$\Omega = \frac{\omega_z\left(J_x - J_z\right)}{J_x} \tag{9.3.452}$$

These equations are solved by:

$$(\omega_1, \omega_2) = \omega_0(\sin(\Omega t), \cos(\Omega t)) \tag{9.3.453}$$

for any constant ω_0. This means that, in the body frame, the direction of the spin is not constant: it precesses about the $\vec{e}_z$ axis with frequency Ω. The direction of the spin depends on the sign on Ω or, in other words, whether $J_x > J_z$ or $J_x < J_z$.

In an inertial frame, this precession of the spin looks like a wobble. For the Earth we have:

$$\frac{J_x - J_z}{J_x} \cong -\frac{1}{300} \tag{9.3.454}$$

Of course, we know the magnitude of the spin ω_z for the Earth: it is $\omega_3=(1\text{ day })^{-1}$. This information is enough to callculate the frequency of the Earth's wobble in absolute value:

$$\Omega_{\text{Earth}}=\frac{\omega_z\left(J_x-J_z\right)}{J_x}\cong\frac{1}{300} \tag{9.3.455}$$

This calculation was first performed by Euler in 11749 (holocene calendar) who predicted that the Earth completes a wobble every 300 days. Despite many searches, this effect wasn't detected until 11891 (holocene calendar) when Seth Carlo Chandler reanalysed the data and saw a wobble with a period of 427 days. It is now known as the "Chandler wobble". It is very small! The angular velocity ω intercepts the surface of the Earth approximately 10 metres from the North pole and precesses around it. More recent measurements place the frequency at 435 days, with the discrepancy between the predicted 300 days and observed 435 days due to the fact that the Earth is not a rigid body, but is flexible because of tidal effects. Less well understood is why these same tidal effects haven't caused the wobble to dampen and disappear completely. There are various hypothesis about what keeps the wobble alive, from earthquakes to fluctuating pressure at the bottom of the ocean.

9.3.8.2.3 Gyroscope

Ok now that we know quite well what is the angular momentum, the moment of inertia and the Lagrangian (see section Analytical Mechanics page 3155) we can now study one of the most fascinating object of macroscopic physics that combines all these concepts!!!

Definition (#9.3.443): A solid (solid of revolution for simplicity...) which can move freely around a fixed point and quickly turning on itself by definition is a "gyroscope".

Out of their recreational use... because they allow configurations considered as pedagogically exceptional... gyroscopes are an important part of inertial navigation systems (before the advent of GPS...) in aviation, aerospace, marine (boat stabilization), movie/television (stabilization of cameras), weapons (ballistic missiles), tunnel mining and many more. The guide instruments through inertia of these systems consist of gyroscopes and accelerometers, which calculate at any time the exact speed and direction of the system in motion (depending on the movement of the gyroscope a voltage variation is caused). The collected signals are transmitted to a computer that records and then corrects any aberrations of the path.

Remark

Gyroscopes based on other operating principles also exist, such as the electronic, microchip-packaged MEMS gyroscopes found in consumer electronics devices, solid-state ring LASERs, fibre optic gyroscopes, and the extremely sensitive quantum gyroscope.

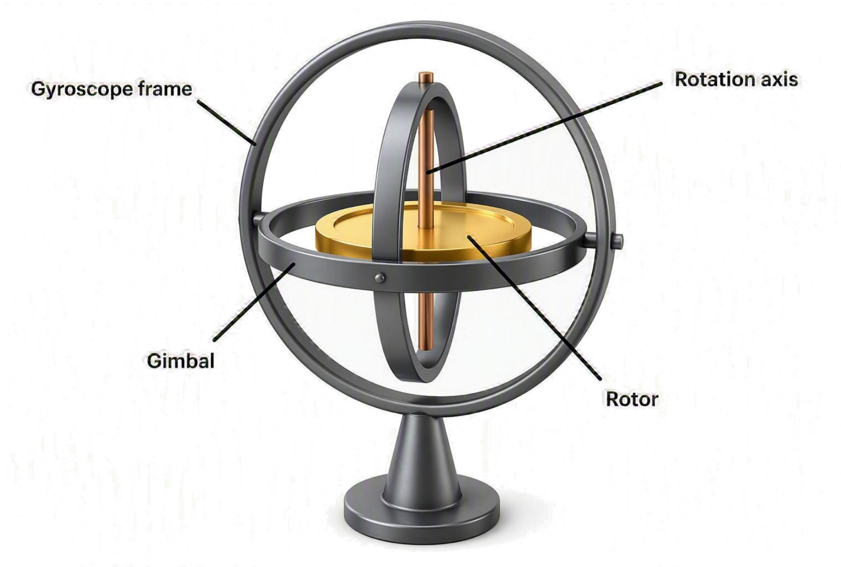


Figure 9.57: **3D Gyroscope**

The planets are another example of famous gyroscopes. The best known example being our Earth rotating around itself and being very massive it's angular momentum makes that its north pole is always (at least in the time scale of a human being...) facing the North star (actually approximately oriented to Polaris) regardless of its position in its orbit (this property of some gyroscopes is named "gyroscopic rigidity").

We will study the gyroscope in two steps:

1. The classical approach (without using Lagrangian mechanics) that is easy to understand but that gives only the possibility to explain the precession angular velocity Ω of the gyroscope.
2. The Lagrangian approach that is more abstract but that opens the door to the understanding of the precession AND the nutation rotation of the gyroscope.

9.3.8.2.3.1 Classical Approach with precession only

The following figure is an example of gyroscope known in laboratories schools named "symmetric weighing gyroscope". This is obviously a special and simplified case but for understanding the basic principle of the gyroscope:

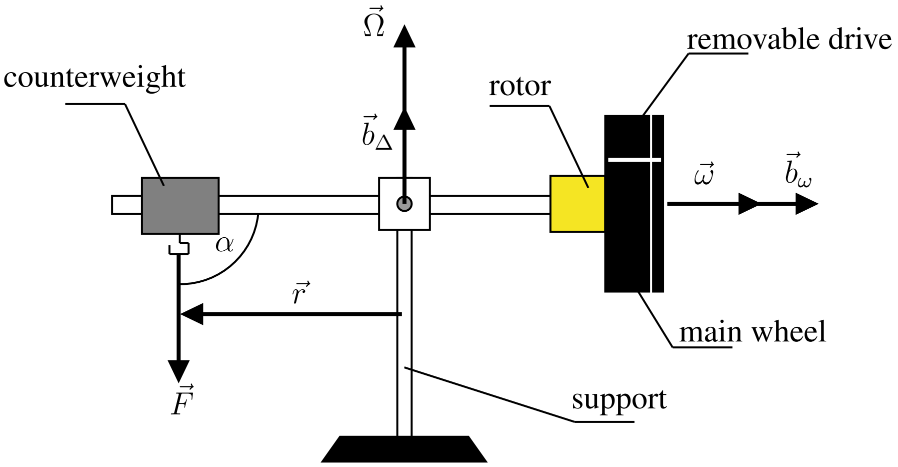


Figure 9.58: Symmetric weighing gyroscope

Consisting of an electric motor whose rotor, the main wheel, form the main mass in rapid angular rotation (a gyroscope of a V2 rocket turn at over $10,000$ rpm). The motor stator is fixed to a rod on which is positioned a counterweight on the opposite side. The assembly is placed on a support on the end of which is a gimbal mounted on a horizontal bearing which allows the gyroscope orientation almost without limitations in all directions.

In this figure we have ω which is the instantaneous angular velocity of the removable disk of radius R, Ω is the rate of precession of the gyroscope (rotation around the support), F is the force of the additional mass m attached to the counterweight and that imbalance the gyroscope, r is the distance of the gyroscope to the counterweight and finally α is the tilt angle that takes the axis of the gyroscope when we do an imbalance by attaching an additional weight to the counterweight.

To begin the theoretical study of this system, remember that have proved above that the angular momentum for a solid having a moment of inertia J is expressed by the following relation:

$$\vec{b}_\omega = J\vec{\omega} \tag{9.3.456}$$

and we saw that all solid in rotation around any axis also has also a momentum that is then customary to write according to what we have seen above:

$$\vec{b}_\Delta = J_\Delta\vec{\Omega} \tag{9.3.457}$$

We also proved above that the rotor, like any rapid mass in rotation, when started or regulated then produced a moment of force given by:

$$\vec{M} = \frac{\mathrm{d}\vec{b}_\omega}{\mathrm{d}t} = J\frac{\mathrm{d}\vec{\omega}}{\mathrm{d}t} \tag{9.3.458}$$

which is vectorially collinear to $\vec{\omega}$ and therefore passes through the axis of symmetry of the gyroscope. As we already know, it is this latter relation that puts the best evidence that the gyroscope always keeps the same direction in space even when we move its support.

In other words a free gyroscope driven at a high speed of rotation has for fundamental property to maintain its axis of rotation in a fixed orientation with respect to absolute space. This is what we named the "first gyroscopic law" or "law of fixity".

Typically the "Foucault gyroscope" shown below, is an excellent practical example of the fixity law, it maintains its orientation regardless of how we handle the support on which it is build:

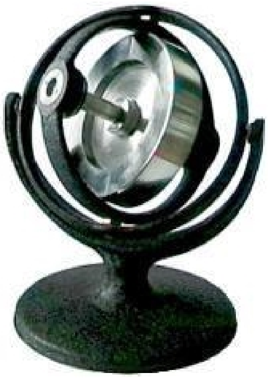


Figure 9.59: **Foucault gyroscope**

If we put the gyroscope a whole day on a table with an engine that maintains the rotation of the central massive disk constant , then we observe the rotation of the Earth because the gyroscope then rotates very slowly on itself in 24 hours!

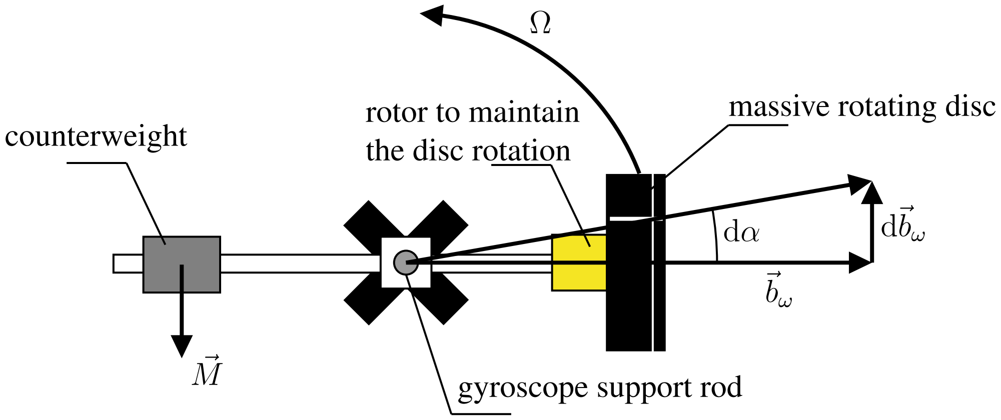
To return to our mathematical considerations ... now let's look at the force momentum of the counterweight that imbalance our symmetrical gyroscope while the disc is rotating, and generates a general rotation of the gyroscope as experience reveals. We then have for the force momentum making rotating the gyroscope about its axis (support rod):

$$M=\|\vec{M}\|=\|\vec{r}\times\vec{F}\|=\|\vec{r}\|\cdot\|\vec{F}\|\sin(\alpha)=rma\sin(\alpha) \qquad (9.3.459)$$

Since the gyroscope does not precess when the system is balanced this mean that the moment of force of the extra weight that imbalance the gyroscope generates an angular momentum according to the relation proved earlier above such that:

$$\frac{\mathrm{d}b_{\omega}}{\mathrm{d}t}=M=rma\sin(\alpha) \qquad (9.3.460)$$

What schematically can be represented as follows (this is our gyroscope seen from above):


Ω
rotor to maintain
the disc rotation
massive rotating disc
counterweight
$\mathrm{d}\alpha$
$\mathrm{d}\vec{b}_\omega$
$\vec{b}_\omega$
$\vec{M}$
gyroscope support rod

Or the real life equivalent (that should help to reader to better imagine what represents the above figure...):

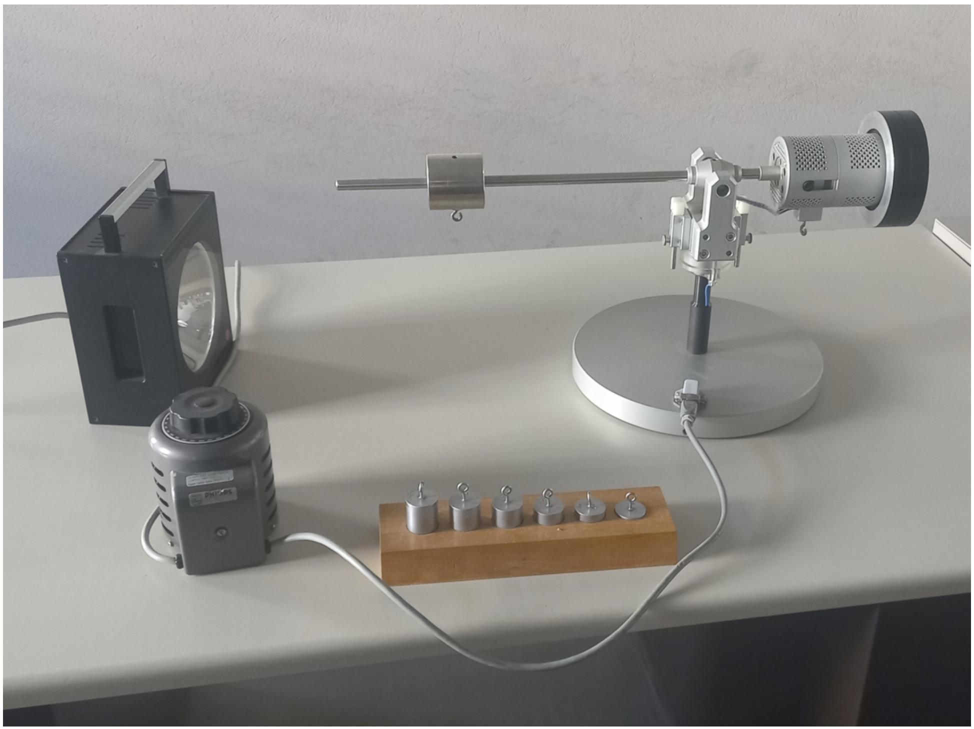


Figure 9.60: Weighted gyroscope with rotor from the physics laboratory of the HES Geneva

We then have, as soon as the gyroscope starts to rotate (in a circular motion):

$$\tan(\mathrm{d}\alpha)=\frac{||\mathrm{d}\vec{b}_\omega||}{\vec{b}_\omega}\cong\frac{\mathrm{d}b_\omega}{b_\omega} \qquad (9.3.461)$$

Taking the Taylor approximation (see section Sequences and Series page 1696) to the first order of the tangent for small angles we get:

$$\mathrm{d}\alpha\cong\frac{\mathrm{d}b_\omega}{b_\omega} \qquad (9.3.462)$$

Let us assume, to simplify the study of the problem, that the variation of total angular momentum with respect to the rotation axis of the gyroscope (that is to say the support rod!) can be assimilated to the angular momentum of the rotor only if that latter is turning fast enough and that its mass is large enough. Which means:

$$\frac{\mathrm{d}\vec{b}_\omega}{\mathrm{d}t} \gg \frac{\mathrm{d}\vec{b}_\Delta}{\mathrm{d}t} \tag{9.3.463}$$

Then we have:

$$\frac{\mathrm{d}\vec{b}}{\mathrm{d}t} = \frac{\mathrm{d}(\vec{b}_\Delta + \vec{b}_\omega)}{\mathrm{d}t} \cong \frac{\mathrm{d}\vec{b}_\omega}{\mathrm{d}t} \tag{9.3.464}$$

and therefore because of this approximation by entire force momentum is assigned to the change in angular momentum of the rotor itself:

$$\mathrm{d}\alpha \cong \frac{M\mathrm{d}t}{b_\omega} \tag{9.3.465}$$

It finally comes for the precession angular velocity:

$$\boxed{\Omega \cong \frac{\mathrm{d}\alpha}{\mathrm{d}t} = \frac{M}{b_\omega} = \frac{rF\sin(\alpha)}{J\omega}} \tag{9.3.466}$$

Therefore when the symmetric weighted gyroscope is balanced (when M is zero in the numerator of the fraction or if you prefer: $\sin(\alpha) = 0$), its angular momentum therefore maintains a fixed orientation regardless of the value of the denominator since its precession Ω will be always equal to zero.

Definition (#9.3.444): The rotational movement resulting from a non-equilibrated gyroscope is named "precession movement" when it is caused intentionally, and "drift" when it is due to a disruptive element.

Let us finally indicate the playful gyroscopes for small children as the spinner below:

that we can roughly represent as following (side view and respectively top view) to make still some classical mathematical analysis (do the test with your children to see if it interests them as much as the toy...):

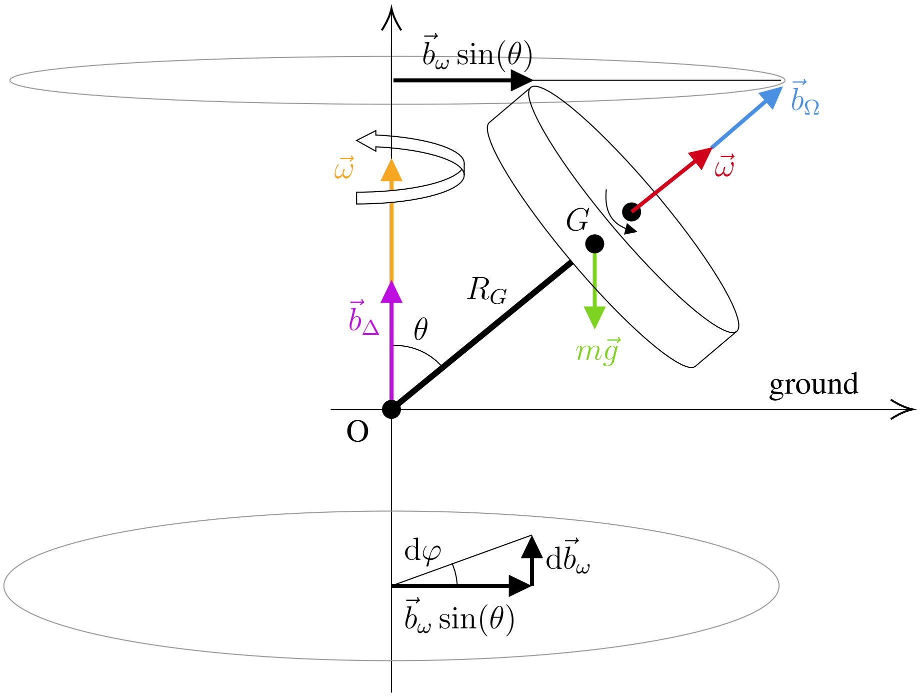


Figure 9.61: **Gyroscope spinner**

where we make the hypothesis that the end of the axis of the spinning top (other name for "gyroscope spinner") is placed on the ground without possibility of slippage and that it has a constant angular velocity ω and that the latter is big enough to not have its angle θ which varies in time (otherwise another effect appear named "nutation" that we can only introduce robustly using Lagrangian Mechanics).

Using the same technique as for the symmetric weighted gyroscope we have (well... we could have also used simply the relation $L=\alpha R$ proved in the section Trigonometry...):

$$\mathrm{d}b_\omega = b_\omega \sin(\theta)\mathrm{d}\varphi \tag{9.3.467}$$

We also have for the moment of force (torque):

$$M = R_G mg \sin(\theta) \tag{9.3.468}$$

But again, the angular momentum changes! Indeed, so we have in this particular case:

$$\mathrm{d}b_\omega = b_\omega \sin(\theta)\mathrm{d}\varphi \tag{9.3.469}$$

it follows that under the same assumptions as the symmetric weighted gyroscope that:

$$\frac{\mathrm{d}b}{\mathrm{d}t} \cong \frac{\mathrm{d}b_\omega}{\mathrm{d}t} = b_\omega \sin(\theta)\frac{\mathrm{d}\varphi}{\mathrm{d}t} = b_\omega \sin(\theta)\Omega \tag{9.3.470}$$

hence in vector form:

$$\frac{\mathrm{d}\vec{b}_\omega}{\mathrm{d}t} = \vec{\Omega} \times \vec{b}_\omega \tag{9.3.471}$$

and we know that this last relation (proved made earlier above) can be completed by writing:

$$\frac{\mathrm{d}\vec{b}_\omega}{\mathrm{d}t} = \vec{\Omega} \times \vec{b}_\omega = \vec{M} \tag{9.3.472}$$

It comes then:

$$\Omega b_\omega \sin(\theta) = R_G mg \sin(\theta) \tag{9.3.473}$$

Hence:

$$\boxed{\Omega = \frac{R_G mg}{b_\omega} = \frac{R_G mg}{J\omega}} \tag{9.3.474}$$

We see that the difference with the symmetric weighted gyroscope is that the speed of precession is then independent of the angle!

Remarks

R1. A cyclist travelling in a straight line is stabilized (fixity law requirement!) by the angular momentum of its wheels perpendicular to the rolling direction (the wheels behaves as a gyroscope).

R2. Probably without realizing it, we tip up while biking in a curve to produce a precession in the wheels and turn more easily. Indeed the precession movement rotates the bicycle wheel in the direction you look without the need to turn the handlebars... !

9.3.8.2.3.2 Lagrangian Approach with precession and nutation

Symmetric spinning top is a case studied by Lagrange in 11788 (holocene calendar). After 110 years, in 11897-11898 (holocene calendar) Felix Klein and Arnold Sommerfeld published two volumes on the topic, whereas in 11897 (holocene calendar) Klein separately published a shorter monograph on the mathematical theory of the top. In the eastern world literature, one of the oldest papers is probably due to Germanovich Appel'rot (11894 according to holocene calendar). Explicit integration of motion equations, to give the nutation in terms of elliptic integrals was given at the end of 119th century (holocene calendar).

The analysis of the symmetrical top (spinning top) is a fascinating topic in Classical Mechanics. It is a simple system that exhibits counter-intuitive balancing behaviour. It's famous "gravity-defying" behaviour to remain perched on its pointed is well known (a simple internet search reveals many references concerning the spinning tops and gyroscopes as if they were "magical" instruments that defy gravity). The following developments first explores the Lagrangian formulation of the symmetrical top in a uniform force-field (e.g. gravity). After establishing a few constraints, a dynamic model is presented. From this model, precession rates and nutation behaviour are deduced!

Figure 9.62: **Symmetrical top gyroscope in equilibrium**

The geometry of the symmetric gyroscopic top in a gravitational field consists of a "heavy" wheel on a narrow stem whose bottom point is fixed to the origin, but is free to rotate (as show in the figure below):

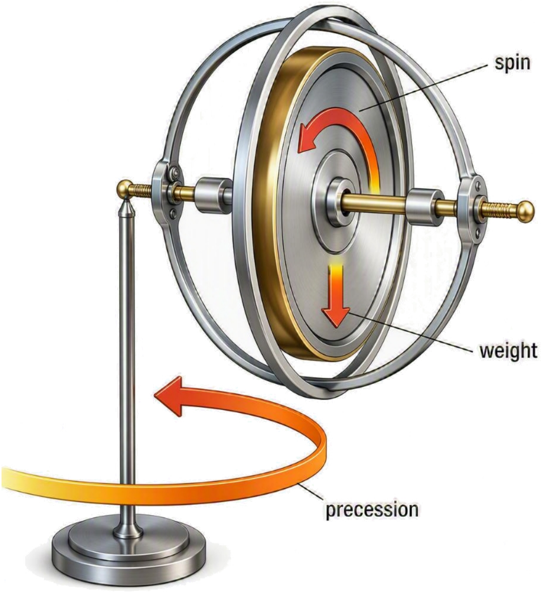


Figure 9.63: **Spinning top in precession only**

The z' axis passes through the center of rotation of the top. The motion of the top can be expressed in terms of the so-named Eulerian angles (ϕ,θ,ψ). The variable θ is the angle that the z' axis makes with the stationary z axis. The variable ϕ is the angle that is formed by the projection of the z' axis in the $x-y$ plane and the x axis. Counter-clockwise rotation yields positive angles. The variable ψ is the angular position around the z' axis. Since our top is circularly symmetric, the choice of origin for is arbitrary. We will only be concerned with the rate of angular rotation of the wheel about z', given by $\dot{\psi}$ ($\omega=\dot{\psi}$ on the above figure). Other dynamical variables of interest are the time derivatives of ϕ and θ. These variables $\dot{\phi}$ and $\dot{\theta}$ describe the precession and nutation of the top. The length of the stem is R_G. We assume the mass of the stem is negligible with respect to that of the wheel m. The acceleration due to uniform gravity is g and is assumed to point downwards along the z axis:

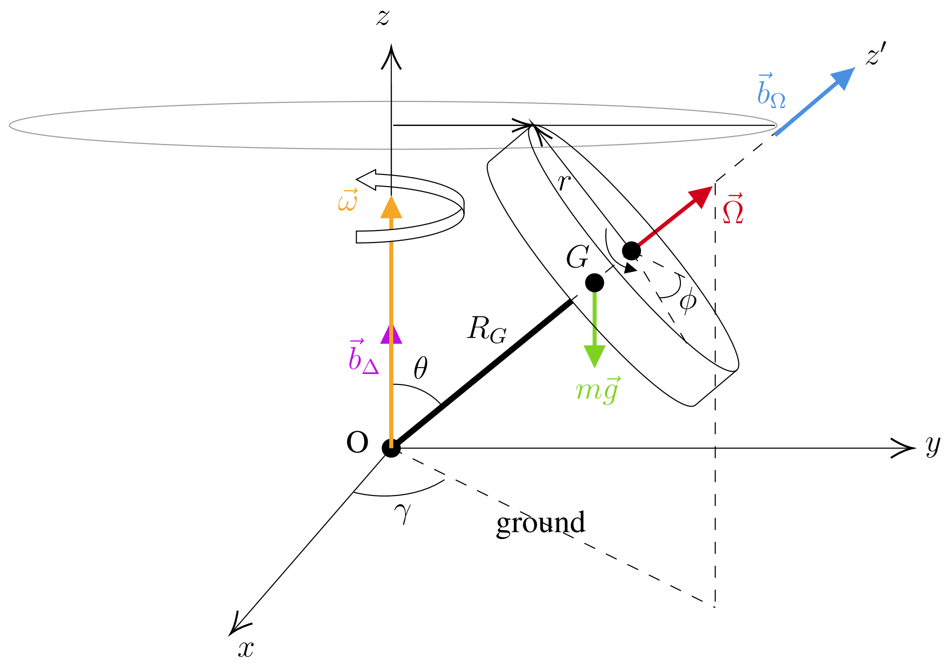


Figure 9.64: **Schematic view of spinning top**

The most convenient way to derive the equations of motion for this system is to use the Lagrangian of Hamiltonian formalism as we have already mentioned it (see section Analytical Mechanics page 3155).

First let us recall that in the section of Euclidean Geometry we have proved the following group of relations:

$$\begin{aligned}\Omega_1&=\sin(\theta)\sin(\gamma)\dot{\phi}+\cos(\gamma)\dot{\theta}\\ \Omega_2&=\sin(\theta)\cos(\gamma)\dot{\phi}-\sin(\gamma)\dot{\theta}\\ \Omega_3&=\cos(\theta)\dot{\phi}+\dot{\gamma}\end{aligned} \tag{9.3.475}$$

named "kinematic Euler equations" and is useful to express the rotation speed (around any axis of rotation) of a body in its own reference frame (most of time its center of mass).

Let us choose O as origin for our reference frame and we will choose the principal axes of inertia as the body reference frame. The total kinetic energy of the object will then be:

$$E_{\text{tot}}=E_c+E_{\text{rot.}}+E_p=\frac{1}{2}m\vec{v}_G^2+\sum_{i=1}^{n}\frac{1}{2}J_i\Omega_i^2+mgR_G\cos(\theta) \tag{9.3.476}$$

As we take $\overrightarrow{\text{O}\Omega}$ as instantaneous axis of rotation relatively to an observer in O we have $v_G=0$.

If for some readers it is not obvious that:

$$E_p=mgR_G\cos(\theta) \tag{9.3.477}$$

just remember first that $E_p=mgh$ but as our comoving reference frame is oriented along $\overrightarrow{\text{O}\Omega}$ we must project the gravitational force on this axis. As you can see with the extreme trivial situations where $\theta=0$ (vertical spinning top) we have as expected:

$$E_p=mgR_G\cos(0)=mgR_G \tag{9.3.478}$$

and for $\theta = \pi/2$ (spinning top rotation axis in the plane of the ground!) then:

$$E_p = mgR_G \cos\left(\frac{\pi}{2}\right) = 0 \tag{9.3.479}$$

Then it remains:

$$E_{\text{tot}} = \sum_{i=1}^{n} \frac{1}{2} J_i \Omega_i^2 + mgR_G \cos(\theta) = \frac{1}{2} J_1 \Omega_1^2 + \frac{1}{2} J_2 \Omega_2^2 + \frac{1}{2} J_3 \Omega_3^2 + mgR_G \cos(\theta) \tag{9.3.480}$$

where the index 123 corresponds to the notation of the comoving reference frame as seen in the section of Euclidean Geometry during our study of the Euler angles.

For J_1, J_2, J_3 following the proof we made for the inertial moment of a rotating cylinder around its main symmetric axis in the section Geometric Shapes (see page 2959) and using also the parallel axis theorem we have by symmetry of our problem:

$$\begin{aligned} J_1 &= J_\theta = \frac{1}{2} mr^2 + mR_G^2 \\ J_2 &= J_\gamma = \frac{1}{2} mr^2 + mR_G^2 \\ J_3 &= J_\phi = \frac{1}{2} mr^2 \end{aligned} \tag{9.3.481}$$

But for the following developments we will not use the explicit relations above. In this way the reader can choose the spinning top geometry he wants and after just put the $J_1 = J_2$, J_3 (so we not consider an asymmetric spinning top otherwise the general formula is rather complicated) corresponding values in the final formula!

For what will follow we will put $I_\phi := J_3$ and $I_0 := J_1 = J_2$. Therefore:

$$\begin{aligned} E_{\text{tot}} &= \frac{1}{2} I_0 \left(\Omega_2^2 + \Omega_3^2\right) + \frac{1}{2} I_\phi \Omega_1^2 + mgR_G \cos(\theta) \\ &= \frac{1}{2} I_0 \left((\sin(\theta)\sin(\gamma)\dot{\phi} + \cos(\gamma)\dot{\theta})^2 + (\sin(\theta)\cos(\gamma)\dot{\phi} - \sin(\gamma)\dot{\theta})^2 \right) + \frac{1}{2} I_\phi (\cos(\theta)\dot{\phi} + \dot{\gamma})^2 \\ &\quad + mgR_G \cos(\theta) \\ &= \frac{1}{2} I_0 \left(\sin^2(\theta)\dot{\phi}^2 + \dot{\theta}^2 \right) + \frac{1}{2} I_\phi (\cos(\theta)\dot{\phi} + \dot{\gamma})^2 + mgR_G \cos(\theta) \end{aligned} \tag{9.3.482}$$

The Lagrangian (see section Analytical Mechanics page 3155) of the spinning top is deceptively simple as given then by:

$$L = \frac{1}{2} I_0 \left(\sin^2(\theta)\dot{\phi}^2 + \dot{\theta}^2 \right) + \frac{1}{2} I_\phi (\cos(\theta)\dot{\phi} + \dot{\gamma})^2 - mgR_G \cos(\theta) \tag{9.3.483}$$

We see that the Lagrangian does not depend explicitly on the variables ψ, ϕ or t. As a consequence, the angular momenta p_ϕ and p_ψ and the total energy E are invariant (i.e. conserved).

Let us recall that in the section of Analytical Mechanics we have proved that for a non-conservative force (that is not derived from a potential) we have:

$$\frac{\partial L(q_i,\dot{q}_i,t)}{\partial q_i}-\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L(q_i,\dot{q}_i,t)}{\partial \dot{q}_i}=F_i \tag{9.3.484}$$

Or equivalently:

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L(q_i,\dot{q}_i,t)}{\partial \dot{q}_i}-\frac{\partial L(q_i,\dot{q}_i,t)}{\partial q_i}=F_i \tag{9.3.485}$$

Therefore to simplify the development we consider the potential gravitational force as a non-conservative force (this doesn't change the result anyway but its more easy to introduce it in this way to students or most readers):

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{\phi}}-\frac{\partial L}{\partial \phi}&=0\\ \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{\theta}}-\frac{\partial L}{\partial \theta}&=mgR_G\cos(\theta)\\ \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{\gamma}}-\frac{\partial L}{\partial \gamma}&=0\end{aligned} \tag{9.3.486}$$

And looking to our Lagrangian we see that it simplifies automatically to:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{\phi}}&=0\\ \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{\theta}}-\frac{\partial L}{\partial \theta}&=mgR_G\cos(\theta)\\ \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{\gamma}}&=0\end{aligned} \tag{9.3.487}$$

So let us focus on what seem the easiest to handle, that is:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{\phi}}&=0\\ \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{\gamma}}&=0\end{aligned} \tag{9.3.488}$$

It follows immediately:

$$\begin{aligned}\frac{\partial L}{\partial \dot{\phi}}&=c_1^{te}\\ \frac{\partial L}{\partial \dot{\gamma}}&=c_2^{te}\end{aligned} \tag{9.3.489}$$

But as we have seen in the section of Analytical Mechanics:

$$\begin{aligned}\frac{\partial L}{\partial \dot{\phi}}&=p_\phi=c_1^{te}\\ \frac{\partial L}{\partial \dot{\gamma}}&=p_\gamma=c_2^{te}\end{aligned} \tag{9.3.490}$$

where p_ϕ and p_γ are the canonical momentum.

Thus explicitly:

$$\begin{aligned}\frac{\partial L}{\partial \dot{\phi}}&=p_\phi=I_0\sin^2(\theta)\dot{\phi}+I_\phi(\cos(\theta)\dot{\phi}+\dot{\gamma})\cos(\theta)=c_1^{te}\\ \frac{\partial L}{\partial \dot{\gamma}}&=p_\gamma=I_\phi(\cos(\theta)\dot{\phi}+\dot{\gamma})=c_2^{te}\end{aligned} \tag{9.3.491}$$

Something important to notice is that since:

$$p_\gamma=I_\phi(\cos(\theta)\dot{\phi}+\dot{\gamma})=c_2^{te} \tag{9.3.492}$$

and as we have obviously $I_\phi=c^{te}$ we get immediately:

$$\cos(\theta)\dot{\phi}+\dot{\gamma}=c_2^{te} \tag{9.3.493}$$

And it is normally obvious that this corresponds therefore to the rotation speed Ω of the spinning top on itself such that:

$$p_\gamma=I_\phi(\cos(\theta)\dot{\phi}+\dot{\gamma})=I_\phi\Omega=c_2^{te} \tag{9.3.494}$$

The values p_ϕ,p_γ are therefore constants of motions and, as we will see it below, are therefore indispensable quantities for deriving other aspect of the top's motion such as precession and nutation.

Now we consider that the reader is familiar with the gyroscope's slow circling behaviour as int apparently defies gravity in its odd leaning without falling over. This is as we know "precession" trough the azimuthal angle γ. For the moment we will ignore nutation $\dot{\theta}=0$ and focus on the leaning angle θ, assumed constant, and the rate of precession $\dot{\gamma}$.

We shall use the energy method to derive the precession rate. This entails finding a point where the total system energy is stationary with respect to the angle θ, i.e.:

$$\frac{\partial E_{\text{tot}}}{\partial \theta}=0 \tag{9.3.495}$$

We can do this because E_{tot} is a conserved quantity set by the initial conditions. The energy must be a "stationary" or extreme value equal to the total system energy. Any deviation in θ from the correct value represents an error in the solution that pushes away from stationary value. This provides us with a handy mathematical trick for finding important physical properties of the system.

Since we ignore for now nutation (only steady precession is assumed), we eliminate the variables $\dot{\phi}$ and $\dot{\gamma}$ in E_{tot} by putting E_{tot} in term of the invariants p_ϕ and p_γ. We get (don't forget that for now we put $\dot{\theta}=0$!):

$$\begin{aligned}E_{\text{tot}}&=\frac{1}{2}I_0\left(\sin^2(\theta)\dot{\phi}^2+\dot{\theta}^2\right)+\frac{1}{2}I_\phi(\cos(\theta)\dot{\phi}+\dot{\gamma})^2+mgR_G\cos(\theta)\\&=\frac{1}{2}I_0\sin^2(\theta)\dot{\phi}^2+\frac{1}{2}I_\phi(\cos(\theta)\dot{\phi}+\dot{\gamma})^2+mgR_G\cos(\theta)\\&=\frac{(p_\phi-p_\gamma\cos(\theta))^2}{2I_0\sin^2(\theta)}+\frac{p_\gamma^2}{2I_\phi}+mgR_G\cos(\theta)\end{aligned}\tag{9.3.496}$$

Indeed as for the first term, this is not obvious! Let us check it:

$$\begin{aligned}\frac{(p_\phi-p_\gamma\cos(\theta))^2}{2I_0\sin^2(\theta)}&=\frac{((I_0\sin^2(\theta)\dot{\phi}+I_\phi(\cos(\theta)\dot{\phi}+\dot{\gamma})\cos(\theta))-(I_\phi(\cos(\theta)\dot{\phi}+\dot{\gamma}))\cos(\theta))^2}{2I_0\sin^2(\theta)}\\&=\frac{(I_0\sin^2(\theta)\dot{\phi})^2}{2I_0\sin^2(\theta)}=\frac{I_0^2\sin^4(\theta)\dot{\phi}^2}{2I_0\sin^2(\theta)}=\frac{1}{2}I_0\sin^2(\theta)\dot{\phi}^2\end{aligned}\tag{9.3.497}$$

Taking the derivative of E_{tot} with respect to the elevation angle θ yields:

$$\begin{aligned}\frac{\partial E_{\text{tot}}}{\partial\theta}&=\frac{\partial}{\partial\theta}\left(\frac{(p_\phi-p_\gamma\cos(\theta))^2}{2I_0\sin^2(\theta)}+\frac{p_\gamma^2}{2I_\phi}+mgR_G\cos(\theta)\right)=\frac{\partial}{\partial\theta}\left(\frac{(p_\phi-p_\gamma\cos(\theta))^2}{2I_0\sin^2(\theta)}+mgR_G\cos(\theta)\right)\\&=\frac{\partial}{\partial\theta}\left(\frac{(p_\phi-p_\gamma\cos(\theta))^2}{2I_0\sin^2(\theta)}\right)-mgR_G\sin(\theta)\\&=\frac{2(p_\phi-p_\gamma\cos(\theta))p_\gamma\sin(\theta)}{2I_0\sin^2(\theta)}+\frac{(p_\phi-p_\gamma\cos(\theta))^2}{2I_0\sin^3(\theta)}(-2\cos(\theta))-mgR_G\sin(\theta)\\&=\frac{(p_\phi-p_\gamma\cos(\theta))p_\gamma\sin(\theta)}{I_0\sin^2(\theta)}-\frac{\cos(\theta)(p_\phi-p_\gamma\cos(\theta))^2}{I_0\sin^3(\theta)}-mgR_G\sin(\theta)\\&=\frac{(p_\phi-p_\gamma\cos(\theta))p_\gamma\sin^2(\theta)-\sin(\theta)\cos(\theta)(p_\phi-p_\gamma\cos(\theta))^2}{I_0\sin^4(\theta)}-mgR_G\sin(\theta)=0\end{aligned}\tag{9.3.498}$$

The solution of the above equation for θ yields the angle at which the gyroscope will lean given the speed of rotation of the wheel. The angle θ is found obviously by solving it for the roots in terms of the invariant quantities.

The precession rate is found by reorganizing:

$$p_\phi=I_0\sin^2(\theta)\dot{\phi}+I_\phi(\cos(\theta)\dot{\phi}+\dot{\gamma})\cos(\theta)\tag{9.3.499}$$

Therefore:

$$p_\phi=I_0\sin^2(\theta)\dot{\phi}+p_\gamma\cos(\theta)\tag{9.3.500}$$

Hence:

$$\dot{\phi}=\frac{p_\phi-p_\gamma\cos(\theta)}{I_0\sin^2(\theta)}\tag{9.3.501}$$

So now we can rewrite:

$$\frac{(p_\phi - p_\gamma\cos(\theta))p_\gamma\sin^2(\theta) - \sin(\theta)\cos(\theta)(p_\phi - p_\gamma\cos(\theta))^2}{I_0\sin^4(\theta)} - mgR_G\sin(\theta) = 0 \tag{9.3.502}$$

We multiply the both sides by:

$$-\frac{I_0\sin(\theta)}{p_\gamma} \tag{9.3.503}$$

So we get:

$$-\frac{(p_\phi - p_\gamma\cos(\theta))p_\gamma\sin^2(\theta) - \sin(\theta)\cos(\theta)(p_\phi - p_\gamma\cos(\theta))^2}{p_\gamma\sin^3(\theta)} + \frac{mgR_GI_0\sin^2(\theta)}{p_\gamma} = 0 \tag{9.3.504}$$

After simplification and rearranging we get the following polynomial equation in $(p_\phi - p_\gamma\cos(\theta))$ to solve:

$$\frac{\cos(\theta)}{p_\gamma\sin^2(\theta)}(p_\phi - p_\gamma\cos(\theta))^2 - (p_\phi - p_\gamma\cos(\theta)) + \frac{mgR_GI_0\sin^2(\theta)}{p_\gamma} = 0 \tag{9.3.505}$$

Using the quadratic formula to solve for the roots of this equation, we have (see section Calculus page 1383):

$$p_\phi - p_\gamma\cos(\theta) = \frac{1 \pm \sqrt{1 - 4\dfrac{\cos(\theta)}{p_\gamma\sin^2(\theta)}\dfrac{mgR_GI_0\sin^2(\theta)}{p_\gamma}}}{2\dfrac{\cos(\theta)}{p_\gamma\sin^2(\theta)}} = \frac{p_\gamma\sin^2(\theta)}{2\cos(\theta)}\left(1 \pm \sqrt{1 - \frac{4mgR_GI_0\cos(\theta)}{p_\gamma^2}}\right) \tag{9.3.506}$$

Using now:

$$\dot{\phi} = \frac{p_\phi - p_\gamma\cos(\theta)}{I_0\sin^2(\theta)} \tag{9.3.507}$$

we get:

$$\dot{\phi} = \frac{p_\gamma}{2I_0\cos(\theta)}\left(1 \pm \sqrt{1 - \frac{4mgR_GI_0\cos(\theta)}{p_\gamma^2}}\right) \tag{9.3.508}$$

Notice that if:

$$|p_\gamma| < \sqrt{4mgR_GI_0\cos(\theta)} \tag{9.3.509}$$

that is to say:

$$\Omega < \frac{\sqrt{4mgR_GI_0\cos(\theta)}}{I_\phi} \tag{9.3.510}$$

there is no well defined uniform precession.

The quadratic equation produces two solutions: the so-named "fast precession" and the "slow precession" solutions.

If the rotation speed of the heavy wheel is large (i.e. $p_\gamma \gg \sqrt{4mgR_G\cos(\theta)}$, the fast precession rate is:

$$\dot{\phi}_{\text{fast}} \cong \frac{p_\gamma}{2I_0\cos(\theta)} = \frac{I_\phi \Omega}{I_0\cos(\theta)} \quad (9.3.511)$$

Or if we take again the notations of our first approach with the symmetric weighing gyroscope:

$$\boxed{\Omega_{\text{fast}} \cong \frac{J\omega}{I_0\cos(\theta)}} \quad (9.3.512)$$

So we see that the gravitational force and the weight does not appear anymore so we are far from the result we get previously with the "classic approach".

The slow precession rate is found by using the small argument approximation for $\sqrt{1-x}$ as proved in the section of Sequences and Series (Maclaurin series page 1696):

$$\sqrt{1-x} \cong 1 - \frac{1}{2}x \quad (9.3.513)$$

Hence we find that:

$$\dot{\phi}_{\text{slow}} \cong \frac{mgR_G}{p_\gamma} = \frac{mgR_G}{I_\phi \Omega} \quad (9.3.514)$$

Or if we take again the notations of our first approach with the symmetric weighing gyroscope:

$$\boxed{\Omega_{\text{slow}} \cong \frac{mgR_G}{J\omega}} \quad (9.3.515)$$

So we fall back on exactly the same result!!!!

Once again it is interesting that the slow precession rate does not depend on the "leaning angle θ". The gyroscope will precess at a rate solely based on the wheel speed, stem length and the gravitational pull[15] on the center of mass of the gyroscope. The slow precession rate is the one most commonly observed experimentally as we already know.

Now that we have an expression for the precession rate, by using again:

$$\dot{\phi} = \frac{p_\phi - p_\gamma\cos(\theta)}{I_0\sin^2(\theta)} \quad (9.3.516)$$

That is:

$$p_\phi - p_\gamma\cos(\theta) - \dot{\phi}I_0\sin^2(\theta) = 0 \quad (9.3.517)$$

Using trigonometric identities:

$$p_\phi - p_\gamma\cos(\theta) - \dot{\phi}I_0(1-\cos^2(\theta)) = 0 \quad (9.3.518)$$

---

[15] Keep in mind that's in the framework of the Newtonian model of gravity as we that in General Relativity there are no "pull".

After rearranging:

$$\dot{\phi}I_0\cos^2(\theta)-p_\gamma\cos(\theta)+(p_\phi-\dot{\phi}I_0)=0 \tag{9.3.519}$$

Or:

$$-\dot{\phi}I_0\cos^2(\theta)+p_\gamma\cos(\theta)+(\dot{\phi}I_0-p_\phi)=0 \tag{9.3.520}$$

Using the "slow" precession rate, i.e. $\dot{\phi}=mgR_G/p_\gamma$ and the quadratic equation to solve for $\cos(\theta)$, we see that:

$$\begin{aligned}\cos(\theta)&=\frac{-p_\gamma\pm\sqrt{p_\gamma^2+4I_0\dot{\phi}\left(I_0\dot{\phi}-p_\phi\right)}}{-2I_0\dot{\phi}}=\frac{-p_\gamma\pm\sqrt{p_\gamma^2+4I_0\dfrac{mgR_G}{p_\gamma}\left(I_0\dfrac{mgR_G}{p_\gamma}-p_\phi\right)}}{-2I_0\dfrac{mgR_G}{p_\gamma}}\\&=\frac{p_\gamma}{2I_0mgR_G}\left(p_\gamma\pm\sqrt{p_\gamma^2+4I_0\frac{mgR_G}{p_\gamma}\left(I_0\frac{mgR_G}{p_\gamma}-p_\phi\right)}\right)\\&=\frac{p_\gamma^2}{2I_0mgR_G}\left(1\pm\sqrt{1+4I_0\frac{mgR_G}{p_\gamma^3}\left(I_0\frac{mgR_G}{p_\gamma}-p_\phi\right)}\right)\\&=\frac{p_\gamma^2}{2I_0mgR_G}\left(1\pm\sqrt{1+4I_0\frac{mgR_G}{p_\gamma^2}\left(\frac{I_0mgR_G/p_\gamma-p_\phi}{p_\gamma}\right)}\right)\end{aligned} \tag{9.3.521}$$

Only the solution with the minus sign gives us a meaningful solution, because $\cos(\theta)$ can only take values between -1 and 1. Therefore:

$$\cos(\theta)=\frac{p_\gamma^2}{2I_0mgR_G}\left(1-\sqrt{1+4I_0\frac{mgR_G}{p_\gamma^2}\left(\frac{I_0mgR_G/p_\gamma-p_\phi}{p_\gamma}\right)}\right) \tag{9.3.522}$$

Moreover, if $p_\gamma^2\gg I_0mgR_G$,we can use the small argument approximation for the square-root, which gives us:

$$\begin{aligned}&\sqrt{1-\left(-4I_0\frac{mgR_G}{p_\gamma}\right)\left(\frac{I_0mgR_G/p_\gamma-p_\phi}{p_\gamma^2}\right)}\cong\sqrt{1-\left(-4I_0\frac{mgR_G}{p_\gamma^2}\right)\left(-\frac{p_\phi}{p_\gamma}\right)}\\&=\sqrt{1-4I_0\frac{mgR_Gp_\phi}{p_\gamma^3}}\cong 1-\frac{1}{2}4I_0\frac{mgR_Gp_\phi}{p_\gamma^3}=1-2I_0\frac{mgR_Gp_\phi}{p_\gamma^3}\end{aligned} \tag{9.3.523}$$

Therefore:

$$\cos(\theta)=\frac{p_\gamma^2}{2I_0mgR_G}\left(2I_0\frac{mgR_Gp_\phi}{p_\gamma^3}\right)=\frac{p_\phi}{p_\gamma} \tag{9.3.524}$$

Hence:

$$\boxed{\theta_{\text{lean}}\cong\arccos\left(\frac{p_\phi}{p_\gamma}\right)} \tag{9.3.525}$$

Now for the nutation who's typical movement is represented in the figure below:

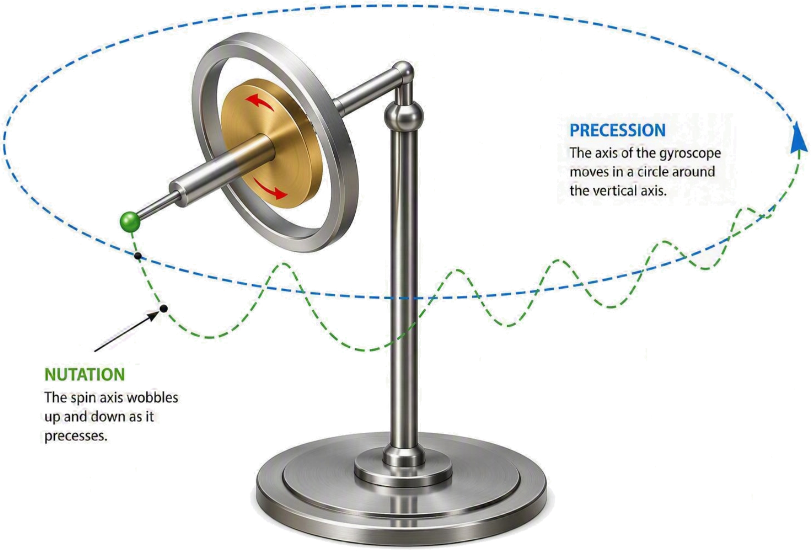

PRECESSION
The axis of the gyroscope moves in a circle around the vertical axis.
NUTATION
The spin axis wobbles up and down as it precesses.

Figure 9.65: **Spinning top with precession AND nutation movement**

don't forget that we have determined for $\dot{\theta}$:

$$E_{\text{tot}}=\frac{(p_\phi-p_\gamma\cos(\theta))^2}{2I_0\sin^2(\theta)}+\frac{p_\gamma^2}{2I_\phi}+mgR_G\cos(\theta) \tag{9.3.526}$$

Re-including the kinetic energy of θ-directed motion (nutation), we have:

$$E_{\text{tot}}=\frac{(p_\phi-p_\gamma\cos(\theta))^2}{2I_0\sin^2(\theta)}+\frac{p_\gamma^2}{2I_\phi}+\frac{1}{2}I_0\dot{\theta}^2+mgR_G\cos(\theta) \tag{9.3.527}$$

A "potential energy" function that provides a "restoring force" can be taken from the above relation as:

$$V(\theta)=\frac{(p_\phi-p_\gamma\cos(\theta))^2}{2I_0\sin^2(\theta)}+\frac{p_\gamma^2}{2I_\phi}+mgR_G\cos(\theta) \tag{9.3.528}$$

Therefore:

$$E_{\text{tot}}-V(\theta)=\frac{1}{2}I_0\dot{\theta}^2 \tag{9.3.529}$$

Hence:

$$\dot{\theta}=\sqrt{\frac{2}{I_0}(E-V(\theta))} \tag{9.3.530}$$

The extreme values of θ can be found by solving the above relation for the roots of $E_{\text{tot}}-V(\theta)$ in θ (these are the points in θ where $\dot{\theta}$ vanishes; i.e. the nutation angle extremes and therefore its variation $\Delta\theta=\theta_{\text{max}}-\theta_{\text{min}}$):

$$E_{\text{tot}}-V(\theta)=E_{\text{tot}}-\frac{(p_\phi-p_\gamma\cos(\theta))^2}{2I_0\sin^2(\theta)}-\frac{p_\gamma^2}{2I_\phi}-mgR_G\cos(\theta)=0 \tag{9.3.531}$$

That is:

$$E_{\text{tot}}-\frac{(p_\phi-p_\gamma\cos(\theta))^2}{2I_0(1-\cos^2(\theta))}-\frac{p_\gamma^2}{2I_\phi}-mgR_G\cos(\theta)=0 \tag{9.3.532}$$

Or that we can rewrite:

$$E_{\text{tot}}(1-\cos^2(\theta))-\frac{(p_\phi-p_\gamma\cos(\theta))^2}{2I_0}-\frac{p_\gamma^2}{2I_\phi}(1-\cos^2(\theta))-mgR_G\cos(\theta)(1-\cos^2(\theta))=0 \tag{9.3.533}$$

and also:

$$2I_0E_{\text{tot}}(1-\cos^2(\theta))-(p_\phi-p_\gamma\cos(\theta))^2-\frac{2I_0p_\gamma^2}{2I_\phi}(1-\cos^2(\theta))-2I_0mgR_G\cos(\theta)(1-\cos^2(\theta))=0 \tag{9.3.534}$$

Grouping, we get:

$$\begin{aligned}2I_0mgR_G\cos^3(\theta)+\left(2I_0p_\phi p_\gamma^2+\frac{2I_0p_\gamma^2}{2I_\phi}-2I_0E_{\text{tot}}\right)\cos^2(\theta)+(4I_0p_\phi p_\gamma-2I_0mgR_G)\cos(\theta)\\+\left(2I_0E_{\text{tot}}-2I_0p_\phi^2-\frac{2I_0p_\gamma^2}{2I_\phi}\right)=0\end{aligned} \tag{9.3.535}$$

Factorizing:

$$\begin{aligned}2I_0mgR_G\cos^3(\theta)+\left(p_\gamma^2+\frac{2I_0p_\gamma^2}{2I_\phi}-2I_0E_{\text{tot}}\right)\cos^2(\theta)+(2p_\phi p_\gamma-2I_0mgR_G)\cos(\theta)\\+\left(2I_0E_{\text{tot}}-p_\phi^2-\frac{2I_0p_\gamma^2}{2I_\phi}\right)=0\end{aligned} \tag{9.3.536}$$

Reorganizing:

$$\begin{aligned}2I_0mgR_G\cos^3(\theta)-\left(2I_0E_{\text{tot}}-\frac{2I_0p_\gamma^2}{2I_\phi}-p_\gamma^2\right)\cos^2(\theta)+(2p_\phi p_\gamma-2I_0mgR_G)\cos(\theta)\\+\left(2I_0E_{\text{tot}}-\frac{2I_0p_\gamma^2}{2I_\phi}-p_\phi^2\right)=0\end{aligned} \tag{9.3.537}$$

Factorizing:

$$\begin{aligned}2I_0mgR_G\cos^3(\theta)-\left\{2I_0\left(E_{\text{tot}}-\frac{p_\gamma^2}{2I_\phi}\right)-p_\gamma^2\right\}\cos^2(\theta)+(2p_\phi p_\gamma-2I_0mgR_G)\cos(\theta)\\+\left\{2I_0\left(E_{\text{tot}}-\frac{p_\gamma^2}{2I_\phi}\right)-p_\phi^2\right\}=0\end{aligned} \tag{9.3.538}$$

Setting $x=\cos(\theta)$ we have:

$$2I_0mgR_Gx^3-\left\{2I_0\left(E_{\text{tot}}-\frac{p_\gamma^2}{2I_\phi}\right)-p_\gamma^2\right\}x^2+2\left(p_\phi p_\gamma-I_0mgR_G\right)x+\left\{2I_0\left(E_{\text{tot}}-\frac{p_\gamma^2}{2I_\phi}\right)-p_\phi^2\right\}=0 \tag{9.3.539}$$

By solving for x, we find one or two roots between -1 and $+1$ and another non-physical root ($|x|>1$). The two physically meaningful roots will yield the span of angles of the nutation. If the roots are equivalent, there is no nutation; only smooth precession.

We will now write the latter relation as in our study of third degree polynomial equations in the section Calculus (see page 1390):

$$ax^3+bx^2+cx+d=0 \tag{9.3.540}$$

with:

$$\begin{aligned}a&=2I_0mgR_G\\ b&=2I_0\left(\frac{p_\gamma^2}{2I_\phi}-E_{\text{tot}}\right)+p_\gamma^2\\ c&=2\left(p_\phi p_\gamma-I_0mgR_G\right)\\ d&=2I_0\left(E_{\text{tot}}-\frac{p_\gamma^2}{2I_\phi}\right)-p_\phi^2\end{aligned} \tag{9.3.541}$$

We use now the trick used in the section Calculus. First we know we have to put:

$$\begin{cases}x=X-\dfrac{b}{3a}\\ p=\dfrac{c}{a}-\dfrac{b^2}{3a^2}\\ q=\dfrac{2b^3}{27a^3}-\dfrac{cb}{3a^2}+\dfrac{d}{a}\end{cases} \tag{9.3.542}$$

So we get the following simplified polynomial (but the explicit expression of the coefficients are quite long...):

$$P=X^3+pX+q \tag{9.3.543}$$

After, as we have proved it in the section Calculus we must calculate the following discriminant:

$$\Delta=q^2+4\frac{p^3}{27} \tag{9.3.544}$$

and determine it's sign.

Caution!! As far as we know, none of the relations derived earlier for the nutation or lean angle applies for a celestial body like Earth or any other planet. A numerical application leads to results that doesn't make any sense in a magnitude of 10^{20}!

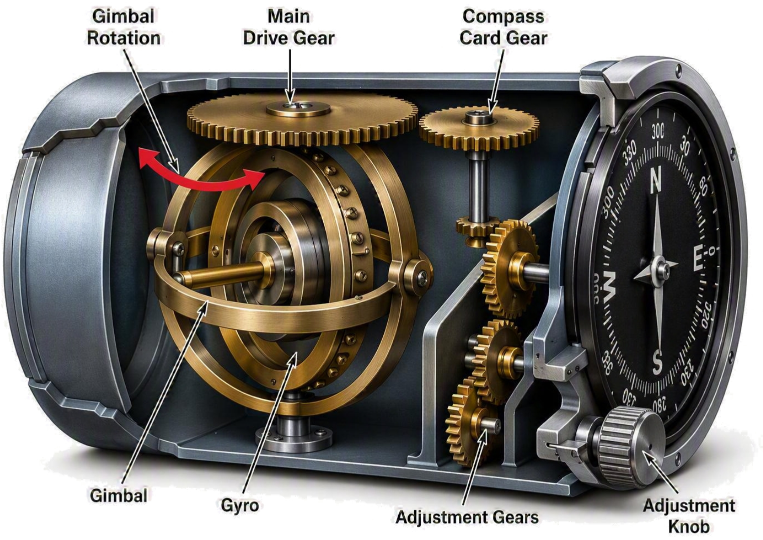

Gimbal Rotation
Main Drive Gear
Compass Card Gear
Gimbal
Gyro
Adjustment Gears
Adjustment Knob
N
E
S
W

Figure 9.66: **Mechanical airplane gyroscope principle**

Forces associated with the rotation of Earth cause the planet to be slightly oblate, displaying a bulge at the equator (see section Astronomy page 5300). The Moon's gravity primarily, and to a lesser degree the Sun's gravity, act on Earth's oblateness to move the axis perpendicular to the plane of Earth's orbit. However, due to gyroscopic action, Earth's poles do not "right themselves" to a position perpendicular to the orbital plane. Instead, they precess at 90 degrees to the force applied. This luni-solar precession effect causes the axis of Earth to describe a circle having a 23.4 degree radius relative to a fixed point in space over about $26,000$ years, a slow wobble reminiscent of the axis of a spinning top swinging around before it falls over.

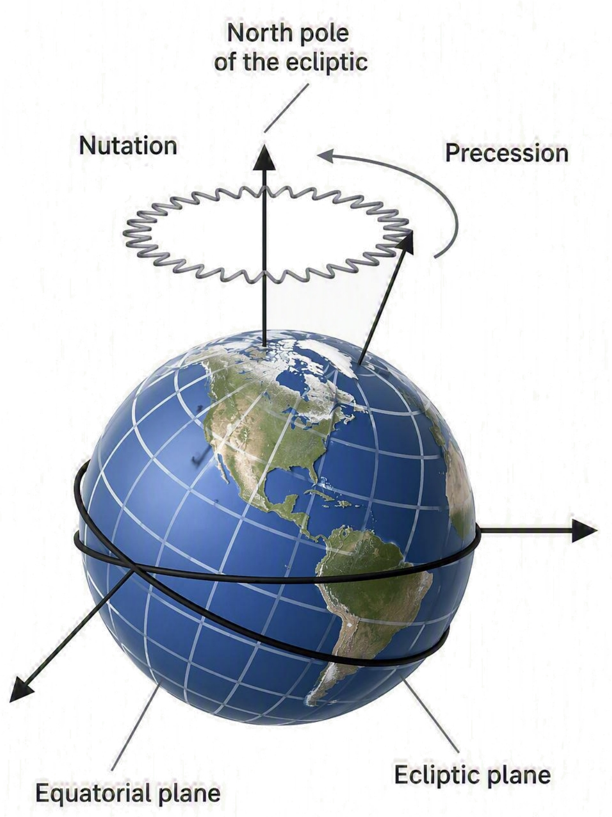

North pole
of the ecliptic
Nutation
Precession
Equatorial plane
Ecliptic plane

Figure 9.67: Mouvement de nutation de la Terre

However, the Earth is not the only planet having precessional and nutational motions. Mars in particular is also rapidly rotating, has an equatorial bulge, and its mean orientation is tilted with respect to the ecliptic (the obliquity of Mars is about 25°). Therefore Mars also changes its orientation in space. All the giants planets (Jupiter, Saturn, Uranus, Neptune) also have a nutation movement!

Like Foucault used his pendulum to prove the rotation of the Earth, in 1185 (holocene calendar) he used the Gyroscope to demonstrate again Earth's rotation:

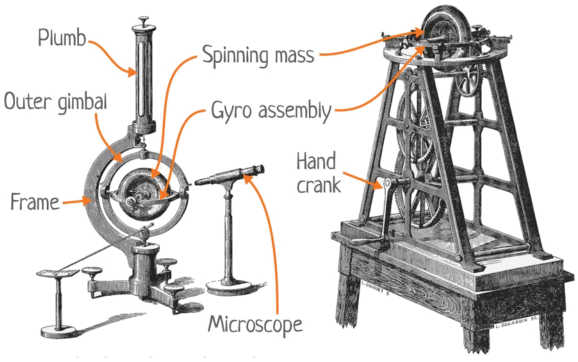

Plumb
Outer gimbal
Frame
Spinning mass
Gyro assembly
Hand
crank
Microscope

Figure 9.68: Foucault gyroscope (author: L. Dujardin)

He used as visible in the illustration above a hand crank with four stages of gears to spin up the gyroscope to the initial speed of $1,200$ rpm, allowing it to persist for ten minutes. This was sufficient to observe the precession from Earth's rotation using a microscope!

Some quite complex mathematical models (maybe one day we will detail them in this textbook) give the possibility to compute the precession-nutation period for planets. The most well-known model (Kinoshita theory) gives for example the following theoretical numerical values for the precession period T_E (in million years) considering only the torque of the Sun (actually a G2V-type star, aka yellow dwarf) but not of the planets satellites:

Planet	Period $T_P(My)$	T_P/T_E
Earth	0.0258	1
Jupiter	1.4960	57.966
Saturn	7.2564	281.26
Uranus	1136.84	44063.6
Neptune	753.49	29204.9

Table 9.18: **Theoretical precession periods of some giant planets**

Here is a nice visual summary of the lean angle (i.e. tilt angle) and rotation periods of some well known celestial bodies as in the 121st century (holocene calendar):

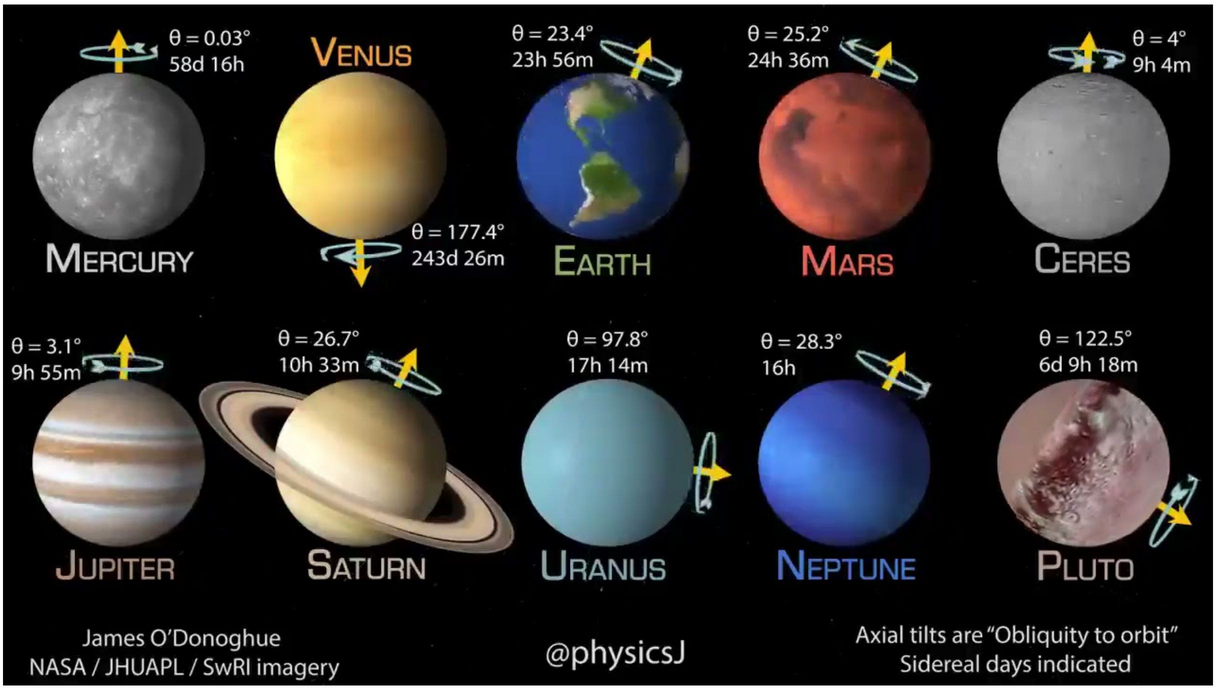

θ = 0.03°
58d 16h
MERCURY
VENUS
θ = 177.4°
243d 26m
θ = 23.4°
23h 56m
EARTH
θ = 25.2°
24h 36m
MARS
θ = 4°
9h 4m
CERES
θ = 3.1°
9h 55m
JUPITER
θ = 26.7°
10h 33m
SATURN
θ = 97.8°
17h 14m
URANUS
θ = 28.3°
16h
NEPTUNE
θ = 122.5°
6d 9h 18m
PLUTO
James O'Donoghue
NASA / JHUAPL / SwRI imagery
@physicsJ
Axial tilts are "Obliquity to orbit"
Sidereal days indicated

Figure 9.69: **Tilt and rotation period of some well known celestial bodies**

9.3.8.2.4 König's kinetic and angular momentum theorems

In kinetics, "König's theorems" or "König's decompositions" is a couple of mathematical relations derived by Johann Samuel König that assists with the calculation of kinetic energy and angular momentum of bodies and systems of particles.

These theorems expresses the kinetic energy and angular momentum of a system of particles in terms of the velocities of the individual particles. Specifically, they states that the kinetic energy and angular moment of a system of particles is the sum of the kinetic energy/angular momentum associated to the movement of the center of mass and the kinetic energy/angular momentum associated to the movement of the particles relative to the center of mass.

As this result is quite obvious many teachers and books don't speak about these theorems and therefore don't prove them. I have put these theorem here only as a tribute to one of the teacher of my Engineering School but for not other reason as we will never mention them directly in any other section or chapter of this textbook (again because it's simply obvious!).

OK. Let us start! So don't forget we have seen so far how to calculate the angular momentum or kinetic energy of a dynamic system compared to a single reference frame (either Galilean or centroid).

The König's theorems give respectively the angular momentum and the total kinetic energy of a dynamic system with respect to a Galilean reference frame R_{gal} or centroid one R_{cen}.

9.3.8.2.4.1 First König's Theorem (König's angular momentum theorem)

Let us use the angular momentum of a body of mass M to prove this theorem (the example being always easily extendable to a discrete or continuous dynamic system of matter).

Let us express the angular momentum $\vec{b}$ of an element m_i of the solid body with respect to the origin O of the Galilean R_{gal} reference frame (denoted thereafter $/\text{O},R$):

$$\vec{b}_{/\text{O},R}=\sum_i m_i\overrightarrow{\text{O}A_i}\times\vec{v}_{i/R'} \tag{9.3.545}$$

Let us express the angular momentum in R_{cen} relatively to its center of mass G (denoted thereafter $/G,R'$):

$$\vec{b}_{/G,R'}=\sum_i m_i\overrightarrow{GA_i}\times\vec{v}_{i/R'} \tag{9.3.546}$$

The reference frame R_{cen} being in translation relatively to R_{gal}, we have:

$$\vec{v}_{i/R}=\vec{v}_{i/R'}+\vec{v}_{G/R} \tag{9.3.547}$$

Without forgetting that:

$$\overrightarrow{\text{O}A_i}=\overrightarrow{\text{O}G}+\overrightarrow{GA_i} \tag{9.3.548}$$

that we insert in the expression of the angular momentum relatively to R_{gal}:

$$\vec{b}_{/\text{O},R}=\sum_i m_i\overrightarrow{\text{O}A_i}\times\vec{v}_{i/R}=\sum_i m_i\left(\overrightarrow{\text{O}G}+\overrightarrow{GA_i}\right)\times\left(\vec{v}_{i/R'}+\vec{v}_{G/R}\right) \tag{9.3.549}$$

From the property of distributivity over addition of the cross product, we have:

$$\vec{b}_{/\text{O},R}=\sum_i m_i\overrightarrow{\text{O}G}\times\vec{v}_{i/R'}+\sum_i m_i\overrightarrow{GA_i}\times\vec{v}_{i/R'}+\sum_i m_i\overrightarrow{\text{O}G}\times\vec{v}_{G/R}+\sum_i m_i\overrightarrow{GA_i}\times\vec{v}_{G/R} \tag{9.3.550}$$

Let us now study the value that take each of the four terms of the previous relation. We have proved by definition of the center of mass that (in a non-relativistic framework):

$$\frac{\sum\limits_i m_i\overrightarrow{GA_i}}{\sum\limits_i}=\vec{0}\Rightarrow\sum_i m_i\overrightarrow{GA_i}=\vec{0}\Rightarrow\frac{\mathrm{d}}{\mathrm{d}t}\sum_i\overrightarrow{GA_i}=\sum_i m_i\left(\frac{\mathrm{d}}{\mathrm{d}t}\overrightarrow{GA_i}\right)=\sum_i m_i\vec{v}_{i/R'}=\vec{0} \tag{9.3.551}$$

hence:

$$\sum_i m_i\overrightarrow{\mathrm{O}G}\times\vec{v}_{i/R'}=\overrightarrow{\mathrm{O}G}\times\left(\sum_i m_i\vec{v}_{i/R'}\right)=\vec{0} \tag{9.3.552}$$

So it comes:

$$\begin{aligned}\vec{b}_{/\mathrm{O},R}&=\sum_i m_i\overrightarrow{GA_i}\times\vec{v}_{i/R'}+\sum_i m_i\overrightarrow{\mathrm{O}G}\times\vec{v}_{G/R}=\vec{b}_{/G,R'}+\overrightarrow{\mathrm{O}G}\times\vec{v}_{G/R}\sum_i m_i\\&=\vec{b}_{/G,R'}+M\overrightarrow{\mathrm{O}G}\times\vec{v}_{G/R}\end{aligned} \tag{9.3.553}$$

Therefore we have finally:

$$\boxed{\vec{b}_{/\mathrm{O},R}=\vec{b}_{/G,R'}+M\overrightarrow{\mathrm{O}G}\times\vec{v}_{G/R}} \tag{9.3.554}$$

This theorem that relates to a fixed point allows an easier application of the theorem of angular momentum.

9.3.8.2.4.2 Second König's Theorem (König's kinetic energy theorem)

To prove this theorem let us use the kinetic energy of a mass M (the example is still easily expandable to a discrete or continuous dynamic system of material).

Let us express the kinetic energy of an element m_i of the solid body with respect to the origin O of the Galilean reference frame R_{gal} (denoted thereafter $/\mathrm{O},R$):

$$E_{c/\mathrm{O},R}=\frac{1}{2}\sum_i m_i v_{i/R}^2 \tag{9.3.555}$$

Let us express the kinetic energy in R_{cent} relatively to its center of mass G (denoted thereafter $/G,R'$):

$$E_{c/G,R'}=\frac{1}{2}\sum_i v_{i/R'}^2 \tag{9.3.556}$$

With same as above:

$$\vec{v}_{i/R}=\vec{v}_{i/R'}+\vec{v}_{G/R} \tag{9.3.557}$$

Therefore we have:

$$E_{c/\mathrm{O},R}=\frac{1}{2}\sum_i m_i\vec{v}_{i/R}^2=\frac{1}{2}\sum_i m_i\left(\vec{v}_{i/R'}+\vec{v}_{G/R}\right)^2 \tag{9.3.558}$$

Hence (the dot products of the vectors are implicit!!!):

$$\begin{aligned}E_{c/\text{O},R}&=\frac{1}{2}\sum_i m_i\vec{v}_{i/R'}^2+\frac{1}{2}\sum_i m_i\vec{v}_{G/R}^2+\frac{1}{2}\sum_i 2m_i\vec{v}_{i/R'}\vec{v}_{G/R}\\&=E_{c,G/R'}+\frac{1}{2}M\vec{v}_{G/R}^2+\vec{v}_{G/R}\sum_i\vec{v}_{i/R'}\end{aligned}\tag{9.3.559}$$

and as for the angular momentum, by the definition of the mass center, we have:

$$\sum_i m_i\vec{v}_{i/R'}=\vec{0}\tag{9.3.560}$$

Hence the second König's theorem:

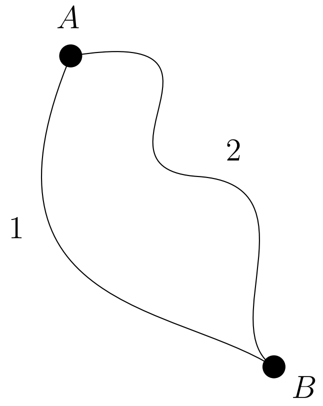
$$\boxed{E_{c/\text{O},R}=E_{c,G/R'}+\frac{1}{2}Mv_{G/R}^2}\tag{9.3.561}$$

9.3.8.3 Gravitational Potential Energy

Gravitational potential energy is energy an object possesses because of its position in a gravitational field. The most common use of gravitational potential energy is for an object near the surface of the Earth. The gravitational potential is also known as the "Newtonian potential" and is fundamental in the study of potential theory.

Since the zero of gravitational potential energy can be chosen at any point (like the choice of the zero of a coordinate system), the potential energy at a height h above that point is equal to the work which would be required to lift the object to that height with no net change in kinetic energy. Since the force required to lift it is equal to its weight, it follows that the gravitational potential energy is equal to its weight mg times the height h to which it is lifted mgh.

Let us first recall that if the work of the force $\vec{F}$ between points A and B is independent of the path taken, we say that this force derives from a potential energy or that force field is a "conservative field" (counter-example: in a motion with friction works necessarily depends on the chosen path). This independence from the path followed that if given two points A and B in space. There are several possible ways to join these two points. If we choose two random points we have:



- On the first path: $\displaystyle\int_A^B \vec{F}\circ \mathrm{d}\vec{r}_1$
- On the second path: $\displaystyle\int_A^B \vec{F}\circ \mathrm{d}\vec{r}_2$

If the field is conservative we have:

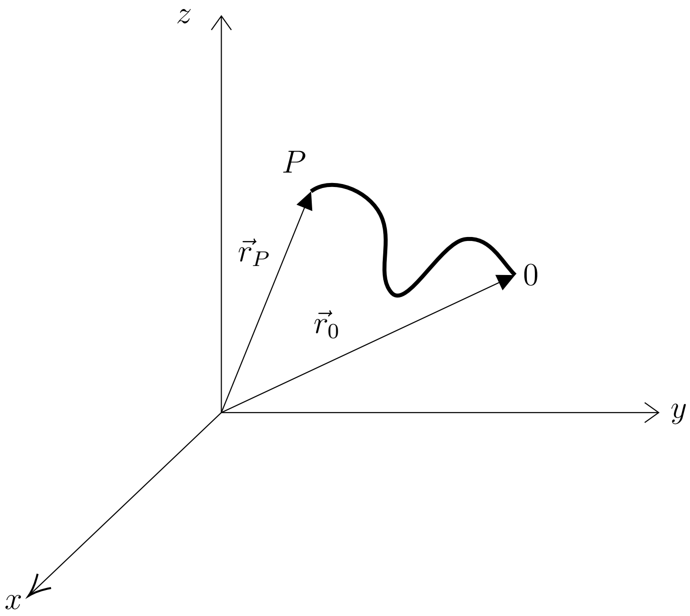
$$\int_A^B \vec{F}\circ \mathrm{d}\vec{r}_1=\int_A^B \vec{F}\circ \mathrm{d}\vec{r}_2=\int_A^B \vec{F}\circ \mathrm{d}\vec{r}=-\int_B^A \vec{F}\circ \mathrm{d}\vec{r} \tag{9.3.562}$$

or in other words that the total work on a closed path (go and come back) is zero. We notice that (see section Differential and Integral Calculus page 1583):

$$\oint \vec{F}\circ \mathrm{d}\vec{r}=0 \tag{9.3.563}$$

The work involved is therefore a function of the location only (E_p) that is to say, depending only on the starting point and ending point. Indeed, if the work depended on the path it would be possible to choose the most generous way when the system provides work and the most economical way when we bring him back to the initial state. So it would be a perpetual motion and the principle of conservation of energy forbids that (see section Thermodynamics page 3666)!

Let us then attach to each point P of the force field a value of the function E_p (a real number) corresponding to the work done by the force field when the mobile moves from one point P to 0, 0 being a reference point arbitrarily chosen:

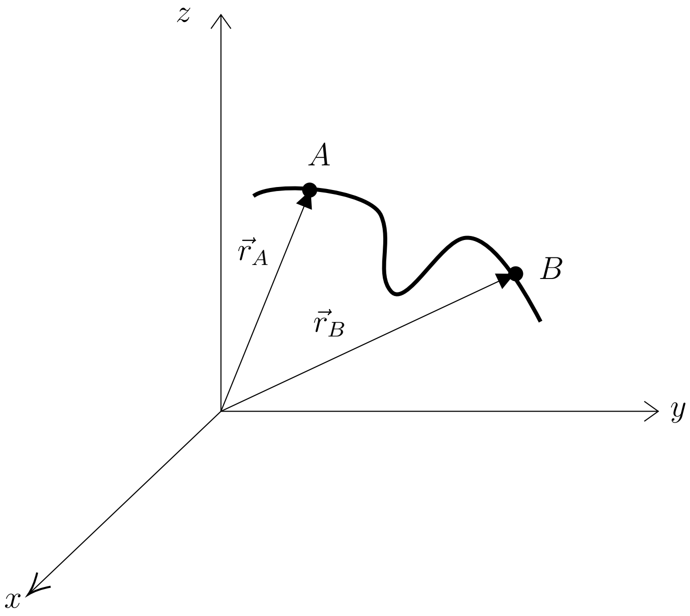


So by definition:

$$E_p(P)=\int_P^0 \vec{F}\circ \mathrm{d}\vec{r} \qquad (9.3.564)$$

with $E_p(0)=0$.

Generalizing this definition a little bit, we say that the work W done by a conservative force when moving from A to B:



is equal to the reduction of potential energy between A and B:

$$W_{A,B}=\int\limits_{A}\vec{F}\circ\mathrm{d}\vec{r}=-(E_p(B)-E_p(A)) \tag{9.3.565}$$

By definition E_p is the potential energy and is measured in Joules [J] as for the kinetic energy.

The above relation is used very often in its differential form is:

$$\boxed{\mathrm{d}W=-\mathrm{d}E_p} \tag{9.3.566}$$

There is also for recall a relation between energy and the gradient of the given force that simply follows from the definition of work:

$$\boxed{\vec{F}=-\vec{\nabla}\cdot E_p} \tag{9.3.567}$$

The most significant application in Classical Mechanics is the work of gravity and the gravitational potential energy near the surface of the Earth (or another planet!). It is therefore a special case where the force is constant...

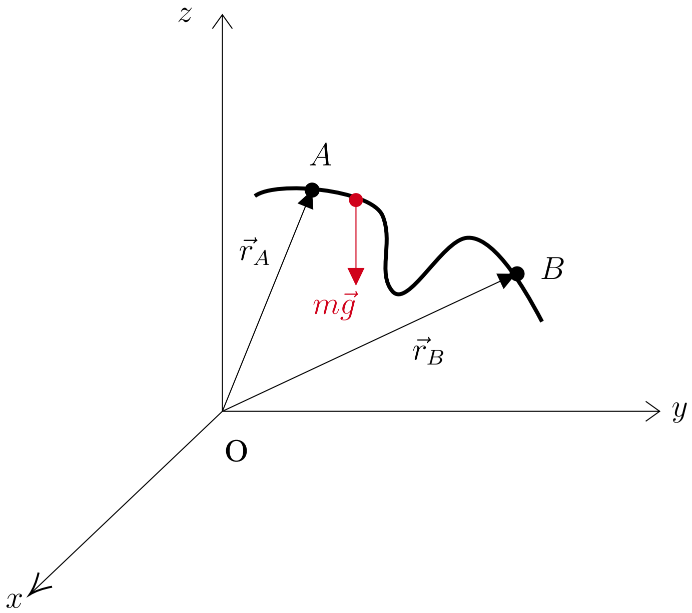


Figure 9.70: **Illustration of gravitational work between two points in constant force field**

Following the figure above given a point of mass m moving following any trajectory AB in a constant gravitational field force $\vec{g}$. The weight $m\vec{g}$ do then the following work:

$$W_{A,B}=\int_A^B \vec{F}\circ\mathrm{d}\vec{r}=\vec{F}\circ(\vec{r}_B-\vec{r}_A)=m\vec{g}\circ(\vec{r}_B-\vec{r}_A) \tag{9.3.568}$$

The difference z_A-z_B is the difference in elevation between points A and B. We notice well that the work does not depend on the path followed but only on the points of departure and arrival. If, conversely, we want to move the massive point from B to A, the work, then provided by an external element is:

$$W_{B,A}=-W_{A,B}=-mg(z_A-z_B) \tag{9.3.569}$$

which shows well that the total work on a closed path is zero is such a special case:

$$\oint \vec{F}\circ\mathrm{d}\vec{r}=0 \tag{9.3.570}$$

Comparing the relations:

$$W_{A,B}=-(E_P(B)-E_P(A)) \qquad \text{and} \qquad W_{A,B}=mg(z_A-z_B) \tag{9.3.571}$$

and identifying, we get for a constant gravitational field:

$$\boxed{E_p(z)=mgz} \tag{9.3.572}$$

which is the gravitational potential energy, z is the altitude of the mass m. We denote this simply this relation this relation under the form:

$$E_p=mgh \tag{9.3.573}$$

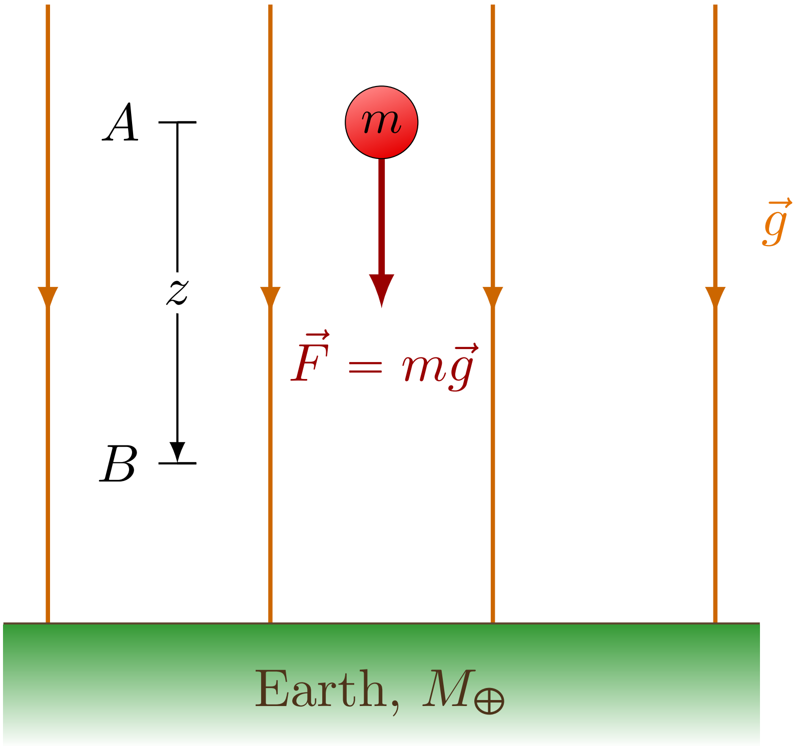


Figure 9.71: **Naive scenario of gravitational potential energy (author: Izaak Neutelings)**

Remark

The choice of the zero level of potential energy is often arbitrary; we set it ourselves for convenience depending on the situation. Only the differences of potential energy are generally interesting as we will see just below.

The previous relation is actually a useful expression near the Earth's surface (or any other planet). At a distance $r \gg R$ where R is the radius of the Earth, the gravitational force weakens and the approximation is no longer valid (if $r \ll R$ also anyway...).

To determine the correct relation, consider two masses m_1, m_2. The first is supposed at rest and fixed and the second is brought from infinity to a given distance from m_1 (the same reasoning applies to the electric field!!!!). The work $\mathrm{d}W$ of the gravitational force at any point being thus:

$$\mathrm{d}W = F\mathrm{d}x \tag{9.3.574}$$

and the potential energy of the system:

$$\mathrm{d}E_p = -\mathrm{d}W \tag{9.3.575}$$

Then:

$$\mathrm{d}E_p = -\mathrm{d}W = -F\mathrm{d}x = -G\frac{m_1m_2}{x^2}\mathrm{d}x \tag{9.3.576}$$

where just after integration (the potential energy in a point):

$$\boxed{E_p = G\frac{m_1m_2}{x}} \tag{9.3.577}$$

Let us see if this is consistent with $E_p = mgz$...:

Proof 9.8.2. At zero height of the planet surface $x=R$, we have:

$$E_p=-G\frac{m_1m_2}{R}=E_{0,p} \tag{9.3.578}$$

where the choice of the sign "$-$" depends only on the selected reference frame that is in this case in line with what is customary to take in schools.

We raise the object of $h\ll R$:

$$E_p(R+h)=-G\frac{m_1m_2}{R+h}=-Gm_1m_2\frac{1}{R+h}=-Gm_1m_2\frac{1}{R(1+h/R)}=-G\frac{m_1m_2}{R}\frac{1}{1+h/R} \tag{9.3.579}$$

We use now the rough approximation (see section Sequences and Series page 1696):

$$\frac{1}{1+x}\cong 1-x \tag{9.3.580}$$

valid when $x\ll 1$, hence:

$$E_p(R+h)\cong -G\frac{m_1m_2}{R}\left(1-\frac{h}{R}\right)=-G\frac{m_1m_2}{R}+G\frac{m_1m_2}{R^2}h=U_0-F_Gh=U_0+mgh \tag{9.3.581}$$

As at the surface of the Earth (or any other planet) we usually put in laboratories that $U_0=0$, we get well eventually:

$$E_p(h)\cong mgh \tag{9.3.582}$$

and we see that it is indeed a rough approximation.

■ Q.E.D.

Remark

We could apply the same development in the study of the Coulomb force and of the Electric field but until now we have never been put a laboratory to the surface of an electric charge... (sic!).

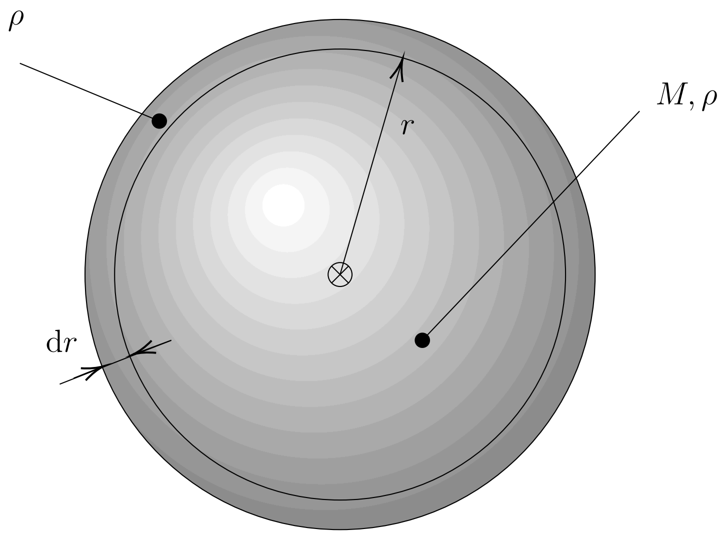
9.3.8.3.1 Gravitational Potential Energy of a Material Sphere

We will calculate here the potential energy of a material sphere. This style of exercise will be very useful in the section of Astrophysics (see page 5539) to determine the internal temperature of the stars and in the section about Cosmogony (see page 5878) to start the study of the Friedmann model.

The potential energy of the spherical ring of inner radius r and thickness $\mathrm{d}r$ is calculated as follows:

$$E_p=-\frac{GMm}{r} \tag{9.3.583}$$

Given a sphere of mass M, of mass density ρ and radius r surrounded by a spherical ring of inner radius r, of the same mass density ρ and thickness $\mathrm{d}r$:



The mass of the sphere of radius r and mass density ρ is equal as we know:

$$M=V\rho=\frac{4}{3}\pi r^3\rho \tag{9.3.584}$$

The mass of the ring surrounding the sphere of radius r, of thickness $\mathrm{d}r$ and mass density ρ is:

$$m=\mathrm{d}V\rho=4\pi r^2\mathrm{d}r\rho \tag{9.3.585}$$

By introducing the last two terms in the expression of the potential energy:

$$\mathrm{d}E_p=-\frac{G\left(\dfrac{4}{3}\pi r^3\rho\right)(4\pi r^2\rho\mathrm{d}r)}{r}=-\frac{1}{3}G(4\pi\rho)^2r^4\mathrm{d}r \tag{9.3.586}$$

By integrating the previous expression between 0 and R, this is equivalent to add successively a range of rings of thickness $\mathrm{d}r$ to get the whole sphere of radius R and therefore the potential energy of the whole sphere:

$$E_p=-\frac{1}{3}G(4\pi\rho)^2\int_0^R r^4\mathrm{d}r=-\frac{1}{3}G(4\pi\rho)^2\frac{R^5}{5} \tag{9.3.587}$$

Which can be written also:

$$E_p=-\frac{3}{5}G\left(\frac{4}{3}\pi\rho R^3\right)^2\frac{1}{R}=-\frac{3}{5}GM^2\frac{1}{R} \tag{9.3.588}$$

Thus finally:

$$\boxed{E_p=-\frac{3}{5}\frac{GM^2}{R}} \tag{9.3.589}$$

Exactly the same calculation can be done in electrostatics by replacing the gravitational constant G by the Coulomb constant k_e and the mass M by the total Electric charge Q. It is also a result we will reuse in the model of the liquid nuclear core (see section Nuclear Physics page 5141).

Remark

As the Newton law and Coulomb law are similar in shape, we get using exactly the same calculation for electrostatic in a condensed stable material:

$$E_p=\frac{3}{5}\frac{Q^2}{4\pi\epsilon_0 a} \tag{9.3.590}$$

It may seem useless to have such a result however... it is used a lot as basis of the study of "polarons". A polaron refers to the local distortion of a molecular structure or lattice caused by a charge carrier moving through a semiconductor, the charge carrier and its disturbed environment treated mathematically as a fermionic quasiparticles. It represents the combination of both the electron or hole and the displacement of atoms from their equilibrium position within the molecule or lattice. Polarons form easily in polarizable materials as excess charge carriers couple with ionic lattice vibrations, leading to localized lattice distortions.

9.3.8.3.2 Gravitational Effective potential

Definition (#445): The "effective potential" (also known as "effective potential energy") combines multiple, perhaps opposing, effects into a single potential. In its basic form, it is the sum of the 'opposing' centrifugal potential energy with the potential energy of a dynamical system. It may be used to determine the orbits of celestial objects (both Newtonian and relativistic as proved at page 5845) and to perform semi-classical atomic calculations (see section Quantum Chemistry page 6187), and often allows problems to be reduced to fewer dimensions.

Consider a particle of mass m orbiting a much heavier object of mass M. Assume Newtonian Mechanics, which is both classical and non-relativistic (the relativistic case based on General Relativity is treated in details in the section of General Relativity at page 5845). The conservation of energy E and angular momentum (see section Classical Mechanics page 3244) that we will denote here L instead of b to respect the traditions (...) give us two constants E and L (gravity being is a central force and thus not a torque which means that angular momentum is conserved), which have values (we expressed the speed in polar coordinates as proven in the section of Vector Calculus at page 1863):

$$\begin{gathered}E_{\text{tot}}=E_c+E_p=\frac{1}{2}mv^2-\frac{GmM}{r}=\frac{1}{2}m\left(\dot{r}^2+r^2\dot{\phi}^2\right)-\frac{GmM}{r}=\frac{1}{2}m\left(v_r^2+v_\phi^2\right)-\frac{GmM}{r}\\ L=mr^2\dot{\phi}\end{gathered} \tag{9.3.591}$$

when the motion of the larger mass is negligible. In these expressions:

- $\dot{r}$ is the derivative of r with respect to time,
- $\dot{\phi}$ is the angular velocity of mass m,
- G is the gravitational constant,
- E_{tot} is the total energy, and

- L is the angular momentum.

Only two variables are needed, since the motion occurs in a plane. Substituting the second expression into the first and rearranging gives:

$$m\dot{r}^2=2E_{\text{tot}}-\frac{L^2}{mr^2}+\frac{2GmM}{r}=2E_{\text{tot}}-\frac{1}{r^2}\left(\frac{L^2}{m}-2GmMr\right) \qquad (9.3.592)$$

That is denoted by tradition:

$$\frac{1}{2}m\dot{r}^2=E-U_{\text{eff}}(r) \qquad (9.3.593)$$

with by definition the "Gravitational non-relativistic effective potential" (as the previous relation has the shape of something like $E_c=E_{\text{tot}}-E_p$):

$$\boxed{U_{\text{eff}}(r):=\frac{L^2}{2mr^2}-\frac{GmM}{r}} \qquad (9.3.594)$$

The original two-variable problem has been reduced to a one-variable problem. For many applications the effective potential can be treated exactly like the potential energy of a one-dimensional system: for instance, an energy diagram using the effective potential determines turning points and locations of stable and unstable equilibria. A similar method may be used in other applications, for instance determining orbits in a general relativistic Schwarzschild metric.

Therefore an effective potential energy is not technically a potential energy but can still be used to obtain qualitative information about motion. So we can say that one introduces an effective potential when a physical effect whose origin is not a potential can be described adequately as if it simply gave rise to a potential!

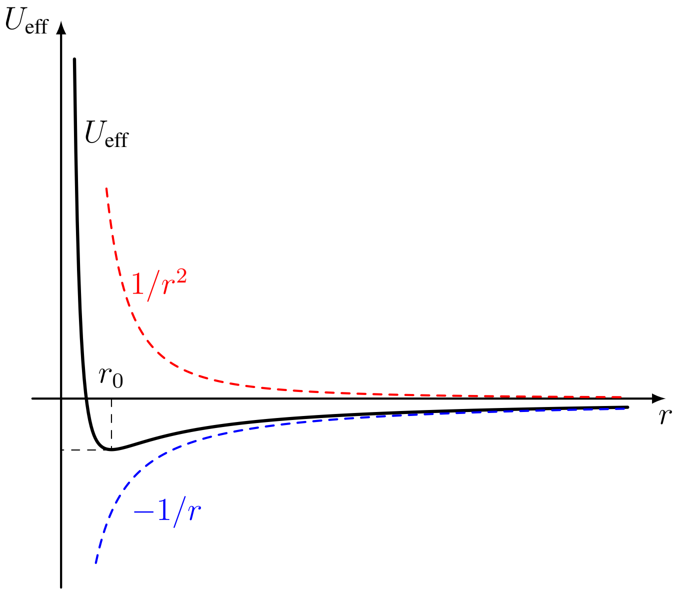


Figure 9.72: Graph of non-relativistic effective gravitational potential (author: Alexandros Tsagkaropoulos)

Given an angular momentum L, we can plot the non-relativistic gravitational effective potential as above and immediately see where the stable circular orbits are (the minima of the potential).

We can also qualitatively see how there is a barrier to approaching the object too closely (which makes sense - if we have angular momentum you wouldn't expect a head on collision).

Doing the same kind of plot for multiple values of L we see that the minimum position of the effective potential increases with angular momentum. Therefore, planets with higher angular momenta are farther from their stars.

9.3.8.4 Conservation of Total Mechanical Energy

Let us compare the relations:

$$W_{A,B} = -(E_p(B) - E_p(A)) = E_c(B) - E_c(A) \tag{9.3.595}$$

since it is the same work!

Which leads to:

$$E_p(A) - E_c(A) = E_c(B) - E_p(A) \tag{9.3.596}$$

sum of the two forms of energy at each point or also, considering any locations A and B, by rewriting the equation in a more general form:

$$\boxed{E_p + E_c = c^{te}} \tag{9.3.597}$$

Remark

We often name the total energy of a system "Hamiltonian of the system" as we have already study it in the section of Analytical Mechanics.

In the absence of friction of external force in the case of mechanical energy, we also write the variation such that:

$$\mathrm{d}E_p + \mathrm{d}E_c = 0 \tag{9.3.598}$$

An increase in kinetic energy therefore results in a decrease of potential energy (and vice versa) since the sum of the two remains constant.

Against-example: If there is friction, heat is released and then total mechanical energy is not constant! Also if the studied object is a rocket propulsed in space.

Furthermore let us come back on the relation:

$$\mathrm{d}W = -\mathrm{d}E_p \Leftrightarrow \mathrm{d}E_p = -\mathrm{d}W \tag{9.3.599}$$

and therefore:

$$\mathrm{d}E_p = -\mathrm{d}W = -\vec{F} \circ \mathrm{d}\vec{r} = -(F_x\mathrm{d}x + F_y\mathrm{d}y + F_z\mathrm{d}z) \tag{9.3.600}$$

On the other hand, E_p being a scalar function dependent on spatial coordinates, let us write its total differential:

$$\mathrm{d}E_p = \frac{\partial E_p}{\partial x}\mathrm{d}x + \frac{\partial E_p}{\partial y}\mathrm{d}y + \frac{\partial E_p}{\partial y}\mathrm{d}y \tag{9.3.601}$$

comparing with the previous equation and identifying term by term, we have:

$$\frac{\partial E_p}{\partial x}=-F_x \qquad \frac{\partial E_p}{\partial y}=-F_y \qquad \frac{\partial E_p}{\partial z}=-F_z \tag{9.3.602}$$

hence the sentence stating that the force derived from a potential energy if the work involved is independent of the followed path. If we express the force $\vec{F}$ in terms of unitary vectors, we get:

$$\vec{F}=F_x\vec{i}+F_y\vec{j}+F_z\vec{k}=-\left(\frac{\partial E_p}{\partial x}\vec{i}+\frac{\partial E_p}{\partial y}\vec{j}+\frac{\partial E_p}{\partial z}\vec{k}\right)=-\overrightarrow{\text{grad}}(E_p) \tag{9.3.603}$$

Finally, the claim (statement) that the force derived from a potential energy E_p can be summarized as:

$$\vec{F}=-\overrightarrow{\text{grad}}(E_p) \tag{9.3.604}$$

In the case of gravity:

$$\vec{F}=m\vec{g}=-\overrightarrow{\text{grad}}(E_p) \tag{9.3.605}$$

which is also written with the nabla operator (see section Vector Calculus page 1869):

$$\boxed{\vec{F}=-\overrightarrow{\text{grad}}(E_p)=-\vec{\nabla}(E_p)} \tag{9.3.606}$$

That is common to write also:

$$\vec{F}=-\vec{\nabla}U \tag{9.3.607}$$

That is to say in the one dimensional case:

$$\boxed{\vec{F}=-\frac{\partial U}{\partial x}} \tag{9.3.608}$$

So what we have seen so far about the potential energy bring us to what we name the "potential theory" and resumed by:

$$\begin{tikzcd}
& \boxed{\vec{F}=-\vec{\nabla}U} \arrow[dl] \arrow[dr, shift right] & \\
\boxed{\vec{\nabla}\times\vec{F}=\vec{0}} \arrow[rr, shift left] & & \boxed{\oint_\Gamma \vec{F}\circ \mathrm{d}\vec{r}=0} \arrow[ul, shift right] \arrow[ll, shift left]
\end{tikzcd}$$

Figure 9.73: **Potential theory summary**

9.3.8.4.1 Generalized Newton law

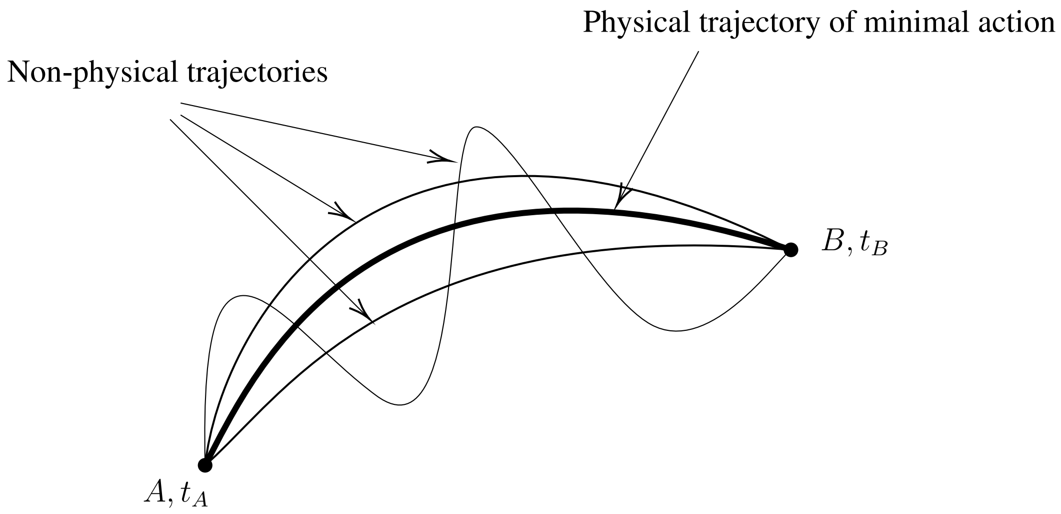
Now that we know explicitly how to express a kinetic and potential energy, let us come back to our principle of least action that we talked about in the section of Analytical Mechanics to proved that we can use it that latter to fall back on the generalized Newton's law demonstrated earlier in this section!

Let us take the case of an object launched into the air and let us choose two points of its motion path in any two different times. An infinite quantity of curves exist between these two points, but the Nature choose only one. So what distinguishes this curve - physical path - from all others? To answer this question we could quite rightly reply that this curve differs from others in that it is a solution of the differential equation of the path ... with appropriate initial conditions. But if we ignore the initial conditions or when the problem can not be reduced to a differential equation, by what method can we then distinguish the physical path of all possible paths?

The principle of last action in this context is expressed by a minimum speed for a minimum of motion path distance.

About speed, it is better in Mechanics to consider the quantity of linear momentum as the latter quantity is directly related to the inertial properties of the body. Pierre Louis Moreau de Maupertuis mathematically translated the principle of least action as follows for recall (see section Analytical Mechanics page 3152): If we consider the motion of a body between two points A in t_1 and B in t_2 by, for a given total energy E, the path selected by Nature is one for which the following quantity A_A^B is minimum:

$$\min A_A^B : \quad A_A^B = \int_{A,t_A}^{B,t_B} m\vec{v} \circ \mathrm{d}\vec{r} \tag{9.3.609}$$


Physical trajectory of minimal action
Non-physical trajectories
B, t_B
A, t_A

The physical path between two points A and B at times t_A and t_B is that for which the action is minimal!

Knowing that:

$$\mathrm{d}\vec{r} = \vec{v}\mathrm{d}t \tag{9.3.610}$$

then we get:

$$A_A^B = \int_{t_A}^{t_B} m\vec{v} \circ \vec{v}\mathrm{d}t = \int_{t_A}^{t_B} mv^2\mathrm{d}t = \int_{t_A}^{t_B} 2T(\vec{r},t)\mathrm{d}t \tag{9.3.611}$$

where T is as we the kinetic energy E_c of the body but as denoted in Analytical Mechanics...

We see that the action takes a surprisingly simple form and is expressed directly in terms of kinetic energy T. A few years later, from an intuition similar to that of Maupertuis, Leonhard Euler came to a very similar statement of the action but starting from the premise that the moving body tend to adopt a state where the potential energy is minimal. The Euler's action was expressed in terms of the potential energy rather than kinetic energy. Who of Maupertuis or Euler was right?

In fact, their personal definition Action were equivalent. We know that in a conservative field if we denote by U the potential energy then the total energy E_{tot} is equal to $T+U$ and the energy is a constant (in an isolated system). We derive immediately that $T=E_{\text{tot}}-U$ and therefore:

$$2T = T + E_{\text{tot}} - U \tag{9.3.612}$$

Hence:

$$A_A^B = \int_{t_A}^{t_B} (T(\vec{r},t) + E_{\text{tot}} - U(\vec{r},t)+)\mathrm{d}t \tag{9.3.613}$$

This relation is true regardless of the initial total energy E_{tot} of the path. We conclude that the value of the constant E_{tot} does not discriminate between different paths and can thus be removed from the formulation of the action. The action of Maupertuis can then be reduced to a new quantity denoted S (we fall back on the notations seen in the section of Analytical Mechanics):

$$S_A^B = \int_{t_A}^{t_B} (T(\vec{r},t) - U(\vec{r},t)+)\mathrm{d}t = \int_{t_A}^{t_B} L(\vec{r},t)\mathrm{d}t \tag{9.3.614}$$

This new formulation of the action was given by Joseph-Louis Lagrange in 11788 (holocene calendar). S is as we already know the "Lagrangian's action" (see section Analytical Mechanics page 3161) and the function:

$$L(\vec{r},t) = T(\vec{r},t) - U(\vec{r},t) \tag{9.3.615}$$

is named as we know the "Mechanics Lagrangrian". Thus formulated, the principle of least action became one of the most powerful tools of mechanics and above.

We have already seen how we express the principle of least action mathematically (see section Analytical Mechanics page 3152). In our case, the action is not a function of analytical variables but of trajectories!

Let us consider the simple case of a body of mass m that is moving on a single dimension (which we represent by the x-axis) of a point of abscissa x_1 at the moment t_1 to a point x_2 at a moment t_2. Let us suppose that the body is subject to a potential U that does not vary with the time that is to say $U=U(x)$. The action on this body on any path Γ going from (x_1,t_1) to (x_2,t_2) is then:

$$S(\Gamma)=\int\limits_{t_1}^{t_2}(T(x,t)-U(x)]\mathrm{d}t=\int\limits_{t_1}^{t_2}\left(\frac{1}{2}m\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2-U(x)\right)\mathrm{d}t \qquad (9.3.616)$$

Given Γ_0 the physical path and S_0 the action on this path. Let us denote by $\bar{x}$ the values of the position x on the physical path. Let us consider now a path Γ very near of Γ_0 such that the positions along the path Γ have the values $x(t)=\bar{x}(t)+\varepsilon(t)$ that we will write, to facilitate the notations: $x=\bar{x}+\varepsilon$.

Remark

The Lagrangian action of a curve $\gamma\in\mathcal{C}(x,y,\tau)$ is often denoted (for a unit mass, i.e. $m=1$):

$$\mathcal{A}_L(\gamma)=\int\limits_a^b L(\gamma,\dot{\gamma})\mathrm{d}t=\int\limits_a^b\left(\frac{1}{2}\|\dot{\gamma}\|^2+U(\gamma)\right)\mathrm{d}t$$

Or when there is no potential:

$$\mathcal{A}_L(\gamma)=\int\limits_a^b L(\gamma,\dot{\gamma})\mathrm{d}t=\int\limits_a^b\frac{1}{2}\|\dot{\gamma}\|^2\mathrm{d}t$$

Let us now calculate the action for this path:

$$S(\Gamma)=\int\limits_{t_1}^{t_2}\left(\frac{1}{2}m\left(\frac{\mathrm{d}(\bar{x}+\varepsilon)}{\mathrm{d}t}\right)^2-U(\bar{x}+\varepsilon)\right)\mathrm{d}t \qquad (9.3.617)$$

As ε is infinitely small, it is possible to express the potential as a limited Taylor development (see section Sequences and Series page 1696):

$$U(\bar{x}+\varepsilon)\cong U(\bar{x})+\frac{1}{1!}\frac{\mathrm{d}U}{\mathrm{d}x}\varepsilon=U(\bar{x})+\varepsilon\frac{\mathrm{d}U}{\mathrm{d}x} \qquad (9.3.618)$$

And for the first term we simply develop the square identity:

$$\left(\frac{\mathrm{d}(\bar{x}+\varepsilon)}{\mathrm{d}t}\right)^2=\left(\frac{\mathrm{d}\bar{x}}{\mathrm{d}t}\right)^2+2\frac{\mathrm{d}\bar{x}}{\mathrm{d}t}\frac{\mathrm{d}\varepsilon}{\mathrm{d}t}+\left(\frac{\mathrm{d}\varepsilon}{\mathrm{d}t}\right)^2 \qquad (9.3.619)$$

As we consider only the variations of the first order (very small one), the last term above can be neglected, which gives for the action on the path Γ:

$$S(\Gamma)=\int\limits_{t_1}^{t_2}\left[\frac{1}{2}m\left(\left(\frac{\mathrm{d}\bar{x}}{\mathrm{d}t}\right)^2+2\frac{\mathrm{d}\bar{x}}{\mathrm{d}t}\frac{\mathrm{d}\varepsilon}{\mathrm{d}t}\right)-U(\bar{x})-\varepsilon\frac{\mathrm{d}U}{\mathrm{d}x}\right]\mathrm{d}t \qquad (9.3.620)$$

Let us now put the variation δ of the action between the physical path Γ_0 and Γ is zero as it seems to be observed in Classical Mechanics labs:

$$\delta S=S(\Gamma)-S(\Gamma_0)=0 \qquad (9.3.621)$$

Therefore:

$$\begin{aligned}\delta S &= S(\Gamma)-S(\Gamma_0)\\ &=\int_{t_1}^{t_2}\left[\frac{1}{2}m\left(\left(\frac{\mathrm{d}\bar{x}}{\mathrm{d}t}\right)^2+2\frac{\mathrm{d}\bar{x}}{\mathrm{d}t}\frac{\mathrm{d}\varepsilon}{\mathrm{d}t}\right)-U(\bar{x})-\varepsilon\frac{\mathrm{d}U}{\mathrm{d}x}\right]\mathrm{d}t-\int_{t_1}^{t_2}\left[\frac{1}{2}m\left(\frac{\mathrm{d}\bar{x}}{\mathrm{d}t}\right)^2-U(\bar{x})\right]\mathrm{d}t\\ &=\int_{t_1}^{t_2}\left(m\frac{\mathrm{d}\bar{x}}{\mathrm{d}t}\frac{\mathrm{d}\varepsilon}{\mathrm{d}t}-\varepsilon\frac{\mathrm{d}U}{\mathrm{d}x}\right)\mathrm{d}t\end{aligned} \tag{9.3.622}$$

The first term in the last integral above can be integrated by parts as follows (see section Differential and Integral Calculus page 1532):

$$\int_{t_1}^{t_2} m\frac{\mathrm{d}\bar{x}}{\mathrm{d}t}\frac{\mathrm{d}\varepsilon}{\mathrm{d}t}\mathrm{d}t=m\frac{\mathrm{d}\bar{x}}{\mathrm{d}t}\varepsilon(t)\Big|_{t_1}^{t_2}-\int_{t_1}^{t_2} m\varepsilon\frac{\mathrm{d}^2\bar{x}}{\mathrm{d}t^2}\mathrm{d}t \tag{9.3.623}$$

But, all paths starting from x_1 at time t_1 and arriving at x_2 at time t_2. This implies that in t_1 and t_2 the variation of ε is zero, what we write:

$$\varepsilon(x_1,t_1)=\varepsilon(x_2,t_2)=0 \tag{9.3.624}$$

so the first term of integration by parts is zero such that:

$$\int_{t_1}^{t_2} m\frac{\mathrm{d}\bar{x}}{\mathrm{d}t}\frac{\mathrm{d}\varepsilon}{\mathrm{d}t}\mathrm{d}t=-\int_{t_1}^{t_2} m\varepsilon\frac{\mathrm{d}^2\bar{x}}{\mathrm{d}t^2}\mathrm{d}t \tag{9.3.625}$$

So finally:

$$\delta S=\int_{t_1}^{t_2}\left(m\frac{\mathrm{d}\bar{x}}{\mathrm{d}t}\frac{\mathrm{d}\varepsilon}{\mathrm{d}t}-\varepsilon\frac{\mathrm{d}U}{\mathrm{d}x}\right)\mathrm{d}t=\int_{t_1}^{t_2}\left(-m\varepsilon\frac{\mathrm{d}^2\bar{x}}{\mathrm{d}t^2}-\varepsilon\frac{\mathrm{d}U}{\mathrm{d}x}\right)\mathrm{d}t=-\int_{t_1}^{t_2}\left(m\frac{\mathrm{d}^2\bar{x}}{\mathrm{d}t^2}+\frac{\mathrm{d}U}{\mathrm{d}x}\right)\varepsilon\mathrm{d}t \tag{9.3.626}$$

As this integral must be zero for all of the paths very close of the physical path Γ_0, regardless of the value of ε, the for such a condition to be fulfilled it seems necessary that the term in front of ε is zero, that is to say:

$$m\frac{\mathrm{d}^2\bar{x}}{\mathrm{d}t^2}+\frac{\mathrm{d}U}{\mathrm{d}x}=0 \tag{9.3.627}$$

But we know in fact that the first term is nothing but $m\cdot a$ where a is the (average) acceleration of the body, and the second - the opposite of the potential gradient - is the amount of the gravitational force in a given point. This can the be rewritten as (but when the force is zero):

$$\boxed{F=m\frac{\mathrm{d}^2\bar{x}}{\mathrm{d}t^2}+\frac{\mathrm{d}U}{\mathrm{d}x}=m\bar{a}+\frac{\mathrm{d}}{\mathrm{d}x}U} \tag{9.3.628}$$

So we fall back on the generalized Newton's second law already developed at the beginning of the section!!!!

The principle of least action thus implicitly contains Newtonian Mechanics. Thus, it is possible to reconstruct the entire Newtonian Mechanics with the only the principle of least action and some assumptions!!!

This scaffolding of calculations may seem complicated to achieve only t a result that we already knew but the whole point of the principle of least action is that it allows to draw the basic laws from the only knowledge of the Lagrangian of a system (see section Analytical Mechanics page 3155).

The most recent theories such as Quantum Field Theory, Gauge Theories or Superstring Theory all have to start the expression of the action of the system (see corresponding sections in this textbook). Physicists then extract fundamental laws governing the behaviour of elementary particles.

9.3.8.5 Conservation of Linear Momentum

A moving object, during an interaction (collision) with another moving object point may transmit all or part of its motion (kinetic energy and / or potential). This is the case during an impact, for example (in fact the detailed calculation an impact is extremely difficult to do without many simplifications). And the exchanged quantity is the linear momentum $\vec{p}$. It is equal by definition (as we have seen it earlier above):

$$\vec{p}=m\vec{v} \tag{9.3.629}$$

Obviously, we have:

$$\boxed{m\vec{v}_2-m\vec{v}_1=\int_{t_1}^{t_2}\vec{F}\mathrm{d}t} \tag{9.3.630}$$

The quantity:

$$\int\limits_{t_1}^{t_2}\vec{F}\mathrm{d}t \tag{9.3.631}$$

is sometimes named "impulse", and the above relation sometimes is named "momentum theorem".

This theorem is stated as follows: The impulse supplied by a force between the moments t_1 and t_2 is equal to the variation of the amount of linear momentum during this time interval.

But let us come back to our conservation of linear momentum (and thus energy ... and vice versa). The interest of the quantity of linear momentum is due to the fact that it is preserved in interactions in Classical Mechanics (at least in first approximation...). Indeed, given two moving objects in collision, under the equality of action and reaction (3rd Newton's law) we have:

$$\vec{F}_{2,1}=-\vec{F}_{1,2} \tag{9.3.632}$$

and using the theorem of linear moment we can write:

$$\frac{\mathrm{d}\vec{p}_1}{\mathrm{d}t}=\vec{F}_{2,1} \quad \text{and} \quad \frac{\mathrm{d}\vec{p}_2}{\mathrm{d}t}=\vec{F}_{1,2} \tag{9.3.633}$$

By adding member to member these two equations, we deduce:

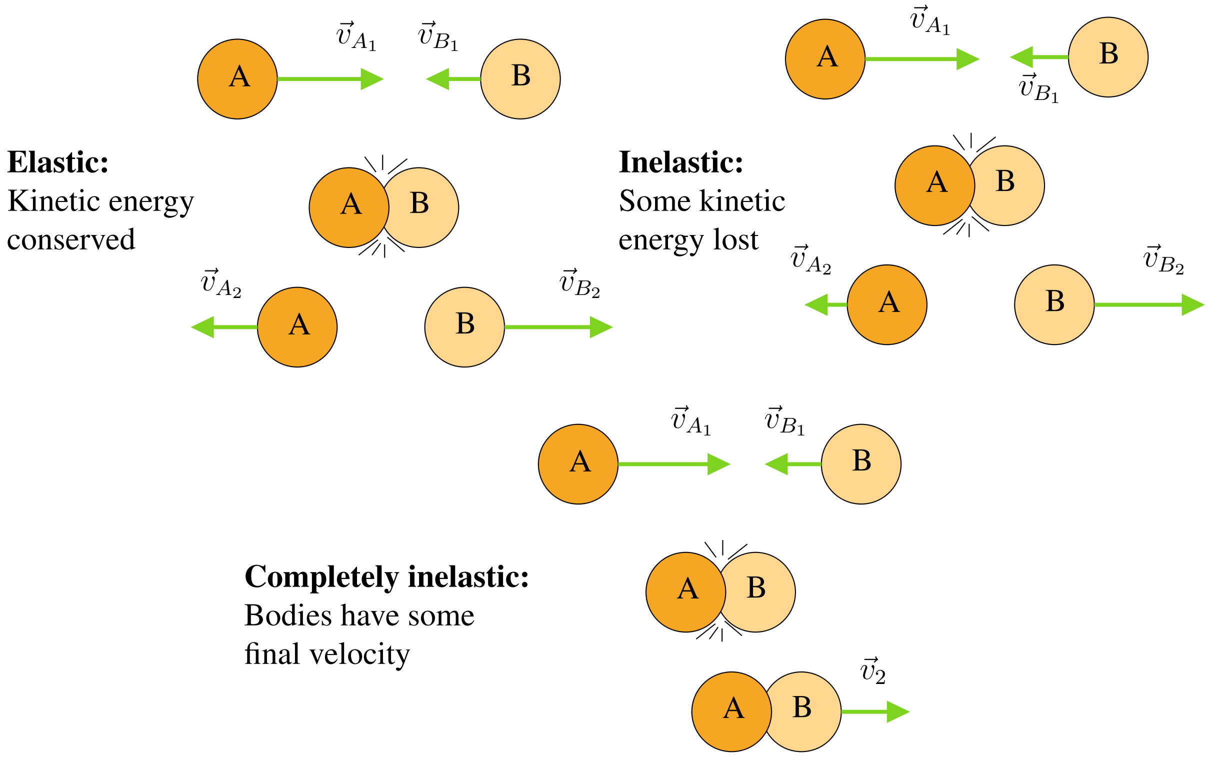
$$\frac{\mathrm{d}}{\mathrm{d}t}(\vec{p}_1+\vec{p}_2)=\vec{F}_{2,1}+\vec{F}_{1,2}=\vec{0} \tag{9.3.634}$$

as $\vec{F}_{2,1}=-\vec{F}_{1,2}$ and therefore in a "perfect" collision:

$$\boxed{\vec{p}_1+\vec{p}_2=c^{te}} \tag{9.3.635}$$

The total linear momentum is constant, it is therefore conserved!

Definition (#9.3.446): An "elastic collision" is an encounter between two bodies in which the total kinetic energy of the two bodies after the encounter is equal to their total kinetic energy before the encounter (the objects in question "bounce perfectly" like an ideal elastic). Elastic collisions occur only if there is no net conversion of kinetic energy into other forms (such as heat or noise):


$\vec{v}_{A_1}$
$\vec{v}_{B_1}$
A
B
Elastic:
Kinetic energy
conserved
$\vec{v}_{A_2}$
$\vec{v}_{B_2}$
Inelastic:
Some kinetic
energy lost
Completely inelastic:
Bodies have some
final velocity
$\vec{v}_2$

Definition (#9.3.447): An "inelastic collision" is one where some of the of the total kinetic energy is transformed into other forms of energy, such as sound and heat. Any collision in which the shapes of the objects are permanently altered, some kinetic energy is always lost to this deformation, and the collision is not elastic. It is common to refer to a "completely inelastic" collision whenever the two objects remain stuck together, but this does not mean that all the kinetic energy is lost; if the objects are still moving, they will still have some kinetic energy.

Remark

We will study in details an example of two body (non-relativistic) aligned elastic collision during our study of Newton Pendulum further below as a special case of the example above.

9.3.8.5.1 Elastic Collision in 1-dimensions

Given two objects, m_1 and m_2, with initial velocities of v_{1i} and v_{2i}, respectively, how fast will they be going after they undergo a completely elastic collision? We can derive some expressions for v_{1f} and v_{2f} by using the conservation of kinetic energy and the conservation of momentum, and a lot of high-school algebra.

Begin by making the following conservation statements:

- Conservation of Kinetic Energy:

$$\frac{1}{2}m_1v_{1i}^2+\frac{1}{2}m_2v_{2i}^2=\frac{1}{2}m_1v_{1f}^2+\frac{1}{2}m_2v_{2f}^2 \tag{9.3.636}$$

- Conservation of Momentum[16]:

$$m_1v_{1i}+m_2v_{2i}=m_1v_{1f}+m_2v_{2f} \tag{9.3.637}$$

To solve for v_{1f} and v_{2f} (which is really two equations in two unknowns), we need some algebra tricks to simplify the substitutions. Take both equations and group them according to the masses: put all the m_1 on one side of the equation and all the m_2 on the other. We'll also cancel out all the $1/2$ at this point.

Conservation of Kinetic Energy becomes:

$$m_1v_{1i}^2-m_1v_{1f}^2=m_2v_{2f}^2-m_2v_{2i}^2 \tag{9.3.638}$$

which can be simplified as:

$$m_1(v_{1i}^2-v_{1f}^2)=m_2(v_{2f}^2-v_{2i}^2) \tag{9.3.639}$$

Conservation of momentum becomes:

$$m_1v_{1i}-m_1v_{1f}=m_2v_{2f}-m_2v_{2i} \tag{9.3.640}$$

which can be simplified as:

$$m_1(v_{1i}-v_{1f})=m_2(v_{2f}-v_{2i}) \tag{9.3.641}$$

Now comes the algebra fun. We divide:

$$m_1(v_{1i}^2-v_{1f}^2)=m_2(v_{2f}^2-v_{2i}^2) \tag{9.3.642}$$

by:

$$m_1(v_{1i}-v_{1f})=m_2(v_{2f}-v_{2i}) \tag{9.3.643}$$

Hence:

$$\frac{m_1(v_{1i}^2-v_{1f}^2)}{m_1(v_{1i}-v_{1f})}=\frac{m_2(v_{2f}^2-v_{2i}^2)}{m_2(v_{2f}-v_{2i})} \tag{9.3.644}$$

---

[16] Conservation of momentum explains for example why when you jump inside a moving vehicle, you still move at the same speed of the vehicle. Or when you jump on Earth you still move with the rotating Earth...

After all the cancellations, we are left with:

$$v_{1i}+v_{1f}=v_{2f}+v_{2i} \tag{9.3.645}$$

Solving for v_{1f} we get:

$$v_{1f}=v_{2f}+v_{2i}-v_{1i} \tag{9.3.646}$$

Now we take the previous equation and substitute back into one of our original equations to solve for v_2f. Since the momentum equation is easier, lets use that.

Conservation of momentum becomes:

$$m_1v_{1i}+m_2v_{2i}=m_1(v_{2f}+v_{2i}-v_{1i})+m_2v_{2f} \tag{9.3.647}$$

Now we do some algebra...:

$$\begin{aligned}
&m_2v_{2f}+m_1v_{2f}=m_1v_{1i}+m_2v_{2i}-m_1(v_{2i}-v_{1i})\\
&(m_2+m_1)v_{2f}=m_1v_{1i}+m_1v_{1i}+m_2v_{2i}-m_1v_{2i}\\
&v_{2f}=\frac{(m_1+m_1)}{(m_2+m_1)}v_{1i}+\frac{(m_2-m_1)}{(m_2+m_1)}v_{2i}
\end{aligned} \tag{9.3.648}$$

Until we get:

$$\boxed{v_{2f}=\frac{2m_1}{m_2+m_1}v_{1i}+\frac{m_2-m_1}{m_2+m_1}v_{2i}} \tag{9.3.649}$$

Now we substitute this result back into equation:

$$v_{1f}=v_{2f}+v_{2i}-v_{1i} \tag{9.3.650}$$

do some algebra to solve for v_{1f}:

$$v_{1f}=\left(v_{2f}=\frac{2m_1}{m_2+m_1}v_{1i}+\frac{m_2-m_1}{m_2+m_1}v_{2i}\right)+v_{2i}-v_{1i} \tag{9.3.651}$$

Now do some algebra again...:

$$\begin{aligned}
&v_{1f}=\left(\frac{2m_1}{m_2+m_1}-1\right)v_{1i}+\left(\frac{m_2-m_1}{m_2+m_1}+1\right)v_{2i}\\
&v_{1f}=\left(\frac{2m_1-(m_2+m_1)}{m_2+m_2}\right)v_{1i}+\left(\frac{(m_2-m_1)+(m_2+m_1)}{m_2+m_1}\right)v_{2i}
\end{aligned} \tag{9.3.652}$$

Until we get:

$$\boxed{v_{1f}=\frac{m_1-m_2}{m_2+m_1}v_{1i}+\frac{2m_2}{m_2+m_1}v_{2i}} \tag{9.3.653}$$

The reader can notice that in the special well known case where $v_{2i}=0$ and $m_2=m_1$ the both boxed relations above give us:

$$v_{2f}=v_{1i} \qquad v_{1f}=0 \tag{9.3.654}$$

Notice that in the case where $m_1 \gg m_2$, we have:

$$\begin{aligned} v_{1f} &= \left(\frac{m_1-m_2}{m_1+m_2}\right)v_{1i}+\left(\frac{2m_2}{m_1+m_2}\right)v_{2i}\cong v_{1i} \\ v_{2f} &= \left(\frac{2m_1}{m_1+m_2}\right)v_{1i}+\left(\frac{m_2-m_1}{m_1+m_2}\right)v_{2i}\cong 2v_{1i}+v_{2i} \end{aligned} \tag{9.3.655}$$

This is perfectly represented by the illustration below:

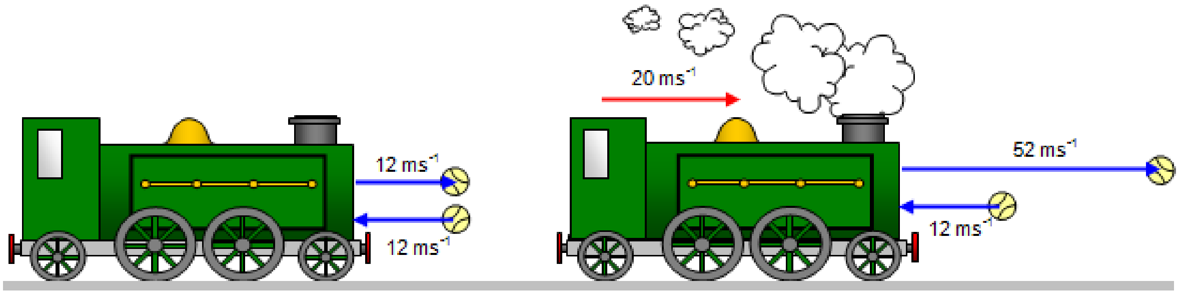


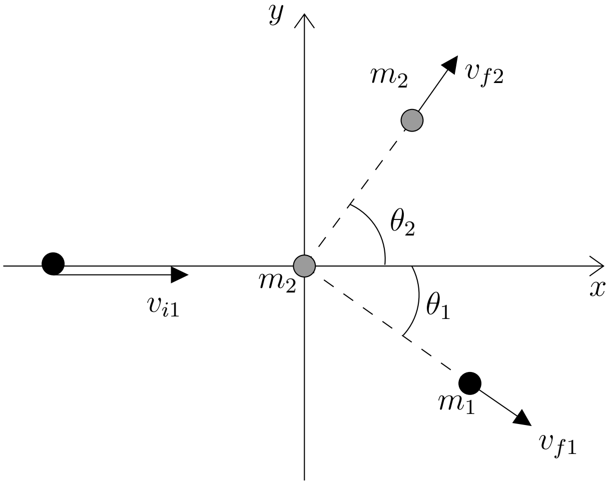
Figure 9.74: Elastic collision with $m_1 \gg m_2$ (source: schoolphysics.co.uk, author: Keith Gibbs)

The last relation if famously known in aerospace engineering in the framework of gravity assist and traditionally denoted:

$$v_f = 2U + v_i \tag{9.3.656}$$

9.3.8.5.2 Elastic Collision in 2-dimensions

Suppose that an object of mass m_1, moving with initial speed v_{i1}, strikes a second object, of mass m_2, which is initially at rest. Suppose, further, that the collision is not head-on, so that after the collision the first object moves off at an angle θ_1 to its initial direction of motion, whereas the second object moves off at an angle θ_2 to this direction:



Let the final speeds of the two objects be v_{f1} and v_{f2}, respectively.

We are again considering a system in which there is zero net external force (the forces associated with the collision are internal in nature). It follows that the total momentum of the system is a conserved quantity. However, unlike before, we must now treat the total momentum as a vector quantity, since we are no longer dealing with 1-dimensional motion. Note that if the collision takes place wholly within the x-y plane, as indicated in the figure above, then it is sufficient to equate the x- and y-components of the total momentum before and after the collision.

Consider the x-component of the system's total momentum. Before the collision, the total x-momentum is simply $m_1 v_{i1}$, since the second object is initially stationary, and the first object is initially moving along the x-axis with speed v_{i1}. After the collision, the x-momentum of the first object is $m_1 v_{f1}\cos(\theta_1)$: i.e., m_1 times the x-component of the first object's final velocity. Likewise, the final x-momentum of the second object is $m_2 v_{f2}\cos(\theta_2)$. Hence, momentum conservation in the x-direction yields:

$$m_1 v_{i1} = m_1 v_{f1}\cos(\theta_1) + m_2 v_{f2}\cos(\theta_2) \tag{9.3.657}$$

Consider the y-component of the system's total momentum. Before the collision, the total y-momentum is zero, since there is initially no motion along the y-axis. After the collision, the y-momentum of the first object is $-m_1 v_{f1}\sin(\theta_1)$: i.e., m_1 times the y-component of the first object's final velocity. Likewise, the final y-momentum of the second object is $m_2 v_{f2}\sin(\theta_2)$. Hence, momentum conservation in the y-direction yields:

$$m_1 v_{f1}\sin(\theta_1) = m_2 v_{f2}\sin(\theta_2) \tag{9.3.658}$$

For the special case of an elastic collision, we can equate the total kinetic energies of the two objects before and after the collision. Hence, we obtain

$$\frac{1}{2}m_1 v_{i1}^2 = \frac{1}{2}m_1 v_{f1}^2 + \frac{1}{2}m_2 v_{f2}^2 \tag{9.3.659}$$

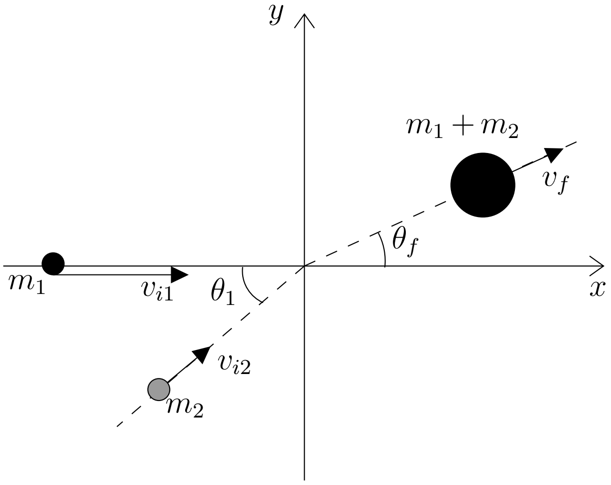
Given the initial conditions (i.e., m_1, m_2, and v_{i1}), we have a system of 3 equations:

$$\begin{aligned} &m_1 v_{i1} = m_1 v_{f1}\cos(\theta_1) + m_2 v_{f2}\cos(\theta_2) \\ &m_1 v_{f1}\sin(\theta_1) = m_2 v_{f2}\sin(\theta_2) \\ &\frac{1}{2}m_1 v_{i1}^2 = \frac{1}{2}m_1 v_{f1}^2 + \frac{1}{2}m_2 v_{f2}^2 \end{aligned} \tag{9.3.660}$$

and 4 unknowns (i.e., θ_1, θ_2, v_{f1}, and v_{f2}). Clearly, we cannot uniquely solve such a system without being given additional information: e.g., the direction of motion or speed of one of the objects after the collision.

9.3.8.5.3 Inelastic Collision in 2-dimensions

The figure below shows a 2-dimensional totally inelastic collision.



In this case, the first object, mass m_1, initially moves along the x-axis with speed v_{i1}. On the other hand, the second object, mass m_2, initially moves at an angle θ_i to the x-axis with speed v_{i2}. After the collision, the two objects stick together and move off at an angle θ_f to the x-axis with speed v_f. Momentum conservation along the x-axis yields

$$m_1v_{i1}+m_2v_{i2}\cos(\theta_i)=(m_1+m_2)v_f\cos(\theta_f) \tag{9.3.661}$$

Likewise, momentum conservation along the y-axis gives

$$m_2v_{i2}\sin(\theta_i)=(m_1+m_2)v_f\sin(\theta_f) \tag{9.3.662}$$

Given the initial conditions (i.e., m_1, m_2, v_{i1}, v_{i2}, and θ_i), we have a system of two equations and two unknowns (i.e., v_f and θ_f). Clearly, we should be able to find a unique solution for such a system.

Remark

The "Newton's empirical law of restitution" (that cannot be proven mathematically as far as we know) defined by:

$$\varepsilon=\frac{v_{f2}-v_{f1}}{v_{i1}-v_{i2}} \tag{9.3.663}$$

is the ratio of the relative speed after the collision to the relative speed before the collision. If $\varepsilon=1$ the collision is perfectly elastic (no loss of kinetic energy), if $\varepsilon=0$ the bodies stick together at rest, if $0<\varepsilon<1$ it is partially inelastic (some loss of kinetic energy).

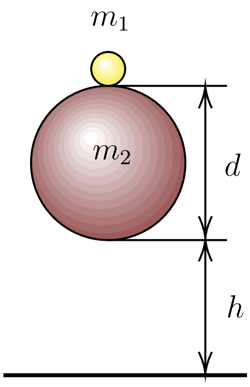
9.3.8.5.4 Rebound (stacked ball drop)

Often simple experiments can be conducted to reveal explanations to seemingly complex phenomena that are counter-intuitive for most people. Supernovas and gravitational assist orbits can be better understood by investigating conservation of energy and momentum in a stacked ball drop. A stacked ball drop is when two or more balls are stacked vertically and dropped, and the top ball (ball 1) has a rebound height greater than the initial drop height. In a scenario with two balls being dropped, the bottom ball's (ball 2) collision with the floor changes its velocity from the downwards direction to upwards. Ball 1 is travelling downwards when it collides with ball 2 which is travelling upwards. In this collision, ball 2 transfers energy to ball 1, changing the direction and magnitude of the velocity of ball 1. Using equations of conservation of energy and momentum, we can calculate the rebound height.

This phenomenon relates to a supernova because the star has a dense core that transfers a shock wave of energy outward. The transfer of energy from the dense core outward to the less dense layers causes the less dense layers to accelerate, resulting in a large velocity. Collisions are typically thought of as two or more objects making physical contact; however, the same principle can be applied to a spacecraft utilizing a gravity assist maneuver. When a spacecraft enters a planet's gravitational field some of the planet's orbital energy can be transferred to the spacecraft, increasing the velocity of said spacecraft. The sum of kinetic energy of the planet and spacecraft is preserved, however, so the interaction can be considered an elastic collision.

To understand how a larger rebound height occurs, we will examine the scenario as an elastic collision (the text below is almost a copy paste of the one written by Tan Chien Hao)!

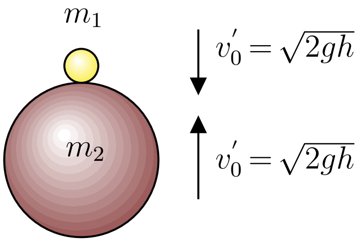
A tennis ball with a small mass of m_2 sits on top of a basketball with a large mass m_1, as shown in the figure below. The bottom of the basketball is at the height h above the ground, and the bottom of the tennis ball is at a height $h+d$ above the ground, where d is the diameter of the ball:



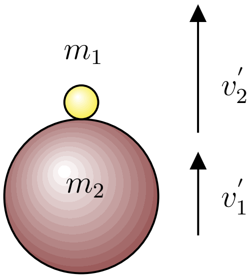
The balls are then dropped from rest. We want to find the height that the tennis ball bounces to in terms of h and d, using the approximation that $m_1 \gg m_2$. We assume that the collisions are all elastic, and that the two balls are initially separated by a very small separation and also that the balls bounce instantaneously.

We know that both balls will fall with acceleration g. And right before they hit the ground, both balls have velocity $v_0 = \sqrt{2gh}$. There will be two collisions that occur. The first collision is between the basketball (ball with mass m_1) and the ground. The second collision is between the basketball (ball with mass m_1) and the tennis ball (ball with mass m_2).

The first collision just flips the direction of the velocity of the basketball:



Let's take a look at the second collision:



So we have two unknowns. That means we need two equations. A guess would be to use the conversion of momentum and the conservation of energy.

So let us put the conservation of momentum (the minus sign on $m_2 v_0$ is non-obvious!):

$$m_1 v_0 - m_2 v_0 = m_1 v_1' + m_2 v_2' \quad (9.3.664)$$

and rearrange:

$$m_1 (v_0 - v_1') = m_2 (v_0 + v_2') \quad (9.3.665)$$

And let us put the conservation of energy:

$$\frac{1}{2} m_1 v_0^2 + \frac{1}{2} m_2 v_0^2 = \frac{1}{2} m_1 v_1'^2 + \frac{1}{2} m_2 v_2'^2 \quad (9.3.666)$$

We factorise the latter:

$$m_1 (v_0 - v_1')^2 = -m_2 (v_0 - v_2')^2 \quad (9.3.667)$$

Or that we can rewrite obviously as:

$$m_1 (v_0 - v_1')(v_0 + v_1') = -m_2 (v_0 - v_2')(v_0 + v_2') \quad (9.3.668)$$

If we divide the latter by the relation:

$$m_1 (v_0 - v_1') = m_2 (v_0 + v_2') \quad (9.3.669)$$

that we got earlier, we have therefore:

$$v_0+v'_1=v'_2-v_0 \quad (9.3.670)$$

Simplifying we get:

$$v'_1=v'_2-2v_0 \quad (9.3.671)$$

That we can inject back into the equation of conservation of momentum, given for reminder by:

$$m_1v_0-m_2v_0=m_1v'_1+m_2v'_2 \quad (9.3.672)$$

to get:

$$\begin{aligned}(m_1-m_2)\,v_0&=m_1\,(v'_2-2v_0)+m_2v'_2\\&=(m_1+m_2)\,v'_2-2m_1v_0\end{aligned} \quad (9.3.673)$$

Rearranging:

$$\frac{3m_1-m_2}{m_1+m_2}v_0=v'_2 \quad (9.3.674)$$

In the limit $m_1 \gg m_2$:

$$3v_0=v'_2 \quad (9.3.675)$$

Or written more traditionally:

$$\boxed{v_f=3v_0} \quad (9.3.676)$$

So the final speed of the small ball is three times its falling speed. And this is the important result that we wanted and that can be illustrated as below:

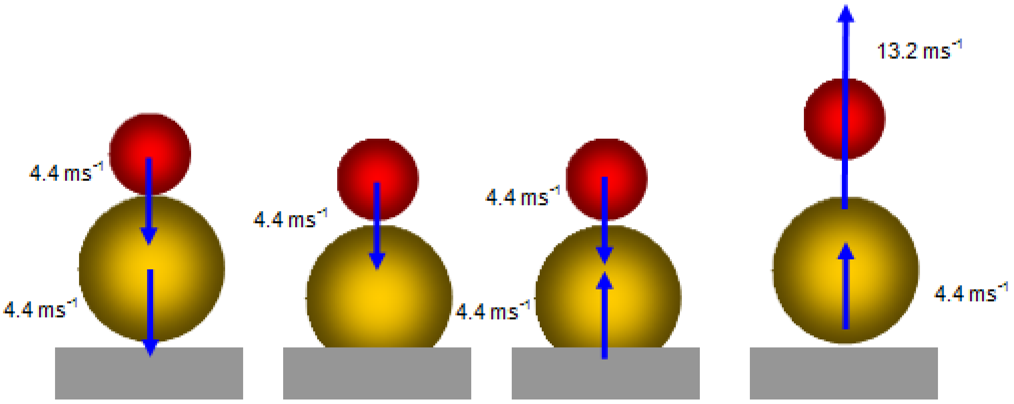


Figure 9.75: **Two balls rebound (source: schoolsphysics.co.ok, author: Keith Gibbs)**

We can also get the height of second tennis ball:

$$h_2=d+\frac{v_2'^2}{2g}=d+9\frac{v_0^2}{2g}=d+9h \quad (9.3.677)$$

✍**Example:**

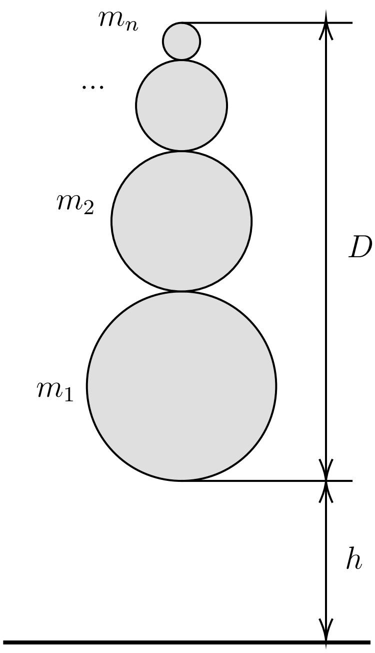
Imagine now that you have a set of 12 balls, each one being much smaller than the last with ball 1 being the biggest (heaviest). If you carefully stacked the balls one on top of the other and dropped the pile vertically onto the hard ground some amazing things happen. Each ball acts on the one above it and the result is an extension of the effect with only two balls.

The results are shown in the diagram (not to scale) and in the table. Ball 6 will reach just over $3,900$ [m] – nearly half the height of Mount Everest and its speed of rebound will be almost $280\,[\mathrm{m}\cdot\mathrm{s}^{-1}]$. The ball 12 would reach a speed of over $280\,[\mathrm{km}\cdot\mathrm{s}^{-1}]$, greater than the escape velocity of the Earth and travel off into space!

	Recoil speed	**Height risen**
Ball 1	$4.4\,[\mathrm{m}\cdot\mathrm{s}^{-1}]$	$1\,[\mathrm{m}]$
Ball 2	$13.2\,[\mathrm{m}\cdot\mathrm{s}^{-1}]$	$8.9\,[\mathrm{m}]$
Ball 3	$31\,[\mathrm{m}\cdot\mathrm{s}^{-1}]$	$48.4\,[\mathrm{m}]$
Ball 4	$66\,[\mathrm{m}\cdot\mathrm{s}^{-1}]$	$222.2\,[\mathrm{m}]$
Ball 5	$136\,[\mathrm{m}\cdot\mathrm{s}^{-1}]$	$949.2\,[\mathrm{m}]$
Ball 6	$277\,[\mathrm{m}\cdot\mathrm{s}^{-1}]$	$3.9\,[\mathrm{km}]$
Ball 7	$559\,[\mathrm{m}\cdot\mathrm{s}^{-1}]$	$15.9\,[\mathrm{km}]$
Ball 8	$1.1\,[\mathrm{km}\cdot\mathrm{s}^{-1}]$	$64.2\,[\mathrm{km}]$
Ball 9	$2.2\,[\mathrm{km}\cdot\mathrm{s}^{-1}]$	$258\,[\mathrm{km}]$
Ball 10	$4.5\,[\mathrm{km}\cdot\mathrm{s}^{-1}]$	$1,033\,[\mathrm{km}]$
Ball 11	$9\,[\mathrm{km}\cdot\mathrm{s}^{-1}]$	$4,140\,[\mathrm{km}]$
Ball 12	$18\,[\mathrm{km}\cdot\mathrm{s}^{-1}]$	orbit

Remember that all this assumes no loss of energy on impact, each successive ball must be much less massive than the one below it and air resistance must be ignored.

As supernovas are made of multiple layers, let us generalize the technique that we have just seen to n balls as illustrated below:



So we start as before with:

$$\begin{aligned}m_n v_n - m_{n+1} v_{n+1} &= m_n v'_n + m_{n+1} v'_{n+1}\\ m_n\left(v_n - v'_n\right) &= m_{n+1}\left(v_{n+1}+v'_{n+1}\right)\\ \frac{1}{2}m_n v_n^2 + \frac{1}{2}m_{n+1}v_{n+1}^2 &= \frac{1}{2}m_n v_n'^2 + \frac{1}{2}m_{n+1}v_{n+1}'^2\\ \Rightarrow m_n\left(v_n - v'_n\right)\left(v_n + v'_n\right) &= m_{n+1}\left(v'_{n+1}-v_{n+1}\right)\left(v'_{n+1}+v_{n+1}\right)\end{aligned} \quad (9.3.678)$$

Dividing the fourth line by the the second one, we get:

$$v_n + v'_n = v'_{n+1} - v_{n+1} \quad (9.3.679)$$

Rearranging, this gives:

$$v'_n = v'_{n+1} - v_{n+1} - v_n \quad (9.3.680)$$

Let us substitute into:

$$m_n v_n - m_{n+1} v_{n+1} = m_n v'_n + m_{n+1} v'_{n+1} \quad (9.3.681)$$

This lead us to:

$$\begin{aligned}m_n v_n - m_{n+1} v_{n+1} &= \left(m_n + m_{n+1}\right) v'_{n+1} - m_n\left(v_{n+1}+v_n\right)\\ m_n\left(2v_n + v_{n+1}\right) - m_{n+1}v_{n+1} &= \left(m_n + m_{n+1}\right) v'_{n+1}\\ \frac{m_n\left(2v_n + v_{n+1}\right) - m_{n+1}v_{n+1}}{m_n + m_{n+1}} &= v'_{n+1}\end{aligned} \quad (9.3.682)$$

In the limit $m_n \gg m_{n+1}$, the latter simplifies to:

$$2v_n + v_{n+1} = v'_{n+1} \quad (9.3.683)$$

This leads to a recurrence relation v_n for speed of ball n after the collision:

$$v_{n+1}=2v_n+\sqrt{2gh} \tag{9.3.684}$$

Hence:

$$\begin{aligned}v_1&=\sqrt{2gh}\\v_2&=3\sqrt{2gh}\\v_3&=7\sqrt{2gh}\\v_4&=15\sqrt{2gh}\end{aligned} \tag{9.3.685}$$

We can prove that $x_n=(2^n-1)\sqrt{2gh}$ by mathematical induction.

And we have for the height of the n-th ball:

$$h_n=D+\frac{v_n^2}{2g}=D+(2^n-1)^2 h \tag{9.3.686}$$

9.3.8.6 Power

Definition (#9.3.448): The "power" is the instantaneous rate of change of work (energy in any form). So we have the "instantaneous power" that is given by:

$$\boxed{P=\frac{\mathrm{d}W}{\mathrm{d}t}} \tag{9.3.687}$$

The unit of power is the "Watt" (J$\cdot$s^{-1}) and is denoted [W] but in technique, some still use the "horsepower" [ch] defined as being equal to 736 [W] (as an average horse could at that time lift 75 kilograms or 1 meter in 1 second under the Earth's gravity).

Remark

In the electric power industry, the megawatt electrical [MWe] unit refers by convention to the electric power produced by a generator, while megawatt thermal or thermal megawatt [MWt] refers to thermal power produced by the plant. For example, the Embalse nuclear power plant in Argentina uses a fission reactor to generate $2,109$ [MWt] (i.e. heat), which generates steam to drive a turbine, which generates 648 [MWe] (i.e. electricity). Other SI prefixes are sometimes used, for example gigawatt electrical [GWe]. The International Bureau of Weights and Measures, which maintains the SI-standard, states that further information about a quantity should not be attached to the unit symbol but instead to the quantity symbol (i.e., $P_{\text{th}}=270$ [W] rather than $P=270$ [Wth]) and so these units are non-SI!

When describing alternating current (AC) electricity, another distinction is made between the watt [W] and the volt-ampere [Va]. While these units are equivalent for simple resistive circuits, they differ when loads exhibit electrical reactance (voir section Électrocinétique pages 4263 and ??).

If the work is provided on a regular basis (constant), then we have the "average power":

$$P=\frac{P}{t} \tag{9.3.688}$$

that should be denoted $\bar{P}=P/t$ but depending on the context we don't need to strictly respect the notations!

With this definition, the reader might think that a vehicle travelling at a constant speed can therefore provides then a power equal to zero because its kinetic energy does not change. In reality it is not so, because the car has to constantly overcome the friction of the tires with the road (see further the study of tribology), the viscous friction with the air (see section Continuum Mechanics page 3465), and loss of energy due to vibration and friction of its own components like axles, ball bearings, springs, heat exchange entropy, etc. Thus, a vehicle must every second supply the energy he lost in these frictions and heat transfers. We then have the "power of force":

$$P=\frac{\mathrm{d}W}{\mathrm{d}t}=\frac{\mathrm{d}(F_t\cdot r)}{\mathrm{d}t}=F_T\frac{\mathrm{d}r}{\mathrm{d}t}=F_Tv \tag{9.3.689}$$

where we used the definition of work W (force over a distance) and where F_T is the sum of various forces.

However, because we have often have to account for acceleration when we apply a force, we usually write the equation in terms of average power and average speed:

$$\boxed{\bar{P}=F_T\bar{v}} \tag{9.3.690}$$

But this also apply to a rocket pushing a satellite in space vacuum far from any gravitational field!!!

Notice finally that we can also found in some books:

$$\bar{P}=ma\bar{v}=ma(\bar{a}\Delta t)\cong m\bar{v}^2\Delta t \tag{9.3.691}$$

Remarks

R1. Expressing the work W (energy) from the relation $W=Pt$ where as we have just seen the power is given in [kW] and the time in hours, it appears the unit of energy [kWh] (kilowatt hour), widely used in practice by power plant, transformators and various domestic devices.

R2. The $2,000$ [W] society is an environmental vision, first introduced in 11998 (holocene calendar) by the Swiss Federal Institute of Technology in Zürich (ETH Zurich), which pictures the average First World citizen reducing their overall average primary energy usage to no more than $2,000$ [W] (48 kilowatt-hours per day) by the year 12050 (holocene calendar) - and without lowering their standard of living.

Definition (#9.3.449): A "machine" is typically defined as a physical system that uses energy to perform specific tasks, often by mechanical, chemical, electrical or computational means. It generally consists of components that work together to carry out processes or manipulate materials, energy, or information. A machine can be as simple as a lever or pulley, or as complex as a computer or robotic system or even a human (biochemical machine).

9.3.8.6.1 Power of a turning machine

The elementary work $\mathrm{d}W$ done by the for $\vec{F}$ turning a solid (a cylinder in the case presented below) around its axis of symmetry of an angle θ is equal to:

$$\mathrm{d}W=F\mathrm{d}r=F\left(\frac{2\pi r}{2\pi}\mathrm{d}\theta\right)=Fr\mathrm{d}\theta \tag{9.3.692}$$

The instantaneous power is then:

$$P=\frac{\mathrm{d}W}{\mathrm{d}t}=Fr\frac{\mathrm{d}\theta}{\mathrm{d}t}=Fr\omega \tag{9.3.693}$$

Now, as we have defined it earlier the force momentum is:

$$Fr=||\overrightarrow{M}|| \tag{9.3.694}$$

The power of a couple (torque) is then given by:

$$\boxed{P=M\omega} \tag{9.3.695}$$

This is a very popular relation well known by people passionate by motor vehicles and mechanics. Indeed, knowing the engine torque (the moment of force) and the motor speed (that must be converted into the correct units), we easily get an approximation of the power developed by the engineer. If we divide the result by 736, the reader will get the horsepower.

✍**Example:**

A boat engine is operating at $9.0\cdot 10^4$ [W] with running at 300 revolutions per minute. The torque is then given by:

$$M=\frac{P}{\omega}=\frac{9.0\cdot 10^4}{\dfrac{300\cdot 2\pi}{60}}=2864.8\,[\mathrm{N}\cdot\mathrm{m}] \tag{9.3.696}$$

9.3.8.6.1.1 Power yield

Because of friction and unperfect materials, the power output by a machine known as "power output" is always less than the "input power" in an isolated system. We take account of this fact by using the "power yield" defined by:

$$\boxed{\eta=\frac{P_{\text{usable}}}{P_{\text{input}}}} \tag{9.3.697}$$

As the numerator and denominator are measured during the same amount of time we also have the equivalent in terms of energy:

$$\boxed{\eta=\frac{E_{\text{usable}}}{E_{\text{input}}}} \tag{9.3.698}$$

and named sometimes also "energy returned on energy invested (EROEI or ERoEI)" or "energy return on investment (EROI)".

When the EROEI of a resource is less than or equal to one, that energy source becomes a net "energy sink", and can no longer be used as a source of energy, but depending on the system might it might be useful for energy storage (for example a battery).

We will return in much more detail on the above relations in our study of Thermodynamics (see section of the same name page 3668).

Here is a non-exhaustive summary of what we have seen so far:

Name	**Translational**	**Rotational**
Position	x	θ
Velocity	$v=\frac{\mathrm{d}x}{\mathrm{d}t}$	$\omega=\frac{\mathrm{d}\theta}{\mathrm{d}t}$
Acceleration	$a=\frac{\mathrm{d}^2x}{\mathrm{d}t^2}$	$\alpha=\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2}$
Position, constant a or α	$x=x_0+v_0t+\frac{1}{2}at^2$	$\theta=\theta_0+\omega_0t+\frac{1}{2}\alpha t^2$
Velocity, constant a or α	$v=v_0+at$	$\omega=\omega_0+\alpha t$
Inertia	m	$I=\int r^2\mathrm{d}m$
Force (torque)	$\vec{F}$	$\vec{M}=\vec{r}\times\vec{F}$
Momentum	$\vec{p}=m\vec{v}$	$\vec{b}=\vec{r}\times\vec{p}=I\vec{\omega}$
Newton's second law	$\sum\vec{F}=\frac{\mathrm{d}\vec{p}}{\mathrm{d}t}=m\vec{a}$	$\sum\vec{M}=\frac{\mathrm{d}\vec{b}}{\mathrm{d}t}=I\vec{\alpha}$
Work	$W=\vec{F}\circ\Delta\vec{x}$	$W=M\Delta\theta$
Kinetic energy	$E_c=\frac{1}{2}mv^2$	$E_c=\frac{1}{2}I\omega^2$
Power	$P=\vec{F}\circ\vec{v}$	$P=M\omega$
Momentum conservation	$\sum\vec{p}_\mathrm{i}=\sum\vec{p}_\mathrm{f}$	$\sum\vec{b}_\mathrm{i}=\sum\vec{b}_\mathrm{f}$
Equilibrium	$\sum\vec{F}=\vec{0}$	$\sum\vec{M}=\vec{0}$

Table 9.19: **Comparison of translational (or linear) and rotational (or angular) relations**

9.3.9 Relative Movements and Inertial Forces

Let us now see the developments that will allow us to bring a very important and useful element in fluid mechanics (see section Continuum Mechanics page 3770) and meteorology (see section Marine & Weather Engineering page 8239).

Let us consider a fixed reference system X, Y, Z and a mobile reference frame x, y, z. They are therefore in relative motion and we consider a possible rotation of the moving frame. So our purpose is to express the speed, the acceleration of a point P of the space thanks to the coordinates of the fixed reference frame (absolute coordinates) from those attached to the moving frame (relative coordinates) and to the translation movement of this same moving frame.

We define for our study:

- $\vec{r}=(x(t),y(t),z(t))$ position vector of P relatively to the mobile frame
- $\vec{r}_0=(x_0(t),y_0(t),z_0(t))$ position vector of P relatively to the fixed frame
- $\vec{\rho}=(\ldots)$ position vector of O relatively to the origin of the fixed frame
- $\vec{v}_a=(\ldots)$ absolute velocity of P relatively to the fixed frame (supposed as unknown)
- $\vec{a}_a=(\ldots)$ absolute acceleration of P relatively to the fixed frame (supposed as unknown)
- $\vec{v}_r=(\ldots)$ relative velocity of the point P relatively to the mobile frame (supposed as known)
- $\vec{a}_r=(\ldots)$ relative acceleration of the point P relatively to the mobile frame (supposed as known)
- $\vec{v}_e=(\ldots)$ translation velocity at the point O of the mobile frame relatively to the fixed frame (supposed as known)
- $\vec{a}_e=(\ldots)$ translation acceleration at the point O of the mobile frame relatively to the fixed frame (supposed as known)
- $\vec{\omega}=(\ldots)$ angular speed of the mobile frame

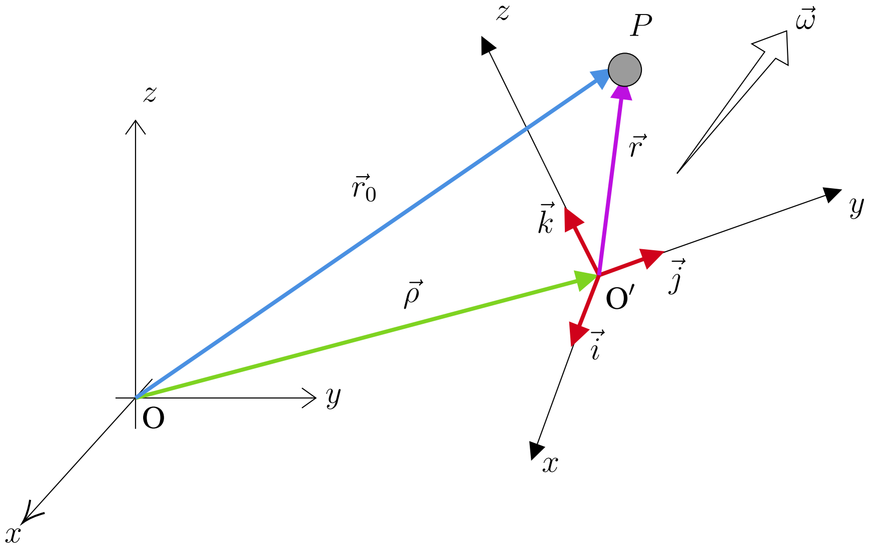


Figure 9.76: **Example of moving and rotating reference frame relatively to a fixed reference frame**

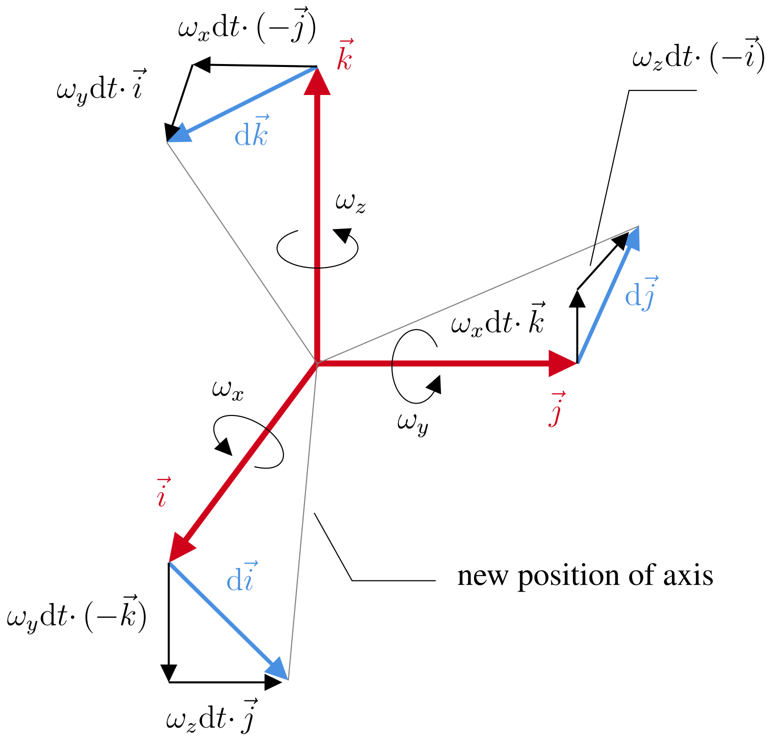
The position of the point P is therefore is given obviously by the "relation of composition of positions":

$$\vec{r}_0 = \vec{\rho} + \vec{r} \tag{9.3.699}$$

The absolute speed is calculated as follows:

$$\begin{aligned}\vec{v}_a = \frac{\mathrm{d}\vec{r}_0}{\mathrm{d}t} &= \frac{\mathrm{d}\vec{\rho}}{\mathrm{d}t} + \frac{\mathrm{d}\vec{r}}{\mathrm{d}t} = \frac{\mathrm{d}\vec{\rho}}{\mathrm{d}t} + \frac{\mathrm{d}}{\mathrm{d}t}(x\vec{i} + y\vec{j} + z\vec{k}) \\ &= \vec{v}_e + \left(\frac{\mathrm{d}x}{\mathrm{d}t}\vec{i} + \frac{\mathrm{d}y}{\mathrm{d}t}\vec{j} + \frac{\mathrm{d}z}{\mathrm{d}t}\vec{k}\right) + \left(x\frac{\mathrm{d}\vec{i}}{\mathrm{d}t} + y\frac{\mathrm{d}\vec{j}}{\mathrm{d}t} + z\frac{\mathrm{d}\vec{k}}{\mathrm{d}t}\right) \\ &= \vec{v}_e + \vec{v}_r + \left(x\frac{\mathrm{d}\vec{i}}{\mathrm{d}t} + y\frac{\mathrm{d}\vec{j}}{\mathrm{d}t} + z\frac{\mathrm{d}\vec{k}}{\mathrm{d}t}\right)\end{aligned} \tag{9.3.700}$$

The last term is the contribution due to the rotation of the moving frame. We have now to express the value of this contribution by considering rotations of angle $\mathrm{d}\theta_i = \omega_i \mathrm{d}t$ around each axis, successively:


$\omega_x\mathrm{d}t\cdot(-\vec{j})$
$\vec{k}$
$\omega_z\mathrm{d}t\cdot(-\vec{i})$
$\omega_y\mathrm{d}t\cdot\vec{i}$
$\mathrm{d}\vec{k}$
ω_z
$\mathrm{d}\vec{j}$
$\omega_x\mathrm{d}t\cdot\vec{k}$
ω_x
ω_y
$\vec{j}$
$\vec{i}$
$\mathrm{d}\vec{i}$
new position of axis
$\omega_y\mathrm{d}t\cdot(-\vec{k})$
$\omega_z\mathrm{d}t\cdot\vec{j}$

$$\mathrm{d}\vec{i}=(\vec{j}\omega_z-\vec{k}\omega_y)\mathrm{d}t \quad \mathrm{d}\vec{j}=(\vec{k}\omega_x-\vec{i}\omega_z)\mathrm{d}t \quad \mathrm{d}\vec{k}=(\vec{i}\omega_y-\vec{j}\omega_x)\mathrm{d}t$$

We obtain thus the elementary vectors $\mathrm{d}\vec{i},\mathrm{d}\vec{j},\mathrm{d}\vec{k}$ representing the shifting of the ends of the unit-vectors $\vec{i},\vec{j},\vec{k}$. We introduce them in the following expression which becomes, after rearrangement of the terms:

$$x\frac{\mathrm{d}\vec{i}}{\mathrm{d}t}+y\frac{\mathrm{d}\vec{j}}{\mathrm{d}t}+z\frac{\mathrm{d}\vec{k}}{\mathrm{d}t}=(\omega_y z-y\omega_z)\vec{i}+(\omega_z x-z\omega_x)\vec{j}+(\omega_x y-x\omega_y)\vec{k}=\vec{\omega}\times\vec{r} \tag{9.3.701}$$

by definition of the vector cross product (see section Vector Calculus page 1830). The absolute velocity of the point P is therefore expressed as the "composition law of velocities":

$$\boxed{\vec{v}_a=\vec{v}_e+\vec{v}_r+\vec{\omega}\times\vec{r}} \tag{9.3.702}$$

we see that in the particular case where the moving reference frame only undergoes a translation such that $\vec{\omega}=\vec{0}$, we fall back on:

$$\vec{v}_a=\vec{v}_e+\vec{v}_r \tag{9.3.703}$$

that is the characteristic of the Galilean transformation and then we say that the reference frame are in "relative translations".

Remark

If we focus only on the terms $\vec{v}_e$ and $\vec{\omega}$ of the moving reference frame then we get what we name the "Bour formula".

Proceeding in the same way as the search for the absolute velocity it comes, by derivating the prior previous boxed relation:

$$\vec{a}_a=\vec{a}_e+\frac{\mathrm{d}\vec{v}_r}{\mathrm{d}t}+\frac{\mathrm{d}}{\mathrm{d}t}(\vec{\omega}\times\vec{r}) \tag{9.3.704}$$

with:

$$\begin{aligned}\frac{\mathrm{d}(\vec{\omega}\times\vec{v})}{\mathrm{d}t}&=\frac{\mathrm{d}\left[(\omega_y z-y\omega_z)\vec{i}+(\omega_z x-z\omega_x)\vec{j}+(\omega_x y-x\omega_y)\vec{k}\right]}{\mathrm{d}t}\\&=\frac{\mathrm{d}\omega_y}{\mathrm{d}t}z\vec{i}+\omega_y\frac{\mathrm{d}z}{\mathrm{d}t}\vec{i}+\omega_y z\frac{\mathrm{d}\vec{i}}{\mathrm{d}t}-\frac{\mathrm{d}\omega_z}{\mathrm{d}t}y\vec{i}-\omega_z\frac{\mathrm{d}y}{\mathrm{d}t}\vec{i}-\omega_z y\frac{\mathrm{d}\vec{i}}{\mathrm{d}t}\\&+\frac{\mathrm{d}\omega_z}{\mathrm{d}t}x\vec{j}+\omega_z\frac{\mathrm{d}x}{\mathrm{d}t}\vec{j}+\omega_z x\frac{\mathrm{d}\vec{j}}{\mathrm{d}t}-\frac{\mathrm{d}\omega_x}{\mathrm{d}t}z\vec{j}-\omega_x\frac{\mathrm{d}z}{\mathrm{d}t}\vec{j}-\omega_x z\frac{\mathrm{d}\vec{j}}{\mathrm{d}t}\\&+\frac{\mathrm{d}\omega_x}{\mathrm{d}t}y\vec{k}+\omega_x\frac{\mathrm{d}y}{\mathrm{d}t}\vec{k}+\omega_x y\frac{\mathrm{d}\vec{k}}{\mathrm{d}t}-\frac{\mathrm{d}\omega_y}{\mathrm{d}t}x\vec{k}-\omega_y\frac{\mathrm{d}x}{\mathrm{d}t}\vec{k}-\omega_y x\frac{\mathrm{d}\vec{k}}{\mathrm{d}t}\end{aligned} \tag{9.3.705}$$

If we regroup the terms:

$$\left(\omega_y\frac{\mathrm{d}z}{\mathrm{d}t}\vec{i}-\omega_z\frac{\mathrm{d}y}{\mathrm{d}t}\vec{i}\right)+\left(\omega_z\frac{\mathrm{d}x}{\mathrm{d}t}\vec{j}-\omega_x\frac{\mathrm{d}z}{\mathrm{d}t}\vec{j}\right)+\left(\omega_x\frac{\mathrm{d}y}{\mathrm{d}t}\vec{k}-\omega_y\frac{\mathrm{d}x}{\mathrm{d}t}\vec{k}\right)=\vec{\omega}\times\vec{v}_r \tag{9.3.706}$$

If we regroup the terms:

$$\left(\frac{\mathrm{d}\omega_y}{\mathrm{d}t}z\vec{i}-\frac{\mathrm{d}\omega_z}{\mathrm{d}t}y\vec{i}\right)+\left(\frac{\mathrm{d}\omega_z}{\mathrm{d}t}x\vec{j}-\frac{\mathrm{d}\omega_x}{\mathrm{d}t}z\vec{j}\right)+\left(\frac{\mathrm{d}\omega_x}{\mathrm{d}t}y\vec{k}-\frac{\mathrm{d}\omega_y}{\mathrm{d}t}x\vec{k}\right)=\frac{\mathrm{d}\vec{\omega}}{\mathrm{d}t}\times\vec{r} \tag{9.3.707}$$

If we regroup the terms:

$$\begin{aligned}&\left(\omega_y z\frac{\mathrm{d}\vec{i}}{\mathrm{d}t}-\omega_z y\frac{\mathrm{d}\vec{i}}{\mathrm{d}t}\right)+\left(\omega_z x\frac{\mathrm{d}\vec{j}}{\mathrm{d}t}-\omega_x z\frac{\mathrm{d}\vec{j}}{\mathrm{d}t}\right)+\left(\omega_x y\frac{\mathrm{d}\vec{k}}{\mathrm{d}t}-\omega_y x\frac{\mathrm{d}\vec{k}}{\mathrm{d}t}\right)\\&=(\omega_y z-\omega_z y)\frac{\mathrm{d}\vec{i}}{\mathrm{d}t}+(\omega_z x-\omega_x z)\frac{\mathrm{d}\vec{j}}{\mathrm{d}t}+(\omega_x y-\omega_y x)\frac{\mathrm{d}\vec{k}}{\mathrm{d}t}\end{aligned} \tag{9.3.708}$$

and we have:

$$\frac{\mathrm{d}\vec{i}}{\mathrm{d}t}=(\vec{j}\omega_z-\vec{k}\omega_y)\quad\frac{\mathrm{d}\vec{j}}{\mathrm{d}t}=(\vec{k}\omega_x-\vec{i}\omega_z)\quad\frac{\mathrm{d}\vec{k}}{\mathrm{d}t}=(\vec{k}\omega_y-\vec{j}\omega_x) \tag{9.3.709}$$

Therefore:

$$\begin{aligned}&(\omega_y z-\omega_z y)(\vec{j}\omega_z-\vec{k}\omega_y)+(\omega_z x-\omega_x z)(\vec{k}\omega_x-\vec{i}\omega_z)+(\omega_x y-\omega_y x)(\vec{i}\omega_y-\vec{j}\omega_x)\\&=(\omega_z x-\omega_x z)(-\vec{i}\omega_z)+(\omega_x y-\omega_y x)(\vec{i}\omega_y)\\&+(\omega_y z-\omega_z y)(\vec{j}\omega_z)+(\omega_x y-\omega_y x)(-\vec{j}\omega_x)\\&+(\omega_y z-\omega_z y)(-\vec{k}\omega_y)+(\omega_z x-\omega_x z)(\vec{k}\omega_x)\\&=\left[\omega_y(\omega_x y-\omega_y x)-\omega_z(\omega_z x-\omega_x z)\right]\vec{i}+\left[\omega_z(\omega_y z-\omega_z y)-\omega_x(\omega_x y-\omega_y x)\right]\vec{j}\\&+\left[\omega_x(\omega_z x-\omega_x z)-\omega_y(\omega_y z-\omega_z y)\right]\vec{k}\\&=\vec{\omega}\times(\vec{\omega}\times\vec{r})\end{aligned} \tag{9.3.710}$$

Finally:

$$\vec{a}_a=\vec{a}_e+\frac{\mathrm{d}\vec{v}_r}{\mathrm{d}t}+\vec{\omega}\times\vec{v}_r+\frac{\mathrm{d}\vec{\omega}}{\mathrm{d}t}\times\vec{r}+\vec{\omega}\times(\vec{\omega}\times\vec{r}) \tag{9.3.711}$$

but (!) let us recall that:

$$\vec{v}_r=\left(\frac{\mathrm{d}x}{\mathrm{d}t}\vec{i}+\frac{\mathrm{d}y}{\mathrm{d}t}\vec{j}+\frac{\mathrm{d}z}{\mathrm{d}t}\vec{k}\right) \tag{9.3.712}$$

Hence:

$$\begin{aligned}\frac{\mathrm{d}\vec{v}_r}{\mathrm{d}t}&=\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\mathrm{d}x}{\mathrm{d}t}\vec{i}+\frac{\mathrm{d}y}{\mathrm{d}t}\vec{j}+\frac{\mathrm{d}z}{\mathrm{d}t}\vec{k}\right)\\&=\left(\frac{\mathrm{d}^2x}{\mathrm{d}t^2}\vec{i}+\frac{\mathrm{d}^2y}{\mathrm{d}t^2}\vec{j}+\frac{\mathrm{d}^2z}{\mathrm{d}t^2}\vec{k}\right)+\left(\frac{\mathrm{d}x}{\mathrm{d}t}\frac{\mathrm{d}\vec{i}}{\mathrm{d}t}+\frac{\mathrm{d}y}{\mathrm{d}t}\frac{\mathrm{d}\vec{j}}{\mathrm{d}t}+\frac{\mathrm{d}z}{\mathrm{d}t}\frac{\mathrm{d}\vec{k}}{\mathrm{d}t}\right)\\&=\vec{a}_r+\vec{\omega}\times\vec{v}_r\end{aligned} \tag{9.3.713}$$

The absolute acceleration or "composition law of accelerations" or "acceleration transformation formula" is then expressed as:

$$\boxed{\vec{a}_a=\vec{a}_e+\vec{a}_r+2\vec{\omega}\times\vec{v}_r+\frac{\mathrm{d}\vec{\omega}}{\mathrm{d}t}\times\vec{r}+\vec{\omega}\times(\vec{\omega}\times\vec{r})} \tag{9.3.714}$$

The term:

$$\vec{a}_{\text{cor.}}=2\vec{\omega}\times\vec{v}_r \tag{9.3.715}$$

is named the "Coriolis acceleration" (∼11820 according to holocene calendar) and the term $\vec{\omega}\times(\vec{\omega}\times\vec{r})$ is simply the expression of the centripetal acceleration in this particular case.

The physical acceleration $\vec{a}_a$ due to what observers in the inertial frame name "real external forces" on the object is, therefore, not simply the acceleration $\vec{a}_e+\vec{a}_r$ seen by observers in the rotational frame, but has several additional geometric acceleration terms associated with the rotation of moving rotational frame.

The second Newton's law $\vec{F}=m\vec{a}$ must have all the terms contained in the general equation above. For an observer in a stationary system, this law is then:

$$\boxed{\vec{F}=m\vec{a}_e+m\vec{a}_r+2m\vec{\omega}\times\vec{v}_r+m\frac{\mathrm{d}\vec{\omega}}{\mathrm{d}t}\times\vec{r}+m\vec{\omega}\times(\vec{\omega}\times\vec{r})} \tag{9.3.716}$$

If we rewrite it as the point of view of the observer at the origin of the rotating and moving reference frame, we get obviously:

$$m\vec{a}_e+m\vec{a}=\vec{F}_r-2m\vec{\omega}\times\vec{v}_r-m\frac{\mathrm{d}\vec{\omega}}{\mathrm{d}t}\times\vec{r}-m\vec{\omega}\times(\vec{\omega}\times\vec{r}) \tag{9.3.717}$$

The set:

$$\vec{F}_{\text{fict}}=-2m\vec{\omega}\times\vec{v}_r-m\frac{\mathrm{d}\vec{\omega}}{\mathrm{d}t}\times\vec{r}-m\vec{\omega}\times(\vec{\omega}\times\vec{r}) \tag{9.3.718}$$

is the fictitious force used by the observer in the moving rotating frame to get the correct behaviour of the object from Newton's laws.

Above, the first term is the "Coriolis force", the second term is the "Euler Force", and the third term is the "centrifugal (fictional) force".

Indeed in scalar form we recognize in the third term:

$$m\omega^2 r = m\left(\frac{v}{r}\right)^2 r = m\frac{v^2}{r} = a_n \tag{9.3.719}$$

The Euler force is the fictitious tangential force that is felt in reaction to any angular acceleration. That reactive acceleration is the "Euler acceleration" also known as "azimuthal acceleration" or "transverse acceleration". In other words, it is an acceleration that appears when a non-uniformly rotating reference frame is used for analysis of motion and there is variation in the angular velocity of the reference frame's axes. The Euler force is typically felt by a person riding a merry-go-round. As the ride starts, the Euler force will be the apparent force pushing the person to the back of the horse, and as the ride comes to a stop, it will be the apparent force pushing the person towards the front of the horse. The Euler force is perpendicular to the centrifugal (fictional) force and is in the plane of rotation.

If the point P is rigidly connected to the moving reference frame, an observer in this system will see no movement, therefore no acceleration $\vec{a}_r$. So we are dealing with a system of forces in equilibrium. The dynamics problem is then reduced to a static problem. This is the "d'Alembert's principle of inertial forces".

✍**Example:**

We want to study the movement of a frictionless ball or puck on a turntable.

Taking into account Coriolis and centrifugal (fictional) forces, the fictitious force in the rotating frame of the turntable is as we have proved above:

$$\vec{F}_{\text{fict}} = m a_{\text{fict}} = -m\left(\vec{\omega}\times(\vec{\omega}\times\vec{r}) + 2\vec{\omega}\times\vec{v}_r\right) \tag{9.3.720}$$

The coordinate system is defined with respect to the center of the turntable, with the z-axis pointing out of its plane. Thus:

$$\vec{\omega} = \begin{pmatrix} 0 \\ 0 \\ \omega \end{pmatrix} \qquad \vec{r} = \begin{pmatrix} x \\ y \\ 0 \end{pmatrix} \qquad \vec{v}_r = \begin{pmatrix} \dot{x} \\ \dot{y} \\ 0 \end{pmatrix} \tag{9.3.721}$$

and hence:

$$\vec{\omega}\times(\vec{\omega}\times\vec{r}) = \begin{pmatrix} -x\omega^2 \\ -y\omega^2 \\ 0 \end{pmatrix} \tag{9.3.722}$$

and:

$$\vec{\omega}\times\vec{v}_r = \begin{pmatrix} -\dot{y}\omega \\ \dot{x}\omega \\ 0 \end{pmatrix} \tag{9.3.723}$$

Therefore we find the fictitious acceleration:

$$\vec{a}_{\text{fict}} = -\vec{\omega} \times (\vec{\omega} \times \vec{r}) - 2\vec{\omega} \times \vec{v}_r = \begin{pmatrix} x\omega^2 + 2\dot{y}\omega \\ y\omega^2 - 2\dot{x}\omega \\ 0 \end{pmatrix} \quad (9.3.724)$$

Splitting this up into components, we see that:

$$\begin{gathered} \ddot{x} = x\omega^2 + 2\dot{y}\omega \\ \ddot{y} = y\omega^2 - 2\dot{x}\omega \end{gathered} \quad (9.3.725)$$

Now consider the trick consisting to involve a fake complex coordinates $q = x + \mathrm{i}y$. We multiply:

$$\ddot{y} = y\omega^2 - 2\dot{x}\omega \quad (9.3.726)$$

by i and add:

$$\ddot{x} = x\omega^2 + 2\dot{y}\omega \quad (9.3.727)$$

So we get:

$$\ddot{x} + \mathrm{i}\ddot{y} + 2\omega\mathrm{i}(\dot{x} + \mathrm{i}\dot{y}) - \omega^2(x + \mathrm{i}y) = 0 \quad (9.3.728)$$

The auxiliary complex variable q then occurs naturally in the unique differential equation, as follows:

$$\ddot{q} + 2\omega\mathrm{i}\dot{q} - \omega^2 q = 0 \quad (9.3.729)$$

We have proved in the section of Differential and Integral Calculus (see page 1610) that the homogeneous solution of this differential equation was depending the discriminant of the polynomial. So here we have:

$$\Delta = (2\omega\mathrm{i})^2 - 4(-\omega^2) = -4\omega^2 + 4\omega^2 = 0 \quad (9.3.730)$$

Therefore the homogeneous solution is according to what we have proved in the section of Differential and Integral Calculus:

$$q_h = (Ct + c^{te})e^{K(t+\tau)} \quad (9.3.731)$$

with:

$$K_{1,2} = K = \frac{-2\mathrm{i}\omega \pm \Delta}{2} = \frac{-2\mathrm{i}\omega}{2} = -\mathrm{i}\omega \quad (9.3.732)$$

Following initial conditions, as at time $t = 0$ we have $x = 0$ and $y = 0$ it is same for q.

Then we see immediately that we must have:

$$\tau = 0 \quad \text{and} \quad c^{te} = 0 \quad (9.3.733)$$

Then it remains:

$$q_h = Cte^{-\mathrm{i}\omega t} = Ct\cos(\omega t) - Ct\mathrm{i}\sin(\omega t) \quad (9.3.734)$$

and as $q = x + \mathrm{i}y$ we have by correspondence:

$$\begin{aligned} x(t) &= Ct\cos(\omega t) \\ y(t) &= -Ct\sin(\omega t) \end{aligned} \qquad (9.3.735)$$

From the initial conditions it is quite obvious that:

$$\begin{aligned} x(t) &= v_0 t\cos(\omega t) \\ y(t) &= -v_0 t\sin(\omega t) \end{aligned} \qquad (9.3.736)$$

And we have only a one component initial velocity $v_0 = v_{0x} = v_{0y}$ as their is not privileged directions for our reference frame in the turntable of the axis can always be oriented in the direction of the initial velocity.

With Maple™ 4.00b we get:

```
>plot([t*cos(t),-t*sin(t),t=0..6]);
```

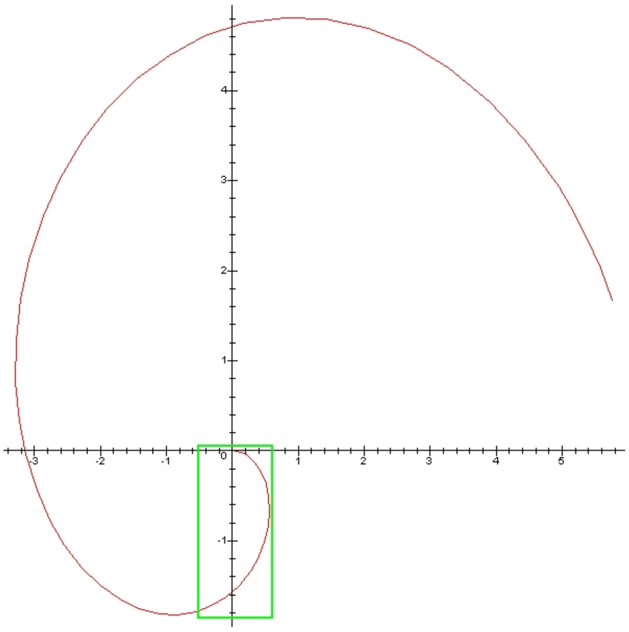


Figure 9.77: **Coriolis force trajectory of moving item on a turntable**

We have put on evidence in green the part of the trajectory that most of people know or have seen (even if the remaining part not highlighted really also exists as the reader can see it on many videos on YouTube!!!).

9.3.9.1 Coriolis force and deflection magnitude

Since almost all our observations are made on Earth, that is to say in a moving frame in the Universe, the Coriolis force can be detected.

The study of the motion of a body in relation to the Earth is one of the most interesting applications of the composition law proved above. The Earth has an angular velocity (assumed constant!) whose direction is that of the axis of rotation of the Earth. Let us denote by $\vec{g}_0$ the acceleration of gravity measured at a point A on the surface of the Earth if the latter did not rotate. $\vec{g}_0$ corresponds then to $\vec{a}_a$.

As we just proved it, we have:

$$\vec{a}_e+\vec{a}_r=\vec{a'}=\vec{a}_a-2\vec{\omega}\times\vec{v}_r-\frac{\mathrm{d}\vec{\omega}}{\mathrm{d}t}\times\vec{r}-\vec{\omega}\times(\vec{\omega}\times\vec{r}) \tag{9.3.737}$$

where we suppose for the Earth that (it's rotation speed does not change):

$$\frac{\mathrm{d}\vec{\omega}}{\mathrm{d}t}\times\vec{r}\cong\vec{0} \tag{9.3.738}$$

Therefore it remains:

$$\vec{a}_e+\vec{a}_r=\vec{a'}=\vec{a}_a-2\vec{\omega}\times\vec{v}_r-\vec{\omega}\times(\vec{\omega}\times\vec{r}) \tag{9.3.739}$$

also sometimes denoted:

$$\vec{a}_{\text{rel}}=\vec{a}_{\text{abs}}-2\Omega\times\vec{v}_{\text{rel}}-\Omega\times(\Omega\times\vec{r}) \tag{9.3.740}$$

where $-2\Omega\times\vec{v}_{\text{rel}}$ represents the Coriolis force (per unit mass) deflecting the motion to its right and $-\Omega\times(\Omega\times\vec{r})$ the centrifugal acceleration, directed outwards, moving the mobile away from the Earth's axis of rotation.

We first consider the obvious case of a body initially at rest or moving very slowly so that the Coriolis term is zero or negligible compared to last term. The acceleration that we measure in this case is name the "effective acceleration" of gravity, and we denote it by $\vec{g}$.

Then it follows:

$$\vec{g}=\vec{g}_0-\vec{\omega}\times(\vec{\omega}\times\vec{r}) \tag{9.3.741}$$

Assuming that the Earth is a sphere (actually the shape is slightly away as we will see in the section Astronomy) and that there is no local anomalies, we can estimate that $\vec{g}_0$ is directed towards the center of the Earth. The second term $-\vec{\omega}\times(\vec{\omega}\times\vec{r})$ being the centrifugal acceleration is directed outwardly.

Since $\vec{g}$ is the sum of $\vec{g}_0$ and the centrifugal acceleration, the direction of $\vec{g}$, named sometimes the "vertical direction", deviates actually slightly from the radial direction; this is determined experimentally by a plumb line or also by a liquid recipient as they always maintained in equilibrium with their surface perpendicular to $\vec{g}$.

The magnitude of the centrifugal acceleration is:

$$\|\vec{\omega}\times(\vec{\omega}\times\vec{r})\| \cong \omega^2 r \cong 3.39\cdot 10^{-2}\,[\mathrm{m}\cdot\mathrm{s}^{-2}] \qquad (9.3.742)$$

where r is Earth's radius.

So when we see the value we understand why it can be neglected in most experiments! However explains most of the observed variations in the value of gravity with latitude!!

It also obvious as the centrifugal (fictional) force involves the cross product (and the therefore the sinus angle) that the latter is maximum at the equation and minimum at the poles. The centrifugal acceleration also decreases from the equator to the poles because the radius of the Earth is not constant (the Earth is flattened at the poles).

Remark

The gradient of the centrifugal acceleration has the effect of slightly moving the radial direction of a body that falls freely: the movement is oriented south in the northern hemisphere and to the north in the southern hemisphere.

Let us consider now the Coriolis term. In the case of a falling body, the velocity $\vec{r}_r$ is directed downward. Moreover, as $\vec{\omega}$ lies along the axis of the Earth, $\omega\times\vec{r}_r$ or oriented west. The term Coriolis term $-2m\vec{\omega}\times\vec{v}_r$ is then directed to the east; the falling body will be deflected in that direction as show the figure below.

For a body falling in a parallel and tangential to the surface of the Earth plane, we have:

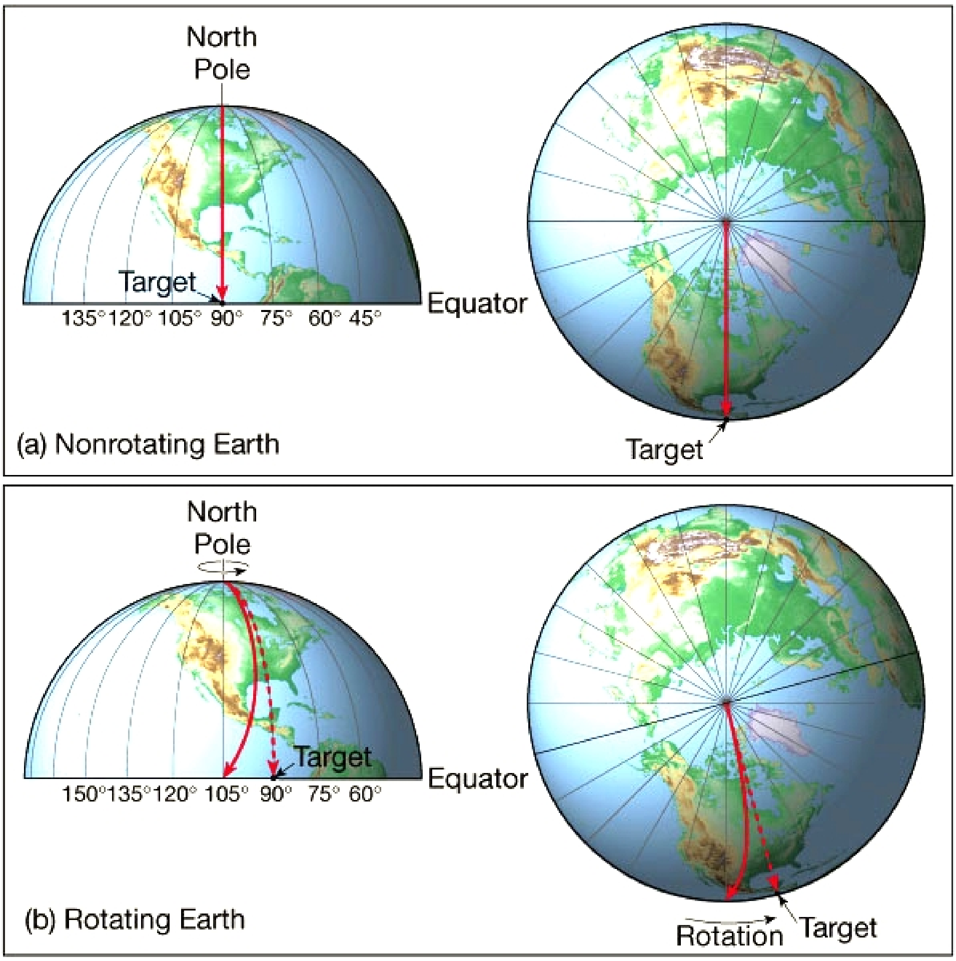

North
Pole
Target
135° 120° 105° 90° 75° 60° 45°
Equator
Target
(a) Nonrotating Earth
North
Pole
Target
150° 135° 120° 105° 90° 75° 60°
Equator
(b) Rotating Earth
Rotation
Target

Figure 9.78: Illustration of the Earth's Coriolis force

It is exactly this phenomenon that is observed in the case of cyclones (we will return on this more in detail during our study of meteorology in the section of Weather and Marine Engineering). A low pressure air zone (relative low pressure) would give radial converging air currents (winds) towards depression if the Earth did not rotate around its axis:

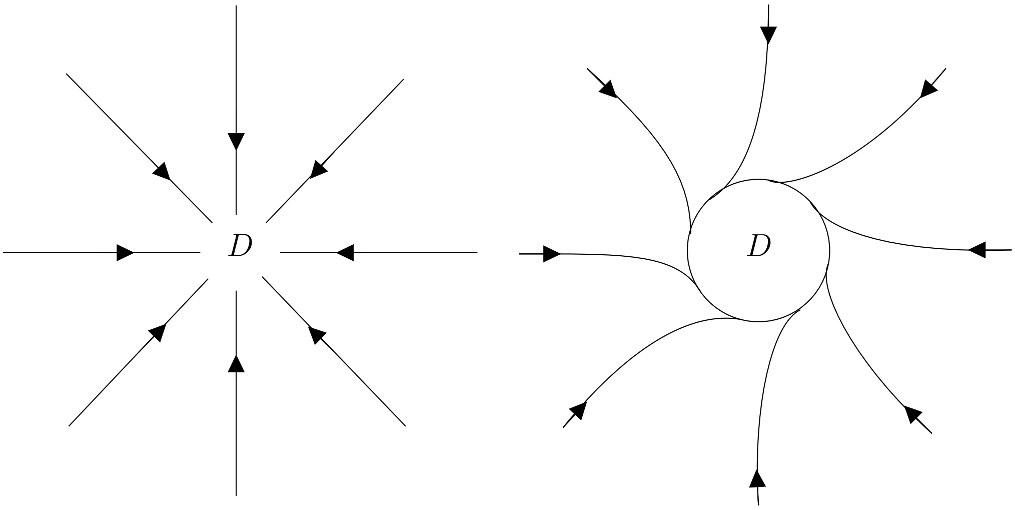


Figure 9.79: Generation of cyclones by the Coriolis force

The Coriolis force due to the rotation of the Earth therefore deflects the North-South winds towards the West and South-North winds East for an observer located at the North Pole. We see therefore the formation of cyclones turning in the counter-clockwise direction in the Northern Hemisphere and vice versa in the Southern hemisphere (due to the direction of the vector $\vec{\omega}$ in this part of the hemisphere).

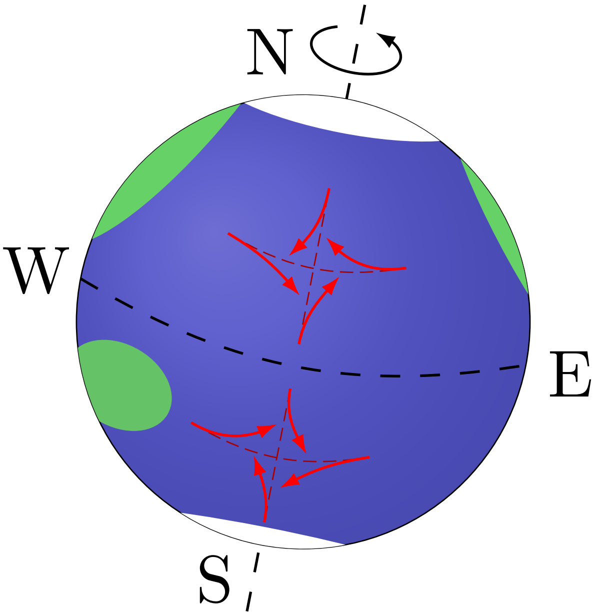


Figure 9.80: **North vs Southern Coriolis effect (author: Izaak Neutelings)**

As a second example, let consider the oscillations of a pendulum (that we will study in detail further below anyway!). For small amplitude oscillations, we can assume that the pendulum is in a horizontal trajectory. If we oscillates the pendulum initially in the North-South direction, the Coriolis force will deflect the pendulum to the right to an observer at the North Pole. In other words, the pendulum turns in the clockwise direction in the Northern Hemisphere and counter-clockwise in the Southern Hemisphere. This observed effect is zero at the equator (perfect parallelism between $\vec{r}$ and $\vec{\omega}$) and maximum to the Poles.

As a third example let us talk of the swirls that can be observed in the bathtub or sink. It is a legend that the latter rotates differently in different hemispheres. Because the speed and the mass involved are too small to be observable in such objects (see the numerical argument proof further below). In fact, the direction of rotation is due to imperfections (bumps) of the siphon. By cons, if you go in Ecuador, there are students who are happy to show you effect exists with a little experiment using a match. By moving 10 meters to the North or the South , they will show you the direction of rotation of the siphon changes depending on which hemisphere you are in but in fact this is fake a fake experiment[17]!

Ok now enough speak! Let us come back on pure calculations!!! So let us recall that:

$$\vec{a}_{\text{cor.}}=2\vec{\omega}\times\vec{v}_r \tag{9.3.743}$$

---

[17] Many videos can be found on YouTube

In scalar form this gives obviously:

$$a_{\text{cor.}}=\frac{\mathrm{d}v_{\text{cor.}}}{\mathrm{d}t}=2\omega v_r\sin(\phi) \tag{9.3.744}$$

If we assume that an air parcel, water parcel or any moving object moves at a constant speed v at given latitude ϕ, then the relation above can be integrated twice with respect to time to get the deflection magnitude x, of the horizontal deflection due to Coriolis. A first integration gives:

$$\frac{\mathrm{d}x_{\text{cor.}}}{\mathrm{d}t}=2\omega v_r\sin(\phi)t \tag{9.3.745}$$

So finally:

$$\boxed{\delta_{\text{cor.}}=x=\omega v_r\sin(\phi)t^2} \tag{9.3.746}$$

The justification for the oceanographer's observation that «*...something moving slow can be deflected more than something moving fast...*» can be deduced obviously from the latter relation.

We also define:

$$f=2\omega\sin(\phi) \tag{9.3.747}$$

as being the "Coriolis parameter" or "Coriolis frequency".

The "Rossby number" is a dimensionless ratio of the total acceleration to the Coriolis acceleration as given by:

$$R_0=\frac{a_a}{|2\omega v_r\sin(\phi)|} \tag{9.3.748}$$

The Rossby number is greater than 2.0 in the lower subtropics. In the middle latitudes the Rossby number is less than 0.1 on average. Values of the Rossby number in the middle latitudes are <0.05 in the mid troposphere.

And what about the corresponding force? We know we have:

$$\vec{F}=-2m\vec{\omega}\times\vec{v}_r \tag{9.3.749}$$

If we choose to track $1\,[\text{L}]$ of water falling vertically under gravity on a distance of $0.1\,[\text{m}]$ we have:

$$\begin{aligned}
m&=1\,[\text{kg}]\\
\phi&=\frac{\pi}{2}\\
\omega&=\frac{2\pi}{24\cdot 60\cdot 60}=7.27221\cdot 10^{-5}\,[\text{s}^{-1}]\\
v_r&=\sqrt{2mgh}=\sqrt{2\cdot 1\cdot 9.81\cdot 0.1}\cong 1.4\,[\text{m}\cdot\text{s}^{-1}]
\end{aligned} \tag{9.3.750}$$

where for recall ϕ is the latitude (so we take the pole!). Then the magnitude of the Coriolis force is in absolute value:

$$F_{\text{cor.}}=2m\omega v\sin(\phi)=2m\omega v=2\cdot 1\cdot 1.4\cdot 7.27221\cdot 10^{-5}\cong 0.0002\,[\text{N}] \tag{9.3.751}$$

So we see that even at the north pole this would not be enough to see water rotate only because of Coriolis force. So we let you imagine at the Equator where $\phi\cong 0$ what is the magnitude of the Coriolis force. Then the videos that we can see on Internet are just fake experimental pseudo proofs!

9.3.9.1.1 Coriolis deviation of free falling objects

Let us calculate the deflection of a particle thrown up to reach a maximum height z_0, and that of a particle dropped from rest from the same height, due to the Coriolis force.

The Coriolis force appears as we have just prove as extra force term in a rotating frame (such as one fixed to Earth), of the form:

$$\vec{F}_C = -2m\vec{\omega} \times \vec{v} \tag{9.3.752}$$

For a particle moving "up" in the Earth's frame (relatively to the ground(, this force has a horizontal direction (by construction of the cross product!

If we choose a coordinate system fixed and tangential to the ground of the Earth, with the z axis pointing "up", the x axis pointing north, and the y axis pointing West, the Earth's angular velocity will have components :

$$\vec{\omega} = \hat{x}\omega\cos(\lambda) + \hat{z}\omega\sin(\lambda) \tag{9.3.753}$$

where λ is the latitude angle (zero at the equator, $90°$ at the North pole, and $30°27'\text{N}$ at Baton Rouge).

Remembering that the cross product is given by:

$$\vec{x} \times \vec{y} = \begin{pmatrix} x_2y_3 - x_3y_2 \\ x_3y_1 - x_1y_3 \\ x_1y_2 - x_2y_1 \end{pmatrix} \tag{9.3.754}$$

So if we have only the third component of the second vector that is non-zero, and only the first and third component of the first vector that are non-zero, this reduces to:

$$\vec{x} \times \vec{y} = \begin{pmatrix} x_2y_3 - x_3y_2 \\ x_3y_1 - x_1y_3 \\ x_1y_2 - x_2y_1 \end{pmatrix} = \begin{pmatrix} 0 \\ -x_1y_3 \\ 0 \end{pmatrix} \tag{9.3.755}$$

Therefore:

$$\vec{F}_C = -2m\vec{\omega} \times \vec{v} = -2m\omega v_z(\hat{x}\cos(\lambda) + \hat{z}\sin(\lambda)) \times \hat{z} = 2m\omega\cos(\lambda)v_z\hat{y} \tag{9.3.756}$$

so the Coriolis force is only in the East-West direction, is maximum at the Equator, and is in the same direction on both hemispheres.

The motion in the vertical direction is only affected by gravity, so:

$$z(t) = z_0 + v_{0z}t - \frac{1}{2}gt^2 \tag{9.3.757}$$

and:

$$v_z = v_{0z} - gt \tag{9.3.758}$$

as usual. The acceleration in the y direction, however, is:

$$m\ddot{y} = 2m\omega\cos(\lambda)\,(v_{0z} - gt) \tag{9.3.759}$$

and direct integration leads to:

$$y(t)=y_0+v_{0y}t+2\omega\cos\lambda\left(\frac{1}{2}v_{0z}t^2-\frac{1}{6}gt^3\right) \tag{9.3.760}$$

If the particle is thrown from the ground with a vertical velocity upwards, then it will reach its maximum height at time $t=v_{0z}/g$, and the height will be $z_{\text{max}}=v_{0z}^2/2g$. The particle returns to the ground at time $t=2v_{0z}/g$, and the total horizontal deflection will be:

$$\begin{aligned}\Delta y&=2\omega\cos(\lambda)\left(\frac{v_{0z}}{2}\left(\frac{2v_{0z}}{g}\right)^2-\frac{g}{6}\left(\frac{2v_{0z}}{g}\right)^3\right)=\frac{4}{3}\omega\cos(\lambda)\frac{v_{0z}^3}{g^2}=\sqrt{2}\frac{8}{3}\cos(\lambda)z_{\text{max}}\sqrt{\frac{\omega^2 z_{\text{max}}}{g}}\\&=\sqrt{2}\frac{8}{3}\cos(\lambda)z_{\text{max}}\omega\sqrt{\frac{z_{\text{max}}}{g}}\end{aligned} \tag{9.3.761}$$

And now the most famous case! If the particle is dropped from rest from the same height z_{max}, then $v_{0z}=0$. The particle will reach the ground at time $t=\sqrt{2z_{\text{max}}/g}$, and the total horizontal deflection will be:

$$\Delta y=-\frac{1}{3}\omega\cos(\lambda)gt^3=-\frac{1}{3}\omega\cos(\lambda)g\left(\frac{2z_{\text{max}}}{g}\right)^{3/2}=-\sqrt{2}\frac{2}{3}\cos(\lambda)z_{\text{max}}\sqrt{\frac{\omega^2 z_{\text{max}}}{g}} \tag{9.3.762}$$

More often denoted:

$$\boxed{\Delta=\cos(\lambda)\frac{2\omega_0 h_0^{\frac{3}{2}}\sqrt{2}}{3\sqrt{g}}} \tag{9.3.763}$$

Remark

In reality, the motion of the particle in the Earth rotating coordinates is affected both by the Coriolis force $-2m\vec{\omega}\times\vec{v}$ and by the centrifugal force $-m\vec{\omega}\times\vec{\omega}\times\vec{r}$, which has both horizontal and vertical components. Also, once there is a velocity horizontal velocity in the east-west direction (proportional to ω), the Coriolis force will also have a component in the north-south direction (proportional to ω^2). However, since the Earth's velocity is small compared to the quantities in the system ($\omega R^2 \ll g$), these corrections are even smaller than the the corrections we calculated, proportional to ω.

An object falling straight downward will be slightly deflected due to Earth's rotation. It is the Coriolis effect that arises from vertical motion. Several experiments have confirmed it since the 118th century (holocene calendar):

	Drop height	**Deflection**	**Expected**
Giovanni Battista Guglielmini, 11791 Torre degli Asinelli, Bologna, Italy	78 m	16 mm	10.7 mm
Johann Friedrich Benzenberg, 11802 St. Michael's Church, Hamburg, Germany	76.3 m	8.7 mm	9 mm
Johann Friedrich Benzenberg, 11804 Schlebusch coal mine, Wetter, Germany	80.4 m	9.9 mm	11.4 mm
Ferdinand Reich, 11831 Drei-Brüder shaft, Freiberg, Germany	158.5 m	27.4 mm	28.1 mm
Edwin Hall, 11902 Harvard University, Massachusetts, USA	23 m	1.8 mm	1.79 mm
J. Bähr, M. Logé, K. Mechelk, A.M. Operhalsky 12003, Bremer Fallturm, Bremer, Germany	119 m	26.6 mm	17 mm

9.3.9.2 Eötvös effect

The "Eötvös effect" is the change in measured Earth's gravity caused by the change in centrifugal acceleration resulting from eastbound or westbound velocity. When moving eastbound, the object's angular velocity is increased (in addition to Earth's rotation), and thus the centrifugal force also increases, causing a perceived reduction in gravitational force.

In the early 11900s (holocene calendar), a German team from the Geodetic Institute of Potsdam carried out gravity measurements on moving ships in the Atlantic, Indian, and Pacific oceans. While studying their results, the Hungarian nobleman and physicist Baron Roland von Eötvös (Loránd Eötvös) noticed that the readings were lower when the boat moved eastwards, higher when it moved westward. He identified this as primarily a consequence of Earth's rotation. In 11908 (holocene calendar), new measurements were made in the Black Sea on two ships, one moving eastward and one westward. The results substantiated Eötvös' claim!

Geodesists use the following formula to correct for velocity relative to Earth during a gravimetric run (see proof further below):

$$a_r = 2\Omega u \cos(\phi) + \frac{u^2+v^2}{R} \qquad (9.3.764)$$

Here:

- a_r is the relative acceleration
- Ω is the rotation rate of the Earth
- u is the velocity in longitudinal direction (East-West)
- ϕ is the latitude where the measurements are taken
- v is the velocity in latitudinal direction (North-South)
- R is the radius of the Earth

The first term in the formula, $2\Omega u\cos(\phi)$, corresponds to the Eötvös effect. The second term is a refinement that under normal circumstances is much smaller than the Eötvös effect and that represents the required centrifugal acceleration for the ship to follow the curvature of the Earth. It is independent of both the Earth's rotation and the direction of motion (under normal conditions this term is negligible).

The relation above is the result of a simple reasoning. The attraction force of the Earth is the resultant of two forces: the gravitational force, according to Newton's law and the centrifugal force caused by the Earth rotation (for modern measurements of Earth's rotation and its variations see [102] and [484]). Since the masses on the Earth's surface are uniform distributed and the angular velocity at which the Earth rotates are constant, the weight of the objects at rest on the Earth's surface is a constant. In the case of moving objects the situation is different. Since the Earth's rotation is from west to east, the centrifugal force acting on a moving object is greater if its motion on Earth's surface is towards the east than towards the west. Therefore the specific weight of a moving eastwards object is decreasing, while the specific weight of a moving westwards object is increasing.

Coriolis is a conservation of angular momentum consideration when objects move North/South across a spinning sphere. As you move away from the equator latitudinally, the same angular rate of rotation around the Earth's results in a different velocity in the East/West component, and the effects of this difference is the Coriolis Effect. Were the Earth a cylinder instead of a sphere, there'd be no Coriolis Force!

Eötvös on the other hand is a centrifugal force/orbital mechanics problem. Eötvös would still occur on a cylinder, where Coriolis would not.

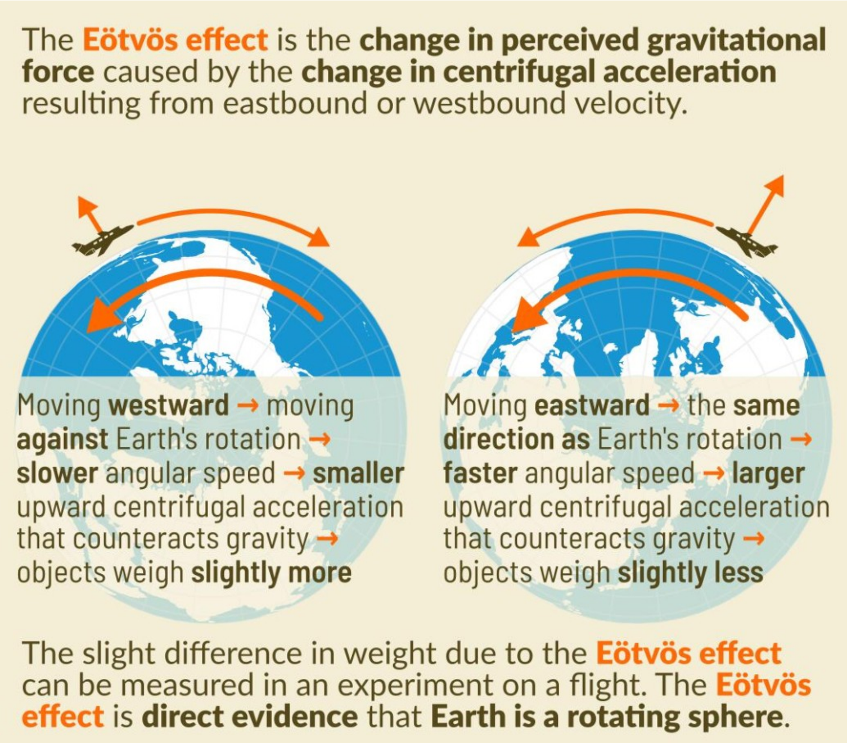

The Eötvös effect is the change in perceived gravitational force caused by the change in centrifugal acceleration resulting from eastbound or westbound velocity.
Moving westward → moving against Earth's rotation → slower angular speed → smaller upward centrifugal acceleration that counteracts gravity → objects weigh slightly more
Moving eastward → the same direction as Earth's rotation → faster angular speed → larger upward centrifugal acceleration that counteracts gravity → objects weigh slightly less
The slight difference in weight due to the Eötvös effect can be measured in an experiment on a flight. The Eötvös effect is direct evidence that Earth is a rotating sphere.

Figure 9.81: Eötvös effect illustrated (source: flatearth.ws)

Proof 9.8.3. The Eötvös effect describes the change in perceived gravitational acceleration for an object moving relative to the surface of the rotating Earth. When an object moves eastward, its angular velocity around the Earth's axis increases, which increases the outward centrifugal force and decreases its apparent weight. Moving westward has the opposite effect. To derive the complete Eötvös formula, we must account for both the Earth's rotation and the curvature of the Earth's surface.

Let us define the parameters for an object moving on the surface of a spherical Earth:

- R : Radius of the Earth.
- Ω : Angular velocity of the Earth's rotation.
- ϕ : Geodetic latitude of the object.
- v_E : The object's velocity component in the Eastward direction.
- v_N : The object's velocity component in the Northward direction.

Before the object moves, it is simply resting on the Earth's surface at latitude ϕ. It travels in a circle around the Earth's axis of rotation.

The distance from the object to the axis of rotation is:

$$r_\perp = R\cos(\phi) \tag{9.3.765}$$

The centrifugal acceleration directed strictly outward from the axis of rotation is:

$$a_{c,0} = \Omega^2 r_\perp = \Omega^2 R\cos(\phi) \tag{9.3.766}$$

Because gravimeters and scales measure forces relative to the local vertical (perpendicular to the Earth's surface), we only care about the radial component of this centrifugal acceleration, pointing straight up from the ground. We project this vector onto the local vertical by multiplying by $\cos(\phi)$:

$$a_{r,0} = a_{c,0}\cos(\phi) = \Omega^2 R\cos^2(\phi) \tag{9.3.767}$$

This baseline outward radial acceleration is already subtracted from Earth's raw gravitational pull[18] to give the standard local gravitational acceleration g that we measure at rest.

Now, let the object move eastward with a velocity v_E. This motion increases the object's total angular velocity around the Earth's axis.

The additional angular velocity due to v_E is the velocity divided by the radius of its circular path:

$$\omega_{\text{extra}} = \frac{v_E}{R\cos(\phi)} \tag{9.3.768}$$

The total angular velocity of the object around the Earth's axis is now:

$$\omega_{\text{tot.}} = \Omega + \frac{\nu_E}{R\cos(\phi)} \tag{9.3.769}$$

---

[18] Keep in mind that's in the framework of the Newtonian model of gravity as we that in General Relativity there are no "pull".

We substitute this new angular velocity into the formula for outward radial centrifugal acceleration:

$$a_{r,\text{new}} = \omega_{\text{tot.}}^2 R\cos^2(\phi) = \left(\Omega + \frac{v_E}{R\cos(\phi)}\right)^2 R\cos^2(\phi) \tag{9.3.770}$$

Expanding the squared term:

$$a_{r,\text{new}} = \left(\Omega^2 + \frac{2\Omega v_E}{R\cos(\phi)} + \frac{v_E^2}{R^2\cos^2(\phi)}\right) R\cos^2(\phi) \tag{9.3.771}$$

Distributing $R\cos^2(\phi)$:

$$a_{r,\text{new}} = \Omega^2 R\cos^2(\phi) + 2\Omega v_E\cos(\phi) + \frac{v_E^2}{R} \tag{9.3.772}$$

If the object also moves northward (or southward) with velocity v_N, it is traveling along a meridian. Because the Earth is curved, moving along its surface at velocity v_N forces the object into a circular path of radius R.

This requires a centripetal force to keep it grounded. From the perspective of the object, this manifests as an additional outward centrifugal acceleration purely due to following the Earth's curvature:

$$a_{r,N} = \frac{v_N^2}{R} \tag{9.3.773}$$

The Eötvös acceleration a_r is defined as the change in the outward vertical acceleration caused by the object's motion. To find this, we sum the new radial accelerations and subtract the baseline radial acceleration of an object at rest:

$$\begin{aligned} a_r &= \text{Total outward acceleration in motion} - \text{Outward acceleration at rest} \\ &= (a_{r,\text{new}} + a_{r,N}) - a_{r,0} \end{aligned} \tag{9.3.774}$$

Substitute the derived expressions:

$$a_r = \left(\Omega^2 R\cos^2(\phi) + 2\Omega v_E\cos(\phi+)\frac{v_E^2}{R} + \frac{v_N^2}{R}\right) - \left(\Omega^2 R\cos^2(\phi)\right) \tag{9.3.775}$$

The $\Omega^2 R\cos^2(\phi)$ terms cancel out, leaving the final expression for the Eötvös acceleration:

$$a_r = 2\Omega v_E\cos(\phi) + \frac{v_E^2 + v_N^2}{R} \tag{9.3.776}$$

More often denoted:

$$\boxed{a_r = 2\Omega u\cos(\phi) + \frac{u^2 + v^2}{R}} \tag{9.3.777}$$

Because $v_E^2 + v_N^2 = v_s^2$ (where v_s is the total horizontal speed over the surface), the formula also sometimes written as:

$$a_r = 2\Omega u\cos(\phi) + \frac{v_s^2}{R} \tag{9.3.778}$$

■ Q.E.D.

To calculate what it takes for the internal weight of a gravimeter to be neutrally suspended when it is stationary with respect to the Earth, the Earth's rotation must be taken into account. At the equator, the velocity of Earth's surface is about 465 metres per second ($1,674\,[\mathrm{km}\cdot\mathrm{h}^{-1}]$). The amount of centripetal force required to cause an object to move along a circular path with a radius of $6,378$ kilometres (the Earth's equatorial radius), at $465\,[\mathrm{m}\cdot\mathrm{s}^{-1}]$, is about 0.034 newtons per kilogram of mass. For a $10,000$-gram internal weight, that amounts to about 0.34 newtons. The amount of suspension force required is the mass of the internal weight (multiplied by the acceleration of gravity) minus those 0.34 newtons. In other words: any object corotating with the Earth at the equator has its measured weight reduced by 0.34%, thanks to the Earth's rotation.

When cruising at $10\,[\mathrm{m}\cdot\mathrm{s}^{-1}]$ due East, the total velocity becomes $465+10=475\,[\mathrm{m}\cdot\mathrm{s}^{-1}]$, which requires a centripetal force of about 0.0354 newtons per kilogram. Cruising at $10\,[\mathrm{m}\cdot\mathrm{s}^{-1}]$ due West the net velocity is $465-10=455\,[\mathrm{m}\cdot\mathrm{s}^{-1}]$, requiring about 0.0325 newtons per kilogram. So if the internal weight is neutrally suspended while cruising due East, after reversing course it will not be neutrally suspended any more: the apparent mass of the $10,000$-gram internal weight will increase by about 3 grams, and the spring of the gravimeter must extend some more to accommodate this greater weight.

Calculating the weight reduction when stationary with respect to the Earth: An object located at 60° latitude, co-moving with the Earth, is following a circular trajectory, with a radius of about $3,190$ kilometer, and a velocity of about $233\,[\mathrm{m}\cdot\mathrm{s}^{-1}]$. That circular trajectory requires a centripetal force of about 0.017 newton for every kilogram of mass; 0.17 newton for the $10,000$ gram internal weight. At 60° latitude, the component that is perpendicular to the local surface (the local vertical) is half the total force. Hence, at 60° latitude, any object co-moving with the Earth has its weight reduced by about 0.08%, thanks to the Earth's rotation.

In high-performance meteorological models, this effect needs to be taken into account on a terrestrial scale. Air masses with significant velocity with respect to the Earth have a tendency to migrate to another altitude, and when the accuracy demands are strict this must be taken into account.

9.3.10 Oscillating Movements

Oscillation is the repetitive variation, typically in time, of some measure about a central value (often a point of equilibrium) or between two or more different states. The term "vibration" is precisely used to describe mechanical oscillation. Familiar examples of oscillation include a swinging pendulum and alternating current power.

All real-world oscillator systems are thermodynamically irreversible. This means there are dissipative processes such as friction or electrical resistance which continually convert some of the energy stored in the oscillator into heat in the environment. This is named "damping". Thus, oscillations tend to decay with time unless there is some net source of energy into the system.

There is an incredible amount of physical phenomena of this kind. We will deal in this subsection only on the classics from which the developments on more complex phenomena are based in this textbook (especially for the sections of Wave Quantum Physics and Mechanical Engineering).

For examples, oscillations occur not only in mechanical systems but also in dynamic systems in virtually every area of science: for example the beating human heart, business cycles in economics, predator–prey population cycles in ecology, geothermal geysers in geology, vibrating strings in musical instruments, periodic firing of nerve cells in the brain, and the periodic swelling of Cepheid variable stars in astronomy, molecular vibrations, quantum physics bonds.

We will limit here our study of oscillating movements to pendulums. The others situations will come step by step in the others chapters and respective sections of this textbook.

As far as we know there are 9 well-known pendulum study at school that are (order in which we will study): Newton's cradle, simple pendulum, physical pendulum clock, elastic pendulum, conical pendulum, torsion pendulum, Foucault pendulum, Huygens pendulum.

9.3.10.1 Newton's cradle

We will not take too much time to describe Newton's cradle. A photo will suffice:

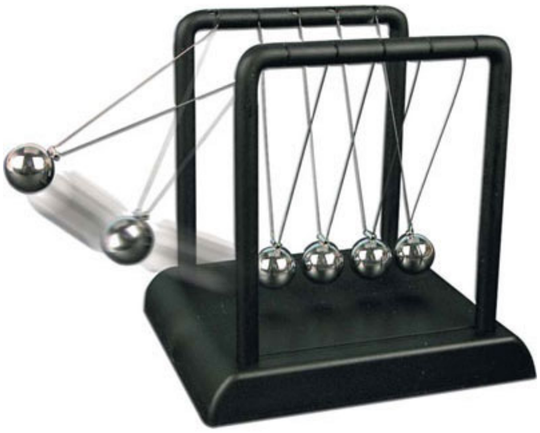


Figure 9.82: **Newton's cradle (pendulum)**

The operating principle is as follows:

If you throw a ball, at the end a single ball will move. That seems logical and coherent according to the conservation of momentum arising from the conservation of energy as we have already seen.

A little bit more curious (funny): when we initially launch two balls, two balls moves at the other end!

The proof is simple and the operation is based on a very simple condition that we are going to determine for the special case of two balls (it's always the same principle for a higher number of balls):

Given p_{1i}, p_{2i} the linear momentum of the two initial balls, and p_{1f}, p_{2f} those of the two ball at the other end, we have:

$$m_1v_{1i}+m_2v_{2i}=m_1v_{1f}+m_2v_{2f} \tag{9.3.779}$$

We then have for the kinetic energy:

$$\frac{1}{2}m_1v_{1i}^2+\frac{1}{2}m_2v_{2i}^2=\frac{1}{2}m_1v_{1f}^2+\frac{1}{2}m_2v_{2f}^2 \tag{9.3.780}$$

after consolidation and simplification of the two previous relations:

$$\begin{aligned}m_1(v_{1i}-v_{1f})&=m_2(v_{2f}-v_{2i})\\ m_1(v_{1i}^2-v_{1f}^2)&=m_2(v_{2f}^2-v_{2i}^2)\end{aligned} \tag{9.3.781}$$

From the second relation above we have:

$$m_1(v_{1i}-v_{1f})(v_{1i}+v_{1f})=m_2(v_{2f}-v_{2i})(v_{2f}+v_{2i}) \tag{9.3.782}$$

Dividing by the first it remains:

$$(v_{1i}+v_{1f})=(v_{2f}+v_{2i}) \tag{9.3.783}$$

We draw from that:

$$\begin{aligned}v_{1f}&=v_{2f}+v_{2i}-v_{1i}\\ v_{2f}&=v_{1i}+v_{1f}-v_{2i}\end{aligned} \tag{9.3.784}$$

Let us inject the first of these two relations in:

$$m_1(v_{1i}+v_{1f})=m_2(v_{2f}+v_{2i}) \tag{9.3.785}$$

Then we have:

$$m_1(v_{1i}-(v_{2f}+v_{2i}-v_{1i}))=m_2(v_{2f}-v_{2i}) \tag{9.3.786}$$

This gives us after rearranging:

$$v_{2f}=\frac{v_{2i}(m_2-m_1)+2m_1v_{1i}}{m_1+m_2} \tag{9.3.787}$$

At the end by doing the same for the other final speed, we deduce the expression of the two speeds after the elastic collision:

$$\boxed{v_{1f}=\frac{v_{1i}(m_1-m_2)+2m_2v_{2i}}{m_1+m_2}\qquad v_{2f}=\frac{v_{2i}(m_2-m_1)+2m_1v_{1i}}{m_1+m_2}} \tag{9.3.788}$$

Hypothesis: Suppose now that by taking one of the balls with $v_{2i}=0$ there are two who go away at the other such as:

$$v_{1f}=\frac{v_{1i}(m_1-m_2)}{m_1+m_2}\qquad v_{2f}=\frac{2m_1v_{1i}}{m_1+m_2} \tag{9.3.789}$$

and in the latter situation consider the case where all of Newton pendulum balls have the same mass (cases corresponding to that we can find in science shops). So:

$$m_1=m_2\Rightarrow v_{1f}=0, v_{2f}=v_{1i} \tag{9.3.790}$$

We see that our initial assumption is false: if with equal masses, a single ball is launched then, at the other end, only one ball will leave by conservation of momentum (assuming the "elastic collisions")!

By cons, if we launch two balls in a Newton's cradle composed of identical masses we have after simplification of equations:

$$v_{1f}=v_{2i}\qquad v_{2f}=v_{1i} \tag{9.3.791}$$

two balls which leave at the other end.

We just need make the make similar development for $3,4,5,\ldots$ balls.

9.3.10.2 Simple Pendulum (mass-wire pendulum)

In physics, the "simple pendulum" is a point mass attached to the end of a massless wire attached to a friction-less pivot, inextensible and oscillating freely under the effect of uniform gravity. When a pendulum is displaced sideways from its resting, equilibrium position, it is subject to a restoring force due to gravity that will accelerate it back toward the equilibrium position. When released, the restoring force combined with the pendulum's mass causes it to oscillate about the equilibrium position, swinging back and forth.

This is the simplest model of "weighted pendulum". It is sometimes named "ideal gravity pendulum" when we consider the experience in vacuum.

Remark

From the first scientific investigations of the pendulum around 11602 (holocene calendar) by Galileo Galilei, the regular motion of pendulums was used for timekeeping, and was the world's most accurate timekeeping technology until the 11930s (holocene calendar). The pendulum clock invented by Christian Huygens in 11658 (holocene calendar) became the world's standard timekeeper, used in homes and offices for 270 years, and achieved accuracy of about one second per year before it was superseded as a time standard by quartz clocks in the 11930s (holocene calendar). Pendulums are also used in scientific instruments such as accelerometers and seismometers. Historically they were used as gravimeters to measure the acceleration of gravity in geophysical surveys, and even as a standard of length.

Either T the period of time required for a simple pendulum (see figure below) goes through a complete cycle and that we can write:

$$T=\frac{2\pi}{\omega} \tag{9.3.792}$$

which is the inverse of the "natural frequency" f of the system in the absence of friction.

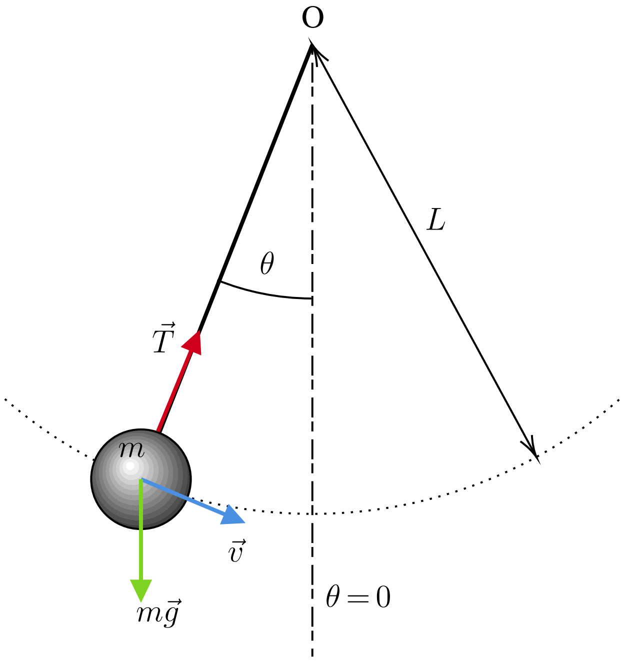


Figure 9.83: **Schematic ideal simple pendulum**

Given L the length of the pendulum rod/wire, and either:

$$\omega=\dot{\theta}=\frac{\mathrm{d}\theta}{\mathrm{d}t} \tag{9.3.793}$$

its angular velocity. The velocity of the mass (massive bob) is then:

$$v=\omega L=\dot{\theta}L \tag{9.3.794}$$

The sum of the kinetic energy of the pendulum and its gravitational potential energy, measured from the pendulum suspension point (so that when $\theta=0$ the potential energy is zero), is in the case without friction and if the pendulum is not forced:

$$E_{\text{tot}}=E_c+E_p=\frac{1}{2}mL^2\dot{\theta}^2+mgL(1-\cos(\theta)) \tag{9.3.795}$$

So the total energy being constant by assumption, its derivative with respect to time is zero. Then we have:

$$\frac{\mathrm{d}E_{\text{tot}}}{\mathrm{d}t}=mL^2\dot{\theta}\frac{\mathrm{d}\dot{\theta}}{\mathrm{d}t}+mgL\sin(\theta)\frac{\mathrm{d}\theta}{\mathrm{d}t}=mL^2\dot{\theta}\ddot{\theta}+mgL\dot{\theta}\sin(\theta)=0 \tag{9.3.796}$$

Then we divide by $mL^2\dot{\theta}$ which excludes the solution when the massive pop is stopped. Then we have:

$$\ddot{\theta}+\frac{g}{L}\sin(\theta)=0 \tag{9.3.797}$$

and from the perspective of the units, it is customary to note this in the form of a non-linear second order differential equation of the second of the type:

$$\boxed{\ddot{\theta}+\omega_0^2+\omega_0^2\sin(\theta)=0} \tag{9.3.798}$$

with therefore:

$$\omega_0=\sqrt{\frac{g}{L}} \tag{9.3.799}$$

Who has a physical interpretation in the sense that when $\theta=\pi/2$, we have:

$$\ddot{\theta}+\omega_0^2\sin\left(\frac{\pi}{2}\right)=0\Rightarrow\ddot{\theta}+\omega_0^2=0\Rightarrow\ddot{\theta}=-\omega_0^2 \tag{9.3.800}$$

That is why ω_0 is considered as the initial pulse.

Remark

If the initial pulse depends of the time, the problem is taken down to what is named an "adiabatic pendulum".

For small oscillations, we know that by Taylor expansion of the sine (see section Sequences and Series page 1703), the differential equation can be written:

$$\ddot{\theta}+\omega_0^2\theta=0 \tag{9.3.801}$$

Which has a trivial solution:

$$\theta(t)=\theta_0\cos(\omega_0 t) \tag{9.3.802}$$

And we also deduce from it the frequency and thus the oscillation period:

$$\omega_0=2\pi f\Leftrightarrow T=2\pi\sqrt{\frac{L}{g}} \tag{9.3.803}$$

So we get the famous relation (don't forget it is an approximation!):

$$\boxed{T\cong 2\pi\sqrt{\frac{L}{g}}} \tag{9.3.804}$$

So the swing period is independent of the amplitude if the latter is small, which explains why the number of swings per minute of a simple pendulum is constant, whatever the force we put to the swinging ... We talk then of "isochronism"!

If:

$$L=l_0+NP \tag{9.3.805}$$

where l_0 is the position of the center of mass of the object and N the number of potential links of the chain (rather than a wire) that we would have taken for the length L of the chain and P is the pitch of the chain.

Which finally gives obviously:

$$T \cong \frac{2\pi}{\omega_0}\sqrt{\frac{l_0+NP}{g}} \tag{9.3.806}$$

Notice that our differential equation can also be reduced to a differential system of equations of the first order (transformation useful for many computer softwares such as Maple™ and MATLAB® for example as they request such an input!):

$$\frac{\mathrm{d}^2\theta}{\mathrm{d}t} = -\omega_0^2\sin(\theta) \underset{\substack{x=\theta \\ y=\mathrm{d}x/\mathrm{d}t}}{\Leftrightarrow} \frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} y \\ -\omega_0^2\sin(x) \end{bmatrix} \tag{9.3.807}$$

With Maple™ 4.00b we can plot thanks to this above relation the phase diagram $\theta-\omega$:

```
>with(DEtools):
>omega:=2:
>phaseportrait([diff(x(t),t)=y(t),diff(y(t),t)=-omega*sin(x(t))],
[x(t),y(t)],t=-10..20,[[x(0)=-2,y(0)=0],[x(0)=3,y(0)=0],[x(0)=1,y(0)=4]],
x=-4..7,y=-7..7,stepsize=0.1,linecolour=[blue,violet,green]);
```

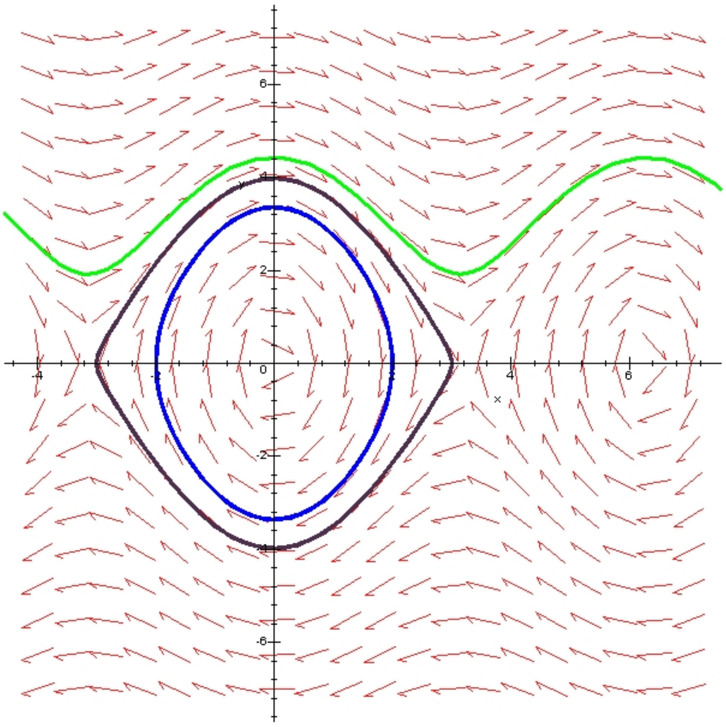


Figure 9.84: **Phase space $\theta-\omega$ of the simple pendulum in Maple™ 4.00b**

We must then know how to read this result:

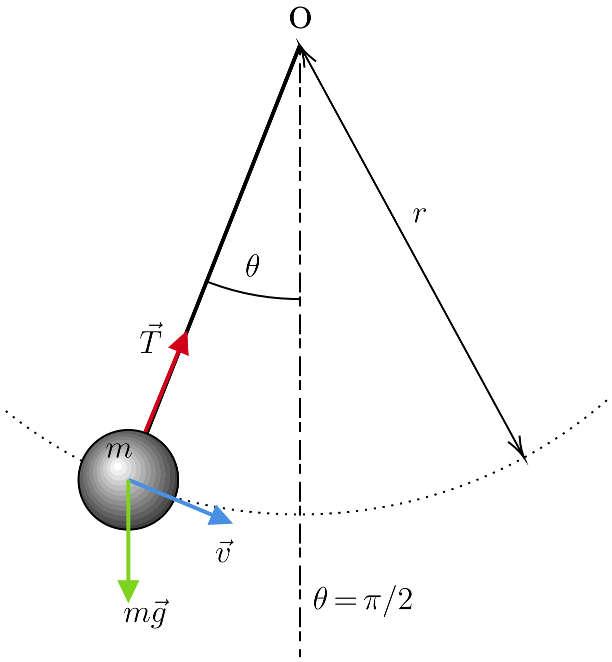
1. If the starting angle `x(0)` is small or large and the starting pulsation `y(0)` is small, the angle remains bounded and the pendulum then has a finite (non-revolving) balanced movement (blue and violet on the plot)

2. Whether the starting angle `x(0)` is small or large if the launching speed `y(0)` is positive and greater than a certain critical value, then the pendulum has an infinite revolving movement (green on the plot)

For the people interested in the Lagrangian approach of the simple ideal pendulum we start as always with the Euler-Lagrange equation (see section Analytical Mechanics page 3162):

$$L=T-V=\frac{1}{2}m(\dot{r}^2+r^2\dot{\theta})+mgr\cos(\theta) \tag{9.3.808}$$

Taking the potential V to be zero when the pendulum is horizontal $\theta=\pi/2$ and expressing the speed in polar coordinates (see section Vector Calculus page 1863).



Now we use our equation of motion:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial\dot{\theta}}\right)-\frac{\partial L}{\partial\theta}=0 \tag{9.3.809}$$

This gives us:

$$\frac{\mathrm{d}}{\mathrm{d}t}(mr^2\dot{\theta})+mgr\sin(\theta)=0 \tag{9.3.810}$$

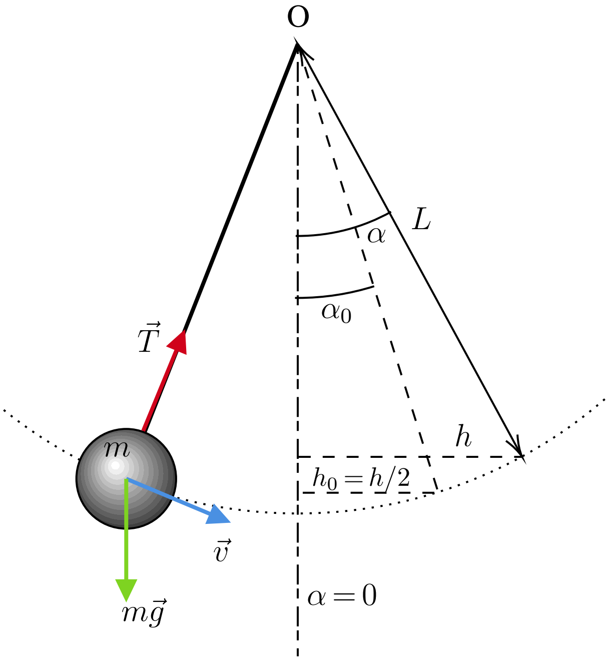
Therefore:

$$mr^2\ddot{\theta}+mgr\sin(\theta)=0 \tag{9.3.811}$$

and we fall back on our ODE...

Finally to close this subject on the simple pendulum, some textbooks (especially computer science textbooks????) give another approach based on the fact that the final relation must be given using an Elliptic Integral (it's a matter of choice). For this purpose let us rewrite the total energy at the angle α_0 where the kinetic energy is equal to the potential energy (obviously the α_0 is where the potential energy is equal to the kinetic energy is a the half height):

$$\frac{1}{2}mL^2\dot{\alpha}^2=mgL(\cos(\alpha_0)-\cos(\alpha)) \tag{9.3.812}$$



That can be rewritten as:

$$(\mathrm{d}t)^2=\frac{L(\mathrm{d}\alpha)^2}{2g(\cos(\alpha)-\cos(\alpha_0))} \tag{9.3.813}$$

therefore:

$$\mathrm{d}t=\sqrt{\frac{L}{2g}}\frac{\mathrm{d}\alpha}{\sqrt{\cos(\alpha)-\cos(\alpha_0)}} \tag{9.3.814}$$

Now if we integrate from $\alpha=0\ldots\alpha_0$, as we need four times this angle to make a complete cycle (indeed! remember that we were are the half height!), we then have:

$$T=4\sqrt{\frac{L}{2g}}\int\limits_0^{\alpha_0}\frac{\mathrm{d}\alpha}{\sqrt{\cos(\alpha)-\cos(\alpha_0)}} \tag{9.3.815}$$

Now let us use some trigonometric identities (see section Trigonometry page 2519):

$$\begin{gathered}\cos(\alpha)=1-2\sin^2\left(\frac{\alpha}{2}\right)\\ \cos(\alpha_0)=1-2\sin^2\left(\frac{\alpha_0}{2}\right)\end{gathered} \tag{9.3.816}$$

By putting:

$$p=\sin\left(\frac{\alpha_0}{2}\right) \tag{9.3.817}$$

Then we have:

$$T=\sqrt{\frac{L}{2g}}\int\limits_0^{\alpha_0}\frac{\mathrm{d}\alpha}{\cos(\alpha)-\cos(\alpha_0)}=2\sqrt{\frac{L}{g}}\int\limits_0^{\alpha_0}\frac{\mathrm{d}\alpha}{p^2-\sin^2\left(\frac{\alpha}{2}\right)} \tag{9.3.818}$$

We now carry out a series of subtle transformations beginning with a change of variable:

$$\sin\left(\frac{\alpha}{2}\right)=p\sin(\theta)=\sin\left(\frac{\alpha_0}{2}\right)\sin(\theta) \tag{9.3.819}$$

So when:

$$\alpha=0\Rightarrow\theta=0 \tag{9.3.820}$$

and:

$$\alpha=\alpha_0\Rightarrow\theta=\frac{\pi}{2} \tag{9.3.821}$$

In addition:

$$\begin{aligned}&\mathrm{d}\left(\sin\left(\frac{\alpha}{2}\right)\right)=\frac{1}{2}\cos\left(\frac{\alpha}{2}\right)\mathrm{d}\alpha\\&\mathrm{d}(p\sin(\theta))=p\cos(\theta)\mathrm{d}\theta\end{aligned} \tag{9.3.822}$$

It follows, since we have put for recall that:

$$\sin\left(\frac{\alpha}{2}\right)=p\sin(\theta) \tag{9.3.823}$$

that we then have:

$$\frac{1}{2}\cos\left(\frac{\alpha}{2}\right)\mathrm{d}\alpha=p\cos(\theta)\mathrm{d}\theta \tag{9.3.824}$$

and therefore:

$$\begin{aligned}\mathrm{d}\alpha&=\frac{2p\cos(\theta)}{\cos\left(\frac{\alpha}{2}\right)}\mathrm{d}\theta=\frac{2p\sqrt{1-\sin^2(\theta)}}{\sqrt{1-\sin^2\left(\frac{\alpha}{2}\right)}}\mathrm{d}\theta=\frac{2\sqrt{p^2-p^2\sin^2(\theta)}}{\sqrt{1-\sin^2\left(\frac{\alpha}{2}\right)}}\mathrm{d}\theta\\&=\frac{2\sqrt{p^2-\sin^2\left(\frac{\alpha}{2}\right)}}{\sqrt{1-\sin^2\left(\frac{\alpha}{2}\right)}}\mathrm{d}\theta=\frac{2\sqrt{p^2-\sin^2\left(\frac{\alpha}{2}\right)}}{\sqrt{1-p^2\sin^2(\theta)}}\mathrm{d}\theta\end{aligned} \tag{9.3.825}$$

Then we have:

$$T=2\sqrt{\frac{L}{g}}\int\limits_0^{\alpha_0}\frac{\mathrm{d}\alpha}{\sqrt{p^2-\sin^2\left(\frac{\alpha}{2}\right)}}=2\sqrt{\frac{L}{g}}\int\limits_0^{\pi/2}\frac{\frac{2\sqrt{p^2-\sin^2\left(\frac{\alpha}{2}\right)}}{\sqrt{1-p^2\sin^2(\theta)}}}{\sqrt{p^2-\sin^2\left(\frac{\alpha}{2}\right)}}\mathrm{d}\theta=4\sqrt{\frac{L}{g}}\int\limits_0^{\pi/2}\frac{\mathrm{d}\theta}{\sqrt{1-p^2\sin^2(\theta)}} \tag{9.3.826}$$

Either in the condensed form of the "elliptic integral of the first kind" (see section Differential and Integral Calculus page 1599):

$$\boxed{T=4\sqrt{\frac{L}{g}}\int\limits_0^{\pi/2}\frac{\mathrm{d}\theta}{\sqrt{1-p^2\sin^2(\theta)}}=4\sqrt{\frac{L}{g}}F\left(\frac{\pi}{2},p^2\right)} \tag{9.3.827}$$

We notice first that if in the integral $p=0$ we fall back on the relation we get earlier:

$$T=2\pi\sqrt{\frac{L}{g}} \tag{9.3.828}$$

We speak then of the "harmonic limit" (small initial angle amplitude).

We can make an approximation of $F(\pi/2,p^2)$ at the neighbourhood of $p=0$. For this we use the Taylor development (see section Sequences and Series page 1703):

$$\frac{1}{\sqrt{1-x^2\sin(\theta)}}=1+\frac{\sin^2(\theta)}{2}p^2+\mathcal{O}(x^4\sin^4(\theta)) \tag{9.3.829}$$

Then we can approach the integral for small x:

$$F\left(\frac{\pi}{2},p^2\right)\cong\int\limits_0^{\pi/2}\left(1+\frac{\sin^2(\theta)}{2}p^2+\mathcal{O}(p^4\sin^4(\theta)\right)\mathrm{d}\theta=\frac{\pi}{2}\left(1+\frac{1}{4}p^2+\mathcal{O}(p^4)\right) \tag{9.3.830}$$

Which gives:

$$F\left(\frac{\pi}{2},p^2\right)\cong4\sqrt{\frac{L}{g}}\frac{\pi}{2}\left(1+\frac{1}{4}p^2+\mathcal{O}(p^4)\right)=2\pi\sqrt{\frac{L}{g}}\left[1+\frac{1}{4}\sin^2\left(\frac{\alpha_0}{2}\right)+\mathcal{O}\left(\sin^2\left(\frac{\alpha_0}{2}\right)\right)\right] \tag{9.3.831}$$

Finally, if we neglect the terms of order 4, we have $\sin^2(\alpha_0/2)\cong\alpha_0^2/2$ and then we get finally the "Borda approximation":

$$T\cong2\pi\sqrt{\frac{L}{g}}\left(1+\frac{\alpha_0^2}{16}\right) \tag{9.3.832}$$

The dependence of the period with the amplitude of the oscillations is therefore quadratic. This an anharmonic effect due to the non-parabolic profile of the potential well in which is trapped the pendulum. However, the Borda formula produces an error greater than 1% as soon as α_0 exceeds $74°$.

There is - and this is less known - an approximate formula based on the arithmetico-geometric mean, which by far exceeds the performance of most of the other formulas found in the literature.

9.3.10.3 Physical Pendulum

We name "physical pendulum" any solid that can oscillate freely in gravity and in vacuum around a friction-less axis A, with a small amplitude ($|\theta_0| < 5°$):

First remember that we have proved during our study of inertia momentum that:

$$\vec{M} = J\frac{\mathrm{d}\vec{\omega}}{\mathrm{d}t} \tag{9.3.833}$$

What we will write here in scalar form as:

$$M = J_A\frac{\mathrm{d}\omega}{\mathrm{d}t} = J_A\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} \tag{9.3.834}$$

where for recall M is the force momentum (torque) that take back the pendulum to its equilibrium position and J_a the moment of inertia of the pendulum from its support axis A.

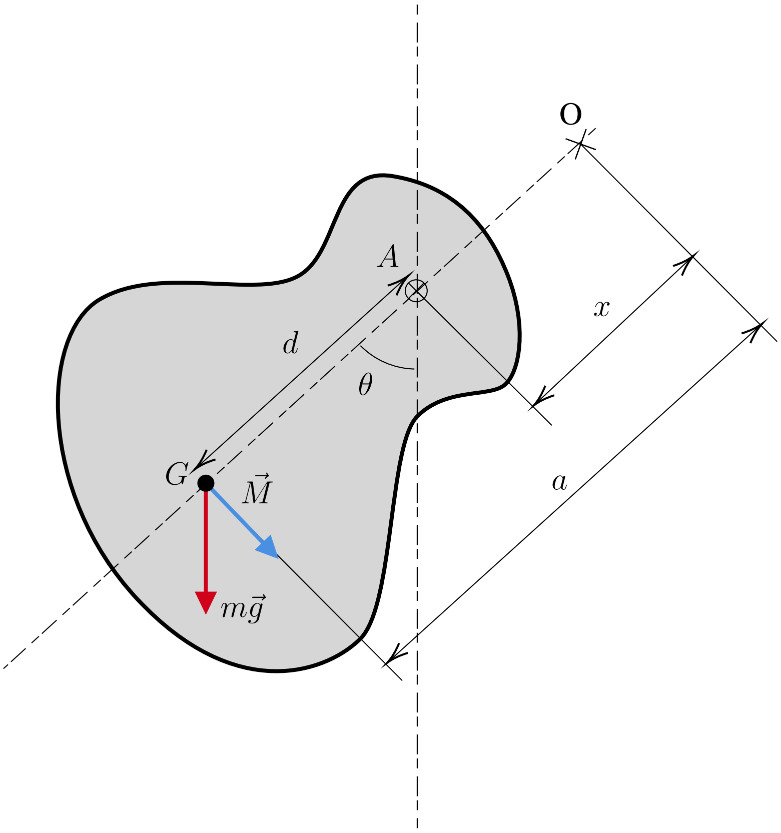


Figure 9.85: **Schematic ideal physical pendulum**

By analysing the force of our pendulum relatively to its centroid G we get another relation for M:

$$M = -Fd = -mg\cos\left(\frac{\pi}{2}-\theta\right)d = -mg\sin(\theta)d \cong -mgd\theta \tag{9.3.835}$$

For $\theta - 0 \leq \theta \leq \theta_0$ and where d is the distance from the friction-less axis of support of the pendulum to its center of mass G. The negative term appears here to express the fact that the period decreases with time (because the torque opposes the angular displacement from equilibrium).

As the angles θ are small, we replaced $\sin(\theta)$ without not too serious error by the first term of its development in Taylor series as for the simple pendulum (see section Sequences and Series page 1696).

Therefore we can write:

$$M=J_A\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2}\cong -mgd\theta \tag{9.3.836}$$

Hence the following differential equation of movement:

$$J_A\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2}+mgd\theta=0\Leftrightarrow\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2}+\frac{mgd}{J_A}\theta=0 \tag{9.3.837}$$

Or in a more condensed way:

$$\ddot{\theta}+\omega_0\theta=0 \tag{9.3.838}$$

So we have exactly the same type of differential equation as the simple pendulum for which a trivial solution is:

$$\theta(t)=\theta_0\cos(\omega_0 t) \tag{9.3.839}$$

but now with:

$$\omega_0=\sqrt{\frac{mgd}{J_A}} \tag{9.3.840}$$

Hence:

$$\boxed{T=2\pi\sqrt{\frac{J_A}{mgd}}} \tag{9.3.841}$$

✍**Example:**

Let us calculate the oscillation period of a swinging rod pivoting about one of its end. As we proved in the section of Geometry, for such a shape we have (see page 2926):

$$I=\frac{ML^2}{3} \tag{9.3.842}$$

Then:

$$T=2\pi\sqrt{\frac{\dfrac{ML^2}{3}}{Mg(L/2)}}=2\pi\sqrt{\frac{2L}{3g}} \tag{9.3.843}$$

Expressing the moment of inertia J_A using Steiner's theorem proven earlier during our study of inertial momentum :

$$J_A=J_G+md^2 \tag{9.3.844}$$

and further introducing the radius of gyration k:

$$J_A = mk^2 + md^2 \tag{9.3.845}$$

Hence:

$$J_A = md\left(\frac{k^2}{d} + d\right) \tag{9.3.846}$$

Given x the position of the axis of rotation A measured relatively to any one origin and a the position of center of gravity G with respect to the same origin we have:

$$d = |x - a| \tag{9.3.847}$$

as visible in the figure above. Therefore:

$$J_A = m|x-a|\left(\frac{k^2}{|x-a|} + |x-a|\right) \tag{9.3.848}$$

which gives us then another well known form of the period of the physical pendulum:

$$\boxed{T = 2\pi\sqrt{\frac{1}{g}\left(\frac{k^2}{|x-a|} + |x-a|\right)}} \tag{9.3.849}$$

As the root square is not comfortable we rise the whole at the square, which ultimately gives us:

$$T^2 = \frac{4\pi^2}{g}\left(\frac{k^2}{|x-a|} + |x-a|\right) \tag{9.3.850}$$

As we know x and T, this relation will allow us to draw a graph to determine the position of G and k if needed.

Thus, by plotting T^2 in function of x:

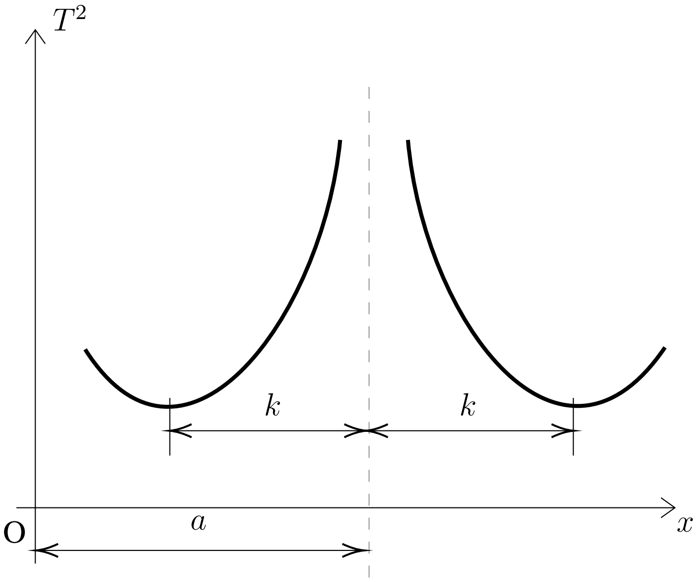


Figure 9.86: **Illustration of the vertical asymptote for determining the center of mass**

The resulting curve has a vertical asymptote ($T = +\infty$) for $x = a$ and two minima.

Differentiating T^2 with respect to x and cancelling the derivatives, we find the position of the minima:

$$x = a \pm k \tag{9.3.851}$$

9.3.10.4 Elastic Pendulum (mass-spring pendulum)

A "spring pendulum" (also named "elastic pendulum" or swinging spring) is a physical system where a piece of mass is connected to a spring so that the resulting motion contains elements of a simple pendulum as well as a spring. The system is much more complex than a simple pendulum, as the properties of the spring add an extra dimension of freedom to the system. For example, when the spring compresses, the shorter radius causes the spring to move faster due to the conservation of angular momentum. It is also possible that the spring has a range that is overtaken by the motion of the pendulum, making it practically neutral to the motion of the pendulum.

9.3.10.4.1 One degree of freedom elastic pendulum with/without friction

First let us consider the natural oscillations of a solid suspended in an elastic spring as it oscillates. After the separation of the solid from the equilibrium position, it will perform harmonic oscillations in the vertical direction, under the assumption that the elastic spring undergoes deformations proportional to the elongation (tension) of the spring.

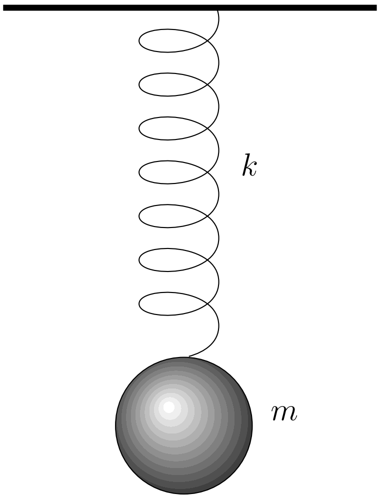


Figure 9.87: **Schematic ideal physical one degree of freedom pendulum**

We will often have to make in this textbook small movements around an equilibrium position.

This type of movement characteristic of what we name a "harmonic oscillator" is very common. It generalizes to all kinds of physical situations, such as RLC circuits (see section Electrical Engineering page 8475), the particle and wave quantum model of the atom (see section Wave Quantum Physics page 4688), quartz resonators or other vibrating structure slightly around its equilibrium point (see section Molecular Chemistry page 6211).

Remark

The harmonic oscillator is important because it's an approximate solution to nearly every system with a minimum of potential energy!

The reasoning comes from Taylor expansion (see section Sequences and Series page 1696). Consider a system with potential energy given by $U(x)$. You can approximate U at $x=x_0$ by:

$$U(x)=U(x_0)+(x-x_0)\left.\frac{\mathrm{d}U}{\mathrm{d}x}\right|_{x_0}+\frac{(x-x_0)^2}{2!}\left.\frac{\mathrm{d}^2U}{\mathrm{d}x^2}\right|_{x_0}+\ldots \tag{9.3.852}$$

The system will tend to settle into the configuration where $U(x)$ has a minimum - but, by definition, that's where the first derivative $\mathrm{d}U/\mathrm{d}x=0$ vanishes. Also a constant offset to a potential energy usually does not affect the physics. That leaves us with:

$$U(x)=\frac{(x-x_0)^2}{2!}\left.\frac{\mathrm{d}^2U}{\mathrm{d}x^2}\right|_{x_0}+\mathcal{O}(x-x_0)^3\cong\frac{1}{2}k(x-x_0)^2 \tag{9.3.853}$$

which is the harmonic oscillator potential for small oscillations around x_0.

We know that the restoring force (Hooke's law) of a spring is proportional and opposite to the strain such that (see section Continuum Mechanics page 3416):

$$F=ma=-k\Delta L \tag{9.3.854}$$

The differential equation of the harmonic oscillator can then be written:

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2}=-\frac{k}{m}x \tag{9.3.855}$$

We will take the simple approach that to try a solution, intuitively of the type... (exactly the same type as for the Simple Pendulum and Physical Pendulum):

$$x(t)=\cos(\omega_0 t) \tag{9.3.856}$$

It is a solution, because indeed:

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}t}&=-\omega_0\sin(\omega t)\\ \frac{\mathrm{d}^2x}{\mathrm{d}t^2}&=-\omega_0^2\cos(\omega t)\end{aligned} \tag{9.3.857}$$

provided that we take the natural frequency:

$$\omega_0=\sqrt{\frac{k}{m}} \tag{9.3.858}$$

So we get the famous period relation for the elastic pendulum:

$$\boxed{T=2\pi\sqrt{\frac{m}{k}}} \tag{9.3.859}$$

We also have the "proper mode":

$$x(t)=\sin(\omega_0 t) \tag{9.3.860}$$

as a solution.

A general solution is as we know (see section Differential and Integral Calculus page 1609):

$$x(t)=A\sin(\omega_0 t)+B\cos(\omega_0 t) \tag{9.3.861}$$

To find A and B, we must specify the initial conditions. Let us take for example at $t=0$:

$$x(0)=x_0=\Delta l \qquad v(0)=v_0 \tag{9.3.862}$$

Then we have:

$$\begin{aligned}x(t=0)&=B=\Delta l\\ v(t)&=A\omega\cos(\omega t)-B\sin(\omega t)\\ v(0)&=A\omega=v_0\end{aligned} \tag{9.3.863}$$

Let us now calculate the work (energy) needed to deform the harmonic oscillator. We have:

$$W=\int_{x_1}^{x_2}F\mathrm{d}x=\int_{x_1}^{x_2}kx\mathrm{d}x=k\int_{x_1}^{x_2}x\mathrm{d}x=\frac{1}{2}k(x_2^2-x_1^2)=\frac{1}{2}k\Delta x^2 \tag{9.3.864}$$

Thus, the elastic potential energy in a spring of material constant k, which have undergone a deformation x from it's initial position is thus given by:

$$\boxed{E_{\text{elastic}}=\frac{1}{2}kx^2} \tag{9.3.865}$$

For a more realistic description, a better modelling, we assume that the oscillator is subjected to an additional force representing friction. Often the approximation in which the friction force is proportional to the speed, and opposite to the speed, is a good approximation. It is not the only possible one and it is not always the best. We'll talk about friction forces (tribology) later.

Thus we consider a frictional force of the form (not to be confused with the notation of the momentum $\vec{b}$ that has absolutely no relation):

$$\vec{F}_f=-b\vec{v} \tag{9.3.866}$$

For our coordinate system:

$$F_f=-b\frac{\mathrm{d}x}{\mathrm{d}t} \tag{9.3.867}$$

The Newton's second law imposes:

$$m\frac{\mathrm{d}^2x}{\mathrm{d}t^2}=-kx-b\frac{\mathrm{d}x}{\mathrm{d}t} \tag{9.3.868}$$

Frequently written in the following important form:

$$\boxed{\frac{\mathrm{d}^2x}{\mathrm{d}t^2}+b\frac{\mathrm{d}x}{\mathrm{d}t}+kx=0} \tag{9.3.869}$$

Remark

For the forced version of this differential equation, introducing the concept of "resonance", the reader should refer to the page 8224.

To comply with a standard notation as part of the oscillator's study, we write:

$$\frac{k}{m}=\omega_0^2 \qquad \frac{b}{2m}=\gamma \tag{9.3.870}$$

hence the differential equation:

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2}+2\gamma\frac{\mathrm{d}x}{\mathrm{d}t}+\omega_0^2x=0 \tag{9.3.871}$$

We take the trail function:

$$x=e^{\lambda t} \tag{9.3.872}$$

Substituting, we find:

$$e^{\lambda t}\left(\lambda^2+2\gamma\lambda+\omega_0^2\right)=0 \tag{9.3.873}$$

As we seek non-zero solutions ($e^{\lambda t}$) we must have:

$$\lambda^2+2\gamma\lambda+\omega_0^2=0 \tag{9.3.874}$$

hence:

$$\begin{aligned}\lambda_1&=-\gamma+\sqrt{\gamma^2-\omega_0^2}\\ \lambda_2&=-\gamma-\sqrt{\gamma^2-\omega_0^2}\end{aligned} \tag{9.3.875}$$

and the general solution is:

$$x(t)=A_1e^{\lambda_1 t}+A_2e^{\lambda_2 t}=e^{-\gamma t}\left(A_1e^{\sqrt{\gamma^2-\omega_0^2}t}+A_2e^{-\sqrt{\gamma^2-\omega_0^2}t}\right) \tag{9.3.876}$$

where the two constants are determined by the initial conditions.

We will see that they corresponds to a low damping. Indeed, we can write with real square roots:

$$\lambda_{1,2}=-\gamma\pm\sqrt{\gamma^2-\omega_0^2}=-\gamma\pm\mathrm{i}\sqrt{\omega_0^2-\gamma^2} \tag{9.3.877}$$

the general solution can then be written as:

$$x(t)=e^{-\gamma t}\left(A_1e^{\mathrm{i}\sqrt{\omega_0^2-\gamma^2}t}+A_2e^{-\mathrm{i}\sqrt{\omega_0^2-\gamma^2}t}\right) \tag{9.3.878}$$

Using the complex (imaginary) properties of exponential and particularly the "Euler's formula" (see section Numbers page 251):

$$\begin{aligned}x(t)&=e^{-\gamma t}\left(A_1e^{\mathrm{i}\sqrt{\omega_0^2-\gamma^2}t}+A_2e^{-\mathrm{i}\sqrt{\omega_0^2-\gamma^2}t}\right)\\ &=e^{-\gamma t}\left(A_1\cos(\sqrt{\omega_0^2-\gamma^2}t)+\mathrm{i}A_1\sin(\sqrt{\omega_0^2-\gamma^2}t)+A_2\cos(-\sqrt{\omega_0^2-\gamma^2}t)+\mathrm{i}A_2\sin(-\sqrt{\omega_0^2-\gamma^2}t)\right)\end{aligned} \tag{9.3.879}$$

Let us choose $A_1 = A_2$ and recall that $\sin(-\alpha) = -\sin(\alpha)$ (see section Trigonometry page 2503). Therefore:

$$x(t) = e^{-\gamma t}\left(A_2\cos(\sqrt{\omega_0^2-\gamma^2}t) + A_2\cos(-\sqrt{\omega_0^2-\gamma^2}t)\right) \tag{9.3.880}$$

and as we also have $\cos(-\alpha) = \cos(\alpha)$. Then:

$$x(t) = e^{-\gamma t}2A_2\cos(\sqrt{\omega_0^2-\gamma^2}t) = e^{-\gamma t}C\cos(\sqrt{\omega_0^2-\gamma^2}t) \tag{9.3.881}$$

Let us put $\omega_1 = \sqrt{\omega_0^2-\gamma^2}$ and as the trigonometric function is periodic at $\phi = \pm 2k\pi$ with $k \in \mathbb{N}$ then:

$$x(t) = e^{-\gamma t}C\cos(\omega_1 t + \phi) \tag{9.3.882}$$

The general appearance of $x(t)$ for a given amplitude is a follows:

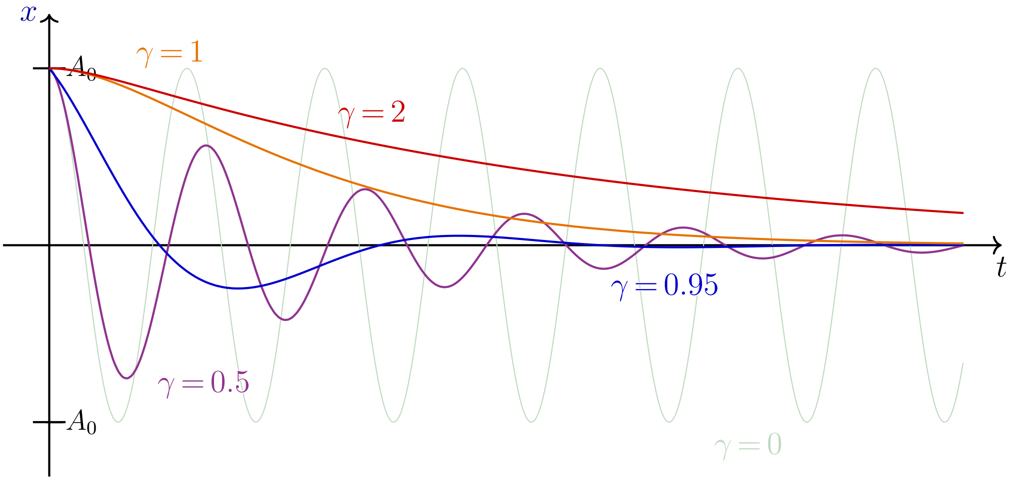


Figure 9.88: **Illustration of the amplitude damping of an elastic pendulum (author: Izaak Neutelings)**

And respectively for the energy:

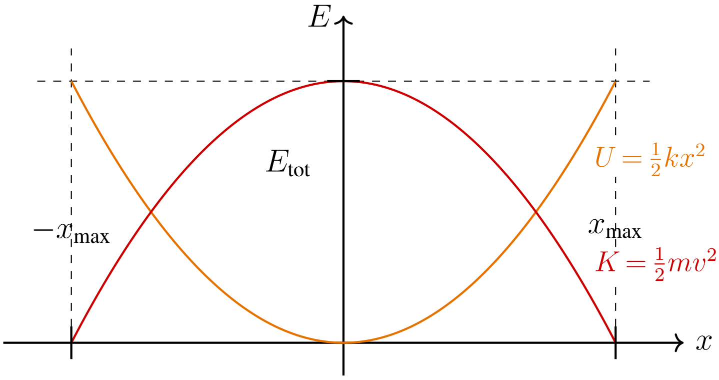


Figure 9.89: **Potential and Kinetic energy balancing of an elastic pendulum (author: Izaak Neutelings)**

When $\omega_0 = \gamma$ we say that there is a "critical damping", when $\omega_0 < \gamma$, we say that there is an "over-critical damping".

The dimensionless ratio:

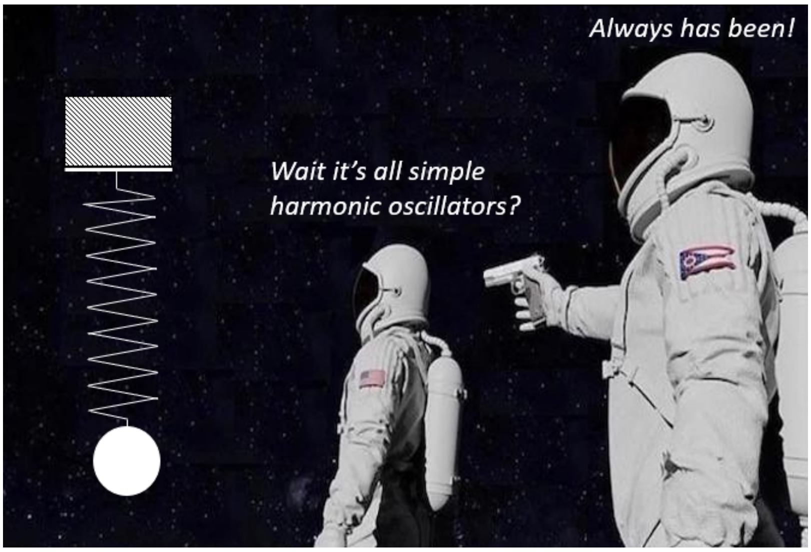
$$Q = \frac{\omega_0}{2\gamma} \tag{9.3.883}$$

meanwhile is named the "quality factor".

9.3.10.4.2 Harmonic oscillator (Lorentz oscillator)

The harmonic oscillator (also named "Lorentz oscillator") that is of main importance in many fields of physics is introduced in this textbook in the section of Music Mathematics at page 8224.

Once you will have finished to read that entire book you will very likely understand that meme:


Always has been!
Wait it's all simple
harmonic oscillators?

9.3.10.4.3 Two degrees of freedom elastic pendulum without friction

Now we consider an elastic pendulum that oscillates in the direction of the spring axes but that also oscillates in a $x - y$ plane:

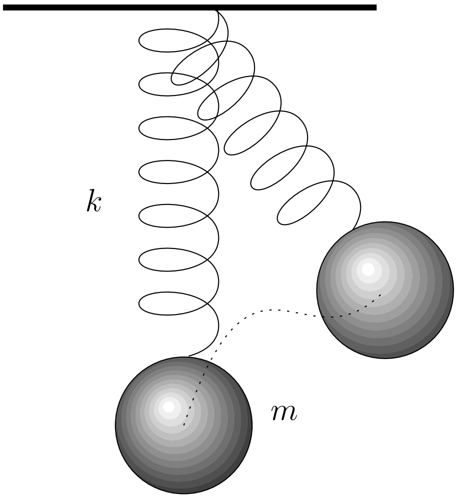


Figure 9.90: **Two degrees of freedom ideal elastic pendulum**

In practice the movement of such a pendulum seems chaotic but if we consider a friction-less pivot and that the experiment occurs in vacuum, it is still predictable!

Solving the equation of motion for such a pendulum cannot be done formally as far as we know and get the system of differential equation using the traditional method is quite awful.

So we will use the lagrangian approach. That is:

$$L = T - V \tag{9.3.884}$$

and:

$$\frac{\partial}{\partial t}\left(\frac{\partial L}{\partial \dot{q}}\right) - \frac{\partial L}{\partial q} \tag{9.3.885}$$

The kinetic energy may be broken up into the radial and tangential parts, so we have (the x component is along the axis of the spring relatively to the equilibrium length l_0 of the spring):

$$\begin{aligned}T &= T_{\text{rad}} + T_{\text{tan}} = \frac{1}{2}m\dot{x}^2 + \frac{1}{2}m(\omega r)^2 = \frac{1}{2}m\dot{x}^2 + \frac{1}{2}mr^2\omega^2 = \frac{1}{2}m\dot{x}^2 + \frac{1}{2}m(l_0+x)^2\dot{\theta}^2 \\ &= \frac{1}{2}m\left(\dot{x}^2 + (l_0+x)^2\dot{\theta}^2\right)\end{aligned} \tag{9.3.886}$$

The potential energy comes from both gravity and the spring, so we have:

$$V = E_p + E_{\text{elastic}} = -mg(l_0+x)\cos(\theta) + \frac{1}{2}kx^2 \tag{9.3.887}$$

The lagrangian is therefore:

$$L = T - V = \frac{1}{2}m\left(\dot{x}^2 + (l_0+x)^2\dot{\theta}^2\right) + mg(l_0+x)\cos(\theta) - \frac{1}{2}kx^2 \tag{9.3.888}$$

9.3.10.5 Conical Pendulum

A "conical pendulum", also sometimes named "circular pendulum",consists of a mass (or bob) fixed on the end of a string (or rod) suspended from a pivot. Its construction is similar to an ordinary pendulum; however, instead of swinging back and forth, the bob of a conical pendulum moves at a constant speed in a circle with the string (or rod) tracing out a cone.

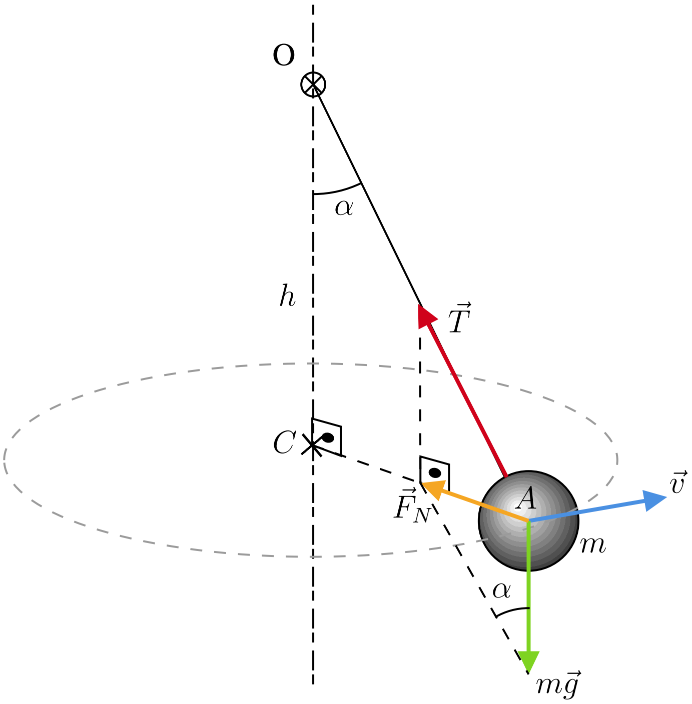


Figure 9.91: **Ideal conical pendulum**

As we can see above, the conical pendulum consist in taking a mass m considered as punctual and suspended to a rope of length $\overline{\text{O}A}=L$ to a friction-less pivot O.

The mass being taking to angle α of the vertical, the aim of this pendulum is frequently (because it is the simplest case) to determine the dependence between the angle and speed when you consider that trajectories are circular.

The mass m moves around the vertical $\overline{\text{O}C}$, describing a circle of radius:

$$r=\overline{CA}=\overline{\text{O}A}\sin(\alpha)=L\sin(\alpha) \tag{9.3.889}$$

The following forces act on the mass m:

$$\begin{aligned}F_g&=mg\\F_N&=m\frac{v^2}{R}=m\frac{(\omega R)^2}{R}=m\omega^2R=m\omega^2L\sin(\alpha)\end{aligned} \tag{9.3.890}$$

From the figure, we see that:

$$\tan(\alpha)=\frac{F_N}{F_G}=\frac{\omega^2L\sin(\alpha)}{g} \tag{9.3.891}$$

or, as:

$$\tan(\alpha)=\frac{\sin(\alpha)}{\cos(\alpha)} \tag{9.3.892}$$

then:

$$\cos(\alpha)=\frac{g}{\omega^2 L} \tag{9.3.893}$$

The angle α is even greater than the angular velocity ω is high, which is confirmed by the experience. For this reason, the conical pendulum was long used as cruise control on steam engines (it closes the steam supply when the speed exceeds a limit set in advance and opens when it falls below this value).

We have also:

$$\cos(\alpha)=\frac{g}{\left(\dfrac{v}{R}\right)^2 L}\Rightarrow v=\sqrt{\frac{R^2 g}{\cos(\alpha)L}}=\sqrt{\frac{g}{\cos(\alpha)L}}R=\sqrt{\frac{g}{\cos(\alpha)L}}L\sin(\alpha) \tag{9.3.894}$$

hence after simplification:

$$\boxed{v=\sqrt{\frac{gL}{\cos(\alpha)}}\sin(\alpha)} \tag{9.3.895}$$

As we have obviously $v=2\pi R/t$ (where t is the period to not confuse it with the T in the figure above that represents the tension) it comes:

$$T=2\pi R\sqrt{\frac{\cos(\alpha)L}{R^2 g}} \tag{9.3.896}$$

That is:

$$\boxed{t=2\pi\sqrt{\frac{\cos(\alpha)L}{g}}} \tag{9.3.897}$$

For small angles θ, $\cos(\theta)\cong 1$, and the period t of a conical pendulum is equal to the period of an ordinary pendulum of the same length. Also, the period for small angles is approximately independent of changes in the angle θ. This means the period of rotation is approximately independent of the force applied to keep it rotating. This property, named "isochronism", is shared with ordinary pendulums and makes both types of pendulums useful for timekeeping.

9.3.10.6 Torsion Pendulum

Consider a disk suspended from a torsion wire attached to its center (see figure below). This setup is known as a "torsion pendulum". A torsion wire is essentially inextensible, but is free to twist about its axis. Of course, as the wire twists it also causes the disk attached to it to rotate in the horizontal plane. Let θ be the angle of rotation of the disk, and let $\theta = 0$ correspond to the case in which the wire is untwisted.

Remark

The torsion pendulum is also a system that was used by Charles-Augustin Coulomb for measuring the elementary charge q_e and Henry Cavendish for measuring the gravitational constant G (see page 5270).

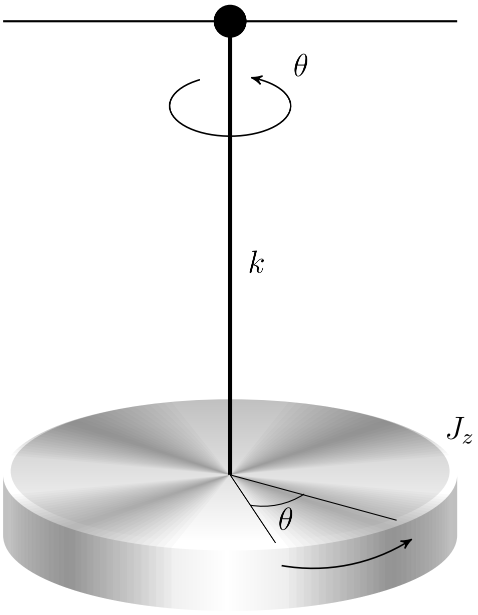


Figure 9.92: **Ideal torsion pendulum**

Any twisting of the wire is inevitably associated with mechanical deformation. The wire resists such deformation by developing a restoring torque, M, which acts to restore the wire to its untwisted state. For relatively small angles of twist, the magnitude of this torque is assumed directly proportional to the twist angle. Hence, we can write:

$$M = -k\theta \tag{9.3.898}$$

where k is the "torsion constant" of that particular material wire (see section Mechanical Engineering page 8404).

So we have:

$$M = J_z \frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} = -k\theta \tag{9.3.899}$$

Hence the differential equation:

$$\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2}+\frac{k}{J_z}\theta=0 \tag{9.3.900}$$

By analogy with the simple pendulum and physical pendulum where we had a similar differential equation to a given factor, we get:

$$\boxed{T=2\pi\sqrt{\frac{J_z}{k}}} \tag{9.3.901}$$

We conclude that when a torsion pendulum is perturbed from its equilibrium state (i.e., $\theta=0$), it executes torsional oscillations about this state at a fixed frequency, ω, which depends only on the torque constant of the wire and the moment of inertia of the disk. Note, in particular, that the frequency is independent of the amplitude of the oscillation (provided θ remains small enough).

Torsion pendulums are often used for time-keeping purposes. For instance, the balance wheel in a mechanical wristwatch is a torsion pendulum in which the restoring torque is provided by a coiled spring.

9.3.10.7 Foucault's Pendulum

The Foucault pendulum is a great experience to account for the rotation of the Earth. While it had long been known that the Earth rotates, the introduction of the Foucault pendulum in 11851 (holocene calendar) was the first simple proof of the rotation in an easy-to-see experiment. Today, Foucault pendulums are popular displays in science museums and universities.

There are several mathematical methods to analyse the behaviour of the Foucault pendulum. We have chosen to present the simplest one which requires little computational pages.

But first let us give a little explanatory text may be relevant as this experience is important and so well known.

Foucault's experiment aims to demonstrate that the Earth rotates on itself as we have just say. We push a pendulum (a mass after a wire). He moved back and forth regularly in the same direction. If we take it in a car and we do not turn too sharply, the pendulum don't care of the turns you take with the car: he continues to fight in the same direction. This is because a simple pendulum always oscillate in the same plane, despite the movements of its support.

This is why the French physicist Léon Foucault had the idea of attaching a heavy pendulum to a 67 meters long wire under the dome of the Pantheon, in the presence of Napoleon III and some scholars:

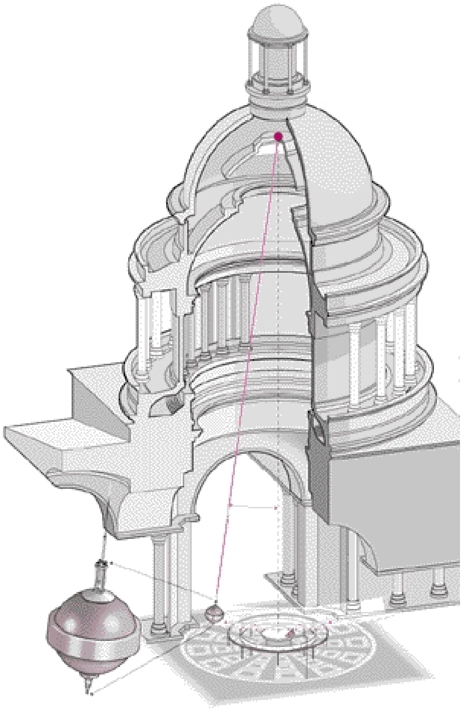


Figure 9.93: **Schematic view of Foucault's pendulum (source: ?)**

At each of its whereabouts, the pendulum came modify a pile of sand where it left a mark. But the track was never in the same place: there were 3 to 4 millimetres of difference between a swing and the next one, 16 seconds later. The pendulum remained in the same plane, but the Pantheon, Paris, the Earth are revolving!

To introduce the maths, consider the figure below:

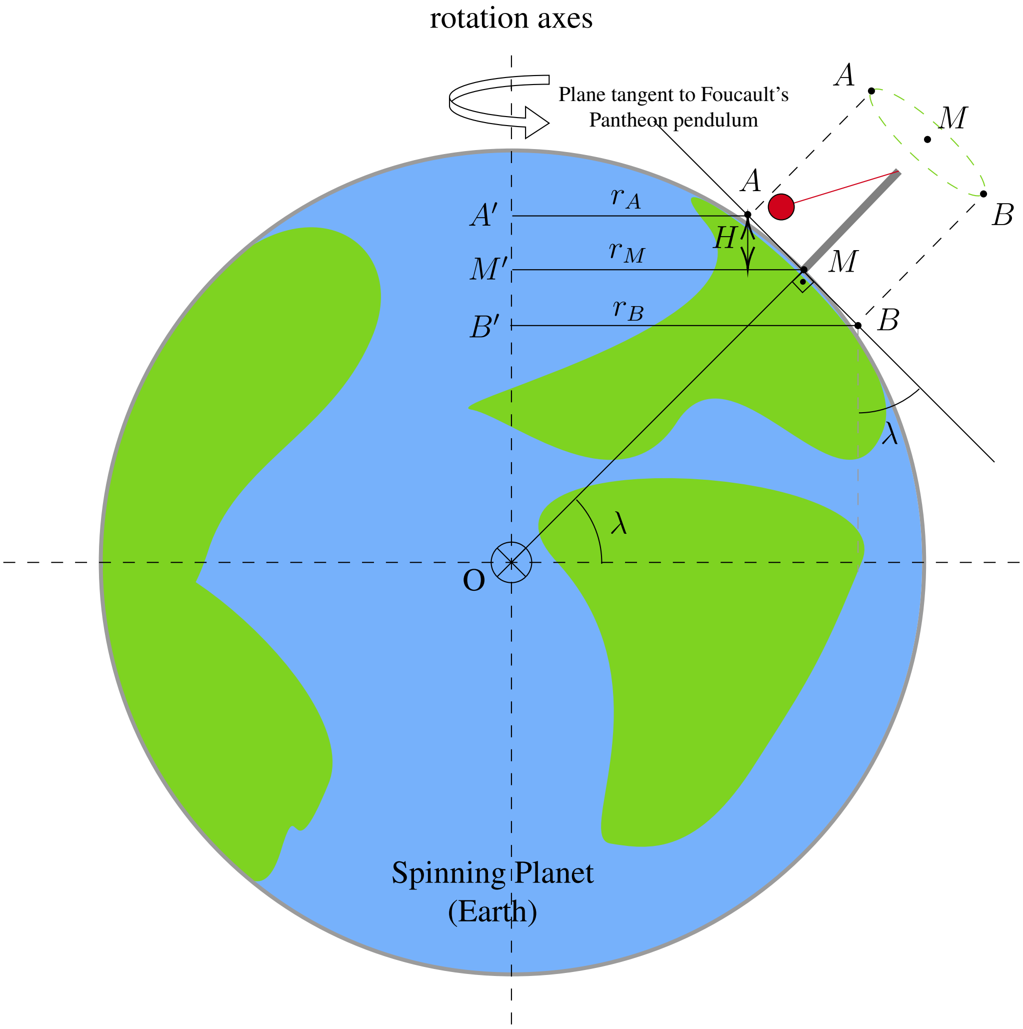


Figure 9.94: **Study diagram of Foucault's pendulum (not to scale obviously!)**

We consider that this is the view from a geocentric reference (Earth) seen in section on a plane containing the axis of rotation.

The size of the pendulum is obviously exaggerated in the figure above. However, it still oscillates in a meridian plane, between A and B (a terrestrial observer sees the line $\overline{AB}$ rotate relative to the Earth's surface by the green circle, seen in perspective, in the reverse direction).

Given T the rotation period of A (or B). The speed of A (v_a), on this circle is due to the fact that in the geocentric frame of reference, the point M, to the vertical of the point of suspension at the latitude λ, and the point on the Earth's surface coinciding with A at a given instant, does not have the same speed in the geocentric frame of reference: the point M being furthest from the axis of rotation of the Earth it is faster than A (same for the speed B that is higher than the speed of M).

Indeed, as $r_B > r_M > r_A$ and that $v = \omega R$, we have $v_B > v_M > v_A$.

The difference is of these speeds can easily be calculated assuming that rotation Earth's reasoning is uniform over a period of one day (sidereal) T_0.

We know that:

$$v = \omega r = \frac{2\pi}{T} \tag{9.3.902}$$

From this it easily follows that:

$$\Delta v = \frac{2\pi}{T_0}(r_m - r_A) \tag{9.3.903}$$

Since in the triangle AHM we have:

$$r_M - r_A = \overline{HM} = \overline{AM}\sin(\lambda) \tag{9.3.904}$$

Then:

$$\Delta v = \frac{2\pi}{T_0}\overline{AM}\sin(\lambda) \tag{9.3.905}$$

But we know Δv is just on green circle:

$$\Delta v = v_a = \frac{2\pi}{T}\overline{AM} \tag{9.3.906}$$

Therefore by equating and simplifying we get:

$$\boxed{T = \frac{T_0}{\sin(\lambda)}} \tag{9.3.907}$$

Remark

The direction of rotation of the pendulum is clockwise for an observer in the Northern Hemisphere; and counter-clockwise for an observer in the Southern Hemisphere.

✍**Example:**

The period of the Pantheon pendulum is 16.5 [s], the maximum amplitude of 6 [m] and the damping time of 6 [h]. We can observe a displacement of several millimetres at each return of the pendulum.

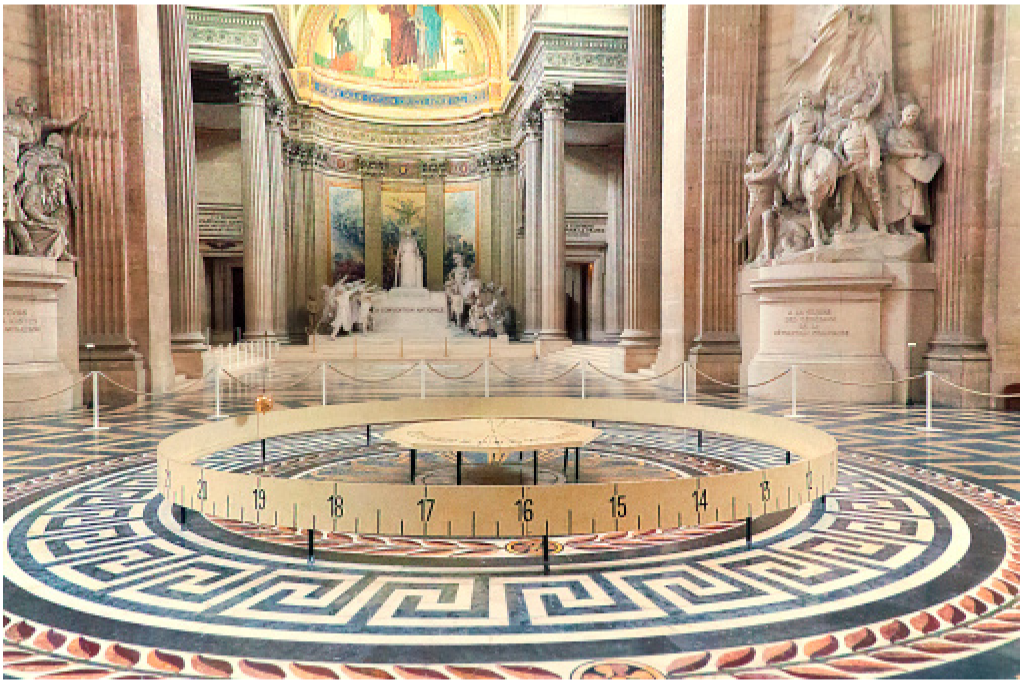


Figure 9.95: **Photo of Foucault's pendulum in Pantheon (Paris/France)**

At the poles (where the angle is $\pi/2$ [rad] and the sinus is equal to 1), the period of the pendulum equals that of the Earth and is therefore 24 [h]. At the equator (where the angle is equal to 0 [rad] and the sinus is equal to 0), the period of rotation of the oscillation plane is infinite: the plane of oscillation is stationary relative to the Earth! A Paris, where the latitude angle is equal $48°52'$ we have:

$$T=\frac{24}{\sin(48°52')}=\frac{24}{0.749}\cong 32 \tag{9.3.908}$$

That is to say in degrees: $11°$ [h^{-1}]. And this corresponds to the value given on Internet relatively to the Foucault's pendulum at the Pantheon!

Place	**Latitude**	Rotation time of the oscillation plane
North Pole	90°N	23 hours 56 minutes
Mourmansk (Russia)	68°58′N	25 hours 42 minutes
Oslo (Norway)	59°55′N	27 hours 45 minutes
Paris (France)	48°50′N	31 hours 57 minutes
Toulouse (France)	43°37′N	34 hours 51 minutes
Alexandria (Egypt)	31°13′N	46 hours 18 minutes
N'Djamena (Tchad)	12°10′N	114 hours 10 minutes
Equator	0°	Infinite : the oscillation plane remains fixed
The Cap (South Africa)	−33°94′S	42 hours 52 minutes
South Pole	90°S	23 hours 56 minutes

Table 9.20: **Foucault pendulum various periods**

However, the importance of Foucault's pendulum is elsewhere ...

The pendulum oscillation plane is fixed in reality and it is the rotation of the Earth on itself that result in an apparent rotation. But finally ... what is the reference system?

Indeed, all motions are relative. If the Earth is rotating, it because it's compared to something else. We can not talk about a movement without defining a frame of reference. The question that therefore arises is then to know relatively to what reference frame the pendulum oscillates in a fixed plane???

The first idea that comes to mind is to say that the pendulum plane is fixed relatively to the Sun. But if Foucault had managed to build a pendulum capable of oscillating long enough, say for a month, he would have noticed that the plane of oscillation also derived from the position of the Sun. Our star is therefore not part of the reference frame we are looking for!

Perhaps it is then necessary to consider the stars near the Sun? But again, if the experience could last long enough, it would show that the plane of oscillation clearly moves relative to the stars after a few years. What object to choose then? The galactic center, the Andromeda galaxy, the local group, the local superclusters? Each of these objects give the illusion of being stationary relative to the plane of the oscillations, but would end, after an increasingly long time, by revealing a drift.

Finally, as a last resort, we can consider the most distant objects, galaxies or quasars located billions of light-years away. With this reference system, and if the experience of Foucault was feasible, the plan of oscillation would finally be fixed and there would be no drift. So it is that in considering the most distant objects in fact the observable Universe as a whole, that we can get a reference frame in which the plane of oscillation stabilizes.

Foucault's pendulum then don't care of the presence of the Earth, the Sun (actually a G2V-type star, aka yellow dwarf) or the Galaxy. A part of its movement is directly dictated by the Universe as a whole. This experience highlights a kind of mysterious link between each point and the entire Universe. Until further notice, the nature of that relationship is unknown.

A similar conclusion was drawn by the Austrian physicist Ernst Mach in the late 119th century (we will talk of the "Mach's principle" in the section of Special Relativity).

According to Newtonian physics, the product of the mass of a body by its acceleration is equal to the force applied on it. Therefore, for a given force, the more massive an object, the more its acceleration is low. From this point of view, the mass is therefore a measure of the inertia of the body, that is to say its ability to resist to a force.

Now suppose that all matter in the Universe disappears, except for that body. The latter is then completely isolated and no force is exerted on it. This means, according to Newtonian physics, the product of its mass by its acceleration is zero. But the acceleration can not be equal to zero. Indeed, as all matter in the Universe is gone, there is no reference system against which to set the speed or acceleration. The latter is indefinite and not zero. From a mathematical point of view, there is only one possibility, that the mass of the body is zero!

This reasoning shows that the mass and inertia of a body are not really the properties of the object itself, but rather the result of an interaction with the rest of the Universe. Just as the Foucault pendulum, Mach's principle shows us that there must be some sort of connection between the local properties of a body and the global properties of the Universe. As in the previous case, the nature of this mysterious connection need to be determined.

9.3.10.8 Huygens' Pendulum (and brachistochrone curve)

We seek to build a pendulum whose period is independent of the amplitude (and not just by approximation as we have seen above!), that it is customary to name "rigorous isochronism". For this we have two cycloidal lamella at symmetrical positions and determined as shown in the figure below:

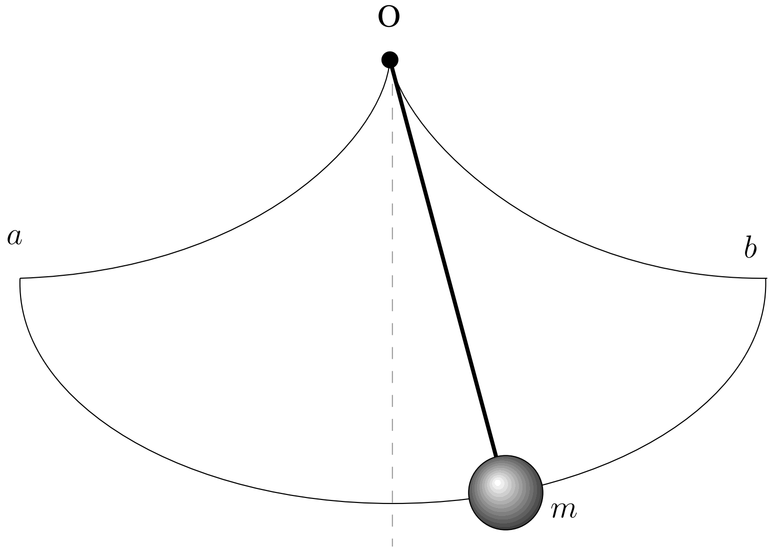


Figure 9.96: **Schematic Huygens' pendulum**

From this moment such pendulum applied to the measurement of time are known as "Coster clock", and at present all of these clocks manufactured under Huygen's license are in museums (the oldest known Huygens-style pendulum clock is dated 11657 - according holocene calendar - and can be seen at the Museum Boerhaave in Leiden) or in private collections.

Figure 9.97: **Reproduction of a coster clock (source: Lamazares Gil)**

The choice of the cycloid (also named "orthocycloïde") is because it is a "brachistochrone curve" (see definition below) and since the work of Christian Huygens in 11659 (holocene calendar), we also know that it is a "tautochronous curve" or "isochrone curve[19]" (some pendulums in modern watches have traditionally this form). That is to say that the bodies that fall into an inverted cycloid reach the lowest point at the same time, in any height they begin to fall (and for the pendulum independently of the starting amplitude - for small amplitudes only - they will reach the lowest point a the same time!).

So contrary to popular belief, the fastest way for a falling body on a solid support is not the straight line!

Indeed, one of the most famous problems of the history of mathematics is the "brachistochrone problem" which is to find the curve along which a particle would slide from one point to another in the shortest time being subject to a uniform field of gravity. This problem was posed by Johann Bernoulli in 11696 (holocene calendar) as a challenge for mathematicians of his day (and that was a really important challenge!!!). The solution was found by Johann Bernoulli himself and his brother Jacques Bernoulli, Isaac Newton, Gottfried Wilhelm Leibniz and the Marquis de l'Hospital. The brachistochrone problem is important in the development of mathematics and proves to be a major application of the method of variation calculus.

To introduce the corresponding maths we consider the two points A and B in the gravity field a material point m moving without friction on a curve of ends A and B. As already mentioned we want to determine the curve, named "brachistochrone", wherein the travel time is minimal when the point m starts from the point with a zero velocity.

Let us consider the figure below:

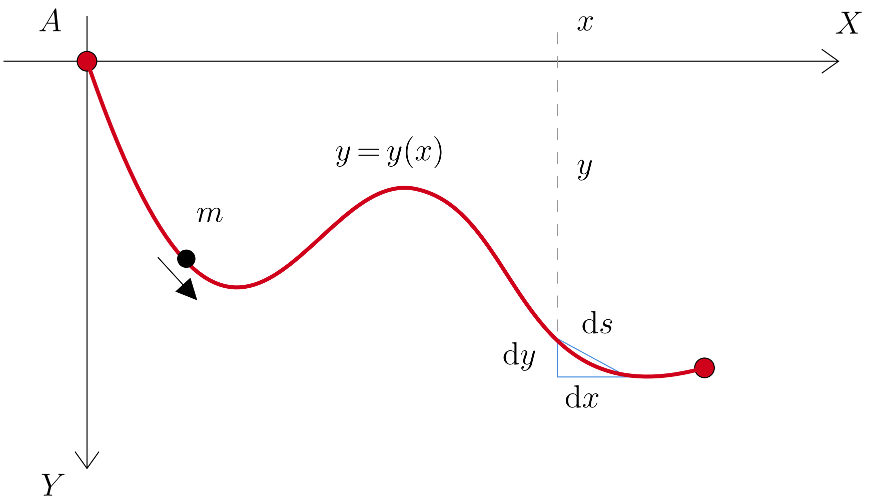


Figure 9.98: **Typical path to attack the brachistochrone problem without bias**

At the x-coordinate on the path plot above, the lost potential energy is $E_p = mgy$, equivalent to the kinetic energy gained by the material point from the start such as:

$$\frac{1}{2}mv^2 = mgh \tag{9.3.909}$$

---

[19] Remember that "brachistochrone" means "shortest time" and "tautochrone" means "same time".

Hence without too many surprises we get the "free fall speed in vacuum":

$$\boxed{v=\sqrt{2gh}} \tag{9.3.910}$$

The speed v is measured along the curve so that we have to rewrite the expression in horizontal and vertical components:

We will put that s is the curvilinear abscissa, $\mathrm{d}s$ the increase of this distance along the curve (see section Differential Geometry page 2850). The quantities $\mathrm{d}x$ and $\mathrm{d}y$ are the horizontal and vertical components of $\mathrm{d}s$.

Therefore:

- $\mathrm{d}s/\mathrm{d}t$ represents the speed along the curve
- $\mathrm{d}x/\mathrm{d}t$ the x component of the speed
- $\mathrm{d}x$ and $\mathrm{d}y$ are given by the Pythagorean theorem in exactly the same way as we did during our study of the Lagrangian formalism in the section of Analytical Mechanics:

$$v=\frac{\mathrm{d}s}{\mathrm{d}t}=\frac{\sqrt{\mathrm{d}x^2+\mathrm{d}y^2}}{\mathrm{d}t}=\sqrt{1+\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2}\frac{\mathrm{d}x}{\mathrm{d}t}=\sqrt{1+(y'(x))^2}\frac{\mathrm{d}x}{\mathrm{d}t} \tag{9.3.911}$$

By inserting the obtained equation from the principles of dynamics:

$$v=\sqrt{2gh}=\sqrt{2gy(x)}=\sqrt{1+(y'(x))^2}\frac{\mathrm{d}x}{\mathrm{d}t} \tag{9.3.912}$$

A simple integration then gives us the expression t to minimize:

$$t=\int\sqrt{\frac{1+(y'(x))^2}{2gy(x)}}\mathrm{d}x \tag{9.3.913}$$

We end up the with a function similar to the one we had in our study of a practical case of the Lagrangian formalism.

The challenge now is to find the minimum reached by t among all possible functions $y(x)$ satisfying the constraints:

$$y(x=0)=0 \qquad y(x=a)=b \tag{9.3.914}$$

The fundamental problem said to be of the "calculus of variations" is to search among all continuously differentiable functions $y=f(x)$ over a given interval $[A,B]$ for which the functions $f(A)$ and $f(B)$ are given known values, which make maximum or minimum the previous integral.

To apply this method, we start from the Euler-Lagrange equation (see section Analytical Mechanics page 3162):

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial Z(x,y,y')}{\partial y'}\right)-\frac{\partial Z(x,y,y')}{\partial y}=0 \tag{9.3.915}$$

which gives the extremum of the integral.

Similarly to an example that we saw during our study of the Lagrangian formalism (see section Analytical Mechanics page 3172) we put in our case:

$$Z(x,y,y')=\sqrt{\frac{1+(y'(x))^2}{2gy(x)}} \tag{9.3.916}$$

Therefore:

$$\begin{aligned}\frac{\partial Z}{\partial y}&=-\frac{1}{4}\frac{\sqrt{2}(1+y'^2}{\sqrt{\dfrac{1+y'^2}{gy}}gy^2}=-\frac{1}{4}\frac{\sqrt{2(1+y'^2}}{y\sqrt{gy}}\\ \frac{\partial Z}{\partial y'}&=\frac{\dfrac{2y'}{2gy}}{2\sqrt{\dfrac{1+y'^2}{2gy}}}=\frac{y'}{2gy}\sqrt{\frac{2gy}{1+y'^2}}=\frac{y'}{\sqrt{1+y'^2}\sqrt{2gy}}\end{aligned} \tag{9.3.917}$$

We wish to inject the last two relations in the Euler-Lagrange equation:

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial Z(x,y,y')}{\partial y'}\right)-\frac{\partial Z(x,y,y')}{\partial y}=0 \tag{9.3.918}$$

but we anticipate relatively easily that the derivative with respect to x will give us an indigestible monster to simplify (I tried ... but I have the excuse of not being good at math).

We will then use the Beltrami's identity proved in the section of Analytical Mechanics (see page 3177) that is written with the above selected notation:

$$Z-y'\frac{\partial Z}{\partial y'}=c^{te} \tag{9.3.919}$$

that we have the right to do as the Beltrami 's condition here is satisfied:

$$\frac{\partial Z}{\partial t}=0 \tag{9.3.920}$$

Which gives us:

$$\sqrt{\frac{1+y'^2}{2gy}}-y'\frac{y'}{\sqrt{1+y'^2}\sqrt{2gy}}=c^{te} \tag{9.3.921}$$

Thus after rearranging:

$$\frac{\sqrt{1+y'^2}\sqrt{1+y'^2}}{\sqrt{2gy}\sqrt{1+y'^2}}-y'\frac{y'}{\sqrt{1+y'^2}\sqrt{2gy}}=\frac{\sqrt{1+y'^2}\sqrt{1+y'^2}-y'^2}{\sqrt{1+y'^2}\sqrt{2gy}}=\frac{1}{\sqrt{1+y'^2}\sqrt{2gy}}=c^{te} \tag{9.3.922}$$

So finally:

$$\frac{1}{\sqrt{1+y'^2}\sqrt{2gy}}=c^{te} \tag{9.3.923}$$

Or written in another way:

$$y(1+y'^2)=\frac{1}{2gc^{te2}}=K \tag{9.3.924}$$

where the constant K is only $-D$, the maximal point reached by the mobile. More explicitly:

$$y\left(1+\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2\right)=\frac{1}{2gc^{te2}}=K \tag{9.3.925}$$

We must solve this differential equation to find the function that provides the fastest path. Let's us focus on it in the following form:

$$y\left(1+\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2\right)=K \tag{9.3.926}$$

and let us recall that on the figure above we have:

$$\frac{\mathrm{d}y}{\mathrm{d}x}=\tan(\theta) \tag{9.3.927}$$

Then we have (see section Trigonometry page 2519):

$$y(1+\tan^2(\theta))=\frac{y}{\cos^2(\theta)}=K \tag{9.3.928}$$

Hence:

$$y=K\cos^2(\theta)=\frac{K}{2}(1+\cos(2\theta)) \tag{9.3.929}$$

Note that we have (still using trigonometric identities):

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}\theta}&=\frac{\mathrm{d}x}{\mathrm{d}y}\frac{\mathrm{d}y}{\mathrm{d}\theta}=-\frac{1}{\tan(\theta)}K\sin(2\theta)=-\frac{1}{\tan(\theta)}K2\tan(\theta)\cos^2(\theta)=-2K\cos^2(\theta)\\&=-2K\frac{(1+\cos(2\theta))}{2}=-K(1+\cos(2\theta))\end{aligned} \tag{9.3.930}$$

Therefore:

$$\frac{\mathrm{d}x}{\mathrm{d}\theta}=-K(1+\cos(2\theta)) \tag{9.3.931}$$

Which is easy to integrate (see section Differential and Integral Calculus page 1534):

$$\begin{aligned}x&=-K\left(\theta+\frac{1}{2}\sin(2\theta)\right)+c^{te}=-\frac{2}{2}K\left(\theta+\frac{1}{2}\sin(\theta)\right)+c^{te}\\&=-\frac{K}{2}(2\theta+\sin(2\theta))+c^{te}\end{aligned} \tag{9.3.932}$$

So we have at this point (for recall):

$$\begin{aligned}x&=-\frac{K}{2}(2\theta+\sin(2\theta))+c^{te}\\y&=\frac{K}{2}(1+\cos(2\theta))\end{aligned} \tag{9.3.933}$$

Let us put to simplify the writing $\alpha=2\theta$, we have then:

$$\begin{aligned}x&=-\frac{K}{2}(\alpha+\sin(\alpha))+c^{te}\\ y&=\frac{K}{2}(1+\cos(\alpha))\end{aligned} \tag{9.3.934}$$

Let us recall now that we have the initial condition:

$$y(x=0)=0 \tag{9.3.935}$$

and as K is not zero, this requires that in the prior previous relation that:

$$\cos(\alpha_{x=0})=-1 \tag{9.3.936}$$

and therefore that:

$$\alpha_{x=0}=-\pi\Leftrightarrow\theta_{x=0}=-\frac{\pi}{2} \tag{9.3.937}$$

Therefore, we must have:

$$x=-\frac{K}{2}(-\pi+\sin(-\pi))+c^{te}=0 \tag{9.3.938}$$

and we deduce immediately that:

$$c^{te}=-\frac{K\pi}{2} \tag{9.3.939}$$

Since then:

$$\begin{aligned}x&=-\frac{K}{2}(\alpha+\sin(\alpha))-\frac{K\pi}{2}\\ y&=\frac{K}{2}(1+\cos(\alpha))\end{aligned} \tag{9.3.940}$$

To close that subject, let us do the traditional variable change:

$$\alpha=\varphi-\pi\qquad a=\frac{K}{2} \tag{9.3.941}$$

It comes then:

$$\begin{aligned}x&=-a((\varphi-\pi)+\sin(\varphi-\pi))-a\pi=-a(\varphi+\sin(\varphi-\pi))=-a(\varphi-\sin(\varphi))\\ y&=a(1+\cos(\alpha)=a(1+\cos(\varphi-\pi))=a(1-\cos(\varphi))\end{aligned} \tag{9.3.942}$$

So in the end:

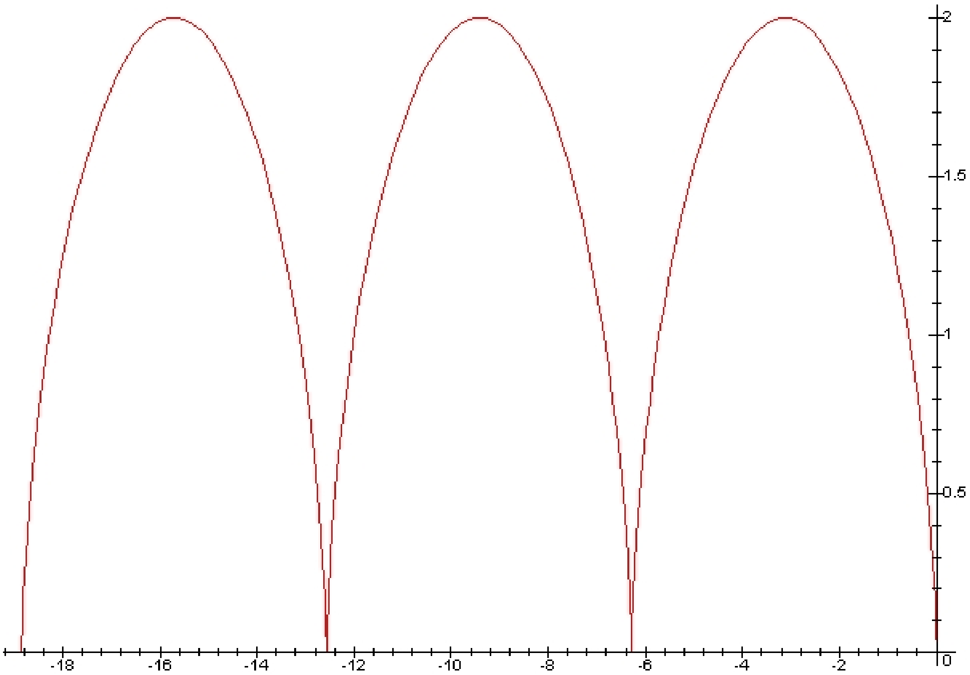
$$\begin{aligned}x&=-a(\varphi-\sin(\varphi))\\ y&=a(1-\cos(\varphi))\end{aligned} \tag{9.3.943}$$

As the sign of the coefficient a of x that has for only effect a translation along the X axis, it is customary to represent that pair of equations in the following form where we recognize the parametric equations of a cycloid curve (see section Analytical Geometry page 2814):

$$\boxed{\begin{aligned}x&=a(\theta-\sin(\theta))\\ y&=a(1-\cos(\theta))\end{aligned}} \tag{9.3.944}$$

By putting $a=1$ we have in Maple™ 4.00b:

```
>plot([theta-sin(theta),1-cos(theta),theta=0..6*Pi]);
```



This has to be compared with the intuitive answers that humans being give when they try to guess the fastest curve (in red below the brachistochrone curve and in blue... most human beings answers):

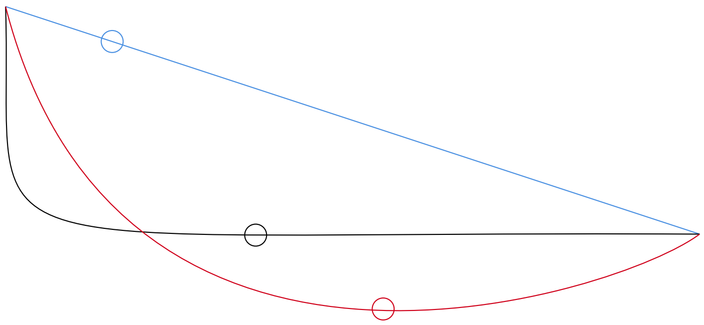


Figure 9.99: **Comparison of the brachistochrone with other intuitive paths**

Given the derivatives:

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}\theta}&=a(1-\cos(\theta))\\ \frac{\mathrm{d}y}{\mathrm{d}\theta}&=a\sin(\theta)\end{aligned} \tag{9.3.945}$$

Therefore:

$$(\mathrm{d}_\theta x)^2+(\mathrm{d}_\theta y)^2=a^2\left[(1-2\cos(\theta)+\cos^2(\theta)+\sin^2(\theta)\right]=2a^2\left(1-\cos(\theta)\right) \tag{9.3.946}$$

The conservation of energy:

$$\frac{1}{2}mv^2=mgy \tag{9.3.947}$$

is then written as we know:

$$v=\frac{\mathrm{d}s}{\mathrm{d}t}=\sqrt{2gy} \tag{9.3.948}$$

hence:

$$\mathrm{d}t=\frac{\mathrm{d}s}{\sqrt{2gy}}=\frac{\sqrt{(\mathrm{d}_\theta x)^2+(\mathrm{d}_\theta y)^2}}{\sqrt{2gy}}=\frac{a\sqrt{2(1-\cos(\theta))\mathrm{d}\theta}}{\sqrt{2ga(1-\cos(\theta)}}=\sqrt{\frac{a}{g}}\mathrm{d}\theta \tag{9.3.949}$$

Therefore the time required to go from top to bottom of the cycloid described by the Huygens' pendulum is:

$$T=\int\limits_0^\pi\sqrt{\frac{a}{g}}\mathrm{d}\theta=\pi\sqrt{\frac{a}{g}} \tag{9.3.950}$$

This term therefore only dependent on fixed parameters and not of the amplitude of the pendulum!

The statement in 11696 (holocene calendar) of the brachistochrone the problem can be considered as the true birth of the calculus of variations, because it is this problem which prompted the search for general methods gradually developed in the context of a real scientific (and business...) competition.

Remark

The brachistochrone of a surface is a curve on which a material punctual mass slide without friction when placed in a uniform gravitational field so that travel time is minimal among all curves joining two fixed points. In other words, the brachistochrone is the shortest lines in time, while the geodesic (see section Tensor Calculus page 5747) are the curves of the shortest distance for a given fixed time interval!

9.3.10.9 Double Pendulum

A double pendulum is a pendulum with another pendulum attached to its end, and is a simple physical system that exhibits rich dynamic behaviour with a strong sensitivity to initial conditions.

The double pendulum problem is a classic application of the Lagrangian formalism and of Chaos Theory and so verbatim a nice school example of non-linear physics.

Here is a figure of how it is often described (the x and y are positive on both axes represented below!):

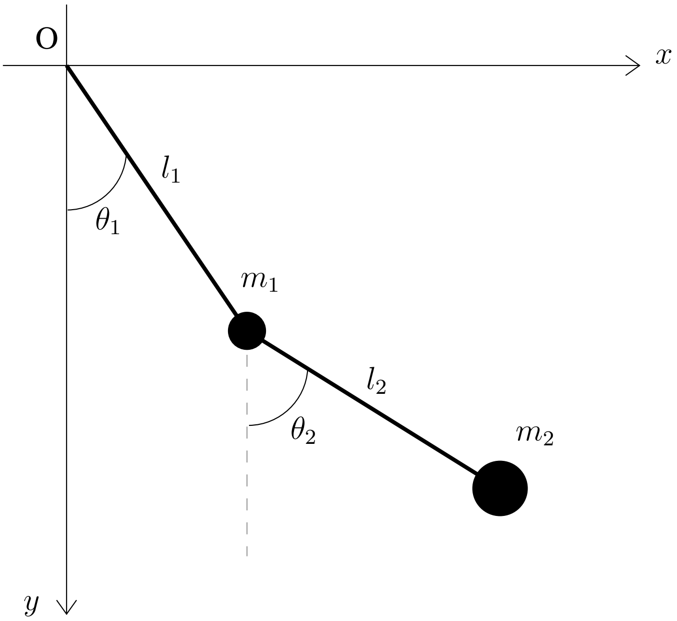


Figure 9.100: **Illustration of the ideal double pendulum**

The position of the mass m_1 will be given by the coordinates (x_1, y_1) and that of the mass m_2 - linked to m_1 and therefore not independent - will be given by the coordinates (x_2, y_2).

The best is to reduce the study of the system through the generalized coordinates (θ_1, θ_2) trough the transformations:

$$x_1 = l_1 \sin(\theta) \qquad y_1 = l_1 \cos(\theta_1) \tag{9.3.951}$$

and:

$$x_2 = x_1 + l_2 \sin(\theta_2) \qquad y_2 = y_1 + l_2 \cos(\theta_2) \tag{9.3.952}$$

So in Cartesian coordinates and using the traditional notation of the Lagrangian formalism, the kinetic energy is then:

$$T = \frac{1}{2} m_1 v_1^2 + \frac{1}{2} m_2 v_2^2 \tag{9.3.953}$$

and as:

$$\begin{aligned} v_1 &= \left\| \frac{\mathrm{d}\vec{r}_1}{\mathrm{d}t} \right\| = \left\| \frac{\mathrm{d}}{\mathrm{d}t} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} \right\| = \left\| \frac{\mathrm{d}}{\mathrm{d}t} \begin{pmatrix} l_1 \sin(\theta_1) \\ l_1 \cos(\theta_1) \end{pmatrix} \right\| = \left\| l_1 \begin{pmatrix} \dot{\theta}_1 \cos(\theta_1) \\ -\dot{\theta}_1 \sin(\theta_1) \end{pmatrix} \right\| \\ &= \sqrt{(l_1 \dot{\theta}_1 \cos(\theta))^2 + (-l_1 \dot{\theta}_1 \sin(\theta_1)^2} = \sqrt{l_1^2 \dot{\theta}_1^2 \cos^2(\theta_1) + l_1^2 \dot{\theta}_1^2 \sin^2(\theta_1)} = \dot{\theta}_1 l_1 \end{aligned} \tag{9.3.954}$$

and:

$$\begin{aligned}v_2&=\left\|\frac{\mathrm{d}\vec{r}_2}{\mathrm{d}t}\right\|=\left\|\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}x_2\\y_2\end{pmatrix}\right\|=\left\|\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}x_1+l_2\sin(\theta_2)\\y_1+l_2\cos(\theta_2)\end{pmatrix}\right\|=\left\|\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}l_1\sin(\theta_1)+l_2\sin(\theta_2)\\l_1\cos(\theta_1)+l_2\cos(\theta_2)\end{pmatrix}\right\|\\&=\left\|\begin{pmatrix}l_1\dot{\theta}_1\cos(\theta_1)+l_2\dot{\theta}_2\cos(\theta_2)\\-l_1\dot{\theta}_1\sin(\theta_1)-l_2\dot{\theta}_2\sin(\theta_2)\end{pmatrix}\right\|=\sqrt{(l_1\dot{\theta}_1\cos(\theta_1)+l_2\dot{\theta}\sin(\theta_2)}\\&=\sqrt{(l_1\dot{\theta}_1\cos(\theta_1)+l_2\dot{\theta}_2\cos(\theta_2))^2+(-l_1\dot{\theta}_1\sin(\theta_1)-l_2\dot{\theta}_2\sin(\theta_2))^2}\\&=\sqrt{\begin{matrix}l_1^2\dot{\theta}_1^2\cos^2(\theta_1)+2l_1l_2\dot{\theta}_1\dot{\theta}_2\cos(\theta_1)\cos(\theta_2)+l_2^2\dot{\theta}_2^2\cos^2(\theta_2)+l_1^2\dot{\theta}_1^2\sin^2(\theta_1)\\+2l_1l_2\dot{\theta}_1\dot{\theta}_2\sin(\theta_1)\sin(\theta_2)+l_2^2\dot{\theta}_2^2\sin^2(\theta_2)\end{matrix}}\\&=\sqrt{l_1^2\dot{\theta}_1^2+l_2^2\dot{\theta}^2+2l_1l_2\dot{\theta}_1\dot{\theta}_2(\cos(\theta_1)\cos(\theta_2)+\sin(\theta_1)\sin(\theta_2))}\\&=\sqrt{l_1^2\dot{\theta}_1^2+l_2^2\dot{\theta}_2^2+2l_1l_2\dot{\theta}_1\dot{\theta}_2\cos(\theta_1-\theta_2)}\end{aligned}\tag{9.3.955}$$

Then we have:

$$\begin{aligned}T&=\frac{1}{2}m_1v_1^2+\frac{1}{2}m_2v_2^2=\frac{1}{2}m_2v_2^2=\frac{1}{2}m_1(\dot{\theta}l_1)^2+\frac{1}{2}m_2\left(\sqrt{l_1^2\dot{\theta}_1^2+l_2^2\dot{\theta}_2^2+2l_1l_2\dot{\theta}_1\dot{\theta}_2\cos(\theta_1-\theta_2)}\right)^2\\&=\frac{1}{2}m_1\dot{\theta}_1^2l_1^2+\frac{1}{2}m_2\left(l_1^2\dot{\theta}_1^2+l_2^2\dot{\theta}_2^2+2l_1l_2\dot{\theta}\dot{\theta}_2\cos(\theta_1-\theta_2)\right)\\&=\frac{1}{2}(m_1+m_2)\dot{\theta}_2l_1^2+\frac{1}{2}m_2\left(l_2^2\dot{\theta}_2^2+2l_1l_2\dot{\theta}\dot{\theta}_2\cos(\theta_1-\theta_2)\right)\end{aligned}\tag{9.3.956}$$

For the potential energy we have (the sign is negative because the masses are below the point 0):

$$V=-m_1gy_1-m_2gy_2=-m_1gl_1\cos(\theta_1)-m_2g(l_1\cos(\theta_1)+l_2\cos(\theta_2)\tag{9.3.957}$$

The Lagrangian (see section Analytical Mechanics page 3152) is therefore given by:

$$\begin{aligned}&L=T-V=\frac{1}{2}(m_1+m_2)\dot{\theta}_1l_1^2+\frac{1}{2}m_2l_2^2\dot{\theta}_2^2+m_2l_1l_2\dot{\theta}_1\dot{\theta}_2\cos(\theta_1-\theta_2)\\&\quad+(m_1+m_2)gl_1\cos(\theta_1)+m_2gl_2\cos(\theta_2))\end{aligned}\tag{9.3.958}$$

We now use the Euler-Lagrange equation with the chosen generalized coordinates:

$$\begin{aligned}\frac{\partial L}{\partial\theta_1}-\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial\dot{\theta}_1}=0\\\frac{\partial L}{\partial\theta_2}-\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial\dot{\theta}_2}=0\end{aligned}\tag{9.3.959}$$

We then have for θ_1:

$$\begin{aligned}
&\frac{\partial L}{\partial \theta_1}=-m_2 l_1 l_2 \dot{\theta}_1 \dot{\theta}_2 \sin(\theta_1-\theta_2)-(m_1+m_2) g l_1 \sin(\theta_1)\\
&\frac{\partial L}{\partial \dot{\theta}_1}=(m_1+m_2)\dot{\theta}_1 l_1^1+m_2 l_1 l_2 \dot{\theta}_2 \cos(\theta_1-\theta_2)\\
&\frac{\mathrm{d}}{\mathrm{d}t}=(m_1+m_2)\ddot{\theta}_1 l_1^2+m_2 l_1 l_2 \ddot{\theta}_2 \cos(\theta_1-\theta_2)+m_2 l_1 l_2 \dot{\theta}_2 \frac{\mathrm{d}}{\mathrm{d}t}\cos(\theta_1-\theta_2)\\
&=(m_1+m_2)\ddot{\theta} l_1^2+m_2 l_1 l_2 \ddot{\theta}_2 \cos(\theta_1-\theta_2)+m_2 l_1 l_2 \dot{\theta}_2(-\dot{\theta}_1 \sin(\theta_1-\theta_2)+\dot{\theta}_2 \sin(\theta_1-\theta_2))\\
&=(m_1+m_2)\ddot{\theta} l_1^2+m_2 l_1 l_2 \ddot{\theta}_2 \cos(\theta_1-\theta_2)+m_2 l_1 l_2 \dot{\theta}_2(\dot{\theta}_2-\dot{\theta}_1)\sin(\theta_1-\theta_2)
\end{aligned} \tag{9.3.960}$$

So we have ultimately for the Euler-Lagrange equation with the generalized coordinate θ_1:

$$\begin{aligned}
&-m_2 l_1 l_2 \dot{\theta}_1 \dot{\theta}_2 \sin(\theta_1-\theta_2)-(m_1+m_2) g l_1 \sin(\theta_1)\\
&-(m_1+m_2)\ddot{\theta}_1 l_1^2-m_2 l_1 l_2 \ddot{\theta}_2 \cos(\theta_1-\theta_2)-m_2 l_1 l_2 \dot{\theta}_2(\dot{\theta}_2-\dot{\theta}_1)\sin(\theta_1-\theta_2)\\
&=-(m_1+m_2) g l_1 \sin(\theta_1)-(m_1+m_2)\ddot{\theta}_1 l_1^2-m_2 l_1 l_2 \ddot{\theta}_2 \cos(\theta_1-\theta_2)-m_2 l_1 l_2 \dot{\theta}_2^2 \sin(\theta_1-\theta_2)\\
&=0
\end{aligned} \tag{9.3.961}$$

We have also for θ_2:

$$\begin{aligned}
&\frac{\partial L}{\partial \theta_2}=-m_2 l_1 l_2 \dot{\theta}_1 \dot{\theta}_2 \sin(\theta_1-\theta_2)-m_2 g l_2 \sin(\theta_2)\\
&\frac{\partial L}{\partial \dot{\theta}_2}=m_2 l_2^2 \dot{\theta}_2+m_2 l_1 l_2 \dot{\theta}_1 \cos(\theta_1-\theta_2)\\
&\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{\theta}_2}=m_2 l_2^2 \ddot{\theta}_2+m_2 l_1 l_2 \ddot{\theta}_1 \cos(\theta_1-\theta_2)+m_2 l_1 l_2 \dot{\theta}_1(\dot{\theta}_2-\dot{\theta}_1)\sin(\theta_1-\theta_2)
\end{aligned} \tag{9.3.962}$$

So we have ultimately for the Euler-Lagrange equation with the generalized coordinate θ_2:

$$\begin{aligned}
&m_2 l_1 l_2 \dot{\theta} \dot{\theta}_2 \sin(\theta_1-\theta_2)-m_2 g l_2 \sin(\theta_2)\\
&-m_2 l_2^2 \ddot{\theta}_2-m_2 l_1 l_2 \ddot{\theta}_1 \cos(\theta_1-\theta_2)-m_2 l_1 l_2 \dot{\theta}_1(\dot{\theta}_2-\dot{\theta}_1)\sin(\theta_1-\theta_2)\\
&=-m_2 g l_2 \sin(\theta_2)-m_2 l_2^2 \ddot{\theta}_2-m_2 l_1 l_2 \ddot{\theta}_1 \cos(\theta_1-\theta_2)+m_2 l_1 l_2 \dot{\theta}_1^2 \sin(\theta_1-\theta_2)=0
\end{aligned} \tag{9.3.963}$$

This gives us the finally the following system of differential equations:

$$\begin{aligned}
&(m_1+m_2) g l_1 \sin(\theta_1)+(m_1+m_2)\ddot{\theta}_1 l_1^2+m_2 l_1 l_2 \ddot{\theta}_2 \cos(\theta_1-\theta_2)+m_2 l_1 l_2 \dot{\theta}_2^2 \sin(\theta_1-\theta_2)=0\\
&m_2 g l_2 \sin(\theta_2)+m_2 l_2^2 \ddot{\theta}_2+m_2 l_1 l_2 \ddot{\theta}_1 \cos(\theta_1-\theta_2)+m_2 l_1 l_2 \dot{\theta}_1^2 \sin(\theta_1-\theta_2)=0
\end{aligned} \tag{9.3.964}$$

With Maple™ 4.00b this gives for the representation of the coordinates of m_2 over time:

```
> with(plots):
> with(plottools):
> Eq1:=(m1+m2)*l1^2*diff(diff(theta1(t),t),t)+m2*l1*l2*cos(theta1(t)-theta2(t))
```

```
*diff(diff(theta2(t),t),t)+m2*l1*l2*sin(theta1(t)-theta2(t))*diff(theta2(t),t)^2
+(m1+m2)*g*l1*sin(theta1(t))=0;
> Eq2:=m2*l1*l2*cos(theta1(t)-theta2(t))*diff(diff(theta1(t),t),t)+m2*l2^2
*diff(diff(theta2(t),t),t)-m2*l1*l2*sin(theta1(t)-theta2(t))*diff(theta1(t),t)^2
+m2*g*l2*sin(theta2(t))=0;
> m1:=2;m2:=3;l1:=6;l2:=4;g:=9.81;
> ff:=dsolve(Eq1,Eq2,theta1(0)=0.5,D(theta1)(0)=4,theta2(0)=1,D(theta2)(0)=
-2,theta1(t),theta2(t),type=numeric,output=listprocedure);
> Theta1:=subs(ff,theta1(t));Theta2:=subs(ff,theta2(t));
> X1:=t->l1*sin(Theta1(t));
> Y1:=t->l1*cos(Theta1(t));
> X2:=t->l1*sin(Theta1(t))+l2*sin(Theta2(t));
> Y2:=t->l1*cos(Theta1(t))+l2*cos(Theta2(t));
> plot([Y2,X2,0..100],numpoints=100);
```

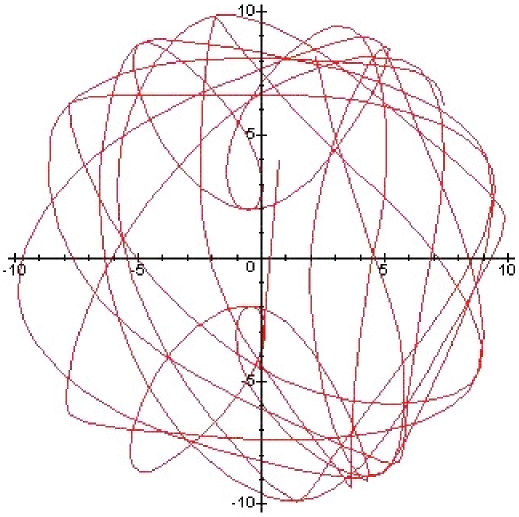


Figure 9.101: **Plot with Maple™ 4.00b of the double pendulum external mass**

And in real life:

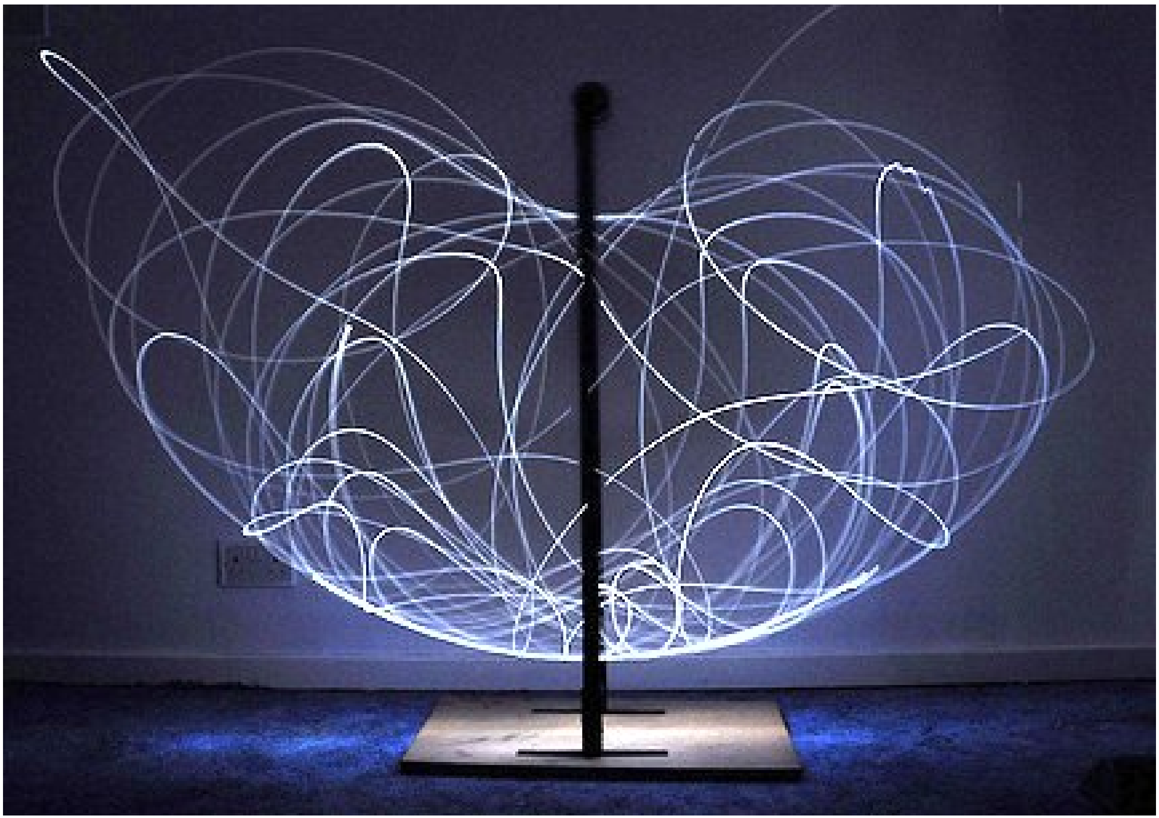
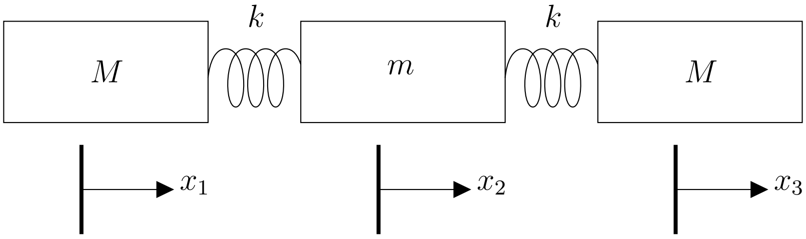


Figure 9.102: **Long exposure of double pendulum exhibiting chaotic motion tracked with a LED (source: Wikipedia)**

9.3.10.10 Triple linear coupled pendulum

Consider three masses on the x -axis joined by springs as shown in the figure below:



The spring forces are assumed to be linear (small displacements, Hooke's law as derived at page 3744), and the mass is constrained to stay on the x-axis.

Using a different coordinate for each mass, Newton's second law yields the set of equations:

$$\begin{aligned}\ddot{x}_1&=-\frac{k}{M}(x_1-x_2)\\ \ddot{x}_2&=-\frac{k}{m}(x_2-x_1)-\frac{k}{m}(x_2-x_3)\\ \ddot{x}_3&=-\frac{k}{M}(x_3-x_2)\end{aligned}\tag{9.3.965}$$

The system of masses is vibrating. We seek the common frequencies, ω, such that all masses vibrate at this same frequency. These are the normal modes. Let:

$$x_i=x_{i0}e^{\mathrm{i}\omega t},\quad i=1,2,3\tag{9.3.966}$$

Substituting this set into the previous set of equalities, we may rewrite it as:

$$\begin{pmatrix}\frac{k}{M} & -\frac{k}{M} & 0\\ -\frac{k}{m} & \frac{2k}{m} & -\frac{k}{m}\\ 0 & -\frac{k}{M} & \frac{k}{M}\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=+\omega^2\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}\tag{9.3.967}$$

with the common factor $e^{\mathrm{i}\omega t}$ divided out. We have a matrix-eigenvalue equation with the matrix asymmetric (see section Linear Algebra page 1997). The secular equation is:

$$\begin{vmatrix}\frac{k}{M}-\omega^2 & -\frac{k}{M} & 0\\ -\frac{k}{m} & \frac{2k}{m}-\omega^2 & -\frac{k}{m}\\ 0 & -\frac{k}{M} & \frac{k}{M}-\omega^2\end{vmatrix}=0\tag{9.3.968}$$

This leads to the following determinant:

$$\omega^2\left(\frac{k}{M}-\omega^2\right)\left(\omega^2-\frac{2k}{m}-\frac{k}{M}\right)=0\tag{9.3.969}$$

The eigenvalues are:

$$\omega^2=0,\quad\frac{k}{M},\quad\frac{k}{M}+\frac{2k}{m}\tag{9.3.970}$$

all real. The corresponding eigenvectors are determined by substituting the eigenvalues back into:

$$\begin{pmatrix} \frac{k}{M} & -\frac{k}{M} & 0 \\ -\frac{k}{m} & \frac{2k}{m} & -\frac{k}{m} \\ 0 & -\frac{k}{M} & \frac{k}{M} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = +\omega^2 \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \tag{9.3.971}$$

one eigenvalue at a time. For $\omega^2 = 0$, the above system, yields:

$$x_1 - x_2 = 0, \quad -x_1 + 2x_2 - x_3 = 0, \quad -x_2 + x_3 = 0 \tag{9.3.972}$$

Then we get:

$$x_1 = x_2 = x_3 \tag{9.3.973}$$

This describes pure translation with no relative motion of the masses and no vibration. For $\omega^2 = k/M$, the above system yields to:

$$x_1 = -x_3, \quad x_2 = 0 \tag{9.3.974}$$

The two outer masses are moving in opposite direction. The central mass is stationary. For $\omega^2 = k/M + 2k/m$, the eigenvector components are:

$$x_1 = x_3, \quad x_2 = -\frac{2M}{m} x_1 \tag{9.3.975}$$

The two outer masses are moving together. The central mass is moving opposite to the two outer ones. The net momentum is zero.

Any displacement of the three masses along the x -axis can be described as a linear combination of these three types of motion: translation plus two forms of vibration.

9.3.10.11 Inverted Pendulum

An inverted pendulum is a pendulum that has its center of mass above its pivot point. It is often implemented with the pivot point mounted on a cart that can move horizontally as shown in the photo below:

Figure 9.103: **A famous business application (SEGWAY™) of mobile inverted pendulum**

Most applications limit the pendulum to 1 degree of freedom by affixing the pole to an axis of rotation. Whereas a normal pendulum is stable when hanging downwards, an inverted pendulum is inherently unstable, and must be actively balanced in order to remain upright; this can be done either by applying a torque at the pivot point, by moving the pivot point horizontally as part of a feedback system, changing the rate of rotation of a mass mounted on the pendulum on an axis parallel to the pivot axis and thereby generating a net torque on the pendulum, or by oscillating the pivot point vertically. A simple demonstration of moving the pivot point in a feedback system is achieved by balancing an upturned broomstick on the end of one's finger. The inverted pendulum is a classic problem in dynamics and control theory and is used as a benchmark for testing control strategies. Arguably the most prevalent example of an inverted pendulum is a human being. A person with an upright body needs to make adjustments constantly to maintain balance whether standing, walking, or running.

The inverted pendulum is a superb academic example of the application of Lagrangian Mechanics to another coupled system than the that of the double pendulum seen previously. Moreover, we can make a very good parallel with the watch systems as they are used in many R&D departments of watch manufacturers.

For the mathematical study let us consider the following experiment (The rod is considered massless!):

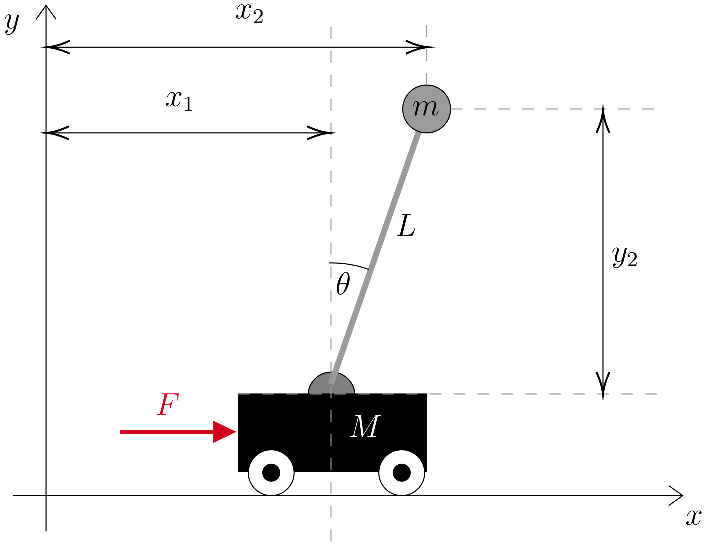


Figure 9.104: **A schematic drawing of the inverted pendulum on a cart**

The position of the mass base carriage M is x_1, the angle of the rod of length L is θ, the scalar force applied on the base of carriage is F. The coordinates of the weight at the extremity of the rod will be (x_2, y_2).

We will deduce the system of differential equations which describes the inverse pendulum dynamics using the Lagrange equations (see section Analytical/Lagrangian Mechanics page 3152).

The kinetic energy of the carriage and the extreme weight of the rod are therefore respectively:

$$\begin{aligned}E_{c,1}&=\frac{1}{2}M\dot{x}^2\\E_{c,2}&=\frac{1}{2}m(\dot{x}_2^2+\dot{y}_2^2)\end{aligned}\tag{9.3.976}$$

With:

$$\begin{aligned}x_2&=x_1+L\sin(\theta)\\y_2&=L\cos(\theta)\end{aligned}\tag{9.3.977}$$

such that:

$$\begin{aligned}\dot{x}_2&=\dot{x}_1+L\dot{\theta}\cos(\theta)\\\dot{y}_2&=-L\dot{\theta}\sin(\theta)\end{aligned}\tag{9.3.978}$$

Thus the total kinetic energy is written:

$$\begin{aligned}E_c=E_{c,1}+E_{c,2}&=\frac{1}{2}M\dot{x}_1^2+\frac{1}{2}m(\dot{x}_2^2+\dot{y}_2^2)=\frac{1}{2}M\dot{x}_1^2+\frac{1}{2}m\left((\dot{x}_1+L\dot{\theta}\cos(\theta))^2+(-L\dot{\theta}\sin(\dot{\theta}))^2\right)\\&=\frac{1}{2}M\dot{x}_1^2+\frac{1}{2}m\left(\dot{x}_1^2+2\dot{x}_1L\dot{\theta}\cos(\theta)+L^2\dot{\theta}^2\cos^2(\theta)+L^2\dot{\theta}^2\sin^2(\theta)\right)\\&=\frac{1}{2}M\dot{x}_1^2+\frac{1}{2}m\left(\dot{x}_1^2+2\dot{x}_1L\dot{\theta}\cos(\theta)+L^2\dot{\theta}^2\right)=\frac{1}{2}(M+m)\dot{x}_1^2+\frac{1}{2}m\left(2\dot{x}_1L\dot{\theta}\cos(\theta)+L^2\dot{\theta}^2\right)\end{aligned}\tag{9.3.979}$$

The potential energy will be considered taking into account only the mass at the extremity of the rod (the carriage being at the level of the ground and the mass of the rod being considered negligible):

$$E_p=mgy_2=mgL\cos(\theta) \tag{9.3.980}$$

The classical Lagrangian is then:

$$L=T-V=E_c-E_p=\frac{1}{2}(M+m)\dot{x}_1^2+m\dot{x}_1L\dot{\theta}\cos(\theta)+\frac{1}{2}mL^2\dot{\theta}^2-mgL\cos(\theta) \tag{9.3.981}$$

We will take as generalized coordinates of the system: (x_1,θ). Which are the variables we have just seen as sufficient to describe the state of the Lagrangian.

Therefore, the Lagrange equations that we have are (compared to the usual writing we have multiplied each line by -1):

$$\begin{gathered}\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{x}_1}-\frac{\partial L}{\partial x_1}=0\\ \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{\theta}}-\frac{\partial L}{\partial \theta}=0\end{gathered} \tag{9.3.982}$$

We inject the preceding results into respectively each of the equations to obtain by already directly calculating the partial derivative:

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\Big((M+m)\dot{x}_1+L\dot{\theta}\cos(\theta)\Big)-0=0\\ &\frac{\mathrm{d}}{\mathrm{d}t}\Big(m\dot{x}_1L\cos(\theta)+mL^2\dot{\theta}\Big)-\Big(-m\dot{x}_1L\dot{\theta}\sin(\theta)+mgL\sin(\theta)\Big)=0\end{aligned} \tag{9.3.983}$$

and therefore:

$$\begin{aligned}&(M+m)\ddot{x}_1+L\ddot{\theta}\cos(\theta)-mL\dot{\theta}^2\sin(\theta)=0\\ &m\ddot{x}_1L\cos(\theta)+mL^2\ddot{\theta}-mgL\dot{\theta}\sin(\theta)=0\end{aligned} \tag{9.3.984}$$

Let us now write this system in matrix form:

$$\begin{bmatrix}M+m & mL\cos(\theta)\\ mL\cos(\theta) & mL^2\end{bmatrix}\begin{pmatrix}\ddot{x}_1\\ \ddot{\theta}\end{pmatrix}=\begin{bmatrix}mL\dot{\theta}^2\sin(\theta)\\ mgL\sin(\theta)\end{bmatrix} \tag{9.3.985}$$

It is a typical Lagrangian system with the square inertial matrix of dimension 2 multiplied by the acceleration vector. The term $mL\theta\dot{\theta}^2\sin(\theta)$ is a centripetal force whereas the term $mgL\sin(\theta)$ is the term of gravity.

We solve this system by inverting the square inertia matrix of dimension 2 as we have demonstrated in the section of Linear Algebra and we get:

$$\begin{aligned}\ddot{x}_1&=\frac{mg\sin(\theta)\cos(\theta)-mL\dot{\theta}^2\sin(\theta)}{m\cos^2(\theta)-(M+m)}\\ \ddot{\theta}&=\frac{-(M+m)g\sin(\theta)+mL\dot{\theta}^2\sin(\theta)\cos(\theta)}{mL\cos^2(\theta)-(M+m)L}\end{aligned} \tag{9.3.986}$$

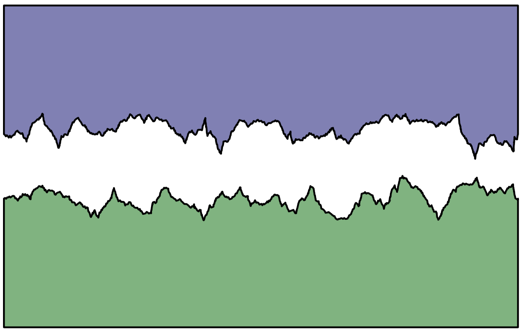
Now, the state of the system would be adequately described in the phase space by the following variables:

$$\begin{pmatrix}\dot{x}_1\\ \ddot{x}_1\\ \dot{\theta}\\ \ddot{\theta}\end{pmatrix}=\begin{bmatrix}\dot{x}_1\\ \dfrac{mg\sin(\theta)\cos(\theta)-mL\dot{\theta}^2\sin(\theta)}{m\cos^2(\theta)-(M+m)}\\ \dot{\theta}\\ \dfrac{-(M+m)g\sin(\theta)+mL\dot{\theta}^2\sin(\theta)\cos(\theta)}{mL\cos^2(\theta)-(M+m)L}\end{bmatrix} \tag{9.3.987}$$

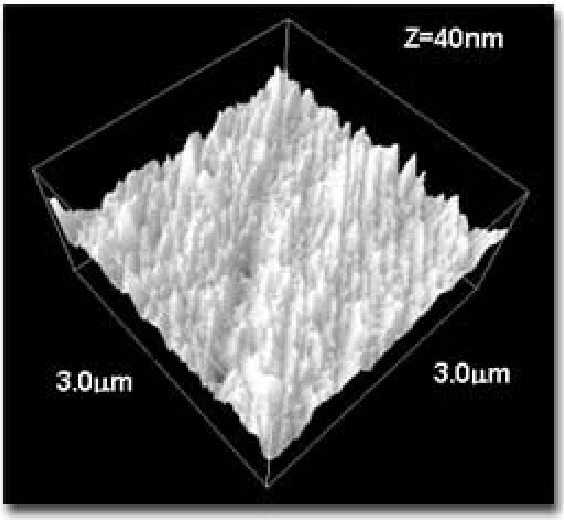
We can obviously simplify this expression by making first-order Taylor developments for small angles. What should follow is only the application of iterative numerical methods.

9.3.11 Tribology

Definition (#9.3.450): "Tribology" is the science of friction (very intuitive concept to everyone because we can feel its effects in everyday life) that occur when two contacting surfaces are moved relative to each other producing a force that opposes to the motion. This field has for purpose to study not only the conditions of friction between the bodies but also the various phenomena and physical and technical aspects in relative movement of the outer surfaces of the bodies. This field is very important in Manufacturing Technology and Aeronautics.



Below we can see a three-dimensional atomic force microscope (AFM) image of a polish mark on a piece of steel. The scan range in X and Y is 3 micrometers and the entire Z range is 40 nanometers:



Definition (#9.3.451): "Roughness" refers to the collection of small defects, peaks or hollows that form the texture of a surface of a fluid, solid or gas.

Most of these phenomena related to friction can be understood as a first approximation based on the phenomenological laws of friction set in the 118th century (holocene calendar) by Guillaume Amontons and Charles August Coulomb (but already highlighted by Leonardo Da Vinci 200 years ago) from the concept of coefficient of friction.

The analysis of a tribological system introduced 4 fundamental components:

1. The body subject to the study (this is usually a solid)
2. The support body on which acts directly the studied body (it can also be a solid, liquid or gas)
3. The intermediate material interposed between the two preceding bodies
4. And finally the environment.

Guillaume Amontons and Charles August Coulomb already observed two types of friction a priori distinct:

1. The "static friction" is the one who shows a resistance to the movement when object put on a planar surface is about the limit to slip and when we imposed to it a tensile force $\vec{F}_T$ (tangent to the plane). This opposition to the tensile force is experimentally measured as being proportional to the weight $\vec{P}$ of the object.

   But comes a limit value of the tangential tensil force from which the object starts to slip. This is what we note:

   $$\boxed{\left|\left|\vec{F}_{T,\text{lim}}\right|\right|=\mu_S\left|\left|\vec{P}\right|\right|} \tag{9.3.988}$$

   where $\vec{F}_{T,\text{lim}}$ is the tensile limit force to move the initially static object, μ_S is the "static friction coefficient" that is dimensionless and expresses the proportionality of the limit friction force with the weight $\vec{P}$ or of its normal shape relatively to the application plane. This is why this last relation is often denoted:

   $$\boxed{\left|\left|\vec{F}_{T,\text{lim}}\right|\right|=\mu_S\left|\left|\vec{N}\right|\right|} \tag{9.3.989}$$

   Remark

   In practice, the friction coefficient can be determined very easily using a simple dynamometer to measure the limiting force and a scale to measure the weight of the object being studied. For example, the static friction coefficient for steel on steel (as in railway applications) ranges from 0.15 to 0.6, while the dynamic friction coefficient ranges from 0.09 to 0.6. The Swiss Re 460 locomotive develops a maximum traction force of about 300 [kN] and has a mass of roughly 80 tons, whereas a full train convoy in Switzerland typically has a mass between 400 and 800 tons.

   We experimentally observe that contrary to the intuition, the tensile limit force is in first approximation independent of the contact surface between the object and the support (within current physical case obviously because the small the contact surface, the higher the pressure is for a same weight, then the contact surface may become plastic at high pressures...!).

   So we can now sate the both Amonton's law:

- "Amontons' first law of friction" also named "Leonardo da Vinci's law of friction" state that the friction force experienced by two dry bodies whose planar surfaces are in contact is independent of the area of contact between the bodies.
- "Amontons' second law of friction" state that the friction force (limiting value for static friction, actual value for kinetic friction) is proportional to the normal force.

In other words, if one kilo of sugar is placed on a table. To move the object, of weight $\vec{P}$ (mass multiplied by the gravitational constant), it must applied a force $\vec{F}$ parallel to the table surface. But experience shows that this object will not move until the force $\vec{F}_T$ is less than a minimum force $\vec{F}_{T,\text{lim}}$. And Amontons and Coulomb showed that this minimum force is directly proportional via a coefficient of static friction to the weight.

We can detail the approach of the above relation by imagining two surfaces having roughness in saw-tooth of an angle α and nested together:

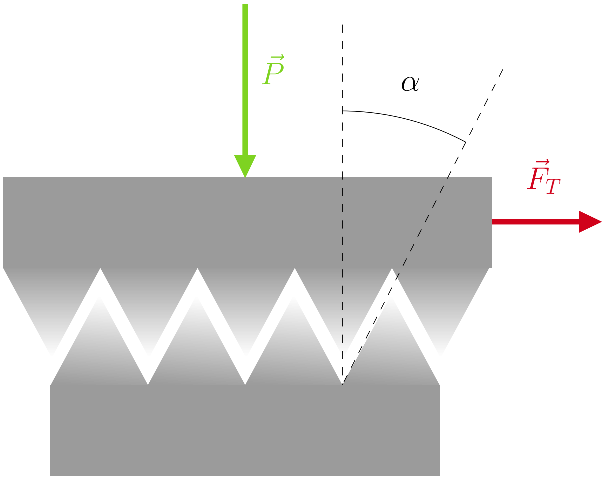


Figure 9.105: Amontons-Coulomb friction saw-tooth model

If we apply a normal force corresponding to the weight $\vec{P}$ and a horizontal force $\vec{F}_T$ we have because of the saw-tooth in the limiting case the following situation:

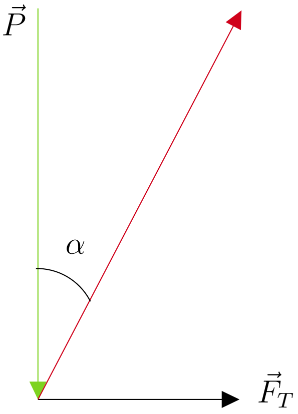


Figure 9.106: Amontons-Coulomb friction saw-tooth free-body diagram

then we see well that the movable piece will begin to move when there will be slip, that is when:

$$\boxed{\|\vec{F}_T\| = \tan(\alpha)\|\vec{P}\|} \tag{9.3.990}$$

As simplistic as it is, this approach allows to relate the (static) friction to the characteristics of roughness. Furthermore typical experimental values of the static coefficients friction of the order of 0.3 correspond surface roughness slope of the order of 15-20 degrees, which is quite compatible with the typical characteristics that can be measured for roughness of surfaces!

However, this argument is based on an implicit assumption: the perfect interlocking between the roughness of both surfaces. We speak in this case of "commensurable surfaces". This is of course not the case in general in the Nature: even at the atomic scale, two ideal surfaces are slightly different in inter-atomic distance that prevent nesting or the presence of air avoid a perfect interlocking. A slight disparity is enough to make very irregular the distribution of contact points between the two surfaces unlike the commensurable case. We speak then of "incommensurable surfaces".

2. The "dynamic friction" is the one who resists when an object placed on a plane is already slipping. This opposition to traction force is also experimentally again proportional to the weight of the object such that:

$$\boxed{\|\vec{F}_T\| = \mu_D \|\vec{P}\|} \tag{9.3.991}$$

but with the "dynamic friction coefficient" μ_D (which exists in several subfamilies: rolling friction coefficient, sliding friction coefficient, ...) - also sometimes named "kinetic friction coefficient" and denoted μ_k - which is usually much smaller than the coefficient of static friction:

$$\mu_S \gg \mu_D \tag{9.3.992}$$

So the above relation states that for two dry solid surfaces sliding against one another, the magnitude of the kinetic friction exerted through the surface is independent of the magnitude of the velocity (i.e., the speed) of the slipping of the surfaces against each other.

So the friction is not the same before our object start to slip than after he start slipping. This corresponds very well with our daily experience of friction (when moving furniture in our homes for examples).

Again, we experimentally observe that contrary to intuition, the pulling force is again independent of the contact area between the object and its support.

Thus, if we a packet of one kilo of sugar on one of its face or on its edge, assuming that at the atomic level its properties are the same everywhere, the friction force is the same (if the surface quality is the same from all sides of the packet of sugar)!

Another quite surprising fact relates to the typical values of the both coefficients of friction, which deviates relatively little $\mu_D \cong \mu_S \cong 0.3$, for very different surfaces from each other. The technologies, however, enables the design of surfaces with coefficients of friction that are much smaller ($\mu_D \cong \mu_S \cong 0.001$) or larger ($\mu_D \cong \mu_S \gg 1$).

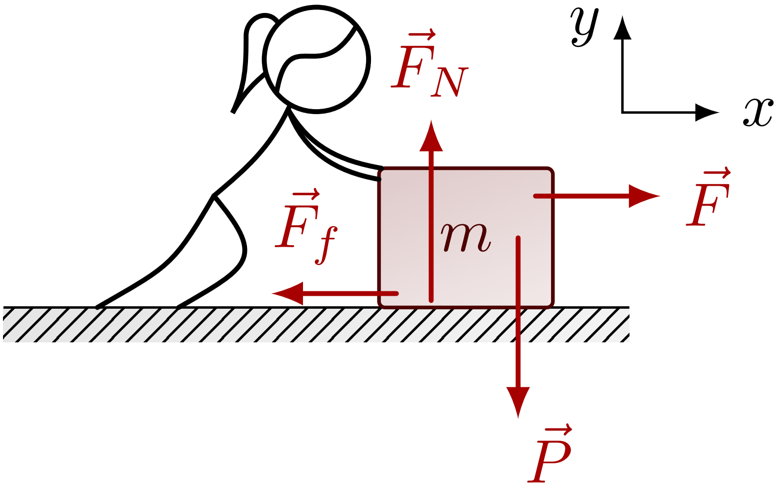
The two above relation relating pressure and pull force are name "Coulomb's friction laws", and are traditionally written:

$$\boxed{\|\vec{F}_{T,\text{lim}}\| = \mu_S \|\vec{N}\| \qquad \|\vec{F}_T\| = \mu_D \|\vec{N}\|} \tag{9.3.993}$$

and with the following notation in many contemporary high-school textbooks:

$$F_{f,\text{lim}}=\mu_S F_N=\mu_S PA \qquad F_f=\mu_D F_N=\mu_D PA \tag{9.3.994}$$

where F_N is the normal force to the support, P is the pressure (weight) and A the contact surface.



✍**Example:**

The mass of a crate is $m=32$ [kg] and the coefficient of static friction between the crate and the floor is $\mu_S=0.53$. Therefore the force necessary to start the crate moving is given as:

$$F=\mu_S mg=0.53\cdot 32\cdot 9.8\cong 166.20\,[\text{N}]$$

The largest stone ever moved by manpower alone, i.e. without the use of animals or machines, is the "Thunder Stone", an enormous boulder of granite serving as the pedestal of the famous Bronze Horseman statue of Peter the Great at St Petersburg, Russia. Before being cut into shape as the pedestal, the Thunder Stone measured 7 [m] by 14 [m] by 9 [m], and was estimated to weight around $1,500$ tonnes. It was moved 6 [km] overland to the Gulf of Finland from the marsh in Lakhta, north of the Gulf by dragging it across the frozen Russian countryside during the winter of 11768 (holocene calendar), a process taking nine months and requiring 400 men, pulling it upon a metallic sledge.

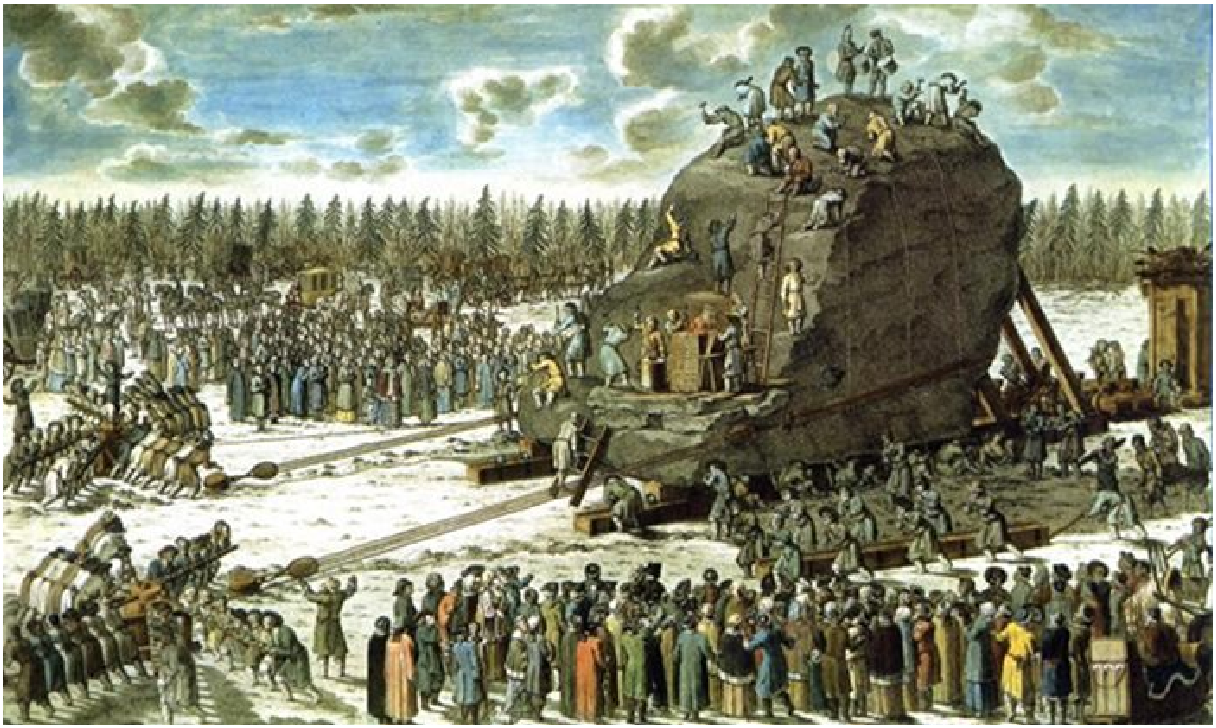


Figure 9.107: **The Transportation of the Thunder-stone in the Presence of Catherine II; Engraving by I. F. Schley of the drawing by Yury Felten, 11770**

Remarks

R1. In 12012 (holocene calendar) a Syrian man was able to move a 150 Tons train by its own force (sometimes you see in the U.S. nine locomotives pulling a lot of cars about 200 but once they are rolling they shut off a couple of engines to save fuel). In 12019 (holocence calendar) a canadian man was able to move a 189 tons airplane (c-17 globemaster III heavy transport).

R2. The reader can easily found on Internet the original video footage of a 250 tonnes excavation of the Mussolini's Monolith that was mostly done on simple wooden sledsm by human strength and with the help of other animals like cows.

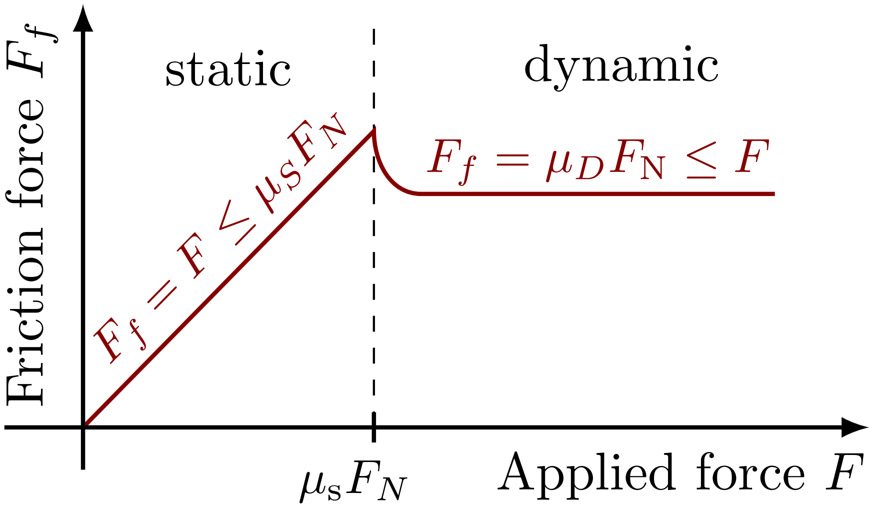


Figure 9.108: **Horizontal tribology scenario configuration (author: Izaak Neutelings)**

The naive origin of friction between two solids is therefore due to the fact that:

- Any solid is never smooth but has asperities which makes the contact surface rough (asperities overlap partially or not and cause more or less friction).
- Impurities between the two contact surfaces are often more important in terms of friction sources than the surface asperities imbrication.
- Friction is weakly dependent on the surface area because the roughness at the atomic scale is such that only a very small percentage of the total area of the two objects are actually in contact (actual contact area is therefore much smaller than the apparent contact surface) which explains that the tangential traction force is proportional to the weight because it forces the actual contact area to increase.

Remark

When two metals (pure - and clean...) of identical nature are in contact trough a smooth surface in a favourable environment (at best: in vacuum) when we can sometimes observe a "cold welding" (in vacuum it's a systematic and not just "sometimes") which is a case of extreme friction because sometimes it becomes almost impossible (or simply just impossible!) to then move one of the elements relative to the other (the result also depends slightly on the initial pressure applied to develop cold welding on the two objects of the same type). Cold welding was recognized as a general physical phenomenon of the materials science in the 11940s (holocene calendar). The reason of this unexpected behaviour is that when the atoms in contact are all of the same nature, they have no way to "know "they are present in different metal parts. When there are other atoms in the oxides and greases and other kinds of surface contaminants, the atoms "know" that they do not belong to the same object (there are widely used industrial adhesives using this principle for emergency repairs...).

Now to see a famous case of treatment of static and dynamic friction let us consider the banked curve situation!

First, the question that appears quite quickly is: What provides the centripetal force on the car for a banked curve? It is the Normal force? Gravity?.... No!!! they are both vertical forces. They have no component in the horizontal direction where we need a centripetal force.

What is it? It's friction!!!! But which friction is it? Static or Dynamic? Well maybe a bit of both and another friction, that of just sinking into the road. In our previous, discussion of friction implicitly considered just ideal rigid surfaces. But real roads do compress a bit under car weight.

In any case, friction is a reactive force that tries to prevent motion between surfaces and only turns on to try to prevent that. It's like the Normal force which only turns on if you press inward on a surface.

In the car-on-a-curve, you angle the car wheels and ideally static friction will be push exactly opposite to the car's current direction.

This static friction can be resolved into two orthogonal components. One component of frictions acts opposite to the wheel direction and the other acts perpendicular to that direction and toward a center of curvature.

Let's analyse the forces and see what relations we get.

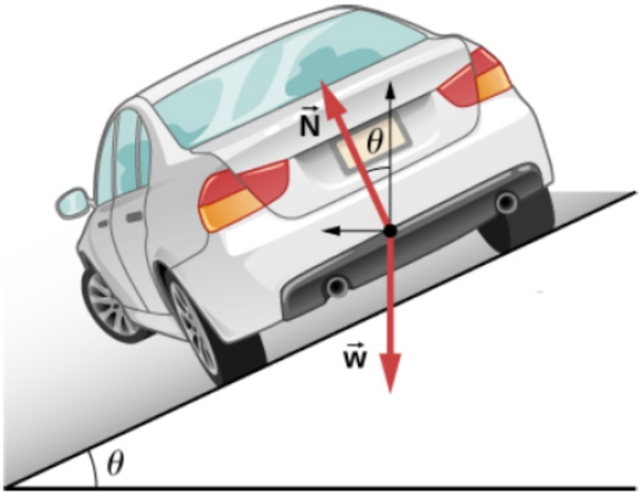


Figure 9.109: **Banked curve analysis (source: OpenStax)**

In this case, we chose the x direction to be the horizontal direction with the x axis toward the curve center. The y is vertical with up positive.

We assume a vehicle is making the turn and that we observe that it is not sliding. This is the key condition!

In the x direction, we have:

$$F_{\text{cen}} = m\frac{v^2}{r} = F_n \sin(\theta) \tag{9.3.995}$$

where F_n is the Normal force (with implicit direction Normal and outward from the surface) and θ is the angle of the normal force from the vertical and also, by geometry, the angle of the incline or banking from the horizontal. Here the x component of the normal force is providing the centripetal force.

In the y direction, we have no motion. The path is rigid, and so vehicle can't sink into the path and in any perturbation to lift off the path, gravity pulls[20] the vehicle back. So we must have balanced forces. So we have:

$$F_n \cos(\theta) = mg \tag{9.3.996}$$

where $F_n \cos(\theta)$ upward component of the Normal force and mg is the force of gravity.

Physics has given us two equations. From these we can solve for two unknowns. The most interesting variables to take to be unknowns are the normal force F_{nor} and the angle of the incline θ (which is also gives the direction of the normal force, of course).

---

[20] Keep in mind that's in the framework of the Newtonian model of gravity as we that in General Relativity there are no "pull".

The normal force magnitude is obtained by adding the squares of the both previous relations. We get:

$$F_n = m\sqrt{\left(\frac{v^2}{r}\right)^2 + g^2} \tag{9.3.997}$$

that is obviously bigger than mg!

To get the incline angle, we divide equation 9.3.995 by equation 9.3.996 and we get:

$$\tan(\theta) = \frac{v^2}{rg} \tag{9.3.998}$$

and:

$$\boxed{\theta = \tan^{-1}\left(\frac{v^2}{rg}\right)} \tag{9.3.999}$$

This is the "banking angle formula". For a given speed and radius, the formula gives the angle one should bank a road at in order for a car to go around the curve with only the Normal force and gravity combining to give the acceleration.

Note the formula is independent of mass! This is because the Normal force is linearly dependent on mass, and so mass cancels out.

The Normal force provides all the force needed for the motion. A car at the right speed would make a properly banked curve ideally even a road were ideal frictionless ice!!!!!!!!!!!!

Note also that if v^2/r goes to infinity, θ goes to 90° and the car plane would be perpendicular to the ground.

The banking angle formula equation always seems a little mysterious. How does the Normal force know to be just strong enough? But the answer is that you have assumed that you are making the curve. Therefore the Normal force must be strong enough by the given condition. It's actually a constraint force!

If you did a banking analysis with the friction force turned on and solved for Normal force and friction force, then you would find that the friction force goes to zero for the banking angle given by the banking angle formula.

Why do civil engineers bank curves using the banking angle formula? Somewhat obviously so that the cars don't need to rely on friction to make the curve. Failure of the Normal force doesn't much happen, except maybe on mud roads. The friction force can fail even on dry roads. Water and ice and snow reduce friction substantially. Another obvious reason is that turning is much easier on a properly banked road. If you go at the specified speed, there is little need to turn the wheels at all and there is no tendency to loss of control.

Yet another reason it is easier on the road structure, the car structure, and the car contents including the passengers. The above analysis actually applies to every bit of the car and contents. If you go at the specified speed, anti-compressional forces do all the acceleration. Antishear forces (of which friction is only one) in the car and the road are never called on. Antishear forces are any force that resists layers sliding over one another including those bonded together inside solids. Calling on such forces repeatedly leads to rapid wear. It's also more comfortable for passengers to be just compressed into their seats making a curve and have no tendency to slide sideways.

Of course, if you drive at the wrong speed around a curve, antishear forces do get called on with all their discomfort, tendency to wear, and chance of failure. So you should always drive at the specified speed.

Too fast and you tend to slide outward. Too slow and you tend to slide inward. It's better to err on the too slow side, of course, because failure is probably less catastrophic. In very icy conditions what are you doing on the road anyway?

Now civil engineers can't use just one banking angle. They must go from straight, unbanked road to maximally banked road continuously. And they can't ask drivers to change speed more than once to the curve speed. So they must vary the radius of curvature to allow a continuous change from unbanked to maximally banked to unbanked again.

Now all of the analysis above is a bit idealized. There are probably many complicating effects that real engineers must include in their designs. One thing, which is probably minor in this context is that cars are not really, point masses. Forces are needed not just to make the car center of mass go round curves, but also rotate the car.

9.3.11.1 Exponential Friction

It is customary to designate as "exponential friction" the friction generated by non-toothed belts (very important case in practice) on a fixed support!

To investigate this case friction let consider the figure below where β is the winding angle of the belt around the pulley where we neglect the thickness and mass of the belt relative to the other elements and where we place ourselves at the limit of slipping and in uniform motion:

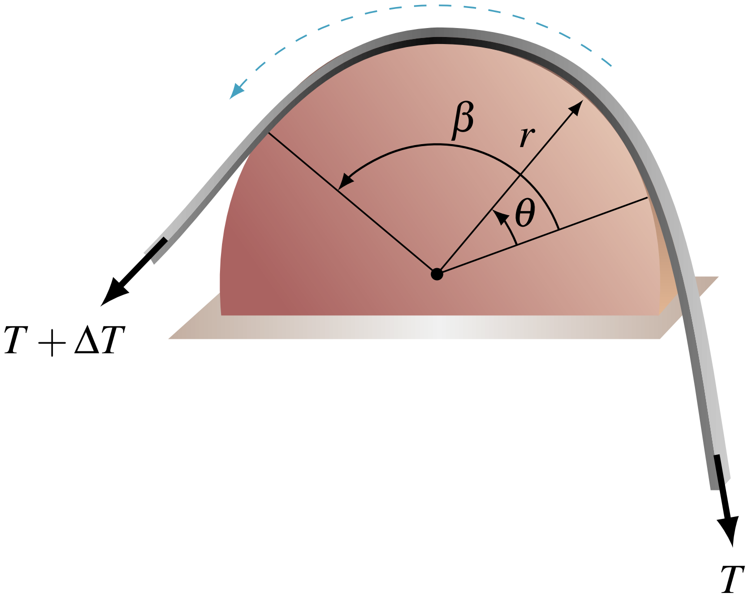


Figure 9.110: Forces between a rope and capstan

By the existence of friction there is obviously a force differential $\mathrm{d}T$ on a point of contact but this is due only to friction since we suppose a uniform rotation. Let's take a closer look to this by isolating a portion of the capstan and applying the fundamental principle of static (with subtlety...):

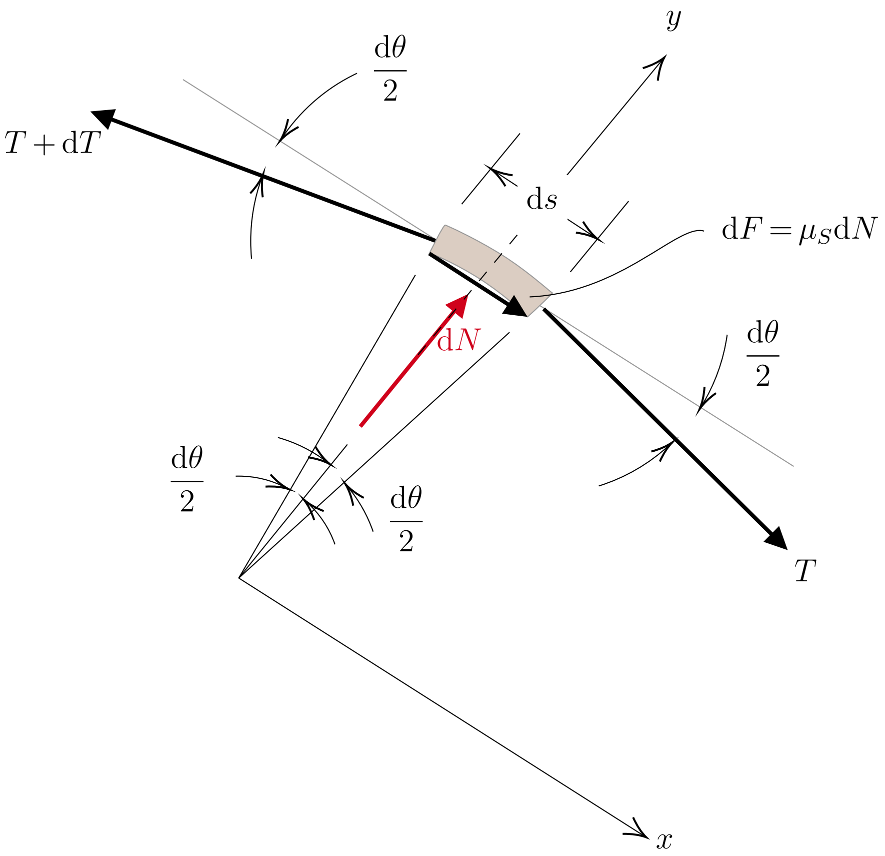


Figure 9.111: **Zoom on a differential element of the rope and capstan**

$$\begin{aligned}\sum F_x=0:\quad & T\cos\left(\frac{d\theta}{2}\right)+\mu_S dN-(T+dT)\cos\left(\frac{d\theta}{2}\right)=0\\ \sum F_y=0:\quad & dN-(T+dT)\sin\left(\frac{d\theta}{2}\right)-T\sin\left(\frac{d\theta}{2}\right)=0\end{aligned}\qquad(9.3.1000)$$

Thus after simplification:

$$\begin{aligned}\sum F_x=0:\quad & \mu_S dN-dT\cos\left(\frac{d\theta}{2}\right)=0\\ \sum F_y=0:\quad & dN-dT\sin\left(\frac{d\theta}{2}\right)-2T\sin\left(\frac{d\theta}{2}\right)=0\end{aligned}\qquad(9.3.1001)$$

and as the angle is very small, we have:

$$\cos\left(\frac{d\theta}{2}\right)\cong\cos(0)=1\qquad(9.3.1002)$$

and for the sine we use the Maclaurin development (see section Sequences and Series page 1703) for small angles:

$$\sin\left(\frac{d\theta}{2}\right)\cong\frac{d\theta}{2}\qquad(9.3.1003)$$

Then we have:

$$\begin{aligned}\sum F_x=0: &\quad \mu_S \mathrm{d}N-\mathrm{d}T=0\\ \sum F_y=0: &\quad \mathrm{d}N-\mathrm{d}T\frac{\mathrm{d}\theta}{2}-2T\frac{\mathrm{d}\theta}{2}=0\end{aligned} \qquad (9.3.1004)$$

Then after rearranging:

$$\begin{aligned}\sum F_x=0: &\quad \mu_S \mathrm{d}N=\mathrm{d}T\\ \sum F_y=0: &\quad \mathrm{d}N=T\mathrm{d}\theta\end{aligned} \qquad (9.3.1005)$$

where the first equation is already known to us as it corresponds to the usual static horizontal friction already seen earlier (Coulomb's law). We can now combine these two relations to get:

$$\mu_S \mathrm{d}\theta=\frac{\mathrm{d}T}{T} \qquad (9.3.1006)$$

By integrating it comes therefore:

$$\int\limits_0^{\beta}\mu_S \mathrm{d}\theta=\int\limits_{T_1}^{T_2}\frac{\mathrm{d}T}{T} \qquad (9.3.1007)$$

Then we have:

$$\ln\left(\frac{T_2}{T_1}\right)=\mu_S\beta \qquad (9.3.1008)$$

and finally we get the "capstan equation" or "belt friction equation", also known as "Eytelwein's formula":

$$\boxed{T_2=T_1e^{\mu_S\beta}} \qquad (9.3.1009)$$

which is very important! For example in practice with a pull force of 500 [N], it produces an actual pull force across the fixed circular support only of 250 [N] because the friction is exponential. This is why it is better to use pulleys as fixed circular supports to raise a mass!

Remark

A "holding capstan" is a ratchet device that can turn only in one direction; once a load is pulled into place in that direction, it can be held with a much smaller force. A powered capstan, also named a "winch", rotates so that the applied tension is multiplied by the friction between rope and capstan. On a tall ship a holding capstan and a powered capstan are used in tandem so that a small force can be used to raise a heavy sail and then the rope can be easily removed from the powered capstan and tied off.

In rock climbing with so-named "top-roping", a lighter person can hold (belay) a heavier person due to this effect.

Obviously because of the friction we will have:

$$T_2>T_1 \qquad (9.3.1010)$$

which obviously implies that:

$$e^{\mu_S\beta}>1 \tag{9.3.1011}$$

and therefore:

$$\mu_S\beta>0 \tag{9.3.1012}$$

For the rest, the reader is referred to go to the section of Civil Engineering where the cylindrical rollers are presented in details.

9.3.11.2 Horizontal Viscous Friction

We have seen that in a first approximation, the frictional force in the case of slipping is proportional to the weight of a body by a friction coefficient which value depends on the nature and condition of the contact surface, but independent of the contact area. We have also already mention that experience shows that in typical cases the friction force is independent of the velocity communicated to the moving body.

By lubricating by a viscous fluid two contact surfaces, the frictional force is reduced and therefore depends on the speed (this is typically the case of car tires which are viscous plastics).

Remark

For recall, the term "viscous" does not necessarily mean that it flows and it drool. This means that the behaviour law depends on the deformation speed (see section Continuum Mechanics page 3775).

Let us now consider a mobile in contact with a plane support via a fluid or viscous material. We know there will be friction and we assume that it is proportional to speed:

$$\boxed{F_f=-kv} \tag{9.3.1013}$$

where k is the "coefficient of viscous friction".

We then by applying Newton's first law:

$$F=m\frac{\mathrm{d}v}{\mathrm{d}t}=-kv \tag{9.3.1014}$$

Thus we have:

$$\frac{1}{v}\mathrm{d}v=-\frac{k}{m}\mathrm{d}t \tag{9.3.1015}$$

By integrating it comes:

$$\ln(v)-\ln(v_0)=-\frac{k}{m}t \tag{9.3.1016}$$

Therefore:

$$\ln(v)=\ln(v_0)-\frac{k}{m}t \tag{9.3.1017}$$

Taking the exponential:

$$\boxed{v = e^{\ln(v_0)} e^{-\frac{k}{m}t} = v_0 e^{-\frac{k}{m}t}} \quad (9.3.1018)$$

Thus, the speed decreases exponentially form an initial velocity to a zero asymptotic value under the assumption of proportionality of friction with speed.

It is a relation very often used in computer animations to represent objects that seem to stop naturally. You simply choose to choose the value of k.

We observe an interesting thing is that heavy moving body decelerates more slowly because of friction forces than a slight moving body!

Let us now show how to calculate the power lost by friction. We know that:

$$P = \frac{\mathrm{d}W}{\mathrm{d}t} = \frac{\mathrm{d}(Fx)}{\mathrm{d}t} = \frac{\mathrm{d}F}{\mathrm{d}t}x + \frac{\mathrm{d}x}{\mathrm{d}t}F \quad (9.3.1019)$$

if the variation in the force is negligible compared to the speed variation we have then:

$$P \cong \frac{\mathrm{d}x}{\mathrm{d}t}F = Fv \quad (9.3.1020)$$

and therefore in the case of friction (in absolute value):

$$\boxed{P \cong |F_f|v = |-kv|v = kv^2} \quad (9.3.1021)$$

Thus, the power dissipated during movement is proportional to the square of speed in the case of Coulomb friction and proportional to the square of the speed in the case of viscous friction.

9.3.11.3 Vertical Viscous Friction

Given a non-deformable solid falling vertically in a uniform gravity field. We assume that air resistance force is proportional to the speed (viscous behaviour at low speeds):

$$F_f = -kv \quad (9.3.1022)$$

and using the fundamental principle of dynamics:

$$mg - kv = m\frac{\mathrm{d}v}{\mathrm{d}t} \quad (9.3.1023)$$

Or written differently (more traditional):

$$m\frac{\mathrm{d}v}{\mathrm{d}t} + kv = mg \quad (9.3.1024)$$

The solution of this linear differential equation of the first order is (see section Integral and Differential Calculus page 1609):

$$v = \frac{mg}{k} + Ce^{-\frac{k}{m}t} \quad (9.3.1025)$$

we can detail if required (on request).

By assuming that at time zero we had a given initial velocity it comes:

$$v(t=0)=v_0=\frac{mg}{k}+C\Rightarrow C=v_0-\frac{mg}{k} \tag{9.3.1026}$$

Therefore:

$$v=\frac{mg}{k}+\left(v_0-\frac{mg}{k}\right)e^{-\frac{k}{m}t}=\frac{mg}{k}\left(1-e^{-\frac{k}{m}t}\right)+v_0e^{-\frac{k}{m}t} \tag{9.3.1027}$$

Thus we see that when time tends to infinity (long enough in other words...) then the speed approaches:

$$v\underset{t\to+\infty}{=}\frac{mg}{k} \tag{9.3.1028}$$

therefore k can be determined experimentally!

It is a relation once again very often used in computer animations to represent objects that seem to slow down at constant speed naturally by falling.

The position at any time may be found by integrating the equation for v. With $v=\mathrm{d}y/\mathrm{d}t$:

$$\int\limits_{t_0}^{t_f}\mathrm{d}y=\int\limits_{t_0}^{t_f}\left[\frac{mg}{k}\left(1-e^{-\frac{k}{m}t}\right)+v_0e^{-\frac{k}{m}t}\right]\mathrm{d}t \tag{9.3.1029}$$

The calculations that follows are obvious (but we can detail on request if needed).

9.3.11.4 Stokes' Vertical Viscous Friction

This is typically the case of the parachutist performing a free fall. We proved in the section of Continuum Mechanics that Stokes' viscous force was given by (viscous behaviour at middle and high speeds):

$$F_f=-\rho v^2L^2C(R_e)=-\rho SCv^2=-kv^2 \tag{9.3.1030}$$

when speed is subsonic (modest in other words...).

The differential equation is the same as before in the case of the presence of a uniform gravitational field with the difference that the speed is this time squared:

$$mg-kv^2=m\frac{\mathrm{d}v}{\mathrm{d}t} \tag{9.3.1031}$$

That is:

$$\frac{k}{m}v^2+\frac{\mathrm{d}v}{\mathrm{d}t}-g=0 \tag{9.3.1032}$$

but we will not try to solve this differential equation. We want only determine the limit of speed and this speed limit is reached when it does not longer varies .... (yes obviously...). So at that moment:

$$\frac{\mathrm{d}v}{\mathrm{d}t}=0 \tag{9.3.1033}$$

and the differential equation becomes:

$$\frac{k}{m}v^2-g=0 \tag{9.3.1034}$$

Therefore:

$$\boxed{v_{\text{lim}}=\sqrt{\frac{gm}{k}}=\sqrt{\frac{gm}{\rho SC}}} \tag{9.3.1035}$$

Thus, it is possible to change the falling speed limit depending on the form factor of the falling object and its apparent exhibition surface and its mass (in the above case study we have neglected the Archimedes force that applies on the parachutist and that also slows down his fall).

✍**Example:**

Let us consider a sphere of radius R, of mass density ρ, released without initial speed a liquid of mass density ρ_{liq}, of viscosity μ (see the classification of viscosity in the norm ISO 3448). The sphere is subjected to its own weight, to a viscous frictional force and to Archimedes' force. We therefore have according to Newton's first law:

$$m\frac{\mathrm{d}v}{\mathrm{d}t}=\underbrace{\frac{4}{3}\pi R^3\rho g}_{\text{attraction force}}-\underbrace{6\pi\mu Rv}_{\text{Stokes' viscous force}}-\underbrace{\frac{4}{3}R^3\rho_{\text{liq}}g}_{\text{Archimedes' force}} \tag{9.3.1036}$$

where the last two terms (viscous Stokes's force and Archimedes force) have been proved in the section of Continuum Mechanics.

Rearranging, we have:

$$m\frac{\mathrm{d}v}{\mathrm{d}t}+6\pi\mu Rv=\frac{4}{3}\pi R^3g(\rho-\rho_{\text{liq}})=\frac{4}{3}\pi R^3g\rho\left(1-\frac{\rho_{\text{liq}}}{\rho}\right) \tag{9.3.1037}$$

or also:

$$\begin{aligned}&\frac{m}{\frac{4}{3}\pi R^3\rho}\frac{\mathrm{d}v}{\mathrm{d}t}+\frac{6\pi\mu R}{\frac{4}{3}\pi R^3\rho}v=\frac{m}{\frac{4}{3}\pi R^3}\frac{1}{\rho}\frac{\mathrm{d}v}{\mathrm{d}t}+\frac{18\mu}{4R^2\rho}v=\rho\frac{1}{\rho}\frac{\mathrm{d}v}{\mathrm{d}t}+\frac{9\mu}{2R^2\rho}v\\&=\frac{\mathrm{d}v}{\mathrm{d}t}+\frac{9\mu}{2R^2\rho}v=\left(1-\frac{\rho_{\text{liq}}}{\rho}\right)g\end{aligned} \tag{9.3.1038}$$

Therefore it remains:

$$\frac{\mathrm{d}v}{\mathrm{d}t}+\frac{9\mu}{2R^2\rho}v=\left(1-\frac{\rho_{\text{liq}}}{\rho}\right)g \tag{9.3.1039}$$

We put now:

$$\boxed{\tau=\frac{2R^2\rho}{9\mu}} \tag{9.3.1040}$$

which will be treated as a time constant. Then we have:

$$\tau\frac{\mathrm{d}v}{\mathrm{d}t}+v=\tau\left(1-\frac{\rho_{\text{liq}}}{\rho}\right)g \tag{9.3.1041}$$

But when the falling speed will become constant, we will have:

$$\frac{\mathrm{d}v}{\mathrm{d}t}=0 \tag{9.3.1042}$$

Which simplifies our previous relation and gives therefore immediately:

$$\boxed{v_{\text{lim}}=\tau\left(1-\frac{\rho_{\text{liq}}}{\rho}\right)g} \tag{9.3.1043}$$

Let us also solve the previous differential equation, beginning with that without the second member (see section Differential and Integral Calculus page 1609):

$$\frac{\mathrm{d}v}{\mathrm{d}t}+\frac{1}{\tau}v=0 \tag{9.3.1044}$$

We then saw in the section of Differential and Integral Calculus that the homogeneous solution was then given by:

$$v_h=Ce^{-\frac{1}{\tau}t} \tag{9.3.1045}$$

We can add the particular solution that is logically when t tends to infinity:

$$v_p=v_{\text{lim}}=\tau\left(1-\frac{\rho_{\text{liq}}}{\rho}\right)g \tag{9.3.1046}$$

Then we have:

$$v=v_p+v_h=v_{\text{lim}}+v_h=\tau\left(1-\frac{\rho_{\text{liq}}}{\rho}\right)g+Ce^{-\frac{1}{\tau}t} \tag{9.3.1047}$$

It remains to us to determine C which is obtained when t tends to 0 because we have then:

$$v(t=0)=v_0=0 \tag{9.3.1048}$$

Therefore:

$$v(t=0)=0=\tau\left(1-\frac{\rho_{\text{liq}}}{\rho}\right)g+Ce^{-\frac{1}{\tau}0}=\tau\left(1-\frac{\rho_{\text{liq}}}{\rho}\right)g+C\Rightarrow C=-\tau\left(1-\frac{\rho_{\text{liq}}}{\rho}\right)g \tag{9.3.1049}$$

Then we have:

$$v=\tau\left(1-\frac{\rho_{\text{liq}}}{\rho}\right)g-\tau\left(1-\frac{\rho_{\text{liq}}}{\rho}\right)ge^{-\frac{1}{\tau}t} \tag{9.3.1050}$$

Finally:

$$\boxed{v=\tau\left(1-\frac{\rho_{\text{liq}}}{\rho}\right)g\left(1-e^{-\frac{1}{\tau}t}\right)} \tag{9.3.1051}$$

9.3.11.5 Stokes' Horizontal Viscous Friction

This is a first approximation where are interested for example at a distance of stop with constantly braking of a mobile without taking into account the coefficient of friction with the ground but only with the ambient air (subsonic speed always...).

We then have according to Newton's first law:

$$m\frac{\mathrm{d}v}{\mathrm{d}v}=-kv^2 \tag{9.3.1052}$$

Suppose we want to know how much time T a mobile which had an initial velocity v_0 will have decelerated to a given final speed v_f.

Then we have:

$$\mathrm{d}t=-\frac{m}{k}\frac{\mathrm{d}v}{v^2} \tag{9.3.1053}$$

Therefore:

$$T=\int\limits_0^T \mathrm{d}t=-\frac{m}{k}\int\limits_{v_0}^{v_f}\frac{1}{v^2}\mathrm{d}v=-\frac{m}{k}\left(-\frac{1}{v}\bigg|_{v_0}^{v_f}\right)=-\frac{m}{k}\left[-\frac{1}{v_f}-\left(-\frac{1}{v_0}\right)\right]=\frac{m}{k}\left(\frac{1}{v_f}-\frac{1}{v_0}\right) \tag{9.3.1054}$$

where we observe already a first problem with this model is that when stopped, the final velocity being equal to zero, it will take an infinite time to get there... but go on, we shall return to this finding.

The evolution law of the speed is determined analogously as:

$$t=-\frac{m}{k}\left[-\frac{1}{v_f}-\left(-\frac{1}{v_0}\right)\right]=\frac{m}{k}\left(\frac{1}{v_f}-\frac{1}{v_0}\right) \tag{9.3.1055}$$

Therefore:

$$t=\frac{m}{k}\frac{1}{v_f}-\frac{m}{k}\frac{1}{v_0}\Rightarrow\frac{m}{k}\frac{1}{v_f}=\frac{m}{k}\frac{1}{v_0}+t\Rightarrow\frac{1}{v_f}=\frac{1}{v_0}+\frac{k}{m}t=\frac{1}{v_0}\left(1+\frac{v_0k}{m}t\right) \tag{9.3.1056}$$

Let us put the time constant:

$$\tau=\frac{m}{v_0k} \tag{9.3.1057}$$

Therefore:

$$\frac{1}{v_f}=\frac{1}{v_0}\left(1+\frac{t}{\tau}\right) \tag{9.3.1058}$$

hence:

$$\boxed{v_f(t)=\frac{v_0}{1+\dfrac{t}{\tau}}} \tag{9.3.1059}$$

The distance travelled at the instant t, leaving the mobile slow down uniquely by the frictional forces is therefore:

$$x=\int_0^t v_f(t)\mathrm{d}t=\int_0^t \frac{v_0}{1+\dfrac{t}{\tau}}\mathrm{d}t=v_0\int_0^t \frac{1}{1+\dfrac{t}{\tau}}\mathrm{d}t=v_0\tau\ln\left(1+\frac{t}{\tau}\right) \tag{9.3.1060}$$

That is:

$$\boxed{x=\frac{m}{k}\ln\left(1+\frac{t}{\tau}\right)} \tag{9.3.1061}$$

The result is pretty but we realize that it's not correct because after an infinite time, the car will have travelled an infinite distance which is clearly unrealistic. This comes from the fact that the model is too simplistic, therefore let us improve it.

The ideal, objectively speaking, would be to take into account the viscous friction tire/ground plus the air friction. We would then have:

$$m\frac{\mathrm{d}v}{\mathrm{d}t}=-kv-k'v^2 \tag{9.3.1062}$$

but the problem with this differential equation is that it will lead us to a singularity if we continue the calculations. It is therefore not usage...

We then try with the following form:

$$m\frac{\mathrm{d}v}{\mathrm{d}t}=-m\alpha-m\beta v^2 \tag{9.3.1063}$$

which expresses that there is a friction force proportional to the weight of the vehicle which is simply the form of the second Coulomb's law:

$$F_{T,\text{lim}}=\mu_D P=\mu_D mg=m(g\mu_D)=m\alpha \tag{9.3.1064}$$

and the second term being the viscous Stokes' friction from which we have take out the mass factor included in the density being implicitly in the constant k of kv^2.

We then have simplifying the mass terms:

$$\frac{\mathrm{d}v}{\mathrm{d}t}=-\alpha-\beta v^2 \tag{9.3.1065}$$

So we already see that in this model we will lose the effect of the mass that normally has for effect to extend the travel distance before stop (in reality!) because of inertia. But let us continue anyway...

So we have to integrate:

$$-\int_0^t \mathrm{d}t=\int_{v_0}^0 \frac{\mathrm{d}v}{\alpha+\beta v^2}=\frac{1}{\beta}\int_{v_0}^0 \frac{\mathrm{d}v}{\dfrac{\alpha}{\beta}+v^2} \tag{9.3.1066}$$

We have proved in the section of Differential and Integral Calculus that:

$$\int \frac{1}{a^2+x^2} = \frac{1}{a}\arctan\left(\frac{v_0}{a}\right) \tag{9.3.1067}$$

Slightly rearranged it gives:

$$\boxed{t = \frac{1}{\sqrt{\alpha\beta}}\arctan\left(\sqrt{\frac{\beta}{\alpha}}v_0\right)} \tag{9.3.1068}$$

So now we fall on a finite time for stop... which is more reassuring result.

Let us now seek for stop distance. We use the fact that:

$$v = \frac{\mathrm{d}x}{\mathrm{d}t} \Rightarrow \mathrm{d}t = \frac{\mathrm{d}x}{v} \tag{9.3.1069}$$

gives us the possibility to write:

$$\frac{\mathrm{d}v}{\mathrm{d}t} = \frac{\mathrm{d}v}{\mathrm{d}x/v} = \frac{v\mathrm{d}v}{v\mathrm{d}x} = -\alpha - \beta v^2 \tag{9.3.1070}$$

Therefore:

$$-\mathrm{d}x = \frac{v\mathrm{d}v}{\alpha+\beta v^2} \tag{9.3.1071}$$

Which brings us to:

$$-\int_0^L \mathrm{d}x = \int_{v_0}^0 \frac{v\mathrm{d}v}{\alpha+\beta v^2} \tag{9.3.1072}$$

Which gives us first:

$$-L = \int_{v_0}^0 \frac{v\mathrm{d}v}{\alpha+\beta v^2} = \frac{1}{2\beta}\int_{v_0}^0 \frac{2\beta\mathrm{d}v}{\alpha+\beta v^2} = \frac{1}{2\beta}\ln(\alpha+\beta v^2)\bigg|_{v_0}^0 = \frac{1}{2\beta}\left[\ln(\alpha) - \ln(\alpha+\beta v_0^2)\right] \tag{9.3.1073}$$

So we have our final result. Better than the last result but however independent of the mass ... but in the meantime it's better than nothing...

Remark

With a dry braking the front axle of a small car (Mercedes Benz Class A 190 Elegance) has a braking force typically of the order of 2.8 [kN].

9.3.11.6 Friction's Heat Factor

The sliding friction intervening between two bodies in contact results in a loss of mechanical energy almost entirely converted into heat energy Q (see section Thermodynamics page 3599). The problem of conversion of friction energy into heat is always very complex and requires detailed studies but is of major importance in the industry.

For simplicity, let assume that the friction component remains constant, applied to the center of gravity of the friction surface, and the speed of the body is linear and constant. The work done by the friction force, for a finite displacement is then:

$$W_f = F_f \Delta L \tag{9.3.1074}$$

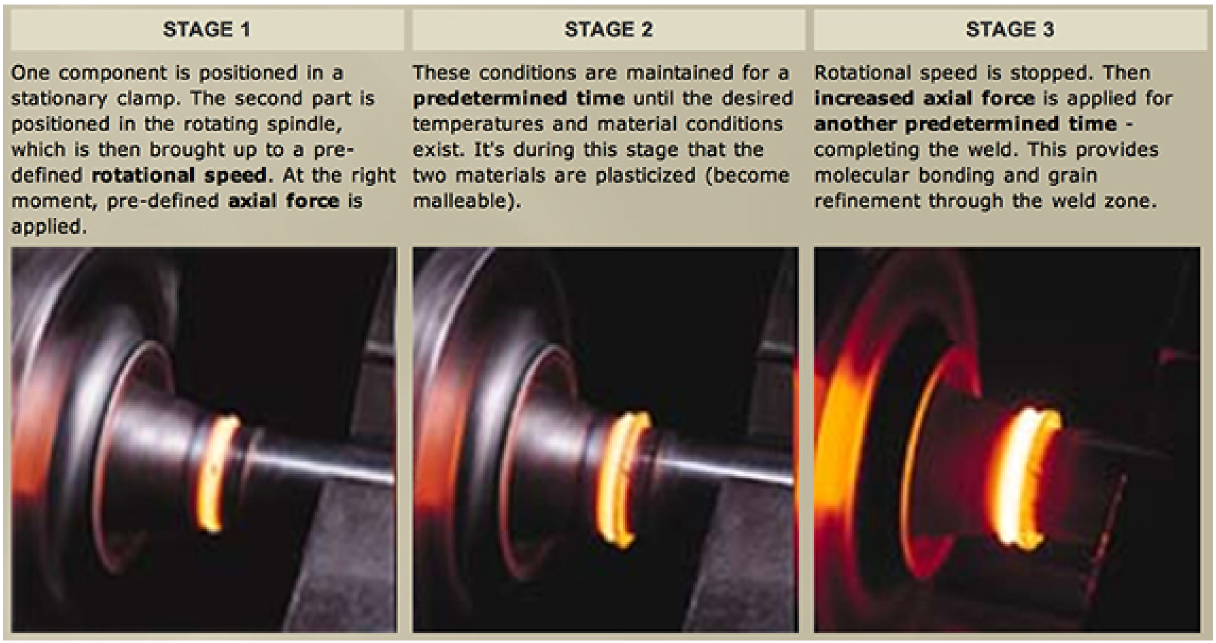
Assuming this mechanical work completely transformed into thermal energy, the quantity of heat produced must be evacuated out of the friction surfaces. This is a fundamental problem in the mechanisms with high heat release such as brakes, clutches, gear transmissions and worm gear and especially plain bearings. The released heat output is then calculated by:

$$\dot{Q} = \frac{F_f \Delta L}{\Delta t} = F_f \frac{\Delta L}{\Delta t} = F_f v \tag{9.3.1075}$$

We just have to divide by the total contact area to the heating power per unit area. Then, in the special case of Coulomb's dynamic friction, we have this last relation that will be written:

$$\boxed{\dot{Q} = F_f v = \mu_D P A v} \tag{9.3.1076}$$

The heat output is proportional to the friction coefficient μ_D and the product of the mean pressure P by the sliding velocity v and the contact area A. This control method of heating is obviously a rough first approach because the coefficient of friction, the pressure and the sliding speed are only part of the factors influencing the temperature of operation of a mechanism.



STAGE 1	STAGE 2	STAGE 3
One component is positioned in a stationary clamp. The second part is positioned in the rotating spindle, which is then brought up to a pre-defined **rotational speed.** At the right moment, pre-defined **axial force** is applied.	These conditions are maintained for a **predetermined time** until the desired temperatures and material conditions exist. It's during this stage that the two materials are plasticized (become malleable).	Rotational speed is stopped. Then **increased axial force** is applied for **another predetermined time** - completing the weld. This provides molecular bonding and grain refinement through the weld zone.

Figure 9.112: **Rotary friction welding principles**

~80 %

★★★★☆

69 votes, 73.33%

9.4. Wave Mechanics

HERE we focus on the study of mathematical properties of vibrating strings that we can also by extension and in a wish of generality assimilate to the immaterial cases of the important concept of "waves". This study is very important because we need some of the results that will be proven here in the section of Thermodynamics, Quantum physics, Astrophysics, Electrodynamics, Acoustics and String theory (to mention only the most important one).

Definitions (#9.4.452):

D1. A "wave" is a transport of energy without transporting material. It is embodied in the propagation of a disturbance of a medium, hence the name "travelling wave". The speed with which the wave progresses depends obviously on the physical properties of the medium.

D2. In the case where the disturbance of the medium (deformation of the wave) is perpendicular to the wave propagation direction, we speak about "transverse wave" or "transverse disturbance" (typical of waves in string for example).

D3. In the case where the disturbance of the medium is parallel to the wave propagation direction, we speak about "longitudinal wave" or "longitudinal disturbance " (typical of springs).

9.4.1 Wave Function

Given a disturbance $y=f(x)$ in a given region of space. If we replace x by $x-b$ we define, in the same region, a disturbance $f(x-b)$ identical to $f(x)$ but translated a distance b in the direction of positive X (to the right if we adopt the conventional representation seen in the section of Functional Analysis).

If t represent a time and if we put $b=vt$, then v can denote the translational speed of the disturbance.

So we name "wave function", the mathematical relation:

$$\boxed{f(x\pm vt)} \tag{9.4.1}$$

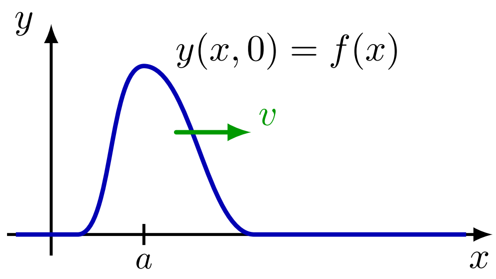
that describes the progress of a disturbance $y(x,t)$ in the space:

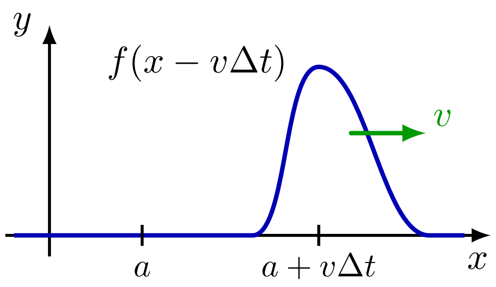
- $y(x,t)=f(x-vt)$ describing a wave that progresses in the direction of $+X$
- $y(x,t)=f(x+vt)$ describing a wave that progresses in the direction of $-X$

v is by definition named "wave phase velocity" and vt the "phase shift". It is constant in a homogeneous medium. The "amplitude of the wave" is the maximum value of the perturbation:

$$y_{\max}=\hat{y} \tag{9.4.2}$$

In the absence of damping, it will retain the same value in each of the points x where the wave passes.





9.4.2 Wave Equation

Theorem 9.9. Without getting into too technical considerations, we will say that every function f whose argument is $x\pm vt$ has the property:

Proof 9.9.1.

$$\begin{gathered}\frac{\partial}{\partial x}f(x-vt)=1\cdot f'(x-vt)\Rightarrow f'(x-vt)=\frac{\partial}{\partial x}f(x-vt)\\ \frac{\partial}{\partial t}f(x-vt)=-v\cdot f'(x-vt)\Rightarrow f'(x-vt)=-\frac{1}{v}\frac{\partial}{\partial t}f(x-vt)\end{gathered} \tag{9.4.3}$$

and therefore the equality follows immediately (this is simply the application of composed derivative as proved in the section of Differential and Integral Calculus).

■ Q.E.D.

Following a request from a reader who did not find this proof very clear, let us make an example. Consider the following function:

$$f(x-vt)=(x-vt)^2 \tag{9.4.4}$$

and therefore:

$$\begin{gathered}\frac{\partial}{\partial x}f(x-vt)=\frac{\partial}{\partial x}(x-vt)^2=\underbrace{1}_{=\frac{\partial}{\partial x}(x-vt)}\cdot\underbrace{2(x-vt)}_{=f'}=2x-vt\\ \frac{\partial}{\partial t}f(x-vt)=\frac{\partial}{\partial t}(x-vt)^2=\underbrace{-v}_{=\frac{\partial}{\partial t}(x-vt)}\cdot\underbrace{2(x-vt)}_{=f'}=-2xv+v^2t\end{gathered} \tag{9.4.5}$$

Returning to the general case, we derive a second time, we obtain another form of the wave equation that we also meet frequently in practice and in the books:

$$\begin{gathered}\frac{\partial^2}{\partial x^2}f(x-vt)=1\cdot\frac{\partial}{\partial x}f'(x-vt)=1\cdot 1\cdot f''(x-vt)\Rightarrow f''(x-vt)=\frac{\partial x^2}{\partial x^2}f(x-vt)\\ \frac{\partial^2}{\partial t^2}f(x-vt)=-v\cdot\frac{\partial}{\partial t}f'(x-vt)=v^2 f''(x-vt)\Rightarrow f''(x-vt)=\frac{1}{v^2}\frac{\partial^2}{\partial t^2}f(x-vt)\end{gathered} \tag{9.4.6}$$

Which brings us to write one of the most important relations in physics named the "wave equation", "propagation equation" or "d'Alembert's equation" and that we see again many times in other sections of this textbook (Electrodynamics, Wave Quantum Physics, General Relativity, Acoustic):

$$\boxed{\frac{\partial x^2}{\partial x^2}f(x\pm vt)=\pm\frac{1}{v^2}\frac{\partial^2}{\partial t^2}f(x\pm vt)} \tag{9.4.7}$$

Remarks

R1. Don't forget (see section Differential and Integral Calculus page 1605) that the sum of the solutions to a differential equation is also a solution of the differential equation!

R2. When two or more waves propagate in a medium, the wave function that results is the algebraic sum of each wave of wave functions. We say then that the waves "interfere" and we name the phenomenon "superposition wave principle" (see further below for mathematical details).

Consider now a string of length L attached by one of its ends to a fixed terminal. Now suppose that a disturbance propagate into the string. When the disturbance reaches the end (terminal), we observe that it changes in sign and that the speed is reversed: the wave is then subject to a reflection with inversion.

To describe the phenomenon, we must impose at the terminal:

$$y(L,t)=0 \qquad \forall t \tag{9.4.8}$$

But any wave function of the form $f(x\pm vt)$, which progresses to the termination, can not verify the above condition:

$$f(x-vt)_{x=L}=f(L-vt)=0 \tag{9.4.9}$$

for all time values t.

The trick is to replace it with another wave function $y(x,t)$ whose shape is similar to f at large distance from the source of the interference, and which is zero at the point of termination for all time values t. For this, we can imagine at the point of termination, a mirror which gives an image of the string of the same length in which we invent a virtual wave:

$$f(2L-(x\pm vt))=f((2L-x)-vt) \tag{9.4.10}$$

symmetrical to $f(x-vt)$, but of opposite sign.

We decide therefore:

- That the two waves progress toward each other to cancel at the endpoint
- Any part of the real wave that pass through the endpoint becomes virtual
- Any part of the virtual wave which enters the string becomes real

At their intersection, the two waves achieve a destructive interference at the endpoint. The algebraic sum of these two wave functions is also a wave function:

$$\boxed{y(x,t)=f(x-vt)-f(2L-(x+vt))} \tag{9.4.11}$$

which has the desired property at $x=L$ for any t:

$$y(L,t)=f(L-vt)-f(2L-(L+vt))=0 \tag{9.4.12}$$

If we now consider a free termination, without friction with its fixed point, we find ourselves in a similar case to the previous one except that the interference is constructive at the endpoint rather than destructive so that the wave function will be written:

$$y(x,t)=f(x-vt)+f(2L-(x+vt)) \tag{9.4.13}$$

Remarks

R1. When the wave arrives on a free or fixed termination, the energy carried is fully returned back.

R2. When the termination is not exactly adapted, only a portion of the energy is absorbed by it, the rest is reflected.

9.4.3 Type of Waves

In theoretical physics (and in practice), we often restrict the study of certain phenomena to particular cases of waves. Primarily, we recognize three of them that we will briefly but carefully develop here:

9.4.3.1 Periodic Waves

If an event produces a wave does that carry only single disturbance produced at a given point, there are many disturbances which are capable to excite a medium in a repeatedly (periodic) way.

The spatial point of the source is then subjected to the same periodic disturbance. The duration of a complete cycle is made identically to our study of pendulum, the "period T".

If the disturbance can propagate as a wave, at speed v, it is described by the wave function that we know:

$$y=f(x,t)=f(x-vt) \tag{9.4.14}$$

At each point of the disturbed system, the periodic wave imposes a "temporal periodicity" of the disturbance which requires us to write:

$$y(x=c^{te})=f(x,t+n_1T)=f(x,t) \qquad \forall x,n_1\in\mathbb{Z} \tag{9.4.15}$$

After several cycles (periods) of disturbances from the source, more disturbances are distributed in space. The distance between two successive disturbances is named the "wavelength λ".

The "spatial periodicity" also requires that:

$$y(t=c^{te})=f(x+n_2\lambda,t)=f(x,t) \qquad \forall t,n_2\in\mathbb{Z} \tag{9.4.16}$$

Therefore, in the case of a periodic wave, λ is the path travelled by the wave during the time T:

$$\boxed{\lambda=vT\Leftrightarrow v=\frac{\lambda}{T}} \tag{9.4.17}$$

Theorem 9.10. If a wave function is periodic in time, it is also in the space, provided that the pulse does not deform as it progresses.

Proof 9.10.1.

$$\begin{aligned}f(x,t)=f(x,t+n_1T)&=f(x-v(t+n_1T))=f(x-vt-n_1vT)=f(x-n_1\lambda-vt)\\&=f(x-n_1\lambda,t)\end{aligned} \tag{9.4.18}$$

By putting $n_2=-n_1=n$ we have:

$$f(x,t+n_1T)=f(x-n_1\lambda,t) \tag{9.4.19}$$

that becomes:

$$f(x,t+nT)=f(x+n\lambda,t) \tag{9.4.20}$$

■ Q.E.D.

9.4.3.2 Harmonic Waves

For these waves, the wave function solution of the d'Alembert's equation is a trigonometric function of the sine or cosine type (or sum of them):

$$y(x,t)=f(x-vt)=\hat{y}\sin(k(x-vt)) \quad \text{or} \quad y(x,t)=f(x-vt)=\hat{y}\cos(k(x-vt)) \tag{9.4.21}$$

The presence of k that represents the number of waves or cycles per unit distance, named "wave number" (sometimes also written "wavenumber" in some textbooks) or "propagation number", is required for two reasons:

- k is expressed in $[\text{m}^{-1}]$
- the value of k must ensure the periodicity of the wave function:

1. Angular periodicity of the mathematical function:

$$\sin(k(x-vt)\pm 2\pi)=\sin(k(x-vt)) \tag{9.4.22}$$

2. Spatial periodicity of the wave function:

$$\sin(k(x\pm\lambda)-vt)=\sin(k(x-vt)) \tag{9.4.23}$$

By equating these two expressions:

$$k(x-vt)\pm 2\pi=k((x\pm\lambda)-vt)=k(x-vt)\pm k\lambda \tag{9.4.24}$$

which implies:

$$\boxed{k=\frac{2\pi}{\lambda}} \tag{9.4.25}$$

Let us introduce:

$$\lambda=vT \tag{9.4.26}$$

in the expression of k:

$$\boxed{kv=\frac{2\pi}{T}=2\pi f=\omega} \tag{9.4.27}$$

hence another important relation, the explicit relation of the "phase velocity":

$$\boxed{v=\frac{\omega}{k}} \tag{9.4.28}$$

$$\begin{tikzcd}
 & \underset{\text{linear wavenumber}}{\tilde{\nu}} \arrow[rr, "x\cdot\tau"] \arrow[dl, "x\cdot v_p"'] \arrow[dd, dotted, leftrightarrow, "\frac{1}{x}"] & & \underset{\text{angular wavenumber}}{k} \arrow[dl, "x\cdot v_p"] \arrow[dd, leftrightarrow, "\frac{1}{x}"] \\
\underset{\text{linear frequency}}{f} \arrow[rr, "x\cdot\tau"] \arrow[dd, leftrightarrow, "\frac{1}{x}"'] & & \underset{\text{angular frequency}}{\omega} \arrow[dd, leftrightarrow, "\frac{1}{x}"] & \\
 & \underset{\text{linear wavelength}}{\lambda} \arrow[rr, dotted, "x/\tau"] \arrow[dl, dotted, "x/v_p"] & & \underset{\text{angular wavelength}}{\lambdabar} \arrow[dl, "x/v_p"] \\
\underset{\text{linear period}}{T} \arrow[rr, "x/\tau"] & & \underset{\text{angular period}}{?} &
\end{tikzcd}$$

v_p = wave speed (phase velocity)
$\tau = 2\pi$

Figure 9.113: **Commutative diagram of harmonic wave properties (source: Wikipedia, author: Waldir)**

The wave function of the harmonic wave can thus be written as:

$$\begin{aligned} y(x,t)&=\hat{y}\sin(kx-\omega t) \quad \text{or} \quad y(x,t)=\hat{y}\sin\left[2\pi\left(\frac{x}{\lambda}-\frac{t}{T}\right)\right] \qquad \text{progressive along } +X \\ y(x,t)&=\hat{y}\sin(kx+\omega t) \quad \text{or} \quad y(x,t)=\hat{y}\sin\left[2\pi\left(\frac{x}{\lambda}+\frac{t}{T}\right)\right] \qquad \text{progressive along } -X \end{aligned} \tag{9.4.29}$$

9.4.3.2.1 Phase velocity and Group velocity

According to what we have just seen above, the phase velocity of a monochromatic wave is equal to the ratio between its pulsation and its wave number (norm of the wave vector).

Let us consider now the simple case consisting of a wave of the superposition of two different pulsations and unit amplitude:

$$\phi(x,t)=\cos(\omega_1 t-k_1 x)+\cos(\omega_2 t-k_x) \tag{9.4.30}$$

Using trigonometric identities (Simpson's formula) we have:

$$\phi(x,t)=2\cos\left(\frac{\omega_2+\omega_1}{2}t-\frac{k_2+k_1}{2}x\right)\cos\left(\frac{\omega_2-\omega_1}{2}t-\frac{k_2-k_1}{2}x\right) \tag{9.4.31}$$

Thus, the wave in question consists of the product of two terms:

- The first term is a monochromatic wave whose velocity phase is:

  $$v_\phi=\frac{\omega_2+\omega_1}{k_2+k_1} \tag{9.4.32}$$

- The second therm is another monochromatic wave whose velocity phase is:

  $$v_\phi=\frac{\omega_2-\omega_1}{k_2-k_1}=\frac{\Delta\omega}{\Delta k} \tag{9.4.33}$$

  This term intervenes as amplitude modulator of the first term.

There is thus a beat phenomenon (i.e. "beating") where a sine wave is modulated in amplitude modulated by a lower pulse sinusoid. The group velocity is then by definition:

$$v_g=\frac{\Delta\omega}{\Delta k} \tag{9.4.34}$$

Finally:

$$\boxed{v_\phi=\frac{\omega}{k} \qquad v_g=\frac{\Delta\omega}{\Delta k}} \tag{9.4.35}$$

Here is an example of a large beating:

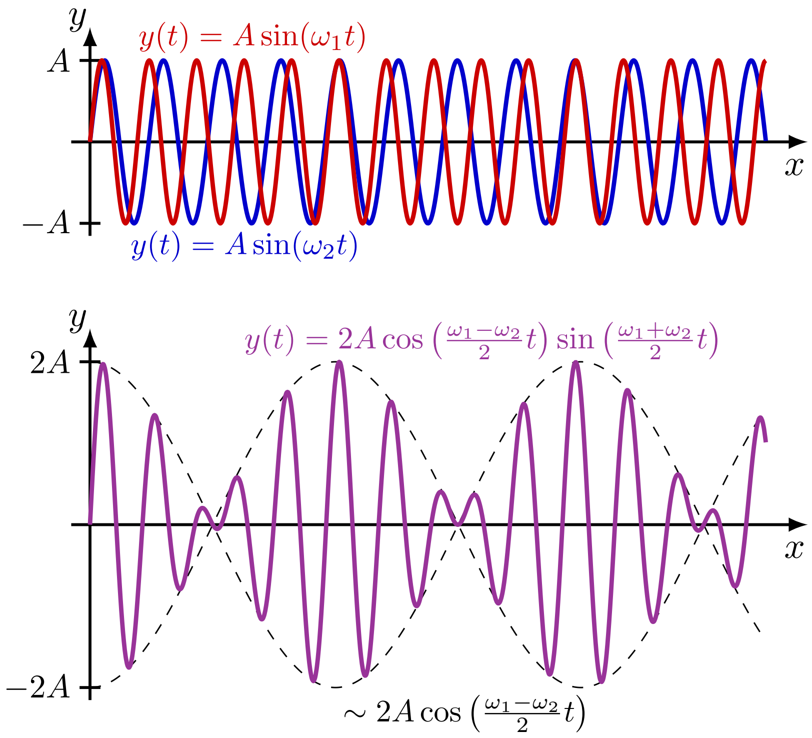


Figure 9.114: Beating with large difference in wavelength (author: Izaak Neutelings)

and with a small beating:

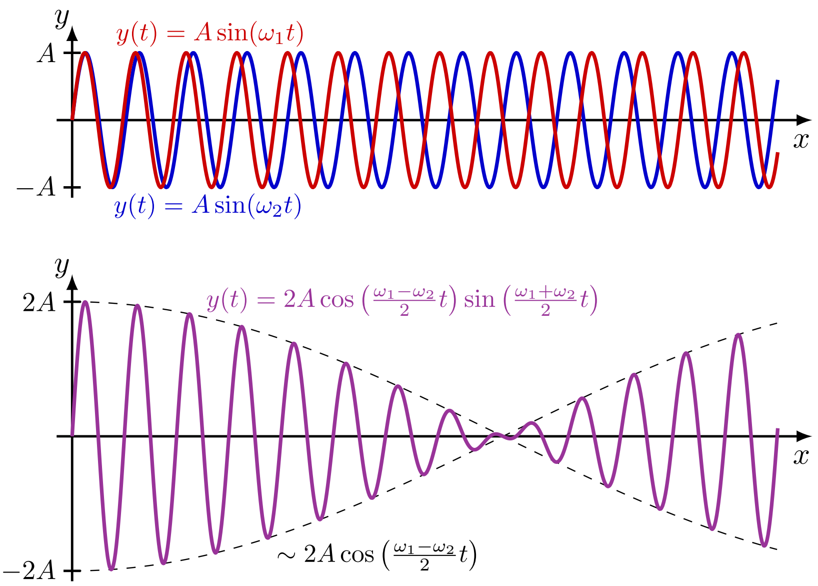


Figure 9.115: Beating small large difference in wavelength (author: Izaak Neutelings)

We can illustrate the both velocities with an animation in Maple™ 4.00b:

```
>restart:with(plots):
>lambda[0]:=1; T[0]:=1; k[0]:=2*Pi/lambda[0]; w[0]:=2*Pi/T[0];
>delta_k:=k[0]/8:  k[1]:=k[0]-delta_k; k[2]:=k[0]+delta_k;
```

```
delta_w:=w[0]/10:  w[1]:=w[0]-delta_w; w[2]:=w[0]+delta_w;
>P1:=animate(cos(k[1]*x-w[1]*t)+cos(k[2]*x-w[2]*t), x=0..1*2*Pi/delta_k,
t=0..2*Pi/delta_w, numpoints=200, frames=15, color=red):
>P2:=animate(2*cos(-1/2*k[1]*x+1/2*w[1]*t+1/2*k[2]*x-1/2*w[2]*t),
-2*cos(-1/2*k[1]*x+1/2*w[1]*t+1/2*k[2]*x-1/2*w[2]*t),
x=0..1*2*Pi/delta_k, t=0..2*Pi/delta_w, numpoints=100, frames=15,
color=blue):
>display(P1,P2);
```

Which gives visually:

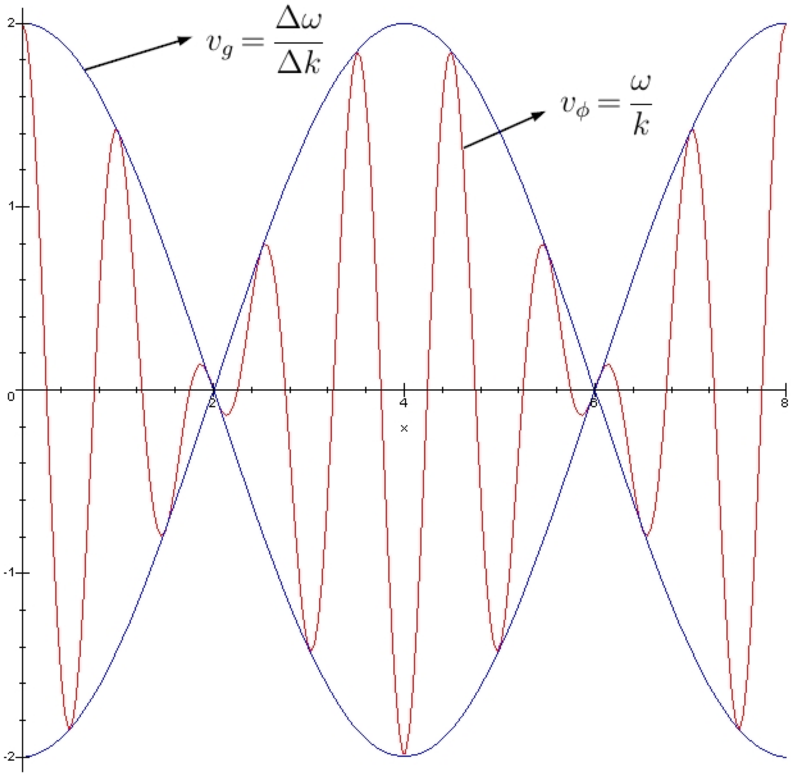


Figure 9.116: **Group and Phase velocities difference**

9.4.3.3 Stationary Waves

Let us imagine a string excited in a harmonic way. Rather than adapting its end to extract the energy of the string, let us assume that this end is fixed. The wave is therefore reflected!!!

A new wave function must be define to describe the superposition of the incoming wave:

$$y_1(x,t)=\hat{y}\sin(kx-\omega t) \tag{9.4.36}$$

and of the reflected wave (symmetric and of opposite sign):

$$y_2(x,t)=\hat{y}\sin(k(2L-x)-\omega t) \tag{9.4.37}$$

by analogy with the result that w have found during our study of the ends of strings:

$$y(x,t)=f(k(2L-x)-vt) \tag{9.4.38}$$

The following trigonometric identity proved in the section of Trigonometry:

$$\sin(p)-\sin(q)=2\sin\left(\frac{p-q}{2}\right)\cos\left(\frac{p+q}{2}\right) \tag{9.4.39}$$

gives us:

$$\boxed{y(x,t)=y_1(x,t)+y_2(x,t)=2\hat{y}\sin(k(x-L))\cos(kL-\omega t)} \tag{9.4.40}$$

This is not a progressive wave as x and t does combine anymore like $(x-vt)$. Some points of the string never move. They satisfy:

$$\sin(k(x-L))=0\Rightarrow k(x-L)=\pm n\pi\Rightarrow\frac{2\pi}{\lambda}(x-L)=\pm n\pi \tag{9.4.41}$$

and are located therefore on:

$$\boxed{x=L\pm n\frac{\lambda}{2}} \tag{9.4.42}$$

For obvious reasons, we will keep only the values of n for which $x<L$.

Each of these points is named a "vibration node".

We observe in such a system, vibration zones, of length $\lambda/2$, in which the string vibrates transversely in an height region of $4\hat{y}$ (two up, two down).

Remark

The points at which the amplitude of vibration is maximum are named sometimes the "antinodes of the vibration".

Since $\sin(k(x-L))=1$, the antinodes are apart of $\lambda/2$ and located at $\lambda/4$ of the nodes.

If we now impose a fixed termination on both ends of a vibrating string, we end up with a configuration named "resonance system".

Most often, we do not observe much until we find the excitation frequency vibrations that places nodes on the two fixed endpoints.

Therefore for a string of length L:

$$x=0=L\pm n\frac{\lambda}{2} \tag{9.4.43}$$

involves:

$$L=n\frac{\lambda}{2}\Leftrightarrow\lambda=\frac{2L}{n} \tag{9.4.44}$$

The string is then the seat of a standing wave whose vibration amplitude is considerably greater than the excitation amplitude (four times)!

We then say that the string came into "resonance" with the exciter.

The relation:

$$\boxed{\lambda=\frac{2L}{n}} \qquad (9.4.45)$$

shows that there are several possible wavelengths whose largest corresponding to $n=1$ and is named the "fundamental wavelength" or "fundamental harmonic" and is obviously equal to $\lambda=2L$:

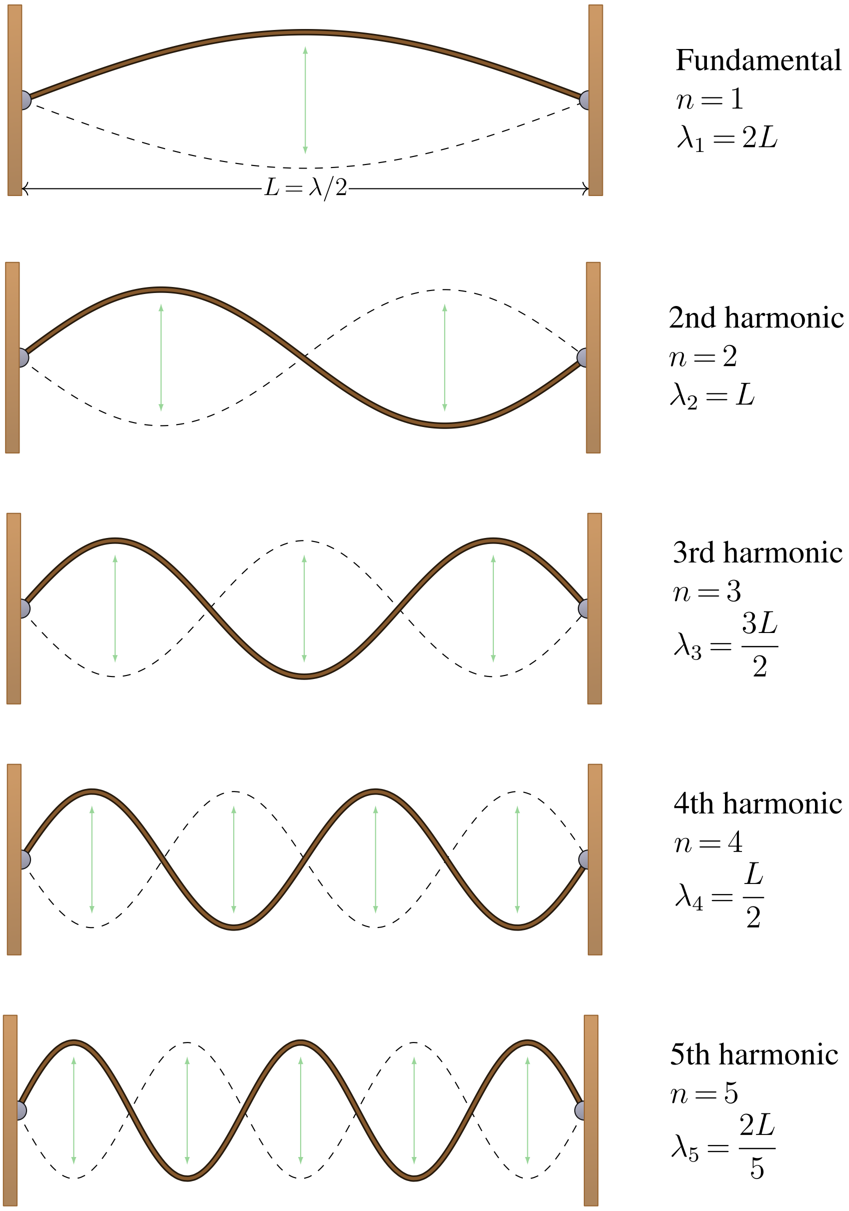

$L=\lambda/2$
Fundamental
$n=1$
$\lambda_1=2L$
2nd harmonic
$n=2$
$\lambda_2=L$
3rd harmonic
$n=3$
$\lambda_3=\frac{3L}{2}$
4th harmonic
$n=4$
$\lambda_4=\frac{L}{2}$
5th harmonic
$n=5$
$\lambda_5=\frac{2L}{5}$

Figure 9.117: **Double fixed standing wave (author: Izaak Neutelings)**

9.4.3.4 Vibration Modes in a Stretched String

We have seen how a wave can move in a string. Let us show now why this is possible and establish the relation $y(x,t)$, giving the shape of the string over time.

Given a wire of diameter $\varnothing$, of length L and mass m, the linear density of the string (assumed constant along it) is then:

$$\rho_l = \frac{m}{L} = \frac{\Delta m}{\Delta L} = \frac{\Delta m}{\Delta x} \qquad (9.4.46)$$

By a small shock, we generate a small transversal perturbation (so we do deform the string and maintain its linear density constant). Let us isolate, in the disturbed area, a string element of length ΔL.

The approximations are the following:

1. Each element of the string may be cut infinitesimally so as to be almost parallel to the x-axis. The angles θ_P, θ_Q are considered small.
2. The string is considered as deformable but non-extensible therefore the norm of the forces in the string is constant at any point regardless of the deformation.

For the rest of the reasoning, we use the figure below:

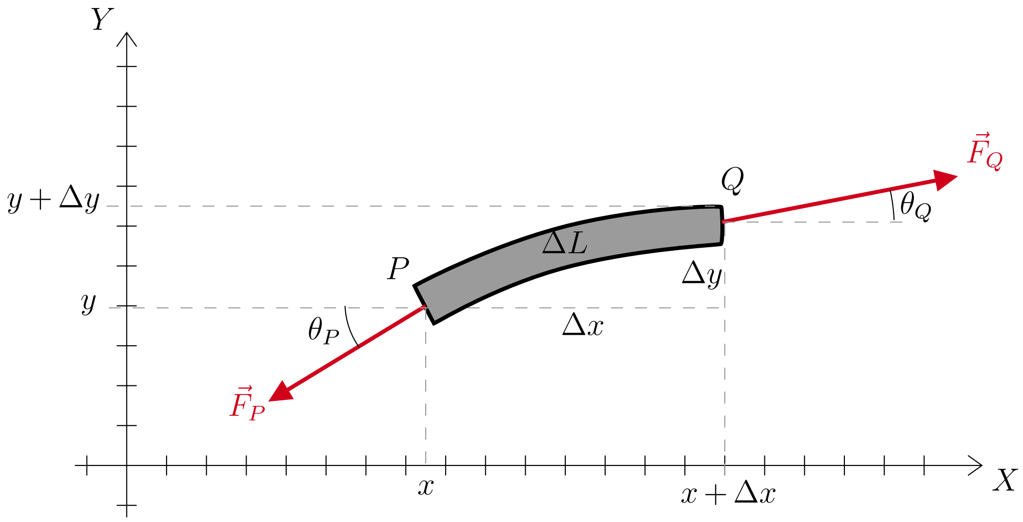


Figure 9.118: **Stretched string scheme**

If the angles are small, the balance of forces gives:

$$\sum F_x = F\cos(\theta_Q) - (F\cos(\theta_P)) - F(\cos(\theta_Q) - \cos(\theta_P)) \cong 0 \qquad (9.4.47)$$

which means that there is no displacement following the x-axis:

$$\sum F_y = F\sin(\theta_Q) - (F\sin(\theta_P)) = F(\sin(\theta_Q) - \sin(\theta_P)) \qquad (9.4.48)$$

If the angles are really small, we have the first term of the Taylor development of $\sin(x)$ and $\tan(x)$ (see section Sequences and Series page 1703) which gives:

$$\sin(x) \cong x \cong \tan(x) \tag{9.4.49}$$

Therefore:

$$\sum F_y = F(\sin(\theta_Q) - \sin(\theta_P)) \cong F(\tan(\theta_Q) - \tan(\theta_P)) \tag{9.4.50}$$

is the acceleration following the y-axis for small angles.

Newton's law applied to the mass $\Delta m = \rho_l \Delta x$ gives (we consider that each point of the mass moves only according to the y-axis since there is no elongation and as we are in the assumption of a transverse excitation):

$$\sum F_y = \Delta m a_y = \rho_l \Delta x \frac{\partial^2 y}{\partial t^2} \tag{9.4.51}$$

The tangents are given by the partial derivatives of the function $y(x)$:

$$\sum F_y = F(\tan(\theta_Q) - \tan(\theta_P)) = F\left(\left.\frac{\partial y}{\partial x}\right|_Q - \left.\frac{\partial y}{\partial x}\right|_P\right) \tag{9.4.52}$$

Which is equalized with the prior-previous relation:

$$\rho_l \Delta x \frac{\partial^2 y}{\partial t^2} = F\left(\left.\frac{\partial y}{\partial x}\right|_Q - \left.\frac{\partial y}{\partial x}\right|_P\right) \tag{9.4.53}$$

and therefore:

$$\frac{\partial^2 y}{\partial t^2} = \frac{F}{\rho_l} \frac{\left(\left.\frac{\partial y}{\partial x}\right|_Q - \left.\frac{\partial y}{\partial x}\right|_P\right)}{\Delta x} \tag{9.4.54}$$

If $\Delta x \to 0$, the two tangents tend to the same value, but the fraction of the right-hand side tends to a finite value:

$$\lim_{\Delta x \to 0} \frac{\left(\left.\frac{\partial y}{\partial x}\right|_Q - \left.\frac{\partial y}{\partial x}\right|_P\right)}{\Delta x} = \frac{\partial y^2}{\partial x^2} \tag{9.4.55}$$

It follows the differential equation:

$$\lim_{\Delta x \to 0} \frac{\left(\left.\frac{\partial y}{\partial x}\right|_Q - \left.\frac{\partial y}{\partial x}\right|_P\right)}{\Delta x} = \frac{\partial y^2}{\partial x^2} \tag{9.4.56}$$

$$\frac{\partial^2 y}{\partial t^2} = \frac{F}{\rho_l} \frac{\partial^2 y}{\partial x^2} \tag{9.4.57}$$

The latter relation is usually written as follows:

$$\boxed{\frac{\partial^2 y}{\partial x^2} - \frac{\rho_l}{F} \frac{\partial^2 y}{\partial t} = 0} \tag{9.4.58}$$

and is named "equation of vibrating strings".

Remark

In some books, the linear density is denoted by μ_0 and the tension force in the string by T_0 which gives:

$$\frac{\partial^2 y}{\partial x^2}-\frac{\mu_0}{T_0}\frac{\partial^2 y}{\partial t}=0 \tag{9.4.59}$$

We see easily that the units of F/ρ_l are those of the square of velocity $[\text{ms}^{-1}]$ as required by dimensional analysis. To simplify the notations, we put:

$$v=\sqrt{\frac{F}{\rho_l}} \tag{9.4.60}$$

and therefore:

$$\frac{\partial^2 y}{\partial x^2}-\frac{1}{v^2}\frac{\partial^2 y}{\partial t}=0 \tag{9.4.61}$$

We will now consider a very interesting case in the context of musicology which is that of the tense string (most string instruments works as this).

9.4.3.4.1 Dirichlet Conditions

The goal is in the framework of the previous differential equation (small deformations in the context of musical instruments) to find a function $y(x,t)$ solution of the latter with the following initial conditions that are typical for a musical instrument :

IC1. $y(0,t)=y(L,t)=0$: the extremities A and B of the string are fixed. These are the "Dirichlet boundary conditions".

IC2. $y(x,0)=f(x)$: the initial shape of the string at excitation.

IC3. $\dfrac{\partial y(x,0)}{\partial t}=0$: initial velocity is zero on any point of the string.

The last two conditions are named "Cauchy boundary conditions".

To solve this linear differential equation, we will make use of the variables separation method by writing (see section Differential and Integral Calculus page 1617):

$$y(x,t)=X(x)T(t) \tag{9.4.62}$$

The differential equation:

$$\frac{\partial^2 y}{\partial x^2}-\frac{\rho_l}{F}\frac{\partial^2 y}{\partial t^2}=0 \tag{9.4.63}$$

Therefore becomes:

$$\begin{gathered}\frac{\partial^2 y}{\partial t^2}=\frac{\partial^2(X(x)T(t))}{\partial t^2}=\frac{\partial}{\partial t}\left(\underbrace{\frac{\partial X(x)}{\partial t}}_{=0}T(t)+X(x)\frac{\partial T(t)}{\partial t}\right)=X(x)\frac{\partial^2 T(t)}{\partial t^2}=\frac{F}{\rho_l}\frac{\partial^2 y}{\partial x^2}\\ \frac{\partial^2 y}{\partial x^2}=\frac{\partial^2(X(x)T(t))}{\partial x^2}=\frac{\partial}{\partial x}\left(\frac{\partial X(x)}{\partial x}T(t)+X(x)\underbrace{\frac{\partial T(t)}{\partial x}}_{=0}\right)=T(t)\frac{\partial^2 X(x)}{\partial x^2}\end{gathered} \tag{9.4.64}$$

So we have:

$$\frac{\dfrac{\partial^2 X}{\partial x^2}}{X}=\frac{\rho_l}{F}\frac{\dfrac{\partial^2 T}{\partial x^2}}{T} \tag{9.4.65}$$

By putting the second relation in the first, we get:

$$\begin{gathered}X(x)\frac{\partial^2 T(t)}{\partial t^2}=\frac{F}{\rho_l}\frac{\partial^2 y}{\partial x^2}\\ \frac{\partial^2 y}{\partial x^2}=T(t)\frac{\partial^2 X(x)}{\partial x^2}\end{gathered} \tag{9.4.66}$$

The left term of the last relation does not contain the variable t and the right does not contain x. The only way to equalize these two expressions is to consider each as a constant, which we will denote by $-\lambda^2$ such that:

$$\frac{\dfrac{\partial^2 X}{\partial x^2}}{X}=\frac{\rho_l}{F}\frac{\dfrac{\partial^2 T}{\partial x^2}}{T}=c^{te}=-\lambda^2 \tag{9.4.67}$$

Thus, we have two differential equations:

$$\frac{\partial^2 X}{\partial x^2}+\lambda^2 X=0 \qquad \text{and} \qquad \frac{\partial^2 T}{\partial t^2}+(v\lambda)^2 T=0 \tag{9.4.68}$$

These two equations are similar, we solve them in a general way (see section Differential and Integral Calculus page 1610):

$$\frac{\partial^2 \alpha(t)}{\partial t^2}+\beta\alpha(t)=0 \qquad \text{with } \beta\in\mathbb{R}^+ \tag{9.4.69}$$

The characteristic equation is then:

$$r^2+\beta=0\Rightarrow\Delta=b^2-4ac=-4\beta \tag{9.4.70}$$

hence:

$$r_{1,2}=\frac{\pm\sqrt{-4\beta}}{2}=\pm\mathrm{i}\sqrt{\beta}=u\pm\mathrm{i}v=0\pm\mathrm{i}v \tag{9.4.71}$$

We know that the general solution if the roots of the characteristic equation are complex ($r_{1,2}\in\mathbb{C}$) is the of the form:

$$\alpha(t)=C_1e^{ut}\cos(vt)+C_2e^{ut}\sin(vt)=C_1\cos(\sqrt{\beta}t)+C_2\sin(\sqrt{B}t) \tag{9.4.72}$$

For our two differential equations, so we have by similitude:

$$\begin{gathered}X(x)=A_1\cos(\lambda x)+B_1\sin(\lambda x)\\ T(t)=A_2\cos(\lambda vt)+B_2\sin(\lambda vt)\end{gathered}\tag{9.4.73}$$

This gives for the solution of our wave equation:

$$y(x,t)=X(x)T(t)=(A_1\cos(\lambda x)+B_1\sin(\lambda x)(A_2\cos(\lambda vt)+B_2\sin(\lambda vt))\tag{9.4.74}$$

Let us determine the constants A_i,B_i by taking into account the initial conditions:

$$y(0,t)=0\Rightarrow 0=A_1(A_2\cos(\lambda vt)+B_2\sin(\lambda vt)),\forall t\Rightarrow A_1=0\tag{9.4.75}$$

Then it only remains:

$$y(x,t)=(B_1\sin(\lambda x))(A_2\cos(\lambda vt)+B_2\sin(\lambda vt))=\sin(\lambda x)(B_1A_2\cos(\lambda vt)+B_1B_2\sin(\lambda vt))\tag{9.4.76}$$

Let us put:

$$B_1B_2=a,\,B_1A_2=b\Rightarrow y(x,t)=\sin(\lambda x)(b\cos(\lambda vt)+a\sin(\lambda vt))\tag{9.4.77}$$

The initial condition $y(L,t)=0,\forall t$ brings us to put that:

$$\begin{aligned}&\sin(\lambda L)=0\Rightarrow\lambda L=n\pi\Rightarrow\lambda=\frac{n\pi}{L},n\in\mathbb{N}\\ &\Rightarrow y(x,t)=\sin\left(\frac{n\pi x}{L}\right)\left[b\cos\left(\frac{n\pi v}{L}t\right)+a\sin\left(\frac{n\pi v}{L}t\right)\right]\end{aligned}\tag{9.4.78}$$

To take into account the zero initial speed, we derivate $y(x,t)$ with respect to time:

$$\begin{aligned}&\frac{\partial y(x,t)}{\partial t}=\sin\left(\frac{n\pi x}{L}\right)\left[-b\frac{n\pi v}{L}\sin\left(\frac{n\pi v}{L}t\right)+a\frac{n\pi v}{L}\cos\left(\frac{n\pi v}{L}t\right)\right]\\ &\Rightarrow\frac{\partial y(x,0)}{\partial t}=\sin\left(\frac{n\pi x}{L}t\right)\left[a\frac{n\pi v}{L}\cos(0)\right]\Rightarrow a=0\end{aligned}\tag{9.4.79}$$

It only remains:

$$y(x,t)=b\sin\left(\frac{n\pi x}{L}t\right)\cos\left(\frac{n\pi v}{L}t\right)\qquad y(x,t)\in[-b,+b]\tag{9.4.80}$$

The constant b represents the amplitude of the transverse displacement of the string. As this amplitude can not be the same everywhere at a given time and a given location for any kind of excitement satisfying the initial conditions, there must be as many values b_n as we choose values of n in $n\pi x/\lambda l$.

The principle of superposition of the solutions of linear differential equations (see section Differential and Integral Calculus page 1605) permits us to write that the linear combination of all solution for the string is therefore:

$$y(x,t)=\sum_{i=1}^{+\infty}b_n\sin\left(\frac{n\pi x}{L}\right)\cos\left(\frac{n\pi v}{L}t\right)\tag{9.4.81}$$

The b_n must be choose to satisfy the initial conditions that gives the shape of the perturbation:

$$f(x)=y(x,0)=\sum_{i=1}^{+\infty} b_n \sin\left(\frac{n\pi x}{L}\right)\cos(0)=\sum_{i=1}^{+\infty} b_n \sin\left(\frac{n\pi x}{L}\right) \tag{9.4.82}$$

This expression for $f(x)$ suggest to compare it at a Fourier series development (see section Sequences and Series page 1758):

$$f(x)=\sum_{n=0}^{+\infty}(a_n\cos(nkx)+b_n\sin(nkx)) \tag{9.4.83}$$

In which $k=\pi/L$ and $a_n=0,\forall n$. Fourier's theorem requires the b_n are given by (see section Sequences and Series page 1727):

$$b_0=0 \qquad b_n=\frac{2}{L}\int_0^L f(x)\sin(nkx)\mathrm{d}x \tag{9.4.84}$$

Let us now imagine a string of length L fixed at its ends and tense. Let us choose the simplest possible excitation form: we scratch the string in its middle in a very quick way, to take it away from a very small distance H from its equilibrium position.

The initial disturbance $y(x,0)$ is then:

$$\begin{aligned} f(x)&=\frac{2Hx}{L} && \text{for } 0\leq x\leq\frac{L}{2}\\ f(x)&=\frac{2H(L-x)}{L} && \text{for } \frac{L}{2}\leq x\leq L \end{aligned} \tag{9.4.85}$$

Let us now compute the Fourier coefficients:

$$\begin{aligned} b_n&=\frac{2}{L}\int\limits_0^L f(x)\sin\left(\frac{n\pi x}{L}\right)\mathrm{d}x\\ &=\frac{2}{L}\left\{\int\limits_0^{L/2}\frac{2Hx}{L}\sin\left(\frac{n\pi x}{L}\right)\mathrm{d}x+\int\limits_{L/2}^{L}\frac{2H(L-x)}{L}\sin\left(\frac{n\pi x}{L}\right)\mathrm{d}x\right\} \end{aligned} \tag{9.4.86}$$

Therefore:

$$b_n=\frac{8H}{\pi^2}\frac{\sin\left(n\frac{\pi}{2}\right)}{n^2} \tag{9.4.87}$$

The wave function becomes:

$$y(x,t)=\sum_{n=1}^{+\infty} b_n\sin\left(\frac{n\pi x}{L}\right)\cos\left(\frac{n\pi v}{L}t\right)=\frac{8H}{\pi^2}\sum_{n=1}^{+\infty}\frac{\sin\left(n\frac{\pi}{2}\right)}{n^2}\sin\left(\frac{n\pi x}{L}\right)\cos\left(\frac{n\pi v}{L}t\right) \tag{9.4.88}$$

Because of the $\sin\left(n\dfrac{\pi}{2}\right)$, the terms for which n is even are all zero. Then it remains:

$$y(x,t)=\frac{8H}{\pi^2}\left[\begin{array}{c}\dfrac{1}{1^2}\sin\left(\dfrac{\pi x}{L}\right)\cos\left(\dfrac{\pi v}{L}t\right)-\dfrac{1}{3^2}\sin\left(\dfrac{3\pi x}{L}\right)\cos\left(\dfrac{3\pi v}{L}t\right)+\\ \dfrac{1}{5^2}\sin\left(\dfrac{\pi x}{L}\right)\cos\left(\dfrac{\pi v}{L}t\right)-\ldots\end{array}\right] \tag{9.4.89}$$

If we retain only the term $n=1$, we keep obviously only:

$$y(x,t)=\frac{8H}{\pi^2}\sin\left(\frac{\pi x}{L}\right)\cos\left(\frac{\pi v}{L}t\right)=\frac{8H}{\pi^2}\sin\left(\frac{2\pi}{2L}x\right)\cos\left(\frac{2\pi}{2L}vt\right)=\frac{8H}{\pi^2}\sin(k_1x)\cos(kvt) \tag{9.4.90}$$

We have:

$$k_1=\frac{2\pi}{2L} \tag{9.4.91}$$

that is the wavenumber corresponding the wavelength:

$$\lambda_1=\frac{2\pi}{k_1}=2L \tag{9.4.92}$$

and:

$$k_1v=\omega_1=2\pi f_1\Rightarrow f_1=\frac{k_1v}{2\pi}=\frac{v}{2L}=\frac{1}{2L}\sqrt{\frac{F}{\rho_l}} \tag{9.4.93}$$

which is the first harmonic frequency vibration of the string.

Thus, for any value n, it is easy to show that the n-th "eigenmode" is given by:

$$y_n(x,t)=\frac{\hat{y}}{n^2}\sin\left(\frac{2\pi}{\lambda_n}x\right)\cos\left(\frac{2\pi}{f_n}t\right) \tag{9.4.94}$$

with:

$$\boxed{\lambda_n=\frac{2L}{n}\qquad f_n=\frac{n}{2L}\sqrt{\frac{F}{\rho_l}}} \tag{9.4.95}$$

relations named "Mersenne's laws" (11644-11648 according to holocene calendar) where the mode of lower frequency with n being equal to 1 is named the "fundamental mode" associated with its "fundamental frequency".

Thus, after being scratched quickly in the middle of its length L, a string held rigidly at both ends can oscillate in several modes. The fundamental mode $n=1$ (fundamental harmonic) is the lowest possible frequency. It corresponds to it the wavelength $\lambda=2L$.

The n higher order frequencies are named "harmonic frequencies". For the same initial displacement H, the maximum amplitude of the vibration decreases according to n^{-2} as we see it in the expression of our function above.

Another way to excite the string is make it oscillates sinusoidally, which means therefore that $y(x,t)$ is of the form:

$$y(x,t)=y(x)\sin(\omega t-\phi) \tag{9.4.96}$$

Substituting this relation in the string wave equation, we get:

$$\frac{\mathrm{d}^2y(x)}{\mathrm{d}x^2}+\omega^2\frac{\rho_l}{F}y(x)=0 \tag{9.4.97}$$

The solution is then reduce to:

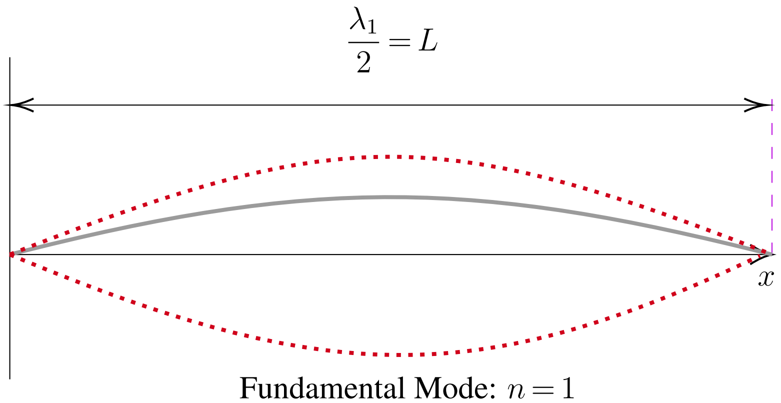
$$y(x)=\sum_{n=1}^{+\infty}b_n\sin\left(\frac{n\pi x}{L}\right) \quad \text{or} \quad y_n(x)=b_n\sin\left(\frac{n\pi x}{L}\right) \quad \forall n\in\mathbb{N}^* \tag{9.4.98}$$

The value $n=0$ can not be included because it gives a string without excitement. By this function in the previous wave equation and simplifying, we get trivially:

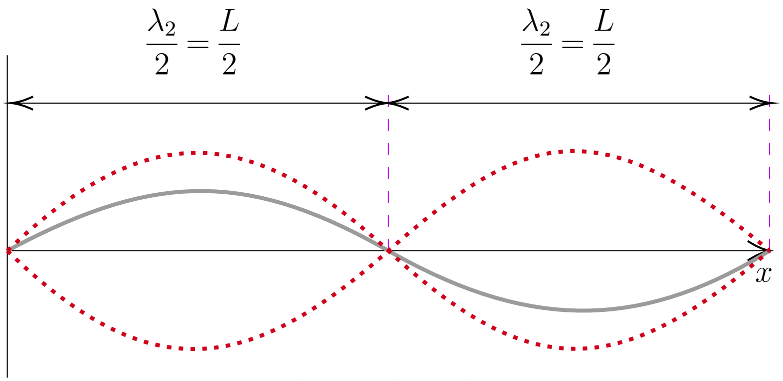
$$\omega_n=\frac{n\pi}{L}\sqrt{\frac{F}{\rho_l}} \tag{9.4.99}$$

These are the "Dirichlet's oscillations frequencies" of a string. The strings of a violin, for example, are Dirichlet strings.

Remember the image seen previously above but focusing this time only on two harmonics:



Fundamental Mode: $n=1$


$\frac{\lambda_2}{2}=\frac{L}{2}$
$\frac{\lambda_2}{2}=\frac{L}{2}$
x

Fundamental Mode: $n=2$

Figure 9.119: First two basic modes of a fixed string

The same analogies, reasoning and developments can be made with the Neumann conditions below.

Remarks

R1. As we have just seen the theory predicts that the vibration can be a linear combination of several modes. This phenomenon is named "simultaneous vibration". It occurs abundantly in a piano.

R2. Many musical instruments are designed to emit sounds with conventional frequencies, being admit that the note recognized by the ear is defined by the fundamental frequency, for example: Do (264 Hertz), The (440 Hertz). For more details see the section of Mathematical Music.

R3. During the construction of musical string instrument, we decide the value of ρ_l (by choosing the diameter and the nature of the string) and we determine its length L by seeking the compromise between the loudness we want to emit and the mechanical strength that the instrument that must support the tension forces F.

9.4.3.4.2 Neumann Conditions

Alternatively to Dirichlet's conditions where the ends are fixed and of equal height, the Neumann conditions assume that the ends are small loops allowed to slide along two bars without friction.

For our string, the Neumann conditions specify the values $\partial y/\partial x$ at the ends. But as the loops are supposed massless and without friction, the derivative $\partial x/\partial x$ must cancel at the end $x=\{0,L\}$ (this is an assumption not easy to guess this is why it has a name...!). If this were not the case, then by the nullity of the mass of the end, the shift will be due to an infinite acceleration, which can not be allowed non-relativistic framework! This is what we require instead of the Dirichlet condition, the Neumann condition is defined by:

IC3. $\dfrac{\partial y}{\partial x}(0,t)=\dfrac{\partial y}{\partial x}(L,t)=0$

the conditions IC1, IC2 remaining the same.

This change of status does not prevent the variables separation method of resolution to be the same as before and that we will fall identically on the following relation to which we will apply the new initial condition (let us know if you want we put the details again):

$$y(x,t)=X(x)T(t)=(A_1\cos(\lambda x)+B_1\sin(\lambda x))(A_2\cos(\lambda vt)+B_2\sin(\lambda vt)) \tag{9.4.100}$$

on which we therefore apply the Neumann conditions:

$$\begin{aligned}&\frac{\partial y}{\partial x}(0,t)=0\\&\frac{\partial y}{\partial x}=\lambda\left(-A_1\sin(\lambda x)+B_1\cos(\lambda x)\right)\left(A_2\cos(\lambda vt)+B_2\sin(\lambda vt)\right)\\&\frac{\partial y}{\partial x}(0,t)=0\Rightarrow\lambda B_1(A_2\cos(\lambda vt)+B_2\sin(\lambda vt))=0,\forall t\Rightarrow B_1=0\end{aligned} \tag{9.4.101}$$

Then it remains:

$$y(x,t)=\cos(\lambda x)(A_1A_2\cos(\lambda vt)+A_1B_2\sin(\lambda vt)) \tag{9.4.102}$$

by putting $b=A_1A_2$,$a=A_1B_2$ the function simplifies in:

$$y(x,t)=\cos(\lambda x)(b\cos(\lambda vt)+a\sin(\lambda vt)) \tag{9.4.103}$$

The initial condition $\partial y/\partial x(L,t)=0,\forall t$ requires:

$$\begin{aligned}&-\sin(\lambda L)=0\Rightarrow\lambda L=n\pi\Rightarrow\lambda=\frac{n\pi}{L}\qquad n\in\mathbb{N}\\&\Rightarrow y(x,t)=\cos\left(\frac{n\pi x}{L}\right)\left[b\cos\left(\frac{n\pi v}{L}t\right)+a\sin\left(\frac{n\pi v}{L}t\right)\right]\end{aligned} \tag{9.4.104}$$

The same developments for the IC2. we have made with the Dirichlet condition IC1 then apply equally (we can write the details on request):

$$y(x,t)=b\cos\left(\frac{n\pi x}{L}\right)\cos\left(\frac{n\pi x}{L}t\right) \tag{9.4.105}$$

then, the analogy with Fourier series applies similarly but with the cosine in place of the sine.

The Neumann frequencies of a string are the same as for the Dirichlet either:

$$\omega_n = \frac{n\pi}{L}\sqrt{\frac{F}{\rho_l}} \tag{9.4.106}$$

The particularity, however, is the value of spatial function that is trivially this time:

$$y(x)=\sum_{n=0}^{+\infty} b_n \cos\left(\frac{n\pi x}{L}\right) \quad \text{or} \quad y_n(x)=b_n\cos\left(\frac{n\pi x}{L}\right) \qquad n\in\mathbb{N} \tag{9.4.107}$$

Indeed, for $n=0$ we have this time an identical amplitude b_n which is transmitted along the string without that it does vibrates!!!

Let us also notice that the function $y(x,t)=at+b$ also fully satisfies the three initial conditions including that of Neumann.

Indeed, we have well:

IC1. $\dfrac{\partial y}{\partial x}(0,t)=\dfrac{\partial y}{\partial x}(L,t)=0$

IC2. $y(x,0)=f(x)=b$

IC3. $\dfrac{\partial y(x,0)}{\partial t}=0$

and in addition, $y(x,t)=at+b$ satisfies also the wave equation:

$$\frac{\partial^2 y}{\partial x^2}-\frac{1}{v^2}\frac{\partial^2 y}{\partial t^2}=0 \tag{9.4.108}$$

9.4.4 Non-relativistic Lagrangian of a String

We will now determine the non-relativistic Lagrangian of a string, calculation that will be us useful to us for our study of String Theory in the chapter of Cosmology.

We keep our string having a constant linear density and tension whose ends are located at $x=0, x=L$ and the speed of for which the transversal speed is not relativistic $v\ll c$.

The kinetic energy is then simply the sum of the kinetic energies of each infinitesimal element of the string. We can then write in Lagrangian notation:

$$T=\int_0^L \frac{1}{2}(\rho_l \mathrm{d}x)\left(\frac{\partial y}{\partial t}\right)^2 \tag{9.4.109}$$

The potential energy is involved in the elongation of the string from which an infinitesimal portion can be seen as varying from $(x,0)$ to $(x+\mathrm{d}x,0)$ when the string is in equilibrium.

When a string is momentarily stressed from (x,y) to $(x+\mathrm{d},y+\mathrm{d}y)$ then the variation of the length $\mathrm{d}l$ of an infinitesimal element of the string is trivially given by:

$$\mathrm{d}l=\sqrt{(\mathrm{d}x)^2+(\mathrm{d}y)^2}-\mathrm{d}x=\mathrm{d}x\left(\sqrt{1+\left(\frac{\partial y}{\partial x}\right)^2}-1\right)\cong\mathrm{d}x\frac{1}{2}\left(\frac{\partial y}{\partial x}\right)^2 \tag{9.4.110}$$

We used above for the approximation the Taylor expansion at the second order (see section Sequences and Series page 1703), which gives us:

$$\sqrt{1+x}=1+\frac{1}{2}x-\frac{1}{2\cdot 4}x^2+\frac{1\cdot 3}{2\cdot 4\cdot 6}x^3-\ldots \qquad |x|\leq 1 \tag{9.4.111}$$

The work done to stretch every infinitesimal element being $F\Delta l$, the total potential energy is then expressed as:

$$V=\int\limits_0^L F\mathrm{d}l\cong\int\limits_0^L\frac{1}{2}F\left(\frac{\partial y}{\partial x}\right)^2\mathrm{d}x \tag{9.4.112}$$

The Lagrangian being defined by $L=T-V$ (see section Analytical Mechanics page 3152), then we have:

$$L=\int\limits_0^L\left[\frac{1}{2}\rho_l\left(\frac{\partial y}{\partial t}\right)-\frac{1}{2}F\left(\frac{\partial y}{\partial x}\right)^2\right]\mathrm{d}x=\int\limits_0^L\mathcal{L}\mathrm{d}x \tag{9.4.113}$$

where $\mathcal{L}$ is defined, well correctly, as the "Lagrangian density of a string":

$$\boxed{\mathcal{L}=\frac{1}{2}\rho_l\left(\frac{\partial y}{\partial t}\right)-\frac{1}{2}F\left(\frac{\partial y}{\partial x}\right)^2} \tag{9.4.114}$$

The action of the string is therefore:

$$S=\int\limits_{t_i}^{t_f}L\mathrm{d}t=\int\limits_{t_i}^{t_f}\mathrm{d}t\int\limits_0^L\mathrm{d}x\left[\frac{1}{2}\rho_l\left(\frac{\partial y}{\partial t}\right)-\frac{1}{2}F\left(\frac{\partial y}{\partial x}\right)^2\right] \tag{9.4.115}$$

In this action, the path of action is the function $y(x,t)$. To find the equations of motion, we must consider the variation of the action when we vary:

$$y(x,t)\rightarrow y(x,t)+\delta y(x,t) \tag{9.4.116}$$

Which gives:

$$\begin{aligned}\delta S&=\int\limits_{t_i}^{t_f}\mathrm{d}t\int\limits_0^L\mathrm{d}x\left[\frac{1}{2}\rho_l\left(\frac{\partial(y(x,t)+\delta y(x,t))}{\partial t}\right)^2-\frac{1}{2}F\left(\frac{\partial(y(x,t)+\delta y(x,t))}{\partial x}\right)^2\right]\\&=\int\limits_{t_i}^{t_f}\mathrm{d}t\int\limits_0^L\mathrm{d}x\left[\begin{array}{l}\frac{1}{2}\rho_l\left(\frac{\partial y(x,t)^2}{\partial t^2}+\frac{2\partial y(x,t)\partial(\delta y(x,t))}{\partial t^2}+\frac{\partial(\delta y(x,t))^2}{\partial t^2}\right)\\-\frac{1}{2}F\left(\frac{\partial y(x,t)^2}{\partial x^2}+\frac{2\partial y(x,t)\partial(\delta y(x,t))}{\partial x^2}+\frac{\partial(\delta y(x,t))^2}{\partial x^2}\right)\end{array}\right]\\&\cong\int\limits_{t_i}^{t_f}\mathrm{d}t\int\limits_0^L\mathrm{d}x\left[\rho_l\frac{\partial y}{\partial t}\frac{\partial(\delta y(x,t))}{\partial t}-F\frac{\partial y}{\partial x}\frac{\partial(\delta y(x,t))}{\partial x}\right]\\&=\int\limits_{t_i}^{t_f}\mathrm{d}t\int\limits_0^L\mathrm{d}x\left[\rho_l\frac{\partial y}{\partial t}\frac{\partial(\delta y)}{\partial t}-F\frac{\partial y}{\partial x}\frac{\partial(\delta y)}{\partial x}\right]\end{aligned}\tag{9.4.117}$$

As:

$$\left(\frac{\partial(\delta y(x,t))}{\partial t}\right)^2\ll\left(\frac{\partial y(x,t)}{\partial t}\right)^2\ll\frac{2\partial y(x,t)\partial(\delta y(x,t))}{\partial t^2}\tag{9.4.118}$$

and this identically for the second term.

We should not have time derivatives acting on the variations. Then using the following trivial relation on the first term:

$$\frac{\partial}{\partial t}\rho_l\left(\frac{\partial y}{\partial t}\delta y\right)=\rho_l\frac{\partial^2y}{\partial t^2}\delta y+\rho_l\frac{\partial y}{\partial t}\frac{\partial(\delta y)}{\partial t}\Rightarrow\rho_l\frac{\partial(\delta y)}{\partial t}=\frac{\partial}{\partial t}\rho_l\left(\frac{\partial y}{\partial t}\delta y\right)-\rho_l\frac{\partial^2y}{\partial t^2}\delta y\tag{9.4.119}$$

and identically on the second, we can rewrite the action:

$$\begin{aligned}\delta S=&\int\limits_{t_i}^{t_f}\mathrm{d}t\int\limits_0^L\mathrm{d}x\left[\rho_l\frac{\partial y}{\partial t}\frac{\partial(\delta y)}{\partial t}-F\frac{\partial y}{\partial x}\frac{\partial(\delta y)}{\partial x}\right]\\=&\int_{t_i}^{t_f}\mathrm{d}t\int\limits_0^L\mathrm{d}x\left[\frac{\partial}{\partial t}\left(\rho_l\frac{\partial y}{\partial t}\delta y\right)-\rho_l\frac{\partial^2y}{\partial t^2}\delta y+\frac{\partial}{\partial x}\left(-F\frac{\partial y}{\partial x}\delta y\right)+F\frac{\partial^2y}{\partial x^2}\delta y\right]\\=&\int\limits_{t_i}^{t_f}\left(\rho_l\frac{\partial y}{\partial t}\delta y\right)\int\limits_0^L\mathrm{d}x-\int\limits_{t_i}^{t_f}\mathrm{d}t\int\limits_0^L\mathrm{d}x\rho_l\frac{\partial^2y}{\partial t^2}\delta y+\int\limits_0^L\partial\left(-F\frac{\partial y}{\partial x}\delta y\right)\int\limits_{t_i}^{t_f}\mathrm{d}t+\int\limits_{t_i}^{t_f}\mathrm{d}t\int\limits_0^L\mathrm{d}xF\frac{\partial^2y}{\partial x^2}\delta y\end{aligned}\tag{9.4.120}$$

As we have seen it in the section of Analytical Mechanics, the correct path is given by $\delta S=0$. Therefore, we must have:

$$\boxed{\delta S=\int\limits_0^L\left[\rho_l\frac{\partial y}{\partial t}\delta y\right]_{t_i}^{t_f}\mathrm{d}x+\int\limits_{t_i}^{t_f}\left[-F\frac{\partial y}{\partial x}\delta y\right]_0^L\mathrm{d}t-\int\limits_{t_i}^{t_f}\mathrm{d}t\int\limits_0^L\mathrm{d}x\left(\rho_l\frac{\partial^2y}{\partial t^2}-F\frac{\partial^2y}{\partial x^2}\right)\delta y=0}\tag{9.4.121}$$

Thus our expression contains three terms. Each of these terms must vanish independently as we shall see:

1. The cancellation of the third term is done following a trivial condition that we are already well known (hopefully ...):

   $$\int\limits_{t_i}^{t_f}\mathrm{d}t\int\limits_0^L\mathrm{d}x\left(\rho_l\frac{\partial^2 y}{\partial t^2}-F\frac{\partial^2 y}{\partial x^2}\right)\delta y=0\Rightarrow\rho_l\frac{\partial^2 y}{\partial t^2}-F\frac{\partial^2 y}{\partial x^2}=0 \tag{9.4.122}$$

   and therefore:

   $$\frac{\partial^2 y}{\partial x^2}-\frac{\rho_l}{F}\frac{\partial^2 y}{\partial t^2}=0 \tag{9.4.123}$$

   So we fall back on the differential equation of a transverse wave as we had proved earlier above. Our hypothesis about the third term can then only be correct as well as the expression of our action.

2. The first term is determined by the configuration of the string at the times t_i,t_f:

   $$\int\limits_0^L\left[\rho_l\frac{\partial y}{\partial t}\delta y\right]_{t_i}^{t_f}\mathrm{d}x \tag{9.4.124}$$

   Now, if we apply the knowledge of these configurations in time, we have by definition:

   $$\delta y(x,t_f)=\delta y(x,t_i)=0 \tag{9.4.125}$$

   (total knowledge of the action path as knowledge of the initial conditions, so no variation). This validates once again the expression of our action and the zero value of this term as expected.

3. The second term is a little more interesting:

   $$\int\limits_{t_i}^{t_f}\left[-F\frac{\partial y}{\partial x}\delta y\right]_0^L\mathrm{d}t \tag{9.4.126}$$

   First, it is only because we know the positions of the ends of the string that we can know its vibration modes, we know that well! So we have to know how the ends behave. For that we will come back on things that are well known to us: the Dirichlet and Neumann conditions of a string.

   Suppose we assume the Dirichlet conditions (see above), the ends are then fixed and we will have necessarily at those ends:

   $$\delta y(0,t)=\delta y(L,t)=0 \tag{9.4.127}$$

   and then the second term vanish well (phew!).

   If, in contrast, we choose the ends to move freely, then the variations:

   $$\delta y(0,t)=\delta y(L,t)=0 \tag{9.4.128}$$

   are unconstrained and therefore, only the Neumann conditions:

   $$\frac{\partial y}{\partial x}(t,0)=\frac{\partial y}{\partial x}(t,L)=0 \tag{9.4.129}$$

(see above for more details) will permit us to have the second term of the action equal to zero.

To be fully aware of the importance of these initial conditions, let us consider the linear momentum p_y carried by the string (there are no other motion of the as we have implicitly assumed a transverse-excitement from since the beginning only in this direction y).

The linear momentum is simply the sum of the linear momentum of each infinitesimal element along the string:

$$p_y = \int\limits_0^L (\rho_l \mathrm{d}x)\frac{\partial y}{\partial t} \tag{9.4.130}$$

Let us check out by curiosity (it's an anticipated curiosity ...) if the momentum is well preserved:

$$\frac{\mathrm{d}p_y}{\mathrm{d}t} = \int\limits_0^L (\rho_l \mathrm{d}x)\frac{\partial^2 y}{\partial t^2} = \int\limits_0^L F\frac{\partial^2 y}{\partial x^2}\mathrm{d}x = F\frac{\partial y}{\partial x}\bigg|_{x=0}^{x=L} \tag{9.4.131}$$

where we have used the transverse wave equation for the substitution.

We see by the result of this little calculation that the momentum is trivially conserved if we impose the Neumann conditions, whereas for the Dirichlet conditions, most of time conservation is not respected! Indeed, it is trivial (there is no need for calculations to realize it normally), when the ends of the string are attached to the wall, the wall constantly exerts a force on the string.

9.4.5 Vibrational modes of a circular membrane

A two-dimensional elastic membrane under tension can support transverse vibrations. The properties of an idealized drumhead can be modelled by the vibrations of a circular membrane of uniform thickness, attached to a rigid frame. Due to the phenomenon of resonance, at certain vibration frequencies, its resonant frequencies, the membrane can store vibrational energy, the surface moving in a characteristic pattern of standing waves. This is named a "normal mode". A membrane has an infinite number of these normal modes, starting with a lowest frequency one named the fundamental mode.

There exist infinitely many ways in which a membrane can vibrate, each depending on the shape of the membrane at some initial time, and the transverse velocity of each point on the membrane at that time. The vibrations of the membrane are given by the solutions of the two-dimensional wave equation with Dirichlet boundary conditions which represent the constraint of the frame. It can be shown that any arbitrarily complex vibration of the membrane can be decomposed into a possibly infinite series of the membrane's normal modes. This is analogous to the decomposition of a time signal into a Fourier series.

We derive the phenomenon of vibration of a circular membrane in tension (typically a drum) in the same manner that the transverse vibration of a string whose equation was for recall:

$$\frac{\partial^2 y}{\partial x^2} - \frac{\rho_l}{F}\frac{\partial^2 y}{\partial t^2} = 0 \tag{9.4.132}$$

But first if we consider a vertical slice of the drum we change y by z as we work in 3D space and that the z-axes is perpendicular to the dream membrane:

$$\frac{\partial^2 z}{\partial x^2}-\frac{\rho_l}{F}\frac{\partial^2 z}{\partial t^2}=0 \tag{9.4.133}$$

By the central symmetry of the problem we can also write (as the general solution is the sum of special solutions and we will consider that the force F takes into account the sum of the special cases):

$$\frac{\partial^2 z}{\partial x^2}+\frac{\partial^2 z}{\partial y^2}-\frac{\rho_l}{F}\frac{\partial^2 z}{\partial t^2}=0 \tag{9.4.134}$$

However, the linear density ρ_l of the string must be replaced by the surface density ρ_s of the membrane denoted σ by tradition. And for the coherence of units we must replace the unidirectional tension force F of the string by a tension force applied on the periphery of the membrane. This force is exerted in all directions of the plane and is described per unit length:

$$\tau=\frac{\mathrm{d}F}{\mathrm{d}l} \tag{9.4.135}$$

So finally:

$$\frac{\partial^2 z}{\partial x^2}+\frac{\partial^2 z}{\partial y^2}-\frac{\sigma}{\frac{F}{l}}\frac{\partial^2 z}{\partial t^2}=0 \tag{9.4.136}$$

Or in a shorter way using the Laplace operator (see section Vector Calculus page 1899):

$$\vec{\nabla}^2 z-\frac{\sigma}{\tau}\frac{\partial^2 z}{\partial t^2}=0 \tag{9.4.137}$$

Or in a more conventional way:

$$\boxed{\vec{\nabla}^2 z=\frac{1}{v^2}\frac{\partial^2 z}{\partial t^2}} \tag{9.4.138}$$

with:

$$\boxed{v=\sqrt{\frac{\tau}{\sigma}}} \tag{9.4.139}$$

We seek the particular solution of the Laplace equation which satisfies the following conditions:

IC1. The membrane is fixed on its circular periphery R (boundary conditions)

IC2. The initial position and speed are given (initial conditions)

The symmetry of the problem suggests using the Laplacian of a scalar field in polar coordinates (see section Vector Calculus page 1899):

$$\frac{\partial^2 z}{\partial r^2}+\frac{1}{r}\frac{\partial z}{\partial r}+\frac{1}{r^2}\frac{\partial^2 z}{\partial \theta^2}=\frac{1}{v^2}\frac{\partial^2 z}{\partial t^2} \tag{9.4.140}$$

And the fixed conditions:

C1. $\forall t \geq 0, \forall \theta \in [0,2\pi], z(r,\theta,t)=0$ (boundary conditions)

C2. $\forall \theta \in [\theta,2\pi], \forall r \in [0,R] \begin{cases} z(r,\theta,0)=f(r,\theta) \\ \partial_t z(r,\theta,0)=g(r,\theta) \end{cases}$ (initial conditions)

where the functions f,g representing the shape of the drum at $t=0$ and the speed of its membrane also at $t=0$ must be given.

Again, to find the solution, we will use the method of separation of the variables (see section Differential and Integral Calculus page 1617) such that:

$$z(r,\theta,t)=X(r,\theta)T(t) \tag{9.4.141}$$

and as identically as for the rope:

$$\frac{\dfrac{\partial^2 X}{\partial x^2}}{x}=\frac{1}{v^2}\frac{\dfrac{\partial^2 T}{\partial t^2}}{T} \tag{9.4.142}$$

and identically that for the rope, we get for the temporal part T a solution of the type:

$$T(t)=A\cos(\lambda vt)+B\sin(\lambda vt) \tag{9.4.143}$$

For $X(r,\theta)$ the method changes with respect to the case of the rope seen above because we now have a differential equation with two variables such that:

$$\frac{\partial^2 X(r,\theta)}{\partial (r,\theta)^2}+\lambda^2 x(r,\theta)=0 \tag{9.4.144}$$

To integrate this last equation, we will seek the solutions also in the following separated form:

$$X(r,\theta)=R(r)\Theta(\theta) \tag{9.4.145}$$

we get by putting this in the differential equation:

$$R''\Theta+\frac{1}{r}R'\Theta+\frac{1}{r^2}R\Theta''+\lambda^2 R\Theta=0 \tag{9.4.146}$$

because for recall we have in polar coordinates (see section Differential and Integral Calculus page 1866):

$$\vec{\nabla}^2 f=\frac{\partial^2 f}{\partial r^2}+\frac{1}{r}\frac{\partial f}{\partial r}+\frac{1}{r^2}\frac{\partial^2 f}{\partial \phi^2} \tag{9.4.147}$$

Hence, by separating the variables:

$$\frac{\Theta''}{\Theta}(\theta)=-\frac{1}{R(r)}\left(r^2R''(r)+rR'(r)+\lambda^2 r^2 R(r)\right) \tag{9.4.148}$$

The left-hand side of the last relation does not contain the variable r and the right-hand side does not contain the variable θ. The only way to equal these two expressions is to consider them each as equal to as constant we will denote μ such that:

$$\frac{\Theta''}{\Theta}(\theta)=\mu=-\frac{1}{R(r)}\left(r^2R''(r)+rR'(r)+\lambda^2 r^2 R(r)\right) \tag{9.4.149}$$

The differential equations verified by R and Θ are then:

$$r^2R''(r)+rR'(r)+\left(\lambda^2r^2+\mu\right)R(r)=0 \qquad \Theta''(\theta)-\mu\Theta(\theta)=0 \tag{9.4.150}$$

The function Θ must obviously be periodic of period 2π, so there exists a natural number n such that $\mu=-n^2$ and therefore identically to the rope, we get quite immediately:

$$\Theta(\theta)=C\cos(n\theta)+D\sin(n\theta) \tag{9.4.151}$$

If we put $\mu=n^2$ in the first differential equation, we get:

$$r^2R''(r)+rR'(r)+(\lambda^2r^2-n^2)R(r)=0 \tag{9.4.152}$$

That we often found in the literature in the following form:

$$R''(r)+\frac{1}{r}R'(r)+\left(\lambda^2-\frac{n^2}{r^2}\right)R(r)=0 \tag{9.4.153}$$

To simplify, we perform the change of variable $x=\lambda r$. That is:

$$x=\lambda r\Rightarrow \mathrm{d}x=\lambda\mathrm{d}r\Rightarrow \mathrm{d}x=\lambda^2\mathrm{d}r^2 \tag{9.4.154}$$

The differential equation therefore becomes:

$$\lambda^2\frac{\mathrm{d}R^2}{\mathrm{d}x^2}+\frac{\lambda^2}{x}\frac{\mathrm{d}R}{\mathrm{d}x}+\left(\lambda^2-\frac{n^2\lambda^2}{x^2}\right)R=0 \tag{9.4.155}$$

After dividing by λ^2 and multiplying by x^2 the differential equation becomes:

$$x^2\frac{\mathrm{d}^2R}{\mathrm{d}x^2}+x\frac{\mathrm{d}R}{\mathrm{d}x}+(x^2-n^2)R=0 \tag{9.4.156}$$

We recognize here the Bessel differential equation of order n as we have introduce it with its special solution in the section of Sequences and Series (see page 1775). Therefore, a special solution is of the type (as the general solution is an infinite sum of Bessel functions by linearity!):

$$R(r)=J_n(x)=J_n(\lambda r) \tag{9.4.157}$$

What ultimately gives us:

$$z(r,\theta,t)=R(r)\Theta(\theta)T(t)=J_n(\lambda r)\left(C\cos(n\theta)+D\sin(n\theta)\right)\left(A\cos(\lambda vt)+B\sin(\lambda vt)\right) \tag{9.4.158}$$

Among the solutions to this equation, let us look for those which satisfy the boundary conditions by putting $R_{\text{membrane}}=L$:

$$\forall\theta\in[0,2\pi],\forall t\geq 0 \qquad R(L)\Theta(\theta)T(t)=0 \tag{9.4.159}$$

Unless Θ or T is the null function, which gives for solution the equilibrium position ... (which probably does not satisfy the initial conditions), we must have $R(L)=0$, that is to say:

$$J_n(\lambda r)=0 \tag{9.4.160}$$

The Bessel function J_n has an infinity of positive zeros that we will denote here by $(\xi_m^n)_{m\in\mathbb{N}}$ (just plot this function with a computer to see it such as with Maple™ 4.00b by writing the command: `>plot(BesselJ(2,x),x=0...100);` where you can change the order value 2 by another value) that provide an infinity of suitable values of λ such that:

$$\lambda_m^n=\frac{\xi_m^n}{L} \tag{9.4.161}$$

This finally corresponds to an infinity of solutions of the initial differential equation that we can write:

$$z_m^n(r,\theta,t)=J_n(\lambda_m^n r)\left[(\alpha_m^n\cos(\omega_m^n t)+\beta_m^n\sin(\omega_m^n t))\cos(n\theta)+(\gamma_m^n\cos(\omega_m^n t)+\delta_m^n\sin(\omega_m^n t))\sin(n\theta)\right] \tag{9.4.162}$$

by having modified the name of the integration constants and having put $\omega_m^n=\lambda_m^n v$ (which verifies the dimensional analysis). Now that this solution satisfies the boundary conditions, we must address the initial conditions.

First for the same reasons as the rope, the general solution is the linear superposition of solutions such as:

$$y(r,\theta,t)=\sum_{n=0}^{+\infty}\sum_{m=1}^{+\infty}y_m^n(r,\theta,t) \tag{9.4.163}$$

We will determine the coefficients α_m^n, β_m^n, γ_m^n, δ_m^n so that the solution $y(r,\theta,t)$ given above also satisfies the initial conditions, that is to say:

$$\begin{aligned}f(r,\theta,0)&=\sum_{n=0}^{+\infty}\sum_{m=1}^{+\infty}J_n(\lambda_m^n r)\left(\alpha_m^n\cos(n\theta)+\gamma_m^n\sin(n\theta)\right)\\ g(r,\theta,0)&=\sum_{n=0}^{+\infty}\sum_{m=1}^{+\infty}\omega_m^n J_n(\lambda_m^n r)\left(\alpha_m^n\cos(n\theta)+\gamma_m^n\sin(n\theta)\right)\end{aligned} \tag{9.4.164}$$

These two relations are similar, let us study the first. It can be written:

$$f(r,\theta,0)=\sum_{n=0}^{+\infty}\left[\left(\sum_{m=1}^{+\infty}\alpha_m^n J_n(\lambda_m^n r)\right)\cos(n\theta)+\left(\sum_{m=1}^{+\infty}\gamma_m^n J_n(\lambda_m^n r)\right)\sin(n\theta)\right] \tag{9.4.165}$$

Which is the Fourier series expansion of the function $f(r,\theta)$ (see section Sequences and Series page 1727). Therefore we have (see section Sequences and Series page 1727):

$$\begin{aligned}\sum_{m=1}^{+\infty}\alpha_m^0 J_n(\lambda_m^0 r)&=\frac{1}{2\pi}\int_{-\pi}^{+\pi}f(r,\theta)\mathrm{d}\theta=f_0(r)\\ \sum_{m=1}^{+\infty}\alpha_m^n J_n(\lambda_m^n r)&=\frac{1}{\pi}\int_{-\pi}^{+\pi}f(r,\theta)\cos(n\theta)\mathrm{d}\theta=f_n(r)\qquad n\in\mathbb{N}^*\\ \sum_{m=1}^{+\infty}\gamma_m^n J_n(\lambda_m^n r)&=\frac{1}{\pi}\int_{-\pi}^{+\pi}f(r,\theta)\sin(n\theta)\mathrm{d}\theta=g_n(r)\qquad n\in\mathbb{N}^*\end{aligned} \tag{9.4.166}$$

Using the orthogonality of the Bessel functions (see section Functional Analysis page 2382) we can deduce from these relations the coefficients α_m^0, α_m^n, γ_m^n (and the same for the others).

For this purpose, let us suppose n fixed and let us put $Z_m(r)=J_n(\lambda_m^n r)$. Let us show that for $m\neq p\Rightarrow Z_m\circ Z_p$ where the dot product is defined for Bessel functions (see section Functional Analysis page 2382):

$$Z_m\circ Z_p=\int\limits_0^L rZ_m(r)Z_p(r)\mathrm{d}r \tag{9.4.167}$$

Since Z_m, Z_p satisfy the differential equation on $R(r)$, we have:

$$\begin{cases} r^2Z'_m+rZ'_m(r)+((\lambda_m^n r)^2-n^2)\,Z_m(r)=0\\ r^2Z'_p+rZ'_p(r)+\left((\lambda_p^n r)^2-n^2\right)\,Z_p(r)=0\end{cases} \tag{9.4.168}$$

By combining these two relations we get:

$$\left((\lambda_p^n)^2-(\lambda_m^n)^2\right)rZ_mZ_p=r(Z''_mZ_p-Z'_pZ_m)+(Z'_mZ_p-Z'_pZ_m)=\left(r(Z'_mZ_p-Z'_pZ_m)\right)' \tag{9.4.169}$$

By integrating member to member between 0 and L and taking into account that:

$$J_n(x)=0 \qquad Z_p(L)=J_n(\lambda_p^n L)=0 \qquad Z_m(L)=J_n(\lambda_m^n L)=0 \tag{9.4.170}$$

we get:

$$\left((\lambda_p^n)^2-(\lambda_m^n)^2\right)\int\limits_0^L rZ_m(r)Z_p(r)\mathrm{d}r=\left(r(Z'_mZ_p-Z'_pZ_m)\right)\Big|_0^L=0 \tag{9.4.171}$$

Hence the result stated earlier since $m\neq p\Rightarrow\lambda_m^n\neq\lambda_p^n$.

The relation:

$$\sum_{m=1}^{+\infty}\alpha_m^nJ_n(\lambda_m^n r)=\frac{1}{\pi}\int\limits_{-\pi}^{+\pi}f(r,\theta)\cos(n\theta)\mathrm{d}\theta=f_n(r) \tag{9.4.172}$$

Can therefore be written:

$$f_n=\sum_{m=1}^{+\infty}\alpha_m^nZ_m \tag{9.4.173}$$

Using the orthogonality Z_m, Z_p for $m\neq p$ we deduce:

$$f_n\circ Z_p=\left(\sum_{m=1}^{+\infty}\alpha_m^nZ_m\right)\circ Z_p=\sum_{m=1}^{+\infty}\alpha(Z_m\circ Z_p)=\alpha_p^n(Z_m\circ Z_p) \tag{9.4.174}$$

The coefficients α_p^n are therefore given by:

$$\alpha_p^n=\frac{f_n\circ Z_p}{Z_\circ Z_p}=\frac{\int_0^L rf_n(r)J_n(\lambda_p^n r)\mathrm{d}r}{\int_0^L r\left(J_n(\lambda_p^n r)\right)^2\mathrm{d}r} \tag{9.4.175}$$

What is not easy to calculate by hand... and this so far... is only for the α_p^n coefficients!!!!!

So we will assume that the reader can imagine that the same process is valid for the other coefficients. To avoid frustration however we will use Maple™ 5.00 to go a little further:

```
> F:=f(r,theta,t);
> wave_eq:=linalg[laplacian](F,[r,theta,z],coords=cylindrical)-diff(F,t,t);
# we take the Laplacian in polar coordinates
> eval(subs(F=A(r,theta)*T(t),wave_eq)); #we apply separation of
variables
> expand(eval(subs(F=A(r,theta)*T(t),wave_eq))/A(r,theta)/T(t));
#we simplify the result
> time_part:=numer(select(has,expand(eval(subs(F=A(r,theta)*T(t),wave_eq))/
A(r,theta)/T(t)),t)-lambda^2); #we take the temporal part
> space_part:=numer(select(has,expand(eval(subs(F=A(r,theta)*T(t),wave_eq))/
A(r,theta)/T(t)),r)+lambda^2); #we take now the spatial part
> sol_time:=dsolve(time_part,T(t)); #we solve the spatial part
> eval(subs(A(r,theta)=phi(r)*U(theta),space_part));
#we split the spatial part in radial and polar part
> expand(eval(subs(A(r,theta)=phi(r)*U(theta),space_part))/phi(r)/U(theta));
#we simplify
> polar_part:=numer(select(has,expand(eval(subs(A(r,theta)=phi(r)*U(theta),
space_part))/phi(r)/U(theta)),theta)+kappa^2);
#we take only the polar part
> radius_part:=numer(select(has,expand(eval(subs(A(r,theta)=phi(r)*U(theta),
space_part))/phi(r)/U(theta)),r)-kappa^2); #we take only the radial part
> dsolve(polar_part,U(theta)); #we solve the angular part
> assume(kappa,integer); #we impose kappa as being an integer number
> dsolve(radius_part,phi(r)); #we solve the radial part that makes appear
the Bessel functions
> dsolve(radius_part,phi(r),series); #the same solution expressed in
limited expansion
> sol_radius:=eval(subs(BesselY=0,dsolve(radius_part,phi(r)))); #we the a
special solution
> sol_lambda:=BesselJZeros(kappa,nu); #we know that the answers are of
the type J Bessel functions therefore the take the zeros in the real set
(requires a Maple™ version great than the 4.00b)
> sol_radius:=BesselJ(kappa,sol_lambda*r);
> plots[animate3d]([r,theta,subs(kappa=2,nu=3,sol_radius*cos(sol_lambda*t)*
cos(kappa*theta))], r=0..1,theta=0..2*Pi,t=0..2*Pi/subs(kappa=2,nu=3,
sol_lambda),coords=cylindrical,frames=20);
```

For those who have a PDF reader compatible with Adobe Flash installed and activated, here is the resulting animated plot (otherwise see here: `https://vimeo.com/575738242`):

and for those who cannot see the animation here is a static decomposed plot:

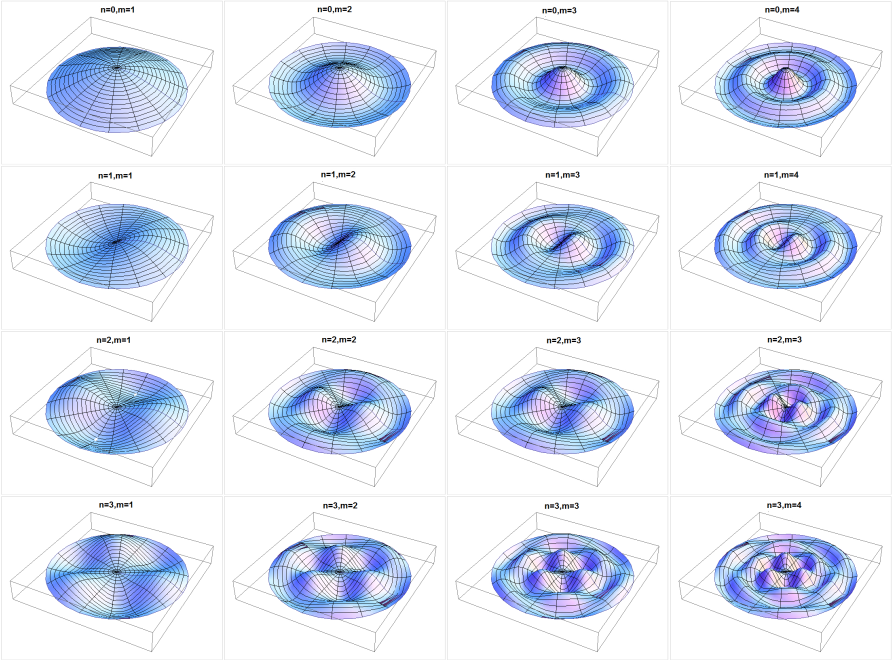

n=0,m=1
n=0,m=2
n=0,m=3
n=0,m=4
n=1,m=1
n=1,m=2
n=1,m=3
n=1,m=4
n=2,m=1
n=2,m=2
n=2,m=3
n=2,m=3
n=3,m=1
n=3,m=2
n=3,m=3
n=3,m=4

Figure 9.120: Some vibration modes of a circular membrane

In real life with a Chladni plate (from Ernst Florens Friedrich Chladni 11756-11827 according to holocene calendar) that is a thin metal sheet - typically square or circular - mounted at its center and set into vibration so that standing-wave patterns form across its surface; when fine sand or powder is sprinkled on the plate, it migrates toward the nodal lines (the regions that remain still), revealing intricate, symmetric patterns that visually map the plate's vibrational modes.

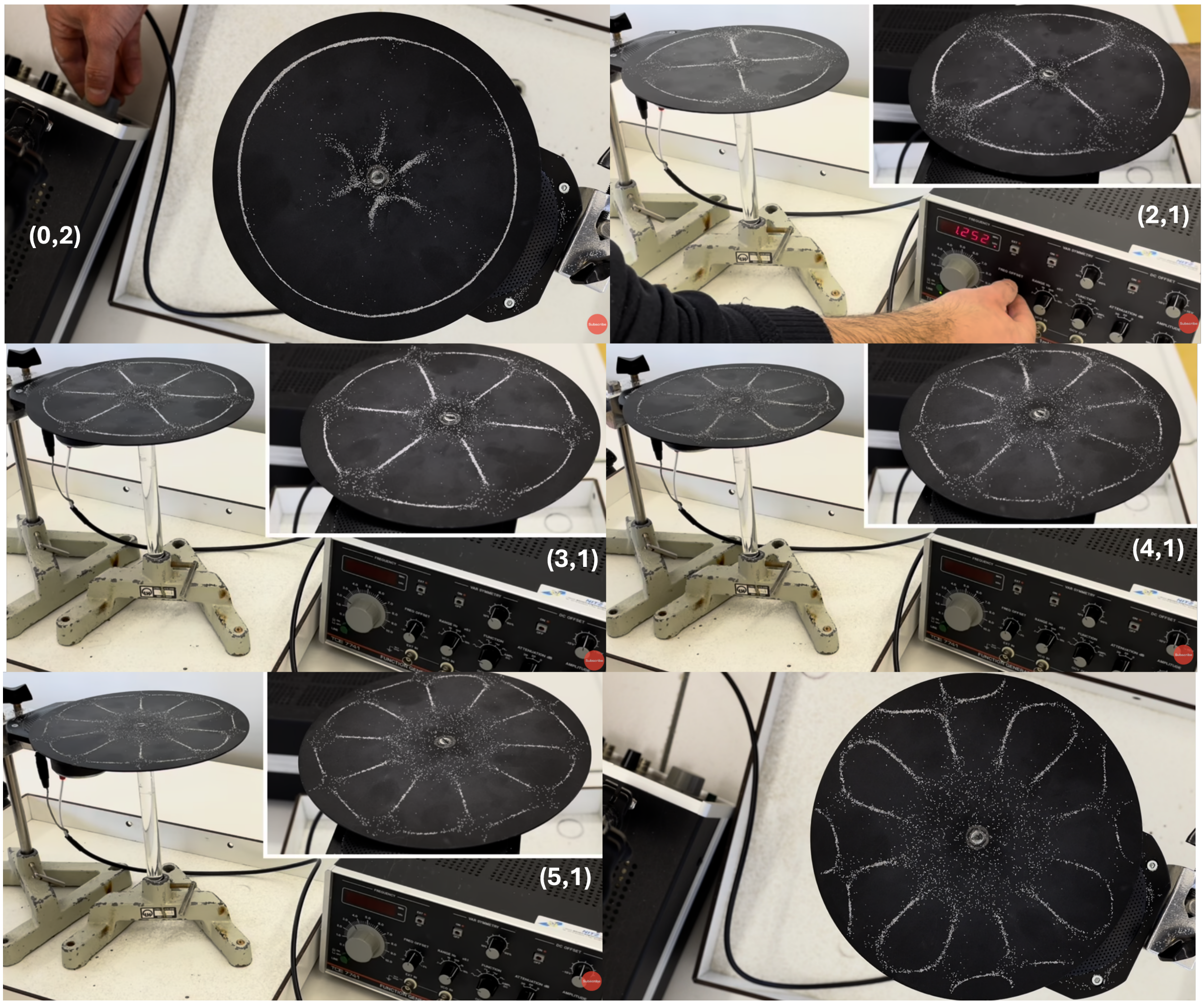


Figure 9.121: **Circular Chladni plate (source: YouTube, author: Labtastic)**

9.4.6 Phasors

There are several ways of expressing the wave functions that we have seen previously. Physicists (as well as electrotechnicians and engineers) use a formulation, named a "phasor" or "Fresnel representation", which simplify of the writing of equation and thus greatly simplifies the study of complex (or simple) problems. Phasors use the properties of complex numbers to express trigonometric wave functions in a simplified form in all phenomena where oscillations occur.

What we name "phasor" is a function f whose value is complex and which, in a space with 1 dimension, is written:

$$f(x,t)=e^{\mathrm{i}(kx-\omega t)} \tag{9.4.176}$$

In all physics applications, t is the time variable.

Since this function is in $\mathbb{C}$, it has a real part that we will denote by g and an imaginary part that we will denote by h. Their identification is easy since, as we have already proved it in our study of complex numbers (see section Numbers page 243):

$$f(x,t)=\cos(kx-\omega t)+\mathrm{i}\sin(kx-\omega t) \tag{9.4.177}$$

Therefore the real and imaginary part are obviously given by:

$$\begin{aligned}g(x,t)&=\Re(f)=\cos(kx-\omega t)\\ h(x,t)&=\Im(f)=\sin(kx-\omega t)\end{aligned} \tag{9.4.178}$$

The module of f is easily computed by calculating:

$$|f|=f\cdot\bar{f}=\cos^2(kx-\omega t)+\sin^2(kx-\omega t)=1 \tag{9.4.179}$$

The real and imaginary parts vary when the position or time varies. The modulus does not change, it is always equal to 1. The change is manifested simply by the simple variation of angle that makes the vector representing f in the complex plane. This is a sufficient reason to speak of a phaser, since the variation of f can be visualized as a simple change of angle or phase.

If we are in a physical space with more than one dimension, say 3, then the expression for f becomes:

$$\boxed{f(\vec{r},t)=e^{\mathrm{i}(\vec{k}\circ\vec{r}-\omega t)}=\cos(\vec{k}\circ\vec{r}-\omega t)+\mathrm{i}\sin(\vec{k}\circ\vec{r}-\omega t)} \tag{9.4.180}$$

The situation is a little more difficult to visualize (...). It is the same as in one dimension, but here everything happens along a direction defined in 3D space by the wave vector $\vec{k}$. Specifically, we will have:

$$\begin{aligned}\cos(\vec{k}\circ\vec{r}-\omega t)&=\cos(k_xx+k_yy+k_zz-\omega t)\\ \sin(\vec{k}\circ\vec{r}-\omega t)&=\sin(k_xx+k_yy+k_zz-\omega t)\end{aligned} \tag{9.4.181}$$

The variation in time remains the same as in the one spatial dimension case but a displacement in space is a little more complicated than in the one dimension case. Here, any spatial displacement in a direction that is not orthonormal to $\vec{k}$ will cause the scalar product to change so that the argument f will vary. Here it is not only the magnitude or the norm of the wave vector $\vec{k}$ which gives the rate of variation of f under a spatial displacement but also the angle that this displacement makes with respect to the direction of $\vec{k}$ since we have an argument that varies as $\vec{k}\circ\vec{r}$ and therefore depends on this angle denoted here φ. Indeed, we have:

$$\vec{k}\circ\vec{r}=\|\vec{k}\|\cdot\|\vec{r}\|\cos(\varphi) \qquad (9.4.182)$$

The quantity $\vec{k}$ is often referred to as the "wave vector". Physically it is most often connected to the equivalent of the (linear) momentum of the wave for which it is obvious that the usual definition of the momentum ($p=mv$) no longer makes sense.

An important part of the systems studied in physics can not be characterized by a point and therefore described by a trajectory. A wave, an elastic band that oscillate do not have a single defined position, they are "continuous medium" over a certain interval. The question we ask ourselves in our attempt to describe them is rather how to describe the displacement of this medium in space and time. For example, for a wave, if we freeze time, how does the amplitude A of this wave vary from one place to another in space? We can also freeze space by looking at one place and ask how the amplitude varies with time?

The space coordinates and time now play a similar role as independent parameters. We measure the amplitude of the phenomenon at any time and in any place. We will therefore seek to obtain an expression of the type:

$$A=A(x,t) \qquad \text{or} \qquad A=A(\vec{r},t) \qquad (9.4.183)$$

in one or three dimensions.

The important point is that we seek to express A, whose correct name is a "field", as a function of time and space coordinates. For example, if the wave is very regular perhaps is it adequately described by $A\sim\cos(\omega t)$ if we look at a single place and by $A\sim\cos(kx)$ if we are able to fix the time?

The quantity ω will characterize the frequency of the variation of the same phenomenon. The name "angular frequency" is easy to understand since the cosine or sine function cycles if its argument changes of 2π over a time of one period $\Delta t=T$.

Let us recall that:

$$T\omega=2\pi\Rightarrow\omega=\frac{2\pi}{T}=2\pi\nu \qquad (9.4.184)$$

In the description of harmonic systems (in Wave Mechanics, Quantum Physics or Quantum Chemistry), the phasor notation can be very useful as we have already said. The equation the most often encountered is the wave equation (which we proved at the beginning of this section). In one dimension, it is written for recall as:

$$\frac{\partial^2}{\partial x^2}A(x,t)-\frac{1}{v}\frac{\partial^2}{\partial t^2}A(x,t)=0 \qquad (9.4.185)$$

We check by simple substitution that some particular solutions are of the type:

$$A(x,t)=A_0\sin(kx-\omega t) \qquad \text{or} \qquad A(x,t)=A_0\cos(kx-\omega t) \qquad (9.4.186)$$

or a linear combination of which the most general form is:

$$A(x,t)=A_0\cos(kx-\omega t)+A_0\sin(kx-\omega t) \tag{9.4.187}$$

A quick and effective way of writing all these relations in a condensed and usefully pictorial way is to write:

$$\boxed{A(x,t)=A_0e^{\mathrm{i}(kx-\omega t)}} \tag{9.4.188}$$

In all three cases, the substitution makes it possible to verify it (the differential equation). Let us take, for example, the sinus solution. Therefore:

$$\begin{aligned}&A(x,t)=A_0\sin(kx-\omega t)\\&\Rightarrow\frac{\partial A}{\partial x}=kA_0\cos(kx-\omega t)\Rightarrow\frac{\partial^2 A}{\partial x^2}=-k^2A_0\sin(kx-\omega t)\\&\Rightarrow\frac{\partial A}{\partial t}=-\omega A_0\cos(kx-\omega t)\Rightarrow\frac{\partial^2 A}{\partial t^2}=-\omega^2A_0\sin(kx-\omega t)\end{aligned} \tag{9.4.189}$$

The replacement in the equation gives:

$$\left(-k^2+\frac{\omega^2}{\nu^2}\right)A_0\sin(kx-\omega t)=0 \tag{9.4.190}$$

Which will be satisfied if and only if:

$$\boxed{k^2=\frac{\omega^2}{v^2}\Leftrightarrow\omega^2=k^2v^2} \tag{9.4.191}$$

The substitution of the phasor as a solution of the wave equation transforms this last differential equation into a simple algebraic equation $\omega=\omega(k)$ named "dispersion relation". This relation is obviously characteristic of the equation that generates it. That which appears above is particularly simple and characterizes a free wave in a non-dispersive medium, as described by the wave equation that we have written.

The phasor solution:

$$A(x,t)=A_0e^{\mathrm{i}(kx-\omega t)} \tag{9.4.192}$$

therefore also satisfies the wave equation, with the same dispersion relation. Is it then also the description of a wave? The answer is YES, and even twice rather than one, as we will see below.

A physical wave is obviously not complex (at least in Classical Mechanics). The phasor solution is complex and therefore has a real part and an imaginary part. We show here that each of the two parts can represent a general real wave. Let us use a more general starting point. Let us assume that the amplitude A_0 is itself a complex number! We can therefore write:

$$A_0=|A_0|e^{\mathrm{i}\delta} \tag{9.4.193}$$

Therefore we have:

$$A(x,t)=A_0e^{\mathrm{i}(kx-\omega t+\delta)}=A_0(\cos(kx-\omega t+\delta)+\mathrm{i}\sin(kx-\omega t+\delta)) \tag{9.4.194}$$

We first study the real part of this expression and using trigonometric identities (see section Trigonometry page 2519):

$$\begin{aligned}\Re(A(x,t))=A_0\cos(kx-\omega t+\delta)&=A_0[\cos(kx-\omega t)\cos(\delta)-\sin(kx-\omega t)\sin(\delta)]\\&=A_0\cos\delta\cos(kx-\omega t)-A_0\sin\delta\sin(kx-\omega t)\end{aligned}\tag{9.4.195}$$

This real part is therefore of the most general (and real) form of the monochromatic solution for the wave equation, namely:

$$A(x,t)=A_1\cos(kx-\omega t)+A_2\sin(kx-\omega t)\tag{9.4.196}$$

Clearly, it is sufficient to identify:

$$\left.\begin{aligned}A_0\cos(\delta)&=A_1\\-A_0\sin(\delta)&=A_2\end{aligned}\right\}=\begin{cases}A_0=\sqrt{A_1^2+A_2^2}\\ \delta=-\arctan(A_2/A_1)\end{cases}\tag{9.4.197}$$

which connect two parameters to two others. The real part of the phasor is therefore sufficient to fully describe the monochromatic wave.

We can do exactly the same thing with the imaginary part of the phasor and prove in an identical way that it is sufficient to describe the monochromatic wave entirely.

The conclusions is that it is therefore possible to use the phasor to do all the mathematical manipulations required by the physical problem and in the end, to keep only the real or imaginary part, according to what was agreed from the beginning.

As we have already said, the most simple general real form of the solution is:

$$A(x,t)=A_1\sin(kx-\omega t)+A_2\cos(kx-\omega t)\tag{9.4.198}$$

This function has the behaviour of a sine (or a cosine) whose amplitude is given by:

$$A_0^2=A_1^2+A_2^2=(A_0\cos(\delta))^2+(-A_0\sin(\delta))^2\tag{9.4.199}$$

Moreover, the initial conditions adjust A_2 such that, at $x=0$ and $t=0$ (initially), the field $A(x,t)$ has the value A_2 therefore. These two conditions set these two parameters. We could have used the equally general form:

$$A(x,t)=A_0\sin(kx-\omega t+\delta)\tag{9.4.200}$$

Here the known amplitude is A_0 and the initial conditions are such that at $x=0$ and $t=0$, the field has the value:

$$A(x,t)=A_0\sin(\delta)\tag{9.4.201}$$

Again, two conditions set two parameters!

The amplitude is an obvious thing, the phase a little less. For a function of the type:

$$A(x,t)=A_0\sin(\arg+\delta)\tag{9.4.202}$$

has any value that one wants when its arg (argument) is, say zero, it suffices to adjust the value of the phase δ. It is like dragging the sinusoidal function along the x-axis, so as to satisfy the initial physical conditions imposed by the system studied.

Of two functions of the sine or cosine type that do not start at the same point in their cycle, we say that they are "out of phase". This becomes vital when there is more than one wave present. Let us imagine the simplest case of two waves of the same amplitude:

$$\begin{aligned}A_1(x,t)&=A_0\sin(\theta+\delta_1)\\A_2(x,t)&=A_0\sin(\theta+\delta_2)\end{aligned}\qquad(9.4.203)$$

Here the argument (θ) is a variable and the phases δ_1 and δ_2 the constant parameters. We consider the physical result of the wave resulting from the addition of these two waves.

If $\delta_2-\delta_1=2k\pi$ for $k\in\mathbb{Z}$ the resulting wave will be a sinusoidal wave of amplitude 2 times A_0. However, if $\delta_2-\delta_1=k\pi\;\;k\in\mathbb{Z}$ the resulting wave will be identically zero everywhere. The difference is therefore considerable and we can produce all intermediate situations! It is therefore important to keep in mind the phase at the origin of the wave or better, its relative phase in relation to other waves of our physical system!

In the phasor, either the real part or the imaginary part, is sufficient to give a general description of the wave (always monochromatic up to now). They are respectively composed of a cosine and a sinus and therefore phase-shifted of $\pi/2$ with respect to each other!

It is desirable to return to the "dispersion relation" that we had obtained. We have seen that for the free monochromatic wave, that which is a solution of the homogeneous wave equation, this relation is written so that the phase of the solution corresponds to this reality:

$$\omega^2=k^2v^2\Rightarrow\omega=\omega(k)\Leftrightarrow k=k(\omega)\qquad(9.4.204)$$

Still in the free case, we can have a physical situation that corresponds to the addition of several free monochromatic waves. The result is not monochromatic and is obviously written as a sum:

$$A(x,t)=\sum_n A_n(x,t)=\sum_n A_n e^{\mathrm{i}(k_nx-\omega_nt+\delta_n)}\qquad(9.4.205)$$

We often name it "wave packet" for obvious reasons (we will use this concept again in the section of Wave Quantum Physics). Since everything is free, each component will satisfy:

$$\omega_n^2=k_n^2v_n^2\qquad(9.4.206)$$

In a non-dispersive wave medium, waves can propagate without deformation. Electromagnetic waves in unbounded free space are non-dispersive as well as non-dissipative and thus can propagate over astronomical distances. Sound waves in air are also nearly non-dispersive even in the ultrasonic frequency range. If not, that is, if high frequency notes (e.g., piccolo) and low frequency notes (e.g., base) propagate at different velocities, they would reach our ears at different times, and music played by an orchestra would not be harmonious. Most waves in material media are dispersive, however, and wave forms originally set up are bound to change in a manner that the wave energy is more spatially spread out or dispersed.

In practice we consider four situation (we give also the Maple™ 4.00b code that permits the user to see the animation by themselves as it is very important for a good a deep comprehension):

- Dispersive wavepacket where the waves in the packet oscillates and the packet becomes larger and larger (we strongly recommend to the reader to change the maximum values of `x` and `t`):

  ```
  >with(plots):
  >animate(sum(.07*(exp(-(.1*k-3)^2)+exp(-(0.1*k+3)^2))*cos(.1*k*x-.1*k/
  sqrt(1+.1*(.1*k)^2)*t),k=1..100),x=-4..60,t=0..100,frames=60,numpoints=200,
  color=red);
  ```

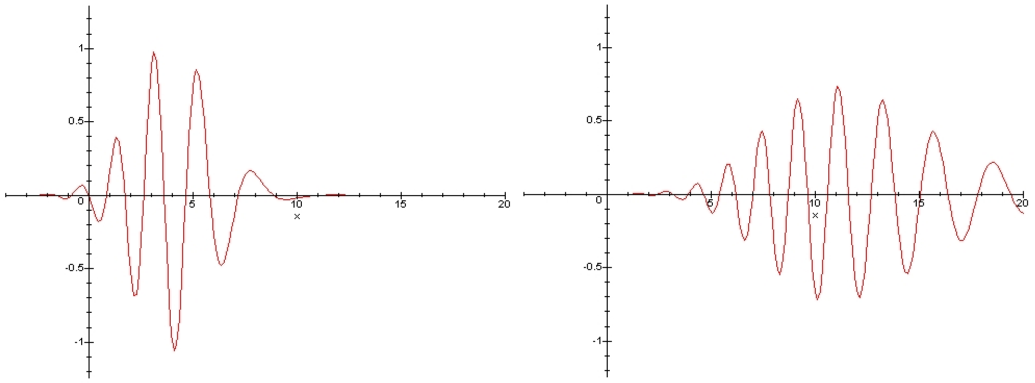


Figure 9.122: **Dispersive Wavepacket in Maple™ 4.00b**

- Non-dispersive wavepacket where the waves in the packet don't oscillates and the packet conserve the same width:

  ```
  >with(plots):
  >animate(sum(.07*(exp(-(.1*k-3)^2)+exp(-(0.1*k+3)^2))*cos(.1*k*x-.1*k/
  sqrt(1+.0*(.1*k)^2)*t),k=1..100),x=-4..20,t=0..30,frames=60,numpoints=200,
  color=red);
  ```

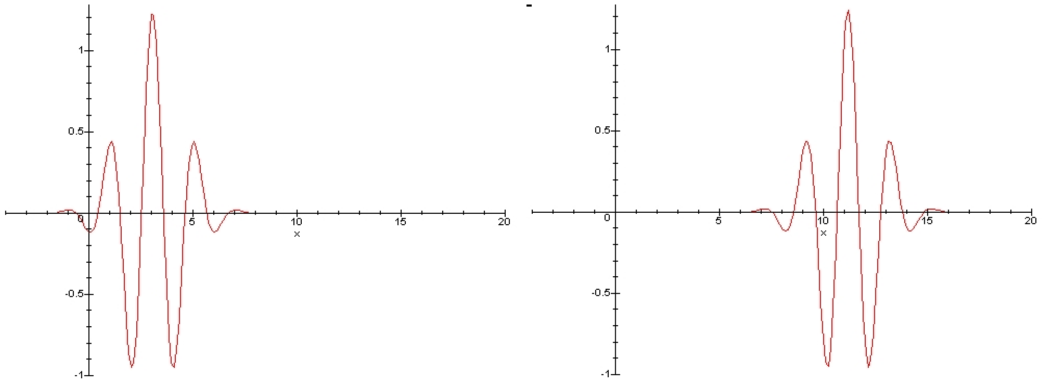


Figure 9.123: **Non-dispersive Wavepacket in Maple™ 4.00b**

- Dispersive sinusoidal wave where we have multiple wavepackets that oscillates in each of the wavepackets and their width doesn't change:

  ```
  >with(plots):
  >animate(sin(x-t)+sin(1.2*x-1.1*t),x=0..50,t=0..63,numpoints=150,
  frames=100,color=red);
  ```

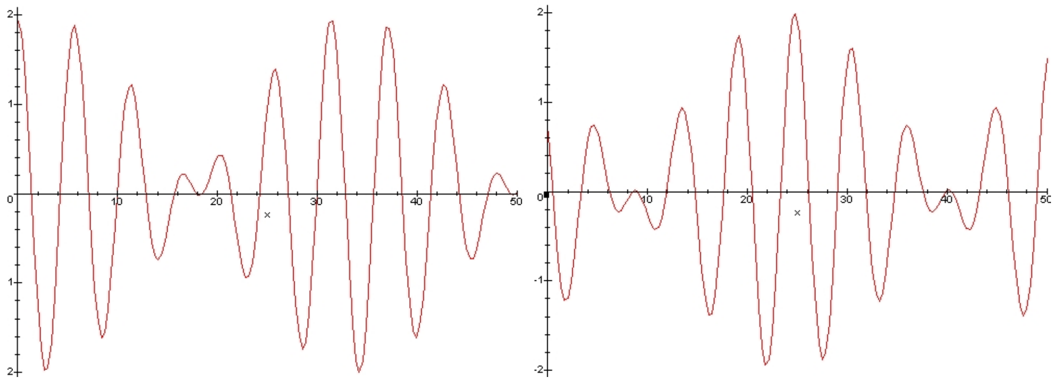


Figure 9.124: **Dispersive sinusoidal waves with Maple™ 4.00b**

- Non-dispersive sinusoidal wave where we have multiple wavepackets where the waves in the packet don't oscillates and their width doesn't change:

  ```
  >with(plots):
  >animate(sin(x-t)+sin(1.2*x-1.2*t),x=0..50,t=0..63,numpoints=150,
  frames=100,color=red);
  ```

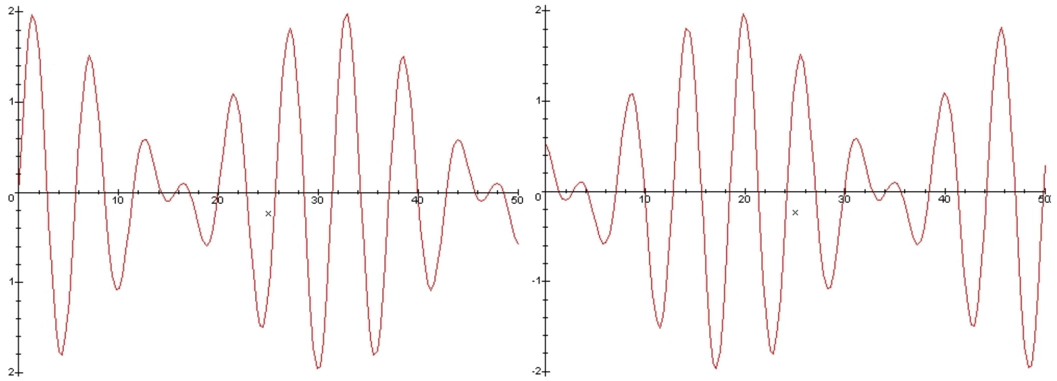


Figure 9.125: **Non-dispersive sinusoidal waves with Maple™ 4.00b**

Here we will notice two things for the free-wave equation:

1. First, even for free waves, it is quadratic, so we can change the sign of the wave vector and / or the pulsation without affecting the equation. We observe trivially that for one or more components of the positive wave vector the wave propagates towards the x-crescents for these positive components and that conversely for one or more negative components the wave propagates towards the x decreasing. Similarly, a negative or positive pulsation means that time varies towards the past or the future.

2. On the other hand, certain types of waves do not obey an equation as simple as the homogeneous above. This is the case for water waves, for example, both because of the nature of the liquid in which they propagate and because of the gravitational recall force. Sometimes, too, a wave that would be totally free, or nearly so, seeks to propagate itself in an environment where the conditions of propagation are seriously affected. For example a sound wave that tries to spread in the mastic or an electromagnetic wave that seeks to propagate in a conductor (a metal). In this case, a large part of the difference between free wave and modified wave by the medium can be described by a change of the dispersion.

9.4.7 Superposition of periodic waves

The superposition of periodic waves is a very common case in electrodynamics, electronics, acoustics and marine engineering. In order not to repeat this study in each of the previously mentioned corresponding sections of this textbook, it seemed more appropriate to us to put all information needed here below.

We will therefore focus on a particular case in order to see the intellectual approach. However, with computers, these calculations by hand have become somewhat obsolete. They must nevertheless be part of the general culture of the engineer or physicist.

Let us first consider two periodic (harmonic) signals generated at the same point of the type:

$$y_1 = A_1 \sin(\phi_1 - \omega_1 t) \quad \text{and} \quad y_2 = A_2 \sin(\phi_2 - \omega_2 t) \tag{9.4.207}$$

For pedagogical reasons, we noticed that it was better to work without phasers. So we will try to make a minimal use of them but for the beginning it will be indispensable. Indeed, the two waves and their sum can be represented in the Gauss plane (see section Numbers page 244) in the following form:

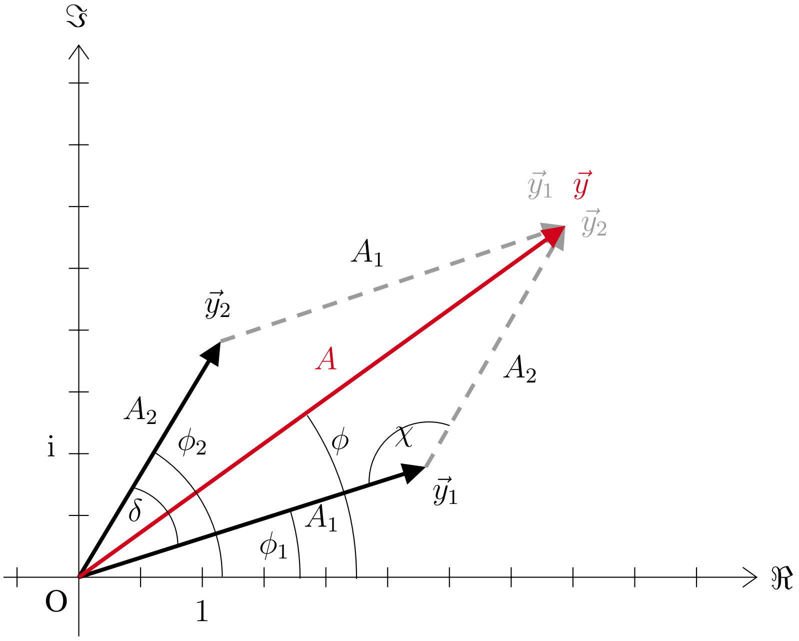


Figure 9.126: **Phase representation of the two harmonics in the Gauss plane**

wither therefore:

$$y = y_1 + y_2 = A_1 \sin(\phi_1 - \omega_1 t) + A_2 \sin(\phi_2 - \omega_2 t) \tag{9.4.208}$$

The amplitude of the sum of the two signals is easily obtained by applying the cosine theorem (see section Trigonometry page 2524). So let us first observe that since we are dealing with a parallelogram, we have (see section Geometric Shapes page 2908):

$$\chi = \pi - \delta \tag{9.4.209}$$

Therefore it comes:

$$A^2 = A_1^2 + A_2^2 - 2A_1A_2\cos(\pi - \delta) = A_1^2 + A_2^2 + 2A_1A_2\cos(\delta) \qquad (9.4.210)$$

hence:

$$A = \sqrt{A_1^2 + A_2^2 + 2A_1A_2\cos(\delta)} \qquad (9.4.211)$$

The term $2A_1A_2\cos(\delta)$ is named the "correlation term". As the phasors are rotating, we have in reality:

$$\delta = (\phi_2 - \omega_2 t) - (\phi_1 - \omega_1 t) = \phi_2 - \phi_1 - (\omega_2 - \omega_1)t \qquad (9.4.212)$$

The phase ϕ of the sum of the waves is obtained simply by doing in the elementary trigonometry in the previous figure:

$$\tan\phi = \frac{A\sin\phi}{A\cos\phi} = \frac{A_1\sin(\phi_1 - \omega_1 t) + A_2\sin(\phi_2 - \omega_2 t)}{A_1\cos(\phi_1 - \omega_1 t) + A_2\cos(\phi_2 - \omega_2 t)} \qquad (9.4.213)$$

Now, let us consider the particular case where the two waves are in phase:

$$\phi_1 = \phi_2 = \pi \qquad (9.4.214)$$

It then comes (see section Trigonometry page 2513):

$$y_1 = A_1\sin(\phi_1 - \omega_1 t) = A_1\sin(\omega_1 t) \quad \text{and} \quad y_2 = A_2\sin(\phi_2 - \omega_2 t) = A_2\sin(\omega_2 t) \qquad (9.4.215)$$

and the amplitude is then written:

$$\begin{aligned} A &= \sqrt{A_1^2 + A_2^2 + 2A_1A_2\cos(\phi_2 - \phi_1 - (\omega_2 - \omega_1)t)} \\ &= \sqrt{A_1^2 + A_2^2 + 2A_1A_2\cos(-(\omega_2 - \omega_1)t)} = \sqrt{A_1^2 + A_2^2 + 2A_1A_2\cos((\omega_2 - \omega_1)t)} \\ &= \sqrt{A_1^2 + A_2^2 + 2A_1A_2\cos(\Delta\omega t)} \end{aligned} \qquad (9.4.216)$$

So the amplitude will have two extremums:

$$\begin{aligned} A_{min} &= \sqrt{A_1^2 + A_2^2 - 2A_1A_2} = \sqrt{(A_1 - A_2)^2} = |A_1 - A_2| \\ A_{max} &= \sqrt{A_1^2 + A_2^2 + 2A_1A_2} = \sqrt{(A_1 + A_2)^2} = A_1 + A_2 \end{aligned} \qquad (9.4.217)$$

Hence:

$$|A_1 - A_2| \leq A \leq A_1 + A_2 \qquad (9.4.218)$$

We then speak in the field of electrodynamics, and of acoustics of "amplitude modulation". The "modulation frequency", also named "beat frequency" is then:

$$f_m = \frac{|\omega_1 - \omega_2|}{2\pi} = |f_2 - f_1| \qquad (9.4.219)$$

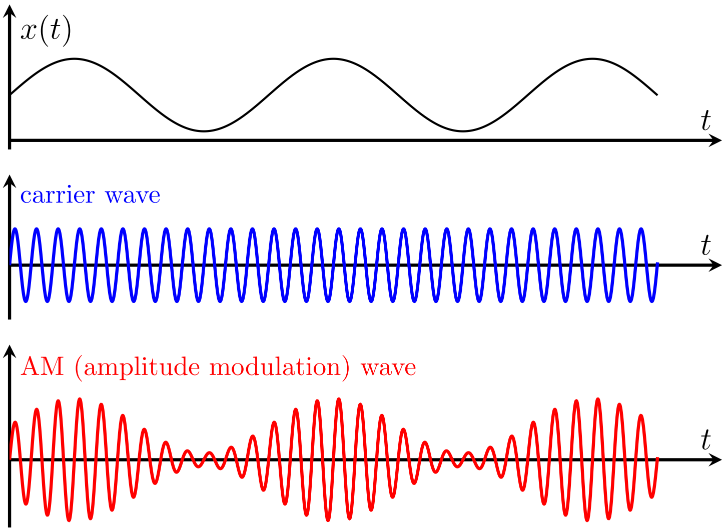

$x(t)$
t
carrier wave
t
AM (amplitude modulation) wave
t

Figure 9.127: **Amplitude modulation (AM) illustrated**

Let us also see an example with Maple™ 4.00b (superimposition of a Bemol Si and of a Do with an intensity ratio of 0.8):

```
>A:=1;B:=0.8;f1:=466.16;f2:=523;w1:=2*Pi*f1;w2:=2*Pi*f2;
>plot(A*sin(w1*t)+B*sin(w2*t),t=0..0.05);
```

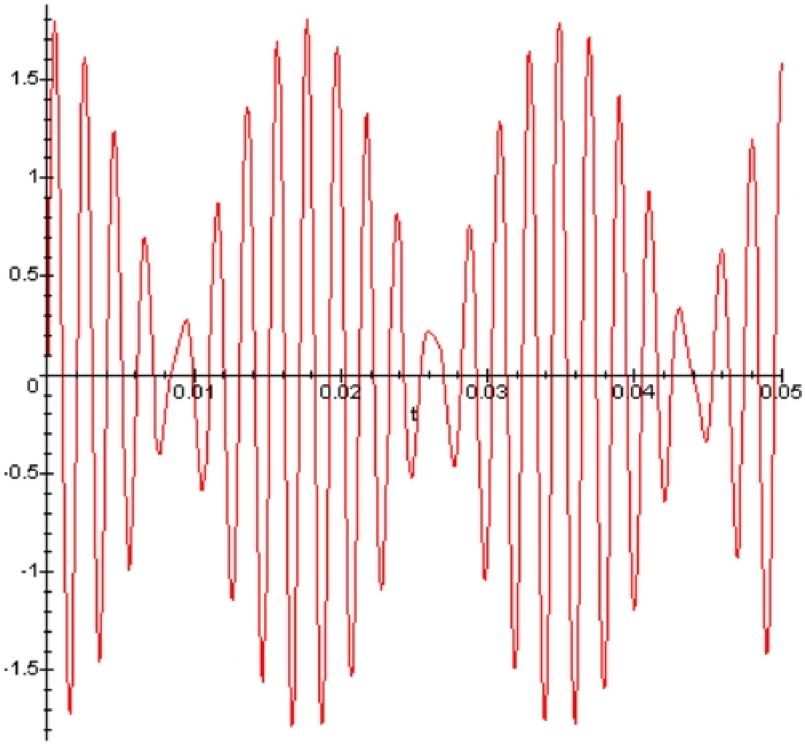


Figure 9.128: **Representation of the amplitude modulation with Maple™ 4.00b**

And if we add the envelope:

```
>plot([A*sin(w1*t)+B*sin(w2*t),sqrt(A^2+B^2+2*A*B*cos((w2-w1)*t))],t=0..0.05);
```

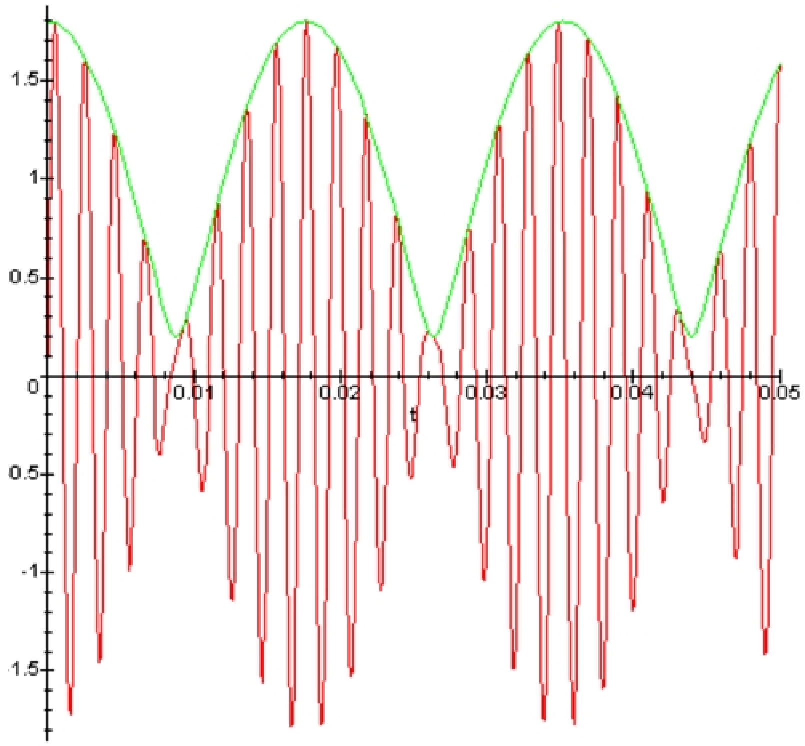


Figure 9.129: **Representation of the envelope of the amplitude modulation with Maple™ 4.00b**

In the general case, it is difficult to simplify the expression of the resulting signal:

$$y=y_1+y_2=A_1\sin(\omega_1 t)+A_2\sin(\omega_2 t) \tag{9.4.220}$$

If the amplitudes are identical $A_2=A_1$ then we have:

$$y=A_1(\sin(\omega_1 t)+\sin(\omega_2 t)) \tag{9.4.221}$$

And thus using one of the Simpson formulas proved in the section of Trigonometry, it comes:

$$y=\underbrace{2A_1\cos\left(\left(\frac{\omega_1-\omega_2}{2}\right)t\right)}_{:=A}\sin\left(\left(\frac{\omega_1+\omega_2}{2}\right)t\right)=A\sin\left(\left(\frac{\omega_1+\omega_2}{2}\right)t\right) \tag{9.4.222}$$

Moreover, if the two frequencies are close, we have:

$$\begin{aligned}y&=2A_1\cos\left(\left(\frac{\omega_1-\omega_2}{2}\right)t\right)\sin\left(\left(\frac{\omega_1+\omega_2}{2}\right)t\right)=2A_1\cos(0)\sin\left(\left(\frac{2\omega_1}{2}\right)t\right)\\&=2A_1\sin(\omega_1 t)\end{aligned} \tag{9.4.223}$$

Now, let consider the case where the waves advance (because until now they were stationary):

$$y=y_1+y_2=A_1\sin(k_1 r-\omega_1 t)+A_2\sin(k_2 r-\omega_2 t) \tag{9.4.224}$$

If the two amplitudes are identical, we have:

$$y=A_1(\sin(k_1 r-\omega_1 t)+\sin(k_2 r-\omega_2 t)) \tag{9.4.225}$$

Using Simpson's formula again, it comes:

$$\begin{aligned}y&=2A_1\cos\left(\frac{(k_1r-\omega_1t)-(k_2r-\omega_2t)}{2}\right)\sin\left(\frac{(k_1r-\omega_1t)+(k_2r-\omega_2t)}{2}t\right)\\&=2A_1\cos\left(-\frac{(k_1r-\omega_1t)-(k_2r-\omega_2t)}{2}\right)\sin\left(\frac{(k_1r-\omega_1t)+(k_2r-\omega_2t)}{2}t\right)\\&=2A_1\cos\left(\left(\frac{k_2-k_1}{2}\right)r-\left(\frac{\omega_2-\omega_1}{2}\right)t\right)\sin\left(\left(\frac{k_2+k_1}{2}\right)r-\left(\frac{\omega_2+\omega_1}{2}\right)t\right)\\&=2A_1\cos\left(\frac{\Delta k}{2}r-\frac{\Delta\omega}{2}t\right)\sin\left(\left(\frac{k_2+k_1}{2}\right)r-\left(\frac{\omega_2+\omega_1}{2}\right)t\right)\end{aligned}\tag{9.4.226}$$

Therefore the amplitude is modulated according to the term:

$$2A_1\cos\left(\frac{\Delta k}{2}r-\frac{\Delta\omega}{2}t\right)\tag{9.4.227}$$

The amplitude is thus described by a wave function. The corresponding wave represents the progression of a carrier wave packet, progressing at the speed:

$$v=\frac{\Delta\omega}{\Delta k}\tag{9.4.228}$$

We also notice that if we have the two progressive waves of direction of opposite propagation such as:

$$k_2=-k_1\tag{9.4.229}$$

then:

$$\begin{aligned}y&=2A_1\cos\left(\frac{\Delta k}{2}r-\frac{\Delta\omega}{2}t\right)\sin\left(\left(\frac{k_2+k_1}{2}\right)r-\left(\frac{\omega_2+\omega_1}{2}\right)t\right)\\&=2A_1\cos\left(k_1r-\frac{\Delta\omega}{2}t\right)\sin\left(\left(\frac{\omega_2+\omega_1}{2}\right)t\right)\end{aligned}\tag{9.4.230}$$

If in addition the two harmonics are synchronous (of the same frequency), then we have:

$$y=2A_1\cos\left(k_1r-\frac{\Delta\omega}{2}t\right)\sin\left(\left(\frac{\omega_2+\omega_1}{2}\right)t\right)=2A_1\cos(k_1r)\sin(\omega_1t)\tag{9.4.231}$$

It is therefore a stationary wave but whose amplitude is modulated as a function of time and position. Indeed the reader will notice that at several points r the wave is always zero.

Now let us consider a system of loudspeakers (TV, stereo, etc.) that emit acoustic waves synchronously with the same amplitude and frequency. Since the source of the two harmonics is not reduced to a single point, we have in function of our distance r_1 the first loudspeaker and the of the second r_2:

$$y=A(\sin(kr_1-\omega t)+\sin(kr_2-\omega t))\tag{9.4.232}$$

Applying Simpson's formula again, it comes:

$$\begin{aligned}y&=2A\cos\left(\frac{(kr_1-\omega t)-(kr_2-\omega t)}{2}\right)\sin\left(\frac{(kr_1-\omega t)+(kr_2-\omega t)}{2}t\right)\\&=2A\cos\left(\frac{k(r_1-r_2)}{2}\right)\sin\left(\frac{k(r_1+r_2)}{2}-\omega t\right)\end{aligned}\tag{9.4.233}$$

If we are equal distance from the two loudspeaker, it comes:

$$y=2A\sin(kr-\omega t) \tag{9.4.234}$$

Thus the amplitude is there maximal ("constructive interference") and is twice the amplitude of each harmonic. In fact, we find maxima of amplitude at each point where:

$$\cos\left(\frac{k(r_1-r_2)}{2}\right)=\pm 1 \tag{9.4.235}$$

thus:

$$\frac{k(r_1-r_2)}{2}=n\pi\Rightarrow r_1-r_2=\frac{2n\pi}{k}=\frac{2n\pi}{\frac{\omega}{v_{\text{phase}}}}=\frac{2n\pi v_{\text{phase}}}{2\pi f}=n\frac{v_{\text{phase}}}{f} \tag{9.4.236}$$

where n is a relative integer (possibly equal to zero!). So, exactly between the two loudspeakers, we find a maxima. For a same n, all the values of r_1, r_2 that satisfy this relation are named "ventral lines".

We have in extenso the amplitude which is zero ("destructive interference") whenever:

$$\cos\left(\frac{k(r_1-r_2)}{2}\right)=0 \tag{9.4.237}$$

Thus:

$$\frac{k(r_1-r_2)}{2}=n\frac{\pi}{2}\Rightarrow r_1-r_2=\frac{n\pi}{k}=\frac{n\pi}{\frac{\omega}{v_{\text{phase}}}}=\frac{n\pi v_{\text{phase}}}{2\pi f}=n\frac{v_{\text{phase}}}{\pi f} \tag{9.4.238}$$

where n is a non-zero relative odd integer. For the same n, all the values r_1, r_2 that satisfy this relation are named "nodal lines".

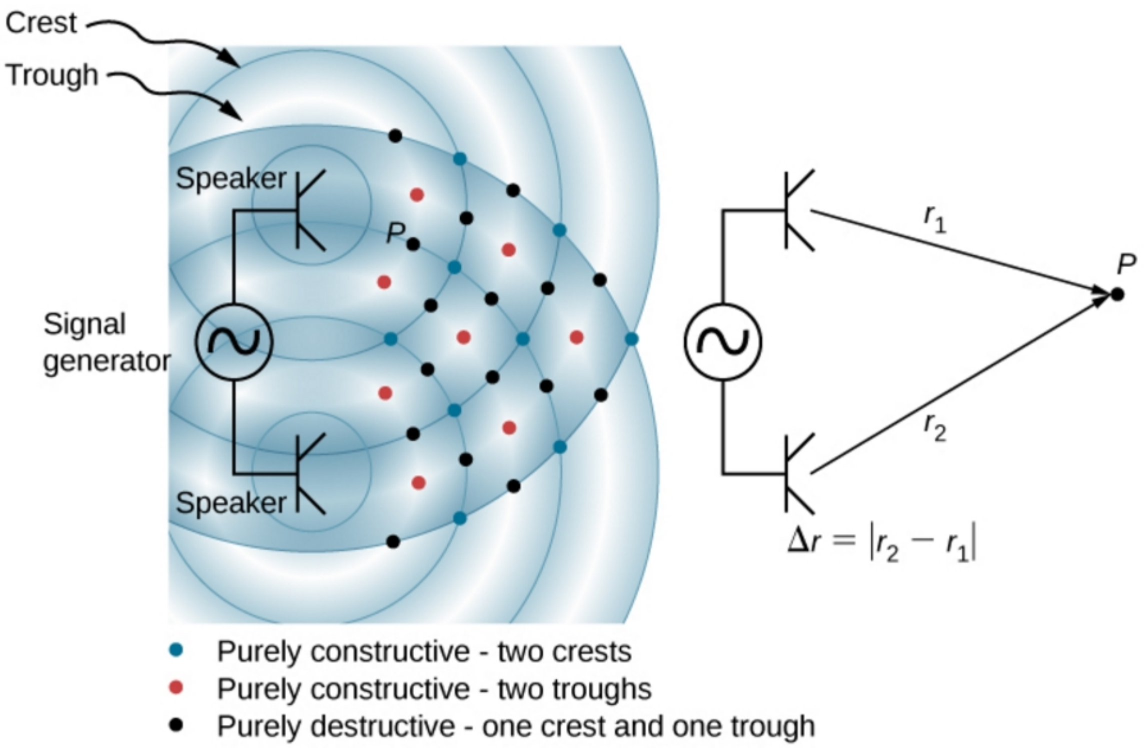

Crest
Trough
Speaker
P
Signal generator
Speaker
r_1
P
r_2
$\Delta r = |r_2 - r_1|$
● Purely constructive - two crests
● Purely constructive - two troughs
● Purely destructive - one crest and one trough

Figure 9.130: Waves interferences of two loudspeakers (source: OpenStax)

There is also a simple elegant way to introduce the superposition principle. Suppose that the wavefunctions ψ_1 and ψ_2 are each separate solutions of the wave equation; it follows that $(\psi_1+\psi_2)$ is also a solution. This is known as the Superposition Principle, and it can easily be proven since it must be true that:

$$\frac{\partial^2\psi_1}{\partial x^2}=\frac{1}{v^2}\frac{\partial^2\psi_1}{\partial t^2} \quad \text{and} \quad \frac{\partial^2\psi_2}{\partial x^2}=\frac{1}{v^2}\frac{\partial^2\psi_2}{\partial t^2} \tag{9.4.239}$$

Adding these yields:

$$\frac{\partial^2\psi_1}{\partial x^2}+\frac{\partial^2\psi_2}{\partial x^2}=\frac{1}{v^2}\frac{\partial^2\psi_1}{\partial t^2}+\frac{1}{v^2}\frac{\partial^2\psi_2}{\partial t^2} \tag{9.4.240}$$

and so:

$$\frac{\partial^2}{\partial x^2}(\psi_1+\psi_2)=\frac{1}{v^2}\frac{\partial^2}{\partial t^2}(\psi_1+\psi_2) \tag{9.4.241}$$

which establishes that $(\psi_1+\psi_2)$ is indeed a solution. What this means is that when two separate waves arrive at the same place in space wherein they overlap, they will simply add to (or subtract from) one another without permanently destroying or disrupting either wave. The resulting disturbance at each point in the region of overlap is the algebraic sum of the individual constituent waves at that location. Once having passed through the region where the two waves coexist, each will move out and away unaffected by the encounter.

9.4.7.1 AM/FM Broadcasting

In telecommunications and signal processing, frequency modulation (FM) is the encoding of information in a carrier wave by varying the instantaneous frequency of the wave. This contrasts with amplitude modulation, in which the amplitude of the carrier wave varies, while the frequency remains constant.

Before a radio station can begin broadcasting, the operator must first determine the frequency that best fits the information it wants to send, and then construct a suitable transmitter and antenna to cover the desired area over which the receivers will be located. Next, the station operator must determine how the information to be broadcast will be superimposed onto the chosen radio frequency. The process of superimposing information onto a radio frequency is named "modulation". There are many different modulation techniques available but the two most popular are "Amplitude Modulation (AM) and Frequency Modulation (FM).

With "Amplitude Modulation", changes in the amplitude of the information signal causes a proportional change in the amplitude of the radio frequency signal (also referred to as a "carrier signal" as we know):

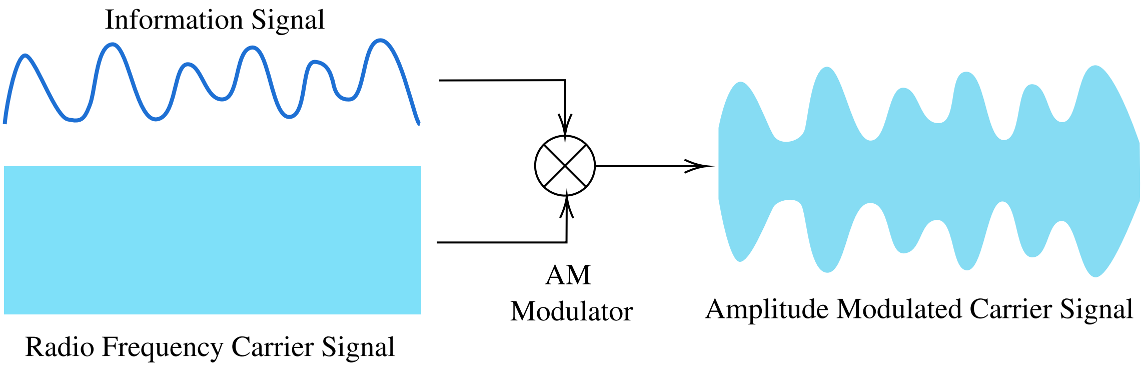

Information Signal
AM
Modulator
Amplitude Modulated Carrier Signal
Radio Frequency Carrier Signal

Figure 9.131: **AM modulation principle (source: netZener)**

With "Frequency Modulation", changes in the amplitude of the information signal causes a proportional shift in the frequency of the "carrier signal". The instantaneous frequency deviation, the difference between the frequency of the carrier and its center frequency, is proportional to the modulating signal.

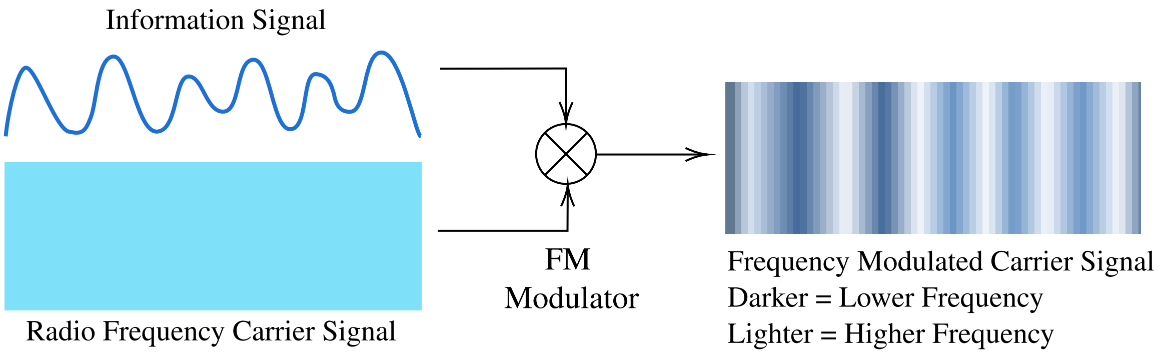

Information Signal
FM
Modulator
Frequency Modulated Carrier Signal
Darker = Lower Frequency
Lighter = Higher Frequency
Radio Frequency Carrier Signal

Figure 9.132: **Waves interferences of two loudspeakers (source: netZener)**

Remark

FM radio is inherently less subject to noise from stray radio sources than AM radio. The reason is that amplitudes of waves add. So an AM receiver would interpret noise added onto the amplitude of its carrier wave as part of the information. A FM receiver can be made to reject amplitudes other than that of the basic carrier wave and only look for variations in frequency. It is thus easier to reject noise from FM, since noise produces a variation in amplitude.

We will study in the section of Electrical Engineering how is build a FM modulator!

~95 %

★★★★☆

21 votes, 69.52%

9.5. Statistical Mechanics

STATISTICAL MECHANICS, also named sometimes "Statistical Thermodynamics" or more generally "Statistical Physics", aims to explain the behaviour of macroscopic systems (consisting of many interacting objects) from their microscopic characteristics using statistical and probabilistic tools. It is in a much more general way that quantum physics describes the properties and evolution of physical systems at the microscopic scale. Statistical Mechanics is therefore built on this quantum description as discussed on mathematical developments that will follow.

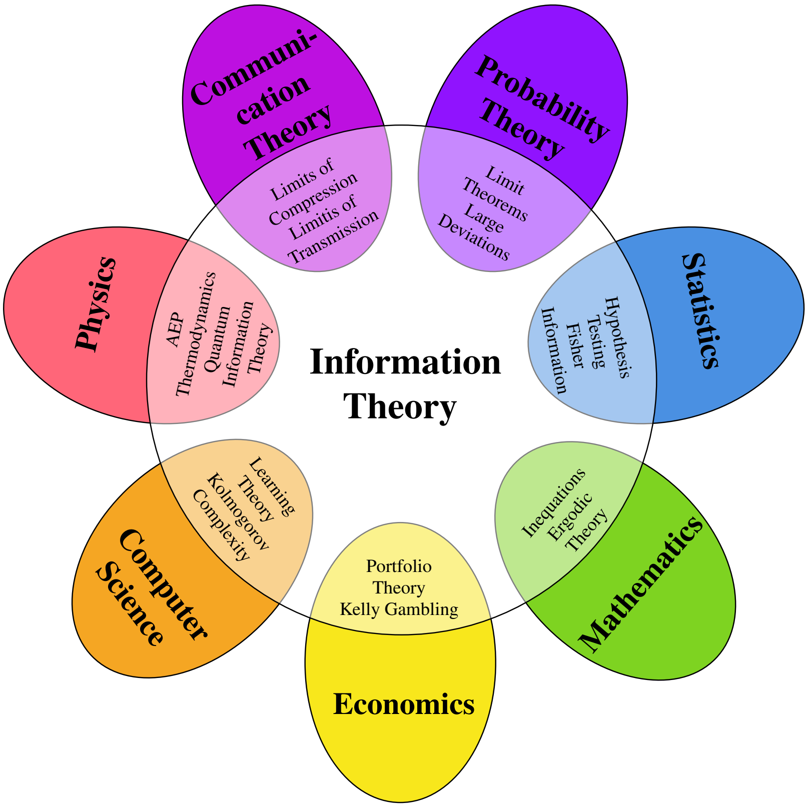
« *Ludwig Boltzmann, who spent much of his life studying statistical Mechanics, died in 11906 (holocene calendar), by his own hand. Paul Ehrenfest, carrying on the work, died similarly in 11933 (holocene calendar). Now it is our turn to study Statistical Mechanics.* » These are the opening sentences of an excellent textbook, *States of Matter* (11975), by David L. Goodstein.

The approach presented here is to address the basic Statistical Mechanics to then deduce the fundamentals of Thermodynamics. Statistical Mechanics is indeed with quantum physics and relativity, a cornerstone of modern physics in explaining phenomena from their constituents! It is important to perceive it immediately as a fundamental theory, and not as a mere attempt to justify retrospectively thermodynamics. Thermodynamics itself gains back more just and deeper understanding of its principles and methods.

9.5.1 Statistical Information Theory

The word "information" is used in very diverse contexts, in totally different sense following scientific disciplines: we can cite as an example thermodynamics with the concept of entropy, applied physics with the theory of signal, biology with the genome theory and quantum physics with the probability of getting information.

Definition (#9.5.453): "Information" is the reduction of uncertainty with a system (formally expressed as Shannon entropy as we will see further below), i.e. the decrease in the number of yes/no choices - measured in bits (binary digits) - required to specify the exact state of a system out of all possible.


Communi-
cation
Theory
Limits of
Compression
Limitis of
Transmission
Probability
Theory
Limit
Theorems
Large
Deviations
Physics
AEP
Thermodynamics
Quantum
Information
Theory
Information
Theory
Statistics
Hypothesis
Testing
Fisher
Information
Learning
Theory
Kolmogorov
Complexity
Computer
Science
Portfolio
Theory
Kelly Gambling
Economics
Inequations
Ergodic
Theory
Mathematics

In physics, information is deeply tied to the state of a system. A new state is equivalent to a new information! Therefore when a star explodes, it creates new atoms hence new information. When particles interact, they produce new states hence new information. When entropy increases, the number of possible microstates grows, hence new information. Similarly inf biological evolution random stochastic mutations introducing variations sometimes create information and sometimes destructs information.

This raises the question whether it is possible to construct a theory of information and if it is unique? Our approach here is not information as such, but the amount of information! When we talk about quantity and measurement, we think to the concept of content or value of information. The information science in purpose must feel concerned by this questioning. If we define the "infometrics" as the set of technical mathematical and statistical measures of information, we wish to have a sufficiently clear definition of the concept of quantity of information that can lead us to define a measure, that is to say a set of well-defined operations, bringing us to clear axioms and whose result is a number.

We are therefore concerned here with the foundations of the statistical information theory also known as "Shannon's theory". The Shannon formula that emerges is certainly one of the fundamental concepts of physics and statistics as it affects all the irreducible brick physics, finance and statistics: information!!

We will show (further below) that an isolated physical system has for most probable state the one that contains the more internal states and is therefore a fortiori the most unpredictable. The most improbable state is the one that is more predictable (concerns about emergence of life...). Therefore, since the unpredictability appears as an essential attribute of the information, we identify the quantitative measurement information to its improbability.

Thus, the amount of information denoted by $h(x)$ given by the realization of an event x of probability $P(x)$ is an increasing function f of its improbability $\dfrac{1}{P(x)}$ following the postulate that the most probable state is the one that is the less predictable:

$$\boxed{h(x)=f\left(\frac{1}{P(x)}\right)} \tag{9.5.1}$$

Moreover, the realization of two independent events x and y intuitively provides an amount of information which is the sum of their respective amounts of information. We have seen in the section of Probabilities that for two independent event we have:

$$P(x,y)=P(x)P(y) \tag{9.5.2}$$

That is to say we are looking for a function f such that :

$$h(x)=f\left(\frac{1}{P(x,y)}\right)=f\left(\frac{1}{P(x)P(y)}\right)=f\left(\frac{1}{P(x)}\right)+f\left(\frac{1}{P(y)}\right)=h(x)+h(y) \tag{9.5.3}$$

The logarithm function (see section Functional Analysis page 2284) is by its properties a natural candidate for f such that:

$$f=\lambda\log_{\alpha}\left(\frac{1}{P(\ldots)}\right) \tag{9.5.4}$$

where λ is obviously a positive number.

Remark

The choice of the base of the logarithm defines the information unit which is completely arbitrary. Thereafter, and unless otherwise specified, "log" will designate the logarithm in base α.

Thus, the "quantity of intrinsic information" of an event x could be therefore using the properties of the logarithm:

$$\boxed{h(x)=-\lambda\log\left(P(x)\right)} \tag{9.5.5}$$

And can be considered as a measure of uncertainty about the event, or as that of the information needed to resolve this uncertainty. Why people decided that this specific measure is what we should use is a good question, and not so obvious. Information theorists try to justify it axiomatically. That is, if we pick this measure, it obeys some nice properties we intuitively want it to obey. Properties like "we want this number of be higher for unlikely events" and "we want this number to be additive for independent events." This is why we get a minus sign and a log

Here is an example that motivates this definition:

✍**Example:**

Agree that knowing the realization of an event that has for probability $1/2$ give us 1 bit of information (think for example to the binary system). We now launch n coins and take knowledge of the result of the experience: we acquire n bits of information. The event in question had the probability $\left(\dfrac{1}{2}\right)^n$. Therefore the number of bits of information is if we chose $\alpha=2$ and $\lambda=1$:

$$h(x)=-\lambda\log\left(P(x)\right)=-\lambda\log_2\left(\left(\frac{1}{2}\right)^n\right)=-n\lambda\log_2\left(\frac{1}{2}\right)=-n\lambda\cdot(-1)=n \quad (9.5.6)$$

This is an example that motivates our choice.

Definitions (#9.5.454):

D1. We define the "intrinsic information by pair" of two events x and y of joint probability $P(x,y)$ (see section Probabilities page 480) by:

$$h(x,y)=-\lambda\log\left(P(x,y)\right) \quad (9.5.7)$$

D2. We define also the "conditional information" of x when y is known by (see section Probabilities page 485):

$$h(x/y)=-\lambda\log\left(P(x/y)\right) \quad (9.5.8)$$

which can therefore also be written (this is just a different way to note that we have already presented in the section of Probabilities):

$$h(x\,|\,y)=-\lambda\log\left(P(x\,|\,y)\right) \quad (9.5.9)$$

This is the amount of information remaining about x once y has been observed.

We have proved in the section of Probabilities that the Bayes' formula allows us to notice immediately that if x and y are independent :

$$h(x/y)=h(x) \quad (9.5.10)$$

This is consistent with common sense.

D3. We also need sometimes to measure the amount of information that a particular variable for example y brings on another for example x. This is particularly the case when we identify x as the selection of a signal applied to the input of a channel and y the corresponding observed output signal (analysis used in code correction error as seen in the section of Error-Correcting code).

Let $P(x)$ the probability that x has been emitted and $P(x/y)$ the a priori probability that x was emitted, given that y was received.

A measurement of this quantity of information, named "mutual information" is denoted by:

$$i(x,y)=\lambda\log\left(\frac{P(x/y)}{P(x)}\right) \quad (9.5.11)$$

It is the logarithmic measure of the increase in the probability of x (thus the decrease in its amount information) due to its conditioning y. We can check if it is coherent with two situations:

(a) If the data of y is equivalent to x (case of a perfect channel), it is equal to the intrinsic information $h(x)$:

$$i(x,y)=\lambda\log\left(\frac{P(x/y)}{P(x)}\right)=\lambda\log\left(\frac{1}{P(x)}\right)=-\lambda\log(P(x)) \tag{9.5.12}$$

(b) It is null if, conversely, x and y are independent.

$$i(x,y)=\lambda\log\left(\frac{P(x/y)}{P(x)}\right)=\lambda\log\left(\frac{P(x)}{P(x)}\right)=-\lambda\log(1)=0 \tag{9.5.13}$$

Obviously we have:

$$\begin{aligned}i(x,y)&=h(x)-h(x/y)=-\lambda\log\left(P(x)\right)-\left[-\lambda\log\left(P(x/y)\right)\right]\\&=\lambda\log\left(P(x/y)\right)-\lambda\log\left(P(x)\right)=\lambda\left(\log\left(P(x/y)\right)-\log\left(P(x)\right)\right)\\&=\lambda\log\left(\frac{P(x/y)}{P(x)}\right)\end{aligned} \tag{9.5.14}$$

In the same way we prove that:

$$i(x,y)=h(y)-h(y/x) \tag{9.5.15}$$

Remark

Notice that in general $h(x/y)\neq h(y/x)$.

And using the formula of compound probabilities (see section Probabilities page 486):

$$\begin{aligned}i(x,y)&=\lambda\log\left(\frac{P(x/y)}{P(x)}\right)=\lambda\log\left(\frac{\dfrac{P(x\cap y)}{P(Y)}}{P(x)}\right)=\lambda\log\left(\frac{P(x\cap y)}{P(x)P(y)}\right)\\&=\lambda\log\left(\frac{P(x,y)}{P(x)P(y)}\right)\end{aligned} \tag{9.5.16}$$

The latter equality justifying the term "mutual". While the intrinsic information is positive, mutual information can be negative. We will see that his average, much more important in practice, can not be negative.

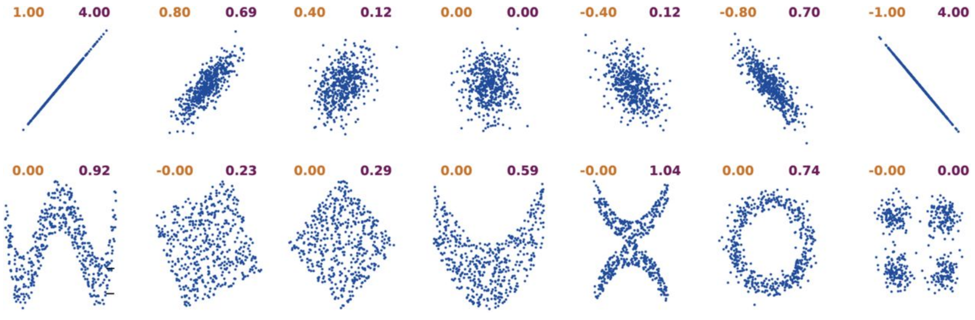


Figure 9.133: Examples of correlation vs mutual information.

> In the figure above, each panel illustrates a scatter plot of samples drawn from a particular bi-variate distribution. For each example, the correlation between the two variables is shown in orange (left) and the MI is shown in purple (right; discrete method, 16 bins, $100,000$ samples, no bias correction). The top row shows linear relationships, for which MI and correlation both detect a relationship (although on different scales, and note that as MI is always positive it does not reveal the direction of the relationship). The bottom row shows a series of distributions for which the correlation is zero.

The individual events are in general less important than the mean effect, we will consider them later as coming from a random source, discrete, finite, stationary, and white (i.e. independent successive events). The events are interpreted as choosing a symbol in the alphabet of the source. Let n be the size of the alphabet, and $x_1,\ldots,x_n$ its symbols. The source is therefore described by the random variable X, which takes values in the alphabet, with respective probabilities $p_1,\ldots p_n$, such as:

$$\sum_{i=1}^{n} P_i = 1 \qquad (9.5.17)$$

Definition (#9.5.455): The average amount of information from this source is the expected mean of the intrinsic information of each symbol of the alphabet of the source (see section Statistics page 634). We will name it "entropy" and denote it by $S(X)$ and given by the relation:

$$\boxed{S(X) = \mathrm{E}(h(x)) = -\lambda\sum_{i=1}^{n} P_i \log(P_i)} \qquad (9.5.18)$$

named "Shannon formula", or "Shannon entropy", and it is indeed a mean because its general expression for a discrete random variable is as we have seen in the section Statistics given by:

$$\mathrm{E}(X) = \sum_{i=1}^{n} P_i x_i \qquad (9.5.19)$$

but with in this case:

$$x_i = -\lambda\log(P_i) \qquad (9.5.20)$$

The reader can also notice that if the n events are of equal probabilities $1/n$ we have the famous relation:

$$S = -\lambda\sum_{i=1}^{n} P_i \log(P_i) = -\lambda\sum_{i=1}^{n}\frac{1}{n}\log\left(\frac{1}{n}\right) = -\lambda n\frac{1}{n}\log\left(\frac{1}{n}\right) = \log(n) \qquad (9.5.21)$$

Remark

For a business real application case of the Shannon formula you can see the subsection on the Diversity Index in the section of Statistics page 644.

However, the relation:

$$S(X) = \mathrm{E}(h(x)) = -\lambda\sum_{i=1}^{n} P_i \log(P_i) \qquad (9.5.22)$$

is an abuse of notation. Indeed, the expectation makes sense if $h(x)$ is a function of x. But in our case $h(x)$ does not depend on the values of x, but only to the associated probabilities. We will therefore sometimes more rigorously denote the entropy of a distribution by:

$$S(P_1,\ldots,P_n) \tag{9.5.23}$$

The "joint and conditional entropies" are defined in similar manner with the suitable notations:

$$S(X,Y)=\mathrm{E}(h(x,y))=-\lambda\sum_{i=1}^{n}\sum_{j=1}^{m}P(x_i,y_j)\log(P(x_i,y_j)) \tag{9.5.24}$$

and (we adopt this time the symbol "|" instead of the "/" to indicate the conditional relation):

$$S(Y|X)=\mathrm{E}(h(y|x))=-\lambda\sum_{i=1}^{n}\sum_{j=1}^{m}P(x_i,y_j)\log(P(y_j,x_i)) \tag{9.5.25}$$

The "mean mutual information", sometimes badly named "mutual information" is defined also directly:

$$I(X,Y)=\mathrm{E}(i(x,y))=\lambda\sum_{i=1}^{n}\sum_{j=1}^{m}P(x_i,y_j)\log\left(\frac{P(x_i|y_j)}{P(x_i)}\right)=\lambda\sum_{i=1}^{n}\sum_{j=1}^{m}P(x_i,y_j)\log\left(\frac{P(x_i,y_j)}{P(x_i)P(y_j)}\right) \tag{9.5.26}$$

Mutual information measures how much information - in the information-theoretic sense à a term an events contains about a class. If a term's distribution is the same in the class as it is en the collection as a whole, then $I(X,Y)=0$. Mutual information reaches its maximum value if the event is a perfect indicator for class membership, that is, if the term is present in a document if and only if the document is in the class.

Remark

The reader must notice that the definition of the amount of information, by a logarithmic measure may seem arbitrary, though reasonable, given the expected properties of such a measure. Shannon, and later Khintchine showed that taking into account certain properties putted as axioms laid the logarithmic function be the only one suitable choice.

✍**Example:**

Consider a classification of $N=801,948$ documents. We build the following contingency table according if they contain respectively the terms *poultry* and *export*:

	***Poultry* detected**	***Poultry* not detected**
***Export* detected**	$N_{11}=49$	$N_{10}=141$
***Export* not detected**	$N_{01}=27,652$	$N_{00}=774,106$

$$\begin{aligned}
I(X,Y)&=\lambda\sum_{i=1}^{n}\sum_{j=1}^{m}P(x_i,y_j)\log\left(\frac{P(x_i|y_j)}{P(x_i)}\right)=\lambda\sum_{i=1}^{n}\sum_{j=1}^{m}P(x_i,y_j)\log\left(\frac{P(x_i,y_j)}{P(x_i)P(y_j)}\right)\\
&=\frac{N_{11}}{N}\log_2\left(\frac{\dfrac{N_{11}}{N}}{\dfrac{N_{1.}}{N}\dfrac{N_{.1}}{N}}\right)+\frac{N_{01}}{N}\log_2\left(\frac{\dfrac{N_{01}}{N}}{\dfrac{N_{0.}}{N}\dfrac{N_{.1}}{N}}\right)\\
&+\frac{N_{10}}{N}\log_2\left(\frac{\dfrac{N_{10}}{N}}{\dfrac{N_{1.}}{N}\dfrac{N_{.0}}{N}}\right)+\frac{N_{00}}{N}\log_2\left(\frac{\dfrac{N_{00}}{N}}{\dfrac{N_{0.}}{N}\dfrac{N_{.0}}{N}}\right)\\
&=\frac{N_{11}}{N}\log_2\left(\frac{NN_{11}}{N_{1.}N_{.1}}\right)+\frac{N_{01}}{N}\log_2\left(\frac{NN_{01}}{N_{0.}N_{.1}}\right)\\
&+\frac{N_{10}}{N}\log_2\left(\frac{NN_{10}}{N_{1.}N_{.0}}\right)+\frac{N_{00}}{N}\log_2\left(\frac{NN_{00}}{N_{0.}N_{.0}}\right)\\
&=\frac{49}{801,948}\log_2\left(\frac{801,948\cdot 49}{(49+27,652)(49+141)}\right)\\
&+\frac{141}{801,948}\log_2\left(\frac{801,948\cdot 141}{(141+774,106)(49+141)}\right)\\
&+\frac{27,652}{801,948}\log_2\left(\frac{801,948\cdot 27,652}{(49+27,652)(27,652+774,106)}\right)\\
&+\frac{774,106}{801,948}\log_2\left(\frac{801,948\cdot 774,106}{(141+774,106)(27,652+774,106)}\right)\cong 0.000105
\end{aligned}\tag{9.5.27}$$

For two variables it is possible to represent the different entropic quantities with an analogy to set theory. In the figure below we see the different quantities, and how the mutual information is the to x and y:

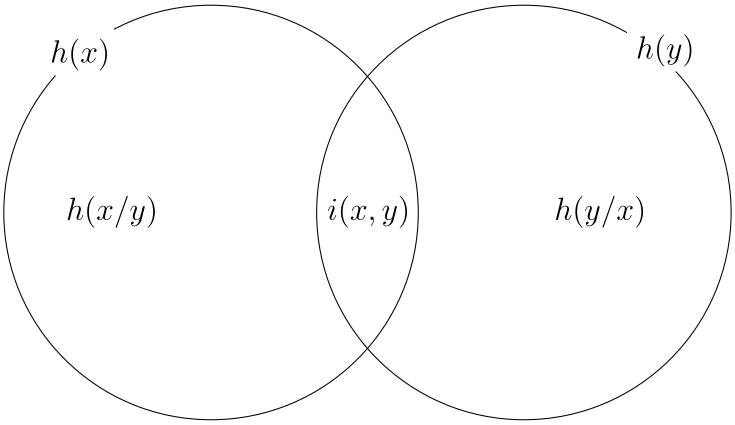


Figure 9.134: **Graphical representation of the conditional entropy and the mutual information**

From which, and using the previous results, we deduce, taking $\lambda = p(x,y)$, that:

$$\begin{aligned}I(x,y) &= \lambda \sum_{x,y} P(x,y)\log\left(\frac{P(x,y)}{P(x)P(y)}\right)\\ &= h(x)-h(x/y)\\ &= h(y)-h(y/x)\\ &= \underbrace{h(x)+h(y)}_{\text{max entropy}} - \underbrace{h(x,y)}_{\text{joint entropy}}\end{aligned} \tag{9.5.28}$$

If x and y are independent (x tell us nothing about y), then $h(y/x)=h(y)$ and therefore we have for the mutual information (the two sets are completely separated):

$$\text{MMI} = I(x,y) = h(y)-h(y/x) = h(y)-h(y) = 0 \tag{9.5.29}$$

or equivalently:

$$\text{MMI} = I(x,y) = h(x)+h(y)-h(x,y) = h(x)+h(y)-(h(x)+h(y)) = 0 \tag{9.5.30}$$

If x tell us everything you y then $h(y/x)=0$ and we have (both sets are completely superposed):

$$\text{MMI} = I(x,y) = h(y)-h(y/x) = h(y) \tag{9.5.31}$$

Remark

Some methods in advanced statistics and therefore data science use the MMI rather than the Pearson correlation coefficient as we know that this latter has some issues for special cases of non-linear regressions (see section Statistics page 615). However the difficulty in practice is that it requires to know the joint probability distribution function for the variables and some form of discretization using kernels.

✍**Examples:**

E1. Given a binary random variable which value is 1 with probability p (and therefore 0 with probability $1-p$). Its entropy is given by:

$$\begin{gathered}S_2(p)=S_2(1-p)\Leftrightarrow S_2(p)-S_2(1-p)=0\\ \Rightarrow -p\lambda\log(p)-(1-p)\lambda\log(1-p)=0\end{gathered} \tag{9.5.32}$$

with obviously $0<p<1$ and with a logarithm in base 2 such that for any event with two equiprobable states, the entropy of obtaining one of the two states is equal to unity. Indeed:

$$S_2(p)=-p\lambda\log_2(p)=-0.5\lambda\log_2(0.5)=-0.5\lambda\cdot -1=1 \tag{9.5.33}$$

Therefore we must have $\lambda=1$.

The entropy is shown in the figure below, in "Shannon" units (unit corresponding to the use of the logarithm with base 2). We remark its symmetry about the value $1/2$ value at which it reaches its maximum, equal to 1 as expected intuitively for such an experience.

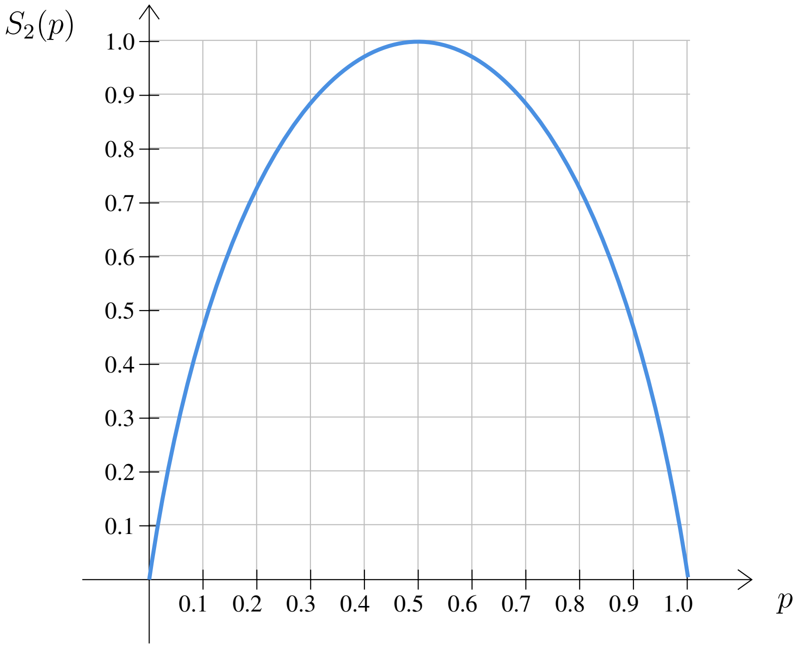


Figure 9.135: **Entropy of a binary variable with $\lambda=1$**

E2. In the field of astronomy where it is often necessary to compress images we seek to calculate the "data redundancy level" to choose a compression ratio eventually applicable. This is done by calculating the information contained in an image using its entropy. Given an image whose pixels are coded on N bits, thus 2^N possible values by pixels, and a proportion (probability) p_i of pixels that can take each color. The entropy of the image is given by (with the notation usage that is traditional in digital imaging):

$$H=-\sum_{i=1}^{2^N} p_i\log_2(p_i) \tag{9.5.34}$$

which is therefore expressed in bits per pixel and represents the number of bits that are necessary to encode the image without loss of information.

E3. We have seen during our study of combinatorics that to give the strength of a password based on the number of combination was not very a convenient way to communicate. That's why most IT professional like to give the entropy of a password in bits! Let us for example consider a password of 8 characters from an alphabet of 94 characters (all characters that can be made with a standard keyboard user). Then we have using the tradition of IT professionals :

$$H=-\lambda\sum_{i=1}^{2^N} p_i\log_2(p_i)=\log_2(94^8)\cong\log_2(2^{52})=52 \tag{9.5.35}$$

Hence according to this empirical tradition such a password has an entropy of 52 bits and such a number is indeed much more convenient to communicate than the huge value of the combinatorics!

The entropy of a discrete random variable is a non-negative number:

$$H(X)=-\sum_{x\in\mathcal{X}} p(x)\cdot\log_b p(x)\geq 0 \tag{9.5.36}$$

The minus sign can be moved into the sum:

$$H(X)=\sum_{x\in\mathcal{X}}[p(x)\cdot(-\log_b p(x))] \tag{9.5.37}$$

Because the co-domain of probability mass functions is $[0,1]$, we can deduce:

$$\begin{array}{rcccl} 0 & \leq & p(x) & \leq & 1\\ -\infty & \leq & \log_b\left(p(x)\right) & \leq & 0\\ 0 & \leq & -\log_b\left(p(x)\right) & \leq & +\infty\\ 0 & \leq & p(x)\cdot\left[-\log_b\left(p(x)\right)\right] & \leq & +\infty \end{array}$$

By convention, $0\cdot\log_b(0)$ is taken to be 0 when calculating entropy, consistent with:

$$\lim_{p\rightarrow 0}[p\log_b(p)]=0 \tag{9.5.38}$$

Taking this together, each term in the sum is positive or zero and thus, the entire sum must also be non-negative.

Remark

Computer scientists use "algorithmic entropy as a special case of entropy as studied in Statistical Mechanics. This viewpoint allows them to apply many techniques developed for use in thermodynamics to the subject of algorithmic information theory. They take the log runtime, length, and output of a program as observables analogous to the energy E, volume V and number of molecules N in a container of gas. The conjugate variables of these observables allow them to define quantities which they name the "algorithmic temperature" T, "algorithmic pressure", P and "algorithmic potential" μ, since they are analogous to the temperature, pressure and chemical potential.

Later we will make the connection between the Statistical Information Theory and Statistical Mechanics!

9.5.1.1 High Entropy Structures

Now we want to understand why there are alloys (or other similar structures) named "high entropy alloys" (HEA) like the $\mathrm{AlCoCrFeNi}$ with body-centered cubic (BCC) lattice and illustrated as:

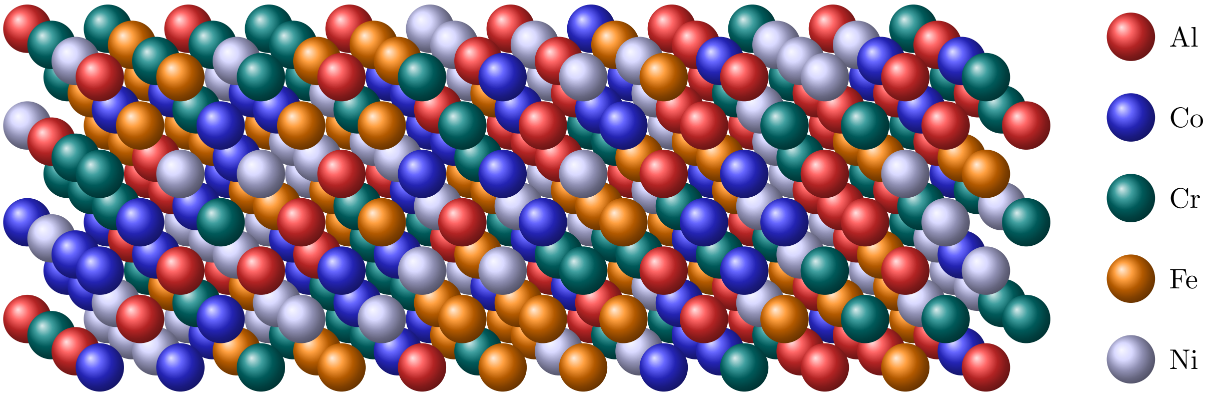


Figure 9.136: $\mathrm{AlCoCrFeNi}$ **High Entropy Alloy (author: Janosh Riebesell)**

To understand let's say we have 3 buckets with 4 balls each. The balls have the following colors:

- Bucket 1: 4 red balls
- Bucket 2: 3 red balls and 1 blue ball
- Bucket 3: 2 red balls and 2 blue balls

And we will judge these three options by how much information we have on the color of a ball drawn at random. In this case, we have the following:

1. In the first bucket, we will know for sure that the ball coming out is red.

2. In the second bucket, we know with 75% certainty that the ball is red, and with 25% certainty that it is blue.
3. In the third bucket, we know with 50% certainty that the ball is red, and with the same certainty that it is blue.

So it makes sense to say that Bucket 1 gives us the most amount of "knowledge" about what ball we will draw (because we know for sure it is red), that Bucket 2 gives us some knowledge, and that Bucket 3 will give us the least amount of knowledge. Well, Entropy is in some way, the opposite of knowledge. So we will say that Bucket 1 has the least amount of entropy, Bucket 2 has medium entropy, and Bucket 3 has the greatest amount of entropy.

How de we compute the entropy for this game? For sure, we will use probabilities!

So now the question is, how do we cook up a formula which gives us a low number for a bucket with 4 red balls, a high number for a bucket with 2 red and 2 blue balls, and a medium number for a bucket with 3 red and 1 blue balls? Well, as a first attempt, let us remember the definition of entropy: If molecules have many possible rearrangements, then the system has high entropy, and if they have very few rearrangements, then the system has low entropy. So a first attempt would be to count the number of rearrangements of these balls. In this case, we have 1 possible rearrangement for Bucket 1, 4 for Bucket 2, and 6 for Bucket 3, this number given by the binomial coefficient.

This number of arrangements won't be part of the formula for entropy, but it gives us an idea, that if there are many arrangements, then entropy is large, and if there are very few arrangements, then entropy is low. In the next section, we will cook up a formula for entropy. The idea is, to consider the probability of drawing the balls in a certain way, from each bucket (each time we draw a ball we put it back in the bucket!!!!).

Consider the following cases:

- The Bucket 1 has 4 red balls (`RRRR`). Then the probability of drawing the balls in the same order (as there are all red and hence indistinguishable) is:

$$1 \cdot 1 \cdot 1 \cdot 1 = 1$$

- The Bucket 2 has 3 red balls and 1 green ball (`RRRG`). Then the probability of drawing the balls in the same order is (don't forget that each time we draw a ball we put it back in the bucket!!!!):

$$0.75 \cdot 0.75 \cdot 0.75 \cdot 0.25 = 0.105$$

- The Bucket 3 has 2 red balls and 2 green ball2 (`RRGG`). Then the probability of drawing the balls in the same order is (don't forget that each time we draw a ball we put it back in the bucket!!!!):

$$0.5 \cdot 0.5 \cdot 0.5 \cdot 0.5 = 0.0625$$

Ok, now we have some measure that gives us different values for the three Buckets. In order to build the entropy formula, we want the opposite, some measure that gives us a low number for

Bucket 1, a medium number for Bucket 2, and a high number for Bucket 3. No problem, this is where the entropy relation will be useful:

$$S(X)=\mathrm{E}(h(x))=-\lambda\sum_{i=1}^{n}P_i\log\left(P_i\right)$$

Taking the logarithm for example in base 10 and $\lambda=1$, we get:

- The Bucket 1:
$$S(X)=\mathrm{E}(h(x))=-\sum_{i=1}^{n}P_i\log\left(P_i\right)=0$$
- The Bucket 2:
$$S(X)=\mathrm{E}(h(x))=-\sum_{i=1}^{n}P_i\log\left(P_i\right)\cong 0.431$$
- The Bucket 3:
$$S(X)=\mathrm{E}(h(x))=-\sum_{i=1}^{n}P_i\log\left(P_i\right)\cong 0.602$$

So indeed the great the mixtures of balls, the higher is the entropy. Hence the name of hight entropy alloys.

9.5.1.2 Usual Shannon entropies

Let us see some common Shannon entropies:

1. For a random variable following a discrete uniform distribution:

   Let X be a random variable following a discrete uniform distribution:
   $$X\sim\mathcal{U}(a,b) \tag{9.5.39}$$
   Then, the (Shannon) entropy of X in nats is:
   $$H(X)=\ln(b-a+1) \tag{9.5.40}$$
   Indeed, the entropy is defined as the probability-weighted average of the logarithmized probabilities for all possible values:
   $$H(X)=-\sum_{x\in\mathcal{X}}p(x)\cdot\log_b\left(p(x)\right) \tag{9.5.41}$$

Entropy is measured in nats by setting $b=e$. Then, with the probability mass function of the discrete uniform distribution (see section Statistics page 649), we have:

$$\begin{aligned}H(X)&=-\sum_{x\in\mathcal{X}}p(x)\cdot\log_e p(x)=-\sum_{x=a}^{b}p(x)\cdot\ln\left(p(x)\right)\\&=-\sum_{x=a}^{b}\frac{1}{b-a+1}\cdot\ln\left(\frac{1}{b-a+1}\right)\\&=-(b-a+1)\cdot\frac{1}{b-a+1}\cdot\ln\left(\frac{1}{b-a+1}\right)\\&=-\ln\left(\frac{1}{b-a+1}\right)=\ln(b-a+1)\end{aligned}\tag{9.5.42}$$

The reasoning is exactly similar for the continuous uniform distribution (see section Statistics page 738) and leads to the result: $H(X)=\ln(b-a)$.

2. For a random variable following a Bernoulli distribution:

   Now let X be a random variable following a Bernoulli distribution:

   $$X\sim\text{Bern}(p)\tag{9.5.43}$$

   Then, the (Shannon) entropy of X in bits is:

   $$H(X)=-p\log_2(p)-(1-p)\log_2(1-p)\tag{9.5.44}$$

   Indeed, the Shannon entropy is defined as the probability-weighted average of the logarithmized probabilities for all possible values:

   $$H(X)=-\sum_{x\in\mathcal{X}}p(x)\cdot\log_b\left(p(x)\right)\tag{9.5.45}$$

   Entropy is measured in bits by setting $b=2$. Since there are only two possible outcomes for a Bernoulli random variable, we have:

   $$\begin{aligned}H(X)&=-P(X=0)\cdot\log_2\left(P(X=0)\right)-P(X=1)\cdot\log_2\left(P(X=1)\right)\\&=-p\log_2(p)-(1-p)\log_2(1-p)\end{aligned}\tag{9.5.46}$$

3. For a random variable following a Normal distribution:

   Now let X be a random variable following a Normal distribution:

   $$X\sim\mathcal{N}\left(\mu,\sigma^2\right)\tag{9.5.47}$$

   Then, the differential entropy of X is:

   $$h(X)=\frac{1}{2}\ln\left(2\pi\sigma^2 e\right)\tag{9.5.48}$$

   Indeed, remember that the differential entropy of a random variable is defined as:

   $$h(X)=-\int_{\mathcal{X}}p(x)\log_b(p(x))\mathrm{d}x\tag{9.5.49}$$

To measure $h(X)$ in nats, we set $b=e$, such that:

$$h(X)=-\mathrm{E}[\ln(p(x))] \tag{9.5.50}$$

With the probability density function of the Normal distribution (see section Statistics page 692), the differential entropy of X is:

$$\begin{aligned}h(X)&=-\mathrm{E}\left[\ln\left(\frac{1}{\sqrt{2\pi}\sigma}\cdot\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]\right)\right]\\&=-\mathrm{E}\left[-\frac{1}{2}\ln\left(2\pi\sigma^2\right)-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]\\&=\frac{1}{2}\ln\left(2\pi\sigma^2\right)+\frac{1}{2}\mathrm{E}\left[\left(\frac{x-\mu}{\sigma}\right)^2\right]\\&=\frac{1}{2}\ln\left(2\pi\sigma^2\right)+\frac{1}{2}\cdot\frac{1}{\sigma^2}\cdot\mathrm{E}\left[(x-\mu)^2\right]\end{aligned} \tag{9.5.51}$$

Note that $\mathrm{E}\left[(x-\mu)^2\right]$ corresponds to the variance of X and the variance of the normal distribution is σ^2. Thus, we can proceed:

$$\begin{aligned}h(X)&=\frac{1}{2}\ln\left(2\pi\sigma^2\right)+\frac{1}{2}\cdot\frac{1}{\sigma^2}\cdot\sigma^2\\&=\frac{1}{2}\ln\left(2\pi\sigma^2\right)+\frac{1}{2}\\&=\frac{1}{2}\ln\left(2\pi\sigma^2 e\right)\end{aligned} \tag{9.5.52}$$

4. For a random variable following a multivariate Normal distribution:

   Let x follow a multivariate normal distribution:

   $$x\sim\mathcal{N}(\mu,\Sigma) \tag{9.5.53}$$

   Then, the differential entropy of x in nats is:

   $$h(x)=\frac{n}{2}\ln(2\pi)+\frac{1}{2}\ln|\Sigma|+\frac{1}{2}n \tag{9.5.54}$$

   Indeed, the differential entropy of a random variable is defined as:

   $$h(X)=-\int_{\mathcal{X}}p(x)\log_b(p(x))\mathrm{d}x \tag{9.5.55}$$

   To measure $h(X)$ in nats, we set $b=e$, such that:

   $$h(X)=-\mathrm{E}[\ln(p(x))] \tag{9.5.56}$$

   With the probability density function of the multivariate normal distribution , the differential entropy of x is:

   $$\begin{aligned}h(x)&=-\mathrm{E}\left[\ln\left(\frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\cdot\exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]\right)\right]\\&=-\mathrm{E}\left[-\frac{n}{2}\ln(2\pi)-\frac{1}{2}\ln|\Sigma|-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]\\&=\frac{n}{2}\ln(2\pi)+\frac{1}{2}\ln|\Sigma|+\frac{1}{2}\mathrm{E}\left[(x-\mu)^T\Sigma^{-1}(x-\mu)\right]\end{aligned} \tag{9.5.57}$$

The last term can be evaluated as:

$$\begin{aligned}\mathrm{E}\left[(x-\mu)^T\Sigma^{-1}(x-\mu)\right]&=\mathrm{E}\left[\mathrm{tr}\left((x-\mu)^T\Sigma^{-1}(x-\mu)\right)\right]\\&=\mathrm{E}\left[\mathrm{tr}\left(\Sigma^{-1}(x-\mu)(x-\mu)^{\mathrm{T}}\right)\right]\\&=\mathrm{tr}\left(\Sigma^{-1}\mathrm{E}\left[(x-\mu)(x-\mu)^T\right]\right)\\&=\mathrm{tr}\left(\Sigma^{-1}\Sigma\right)=\mathrm{tr}\left(I_n\right)=n\end{aligned} \tag{9.5.58}$$

such that the differential entropy is:

$$h(x)=\frac{n}{2}\ln(2\pi)+\frac{1}{2}\ln|\Sigma|+\frac{1}{2}n \tag{9.5.59}$$

9.5.1.3 Perplexity

"Perplexity" is a concept used in Text Mining and some Machine Learning algorithms (we will see much later a usage of that latter during our study of the T-SNE at page 6939) and defined as:

$$\boxed{\mathrm{Perp}(p)=b^{S(p)}=b^{\mathrm{E}[\log_b(p)]}} \tag{9.5.60}$$

where b is the base of the logarithm used. We can already point out that perplexity is invariant with the base we use to define entropy. So in this sense, perplexity is infinitely more unique/less arbitrary than entropy as a measurement.

For discrete case, suppose we have n discrete states:

$$\mathrm{Perp}(p)=b^{S(p)}=b^{-\sum_{i=1}^n p(x_i)\log_b p(x_i)} \tag{9.5.61}$$

So it has the exact same information to Shannon entropy. And we will explain now why people invented this!

A very common version of the above relation is in base $b=2$. We then have:

$$\mathrm{Perp}(2)=2^{S(p)}=2^{-\sum_{i=1}^n p(x_i)\log_2 p(x_i)} \tag{9.5.62}$$

Now to understand the origin consider N events equally likely. We then have using the definition and properties of exponentiation and logarithm we get:

$$\mathrm{Perp}(2)=2^{S(p)}=2^{-\sum_{i=1}^n \frac{1}{N}\log_2 p(x_i)}=\left(\prod_{i=1}^n\frac{1}{p(x_i)}\right)^{\frac{1}{N}}=\sqrt{\prod_{i=1}^n\frac{1}{p(x_i)}} \tag{9.5.63}$$

So the perplexity can be seen as the geometric average of the inverse probabilities!

Let's play with this a bit. Let's say you're just looking at a coin. When the coin is fair, entropy is at a maximum, and perplexity is at a maximum of:

$$\frac{1}{\frac{1}{2}^{\frac{1}{2}}\cdot\frac{1}{2}^{\frac{1}{2}}}=2 \tag{9.5.64}$$

Now what happens when we look at an N sided dice? Perplexity is:

$$\frac{1}{\left(\frac{1}{N}^{\frac{1}{N}}\right)^{N}}=N \tag{9.5.65}$$

So perplexity represents the number of sides of a fair die that when rolled, produces a sequence with the same entropy as your given probability distribution.

Therefore perplexity of whatever you're evaluating, on the data we are evaluating it on, sort of tells « *us this thing is right about as often as a x-sided die would be*».

9.5.1.4 Kullback–Leibler divergence

In mathematical statistics, the Kullback–Leibler divergence (also called "relative entropy") is a measure of how one probability distribution diverges from a second, expected probability distribution. Applications include characterizing the relative (Shannon) entropy in information systems, randomness in continuous time-series, and information gain when comparing statistical models of inference. In contrast to variation of information, it is a distribution-wise asymmetric measure and thus does not qualify as a statistical metric of spread. In the simple case, a Kullback–Leibler divergence of 0 indicates that we can expect similar, if not the same, behaviour of two different distributions, while a Kullback–Leibler divergence of 1 indicates that the two distributions behave in such a different manner that the expectation given the first distribution approaches zero. In simplified terms, it is a measure of surprise, with diverse applications such as in statistics and Machine Learning (see section Numerical Methods page 6768).

Given the Shannon entropy introduced earlier before:

$$S(X)=-\lambda\sum_{i=1}^{n}p_i(x)\log(p_i(x)) \tag{9.5.66}$$

Kullback-Leibler Divergence is just a slight modification of our formula for entropy. Rather than just having our probability distribution p_X we add in an approximating distribution p_Y. Then we look at the difference of the log values for each (and we put $\lambda=1$):

$$D_{\text{KL}}(p_X,p_Y)=\sum_{i=1}^{n}p_X(x_i)\left(\log(p_X(x_i)-p_Y(x_i)\right) \tag{9.5.67}$$

Essentially, what we're looking at with the KL divergence is the expectation of the log difference between the probability of data in the original distribution with the approximating distribution. So the Kullback–Leibler divergence is measure of entropy increase due to the use of an approximation to the true distribution rather than the true distribution itself.

The "discrete Kullback-Leibler divergence" between two distributions X and Y is given by:

$$\boxed{D_{\text{KL}}(p_X,p_Y)=\sum_{x\in S}p_X(x)\log\left(\frac{p_X(x)}{p_Y(x)}\right)} \tag{9.5.68}$$

also sometimes denoted:

$$\mathrm{KL}(p_X||p_Y)=\sum_{x\in S}p_X(x)\log\left(\frac{p_X(x)}{p_Y(x)}\right)\tag{9.5.69}$$

and the "continuous Kullback-Leibler divergence" between two distributions X and Y is given by:

$$\boxed{D_{\mathrm{KL}}(p_X,p_Y)=\int p_X(x)\log\left(\frac{p_X(x)}{p_Y(x)}\right)\mathrm{d}x}\tag{9.5.70}$$

The Kullback-Leibler divergence is a kind of distance between the probability distributions of X and Y, although $D_{\mathrm{KL}}(p_X,p_Y)\neq D_{\mathrm{KL}}(p_Y,p_X)$ (but it is therefore not a true "distance" because it is asymmetric!!!!).

As it is quite abstract let us do a companion example of two Normal distributions:

$$p_X=\mathcal{N}(\mu,\sigma^2)\qquad p_Y=\mathcal{N}(m,s^2)\tag{9.5.71}$$

So now we want obviously to calculate:

$$\begin{aligned}D_{\mathrm{KL}}(p_X,p_Y)&=\int p_X(x)\log\left(\frac{p_X(x)}{p_Y(x)}\right)\mathrm{d}x=\int p_X(x)\log\left(\frac{\dfrac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\dfrac{1}{\sqrt{2\pi s^2}}e^{-\frac{(x-m)^2}{2s^2}}}\right)\mathrm{d}x\\&=\int p_X(x)\log\left(\sqrt{\frac{s^2}{\sigma^2}}\right)\mathrm{d}x+\int p_X(x)\log\left(-\frac{(x-\mu)^2}{2\sigma^2}+\frac{(x-m)^2}{2s^2}\right)\mathrm{d}x\\&=\frac{1}{2}\log\left(\frac{s^2}{\sigma^2}\right)\int p_X(x)\mathrm{d}x+\frac{1}{2\sigma^2}\left[-\int(x-\mu)^2p_X(x)\mathrm{d}x\right]+\frac{1}{2s^2}\left[\int(x-m)^2p_X(x)\mathrm{d}x\right]\end{aligned}\tag{9.5.72}$$

But we have:

$$\int p_X(x)\mathrm{d}x=1\tag{9.5.73}$$

and:

$$\int(x-\mu)^2p_X(x)\mathrm{d}x=\int x^2p_X(x)\mathrm{d}x-\int 2\mu x\mathrm{d}x+\int\mu^2p_X(x)\mathrm{d}x=\sigma^2-2\mu^2+\mu^2=\sigma^2\tag{9.5.74}$$

Therefore:

$$D_{\mathrm{KL}}(p_X,p_Y)=\frac{1}{2}\log\left(\frac{s^2}{\sigma^2}\right)-\frac{\sigma^2}{2\sigma^2}+\frac{1}{2s^2}\left[\int(x-m)^2p_X(x)\mathrm{d}x\right]\tag{9.5.75}$$

But we have:

$$\begin{aligned}\int(x-m)^2p_X(x)\mathrm{d}x&=\int(x-\mu+\mu-m)^2p_X(x)\mathrm{d}x\\&=\int(x-\mu)^2p_X(x)\mathrm{d}x+(\mu-m)^2\int p_X(x)\mathrm{d}x+2(\mu-m)\int(x-\mu)p_X(x)\mathrm{d}x\\&=\sigma^2+(\mu-m)^2+2(\mu-m)\left[\int xp_X(x)\mathrm{d}x-\mu\int p_X(x)\mathrm{d}x\right]\\&=\sigma^2+(\mu-m)^2+2(\mu-m)(\mu-\mu)=\sigma^2+(\mu-m)^2\end{aligned}\tag{9.5.76}$$

Therefore:

$$\begin{aligned}D_{\text{KL}}(p_X,p_Y)=\frac{1}{2}\log\left(\frac{s^2}{\sigma^2}\right)-\frac{1}{2}+\frac{1}{2s^2}\left[\int(x-m)^2p_X(x)\mathrm{d}x\right]&=\frac{1}{2}\log\left(\frac{s^2}{\sigma^2}\right)-\frac{1}{2s^2}\left[\sigma^2+(\mu-m)^2\right]\\&=\frac{1}{2}\left[\log\left(\frac{s^2}{\sigma^2}\right)-1+\frac{\sigma^2+(\mu-m)^2}{s^2}\right]\end{aligned} \tag{9.5.77}$$

So finally:

$$D_{\text{KL}}\left(\mathcal{N}(\mu,\sigma),\mathcal{N}(m,s)\right)=\frac{1}{2}\left[\log\left(\frac{s^2}{\sigma^2}\right)-1+\frac{\sigma^2+(\mu-m)^2}{s^2}\right] \tag{9.5.78}$$

Let us take another example based on two discrete uniform distribution $p=\mathcal{U}(a_1,b_1)$ and $q=\mathcal{U}(a_2,b_2)$ and as we know the cumulative distribution function of the latter (see section Statistics page 649), then we have:

$$\begin{aligned}D_{\text{KL}}(p_X,p_Y)&=\sum_{x=a_1}^{b_1}\frac{1}{b_1-a_1+1}\cdot\ln\left(\frac{\frac{1}{b_1-a_1+1}}{\frac{1}{b_2-a_2+1}}\right)\\&=\frac{1}{b_1-a_1+1}\cdot\ln\left(\frac{b_2-a_2+1}{b_1-a_1+1}\right)\sum_{x=a_1}^{b_1}1\\&=\frac{1}{b_1-a_1+1}\cdot\ln\left(\frac{b_2-a_2+1}{b_1-a_1+1}\right)\cdot(b_1-a_1+1)\\&=\ln\left(\frac{b_2-a_2+1}{b_1-a_1+1}\right)\end{aligned} \tag{9.5.79}$$

The reasoning is exactly similar for the continuous uniform distribution (see section Statistics page 738) and leads to the result:

$$D_{\text{KL}}(p_X,p_Y)=\ln\left(\frac{b_2-a_2}{b_1-a_1}\right) \tag{9.5.80}$$

Taking another example based on the Bernoulli distributions:

$$P(X=x)=\begin{cases}p & \text{if } x=1\\1-p & \text{if } x=0\\0 & \text{otherwise}\end{cases}\qquad P(X=x)=\begin{cases}q & \text{if } x=1\\1-q & \text{if } x=0\\0 & \text{otherwise}\end{cases} \tag{9.5.81}$$

we have straightforward:

$$D_{\text{KL}}(p_X,p_Y)=p\log\left(\frac{p}{q}\right)+(1-p)\log\left(\frac{1-p}{1-q}\right) \tag{9.5.82}$$

The Kullback–Leibler divergence has the following properties:

P1. It is positive definite such that $D_{\text{KL}}\geq 0$

Proof 9.10.2. We start from:

$$D_{\mathrm{KL}}(p_X,p_Y)=\sum_{x\in S}p_X(x)\log\left(\frac{p_X(x)}{p_Y(x)}\right)=-\sum_{x\in S}p_X(x)\log\left(\frac{p_Y(x)}{p_X(x)}\right) \tag{9.5.83}$$

As $-\log(z)$ is a convex function we can apply the Jensen's equality (see section Statistics page 576) that is given for recall by:

$$f(\mathrm{E}(X))\leq \mathrm{E}(f(X)) \tag{9.5.84}$$

Therefore:

$$-\log\left(\sum_{x\in S}p_X(x)\frac{p_Y(x)}{p_X(x)}\right)\leq -\sum_{x\in S}p_X(x)\log\left(\frac{p_Y(x)}{p_X(x)}\right) \tag{9.5.85}$$

Hence if we simplify:

$$-\log\left(\sum_{x\in S}p_Y(x)\right)\leq -\sum_{x\in S}p_X(x)\log\left(\frac{p_Y(x)}{p_X(x)}\right) \tag{9.5.86}$$

This is equal to write:

$$-\log(1)\leq -\sum_{x\in S}p_X(x)\log\left(\frac{p_Y(x)}{p_X(x)}\right) \tag{9.5.87}$$

Hence:

$$0\leq -\sum_{x\in S}p_X(x)\log\left(\frac{p_Y(x)}{p_X(x)}\right) \tag{9.5.88}$$

Therefore:

$$\sum_{x\in S}p_X(x)\log\left(\frac{p_X(x)}{p_Y(x)}\right)\geq 0 \tag{9.5.89}$$

■ Q.E.D.

P2. It is zero if both distribution are equal $D_{\mathrm{KL}}(p_X,p_Y)=0$ if $p_X=p_Y$.

Proof 9.10.3. The proof is immediate:

$$D_{\mathrm{KL}}(p_X,p_x)=\sum_{x\in S}p_X(x)\log\left(\frac{p_X(x)}{p_X(x)}\right)=\sum_{x\in S}p_X(x)\log(1)=\sum_{x\in S}p_X(x)\cdot 0=0 \tag{9.5.90}$$

■ Q.E.D.

Thanks to the first property we can prove that the discrete uniform distribution maximizes (Shannon) entropy for a random variable with finite support. Let us remind first that the Shannon entropy is defined by:

$$H(X)=-\sum_{x\in\mathcal{X}}p(x)\log(p(x)) \tag{9.5.91}$$

And that we have for the Kullback-Leibler divergence:

$$\begin{aligned}0\leq D_{\text{KL}}(f,g)&=\sum_{x\in\mathcal{X}}f(x)\log\left(\frac{f(x)}{g(x)}\right)\\&=\sum_{x\in\mathcal{X}}f(x)\log\left(f(x)\right)-\sum_{x\in\mathcal{X}}f(x)\log\left(g(x)\right)\\&=-H[f(x)]-\sum_{x\in\mathcal{X}}f(x)\log\left(g(x)\right)\end{aligned}\tag{9.5.92}$$

By plugging the probability mass function of the discrete uniform distribution (see section Statistics page 649) into the second term, we get:

$$\sum_{x\in\mathcal{X}}f(x)\log\left(g(x)\right)=\sum_{x=1}^{n}f(x)\log\left(\frac{1}{n-1+1}\right)=\log\left(\frac{1}{n}\right)\sum_{x=1}^{n}f(x)=-\log(n)\tag{9.5.93}$$

This is actually the negative of the Shannon entropy of the discrete uniform distribution (proven earlier above) such that:

$$\sum_{x\in\mathcal{X}}f(x)\log g(x)=-H\left(\mathcal{U}(1,n)\right)=-\text{H}[g(x)]\tag{9.5.94}$$

Combining the previous results above with the previous one, we get:

$$\begin{aligned}0&\leq D_{\text{KL}}(f,g)\\0&\leq -H(f(x))-(-H(g(x)))\\H(g(x))&\geq H(f(x))\end{aligned}\tag{9.5.95}$$

which means that the entropy of the discrete uniform distribution $\mathcal{U}(a,b)$ will be larger than or equal to any other distribution defined on the same set of values $\{a,\ldots,b\}$! Obviously the same reasoning and conclusion applies for the continuous uniform distribution.

Let X be a random variable. Assume two Bernoulli distributions P and Q specifying the probability distribution of X as:

$$\begin{aligned}P&:X\sim\text{Bern}\left(p_1\right)\\Q&:X\sim\text{Bern}\left(p_2\right)\end{aligned}\tag{9.5.96}$$

Then, the Kullback-Leibler divergence of P from Q is given by:

$$D_{\text{KL}}(p,q)=\ln\left(\frac{1-p_1}{1-p_2}\right)+p_1\cdot\ln\left(\frac{p_1\left(1-p_2\right)}{p_2\left(1-p_1\right)}\right)\tag{9.5.97}$$

Indeed, remember that for a discrete random variable the Kullback-Leibler divergence is given by:

$$D_{\text{KL}}(p,q)=\sum_{x\in\mathcal{X}}p(x)\ln\left(\frac{p(x)}{q(x)}\right)\tag{9.5.98}$$

which, applied to the Bernoulli distributions in, yields:

$$\begin{aligned}D_{\text{KL}}(p,q)&=\sum_{x\in\{0,1\}}p(x)\ln\left(\frac{p(x)}{q(x)}\right)\\&=p(X=0)\cdot\ln\left(\frac{p(X=0)}{q(X=0)}\right)+p(X=1)\cdot\ln\left(\frac{p(X=1)}{q(X=1)}\right)\end{aligned}\tag{9.5.99}$$

Using the probability mass function of the Bernoulli distribution (see section Probabilities page 652), this becomes:

$$\begin{aligned}D_{\text{KL}}(p,q)&=(1-p_1)\cdot\ln\left(\frac{1-p_1}{1-p_2}\right)+p_1\cdot\ln\left(\frac{p_1}{p_2}\right)\\&=\ln\left(\frac{1-p_1}{1-p_2}\right)+p_1\cdot\ln\left(\frac{p_1}{p_2}\right)-p_1\cdot\ln\left(\frac{1-p_1}{1-p_2}\right)\\&=\ln\left(\frac{1-p_1}{1-p_2}\right)+p_1\cdot\left[\ln\left(\frac{p_1}{p_2}\right)+\ln\left(\frac{1-p_2}{1-p_1}\right)\right]\\&=\ln\left(\frac{1-p_1}{1-p_2}\right)+p_1\cdot\ln\left(\frac{p_1\left(1-p_2\right)}{p_2\left(1-p_1\right)}\right)\end{aligned}\tag{9.5.100}$$

Now let X be a random variable. Assume two binomial distributions P and Q specifying the probability distribution of X as:

$$\begin{aligned}P&:X\sim\text{Bin}\left(n,p_1\right)\\Q&:X\sim\text{Bin}\left(n,p_2\right)\end{aligned}\tag{9.5.101}$$

Then, the Kullback-Leibler divergence of P from Q is given by:

$$D_{\text{KL}}(p,q)=np_1\cdot\ln\left(\frac{p_1}{p_2}\right)+n\left(1-p_1\right)\cdot\ln\left(\frac{1-p_1}{1-p_2}\right)\tag{9.5.102}$$

Indeed, as the Kullback-Leibler divergence for a discrete random variable is given by:

$$D_{\text{KL}}(p,q)=\sum_{x\in\mathcal{X}}p(x)\ln\left(\frac{p(x)}{q(x)}\right)\tag{9.5.103}$$

which, applied to the binomial distributions yields:

$$\begin{aligned}D_{\text{KL}}(p,q)&=\sum_{x=0}^{n}p(x)\ln\left(\frac{p(x)}{q(x)}\right)\\&=p(X=0)\cdot\ln\left(\frac{p(X=0)}{q(X=0)}\right)+\ldots+p(X=n)\cdot\ln\left(\frac{p(X=n)}{q(X=n)}\right)\end{aligned}\tag{9.5.104}$$

Using the probability mass function of the binomial distribution (see section Statistics page 658), this becomes:

$$\begin{aligned}D_{\text{KL}}(p,q)&=\sum_{x=0}^{n}\binom{n}{x}p_1^x\left(1-p_1\right)^{n-x}\cdot\ln\left(\frac{\binom{n}{x}p_1^x\left(1-p_1\right)^{n-x}}{\binom{n}{x}p_2^x\left(1-p_2\right)^{n-x}}\right)\\&=\sum_{x=0}^{n}\binom{n}{x}p_1^x\left(1-p_1\right)^{n-x}\cdot\left[x\cdot\ln\left(\frac{p_1}{p_2}\right)+(n-x)\cdot\ln\left(\frac{1-p_1}{1-p_2}\right)\right]\\&=\ln\left(\frac{p_1}{p_2}\right)\cdot\sum_{x=0}^{n}\binom{n}{x}p_1^x\left(1-p_1\right)^{n-x}x+\ln\left(\frac{1-p_1}{1-p_2}\right)\cdot\sum_{x=0}^{n}\binom{n}{x}p_1^x\left(1-p_1\right)^{n-x}(n-x)\end{aligned}\tag{9.5.105}$$

We can now see that some terms in this sum are expected values with respect to binomial distributions:

$$\begin{gathered}\sum_{x=0}^{n}\left(\begin{array}{c}n\\x\end{array}\right)p_1^x\left(1-p_1\right)^{n-x}x=\mathrm{E}(x)_{\mathrm{Bin}(n,p_1)}\\ \sum_{x=0}^{n}\left(\begin{array}{c}n\\x\end{array}\right)p_1^x\left(1-p_1\right)^{n-x}(n-x)=\mathrm{E}(n-x)_{\mathrm{Bin}(n,p_1)}\end{gathered}\tag{9.5.106}$$

Using the expected value of the binomial distribution, these can be simplified to:

$$\begin{aligned}\mathrm{E}(x)_{\mathrm{Bin}(n,p_1)}&=np_1\\ \mathrm{E}(n-x)_{\mathrm{Bin}(n,p_1)}&=n-np_1\end{aligned}\tag{9.5.107}$$

such that the Kullback-Leibler divergence finally becomes:

$$D_{\mathrm{KL}}(p,q)=np_1\cdot\ln\left(\frac{p_1}{p_2}\right)+n\left(1-p_1\right)\cdot\ln\left(\frac{1-p_1}{1-p_2}\right)\tag{9.5.108}$$

Remark

For a practical numerical application, the reader can refer to our R companion book! The reader may have notice that the Kullback-Leibler divergence is not a distance (i.e. a "metric") as obviously $D_{\mathrm{KL}}(p_X,p_Y)=D_{\mathrm{KL}}(p_Y,p_X)$ (i.e. the KL divergence is not symmetric). This is why there is another approach named the "Wasserstein metric" that doesn't measure the divergence but the distance between two probability distributions!

In an approach to statistical inference based on information theory, the true but unknown distribution is compared with information in the sample using a divergence measure between the population distribution and the sample distribution. The divergence measure may also be used to compare two hypothesized distributions. A general type of divergence measure is named the "ϕ-divergence" and for the PDFs p and q of random variables with a common support D is defined as:

$$\boxed{\int_D\phi\left(\frac{p(y)}{q(y)}\right)\mathrm{d}y}\tag{9.5.109}$$

if the integral exists. The ϕ-divergence is also called the f-divergence. The ϕ-divergence is in general not a metric because it is not symmetric. One function is taken as the base from which the other function is measured. A specific instance of ϕ-divergence that is widely used is the Kullback-Leibler measure introduced just earlier:

$$\int_{\mathbb{R}}p(y)\log\left(\frac{p(y)}{q(y)}\right)\mathrm{d}y\tag{9.5.110}$$

9.5.2 Boltzmann law

First we will demonstrate through a simple case that for any system closed and isolated system, the most likely state is the state is that of equilibrium!

Let us consider an isolated system (a system is said to be "isolated" when it is impervious to any flow - heat (adiabatic), matter, fields, etc.) populated by N distinguishable particles. This system is divided into 2 compartments (or levels) identical and separated by an impermeable wall. Each compartment is assumed to contain a quantity N_1, N_2 number of particles.

For a given configuration of the system, we speak of "microstate" in the sense that it is possible thanks to the quantity of particles to measure a macroscopic known quantity such as energy, mass, pressure, etc.

If we fix this particular system, it is of course possible for number N of particles to imagine a given number of macro-states (don't forget that there are two compartments). Such as:

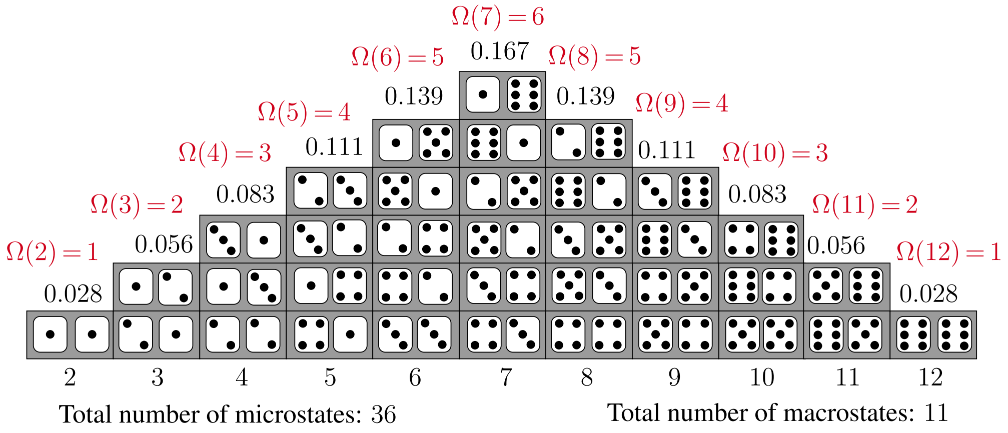
- One particle: 2 macrostate (2 possible configurations, therefore 1 configuration by macrostate)
- Two particles: 3 macrostate (4 possible configurations by permutations of the compartments
- Three particles: 4 macrostate (8 possible configurations by permutations of the compartments)
- Four particles: 5 macrostate (16 possible configurations by permutations of the compartments)
- n particles: $n+1$ macrostate (2^n possible configurations by permutations of the compartments

Definition (#9.5.456): We name "microstate", a configuration of permutation of macrostate. In the context of the Einstein model of a solid, the term "macrostate" refers to the total amount of energy in the system.

The probability of finding a system in a given state depends upon the multiplicity of that state. That is to say, it is proportional to the number of ways you can produce that state. Here a "state" is defined by some measurable property which would allow you to distinguish it from other states.

✍**Example:**

In throwing a pair of dice, that measurable property is the sum of the number of dots facing up. The multiplicity for two dots showing is just one, because there is only one arrangement of the dice which will give that state. The multiplicity for seven dots showing is six, because there are six arrangements of the dice which will show a total of seven dots:


Ω(7) = 6
0.167
Ω(6) = 5
Ω(8) = 5
0.139
0.139
Ω(5) = 4
Ω(9) = 4
0.111
0.111
Ω(4) = 3
Ω(10) = 3
0.083
0.083
Ω(3) = 2
Ω(11) = 2
0.056
0.056
Ω(2) = 1
Ω(12) = 1
0.028
0.028
2
3
4
5
6
7
8
9
10
11
12
Total number of microstates: 36
Total number of macrostates: 11

Let us now determine using combinatorial analysis (see section Probabilities page 522) the number of possible microstates for each macro-state named "multiplicity" and denoted by Ω. By analogy, this corresponds to imagine that the system is a stem on which are strung balls (particles) and the stem is separated by an imaginary boundary in one of its points (Chinese abacus). In such a situation, we have:

$$\Omega = \frac{N!}{N_1!N_2!} \tag{9.5.111}$$

This gives us all the possible arrangements of the "lefts particles" with "right particles" (of the border) for a given macro-state (the number of ways in which particles can be shared between the two compartments). But we also have in this particular case:

$$\Omega = \frac{N!}{N_1!N_2!} = \frac{N!}{N_1!(N-N_1)!} = \frac{N!}{(N-N_2)!N_2!} \tag{9.5.112}$$

But this corresponds to the combinatorial such that:

$$C_{N_1}^{N} = \frac{N!}{N_1!(N-N_1)!} \qquad C_{N_2}^{N} = \frac{N!}{(N-N_2)!N_2!} \tag{9.5.113}$$

and therefore:

$$\Omega = \frac{N!}{N_1!N_2} = C_{N_1}^{N} = C_{N_2}^{N} \tag{9.5.114}$$

We finally have for all macrostates of a system of N particles, a total of:

$$\sum_{N_1=0}^{N} C_{N_1}^{N} = \sum_{N_2=0}^{N} C_{N_2}^{N} \tag{9.5.115}$$

possible microstates (configurations). But we have seen in the initial example that:

$$\sum_{N_1=0}^{N} C_{N_1}^{N} = \sum_{N_2=0}^{N} C_{N_2}^{N} = 2^N \tag{9.5.116}$$

Thus, the probability of existence of a given microstate is 2^{-N} and is equally likely!!

We can now state the first postulate of Statistical Mechanics ("Gibbs postulate") all discernible and accessible microstates of an isolated system are equally likely.

Let us go back to our original question about equilibrium:

The notion of equilibrium associated to a macro-state is provided by classical thermodynamics. We say that a system is said to be in equilibrium when its state is characterized by the temporal independence of macroscopic quantities (mass, power, pressure, ...) and the constancy of the thermodynamic potentials (internal energy, enthalpy, Gibbs energy, ...).

To find out why the equilibrium is the most probable state, we only need to find what is the pair (N_1, N_2) that maximizes:

$$\Omega_{\text{max}} := \Omega = \frac{N!}{N_1!N_2!} \tag{9.5.117}$$

since all microstates are equally probable anyway. It is easy to check that this maximum is given for:

$$N_1 = N_2 \tag{9.5.118}$$

We can now announce the second postulate of Statistical Mechanics: the equilibrium state is the state which is the one with the largest number of configurations (microstates) and is the most likely state!!

Or in other words: A system reaches its equilibrium when its entropy is maximum and the entropy can only grow (fundamental postulate in physics!)!!

Remark

Some physicists have supposed that this fundamental postulate of increasing entropy is even stronger that the principle of the Darwin's natural selection model. Indeed, if we imagine molecules dispersed in a primordial ocean, by the principle of entropy, they will self-organize naturally to maximize the dissipation of heat from the ocean. What is powerful is that principle provides a physical basis for Darwin's natural selection phenomenon and helps explain some evolutionary trends that mere natural selection is not able to explain to this day: why some organisms have a characteristic X rather than Y is not due to the fact that X is better suited than Y but because the surrounding physical constraints makes that X is more likely to move towards maximum entropy than Y.

Let us now consider the following system:

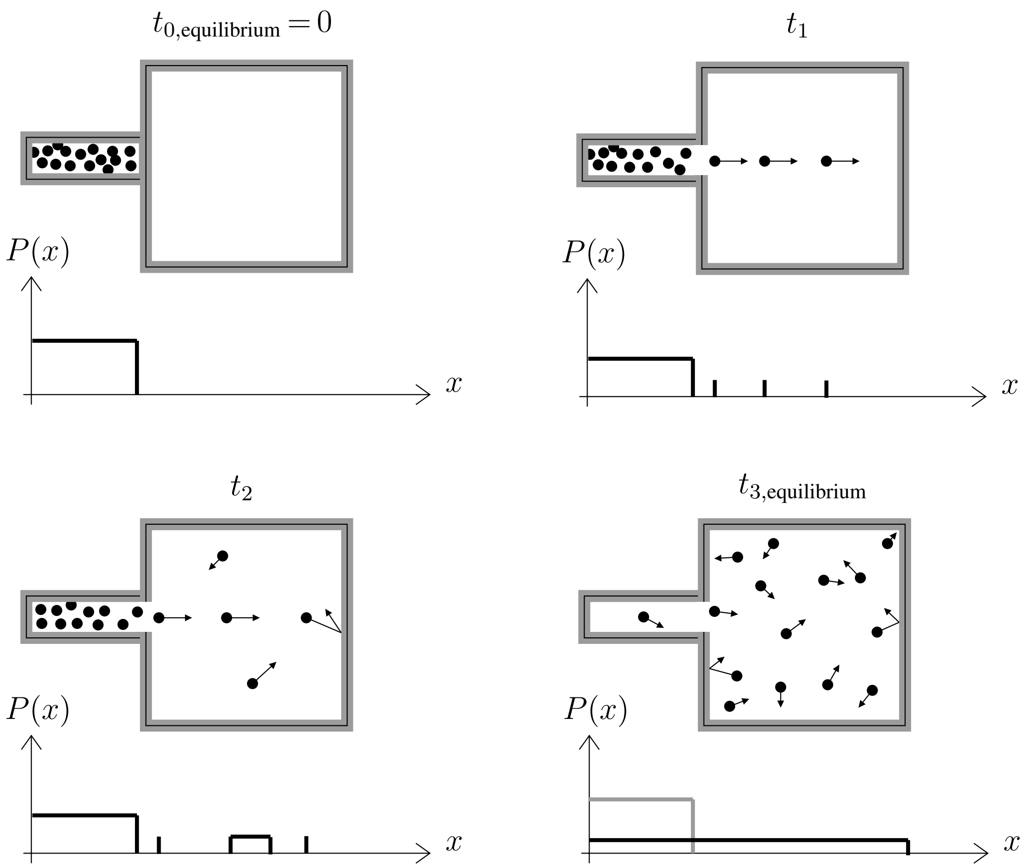


Figure 9.137: Example of change of entropy of a closed system

The distribution function $P(x)$ which describes the position of the particle along the x axis at the equilibrium $(t=0)$ will change to another distribution function corresponding to a new equilibrium at t_3. At equilibrium $P(x)$ is constant (uniformly distributed). But between the two equilibria, it evolves and becomes wider. So we lose the information on the position of particles. We can therefore re-state the second assumption, saying that a system that is out of equilibrium always operates in the sense of a loss of information (an enlargement of the corresponding probabilistic distribution function).

Meanwhile, the second principle of classical thermodynamics tells us that natural evolution must correspond to an increase of entropy $(\mathrm{d}S > 0)$. So there must be a close link between the information we have about the state of each particle and the entropy of the system.

The case we just described above shows clearly that the parameters or concepts: number of configurations, disorder, equilibrium, amount of information and entropy of an isolated system can be used to represent the state of a system. These parameters have the same function. Mathematical relations must therefore connect each other!

Let us recall that we have prove that the statistical information metric entropy of a system is given by:

$$S(X) = \mathrm{E}\{h(x)\} = -\lambda \sum_{i=1}^{n} p_i \log(p_i) \tag{9.5.119}$$

If we apply this relation to the case of a physical system in equilibrium for which we wish to calculate the entropy, we have proved earlier that:

$$p_1,p_2,\ldots,p_n=c^{te} \tag{9.5.120}$$

We still need to know what to what corresponds this constant probability. We have previously shown that at equilibrium, we had:

$$\Omega=\Omega_{\text{max}} \tag{9.5.121}$$

which is therefore the number of microstates at equilibrium. Thus, the probability of drawing a microstate among all of them is:

$$p=\frac{1}{\Omega_{\text{max}}} \tag{9.5.122}$$

that we simply write (a little dangerous ...) by tradition:

$$p=\frac{1}{\Omega} \tag{9.5.123}$$

We therefore have:

$$\begin{aligned}S&=-\lambda\sum_{i=1}^{n}p_i\log(p_i)=-\lambda\left(\frac{1}{\Omega}\log\left(\frac{1}{\Omega}\right)+\frac{1}{\Omega}\log\left(\frac{1}{\Omega}\right)+\ldots+\frac{1}{\Omega}\log\left(\frac{1}{\Omega}\right)\right)\\&=-\lambda\left(-\frac{1}{\Omega}\log(\Omega)-\frac{1}{\Omega}\log(\Omega)-\ldots-\frac{1}{\Omega}\log(\Omega)\right)=\lambda\log(\Omega)\left(\frac{1}{\Omega}+\frac{1}{\Omega}+\ldots+\frac{1}{\Omega}\right)\end{aligned} \tag{9.5.124}$$

As the probabilities of microstates are equally likely and that we sum over all of these, it comes:

$$\left(\frac{1}{\Omega}+\frac{1}{\Omega}+\ldots+\frac{1}{\Omega}\right)=1 \tag{9.5.125}$$

and therefore (without forgetting that in this special case the probabilities all have the same values!):

$$S=-\lambda\sum_{i=1}^{n}p_i\log(p_i)=\lambda\log(\Omega)=c^{te}\log(\Omega) \tag{9.5.126}$$

Since the equilibrium is related to the maximum disorder, and the disorder is related to the missing information, it seems reasonable to link the statistical entropy of information with statistical entropy of thermodynamics in physics. For this, we need that the constant allows us to get the right units and it comes naturally to choose this constant as it is equal to the Boltzmann constant k has the same units as the thermodynamic entropy. So:

$$S=k\log(\Omega) \tag{9.5.127}$$

We still have to choose the base of the logarithm. The experience shows that we must choose the natural logarithm that gives us the possibility to fall back on the results of Classical Mechanics after developments.

So we finally get the "Boltzmann law":

$$\boxed{S=k\ln(\Omega)} \tag{9.5.128}$$

which gives us the thermodynamic entropy of a system in equilibrium!

By the mathematical properties of the mean (see section Statistics page 571), especially the multiplication by a constant, we have for a set of N subsystems:

$$S(N\Omega)=NS(\Omega)=Nk\ln(\Omega)=-kN\sum_{i=1}^{n}p_i\log(p_i) \tag{9.5.129}$$

since as we pointed out earlier in our presentation of the Shannon formula, the entropy S is an expected mean (without forgetting that in this case the probabilities are of equal value!).

Before we continue we would like to quote John Wheeler because what he said in his book *Black Holes* is quite representative of the general trend of the physics of the 121st century (and some high level theories going also in this direction are quite promising):

quote
9.5:3554.#63

> ❞ *I think of my lifetime in physics as divided into three periods. In the first period... I was in the grip of the idea that Everything is Particles... I call my second period Everything is Fields. ... Now I am in the grip of a new vision, that Everything is Information* ❞
>
> John Archibald Wheeler

9.5.3 Statistical Physics Distributions

We will distinguish in this textbook 4 different statistics that come or not from quantum effects that will lead us to 4 well known distinct statistical distributions. These are the: Maxwell distribution, Maxwell-Boltzmann distribution, Fermi-Dirac distribution and Bose-Einstein distribution. They have numerous applications in physics as the black-body radiation that will be study in the section of Thermodynamics but also in Electrokinetic.

9.5.3.1 Maxwell Distribution (velocity distribution)

For a gas at equilibrium, let us ask the following question: What is the probability that a molecule has its velocity components between v_x and $v_x+\mathrm{d}x$, and v_y and $v_y+\mathrm{d}y$, and also v_z and $v_z+\mathrm{d}z$ in a conventional cartesian coordinate system?

This probability d^3P depends of $\vec{v}$ (that is to say of: (v_x,v_y,v_z)) and of $\mathrm{d}v_x,\mathrm{d}v_y,\mathrm{d}v_z$. It does not depend on the position of the molecule since the gas is assumed to equilibrium with respect to its center of mass!

We postulate in a first time that $\mathrm{d}P^3$ is proportional to each of the intervals $\mathrm{d}v_x,\mathrm{d}v_y,\mathrm{d}v_z$ such as:

$$\mathrm{d}^3P=f(\vec{v})\mathrm{d}v_x\mathrm{d}v_y\mathrm{d}v_z \tag{9.5.130}$$

and there is no preferred spatial directions. We can make a circular rotation of the Cartesian axes, the probability will be unchanged (isotropy of space).

The isotropy brings that the function $f(\vec{v})$ does not depend on vector $\vec{v}$, at best it dependent on the norm of the speed, such as:

$$\mathrm{d}^3P=f(v)\mathrm{d}v_x\mathrm{d}v_y\mathrm{d}v_z \tag{9.5.131}$$

Given $\mathrm{d}P_x$ (also often denoted $\mathrm{d}P(v_x)$ depending on the context) the probability for a molecule to have its component along the Ox axis between v_x and $v_x+\mathrm{d}v_x$ then:

$$\mathrm{d}P_x=\mathrm{d}P(v_x)=\varphi_x(v_x)\mathrm{d}v_x \tag{9.5.132}$$

also same with:

$$\mathrm{d}P_y=\mathrm{d}P(v_y)=\varphi_y(v_y)\mathrm{d}v_y \qquad \mathrm{d}P_z=\mathrm{d}P(v_z)=\varphi_z(v_z)\mathrm{d}v_z \tag{9.5.133}$$

The isotropy of space brings to write:

$$\varphi(v_x)=\varphi(v_y)=\varphi(v_z)=\varphi \tag{9.5.134}$$

The law of compound probabilities (see section Probabilities page 486) bring us to write:

$$\mathrm{d}^3P=\mathrm{d}P_x\mathrm{d}P_y\mathrm{d}P_z\Rightarrow f(v)=\varphi(v_x)\varphi(v_y)\varphi(v_z) \tag{9.5.135}$$

Determining the functions f and φ reveals within the following mathematical methodology:

$$\frac{\partial f}{\partial v_x}=\frac{\partial f}{\partial v}\frac{\partial v}{\partial v_x}=\frac{\partial}{\partial v_x}[\varphi(v_x)\varphi(v_y)\varphi(v_z)]=\varphi(v_y)\varphi(v_z)\frac{\mathrm{d}\varphi(v_x)}{\mathrm{d}v_x}=\frac{f}{\varphi(v_x)}\frac{\mathrm{d}\varphi(v_x)}{\mathrm{d}v_x} \tag{9.5.136}$$

Since:

$$v=\sqrt{v_x^2+v_y^2+v_z^2}\Rightarrow\frac{\partial v}{\partial v_x}=\frac{v_x}{\sqrt{v_x^2+v_y^2+v_z^2}}=\frac{v_x}{v} \tag{9.5.137}$$

Which gives us finally:

$$\frac{\mathrm{d}f}{\mathrm{d}v}\frac{\partial v}{\partial v_x}=\frac{v_x}{v}\frac{\mathrm{d}f}{\mathrm{d}v}=\frac{f}{\varphi(v_x)}\frac{\varphi(\mathrm{v_x})}{\mathrm{d}v_x}\Leftrightarrow\frac{1}{v}\frac{\mathrm{d}f}{\mathrm{d}v}=\frac{f}{v_x\varphi(v_x)}\frac{\mathrm{d}\varphi(v_x)}{\mathrm{d}v_x} \tag{9.5.138}$$

The left side of the last equality depends only on v, the right one only on v_x. The result can only be a constant that we will denote x_x. It follows that:

$$\frac{f}{v_x\varphi(v_x)}\frac{\mathrm{d}\varphi(v_x)}{\mathrm{d}v_x}=\alpha\leftrightarrow\frac{1}{\varphi(v_x)}\mathrm{d}\varphi(v_x)=\alpha v_x\mathrm{d}v_x \tag{9.5.139}$$

By integrating:

$$\int\frac{1}{\varphi(v_x)}\mathrm{d}\varphi(v_x)=\ln(\varphi(v_x))+c_1^{te}=\int\alpha v_x\mathrm{d}v_x=\frac{\alpha}{2}v_x^2+c_2^{te} \tag{9.5.140}$$

Therefore:

$$\ln(\varphi(v_x))=\frac{\alpha}{2}v_x^2+(c_2^{te}-c_1^{te})=\frac{\alpha}{2}v_x^2+c^{te} \tag{9.5.141}$$

and let us put $c^{te}=\ln(A)$. It comes therefore:

$$\ln(\varphi(v_x))=\frac{\alpha}{2}v_x^2+\ln(A)\Rightarrow\ln\left(\frac{\varphi(v_x)}{A}\right)=\frac{\alpha}{2}v_x^2\Rightarrow\varphi(v_x)=Ae^{\frac{\alpha}{2}v_x^2} \tag{9.5.142}$$

Identically, we have:

$$\varphi(v_x)=Ae^{\frac{\alpha}{2}v_x^2}\qquad\varphi(v_y)=Ae^{\frac{\alpha}{2}v_y^2}\qquad\varphi(v_z)=Ae^{\frac{\alpha}{2}v_z^2} \tag{9.5.143}$$

The law of compound probabilities involving:

$$f(v)=\varphi(v_x)\varphi(v_y)\varphi(v_z) \tag{9.5.144}$$

We finally have:

$$f(v)=A^3e^{\frac{\alpha}{2}(v_x^2+v_y^2+v_z^2)}=A^3e^{\frac{\alpha v^2}{2}} \tag{9.5.145}$$

Let us notice that α is necessarily negative otherwise the probability for a molecule to have an infinite velocity component would be infinite which would mean that all the molecules will have an infinite velocity and therefore an infinite energy!!!

We put:

$$\frac{\alpha}{2}=-B \tag{9.5.146}$$

with $B>0$.

Finally, it comes:

$$\mathrm{d}^3P=\mathrm{d}P_x\mathrm{d}P_y\mathrm{d}P_z=A^3e^{-Bv^2}\mathrm{d}v_x\mathrm{d}v_y\mathrm{d}v_z \tag{9.5.147}$$

It remains to us to normalize A. The probability of a molecule having a velocity component between speed $[-\infty,+\infty]$ or a velocity norm between $[0,+\infty]$ is equal to 1 (100% chance). So using exactly the same method computation than that seen in the section Statistics for the Normal distribution (see page 695), we have:

$$\int\limits_{-\infty}^{+\infty}\mathrm{d}P(v_x)=\int\limits_{-\infty}^{+\infty}Ae^{-Bv_x^2}\mathrm{d}v_x=A\int\limits_{-\infty}^{+\infty}e^{-Bv_x^2}\mathrm{d}v_x=1=A\sqrt{\frac{\pi}{B}}\Rightarrow A=\sqrt{\frac{B}{\pi}} \tag{9.5.148}$$

We have so far spoken in terms of probabilities. An equivalent language is to seek in a box containing N molecules, the number $\mathrm{d}N$ of molecules having given characteristics, that is to say for example, the number of molecules $\mathrm{d}N(v_x)$ having a velocity component between v_x and $v_x+\mathrm{d}v_2$. This number being obviously equal in one dimension to:

$$\mathrm{d}N(v_x)=N\mathrm{d}P_x=NA^{-Bv_x^2}\mathrm{d}v_x \tag{9.5.149}$$

More generally:

$$\mathrm{d}N(v)=N\mathrm{d}^3P=N\left(\frac{B}{\pi}\right)^{3/2}e^{-Bv^2}\mathrm{d}v_x\mathrm{d}v_y\mathrm{d}v_z \tag{9.5.150}$$

To get $\mathrm{d}P$ we put ourselves in the configuration space of velocities that is to say, in a Cartesian coordinate system of coordinates $(\mathrm{O}, v_x, v_y, v_z)$. The components (v_x, v_y, v_z) are not independent as related by the relation:

$$v = v_x^2 + v_y^2 + v_z^2 \tag{9.5.151}$$

The extremity of the velocity vectors $\vec{v}$ having a velocity norm of v, that is to say of velocity components bounded as the relation above, is in velocity space of configuration on a sphere of radius v. It will be the same for velocity vectors of having a norm equal to $v + \mathrm{d}v$.

In the velocity space configuration, we outline a portion of space between the sphere of radius v and the sphere of radius $v + \mathrm{d}v$, equal in volume to $4\pi v^2 \mathrm{d}v^2$.

The probability $\mathrm{d}P^3$ is proportional to $\mathrm{d}v_x \mathrm{d}v_y \mathrm{d}v_z$ that is to say to the elementary volume in the velocity configuration space. To get $\mathrm{d}P(v)$, we must integrate $\mathrm{d}P^3$ at all possible speeds vectors, that is to say having their extremities between the two spheres. This integration is particularly simple because the norm of $\vec{v}$ in this configuration space (inter-volume) is constant. So we get:

$$\mathrm{d}P(v) = \left(\frac{B}{\pi}\right)^{3/2} e^{-Bv^2} 4\pi v^2 \mathrm{d}v \tag{9.5.152}$$

and therefore, in a box containing N molecules, the number of molecules $\mathrm{d}N(v)$ have a norm between v and $v + \mathrm{d}v$ is equal to:

$$\mathrm{d}N(v) = N\mathrm{d}P(v) \tag{9.5.153}$$

Let us recall now that according to what we have studied in the section Statistics, the average value $\langle G \rangle$ of a quantity G is the product of G by all the probabilities of G integrated with all possible values of this quantity such that in one dimension we have (according to the x axis):

$$\begin{aligned}\langle v_x^2 \rangle &= \int_0^{+\infty} v_x^2 \mathrm{d}P(v_x) = \left(\frac{B}{\pi}\right)^{1/2} \int_0^{+\infty} v_x^2 e^{-Bv_x^2} \mathrm{d}v_x = \left(\frac{\sqrt{\pi}\operatorname{erf}(\sqrt{B}v_x)}{4B^{3/2}} - \frac{v_x e^{-Bv_x^2}}{2B}\right)\Bigg|_0^{+\infty} \\ &= \left(\frac{B}{\pi}\right)^{1/2} \frac{\sqrt{\pi}}{4B^{3/2}} = \frac{1}{4B}\end{aligned} \tag{9.5.154}$$

where the integral is calculated by making the change $v_x^2 = x$ and where we have then used an integration by parts (for the erf function see page 3697).

We have respectively generally :

$$\begin{aligned}\langle v^2 \rangle &= \int_0^{+\infty} v^2 P(v) \mathrm{d}v = 4\pi \left(\frac{B}{\pi}\right)^{3/2} \int_0^{+\infty} v^4 e^{-Bv^2} \mathrm{d}v = \left(\frac{3\sqrt{\pi}\operatorname{erf}(\sqrt{B}v)}{8B^{5/2}} - \frac{ve^{-Bv^2}(2Bv^2+3)}{4B^2}\right)\Bigg|_0^{+\infty} \\ &= 4\pi \left(\frac{B}{\pi}\right)^{3/2} \frac{3\sqrt{\pi}}{8B^{5/2}} = \frac{3}{2B}\end{aligned} \tag{9.5.155}$$

where the integral has been calculated in the same manner.

Therefore:

$$\langle v_x^2\rangle = \langle v_y^2\rangle = \langle v_z^2\rangle = \frac{1}{2B} \tag{9.5.156}$$

or:

$$\langle v_x^2\rangle + \langle v_y^2\rangle + \langle v_z^2\rangle = \frac{3}{2B} \tag{9.5.157}$$

Thermodynamics or Fluid Mechanics (see virial theorem page 3897) gives us for a perfect monoatomic gas (see the sections of Continuum Mechanics and Thermodynamics) of constant heat capacity and volume:

$$E_c = \frac{3}{2}NkT \tag{9.5.158}$$

If we hypothesize that the part of the energy associated with the temperature is due to the kinetic motion of molecules, we can write:

$$E_c = \frac{3}{2}NkT := N\frac{1}{2}m\langle v^2\rangle \Rightarrow \langle v^2\rangle = \frac{3kT}{m} \tag{9.5.159}$$

We then define the "average thermal velocity" or "root-mean-square (rms) speed":

$$\boxed{\langle v_{\text{th}}\rangle = \sqrt{\langle v^2\rangle} = \sqrt{\frac{3kT}{m}}} \tag{9.5.160}$$

A numerical application (the corresponding integral cannot be solved in closed form) for the electron in the case of semiconductors (where we approximate some relations by the Maxwell-Boltzmann distribution) gives an average thermal velocity at room temperature of $120,000\,[\text{m}\cdot\text{s}^{-1}]$.

But to return to our distribution .... So we have:

$$\langle v^2\rangle = \frac{3}{2B} = \frac{3kT}{m} \Rightarrow B = \frac{m}{2kT} \tag{9.5.161}$$

Then we finally have:

$$\boxed{\mathrm{d}P(v) = 4\pi v^2\left(\frac{m}{2\pi kT}\right)^{3/2} e^{-\frac{m}{2kT}v^2}\mathrm{d}v} \tag{9.5.162}$$

that is the velocity distribution in a monoatomic gas which plot looks like (the units in ordinate axes are arbitrary):

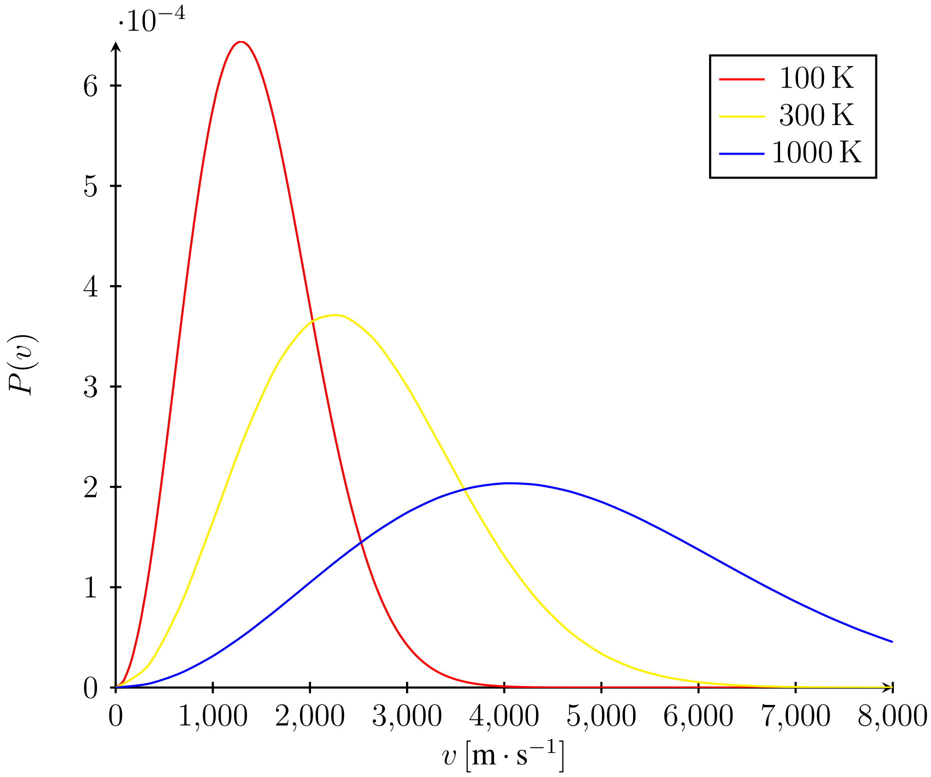


Figure 9.138: **Maxwell velocity distribution for a monoatomic ideal gas (author: Janosh Riebesell)**

Using the fact that $E=1/2mv^2$ (and therefore $E=mv\mathrm{d}v$) the previous relation is often used in the following shape:

$$\mathrm{d}P(E)=\frac{2}{\sqrt{\pi}}\frac{E^{1/2}}{(kT)^{3/2}}e^{-\frac{E}{kT}}\mathrm{d}E \qquad (9.5.163)$$

The previous relation thus gives the proportion of gas molecules having at a given time t a velocity v. We see that the probability density extends to infinity and we find a non-zero probability (approaching zero) that the particle have extremely high speed. In fact, if we have a system of identical particles at the same temperature T, then we will find a small given number of them with these very high speeds. At the heart of the stars it is these rare and very fast particles that are able to overcome the Coulomb repulsion between the nuclei and make them approach one another to initiate nuclear fusion mechanism which is the source energy of the star.

We get then for a single spatial direction and using a common notation in the English literature (by adopting the preceding developments we fall relatively easily on this result) the following distribution function:

$$f(v)=4\pi v^2\left(\frac{m}{2\pi kT}\right)^{3/2}e^{-\frac{m}{2kT}v^2} \qquad (9.5.164)$$

and for the three spatial dimensions still using the notation common in English literature (some terms are simply placed under the root):

$$f(v)=\sqrt{\frac{2}{\pi}\left(\frac{m}{kT}\right)^3}v^2e^{-\frac{m}{2kT}v^2} \qquad (9.5.165)$$

So this distribution function gives us the possibility to define the most likely velocity (the "modal" value as we say in statistics) sometimes denoted v_m, which is the maximum of the curve $f(v)$, or where the first derivative is equal to zero. Thus, for a single spatial dimension we have for the proportion of gas molecules:

$$\begin{aligned}\frac{\partial f(v)}{\partial v}&=\frac{\partial}{\partial v}\left(4\pi v^2\left(\frac{m}{2\pi kT}\right)^{1/2}e^{-\frac{mv^2}{2kT}}\right)\\&=8\pi v\left(\frac{m}{2\pi kT}\right)^{1/2}e^{-\frac{mv^2}{2kT}}-8\frac{8}{2kT}\pi v^3\left(\frac{m}{2\pi kT}\right)^{1/2}e^{-\frac{mv^2}{2kT}}\\&=8\pi v\left[1-\frac{2}{2kT}v^2\right]\left(\frac{m}{2\pi kT}\right)^{1/2}e^{-\frac{mv^2}{2kT}}=0\end{aligned}\tag{9.5.166}$$

Therefore (you notice that the result would have been the same if we had taken all spatial dimensions ... which is consistent!):

$$\begin{aligned}1-\frac{m}{2kT}v^2&=0\\\Rightarrow\frac{m}{2kT}v^2&=1\\\Rightarrow v^2&=\frac{2kT}{m}\end{aligned}\tag{9.5.167}$$

hence the most likely speed in one dimension (caution! this is not the maximum speed since we know that the distribution extends to the infinite speed):

$$v=\sqrt{\frac{2kT}{m}}\tag{9.5.168}$$

For the average speed (the mean) and using directly all spatial dimensions we have:

$$\begin{aligned}\bar{v}=\mathrm{E}(v)=\int_0^{+\infty}vf(v)\mathrm{d}v&=\int_0^{+\infty}v4\pi v^2\left(\frac{m}{2\pi kT}\right)^{3/2}e^{-\frac{mv^2}{2kT}}\mathrm{d}v\\=4\pi\left(\frac{m}{2\pi kT}\right)^{3/2}&\int_0^{+\infty}v^3e^{-\frac{m}{2kT}v^2}\mathrm{d}v\end{aligned}\tag{9.5.169}$$

We then have an integral of the type:

$$\int_0^{+\infty}x^3e^{-ax^2}\mathrm{d}x\tag{9.5.170}$$

The best is to decompose $x^3e^{-ax^2}$ in $x^2xe^{-ax^2}$ and then we integrate by parts. Then we have:

$$\int_0^{+\infty}x^2\left(xe^{-ax^2}\right)\mathrm{d}x=\underbrace{-x^2\frac{1}{2a}e^{-ax^2}\Big|_0^{+\infty}}_{=0}+\int_0^{+\infty}2x\frac{1}{2a}e^{-ax^2}\mathrm{d}x\tag{9.5.171}$$

The last integral is easily calculable. This is the same as the one we already solved in the section Statistics for our introduction of the Normal law (see page 695). So:

$$\int_0^{+\infty}\frac{1}{a}xe^{-ax^2}\mathrm{d}x=\frac{1}{a}\left(-\frac{1}{2a}e^{-ax^2}\Big|_0^{+\infty}\right)=\frac{1}{2a^2}\tag{9.5.172}$$

Therefore it comes:

$$\begin{aligned}\bar{v}&=4\pi\left(\frac{m}{2\pi kT}\right)^{3/2}\int\limits_0^{+\infty}v^3e^{-\frac{m}{2kT}v^2}\mathrm{d}v=4\pi\left(\frac{m}{2\pi kT}\right)^{3/2}\frac{1}{2\left(\frac{m}{2kT}\right)^2}\\&=\left(\frac{m}{2\pi kT}\right)^{3/2}\frac{8\pi k^2T^2}{m^2}=\sqrt{\frac{8kT}{\pi m}}\end{aligned} \tag{9.5.173}$$

but this time the result is dependent on the number of spatial dimensions that we take!

9.5.3.2 Maxwell-Boltzmann Distribution

In corpuscular quantum physics, we learn that the energy of a particle is quantized, that is, the possible values for the energy form a discrete spectrum (see section Corpuscular Quantum Physics page 4627). Even if, in a number of common situations, for a particle and even more for a system consisting of a large number of particles, the energy levels are so tight that we can handle, mathematically, this spectrum as being continuous ("continuum approximation"), the fact remains that they are rigorously quantified. This quantified approach of physical distributions at the corpuscular level of the material is often referred as "quantum statistics".

A particle having a level of energy E_i, can be in different sub-states.

We know (we did not use the Schrödinger equation in the section of Corpuscular Quantum Physics to prove this strictly) that, to describe an atom, we introduce four quantum numbers, namely:

1. the principal quantum number that quantifies energy $n=n_r+n_\theta$
2. the azimuthal quantum number that quantifies the momentum n_θ
3. the quantum magnetic number that quantifies the magnetic moment m_l
4. the tree spin that quantifies the "proper rotation" (...) of the electrons s

Thus for a same energy (for a particular value of the principal quantum number), an atom or an electron can have different values of the secondary, magnetic or spin quantum numbers.

To qualify the possibility of sub-states corresponding to a same energy, we use the term "degeneration" and we translate by the variable g_i the number of degeneration corresponding to the same level of energy E_i.

We will consider a system of N particles that are placed on K different energy levels E_i. We find n_i particles on the energy level E_i. We then always have the following relations for the total energy (which we denote by the same letter as that used in thermodynamics) and for the number of particles:

$$\begin{aligned}&\sum_{i=1}^{K} n_i E_i = U\\&\sum_{i=1}^{K} n_i = N \Rightarrow \sum_{i=1}^{K} \mathrm{d}n_i = 0\end{aligned} \tag{9.5.174}$$

We will assume (this is important) these quantities as being constant and that the system is completely determined by the distribution $\{n_i\}$ of the particles on the K different energy levels.

It exist many possible microscopic configurations - also named "complexion" - Ω that are compatible with the distribution $\{n_i\}$. Their number is given by (see permutations with repetition in the section of Probabilities page 525):

$$\Omega = \frac{N!}{n_1!n_2!\ldots n_K!} \tag{9.5.175}$$

But we have neglected the possible degeneration of the levels i. If there exist g_i sublevels with an energy E_i, we have then:

$$\Omega = N!\frac{g_1^{n_1}g_2^{n_2}\ldots g_K^{n_K}}{n_1!n_2!\ldots n_K!} = N!\prod_i \frac{g_i^{n_i}}{n_i!} \tag{9.5.176}$$

This can be checked with the following figure depicting a unique energy level for two distinguishable particles having three possible (quantum) states (i.e. sublevels):

$$\Omega_{\text{M.B.}} = N!\frac{g^n}{n!} = 2!\frac{3^2}{2!} = 9$$

Example:
2 particles: A and B
3 states: 1, 2 and 3

9 distinct states

1	2	3
AB	$\ldots$	$\ldots$
$\ldots$	AB	$\ldots$
$\ldots$	$\ldots$	AB
A	B	$\ldots$
B	A	$\ldots$
A	$\ldots$	B
B	$\ldots$	A
$\ldots$	A	B
$\ldots$	B	B

Discernable ($A \neq B$)

Figure 9.139: $\Omega_{\text{M.B.}}$ **states in the Maxwell-Boltzmann distribution**

Remark

The number g_i is then the degeneration of the energy state E_i, i.e. the number of states with this energy.

Taking the logarithm, it comes:

$$\ln(\Omega)=\ln(N!)+\sum_{i=1}^{K}(n_i\ln(g_i)-\ln(n_i!)) \tag{9.5.177}$$

and using Stirling's formula (see section Numerical Methods page 6301):

$$n!=e^{n\ln(n)}e^{-n}\Rightarrow\ln(n!)\cong n\ln(n)-n \tag{9.5.178}$$

keeping in mind that the latter is a rough approximation for smaller values of n but becomes an acceptable approximation for n greater than $1,000$. Then we can write:

$$\ln(\Omega)=N\ln(N)-N+\sum_{i=1}^{K}(n_i\ln(g_i)-n_i\ln(n_i)+n_i)=N\ln(N)+\sum_{i=1}^{K}n_i\ln\left(\frac{g_i}{n_i}\right) \tag{9.5.179}$$

We now look for the most probable distribution, that is to say one that maximizes Ω (and thus that implicitly maximizes entropy as it can only increase as show us every day life experience so far for an isolated system). To find the extremum, we will differentiate this expression such that:

$$\frac{\partial\ln(\Omega)}{\partial n_i}=0 \tag{9.5.180}$$

so all other parameters other than n_i are fixed.

But, as:

$$\begin{gathered}\sum_{i=1}^{K}n_iE_i=U\Leftrightarrow U-\sum_{i=1}^{K}n_iE_i=0\Leftrightarrow\beta\left(U-\sum_{i=1}^{K}n_iE_i\right)=0\\ \sum_{i=1}^{K}n_i=N\Leftrightarrow N-\sum_{i=1}^{K}n_i=0\Leftrightarrow\alpha\left(N-\sum_{i=1}^{K}n_i\right)=0\end{gathered} \tag{9.5.181}$$

where α,β are constants that help to ensure consistency of dimensional analysis (in other words: that the units are coherent ...).

It is therefore equivalent and necessary to take into account also these intrinsic parameters (the fact to in such a clever way null terms is named "method of Lagrange multipliers" as seen in the section of Theoretical Computing) to write:

$$\frac{\partial\ln(\Omega)}{\partial n_i}=0 \tag{9.5.182}$$

in a more complete form below:

$$\frac{\partial}{\partial n_i}\left(\ln(\Omega)+\alpha\left(N-\sum_{i=1}^{K}n_i\right)+\beta\left(U-\sum_{i=1}^{K}n_iE_i\right)\right)=0 \tag{9.5.183}$$

So after rearrangement:

$$\frac{\partial \ln(\Omega)}{\partial n_i}=-\frac{\partial}{\partial n_i}\left[\alpha\left(N-\sum_{i=1}^{K}n_i\right)+\beta\left(U-\sum_{i=1}^{K}n_iE_i\right)\right] \tag{9.5.184}$$

giving after a first simplification (elimination of the derivative of constants):

$$\frac{\partial \ln(\Omega)}{\partial n_i}=\frac{\partial}{\partial n_i}\left[\alpha\sum_{i=1}^{K}n_i+\beta\sum_{i=1}^{K}n_iE_i\right]=\frac{\partial}{\partial n_i}\sum_{i=1}^{K}\left[\alpha n_i+\beta n_iE_i\right] \tag{9.5.185}$$

Which brings us to write (since all the terms of the sum indexed on a different value i from that of the partial derivative are equal zero):

$$\frac{\partial \ln(\Omega)}{\partial n_i}\frac{\partial}{\partial n_i}(\alpha n_i+\beta n_iE_i)=\alpha+\beta E_i \tag{9.5.186}$$

Which gives us finally:

$$\frac{\partial \ln(\Omega)}{\partial n_i}=\alpha+\beta E_i \tag{9.5.187}$$

which will also therefore have to be equal zero. Thus, we have cleverly show a result which is nonetheless very powerful. By imposing to maximize the entropy we revealed two terms (each including a constant that we will have to determine later!) and whose sum has to be equal zero to ensure the experimental validity of theoretical model when the entropy of the isolated system is maximal.

But this latter relation is equivalent as to seek (which will be very useful later!):

$$\mathrm{d}\ln(\Omega)=\mathrm{d}\left(N\ln(N)+\sum_{i=1}^{K}n_i\ln\left(\frac{g_i}{n_i}\right)\right)=0 \tag{9.5.188}$$

Indeed, let's use again the clever method of Lagrange multipliers as seen in the section of Theoretical Computing (adding zero terms assuming energy levels E_i and the number degeneration g_i as fixed) and remembering (see section Differential and Integral Calculus page 1534) that :

$$\frac{\mathrm{d}\left(x\ln\left(\frac{a}{x}\right)\right)}{\mathrm{d}x}=\ln\left(\frac{a}{x}\right)-1\Leftrightarrow \mathrm{d}\left(x\ln\left(\frac{a}{x}\right)\right)=\ln\left(\frac{a}{x}\right)\mathrm{d}x-\mathrm{d}x \tag{9.5.189}$$

Therefore we have:

$$\begin{aligned}
&\mathrm{d}\left(N\ln(N)+\sum_{i=1}^{K}n_i\ln\left(\frac{g_i}{n_i}\right)+\alpha\left(N-\sum_{i=1}^{K}n_i\right)+\beta\left(U-\sum_{i=1}^{K}n_iE_i\right)\right)=0\\
&\Rightarrow\sum_{i=1}^{K}\mathrm{d}n_i\ln\left(\frac{g_i}{n_i}\right)-\underbrace{\sum_{i=1}^{K}\mathrm{d}n_i}_{\to 0}-\alpha\sum_{i=1}^{K}\mathrm{d}n_i-\beta\sum_{i=1}^{K}E_i\mathrm{d}n_i=\sum_{i=1}^{K}\left[\left(\ln\left(\frac{g_i}{n_i}\right)-\alpha-\beta E_i\right)\mathrm{d}n_i\right]=0\\
&\Rightarrow\left(\ln\left(\frac{g_i}{n_i}\right)-\alpha-\beta E_i\right)\mathrm{d}n_i=0\\
&\Rightarrow\ln\left(\frac{g_i}{n_i}\right)=\alpha+\beta E_i\\
&\Rightarrow\ln\left(\frac{n_i}{g_i}\right)=-\alpha-\beta E_i
\end{aligned} \tag{9.5.190}$$

Therefore (relation often named "Boltzmann distribution"):

$$n_i = g_i e^{-\alpha-\beta E_i} \tag{9.5.191}$$

Let us recall that we have for the entropy of a system:

$$S = k\ln(\Omega) \tag{9.5.192}$$

and therefore:

$$\mathrm{d}S = k\mathrm{d}(\ln(\Omega)) = k\sum_i \frac{\partial \ln(\Omega)}{\partial n_i}\mathrm{d}n_i \tag{9.5.193}$$

and as:

$$\frac{\partial \ln(\Omega)}{\partial n_i} = \alpha + \beta E_i \tag{9.5.194}$$

Then we have:

$$\frac{\mathrm{d}S}{k} = \sum_{i=1}^{K}(\alpha+\beta E_i)\mathrm{d}n_i = \alpha\sum_{i=1}^{K}\mathrm{d}n_i + \beta\sum_{i=1}^{K}E_i\mathrm{d}n_i = \alpha\mathrm{d}N + \beta\mathrm{d}U \tag{9.5.195}$$

Thermodynamics leads us to (see section Thermodynamics page 3660):

$$\mathrm{d}U = T\mathrm{d}S + \sum_{i=1}^{K}\mu_i\mathrm{d}N_i \tag{9.5.196}$$

So we see appearing the chemical potential μ which has an influences when the number of particles considered varies (i.e. the chemical potential is defined as the change in the system energy on adding a particle).

If all the particles (or particle groups) are identical $\mu_i = c^{te} = \mu$ thus:

$$\mathrm{d}U = T\mathrm{d}S + \mu\sum_{i=1}^{K}\mathrm{d}N_i = T\mathrm{d}S + \mu\mathrm{d}N \tag{9.5.197}$$

Which finally bring us to:

$$\frac{\mathrm{d}S}{k} = \frac{\mathrm{d}U}{kT} - \frac{\mu}{kT}\mathrm{d}N \tag{9.5.198}$$

and bring is to identify:

$$\frac{\mathrm{d}S}{k} = \frac{\mathrm{d}U}{kT} - \frac{\mu}{kT}\mathrm{d}N \equiv \beta\mathrm{d}U + \alpha\mathrm{d}N \tag{9.5.199}$$

and therefore:

$$\alpha = -\frac{\mu}{kT} \quad \text{and} \quad \beta = \frac{1}{kT} \tag{9.5.200}$$

We can then rewrite the relation:

$$n_i = g_i e^{-\alpha-\beta E_i} \tag{9.5.201}$$

in the final traditional form of "Maxwell-Boltzmann function":

$$\boxed{n_i=g_ie^{-\frac{E_i+\mu}{kT}}=g_ie^{\frac{\mu}{kT}}e^{\frac{-E_i}{kT}}} \tag{9.5.202}$$

where $e^{\frac{-E_i}{kT}}$ is often named the "Boltzmann factor".

Example:

In a sample at room temperature (about $300\,[\text{K}]$), almost all atoms are in their electronic ground state. For example, if the first excited state is $1\,[\text{eV}]$ above the ground state, the Boltzmann factor is:

$$e^{-E/kT}=e^{-\frac{1.6\cdot 10^{-19}\,[\text{J}]}{(1.4\cdot 10^{-25}\,[\text{J}\cdot\text{K}^{-1}])(300\,[\text{K}])}}=e^{-38}=3\cdot 10^{-17}$$

This extremely tiny number represents the fact that at room temperature an insignificant fraction of the atoms will be excited thermally to $1\,[\text{eV}]$ above the ground state. Therefore, it is reasonable to assume that essentially all the atoms in an ordinary room-temperature absorption measurement start out in the electronic ground state.

On the other hand, the surface temperature of the Sun (actually a G2V-type star, aka yellow dwarf) is about $5,000\,[\text{K}]$, and with $E=1\,[\text{eV}]$ the Boltzmann factor is $e^{-E/kT}=e^{-2.3}=0.1$, so a sizable fraction of atoms can be in an excited state $1\,[\text{eV}]$ above the ground state, due to thermal excitation.

That we find most frequently in the literature as follows:

$$\boxed{n_i=g_ie^{\frac{\mu-E_i}{kT}}} \tag{9.5.203}$$

and especially in this form:

$$\boxed{f_{MB}(E,T,\mu)=e^{\frac{\mu-E_i}{kT}}} \tag{9.5.204}$$

We will use this relation in our study of semiconductor theory (see section Electrokinetics page 4279) as well as in our study of plasmas (see section Continuum Mechanics page 3945).

As $N=\sum_{i=1}^{K}n_i$, we have:

$$N=e^{\frac{\mu}{kT}}\sum_{i=1}^{K}g_ie^{\frac{-E_i}{kT}} \tag{9.5.205}$$

We can then calculate n_i/N and we get the discrete formulation of the "Maxwell-Boltzmann statistics":

$$\boxed{P=\frac{n_i}{N}=\frac{g_ie^{\frac{-E_i}{kT}}}{\sum\limits_{i=1}^{K}g_ie^{\frac{-E_i}{kT}}}} \tag{9.5.206}$$

This relation therefore gives the ratio (or "proportion" or "probability") between the number of particles which by hypothesis does not interact with each other (otherwise the model is too complex) in a given energy state E_i and the number of particles that do not interact with each other and can take different discrete energy states E_i. Thus, knowing N, it is possible using this relation to know the proportion of particles in a particular energy state. Obviously, if there is only one sub energy level, this relation reduces to (it is also in this form that we find it often in the literature):

$$P=\frac{n_i}{N}=\frac{e^{\frac{-E_i}{kT}}}{\sum\limits_{i=1}^{K} e^{\frac{-E_i}{kT}}} \tag{9.5.207}$$

Remark

The Maxwell-Boltzmann statistics applies in the absence of interaction between particles and is therefore valid for an ideal gas, but does not apply, for example, for a liquid. Moreover, it applies only to high temperatures when quantum effects are negligible. At low temperatures we use the Bose-Einstein statistics for bosons (referring to the name of Bose...) and the Fermi-Dirac statistics for fermions (referring to the name of Fermi...) that we will study and prove further below.

We name the term in the denominator, the "canonical partition function". It is most often denoted Z_c such that:

$$P=\frac{n_i}{N}=\frac{e^{\frac{-E_i}{kT}}}{Z_c} \tag{9.5.208}$$

and defined in the most general case by:

$$Z_c=\sum_{i=1}^{K} g_i e^{-\beta E_i} \tag{9.5.209}$$

Therefore (obvious):

$$\frac{n_i}{N}=\frac{e^{\frac{-E_i}{kT}}}{Z_c} \Leftrightarrow n_i=\frac{N}{Z}g_i e^{-\beta E_i} \tag{9.5.210}$$

The microcanonical, canonical and grand canonical sets correspond to systems subject to various constraints which are respectively:

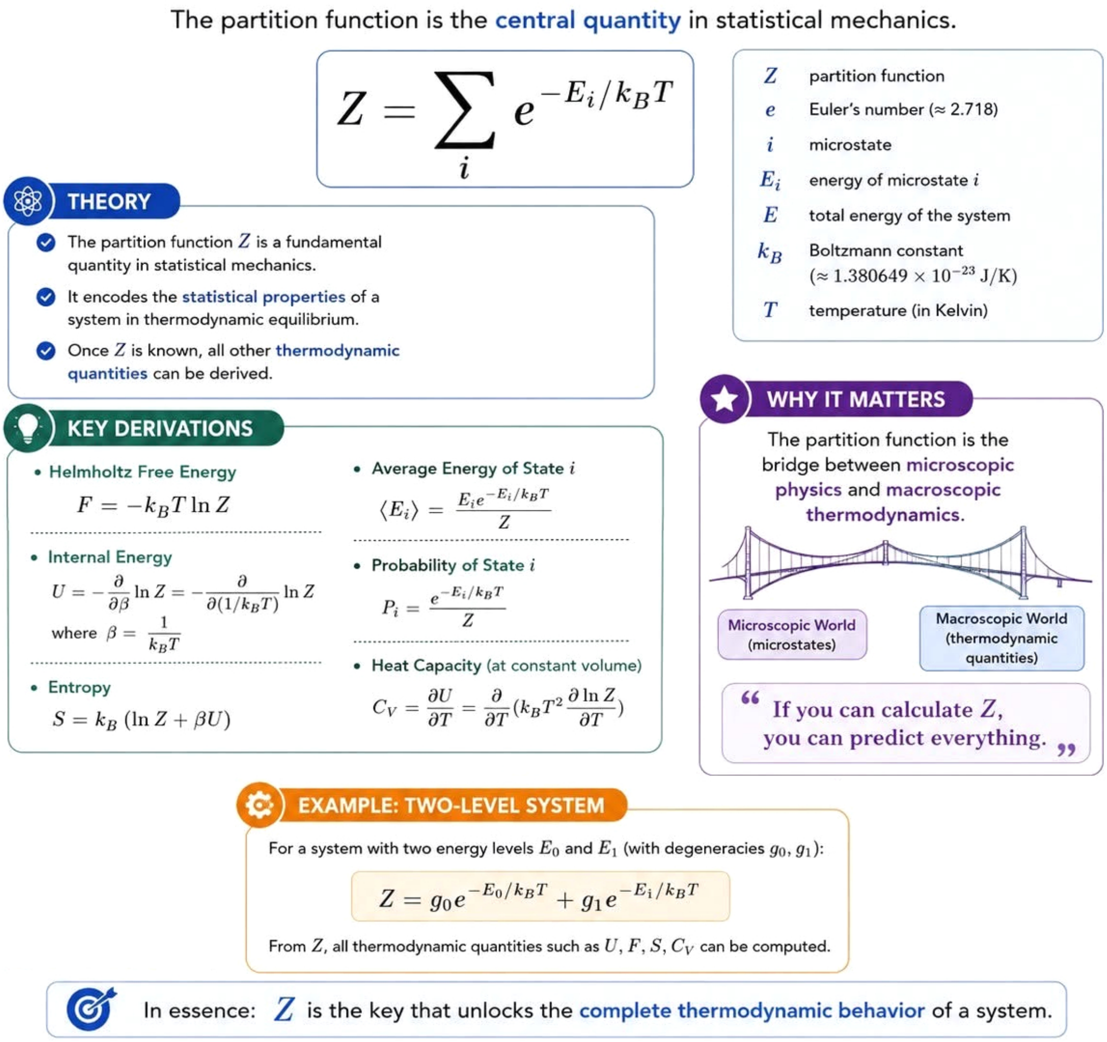
- Microcanonical: fixed energy, fixed number of particles, fixed volume
- Canonical: fixed mean energy = fixed temperature, fixed number of particles, fixed volume
- Grand canonical: fixed mean energy = temperature fixed, average number of particles fixed = fixed chemical potential[21], fixed volume

---

[21] Let us remind that the chemical potential is defined as the change in the system energy on adding a particle

This is true in quantum physics as in classical physics.

Here is an infographic summarizing the future implications of this canonical partition function, which we will demonstrate in detail later in this textbook.



The partition function is the **central quantity** in statistical mechanics.

$$Z=\sum_i e^{-E_i/k_BT}$$

Z	partition function
e	Euler's number (≈ 2.718)
i	microstate
E_i	energy of microstate i
E	total energy of the system
k_B	Boltzmann constant ($\approx 1.380649\times 10^{-23}\,\mathrm{J/K}$)
T	temperature (in Kelvin)

THEORY

- ✓ The partition function Z is a fundamental quantity in statistical mechanics.
- ✓ It encodes the **statistical properties** of a system in thermodynamic equilibrium.
- ✓ Once Z is known, all other **thermodynamic quantities** can be derived.

KEY DERIVATIONS

- **Helmholtz Free Energy**

$$F=-k_BT\ln Z$$

- **Internal Energy**

$$U=-\frac{\partial}{\partial\beta}\ln Z=-\frac{\partial}{\partial(1/k_BT)}\ln Z$$

where $\beta=\dfrac{1}{k_BT}$

- **Entropy**

$$S=k_B\left(\ln Z+\beta U\right)$$

- **Average Energy of State i**

$$\langle E_i\rangle=\frac{E_ie^{-E_i/k_BT}}{Z}$$

- **Probability of State i**

$$P_i=\frac{e^{-E_i/k_BT}}{Z}$$

- **Heat Capacity** (at constant volume)

$$C_V=\frac{\partial U}{\partial T}=\frac{\partial}{\partial T}(k_BT^2\frac{\partial\ln Z}{\partial T})$$

WHY IT MATTERS

The partition function is the bridge between **microscopic physics** and **macroscopic thermodynamics**.

> "If you can calculate Z, you can predict everything."

EXAMPLE: TWO-LEVEL SYSTEM

For a system with two energy levels E_0 and E_1 (with degeneracies g_0, g_1):

$$Z=g_0e^{-E_0/k_BT}+g_1e^{-E_1/k_BT}$$

From Z, all thermodynamic quantities such as U, F, S, C_V can be computed.

In essence: Z is the key that unlocks the **complete thermodynamic behavior** of a system.

9.5.3.3 Boltzmann Distribution

Let us now see a first interesting application of the Maxwell-Boltzmann distribution function.

Consider the interesting case where :

- The volume and number of particles are fixed (microcanonical system)
- The fact that the number of particles is fixed, the chemical potential[22] is zero
- A temperature high enough so that the energy states are in endless quantity

---

[22] Let us remind that the chemical potential is defined as the change in the system energy on adding a particle.

It is quite obvious that if the energy states are in infinite quantity, the degeneration factor g_i will be equal to the unit (because there is no place for degeneration levels other than the level itself).

We then have the Maxwell-Boltzmann distribution that becomes in theses condition the "Boltzmann distribution":

$$P=\frac{e^{-\frac{E}{kT}}}{Z_{mc}}=\frac{1}{Z_{mc}}e^{-\frac{E}{kT}} \tag{9.5.211}$$

that we can find more of the in the literature under the form:

$$\boxed{f_B(E,T)=e^{-\frac{E}{kT}}} \tag{9.5.212}$$

If we consider that in a broadly neutral gas, energy is defined only by the kinetic energy (in the case of the ideal gas!).

So with the transition to the continuum, we have:

$$Z_{mc}=\int\limits_{-\infty}^{+\infty} e^{-\frac{E}{kT}}\mathrm{d}v \tag{9.5.213}$$

Therefore, we can integrate the Boltzmann function above on all speeds and we should have:

$$P(-\infty<v<+\infty)=\int\limits_{-\infty}^{+\infty}\frac{1}{Z_{mc}}e^{-\frac{E}{kT}}\mathrm{d}v=\frac{1}{Z_{mc}}\int\limits_{-\infty}^{+\infty} e^{-\frac{\frac{1}{2}mv^2}{kT}}\mathrm{d}v=1 \tag{9.5.214}$$

With the microcanonical partition function:

$$Z_{mc}=\int\limits_{-\infty}^{+\infty} e^{-\frac{mv^2}{2kT}}\mathrm{d}v \tag{9.5.215}$$

for the normalization condition to be satisfied (and anyway it matches to the developments made earlier above!).

we recognize here an integral that is familiar to and already proven in the section Statistics during our study of the Normal distribution. Then we have:

$$Z_{mc}=\sqrt{\frac{2\pi kT}{m}} \tag{9.5.216}$$

Therefore, the probability density function can be written:

$$P(-\infty<v<+\infty)=\frac{1}{\sqrt{\frac{2\pi kT}{m}}}\int\limits_{-\infty}^{+\infty} e^{-\frac{mv^2}{2kT}}\mathrm{d}v \tag{9.5.217}$$

Now, since the speed is treated as a random variable, let us focus on the calculation of the average value of the square of the speed which traditionally is denoted as follows (the notation has no relation to the ket-bra notation as in Quantum Physics or in Functional Analysis with the dot!):

$$\langle v^2\rangle = \frac{1}{\sqrt{\dfrac{2\pi kT}{m}}}\int\limits_{-\infty}^{+\infty} v^2 e^{-\frac{mv^2}{2kT}}\,\mathrm{d}v \tag{9.5.218}$$

In statistics we would write this:

$$\mathrm{E}(v^2) = \frac{1}{\sqrt{\dfrac{2\pi kT}{m}}}\int\limits_{-\infty}^{+\infty} v^2 e^{-\frac{mv^2}{2kT}}\,\mathrm{d}v \tag{9.5.219}$$

Let us focus on:

$$\int\limits_{-\infty}^{+\infty} v^2 e^{-\frac{mv^2}{2kT}}\,\mathrm{d}v \tag{9.5.220}$$

But if we put:

$$\sigma = \sqrt{\frac{2kT}{m}} \tag{9.5.221}$$

Then we have:

$$\int\limits_{-\infty}^{+\infty} v^2 e^{-\left(\frac{v}{\sigma}\right)^2}\,\mathrm{d}v \tag{9.5.222}$$

Thus with a simple change of variable:

$$\sigma^3 \int\limits_{-\infty}^{+\infty} y^2 e^{y^2}\,\mathrm{d}y \tag{9.5.223}$$

and this is exactly the integral of the Normal centered law as proved in the section Statistics as being equal to:

$$\int\limits_{-\infty}^{+\infty} v^2 e^{-\left(\frac{v}{\sigma}\right)^2}\,\mathrm{d}v = \sigma^2\frac{\sqrt{\pi}}{2} \tag{9.5.224}$$

Therefore:

$$\begin{aligned}\mathrm{E}(v^2) &= \frac{1}{\sqrt{\dfrac{2\pi kT}{m}}}\int\limits_{-\infty}^{+\infty} v^2 e^{-\frac{mv^2}{2kT}}\,\mathrm{d}v = \frac{1}{\sqrt{\dfrac{2\pi kT}{m}}}\left(\frac{2kT}{m}\right)^{3/2}\frac{\sqrt{\pi}}{2}\\ &= \frac{1}{\sqrt{\dfrac{2kT}{m}}}\left(\frac{2kT}{m}\right)\left(\frac{2kT}{m}\right)^{1/2}\frac{1}{2} = \left(\frac{2kT}{m}\right)\frac{1}{2} = \frac{kT}{m}\end{aligned} \tag{9.5.225}$$

and finally:

$$\langle E\rangle=\frac{1}{2}m\langle v^2\rangle=\frac{1}{2}m\frac{kT}{m}=\frac{1}{2}kT \tag{9.5.226}$$

for a particle moving in one spatial dimension. We therefore fall back on the same result as the Maxwell distribution of speeds derived at page 3558 in the three multidimensional case but with a different mathematical approach.

9.5.3.4 Fermi-Dirac Distribution

The principle of indistinguishability can have very important consequences for the statistics. We distinguish two types of indistinguishable particles: bosons (like the photon) and fermions (like the electron).

Let us make two recalls:

1. The bosons correspond to particles whose representative wave function is always symmetrical, while that of fermions is antisymmetric.
2. The Pauli exclusion principle requires that two fermions can not be in the same quantum state. But the bosons can!

Their respective properties have for important consequence that the minimum energy of a set of N bosons is equal to N times the minimum energy of each boson. While for a set of fermions, the minimum energy is equal to the sum of the N lower energies.

This both types of particles cause two kinds of statistics: Fermi-Dirac statistics for fermions (that we will prove first) and the Bose-Einstein statistics for the bosons (which follows).

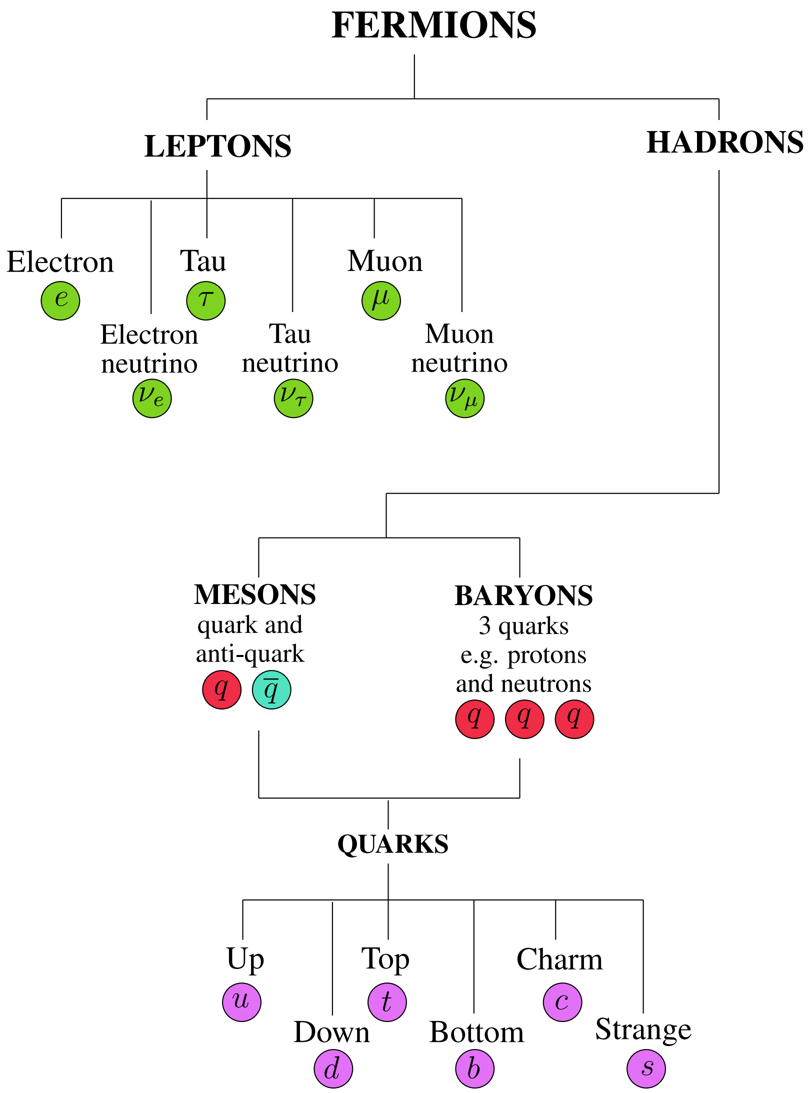

FERMIONS
LEPTONS
HADRONS
Electron
e
Tau
τ
Muon
μ
Electron neutrino
ν_e
Tau neutrino
ν_τ
Muon neutrino
ν_μ
MESONS
quark and anti-quark
q $\bar{q}$
BARYONS
3 quarks
e.g. protons and neutrons
q q q
QUARKS
Up
u
Top
t
Charm
c
Down
d
Bottom
b
Strange
s

Figure 9.140: **Fermions, Hadrons, Leptons, Mesons, Baryons, Quarks**

So there is only one way to distribute N fermions on the accessible energy states (instead of the $N!$ for distinguishable particles). It can not be more than n_i particles in an energy level E_i there are degenerations levels g_i. So:

$$0 \leq n_i \leq g_i \tag{9.5.227}$$

The number of possible combinations for a level E_i, degenerated g_i times and having particle n_i particle is then the combinatorial $C_{n_i}^{g_i}$. The total number of configurations is therefore:

$$\Omega = \prod_i C_{n_i}^{g_i} = \prod_i \frac{g_i!}{n_i!(g_i - n_i)!} \tag{9.5.228}$$

This can be checked with the following figure:

$$\Omega_{\text{F.D.}} = \frac{g!}{n!(g-n)!} = \frac{3!}{2!(3-2)!} = 3$$

Example:
2 particles: A and B
3 states: 1, 2 and 3

$$\text{3 distinct states}\left\{\begin{array}{ccc}\hline 1 & 2 & 3\\ \hline A & A & \dots \\ A & \dots & A\\ \dots & A & A\end{array}\right\}\text{Indistinguishable } (A=B)$$

Figure 9.141: $\Omega_{\text{B.E.}}$ **states in the Fermi-Dirac distribution**

The statistics is therefore very different from the classical Maxwell-Boltzmann case. By taking the logarithm of the number of micro-states and making use of the Stirling's formula approximation as above, we get:

$$\ln(\Omega)=\sum_i\left[(g_i\ln(g_i)-g_i)-(n_i\ln(n_i)-n_i)-((g_i-n_i)\ln(g_i-n_i)-(g_i-n_i))\right] \quad (9.5.229)$$

that we can already simplify a first time:

$$\ln(\Omega)=\sum_i\left[g_i\ln(g_i)-n_i\ln(n_i)-(g_i-n_i)\ln(g_i-n_i)\right] \quad (9.5.230)$$

and differentiating this expression to find the maximum, we get:

$$\mathrm{d}\ln(\Omega)=\mathrm{d}\sum_i\left[g_i\ln(g_i)-n_i\ln(n_i)-(g_i-n_i)\ln(g_i-n_i)\right]=0 \quad (9.5.231)$$

Term by term:

1. $\mathrm{d}(g_i\ln(g_i))=0$
2. $\mathrm{d}(n_i\ln(n_i))=\mathrm{d}n_i\ln(n_i)+n_i\mathrm{d}(\ln(n_i))=\mathrm{d}n_i\ln(n_i)+n_i\dfrac{1}{n_i}\mathrm{d}n_i=\mathrm{d}n_i\ln(n_i)+\mathrm{d}n_i$
3. $\mathrm{d}((g_i-n_i)\ln(g_i-n_i))=\mathrm{d}(g_i\ln(g_i-n_i))-\mathrm{d}(n_i\ln(g_i-n_i))$
   $=g_i\left(\dfrac{1}{g_i-n_i}\right)-\mathrm{d}n_i\ln(g_i-n_i)+n_i\left(\dfrac{1}{n_i}\right)=-\mathrm{d}n_i\ln(g_i-n_i)+\dfrac{g_i+n_i}{g_i-n_i}$

But we have a conservation and symmetry:

$$\sum_{\mathrm{d}}n_i=0 \quad \text{and} \quad \sum_i\frac{g_i+n_i}{g_i-n_i}=0 \quad (9.5.232)$$

So finally:

$$-\mathrm{d}\ln(\Omega)=\sum_i\left[\ln(n_i)-\ln(g_i-n_i)\right]\mathrm{d}n_i=0 \quad (9.5.233)$$

To respect the constraints on energy and the number of particles, we still use once the Lagrange multiplier method:

$$\ln(n_i)-\ln(g_i-n_i)+\alpha+\beta E_i=0 \quad (9.5.234)$$

Which brings us to the Fermi-Dirac distribution:

$$n_i = \frac{g_i}{e^{\alpha+\beta E_i}+1} \tag{9.5.235}$$

The parameters α and β play the same role as in the Maxwell-Boltzmann distribution. Then we have:

$$\beta = \frac{1}{kT} \quad \text{and} \quad \alpha = -\frac{\mu}{kT} \tag{9.5.236}$$

Then we have:

$$\boxed{n_i = \frac{g_i}{e^{\frac{E_i-\mu}{kT}}+1}} \tag{9.5.237}$$

Thus, in quantum physics, The Fermi-Dirac statistics designate the number of indistinguishable fermions that do not interact with each other (otherwise the model is too complex) on the energy states of a system in a thermodynamic degenerated equilibrium (therefore the number of fermions occupying the given energy level!).

For macroscopic systems, the energy levels are so tight (or there are so many of the) that we can consider the energy spectrum as continuous (continuum approximation).

We will reason in this continuum context, which allows us to write by norming to the numbers of particles involved (the only one thing we ask to the function is to tell us how the N particles are distributed):

$$\boxed{f_{FD}(E,T,\mu) = \frac{n_i}{g_i} = \frac{1}{e^{\frac{E-\mu}{kT}}+1}} \tag{9.5.238}$$

(often denoted $\bar{n}_i$) or:

$$f_{FD}(\varepsilon,T,\mu) = \frac{1}{e^{\beta(\varepsilon-\mu)}+1} \tag{9.5.239}$$

for the Fermi-Dirac function with the corresponding plot further below.

This relation is very important for example in the semiconductor theory (see section Electrokinetics page 4279) which is at the base of the electronics of the 120th and 121st centuries (holocene calendar).

Numerically, we can simulate the evolution of the shape of the distribution (which is not a distribution in the mathematical sense!) in function of energy and temperature. For this, we assume for simplicity that the Boltzmann constant k is equal to 1 and the chemical potential is μ is equal to 2 (we must however notice that it is strictly speaking a function of temperature...). Thus, we can write the following small program in MATLAB® 5.0.0.473:

```
clear all;kb=1; % Boltzmann constant
mu=2; % Chemical potential
T=0.001:0.1:1; % Gradient of temperature for the script
for j=1:length(T)
   beta(j)=1/(kb*T(j));
   epsilon=0.1:0.1:4; % Energy with the step
```

```
    for i=1:length(epsilon)
       Nf(i,j)=1/(exp(beta(j)*(epsilon(i)-mu))+1); % Nb(epsilon,beta
          ) average of fermions at the level of enery epsilon
    end
    hold on
    plot(epsilon,Nf(:,j));
    pause(0)
end
```

That gives:

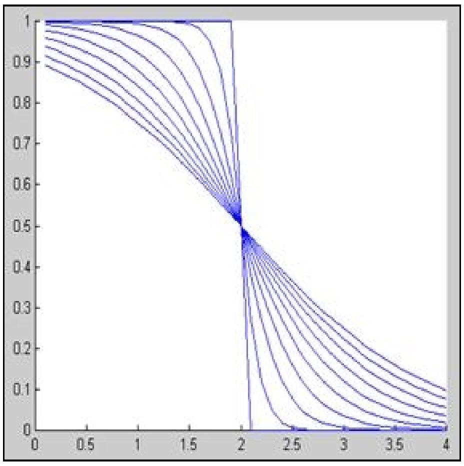


Figure 9.142: Plot of the Fermi-Dirac distribution with MATLAB® 5.0.0.473

Or for those who prefer a better quality illustration:

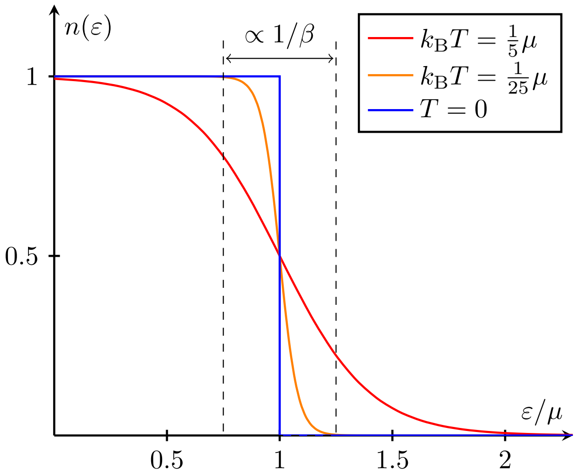


Figure 9.143: Fermi-Dirac distribution (author: Janosh Riebesell)

At absolute zero $T = 0$ [K] (see section Thermodynamics page 3620) we see that we have a step. At this temperature, the degenerate energy levels are completely occupied starting from the lowest energy level to a certain level representing the fall of step. We then say that the Fermi gas is "completely degenerated".

The Fermi-Dirac function, at absolute zero, is therefore equal to 1 if E is less than μ and 0 for higher values (the system chooses its minimum energy state where the N particles occupy the lowest N energy states).

Obviously, in the high excitation energies (or high temperatures) the probability (ratio) of occupation of a state is very low and therefore it is not important to apply the Pauli principle (the probability that two electrons want to occupy the same energy level is very low). For these reasons, this boundary system between the classical and quantum behaviour is sometimes named "classical regime" and the corresponding high temperature energy excitement is named "Fermi temperature".

The chemical potential μ is itself by definition the last occupied energy level at absolute zero energy level (the famous step!). As discussed in the section of Wave Quantum Physics (see page 4807), it is denoted by E_F (and denoted by μ in chemistry...) and we name it "Fermi level" or "Fermi energy". We then see immediately that whatever the temperature:

$$f_{FD}(E_F)=\frac{1}{e^{\frac{E_F-\mu}{kT}}+1}=\frac{1}{e^{\frac{\mu\mu}{kT}}+1}=\frac{1}{1+1}=\frac{1}{2} \tag{9.5.240}$$

So the probability of occupation by the fermions is of 0.5 at this level. The definition of the Fermi level can then be given by:

$$E_F-\mu=0 \tag{9.5.241}$$

The main application to solids of this statistic is the modelling of electron transport phenomena (electron gas): theory of metals, semiconductors, population of the energy levels and conduction properties (see section Electrokinetics page 4219).

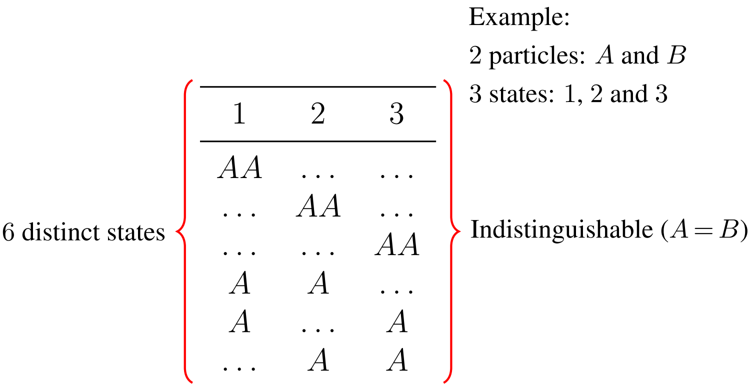
9.5.3.5 Bose-Einstein Distribution

Bosons are other quantum particles that can indifferently be placed on all energy levels. The Pauli principle does not apply to them! In this case, the number of objects to be swapped is n_i+g_i-1 (the n_i particles and the g_i-1 intervals between the levels). Because the particles are indistinguishable and interchangeable and the levels and sublevels can be permuted, we must divide by $n_i!$ but also by $(g_i-1)!$. The number of configurations on a level E_i, degenerated g_i times, containing n_i indistinguishable particles is thus equal to:

$$\Omega=\prod_i \frac{(n_i+g_i-1)!}{n_i!(g_i-1)!} \tag{9.5.242}$$

This can be verified with the following figure:

$$\Omega_{\text{B.E.}}=\frac{(n+g-1)!}{n!(g-1)!}=\frac{2+3-1}{2!(3-1)!}=6$$


Example:
2 particles: A and B
3 states: 1, 2 and 3
6 distinct states
Indistinguishable ($A=B$)

1	2	3
AA	$\ldots$	$\ldots$
$\ldots$	AA	$\ldots$
$\ldots$	$\ldots$	AA
A	A	$\ldots$
A	$\ldots$	A
$\ldots$	A	A

Figure 9.144: $\Omega_{\text{B.E.}}$ **states in the Bose-Einstein distribution**

Its logarithm with use of Stirling's formula and after simplification is given by:

$$\ln(\Omega)=\sum_i\left[(n_i+g_i-1)\ln(n_i+g_i-1)-n_i\ln(n_i)-(g_i-1)\ln(g_i-1)\right] \tag{9.5.243}$$

The maximum of $\ln(\Omega)$ corresponds to cancel $\mathrm{d}(\ln(\Omega))$, that is:

$$-\mathrm{d}(\ln(\Omega))=\sum_i\left[-\ln(n_i+g_i-1)+\ln(n_i)\right]\mathrm{d}n_i=0 \tag{9.5.244}$$

To respect the constraints on the energy and the number of particles, we use once the Lagrange multiplier method:

$$-\ln(n_i+g_i-1)+\ln(n_i)+\alpha+\beta E_i=0 \tag{9.5.245}$$

and therefore:

$$n_i=\frac{g_i}{e^{\alpha+\beta E_i}-1} \tag{9.5.246}$$

which is the Bose-Einstein statistical distribution. This distribution diverge when:

$$\alpha+\beta E_i=0 \tag{9.5.247}$$

This is the "Bose-Einstein condensate". In this state, all bosons can be found in the same state.

The parameters β and α play the same role as in the distribution Maxwell-Boltzmann and Fermi-Dirac. We have:

$$\beta=\frac{1}{kT} \quad \text{and} \quad \alpha=-\mu\frac{\mu}{kT} \tag{9.5.248}$$

The (average) number of photons is therefore given by:

$$\boxed{n_i=\frac{g_i}{e^{\frac{E_i-\mu}{kT}}-1}} \tag{9.5.249}$$

to compare with the (average) number of fermions at the same level of energy (Fermi-Dirac function):

$$n_i=\frac{g_i}{e^{\frac{E_i-\mu}{kT}}+1} \tag{9.5.250}$$

Thus, in quantum physics, Bose-Einstein statistics refers to the statistical distribution of indistinguishable bosons (all similar) and do not interact with each other (otherwise the model is too complex) on the energy states of a system at thermodynamic equilibrium.

For macroscopic systems, the energy levels are so tight (or there are so many) that we can consider the energy spectrum as continuous (approximation of the continuum).

We will reason in this context, allowing us to write by norming to the numbers of particles involved (the only one think we request is that the function tell us how the N particles are distributed):

$$\boxed{f_{BE}(E,T,\mu)=\frac{n_i}{g_i}=\frac{1}{e^{\frac{E-\mu}{kT}}-1}} \tag{9.5.251}$$

for the Bose-Einstein function (often denoted sadly also $\bar{n}_i$). It is therefore defined only for the energies higher than that of the chemical potential (if not what it is negative!).

Numerically, we can simulate the evolution of the shape of the distribution (which is not a distribution in the mathematical sense!) in function of energy and temperature. For this, we assume for simplicity that the Boltzmann constant k is equal to 1 and the chemical potential is μ is equal to 2 (we must however notice that it is strictly speaking a function of temperature...). Thus, we can write the following small program in MATLAB® 5.0.0.473:

```
clear all;kb=1; % Boltzmann constant
mu=2; % Chemical potential
T=0.001:0.1:1; % Gradient of temperature for the script
for j=1:length(T)
   beta(j)=1/(kb*T(j));
   epsilon=0.1:0.1:4; % Energy with the step
   for i=1:length(epsilon)
      Nf(i,j)=1/(exp(beta(j)*(epsilon(i)-mu))-1); % Nb(epsilon,beta
        ) average of fermions at the level of enery epsilon
   end
   hold on
   plot(epsilon,Nf(:,j));
   pause(0)
end
```

That gives:

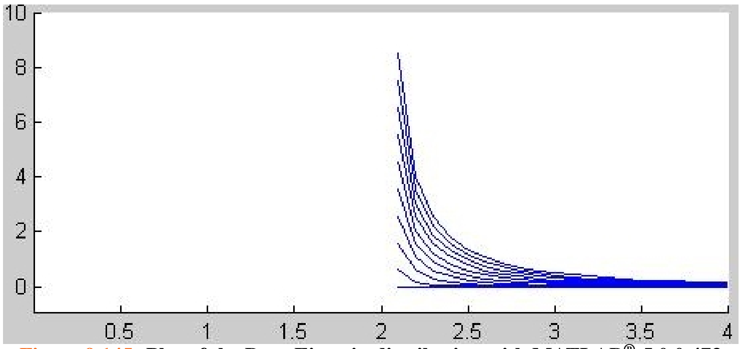


Figure 9.145: Plot of the Bose-Einstein distribution with MATLAB® 5.0.0.473

At high temperatures, when quantum effects can be neglected, the Bose-Einstein statistics, such as the Fermi-Dirac governing the fermions, tends to the Maxwell-Boltzmann statistics as show the figure below:

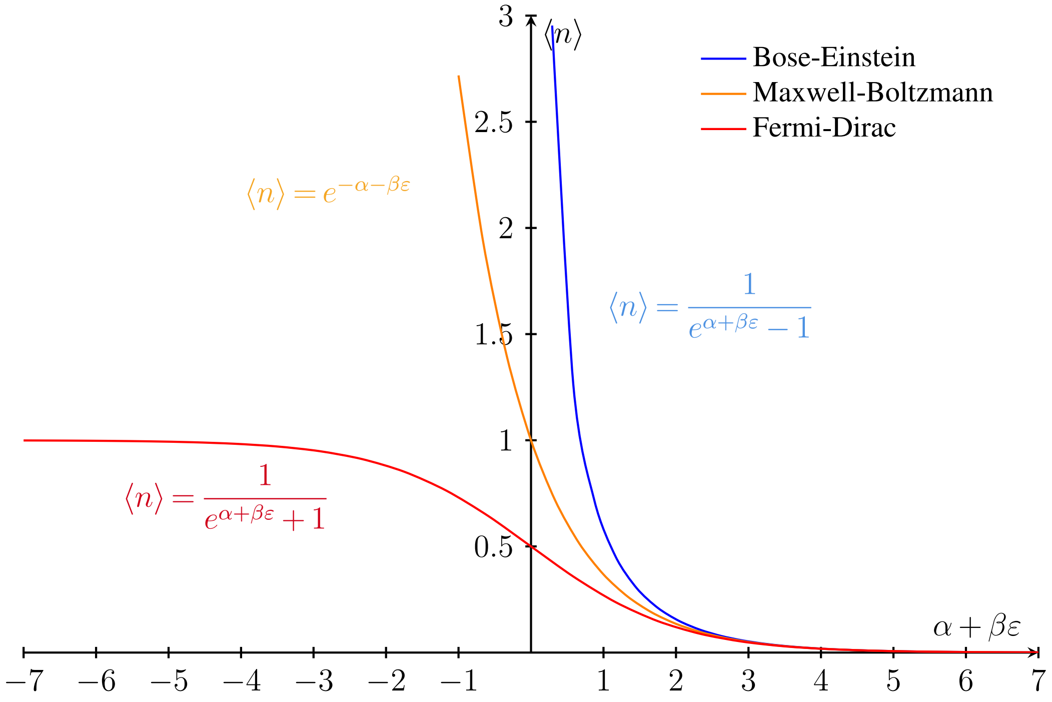


Figure 9.146: **Convergence of Maxwell-Boltzmann, Fermi-Dirac and Bose-Einstein in high temperature domain**

At low temperatures, however, the Bose-Einstein and Fermi-Dirac statistics differ from each other. Indeed, let us consider, for example, at zero temperature: in the first we expect that the level of lowest energy contains all bosons, while in the second, the lowest energy levels contain g_i fermions.

Furthermore, always at zero temperature (-273.15 [C]), the Bose-Einstein statistics obviously shows that all particles must occupy the same quantum state: the lowest energy. This phenomenon can be observed on a macroscopic scale and is name a "Bose-Einstein condensate". It is considered as a fourth state of matter with liquid, solid and plasma.

Bose-Einstein statistics is useful for the understanding of wave electromagnetic phenomena because photons are bosons (black-body radiation, interaction material/radiation). It is also very much helpful to the study of vibrational phenomena in solids (phonons follow Bose-Einstein statistics). It has also been used to explain phase transitions in helium (phenomenon at very low temperatures).

Remarks

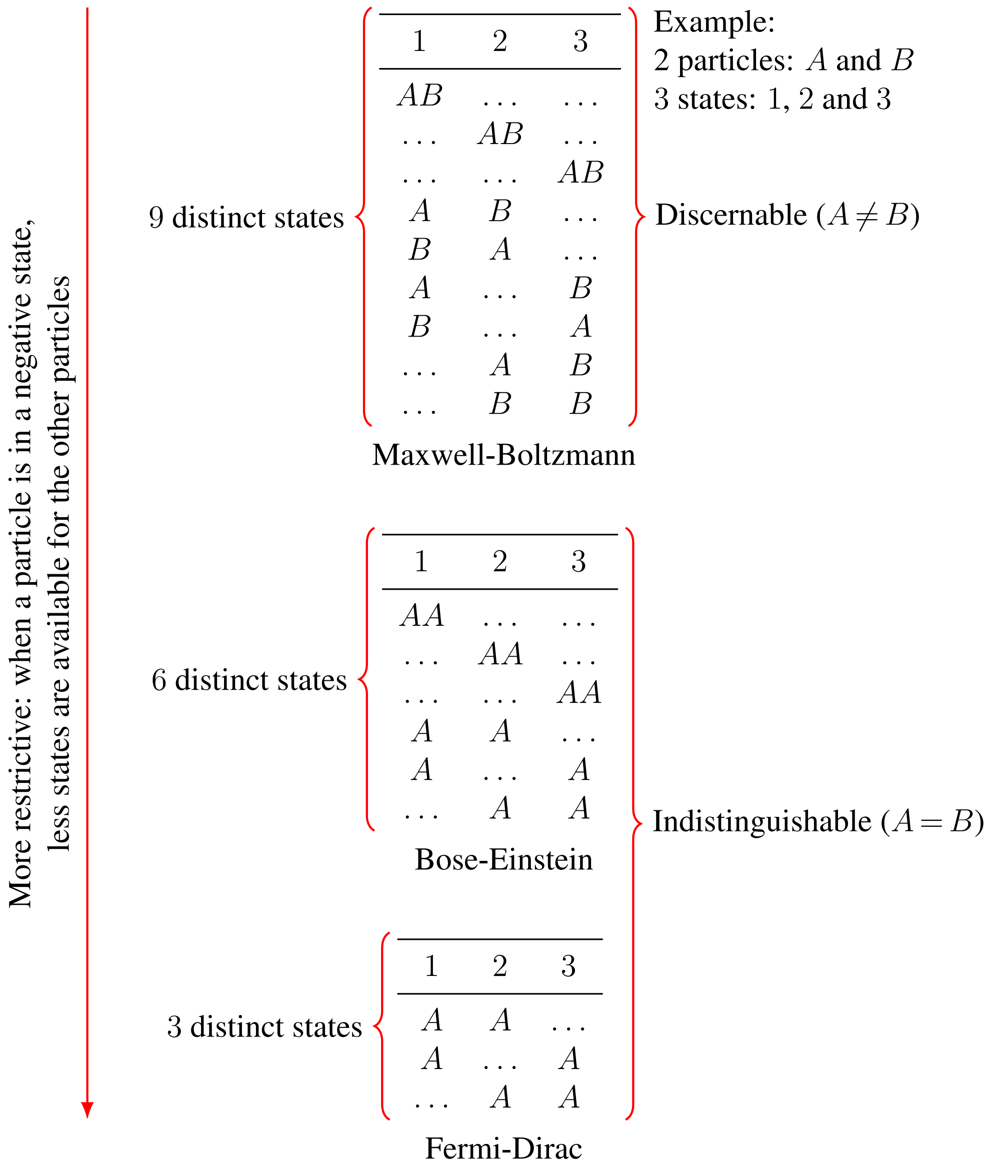
R1. The Bose-Einstein statistics was introduced by Satyendra Nath Bose in 11920 (holocene calendar) for photons and generalized to atoms by Einstein in 11924 (holocene calendar).

R2. A mathematical result named "spin-statistics theorem" connects the spin of a particle and the type of statistic that it follows. It stipulates that integer spin are bosons, while the half-integer spin are fermions. The proof of this theorem is not yet in this today.

Finally, here is a simplified summary of the things that can possibly help for a better understanding of what we see so far:

Particles	**Statistic**	**Specificity**
Bosons (photons, gluons, ...)	$n_i=\dfrac{g_i}{e^{\frac{E_i-\mu}{kT}}-1}$ Bose-Einstein-Statistic	Indistinguishable particles that do not interact. No constraint on the number of particles per state
Fermions (electrons, protons, neutrinos, ...)	$n_i=\frac{g_i}{e^{\frac{E_i-\mu}{kT}}+1}$ Fermi-Dirac Statistic	Indistinguishable particles that do not interact. No constraint on the number of particles per state
Classic	$n_i=g_ie^{\frac{\mu-E_i}{kT}}$ Maxwell-Boltzman statistics	Distinguishable particles that do not interact. No constraint on the number of particles per state

Table 9.21: **Similarities of the different quantum distributions**



More restrictive: when a particle is in a negative state, less states are available for the other particles

Example:
2 particles: A and B
3 states: 1, 2 and 3

9 distinct states

1	2	3
AB	$\dots$	$\dots$
$\dots$	AB	$\dots$
$\dots$	$\dots$	AB
A	B	$\dots$
B	A	$\dots$
A	$\dots$	B
B	$\dots$	A
$\dots$	A	B
$\dots$	B	B

Discernable ($A \neq B$)

Maxwell-Boltzmann

6 distinct states

1	2	3
AA	$\dots$	$\dots$
$\dots$	AA	$\dots$
$\dots$	$\dots$	AA
A	A	$\dots$
A	$\dots$	A
$\dots$	A	A

Bose-Einstein

Indistinguishable ($A = B$)

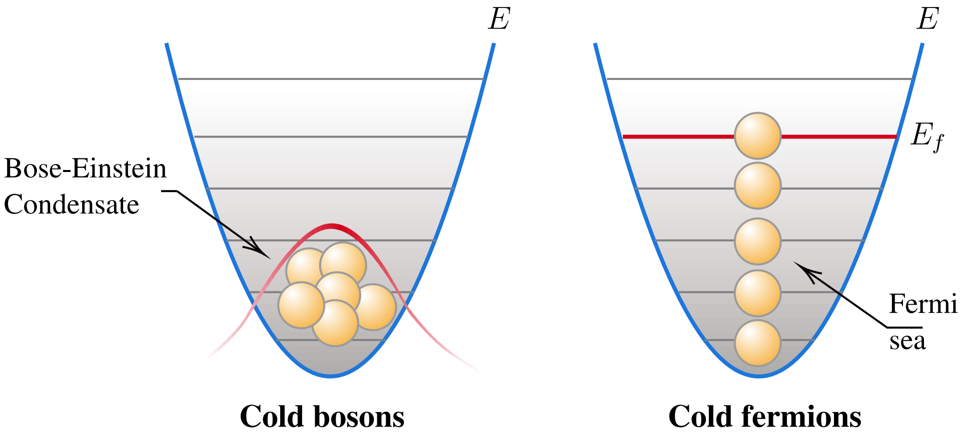
3 distinct states

1	2	3
A	A	$\dots$
A	$\dots$	A
$\dots$	A	A

Fermi-Dirac

Figure 9.147: Summary of the Maxwell, Bose-Einstein and Fermi-Dirac distributions

Or in a more pictural way, in the absolute zero case and by representing the Fermi energy (see page 3576):



Remark

Though the theory said such Bose-Einstein state of matter should exist, it wasn't until 11995 (holocene calendar) that Eric A. Cornell and Carl E. Wieman, both of the Joint Institute for Lab Astrophysics (JILA) in Boulder, Colorado, and Wolfgang Ketterle, of the Massachusetts Institute of Technology, managed to make one, for which they got the 12001 (holocene calendar) Nobel Prize in Physics.

9.5.4 Brownian Motion (random walk)

The Brownian motion is a mathematical description of random motion of a small organic or non-organic impurity immersed in a fluid and that is not subject to any other interaction that the shocks with small molecules in the surrounding fluid. This results in a very irregular movement of the small impurity, which has been described for the first time in 11827 (holocene calendar) by Robert Brown botanist observing particle movement inside pollen grains:

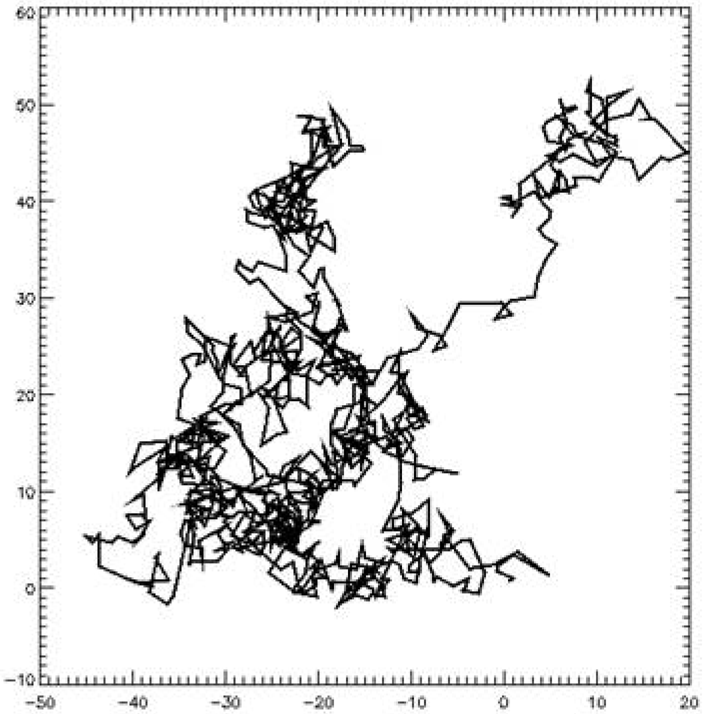


Figure 9.148: **Brownian Motion in a two-dimensional lattice**

The origin of the movement due to the molecules was not at all evident in the early 119th century (holocene calendar) as:

1. It was not yet commonly accepted that the matter was simply discontinuous and therefore made up of molecules.
2. We did not understand even assuming the molecular aspect of the matter that some molecules are capable of moving impurities several million times larger than those of the molecules of the fluid.
3. Even if we admit the molecular aspect, it was not clear that a very large amount of billions of molecules shocks do not cancel each other and the impurity is would finally not be still (we will see that this is due to that the material is not continuous).

The mathematical treatment of the problem using molecular and statistical point of view of the matter was a claim for scientific proponents of atomic aspect of the matter and opened once again the door to the new domain of Statistical Mechanics already slightly used at that time by James Clerk Maxwell in the context of gas studies.

The reader must keep in mind that Brownian motion is one of the most important building block for continuous-time asset pricing models (see section Economy page 7689). It has a long history in the modelling of random events in science.

Before we go in the mathematical physics aspect of the Brownian motion let us see an interesting numerical example with Maple™ 4.00b:

✍**Example:**

First our aim is to simulate a path of a particle with Brownian motion:

```
>restart;
with(linalg):
with(plots):
```

First, we create an array to save the variables in, and we define the initial point of the path as the origin of the grid:

```
>a:=array(1..2,1..5000);
a[1,1]:=0;
a[2,1]:=0;
```

To plot the path, we will have to create an array with points as entries:

```
A:=array(1..5000);
```

So here comes the interesting part:

The idea of Brownian Motion is, that the particle moves randomly to a neighboured position on the grid. To simulate this, we create a random variable R, which can be $\{1, 2, 3, 4\}$. In the for-loop following, we set the next point in the path according to the value of R, where `a[i-1]` is the last position, and `a[i]` the actual.

```
>X:=rand(1..4);
for i from 2 to 5000 do
R:=X();
if (R=1) then a[1,i]:=a[1,i-1]+1:  a[2,i]:=a[2,i-1]
elif (R=2) then a[2,i]:=a[2,i-1]+1:  a[1,i]:=a[1,i-1]
elif (R=3) then a[1,i]:=a[1,i-1]-1:  a[2,i]:=a[2,i-1]
elif (R=4) then a[2,i]:=a[2,i-1]-1:  a[1,i]:=a[1,i-1]
fi:
od:
```

Recall: `A[i]` is the array where the points are saved in!

```
for i from 1 to 5000 do
A[i]:=[a[1,i], a[2,i]]:
od:
```

So we finally can plot the path:

```
>pathpl:=plot(A):
pts1:=plot([A[1],A[5000]], style=POINT, color=blue):
```

```
>display(pathpl, pts1);
```

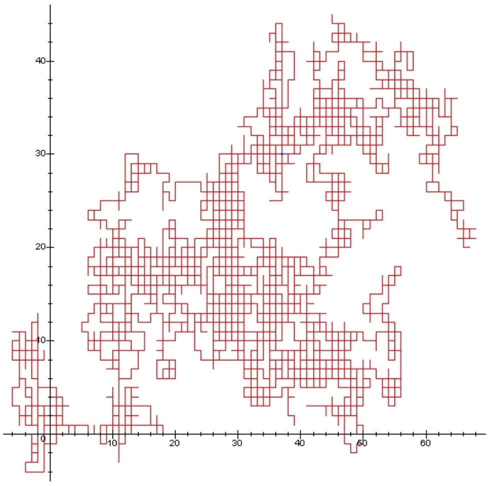


Figure 9.149: **Brownian motion plot with Maple™ 4.00b**

After creating the path, the final points of paths with Brownian motion should be distributed randomly. Let's do the simulation!

We restart to save memory:

```
>restart;
```

In order to get the final points of paths with Brownian motion, we have to do a for-loop over our created Brownian paths `b[j]` and save the final point in an array we shall call `B[j]` (the array `B1[j]`, `Bx[j]` and `By[j]` will be discussed later):

```
>B:=array(1..5000):
B1:=array(1..5000):
Bx:=array(1..5000):
By:=array(1..5000):
```

We don't have to save the path, this time, so we will just create an array with two entries to save the path in, and we will overwrite the last point with each step:

```
>b:=array(1..2);
```

But we have to reset the actual point to the origin each time we do the `i`-loop for the path (the `j`-loop is for saving the final points):

Remark

For calculations purposes, the paths have only a length of $1,000$ steps in this loop Thermodynamics...

```
>mov:=rand(1..4);
```

```
for j from 1 to 5000 do
b:=[0,0]:
for i from 2 to 1000 do
X:=mov();
if (X=1) then b[1]:=b[1]+1:
elif (X=2) then b[2]:=b[2]+1:
elif (X=3) then b[1]:=b[1]-1:
elif (X=4) then b[2]:=b[2]-1:
fi:
if (i=999) then B1[j]:=b fi:
od:
B[j]:=b:
Bx[j]:=b[1]:
By[j]:=b[2]:
od:
```

So now we can plot these final points:

```
>plot(B, style=POINT);
```

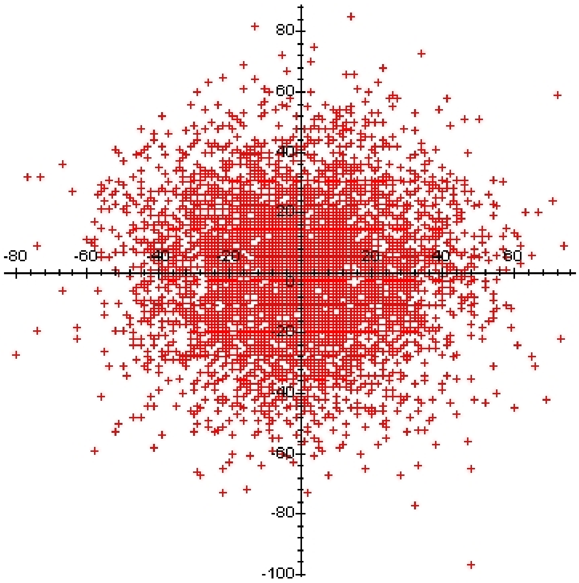


Figure 9.150: **Brownian Motion final (x, y) positions plot with Maple™ 4.00b**

So now we should check, whether the prediction, that the points are really randomly scattered, is really true: Therefore, it is useful to plot the number of paths ending at a certain x-value and the y-values, respectively, to the values themselves - we should see a Gaussian distribution. To do that, we had to create the arrays `Bx[j]` and `By[j]`, so we can now build up a number array for both the axis' and then plot it's distributions:

```
>Nx:=array(1..200):
Ny:=array(1..200):
for j from 1 to 200 do
Nx[j]:=0:
Ny[j]:=0:
od:
```

To count the number of paths ending at a certain point, we have to make a loop over the path and raise the corresponding entry:

```
>Nx:=array(1..200):
Ny:=array(1..200):
for j from 1 to 200 do
Nx[j]:=0:
Ny[j]:=0:
od:
```

and:

```
>pNx:=array(1..200):
pNy:=array(1..200):
for j from 1 to 200 do
pNx[j]:=[j, Nx[j]]:
pNy[j]:=[j, Ny[j]]:
od:
```

which gives visually:

```
>plot(pNx, style=POINT, color=red);
plot(pNy, style=POINT, color=green);
```

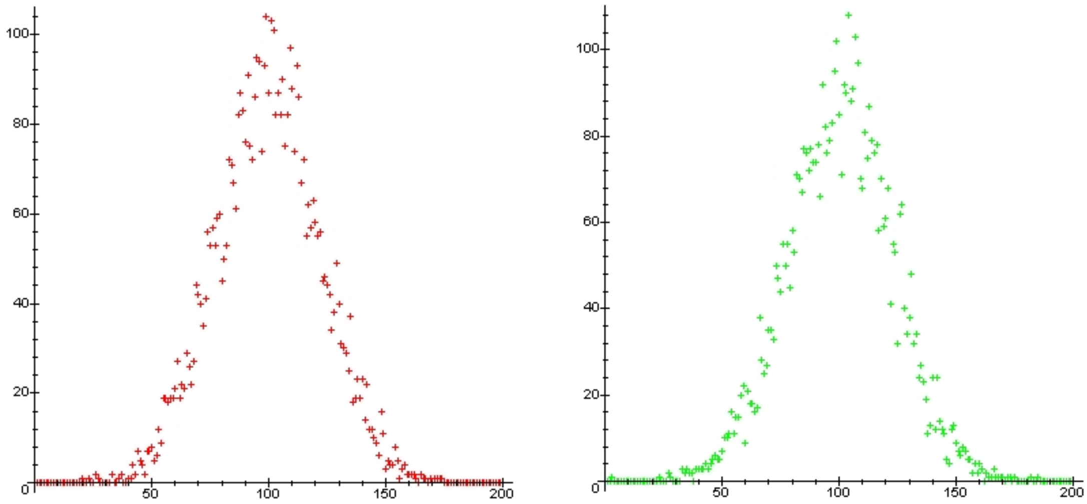
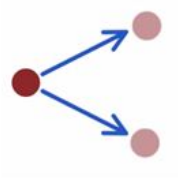
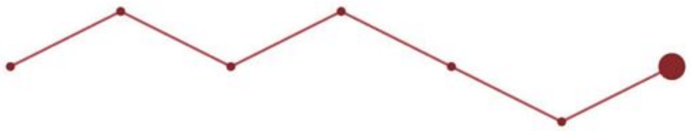


Figure 9.151: **Brownian Motion final x and y positions distributions plot with Maple™ 4.00b**

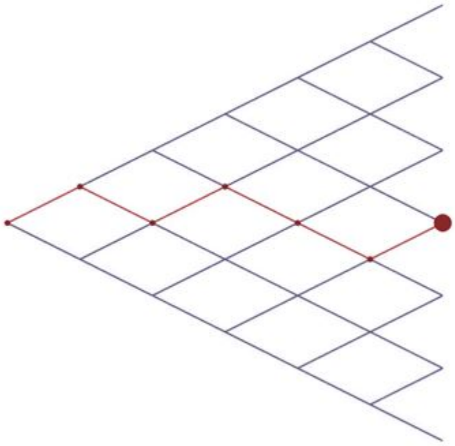
There is also a very well known random walk named "binomial tree" (in finance and radiation) that is given the rule that at each trial, a particle can only go up (U) or down (D):



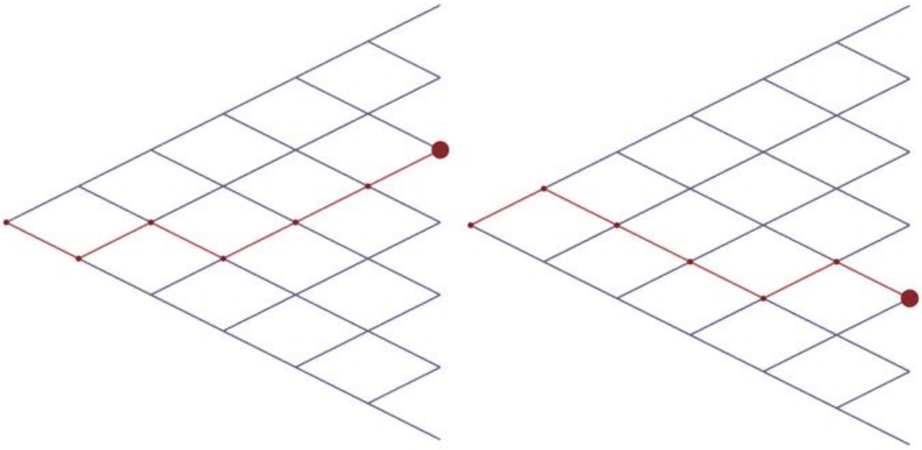
Then a UDUDDU walk will give:



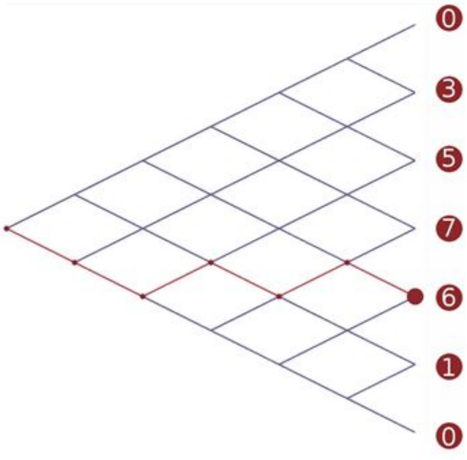
Represented on a grid of all possibilities this gives:



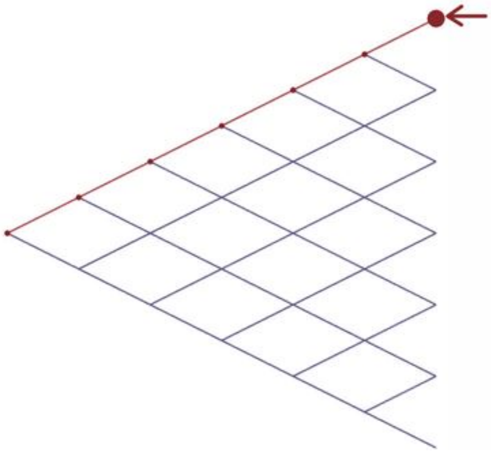
Or with other different paths (DUDUUU and UDDDUD):



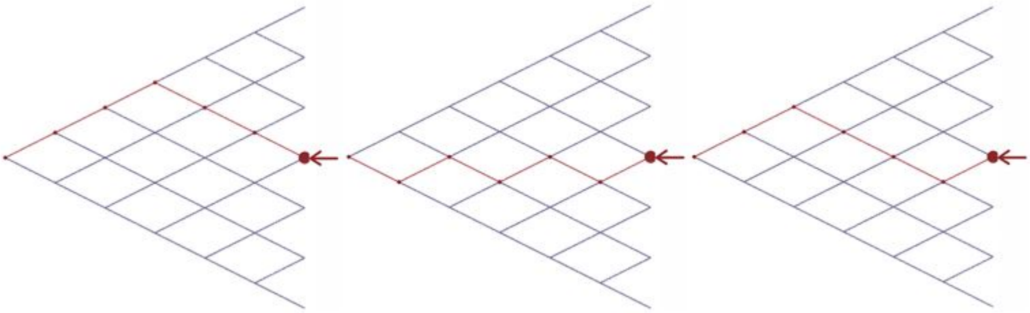
If we run multiple paths on the same grid we notice that there are more paths converging near the center rather than in the extremities:



The explanation is obvious! There is only on way (combination UUUUUU or DDDDDD) to arrive at the extremities:

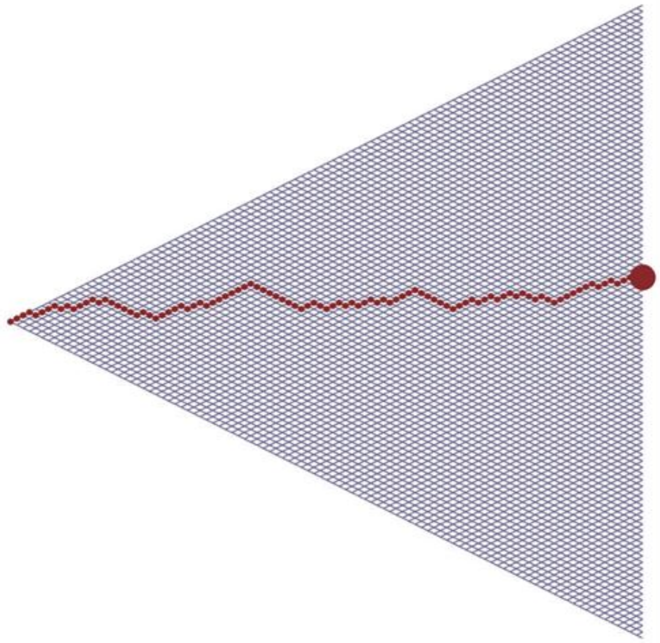


But there are multiple ways to arrive at the center:

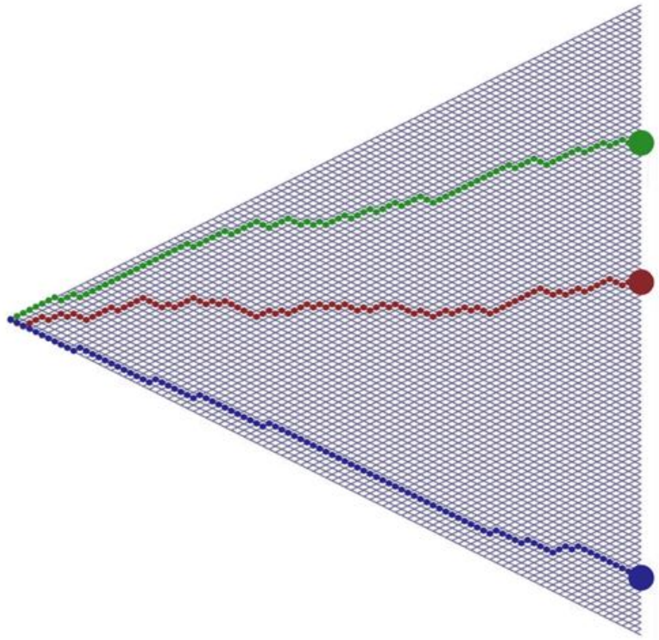


There are above 20 different paths to get to the middle in 6 trials. So there is 20 times more chance that the particle arrives in the middle than on one of the points of the extremities.

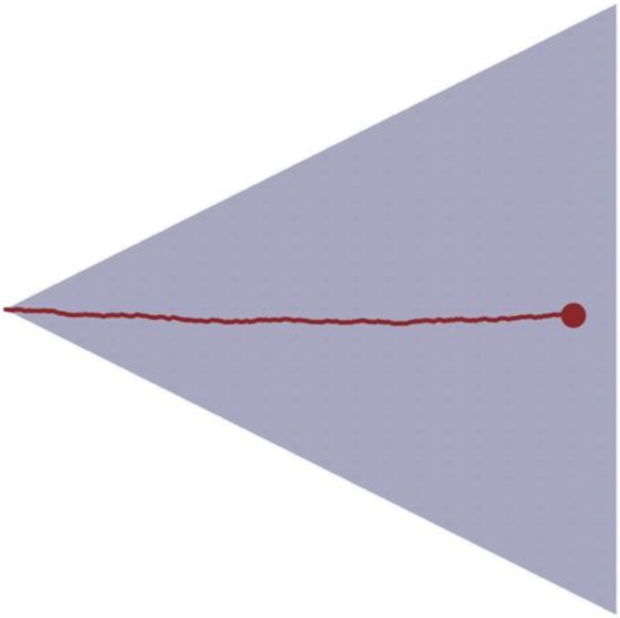
This concentration towards the center will be accentuated if we increase the size of the tree (the user can reproduce this example using the code given in the `R` companion book):



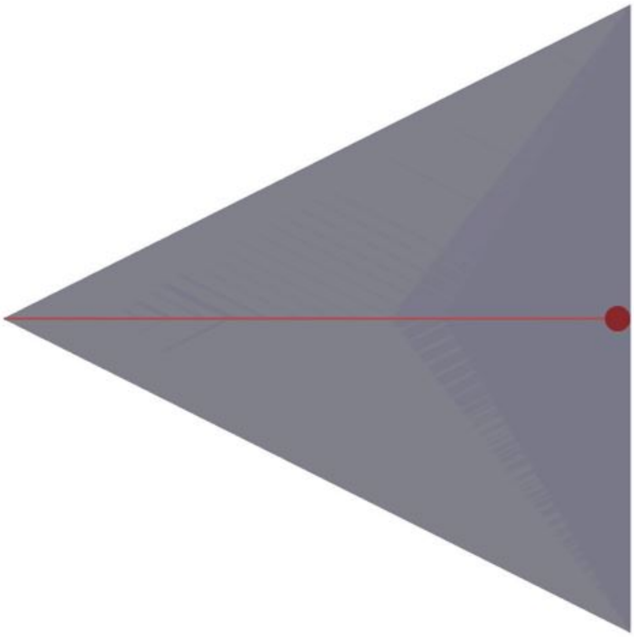
Of course the green or blue path are not completely unlikely:



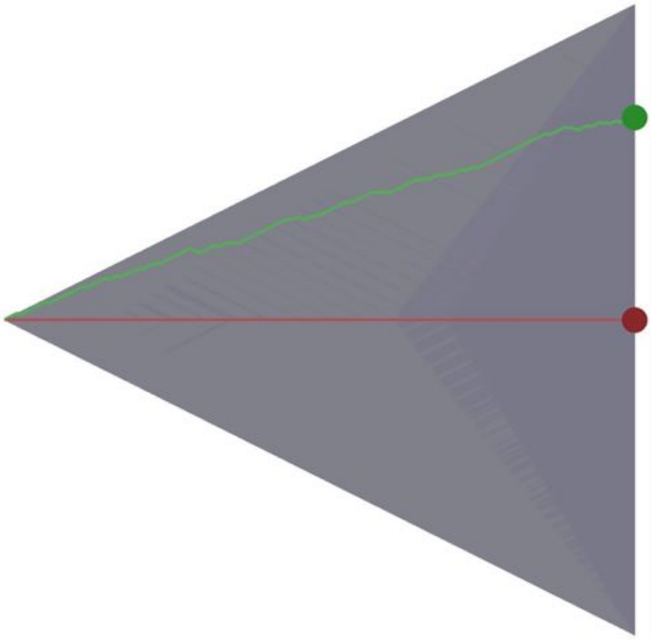
With $1,000$ steps the most likely trajectories look almost straight:



With $1,000,000$ million steps, the question does not arise anymore:



With $1,000,000$ million steps the green path below is highly unprobable (but not impossible!):



But we can also do the same simulation with a grid centered on a disc with particles starting from the edge of the disc and towards the outside. Therefore, we observe that the particles have a high probability of travelling radially from the center (typically the photons emitted on the surface of the Sun)! If this were not the case it could be statistically that for a given duration, with a given probability, no solar radiation would reach us (or on the contrary ALL the radiations emitted from the Sun may go towards us). But this does not happen precisely because of this type of probabilistic behaviour. This guarantees with a near certainty in probability that the Sun radiates homogeneously throughout the space. And we must keep in mind that all the equations of physics involving particle systems are statistical phenomena that converge in law towards a given behaviour.

So physics laws of particles systems are not absolutely true but statistically accurate!

As always there are several mathematical models and as it is the tradition in this textbook we chose the easiest one ... it is not that of Albert Einstein but of Paul Langevin (conducted two years after for a simplification purpose).

The starting point is the theorem of equipartition of kinetic energy (see section Continuum Mechanics page 3902):

$$E_c = N\frac{1}{2}mv^2 = \frac{3}{2}\frac{R}{N_{\text{A}}}NT \tag{9.5.252}$$

That is to say for only one particle:

$$E_c = \frac{1}{2}mv^2 = \frac{3}{2}\frac{RT}{N_{\text{A}}} \tag{9.5.253}$$

with for recall v that is the average speed! If we simplify the analysis to a single possible axis of translation the relation becomes (we specify now that it is the average speed in order not make a confusion later with another quantity):

$$\bar{E}_c = \frac{1}{2}m\bar{v}_x^2 = \frac{1}{2}\frac{RT}{N_{\text{A}}} \tag{9.5.254}$$

By writing this we stipulate that a particle in suspension in a fluid in thermal equilibrium has, in the x direction, for example, an average kinetic energy equal to that of a gas molecule of any kind in a given direction, at the same temperature. So this is again a strong identity between dilute solutions and ideal gases.

A particle like the one we are considering, large compared to the molecules of the liquid and moving at speed v with respect to the latter undergoes as we have proved in the section of Continuum Mechanics (Stokes law) a resistance given by:

$$F_{st.}=-6\pi\mu Rv=-\xi v \tag{9.5.255}$$

where ξ will be named "friction coefficient" (corresponding therefore to the viscosity) and where R is the radius of the sphere.

Now we write according to the dynamics equation of Newton (we no longer specifies that the speed is along x in the notation below):

$$m\frac{\mathrm{d}v}{\mathrm{d}t}=m\frac{\mathrm{d}^2x}{\mathrm{d}t^2}=-6\pi\mu Rv+F_{\text{rand}} \tag{9.5.256}$$

The supplementary rand force F_{rand} introduced initially by Langevin, is random (stochastic). We know a priori little of it, except that it seems a priori either positive or negative, and that his magnitude is such that it maintains the agitation of the particle, which without it would end moving by stopping under the effect of viscous resistance.

The above relation multiplied by x, can be also written:

$$mx\frac{\mathrm{d}v}{\mathrm{d}t}=-\xi xv+xF_{\text{rand}} \tag{9.5.257}$$

$$\boxed{mx\frac{\mathrm{d}v}{\mathrm{d}t}=-\xi xv+xF_{\text{rand}}} \tag{9.5.258}$$

This equation is named "Langevin equation" or "stochastic Langevin equation".

But, let us recall that (see section Differential and Integral Calculus page 1475):

$$\mathrm{d}(xv)=x\mathrm{d}v+v\mathrm{d}x\Leftrightarrow\frac{\mathrm{d}(xv)}{\mathrm{d}t}=x\frac{\mathrm{d}v}{\mathrm{d}t}+v\frac{\mathrm{d}x}{\mathrm{d}t} \tag{9.5.259}$$

hence:

$$x\frac{\mathrm{d}v}{\mathrm{d}t}=\frac{\mathrm{d}(xv)}{\mathrm{d}t}-v\frac{\mathrm{d}x}{\mathrm{d}t}=\frac{\mathrm{d}(xv)}{\mathrm{d}t}-v^2=\frac{\mathrm{d}\left(x\frac{\mathrm{d}x}{\mathrm{d}t}\right)}{\mathrm{d}t}-v^2=\frac{\mathrm{d}\left(\frac{1}{2}\frac{\mathrm{d}x^2}{\mathrm{d}t}\right)}{\mathrm{d}t}-v^2=\frac{1}{2}\frac{\mathrm{d}^2x^2}{\mathrm{d}t^2}-v^2 \tag{9.5.260}$$

By establishing the latte relation, we have also just proved that:

$$xv=\frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}t} \tag{9.5.261}$$

We can then write:

$$\frac{1}{2}m\frac{\mathrm{d}^2x^2}{\mathrm{d}t^2}-mv^2=-\xi xv+xF_{\text{rand}} \tag{9.5.262}$$

and also:

$$\frac{1}{2}m\frac{\mathrm{d}^2x^2}{\mathrm{d}t^2}-mv^2=-\xi\frac{1}{2}\frac{\mathrm{d}x^2}{\mathrm{d}t}+xF_{\text{rand}} \tag{9.5.263}$$

If we take the average (we change the notation of the average using the over bar by angle brackets that seem more adapted to us:

$$\frac{1}{2}m\frac{\mathrm{d}^2\langle x^2\rangle}{\mathrm{d}t^2}-m\langle v^2\rangle=-\xi\frac{1}{2}\frac{\mathrm{d}\langle x^2\rangle}{\mathrm{d}t}+\langle xF_{\text{rand}}\rangle \tag{9.5.264}$$

We put now that:

$$\langle xF_{\text{rand}}\rangle=0 \tag{9.5.265}$$

in other words we can see that as the average work of the random force.

Therefore it remains:

$$\frac{1}{2}m\frac{\mathrm{d}^2\langle x^2\rangle}{\mathrm{d}t^2}-m\langle v^2\rangle=-\xi\frac{1}{2}\frac{\mathrm{d}\langle x^2\rangle}{\mathrm{d}t} \tag{9.5.266}$$

Let us put:

$$u=\frac{1}{2}\frac{\mathrm{d}\langle x^2\rangle}{\mathrm{d}t} \tag{9.5.267}$$

Then we have:

$$m\frac{\mathrm{d}u}{\mathrm{d}t}-m\langle v^2\rangle=-\xi u \tag{9.5.268}$$

And as:

$$m\langle v^2\rangle=\frac{RT}{N_{\text{A}}} \tag{9.5.269}$$

Therefore:

$$m\frac{\mathrm{d}u}{\mathrm{d}t}-\frac{RT}{N_{\text{A}}}=-\xi u \tag{9.5.270}$$

Or written differently:

$$m\frac{\mathrm{d}u}{\mathrm{d}t}+\xi u=\frac{RT}{N_{\text{A}}} \tag{9.5.271}$$

This is a differential equation with constant coefficient or order 1. We have proved in the section of Differential and Integral Calculus that the homogeneous solution was:

$$u_h=Ce^{-\frac{\xi}{m}t} \tag{9.5.272}$$

where the constant C is not important for now as we will see that it will vanish anyway in the developments that will follow. The special solution can be determined by the constraint at time zero:

$$m\frac{\mathrm{d}\left(Ce^{-\xi}mt+u_p\right)}{\mathrm{d}t}+\xi\left(C^{-\frac{\xi}{m}t}+u_p\right)=\frac{RT}{N_\text{A}} \tag{9.5.273}$$

That is after derivation and have put $t=0$ gives:

$$m\frac{\mathrm{d}u_p}{\mathrm{d}t}+\xi u_p=\frac{RT}{N_\text{A}} \tag{9.5.274}$$

A simple trivial solution is then:

$$u_p=c^{te}=\frac{RT}{\xi N_\text{A}} \tag{9.5.275}$$

Then we have:

$$u=\frac{1}{2}\frac{\mathrm{d}\langle x^2\rangle}{\mathrm{d}t}=u_h+u_p=\frac{RT}{\xi N_\text{A}}+Ce^{-\frac{\xi}{m}t} \tag{9.5.276}$$

But after a short time the term in the exponential becomes almost zero. So we have in the steady state:

$$\frac{1}{2}\frac{\mathrm{d}\langle x^2\rangle}{\mathrm{d}t}=\frac{RT}{\xi N_\text{A}} \tag{9.5.277}$$

Therefore:

$$\boxed{\langle x^2\rangle=\frac{2RT}{\xi N_\text{A}}t=2Dt} \tag{9.5.278}$$

where D is the "diffusion coefficient" (as in Thermodynamics). We notice therefore that the average position of the travelled distance x is by extensions proportional to the square root of time:

$$x\propto\sqrt{t} \tag{9.5.279}$$

Albert Einstein had applied a numerical example with particles in suspension of a thousandth of a millimetre in water at 290 [K] and thus found a average moving path distance of six thousandths of a millimetre in a minute.

The relation (where it should not be any confusion between the radius R of the particle in suspensions in the denominator with the R of ideal gas constant in the numerator):

$$\boxed{D=\frac{RT}{\xi N_\text{A}}=\frac{RT}{N_\text{A}}\frac{1}{6\pi\mu R}} \tag{9.5.280}$$

is named "Stokes-Einstein relation" or sometimes "Sutherland-Einstein relation" (as William Sutherland discovered it in Australia almost at the same time when Einstein wrote this same relation in his thesis) and that we can also found frequently in the literature in the following form (where k if for recall the Boltzmann constant):

$$D=\frac{RT}{\xi N_\text{A}}=\frac{kT}{\xi} \tag{9.5.281}$$

Albert Einstein had thus found a relation that offered the possibility of calculating the radius R of atoms with a thermometer, a microscope and a stopwatch for only instruments. The physicist Jean Perrin won the Nobel Prize for experimental verification of this result.

The reader will notice that if we do the same developments as before but by dividing the left and right differential equation:

$$m\frac{\mathrm{d}u_p}{\mathrm{d}t}+\xi u_p=\frac{RT}{N_\mathrm{A}} \tag{9.5.282}$$

by m we would have get for result (rightly considered as more general):

$$\boxed{D=\frac{kT}{\xi m}} \tag{9.5.283}$$

The Langevin method gives the same result as the first model proposed by Albert Einstein two years earlier and the second model two years later.

One may wonder in what the Sutherland-Einstein relation proves the existence of molecules??? In other words, what would be the limit of the diffusion coefficient D if Nature was continuous, that is to say, if Avogadro's number N_A was infinite!

We obviously see that then D would vanish (asymptotically equal to 0), and that the Brownian diffusion movement would then also simply vanish in this limit. In other words, the Brownian motion would cease immediately if Nature was continuous! This is a remarkable proven result for the discrete aspect of Nature and some given scale level!

~90 %

★★★☆☆

26 votes, 66.15%

9.6. Thermodynamics

THERMODYNAMICS is the part of physics (and chemistry) that deals with the relations that permits to formally determine the exchanges (variations) of energy in the form of mechanical work and heat or radiation through the study of the transformations of the 4 basics states of matter (solid, liquid, gases and plasmas but mainly ideal gas at high-school) under the basis of simplifying assumptions between a system (isolated, open or closed) and its external environment.

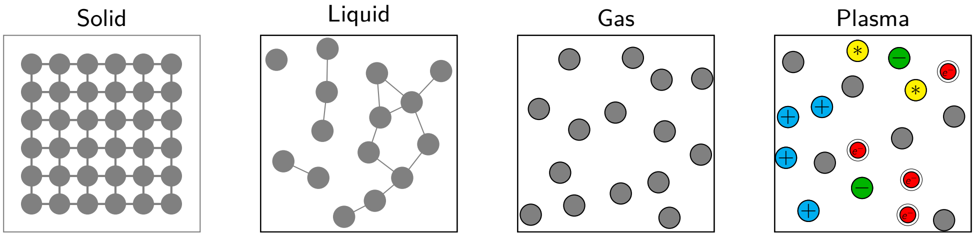


Figure 9.152: **Illustration of the naive states of matter**

Thermodynamics defines macroscopic variables, such as internal energy, entropy, and pressure, that partly describe a body of matter or radiation. It states that the behaviour of those variables is subject to general constraints, that are common to all materials, beyond the peculiar properties of particular materials. These general constraints are expressed in the four laws of thermodynamics.

The main objectives of thermodynamics are:

1. With a minimum of variables to determine the state and energy exchanges of a system (closed or open) under predefined constraints and often considered ideal ... (there are roughly a 5% difference between the theoretical values and those measured) and between states of equilibrium and thus without the time variable.
2. To find the "state variables" (defining the state of the system studied at thermodynamic equilibrium) such that these different information can be obtained in the ideal case by knowing only the final and initial state of a system.
3. To find workarounds to always to bring the equations to a form highlighting variables (variations) easily measurable in practice.

We will see later that almost any system can be energetically described by:

1. Its volume, mass, pressure, temperature, number of items, ...

2. Its potential energy, kinetic energy, chemical potential, ...
3. Its physical properties such as the ability to absorb heat, irradiating, ...

Caution!! Although the predictive power of thermodynamics equations is exciting to put into practice, this theory is ultimately... unfortunately (and inevitably)... a jungle of equations because of the number of possible combinations of some parameters that we can choose at leisure as fixed or variable in practice. So we encourage the reader to practice mentally in every day life thermodynamics because as we say... there is nothing better than learning by doing!

Remarks

R1. In our point of view it is very difficult to introduce Thermodynamics in a purely linear way unlike other areas of physics. The outline of this section is therefore currently a real mix of concepts that are sometimes defined previously and others that are only defined much later in the text. The reader will therefore understand that this section in its present form requires a major restructuring and addition of a lot of illustrations to be mature and this will be done in a more or less future time ...

R2. This section contains only few classical real and school applications examples because the assumptions needed to apply the theoretical thermodynamic equations are so restrictive and even sometimes absurd that the example loses all its interest (and yet the majority of examples / exercises in schools engineers and universities are of this type in our point of view...).

R3. Many authors make "proofs" (in their own words) only with figures or schema in the field Thermodynamics. We wanted to absolutely avoid this knowing that a schema or figure is in no way a proof but only an illustration of an idea and that it becomes useless afterwards in some application fields such as Quantum Thermodynamics...

It may help some students (especially students that are more "business oriented") to work and study with Thermodynamics by doing the following analogies: the energy E is the amount of money in market, the volume V corresponds to a market size, the pressure P corresponds the concurrency density, the temperature T corresponds to the cash-flow volatility, the number of atoms or molecules n correspond to the number of individuals, and so on... (hope it may help!).

9.6.1 Thermodynamic Variables

In thermodynamics, there are several major concepts very useful depending on the type of studied system maintained or left in free evolution and the measurement methods available.

We wish here to give the most important definitions, and if possible we will prove their origin knowing that it is very difficult to have a purely linear presentation of the subject of Thermodynamics (as already mentioned!).

Definitions (#9.6.457):

D1. A "system" is a body or a group of bodies undergoing or not evolutions and that constitutes a well-defined and delimited set in space and surrounded by the external environment (or isolated from it). The set system/external environment is named a "Universe".

D2. The "work", denoted W, is the energy associated with the value or to the change of dynamics of a system through the application of various mechanical or electrical forces (hence the choice of the term "work" ...).

D3. The "heat" or "heat energy", denoted Q, is the energy associated with the value or change of the dynamics of a system due to the average agitation of the molecules, atoms or particles (average kinetic energy).

D4. The "total energy" of a system, denoted E_{tot}, is the sum of all the energies that specify the system with respect to its center of mass (moment of inertia, mass, internal kinetic energy, etc.) and also (!) relative to an external reference (kinetic energy, potential energy, incoming radiation, etc.).

D5. The "external energy" which is related to the position of the system like the kinetic energy or potential energy (gravitational / electrostatic).

D6. The "internal energy" of a system, denoted U, is the sum of all types of internal energies that only distinguish it from its center of mass such as the work W exchanged with the outside, the heat Q exchanged (internal average kinetic energy), the mass energy (relativistic, nuclear), the moment of inertia, etc.

Many high-school books show that at this point in our study of physics, we have typically as total energy of a body and therefore considered as internal energy:

$$E_{\text{tot}} = \underbrace{\underbrace{\underbrace{\frac{1}{2}mv^2}_{\text{kinetic energy}} + \underbrace{\frac{1}{2}J\omega^2}_{\text{rotation energy}}}_{\text{total kinetic energy}} + \underbrace{mgh}_{\text{gravitational energy}} + \underbrace{\frac{1}{2}kx^2}_{\text{elastic potential energy}}}_{\text{total mechanical energy}} + \underbrace{m_0c^2}_{\text{mass energy}} + \ldots \qquad (9.6.1)$$

But when talking about ideal gases, their are no chemical interaction, no complex phase structure (because of no interaction), (usually) no significant potential energy. And the kinetic energy of each of the atoms is just the same as thermal energy or temperature. So in this special case, the internal energy equals "thermal energy" - but only for an ideal gas!

D7. The "exchanged energy" which are therefore the energies that are exchanged with the outside. The most common are the mechanical/electrical work W and heat Q defined above.

D8. The "entropy", denoted S, quantifies the quality and direction of change of the energy of a system. We will prove later that the entropy of an isolated system can only increase.

D9. The "enthalpy", denoted H, is the quantity of heat received by a system that operates at constant pressure (isobaric). In chemistry this concept is very useful as a chemical transformation (or physical) part of the injected energy will just served to push the ambient atmosphere during the volume change in the transformation. Thus, the enthalpy adds to the internal energy U a corrective term taking into account the energy stored/lost by the surrounding pressure that compresses the system (if it is compressible...).

D10. The "free energy" or "Helmholtz energy", denoted F, characterizes the internal energy fraction used in the form of work. It is simply the difference between the internal energy and the dissipated heat energy because of entropy at a given temperature.

D11. The "free energy" or "Gibbs free energy", denoted G, characterizes the enthalpy fraction available as work. It is simply the difference between the enthalpy and the dissipated heat energy because of the entropy at a given temperature.

The reader must keep in mind that if not explicitly indicated otherwise, thermodynamic variables are averages of all possible values of the system (otherwise we should deal all the time with statistical distribution and this would be a pain in the ass to do so...).

For summary, and in order of appearance and importance we have the following types of energy:

Symbol	Description
W	Work
Q	Heat
P	Pressure
T	Absolute temperature
V	Volume
E_{tot}	Total Energy
U	Internal Energy
S	Entropy
H	Enthalpy
F	Free energy
G	Free enthalpy

Table 9.22: **Most famous type of energy in thermodynamics**

Let us indicate that we will not return back on the concept of temperature T that is, as we have already seen in the section of Continuum Mechanics (virial theorem page 3897) and Statistical Mechanics (black-body radiation), a parameter that can advantageously link the average movement of different bodies with their average kinetic energy (meaning their "excitement" or "disorder") or respectively the radiation of certain bodies with their emission energy. But let us still recall the origin of the zero Kelvin, because this is a redundant question on the web:

We have proved in the section of Continuum Mechanics that at constant pressure P ("isobaric" system), the volume of a fixed amount of ideal gas is proportional to the absolute temperature. This is the "Gay-Lussac's law" (in the case of perfect gases ...!):

$$V = c^{te}T \tag{9.6.2}$$

and we have then mentioned that the experimental measurements then give a straight right, which extrapolated (somewhat abruptly ...) in negative values of temperature in Celsius gives zero volume for an ideal gas (neglecting quantum aspects to this border ...) at a temperature systematically of -273.15 [°C]:

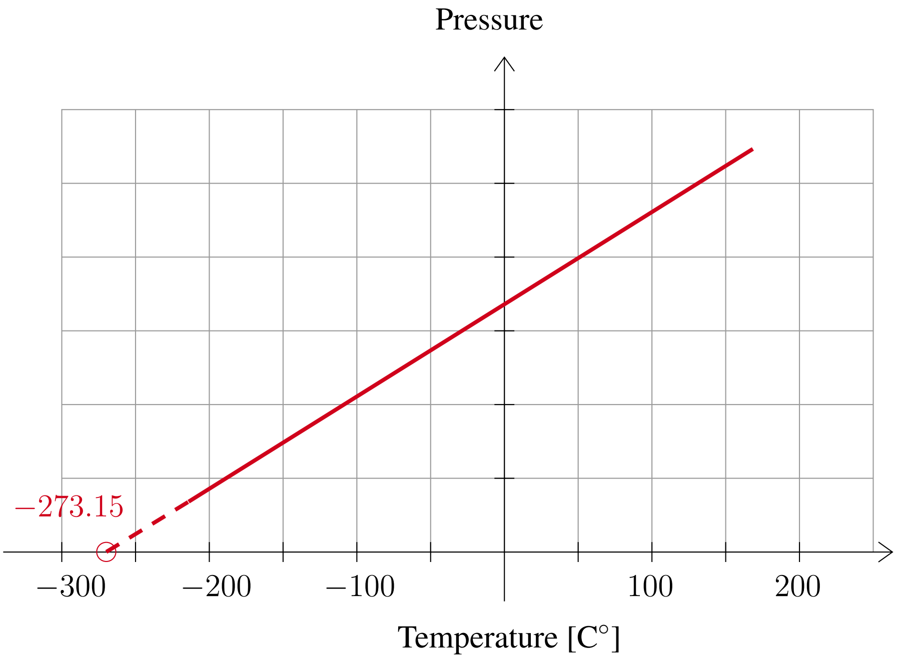


Figure 9.153: Representation of the Gay-Lussac law with pressure

hence the comfortable choice to put the 0 [K] at this value and redefine a new temperature scale!

Moreover, a difficulty is to talk about temperature change, as for example when we study the sensitivity of water to move from solid to liquid phase. We indeed observe experimentally that under ideal laboratory conditions when we move from -0.1 [°C] to $+0.1$ [°C] it is enough to observe the passage of qualitative change of water from liquid to solid state. Speak then of sensibility of $\%$ of temperature is difficult with the traditional scale in this case. But if we talk in absolute scale then this corresponds to a temperature changes from 273.05 to 273.25 [K], that is to say a phase transition sensitivity to water of 0.1% in temperature around its melting point.

9.6.2 Thermodynamics Systems

Most of time a thermodynamic system is the set of bodies located inside an imaginary closed surface defined by a "border" and often considered extensible without loss of energy.

The border can also under indication of its characteristics be material!

Let us notice that the border can be limited to an elementary surface $\mathrm{d}A$ (we take A to avoid confusion with the entropy S) associated with its normal vector wrapping a fluid particle (see the section Marine and Weather Engineering for example!). We name it in this case "particulate border".

It is often useful in thermodynamics, to the energy accounting that is transferred between the thermodynamic system and the external environment, that is to say, consider everything that crosses the border!

The main type of transfers (but not the only one!) that may be made are:

1. The "work transfer W": (Mechanical) Work that operate most of time at macroscopic level done by a force over a distance. When no work transfer (energy) is operated on a macroscopic scale, the system is named "system without work".
2. The "heat-transfer Q": (agitation) Energy from the variation in the number of microstates at the microscopic scale (see section Mechanical Statistics page 3549). When no heat-transfer (energy) is operated at the microscopic level, the system is named "adiabatic system", otherwise it is said "diathermal".
3. The "mass transfer M": mass injected into the system. When no mass transfer is made, the system is named a "closed system".

Definitions (#9.6.458):

D1. An "isolated system" can not exchange neither work nor heat nor mass with its surroundings.

D2. An "open system" can exchange work, heat and mass with its surroundings.

D3. A "closed system" can share work and heat but no matter with its surroundings.

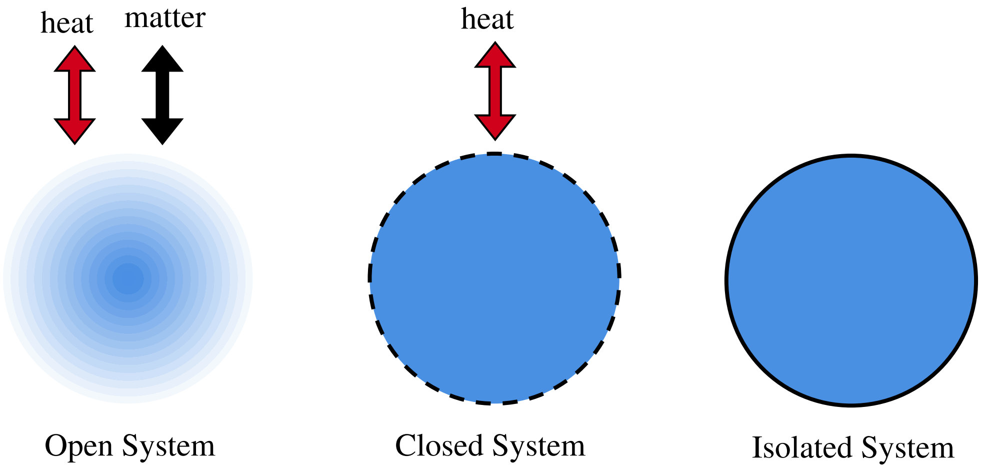


Figure 9.154: **Common type of thermodynamic systems**

Remarks

R1. Some systems count a bilan of external actions applied to them that is equal to zero. They are then say to be "pseudo-isolated".

R2. We speak of "homogeneous system" if the nature of its constituents is equal in every point, it is a "uniform system" or "isotropic system" if its characteristics are equal in every point.

9.6.3 Thermodynamic Transformations

Definition (#9.6.459): A "thermodynamic transformation" is the operation in which the system state changes passing from an initial state to a final state.

We distinguish at least two main types:

1. The "quasi-static thermodynamic transformation" that brings a system from an initial state to a final state through a succession of states that are exclusively equilibrium states (assumed as very slow transformation).

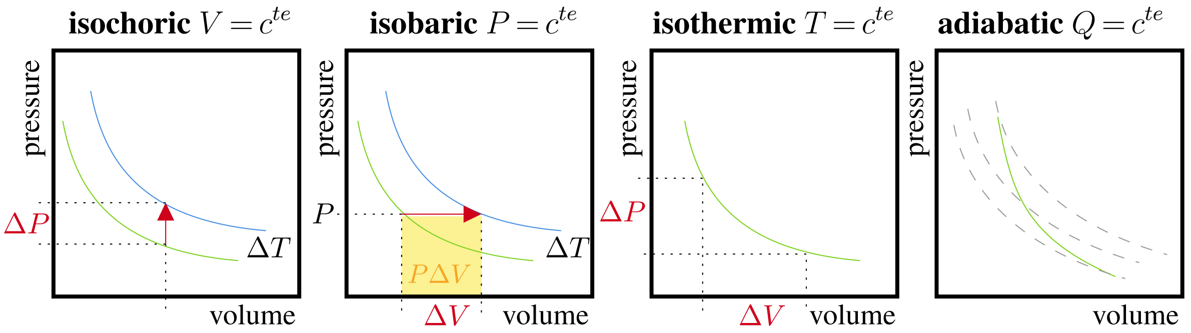
2. The transformations where all state variables change simultaneously are named "polytropic transformations".

It is possible that some variables remain constant during a thermodynamic transformation. In this case, we use very specific denominations:

Symbol	Description	Denomination
Q	Heat	Adiabatic
P	Pressure	Isobar
T	Absolute temperature	Isotherm
V	Volume	Isochore
U	Internal Energy	Isoenergetic
S	Entropy	Isentropic
H	Enthalpy	Isenthalpic
G	Free enthalpy	Extensive

Table 9.23: **Denomination of common constant systems**

Or visually with more relationships that we will prove further below:


isochoric $V = c^{te}$
pressure
ΔP
ΔT
volume
isobaric $P = c^{te}$
pressure
P
$P\Delta V$
ΔT
ΔV
volume
isothermic $T = c^{te}$
pressure
ΔP
ΔV
volume
adiabatic $Q = c^{te}$
pressure
volume

$$\begin{aligned}&\Delta P = nr\Delta T\\ &\mathrm{d}Q = \mathrm{d}U\\ &\mathrm{d}W = 0\\ &\mathrm{d}Q = nC_V\mathrm{d}T\end{aligned}$$

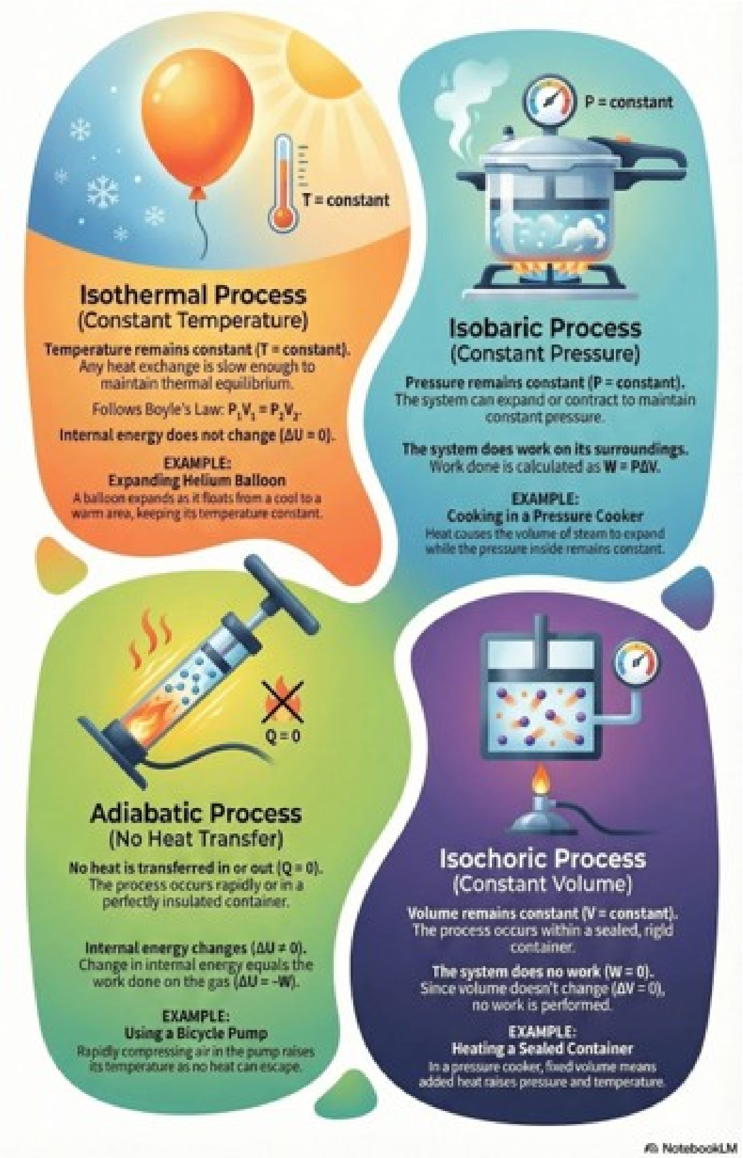
$$\begin{aligned}&\Delta \mathrm{V} = \mathrm{nr}\Delta \mathrm{T}\\ &\mathrm{d}H = \mathrm{d}Q - V\mathrm{d}P = \mathrm{d}Q\\ &\mathrm{d}W = P\mathrm{d}V\\ &\mathrm{d}U = \mathrm{d}Q - W\\ &\mathrm{d}Q = nC_V\mathrm{d}T\end{aligned}$$

$$\begin{aligned}&\Delta \mathrm{P}\Delta \mathrm{V} = \mathrm{c}^{\mathrm{te}}\\ &\mathrm{d}Q = -P\mathrm{d}V\\ &\mathrm{d}U = \mathrm{d}H = 0\\ &\mathrm{d}Q = \mathrm{d}W\\ &\Delta W = nR\ln\left(\frac{V_2}{V_1}\right)\end{aligned}$$

$$\begin{aligned}&\Delta P\Delta V = nR\Delta T\\ &\mathrm{d}Q = 0\\ &\mathrm{d}U = -P\mathrm{d}V\\ &\mathrm{d}H = V\mathrm{d}P\\ &\mathrm{d}U = -\mathrm{d}W\\ &\Delta W = -\frac{C_V}{R}(P_2V_2 - P_1V_1)\end{aligned}$$

Figure 9.155: **Most common type of thermodynamic processes illustrated**

Caution!!! An adiabatic transformation may be reversible (i.e. both the system and its environment can return to exactly the states they were in by following the reverse path) or irreversible (i.e. both the system and its environment cannot return together to exactly the states that they were in by following the reverse path), while an isothermal transformation is reversible for ideal gas.



Isothermal Process
(Constant Temperature)

Temperature remains constant (T = constant). Any heat exchange is slow enough to maintain thermal equilibrium.

Follows Boyle's Law: $P_1V_1 = P_2V_2$.

Internal energy does not change ($\Delta U = 0$).

EXAMPLE:
Expanding Helium Balloon
A balloon expands as it floats from a cool to a warm area, keeping its temperature constant.

Isobaric Process
(Constant Pressure)

Pressure remains constant (P = constant). The system can expand or contract to maintain constant pressure.

The system does work on its surroundings. Work done is calculated as $W = P\Delta V$.

EXAMPLE:
Cooking in a Pressure Cooker
Heat causes the volume of steam to expand while the pressure inside remains constant.

Adiabatic Process
(No Heat Transfer)

No heat is transferred in or out ($Q = 0$). The process occurs rapidly or in a perfectly insulated container.

Internal energy changes ($\Delta U \neq 0$). Change in internal energy equals the work done on the gas ($\Delta U = -W$).

EXAMPLE:
Using a Bicycle Pump
Rapidly compressing air in the pump raises its temperature as no heat can escape.

Isochoric Process
(Constant Volume)

Volume remains constant (V = constant). The process occurs within a sealed, rigid container.

The system does no work ($W = 0$). Since volume doesn't change ($\Delta V = 0$), no work is performed.

EXAMPLE:
Heating a Sealed Container
In a pressure cooker, fixed volume means added heat raises pressure and temperature.

Figure 9.156: **Main thermodynamics processes at a glance**

Remarks

R1. The studies of isobars systems are of great practical interest since all systems in contact with the atmosphere are often, at equilibrium, naturally or forcibly at constant pressure (that is to say at the same pressure as the surrounding atmosphere!) in laboratories.

R2. It is quite difficult to build useful isothermal transformations. Keep the temperature constant requires excellent thermal contact or isolation and a long reaction time to keep the temperature uniform. This type of transformation will be in the real world, rather slow.

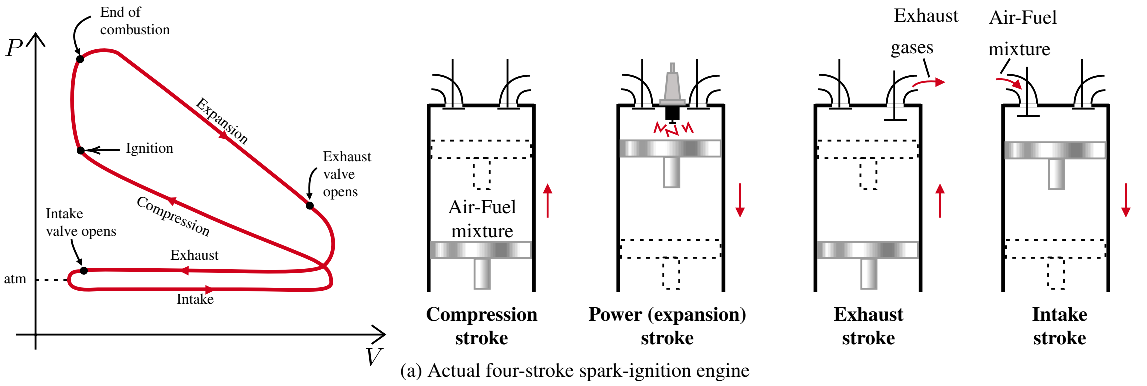
R3. Many real transformations are considered as adiabatic. It is enough for this that the container is well isolated or even the transformation is fast enough such that the heat exchanges are negligible.

Definition (#9.6.460): A "thermodynamic cycle" consists of a linked sequence of thermodynamic processes that involve transfer of heat and work into and out of the system, while varying pressure, temperature, and other state variables within the system, and that eventually returns the system to its initial state.

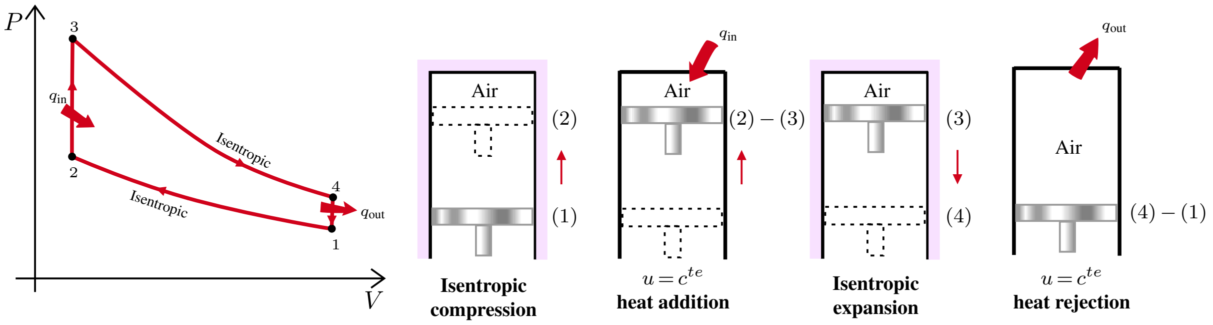
We also distinguish two main thermodynamic transformations cycles that are:

- The "closed thermodynamic cycle": the system described a series of transformations such as the final state and the initial state of the process are the same and the quantities and properties of the elements involved in the cycle are always the same.
- The "open thermodynamic cycle": the system describes a series of transformations such that the final state and the initial state of the process are also the same and the quantity and properties of the elements involved in the cycle are not always the same.

Below a figure with two illustrative examples (we will come back on the technical aspects further below):


P
End of combustion
Expansion
Ignition
Exhaust valve opens
Intake valve opens
Compression
Exhaust
atm
Intake
V
Exhaust gases
Air-Fuel mixture
Air-Fuel mixture
Compression stroke
Power (expansion) stroke
Exhaust stroke
Intake stroke

(a) Actual four-stroke spark-ignition engine


P
3
q_{in}
2
Isentropic
Isentropic
4
q_{out}
1
V
Air
(2)
(1)
Isentropic compression
q_{in}
Air
(2) − (3)
$u = c^{te}$
heat addition
Air
(3)
(4)
Isentropic expansion
q_{out}
Air
(4) − (1)
$u = c^{te}$
heat rejection

(b) Ideal Otto cycle

Figure 9.157: **Two typical thermodynamics cycles**

9.6.4 State Variables

Definition (#9.6.461): The "thermodynamic state" of a system is the set of properties that characterize it, regardless of the shape of its border. The variables that describe the state of the system by knowing only the final and initial state of the latter are named most of time "state functions" or more frequently "state variables" and sometimes even "state quantities"... Intuitively, the state of a system describes enough about the system to determine its future behaviour in the absence of any external forces affecting the system.

Until now in fact we already worked with state variables. Indeed, in mechanical systems, the position coordinates and velocities of mechanical parts are typical state variables. Knowing these, it is possible to determine the future state of the objects in the system.

In thermodynamics state variables include temperature, pressure, volume, internal energy, enthalpy, and entropy. In contrast heat and work are not state functions, but "process functions".

As we will see in the chapter of Electrodynamics, in electronic circuits, the voltages of the nodes and the currents through components in the circuit are usually the state variables. In Social Sciences the population sizes (or concentrations) of plants, animals and resources (nutrients, organic material) are typical state variables.

Remark

Some state variables play a special role in the definition of equilibrium states of a system. These are accessible quantities, at the macroscopic scale, directly or indirectly through measuring instruments. These special state functions (such as pressure, temperature, volume, etc.) are named sometimes "state variables of thermodynamic system at equilibrium".

The previous definition implicitly assumes the existence of a state and state variables, that is to say, the system characteristic parameters are defined (or theoretically accessible in the sense of measurement) at any time and at any point of the system. This is far from being easy if we consider the rapid changes of events such as explosions.

This difficulty can be studied by hiding us behind the "hypothesis of the local sate": We assume that at any time, the characteristic quantities have locally the same expressions as in a stationary configuration, which implies that the time necessary to state changes are negligible relatively to the characteristics periods of evolution.

The choice of the state variables depends on the nature of the problem studied. Nevertheless, we can separate all of these state variables at least into two groups:

D1. State variables are say to be "extensive quantities", if their are proportion to the quantity of matter and therefore the to the number of atoms/molecules of the system that are used to define them (therefore there are additive). This is typically the case of the mass M, the volume V, the Entropy S, the energy E, etc.

This means the system could be divided into any number of subsystems, and the extensive property measured for each subsystem; the value of the property for the system would be the sum of the property for each subsystem.

D2. State variables are say to be "intensive quantities", if their are independent to the quantity of matter and therefore the to the number of atoms/molecules of the system that are used to define them (therefore there are not additive). This is typically the case of the pressure P, the temperature T, the density ρ, the color, etc.

An intensive property is a therefore a bulk property, meaning that it is a physical property of a system that does not depend on the system size or the amount of material in the system

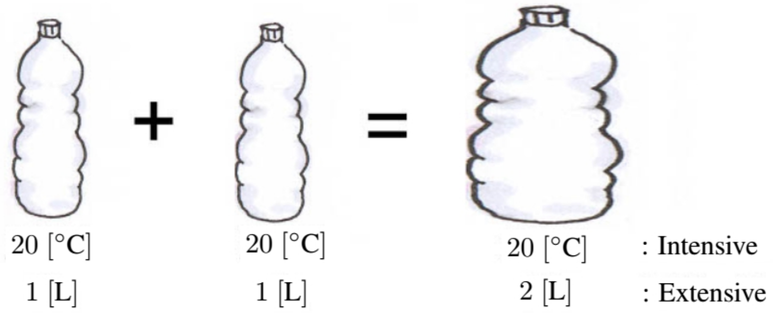


Figure 9.158: **Intensive vs Extensive variables**

In general, extensive variable depends on the point of view considered in the studied system while extensive variable are defined on the whole of the system.

Let us now show with a particular example that the ratio of two extensive variables is an intensive variable. For this, let us recall the ideal gas law (see section Continuum Mechanics page 3903) which is an equation of state (we will come back on state equations further below):

$$PV=N_{\text{Av}}nkT=nRT \tag{9.6.3}$$

which therefore includes three thermodynamic variables P, V and T (which leaves to us two choices of independent control variables and one dependent variable).

We can now calculate the concentration C (not to be confused with the notation of the heat capacity we will see later!), Which is therefore an intensive magnitude, by the ratio of two extensive quantities n, V such that:

$$C=\frac{n}{V}=\frac{P}{RT} \tag{9.6.4}$$

The state variables have also a special property: their variations are not dependent on the nature of the process that affects the system, but only the final and initial state of the system at equilibrium (which is very useful in practice ...!). This is the related to the concept of path integral that we have already discussed in detail in the section of Classical Mechanics and Differential and Integral Calculus.

Let us list the most familiar extensive and intensive quantities in thermodynamics:

Symbol	Legend	Quantity
Q	Heat	Extensive
P	Pressure	Intensive
T	Absolute temperature	Intensive
V	Volume	Extensive
E	Total Energy	Extensive
U	Internal Energy	Extensive
S	Entropy	Extensive
H	Enthalpy	Extensive
F	Free Energy	Extensive
G	Free enthalpy	Extensive

Table 9.24: **Common extensives and intensives state variables**

In contrast, the work W and heat Q are not a state variable because they depend on the nature of the transformation. However there are special cases where the heat and work are no longer dependent on the path followed when changes are made either at constant pressure or at constant volume (we will see this later).

9.6.4.1 Phases

Definitions (#9.6.462):

D1. A system in which various intensive variables vary continuously constitutes a "phase". In other words, a phase is a region of space (a thermodynamic system), throughout which all physical properties of a material are essentially uniform

We can therefore consider that any intensive magnitude depends only on the coordinates of the studied point: the system consists of a single phase if the intensive variable is continuous throughout the system. This is the case of gases, liquids and of some homogeneous solids.

D2. A "phase transition" is the transformation of a thermodynamic system from one phase or state of matter to another one by heat transfer. The term is most commonly used to describe transitions between solid, liquid and gaseous states of matter, and, in rare cases, plasma. A phase of a thermodynamic system and the states of matter have uniform physical properties. During a phase transition of a given medium certain properties of the medium change, often discontinuously, as a result of the change of some external condition, such as temperature, pressure, or others. For example, a liquid may become gas upon heating to the boiling point, resulting in an abrupt change in volume. The measurement of the external conditions at which the transformation occurs is termed the phase transition. Phase transitions are common in nature and used today in many technologies.

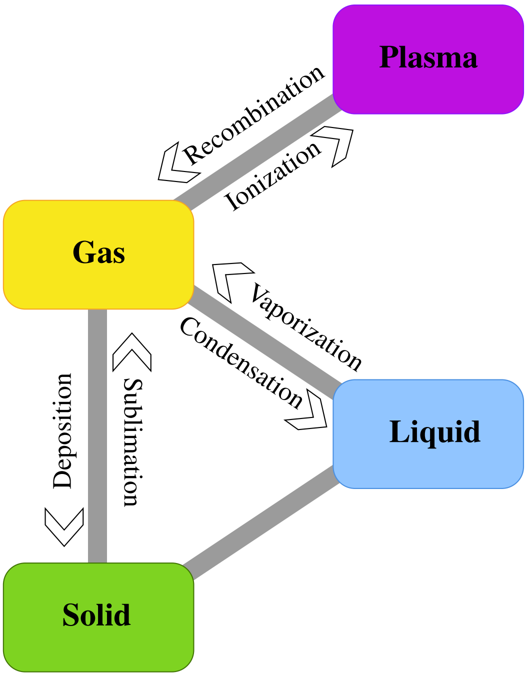


Figure 9.159: **Common phase transitions**

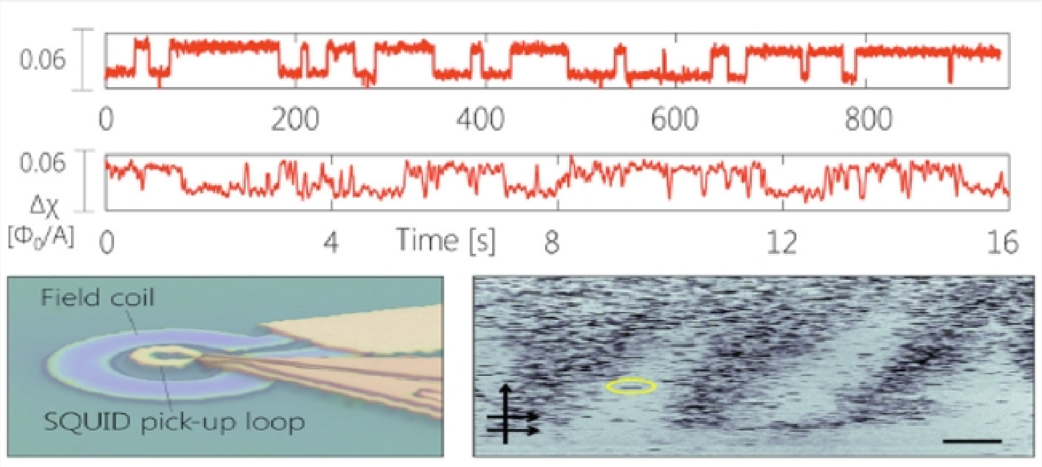
D3. If the intensive quantity has one discontinuity (or more), the system is said to be "multiphase". However, if the intensive quantities have the same value at any point in the system, the phase is said to be an "uniform phase": the system will have same temperature, pressure and composition in each of its points.

Remark

A "quantum phase transition" (QPT) occurs between two competing phases of matter at zero temperature, driven by quantum fluctuations. The presence of these fluctuations is well established (QPT represents quantum tunneling events between different states!), they have been locally imaged in space and their local dynamics has studied as they may be crucial for understanding why the structure of our universe is not homogeneous.

Scanning superconducting quantum interference device (SQUID) can be used to image quantum fluctuations in the vicinity of the QPT from a superconductor to an insulator (see [322]). Fluctuations of the diamagnetic response in both space and time that survive well below the transition temperature (in the figure below at $0.6T_c$), demonstrating their quantum nature.

By parking a SQUID (Superconducting Quantum Interference Device) over a superconducting elements it is possible to generated a map of the local diamagnetic susceptibility χ, revealing emergent puddles or regions where superconductivity is suppressed. The images reveal small streaks in the scan direction, which reflect dynamic events that can be tracked with an acquisition per pixel of 40 [ms]. The streaks manifest themselves as telegraph-like noise in the susceptibility signal with discrete jumps over a wide range of time scales (taken from [322]):



For a homogeneous system, it may be convenient to bring the extensive variables to the unit of mass. We speak then of "mass quantities" (or "specific quantities"), usually denoted in lower case.

We use if possible to avoid confusion the following scoring rules:

1. Any non-mass quantity is represented by an uppercase Latin letter $V, H, U, S, G, F, \ldots$ and then named respectively "volume", "enthalpy", "internal energy", "entropy", "free enthalpy", "free energy", and so on...
2. Any mass quantity is represented by a lowercase Latin letter $v, h, u, s, g, f, \ldots$ and then named respectively "specific volume", "specific enthalpy", "specific internal energy", "specific entropy", "specific free enthalpy", "specific free energy", and so on...

Otherwise, if we do not follow this notation, we will have to specify what type of variable we are dealing with.

Notice by the way that all specific variables are intensive variables!

9.6.5 Equation of State

In physics, particularly in thermodynamics, the "equation of state" of a system at thermodynamic equilibrium is a relation between different physical parameters, so named "state variables" already defined previously that determine its status. This may be for example of a relation between temperature, pressure and volume. From the characteristic equation of state of a physical system, it is possible to determine all the thermodynamic quantities describing this system and hence predict its properties.

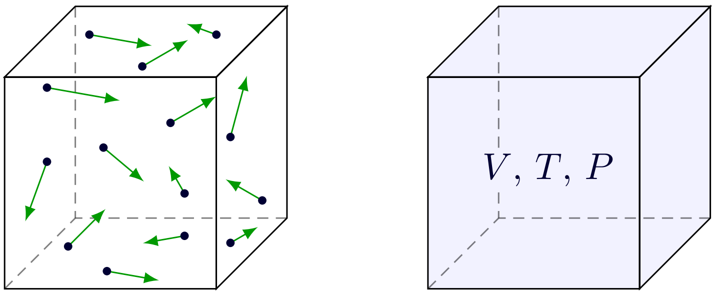
The equations of state are generally restricted to a type of behaviour or to given physical phenomena. One body can have several equations of state, for example the magnetic state or its thermodynamic state.

For a body to be characterized by an equation of state at a given time, it is necessary that the state of the body depends only of the values of the known parameters at this time. The body having a hysteresis phenomenon can not therefore be characterized by an equation of state.

9.6.5.1 Ideal Gas law

We have proved in the section of Continuum Mechanics (see page 3903) that the state equation of an ideal gas was (relation already introduced previously a little bit fortuitously earlier above):

$$\boxed{PV = nRT} \tag{9.6.5}$$



That is to say also:

$$V(P,T) = \frac{nRT}{P} \tag{9.6.6}$$

and the equation of state of real gases (also proved in the section of Continuum Mechanics):

$$\left(P + a\left(\frac{N}{V}\right)^2\right)(V - Nb) \cong kNT \tag{9.6.7}$$

To determine the particular and idealized equations of state of solids and liquids, it will first introduce several factors, named "thermo-elastic coefficients", used to characterize some global properties of these continuous media.

These coefficients should be defined carefully, depending on how we measure them, that is to say, depending on what is actually measured.

Let us take, for example, the concept of (volumetric) compressibility of a sample. It is reasonable to say that the underlying idea is the measurement of volume change under a change in pressure. If the pressure increases, the volume will decrease, so it seems reasonable to set a compressibility factor by:

$$\chi = -\frac{\partial V}{\partial P} \tag{9.6.8}$$

where the sign "$-$" is used to ensure a positive χ in the case of compression.

But this definition does not mean anything. We have not specified the conditions under which the measures are taken. Clearly, we increase (or diminish) the pressure measured using a manometer for example, and we measure with a ruler (!) the volume change that results. It sounds simple, but it is incomplete. If we increase the pressure fast enough, the volume will decrease but the temperature will increase. Do we measure the volume change immediately or we wait that the temperature is returned that it was by placing the sample in thermal contact with a thermostat that we used before measuring the volume change. In the latter case, the measurement has been made at constant temperature.

In other words, do we measure the volume change of a receptacle in perfect thermal contact with the external environment (we leave it the time to go to equilibrium) or thermally isolated from this environment. It is very different!

Furthermore, V being an extensive variable, there will be an advantage in defining the compressibility as an intensive quantity, that is to say independent of the amount of gas in question (here isothermal), which makes easier the construction of tables of values. We then have intuitively the following definition of the "(relative) isothermal coefficient of compressibility":

$$\chi_T=-\frac{1}{V}\left(\frac{\partial V}{\partial P}\right)_T \tag{9.6.9}$$

on which will be come back later more in details.

In the same intuitive we, we can define the "(relative) isobaric coefficient of compressibility" (more rarely named "(relative) volume isobaric coefficient of compressibility"):

$$\alpha_{V,P}=-\frac{1}{V}\left(\frac{\partial V}{\partial T}\right)_P \tag{9.6.10}$$

on which will be come back later more in details.

For example, for a perfect gas we have:

$$\alpha_{V,P}=\frac{1}{V}\left(\frac{\partial\left(\dfrac{nRT}{P}\right)}{\partial T}\right)_P=\frac{1}{V}\frac{nR}{P}=\frac{nR}{VP}=\frac{nR}{nRT}=\frac{1}{T} \tag{9.6.11}$$

We will later another approach to fall back on this equality.

9.6.5.2 State equation of a Liquid

Now the (macroscopic) state variables of a liquid are quite trivially the volume V, the pressure P and temperature T. Its equation of state can then be written in the form:

$$V=f(T,P) \tag{9.6.12}$$

Under differential form this give:

$$\mathrm{d}V=\left(\frac{\partial V}{\partial T}\right)_P\mathrm{d}T+\left(\frac{\partial V}{\partial P}\right)_T\mathrm{d}P \tag{9.6.13}$$

By introducing the thermoelastic coefficients as we did above, we get:

$$\frac{\mathrm{d}V}{V}=\alpha_{V,P}\mathrm{d}T-\chi_T\mathrm{d}P \tag{9.6.14}$$

Experience shows that for liquids, the thermoelastic coefficients vary very little with temperature and pressure (yes ... the experience is still useful sometimes ...). We can therefore assume them as constant, provided that the variations of the temperature and pressure are moderate. As the changes in volume of a liquid are very low when the transformations are slow, we can do the same approximation as that proved in the section of Continuum Mechanics with the true longitudinal deformation. Which means:

$$\frac{\mathrm{d}V}{V}\cong\frac{\mathrm{d}V}{V_0} \tag{9.6.15}$$

where V_0 is the constant reference volume. Taking into account these remarks, we get trivially by integration the general state equation of a liquid:

$$\boxed{V(T,P)=V_0[1+\alpha_{V,P}(T-T_0)-\chi_T(P-P_0)]} \tag{9.6.16}$$

Let us notice that the approximation of an (assumed...) incompressible liquid[23], commonly used in fluid dynamics, assume that:

$$\alpha_{V,P}=0 \qquad \chi_T=0 \tag{9.6.17}$$

The state equation of the liquid the becomes obviously:

$$V(T,P)=V_0=c^{te} \tag{9.6.18}$$

Let us also notice that in small classes, solids are considered sometimes as incompressible liquids. We then have the famous relation known by many students under the name of "volumetric dilatation formula":

$$\boxed{V(T,P)=V_0[1+\alpha_{V,P}(T-T_0)]} \tag{9.6.19}$$

These two formulations are exactly the same that the technical formulation of the Gay-Lussac law for gas (see section Continuum Mechanics page 3892).

We can do again the same reasoning with an element of length but by defining another thermoelastic coefficient: a linear one this time! Therefore, it comes another famous relation known by many students under the name of "length dilatation formula":

$$\boxed{\Delta L=\alpha_{L,P}L_0\Delta T} \tag{9.6.20}$$

We can then determine the relation between $\alpha_{L,P}$ and $\alpha_{V,P}$ (which are both much small than the unit as far as we know). Indeed, for a cube of sides L, we have:

$$\begin{aligned}\Delta V&=(L_0+\Delta L)^3-L_0^3=(L_0+\alpha_{L,P}L_0\Delta T)^3-L_0^3=L_0^3(1+\alpha_{L,P}\Delta T)^3-L_0^3\\&=L_0^3\left(1+3\alpha_{L,P}\Delta T+\underbrace{3\alpha_{L,P}^2\Delta^2}_{\to 0}+\underbrace{\alpha_L^3\Delta T^3}_{\to 0}\right)-L_0^3\cong 3\alpha_{L,P}L_0^3\Delta T\end{aligned} \tag{9.6.21}$$

---

[23] Keep in mind that even water is compressible otherwise sound wouldn't propagate in the latter...! In deep oceans for example the water is compressed by a factor of approximately 1.9%.

and therefore it comes:

$$\boxed{3\alpha_{L,P} \cong \alpha_{V,P}} \tag{9.6.22}$$

We have for example for water at room temperature:

$$\alpha_{V,P} \cong 2.6 \cdot 10^{-4}\,[\text{K}^{-1}] \qquad \chi_V \cong 4.5 \cdot 10^{-7}\,[\text{Pa}^{-1}] \tag{9.6.23}$$

✍**Example:**

A tank of volume V, with walls supposed non-deformable, is completely filled with water initially in thermodynamic equilibrium with the initial pressure P_0 of 1 [bar] and at the initial temperature T_0 of 298 [K]. After having received a certain amount of heat, the liquid reaches a new state of thermodynamic equilibrium, at the temperature of 398 [K]. We then have since $V=V_0$, the liquid state equation that reduces to:

$$\frac{V(T,P)}{V_0}=1=[1+\alpha_{V,P}(T-T_0)-\chi_V(P-P_0)] \tag{9.6.24}$$

Or after rearrangement:

$$P=\frac{\alpha_{V,P}}{\chi_V}(T-T_0)+P_0 \cong 1.58\,[\text{bar}] \tag{9.6.25}$$

Remark

The Eiffel Tower for example was well designed: it can oscillates because of the wind of ± 28 [cm], it can dilates vertically of ± 17 [cm] and because of the thermal radiation and of the path of the Sun (actually a G2V-type star, aka yellow dwarf) in the sky during a day it describes and ellipse of ± 18 [cm].

9.6.5.3 State equation of Solids

Let us now consider the special case of a beam (or a string) composed of an elastic material. The typical state variables of the system are then its length L, its temperature T and the tensile force F (positive in the case of traction and negative in the case of compression by convention). The equation of state is thus put in the following differential form:

$$\mathrm{d}L=\left(\frac{\partial L}{\partial T}\right)_F \mathrm{d}T+\left(\frac{\partial L}{\partial F}\right)_T \mathrm{d}F \tag{9.6.26}$$

By analogy with isobaric volumetric expansion coefficient $\alpha_{V,P}$, we define the "coefficient of linear expansion":

$$\alpha_{L,F}=\frac{1}{L}\left(\frac{\partial L}{\partial T}\right)_F \tag{9.6.27}$$

Furthermore, we proved in the section of Continuum Mechanics, that for an isothermal tensile test (without tangential component) performed on an elastic material we have (implicitly it is the Hooke's law for recall):

$$\frac{|\Delta L|}{L_0}=\frac{F/S}{E}=\frac{F}{ES} \tag{9.6.28}$$

But we can also use the true longitudinal deformation afterwards it is then obviously necessary to adapt the right term accordingly:

$$\frac{\mathrm{d}L}{L}=\frac{\mathrm{d}F}{ES} \tag{9.6.29}$$

Therefore it comes:

$$\frac{1}{ES}=\frac{1}{L}\left(\frac{\partial L}{\partial F}\right)_T \tag{9.6.30}$$

The state equation of solids can then be written by injecting inside the latter relation:

$$\frac{\mathrm{d}L}{L}=\alpha_{L,F}\mathrm{d}T+\frac{\mathrm{d}F}{ES} \tag{9.6.31}$$

This relation gives us the variation of relative elongation corresponding to an infinitesimal change of the tensioning force and of the temperature, knowing the linear expansion coefficient .

The above state equation is easily integrated assuming that the coefficient of linear expansion, Young's modulus and sectional area are constant (let us know as always if you want the details):

$$\boxed{L=L_0e^{\alpha_L\Delta T+\frac{\Delta F}{ES}}=L_0e^{\alpha_L\Delta T}e^{\frac{\Delta F}{ES}}} \tag{9.6.32}$$

If the deformations are very small, we have proved in the section of Continuum Mechanics that the relative variation of elongation could be written:

$$\frac{\mathrm{d}L}{L}\cong\frac{\mathrm{d}L}{L_0} \tag{9.6.33}$$

$$\frac{\mathrm{d}L}{L_0}=\alpha_{L,F}\mathrm{d}T+\frac{\mathrm{d}F}{ES} \tag{9.6.34}$$

So after integration:

$$\boxed{\frac{\Delta L}{L_0}=\alpha_{L,F}\Delta T+\frac{\Delta F}{ES}} \tag{9.6.35}$$

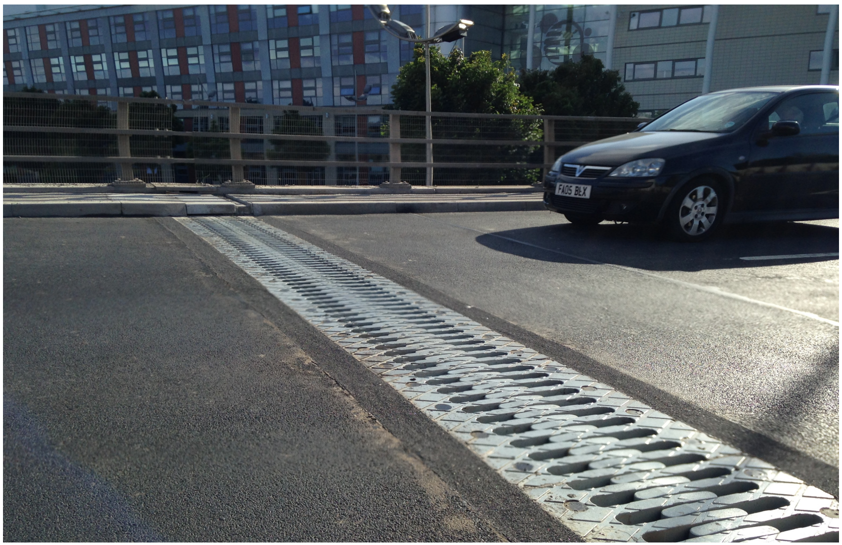


Figure 9.160: **Thermal expansion joints like these allow bridges to change length without buckling (source: ?)**

✍**Example:**

A cylindrical steel beam of radius 15 [mm] is initially in a state of thermodynamic equilibrium characterized by an initial length of 1 [m] at the initial temperature of 300 [K] and a zero initial tension force. The beam is heated to a new equilibrium in which the final temperature is 400 [K]. To calculate the new length in the particular case that the beam is free to move and that the deformation is considered as small:

$$\frac{L-L_0}{L_0}=\alpha_L(T-T_0)+\frac{\overbrace{\Delta F}^{}_{\cong 0}}{ES} \tag{9.6.36}$$

Therefore:

$$L=L_0\alpha_L(T-T_0)+L_0=1\cdot 12\cdot 10^{-6}\cdot 100+1=1.0012\,[\text{m}] \tag{9.6.37}$$

If the beam is embed at the both sides, we can then calculate the internal stress force generated ("thermal stress") under the assumption of small deformations. We then use:

$$\frac{\underbrace{\Delta L}_{\cong 0}}{L_0}=\alpha_L(T-T_0)+\frac{\Delta F}{ES} \tag{9.6.38}$$

Therefore:

$$\begin{aligned}F&=-\alpha_L ES(T-T_0)+F_0=-\alpha_L ES(T-T_0)\\&=-12\cdot 10^{-6}\cdot 2\cdot 10^{11}\pi\cdot 0.015^2\cdot 100\cong 170\,[\text{kN}]\end{aligned} \tag{9.6.39}$$

We see here a phenomenon well known in mechanical engineering: the thermal deformation of a solid is usually very small, but prevent it can generate tremendous efforts, that could destroy a mechanism.

9.6.6 Laws of Thermodynamics

The principles of thermodynamics are the bricks of thermal or mechanical energy. Each principle involves a lot of concepts that we will try to present and define as best as we can further below. This is the "sensitive" part of this field of physics.

Thermodynamics is based on four basic laws (some of which are provable):

P0. The "zeroth law of thermodynamics" or "principle of thermal equilibrium" is defined by the fact that if two thermodynamic systems 1 and 2 are in thermodynamic equilibrium with a third one 3, they are themselves in thermodynamic equilibrium (it is an "assertion" in the language of Proof Theory). This can be denoted formally by:

$$\boxed{\lim_{t\to+\infty} T_{M_1}(t,T_{M_2})=\lim_{t\to+\infty} T_{M_2}(t,T_{M_1})} \tag{9.6.40}$$

So if two materials in contacts are in thermal equilibrium, they have the same temperature and thus the state remains stationary. Two materials at the same temperature in contact are in thermal equilibrium.

> **Remark**
>
> We used this principle implicitly (when we stated that the steady state was the most likely one) in the section of Statistical Mechanics to prove the Boltzmann law.

P1. The "first law of thermodynamics" or "conservation principle" refers to the conservative nature of energy and states that during any of a system transformation, the change in its total energy is equal to the sum of changes in all types of energy defining it:

$$\boxed{\Delta E_{\text{tot}}=\Delta U+\Delta E_c+\Delta E_p} \tag{9.6.41}$$

If the system is isolated then the change in total energy will obviously always be zero (in average...)!

> **Remark**
>
> Depending on the authors, the kinetic energy and potential energy are part of the internal energy U.

Of course, when the system studied (the container and its content) are in the same repository as the observer, it remains only the term of internal energy variation:

$$\boxed{\Delta E_{\text{tot}}=\Delta U} \tag{9.6.42}$$

or in differential terms:

$$\mathrm{d}E_{\text{tot}}=\mathrm{d}U \tag{9.6.43}$$

It remains to determine the exact expression of the change in internal energy U and variables on which it depends (developments that we will do a little further below).

Remark

This principle is provable if we accept Noether's theorem (see section Principia page 3132) on time invariance of physics laws of physics as a higher principle.

P2. The "second law of thermodynamics" also named "Carnot-Clausius' principle" or "principle of evolution" refers to the irreversible nature and is associated with the concept of entropy. This principle, the most famous one, states that heat can only pass itself from a warm body to a cooler body (or a body will inevitably be cooler with time) in a closed system (i.e. without increase or decrease of energy) in a non-changing space-time. The reverse operation requiring the contribution of mechanical work took on the external system that is given by equation (proved a little further below):

$$\boxed{\mathrm{d}S=\frac{\delta Q}{T}\geq 0} \tag{9.6.44}$$

The left hand side of equality states that entropy can only increase and that the amount of heat exchange with the external system is always positive (numerator term right to equality). Therefore the body delivers heat to the external system and cooled down inexorably.

The special notation δQ is here to indicate that heat is an inaccurate total differential. This is a concept to which we will return further below.

Remark

We will prove roughly ... this relation further below by using the Boltzmann law on the access to the information in a system such as studied in the section of Statistical Mechanics.

P3. The "third law of thermodynamics" also named "Nernst principle" concern the properties of matter in the vicinity of absolute zero and states that at the limit of the absolute zero temperature that can not be reached (see section Wave Quantum Physics page 4814), the equilibrium entropy of a system approaches a constant independent of the other intensive parameters, constant that is taken as zero. Or in other words: The entropy of a perfect crystal at absolute zero is exactly equal to zero.

Proof 9.10.4. This statement is based on the proof of the "Nernst theorem" which formal result is simply based on the second law giving the entropy (see section Statistical Mechanics page 3553):

$$S=k\ln(\Omega) \tag{9.6.45}$$

We see that if we consider a crystal at absolute zero temperature, the particles position against each other are perfectly uniquely defined. Therefore, there is only one possible complexion for crystal at absolute zero. The number of complexions Ω is equal to 1. This gives us:

$$\boxed{S=k\ln(1)=0} \tag{9.6.46}$$

■ Q.E.D.

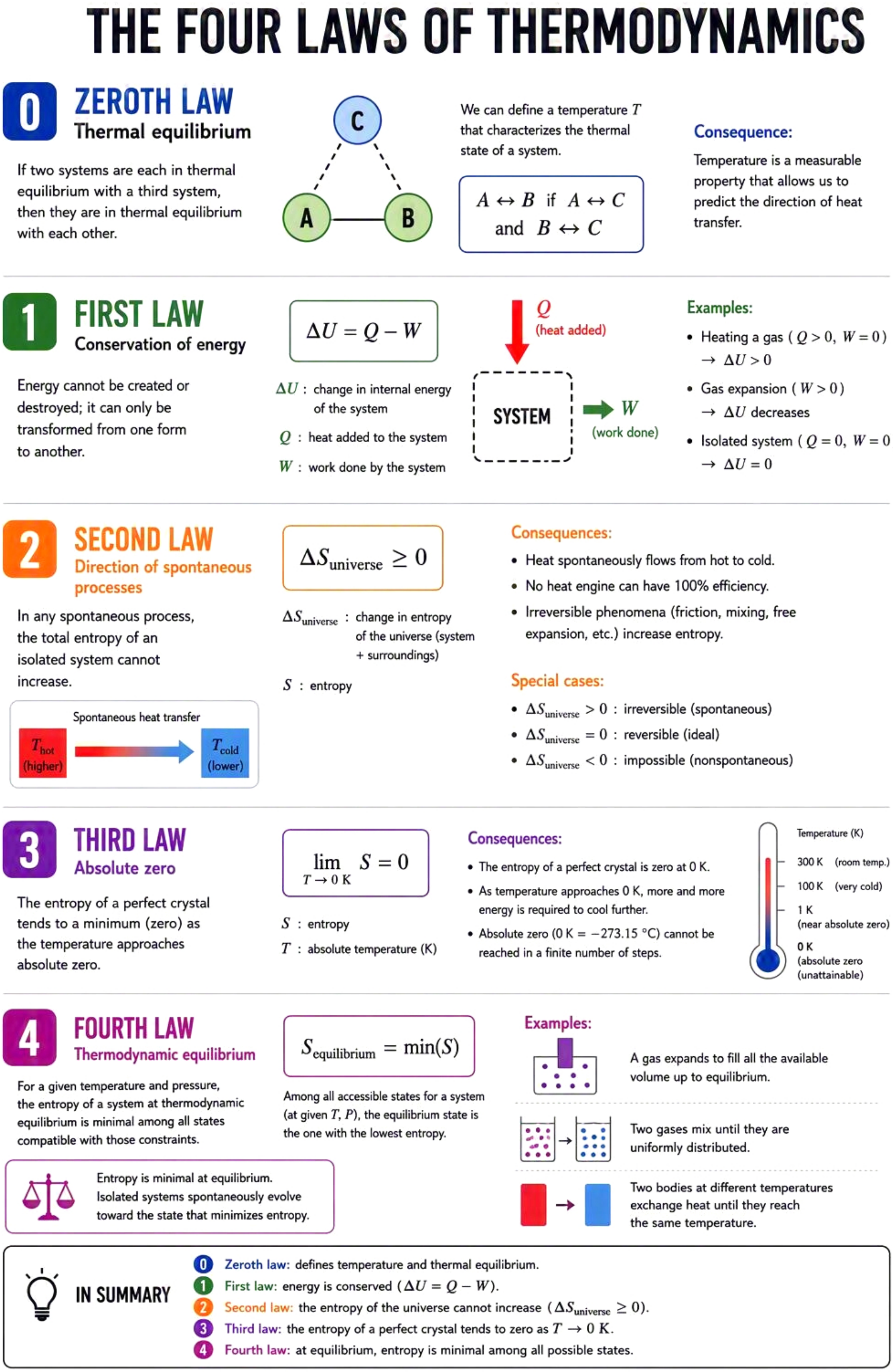
Let us recall that the "absolute zero" seems to be the lowest temperature experimentally reachable (theoretically - using quantum information theory - the value is derived by the fact we should need an infinite amount of energy to reach it) available in physics. At this temperature a substance no longer contains, at a macroscopic scale, the thermal energy (or heat) required for the occupation of several microscopic energy levels (see section Statistical Mechanics page 3549). The particles that make it up (atoms, molecules) are all in the same minimum energy state (ground state). This results in a null entropy due to the indistinguishability of the particles in the same fundamental energy level of entropy (following the third law of thermodynamics) and a total immobility in the conventional sense following the virial theorem (see section Continuum Mechanics page 3897).

However, in the quantum point of view, we will see in the section of Wave Quantum Physics that particles always have a non-zero linear momentum following the Heisenberg uncertainty relations and that following the Fermi-Dirac and Bose-Einstein distributions (see section Statistical Mechanics page 3571 and page 3576), there are still different energy levels when we take into account the Pauli exclusion principle. Hence the emergence of quantum thermodynamics.

Remark

A common question that arise is why the zeroth law of thermodynamics is named "zeroth law", why didn't physicists start the counting from one as in almost all other fields of physics? Indeed laws of physics are numbered according to their fundamental nature, the more fundamental the law, the higher it is in the pecking order of the laws. E.g., Newton gave the three laws of motion in one go, but he numbered the First Law of motion as the "first law" because it was the most fundamental of all the three law. Without understanding the first law, you cannot go to the second and the third laws.

As far as Thermodynamics is concerned, the First Law had been in currency and had been accepted for long. But when some scientist postulated the (now named) "Zeroth law", the scientific community came to the conclusion that this new law is more fundamental than the First Law because it provides the basis for defining an entity named "Temperature". So they decided to name it the "Zeroth Law", i.e. more fundamental than the First Law.



THE FOUR LAWS OF THERMODYNAMICS

0 ZEROTH LAW
Thermal equilibrium

If two systems are each in thermal equilibrium with a third system, then they are in thermal equilibrium with each other.

We can define a temperature T that characterizes the thermal state of a system.

$$\boxed{A \leftrightarrow B \text{ if } A \leftrightarrow C \text{ and } B \leftrightarrow C}$$

Consequence:

Temperature is a measurable property that allows us to predict the direction of heat transfer.

1 FIRST LAW
Conservation of energy

Energy cannot be created or destroyed; it can only be transformed from one form to another.

$$\boxed{\Delta U = Q - W}$$

ΔU : change in internal energy of the system

Q : heat added to the system

W : work done by the system

Examples:

- Heating a gas ($Q > 0,\ W = 0$) → $\Delta U > 0$
- Gas expansion ($W > 0$) → ΔU decreases
- Isolated system ($Q = 0,\ W = 0$) → $\Delta U = 0$

2 SECOND LAW
Direction of spontaneous processes

In any spontaneous process, the total entropy of an isolated system cannot increase.

Spontaneous heat transfer
T_{hot} (higher)
T_{cold} (lower)

$$\boxed{\Delta S_{\text{universe}} \geq 0}$$

$\Delta S_{\text{universe}}$: change in entropy of the universe (system + surroundings)

S : entropy

Consequences:

- Heat spontaneously flows from hot to cold.
- No heat engine can have 100% efficiency.
- Irreversible phenomena (friction, mixing, free expansion, etc.) increase entropy.

Special cases:

- $\Delta S_{\text{universe}} > 0$: irreversible (spontaneous)
- $\Delta S_{\text{universe}} = 0$: reversible (ideal)
- $\Delta S_{\text{universe}} < 0$: impossible (nonspontaneous)

3 THIRD LAW
Absolute zero

The entropy of a perfect crystal tends to a minimum (zero) as the temperature approaches absolute zero.

$$\boxed{\lim_{T \to 0\,\text{K}} S = 0}$$

S : entropy

T : absolute temperature (K)

Consequences:

- The entropy of a perfect crystal is zero at 0 K.
- As temperature approaches 0 K, more and more energy is required to cool further.
- Absolute zero (0 K = −273.15 °C) cannot be reached in a finite number of steps.

4 FOURTH LAW
Thermodynamic equilibrium

For a given temperature and pressure, the entropy of a system at thermodynamic equilibrium is minimal among all states compatible with those constraints.

$$\boxed{S_{\text{equilibrium}} = \min(S)}$$

Among all accessible states for a system (at given T, P), the equilibrium state is the one with the lowest entropy.

Entropy is minimal at equilibrium. Isolated systems spontaneously evolve toward the state that minimizes entropy.

Examples:

A gas expands to fill all the available volume up to equilibrium.

Two gases mix until they are uniformly distributed.

Two bodies at different temperatures exchange heat until they reach the same temperature.

IN SUMMARY

- ⓪ Zeroth law: defines temperature and thermal equilibrium.
- ① First law: energy is conserved ($\Delta U = Q - W$).
- ② Second law: the entropy of the universe cannot increase ($\Delta S_{\text{universe}} \geq 0$).
- ③ Third law: the entropy of a perfect crystal tends to zero as $T \to 0$ K.
- ④ Fourth law: at equilibrium, entropy is minimal among all possible states.

Figure 9.161: Summary of the four laws of the thermodynamics

9.6.7 Calorific Capacities (heat capacity)

In a given transformation, an intake of heat ΔQ generally results in a raising ΔT of the temperature of the system.

Definition (#9.6.463): We name "heat capacity" (or "calorific capacity", or "thermal capacity") C of system, the quantity:

$$\boxed{C=\frac{\Delta Q}{\Delta T}} \tag{9.6.47}$$

The heat capacity is therefore by its units, the energy it takes to bring a body to increase its temperature by 1 Kelvin. Under infinitesimal form, we have:

$$\boxed{C=\frac{\delta Q}{\partial T}} \tag{9.6.48}$$

We see that the educational benefit of having a notation using discrete variations rather than infinitesimal allows us not have to ask the question if we have an exact total differential or inaccurate one... (see below for the distinction).

Heat capacity is obviously an extensive property of matter, meaning it is proportional to the size of the system. When expressing the same phenomenon as an intensive property, the heat capacity is divided by the amount of substance, mass, or volume, so that the quantity is independent of the size or extent of the sample:

$$\boxed{c=\frac{\Delta Q}{\Delta T}\frac{1}{M}} \tag{9.6.49}$$

and then we speak of "specific heat".

The molar heat capacity is the heat capacity per unit amount (SI unit: mole) of a pure substance (see section Analytical Chemistry page 6228).

In the case of an ideal gas, we assume that all of the exchanged heat (energy) is changed into kinetic energy of the atoms or molecules (thermal energy), so we can write:

$$C=\frac{\Delta Q}{\Delta T}=\frac{\Delta U}{\Delta T}=\frac{\Delta E_c}{\Delta T} \tag{9.6.50}$$

and we will prove later how to determine the exact value of this capability for such gas.

However, let us already see how to determine this heat capacity for crystalline solids (which we will assume at constant volume). To do this, remember that we proved in the chapter of Electrokinetics (see page 4279) that in a crystal, the density of states in a volume under certain very specific conditions was:

$$D(\omega)=\frac{\mathrm{d}N}{\mathrm{d}k}\frac{\mathrm{d}k}{\mathrm{d}\omega}=\frac{Vk^2}{2\pi^2}\frac{\mathrm{d}k}{\mathrm{d}\omega} \tag{9.6.51}$$

and the number of related states was given by:

$$N=\frac{k^3}{6\pi^2}V \tag{9.6.52}$$

However, the prior-previous relation is independent of the fact that in the field of electrokinetic we consider the electron or any other particle vector of some property (electric charge, heat or other). Thus in the context of the thermodynamic crystalline solids, a model consist to assume the heat vector as a particle having similar wave properties to that of an electron that it is customary to name a "phonon".

Assuming that the phonons are identical in all three spatial directions and remembering the Planck-Einstein relation (see section Corpuscular Quantum Physics page 4627) valid for any particle by the principle of wave-particle duality:

$$E=h\nu \tag{9.6.53}$$

and (see section Statistical Mechanics page 3576) from the Bose-Einstein statistics for bosons (particles not subjected to the Pauli exclusion principle for recall ...):

$$f_{\text{BE}}(E,T,\mu)=\frac{1}{e^{\frac{E-\mu}{kT}}-1} \tag{9.6.54}$$

which we write in the present case as following (since the chemical potential is zero in our case):

$$n(\omega)=\frac{1}{e^{\frac{E}{kT}}-1}=\frac{1}{e^{\frac{h\nu}{kT}}-1}=\frac{1}{e^{\frac{\hbar\omega}{kT}}-1} \tag{9.6.55}$$

we have by combining these three big results:

$$U=3\int\limits_{\omega}h\nu(\omega)D(\omega)\mathrm{d}\omega=3\int\limits_{\omega}h\frac{\omega}{2\pi}n(\omega)D(\omega)\mathrm{d}\omega=3\int\limits_{\omega}\hbar\omega n(\omega)D(\omega)\mathrm{d}\omega=3\int\limits_{\omega}\frac{\hbar\omega}{e^{\frac{\hbar\omega}{kT}}-1}D(\omega)\mathrm{d}\omega \tag{9.6.56}$$

Then we have:

$$\begin{aligned}C_V&=\left(\frac{\delta Q}{\partial T}\right)_V=\left(\frac{\mathrm{d}U-\delta W}{\partial T}\right)=\left(\frac{\mathrm{d}U-P\underbrace{\mathrm{d}V}_{=0}}{\partial T}\right)=\left(\frac{\mathrm{d}U}{\partial T}\right)_V\\&=3\int\limits_{\omega}\hbar\omega\frac{\partial}{\partial T}\left(\frac{1}{e^{\hbar\omega/kT}-1}\right)D(\omega)\mathrm{d}\omega=3\int\limits_{\omega}\hbar\omega\left(\frac{\hbar\omega}{kT}\frac{e^{\hbar\omega/kT}}{\left(e^{\hbar\omega/kT}-1\right)^2}\right)D(\omega)\mathrm{d}\omega\\&=3\int\limits_{\omega}\hbar\omega\left[-\frac{1}{\left(e^{\hbar\omega/kT}-1\right)^2}\left(-\frac{\hbar\omega}{kT^2}e^{\hbar\omega/kT}\right)\right]D(\omega)\mathrm{d}\omega\\&=3k\int\limits_{\omega}\left(\frac{\hbar\omega}{kT}\right)^2\frac{e^{\hbar\omega/kT}}{\left(e^{\hbar\omega/kT}-1\right)^2}D(\omega)\mathrm{d}\omega\end{aligned} \tag{9.6.57}$$

From here there are two main models for predicting the heat capacity of crystals, the "Debye model" and the "Einstein model", each making assumptions about the density of phonon modes. The Debye model treats the vibrations of the atomic lattice (heat) - named "Debye solid" as phonons in a box, in contrast to the Einstein (photoelectron) model, which treats the solid - named "Einstein solid" as many individual, non-interacting quantum harmonic oscillators.

In the Einstein model a single frequency mode (respectively pulsation) is supposed possible, so that the vibration of the solid is represented by N harmonic oscillators at this frequency. This particular pulse is often associated with what we name the "Einstein temperature" (also often denoted T_E):

$$\Theta_E=\frac{\hbar\omega_E}{k} \tag{9.6.58}$$

The density of modes is then symbolized by N Dirac distributions (see section Differential and Integral Calculus page 1569) around such that:

$$\begin{aligned}C_V&=3k\int\limits_{\omega}\frac{e^{\hbar\omega/kT}}{(e^{\hbar\omega/kT}-1)^2}N\delta(\omega_E)\mathrm{d}\omega=3kN\left(\frac{\hbar\omega_E}{kT}\right)^2\frac{e^{\hbar\omega_E/kT}}{(e^{\hbar\omega_E/kT}-1)^2}\\&=3kN\left(\frac{\Theta}{T}\right)^2\frac{e^{\Theta/T}}{(e^{\Theta/T}-1)^2}\end{aligned} \tag{9.6.59}$$

This model provides a limit value of the heat capacity at high temperatures (we use a Taylor series approximation):

$$\begin{aligned}C_{V,T\rightarrow+\infty}&\cong3kN\left(\left(\frac{\hbar\omega}{kT}\right)^2\frac{1+\dfrac{\hbar\omega}{kT}}{\left(\left(1+\dfrac{\hbar\omega}{kT}\right)-1\right)^2}\right)=3kN\left(\left(\frac{\hbar\omega}{kT}\right)^2\frac{1+\dfrac{\hbar\omega}{kT}}{\left(\dfrac{\hbar\omega}{kT}\right)^2}\right)\\&=3kN\left(1+\frac{\hbar\omega}{kT}\right)\cong3kN\end{aligned} \tag{9.6.60}$$

It is customary when we presented this relation to consider a quantity of material corresponding to a mole to make appear the ideal gas constant:

$$\boxed{C_{V,T\rightarrow+\infty}\cong3kN_{\text{A}}=3R} \tag{9.6.61}$$

This value was originally given by Dulong and Petit and is in fairly good agreement with experimental results. We often talk about the "Dulong-Petit law".

In contrast at low temperatures, it is necessary to use the Debye model that consist to assume as linear dispersion relation with the speed of propagation of phonons (speed of sound in the crystal) supposed to be independent of temperature:

$$\omega(k)=vk \tag{9.6.62}$$

The density of states is then written:

$$D(\omega)=\frac{Vk^2}{2\pi^2}\frac{\mathrm{d}k}{\mathrm{d}\omega}=\frac{Vk^2}{2\pi^2}\frac{\mathrm{d}\left(\dfrac{\omega}{v}\right)}{\mathrm{d}\omega}=\frac{Vk^2}{2\pi^2v}=\frac{V\left(\dfrac{\omega}{v}\right)^2}{2\pi^2v}=\frac{V\omega^2}{2\pi^2v^3} \tag{9.6.63}$$

We then have for the heat capacity:

$$U=3\int\limits_{\omega}\hbar\omega\frac{1}{e^{\hbar\omega/kT}-1}\frac{V\omega^2}{2\pi^2v^3}\mathrm{d}\omega=\frac{3V\hbar}{2\pi^2v^3}\int\limits_{\omega}\frac{\omega^3}{e^{\hbar\omega/kT}-1}\mathrm{d}\omega \tag{9.6.64}$$

In the Debye model, only the acoustic phonon branch is considered, this is equivalent to schematize the crystal as a mono-atomic network formed of heavy atoms. Therefore, the resonance modes are limited to a certain frequency (pulse), with the corresponding volume, known as "cutting frequency" (we change the notation of the Boltzmann constant in K_B so to confuse it with the wave number k):

$$\begin{aligned}U&=\frac{3V\hbar}{2\pi v^3}\int\limits_0^{\omega_D}\frac{\omega^3}{e^{\hbar\omega/kT}-1}\mathrm{d}\omega=\frac{3V\hbar}{2\pi^2v^3}\int\limits_0^{x_D}\frac{\left(\frac{k_BT}{\hbar}x\right)^3}{e^x-1}\frac{k_BT}{\hbar}\mathrm{d}x\\&=\frac{3V\hbar}{2\pi^2v^3}\left(\frac{k_BT}{\hbar}\right)^4\int\limits_0^{x_D}\frac{x^3}{e^x-1}\mathrm{d}x=\frac{3V\hbar}{2\pi^2}\frac{k_BT}{\hbar}\left(\frac{k_BT}{v\hbar}\right)^3\int\limits_0^{x_D}\frac{x^3}{e^x-1}\mathrm{d}x\\&=\frac{3V}{2\pi^2}k_BT\left(\frac{k_BT}{v\hbar}\right)^3\int\limits_0^{x_D}\frac{x^3}{e^x-1}\mathrm{d}x=\frac{3\dfrac{N6\pi^2}{k^3}}{2\pi^2}k_BT\left(\frac{k_BT}{v\hbar}\right)^3\int\limits_0^{x_D}\frac{x^3}{e^x-1}\mathrm{d}x\\&=9Nk_BT\left(\frac{T}{T_D}\right)^3\int\limits_0^{x_D}\frac{x^3}{e^x-1}\mathrm{d}x\end{aligned}\tag{9.6.65}$$

where we have (we indicate here the two most common forms):

$$\boxed{T_D=\frac{v\hbar}{k_B}\left(6\pi^2\frac{N}{V}\right)^{1/3}=\frac{v_{\text{sound}}h}{2k_BL}\left(\frac{6N}{\pi}\right)^{1/3}}\tag{9.6.66}$$

We then have for the heat capacity:

$$\begin{aligned}C_V&=\left(\frac{\partial U}{\partial T}\right)_V=\frac{3V\hbar}{2\pi^2v^3}\frac{\partial}{\partial T}\int\limits_0^{\omega_D}\frac{\omega^3}{e^{\hbar\omega/k_BT}-1}\mathrm{d}\omega=\frac{3V\hbar}{2\pi^2v^3}\int\limits_0^{\omega_D}\frac{\partial}{\partial T}\frac{\omega^3}{e^{\hbar\omega/k_BT}-1}\mathrm{d}\omega\\&=\frac{3V\hbar^2}{2\pi^2v^3}\int\limits_0^{\omega_D}\left(-\frac{\omega^3}{\left(e^{\hbar\omega/k_BT}\right)^2}\cdot\left(-\frac{\hbar\omega}{k_BT^2e^{\hbar\omega/k_BT}}\right)\right)\mathrm{d}\omega=\frac{3V\hbar}{2\pi^2v^3}\frac{\hbar}{k_BT^2}\int\limits_0^{\omega_D}\frac{\omega^4e^{\hbar\omega/k_BT}}{\left(e^{\hbar\omega/k_BT}-1\right)^2}\mathrm{d}\omega\\&=\frac{3V\hbar^2}{2\pi^2v^3k_BT^2}\int\limits_0^{x_D}\frac{\left(\dfrac{k_BT}{\hbar}x\right)^4e^x}{(e^x-1)^2}\frac{k_BT}{\hbar}\mathrm{d}x=\frac{3V\hbar}{2\pi^2v^3T}\left(\frac{k_BT}{\hbar}\right)^4\int\limits_0^{x_D}\frac{x^4e^x}{(e^x-1)^2}\mathrm{d}x\\&=9Nk_B\left(\frac{T}{T_D}\right)^3\int\limits_0^{x_D}\frac{x^4e^x}{(e^x-1)^2}\mathrm{d}x\end{aligned}\tag{9.6.67}$$

If the absolute temperature is small, then we have:

$$C_{V,T\rightarrow 0}=9N_k\left(\frac{T}{T_D}\right)^3\int\limits_0^{+\infty}\frac{x^4e^x}{(e^x-1)^2}\mathrm{d}x\tag{9.6.68}$$

What is quite difficult to calculate. So we prefer rather use the following approach:

$$C_{V,T\rightarrow 0}=\left(\frac{\partial U}{\partial T}\right)_{V,T\rightarrow 0}=9Nk_B\frac{\partial}{\partial T}\left[T\left(\frac{T}{T_D}\right)^3\int\limits_0^{+\infty}\frac{x^3}{e^x-1}\mathrm{d}x\right]\tag{9.6.69}$$

And this last integral is known to us. Indeed, we will prove it in details below using the Riemann zeta function during our proof of the Stefan-Boltzmann constant (see page 3726):

$$\begin{aligned} C_{V,T\to 0} &= 9Nk_B\frac{\partial}{\partial T}\left[T\left(\frac{T}{T_D}\right)^3\int\limits_0^{+\infty}\frac{x^3}{e^x-1}\mathrm{d}x\right] \\ &= 9Nk_B\frac{\pi^2}{15T_D^3}\frac{\partial}{\partial T}T^4 = 3Nk_B\frac{\pi^2}{5T_D^3}4T^3 = \frac{12Nk_B\pi^2}{5}\left(\frac{T}{T_D}\right)^3 \end{aligned} \tag{9.6.70}$$

Therefore:

$$\boxed{C_{V,T\to 0} = \frac{12Nk_B\pi^2}{5}\left(\frac{T}{T_D}\right)^3} \tag{9.6.71}$$

and experimental measurement correspond quite well to that behaviour! The Debye model correctly predicts the low-temperature dependence of the heat capacity of solids, which is proportional to T^3. Similarly to the Einstein photoelectron model, it recovers the Dulong–Petit law at high temperatures. Due to simplifying assumptions, its accuracy suffers at intermediate temperatures. But the Debye model is correct at low temperatures whereas the Einstein model is not.

Even though the Debye model is not completely correct, it gives a good approximation for the low temperature heat capacity of insulating, crystalline solids where other contributions (such as highly mobile conduction electrons) are negligible. For metals, the electron contribution to the heat is proportional to T, which at low temperatures dominates the Debye T^3 result for lattice vibrations. In this case, the Debye model can only be said to approximate the lattice contribution to the specific heat. The following table lists Debye temperatures for several pure elements and sapphire:

Aluminium	428 [K]	Copper	343 [K]	Nickel	450 [K]	Silver	215 [K]
Beryllium	1440 [K]	Germanium	374 [K]	Platinum	240 [K]	Tantalum	240 [K]
Cadmium	209 [K]	Gold	170 [K]	Rubidium	56 [K]	Tin (white)	200 [K]
Caesium	38 [K]	Iron	470 [K]	Sapphire	1047 [K]	Titanium	420 [K]
Carbon	2230 [K]	Lead	105 [K]	Selenium	90 [K]	Tungsten	400 [K]
Chromium	630 [K]	Manganese	410 [K]	Silicium	645 [K]	Zinc	327 [K]

Table 9.25: **Some Debye temperatures**

To visualize the difference between the models, one would naturally plot the two on the same set of axes, but this is not immediately possible as both the Einstein model and the Debye model provide a functional form for the heat capacity. As models, they require scales to relate them to their real-world counterparts. One can see that the scale of the Einstein model is given by Θ (i.e. T_E in what will follow) and T_D for the Debye model. Because the two methods approach the problem from different directions and different geometries, Einstein and Debye scales are not the same, that is to say:

$$\frac{h\nu}{k_B}\neq T_D \tag{9.6.72}$$

which means that plotting them on the same set of axes makes no sense. They are two models of the same thing, but of different scales. If one defines the Einstein temperature as we did above:

$$T_E \overset{\text{def}}{=} \frac{h\nu}{k_B} \tag{9.6.73}$$

then it's obvious that:

$$T_E \neq T_D \tag{9.6.74}$$

and, to relate the two, we must seek the ratio

$$\frac{T_E}{T_D} \tag{9.6.75}$$

The Einstein solid is composed of single-frequency quantum harmonic oscillators, $\varepsilon = \hbar\omega = h\nu$. That frequency, if it indeed existed, would be related to the speed of sound in the solid. If one imagines the propagation of sound as a sequence of atoms hitting one another, then it becomes obvious that the frequency of oscillation must correspond to the minimum wavelength sustainable by the atomic lattice, λ_{min}, where:

$$\nu = \frac{\lambda\nu}{\lambda} = \frac{\lambda\nu\sqrt[3]{N}}{2L} = \frac{\lambda\nu}{2}\sqrt[3]{\frac{N}{V}} \tag{9.6.76}$$

which makes the Einstein temperature:

$$T_E = \frac{h\nu}{k_B} = \frac{h\lambda\nu}{2k}\sqrt[3]{\frac{N}{V}} = \frac{h\lambda\nu}{2k_B}\left(\frac{N}{V}\right)^{1/3} = \frac{hv}{2k_B}\left(\frac{N}{V}\right)^{1/3} \tag{9.6.77}$$

and the sought ratio is therefore:

$$\frac{T_\mathrm{E}}{T_\mathrm{D}} = \frac{\dfrac{hv}{2k_B}\left(\dfrac{N}{V}\right)^{1/3}}{\dfrac{v\hbar}{k_B}\left(6\pi^2\dfrac{N}{V}\right)^{1/3}} = \sqrt[3]{\frac{\pi}{6}} = 0.805995977\ldots \tag{9.6.78}$$

Using the ratio, both models can be plotted on the same graph:

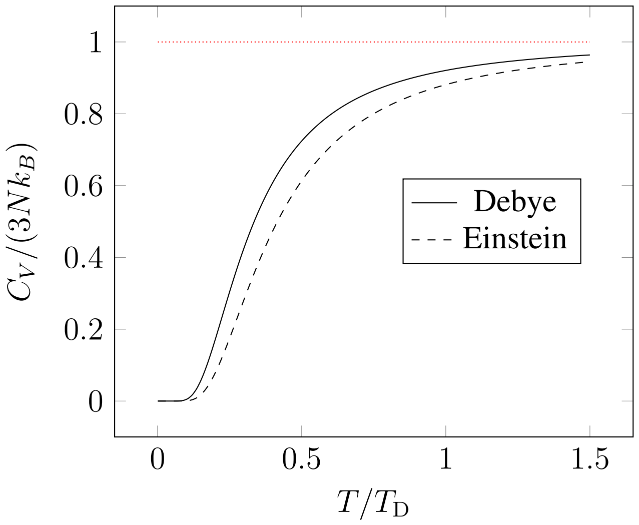


Figure 9.162: **Debye vs Einstein model**

This ratio is the cube root of the ratio of the volume of one octant of a 3-dimensional sphere to the volume of the cube that contains it, which is just the correction factor used by Debye when approximating the energy integral above. Alternatively, the ratio of the two temperatures can be seen to be the ratio of Einstein's single frequency at which all oscillators oscillate and Debye's maximum frequency. Einstein's single frequency can then be seen to be a mean of the frequencies available to the Debye model.

In thermodynamics, we must also always specify what is considered transformation as the ΔQ depends on the nature of the transformation. We distinguish in particular the heat capacity at constant volume (isochoric) and denoted:

$$C_V=\left(\frac{\delta Q}{\partial T}\right)_V \tag{9.6.79}$$

and heat capacity at constant pressure (isobaric) often used in chemistry and denoted:

$$C_P=\left(\frac{\delta Q}{\partial T}\right)_P \tag{9.6.80}$$

The difference between the specific heat at constant pressure C_P and constant volume specific heat C_V is obviously due to the work that must be provided to expand the body in the presence of external pressure. Both term also highlight a fundamental difference in the experimental setups measurement, that is to say physically different situations. In practice, to use the concept of heat capacity, we need to identify for our experience what C we must use depending on how the installation / machine will have been build.

Then it is quite intuitive that the specific heat at constant pressure is always strictly greater than the specific heat at constant volume $C_P>C_V$. We will also prove further below that we specifically for a perfect gas:

$$C_P-C_V=nR=Nk \tag{9.6.81}$$

If instead of considering the whole system, we report measurements to a unit of mass (which is much more useful in practice!), Then we define the "specific heat capacity" at constant volume (isochoric):

$$c_V=\frac{1}{M}\left(\frac{\delta Q}{\partial T}\right)_V \tag{9.6.82}$$

and the specific heat capacity at constant pressure (isobaric):

$$c_P=\frac{1}{M}\left(\frac{\delta Q}{\partial T}\right)_P \tag{9.6.83}$$

If instead of considering the whole system, we report the measurement to one mole of the considered medium (which is much more useful in chemistry!), We define the "molar heat capacity" at constant volume (isochoric):

$$c_V^M=\frac{1}{n}\left(\frac{\delta Q}{\partial T}\right)_V \tag{9.6.84}$$

and the molar heat capacity at constant pressure (isobaric):

$$c_P^M = \frac{1}{n}\left(\frac{\delta Q}{\partial T}\right)_P \tag{9.6.85}$$

where n is the number of moles of the system.

Here are some typical value for some gases at 293 [K] (i.e. 20 [°C]):

Gas	**Formula**	c_P^M **[J·mol⁻¹·K⁻¹]**
Air		29.1
Ammonia	NH_3	36.8
Argon	Ar	20.9
Nitrogen	N_2	29.1
Chlorine	Cl_2	33.8
Carbon dioxide	CO_2	36.8
Sulphur dioxid	SO_2	41.0
Ethine (acetylene)	C_2H_2	43.8
Hydrogen sulfide	H_2S	35.7
Helium	He	20.9
Hydrogen	H_2	28.9
Methan	CH_4	35.6
Neon	Ne	20.8
Oxygen	O_2	29.3

Table 9.26: **Some gases molar heat capacities**

So we have of course by extension, by going to the discrete case, the relation that is very useful in practice:

$$\boxed{\Delta Q = mc\Delta T} \tag{9.6.86}$$

where c is the specific polytropic heat capacity (and not the speed of light...!).

This is a very important relation in practice and whose everyday applications are huge because it allows in idealized cases to calculate the heat used or provided when a certain amount of material undergoes a change in temperature or calculate the final temperature of two bodies in contact.

Think about it when you take your shower or you heat a pot of water. You know then ΔT and the amount of water M used for your shower/bath or pan. This will give you the total energy used ΔQ and knowing the cost of energy in your city you quickly deduce the cost of a shower, a bath or a pot of boiling water!

Another nice application is to calculate the temperature increase (in an ideal case) of an object dropped from a certain height under the assumption that all its potential energy is converted into heat after touching the ground (the only problem being to find the value of the massic polytropic heat capacity in the tables...):

$$\Delta Q = mgh = mc_j\Delta T \tag{9.6.87}$$

where c_j is the massic polytropic energy capacity (it is simply c but given in joules per kilogram per kelvin rather than calories per kilogram per kelvin). So after rearrangement we have:

$$\boxed{\Delta T=\frac{mgh}{mc_j}} \tag{9.6.88}$$

The major case of the final equilibrium temperature of two bodies in contact is easy to obtain (if the system is completely isolated from the outside). Indeed, as the heat lost by one of the two bodies will be won by the other, we have by choosing properly ΔT for each of the two bodies so that ΔQ is the same sign:

$$c_1m_1(T_1-T_f)=c_2m_2(T_f-T_2) \tag{9.6.89}$$

Or after an elementary rearrangement:

$$\boxed{T_f=\frac{c_1m_1T_1+c_2m_2T_2}{c_1m_1+c_2m_2}} \tag{9.6.90}$$

✍**Example:**

A 0.3 [kg] steel bar has an initial temperature of $1,200$ [K] is immersed in a tank of water at 300 [K] (the tank is closed, isolated from the outside and full) containing 9 [kg] of water. Taking the specific polytropic heat capacity of water that is $4.2\,[\text{KJ}\cdot\text{kg}^{-1}\cdot\text{K}^{-1}]$ and that for steel $0.42\,[\text{KJ}\cdot\text{kg}^{-1}\cdot\text{K}^{-1}]$, the final equilibrium temperature (of the bar and water) is then:

$$T_f=\frac{4.2\cdot 10^3\cdot 9\cdot 300+0.42\cdot 10^3\cdot 0.3\cdot 1200}{4.2\cdot 10^3\cdot 9+0.42\cdot 10^3\cdot 0.3}=\frac{4.2\cdot 9\cdot 300+0.42\cdot 0.3\cdot 1200}{4.2\cdot 9+0.42\cdot 0.3}\cong 303\,[\text{K}] \tag{9.6.91}$$

Let us now consider the case where we have a system without exchange of mechanical work but only of heat by the following relation:

$$\Delta U=C\Delta T=C(T_2-T_1) \tag{9.6.92}$$

However, power is energy divided by time (see section Classical Mechanics page 3379). It comes therefore (take care not confuse the notation of power with that of pressure!):

$$P=\frac{C(T_2-T_1)}{t} \tag{9.6.93}$$

and if the heat capacity is given in the massic form at constant pressure:

$$P=\frac{mc_p(T_2-T_1)}{t}=\frac{m}{t}c_p(T_2-T_1) \tag{9.6.94}$$

Which already allows to calculate the power to be supplied to heat a quantity of material at a given temperature in a given time without state change otherwise we must we must take into account the latent heat.

The previous fraction can also be expressed as the flow rate of a fluid $\dot{V}$ $[\mathrm{m}^3\cdot\mathrm{s}^{-1}]$ multiplied by the fluid density:

$$\boxed{P=\dot{V}\rho c_p(T_2-T_1)} \tag{9.6.95}$$

which allows to know the power supplied by a water flow into a radiator at a certain temperature and leaving it with the same flow but with a different temperature. Those working with surfaces prefer to write this relation as follows:

$$\boxed{P=\dot{Q}=h_c S\Delta T} \tag{9.6.96}$$

which is often named "Newton's law of heat transfer by convection" or more simply "Newton's law of cooling" and where h_c is the "convection coefficient".

We will also find sometimes the prior-previous relation in the literature as follows:

$$E=mc_p(T_2-T_1)=\rho V c_p(T_2-T_1) \tag{9.6.97}$$

after having obviously multiplied both sides of the equality by the time.

✍**Example:**

A microprocessor surface of $25\,[\mathrm{mm}^2]$ is powered by a power source of 0.224 [W]. The ambient atmosphere is at 293 [K] $(20\,[^\circ\mathrm{C}])$ and the convection coefficient is given at the standard conditions of temperature and pressure as being equal to $h_c=150\,[\mathrm{W}\cdot\mathrm{m}^{-2}]$.

Considering that all the heat transfer takes place between the air and the microprocessor, the temperature of the latter can then be calculated as follows:

$$P=h_c S\Delta T=h_C S(T_2-T_1) \tag{9.6.98}$$

Therefore:

$$T_2=\frac{P}{h_C S}+T_1\cong\frac{0.225}{150\cdot 25\cdot 10^{-6}}+293=353\,[\mathrm{K}]=80\,[^\circ\mathrm{C}] \tag{9.6.99}$$

As we have know now that we have:

$$Q=m\cdot c\cdot(T_c-T_e) \tag{9.6.100}$$

where T_c is the core temperature of an object and T_e the temperature of its environment.

As we already have (we want to focus on the decrease of temperature hence the minus sign):

$$\frac{\mathrm{d}Q}{\mathrm{d}t}=-h_c\cdot S\cdot(T_c-T_e) \tag{9.6.101}$$

Injecting the first into to second one we get:

$$\frac{\mathrm{d}\left(m\cdot c\cdot(T_c-T_e)\right)}{\mathrm{d}t}=-h_c\cdot S\cdot(T_c-T_e) \tag{9.6.102}$$

Therefore after simplification and assuming T_e is constant (hence is derivative is zero):

$$\frac{\mathrm{d}T_c}{\mathrm{d}t}=-\frac{h_c S}{mc}(T_c-T_e) \tag{9.6.103}$$

Remarks

The relation above is often denoted in undergraduate textbooks:

$$T' \propto k\Delta T$$

And explained the by the analogy that a very hot (compared to its environment) drink does not stay hot for long. A tepid (compared to its environment) drink stays tepid for longer.

Rearranging:

$$\frac{1}{(T_c - T_e)}\mathrm{d}T_c = -\frac{h_c S}{mc}\mathrm{d}t \tag{9.6.104}$$

Taking primitives on both sides we get:

$$\ln(T_c - T_e) = -\frac{h_c S}{mc}t + c^{te} \tag{9.6.105}$$

Taking the exponential on both sides and rearranging we get:

$$T_c = e^{-\frac{h_c S}{mc}t + c^{te}} + T_e = e^{c^{te}}e^{-\frac{h_c S}{mc}t} + T_e := T_0 e^{-\frac{h_c S}{mc}t} + T_e \tag{9.6.106}$$

So finally we get the famous exponential time decrease of temperature:

$$\boxed{T_c = T_e + T_0 e^{-\frac{h_c S}{mc}t}} \tag{9.6.107}$$

In some transformations, we have a contribution or a heat withdrawal while the system temperature remains constant (typically in phase change transformations of a body). This comes from another phenomenon:

Definition (#9.6.464): The "enthalpy change of state" or "latent heat", denoted by L and is given in Joules per kilogram, corresponds to the quantity of heat required for a unit material quantity or a unit of mass of a body to changes its state. This is a very important value in practice to perform calculations.

So we of course by extension we have the very useful relation in practice:

$$\boxed{Q = mL} \tag{9.6.108}$$

which thus gives the quantity of heat required to change phase a certain quantity of mass m (the sign of Q is then negative then by convention). Conversely, if the amount of material returns to a lower energy level, the amount of latent heat will be given to the environment and not absorbed and the sign of Q will be positive!

Water has the following values for latent heat:

- The latent heat of vaporization of water is 540 calories per gram (changes 1 [g] water at 100 [°C] into 1 [g] steam at 100 [°C])
- The latent heat of fusion of water is 80 calories per gram (changes 1 [g] of ice at 0 [°C] into 1 [g] of water at 0 [°C])

The process also works in reverse. Above, heat was added to the substance, changing its phase from solid to liquid or liquid to gas. What if we cool the substance instead?

In this case, heat is released as the substance changes phase from gas to liquid and liquid to solid. It makes sense that the amount of heat required to change a gram of substance from liquid to gas is the same amount that is released when the substance changes from a gas back into a liquid. Because the values are the same, and the only difference is the direction of the process, we only use a single term for each phase change:

- "Latent heat of vaporization" or "enthalpy of vaporization": phase changes from liquid to gas OR gas to liquid;
- "Latent heat of fusion" or "enthalpy of fusion": phase changes from solid to liquid OR liquid to solid

The following picture shows what happens to 1 gram (1 ml) of water. The orange arrow on the top represents adding heat, from left to right, and the blue arrows represent releasing heat from right to left on the bottom:

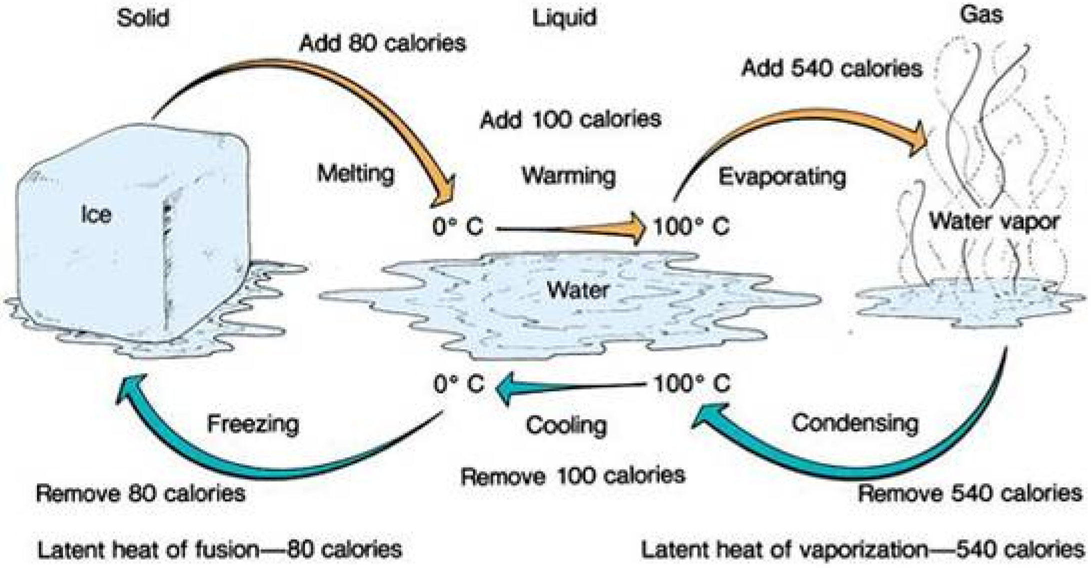

Solid
Add 80 calories
Liquid
Gas
Add 540 calories
Add 100 calories
Melting
Warming
Evaporating
Ice
0° C
100° C
Water vapor
Water
0° C
100° C
Freezing
Cooling
Condensing
Remove 80 calories
Remove 100 calories
Remove 540 calories
Latent heat of fusion—80 calories
Latent heat of vaporization—540 calories

Figure 9.163: **Water latent heat process (source: ?)**

We will prove later that at constant pressure the latter relation is then written more explicitly as follows:

$$\Delta H=\Delta Q=\int_{T_1}^{T_2} c_P \mathrm{d}T=\int_{T_1}^{T_2} C_P \mathrm{d}T \cong C_P \Delta T \qquad (9.6.109)$$

The enthalpy exchanged during the change of state results from the modification (rupture or establishment) of inter-atomic or intermolecular bonds. As we know are three main physical states for any pure substance: solid, liquid and gas. The links are stronger in solid form in the liquid state and these links are virtually absent in the gaseous state. There are, as we already know a fourth state obtained at very high temperature where the material is in the form of a plasma of ions and electrons (see section Continuum Mechanics page 3945).

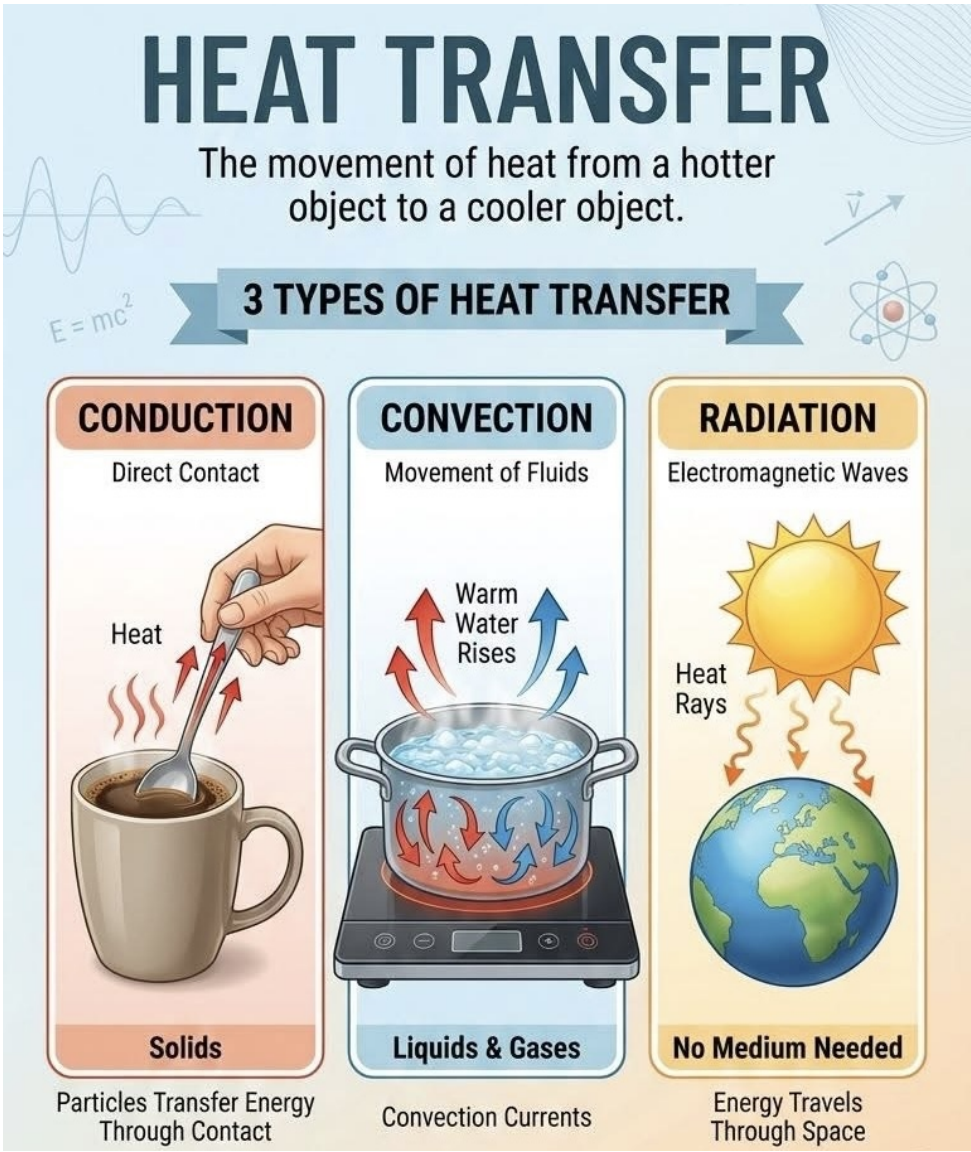
We differentiate three main modes of heat transfer:

- "convection": fluid coming alternately from a hot body and a cold body (which is typical of the movements of the Earth's atmosphere).
- "radiation": a body at a given temperature emits electromagnetic radiation that can be absorbed by another body (the electromagnetic radiation being not necessarily in the visible spectrum part!).
- "conduction": energy transfer occurs from hot areas to cold areas by collision of the excited particles (hot zone) with neighbouring particles less excited and so on....

  That's why in space there is almost no heat transfer by conduction and by convection!

  To understand how density matters for heat transfer by conduction, consider that when baking bread in the oven, both air and metal plates are of the same temperature. However, while you can still put your hand into that oven (to take the bread out), you would be foolish trying to touch the metal plates...

  Or the analogy of a sauna: You often have around 100 degree Celsius in a sauna, and believe me - it is very nice (if you are into that kind of thing). But 100 degree water is boiling and lethal!


HEAT TRANSFER
The movement of heat from a hotter object to a cooler object.
3 TYPES OF HEAT TRANSFER
E = mc²
CONDUCTION
Direct Contact
Heat
Solids
Particles Transfer Energy Through Contact
CONVECTION
Movement of Fluids
Warm Water Rises
Liquids & Gases
Convection Currents
RADIATION
Electromagnetic Waves
Heat Rays
No Medium Needed
Energy Travels Through Space

9.6.7.1 Newton's Law of Cooling

We have introduce earlier above the "Newton's law of cooling" that explains the rate at which an object/entity changes its temperature when it is exposed to radiation[24]. This change is almost proportional to the difference between the object's temperature and its surroundings temperature, given that this difference is quite small. Newton's law of cooling is included in virtually every introductory textbook on differential equations in the USA (but almost never in Europe).

We know that the formula of this law can be used to calculate how quickly a substance, at a particular temperature, would cool in a particular environment. Newton's law of cooling models the average temperature in an object by a simple ordinary differential equation as we will see it, while the heat equation is a partial differential equation that models the temperature as a function of both space and time. A series solution of the heat equation, in the case of a spherical body, shows that Newton's law gives an accurate approximation to the average temperature if the body is not too large and it conducts heat much faster than it gains heat from the surroundings (for the mathematical proof on a few dozens of pages see [207]). Finite element simulation confirms and extends the analysis.

---

[24] Newton did not originally state his law in the above form. Rather, using today's terms, Newton noted after some mathematical manipulation that the rate of temperature change of a body is proportional to the difference in temperatures between the body and its surroundings.

It also allows us to determine how the rate of cooling of an object depends not only on the temperature difference between the substance and the surrounding but also on the constant of the substance.

So let us introduce this law using an easier way than earlier above more adapted to college level students.

According to Newton's assumption we put:

$$\frac{\mathrm{d}T}{\mathrm{d}t} \propto (T-T_s) \Rightarrow \frac{\mathrm{d}T}{\mathrm{d}t} = -k\,(T-T_s) \tag{9.6.110}$$

where T is the temperature of the object, T_s the temperature of it's surroundings.

The negative sing on the right hand side implies that the rate of temperature change decreases.

Let T_f is the temperature at time t and T_0 is the temperature at the beginning.

Applying Integration on both sides:

$$\begin{aligned} \int_{T_f}^{T_0} \frac{\mathrm{d}T}{(T-T_s)} &= \int -k\mathrm{d}t \\ &\Rightarrow \ln\left(\frac{T_f-T_s}{T_0-T_s}\right) = -kt \\ &\Rightarrow \frac{T_f-T_s}{T_0-T_s} = e^{-kt} \\ &\Rightarrow T_f - T_s = (T_0-T_s)\,e^{-kt} \end{aligned} \tag{9.6.111}$$

Therefore:

$$\boxed{T_f = T_s + (T_0-T_s)\,e^{-kt}} \tag{9.6.112}$$

If we derive with rigour this relation from the heat equation (details available in [207]), then we see that the limitations of this relation are:

- The difference in temperatures between the surroundings and the object must be small
- The loss of heat should be via radiation only
- The temperature of the surroundings must remain constant during the cooling of the object

9.6.8 Internal Energy

In the section of Classical Mechanics, we have proved (second König's theorem) that the total energy of a body with respect to an external inertial frame R' to the center of mass was given by the sum of kinetic and potential energy compared to that same inertial frame, added to the kinetic energy and potential energy of that body relatively the reference frame of the center of mass R (barycentric repository).

$$E_{\text{tot}}=\left(E_{c,G/R}+E_{p,G/R}\right)+\left(E_{c,G/R'}+E_{p,G/R'}\right)=\left(E_{c,G/R}+E_{p,G/R}\right)+U \tag{9.6.113}$$

where U is the energy quantity or "internal energy" (also sometimes name "proper energy" or of the object itself relatively to its own reference frame.

In a more simplified form, the previous relation is traditionally written in thermodynamics for a closed isolated system:

$$E_{\text{tot}}=E_c+E_p+U \tag{9.6.114}$$

The first principle of thermodynamics is then written:

$$\boxed{Q+W=E_{\text{tot}}=E_c+E_p+U} \tag{9.6.115}$$

Therefore the total energy is the sum of the macroscopic mechanical energy (kinetic, potential or rotational) and microscopic (internal energy).

Generally in thermodynamic studied systems are generally at rest ($v_G=0$) with respect to the experimenter and therefore $\Delta E_c=0$. Similarly, we have generally a constant and isotropic potential field in the experimental room or area which implies equation $\Delta E_p=0$.

As a result of these simplifications the law of conservation of energy[25], or first principle of thermodynamics, is reduced as we have already mentioned to the internal energy as:

$$\boxed{Q+W=U} \tag{9.6.116}$$

or more rigorously:

$$\Delta Q+\Delta W=\Delta U \tag{9.6.117}$$

---

[25] It is strongly advice that the reader remembers the situation where this principle doesn't apply as seen during our study of Noether's theorem!

✍**Example:**

A piston contains 0.9 [kg] of air at a temperature of 300 [K] and a pressure of 1 [bar]. This air is compressed to a pressure of 6 [bar] to a temperature of 470 [K]. During the compression phase 20 [kJ] heat is exchanged between the external environment (heat conduction).

Using the fact that the internal massic energy of air is at 300 [K] given in the thermodynamic tables is equal to $u_1 = 214.07\,[\text{kJ}\cdot\text{kg}^{-1}]$ at 470 [K] as being equal to $u_2 = 337.32\,[\text{kJ}\cdot\text{kg}^{-1}]$, then we have the work done for the compression knowing that:

$$\Delta Q + \Delta W = \Delta U \tag{9.6.118}$$

that was equal to:

$$\Delta W = \Delta U - \Delta Q = m(u_2 - u_1) - \Delta Q = 0.9(337.32 - 214.07) - (-20) \cong 131\,[\text{kJ}] \tag{9.6.119}$$

Remark

The internal energy U is often given to a given additive constant, that is why some it is sometimes rightly named the "internal over-energy". As we have proved it in the section of Classical Mechanics, the total energy of a system is the sum of all elementary energies of it. Thus, the internal energy is an extensive quantity.

Some may ask what is exactly the internal massic energy?! It's quite easy in fact. For example when you touch something hot like a stove, you might think that there is heat in the stove but heat is only the energy being transfer. The energy in the stove is named "Internal Energy". It is made up of the energy of the fast moving atoms and the energy stored in the bonds that hold the atoms of the stove together. The hotter something is the more internal energy the object has. The diagram on the right shows an image of atoms held together with bonds, this is what a molecule of a solid or liquid would look like. When heat is transferred to the atoms they move very quickly, if enough heat is transferred then the bonds will break and the molecule will change state (melt or boil).

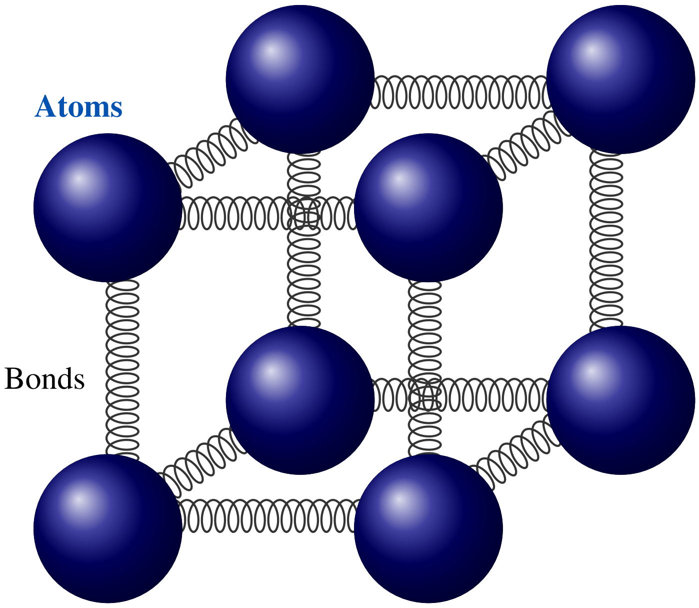


Figure 9.164: **Vibration bonds source of the internal energy in solid and liquids (author: Izaak Neutelings)**

This is why the internal energy is defined as: the energy associated with the random, disordered motion of molecules.

Typically the internal energy can be related to the following three internal energies with the corresponding energies proportions for the three common states of matter:

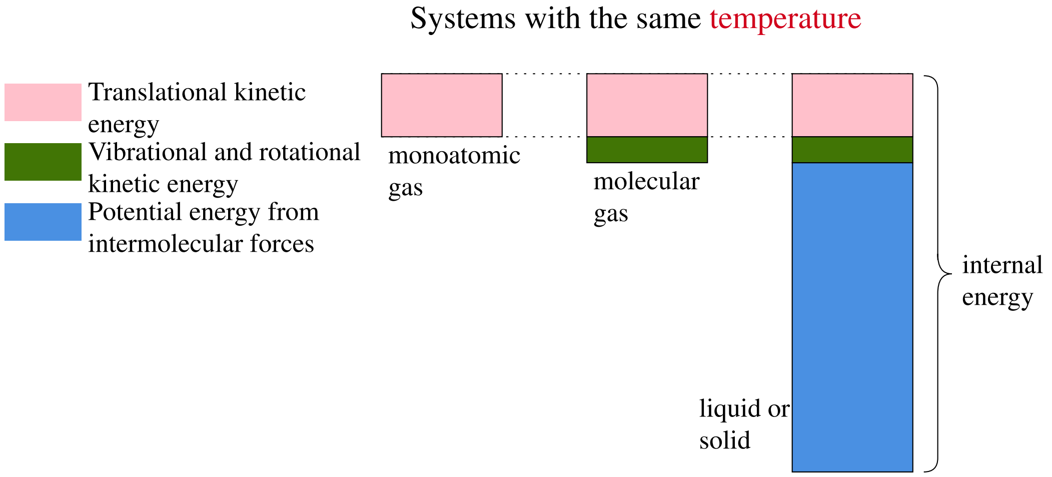


Figure 9.165: **Proportion of internal energy type by common state of matter**

Let us wow consider a system described by the thermodynamic state variables:

$$x_1, x_2, \ldots, x_n \tag{9.6.120}$$

In an elementary transformation $\mathrm{d}x_1, \mathrm{d}x_2, \ldots, \mathrm{d}x_n$ the elementary work $\mathrm{d}W$ (if we interested only to that form of energy) will be in the form (see section Differential and Integral Calculus page 1466) of an exact total differential:

$$\mathrm{d}W = A_1\mathrm{d}x_1 + A_2\mathrm{d}x_2 + \ldots + A_n\mathrm{d}x_n \tag{9.6.121}$$

if we have the following relation, proved in the section of Differential Calculus and Integral, which is satisfied (which are the coefficients of the $\mathrm{d}x_i$ above):

$$\frac{\partial A_i}{\partial x_j}=\frac{\partial A_j}{x_i} \tag{9.6.122}$$

Otherwise, if this relation is not satisfied, which would imply:

$$\frac{\partial A_i}{\partial x_j}\neq\frac{\partial A_j}{x_i} \tag{9.6.123}$$

whereas in the latter case, we have an inexact total differential (the differential depends of the path!) which would bring us to write:

$$\delta A=A_1\mathrm{d}x_1+A_2\mathrm{d}x_2+\ldots+A_n\mathrm{d}x_n \tag{9.6.124}$$

Since the mechanical work W is in one way or another a force over a distance, we have already proved in the section of Classical Mechanics that it depends on the path travelled and not only the starting and final point.

In the same transformation, the quantity of heat $\mathrm{d}Q$ can be put in a similar form:

$$\mathrm{d}Q=B_1\mathrm{d}x_1+B_2\mathrm{d}x_2+\ldots+B_n\mathrm{d}x_n \tag{9.6.125}$$

if again the following relation holds for the state variables:

$$\frac{\partial B_i}{\partial x_j}=\frac{\partial B_j}{x_i} \tag{9.6.126}$$

or otherwise:

$$\delta Q=B_1\mathrm{d}x_1+B_2\mathrm{d}x_2+\ldots+B_n\mathrm{d}x_n \tag{9.6.127}$$

The characteristic quantities of a transformation of the system but whose value depends not only on the initial and final states, but also t the path followed are named "transfer quantities".

Let us recall that knowing what has already been studied in the section of Differential and Integral Calculus then we have for a thermodynamic cycle in an isolated system:

$$\boxed{\Delta U=\oint\mathrm{d}U=0\qquad\Delta Q=\oint\delta Q=0\qquad\Delta W=\oint\delta W\neq 0} \tag{9.6.128}$$

and that is the fact that the internal energy is an exact total differential which logically requires that the heat Q is an inaccurate total differential by the property of linearity of integral (otherwise it would not cancel by summing the variations of the W a closed cycle). Indeed, according to the first law of thermodynamics, if the energy is conservative for a closed isolated system then we have for the internal energy (in discrete form):

$$\Delta U=0 \tag{9.6.129}$$

This means that regardless of the transformation, the variations of work and heat are given within this conservative system by (in discrete form):

$$\boxed{\Delta Q+\Delta W=0} \tag{9.6.130}$$

or more technically:

$$\oint \mathrm{d}U=\oint(\delta U+\delta W)=\underbrace{\oint \delta Q}_{\neq 0}+\underbrace{\oint \delta Q}_{\neq 0}=0\Rightarrow\oint \delta Q=-\oint \delta W \qquad (9.6.131)$$

Therefore regardless of the ways in which transformations occur inside the system, the final and initial state are identical in terms of energy in an isolated system. This assumes that the thermodynamic transformations are independent of the manner in which the phenomena take place inside the system. The term ΔU is then an exact total differential (see section Differential and Integral Calculus page 1466) that we will write $\mathrm{d}U$.

9.6.8.1 Work (energy) of Mechanical Forces

Mechanical work is the amount of energy transferred by a force. Like energy, it is a scalar quantity, with SI units of joules. Heat conduction is not considered to be a form of work, since there is no macroscopically measurable force, only microscopic forces occurring in atomic collisions. In the 11830s (holocene calendar), French mathematician Gaspard-Gustave Coriolis coined the term work for the product of force and distance.

Let us recall now that we have seen in the section of Classical Mechanics that, by definition, the work W is given by a force over a distance (whatever the origin of this force: mechanical, radiant, nuclear, etc.). It follows therefore a quite important relation in thermodynamics:

$$\boxed{\mathrm{d}W=\vec{F}\circ\mathrm{d}\vec{l}=PS\mathrm{d}x=P\mathrm{d}V} \qquad (9.6.132)$$

which expresses, at constant pressure (isobaric), the variation of the energy due to the work of the external pressure forces on a system (typically a thermodynamic gas..) whose volume has varied (without being restricted by rigid boundary!).

Obviously, if the volume variation is zero or there is no pressure... the change in energy due to work pressure forces will be zero at equilibrium...

At equilibrium, the pressure P is taken as the considered system (internal pressure) or that of the surrounding atmosphere as anyway the pressure will be equal if the system is in contact with the atmosphere (otherwise there would be no equilibrium at the opposite of what we have just assumed at the beginning of this paragraph...).

Similarly, the change in volume is taken as being either that of the considered system (proper volume) or the variation of the surrounding atmosphere since in any case the volume change will be the same for both!

Moreover, we see that the path is appears explicitly in the previous expression of the work and therefore that it is a quantity that depends on the path travelled (see section Differential and Integral Calculus page 1583). This implies that we must write δW instead of equation $\mathrm{d}W$. There are of course many other ways of expressing the work, but by it's own definition, it will always be an inexact differential.

This result has a direct implication on the expression of heat change for which the virial theorem (see section Continuum Mechanics page 3897): that it is given by the thermal agitation.

It should be recalled that this agitation is given by the average kinetic energy and kinetic energy exist by applying a force over a distance for each particle. Thus heat is also an inexact differential δQ!

Finally, we have in the case of a fluid (liquid or gas) having an internal polytropic energy change:

$$\boxed{\mathrm{d}U=\delta Q+\delta W} \tag{9.6.133}$$

During its evolution, the volume V of the system may vary. If we consider an infinitesimal change during which the volume varies of $\mathrm{d}V$ and if we denote by P_e the external pressure on the system, we will write in a more general context (the negative sign in front of the pressure is a convention!):

$$\delta W=-P_e\mathrm{d}V+\mathrm{d}T \tag{9.6.134}$$

where $-P_e\mathrm{d}V$ is the work named "outside delivery" (if $\mathrm{d}V>0$, the increase of the volume of the system expresses the supply of work to the external environment, which is what the sign $-$ explains indicating that the internal energy of the system decreases) and where $\mathrm{d}T$ (not to be confused with temperature variation!!!) is another form of energy (if the system is reactive to the chemical level for example, it can provide a potential chemical energy too).

Except in special systems (batteries, accumulators, ...) the only energy exchanged is the work of the pressure forces so that we keep only:

$$\delta W=-P_e\mathrm{d}V \tag{9.6.135}$$

and then we can write $\mathrm{d}W$ at leisure instead of using the inexact differential δW since this work is dependent only of on state variable (the notation $\mathrm{d}W$ being often used in physics in this case)!

In addition, most systems are studied by hypothesis, at equilibrium, at the same internal pressure as the external pressure (ambient atmosphere). This allows us in this case to write:

$$P_e=P \tag{9.6.136}$$

where P is the pressure inside and outside of the system (the system is then said "isobaric" as we already know it and rarely "monobaric").

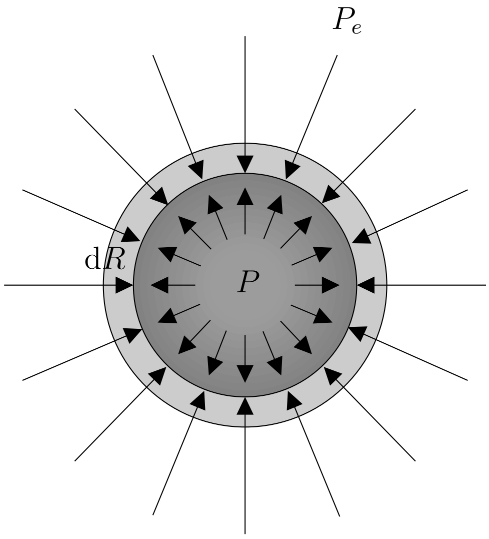


Figure 9.166: **Example of a sphere at pressure equilibrium with the environment**

Therefore, we have in this particular case:

$$\delta W = -P\mathrm{d}V \qquad \Delta W = -\int_1^2 P\mathrm{d}V \tag{9.6.137}$$

The heat quantity Q_V into play in an isochoric transformation (at constant volume) then reduces obviously the change in internal energy such that (caution with the traditional notation Q_V that can makes us forget that we are dealing with a heat variation !):

$$\delta Q := Q_V = U_2 - U_1 \tag{9.6.138}$$

If we consider a system evolving from a state 1 to a state 2 in a surrounding isobaric atmosphere (constant pressure) and therefore that its internal pressure is in equilibrium with the atmosphere, then we have the "mechanical forces at constant pressure" relation:

$$\boxed{\Delta W = -\int_1^2 P\mathrm{d}V = -P(V_2 - V_1)} \tag{9.6.139}$$

Or if the system evolves from a state 1 to a state 2 in an isotherm environment, then we have the "mechanical forces at constant temperature" relation for a perfect gas:

$$\boxed{\Delta W = -\int_1^2 P\mathrm{d}V = -\int_1^2 \left(\frac{nRT}{V}\right)\mathrm{d}V = -nRT\int_1^2 \frac{\mathrm{d}V}{V} = -nRT\ln\left(\frac{V_2}{V_1}\right)} \tag{9.6.140}$$

If we want to produce a reasonable amount of work it seems a priori that the volume will have to reach unrealistic dimensions. So this is why we build thermal machines with a cyclic return.

OK... with what we have seen so far we can already make the following summary:

Process	What is constant	The first law predicts
Isothermal	$T =$ constant	$\Delta T = 0$ makes $\Delta U = 0$, so $\Delta Q = \Delta W$
Isobaric	$P =$ constant	$\Delta P = 0$ makes $\Delta Q = \Delta U + W = \Delta U + P\Delta V$
Isochoric	$V =$ constant	$\Delta V = 0$ makes $\Delta W = 0$ so $\Delta Q = \Delta U$
Adiabatic	Q =constant	$\Delta Q = 0$ makes $\Delta U = -\Delta W$

Table 9.27: **Summary of basic thermodynamics processes**

9.6.8.2 Enthalpy

Let us consider the isobaric case very common in chemistry labs since the beakers are open to the ambient atmosphere. If we consider such a situation, then we have for the variation of internal energy:

$$U_2 - U_1 = \Delta Q - P(V_2 - V_1) \rightarrow \Delta Q = U_2 - U_1 + P(V_2 - V_1) = (U_2 + PV_2) - (U_1 + PV_1) \tag{9.6.141}$$

so where the pressure P and the volume V are the internal state variables to the studied system! It is important to notice again that if the change in heat ΔQ is zero and that the internal volume increases the internal energy U variation is then negative, which will be interpreted by the experimenter as an energy taken to the external system and so it is common to speak of "loss" or "endothermic reaction"!

The heat quantity Q_P into play in an isobaric transformation (the index of Q_P is indicated in this purpose... can makes us forget that we are dealing with a variation of heat) is equal to the variation of two terms in identical form:

$$\Delta Q := Q_P = (U_2 + PV_2) - (U_1 + PV_1) = H_2 - H_1 = \Delta H \tag{9.6.142}$$

where we therefore define a new convenient function of state, named "enthalpy H" (extensive quantity) in an isobaric transformation as given by:

$$\boxed{H = U + PV = U + nRT} \tag{9.6.143}$$

where n is the number of moles internal to the ideal gas undergoing the volume change! We see further that if the surrounding pressure is zero, the enthalpy is equal to the internal energy. The fact that there are external pressures force (or internal) adds an energy to the system which therefore defines the concept of enthalpy!

The prior previous relation expresses the fact that when a system evolves at constant pressure, the heat received (or exchanged by the system with the external environment) is equal to the enthalpy change.

It also comes with this new definition a new possible writing of to the heat capacity at constant pressure frequently used in chemistry:

$$C_P = \left(\frac{\delta Q}{\partial T}\right)_P = \frac{\partial H}{T} \tag{9.6.144}$$

So get the very important relation in practice that we had specified earlier above for the latent heat (enthalpy of state change):

$$\Delta H = m\int_{T_1}^{T_2} c_P \mathrm{d}T = \int_{T_1}^{T_2} C_P \mathrm{d}T \cong C_P \Delta T \tag{9.6.145}$$

where we must not forget that strictly speaking the fixed pressure heat capacity is not constant, it is quite often considered in practice as constant for a given temperature range.

More explicitly, we have also for the prior previous relation using the ideal gas law:

$$H = U + PV = \frac{\mathrm{df}}{2}nRT + nRT = \frac{\mathrm{df}+2}{2}nRT \tag{9.6.146}$$

where for recall (see section Continuum Mechanics page 3897) the number of degrees of freedom df for a perfect monoatomic gas is equal to 3. The last term of this relation is obviously not valid for fluids!

Sometimes we obviously also write the definition of enthalpy in the form of a variation such that:

$$\Delta H = \Delta U + P\Delta V = nR\Delta T \tag{9.6.147}$$

in the particular case of application to the field of validity of the ideal gas (see section Continuum Mechanics page 3903) regarding to the last term.

It is important to remember that the term PV in the previous relations represents the work of the pressure forces from the surrounding atmosphere on the system or by equivalence, respectively of the system on the surrounding atmosphere that surrounds it. But the variable n is always the number of moles of an ideal gas of the system studied and not the surrounding atmosphere.

The enthalpy is a concept used extensively in thermal chemistry (see section Thermochemistry page 6249) and we will use it constantly in its study.

In practice it is difficult (if not impossible) to know the internal energy. We then calculate rather the relation relatively to one mole (n being then equal to 1) of a perfect gas:

$$H-U=RT=PV \tag{9.6.148}$$

That is a strictly positive value giving the energy (or rather the surplus of energy) available in the ideal gas because of the surrounding pressure (there is only to put the pressure as being equal to zero to see it!).

✍**Example:**

Let us take a molar unit volume of ideal gas at standard temperature and pressure (whatever that perfect gas, the molar volume under these conditions will always 24 litres according to the ideal gas law!).

The work of the pressure forces that were necessary to bring this ideal gas to the above conditions is then:

$$H-U=PV\cong 1.015\cdot 10^5\frac{24}{1000}=2,436\,[\text{J}] \tag{9.6.149}$$

or:

$$H-U=RT\cong 293.15\cdot 8.31=2,436\,[\text{J}] \tag{9.6.150}$$

which is energy in the form of mechanical forces of pressure that we could recover from a molar volume of air in comparison to the vacuum state.

The internal energy into the form of heat and mechanical work that we could recover in theory (the problem is how to...) of the molar volume is given for a monoatomic gas (see further below for the detailed proof), by :

$$H=Q=c_P^M RT=U+PV=\frac{\text{df}+2}{2}RT=\frac{5}{2}RT\cong 6\,[\text{kJ}] \tag{9.6.151}$$

and we could also calculate the internal energy of the electrons binding, and that of the nuclei of atoms... and the equivalence mass and energy, but we would go out of common industrial case...

For water, we can only use the first relation, because the second one is valid only for ideal gases, the molar volume is $1,000$ times smaller, we can only extract of it an energy $1,000$ times smaller (about 2 [J]). This is why in the case of fluids, we consider that there is no difference between the enthalpy and internal energy.

Let us now see other theoretical implications of some previously seen items that will be very useful both in acoustic (see section Music Mathematics page 8195) or in Continuum Mechanics (see section of the same name page 3734).

9.6.8.3 Laplace's law

We have proved in the section of Continuum Mechanics with the virial theorem (see page 3897) that the internal energy (kinetic energy) of a perfect monoatomic gas was given by:

$$U=\frac{3}{2}kNT \tag{9.6.152}$$

where we took the notation of the section of Continuum Mechanics (an uppercase N instead of an uppercase n).

Therefore we have:

$$\mathrm{d}U=\frac{3}{2}kN\mathrm{d}T \tag{9.6.153}$$

If the transformation process is at constant volume, we will assume that there is no mechanical work (force over a distance) provided (inelastic collisions on the walls) and then:

$$\mathrm{d}U=\delta Q+\delta W=\delta Q_V=\frac{3}{2}kN\mathrm{d}T \tag{9.6.154}$$

therefore where $\mathrm{d}W$ is zero!

It comes then the possibility to calculate the heat capacity at constant volume:

$$\boxed{C_V=\left(\frac{\delta Q}{\mathrm{d}T}\right)_V=\frac{3}{2}kN=\frac{3}{2}\frac{R}{N_{\text{A}}}N=\frac{3}{2}R\frac{N}{N_{\text{A}}}} \tag{9.6.155}$$

where we then observed, that for an ideal gas, the volumetric heat capacity is independent of the volume !! We have verbatim for a mole:

$$C_V^M=\frac{C_V}{\dfrac{N}{N_{\text{A}}}}=\frac{C_V}{N}N_{\text{A}}=\frac{3}{2}R \tag{9.6.156}$$

so that we can write for an ideal monoatomic gas at constant volume the internal energy:

$$U=\frac{3}{2}R\frac{N}{N_{\text{A}}}T=C_VT \tag{9.6.157}$$

which allows to know in practice, and from table, the internal energy variation of a gas depending on its temperature variation since:

$$\Delta U=C_V\Delta T \tag{9.6.158}$$

which of course corresponds to a supply work or absorbed work. Under its differential form we name it the "first Joule's law" (internal energy of an ideal gas depends only on the temperature)

$$\boxed{\mathrm{d}U=nC_v\mathrm{d}T} \tag{9.6.159}$$

An interesting case of the prior previous relation is to remember that as we are at constant volume we have then:

$$\left.\begin{array}{l}P_1V_1=nRT_1\\P_2V_2=nRT_2\end{array}\right\}\frac{P_1}{T_1}=\frac{P_2}{T_2}\Rightarrow T_2=\frac{P_2}{P_1}T_1 \tag{9.6.160}$$

This allows us to write:

$$\Delta U=C_V\Delta T=C_V(T_2-T_1)=C_V\left(\frac{P_2-P_1}{P_1}\right)T_1 \tag{9.6.161}$$

With the important part:

$$\boxed{\Delta U=C_V(T_2-T_1)=C_V\left(\frac{P_2-P_1}{P_1}\right)T_1} \tag{9.6.162}$$

So if there is decompression (gas expansion), the latter relation is negative. In this case the gas loses heat to the environment this is why when a container of a gas release a gas it cools down!

If the process takes place at constant pressure (constant kinetic energy of the gas atoms) then we have (see virial theorem in the section of Continuum Mechanics page 3897):

$$\delta W=-P\mathrm{d}V=-kN\mathrm{d}T \tag{9.6.163}$$

(the collisions that push the wall of the volume make the system loose hence the sign "$-$").

Therefore:

$$\mathrm{d}U=\delta Q_P+\delta W\Rightarrow\delta Q_P=\mathrm{d}U-\delta W=\frac{3}{2}kN\mathrm{d}T+kN\mathrm{d}T=\frac{5}{2}kN\mathrm{d}T \tag{9.6.164}$$

Thus we can also calculate the heat capacity at constant pressure:

$$\boxed{C_P=\left(\frac{\delta Q}{t}\right)_P=\frac{5}{2}kN=\frac{5}{2}kN=\frac{5}{2}\frac{R}{N_\mathrm{A}}N=\frac{5}{2}R\frac{N}{N_\mathrm{A}}} \tag{9.6.165}$$

where can therefore observed, for an ideal gas, that the heat capacity at constant pressure is also independent of the pressure !! We have verbatim for a mole:

$$c_P^M=\frac{C_P}{\dfrac{N}{N_\mathrm{A}}}=\frac{C_P}{N}N_\mathrm{A}=\frac{5}{2}R \tag{9.6.166}$$

Of the two boxed previous results we get for an ideal mono-atomic gas the "heat capacity ratio" or also named "adiabatic index" or "ratio of specific heats" or "Poisson constant":

$$\boxed{\gamma=\frac{C_P}{C_V}=\frac{5}{3}} \tag{9.6.167}$$

with the "Mayer relation":

$$\boxed{C_P-C_V=\frac{5}{2}R\frac{N}{N_\mathrm{A}}-\frac{3}{2}R\frac{N}{N_\mathrm{A}}=nR} \tag{9.6.168}$$

which gives us the same time as R is positive:

$$C_P > C_V \tag{9.6.169}$$

We get also a relation that will be useful later:

$$C_P = C_V + nR = \frac{C_P}{\gamma} + nR \Rightarrow C_P\left(1 - \frac{1}{\gamma}\right) = nR \Rightarrow = nR\frac{\gamma}{\gamma - 1} \tag{9.6.170}$$

If the ideal gas is diatomic, there are 5 degrees of freedom (3 for the position of the first atom, 3 for the position of the second one, -1 for the constraint that the distance between them is fixed) and we then have:

$$\mathrm{d}U \cong \frac{5}{2}kN\mathrm{d}T \tag{9.6.171}$$

By doing the same developments, we get (value that will use in the section Music Mathematics but which is useful in many other fields of physics):

$$\boxed{\gamma = \frac{C_P}{C_V} = \frac{7}{5}} \tag{9.6.172}$$

When an isolated system of ideal gas undergoes an adiabatic transformation at constant pressure, the internal energy change in the system will be drawn by the variation of internal work. Which traditionally is written by a negative sign such that(using the result above):

$$\mathrm{d}U = -\delta W = -P\mathrm{d}V = C_v\mathrm{d}T \tag{9.6.173}$$

Remark

Warning!!! Let us recall that the choice of the sign for the work W is only a sign convention !! Thus, in this case study it is the tradition to put a "$-$" instead of a "$+$". But this does not change the results that will follow !!!

Hence:

$$P\mathrm{d}V + C_V\mathrm{d}T = 0 \tag{9.6.174}$$

Let us now take the ideal gas equation $PV = nRT$ (without collisions) and differentiate. We get:

$$P\mathrm{d}V + V\mathrm{d}P = Nk\mathrm{d}T \Rightarrow \mathrm{d}T = \frac{P\mathrm{d}V + V\mathrm{d}P}{Nk} \tag{9.6.175}$$

Therefore by eliminating $\mathrm{d}T$ between the last two relations, we get:

$$P\mathrm{d}V + C_V\frac{P\mathrm{d}V + V\mathrm{d}P}{Nk} = 0 \tag{9.6.176}$$

Thus after simplification and rearrangement of the terms:

$$(Nk + C_V)P\mathrm{d}V + C_V V\mathrm{d}P = 0 \tag{9.6.177}$$

What reported to moles quantities is written (as customary for chemists)

$$(R + c_V^M)p\mathrm{d}V + c_V^M V\mathrm{d}p = 0 \tag{9.6.178}$$

anyway ... by remembering that:

$$C_P - C_V = nR = Nk \tag{9.6.179}$$

We have:

$$C_P P\mathrm{d}V + C_V V\mathrm{d}P = 0 \tag{9.6.180}$$

Using the definition of the Laplace coefficient, also named "adiabatic coefficient" already met earlier:

$$\gamma = \frac{C_P}{C_V} \tag{9.6.181}$$

we get the expression:

$$\gamma P\mathrm{d}V + V\mathrm{d}P = 0 \tag{9.6.182}$$

Rearranging we get:

$$\gamma\frac{\mathrm{d}V}{V} + \frac{\mathrm{d}P}{P} = 0 \tag{9.6.183}$$

We get by integration:

$$\gamma\int\frac{\mathrm{d}V}{V} + \int\frac{\mathrm{d}P}{P} = 0 \tag{9.6.184}$$

therefore:

$$\ln(P) + \gamma\ln(V) = c^{te} \tag{9.6.185}$$

which is equivalent using the properties of logarithms to:

$$\boxed{PV^{\gamma} = c^{te}} \tag{9.6.186}$$

This is the famous "Laplace's thermodynamcis law" or "polytropic gas equation" which gives the relation between pressure and volume in an adiabatic transformation of a gas (which does not mean that the temperature is constant for recall but only that exchange of heat with the outside system is zero or at least negligible!). We will reuse besides this relation in the section of Marine & Weather Engineering.

Thus, we also have the following information that may be useful in the industry:

$$\begin{aligned} W &= \int_{V_1}^{V_2} P\mathrm{d}V = c^{te}\int_{V_1}^{V_2} V^{-\gamma}\mathrm{d}V = c^{te}\frac{V^{1-\gamma}\Big|_{V_1}^{V_2}}{1-\gamma} = c^{te}\frac{V_2^{1-\gamma} - V_1^{1-\gamma}}{1-\gamma} = \frac{c^{te}V_2^{-\gamma}V_2 - c^{te}V_1^{-\gamma}V_1}{1-\gamma} \\ &= \frac{P_2V_2 - P_1V_1}{1-\gamma} = \frac{Nk(T_2 - T_1)}{1-\gamma} \end{aligned} \tag{9.6.187}$$

As well as:

$$\frac{P_2}{P_1}=\left(\frac{V_1}{V_2}\right)^{\gamma}=\left(\frac{\dfrac{N_1kT_1}{P_1}}{\dfrac{N_2kT_2}{P_2}}\right)^{\gamma}\underset{N_1=N_2}{=}\left(\frac{T_1P_2}{T_2P_1}\right)^{\gamma}\Rightarrow\frac{T_1}{T_2}=\left(\frac{P_2P_1^{\gamma}}{P_1P_2^{\gamma}}\right)^{1/\gamma}=\left(\frac{P_1^{\gamma-1}}{P_2^{\gamma-1}}\right)^{1/\gamma}=\left(\frac{P_1}{P_2}\right)^{\gamma-1/\gamma} \tag{9.6.188}$$

verbatim:

$$\frac{T_1}{T_2}=\left(\frac{P_1}{P_2}\right)^{\gamma-1/\gamma}=\left(\left(\frac{V_2}{V_1}\right)^{\gamma}\right)^{\gamma-1/\gamma}=\left(\frac{V_2}{V_1}\right)^{\gamma-1} \tag{9.6.189}$$

So in the end we have the very useful relations in practice for the polytropic case:

$$\boxed{\frac{T_1}{T_2}=\left(\frac{P_1}{P_2}\right)^{\gamma-1/\gamma}=\left(\frac{V_2}{V_1}\right)^{\gamma-1}} \tag{9.6.190}$$

The problem in practice... is to have in the value of the adiabatic coefficient ...

OK so far we can make the following small summary hoping it helps:

	ΔQ	ΔW	ΔU	ΔH
Isothermal	$\Delta Q=nRT\ln\left(\frac{V_f}{V_i}\right)$ $=nRT\ln\left(\frac{P_f}{P_i}\right)$	$-\Delta Q$	0	0
Isochoric	$\Delta Q=nC_v\Delta T=\Delta U$	0	$\Delta U=nC_v\Delta T$	$\Delta H=nC_P\Delta T$
Isobaric	$\Delta Q=nC_P\Delta T=\Delta H$	$\Delta W=-P\Delta V$ $=-nR\Delta T$	$\Delta U=nC_v\Delta T$	$\Delta H=nC_P\Delta T$
Adiabatic	0	$\Delta W=nC_V\Delta T$	$\Delta U=nC_v\Delta T$	$\Delta H=nC_P\Delta T$

Table 9.28: **Main ideal gas processes (reversible processes)**

We will use the polytropic gas equation above for an adiabatic expansion to calculate the change ΔP in pressure when we change the volume of the gas in a cylinder by an amount ΔV. Before we compress we have:

$$PV^{\gamma}=P_0V_0^{\gamma} \tag{9.6.191}$$

where P_0 and V_0 are our original pressure and volume.

After the expansion, V goes to $V_0+\Delta V$ and P goes to $P_0+\Delta P$, where we know that ΔP is negative for an expansion. Thus after expansion we have:

$$PV^{\gamma}=(P_0+\Delta P)(V_0+\Delta V)^{\gamma} \tag{9.6.192}$$

With $PV^{\gamma}=P_0V_0^{\gamma}=c^{te}$, we get:

$$P_0V_0^{\gamma}=(P_0+\Delta P)(V_0+\Delta V)^{\gamma} \tag{9.6.193}$$

We can use the fact that ΔV is very small compared to V_0 to get:

$$(V_0+\Delta V)^{\gamma}=\left[V_0\left(1+\frac{\Delta V}{V_0}\right)\right]=V_0^{\gamma}\left(1+\frac{\Delta V}{V_0}\right)^{\gamma} \tag{9.6.194}$$

Using the binomial series (see section Calculus page 1364):

$$(1+\alpha)^{\gamma}\cong 1+\gamma\alpha \tag{9.6.195}$$

for a small α, we have:

$$\left(1+\frac{\Delta V}{V_0}\right)^{\gamma}\cong 1+\gamma\frac{\Delta V}{V_0} \tag{9.6.196}$$

Therefore:

$$P_0V_0^{\gamma}=(P_0+\Delta P)(V_0+\Delta V)^{\gamma}=P_0\left(1+\frac{\Delta P}{P_0}\right)V_0^{\gamma}\left(1+\gamma\frac{\Delta V}{V_0}\right) \tag{9.6.197}$$

Multiplying this out gives:

$$P_0V_0^{\gamma}=P_0V_0^{\gamma}\left(1+\frac{\Delta P}{P_0}+\frac{\gamma\Delta V}{V_0}+\frac{\gamma}{P_0V_0}\Delta P\Delta V\right) \tag{9.6.198}$$

The factors $P_0V_0^{\gamma}$ cancel, and we can neglect the second order term $\Delta P\Delta V$, giving:

$$1=1+\frac{\Delta P}{P_0}+\frac{\gamma\Delta V}{V_0} \tag{9.6.199}$$

After cancelling the 1's and multiplying through by P_0 we get for the pressure change ΔP:

$$\boxed{\Delta P=-\gamma P_0\left(\frac{\Delta V}{V_0}\right)} \tag{9.6.200}$$

9.6.8.4 Saint-Venant Thermodynamic Equation

There are several equations that bear the name of "Saint-Venant equation" in physics ... one that concerns us here is very useful in the field of aero-dynamics and is more "thermodynamics" oriented that the equation of the same name in the field of incompressible fluid mechanics.

Let us see what it is!

We will consider a compressible gas or fluid always using the first law of thermodynamics but by highlighting from the internal energy explicitly three macroscopic energies commonly used in the field of Continuum Mechanics:

$$\Delta U=\Delta E_c+\Delta E_p+\Delta H=\Delta Q+\Delta W \tag{9.6.201}$$

To introduce the Saint-Venant equation for thermodynamics as simple as possible, we will consider only the cases where the changes in potential energy, heat exchanges and work are negligible or zero (which is a strong hypothesis but not non-existent in some practical cases). Then we have:

$$\Delta E_p+\Delta H=\frac{1}{2}m(v_2^2-v_1^2)+(H_2-H_1)=0 \tag{9.6.202}$$

which seems sometimes be named "Thomson equation" or "Zeuner formula" (we give names to anything sometimes...).

Which brings us to write:

$$\frac{1}{2}mv^2+H=c^{te} \tag{9.6.203}$$

Now we have prove earlier above that:

$$\Delta H\cong C_P\Delta T=mc_P\Delta T=nR\frac{\gamma}{\gamma-1}\Delta T=nR\frac{\gamma}{\gamma-1}\Delta\left(\frac{PV}{nR}\right)=\frac{\gamma}{\gamma-1}\Delta(PV) \tag{9.6.204}$$

Often the latter relation is expressed in massic form such that:

$$\frac{\Delta H}{n}=\Delta h\cong\frac{\gamma}{\gamma-1}\Delta\left(P\frac{V}{m}\right)=\frac{\gamma}{\gamma-1}\Delta\left(\frac{P}{\rho}\right) \tag{9.6.205}$$

or written differently:

$$h=\frac{\gamma}{\gamma-1}\frac{P}{\rho} \tag{9.6.206}$$

It comes therefore the "Saint-Venant equation for thermodynamics":

$$\boxed{\frac{1}{2}v^2+h=\frac{1}{2}v^2+\frac{\gamma}{\gamma-1}\frac{P}{\rho}=c^{te}} \tag{9.6.207}$$

If we rewrite this relation between two points but we choose one of the two points has zero speed (whatever which one) we can always bring us back to a relation of the form:

$$\frac{\gamma}{\gamma-1}\frac{P_{\text{eg}}}{\rho_{\text{eg}}}=\frac{\gamma}{\gamma-1}\frac{P}{\rho}+\frac{1}{2}v^2 \tag{9.6.208}$$

But we have proved in the section of Music Mathematics that the speed of sound was given by:

$$v_{\text{sound}}=\sqrt{\frac{\gamma P}{\rho}}=\sqrt{\frac{\gamma RT}{M_m}}=\sqrt{\gamma rT} \tag{9.6.209}$$

where r is the ideal gas massic constant ...

Then we can write:

$$\frac{1}{\gamma-1}v_{\text{sound,eg}}^2=\frac{1}{\gamma-1}v_{\text{sound}}^2+\frac{1}{2}v^2\Leftrightarrow\frac{2}{\gamma-1}\frac{v_{\text{sound,eg}}^2}{v_{\text{sound}}}=\frac{2}{\gamma-1}+\left(\frac{v}{v_{\text{sound}}}\right)^2 \tag{9.6.210}$$

We then introduce the "Mach number":

$$\boxed{M=\frac{v}{v_{\text{sound}}}} \tag{9.6.211}$$

Therefore:

$$\frac{2}{\gamma-1}\frac{v_{\text{sound,eg}}^2}{v_{\text{sound}}^2}=\frac{2}{\gamma-1}+M^2 \tag{9.6.212}$$

But we also have:

$$\frac{v_{\text{sound,eg}}^2}{v_{\text{sound}}^2}=\frac{\gamma r T_{\text{eg}}}{\gamma r T}=\frac{T_{\text{eg}}}{T} \tag{9.6.213}$$

Hence after a small rearrangement:

$$\frac{v_{\text{sound,eg}}^2}{v_{\text{sound}}^2}=\frac{T_{\text{eg}}}{T}=1+\frac{\gamma-1}{2}M^2 \tag{9.6.214}$$

But using the relations proved above and obtained from the Laplace's law we get:

$$\frac{T_{\text{eg}}}{T}=\left(\frac{P_{\text{eg}}}{P}\right)^{\dfrac{\gamma-1}{\gamma}}=\left(\frac{V}{V_{\text{eq}}}\right)^{\gamma-1}=\left(\frac{\rho_{\text{eg}}}{\rho}\right)^{\gamma-1}=1+\frac{\gamma-1}{2}M^2 \tag{9.6.215}$$

✍**Example:**

Compressed air contained in a large tank escapes outward through a hole at a Mach number equal to $M=0.77$. Relaxation occurs in the atmosphere where there is an atmospheric pressure of 1.014 [bar]. The massic ratio of air at the NTP is given as being worth $\gamma=1.4$. We then have the pressure inside the tank which is given by:

$$\left(\frac{P_{\text{eg}}}{P}\right)^{\frac{\gamma-1}{\gamma}}=1+\frac{\gamma-1}{2}M^2\Rightarrow P_{\text{eg}}=P\left(1+\frac{\gamma+1}{2}M^2\right)^{\frac{\gamma}{\gamma-1}}\cong 1.5\,[\text{bar}] \tag{9.6.216}$$

We can also have fun by calculating the internal necessary pressure of the tank for the flow to become supersonic (and then do not forget that ... action = reaction... lol).

9.6.8.5 Thermoelastics coefficients

If we differentiate $V(P,T)$ we have (see Differential and Integral Calculus):

$$\mathrm{d}V=\frac{\partial V}{\partial T}\mathrm{d}T+\frac{\partial V}{\partial P}\mathrm{d}P \tag{9.6.217}$$

or written differently:

$$\mathrm{d}V=\left(\frac{\partial V}{\partial T}\right)_P\mathrm{d}T+\left(\frac{\partial V}{\partial P}\right)_T\mathrm{d}P \tag{9.6.218}$$

We have introduced naturally in the section Mathematical Music a coefficient named "coefficient of isothermal compressibility" that was for recall (the index T indicates that the temperature is kept constant):

$$\boxed{\chi_T=-\frac{1}{V}\left(\frac{\partial V}{\partial P}\right)_T} \qquad (9.6.219)$$

that we can reintroduce here:

$$\mathrm{d}V=\left(\frac{\partial V}{\partial T}\right)_P\mathrm{d}T-\chi_T V\mathrm{d}P \qquad (9.6.220)$$

Also it would be interesting to have a different coefficient for the first term that it would suffice to define by analogy (the index V is there to indicate that this is a volumic coefficient and the P to indicate that we are at constant pressure):

$$\boxed{\alpha_{V,P}=\frac{1}{V}\left(\frac{\partial V}{\partial T}\right)_P} \qquad (9.6.221)$$

named "isobaric volume expansion coefficient".

Thus we have:

$$\mathrm{d}V=\alpha_{V,P}V\mathrm{d}T-\chi_T V\mathrm{d}P \qquad (9.6.222)$$

Therefore we have the mechanical work (the total differential is inaccurate because we have more than one state variable) by multiplying the pressure to have the correct units:

$$\delta W=\alpha_{V,P}PV\mathrm{d}T-\chi_T PV\mathrm{d}P \qquad (9.6.223)$$

For an isothermal transformation the first term is zero, and for an isobaric transformation, it is the second which is zero.

The fact to know the thermoelastic coefficients (measured experimentally) must bring us to fall back on the ideal gas law by integration of $V(P,T)$, which is minimum we should expect, since V is a state function. In the case of the ideal gas, for example, we can write intuitively by the dimensions of the constants (we will introduce further below during our study of the equations of state the reverse mathematical path for pedagogical reasons):

$$\alpha=\frac{1}{T} \qquad \chi=\frac{1}{P} \qquad (9.6.224)$$

Therefore we have:

$$\mathrm{d}V=\frac{V}{T}\mathrm{d}T-\frac{V}{P}\mathrm{d}P \qquad (9.6.225)$$

and therefore by multiplying by the pressure:

$$P\mathrm{d}V=\frac{PV}{T}\mathrm{d}T-V\mathrm{d}P\Rightarrow P\mathrm{d}V+V\mathrm{d}P=PV\frac{\mathrm{d}T}{T} \qquad (9.6.226)$$

Which leads to:

$$\mathrm{d}(PV)=PV\frac{\mathrm{d}T}{T} \qquad (9.6.227)$$

Therefore:

$$\frac{\mathrm{d}(PV)}{PV}=\frac{\mathrm{d}T}{T} \tag{9.6.228}$$

Giving immediately after integration:

$$\ln(PV)=\ln(T)+c^{te}\Rightarrow\ln\left(\frac{PV}{T}\right)=c^{te} \tag{9.6.229}$$

Remark

Notice that in the above relation we have something typical in Thermodynamics and Astrophysics: a logarithm of a non-dimensionless argument...!! The use for example of $\ln(T)$ etc. is always a shorthand for $\ln\left(T/T_0\right)$ which is okay to use if for some reason we don't care about T_0, i.e. because it cancels out or we are interested in the asymptotic behaviour only. So it's a lazy shorthand - the kind of thing much beloved by physicists...

Therefore:

$$PV=c^{te} \tag{9.6.230}$$

That said, we have differentiated V to get two coefficients such as:

$$\mathrm{d}V=\alpha V\mathrm{d}T-\chi V\mathrm{d}P \tag{9.6.231}$$

We could do the same for the pressure and temperature, and then we have a total of three relations:

$$\begin{aligned}\mathrm{d}V&=\left(\frac{\partial V}{\partial P}\right)_T\mathrm{d}P+\left(\frac{\partial V}{\partial T}\right)_P\mathrm{d}T\\ \mathrm{d}P&=\left(\frac{\partial P}{\partial T}\right)_V\mathrm{d}T+\left(\frac{\partial P}{\partial V}\right)_T\mathrm{d}V\\ \mathrm{d}T&=\left(\frac{\partial T}{\partial P}\right)_V\mathrm{d}P+\left(\frac{\partial T}{\partial V}\right)_P\mathrm{d}V\end{aligned} \tag{9.6.232}$$

but among the 6 factors that we see in these three relations, 4 are already defined (some are the opposite of the coefficients defined above). By cons a definition is missing for a single coefficient for the both missing factors. We choose the one that in practice is most often used by analogy to the other factors:

$$\boxed{\beta=\frac{1}{P}\left(\frac{\partial P}{\partial T}\right)_V} \tag{9.6.233}$$

named "coefficient of increase of isochoric pressure".

We have therefore three very used coefficients in practice:

$$\chi_T=-\frac{1}{V}\left(\frac{\partial V}{\partial P}\right)_V\qquad\alpha_{V,P}=\frac{1}{V}\left(\frac{\partial V}{\partial T}\right)_P\qquad\beta=\frac{1}{P}\left(\frac{\partial P}{\partial T}\right)_V \tag{9.6.234}$$

respectively in the order:

1. Coefficient of isothermal compressibility χ_T
2. Isobaric expansion coefficient $\alpha_{V,P}$
3. Isochoric pressure increasing coefficient β

We will see these coefficients again in our study of convective motions in the section of Weather & Marine Engineering.

9.6.9 Heat

Before continuing, we have at all costs to eliminate one of the greatest difficulties in the understanding of thermodynamics (apart from that of differentiation between correct and incorrect total differential that has already been set in the section of Differential and Integral Calculus):

- The difference between heat and temperature
- The difference between the energy-work and energy-heat

As we saw in the section of Continuum Mechanics, the temperature characterize a state of thermodynamic equilibrium and translates the existence of a thermal motion (virial theorem page 3897) and can vary when the outside provides a work $W>0$. However, experience shows us that it is by "heating" a system that we increase the more easily its temperature. But then what is the "heat"?

Let us consider a thermodynamic system in equilibrium, and let us write its total internal energy E as the sum of the products of the energy E of a microstate i by the population size P of the same microstate i:

$$E=\sum_i E_i P_i \tag{9.6.235}$$

Its variation during an infinitesimal transformation (which we assume with a constant number of particles) is (total differential):

$$\mathrm{d}E=\left(\frac{\partial E_i}{\partial P_i}\right)_{E_i}\mathrm{d}P_i+\left(\frac{\partial P_i}{\partial E_i}\right)\mathrm{d}E_i=\sum_i E_i\mathrm{d}P_i+\sum_i P_i\mathrm{d}E_i \tag{9.6.236}$$

where $\mathrm{d}E_i$ is the variation of the energy E_i of the microstate i caused by the transformation and $\mathrm{d}P_i$ the variation of the population P_i of this same microstate i.

We have seen earlier above that under certain conditions, if the system during the infinitesimal transformation undergoes a variation of work δW and of heat δQ, its internal energy varies of:

$$\mathrm{d}E=\mathrm{d}U=\delta W+\delta Q \tag{9.6.237}$$

We can now compare this last relation (expressing the first law of thermodynamics) with the one before (resulting from Statistical Mechanics) and identify the terms as follows:

$$\boxed{\mathrm{d}U=\sum_i E_i\mathrm{d}P_i+\sum_i P_i\mathrm{d}E_i=\delta W+\delta Q} \tag{9.6.238}$$

Let us examine the case of processing in which the system receives only heat:

$$\delta W = 0 \quad (9.6.239)$$

In this situation no external parameters to the system generally varies in the transformation so that $\mathrm{d}E_i = 0$. We conclude:

$$\delta Q = \sum_i E_i \mathrm{d}P_i \quad (9.6.240)$$

Thus, when a system receives only heat, its energy vary by the change of the populations of its microstates: If the amount of heat received is positive, the probability of the high energy microstates increases, at the expense of the microstates of lower energy. Finally, if we take into account that the system must be in equilibrium (the most probable state as we have proved it in the section of Statistical Mechanics) in the initial and final states of transformation, we find, as we should expect, that the system temperature varies: it has increased if the amount of heat received is positive (this can be proved by choosing a canonical distribution to describe the system of macroscopic equilibrium states as done in the section of Statistical Mechanics).

This explains the frequent confusion between these two very different concepts that are temperature and heat. This confusion is compounded by the gap between everyday language and scientific terminology. In everyday language, when we talk about "heat" of a body, we say actually that its temperature is high. The confusion is unfortunate because the concept of heat is present in physics but its meaning is different.

Thus, heat a system, is provide heat to it, is increasing its internal energy (the number of high-energy micro-states) by means that are not purely mechanical. The heat is thus a particular form of energy!

9.6.9.1 Entropy

An isolated macroscopic system tends towards equilibrium. It reached it in a finite time (which can be extremely long). The steady state is unique: the exceptions to this claim are too special to merit a digression at least for now.

The very existence of a state of equilibrium is fundamental to thermodynamics. However, the equilibrium walk process is not due to a dogma: there must be no dogma in physics as we know (and in science in general)! Like any other law, it is subject verified experimentally and must be analysed. One question in particular arises: what is the microscopic counterpart of the equilibrium walk process, that is a macroscopic process also.

We proved in the section Statistical Mechanics that by definition: the equilibrium state is the state which is have the largest number of configurations (microstates) and is the most probable state.

What had brought us to the following relation (see page 3525):

$$S = \lambda \ln(\Omega) \tag{9.6.241}$$

where S represents the statistical expected mean of information on microstates and that we named for recall "entropy". Clearly S has the units of λ that as we showed it is a constant.

The question that arises in thermodynamics is: what is the constant that permits characterize for a gas, liquid or solid the expected number of microstates.

Then it comes when looking at all the relation in thermodynamics one constant that always appears when it comes to characterize a thermodynamic state. This is the Boltzmann constant:

$$\lambda = k = 1.3806 \cdot 10^{-23}\,[\mathrm{J} \cdot \mathrm{K}^{-1}] \tag{9.6.242}$$

So S has units corresponding to the ratio J / K that thus permits to measure the degree of disorder in a system at the microscopic level. Intuitively: the higher the entropy of the system, the less its elements are ordered and capable of producing mechanical effects, and the greater the part of energy unused or used inconsistently.

The entropy is an extensive quantity. Indeed, we have proved that the choice of the logarithm in the Boltzmann law came from the fact that the entropy of a macro state was the expected mean of all microstates:

$$\begin{aligned} S &= -\lambda \sum_{i=1}^{n} p_i \log(p_i) = -\lambda \left(\frac{1}{\Omega} \log\left(\frac{1}{\Omega}\right) + \frac{1}{\Omega} \log\left(\frac{1}{\Omega}\right) + \ldots + \frac{1}{\Omega} \log\left(\frac{1}{\Omega}\right) \right) \\ &= -\lambda \left(-\frac{1}{\Omega} \log(\Omega) - \frac{1}{\Omega} \log(\Omega) - \ldots - \frac{1}{\Omega} \log(\Omega) \right) = \lambda \log(\Omega) \left(\frac{1}{\Omega} + \frac{1}{\Omega} + \ldots + \frac{1}{\Omega} \right) \end{aligned} \tag{9.6.243}$$

what had brought us to:

$$S = \lambda \ln(\Omega) \tag{9.6.244}$$

and shows well that entropy is an extensive quantity because summable on micro-states (complexions).

What remains difficult now is whether to know if the energy in the units of entropy comes from the work W or from heat Q or from both? In fact the answer is quite simple because in our development of the Boltzmann law, at any time the system (ideal) studied has provided (mechanical) work. So the only energy involved is that of heat.

Therefore:

$$S=\frac{Q}{T} \tag{9.6.245}$$

However, entropy can not be given rigorously by:

$$S=\frac{\Delta Q}{\Delta T} \tag{9.6.246}$$

because that is the definition of specific heat! Furthermore, if the reader remembers our developments in the section of Statistical Mechanics, the study was done in an isolated system with two cavities. So if the transition from one cavity to the other is very slow (so that there is not an expansion of the gas) the temperature will remain constant (isothermal expansion). Which implies the definition (so we fall back on the discrete form of the second law of thermodynamics already introduced at the beginning of this section):

$$\Delta S=\frac{\Delta Q}{T} \tag{9.6.247}$$

To move to the differential form, it should be remembered that heat Q is an inexact differential. So for a reversible transformation:

$$\mathrm{d}S=\frac{\delta Q}{T} \tag{9.6.248}$$

Therefore the entropy is an exact total differential (heat depends on how the transformation takes place, but S depends only on the final and initial state of the heat Q) and as:

$$S=\lambda\ln(\Omega)\geq 0 \tag{9.6.249}$$

Finally we have in an adiabatic (constant energy), closed and finite system within a static space-time:

$$\boxed{S=\frac{\delta Q}{T}\geq 0} \tag{9.6.250}$$

That's the second law of thermodynamics! If the system is finite and in an adiabatic transformation (without heat and work exchange - energy - with the outside), then the numerator is zero and the entropy is also zero verbatim (this is why an adiabatic process is necessarily isentropic!). Otherwise, the system takes the entropy to the Universe in a natural evolution.

This means that the entropy (expected mean of the intrinsic information) in a system in contact with the outside can only increase or remain constant.

What is important is that any processes (non-adiabatic) converting energy from one form to another in an isolated system necessarily loses some form of heat.

Regarding the Universe ... the question is whether this is an isolated thermodynamic system or not and finite or not and static or not...!? There are therefore a few difficulties with our Universe:

- We can observe and do measurements only for the observable Universe! The Universe in it's entirety isn't observable and therefore we can't measure the value of its entropy (there are only actually speculative theoretical models that try to speculate the entropy for the observable Universe).

- Modern measurements give evidence that our observable Universe is flat and therefore the entire Universe may be infinite (hence it is not closed). And as the second law of thermodynamics holds only for finite closed systems therefore we don't know if the latter can be applied to our Universe or not.

- As we have already prove it with the Noeother theorems (see page 3132), if space-time isn't static then energy isn't conversed and therefore the second law of thermodynamics doesn't applied. And actually all empirical measurement show that the space-time structure of our observable Universe is dynamic. Therefore for the observable Universe the second principle of thermodynamics doesn't hold.

Hence although it is attempting to apply thermodynamics to the whole Universe, and some of the founding fathers of Thermodynamics did it and also some scientists still do it because of the publish or perish paradigm..., a closer inspection of the problem suggest some caution as show with our enumeration above!

> Caution!!! There is no principle of physics yet (in the beginning in the 121st century according to holocene calendar) that says physicals laws or constants have to be the same everywhere and always. Indeed, the reader must remember that the "laws" of physics are **descriptive**, not **imperative** and therefore they are some observed regularity in what we can observe experimentally with sufficiently wide occurrence or relevance! Asking whether the laws can change is no more and no less than asking whether regularities which have been observed up till now might not be quite as universal as was thought. The answer to that is: YES! This could drive the reader a bit mad. Indeed, how can we possibly do Science if the rules keep changing? But if you are a professional active scientist, you have to face up to that possibility - and you have a responsibility to check out if there is evidence beyond reasonable doubt about it! We will repeat this warning multiple times in this textbook as its seems that this is something necessary!

We then have the very useful relation in fluid mechanics (and cosmology):

$$\boxed{\mathrm{d}U=\delta Q+\delta W=T\mathrm{d}S-P\mathrm{d}V} \tag{9.6.251}$$

which is named "thermodynamic identity" relatively to the internal energy U or "characteristic function of a fluid at equilibrium" or more often "fundamental Gibbs' equation" (we will notice that it is a sum of products of intensive and extensive quantities). If the amount of material is not constant (material inserted or extracted from the system), the latter relation is then written for a variation of moles $\mathrm{d}n$:

$$\boxed{\mathrm{d}U=T\mathrm{d}S-P\mathrm{d}V+\mu\mathrm{d}n} \tag{9.6.252}$$

where μ is a coefficient of proportionality named "chemical potential" and can therefore be measured by the fact that the previous relation leads to the definition:

$$\boxed{\mu=\left(\frac{\partial U}{\partial n}\right)_{S,V}} \tag{9.6.253}$$

which means that the chemical potential μ is, as we already know, the internal energy variation of the gas when we inject it in or withdraw from it $\mathrm{d}n$ moles of gas, while keeping constant the volume of the tank and prohibiting heat exchange with the environment.

We can often found in the prior previous relations follows after a simple rearrangement in many text books (and assuming constant the quantity of material):

$$\mathrm{d}S=\frac{\mathrm{d}U}{T}+\frac{P}{T}\mathrm{d}V \tag{9.6.254}$$

The latter relation is often assimilated to the first law of thermodynamics for closed systems for whose the overall variation of potential and kinetic energy is constant.

Let us notice that by using the relation between the internal variation of energy of an ideal monoatomic gas in function of its temperature at constant volume as proved earlier above (first Joule's law):

$$\mathrm{d}U=C_V\mathrm{d}T \tag{9.6.255}$$

and the ideal gas equation (see section Continuum Mechanics page 3903):

$$PV=nRT \tag{9.6.256}$$

We then have using the molar heat capacity at constant volume (remember that we can inject into this relation only for an ideal gas because it is then independent of the volume!):

$$\mathrm{d}S=nc_V^M\frac{\mathrm{d}T}{T}+nR\frac{\mathrm{d}V}{V} \tag{9.6.257}$$

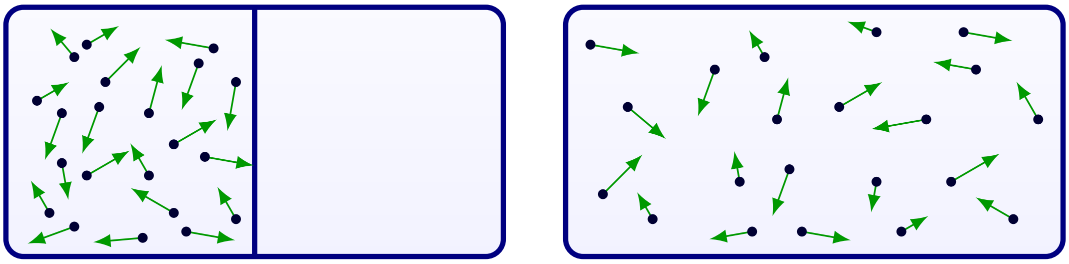
So integrating between two states, we get:

$$\begin{aligned}\Delta S&=\int\limits_1^2\mathrm{d}S=\int\limits_1^2 nc_V^M\frac{\mathrm{d}T}{T}+nR\frac{\mathrm{d}V}{V}=\int\limits_1^2 nc_V^M\frac{\mathrm{d}T}{T}+\int\limits_1^2 nR\frac{\mathrm{d}V}{V}\\&=nc_V^M\int\limits_1^2\frac{\mathrm{d}T}{T}+nR\int\limits_1^2\frac{\mathrm{d}V}{V}=nc_V^M\ln(T)|_1^2+nR\ln(V)|_1^2=nc_V^m\ln\left(\frac{T_2}{T_1}\right)+nR\ln\left(\frac{V_2}{V_1}\right)\end{aligned} \tag{9.6.258}$$

Thus the entropy change of an ideal gas at constant pressure:

$$\Delta S=nc_V\ln\left(\frac{T_2}{T_1}\right)+nR\ln\left(\frac{V_2}{V_1}\right) \tag{9.6.259}$$

And for the famous case at constant temperature known as "Joule expansion" as illustrated below:



We have therefore:

$$\Delta S = nR\ln\left(\frac{V_2}{V_1}\right) \geq 0 \tag{9.6.260}$$

And as we have also proved above that:

$$c_V^M = \frac{3}{2}R \tag{9.6.261}$$

It comes therefore by using the algebraic properties of the logarithm:

$$\begin{aligned}\Delta S &= \frac{3nR}{2}\ln\left(\frac{T_2}{T_1}\right) + nR\ln\left(\frac{V_2}{V_1}\right) = nr\left[\frac{3}{2}\ln\left(\frac{T_2}{T_1}\right) + \ln\left(\frac{V_2}{V_1}\right)\right] \\ &= nR\left[\ln\left(\left(\frac{T_2}{T_1}\right)^{3/2}\right) + \ln\left(\frac{V_2}{V_1}\right)\right] = nR\ln\left(\left(\frac{T_2}{T_1}\right)^{3/2}\frac{V_2}{V_1}\right)\end{aligned} \tag{9.6.262}$$

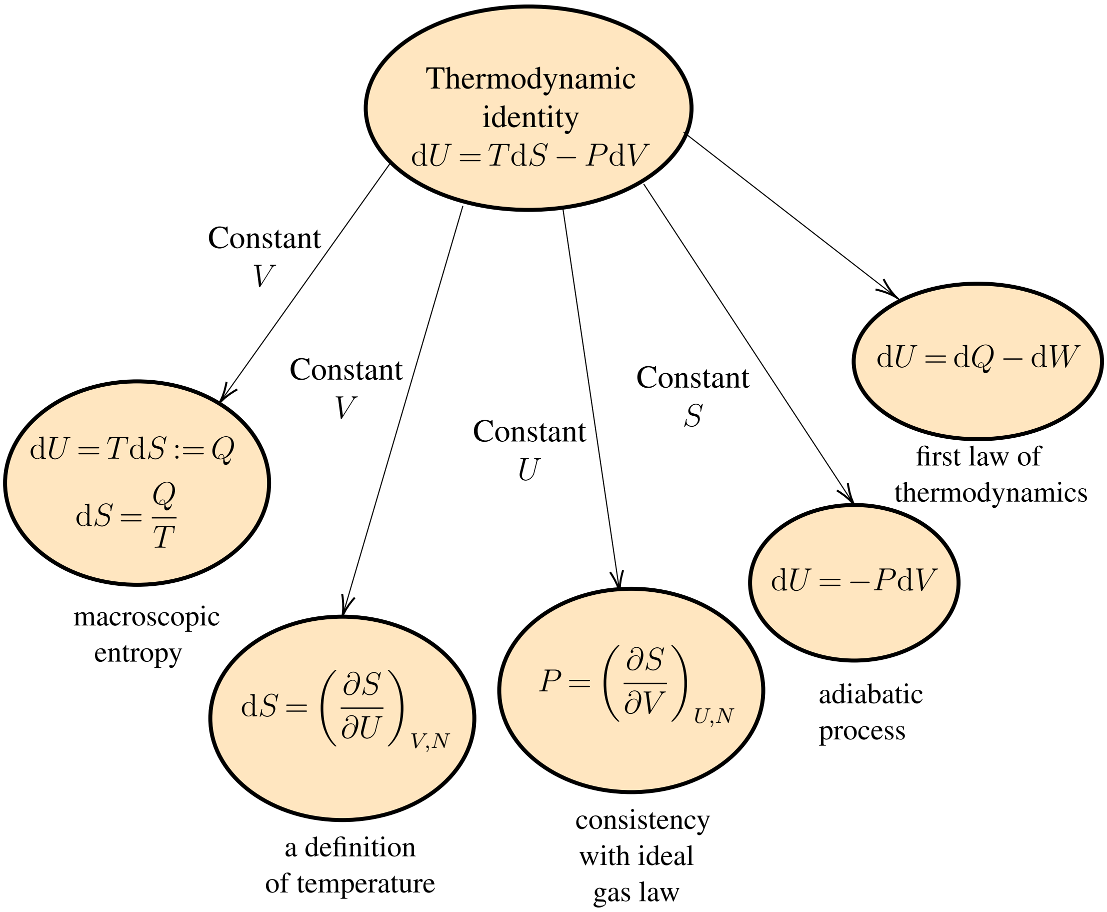
And as in practice it is easier to measure the pressure than the volume, we use the ideal gas law where at constant temperature we have (see section Continuum Mechanics page 3890):

$$P_1V_1 = P_2V_2 \tag{9.6.263}$$

to get the very important next result in practice considered the entropy change of an ideal gas at constant volume:

$$\boxed{\Delta S = nR\ln\left(\left(\frac{T_2}{T_1}\right)^{3/2}\frac{P_1}{P_2}\right)} \tag{9.6.264}$$

The reader must not forget that if the calculation with numerical applications give negative values for the change in entropy then the interpretation is quite simple: that means that the direction of the arbitrary considered transformation is not the real (natural) direction in which would conduct the system left to itself in the Universe (i.e. it is then an artificial transformation).


Thermodynamic identity
$\mathrm{d}U = T\mathrm{d}S - P\mathrm{d}V$
Constant V
Constant V
Constant U
Constant S
$\mathrm{d}U = \mathrm{d}Q - \mathrm{d}W$
first law of thermodynamics
$\mathrm{d}U = T\mathrm{d}S := Q$
$\mathrm{d}S = \dfrac{Q}{T}$
macroscopic entropy
$\mathrm{d}U = -P\mathrm{d}V$
adiabatic process
$\mathrm{d}S = \left(\dfrac{\partial S}{\partial U}\right)_{V,N}$
a definition of temperature
$P = \left(\dfrac{\partial S}{\partial V}\right)_{U,N}$
consistency with ideal gas law

Now let us derive the entropy expression based on the canonical partition function as promised in the section of Statistical Mechanics!

By definition (see section Statistical Mechanics page **??**) the canonical partition function is $Z_e = \sum_i e^{-\beta E_i}$, where $\beta = \frac{1}{k_B T}$ and E_i are the energy levels.

The probability P_i of finding the system in a specific microstate i is given by the Boltzmann distribution (still see **??**):

$$p_i = \frac{e^{-\beta E_i}}{Z} \tag{9.6.265}$$

Statistical mechanics defines entropy via the Shannon entropy formula (see section Statistical Mechanics page 3530):

$$S = -k_B \sum_i p_i \ln(p_i) \tag{9.6.266}$$

Therefore:

$$S = -k_B \sum_i p_i \left(-\beta E_i - \ln(Z)\right) = k_B \beta \sum_i p_i E_i + k_B \ln(Z) \sum_i p_i \tag{9.6.267}$$

Note that the sum of all probabilities equals one ($\sum_i p_i = 1$) and the average energy is $U = \sum_i p_i E_i$, therefore:

$$S = k_B \beta U + k_B \ln(Z) \tag{9.6.268}$$

Replacing β with $1/k_BT$, leads to:

$$S = \frac{U}{T} + k_B \ln(Z) \tag{9.6.269}$$

Alternatively, since (see section page **??**):

$$U = -\frac{\partial \ln(Z)}{\partial \beta} \tag{9.6.270}$$

this can be written directly in terms of temperature derivatives as:

$$\boxed{S = k_B \ln(Z) + k_BT \left(\frac{\partial \ln(Z)}{\partial T}\right)_V} \tag{9.6.271}$$

Now let us derive the entropy expression based on the canonical partition function as promised in the section of Statistical Mechanics!

By definition (see section Statistical Mechanics page **??**) the canonical partition function is $Z_c = \sum_i e^{-\beta E_i}$, where $\beta = \frac{1}{kT}$ and E_i are the energy levels.

The probability P_i of finding the system in a specific microstate i is given by the Boltzmann distribution (still see **??**):

$$p_i = \frac{e^{-\beta E_i}}{Z_c} \tag{9.6.272}$$

Statistical mechanics defines entropy via the Shannon entropy formula (see section Statistical Mechanics page 3530):

$$S = -k \sum_i p_i \ln(p_i) \tag{9.6.273}$$

Therefore:

$$S = -k \sum_i p_i \left(-\beta E_i - \ln(Z_c)\right) = k\beta \sum_i p_i E_i + k \ln(Z_c) \sum_i p_i \tag{9.6.274}$$

Note that the sum of all probabilities equals one ($\sum_i p_i = 1$) and the average energy is $U = \sum_i p_i E_i$, therefore:

$$S = k\beta U + k \ln(Z_c) \tag{9.6.275}$$

Replacing β with $1/kT$, leads to:

$$S = \frac{U}{T} + k \ln(Z_c) \tag{9.6.276}$$

Alternatively, since (see section page 3911):

$$U = -\frac{\partial \ln(Z_c)}{\partial \beta} \tag{9.6.277}$$

this can be written directly in terms of temperature derivatives as:

$$\boxed{S = k \ln(Z_c) + kT \left(\frac{\partial \ln(Z_c)}{\partial T}\right)_V} \tag{9.6.278}$$

9.6.9.1.1 Life, evolution, and the second law of thermodynamics

Some people misunderstand the second law of thermodynamics, stated in terms of entropy, to say that the process of the evolution of life violates this law. Over time, complex organisms evolved from much simpler ancestors, representing a large decrease in entropy of the Earth's biosphere. It is a fact that the huge majority of living organisms have evolved (for more information see the excellent [453]) to be highly structured, and much lower in entropy than the substances from which they grow. But it is always possible for the entropy of one part of the universe to decrease, provided the total change in entropy of the universe increases. In equation form, we can write this as $\Delta S_{\text{tot}} = \Delta S_{\text{syst}} + \Delta S_{\text{envir}} > 0$.

Thus ΔS_{syst} can be negative as long as ΔS_{envir} is positive and greater in magnitude.

How is it possible for a system to decrease its entropy? Energy transfer is necessary! If we pick up marbles that are scattered about the room and put them into a cup, our work has decreased the entropy of that system. If we gather iron ore from the ground and convert it into steel and build a bridge, our work has decreased the entropy of that system. Energy coming from the Sun (actually a G2V-type star, aka yellow dwarf) can decrease the entropy of local systems on Earth that is, ΔS_{syst} is negative. But the overall entropy of the rest of the universe increases by a greater amount-that is, ΔS_{envir} is positive and greater in magnitude. Thus, $\Delta S_{\text{tot}} = \Delta S_{\text{syst}} + \Delta S_{\text{envir}} > 0$, and the second law of thermodynamics is not violated.

Every time a plant stores some solar energy in the form of chemical potential energy, or an updraft of warm air lifts a soaring bird, the Earth can be viewed as a heat engine operating between a hot reservoir supplied by the Sun and a cold reservoir supplied by dark outer space-a heat engine of high complexity, causing local decreases in entropy as it uses part of the heat transfer from the Sun into deep space. There is a large total increase in entropy resulting from this massive heat transfer. A small part of this heat transfer is stored in structured systems on Earth, producing much smaller local decreases in entropy.

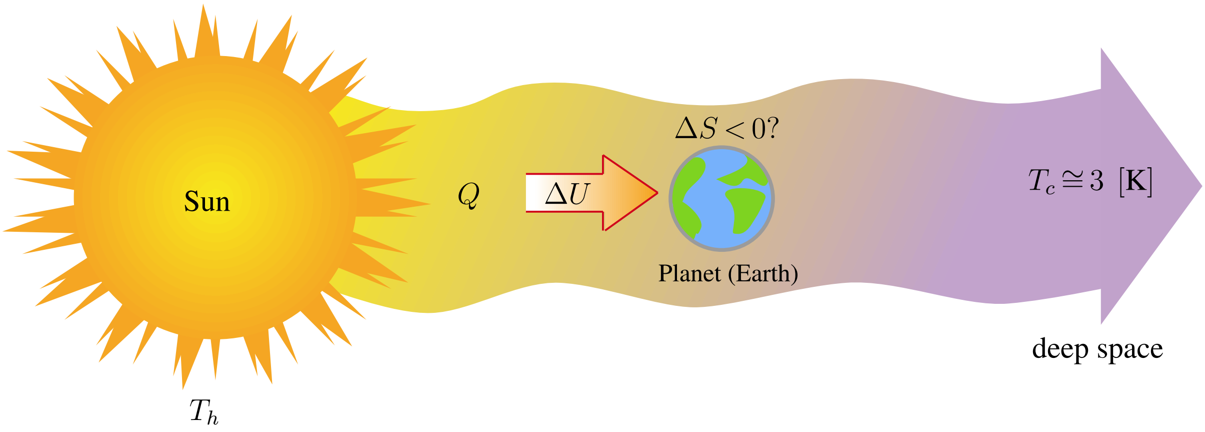


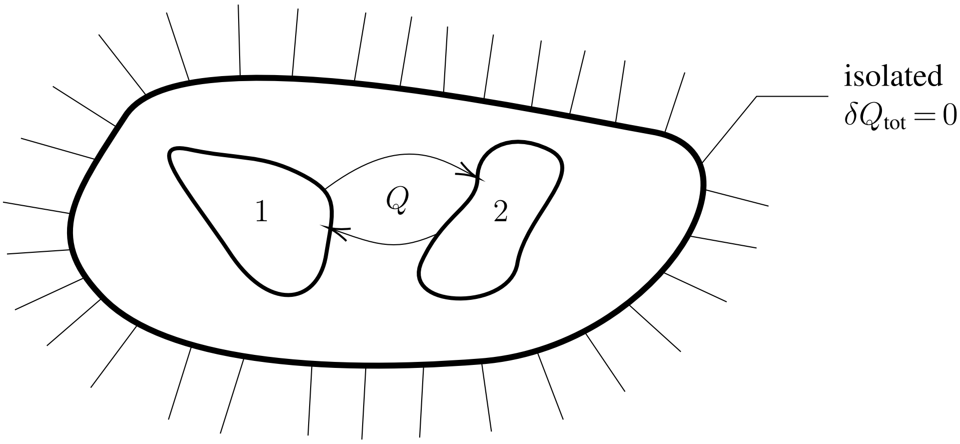
Figure 9.167: **Earth's entropy may decrease in the process of intercepting a small part of the heat transfer from the Sun into deep space (not to scale)**

Sun is a (nice) source of low entropy for Earth: A concentrate of energy that permits life to be structure itself! For each photon of the Sun of a given energy received the Earth will re-emit the equivalent concentrate energy (low entropy) in the form of an average of twenty photons of dissipated energy (higher entropy). The Sun get itself its low entropy it gives to Earth from gravity. As in a closed system... gravity can decrease entropy by concentrating matter in a small volume of space!

9.6.9.1.2 Heat Flow

Now that we have introduced the basics of thermodynamics let do on an example of very important application of the first and second laws of thermodynamics.

Let us consider two closed systems 1 and 2 being in thermal contact and forming a total system that is assumed thermally isolated from the environment:



In both systems, there are no mechanical parts, such as pistons or other, that can work. Therefore:

$$\delta W_1=\delta W_2=\delta W_{\text{tot}}=0 \tag{9.6.279}$$

The conditions and assumptions being known, we use ingredients of thermodynamics to make our recipes. It then comes by applying the first principle as proved above in a first time for the whole:

$$\begin{aligned}\mathrm{d}U_{\text{tot}}&=\mathrm{d}U_1+\mathrm{d}U_2\\ \mathrm{d}U_{\text{tot}}&=\underbrace{\delta Q_{\text{tot}}}_{=0}+\underbrace{\delta W_{\text{tot}}}_{=0}=0\end{aligned} \tag{9.6.280}$$

We deduce that:

$$\mathrm{d}U_1+\mathrm{d}U_2=0 \tag{9.6.281}$$

We also have for each sub-system:

$$\begin{aligned}\mathrm{d}W_1&=0=\mathrm{d}U_1-\delta Q_1\\ \mathrm{d}W_2&=0=\mathrm{d}U_2-\delta Q_2\end{aligned} \tag{9.6.282}$$

Hence we deduce from the latter system of equations by summing:

$$\mathrm{d}U_1+\mathrm{d}U_2=\delta Q_1+\delta Q_2 \tag{9.6.283}$$

and as we have proved just before:

$$\mathrm{d}U_1+\mathrm{d}U_2=0 \tag{9.6.284}$$

It comes:

$$\mathrm{d}U_1+\mathrm{d}U_2=0=\delta Q_1+\delta Q_2 \tag{9.6.285}$$

and therefore (obvious!):

$$\delta Q_1=-\delta Q_2 \tag{9.6.286}$$

Let us now apply the second principle (principle of evolution). We have:

$$\mathrm{d}S_{\text{tot}}=\mathrm{d}S_1+\mathrm{d}S_2=\frac{\delta Q_1}{T_1}+\frac{\delta Q_2}{T_2}=\frac{\delta Q_{\text{tot}}}{T_1} \tag{9.6.287}$$

But with the prior previous relation, we have:

$$\mathrm{d}S_{\text{tot}}=\frac{\delta Q_1}{T_1}+\frac{-Q_1}{T_2}=\delta Q_1\left(\frac{1}{T_1}-\frac{1}{T_2}\right) \tag{9.6.288}$$

We now have two possible scenarios. For the first, we have:

$$T_2>T_1\rightarrow\left(\frac{1}{T_1}-\frac{1}{T_2}\right)>0 \tag{9.6.289}$$

and when:

$$T_2<T_1\Rightarrow\delta Q_1>0 \tag{9.6.290}$$

Therefore:

$$\mathrm{d}S_{\text{tot}}>0 \tag{9.6.291}$$

For the second scenario, we have:

$$T_1>T_2\rightarrow\left(\frac{1}{T_1}-\frac{1}{T_2}\right)y0 \tag{9.6.292}$$

and when:

$$T_1>T_2\Rightarrow\delta Q_1<0 \tag{9.6.293}$$

Therefore:

$$\mathrm{d}S_{\text{tot}}>0 \tag{9.6.294}$$

We conclude that whatever happens, entropy increases for an isolated system in a static space-time without gravity. This result is therefore consistent with what we communicated during our presentation of the second principle. So this result proves the impossibility for an isolated dithermal thermal machine to provide a perpetual work because the entropy can only increase in all cases!

We also see, as we have already mentioned it, that entropy defines the arrow of time in the Universe (or vice versa).

9.6.9.2 Carnot Cycle

The Carnot cycle's main objective is to calculate the thermal efficiency of a thermodynamic cycle of a dithermal idealized machine ... (so the model is far from the practice!) Based on a heat source from which we take heat to turn it into work and then from which we use work to provide heat and so on in loop(think of a locomotive for example!). We speak then of "reverse cycle".

The idea is pretty simple ... but as always someone had to think to it first!

For information we have chosen here a mathematical approach of the Carnot cycle which allows to avoid any integral and that is closer to the reasoning made at the time of Carnot by colas Léonard Sadi Carnot himself. This approach can therefore make bristle the hair to some readers ...

From the perspective of the internal energy, on a thermodynamic cycle, there will be no variations (since at each cycle beginning we must found ourselves in the situation of the previous cycle). Therefore:

$$\Delta U=0 \qquad (9.6.295)$$

This is so far simply a recall of the principle of conservation of energy (first law of thermodynamics).

This means that on a cycle the heat balance Q_e borrowed by the machine to produce a (mechanical) work W_p and of the heat Q_i re-injected and that required a (mechanical) work W_i is zero on a cycle such that:

$$Q_e+Q_i+W_p+W_i=0 \qquad (9.6.296)$$

where what is obviously borrowed will, as it is customary, have a positive sign and what is injected a negative sign !!!!! It is customary to notice this last relation in the following condensed form:

$$Q_e+Q_i+W=0 \Leftrightarrow W=-(Q_e+Q_i) \qquad (9.6.297)$$

The "thermal efficiency" η is defined as the ratio of the work that machine can really provide (i.e. the difference between what it gives W_p and what we must give to it W_i so that it works) on heat that was injected (with the working equation). So:

$$\boxed{\eta=\frac{|W_p+W_i|}{Q_i}=\frac{|W|}{Q_i}=\frac{|-(Q_e+Q_i)|}{Q_i}=\frac{Q_e+Q_i}{Q_i}=1+\frac{Q_e}{Q_i}} \qquad (9.6.298)$$

which is always less than 1 (and greater than 0), since the amount of heat removed (positive sign by convention) is necessarily less than or equal to the amount of injected heat (negative sign by convention).

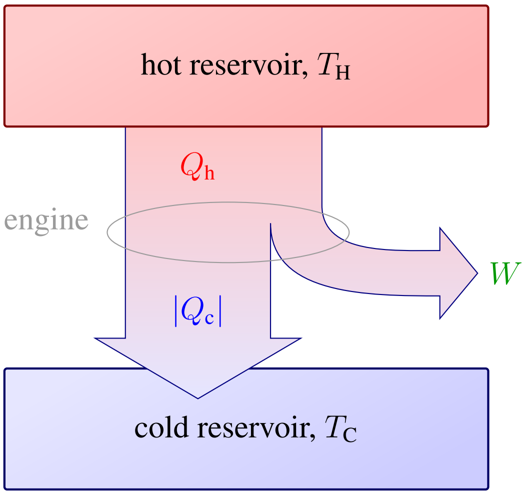

hot reservoir, T_{H}
Q_{h}
engine
W
$|Q_{\mathrm{c}}|$
cold reservoir, T_{C}

Figure 9.168: **A schematic representation ("heat bath scheme") of a heat engine**

However, we have proved that in the conditions of isothermal expansion or isothermal compression (so the cycle is assumed to be very slow in the Cournot model), we had:

$$\Delta Q = T\Delta S \tag{9.6.299}$$

But, in our case Q_e, Q_i are in reality an abusive notation in thermodynamics as it is heat variation (we should notice them objectively ΔQ_e,ΔQ_i). We then have for the reversible thermal efficiency for isothermal compression or expansion:

$$\eta = 1 + \frac{T_e \Delta S_e}{T_i \Delta S_i} \tag{9.6.300}$$

The transformation being quasi-static, the relation obtained is obviously the theoretical maximum efficiency for an engine operating between these temperatures. It is never reached in a real cycle! We then say that the "Carnot efficiency" is a maximum (the entropy injected being in a real case always greater than the entropy extracted by construction!).

This leads us to the "Kelvin-Planck statement" that says:

- First alternative: it (seems as far as we know) that it is impossible to design a thermal machine that describes a cycle and that would have for only effect to produce work and to exchange heat with a single thermal tank.
- Second alternative: it is impossible for a thermal machine making a cyclic process to convert entirely into work the heat it absorbs.

In the case of a steam engine, the maximum theoretical efficiency would be (remember that the temperature is in Kelvin, and always positive!):

$$\eta = 1 - \frac{T_e}{T_i} = 1 - \frac{273+25}{273+100} \cong 0.2 = 20\% \tag{9.6.301}$$

For example in the case of a central nuclear working with pressurized water, we find the famous performance often mentioned in the press:

$$\eta = 1 - \frac{T_e}{T_i} \cong 1 - \frac{273+35}{273+285} \cong 45\% \tag{9.6.302}$$

The real thermal efficiency of the conversion of nuclear fission heat to electricity is in the order of only 33% because the real thermal efficiency is always less than the theoretical maximum thermal efficiency! At the time of Carnot the theoretical efficiency of a steam engine was 32%. However in practice because of heat dissipation, friction, etc. the practical efficiency of a steam engine was around 3% only!

Remark

Notice that to reach a 100% efficiency the last relation would require absolute zero to one side or infinite temperature to the other side...

Here are some other values:

Type	Carnot efficiency
Electric generator	$70-99\%$
Dry battery	90%
Industrial electric motor	$60-90\%$
Hydrogen-oxygen fuel cell	60%
Diesel motor	$34-50\%$
Liquid fuel rocket	47%
Steam turbine	$34-46\%$
Fossil fuel power station	40%
GE90 (Boeing 777)	36.1%
Nuclear Fuel Power Plant	32%
Petrol engine	27%
Human body	25%
Solar Cell	$20-30\%$
Piston engine	$10-25\%$

Table 9.29: Some typical Carnot efficiencies

The relation:

$$\eta=1-\frac{T_e}{T_i} \tag{9.6.303}$$

has a very useful interest in practice. Indeed, if an engineer said to have built a machine that is able to convert for example 1000 [kJ] in input in 410 [kJ] of pure work with an efficiency of 41%, then other engineers can check if this thermal efficiency is overestimated by measuring the temperature of the input source and that of the output.

Remark

We may notice the interest of the condenser. If we let escape to the atmosphere the steam coming out of turbines, firstly we would lose purified water instead of recycling it (for reuse in the cycle), on the other part we would have a heat source of higher temperature: 100 [°C] instead of 100 [°C] we would lose about 12% of the Carnot efficiency.

Finally, we implicitly also find one of the historical statements of the second law of thermodynamics considering where equation. In this case, it is not effective and the engine does not therefore provide any work.

Let us return briefly to the fact that in the ideal case of the cycle considered above which for recall is therefore an ideal thermodynamic cycle consisting of four reversible processes: isothermal expansion, adiabatic expansion (so isentropic because reversible), isothermal compression, and adiabatic compression... :

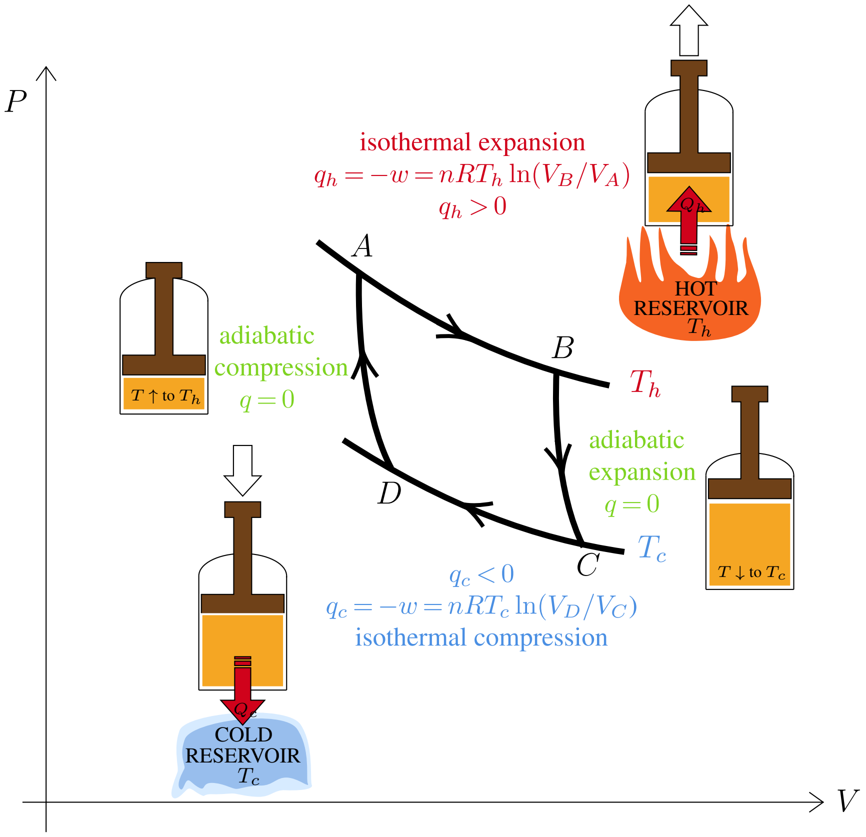

P
V
isothermal expansion
$q_h=-w=nRT_h\ln(V_B/V_A)$
$q_h>0$
A
B
C
D
T_h
T_c
HOT RESERVOIR T_h
Q_h
adiabatic compression
$q=0$
$T\uparrow$ to T_h
adiabatic expansion
$q=0$
$T\downarrow$ to T_c
$q_c<0$
$q_c=-w=nRT_c\ln(V_D/V_C)$
isothermal compression
Q_c
COLD RESERVOIR T_c

Figure 9.169: **Carnot cycle illustration**

we were led to write:

$$\Delta S_{\text{cyc.}}=\Delta S_e+\Delta S_i=0 \tag{9.6.304}$$

But the second law of thermodynamics says that the entropy change is always positive or zero, what we wrote for recall:

$$\mathrm{d}S=\frac{\delta Q}{T}\geq 0 \tag{9.6.305}$$

but in the above reversible case, for the cyclic variation of entropy of the reversible process to be zero that mean that one of the both term violates must violate this principle (at least in the case of a cycle) and therefore has a negative entropy (so that the sum is zero) such that we are led to write:

$$\mathrm{d}S=\frac{\delta Q}{T}\leq 0 \tag{9.6.306}$$

It is a violation of the reality that thermodynamicists name "Clausius inequality". It is then customary to summarize the thermodynamic cycles as follows:

- $\displaystyle\oint\frac{\delta Q}{T}=0$: no irreversibility in the system

- $\oint \dfrac{\delta Q}{T} > 0$: irreversibilities in the system (the reality of our Universe)
- $\oint \dfrac{\delta Q}{T} < 0$: impossible (at least as far as we know...)

As the second relation listed above is not possible we then have for any system exchanging heat with external reservoirs and undergoing a cyclic process:

$$\boxed{\Delta S = \oint \frac{\delta Q}{T} \leq 0} \tag{9.6.307}$$

that is also named "Clausius theorem".

This theorem can be proved but all the proofs I know so far are not aesthetic in my point of view. But it's quite obvious and this relation is just a special case of the second principle of thermodynamics.

Especially the case if during a cycle no information is lost (as we have proved in the section of Statistical Mechanics, entropy is related to information!) is quite obvious that:

$$\Delta S = \oint \frac{\delta Q}{T} \leq 0 \tag{9.6.308}$$

9.6.10 Maxwell relations

Maxwell's relations are a set of equations in thermodynamics which are derivable from the symmetry of second derivatives and from the definitions of the thermodynamic potentials. These relations can be useful when we are not able to directly measure a given thermodynamics property but that we can workaround this issue by expressing it in terms of other thermodynamic variables.

Let us return first to what we have already mentioned a little earlier above but by restricting ourselves to two variables. That is to say the exact total differential:

$$\mathrm{d}z = M\mathrm{d}x + N\mathrm{d}y = \frac{\partial z}{\partial x}\mathrm{d}x + \frac{\partial z}{\partial y}\mathrm{d}y = \left(\frac{\partial z}{\partial x}\right)_y \mathrm{d}x + \left(\frac{\partial z}{\partial y}\right)_x \mathrm{d}y \tag{9.6.309}$$

When we assume that we are authorized to write the differential of a state variable in thermodynamics as above, and especially for a simple compressible system that is completely specified by two independent, intensive properties, then we say that we are using the "state postulate".

Remark

Two properties are considered independent if one can be varied while the other is held constant. For example, temperature and specific volume are always independent. However, temperature and pressure are independent only for a single-phase system; for a multiphase system (such as a mixture of gas and liquid) this is not the case. (e.g., boiling point (temperature) depends on elevation (ambient pressure)).

So we have also:

$$\begin{aligned}\mathrm{d}x&=\left(\frac{\partial x}{\partial y}\right)_z\mathrm{d}y+\left(\frac{\partial x}{\partial z}\right)_y\mathrm{d}z\\ \mathrm{d}y&=\left(\frac{\partial y}{\partial x}\right)_z\mathrm{d}x+\left(\frac{\partial y}{\partial z}\right)_x\mathrm{d}z\\ \mathrm{d}z&=\left(\frac{\partial z}{\partial x}\right)_y\mathrm{d}x+\left(\frac{\partial z}{\partial y}\right)_x\mathrm{d}y\end{aligned}\tag{9.6.310}$$

By inserting $\mathrm{d}y$ in $\mathrm{d}x$:

$$\begin{aligned}\mathrm{d}x&=\left(\frac{\partial x}{\partial y}\right)_z\left(\left(\frac{\partial y}{\partial x}\right)_z\mathrm{d}x+\left(\frac{\partial y}{\partial z}\right)_x\mathrm{d}z\right)+\left(\frac{\partial x}{\partial z}\right)_y\mathrm{d}z\\ &=\left(\frac{\partial x}{\partial y}\right)_z\left(\frac{\partial y}{\partial x}\right)_z\mathrm{d}x\left(\left(\frac{\partial x}{\partial y}\right)_z\left(\frac{\partial y}{\partial z}\right)_x+\left(\frac{\partial x}{\partial z}\right)_y\right)\mathrm{d}z\end{aligned}\tag{9.6.311}$$

or also:

$$\left[\left(\frac{\partial x}{\partial y}\right)_z\left(\frac{\partial y}{\partial x}\right)_z-1\right]\mathrm{d}x+\left(\left(\frac{\partial x}{\partial y}\right)_z\left(\frac{\partial y}{\partial z}\right)_x+\left(\frac{\partial x}{\partial z}\right)_y\right)\mathrm{d}z=0\tag{9.6.312}$$

as the terms in parenthesis are functions and that $\mathrm{d}x$ and $\mathrm{d}z$ are by cons arbitrary, the only solution to this relation is:

$$\left(\frac{\partial x}{\partial y}\right)_z\left(\frac{\partial y}{\partial x}\right)_z-1=0\qquad\left(\frac{\partial x}{\partial y}\right)_z\left(\frac{\partial y}{\partial z}\right)_x+\left(\frac{\partial x}{\partial z}\right)_y=0\tag{9.6.313}$$

Multiplying the second relation by $(\partial z/\partial x)_y$:

$$\left(\frac{\partial z}{\partial x}\right)_y\left(\frac{\partial x}{\partial y}\right)_z\left(\frac{\partial y}{\partial z}\right)_x+\left(\frac{\partial z}{\partial x}\right)_y\left(\frac{\partial x}{\partial z}\right)_y=0\tag{9.6.314}$$

Then we get after simplification:

$$\left(\frac{\partial z}{\partial x}\right)_y\left(\frac{\partial x}{\partial y}\right)_z\left(\frac{\partial y}{\partial z}\right)_x+1=0\tag{9.6.315}$$

Let us now turn to the facts. Let us recall the fundamental Gibbs' equation:

$$\mathrm{d}U=T\mathrm{d}S-P\mathrm{d}V\tag{9.6.316}$$

relation very useful in fluids where the pressure is constant and where the heat change is made by that of entropy. And the relation defining the enthalpy:

$$H=U+PV\tag{9.6.317}$$

from which will modify the differential:

$$\mathrm{d}H=\mathrm{d}U+V\mathrm{d}P+P\mathrm{d}V\tag{9.6.318}$$

by injecting in it (first principle):

$$\mathrm{d}U=\delta Q-P\mathrm{d}V \tag{9.6.319}$$

we have (notice that the transformation is isobaric and that the enthalpy change then represents only the amount of heat received by the closed system):

$$\mathrm{d}H=\delta Q+V\mathrm{d}P \tag{9.6.320}$$

Thus, having just added PV to the internal energy U we do have a state function (H) where we control the pressure and the entropy as independent variables, while for U alone, we control the entropy and volume as independent variables.

So we will use the following two relations that will be helpful:

$$\begin{aligned}\mathrm{d}U&=T\mathrm{d}S-P\mathrm{d}V\\ \mathrm{d}H&=T\mathrm{d}S+V\mathrm{d}P\end{aligned} \tag{9.6.321}$$

We now introduce a new quantity named "free energy" (which is really available in the system) and will be given by the "Gibbs-Helmholtz relation":

$$\boxed{F=U-TS} \tag{9.6.322}$$

and giving simply the difference between the internal energy and the heat energy dissipated because of the entropy at a given temperature. The idea being this time to control the volume and temperature as independent variables (see just a little further below when we take the differential), which is obviously an important situation in chemistry!

We also introduce another new quantity that we name "free enthalpy" or "Gibbs energy" (which is that really available in the system) and will be given identically by:

$$\boxed{G=H-TS=U+PV-TS} \tag{9.6.323}$$

which is simply the difference between the enthalpy and the dissipated heat energy because of the entropy at a given temperature.

So we have for the free energy the differential form:

$$\mathrm{d}F=\mathrm{d}U-S\mathrm{d}T-T\mathrm{d}S \tag{9.6.324}$$

by injecting in it the first principle and the second principle:

$$\mathrm{d}F=(\delta Q-P\mathrm{d}V)-S\mathrm{d}T-T\mathrm{d}S=(T\mathrm{d}S-P\mathrm{d}V)-S\mathrm{d}T-T\mathrm{d}S=-P\mathrm{d}V-S\mathrm{d}T \tag{9.6.325}$$

Thus we see that if the transformation is isothermal $\mathrm{d}F$ reduces to the work received by the closed system (since $\mathrm{d}T$ is zero).

Similarly for the free enthalpy:

$$\mathrm{d}G=\mathrm{d}H-S\mathrm{d}T-T\mathrm{d}S \tag{9.6.326}$$

by injecting:

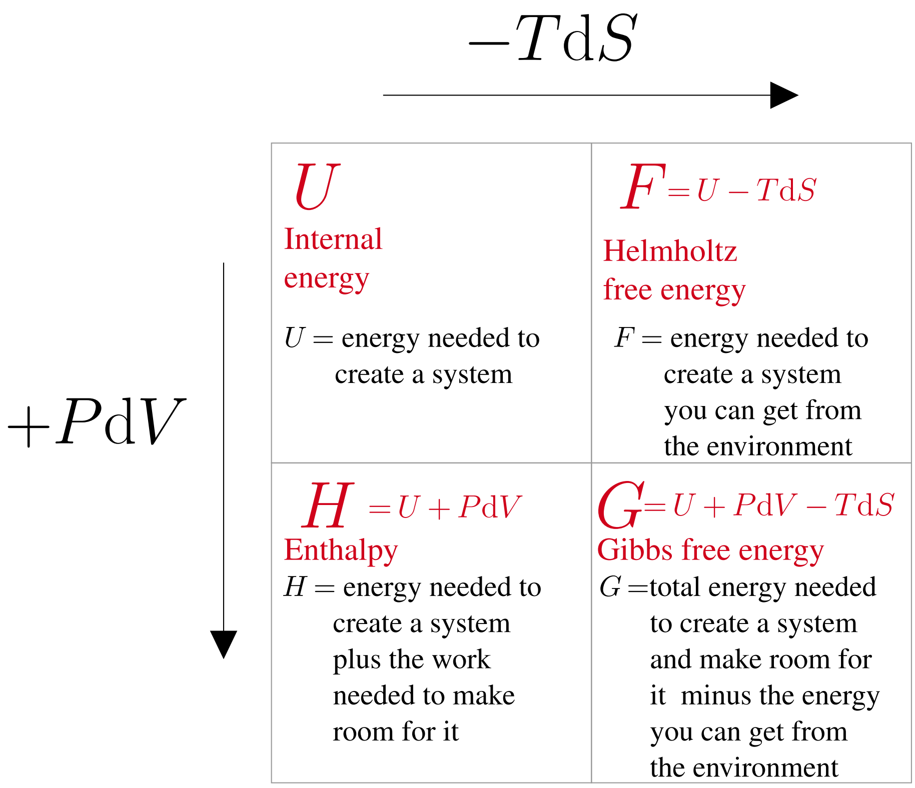
$$\mathrm{d}G=(T\mathrm{d}S+V\mathrm{d}P)-S\mathrm{d}T-T\mathrm{d}S=V\mathrm{d}P-S\mathrm{d}T \tag{9.6.327}$$

Therefore, in the case of an isobaric and isothermal transformation the free energy variation is equal zero.

So we have four relations:

$$\boxed{\begin{aligned} \mathrm{d}U &= T\mathrm{d}S - P\mathrm{d}V \\ \mathrm{d}H &= T\mathrm{d}S + V\mathrm{d}P \\ \mathrm{d}F &= -P\mathrm{d}V - S\mathrm{d}T \\ \mathrm{d}G &= V\mathrm{d}P - S\mathrm{d}T \end{aligned}} \tag{9.6.328}$$

named "Gibbs equations".



U Internal energy $U =$ energy needed to create a system	$F = U - T\mathrm{d}S$ Helmholtz free energy $F =$ energy needed to create a system you can get from the environment
$H = U + P\mathrm{d}V$ Enthalpy $H =$ energy needed to create a system plus the work needed to make room for it	$G = U + P\mathrm{d}V - T\mathrm{d}S$ Gibbs free energy $G =$total energy needed to create a system and make room for it minus the energy you can get from the environment

We notice that all these equations are of the form:

$$\mathrm{d}z = M\mathrm{d}x + N\mathrm{d}y \tag{9.6.329}$$

But, let recall that according to the Schwarz theorem (see section Differential and Integral Calculus page 1469) if $\mathrm{d}z$ is an exact total differential, then we have:

$$\left(\frac{\partial M}{\partial y}\right)_x = \left(\frac{\partial N}{\partial x}\right)_y \tag{9.6.330}$$

This gives us immediately the four most common "Maxwell relations":

$$\boxed{\begin{aligned}\left(\frac{\partial T}{\partial V}\right)_S&=-\left(\frac{\partial P}{\partial S}\right)_V\\\left(\frac{\partial T}{\partial P}\right)_S&=-\left(\frac{\partial V}{\partial S}\right)_P\\\left(\frac{\partial P}{\partial T}\right)_V&=\left(\frac{\partial S}{\partial V}\right)_T\\\left(\frac{\partial V}{\partial T}\right)_P&=-\left(\frac{\partial S}{\partial P}\right)_T\end{aligned}} \tag{9.6.331}$$

Remark

Notice that from:

$$\left(\frac{\partial V}{\partial T}\right)_P=-\left(\frac{\partial S}{\partial P}\right)_T \tag{9.6.332}$$

we get:

$$\Delta S=-\int_{P_1}^{P_2}\left(\frac{\partial V}{\partial T}\right)_P\mathrm{d}P \tag{9.6.333}$$

where we note that T is held constant during integration over P.

And as for an ideal gas we have:

$$\left(\frac{\partial V}{\partial T}\right)_p=\frac{nR}{P} \tag{9.6.334}$$

Therefore (isothermal):

$$\Delta S=-nR\int_{P_1}^{P_2}\frac{\mathrm{d}P}{P}=-nR\ln\left(\frac{P_2}{P_1}\right) \tag{9.6.335}$$

Furthermore, by the definition of partial derivatives and the four relations:

$$\begin{aligned}\mathrm{d}U&=T\mathrm{d}S-P\mathrm{d}V\\\mathrm{d}H&=T\mathrm{d}S+V\mathrm{d}P\\\mathrm{d}F&=-P\mathrm{d}V-S\mathrm{d}T\\\mathrm{d}G&=V\mathrm{d}P-S\mathrm{d}T\end{aligned} \tag{9.6.336}$$

We get immediately:

$$\boxed{\begin{aligned}\left(\frac{\partial U}{\partial S}\right)_V&=T & \left(\frac{\partial U}{\partial V}\right)_S&=-P\\ \left(\frac{\partial H}{\partial S}\right)_V&=T & \left(\frac{\partial H}{\partial P}\right)_S&=V\\ \left(\frac{\partial F}{\partial T}\right)_V&=-S & \left(\frac{\partial F}{\partial V}\right)_T&=-P\\ \left(\frac{\partial G}{\partial T}\right)_P&=-S & \left(\frac{\partial G}{\partial P}\right)_T&=V\end{aligned}} \tag{9.6.337}$$

All these relationships are used to calculate the thermodynamic variables not directly measurable from the experimental data.

For example, from the last one:

$$\left(\frac{\partial G}{\partial P}\right)_T=V \tag{9.6.338}$$

we get immediately:

$$\Delta G=\int_{P_1}^{P_2} V\,\mathrm{d}P \tag{9.6.339}$$

and considering and ideal gas again with $V=nRT/P$, we get (at constant temperature):

$$\Delta G=nRT\int_{P_1}^{P_2}\frac{1}{P}\mathrm{d}P=nRT\ln\left(\frac{P_2}{P_1}\right) \tag{9.6.340}$$

To close the subject about the definitions of these thermodynamic variables, let us indicate that the fact of creating (defining):

$$\begin{aligned}H&=U+PV\\F&=U-TS\\G&=H-TS\end{aligned} \tag{9.6.341}$$

or in other words, that fact that for every pair X,Y of conjugated thermodynamic variables and function f we put:

$$g=f-XY \tag{9.6.342}$$

is named a "Legendre transformation" (for the mathematical details of the transformation see the section of Analytical Mechanics). This type of transformation allows to change the set of independent variables to get a set of best suited variables to the problem considered.

Now let see a relation that will be useful to use in the section of Weather and Marine!

We know that the specific heat is given by definition at constant pressure by:

$$C_P=\frac{\Delta Q}{\Delta T} \tag{9.6.343}$$

But at constant pressure, the heat variation can be written by definition with the enthalpy:

$$C_P=\left(\frac{\partial H}{\partial T}\right)_P \tag{9.6.344}$$

Now remember that the enthalpy is given by:

$$\mathrm{d}H=T\mathrm{d}S+V\mathrm{d}P \tag{9.6.345}$$

as $\mathrm{d}S$ is an exact differential, we can write it as a function of the parameters of temperature and pressure only:

$$\mathrm{d}S=\left(\frac{\partial S}{\partial T}\right)_P\mathrm{d}T+\left(\frac{\partial S}{\partial P}\right)_T\mathrm{d}P \tag{9.6.346}$$

Therefore we have:

$$\mathrm{d}H=\left[\left(\frac{\partial S}{\partial T}\right)_P\mathrm{d}T+\left(\frac{\partial S}{\partial P}\right)_T\mathrm{d}P\right]+V\mathrm{d}P=T\left(\frac{\partial S}{\partial T}\right)_P\mathrm{d}T+\left[T\left(\frac{\partial S}{\partial P}\right)_T+V\right]\mathrm{d}P \tag{9.6.347}$$

As by the way $\mathrm{d}H$ is an exact differential, we can also write it in function of the parameters of temperature and pressure only:

$$\mathrm{d}H=\left(\frac{\partial H}{\partial T}\right)_P\mathrm{d}T+\left(\frac{\partial H}{\partial P}\right)_T\mathrm{d}P=C_P\mathrm{d}T+\left(\frac{\partial H}{\partial P}\right)_T\mathrm{d}P \tag{9.6.348}$$

We have then the two relations to identify:

$$\begin{aligned}\mathrm{d}H&=T\left(\frac{\partial S}{\partial T}\right)_P\mathrm{d}T+\left[T\left(\frac{\partial S}{\partial P}\right)_T+V\right]\mathrm{d}P\\ \mathrm{d}H&=C_P\mathrm{d}T+\left(\frac{\partial H}{\partial P}\right)_T\mathrm{d}P\end{aligned} \tag{9.6.349}$$

Remark

If the Enthalpy of a gas depends only of the temperature (quite unrealistic), the last relation above reduces to:

$$\mathrm{d}=C_P\mathrm{d}T \tag{9.6.350}$$

This is named the "second Joule's law".

It comes therefore:

$$\boxed{T\left(\frac{\partial S}{\partial T}\right)_P=C_P\qquad T\left(\frac{\partial S}{\partial P}\right)_T+V=\left(\frac{\partial H}{\partial P}\right)_T} \tag{9.6.351}$$

9.6.11 Continuity Equation

A continuity equation in physics is an equation that describes the transport of some quantity. It is particularly simple and particularly powerful when applied to a conserved quantity, but it can be generalized to apply to any extensive quantity. Since mass, energy, momentum, electric charge and other natural quantities are conserved under their respective appropriate conditions, a variety of physical phenomena may be described using continuity equations.

Continuity equations are a stronger, local form of conservation laws. For example, a weak version of the law of conservation of energy states that energy can neither be created nor destroyed i.e., the total amount of energy is fixed[26]. This statement does not immediately rule out the possibility that energy could disappear from a field in Canada while simultaneously appearing in a room in Indonesia. A stronger statement is that energy is locally conserved: Energy can neither be created nor destroyed, nor can it "teleport" from one place to another - it can only move by a continuous flow. A continuity equation is the mathematical way to express this kind of statement. For example, the continuity equation for electric charge states that the amount of electric charge at any point can only change by the amount of electric current flowing into or out of that point.

Continuity equations more generally can include "source" and "sink" terms, which allow them to describe quantities that are often but not always conserved, such as the density of a molecular species which can be generated or destroyed by chemical reactions. In an everyday example, there is a continuity equation for the number of people alive; it has a "source term" to account for people being born, and a "sink term" to account for people dying. Any continuity equation can be expressed in an "integral form" (in terms of a flux integral), which applies to any finite region, or in a "differential form" (in terms of the divergence operator) which applies at a point. Continuity equations underlie more specific transport equations such as the convection–diffusion equation, Boltzmann transport equation, and Navier–Stokes equations.

Now let us consider generally an open system, limited by any boundary Ω (deformable or not) and animated by any movement (moving or stationary) relative to a reference frame considered as fixed.

This system is shown in the figure below, is capable of transferring energy (or mass, or electric charges, etc.) between itself and the outside. This system may be inertial or not.

Given an extensive quantity A (like mass or charge). The corresponding quantitative magnitude will be denoted by a (it can express for example the isotropy or anisotropy of the system).

---

[26] It is strongly advice that the reader remembers the situation where this principle doesn't apply as seen during our study of Noether's theorem!

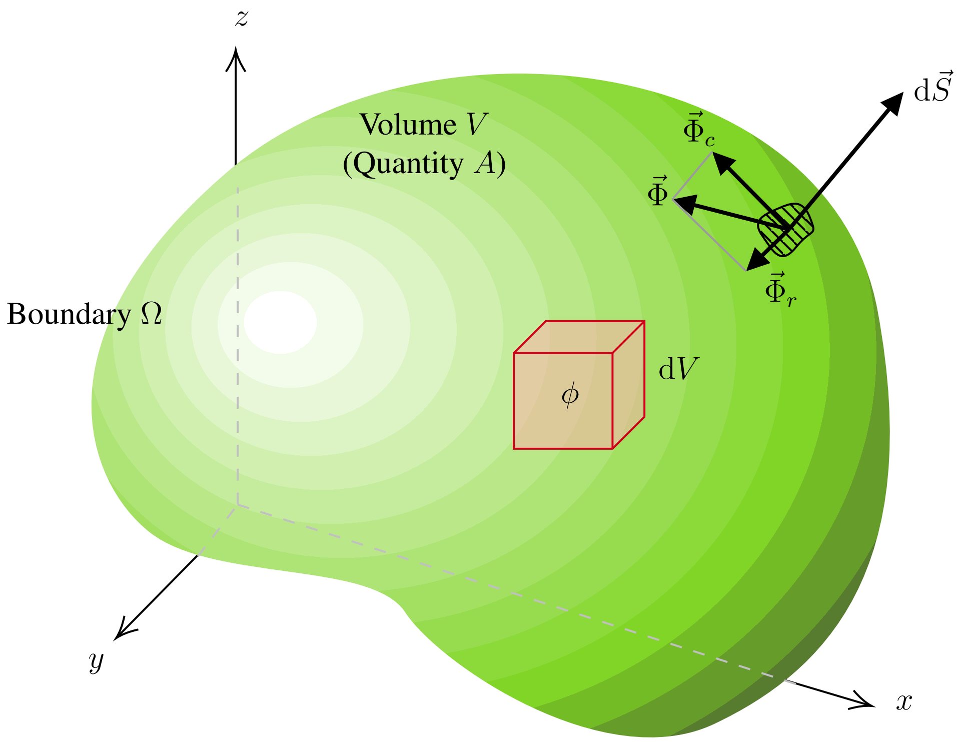


Figure 9.170: **Open system in translation or not in a repository and transferring energy**

In general, the value of A within the system is, at any instant:

$$A = \iiint_V \rho a \mathrm{d}V \qquad (9.6.352)$$

ρ being the density of an extensive quantity A.

The spatial variation rate of A is given by the derivative $\mathrm{d}A/\mathrm{d}t$. The causes of variations of A can be related to two different phenomena: the flows and the sources or sinks.

Counting positively what enters the system, the flow of A through the boundary Ω is given by the surface integral:

in which we define:

- $\vec{\Phi}$ as the total surfacic flux vector (or the "current density vector") relatively to A
- $\mathrm{d}\vec{S}$ as the boundary element, expressed by a vector normal to the surface and directed towards the outside

$$\iint_S \vec{\Phi} \circ \mathrm{d}S \qquad (9.6.353)$$

Let us notice that, at the contrary to the usual meaning in physics, the concept of flow Φ already contains here the derivation with respect to time. Furthermore, to simplify the text, the expression "surfacic vector flow" is reduced here the term "flow" in all that follows.

This flow can be split into multiple flows, according to the relation:

$$\vec{\Phi}=\vec{\Phi}_a-\vec{\Phi}_d+\vec{\Phi}_c=\vec{\Phi}_r+\vec{\Phi}_c \qquad (9.6.354)$$

where:

- The term $\vec{\Phi}_a$ is a flow by absolute translation, characterizing a flow from a fluid flow. We have then relation:

  $$\vec{\Phi}_a=\rho a\vec{v}_a \qquad (9.6.355)$$

  where $\vec{v}_a$ the absolute velocity of a fluid particle with respect to the fixed reference frame.

- The term $\vec{\Phi}$ is an apparent displacement flow, involved only when the boundary Ω moves. We have the relation:

  $$\vec{\Phi}_d=\rho a\vec{v}_d \qquad (9.6.356)$$

  where $\vec{v}_d$ is apparent speed of movement of a point on the boundary Ω, relatively to the fixed reference frame.

- The term $\vec{\Phi}_c$ is the total flow by conduction, characterizing a flow linked to a transfer phenomenon at the microscopic level, without fluid movement (eg heat conduction, electrical conduction, mechanical work).

Under the principle of composition of velocities (absolute speed is the sum of the relative speed and its apparent speed), we have:

$$\vec{v_r}=\vec{v}_a-\vec{v}_d \qquad (9.6.357)$$

and therefore:

$$\vec{\Phi}_r=\vec{\Phi}_a-\vec{\Phi}_d \qquad (9.6.358)$$

The corresponding flow of A is then given by (under the assumption that a is given in unit of our quantity of interest by unit mass):

$$\iint_{S_f} a\frac{\mathrm{d}M}{\mathrm{d}t}=\iint_{S_f}\vec{\Phi}_r\circ\mathrm{d}\vec{S} \qquad (9.6.359)$$

where S_f denotes the boundary portion through which the mass flow occurs.

If we count positively the effect of a source, the rate of increase of A is given by:

$$\iiint_V \phi\mathrm{d}V \qquad (9.6.360)$$

where ϕ is the volume flow from a source of A.

Taking into account both flows and sources, we have the spatial rate variation of A:

$$\frac{\mathrm{d}A}{\mathrm{d}t}=\iint_S a\mathrm{d}M+\iint_S\vec{\Phi}_c\circ\mathrm{d}\vec{S}+\iint_V\phi\mathrm{d}V \qquad (9.6.361)$$

The spatial balance of A is finally expressed by the relation: In the particular case of a system in continuous operation (eg in the case of a flowing fluid or a solid in which we can observe a thermal conduction, an electrical conduction, a nuclear reaction, etc.), all local variables are constant at any point of the system. If, furthermore, we choose a boundary Ω dimensionally stable, it is possible to reason with respect to a reference frame linked to the considered system. We then have, at any fixed point of the system relative to this reference frame:

$$\rho = c^{te} \qquad a = c^{te} \qquad V = c^{te} \tag{9.6.362}$$

It results for the whole system:

$$A = c^{te} \Leftrightarrow \frac{\mathrm{d}A}{\mathrm{d}t} = 0 \tag{9.6.363}$$

So in the case of a steady system, with a dimensionally stable boundary Ω, linked to the system, the spatial variation rate of any extensive scalar quantity is zero.

The spatial variation rate of A starting (for recall) from:

$$A = \iiint_V \rho a \mathrm{d}V \tag{9.6.364}$$

given by:

$$\Phi = \frac{\mathrm{d}A}{\mathrm{d}t} = \iiint_V \rho a \mathrm{d}V = \iiint_V \frac{\partial(\rho a)}{\partial t}\mathrm{d}V + \iiint_V \rho a \frac{\partial}{\partial t}(\mathrm{d}V) \tag{9.6.365}$$

The elementary variation in the volume V is due to the displacement (in the sense of the deformation!) of the boundary Ω, so that:

$$\partial(\mathrm{d}V) = \mathrm{d}(\mathrm{d}V) = (\mathrm{d}\vec{S} \circ \vec{v}_d)\mathrm{d}t \tag{9.6.366}$$

Therefore we can write that:

$$\iiint_V \rho a \frac{\partial}{\partial t}(\mathrm{d}V) = \iiint_V \rho a \frac{\partial}{\partial t}(\mathrm{d}\vec{S} \circ \vec{v}_d)\mathrm{d}t = \iiint_V \vec{\Phi}_d \circ \mathrm{d}\vec{S} \tag{9.6.367}$$

Hence:

$$\frac{\mathrm{d}A}{\mathrm{d}t} = \iiint_V \frac{\partial(\rho a)}{\partial t}\mathrm{d}V + \iiint_V \vec{\Phi}_d \circ \mathrm{d}\vec{S} \tag{9.6.368}$$

Considering the flow of sources and wells, we have for the first term:

$$\iiint_V \frac{\partial(\rho a)}{\partial t}\mathrm{d}V = \iint_S \vec{\Phi} \circ \mathrm{d}\vec{S} + \iiint_V \phi \mathrm{d}V = \iint_S (\vec{\Phi}_a + \vec{\Phi}_c) \circ \mathrm{d}\vec{S} + \iiint_V \phi \mathrm{d}V \tag{9.6.369}$$

The Gauss-Ostrogradsky theorem (see section Vector Calculus page 1882) will allow us to write the surface integral into a volume integral, and grouping all terms under the same integral sign, we get immediately:

$$\iiint_V \left[\frac{\partial(\rho a)}{\partial t} - \vec{\nabla} \circ (\vec{\Phi}_a + \vec{\Phi}_c) - \phi\right] \mathrm{d}V = 0 \tag{9.6.370}$$

As the limits of integration (boundary Ω) are arbitrary, the expression in brackets is identically zero such that we are led to write:

$$\frac{\partial(\rho a)}{\partial t}-\vec{\nabla}\circ(\vec{\Phi}_a+\vec{\Phi}_c)-\phi \tag{9.6.371}$$

Let us consider now the special case where the scalar extensive quantity is the mass M, then we have:

$$A=M=\iiint_V \rho \mathrm{d}V \tag{9.6.372}$$

with $a=1$.

Since the mass is not likely to be transferred by a conduction phenomenon (in a classic case (non-quantum)) we have $\vec{\Phi}_c$ that is zero (in many text books even when there is conduction we consider $\vec{\Phi}_c \ll \vec{\Phi}_a$...). Since the mass is conservative, there is no source or well so that ϕ is also zero.

We have therefore:

$$\frac{\partial(\rho a)}{\partial t}-\vec{\nabla}\circ(\vec{\Phi}_a+\vec{\Phi}_c)-\phi=\frac{\partial\rho}{\partial t}-\vec{\nabla}\circ(\rho\vec{v}_a)=0 \tag{9.6.373}$$

Since the mass is not likely to be transferred by a conduction phenomenon (in a classic case (non-quantum)) we have $\vec{\Phi}_c$ that is zero (in many text books even when there is conduction we consider $\vec{\Phi}_c \ll \vec{\Phi}_a$...). Since the mass is conservative, there is no source or well so that ϕ is also zero.

We have therefore:

$$\frac{\partial(\rho a)}{\partial t}-\vec{\nabla}\circ(\vec{\Phi}_a+\vec{\Phi}_c)-\phi=\frac{\partial\rho}{\partial t}-\vec{\nabla}\circ(\rho\vec{v}_a)=0 \tag{9.6.374}$$

The relation:

$$\boxed{\frac{\partial\rho}{\partial t}-\vec{\nabla}\circ(\rho\vec{v}_a)=0} \tag{9.6.375}$$

is named "continuity equation" or "(mass) conservation equation".

The "$-$" sign is here because we defined the incoming flow as positive. It is possible that in the literature as well as in this textbook, you find a "$+$" instead of that sign such that:

$$\frac{\partial\rho}{\partial t}-\vec{\nabla}\circ(\rho\vec{v}_a)=0 \tag{9.6.376}$$

Thus, a fluid element in a divergent flow field will experience a reduction in density!

There is another much more common form in which we can find the continuity equation. The reader will have noticed that the term $\rho\vec{v}_a$ has units of a massic surface current density $[\mathrm{kg}\cdot\mathrm{m}^{-2}\cdot\mathrm{s}^{-1}]$ which bring us in analogy with electronics (see section Electrokinetics page 4229) to write:

$$(\rho\vec{v}_a)=\vec{j} \tag{9.6.377}$$

Which leads to the following form of the continuity equation:

$$\boxed{\frac{\partial\rho}{\partial t}-\vec{\nabla}\circ\vec{j}=0} \tag{9.6.378}$$

9.6.11.1 Heat Equation

Let us apply this result to the diffusion of heat that as important applications in many fields of study (wave mechanics, Schrödinger equation in quantum physics, thermodynamics, fluid dynamics, Black & Scholes equation in finance, for resistance study in electrokinetics, electromagnetic wave propagation in a medium, in chemistry for some reactions, in nuclear neutronics, etc.) and there is considerable literature on the various solutions of this differential equation of the second order. In the section of Functional Analysis will prove how to solve this equation with the Fourier transform and the Laplace transform.

As for the conservation of mass equation, we can write for the heat in the absence of sources:

$$\boxed{\frac{\partial q}{\partial t}+\vec{\nabla}\circ\vec{\Phi}_q=0} \qquad (9.6.379)$$

where q is the amount of heat per unit volume (do not forget it otherwise we would have taken a capital Q!) and $\vec{\Phi}_q$ is the heat flow whose incoming quantity was defined as negative.

A temperature variation causing a variation in the quantity of heat is defined as a first approximation by the following physical law (this follows from the definition of the specific massic heat...):

$$\mathrm{d}q=\rho c_P\mathrm{d}T \qquad (9.6.380)$$

where ρ is the density of matter and c_P is the specific heat capacity. Or equivalently (since in Thermodynamics, as we have already mentioned it, lower case is reported to the mass):

$$\mathrm{d}Q=\rho C_P\mathrm{d}T \qquad (9.6.381)$$

So the heat equation is a parabolic partial differential equation that describes the distribution of heat (or variation in temperature) in a given region over time.

The heat flux being trivially induced by spatial temperature difference, we obtain the "Fourier law" which expresses the heat flow in proportion to the spatial temperature gradient:

$$\boxed{\vec{\Phi}_q=-\kappa\vec{\nabla}T} \qquad (9.6.382)$$

The "$-$" sign is simply due to the fact the heat flow goes from the warmer to the colder and κ is the "heat transmission coefficient" expressing the "thermal conductivity" of the material dependent on its atomic and structural macroscopic properties. In the 121st century (holocene calendar) where climate change is quite a big issue, the research for good insulating material (low thermal conductivity materials) to keep houses and buildings warm in Winter (in the purpose to decrease energy loss by radiation) is an important area of research.

Material	**Thermal Conductivity** $\left[\mathrm{W}\cdot\mathrm{m}^{-1}\cdot\mathrm{K}^{-1}\right]$
Glass wool	0.04
Rock wool	0.045
Concrete, light	0.42
Cellulose	0.039
PVC	0.19
Cement	1.01
Cement mortar	1.73
Granite	$1.7-4$
Oak tree	0.17
Marble stone	$2.08-2.94$

Table 9.30: **Some Thermal conductivities**

Or for some ceramics:

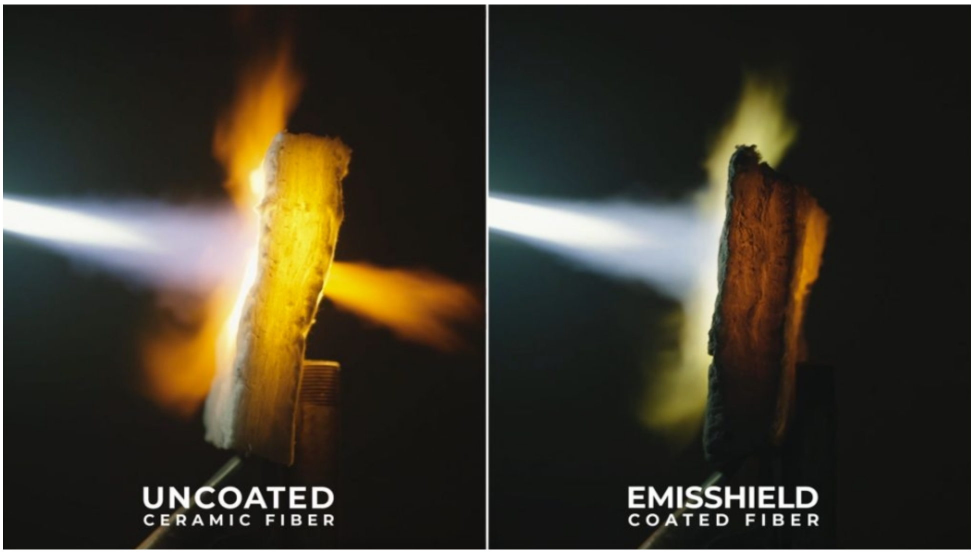
Material Name	**Melting Point** [°**C**]	**Density** [**g** $\cdot$ **cm**$^{-3}$]	**Specific Heat** [**J** $\cdot$ **kg**$^{-1}$ $\cdot$ °**C**$^{-1}$]	**Linear Coefficient of expansion** 10^{-6} [°**C**$^{-1}$]	**Thermal Conductivity** [**W** $\cdot$ **m**$^{-1}$ $\cdot$ °**C**$^{-1}$]
Alumina (Al_2O_3)	2050	3.96	1050	8.0	4@1315
Beryllia (BeO)	2550	3.0	2180	7.5	29@100
Magnesia (MgO)	2850	3.60	1170	13.5	59@1100
Thoria (ThO_2)	3220	9.75	290	9.5	3@1000
Zirconia (ZrO_2)	2700	5.7	590	7.5	3@1315
Zircon (ZrO_2SiO_2)	2500	4.6	630	4.5	4@1200
Spinel ($MgOAl_2O_3$)	2130	3.60	1050	8.5	2@1315
Mullite ($3Al_2O_32SiO_2$)	1850	2.8	840	5.0	4@1200
Sillimanite ($Al_2O_3SiO_2$)	1800	3.2	840	5.0	2@1300
Silicon Carbide (SiC)	2450	3.2	840	4.5	13@1100
Silicon Nitride (Si_3N_4)	1900	3.18	1050	2.6	9.5@1200
Carbon Graphite (C)	3600	2.2	1600	2.2	63@900
Quartizite (SiO_2)	1400	2.65	1170	8.6	2.6@1200
Boron Carbide (B_4C)	2350	2.5	2090	5.7	17.3@800
Boron Nitride (BN)	2721	2.1	1570	7.5	26@900
Titanium Carbide (TiC)	3140	6.5	1050	6.9	40@1100
Tungsten Carbide (WC)	2780	14.3	300	6.3	43.3@1100
Hafnium Carbide (HfC)	3958	12.2	200	6.73	27@800

Table 9.31: **Thermal conductivities and specific heat of some ceramics**

As we can see above, Ceramic is incredibly resistant, and it can withstand temperatures of $4,000$ degrees Celsius. Ceramic coating (like Emmishield™ Wessex's PCCM[27] product) that can withstand up to $2,000$ degrees Celsius is often used on spatial applications to increase the resistance to heat.

---

[27] protective coating for ceramic materials



Remark

AVCOAT 5026-39 is a NASA code (see [386]) for a specific ablative heat shield material developed by Avco (acquired by Textron in 1984). It is an epoxy novolac resin with special additives in a fiberglass honeycomb matrix. In fabrication, the empty honeycomb is bonded to the primary structure and the resin is gunned into each cell individually. AVCOAT was used for the heat shield on NASA's Apollo command module (which landed astronauts on the Moon between 11969 and 11972 according to holocene calendar). It has a density of $0.51\;[\text{g}\cdot\text{cm}^{-3}]$, it has a thermal conductivity of $0.16442\;[\text{W}\cdot\text{m}^{-1}]$ at $400\;[\text{K}]$, a melting point of $2,760\;[\text{K}]$ and a linear coefficient of expansion of $0.16442\cdot 10^{-2}\;[^\circ\text{C}^{-1}]$ at $400\;[\text{K}]$.

We find more often in practice this relation in the following form in the one-dimensional case (we will do a small practical example further with this form of writing):

$$q=\Phi_q S=-\kappa S\frac{\mathrm{d}T}{\mathrm{d}x}\cong-\kappa\frac{\Delta T}{\Delta x}\qquad(9.6.383)$$

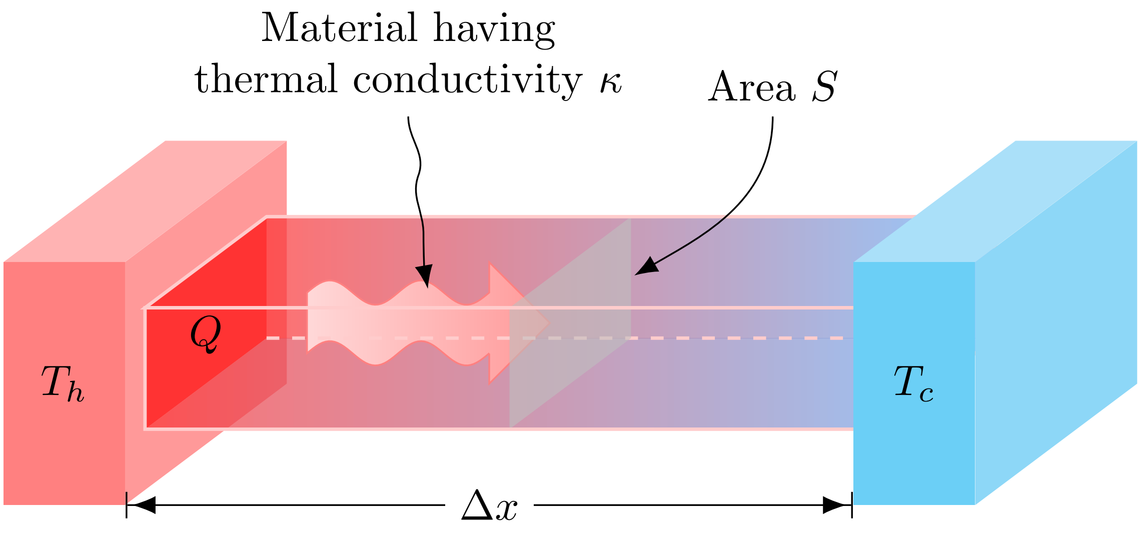


Figure 9.171: Heat conduction occurs through any material, represented here by a rectangular bar

By inserting the two prior previous relations in the heat conservation equation, we have:

$$\frac{\rho C_P \mathrm{d}T}{\partial t}+\vec{\nabla}\circ(-\kappa\vec{\nabla}T)=0 \tag{9.6.384}$$

after rearranging a little bit:

$$\frac{\partial T}{\partial t}-\frac{\kappa}{\rho c_P}\vec{\nabla}\circ(\vec{\nabla}T)=0 \tag{9.6.385}$$

More aesthetically and more generally (but still in one dimension), we find it in the condensed form of the "heat diffusion equation" or more simply named "heat equation":

$$\boxed{\frac{\partial T}{\partial t}-D_T\frac{\partial^2 T}{\partial x^2}=0} \tag{9.6.386}$$

where the coefficient of proportionality is named in the context of heat study: "thermal diffusivity":

$$D_T=\frac{\kappa}{c_P\rho} \tag{9.6.387}$$

It is possible to prove its microscopic origin as we will prove it further below.

We must always pay attention to the units of κ following that we work the with the massic specific heat capacity c_P or the heat capacity C_P in the denominator!

So under completely explicit form, we have in one dimension:

$$\frac{\partial T(x,t)}{\partial t}-\frac{\kappa}{\rho c_P}\frac{\partial^2 T(x,t)}{\partial x^2}=0 \tag{9.6.388}$$

Remarks

R1. We will see again this equation in the section of Theoretical Computing to introduce the reader to the concept of solving differential equations using the finite element method with a simple example of resolution using Microsoft® Excel® or in the MATLAB® companion book.

R2. It is by studying this equation that Fourier introduced the series and transformation that bear his name, and which became so important in the study of phenomena of propagation / diffusion.

R3. Often practitioners write:

$$q = -\kappa S\frac{\Delta T}{\Delta x} = -\frac{\Delta T}{\dfrac{\Delta x}{\kappa S}} = -\frac{\Delta T}{R_{\text{th.}}} \qquad (9.6.389)$$

and thus we can introduce the concept of "thermal resistance $R_{\text{th.}}$" in the denominator. In short nothing spectacular but of frequent use.

Let us insist that all relations of the type:

$$\frac{\partial u}{\partial t} - D\frac{\partial^2 u}{\partial x^2} = 0 \qquad (9.6.390)$$

are named "diffusion equations" of physical parameter D and is a parabolic partial differential equation (see section Differential and Integral Calculus page 1625). We will immediately see how to solve this differential partial equation using and this will lead us to understand why Fourier introduced the famous transformations that bear today his name. But let us remind the reader that we find this partial differential equation in multiple contexts.

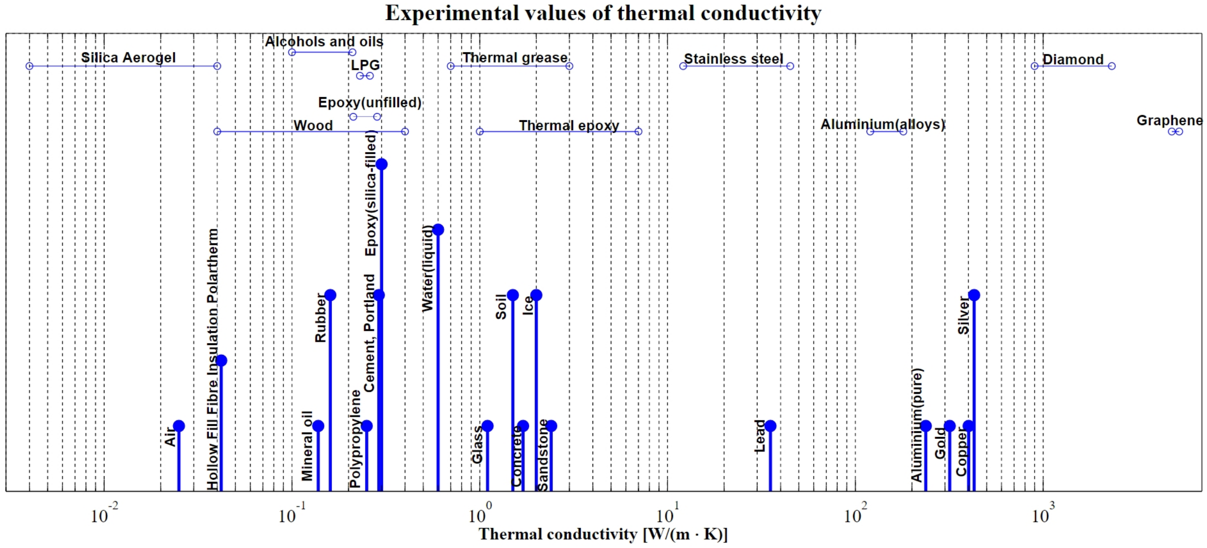


Figure 9.172: Thermal conductivity values for various materials (source: Wikipedia)

✍**Examples:**

E1. A parallelepiped made of steel (having a thermal conductivity of 17 $[\mathrm{W}\cdot\mathrm{m}^{-1}\cdot\mathrm{K}^{-1}]$) has one side thereof maintained at 331.15 $[\mathrm{K}]$ and the other at 297.15 $[\mathrm{K}]$, both faces being separated by 1 $[\mathrm{cm}]$ and having a surface 1 $[\mathrm{m}^2]$: Assuming a situation at equilibrium the heat transfer is given by:

$$q=-\kappa S\frac{\Delta T}{\Delta x}=-17.1\left(\frac{331.15-297.15}{0-0.01}\right)=34\,[\mathrm{W}] \tag{9.6.391}$$

E2. For a cylindrical tube carrying a heated fluid (very important case in practice) we want to determine the amount of heat exchanged per unit of time with the outside environment. Knowing that the surface in contact with the fluid (with inner radius r_i) is not the same that the one in contact with the outside environment (external radius r_e) we can not simply calculate ratio of the temperature differences with the distance between the two surfaces. Therefore we have to proceed as following:

$$\begin{aligned}q&=-\kappa S\frac{\mathrm{d}T}{\mathrm{d}r}=-\kappa 2\pi r\frac{\Delta T}{\mathrm{d}r}=-\kappa 2\pi\frac{\Delta T}{\dfrac{1}{r}\mathrm{d}r}=-\kappa 2\pi\frac{\Delta T}{\dfrac{1}{r}\mathrm{d}r}=-\kappa 2\pi\frac{T_e-T_i}{\displaystyle\int_{r_i}^{r_e}\frac{1}{r}\mathrm{d}r}\\&=\kappa 2\pi\frac{T_i-T_e}{\ln(r)|_{r_i}^{r_e}}=\kappa 2\pi\frac{T_i-T_e}{\ln(r_e)-\ln(r_i)}=\kappa 2\pi\frac{T_i-T_e}{\ln\left(\dfrac{r_e}{r_i}\right)}\end{aligned} \tag{9.6.392}$$

Let us solve the general form of the diffusion equation without sources:

$$\frac{\partial u}{\partial t}=D\frac{\partial^2 u}{\partial x^2} \tag{9.6.393}$$

To solve this differential equation of the second order, we will use the method of separation of variables (see section Differential and Integral Calculus page 1617) rather than attacking directly by Fourier transforms (which requires as prior the proof of the dominated convergence theorem which makes me horror).

We assume therefore using the separation of variables method that:

$$u(x,t)=X(x)T(t) \tag{9.6.394}$$

And where $T(t)$ by the principle of the second law of thermodynamics must decrease as t increases.

We get immediately:

$$\begin{aligned}\frac{\partial u}{\partial t}&=\frac{\partial X(x)\partial T(t)}{\partial t}=\frac{\partial X(x)}{\partial t}T(t)+X(x)\frac{\partial T(t)}{\partial t}=X\frac{\partial T}{\partial t}\\\frac{\partial^2 u}{\partial x^2}&=\frac{\partial^2 X(x)\partial T(t)}{\partial x^2}=\frac{\partial^2 X(x)}{\partial x^2}T(t)+X(x)\frac{\partial^2 T(t)}{\partial x^2}=T\frac{\partial^2 X}{\partial x^2}\end{aligned} \tag{9.6.395}$$

hence the diffusion equation:

$$\frac{\partial u}{\partial t}=D\frac{\partial^2 u}{\partial x^2}\Rightarrow X\frac{\partial T}{\partial t}=DT\frac{\partial^2 X}{\partial x^2} \tag{9.6.396}$$

what rearranged and condensed is also written:

$$\frac{1}{D}\frac{T'}{T}=\frac{X''}{X} \tag{9.6.397}$$

Which can be written by assuming that each of the expressions on the left and right of equality are assimilated to functions:

$$G(t)=F(x) \tag{9.6.398}$$

so that equality is true for all t and x, the functions G and F functions must be constant. So we have the right to write:

$$\frac{1}{D}\frac{T'}{T}=\frac{X''}{X}=c^{te}=-\lambda^2 \tag{9.6.399}$$

The fact to write the constant negative and squared is a simple anticipation of the result historically already known... But to understand why the constant is necessarily negative, just think that T' is necessarily negative (the temperature of a isolated system will not rise alone but naturally decrease based on the principle of entropy) and as D and T are positive then verbatim ...

What gives us the system of two independent differential equations of the second order:

$$\begin{aligned}X''+\lambda^2X&=0\\T'+\lambda^2DT&=0\end{aligned} \tag{9.6.400}$$

That some text books write as following:

$$\begin{aligned}X_{xx}+\lambda^2X&=0\\T_t+\lambda^2DT&=0\end{aligned} \tag{9.6.401}$$

We solve the second differential equation first:

$$\frac{1}{T(t)}\frac{\partial T(t)}{\partial t}=-\lambda^2D\Rightarrow \ln(T(t))=-\lambda^2Dt+c^{te} \tag{9.6.402}$$

Therefore:

$$T(t)=e^{-\lambda^2Dt+c^{te}}=e^{c^{te}}e^{-\lambda^2Dt}=A(\lambda)e^{-\lambda^2Dt} \tag{9.6.403}$$

And for the first differential equation:

$$\frac{\partial X(x)^2}{\partial x^2}+\lambda^2X(x)=0 \tag{9.6.404}$$

We then have the characteristic polynomial:

$$r^2+\lambda^2=0 \tag{9.6.405}$$

Thus the roots:

$$r_{1,2}=\pm\sqrt{-\lambda^2}=\pm\mathrm{i}\lambda \tag{9.6.406}$$

So as the discriminant is negative (see section Differential and Integral Calculus page 1611) we have:

$$X(x)=A'(\lambda)e^{\mathrm{i}(\lambda x+c^{te})}+B'(\lambda)e^{-\mathrm{i}(\lambda x+c^{te})} \tag{9.6.407}$$

Therefore:

$$\begin{aligned}u(x,t)&=T(t)X(x)=A(\lambda)e^{-\lambda^2Dt}\left(A'(\lambda)e^{\mathrm{i}(\lambda x+c^{te})}+B'(\lambda)e^{-\mathrm{i}(\lambda x+c^{te})}\right)\\&=\left(C_1(\lambda)e^{\mathrm{i}(\lambda x+c^{te})}+C_2(\lambda)e^{-\mathrm{i}(\lambda x+c^{te})}\right)e^{-\lambda^2Dt}\end{aligned} \tag{9.6.408}$$

If we move the exponential of the constants in factors such as:

$$\begin{aligned}C_1(\lambda)e^{\mathrm{i}c^{te}}&={C'}_1(\lambda)\\C_2(\lambda)e^{-\mathrm{i}c^{te}}&={C'}_2(\lambda)\end{aligned} \tag{9.6.409}$$

this simplifies the expression:

$$u(x,t)=\left({C'}_1(\lambda)e^{\mathrm{i}\lambda x}+{C'}_2(\lambda)e^{-\mathrm{i}\lambda x}\right)e^{-\lambda^2Dt} \tag{9.6.410}$$

whose coefficients will be determined case by case by the initial conditions of which the most famous school cases and subject to an extensive literature are:

- Dirichlet boundaries conditions:

$$x=(0,L)\qquad u(0,t)=0=u(L,t) \tag{9.6.411}$$

- Neumann boundaries conditions:

$$x=(0,L)\qquad \frac{\mathrm{d}u}{\mathrm{d}x}(0,t)=0=\frac{\mathrm{d}u}{\mathrm{d}x}(L,t) \tag{9.6.412}$$

Since for each possible value of λ we get a solution, it appears that doing the sum of all these solutions, we get the general solution. We therefore finite Fourier series:

$$u(x,t)=\sum_{\lambda}\left({C'}_1e^{\mathrm{i}\lambda x}+{C'}_2(\lambda)e^{-\mathrm{i}\lambda x}\right)e^{-\lambda^2Dt} \tag{9.6.413}$$

because of the presence of the square λ in the temporal exponential in factor (right) of the parenthesis, if we sum over all the λ from $-\infty,+\infty$ then the sum of all terms of the parenthesis cancel for each pair of positive and negative λ of the same value. Then we must summon only on the semi-positive infinite line (either left or right it does not matter) if we do not want the result to be zero. If we choose arbitrarily the positive infinite line, then we get:

$$u(x,t)=\sum_{\lambda=0}^{+\infty}\left({C'}_1(\lambda)e^{\mathrm{i}\lambda x}+{C'}_2(\lambda)e^{-\mathrm{i}\lambda x}\right)e^{-\lambda^2Dt} \tag{9.6.414}$$

Nothing now prevents us from taking out a factor λ of the constant $C'_1(\lambda)$, $C'_2(\lambda)$, the only effect it will have is to change their normalisation given by the initial conditions .. Then we have:

$$\begin{aligned} u(x,t) &= \sum_{\lambda=0}^{+\infty} \left(\lambda C''_1(\lambda)e^{\mathrm{i}\lambda x} + \lambda C''_2(\lambda)e^{-\mathrm{i}\lambda x}\right) e^{-\lambda^2 Dt} \\ &= \sum_{\lambda=0}^{+\infty} \left(C''_1(\lambda)e^{\mathrm{i}\lambda x} + C''_2(\lambda)e^{-\mathrm{i}\lambda x}\right) e^{-\lambda^2 Dt}\lambda \end{aligned} \tag{9.6.415}$$

As λ is a continuous real parameter (and also positive) in by making it tend to zero the sum can be changed into an integral (doing this in the physicist way...) and then we get:

$$u(x,t) = \int\limits_0^{+\infty} C''_1(\lambda)e^{-\lambda^2 Dt+\mathrm{i}\lambda x}\mathrm{d}\lambda + \int\limits_0^{+\infty} C''_2(\lambda)e^{-\lambda^2 Dt-\mathrm{i}\lambda x}\mathrm{d}\lambda \tag{9.6.416}$$

And making a change of variable for the second integral, we have:

$$\begin{aligned} u(x,t) &= \int\limits_0^{+\infty} C''_1(\lambda)e^{-\lambda^2 Dt+\mathrm{i}(\lambda x+c^{te})}\mathrm{d}\lambda - \int\limits_0^{-\infty} C''_2(-\lambda)e^{-\lambda^2 Dt-\mathrm{i}(-\lambda x+c^{te})}\mathrm{d}t \\ &= \int\limits_0^{+\infty} C''_1(\lambda)e^{-\lambda^2 Dt+\mathrm{i}(\lambda x+c^{te})}\mathrm{d}\lambda + \int\limits_{-\infty}^{0} C''_2(-\lambda)e^{-\lambda^2 Dt+\mathrm{i}(\lambda x-c^{te}}\mathrm{d}\lambda \end{aligned} \tag{9.6.417}$$

By grouping the constants, we have:

$$\boxed{u(x,t) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} F(\lambda)e^{-\lambda^2 Dt+\mathrm{i}\lambda x}\mathrm{d}\lambda} \tag{9.6.418}$$

with:

$$\begin{aligned} F(\lambda) &= \sqrt{2\pi}C''_1(\lambda) \quad \text{for } \lambda > 0 \\ F(\lambda) &= \sqrt{2\pi}C''_2(\lambda) \quad \text{for } \lambda < 0 \end{aligned} \tag{9.6.419}$$

We have now something interesting! At the time $t=0$, the previous integral is written:

$$u(x,t) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} F(\lambda)e^{\mathrm{i}\lambda x}\mathrm{d}\lambda \tag{9.6.420}$$

and changing the notation for the parameter λ and the notation for x, we have:

$$u(x,t) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} F(\omega)e^{\mathrm{i}\omega x}\mathrm{d}\lambda \tag{9.6.421}$$

And so what???? are you perhaps thinking just now...? Well as we saw in our study of the Fourier transforms (see section Sequences and Series page 1758) the latter relation is an inverse Fourier transform which is an infinite sum of real trigonometric functions! This means that any function describing the temperature distribution at time t zero for an infinitely long bar and whose ends at the same time tend to zero (thus implicitly it is a function of infinite period) such that:

$$\begin{gathered}u(x,0)=f(x)\\ f(\pm\infty)\rightarrow 0\end{gathered}\tag{9.6.422}$$

Satisfies the equation of Heat as an infinite trigonometric series. Let us return to our initial notation:

$$u(x,t)=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}F(\lambda)e^{-\lambda^2Dt+\mathrm{i}\lambda x}\mathrm{d}\lambda\tag{9.6.423}$$

As a recall of the section Sequences and Series (with the notation adapted here) we have by applying the inverse Fourier transform:

$$u(x,0)=\mathcal{F}^{-1}(F(\lambda))(x)=f(x)=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}F(\lambda)e^{-\mathrm{i}\lambda x}\mathrm{d}x\tag{9.6.424}$$

Therefore $F(\lambda)$ is the Fourier transform of the desired function such that:

$$F(\lambda)=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}f(x)e^{-\mathrm{i}\lambda x}\mathrm{d}x\tag{9.6.425}$$

Which then gives by injecting the Fourier transform with a change of variable that imposes itself:

$$\begin{aligned}u(x,t)&=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}F(\lambda)e^{-\lambda^2Dt+\mathrm{i}\lambda x}\mathrm{d}\lambda=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}\left(\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}f(v)e^{-\mathrm{i}\lambda v}\mathrm{d}v\right)e^{-\lambda^2Dt+\mathrm{i}\lambda x}\mathrm{d}\lambda\\&=\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty}\left(\int\limits_{-\infty}^{+\infty}f(v)e^{-\mathrm{i}\lambda v}\mathrm{d}v\right)e^{-\lambda^2Dt+\mathrm{i}\lambda x}\mathrm{d}\lambda\end{aligned}\tag{9.6.426}$$

Always keeping in mind that if we put t as being null below, we then fall back well on:

$$u(x,0)=f(x)\tag{9.6.427}$$

Since in the last equality the term $e^{-\lambda^2Dt+\mathrm{i}\lambda x}$ is independent of the variable v, we can put it in the second integral such that:

$$u(x,t)=\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty}\left(\int\limits_{-\infty}^{+\infty}f(v)e^{-\mathrm{i}\lambda v\mathrm{d}v}\right)e^{-\lambda^2Dt+\mathrm{i}\lambda x}\mathrm{d}\lambda=\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty}f(v)e^{-\mathrm{i}\lambda v}e^{-\lambda^2Dt+\mathrm{i}\lambda x}\mathrm{d}v\mathrm{d}\lambda\tag{9.6.428}$$

We then recognize a double integral of which we can change the order of integration by applying the Fubini theorem (see section Differential and Integral Calculus page 1521):

$$\begin{aligned}u(x,t)&=\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty}f(v)e^{-\mathrm{i}\lambda v}e^{-\lambda^2Dt+\mathrm{i}\lambda x}\mathrm{d}\lambda\mathrm{d}v=\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty}f(v)\int\limits_{-\infty}^{+\infty}e^{-\mathrm{i}\lambda v}e^{-\lambda^2Dt+\mathrm{i}\lambda x}\mathrm{d}\lambda\mathrm{d}v\\&=\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty}f(v)\int\limits_{-\infty}^{+\infty}e^{-\mathrm{i}\lambda v}e^{-\lambda^2Dt-\mathrm{i}\lambda(v-x)}\mathrm{d}\lambda\mathrm{d}v\end{aligned}\tag{9.6.429}$$

It then comes in the case that interests us by making a correspondence term by term:

$$\begin{aligned}u(x,t)&=\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty}f(v)\int\limits_{-\infty}^{+\infty}e^{-Dt\lambda^2-\mathrm{i}(v-x)\lambda}\mathrm{d}\lambda\mathrm{d}v=\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty}f(v)\left(\sqrt{2\pi}\frac{1}{\sqrt{2Dt}}e^{-\frac{(v-x)^2}{4Dt}}\right)\mathrm{d}v\\&=\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty}f(v)\frac{\sqrt{\pi}}{\sqrt{Dt}}e^{-\frac{(v-x)^2}{4Dt}}\mathrm{d}v\end{aligned}\tag{9.6.430}$$

Therefore:

$$\boxed{u(x,t)=\frac{1}{2\sqrt{\pi Dt}}\int\limits_{-\infty}^{+\infty}f(v)e^{-\frac{(v-x)^2}{4Dt}}\mathrm{d}v}\tag{9.6.431}$$

Thus any function $f(v)$ injected in the above relation will satisfy the initial conditions we have imposed to ourselves:

$$u(x,0)=f(x)f(\pm\infty)\rightarrow 0\tag{9.6.432}$$

Remark

The above solution is sometimes named "Gauss-Weierstrass integral form of the solution" as it is similar to the definition of the "Weierstrass transform" defined by:

$$W[f](x)=\frac{1}{\sqrt{4\pi}}\int\limits_{-\infty}^{+\infty}f(y)e^{-\frac{(x-y)^2}{4}}\mathrm{d}y=\frac{1}{\sqrt{4\pi}}\int\limits_{-\infty}^{+\infty}f(x-y)e^{-\frac{y^2}{4}}\mathrm{d}y\tag{9.6.433}$$

The case with the heat equation having $Dt=1$ is named the "standard Weierstrass transform". By varying Dt (especially by putting $D=1$ and making t varying) we speak of "generalized Weierstrass transform".

For information, many physicists name the following expression visible in the previous boxed relation:

$$K(t,x,v)=\frac{1}{\sqrt{4\pi t}}e^{-\frac{(v-x)^2}{4t}}\tag{9.6.434}$$

The "heat kernel". Thus the final result is in reality only a convolution of $f(v)$ with this "kernel".

In practice this integral is not analytically calculable since it is a convolution with a Gaussian (see section Statistics page 692). This is the reasons for which the heat equation is almost always solved at the initial conditions mentioned above using numerical methods (see section on Numerical Methods page 6753).

But let us solve some simplistic but real cases:

✍**Examples:**

E1. When two ends of a system of size L are maintained at two different temperatures T_1 and T_2, the solution of the heat equation is stationary (independent of time). We then have:

$$\frac{\partial^2 T(x)}{\partial x^2}=0 \tag{9.6.435}$$

The solution to this differential equation is very simple and does not require the use of the previous result (it is elementary integral calculus with known initial conditions):

$$T(x)=T_1+(T_2-T_1)\frac{x}{L} \tag{9.6.436}$$

This is a situation we find in everyday life ...

E2. Let us consider the case where we put in contact two bars of infinite length (since with the Fourier transform it is always necessary that one of the variables range the whole set of real number) and of opposite temperature in sign (on the Celsius scale therefore...) such as:

$$u(x,0)=\begin{cases}+T_0 & \text{if } x>0\\ -T_0 & \text{if } x<0\end{cases} \tag{9.6.437}$$

We then have:

$$\begin{aligned}u(x,t)&=\frac{1}{2\sqrt{xDt}}\int_{-\infty}^{+\infty} f(v)e^{-\frac{(v-x)^2}{4Dt}}\mathrm{d}v\\&=\frac{1}{2\sqrt{\pi Dt}}\left(\int_{-\infty}^{0} -T_0 e^{-\frac{(v-x)^2}{4Dt}}\mathrm{d}v+\int_{0}^{+\infty} T_0 e^{-\frac{(v-x)^2}{4Dt}}\mathrm{d}v\right)\\&=\frac{T_0}{2\sqrt{\pi Dt}}\left(\int_{0}^{+\infty} e^{-\frac{(v-x)^2}{4Dt}}\mathrm{d}v-\int_{-\infty}^{0} e^{-\frac{(v-x)^2}{4Dt}}\mathrm{d}v\right)\end{aligned} \tag{9.6.438}$$

And let us make the following change of variable:

$$\frac{v-x}{\sqrt{4Dt}}=\frac{v-x}{2\sqrt{Dt}}=\tau \tag{9.6.439}$$

Hence:

$$\mathrm{d}v=-2\sqrt{Dt}\mathrm{d}\tau \tag{9.6.440}$$

Then we have:

$$\begin{aligned}u(x,t)&=\frac{T_0}{2\sqrt{\pi Dt}}\left(\int\limits_{-x/2\sqrt{Dt}}^{+\infty}-2\sqrt{Dt}e^{-\tau^2}\mathrm{d}\tau-\int\limits_{-\infty}^{-x/2\sqrt{Dt}}-2\sqrt{Dt}e^{-\tau^2}\mathrm{d}\tau\right)\\&=\frac{T_0}{\sqrt{\pi}}\left(\int\limits_{-x/2\sqrt{Dt}}^{+\infty}-e^{-\tau^2}\mathrm{d}\tau-\int\limits_{-\infty}^{-x/2\sqrt{2Dt}}-e^{-\tau^2}\mathrm{d}\tau\right)\\&=\frac{T_0}{\sqrt{\pi}}\left(\int\limits_{-x/2\sqrt{Dt}}^{+\infty}-e^{-\tau^2}\mathrm{d}\tau+\int\limits_{-\infty}^{-x/2\sqrt{2Dt}}e^{-\tau^2}\mathrm{d}\tau\right)\\&=\frac{T_0}{\sqrt{\pi}}\left(\int\limits_{-\infty}^{-x/2\sqrt{Dt}}e^{-\tau^2}\mathrm{d}\tau-\int\limits_{-x/2\sqrt{Dt}}^{+\infty}e^{-\tau^2}\mathrm{d}\tau\right)\end{aligned} \tag{9.6.441}$$

Since the integrated function is an even function (see section Functional Analysis page 2212) we have:

$$\int\limits_{-\infty}^{-x/2\sqrt{Dt}}e^{-\tau^2}\mathrm{d}\tau=\int\limits_{x/2\sqrt{Dt}}e^{-\tau^2}\mathrm{d}\tau \tag{9.6.442}$$

Therefore:

$$\begin{aligned}u(x,t)&=\frac{T_0}{\sqrt{\pi}}\left(\int\limits_{-x/2\sqrt{Dt}}^{+\infty}e^{-\tau^2}\mathrm{d}\tau-\int\limits_{-x/2\sqrt{Dt}}^{+\infty}e^{-\tau^2}\mathrm{d}\tau\right)=\frac{T_0}{\sqrt{\pi}}\left(\int\limits_{-x/2\sqrt{Dt}}^{x/2\sqrt{Dt}}e^{-\tau^2}\mathrm{d}\tau\right)\\&=\frac{2T_0}{\sqrt{\pi}}\int\limits_{0}^{x/2\sqrt{Dt}}e^{-\tau^2}\mathrm{d}\tau\end{aligned} \tag{9.6.443}$$

What it is customary to write:

$$u(x,t)=T_0\mathrm{erf}\left(\frac{x}{2\sqrt{Dt}}\right) \tag{9.6.444}$$

Where "erf" is named the "Gauss error function" and can not be analytically computed accurately (we must go through limited series developments). It can however be found in spreadsheets like Microsoft® Excel® under the name `ERF( )`.

9.6.11.1.1 Fick's laws of diffusion

Equations based on Fick's law have been commonly used to model transport processes in foods, neurons, biopolymers, pharmaceuticals, porous soils, population dynamics, nuclear materials, semiconductor doping process, etc. Theory of all voltammetric methods is based on solutions of Fick's equation. A large amount of experimental research in polymer science and food science has shown that a more general approach is required to describe transport of components in materials undergoing glass transition. In the vicinity of glass transition the flow behaviour becomes "non-Fickian".

When two miscible liquids are brought into contact, and diffusion takes place, the macroscopic (or average) concentration evolves following Fick's law. On a mesoscopic scale, that is, between the macroscopic scale described by Fick's law and molecular scale, where molecular random walks take place, fluctuations cannot be neglected. Such situations can be successfully modelled with Landau-Lifshitz fluctuating hydrodynamics. In this theoretical framework, diffusion is due to fluctuations whose dimensions range from the molecular scale to the macroscopic scale.

Integrated circuit fabrication technologies, model processes like CVD (chemical vapor deposition processes), thermal oxidation, wet oxidation, doping, etc. use diffusion equations obtained from Fick's law.

We have just seen the proof of the heat propagation equation proposed by Fourier in 11822 (holocene calendar) obtained from the equation of continuity. We had obtained (it is highly recommended to the reader to refer to it again only to read a second time the remarks relating to the proof):

$$\frac{\partial T(x,t)}{\partial t}-D_T\frac{\partial^2 T(x,t)}{\partial x^2}=0 \tag{9.6.445}$$

Based on the same hypotheses as Fourier, Adolf Fick proposed in 11855 (holocene calendar) that a flow of particles could diffuse through a material according to a similar law, the "Fick's second law" or "Fick's law of diffusion", of the form:

$$\boxed{\frac{\partial \rho(x,t)}{\partial t}-D_p\frac{\partial^2 \rho(x,t)}{\partial x^2}=0} \tag{9.6.446}$$

Where the constant of proportionality D_p is the "diffusion coefficient of matter" and ρ the density of particles per unit volume (and not the mass density!).

Remark

In practice, diffusion plays an essential role in the manufacture of ceramics, semiconductors (doping), solar cells and in the solidification of metals (carbon and heat treatment). Because since two heated materials are taken into contact, their atoms diffuse into each other (also a little bit when not heated).

It must be understood that everything diffuses in everything as we have already mention it in our study of Tribology in the section of Classical Mechanics! So think about pesticides on fruits and vegetables, pollution in ground water, PET in drinks, etc.

Therefore, the relation of the heat surfacic flow of that we have used above to obtain the Fourier law and which was:

$$\vec{\Phi}_q = -\kappa \vec{\nabla} T \tag{9.6.447}$$

can then also be written (we will prove it) in the case of the mass in the form of a flow of particles named the "first Fick's law":

$$\boxed{\vec{\Phi}_p = -D \vec{\nabla} \rho} \tag{9.6.448}$$

where D is the transport coefficient of the material (to be determined ...). We find in the literature this last relation frequently in unidimensional form as following:

$$\Phi_p = -D \frac{\partial \rho}{\partial x} \tag{9.6.449}$$

Either in discrete form:

$$\Phi_p = -D \frac{\Delta \rho}{\Delta x} \tag{9.6.450}$$

If the variation of the distance is put in correspondence with a length L travelled by the diffusion, then we have the notation even more simplified:

$$\Phi_p = -D \frac{\Delta \rho}{L} \tag{9.6.451}$$

Below we can see the gradient density between salt water and fresh water (the transition zone known since more than 2000 years is named "halocline"):

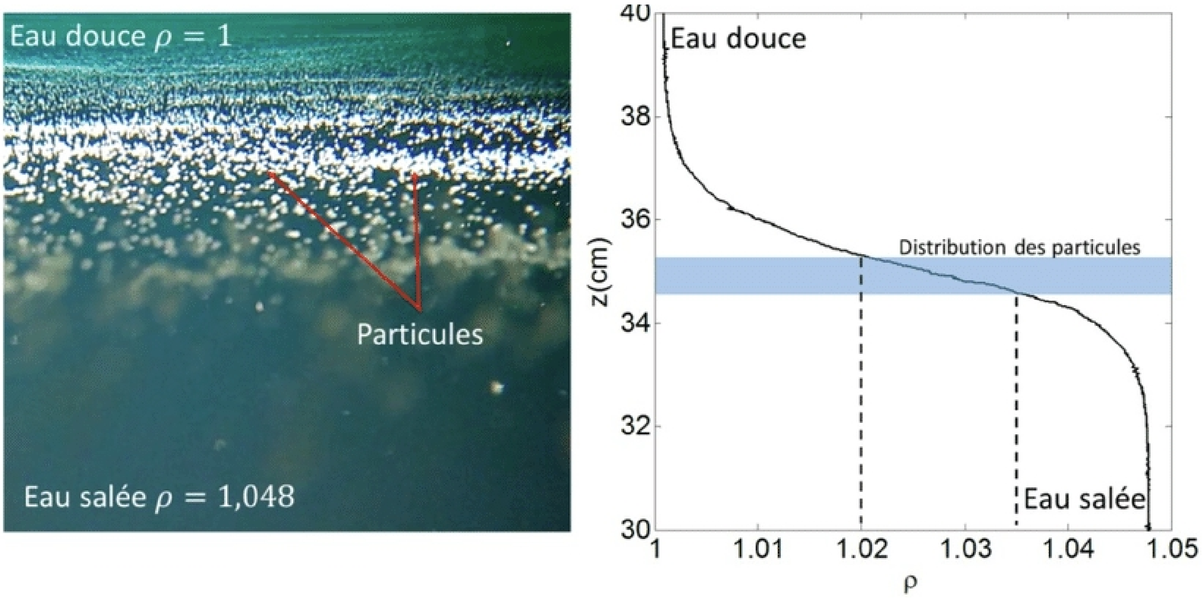
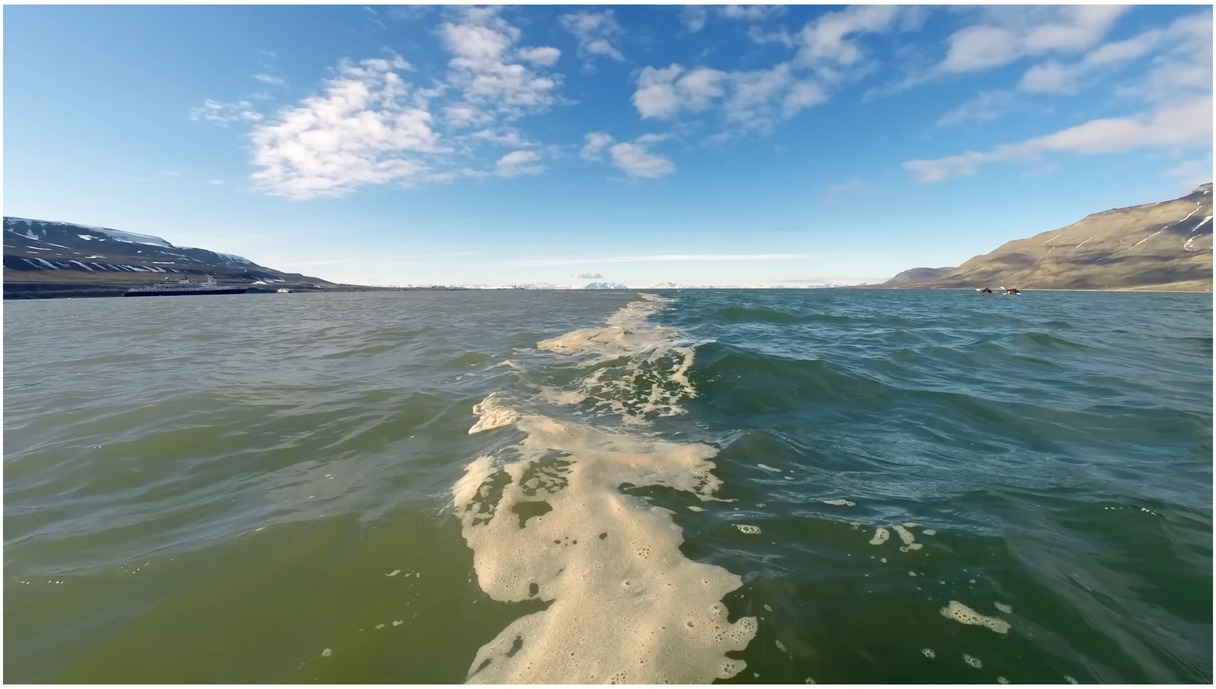
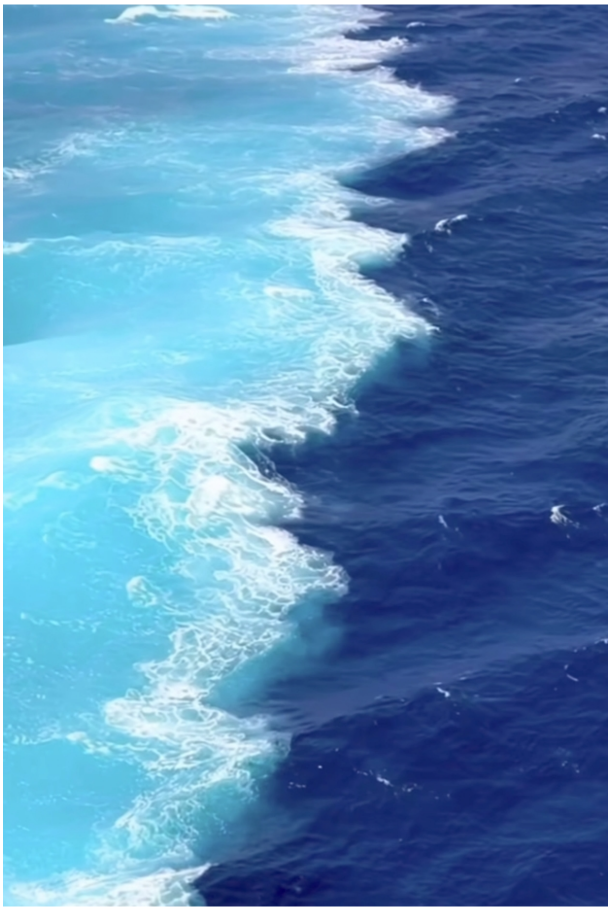


Figure 9.173: **Corresponding density profile for the transition between salt water having a density of $\rho_2 = 1.048$ and fresh water $\rho_1 = 1$ that takes place over approximately $5[\text{cm}]$ (source: [30])**

Or with the eyes you can see the gradient diffusion in the photo below (Tasman Sea - that is a salt sea - meets the Pacific Ocean at the very northern tip of New Zealand):



and also on the photo below (Baltic Sea/North Sea in Denmark near Skagen):



Remark

In the field of electrokinetics, the first Fick's law is often denoted by engineers as following:

$$J_{\text{diff}} = -qD\frac{\mathrm{d}\rho}{\mathrm{d}x} \tag{9.6.452}$$

and named the "diffusion current".

To fall back on the famous relation known by small classes of chemistry it is necessary to know that the chemist is accustomed to write the density by the letter C to indicate that it is a **C**oncentration. Therefore, we have:

$$\Phi_p = -D\frac{\Delta C}{L} \tag{9.6.453}$$

But this is not the only change of notation made by chemists. Indeed, it must be remembered that the surfacic flow of particles (mass) is defined by the quantity of mass passing through a constant surface per unit time. We then have:

$$\frac{\dfrac{\Delta m}{\Delta t}}{S} = -D\frac{\Delta C}{L} \tag{9.6.454}$$

Since the origin of time is often taken as zero and the initial mass is zero then we get finally after a small rearrangement the classical relation known by chemists and biologists (to a given sign since it is only a matter of convention) :

$$\boxed{m = -\frac{DS\Delta Ct}{L}} \tag{9.6.455}$$

✍**Example:**

Plants absorb water for photosynthesis. We know from experimental measurements that the diffusion coefficient of water in the air is $D \cong 2.4\cdot 10^{-5}\;[\text{m}^2\cdot\text{s}^{-1}]$. The pores of absorption of plants have a section of the order of $S = 8.0\cdot 10^{-11}\;[\text{m}^2]$. The diffusion distance is of the order of $L \cong 2.5\cdot 10^{-5}\;[\text{m}]$. The water vapour density inside the plant is of the order of $C_2 \cong 0.022\;[\text{kg}\cdot\text{m}^{-3}]$ and outside the plan of $C_1 \cong 0.011\;[\text{kg}\cdot\text{m}^{-3}]$.

Consequently, the mass of water absorbed in one hour is approximately:

$$m = \frac{DS\Delta Ct}{L} \cong \frac{2.4\cdot 10^{-5}\cdot 8\cdot 10^{-11}\cdot(0.022-0.011)\cdot 3600}{2.5\cdot 10^{-5}} \cong 3\cdot 10^{-9}\;[\text{kg}] \tag{9.6.456}$$

This result must be multiplied by the millions of pores that a plant has.

Remark

In the facts, Fick proved at first the first law and proceeding in exactly the same way to the heat equation he obtained the second law that bears his name (that is to say in the opposite order presented above).

We can also estimate the flow of thermal energy transported by these same particles along the x-axis. Indeed, in each slice of fluid, n particles carry each an energy E corresponding to a given amount of heat Q (according to Joule's law). We therefore have a surfacic flux of energy Φ_q whose first component is given by the same type of balance as the previous developments:

$$\Phi_q \cong -\frac{1}{3}\frac{\partial(nQ)}{\partial x}\lambda_x\bar{v} = -\frac{1}{3}\lambda_x\bar{v}\frac{\partial(nQ)}{\partial T}\frac{\partial T}{\partial x} \tag{9.6.457}$$

We find in it immediately the definition of the calorific capacity (if we divide by mass we would have the massic calorific capacity according to what we saw earlier above). Thus, in the one-dimensional case:

$$\boxed{\kappa = \frac{1}{3}\lambda_x\bar{v}C = CD} \tag{9.6.458}$$

There is therefore a simple ratio of proportionality between κ and C.

Fourier's law	**Fick's law**	**Ohm's law**
Thermodynamics	Thermodynamics	Electrokinetics
$\vec{j}_q = -\kappa\vec{\nabla}T$	$\vec{j} = -D\vec{\nabla}\rho$	$\vec{j} = \sigma\vec{E} = -\sigma\vec{\nabla}U$
Thermic current density	Particle current density	Electric current density
T: Temperature	ρ: concentration	U: electric potential
κ: Thermic conductivity	D: coefficient of diffusion	σ: electric conductivity
Thermic flux: $\Phi_q = \iint_S \vec{j}_q \circ \mathrm{d}\vec{S}$	Particle flux: $\Phi_p = \iint_S \vec{j}_p \circ \mathrm{d}\vec{S}$	Electric flux: $I = \iint_S \vec{j} \circ \mathrm{d}\vec{S}$

Table 9.32: **Similarities of different diffusion laws in physics**

Here are some self-diffusion coefficient D_A of some gases at $T = 273$ [K] and $P = 0.1$ [MPa]:

Gas	D_A [cm$^2\cdot$s^{-1}]	Gas	D$_A$ [cm$^2\cdot$s^{-1}]	Gas	D_A [cm$^2\cdot$s^{-1}]
H_2	1.604	O_2	0.192	H_2O	0.276
He	1.386	N_2	0.155	CH_4	0.188
Ar	0.157	CO_2	0.106	NH_3	0.192

And some interdiffusion coeffients D_{AB}:

Diffusing gas A	Medium B	Pressure mmH$_2$O	Temperature [K]	D_{AB} [cm$^2\cdot$s^{-1}]
H_2	He	760	298	1.64
Ar	He	751	288	0.703
N_2	CO_2	760	288	0.158
H_2O	air	760	273	0.219
H_2	N_2	760	288	0.743
H_2O	CO_2	760	273	0.146

And some diffusion coefficients of gases in liquids:

Gas	Liquid	Temperature [K]	D (10^{-5}) $[\text{cm}\cdot\text{s}^{-1}]$
Air	water	293	2.5
CO_2	water	298	1.92
H_2	water	298	4.50
NH_3	water	285	1.64
CO_2	ethanol	298	3.42
CO_2	heptane	298	6.03

9.6.12 Thermal radiation

Thermal radiation is electromagnetic radiation generated by the thermal motion of charged particles in matter. All matter with a temperature greater than absolute zero emits thermal radiation. When the temperature of a body is greater than absolute zero, inter-atomic collisions cause the kinetic energy of the atoms or molecules to change. This results in charge-acceleration and/or dipole oscillation which produces electromagnetic radiation, and the wide spectrum of radiation reflects the wide spectrum of energies and accelerations that occur even at a single temperature.

Examples of thermal radiation include the visible light and infra-red light emitted by an incandescent light bulb, the infra-red radiation emitted by animals is detectable with an infra-red camera, and the cosmic microwave background radiation. Thermal radiation is different from thermal convection and thermal conduction person near a raging bonfire feels radiant heating from the fire, even if the surrounding air is very cold.

Sunlight is part of thermal radiation generated by the hot plasma of the Sun. The Earth also emits thermal radiation, but at a much lower intensity and different spectral distribution (infra-red rather than visible) because it is cooler. The Earth's absorption of solar radiation, followed by its outgoing thermal radiation are the two most important processes that determine the temperature and climate of the Earth.

If a radiation-emitting object meets the physical characteristics of a black-body in thermodynamic equilibrium, the radiation is named "black-body radiation". Planck's law describes the spectrum of black-body radiation, which depends only on the object's temperature. Wien's displacement law determines the most likely frequency of the emitted radiation, and the Stefan–Boltzmann law gives the radiant intensity.

Thermal radiation is one of the fundamental mechanisms of heat transfer.

9.6.12.1 Black Body radiation

Black-body radiation is the type of electromagnetic radiation within or surrounding a body in thermodynamic equilibrium with its environment, or emitted by a black-body (an opaque and non-reflective body), assumed for the sake of calculations and theory to be held at constant, uniform temperature. The radiation has a specific spectrum and intensity that depends only on the temperature of the body.

The thermal radiation spontaneously emitted by many ordinary objects can be approximated as black-body radiation. A perfectly insulated enclosure that is in thermal equilibrium internally contains black-body radiation and will emit it through a hole made in its wall, provided the hole is small enough to have negligible effect upon the equilibrium.

A black-body at room temperature appears black, as most of the energy it radiates is infra-red and cannot be perceived by the human eye. Because the human eye cannot perceive color at very low light intensities, a black-body, viewed in the dark at the lowest just faintly visible temperature, subjectively appears grey (but only because the human eye is sensitive only to black and white at very low intensities - in reality, the frequency of the light in the visible range would still be red, although the intensity would be too low to discern as red), even though its objective physical spectrum peaks in the infra-red range. When it becomes a little hotter, it appears dull red. As its temperature increases further it eventually becomes blue-white.

Although planets and stars are neither in thermal equilibrium with their surroundings nor perfect black bodies, black-body radiation is used as a first approximation for the energy they emit.[6] Black holes are near-perfect black bodies, in the sense that they absorb all the radiation that falls on them. It has been proposed that they emit black-body radiation (named "Hawking radiation"), with a temperature that depends on the mass of the Black Hole.

The study of the black-body is the basis of the famous theory of Wave Quantum Physics (see section of the same name page 4616), one of the pillars of actual modern physics. Indeed, some experimental results could not be explained without the introduction of the famous Planck constant, the use of quantification of energy (quanta) and the acceptance of the atomic model and Boltzmann statistical theory (implicitly the second principle of thermodynamics of which Max Planck was a specialist).

Before starting we will make an exceptionally (for this textbook) an historical hook which that we suppose to be very useful to understand why the study of thermal radiation is so important in physics.

The fact that all heated objects emit a light of the same color at the same temperature is a curiosity that had been known for centuries in some fields, long before Gustav Krichhoff began his scientific and theoretical investigations on the nature of this strange correlation. To simplify his analysis, he elaborated the concept of a perfectly absorbing and perfectly emissive object. Thus, as a perfect transmitter, it would be anything but black if its temperature were high enough so that it could radiate in wavelengths of the visible part of the spectrum. Therefore, with respect to experience feedback, the radiation of this object would have an intensity and a spectral extent independent of the type of material of which it is composed. The objective was then to measure the spectral distribution of the radiated energy for each wave length and for each temperature plateau and to derive an equation and therefore dependent of two variables (wavelength and temperature) to reproduce this distribution and therefore this for any material placed under the conditions of a perfectly emessive and receptive body.

Definition (#9.6.465): A "black-body" (or "integral receiver") is defined as a body having an "energy absorption coefficient α" (see section Optical Geometry page 4351) and an "emissivity coefficient ε" equal to the unity.

Definition (#9.6.466): The emissive power of an arbitrary opaque body of fixed size and shape at a definite temperature can be described by a dimensionless ratio, named "emissivity", the ratio of the emissive power of the body to the emissive power of a black-body of the same size and shape at the same fixed temperature[28].

The first principle of thermodynamics establishes an equivalence between work and heat as modes of energy transfer between a system and its environment (and in fact the balance in terms of internal energy). We are interested here in the heat, which we can be define as "the energy one body communicates to another because of their temperature difference".

Heat is, for recall, communicated from one place to another in three different ways, as we have already mentioned earlier above:

1. By conduction: it is a transfer of heat in a set of material points in contact which takes place without macroscopic movements, under the influence of a temperature gradient. Conduction is therefore the result of molecular collisions. We observe it mainly in solids: in metals, it involves the free electrons which make them good conductors of heat. On the other hand, in insulators, conduction is badly done. Hence the strong correspondence between the thermal and electrical properties of solids.
2. By convection: convection involves the transport of heat by a part of a fluid that mixes with another fluid or material. It has its origin in a macroscopic transport of matter and therefore does not concern solids most of time.
3. By radiation: conduction and convection assume the presence of matter. Radiation, on the other hand, allows a transfer of energy that can be made through the vacuum. This is electromagnetic radiation. It should be noted that radiation is not a mode of heat transfer but energy, which can be transformed into heat in contact with a body.

The thermal radiation emitted by a body at a given temperature results from a conversion of the internal energy of the body into radiation (excited quantum state level back to ground level). Conversely, absorption is the transformation of the incident energy into internal energy.

When a surface is subjected to absorbed radiation, we perform the energy balance according to the Kirchhoff law seen in our study of photometry (see section Optical Geometry page 4351):

$$\alpha+\rho+\tau=1 \tag{9.6.459}$$

where let us recall that α is the fraction of the absorbed radiation, ρ is the reflected part (scattered) and τ the transmitted part (which crosses the surface). This balance schema results from the principle of energy conservation.

---

[28] With this definition, Kirchhoff's law states that for an arbitrary body emitting and absorbing thermal radiation in thermodynamic equilibrium, the emissivity is equal to the absorptivity.

We will now examine the mechanisms of absorption and emission and establish a link between heat and radiant energy before we to focus directly to the black-body:

9.6.12.1.1 Stefan–Boltzmann law

In our study of photometry, we defined the concept of emittance (energy irradiated by a non-punctual body per unit area) for the entire spectrum.

What we had omitted to specify, however, is that in order for a body to radiate (apart from the fact that it can itself be illuminated by another body) it must be heated (that we provide an excitation energy to the constituents of the body in question - that is to say to the electrons).

So we should be able to establish a relation between the temperature of a body and its emittance.

In 11879 (holocene calendar), the physicist Jožef Stefan was able to establish experimentally that the total emittance of the black-body (or "energetic excitance" also named "radiant exitance" of the black-body) at a temperature T increased in proportion to the fourth power of the temperature such that:

$$\boxed{M(T)=\sigma T^4} \tag{9.6.460}$$

Where $M(T)$ is the integration on all wavelengths (or frequencies ... whatever) of $M(\lambda,T)$:

$$M(T)=\int\limits_0^{+\infty} M(\lambda,T)\mathrm{d}\lambda \tag{9.6.461}$$

with $M(\lambda,T)$ given by the Planck law which we prove further below.

Let us also recall that (this will be shown during our proof of Planck's law):

$$\sigma\cong 5.670374419\cdot 10^{-8}\;[\mathrm{W}\cdot\mathrm{m}^{-2}\cdot\mathrm{K}^{-4}] \tag{9.6.462}$$

is the "Stefan constant" (not to be confused with Stefan-Boltzmann's constant !!!).

In 11884 (holocene calendar), Ludwig Boltzmann indirectly proved Stefan's law on the basis of the study of the black-body in thermal equilibrium (where we consider that the border of the black-body wall define the terminations of electromagnetic waves) from the theory of electromagnetism and using a powerful thermodynamic reasoning.

Proof 9.10.5. In a first step, Boltzmann determined the radiation pressure in such an enclosure (or in such a body).

Here are the developments that led him to determine the radiation pressure $P(T)$ at the thermodynamic equilibrium temperature T for the corresponding internal energy density $\rho(T)$:

Let us recall the expression of the "Einstein relation" that we have proved during our study of Special Relativity:

$$p=\frac{h\nu}{c} \tag{9.6.463}$$

Let us consider now a chamber of volume V whose walls are reflective for the photons (case of the black-body):

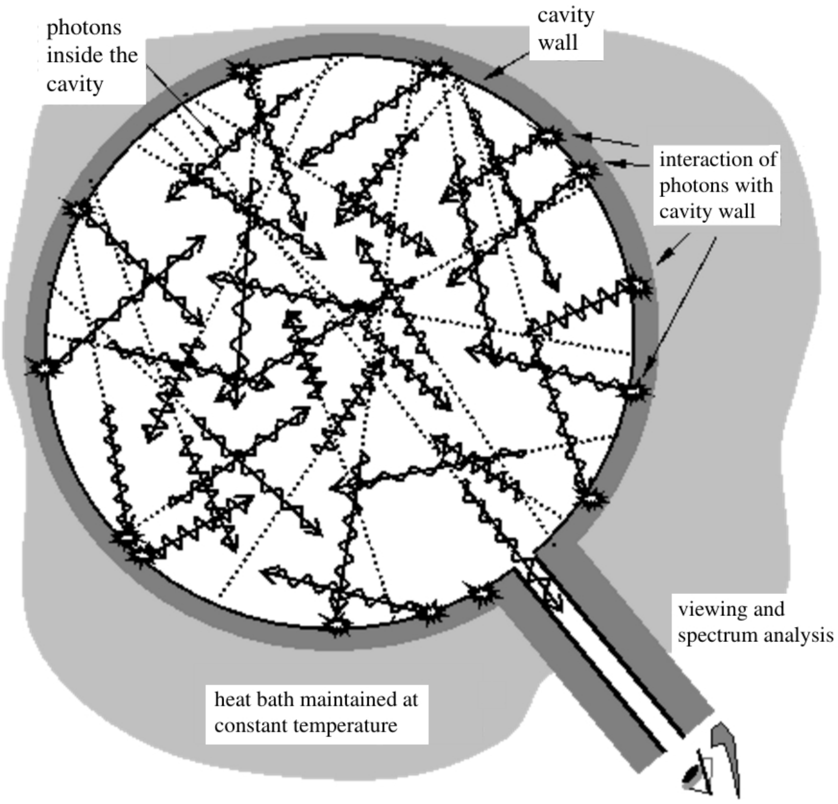

photons
inside the
cavity
cavity
wall
interaction of
photons with
cavity wall
viewing and
spectrum analysis
heat bath maintained at
constant temperature

Figure 9.174: **Conceptual black-body**

We study the variation of the momentum before and after the collision on an infinitely small surface $\mathrm{d}s$ (which makes it possible to consider the trajectories before and after the impact as rectilinear and symmetrical with respect to the oriented axis $\mathrm{O}x$ perpendicular to the surface of the black-body coinciding with the surface $\mathrm{d}s$).

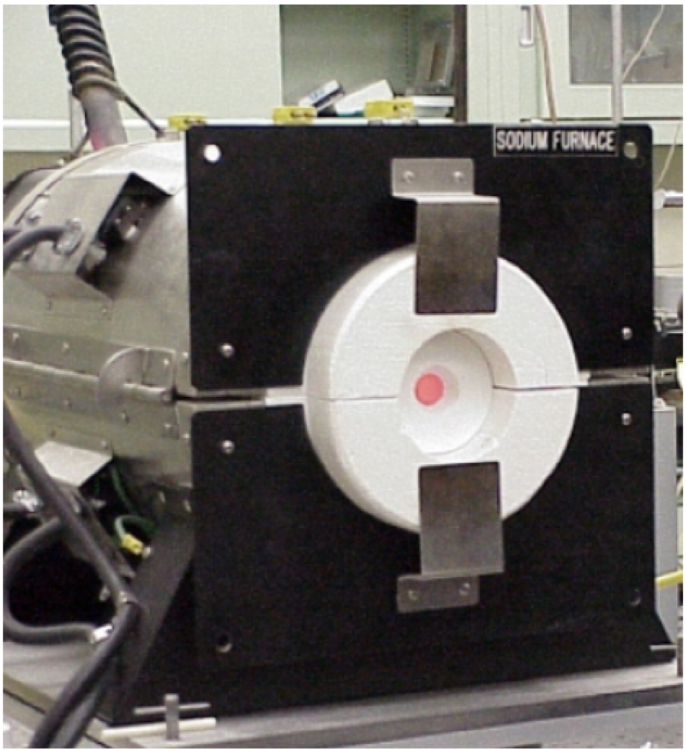


Figure 9.175: **Black-body laboratory radiator**

Thus, we have before collision for the linear momentum:

$$\vec{p}=p_x\vec{u}_x+p_y\vec{u}_y \tag{9.6.464}$$

and after collision:

$$\vec{p}\,'=-p'_x\vec{u}_x+p'_y\vec{u}_y \tag{9.6.465}$$

If the collision is elastic (which is comforting relative to the photon ...):

$$|p'_x|=|p_x| \qquad |p'_y|=|p_y| \tag{9.6.466}$$

Then we have:

$$\vec{p}\,'=-p_x\vec{u}_x+p_y\vec{u}_y \tag{9.6.467}$$

The variation of the momentum is then:

$$\mathrm{d}\vec{p}=\vec{p}\,'-\vec{p}=(-p_x\vec{u}_x+p_y\vec{u}_y)-(p_x\vec{u}_y+p_y\vec{u}_y)=-2p_x\vec{u}_x \tag{9.6.468}$$

As:

$$p_x=p\cos(\alpha) \tag{9.6.469}$$

then we have:

$$\mathrm{d}\vec{p}=-2p\vec{u}_x\cos(\alpha) \tag{9.6.470}$$

By considering only the norm of the expression and that it is a single photon:

$$\mathrm{d}p=2\frac{h\nu}{c}\cos(\alpha) \tag{9.6.471}$$

Remark

We assume that after its rebound, the photon conserves its frequency which leads us to suppose that the black-body has stationary waves at thermodynamic equilibrium (see section Wave Mechanics page 3520).

Until now, we have reasoned on a single photon, but the enclosure actually contains a gas of photons. The internal volume energy u of the radiation contains a volumetric density n of photons of identical frequency ν. Consequently, the quantity n of photons per unit volume in the enclosure is:

$$n=\frac{u}{h\nu}\,[\mathrm{m}^{-3}] \tag{9.6.472}$$

Remark

We specify the units, because we noticed that the following development sometimes presented some problems of understanding to some readers if the units were not given explicitly.

We consider that during a time interval $\mathrm{d}t$ the number of photons potentially striking the surface $\mathrm{d}s$ at an angle of incidence α is contained in a cylinder of generator $c\mathrm{d}t$ whose axis is necessarily inclined by an angle α and having as base area $\mathrm{d}s$. The volume of this cylinder is then by the projection of the base surface: The number of photons $\mathrm{d}n_c$ that potentially strike the wall $\mathrm{d}s$ per unit time is:

$$\mathrm{d}n_c=n\mathrm{d}V=\frac{u}{h\nu}\mathrm{d}V=\frac{uc}{h\nu}\cos(\alpha)\mathrm{d}s\mathrm{d}t \tag{9.6.473}$$

In this last expression, we assumed that all the photons of $\mathrm{d}V$ had a linear momentum in the direction subtended by α. In fact, the photons actually arriving at $\mathrm{d}s$ are contained in a solid angle $\mathrm{d}\Omega$ between two cones of half-angle at the apex of α and $\alpha+\mathrm{d}\alpha$ (for reasons of the geometry of the black-body's experiment which was, except error, spherical and by the way this spherical symmetry facilitates the calculations...).

The relation between $\mathrm{d}\Omega$ and α is as we have proved in the section of Trigonometry:

$$\mathrm{d}\Omega=\mathrm{d}(2\pi(1-\cos(\alpha))=2\pi\sin(\alpha)\mathrm{d}\alpha \tag{9.6.474}$$

Knowing that in the whole volume (for recall), the solid angle is:

$$\Omega=\frac{4\pi r^2}{r^2} \tag{9.6.475}$$

the number $\mathrm{d}n$ included in the elementary solid angle $\mathrm{d}\Omega$ which reaches the surface $\mathrm{d}s$ at an angle of incidence between α and $\mathrm{d}\alpha$ is then:

$$\mathrm{d}n=\frac{\mathrm{d}\Omega}{\Omega}\mathrm{d}n_c=\frac{2\pi\sin(\alpha)\mathrm{d}\alpha}{4\pi}\frac{uc\cos(\alpha)}{h\nu}\mathrm{d}s\mathrm{d}t=\frac{uc\sin(\alpha)\cos(\alpha)\mathrm{d}\alpha\mathrm{d}s\mathrm{d}t}{2h\nu} \tag{9.6.476}$$

Using now the definition of the pressure P:

$$P=\frac{\mathrm{d}F}{\mathrm{d}s}=\frac{F\mathrm{d}n}{\mathrm{d}s}=\frac{\left(\dfrac{\mathrm{d}p}{\mathrm{d}t}\right)\mathrm{d}n}{\mathrm{d}s} \tag{9.6.477}$$

by substituting what is covenant:

$$P=\frac{\frac{\mathrm{d}p}{\mathrm{d}t}\mathrm{d}n}{\mathrm{d}s}=\frac{\frac{2\frac{h\nu}{c}\cos(\alpha)}{\mathrm{d}t}\frac{uc\sin(\alpha)\cos(\alpha)}{2h\nu}}{\mathrm{d}s} \tag{9.6.478}$$

Which gives after simplification:

$$P=u\sin(\alpha)\cos^2(\alpha)\mathrm{d}\alpha=u\cos^2(\alpha)\mathrm{d}(-\cos(\alpha) \tag{9.6.479}$$

The total pressure of radiation being given in this case by:

$$P=nh\nu\int\limits_0^{\pi/2}\cos^2(\alpha)\mathrm{d}(-\cos(\alpha))=-nh\nu\frac{\cos^2(\alpha)}{3}\bigg|_0^{\pi/2}=\frac{nh\nu}{3} \tag{9.6.480}$$

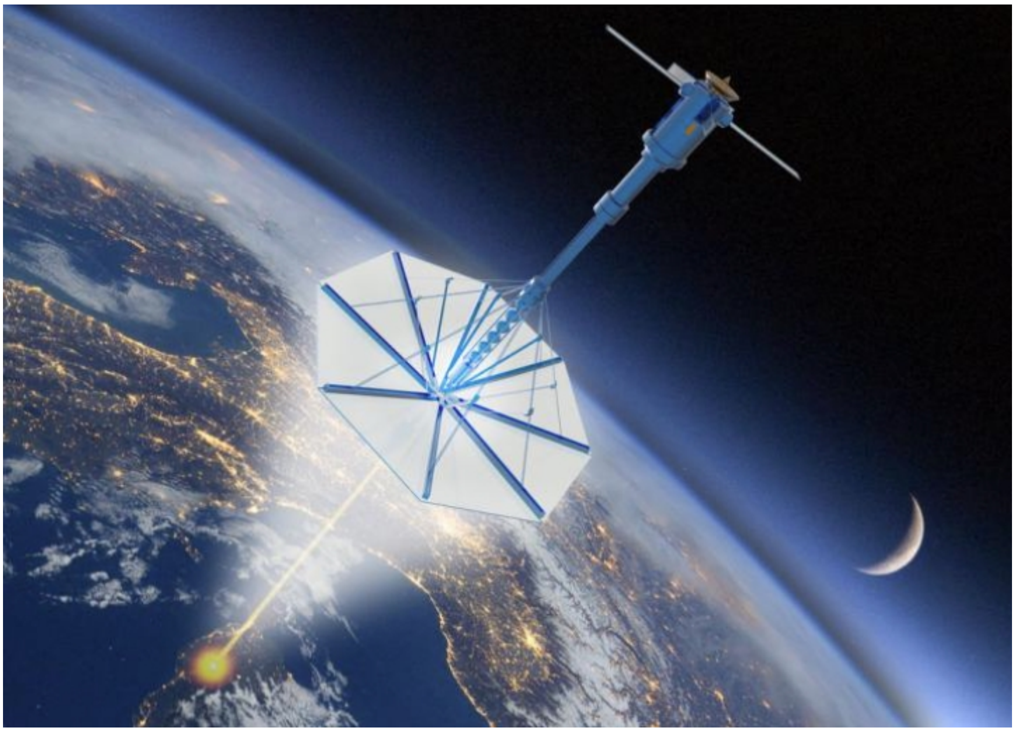
Which is equivalent to write (at equilibrium temperature for a given temperature):

$$\boxed{P(T)=\frac{\rho(T)}{3}} \tag{9.6.481}$$

Relation that is very important and will be useful to us in the section Cosmology during our study of some Friedman Universe models . This relation is also important for rocket science using propulsion by radiation!

As we have seen it in the table of most common pressure level at the beginning of this chapter, remember that the radiation pressure at the surface of the Sun (actually a G2V-type star, aka yellow dwarf) is 0.2 [Pa] and 0.05 [μPa] at Earth level. It quite obviously decrease as the inverse of the square of distance from the Sun!

LASER powered space objects with a LASER based on Earth is very unlikely to happen as the power of a LASER able to counterbalance Earth or Sun gravity is far beyond the technology available in the 121st century (holocene calendar) and even if it was, such a LASER beam would have too many side effects on the local atmosphere (ionization) to be tolerated by humans and politics. We may solve the problem of the ionized atmosphere by boarding the LASER on the space vehicle but there will be therefore the problem of the source power of the LASER that is still too huge to be reliable on a long period of time as required for interstellar travels (at least in the 121st century and very likely for the few next centuries too excepted if we found a new type of extra light source of energy...).



The total energy is the energy density multiplied by the considered volume:

$$E=\rho(T)V \tag{9.6.482}$$

Let us suppose that the volume could vary. The work done by the radiation pressure during a dilatation $\mathrm{d}V$ is obviously given by (dimensional analysis):

$$\mathrm{d}W=-P(T)\mathrm{d}V=-\frac{\rho(T)}{3}\mathrm{d}V \tag{9.6.483}$$

The variation of internal energy of the system following the first principle of thermodynamics is:

$$\mathrm{d}E=\delta Q+\delta W \tag{9.6.484}$$

But as $E=\rho(T)V$, we have:

$$\mathrm{d}E=\mathrm{d}\rho(T)V+\rho(T)\mathrm{d}V=\mathrm{d}\rho(T)V\frac{\mathrm{d}T}{\mathrm{d}T}+\rho(T)\mathrm{d}V \tag{9.6.485}$$

Hence:

$$\delta Q=\mathrm{d}E-\delta w=\rho(T)\mathrm{d}V+\frac{\mathrm{d}\rho(T)V\mathrm{d}T}{\mathrm{d}T}-\left(-\frac{\rho(T)}{3}\mathrm{d}V\right)=\frac{4}{3}\rho(T)\mathrm{d}V+\frac{\mathrm{d}\rho(T)V\mathrm{d}T}{\mathrm{d}T} \tag{9.6.486}$$

and following the second principle of thermodynamics (do not confuse here the notation of the entropy S with that of a surface S...) we have for an ideal reversible system:

$$\mathrm{d}S=\frac{\delta Q}{T} \tag{9.6.487}$$

We have then:

$$\mathrm{d}S=\frac{4}{3}\frac{\rho(T)\mathrm{d}V}{T}+\frac{\mathrm{d}\rho(T)V\mathrm{d}T}{T\mathrm{d}T} \tag{9.6.488}$$

Rearranged this can be written:

$$\mathrm{d}S=\frac{4}{3}\frac{\rho(T)}{T}\mathrm{d}V+\frac{\mathrm{d}\rho(T)V}{T\mathrm{d}T}\mathrm{d}T=\frac{\partial S}{\partial V}\mathrm{d}V+\frac{\partial S}{\partial T}\mathrm{d}T \tag{9.6.489}$$

As $\mathrm{d}S$ is a total exact differential, we have proved in the section of Differential and Integral Calculus that tan S satisfy the Schwarz' theorem:

$$\frac{\partial^2 S}{\partial T\partial V}=\frac{\partial^2 S}{\partial V\partial T} \tag{9.6.490}$$

We have in this case:

$$\frac{\mathrm{d}}{\mathrm{d}T}\left(\frac{4}{3}\frac{\rho(T)}{T}\right)=\frac{\mathrm{d}}{\mathrm{d}V}\left[\frac{\mathrm{d}\rho(T)V}{T\mathrm{d}T}\right]=\frac{\mathrm{d}\rho(T)}{T\mathrm{d}T} \tag{9.6.491}$$

Which bring us to write:

$$\frac{\mathrm{d}\rho(T)}{T\mathrm{d}T}=\frac{4}{3}\frac{\mathrm{d}}{\mathrm{d}T}\left(\frac{\rho(T)}{T}\right) \tag{9.6.492}$$

By calculating the derivative of the right member:

$$\frac{\mathrm{d}\rho(T)}{T\mathrm{d}T}=\frac{4}{3}\frac{\frac{\rho(T)}{\mathrm{d}T}T-\rho(T)\frac{\mathrm{d}T}{\mathrm{d}T}}{T^2} \tag{9.6.493}$$

By simplifying:

$$\frac{\mathrm{d}\rho(T)}{\mathrm{d}T}=\frac{4}{3}\frac{\rho(T)}{\mathrm{d}T}-\frac{4}{3}\frac{\rho(T)}{T} \tag{9.6.494}$$

after rearranging:

$$\frac{\mathrm{d}\rho(T)}{\mathrm{d}T}\left(1-\frac{4}{3}\right)=-\frac{4}{3}\frac{\rho(T)}{T} \tag{9.6.495}$$

Thus:

$$-\frac{1}{3}\frac{\mathrm{d}\rho(T)}{\mathrm{d}T}=-\frac{4}{3}\frac{\rho(T)}{T} \tag{9.6.496}$$

which become the equation:

$$\frac{\mathrm{d}\rho(T)}{\rho(T)}=4\frac{\mathrm{d}T}{T} \tag{9.6.497}$$

Which gives after a trivial integration:

$$\ln(\rho(T))=4\ln(T)+c^{te} \tag{9.6.498}$$

Finally:

$$\rho(T)=c^{te}T^4 \tag{9.6.499}$$

which is commonly written:

$$\boxed{\rho(T)=\sigma_{\text{S.B.}}T^4} \tag{9.6.500}$$

with:

$$\sigma_{\text{S.B.}}=7.56\cdot 10^{-16}\,[\text{J}\cdot\text{m}^{-3}\cdot\text{K}^{-4}] \tag{9.6.501}$$

being the Stefan-Boltzmann constant with the value and the units as they were given at the time of Stefan and Boltzmann and initially experimentally (very far from the present known value known with different units than those in use today)! We will prove however further below the theoretical value with the common units used in the 120th century (holocene calendar).

We see above the correspondence which exists between the relation which we had posed at the beginning and that which we have just obtained:

$$\rho(T)=\sigma_{\text{S.B.}}T^4 \qquad M(T)=\sigma T^4 \tag{9.6.502}$$

Since we have not yet proved Planck's law at this point, we can make a daring reasoning but which we will later justify with a proof as support.

Remark

The last two relations gives us a fundamental information that all bodies that are not at zero kelvin (absolute zero) radiate!

$M(T)$ and $\rho(T)$ are differentiated at the level of the units by the dimensions of a velocity. Now, intuitively and roughly (...), the speed which can immediately appear to us as trivial in this case of study is the speed of light c. Thus, we notice that:

$$\frac{\sigma}{\sigma_{\text{S.B.}}}=\frac{c}{4} \tag{9.6.503}$$

Remark

The product $\frac{4}{c}\sigma$ is sometimes known as the "radiation constant" or "radiation density constant".

Which gives us the Stefan–Boltzmann law expressed as a formula for radiation energy density :

$$\boxed{M(T)=\frac{c}{4}\rho(T)=\frac{c}{4}\sigma_{\text{S.B.}}T^4} \tag{9.6.504}$$

Curious isn't it...? But we will prove it later because our philosophy in this textbook is to never (or as little as possible) leave room for intuition.

■ Q.E.D.

Remark

When we will study Planck's law, the $\rho(T)$ will be denoted $R(T)$ to not to confuse then energy density with the radiance (because the notation can unfortunately be confusing).

Let us consider now an isolated chamber or cavity (like a furnace) in thermal equilibrium at a certain temperature T. This cavity will surely be filled with electromagnetic radiations of different wavelengths. Let us suppose that there exists a distribution function $M(T)$ which depends only on the temperature.

Logically, the total quantity of electromagnetic energy at all wavelengths absorbed by the walls of the cavity must be equal to that emitted by the walls otherwise the body forming the cavity would see its temperature change. Gustav Robert Kirchhoff reasoned that if the body forming the cavity is made of different materials (thus behaving in different ways with temperature), then the equilibrium between emitted radiation and absorbed radiation must apply for each wavelength (or domain of wave length).

We thus see that $M(T)$ is a universal function, the same for all the cavities, regardless of their composition, their geometry or the color of their walls. Kirchhoff did not give this function, but he pointed out that a perfectly absorbing body, that is to say a body for which $\alpha=1$ will appear, how to say... : black.

It comes then that the radiation stored in equilibrium in an isolated cavity in thermodynamic equilibrium is in all aspects the same as that emitted by a perfectly black-body at the same temperature.

Obviously, if the cavity is closed, we can not measure the current of energy that escapes. But let us create a very small hole in this cavity (small enough not to disturb the equilibrium of the electromagnetic radiation inside), then the electromagnetic energy escaping from this small hole is the same as that emitted by a perfectly body black.

However, no object is actually a black-body. Coal black has an absorption coefficient very close to 1 but only for certain frequencies (including, of course, the visible frequencies). Its absorption coefficient is much smaller in the far infra-red. However most objects approach the behaviour of a black-body in certain frequency ranges. The human body, for example, is almost a black-body in the infra-red (hence the military night glasses ...). To treat the various bodies, named "gray body", we introduce a factor named "total emissivity" that connects the emittance emitted by the body to that emitted by a perfect black-body for which $\varepsilon=1$ (for steel the emittance is approximately equal to 0.06). Then we have:

$$\boxed{M(T)=\varepsilon\sigma T^4} \tag{9.6.505}$$

Practitioners of thermodynamics often prefer the following writing (which is actually more coherent in the context of pure thermodynamics):

$$Q=\varepsilon\sigma S T^4 \tag{9.6.506}$$

It is a very sympathetic relation because in all cases by putting $\varepsilon=1$ the curious can know for a given surface what it is the energy (or heat) at the worst (or at best following the point of view ...) emitted by a body of a certain surface at a certain temperature.

Remark

The Stefan-Boltzmann relation gives us the power emitted by a body per unit of surface by expressing it proportionally to the fourth power of the temperature. This exponent gives us the reason why it becomes more and more difficult to increase the temperature of a body by heating it, the latter losing more and more rapidly the energy that we provide for its heating.

In the figure below we have depicted the situation in which the heat bath of the previous figure is produced by an ideal gas of indefinite extent for which there would be no fringe effects:

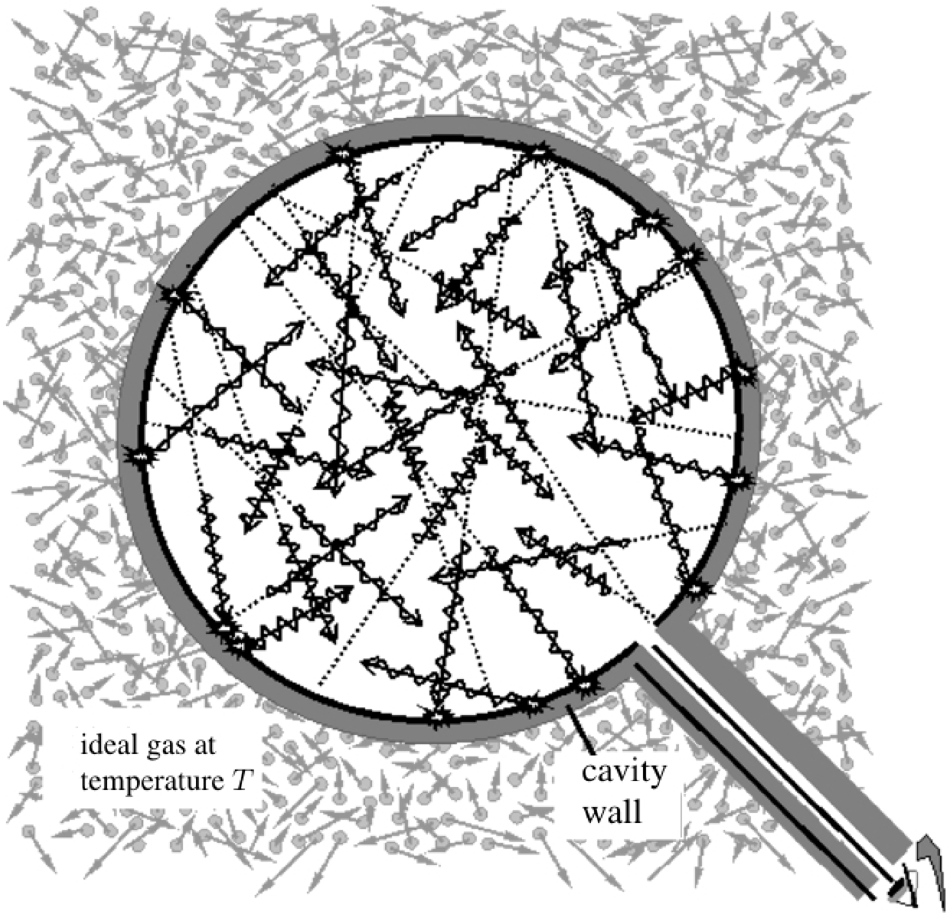


Figure 9.176: **Conceptual black-body with outside gas bath**

The temperature of the gas is maintained at the required value of the cavity wall. In an equilibrium situation characterized by the conservation of energy, total energy will be partitioned equally among all constituents of a gas. This includes not only particulate components of the gas, but also the thermalized photons of electromagnetic energy.

In this figure, as in the prior previous figure, all photons are seen to have originated as emissions or reflections off the wall of the cavity, and are assumed to remain unchanged during transmission unless and until they interact with the cavity wall in a subsequent scattering event. However, for a "bath" comprised of the stationary state ideal gas shown in the figure above, the situation is similar to taking away the cavity wall altogether and allowing the gas to move freely throughout what was formerly the cavity as is now shown in the figure further below.

Photons would then scatter off of particles in the gas rather than the wall of the cavity, but to similar effect. However, in order that it indeed be to similar effect, the gas must be dense enough or its extent great enough that the same density of photons is realized within the conceptual cavity walls as shown.

9.6.12.1.2 Planck's law

To establish Planck's law, we are not going to make the original development of Max Planck even if interesting of the intuitively point of view it was only a succession of tinkering (admitted by Max Planck himself). We will introduce this law with a modern and as simple as possible approach in the point of view of the mathematical formalism as done by Albert Einstein itself!

For this we now consider the black-body as an isolated system at thermal equilibrium, in which the radiation is in the stationary state and reflected totally by the walls. The photons can therefore be considered as particles not interacting with one another in a potential well with rectilinear walls.

Thus, identically to what we have seen in the section of Wave Quantum Physics, the solution of the problem is that of a potential well with rectilinear walls for which we have obtained for wave function:

$$\Psi_x=\Psi_{x,0}\sin(\omega x)=\Psi_{x,0}\sin\left(x\sqrt{\frac{8\pi^2 m}{h^2}(E_{\text{tot}}-E_p)}\right) \tag{9.6.507}$$

function to which the boundary conditions should be applied.

The conditions we imposed during our study of this case in the section of Wave Quantum Physics were too restrictive (this is why they are named "strict boundary conditions"). Indeed, atoms of the wall absorb and emit radiation irrespective of how the radiation incident. But equilibrium imposes at least that boundary conditions are periodic by the very definition of equilibrium. This is why we impose what we name the "periodic boundary conditions":

- For $x=0$ and $x=L_x$, we have: $\Psi(0)=\Psi(L_x)=c^{te}$
- The wave function Ψ must have an integer number of half wavelengths over the length L_x
- In the black-body, $U=0$ therefore $E_p=0$
- If at the ends ($x=0$ and $x=L_x$) we have $\Psi=c^{te}$ the argument of the sine has the same value $n2\pi$ (to a given integer multiplicative factor) in 0 and in L_x.

So we must have:

$$L_x\sqrt{\frac{8\pi^2 m}{h^2}(E_{\text{tot}}-E_p)}=n_x 2\pi \tag{9.6.508}$$

and as $E_p = 0$, after some elementary algebraic simplifications, we get:

$$\frac{8\pi^2 m}{h^2}E_{\text{tot}} = \frac{n_x^2}{L_x^2}4\pi^2 \Rightarrow E_{nx} = \frac{n_x^2}{L_x^2}4\pi^2\frac{h^2}{8\pi^2 m} = \frac{n_x^2}{L_x^2}\frac{h^2}{2m} \tag{9.6.509}$$

Where E_{nx} is the total energy of the quantum level n according to x (see section Wave Quantum Physics page 4616).

The total energy of the particle thus presents a discrete sequence of values, the only ones permitted. The value of L_x is determined using the Bohr or Sommerfeld model, depending on the case.

Since the corresponding wave functions in the well are such that $U = 0 \Leftrightarrow E_p = 0$, we have therefore:

$$\begin{aligned}\Psi_x &= \Psi_{x,0}\sin\left(x\sqrt{\frac{8\pi^2 m}{h^2}E_{\text{tot}}}\right) = \Psi_{x,0}\sin\left(\frac{8\pi^2 m}{h^2}\frac{n_x^2}{L_x^2}\frac{h^2}{2m}\right) = \Psi_{x,0}\sin\left(x\sqrt{\frac{4\pi^2 n_x^2}{L_x^2}}\right)\\ &= \Psi_{x,0}\sin\left(\frac{2\pi n_x}{L_x}x\right) = \Psi_{x,0}\sin(k_x x)\end{aligned} \tag{9.6.510}$$

Thus, the total energy can be written:

$$E_{nx} = \frac{n_x^2}{L_x^2}\frac{2\pi^2\hbar^2}{m} = \frac{\hbar^2}{2m}\frac{4\pi^2 n_x^2}{L_x^2} = \frac{\hbar^2}{2m}\left(\frac{2\pi n_x}{L_x}\right)^2 = \frac{\hbar^2}{2m}k_x^2 \tag{9.6.511}$$

Thus, given that the wave function is a conditional probability, we have in the form of a phasor:

$$\Psi = \Psi_x\Psi_y\Psi_z = c^{te}e^{\mathrm{i}k_x x}e^{\mathrm{i}k_y y}e^{\mathrm{i}k_z z} = c^{te}e^{\mathrm{i}\vec{k}\circ\vec{r}} \tag{9.6.512}$$

And the associated discrete energies are then:

$$E_n = \frac{\hbar^2}{2m}(k_x^2 + k_y^2 + k_z^2) = \frac{\hbar^2\|\vec{k}\|^2}{2m} \tag{9.6.513}$$

The vector $\vec{k}$ is therefore defined by:

$$\vec{k} = \left(n_x\frac{2\pi}{L_x}, n_y\frac{2\pi}{L_y}, n_z\frac{2\pi}{L_z}\right) \tag{9.6.514}$$

Remark

We can easily see that the differences in energy between consecutive levels are all smaller as the dimensions of the black-body (assimilated to a box) L_x,L_y,L_z are larger. For macroscopic dimensions, these differences are totally inappreciable. This observation will allow us a little further to make a small approximation.

Explanation: For an electron ($m \cong 10^{-30}$ [kg]) enclosed in a cubic box of side $L = 1$ [m], the difference between two consecutive levels is:

$$E_{nx+1,ny,nz} - E_{nx,ny,nz} = (2n_x + 1)\frac{2\pi^2\hbar^2}{mL_x^2} \tag{9.6.515}$$

Therefore approximately 10^{-35} [J]...

The vectors $\vec{k}$ that interest us (since each represent respectively a possible micro-state), immersed in the phase space of the wave numbers, have their end situated in one of the nodes of a three-dimensional network constituted by elementary meshes whose edges are parallel to the axes and which measure respectively $2\pi/L_x$, $2\pi/L_y$ and $2\pi/L_z$. We want to evaluate the number of vectors for which this end falls in the interval between the two spheres centered at the origin and radii of standard k and $k+\mathrm{d}k$. The infinitesimal volume of the spherical shell between the two spheres is thus trivially given by:

$$\mathrm{d}V=4\pi k^2\mathrm{d}k \tag{9.6.516}$$

The number of elementary cells (microstates) included in this region of the space of $\vec{k}$ is roughly equal to the number of times its volume contains that of the elementary cell, which is:

$$\frac{2\pi}{L_x}\frac{2\pi}{L_y}\frac{2\pi}{L_z}=\frac{8\pi^3}{V} \tag{9.6.517}$$

We thus obtain the number of micro-states in the volume $\mathrm{d}V$ (hence the density of micro-states):

$$\frac{4\pi k^2\mathrm{d}k}{\dfrac{8\pi^3}{V}}=\frac{Vk^2\mathrm{d}k}{2\pi^2} \tag{9.6.518}$$

We must not forget the following relations (see sections of Wave Mechanics page 3479, Wave Quantum Physics page 4749 and Special Relativity page 5687):

$$k=\frac{2\pi}{\lambda}=\frac{\omega}{c} \tag{9.6.519}$$

Therefore as:

$$k^2=\frac{\omega^2}{c^2} \qquad \mathrm{d}k=\frac{1}{c}\mathrm{d}\omega \tag{9.6.520}$$

and (for recall):

$$\omega=2\pi\nu \tag{9.6.521}$$

Therefore it comes:

$$\frac{Vk^2\mathrm{d}k}{2\pi^2}=\frac{V\omega^2}{2\pi^2c^3}\mathrm{d}\omega=\frac{V4\pi^2\nu^2}{2\pi^2c^3}2\pi\mathrm{d}\nu=\frac{V4\pi^3\nu^2}{\pi^2c^3}\mathrm{d}\nu=\frac{V4\pi\nu^2}{c^3}\mathrm{d}\nu \tag{9.6.522}$$

But when physicists had developed this theoretical model (apparently Bose would have been the first to do so in a modern way), they had noticed that the final result did not correspond to the experience with a factor of 2. From then on, they multiplied empirically the density of micro-states by a factor 2 such that:

$$2\frac{V4\pi\nu^2}{c^3}\mathrm{d}\nu=\frac{V8\pi\nu^2}{c^3}\mathrm{d}\nu \tag{9.6.523}$$

This factor 2 would be explained today (I have never seen the proof of this personally ...) by the spin 1 of the photon. This permits a priori three values for its projection: -1, 0, 1. The value 0

would be forbidden by the quantum theoretical field (if someone has the proof of this statement, I am a taker!).

In a black-body in thermodynamic equilibrium, photons (which are bosons) form a gas whose constituents do not interact chemically with each other. This type of situation is typically described by the Bose-Einstein distribution that we have proved in the section of Statistical Mechanics. Thus, since $\mu=0$, we have in the case of a discrete spectrum of energy states (it is thus here that appears the concept of quantification that Planck had originally introduced by tinkering):

$$n_i=\frac{g_i}{e^{\frac{E_i}{kT}}-1}=\frac{g_i(\nu)}{e^{\frac{h\nu_i}{kT}}-1} \tag{9.6.524}$$

and in a case that we consider as continuous:

$$\mathrm{d}n=\frac{g(\nu)}{e^{\frac{h\nu}{kT}}-1}=\frac{\frac{V8\pi\nu^2}{c^3}\mathrm{d}\nu}{e^{\frac{h\nu}{kT}}-1} \tag{9.6.525}$$

Before continuing it is important for the reader to realize that the analysis of the radiation of the black-body thus comes for the two previous relations to be imposed:

- First a quantification of energy levels in packets $h\nu$ whereas at the time of the discovery of the phenomenon the Nature was considered continuous in all its phenomena.
- Secondly, consider Boltzmann's theory of entropy and the resulting statistical distribution of Bose-Einstein is a pillar of the study of physical systems.

In the black-body, we have for internal energy:

$$U=\sum_i E_i n_i=\int\limits_0^{+\infty} h\nu\frac{\frac{V8\pi\nu^2}{c^3}\mathrm{d}\nu}{e^{\frac{h\nu}{kT}}-1}=V\int\limits_0^{+\infty}\frac{8\pi h}{c^3}\frac{\nu^3}{e^{\frac{h\nu}{kT}}-1}\mathrm{d}\nu \tag{9.6.526}$$

Hence:

$$\frac{U}{V}=\int\limits_0^{+\infty}\frac{8\pi h}{c^3}\frac{\nu^3}{e^{\frac{h\nu}{kT}}-1}\mathrm{d}\nu=\int\limits_0^{+\infty}\rho(\nu,T)\mathrm{d}\nu\;[\mathrm{J}\cdot\mathrm{m}^{-3}] \tag{9.6.527}$$

The radiation of a black-body (its "monochromatic brightness" as some say ...) is thus given by the "Planck's law":

$$\boxed{R(\nu,T)=\frac{8\pi h\nu^3}{c^3}\frac{1}{e^{\frac{h\nu}{kT}}-1}=\frac{2\hbar\omega^3}{\pi c^3}\frac{1}{e^{\frac{\hbar\omega}{kT}}-1}\;[\mathrm{J}\cdot\mathrm{s}\cdot\mathrm{m}^{-3}]} \tag{9.6.528}$$

The first equality is often written in the following form which allows an interpretation of the result:

$$\boxed{R(\nu,T)=\frac{8\pi\nu^2}{c^3}h\nu\frac{1}{e^{\frac{h\nu}{kT}}-1}} \tag{9.6.529}$$

Where:

- $8\pi\nu^2/c^3$ corresponds to the volumetric density of modes of the radiation of the black-body
- $h\nu$ to the mean energy per quantum of energy (!!!!!)
- $\dfrac{1}{e^{\frac{h\nu}{kT}}-1}$ represents the population of the modes

Finally, let us see another common way of writing Planck's law. Since:

$$R(\nu,T)\mathrm{d}\nu=R(\lambda,T)\mathrm{d}\lambda \qquad \text{and} \qquad v=\frac{c}{\lambda}\Rightarrow\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}=\frac{-c}{\lambda^2} \tag{9.6.530}$$

therefore:

$$R(\nu,T)\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}=\frac{-c}{\lambda^2}R(\nu,T)=R(\lambda,T) \tag{9.6.531}$$

but, as $R(\lambda,T)\geq 0$, $R(\nu,T)\geq 0$, it is convenient to take the absolute value:

$$R(\nu,T)\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}=\left|-\frac{c}{\lambda^2}\right|R(\nu,T)=\frac{c}{\lambda}^2R(\nu,T) \tag{9.6.532}$$

Finally, we get another form of Planck's law which expresses the energy flux density for a given wavelength:

$$\boxed{R(\lambda,T)=\frac{8\pi ch}{\lambda^5}\frac{1}{e^{\frac{hc}{\lambda kT}}-1}} \tag{9.6.533}$$

Remark

Max Planck proposed this law by a succession of theoretical trial and errors and analogies in 11900 (holocene calendar) without knowing the statistical distribution of Bose-Einstein which is remarkable experimentally speaking!

If $h\nu\ll kT$ (hence in the domain of long wavelengths), the Taylor development of e^x for small x (see section Sequences and Series page 1703) gives:

$$e^x=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\ldots\cong 1+x \tag{9.6.534}$$

Which gives us:

$$e^{\frac{hc}{\lambda kT}}\cong 1+\frac{hc}{\lambda kT} \tag{9.6.535}$$

and the Planck's law thus becomes the "Rayleigh-Jeans law" (which had been discovered before Planck's law):

$$\boxed{R(\lambda,T)=\frac{8\pi h}{\lambda^3}\frac{1}{\dfrac{hc}{\lambda kT}}=\frac{8\pi h}{\lambda^3}\frac{\lambda kT}{hc}=\frac{8\pi}{\lambda^2}\frac{kT}{c}} \tag{9.6.536}$$

That can also sometimes be found in the specialized literature in the form:

$$R(\nu,T)=\frac{8\pi}{\left(\dfrac{c}{\nu}\right)^2}\frac{kT}{c}=\frac{8\pi\nu^2}{c^3}kT \tag{9.6.537}$$

Conversely, if $h\nu \gg kT$ we have:

$$e^{\frac{hc}{\lambda kT}} - 1 \cong e^{\frac{hc}{\lambda kT}} \tag{9.6.538}$$

And the Planck's law therefore becomes:

$$\boxed{R(\lambda,T) = \frac{8\pi h}{\lambda^3}\frac{1}{e^{\frac{hc}{\lambda kT}}} = \frac{8\pi h}{\lambda^3}e^{-\frac{hc}{\lambda kT}}} \tag{9.6.539}$$

Which is nothing but the "Wien's first law" (which had been discovered before the Rayleigh-Jeans law). This law effectively describes the presence of a maximum of radiation, but contrary to the Rayleigh-Jeans law, it provides false values for large wavelengths (i.e. small frequencies). Moreover, it implies that the intensity of radiation tends towards zero with the increase of the temperature, which also contradicts the experiment.

Here is a diagram showing the differences between the three laws at 300 Kelvins:

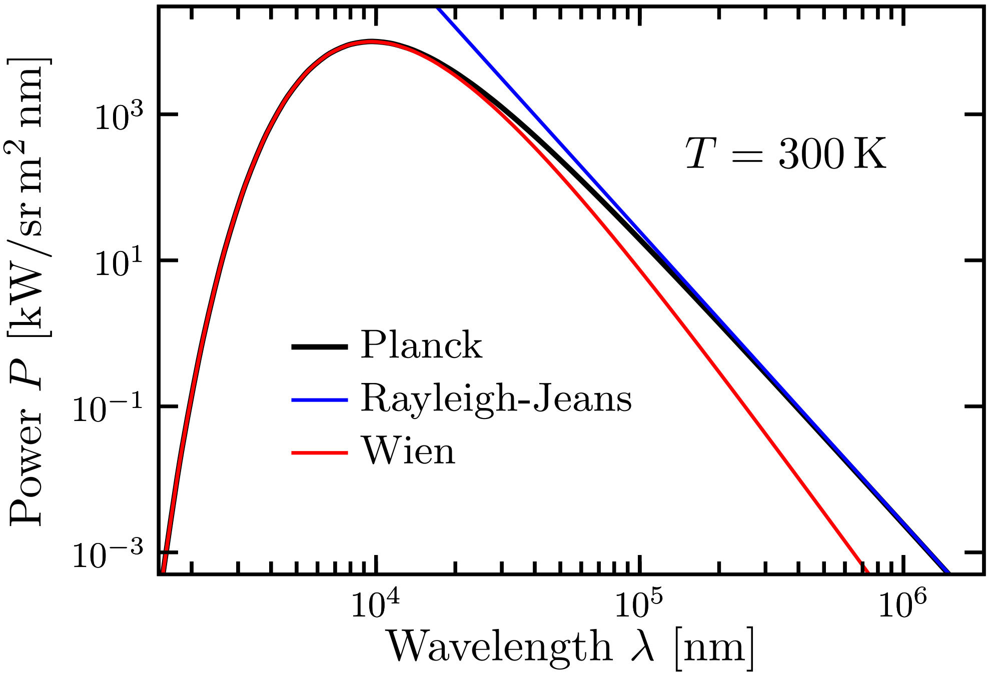
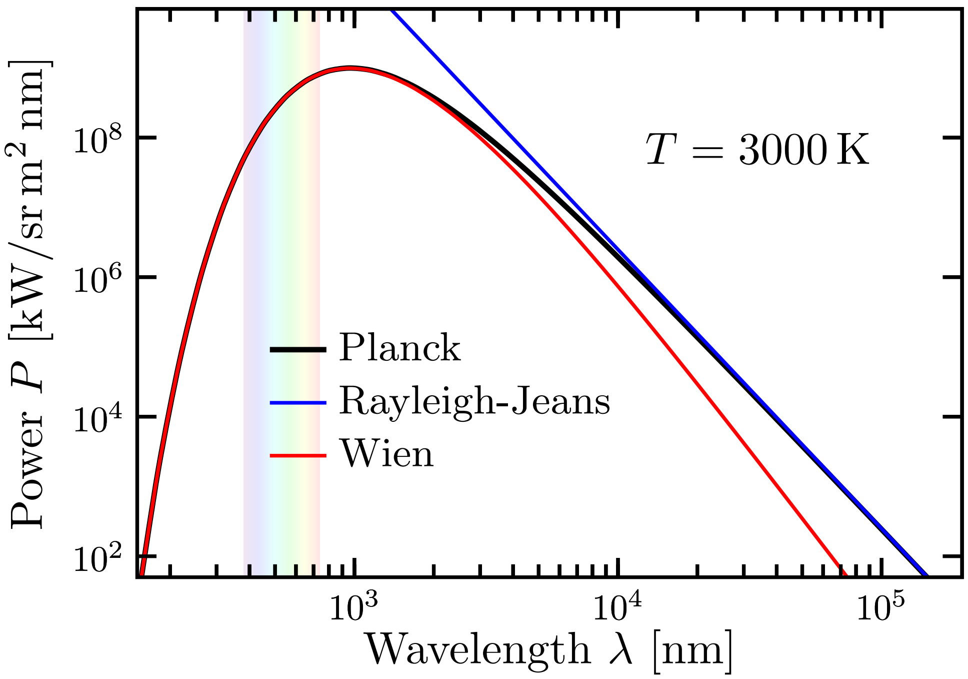
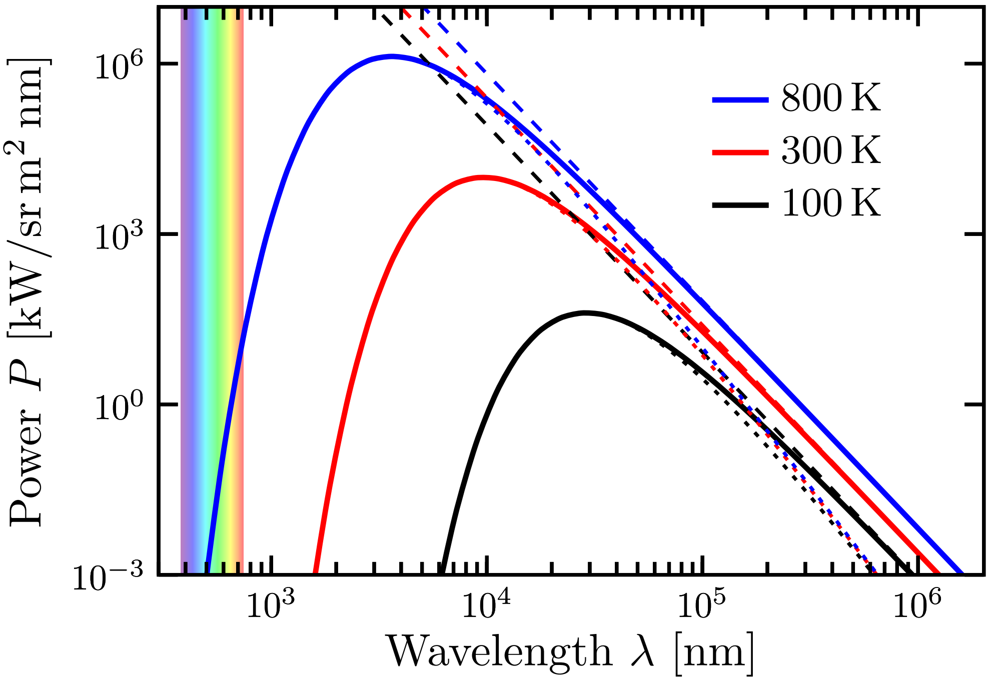


Figure 9.177: **Plot of the three radiation laws (Planck's, Rayleigh-Jeans, Wien) in a log-log scale for comparison (author: Izaak Neutelings)**

And the same kind of plot at $3,000$ Kelvins:



Or for various temperatures:



We see above that the Rayleigh-Jeans was good at low frequencies whereas that of Wien was good at the highs. Historically it took numerous tests and quite a few years to the physicists to find the adequate model that finally led Max Planck to consider the quantification of the states of energy (which he did not realize himself only quite a long time after having elaborated his model).

Another form of the previous diagram only showing the Rayleigh-Jean law and Planck's law in function of wavelength rather than in frequencies is most well known:

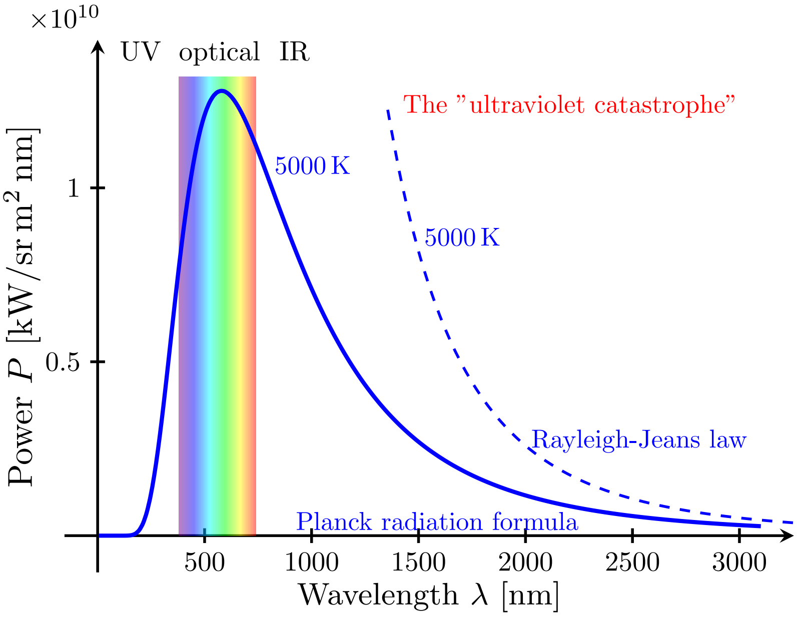


Figure 9.178: Plot of Planck's and Rayleigh-Jeans law (author: Izaak Neutelings)

as it highlights the "ultraviolet catastrophe", also named the "Rayleigh–Jeans catastrophe". Indeed, it was the prediction of late 119th century/early 120th century (holocene calendar) classical physics , based on mathematical developments of Rayleigh and Jeans using the equipartition theorem (see section Continuum Mechanics page 3902), that an ideal black-body at thermal equilibrium will emit radiation in all frequency ranges, emitting more energy as the frequency increases. By calculating the total amount of radiated energy, (i.e., the sum of emissions in all frequency ranges), it can be shown that a black-body would release an infinite amount of energy, contradicting the principles of conservation of energy and indicating that a new model for the behaviour of black bodies was needed.

Remark

Many popular histories of physics, as well as a number of physics textbooks, present an incorrect version of the history of the ultraviolet catastrophe. In that version, the "catastrophe" was first noticed by Planck, who developed his formula in response. In fact Planck never concerned himself with this aspect of the problem, because he did not think that the equipartition theorem (see section Continuum Mechanics page 3902) was fundamental – his motivation for introducing "quanta" was entirely different. That Planck's proposal happened to provide a solution for it was realized only later, as stated above. Though the true sequence of events has been known to historians for many decades, the historically incorrect version persists, in part because Planck's actual motivations for the proposal of the quantum are complicated and difficult to summarize for a lay audience.

We can also prove again the Stefan's law (we have already done this but with another approach) but this time by explaining the origin of the Stefan-Boltzmann constant $\sigma_{\text{S.B.}}$.

Let us first recall that the energy flux (see section Geometric Optics page 4343) is given by:

$$\mathrm{d}\Phi=L(\theta)\mathrm{d}S\mathrm{d}\Omega\cos(\theta) \qquad (9.6.540)$$

Since the luminance depends on the frequency and therefore the temperature of the emitter body, we can add:

$$\mathrm{d}\Phi(\nu,T)=L(\theta,\nu,T)\mathrm{d}S\mathrm{d}\Omega\cos(\theta)\mathrm{d}\nu \qquad (9.6.541)$$

The energy radiated through a given elementary surface $\mathrm{d}S\cos(\theta)$ is therefore:

$$\mathrm{d}\Phi(\nu,T)\mathrm{d}\nu\mathrm{d}t=L(\theta,\nu,T)\mathrm{d}S\mathrm{d}\Omega\cos(\theta)\mathrm{d}\nu\mathrm{d}t \qquad (9.6.542)$$

If the volume of emission is considered as an elementary volume assimilated to a cylinder of height $c\mathrm{d}t$ and of roof having for surface $\mathrm{d}S\cos(\theta)$ (see section Geometric Optics page ??) the energy density per unit of frequency and steradian is then given by:

$$\mathrm{d}U(\nu,T)=\frac{L(\theta,\nu,T)\mathrm{d}S\mathrm{d}\Omega\cos(\theta)\mathrm{d}\nu\mathrm{d}t}{c\mathrm{d}t\mathrm{d}S\cos(\theta)}=\frac{1}{c}L(\theta,\nu,T)\mathrm{d}\Omega\mathrm{d}\nu \qquad (9.6.543)$$

Given the isotropy of the black-body at equilibrium, we have by integrating on the whole of the solid angle the energy density per unit of frequency only:

$$\mathrm{d}U(\nu,T)=\frac{4\pi}{c}L(\theta,\nu,T)\mathrm{d}\nu \qquad (9.6.544)$$

Dimensional analysis gives us:

$$\frac{4\pi}{c}L(\theta,\nu,T)=R(\nu,T)\Leftrightarrow L(\theta,\nu,T)=\frac{c}{4\pi}R(\nu,T) \qquad (9.6.545)$$

Finally, it is useful to consider the total power emitted per unit area (i.e. emittance):

$$\mathrm{d}M(\nu,T)=L(\theta)\mathrm{d}\Omega\cos(\theta) \qquad (9.6.546)$$

If we integrate on the half surface of a sphere (relatively to the surface point of the emitter):

$$\mathrm{d}M(\nu,T)=\int\limits_0^{2\pi}L(\theta)\mathrm{d}\Omega\cos(\theta)=\int\limits_0^{2\pi}L(\theta)\cos(\theta)(2\pi\sin(\theta)\mathrm{d}\theta) \qquad (9.6.547)$$

Indeed for a sphere (see section Trigonometry page 2549):

$$\mathrm{d}\Omega=2\pi\sin(\theta)\mathrm{d}\theta \qquad (9.6.548)$$

Since the luminance is independent of θ (isotropy of the radiation of the black-body), integration is elementary, and we find:

$$M(\nu,T)=\pi L=\pi\frac{c}{4\pi}R(\nu,T)=\frac{c}{4}R(\nu,T)\;[\mathrm{J}\cdot\mathrm{m}^{-2}] \qquad (9.6.549)$$

The total emittance is then given by:

$$\boxed{M(T)=\frac{c}{4}\int\limits_0^{+\infty}R(\nu,T)\mathrm{d}\nu=\frac{c}{4}\int\limits_0^{+\infty}\frac{h\nu^3}{c^3}\frac{1}{e^{\frac{h\nu}{kT}}-1}\mathrm{d}\nu=2\pi\frac{h}{c^2}\int\limits_0^{+\infty}\frac{\nu^3}{e^{\frac{h\nu}{kT}}-1}\mathrm{d}\nu} \qquad (9.6.550)$$

Remark

We will see in the section of Optics (ray optics) at the page 4340, how we can deduce the corresponding number of photons for important calculations about CCD (semiconductor Charge-Coupled Devices) cameras.

By putting $x=h\nu/kT$, we can simplify the integrand so that:

$$M(T)=\frac{2\pi h}{c^2}\frac{(kT)^4}{h^4}\int\limits_0^{+\infty}\frac{x^3}{e^x-1}\mathrm{d}x=\frac{2\pi h}{c^2}\frac{(kT)^4}{h^4}\left(\frac{\pi^4}{15}\right)=\frac{2\pi^5k^4}{15c^2h^3}T^4 \tag{9.6.551}$$

Let us prove, because its far to be easy that:

$$\int\limits_0^{+\infty}\frac{x^3}{e^x-1}\mathrm{d}x=\frac{\pi^4}{15} \tag{9.6.552}$$

Let us write the integrand in the following form:

$$x^3(e^x-1)^{-1}=x^3e^{-x}(e^x-1)^{-1}e^x=x^3e^{-x}(1-e^{-x})^{-1} \tag{9.6.553}$$

For the term:

$$\frac{1}{1-e^{-x}} \tag{9.6.554}$$

Let us put $t=e^{-x}$, then we have:

$$\frac{1}{1-t} \tag{9.6.555}$$

And we have proved in the section of Functional Analysis, using results of the section Sequences and Series, that under the condition $|t|<1$, the last fraction can be written:

$$\frac{1}{1-t}=\sum_{n=0}^{+\infty}t^n=1+t+t^3+\ldots \tag{9.6.556}$$

Since then:

$$x^3e^{-x}(1-e^{-x})^{-1}=x^3e^{-x}(1+e^{-x}+e^{-2x}+e^{-3x}+\ldots)=x^3e^{-x}+x^3e^{-2x}+x^3e^{-3x}+\ldots \tag{9.6.557}$$

Hence:

$$\int\limits_0^{+\infty}\frac{x^3}{e^x-1}\mathrm{d}x=\int\limits_0^{+\infty}\left(x^3e^{-x}+x^3e^{-2x}+x^3e^{-3x}+\ldots\right)\mathrm{d}x \tag{9.6.558}$$

We can therefore replace our integral by a sum of defined integrals. We see that by making integrations by successive parts, we get:

$$\begin{aligned}\int_0^{+\infty} x^3 e^{-nx}\mathrm{d}x &= -\frac{1}{n}\left[x^3e^{-nx}\right]_0^{+\infty}+\frac{1}{n}\int_0^{+\infty}3x^2e^{-nx}\mathrm{d}x=\frac{3}{n}\int_0^{+\infty}x^2e^{-nx}\mathrm{d}x\\ &=-\frac{3}{n^2}\left[x^2e^{-nx}\right]_0^{+\infty}+\frac{3}{n^2}\int_0^{+\infty}2xe^{-nx}\mathrm{d}x=\frac{3}{n^2}\int_0^{+\infty}2xe^{-nx}\mathrm{d}x\\ &=-\frac{6}{n^3}\left[xe^{-nx}\right]_0^{+\infty}+\frac{6}{n^3}\int_0^{+\infty}e^{-nx}\mathrm{d}x=\frac{6}{n^3}\int_0^{+\infty}e^{-nx}\mathrm{d}x=\frac{6}{n^3}\left[-\frac{1}{n}e^{-nx}\right]_0^{+\infty}\\ &=-\frac{6}{n^4}\left[e^{-nx}\right]_0^{+\infty}=-\frac{6}{n^4}\left(e^{-\infty x}-e^{-0x}\right)=-\frac{6}{n^4}(0-1)=\frac{6}{n^4}\end{aligned}\tag{9.6.559}$$

Therefore:

$$\int_0^{+\infty}\frac{x^3}{e^x-1}\mathrm{d}x=6\left(1+\frac{1}{2^4}+\frac{1}{3^4}+\frac{1}{4^4}+\ldots\right)=6\sum_{n=1}^{+\infty}\frac{1}{n^4}\tag{9.6.560}$$

However, we have proved in the section of Sequences and Series in our study of the Riemann zeta function that:

$$\zeta(4)=\sum_{n=1}^{+\infty}\frac{1}{n^4}=\frac{\pi^4}{90}\tag{9.6.561}$$

Therefore:

$$\int_0^{+\infty}\frac{x^3}{e^x-1}\mathrm{d}x=6\sum_{n=1}^{+\infty}\frac{1}{n^4}=6\zeta(4)=\frac{6\pi^4}{90}=\frac{\pi^4}{15}\tag{9.6.562}$$

Finally, the Stefan-Boltzmann constant can be expressed analytically in the form:

$$\boxed{\sigma\cong\frac{2\pi^5k^4}{15c^2h^3}=\frac{\pi^2k^4}{60c^2\hbar^2}=5.67051\cdot10^{-8}\,[\mathrm{W}\cdot\mathrm{m}^{-2}\cdot\mathrm{K}^4]}\tag{9.6.563}$$

Honestly ... it was hard to guess ... It must be notice that since Planck's constant was not introduced until a couple decades after the Stefan-Boltzmann law was developed, one might more appropriately say that the Stefan-Boltzmann constant pins down Planck's constant.

For summary, we have then:

$$\boxed{M(T)=\frac{2\pi^5k^4}{15c^2h^3}T^4=\sigma T^4\qquad\text{and}\qquad\rho(T)=\frac{4}{c}M(T)=\frac{8\pi^5k^4}{15c^3h^3}T^4=\sigma_{\text{S.B.}}T^4}\tag{9.6.564}$$

where the relation on the left is the famous "Stefan-Boltzmann law".

Let us determine now for what frequency we have the maximum energy density. In other words, this is equivalent to seek where the derivative:

$$\frac{\mathrm{d}R(\nu,T)}{\mathrm{d}\nu}=\frac{\mathrm{d}}{\mathrm{d}\nu}\left(8\pi\frac{h\nu^3}{c^3}\frac{1}{e^{\frac{h\nu}{kT}}-1}\right) \tag{9.6.565}$$

cancels. Therefore:

$$\frac{\mathrm{d}R(\nu,T)}{\mathrm{d}\nu}=\frac{8\cdot 3\pi h\nu^3}{c^3}\frac{1}{e^{\frac{h\nu}{kT}}-1}-8\pi\frac{h\nu^3}{c^3}\frac{h}{kT}\frac{e^{\frac{h\nu}{kT}}}{\left(e^{\frac{h\nu}{kT}}-1\right)^2} \tag{9.6.566}$$

Let us divide by $R(\nu,T)$:

$$\begin{aligned}&\frac{1}{R(\nu,T)}\frac{\mathrm{d}R(\nu,T)}{\mathrm{d}\nu}=\frac{3}{\nu}-\frac{h}{kT}\frac{e^{\frac{h\nu}{kT}}}{e^{\frac{h\nu}{kT}}-1}=0\Rightarrow\frac{3}{\nu}e^{\frac{h\nu}{kT}}-\frac{3}{\nu}-\frac{h}{kT}e^{\frac{h\nu}{kT}}=0\\&\Rightarrow\frac{3}{\nu}-\frac{3}{\nu}e^{-\frac{h\nu}{kT}}-\frac{h}{kT}=0\Leftrightarrow e^{-\frac{h\nu}{kT}}=1-\frac{1}{3}\frac{h\nu}{kT}\end{aligned} \tag{9.6.567}$$

The last relation admits a single positive root that we can determine with Maple™ 4.00b using the command:

```
>evalf(solve(exp(-x)-1+1/3*x=0,x));
```

$$\frac{h\nu}{kT}=2.82144 \tag{9.6.568}$$

What gives us the "Wien's second law" or "Wien's law of displacement" which says that as the temperature of a black-body increases, the wavelength at which the intensity of the radiation Is the strongest becomes shorter and shorter:

$$\boxed{\nu_{\max}=2.82144\frac{kT}{h}=a\frac{kT}{h}} \tag{9.6.569}$$

where a is named the "Wien's constant". The latter is often rearranged and condensed as following:

$$T=\frac{2,897.8\cdot 10^{-6}\,[\mathrm{m}\cdot\mathrm{K}]}{\lambda_{\max}} \tag{9.6.570}$$

The figure below shows this result using the logarithmic scales:

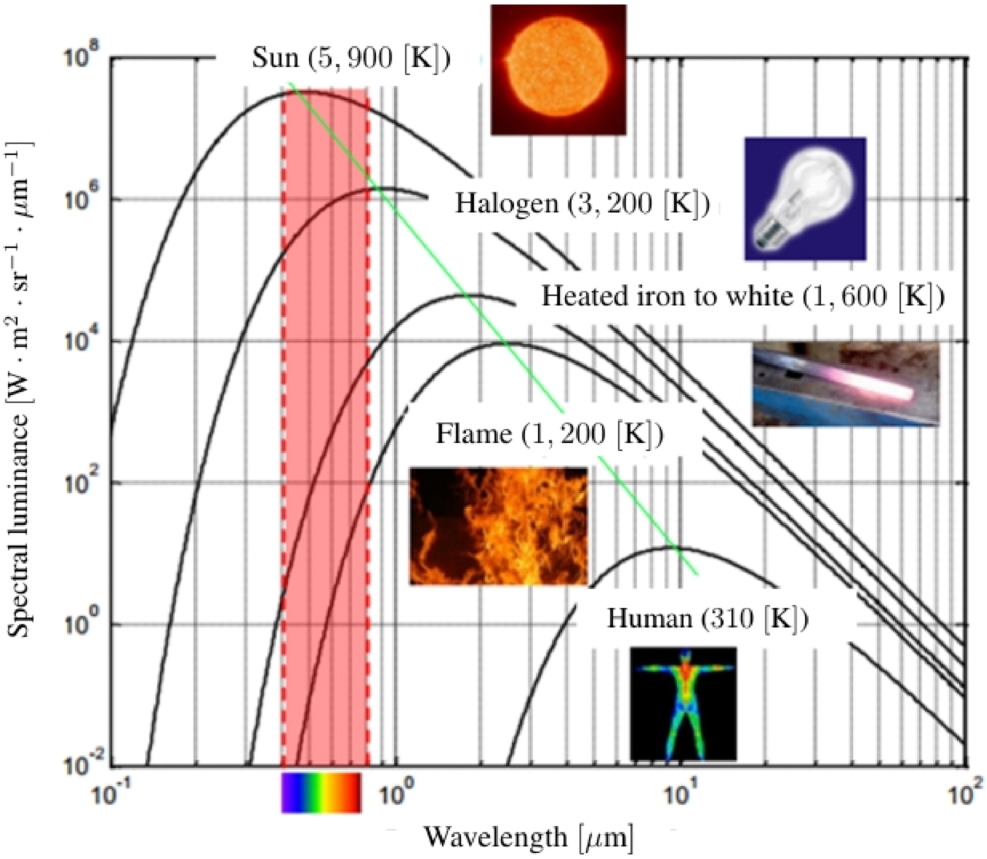


Figure 9.179: Representation of Planck's law (distribution) with Wien's displacement

Thus, not only does the increase in temperature lead to an increase in the total amount of radiated energy but the Wien displacement law tells us that the wavelength at which the maximum amount of radiation is emitted multiplied by the temperature of the black-body is always a constant (or otherwise seen: the maximum wavelength is inversely proportional to the temperature as shown in the green line in the figure above)! This is an extremely simple result that tells us two important things:

1. If the temperature doubles, then the maximum wavelength will be half the previous wavelength.
2. If we know the constant (which is our case) then we can calculate the maximum wavelength for any temperature at which a black-body can be found.

It is easy to understand now why any material at low temperatures emits mainly long wavelength radiation in the infra-red part of the spectrum and as the temperature rises, there is more radiated energy in each region of the spectrum and the wavelength peak decreases as it moves toward shorter wavelengths. As a result, the color of the emitted light changes from red to orange, then yellow to bluish-white (and further into the ultraviolet).

Remark

Obviously the second Wien's law is also sometimes given in the literature not in function of the frequency, but to in function of the wavelength ...

Let us insist on the fact that the Planck's law is valid only in the cases where the radiation is in thermal equilibrium. This restriction is important in practice since the phenomena of emission or absorption of radiation by the material occur most often under conditions out of equilibrium: in the case, for example, of lighting by an electric lamp or electric heating by infra-red radiation, the is an irreversible (and therefore out of equilibrium) transformation of electrical energy into radiation energy. Similarly, solar radiation is produced by the nuclear reactions which take place inside the Sun (actually a G2V-type star, aka yellow dwarf) and which gradually consume its substance; At the microscopic level also, the emission of a photon by an excited atom is very often an irreversible return of the atom to its fundamental state (spontaneous emission out of equilibrium). In the case of the black-body, on the contrary, the radiation is confined within a closed enclosure (we eventually leave a negligible fraction of this radiation to escape outside to be subjected to the measurements) and we can thus achieve thermal equilibrium with the walls.

Planck's law, which we have demonstrated above, is perfectly verified by experience in the whole range of temperatures accessible at this date for a black-body and remains a quite good approximation also for Stars and other objects that are not at equilibrium:

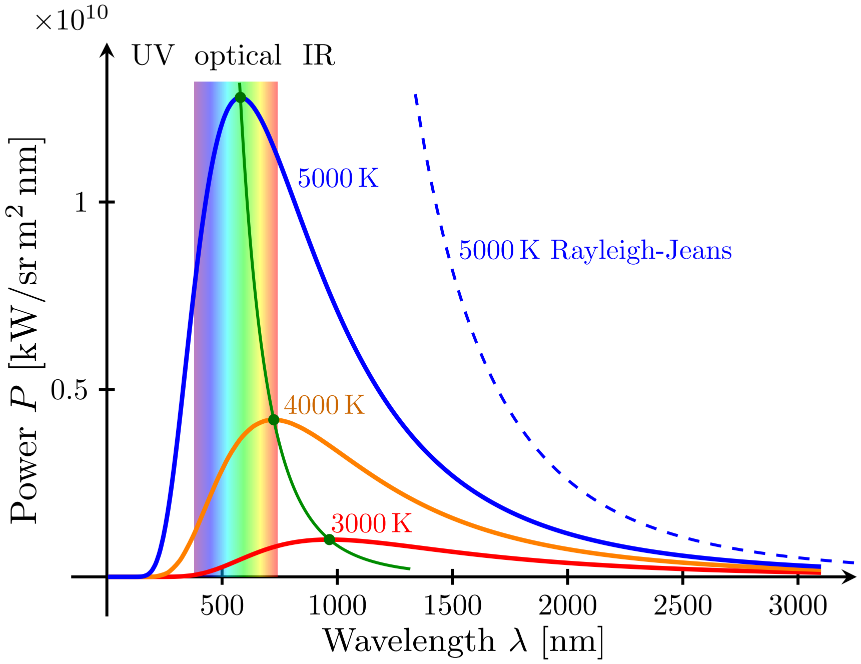


Figure 9.180: Planck's law at various temperatures in function of wave length (source: ?)

We notice from the graph above that a body heated above $5,000$ [K] has an emission peak in the middle of the visible spectrum. In the field of colorimetry, we associate a temperature with a color by looking for the temperature of the black-body for which the peak of radiation has its maximum in the wavelength of the given color. Another typical representation (and more complete) for the Sun is the following (TAO on this figure means: Top Of Atmosphere):

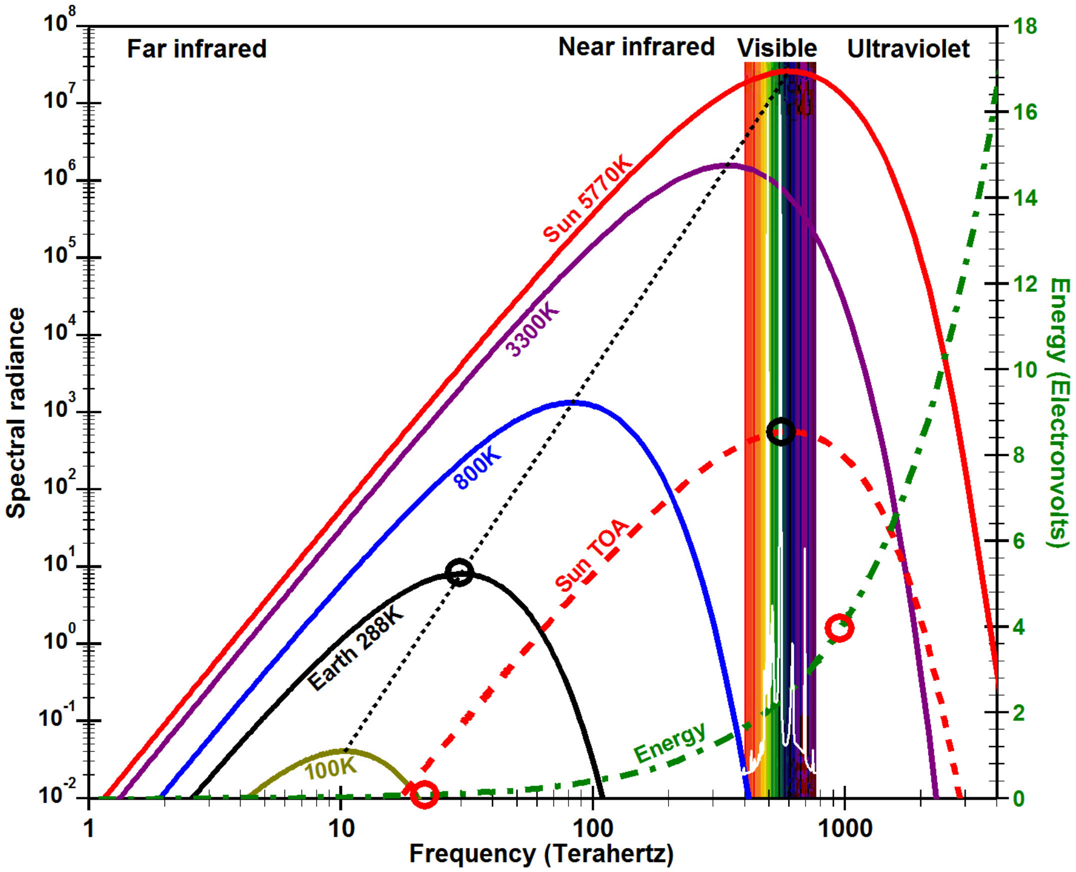


Figure 9.181: Sun's radiation spectrum in function of frequency and logarithmic scale (source: ?)

Or a more detailed Sun spectrum at sea level:

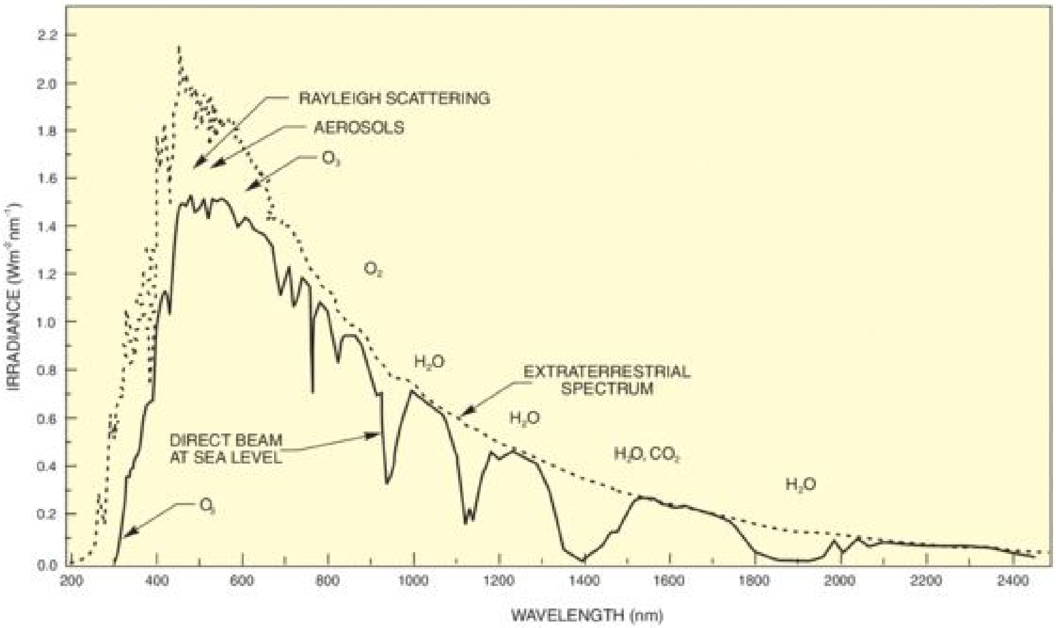


Figure 9.182: Sun's radiation spectrum in function of frequency at sea level (source: ?)

It should be notice that many light sources emit a luminous flux that does not follow the law of the black-body (a bulb filament, for example) and that the Wien law does not apply to them. On the other hand, it remains true that they emit at a wavelength all the shorter as they are hot.

It should also be keep in mind that the luminous flux coming from an object is not necessarily of a thermal nature! In other words its color does not always indicate its temperature. For example, the color of the sky comes from blue sunlight diffused by the air (Rayleigh scattering), and not from a hypothetical temperature of $15,000$ [K]. Similarly a tree is green, not because it is $8,000$ [K] hot, but because it reflects the green light that makes up the light of day...

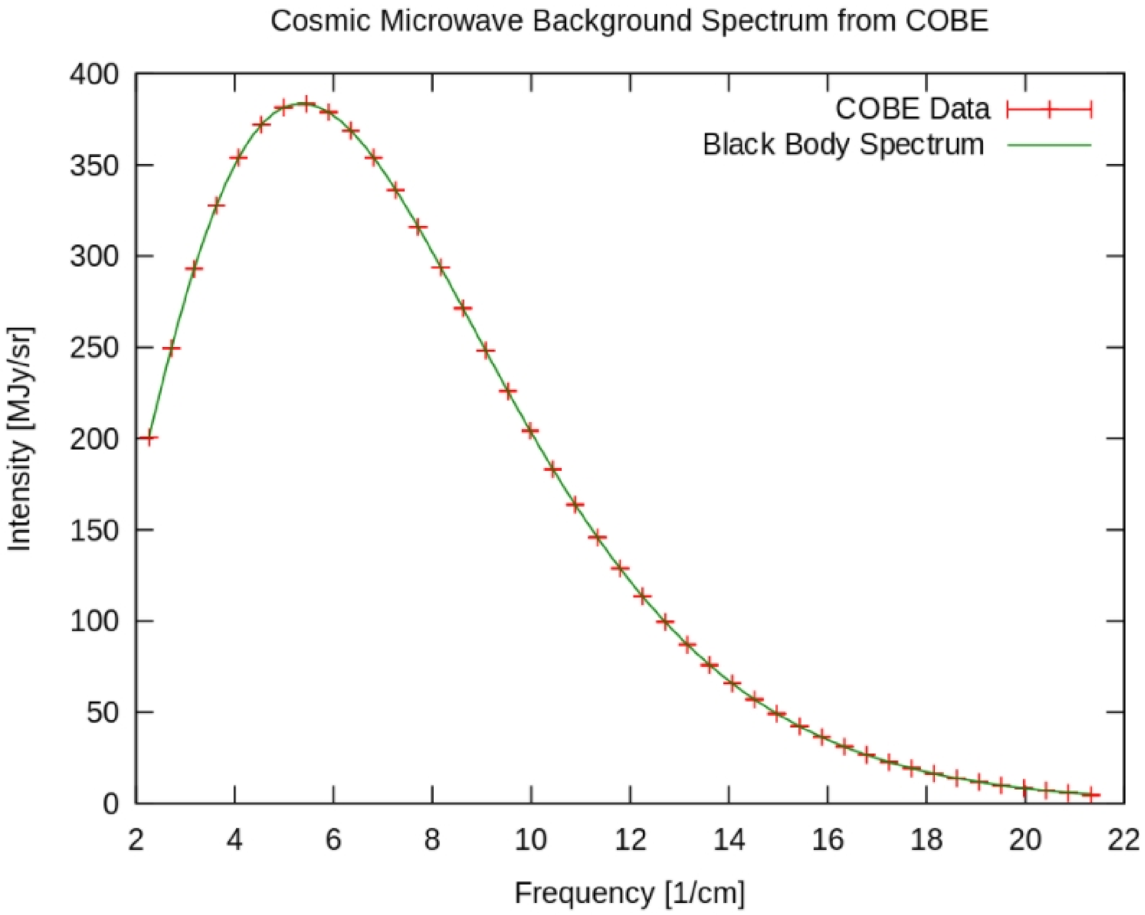


Figure 9.183: Comparison of theoretical black-body and our Universe cosmic microwave background (source: Wikipedia)

9.6.12.2 Summing up all types of heat transfer

So we have seen earlier above the definition that the rate at which the temperature of a solid object of mass m [kg] and specific heat capacity C_p $[\mathrm{J}\cdot\mathrm{kg}^{-1}\cdot\mathrm{K}^{-1}]$ will change $\mathrm{d}T/\mathrm{d}t$ $[\mathrm{K}\cdot\mathrm{s}^{-1}]$ is written as below (see page 3684):

$$mC_p\frac{\mathrm{d}T}{\mathrm{d}t}=\dot{q} \tag{9.6.571}$$

The term $\dot{q}\,[\mathrm{J}\cdot\mathrm{s}^{-1}]$ is the rate at which heat enters (to heat up) or leaves (to cool down) the object.

To start, let us assume the object is initially at T_o and the surroundings is held at constant T_s.

Let us enumerate the different types of heat flows we have seen so far:

1. Heat can enter/leave the object by radiation to a surrounding fluid (gas or liquid). The radiation rate is proportional to the fourth power of absolute temperature $\dot{q}_r \propto T^4$. The initial overall rate of heat transfer is expressed to first order as below (see page 3706):

   $$\dot{q}_{rT} \propto \left|T_o^4 - T_s^4\right|$$

   We can use this to determine the initial temperatures of the object where the cold object receives more radiation heat flow from the surroundings than the hot object looses by radiation heat flow to the surroundings. Thus, we can determine for which cases a cold or hot object changes temperature faster than a hot object in the same air temperatures.

2. Heat can enter/leave the object by convection when the surroundings is a fluid (gas or liquid). The rate of heat transfer is proportional to temperature difference. The net rate of heat transfer by convection is determined as below (see page 3635):

   $$\dot{q}_{hT} \propto |T_o - T_s|$$

   We learn from this that a hot or cold object will cool or heat by convection at the same rate as long as the temperature difference between the object and the surrounding fluid is the same.

3. Finally, heat can enter/leave the object to the surroundings by conduction in the surroundings. This happens with a temperature gradient $\mathrm{d}T_s/\mathrm{d}z\,[\mathrm{K}\cdot\mathrm{m}^{-1}]$ in the surroundings. A fluid does not typically sustain a temperature gradient well; the gradient dissipates and the flow is by convection. Therefore making a statement that you want to consider conduction in the surroundings is akin to saying that you define the surroundings also as a solid. The net rate of conductive flow is written as below (see page 3684):

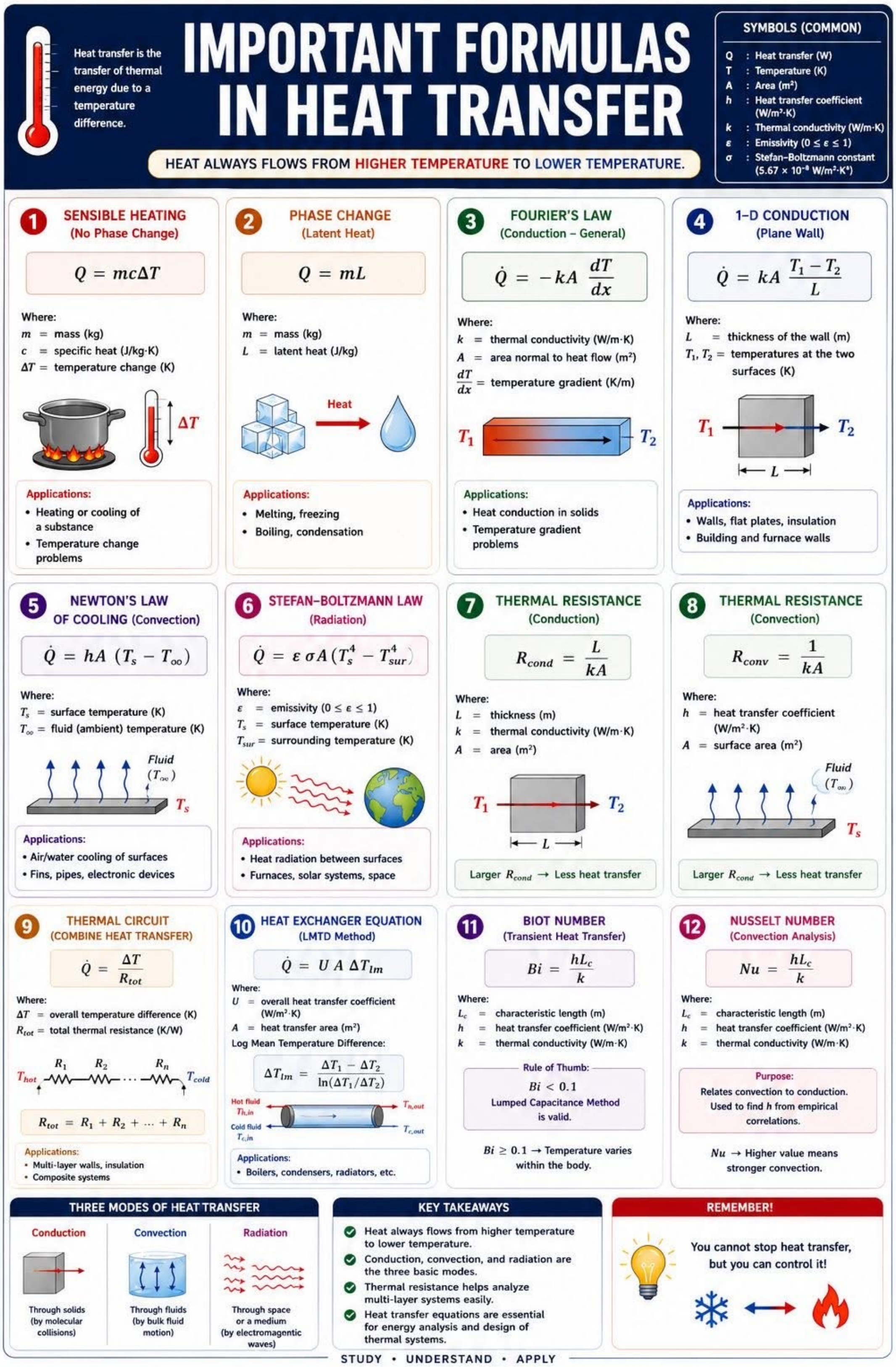
   $$\dot{q}_{kT} \propto |\mathrm{d}T_s/\mathrm{d}z|$$

Combining we get:

$$\boxed{mC_p\frac{\mathrm{d}T_o}{\mathrm{d}t} = \sigma S\left(\varepsilon_s T_s^4 - \varepsilon_o T_o^4\right) + h_s A\left(T_s - T_o\right) - k_s A\frac{\mathrm{d}T_s}{\mathrm{d}z}} \tag{9.6.572}$$

The various terms include the Stefan-Boltzmann factor σ, object area S, convection coefficient h_s, and thermal conductivity k_s. This equation defines the rate of temperature change of the object at any point in time. The $/pm$ signs are set so that, when the object is hotter than the surroundings, the object will cool down (and vice-versa). It is a rather unwieldy equation that is typically solved by considering only one of the three cases (radiation, convection, or conduction) at a time. It also presumes the object heats or cools uniformly throughout. The case when we have a thermal gradient in the solid object is yet another equation.

As a general rule, the rate at which a hot object changes temperature (cools down) cannot be stated to be faster (or slower) than the rate at which a cold object changes temperature (heats up). The rate of heat transfer depends on the type of surroundings (vacuum, fluid, or solid), the temperature of the surroundings, and the type of heat transfer (radiation, convection, or conduction)!



Heat transfer is the transfer of thermal energy due to a temperature difference.

IMPORTANT FORMULAS IN HEAT TRANSFER

HEAT ALWAYS FLOWS FROM HIGHER TEMPERATURE TO LOWER TEMPERATURE.

SYMBOLS (COMMON)

Q : Heat transfer (W)
T : Temperature (K)
A : Area (m²)
h : Heat transfer coefficient (W/m²·K)
k : Thermal conductivity (W/m·K)
ε : Emissivity ($0 \le \varepsilon \le 1$)
σ : Stefan–Boltzmann constant (5.67×10^{-8} W/m²·K⁴)

① SENSIBLE HEATING (No Phase Change)

$$Q = mc\Delta T$$

Where:
m = mass (kg)
c = specific heat (J/kg·K)
ΔT = temperature change (K)

Applications:
- Heating or cooling of a substance
- Temperature change problems

② PHASE CHANGE (Latent Heat)

$$Q = mL$$

Where:
m = mass (kg)
L = latent heat (J/kg)

Applications:
- Melting, freezing
- Boiling, condensation

③ FOURIER'S LAW (Conduction – General)

$$\dot{Q} = -kA\,\frac{dT}{dx}$$

Where:
k = thermal conductivity (W/m·K)
A = area normal to heat flow (m²)
$\frac{dT}{dx}$ = temperature gradient (K/m)

Applications:
- Heat conduction in solids
- Temperature gradient problems

④ 1–D CONDUCTION (Plane Wall)

$$\dot{Q} = kA\,\frac{T_1 - T_2}{L}$$

Where:
L = thickness of the wall (m)
T_1, T_2 = temperatures at the two surfaces (K)

Applications:
- Walls, flat plates, insulation
- Building and furnace walls

⑤ NEWTON'S LAW OF COOLING (Convection)

$$\dot{Q} = hA\,(T_s - T_{oo})$$

Where:
T_s = surface temperature (K)
T_∞ = fluid (ambient) temperature (K)

Applications:
- Air/water cooling of surfaces
- Fins, pipes, electronic devices

⑥ STEFAN–BOLTZMANN LAW (Radiation)

$$\dot{Q} = \varepsilon\,\sigma A\,(T_s^4 - T_{sur}^4)$$

Where:
ε = emissivity ($0 \le \varepsilon \le 1$)
T_s = surface temperature (K)
T_{sur} = surrounding temperature (K)

Applications:
- Heat radiation between surfaces
- Furnaces, solar systems, space

⑦ THERMAL RESISTANCE (Conduction)

$$R_{cond} = \frac{L}{kA}$$

Where:
L = thickness (m)
k = thermal conductivity (W/m·K)
A = area (m²)

Larger R_{cond} → Less heat transfer

⑧ THERMAL RESISTANCE (Convection)

$$R_{conv} = \frac{1}{kA}$$

Where:
h = heat transfer coefficient (W/m²·K)
A = surface area (m²)

Larger R_{cond} → Less heat transfer

⑨ THERMAL CIRCUIT (COMBINE HEAT TRANSFER)

$$\dot{Q} = \frac{\Delta T}{R_{tot}}$$

Where:
ΔT = overall temperature difference (K)
R_{tot} = total thermal resistance (K/W)

$$R_{tot} = R_1 + R_2 + \dots + R_n$$

Applications:
- Multi-layer walls, insulation
- Composite systems

⑩ HEAT EXCHANGER EQUATION (LMTD Method)

$$\dot{Q} = U\,A\,\Delta T_{lm}$$

Where:
U = overall heat transfer coefficient (W/m²·K)
A = heat transfer area (m²)

Log Mean Temperature Difference:

$$\Delta T_{lm} = \frac{\Delta T_1 - \Delta T_2}{\ln(\Delta T_1/\Delta T_2)}$$

Applications:
- Boilers, condensers, radiators, etc.

⑪ BIOT NUMBER (Transient Heat Transfer)

$$Bi = \frac{hL_c}{k}$$

Where:
L_c = characteristic length (m)
h = heat transfer coefficient (W/m²·K)
k = thermal conductivity (W/m·K)

Rule of Thumb:
$Bi < 0.1$
Lumped Capacitance Method is valid.

$Bi \ge 0.1$ → Temperature varies within the body.

⑫ NUSSELT NUMBER (Convection Analysis)

$$Nu = \frac{hL_c}{k}$$

Where:
L_c = characteristic length (m)
h = heat transfer coefficient (W/m²·K)
k = thermal conductivity (W/m·K)

Purpose:
Relates convection to conduction. Used to find h from empirical correlations.

Nu → Higher value means stronger convection.

THREE MODES OF HEAT TRANSFER

Conduction	Convection	Radiation
Through solids (by molecular collisions)	Through fluids (by bulk fluid motion)	Through space or a medium (by electromagentic waves)

KEY TAKEAWAYS

- ✓ Heat always flows from higher temperature to lower temperature.
- ✓ Conduction, convection, and radiation are the three basic modes.
- ✓ Thermal resistance helps analyze multi-layer systems easily.
- ✓ Heat transfer equations are essential for energy analysis and design of thermal systems.

REMEMBER!

You cannot stop heat transfer, but you can control it!

STUDY • UNDERSTAND • APPLY

~90 %

★★★☆☆

71 votes, 66.76%

9.7. Continuum Mechanics

IN a strict sense, Continuum Mechanics is the branch of mechanics that has for purpose the study of motion, deformation and stress fields inside continuum materials whether they are solids, liquids, gases or plasmas.

Modelling an object as a continuum assumes that the substance of the object completely fills the space it occupies. Modelling objects in this way ignores the fact that matter is made of atoms, and so is not continuous; however, on length scales much greater than that of inter-atomic distances, such models are highly accurate. Fundamental physical laws such as the conservation of mass, the conservation of momentum, and the conservation of energy may be applied to such models to derive differential equations describing the behaviour of such objects, and some information about the particular material studied is added through constitutive relations.

Continuum Mechanics deals with physical properties of solids and fluids which are independent of any particular coordinate system in which they are observed. These physical properties are then represented by tensors, which are mathematical objects that have the required property of being independent of coordinate system. These tensors can be expressed in coordinate systems for computational convenience.

Definitions (#9.7.467):

D1. We name "medium", any fluid (solid, liquid, gas or plasma depending on what we have seen in the section of Thermodynamics), deformable or not, when we consider it a macroscopic point of view, as opposed to a corpuscular description.

D2. We name "continuous medium", a medium such that if M and M' belong to a medium and M' belongs to the neighbourhood of M, then whatever the deformation undergone by that environment, $\mathrm{d}M'$ will still belong to the neighbourhood of $\mathrm{d}M$.

This field is often seen as the engineering science for understanding and describing the material world around us and currents phenomena that take place: liquid movements, gas, flying airplanes, helicopters, rockets, satellites, sailing boats, solids deformations, internal structure of stars, etc. By its relations with thermal mechanics (thermodynamics), it extends to the thermal analysis, energy analysis, acoustic and even General Relativity (the Universe being considered as a continuum material on long distances)!

Taking into account the behaviour of continuum, it encompasses hydrodynamics, gas dynamics, elasticity, acoustics, plasticity and other behaviours. It is the key to what we now name "modelling", which is nothing but the art of analysing a physical phenomenon and describe it mathematically, allowing their study with rigour specific to this discipline.

This section of the book is divided into four main parts: solid, liquid, gas and plasma (some concepts have been deliberately developed in the section of Music Mathematics page 8195). In each part, we introduce the specific mathematical tools to study a particular continuum with a complexity (all relative) growing. However, by choice and as always in this textbook, it was decided to present the theorems with the simplest possible mathematical tools but all arriving at the same results. For example, the proof of the Navier-Stokes equation that would take 150 pages of rigorous mathematical developments do takes no more than 27. There is therefore a significant advantage both for the writer (...) and for the reader do so.

Remark

Regarding the Navier-Stokes, we also give practical closed-forms examples of these in our study of Meteorology (see section Marine & Weather Engineering page 8239) and when we will have time we will give an easy example of a liquid wave simulation in MATLAB® in the section of Theoretical Computing.

9.7.1 Rigid Bodies

Atoms of the same element or of different elements are assembled in specific structures. This conditions the strength of their electrical interactions, which define the final structure of the substance. Under normal conditions on our planet, matter exists in solid, liquid, gas or plasma. If the inter-atomic forces are sufficiently intense, the particle collection retains its shape and volume.

This property of retaining the shape and volume, as well as elastic properties distinguish solids from other states of matter.

9.7.1.1 Pressures

The concepts of "compression" and "constraint" (that we can include improperly into the global term "pressure") are of prime importance in Continuum Mechanics. It is therefore necessary to define the different types of pressure with a minimum of rigour!

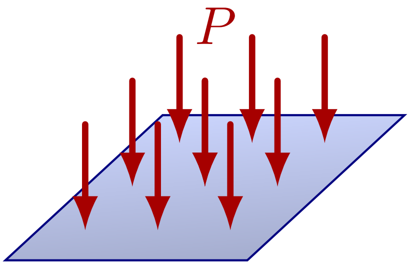
Definitions (#9.7.468):

D1. We name "compression pressure" traditionally denoted P, the ratio expressed by the force F which is exerted (applied) on a surface element S to the normal thereof. Thus, in scalar form:

$$\boxed{P=\frac{F}{S}} \tag{9.7.1}$$

definition that is valid only if This is incorrect if the force being applied is perpendicular to the surface.

Pressure in reality is there even if no surfaces or forces are present! In the middle of a gas bottle, gas has pressure. You can imagine putting a virtual surface there, and no matter how you orient it, you'd get the force of same magnitude, but that's just a side-effect.



Remark

If a force acts on a finished surface, then we also speak of "distributed force".

D2. We name "constraint pressure" the ratio expressed by the force F that pulls on a surface element S not necessarily perpendicularly, traction force which can then be divided into two respectively tangent and normal vectors. So as into vector form:

$$\boxed{\vec{\sigma}+\vec{\tau}=\frac{\mathrm{d}\vec{F}}{\mathrm{d}S}=\vec{T}} \tag{9.7.2}$$

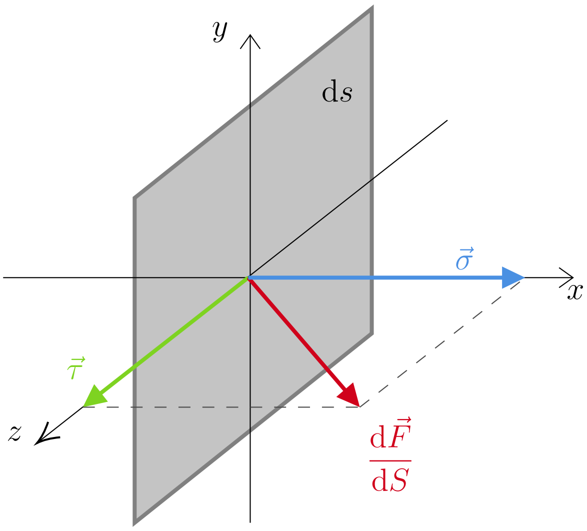


Figure 9.184: Illustration of normal and tangential constraints (stresses)

where $\vec{\sigma}$ and $\vec{\tau}$ are respectively the "normal constraint" and the "tangential constraint" (also sometimes denoted with an S as indices to indicates that this is relatively to a surface).

Both variables above are measured in $[\mathrm{N}\cdot\mathrm{m}^{-2}]$ or "Pascals [Pa]" and by definition (as already mentioned during our study of Units), $100,000$ [Pa] is equal 1 [bar].

Conversion factors for pressure measurement units							
	1MPa	1 at	1psi	1 Pascal	1 bar	1 atm	1 Torr
MPa $[\text{N/mm}^2]$	1	10.2	$145,038$	10^6	10	9.87	$7,500$
at $[\text{kg/cm}^2]$	0.0981	1	14.2	$98,100$	0.981	0.968	736
psi $[\text{lbf/in}^2]$	$6.89\cdot 10^{-3}$	$7.02\cdot 10^{-2}$	1	$6,890$	$6.89\cdot 10^{-2}$	$6.8\cdot 10^{-2}$	51.80
Pa $[\text{N/m}^2]$	10^{-6}	$1.02\cdot 10^{-5}$	$1.45\cdot 10^{-4}$	1	10^{-5}	$9.87\cdot 10^{-6}$	0.0075
bar $[\text{daN/cm}^2]$	0.1	1.02	14.5	$100,000$	1	0.987	752
atm [760 torr]	0,1013	1,0332	14,7	101325	1,013	1	760
torr	$1.33\cdot 10^{-4}$	$1.36\cdot 10^{-3}$	$1.93\cdot 10^{-2}$	133	0.00133	0.00132	1

Table 9.33: **Conversion factors for pressure measurement units**

Keep in mind that the standard atmosphere at the ground level has a pressure of $101,325\;[\text{N}\cdot\text{m}^{-2}]$ and in this early 121st century (according to holocene calendar) the most powerful industrial press belongs to GIVA's Forgiatura a Vienna facility in Italy, with an estimated mass equivalent of $100,000$ tons on a surface of maximum $90\;[\text{m}^2]$ hence a pressure of $\sim 10^7\;[\text{N}\cdot\text{m}^{-2}]$ (one hundred times the atmospheric pressure) or on a typical working surface of $1\;[\text{m}^2]$ this gives a pressure of $\sim 10^8\;[\text{N}\cdot\text{m}^{-2}]$ (one thousand times the atmospheric pressure):

Figure 9.185: **Hydraulic press owned by the GIVA's Forgiatura in Italy**

We could very well include the above two definitions in only one and work with the signs of the forces. But to be consistent with what is taught in schools, we will keep these two definitions that identify by definition the fact that the forces are opposed relative to a surface element S.

Remark

We will give the definition of hydrostatic pressure further below during our study of liquids.

Caution! A common mistake is to think that pressure is a vector... Indeed strictly speaking, the pressure p is proportionality factor defined by:

$$\vec{F}=p\vec{S}=p\vec{n}S \qquad (9.7.3)$$

where the surface $\vec{S}$ is the one that gives the direction. We have to recall that pressure is defined everywhere in the bulk volume, not just at the surface. A volume of gas has pressure defined everywhere. And the force direction is determined by us - by the way we orient your surface that you put into the gas.

Therefore it is incorrect (although rather usual) to say "*the pressure is directed in such or such direction*". The pressure, as a scalar, has no direction. The force given by the previous relationship to the quantity has a direction, but the pressure does not. If we change the orientation of the surface element, the direction of the normal force changes accordingly, but the pressure remains the same.

If $\vec{F}$ is just caused by the pressure, it can't be anything else but perpendicular to the area, otherwise you have other forces present in the system, or the liquid is not isotropic.

But as reality is always more complex than it seems, we will se later that in reality pressure is a tensor (the "Cauchy stress tensor" and then pressure is the average of the stress tensor over all orientations and in a fluid which isn't moving, this is the only stress that is allowed), but for gasses, it's isotropic, so it acts as a scalar.

9.7.1.2 Elasticity of Solids

In one way or another, a compressive or tension constraint can deform the triple: height, width, thickness of a body. Directly address to the study of a case that deform these three parameters is a bit long and will be discussed further below in the part on the determination of the expression of Young's shear modulus.

But it could be useful, at least only from the point of view of the vocabulary to give an example from the most simplistic case in can be. If we imagine a one dimension elastic body (having neither height nor width, but just a length) under the application of two perfectly collinear constraints but antagonists forces, we can imagine that the considerated body is lengthen from a given factor.

Definition (#9.7.469): The "normal deformation" under axial and antagonists forces is given by the ratio between the change in body length to its original length (i.e.: the relative elongation) such that:

$$\boxed{\varepsilon=\frac{|\Delta L|}{L_0}} \qquad (9.7.4)$$

This relation is an extremely simplified form of all kinds of deformations that may exist. Indeed, when we rigorously studies the deformation, we must take into account the accumulated distortions and then we speak of "true longitudinal deformation", or even of "longitudinal rational deformation", by the change in length such that:

$$\mathrm{d}\varepsilon=\frac{\mathrm{d}L}{L} \tag{9.7.5}$$

Therefore after integration:

$$\varepsilon=\int_{L_0}^{L}\frac{\mathrm{d}L}{L}=\ln(L)|_{L_0}^{L}=\ln\left(\frac{L}{L_0}\right) \tag{9.7.6}$$

We can consider in all generality that:

$$L=L_0+\Delta L \tag{9.7.7}$$

Therefore it comes (the same reasoning applying for a surface or a volume)

$$\varepsilon=\ln\left(\frac{L_0+\Delta L}{L_0}\right)=\ln\left(1+\frac{\Delta L}{L_0}\right) \tag{9.7.8}$$

But we see above a well-known expression of usual Taylor series (see section Sequences and Series page 1703) and then we can write for small deformations:

$$\varepsilon\cong\frac{\Delta L}{L_0} \tag{9.7.9}$$

and we fall back on our initial naive definition. We also see that this approximation is also equivalent to put from start that:

$$\mathrm{d}\varepsilon=\frac{\mathrm{d}L}{L}\cong\frac{\Delta L}{L_0} \tag{9.7.10}$$

which is a writing (approximation) that we will see again in the section of Thermodynamics in our determination of the liquid state equation but with the volume V instead of the length L.

There is necessarily a relation between compression and tensile forces and the variation in size of a body. This relation is dependent on the atomic structure of the material and should rigorously use quantum physics to be determined (we will omit this approach in the section of this textbook). We observe, however, depending on the materials various characteristics that interest a the highest point engineers:

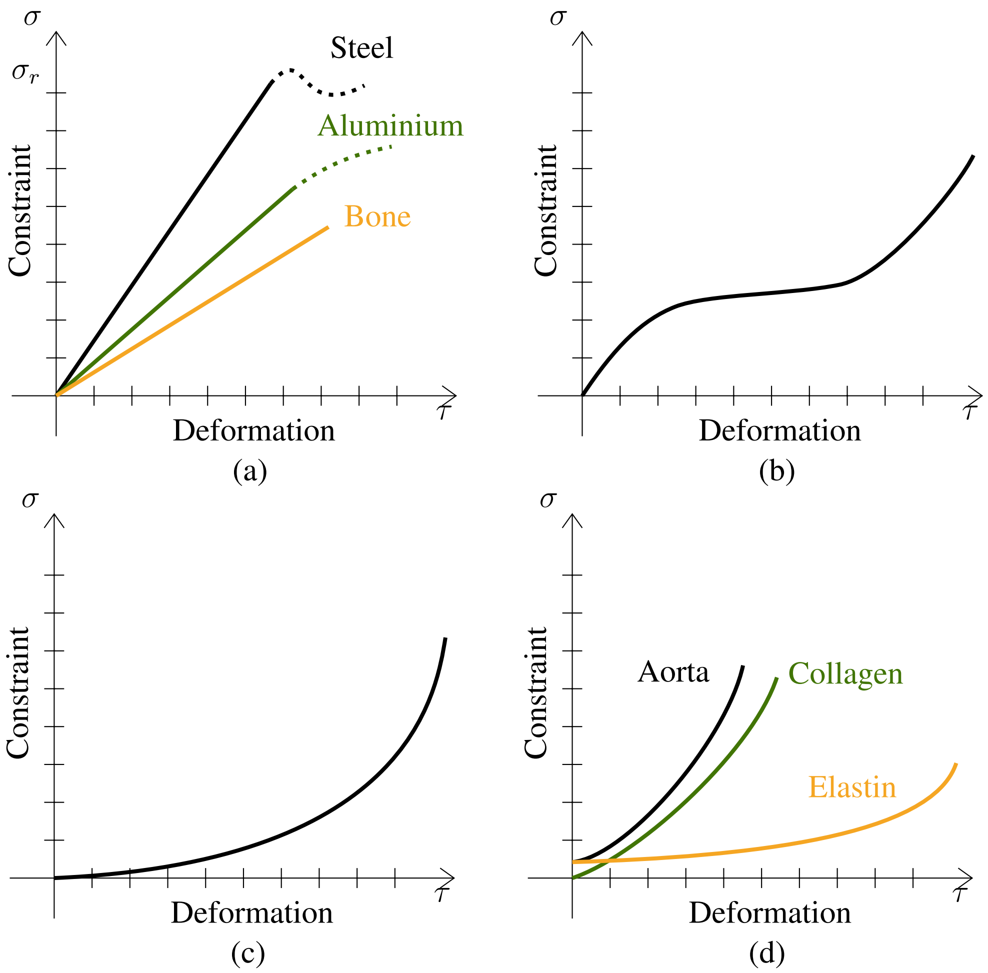


Figure 9.186: **Behaviour under stress/compression for some materials**

The above figures represent the variation of the compressive stress versus strain stress for some materials (typically we represent these characteristics by inverting the axes). We can observe that:

- Ductile materials such as mild steel (a) ceases to be linear in the elasticity limit denoted σ_r above.
- Under traction, rubber polymers (b) first lengthen their molecules and pulling their chemical bonds (see section Quantum Chemistry page 6190).
- Most organic materials (c) are under constraint, even when if they are not distorted. The skin, for example, is as a big rubber glove enveloping the body.
- Elastin (d) is usually enhanced with collagen in biological systems such as arteries. A tendon is primarily made of collagen.

In a more general case, the engineers used to define the points shown below in their material constraint:

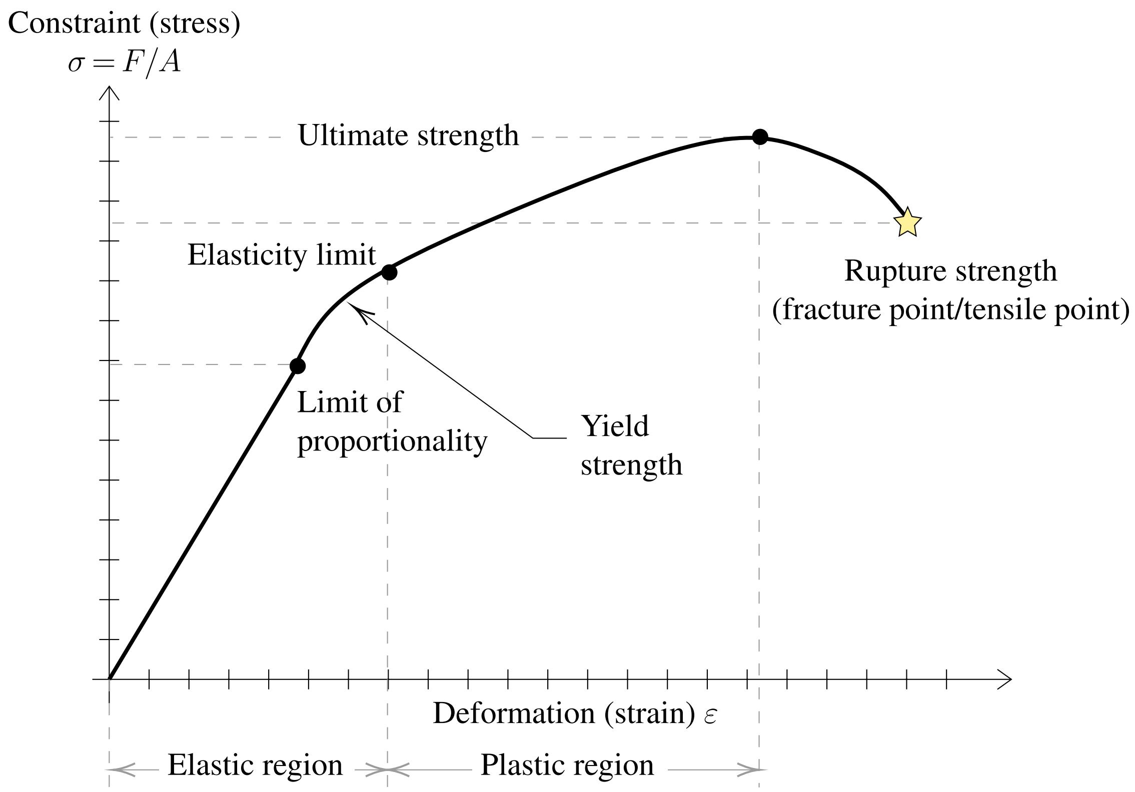


Figure 9.187: **Tensile Stress-Strain Diagram**

The above characteristic includes a linear portion as is the case of a certain class of materials. This means that the slope of the characteristic is a constant which reflects the elastic deformation of the material under the effect of increasing stress. This elastic stress by deformation unit defines the "Young's modulus" (there is no tangential component in this case study!):

$$\boxed{E = \frac{\sigma}{\varepsilon}} \tag{9.7.11}$$

also denoted sometimes by the letter Y to avoid obviously any confusion with energy. This relation is valid both compressive or in tension stresses. We will return to this relation in the following paragraphs.

Remarks

R1. "Rheology" is field of the mechanics that studies plasticity, elasticity, viscosity and the fluidity characteristics of the deformable body. This is a very important branch of industrial engineering.

R2. Be careful the calculations that will follow are relatively long and difficult, even if we tried to simplify them to the maximum as we always do in this textbook. However, all the results will be infinitely useful either to determine the Navier-Stokes equation or for the study of material strength (see section Mechanical Engineering page 8359)!

Definition (#9.7.470): The "toughness" is the ability of a material to absorb energy and plastically deform without fracturing. One definition of material toughness is the amount of energy per unit volume that a material can absorb before rupturing. It is also defined as a material's resistance to fracture when stressed.

Toughness can be determined by integrating the stress-strain curve as seen above. It is the energy of mechanical deformation per unit volume prior to fracture. The explicit mathematical description is:

$$\frac{\text{Energy}}{\text{Volume}}=\int_0^{\varepsilon_f}\sigma\mathrm{d}\varepsilon \tag{9.7.12}$$

where ε is strain, ε_f is the strain upon failure and σ is stress.

Typically in laboratories the toughness is multiplied by the length of the test specimen. This is why the toughness is given many times in Energy units by Surface units as in the figure below:

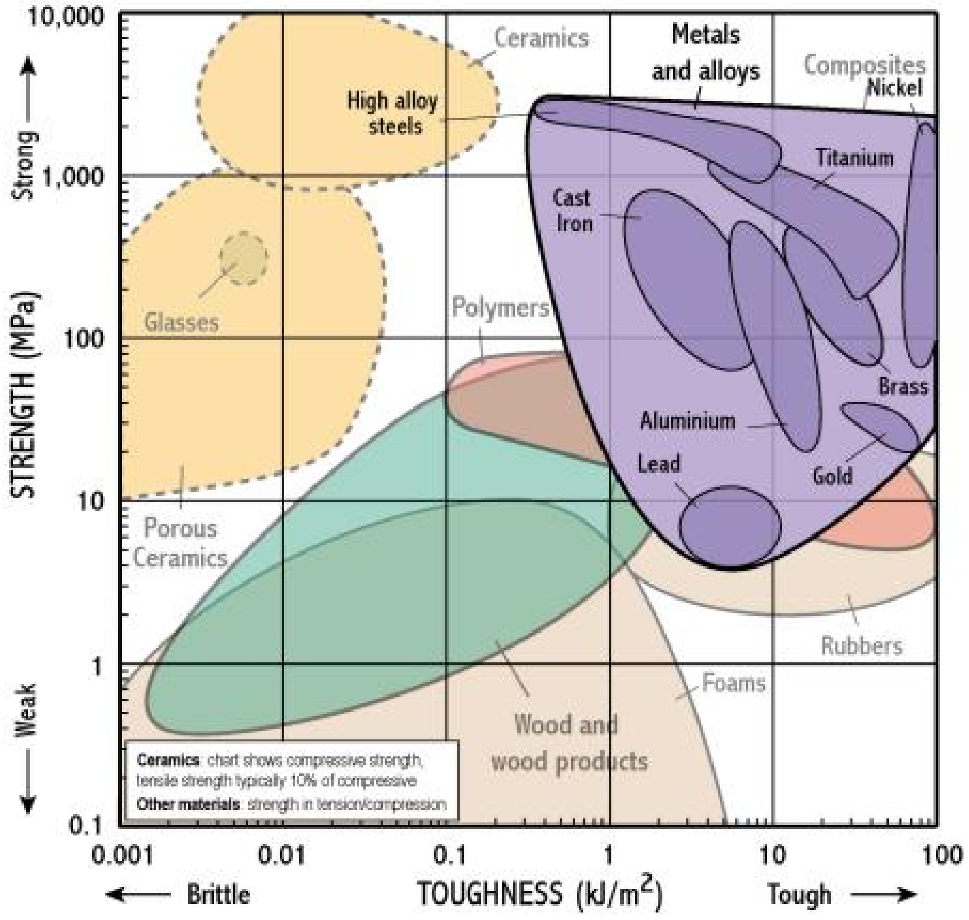


Figure 9.188: **Toughness schema of common material (source: ?)**

For example, brittle materials (like ceramics) that are strong but with limited ductility are not tough; conversely, very ductile materials with low strengths are also not tough. To be tough, a material should withstand both high stresses and high strains. Generally speaking, strength indicates how much force the material can support, while toughness indicates how much energy a material can absorb before rupturing.

For general knowledge here is a diagram of strength vs density for various materials (sadly the source in unknown):

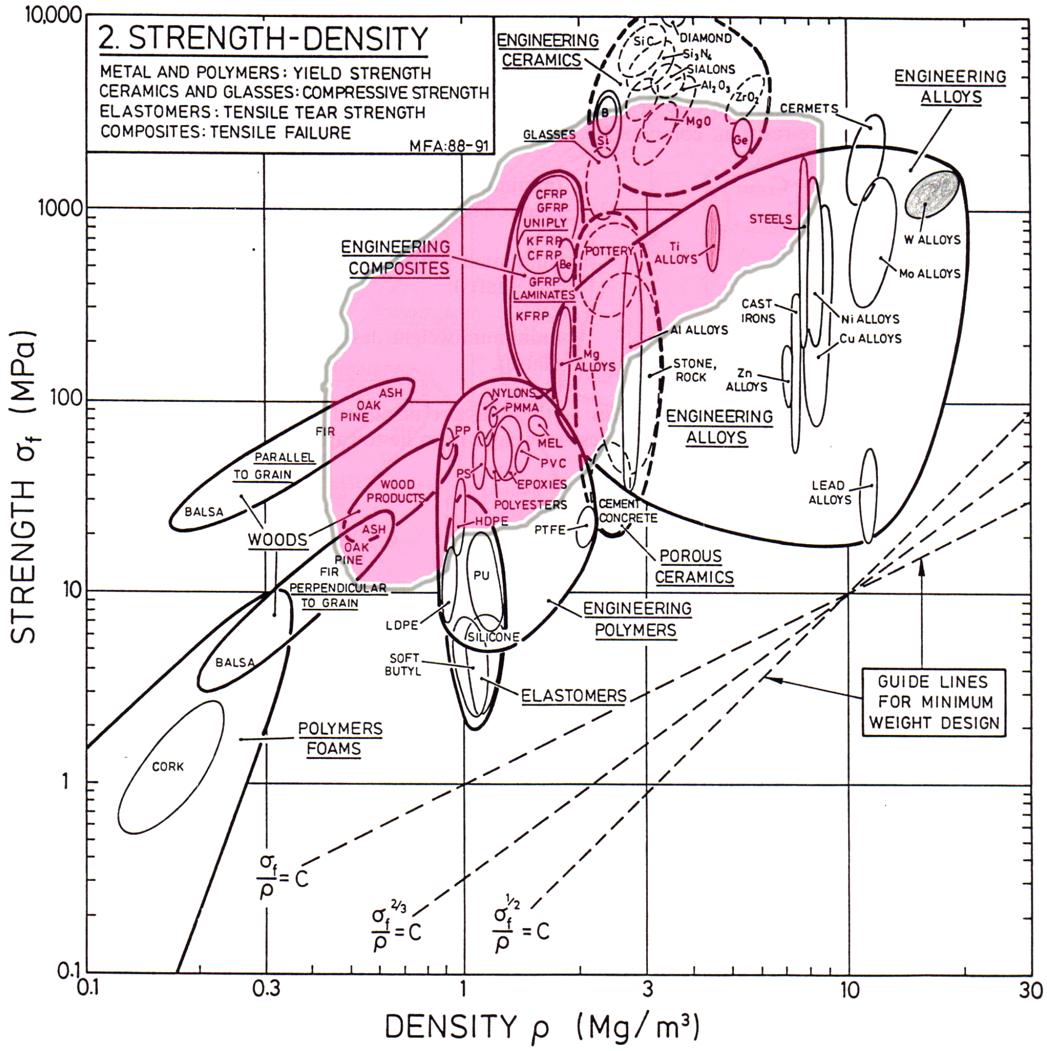
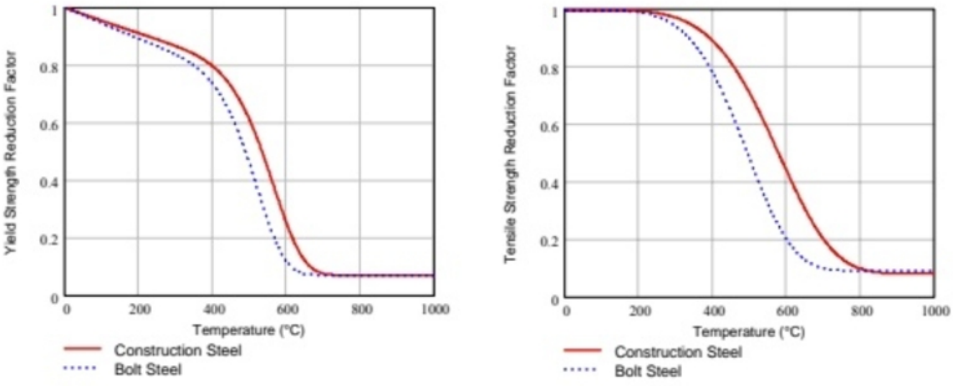


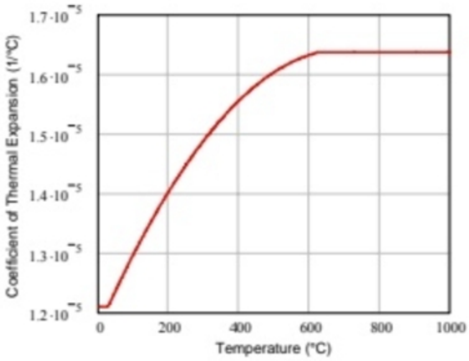
Figure 9.189: Strength vs Density of some materials (source: ?)

The reader must keep in mind that obviously, as always in physics, these properties are temperature dependant:



Yield strength reduction factor

Tensile strength reduction factor



Instantaneous coefficient of thermal expansion

Figure 9.190: **Variation of various steel strengths as function of temperature (source: NIST)**

9.7.1.2.1 Hooke's law

Hooke's law is a principle of physics that states that the force F needed to extend or compress a spring by some distance ΔL is proportional to that distance. An elastic body or material for which this assumption can be assumed is said to be "linear-elastic" or "Hookean".

On the other hand, Hooke's law is an accurate approximation for most solid bodies, as long as the forces and deformations are small enough. For this reason, Hooke's law is extensively used in all branches of science and engineering, and is the foundation of many disciplines such as seismology, molecular mechanics and acoustics. It is also the fundamental principle behind the spring scale, the manometer, and the balance wheel of the mechanical clock.

The modern theory of elasticity generalizes Hooke's law to say that the strain (deformation) of an elastic object or material is proportional to the stress applied to it.

Taking into consideration the definitions given above:

$$\sigma=\frac{F}{S} \qquad \varepsilon=\frac{|\Delta L|}{L_0} \qquad E=\frac{\sigma}{\varepsilon} \tag{9.7.13}$$

we get the relation:

$$\boxed{\frac{|\Delta L|}{L_0}=\frac{F}{ES}} \tag{9.7.14}$$

which by definition is the "linear Hooke's Law" in normal stress only!

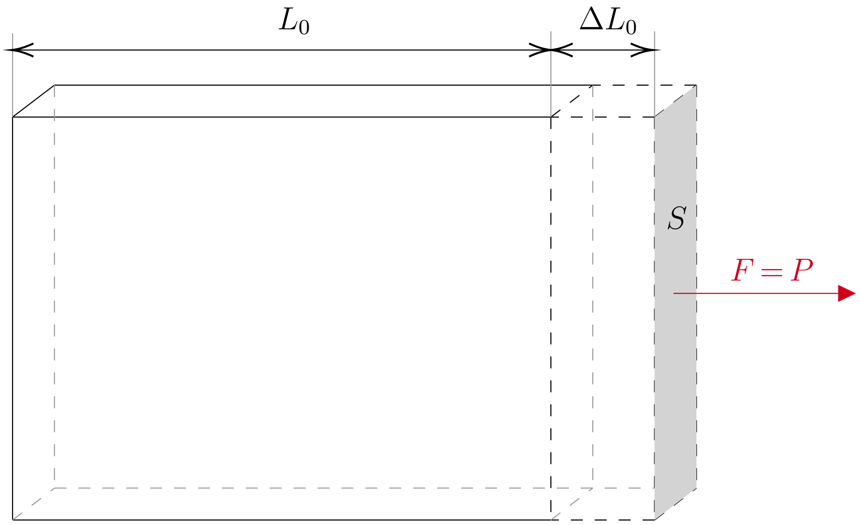


Figure 9.191: **Illustration of the effect of a normal stress**

✍**Example:**

Consider a suspension cable used to carry gondolas at ski resorts and that the cable has an unsupported span of 3 [km]. We assume that the cable has a radius of $r=2.3$ [cm] and its tension is $3.0\cdot 10^6$ [N]. The stretch of the steel cable is therefore equal to:

$$|\Delta L|=\frac{F}{ES}L_0=\frac{3\cdot 10^6}{210\cdot 10^9\cdot\pi(0.028)^2}\cdot 3000\cong 17.5\,[\text{m}] \tag{9.7.15}$$

It is quite intuitive to assume that the higher the bonding force of the atoms constituting the study material, the greater is the force to be applied to take them away, so to stretch the body. Solids, which have big bond strengths, have a high melting temperature (this is detailed in the section of Quantum Chemistry).

If we denote:

$$k:=\frac{ES}{L_0} \tag{9.7.16}$$

We end up with a relation that we already know:

$$\boxed{F=k|\Delta L|} \tag{9.7.17}$$

which is the restoring force of springs (see sections of Classical Mechanics page 3417 and Mechanical Engineering page 3416) and just a variant notation of Hooke's law.

There is also another way to introduced Hooke's law that is quite interesting and used in many other fields of physics (especially in Solid State Physics) as it used the microscopic picture of a solid.

Let us consider for this purpose a solid piece of some simple material. If we look at pure elemental solids – for example metals – we often find that when they form a solid the atoms arrange themselves into a regular lattice that is "close packed" – arranged so that the atoms more or less touch each other with a minimum of wasted volume. There are a number of kinds of lattices that appear (determined by the subtleties of the quantum mechanical interactions between the atoms) but in many cases the lattice is a variant of the cubic lattice where there are atoms on the corners of a regular cartesian grid in three dimensions (and sometimes additional atoms in the center of the cubes or on the faces of the cubes).

Not all materials are so regular. Materials made out of a mixture of atoms, out of molecules made of a mixture of atoms, out of a mixture of molecules, or even out of living cells made out of a mixture of molecules. The resulting materials can be ordered, structured (not exactly the same thing as ordered, especially in the case of solids formed by life processes such as bone or coral), or disordered (amorphous).

As usual, we will deal with all of this complexity by ignoring most of it for now and considering an "ideal" case where a single kind of atoms lined up in a regular simple cubic lattice is sufficient to help us understand properties that will hold, with different values of course, even for amorphous or structured solids.

This is illustrated in the figure below:

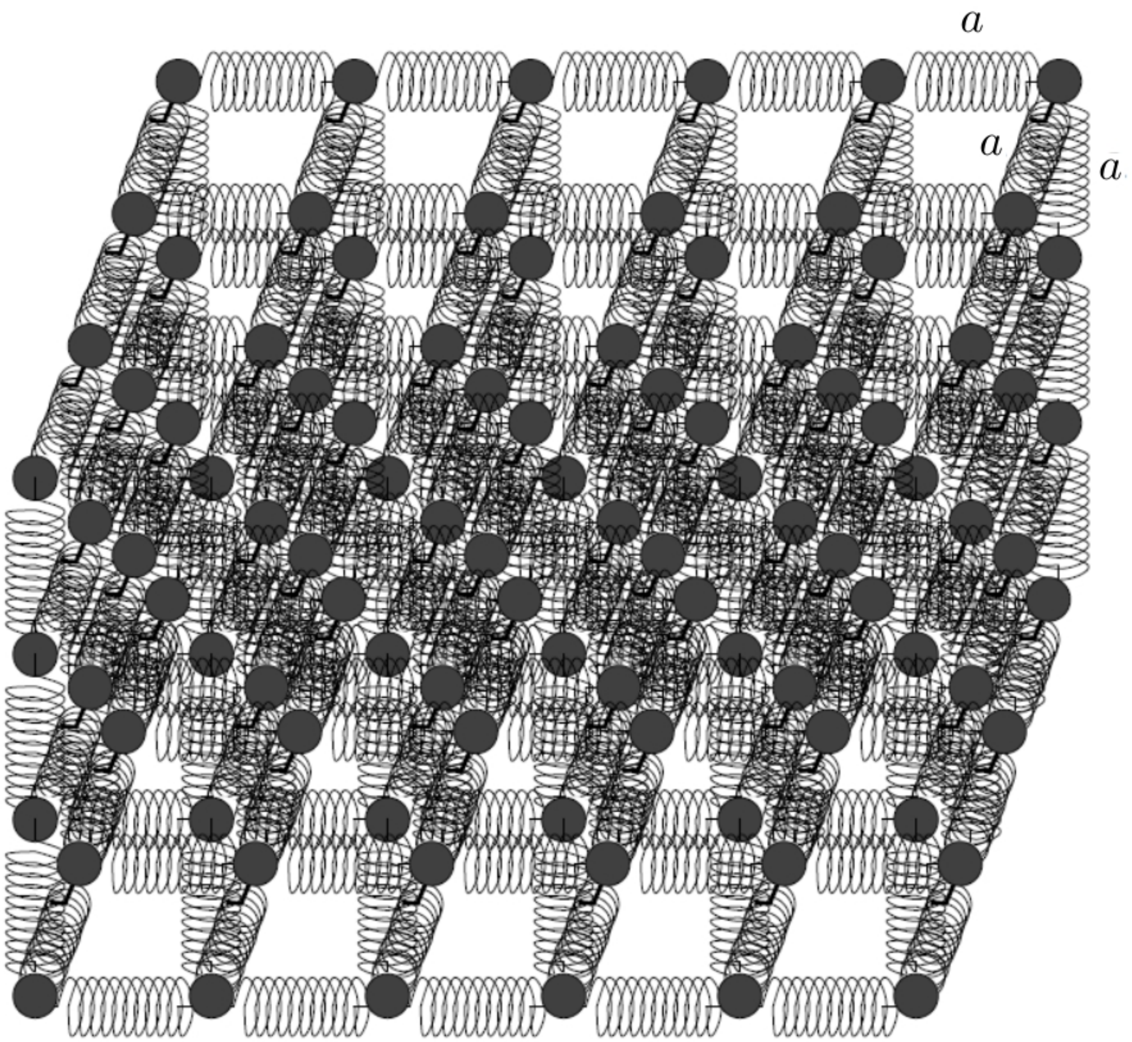


Figure 9.192: Idealized simple cubic lattice of atoms separated by the "springs" of inter-atomic forces that hold them in equilibrium positions

which is basically a mental cartoon model for a generic solid – lots of atoms in a regular cubic lattice with a cube side a, where the inter-atomic forces that hold each atom in position is represented by a spring, a concept that is valid as long as we don't compress or stretch these inter-atomic bonds by too much.

We need to quantify the numbers of atoms and bonds in a way that helps us understand how stretching or compressing forces are distributed among all of the bonds. Suppose we have N_x atoms in the x-direction, so that the length of the solid is $L = N_x a$. Suppose also that we have N_y and N_z atoms in the y- and z-directions respectively, so that the cross-sectional surface is therefore:

$$S = N_y N_z a^2 \tag{9.7.18}$$

Now let us imagine applying a force (magnitude F) uniformly to all of the atoms on the left and right ends (see figure below) that stretches all of these bonds by a small amount Δx, presumed to be "small" in precisely the sense that leaves the inter-atomic bonds still behaving like springs. The force F has to be distributed equally among all of the atoms on the end areas on both sides, so that the force applied to each chain of atoms in the x-direction end to end is:

$$F_{\text{chain}} = \frac{F}{N_y N_z} = \frac{F}{\dfrac{S}{a^2}} = \frac{Fa^2}{S} \tag{9.7.19}$$

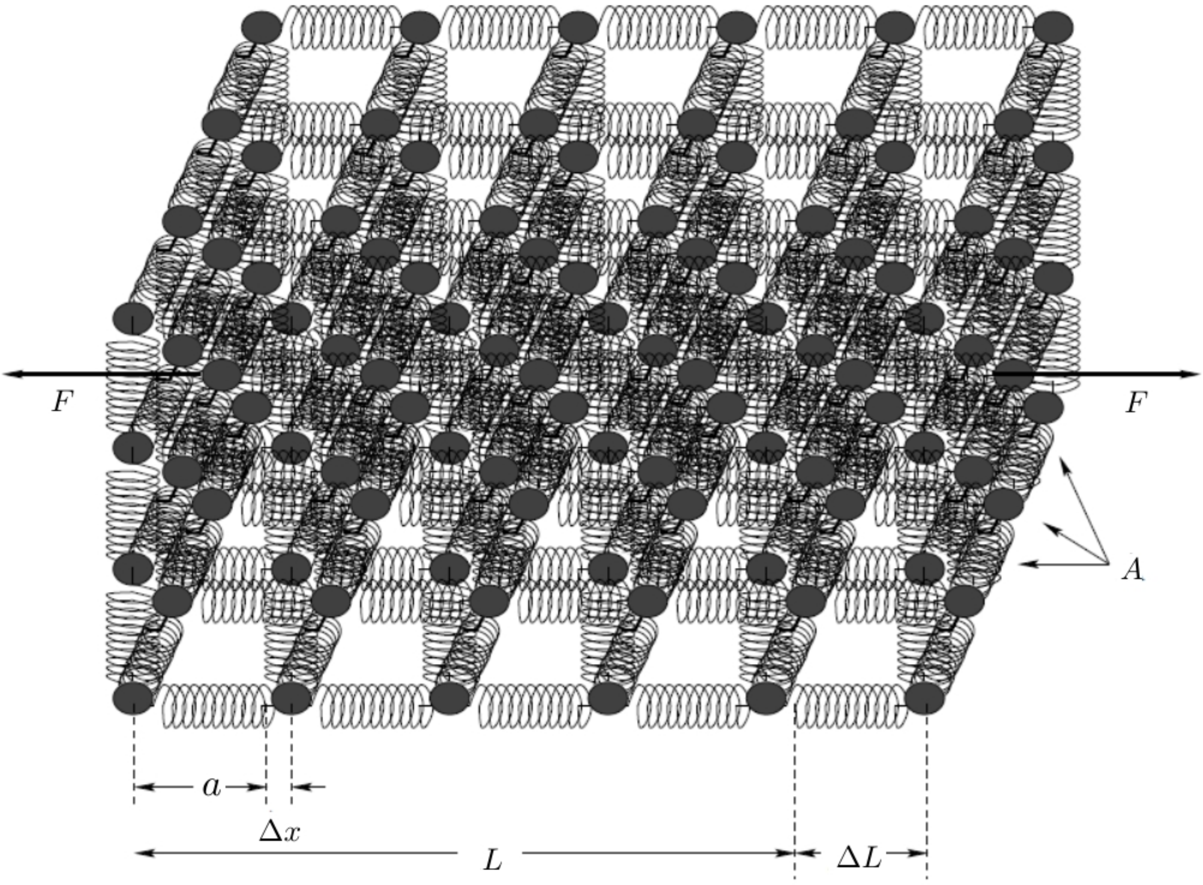


Figure 9.193: The same lattice stretched by an amount ΔL as a force F is applied to both ends, spread out uniformly across the cross-sectional area of the faces A.

Each spring in the chain is stretched by this force between the atoms on the ends, so that:

$$F_{\text{chain}}=-\frac{Fa^2}{S}=-K\Delta x \tag{9.7.20}$$

The negative sign just means that the springs are trying to go back to their equilibrium length and hence oppose the applied force. We multiply this by 1 in the form $(aN_x)/(aN_x)$ to get:

$$\frac{Fa^2}{S}=-K\frac{aN_x}{aN_x}\Delta x \tag{9.7.21}$$

giving as$L=aN_x$:

$$\frac{Fa^2}{S}=-Ka\frac{N_x\Delta x}{L}=-Ka\frac{\Delta L}{L} \tag{9.7.22}$$

Finally, we rearrange this by putting all of the extra terms together as follows (we see that we fall back on Hooke's law):

$$\frac{F}{S}=-\frac{K}{a}\frac{\Delta L}{L}=-E\frac{\Delta L}{L} \tag{9.7.23}$$

In words, we define $\sigma=F/S$ to be the "Stress", $\varepsilon=\Delta L/L$ to be the "Strain", and $E=K/a$ to be Young's Modulus. With these definitions, the equation above states that: Compressive or extensive Stress applied to a solid equals Young's Modulus times the Strain.

But there are several types of constraints with their respective modules. So here are the most important definitions in the linear part of their characteristic associated with the explanatory diagram:

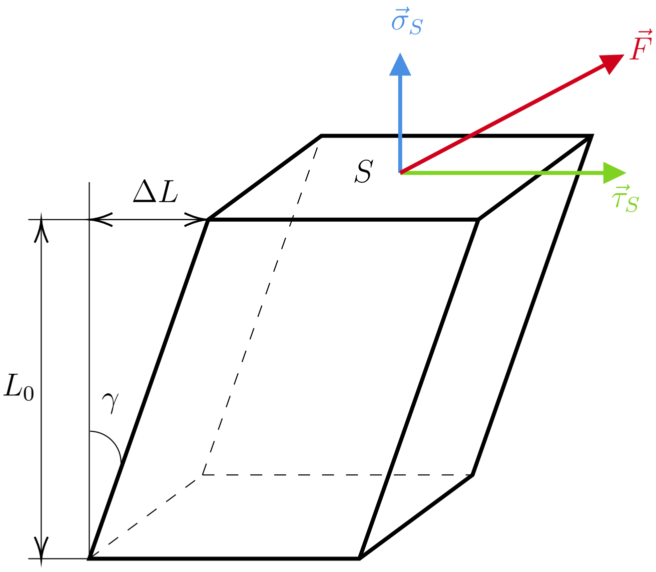


Figure 9.194: **Illustration of the effect of a shear stress**

Definitions (#9.7.471):

D1. We define the "shear modulus" or "rigidity modulus" by the ratio of the normal component of the force (compression pressure) to the shear strain:

$$G=\frac{\sigma_S}{\varepsilon_S}=\frac{F_N/S}{\gamma} \tag{9.7.24}$$

where the numerator is named the "shear stress" and where γ is the "angle of deformation". Generally this angle being small, we have the approximation:

$$\varepsilon_S = \gamma \cong \tan(\gamma) = \frac{|\Delta L|}{L_0} \tag{9.7.25}$$

S is the surface of the upper or lower face of the deformed body shown in the figure above.

D2. We define the "sliding elastic modulus", also named "sliding modulus" and most commonly "shear modulus" or "Coulomb modulus" by the ratio of the tangential component of the force (shear stress) to the shear strain:

$$G = \frac{\tau_S}{\varepsilon_S} = \frac{F_T/S}{\gamma} = \frac{E}{2(1+\eta)} \tag{9.7.26}$$

where η is the "Poisson's coefficient" which we will prove later in detail where it comes from.

Notice well that even if the numerator of the above definition is a force divided by an area, this is not a pressure because the force is tangential to the surface (hence the T in index of F).

This is because any force can be decomposed into a normal and tangential force (see the above definition of the compression pressure and constraint pressure) that we have two distinct definitions above. In the vast majority of laboratories situations, we arrange to have a purely tangential force (again... hence the T in index of the F) or purely normal force (hence the N in index of the F) to the surface S.

In practice many times make usage of only the second definition and thus to such a level that the latter is often just named "rigidity modulus" also...

✍**Example:**

A very interesting thing (for the parenthesis ...) if we consider that the tectonic plates are in a shearing situation between themselves we then have according to the sliding modulus:

$$F_T = \gamma G S = \frac{|\Delta L|}{L_0} G S \tag{9.7.27}$$

But for a tectonic plate inf friction of length L_0 over a height H:

$$F_T = \frac{|\Delta L|}{L_0} G(L_0 H) = G\left(\frac{|\Delta L|}{L_0}\right)(L_0 H) \tag{9.7.28}$$

and since energy is a force multiplied by a distance, we get:

$$E = F_T \Delta L = \left[G\frac{|\Delta L|}{L_0}(L_0 H)\right]\Delta L = G\Delta L^2 H \tag{9.7.29}$$

which is typically the energy liberated by the shear of the friction of two tectonic plates whose contact surfaces have an average height H, an initial length L_0 and undergoing a deformation ΔL.

Typically for an earthquake of the type of Sumatra in the year 12009 according to holocene calendar (9.1 to 9.3 on Richter scale, 28 [km] underground, 200,000 deaths), we had:

$$\Delta L \cong 10\,[\text{m}] \quad H \cong 10,000\,[\text{m}] \quad L_0 \cong 10,000\,[\text{m}] \quad G \cong 10^{10}\,[\text{Pa}] \tag{9.7.30}$$

Therefore it comes:

$$E \cong 10^{16}\,[\text{J}] \tag{9.7.31}$$

in other words ... a thousand times the energy of the Hiroshima nuclear bomb in.

If we denote by M the empirical definition of magnitude of the Richter's scale magnitude, we get:

$$E \cong 10^{5+1.44M} = 10^{16} \Rightarrow M \cong 7.6 \tag{9.7.32}$$

while estimations give a range of 6.2 to 9.2 ... so we are not too bad in the theoretical approach.

D3. We define the "omnidirectional modulus", as the ratio of the volume constraint named "bulk stress" on the volumetric strain named "bulk strain" (we will prove further below the mathematical developments that lead to the last term of this relation):

$$\kappa = \frac{\varepsilon_v}{\sigma} = \frac{F/S}{\Delta V/V_0} = \frac{P}{\Delta V/V_0} = \frac{3(1-2\eta)}{E} \tag{9.7.33}$$

In practice it is much more interesting to use:

$$\Delta V = \frac{PV_0}{\kappa} \tag{9.7.34}$$

and to seek what material (depending on the context) maximize or minimize ΔV.

We could define many other modules such as the bending modulus, the pure bending modulus, the composed bending modulus, the twisting modulus... We will study some of them further below.

For each of the different definitions of modules that we can imagine, we can define a Hooke's law that is suitable to it. However, all this may seem rather arbitrary, but in fact it is not! Because all definitions of modules that we saw earlier are a special case of a generalized mathematical relation we will be prove in this textbook in a near future.

9.7.1.2.2 Shear Modulus

The necessary condition for a rigid solid to be in static equilibrium as we saw it in the section of Classical Mechanics, is that the resultant forces that are applied to the body is equal zero:

$$\vec{F}_{\text{ext}} = \vec{0} \tag{9.7.35}$$

However, when a solid is subjected to stresses and it can tolerate them, there can be a deformation which may be followed by a breaking or a similar modification. More precisely, there is "deformation" of a body (not necessarily solid) when the distance between the elements it is made of have changed.

When in the theoretical study of elasticity, we exclude modifications of the body studied such as breaks, we say that we restrict ourselves to the "elastic deformations".

The geometry and the physics of deformations can be complex. Their description is derived from that of a number of elementary deformations we will specify further the characteristics.

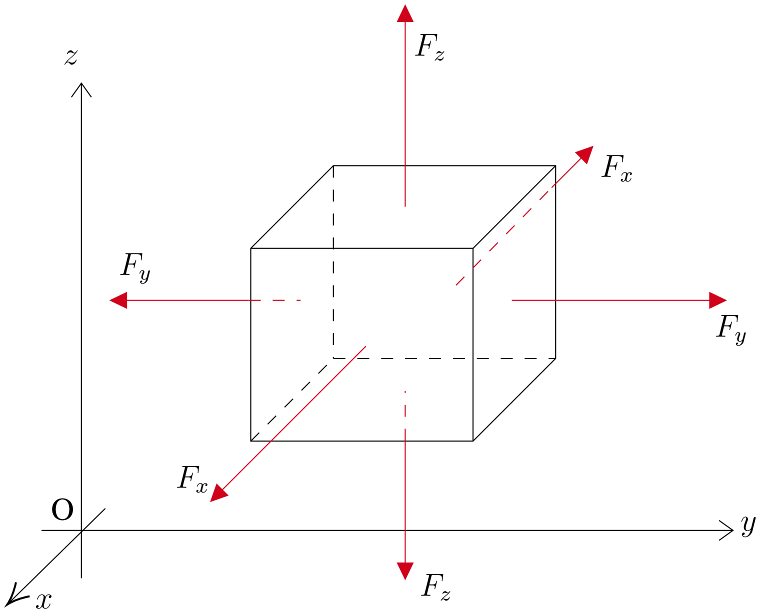


Figure 9.195: **Cube under normflexural modulusal stresses**

The scalar tensile stresses forces F_x, F_y, F_z generate on their faces "normal" tensions (that is to say perpendicular... for recall...):

$$\sigma_x = \frac{F_x}{S_x} \qquad \sigma_y = \frac{F_y}{S_y} \qquad \sigma_z = \frac{F_z}{S_z} \tag{9.7.36}$$

Assuming the only the force F_x acts, the unit deformation is by definition:

$$\varepsilon'_x = \frac{\sigma_x}{E} \tag{9.7.37}$$

When a parallelepiped is subjected to a tensile stress F_y, there is intuitively contraction of dimensions along the x direction. Contraction also observable also intuitively for F_z.

We then have if only the force F_y acts:

$$\varepsilon''_x = -\eta\frac{\sigma_y}{E} \tag{9.7.38}$$

where the sign "$-$" indicates a contraction and where η is a coefficient named "Poisson's coefficient".

If the Poisson coefficient of the studied matter is almost zero, this means that when you "crush" (compress) the material, its height decreases, but its width remains almost constant ...

One remark on the way: There are products that have a negative Poisson coefficient: when compressed, their width decreases! In general, it is porous media, inverted honeycomb style like expanded graphite that has this original property.

If only the force F_z acts:

$$\varepsilon'''_x = -\eta\frac{\sigma_z}{E} \tag{9.7.39}$$

By accepting the principle of superposition of forces, the effect of several forces acting simultaneously is equal to the sum of the effects produced by each of the superimposed forces acting separately (this is obviously a strong assumption!). Since then:

$$\varepsilon_x = \varepsilon'_x + \varepsilon''_x + \varepsilon'''_x \tag{9.7.40}$$

This can be considered as acceptable, given the linearity of the equations linking the unit deformation and the normal stresses. We then get:

$$\begin{aligned}\varepsilon_x &= \frac{1}{E}\left[\sigma_x - \eta(\sigma_y + \sigma_z)\right]\\ \varepsilon_y &= \frac{1}{E}\left[\sigma_y - \eta(\sigma_x + \sigma_z)\right]\\ \varepsilon_z &= \frac{1}{E}\left[\sigma_z - \eta(\sigma_x + \sigma_y)\right]\end{aligned} \tag{9.7.41}$$

having proceed in the same way for the other two directions following the y-axis and z-axis.

From the previous relations, it is easy to find the equations linking σ to ε by solving the following linear system (see section Linear Algebra page 1929):

$$\begin{aligned}\sigma_x &= \frac{E}{1+\eta}\left[\varepsilon_x + \frac{\eta}{1-2\eta}(\varepsilon_x + \varepsilon_y + \varepsilon_z)\right]\\ \sigma_y &= \frac{E}{1+\eta}\left[\varepsilon_y + \frac{\eta}{1-2\eta}(\varepsilon_x + \varepsilon_y + \varepsilon_z)\right]\\ \sigma_z &= \frac{E}{1+\eta}\left[\varepsilon_z + \frac{\eta}{1-2\eta}(\varepsilon_x + \varepsilon_y + \varepsilon_z)\right]\end{aligned} \tag{9.7.42}$$

Given a material subject to various constraints. Inside of it, we operate in thought, to the extraction of a rectangular parallelepiped:

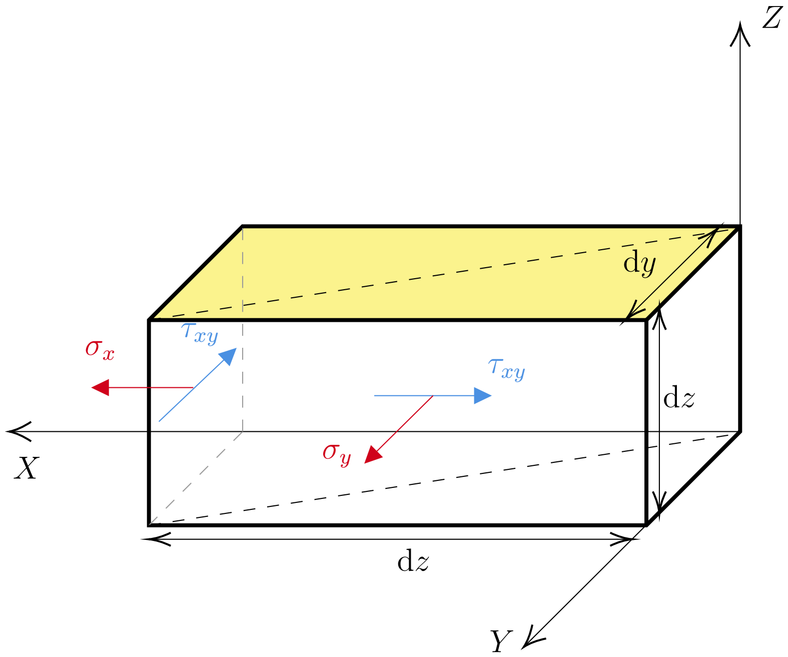

Z
dy
τ_{xy}
σ_x
τ_{xy}
dz
X
σ_y
dz
Y

Figure 9.196: **Parallelepiped rectangle p for the theoretical study**

The surfaces thereof are solicited by normal σ and tangential τ stresses (in the figure below the solid is in static equilibrium). Here is the yellow face:

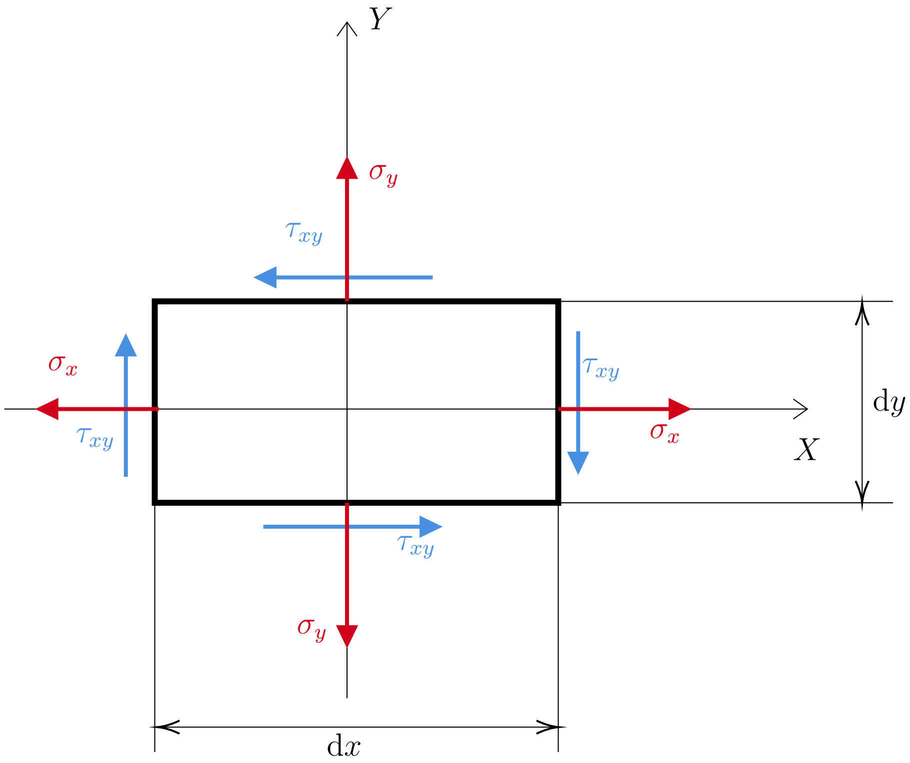

Y
σ_y
τ_{xy}
σ_x
τ_{xy}
τ_{xy}
σ_x
X
dy
τ_{xy}
σ_y
dx

Figure 9.197: **Generic illustration of a material under normal and tangential stresses**

The normal σ and tangential τ constraints represent the actions of the parallelepiped material extracted mentally on the faces of the examined item.

It is interesting (in the sense that it facilitates the analysis) to seek the constraints that exist in a plane making an angle θ with the x-axis. To do this, we imagine a material triangle having a top angle θ extracted from matter mentally. We neglect the influence of gravity.

Therefore:

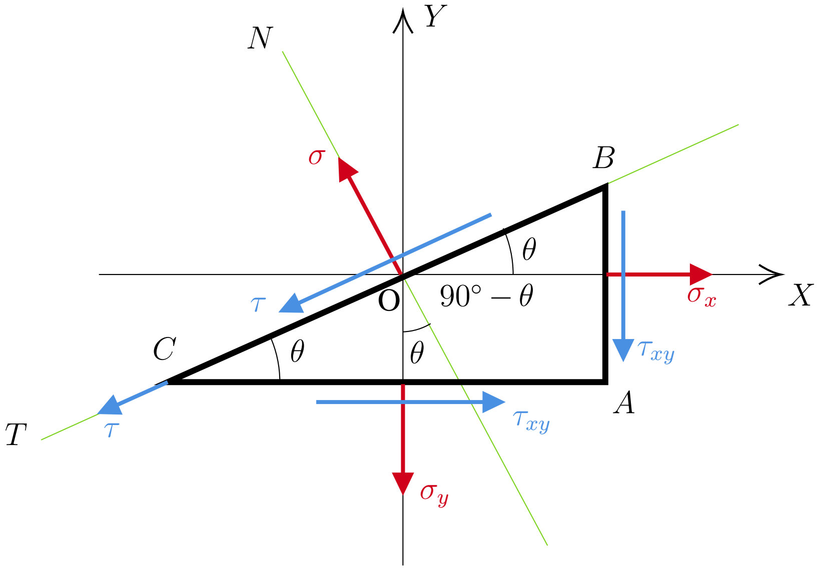


Figure 9.198: Search of constraints expressions in an oblique plane

Let us put:

$$\overline{CB} = \mathrm{d}s \qquad \overline{CA} = \mathrm{d}x \qquad \overline{AB} = \mathrm{d}y \tag{9.7.43}$$

and $\mathrm{d}z$ being the thickness of the solid (not shown in the diagram above).

On the length $\mathrm{d}s$ constraints appear and are decomposed under normal σ and tangential τ stresses (the latter being also named "shear stress" or "bending stress" as we already know).

The problem is to establish the relations between σ, τ and σ_x, σ_y and τ.

The sign conventions are:

- The constraints σ applying a traction are positive while the constraints σ exerting compression are negative.
- Constraints τ that tends to rotate the cuboid in the clockwise direction are positive. Counter-clockwise, they will be negative.

The projection equation of equilibrium following the direction ON is then:

$$\sigma_z \mathrm{d}z\mathrm{d}s - \left[\sigma_y \mathrm{d}z\mathrm{d}x\cos(\theta) + \sigma_x \mathrm{d}z\mathrm{d}y\cos\left(\frac{\pi}{2}-\theta\right) + \tau_{xy}\mathrm{d}x\mathrm{d}y\cos(\theta) + \tau_{xy}\mathrm{d}z\mathrm{d}x\cos\left(\frac{\pi}{2}-\theta\right)\right] = 0 \tag{9.7.44}$$

Let us recall that:

$$\sigma_i = \frac{F_{iN}}{S} \Leftrightarrow F_{iN} = S\sigma_i \tag{9.7.45}$$

As $\mathrm{d}x=\mathrm{d}s\cos(\theta)$ and $\mathrm{d}y=\mathrm{d}s\sin(\theta)$ we have:

$$\begin{aligned}&\sigma\mathrm{d}s=\sigma_x\mathrm{d}s\sin^2(\theta)+\sigma_y\mathrm{d}s\cos^2(\theta)+\tau_{xy}\mathrm{d}s\sin(\theta)\cos(\theta)+\tau_{xy}\mathrm{d}s\sin(\theta)\cos(\theta)\\&\sigma=\sigma_x\sin^2(\theta)+\sigma_y\cos^2(\theta)+\tau_{xy}\sin(2\theta)\end{aligned}\tag{9.7.46}$$

As:

$$\sin^2(\theta)=\frac{1-\cos(2\theta)}{2}\quad\text{and}\quad\cos^2(\theta)=\frac{1+\cos(2\theta)}{2}\tag{9.7.47}$$

then:

$$\sigma=\frac{\sigma_x}{2}(1-\cos(2\theta))+\frac{\sigma_y}{2}(1+\cos(2\theta))+\tau_{xy}\sin(2\theta)\tag{9.7.48}$$

Finally:

$$\sigma=\frac{1}{2}(\sigma_x+\sigma_y)-\frac{1}{2}(\sigma_x-\sigma_y)\cos(2\theta)+\tau_{xy}\sin(2\theta)\tag{9.7.49}$$

As conclusion we can say that in function of σ_x, σ_y and τ_{xy}, it is possible to calculate the normal stress that exists on a plane surface of angle θ.

The projection equation of equilibrium is following the direction OT is:

$$\begin{aligned}&\tau\mathrm{d}z\mathrm{d}s-\sigma_x\mathrm{d}z\mathrm{d}y\cos(\theta)+\tau_{xy}\mathrm{d}z\mathrm{d}y\cos\left(\frac{\pi}{2}-\theta\right)+\sigma_y\mathrm{d}z\mathrm{d}x\mathrm{d}x\cos\left(\frac{\pi}{2}-\theta\right)-\tau_{xy}\mathrm{d}z\mathrm{d}x\cos(\theta)=0\\&\tau\mathrm{d}s=\sigma_x\mathrm{d}\sin(\theta)\cos(\theta)-\sigma_y\mathrm{d}s\sin(\theta)\cos(\theta)-\tau_{xy}\mathrm{d}s\sin^2(\theta)+\tau_{xy}\mathrm{d}s\cos^2(\theta)=0\\&\tau=\frac{1}{2}(\sigma_x-\sigma_y)\sin(2\theta)+\tau_{xy}(\cos^2(\theta)-\sin^2(\theta))\end{aligned}\tag{9.7.50}$$

as $\cos^2(\theta)-\sin^2(\theta)=\cos(2\theta)$ then finally:

$$\tau=\frac{1}{2}(\sigma_x-\sigma_y)\sin(2\theta)+\tau_{xy}\cos(2\theta)\tag{9.7.51}$$

The conclusion is that in function of σ_x, σ_y and τ_{xy} it is possible to calculate the tangential stress τ that exists on a plane surface of angle θ.

So finally with to both relations:

$$\begin{aligned}\sigma&=\frac{1}{2}(\sigma_x+\sigma_y)-\frac{1}{2}(\sigma_x-\sigma_y)\cos(2\theta)+\tau_{xy}\sin(2\theta)\\\tau&=\frac{1}{2}(\sigma_x-\sigma_y)\sin(2\theta)+\tau_{xy}\cos(2\theta)\end{aligned}\tag{9.7.52}$$

gives the values of the normal and tangential stresses in any cut in function of the normal and tangential stresses to the primitives axes.

Given now the following situation:

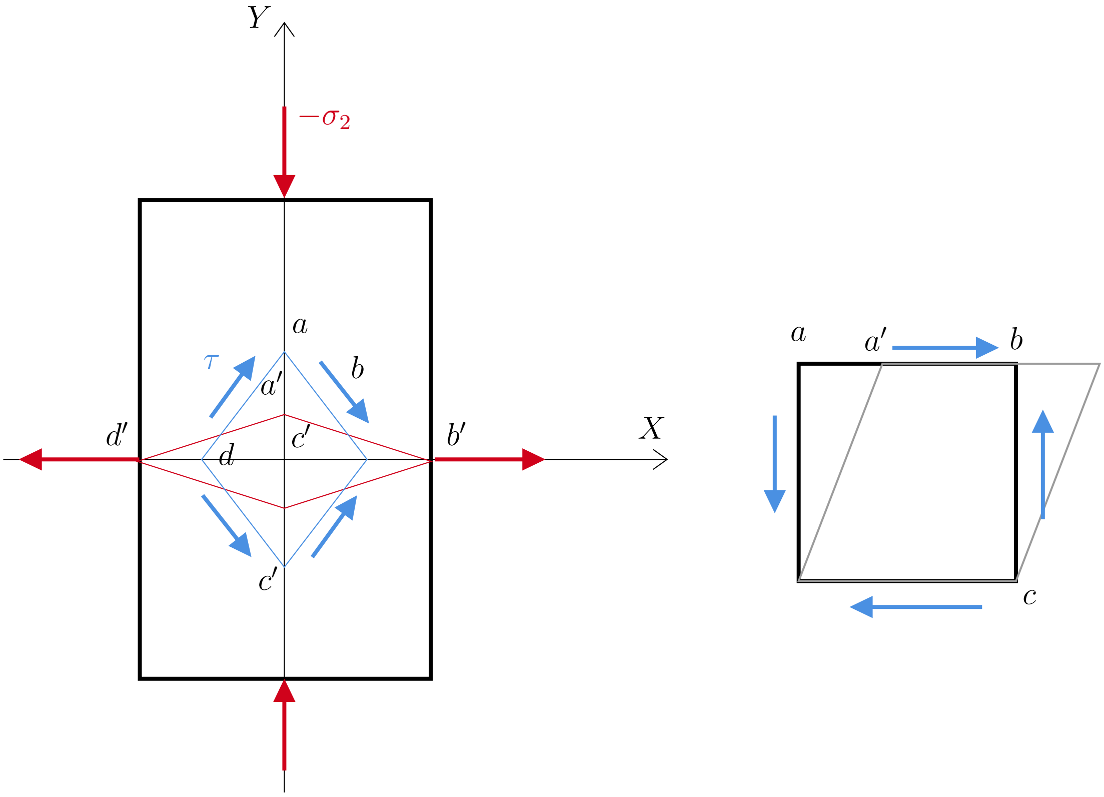


Figure 9.199: Situation for coming back to the study of the three dimensional case

On the figure above on the left the reader must imagine a block of material from which is virtually extracted a small square plane (in blue on the left figure) from which we will study in a first time only of the right triangles that compose it for afterwards study the whole.

Before solicitation, we therefore consider the diamond $abcd$ that is in fact originally a square with an angle of $\pi/4$ in the x direction. On the request of a reader here is perspective diagram of this situation:

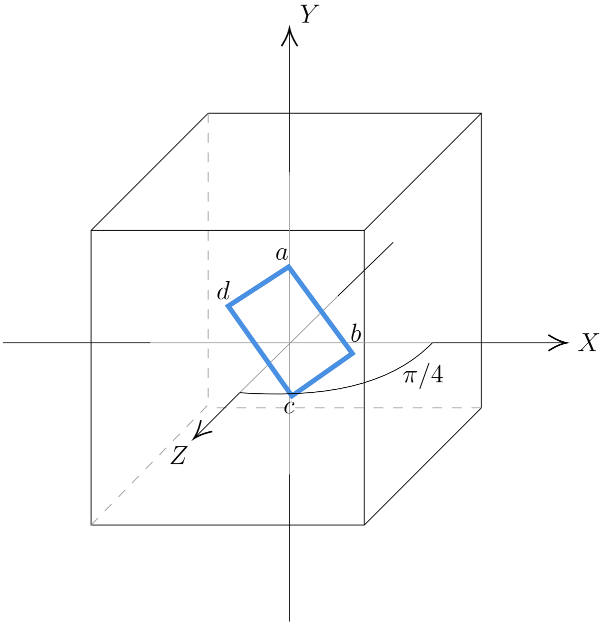


Figure 9.200: **Initial material block situation**

During solicitation, this diamond $abcd$ is deformed under the action of tangential stresses decomposed in pure shear stresses and becomes the diamond $a'b'c'd'$. Here is also a schema of this situation as requested by a reader:

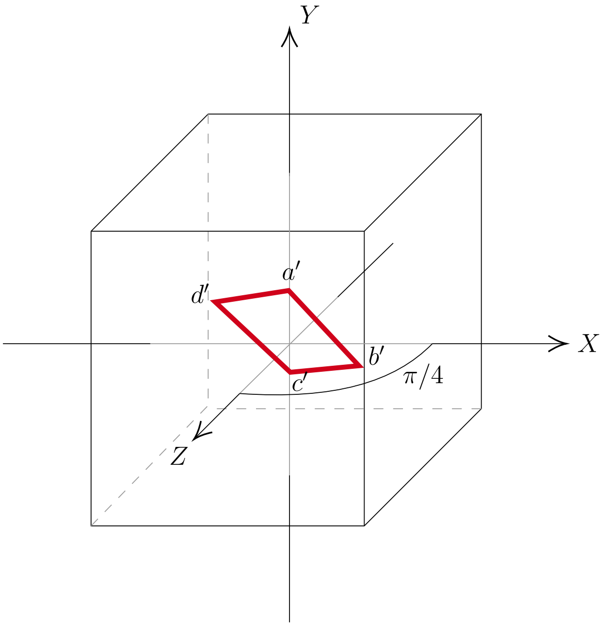


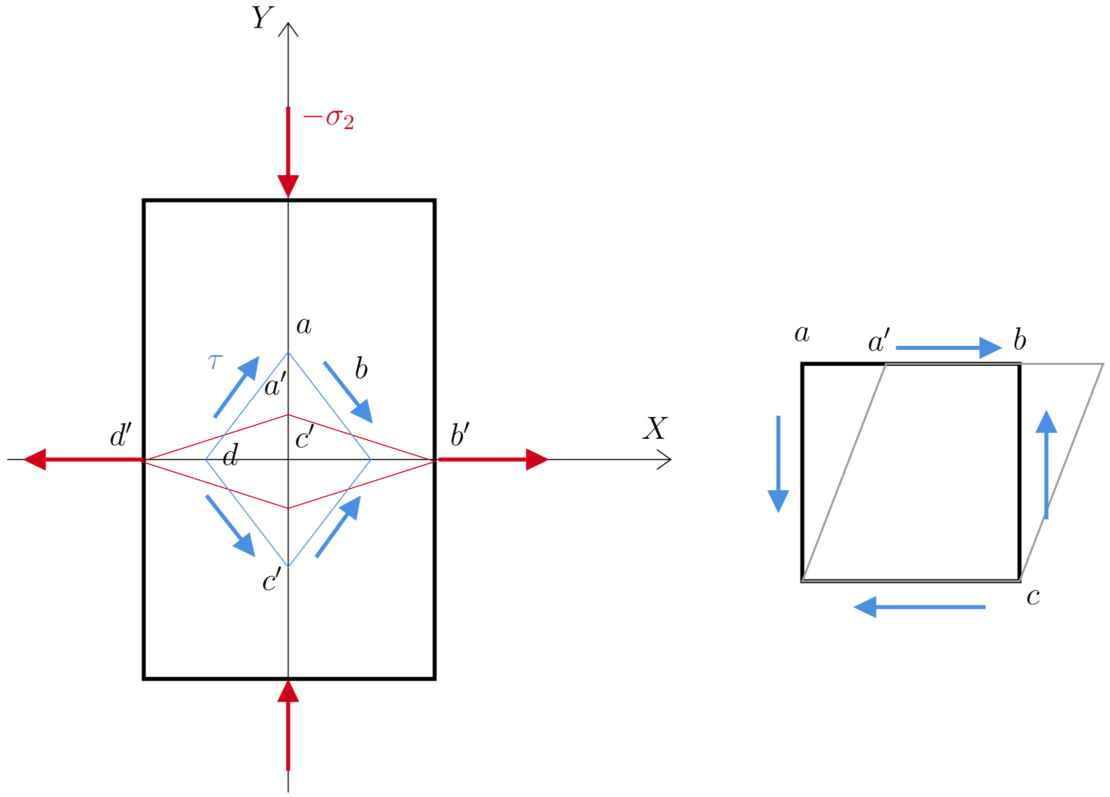
Figure 9.201: **Final material block situation**

The diagonal $\overline{bd}$ is then extended and the diagonal $\overline{ac}$ is compressed. The angle $\hat{a}$ that had for value $\pi/2$ is after deformation equal to $\pi/2+\gamma$ (in a'). Similarly, the angle b which value was $\pi/2$ is after deformation equal to $\pi/2-\gamma$.

Remark

The angle γ is named "slip angle" and we consider it to be small.

We need only to look at the effect of the deformation by isolating the diamond and by applying on it a rotation of $\pi/4$. After deformation, we have the shape indicated by the red lines in the left part of the figure below (figure that we already know):



The slip angle being small, we have:

$$\frac{\overline{aa'}}{\overline{ad}}=\tan(\gamma)\cong\gamma \quad (9.7.53)$$

Therefore γ represents the slip of the side ab with respect to dc divided by the distance between the two planes ab and dc. The analysis that has been done is still valid whatever the solid or liquid body considered.

Given, now, the case of an elastic solid obeying Hooke's law. The problem will be to establish the relation between the slip angle γ and the tangential constraints τ acting on the sides of the diamond.

Given the right triangle Oab. The lengthening of the side $\overline{\text{O}b}$ and shortening of side $\overline{\text{O}a}$ during the deformation are obtained from the following equations:

$$\begin{aligned}\varepsilon_x&=\frac{1}{E}\left[\sigma_x-\eta(\sigma_y+\sigma_z)\right]\\ \varepsilon_y&=\frac{1}{E}\left[\sigma_y-\eta(\sigma_x+\sigma_z)\right]\\ \varepsilon_z&=\frac{1}{E}\left[\sigma_z-\eta(\sigma_x+\sigma_y)\right]\end{aligned} \tag{9.7.54}$$

As:

$$\sigma_x=\sigma_1=\tau \qquad \sigma_y=\sigma_2=-\tau \qquad \sigma_z=0 \tag{9.7.55}$$

We have:

$$\varepsilon_x=\frac{1}{E}[\tau+\eta\tau]=\frac{(1+\eta)}{E}\tau \quad \text{and} \quad \varepsilon_y=\frac{1}{E}[-\tau-\eta\tau]=-\frac{(1+\eta)}{E}\tau \tag{9.7.56}$$

Therefore:

$$\overline{\mathrm{O}a'}=\overline{\mathrm{O}a}+\overline{\mathrm{O}a}\varepsilon_y=\overline{\mathrm{O}a}(1+\varepsilon_y)=\overline{\mathrm{O}a}\left[1-\frac{(1+\eta)}{E}\tau\right] \tag{9.7.57}$$

Then the length $\overline{\mathrm{O}a'}$ decrease if τ increase.

We also have:

$$\overline{\mathrm{O}b'}=\overline{\mathrm{O}b}+\overline{\mathrm{O}b}\varepsilon_x=\overline{\mathrm{O}b}(1+\varepsilon_x)=\overline{\mathrm{O}b}\left[1+\frac{(1+\eta)}{E}\tau\right] \tag{9.7.58}$$

Then the length $\overline{\mathrm{O}b'}$ decrease if τ increase.

For the angle of the right triangle $\mathrm{O}a'b'$, we have:

$$\tan\left(\widehat{\mathrm{O}a'b'}\right)=\tan\left(\frac{1}{2}\left(\frac{\pi}{2}+\gamma\right)\right)=\tan\left(\frac{\pi}{4}+\frac{\gamma}{2}\right)=\frac{\overline{\mathrm{O}b'}}{\overline{\mathrm{O}a'}}=\frac{\overline{\mathrm{O}b}\left[1+\dfrac{(1+\eta)}{E}\tau\right]}{\overline{\mathrm{O}a}\left[1-\dfrac{(1+\eta)}{E}\tau\right]}=\frac{1+\dfrac{(1+\eta)}{E}\tau}{1-\dfrac{(1+\eta)}{E}\tau} \tag{9.7.59}$$

But using trigonometric identities (see section Trigonometry page 2519):

$$\tan\left(\frac{\pi}{4}+\frac{\gamma}{2}\right)=\frac{\tan\left(\dfrac{\pi}{4}\right)+\tan\left(\dfrac{\gamma}{2}\right)}{1-\tan\left(\dfrac{\pi}{4}\right)\tan\left(\dfrac{\gamma}{2}\right)}=\frac{1+\tan\left(\dfrac{\gamma}{2}\right)}{1-\tan\left(\dfrac{\gamma}{2}\right)} \tag{9.7.60}$$

As $\tan\left(\dfrac{\gamma}{2}\right)\cong\dfrac{\gamma}{2}$ when γ is small as proved in the section of sequences and series, we have:

$$\frac{1+\dfrac{\gamma}{2}}{1-\dfrac{\gamma}{2}}=\frac{1+\dfrac{(1+\eta)}{E}\tau}{1-\dfrac{(1+\eta)}{E}\tau} \tag{9.7.61}$$

Therefore:

$$\frac{\gamma}{2}=\frac{1+\eta}{E}\tau \tag{9.7.62}$$

Finally, we fall back on the relation giving the "shear modulus" or "Coulomb module" we had given earlier without proof:

$$\boxed{\frac{\tau}{\gamma}=G=\frac{E}{2(1+\eta)}} \tag{9.7.63}$$

9.7.1.2.3 Compressibility Modulus (bulk modulus)

We still have to see the origin of mathematical expression from another modulus as important as the shear modulus: the bulk modulus.

Given the system of equations determined in the previous study:

$$\begin{aligned}\varepsilon_x&=\frac{1}{E}\left[\sigma_x-\eta(\sigma_y+\sigma_z)\right]\\ \varepsilon_y&=\frac{1}{E}\left[\sigma_y-\eta(\sigma_x+\sigma_z)\right]\\ \varepsilon_z&=\frac{1}{E}\left[\sigma_z-\eta(\sigma_x+\sigma_y)\right]\end{aligned} \tag{9.7.64}$$

If the forces applied on the cube are equal in intensity, we have:

$$\sigma_x=\sigma_y=\sigma_z=\sigma \tag{9.7.65}$$

Which gives us:

$$\begin{aligned}\varepsilon_x&=\frac{1}{E}\left[\sigma-\eta(\sigma+\sigma)\right]=\frac{\sigma}{E}(1-2\eta)\\ \varepsilon_y&=\frac{1}{E}\left[\sigma-\eta(\sigma+\sigma)\right]=\frac{\sigma}{E}(1-2\eta)\\ \varepsilon_z&=\frac{1}{E}\left[\sigma-\eta(\sigma+\sigma)\right]=\frac{\sigma}{E}(1-2\eta)\end{aligned} \tag{9.7.66}$$

Summing the terms according to the principle of linear superposition of forces:

$$\varepsilon_x+\varepsilon_y+\varepsilon_z=\frac{3\sigma}{E}(1-2\eta) \tag{9.7.67}$$

But:

$$\varepsilon_x+\varepsilon_y+\varepsilon_z=\frac{|\Delta L_x|}{L_{x0}}+\frac{|\Delta L_y|}{L_{y0}}+\frac{|\Delta L_z|}{L_{z0}} \tag{9.7.68}$$

That is:

$$\varepsilon_x L_{x0}L_{y0}L_{z0}+\varepsilon_y L_{x0}L_{y0}L_{z0}+\varepsilon_z L_{x0}L_{y0}L_{z0}=|\Delta L_x|L_{y0}L_{z0}+|\Delta L_y|L_{x0}L_{z0}+|\Delta L_z|L_{x0}L_{y0} \tag{9.7.69}$$

Therefore:

$$V_0(\varepsilon_x+\varepsilon_y+\varepsilon_z)=\Delta V \tag{9.7.70}$$

So:

$$\varepsilon_x+\varepsilon_y+\varepsilon_z=\frac{\Delta V}{V_0} \tag{9.7.71}$$

Finally, we find for the relative change in volume of a solid after deformation:

$$\boxed{\frac{\Delta V}{V_0}=\sigma\frac{3(1-2\eta)}{E}} \tag{9.7.72}$$

That we also write sometimes

$$\varepsilon_V=\sigma\frac{3(1-2\eta)}{E}=\sigma\kappa \tag{9.7.73}$$

Therefore:

$$\boxed{\kappa=\frac{\varepsilon_V}{\sigma}} \tag{9.7.74}$$

with κ being by definition the "compression coefficient".

We then define the "bulk modulus" by:

$$\boxed{K=\frac{E}{3(1-2\eta)}} \tag{9.7.75}$$

Let us notice that when the Poisson coefficient η is about 0.33 then K is equal to E (metallic materials are close to this case) and when the Poisson's ratio approaches 0.5 then K approaches infinity and thus the material is incompressible (elastomers approach such an incompressible behaviour).

9.7.1.2.4 Flexural Modulus (bending modulus)

The "flexural modulus" or "bending modulus" is the ratio of stress to strain in flexural deformation, or the tendency for a material to bend. It is determined from the slope of a stress-strain curve produced by a flexural test, and uses units of force per area. It is an intensive property

To study the flexural modulus we consider the following situation:

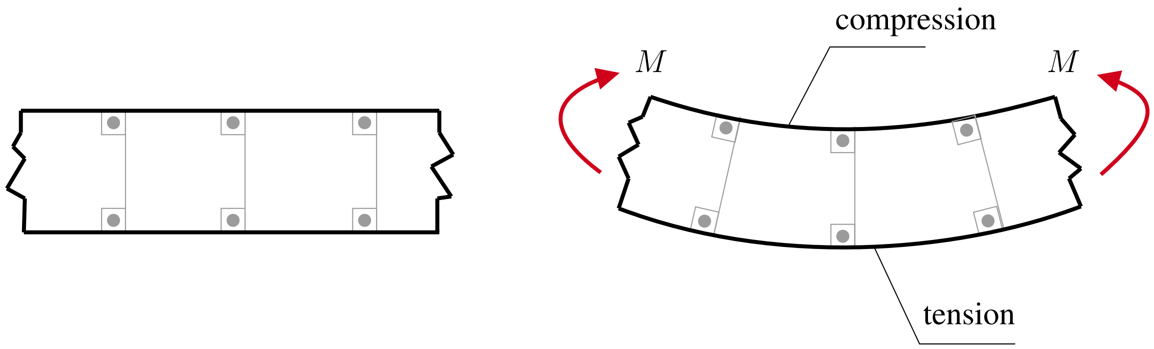


Figure 9.202: Example of a bar under flexural deformation

The above left figure shows a material in a static state. The right figure shows the same material but subject to a couple force momentum M (the height variation of the middle point is named the "deflection").

As the material undergoes on its surface a compression and at the opposite side a tension, so it should exist a border (a line or plane) where no constraint exists. This line or plane (it's rare that we are dealing with a material having only two dimensions ...) is named "neutral plane" (N.P.). This neutral plane will serve us as a reference for defining the bending stress.

Now that this plane is defined, let us consider the following figures:

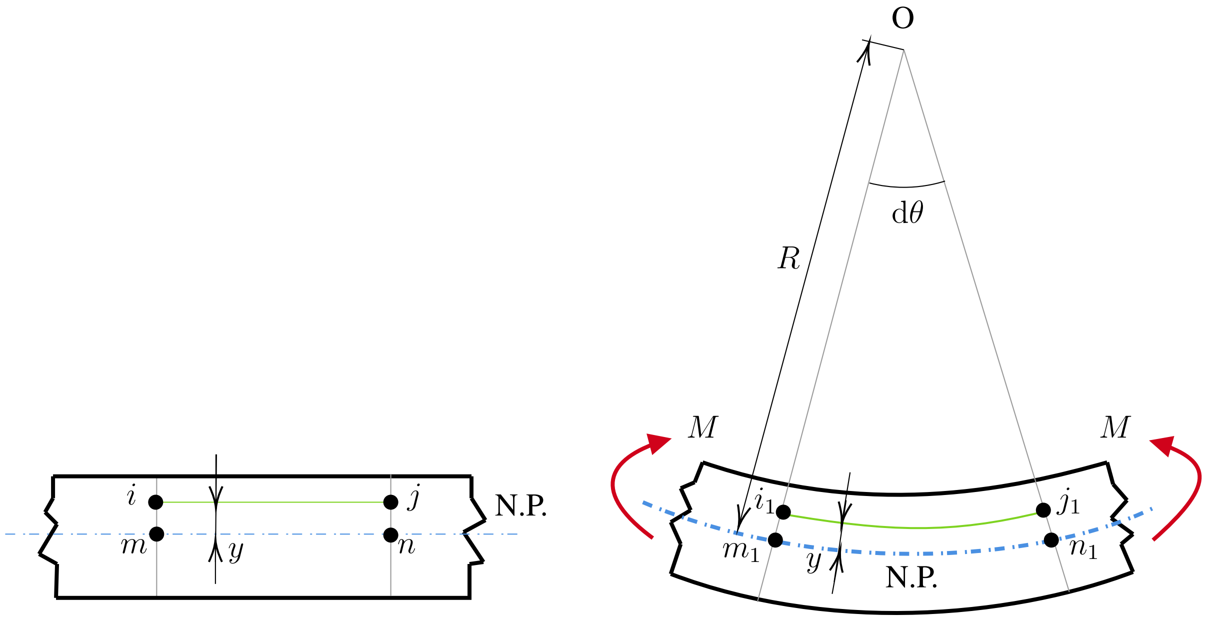


Figure 9.203: Illustration of the neutral plane to determine the flexural modulus

Given R be the bar (cylinder, plate, parallelepiped, ...) radius of curvature. The deformation on the segment $\overline{i_1j_1}$ defined by the relation:

$$\varepsilon_{NP}=\frac{\overline{i_1j_1}-\overline{ij}}{ij} \tag{9.7.76}$$

The lengths $\overline{mn}$ and $\overline{ij}$ are defined by:

$$\overline{ij}=\overline{mn}=\overline{m_1n_1}=R\mathrm{d}\theta \tag{9.7.77}$$

and the length $\overline{i_1j_1}$ by:

$$\overline{i_1j_1}=(R-y)\mathrm{d}\theta \tag{9.7.78}$$

thus expression of the deformation becomes:

$$\varepsilon_{NP}=\frac{(R-y)\mathrm{d}\theta-R\mathrm{d}\theta}{R\mathrm{d}\theta}=-\frac{y}{R} \tag{9.7.79}$$

indicating that the deformation varies linearly with y.

We can define the flexural modulus by:

$$E=\frac{\sigma}{\varepsilon_{NP}}=\frac{-\sigma R}{y} \tag{9.7.80}$$

Let us consider the static state of the bar. The sum of traction and compression constraints are therefore zero. Indeed, we see it well if we look at the diagram below:

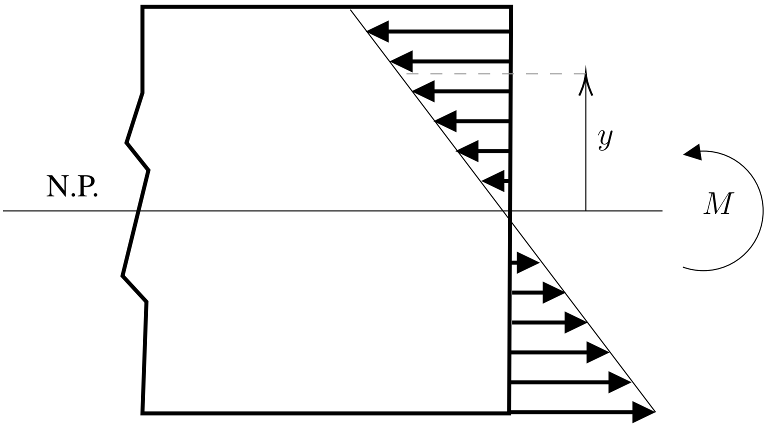


Figure 9.204: **Zoom on the flexion plane**

Let us consider $\sigma\mathrm{d}S$ the force acting on a surface element $\mathrm{d}S$. We can consider the equilibrium of forces in the static state such that:

$$\sum F=\int\limits_S\sigma\mathrm{d}S=0 \tag{9.7.81}$$

Substituting the expression of the constraint obtained previously:

$$\int\limits_S-\frac{Ey}{R}\mathrm{d}S=-\frac{E}{R}\int_S y\mathrm{d}S=0 \tag{9.7.82}$$

by assuming linear the strain characteristic in a first approximation, therefore $E=c^{te}$.

By simplifying a little bit:

$$\int y\mathrm{d}S=0 \tag{9.7.83}$$

If we multiply the integral by σ then the relation should be equal to the force momentum (torque) applied such that:

$$\int_S y\mathrm{d}S=M \tag{9.7.84}$$

By substituting by the expression of the stress obtained previously:

$$M=\int_S \sigma y\mathrm{d}S=\int_S -\frac{Ey^2}{R}\mathrm{d}S=-\frac{E}{R}\int_S y^2\mathrm{d}S \tag{9.7.85}$$

Which brings us to define the term:

$$\boxed{I=\int y^2\mathrm{d}S} \tag{9.7.86}$$

that engineers named the "moment of inertia of the bar relative to the neutral plane" or "static moment of inertia". This term is a measure of the rigidity of the cross section of the bar from a geometric point of view, without consideration of material properties.

Substituting this relationship in the bending stress equation, we get the general "flexural modulus" relation:

$$\boxed{E=\frac{MR}{I}=-\frac{\sigma R}{y}} \tag{9.7.87}$$

The challenge for the engineer is often to locate the plane mathematically neutral ... We will see how to calculate the latter in some special cases in the section of Mechanical Engineering.

9.7.1.2.5 Tranverse Wave in Solids

The transverse sound waves or "S-waves" ("shear waves"), at the opposition of "P-waves" that are compression/dilation type wave[29], for occur only in solids and typical of seismic events. The successive layers of the medium move laterally without volume, density or pressure, change:

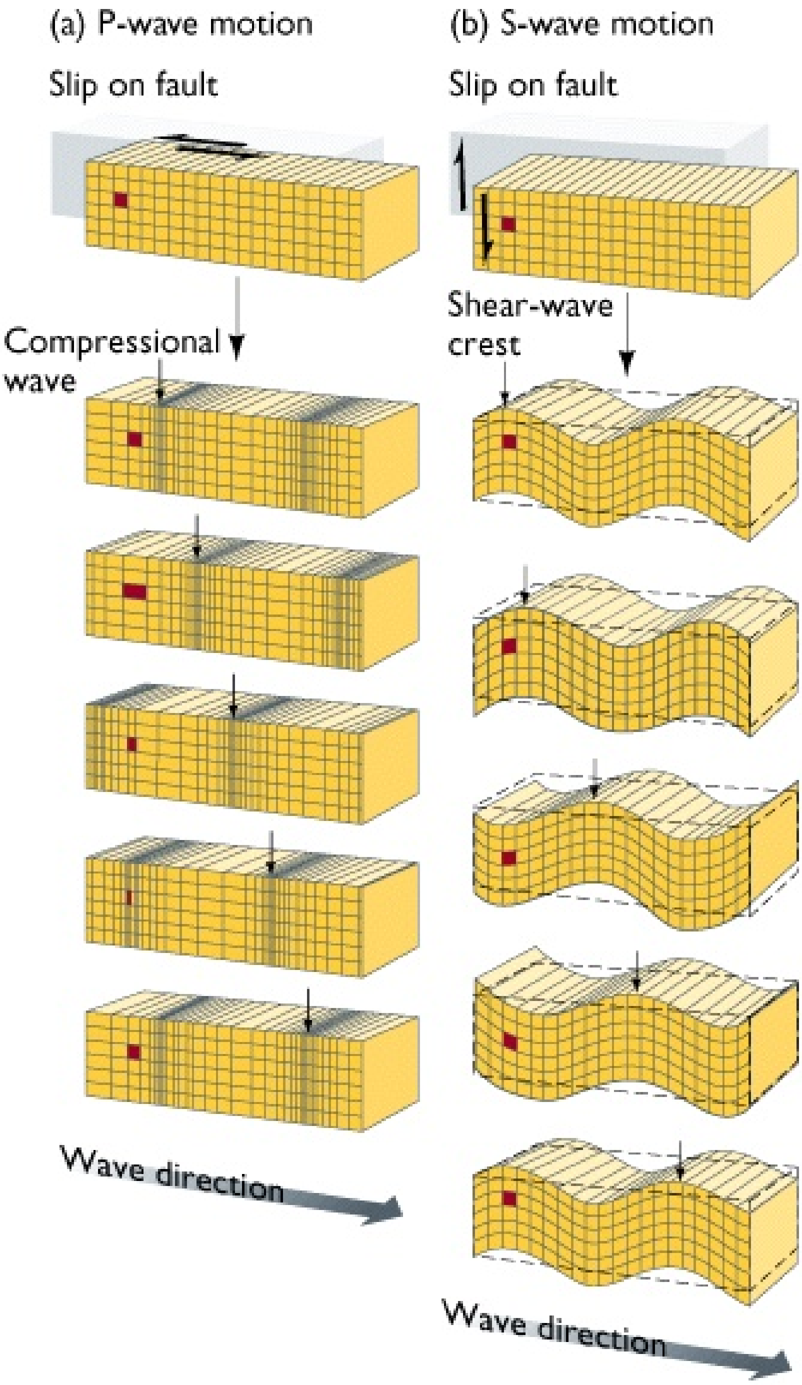

(a) P-wave motion
Slip on fault
Compressional wave
Wave direction
(b) S-wave motion
Slip on fault
Shear-wave crest
Wave direction

Figure 9.205: **P-wave and S-Wave in solids**

Or for those who prefer a better quality illustration:

---

[29] "S" stands for "secondary" and the "P" form "primary"

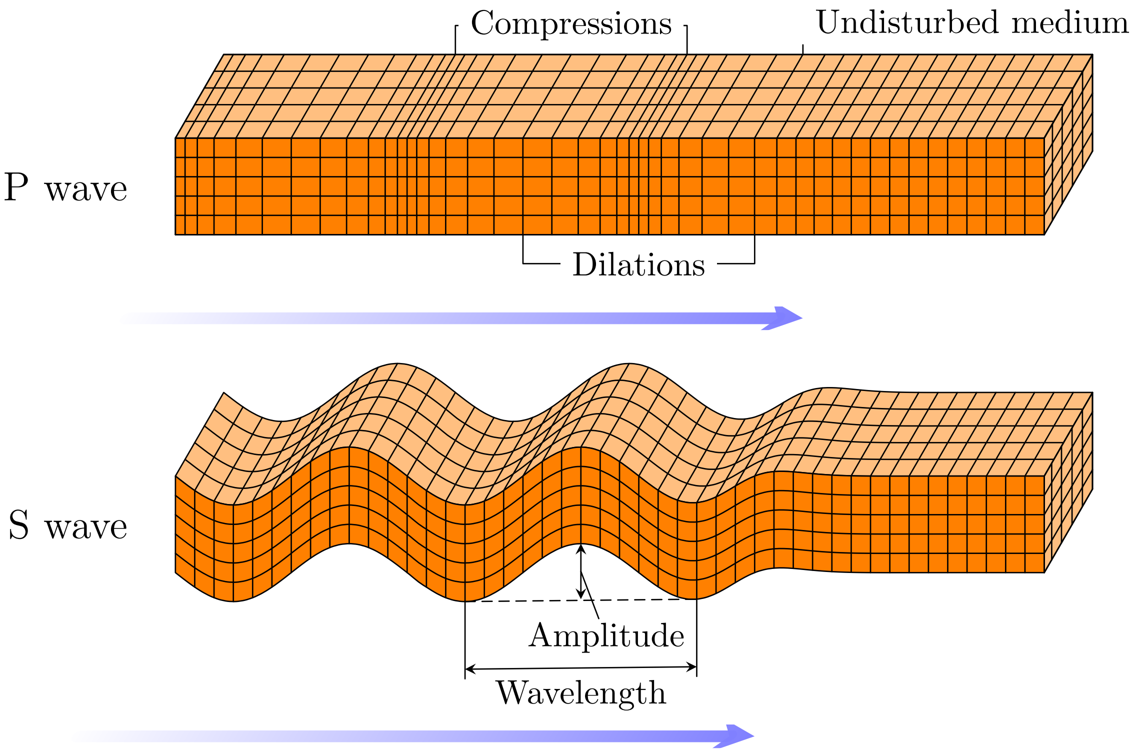

Compressions
Undisturbed medium
P wave
Dilations
S wave
Amplitude
Wavelength

Figure 9.206: P-wave and S-Wave in solids made with TikZ

and focusing only on the S-waves we have for the front view schematically:

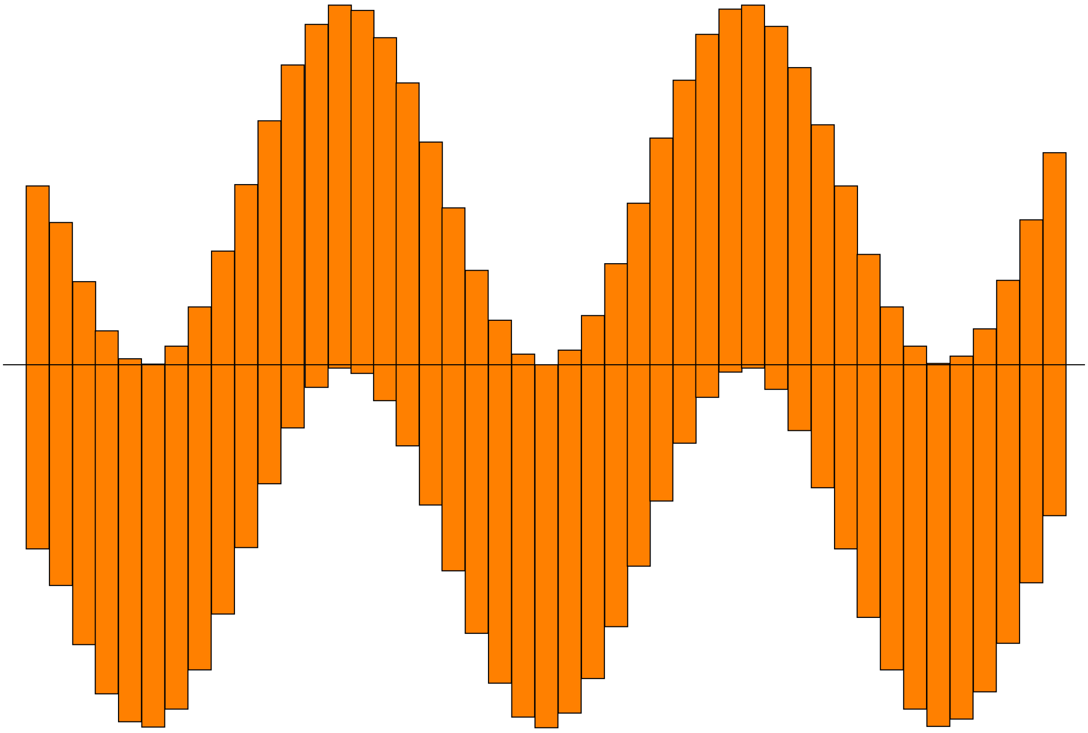


Figure 9.207: S-Wave front view

The medium is deformed in the same way that you can distort a book or ream of paper laid both on a flat table by pushing up horizontally the cover or respectively the first paper page. Neither the book nor the train not change volume!

The way to obtain the wave equation for transverse waves is almost the same as for a string (see section Wave Mechanics page 3487). Let us take three contiguous thin planar layers of the medium (see figure below):

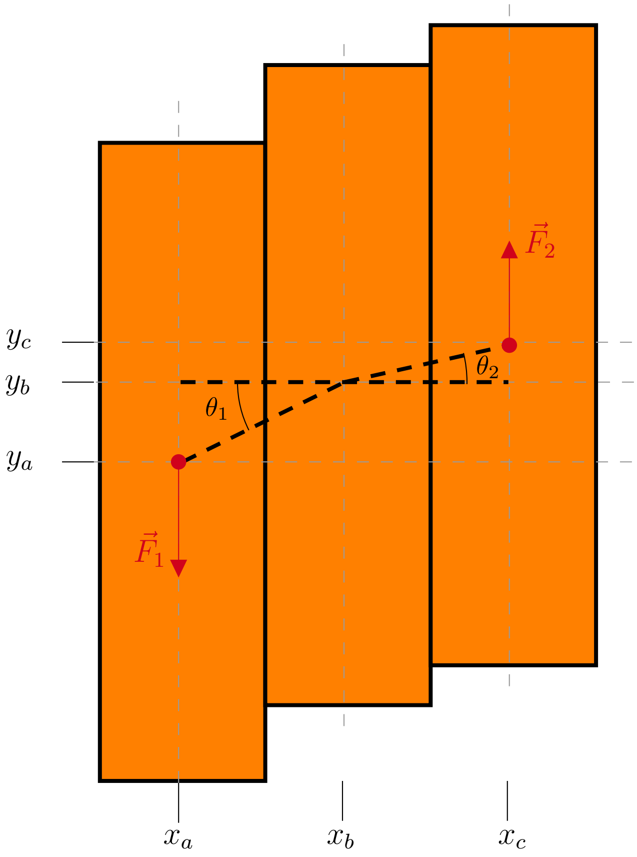


Figure 9.208: **Zoom on three layers of a shear wave**

The centers of the layers being on x_a, x_b, x_c with:

$$x_c - x_b = x_b - x_a = \Delta x \tag{9.7.88}$$

The transversely displacement of the three adjacent layers is y_a, y_b, y_c. The angles of deformation respectively between the layer b and the layer a, and, between the layer c and layer b are in the first order Taylor approximation (see section Sequences and Series page 1703):

$$\begin{aligned}\theta_1 &= \arctan\left(\frac{y_b - y_a}{x_b - x_a}\right) \cong \frac{y_b - y_a}{x_b - x_a}\\ \theta_2 &= \arctan\left(\frac{y_c - y_b}{x_c - x_b}\right) \cong \frac{y_c - y_b}{x_c - x_b}\end{aligned} \tag{9.7.89}$$

If we calculate the forces between the layers for a piece of layer of surface S, we get:

$$\begin{aligned}\frac{F_1}{S}&=G\frac{y_b-y_a}{\mathrm{d}x}\\ \frac{F_2}{S}&=G\frac{y_c-y_b}{\mathrm{d}x}\end{aligned} \tag{9.7.90}$$

where G is as we know the shear modulus of the medium. The resultant force is then:

$$F_b=F_2-F_1=SG\left(\frac{y_c-y_b}{\mathrm{d}x}-\frac{y_b-y_a}{\mathrm{d}x}\right) \tag{9.7.91}$$

The force of the layer F_b will be equal at any time to the product of the mass of the piece of layer b, of thickness $\mathrm{d}s$, of surface S and density ρ, multiplied by the acceleration of the layer given by:

$$\left(\frac{\partial^2 y}{\partial t^2}\right)_b \tag{9.7.92}$$

Then we have:

$$F_b=ma_b=\rho S\mathrm{d}x\left(\frac{\partial^2 y}{\partial t^2}\right)_b=SG\left(\frac{y_c-y_b}{\mathrm{d}x}-\frac{y_b-y_a}{\mathrm{d}x}\right) \tag{9.7.93}$$

Which gives:

$$\left(\frac{\partial^2 y}{\partial t^2}\right)_b=\frac{G}{\rho}\frac{\left(\dfrac{y_c-y_b}{\mathrm{d}x}-\dfrac{y_b-y_a}{\mathrm{d}x}\right)}{\mathrm{d}x}=\frac{G}{\rho}\left(\frac{\partial^2 y}{\partial x^2}\right)_b \tag{9.7.94}$$

What we have just deduce for any value x_b is true for any coordinate:

$$\boxed{\frac{\partial^2 y}{\partial t^2}=\frac{G}{\rho}\frac{\partial^2 y}{\partial x^2}} \tag{9.7.95}$$

and the transverse wave velocity is therefore:

$$\boxed{v=\sqrt{\frac{G}{\rho}}} \tag{9.7.96}$$

Indeed, the ratio G/ρ has the units of the square of a speed:

$$\frac{G}{\rho}=\frac{\dfrac{\mathrm{N}}{\mathrm{m}^2}}{\dfrac{\mathrm{kg}}{\mathrm{m}^3}}=\frac{\dfrac{\mathrm{kg}\cdot\mathrm{m}}{\mathrm{m}^2\cdot\mathrm{s}^2}}{\dfrac{\mathrm{kg}}{\mathrm{m}^3}}=\frac{\mathrm{kg}\cdot\mathrm{m}\cdot\mathrm{m}^3}{\mathrm{m}^3\cdot\mathrm{s}^2\cdot\mathrm{kg}}=\frac{\mathrm{m}^2}{\mathrm{s}^2} \tag{9.7.97}$$

It we finish therefore with a wave equation (recall) of the form of a Poisson's equation (more particularly it is a "d'Alembert's equation" as we know), i.e.:

$$\frac{\partial^2 y}{\partial x^2}=\frac{1}{v^2}\frac{\partial^2 y}{\partial t^2} \tag{9.7.98}$$

Transverse waves propagate only in solids and therefore we can not hear them unless we transform them into longitudinal waves (P waves) by mechanical or electrical means. The transverse waves can be transmitted along a bar or rod of any or even a metal wire, and this without the need that the latter is under stress. Even if the metal wire is under stress, the speed of shear waves is not dependent on the stress (obviously in a given limit...). This is the high shear modulus G of steel that gives electric guitars that characteristic sound.

Another remarkable case of transverse waves (shear waves) is the seismic wave has we have already mention it. There are seismic shear waves and also longitudinal waves or pressure as we have show it also earlier. Shear waves propagate in the Earth's crust at the speed of $3,600\,[\text{m}\cdot\text{s}^{-1}]$ and the pressure wave at the speed of $6,000\,[\text{m}\cdot\text{s}^{-1}]$. During an earthquake or an atomic explosion, the two types of waves will be produced, but as the waves travel at different speeds, they will not arrive at the same time at remote sensing stations. It is from this difference in arrival times that we can determined the distance at the epicenter. The direction is obtained from the direction of the oscillations. Only sufficiently distant stations to receive both types of wave separately can make the determination of the epicenter.

To summarize, we have for longitudinal waves in a solid (see section Music Mathematics page 8202):

$$\boxed{\frac{\partial^2\xi}{\partial t^2}=\frac{E}{\rho}\frac{\partial^2\xi}{\partial x^2}} \tag{9.7.99}$$

and for transverse waves:

$$\boxed{\frac{\partial^2 y}{\partial t^2}=\frac{G}{\rho}\frac{\partial^2 y}{\partial x^2}} \tag{9.7.100}$$

For the details of the mathematical developments in the gas and solid about propagation waves, the reader should read the section of Mathematical Music (Acoustic) page 8195.

9.7.2 Liquids (fluid mechanics)

A liquid is a nearly incompressible fluid[30] that conforms to the shape of its container but retains a (nearly) constant volume independent of pressure. As such, it is one of the four fundamental states of matter (the others being solid, gas, and plasma), and is the only state with a definite volume but no fixed shape. A liquid is made up of tiny vibrating particles of matter, such as atoms, held together by intermolecular bonds. Water is, by far, the most common liquid on Earth. Like a gas, a liquid is able to flow and take the shape of a container. Most liquids resist compression, although others can be compressed. Unlike a gas, a liquid does not disperse to fill every space of a container, and maintains a fairly constant density. A distinctive property of the liquid state is surface tension, leading to wetting phenomena.

The density of a liquid is usually close to that of a solid, and much higher than in a gas. Therefore, liquid and solid are both termed condensed matter. On the other hand, as liquids and gases share the ability to flow, they are both named "fluids". Although liquid water is abundant on Earth, this state of matter is actually the least common in the known universe, because liquids require a relatively narrow temperature/pressure range to exist. Most known matter in the universe is in gaseous form (with traces of detectable solid matter) as interstellar clouds or in plasma form within stars.

The distinction between liquid and gas is subtle. However, we can say that the proper volume of liquid manifests the existence of cohesiveness due to a fairly high density (van der Waals bonds); this cohesion disappears with the proper volume of a gas.

If we compare the fluids with solids, the first that comes concerns the isotropy (properties are the same in all spatial directions) of conventional fluids that is always satisfied (at least if we do not act on the fluid directly!).

In what follows, we will address the fluid mechanics theory in increasing difficulty and redundancy by trying to keep it simple and stupid as possible. First, we will be prove that the properties of a static fluid are isotropic (Pascal's theorem). Using this result, it will be easier for us to understand Bernoulli's theorem that will allow us, among others, to define the concept of "hydrostatic pressure". Then we will build a very important model for fluid dynamics, known as the "Navier-Stokes equations", which has usages almost all possible existing areas of science (astrophysics, quantum mechanics, meteorology, ...). This fluid dynamics model is quite big in theoretical developments and experimental results and can be considered as difficult to approach. However, for ease of reading, we chose not to address it by use of tensor calculus. We have made sure that the tensor variables appear by themselves following the simple previous vector analysis results we will get. Once the Navier-Stokes equations determined and proved, we will see that we can find the expression of Bernoulli's theorem from these equations.

---

[30] In fact nothing as far as we know is absolutely incompressible but when in the study we do the relative variations of pressure and speed a small enough we consider the fluid as being incompressible.

The Navier-Stokes equations and finally used with Finite Element Methods to make numerical simulations of shapes to optimize the penetration or movement of objects in water or in the air:

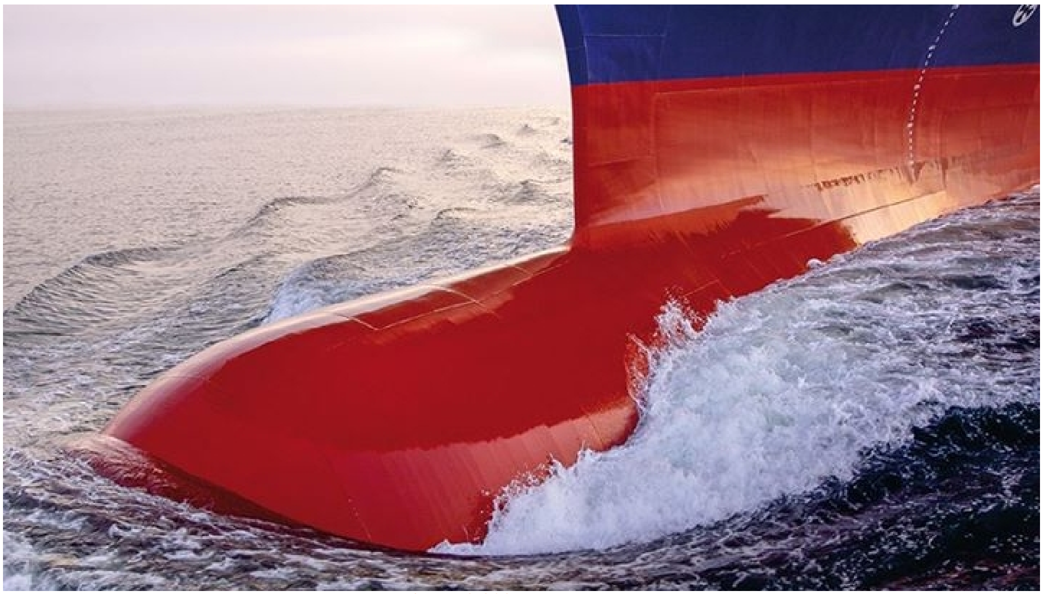


Figure 9.209: **Bulbous bow**

On the picture above we see a bulbous bow that is a protruding bulb at the bow (or front) of a ship and using the concept of destructive interference of waves to cancel the wave produced at the back by the bow. Large ships with bulbous bows generally have a twelve to fifteen percent better fuel efficiency than similar vessels without them.

Fluid dynamics, or "hydrodynamic" is by far the field of Classical Mechanics the less easy to understand as regards to the descriptions and predictions. This is why the Bernoulli's theorem is often used, not to explain in detail the behaviour of a fluid, but to make a qualitative description for it.

Caution! To determine whether Statistical Mechanics or the Continuum Mechanics formulation of fluid dynamics should be used to model a situation, we use the "Knudsen number". This is a dimensionless number defined as the ratio of the molecular mean free path length to a representative physical length scale. This length scale could be, for example, the radius of a body in a fluid. If the Knudsen number is near or greater than one, the mean free path of a molecule is comparable to a length scale of the problem, and the continuum assumption of fluid mechanics is no longer a good approximation. In such cases, statistical methods should be used.

Commonly for the average individual, "fluids" are always associated with liquids. However in physics "fluid" is a liquid, gas, or other material that may continuously move and deform (flow) under an applied shear stress, or external force. They have zero shear modulus, or, in simpler terms, are substances which cannot resist any shear force applied to them. However in geology it's a little bit more complicated than this because some things that we consider to be solid will flow if given enough time.

Below we can see the gas dioxide nitrogen (NO_2) flowing from an Erlenmeyer flask into a beaker initially full of air (dioxide nitrogen has a mass and therefore a weight due to gravity greater than those of the different components that air is made of):

100
APPROX
80
60
40
20

9.7.2.1 Pascal's Theorem

Definition (#9.7.472): The "Pascal's law" or the "principle of transmission of fluid-pressure" (also named "Pascal's Principle") is a principle in fluid mechanics that states that a pressure change occurring anywhere in a confined incompressible fluid is transmitted throughout the fluid such that the same change occurs everywhere.

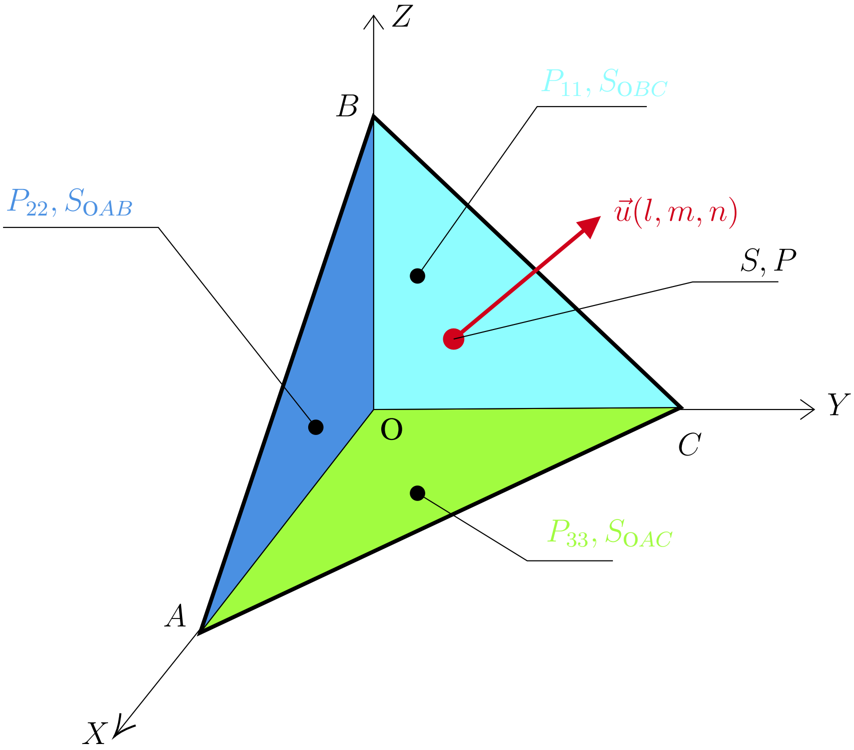


Figure 9.210: **Elementary tetrahedron to introduce Pascal's theorem**

If we consider the forces acting, in the absence of movement on an elementary tetrahedron OABC of volume V, it is always possible to choose a sufficiently small volume in order to have a uniform pressure on all the faces of the tetrahedron .

Given $P_{11}, P_{22}, P_{33}, P$, the fluid reaction pressures due to external constraints soliciting the respective faces OBC, OAB, OAC and ABC of respecting surfaces $S_{\text{O}BC}, S_{\text{O}AB}, S_{\text{O}AC}, S$. Given also the cosine direction (see section Vector Calculus page 1811) of the unit vector normal $\vec{n}$ to the surface ABC.

The system being in equilibrium, the resultant $\vec{R}$ of the of the forces of reaction of the system is zero. So we have the following equations resulting from the projection along the three coordinate axes:

$$\begin{aligned}R_x &= P_{11}S_{\text{O}BC} - PS_{\text{O}BC} = 0\\ R_y &= P_{22}S_{\text{O}AB} - PS_{\text{O}AB} = 0\\ R_y &= P_{33}S_{\text{O}AC} - PS_{\text{O}AC} = 0\end{aligned} \tag{9.7.101}$$

By elementary simplification, it comes:

$$\begin{aligned} P-P_{11}&=0\\ P-P_{22}&=0\\ p-P_{33}&=0 \end{aligned} \tag{9.7.102}$$

We then get the following relation:

$$\boxed{P=p_{11}=p_{22}=p_{33}} \tag{9.7.103}$$

The major conclusion is that in any point of a fluid pressure is independent of the direction of the normal $\vec{n}$ to the elementary surface on which it is exercised.

In other words, by the Newton's principle of action and reaction, we are led to state the "Pascal's theorem" and "Pascal's principle": The incompressible fluids (in equilibrium) transmit integrally and in all directions, the pressures applied to them.

This theorem is fundamental both in fluid mechanics and gas mechanics and the practical implications are huge. This theorem explains among others the famous "hydrostatic paradox" as what the pressure of a liquid on the bottom of a container is independent of its shape, and also of the bottom surface but only depends on the water level in the recipient. This theorem is the basis for the design of many hydraulic machines!

Let's look at a very clear illustration. Consider the following diagram (we made an example in U, although if the majority of hydraulic systems are linear):

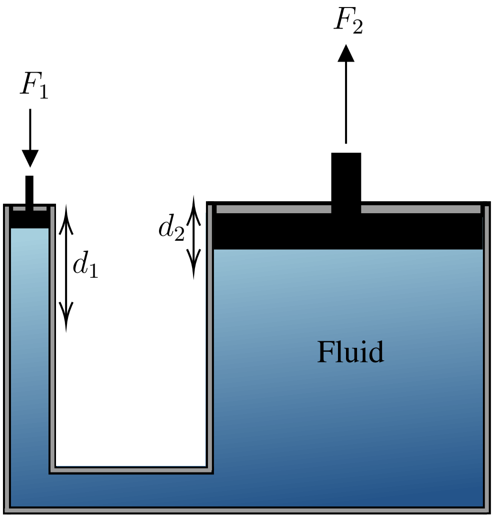


Figure 9.211: **Typical hydraulic machine (elevator, hydraulic jack)**

What is happening exactly? Well on the left, we have the force F_1 that generates the pressure:

$$P_1=\frac{F_1}{S_1} \tag{9.7.104}$$

and by the Pascal's principle, we must find the same pressure on the other side such that:

$$P_2=P_1=\frac{F_2}{S_2} \tag{9.7.105}$$

It follows immediately the very important relation in practice:

$$\boxed{P_1S_2=F_1\frac{S_2}{S_1}=F_2} \qquad (9.7.106)$$

So we have a force multiplier! Warning! The principle of the hydraulic machine is to have a fantastic force multiplier but in no case it multiplies the work W! So by the principle of conservation of energy, we have:

$$W_1=F_1d_1=W_2=F_2d_2 \qquad (9.7.107)$$

and then it comes:

$$d_1=\frac{F_1}{F_2}d_2=\frac{P_1/S_1}{P_2/S_2}d_2=\frac{P_1/S_1}{P_1/S_2}d_2=\frac{S_2}{S_1}d_2 \qquad (9.7.108)$$

Therefore, in counterpart of the force multiplication effect, the energy conservation, the primary force must travel a greater distance.

9.7.2.2 Viscosity

The viscosity of a fluid is a measure of its resistance to gradual deformation by shear stress or tensile stress. For liquids, it corresponds to the informal concept of "thickness"; for example, honey has a much higher viscosity than water.

Viscosity is a property arising from collisions between neighbouring particles in a fluid that are moving at different velocities. When the fluid is forced through a tube, the particles which compose the fluid generally move more quickly near the tube's axis and more slowly near its walls; therefore some stress (such as a pressure difference between the two ends of the tube) is needed to overcome the friction between particle layers to keep the fluid moving. For a given velocity pattern, the stress required is proportional to the fluid's viscosity.

A fluid that has no resistance to shear stress is known as an ideal or inviscid fluid. Zero viscosity is observed only at very low temperatures in superfluids. Otherwise, all fluids have positive viscosity, and are technically said to be viscous or viscid. In common parlance, however, a liquid is said to be viscous if its viscosity is substantially greater than that of water, and may be described as mobile if the viscosity is noticeably less than water. A fluid with a relatively high viscosity, such as pitch, may appear to be a solid.

In fluid mechanics, it is useful to consider several types of fluids with characteristics that differentiate them. This is especially useful for simulations while remaining consistent with the experimental observations (see section Marine & Weather Engineering page 8239).

Definitions (#9.7.473): We define the "viscosity" μ by the internal forces opposing the movement of the various layers composing the fluid. We distinguish two type of viscosity:

D1. The "dynamic viscosity μ_d" also simply named "coefficient of viscosity":

$$\boxed{\mu_d=\frac{\mathrm{d}F}{\partial_z v\mathrm{d}S}} \qquad (9.7.109)$$

which unit is the Poiseuille: [PI] (equivalent to the pascal-second $[\text{Pa}\cdot\text{s}]$, or $[\text{N}\cdot\text{s}\cdot\text{m}^{-2}]$, or $[\text{kg}\cdot\text{m}^{-1}\cdot\text{s}^{-1}]$), $\mathrm{d}F$ the variation of the frictional force between two infinitely neighbouring layers, $\partial_z v$ being the variation in speed by the distance between two infinitely neighbouring layers and in $[\text{s}^{-1}]$ and $\mathrm{d}S$ being the considered surface $[\text{m}^2]$.

If a fluid is placed between two plates with a distance of 1 [m], and one plate is pushed sideways with a shear stress of one [Pa], and it moves at v meter per second, then it has a viscosity of $1/v$ $[\text{Pa}\cdot\text{s}]$.

For example, water at 293 [K] has a dynamic viscosity of 0.001 $[\text{Pa}\cdot\text{s}]$ while a typical motor oil could have a viscosity of about 0.25 $[\text{Pa}\cdot\text{s}]$. Here are some more dynamics viscosity values:

Material	**Temperature [K]**	μ_d **[Pa · s]**
Water	273	$1.8\cdot10^{-3}$
Water	293	$1.0\cdot10^{-3}$
Water	313	$0.7\cdot10^{-3}$
Alcohol (ethanol)	293	$1.2\cdot10^{-3}$
Glycerine	293	1.5
Light machine oil	313	$3.4\cdot10^{-2}$
Weight machine oil	313	$13.0\cdot10^{-2}$
Engine oil	313	$2.4\cdot10^{-1}$
Engine oil	373	$0.2\cdot10^{1}$
Mercury	293	$1.6\cdot10^{-3}$
Air	273	$1.7\cdot10^{-5}$
Air	293	$1.8\cdot10^{-5}$
Air	313	$1.9\cdot10^{-5}$
Water Vapour	393	$1.25\cdot10^{-5}$
Water Vapour	493	$1.64\cdot10^{-5}$
Ammonia	293	$1\cdot10^{-5}$
Nitrogen	293	$1.75\cdot10^{-5}$
Carbon dioxide	293	$1.5\cdot10^{-5}$
Hydrogen	293	$0.9\cdot10^{-5}$
Oxygen	293	$2\cdot10^{-5}$

Table 9.34: **Some dynamic viscosity values**

Conclusion: The Poiseuille is the viscosity of a fluid needing a force of 1 Newton to slide at a speed of 1 meter per second, two fluid layers of 1 square meter by a distance of 1 meter.

A transformation of the definition of dynamic viscosity gives (this relations must be remembered for later use!!):

$$\boxed{\tau=\frac{\mathrm{d}F}{\mathrm{d}S}=\mu_d\partial_z v} \tag{9.7.110}$$

That is:

$$\partial_z v=\frac{1}{\mu_d}\tau \tag{9.7.111}$$

D2. The "kinematic viscosity μ_k":

$$\boxed{\mu_k=\frac{\mu_d}{\rho}} \tag{9.7.112}$$

By definition fluids having the following characteristics:

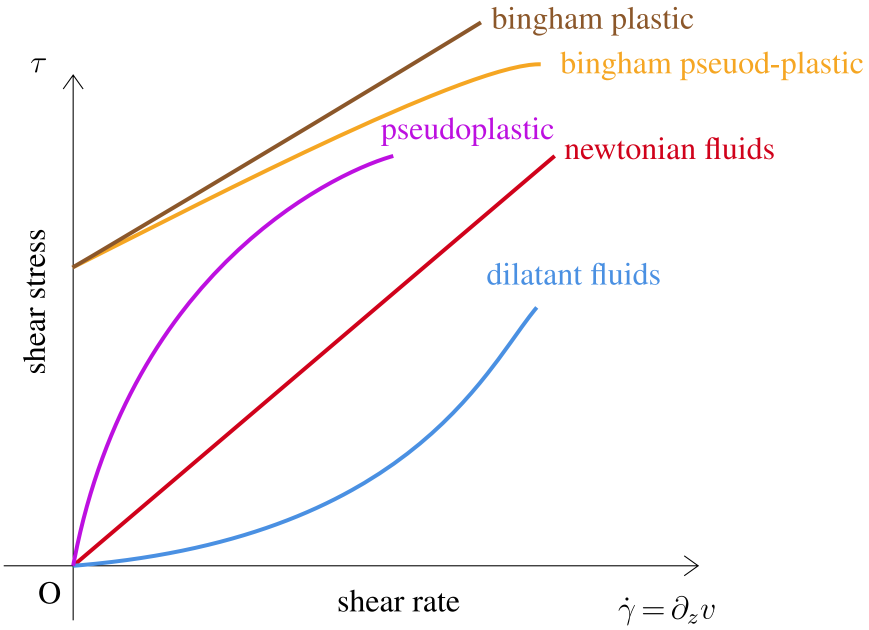


Figure 9.212: **Very approximative viscosity characteristics of different typical fluids**

There are still three other main types of fluids (among others!) not shown in the figure and whose viscosity is assumed to be zero. We then speak of "Pascalian fluids":

- "Perfect fluids": A fluid that can be completely characterized by its rest frame mass density ρ and isotropic pressure p. Specifically, perfect fluids have no shear stresses, viscosity, or heat conduction.
- "Semi-perfect fluids": A fluid that eventually conduct heat and whose heat capacity is only dependent of temperature.
- "Real fluids": A fluid with all possible complications that we observe in real life...

Remarks

R1. The behaviour of a perfect fluid is very different from that of a real fluid as small as the viscosity can be for the latter. Indeed, the perfect fluid because it has no viscosity, never dissipate kinetic energy. While a real fluid with a small viscosity effectively dissipates it very efficiently through turbulences.

R2. We will come back on the properties of the dynamic and kinematic viscosity during the proof of the Navier-Stokes-(Reynolds) equations.

R3. The fluids that are not Newtonians are named very generally in the literature "non-Newtonian fluids" ... and we will not deal with Ferrofluids in this textbook because they are theoretically outside the range of this textbook.

Non-Newtonian fluids thus have a deformation which depends on the force that we apply to them. The best example is the wet sand by the sea: when we hit the sand, it has the high viscosity of a solid, whereas when we quietly press on it, it behaves like a paste. Moreover, some non-Newtonian fluids have properties such that it is possible for an individual to run on it without sinking or sink staying in position...

9.7.2.2.1 Poiseuille's law

In 11835 (holocene calendar) a French doctor, Jean Leonard Marie Poiseuille, made a series of careful experiments to determine how a viscous fluid flows through a narrow pipe. His goal was to understand the dynamics of blood flow in humans. Blood plasma behaves like a Newtonian fluid, while blood does not. Almost half of the volume of the normal blood is made large enough cells to disrupt laminar flow, especially when they come in contact with the vessel walls, a phenomenon that becomes important in very narrow capillaries. However, the analysis of Poiseuille applies to flow in the veins and larger arteries and it has great value, although it is a little bit simplistic.

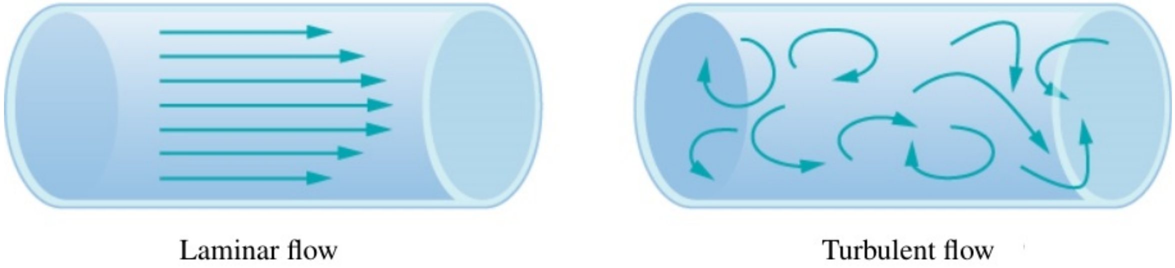


Figure 9.213: Laminar vs Turbulent flow (source: OpenStax)

The result of Poiseuille can be established by considering the fluid in a pipe as formed by concentric cylindrical layers oriented along an x-axis of radius r moving at speeds which decrease from the center (circular symmetry assumed).

Then the relation defining the viscosity is:

$$\boxed{F=\mu_d S\frac{\mathrm{d}v_x}{\mathrm{d}r}} \tag{9.7.113}$$

Remark

In some textbooks the definition is given as:

$$\tau=\frac{F}{S}=\mu_d\frac{\mathrm{d}v_x}{\mathrm{d}r} \tag{9.7.114}$$

That is the stress τ of a Newtonian fluid is assumed to be directly proportional to the rate at which the velocity field is changing as we move up.

What gives us the viscous force on the cylinder. The contact surface of each cylindrical layer of length l is given obviously by $S=2\pi rl$ and therefore:

$$F=\mu_d 2\pi rl\frac{\mathrm{d}v_x}{\mathrm{d}r} \tag{9.7.115}$$

The origin of the acceleration (i.e. the force) can be done only by a pressure difference such as:

$$\Delta PS=\Delta P\pi r^2=F \tag{9.7.116}$$

which brings us to write:

$$\Delta P\pi r^2=\mu_d 2\pi rl\frac{\mathrm{d}v_x}{\mathrm{d}r}\Rightarrow \mathrm{d}v_x=\frac{\Delta P}{2\mu_d l}r\mathrm{d}r \tag{9.7.117}$$

By integrating term by term, we get:

$$\int_0^{v_x} \mathrm{d}v_x = \frac{\Delta P}{2\mu_d} l \int_r^R r\mathrm{d}r \tag{9.7.118}$$

So we get the famous "law of viscous laminar flow" discovered by the French physicist Jean-Louis Marie Poiseuille in 11840 (holocene calendar):

$$\boxed{v_x = \frac{\Delta P}{4\mu_d l}(R^2 - r^2)} \tag{9.7.119}$$

widely used in medicine to calculate the blood flow speed in the veins, in hydraulic systems (to determine the pressure required to have a certain speed) and in wind tunnels (for the same reason as in hydraulic systems).

Turbulent air is not ideal for jet engines because it's less stable and is therefore more likely to induce vibration on fan blades, causing accelerated wear. So when it comes to fighter jets with intakes mounted rear of the nose (often just under leading edge of the wings), you need to make an effort to raise the intakes out of the way of the boundary layer between turbulent and laminar air:

The representing curve of the speed in function of r is a parabola whose summit is on the cylinder center axis ($r = 0$). That's why the above relation is often denoted:

$$v_x = \frac{v_0}{R^2}(R^2 - r^2) \tag{9.7.120}$$

So the speed is greater at the center than at the edges (that is quite intuitive). Differentiating this last relationship with respect to r, we immediately get the velocity gradient in function of r (calculation useful in practice).

The "volume flow" J (sometimes denoted D in small classes) carried by a cylindrical layer between r and $r + \mathrm{d}r$ is:

$$\boxed{J = \frac{\mathrm{d}V}{\mathrm{d}t} = v_x \mathrm{d}S = v_x(2\pi r\mathrm{d}r)} \tag{9.7.121}$$

Thus, the total flow is:

$$J=\int\limits_{0}^{R}\frac{\Delta P}{4\mu_d l}(R^2-r^2)2\pi r\mathrm{d}r \tag{9.7.122}$$

and we get the "Poiseuille law" for the viscous laminar flow:

$$J=\frac{\pi R^4}{8\mu_d}\frac{\Delta P}{l} \tag{9.7.123}$$

Some books introduce the "resistance to flow" denoted R_f such that the Poiseuille law is finally written:

$$\boxed{J=\frac{\pi R^4}{8\mu_d}\frac{\Delta P}{l}=\frac{\Delta P}{R_f}} \tag{9.7.124}$$

So we find the logical result that the flow rate increases with the pressure gradient $\Delta P/l$ and the radius of the tube, and decreases with the viscosity.

We find also a relation analogous to Ohm's law (see section Electrokinetics page 4233) where the potential difference is given by the resistance multiplied by the current so that the difference in pressure is given by the viscous resistance multiplied by the flaw. We will also come back on this relationship when we deal with the losses pressure in the pipes.

Remark

In the literature, the flow is sometimes denoted J, Q or D or even $\dot{V}$ ... In short, there is no standard notation as far as we know and we must do with ...

✍**Example:**

An air conditioning system is being designed to supply air at a gauge pressure of 0.054 [Pa] at a temperature of 293 [K]. The air is sent through an insulated, round conduit with a diameter of 0.18 [m]. The conduit is 20 [m] long and is open to a room at atmospheric pressure 101.30 [kPa]. The room has a length of 12 [m], a width of 6 [m], and a height of 3 [m] and we assume a constant pressure difference and using the viscosity $\mu_d=0.0181\;[\text{mPa}\cdot\text{s}]$.

The volume flow rate through the pipe, assuming laminar flow is then given by:

$$J=\frac{0.054\cdot\pi(0.09)^4}{8\cdot 0.0181\cdot 10^{-3}\cdot 20}\cong 3.84\cdot 10^{-3}\;[\text{m}^3\cdot\text{s}^{-1}] \tag{9.7.125}$$

The length of time to completely replace the air in the room is then by assuming constant flow equal to:

$$\Delta t=\frac{\Delta V}{J}=\frac{12\cdot 6\cdot 3}{3.84\cdot 10^{-3}}\cong 5.63\cdot 10^4=15.63\;[\text{h}] \tag{9.7.126}$$

The previous relation:

$$J=\frac{\pi R^4}{8\mu_d}\frac{\Delta P}{l}=\frac{\Delta P}{R_f} \tag{9.7.127}$$

Is also denoted in some textbooks:

$$\Phi=\frac{\pi R^4}{8\mu_d}\frac{\Delta P}{l} \tag{9.7.128}$$

and after rearranging we get:

$$\mu_d=\frac{\pi R^4}{8l\Phi}\Delta P \tag{9.7.129}$$

This is an expression well known to experimentally measure the dynamic viscosity of a fluid based on simple parameters (and dividing by ρ by definition we get the kinematic viscosity). Indeed, we just need a tank that has a source of liquid that maintains the liquid hight constant inside the tank and pipe at the bottom from which the liquid flows and know the difference of pressure between the top of the liquid in the tank and pipe separated from a vertical distance h. That's why the above relation can often be found as following in the literature:

$$\boxed{\mu_d=\frac{\pi R^4}{8l\Phi}\Delta P=\frac{\pi R^4}{8l\Phi}mgh} \tag{9.7.130}$$

Although rather a mess of constants appears in our formula for the kinematic viscosity μ_d, all are quite easily measured.

Remark

It seems that there is no analytical method to calculate the pressure drop of a pipe elbow.

9.7.2.2.2 Viscous force on a fluid element

Suppose again that we have a simple x directed velocity field whose velocity profile is the same as in the previous figure above (a horizontally oriented parabola). But as the flow is symmetric we can focus only on half of the pipe (i.e. half of the parabola).

Now let us consider in the bottom symmetric part of the pipe a small volume element with sides Δx, Δy and Δz, the bottom of which is located at y and the top at $y+\Delta y$. The fluid below at y is dragging the fluid above, something that we can write:

$$\tau_-(y)=-\mu_d\frac{\partial_x v_x(y)}{\partial y} \tag{9.7.131}$$

The total force at the bottom is as we already know the stress times the area $\Delta x\Delta y$ upon which it is acting:

$$\Delta F_-(y)=\tau_-(y)\Delta x\Delta z=-\mu_d\frac{\partial_x v_x(y)}{\partial y}\Delta x\Delta z \tag{9.7.132}$$

Respectively the faster fluid pulls forward the slower one with a total force:

$$\Delta F_+(y+\Delta y)=\tau_+(y+\Delta y)\Delta x\Delta z=+\mu_d\frac{\partial_x v_x(y+\Delta y)}{\partial y}\Delta x\Delta z \tag{9.7.133}$$

With the two relation above we see that the total viscous force on the fluid in our volume element can be written:

$$\Delta F_x=\Delta F_-(y)+\Delta F_+(y+\Delta y)=\mu_d\left[-\frac{\partial_x v_x(y)}{\partial y}+\frac{\partial_x v_x(y+\Delta y)}{\partial y}\right]\Delta x\Delta z \tag{9.7.134}$$

Multiplying the right side by $\Delta y/\Delta y$ gives obviously:

$$\Delta F_x=\mu_d\left[\frac{\dfrac{\partial_x v_x(y+\Delta y)}{\partial y}-\dfrac{\partial_x v_x(y)}{\partial y}}{\Delta y}\right]\Delta x\Delta y\Delta z \tag{9.7.135}$$

The quantity in the square brackets should be recognized as the second derivative of $v_x(y)$ with respect to y. Dividing through both sides by the volume $\Delta x\Delta y\Delta z$ gives us the viscous force per unit volume:

$$f_{v_x}=\mu_d\frac{\partial^2 v_x(y)}{\partial y^2} \tag{9.7.136}$$

This is the formula for the viscous force per unit volume acting on the fluid particles when we have a purely x directed flow of a Newtonian fluid whose speed varies only in the y direction.

Now remember that we get earlier the law of viscous laminar flow:

$$v_x=\frac{\Delta P}{4\mu_d l}(R^2-r^2) \tag{9.7.137}$$

Applying the result we just get, this lead us to:

$$f_{v_x}=\mu_d\frac{\partial^2 v_x(r)}{\partial r^2}=\frac{\Delta P}{2l} \tag{9.7.138}$$

9.7.2.3 Bernoulli's Theorem

When we discuss of the motion of a fluid, the continuity equation (see section Thermodynamics page 3679), which expresses the conservation of mass (volume) of the fluid is an important concept.

$$\boxed{\frac{\partial\rho}{\partial t}-\vec{\nabla}\circ(\rho\vec{v}_a)=0} \tag{9.7.139}$$

Let us consider this equation in the particular case that interest us here: a non-viscous fluid in laminar flow moving inside a tube following parallel streams lines (the fluid motion is irrotational type - see section Vector Calculus page 1889), delimited by the surface A_1:

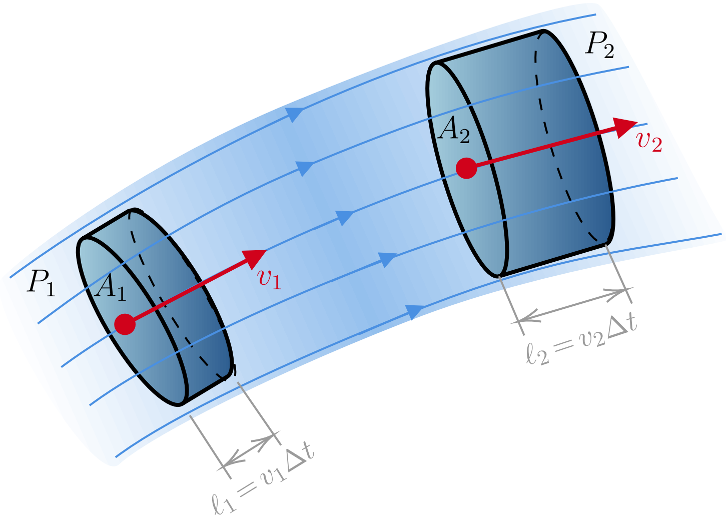


Figure 9.214: Non-viscous fluid in laminar flow from in a tube of parallel lines of current lines converging

We consider that we are in steady state (the appearance of movement is independent of time) and the mass is not provided by a source or removed by a well within the region. The volume of fluid which crosses A_1 in the interval corresponds Δt corresponds to a cylinder of base A_1, and length $v_1\Delta t$ and therefore volume of volume $A_1v_1\Delta t$. The mass of fluid which has passed through A_1 for the time Δt is therefore:

$$M_1 = \rho_1 A_1 v_1 \Delta t \tag{9.7.140}$$

And also:

$$M_2 = \rho_2 A_2 v_2 \Delta t \tag{9.7.141}$$

is the mass of fluid which has passed through A_2 during the same time interval. With the our assumptions, the mass conservation equation requires that two masses are the same, or expressed differently that:

$$\rho_1 A_1 v_1 \Delta t = \rho_2 A_2 v_2 \Delta t \tag{9.7.142}$$

Hence:

$$\boxed{\rho_1 A_1 v_1 = \rho_2 A_2 v_2} \tag{9.7.143}$$

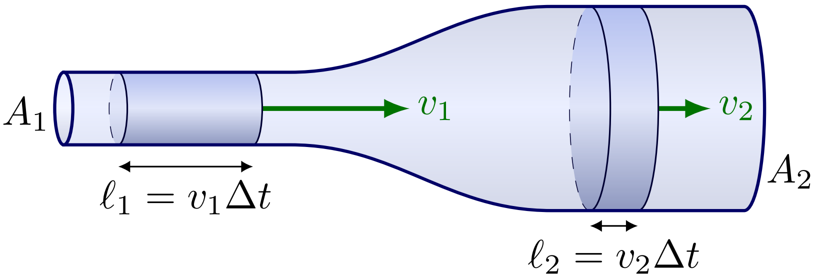


Figure 9.215: Bernoulli's theorem illustrated (author: Izaak Neutelings)

This is the shape of the continuity equation that interest us in our context. In addition, if the fluid is incompressible, the density is everywhere the same and the above relation reduces to:

$$A_1 v_1 = A_2 v_2 \tag{9.7.144}$$

Thus, the ratio of input/output speed in a tubular pipe will be usually (not included frictions!):

$$\frac{v_2}{v_1} = \frac{v_{\text{out}}}{v_{\text{in}}} = \frac{A_1}{A_2} = \frac{\pi R_1^2}{\pi R_2^2} = \frac{R_1^2}{R_2^2} = \frac{R_{\text{in}}^2}{R_{\text{out}}^2} \tag{9.7.145}$$

The continuity equation is indirectly known to all children and adults who play with or use a garden hose (thumb serves as a "convergent"):

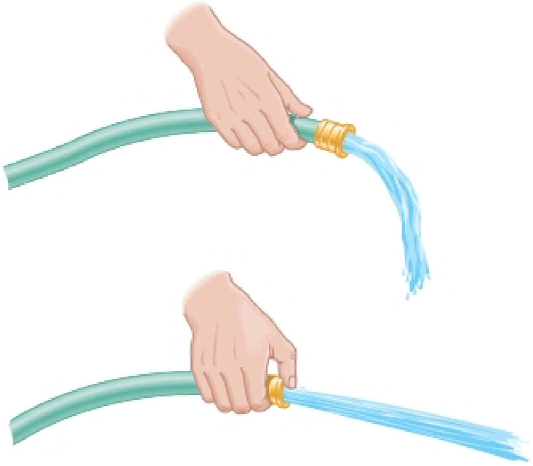


Figure 9.216: Application in everyday life of the continuity equation

Let us now consider a region in a fluid where there is a steady flow[31] as shown in the figure below:

---

[31] That is, the pressure, velocity, density, etc. can change from place to place, but they must not change as a function of time at any particular place (in your frame of reference).

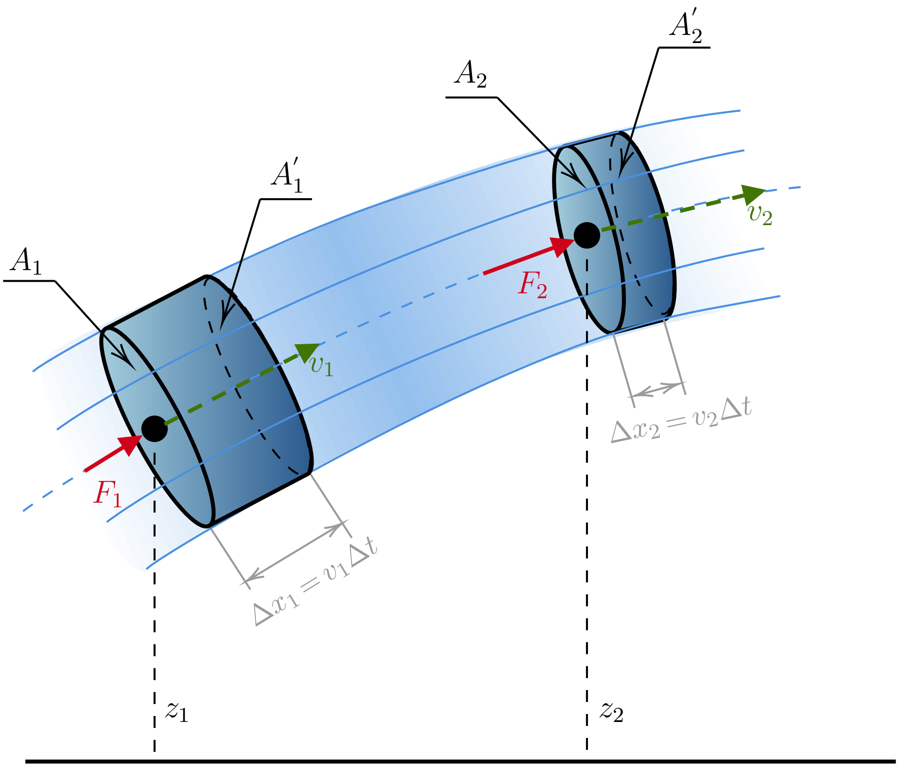


Figure 9.217: **Zoom of a fluid region for the Bernoulli's theorem**

During a short time interval Δt, the fluid which initially crossed the section A_1 has moved to a surface A'_1 at a distance $\Delta x_1 = v_1\Delta t$ while the fluid crossing the section A_2 is found on A'_2 at a distance $\Delta x_2 = v_2\Delta t$.

The both small cylinders have the same volume as the fluid is assumed to be incompressible and that the fluid passing through A_1 pushes the volume that pass through A_2 and therefore the equation continuity remains valid.

Given F_1 and F_2 the forces applied indirectly on the surface A_1 and A_2 due to the pressure prevailing in the fluid. Because of these forces, the fluid produce or receives work by moving the two volumes. In A_1, the surface is pushed by the fluid and the work performed on the fluid is $F_1\Delta x$ when on A_2 the fluid pushes the surface and the work done by the fluid is $F_2\Delta x$. The total work exerted on the fluid volume situated between A_1 and A_2 is therefore equal to:

$$W = F_1\Delta x_1 - F_2\Delta x_2 = P_1 A_1\Delta x_1 - P_2 A_2\Delta x_2 \tag{9.7.146}$$

where P_1 and P_2 are obviously the respective pressures on A_1 and A_2 and by writing obviously:

$$F = PA \tag{9.7.147}$$

according to the definition of the pressure. As:

$$A_1\Delta x_1 = A_2\Delta x_2 \tag{9.7.148}$$

according to the continuity equation and the assumption of incompressibility, we can write that:

$$W=(P_1-P_2)A_1\Delta x_1 \tag{9.7.149}$$

The external work applied on the system changes its internal energy as established in the section Thermodynamics ($W=\Delta U$). For the considered volume of fluid, the internal energy of the volumes highlighted in the figure above includes kinetic energy and gravitational energy. The fluid between A_1 and A_2 gains energy in the volume $A_2\Delta x_2$. Suppose the two volumes have equal mass m, again because of the continuity equation. Then the net energy gain is:

$$\Delta U=U_2-U_1=\left(\frac{1}{2}mv_2^2+mgz_2\right)-\left(\frac{1}{2}mv_1^2+mgz_1\right) \tag{9.7.150}$$

Since we have already assumed the fluid as incompressible, the density ρ is the same everywhere and m can be replaced by $\rho A_1\Delta x_1$ at both ends. Hence:

$$\Delta U=\left[\left(\frac{1}{2}\rho v_2^2+\rho g z_2\right)-\left(\frac{1}{2}\rho v_1^2+\rho g z_1\right)\right]A_1\Delta x_1 \tag{9.7.151}$$

Combining this relation with $W=(P_1-P_2)A_1\Delta x_1$ we get:

$$P_1-P_2=\left(\frac{1}{2}\rho v_2^2+\rho g z_2\right)-\left(\frac{1}{2}\rho v_1^2+\rho g z_1\right) \tag{9.7.152}$$

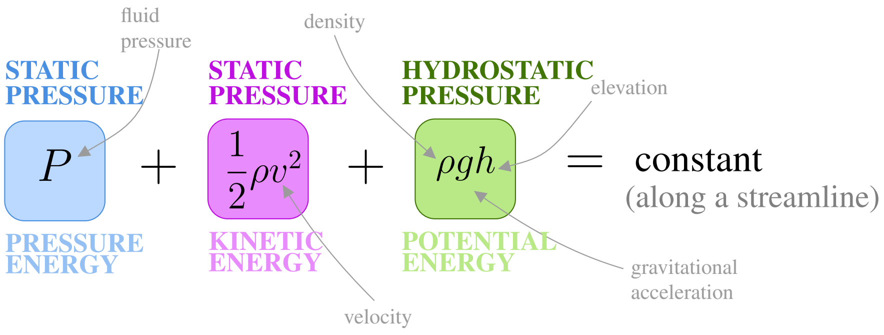
or:

$$\boxed{\frac{1}{2}\rho v_1^2+\rho g z_1+P_1=\frac{1}{2}\rho v_2^2+\rho g z_2+P_2} \tag{9.7.153}$$

As the above equation relates the magnitudes taken between two arbitrary points along a streamline, we can generalize and write:

$$\boxed{\frac{1}{2}\rho v^2+\rho g z+P=c^{te}} \tag{9.7.154}$$

This result, known as the "Bernoulli's theorem", expresses the constancy of the pressure along a streamline in an incompressible fluid, irrotational and non-viscous and where external volumetric forces are derived from a potential energy (we'll come back on this more in detail after having determined the Navier-Stokes equations). The above relation can also obviously be interpreted as saying the enthalpy of the parcel remains constant as it flows along a streamline.


fluid pressure
density
STATIC PRESSURE
STATIC PRESSURE
HYDROSTATIC PRESSURE
elevation
P
$+$
$\frac{1}{2}\rho v^2$
$+$
$\rho g h$
$=$ constant
(along a streamline)
PRESSURE ENERGY
KINETIC ENERGY
POTENTIAL ENERGY
gravitational acceleration
velocity

We will see further below during our study of gas the generalized version of the above expression involving the polytropic constant γ (heat capacity ratio).

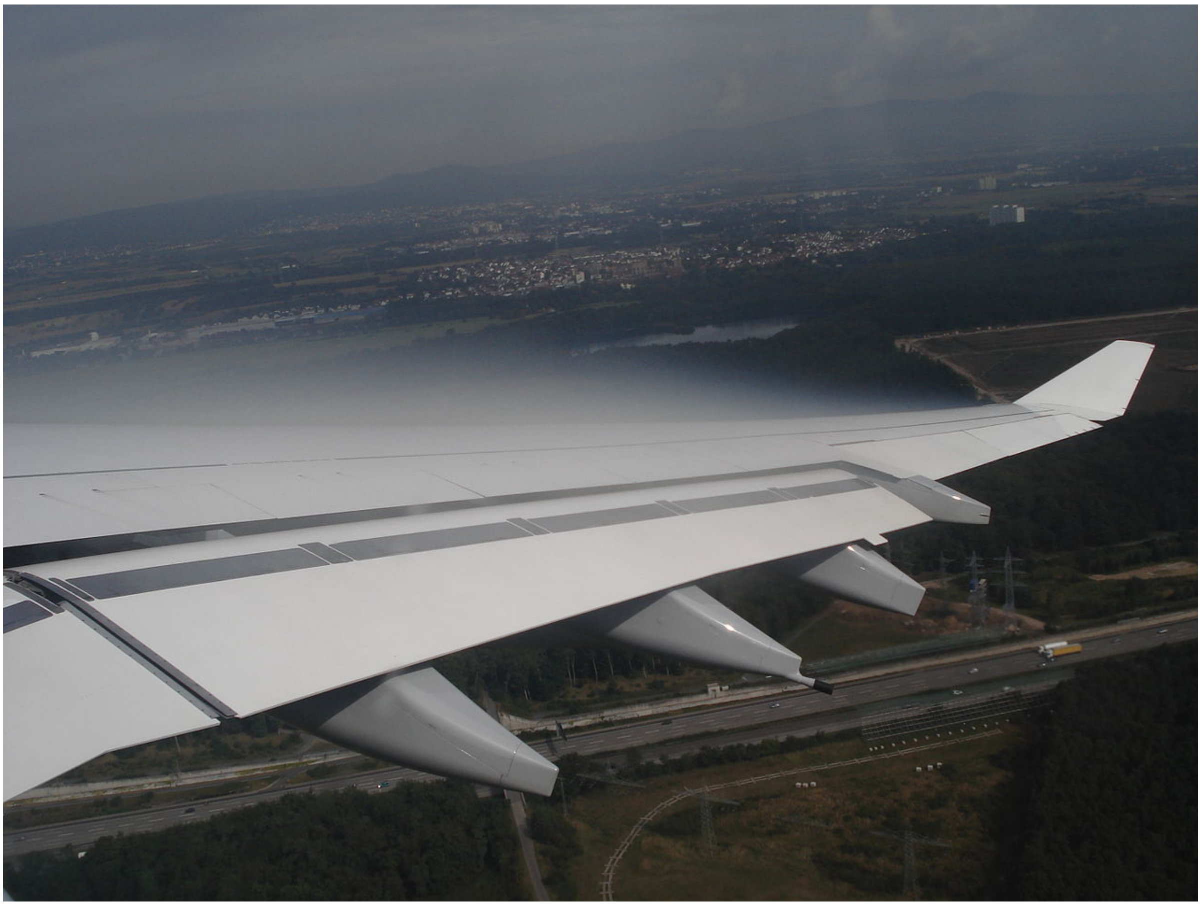


Figure 9.218: **Condensation visible over the upper surface of an Airbus A340 wing caused by the fall in temperature accompanying the fall in pressure, both due to acceleration of the air (source: Wikipedia)**

The conservation of the left quantity expresses the conservation of energy along a stream line and we find respectively the volumetric kinetic energy, the gravitational volumetric potential energy and the pressure.

Remark

Form a stream line to the other, that is the value of the constant c^{te} that change. In addition, use of Bernoulli's theorem requires to know (or at least to have an idea) of the shape of the stream line.

The reader must keep in mind that the following assumptions must be met for this Bernoulli equation to apply:

- the flow must be steady, i.e. the fluid velocity at a point cannot change with time
- the flow must be incompressible - even though pressure varies, the density must remain constant along a streamline
- friction by viscous forces has to be negligible (no thermic dissipation)

Let us notice also an another elegant and easy way to find this relation. Indeed, the conservation of energy gives us along a stream line:

$$E_c+E_p+E_P=c^{te} \tag{9.7.155}$$

with respectively the kinetic energy E_c, the potential energy E_p and the pressure energy E_P (that latter is not intuitive). Hence:

$$\frac{1}{2}mv^2+mgz+Fd=c^{te} \tag{9.7.156}$$

and if we divide all by the volume, Then We get:

$$\frac{1}{2}\rho v^2+\rho gz+P=c^{te} \tag{9.7.157}$$

that's it... Therefore Bernoulli equation is simply a statement of the principle of conservation of energy in fluids!

The term:

$$q=\frac{1}{2}\rho v^2 \tag{9.7.158}$$

is quite famous and named "dynamic pressure". It is especially famous to rocket science lovers that know the famous "Max q". The max q condition is the point when an aerospace vehicle's atmospheric flight reaches maximum dynamic pressure.

Now consider now two important applications of Bernoulli's theorem.

If the fluid moves in a horizontal plane, the gravitational potential energy remains constant and then the Bernoulli equation reduces then to (or also in absence of gravitational field):

$$\boxed{\frac{1}{2}\rho v^2+P=\frac{1}{2}\rho\left(\frac{\mathrm{d}x}{\mathrm{d}x}\right)^2+P=c^{te}} \tag{9.7.159}$$

Thus, in a closed horizontal pipe (wind tunnel), the velocity increase as the pressure decrease and vice versa. We also use this effect to participate in the thrust of an airplane at least in the case of wind tunnel tests in a closed cylinder (caution! this parameter is minor in the case of a displacement in the atmosphere which is an open volume , because it is not what contributes the most to the flight of an airplane, it is the Magnus effect whose proof is given in the Aerospace section).

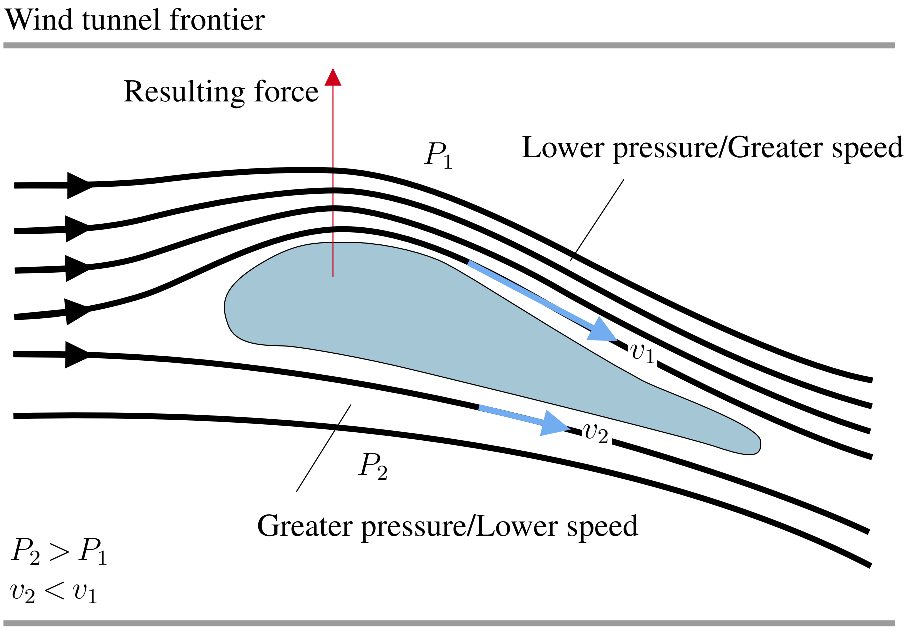


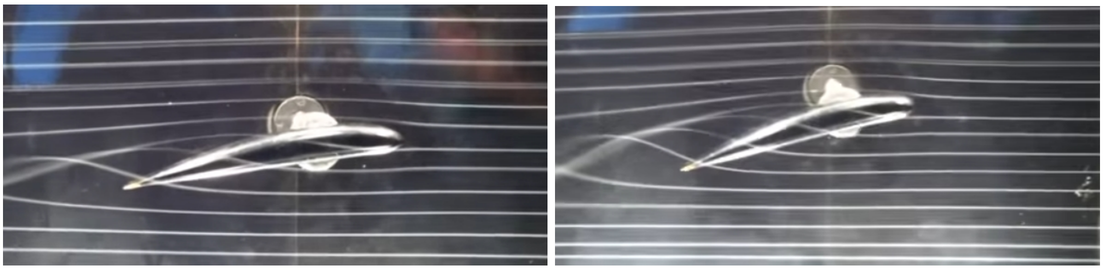
Figure 9.219: Illustration of the profile of a wing with the corresponding pressures and speeds in a wind tunnel

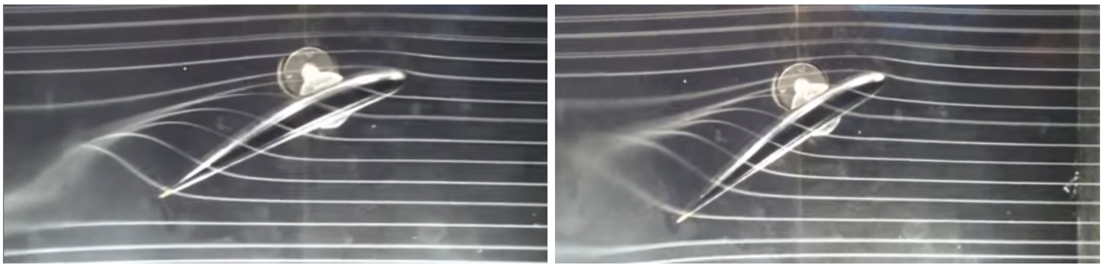
The profile of a wing is most of time constructed in such a way that the air has can in some given configuration of the wind greater velocity above the wing surface than under it, producing a higher pressure below that of the wing, Above in the case of an air flow in a closed cylinder (as was the case in the first wind tunnel tests in the early 1900s according to holocene calendar!). This results in a resultant upward force (but it's not the only force from which the lift origins!).

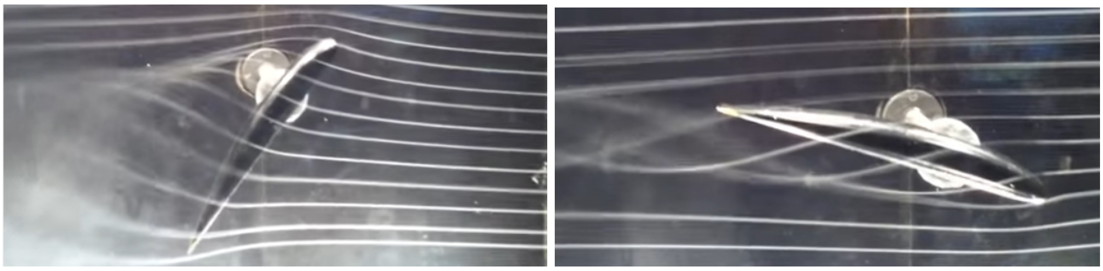
In other words, a specialist in aerodynamics (for aircraft) or in hydrodynamics (for roll stabilizers of large boats) would say in the case of closed volume tests:

- At the "extrados": By curvature effect, the air (water) particles will pass through a smaller transverse surface in the test cylinder and by Venturi effect will then accelerate. Their velocity will therefore first increase strongly and then diminish in order to find at the trailing edge the initial velocity of the flow. All the extrados is therefore the seat of a generalized local depression.
- In the "intrados": By curvature, the air (water) particles will pass through a larger transverse surface in the test cylinder and by Venturi effect will then decelerate. Their velocity will therefore decrease strongly and then increase in order to recover the initial velocity of the flow. All the intrados is therefore the center of a generalized local overpressure.

The lift of a wing with regard to its geometry is due in part to the depression on the extrados, not to the overpressure on the intrados. The wing does not mainly "rest" on the air, but is sucked upward little bit by the latter (for more details see the Aerospace section)!







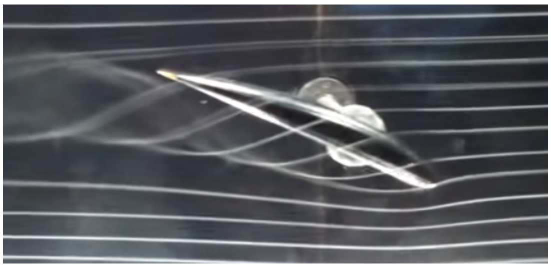


Figure 9.223: **Wing streamlines in wind tunnel in real life**

Another thing, if the fluid is not in motion, we have then Bernoulli equation which is obviously written:

$$\boxed{\rho g z + P = c^{te}} \quad (9.7.160)$$

and that is sometimes named the "Laplace hydrostatic equation" (used in communicating vessels study).

9.7.2.3.1 Torricelli's barometer

Take a glass tube about one meter long opened at one end having uniform diameter. Fill the whole tube with pure mercury. Close the open end of the tube tightly with finger and invert the tube. Take this inverted end of the tube along with your finger inside the pure mercury contained in a container. Now remove the finger and take necessary steps to keep the tube vertical. It will be observed that the mercury has come down to some extent and has attained a steady position. It might appear apparently that the mercury inside the tube is in standing position by itself. But in reality it is not so. It happens so due to air pressure.

The force of gravity on air is directed downward. The force of gravity on the mercury in the bowl and the tube is directed downward. Newton's third law: For every force, there is an equal and opposite force. The mercury in the tube drops down until it is balanced by the force exerted by air on the mercury in the bowl.

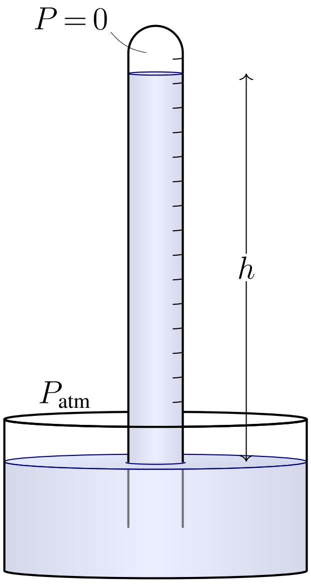


Figure 9.224: **Torricelli's experiment (author: Izaak Neutelings)**

The space that remains just above the mercury column in the tube up to its closed end is totally empty. This empty space (assumed as because we will see during study of Quantum Field Theory that there are not such things as "empty space") is termed as "Torricelli's vacuum" (because in reality it's not a vacuum as it contains mercury vapor).

Atmosphere is always exerting pressure on the mercury of the container. This pressure is convoyed through the mercury and acts in the upward direction in the tube. This very pressure holds the mercury column of the tube. In absence of this pressure the mercury of the tube will come down due to the force of gravity. Therefore, atmospheric pressure is equal to the pressure of the mercury column of the tube. Generally the height of the mercury column that remains in the tube in about 76 [cm] i.e the air pressure supports the pressure of 76 [cm] column of mercury. In this way using the height of a liquid the atmospheric pressure can be measured.

We can determine the 76 [cm] theoretically using the relation derived earlier as the vacuum on the top of the tube exerts an equal force to that of the pressure of the surrounding of the atmosphere. That means:

$$P_{\text{atm}}=P_{\text{Hg}} \qquad (9.7.161)$$

As derived earlier above we know that we have:

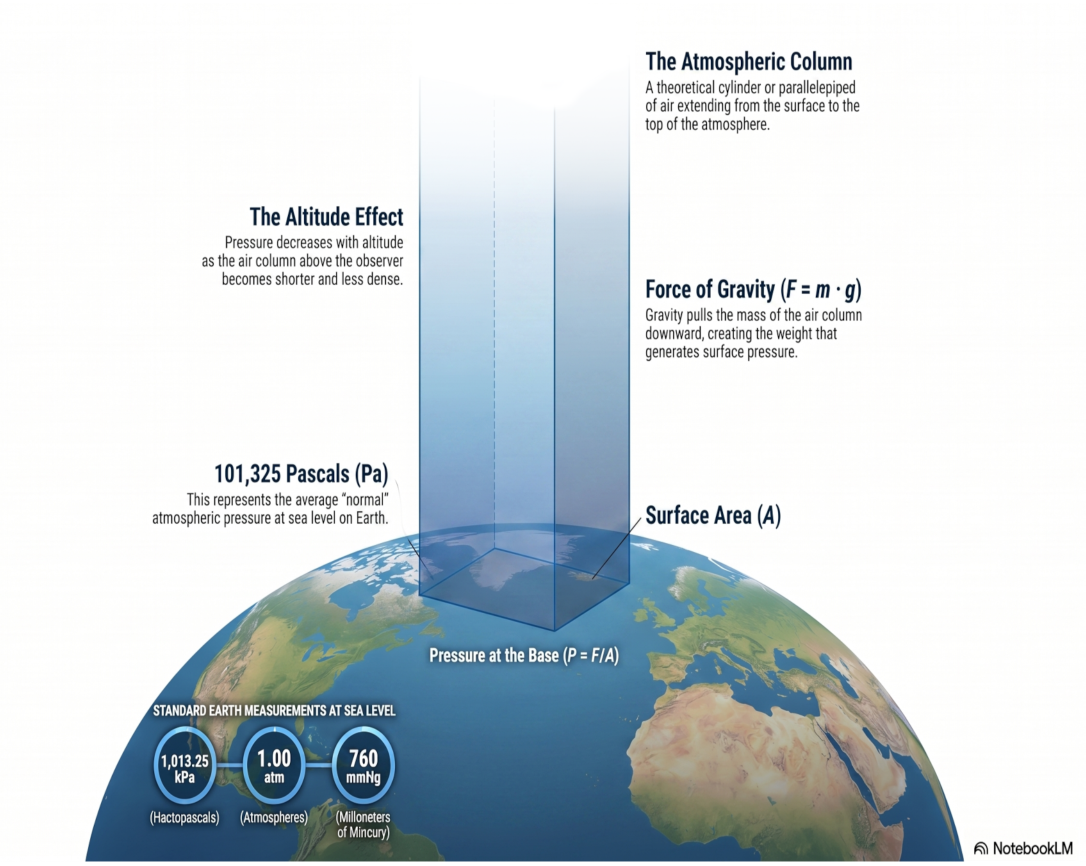
$$P_{\text{Hg}}=\rho_{\text{Hg}}gh \qquad (9.7.162)$$

Hence:

$$P_{\text{atm}}=P_{\text{Hg}}=\rho_{\text{Hg}}gh \qquad (9.7.163)$$

Therefore using the standard pressure and temperature at sea level and mercury density at standard pressure and temperature and average gravity value on Earth we get:

$$h=\frac{P_{\text{atm}}}{\rho_{\text{Hg}}g}\cong\frac{101,325}{13,545.848\cdot 9.80665}\cong 0.762\,[\text{m}]\cong 76\,[\text{cm}] \qquad (9.7.164)$$


The Atmospheric Column
A theoretical cylinder or parallelepiped of air extending from the surface to the top of the atmosphere.
The Altitude Effect
Pressure decreases with altitude as the air column above the observer becomes shorter and less dense.
Force of Gravity (F = m · g)
Gravity pulls the mass of the air column downward, creating the weight that generates surface pressure.
101,325 Pascals (Pa)
This represents the average "normal" atmospheric pressure at sea level on Earth.
Surface Area (A)
Pressure at the Base (P = F/A)
STANDARD EARTH MEASUREMENTS AT SEA LEVEL
1,013.25 kPa
(Hactopascals)
1.00 atm
(Atmospheres)
760 mmHg
(Milloneters of Mincury)
NotebookLM

9.7.2.3.2 Torricelli's law

The "Torricelli's law", also known as "Torricelli's theorem", is a theorem in fluid dynamics relating the speed of fluid flowing out of an opening to the height of fluid above the opening assuming no air resistance, viscosity, or other hindrance to the fluid flow. This is a classic case study in high schools.

For this study let us consider a closed volume (recipient) containing a liquid of equal density ρ and having small orifice of surface S_2 from which the liquid flows outwards. We want to determine the speed v_2 at which the liquid flow of this orifice:

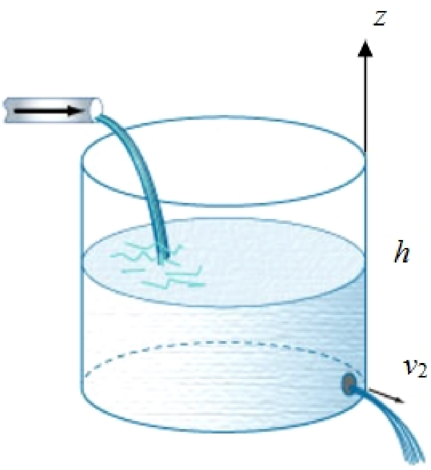


Figure 9.225: **Torricelli experiments configuration**

The volume is assumed to be big enough that neither the liquid level nor the pressure P above its surface S_1 will appreciably vary during the flow. Since the liquid exhausting virtual tube runs from the region of the surface of the liquid to the open orifice in the open air, we have:

$$P_1 = P \tag{9.7.165}$$

A liquid flowing in the open air is at atmospheric pressure, $P_2 = P_A$, because the liquid is surrounded by free air and nothing is present to maintain a pressure difference. According to Bernoulli's theorem, with:

$$z_1 - z_2 = h \tag{9.7.166}$$

we have therefore found a current line (streamline) and we can write:

$$P + \frac{1}{2}\rho v_1^2 + \rho g h = P_A + \frac{1}{2}\rho v_2^2 \tag{9.7.167}$$

hence:

$$v_2^2 = v_1^2 + \frac{2(P - P_A)}{\rho} + 2gh \tag{9.7.168}$$

From the continuity equation principle (flow conservation):

$$v_1 S_1 = v_2 S_2 \tag{9.7.169}$$

we deduce that if $S_1 \gg S_2$ then $v_2 \gg v_1$ and v_1^2 is then negligible relatively to v_2^2. In the particular but frequent case where the recipient is open to the open air ($P = P_A$), the pressure energy density disappears. The fluid flows only under the effect of gravity, without being pushed by a difference in pressure. We then find (by multiplying this result by the surface of the orifice, we get the flow rate):

$$\boxed{v_2 = \sqrt{2gh}} \tag{9.7.170}$$

which is independent of the sections! Obviously by using this last relation, in the case where the flow and the surface of the orifice are given, we can deduce the flow velocity in $\mathrm{m}^2 \cdot \mathrm{s}^{-1}$.

This relation constitutes the "Torricelli theorem". Curiously, we have already seen this relation in the section of Classical Mechanics for the rate of free fall of a body (see page 3435). Therefore it comes the observation made by Torricelli: if the jet is directed directly upwards, it almost reaches the level of the surface of the liquid in the volume. The reason why the jet would not actually reach this level is because some energy loss due to friction.

From this result we can study an interesting case in practice and often also used in numerical modelling courses. It is the exercise consisting in determining the time for a tank (recipient) of the following type to empty:

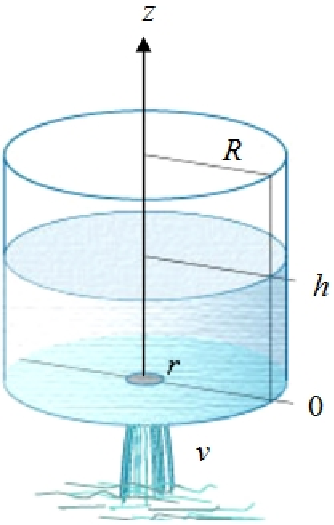


Figure 9.226: **Tank emptying**

We start from the theorem of Torricelli in the case where the orifice is much smaller than the diameter of the tank (we will see after that if not the speed is not the same!):

$$v=\sqrt{2gz} \tag{9.7.171}$$

And if we denoted the flow rate by D then we have trivially:

$$D=-\pi r^2\sqrt{2gz} \tag{9.7.172}$$

where the sign "$-$" is there to indicate that the tank is emptying. We also have the flow rate which is defined by:

$$D=\frac{\mathrm{d}V}{\mathrm{d}t} \tag{9.7.173}$$

and the variation of volume V in the tank be expressed as:

$$\mathrm{d}V=S(z)\mathrm{d}z \tag{9.7.174}$$

We then have explicitly:

$$\frac{S(z)\mathrm{d}z}{\mathrm{d}t}=D(z)=-\pi r^2\sqrt{2gz} \tag{9.7.175}$$

In our particular case, the surface does not depend on the z. We have then:

$$\frac{\mathrm{d}z}{\mathrm{d}t}=-\frac{\pi r^2}{S(z)}\sqrt{2gz}=-\frac{\pi r^2}{\pi R^2}\sqrt{2gz}=-\frac{r^2}{R^2}\sqrt{2gz} \quad (9.7.176)$$

Either after rearrangement:

$$\frac{\mathrm{d}z}{\sqrt{z}}=-\frac{r^2}{R^2}\sqrt{2g}\mathrm{d}t \quad (9.7.177)$$

Now we integrate on the left of H to 0 (H being the initial height of the fluid) and on the right from the time 0 to T. We then have:

$$\int_H^0 \frac{\mathrm{d}z}{\sqrt{z}}=-\int_0^T \frac{r^2}{R^2}\sqrt{2g}\mathrm{d}t \quad (9.7.178)$$

Either after rearrangement we deduce that the time T of emptying is:

$$\boxed{T=\frac{R^2}{r^2}\sqrt{\frac{2H}{g}}} \quad (9.7.179)$$

This formula can be used to calibrate a water clock.

And we notice that if the radius of the orifice of evacuation is equal to the radius of the cylinder we fall back on the expression of the free fall time of a point mass as proved in the section of Classical Mechanics:

$$T=\sqrt{\frac{2H}{g}} \quad (9.7.180)$$

The figure below show set of Torricelli jets, vertically aligned, leaving the reservoir horizontally:

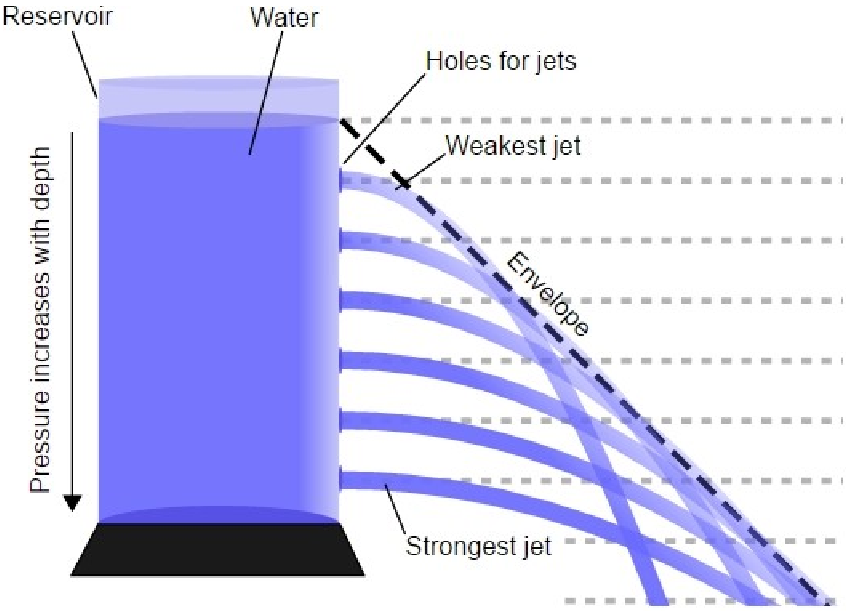


Figure 9.227: **Torricelli jets envelope illustration (source: Wikipedia, author: Matt Cook)**

In this case, the jets have an envelope (a concept also due to Torricelli) which is a line descending at $45°$ from the water's surface over the jets. Each jet reaches farther than any other jet at the point where it touches the envelope, which is at twice the depth of the jet's source. The depth at which two jets cross is the sum of their source depths. Every jet (even if not leaving horizontally) takes a parabolic path whose directrix is the surface of the water.

Remark

Some Abrahamic religions believe that there was a flood on the whole Earth that covered all mountains (including therefore Mt Everest at $8,849$ [m]) thanks to a rain falling during 40 days according to their myths and legends. How fast would the rain have to fall to make $8,849$ [m] in 40 days? That's easy! Forty days times 24 hours per day is 960 hours. Dividing $8,849$ by 960 and we find a rainfall of about 9 meters per hours! On a surface of $1\,[\text{m}^2]$ that represents a volume of $9,000\,[\text{L}\cdot\text{h}^{-1}]$, or $150\,[\text{L}\cdot\text{min}^{-1}]$, or $2.5\,[\text{L}\cdot\text{s}^{-1}]$. Practice shows that a debit rate between $0.5\,[\text{L}\cdot\text{min}^{-1}]$ and $60\,[\text{L}\cdot\text{min}^{-1}]$ per square meter is enough to cut the strongest wood. So if the Arch was made of wood it was cut to pieces by the rain...

We don't even need to say that according to the same abrahamic myths and legends that the roughly $4,500,000,000$ cubic kilometre of water - a cube of about $1,650$ [km] side - receded in 150 days naturally is simply impossible.

9.7.2.3.3 Siphon

A "siphon" is any of a wide variety of devices that involve the flow of liquids through tubes. In a narrower sense, the word refers particularly to a tube in an inverted "U" shape, which causes a liquid to flow upward, above the surface of a reservoir, with no pump, but powered by the fall of the liquid as it flows down the tube under the pull of gravity[32], then discharging at a level lower than the surface of the reservoir from which it came.

There are two leading theories about how siphons cause liquid to flow uphill, against gravity, without being pumped, and powered only by gravity. The traditional theory for centuries was that gravity pulling the liquid down on the exit side of the siphon resulted in reduced pressure at the top of the siphon. Then atmospheric pressure was able to push the liquid from the upper reservoir, up into the reduced pressure at the top of the siphon, like in a barometer or drinking straw, and then over. However, it has been demonstrated that siphons can operate in a vacuum and to heights exceeding the barometric height of the liquid. Consequently, the cohesion tension theory of siphon operation has been advocated, where the liquid is pulled over the siphon in a way similar to the chain fountain. It need not be one theory or the other that is correct, but rather both theories may be correct in different circumstances of ambient pressure. The atmospheric pressure with gravity theory cannot explain siphons in vacuum, where there is no significant atmospheric pressure. But the cohesion tension with gravity theory cannot explain CO_2 gas siphons, siphons working despite bubbles, and the flying droplet siphon, where gases do not exert significant pulling forces, and liquids not in contact cannot exert a cohesive tension force.

All known published theories in modern times recognize Bernoulli's equation as a decent approximation to idealized, friction-free siphon operation. And this is what we will study here considering the following configuration:

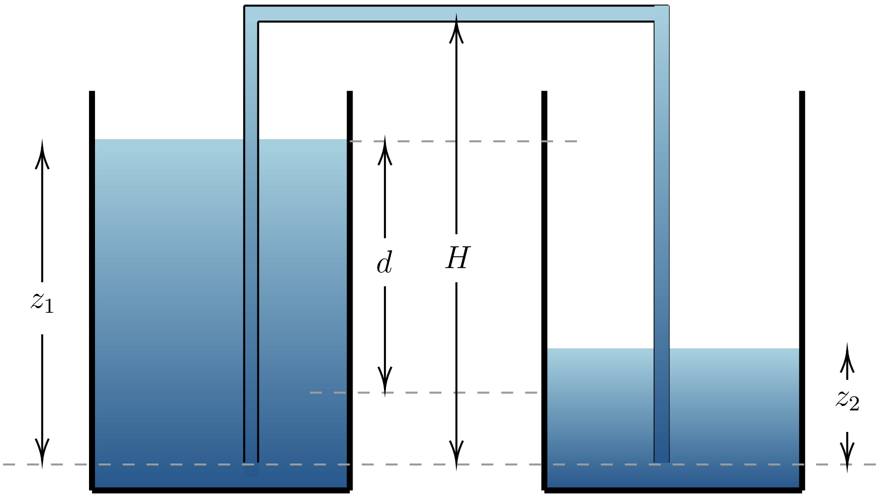


Figure 9.228: **Special case of siphon**

Let us consider that the figure above is a snapshot of an observation and the we see the fluids flows from left to right.

The first thing we want to do is to find the velocity v of the fluid in the tube.

---

[32] Keep in mind that's in the framework of the Newtonian model of gravity as we that in General Relativity there are no "pull".

Begin by applying Bernoulli's theorem to the top of the fluid in the left container and to the bottom right hand side of the tube. The pressure at the bottom left is just $P_0 + \rho g z_2$, so at the point where fluid is entering the second tank:

$$P_0 + \rho g z_1 + \frac{1}{2}\rho v_1^2 = P_0 + \rho g z_2 + \frac{1}{2}\rho v^2 \tag{9.7.181}$$

and (solving for v, cancel P_0, assume $v_1 \ll v$, reorganize and):

$$\frac{1}{2}\rho v^2 = \rho g \left(z_1 - z_2\right) \rightarrow v \cong \sqrt{2gd} \tag{9.7.182}$$

which solves the question.

Now let us found the pressure at the top of the tube.

To find the pressure P_H at the top of the tube (at height H) we have to apply Bernoulli's theorem to the top of the left container (again) but also to the fluid in the tube at height H :

$$P_0 + \rho g z_1 + \frac{1}{2}\rho v_1^2 = P_H + \rho g H + \frac{1}{2}\rho v^2 \tag{9.7.183}$$

Rearranging and using the fact that $\frac{1}{2}\rho v^2 = \rho g \left(z_1 - z_2\right)$ and still assuming $v_1 \ll v$:

$$\begin{aligned} P_H &\cong P_0 - \rho g H + \rho g z_1 - \frac{1}{2}\rho v^2 \\ &\cong P_0 - \rho g H + \rho g z_1 - \rho g \left(z_1 - z_2\right) \\ &\cong P_0 - \rho g H + \rho g z_2 = P_0 - \rho g \left(H - z_2\right) \end{aligned} \tag{9.7.184}$$

One interesting (and correct!) consequence of this is that there is a limit to the height over which one can siphon a fluid. The pressure in the tube cannot be negative, as (before) it reaches 0 (a vacuum) the fluid will chemically "come apart" and fall down the two columns as air pressure no longer suffices to hold it up in the tube and the picture will become that of two barometers with a shared vacuum space. We see that:

$$P_{\text{min}} \approx 0 = P_0 - \rho g \left(H_{\text{max}} - z_2\right) \tag{9.7.185}$$

or when the difference of H and z_2 (the height of fluid in the lower of the two vessels) is equal to $P_0/\rho g$ (about 10 meters for water) the siphon will break.

9.7.2.3.4 Communicating vessels

Let's talk a little bit about the communicating vessels because like this it seems to be nothing quite interesting but... still in this beginning of the 121st century (holocene calendar), most buildings have a supply of running water based on this principle (reserve lakes in height, Water castles and so on...) and many people have used the principle of communicating vessels to siphon a fluid from one tank / container to another at least once in their lives.

We can first introduce the concept intuitively based on the experimental observation of the 4 communicating vessels of the following figure with a non-viscous fluid:

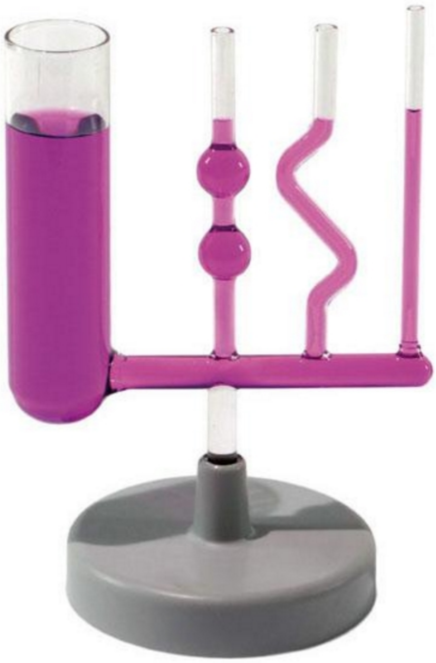


Figure 9.229: Photo of the principle of communicating vessels

we see that the fluid stabilizes at the same height for the four vessels independently of heir shape!

The equilibrium of vessel is explained be the development that follows!

Let us recall that we have by definition:

$$P=\frac{F}{S} \tag{9.7.186}$$

and therefore:

$$\mathrm{d}\vec{F}=-P\mathrm{d}\vec{S} \tag{9.7.187}$$

Using the gradient theorem proved in the section of Vector Calculus, we have:

$$\vec{F}=\oiint_S -p\mathrm{d}\vec{S}=\iiint_V \vec{\nabla}(-p)\mathrm{d}V=\iiint_V -\vec{\nabla}p\mathrm{d}V \tag{9.7.188}$$

In a volume of fluid in an homogeneous gravitational field we have:

$$\vec{F}=\iiint_V \vec{g}\mathrm{d}m=\iiint_V \rho\vec{g}\mathrm{d}V=\vec{g}\iiint_V \rho\mathrm{d}V=\vec{g}m \tag{9.7.189}$$

Therefore putting the two previous triple integrals:

$$\vec{F}=\iiint_V -\vec{\nabla}p\mathrm{d}V \qquad \text{and} \qquad \vec{F}=\iiint_V \rho\vec{g}\mathrm{d}V \tag{9.7.190}$$

together we identify the "hydrostatic balance":

$$-\vec{\nabla}p=\rho\vec{g} \tag{9.7.191}$$

In one dimension that can be rewritten:

$$\mathrm{d}p=-\rho g\mathrm{d}z \tag{9.7.192}$$

That is:

$$\boxed{P=P_0+\rho g(z-z_0)=P_0+\rho gh} \tag{9.7.193}$$

That is the "fundamental theorem of hydrostatics" or sometimes also named "Stevin's law" (or "Pascal-Stevin law"): the pressure at a point in a liquid in statics equilibrium depends only on the depth at that point! This is a quite counter-intuitive result form some people and that's why it is also named the "hydrostatic paradox" in some countries.

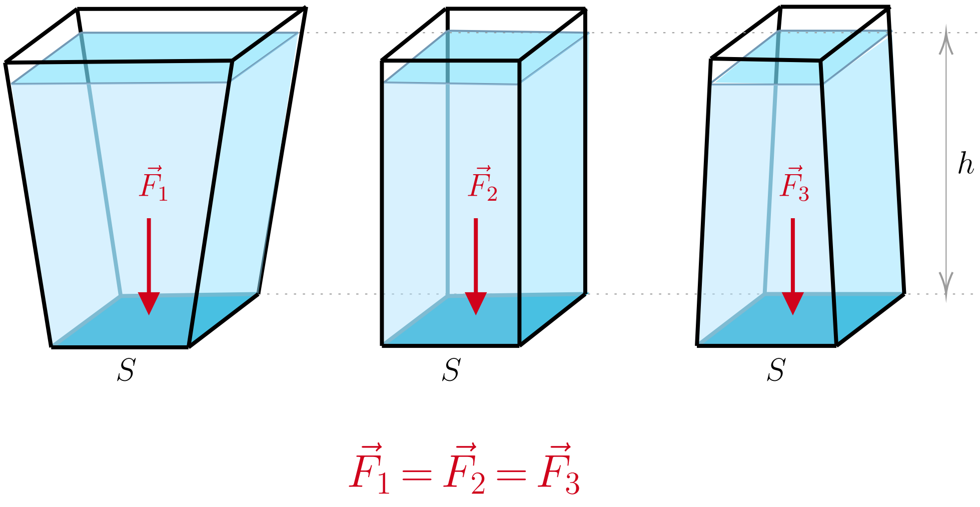


Figure 9.230: **Hydrostatic paradox illustrated**

Obviously the relation above on Earth is often denoted:

$$P=P_{\text{atm}}+\rho gh \tag{9.7.194}$$

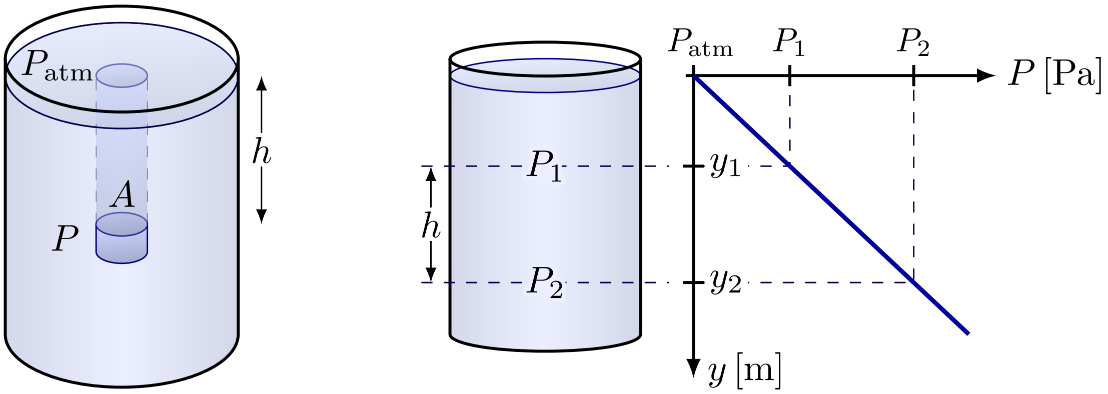
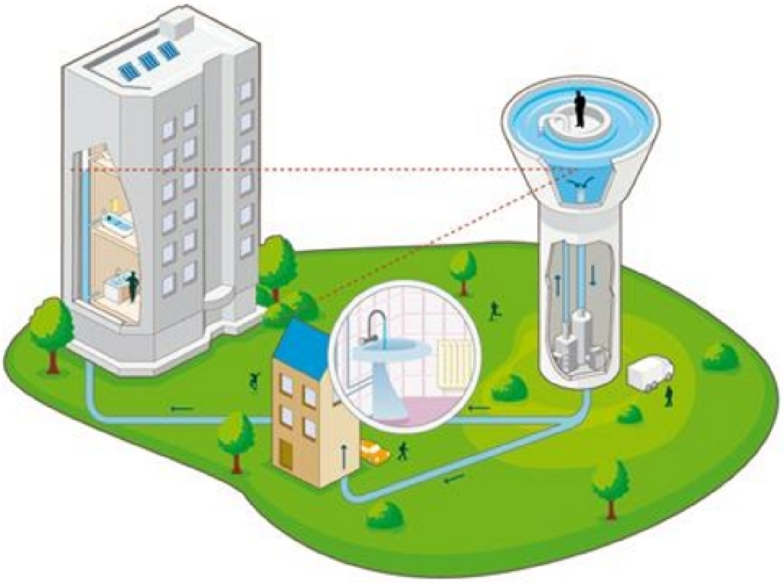


Figure 9.231: **Fundamental theorem of hydrostatics illustrated (author: Izaak Neutelings)**

Remark

We will prove far far far further below that we fall back on this relation from the Navier-Stokes equation!

The same theorems applies for a water tower (essential if it is to reduce the costs of supply and if the location does not allow to play with the topography of the places):



© AGENCE RESEAUX - Illustration :Kler Roger - 2009

Figure 9.232: **Application of communicating vessels with water towers**

So if we want for example on the 4th floor of a building to have a minimum flow at a given tap, we will have to apply Torricelli's theorem to know how many meters h above the tap level we must found the water level in a water tower.

We had obtained earlier above:

$$v=\sqrt{2gh} \tag{9.7.195}$$

Therefore:

$$h=\frac{v^2}{2g} \tag{9.7.196}$$

Either by using the flow definition:

$$h=\frac{\left(\dfrac{D}{S}\right)^2}{2g} \tag{9.7.197}$$

Where D is the flow rate to be supplied in cubic meters per second and S is the cross-section through which we must ensure the same flow! Thus, a flow rate of 0.05 litres per second (typical of a kitchen or bathroom tap), that is to say $5\cdot 10^{-5}\;[\mathrm{m}^3\cdot\mathrm{s}^{-1}]$, through a section having a radius of 0.5 centimetres will give:

$$h=\frac{\left(\dfrac{0.05\cdot 10^{-3}}{\pi 0.005^2}\right)^2}{2\cdot 9.81}\cong 2\;[\mathrm{m}] \tag{9.7.198}$$

This value that gives the minimum height difference between the water tower water level and the 4th floor where is our tap of interest obviously does not take into account the pressure drop in the pipe!

9.7.2.3.5 Venturi effect

Some practical applications of fluid mechanics result from the interdependence of pressure and velocity and there is a category of situations in which the potential gravitational energy variation is negligible. The Bernoulli equation then connects the pressure difference to the kinetic energy difference and therefore the variation of the square of the velocity.

To introduce the Venturi effect let us consider an incompressible (!) fluid, non-viscous and of density ρ. The fluid flows in steady state in a cylindrical pipe of radius R_1 and cross section S_1 followed by a cylindrical tube of radius R_2 and cross section S_2 and which then takes back the initial geometry of radius R_1 and of section S_1 (it is typically connection which it is customary to name "Venturi tube"). The connection is made by a long cylindrical pipe so that we stay in laminar regime:

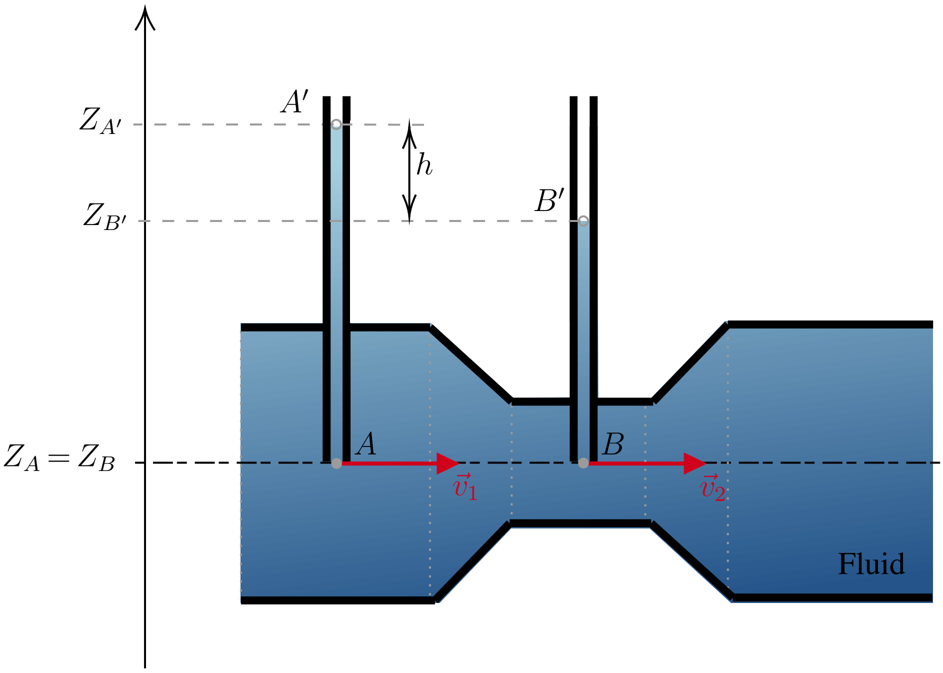


Figure 9.233: Venturi tube configuration study

We know (continuity equation) that:

$$S_1 v_1 = S_2 v_2 \tag{9.7.199}$$

Which means, as we have seen it already, that a diminution of the section through which the fluid passes, results in an increase of its velocity.

In any situation where the incoming flux is about the same level as the decreasing radius we have $Z_a \cong Z_B$ and then the Bernoulli's equation:

$$P_1 + \frac{1}{2}\rho v_1^2 + \rho g Z_A = P_2 + \frac{1}{2}\rho v_2^2 + \rho g Z_B \tag{9.7.200}$$

becomes:

$$P_1 + \frac{1}{2}\rho v_1^2 = P_2 + \frac{1}{2}\rho v_2^2 \tag{9.7.201}$$

Using the equation of continuity, to eliminate (arbitrarily) v_1, we get after rearranging:

$$\boxed{P_1 - P_2 = \frac{1}{2}\rho v_2^2 \left(1 - \frac{S_2^2}{S_1^2}\right)} \tag{9.7.202}$$

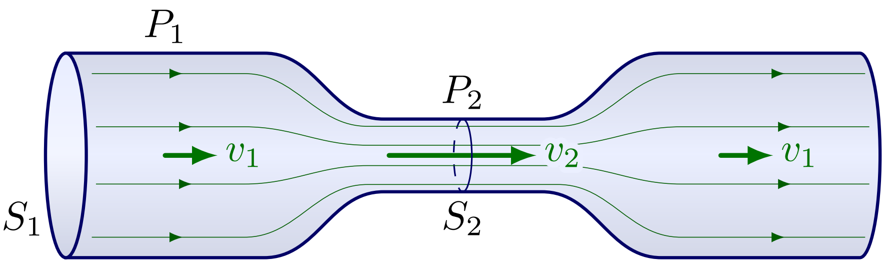


Figure 9.234: Venturi effect illustrated (author: Izaak Neutelings)

We can also write using the fundamental equation of hydrostatic proven just earlier above:

$$P_1 - P_2 = \rho g h = \frac{1}{2}\rho v_2^2 \left(1 - \frac{S_2^2}{S_1^2}\right) \tag{9.7.203}$$

As $S_1 > S_2$ the second member of the relation is positive and therefore $P_1 > P_2$: there is therefore a pressure drop in the tight part of the tube. On arriving at the divergent region again in $S_3 = S_2$, the pressure of the fluid increases again and the speed takes back its initial value. This decrease in pressure which accompanies the increase in speed is named the "Bernoulli effect" or "Venturi effect".

Thus, the velocity of the fluid increases in a bottleneck to satisfy the equation of continuity (conservation of the flux / mass) and the fact that it is incompressible (otherwise there would be a sort of plug).

Remark

Paradoxically, the Venturi effect also occurs when atmospheric air cross a summit or crest or also in the streets of cities. Indeed, the air that arrives on the mountain or the crest tends to "crash" on it. The cross-section of air flow at the top is therefore lower than at the base. There is therefore also a Venturi effect: the wind speed is higher on the summits and the ridges than at the bottom (the professionals of the glider know something ...). The Venturi effect is also a dangerous phenomenon to take into account when building maintenance works occurs in cities near tunnels. Indeed, as the speed of wind can increase dramatically near the exit of the tunnel it can damage or simply take down all maintenance installation (in such a configuration an accident happened in Fribourg in Switzerland a few years ago).

The Venturi tube is used in aerospace to measure the relative speed in flight, but also in automotive carburettors and in airbrushes. The Venturi tube is also used in industrial sites to measure the flow of a fluid in a pipe.

From the last relation, we derive trivially a very useful equality in practice:

$$v_2 = \sqrt{\frac{2\rho g h}{1 - \dfrac{S_2^2}{S_1^2}}} = \sqrt{\frac{\Delta P}{1 - \dfrac{S_2^2}{S_1^2}}} \tag{9.7.204}$$

that it is sufficient to multiply by the surface S_2 to have the flow rate in the connection.

Obviously we have identically:

$$v_1 = \sqrt{\frac{2\rho g h}{\dfrac{S_1^2}{S_2^2} - 1}} = \sqrt{\frac{\Delta P}{\dfrac{S_1^2}{S_2^2} - 1}} \tag{9.7.205}$$

Here is a typical Venturi tube to measures the D_1 flow rate thanks to pressure differential:

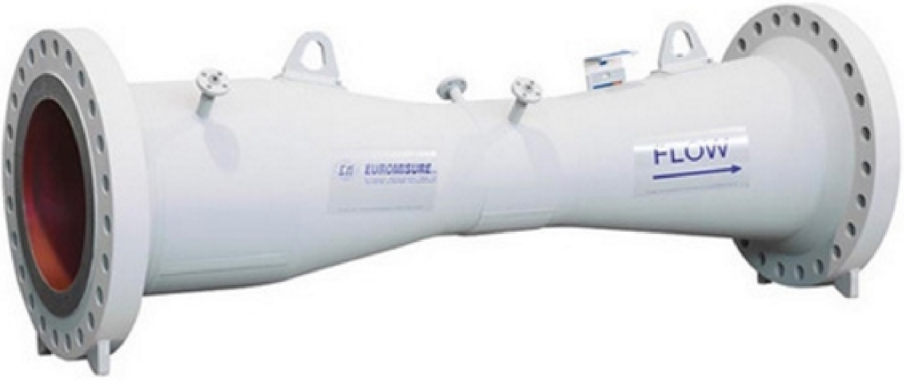


Figure 9.235: **Industrial Venturi tube (source: WIKA)**

For (old) airplanes as already mentioned Venturi tube can be used to measure relative velocity of the plane:

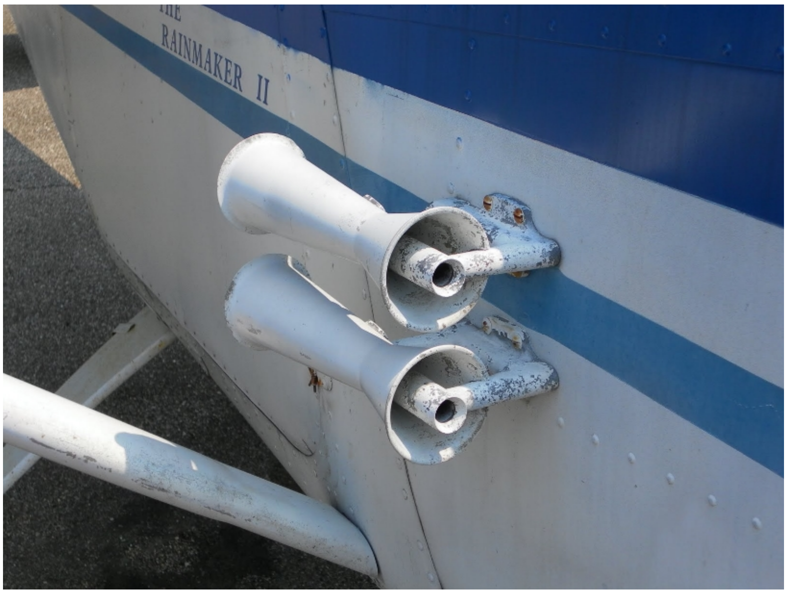


Figure 9.236: **Venturi tube on a CESSNA**

9.7.2.3.6 Pitot Tube

The Pitot tube is used to measure the flow velocity of a perfect subsonic fluid / gas (assumed to be in continuous flow). The Pitot tube most commonly used in aeronautics consists in forming in a tube a pressure tapping orifice at A and at B as visible in the figure below:

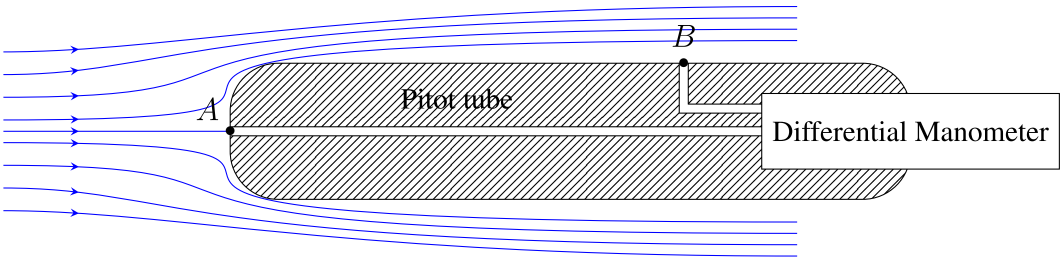


Figure 9.237: **Horizontal Pitot tube (source: http://femto-physique.fr, author: Jimmy Roussel)**

Point A is a stop point because the velocity is zero (there is no flow in the orifice, it is just a pressure tap). Far from the obstacle (the Pitot tube) the flow is assumed to in uniform speed vand pressure P_B

At A (breakpoint), using the Bernoulli equation along the current line and considering the variation in height between A and B negligible, the pressure is equal to:

$$P_A = P_B + \frac{1}{2}\rho v^2 \tag{9.7.206}$$

Therefore have:

$$\boxed{v = \sqrt{2(P_A - P_B)}\rho} \tag{9.7.207}$$

So for aircraft from the difference of a measure of pressure and knowledge of gas density, it is possible to know the speed! By multiplying by the section of the conduit we also have the corresponding flow rate.

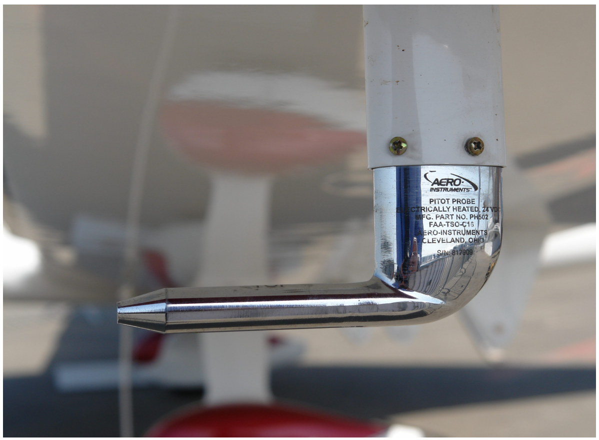


Figure 9.238: **Real Pitot tube**

Likely due to the aircraft's Pitot tubes being obstructed by ice crystals – caused the autopilot to disconnect, after which the crew reacted incorrectly and ultimately caused the aircraft to enter an aerodynamic stall from which it did not recover - the Air France Flight 447 (AF447/AFR447) from Rio de Janeiro to Paris was the deadliest in the history of Air France on 1 June 12009 (holocene calendar) with 228 deaths.

Remark

In aeronautics, the dynamic pressure is added to the static pressure (of ambient atmosphere) to give the total pressure which can be measured at the point of zero velocity of the Pitot tube. By removing the static pressure, one finds the "dynamic pressure".

There is another variant of the Pitot tube more used in the industrial field of which here is a diagram:

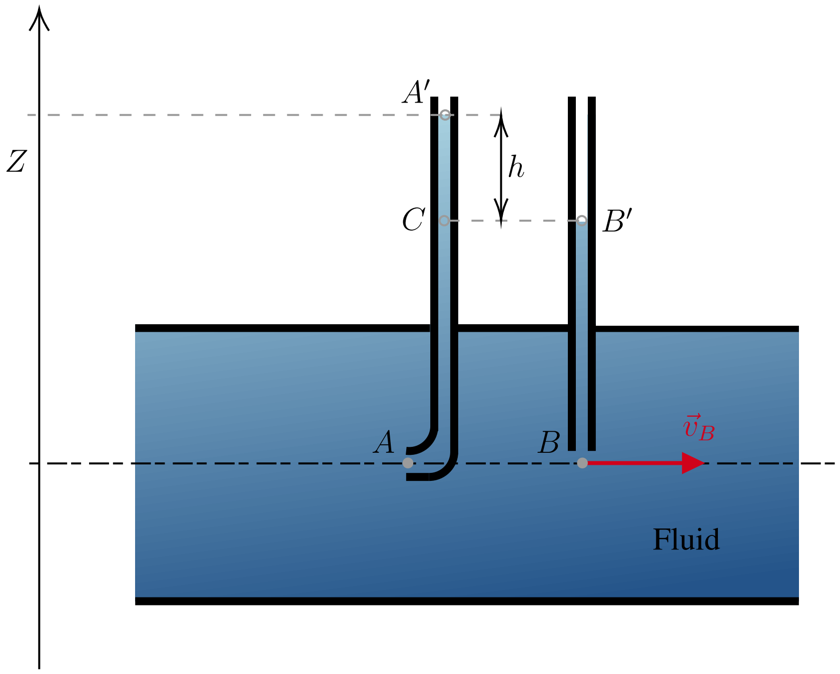


Figure 9.239: **Vertical Pitot tube**

We therefore also apply the Bernoulli theorem on the axis of the pipe. We have then in it (we detail one the developments a little more than in the previous example):

$$P_A+\rho g Z_A+\frac{1}{2}\rho v_A^2=P_B+\rho g Z_B+\frac{1}{2}\rho v_B^2 \tag{9.7.208}$$

but:

$$Z_A=Z_B \qquad v_A=0 \qquad v_B=v \tag{9.7.209}$$

and therefore:

$$P_A=P_B+\frac{1}{2}\rho v_B^2 \tag{9.7.210}$$

But we also have using the fundamental theorem of hydrostatics:

$$\begin{aligned}P_A&=P_{A'}+\rho g(Z_{A'}-Z_A)\cong P_{\text{atm}}+\rho g(Z_{A'}-Z_A)\\P_B&=P_{B'}+\rho g(Z_{B'}-Z_B)\cong P_{\text{atm}}+\rho g(Z_{B'}-Z_B)=P_{\text{atm}}+\rho g(Z_{B'}-Z_A)\end{aligned} \tag{9.7.211}$$

We have then:

$$P_A=P_B+\frac{1}{2}\rho v_B^2\Rightarrow P_{\text{atm}}+\rho g(Z_{A'}-Z_A)=(P_{\text{atm}}+\rho g(Z_{B'}-Z_A))+\frac{1}{2}\rho v_B^2 \tag{9.7.212}$$

Therefore:

$$\boxed{v=\sqrt{2g(Z_{A'}-Z_{B'})}=\sqrt{2gh}} \tag{9.7.213}$$

9.7.2.3.7 Pressure drop (pressure loss)

Pressure drop is defined as the difference in pressure between two points of a fluid carrying network. Pressure drop occurs when frictional forces, caused by the resistance to flow, act on a fluid as it flows through the tube. The main determinants of resistance to fluid flow are fluid velocity through the pipe and fluid viscosity.

When there is no machine (no pump or turbine) in the flow of a perfect fluid between points (1) and (2) of the same current line, the Bernoulli relation can be written In the following form:

$$\frac{1}{2}\rho(v_2^2-v_1^2)+\rho g(z_2-z_1)+(P_2-P_1)=0 \tag{9.7.214}$$

When the fluid passes through a hydraulic machine, it exchanges energy with this machine in form of work for a given time. The power P exchanged is then (see section Classical Mechanics page 3379):

$$P=\frac{\Delta W}{\Delta t} \tag{9.7.215}$$

Where by convention, if $P>0$ the energy is received by the fluid (pump) otherwise, if $P<0$ the energy is supplied by the fluid (turbine).

If the volumetric flow is Q_V then Bernoulli's relation is written logically:

$$\frac{1}{2}\rho(v_2^2-v_1^2)+\rho g(z_2-z_1)+(P_2-P_1)=w \tag{9.7.216}$$

where:

$$w=\frac{P}{Q_V}\;[\mathrm{J}\cdot\mathrm{m}^{-3}] \tag{9.7.217}$$

A perfect fluid does not exist. During a flow in a pipe, the friction forces dissipate part of the kinetic and potential energy, which results in the existence of pressure losses which must be taken into account.

Let us consider a stationary and incompressible horizontal cylindrical flow. If we apply the Bernoulli relation between input and output, we obtain:

$$P_1=P_2 \tag{9.7.218}$$

However, experimentally, we observe that we must impose a higher input pressure to maintain the steady state. Indeed, the viscosity forces resist to the flow. It is therefore necessary to impose an overpressure ΔP_f which we name the "pressure loss" and that is due to the existence of frictional forces (viscosity) or singular losses (geometry of the distribution circuits).

The "generalized Bernoulli equation (for liquids)" is then written in this case of study which is part of the field of process engineering:

$$\boxed{\frac{1}{2}\rho v_1^2+\rho g z_1+P_1+\frac{P}{Q_V}=\frac{1}{2}\rho v_2^2+\rho g z_2+P_2+\Delta P_f} \qquad (9.7.219)$$

This relationship is often used in the theoretical (...) study of pipeline problems.

In practice, the pressure drop is calculated from the Poiseuille law which we have already demonstrated earlier above:

$$J=\frac{\pi R^4}{8\mu_d}\frac{\Delta P}{l} \qquad (9.7.220)$$

where for recall J is the traditional notation of the volumetric flow in this particular field.

What is interesting to observe with the Poiseulle law in the case of the analysis of the pressure losses is that the volumetric flow is therefore inversely proportional to the distance l of the horizontal pipe. Thus, the loss of pressure and therefore the pressure at the end of a pipe is very easily calculated by rearranging the above relationship:

$$\boxed{P_1=P_2-\frac{J8\mu_d l}{\pi R^4}} \qquad (9.7.221)$$

So obviously the pressure at the end of a pipe is less than the pressure at its beginning.

9.7.2.4 Navier-Stokes Equations

Let us consider an elementary parallelepiped extracted from a static fluid at equilibrium of dimensions $\mathrm{d}x$, $\mathrm{d}y$, $\mathrm{d}z$ represented in the figure below. The equilibrium material composing the parallelepiped is generally subjected to volumetric forces in all directions (Pascal's theorem) whose components on the three orthogonal axes are represented in the figure below (these forces may be of Gravitational, Electromagnetic or Inertial origin...).

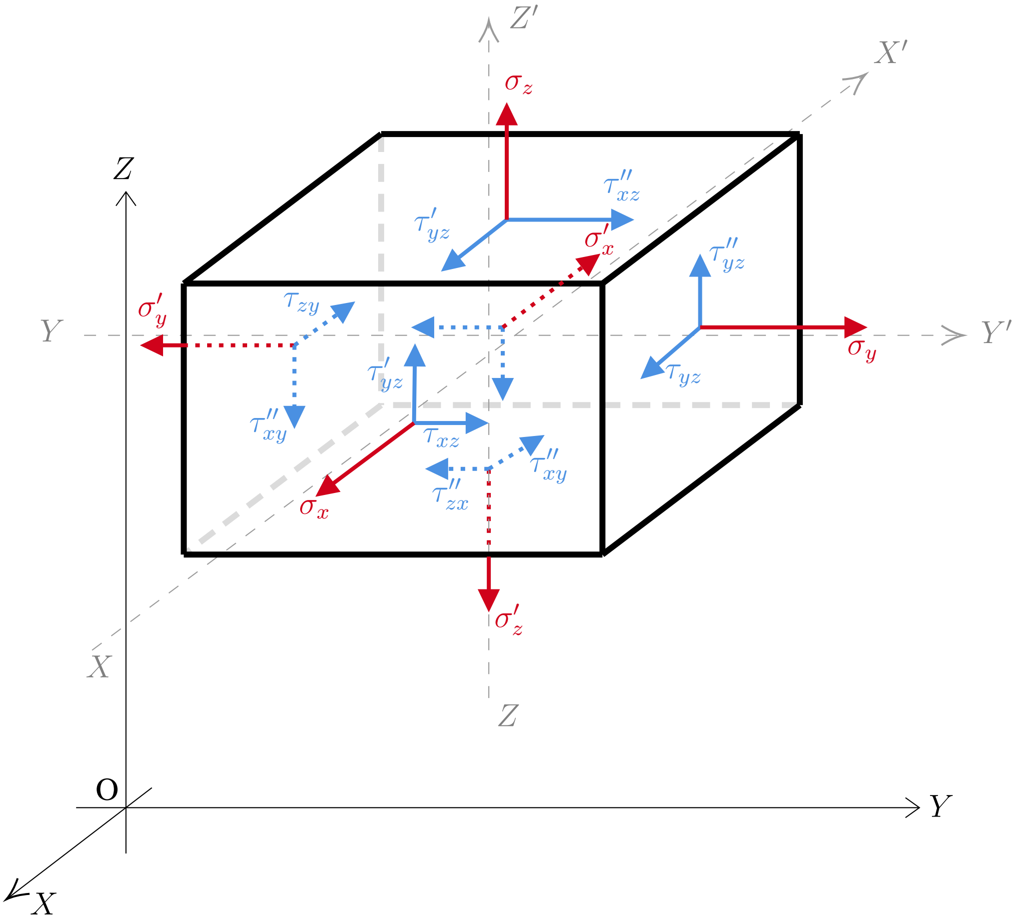


Figure 9.240: Elementary parallelepiped extracted from a static fluid at equilibrium

Remarks

R1. It is important to notice that the components of all the vectors visible in the figure above are expressed in newtons per unit area or in other words per unit pressure (which is the unit for recall... of a constraint...).

R2. It is important to be very attentive to what follows because some of the results we will get here will be reused in the section of General Relativity (see page 5720) to understand the energy-momentum tensor!!

We can, as we have represented it above, decompose and translate all the forces to which the parallelepiped is subjected to the centers of the faces of the latter. We obviously represent each of the constraints on each of the faces as the sum of the normal and tangential stresses as we did for the study of constrained solids (according to the three axes always, hence the sum of three components!).

In total, we find ourselves with 18 components of normal and tangential stresses:

$$\begin{aligned}\vec{\sigma}_x+\vec{\tau}'_{yx}+\vec{\tau}_{xz}&=\frac{\vec{F}_{+x}}{\mathrm{d}S} & \vec{\sigma}'_x+\vec{\tau}'_{xy}+\vec{\tau}_{zx}&=\frac{\vec{F}_{-x}}{\mathrm{d}S}\\ \vec{\sigma}_y+\vec{\tau}'_{yz}+\vec{\tau}_{yx}&=\frac{\vec{F}_{+y}}{\mathrm{d}S} & \vec{\sigma}'_y+\vec{\tau}'_{zy}+\vec{\tau}''_{xy}&=\frac{\vec{F}_{-y}}{\mathrm{d}S}\\ \vec{\sigma}_z+\vec{\tau}''_{xz}+\vec{\tau}'_{yz}&=\frac{\vec{F}_{+z}}{\mathrm{d}S} & \vec{\sigma}'_z+\vec{\tau}''_{zx}+\vec{\tau}'_{zy}&=\frac{\vec{F}_{-z}}{\mathrm{d}S}\end{aligned}\tag{9.7.222}$$

We try to minimize the number of normal components in order to determine the sufficient constraints on each of the axes. Therefore we will put:

$$\begin{aligned}\vec{\sigma}_x+\vec{\sigma}'_x&=\vec{\sigma}_X\\ \vec{\sigma}_y+\vec{\sigma}'_y&=\vec{\sigma}_Y\\ \vec{\sigma}_z+\vec{\sigma}'_z&=\vec{\sigma}_Z\end{aligned}\tag{9.7.223}$$

Thus, three components are sufficient to know the normal stress forces on the surfaces along each axis.

If we make the sum of the moments of forces with respect to the centers of gravity for each axis of symmetry of the parallelepiped (XX',YY',ZZ'), it is evident that of the 12 tangential components 6 are sufficient to describe the whole of system.

Therefore for the $X\mathrm{O}Y$ plane passing through the center of gravity we have:

$$\begin{aligned}&(\vec{\tau}_{zy}\mathrm{d}x\mathrm{d}x)\frac{\mathrm{d}y}{2}+(\vec{\tau}_{zx}\mathrm{d}y\mathrm{d}z)\frac{\mathrm{d}x}{2}+(\vec{\tau}\mathrm{d}x\mathrm{d}z)\frac{\mathrm{d}y}{2}+(\vec{\tau}_{xz}\mathrm{d}z\mathrm{d}y)\frac{\mathrm{d}x}{2}\\ &=\frac{1}{2}\mathrm{d}V\left[(\vec{\tau}_{zy}+\vec{\tau}_{yz})+(\vec{\tau}_{zx}+\vec{\tau}_{xz})\right]=\frac{1}{2}\mathrm{d}V\left[\vec{\tau}_{ZY}+\vec{\tau}_{ZX}\right]=\frac{1}{2}\mathrm{d}V\vec{\tau}_{X\mathrm{O}Y}\end{aligned}\tag{9.7.224}$$

For the $X\mathrm{O}Z$ plane:

$$\begin{aligned}&(\vec{\tau}'_{yx}\mathrm{d}y\mathrm{d}z)\frac{\mathrm{d}x}{2}+(\vec{\tau}'_{yz}\mathrm{d}x\mathrm{d}y)\frac{\mathrm{d}z}{2}+(\vec{\tau}'_{xy}\mathrm{d}z\mathrm{d}y)\frac{\mathrm{d}x}{2}+(\vec{\tau}'_{zy}\mathrm{d}y\mathrm{d}x)\frac{\mathrm{d}z}{2}\\ &=\frac{1}{2}\mathrm{d}V\left[(\vec{\tau}'_{yx}+\vec{\tau}'_{xy})+(\vec{\tau}''_{yz}+\vec{\tau}'_{zy})\right]=\frac{1}{2}\mathrm{d}V\left[\vec{\tau}_{YX}+\vec{\tau}_{YZ}\right]=\frac{1}{2}\mathrm{d}V\vec{\tau}_{X\mathrm{O}Z}\end{aligned}\tag{9.7.225}$$

For the $Z\mathrm{O}Y$ plane:

$$\begin{aligned}&(\vec{\tau}''_{xz}\mathrm{d}x\mathrm{d}y)\frac{\mathrm{d}z}{2}+(\vec{\tau}''_{yx}\mathrm{d}x\mathrm{d}z)\frac{\mathrm{d}y}{2}+(\vec{\tau}''_{zx}\mathrm{d}y\mathrm{d}x)\frac{\mathrm{d}z}{2}+(\vec{\tau}''_{xy}\mathrm{d}z\mathrm{d}x)\frac{\mathrm{d}y}{2}\\ &=\frac{1}{2}\mathrm{d}V\left[(\vec{\tau}''_{xz}+\vec{\tau}''_{zx})+(\vec{\tau}''_{yx}+\vec{\tau}''_{xy})\right]=\frac{1}{2}\mathrm{d}V\left[\vec{\tau}_{ZX}+\vec{\tau}_{XY}\right]=\frac{1}{2}\mathrm{d}V\vec{\tau}_{Z\mathrm{O}Y}\end{aligned}\tag{9.7.226}$$

We can obtain the same components of equilibrium by considering this time a static regular elementary tetrahedron (extracted from the cube). The aim is to demonstrate that we find the 6 components determined previously.

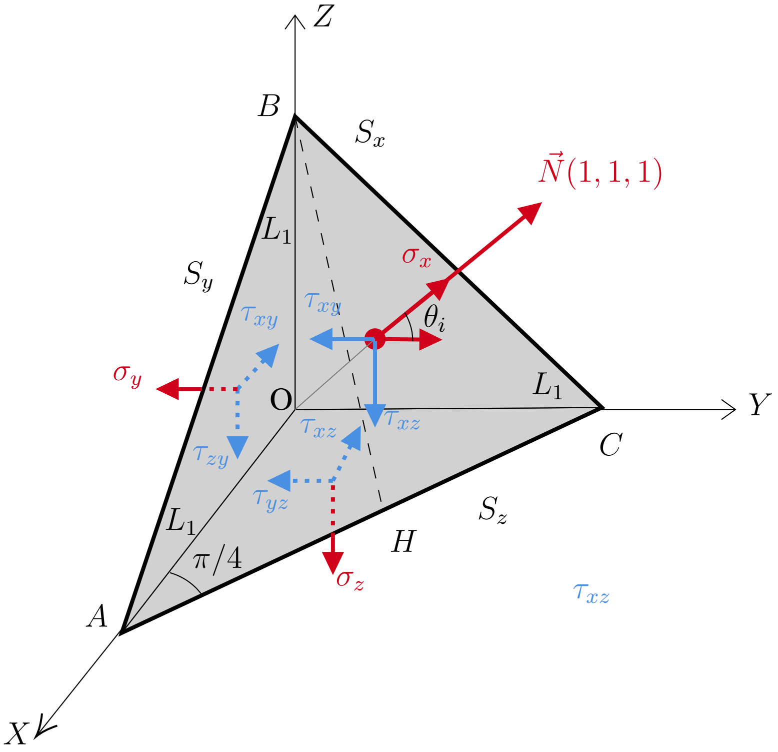

Z
B
S_x
$\vec{N}(1,1,1)$
L_1
S_y
σ_x
τ_{xy}
τ_{xy}
θ_i
σ_y
O
L_1
Y
τ_{xz}
τ_{xz}
C
τ_{zy}
τ_{yz}
S_z
L_1
H
$\pi/4$
σ_z
τ_{xz}
A
X

Figure 9.241: Regular elemental tetrahedron extracted from a fluid at equilibrium

Remark

It is important to notice again that the components of all the vectors visible in the figure above are expressed in newtons per unit area or in other words per unit of pressure (which is the unit of constraints for recall...).

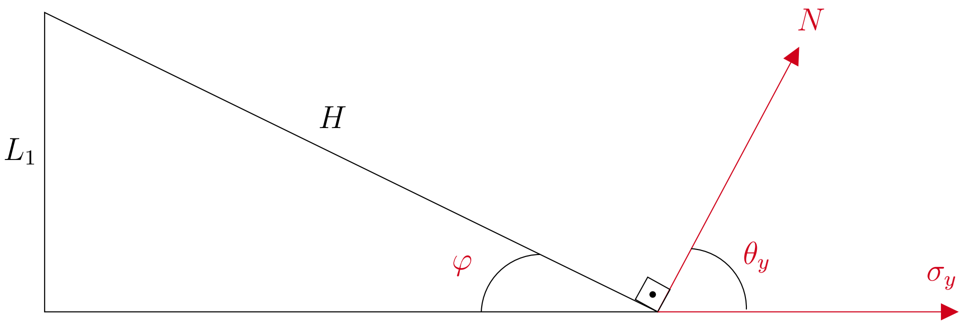
To know the surface of the faces OAC, OBC, OAB, we multiply the surface ABC (denoted hereafter: S) by the cosine of the angle that form the vectors $\vec{\sigma}_i$ and $\vec{N}$.

Indeed, given the surfaces:

$$S_y=\frac{L_1\cdot L_1}{2} \qquad \text{and} \qquad S=\frac{H\cdot\overline{AC}}{2}=\frac{H\cdot 2L_1\cos\left(\dfrac{\pi}{4}\right)}{2}=HL_1\frac{\sqrt{2}}{2} \tag{9.7.227}$$

However, let us try to express the S_i according to S. The schema below (slice of the tetrahedron) should help to understand the reasoning:

$$\varphi=\pi-\frac{\pi}{2}-\theta_y=\frac{\pi}{2}-\theta_y \tag{9.7.228}$$


N
H
L_1
φ
θ_y
σ_y

and therefore:

$$\sin(\varphi)=\frac{L_1}{H}=\sin\left(\frac{\pi}{2}-\theta_y\right)=\cos(\theta_y) \tag{9.7.229}$$

Finally:

$$S_y=\frac{H\cos(\theta_y)L_1}{2} \tag{9.7.230}$$

The ratio:

$$\frac{S_y}{S}=\frac{\dfrac{H\cos(\theta_y)L_1}{2}}{HL_1\dfrac{\sqrt{2}}{2}}=\frac{\cos(\theta_y)}{\sqrt{2}} \tag{9.7.231}$$

hence:

$$S_y=\frac{S}{\sqrt{2}}\cos(\theta_y) \tag{9.7.232}$$

The principle of analysis being the same for all other surfaces such as:

$$S_x=\frac{S}{\sqrt{2}}\cos(\theta_x)\qquad S_y=\frac{S}{\sqrt{2}}\cos(\theta_y)\qquad S_z=\frac{S}{\sqrt{2}}\cos(\theta_z) \tag{9.7.233}$$

We will write:

$$\frac{\cos(\theta_x)}{\sqrt{2}}=l\qquad\frac{\cos(\theta_y)}{\sqrt{2}}=m\qquad\frac{\cos(\theta_z)}{\sqrt{2}}=n \tag{9.7.234}$$

such that:

$$S_x=Sl\qquad S_y=Sm\qquad S_z=Sn \tag{9.7.235}$$

Remark

We can easily know the values of the θ_i using vector analysis. Indeed, the plane ABC being of equation (see section Analytical Geometry page 2800):

$$ax+by+cz+d=0\qquad a=b=c=L_1 \tag{9.7.236}$$

by simplifying by L_1:

$$1\cdot x+1\cdot y+1\cdot z+d'=0 \tag{9.7.237}$$

The vector normal to the plane being indeed:

$$\vec{N}=(1,1,1) \tag{9.7.238}$$

To know the cosines of the angle of the normal vector with the $\vec{\sigma}_i$, it is enough to assimilate these to the basis vectors $(\vec{i},\vec{j},\vec{k})$ such that (elementary trigonometry):

$$\cos(\theta_i)=\frac{\|\vec{i}\|}{\|\vec{N}\|}=\frac{||(1,0,0)||}{||(1,1,1)||}=\frac{1}{\sqrt{3}} \tag{9.7.239}$$

and proceeding in the same way for all the other θ.

The equilibrium of forces gives us:

$$N_xS-\sigma_xSl-\tau_{xy}Sm-\tau_{xz}Sn=0 \tag{9.7.240}$$

After simplification:

$$N_x=\sigma_xl+\tau_{xy}m+\tau_{xz}n \tag{9.7.241}$$

Following to the other axes:

$$\begin{aligned}N_y&=\tau_{yx}l+\sigma_ym+\tau_{yz}n\\N_z&=\tau_{zx}l+\tau_{zy}m+\sigma_zn\end{aligned} \tag{9.7.242}$$

In summary:

$$\begin{aligned}N_x&=\sigma_xl+\tau_{xy}m+\tau_{xz}nN_y&&=\tau_{yx}l+\sigma_ym+\tau_{yz}n\\N_z&=\tau_{zx}l+\tau_{zy}m+\sigma_zn\end{aligned} \tag{9.7.243}$$

Using the matrix representation, we get:

$$\boxed{\begin{bmatrix}N_x\\N_y\\N_z\end{bmatrix}=\begin{bmatrix}\sigma_{xx}&\tau_{xy}&\tau_{xz}\\\tau_{yx}&\sigma_{yy}&\tau_{yz}\\\tau_{zx}&\tau_{zy}&\sigma_{zz}\end{bmatrix}\begin{bmatrix}l\\m\\n\end{bmatrix}} \tag{9.7.244}$$

Either in index notation the normal and tangential constraints are given by the following traditional relation (where we do not distinguish what is tangential from what is normal so there is a loss of clarity that is typical habit of solid states physicists...):

$$\boxed{N_i=\sigma_{ij}n_j=\sum_j\sigma_{ij}n_j} \tag{9.7.245}$$

We see appear a mathematical quantity σ_{ij} having 9 components (corresponding for recall to the vector of the stress-forces projected into the nine 2×2 combinations of axis xx, zz, yy and planes $xy=yx$, $yz=zy$, $zx=xz$), whereas a vector in the same space $\mathbb{R}^3$ possesses 3 components. We know this kind of mathematical being that we have already studied in the in the section of Tensor Calculus. The quantity σ_{ij} is named "second order stress tensor" or "Cauchy stress tensor". Moreover, some components can be equal ($\sigma_{ij}=\sigma_{ji}$, if $i\neq j$), which would make it symmetrical. It then has only the 6 distinct components, relative to the numbers of components sufficient to completely describe a system at equilibrium.

To study the deformations of a continuous medium such as a fluid, we will first consider the case of very small deformations. The small displacements $\vec{D}$ of a point will be represented by (u,v,w) parallel to the axes of a referential OXYZ. We assume that these components are very small quantities continuously varying in the volume of the considerated body.

Let us consider a linear segment $\overline{\text{O}P}$ located in a solid before deformation. In an OXYZ repository, we will denote by (x_0,y_0,z_0) and (x_p,y_p,z_p) the coordinates of O and P.

During the deformation, the line $\overline{\text{O}P}$ becomes $\overline{\text{O}'P'}$ as shown below:

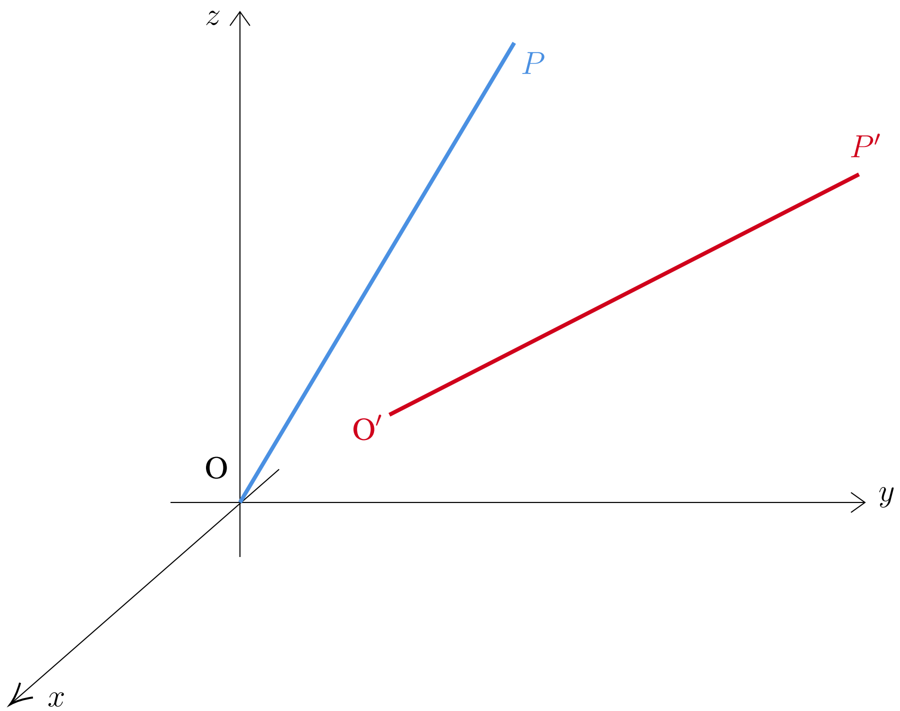


Figure 9.242: **Linear segment in a solid of fluid before and after deformation**

Given u_0, v_0, w_0 the displacements of the point O parallel to the axes OX, OY, OZ and u_p, v_p, w_p the displacements of the point P parallel to the same axes.

The coordinates of the points O′ and P' are then:

$$(x_0+u_0,y_0+v_0,z_0+w_0) \qquad \text{and} \qquad (x_p+u_p,y_p+v_p,z_p+w_p) \tag{9.7.246}$$

Before deformation, given L the length $\overline{\mathrm{O}P}$:

$$L^2=(x_p-x_0)^2+(y_p-y_0)^2+(z_p-z_0)^2 \tag{9.7.247}$$

After deformation, we have a length L' being equal to:

$$L'^2=[(x_p+u_p)-(x_0+u_0)]^2+[(y_p+v_p)-(y_0+v_0)]^2+[(z_p+w_p)-(z_0+w_0)]^2 \tag{9.7.248}$$

If ΔL is the elongation of the element $\overline{\mathrm{O}P}$ during the deformation, we have:

$$L'^2=(L+\Delta L)^2=L^2+2L\Delta L+\Delta L^2 \tag{9.7.249}$$

By carrying out the following transformations:

$$\begin{aligned}L'^2=L^2+2L\Delta L+\Delta L^2=&\\ &\left[(x_p+u_p)^2+(x_0+u_0)^2-2(x_p+u_p)(x_0+u_0)\right]+\\ &\left[(y_p+v_p)^2+(y_0+v_0)^2-2(y_p+v_p)(y_0+v_0)\right]+\\ &\left[(z_p+w_p)^2+(z_0+w_0)^2-2(z_p+w_p)(z_0+w_0)\right]\end{aligned} \tag{9.7.250}$$

By developing:

$$\begin{aligned}&[x_p^2+u_p^2+2x_pu_p+x_0^2+u_0^2+2x_0u_0-2x_px_0-2x_pu_0-2x_0u_p-2u_pu_0]+\\ &[y_p^2+v_p^2+2y_pv_p+y_0^2+v_0^2+2y_0v_0-2y_py_0-2y_pv_0-2y_0v_p-2v_pv_0]+\\ &[z_p^2+w_p^2+2z_pw_p+z_0^2+w_0^2+2z_0w_0-2z_pz_0-2z_pw_0-2z_0w_p-2w_pw_0]\end{aligned} \tag{9.7.251}$$

Therefore:

$$\begin{aligned}&L^2+2L\Delta L+\Delta L^2=\\&\left[(x_p-x_0)^2+(y_p-y_0)^2+(z_p-z_0)^2\right]+\\&2\left[(x_p-x_0)(u_p-u_0)+(y_p-y_0)(v_p-v_0)+(z_p-z_0)(w_p-w_0)\right]+\\&\left[(u_p-u_0)^2+(v_p-v_0)^2+(w_p-w_0)^2\right]\end{aligned} \tag{9.7.252}$$

By neglecting the terms of displacement of higher order and taking into account the relation:

$$L^2=(x_p-x_0)^2+(y_p-y_0)^2+(z_p-z_0)^2 \tag{9.7.253}$$

It comes that L^2 disappears with $(x_p-x_0)^2+(y_p-y_0)^2+(z_p-z_0)^2$ as well as the terms squared, we have then:

$$\frac{\Delta L}{L}=\frac{1}{L^2}\left[(x_p-x_0)(u_p-u_0)+(y_p-y_0)(v_p-v_0)+(z_p-z_0)(w_p-w_0)\right] \tag{9.7.254}$$

But the Analytical Geometry (elementary trigonometry: ratio of the sides opposite and adjacent to the hypotenuse) gives the following relations:

$$\frac{x_p-x_0}{L}=l \qquad \frac{y_p-y_0}{L}=m \qquad \frac{z_p-z_0}{L}=n \tag{9.7.255}$$

which are the direction cosines of line L.

We can then write:

$$\varepsilon=\frac{\Delta L}{L}=\frac{l}{L}(u_p-u_0)+\frac{m}{L}(v_p-v_0)+\frac{n}{L}(w_p-w_0) \tag{9.7.256}$$

The variation u_p being a small displacement, we use a Taylor series development (see section Sequences and Series page 1714) whose upper order terms are neglected (linearisation of equations):

$$u_p=u+x_p\partial_x u+y_p\partial_y u+z_p\partial_z u \tag{9.7.257}$$

We also have:

$$u_0=u+x_0\partial_x u+y_0\partial_y u+z_0\partial_z u \tag{9.7.258}$$

The difference gives:

$$u_p-u_0=(x_p-x_0)\partial_x u+(y_p-y_0)\partial_y u+(z_p-z_0)\partial_z u \tag{9.7.259}$$

So we can now write:

$$\begin{aligned}u_p-u_0&=lL\partial_x u+mL\partial_y u+nL\partial_z u\\v_p-v_0&=lL\partial_x v+mL\partial_y v+nL\partial_z v\\w_p-w_0&=lL\partial_x w+mL\partial_y w+nL\partial_z w\end{aligned} \tag{9.7.260}$$

Finally:

$$\begin{aligned}\varepsilon&=\frac{l}{L}(lL\partial_x u+mL\partial_y u+nL\partial_z u)+\frac{m}{L}(lL\partial_x v+mL\partial_y v+nL\partial_z v)+\frac{n}{L}(lL\partial_x w+mL\partial_y w+nL\partial_z w)\\&=l^2\partial_x u+ml\partial_y u+nl\partial_z u+ml\partial_x v+m^2\partial_y v+mn\partial_z v+nl\partial_x w+mn\partial_y w+n^2\partial_z w\end{aligned} \tag{9.7.261}$$

By grouping, we have:

$$\varepsilon = l^2\partial_x u + m^2\partial_y v + n^2\partial_z w + mn(\partial_z v + \partial_y w) + nl(\partial_z u + \partial_x w) + ml(\partial_y u + \partial_x v) \quad (9.7.262)$$

This expression allows, at any point, the computation of the deformation ε in a direction having as cosine direction l, m, n as a function of the displacements u, v, w at this point.

Let us consider the case where the line L coincides with the axis OX, we have $l=1$, $m=n=0$ the preceding equation then becomes:

$$\varepsilon_x = \partial_x u \quad (9.7.263)$$

We have, if L coincides with the OY axis ($m=1$, $l=n=0$) or with the OZ axis ($n=1$, $l=m=0$):

$$\varepsilon_y = \partial_y v \qquad \text{and} \qquad \varepsilon_z = \partial_z w \quad (9.7.264)$$

The quantities ε_x, ε_y, ε_z are named "normal deformations" and have no dimensional units.

For the interpretation of the terms $(\partial_z v + \partial_y w)$, $(\partial_z u + \partial_x w)$, $(\partial_y u + \partial_x v)$, we refer to the following figure:

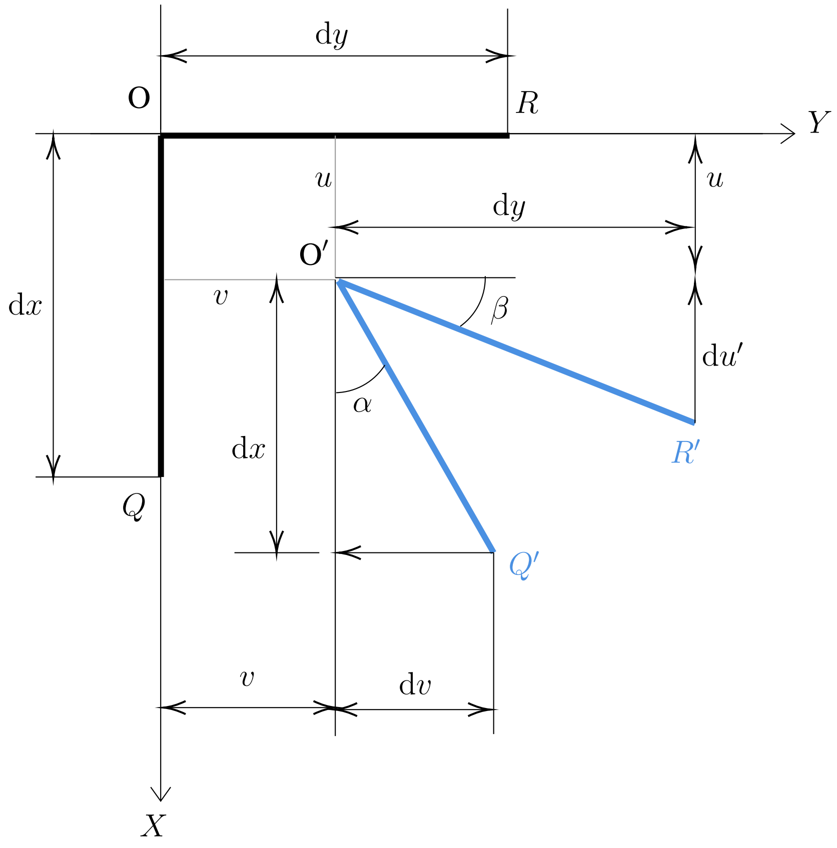


Figure 9.243: Situation that allows to make emerge shear stresses

Given two line segments $\overline{\text{O}R}$ and $\overline{\text{O}Q}$ located in the plane XOY. Before deformation $\overline{\text{O}R}$ and $\overline{\text{O}Q}$ coincided with the orthonormal referential YOX. After deformation, they can take the position $\overline{\text{O}'R'}$ and $\overline{\text{O}'Q'}$. The components of the displacement of O are u and v.

- The displacement component of R' is calculated as follows:

$$\mathrm{d}u = \mathrm{d}y\tan(\beta) \tag{9.7.265}$$

  with:

$$\beta \cong \tan(\beta) = \frac{\mathrm{d}u}{\mathrm{d}y} \tag{9.7.266}$$

  as the angle is small.

  In general, as $u = f(x,y,z)$, we will write:

$$\beta = \partial_y u \tag{9.7.267}$$

- The component of the displacement of Q' is it:

$$\alpha = \partial_x v \tag{9.7.268}$$

As before deformation, the angle $\widehat{Q\mathrm{O}R}$ is of $\pi/2$, after deformation, the right angle is reduced of a quantity $\alpha+\beta$. This reduction equation is named the "shear deformation" or "tangential deformation" and is denoted by γ_{xy}.

We will proceed in the same way for the other terms, hence:

$$\begin{aligned}\gamma_{xy} &= \partial_y u + \partial_x v\\ \gamma_{xz} &= \partial_z u + \partial_x w\\ \gamma_{yz} &= \partial_z v + \partial_y w\end{aligned} \tag{9.7.269}$$

Seeing what precedes, it is customary to define the "matrix of differential operators":

$$\boxed{\begin{bmatrix}\varepsilon_x\\ \varepsilon_y\\ \varepsilon_z\end{bmatrix} = D_{ij}\begin{bmatrix}u\\ v\\ w\end{bmatrix} = \begin{bmatrix}\partial_x & 0 & 0\\ 0 & \partial_y & 0\\ 0 & 0 & \partial_z\\ \partial_y & \partial_x & 0\\ 0 & \partial_z & \partial_y\\ \partial_z & 0 & \partial_x\end{bmatrix}\begin{bmatrix}u\\ v\\ w\end{bmatrix}} \tag{9.7.270}$$

Given the quadruplet of groups of equations previously demonstrated in this section (see deformations of solids):

$$\begin{aligned}
\sigma_x &= \frac{E}{1+\eta}\left[\varepsilon_x+\frac{\eta}{1-2\eta}(\varepsilon_x+\varepsilon_y+\varepsilon_z)\right]\\
\sigma_y &= \frac{E}{1+\eta}\left[\varepsilon_y+\frac{\eta}{1-2\eta}(\varepsilon_x+\varepsilon_y+\varepsilon_z)\right]\\
\sigma_x &= \frac{E}{1+\eta}\left[\varepsilon_x+\frac{\eta}{1-2\eta}(\varepsilon_x+\varepsilon_y+\varepsilon_z)\right]
\end{aligned}$$

$$\gamma_{xy}=\partial_y u+\partial_x v \qquad \gamma_{xz}=\partial_z u+\partial_x w \qquad \gamma_{yz}=\partial_z v+\partial_y w \tag{9.7.271}$$

$$\begin{bmatrix} N_x\\ N_y\\ N_z\end{bmatrix}=\begin{bmatrix}\sigma_{xx} & \tau_{xy} & \tau_{xz}\\ \tau_{yx} & \sigma_{yy} & \tau_{yz}\\ \tau_{zx} & \tau_{zy} & \sigma_{zz}\end{bmatrix}\begin{bmatrix} l\\ m\\ n\end{bmatrix}$$

$$\tau=\frac{E}{2(1+\eta)}\gamma$$

We can summarize:

$$\begin{aligned}
\sigma_{xx} &= \frac{E}{1+\eta}\left[\varepsilon_{xx}+\frac{\eta}{1-2\eta}(\varepsilon_{xx}+\varepsilon_{yy}+\varepsilon_{zz})\right]\\
\sigma_{yy} &= \frac{E}{1+\eta}\left[\varepsilon_{yy}+\frac{\eta}{1-2\eta}(\varepsilon_{xx}+\varepsilon_{yy}+\varepsilon_{zz})\right]\\
\sigma_{zz} &= \frac{E}{1+\eta}\left[\varepsilon_{zz}+\frac{\eta}{1-2\eta}(\varepsilon_{xx}+\varepsilon_{yy}+\varepsilon_{zz})\right]
\end{aligned}$$

$$\begin{aligned}
\tau_{xy}=\tau_{yx} &= \frac{E}{2(1+\eta)}\gamma_{xy}=\frac{E}{2(1+\eta)}(\partial_y u+\partial_x v)\\
\tau_{yz}=\tau_{zy} &= \frac{E}{2(1+\eta)}\gamma_{yz}=\frac{E}{2(1+\eta)}(\partial_z v+\partial_y w)\\
\tau_{xz}=\tau_{zy} &= \frac{E}{2(1+\eta)}\gamma_{zx}=\frac{E}{2(1+\eta)}(\partial_z u+\partial_x w)
\end{aligned}\tag{9.7.272}$$

At the view of the previous relations, it is customary to define the "matrix of transformation of deformation stresses":

$$\begin{bmatrix}\sigma_x\\ \sigma_y\\ \sigma_z\\ \tau_{xy}\\ \tau_{yz}\\ \tau_{zx}\end{bmatrix}=C_{ij}\begin{bmatrix}\varepsilon_x\\ \varepsilon_y\\ \varepsilon_z\\ \gamma_{xy}\\ \gamma_{yz}\\ \gamma_{yz}\\ \gamma_{zx}\end{bmatrix}=\frac{E}{(1+\eta)(1-2\eta)}\begin{bmatrix}1-\eta & \eta & \eta & 0 & 0 & 0\\ \eta & 1-\eta & \eta & 0 & 0 & 0\\ \eta & \eta & & 1-\eta & 0 & 0\\ 0 & 0 & 0 & \dfrac{1-2\eta}{2} & 0 & 0\\ 0 & 0 & 0 & 0 & \dfrac{1-2\eta}{2} & 0\\ 0 & 0 & 0 & 0 & 0 & \dfrac{1-2\eta}{2}\end{bmatrix}\begin{bmatrix}\varepsilon_x\\ \varepsilon_y\\ \varepsilon_z\\ \gamma_{xy}\\ \gamma_{yz}\\ \gamma_{zx}\end{bmatrix}\tag{9.7.273}$$

where it is possible to invert the matrix to thus obtain the "matrix of transformation of the stresses strains".

Generally, we put to simplify the notations (the reader must however not think that the shear deformation becomes a normal deformation!!!! it is only a convention of writing that must be remembered!):

$$\gamma_{xy}=2\varepsilon_{xy} \qquad \gamma_{xz}=2\varepsilon_{xz} \qquad \gamma_{yz}=2\varepsilon_{yz} \tag{9.7.274}$$

Similarly, we put:

$$\varepsilon=\varepsilon_{xx}+\varepsilon_{yy}+\varepsilon_{zz} \tag{9.7.275}$$

Thus finally:

$$\begin{aligned}\sigma_{xx}&=\frac{E}{1+\eta}\left(\varepsilon_{xx}+\frac{\eta}{1-2\eta}\varepsilon\right)\\ \sigma_{yy}&=\frac{E}{1+\eta}\left(\varepsilon_{yy}+\frac{\eta}{1-2\eta}\varepsilon\right)\\ \sigma_{zz}&=\frac{E}{1+\eta}\left(\varepsilon_{zz}+\frac{\eta}{1-2\eta}\varepsilon\right)\end{aligned} \tag{9.7.276}$$

$$\begin{aligned}\tau_{xy}=\tau_{yx}&=\frac{E}{2(1+\eta)}2\varepsilon_{xy}\\ \tau_{yz}=\tau_{zy}&=\frac{E}{2(1+\eta)}2\varepsilon_{yz}\\ \tau_{xz}=\tau_{zx}&=\frac{E}{2(1+\eta)}2\varepsilon_{zy}\end{aligned}$$

Taking into account that:

$$G=\frac{E}{2(1+\eta)} \tag{9.7.277}$$

We get the shear tensions as follows:

$$\begin{aligned}\sigma_{xx}&=2G\left(\varepsilon_{xx}+\frac{\eta}{1-2\eta}\varepsilon\right)\\ \sigma_{yy}&=2G\left(\varepsilon_{yy}+\frac{\eta}{1-2\eta}\varepsilon\right)\\ \sigma_{zz}&=2G\left(\varepsilon_{zz}+\frac{\eta}{1-2\eta}\varepsilon\right)\end{aligned} \tag{9.7.278}$$

$$\begin{aligned}\tau_{xy}=\tau_{xy}&=2G\varepsilon_{xy}\\ \tau_{yz}=\tau_{zy}&=2G\varepsilon_{yz}\\ \tau_{xz}=\tau_{zx}&=2G\varepsilon_{zy}\end{aligned}$$

Consider now, for example, a fluid flowing in the direction of OY with a velocity gradient in the direction of x:

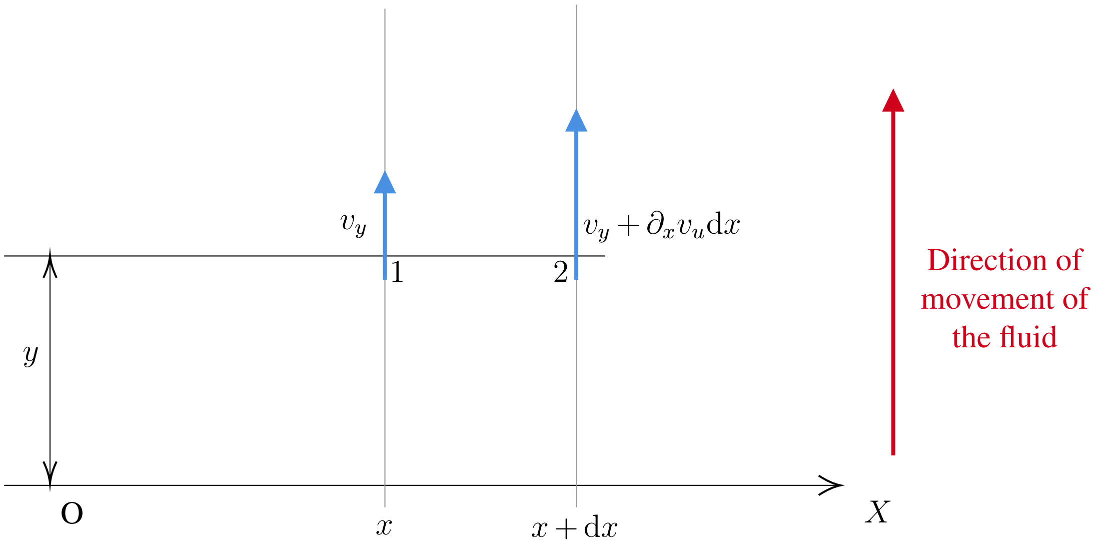

v_y
$v_y+\partial_x v_u \mathrm{d}x$
1
2
y
O
x
$x+\mathrm{d}x$
X
Direction of movement of the fluid

Figure 9.244: **Situation that allows the emergence of normal constraints**

By placing ourselves at the level of y and at point 1 of the abscissa x, we have a velocity v_y and at point 2 of abscissa $x+\mathrm{d}x$, a speed:

$$v_y+\partial_y \mathrm{d}x \tag{9.7.279}$$

with $\partial_x v_y \neq 0$.

In the direction of x, there is no velocity component therefore:

$$v_x=0 \tag{9.7.280}$$

with $\partial_y v_x=0$.

We now assume that the shear tensions are proportional to $\partial_x v_y$ to a given factor such that:

$$\tau_{xy}=\mu_d \partial_x v_y \tag{9.7.281}$$

with for recall:

$$\mu_d=\frac{\mathrm{d}F}{\partial_z v \mathrm{d}S} \tag{9.7.282}$$

It is therefore possible to consider displacements per unit of time by putting:

$$u \rightarrow v_x \qquad v \rightarrow v_y \qquad w \rightarrow v_z \tag{9.7.283}$$

By relating the previous relation of μ_d of:

$$\begin{aligned}\tau_{xy}&=G(\partial_y u+\partial_y v)=\mu_d(\partial_y v_x+\partial_x v_y)\\ \tau_{yz}&=G(\partial_z v+\partial_y w)=\mu_d(\partial_z v_y+\partial_y v_z)\\ \tau_{zx}&=G(\partial_x w+\partial_z u)=\mu_d(\partial_x v_z+\partial_z v_x)\end{aligned} \tag{9.7.284}$$

we can say that G initially valid in a solid elastic medium considered by its displacements is the analogue of μ_d in the case of a viscous fluid considered by its displacements per unit of time. So we see that the units are conserved.

Considering also the deformations per unit of time for the normal stresses (we will return to this in detail later), we then have the system of equations:

$$\begin{aligned}
\sigma_{xx}&=2\mu_d\left(\dot{\varepsilon}_{xx}+\frac{\eta}{1-2\eta}\dot{\varepsilon}\right)=2\mu_d\theta_{xx}+2\mu_d\frac{\eta}{1-2\eta}\dot{\varepsilon}=2\mu_d\theta_{xx}+\alpha(\theta_{xx}+\theta_{yy}+\theta_{zz})\\
\sigma_{yy}&=2\mu_d\left(\dot{\varepsilon}_{yy}+\frac{\eta}{1-2\eta}\dot{\varepsilon}\right)=2\mu_d\theta_{yy}+2\mu_d\frac{\eta}{1-2\eta}\dot{\varepsilon}=2\mu_d\theta_{yy}+\alpha(\theta_{xx}+\theta_{yy}+\theta_{zz})\\
\sigma_{zz}&=2\mu_d\left(\dot{\varepsilon}_{zz}+\frac{\eta}{1-2\eta}\dot{\varepsilon}\right)=2\mu_d\theta_{zz}+2\mu_d\frac{\eta}{1-2\eta}\dot{\varepsilon}=2\mu_d\theta_{zz}+\alpha(\theta_{xx}+\theta_{yy}+\theta_{zz})
\end{aligned}\tag{9.7.285}$$

$$\begin{aligned}
\tau_{xy}&=\tau_{xy}=2\mu_d\dot{\varepsilon}_{xy}=2\mu_d\theta_{xy}\\
\tau_{yz}&=\tau_{zy}=2\mu_d\dot{\varepsilon}_{yz}=2\mu_d\theta_{yz}\\
\tau_{xz}&=\tau_{xz}=2\mu_d\dot{\varepsilon}_{zy}=2\mu_d\theta_{xz}
\end{aligned}$$

Thus, we get a condensed notation:

$$P_{ij}=2\mu_d\theta_{ij}+\alpha(\theta_{11}+\theta_{22}+\theta_{33})\delta_{ij}\tag{9.7.286}$$

where δ_{ij} is the Kronecker's symbol:

$$\delta_{ij}=\begin{cases}\delta_{ij}=1 & i=j\\ \delta_{ij}=0 & i\neq j\end{cases}\tag{9.7.287}$$

The tensor P_{ij} thus describes in part the set of stresses of a viscous fluid in which we assumed, within the framework of the Newtonian fluid hypothesis, that there are linear relations between the tensions and the normal deformations!!!!

We now put the sum of dynamic constraints in a general form that we will justify:

$$P_{ij}=2\mu_d\theta_{ij}+\alpha(\theta_{11}+\theta_{22}+\theta_{33})\delta_{ij}-p\delta_{ij}\tag{9.7.288}$$

where the term $p\delta_{ij}$ is justified by the fact that in the static case a constant dynamic pressure p always exists at a point of a fluid, situation that we do not have in the case of a solid. To justify the negative sign, we observe that in the expression of P_{ij}, the first two terms of the right-hand side correspond, in the previous study, to extension constraints while the pressure p corresponds to a compression of the fluid.

We now have to determine the coefficient α. Given $i=j$, then we have $\delta_{ij}=1$. It comes successively and by addition:

$$\begin{array}{lr}
 & P_{11}=2\mu_d\theta_{11}+\alpha(\theta_{11}+\theta_{22}+\theta_{33})-p\\
 & P_{22}=2\mu_d\theta_{22}+\alpha(\theta_{11}+\theta_{22}+\theta_{33})-p\\
+ & P_{33}=2\mu_d\theta_{33}+\alpha(\theta_{11}+\theta_{22}+\theta_{33})-p\\
\hline
\multicolumn{2}{r}{P_{11}+P_{22}+P_{33}=2\mu_d(\theta_{11}+\theta_{22}+\theta_{33})+3\alpha(\theta_{11}+\theta_{22}+\theta_{33})-3p}
\end{array}\tag{9.7.289}$$

This expression must satisfy also a fluid which is also in a static (at rest) situation such that:

$$P_{11}+P_{22}+P_{33}=-3p \tag{9.7.290}$$

It comes then that in the static case:

$$2\mu_d(\theta_{11}+\theta_{22}+\theta_{33})+3\alpha(\theta_{11}+\theta_{22}+\theta_{33})=0 \tag{9.7.291}$$

Since:

$$\theta_{11}+\theta_{22}+\theta_{33}\neq 0 \tag{9.7.292}$$

We have then:

$$2\mu_d+3\alpha=0\Rightarrow\alpha=-\frac{2}{3}\mu_d \tag{9.7.293}$$

The general expression of the constraints is then written:

$$P_{ij}=2\mu_d\theta_{ij}-\frac{2}{3}\mu_d(\theta_{11}+\theta_{22}+\theta_{33})\delta_{ij}-p\delta_{ij} \tag{9.7.294}$$

Presently, we will introduce the operators of the vector analysis in order to have a more general expression. In this way, we can adapt the formulation to any coordinate system (Cartesian, cylindrical, spherical, etc.) which will facilitate the solving of practical problems.

We have seen that for a solid we had:

$$\begin{aligned}\gamma_{xy}&=2\varepsilon_{xy}=\partial_y u+\partial_x v\\ \gamma_{xz}&=2\varepsilon_{xz}=\partial_z u+\partial_x w\\ \gamma_{yz}&=2\varepsilon_{yz}=\partial_z v+\partial_y w\end{aligned} \tag{9.7.295}$$

We will determine these equations in an index form by always considering displacements per unit of time (velocities).

$$\dot{\varepsilon}_{ij}=\frac{1}{2}(\partial_{x^j}v_i+\partial_{x^i}v_j) \tag{9.7.296}$$

such that $i,j=\{1,2,3\}$ and that:

$$1\rightarrow x\qquad x^1=x\qquad 2\rightarrow y\qquad x^2=y\qquad 3\rightarrow z\qquad x^3=z \tag{9.7.297}$$

For $i=j=1$, we have therefore:

$$\dot{\varepsilon}_{11}=\frac{1}{2}(\partial_{x^1}v_1+\partial_{x^1}v_1) \tag{9.7.298}$$

or:

$$\dot{\varepsilon}_{xx}=\partial_x v_x \tag{9.7.299}$$

$$\begin{gathered}\dot{\varepsilon}_{xx}=\partial_x v_x\qquad\dot{\varepsilon}_{yy}=\partial_y v_y\qquad\dot{\varepsilon}_{zz}=\partial_z v_z\\ \dot{\varepsilon}_{xy}=\frac{1}{2}(\partial_y v_x+\partial_x v_y)\qquad\dot{\varepsilon}_{xz}=\frac{1}{2}(\partial_z v_x+\partial_x v_z)\qquad\dot{\varepsilon}_{yz}=\frac{1}{2}(\partial_z v_y+\partial_y v_z)\end{gathered} \tag{9.7.300}$$

By doing the sum of the terms of the normal deformations rate we get:

$$\dot{\varepsilon}_{xx}+\dot{\varepsilon}_{yy}+\dot{\varepsilon}_{zz}=\partial_x v_x+\partial_y v_y+\partial_z v_z \tag{9.7.301}$$

But, the tools of Vector Calculus allow us to write:

$$\dot{\varepsilon}_{xx}+\dot{\varepsilon}_{yy}+\dot{\varepsilon}_{zz}=\partial_x v_x+\partial_y v_y+\partial_z v_z=\vec{\nabla}\circ\vec{v} \tag{9.7.302}$$

For a fluid, we will then have by analogy:

$$\theta_{11}+\theta_{22}+\theta_{33}=\vec{\nabla}\circ\vec{v} \tag{9.7.303}$$

The general dynamic equation of stresses will then be written in the following form of a Newtonian fluid (under small deformations):

$$\boxed{P_{ij}=2\mu_d\theta_{ij}-\frac{2}{3}\mu_d(\vec{\nabla}\circ\vec{v})\delta_{ij}-p\delta_{ij}} \tag{9.7.304}$$

This stress tensors is written by some teachers and textbooks authors as following (this notion is a little bit dangerous but it can be justified in the context of an in-deep study of fluids dynamics):

$$\sigma=-pI+\lambda(\vec{\nabla}\circ\vec{v})I+2\mu D \tag{9.7.305}$$

or also to differentiate vector and tensor:

$$\underline{\underline{\sigma}}=-p\underline{\underline{I}}+\lambda(\underline{\nabla}\circ\underline{v})\underline{\underline{I}}+2\mu\underline{\underline{D}} \tag{9.7.306}$$

If the normal constraints are negligible (incompressible fluid) the second term simplifies and we have then (relation that we will use in the section of Marine & Weather Engineering):

$$\underline{\underline{\sigma}}=-p\underline{\underline{I}}+2\mu\underline{\underline{D}} \tag{9.7.307}$$

It is now useful to go back in a developed form for the preceding equation, remembering that (see above):

$$\begin{aligned}P_{xx}&=-p-\frac{2}{3}\mu_d(\vec{\nabla}\circ\vec{v})+2\mu_d\partial_x v_x\\P_{yy}&=-p-\frac{2}{3}\mu_d(\vec{\nabla}\circ\vec{v})+2\mu_d\partial_y v_y\\P_{zz}&=-p-\frac{2}{3}\mu_d(\vec{\nabla}\circ\vec{v})+2\mu_d\partial_z v_z\end{aligned} \tag{9.7.308}$$

$$\begin{aligned}P_{xy}&=\mu_d(\partial_x v_x+\partial_x v_y)\\P_{xz}&=\mu_d(\partial_z v_x+\partial_x v_z)\\P_{yz}&=\mu_d(\partial_z y+\partial_y v_z)\end{aligned}$$

Let us now write the system of Newton equations (sums of dynamic internal and external constraints to a volume element of a fluid) which is:

$$\frac{F_T}{V}=\frac{F_{\text{int}}}{V}+\frac{F_{\text{ext}}}{V}=\rho f_i+\frac{\partial}{\partial x_i}\sum_{j=1}^{3}P_{ij} \tag{9.7.309}$$

where:

- $\dfrac{\partial}{\partial x_i}\sum_{j=1}^{3}P_{ij}$ is the sum of external forces by unit volume
- f_i is the traditional (unfortunate ...) notation of the mass acceleration in $[\mathrm{m}\cdot\mathrm{s}^{-2}]$
- ρ is the fluid density

The previous relation can be rewritten in vector form:

$$\rho \mathrm{d}_t\vec{v}=\vec{\nabla}(\vec{\sigma})+\rho\vec{f} \tag{9.7.310}$$

with obviously using tensor summation convention:

$$\vec{\nabla}(\vec{\sigma})_i=(\mathrm{div}\vec{\sigma})_i=\frac{\partial\sigma_{ij}}{\partial x_j}=\frac{\partial P_{ij}}{\partial x_j} \tag{9.7.311}$$

Then we have:

$$\begin{aligned}\rho\mathrm{d}_t v_x&=\partial_x P_{xx}+\partial_y P_{xy}+\partial_z P_{xz}+\rho f_x\\ \rho\mathrm{d}_t v_y&=\partial_x P_{yx}+\partial_y P_{yy}+\partial_z P_{yz}+\rho f_y\\ \rho\mathrm{d}_t v_z&=\partial_x P_{zx}+\partial_z P_{zz}+\partial_z P_{zz}+\rho f_z\end{aligned} \tag{9.7.312}$$

By introducing the expressions of P_{ij} obtained in the above relation, we arrive at the equations:

$$\begin{aligned}\rho\mathrm{d}_t v_x&=-\partial_x p-\partial_x\left(\frac{2}{3}\mu_d(\vec{\nabla}\circ\vec{v})-2\mu_d\partial_x v_x\right)+\partial_y\left[\mu_d\left(\partial_y v_x+\partial_x v_y\right)\right]+\partial_z\left[\mu_d\left(\partial_z v_x+\partial_x v_z\right)\right]+\rho f_x\\ \rho\mathrm{d}_t v_y&=\partial_x\left[\mu_d\left(\partial_y v_x+\partial_x v_y\right)\right]-\partial_y p-\partial_y\left(\frac{2}{3}\mu_d(\vec{\nabla}\circ\vec{v})-2\mu_d\partial_y v_y\right)+\partial_z\left[\mu_d\left(\partial_z v_y+\partial_y v_z\right)\right]+\rho f_y\\ \rho\mathrm{d}_t v_z&=\partial_x\left[\mu_d\left(\partial_z v_x+\partial_x v_z\right)\right]+\partial_y\left[\mu_d(\partial_z v_y+\partial_y v_z))\right]-\partial_z p-\partial_z\left(\frac{2}{3}\mu_d(\vec{\nabla}\circ\vec{v})-2\mu_d\partial_d\partial_z v_z\right)+\rho f_z\end{aligned} \tag{9.7.313}$$

These are the "Navier-Stokes equations of the dynamics of Newtonian fluids" for small deformations. There opne condensed forms which we shall now proceed to determine!

By taking Navier-Stokes's first equation and expanding it, it comes:

$$\begin{aligned}\rho\mathrm{d}_t v_x&=-\partial_x p-\partial_x\left(\frac{2}{3}\mu_d(\vec{\nabla}\circ\vec{v})-2\mu_d\partial_x v_x\right)+\partial_y\left[\mu_d\left(\partial_y v_x+\partial_x v_y\right)\right]+\partial_z\left[\mu_d\left(\partial_z v_x+\partial_x v_z\right)\right]+\rho f_x\\ &=-\partial_x p-\frac{2}{3}\mu_d\partial_x(\vec{\nabla}\circ\vec{v})+2\mu_d\partial_x^2 v_x+\mu_d\partial_y^2 v_x+\mu_d\partial_{yx}^2 v_y+\mu_d\partial_z^2 v_x+\mu_d\partial_{zx}^2 v_z+\rho f_x\\ &=-\partial_x p-\frac{2}{3}\mu\partial_x(\vec{\nabla}\circ\vec{v})+\underbrace{\underline{\mu_d\partial_x^2 v_x}+\underline{\underline{\mu_d\partial_y^2 v_x}}}_{2\mu_d\partial_x^2 v_x}+\underline{\underline{\mu_d\partial_y^2 v_x}}+\underline{\mu_d\partial_y^2 v_x}+\underline{\mu_d\partial_{yx}^2 v_y}+\underline{\underline{\mu_d\partial_z^2 v_x}}+\rho f_x\\ &=-\partial_x p-\frac{2}{3}\mu_d\partial_x(\vec{\nabla}\circ\vec{v})+\mu_d(\partial_x^2 v_x+\partial_{yx}^2 v_y+\partial_{zx}^2 v_z)+\mu_d(\partial_x^2 v_x+\partial_y^2 v_x+\partial_z^2 v_x)+\rho f_x\end{aligned} \tag{9.7.314}$$

As:

$$\partial_x^2 v_x+\partial_{yx}^2 v_y+\partial_{zx}^2 v_z=\partial(\partial_x v_x+\partial_y v_y+\partial_z v_z)=\partial_x(\vec{\nabla}\circ\vec{v}) \tag{9.7.315}$$

and that:

$$\partial_x^2 v_x + \partial_y^2 v_x + \partial_z^2 v_x = \vec{\nabla} \circ (\vec{\nabla}(v_x)) \tag{9.7.316}$$

We get:

$$\rho \mathrm{d}_t v_x = -\partial_x p - \frac{2}{3}\mu_d \partial_x (\vec{\nabla} \circ \vec{v}) + \mu_d \partial_x (\vec{\nabla} \circ \vec{v}) + \mu_d \vec{\nabla} \circ (\vec{\nabla}(v_x)) + \rho f_x \tag{9.7.317}$$

Simplifying, it comes therefore:

$$\rho \mathrm{d}_t v_x = -\partial_x p + \frac{1}{3}\mu_d \partial_x (\vec{\nabla} \circ \vec{v}) + \mu_d \vec{\nabla} \circ (\vec{\nabla}(v_x)) + \rho f_x \tag{9.7.318}$$

By operating in the same way for the other two components, we can reduce the system of Navier-Stokes equations to a single vector equation:

$$\rho \mathrm{d}_t \vec{v} = -\vec{\nabla} p + \frac{1}{3}\mu_d \vec{\nabla}(\vec{\nabla} \circ \vec{v}) + \mu_d \vec{\nabla} \circ (\vec{\nabla} \vec{v}) + \rho \vec{f} \tag{9.7.319}$$

As (see section Vector Calculus page 1903 and page 1911) the vector Laplacian is given by (recalling the different existing notations):

$$\Delta \vec{v} = \vec{\nabla}^2 \vec{v} = \vec{\nabla} \circ (\vec{\nabla} \vec{v}) = \vec{\nabla}(\vec{\nabla} \circ \vec{v}) - \vec{\nabla} \times (\vec{\nabla} \times \vec{v}) \tag{9.7.320}$$

We have:

$$\rho \mathrm{d}_t \vec{v} = -\vec{\nabla} p + \frac{1}{3}\mu_d \vec{\nabla}(\vec{\nabla} \circ \vec{v}) + \mu_d \left[\vec{\nabla}(\vec{\nabla} \circ \vec{v}) - \vec{\nabla} \times (\vec{\nabla} \times \vec{v})\right] + \rho \vec{f} \tag{9.7.321}$$

Thus finally:

$$\rho_{\mathrm{d}} \vec{v} = -\vec{\nabla} p + \frac{4}{3}\vec{\nabla}(\vec{\nabla} \circ \vec{v}) - \mu_d \vec{\nabla} \times (\vec{\nabla} \times \vec{v}) + \rho \vec{f} \tag{9.7.322}$$

Remark

We can also sometimes find in the literature an equation containing a second viscosity ${\mu'}_d$, whereas μ_d only manifests itself rigorously during pure shearing according to our hypotheses, ${\mu'}_d$ appears during an omnidirectional compression accompanied by a variation of density.

The preceding equation is then written:

$$\boxed{\rho_{\mathrm{d}} \vec{v} = -\vec{\nabla} p + \frac{4}{3}(\mu_d + {\mu'}_d)\vec{\nabla}(\vec{\nabla} \circ \vec{v}) - \mu_d \vec{\nabla} \times (\vec{\nabla} \times \vec{v}) + \rho \vec{f}} \tag{9.7.323}$$

This is the "Navier-Stokes equation" (still for Newtonian fluids in small deformations) or also named "equation of motion for a Newtonian fluid".

9.7.2.4.1 Incompressible flow

In fluid mechanics or more generally Continuum Mechanics, an "incompressible flow" or "isochoric flow" refers to a flow in which the material density is constant within a fluid parcel an infinitesimal volume that moves with the flow velocity. An equivalent statement that implies incompressibility is that the divergence of the flow velocity is zero (see the derivation below, which illustrates why these conditions are equivalent).

Incompressible flow does not imply that the fluid itself is incompressible. It is shown in the derivation below that (under the right conditions) even compressible fluids can – to a good approximation – be modelled as an incompressible flow. Incompressible flow implies that the density remains constant within a parcel of fluid that moves with the flow velocity.

In an incompressible fluid, we have by definition $\rho=c^{te}$. The conservation equation which is (see section Thermodynamics page 3683):

$$\partial_t\rho+\vec{\nabla}\circ(\rho\vec{v})=0 \tag{9.7.324}$$

is then written:

$$\rho\vec{\nabla}\circ\vec{v}=0 \tag{9.7.325}$$

Therefore the divergence of the velocity field is zero:

$$\vec{\nabla}\circ\vec{v}=0 \tag{9.7.326}$$

The Navier-Stokes equation in the form:

$$\rho\mathrm{d}_t\vec{v}=-\vec{\nabla}p+\frac{1}{3}\mu_d\vec{\nabla}(\vec{\nabla}\circ\vec{v})+\mu_d\vec{\nabla}\circ(\vec{\nabla}\vec{v})+\rho\vec{f} \tag{9.7.327}$$

is then written:

$$\rho\mathrm{d}_t\vec{v}=-\vec{\nabla}p+\mu_d\vec{\nabla}\circ(\vec{\nabla}\vec{v})+\rho\vec{f} \tag{9.7.328}$$

or differently using the traditional notation of the Laplacian of a vector field:

$$\rho\mathrm{d}_t\vec{v}=-\vec{\nabla}p+\mu_d\vec{\nabla}^2\vec{v}+\rho\vec{f} \tag{9.7.329}$$

If, moreover the viscosity μ_d is negligible, we have for a perfect fluid:

$$\boxed{\rho\mathrm{d}_t\vec{v}=-\vec{\nabla}p+\rho\vec{f}} \tag{9.7.330}$$

This equation is named the "Euler equation of the 1st form" or "local equation of the balance conservation of the momentum". We will reuse this relation as part of our study of gravity waves in the section of Weather & Marine Engineering and also in the section of Cosmology to study the Einstein Universe models! When $\vec{v}=\vec{0}$ we have what is commonly named an "hydrostatic balance".

Remark

It was proved in 11996 (holocene calendar) by Vladimir Scheffer (but the demonstration is horribly complex and comprehensible only by some people on the planet) that the Euler equation in the plan allows a spontaneous creation of energy! So an energy creation from nothing! A year later another mathematician presents a new proof of the same result. A few years later, two young mathematicians proved that it was useless to impose a criterion to the solutions to avoid this paradox ...

There exists a second form of the Euler equation in the framework of an incompressible fluid and of negligible viscosity which we will immediately determine (often used in industry):

If $\vec{v}=f(x,y,z,t)$ we can write:

$$\mathrm{d}_t\vec{v}=\frac{\partial\vec{v}}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t}+\frac{\partial\vec{v}}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t}+\frac{\partial\vec{v}}{\partial z}\frac{\mathrm{d}z}{\mathrm{d}t}+\frac{\partial\vec{v}}{\partial t} \tag{9.7.331}$$

This can also be written:

$$\begin{aligned}\mathrm{d}_t\vec{v}&=(\partial_x\vec{v})v_x+(\partial_y\vec{v})v_y+(\partial_z\vec{v})v_z+\partial_t\vec{v}\\&=\left[\partial_x(v_x\vec{i}+v_y\vec{j}+v_z\vec{k})\right]v_x+\left[\partial_y(v_x\vec{i}+v_y\vec{j}+v_z\vec{k})\right]v_y+\left[\partial_z(v_x\vec{i}+v_y\vec{j}+v_z\vec{k})\right]v_z+\partial_t\vec{v}\\&=[v_x\partial_x v_x]\vec{i}+[v_x\partial_x v_y]\vec{j}+[v_x\partial_x v_z]\vec{k}+[v_y\partial_y v_x]\vec{i}+[v_y\partial_y v_y]\vec{j}+[v_y\partial_y v_z]\vec{k}+[v_z\partial_z v_x]\vec{i}\\&\quad+[v_z\partial_z v_y]\vec{j}+[v_z\partial_z v_z]\vec{k}+\partial_t\vec{v}\end{aligned} \tag{9.7.332}$$

What is also written:

$$(v_x\partial_x+v_y\partial_y+v_z\partial_z)(v_x\vec{i}+v_y\vec{j}+v_z\vec{k})+\partial_t\vec{v} \tag{9.7.333}$$

The first factor can be considered as the following scalar product:

$$(v_x\partial_x+v_y\partial_y+v_z\partial_z)=\left[\underbrace{(v_x\vec{i}+v_y\vec{j}+v_z\vec{k})}_{=\vec{v}}\underbrace{(\partial_x\vec{i}+\partial_y\vec{j}+\partial_z\vec{k})}_{\vec{\nabla}}\right]=\vec{v}\circ\vec{\nabla} \tag{9.7.334}$$

Thus:

$$\mathrm{d}_t\vec{v}=(v_x\partial_x+v_y\partial_y+v_z\partial_z)(v_x\vec{i}+v_y\vec{j}+v_z\vec{k})+\partial_t\vec{v}=(\vec{v}\circ\vec{\nabla})\vec{v}+\partial_t\vec{v} \tag{9.7.335}$$

The "material derivative" (there are many other names for the material derivative[33]) can then take the following condensed form (we thus find the same result as in our study of the second generalized Newton force proved in the section of Classical Mechanics!!!):

$$\mathrm{d}_t\vec{v}=\partial_t\vec{v}+(\vec{v}\circ\vec{\nabla})\vec{v} \tag{9.7.336}$$

That is often written in text books:

$$\boxed{\frac{D\vec{v}}{Dt}=\frac{\partial\vec{v}}{\partial t}+\vec{v}\vec{\nabla}\circ\vec{v}} \tag{9.7.337}$$

where:

---

[33] advective derivative, convective derivative, derivative following the motion, hydrodynamic derivative, Lagrangian derivative, Lagrangian acceleration, particle derivative, substantial derivative, substantive derivative, Stokes derivative, total derivative, ...

- $\partial\vec{v}/\partial t$ is the "Eulerian derivative". That is the rate of change at a fixed point.
- $\vec{\nabla}\circ\vec{v}$ is the "advection term". Advection is the transport of a scalar or vectorial quantity preserved by a vector field. An example is the transport of polluting matter by the flow of a river. In meteorology and oceanography, advection refers mainly to the horizontal transport of certain properties by the fluids considered, including transport by wind or currents: advection of water vapor, heat, salinity, etc.
- $\vec{v}\vec{\nabla}\circ\vec{v}$ is the "convective acceleration". Convective is the acceleration defined as the rate of change of velocity due to the change of position of fluid particles in a fluid flow.

If the material derivative is zero then the field is conserved following the motion and the local rate of change is entirely due to advection.

Remark

The x component of the material derivative is therefore (we will use this again in the section of Marine & Weather Engineering):

$$\mathrm{d}_t v_x=\partial_t v_x+v_x\partial_x v_x+v_y\partial_y v_x+v_z\partial_z v_x \tag{9.7.338}$$

what some experts in the field generally note for any component:

$$\frac{D}{Dt}=\partial_t+v_x\partial_x+v_y\partial_y+v_z\partial_z=\partial_t+\vec{v}\circ\vec{\nabla} \tag{9.7.339}$$

The Euler equation of the first form:

$$\rho\mathrm{d}_t\vec{v}=-\vec{\nabla}p+\rho\vec{f} \tag{9.7.340}$$

becomes taking into account the material derivative:

$$\rho\left[\partial_t\vec{v}+(\vec{v}\circ\vec{\nabla})\vec{v}\right]=\rho\vec{f}-\vec{\nabla}p \tag{9.7.341}$$

Remark

Notice that if we didn't neglect the viscosity μ_d we would simply have:

$$\rho\left[\partial_t\vec{v}+(\vec{v}\circ\vec{\nabla})\vec{v}\right]=\rho\vec{f}+\mu_d\vec{\nabla}^2\vec{v}-\vec{\nabla}p \tag{9.7.342}$$

This is a very common form of the Navier-Stokes equation that we can also found in plenty of textbooks!

Or (common form in the literature):

$$\partial_t\vec{v}+(\vec{v}\circ\vec{\Delta})\vec{v}=-\frac{1}{\rho}\vec{\nabla}p+\vec{f} \tag{9.7.343}$$

We have proved in the section of Vector Calculus (see 1911) that:

$$\vec{\nabla}(\vec{u}\circ\vec{v})=\vec{u}\times(\vec{\nabla}\times\vec{v})+\vec{v}\times(\vec{\nabla}\times\vec{u})+(\vec{u}\circ\vec{\nabla})\vec{v}+(\vec{v}\circ\vec{\nabla})\vec{u} \tag{9.7.344}$$

If we put $\vec{u}=\vec{v}$, we have:

$$\vec{\nabla}\|\vec{v}\|^2=\vec{\nabla}v^2=2\vec{v}\times(\vec{\nabla}\times\vec{v})+2(\vec{v}\circ\nabla)\vec{v} \tag{9.7.345}$$

Therefore:

$$(\vec{v}\circ\vec{\nabla})\vec{v}=\frac{1}{2}\vec{\nabla}v^2-\vec{v}\times(\vec{\nabla}\times\vec{v}) \tag{9.7.346}$$

Notice that we can therefore rewrite the acceleration (material derivative):

$$\frac{D\vec{v}}{Dt}=\frac{\partial\vec{v}}{\partial t}+\vec{v}\vec{\nabla}\circ\vec{v} \tag{9.7.347}$$

as:

$$\frac{D\vec{v}}{Dt}=\frac{\partial\vec{v}}{\partial t}+\frac{1}{2}\vec{\nabla}v^2-\vec{v}\times(\vec{\nabla}\times\vec{v}) \tag{9.7.348}$$

as we will see and explain later the rotational of the vector speed is related to the concept of "vorticity".

Finally, we get a new equation named the "Euler equation of the 2nd form" which is written:

$$\boxed{\rho\partial_t\vec{v}=\rho\vec{f}-\vec{\nabla}p-\frac{\rho}{2}\vec{\nabla}v^2+\rho\vec{v}\times(\vec{\nabla}\times\vec{v})} \tag{9.7.349}$$

Also sometimes rearranged as:

$$\rho\partial_t\vec{v}+\rho\left(\frac{1}{2}\vec{\nabla}v^2-\vec{v}\times(\vec{\nabla}\times\vec{v})\right)-\rho\vec{f}=\vec{0} \tag{9.7.350}$$

Although the two Euler equations of are very important, there exists a variation form very useful in meteorology which we will now determine.

We always rely on the flow of an incompressible and non-viscous fluid, but whose volumic forces are derived this time from a potential $\vec{f}=-\vec{\nabla}U$ where U is typicall the gravitional potentiel for example.

In this case, we use the Euler equation in its first form:

$$\rho\mathrm{d}_t\vec{v}=-\vec{\nabla}p+\rho\vec{f} \tag{9.7.351}$$

Since the volumic forces $\vec{f}$ derive from a potential U, we have:

$$\rho\mathrm{d}_t\vec{v}=-\vec{\nabla}(p+\rho U) \tag{9.7.352}$$

Let us recall the relation (see section Vector Calculus page 1911):

$$(\vec{v}\circ\vec{\nabla})\vec{v}=\frac{1}{2}\vec{\nabla}v^2-\vec{v}\times(\vec{\nabla}\times\vec{v}) \tag{9.7.353}$$

Given $\vec{\nabla}\times\vec{v}$ be a vector $\vec{A}$, it comes:

$$-\vec{v}\times\vec{A}=\vec{A}\times\vec{v} \tag{9.7.354}$$

Therefore:

$$-\vec{v}\times(\vec{\nabla}\times\vec{v})=(\vec{\nabla}\times\vec{v})\times\vec{v} \tag{9.7.355}$$

so we can also write:

$$(\vec{v}\circ\vec{\nabla})\vec{v}=\frac{1}{2}\vec{\nabla}v^2+(\vec{\nabla}\times\vec{v})\times\vec{v} \tag{9.7.356}$$

By taking up the material derivative[34]:

$$\mathrm{d}_t\vec{v}=\partial_t\vec{v}+(\vec{v}\circ\vec{\nabla})\vec{v} \tag{9.7.357}$$

the equation:

$$\rho\mathrm{d}_t\vec{v}=-\vec{\nabla}(p+\rho U) \tag{9.7.358}$$

then becomes:

$$\rho\mathrm{d}_t\vec{v}=\rho\left[\partial_t\vec{v}+(\vec{v}\circ\vec{\nabla})\vec{v}\right]=-\vec{\nabla}(p+\rho U) \tag{9.7.359}$$

and using:

$$(\vec{v}\circ\vec{\nabla})\vec{v}=\frac{1}{2}\vec{\nabla}v^2+(\vec{\nabla}\times\vec{v})\times\vec{v} \tag{9.7.360}$$

the latter becomes:

$$\rho\partial_t\vec{v}+\rho\left[\frac{1}{2}v^2+(\vec{\nabla}\times\vec{v})\times\vec{v}\right]=-\vec{\nabla}(p+\rho U) \tag{9.7.361}$$

and since:

$$\frac{1}{2}\vec{\nabla}v^2=\vec{\nabla}\left(\frac{v^2}{2}\right) \tag{9.7.362}$$

we can finally write:

$$\rho\partial_t\vec{v}+\rho\left[\frac{1}{2}\vec{\nabla}v^2+(\vec{\nabla}\times\vec{v})\times\vec{v}\right]+\vec{\nabla}(p+\rho U)=\vec{0} \tag{9.7.363}$$

Rearranging and simplifying, we get:

$$\partial_t\vec{v}+\vec{\nabla}\left[\frac{v^2}{2}+\frac{p}{\rho}+U\right]+(\vec{\nabla}\times\vec{v})\times\vec{v}=\vec{0} \tag{9.7.364}$$

✍**Example:**

E1. Let us consider a bucket filled with a Newtonian fluid rotating at an angular velocity ω.

---

[34] Also sometimes, for recall, denoted $\frac{\mathrm{D}\vec{v}}{\mathrm{D}t}=\partial_t\vec{v}+(\vec{v}\circ\vec{\nabla})\vec{v}$

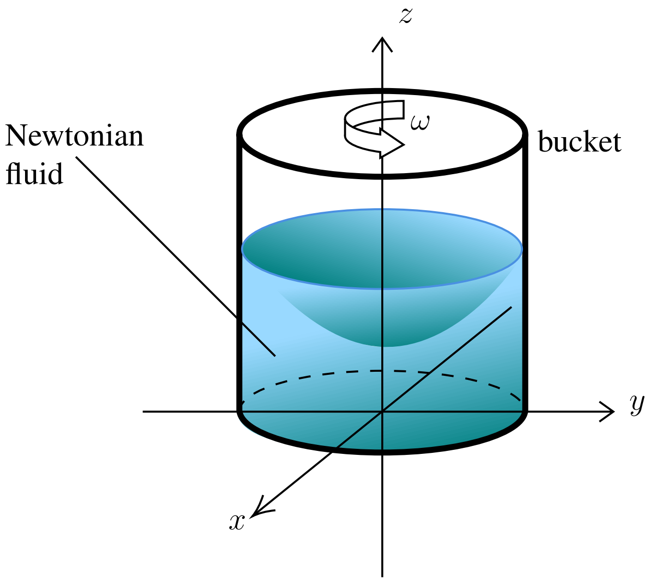


Figure 9.245: Rotating bucket

Given incompressibility assumption and steady flow !

Consider Euler's equation in a uniform constant gravitational field (see page 3830):

$$\frac{\partial \vec{v}}{\partial t}+(\vec{v}\circ\vec{\nabla})\vec{v}=-\vec{\nabla}\frac{p}{\rho}+\vec{g} \tag{9.7.365}$$

As the flow is steady (i.e. $\partial\vec{v}/\partial t=\vec{0}$), the latter reduces to:

$$(\vec{v}\circ\vec{\nabla})\vec{v}=-\vec{\nabla}\frac{p}{\rho}+\vec{g} \tag{9.7.366}$$

As the fluid is rotating circularly following the axis of symmetry z of the configuration we have:

$$\begin{aligned} v_x&=y\omega\\ v_y&=x\omega\\ v_z&=0 \end{aligned} \tag{9.7.367}$$

Where we are assuming obvioulsy the z-axis to be the axis of rotation. Hence:

$$(\vec{v}\circ\vec{\nabla})\vec{v}=\left(\begin{pmatrix} x\omega\\ y\omega\\ 0\end{pmatrix}\circ\begin{pmatrix}\partial_x\\ \partial_y\\ \partial_z\end{pmatrix}\right)\begin{pmatrix} x\omega\\ y\omega\\ 0\end{pmatrix}=\begin{pmatrix} x\omega^2\\ y\omega^2\\ 0\end{pmatrix} \tag{9.7.368}$$

Upon substitution of each component individually into Euler's equation above, we arrive at the centripetal acceleration in terms of surface pressure (assuming an inertial frame):

$$\begin{aligned} x\omega^2&=-\frac{1}{\rho}\frac{\partial p}{\partial x}\\ y\omega^2&=-\frac{1}{\rho}\frac{\partial p}{\partial y}\\ 0&=-\frac{1}{\rho}\frac{\partial p}{\partial z}+g \end{aligned} \tag{9.7.369}$$

Now, we evaluate the integral of these expressions:

$$\begin{aligned}\frac{x^2\omega^2}{2}&=\frac{p}{\rho}\\ \frac{y^2\omega^2}{2}&=\frac{p}{\rho}\\ \frac{p}{\rho}&=-gz\end{aligned} \tag{9.7.370}$$

We know that the general solution is the sum of all particular solution. We then have:

$$\frac{p}{\rho}=\frac{\omega^2\left(x^2+y^2\right)}{2}-gz+c^{te} \tag{9.7.371}$$

Indeed we can check that's correct:

$$\begin{aligned}x\omega^2&=-\frac{1}{\rho}\frac{\partial p}{\partial x}=-\frac{1}{\rho}\frac{\partial\left(\rho\frac{\omega^2\left(x^2+y^2\right)}{2}-\rho gz+c^{te}\right)}{\partial x}\\ y\omega^2&=-\frac{1}{\rho}\frac{\partial\left(\rho\frac{\omega^2\left(x^2+y^2\right)}{2}-\rho gz+c^{te}\right)}{\partial y}\\ 0&=-\frac{1}{\rho}\frac{\partial\left(\rho\frac{\omega^2\left(x^2+y^2\right)}{2}-\rho gz+c^{te}\right)}{\partial z}+g\end{aligned} \tag{9.7.372}$$

Given $p=$ constant at a free surface (we can then choose $c^{te}=p/\rho$ so that they cancel), i.e. at the interface between the liquid and the gas, and solving for z:

$$\boxed{z=\frac{\omega^2\left(x^2+y^2\right)}{2g}} \tag{9.7.373}$$

The surface adopts the shape of a paraboloid! This is the proof water can be curved by a force like the centripetal acceleration (like the water/seas/ocean can be curved by gravity)!

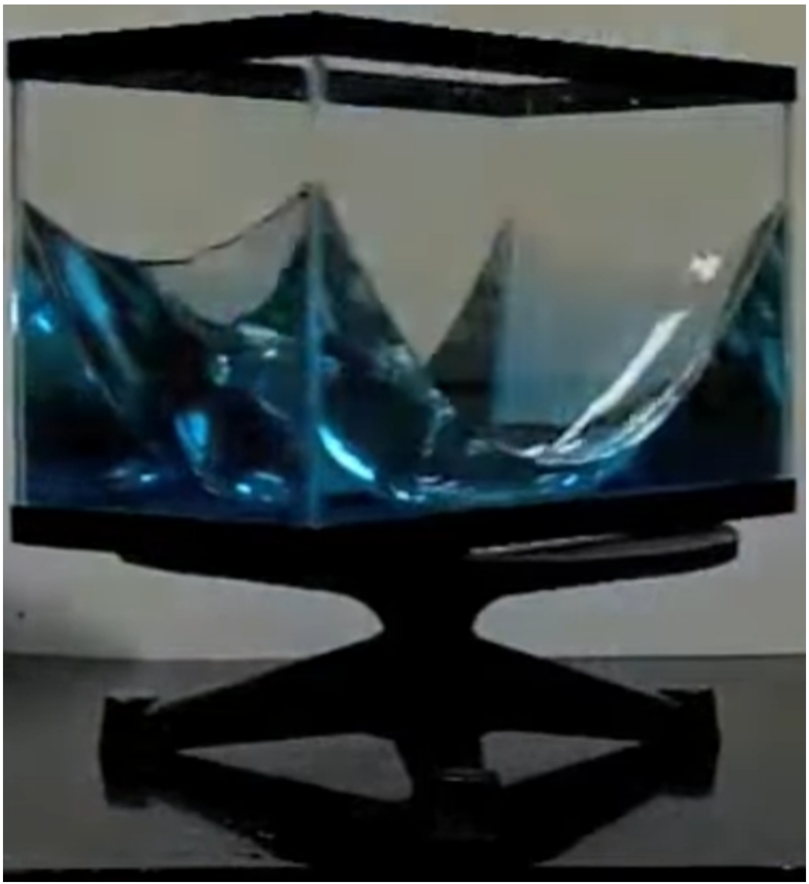
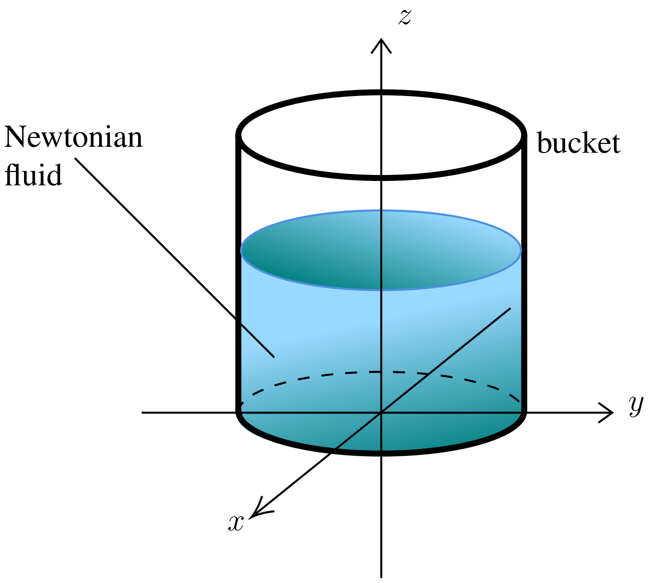


Figure 9.246: **Rotating bucket experiment (source: YouTube, author: Dan Russell)**

Notice that if $\omega=0$ we have $z=0$ for all (x,y), that means the liquid is obviously at the same level and therefore flat:



Let us now denote z_0 the height between the bottom of the parabola and the bottom of the bucket and let us denote $x^2+y^2=r^2$ the radius distance from the vertical axis of rotation of the bucket.

Therefore:

$$z=z_0+\frac{\omega^2r^2}{2g} \tag{9.7.374}$$

Now let us denote H the height of the fluid at rest (when $\omega=0$). Therefore obviously the total mass of the fluid in the cylinder is simply given by:

$$m=\rho\pi R^2H \tag{9.7.375}$$

where R is the radius of the bucket. When $\omega\neq 0$ we have using the shell method:

$$\begin{aligned}m&=\rho\int\limits_0^R(2\pi r)z\mathrm{d}r=\rho\int\limits_0^R(2\pi r)\left[z_0+\frac{\omega^2r^2}{2g}\right]\mathrm{d}r\\&=\rho(2\pi)\int\limits_0^R\left(z_0r+\frac{\omega^2r^3}{2g}\right)\mathrm{d}r=2\pi\rho\left(\frac{z_0R^2}{2}+\frac{\omega^2R^4}{8g}\right)\end{aligned} \tag{9.7.376}$$

Equating the two masses (the non-spinning case and the spinning case), we get:

$$m=\rho\pi R^2H=2\pi\rho\left(\frac{z_0R^2}{2}+\frac{\omega^2R^4}{8g}\right) \tag{9.7.377}$$

Hence after simplification and rearrangement:

$$H=z_0+\frac{\omega^2R^2}{4g} \tag{9.7.378}$$

So finally at the center of the container:

$$z_0=H-\frac{\omega^2R^2}{4g} \tag{9.7.379}$$

We can seek for the maximum ω such that $z_0=0$. Therefore we get immediately:

$$\omega_{\text{max}}=\frac{2}{R}\sqrt{gH} \tag{9.7.380}$$

9.7.2.4.1.1 Vorticity and vorticity equation

Let us generalize this last relation by making emerge possible rotations. For this we know that (see section Classical Mechancis page 3280):

$$\vec{v}=\vec{\omega}\times\vec{r} \tag{9.7.381}$$

therefore:

$$\vec{\nabla}\times\vec{v}=\vec{\nabla}\times(\vec{\omega}\times\vec{r}) \tag{9.7.382}$$

By writing the vector product $\vec{\omega}\times\vec{r}$ in developed form, we have:

$$\vec{\nabla}\times\left[(\omega_y z-\omega_z y)\vec{i}+(\omega_z x-\omega_x z)\vec{j}+(\omega_x y-\omega_y k)\vec{k}\right] \tag{9.7.383}$$

Which gives:

$$\begin{aligned}\vec{\nabla}\times\vec{v}=&\,\partial_y(\omega_x y-\omega_y x)\vec{i}+\partial_x(\omega_z x-\omega_x z)\vec{k}+\partial_z(\omega_y z-\omega_z y)\vec{j}\\&-\partial_y(\omega_y z-\omega_z y)\vec{k}-\partial_z(\omega_z x-\omega_x z)\vec{i}-\partial_x(\omega_x y-\omega_y x)\vec{j}\end{aligned} \tag{9.7.384}$$

Let us suppose that $\vec{\omega}$ is a constant angular velocity vector, we then have:

$$\vec{\nabla}\times\vec{v}=\omega_x\vec{i}+\omega_z\vec{k}+\omega_y\vec{j}+\omega_z\vec{k}+\omega_x\vec{i}+\omega_y\vec{j}=2(\omega_x\vec{i}+\omega_y\vec{j}+\omega_z\vec{k})=2\vec{\omega} \tag{9.7.385}$$

Definition (#9.7.474): We say that we have a "vortex flow" if the rotation of the velocity field is non-zero:

$$\vec{\nabla}\times\vec{v}\neq\vec{0} \tag{9.7.386}$$

everywhere or at certain points. We also define from the prior-previous relationship the "vorticity" by:

$$\boxed{\vec{\omega}=\frac{1}{2}\vec{\nabla}\times\vec{v}} \tag{9.7.387}$$

Notice that in the special case of a 2-D flow:

$$\vec{u}(x,y,t)=\begin{pmatrix}u(x,y,t)\\v(x,y,t)\\0\end{pmatrix} \tag{9.7.388}$$

the vorticity:

$$\vec{\omega}=\vec{\nabla}\times\vec{u}=\begin{pmatrix}0\\0\\\frac{\partial v}{\partial x}-\frac{\partial u}{\partial y}\end{pmatrix} \tag{9.7.389}$$

is perpendicular to the plane of motion, its magnitude is:

$$\omega=\frac{\partial v}{\partial x}-\frac{\partial u}{\partial y} \tag{9.7.390}$$

Caution! At this point, we have a small problem with conventions... In a huge majority of textbooks (and even Wikipedia) the vorticity is not derived as we did just above but rather just defined! And the most common definition of vorticity is sadly then given by $\vec{\omega}=\vec{\nabla}\times\vec{v}$. So therefore the consequence is that obviously the vorticity is equal twice the value of angular velocity and many textbooks authors have problems explaining this difference of a factor 2. Even worst... all related relation that we will derived later below will always have multiple times a factor 2 that differs with the huge majority of textbooks. That's the problem when there are no ISO standards for everything in physics...

Notice that we have therefore:

$$(\vec{v}\circ\vec{\nabla})\vec{v}=\frac{1}{2}\vec{\nabla}v^2+(\vec{\nabla}\times\vec{v})\times\vec{v}\Leftrightarrow(\vec{v}\circ\vec{\nabla})\vec{v}=\frac{1}{2}\vec{\nabla}v^2+2\vec{\omega}\times\vec{v} \qquad (9.7.391)$$

Example of partially vortex flow simulation (at some points):

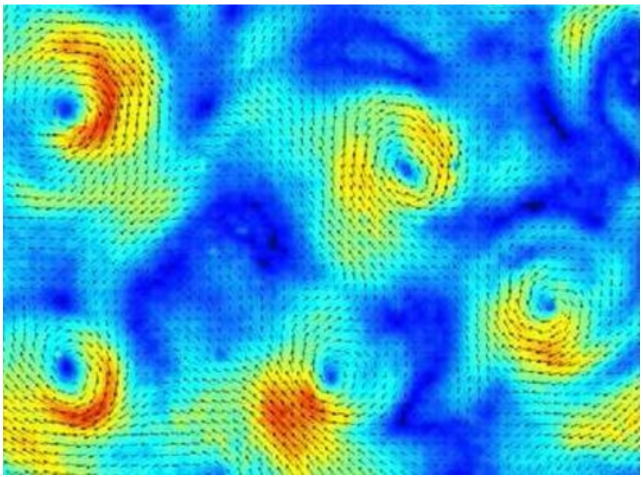


Figure 9.247: Example of a simulated vortex flow

and in real life at large scale levels[35] (Jupiter has also a famous one...):

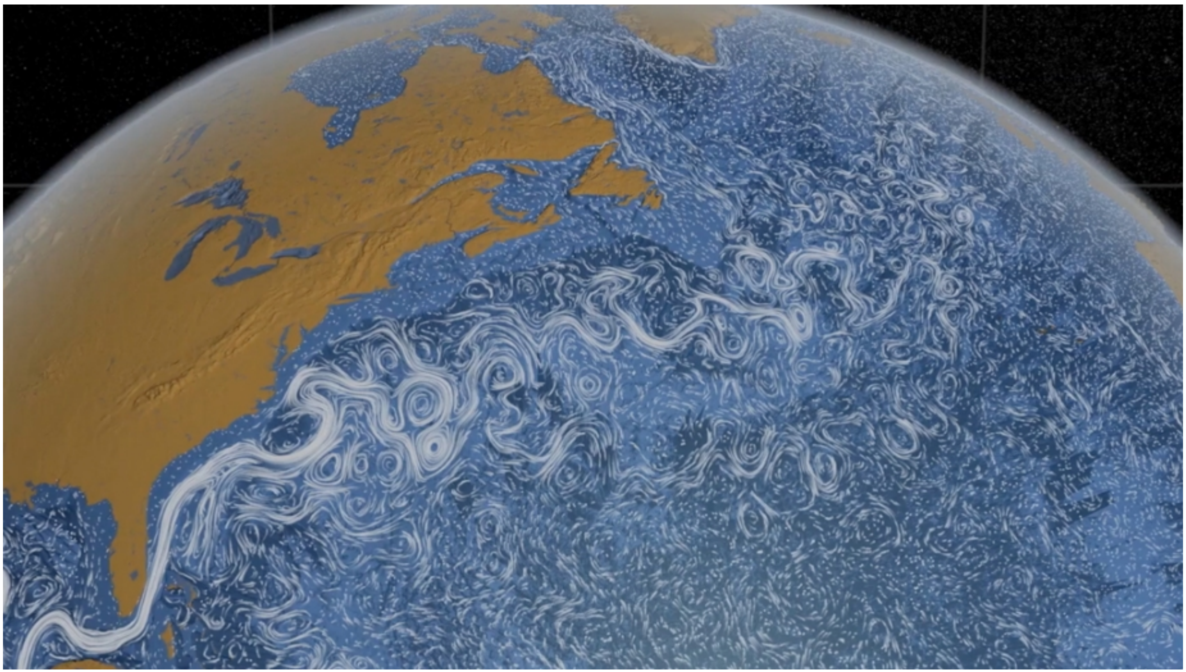


Figure 9.248: Earth currents and vortex between June 12005 to December 12007 (source: NASA/SVS)

---

[35] This a rebuild pic of ocean surface currents between June 12005 (holocene calendar) to December 12007 (holocene calendar) from NASA satellites on our planet. The Pacific carry warm waters across thousands of kilometers at speeds greater than six kilometers per hour; coastal currents like the Agulhas in the Southern Hemisphere move equatorial waters toward Earth's poles; and how thousands of other ocean currents are confined to particular regions and form slow-moving, circular pools.

✍**Examples:**

E1. Free vortex flow is one in which the fluid mass rotates without any external impressed contact force. The whole fluid mass rotates either due to fluid pressure itself or the gravity or due to rotation previously imparted. The free vortex motion is also named "potential vortex" or "irrotational vortex".

Typical well known practical cases are: flow around a circular bend, a whirlpool in a river, flow of liquid in a centrifugal pump casing after it has left the impeller, flow of water in a turbine casing before it enters the guide vanes, flow of liquid through a hole/outlet provided at the bottom of a shallow vessel (e.g., wash basin, bath tub, etc.).

In free vortex the relation between velocity and radius is obtained by putting the value of external torque equal to zero, or, the time rate change of angular momentum (i.e., moment of momentum) must be zero! That is:

$$\frac{\partial}{\partial t}(mvr)=0$$

Integrating, we get:

$$mvr=c^{te}$$

Since m is constant:

$$mvr=C^{te}$$

where C^{te} is a constant and is known as "strength of vortex". Therefore:

$$v=\frac{C^{te}}{r}\propto\frac{1}{r}$$

ie tangential velocity is inversely proportional to distance r.

Now, if we look at a rotating flow in cylindrical coordinates we know from our study of Differential Operators that:

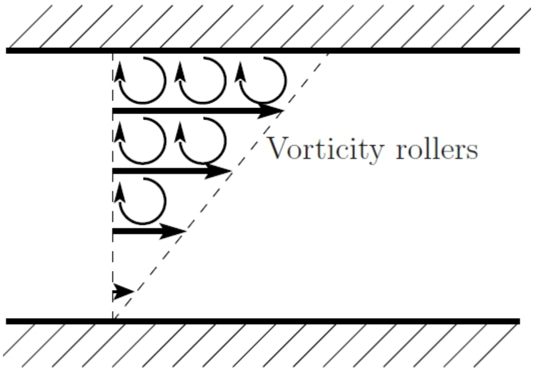
$$\vec{\nabla}\times\vec{v}=\frac{1}{r}\frac{\partial(rv_\theta)}{\partial r}$$

in case of a free vortex we have just seen that $v_\theta\propto 1/r$ and therefore the derivative in the above equation vanishes for a vortex centered at $r=0$. In other words, the vorticity is zero everywhere, i.e. $\vec{\omega}=\vec{0}$.

Although the flow is rotating globally, there is no local rotation. Crossed vanes placed in the flow would move in a circle, but without spinning.

E2. Let us consider again a 2-D flow in a pipe where the flow speed decrease from top to down in proportion to a given constant k:

$$\vec{u}(x,y,t)=\begin{pmatrix}u(x,y,t)\\v(x,y,t)\\0\end{pmatrix}=\begin{pmatrix}ky\\0\\0\end{pmatrix}$$



Therefore we have $\vec{\omega}=(0,0,-k)$. So in this case the vorticity is not a measure of global rotation: a shear flow has no global rotation but a non-zero vorticity!!!

The equation:

$$\partial_t\vec{v}+\vec{\nabla}\left[\frac{v^2}{2}+\frac{p}{\rho}+U\right]+(\vec{\nabla}\times\vec{v})\times\vec{v}=\vec{0} \tag{9.7.392}$$

is then written:

$$\boxed{\partial_t\vec{v}+\vec{\nabla}\left[\frac{v^2}{2}+\frac{p}{\rho}+U\right]+2\vec{\omega}\times\vec{v}=\vec{0}} \tag{9.7.393}$$

and sometimes named the "momentum balance equation". We find in this equation, used in meteorology, the Coriolis acceleration that we had determined in the section of Classical Mechanics page 3388!

If the flow is at constant velocity then:

$$\mathrm{d}_t\vec{v}=\vec{0} \tag{9.7.394}$$

and the fluid is then not rotational (non-turbulent):

$$\vec{\omega}\times\vec{v}=\vec{0} \tag{9.7.395}$$

then the above equation reduces to:

$$\vec{\nabla}\left[\frac{v^2}{2}+\frac{p}{\rho}+U\right]=\vec{0} \tag{9.7.396}$$

In classical dynamics of the rigid material point (see section Classical Mechanics page 3351), we have shown that in the case of a terrestrial gravitational potential the "geopotential" was given by:

$$U=g(z-z_0)\;[\mathrm{J}\cdot\mathrm{kg}^{-1}] \tag{9.7.397}$$

z being the altitude of a point of the fluid with respect to an reference level z_0. If we take z_0 for the ground level, the prior-previous relation becomes then in the case of a flow then say to be "potential flow":

$$\vec{\nabla}\left[\frac{v^2}{2}+\frac{p}{\rho}+gz\right]=\vec{0} \tag{9.7.398}$$

The term in brackets to satisfy this relationship must be such that:

$$\frac{1}{2}\rho v^2+\rho gz+p=c^{te} \quad (9.7.399)$$

We thus fall back well on Bernoulli's theorem, which confirms our model of Newtonian fluids according to the Navier-Stokes model.

If the flow is inversely a vortex flow, such that:

$$\vec{\nabla}\times\vec{v}=\vec{0} \quad (9.7.400)$$

Now, as we have proved in the section of Vector Calculus, if the rotational of a vector variable is zero, the variable can be expressed as the gradient of a scalar potential! Consequently, in the case of the vortex, the velocity field then derives from a velocity potential:

$$\vec{v}=\vec{\nabla}\varphi \quad (9.7.401)$$

And if in addition the fluid is incompressible, as we saw above:

$$\vec{\nabla}\circ\vec{v}=0 \quad (9.7.402)$$

Then by combining the last two relations it comes (divergence of the gradient):

$$\vec{\nabla}\circ(\vec{\nabla}\circ\vec{v})=0 \quad (9.7.403)$$

Thus explicitly:

$$\frac{\partial^2\varphi}{\partial x^2}+\frac{\partial^2\varphi}{\partial y^2}+\frac{\partial^2\varphi}{\partial z^2}=0 \quad (9.7.404)$$

The potential of the velocities in the case of an uncompressible and non-vortex liquid thus verifies the Laplace's equation:

$$\nabla^2\varphi=0 \qquad \text{or} \qquad \Delta\varphi=0 \quad (9.7.405)$$

Remember that if we didn't neglect the viscosity μ_d we would simply have:

$$\rho\left[\partial_t\vec{v}+(\vec{v}\circ\vec{\nabla})\vec{v}\right]=\rho\vec{f}+\mu_d\vec{\nabla}^2\vec{v}-\vec{\nabla}p \quad (9.7.406)$$

and using:

$$(\vec{v}\circ\vec{\nabla})\vec{v}=\frac{1}{2}\vec{\nabla}v^2+2\vec{\omega}\times\vec{v} \quad (9.7.407)$$

This lead us then obviously to:

$$\rho\left[\partial_t\vec{v}+\frac{1}{2}\vec{\nabla}v^2+2\vec{\omega}\times\vec{v}\right]=\rho\vec{f}+\mu_d\vec{\nabla}^2\vec{v}-\vec{\nabla}p \quad (9.7.408)$$

Often rearranged as following:

$$\partial_t\vec{v}-2\vec{v}\times\vec{\omega}=-\vec{\nabla}\left(\frac{1}{2}v^2+\frac{p}{\rho}+\vec{f}\right)+\mu_d\vec{\nabla}^2\vec{v} \quad (9.7.409)$$

and considering a gravitational potential:

$$\partial_t \vec{v} - 2\vec{v} \times \vec{\omega} = -\vec{\nabla}\left(\frac{1}{2}v^2 + \frac{p}{\rho} + g\vec{z}\right) + \mu_d \vec{\nabla}^2 \vec{v} \tag{9.7.410}$$

In the relation above, between the parenthesis, we see the Bernoulli terms. We can now also use this equation to derive Bernoulli's equation and to state the conditions under which it applies!

Suppose then the following conditions:

$$\begin{aligned} \rho &= c^{te} && \text{incompressible fluid} \\ \partial_t \vec{v} &= \vec{0} && \text{steady flow} \\ \mu_d \vec{\nabla}^2 v &= \vec{0} && \text{neglect viscosity} \end{aligned} \tag{9.7.411}$$

Under these conditions the equation above simplifies to:

$$2\vec{v} \times \vec{\omega} = \vec{\nabla}\left(\frac{1}{2}v^2 + \frac{p}{\rho} + g\vec{z}\right) \tag{9.7.412}$$

In the special case where $\vec{\omega} = \frac{1}{2}\vec{\nabla} \times \vec{v}$ is zero everywhere, that is the fluid is not rotating, we have:

$$\vec{0} = \vec{\nabla}\left(\frac{1}{2}v^2 + \frac{p}{\rho} + gz\right) \tag{9.7.413}$$

Therefore:

$$\frac{1}{2}v^2 + \frac{p}{\rho} + gz = c^{te} \tag{9.7.414}$$

and we fall back on Bernoulli's equation!

Notice that $\vec{v} \times \vec{\omega}$, being a cross product, it is always perpendicular to $\vec{v}$, i.e., always zero in the direction of $\vec{v}$. That means we can use Bernoulli's equation only along a streamline!

Now let's go back to (remember $\vec{f} = -\vec{\nabla}U$):

$$\partial_t \vec{v} - 2\vec{v} \times \vec{\omega} = -\vec{\nabla}\left(\frac{1}{2}v^2 + \frac{p}{\rho} + U\right) + \mu_d \vec{\nabla}^2 \vec{v} \tag{9.7.415}$$

to see how the viscosity term may be influenced by $\vec{\omega}$.

For that purpose, remember that we have proved in the section of Vector Calculus at page 1914, that:

$$\vec{\nabla} \times (\vec{\nabla} \times \vec{v}) = \vec{\nabla}(\vec{\nabla} \circ \vec{v}) - \Delta \vec{v} = \vec{\nabla}(\vec{\nabla} \circ \vec{v}) - \vec{\nabla}^2 \vec{v} \tag{9.7.416}$$

Hence:

$$\vec{\nabla}^2 \vec{v} = \vec{\nabla}(\vec{\nabla} \circ \vec{v}) - \vec{\nabla} \times (\vec{\nabla} \times \vec{v}) \tag{9.7.417}$$

If we assume the fluid velocity non divergent (i.e. $\vec{\nabla}\circ\vec{v}=0$), then we get obviously:

$$\vec{\nabla}^2\vec{v}=-\vec{\nabla}\times(\vec{\nabla}\times\vec{v}) \qquad (9.7.418)$$

Using the vorticity relation proven earlier above:

$$\vec{\omega}=\frac{1}{2}\vec{\nabla}\times\vec{v} \qquad (9.7.419)$$

Therefore:

$$\vec{\nabla}^2\vec{v}=-2\vec{\nabla}\times\vec{\omega} \qquad (9.7.420)$$

Hence:

$$\mu_d\vec{\nabla}^2\vec{v}=-2\mu_d\vec{\nabla}\times\vec{\omega} \qquad (9.7.421)$$

From this relation we see that there are no viscous forces where the vorticity $\vec{\omega}$ is zero, or even when $\vec{\omega}$ is constant as in solid body rotation!

With these assumptions and relations we can rewrite:

$$\partial_t\vec{v}-2\vec{v}\times\vec{\omega}=-\vec{\nabla}\left(\frac{1}{2}v^2+\frac{p}{\rho}+U\right)+\mu_d\vec{\nabla}^2\vec{v} \qquad (9.7.422)$$

as following:

$$\partial_t\vec{v}-2\vec{v}\times\vec{\omega}=-\vec{\nabla}\left(\frac{1}{2}v^2+\frac{p}{\rho}+U\right)-2\mu_d\vec{\nabla}\times\vec{\omega} \qquad (9.7.423)$$

Now let us take the curl of both sides of the relation above:

$$\vec{\nabla}\times[\partial_t\vec{v}-2\vec{v}\times\vec{\omega}]=\vec{\nabla}\times\left[-\vec{\nabla}\left(\frac{1}{2}v^2+\frac{p}{\rho}+U\right)-2\mu_d\vec{\nabla}\times\vec{\omega}\right] \qquad (9.7.424)$$

Let us recall that the curl of a gradient is identically zero (vector identity: $\vec{\nabla}\times\vec{\nabla}\cdot U=0$ as proved in the section of Vector Calculus page 1912). Thus the Bernoulli terms all gout out in the above relation:

$$\vec{\nabla}\times\vec{\nabla}\left(\frac{1}{2}v^2+\frac{p}{\rho}+U\right)=0 \qquad (9.7.425)$$

Therefore it reduces to:

$$\vec{\nabla}\times[\partial_t\vec{v}-2\vec{v}\times\vec{\omega}]=-2\mu_d\vec{\nabla}\times\vec{\nabla}\times\vec{\omega} \qquad (9.7.426)$$

Next, we note that because we can interchange the order of partial differentiation, we get:

$$\vec{\nabla}\times\frac{\partial\vec{v}}{\partial t}=\frac{\partial}{\partial t}\left(\vec{\nabla}\times\vec{v}\right)=2\frac{\partial\vec{\omega}}{\partial t} \qquad (9.7.427)$$

Putting it back in the prior-previous relation and simplifying in both sides by the factor 2 we get the "vorticity equation":

$$\boxed{\partial_t\vec{\omega}+\vec{\nabla}\times(\vec{\omega}\times\vec{v})=-\mu_d\vec{\nabla}\times(\vec{\nabla}\times\vec{\omega})} \qquad (9.7.428)$$

This equation is the differential equation for the dynamical behaviour of the vorticity field $\vec{\omega}$. The only restriction is that it applies to constant density fluids. If we wish to work with non constant density fluids we have to go back and work and use a more general formula for the viscous force.

We notice that if $\vec{\omega}=\vec{0}$ everywhere initially, then $\vec{\omega}$ remains zero. Thus, flows that start off irrotational remain so.

Remark

An important simplification we obtained in going to an equation for the vorticity field $\vec{\omega}$ was the elimination of the Bernoulli terms. This removes the pressure and gravitational forces from the equation for $\vec{\omega}$, implying that pressure and gravity have no direct effect on the behaviour of vorticity. We see this result in the case of the motion of a smoke ring. The ring moves in a straight line across a room completely unaffected by gravity (pressure and gravity can have an indirect effect in that they affect the velocity field $\vec{v}$ which appears in the $\vec{\nabla}\times\vec{\omega}\times\vec{v}$ term).

The latter relation can be found in another very common shape if we know the identity:

$$\vec{\nabla}\times\left(\vec{f}\times\vec{g}\right)=\left(\vec{g}\circ\vec{\nabla}\right)\vec{f}-\vec{g}\left(\vec{\nabla}\circ\vec{f}\right)-\left(\vec{f}\circ\vec{\nabla}\right)\vec{g}+\vec{f}\left(\vec{\nabla}\circ\vec{g}\right) \tag{9.7.429}$$

already proved in the section of Vector Calculus page 1912 and we neglect μ_d. Therefore:

$$\partial_t\vec{\omega}+\vec{\nabla}\times(\vec{\omega}\times\vec{v})=\vec{0} \tag{9.7.430}$$

Becomes:

$$\partial_t\vec{\omega}+\left(\vec{v}\circ\vec{\nabla}\right)\vec{\omega}-\vec{v}\left(\vec{\nabla}\circ\vec{\omega}\right)-\left(\vec{\omega}\circ\vec{\nabla}\right)\vec{v}+\vec{\omega}\left(\vec{\nabla}\circ\vec{v}\right)=\vec{0} \tag{9.7.431}$$

For incompressible flows $\vec{\nabla}\circ\vec{v}=0$ and, in addition, $\vec{\nabla}\circ\vec{\omega}=0$ as $\vec{\omega}=\frac{1}{2}\vec{\nabla}\times\vec{v}$. Hence:

$$\partial_t\vec{\omega}+\left(\vec{v}\circ\vec{\nabla}\right)\vec{\omega}-\left(\vec{\omega}\circ\vec{\nabla}\right)\vec{v}=\vec{0} \tag{9.7.432}$$

Rearranging we get the other famous shape of a special case of the vorticity equation:

$$\partial_t\vec{\omega}+\left(\vec{v}\circ\vec{\nabla}\right)\vec{\omega}=\left(\vec{\omega}\circ\vec{\nabla}\right)\vec{v} \tag{9.7.433}$$

And some specialists the introduce the "material derivative of the rotational" as:

$$\boxed{\frac{\mathrm{D}\tilde{\omega}}{\mathrm{D}t}=\partial_t\vec{\omega}+\left(\vec{v}\circ\vec{\nabla}\right)\vec{\omega}=\left(\vec{\omega}\circ\vec{\nabla}\right)\vec{v}} \tag{9.7.434}$$

In a two-dimensional planar flow, $\vec{v}=(u(x,y),v(x,y),0)$, the vector vorticity has only one non-zero component, $\vec{\omega}=(\partial v/\partial x-\partial u/\partial y)\hat{\vec{e}}_z$, so that:

$$(\vec{\omega}\circ\vec{\nabla})\vec{u}=\omega\frac{\mathrm{d}}{\mathrm{d}z}\vec{v}(x,y)=0 \tag{9.7.435}$$

Hence, the vorticity equation, reduces to:

$$\frac{\mathrm{D}\omega_z}{\mathrm{D}t}=\frac{\partial\omega_z}{\partial t}+(\vec{u}\circ\vec{\nabla})\omega_z=0 \tag{9.7.436}$$

shows that the vorticity of a fluid particle remains constant. If, in addition the flow is steady, $\partial\omega_z/\partial t=0$ then the vorticity is constant along streamlines.

9.7.2.4.1.2 Circulation

Consider a closed curve $\mathcal{C}$ in the flow. The circulation around $\mathcal{C}$ is the line integral of the tangential velocity around $\mathcal{C}$:

$$\Gamma=\oint_{\mathcal{C}}\vec{v}\circ\mathrm{d}\vec{l} \qquad (9.7.437)$$

By Stokes theorem, for any surface $\mathcal{S}$ spanning the curve $\mathcal{C}$:

$$\Gamma=\oint_{\mathcal{C}}\vec{v}\circ\mathrm{d}\vec{l}=\int_{\mathcal{S}}(\vec{\nabla}\times\vec{v})\circ\vec{n}\mathrm{d}S=\int_{\mathcal{S}}\omega\cdot\vec{n}\mathrm{d}S \qquad (9.7.438)$$

So, the circulation around $\mathcal{C}$ is equal to the vorticity flux through the surface $\mathcal{S}$: it is the strength of the vortex tube.

9.7.2.4.2 Compressible flow

In this case ρ_0 is a function of the pressure p (case of the "barotropic fluids"). We also consider that the viscosity is negligible. He then comes:

$$\frac{\vec{\nabla}p}{\rho}=\vec{\nabla}\int\frac{\mathrm{d}p}{\rho} \tag{9.7.439}$$

The equation:

$$\partial_t\vec{v}+\vec{\nabla}\left[\frac{v^2}{2}+\frac{p}{\rho}+U\right]+2\vec{\omega}\times\vec{v}=\vec{0} \tag{9.7.440}$$

is then written:

$$\boxed{\partial_t\vec{v}+\vec{\nabla}\left[\frac{v^2}{2}+\int\frac{\mathrm{d}p}{\rho}+U\right]+2\vec{\omega}\times\vec{v}=\vec{0}} \tag{9.7.441}$$

9.7.2.4.3 Static flow

The fluid on which the net force is zero is considered as a "static flow". Means the bulk of fluid moving with constant velocity would be a static fluid. This also includes the stationary fluid. i.e. zero velocity. Non ideal fluids (with viscocity and turbulence present) would never have a flow with constant bulk velocity due to presence of non zero shear forces. But in some cases, like the fluid in a tank moving with constant acceleration is considered as static case with the assumption that the internal motion within the fluid is negligible.

In the static case:

$$\vec{v}=\vec{0}\qquad\text{and}\qquad\mu_D=0 \tag{9.7.442}$$

and the equation:

$$\rho\mathrm{d}_t\vec{v}=-\vec{\nabla}p+\frac{4}{3}(\mu_d+\mu'_d)\vec{\nabla}(\vec{\nabla}\circ\vec{v})-\mu_d\vec{\nabla}\times(\vec{\nabla}\times\vec{v})+\rho\vec{f} \tag{9.7.443}$$

becomes simply:

$$\boxed{\vec{\nabla}p-\rho\vec{f}=\vec{0}} \tag{9.7.444}$$

which is the "equation of the static fluid" or the "fundamental law of the hydrostatic". In discrete and more classical (rearranged) form, we fall back on the relation that must textbooks provides and that we have also proved earlier above using another approach (gradient theorem):

$$p_2-p_1=\rho g(h_2-h_1) \tag{9.7.445}$$

or by taking the initial height as being equal to zero:

$$p_2=p_1+\rho g h_2 \tag{9.7.446}$$

Remark

As the viscosities disappear, the static fluids are the same for viscous or non-viscous fluids.

9.7.2.4.4 Reynolds number

The "Reynolds number (Re)" is an important dimensionless quantity in fluid mechanics that is used to help predict flow patterns in different fluid flow situations. It is widely used in many applications ranging from liquid flow in a pipe to the passage of air over an aircraft wing. The Reynolds Number is valuable as a guide to the transition point from laminar to turbulent flow in a particular flow situation, and for the scaling of similar but different-sized flow situations, such as between an aircraft model in a wind tunnel and the full size version.

Let us consider first, for simplicity, the incompressible case.

The equation of continuity, or conservation of mass, (see section Thermodynamics page 3683) that is for recall:

$$\rho\frac{\partial\rho}{\partial t}-\vec{\nabla}\circ(\rho\vec{v}_a)=0 \tag{9.7.447}$$

is then written in this particular case:

$$\rho\vec{\nabla}\circ\vec{v}_a=0\Rightarrow\vec{\nabla}\circ\vec{v}_a=\frac{\partial v_x}{\partial_x}+\frac{\partial v_y}{\partial_y}+\frac{\partial v_z}{\partial_z}=0 \tag{9.7.448}$$

We now choose several quantities of dimensionless references denoted by an index r such that:

$$\begin{gathered}
x'=\frac{x}{L_r},y'=\frac{y}{L_r},z'=\frac{z}{L_r} \qquad\qquad \rho'=\frac{\rho}{\rho_r},t'=\frac{t}{t_r}\\
\text{and}\\
v'_x=\frac{v_x}{v_r},v'_y=\frac{v_y}{v_r},v'_z=\frac{v_z}{v_r} \qquad\qquad f'_i=\frac{f_i}{f_r},p'=\frac{p}{p_r},\mu'_d=\frac{\mu_d}{\mu_r}
\end{gathered} \tag{9.7.449}$$

By these definitions, we have for example:

$$\frac{\partial v_x}{\partial x}=\frac{\partial(v_r v'_x)}{\partial(L_r x')}=\frac{v_r}{L_r}\frac{\partial v'_x}{\partial x'} \tag{9.7.450}$$

Therefore the equation of the deformations per unit of time becomes:

$$\rho\vec{\nabla}\circ\vec{v}=\frac{v_r}{L_r}\left(\frac{\partial v'_x}{\partial x'}+\frac{\partial v'_y}{\partial y'}+\frac{\partial v'_z}{\partial z'}\right)=0 \tag{9.7.451}$$

But we also have:

$$\rho\frac{\mathrm{d}\vec{v}}{\mathrm{d}t}=\frac{\partial\vec{v}}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t}+\frac{\partial\vec{v}}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t}+\frac{\partial\vec{v}}{\partial z}\frac{\mathrm{d}z}{\mathrm{d}t}+\frac{\partial\vec{v}}{\partial t} \tag{9.7.452}$$

Let us restrict ourselves to the study of only one component:

$$\rho\frac{\mathrm{d}v_x}{\mathrm{d}t}=\frac{\partial v_x}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t}+\frac{\partial v_x}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t}+\frac{\partial v_x}{\partial z}\frac{\mathrm{d}z}{\mathrm{d}t}+\frac{\partial v_x}{\partial t} \tag{9.7.453}$$

By multiplying this last relation by the density ρ and by definition of the speed we get:

$$\rho\frac{\mathrm{d}v_x}{\mathrm{d}t}=\rho\left[\frac{\partial v_x}{\partial_x}\frac{\mathrm{d}x}{\mathrm{d}t}+\frac{\partial v_x}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t}+\frac{\partial v_x}{\partial z}\frac{\mathrm{d}z}{\mathrm{d}t}+\frac{\partial v_x}{\partial t}\right]=\rho\left[\frac{\partial v_x}{\partial x}v_x+\frac{\partial v_x}{\partial y}v_y+\frac{\partial v_x}{\partial z}v_z+\frac{\partial v_x}{\partial t}\right] \tag{9.7.454}$$

Let us now take one of the possible formulations of the Navier-Stokes equation proved earlier above:

$$\rho\mathrm{d}_t v_x=-\partial_x p+\frac{1}{3}\mu_d\partial_x\partial_x(\vec{\nabla}\circ\vec{v})+\mu_d\vec{\nabla}\circ(\vec{\nabla}(v_x))+\rho f_x \tag{9.7.455}$$

By not forgetting that for an incompressible fluid we have:

$$\vec{\nabla}\circ\vec{v}=0 \tag{9.7.456}$$

the previous Navier-Stokes equation, reduces then to:

$$\rho\frac{\mathrm{d}v_x}{\mathrm{d}t}=-\frac{\partial p}{\partial x}+\mu_d\left(\frac{\partial}{\partial x}\frac{\partial}{\partial x}v_x+\frac{\partial}{\partial y}\frac{\partial}{\partial y}v_y+\frac{\partial}{\partial z}\frac{\partial}{\partial z}v_x\right)+\rho f_x \tag{9.7.457}$$

But, for a fluid we had assumed above that the shear tensions were given by:

$$\tau_{xy}=\mu_d\partial_x v_y \tag{9.7.458}$$

The terms where the viscosity coefficients appear can be rewritten such that:

$$\rho\frac{\mathrm{d}v_x}{\mathrm{d}t}=-\frac{\partial p}{\partial x}+\left(\frac{\partial\tau_{xx}}{\partial x}+\frac{\partial\tau_{yx}}{\partial y}+\frac{\partial\tau_{zx}}{\partial z}\right)+\rho f_x \tag{9.7.459}$$

Thus by correspondence:

$$\rho\left[\frac{\partial v_x}{\partial x}v_x+\frac{\partial v_x}{\partial y}v_y+\frac{\partial v_x}{\partial z}v_z+\frac{\partial v_x}{\partial t}\right]=-\frac{\partial p}{\partial x}+\left(\frac{\partial\tau_{xx}}{\partial x}+\frac{\partial\tau_{yx}}{\partial y}+\frac{\partial\tau_{zx}}{\partial z}\right)+\rho f_x \tag{9.7.460}$$

that we can write in even more condensed form by using a rather abusive tensor notation:

$$\rho\frac{\mathrm{d}v_x}{\mathrm{d}t}=-\frac{\partial p}{\partial x}+\frac{\partial\tau_{ij}}{\partial x_i}+\rho f_x \tag{9.7.461}$$

By introducing the dimensionless variables:

$$\rho'\rho_r\left[\frac{v_r^2}{L_r}\left(\frac{\partial v'_x}{\partial x'}v'_x+\frac{\partial v'_y}{\partial y'}v'_y+\frac{\partial v'_z}{\partial z'}v'_z\right)+\frac{v_r}{t_r}\frac{\partial v'_x}{\partial t'}\right]=-\frac{p_r}{L_r}\frac{\partial p'}{\partial x'}+\frac{v_r\mu_r}{L_r}\frac{1}{L_r}\frac{\tau'_{ij}}{\partial x'_i}+\rho_r\rho' f_r f'_x \tag{9.7.462}$$

Let us now multiply this last relation by $\dfrac{L_r}{v^2}$ and let us divide it by ρ_r on both sides of equality such that it becomes:

$$\rho'\left[\left(\frac{\partial v'_x}{\partial x'}v'_x+\frac{\partial v'_y}{\partial y'}v'_y+\frac{\partial v'_z}{\partial z'}v'_z\right)+\frac{L_r}{t_r v_r}\frac{\partial v'_x}{\partial t'}\right]=-\frac{p_r}{\rho_r v_r^2}\frac{\partial p'}{\partial x'}+\frac{\mu_r}{\rho_r L_r v_r}\frac{1}{L_r}\frac{\tau'_{ij}}{\partial x'_i}+\rho'\frac{L_r}{v_r^2}f_r f'_x \tag{9.7.463}$$

At the dimensional level, let us notice that we have:

$$t_r=\frac{L_r}{v_r}\qquad \frac{p_r}{\rho_r v_r^2}=1\qquad \frac{L_r}{v_r^2}f_r=1 \tag{9.7.464}$$

Finally:

$$\boxed{\rho'\left[\left(\frac{\partial v'_x}{\partial x'}v'_x+\frac{\partial v'_y}{\partial y'}v'_y+\frac{\partial v'_z}{\partial z'}v'_z\right)+\frac{\partial v'_x}{\partial t'}\right]=-\frac{\partial p'}{\partial x'}+\frac{1}{\text{Re}}\frac{\tau'_{ij}}{\partial x'_i}+\rho' f'_x} \tag{9.7.465}$$

This differential equation expressed in relative and dimensionless variables is named the "dimensionless Navier-Stokes-Reynolds equation".

The term Re, as we know is the Reynolds number, symbolically represented the ratio of the inertial forces to the viscous forces:

$$\boxed{\text{Re}=\frac{\rho_r L_r v_r}{\mu_r}=\frac{L_r v_r}{\mu_{rc}}} \tag{9.7.466}$$

where μ_{rc} is the "relative kinematic viscosity". The dynamic viscosity is therefore a term inversely proportional to the value of the Reynolds number.

We also often find this last relation in the following form (term-to-term identification with the first equation of the previous expression where $2R$ is the diameter of the circular section of the tube that transports the fluid):

$$N_{\text{Re}}=\frac{\rho 2R\bar{u}}{\mu} \tag{9.7.467}$$

If instead of the velocity it is the mass flow which is measured through a circular section, we then have using the following relation between the mass flow rate and fluid velocity in a cylindrical tube:

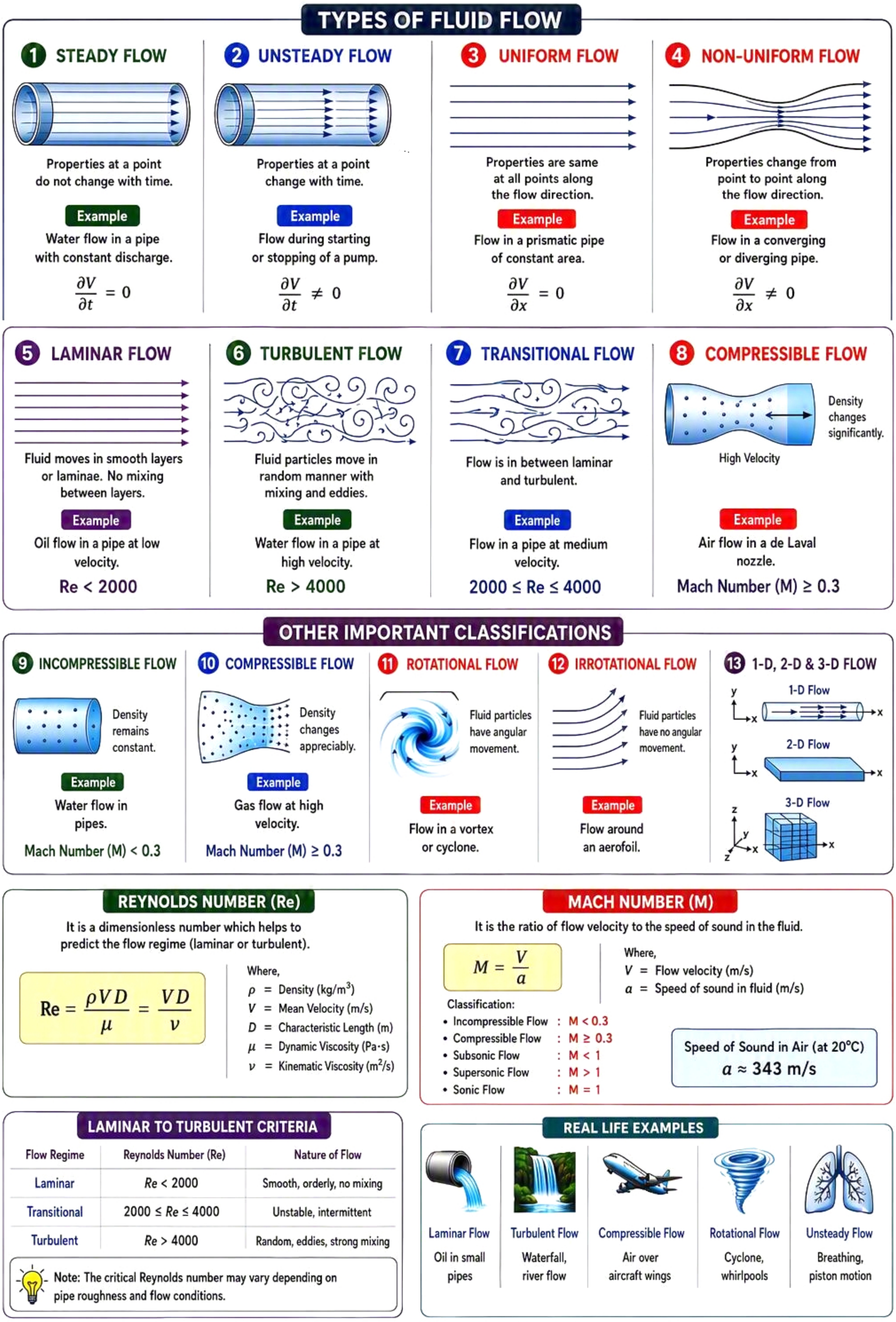
$$J=\frac{\Delta V}{\Delta t}=\frac{S\Delta d}{\Delta t}=Sv=\pi R^2 v\Rightarrow \dot{m}=J\rho=\pi R^2 v\rho\Rightarrow v=\frac{\dot{m}}{\pi R^2\rho} \tag{9.7.468}$$

Another writing for the prior-previous relation that we will often using in industrial practice:

$$N_{\text{Re}}=\frac{\rho 2R\left(\dfrac{\dot{m}}{\pi R^2\rho}\right)}{\mu}=\frac{2\dot{m}}{\mu\pi R}=\frac{2\dot{m}}{\mu\pi\dfrac{\varnothing}{2}}=\frac{4\dot{m}}{\mu\pi\varnothing} \tag{9.7.469}$$

In practice we regularly calculate the time it would take to fill a given volume tank with a fluid having a known mass flow rate, a known density, a known viscosity and a known number of Reynolds in a cylindrical pipe whose radii are also known. Therefore it comes:

$$t=\frac{V}{J}=\frac{V}{\dfrac{\Delta V}{\Delta t}}=\frac{V}{Sv}=\frac{V}{S\dfrac{\dot{m}}{\pi R^2\rho}}=\frac{V\rho}{\dot{m}}=\frac{4V\rho}{N_{\text{Re}}\mu\pi\varnothing} \tag{9.7.470}$$



TYPES OF FLUID FLOW

① STEADY FLOW

Properties at a point do not change with time.

Example

Water flow in a pipe with constant discharge.

$$\frac{\partial V}{\partial t} = 0$$

② UNSTEADY FLOW

Properties at a point change with time.

Example

Flow during starting or stopping of a pump.

$$\frac{\partial V}{\partial t} \neq 0$$

③ UNIFORM FLOW

Properties are same at all points along the flow direction.

Example

Flow in a prismatic pipe of constant area.

$$\frac{\partial V}{\partial x} = 0$$

④ NON-UNIFORM FLOW

Properties change from point to point along the flow direction.

Example

Flow in a converging or diverging pipe.

$$\frac{\partial V}{\partial x} \neq 0$$

⑤ LAMINAR FLOW

Fluid moves in smooth layers or laminae. No mixing between layers.

Example

Oil flow in a pipe at low velocity.

Re < 2000

⑥ TURBULENT FLOW

Fluid particles move in random manner with mixing and eddies.

Example

Water flow in a pipe at high velocity.

Re > 4000

⑦ TRANSITIONAL FLOW

Flow is in between laminar and turbulent.

Example

Flow in a pipe at medium velocity.

2000 ≤ Re ≤ 4000

⑧ COMPRESSIBLE FLOW

Density changes significantly.

Example

Air flow in a de Laval nozzle.

Mach Number (M) ≥ 0.3

OTHER IMPORTANT CLASSIFICATIONS

⑨ INCOMPRESSIBLE FLOW

Density remains constant.

Example

Water flow in pipes.

Mach Number (M) < 0.3

⑩ COMPRESSIBLE FLOW

Density changes appreciably.

Example

Gas flow at high velocity.

Mach Number (M) ≥ 0.3

⑪ ROTATIONAL FLOW

Fluid particles have angular movement.

Example

Flow in a vortex or cyclone.

⑫ IRROTATIONAL FLOW

Fluid particles have no angular movement.

Example

Flow around an aerofoil.

⑬ 1-D, 2-D & 3-D FLOW

REYNOLDS NUMBER (Re)

It is a dimensionless number which helps to predict the flow regime (laminar or turbulent).

$$\mathrm{Re} = \frac{\rho V D}{\mu} = \frac{V D}{\nu}$$

Where,

ρ = Density (kg/m^3)

V = Mean Velocity (m/s)

D = Characteristic Length (m)

μ = Dynamic Viscosity (Pa·s)

ν = Kinematic Viscosity (m^2/s)

MACH NUMBER (M)

It is the ratio of flow velocity to the speed of sound in the fluid.

$$M = \frac{V}{a}$$

Where,

V = Flow velocity (m/s)

a = Speed of sound in fluid (m/s)

Classification:

- Incompressible Flow : M < 0.3
- Compressible Flow : M ≥ 0.3
- Subsonic Flow : M < 1
- Supersonic Flow : M > 1
- Sonic Flow : M = 1

Speed of Sound in Air (at 20°C)

$a \approx 343$ m/s

LAMINAR TO TURBULENT CRITERIA

Flow Regime	Reynolds Number (Re)	Nature of Flow
Laminar	$Re < 2000$	Smooth, orderly, no mixing
Transitional	$2000 \le Re \le 4000$	Unstable, intermittent
Turbulent	$Re > 4000$	Random, eddies, strong mixing

Note: The critical Reynolds number may vary depending on pipe roughness and flow conditions.

REAL LIFE EXAMPLES

Laminar Flow	Turbulent Flow	Compressible Flow	Rotational Flow	Unsteady Flow
Oil in small pipes	Waterfall, river flow	Air over aircraft wings	Cyclone, whirlpools	Breathing, piston motion

9.7.2.4.5 Boussinesq approximation (buoyancy)

In fluid dynamics, the Boussinesq approximation is used in the field of buoyancy-driven flow (also known as natural convection). It ignores density differences except where they appear in terms multiplied by g, the acceleration due to gravity. The essence of the Boussinesq approximation is that the difference in inertia is negligible but gravity is sufficiently strong to make the specific weight appreciably different between the two fluids. Sound waves are impossible/neglected when the Boussinesq approximation is used since sound waves move via density variations.

Boussinesq flows are common in nature (such as atmospheric fronts, oceanic circulation, katabatic winds), industry (dense gas dispersion, fume cupboard ventilation), and the built environment (natural ventilation, central heating). The approximation is extremely accurate for many such flows, and makes the mathematics and physics simpler.

Given the relation already proved previously:

$$\partial_t\vec{v}+\vec{\nabla}\left[\frac{v^2}{2}+\frac{p}{\rho}+U\right]+2\vec{\omega}\times\vec{v}=\vec{0} \tag{9.7.471}$$

Putting in it again the term containing the viscosity:

$$\partial_t\vec{v}+\vec{\nabla}\left[\frac{v^2}{2}+\frac{p}{\rho}+U\right]+2\vec{\omega}\times\vec{v}-\mu_d\vec{\nabla}^2 v=\vec{0} \tag{9.7.472}$$

without forgetting that at the notation level (we know ... it's a bit annoying):

$$\Delta\cdot v=\vec{\nabla}^2 v=\vec{\nabla}\circ(\vec{\nabla}\cdot\vec{v})=\vec{\nabla}(\vec{\nabla}\circ\vec{v})-\vec{\nabla}\times(\vec{\nabla}\times\vec{v}) \tag{9.7.473}$$

If the potential is a gravitational one, it goes without saying that:

$$\vec{\nabla}\cdot U=\vec{\nabla}\cdot(gz)=\vec{g} \tag{9.7.474}$$

Therefore:

$$\partial_t\vec{v}+\vec{\nabla}\frac{v^2}{2}+\frac{1}{\rho}\vec{\nabla}p+\vec{g}+2\vec{\omega}\times\vec{v}-\mu_d\vec{\nabla}^2 v=\vec{0} \tag{9.7.475}$$

$$\partial_t\vec{v}+\vec{\nabla}\frac{v^2}{2}+\frac{1}{\rho}\vec{\nabla}p+\vec{g}+2\vec{\omega}\times\vec{v}-\mu_d\vec{\nabla}v=\vec{0} \tag{9.7.476}$$

If we can consider the context of the experiment such that the volume density is less than or equal to that of the water and the velocities are small, then we can eliminate the second degree terms such that the previous relation will be written:

$$\partial_t\vec{v}+2\vec{\omega}\times\vec{v}+\frac{1}{\rho}\vec{\nabla}p+\vec{g}-\mu_d\vec{\nabla}^2 v=\vec{0} \tag{9.7.477}$$

We place ourselves within the framework of a weakly turbulent fluid, in which the pressure and the density are written:

$$p=p_0+p' \qquad \rho=\rho_0+\rho' \qquad \frac{\rho'}{\rho_0}\ll 1 \tag{9.7.478}$$

where p', ρ' represents the turbulent increasing terms with respect to the static values of the fluid.

We also neglect the friction on the bounds and therefore the viscosity, assuming that the effect of the turbulence quickly becomes preponderant on the value of the friction.

So we have the system of equations:

$$\partial_t\vec{v}+2\vec{\omega}\times\vec{v}+\frac{1}{\rho}\vec{\nabla}p+\vec{g}=\vec{0} \tag{9.7.479}$$

that can be written:

$$\rho\partial_t\vec{v}+\rho 2\vec{\omega}\times\vec{v}+\vec{\nabla}p+\rho\vec{g}=\vec{0} \tag{9.7.480}$$

and even:

$$\rho\partial_t\vec{v}+\rho 2\vec{\omega}\times\vec{v}+\vec{\nabla}(p_0+p')+(\rho_0+\rho')\vec{g}=\vec{0} \tag{9.7.481}$$

which can also written:

$$\rho\partial_t\vec{v}+\rho 2\vec{\omega}\times\vec{v}+(\vec{\nabla}p'+\rho'\vec{g})+(\vec{\nabla}p_0+\rho_0\vec{g})=\vec{0} \tag{9.7.482}$$

But in the static case we know (we proved it many times) that we have:

$$\vec{\nabla}p_0+\rho\vec{g}=\vec{0} \tag{9.7.483}$$

It then remains to us:

$$\rho\partial_t\vec{v}+\rho 2\vec{\omega}\times\vec{v}+\vec{\nabla}p'+\rho'\vec{g}=\vec{0} \tag{9.7.484}$$

Therefore it remains:

$$\rho\partial_t\vec{v}+\rho 2\vec{\omega}\times\vec{v}+\vec{\nabla}p'+\rho'\vec{g}=\vec{0} \tag{9.7.485}$$

By dividing the whole by ρ:

$$\partial_t\vec{v}+2\vec{\omega}\times\vec{v}+\frac{1}{\rho}\vec{\nabla}p'+\frac{\rho'}{\rho}\vec{g}=\vec{0} \tag{9.7.486}$$

but again we can write this:

$$\partial_t\vec{v}+2\vec{\omega}\times\vec{v}+\frac{1}{\rho'+\rho_0}\vec{\nabla}p'+\frac{\rho'}{\rho+\rho_0}\vec{g}=\vec{0} \tag{9.7.487}$$

The Boussinesq approximation consisting of supposing that the fluid is incompressible and that the system is at constant temperature and not very turbulent, we have:

$$\rho_0\gg\rho' \tag{9.7.488}$$

Which gives us:

$$\boxed{\partial_t\vec{v}+2\vec{\omega}\times\vec{v}+\frac{1}{\rho_0}\vec{\nabla}p'+\frac{\rho'}{\rho_0}\vec{g}=\vec{0}} \tag{9.7.489}$$

This equation is named the "Boussinesq equation" and will allow us to introduce the theory of chaos in the field of meteorology and fluids in the particular case of convection cells (see section Weather & Marine Engineering page 8328).

9.7.2.4.6 Stokes' law

In 11851 (holocene calendar), George Gabriel Stokes derived an expression, now known as Stokes' law, for the frictional force - also named drag force - exerted on spherical objects with very small Reynolds numbers (i.e. very small particles) in a viscous fluid. Stokes' law is derived by solving the Stokes flow limit for small Reynolds numbers of the Navier–Stokes equations.

The complexity of hydrodynamics is a suitable framework for the application of dimensional analysis which we introduced at the very beginning of our study of Mechanics (see section Principia page 3080). The example analyzed here clearly shows the possibilities, but also the limitations of this method.

We consider a solid of any form immersed in an incompressible fluid with a uniform velocity at great distance (the problem is equivalent to that of a solid which moves at a constant velocity in a fluid at rest). We try to express the force F that the fluid exerts on the obstacle, supposed to be immobile (and in particular devoid of any rotational motion).

The analytical solution is too complex to waste time solving this kind of practical problem. Dimensional analysis should be used.

The relevant parameters are in our study:

- L the linear dimension of the obstacle
- v the fluid velocity at far distances
- ρ the fluid density
- μ the viscosity coefficient of the fluid

As a matter of fact, all these parameters are constants, even if the velocity varies in direction and norm in the vicinity of the obstacle: at long distances, it is uniform and its value v is indeed a relevant parameter.

We might ask ourselves if pressure should not be one of those parameters. This is not the case. The pressure is conditioned by the value of the velocity and by those of the constant parameters as we saw in Bernoulli's theorem. There is no need to add a redundant term.

Without seeking the unique combination without dimension of the first four, we apply the systematic approach. We want to determine A, B, C, D such that:

$$F = L^A v^B \rho^C \mu^D \tag{9.7.490}$$

As:

$$F:[\mathrm{kg}\cdot\mathrm{m}\cdot\mathrm{s}^{-2}] \qquad \rho:[\mathrm{kg}\cdot\mathrm{m}^{-3}] \qquad v:[\mathrm{m}\cdot\mathrm{s}^{-1}] \qquad \mu:[\mathrm{kg}\cdot\mathrm{m}^{-1}\cdot\mathrm{s}^{-1}] \tag{9.7.491}$$

It comes:

$$\mathrm{kg}\cdot\mathrm{m}\cdot\mathrm{s}^{-2} = \mathrm{kg}^{C+D}\cdot\mathrm{m}^{A+B-3C-D}\cdot\mathrm{s}^{-B-D} \tag{9.7.492}$$

The dimensionality system is then written:

$$\begin{aligned}&C+D=1\rightarrow C=1-D\\&A+B-3C-D=1\\&B+D=2\Rightarrow B=2-D\end{aligned}\tag{9.7.493}$$

Therefore:

$$A=1-B+3C+D=2-D\tag{9.7.494}$$

Hence:

$$F=\rho v^2L^2\left(\frac{Lv\rho}{\mu}\right)^{-D}\tag{9.7.495}$$

and curiously we fall back here in the parenthesis on what we had seen in our development of the Reynolds number:

$$\text{Re}=\frac{\rho_r L_r v_r}{\mu_{rd}}=\frac{L_r v_r}{\mu_{rc}}\tag{9.7.496}$$

Thus the force exerted by the fluid is then written:

$$F=\rho v^2L^2\text{Re}^{-D}\tag{9.7.497}$$

In the literature, we find the notation for the drag force:

$$\boxed{F=\rho v^2L^2C(\text{Re})}\tag{9.7.498}$$

Where C depends obviously on Re. We also find this relation in some textbooks in the following form:

$$F_D=\frac{1}{2}\rho V^2C_DA\tag{9.7.499}$$

Or in even more condensed form:

$$F_D=C_Dv^2\tag{9.7.500}$$

where appears the "drag coefficient C_D" which is determined in wind tunnels or experimentally based on the fact that an object in free fall in a perfect liquid or gas will follow the differential equation:

$$m\frac{\mathrm{d}v}{\mathrm{d}t}=mg-C_Dv^2\tag{9.7.501}$$

Differential equation that we can solve ingeniously by writing:

$$m\frac{\mathrm{d}x}{\mathrm{d}x}\frac{\mathrm{d}v}{\mathrm{d}t}=mv\frac{\mathrm{d}v}{\mathrm{d}x}=mg-C_Dv^2\tag{9.7.502}$$

It then comes after rearrangement of the terms:

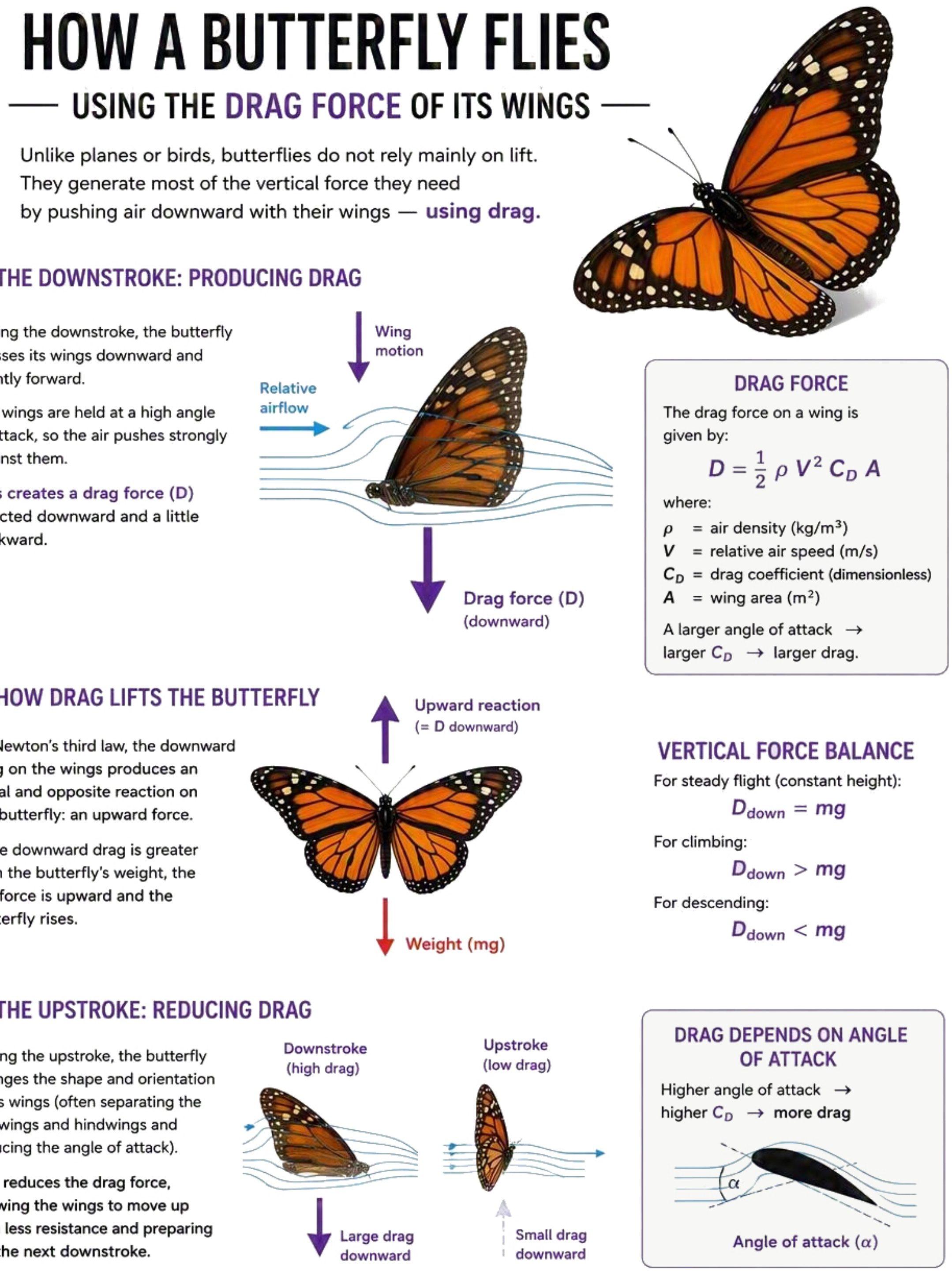
$$\frac{v\mathrm{d}v}{1-\dfrac{C_D}{mg}v^2}=g\mathrm{d}x\tag{9.7.503}$$

Either after integration:

$$\ln\left(1-\frac{C_D}{mg}v^2\right)=-2\frac{C_D}{m}x \tag{9.7.504}$$

Then it comes a relation which makes it possible to determine experimentally the drag coefficient for an object in free fall without too many difficulties:

$$v^2=\frac{mg}{C_D}\left(1+e^{-2\frac{C_D}{m}x}\right) \tag{9.7.505}$$



HOW A BUTTERFLY FLIES

— USING THE DRAG FORCE OF ITS WINGS —

Unlike planes or birds, butterflies do not rely mainly on lift. They generate most of the vertical force they need by pushing air downward with their wings — **using drag.**

1 THE DOWNSTROKE: PRODUCING DRAG

During the downstroke, the butterfly presses its wings downward and slightly forward.

The wings are held at a high angle of attack, so the air pushes strongly against them.

This creates a drag force (D) directed downward and a little backward.

DRAG FORCE

The drag force on a wing is given by:

$$D=\frac{1}{2}\rho V^2 C_D A$$

where:

ρ = air density (kg/m³)
V = relative air speed (m/s)
C_D = drag coefficient (dimensionless)
A = wing area (m²)

A larger angle of attack → larger C_D → larger drag.

2 HOW DRAG LIFTS THE BUTTERFLY

By Newton's third law, the downward drag on the wings produces an equal and opposite reaction on the butterfly: an upward force.

If the downward drag is greater than the butterfly's weight, the net force is upward and the butterfly rises.

VERTICAL FORCE BALANCE

For steady flight (constant height):

$$D_{\text{down}}=mg$$

For climbing:

$$D_{\text{down}}>mg$$

For descending:

$$D_{\text{down}}<mg$$

3 THE UPSTROKE: REDUCING DRAG

During the upstroke, the butterfly changes the shape and orientation of its wings (often separating the forewings and hindwings and reducing the angle of attack).

This reduces the drag force, allowing the wings to move up with less resistance and preparing for the next downstroke.

DRAG DEPENDS ON ANGLE OF ATTACK

Higher angle of attack → higher C_D → more drag

So if we want to check for a butterfly like the Gliding Swallowtail Butterfly that (we replace the drag coefficient C_D by the lift coefficient C_L):

$$\frac{1}{2}\rho V^2 C_L A \overset{?}{>} mg \tag{9.7.506}$$

we just have to replace the experimental numerical values taken from [431] (lift coeficient measured by fixing a real gliding swallowtail butterfly in a small wind tunnel), [598] and [98] to get:

$$\frac{1}{2}\cdot 1.225\cdot 1.4^2\cdot 0.65\cdot(7.011\cdot 10^{-3})\overset{?}{>}(4\cdot 10^{-5})\cdot 9.81 \tag{9.7.507}$$

That indeed satisfies observed reality (i.e. the Gliding Swallowtail Butterfly can fly...):

$$0.005471 > 0.000392 \tag{9.7.508}$$

So the lift force of this butterfly if approximately 13 times greater than the force of gravity!

Remark

Obviously C_D contains a mass unit, so nothing prevents us to extract a mass term and also a factor $1/2$, and also the surface of the object S (obviously we can guess that the surface has an incidence on the drag force..) to introduce a new drag coefficient denoted C such that we define the "drag force" as:

$$F_D = \frac{1}{2}Cmv^2 = \frac{1}{2}C\rho S v^2 \tag{9.7.509}$$

Then the "terminal velocity" (equilibrium between the gravitational force and drag force of a falling object in gas or liquid) is simply given by:

$$mg = F_D \Leftrightarrow mg = \frac{1}{2}C\rho S v^2 \tag{9.7.510}$$

Hence:

$$\boxed{v = \sqrt{\frac{2mg}{\rho C S}}} \tag{9.7.511}$$

✍**Example:**

To keep an object in motion in the presence of drag (aerodynamic or otherwise) requires an ongoing input of energy. Work must be done over some time. Power must be used. Recall the following chain of reasoning that starts from the definition of power as the rate at which work is done:

$$P = \frac{W}{t} = \frac{F\cdot x}{t} = F\cdot v \tag{9.7.512}$$

or more specifically, in the case of pressure drag:

$$P = F_D\cdot v = \left(\frac{1}{2}C\rho S v^2\right)v = \frac{1}{2}C\rho S v^3 \tag{9.7.513}$$

Therefore we have for maximum speed:

$$v=\sqrt[3]{\frac{2P}{\rho CS}} \tag{9.7.514}$$

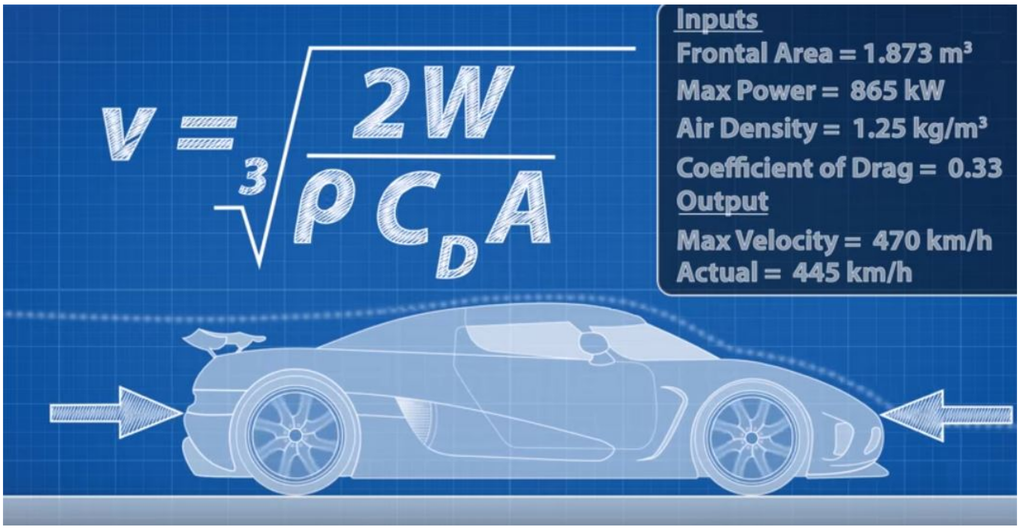

$V=\sqrt[3]{\frac{2W}{\rho C_D A}}$
Inputs
Frontal Area = 1.873 m³
Max Power = 865 kW
Air Density = 1.25 kg/m³
Coefficient of Drag = 0.33
Output
Max Velocity = 470 km/h
Actual = 445 km/h

Figure 9.249: **Koenigsegg Agera RS theoretical maximum speed (source: Real Engineering/YouTube)**

The limits of the dimensional analytical method (and just analytic one...) appear when we confront this model with the experiment (obviously we could make numerical models of the Navier-Stokes-Reynolds equation for the computer and thus the honor would be safe):

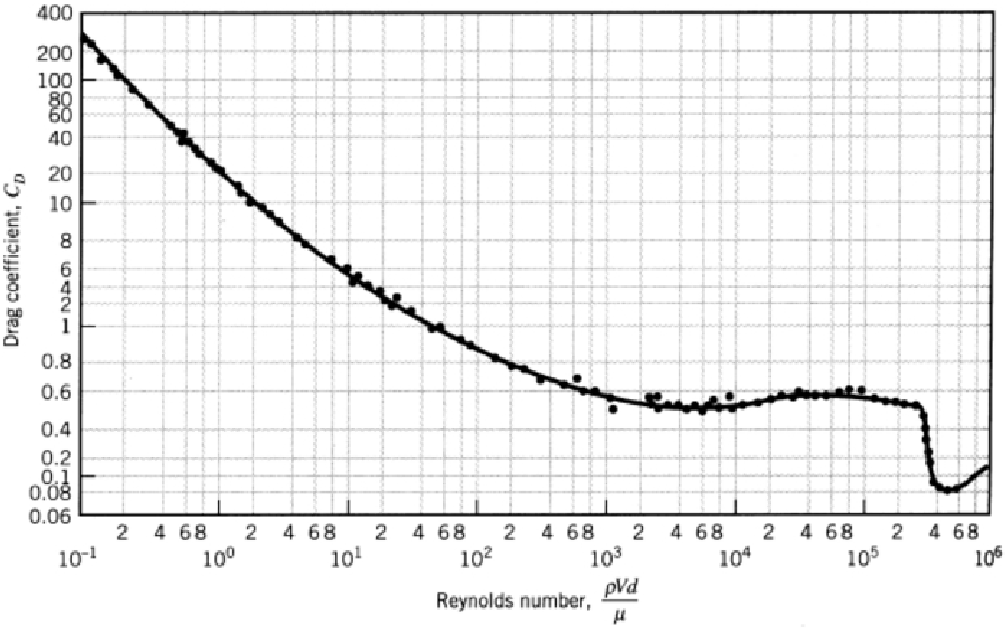


Figure 9.250: **Effect of Reynolds number on the drag coefficient C_D of a smooth sphere of diameter d**

or more explicitly:

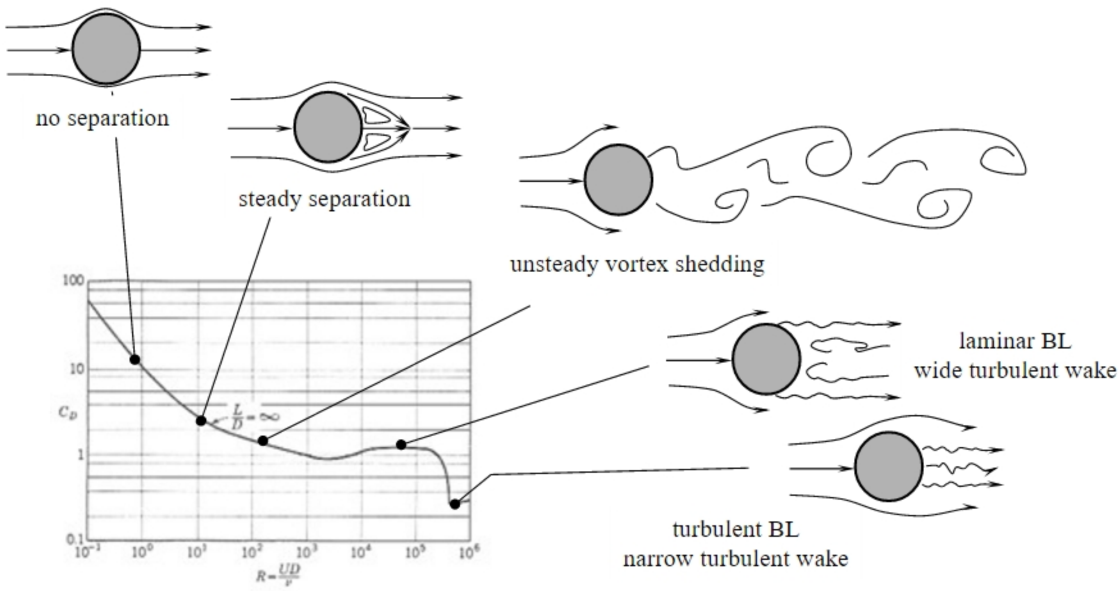


Figure 9.251: Effect of Reynolds number on the drag coefficient C_D of a smooth sphere with regime

Or comparing the drag coefficient of a sphere and a cylinder:

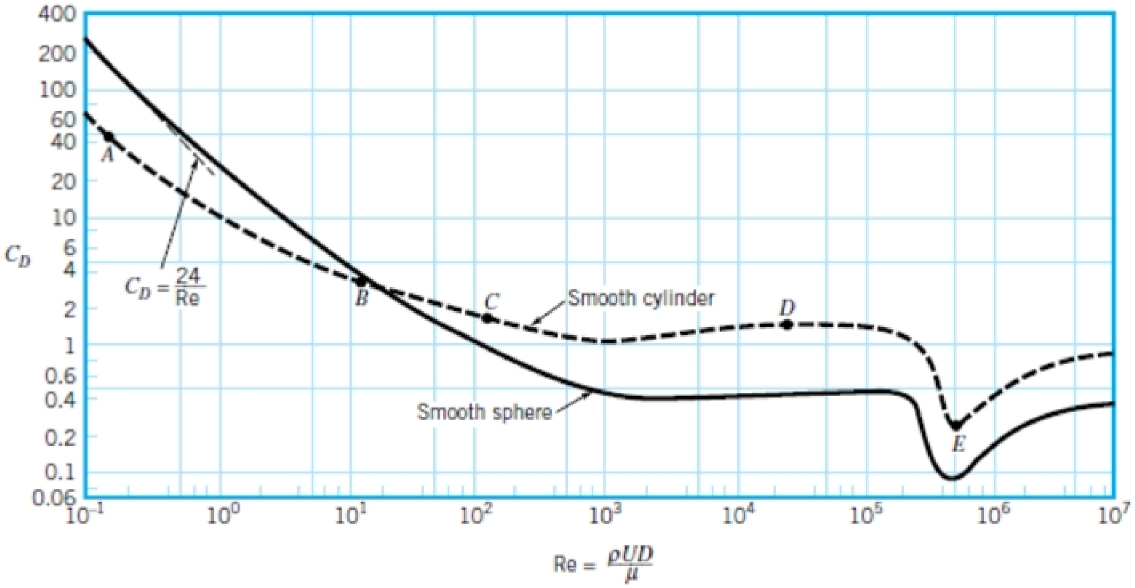


Figure 9.252: Drag coefficient comparison of a sphere and cylinder

The curves above have two remarkable characteristics:

1. It was obtained by independently modifying the values of the four parameters. We find that C_D depends only on the Reynolds number Re: it is a success of the dimensional analysis!
2. It seems useless to hope to find a simple analytic function which reproduces the experimental curve. We must therefore take a closer look at the various regimes corresponding to this complex curve.

The figure below schematizes the flow of a viscous fluid around a cylinder for different values of the Reynolds number:

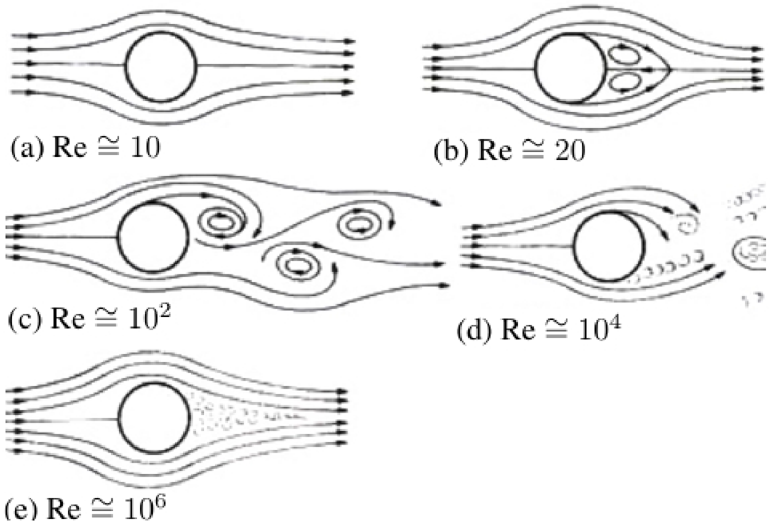


Figure 9.253: **Various fluid regimes for cylinder with corresponding Reynolds number**

The regime corresponding to figure (a) is name "steady state" or "steady regime". We can speak of a "quasi-static" movement on each point of the fluid where the acceleration is negligible. We must therefore expect that the inertia of the fluid does not intervene in the expression of force. For this, it is necessary and sufficient that:

$$\lim_{\text{Re}\to 0} C(\text{Re}) \to C'\frac{1}{\text{Re}} = C'\frac{\mu}{Lv\rho} \tag{9.7.515}$$

where C is independent of Re.

Therefore we have:

$$F = \rho v^2 L^2 C' \frac{\mu}{Lv\rho} = C'vL\mu \tag{9.7.516}$$

The parameter C' without dimensions can depend only on the geometry of the obstacle. In the case where the obstacle is spherical (very important case in physics with $L = R$), C' has been determined experimentally as being equall to 6π such that:

$$\boxed{F = 6\pi\mu R v} \tag{9.7.517}$$

known as the "Stokes' law" or "Stokes formula". Attention ... this law only applies well for small speeds and small spheres.

In the regime described by (b), two vortices takes place symmetrically behind the cylinder. When Re increases beyond 40, we distinguish the alley of "Kármán vortex street".

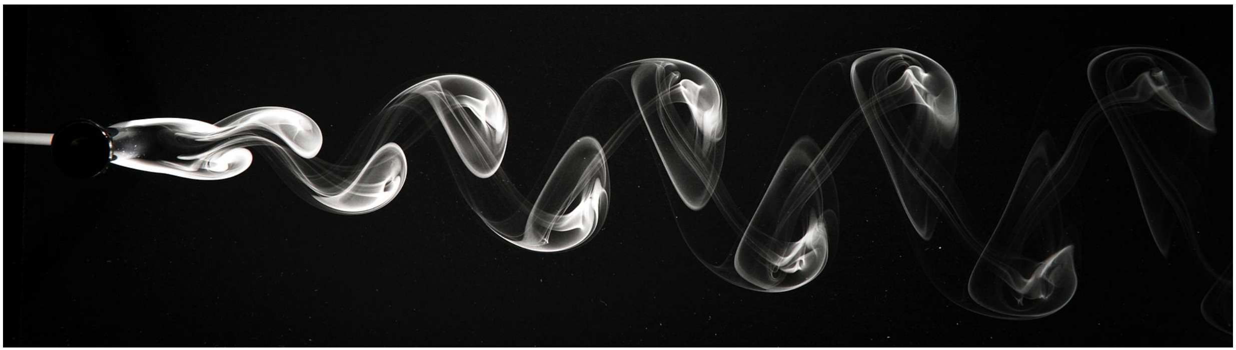


Figure 9.254: Visualisation of the vortex street behind a circular cylinder in air; the flow is made visible through release of oil vapour in the air near the cylinder (source: Wikipedia, author: Jürgen Wagner)

And at the macroscopic level on Earth typically (there is quantity of such photos on the Internet):

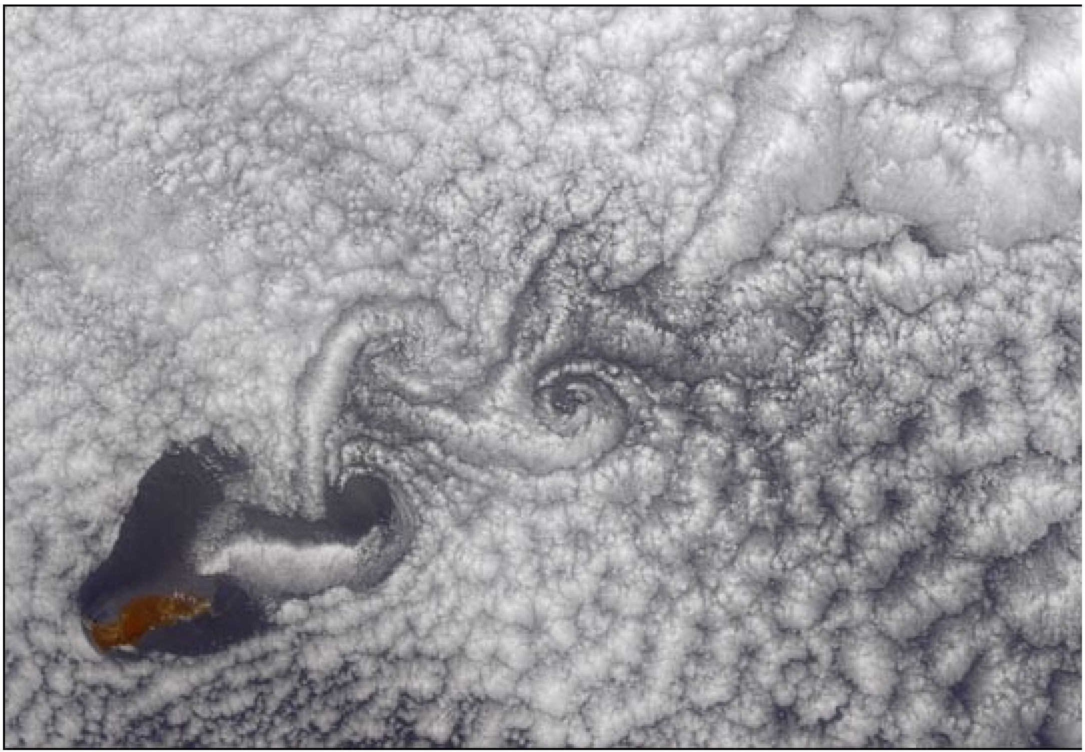


Figure 9.255: Visualization of the vortex street around Guadalupe island (source: NASA/GSFC/JPL, MISR Team)

And at a planetary level:

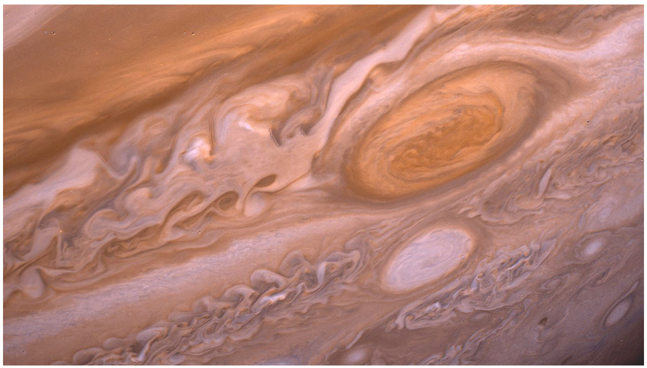


Figure 9.256: Visualisation of the vortex street around Jupiter red spot

9.7.2.4.6.1 Why don't clouds fall?

Clouds can hold an enormous amount of water. When this water falls as rain it clearly has a significant mass so why don't clouds fall? In fact, the small water droplets that make up clouds do fall slowly. However, the drag force of the air dominates over the gravitational force for small particles. The drag force increases as the size of an object decreases. The force needed to move a sphere through a viscous medium is given by Stokes's law:

$$F = 6\pi\mu R v \tag{9.7.518}$$

Here, R is the radius of the sphere, v is the velocity, and μ is the viscosity. The viscosity of air is about $0.018 \cdot 10^{-3}$ [Pa · s] and the viscosity of water is about $1.8 \cdot 10^{-3}$ [Pa · s]. Stokes's law is valid if the Reynolds number $N_{\text{Reynolds}} = 2R\rho v/\mu$ is less than about 2000. Here ρ is the mass density.

A spherical particle falling under the force of gravity will reach terminal velocity when the gravitational force matches the drag force:

$$mg = 6\pi\mu R v \tag{9.7.519}$$

Solving this for the terminal velocity yields:

$$v_{\text{terminal}} = \frac{mg}{6\pi\mu v} = \frac{\left(\rho\frac{4}{3}\pi R^3\right) g}{6\pi\mu R} = \frac{2g\rho R^2}{9\eta} \tag{9.7.520}$$

A water droplet with a 10 [nm] radius falls at $12\,[\text{nm} \cdot \text{s}^{-1}]$ in air. It would take 2.6 years for this droplet to fall one meter. It is only when the small droplets begin to coalesce into larger droplets that they fall with significant speed.

In some sense, the inverse effect to rain is the rising of bubbles in beer. Bubbles are lighter than the surrounding liquid so gravity pushes them up. They rise with a constant velocity which is described by Stokes law. If you are the type that carefully observes your beer, you will have noticed that sometimes the bubbles move down. This is because of the circulation of liquid in the beer glass. The rising bubbles in the center of the glass drag some liquid along with them. After this liquid reaches the top of the glass, it returns to the bottom along the sides of the glass. This downward flow can drag bubbles, especially small bubbles, downwards against the force of gravity.

In a similar manner, small rain drops can be pull up by air currents against the force of gravity.

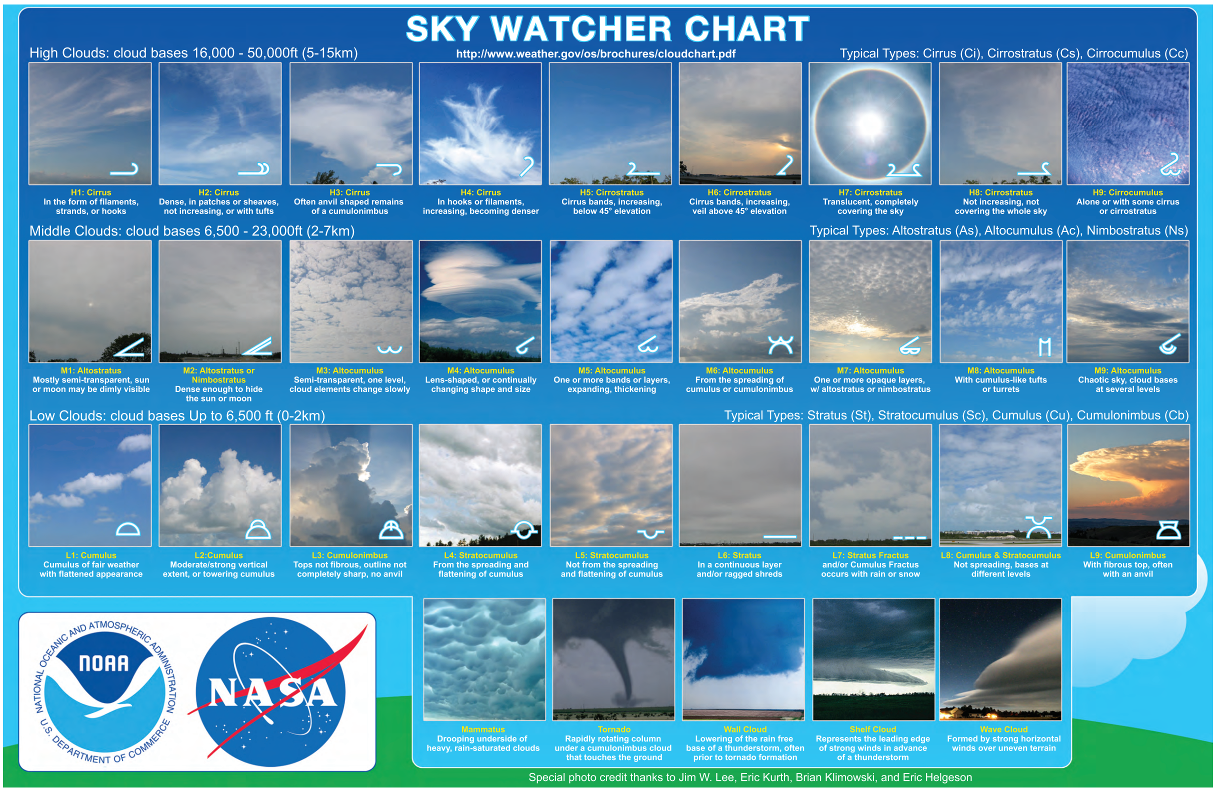

SKY WATCHER CHART
High Clouds: cloud bases 16,000 - 50,000ft (5-15km)
http://www.weather.gov/os/brochures/cloudchart.pdf
Typical Types: Cirrus (Ci), Cirrostratus (Cs), Cirrocumulus (Cc)
H1: Cirrus
In the form of filaments, strands, or hooks
H2: Cirrus
Dense, in patches or sheaves, not increasing, or with tufts
H3: Cirrus
Often anvil shaped remains of a cumulonimbus
H4: Cirrus
In hooks or filaments, increasing, becoming denser
H5: Cirrostratus
Cirrus bands, increasing, below 45° elevation
H6: Cirrostratus
Cirrus bands, increasing, veil above 45° elevation
H7: Cirrostratus
Translucent, completely covering the sky
H8: Cirrostratus
Not increasing, not covering the whole sky
H9: Cirrocumulus
Alone or with some cirrus or cirrostratus
Middle Clouds: cloud bases 6,500 - 23,000ft (2-7km)
Typical Types: Altostratus (As), Altocumulus (Ac), Nimbostratus (Ns)
M1: Altostratus
Mostly semi-transparent, sun or moon may be dimly visible
M2: Altostratus or Nimbostratus
Dense enough to hide the sun or moon
M3: Altocumulus
Semi-transparent, one level, cloud elements change slowly
M4: Altocumulus
Lens-shaped, or continually changing shape and size
M5: Altocumulus
One or more bands or layers, expanding, thickening
M6: Altocumulus
From the spreading of cumulus or cumulonimbus
M7: Altocumulus
One or more opaque layers, w/ altostratus or nimbostratus
M8: Altocumulus
With cumulus-like tufts or turrets
M9: Altocumulus
Chaotic sky, cloud bases at several levels
Low Clouds: cloud bases Up to 6,500 ft (0-2km)
Typical Types: Stratus (St), Stratocumulus (Sc), Cumulus (Cu), Cumulonimbus (Cb)
L1: Cumulus
Cumulus of fair weather with flattened appearance
L2:Cumulus
Moderate/strong vertical extent, or towering cumulus
L3: Cumulonimbus
Tops not fibrous, outline not completely sharp, no anvil
L4: Stratocumulus
From the spreading and flattening of cumulus
L5: Stratocumulus
Not from the spreading and flattening of cumulus
L6: Stratus
In a continuous layer and/or ragged shreds
L7: Stratus Fractus
and/or Cumulus Fractus occurs with rain or snow
L8: Cumulus & Stratocumulus
Not spreading, bases at different levels
L9: Cumulonimbus
With fibrous top, often with an anvil
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
NOAA
U.S. DEPARTMENT OF COMMERCE
NASA
Mammatus
Drooping underside of heavy, rain-saturated clouds
Tornado
Rapidly rotating column under a cumulonimbus cloud that touches the ground
Wall Cloud
Lowering of the rain free base of a thunderstorm, often prior to tornado formation
Shelf Cloud
Represents the leading edge of strong winds in advance of a thunderstorm
Wave Cloud
Formed by strong horizontal winds over uneven terrain
Special photo credit thanks to Jim W. Lee, Eric Kurth, Brian Klimowski, and Eric Helgeson

Figure 9.257: Cloud's sky watcher chart (source: NOAA, NASA)

9.7.2.4.7 Betz's law

"Betz's law" indicates the maximum power that can be extracted from the wind, independent of the design of a wind turbine in open flow. It was published in 1919 by the German physicist Albert Betz[36]. The law is derived from the principles of conservation of mass and momentum of the air stream flowing through an idealized "actuator disk" that extracts energy from the wind stream. According to Betz's law, no turbine can capture more than $16/27$ (59.3%) of the kinetic energy in wind. The factor $16/27$ (0.593) is known as "Betz's coefficient". Practical utility-scale wind turbines achieve at peak 75 to 80% of the Betz limit.

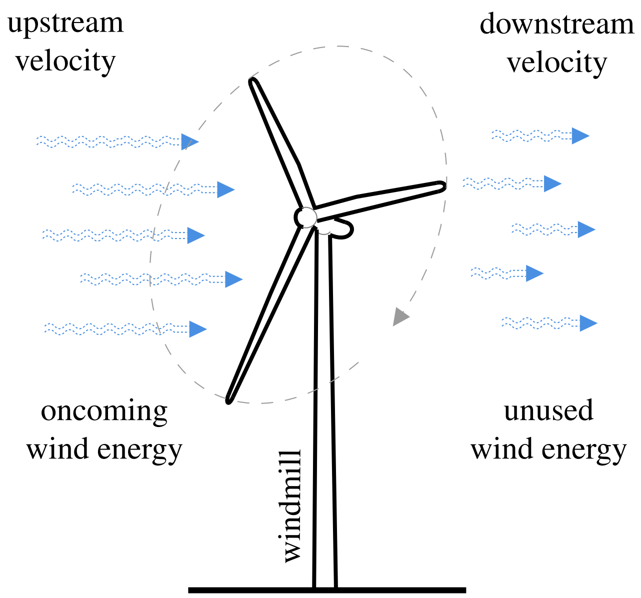
Betz's law applies to all Newtonian fluids, including wind. If all of the energy coming from wind movement through a turbine were extracted as useful energy, the wind speed afterwards would drop to zero. If the wind stopped moving at the exit of the turbine, then no more fresh wind could get in; it would be blocked. In order to keep the wind moving through the turbine, there has to be some wind movement, however small, on the other side with some wind speed greater than zero. Betz's law shows that as air flows through a certain area, and as wind speed slows from losing energy to extraction from a turbine, the airflow must distribute to a wider area. As a result, geometry limits any turbine efficiency to a maximum of 59.3%.

The Betz Limit is the maximum possible energy that may be derived by means of an infinitely thin rotor from a fluid flowing at a certain speed.

---

[36] British scientist Frederick W. Lanchester derived the same maximum in 1915. The leader of the Russian aerodynamic school, Nikolay Zhukowsky, also published the same result for an ideal wind turbine in 1920, the same year as Betz did. It is thus an example of Stigler's law, which posits that no scientific discovery is named after its actual discoverer.

In order to calculate the maximum theoretical efficiency of a thin rotor (of, for example, a windmill) one imagines it to be replaced by a disc that withdraws energy from the fluid passing through it. At a certain distance behind this disc the fluid that has passed through flows with a reduced velocity.



The assumptions are:

- The rotor does not possess a hub and is ideal, with an infinite number of blades, which have no drag. Any resulting drag would only lower this idealized value.
- The flow into and out of the rotor is axial. This is a control-volume analysis, and to construct a solution, the control volume must contain all flow going in and out, failure to account for that flow would violate the conservation equations.
- The flow is non-compressible. Density remains constant, and there is no heat transfer.
- Uniform thrust is exerted on the disc or rotor.

Applying conservation of mass to this control volume, the mass flow rate (the mass of fluid flowing per unit time) is given by:

$$\dot{m} = \rho A_1 v_1 = \rho S v = \rho A_2 v_2 \tag{9.7.521}$$

where v_1 is the speed in the front of the rotor, v_2 is the speed downstream of the rotor, v is the speed at the fluid power device, ρ is the fluid density, the area of the turbine is given by S, and A_1 and A_2 are the areas of the fluid before and after reaching the turbine.

So the density times the area and speed should be equal in each of the three regions: before, while going through the turbine and afterwards!

The force exerted on the wind by the rotor is the mass of air multiplied by its acceleration. In terms of the density, surface area and velocities, this can be written as:

$$F = ma = m\frac{\mathrm{d}v}{\mathrm{d}t} = \dot{m}\Delta v = \rho S v\left(v_1 - v_2\right) \tag{9.7.522}$$

The work done by the force may be written incrementally as:

$$\mathrm{d}E = F\mathrm{d}x \tag{9.7.523}$$

and the power (rate of work done) of the wind is (the reader must try to not confuse the P of the Power used here with the P of pressure...):

$$P = \frac{\mathrm{d}E}{\mathrm{d}t} = F\frac{\mathrm{d}x}{\mathrm{d}t} = Fv \tag{9.7.524}$$

Now substituting the force F computed above into the power equation will yield the power extracted from the wind:

$$P = \rho S v^2\left(v_1 - v_2\right) \tag{9.7.525}$$

However, power can be computed another way, by using the kinetic energy. Applying the conservation of energy equation to the control volume yields:

$$P = \frac{\Delta E}{\Delta t} = \frac{1}{2}\dot{m}\left(v_1^2 - v_2^2\right) \tag{9.7.526}$$

Looking back at the continuity equation, a substitution for the mass flow rate yields:

$$P = \frac{1}{2}\rho S v\left(v_1^2 - v_2^2\right) \tag{9.7.527}$$

Both of these expressions for power are completely valid, one was derived by examining the incremental work done, and the other by the conservation of energy. Equating these two expressions yields:

$$P = \frac{1}{2}\rho S v\left(v_1^2 - v_2^2\right) = \rho S v^2\left(v_1 - v_2\right) \tag{9.7.528}$$

For all v and S the density cannot be 0. Examining the two equated expressions yields an interesting result, namely:

$$\frac{1}{2}\left(v_1^2 - v_2^2\right) = \frac{1}{2}\left(v_1 - v_2\right)\left(v_1 + v_2\right) = v\left(v_1 - v_2\right) \tag{9.7.529}$$

or:

$$v = \frac{1}{2}\left(v_1 + v_2\right) \tag{9.7.530}$$

Therefore, the wind velocity at the rotor may be taken as the average of the upstream and downstream velocities.

Returning to the previous expression for power based on kinetic energy:

$$\begin{aligned} P &= \frac{1}{2}\dot{m}\left(v_1^2 - v_2^2\right) = \frac{1}{2}\rho S v\left(v_1^2 - v_2^2\right) = \frac{1}{4}\rho S\left(v_1 + v_2\right)\left(v_1^2 - v_2^2\right) \\ &= \frac{1}{4}\rho S v_1^3\left(1 - \left(\frac{v_2}{v_1}\right)^2 + \left(\frac{v_2}{v_1}\right) - \left(\frac{v_2}{v_1}\right)^3\right) \end{aligned} \tag{9.7.531}$$

By differentiating P with respect to $\frac{v_2}{v_1}$ for a given fluid speed v_1 and a given area S, one finds the maximum or minimum value for P:

$$P_{\max} \Rightarrow \frac{\partial P}{\partial(v_2/v_1)} = \frac{1}{4}\rho S v_1^3 \left(-2\frac{v_2}{v_1} + 1 - 3\left(\frac{v_2}{v_1}\right)^2\right) = 0 \tag{9.7.532}$$

Therefore:

$$-2\frac{v_2}{v_1} + 1 - 3\left(\frac{v_2}{v_1}\right)^2 = 0 \tag{9.7.533}$$

Hence:

$$3\left(\frac{v_2}{v_1}\right)^2 + 2\frac{v_2}{v_1} - 1 = 0 \tag{9.7.534}$$

This is an equation of the second degree in $v_2/v1$ that is solved easily... The solutions are:

$$\left(\frac{v_2}{v_1}\right)_{1,2} = \frac{-2 \pm \sqrt{2^2 - 4\cdot 3\cdot(-1)}}{2\cdot 3} = \frac{-2 \pm \sqrt{16}}{6} = \{1/3, -1\} \tag{9.7.535}$$

Practical considerations lead us to reject the negative solution. Therefore we keep only:

$$P_{\max} \Rightarrow \frac{v_2}{v_1} = \frac{1}{3} \tag{9.7.536}$$

Therefore:

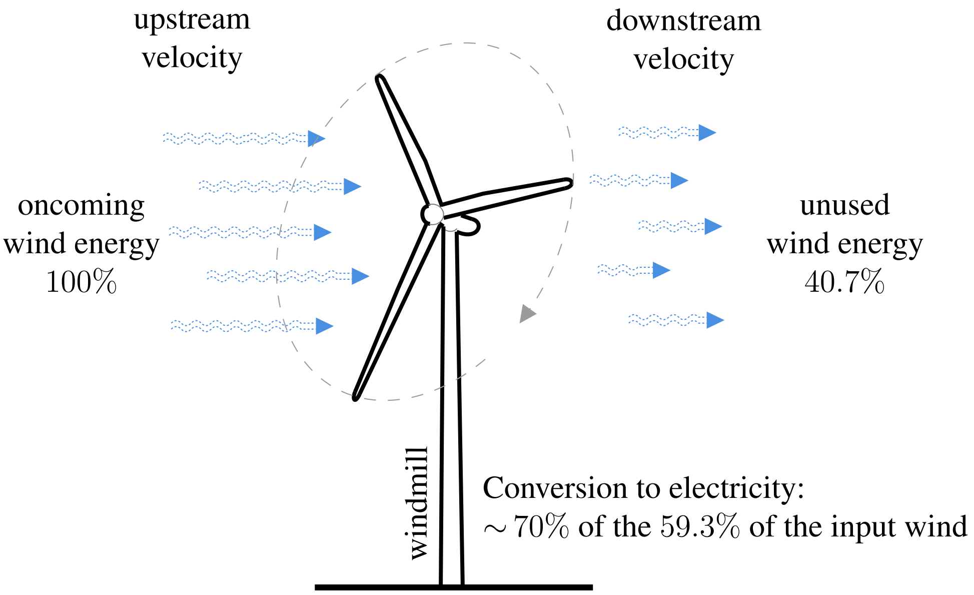
$$\begin{aligned} P_{\max} &= \frac{1}{4}\rho S v_1^3 \left(1 - \left(\frac{v_2}{v_1}\right)^2 + \left(\frac{v_2}{v_1}\right) - \left(\frac{v_2}{v_1}\right)^3\right) \\ &= \frac{1}{4}\rho S v_1^3 \left(1 - \frac{5}{27}\right) = \frac{1}{4}\rho S v_1^3 \frac{32}{27} = \frac{8}{27}\rho S v_1^3 \end{aligned} \tag{9.7.537}$$

Hence:

$$\frac{\text{Max power extracted}}{\text{Max power available}} = \frac{\frac{8}{27}\rho S v_1^3}{\frac{1}{2}\rho S v_1^3} = \frac{16}{27} \cong 0.593 = 59.3\% := C_p \tag{9.7.538}$$

C_p is referred to as the "Betz's limit" and states that the maximum power an ideal rotor can extract from wind is 59.3%.

An ideal rotor of the type described above is named an "actuator disk". The actuator disk induces a reduction of the free-stream wind speed.



This is an important limit because it defines the upper limit of the efficiency of any rotor disk type energy extracting device that is placed in the flow of a fluid. A large fraction of the 59.3% of total wind energy that is extracted from wind is transferred to the turbine, but some of it is used to overcome viscous drag on blades and generate vortices in the wake. Within the turbine, most of the energy is converted into useful electrical energy, while some of it is lost in gearbox, bearings, generator, power converter, transmission and others. Most practical rotors with three blades reach an overall efficiency of about 50%.

Often, inventors claim to have developed a rotor that achieves an efficiency that is greater than 59.3%; such claims are suspect and must be analysed. There are turbine configurations that may violate the Betz limit.

✍**Example:**

Let us consider a wind turbine with a rotor diameter of $70\,[\text{m}]$ and assume that $v_1=12\,[\text{m}\cdot\text{s}^{-1}]$ with air density of $1.225\,[\text{kg}\cdot\text{m}^{-3}]$. Therefore:

$$P_{\text{Betz}}(v_1)=\frac{16}{27}\frac{\rho S v_1^3}{2}=\frac{16}{27}\frac{1.225\pi\left(\frac{70}{2}\right)^2 12^3}{2}=2.4\,[\text{MW}]$$

For more information about this result and its implications, the reader can refer to the excellent textbooks [285].

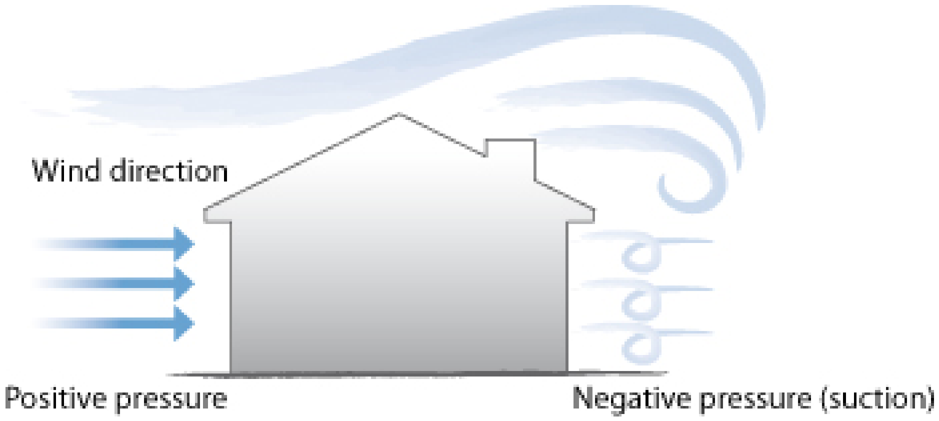
Remark

In the technical field study of pumps, tires or compressed air tanks, we measure the pressure with "manometers" whose zero corresponds to the atmospheric pressure.

9.7.2.4.7.1 Wind pressure

The "wind pressure", alos named "velocity pressure", is the total force exerted upon a structure by wind.

For a flat surface it consists of two factors, the first being the dynamic pressure exerted on the windward side of the surface (i.e. wind load). This is equal to $(1/2)\rho v^2$ as seen just earlier above[37], where ρ is the air density and v is the wind speed normal to the surface. The second factor is the pressure decrease, or "suction", produced on the leeward side of the surface (that latter is often forgotten as many beginners consider that wall as having a infinite surface size), which is equal to $(1/2)C\rho v^2$, where C is a structural constant varying from -0.3 for cylindrical objects to 1.0 for long plates (see page 3856 for the derivation of this force).



The wind pressure P is the sum of these two, explicitly:

$$P_{\text{wind}} = \frac{1}{2}(1+C)\rho v^2 \tag{9.7.539}$$

This formula should be used with caution. The actual wind pressure depends on the shape of the object and the nature of the environment. For objects not presenting a flat surface to the wind, such as bridges and chimneys, complicating aerodynamic factors are introduced that make the above formula invalid.

If we just want a lower bound (lower limit) we can get rid of the drage force to get:

$$P_{\text{wind,LL}} = \frac{1}{2}\rho v^2 \tag{9.7.540}$$

If we consider that the Betz's limit may apply to any flat obstacle, we could we even write:

$$P_{\text{wind}} = \frac{1}{2}(C_p+C)\rho v^2 \tag{9.7.541}$$

Therefore a theoretical lower bound would be:

$$P_{\text{wind,LL}} = \frac{1}{2}C_p\rho v^2 \tag{9.7.542}$$

---

[37] Dynamic pressure based on $(1/2)\rho v^2$ is for incompressible flow. It quickly becomes meaningless above Mach 0.3.

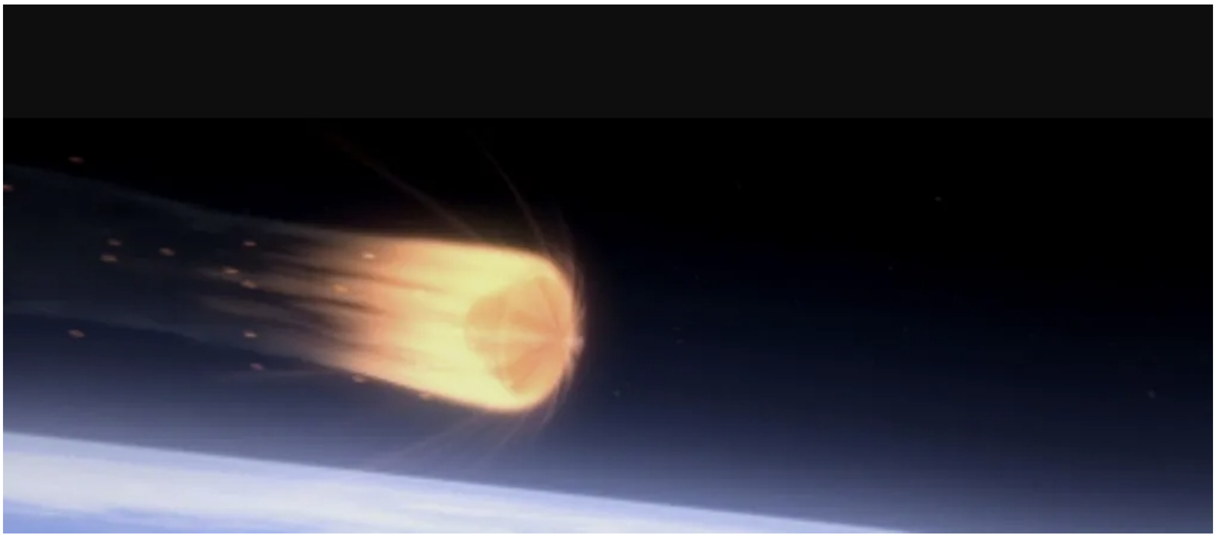
✍**Example:**

What is the lower bound force of a wind with a speed of $100\,[\text{km}\cdot\text{h}^{-1}]$ (i.e. $27.8\,[\text{m}\cdot\text{s}^{-1}]$) on a 20 square meter wall given the air density is $1.225\,[\text{kg}\cdot\text{m}^{-3}]$ and $C_p = 0.593$?

The solution is simply given by:

$$F_{\text{wind,LL}} = P_{\text{wind,LL}} \cdot S = \frac{1}{2}C_p\rho v^2 S \cong 5614\,[\text{N}]$$

We can't apply the calculations above (as far as we know) to compute the thermal shield temperature of a spacecraft that re-enters Earth's atmosphere deriving the encountered pressure at a typical speed of $28,000\,[\text{km}\cdot\text{h}^{-1}]$ (i.e. $7,777\,[\text{m}\cdot\text{s}^{-1}]$) with a thermal shield with a diameter of $3.3\,[\text{m}]$ with a $C = -0.2$ at a height of $80\,[\text{km}]$ (corresponding to an air density of approximately $3\cdot 10^{-5}\,[\text{kg}\cdot\text{m}^{-3}]$) as we don't know on which volume the result should be applied to!



By re-entering atmosphere spaceship transfers it's kinetic energy to impacting air molecules, thus molecules starts to move at a spaceship speed too. Knowing average molecules speed (which would be an aircraft speed) we can calculate shock-wave zone gas temperature (see section Statistical Mechanics page 3558):

$$T = \frac{m\bar{v}^2}{3k} \tag{9.7.543}$$

Typical re-entry speed of Space Shuttle is about $7,823\,[\text{m}\cdot\text{s}^{-1}]$, this gives shock zone temperature $\sim 68,000$ [K] that seems more than enough to heat Nitrogen into a plasma...!? Obviously the last relation is not accurate as if you apply it for commercial airplanes you would get a temperature of ~ 70 [K], for the Concorde at Mach 2 you would get ~ 490 [K] and for the SR-71 Blackbird you would get $\sim 1,000$ [K]...

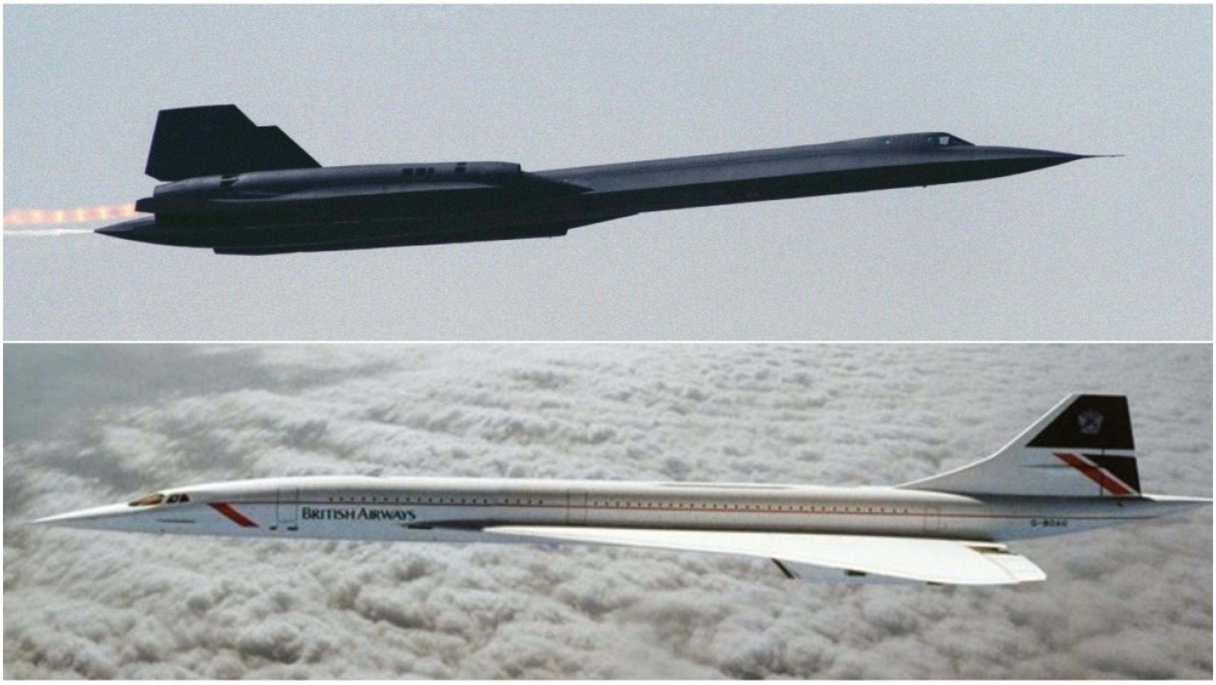


Figure 9.258: **SR-71 Blackbird on the top and Concorde on the bottom**

The reason why the spacecraft itself doesn't heat up at this temperature is that spacecrafts are protected with a heat shield composed of layers from special materials dissipating heat received. Also there's an ablative (process that utilizes a sacrificial liner which cools something through chemical reactions and the enthalpy of vaporization) cooling mechanism in action and besides at these altitudes air density is very small, so heat is transferred from plasma to spacecraft at low rates.

FDM simulations (see section Numerical Methods page 6752) show the air in the bow shock of the stardust probe reaches temperatures of around $50,000$ [K] at 71 [km] altitude, falling to $10,000$ [K] at 51 [km].

9.7.2.5 Levelling law

We will prove now that a fluid (gas or liquid) will be levelled if and only if it is a stationary compressible inviscid fluids in a gravitational potential with an isobaric surface.

As proven at page 3830, the Navier-Stokes equation can be written as:

$$\rho\left[\partial_t\vec{v}+(\vec{v}\circ\vec{\nabla})\vec{v}\right]=\rho\vec{f}-\vec{\nabla}p \tag{9.7.544}$$

Let us consider a stationary and incompressible fluid[38] (hydrostatics), where velocity $\vec{v}=0$ everywhere, and all time derivatives are zero $\left(\frac{\partial}{\partial t}=0\right)$. The equation simplifies dramatically to the hydrostatic equation:

$$\vec{\nabla}p=\rho\vec{f} \tag{9.7.545}$$

---

[38] Oceans and seas and rivers are obviously not stationary!

For reminder, that tells us that any internal pressure gradient must perfectly balance the external body forces.

In any standard real-world scenario, the primary body force is gravity. Gravity is a conservative force field, which means it can be expressed as the gradient of a scalar potential function, φ :

$$\vec{f}=-\vec{\nabla}\varphi \tag{9.7.546}$$

Hence:

$$\vec{\nabla}p=-\rho\vec{\nabla}\varphi \tag{9.7.547}$$

For uniform gravity near Earth's surface, $\varphi=gz$, making $\vec{f}=-g\hat{\vec{z}}$.

Now, let's take the curl $(\vec{\nabla}\times)$ of both sides of the equation. A fundamental identity in vector calculus states that the curl of any gradient is zero ($\vec{\nabla}\times(\vec{\nabla}f)=\vec{0}$) (see section Vector Calculus page 1361).

So:

$$\vec{\nabla}\times(\vec{\nabla}p)=\vec{\nabla}\times(-\rho\vec{\nabla}\varphi) \tag{9.7.548}$$

Therefore (as $\vec{\nabla}\times(\vec{\nabla}f)=\vec{0}$) :

$$\vec{0}=-\rho\vec{\nabla}\times(\vec{\nabla}\varphi) \tag{9.7.549}$$

Hence using chain rule:

$$\vec{0}=-\vec{\nabla}\rho\times(\vec{\nabla}\varphi)-\rho(\vec{\nabla}\times\vec{\nabla}\varphi) \tag{9.7.550}$$

Given the identity ($\vec{\nabla}\times(\vec{\nabla}f)=\vec{0}$) and eliminating the -1 factor:

$$\vec{0}=\vec{\nabla}\rho\times(\vec{\nabla}\varphi) \tag{9.7.551}$$

As this proof is for fluids (a gas, a fluid and a plasma are fluids), we reality show us the density is not constant when the fluid is submitted to a gravitation potential. Hence taking the case of compressible fluid the term $\vec{\nabla}\rho$ is not zero!

For $\vec{\nabla}\rho\times(\vec{\nabla}\varphi)$ to be zero the gradient of density $(\vec{\nabla}\rho)$ must be parallel to the gradient of the potential $(\vec{\nabla}\varphi)$. This proves that density layers must align perfectly with the equipotential surfaces under our previous considerations that are based on our observations of reality, i.e.:

$$\rho=\rho(\varphi) \tag{9.7.552}$$

So we have to rewrite:

$$\vec{\nabla}p=-\rho\vec{\nabla}\varphi \tag{9.7.553}$$

as:

$$\vec{\nabla}p=-\rho(\varphi)\vec{\nabla}\varphi \tag{9.7.554}$$

It follows by the chain rule that pressure must also be a strict function of the potential, i.e.:

$$p=p(\varphi) \tag{9.7.555}$$

This means that surfaces of constant potential (equipotentials) are automatically surfaces of constant pressure (isobars) and constant density (isopycnals). So we can obviously rewrite:

$$\vec{\nabla}p(\varphi)=-\rho(\varphi)\vec{\nabla}\varphi \tag{9.7.556}$$

Let's look specifically at the top surface of our stationnary compressible fluid, exposed to the atmosphere!

The atmosphere exerts a uniform ambient pressure (p_{atm}) across the entire open surface. Therefore, the boundary condition for the free surface is:

$$p(x,y,z)=p_{\text{atm}}=c^{te} \tag{9.7.557}$$

Because we established that p is strictly a function of $\varphi(p=p(\varphi))$, if p is constant across the surface, then φ must also be constant across the surface:

$$\varphi(x,y,z)=C^{te} \tag{9.7.558}$$

Where for reminder and by definition, a surface where $\varphi=C^{te}$ is an "equipotential surface".

If we take a small displacement vector $\mathrm{d}\vec{r}=(\mathrm{d}\vec{x},\mathrm{d}\vec{y},\mathrm{d}\vec{z})$ tangent to the fluid (gas, liquid, plasma) surface, the change in potential along the surface must be zero:

$$\vec{\nabla}\varphi\circ\mathrm{d}\vec{r}=0 \tag{9.7.559}$$

Because the dot product is zero, the force vector $\vec{f}$ (which points in the direction of $-\nabla\varphi$ by definition) is strictly perpendicular to the fluid surface at every single point!

If the surface were tilted or wavy, $\vec{\nabla}\varphi\circ\mathrm{d}\vec{r}$ would not be zero. Gravity would have a parallel component along the slope, creating a shear stress that forces the fluid to flow until it flattens out.

Mathematically, a conclusion, a fluid surface will be perfectly level if:

- We have an inviscid fluid (with no viscosity)
- We have a stationary fluid (that never happens in reality)
- The pressure gradient is aligned with gravity (that almost never happens in reality)
- Isobaric lines matches with equipotential of the gravitational field

On a localized scale, this is presented as a very good apparent flat plane with human eye (but not with precision instruments as we would always measure at tiny curvature even in an Olympic swimming pool for example). On a planetary scale, this equipotential surface curves with the planet's mass, forming almost a perfect spherical geoid.

9.7.2.6 Hydrostatic Pressure

We have previously demonstrated without difficulty that:

$$P = -z\rho g - \frac{1}{2}\rho v^2 - c^{te} \tag{9.7.560}$$

If the fluid velocity is zero:

$$P = -z\rho g - c^{te} \tag{9.7.561}$$

This gives in differential form:

$$\frac{\mathrm{d}P}{\mathrm{d}z} = -\rho g \tag{9.7.562}$$

If we measure the liquid pressure from its upper area z_0:

$$\int_{z_0}^{z} \mathrm{d}P = P = -\int_{z_0}^{z} \rho g \mathrm{d}z = -(z\rho g - z_0\rho g) + c^{te} \tag{9.7.563}$$

If we take z_0 as a reference, we can put that:

$$h = z_0 - z \tag{9.7.564}$$

hence:

$$P = c^{te} + \rho g h \tag{9.7.565}$$

If we find ourselves in the case of a receptacle filled with a fluid in contact with the atmosphere, in order to calculate the pressure in this fluid at a given height h, we should take in consideration the atmospheric pressure, which also is "put" on the fluid. Thus, the "hydrostatic pressure" is given by:

$$P = P_0 + h\rho g \tag{9.7.566}$$

The consequence of this relation is that in a liquid at rest, homogeneous, the equipotential gravific lines are confused with isobaric surfaces. Otherwise, there would be transverse movements.

From this relation we can calculate the famous "water pump suction limit". Using:

$$h_{\max} = \frac{P - \overbrace{P_0}^{=0}}{\rho \cdot g} = \frac{P}{\rho g} = 10.34\,[\mathrm{m}] \tag{9.7.567}$$

The reader must keep in mind that the tube is only subjected to a partial vacuum (just like a real-world pump produces), the pressure differential is less, meaning that the water is "raised" or "lifted" up the tube to a reduced height of typically 7 to 8 meters.

Remark

In the technical field study of pumps, tires or compressed air tanks, we measure the pressure with "manometers" whose zero corresponds to the atmospheric pressure.

Finally here are some common drag coefficient values (most model names are based on the Gregorian calendar):

C_D	**Item**
0.001	Laminar flat plate parallel to the flow (Re<10^6)
0.005	Turbulent flat plate parallel to the flow (Re>10^6)
0.021	F-4 Phantom II (subsonic)
0.022	Learjet 24
0.024	Boeing 787
0.0265	Airbus A380
0.022	Cessna 172/182
0.031	Boeing 747
0.044	F-4 Phantom II (supersonic)
0.048	F-104 Starfighter
0.075	Pac-car
0.095	X-15 (not confirmed)
0.1	Smooth sphere (Re=10^6)
0.15	Schlörwagen Model 1939
0.18	Mercedes-Benz T80 Model 1939
0.19	General Motors EV1 Model 1996
0.212	Tatra 77A Model 1935
0.24	Tesla Model S
0.25	Toyota Prius (4th Generation)
0.26	BMW i8
0.28	Dodge Charger Daytona Model 1969
0.29	Mazda3 Model 2007
0.295	bullet (not ogive, at subsonic velocity)
0.3	Audi 100 C3 Model 1983
0.324	Ford Focus Mk2/2.5 Model 2004-2011
0.35	Maserati Quattroporte V Model 2003–2012
0.36	Citroen CX Model 1974-1991
0.48	Volkswagen Beetle
0.58	Jeep Wrangler TJ Model 1997-2005
0.75	Typical model rocket
1	Coffee filter, face-up
$1-1.3$	Wires and cables or typical adult human
1.28	Flat plate perpendicular to flow
$1.3-1.5$	Empire State Building
$1.8-2.0$	Eiffel Tower

Table 9.35: **Drag coefficient C_D examples (source: Wikipedia)**

9.7.2.7 Archimedes' principle

The Archimedes' principle, almost worldwide known phenomenon ..., is often rebellious to the first intuition. By the way, we have too much tendency in schools to present the Archimedes force as a "principle" and this wrongly because a simple mathematical analysis is enough to prove it.

Indeed, if we isolate an arbitrary portion ΔV of a fluid in static equilibrium, the conditions of this equilibrium are necessarily written (otherwise the volume dissociates and is no longer in static equilibrium):

$$\boxed{\vec{P}+\oint(\vec{n}p)\mathrm{d}S=\vec{0}} \tag{9.7.568}$$

where P denotes the weight ($\vec{P}=m\vec{a}$ in a first approximation ...) of equation whereas the term equation describes the resultant of the forces of pressure exerted on the surface of equation.

where P denotes the weight ($\vec{P}=m\vec{a}$ in a first approximation ...) of ΔV whereas the term $\oint(\vec{n}p)\mathrm{d}S$ describes the resultant of the forces of pressure exerted on the surface of ΔV

Each surface element $\mathrm{d}S$ therefore undergoes a force:

$$\vec{F}_{\rightarrow\mathrm{d}S}=(\vec{n}p)\mathrm{d}S \tag{9.7.569}$$

where p is the pressure applied locally on $\mathrm{d}S$. And for $\vec{n}$, we know that it is a unit vector directed normally (at the perpendicular) of $\mathrm{d}S$ but this time towards the interior of ΔS. The resultant of all these forces is denoted historically as follows:

$$\vec{F}_A=\oint(\vec{n}p)\mathrm{d}S \tag{9.7.570}$$

which therefore expresses, as you can guess, the famous "Archimedes' force" that the rest of the fluid exerts on the element. The integral bears on the whole surface (this surface is closed, hence the corresponding curvilinear integral) of the element ΔV.

The equilibrium condition therefore requires that:

$$\vec{F}_A=-\vec{P} \tag{9.7.571}$$

We easily understand that $\vec{F}_A$ is directed upwards: under the effect of the gravitational field and therefore p increases with depth.

If we replace the fluid contained in the volume with any fluid or solid object but that occupies the same volume, the Archimedes' force is not altered. Because of the relation $\vec{F}_A=-\vec{P}$ we are accustomed to say that it is equivalent to the weight of the displaced fluid.

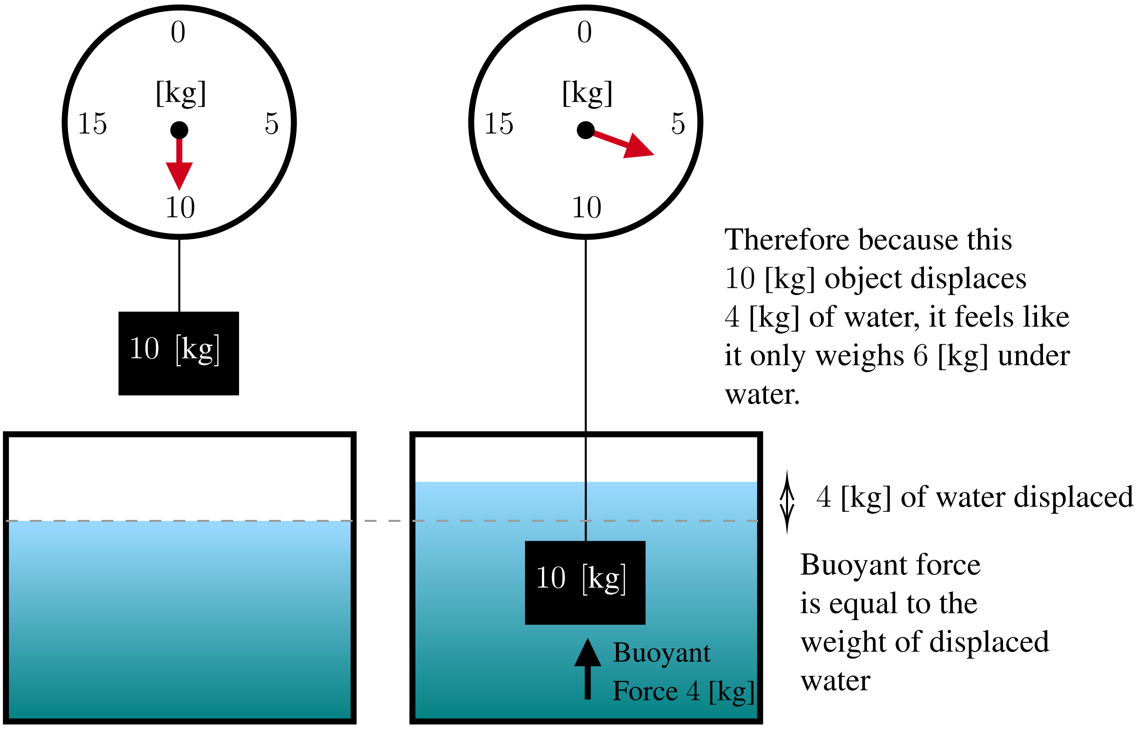

0
[kg]
15
5
10
10 [kg]
0
[kg]
15
5
10
Therefore because this 10 [kg] object displaces 4 [kg] of water, it feels like it only weighs 6 [kg] under water.
4 [kg] of water displaced
10 [kg]
Buoyant Force 4 [kg]
Buoyant force is equal to the weight of displaced water

Figure 9.259: **Archimedes principle**

In other words, for an object floating on a liquid surface (like a boat) or floating submerged in a fluid (like a submarine in water or dirigible in air) the weight of the displaced liquid equals the weight of the object.

In the case where the direction and the intensity in time of $\vec{g}$ are uniform and constant we can write:

$$F_A = -P = \int\limits_{\Delta V} \rho g \mathrm{d}V = \rho g V \tag{9.7.572}$$

and we find again the relation of the "Archimedes' law" well known to all schoolchildren:

$$\boxed{F_A = \rho g V} \tag{9.7.573}$$

There is another possibility to arrive at this proof which requires fewer mathematical tools and is therefore more affordable:

For this let us consider a cylinder of volume V immersed in a liquid at the vertical. The horizontal components of the pressure forces vanish, but the vertical component P_1 at the top of the cylinder (close to the surface) is less in intensity (except external cause) than that at its base P_2 We can therefore write:

$$F_A = (P_2 - P_1)S = (h_2 \rho g - h_1 \rho g)S = \Delta h S \rho g = V \rho g \tag{9.7.574}$$

It is a bit simpler and it keeps in a single line without integrals...

The most general and beautiful proof of the Archimedes principle relies on Gauss' gradient theorem (see section Vector Calculus page 1882) given for recall by:

$$\oiint_S f \mathrm{d}\vec{S} = \iiint_V \vec{\nabla}(f) \mathrm{d}V \tag{9.7.575}$$

That we will rewrite here as:

$$\oint_{\partial V} p(\vec{r})\mathrm{d}\vec{A} = \int_V \vec{\nabla} p(\vec{r})\mathrm{d}V \tag{9.7.576}$$

Now, as the position-dependent function we choose the pressure:

$$p(\vec{r}) = p_0 - \rho g z \tag{9.7.577}$$

where z is the vertical position coordinate and p_0 is the pressure at zero-level $(z = 0)$. We need a minus sign here, because pressure increases when going down in the liquid (i.e. in negative z direction).

Then de gradient is:

$$\vec{\nabla} p(\vec{r}) = -\rho g \hat{z} \tag{9.7.578}$$

where $\hat{z}$ is the unit-vector in z-direction (i.e. upwards).

Inserting the latter into to Gauss' gradient theorem, we get:

$$\oint_{\partial V} p(\vec{r})\mathrm{d}\vec{A} = \int_V (-\rho g \hat{z})\mathrm{d}V \tag{9.7.579}$$

Now on the left side $p\mathrm{d}\vec{A}$ obviously is the pressure force acting on the surface area element $\mathrm{d}\vec{A}$ (except for a minus sign, because the force element $\mathrm{d}\vec{F}$ points inside the body, while the area element $\mathrm{d}\vec{A}$ points outside). And on the right side, the constants ($-\rho g \hat{z}$) can be factored out. So we get:

$$-\oint_{\partial V} \mathrm{d}\vec{F} = -\rho g \hat{z} \int_V \mathrm{d}V \tag{9.7.580}$$

or finally:

$$\vec{F} = \rho g \hat{z} V \tag{9.7.581}$$

This is just Archimedes' principle (buoyant force is pointing upwards and equal to the weight of the displaced liquid)!

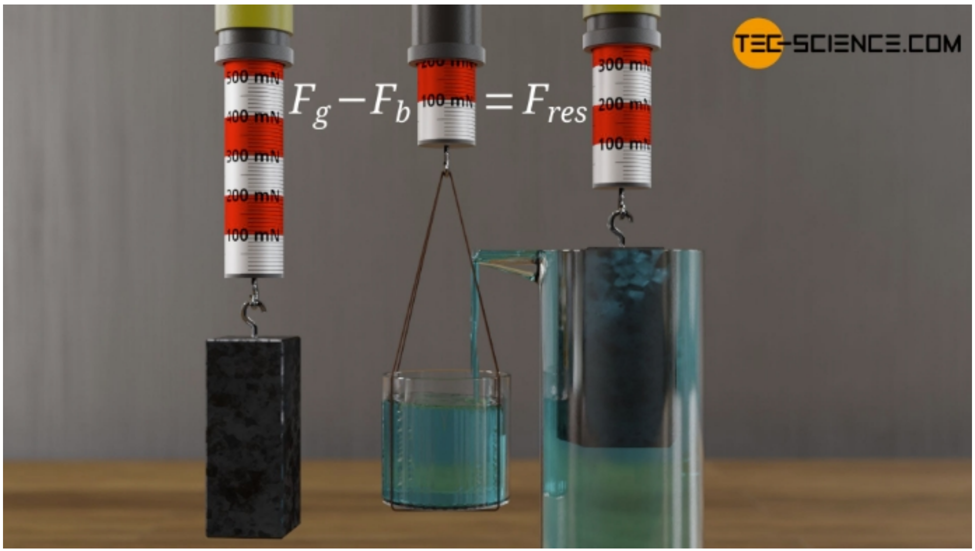


Figure 9.260: **Demonstration of the Archimedes' principle (source:** `https://www.tec-science.com`**)**

Let us derive now the Archimedes principle for a sphere with direct hydrostatic pressure calculation for the fun and the beauty!

We will need to integrate the pressure at each position to recover Archimedes law, using the fact that at a given height the pressure is $p(z)=\rho g z$ where ρ is the water's density.

First of all, we have the weight of the sphere (with density ρ_s) which gives:

$$F_w=-mg=-\rho_s\frac{4}{3}\pi R^3 g \qquad (9.7.582)$$

pointed towards the bottom (minus sign). This is the first force acting on the sphere, we will need it later.

On the other hand, the force due to hydrostatic pressure is:

$$p(z)=\rho g(R-z) \qquad (9.7.583)$$

where z is the height. We set $z=0$ at the equatorial plane, which will be useful later on when we go to spherical coordinates, so that $p(z=R)=0$ at the north pole of the sphere and the pressure grows towards the bottom of the sphere, reaching $p(z=-R)=\rho g 2R$ at south pole (towards the top). We can always set the z where we want as this simply yields an extra constant external pressure p_0 which balances on the top and bottom, so we neglect it.

The pressure acting on the sphere is directed radially (it's actually directed isotropically, but only the radial part acts on the sphere).

This means that the force due to water acting on an infinitesimal surface of the sphere $\mathrm{d}A$ is given by:

$$\mathrm{d}F=p(z)\mathrm{d}A \qquad (9.7.584)$$

and we can write:

$$\mathrm{d}A=R^2\sin(\theta)\mathrm{d}\theta\mathrm{d}\phi \qquad (9.7.585)$$

using spherical coordinates. θ is the angle with the vertical, as usual convention, with $\theta=0$ being the north pole and $\theta=\pi$ the south pole.

Because we are only interested in the z direction we only consider the part of the force directed along the z direction (the other components disappear by symmetry as they all are on the same z level).

So we only need:

$$\mathrm{d}F_z=\mathrm{d}F\cos(\theta) \qquad (9.7.586)$$

where, as before, θ is the angle with respect to the vertical director. At $\theta=0$ (north pole) the force pushes towards the bottom.

Now we integrate on the sphere only along θ as the force does not depend on ϕ. The integral on ϕ gives us a 2π factor:

$$F_z=\int\limits_0^{2\pi}\int\limits_0^{\pi}\mathrm{d}F_z=\int\limits_0^{2\pi}\int\limits_0^{\pi}\mathrm{d}F\cos(\theta)=2\pi\int\limits_0^{\pi}p(z)\mathrm{d}A(\theta)\cos(\theta) \qquad (9.7.587)$$

where $\mathrm{d}A$ has already been integrated in ϕ.

Putting all together and using $z=R\cos(\theta)$ and[39] $\mathrm{d}A(\theta)=R^2\sin(\theta)\mathrm{d}\theta$ (see section Differential and Integral Calculus page 1524) we get:

$$\begin{aligned}F_z&=2\pi\int_0^{\pi}\rho g(z-R)R^2\sin(\theta)\cos(\theta)\mathrm{d}\theta\\&=2\pi\int_0^{\pi}\rho g(R\cos(\theta)-R)R^2\sin(\theta)\cos(\theta)\mathrm{d}\theta\end{aligned}\tag{9.7.588}$$

which we then rewrite as:

$$F_z=2\pi\int_0^{\pi}\rho gR(\cos(\theta)-1)R^2\sin(\theta)\cos(\theta)\mathrm{d}\theta\tag{9.7.589}$$

and so:

$$F_z=2\pi R^3\rho g\int_0^{\pi}(\cos(\theta)-1)\sin(\theta)\cos(\theta)\mathrm{d}\theta\tag{9.7.590}$$

The integral is not that hard. Indeed, let us substitute $u=\cos(\theta)\longrightarrow\mathrm{d}u=-\sin(\theta)\mathrm{d}\theta$, therefore:

$$\begin{aligned}\int_0^{\pi}(\cos(\theta)-1)\sin(\theta)\cos(\theta)\mathrm{d}\theta&=-\int_0^{\pi}(u-1)u\mathrm{d}u=-\int_0^{\pi}\left(u^2-u\right)\mathrm{d}u=-\int_0^{\pi}u^2\mathrm{d}u+\int_0^{\pi}u\mathrm{d}u\\&=-\frac{u^3}{3}+\frac{u^2}{2}\bigg|_0^{\pi}=-\frac{\cos^3(\theta)}{3}+\frac{\cos^2(\theta)}{2}\bigg|_0^{\pi}\\&=\left(\frac{1}{3}+\frac{1}{2}\right)-\left(-\frac{1}{3}+\frac{1}{2}\right)=\frac{2}{3}\end{aligned}\tag{9.7.591}$$

So we get:

$$\int_0^{\pi}(\cos(\theta)-1)\sin(\theta)\cos(\theta)\mathrm{d}\theta=\frac{2}{3}\tag{9.7.592}$$

and this means:

$$F_z=2\pi R^3\rho g\int_0^{\pi}(\cos(\theta)-1)\sin(\theta)\cos(\theta)\mathrm{d}\theta=\frac{4}{3}\pi R^3\rho g\tag{9.7.593}$$

Which means that the total force acting on the sphere, is:

$$F_b=F_w+F_z=-\rho_s\frac{4}{3}\pi R^3+\frac{4}{3}\pi R^3\rho g=(\rho-\rho_s)\,\frac{4}{3}\pi R^3\tag{9.7.594}$$

---

[39] Remember that the spherical surface element is $\mathrm{d}A=R^2\sin(\theta)\mathrm{d}\theta\mathrm{d}\phi$ but as the force doesn't depend of ϕ the term $\mathrm{d}\phi$ is unnecessary.

i.e. Archimedes law: the sphere weight minus the equivalent weight in water!

It should be remembered that the Archimedes principle is a force which applies to fluids and therefore also to gases. It is thus thanks to Archimedes' force that a hot air balloon or an airship can rise in the air (in both cases, a gas of lower density than air is used, eg. heated air or helium).

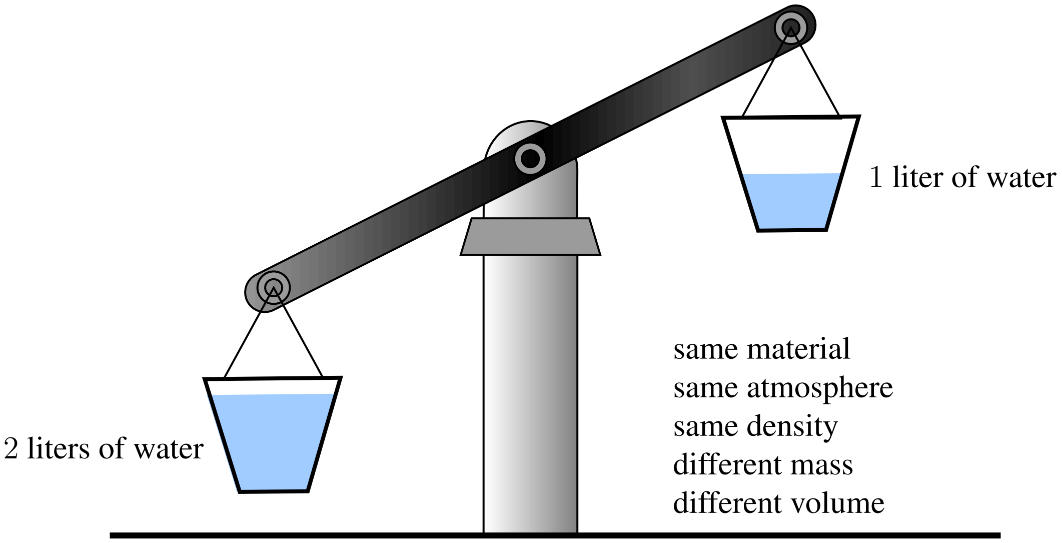
Figure 9.261: **The hot air inside these balloons is less dense than the surrounding cool air. This results in a buoyant force that causes the balloons to rise when their guy lines are untied (source: Anthony Quintano)**

It is also funny, after the proof of the law of perfect gases (see below), to determine the pressure that our atmosphere should have to have the same density as water and so that a human can then float in the air...

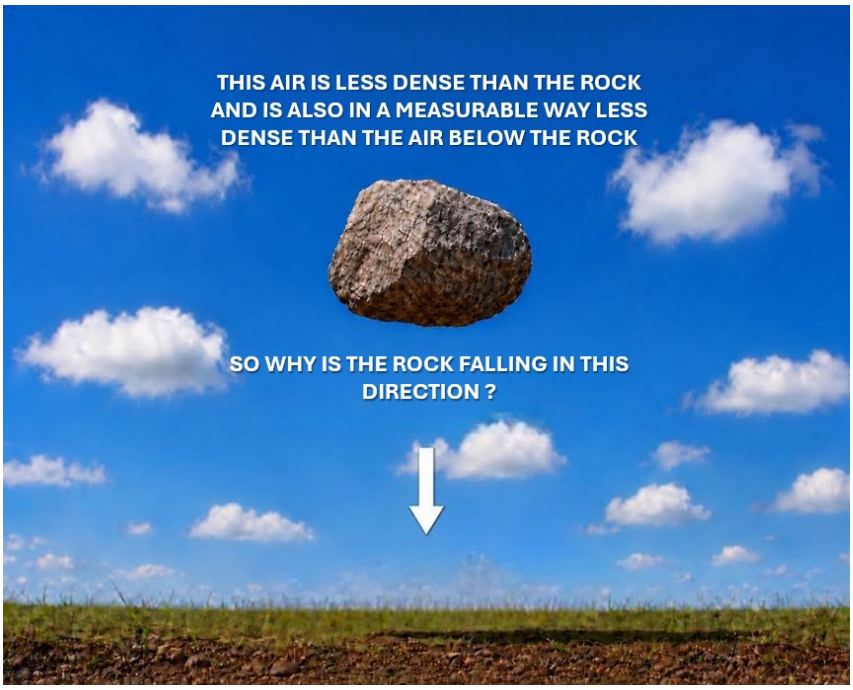
The weightings that are made on a balance will, of course, give the weight in air in most laboratories. When an object displaces its volume in air, it will be buoyed up by the weight of air displaced according to Achimedes's principle. The density of air is 0.0012 [g] (1.2 [mg]) per millilitre. If the density of the weights and the density of the object being weighed are the same, then they will be buoyed up by the same amount, and the recorded weight will be equal to the weight in a vacuum, where there is no buoyancy. If the densities are markedly different, the differences in the buoyancies will lead to a small error in the weighing: One will be buoyed up more than the other, and an unbalance will result. Such a situation arises in the weighing of very dense objects [e.g., platinum vessels (density = 21.4) or mercury (density=13.6)] or light, bulky objects [e.g., water (density $\cong 1$)]; and in very careful work, a correction should be made for this error.

Note that in most cases, a correction is not necessary because the error resulting from the buoyancy will cancel out in percent composition calculations. The same error will occur in the numerator (as the concentration of a standard solution or weight of a gravimetric precipitate) and in the denominator (as the weight of the sample). Of course, all weightings must be made with the materials in the same type of container (same density) to keep the error constant. Furthermore for most objects weighed, buoyancy errors can be neglected as they are only about one part per thousand!

We know experimentally and theoretically that object falls down because of Earth's gravitational attraction (in the point of view of Newtonian gravity). But some uneducated people claim that an object falls down due to its density, not gravity.


1 liter of water
2 liters of water
same material
same atmosphere
same density
different mass
different volume

By putting two objects of an equal density & different mass against each other using a balance like in the figure above, it is easy to see that the object's density is not the reason it falls down. Or more absurd as depicted in the image below:


THIS AIR IS LESS DENSE THAN THE ROCK AND IS ALSO IN A MEASURABLE WAY LESS DENSE THAN THE AIR BELOW THE ROCK
SO WHY IS THE ROCK FALLING IN THIS DIRECTION ?

9.7.2.8 Surface tension

"Surface tension" is the tendency of liquid surfaces at rest to shrink into the minimum surface area possible. Surface tension is what allows objects with a higher density than water such as razor blades and insects (e.g. water striders) to float on a water surface without becoming even partly submerged.

Because of the relatively high attraction of water molecules to each other through a web of hydrogen bonds, water has a higher surface tension (72.8 millinewtons [mN] per meter at $20\ [^\circ C]$) than most other liquids. Surface tension is an important factor in the phenomenon of capillarity.

In other words, from an energetic point of view, two interacting molecules always try without external influence to decrease the fluid's energy locally (see section Analytical Mechanics page 3152). Molecules at the surface of the liquid have less opportunity to lower their energy through bonds than molecules in the center of the liquid. The surface of the liquid is therefore a more energetic area than the interior. As any system seeks to have a minimum energy, this surface is reduced to a minimum. In weightlessness, the liquid forms a sphere, which offers a minimal surface area at constant volume (see section Geometric Shapes page 2993). On Earth, under the effect of gravity, the liquid forms a drop, and depending on its affinity with the surface on which it is placed, it spreads more or less, forming a dome, slightly flattened due to gravity.

The energy difference therefore generates forces that keep the fluid in a given geometric configuration; these are the capillary forces that are tangent to the surface. If we consider a portion of the surface, this force is exerted on the contour of the surface and keeps it in a given position. Hence, the surface tension has the dimension of force per unit length and is defined by:

$$\boxed{\gamma=\frac{F_T}{L}} \tag{9.7.595}$$

or of energy per unit area (it is also sometimes denoted T instead of γ). The two are equivalent, but when referring to energy per unit of area, it is common to use the term surface energy, which is a more general term in the sense that it applies also to solids.

For really large volumes of water (if you, for instance, sucked all the water out of the oceans and placed it somewhere far away in space in the standard mad-scientist way), then you also get a sphere, but for a different reason: the mass of water wants to minimize its (self-)gravitational potential energy and this is also done when it is spherical (see section Geometric Shapes page 2993). If such a volume is in the presence of external gravitational field (for instance if it was orbiting the Earth) then it would not be completely spherical: this is one of the reasons the Moon has a slightly odd shape, for instance.

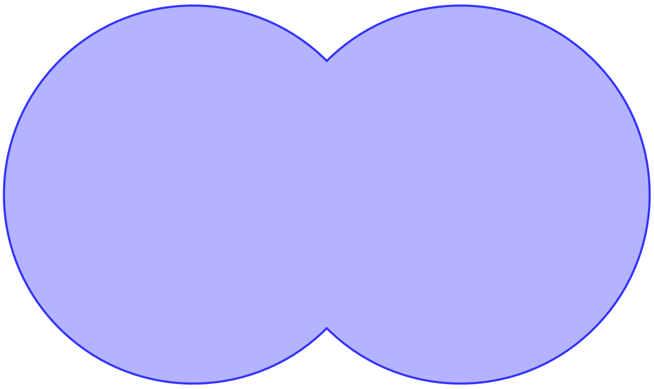
It is interesting to try to quantify the differences between the effects. One way to do this is to consider a spherical ball of water (or anything else, but we will stick to water because it is a very common asked case) and consider what force we would need to bisect the sphere and move the two halves apart. Then we can compute the force that would be needed to break the surface tension, and that needed to overcome the gravitational attraction of the two halves.

Let the radius of the ball be R, and surface tension be γ. So the total force we need to exert when splitting the sphere is simply the total force exerted by surface tension around a circumference of the sphere, and we can see immediately that this goes like R:

$$F_T = 2\pi R\gamma \tag{9.7.596}$$

Let us calculate now the equivalent force caused by gravity.

If we think about the surface which cuts the initial sphere into two hemispheres, then what is preventing the ball collapsing inwards across this surface at equilibrium (if we omit surface tension!) is pressure. So the gravitational force between the two halves of the initial sphere, when they are touching, must be equal to the integral of the pressure over that surface:



Let's assume the density is uniform, which it won't be for really large objects but it will be for reasonably small ones. Then we can work out the gravitational acceleration at radius r from the centre, relying on the shell theorem (see section Astronomy page 5281) and knowing the mass inside r is $m(r) = 4/3\pi r^3$:

$$g(r) = \frac{4\pi}{3}G\rho r \tag{9.7.597}$$

And this gives us the pressure at r, just by integrating g from r to R :

$$p(r_0) = \int\limits_{r_0}^{R} \rho g(r)\mathrm{d}r = \frac{4\pi}{3}G\rho^2 \int\limits_{r_0}^{R} r\mathrm{d}r = \frac{2\pi}{3}G\rho^2\left[R^2 - r_0^2\right] \tag{9.7.598}$$

or:

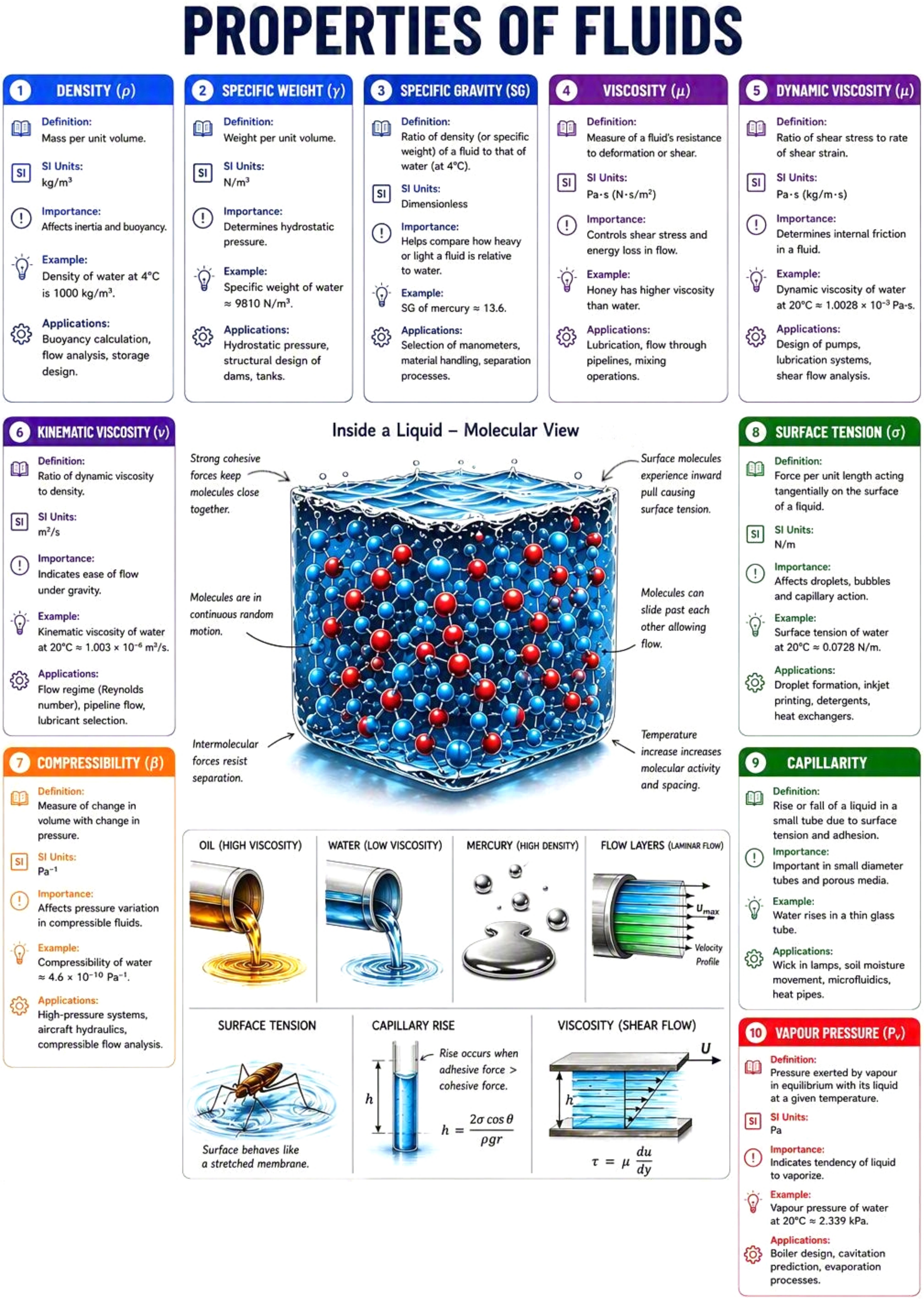
$$p(r) = \frac{2\pi}{3}G\rho^2\left[R^2 - r^2\right] \tag{9.7.599}$$

And finally we can integrate this over the surface to get the total force:

$$F_G = \int\limits_{0}^{R} 2\pi r p(r)\mathrm{d}r = \frac{4\pi^2}{3}G\rho^2 \int_0^R R^2 r - r^3\mathrm{d}r = \frac{\pi^2}{3}G\rho^2 R^4 \tag{9.7.600}$$

So, then, given $\rho = 10^3\,[\mathrm{kg}\cdot\mathrm{m}^{-3}]$ (density of water at ambiant temparature), $G = 6.7\cdot 10^{-11}\,[\mathrm{m}^3\cdot\mathrm{kg}^{-1}\cdot\mathrm{s}^{-2}]$, we can solve for the radius R where $F_T = F_G$, and the answer is about $R \cong 12.8\,[\mathrm{m}]$.

So if this is right, it means that gravity starts beating surface tension for a ball of water which is about 13 [m] in radius.



PROPERTIES OF FLUIDS

1 DENSITY (ρ)

Definition: Mass per unit volume.

SI Units: kg/m^3

Importance: Affects inertia and buoyancy.

Example: Density of water at 4°C is 1000 kg/m^3.

Applications: Buoyancy calculation, flow analysis, storage design.

2 SPECIFIC WEIGHT (γ)

Definition: Weight per unit volume.

SI Units: N/m^3

Importance: Determines hydrostatic pressure.

Example: Specific weight of water $\approx$ 9810 N/m^3.

Applications: Hydrostatic pressure, structural design of dams, tanks.

3 SPECIFIC GRAVITY (SG)

Definition: Ratio of density (or specific weight) of a fluid to that of water (at 4°C).

SI Units: Dimensionless

Importance: Helps compare how heavy or light a fluid is relative to water.

Example: SG of mercury $\approx$ 13.6.

Applications: Selection of manometers, material handling, separation processes.

4 VISCOSITY (μ)

Definition: Measure of a fluid's resistance to deformation or shear.

SI Units: Pa·s (N·s/m^2)

Importance: Controls shear stress and energy loss in flow.

Example: Honey has higher viscosity than water.

Applications: Lubrication, flow through pipelines, mixing operations.

5 DYNAMIC VISCOSITY (μ)

Definition: Ratio of shear stress to rate of shear strain.

SI Units: Pa·s (kg/m·s)

Importance: Determines internal friction in a fluid.

Example: Dynamic viscosity of water at 20°C $\approx 1.0028 \times 10^{-3}$ Pa·s.

Applications: Design of pumps, lubrication systems, shear flow analysis.

6 KINEMATIC VISCOSITY (ν)

Definition: Ratio of dynamic viscosity to density.

SI Units: m^2/s

Importance: Indicates ease of flow under gravity.

Example: Kinematic viscosity of water at 20°C $\approx 1.003 \times 10^{-6}$ m^2/s.

Applications: Flow regime (Reynolds number), pipeline flow, lubricant selection.

7 COMPRESSIBILITY (β)

Definition: Measure of change in volume with change in pressure.

SI Units: Pa^{-1}

Importance: Affects pressure variation in compressible fluids.

Example: Compressibility of water $\approx 4.6 \times 10^{-10}$ Pa^{-1}.

Applications: High-pressure systems, aircraft hydraulics, compressible flow analysis.

Inside a Liquid – Molecular View

Strong cohesive forces keep molecules close together.

Surface molecules experience inward pull causing surface tension.

Molecules are in continuous random motion.

Molecules can slide past each other allowing flow.

Intermolecular forces resist separation.

Temperature increases molecular activity and spacing.

OIL (HIGH VISCOSITY)

WATER (LOW VISCOSITY)

MERCURY (HIGH DENSITY)

FLOW LAYERS (LAMINAR FLOW)

SURFACE TENSION

Surface behaves like a stretched membrane.

CAPILLARY RISE

Rise occurs when adhesive force > cohesive force.

$$h = \frac{2\sigma\cos\theta}{\rho g r}$$

VISCOSITY (SHEAR FLOW)

$$\tau = \mu \frac{du}{dy}$$

8 SURFACE TENSION (σ)

Definition: Force per unit length acting tangentially on the surface of a liquid.

SI Units: N/m

Importance: Affects droplets, bubbles and capillary action.

Example: Surface tension of water at 20°C $\approx$ 0.0728 N/m.

Applications: Droplet formation, inkjet printing, detergents, heat exchangers.

9 CAPILLARITY

Definition: Rise or fall of a liquid in a small tube due to surface tension and adhesion.

Importance: Important in small diameter tubes and porous media.

Example: Water rises in a thin glass tube.

Applications: Wick in lamps, soil moisture movement, microfluidics, heat pipes.

10 VAPOUR PRESSURE (P_v)

Definition: Pressure exerted by vapour in equilibrium with its liquid at a given temperature.

SI Units: Pa

Importance: Indicates tendency of liquid to vaporize.

Example: Vapour pressure of water at 20°C $\approx$ 2.339 kPa.

Applications: Boiler design, cavitation prediction, evaporation processes.

9.7.2.9 Speed of sound in a liquid

Let us take a moment to calculate the speed of sound in a liquid. We have proved in the our study of the longitudinal sound waves of the Mathematical Music section that:

$$\frac{\mathrm{d}P}{\mathrm{d}\rho}=-\gamma\frac{P}{\rho_0} \tag{9.7.601}$$

where, for recall, γ is the "Laplace coefficient", also named "adiabatic coefficient" defined by:

$$\gamma=\frac{C_P}{C_V} \tag{9.7.602}$$

and we had proved that the speed of the sound wave was given by:

$$v=\sqrt{\frac{\gamma}{\rho_0}} \tag{9.7.603}$$

Combining it comes:

$$v=\sqrt{\frac{1}{-\dfrac{\mathrm{d}\rho}{\mathrm{d}P}}}=\sqrt{\frac{1}{-\dfrac{\mathrm{d}\left(\dfrac{m}{V}\right)}{\mathrm{d}P}}}=\sqrt{\frac{1}{-\dfrac{-\dfrac{m}{V^2}\mathrm{d}V}{\mathrm{d}P}}}=\sqrt{\frac{1}{\dfrac{m}{V}\dfrac{\dfrac{\Delta V}{V}}{\Delta P}}}=\sqrt{\frac{1}{\rho\dfrac{\dfrac{\Delta V}{V}}{\Delta P}}}=\sqrt{\frac{B}{\rho}} \tag{9.7.604}$$

The fraction:

$$\boxed{B=-\frac{\Delta P}{\Delta V}V} \tag{9.7.605}$$

that is to say the ratio between a variation in pressure and the relative variation in volume which it entails is named the "volumetric modulus of elasticity". Notice that the sign $-$ is required for B to be positive: when the pressure increases, the volume decreases!

We have for example for water:

$$v=\sqrt{\frac{1}{\dfrac{\rho}{B}}}\cong\sqrt{\frac{1}{\dfrac{1000}{2\cdot 10^9}}}=1,414\,[\mathrm{m}\cdot\mathrm{s}^{-1}] \tag{9.7.606}$$

The measured value being in real life in standard condition of temperature and pressure: $1,441\,[\mathrm{m}\cdot\mathrm{s}^{-1}]$. It may seem surprising that the velocity of sound in a liquid, which is much more difficult to compress than a gas is only 5 times greater than in a gas. The reason is that the density of a liquid is about a thousand times higher than that of a gas. In both, the two properties partially compensate.

9.7.2.10 Summary of fluid dynamics laws

Here is a summary of the gas laws proven so far we have seen so far on request of a few readers (for reminder these laws assume ideal fluid behavior, i.e. particles have no volume and exert no intermolecular forces on one another and temperatures are always in Kelvin):

- **Continuity Law:** Governs that mass entering a system must equal mass leaving the system plus any accumulation. For an incompressible fluid, the volumetric flow rate remains constant across variable cross-sections

$$\frac{\partial\rho}{\partial t}+\nabla\circ(\rho\vec{u})=0 \tag{9.7.607}$$

- **Navier-Stokes Law:** Describes how the velocity field of a viscous fluid changes under the influence of internal forces (pressure and viscosity) and external forces (gravity)

$$\rho_{\mathrm{d}}\vec{v}=-\vec{\nabla}p+\frac{4}{3}\left(\mu_d+\mu'_d\right)\vec{\nabla}(\vec{\nabla}\circ\vec{v})-\mu_d\vec{\nabla}\times(\vec{\nabla}\times\vec{v})+\rho\vec{f} \tag{9.7.608}$$

- **Bernoulli's Law:** States that for an inviscid, incompressible, and irrotational flow along a streamline, an increase in fluid speed occurs simultaneously with a decrease in static pressure or potential gravitational energy

$$p+\frac{1}{2}\rho v^2+\rho gh=c^{te} \tag{9.7.609}$$

- **Archimedes' Law:** Any body completely or partially submerged in a fluid at rest is acted upon by an upward, or buoyant, force proportional to the weight of the fluid displaced by the body and by the local acceleration of gravity:

$$F_B=\rho gV_{\text{displaced}} \tag{9.7.610}$$

- **Pascal's Law:** A pressure change applied to an enclosed, incompressible fluid at rest is transmitted undiminished throughout the fluid and to the walls of its container

$$\Delta p=\rho g\Delta h \tag{9.7.611}$$

- **Newton's viscosity Law:** Dictates that the shear stress between adjacent fluid layers is directly proportional to the negative velocity gradient perpendicular to the direction of flow (fluids adhering to this are classified as "Newtonian")

$$\tau=\mu\frac{\partial u}{\partial y} \tag{9.7.612}$$

- **Hagen-Poiseuille Law:** Calculates the pressure drop in an incompressible, Newtonian fluid undergoing laminar flow through a long cylindrical pipe of constant cross-section

$$J=\frac{\pi R^4}{8\mu_d}\frac{\Delta P}{l} \tag{9.7.613}$$

- **Stokes' Drag Law:** Gives the frictional force-also known as drag force-exerted on spherical objects moving at very ow Reynolds numbers (creeping flow) through a viscous fluid

$$\Delta p = \frac{8\mu LQ}{\pi R^4} \tag{9.7.614}$$

- **Torricelli's Law:** An application of Bernoulli's principle that determines the efflux velocity of a liquid leaving an open orifice under the sole action of gravity

$$v = \sqrt{2gh} \tag{9.7.615}$$

- **Isoperimetric Law:** Any fluid under its own mass will tend to a collapse into a sphere to minimize its energy state (see the section of Geometry for the detailed proof)

$$4\pi\mathcal{A}(\text{Int}(\gamma)) \leq \mathcal{L}(\gamma)^2 \tag{9.7.616}$$

- **Levelling Law:** A stationary compressible inviscid fluids in a gravitational potential and having an isobaric surface will always be at level (see definition of "level" at page 5260)

$$\vec{0} = \vec{\nabla}\rho \times (\vec{\nabla}\varphi) \Leftrightarrow \nabla\rho \parallel \vec{\nabla}\varphi \tag{9.7.617}$$

The reader must keep in mind there are plenty of other gas laws but that we didn't prove so fare in this textbook!

9.7.3 Gas

The solids have a well defined shape and are difficult to compress. Liquids can flow freely and their flow is limited by self-formed surfaces. The gases dilate freely to occupy the volume of the receptacle containing them, and have a density about a thousand times lower than that of liquids and solids. They conduct heat and electricity only a little, unless we ionize them (forming then plasma). The molecules of a neutral gas move along rectilinear trajectories which change direction with each collision with another molecule. Unlike solids and liquids, interactions between molecules remain weak. The macroscopic properties of a gas are thus directly deduced from the properties of the molecules that compose it (or atoms in the case of a monoatomic gas).

Definition (#9.7.475): A "gas" is a diluted fluid, in the sense that its state of matter characterized by the fact that it doesn't have a fixed shape or volume (a gas will expand to fill the entire volume of its container), it can be compressed (meaning their volume can be decreased under pressure), the molecules in the gas can move freely and are widely spaced compared to liquids and solids, the behavior of the gas is significantly influenced by changes in temperature and pressure.

9.7.3.1 Types of Gas

In gas theory (we often speak of "kinetic theory of gases") we always consider two types of neutral gases:

9.7.3.1.1 Perfect Gas

The "perfect gas" is a model in which we neglect the molecular interactions of the gas, with the exception of collisions, and whose own volume is negligible relatively to the volume of the container. We deduce from these considerations that the behaviour of a real gas can be approached by a perfect gas if the volume it occupies is sufficiently large, which will occur at low pressure and at high temperature. In practice, the perfect gas model gives satisfactory results for many conventional gases taken at atmospheric pressure and at ambient temperature (nitrogen, oxygen, steam, etc.).

When a gas is at low pressure, the interactions between its molecules are weak. Thus, the properties of a real gas at low pressure approximate those of a perfect gas. We can then describe the behaviour of the gas by the "equation of state of perfect gases" that we will prove in detail further below during our study of the virial theorem (see section Continuum Mechanics page 3897):

$$\boxed{PV = nRT} \tag{9.7.618}$$

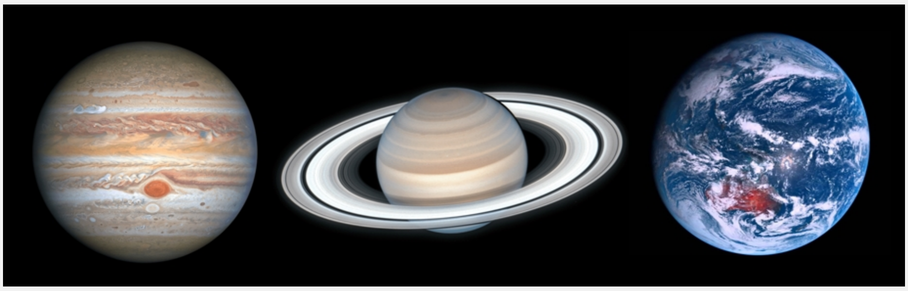
With n the number of moles of gas, P the gas pressure, V the volume occupied by the n moles and T the absolute temperature of the gas. The constant R being the "constant of the perfect gases":

$$R = 8.3144598(48)\,[\mathrm{J}\cdot\mathrm{mol}^{-1}\cdot\mathrm{K}^{-1}] \tag{9.7.619}$$

Remark

The pressure of a gas is defined as the force the gas would exert upon a surface or container. However, there is no need of a container for pressure to exist! For instance, the air you are breathing right now (unless you're in an airplane or submarine) has pressure due to the column of atmosphere above you that exist due to Earth's gravity. Stars are balls of gas (plasma, actually) that are also pressurized by gravity; no containers to be seen!

We can have a vacuum environment next to a pressurized system without a solid barrier between them due to gravity which is scientifically proven by multiple independent empirical observations:



On Earth itself we also don't need a closed container to keep a gas that has a molar mass much greater than air, like sulfur hexafluoride (SF_6) that has a molar mass of $146.06\,[\text{g}\cdot\text{mol}^{-1}]$ considerably higher than the density of oxygen O_2 with its $31.998\,[\text{g}\cdot\text{mol}^{-1}]$ or even CO_2 with its $44.009\,[\text{g}\cdot\text{mol}^{-1}]$. It is therefore allowing experiments like the one visible in the photo below, where soap bubble are suspended over a tank (container without top):

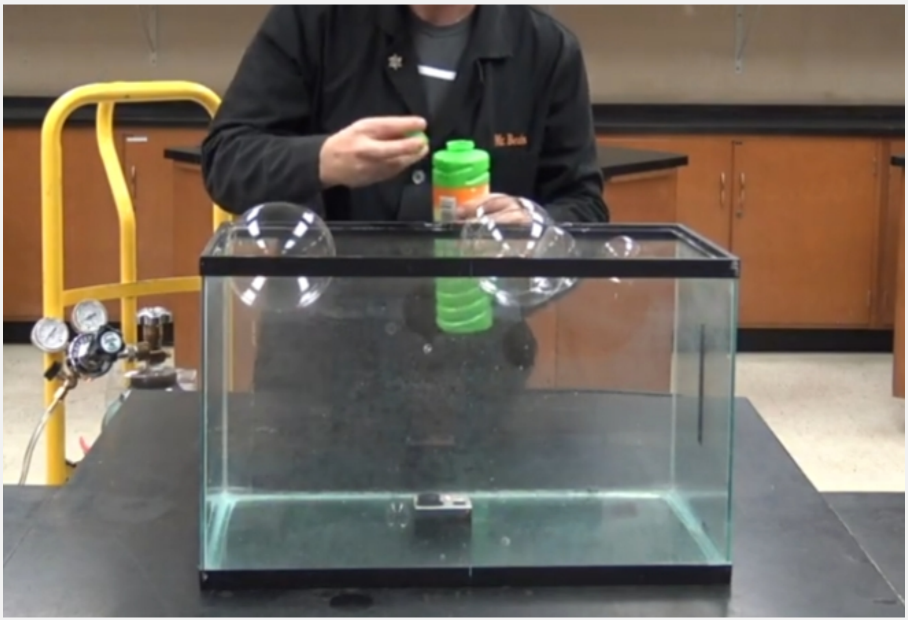


Figure 9.262: Example of a gas (SF_6) contained in a container without top (author: Craig Beals, source: `https://www.youtube.com/watch?v=KHmNiOsTYfA`)

This equation shows trivially that:

- At constant temperature T ("isothermal" system), the volume of a fixed quantity of an ideal gas is inversely proportional to its pressure (that exclude the range of temperatures and pressures where the gas behaves like a liquid or like a solide obviously). This is the "Boyle-Mariotte's law":

$$\boxed{V=\frac{c^{te}}{P}} \tag{9.7.620}$$

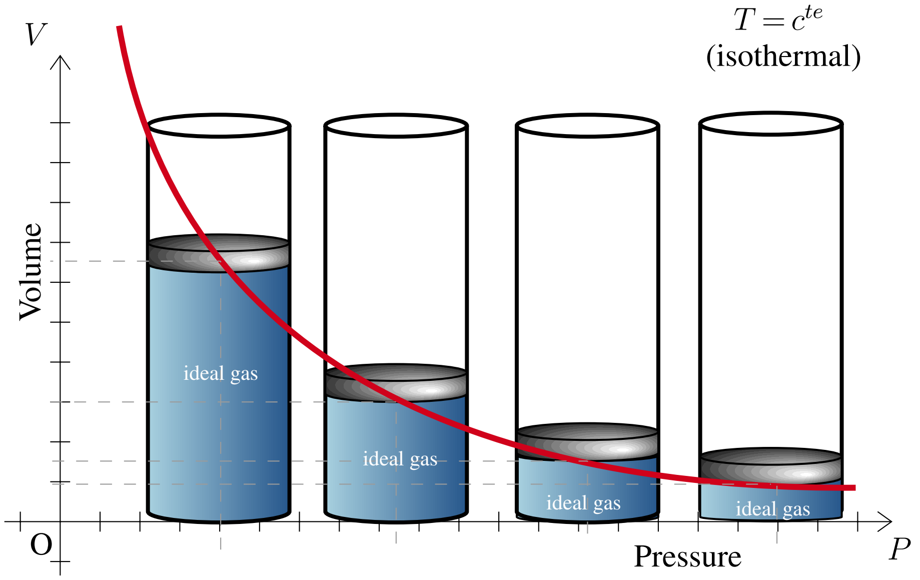


Figure 9.263: Boyle-Mariotte's law illustration

From the Boyle-Mariotte's law, we deduce immediately that at constant temperature (i.e. during an isothermal transformation):

$$\boxed{P_1V_1=P_2V_2\Leftrightarrow\frac{P_1}{P_2}=\frac{V_2}{V_1}} \tag{9.7.621}$$

A real PV diagram taken from an old handbook for Oxygen is given below for people interested (the reader can notice that when the temperature is too low... the gas doesn't behave like an ideal gas anymore and at some very low temperature it isn't a gas at all anymore!):

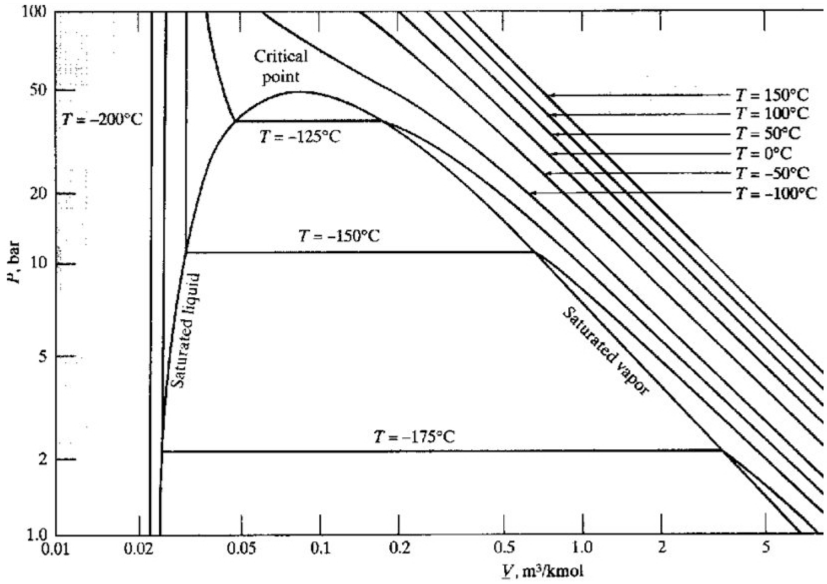


Figure 9.264: PV **diagram for oxygen**

- At constant pressure P ("isobaric" system), the volume of a fixed quantity of ideal gas is proportional to the absolute temperature (therefore and again... this doesn't include the ranges of pressure and temperatures where the gas isn't a gas anymore!). This is the "Charles' law" (in the case of perfect gases ...):

$$\boxed{V = c^{te}T} \tag{9.7.622}$$

Obviously the Kelvin scale must be used because zero on the Kelvin scale corresponds to a complete stop of molecular motion!

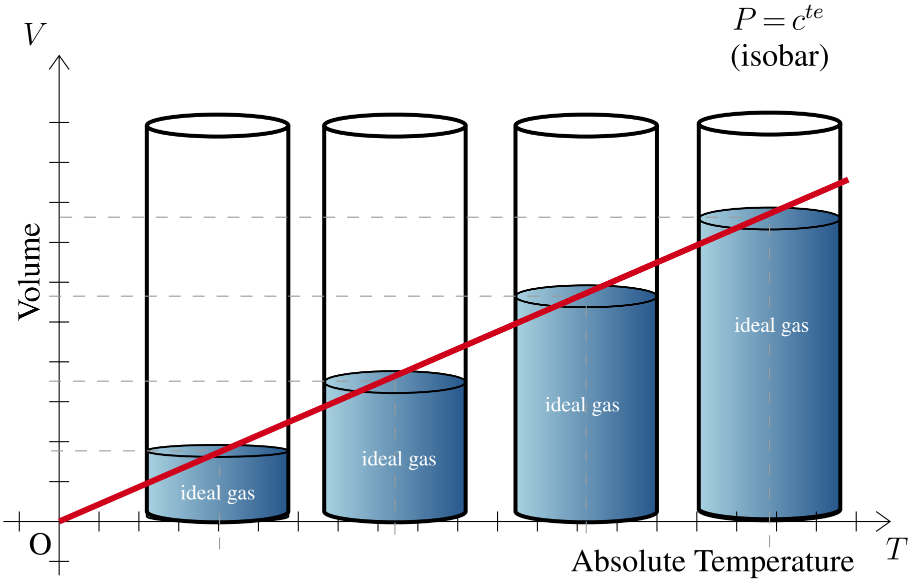


Figure 9.265: **Charles' law illustration**

The experiment is performed at a constant atmospheric pressure. The experiment proceeds by placing an empty flask in a boiling water bath. As the temperature increases, the air inside the flask expands. Afterwards, the gas is cooled in a water bath by maintaining the amount of the air in the flask constant. By measuring/determining the initial and final temperature and volume, we verify Charles' law.

Notice that the line goes exactly toward the origin, meaning that as the absolute temperature of the gas approaches zero, its volume approaches zero. However, when a gas is brought to extremely cold temperatures, its molecules would eventually condense into the liquid state before reaching absolute zero. The temperature at which this change into the liquid state occurs varies for different gases.

From the Charles's law, we deduce immediately that at constant pressure (i.e. during an isobaric transformation):

$$\boxed{\frac{V_1}{T_1}=\frac{V_2}{T_2}\Leftrightarrow V_1T_2=V_2T_1} \tag{9.7.623}$$

It is this relation which is often used in the labs of small classes to show that with an extrapolation of the measured line, the volume theoretically becomes zero at a temperature of about -273.15 [°C]. Notice by the way seriously from a certain temperature, it is necessary to use quantum models and moreover this relation is valid only for the gases (thus, when the water vapour becomes liquid ... it is no longer valid).

Let also recall that we have proved a more elaborate expression of the Gay-Lussac's law in the section of Thermodynamics by making use of equations of state and thermoelastic coefficients ...:

$$\Delta V=\alpha_V V_0\Delta T \tag{9.7.624}$$

This show us that the Gay-Lussac's law then no longer really have the status of "law" but rather that of a "relation" as in can be proved! We have also proved in the section of Thermodynamics that this law is only approached to the high pressures for the gases since they must be considered as incompressible like liquids to obtain the relation in question.

- At constant volume V ("isochore" system), the pressure of a fixed quantity of gas is proportional to its absolute temperature. This is the "Gay-Lussac's law":

$$\boxed{P=c^{te}T} \tag{9.7.625}$$

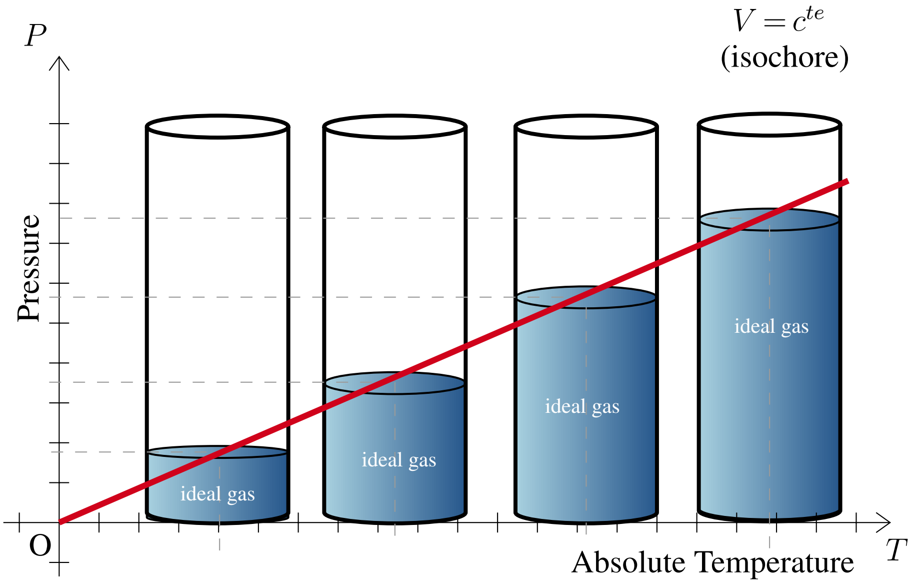


Figure 9.266: **Gay-Lussac's law illustration**

From the Gay-Lussac's law, we deduce immediately that at constant volume (i.e. during an isochoric transformation):

$$\boxed{\frac{P_1}{T_1}=\frac{P_2}{T_2}\Leftrightarrow P_1T_2=P_2T_1} \tag{9.7.626}$$

The reader may have noticed that the underling idea of Gay-Lussac's law is very similar to Charles's law. Indeed, the only difference being the type of container. Whereas the container in a Charles's law experiment is flexible, it is rigid in a Gay-Lussac's law experiment!

Subsequently, Amedeo Avogadro, to whom we owe the word of "molecule" affirms the molecular concept of gases and concludes that equal volumes of different gases, taken under the same conditions of temperature and pressure, contain the same number of molecules (this observation now follows automatically from the equation of perfect gases which will be proved further below in details). This property was named the "Avogadro and Ampere law" and is often illustrated as follow:

$$\boxed{V=nc^{te}} \tag{9.7.627}$$

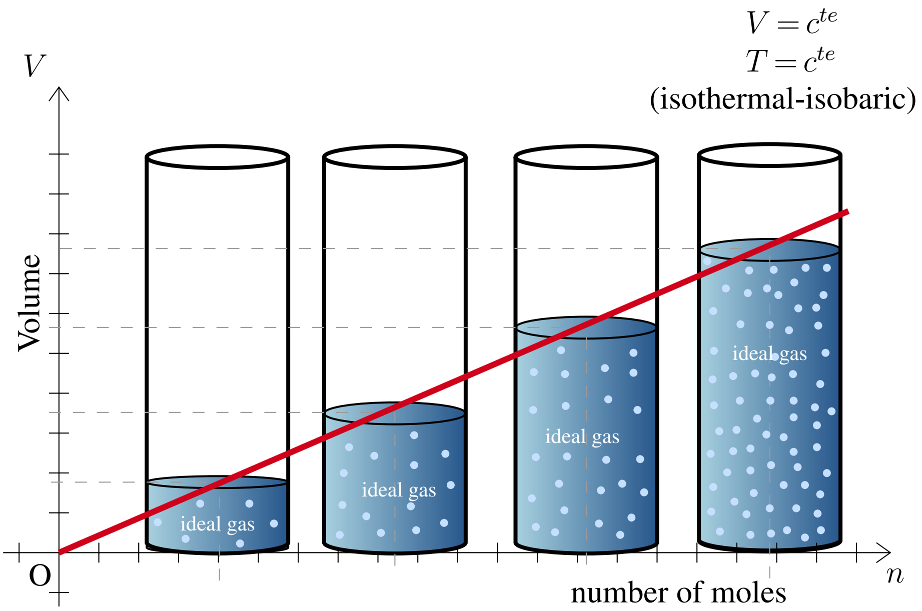


Figure 9.267: **Avogadro's law illustration**

In other words, Avogadro's Law states that the volume of a gas is directly proportional to the number of moles (or number of particles) of gas when the temperature and pressure are held constant.

From the Avogadro's law, we deduce immediately that at constant temperature and constant volume we then have:

$$\boxed{\frac{V_1}{n_1} = \frac{V_2}{n_2} \Leftrightarrow V_1 n_2 = V_2 n_1} \qquad (9.7.628)$$

If the container holding the gas is rigid rather than flexible, pressure can be substituted for volume in Avogadro's Law. Adding gas to a rigid container makes the pressure increase.

✍**Example:**

The air in a diving cylinder is initially at the pressure P_i in a volume V_i corresponding to that of the diving cylinder that we assume to be $0.015\,[\mathrm{m}^3 \cdot \mathrm{s}^{-1}]$. A diver never breathes air directly into the bottle because of the pressure that is about 200 times higher than the atmospheric pressure $20.02 \cdot 10^7$ [Pa] would blow his lungs. A valve makes it possible to adjust the pressure of the air sent to the plunger so that it is equal to the ambient pressure P_t of the water depth. It is interesting to calculate the time t that the diver can stay underwater at a depth of 10 meters if it is assumed to consume 0.03 cubic meters per minute.

To answer this question we must therefore first know the pressure P_f at 10 meters depth. For this we will use the Hydrostatic theorem proved during our study of fluids:

$$P_f = P_{\text{atm}} + \rho_{\text{water}} g h = 1.01 \cdot 10^5 + 1025 \cdot 9.81 \cdot 10 \cong 2.01 \cdot 10^5 \text{ [Pa]} \qquad (9.7.629)$$

Afterwards the volume in the bottle should be set to the above pressure volume using the

Boyle-Mariotte's law:

$$P_iV_i = P_fV_f \tag{9.7.630}$$

We then have as equivalent volume:

$$V_f = \frac{P_iV_i}{P_f} \cong \frac{2.02 \cdot 10^7 \cdot 0.015}{2.01 \cdot 10^5} = \frac{2.02 \cdot 10^7 \cdot 0.015}{2.01 \cdot 10^5} = 1.507\,[\text{m}^3] \tag{9.7.631}$$

But the trap is that the usable volume is smaller because it still requires air to push the remaining air. We must then subtract the two volumes:

$$V_{\text{usable}} = V_f - V_i = 1.507 - 0.015 = 1.492\,[\text{m}^3] \tag{9.7.632}$$

And dividing by a standard volume consumption rate C we have at the end:

$$t = \frac{V_f - V_i}{C} = \frac{\dfrac{P_iV_i}{P_f} - V_i}{C} = \frac{\dfrac{P_iV_i}{P_{\text{atm}} + \rho_{\text{water}}gh} - V_i}{C} = \frac{1.492\,[\text{m}^3]}{0.03\,\left[\text{m}^3 \cdot \text{min}^{-1}\right]} = 49.7\,[\text{min}] \tag{9.7.633}$$

So to summarize so far:

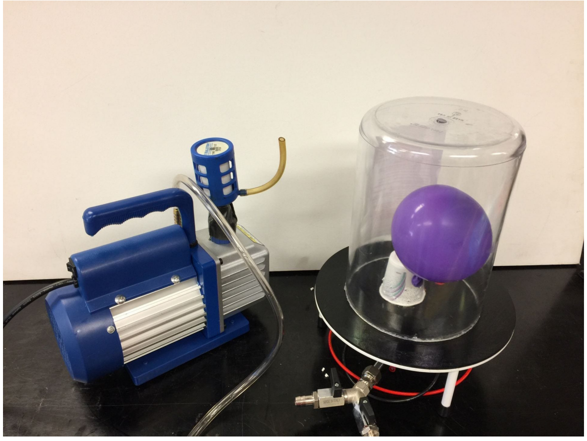
Gas Law	**Formula**
Boyle's law	$P_1V_1 = P_2V_2$
Charles's law	$\frac{V_1}{T_1} = \frac{V_2}{T_2}$
Gay-Lussac's law	$\frac{P_1}{T_1} = \frac{P_2}{T_2}$
Avogadro's law	$\frac{V_1}{n_1} = \frac{V_2}{n_2}$
Combined law	$\frac{PV_1}{T_1} = \frac{P_2V_2}{T_2}$
Ideal gas law	$PV = nRT$

Table 9.36: **Summary of ideal gas laws**

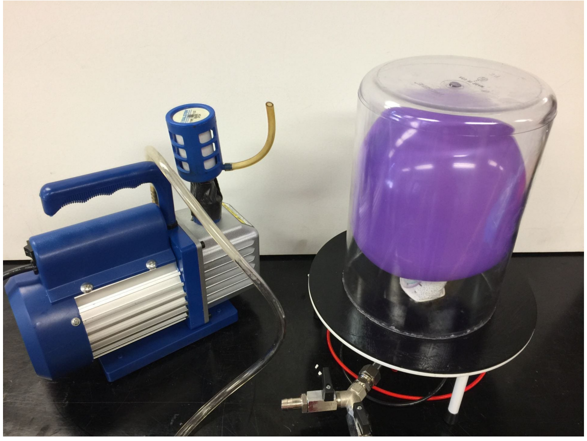
Caution! Collisions between gas particles or collisions with the walls the container are perfectly elastic! None of the energy of gas particles is lost when it collides with another particle or with the walls of the container. More generally, they do lose or gain energy when they crash into walls. If the walls are the same temperature as the gas they will on average gain as much as the lose. If the walls are colder they will lose more than they gain heating the walls. If the wall is hotter than the gas they will gain more than they lose heating the gas.

Let us illustrate now the effects of a change in atmospheric pressure, especially the expansion in a Vacuum as it often misunderstood by some children (and even adults...)!

Below we can see a latex balloon that rests on a stand in a bell-jar. The balloon expands when the air is removed from the jar by a vacuum pump:



(a) Latex ballon filled with air in the in the bell-jar filled with air

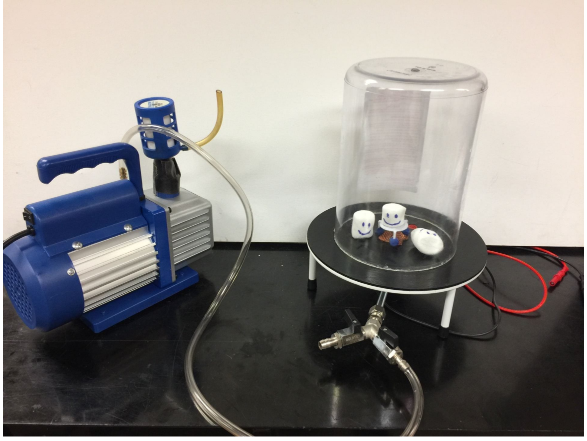


(b) Latex balloon filled with air in pseudo-vacuum bell-jar

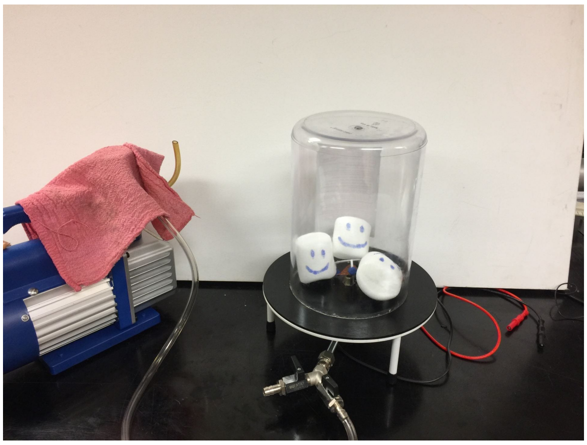
The latex balloon will not float in air, since the latex balloon is only filled with air, and hence, not light enough by Archimedes' Principle to float. In this case, the latex balloon rests on a stand that is inside of the bell-jar. The latex balloon is tied off, which makes the latex balloon behave as a closed system, and exerts a constant pressure from the inside of the latex balloon in the outward, radial direction.

When the latex balloon is in the bell-jar full of air, the balloon is at equilibrium, since the pressure exerted by the air inside of the balloon is equal to the pressure of the atmosphere that is acting on the outside of the balloon. When the atmosphere (air) inside of the bell-jar is pumped out, the system is no longer in equilibrium, since the atmospheric pressure in the bell-jar is decreasing, while the pressure inside of the latex balloon remains constant. The pressure inside of the latex balloon is greater than the atmospheric pressure, and the latex balloon with start to expand, until it reaches equilibrium.

Now the same experiment with an object having a stronger molecular structure:



(a) Marshmallows in the bell-jar, filled with air



(b) Marshmallows in pseudo-vacuum bell-jar

Marshmallows have little pockets of air that are trapped inside of them. When the marshmallows is in the bell-jar full of air, the pressure exerted by the air pockets inside of the marshmallows are in equilibrium with the atmospheric pressure inside of the bell-jar. As the atmospheric pressure in the bell-jar decreases due to air being sucked out by the vacuum pump, the system is no longer in an equilibrium state, and the pressure coming from the air pockets inside of the marshmallow is greater than the decreasing atmospheric pressure. The pressure of the air pockets cause the marshmallows to expand less significantly than the balloon.

Now let us come back to the equation of state of perfect gases:

$$PV = nRT \tag{9.7.634}$$

9.7.3.1.2 Real Gas

The perfect gas equation is obviously approximate. For example, a perfect gas could neither liquefy nor solidify, regardless of the cooling and compression to which it is subjected. Real gases, especially under conditions of pressure and temperature close to the transition to the liquid state, can present considerable deviations with the law of perfect gases!

It must therefore be adapted to real cases. For this we use the Van der Waals equation of state which is particularly useful and well known. It can be obtained qualitatively once the perfect gas equation is demonstrated and is then given by (see below how we prove it):

$$\boxed{\left(P+\frac{a}{V^2}\right)(V-b)=RT} \tag{9.7.635}$$

for one mole, a and b being adaptable parameters determined by experimental measurements carried out on the concerned gas (a is a measure of the average attraction between particles, b the volume excluded by a mole of particles). These are parameters that vary from one gas to another.

The Van der Waals equation can also be interpreted at the microscopic level. Molecules interact with each other. This interaction is strongly repulsive for the molecules close to one another, becomes slightly attractive for a mean distance and disappears when the distance is important. At high pressure, the perfect gas law must be rectified to take into account attractive or repulsive forces.

9.7.3.2 Virial Theorem

We will here discuss a study of perfect gases via a particular method. It makes it possible to obtain an interesting result and particularly for astrophysics (see section Astrophysics page 5539). The virial theorem[40] also makes it possible to obtain other very interesting results but which pedagogically are somewhat difficult to access. The reader who would be interested in this second part of the results can directly go back a little further where the concepts of pressure and kinetic temperature are treated.

By definition, the "virial expression $V_{\text{ir.}}$" of a material point is the scalar:

$$\boxed{V_{\text{ir.}} = \vec{F}\circ\vec{r}} \tag{9.7.636}$$

---

[40] The word *virial* derives from *vis*, the Latin word for *force* or *energy*, and was given its technical definition by Rudolf Clausius in 1870.

By definition, the "virial" of a system composed of N material points is:

$$V_{\text{ir.}}=\sum_{i=1}^{N}\vec{F}_i\circ\vec{r}_i \tag{9.7.637}$$

Subjected to a central force, the virial is written (by the properties of the scalar product):

$$V_{\text{ir.}}=\sum_{i=1}^{N}F_i r_i \tag{9.7.638}$$

The "virial theorem" states: For a system in equilibrium (!), the internal energy is equal to the opposite of its total half-virial when all the particles are located with respect to its center of mass.

Proof 9.10.6. Given the differentiation (see section Differential and Integral Calculus page 1464):

$$\mathrm{d}_t r_i^2=2r_i\mathrm{d}_t r_i \tag{9.7.639}$$

Its second derivative:

$$\mathrm{d}_t^2 r_i^2=\mathrm{d}_t(2r_i\mathrm{d}_t ri)=2\mathrm{d}_t r_i\mathrm{d}_t r_i+2r_i\mathrm{d}_t^2 r_i=2(\mathrm{d}_t r_i)^2+2r_i\mathrm{d}_t^2 r_i \tag{9.7.640}$$

Multiplying by m_i and summing on i:

$$\mathrm{d}_t^2\sum_{i=1}^{N}m_i r_i^2=2\sum_{i=1}^{N}\left[m_i(\mathrm{d}_t r_i)^2+\vec{r}_i m_i\mathrm{d}_t^2 r_i\right] \tag{9.7.641}$$

But:

$$m_i\mathrm{d}_r^2 r_i=F_i \tag{9.7.642}$$

and:

$$\mathrm{d}_t r_i=v\Rightarrow(\mathrm{d}_t r_i)^2=v_i^2 \tag{9.7.643}$$

Therefore:

$$\mathrm{d}_t^2\sum_{i=1}^{N}m_i r_i^2=2\left[\sum_{i=1}^{N}m_i v_i^2+\sum_{i=1}^{N}r_i F_i\right] \tag{9.7.644}$$

This last expression is valid whatever the position of the adopted coordinate system. However, it is interesting to place its origin at the center of mass (centroid) of the system because we are no longer dependent on its movement.

If the system is in equilibrium, the macroscopic quantities which characterize it are not dependent on time. We then conclude that the sum of any quantity attached to any material point of the system is in fact a quantity of that system.

Thus, $\sum m_i r_i$ is a macroscopic quantity independent of time. That implies that:

$$\sum m_i v_i^2+\sum r_i F_i=0 \tag{9.7.645}$$

What is written again (we multiply by $1/2$ on both sides):

$$\frac{1}{2}\sum m_i v_i^2 = -\frac{1}{2}\sum r_i F_i \qquad (9.7.646)$$

So we have finally:

$$E_c = \frac{1}{2}V_{\text{ir.}} \qquad (9.7.647)$$

This expression of kinetic energy is known as "virial theorem" and the right-hand side is therefore named the "virial of the system".

■ Q.E.D.

We write that:

$$E_c = N\langle E_c\rangle \qquad (9.7.648)$$

where E_c is the total kinetic energy associated with all the material points of the system. We name it the "internal energy of the system" and thermodynamicists often denote it with the letter U as we know it...! The term $\langle E_c\rangle$ is the energy of any material point in the system.

It is possible to fall back on the expression of the virial from a system of particles (cloud in accretion). Strictly speaking, equilibrium does not exist in such a case. Nevertheless, we can assume that if the gravitational contraction is sufficiently slow then its different phases can be considered as a succession of equilibrium states (typical thermodynamic approach that we already know).

In the case of a central force and derived from a potential, we can write (see section Classical Mechanics page 3351):

$$F = -\frac{\partial E_p}{\partial r} \qquad (9.7.649)$$

and therefore:

$$Fr = -r\frac{\partial E_p}{\partial r} \qquad (9.7.650)$$

If the potential energy is of the form k/r (which is the case for the electrical and gravitational potential) then it comes:

$$Fr = E_p \qquad (9.7.651)$$

and it remains:

$$\boxed{E_c = -\frac{1}{2}E_p} \qquad (9.7.652)$$

sometimes denoted and rewritten (especially in Astrophysics):

$$2T + U = 0 \qquad (9.7.653)$$

In summary, the virial theorem gives us a relation between total kinetic and potential energies. To be valid, the mobile must describe a trajectory around the center of central force and remain indefinitely in a finite volume (bound state). This type of reasoning can be applied to a very large number of phenomena, from the structure of certain galaxies to the release of energy in nuclear explosions through the study of the Sun (actually a G2V-type star, aka yellow dwarf) and the behaviour of real gases. This is the first result we were interested in.

Remark

Astrophysicists use (veeery roughly and stupidly as the assumptions are not satisfied!) the relation:

$$2T+U=0 \tag{9.7.654}$$

for galaxy clusters.... with:

$$T=\frac{1}{2}\sum_i m_i v_i^2 \cong \frac{1}{2}M_D v^2 \tag{9.7.655}$$

where M_D is the cluster mass and:

$$U=\sum_{i,j}-\frac{1}{2}\frac{Gm_i m_j}{|\vec{r}_i-\vec{r}_j|}\cong -\frac{1}{2}\frac{M_D^2 G}{R} \tag{9.7.656}$$

where R is the estimated galaxy cluster radius. Therefore:

$$\boxed{M_D \cong \frac{2Rv^2}{G}} \tag{9.7.657}$$

and according to the bad astrophysicists that do this huge approximation, the observed M_D is not compliant with the calculated one, so there must be something missing that we can't see and that they name the "dark matter"... It should be interesting to do the calculations without so huge approximations and with a model that have assumptions that are compatible with galaxies...

Since June 12025 (holocene calendar) it is known thanks to FRB (Fast Radio Bursts) that approximately 76% of the Universe's baryonic matter lies in the IGM (InterGalactic Medium). About 15% resides in galaxy halos, and a small fraction is burrowed in stars or amid cold galactic galactic gas ([109]).

In a gaseous system, the potential energy can be written as the sum of the energy of the forces acting from the outside and those which are internal even to the gas. Such as:

$$E_p=\sum_{i=1}^N \vec{F}_i\circ\vec{r}_i=\sum_{i=1}^N \vec{F}_{i,\text{int.}}+\sum_{i=1}^N \vec{F}_{i,\text{ext.}} \tag{9.7.658}$$

But the internal forces can be written as:

$$\vec{F}_{i,\text{int.}}=\sum \vec{F}_{ij} \tag{9.7.659}$$

It is not necessary in this sum to take the force exerted by each of the particles upon itself. Such as:

$$\vec{F}_{i,\text{int.}}=\sum_{i\neq j} \vec{F}_{ij} \tag{9.7.660}$$

What give us:

$$E_p=\sum_{i=1}^N \vec{F}_i\circ\vec{r}_i=\sum_{j\neq i} \vec{F}_{ij}\circ\vec{r}_i+\sum_{i=1}^N \vec{F}_{i,\text{ext.}} \tag{9.7.661}$$

In the double sum, we can group the terms two by two and use the principle of action-reaction such as:

$$\vec{F}_{ij}=-\vec{F}_{ij} \tag{9.7.662}$$

To get by definition:

$$\vec{F}_{ij}\circ r_i+\vec{F}_{ji}\circ\vec{r}_j=\vec{F}_{ij}(\vec{r}_i-\vec{r}_j):=\vec{F}_{ij}\circ\vec{r}_{ij} \tag{9.7.663}$$

What ultimately gives us:

$$E_p=\sum_{i\neq j}\vec{F}_{ij}\circ\vec{r}_{ij}+\sum_{i=1}^{N}\vec{F}_{i,\text{ext.}}\circ\vec{r}_i \tag{9.7.664}$$

And:

$$E_c=\frac{1}{2}E_p=\frac{1}{2}\left(\sum_{i\neq j}\vec{F}_{ij}\circ\vec{r}_{ij}+\sum_{i=1}^{N}\vec{F}_{i,\text{ext.}}\circ\vec{r}_i\right) \tag{9.7.665}$$

The first term on the right involves the internal forces (interactions) between the (pairs of) particles and the second term on the right involves external forces.

Let us consider now a gas contained in a tank. Its molecules are subject to external forces only when they hit a wall and we imagine that on average this force is perpendicular to the wall (elastic shocks).

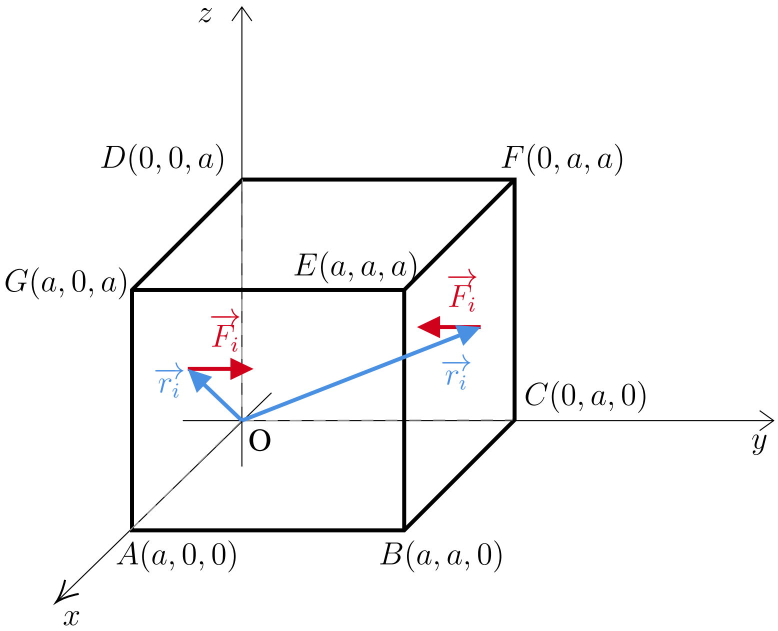


Figure 9.270: **Gas contained in a tank**

For all the faces contained in the planes defined by the axes, we always have:

$$\vec{r}_i\circ\vec{F}_i=0 \tag{9.7.666}$$

Since on average $\vec{F}_i$ is always perpendicular to $\vec{r}_i$.

For the other faces ($BCFE$ for example) we have:

$$\vec{F}_i = -F_i\vec{u}_y \qquad (9.7.667)$$

and thus:

$$\vec{r}_i \circ \vec{F}_i = r_i F_i \cos(\alpha) = -aF_i \qquad (9.7.668)$$

where we denote by a the coordinate according to Oy of the end of $\vec{r}_i$ (not to be confused with acceleration!). Therefore for each face:

$$\sum_{i=1}^{N} \vec{F}_i \circ \vec{r}_i = -a\sum_{i=1}^{N} F_i = -aF = -Pa^3 = -PV \qquad (9.7.669)$$

Indeed because the internal pressure of the system on the walls is defined by:

$$P = \frac{F}{S} = \frac{F}{a^2} \qquad (9.7.670)$$

By the virial theorem, by adding the non-zero contributions of the faces $BCFE$, $DEFG$ and $ABED$, it comes:

$$E_c = \frac{3}{2}PV - \frac{1}{2}\sum_{j\neq 1} \vec{F}_{ij} \circ \vec{r}_{ij} \qquad (9.7.671)$$

If the average kinetic energy of a molecule is:

$$E_c = \frac{1}{2}mv^2 = \frac{3}{2}kT \qquad (9.7.672)$$

The total mean kinetic energy for N molecules is then (we will return to this relation later but with another approach):

$$\boxed{E_c = \frac{N}{2}mv^2 = \frac{3}{2}kNT = \frac{3}{2}\frac{R}{N_\text{A}}NT} \qquad (9.7.673)$$

that thermodynamicists often write:

$$U = n\frac{\text{df}}{2}RT \qquad (9.7.674)$$

where U is as we know the internal energy of the gas and df the number of degrees of freedom of the constituents.

The prior relation, named the "equipartition theorem of energy" (as it assumes that the total kinetic energy is equally distributed by each degree of freedom) is important because it allows:

1. A microscopic interpretation of temperature T and to determine the internal energy of a monoatomic perfect gas (and by extension of other gases with other degrees of freedom).
2. To notice that the temperature system in Celsius is not adapted to the physical reality. Indeed, in the system using the Celsius, at 0 [°C] everything should be immobile (zero kinetic energy) and it is obvious that this is not the case for all substances. Therefore, a new temperature must be introduced that matches the measured kinetic energy with the traditional temperature. It will be the "absolute temperature" measured in Kelvin [K] whose equivalence kinetic energy / temperature measured is such that 0 [°C] corresponds to 273.15 [K].

Internal energy is a contribution to energy that does not appear in Classical Mechanics as we already know. From the macroscopic point of view, a container at rest that contains a fluid does not possess kinetic energy, while its potential energy is constant. We can ignore it by giving the value zero to this constant.

From the microscopic point of view, however, things change! Indeed, the atoms or molecules of the fluid are in motion and interact. It is necessary to associate to them an energy (the internal energy) which is the sum of the contributions relative to each atom.

Since then:

$$\boxed{PV=\frac{1}{3}Nmv^2+\frac{1}{3}\sum_{i\neq j}\vec{F}_{ij}\circ\vec{r}_{ij}=NkT+\frac{1}{3}\sum_{i\neq j}\vec{F}_{ij}\circ\vec{r}_{ij}} \tag{9.7.675}$$

It is the "general equation of state of a real gas", that is, the equation of state which takes into account the interactions between molecules. It is interesting to notice that finally this relation can allow us to calculate the energy of the gas even if there are no walls!

If the gas is perfect, there are no interactions between the N molecules (by hypothesis) and then we fall back on the "ideal gas law" as follows:

$$PV=NkT \tag{9.7.676}$$

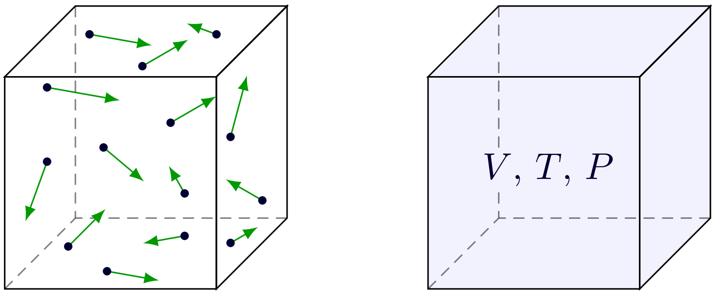
where P is the absolute pressure of the gas, N is the number of molecules in the given volume V and k (written k_B if there is a risk of confusion) is the Boltzmann constant:

$$k=1.38064852(79)\cdot 10^{-23}\,[\mathbf{J}\cdot\mathbf{K}^{-1}] \tag{9.7.677}$$

That we find much more frequently in the form:

$$\boxed{PV=NkT=n\frac{N}{n}T=nRT} \tag{9.7.678}$$

if n is expressed in moles with R the constant of the perfect gases. Obviously for $n=1$, N is equal to the number of particles in one mole (Avogadro's number: N_{A}).



Remark

If you assume air is an ideal gas (as it is often done at standard temperature and pressure), then one mole occupies at the sea level:

$$V=\frac{nRT}{P}=\frac{1\cdot 8,314\cdot 298}{101,325}\cong 0.0245\,[\text{m}^3] \tag{9.7.679}$$

Thus, each atom occupies statistically a volume of:

$$\frac{V}{N_a}=\frac{V}{6.023\cdot 10^{23}}\cong 4\cdot 10^{-26}\,[\text{m}^3] \tag{9.7.680}$$

The spacing between atoms is on the order of the cube root of that, or ~ 3 [nm] (the same calculation for the stratosphere leads to a value of ~ 6 [nm]). However this distance isn't the size of the atoms (if you assume wrongly that it is the case you would get a sectional surface of the order of 10^{-17})! Indeed, advanced calculations give for O_2 a Rayleigh cross-section of approximately $6\cdot 10^{-25}\,[\text{m}^{-2}]$ for light wavelength between 185 and 187 [nm] (see [494]). Keep in mind that Hughes and Jauncey, Phys. Rev. (36 1930), 773 have shown experimentally that for visible light, light to light scattering is $\sim 10^{-64}\,[\text{m}^{-2}]$.

If the temperature is constant, we fall back on "Boyle-Mariotte's law":

$$PV=c^{te} \tag{9.7.681}$$

Before continuing we can also come back on Avogadro statement: "equal volumes of all gases, at the same temperature and pressure, have the same number of molecules". That is to say for a given mass of an ideal gas, the volume and amount (moles) of the gas are directly proportional if the temperature and pressure are constant which can be written as:

$$V\propto n \tag{9.7.682}$$

or:

$$\frac{V}{n}=c^{te} \tag{9.7.683}$$

This law describes how, under the same condition of temperature and pressure, equal volumes of all gases contain the same number of molecules. For comparing the same substance under two different sets of conditions, the law can be usefully expressed as follows:

$$\frac{V_1}{n_1}=\frac{V_2}{n_2}=c^{te} \tag{9.7.684}$$

The most significant consequence of Avogadro's law is that the ideal gas constant has the same value for all gases. This means that:

$$\frac{P_1V_1}{T_1n_1}=\frac{P_2V_2}{T_2n_2}=c^{te} \tag{9.7.685}$$

To arrive at the perfect gas equation, it is useful to remember that we have made implicitly three hypotheses:

H1. The molecules are assimilated to hard spheres whose diameter is negligible in comparison with the average distance between them. That is what we name the "structural hypothesis".

H2. Ultimately, and this is what we have learned, if we consider molecules as punctual, the possibility of interaction between the particles vanishes. The only remaining interactions will be shocks on the walls of the tank that contains the gas. These shocks are perfectly elastic so that we can apply the laws of conservation of the momentum of kinetic energy. This is what we name the "interactive hypothesis limit"

H3. The gas is studied in a state of thermodynamic equilibrium, which results in the homogeneity of intensive and extensive variables. That is what we name the "hypothesis of molecular chaos".

In a particular case, if the interactions are derived from a central potential:

$$\vec{F}_{ij}=-\frac{\partial E_(r_{ij})}{\partial r_{ij}}\vec{u}_{ij} \tag{9.7.686}$$

We can rewrite this as follows:

$$F_{ij}r_{ij}=-r_{ij}\frac{\partial E_(r_{ij})}{\partial r_{ij}} \tag{9.7.687}$$

If in addition the potential energy is of the type (be careful not to confuse the parameter k with the Boltzmann constant denoted in the same way!):

$$E_p(r_{ij})=\frac{k}{\partial r_{ij}} \tag{9.7.688}$$

Then we have:

$$F_{ij}r_{ij}=-r_{ij}\frac{\partial E_(r_{ij})}{\partial r_{ij}}=-r_{ij}\left(-\frac{k}{r_{ij}^2}\right)=\frac{k}{r_{ij}}=E_p(r_{ij}) \tag{9.7.689}$$

And therefore:

$$PV=NkT+\frac{1}{3}\sum_{i\neq j}\vec{F}_{ij}\circ\vec{r}=NkT+\frac{1}{3}E_p \tag{9.7.690}$$

Where E_p is then mean total energy of the system.

In fact, care must be taken that we have not rigorously proved the ideal gas law. Indeed, when we had put earlier above:

$$E_c=\frac{1}{2}mv^2=\frac{3}{2}kT \tag{9.7.691}$$

This implicitly assumed that the perfect gas equation was already known (...).

This approach to gas kinetics is interesting because it is useful in astrophysics. However, this is by far not the simplest in a school context and pedagogically... We propose the reader to come back to these same results further below via the concepts of pressure and kinetic temperature once the Van der Waals equation is determined.

In 11875 (holocene calendar) the Dutch scientist J.D. Van der Waals tried to replace the perfect gas equation by a relation which would take into account the intermolecular forces and the size of the molecules. The first and most obvious correction to the equation of perfect gases:

$$PV=kNT \tag{9.7.692}$$

is to subtract the volume of the gas molecules from volume V. We can do this logically by replacing V by $V-Nb$ where b is a very small constant, which obviously represents the average volume per molecule (tables exist for this). So:

$$P=\frac{kNT}{V-Nb} \tag{9.7.693}$$

where the term Nb is commonly referred to as the "covolume".

To take account of the intermolecular forces that we have previously neglected, we can try an approximate approach by knowing already that the attraction force of each molecule will be on $N-1$ molecules. Consequently, the numerator of the attraction force will trivially contain (by the sum of all terms) if the gas is isotropic and homogeneous a term of the type $N(N-1)$ for the influence of all the molecules together which if N is very large can be approximated by N^2.

We also know that in the numerator there will be a mass term for each particle. If we know the ratio N/V then we only need to know the mass density of the gas (but this is not an extensive variable so we will avoid making it appear explicitly). Thus, the term N^2 can be directly written $(N/V)^2$.

Gas	**Chemical Formula**	**Molar mass** $[\mathbf{g\cdot mol^{-1}}]$	**Density** $[\mathbf{kg\cdot m^{-3}}]$
Air	-	28.96	1.2740
Ammonia	NH_3	17.03	0.7491
Argon	Ar	39.95	1.7572
Bio Gas	$35\%\ CO_2\ 65\%CH_4$	25.83	1.1363
Butane	C_4H_{10}	58.12	2.5567
Butylene	C_4H_8	56.11	2.4681
Carbon Dioxide	CO_2	44.01	1.9359
Carbon Monoxide	CO	28.01	1.2321
Chlorine	Cl_2	70.91	3.1205
Ethane	C_2H_6	30.07	1.3227
Ethylene	C_2H_4	28.05	1.2340
Helium	He	4.00	0.1761
Hexane	C_6H_{14}	86.18	3.7908
Hydrogen	H_2	2.02	0.0887
Methane	CH_4	16.04	0.7057
Nitrogen	NH_2	28.01	1.2323
Oxygen	O_2	32.00	1.4076
Pentane	C_5H_{12}	72.15	3.1738
Propane	C_3H_8	44.10	1.9397
Propylène	C_3H_6	42.08	1.8510
Sulfur Dioxide	CO_2	64.06	2.8191

Table 9.37: Densities of various typical gases at normal pressure and temperature

Remark

A gas can have a density higher than a liquid under certain conditions. This typically occurs at very high pressures and temperatures, where the gas reaches a state known as a supercritical fluid. In this state, the gas does not condense into a liquid but has a density similar to that of a liquid.

Following this reasoning, Van der Waals added to the right-hand side of the relation above a negative term proportional to the quantity $(N/V)^2$. The presence of this term results in a lowering of the pressure as the density of the gas increases. The modified relation is thus:

$$P \cong \frac{kNT}{V-Nb} - a\left(\frac{N}{V}\right)^2 \tag{9.7.694}$$

where a is a constant of proportionality. We can rewrite this relation in the form:

$$\boxed{\left(P + a\left(\frac{N}{V}\right)^2\right)(V-Nb) \cong kNT} \tag{9.7.695}$$

which is named the "Van der Waals state equation" or "Van der Waals real gas equation" (we find in the literature several equivalent ways of writing this last relation). It is an excellent description of the equation of state in a wide range of the variables P, V, T, the values a and b being characteristic of each gas. The constants a and b are determined experimentally as we have already mentioned.

We can represent the Van der Waals equation in a P-V diagram (hence constant temperature). Here is an example with CO_2 (carbon dioxide):

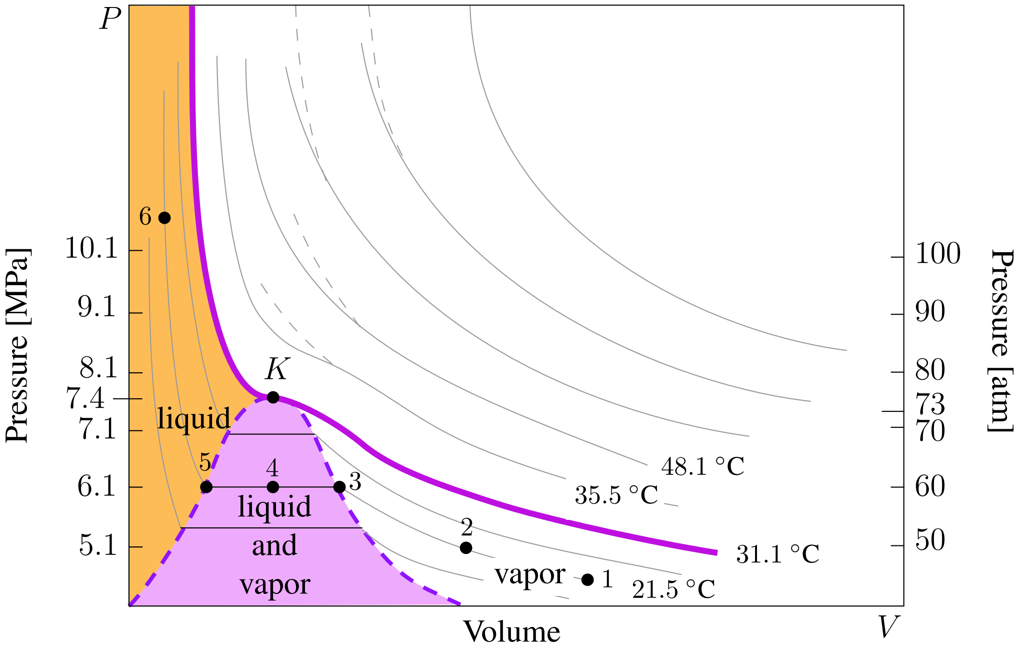


Figure 9.271: **Van der Waals P-V diagram for CO_2**

Which corresponds to:

$$P \cong \frac{kNT}{V-Nb} - a\frac{N^2}{V^2} \tag{9.7.696}$$

If we set $a=0$ and $b=0$ we see that we fall back on a famous relation...

The bottom figure shows the general case. One can pass from state F (vapour) to state I (liquid) by following a path passing through states A and E (liquefaction stage), then a phase transition is observed. If we follow another path, for example by using the isotherm $\overline{GH}$ ($T>T_k$, red path), there is no liquefaction stage, we pass continuously from the gaseous state to the liquid state:

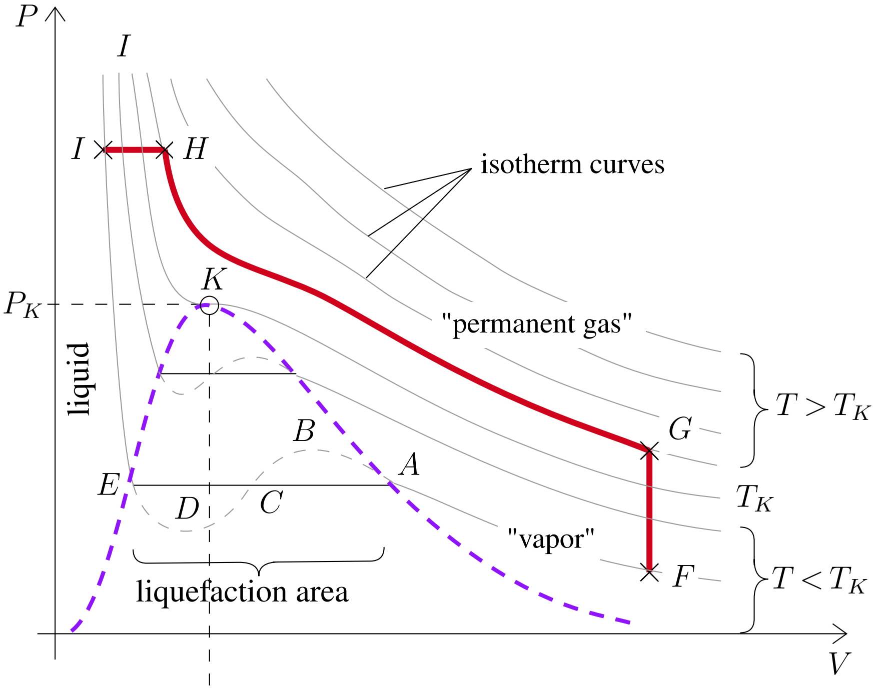


Figure 9.272: Van der Waals general P-V diagram

The Van der Waals equation can be put in the form of a virial type development using a Taylor-Maclaurin development (see section Sequences and Series page 1703):

$$P\cong\frac{kNT}{V-Nb}-a\left(\frac{N}{V}\right)^2=\frac{kNT}{V(1-\dfrac{Nb}{V})}-a\left(\frac{N}{V}\right)^2=\frac{kNT}{V(1-\dfrac{Nb}{V})}-\frac{akNT}{V}\left(\frac{N}{kTV}\right)$$
$$=\frac{kNT}{V}\left(1+\frac{Nb}{V}+\frac{N^2b^2}{V^2}+\ldots\right)-\frac{akNT}{V}\left(\frac{N}{kTV}\right)\tag{9.7.697}$$

or by rearranging the terms:

$$P\cong\frac{kNT}{V}\left(1+\left(b-\frac{a}{kT}\right)\frac{N}{V}+b^2\frac{N^2}{V^2}+\ldots\right)\tag{9.7.698}$$

Let us notice that there exists a temperature for which the term $\left(b-\dfrac{a}{kT}\right)$ is zero. We name it the "Boyle temperature" of the gas: this is the temperature at which the actual gas behaves like an ideal gas.

Form information here is for comparison a PVT diagram of the ideal gas law:

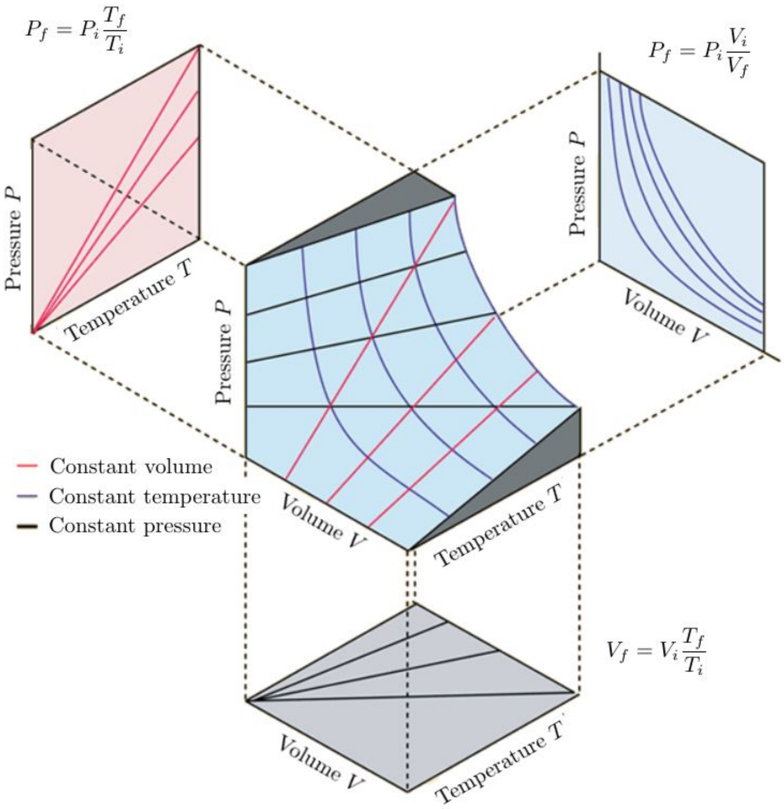

$P_f = P_i\frac{T_f}{T_i}$
Pressure P
Temperature T
$P_f = P_i\frac{V_i}{V_f}$
Pressure P
Volume V
Pressure P
Volume V
Temperature T
Constant volume
Constant temperature
Constant pressure
$V_f = V_i\frac{T_f}{T_i}$
Volume V
Temperature T

Figure 9.273: PVT **diagram of the ideal gas law**

And below a more realistic one where we can identify the Van der Waals law application domain:

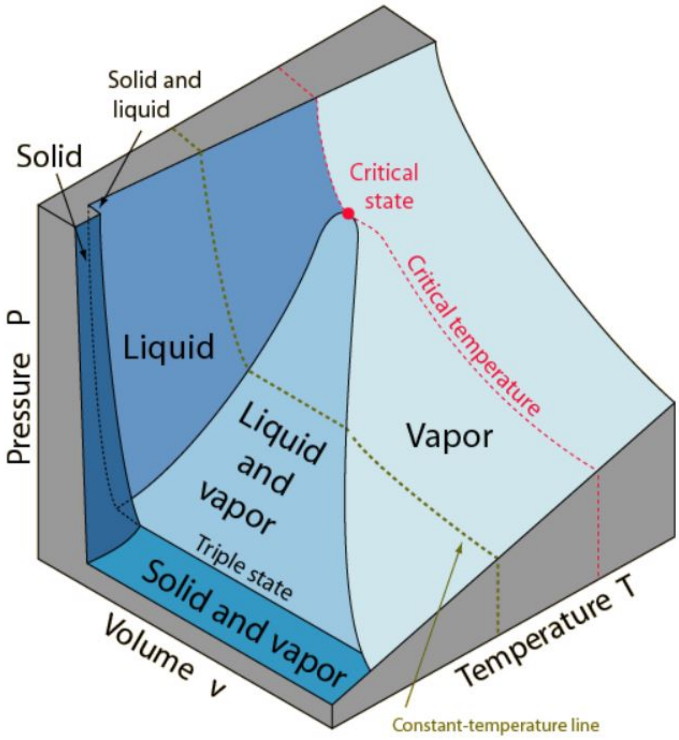

Solid and liquid
Solid
Critical state
Critical temperature
Pressure P
Liquid
Liquid and vapor
Vapor
Triple state
Solid and vapor
Volume v
Temperature T
Constant-temperature line

Figure 9.274: PVT **phase diagram for real gas**

Notice that in the last figure there is a line named "triple state", seen in the PV side, it reduces to a well known point named the "triple point". Therefore the triple point of a substance is the temperature and pressure at which the three phases (gas, liquid, and solid) of that substance coexist in thermodynamic equilibrium. Most curious is the "critical point[41]" that results in a "critical state", also commonly named "critical superfluid", where vapour (gas) and liquid coexist and cannot be visually differentiate (this happens in a value of pressure and temperature that is not natural on Earth but possible on Saturn and Neptune).

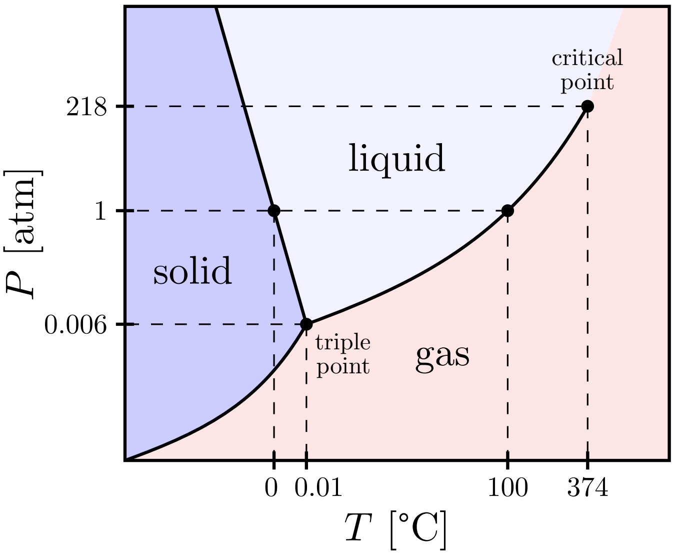


Figure 9.275: PVT **phase diagram for water** $\mathrm{H_2O}$ **(author: Izaak Neutelings)**

Here are some triple point PV values for some substances:

Substance	**Formula**	**Temperature [K]**	**Pressure [kPa]**
Carbon (graphite)	C	$4,765$	$10,132$
Carbon dioxide	CO_2	216.55	517
Deuterium	^2H	18.63	17.1
Hydrogen	H	13.84	7.04
Mercury	Hg	234.2	$1.65 \cdot 10^{-7}$
Methane	CH_4	90.68	11.7
Oxygen	O	54.36	0.152
Titanium	Ti	$1,941$	$5.3 \cdot 10^{-3}$
Water	H_2O	273.16	0.611657

Table 9.38: **PV triple point for some substances**

And.... keep always in mind that the phase diagram show always the state of a given substance for a given pressure, temperature (and volume) when its surrounding is the substance itself and not the air (common misconception)!!!!!!!

---

[41] There are nice videos on YouTube where you can see these states appears (i.e. the critical point and the triple point).

9.7.3.2.1 Equipartition theorem of energy derivation based on quantum physics and statistical mechanics

What we have seen above can also be demonstrated from purely quantum and statistical considerations. Let us see now for general culture (some parts of the development below seems to arise magically but the reader must keep in mind that the physicists who developed it know to what result they had to arrive... and therefore there wave likely some trial and errors) and also because some of the steps used below will be useful to us to derive the Sackur-Tetrode equation.

The internal energy U can be expressed using the canonical partition function Z_c.

At thermodynamic equilibrium as for the Maxwell-Boltzmann distribution (see section Statistical Mechanics 3561) we have:

$$U=\sum_{i=1}^{K}n_iE_i=\frac{N}{Z_c}\sum_{i=1}^{K}g_i\,e^{-E_i/(kT)}E_i \qquad (9.7.699)$$

with as a reminder by definition:

$$Z_c(T,V)=\sum_{i=1}^{K}g_i\,e^{-E_i/(kT)} \qquad (9.7.700)$$

We derive the partition function with respect to the temperature T at a constant volume V:

$$\left(\frac{\partial Z_c}{\partial T}\right)_V=\sum_{i=1}^{K}g_i\frac{E_i}{kT^2}e^{-E_i/kT}=\frac{1}{kT^2}\sum_{i=1}^{K}g_i\,e^{-E_i/kT}E_i \qquad (9.7.701)$$

Hence:

$$\sum_{i=1}^{K}g_i\,e^{-E_i/kT}E_i=kT^2\left(\frac{\partial Z_c}{\partial T}\right)_V \qquad (9.7.702)$$

Therefore:

$$U=\frac{NkT^2}{Z_c}\left(\frac{\partial Z_c}{\partial T}\right)_V=NkT^2\left(\frac{\partial\ln(Z_c)}{\partial T}\right)_V \qquad (9.7.703)$$

We consider the simplest gas, a monoatomic ideal gas of zero electric charge.

If we do not neglect the potential gravitational energy mgz in front of the kinetic energy of translation, the energy of the particles of this gas is written:

$$E_i=\frac{1}{2}mv_i^2=\frac{p_i^2}{2m} \qquad (9.7.704)$$

the speed v_i of each particle can vary from $-\infty$ to $+\infty$.

We will use the ideal gas model, where the potential interaction energy is neglected, and the particles are considered to be free. This is an approximation because without interaction between the particles, the gas could not evolve over time, and tend towards thermodynamic equilibrium.

The gas is in a chamber that is supposed to be parallel to the sides a, b, c, the walls of which are modelled by barriers of infinite potential. We have here in our model of potential energy, a three-dimensional flat well, of infinite depth:

$$U_i(\vec{r},t)=0 \text{ if } \begin{cases} a>x>0 \\ b>y>0 \\ c>z>0 \end{cases} \qquad U_i(\vec{r},t)=+\infty \text{ otherwise} \tag{9.7.705}$$

If moreover we go neglect the potential energy of gravity in front of the kinetic energy, then the potential energy is zero and the Schrödinger equation of the quanton in the enclosure is written (see section Quantum Wave Physics page 4772):

$$\frac{-\hbar^2}{2m}\Delta\psi_i=\mathrm{i}\hbar\frac{\partial\psi_i}{\partial t} \tag{9.7.706}$$

If the space and time variables are linked, they are only linked in the expression of the potential energy.

When the potential interaction energy U is independent of time and depends only on space, which is the case here since it is zero, the Schrödinger equation can be solved by the method of separation of space variables $\vec{r}$ and time t.

We then look for a solution of the form:

$$\psi_i(\vec{r},t)=\varphi_i(\vec{r})\,f_i(t) \tag{9.7.707}$$

Since the Laplacian Δ only applies to space variables, the Schrödinger equation becomes:

$$\frac{-\hbar^2}{2m}f_i(t)\,\Delta\varphi_i(\vec{r})=\mathrm{i}\hbar\,\varphi_i(\vec{r})\frac{\mathrm{d}f_i(t)}{\mathrm{d}t} \tag{9.7.708}$$

by dividing by $\varphi_i(\vec{r})\,f_i(t)$:

$$-\frac{\hbar^2}{2m}\frac{\Delta\varphi_i(\vec{r})}{\varphi_i(\vec{r})}=\mathrm{i}\hbar\frac{1}{f_i(t)}\frac{\mathrm{d}f_i(t)}{\mathrm{d}t} \tag{9.7.709}$$

We have a function of $\vec{r}$ which is equal to a function of t.

The only possibility is that they are equal to the same constant in space and time. The dimensional analysis shows that this constant is an energy, which can only be the internal energy ε_i of the quanton.

For the temporal part we have:

$$\begin{aligned} \mathrm{i}\hbar\frac{1}{f_i(t)}\frac{\mathrm{d}f_i(t)}{\mathrm{d}t}&=E_i \\ \frac{\mathrm{d}f_i(t)}{f_i(t)}&=-\frac{\mathrm{i}}{\hbar}E_i\,\mathrm{d}t \\ f_i(t)&=Ce^{-\mathrm{i}E_it/\hbar} \end{aligned} \tag{9.7.710}$$

and for the spatial part:

$$-\frac{\hbar^2}{2m}\frac{\Delta\varphi_i(\vec{r})}{\varphi_i(\vec{r})}=E_i \tag{9.7.711}$$

where we recognize the time independent Schrödinger equation.

The Laplacian being the sum of three operators each acting on a single space coordinate, we seek a solution with the space variables separated:

$$\varphi_i(\vec{r})=\varphi_{i1}(x)\,\varphi_{i2}(y)\,\varphi_{i3}(z) \tag{9.7.712}$$

Therefore:

$$\Delta\varphi_i(\vec{r})=\varphi_{i1}''(x)\,\varphi_{i2}(y)\,\varphi_{i3}(z)+\varphi_{i1}(x)\,\varphi_{i2}''(y)\,\varphi_{i3}(z)+\varphi_{i1}(x)\,\varphi_{i2}(y)\,\varphi_{i3}''(z) \tag{9.7.713}$$

Hence:

$$-\frac{\hbar^2}{2m}\left(\frac{\varphi_{i1}''(x)}{\varphi_{i1}(x)}+\frac{\varphi_{i2}''(y)}{\varphi_{i2}(y)}+\frac{\varphi_{i3}''(z)}{\varphi_{i3}(z)}\right)=E_i \tag{9.7.714}$$

The space coordinates being independent:

$$-\frac{\hbar^2}{2m}\frac{\varphi_{i1}''(x)}{\varphi_{i1}(x)}=E_{ix}\qquad -\frac{\hbar^2}{2m}\frac{\varphi_{i2}''(y)}{\varphi_{i2}(y)}=E_{iy}\qquad -\frac{\hbar^2}{2m}\frac{\varphi_{i3}''(z)}{\varphi_{i3}(z)}=E_{iz} \tag{9.7.715}$$

with:

$$E_i=E_{ix}+E_{iy}+E_{iz} \tag{9.7.716}$$

By defining:

$$k_{ix}^2=2mE_{ix}/\hbar^2\qquad k_{iy}^2=2mE_{iy}/\hbar^2\qquad k_{iz}^2=2mE_{iz}/\hbar^2 \tag{9.7.717}$$

we have the simple following particular solution:

$$\frac{\varphi_{i1}''(x)}{\varphi_{i1}(x)}=-k_{ix}^2\qquad \frac{\varphi_{i2}''(y)}{\varphi_{i2}(y)}=-k_{iy}^2\qquad \frac{\varphi_{i3}''(z)}{\varphi_{i3}(z)}=-k_{iz}^2 \tag{9.7.718}$$

Whose solutions are:

$$\begin{aligned}\varphi_{i1}(x)&=A_1\sin(k_{ix}x+\alpha_{ix})\\ \varphi_{i2}(y)&=A_2\sin(k_{iy}y+\alpha_{iy})\\ \varphi_{i3}(z)&=A_3\sin(k_{iz}z+\alpha_{iz})\end{aligned} \tag{9.7.719}$$

We make the non-trivial assumption that the wave function is zero on the walls of the enclosure.

For the first relation, the boundary condition at $x=0$ is written:

$$\varphi_{i1}(0)=A_1\sin(\alpha_{ix})=0\qquad \forall p_{ix}\in\mathbb{Z},\quad \alpha_{ix}=p_{ix}\pi \tag{9.7.720}$$

and, on $x=a$:

$$\varphi_{i1}(a)=A_1\sin(k_{ix}a+p_{ix}\pi)=0\Leftrightarrow k_{ix}=n_{ix}\frac{\pi}{a} \tag{9.7.721}$$

where we define in a very tricky way.... n_{ix} as an integer such that:

$$\begin{aligned}n_{ix}&=\frac{a}{\pi}k_{ix}\\ &=\frac{a}{\pi\hbar}\sqrt{2mE_{ix}}\\ &\leqslant\frac{2a}{h}\sqrt{2mE_i}\end{aligned} \tag{9.7.722}$$

n_{ix} is therefore a positive integer that has for maximum value $2a\sqrt{2mE_i}/h$.

For the three relations we then define:

$$k_{ix} = n_{ix}\frac{\pi}{a} \qquad k_{iy} = n_{iy}\frac{\pi}{b} \qquad k_{iz} = n_{iz}\frac{\pi}{c} \tag{9.7.723}$$

The energy E_i of the i^{th} quanton is therefore quantified, and has for expression:

$$\begin{aligned} E_i &= E_{ix} + E_{iy} + E_{iz} \\ &= \frac{\hbar^2}{2m}\left(k_{ix}^2 + k_{iy}^2 + k_{iz}^2\right) \\ &= \frac{\hbar^2\pi^2}{2m}\left(\frac{n_{ix}^2}{a^2} + \frac{n_{iy}^2}{b^2} + \frac{n_{iz}^2}{c^2}\right) \\ &= \frac{h^2}{8m}\left(\frac{n_{ix}^2}{a^2} + \frac{n_{iy}^2}{b^2} + \frac{n_{iz}^2}{c^2}\right) \end{aligned} \tag{9.7.724}$$

The above equality gives the set of quantum states accessible to a quanton of energy E_i. We can write it after rearrangement:

$$\frac{h^2}{8m}\left(\frac{n_{ix}^2}{a^2\varepsilon_i} + \frac{n_{iy}^2}{b^2\varepsilon_i} + \frac{n_{iz}^2}{c^2\varepsilon_i}\right) = 1 \tag{9.7.725}$$

which is of the shape:

$$\frac{x^2}{A^2} + \frac{y^2}{B^2} + \frac{z^2}{C^2} = 1 \tag{9.7.726}$$

with:

$$x = n_{ix} \qquad y = n_{iy} \qquad z = n_{iz} \tag{9.7.727}$$

$$A^2 = \frac{8ma^2\,E_i}{h^2} \qquad B^2 = \frac{8mb^2\,E_i}{h^2} \qquad C^2 = \frac{8mc^2\,E_i}{h^2} \tag{9.7.728}$$

It is the equation of an ellipsoid of parameters A, B, C (lengths of the three half-axes). Each point of integer coordinates (n_{ix}, n_{iy}, n_{iz}) inside the ellipsoid, represents a quantum state accessible to a quanton of energy E_i. The quantum numbers n_{ix}, n_{iy} and n_{iz} being positive, we only take into account the eighth $(\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2})$ of the ellipsoid volume. The factor of degeneration is then given by:

$$\begin{aligned} g_i &= \frac{1}{8}\cdot V_{\text{ellipsoid}} = \frac{1}{8}\cdot\frac{4}{3}\pi\,ABC \\ &= \frac{\pi}{6}\cdot\frac{a}{h}\sqrt{8mE_i}\cdot\frac{b}{h}\sqrt{8mE_i}\cdot\frac{c}{h}\sqrt{8mE_i} \\ &= \frac{\pi\,abc\,(8mE_i)^{3/2}}{6h^3} \\ &= \frac{4\pi\,V\,(2mE_i)^{3/2}}{3h^3} \\ &= \frac{4\pi\,V\,p^3}{3h^3} \\ &= \frac{1}{h^3}\cdot V\cdot\frac{4}{3}\pi p^3 \end{aligned} \tag{9.7.729}$$

Let us interpret this result: Let us consider the 6-dimensional hyperspace $(\vec{r},\vec{p})$, named as we know the "phase space". The hyper-volume of this space is constituted by the multiplication of the volume V of ordinary space and the volume of the space of the quantities of movement $4/3\cdot\pi p^3$.

In Cartesian coordinates the degeneracy factor is written:

$$g_i=\frac{\displaystyle\int_x\int_y\int_z\int_{p_x}\int_{p_y}\int_{p_z}\mathrm{d}x\mathrm{d}y\mathrm{d}z\mathrm{d}p_x\mathrm{d}p_y\mathrm{d}p_z}{h^3} \qquad (9.7.730)$$

Let's go back to the partition function:

$$\begin{aligned}Z_c&=\frac{1}{h^3}\int_x\int_y\int_z\int_{p_x}\int_{p_y}\int_{p_z}e^{-E_i/(kT)}\mathrm{d}x\mathrm{d}y\mathrm{d}z\mathrm{d}p_x\mathrm{d}p_y\mathrm{d}p_z\\&=\frac{m^3V}{h^3}\int_{v_x}\int_{v_y}\int_{v_z}e^{-m(v_x^2+v_y^2+v_z^2)/(2kT)}\mathrm{d}v_x\mathrm{d}v_y\mathrm{d}v_z\\&=\frac{m^3V}{h^3}\left[\int_{-\infty}^{+\infty}e^{-mv^2/(2kT)}\mathrm{d}v\right]^3\end{aligned} \qquad (9.7.731)$$

We put $a=m/(2kT)$:

$$Z_c=\frac{m^3V}{h^3}\left[\int_{-\infty}^{+\infty}e^{-av^2}\mathrm{d}v\right]^3 \qquad (9.7.732)$$

Using the result of the Gauss integral (see section Statistics page 695):

$$\int_{-\infty}^{+\infty}e^{-ax^2}\mathrm{d}x=\sqrt{\frac{\pi}{a}} \qquad (9.7.733)$$

Therefore we get the following canonical partition function for a monoatomic classical ideal gas:

$$Z_c=\frac{m^3V}{h^3}\left(\sqrt{\frac{2kT\pi}{m}}\right)^3=\frac{V}{h^3}(2\pi mkT)^{3/2} \qquad (9.7.734)$$

Z_c is therefore a function of the temperature T and the volume V :

$$Z_c=Z_c(T,V) \qquad (9.7.735)$$

Let us now express the internal energy of translation U. Starting from;

$$Z_c=\frac{V}{h^3}(2\pi mkT)^{3/2} \qquad (9.7.736)$$

Taking the natural logarithm:

$$\ln(Z_c)=\ln\left(\frac{V}{h^3}\right)+\frac{3}{2}\ln(2\pi mkT) \qquad (9.7.737)$$

Therefore:

$$\left(\frac{\partial \ln(Z_c)}{\partial T}\right)_V = \frac{3}{2T} \tag{9.7.738}$$

Remembering that just earlier above we have derived:

$$U = \frac{NkT^2}{Z_c}\left(\frac{\partial Z_c}{\partial T}\right)_V = NkT^2\left(\frac{\partial \ln(Z_c)}{\partial T}\right)_V \tag{9.7.739}$$

So finally:

$$\boxed{U = \frac{3}{2}NkT} \tag{9.7.740}$$

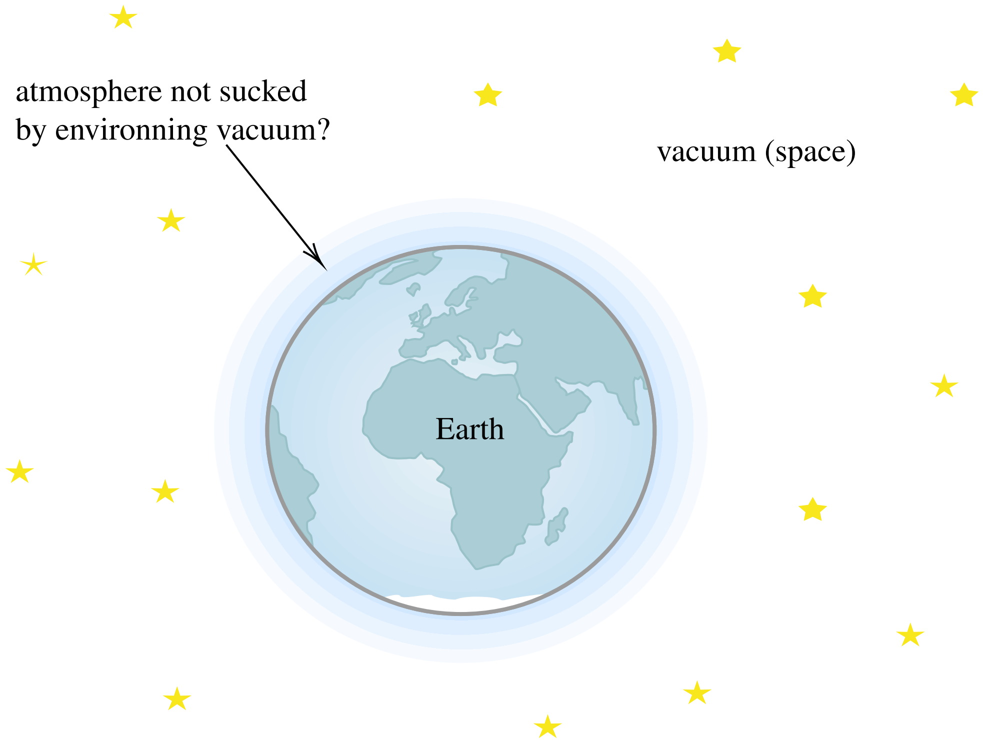
Now that we have the two relations, i.e. the equipartition energy for a monoatomic particle:

$$E_c = \frac{3}{2}NkT \tag{9.7.741}$$

But also the escape velocity (see section Mechanics page 3234):

$$v_e = \sqrt{\frac{2Gm}{R^2}R} = \sqrt{\frac{2Gm}{R}} \tag{9.7.742}$$

We can now easily understand why Earth's atmosphere is not sucked by its surrounding vacuum!



Anyway the question is quite dumb as vacuum does not suck like many believe, it's the atmosphere pushing into the space of the vacuum. A sealed empty container under water does not such the water in when we open it. It's the water pushing into the container! A vacuum doesn't suck, it just doesn't push back.

For Earth that gives us an escape velocity of $11,179.365\,[\mathrm{m}\cdot\mathrm{s}^{-1}]$. And using the equipartition energy in the most favorable (highest temperature ever measure of troposphere and lightest significant molecule that is nitrogen N_2) case we have:

$$v=\sqrt{\frac{3kT}{m}}=\sqrt{\frac{3\cdot 1.380649\cdot 10^{-23}\cdot 329.15}{4.652\cdot 10^{-26}}}\cong 541\,[\mathrm{m}\cdot\mathrm{s}^{-1}] \qquad (9.7.743)$$

So it's far from being enough for the atmosphere to escape Earth's attraction!

9.7.3.3 Kinetic Pressures (kinetic theory of gases)

The kinetic theory describe a gas as a large number of submicroscopic particles (atoms or molecules), all of which are in constant rapid motion that has randomness arising from their many collisions with each other and with the walls of the container.

Kinetic theory explains macroscopic properties of gases, such as pressure, temperature, viscosity, thermal conductivity, and volume, by considering their molecular composition and motion. The theory posits that gas pressure is due to the impacts, on the walls of a container, of molecules or atoms moving at different velocities.

Kinetic theory defines temperature in its own way, not identical with the thermodynamic definition.

Let us seek the number of molecules of a perfect gas, all supposed to be animated by an equal speed v, which strike a surface S for a duration $\mathrm{d}t$.

If the gas studied is in a state of thermodynamic equilibrium, this will result in the homogeneity of the intensive variables (hypothesis of molecular chaos).

It follows that the density of the particles is constant:

$$N_v = \frac{N}{V} = c^{te} \tag{9.7.744}$$

If we admit that the molecules are in motion, we shall assume that there is isotropy of velocities. In other words, since the velocities can be described in a system of 3 orthogonal axes (three spatial dimensions), there are 6 possible primary directions in total (2 directions per axis: forward, backward ...) .

This translates into equivalence between the different directions. There is:

$$\frac{N_v}{6} \tag{9.7.745}$$

particles having a velocity v along one of the primary directions.

Thus, during a time $\mathrm{d}t$, the surface S of the wall is percussed only by a part of the molecules contained in the volume $Sv\mathrm{d}t$ Indeed, only $1/6$ of the molecules contained in this volume are actually directed towards the surface S.

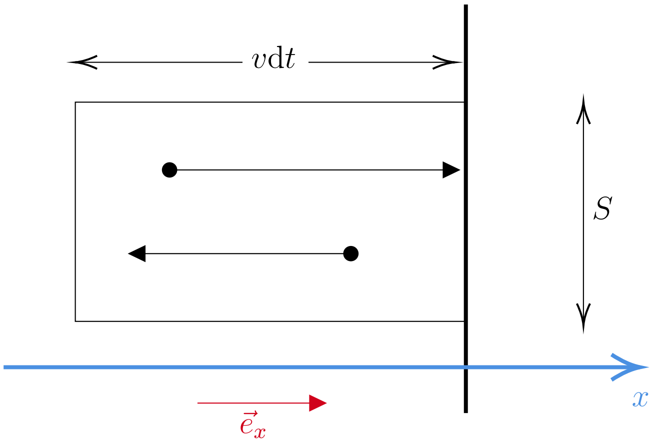


Figure 9.276: One of the walls of the enclosure (tank)

The number of molecules which strike the wall during the duration $\mathrm{d}t$ is thus:

$$\frac{1}{6}N_vSv\mathrm{d}t \tag{9.7.746}$$

Let us now study the dynamics of the impact of a particle on the wall!

The particle of a mass m arriving on the wall with the speed $\vec{v}_x$ goes away again with a speed $-\vec{v}_x$ if the shock is perfectly elastic.

The variation of the momentum of the particle is thus:

$$\Delta\vec{p}_{\text{wall}}=-m(\vec{v}_f-\vec{v}_i)=2mv\vec{e}_x \tag{9.7.747}$$

$$\Delta\vec{p}_{\text{molecule}}=m(\vec{v}_f-\vec{v}_i)=m((-\vec{v}_x)-\vec{v}_x)=-2mv\vec{e}_x \tag{9.7.748}$$

Following the conservation of the linear momentum, this linear momentum is transferred to the wall: the variation of the quantity of motion undergone by the wall is the opposite of that of the particle:

$$\Delta\vec{p}_{\text{wall}}=-m(\vec{v}_f-\vec{v}_i)=2mv\vec{e}_x \tag{9.7.749}$$

The total variation of the linear momentum of all the particles which strike the wall during the period $\mathrm{d}t$ is then equal in norm to:

$$\mathrm{d}p=(2mv)\frac{1}{6}N_vSv\mathrm{d}t \tag{9.7.750}$$

Applying the fundamental principle of dynamics, we can pass to the force undergone by the wall during the striking of these molecules:

$$F=\frac{\mathrm{d}p}{\mathrm{d}t} \tag{9.7.751}$$

therefore:

$$F=\frac{1}{3}N_vSmv^2 \tag{9.7.752}$$

Now by defining the pressure and taking the module: But by definition of the pressure and taking the module:

$$P=\frac{F}{S} \tag{9.7.753}$$

we then have the "kinetic pressure" given by:

$$\boxed{P=\frac{1}{3}N_v mv^2} \tag{9.7.754}$$

The kinetic pressure is therefore nothing else than the frequency of the shocks of the particles (molecules) on the wall. The more molecules (term in N_v) and the quicker they are (term in v), the more the number of shocks increases. This is consistent with experience and intuition.

9.7.3.4 Kinetic Temperature

Now comes the proof we are waiting from since the beginning of our study of gas: the proof that temperature (at a molecular or atomic level) is nothing more that a quantity related to the average velocity but with a conversion factor that is nothing else than the Boltzmann constant! Obviously we will see that there is a more precise definition of temperature in the section of Thermodynamics as the result that follows doesn't apply at the quantum level.

For this purpose we start again from:

$$P=\frac{1}{3}N_v mv^2 \tag{9.7.755}$$

By replacing N_v by the quotient N/V, the preceding relation can be written as:

$$PV=\frac{1}{3}Nmv^2 \tag{9.7.756}$$

If we identify it with the perfect gas state equation presented earlier above:

$$PV=nRT \tag{9.7.757}$$

it comes:

$$\frac{1}{3}Nmv^2=nRT \tag{9.7.758}$$

But the quantity of moles n is equal to the ratio of the number of particles to the number of Avogadro as we know:

$$n=\frac{N}{N_{\text{A}}} \tag{9.7.759}$$

which makes it possible to define the "kinetic temperature" by:

$$T_k=\frac{1}{3}\frac{N}{n}\frac{mv^2}{R}=\frac{1}{3}\frac{N_{\text{A}}mv^2}{R} \tag{9.7.760}$$

We then fall back on a way to introduce "Boltzmann constant" and such that:

$$k=\frac{R}{N_{\text{A}}}=1.38064852\cdot 10^{-23}\,[\text{J}^2\cdot\text{K}^{-1}] \tag{9.7.761}$$

The expression of kinetic temperature then takes the form:

$$\boxed{T_k=\frac{mv^2}{3k}} \tag{9.7.762}$$

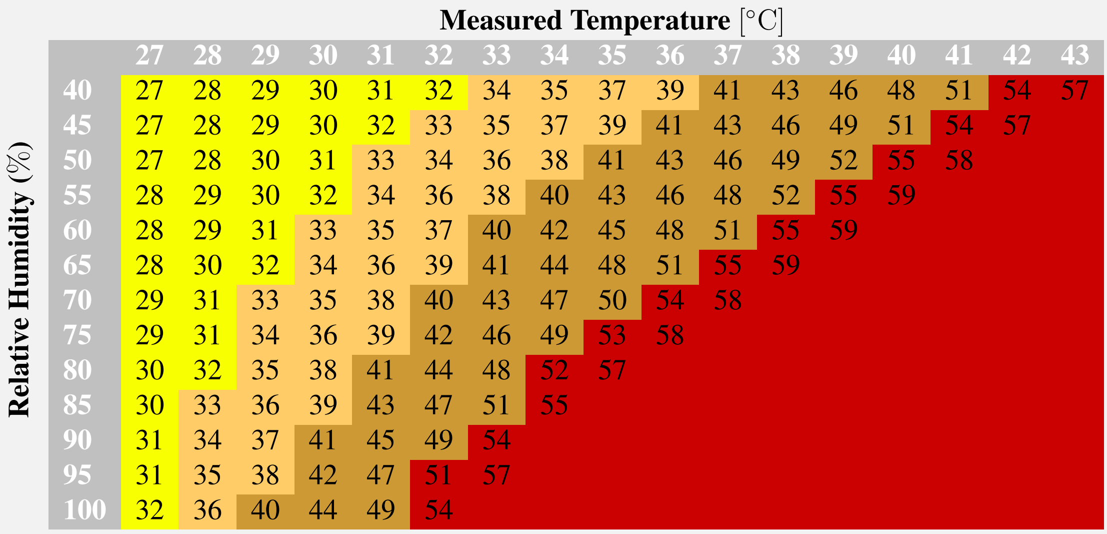
We can see that it seems the temperature can range from $]0,+\infty[$. However as we will see in the section of Quantum Mechanics an absolute temperature of zero cannot be reached because of the Heisenberg uncertainty principle. We also don't have experimental evidence yet for an upper limit for temperature.

Notice that the physical relation above doesn't gives us any threshold to define what is "hot" and what is "cold". Indeed, these two concepts are completely subjective to every human is not even precisely defined by any ISO standard (don't forget as already mentioned in the Introduction chapter of that book that it is not because a word exist in human vocabulary that the underlying concept has a physical reality!). The nearest thing we can find is the ISO 13732-1 *Ergonomics of the thermal environment - Methods for the assessment of human responses to contact with surfaces*. Especially the chapter *5.2 Identification of hot touchable surface* and the Part 3 treating of *Cold surfaces*. But no precise temperatures are given, only a protocol to define it statistical based on a test group.

This relationship shows that the kinetic temperature is only the representation of the kinetic energy of the particles (whose temperature is in fact a measure of the average velocity of the molecules). Imaginatively, it is the image of the violence of shocks. We can also better understand the origin of the definition of absolute temperature (otherwise, the velocity should be a complex number) and the argument that everything is immobile at absolute zero!

Remark

The kinetic temperature must not be confused with the "heat index" (HI), also named "feel/like temperature", that combines air temperature and relative humidity, in shaded areas, to posit a human-perceived equivalent temperature, as how hot it would feel if the humidity were some other value in the shade. Indeed, the human body normally cools itself by perspiration, or sweating. Heat is removed from the body by evaporation of that sweat. However, high relative humidity reduces the evaporation rate. This results in a lower rate of heat removal from the body, hence the sensation of being overheated (indeed you won't feel the same in a 323 [K] standard air without wind than in water being at the same temperature... and that's also why most human fail by estimating the temperature of different materials even if they are all... at the same temperature!). This effect is subjective, with different individuals perceiving heat differently for various reasons (such as differences in body shape, metabolic differences, differences in hydration, pregnancy, menopause, effects of drugs and/or drug withdrawal); its measurement has been based on subjective descriptions of how hot subjects feel for a given temperature and humidity (it's strongly related to the "heat capacity", i.e. the amount of energy that must be added or removed, in the form of heat, to one unit of mass of a substance in order to cause an increase or decrease of one unit in its temperature). This results in a heat index that relates one combination of temperature and humidity to another. The National Oceanic and Atmospheric Administration provides a double entry table (relative humidity + temperature) to extract a feel-like temperature, some other countries use - quite different - empirical formulas like the one below:



Relative Humidity (%) \ Measured Temperature [°C]	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43
40	27	28	29	30	31	32	34	35	37	39	41	43	46	48	51	54	57
45	27	28	29	30	32	33	35	37	39	41	43	46	49	51	54	57	
50	27	28	30	31	33	34	36	38	41	43	46	49	52	55	58		
55	28	29	30	32	34	36	38	40	43	46	48	52	55	59			
60	28	29	31	33	35	37	40	42	45	48	51	55	59				
65	28	30	32	34	36	39	41	44	48	51	55	59					
70	29	31	33	35	38	40	43	47	50	54	58						
75	29	31	34	36	39	42	46	49	53	58							
80	30	32	35	38	41	44	48	52	57								
85	30	33	36	39	43	47	51	55									
90	31	34	37	41	45	49	54										
95	31	35	38	42	47	51	57										
100	32	36	40	44	49	54											

🟨 Discomfort
🟧 Extreme discomfort
🟫 Danger
🟥 Extreme Danger

Table 9.39: **Temperature felt based on measured temperature and humidity**

Thus the energy of a perfect gas is reduced to the sum of the kinetic energies of the particles which constitute it:

$$U=\sum_i \frac{1}{2}mv_i^2=\frac{N}{2}mv^2=\frac{3}{2}NkT=\frac{3}{2}nRT \tag{9.7.763}$$

The atoms of a monoatomic perfect gas can be assimilated to material points. Their kinetic energy is a kinetic energy of translation whose mean value, per atom, is written:

$$E_c=\frac{1}{2}mv^2=\frac{3}{2}kT \tag{9.7.764}$$

In the velocity space, all directions are equivalent! There is isotropy of the velocity distribution. In Cartesian coordinates, it comes:

$$v^2=v_x^2+v_y^2+v_z^2 \tag{9.7.765}$$

and as a result of isotropy:

$$v_x^2=v_y^2=v_z^2=\frac{1}{3}v^2 \tag{9.7.766}$$

hence:

$$E_{c,x}=E_{c,y}=E_{c,z}=\frac{1}{2}kT \tag{9.7.767}$$

Thus, the average kinetic energy per degree of freedom of translation is equal to:

$$E_{c,\text{df}}=\frac{1}{2}kT \tag{9.7.768}$$

The reader will of course notice that we obtain here the same results and conclusions as in our study of the virial theorem with the difference that the approach here is simpler and therefore more didactic.

9.7.3.5 Amagat and Dalton's law

Let us consider a mixture of two perfect gases in a volume V imposed constant at constant temperature T also imposed. We can write for each of the gases:

$$P_1V = n_1RT \qquad P_2V = n_2RT \tag{9.7.769}$$

Where P_1, P_2 are named "partial pressures". The whole mixture will be written:

$$PV = nRT = (n_1 + n_2)RT \tag{9.7.770}$$

But:

$$(n_1 + n_2)RT = n_1RT + n_2RT = P_1V + P_2V \tag{9.7.771}$$

and therefore:

$$(P_1 + P_2)V = (n_1 + n_2)RT \tag{9.7.772}$$

Finally we conclude that:

$$\boxed{P_{V,T} = P_1 + P_2} \tag{9.7.773}$$

The pressure of a mixture of two perfect gases is the sum of the partial pressures, this is the "Dalton's law" also named "Dalton's law of partial pressures".

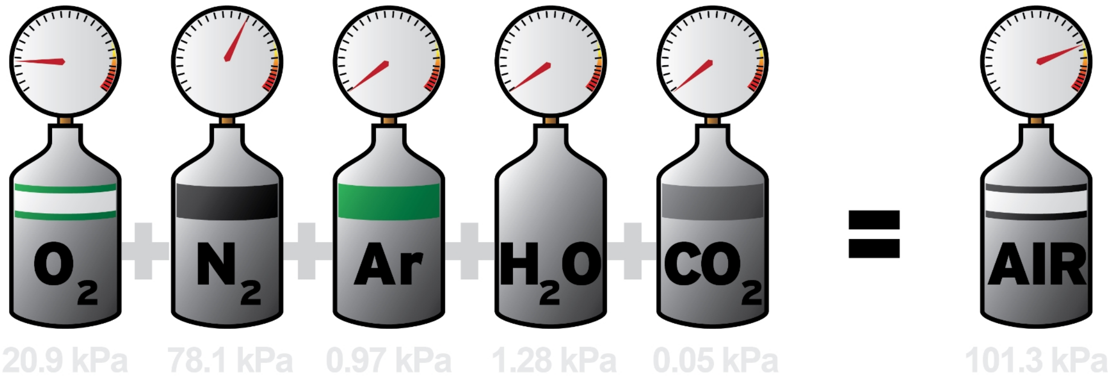


Figure 9.277: **Dalton's law illustration for air at normal temperature**

> **Remark**
>
> We mentioned in the section Thermodynamics that pressure was an intensive variable and therefore not additive. But this is not the case in the situation of perfect gases when the volume is forced as being constant.

Thus, the partial pressure P_i of a gas in a mixture is equal to the product of the total pressure P of the mixture by the percentage (concentration) c_i of the gas considered in the mixture:

$$P_i = c_iP \tag{9.7.774}$$

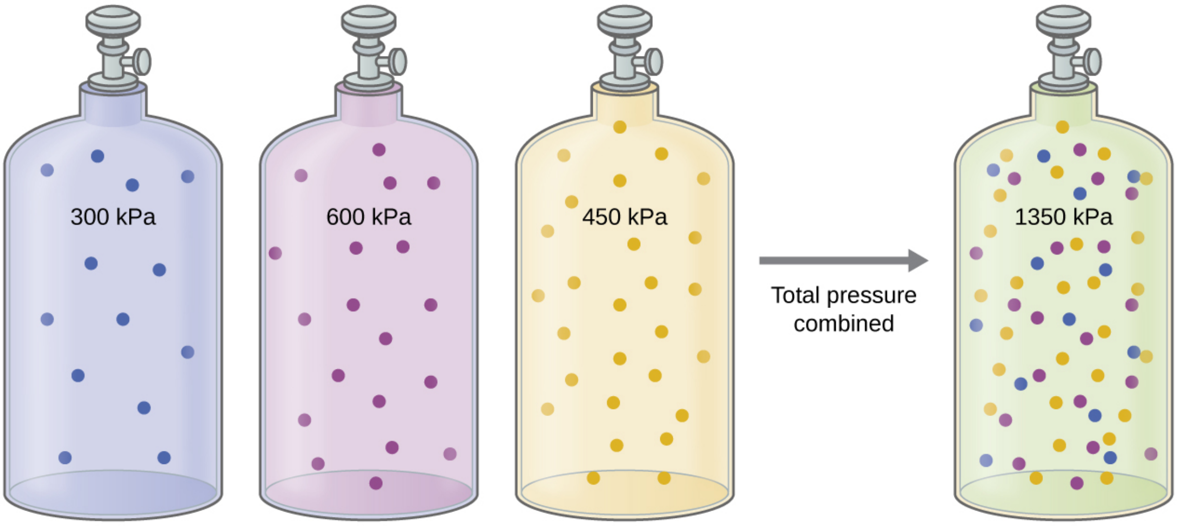


Figure 9.278: Numerical application of Dalton's law (source: OpenStax)

We can make the same reasoning at constant pressure P and constant temperature T. We then get:

$$P(V_1+V_2)=(n_1+n_2)RT \quad (9.7.775)$$

hence:

$$\boxed{V_{T,P}=V_1+V_2} \quad (9.7.776)$$

The volume of a mixture of two perfect gases at constant pressure and constant temperature is the sum of the partial volumes, it is the "Amagat's law".

9.7.3.6 Mean free path (in kinetic theory)

The mean free path is the average distance travelled by a moving particle (such as an atom, a molecule, a photon) between successive impacts (collisions), which modify its direction or energy or other particle properties.

We will now see a very interesting case study of gases which makes it possible to clarify a lot of misunderstandings in everyday life (smoke in restaurants, heat near a radiator, ...). However, the phenomena are in reality more complex, it is also necessary to take into account diffusion, convection, etc (see section Thermodynamics page 3634).

Consider a molecule, which moves at the average speed $\bar{v}$. Its sphere of influence V_i then sweeps, during the unit of time of its displacement, a volume (cylinder) given by:

$$V_i=\pi r^2\bar{v} \quad (9.7.777)$$

Remark

In the case of an atom or molecule as a sphere of influence, it is often referred to as the "Van der Waals volume" and the associated "Van der Waals radius".

If the unit of volume contains n molecules (n has therefore the units of a volumetric density), the number of shocks (collisions) during the unit of time in this same unit of volume will then be obviously:

$$C=n\pi r^2\bar{v} \tag{9.7.778}$$

at least if the other spheres of influence were immobile ... So to take into account the movement of other spheres of influence, consider the figure below with three simplistic scenarios:

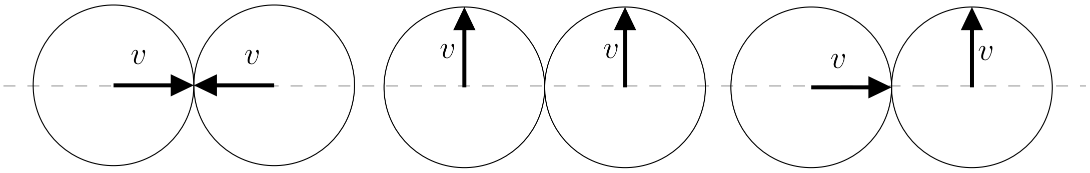


Figure 9.279: **Three collision scenarios**

From left to right we have:

1. The relative speed of the molecules is $2\bar{v}$
2. The relative speed of the molecules is zero
3. The relative speed of the molecules equals (Pythagoras) $\sqrt{2}\bar{v}$

The first two cases are somewhat extreme ... The third one will be considered as an average and can serve as a new basis for the previous calculation.

Thus, using the relative speed of the last scenario, the number of collisions during the unit of time in the unit volume becomes:

$$C=n\pi r^2\sqrt{2}\bar{v} \tag{9.7.779}$$

Thus, between two collisions a molecule travels an average distance:

$$\lambda=\frac{\bar{v}}{C}=\frac{1}{n\pi r^2\sqrt{2}} \tag{9.7.780}$$

It is therefore the expression of the mean free path as a function of the molecular (particle) density n (and not the number of moles!!!) and the radius r of the sphere of influence, the intrinsic parameter of the gas under consideration. The above relation is often written $\gamma=(n\sigma)^{-1}$ where σ is named the "scattering cross-section".

Given the ideal gas law:

$$PV=NkT \tag{9.7.781}$$

we have:

$$P=\frac{N}{V}kT=nkT \tag{9.7.782}$$

and therefore:

$$n=\frac{P}{kT} \tag{9.7.783}$$

Where n is still the molecular (particles) density (and not the number of moles !!!) and therefore we can write:

$$\boxed{\lambda=\frac{k}{\pi r^2\sqrt{2}}\frac{T}{P}} \tag{9.7.784}$$

A numerical application gives for the element at normal conditions of temperature and pressure a free mean path which for the majority of the current molecules is several thousands of times the diameter of a molecule of standard size (the mean free path is therefore of the micrometer[42]).

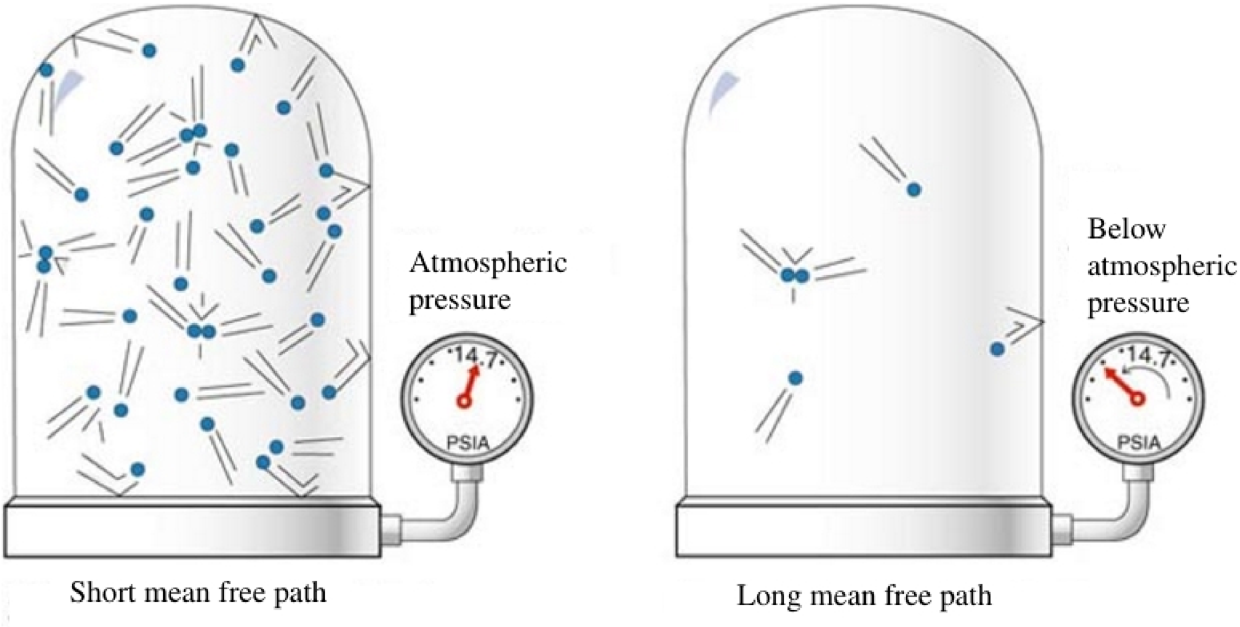


Figure 9.280: Simplified illustration of free mean path (source: ?)

Using the relation of the most probable mean velocity under the hypothesis of a Maxwell distribution of the velocities (see section Statistical Mechanics page 3554) we have:

$$\bar{v}=\sqrt{\frac{2kT}{m}}=\sqrt{\frac{2RT}{M_m}}\cong 389\,[\mathrm{m}\cdot\mathrm{s}^{-1}] \tag{9.7.785}$$

For a molecule with a molar mass of 30 [g] at normal temperature. This gives a number of collisions for one mole of molecules:

$$C=n\pi r^2\sqrt{2}\bar{v}\cong 15\cdot 10^6\,[\mathrm{s}^{-1}] \tag{9.7.786}$$

---

[42] In the section Principia we have given a table with a lot of examples of vacuum levels and their corresponding mean free path for recall...

This high frequency of collisions explains, under normal conditions of temperature and pressure, the rapidity with which the statistical equilibrium is established within a gas and also the speed of sound propagation.

At ambient temperature and for a vacuum of $7.5 \cdot 10^{-10}$ [Pa], we get with the same values a mean free path of approximately 1 [km].

The following table lists some typical values for air at different pressures at room temperature (remember by the way that 1 cubic meter is $1,000$ litres):

Pressure range	**Pressure** (hPa)	**Number density** (molecules $\cdot \mathrm{m}^{-3}$)	**Mean free path**
Ambient pressure	$1,013$	$2.7 \cdot 10^{25}$	$64 - 68$ nm[5]
Low vacuum	$300 - 1$	$10^{25} - 10^{22}$	$0.1 - 100\,\mu$m
Medium vacuum	$1 - 10^{-3}$	$10^{22} - 10^{19}$	$0.1 - 100$ mm
High vacuum	$10^{-3} - 10^{-7}$	$10^{19} - 10^{15}$	10 cm $- 1$ km
Ultra-high vacuum	$10^{-7} - 10^{-12}$	$10^{15} - 10^{10}$	1 km $- 10^5$ km
Extremely high vacuum	$< 10^{-12}$	$< 10^{10}$	$> 10^5$ km

Table 9.40: **Some typical free mean path for air**

Definition (#9.7.476): A "vacuum" is a region of space in which the density of matter is extremely low, meaning that it contains far fewer particles (atoms, molecules, ions, or dust) than ordinary matter. A perfect vacuum, defined as the complete absence of all matter, is an ideal concept that cannot be fully achieved in practice[43]. Even the best laboratory vacuums and the vast voids of interstellar space contain some particles and electromagnetic fields. In modern physics, the quantum vacuum is not truly empty either, as it is considered the lowest-energy state of a field and is characterized by transient quantum fluctuations.

Since the dimensions of the containers containing the gases being almost always less than this order of magnitude, it appears that when the vacuum is produced in a close volume, the intermolecular collisions are negligible vis-à-vis the molecule-wall collisions.

---

[43] The standard for vacuum technology vocabulary ISO 3529-1 defines a vacuum as a space with a gas pressure below atmospheric levels

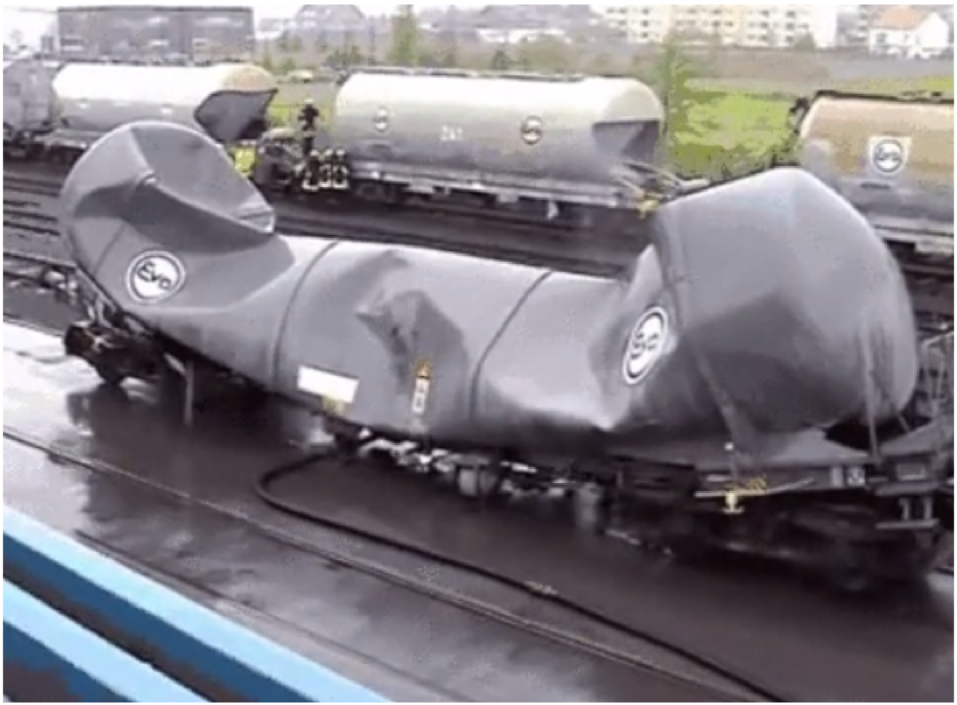


Figure 9.281: **Tank trailer vacuum implosion to atmospheric pressure (source: YouTube)**

Typical vacuums are designated according to the following convention:

Name	Pressure in Pascals
Coarse vacuum	10^5 to 100
Average vacuum	100 to 1
High vacuum (HV)	1 to 10^{-2}
Ultra high vacuum (UHV)	10^{-5} to 10^{-7}
Extreme high vacuum (XHV)	$< 10^{-7}$
Absolute emptiness	0 (no particles)

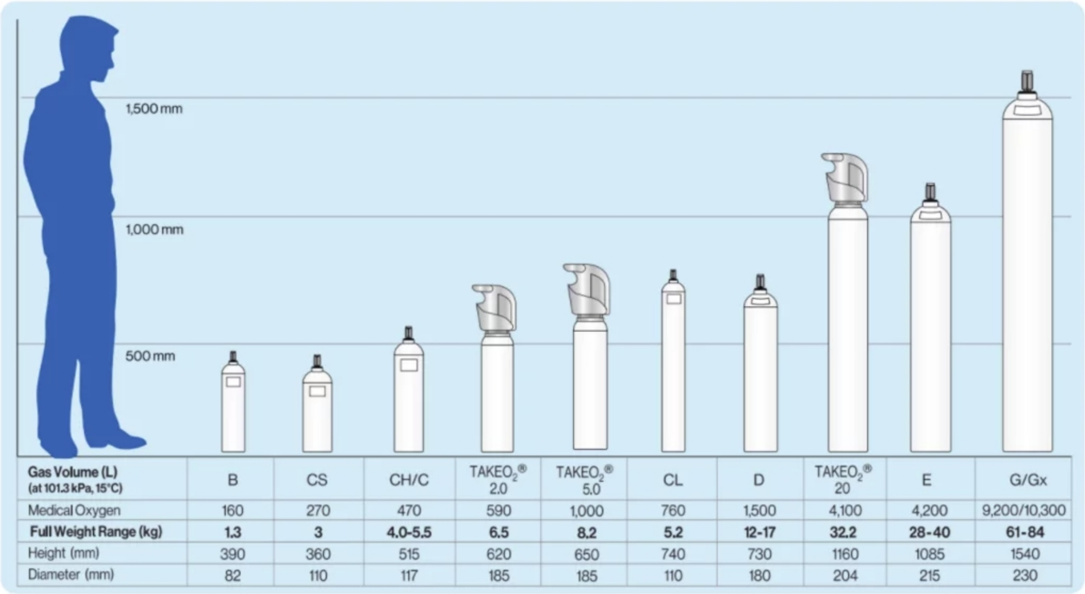
Table 9.41: **Different types of vacuum**

For a list of typical pressures (and vacuums) the reader can refer to the table at page 3101.

Caution! A vacuum on Earth does not suck like many believe, it's the atmosphere pushing into the space of the vacuum. A sealed empty container under water does not such the water in when we open it. It's the water pushing into the container! A vacuum doesn't suck, it just doesn't push back.

For people interesting in vacuum technologies and related mathematics, we strongly recommend the textbook *Vacuum and Ultravacuum: Physics and Technology* written by Igor Bello with 1063 pages (see [48]).

For information, an average human resting at sea level consumes roughly 378 to 660 liters of pure oxygen per day. As we know nowadays the technology to compress gases inside small volumes, in the image below the reader can see typical pressurised containers with their sizes and respective capacities:



Gas Volume (L) (at 101.3 kPa, 15°C)	B	CS	CH/C	TAKEO$_2$® 2.0	TAKEO$_2$® 5.0	CL	D	TAKEO$_2$® 20	E	G/Gx
Medical Oxygen	160	270	470	590	1,000	760	1,500	4,100	4,200	9,200/10,300
Full Weight Range (kg)	**1.3**	**3**	**4.0-5.5**	**6.5**	**8.2**	**5.2**	**12-17**	**32.2**	**28-40**	**61-84**
Height (mm)	390	360	515	620	650	740	730	1160	1085	1540
Diameter (mm)	82	110	117	185	185	110	180	204	215	230

Figure 9.282: **Typical pressurised containers with their sizes and respective capacities**

9.7.3.7 Sackur-Tetrode equation

The Sackur–Tetrode equation is an expression for the entropy of a monoatomic classical ideal gas which incorporates quantum considerations. The derivation will be based on the Boltzmann formula:

$$S=k\ln(\Omega) \tag{9.7.787}$$

Before we move to the maths let us recall that since classically, entropy is not directly measurable, only entropy differences are, and thus entropy is time dependent!

9.7.3.7.1 Case for distinguishable particles

We have proved in the section of Statistical Mechanics (see page 3562) that the complexion of the Maxwell-Boltzmann distribution was given by:

$$\Omega_{\text{M.B.}}=N!\prod_i \frac{g_i^{n_i}}{n_i!} \tag{9.7.788}$$

Therefore:

$$S=k\ln\left(N!\prod_{i=1}^K \frac{g_i^{n_i}}{n_i!}\right)=k\left[\ln(N!)+\sum_{i=1}^K \ln\left(\frac{g_i^{n_i}}{n_i!}\right)\right] \tag{9.7.789}$$

Using Stirling's formula approximation with $n_i \gg 1$ (see section Numerical Methods page 6301):

$$\begin{aligned}S&\cong k\left[N\ln(N)-N+\sum_{i=1}^K (n_i\ln(g_i)-n_i\ln(n_i)+n_i)\right]\\&=k\left(N\ln(N)+\sum_{i=1}^K n_i\ln\left(\frac{g_i}{n_i}\right)\right)\end{aligned} \tag{9.7.790}$$

We are looking for entropy at thermodynamic equilibrium, so with $n_i=n_i^0$. Therefore:

$$n_i^0=\frac{N}{Z_c}g_ie^{-E_i/(kT)} \tag{9.7.791}$$

Rearranging:

$$\frac{g_i}{n_i^0}=\frac{Z_c}{N}e^{E_i/(kT)} \tag{9.7.792}$$

Taking the logarithm:

$$\ln\left(\frac{g_i}{n_i^0}\right)=\ln\left(\frac{Z_c}{N}\right)+\frac{E_i}{kT} \tag{9.7.793}$$

Therefore:

$$\begin{aligned}S &\cong \left[N\ln(N)+\sum_{i=1}^{K} n_i\left(\ln\left(\frac{Z_c}{N}\right)+\frac{E_i}{kT}\right)\right]\\ &= k\left(N\ln(N)+N\ln\left(\frac{Z_c}{N}\right)+\frac{U}{kT}\right)\\ &= kN\ln(Z_c)+\frac{U}{T}\end{aligned} \qquad (9.7.794)$$

Let us calculate the entropy for a monoatomic ideal gas, from the expression of its canonical partition function Z_c (see section Continuum Mechanics page 3915) and from the expression of the internal energy U (see section Continuum Mechanics page 3902):

$$\begin{aligned}S &\cong kN\left\{\ln\left[\frac{V}{h^3}(2\pi mkT)^{3/2}\right]+1\right\}+\frac{(3/2)\,NkT}{T}\\ &= kN\left[\ln(V)+\frac{3}{2}\ln\left(\frac{2\pi mkT}{h^2}\right)+\frac{3}{2}\right]\end{aligned} \qquad (9.7.795)$$

Therefore:

$$\boxed{S_{\text{ST}}^{\text{dist.}}(V,N,T)\cong kN\left\{\ln\left[V\left(\frac{2\pi mkT}{h^2}\right)^{3/2}\right]+\frac{3}{2}\right\}} \qquad (9.7.796)$$

That is the "Sackur–Tetrode equation for distinguishable particles " (there are several ways to express it depending on how we play with the two logarithms).

Let us consider now a chamber containing a monoatomic ideal gas at the temperature T. Let's separate the enclosure into two volumes (V_1,N_1) and (V_2,N_2) thanks to a removable partition.

Let us calculate the change in entropy when we remove the partition. The initial entropy is the sum of the entropies of the 2 gases in their respective volumes, (V_1,N_1) and (V_1,N_2):

$$S_{\text{initial}}=kN_1\left[\ln(V_1)+\frac{3}{2}\ln\left(\frac{2\pi mkT}{h^2}\right)+\frac{3}{2}\right]+kN_2\left[\ln(V_2)+\frac{3}{2}\ln\left(\frac{2\pi mkT}{h^2}\right)+\frac{3}{2}\right] \qquad (9.7.797)$$

The final entropy is the sum of the entropies of the gases in the final volume V_1+V_2:

$$\begin{aligned}S_{\text{final}}=kN_1&\left[\ln(V_1+V_2)+\frac{3}{2}\ln\left(\frac{2\pi mkT}{h^2}\right)+\frac{3}{2}\right]\\ +kN_2&\left[\ln(V_1+V_2)+\frac{3}{2}\ln\left(\frac{2\pi mkT}{h^2}\right)+\frac{3}{2}\right]\end{aligned} \qquad (9.7.798)$$

The entropy variation ΔS is written:

$$S_{\text{final}}-S_{\text{initial}}=kN_1\ln\left(\frac{V_1+V_2}{V_1}\right)+kN_2\ln\left(\frac{V_1+V_2}{V_2}\right)\geq 0 \qquad (9.7.799)$$

This entropy variation is positive, which means that each time you put back and remove the partition, the entropy increases! Increase in entropy seems inevitable because of the action of mixture.

In classical thermodynamics, entropy is an extensive variable. If we double the extensive variables, the number of particles N at the same time as the volume V, the entropy must double in value. However, we see that this is not the case if we use the above Sackur–Tetrode equation as we would have:

$$S_{\text{ST}}^{\text{dist.}}(2V,2N,T)\cong k2N\left\{\ln\left[2V\left(\frac{2\pi mkT}{h^2}\right)^{3/2}\right]+\frac{3}{2}\right\}\neq 2S_{\text{ST}}^{\text{dist.}}(V,N,T) \tag{9.7.800}$$

To solve this problem, we have to consider that the particles of the gas are indistinguishable, because under this assumption, we make sure that absolutely nothing happens when we remove the partition.

9.7.3.7.2 Case for indistinguishable particles

We start always from the Boltzmann formula:

$$S=k\ln(\Omega) \tag{9.7.801}$$

But we take a corrected complexion Ω from Maxwell-Boltzmann for indistinguishable particles which is between the classical Maxwell-Boltzmann and the Fermi-Dirac (it gives non-integer values but it is a satisfactory approximation to avoid to having deal with heavy mathematical artillery):

$$\begin{aligned}S&\cong k\ln\left(\prod_{i=1}^{K}\frac{g_i^{n_i}}{n_i!}\right)\\&=k\sum_{i=1}^{K}\left(n_i\ln(g_i)-n_i\ln(n_i)+n_i\right)\\&=k\left(N+\sum_{i=1}^{K}n_i\ln\left(\frac{g_i}{n_i}\right)\right)\end{aligned} \tag{9.7.802}$$

Like before, we look for entropy at thermodynamic equilibrium, so with $n_i=n_i^0$:

$$\begin{aligned}n_i^0&=\frac{N}{Z_c}g_i\,e^{-E_i/(kT)}\\\frac{g_i}{n_i^0}&=\frac{Z_c}{N}e^{E_i/(kT)}\\\ln\left(\frac{g_i}{n_i^0}\right)&=\ln\left(\frac{Z_c}{N}\right)+\frac{E_i}{kT}\end{aligned} \tag{9.7.803}$$

Therefore:

$$\begin{aligned}S&\cong k\left[N+\sum_{i=1}^{K}n_i\left(\ln\left(\frac{Z_c}{N}\right)+\frac{E_i}{kT}\right)\right]\\&=k\left(N+N\ln\left(\frac{Z_c}{N}\right)+\frac{U}{kT}\right)\\&=kN\left(\ln\left(\frac{Z_c}{N}\right)+1\right)+\frac{U}{T}\end{aligned} \tag{9.7.804}$$

Let us calculate again the entropy for a monoatomic ideal gas, from the expression of its canonical partition function Z_c (see section Continuum Mechanics page ??) and from the expression of the internal energy U (see section Continuum Mechanics page 3902):

$$\begin{aligned}S&\cong kN\left\{\ln\left[\dfrac{V}{Nh^3}(2\pi mkT)^{3/2}\right]+1\right\}+\dfrac{(3/2)\,NkT}{T}\\&=kN\left[\ln\left(\dfrac{V}{N}\right)+\dfrac{3}{2}\ln\left(\dfrac{2\pi mkT}{h^2}\right)+1\right]+\dfrac{3}{2}kN\\&=kN\left[\ln\left(\dfrac{V}{N}\right)+\dfrac{3}{2}\ln\left(\dfrac{2\pi mkT}{h^2}\right)+\dfrac{5}{2}\right]\end{aligned}\tag{9.7.805}$$

Therefore:

$$\boxed{S_{\text{ST}}^{\text{ind.}}(V,N,T)\cong kN\left\{\ln\left[\dfrac{V}{N}\left(\dfrac{2\pi mkT}{h^2}\right)^{3/2}\right]+\dfrac{5}{2}\right\}}\tag{9.7.806}$$

that is the "Sackur-Tetrode formula for indistinguishable particles".

Remark

Notice that for both versions of the formula, the Sackur–Tetrode formula approaches negative infinity as the temperature approaches zero. That seems obviously wrong but it's normal as the assumption that we made a unrealistic for situation where quantum effects predominate.

The entropy variation with the two volume case introduced earlier above is also positive, which means that each time you put back and remove the partition, the entropy also increases! Increase in entropy seems inevitable because of the action of mixture.

We can verify that entropy is indeed an extensive variable with the corrected Maxwell-Boltzmann statistics: its value is doubled when we double both the number of particles N and the volume V. It respects the additive character attributed to it by macroscopic thermodynamics!

$$S_{\text{ST}}^{\text{ind.}}(2V,2N,T)\cong k2N\left\{\ln\left[\dfrac{2V}{2N}\left(\dfrac{2\pi mkT}{h^2}\right)^{3/2}\right]+\dfrac{5}{2}\right\}=2S_{\text{ST}}^{\text{ind.}}(V,N,T)\tag{9.7.807}$$

From these both last relation we can deduce all the other properties of ideal gases:

- The "caloric equation" of state for the ideal gas:

  $$\dfrac{1}{T}=\left(\dfrac{\partial S_{\text{ST}}(V,N,T)}{\partial E}\right)_{V,N}=Nk\dfrac{3}{2}\dfrac{1}{E}\tag{9.7.808}$$

  Therefore:

  $$E=\dfrac{3}{2}NkT=U\tag{9.7.809}$$

  The prior-previous relation is often denoted:

  $$\boxed{\dfrac{1}{T}=\dfrac{\partial S}{\partial E}=\dfrac{\partial k_B\ln\left(\Omega(E)\right)}{\partial E}}\tag{9.7.810}$$

and considered as the most general definition of "thermodynamic temperature". It can be compared by the reader with the naive definition of the kinetic temperature as we saw it earlier at page 3921.

Remark

Notice that in ISO 80000-5:2007 *Quantities and units — Part 5: Thermodynamics*, the thermodynamics temperature unit Kelvin is defined as (...): "*unit of thermodynamic temperature that is the fraction* $1/273.16$ *of the thermodynamic temperature of the triple point of water*".

Remember as already mentioned during our study of kinetic temperature that the thermodynamic relation above doesn't gives us any threshold to define what is "hot" and what is "cold". Indeed, these two concepts are completely subjective to every human is not even precisely defined by any ISO standard (don't forget as already mentioned in the Introduction chapter of that book that it is not because a word exist in human vocabulary that the underlying concept has a physical reality!). The nearest thing we can find is the ISO 13732-1 *Ergonomics of the thermal environment - Methods for the assessment of human responses to contact with surfaces*. Especially the chapter *5.2 Identification of hot touchable surface* and the Part 3 treating of *Cold surfaces*. But no precise temperatures are given, only a protocol to define it statistical based on a test group.

- The "thermal equation of state" for ideal gas:

$$\frac{P}{T}=\left(\frac{\partial S_{\text{ST}}(V,N,T)}{\partial V}\right)_{E,T}=\frac{kN}{V} \tag{9.7.811}$$

Therefore:

$$PV=NkT \tag{9.7.812}$$

9.7.3.8 Generalized Bernoulli equation (for gases)

To determine the "generalized Bernoulli equation (for gases)" (quite important for our study later of aerospace engineering) we consider the behaviour of a particular parcel of fluid as it flows along a stream line. We assume the system is non-dissipative; that is, we assume that viscosity and thermal conductivity make a negligible contribution to the energy budget. For simplicity, we assume a steady pressure distribution.

We start by invoking the law of conservation of energy. As always, this law says that any change in the energy within the parcel must be exactly balanced by the transfer of energy across the boundary of the parcel. That is:

$$\Delta(\text{change inside boundary}) = -\Phi(\text{energy outward across boundary}) \tag{9.7.813}$$

The energy inside the parcel can take various forms, including kinetic energy, gravitational energy, and the energy stored in the springiness of the fluid. In a gas such as air, the latter is sometimes named the "thermal energy", but of course the laws of thermodynamics apply to all forms of energy, not just the air-spring energy. The relation below restates what we have just said, expressing it in mathematical terms:

$$\mathrm{d}E = -P\mathrm{d}V + \mathrm{d}(mgh) + \mathrm{d}\left(\frac{1}{2}mv^2\right) = -P\mathrm{d}V + mg\mathrm{d}h + mv\mathrm{d}v \tag{9.7.814}$$

This last relation could describe almost any parcel under almost any conditions. To make progress, we need to invoke the fact that our parcel is flowing along a streamline, in thermodynamic equilibrium with the surrounding fluid. We continue to assume that viscosity and thermal conductivity are not significant, so no entropy is being created and no entropy is flowing across the boundary of the parcel.

In thermodynamic equilibrium, the total entropy of the parcel and its surroundings is at a maximum with respect to small changes in P, V, h, et cetera. In the present situation, where the environment exerts a pressure on the parcel, this is provably equivalent to saying that the enthalpy of the parcel not counting the surroundings is at a minimum. Therefore we are interested in (see section Thermodynamcis page 3643):

$$\mathrm{d}H = \mathrm{d}(E + PV) = \mathrm{d}E + P\mathrm{d}V + V\mathrm{d}P = 0 \tag{9.7.815}$$

You may ask, how can we assert thermodynamical equilibrium, when the parcel is not even in mechanical equilibrium? That is, what about the fact that the parcel is accelerating and decelerating as it flows along? That's a good question, for which we have a good answer: Our derivation hinges on the fact that even though the parcel is exchanging energy with the surroundings, it does so in a way that does not exchange any entropy. The only way this can happen is if $\mathrm{d}H = 0$. This depends on our assumption that viscosity and thermal conductivity are negligible.

Therefore:

$$\mathrm{d}H = (-P\mathrm{d}V + mg\mathrm{d}h + mv\mathrm{d}v) + P\mathrm{d}V + V\mathrm{d}P = V\mathrm{d}P + mg\mathrm{d}h + mv\mathrm{d}v = 0 \tag{9.7.816}$$

We wish to integrate this relation, but before we can integrate the $V\mathrm{d}P$ term, we need to know the equation of state. We assume a polytropic equation of state, namely (see section Thermodynamics page 3649):

$$PV^{\gamma} = P_0V_0^{\gamma} \tag{9.7.817}$$

Rearranging a bit:

$$V=V_0P_0^{1/\gamma}P^{-1/\gamma} \tag{9.7.818}$$

Therefore:

$$\mathrm{d}H=V_0P_0^{1/\gamma}P^{-1/\gamma}\mathrm{d}P+mg\mathrm{d}h+mv\mathrm{d}v=0 \tag{9.7.819}$$

After integration:

$$H=V_0P_0^{1/\gamma}\frac{\gamma}{\gamma-1}P^{\frac{\gamma-1}{\gamma}}+mgh+\frac{1}{2}mv^2=c^{te} \tag{9.7.820}$$

where we have absorbed a constant of integration into our definition of H.

As:

$$V_0P_0^{1/\gamma}P^{\frac{\gamma-1}{\gamma}}=V_0P_0^{1/\gamma}P^{\frac{-1}{\gamma}}P=VP \tag{9.7.821}$$

Then the prior-previous relation can be simplified as:

$$H=\frac{\gamma}{\gamma-1}PV+mgh+\frac{1}{2}mv^2=c^{te} \tag{9.7.822}$$

We can make these equations more practical by eliminating the explicit dependence on V in favour of P. We can do this by using:

$$PV^\gamma=P_0V_0^\gamma \tag{9.7.823}$$

Then:

$$\begin{aligned}H&=V_0P_0^{1/\gamma}\frac{\gamma}{\gamma-1}P^{\frac{\gamma-1}{\gamma}}+mgh+\frac{1}{2}mv^2=\frac{\gamma}{\gamma-1}V_0P_0\left(\frac{P}{P_0}\right)^{\dfrac{\gamma-1}{\gamma}}+mgh+\frac{1}{2}mv^2\\&=\frac{\gamma}{\gamma-1}\frac{\rho_0}{m}P_0\left(\frac{P}{P_0}\right)^{\dfrac{\gamma-1}{\gamma}}+mgh+\frac{1}{2}mv^2\end{aligned} \tag{9.7.824}$$

But as:

$$\frac{P}{P_0}=\frac{V_0^\gamma}{V^\gamma}=\left(\frac{m/\rho_0}{m/\rho}\right)^\gamma=\left(\frac{\rho}{\rho_0}\right)^\gamma \tag{9.7.825}$$

Therefore:

$$\begin{aligned}H&=\frac{\gamma}{\gamma-1}\frac{\rho_0}{m}P_0\left(\frac{P}{P_0}\right)^{\frac{\gamma-1}{\gamma}}+mgh+\frac{1}{2}mv^2=\frac{\gamma}{\gamma-1}\frac{\rho_0}{m}P_0\left(\left(\frac{\rho}{\rho_0}\right)^\gamma\right)^{\frac{\gamma-1}{\gamma}}+mgh+\frac{1}{2}mv^2\\&=\frac{\gamma}{\gamma-1}\frac{\rho_0}{m}P_0\left(\frac{\rho}{\rho_0}\right)^{\gamma-1}+mgh+\frac{1}{2}mv^2\end{aligned} \tag{9.7.826}$$

That's it!

Now as $H=E+PV$ and:

$$H=\frac{\gamma}{\gamma-1}PV+mgh+\frac{1}{2}mv^2=c^{te} \tag{9.7.827}$$

We get:

$$\boxed{E=H-PV=\frac{1}{\gamma-1}PV+mgh+\frac{1}{2}mv^2=c^{te}} \tag{9.7.828}$$

Any of the results in this section can be considered the general form of the Bernoulli equation.

For air tables give $\gamma\cong 1.4$ at 293 [K]. It is interesting to notice that for when $\gamma=2$ we fall back on the Bernoulli equation that we have determined during our study of liquids. But obviously we do not fall back on the generalized Bernoulli equation for liquids (see page 3810).

9.7.3.9 Magnus (gravitational) Effect

The (gravititional) "Magnus effect", named after Heinrich Gustav Magnusl, the German physicist who investigated it, is observable phenomenon that is commonly associated with a spinning object moving through a fluid[44]. The path of the spinning object is deflected in a manner that is not present when the object is not spinning. The deflection can be explained by the difference in pressure of the fluid on opposite sides of the spinning object. The Magnus effect is dependent on the speed of rotation. The special case of the force on a rotating cylinder is known as the "Kutta–Joukowski lift", after Martin Kutta and Nikolai Zhukovsky (or Joukowski), who first analyzed the effect.

Remark

The German physicist Heinrich Gustav Magnus described the effect in 11852 (holocene calendar). However, in 11672 (holocene calendar), Isaac Newton had described it and correctly inferred the cause after observing tennis players in his Cambridge college. In 11742 (holocene calendar), Benjamin Robins, a British mathematician, ballistics researcher, and military engineer, explained deviations in the trajectories of musket balls in terms of the Magnus effect.

The most readily observable case of the Magnus effect is when a spinning sphere (or cylinder) curves away from the arc it would follow if it were not spinning. It is often used by association football and volleyball players, baseball pitchers, and cricket bowlers. Consequently, the phenomenon is important in the study of the physics of many ball sports. It is also an important factor in the study of the effects of spinning on guided missiles - and has some engineering uses, for instance in the design of rotor ships and Flettner aeroplanes.

Topspin in ball games is defined as spin about a horizontal axis perpendicular to the direction of travel that moves the top surface of the ball in the direction of travel. Under the Magnus effect, topspin produces a downward swerve of a moving ball, greater than would be produced by gravity alone. Backspin produces an upwards force that prolongs the flight of a moving ball.

---

[44] The Magnus effect is due to the fluid dynamics around a rotating object, so it should theoretically work in any fluid. It is by the way also experimentally observed in water and not only in the air!

Likewise side-spin causes swerve to either side as seen during some baseball pitches, e.g. slider. The overall behaviour is similar to that around an aerofoil (see lift force), but with a circulation generated by mechanical rotation rather than shape of the foil.

9.7.3.9.1 Naive approach for the spherical and cylinder rotating case

The Magnus effect explains commonly observed deviations from the typical trajectories or paths of spinning balls in sport, notably association football, table tennis, tennis, volleyball, golf, baseball, and cricket.

The curved path of a golf ball known as slice or hook is largely due to the ball's spinning motion (about its vertical axis) and the Magnus effect, causing a horizontal force that moves the ball from a straight line in its trajectory. Backspin (upper surface rotating backwards from the direction of movement) on a golf ball causes a vertical force that counteracts the force of gravity slightly, and enables the ball to remain airborne a little longer than it would were the ball not spinning: this allows the ball to travel farther than a ball not spinning about its horizontal axis.

In table tennis, the Magnus effect is easily observed, because of the small mass and low density of the ball. An experienced player can place a wide variety of spins on the ball. Table tennis rackets usually have a surface made of rubber to give the racket maximum grip on the ball to impart a spin.

A rotating ball moving through the air will curl due to the Magnus effect, a force acting perpendicular to the direction and the rotational axis of the ball. Here we present some practical experiments, simulations and methods to calculate the trajectory.

In June 11997 (holocene calendar), Roberto Carlos scored an infamous goal in a 35 [m] free kick that still baffles the viewer. How can the ball behave like this, going off in one direction and then magically curling toward the goal? The answer is that the ball is spinning in the air and is subjected to the Magnus force. To analyse the trajectory of a ball, we need to assess three forces acting on the ball: gravity F_g, the Magnus force F_M (in the direction where the high velocity of the air flow and therefore the low pressure is!) and the drag force F_D:

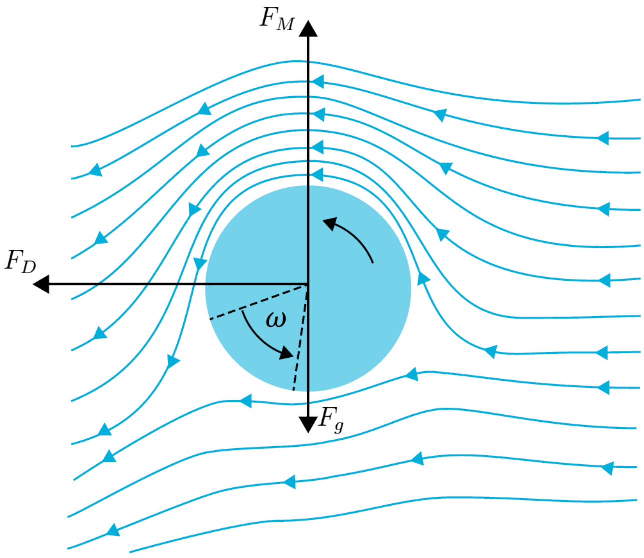


Figure 9.283: **Magnus effect on a rotating sphere/ball**

The gravitational force is simply given by Newton's second law, $F_g = mg$, where m is the mass of the ball and g is the gravitational acceleration.

The Magnus force F_M occurs due to differences in pressure on opposing sides of the ball. The changes in pressure can be described via the Bernoulli principle. For a point on a surface moving through a medium with the velocity v, the total pressure p is equal to the surrounding static pressure p_0 plus the dynamic pressure q, where ρ is the density of the medium, in our case the density of air. But when a ball or cylinder with a radius R is rotating (with an angular velocity of ω in radians per second), a point on the surface on one side of the ball is subjected to a higher flow of air $(v+\omega R)$ (in the most optimal case[45]) than the opposing point on the other side $(v-\omega R)$ (in the most optimal case too). Hence we can derive the difference in pressure for the Bernoulli equation:

$$\frac{1}{2}\rho v^2 + \rho g z + P = c^{te} \tag{9.7.829}$$

Conservation of energy lead us to write:

$$\frac{1}{2}\rho v_1^2 + \rho g z_1 + P_1 = \frac{1}{2}\rho v_2^2 + \rho g z_2 + P_2 \tag{9.7.830}$$

Hence explicitly:

$$\frac{1}{2}\rho(v-\omega R)^2 + \rho g z_1 + P_1 = \frac{1}{2}\rho(v+\omega R)^2 + \rho g z_2 + P_2 \tag{9.7.831}$$

Assuming $z_1 \cong z_2$, this simplifies to:

$$\frac{1}{2}\rho(v-\omega R)^2 + P_1 = \frac{1}{2}\rho(v+\omega R)^2 + P_2 \tag{9.7.832}$$

If we calculate the difference:

$$\Delta P = P_2 - P_1 = 2\rho\omega v R \tag{9.7.833}$$

Therefore:

$$F_M = \Delta P \cdot S = (2\rho\omega v R)\cdot S \tag{9.7.834}$$

Now the reader must keep in mind the symmetry of objects like a sphere or cylinder spinning in an air flow (i.e. to their relative movement in the air)! Because of that symmetry, if we consider the case of the sphere/ball that only half of surface geometry experiments a lift force, every force is replicated on every point of the surface and the resultant of every of theses forces points perpendicular to the cross-section of the sphere/ball. The idea is same for the cylinder, only it cross-section (perpendicular to the airflow) must be taken into account as all resultant forces are perpendicular to that cross-section. Hence the latter relation, will be denoted:

$$F_M = \Delta P \cdot S = (2\rho\omega v R)\cdot S_{\perp\vec{\Phi}} \tag{9.7.835}$$

Therefore considering that only half of the surface is submitted to the force difference:

---

[45] Because in reality it's a fraction of the term ωR, because the rotational speed ($\vec{v} = \vec{\omega}\times\vec{r}$) isn't constant everywhere!

- For a cylinder (or radius r and height h):

$$F_{M,\text{cyl.}}=(2\rho\omega vR)(2Rh)=4\rho\omega vR^2h \qquad (9.7.836)$$

- For a sphere :

$$F_{M,\text{sph.}}=(2\rho\omega vR)(\pi R^2)=2\rho\omega v\pi R^3 \qquad (9.7.837)$$

But as the flow isn't completely influenced by the rotation (only a fraction of the ωR), we should keep in mind that these both relations above are upper-bound, hence approximation of the naive gravitational Magnus effect. We should write instead, explicitly these are approximation and as fraction of the maximum. Therefore:

$$\boxed{\begin{aligned}F_{M,\text{cyl.}}&\cong 4C_s\rho\omega vR^2h\\ F_{M,\text{sph.}}&\cong 2C_s\rho\omega v\pi R^3\end{aligned}} \qquad (9.7.838)$$

where C_s is a coefficient taking into account the fraction of ωR involved.

Remark

A treatment taking into account the fact that the rotational speed of the sphere isn't constant everywhere (i.e. $\vec{v}=\vec{\omega}\times\vec{r}$) leads to: $\tilde{F}_{M,\text{sph.}}\cong 4/3C_s\rho\omega v\pi R^3<F_{M,\text{sph.}}$.

Finally, we have to assess the drag force F_D. Drag is complicated, as the airflow can be laminar or turbulent, depending to a large extent on the shape of the object and the nature of the fluid it is moving in. For our experiments it suffices to assume that the flow is laminar (as in the figure above) and use the standard drag equation where the force is directed in the opposite direction to v and proportional to the velocity (valid for reminder only for small sphere and low speeds as derived on page 3853):

$$F_D=\beta v \qquad (9.7.839)$$

β is a constant that depends on the properties of the fluid and the dimensions of the object; for a football in air it is $\beta=0.142\,[\text{kg}\cdot\text{s}^{-1}]$.

Experimental observation shows that the Magnus force behaves quite like a centripetal force. In the most simple configuration we can consider it as the centripetal force of a circular trajectory (instead of an elliptic, parabolic or hyperbolic one) of radius r. Then we have for the ball case, in the approximation of a circular trajectory:

$$F_{M,\text{sph.}}\cong 2C_s\rho\omega v\pi R^3\cong\frac{mv^2}{r} \qquad (9.7.840)$$

Rearranging:

$$r=\frac{mv}{2C_s\pi\rho\omega R^3} \qquad (9.7.841)$$

Now remember that during our study of the ballistic equation without friction in the section of Classical Mechanics (see page 3265), we have derived:

$$y=x\tan(\beta)-\frac{g}{2v_0^2\cos^2(\beta)}x^2 \qquad (9.7.842)$$

If we put $\beta = 0$, this simplifies to:

$$y = -\frac{g}{2v_0^2}x^2 \tag{9.7.843}$$

And we can write a instead of g to consider the case where acceleration may be something else that the vertical acceleration of gravity. Then (we can get rid also of the negative sign that is not very relevant):

$$y = \frac{a}{2v_0^2}x^2 \tag{9.7.844}$$

Using:

$$F_{M,\text{sph.}} \cong 2C_s\rho\omega v_0\pi R^3 = ma \Rightarrow a \cong \frac{2C_s\rho\omega v_0\pi R^3}{m} \tag{9.7.845}$$

Hence for a perfect sphere in a laminar regime and assuming ... no friction, we get:

$$\boxed{y \cong \frac{C_s\rho\omega\pi R^3}{v_0 m}x^2} \tag{9.7.846}$$

If the reader wants to play with ta formula, the average penalty kick goes flying towards the goal at a speed $136\,[\text{km}\cdot\text{h}^{-1}]$ (i.e. $v_0 \cong 37\,[\text{m}\cdot\text{s}^{-1}]$), density of air under normal surface temperature and pressure $1.225\,[\text{kg}\cdot\text{m}^{-3}]$, the mass of soccer ball $0.450\,[\text{kg}]$, the average radius is $0.1098\,[\text{m}]$, and $C_s \cong 0.13$ and $\omega \cong 12\pi\,[\text{rad}\cdot\text{s}^{-1}]$. Hence:

$$y \cong 0.00367C_s x^2 \tag{9.7.847}$$

The famous kick of Roberto Carlos in 11997 (holocene calendar) was at a distance $x = 35$. Hence $y \cong 4.5C_s$ (without wind and remembering all approximations and oversimplified assumptions we did earlier...!). That is a deflection angle, if $C_s = 1$ of approximately $\delta = \sin(4.5/35) \cong 7.37°$. A more realistic treatment would give... $\delta = 0.39°$.

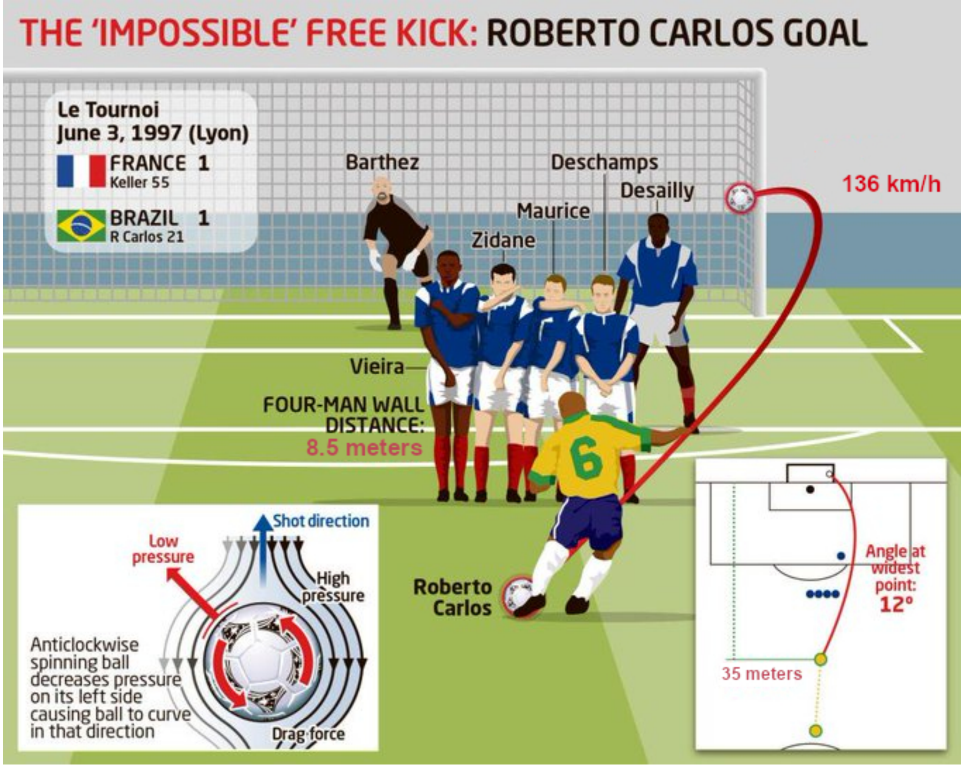

THE 'IMPOSSIBLE' FREE KICK: ROBERTO CARLOS GOAL
Le Tournoi
June 3, 1997 (Lyon)
FRANCE 1
Keller 55
BRAZIL 1
R Carlos 21
Barthez
Deschamps
Desailly
Maurice
Zidane
136 km/h
Vieira
FOUR-MAN WALL
DISTANCE:
8.5 meters
Low pressure
Shot direction
High pressure
Anticlockwise spinning ball decreases pressure on its left side causing ball to curve in that direction
Drag force
Roberto Carlos
Angle at widest point:
12°
35 meters

Figure 9.284: **Roberto Carlos "impossible free kick"**

So our model is very bad or an external factor is missing. In 12017 (holocene calendar) Roberto Carlos revealed his iconic free kick was wind-aided by a wind that came in from the right...

Anyway, an acceptable treatment - i.e. more realistic - of the topic would need an entire textbook. That's why for such cases it is better to move to numerical Finite Elements Methods (computer simulations).

9.7.3.10 Summary of gas laws

Here is a summary of the gas laws proven so far we have seen so far on request of a few readers (for reminder these laws assume ideal gas behavior, i.e. particles have no volume and exert no intermolecular forces on one another and temperatures are always in Kelvin):

1. **Boyle's Law:** If you compress a gas into half its volume, the particles hit the walls twice as often, doubling the pressure

$$P_1V_1=P_2V_2 \tag{9.7.848}$$

2. **Charles's Law:** Heating a gas increases particle kinetic energy. To keep pressure constant, the gas must expand

$$\frac{V_1}{T_1}=\frac{V_2}{T_2} \tag{9.7.849}$$

3. **Gay-Lussac's Law:** Heating a gas in a rigid container causes particles to collide more frequently, driving up the pressure

$$\frac{P_1}{T_1}=\frac{P_2}{T_2} \tag{9.7.850}$$

4. **Avogadro's Law:** Equal volumes of gases under identical conditions contain an equal number of molecules, regardless of the gas's chemical identity

$$\frac{V_1}{n_1}=\frac{V_2}{n_2} \tag{9.7.851}$$

5. **Combined Gas Law**: Used to calculate changes in a fixed amount of gas when pressure, volume, and temperature vary simultaneously (merges Boyle's, Charles's, and Gay-Lussac's laws into one equation)

$$\frac{P_1V_1}{T_1}=\frac{P_2V_2}{T_2} \tag{9.7.852}$$

6. **Ideal Gas Law:** he foundational equation of state for a hypothetical ideal gas, combining all the fundamental empirical laws

$$PV=nRT \tag{9.7.853}$$

7. **Van der Waals Law:** Improves the ideal gas law to more realistic gas including two supplementary physical property of the gas

$$\left(P+\frac{an^2}{V^2}\right)(V-nb)=nRT \tag{9.7.854}$$

8. **Dalton's Law of Partial Pressures:** The total pressure exerted by a mixture of non-reactive gases is equal to the sum of the individual pressures that each gas would exert if it were alone in a isolated system

$$P_{\text{total}}=P_1+P_2+P_3+\ldots \tag{9.7.855}$$

9. **Amagat's Law of Partial Volumes:** The total volume of a gas mixture is equal to the sum of the partial volumes of its components (the volume each gas would occupy if maintained at the total pressure and temperature of the mixture)

$$V_{\text{total}}=V_1+V_2+V_3+\ldots \tag{9.7.856}$$

10. **Barometric Law:** If we assume the temperature remains uniform throughout the air column, the pressure P at a given height h is given by the classic exponential equation

$$P(h)=P_0\exp\left(-\frac{Mgh}{RT}\right) \tag{9.7.857}$$

    Hence without gravity (g) there are no pressure gradient in a closed container.

11. **Isoperimetric Law:** Any gas under its own mass will tend to a collapse into a sphere to minimize its energy state (see the section of Geometry for the detailed proof)

$$4\pi\mathcal{A}(\text{Int}(\gamma))\leq\mathcal{L}(\gamma)^2 \tag{9.7.858}$$

12. **Planetary Stability Gas Law:** Any gas satisfying the following relation (see section Astronomy for the detailed proof):

$$M_{\text{min}}=\frac{3kTRN_A}{2GM_m} \tag{9.7.859}$$

    will be maintained around an object of mass M_{min} in vacuum without needing a container (assuming there are no external influences like other bodies source of gravity or stellar wind, etc.)

The reader must keep in mind there are plenty of other gas laws but that we didn't prove so fare in this textbook!

9.7.4 Plasmas

We define a "plasma" as a state of matter in which certain electronic links have been broken, causing the emergence of free, negatively charged and positively charged ions. Low-ionized gases named "plasmas" by language abuse have the same mechanical properties (flows, acoustic waves, etc.) as neutral gases. On the other hand, their electromagnetic properties (electrical conductivity, refractive index) differ by the presence of free electrons within them.

Plasma has become increasingly important as their application to Tokamak (nuclear fusion reactor) and Pulsed plasma thruster for space engine have become technically possible.

Figure 9.285: **Some well known everyday applications of plasma in the 120th and 121st century**

Remark

Plasma is also named as we already know (see section Thermodynamics page 3609) the "fourth (classical) state of matter" (after solid, liquid and gaseous states).

Caution! Fire is a chemical reaction - combustion - that produces hot gases and light, while plasma is a state of matter made of ionized particles that respond strongly to electric and magnetic fields; ordinary flames are only weakly ionized and therefore not considered true plasma, which requires far higher temperatures and full ionization, and unlike plasma, combustion can only occur if fuel, oxygen, and sufficient heat are all present. Therefore stars are not ball of fire, but ball of plasma!

In their normal state, gases are electrical insulators. This is because they do not contain free charged particles, but only neutral molecules. However, if we apply a fairly strong electric fields to them, they become conductors. The complex phenomena that occur then are named "discharges in gases" and are due to the appearance of electrons and free ions.

The result of a discharge in a gas is therefore the production of an ionized gas containing, for example, an average density of electrons $n_e$equations, positive ion n_i and n_0 neutral items (atoms or molecules). In general, the gas is macroscopically neutral. We have then:

$$n_e = n_i = n \tag{9.7.860}$$

Or otherwise expressed, neutrality is also written:

$$q_i n_i + q_e n_e = 0 \tag{9.7.861}$$

This neutrality is the consequence of the very intense electrostatic forces that appear as soon as one has equation. The particle density is therefore the first fundamental quantity.

The "degree of ionization" of a gas is defined by the ratio:

$$\alpha = \frac{n}{n_0 + n} \tag{9.7.862}$$

where n_0 is the density of neutral particles and n that of the electrons (or positive ions). The value of the degree of ionization in the various types of ionized gas varies in practice from very low values, from the order of 10^{-10}, for example, to 1.

The second fundamental quantity is the temperature T. When heating a gas at a sufficiently high temperature (of the order of 10^4 [T]), the mean kinetic energy (see virial theorem page 3897):

$$E_c = \frac{3}{2} kT \tag{9.7.863}$$

of its molecules can become of the same order as their ionization energy E_i. Under these conditions, when two molecules collide, there can be ionization of one of them.

If the gas is in thermodynamic equilibrium, the collision ionization is counterbalanced by recombination processes between electrons and ions and the result is that the three equation variables α, n and T are not independent: ionization is determined by pressure and temperature, we say then the gas is in "equilibrium state of thermal ionization".

At higher temperatures, the gas atoms can be ionized several times. In many cases, ionization is due to an external electric field (+ an external magnetic fields sometimes to stabilize the process), and the gas is not in thermodynamic equilibrium. It will often reach a stationary state that can be characterized by the parameters α, n, T_e (temperature of the electrons), T_i (temperature of the ions) and T_0 (temperature of the molecules).

The three temperatures thus introduced are defined by the condition that $3/2kT_a$ represents the average kinetic energy of the particles of species a, in a frame where they have a mean zero velocity. The difference between T_e, T_i and T_0 can be important! for example, in a typical discharge tube, we can have: $T_0 \cong T_i \cong 300$ [K] and $T_e \cong 3 \cdot 10^4$ [K]. The high value of T_e is due to the action of the electric field on the electrons, and the ionization is then produced by the collisions of these hot electrons on the neutral molecules of the gas.

In conclusion, there are only two basic quantities to characterize a plasma: the density and the electronic temperature. We shall now turn to two other important but non-fundamental magnitudes in the sense that they are expressed on the basis of density and temperature.

9.7.4.1 Plasma Frequency

The "plasma oscillations", also known as "Langmuir waves" (after Irving Langmuir) or "plasma frequency", are rapid oscillations of the electron density in conducting media such as plasma or metals. The oscillations can be described as an instability in the dielectric function of a free electron gas. The frequency only depends weakly on the wavelength of the oscillation. The quasi-particle resulting from the quantization of these oscillations is the "plasmon".

If in an initially neutral plasma we produce a local perturbation in the form of an excess of positive or negative electrical charge, it will tend to return to the state of equilibrium of neutrality. However, we can easily see that the initial perturbation generates an undamped pendular oscillation of the plasma around its equilibrium state. Consider, for example, the situation shown in the figure below:

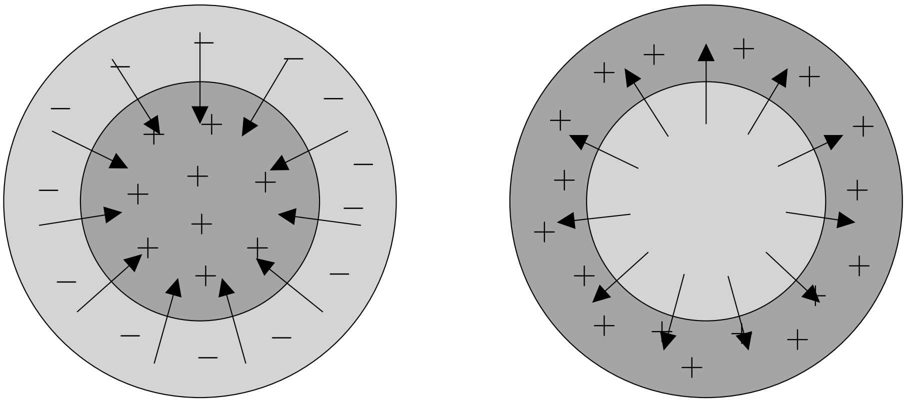


Figure 9.286: **Plasma electronic oscillation**

At the initial moment the region at the center contains an electron deficit and the region all around an excess of electrons. This produces an electric field tending to generate a movement of the electrons in the direction of the arrows. In this movement, these will acquire a certain kinetic energy and they will be able, after a given time, to go beyond the position of equilibrium. Too many electrons left the outer region, there will be a defect of electrons in this region and an electric field tending to bring them back towards it. After a certain time, the initial situation is reconstructed and the cycle begins again. The vibrations thus produced are the "electronic plasma oscillations" we have defined previously.

We can quantitatively study this problem by posing the general equations of an electronic charge oscillation and with the following simplifying assumptions:

- H1. The ions are assumed immobile since they are much heavier than electrons, and their quantity is equal to n_{i0}
- H2. Thermal agitation is negligible (assumption named "cold electron plasma oscillation model").
- H3. The collisions are negligible (we speak then of model of "free transport")
- H4. The oscillations are of small amplitude
- H5. There is no electric or magnetic field imposed by external sources

Now, let us recall that we have proved in the section of Electrodynamics that (equation of conservation of the electric charge):

$$\frac{\partial \rho_Q}{\partial t}+\vec{\nabla}\circ\vec{j}=0 \tag{9.7.864}$$

and in the section of Electrokinetics that:

$$\vec{j}=\rho_Q\vec{v} \tag{9.7.865}$$

Therefore, by adopting the aforementioned notations, it comes locally:

$$\boxed{\frac{n_e}{\partial t}+\vec{\nabla}\circ(n_e\vec{v}_e)=0} \tag{9.7.866}$$

which constitutes the "hydrodynamic equation of electrons".

Remark

A plasma is theoretically generally neutral, but we may theoretically have a non-neutral volume. It is this hypothesis which allows us to postulate that the divergence of the current is not zero!

Let us now recall the Coulomb's force (see section Electrostatics page 3968):

$$\vec{F}_e=m_e\vec{a}_e=m_e\frac{\mathrm{d}\vec{v}_e}{\mathrm{d}t}=q_e\vec{E} \tag{9.7.867}$$

where we have clearly neglected the kinetic pressure term and the collision term (hypotheses H2 and H3) and neglected the magnetic field associated with the oscillation.

We can simplify these equations using the hypothesis H4 in the form:

$$n_e(\vec{r},t)=n_{e0}+n_{\text{el}}(\vec{r},t) \tag{9.7.868}$$

where $n_{\text{el}}(\vec{r},t)$ is a small perturbation around the equilibrium.

Let us suppose moreover that variable quantities vary at the pulsation ω, we can therefore write:

$$E(\vec{r},t)=E(\vec{r})e^{-\mathrm{i}\omega t}\qquad\text{and}\qquad n_{\text{el}}(\vec{r},t)=n_{\text{el}}(\vec{r})e^{-\mathrm{i}\omega t} \tag{9.7.869}$$

Consequently, the hydrodynamic equation of the electrons becomes:

$$\begin{aligned}\frac{\partial n_e}{\partial t}+\vec{\nabla}\circ(n_e\vec{v}_e)&=\frac{\partial(n_{e0}+n_{\text{el}})}{\partial t}+\vec{\nabla}((n_{e0}+n_{\text{el}})\vec{v}_e)\\&=\frac{\partial n_{\text{el}}}{\partial t}+\vec{\nabla}\circ(n_{e0}\vec{v}_e)+\underbrace{\vec{\nabla}\circ(n_{\text{el}}\vec{v}_e)}_{\cong 0}=\frac{\partial n_{\text{el}}}{\partial t}+\vec{\nabla}\circ(n_{e0}\vec{v}_e)\\&=\frac{\partial n_{\text{el}}}{\partial t}+n_{e0}\vec{\nabla}\circ\vec{v}_e=\frac{\partial n_{\text{el}}(\vec{r})e^{-\mathrm{i}\omega t}}{\partial t}+n_{e0}\vec{\nabla}\circ\vec{v}_e=-\mathrm{i}\omega n_{\text{el}}(\vec{r})e^{-\mathrm{i}\omega t}+n_{e0}\vec{\nabla}\circ\vec{v}_e\\&=-\mathrm{i}\omega n_{el}+n_{e0}\vec{\nabla}\circ\vec{v}_e\end{aligned}\tag{9.7.870}$$

Thus finally:

$$-\mathrm{i}\omega n_{\text{el}}+n_{e0}\vec{\nabla}\circ\vec{v}_e=0\tag{9.7.871}$$

From the expression of Coulomb's force we also deduce:

$$\mathrm{d}\vec{v}_e=\frac{q_e}{m_e}\vec{E}\mathrm{d}t\tag{9.7.872}$$

From where we get:

$$\vec{v}_e=\frac{q_e}{m_e}\int\vec{E}\mathrm{d}t=\frac{q_e}{m_e}\int E(\vec{r})e^{-\mathrm{i}\omega t}\mathrm{d}t=\frac{q_e}{m_e}E(\vec{r})\int e^{-\mathrm{d}\omega t}\mathrm{d}t=-\frac{q_e}{\mathrm{i}\omega m_e}E(\vec{r})\tag{9.7.873}$$

But on the other hand we have the Gauss's law (see section Electrodynamics page 4097):

$$\vec{\nabla}\circ\vec{E}=\frac{n_{i0}q_i+n_{e0}q_e+n_{\text{e1}}q_e}{\varepsilon_0}=\frac{n_{\text{el}}q_e}{\varepsilon_0}\tag{9.7.874}$$

indeed, as taking into account the neutrality condition of the undisturbed plasma, we have:

$$(n_{i0}+n_{e0}q_e)=0\tag{9.7.875}$$

Therefore:

$$\vec{\nabla}\circ\vec{E}-\frac{n_{\text{e1}}q_e}{\varepsilon_0}=0\tag{9.7.876}$$

From the relation proved previously:

$$\vec{v}_e=\frac{q_e}{-\mathrm{i}\omega m_e}\vec{E}\tag{9.7.877}$$

replacing in the following relation (also proved just earlier above):

$$-\mathrm{i}\omega n_{\text{el}}+n_{e0}\vec{\nabla}\circ\vec{v}_e=0\tag{9.7.878}$$

we get:

$$n_{\text{el}}=\frac{n_{e0}q_e}{\omega^2 m_e}\vec{\nabla}\circ\vec{E}\tag{9.7.879}$$

Finally, replacing the latter expression in Gauss's law, we get:

$$\nabla\circ\vec{E}-\frac{n_{e0}q_e^2}{\omega^2 m_e\varepsilon_0}\nabla\circ\vec{E}=\nabla\circ\vec{E}\left(1-\frac{n_{e0}q_e^2}{\omega^2 m_e\varepsilon_0}\right)=0\tag{9.7.880}$$

But in space charge oscillations we have by definition:

$$\nabla \circ \vec{E} \neq 0 \tag{9.7.881}$$

Which leads us to:

$$\frac{n_{e0}q_e^2}{\omega^2 m_e \varepsilon_0} = 1 \Leftrightarrow \omega = \sqrt{\frac{n_{e0}q_e^2}{m_e \varepsilon_0}} \tag{9.7.882}$$

Finally, we get the "plasma frequency for cold electrons" or "Langmuir frequency for cold electrons" given by:

$$\boxed{f_{\text{ce}} = \sqrt{\frac{n_{e0}q_e^2}{4\pi^2 m_e \varepsilon_0}}} \tag{9.7.883}$$

So we fall back on a relation we already derived in the section of Electrokinetics at page 4240.

In physics, the plasma frequency is thus the characteristic frequency of plasma waves, that is to say the oscillations of the electric charges present in conductive media, such as metal or plasmas. Like the electromagnetic wave which, quantified, is described by photons, this plasma wave is quantified in "plasmons" as we have already mention it before.

The oscillations of the electric charges can be understood by the following reasoning: if the electrons of a zone of the plasma are moved, then the ions of this zone, having little moved due to their large mass, will exert on these electrons an attractive Coulomb force. So they're going back to their original position, and so on...

Notice that f_{ce} depends only on physical constants and concentration of electrons n_{e0}. The numeric expression for plasma ordinary frequency is then:

$$f_{\text{ce}} \cong 8.98\sqrt{ne} \tag{9.7.884}$$

9.7.4.2 Lawson Criterion for D-T

In nuclear fusion research, the "Lawson criterion", first derived on fusion reactors (initially classified) by John D. Lawson in 11955 (holocene calendar) and published in 11957 (holocene calendar), is a general system measure that defines the conditions needed for a fusion reactor to reach ignition, that is, that the heating of the plasma by the products of the fusion reactions is sufficient to maintain the temperature of the plasma against all losses without external power input. As originally formulated the Lawson criterion gives a minimum required value for the product of the plasma (electron) density n_e and the "energy confinement time" τ_E.

We are interested in how long the plasma contains energy. We use the quantity τ_E, named the "confinement time". This is a measure of the rate at which the system loses energy to the environment. It can be defined as the energy content of the plasma divided by the power loss:

$$\tau_E = \frac{E}{P_{\text{loss}}} \tag{9.7.885}$$

where obviously E is the energy content of the plasma, and P_{loss} is the power loss.

The thermal energy of a plasma can be defined as (see virial theorem page 3897):

$$E=\int\frac{3}{2}k(n_{e-}T_{e-}+(n_D+n_T))T_{\text{ions}})\mathrm{d}V \tag{9.7.886}$$

Where k is Boltzmann's constant, n_{e-} is the electron density, T_{e-} is the electron temperature, n_D and n_T are the ion densities of deuterium and tritium respectively and T_{ions} is their temperature. The integral is over the volume. If we assume that all temperatures are the same, and that the densities of tritium and deuterium are equal (that why the $1/2$ vanish!), we get:

$$\frac{E}{V}=3n_ekT \tag{9.7.887}$$

Where in this equation, n_e is the electron density and V is the volume.

The number of fusions per volume per time is given by:

$$f=n_Dn_T\langle\sigma v\rangle=\left(\frac{1}{2}n_e\right)\left(\frac{1}{2}n_e\right)\langle\sigma v\rangle=\frac{1}{4}n_e^2\langle\sigma v\rangle \tag{9.7.888}$$

Where f is the number of fusions per volume per time, n_D and n_T are the ion densities of deuterium and tritium respectively, σ is the fusion cross-section (in [m^2] for recall..), v is the relative velocity. The angle brackets mean an average over the Maxwellian velocity distribution. We are also assuming that the densities of the ions are equal, and that their total density is equal to the total electron density n_e. This implies that (as already assumed in the previous above!):

$$n_D=n_T=\frac{1}{2}n_e \tag{9.7.889}$$

For ignition state, we require that the same amount of energy be produced and kept within the plasma as that which leaves the plasma. The rate of heating per volume is defined as the product of f and E_{ch}, the energy of the charged fusion products. We are assuming here (quite accurately) that the neutron emissions contribute nothing to plasma heating.

So we require:

$$fE_{\text{ch}}\geq P_{\text{loss}} \tag{9.7.890}$$

Subbing in our definitions of f and P_{loss} we get:

$$\frac{1}{4}n_e^2\langle\sigma v\rangle E_{\text{ch}}\geq\frac{3n_ekT}{\tau_E} \tag{9.7.891}$$

Rearranging leads to the standard definition of the Lawson Criterion.

$$n_e\tau_E\geq\frac{12}{E_{\text{ch}}}\frac{kT}{\langle\sigma v\rangle} \tag{9.7.892}$$

The right hand side of this equation is called the L function. Where:

$$\boxed{L=\frac{12}{E_{\text{ch}}}\frac{kT}{\langle\sigma v\rangle}} \tag{9.7.893}$$

The minimum value of L for D-T is reached near the temperature of $25keV$.

The quantity $T/\langle\sigma v\rangle$ is a function of temperature with an absolute minimum. Replacing the function with its minimum value provides an absolute lower limit for the product $n_e\tau_E$. This is the Lawson criterion!

This image, now modified slightly, was originally from wikimedia commons.

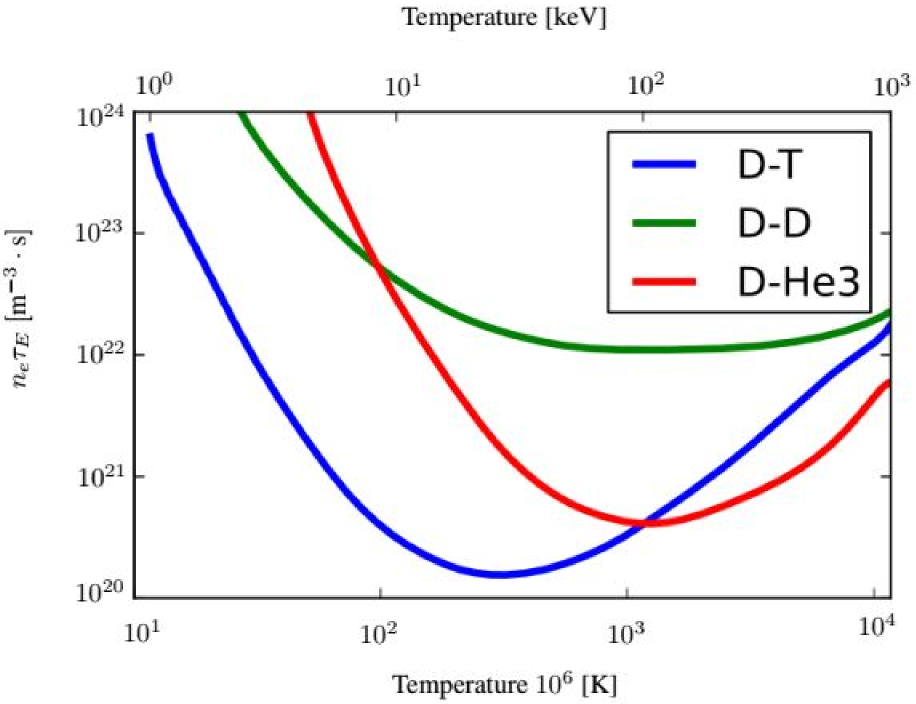


Figure 9.287: **Minimum value of $n_e\tau_E$ required for self-heating, for three fusion reactions (source: Wikipedia)**

For D-T the absolute minimum for the product $n_e\tau_E$ is:

$$n_e\tau_E \geq 1.5\cdot 10^{20}\,[\mathrm{s}\cdot\mathrm{m}^{-3}] \tag{9.7.894}$$

The Lawson criterion applies to inertial confinement fusion[46] (ICF) as well as to magnetic confinement fusion (MCF) but in the inertial case it is more usefully expressed in a different form. A good approximation for the inertial confinement time τ_E is the time that it takes an ion to travel over a distance R at its average thermal speed in three dimensions (see section Probabilities page):

$$\langle v_{\text{th}}\rangle=\sqrt{\langle v^2\rangle}=\sqrt{\frac{3kT}{m_i}} \tag{9.7.895}$$

where m_i denotes mean ionic mass. The inertial confinement time τ_E can thus be approximated as:

$$\tau_E\cong\frac{R}{\langle v_{\text{th}}\rangle}=\frac{R}{\sqrt{\dfrac{3kT}{m_i}}}=R\sqrt{\frac{m_i}{3kT}} \tag{9.7.896}$$

---

[46] Type of fusion energy research that attempts to initiate nuclear fusion reactions by heating and compressing a fuel target, typically in the form of a pellet that most often contains a mixture of deuterium and tritium.

By substitution of the above expression into the Lawson criterion relation, we get:

$$n_e \tau_E \approx n \cdot R \cdot \sqrt{\frac{m_i}{3kT}} \geq \frac{12}{E_{\text{ch}}} \frac{kT}{\langle \sigma v \rangle} \tag{9.7.897}$$

Hence:

$$n_e \cdot R \gtrapprox \frac{12}{E_{\text{ch}}} \frac{\sqrt{3}\,(kT)^{3/2}}{\langle \sigma v \rangle \cdot m_i^{1/2}} \tag{9.7.898}$$

Finally:

$$R \gtrapprox \frac{12}{E_{\text{ch}}} \frac{\sqrt{3}\,(kT)^{3/2}}{\langle \sigma v \rangle \cdot m_i^{1/2} n_e} \tag{9.7.899}$$

That you can find also sometimes in the literature in the case of the one dimension average thermal speed as:

$$R \gtrapprox \frac{12}{E_{\text{ch}}} \frac{(kT)^{3/2}}{\langle \sigma v \rangle \cdot m_i^{1/2} n_e} \tag{9.7.900}$$

9.7.4.3 Saha ionization equation

The Saha ionisation equation is an expression that relates the ionisation state of a gas in thermal equilibrium to the temperature and pressure. The equation is a result of combining ideas of Quantum Mechanics and Statistical Mechanics and is used to explain the spectral classification of stars or the recombination temperature after the Big Ban. The expression was developed by Indian physicist Meghnad Saha in 11920 (holocene calendar).

For a gas at a high enough temperature and or density, the thermal collisions of the atoms will ionize some of the atoms, making an ionized gas. When several or more of the electrons that are normally bound to the atom in orbits around the atomic nucleus are freed, they form an independent electron gas cloud co-existing with the surrounding gas of atomic ions and neutral atoms. In turn, this generates an electric field, where the motion of charges generates currents, making a localised magnetic field, and generates the state of plasma.

The Saha equation describes the degree of ionization for any gas in thermal equilibrium as a function of the temperature, density, and ionization energies of the atoms and only holds for weakly ionized plasmas for which the Debye length is large. This means that the screening of the Coulomb interaction of ions and electrons by other ions and electrons is negligible. The subsequent lowering of the ionization potentials and the "cutoff" of the partition function is therefore also negligible.

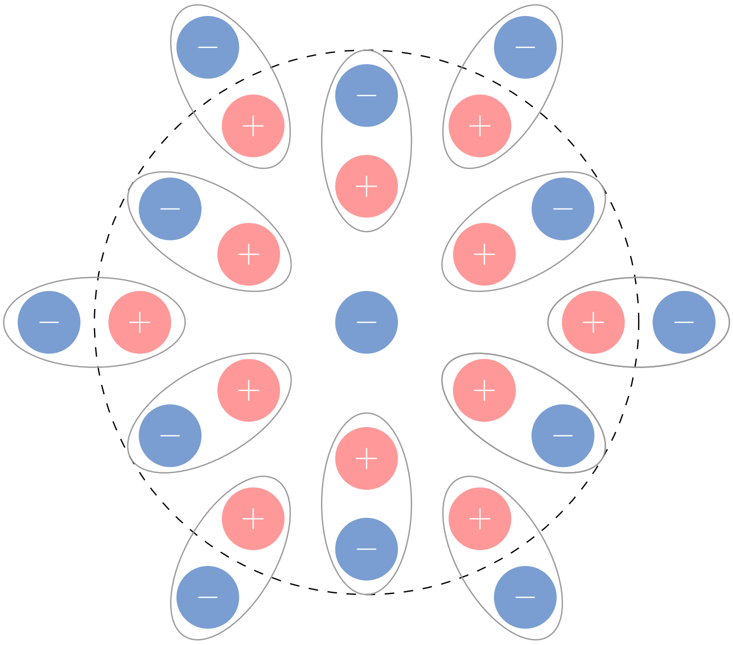


Figure 9.288: **Example of a naive case of screening of an electron by multiple dipoles**

So the question is to know what is the equilibrium ionization state of a gas at a given temperature?

Consider the Hydrogen example that will interest us the most for our study of the CMB (cosmological microwave background):

$$e+p \rightleftharpoons \mathrm{H}+\gamma \tag{9.7.901}$$

Let us define the total number of baryons:

$$N_{\mathrm{b}}=N_p+N_{\mathrm{H}} \tag{9.7.902}$$

and an ionization fraction (assuming charge neutrality $N_e=N_p$):

$$x_e:=\frac{N_p}{N_b}=\frac{N_e}{n_b}=\frac{N_p}{N_p+N_{\mathrm{H}}}=\frac{N_e}{N_e+N_{\mathrm{H}}} \tag{9.7.903}$$

We can also write:

$$1-x_e=1-\frac{N_e}{(N_p+N_{\mathrm{H}})}=\frac{N_p+N_{\mathrm{H}}-N_e}{(N_p+n_{\mathrm{H}})}=\frac{N_{\mathrm{H}}}{(N_p+N_{\mathrm{H}})} \tag{9.7.904}$$

Dividing (if we divide the numerator and denominator by the volume V^2 we get densities obviously):

$$x_e^2=\frac{N_e^2}{(N_p+N_{\mathrm{H}})^2}=\frac{n_e^2}{(n_p+n_{\mathrm{H}})^2} \tag{9.7.905}$$

by the above relation, we get:

$$\frac{x_e^2}{1-x_e}=\frac{n_e^2}{(n_p+n_{\mathrm{H}})^2}\cdot\frac{(n_p+n_{\mathrm{H}})}{n_{\mathrm{H}}}=\frac{n_e^2}{(n_p+n_{\mathrm{H}})}\cdot\frac{1}{n_{\mathrm{H}}} \tag{9.7.906}$$

That latter result can also be often in the literature written as (still assuming charge neutrality!):

$$\frac{n_p n_e}{n_{\mathrm{H}} n_b}=\frac{x_e^2}{1-x_e} \tag{9.7.907}$$

We will now consider the number densities defined by a distribution function in thermal equilibrium. e and p are non-relativistic at the energy scales of recombination.

As we have proved it in the section of Electrokinetics that Maxwell-Boltzmann density states given for electrons by (see page 4294):

$$\rho_C(E)\cong 2\left(\frac{m_e kT}{2\pi\hbar^2}\right)^{3/2}\exp\left(\frac{\mu-E_C}{kT}\right) \tag{9.7.908}$$

So adapting that result to our case (also the notation!), and degeneracy level, it becomes:

$$n\cong g\left(\frac{m_e kT}{2\pi\hbar^2}\right)^{3/2}\exp\left(\frac{\mu-mc^2}{kT}\right)=g\left(\frac{m_e kT}{2\pi\hbar^2}\right)^{3/2}e^{-\left(mc^2-\mu\right)/kT} \tag{9.7.909}$$

Therefore, for particles and atoms involved in Hydrogen binding energy:

$$\begin{aligned}n_p&\cong g_p e^{-\left(m_p c^2-\mu_p\right)/kT}\left(m_p kT/2\pi\hbar^2\right)^{3/2}\\ n_e&\cong g_e e^{-\left(m_e c^2-\mu_e\right)/kT}\left(m_e kT/2\pi\hbar^2\right)^{3/2}\\ n_{\mathrm{H}}&\cong g_{\mathrm{H}} e^{-\left(m_{\mathrm{H}} c^2-\mu_{\mathrm{H}}\right)/kT}\left(m_{\mathrm{H}} kT/2\pi\hbar^2\right)^{3/2}\end{aligned} \tag{9.7.910}$$

We know (and we have calculated) the Hydrogen first level binding energy. We will denote it here by $B=13.605[\mathrm{eV}]$ (doing the strong assumption that ionization energy that we know today had the same value a short time after the Big Bang...). Therefore we have:

$$m_{\mathrm{H}}=m_p+m_e-B/c^2 \tag{9.7.911}$$

hence the "Saha ionization equation for Hydrogen plasma":

$$\boxed{\frac{n_p n_e}{n_{\mathrm{H}} n_b}=\frac{x_e^2}{1-x_e}\cong\frac{g_p g_e}{g_{\mathrm{H}} n_b}e^{-B/kT}e^{\left(\mu_p+\mu_e-\mu_{\mathrm{H}}\right)/kT}\left(\frac{m_e m_p}{m_{\mathrm{H}}}\right)^{3/2}\left(\frac{kT}{2\pi\hbar^2}\right)^{3/2}} \tag{9.7.912}$$

Remark

Some textbooks do the following complementary approximation:

$$\frac{n_p n_e}{n_{\mathrm{H}} n_b}=\frac{x_e^2}{1-x_e}\cong\frac{g_p g_e}{g_{\mathrm{H}} n_b}e^{-B/kT}e^{\left(\mu_p+\mu_e-\mu_{\mathrm{H}}\right)/kT}\left(\frac{m_e kT}{2\pi\hbar^2}\right)^{3/2} \tag{9.7.913}$$

Putting in spin degeneracy of the different components $g_p=2$, $g_e=2$, $g_{\mathrm{H}}=4$ and assuming equilibrium $\mu_p+\mu_e=\mu_{\mathrm{H}}$, we get[47]:

$$\frac{x_e^2}{1-x_e}\cong\frac{1}{n_b}e^{-B/kT}\left(\frac{m_e m_p}{m_{\mathrm{H}}}\right)^{3/2}\left(\frac{kT}{2\pi\hbar^2}\right)^{3/2} \tag{9.7.914}$$

---

[47] The reader must remember that $\mu_\gamma=0$ because photons are freely generated and destroyed in recombination reactions

So it's a quadratic equation involving T and the total density and without any of the μ_i. We have an exponential dominant factor: ionization drops quickly as kT drops below B - exactly where the sharp transition occurs depends on the density n_b.

After simplification and assuming $n_e=n_p$ we may also get:

$$\frac{n_e^2}{n_{\text{H}}}\cong e^{-B/kT}\left(\frac{m_e m_p}{m_{\text{H}}}\right)^{3/2}\left(\frac{kT}{2\pi\hbar^2}\right)^{3/2} \tag{9.7.915}$$

But another common shape is using the baryon to photon ratio $\eta=n_\gamma/n_b$ (see page 5972) to trade the rightover n_b. Then we get:

$$\frac{x_e^2}{1-x_e}\cong\frac{\eta}{n_\gamma}e^{-B/kT}\left(\frac{m_e m_p}{m_{\text{H}}}\right)^{3/2}\left(\frac{kT}{2\pi\hbar^2}\right)^{3/2} \tag{9.7.916}$$

We will use that relation in the section of Cosmology for our study of the Cosmological Microwave Background matter/photon decoupling (see page 5973).

9.7.4.4 Ionospheric propagation

The text and proofs that follow are almost a copy/paste of the course notes *Ionospheric Propagation* by Prof. Sean Victor Hum. He kindly authorized us to reproduce it here with slight modifications!

The ionosphere exists between about 90 and 1000 [km] above the earth's surface. Radiation from the Sun (actually a G2V-type star, aka yellow dwarf) ionizes atoms and molecules here, liberating electrons from molecules and creating a space of free electron and ions. Subjected to an external electric field from a radio signal, these free and ions will experience a force and be pushed into motion. However, since the mass of the ions is much larger than the mass of the electrons, ionic motions are relatively small and will be ignored here.

Free electron densities on the order of 10^{10} to 10^{12} electrons per cubic metre are produced by ionization from the sun's rays. Layers of high densities of electrons are given special names called the D, E, and F layers (during the day the F layer splits into two layers called the F_1 and F_2 layers, while the D layer vanishes completely at night). These regions do not have sharp boundaries, and the altitudes at which they occur vary during the course of a day and from season to season. The D region is the lowest, starting about 60 or 70 [km] above the ground and extending upward to about 90 [km]. Next higher is the E region, starting at about 90 or 100 [km] up and extending to 120 or 150 [km]. The uppermost part of the ionosphere, the F region, starts about 150 [km] and extends far upward, sometimes as high as 500 [km] above the surface of our home planet.

The regions of the ionosphere are not considered separate layers, such as the more familiar troposphere and stratosphere. Instead, they are ionized regions embedded within the standard atmospheric layers. The D region usually forms in the upper part of the mesosphere, while the E region typically appears in the lower thermosphere and the F region is found in the upper reaches of the thermosphere.

Before communication via satellites became common, the operators of radio communication systems often used the ionosphere to extend the range of their transmissions. Radio waves generally travel in straight lines, so unless a tall transmission tower can "see" the top of a receiver tower, the curvature of the Earth limits the range of radio transmissions to stations that are not over the horizon. However, some frequencies of radio waves bounce or reflect off of the electrically charged particles in certain ionosphere layers. Pre-satellite radio communications often took advantage of this phenomenon, bouncing radio waves off of the "sky" to extend the range of the signals. Radio operators had to account for the constant changes in the ionosphere, particularly the shifts or disappearance of the layers between day and night, to effectively take advantage of these mirror-like reflections of radio waves.

The ionosphere regions can absorb or dampen radio signals, or they can bend radio waves, as well as reflecting the signals as described above. The specific behavior depends on both the frequency of the radio signal as well as the characteristics of the ionosphere region involved. Since Global Positioning System (GPS) satellites use radio signals to determine locations, the accuracy of GPS can be severely reduced when those signals bend as they pass through ionosphere regions. Similarly, some radio communications can be disrupted if the frequency used is one that an ionosphere layer dampens or absorbs entirely, resulting in a weakened signal or even total loss of communications. Scientists constantly measure and produce computer models of the ever-changing ionosphere so that people in charge of radio communications can anticipate disruptions.

Radio waves below 40 [MHz] are significantly affected by the ionosphere, primarily because radio waves in this frequency range are effectively reflected by the ionosphere. The E and F layers are the most important for this process. For frequencies beyond 40 [MHz], the wave tend to penetrate through the atmosphere versus being reflected.

The major usefulness of the ionosphere is that the reflections enable wave propagation over a much larger distance than would be possible with line-of-sight or even atmospheric refraction effects. This is shown graphically below:

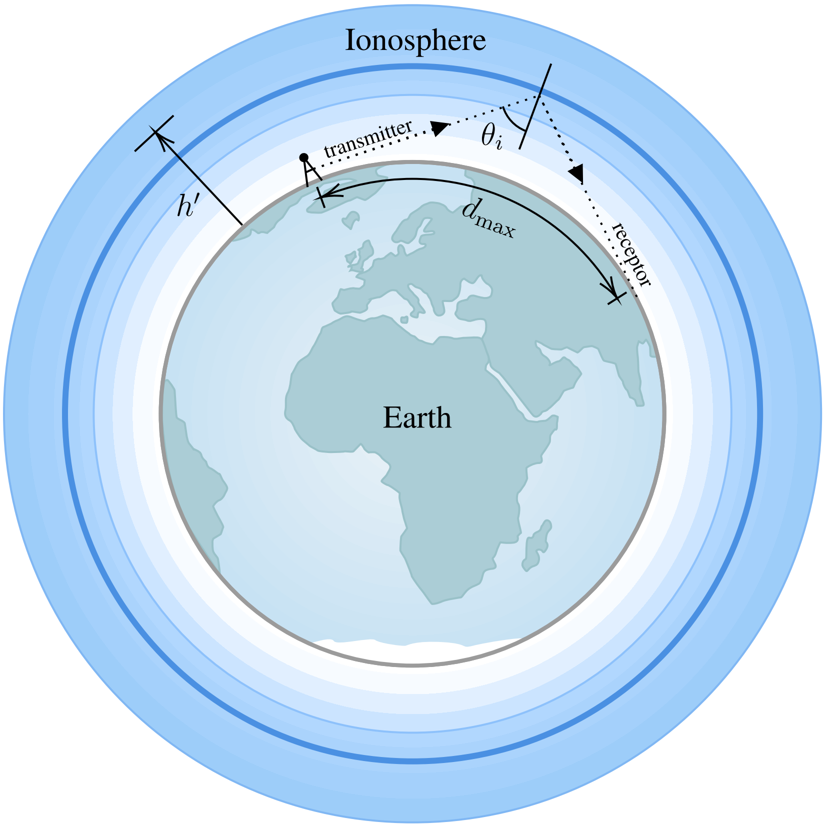


Figure 9.289: **A single skip of a radio wave using the ionosphere**

The skip distance $d_{\max}$ can be very large, allowing very large communication distances. This is further enhanced by multiple reflections between the ionosphere and the ground, leading to multiple skips. This form of propagation allows shortwave and amateur radio signals to propagate worldwide. Since the D layer disappears at night, the best time for long-range communications is at night, since the skip distance is larger as the E, and F regions are at higher altitudes.

Where does the reflection come from? The reflections from the ionosphere are actually produced by refraction as the wave propagates through the ionosphere. The ionosphere is a concentrated region highly charged ions and electrons that collective form an ionized gas or plasma. This gas has a dielectric constant that is a function of various parameters, including the electron concentration and the frequency of operation. We now derive the dielectric constant of a plasma.

The electric field will produce a force on a given electron and displace it along a vector $\vec{r}$, as shown on the left side of the figure below:

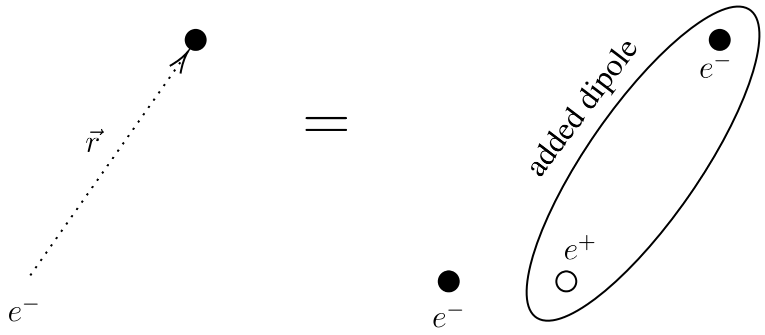


Figure 9.290: **Representation of a moving electron as a dipole moment**

The displacement of an electron along this path can be modelled in an equivalent situation where the original electron remained stationary and an equivalent electric dipole is added, as shown in the right half of the figure. The dipole moment of this dipole is then equal to $-e\vec{r}$ (see section Electrostatics page 3986). If there are N electrons per unit volume, each displaced by $\vec{r}$ on average, then the volume polarization is:

$$\vec{P}=-Ne\vec{r} \tag{9.7.917}$$

The equation of motion for a single electron of mass $m_e=9.109\cdot 10^{-31}$ [kg], charge $e=1.6021\cdot 10^{-19}$ [C], with velocity $\vec{v}=\mathrm{d}\vec{r}/\mathrm{d}t$, and acted upon by an electric field $\vec{E}$, is:

$$m_e\frac{\mathrm{d}\vec{v}}{\mathrm{d}t}=-e\vec{E} \tag{9.7.918}$$

The electron also experiences a frictional force resulting from collisions with neutral molecules. This force is added to the electric field force above, yielding:

$$m_e\frac{\mathrm{d}\vec{v}}{\mathrm{d}t}=-e\vec{E}-\nu m_e\vec{v} \tag{9.7.919}$$

where ν is the electron collision frequency. Re-writing the equation in terms of $\vec{r}$:

$$m_e\frac{\mathrm{d}^2\vec{r}}{\mathrm{d}t^2}=-e\vec{E}-\nu m_e\frac{\mathrm{d}\vec{r}}{\mathrm{d}t} \tag{9.7.920}$$

We know that for sinusoidal fields, we can write equations in terms of phasors (see section Wave Mechanics page 3510) and replace $\mathrm{d}/\mathrm{d}t$ with $\mathrm{i}\omega$. Hence we can write the equation of motion on the electron in terms of phasors as:

$$-\omega^2 m_e\vec{r}=-e\vec{E}-\mathrm{i}\omega\nu m_e\vec{r} \tag{9.7.921}$$

Or:

$$\vec{r}=\frac{e\vec{E}}{\omega^2 m_e-\mathrm{i}\omega\nu m_e}=\frac{e\vec{E}}{m_e\omega^2\left(1-\mathrm{i}\frac{\nu}{\omega}\right)} \tag{9.7.922}$$

Substituting this into $\vec{P}=-Ne\vec{r}$, we get:

$$\vec{P}=-\frac{Ne^2\vec{E}}{m_e\omega^2\left(1-\mathrm{i}\frac{\nu}{\omega}\right)} \tag{9.7.923}$$

The electric flux density can then be found as (see section Electromagnetism page 4108):

$$\vec{D}=\varepsilon_0\vec{E}+\vec{P}=\varepsilon_0\vec{E}-\frac{Ne^2\vec{E}}{m_e\omega^2\left(1-\mathrm{i}\frac{\nu}{\omega}\right)}\equiv\varepsilon_r\varepsilon_0\vec{E} \tag{9.7.924}$$

The "effective relative dielectric constant" of the plasma is then defined by:

$$\varepsilon_r=1-\frac{Ne^2}{m_e\omega^2\varepsilon_0\left(1-\mathrm{i}\frac{\nu}{\omega}\right)} \tag{9.7.925}$$

An "angular plasma frequency" can be defined such that:

$$\omega_p^2=\frac{Ne^2}{m_e\varepsilon_0}\cong 3,183N \tag{9.7.926}$$

We see that in the presence of electron collisions, the dielectric constant can in general be complex. If we ignore collisions for the moment (i.e. $\nu=0$), then:

$$\varepsilon_r=1-\frac{\omega_p^2}{\omega^2}\approx 1-\frac{81N}{f^2} \tag{9.7.927}$$

This relation is knows as the "Eccles relation".

Let us now consider an infinite plane-wave, of frequency, ω so that the wave electric field takes the form:

$$\vec{E}=E_0e^{\mathrm{i}(kx-\omega t)}\hat{\vec{z}} \tag{9.7.928}$$

where it is understood that the physical electric field is the real part of the above expression. A peak or trough of the above wave travels at the so-named "phase velocity", which is for recall given by (see section Mechanics page 3480):

$$v_p=\frac{\omega}{k} \tag{9.7.929}$$

Now, we have also seen that the phase velocity of electromagnetic waves in a dielectric medium is $v_p=c/n=c/\sqrt{\varepsilon}$ (see section Electromagnetism page 4243), so:

$$\omega^2=\frac{k^2c^2}{\varepsilon} \tag{9.7.930}$$

It follows from :

$$\varepsilon_r=1-\frac{\omega_p^2}{\omega^2} \tag{9.7.931}$$

that:

$$\omega^2=k^2c^2+\omega_p^2 \tag{9.7.932}$$

in a plasma. Hence:

$$k=\sqrt{\frac{\omega^2-\omega_p^2}{c^2}}=\frac{\mathrm{i}\sqrt{\omega_p^2-\omega^2}}{c}:=\mathrm{i}|k| \tag{9.7.933}$$

Therefore, the wave electric field takes the form:

$$E_z=E_0e^{\mathrm{i}(\mathrm{i}|k|x-\omega t)}=E_0e^{-|k|x}e^{-\mathrm{i}\omega t} \tag{9.7.934}$$

Indeed, it can be seen that for $\omega<\omega_p$ electromagnetic waves in a plasma take the form of decaying standing waves, rather than travelling waves. We conclude that an electromagnetic wave, of frequency less than the plasma frequency, which is incident on a plasma will not propagate through the plasma. Instead, it will be totally reflected.

We can be sure that the incident wave is reflected by the plasma, rather than absorbed, by considering the energy flux of the wave in the plasma. We have demonstrated that the Poynting vector can be written (see section Electromagnetism page 4156):

$$\vec{S}=\frac{E^2}{\mu_0\omega}\vec{k} \tag{9.7.935}$$

For a wave with a real frequency and a complex $\vec{k}$-vector, the above relation reduces to:

$$\vec{S}=\frac{E^2}{\mu_0\omega}\Re(\vec{k}) \tag{9.7.936}$$

However, according to:

$$k=\sqrt{\frac{\omega^2-\omega_p^2}{c^2}}=\frac{\mathrm{i}\sqrt{\omega_p^2-\omega^2}}{c}:=\mathrm{i}|k| \tag{9.7.937}$$

the $\vec{k}$-vector for a low frequency electromagnetic wave in a plasma is purely imaginary. It follows that the associated energy flux is zero. Hence, any low frequency wave which is incident on the plasma must be totally reflected, since if there were any absorption of the wave energy then there would be a net energy flux into the plasma.

From these observations we can observe two propagation mechanisms for a wave entering the ionosphere. If it is below the plasma frequency, it is simply reflected off the ionosphere. When it is above the plasma frequency, then refraction analogous to that which occurs with atmospheric refraction occurs, though it is much more pronounced because of the large changes in refractive index. In fact, provided the angle at which the wave enters the atmosphere is not too steep, the wave will be bent quite strongly as it enters ionosphere. If the angle of incidence reaches the critical angle (see section Optics page 4369) for total internal reflection, the wave is reflected off the ionosphere and starts to return to Earth, going through the reverse process of refraction as it does so.

The world ground-to-ground radio transmission worldwide record using ionospheric propagation at the day we write these lines was done the 1st January 12020 (holocene calendar), by the amateur radio Nick G4KUX (radio nickname) in the county of Durham in the north of England who managed to work D41CV also on FT8 on 432 [MHz] to $4,644$ kilometres. Nick, G4KUX lives in an elevated location to the west of Bishop Auckland in County Durham in the north of England. His location is about 360 metres above sea level and he has a good take off for radio in all directions. For the record breaking contact with D41CV on Cape Verde Island, Nick was using 400 Watts into a UHF Log-Periodic antenna.

For more detailed and modern approaches than the above one see *The propagation of radio waves over the surface of the earth and in the upper atmosphere* ([419]) and *The ancient and modern history of EM ground-wave propagation* ([565]).

∼90 %

★★★☆☆

27 votes, 80.74%

10

Electromagnetism

Electrodynamics is the field of physics that study the dynamic action of electric currents and the propagation of electromagnetic waves.

Contents

10.1 Electrostatics . . . 3968
10.1.1 Electric Force . . . 3969
10.1.2 Electric Potential . . . 3976
10.1.2.1 Path Independence . . . 3979
10.1.3 Equipotential and Field lines . . . 3980
10.1.3.1 Infinite straight wire . . . 3982
10.1.3.2 Electric Rigid Dipole . . . 3984
10.1.4 Electric Flux . . . 3996
10.1.4.1 Capacitor . . . 3997
10.1.5 Electrostatic potential energy . . . 4005
10.2 Magnetostatics . . . 4008
10.2.1 Ampère's theorem . . . 4011
10.2.1.1 Infinitely long solenoid . . . 4014
10.2.1.2 Toroidal coils . . . 4018
10.2.1.3 Electromagnet . . . 4019
10.2.2 Maxwell-Ampère Relation . . . 4022
10.2.3 Biot-Savart law . . . 4023
10.2.3.1 Magnetic field for a current loop . . . 4025
10.2.3.2 Magnetic field for an infinite wire . . . 4026
10.2.4 Magnetic dipole . . . 4030
10.2.4.1 Magnetic torque . . . 4050
10.2.5 Lorentz law (Lorentz force) . . . 4052

10.2.5.1 Magnetic Vector Potential . . . 4057
10.2.5.2 Work of Magnetic Field . . . 4062
10.2.5.3 Classical Hall effect . . . 4065
10.2.5.4 Larmor radius . . . 4068
10.2.5.5 Energy of a magnetic dipole . . . 4081
10.2.6 Langevin treatment of Diamagnetism and Paramagnetism . . . 4085
10.2.6.1 Langevin model of diamagnetism . . . 4085
10.2.6.2 Langevin model of paramagnetism . . . 4088
10.3 Electrodynamics . . . 4094
10.3.1 Maxwell Equations . . . 4095
10.3.1.1 First Maxwell Equation (constant electric flow) . . . 4095
10.3.1.2 Second Maxwell Equation (non-existence of magnetic monopole) . . . 4099
10.3.1.3 Third Maxwell Equation . . . 4101
10.3.1.4 Fourth Maxwell Equation . . . 4106
10.3.1.5 Magnetic Monopoles . . . 4110
10.3.2 Charge conservation equation . . . 4114
10.3.3 Gauge Theory . . . 4115
10.3.3.1 Electromagnetic field tensor . . . 4121
10.3.4 Electromagnetic wave equation . . . 4138
10.3.4.1 Foucault light speed experiment . . . 4145
10.3.4.2 Helmholtz equation . . . 4152
10.3.4.3 Energy flow transportation (Poynting vector) . . . 4156
10.3.4.4 Emissions . . . 4160
10.3.5 Synchrotron radiation (bremsstrahlung) . . . 4163
10.3.5.1 Liénard-Wiechert potentials . . . 4168
10.3.5.2 Retarded Electric and Magnetic fields . . . 4175
10.3.5.3 Power radiated through an elementary surface . . . 4188
10.3.6 Dipole antenna radiation pattern derivation (Hertz Electric Dipole) . . . 4203
10.3.6.1 The $L=\lambda/2$ (half-wave) far-field resonant antenna . . . 4214
10.4 Electrokinetics . . . 4219
10.4.1 Kirchhoff's laws . . . 4221
10.4.1.1 Mesh law (Kirchhoff's Loop Law) . . . 4221
10.4.1.2 Nodes law (Kirchhoff's Point Law) . . . 4222
10.4.2 Drude model . . . 4226
10.4.3 Ohm's law . . . 4233
10.4.3.1 Complex Conductivity . . . 4238
10.4.3.2 Simple (harmonic oscillator) model of dielectric and conductors (and plasmas) 4240

10.4.3.3 Equivalent Resistance 4246
10.4.3.4 Equivalent Capacities 4248
10.4.3.5 Equivalent Inductance 4250
10.4.4 Electromotive Force 4253
10.4.4.1 Faraday's law of induction 4258
10.4.4.2 Inductance 4261
10.4.5 Electromagnetic Waves in Conductors 4268
10.4.5.1 Skin effect 4268
10.4.5.2 Phase and group velocity of EM waves inside conductors (modified Maxwell equations) 4272
10.4.6 Semiconductors 4279
10.4.6.1 Non-degenerated statistic density of negative electric charge carriers 4291
10.4.6.2 Non-degenerated statistic density of positive electric charge carriers 4296
10.4.6.3 Energy bands (conduction band, band gap, valence band) 4300
10.4.6.4 Ohm's law of Semiconductors 4304
10.4.7 Superconductivity 4307
10.4.7.1 Meissner effect (second London's equation) 4308
10.5 Optics (ray optics) 4314
10.5.1 Sources and Shadows 4315
10.5.2 Colors 4325
10.5.3 Radiometry/Photometry 4337
10.5.3.1 Energy flow 4337
10.5.3.2 Light Intensity (Radiant Intensity) 4339
10.5.3.3 Energy Emittance (Radiant Emittance) 4340
10.5.3.4 Radiance and Luminance 4343
10.5.3.5 Kirchhoff's law of Radiation 4351
10.5.3.6 Spectral Decomposition 4352
10.5.4 Law of Refraction and Reflection 4353
10.5.4.1 Refractive index 4358
10.5.4.2 Snell's law 4364
10.5.4.3 Rainbow 4372
10.5.4.4 The 22° Halo 4385
10.5.4.5 Optical Mirages 4397
10.5.4.6 Cherenkov radiation 4414
10.5.5 Descartes' Formulas 4416
10.5.5.1 Stigmatism 4420
10.5.5.2 Lenses 4426

10.5.5.3 Triangular Prism 4462
10.5.5.4 Pentaprism 4466

10.6 Wave Optics 4470

10.6.1 Huygens' principle 4470
10.6.1.1 Normal refraction 4479
10.6.1.2 Thin film interference 4481
10.6.1.3 Fresnel coefficients 4485
10.6.2 Fraunhofer Diffraction 4492
10.6.2.1 Case of a small rectangular aperture (slit) 4492
10.6.2.2 Case of a network of small rectangular apertures (slits) 4502
10.6.2.3 Young's interference experiment 4509
10.6.3 Fresnel diffraction 4517
10.6.3.1 Case of a rectangular aperture 4517
10.6.3.2 Case of a circular aperture 4521
10.6.3.3 Case of a circular obstacle 4536
10.6.4 Light polarization 4538
10.6.4.1 Linear polarization 4543
10.6.4.2 Elliptical polarization 4544
10.6.4.3 Circular polarization 4546
10.6.4.4 Natural polarization 4547
10.6.4.5 Malus' law 4550
10.6.5 Coherence and interference (interferometry) 4554
10.6.5.1 Michelson Interferometer 4562
10.6.5.2 Sagnac interferometer (Sagnac effect) 4579
10.6.6 LASER 4587
10.6.6.1 LASER Dispersion/Divergence 4595
10.6.7 Fiber Optics 4604

10.1. Electrostatics

So far we have focus only on the gravitational interaction and the characteristic quantity of matter, named "mass" associated with it. We discussed the electromagnetic interaction, analysing macroscopic phenomena such as friction, cohesion, elasticity, the forces of contact, etc. Now we look at electronic forces and the characteristic of matter named "electric charge" associated with them. The electromagnetic interaction binds matter in all its observable forms. It is this that holds the electrons to the nucleus in the atom, which holds together the atoms in molecules, molecules into objects and even your nose to your face...

The "electric charge" generates the "electric force" or "Coulomb force" and we are only beginning to understand this force thanks to Quantum Field Theory (see further below in this textbook). The electric charge is a fundamental concept, which can not be described in terms of more simple and fundamentals concepts. We know it by its effects and unfortunately not by what it is (this was also the case for the mass before the discovery of the Higgs Boson).

Experience has also shown that even if the electric charge is an additive quantity such as the mass, however, it also has negative value (and not exclusively positive as known nowadays for the mass). Thus, in the current language, and as confirmed by experience, two identical sign electric charges repel and two opposite sign electric charges attract (we will see a schematic figure of this further below).

Figure 10.1: **Electric charges exist all around us. They can cause objects to be repelled from each other or to be attracted to each other (credit: modification of work by Sean McGrath)**

People should be wary of charlatans - especially self-proclaimed telekinetics - who exploit simple electrostatic tricks to create the illusion of mystical powers. By subtly charging objects or manipulating lightweight materials, these performers can make items move or react in ways that seem supernatural to the untrained eye. Their goal is not to reveal hidden truths about the universe but to deceive, impress, and ultimately extract money from individuals who may not be familiar with the underlying physics (that's why the will reject to accept to be tested with an electrostatic fieldmeter). Understanding that these effects rely on basic electrostatics, not paranormal abilities, is essential to avoiding manipulation and protecting oneself from financial exploitation.

Let us now see the classic force that is associated with the electric charge:

10.1.1 Electric Force

It has experimentally been established by Coulomb that a reference point particle undergoes a force $\vec{F}$ of an intensity proportional to its charge q, when placed in the neighbourhood of one or more electrical charges Q_i in a medium of absolute electrical permittivity ε (permittivity to the electric field of course...!) given in vector notation and in non-relativistic case by:

$$\vec{F}=q\left(\frac{1}{4\pi\varepsilon}\sum_i\frac{Q_i}{||\vec{r}-\vec{r}_i||^3}(\vec{r}-\vec{r}_i)\right)=q\left(\frac{1}{4\pi\varepsilon}\sum_i\frac{Q_i}{||\vec{r}-\vec{r}_i||^2}\vec{u}_i\right) \tag{10.1.1}$$

where $\vec{r}_i$ is the vector position of the sample charge Q_i and $\vec{r}$ of the reference point charge particle relatively to a same orthonormal vector basis.

In other words, two point electric charged particles attract or repel each other in a force directly proportional to their electric charge and inversely proportional to the square of the distance that separates them.

In the case of a system with two particles separated by a distance r, we have the same simplified relation and we fall back on the most common form of the electric force or "Coulomb force" as given in most textbooks (as scalar and non-relativistic):

$$\boxed{F=\frac{1}{4\pi\varepsilon}\frac{Qq}{r^2}=\frac{1}{4\pi\varepsilon}\frac{q_1q_2}{r^2}} \tag{10.1.2}$$

Hence recalling Newton's second law:

$$F=ma=\frac{1}{4\pi\varepsilon}\frac{Qq}{r^2}=\frac{1}{4\pi\varepsilon}\frac{q_1q_2}{r^2} \tag{10.1.3}$$

Frequently, these relation is defined as the "Coulomb's law" in most schools and admitted as unprovable. In fact, it is not! This relation can be proved as we will see in the study of quantum fields physics (see section Quantum Field Theory page 5164) using the Klein-Gordon equation in the context of a potential field with spherical symmetry (proof performed by Hideki Yukawa).

The law was discovered in 11785 (holocene calendar) by the French physicist Charles-Augustin de Coulomb through a series of memoirs on electricity and magnetism (see [115]). Especially he used a torsion balance (based on torsion pendulum) that is today named the "Coulomb balance" that plenty of high-school around the World have in their physics lab:

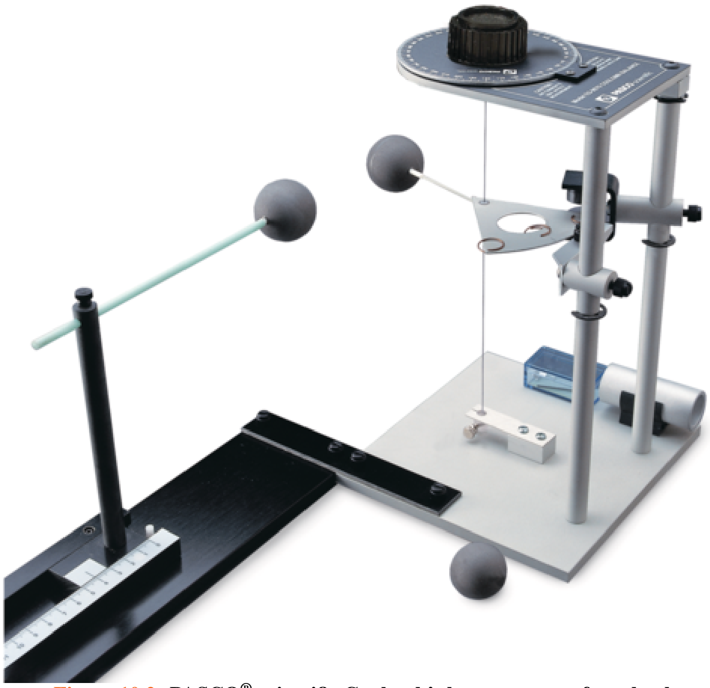


Figure 10.2: **PASCO® scientific Coulomb's law apparatus for schools**

Remark

Don't forget that as 2π appears frequently in many theorems of physics and mathematics (more than π alone), the value 4π is frequently denoted by 2τ as by definition $\tau = 2\pi$.

In either formulation, Coulomb's law is fully accurate only when the objects are stationary, and remains approximately correct only for slow movement. These conditions are collectively known as the electrostatic approximation. When movement takes place, magnetic fields that alter the force on the two objects are produced. The magnetic interaction between moving charges may be thought of as a manifestation of the force from the electrostatic field but with Einstein's theory of relativity taken into consideration.

Remark

For the relativistic form of Coulomb's law, the reader is referred to the section Special Relativity (see 5699) where it is proved that (vector form):

$$\vec{F} = \frac{q(1-v^2/c^2)}{\left[1-(v^2/c^2)\sin^2(\theta)\right]^{2/3}} \frac{Q\vec{r}}{4\pi\varepsilon r^3} \qquad (10.1.4)$$

The value of electric permittivity in vacuum (as we will see during our study of Quantum Field Theory, vacuum isn't really empty) is in turn given experimentally by the "dielectric constant" or simple "electric constant":

$$\varepsilon_0 \cong 8.8541878176\ldots \cdot 10^{-12} \;\left[\mathrm{C}^2 \cdot \mathrm{N}^{-1} \cdot \mathrm{m}^{-2}\right] \qquad (10.1.5)$$

and relatively to medium considered, we define a "relative dielectric permittivity" ε that makes it easier to determine the properties of a material with respect to the electric field so that we have the "absolute electrical permittivity":

$$\varepsilon = \varepsilon_0 \varepsilon_r \tag{10.1.6}$$

It should be mentioned that some authors define the permittivity of vacuum from the speed of light and the magnetic permeability of vacuum (see section Magnetostatics page 4011). Therefore, the value of the electrical permittivity of vacuum is obviously correct by definition. But this only makes sense once known the Maxwell's theory and this will be presented and proved later in the section Electrodynamics (we follow the steps in the historical scientific discoveries).

The SI unit of permittivity is (sadly) the "farad" per meter, such that:

$$\frac{\mathrm{F}}{\mathrm{m}} = \frac{\mathrm{C}}{\mathrm{V}\cdot\mathrm{m}} = \frac{\mathrm{C}^2}{\mathrm{N}\cdot\mathrm{m}^2} = \frac{\mathrm{C}^2\cdot\mathrm{s}^2}{\mathrm{kg}\cdot\mathrm{m}^3} = \frac{\mathrm{A}^2\cdot\mathrm{s}^4}{\mathrm{kg}\cdot\mathrm{m}^3}$$

It also appears in the Coulomb force constant, as the "Coulomb constant":

$$K_e = \frac{1}{4\pi\varepsilon_0} \tag{10.1.7}$$

The factor into parentheses in:

$$\vec{F} = q\left(\frac{1}{4\pi\varepsilon}\sum_i \frac{Q_i}{\|\vec{r}-\vec{r}_i\|^3}(\vec{r}-\vec{r}_i)\right) \tag{10.1.8}$$

depends only on the distribution of charges Q_i in the volume and the absolute electrical permittivity of the medium ε. Since its value varies from one place to another and depends on the position vector $\vec{r}$ of the reference electric charge, it forms a set of vectors, which property is this of a multitude of electrical field lines hence the use of the term "electric field".

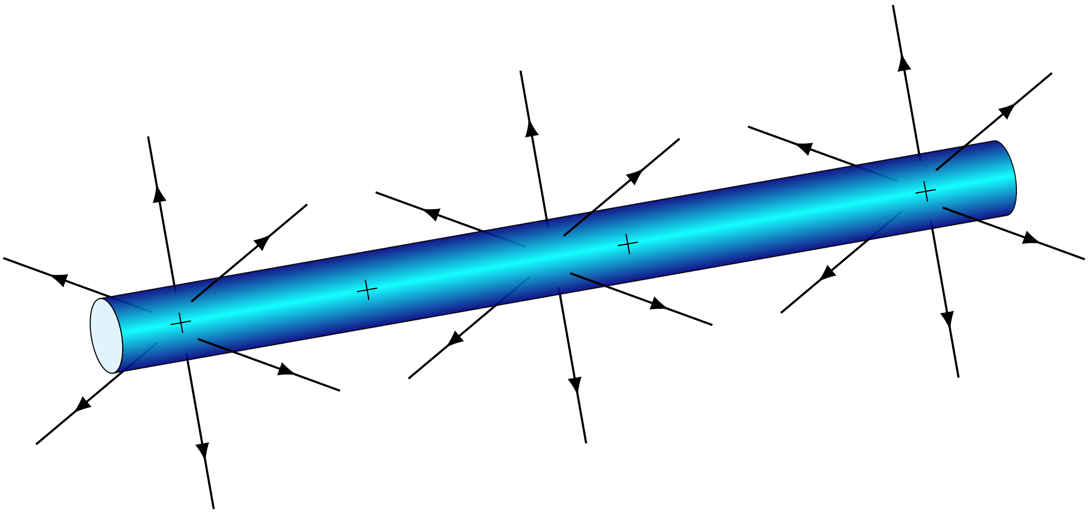


Figure 10.3: **Electric field of an infinite line charge**

The set of these vectors $\vec{E}$ carries therefore the name "electric field", at the point $\vec{r}$,developed in the electric charge distribution Q_i:

$$\vec{E}=\frac{1}{4\pi\varepsilon}\sum_i \frac{Q_i}{\|\vec{r}-\vec{r}_i\|^3}(\vec{r}-\vec{r}_i) \qquad (10.1.9)$$

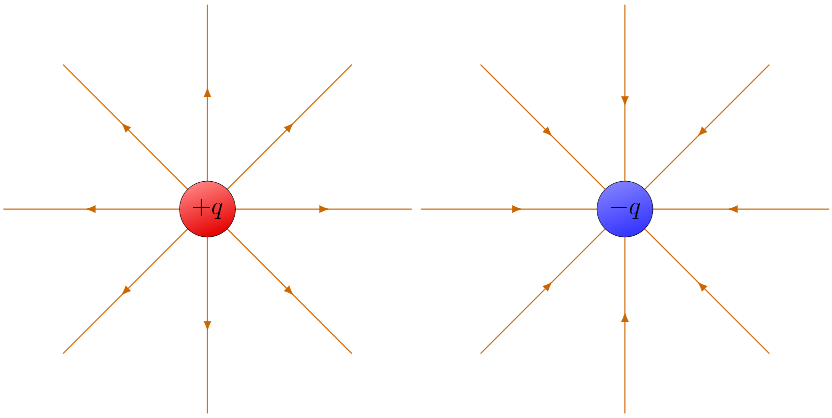
Engineers often use another notation that allows to characterize only the geometry of the field regardless the environment (medium) and for this purpose they introduce the concept of "displacement field":

$$\boxed{\vec{D}=\varepsilon\vec{E}=\varepsilon_0(1+\chi)\vec{E}} \qquad (10.1.10)$$

where χ is the "electric susceptibility". We will meet this vector again in the section Electrodynamics during our synthesis of Maxwell's equations.

Coulomb force acting on the sample charge q, is then written in a conventional way (hence the electric field can simple be interpreted as the force per unit charge!):

$$\boxed{\vec{F}=q\vec{E}(\vec{r})} \qquad (10.1.11)$$



One configuration is of particular interest: two separated point charges of opposite charge. In the limit of vanishing separation, it is named "dipole". Its field fundamentally differs from that of just a single charge even though it is just the sum of the charge. The dipole as a concept is extremely important throughout electrodynamics. It is applied for example to explain the emission of electromagnetic radiation or as a model for molecules.

Let us first consider the case of opposite electric charges. For the given problem we have $\vec{r}_1=-d/2\vec{e}_x$ and $\vec{r}_2=d/2\vec{e}_x$. So the charges lie on the x-axis with a separation d. Remember that $\vec{e}_x$ is the unit vector in x direction (see section Vector Calculus page 1794). The electric field is then given by:

$$\begin{aligned}\vec{E}(\vec{r})=\vec{E}(x,y,z)&=\frac{1}{4\pi\varepsilon}\left(Q\frac{\vec{r}-\vec{r}_1}{\|\vec{r}-\vec{r}_1\|^3}+(-Q)\frac{\vec{r}-\vec{r}_2}{\|\vec{r}-\vec{r}_2\|^3}\right)\\&=\frac{Q}{4\pi\varepsilon}\left(\frac{(x-d/2)\vec{e}_x+y\vec{e}_y+z\vec{e}_z}{[(x-d/2)^2+y^2+z^2]^{3/2}}-\frac{(x+d/2)\vec{e}_x+y\vec{e}_y+z\vec{e}_z}{[(x+d/2)^2+y^2+z^2]^{3/2}}\right)\end{aligned} \qquad (10.1.12)$$

Imagine we are interested in the magnitude and the direction of the field only about the line of the y-axis. To figure it out both, we simply calculate:

$$\begin{aligned}\vec{E}(x=0,y,z=0)&=\frac{Q}{4\pi\varepsilon}\left(\frac{-d/2\vec{e}_x+y\vec{e}_y}{[(-d/2)^2+y^2]^{3/2}}-\frac{(+d/2)\vec{e}_x+y\vec{e}_y}{[(+d/2)^2+y^2]^{3/2}}\right)\\&=\frac{Q}{4\pi\varepsilon}\frac{d\vec{e}_x}{[(-d/2)^2+y^2]^{3/2}}\end{aligned}\tag{10.1.13}$$

We see that the electric field has only a component in x-direction. Because of the symmetric choice of the coordinate system we could have guessed this in the first place.

The magnitude is therefore given by the norm of the electric field:

$$\|\vec{E}(x=0,y,z=0)\|=\frac{Q}{4\pi\varepsilon}\frac{d}{[(d/2)^2+y^2]^{3/2}}=\frac{Q}{4\pi\varepsilon}\frac{1}{d^2}\frac{1}{[(1/2)^2+(y/d)^2]^{3/2}}\tag{10.1.14}$$

Let us now consider the case of equal charges (positive for example):

$$\begin{aligned}\vec{E}(x=0,y,z=0)&=\frac{Q}{4\pi\varepsilon}\left(\frac{-d/2\vec{e}_x+y\vec{e}_y}{[(-d/2)^2+y^2]^{3/2}}+\frac{(+d/2)\vec{e}_x+y\vec{e}_y}{[(+d/2)^2+y^2]^{3/2}}\right)\\&=\frac{2Q}{4\pi\varepsilon}\frac{y\vec{e}_y}{[(d/2)^2+y^2]^{3/2}}\end{aligned}\tag{10.1.15}$$

The direction of the field is in this case always parallel to the y-axis but changing sign at $y=0$. Its magnitude is given by:

$$\|\vec{E}(x=0,y,z=0)\|=\frac{2Q}{4\pi\varepsilon}\frac{|y|}{[(d/2)^2+y^2]^{3/2}}=\frac{2Q}{4\pi\varepsilon}\frac{1}{y^2}\frac{1}{[(d/2y)^2+1]^{3/2}}\tag{10.1.16}$$

We find that for equal charges the magnitude of the electric field decreases for large y.

With Maple™ 4.00b we can easily plot these two magnitudes:

```
>plot(1/(0.5^2+y^2)^(3/2),1/(y^2)*1/((1/(2*y))^2+1)^(3/2),y=-5..5);
```

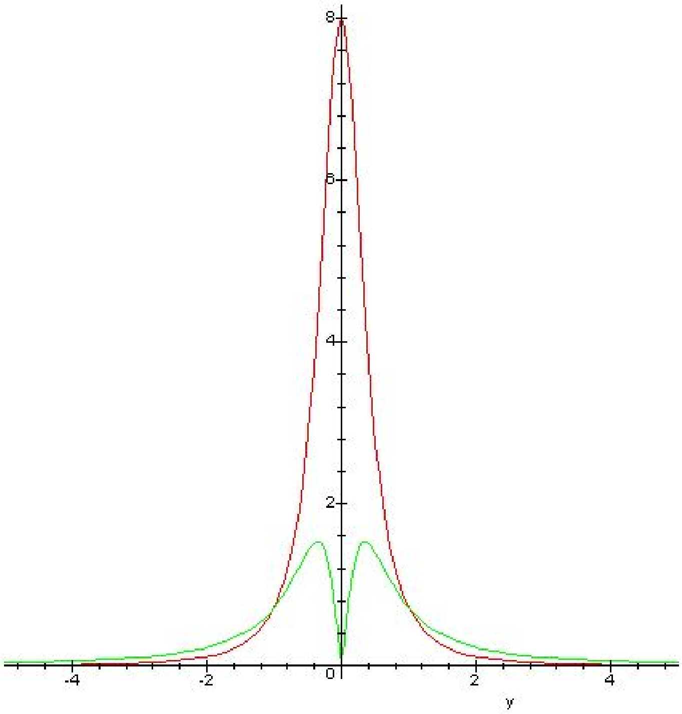


Figure 10.4: **Dipole profile for $q_1 = -q_2$ and $q_1 = q_2$ along y-axis**

This is how looks like some lines of electric field of two identical charges particles ($q_1 = q_2$) with a nice perspective effect:

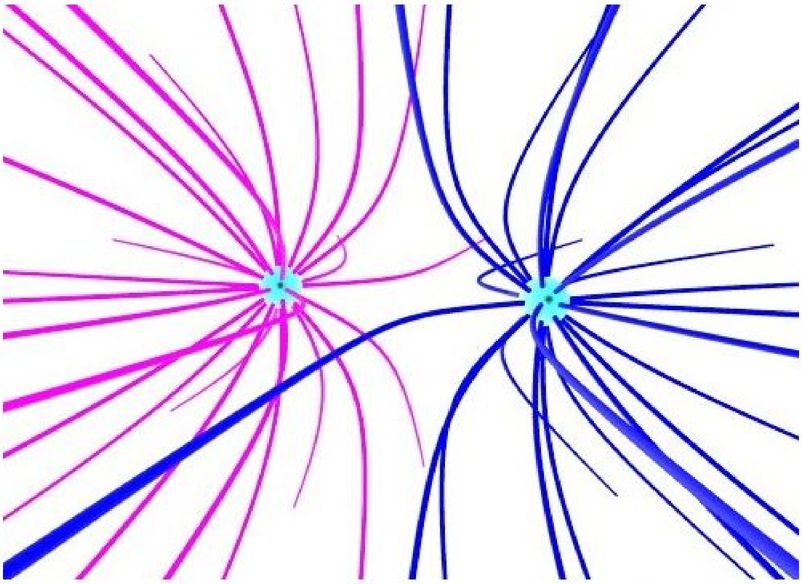


Figure 10.5: **3D dipole profile with $q_1 = q_2$**

Or more traditionally in 2D:

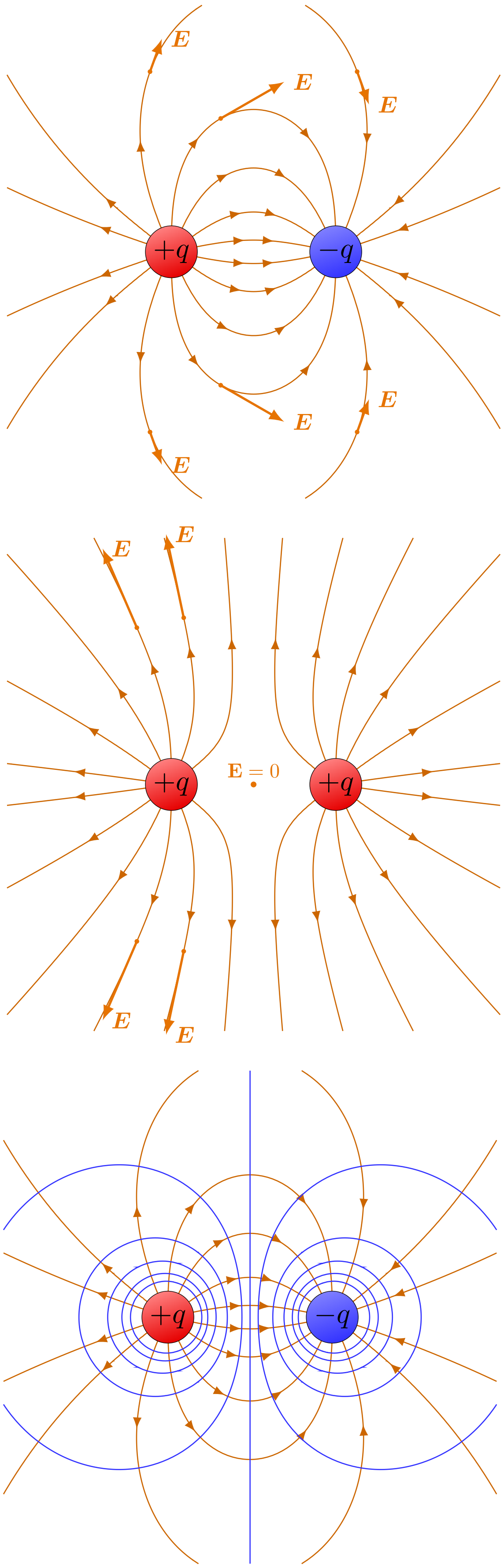

E
E
E
E
E
E
$+q$
$-q$
$+2q$
$-q$
E
E
E
E
$+q$
$\mathbf{E} = 0$
$+q$
$+2q$
$\mathbf{E} = 0$
$+q$
$+q$
$-q$
$+q$
$+q$

Figure 10.6: **Various typical electric fields configurations (author: Izaak Neutelings)**

10.1.2 Electric Potential

Given two points A and B in a region of space where there is an electric field $\vec{E}(x,y,z)$ and given Γ a path connecting these two points. So, in the particular case where the source of a field $\vec{E}$ is a sphere or a punctual body and we put a charge as its neighbourhood, we have for the work done by the force to move the charge from point A to point B:

$$\begin{aligned}\mathrm{d}W_{A,B}=\vec{F}\circ\mathrm{d}\vec{r}\Rightarrow\int\limits_A^B\vec{F}\circ\mathrm{d}\vec{r}=q\int\limits_A^B\vec{E}\circ\mathrm{d}\vec{r}=q\int\limits_A^B E\cos(\alpha)\mathrm{d}r=q\int\limits_A^B E\mathrm{d}r=q\frac{Q}{4\pi\varepsilon}\int\limits_A^B\frac{1}{r^2}\mathrm{d}r\\=q\frac{Q}{4\pi\varepsilon}\left(-\frac{1}{r_B}+\frac{1}{r_A}\right)=q\frac{Q}{4\pi\varepsilon}\left(\frac{1}{r_A}-\frac{1}{r_B}\right)\end{aligned}\tag{10.1.17}$$

Moreover, this work is as we shall see, equivalent to the potential energy. We thus define the "potential difference" or simply the "potential" by the relation:

$$-\Delta U=\frac{Q}{4\pi\varepsilon}\left(\frac{1}{r_B}-\frac{1}{r_A}\right)=U_B-U_A\tag{10.1.18}$$

and therefore:

$$\boxed{\Delta U=-\int\vec{E}\circ\mathrm{d}\vec{r}=-\int E\mathrm{d}r}\tag{10.1.19}$$

measured in Volts [V], or equivalently in $\left[\mathrm{J}\cdot\mathrm{C}^{-1}\right]$.

In other words, the "difference in electric potential" between two points (i.e., voltage) in a static electric field is defined as the work needed per unit of charge to move a test charge between the two points.

The work can then be written:

$$\Delta W_{A,B}=q\Delta U\tag{10.1.20}$$

Remarks

R1. The potential is often named "voltage" by electricians, electrical engineers and other engineers and the unit of measurement of the potential that is the "Volt" denoted by [V].

R2. The potential difference can either exists between two terminals of opposite charge $(+,-)$, or between two terminals of the type $(+,\text{neutral})$ or also $(-,\text{neutral})$. The latter two cases represents typically the configuration used by trains, trams, storms and almost all electromechanical appliances.

Theorem 10.1. We will now show in the more general framework that exists that the stationary vector field $\vec{E}$ derivates from a potential field:

Proof 10.1.1. Given a charge Q located relative to a reference frame by the vector $\vec{r}_Q$. Then in each point of space there exists a field $\vec{E}$ such as:

$$\vec{E}(\vec{r})=\underbrace{\frac{1}{4\pi\varepsilon}}_{=k_e}\frac{Q}{\|\vec{r}-\vec{r}_Q\|^3}(\vec{r}-\vec{r}_Q)\tag{10.1.21}$$

let us develop this expression:

$$\vec{E}(x,y,z)=\frac{k_eQ}{[(x-r_{Q_x})^2+(y-r_{Q_y})^2+(z-r_{Q_z})^2]^{3/2}}\begin{pmatrix}x-r_{Q_x}\\y-r_{Q_y}\\z-r_{Q_z}\end{pmatrix}\tag{10.1.22}$$

If $\vec{E}$ is a stationary potential field then, there must be a potential $\Phi(x,y,z)$ of this field which satisfies:

$$\frac{\partial\Phi}{\partial x}(x,y,z)=E_x(x,y,z).\ \frac{\partial\Phi}{\partial y}(x,y,z)=E_y(x,y,z),\ \frac{\partial\Phi}{\partial z}(x,y,z)=E_z(x,y,z)\tag{10.1.23}$$

Let us look to the potential Φ exists for a Coulomb field. Then we must have for the field in x:

$$\frac{\partial\Phi}{\partial x}=E_x=\frac{KQ}{[(x-r_{Q_x})^2+(y-r_{Q_y})^2+(z-r_{Q_z})^2]^{3/2}}(x-r_{Q_x})\tag{10.1.24}$$

therefore:

$$\Phi(x,y,z)=-k_eQ[(x-r_{Q_x})^2+(y-r_{Q_y})^2+(z-r_{Q_z})^2]^{-1/2}+c^{te}\tag{10.1.25}$$

and if we do the same development for each component, we also get the same result. Thus the electric potential is a scalar field and not a vector field (the electric field is it obviously a vector field)!

■ Q.E.D.

The potential $\Phi(x,y,z)$ is named in the case of a Coulomb field of "Coulomb potential" and is conventionally chosen such that:

$$\Phi(x,y,z)=-U\tag{10.1.26}$$

As we can see it by the prior-previous expression of $\Phi(x,y,z)$, c^{te} is an arbitrary constant, which imposes in the case of absence of charges that:

$$U=0\tag{10.1.27}$$

Which finally gives us:

$$E_x=-\frac{\partial U}{\partial x},\ E_y=-\frac{\partial U}{\partial y},\ E_z=-\frac{\partial U}{\partial z}\tag{10.1.28}$$

Giving for all components:

$$\vec{E}=-\begin{pmatrix}\dfrac{\partial}{\partial x}\\\dfrac{\partial}{\partial y}\\\dfrac{\partial}{\partial z}\end{pmatrix}\cdot U\tag{10.1.29}$$

that we write in a more condensed way:

$$\boxed{\vec{E}=-\vec{\nabla}\cdot U=-\overrightarrow{\text{grad}}(U)}\tag{10.1.30}$$

A common concept that you can find in many electromagnetic compatibility standards for humans (especially) but also for machines is the "step voltage" that is just the difference in potential measured between two point distant of 1 meter. In other words it's just the electric field value. Near some high-voltage installations, the step voltage can kill humans or other animals.

The resistance of the human body is roughly between $1,000$ and $1,000$ ohms. Then we have the following typical table:

Current	**Required Voltage**		**Comments**	
	$1,000$ $[\Omega]$	$100,000$ $[\Omega]$		
1 $[\text{mA}]$	1 $[\text{V}]$	100 $[\text{V}]$	Threshold of feeling, tingling sensation	
15 $[\text{mA}]$	5 $[\text{V}]$	500 $[\text{V}]$	Maximum current level a human can withstand before sustaining injury	
$15-20$ $[\text{mA}]$	10 $[\text{V}]$	$1,000$ $[\text{V}]$	Beginning of sustained muscular contraction ("can't let go" current)	extreme pain
$100-300$ $[\text{mA}]$	100 $[\text{V}]$	$10,000$ $[\text{V}]$	Ventricular fibrilation, fatal if continued. Respiratory function continues.	burns, permanent damage, or death
6 $[\text{A}]$	$6,000$ $[\text{V}]$	$600,000$ $[\text{V}]$	Sustained ventricular contraction followed by normal heart rhythm (defibrilation). Temporary respiratory paralysis and burns	

Figure 10.7: **Various voltage values relative to human body health**

Mobile phone 3G and 4G antenna seem to have a step voltage (depending on various countries regulations) between $11\,[\text{V}\cdot\text{m}^{-1}]$ and $58\,[\text{V}\cdot\text{m}^{-1}]$ at the day we write these lines. This is a higher value than the step voltage of a hair dryer, a toaster, a mixer, a fridge or even in comparison of Earth natural step voltage that is over land in average around[1] $120\,[\text{V}\cdot\text{m}^{-1}]$.

Definition (#10.1.477): The "electricity" is the set of physical phenomena associated with the presence and motion of matter possessing an electric charge. Electricity can be directly observed thanks to thunder for example or electric arcs.

> Remark
>
> The same developments and results (and those that follow) are applicable regarding the gravitational potential field. However, they rarely are made in the literature or schools because the human being does not control (at least until now) the gravitational field with an ease and intensity equivalent to that of the electric field...

It therefore follows that:

$$\vec{F}=q\vec{E}=-q\vec{\nabla}\cdot U \qquad (10.1.31)$$

and so for example for a spherical symmetric potential (case that we will find in many other sections of this textbook), it comes:

---

[1] According to the study Atmospheric Electricity, Grupo de Eletricidade Atmosférica (ELAT), Instituto Nacional de Pesquisas Espaciais, June 12005 (holocene calendar).

$$F=\|\vec{F}\|=\left|-q\frac{\mathrm{d}U}{\mathrm{d}R}\right| \tag{10.1.32}$$

10.1.2.1 Path Independence

Let us prove now that the potential difference between two points A and B is independent of the path Γ travelled such as we did for the gravitational potential field in the section of Classical Mechanics.

Given Γ a path between two points A and B and a field $\vec{E}$ and let us make so that we can express the field on x, y and z with respect to a single variable t (which has nothing to do with time...) that would account for its magnitude change for any movement between these two points:

$$\vec{E}\circ\mathrm{d}\vec{r}=-\left(\vec{\nabla}\cdot U(r(t))\right)\circ\mathrm{d}\vec{r} \tag{10.1.33}$$

So with A corresponding to the value t_1 of the parametrization and B to the value t_2.

But, we know that (see section Differential and Integral Calculus page 1466) that:

$$\mathrm{d}f=\frac{\partial f}{\partial x}\mathrm{d}x+\frac{\partial f}{\partial y}\mathrm{d}y+\frac{\partial f}{\partial z}\mathrm{d}z=\partial_x f\mathrm{d}x+\partial_y f\mathrm{d}y+\partial_z f\mathrm{d}z=(\vec{\nabla}\cdot f)\circ\mathrm{d}\vec{r} \tag{10.1.34}$$

Therefore it comes:

$$\vec{E}\circ\mathrm{d}\vec{r}=-\mathrm{d}U(r(t)) \tag{10.1.35}$$

Therefore:

$$\underbrace{-\int_{\Gamma}\vec{E}\circ\mathrm{d}\vec{r}}_{=\Delta U}=\int_{t_1}^{t_2}\mathrm{d}U(r(t))=U(r(t))\Big|_{t_1}^{t_2}=U(r(t_2))-U(r(t_1)) \tag{10.1.36}$$

This last expression shows that U is independent of the path Γ regardless of the way we parametrize it.

The Coulomb field is therefore a "conservative vector field" Indeed, if we consider a closed path Γ and A and B are two confused points of the part then the potential difference will be zero!

Notice that sometimes we also say that the gradient of the potential is conservative.

10.1.3 Equipotential and Field lines

Now we can from what we have built, define the "equipotentials" and "field lines".

Given a Coulomb field defined relatively to a given repository (reference frame). Then at each point (x, y, z) of space, we can associate an electric field vector $\vec{E}(x, y, z)$ and an electric potential.

Definition (#10.1.478): The "field lines" is a family of curves for which the vector $\vec{E}(x, y, z)$ is tangent and constant at each point and the "equipotentials" as the lines for which the potential $U(x, y, z)$ are also constant.

Theorem 10.2. In this case, and that is what we will show, all field lines are perpendicular to all equipotentials if the vector field derived from a potential.

Let us use the property conservation of the Coulomb field for the proof:

Proof 10.2.1.

$$\Delta U = -\oint_{\Lambda} \vec{E} \circ \mathrm{d}r = U(B) - U(A) = 0 \tag{10.1.37}$$

As we are in the presence of an electric field, this therefore derives from a potential as we know. This implies that if the field is not null the potential there is also not. So in the line integral:

$$\Delta U = -\oint_{\Lambda} \vec{E} \circ \mathrm{d}r \tag{10.1.38}$$

one of the terms is zero! This is not the electric field $\vec{E}$ as we are in presence of one, which discredits the potential U and as the charge moves, $\mathrm{d}\vec{r}$ is not null either. Then write the line integral in another way:

$$\Delta U = -\oint_{\Lambda} \vec{E} \circ \mathrm{d}r = -\oint_{\Lambda} E \circ \mathrm{d}r \cos(\alpha) = 0 \Leftrightarrow \frac{\mathrm{d}U}{\mathrm{d}r} = -E\cos(\alpha) = 0 \Leftrightarrow \frac{\left(\dfrac{\mathrm{d}U}{\mathrm{d}r}\right)}{E} = \cos(\alpha) = 0 \tag{10.1.39}$$

Therefore:

$$\alpha = \frac{\pi}{2} \tag{10.1.40}$$

We can then conclude that the equipotentials are perpendicular to the lines of the electric field and vice versa. This is what was to be proved.

■ Q.E.D.

Here are examples of level lines including field lines and equipotential lines obtained using Maple™ 4.00b (we will show in our study of differential equations how to get the mathematical functions of the field lines):

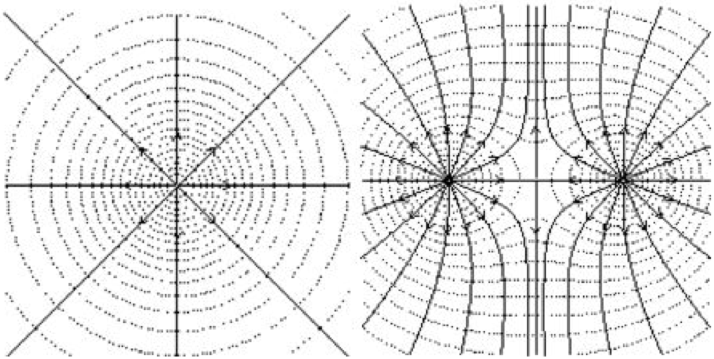


Figure 10.8: **Left: a single charge - Right: two charges of the same sign**

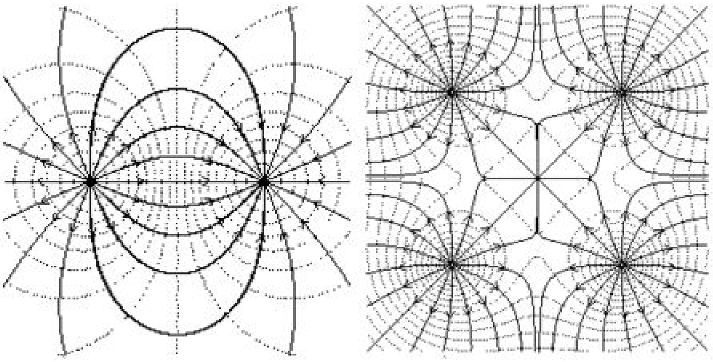


Figure 10.9: **Left: two charges of opposite signs - Right: four charges of the same sign**

> **Remark**
>
> Apart from the opposite charges, we recall that the same results are applicable to the masses with the gravitational field!

Two applications of these results are very important (for which we will limit ourselves to the study of the most important properties):

1. The determination of the field lines and equipotential lines for an infinite straight wire such as we can consider in approximation in the electric circuits or high voltage overhead lines (to determine the influence of fields from wires with their environment - this study is part of the branch of electro-engineering we name EMC for "ElectroMagnetic Compatibility"). The results can also be used to determine the "voltage step" for some rectilinear systems which determines for a given distance, the potential per meter for which a mammal may be killed by electroshock near such a wire. An extension (which I do not wish to deal even if the subject is exciting but very controversial) is also the influence of this type of potential on the functioning of the human brain in the case of the use of portable phones (antennas transmitting a potential) or about houses near high voltage power lines...

   > **Remark**
   >
   > We will determine in the section of Magnetostatics the Biot and Savart law giving the magnetic field for such a wire carrying a given current intensity.

2. The determination of the field lines and equipotential of an electric dipole has a huge importance in chemistry. We will also see what is the dynamics of it when immersed in a

uniform electric field and the energy of interaction between dipoles (as is often the case in chemistry).

10.1.3.1 Infinite straight wire

Given:

$$\vec{u}_r=\frac{\vec{r}}{\|\vec{r}\|} \tag{10.1.41}$$

We have:

$$\mathrm{d}\vec{E}=\frac{1}{4\pi\varepsilon}\frac{\mathrm{d}q}{r^2}\vec{u}_r \tag{10.1.42}$$

by making use of the concept of linear charge density as we have defined it in the section of Principles of Mechanics, we have:

$$\vec{E}=\frac{1}{4\pi\varepsilon}\int\limits_L\frac{\gamma\mathrm{d}l}{r^2}\vec{u}_r \tag{10.1.43}$$

Let us consider an infinite line (wire) of negligible section, and carrying a linear continuous density charge γ. The goal is then to calculate the electric field and potential at any point M of the space outside this line (wire) in order to know the influences of charges of this line (wire) on the environment by considering the influence of the electric field (if the charges were moving we should also take into account the influence of the magnetic field, which we will do in the section Magnetostatics).

For this, the method is to cut the line (wire) into small elements of line (wire) $\mathrm{d}l$, each element carrying a charge load $\mathrm{d}q$. The electric field generated by the charge load on P at point M located at a distance x and of orthogonal projection H on the line is:

$$\mathrm{d}\vec{E}_P(M)=\frac{1}{4\pi\varepsilon}\frac{\lambda\mathrm{d}l}{x^2}\vec{u}_x \tag{10.1.44}$$

The trick now is to take the symmetric P' of P with respect to H (the orthogonal projection of M on the wire)

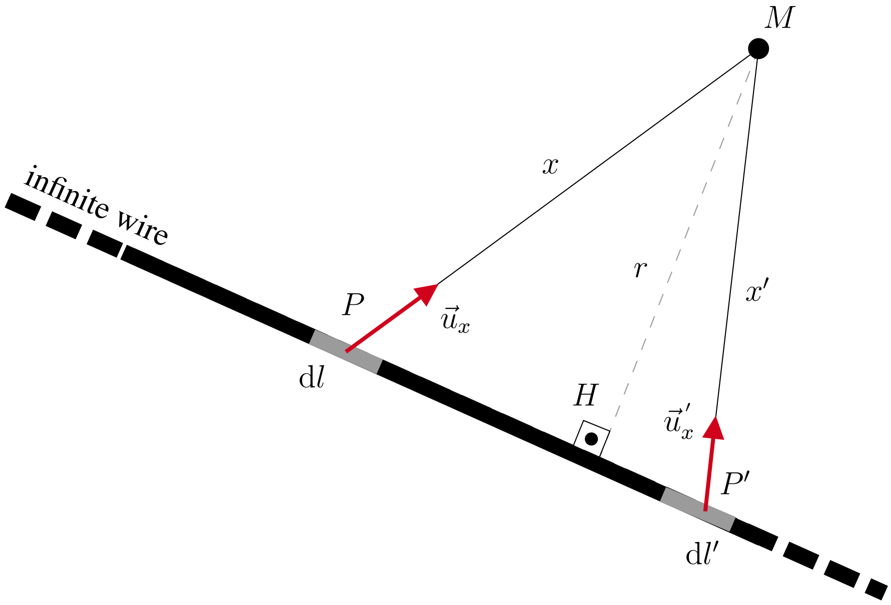


Figure 10.10: Configuration for the analysis of the electric field of an infinite straight wire

for which we have identically:

$$\mathrm{d}\vec{E}_{p'}(M)=\frac{1}{4\pi\varepsilon}\frac{\gamma \mathrm{d}l'}{x'^2}\vec{u}_x \qquad (10.1.45)$$

The total field is therefore:

$$\mathrm{d}\vec{E}(M)=\mathrm{d}\vec{E}_P(M)+\mathrm{d}\vec{E}_{P'}(M)=\frac{1}{4\pi\varepsilon}\frac{\gamma \mathrm{d}l}{x^2}\vec{u}_x+\frac{1}{4\pi\varepsilon}\frac{\gamma \mathrm{d}l'}{x'^2}\vec{u}_{x'} \qquad (10.1.46)$$

But we have:

$$x=x' \quad \mathrm{d}l=\mathrm{d}l' \quad \vec{u}_x=\frac{\overrightarrow{PM}}{\overline{\overline{PM}}} \quad \vec{u}_{x'}=\frac{\overrightarrow{P'M}}{\overline{\overline{P'M}}} \qquad (10.1.47)$$

Therefore:

$$\mathrm{d}\vec{E}(M)=\frac{1}{4\pi\varepsilon}\frac{\gamma \mathrm{d}l}{x^2}\left(\frac{\overrightarrow{PM}}{\overline{\overline{PM}}}+\frac{\overrightarrow{P'M}}{\overline{\overline{P'M}}}\right)=\frac{1}{4\pi\varepsilon}\frac{\gamma \mathrm{d}l}{x^3}\left(\overrightarrow{PM}+\overrightarrow{P'M}\right)=\frac{2}{4\pi\varepsilon}\frac{\gamma \mathrm{d}l}{x^3}\overrightarrow{HM} \qquad (10.1.48)$$

As we might expect, the latter relation shows that the field is perpendicular to the line (to the wire...). The norm of $\mathrm{d}\vec{E}(M)$ is:

$$||\mathrm{d}\vec{E}(M)||=\frac{2}{4\pi\varepsilon}\frac{\gamma \mathrm{d}l}{x^3}\|\overrightarrow{HM}\|=\frac{2}{4\pi\varepsilon}\frac{\gamma \mathrm{d}l}{x^3}r \qquad (10.1.49)$$

This relation has three dependent variables r, $\mathrm{d}l$, x. The norm of the total field on a point is equal to the sum of the norms of all the vectors $\mathrm{d}\vec{E}(M)$ of the entire length of the wire since all vectors have the same direction.

For this calculation, we will make a change of variable, and put r, $\mathrm{d}l$, x according to the angle α between the line (wire) and the vector $\overrightarrow{PM}$. In the rectangle triangle HMP:

$$\cos(\alpha)=\frac{\overline{HP}}{\overline{\overline{PM}}}=\frac{z}{x} \qquad (10.1.50)$$

if we take the origin of z in H. We also have:

$$\sin(\alpha)=\frac{\overline{HM}}{\overline{PM}}=\frac{r}{x}\qquad \tan(\alpha)=\frac{\overline{HM}}{\overline{HP}}=\frac{r}{z}\tag{10.1.51}$$

and:

$$\mathrm{d}l=\mathrm{d}z=-r\frac{(1+\tan^2(\alpha))}{\tan^2(\alpha)}\mathrm{d}\alpha=-r\frac{\left(1+\dfrac{\sin^2(\alpha)}{\cos^2(\alpha)}\right)}{\dfrac{\sin^2(\alpha)}{\cos^2(\alpha)}}\mathrm{d}\alpha=-r\frac{sin^2(\alpha)+\cos^2(\alpha)}{\sin^2(\alpha)}\mathrm{d}\alpha=\frac{-r}{\sin^2(\alpha)}\mathrm{d}\alpha\tag{10.1.52}$$

therefore:

$$\mathrm{d}E=-\frac{\gamma}{2\pi\varepsilon}\frac{r}{\sin^2(\alpha)}\mathrm{d}\alpha\frac{\sin^3(\alpha)}{r^3}r=-\frac{\gamma}{2\pi\varepsilon r}\sin(\alpha)\mathrm{d}\alpha\tag{10.1.53}$$

$$E=\frac{\gamma}{2\pi\varepsilon r}\cos(\alpha)|_0^{\pi/2}\tag{10.1.54}$$

and therefore:

$$\vec{E}=-\frac{\gamma}{2\pi\varepsilon r}\vec{u}_r\tag{10.1.55}$$

The potential is easily deduced by taking the primitive of E since:

$$\vec{E}=-\begin{pmatrix}\dfrac{\partial}{\partial x}\\ \dfrac{\partial}{\partial y}\\ \dfrac{\partial}{\partial z}\end{pmatrix}\cdot U\tag{10.1.56}$$

Then we have:

$$U=\frac{\gamma}{2\pi\varepsilon}\ln(r)+c^{te}\tag{10.1.57}$$

The constant is undetermined since as r approaches infinity, U tends therefore to zero and leads to an infinite constant. This uncertainty is mainly due to the approximation of the infinite wire length.

10.1.3.2 Electric Rigid Dipole

An important and interesting disposition of electric charges is that forming an "electric dipole" rigorously named "rigid electric dipole" or "electrostatic dipole". It consists of two equal and opposite electric charges $+q,-q$ separated by a very small distance.

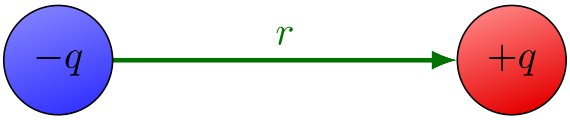


Figure 10.11: **Naive illustration of an electric dipole (author: Izaak Neutelings)**

We will seek to determine the potential and the electric field at a point M of the dipole environment.

To determine this let us consider an electric charge q_i on a point A_i and a very distant point M from A_i. Let us take a Galilean reference frame in O:

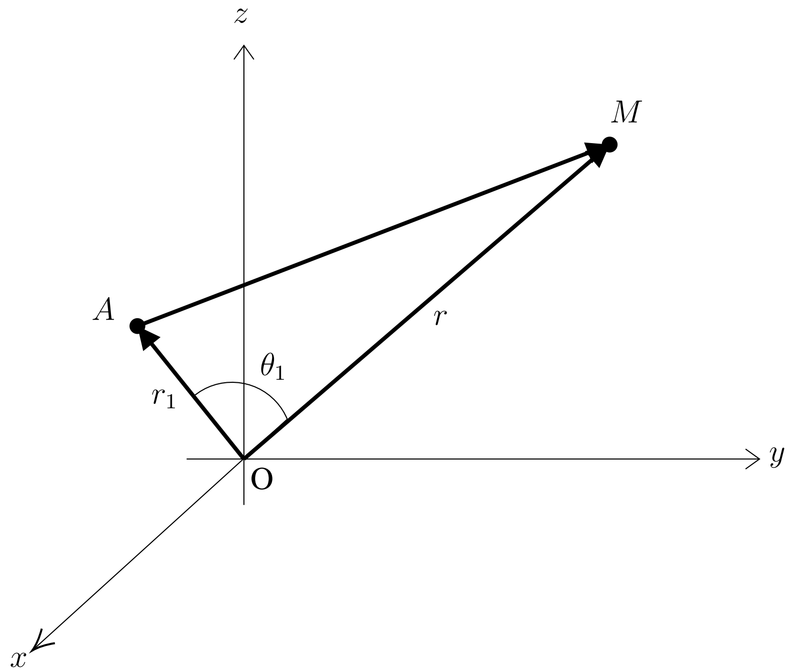


Figure 10.12: **Electric field at a distant point M of the dipole**

The electric potential generated at the M by the charge q_i is:

$$U_i(M)=\frac{1}{4\pi\varepsilon}\frac{q_i}{\overline{A_iM}} \quad (10.1.58)$$

In the triangle OA_iM, the distance $\overline{A_iM}$ can be written with the cosine theorem:

$$\overline{A_iM}=\sqrt{r_i^2+r^2-2r_ir\cos(\theta_i)} \quad (10.1.59)$$

The potential becomes:

$$U_i(M)=\frac{1}{4\pi\varepsilon}\frac{q_i}{\sqrt{r_i^2+r^2-2r_ir\cos(\theta_i)}} \quad (10.1.60)$$

At great distance, r becomes much higher than r_i, the quantity:

$$u=\frac{r_i^2}{r^2}-\frac{2r_i\cos(\theta_i)}{r} \quad (10.1.61)$$

tends to zero. So we can make a Maclaurin development (see section Sequences and Series page 1703) of $(1+u)^{-1/2}$ at the neighbourhood of $u=0$. To avoid making heavy calculations, we will limited ourselves to the order two on r:

$$(1+u)^{-1/2}=1-\frac{1}{2}u-\frac{1}{2}\left(-\frac{3}{2}u^2\right)\frac{1}{2!}=1-\frac{1}{2}u+\frac{3}{8}u^2 \quad (10.1.62)$$

therefore:

$$I=\left(1+\frac{r_i^2}{r^2}-\frac{2r_i\cos(\theta_i)}{r}\right)^{-1/2}\cong 1-\frac{1}{2}\left(\frac{r_i^2}{r^2}-\frac{2r_i\cos(\theta_i)}{r}\right)+\frac{3}{8}\left(\frac{r_i^2}{r^2}-\frac{2r_i\cos(\theta_i)}{r}\right)^2 \tag{10.1.63}$$

Keeping only the terms of the second order in r:

$$I=1-\frac{1}{2}\frac{r_i^2}{r^2}+\frac{r_i\cos(\theta_i)}{r}+\frac{3}{2}\frac{r_i^2\cos^2(\theta_i)}{r^2} \tag{10.1.64}$$

The potential becomes:

$$\begin{aligned}U_i(M)&=\frac{1}{4\pi\varepsilon}\frac{q_i}{r}\left(1-\frac{1}{2}\frac{r_i^2}{r^2}+\frac{r_i\cos(\theta_i)}{r}+\frac{3}{2}\frac{r_i^2\cos^2(\theta_i)}{r^2}\right)\\&=\frac{1}{4\pi\varepsilon}\frac{q_i}{r}+\frac{1}{4\pi\varepsilon}\frac{q_ir_i\cos(\theta_i)}{r^2}+\frac{1}{4\pi\varepsilon}\frac{q_ir_i^2}{r^3}\left(-\frac{1}{2}+\frac{3}{2}\cos^2(\theta_i)\right)=U_i^0+U_i^1+U_i^2\end{aligned} \tag{10.1.65}$$

We kept in the expression of the potential three terms. The term U_i^0 is the potential generated by a charge that would be in O. In other words, at the zero order, the potential generated by a charge located at a point near O is identical to the potential generated by a charge that would be in O. The terms U_i^1,U_i^2 are correction terms, at the first order and second order respectively. We notice that these terms vary on $1/r^2$ and $1/r^3$, thus decreasing faster than the first. These two terms are thus more effective at smaller distances. We see that the terms U_i^1,U_i^2 involve the quantity q_ir_i. This quantity is what we define as the "dipole moment" of the electric rigid dipole:

$$\boxed{\vec{p}_i=q_i\vec{r}_i} \tag{10.1.66}$$

Remark

The dipole moment is expressed in Coulomb meter, but for convenience (...) it is expressed in Debye [D] by some engineers.

The potential generated at great distance by a discrete charge distribution is obtained by summing all individual contributions:

$$U(M)=\sum_i U_i(M)=\frac{1}{4\pi\varepsilon}\frac{1}{r}\sum_i q_i+\frac{1}{4\pi\varepsilon}\frac{1}{r^2}\sum_i q_ir_i\cos(\theta_i)+\frac{1}{4\pi\varepsilon}\sum_i\frac{q_ir_i^2}{r^3}\left(-\frac{1}{2}+\frac{3}{2}\cos^2(\theta_i)\right) \tag{10.1.67}$$

Which can also be written:

$$U(M)=U_0(M)+U_1(M)+U_2(M) \tag{10.1.68}$$

By definition, $U_0(M)$ is unipolar or monopolar term, $U_1(M)$ is the dipolar term, $U_2(M)$ is the quadrupolar term. If the electric charge distribution is a globally zero, as is the case of an ideal atom or of an ideal non-ionized molecule, it only remain the multipolar contributions.

Let us return to the particular case of the dipole:

The term monopolar term is zero, since the sum of the electric charges is zero. If we neglect the terms higher than the first order, there remain the dipole contribution.

The angles θ_1 and θ_2 of the dipole are complementary, therefore $\cos(\theta_1) = -\cos(\theta_2)$. But as $q_1 = -q_2 = q$, the product $q_i\cos(\theta_i)$ is constant.

If the two charges of the dipole are at a constant distance from each other and equidistant from the origin O. We will put that $r_1 = r_2 = d$.

The potential is therefore reduce to:

$$U(M) = \frac{1}{4\pi\varepsilon}\frac{2qd\cos(\theta)}{r^2} = \frac{1}{4\pi\varepsilon}\frac{qa\cos(\theta)}{r^2} \tag{10.1.69}$$

where a is simply the constant distance between the two electric charges. It is customary in the case of the study of the electric dipole to write the above relation as:

$$U(M) = \frac{1}{4\pi\varepsilon}\frac{\vec{p}\circ\vec{u}_r}{r^2} \tag{10.1.70}$$

where $\vec{p}$ is the definition of the dipolar moment and:

$$\vec{u}_r = \frac{\overrightarrow{\text{O}M}}{\left\|\overrightarrow{\text{O}M}\right\|} \tag{10.1.71}$$

Let us recall now that we proved at the beginning of this section that:

$$\vec{E} = -\vec{\nabla}\cdot U \tag{10.1.72}$$

and as we saw in section of Vector Calculus, the gradient in spherical coordinates leads us to write:

$$\vec{E} = -\vec{\nabla}\cdot U = -\begin{pmatrix} \dfrac{\partial U}{\partial r} \\ \dfrac{1}{r}\dfrac{\partial U}{\partial\theta}\dfrac{1}{r\sin(\theta)}\dfrac{\partial U}{\partial\phi} \end{pmatrix} \tag{10.1.73}$$

therefore:

$$\boxed{E_r = \frac{p\cos(\theta)}{2\pi\varepsilon r^3} \quad E_\theta = \frac{p\sin(\theta)}{4\pi\varepsilon r^3} \quad E_\phi = 0} \tag{10.1.74}$$

To determine the equation of the equipotential, remember that these lines (or "surfaces" when in space) are obtained through by the constraint:

$$U = c^{te} := U_0 \tag{10.1.75}$$

Therefore:

$$U = U_0 \Leftrightarrow \frac{p\cos(\theta)}{4\pi\varepsilon r^2} = U_0 \Leftrightarrow r = \sqrt{\frac{p\cos(\theta)}{4\pi\varepsilon U_0}} = r_0\sqrt{\cos(\theta)} \tag{10.1.76}$$

with:

$$r_0 = \sqrt{\frac{p}{4\pi\varepsilon U_0}} \tag{10.1.77}$$

The electric field must by definition be tangential to the field lines, thus parallel to the elementary displacement:

$$\vec{E}\times \mathrm{d}\vec{r}=\vec{0}\Leftrightarrow \begin{pmatrix}E_r\\E_\theta\\E_\phi\end{pmatrix}\times\begin{pmatrix}\mathrm{d}r\\r\mathrm{d}\theta\\r\sin(\theta)\mathrm{d}\phi\end{pmatrix}=\begin{pmatrix}E_r\\E_\theta\\E_\phi\end{pmatrix}\times\begin{pmatrix}\mathrm{d}r\\r\mathrm{d}\theta\\r\sin(\theta)\mathrm{d}\phi\end{pmatrix}=\begin{pmatrix}E_\theta r\sin(\theta)\mathrm{d}\phi-E_\phi r\mathrm{d}\theta\\E_\phi\mathrm{d}r-E_r r\sin(\theta)\mathrm{d}\phi\\E_r r\mathrm{d}\theta-E_\theta\mathrm{d}r\end{pmatrix}=\vec{0} \quad (10.1.78)$$

Since $E_\phi=0$, we have:

$$\vec{E}\times \mathrm{d}\vec{r}=\vec{0}\Leftrightarrow \begin{pmatrix}E_r\\E_\theta\\E_\phi\end{pmatrix}\times\begin{pmatrix}\mathrm{d}r\\r\mathrm{d}\theta\\r\sin(\theta)\mathrm{d}\phi\end{pmatrix}=\begin{pmatrix}E_r\\E_\theta\\E_\phi\end{pmatrix}\times\begin{pmatrix}\mathrm{d}r\\r\mathrm{d}\theta\\r\sin(\theta)\mathrm{d}\phi\end{pmatrix}=\begin{pmatrix}E_\theta r\sin(\theta)\mathrm{d}\phi\\-E_r r\sin(\theta)\mathrm{d}\phi\\E_r r\mathrm{d}\theta-E_\theta\mathrm{d}r\end{pmatrix}=\vec{0} \quad (10.1.79)$$

So finally there remains only:

$$E_r r\mathrm{d}\theta-E_\theta\mathrm{d}r=0\Leftrightarrow E_r r\mathrm{d}\theta=E_\theta\mathrm{d}r \quad (10.1.80)$$

Which is a differential equation that can be easily integrate:

$$\frac{\mathrm{d}r}{r}=\frac{E_r}{E_\theta}\mathrm{d}\theta=2\frac{\cos(\theta)\mathrm{d}\theta}{\sin(\theta)}=2\frac{\mathrm{d}\theta}{\sin(\theta)}\Leftrightarrow\ln(r)=\ln(\sin^2(\theta))+c^{te} \quad (10.1.81)$$

Which is equivalent to write:

$$r=c^{te}\sin^2(\theta) \quad (10.1.82)$$

The plot of the field and equipotential lines then gives in spherical coordinates (remember that the vertical component is zero by symmetry):

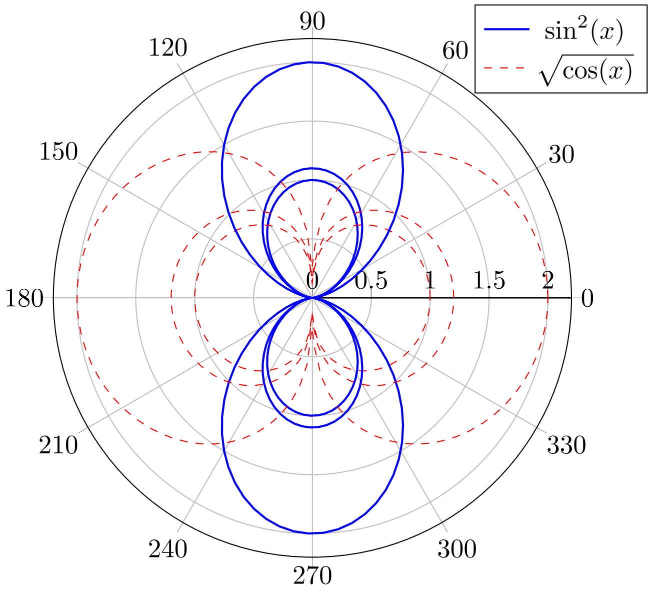


Figure 10.13: **Polar plot of the field lines and equipotentials of an electric dipole**

Even if in an electric dipole the two charges are equal and opposite, giving a zero net charge, the fact they are slightly displaced is sufficient to produce a non identically zero electric field.

In atoms, the center of mass of the electrons coincides with the core, and therefore the average electric dipole moment of the atom is zero. But if an external field is applied, the movement of electrons is distorted and the electron mass center is displaced by a given distance from the core. The atom is then polarised and becomes an electric dipole of moment $\vec{p}$. This moment is proportional to the external applied field $\vec{E}$.

As we did not found in Maple™ 17.00 how to do the polar plot above here is another Maple™ 17.00 code to have fun with a dipole:

```
>V:=1/sqrt((x-1)^2+y^2+z^2)-2/sqrt((x+1)^2+y^2+z^2):
>with(LinearAlgebra); with(VectorCalculus); with(plots);
>with(plottools);
>Efield := Gradient(-V, [x, y, z]);
>fieldplot3d([Efield[1], Efield[2], Efield[3]],x =-1.5..1.5, y
=-1.5..1.5, z =-1.5..1.5);
>NormEfield := Normalize(Efield,2);
>p1:=sphere([1, 0, 0],0.75,color=red);
>p2:=sphere([-1, 0, 0],1.5,color=blue);
>p3:=fieldplot3d(NormEfield,x=-4.5..4.5,y=-4.5..4.5,z
=-4.5..4.5,color=black);
>display([p1, p2, p3], scaling = constrained);
```

Giving at first (obviously the dimensions are fictitious):

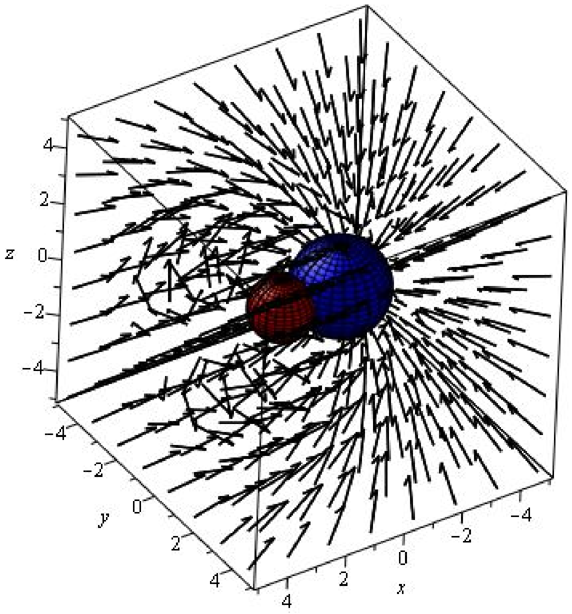


Figure 10.14: **Electric field lines for two charges of opposite signs in Maple™ 17.00**

Or to get the view of the potential with the equipotentials:

```
>z:= 0;
>plot3d(V,x =-1.5..1.5, y=-1.5..1.5, style=patchcontour,contours=200);
```

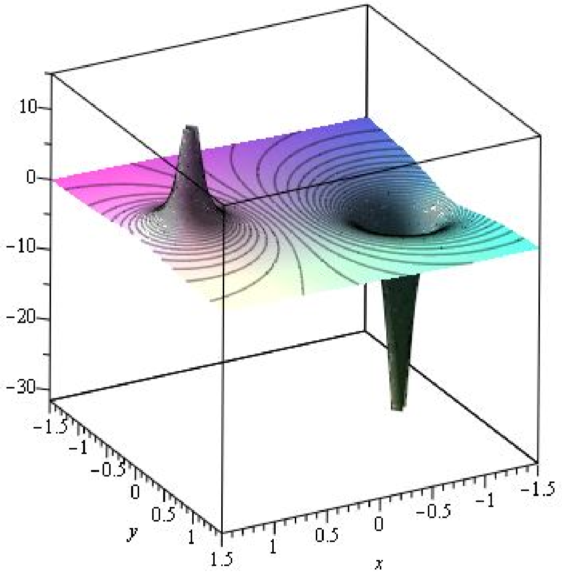


Figure 10.15: **Potential well and equipotentials of the dipole in Maple™ 17.00**

Or to get a 2D profile of the electric field and equipotentials:

```
>p4:=implicitplot(seq(V=(1/10)*b,b=-10..10),x=-5..5,y=-4..4);
>p5:=fieldplot([NormEfield[1],NormEfield[2]],x=-5 ..  5,y=-4..4);
>display([p4, p5], scaling = constrained);
```

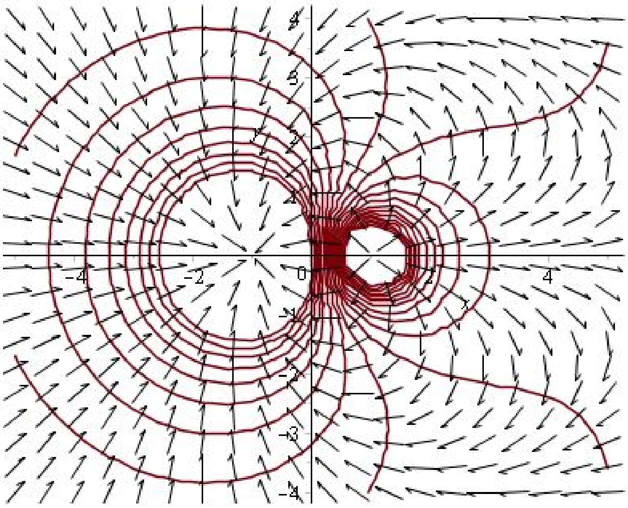


Figure 10.16: **2D profile of electric field and potential of the dipole in Maple™ 17.00**

Remark

Molecules also may have a permanent electric moment. Such molecules are named "polar molecules" For example, in the HCl molecule the electron of the hydrogen atom spends more time to move around the chlorine that around the hydrogen atom. Also, the center of negative charges does he not coincide with that of the positive charge of the center and the molecule has a dipole moment. By cons, in the molecule CO_2, all the atoms are aligned, and the resulting electric dipole moment is zero for reasons of symmetry.

When an electric dipole is positioned in an electric field, a force is exerted on each of the charges of the dipole. The resulting force is:

$$\vec{F}=q\vec{E}_{+}-q\vec{E}_{-}=q(\vec{E}_{+}-\vec{E}_{-}) \tag{10.1.83}$$

Let us consider the particular case where the electric field is directed along the x-axis and where the dipole is oriented parallel to this field.

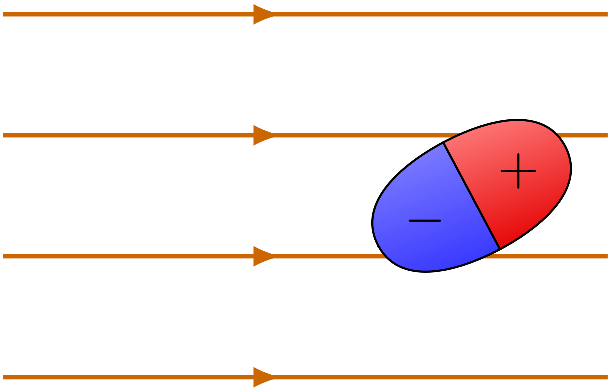


Figure 10.17: **Naive illustration of an electric dipole in a uniform electric field (author: Izaak Neutelings)**

If we only consider the quantities:

$$\vec{E}_{+}-\vec{E}_{-}=\left(\frac{\mathrm{d}E}{\mathrm{d}x}\right)a \tag{10.1.84}$$

with a being the distance between the two charges, and therefore: dipole is oriented parallel to this field. If we only consider the quantities:

$$F=q\frac{\mathrm{d}E}{\mathrm{d}x}a=p\frac{\mathrm{d}E}{\mathrm{d}x} \tag{10.1.85}$$

This result show that an electric dipole oriented parallel to the field tends to move in the direction in which the field increases (as the gradient thereof). We note that if the electric field is uniform, the resultant force on the dipole is zero!!!

The potential energy of the dipole is:

$$E_P=qU_{+}-qU_{-}=q(U_{+}-U_{-})=-qa\left(-\frac{U_{+}-U_{-}}{a}\right) \tag{10.1.86}$$

If we use the equation:

$$\vec{E}=-\vec{\nabla}\cdot U \tag{10.1.87}$$

to describe the uniform electric field and if θ is the angle between the dipole and the electric field, the last factor $(U_{+}-U_{-})/a$ is just the component $E_a=E\cos(\theta)$ of the field $\vec{E}$ parallel to a. So:

$$E_P=-qaE_a \tag{10.1.88}$$

or:

$$E_P = -pE\cos(\theta) = -\vec{p} \circ \vec{E} \tag{10.1.89}$$

The potential energy is minimum for $\theta = 0$, which shows that the dipole is in equilibrium when it is oriented parallel to the field.

These configurations of a dipole placed in an electric field have very important applications. For example, the electric field of an ion in a solution polarized the solvent molecules that surround the ions and they are oriented as in the figure below:

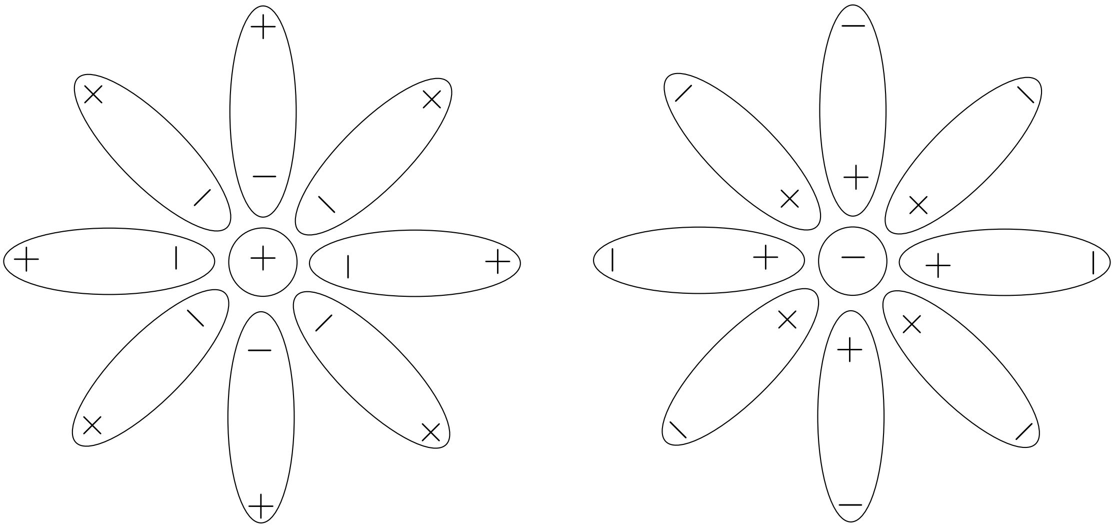


Figure 10.18: Example of what happens in a solution with an ion

In a solvent with polar molecules such as water, ions of an electrolyte in solution are surrounded by a number of these molecules due to the dipole-charge interaction. This phenomenon is named "solvation" of the ion, specifically "hydration" if the solvent is water.

For example with Salz (NaCl):

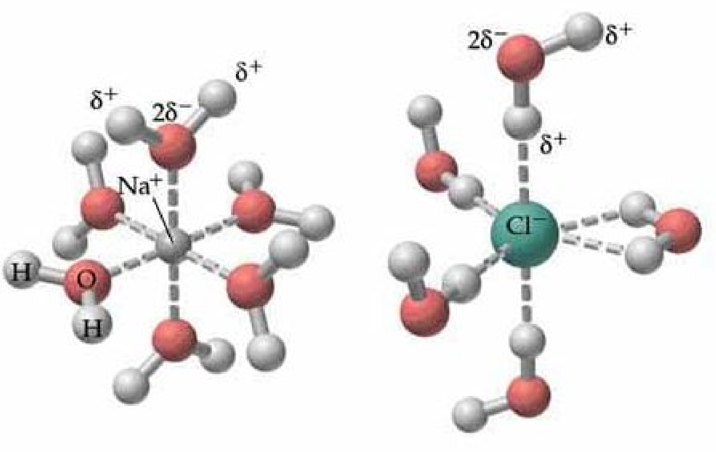


Figure 10.19: Hydration of salz (source: ?)

These oriented molecules become more or less dependant of the ion, increasing its effective mass and decreasing its effective charge, which is partially obscured by the molecules (screening). The net effect is that the mobility of a ion in an external field is reduced. Similarly, when a gas or liquid, which molecules are permanent dipoles is placed in an electric field, the molecules as a result of the force couple applied due to the electric field, tend to align with their dipoles parallel. We then say that the substance was "polarized".

It may therefore be interesting to determine the vector electric field produced by a dipole rather than its potential. The electrostatic field generated at a point M by the doublet is obtained by performing the vector sum of the fields generated in this point by positive P and negative N charges, hence:

$$\vec{E}=\vec{E}_P+\vec{E}_N=\frac{q}{4\pi\varepsilon}\left(\frac{\overrightarrow{PM}}{\left\|\overrightarrow{PM}\right\|^3}+\frac{\overrightarrow{NM}}{\left\|\overrightarrow{NM}\right\|^3}\right)=\frac{q}{4\pi\varepsilon}\left(\frac{\vec{r}-a\dfrac{\vec{e}_z}{2}}{\left\|\vec{r}-a\dfrac{\vec{e}_z}{2}\right\|^2}-\frac{\vec{r}+a\dfrac{\vec{e}_z}{2}}{\left\|\vec{r}+a\dfrac{\vec{e}_z}{2}\right\|^2}\right)$$
$$=\frac{q}{4\pi\varepsilon}\left(\left(\vec{r}-\frac{a}{2}\vec{e}_z\right)\left\|\vec{r}-a\frac{\vec{e}_z}{2}\right\|^{-3}-\left(\vec{r}+\frac{a}{2}\vec{e}_z\right)\left\|\vec{r}+a\frac{\vec{e}_z}{2}\right\|^{-3}\right) \tag{10.1.90}$$

The distribution of charges being invariant by rotation about the z axis of the doublet, the topography is independent of the azimuthal angle π of the spherical coordinates. We can then represent it in any meridian plane passing through the axis NP. The field $\vec{E}$ is the given by:

$$\vec{E}=\frac{q}{4\pi\varepsilon}\left(\left(\vec{r}-\frac{a}{2}\vec{e}_z\right)\left\|\vec{r}-a\frac{\vec{e}_z}{2}\right\|^{-3}-\left(\vec{r}+\frac{a}{2}\vec{e}_z\right)\left\|\vec{r}+a\frac{\vec{e}_z}{2}\right\|^{-3}\right) \tag{10.1.91}$$

Having:

$$\left\|\vec{r}+a\frac{\vec{e}_z}{2}\right\|^3=\left(\left\|\vec{r}+a\frac{\vec{e}_z}{2}\right\|^2\right)^{3/2} \tag{10.1.92}$$

vectorially, we get:

$$\vec{r}+\frac{a}{2}\vec{e}_z=\begin{pmatrix} r_x \\ r_y \\ r_z+\dfrac{a}{2}\end{pmatrix}\Rightarrow\left\|\vec{r}+a\frac{\vec{e}_z}{2}\right\|^2=\underbrace{r_x^2+r_y^2+r_z^2}_{r^2}+2r_z\frac{a}{2}+\frac{a^2}{4}=r^2+ar_z+\frac{a^2}{4} \tag{10.1.93}$$

The dot product being the multiplication of components one by one, we have:

$$\vec{r}\circ\vec{e}_z=r_z \tag{10.1.94}$$

Hence:

$$\left\|\vec{r}+a\frac{\vec{e}_z}{2}\right\|^2=r^2+a\vec{r}\circ\vec{e}_z+\frac{a^2}{4} \tag{10.1.95}$$

Finally:

$$\left(\left\|\vec{r}+a\frac{\vec{e}_z}{2}\right\|^2\right)^{3/2}=\left(r^2+a\vec{r}\circ\vec{e}_z+\frac{a^2}{4}\right)^{3/2}=r^3\left(1+a\frac{\vec{r}\circ\vec{e}_z}{r^2}+\frac{a^2}{4r^2}\right)^{3/2} \tag{10.1.96}$$

So by a limited Maclaurin series development as we did at the beginning:

$$\begin{aligned}\vec{E}&=\frac{q}{4\pi\varepsilon r^3}\left(\left(\vec{r}-\frac{a}{2}\vec{e}_z\right)\left[1-a\frac{\vec{r}\circ\vec{e}_z}{r^2}+\frac{a^2}{4r^2}\right]^{-3/2}-\left(\vec{r}+\frac{a}{2}\vec{e}_z\right)\left[1+a\frac{\vec{r}\circ\vec{e}_z}{r^2}+\frac{a^2}{4r^2}\right]^{-3/2}\right)\\&\cong\frac{q}{4\pi\varepsilon r^3}\left(\left(\vec{r}-\frac{a}{2}\vec{e}_z\right)\left[1+\frac{3}{2}a\frac{\vec{r}\circ\vec{e}_z}{r^2}+\ldots\right]-\left(\vec{r}+\frac{a}{2}\vec{e}_z\right)[]\left[1-\frac{3}{2}a\frac{\vec{r}\circ\vec{e}_z}{r^2}+\ldots\right]\right)\\&=-\frac{q}{4\pi\varepsilon r^3}a\vec{e}_z+\frac{3q\vec{r}}{4\pi\varepsilon r^5}a\vec{r}\circ\vec{e}_z+\ldots\cong\frac{1}{4\pi\varepsilon}\left[\frac{3(\vec{p}\circ\vec{r})\cdot\vec{r}}{r^5}-\frac{\vec{p}}{r^3}\right]\end{aligned} \tag{10.1.97}$$

By introducing:

$$\vec{u}_r=\frac{\vec{r}}{\|\vec{r}\|}=\frac{\vec{r}}{r} \tag{10.1.98}$$

We can simplify the notations:

$$\vec{E}\cong\frac{3(\vec{p}\circ\vec{u}_r)\vec{u}_r-\vec{p}}{4\pi\varepsilon r^3} \tag{10.1.99}$$

It may also be relevant to calculate the energy of interaction between two electric dipoles. If we denote by $\vec{p}_1$ the dipole moment, we can write:

$$\vec{E}_1\cong\frac{3(\vec{p}_1\circ\vec{u}_r)\vec{u}_r-\vec{p}_1}{4\pi\varepsilon r^3} \tag{10.1.100}$$

If we denote by $\vec{p}_2$ the moment of the second dipole and if we use the relation:

$$E_p=-\vec{p}\circ\vec{E} \tag{10.1.101}$$

we find that the interaction energy between two dipoles is:

$$E_{P,1,2}=-\vec{p}_2\circ\vec{E}_1-\frac{3((\vec{p}_1\circ\vec{u}_r)\vec{u}_r)\circ cp_2-\vec{p}_1\circ\vec{p}_2}{4\pi\varepsilon r^3} \tag{10.1.102}$$

We can draw out several important conclusions from this result. The energy of interaction $E_{P,1,2}$ is symmetric relatively to the two dipoles, because the permutation of $\vec{p}_1$ and of $\vec{p}_2$ leaves it unchanged. This is an expected result. The interaction between two dipoles is not central as it depends of the angles which the vector position or the unitary vector $\vec{u}_r$ do with $\vec{p}_1$ and $\vec{p}_2$.

An atom, molecule or a ion, whose dipole moment is zero in the fundamental state, acquire a dipole moment under the action of the applied non-uniform electric field as we have seen since the opposing signs charges are solicited in opposed directions. The centers of gravity of positive and negative charges do not coincide anymore, it appears an "induced dipole moment".

In an experiment with linear approximation for weak excitatory electric fields, the induced dipole moment is proportional to the applied field $\vec{E}$, which we translate by (it is in fact an approximation of the Langevin-Debye relation we will prove later):

$$\vec{p}=\alpha\varepsilon\vec{E} \tag{10.1.103}$$

The quantity α, which physical dimension is that of a volume is the "polarizability" of the structure. The dipole-dipole electrostatic interaction was introduced by J.D. Van der Waals in 1873, in the case of molecules, to interpret real deviations from the ideal gas.

The Van der Waals forces are repulsive as the distance between molecules is very low because they oppose the interpenetration of electron clouds, what we express by introducing their volume (covolume).

By cons, they are attractive when the distance is sufficient. We attribute this attraction to three types of interactions involving rigid or induced dipoles:

1. The forces between polar molecules (rigid dipoles), say to be "Keesom forces".
2. The forces between a polar molecule (rigid dipole) and a polarizable molecule (induced dipole) named "Debye forces".
3. The average forces between induced dipoles which appear even when the molecules are not polar, named "London forces".

In all these three cases, the electrostatic energy is negative (attraction) and varies as r^{-6}. To prove this statement, let us calculate the interaction energy between two rigid dipoles, of dipolar moments $\vec{p}_1$ and $\vec{p}_2$:

$$E_{P,2,1}=-\vec{p}_1\circ\vec{E}_2 \tag{10.1.104}$$

with:

$$\vec{E}_2\cong\frac{3(\vec{p}_2\circ\vec{u}_r)\vec{u}_r-\vec{p}_2}{4\pi\varepsilon r^3} \tag{10.1.105}$$

and:

$$\vec{p}_1=\alpha\varepsilon\vec{E}_2 \tag{10.1.106}$$

Therefore:

$$\vec{E}_2=-\alpha\varepsilon\vec{E}_2\circ\vec{E}_2=-\alpha\varepsilon E_2^2\cong-\alpha\varepsilon\left(\frac{2p_2}{4\pi\varepsilon r^3}\right)^2=-\frac{c^{te}}{r^6}=-\frac{C_6}{r^6} \tag{10.1.107}$$

This is the van der Waals interaction potential between two dipoles atoms where C_6 is the "Van der Waals constant".

Hence:

$$\vec{F}=-\vec{\nabla}\cdot E_P=\frac{\overrightarrow{C^{te}}}{r^7} \tag{10.1.108}$$

Thus, the radial dependence of the force is indeed in r^{-7}. This very rapid decay of the Van der Waals force with distance helps explain its very short range and therefore its influence when the medium is sufficiently dense. However the decay is slower than the chemical bond that decrease exponentially (see section Quantum Chemistry page 6190).

Remark

The interaction between polar molecules, of the Keesom type, is made very high in the presence of hydrogen, because the latter, thanks to its small size, also interacts with the atoms of other molecules. It is that one which is at the origin of the "hydrogen bonding".

10.1.4 Electric Flux

Given $\vec{V}$ a vector field and S a surface named "Gauss surface" in space. If we divide this area into a number N of smaller surfaces $\mathrm{d}S$ each traversed by a field $\vec{V}_i$ and having a unit perpendicular vector $\vec{n}_i$ (special case) on their surface, then we can form the sum of:

$$\sum_{i=1}^{N} \vec{V}_i \circ \vec{n}_i \mathrm{d}S \tag{10.1.109}$$

When N tends to infinity and all $\mathrm{d}S$ to zero, we get for this sum:

$$\Phi_V = \int_{\Sigma} \vec{V}_i \circ \mathrm{d}\vec{S} \tag{10.1.110}$$

The value of this integral thus gives the flow Φ_V of the field $\vec{V}$ through the surface S delimited by a domain Λ and where:

$$\mathrm{d}\vec{S} = \vec{n}\mathrm{d}S \tag{10.1.111}$$

In the case of the electrostatic field, we write:

$$\boxed{\Phi_E = \int_{\Sigma} \vec{E} \circ \mathrm{d}\vec{S}} \tag{10.1.112}$$

This expression, define the "electric flow" or also named "electric flux".

The inevitable question that arises is then: what is its physical meaning? The flux of a fluid is the amount of fluid (especially the quantity of volume) passing through a surface by second. Then there is a flux (flow) of something. About the electrical flux (flow), from the classical point of view, nothing flows, the electric field is already established and is static, but is flow through the surface. The value of the electric field at any point in space is the field intensity at that point, while the flow can be considered the quantity of field which passes through the surface S.

There is a hundred years, physicists identified the flow with the number of electric field lines passing through the surface. But the least we can say is that the simplistic view that the field lines have a distinct reality and that we can count is misleading. We will see during our study of Quantum Field Theory, in the section of the same name, that the latter supports a virtual photons current that is the nature of electromagnetic interactions. Despite this, the physicists did not hurry to associate the flow of virtual photons of the 120th century (holocene calendar) to the continuous electric field lines of the 119th century. Whatever its nature, the concept of flow is powerful and of great practical use, both in electricity and magnetism.

As will we prove it in the framework of Maxwell's equations (see section Electrodynamics page 4095), solving this integral gives in the general case (this is the "Gauss law for electrostatic" also named "Gauss theorem for electrostatic"):

$$\boxed{\Phi_E=\oint\limits_{\Sigma}\vec{E}\circ\mathrm{d}\vec{S}=\frac{Q}{\varepsilon}=\frac{\sum_i q_i}{\varepsilon}} \tag{10.1.113}$$

In the case where the surface is not closed and reduce to a plane, the previous close surface integral is reduced to:

$$\Phi_E=\vec{E}\circ\vec{S}=ES\cos(\alpha) \tag{10.1.114}$$

10.1.4.1 Capacitor

A capacitor (originally known as a condenser) is a passive two-terminal electrical component used to store electrical energy temporarily in an electric field. The forms of practical capacitors vary widely, but all contain at least two electrical conductors (plates) separated by a dielectric (i.e. an insulator that can store energy by becoming polarized). The conductors can be thin films, foils or sintered beads of metal or conductive electrolyte, etc. The non-conducting dielectric acts to increase the capacitor's charge capacity. Materials commonly used as dielectrics include glass, ceramic, plastic film, air, vacuum, paper, mica, and oxide layers. Capacitors are widely used as parts of electrical circuits in many common electrical devices. Unlike a resistor, an ideal capacitor does not dissipate energy. Instead, a capacitor stores energy in the form of an electrostatic field between its plates.

When there is a potential difference across the conductors (e.g., when a capacitor is attached across a battery), an electric field develops across the dielectric, causing positive charge $+Q$ to collect on one plate and negative charge $-Q$ to collect on the other plate. If a battery has been attached to a capacitor for a sufficient amount of time, no current can flow through the capacitor. However, if a time-varying voltage is applied across the leads of the capacitor, a displacement current can flow.

An ideal capacitor is characterized by a single constant value, its capacitance (in practice we can see a voltage rating on any capacitor that is the maximum amount of voltage that a capacitor can safely be exposed to and can store). Capacitance is defined as the ratio of the electric charge Q on each conductor to the potential difference U between them. The SI unit of capacitance is the farad [F], which is equal to one coulomb per volt (1 [C·V^{-1}]). Typical capacitance values range from about 1 [pF] to about 1 [mF], but "super capacitors" of a few centimetres in 12021 (holocene calendar) can store $30,000$ [F].

The larger the surface area of the "plates" (conductors) and the narrower the gap between them, the greater the capacitance is. In practice, the dielectric between the plates passes a small amount of leakage current and also has an electric field strength limit, known as the breakdown voltage. The conductors and leads introduce an undesired inductance and resistance.

Capacitors are widely used in electronic circuits for blocking direct current while allowing alternating current to pass. In analogue filter networks, they smooth the output of power supplies. In resonant circuits they tune radios to particular frequencies. In electric power transmission systems, they stabilize voltage and power flow.

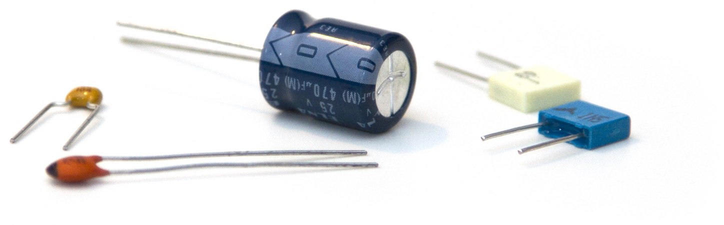


Figure 10.20: **Some capacitive dipoles**
(source: `http://www.e-style.ch`, author: Martin Bircher)

As a direct application of Gauss's theorem, very useful in electronics and for engineers, consider a large thin a flat sheet, wearing a homogeneous surfacic electronic charge density σ and immersed in an environment of absolute electrical permittivity ε. In the area close to its center, the electric field resulting from all the fields of the fields is normal, uniform, constant and go away from the sheet. Let us consider a Gaussian surface with a cylinder shape limited by the bases $S_1=S_2=S$ and its tubular surface S_3 symmetrical with respect to the sheet. It therefore encloses an electric charge σS. It follows that:

$$\oint \vec{E}\circ\mathrm{d}\vec{S}=E_1S_1+E_2S_2+E_3S_3=\frac{\sum_i q_i}{\varepsilon} \tag{10.1.115}$$

and as $E=E_1=E_2$ and $E_3=0$, we find:

$$ES+ES=2ES=\frac{Q}{\varepsilon}\Rightarrow E=\frac{Q/S}{2\varepsilon}=\frac{Q/S}{2\varepsilon}=\frac{\sigma}{2\varepsilon} \tag{10.1.116}$$

Finally, the electric field of a large and thin loaded flat sheet is:

$$E=\frac{\sigma}{2\varepsilon} \tag{10.1.117}$$

If we put face to face two identical plates but with opposite charges, the algebraic sum will of course be:

$$E=\frac{\sigma}{\varepsilon} \tag{10.1.118}$$

Excepted of the extremities where the side effect is important, the overall field is everywhere the vector sum of uniform fields from two opposing thin loaded layers. We name such a system a "plane and parallel capacitor":

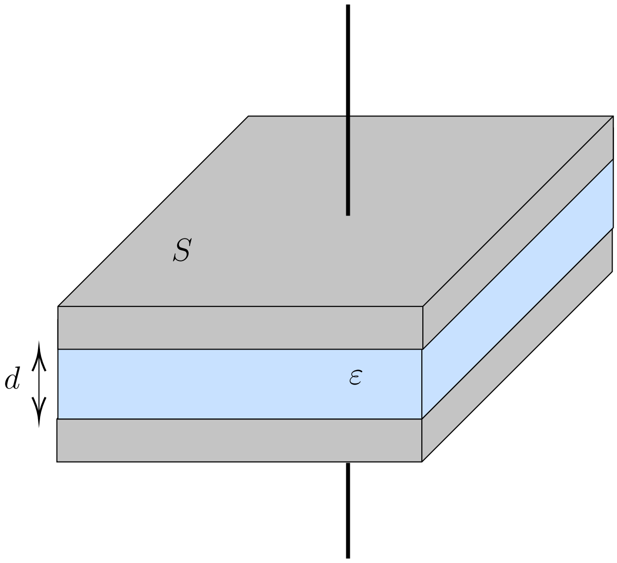


Figure 10.21: **Schematic diagram of plane and parallel capacitor**

The result is remarkable because it is independent of the distance between the planes (in fact never forget it is an approximation for very small distances!). The calculation of electric potential is therefore simplified. Thus:

$$\Delta U=-\int_0^d E\mathrm{d}r=-\frac{\sigma}{\varepsilon}d \Rightarrow U=|\Delta U|=\frac{\sigma d}{\varepsilon}=\frac{Qd}{\varepsilon S} \tag{10.1.119}$$

Thus, the capacitance of the plane and parallel capacitor is consequently:

$$\boxed{C=\frac{Q}{U}=\frac{\varepsilon S}{d}} \tag{10.1.120}$$

Let us now recall that we have shown previously that:

$$E=\frac{\sigma}{2\varepsilon}=\frac{Q}{2S\varepsilon} \tag{10.1.121}$$

Since each plate of a parallel and plane capacitor contributes equally, the total electric field between the plates would be :

$$E_{\text{tot}}=2\frac{Q}{2S\varepsilon}=\frac{Q}{S\varepsilon} \tag{10.1.122}$$

The potential difference is:

$$U=E_{\text{tot}}d=d\frac{Q}{S\varepsilon} \tag{10.1.123}$$

Solving for Q yields:

$$Q=S\varepsilon\frac{U}{d} \tag{10.1.124}$$

The plates are oppositely charged, so the attractive force F_{att} between the two plates is equal to the electric field produced by one of the plates times the charge on the other:

$$\boxed{F_{\text{att}}=Q\frac{Q}{2S\varepsilon}=\frac{\varepsilon SU^2}{2d^2}} \tag{10.1.125}$$

This is therefore the force between the plates of a parallel (infinite) plate capacitor.

Let us see a second academic example that is the "cylindrical capacitor":

The plates of a cylindrical capacitor are two infinite rolled cylinders (or very long relatively to their diameter), coaxial of respective radius R_1 and R_2. It is therefore the very important case of the coaxial cable (which dielectric is often polyethylene) that we can find in many laboratories (and not only!):

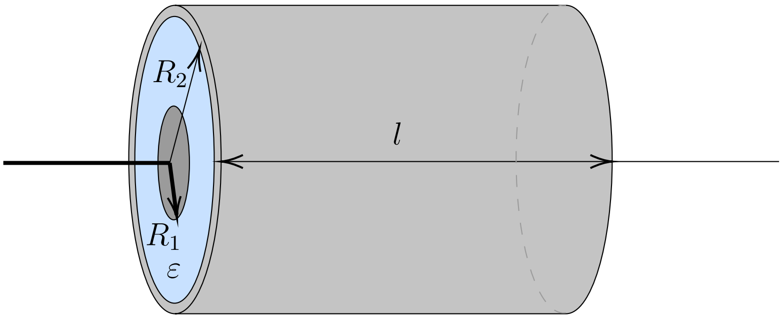


Figure 10.22: Schematic diagram of cylindrical capacitor

By Gauss theorem, we know that:

$$\vec{E}\circ\vec{S}=\frac{Q}{\varepsilon} \tag{10.1.126}$$

And since the field $\vec{E}$ is collinear at every point of the surface $\vec{S}$, it immediately comes by knowing the expression of the surface of the cylinder (see section Geometric Shapes page 2959):

$$E=\frac{Q}{2\pi rl\varepsilon} \tag{10.1.127}$$

But:

$$\begin{aligned}\Delta U&=-\int\limits_{R_1}^{R_2}E\mathrm{d}r=-\int\limits_{R_1}^{R_2}\frac{Q}{2\pi l\varepsilon}\frac{1}{r}\mathrm{d}r=-\frac{Q}{2\pi\varepsilon l}\left[\ln(r)\right]\Big|_{R_1}^{R_2}\\&=-\frac{Q}{2\pi\varepsilon l}\left[\ln(R_2)-\ln(R_1)\right]=-\frac{Q}{2\pi\varepsilon l}\ln\left(\frac{R_2}{R_1}\right)\end{aligned} \tag{10.1.128}$$

Therefore:

$$\boxed{C=\frac{Q}{U}=\frac{Q}{|\Delta U|}=2\pi\varepsilon l\frac{1}{\ln\left(\dfrac{R_2}{R_1}\right)}} \tag{10.1.129}$$

Or often written in the form of the "coaxial capacity per unit length":

$$C=\frac{Q}{U}=\frac{Q}{|\Delta U|}=2\pi\varepsilon\frac{1}{\ln\left(\dfrac{R_2}{R_1}\right)} \tag{10.1.130}$$

Let us also calculate the capacity of a spherical capacitor which corresponds in a first approximation to some of Van Der Graaf generators that we have in the labs of some schools, museums or even research centers...

A "spherical capacitor" consists of two concentric spheres of radius R_1 and R_2 with $R_1 < R_2$.

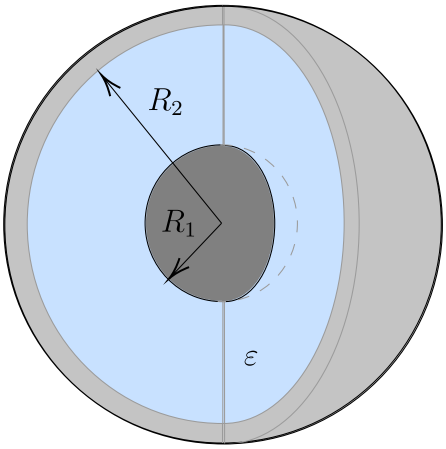


Figure 10.23: Schematic diagram of spherical capacitor

We now have immediately (by Gauss theorem):

$$\vec{E}\circ\vec{S}=\frac{Q}{\varepsilon} \tag{10.1.131}$$

And therefore since the field is assumed to be collinear at any point on the surface it comes immediately knowing the expression of the surface of a sphere:

$$E=\frac{Q}{4\pi r^2\varepsilon} \tag{10.1.132}$$

But:

$$\Delta U=-\int\limits_{R_1}^{R_2} E\mathrm{d}r=-\int\limits_{R_1}^{R_2}\frac{Q}{4\pi\varepsilon}\frac{1}{r^2}\mathrm{d}r=-\frac{Q}{4\pi\varepsilon}\left[\frac{1}{r}\right]\Bigg|_{R_1}^{R_2}=-\frac{Q}{4\pi\varepsilon}\left(\frac{1}{R_2}-\frac{1}{R_1}\right) \tag{10.1.133}$$

Then we have:

$$C=\frac{Q}{U}=\frac{Q}{|\Delta U|}=4\pi\varepsilon\left(\frac{1}{R_2}-\frac{1}{R_1}\right)^{-1} \tag{10.1.134}$$

Then we have:

$$\boxed{C=4\pi\varepsilon\frac{R_1R_2}{R_1-R_2}} \tag{10.1.135}$$

And that's all for the classical examples...

Well! We have just seen that the capacity was defined as:

$$Q=CU \tag{10.1.136}$$

or in alternative current notation (when the current or the voltage varies as function of the time, it is the usage to write them in lowercase):

$$q=Cu\Leftrightarrow\frac{\mathrm{d}q}{\mathrm{d}t}=i=C\frac{\mathrm{d}u}{\mathrm{d}t}\tag{10.1.137}$$

We then have for the instantaneous power (see section Electrokinetics page 4253):

$$p=\Delta ui=ui=(ri-e)i=ri^2-ei\tag{10.1.138}$$

Assuming an ideal capacitor (which does not dissipate energy by Joule effect) we get:

$$p=ui=uC\frac{\mathrm{d}u}{\mathrm{d}t}=Cu\frac{\mathrm{d}u}{\mathrm{d}t}=\frac{\mathrm{d}\left(\dfrac{1}{2}Cu^2\right)}{\mathrm{d}t}\tag{10.1.139}$$

So by integration in a given time interval from 0 to t we have for the second term:

$$E_C=\frac{1}{2}Cu^2\tag{10.1.140}$$

This energy is always positive and is stored in electrostatic form in the capacitor.

In the context of a sine wave alternative current, the average power will be equal to zero. We can generalize this by assuming that a perfect capacitor does not dissipate power no Joule effect.

Remark

Some scientific experiments requiring enormous energy use thousands of giant capacitors charged in the long term to accelerate particles or to make operate megajoules LASER. However we can not actually store the surplus of electric power of some power plants, which is why we use this surplus to pull back the water in the dams that can use their pool as a reserve of potential energy to produce a complement to electricity at the time of peak consumption (inverse transformation). We can also found huge capacitor that have for purpose to manage critical voltage collapse in a changing environment (this are the big cylinder in transformation stations).

Obviously a real capacitor has an internal resistance often depicted as:

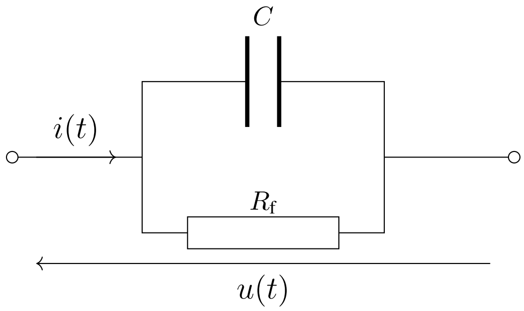


Figure 10.24: **Real capacitor (author: Jimmy Roussel)**

10.1.4.1.1 Dielectric strength

The term "dielectric strength" denoted $\mathcal{R}$ and (measured in SI units volts per meter [V·m^{-2}]) has the following meanings:

- Of an insulating material, the maximum electric field that a pure material can withstand under ideal conditions without breaking down (i.e., without experiencing failure of its insulating properties).
- For a specific configuration of dielectric material and electrodes, the minimum applied electric field (i.e., the applied voltage divided by electrode separation distance) that results in breakdown.

The theoretical dielectric strength of a material is an intrinsic property of the bulk material and is independent of the configuration of the material or the electrodes with which the field is applied. This "intrinsic dielectric strength" corresponds to what would be measured using pure materials under ideal laboratory conditions. At breakdown, the electric field frees bound electrons. If the applied electric field is sufficiently high, free electrons from background radiation may become accelerated to velocities that can liberate additional electrons during collisions with neutral atoms or molecules in a process named "avalanche breakdown". Breakdown occurs quite abruptly (typically in nanoseconds), resulting in the formation of an electrically conductive path and a disruptive discharge through the material. For solid materials, a breakdown event severely degrades, or even destroys, its insulating capability.

The factors affecting apparent dielectric strength are typically:

- it decreases with increased sample thickness.
- it decreases with increased operating temperature.
- it decreases with increased frequency.
- for gases (e.g. nitrogen, sulphur hexafluoride) it normally decreases with increased humidity.
- for air, dielectric strength increases slightly as humidity increases

For a capacitor used in electronics, if we exceed the dielectric strength value, we observe the destruction of the element. This maximum value of the voltage applied to the terminals, is named sometimes the "breakdown voltage" of the capacitor and denoted U_c. We can define the dielectric strength of the material (medium) as:

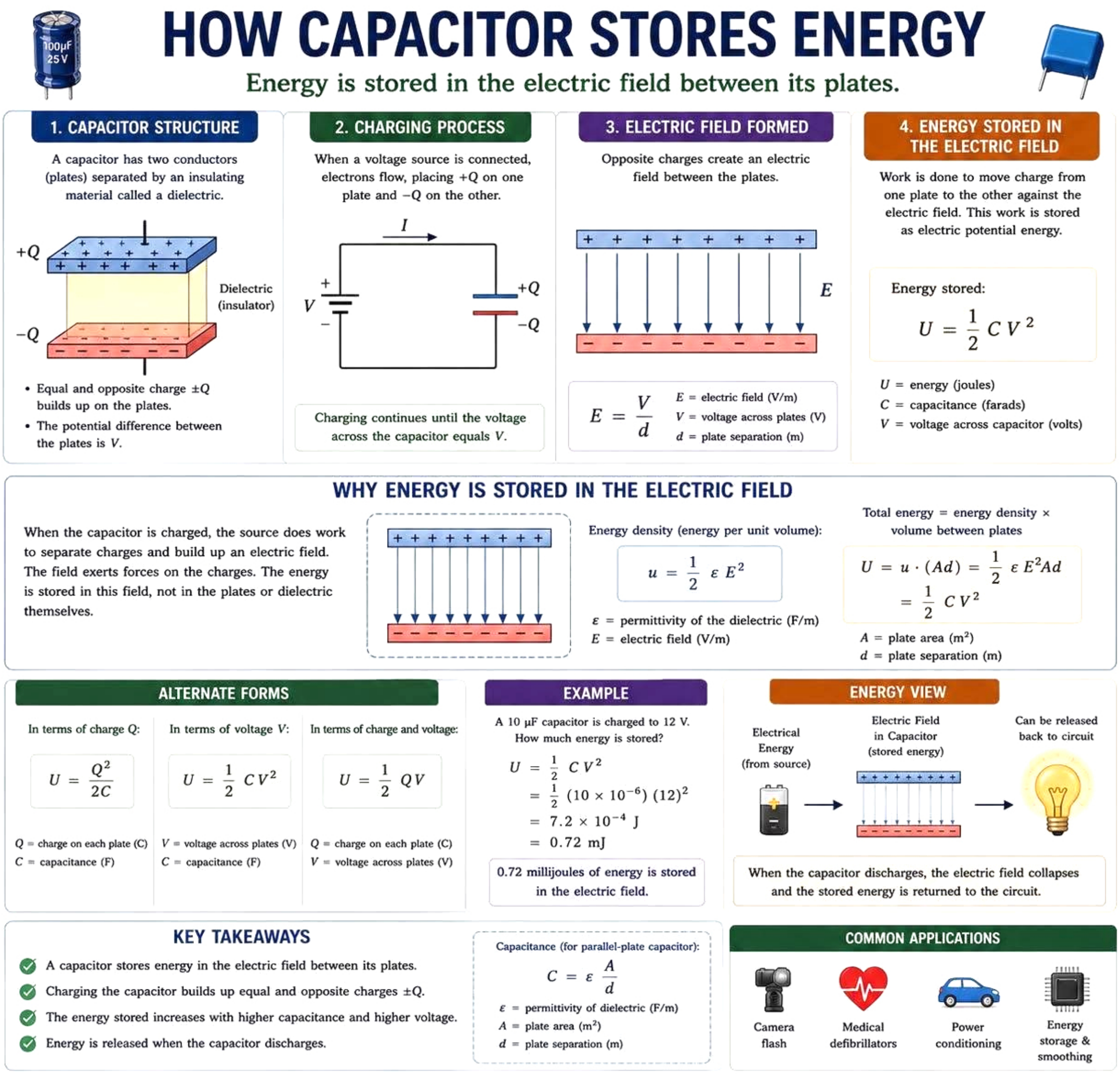
$$\boxed{\mathcal{R}=\frac{U_c}{d}} \tag{10.1.141}$$

✍**Example:**

Two common dielectric strength values that have to be know by students are:

$$\mathcal{R}_{\text{air}} \cong 3 \cdot 10^9 \;[\text{V} \cdot \text{m}^{-1}] \qquad \mathcal{R}_{\text{vacuum}} \cong 10^{21} \;[\text{V} \cdot \text{m}^{-1}] \tag{10.1.142}$$

When we talk of dielectric strength, we also speak of the dielectric is an insulator or a substance that does not conduct electricity and is polarizable by an electric field. In most cases, the dielectric properties are due to the polarization of the substance. When the dielectric (in this case, air is the dielectric) is placed in an electric field, the electrons and protons of its atoms are reoriented and, in some cases, at the molecular level, a polarization is induced (as we have seen in our study of the dipoles above). This polarization produces a potential difference, or voltage, between both terminals of the dielectric; it then stores the energy which becomes available when the electric field is removed. The effectiveness of a dielectric is its relative ability to store energy compared to that of vacuum. It is expressed by the relative electric permittivity ε_r, determined relative to that of a vacuum ε_0. The dielectric strength is then ability of a dielectric to withstand electric fields without losing its insulating properties. An effective dielectric releases a big fraction of the energy it had stored when the electric field is reversed.


100µF
25 V
+Q
−Q
Dielectric (insulator)
I
V
+
−
+Q
−Q
E
Electrical Energy (from source)
Electric Field in Capacitor (stored energy)
Can be released back to circuit
Camera flash
Medical defibrillators
Power conditioning
Energy storage & smoothing

HOW CAPACITOR STORES ENERGY

Energy is stored in the electric field between its plates.

1. CAPACITOR STRUCTURE

A capacitor has two conductors (plates) separated by an insulating material called a dielectric.

- Equal and opposite charge ±Q builds up on the plates.
- The potential difference between the plates is V.

2. CHARGING PROCESS

When a voltage source is connected, electrons flow, placing +Q on one plate and −Q on the other.

Charging continues until the voltage across the capacitor equals V.

3. ELECTRIC FIELD FORMED

Opposite charges create an electric field between the plates.

$$E = \frac{V}{d}$$

E = electric field (V/m)
V = voltage across plates (V)
d = plate separation (m)

4. ENERGY STORED IN THE ELECTRIC FIELD

Work is done to move charge from one plate to the other against the electric field. This work is stored as electric potential energy.

Energy stored:

$$U = \frac{1}{2} C V^2$$

U = energy (joules)
C = capacitance (farads)
V = voltage across capacitor (volts)

WHY ENERGY IS STORED IN THE ELECTRIC FIELD

When the capacitor is charged, the source does work to separate charges and build up an electric field. The field exerts forces on the charges. The energy is stored in this field, not in the plates or dielectric themselves.

Energy density (energy per unit volume):

$$u = \frac{1}{2} \varepsilon E^2$$

ε = permittivity of the dielectric (F/m)
E = electric field (V/m)

Total energy = energy density × volume between plates

$$U = u \cdot (Ad) = \frac{1}{2} \varepsilon E^2 Ad = \frac{1}{2} C V^2$$

A = plate area (m^2)
d = plate separation (m)

ALTERNATE FORMS

In terms of charge Q:

$$U = \frac{Q^2}{2C}$$

Q = charge on each plate (C)
C = capacitance (F)

In terms of voltage V:

$$U = \frac{1}{2} C V^2$$

V = voltage across plates (V)
C = capacitance (F)

In terms of charge and voltage:

$$U = \frac{1}{2} Q V$$

Q = charge on each plate (C)
V = voltage across plates (V)

EXAMPLE

A 10 µF capacitor is charged to 12 V. How much energy is stored?

$$\begin{aligned} U &= \frac{1}{2} C V^2 \\ &= \frac{1}{2} (10 \times 10^{-6}) (12)^2 \\ &= 7.2 \times 10^{-4} \text{ J} \\ &= 0.72 \text{ mJ} \end{aligned}$$

0.72 millijoules of energy is stored in the electric field.

ENERGY VIEW

When the capacitor discharges, the electric field collapses and the stored energy is returned to the circuit.

KEY TAKEAWAYS

- ✓ A capacitor stores energy in the electric field between its plates.
- ✓ Charging the capacitor builds up equal and opposite charges ±Q.
- ✓ The energy stored increases with higher capacitance and higher voltage.
- ✓ Energy is released when the capacitor discharges.

Capacitance (for parallel-plate capacitor):

$$C = \varepsilon \frac{A}{d}$$

ε = permittivity of dielectric (F/m)
A = plate area (m^2)
d = plate separation (m)

COMMON APPLICATIONS

10.1.5 Electrostatic potential energy

Let us consider two electric charges q_1, q_2. The first is supposed to be at rest and fixed; the second is brought from infinity to a distance a from q_1 (the same reasoning was applied to the gravitational field in the section of Classical Mechanics!). Suppose that the two charges are of the same sign. As q_1, q_2 tend to repel each other, we have to provide a potential energy E_p to take q_2 (infinitely slowly) of q_1. The work $\mathrm{d}W$ done by the electrostatic force at any point is by definition as we know:

$$\mathrm{d}W = F \mathrm{d}x \tag{10.1.143}$$

The potential energy of the system is:

$$\mathrm{d}E_p = -\mathrm{d}W \tag{10.1.144}$$

as the force F is resistive (hence the origin of the sign "$-$").

Therefore:

$$\mathrm{d}E_p=-\mathrm{d}W=-F\mathrm{d}x=-\frac{q_1q_2}{4\pi\varepsilon}\frac{1}{x^2}\mathrm{d}x \tag{10.1.145}$$

We then get the potential energy at a point (x in the numerator is simplified with an x in the denominator) to a given sign convention:

$$\boxed{E_p=\frac{q_1q_2}{4\pi\varepsilon}\frac{1}{x}} \tag{10.1.146}$$

If we have $q_2=Zq_1$ where $Z\in\mathbb{N}$ and we denote $q_1=e$ then we fall back on a well known way to write the previous relation:

$$E_p=\frac{Ze^2}{4\pi\varepsilon}\frac{1}{x} \tag{10.1.147}$$

If the path is not linear we have obviously as in Classical Mechanics with the electric field and using the notations seen at the beginning of the section:

$$E_p=-W=-\int\limits_{\Gamma}q\vec{E}\circ\mathrm{d}\vec{s}=\int\limits_{\Gamma}q\vec{\nabla}\cdot U\circ\mathrm{d}\vec{s} \tag{10.1.148}$$

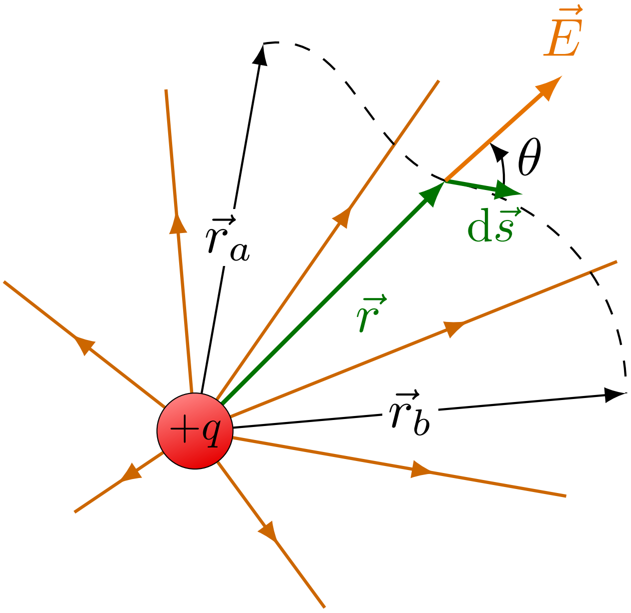


Figure 10.25: **Illustration of the work path in an electric field (author: Izaak Neutelings)**

Notice that:

$$E_p=\frac{q_1q_2}{4\pi\varepsilon}\frac{1}{x} \tag{10.1.149}$$

can also be written as:

$$E_p=q_1U_2=q_2U_1=\frac{1}{2}(q_1U_2+q_2U_1) \tag{10.1.150}$$

Caution! When we do physics, we have to be sure what potential energy we are talking about. This is a major problem! If you take for example the potential energy due the gravitational interaction, it can take any value depending on the reference point. If the reference is at sea level, a point below the sea level will have a negative potential energy, by cons if the reference is the center of the Earth, there will be only positive potential energies. That is why we are writing rather the potential energy in the form of height difference relative to a reference in mechanics. For the potential energy of the electron, we must know with what it interacts. If it is with a negative charge, the product of the electric charge values will be positive and therefore the potential energy of interaction will be positive, if it interacts with a positive charge, the product will be negative and the potential energy will be negative. In short, we must know what we are talking about. This is an example where words are important in physics too (same problems as the foundations of mathematics!).

In general, if the potential energy decreases when the distance increase, the force is repulsive if it increases when the distance increase, the force is attractive.

~90 %

★★★☆☆

75 votes, 66.13%

10.2. Magnetostatics

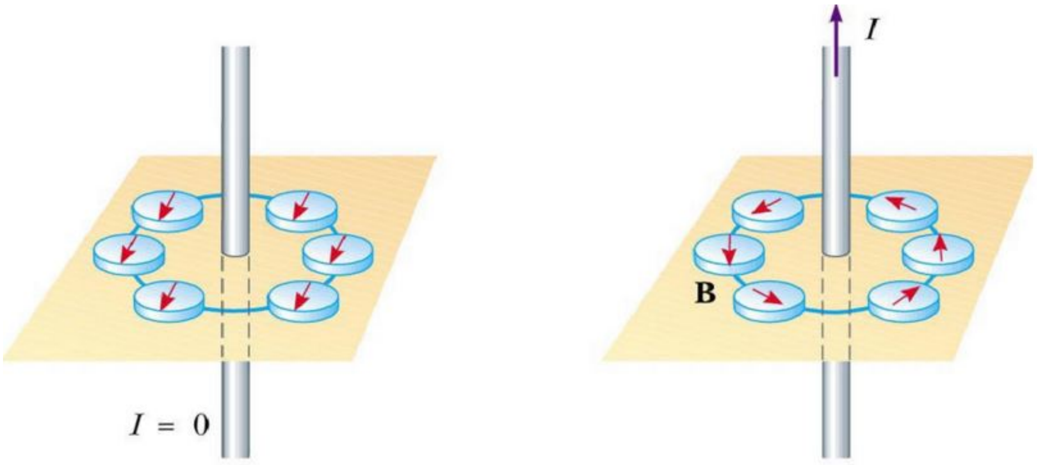
MAGNETS are known since ancient times (without that it was known at that the origin of their properties) under the name "magnetite": black stones found near the city of Magnesia (Turkey). This is from this stone that also comes the current name of the "magnetic field". The Chinese were the first to use the properties of the individual magnets more than 1,000 years before compasses. They consisted of a magnetite needle resting on straw floating on the water contained in a graduated container.

Magnetostatics is the study of magnetic fields in systems where the currents are steady (not changing with time). It is the magnetic analogue of electrostatics, where the charges are stationary. The magnetization need not be static; the equations of magnetostatics can be used to predict fast magnetic switching events that occur on time scales of nanoseconds or less. It is even a good approximation when the currents are not static - as long as the currents do not alternate rapidly.

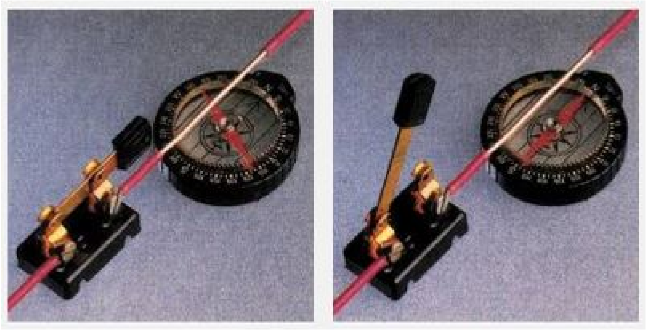
As for the electric field, a good/better understanding of the origin of this field can be done through modern theories such as wave quantum physics and quantum field theory. The beginner reader will have therefore to be patient, as for the study of the electric field, until to get in this textbook the knowledge to study these modern theories.

The quantitative study of the interactions between magnets and currents was made by the physicists Biot and Savart in 11820 only according to holocene calendar (nobody knew before that the compass was influenced by the current inside the core of the Earth). They measured the amplitude of the oscillations of a magnetic needle according to its distance to a straight current. They found that the force acting on a pole is directed perpendicular to the direction connecting the center to this conductor wire and it varies inversely with the distance. This is the first case that we will study.

The idea of the experiment was the following:



Or in real life with a much simple configuration:



Given a movement of electric charges (i.e.: a current I) generating in the space a vector field whose effects are measurable and whose properties differ from those of the electrostatic field. We deduce the existence of a new vector field we name (temporarily) "magnetic field" and which we denote by $\vec{B}$.

The physical units of the magnetic field will naturally been deduced from the time we will be able to connect this magnetic field to something known as a Force for example. This is what we will see further below during our study of the "Laplace's force law".

The simplest case of study consisting of an indefinite straight conductor wire (example that we can also assimilate to a simple move of charges without having necessarily a wire to carry them) carrying a current I (see the section Electrokinetics for the concept of "current") shows that the magnetic field lines are circles having for axis the conductor wire itself:

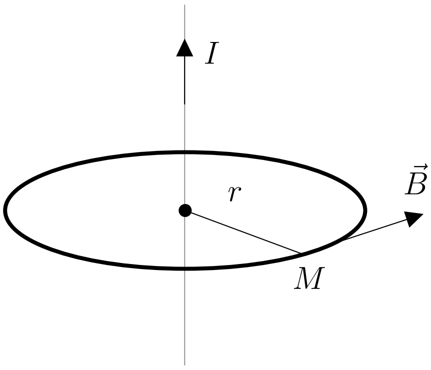
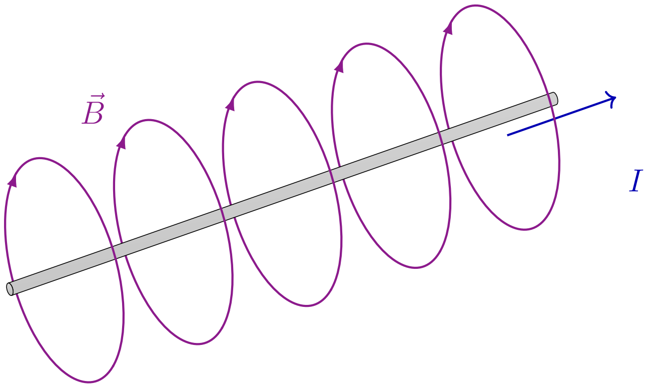


Figure 10.26: **Magnetic field around a straight infinite wire**

Caution!! As we will prove it in the section of Special Relativity (see page 5694), the magnetic field and electric fields from moving charges (do not confuse it with the magnetic field of spin!) are not different things they are the two sides of the same coin. A magnetic field produced by moving charges is just an electric field viewed from a different moving frame of reference!!! More explicitly the motion of a charged particle (which has dimensions and therefore finite charge density) causes a Lorentz contraction along the line of its motion. This causes the particle to see all stationary charges as denser in that direction and, therefore, it experiences an increased electrostatic force from them that is interpreted as a new kind of field: the magnetic field!

Or in perspective (if it may help the reader to understand better the concept):



The direction of $\vec{B}$ is usually defined by the intermittance of "Ampère's observer", that is to say, an observer that would be positioned along the wire, so that the current goes to its feet towards its head and who would look at the point M where we evaluate the magnetic field. The field $\vec{B}$ is directed from the right to the left of this observer. This is represented many times by the following figure:

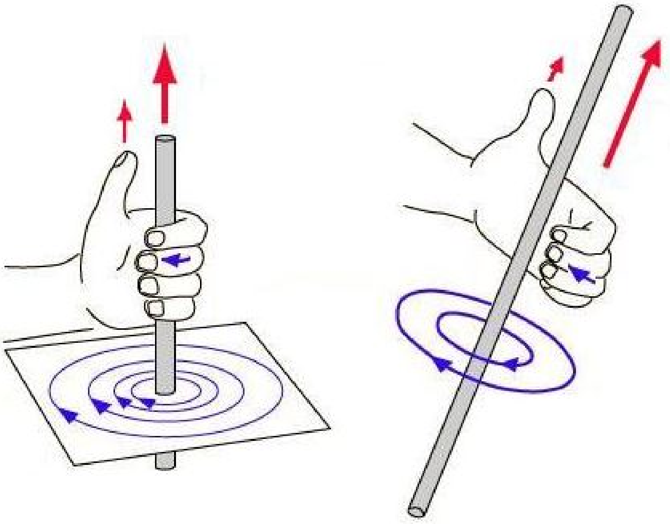


Figure 10.27: Ampère observer hand rule

It is said that is was experimentally established by Biot and Savart in 11820 (holocene calendar) that the norm of the magnetic field $\vec{B}$ at the distance r of the wire is proportional to the current I running through it and inversely proportional to r such that:

$$B = kIr^{-1} \tag{10.2.1}$$

This relation is traditionally considered as the pillar of the study of the magnetic field.

The proportionality coefficient k depends as always to selected unit system (as for other study fields). For the set of all its consequences, it is advantageous to write the above expression in a form that makes appear the perimeter of the circle of radius r. So we put:

$$k = \frac{\mu_0}{2\pi} \tag{10.2.2}$$

Remark

Don't forget that as 2π appears frequently in many theorems of physics and mathematics (more than π alone), this value is frequently denoted by the letter τ.

Thus we obtain the value of the magnetic field at a distance r from a straight conductor wire through which a constant current I pass through:

$$\boxed{B = \frac{\mu_0 I}{2\pi r}} \tag{10.2.3}$$

where μ_0 is a new constant which we name "magnetic permeability of vacuum" (again as for the electrical permittivity, there is a "relative magnetic permeability") and whose value is given as usual in this textbook with the other physical constants in the section Principles of Mechanics chapter.

The units of this constant, even if given in the section Principias of the Mechanics chapter, will be deducted automatically as soon as we have managed to link the magnetic field with the concept already known of "Force" (see below). This is what we will see when we will study the "Laplace's force law".

10.2.1 Ampère's theorem

In classical electromagnetism, the "Ampère's theorem" or "Ampère's circuital law" or simply "Ampère's law" relates the integrated magnetic field around a closed loop to the electric current passing through the loop.

It is interesting to calculate the "circulation of the magnetic field" $\vec{B}$ in vacuum (or not) along a contour Λ that turns once in the positive direction (anti-clockwise) around the wire oriented in the direction of the current (Ampère observer):

$$C_B = \oint\limits_{\Gamma} \vec{B} \circ \mathrm{d}\vec{l} = \oint\limits_{\Gamma} B \cdot \mathrm{d}l = \int\limits_0^{2\pi} B \cdot r \cdot \mathrm{d}\theta = \frac{\mu_0 \cdot I}{2\pi} \int\limits_0^{2\pi} \mathrm{d}\theta = \mu_0 \cdot I \tag{10.2.4}$$

Remark

The field is collinear along the path as we have seen previously hence the fact that the dot product can be written as a simple product of norms.

We obtain then by definition the "Ampère's law" (or erroneously named "Ampère's theorem" because this result is not provable ... at least as far as we know...):

$$\boxed{C_B=\oint\limits_{\Gamma}\vec{B}\circ \mathrm{d}\vec{l}=\mu_0 I} \tag{10.2.5}$$

where the current I in a high symmetry system can be assimilated to a simple algebraic sum of currents snared by the path such that we have:

$$C_B=\oint\limits_{\Gamma}\vec{B}\circ \mathrm{d}\vec{l}=\mu_0\sum_k I_k \tag{10.2.6}$$

Warning!!! This is not because the circulation of the magnetic field is zero in a region of space that the magnetic field is zero at any point!

Remarks

R1. The Ampère's law will gives us the possibility to determine the fourth Maxwell equation that we will prove in the section of Electrodynamics.

R2. The prior-previous relation is sometimes wrongly named "Ampère theorem" when in fact this result is not provable as already mentioned. Some physicists, however, use the fourth Maxwell equation to prove the prior-previous relation but then this is the snake biting its tail...

The expression that we obtained can be simplified even more if we introduce a new physical being named "magnetic field intensity" or more commonly "magnetic excitation" and which is denoted by the letter $\vec{H}$ (which is intrinsically independent of the propagation medium!).

If we consider that we are always in the vacuum where there is no magnetic dipole then we define it in the vacuum by:

$$\vec{B}=\mu_0\vec{H} \tag{10.2.7}$$

Therefore, we are often taken to speak about "magnetic induction" for $\vec{B}$ and "magnetic field" for $\vec{H}$. But both are happily confused following the authors/professors/teachers and especially depending on the contexts (as will be the case also in this textbook). When we deal with magnets that have intrinsic magnetization by the properties of the material that are made from, we denote distinctly the external magnetic field by:

$$\boxed{\vec{B}=\mu_0(\vec{H}+\vec{M})} \tag{10.2.8}$$

which is a more general form of the previous relation. It is therefore usual to define the "magnetic susceptibility" as being the dimensionless ratio:

$$\chi_m=\frac{M}{H} \tag{10.2.9}$$

Thus, the magnetic susceptibility indicates the amplitude with which a material responds magnetically to the presence of a magnetic excitation. We then get by this definition, the relation between relative magnetic permeability and magnetic susceptibility:

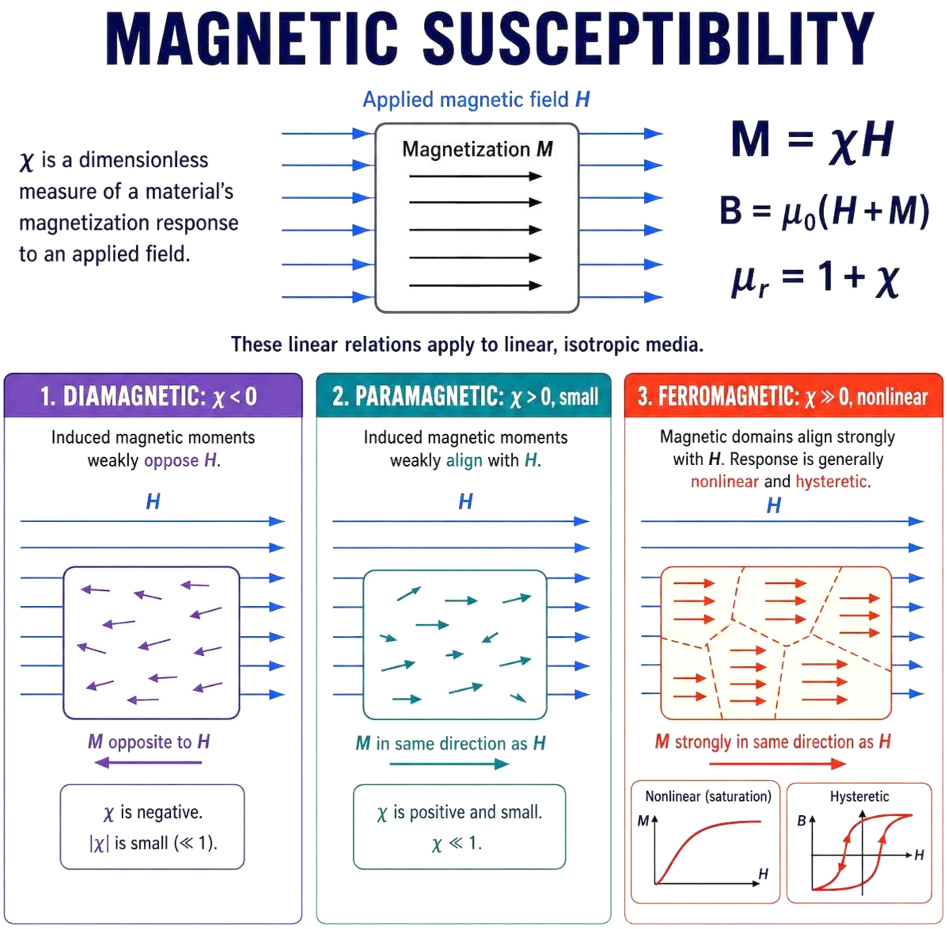
$$\vec{B}=\mu_0(\vec{H}+\vec{M})=\mu_0(\vec{H}+\chi_m)\vec{H}=\mu_0\mu_r\vec{H}=\mu\vec{H} \tag{10.2.10}$$

Therefore:

$$\boxed{\mu_r=\frac{\mu}{\mu_0}=1+\chi_m} \tag{10.2.11}$$

where μ is named the "absolute magnetic permeability".

It is customary to name materials that have a positive magnetic susceptibility "paramagnetic materials" (contributing to increase the magnetic field) and those with a negative magnetic susceptibility "diamagnetic materials" (tend to oppose to the magnetic field). We will see later the Langevin theoretical models to explain quantitatively with a relatively good approximation the two phenomena (in both cases the magnetic susceptibility has a value which is very low).


MAGNETIC SUSCEPTIBILITY
Applied magnetic field H
χ is a dimensionless measure of a material's magnetization response to an applied field.
Magnetization M
M = χH
B = μ0(H+M)
μr = 1 + χ
These linear relations apply to linear, isotropic media.
1. DIAMAGNETIC: χ < 0
Induced magnetic moments weakly oppose H.
H
M opposite to H
χ is negative.
|χ| is small (≪ 1).
2. PARAMAGNETIC: χ > 0, small
Induced magnetic moments weakly align with H.
H
M in same direction as H
χ is positive and small.
χ ≪ 1.
3. FERROMAGNETIC: χ ≫ 0, nonlinear
Magnetic domains align strongly with H. Response is generally nonlinear and hysteretic.
H
M strongly in same direction as H
Nonlinear (saturation)
M
H
Hysteretic
B
H

So finally we can also write the Ampère's Law as following:

$$\oint_{\Gamma}\vec{H}\circ\mathrm{d}\vec{l}=I \tag{10.2.12}$$

The interest of the Ampère's law and of the concept of circulation of the magnetic field may appears like this more evident.

This latter relation obviously is obviously very useful in theoretical physics because it will allow us to determine other important powerful results. Otherwise, in practice, the physicist or electrician/electrical engineer will often be faced with having to use electromagnets for small and medium experiences which he may wish to recalibrate the nominal values, or even solenoids.

10.2.1.1 Infinitely long solenoid

Also a particularly important application in electronics and electrical engineering is the calculation of the induction field in a coil of wire through which pass a current that we will consider as constant in a first time. This is nothing more than an induction coil more technically named an "inductance". Let's see what it is:

A solenoid is a coil of wire formed by a conductor wire wound helically through which pass a current of intensity I. In what follows, we assume that the induction field $\vec{B}$ of a solenoid is zero between the coils and parallel to the axis of the solenoid.

Let us consider the following figure and let us look in approximation only to the inside part of the solenoid admitting that the external field is zero by the infinite length of it and that the joins of the coils are perfect (not letting dissipate any magnetic field)...:

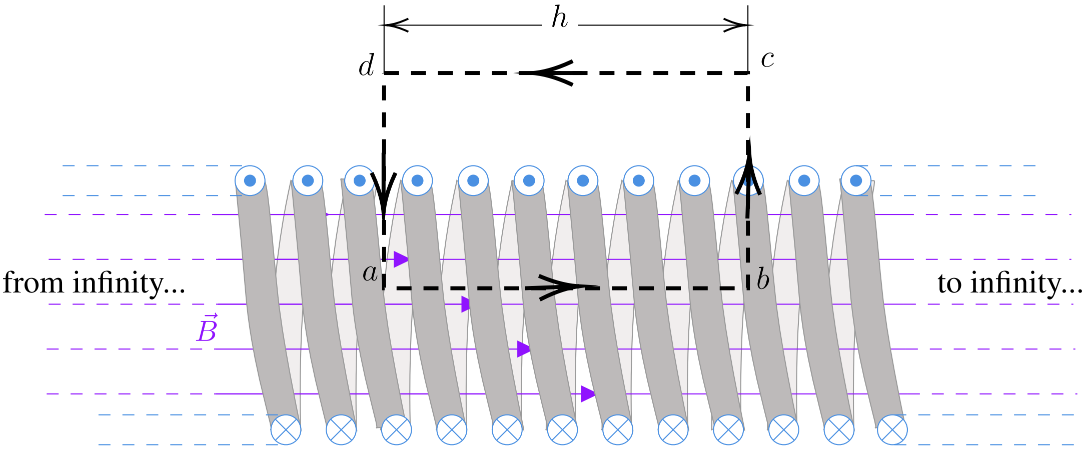


Figure 10.28: Infinite solenoid (coil of wire)

Let us apply the Ampère law to a rectangular path $abcd$. Therefore:

$$\oint \vec{B} \circ \mathrm{d}\vec{l} = \int_a^b \vec{B} \circ \mathrm{d}\vec{l} \qquad (10.2.13)$$

The first integral of the right member gives $B \cdot h$ where B is the norm of $\vec{B}$ inside the solenoid and h the length of the segment $\overline{ab}$. We can notice that the segment $\overline{ab}$, even if it is parallel to the axis of the solenoid, don't need not coincide with it.

The second and fourth integral is equal zero because for these two segments $\vec{B}$ and $\mathrm{d}\vec{l}$ are everywhere perpendicular: as $\vec{B}\circ\mathrm{d}\vec{l}$ is zero everywhere, the two integrals are equal to zero. The third integral is also equal to zero since the segment is calculated outside the solenoid where we assumed that the magnetic field is equal to zero as we consider the solenoid to be perfect.

Thus, the integral $\oint\vec{B}\circ\mathrm{d}\vec{l}$ for the entire rectangular path is equal to $B\cdot h$ such that:

$$\oint\vec{B}\circ\mathrm{d}\vec{l}=B\cdot h=\mu_0 I \tag{10.2.14}$$

but the current I is the sum of the currents I_0 passing through the N turns contained in the path of integration. But in electronics, we used to work with the value n (we choose the lowercase letter by analogy to Thermodynamics where lowercase letters represent densities) which is the number of turns per unit length:

$$n=\frac{N}{h} \tag{10.2.15}$$

Therefore we have:

$$B\cdot h=\mu_0\cdot I=\mu_0\cdot I_0\cdot N\Rightarrow B=\mu_0 I_0 n \tag{10.2.16}$$

That is:

$$\boxed{B=\mu_0\frac{I_0N}{h}=\mu_0 I_0 n} \tag{10.2.17}$$

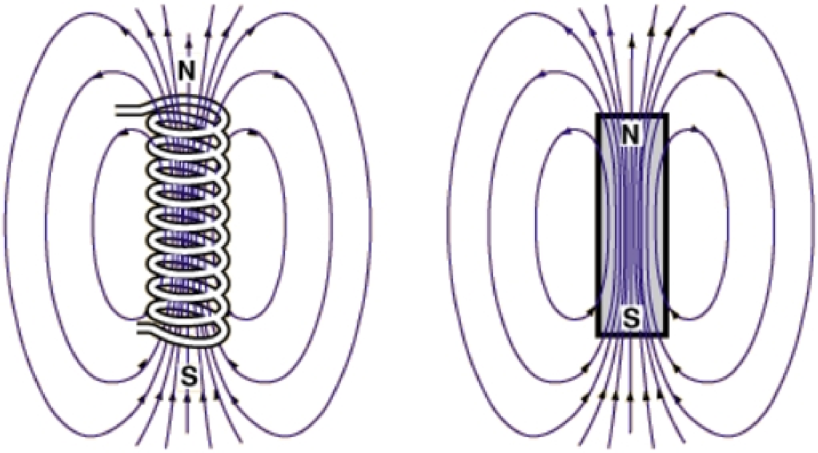


Figure 10.29: **Linear solenoid analogy with (dipole) magnet (source: ?)**

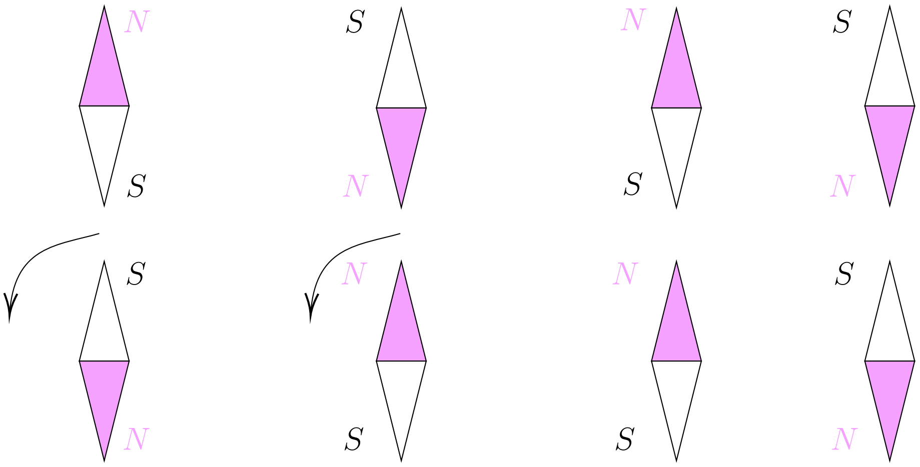
Even if this relation has been established for an ideal infinite solenoid, it gives a pretty good magnitude (but not exact!) of magnetic induction field for interior points near the center of a real solenoid! This relation also shows that the magnetic field is independent in approximation of the solenoid diameter and that this latter is uniform across the section thereof!!! In laboratories, a solenoid is a convenient device for producing a uniform field induction in the same way that the plane capacitor is used to produce a uniform electric field.

If we approach two solenoids (or magnetic needle) free to orient themselves we notice that :

- Two poles of the same nature repel each other
- Two poles of different nature attract

This property is used to determine the nature of the poles of any magnet: it suffices to approach a compass towards one of the ends of the magnet and observe which pole is attracted. The unknown magnet pole will then be of a different nature.

Here is an illustration of the magnetic actions between two magnets:



a) Unstable positions: the poles of the same nature repel each other

b) Stable positions: the poles of different nature attracts

As magnetic field can't be perfectly isolated because any lab on Earth and in space is inside the Earth's natural magnetic field (same as for the gravitational field), the world isolation record known so far is held by the shielding room at the Paul Scherrer Institute in Switzerland (see photo below) with a magnetic field measured as less than 150 picotesla (for comparison the Earth's magnetic field is 30 to 60 microtesla, several hundred thousand times stronger):

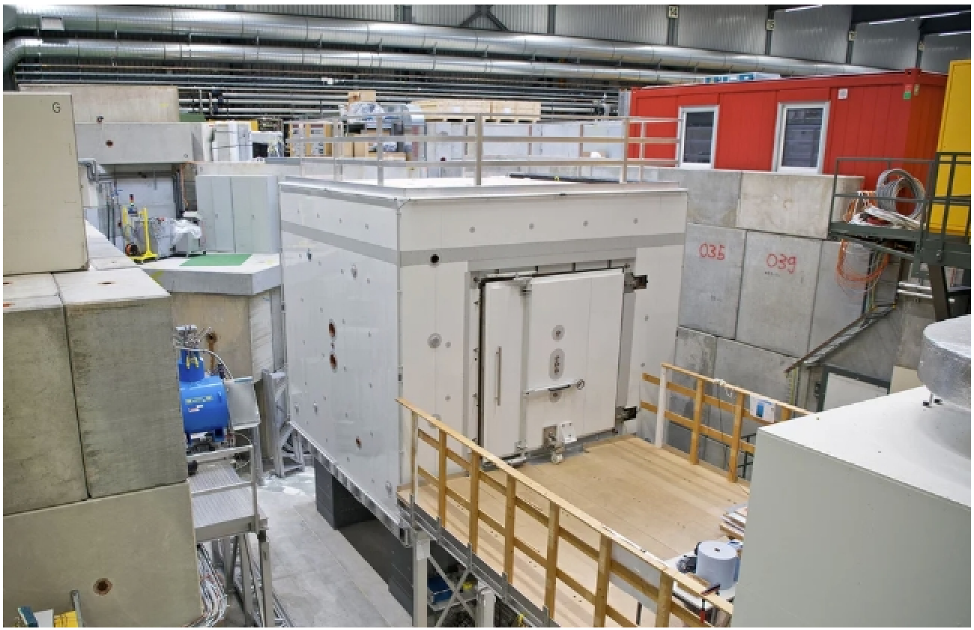


Figure 10.30: **The shielding room at Paul Scherrer Institute**

10.2.1.2 Toroidal coils

The toroidal coil is another important example of the application of Ampère's law. Indeed, we especially find this configuration in low power electronics (e.g. computers) where the inductors are mostly toroidal or in the energy production with the famous Tokomak that are (very ...) schematically reduced to toroidal coils.

Figure 10.31: **Toroidal coil**

For symmetry reasons, it is clear that the magnetic induction lines form concentric circles inside the coil. Let us apply Ampère's law to the radius r of integration of the circular path:

$$\oint \vec{B} \circ \mathrm{d}\vec{l} = \mu_0 I \tag{10.2.18}$$

That is to day:

$$B 2\pi r = \mu_0 I = \mu_0 I_0 N \tag{10.2.19}$$

If follows then:

$$\boxed{B = \frac{\mu_0}{2\pi} \frac{I_0 N}{r}} \tag{10.2.20}$$

Thus, unlike B inside a solenoid, B is not constant within the toroidal coil.

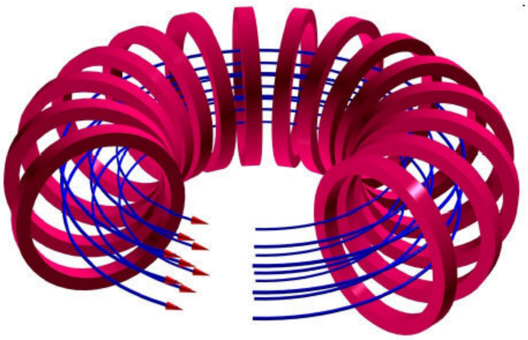


Figure 10.32: **Toroidal field**

10.2.1.3 Electromagnet

Let us determine for example (important and interesting one!) the magnetic field in the air gap of length L_a and section S_a of an electromagnet of length L_{Fe} and section S_{Fe} as shown below:

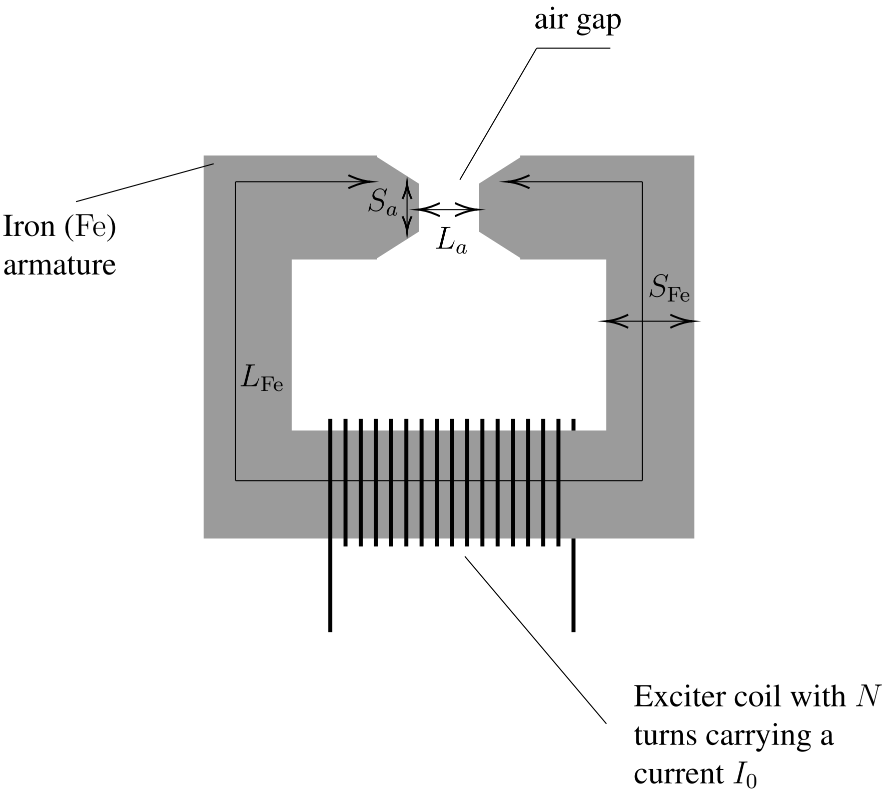


Figure 10.33: Schematic rectangular electromagnet

The Ampère's law gives us in vacuum:

$$\oint_{\sigma} \vec{B} \circ \mathrm{d}\vec{l} = \mu_0 \sum_k I_k \tag{10.2.21}$$

in the case of the electromagnet, we can write that the circulation of the field is the sum of the circulation of the air gap field and the magnet field itself:

$$\oint_{\Sigma} \vec{H} \circ \mathrm{d}\vec{l} = H_{\text{Fe}} L_{\text{Fe}} + H_{\text{airgap}} L_{\text{airgap}} = \sum_k I_k = N I_0 \tag{10.2.22}$$

where N corresponds to the number of current loops surrounding the magnet and which allows the production of the magnetic field.

Let us recall that we have by definition:

$$H_{\text{Fe}} = \frac{B_{\text{Fe}}}{\mu_{\text{Fe}}} \quad \text{and} \quad H_{\text{airgap}} = \frac{B_{\text{airgap}}}{\mu_{\text{airgap}}} \tag{10.2.23}$$

Therefore:

$$\oint_{\Sigma} \vec{H} \circ \mathrm{d}\vec{l} = H_{\text{Fe}} L_{\text{Fe}} + H_{\text{airgap}} L_{\text{airgap}} = \frac{B_{\text{Fe}}}{\mu_{\text{Fe}}} L_{\text{Fe}} + \frac{B_{\text{airgap}}}{\mu_{\text{airgap}}} = N I_0 \tag{10.2.24}$$

If the airgap is not to big $L_{\text{Fe}} >> L_{\text{airgap}}$ we can write:

$$B_{\text{airgap}} \cong B_{\text{Fe}} \tag{10.2.25}$$

Therefore:

$$NI_0 = B_{\text{airgap}}\left(\frac{L_{\text{Fe}}}{\mu_{\text{Fe}}}+\frac{L_{\text{airgap}}}{\mu_{\text{airgap}}}\right) \Leftrightarrow B_{\text{airgap}} = \frac{NI_0}{\dfrac{L_{\text{Fe}}}{\mu_{\text{Fe}}}+\dfrac{L_{\text{airgap}}}{\mu_{\text{airgap}}}} = \frac{NI_0}{\dfrac{L_{\text{Fe}}}{\mu_{\text{Fe}}\mu_0}+\dfrac{L_{\text{airgap}}}{\mu_{\text{airgap}}\mu_0}} \tag{10.2.26}$$

Finally:

$$\boxed{B_{\text{airgap}} \cong \frac{\mu_0 NI_0}{\dfrac{L_{\text{Fe}}}{\mu_{\text{Fe}}}+L_{\text{airgap}}}} \tag{10.2.27}$$

The relation is the same for an electromagnet having two coils! The reader will have perhaps notice on the way that this relation can also be used experimentally if we seek to determine the value of the relative magnetic permeability of the iron when all other parameters are known.

10.2.1.3.1 Strength of a magnet or electromagnet

If we have know the norm of the magnetic field B produced by a magnet at its surface, we can calculate with a certain approximation the force required to detach it from an iron surface.

For this, we will denote by F the necessary force to take off the magnet from a distance d of an iron surface. We will assume the distance d to be sufficiently small distance so that we can accept that the in the entire volume between the magnet and the iron, the magnetic field is constant.

Thus, the work done by the force F is (see section Classical Mechanics page 3293):

$$W = Fd \tag{10.2.28}$$

This work has been turned into magnetic field energy in the volume defined between the magnet and iron. The volumetric energy density due to the magnetic field of air being (see section Electrodynamics page 4156):

$$u_B = \frac{1}{2\mu}B^2 \tag{10.2.29}$$

The volume of the space generated between the magnet and the iron being equal to $V = Sd$ where S is the surface of the magnet that was bonded to iron. We then have the following dimensional equivalence:

$$Fd = \left(\frac{1}{2\mu_0}B^2\right)Sd \tag{10.2.30}$$

Hence we deduce the contact force for small values of d:

$$\boxed{F = \left(\frac{1}{2\mu_0}B^2\right)S} \tag{10.2.31}$$

where B is the limit value of the magnetic field which causes our material to stick to the magnet (so that by raising the magnet, the associated material follows).

If we look at a lifting electromagnet of radius 0.75 [m] capable of lifting 200 [kg]:

Figure 10.34: **Lifting electromagnet**

Then we have:

$$B=\sqrt{\frac{2\mu_0 mg}{S}}=\sqrt{\frac{2\mu_0 mg}{\pi R^2}}\cong 0.05\;[\text{T}] \tag{10.2.32}$$

We can also use roughly the same calculation to determine the magnetic field of the electromagnet of the famous worldwide following playful toy known by people passionate in physics:

Figure 10.35: **Playful lifting electromagnet**

10.2.2 Maxwell-Ampère Relation

Given $\vec{J}$ the current density at any point of the space in the case of a three-dimensional distribution and let S be a closed surface which is based on any contour Γ. The current I that pass through Γ is of course given by:

$$I=\oint\limits_S \vec{J}\circ\mathrm{d}\vec{S} \tag{10.2.33}$$

According to Ampère's law, the flow of the magnetic field along Γ is equal to this integral. It may therefore take here, given the selected contour Γ, an infinite continuous variable values. On the other hand, the Stokes' theorem (see section Vector Calculus page 1889) gives us that:

$$\int\limits_\Gamma \vec{B}\circ\mathrm{d}\vec{l}=\oint\limits_S \vec{\nabla}\times\vec{B}\circ\mathrm{d}\vec{S} \tag{10.2.34}$$

Therefore:

$$\int\limits_\Gamma \vec{B}\circ\mathrm{d}\vec{l}=\mu_0 I=\mu_0\oint\limits_S \vec{J}\circ\mathrm{d}\vec{S}=\oint\limits_S \vec{\nabla}\times\vec{B}\circ\mathrm{d}\vec{S} \tag{10.2.35}$$

and finally we conclude that:

$$\boxed{\vec{\nabla}\times\vec{B}=\mu_0\vec{J}} \tag{10.2.36}$$

We can make a bold comparison of this result with the relation below (proven in the section Electrodynamics), by extension of the static and dynamic electric charge:

$$\vec{\nabla}\circ\vec{E}=\frac{\rho}{\varepsilon} \tag{10.2.37}$$

which is nothing else than the first Maxwell equation (see section Electrodynamics page 4095). Therefore, as we have seen in the section Electrostatics, we have:

$$\vec{E}=\frac{1}{4\pi\varepsilon}\iiint\limits_V \frac{\rho(\vec{r})}{r^2}\mathrm{d}V \tag{10.2.38}$$

By analogy, the idea is to put (this assumption is verified a little further below in the text by the remarkable results obtained):

$$\vec{B}=\frac{\mu_0}{4\pi}\oint_V \frac{J(\vec{r})}{r^2}\mathrm{d}V \tag{10.2.39}$$

relation that we can write in a more elegant way by assuming the current not dependent on the observer's position in space and collinear with perpendicular vector perpendicular of the crossing surface:

$$\vec{B}=\frac{\mu_0}{4\pi}\oint_V \frac{J(\vec{r})}{r^2}\mathrm{d}V=\frac{\mu_0}{4\pi}\oint_V \frac{\dfrac{I(\vec{r})}{S}}{r^2}\mathrm{d}V=\frac{\mu_0}{4\pi}\oint_\Gamma \frac{I(\vec{r})}{r^2}\mathrm{d}l=\frac{\mu_0}{4\pi}I\oint_\Gamma \frac{\mathrm{d}\vec{l}}{r^2} \tag{10.2.40}$$

where Γ represents the perimeter of the wire in which the current I flows.

10.2.3 Biot-Savart law

Form the latest development, we get:

$$\mathrm{d}\vec{B}=\frac{\mu_0}{4\pi}\frac{I\mathrm{d}\vec{l}}{r^2} \tag{10.2.41}$$

Remember that at the last stage of our previous developments (we have specified it implicitly) the integration path is perpendicular to the electric current! But the magnetic field can not be zero at any point of the current line. Therefore, we are led to write what is hidden:

$$\mathrm{d}\vec{B}=\frac{\mu_0 I\mathrm{d}\vec{l}}{4\pi r^2}\cdot 1=\frac{\mu_0 I\mathrm{d}\vec{l}}{4\pi r^2}\cdot\sin\left(\frac{\pi}{2}\right) \tag{10.2.42}$$

The above relation allows us thus, by extension, to write in a more general form:

$$\boxed{\mathrm{d}\vec{B}=\frac{\mu_0 I\sin(\alpha)}{4\pi r^2}\cdot\vec{l}=\frac{\mu_0 I\vec{l}\times\vec{r}}{4\pi r^2}} \tag{10.2.43}$$

which is nothing else than the "Biot-Savart law" often presented first in high-school classes as starting point from the study of magnetism (originally it was experimentally determined in 11820 (holocene calendar) by Jean-Baptiste Biot and Félix Savart with the mathematical help of Simon de Laplace).

This latter relation may as well be written (very important form):

$$\mathrm{d}\vec{B}=\frac{\mu_0}{4\pi r^3}I\mathrm{d}\vec{l}\times\vec{r}=\frac{\mu_0}{4\pi r^3}\frac{\mathrm{d}Q}{\mathrm{d}t}\mathrm{d}\vec{l}\times\vec{r}=\frac{\mu_0}{4\pi r^3}\mathrm{d}Q\frac{\mathrm{d}\vec{l}}{\mathrm{d}t}\times\vec{r}=\frac{\mu_0\mathrm{d}Q}{4\pi r^3}\vec{v}\times\vec{r} \tag{10.2.44}$$

Therefore:

$$\boxed{\vec{B}=\frac{\mu_0 Q}{4\pi r^3}\vec{v}\times\vec{r}} \tag{10.2.45}$$

Here we fall back on the non-relativistic approximation of the magnetic field as we have determined it in our study of Special Relativity (see section Special Relativity page 5694), where we have prove that:

$$\vec{B}=\frac{Q}{c^2\varepsilon_0 4\pi r^3}\frac{1-v^2/c^2}{\left[(1-(v^2/c^2)\sin^2(\theta)\right]^{3/2}}\vec{v}\times\vec{r} \tag{10.2.46}$$

Another important form of the expression for the magnetic field is:

$$\mathrm{d}\vec{B}=\frac{\mu_0}{4\pi}\frac{I\mathrm{d}\vec{l}\times\vec{r}}{r^3}=\frac{\mu_0}{4\pi}\frac{(J\cdot S)\mathrm{d}\vec{l}\times\vec{r}}{r^3} \tag{10.2.47}$$

As the current density $\vec{J}$ is collinear $\mathrm{d}\vec{l}$, we can write:

$$\mathrm{d}\vec{B}=\frac{\mu_0}{4\pi}\frac{(S\mathrm{d}l)\vec{J}\times\vec{r}}{r^3}=\frac{\mu_0}{4\pi}\frac{\mathrm{d}V\vec{J}\times\vec{r}}{r^3}\mathrm{d}V=\frac{\mu_0}{4\pi}\frac{\vec{J}\times\vec{r}}{r^3}\mathrm{d}V \tag{10.2.48}$$

Therefore:

$$\boxed{\vec{B}=\frac{\mu_0}{4\pi}\oint\limits_V \frac{\vec{J}\times\vec{r}}{r^3}\mathrm{d}V} \tag{10.2.49}$$

An important remark is necessary to our level of our discussion in the context of pre-university academic studies, mathematical formulations of magnetic and electric fields are considered as unprovable laws from which we later deduce the Maxwell's equations (furthermore the developments are not of the most aesthetic and rigorous type). The experimental aspect of such important relations can give a negative image of theoretical physics to students. It is therefore clear that at university, we have an approach somewhat less pragmatic.

Indeed, we postulate the Schrödinger equation (see section Wave Quantum Physics page 4759) which we use to prove the non-relativistic formulation of Coulomb's law with the Yukawa theory (see section Quantum Field Theory page 5164). Then, during the study of relativity (see section Special Relativity page 5694), we determine the relativistic form of Coulomb's law. Then we admit the existence of the magnetic field whose expression is experimentally given by the Lorentz force (see further below in this section) and by the properties of Lorentz transformations and knowledge of the relativistic expression of the Coulomb's law, we determine the expression of relativistic magnetic field. Then, by non-relativistic approximation, we fall back on the Biot-Savart law. This approach is much more welcomed by students but not necessarily accessible to all levels.

Now let us come back on the Biot-Savart law. An important example in astrophysics of the Biot-Savart law in the context of plasma jets accretion disks are the unique circular current loops (we also have to add to this the Lorentz force in the relativistic framework for understanding the dynamics of these jets).

10.2.3.1 Magnetic field for a current loop

The figure below is a good example of a current loop seen from profile:

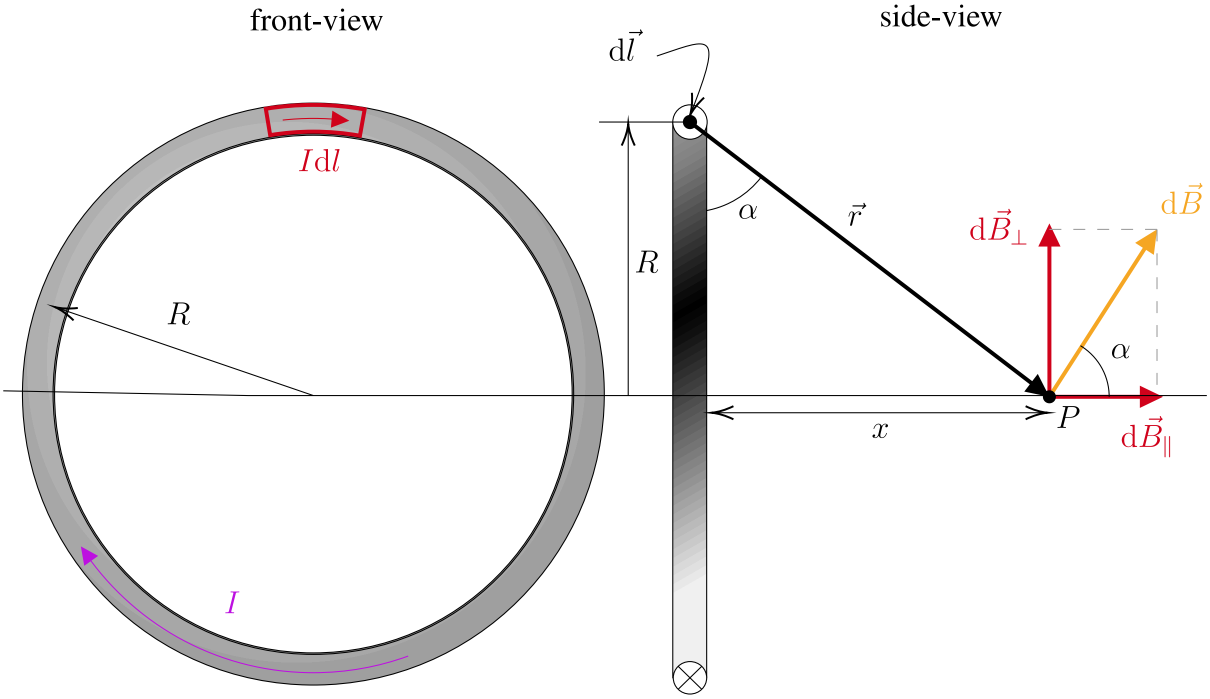


Figure 10.36: **Magnetic field for a current loop**

So we have a circular loop of radius R traversed by an electric current of intensity I. The objective is to calculate $\vec{B}$ at a point P on the axis of the loop.

The vector $\mathrm{d}\vec{l}$ corresponding to an elementary electric current at the top of the loop exits perpendicularly from the plane of the page (screen). The angle θ between this vector and $\vec{r}$ is therefore $\pi/2$. The plane formed by $\mathrm{d}\vec{l}$ and $\vec{r}$ is normal to the figure. The vector $\mathrm{d}\vec{B}$ produced by this elementary electric current is normal to this plane by the form of the Biot-Savart law. It is therefore in the plane of the figure and at a right angle with the vector $\vec{r}$ as shown on the figure.

Let us decompose $\mathrm{d}\vec{B}$ into two parts: the first $\mathrm{d}\vec{B}_{||}$ is along the axis of the loop and the second $\mathrm{d}\vec{B}_\perp$ is perpendicular to this axis. Only the component $\mathrm{d}\vec{B}_{||}$ contributes to the total magnetic induction at the point P. This is because the components $\mathrm{d}\vec{B}_{||}$ of all elementary currents are positioned on the axis and they are added directly. As for the components $\mathrm{d}\vec{B}_\perp$, they are directed in different directions perpendicular to that axis so that, by symmetry, their contribution is zero on this axis (be really careful with this special case!).

We get:

$$B=\int \mathrm{d}B_{||} \tag{10.2.50}$$

It is a scalar integral performed on all elementary currents. We get by the Biot-Savart law:

$$\mathrm{d}B=\frac{\mu_0 I}{4\pi}\frac{\mathrm{d}l\sin(\pi/2)}{r^2} \tag{10.2.51}$$

In addition, we have according to the figure above:

$$\mathrm{d}B_{||}=\mathrm{d}B\cos(\alpha) \tag{10.2.52}$$

Combining these relations, we get:

$$\mathrm{d}B_{||}=\frac{\mu_0 I}{4\pi}\frac{\cos(\alpha)\mathrm{d}l}{r^2} \tag{10.2.53}$$

The figure reveals that r and α are not independent variables. We can express them in function of the new variable x, the distance between the center of the loop and the point P. The relations between these variables are:

$$\begin{aligned} r&=\sqrt{R^2+x^2}\\ \cos(\alpha)&=\frac{R}{r}=\frac{R}{\sqrt{R^2+x^2}} \end{aligned} \tag{10.2.54}$$

Substituting these values in the expression of $\mathrm{d}B_{||}$, we obtain:

$$\mathrm{d}B_{||}=\frac{\mu_0 IR}{4\pi(R^2+x^2)^{3/2}}\mathrm{d}l \tag{10.2.55}$$

We notice that for all the elementary currents, I, R, x have respectively the same values. Integrating this differential gives:

$$B=\int\mathrm{d}B_{||}=\frac{\mu_0 IR}{4\pi(R^2+x^2)^{3/2}}\oint\mathrm{d}l=\frac{\mu_0 IR}{4\pi(R^2+x^2)^{3/2}}2\pi R=\frac{\mu_0 IR^2}{2(R^2+x^2)^{3/2}} \tag{10.2.56}$$

An important point of this relation is therefore $x=0$ we get:

$$\boxed{B=\frac{\mu_0 I}{2R}} \tag{10.2.57}$$

Another important application of the Biot-Savart law consists to take the previous example, but for any continuous falt shape and considered as punctual and we would like to know the value of the field elsewhere than on the axis of symmetry. The results will be very useful when we will study the Corpuscular Quantum Physics and therefore the magnetic properties of metals.

10.2.3.2 Magnetic field for an infinite wire

Let us also prove (it is an interesting example!) that from the Biot-Savart law:

$$\mathrm{d}B=\frac{\mu_0}{4\pi}\frac{I\mathrm{d}l\sin(\theta)}{r^2} \tag{10.2.58}$$

we can get for an infinite straight wire of radius R the relation:

$$B=\frac{\mu_0 I}{2\pi r} \tag{10.2.59}$$

for $r > R$ that we get already earlier with the Ampère's theorem (which shows the equivalence between the two ways of calculating!).

Let us choose for the infinite straight wire below x as variable:

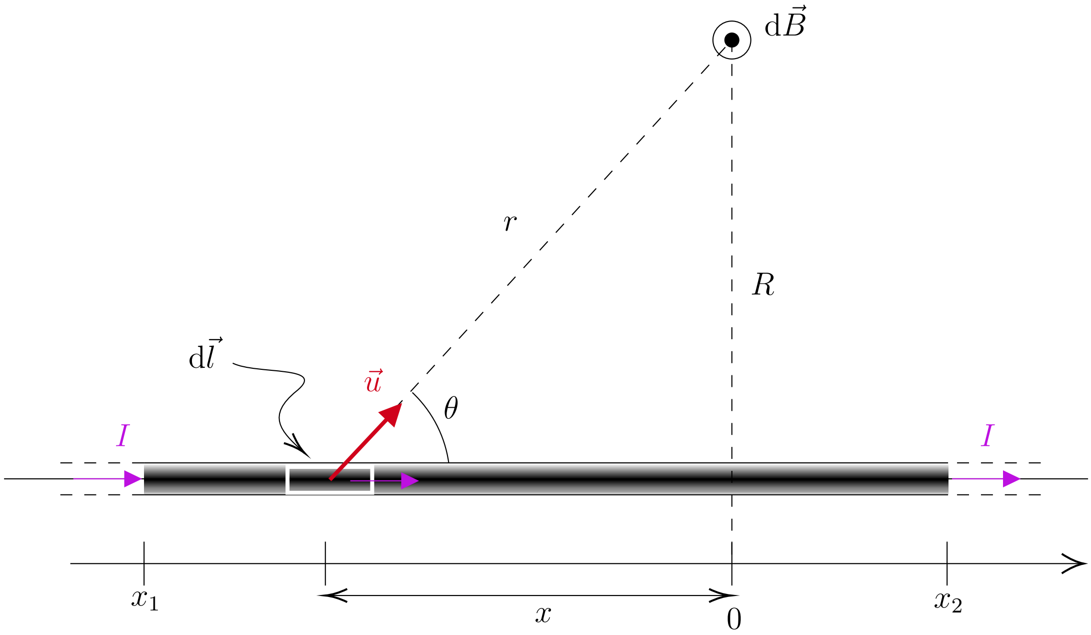


Figure 10.37: **Infinite straight wire**

We then have from the figure above:

$$\begin{aligned} \mathrm{d}l &= \mathrm{d}x \\ r &= \sqrt{x^2+R^2} \\ \sin(\theta) &= \frac{\theta}{r} = \frac{R}{x^2+R^2} \end{aligned} \tag{10.2.60}$$

hence:

$$B = \frac{\mu_0}{4\pi}\frac{I\mathrm{d}x\sin(\theta)}{r^2} = \frac{\mu_0}{4\pi}\frac{I\mathrm{d}x}{r^2}\frac{R}{r} = \frac{\mu_0}{4\pi}IR\frac{\mathrm{d}x}{(x^2+R^2)^{3/2}} \tag{10.2.61}$$

By integrating:

$$\begin{aligned} B &= \frac{\mu_0}{4\pi}IR\int_{x_1}^{x_2}\int\limits_{x_1}^{x_2}\frac{1}{(x^2+R^2)^{3/2}}\mathrm{d}x = \frac{\mu_0}{4\pi}IR\left[\frac{x}{R^2(x^2+R^2)^{1/2}}\right]\Bigg|_{x_1}^{x_2} \\ &= \frac{\mu_0 I}{4\pi R}\left(\frac{x_2}{(x_2^2+R^2)^{1/2}-\dfrac{x_1}{(x_1^2+R^2)^{1/2}}}\right) \end{aligned} \tag{10.2.62}$$

For the remaining part of the development, the trick is to use the following configuration:

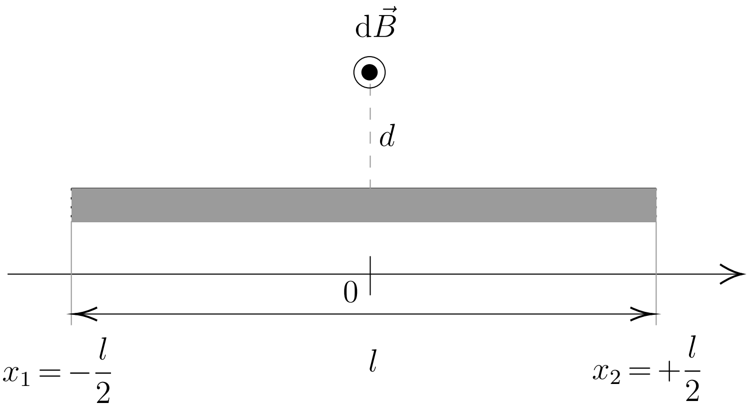


Figure 10.38: **Piece of the infinite wire**

Therefore:

$$x_2 = \frac{l}{2} \qquad x_1 = -\frac{l}{2} \qquad R = d \tag{10.2.63}$$

Which gives us:

$$B = \frac{\mu_0 I}{4\pi d}\left[\frac{l/2}{\left(\frac{l^2}{4}+d^2\right)^{1/2}} - \frac{-l/2}{\left(\frac{l^2}{4}+d^2\right)^{1/2}}\right] = \frac{\mu_0 I}{4\pi d}\frac{l}{\sqrt{\frac{l^2}{4}+d^2}} \tag{10.2.64}$$

After simplification:

$$B = \frac{\mu_0 I l}{2\pi\sqrt{l^2+4d^2}} = \frac{\mu_0 I l}{2\pi d l\sqrt{l^2+\frac{4d^2}{l^2}}} = \frac{\mu_0 I}{2\pi l\sqrt{l^2+\frac{4d^2}{l^2}}} \tag{10.2.65}$$

and so when the wire length tends is very big relatively to l we get for the magnetic field outside the wire ($d > R$):

$$B = \frac{\mu_0 I}{2\pi l\sqrt{l^2+\underbrace{\frac{4d^2}{l^2}}_{\to 0}}} \cong \frac{\mu_0 I}{2\pi d} \tag{10.2.66}$$

Inside of the cable, the circulation integral around a circular path of radius, $r < R$, is the same:

$$\oint \vec{B}\cdot \mathrm{d}\vec{l} = B2\pi r \tag{10.2.67}$$

However, in this case, the smaller Amperian loop does not enclose all of the current flowing through the cable. We are told that the current density, J, is uniform in the cable. We can thus determine the current per unit area (i.e. the current density) that flows through the whole cable, and use that to determine how much current flows through the surface with area πr^2 that is defined by the Amperian loop:

$$J = \frac{I}{A} = \frac{I}{\pi R^2} \tag{10.2.68}$$

And by definition:

$$I_{\text{enc}} = J\left(\pi r^2\right) = \frac{I}{\pi R^2}\left(\pi r^2\right) = I\frac{r^2}{R^2} \qquad (10.2.69)$$

Finally, we can apply Ampère's Law to determine the magnitude of the magnetic field inside the cable:

$$\oint \vec{B}\cdot d\vec{l} = \mu_0 I_{\text{enc}} \qquad (10.2.70)$$

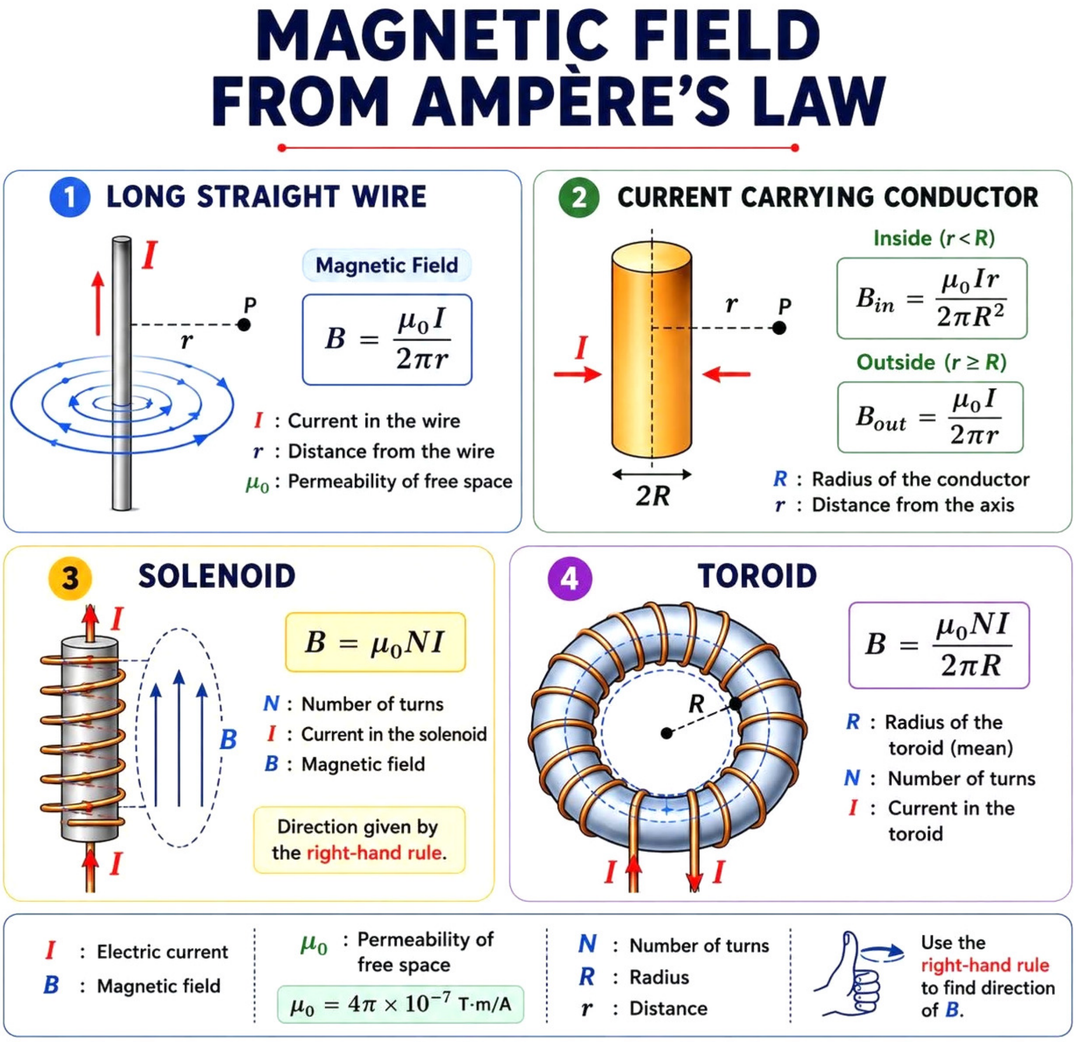
Therefore:

$$B(2\pi r) = \mu_0 I \frac{r^2}{R^2} \qquad (10.2.71)$$

Hence:

$$B = \frac{\mu_0 I}{2\pi R^2} r \qquad (10.2.72)$$

and we find that the magnetic field is zero at the center of the cable $(r = 0)$, and increases linearly up to the edge of the cable $(r = R)$.

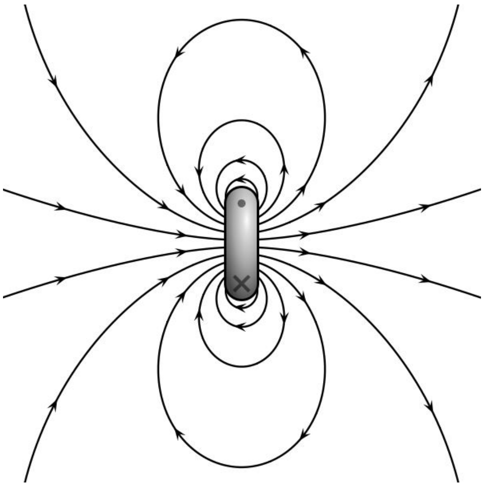

MAGNETIC FIELD
FROM AMPÈRE'S LAW
1 LONG STRAIGHT WIRE
I
P
r
Magnetic Field
$B = \frac{\mu_0 I}{2\pi r}$
I : Current in the wire
r : Distance from the wire
μ_0 : Permeability of free space
2 CURRENT CARRYING CONDUCTOR
I
r
P
2R
Inside (r < R)
$B_{in} = \frac{\mu_0 I r}{2\pi R^2}$
Outside (r ≥ R)
$B_{out} = \frac{\mu_0 I}{2\pi r}$
R : Radius of the conductor
r : Distance from the axis
3 SOLENOID
I
B
I
$B = \mu_0 N I$
N : Number of turns
I : Current in the solenoid
B : Magnetic field
Direction given by the right-hand rule.
4 TOROID
R
I
I
$B = \frac{\mu_0 N I}{2\pi R}$
R : Radius of the toroid (mean)
N : Number of turns
I : Current in the toroid
I : Electric current
B : Magnetic field
μ_0 : Permeability of free space
$\mu_0 = 4\pi \times 10^{-7}$ T·m/A
N : Number of turns
R : Radius
r : Distance
Use the right-hand rule to find direction of B.

10.2.4 Magnetic dipole

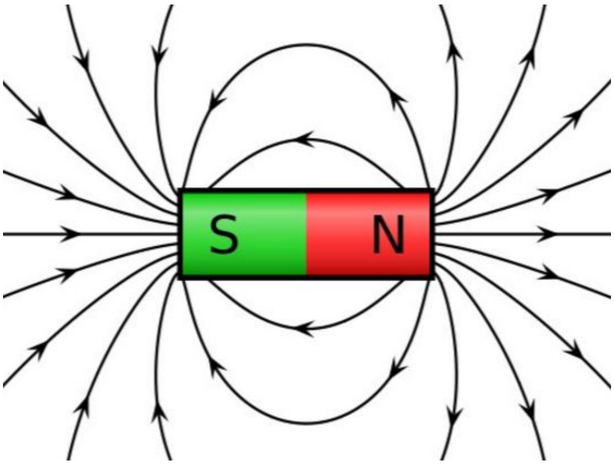
The magnetic dipole has as its electrostatic counterpart enormous importance in the study of magnetic properties of materials for which it allows to develop good theoretical models.

Before reading what follows, we strongly advise the reader (this even more than an advice) to read absolutely everything about the development of the rigid electric dipole moment in the section of Electrostatic. Indeed, most calculations that follow have the same reasoning, mathematical developments and approximations to some tiny nuances. We do not therefore wish to make the same intermediate calculations already present when calculating the electric dipole moment (however, if you really have trouble on reading what follow, we are ready to complete the missing as always in this textbook).

A magnetic dipole is the limit of either a closed loop of electric current or a pair of poles as the dimensions of the source are reduced to zero while keeping the magnetic moment constant.



(a) Magnetic dipole as a current loop (source: Wikipedia)



(b) Magnetic dipole as a magnet

Figure 10.39: Microscopic dipoles point of views

It is a magnetic analogue of the electric dipole, but the analogy is not complete. In particular, a magnetic monopole, the magnetic analogue of an electric charge, has never been observed. Moreover, one form of magnetic dipole moment is associated with a fundamental quantum property: the spin of elementary particles.

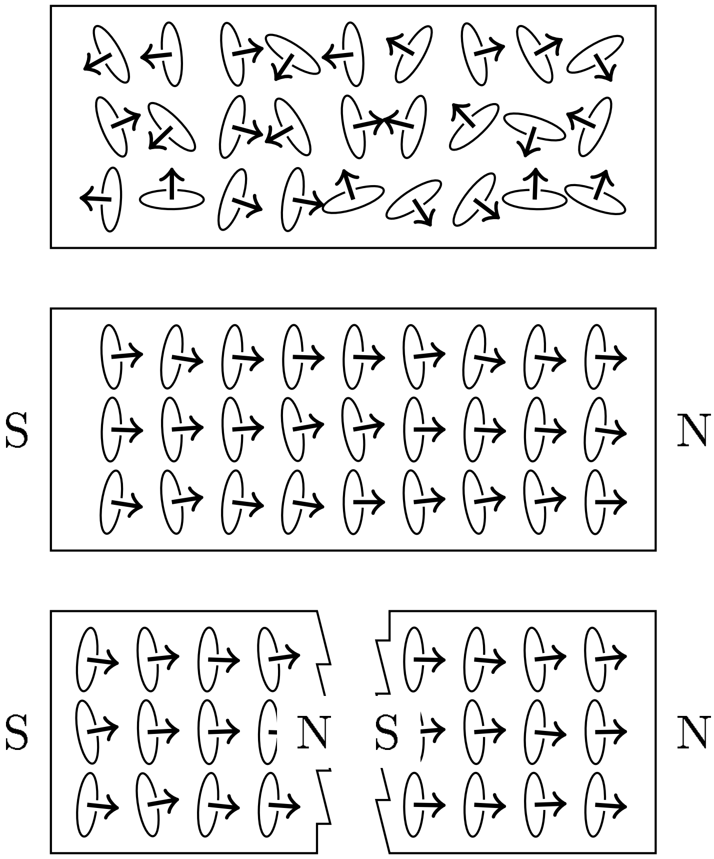


Figure 10.40: **Magnetic dipole (magnet) broken with small spin naively highlighted (author: Jimmy Roussel)**

Some planets as the Earth have a magnetic field quite similar to a magnetic dipole (but quite not homogeneous and not static):

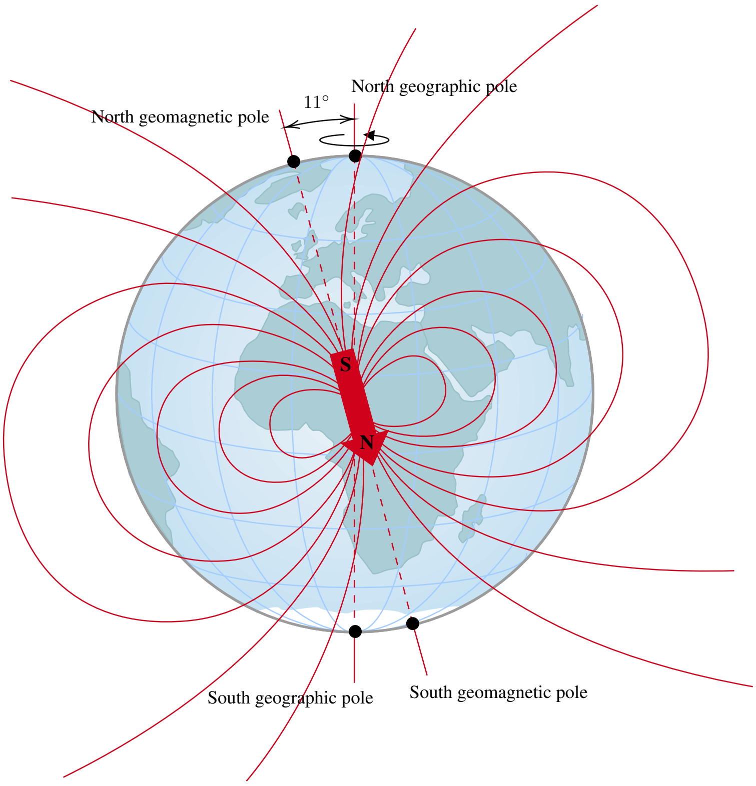


Figure 10.41: Earth viewed as an ideal magnetic dipole

The magnetic dipole, however, has a significant difference in relation to the practical case we impose us as part of its study ... there are no two electric charges! Indeed, electric charges at rest emit in a first approximation (this is experimental and theoretical fact...) an intrinsic magnetic field far too low to be considered interesting in the context of the study of the magnetic properties of materials. However, we should clarify something interesting (very cool), the elementary Coulomb electric charges are modelled sometimes (wrongly!) by physicists as a sphere rotating on itself (the "spin") and are represented as a superposition of small circular loops (oh... a loop we already know something like this) that are infinitely small so an observer in the rest frame (center of the electric charge) can interpret the overall Coulomb electric charge as a current moving in the different current loops, thereby inducing an intrinsic magnetic field (not pretty!?).

In the image below we can see a "standard magnetic compass" and a "dip circle" that differ in purpose and orientation. A magnetic compass is designed to indicate horizontal direction by aligning its needle with the horizontal component of Earth's magnetic field, making it ideal for navigation. In contrast, a dip circle measures the angle of dip (magnetic inclination), which is the angle between the magnetic field and the horizontal plane. Its needle pivots in a vertical plane, allowing it to study the vertical component of Earth's magnetic field, often used in geomagnetic research. Essentially, the compass focuses on direction, while the dip circle explores magnetic field inclination.

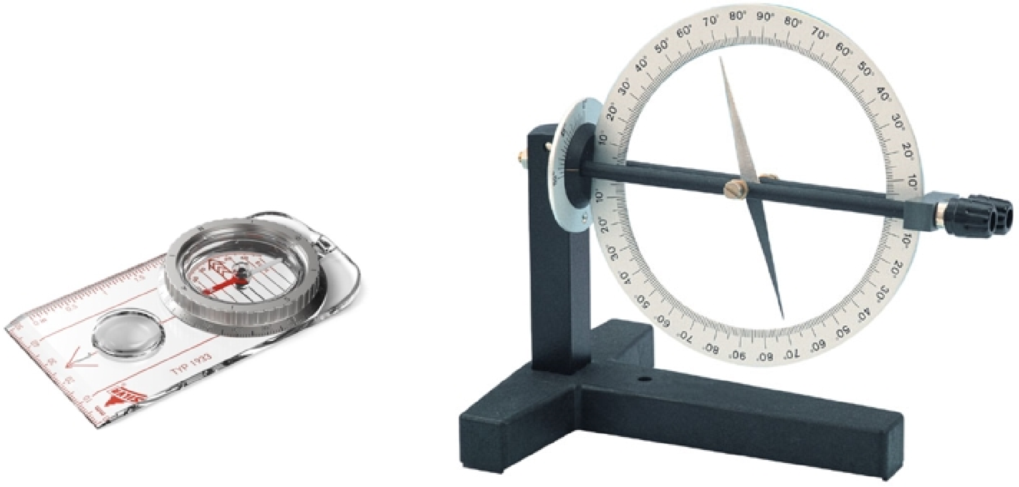
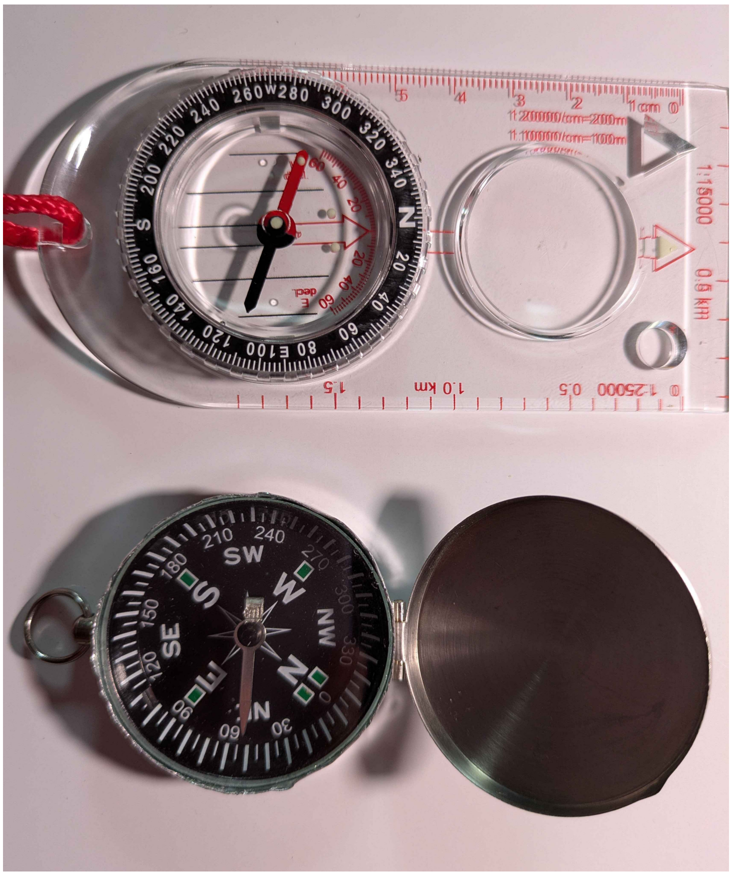


Figure 10.42: **Standard magnetic compass and Dip circle**

And below we have side by side a classic compass with its complete needle which respectively indicates North and South and a compass whose needle has been cut (it remains however a magnetic dipole) and which aligns with the South but with a little less accuracy than the compass above since its center of moment of force has been moved outside its support.



Remark

Nothing special visible to the human eye happens to a compass at vicinity of the North or South magnetic pole of our planet. Indeed, a magnetic field that is perfectly homogeneous and perfectly perpendicular to the ground (or perfectly horizontal) exists only in laboratories in a well constraint and controlled environment. On the North and South magnetic poles of our planet there will therefore still be a horizontal component that will make the compass work as usual.

In short, let us consider a flat loop (help... another loop again...), of any shape, center on O, through which travel a permanent and constant current I and where one of the points of the loop is denoted by P. We want to calculate the magnetic field generated by this coil at any point M in space, located at a distance r of the loop (specifically, at large distances compared to the size of the loop).

Remark

Personally there are some steps in the calculation that I find... well... very far unconvincing... but... there are so many approximations anyway that these problematic steps can bee see as details only... hummm...

We put:

$$\vec{r}=\overrightarrow{\text{O}M} \quad \vec{r}\,'=\overrightarrow{PM} \quad \vec{\rho}=\overrightarrow{\text{O}P}=\vec{r}-\vec{r}\,' \quad \vec{u}=\frac{\vec{r}}{r} \tag{10.2.73}$$

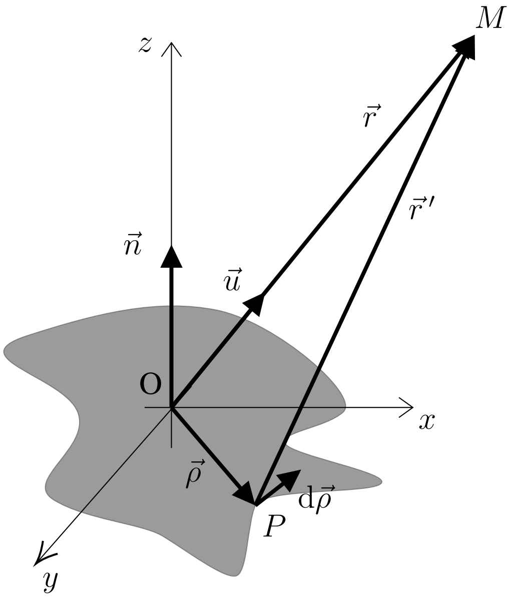


Figure 10.43: **Magnetic dipole study configuration**

So we will use the Biot-Savart law:

$$\vec{B}=\frac{\mu_0}{4\pi}\oint\frac{\vec{J}\times\vec{r}}{r^3}\mathrm{d}V \tag{10.2.74}$$

under the assumption that the point M is located at great distance from the loop. What gives us the right to write:

$$\vec{B}(M)\cong\frac{\mu_0 I}{4\pi}\oint\frac{\mathrm{d}\vec{l}\times\vec{r}'}{r^3}\mathrm{d}V \tag{10.2.75}$$

But as $\mathrm{d}\vec{l}=\mathrm{d}\vec{\rho}$ therefore:

$$\vec{B}(M)\cong\frac{\mu_0 I}{4\pi}\oint\frac{\mathrm{d}\vec{\rho}\times\vec{r}'}{r^3}\mathrm{d}V \tag{10.2.76}$$

Let us evaluate the term $\vec{r}'/r^3$ for points M located far away from the current loop (at the denominator we used the cosine theorem as during our study of the rigid electric dipole moment in the section Electrostatics):

$$\begin{aligned}\frac{\vec{r}'}{r^3}&=\frac{\vec{r}-\vec{\rho}}{(r^2+\rho^2-2\vec{r}\circ\vec{\rho})^{3/2}}\cong\frac{\vec{r}-\vec{\rho}}{r^3\left(1-2\dfrac{\vec{r}\circ\vec{\rho}}{r^2}\right)^{3/2}}\cong\frac{\vec{r}-\vec{\rho}}{r^2}\left(1+3\frac{\vec{r}\circ\vec{\rho}}{r^2}\right)\\&\cong\frac{\vec{u}}{r^2}-\frac{\vec{\rho}}{r^3}+3\frac{\vec{u}\circ\vec{\rho}}{r^3}\vec{u}\end{aligned} \tag{10.2.77}$$

where we did like the rigid electric dipole moment a Taylor limited development to order 1.

Remark

The last approximation is very rough in the sense that it is a clever choice of the terms to neglect to achieve a visually aesthetic result and to define the magnetic dipole moment (see further below) ...

Putting this expression into the Biot-Savart law, we get:

$$\vec{B}(M)\cong\frac{\mu_0 I}{4\pi r^2}\left[\oint\mathrm{d}\vec{\rho}\times\vec{u}-\frac{1}{r}\oint\mathrm{d}\vec{\rho}\times\vec{\rho}+\frac{3}{r}\oint\mathrm{d}\vec{\rho}\times\vec{u}(\vec{u}\circ\vec{\rho})\right] \tag{10.2.78}$$

Let us evaluate each term involved in the parenthesis separately:

1. $\displaystyle\oint\mathrm{d}\vec{\rho}\times\vec{u}=\left(\oint\mathrm{d}\vec{\rho}\right)\times\vec{u}=(\vec{\rho}(P_0)-\vec{\rho}(P_0))\times\vec{u}=\vec{0}\times\vec{u}=\vec{0}$

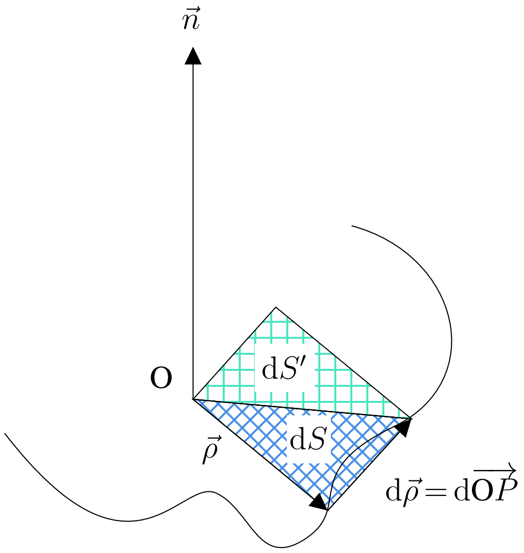
   since the vector $\vec{u}$ is independent of the point P on the current loop and we do a curvilinear integration along the entire turn, returning to the starting point.

2. $\displaystyle -\frac{1}{r}\oint\mathrm{d}\vec{\rho}\times\vec{\rho}$

   By the properties of the vector product:

   $$-\frac{1}{r}\oint\mathrm{d}\vec{\rho}\times\vec{\rho}=\frac{1}{r}\oint\vec{\rho}\times\mathrm{d}\vec{\rho} \tag{10.2.79}$$

   But since $\mathrm{d}\vec{\rho}$ and $\vec{\rho}$ are perpendicular and in a same plane, we have $\vec{\rho}\times\mathrm{d}\vec{\rho}$ which is the infinitesimal surface $\mathrm{d}S'$ of a rectangle and it represents nothing as the abscissa is curved relatively O. Indeed:


$\vec{n}$
O
$\mathrm{d}S'$
$\mathrm{d}S$
$\vec{\rho}$
$\mathrm{d}\vec{\rho}=\mathrm{d}\overrightarrow{\mathrm{O}P}$

So we can write:

$$\vec{\rho}\times\mathrm{d}\vec{\rho}=\mathrm{d}S\vec{n}=2\mathrm{d}S\vec{n}\Leftrightarrow\vec{\rho}\times\mathrm{d}\vec{\rho}=2\mathrm{d}S\vec{n} \tag{10.2.80}$$

where $\vec{n}$ is the vector normal to the plane of the loop (base vector of the z-axis). This result is generally valid regardless of the surface.

Hence:

$$-\frac{1}{r}\oint\mathrm{d}\vec{\rho}\times\vec{\rho}=\frac{1}{r}\oint\vec{\rho}\times\mathrm{d}\vec{\rho}=\frac{1}{r}\oint 2\mathrm{d}S\vec{n}=\frac{1}{r}\vec{n}2\oint\mathrm{d}S=\frac{2S}{r}\vec{n} \tag{10.2.81}$$

3. $\oint\mathrm{d}\vec{\rho}\times\vec{u}(\vec{u}\circ\vec{\rho})=-\vec{u}\times\oint\mathrm{d}\vec{\rho}(\vec{u}\circ\vec{\rho})$ by the properties of the cross product (see section Vector Calculus page 1830).

We will use these relations to calculate the unknown integral of the start of our study. If we decompose the vector $\vec{\rho}$ and $\vec{u}$ in the base $(\vec{e}_1,\vec{e}_2)$ generating the plane of the loop, we get:

$$\mathrm{d}\vec{\rho}(\vec{u}\circ\vec{\rho})=\mathrm{d}\rho_1(\rho_1u_1+\rho_2u_2)\vec{e}_1+\mathrm{d}\rho_2(\rho_1u_1+\rho_2u_2)\vec{e}_2 \tag{10.2.82}$$

as $\mathrm{d}\rho_3=0$ and $\rho_3=0$.

We also have:

$$\begin{gathered}\oint\rho_1u_1\mathrm{d}\rho_1=\frac{u_1}{2}(\rho_1^2(P_0)-\rho_1^2(P_0))=0\\ \oint\rho_2u_2\mathrm{d}\rho_2=\frac{u_2}{2}(\rho_2^2(P_0)-\rho_2^2(P_0))=0\end{gathered} \tag{10.2.83}$$

Hence:

$$\oint\mathrm{d}\vec{\rho}(\vec{u}\circ\vec{\rho})=u_2\oint\rho_2\mathrm{d}\rho_1\vec{e}_1+u_1\oint\rho_1\mathrm{d}\rho_2\vec{e}_2=u_2\vec{e}_1\oint\rho_2\mathrm{d}\rho_1+u_1\vec{e}_2\oint\rho_1\mathrm{d}\rho_2 \tag{10.2.84}$$

Let us recall that:

$$\vec{\rho}\times\mathrm{d}\vec{\rho}=\mathrm{d}S'\vec{n}=2\mathrm{d}S\vec{n}\Leftrightarrow\frac{\vec{\rho}\times\mathrm{d}\vec{\rho}}{2}=\mathrm{d}S\vec{n} \tag{10.2.85}$$

In the form of components (only the third term is non-zero since $\mathrm{d}\rho_3=0$ and $\rho_3=0$), we have:

$$\frac{1}{2}(\rho_1\mathrm{d}\rho_2-\rho_2\mathrm{d}\rho_1)=\mathrm{d}S \tag{10.2.86}$$

hence:

$$\rho_1\mathrm{d}\rho_2-\rho_2\mathrm{d}\rho_1=2\mathrm{d}S\Leftrightarrow\rho_1\mathrm{d}\rho_2=\mathrm{d}S\qquad\rho_2\mathrm{d}\rho_1=-\mathrm{d}S \tag{10.2.87}$$

Which brings us to write:

$$\oint\mathrm{d}\vec{\rho}(\vec{u}\circ\vec{\rho})=u_2\vec{e}_1\oint\rho_2\mathrm{d}\rho_1+u_1\vec{e}_2\oint\rho_1\mathrm{d}\rho_2=-u_2\vec{e}_1\oint\mathrm{d}S+u_1\vec{e}_2\oint\mathrm{d}S \tag{10.2.88}$$

Therefore:

$$-(u_2\vec{e}_1S-u_1\vec{e}_2S)=-S(u_2\vec{e}_1-u_1\vec{e}_2)=-S\left(u_2\begin{pmatrix}1\\0\\0\end{pmatrix}-u_1\begin{pmatrix}0\\1\\0\end{pmatrix}\right)=S\begin{pmatrix}-u_2\\u_1\\0\end{pmatrix} \tag{10.2.89}$$

We notice that the latter relation is equal to:

$$S(\vec{n}\times\vec{u})=S\begin{pmatrix}0\\0\\1\end{pmatrix}\times\begin{pmatrix}u_1\\u_2\\u_3\end{pmatrix}=S\begin{pmatrix}-1\cdot u_2\\1\cdot u_1\\0\end{pmatrix}=S\begin{pmatrix}-u_2\\u_1\\0\end{pmatrix} \tag{10.2.90}$$

Then finally:

$$\oint\mathrm{d}\vec{\rho}(\vec{u}\circ\vec{\rho})=S\vec{n}\times\vec{u} \tag{10.2.91}$$

By bringing together these results we obtain for the magnetic field:

$$\vec{B}(M)\cong\frac{\mu_0I}{4\pi r^2}\left[\frac{2S}{r}\vec{n}-\frac{3\vec{u}}{r}\times(S\vec{n}\times\vec{u})\right] \tag{10.2.92}$$

We see therefore an important quantity that appears because it fully describes the spire seen from a great distance, namely the "magnetic local dipole moment" or simply "magnetic dipole moment":

$$\boxed{\vec{\mu}_l=IS\vec{n}} \tag{10.2.93}$$

also often denoted by $\vec{\mathcal{M}}$ by some practitioners. The magnetic moment has then for units $[\mathrm{A}\cdot\mathrm{m}^2]=[\mathrm{J}\cdot\mathrm{T}^{-1}]=[\mathrm{N}\cdot\mathrm{m}\cdot\mathrm{T}^{-1}]$.

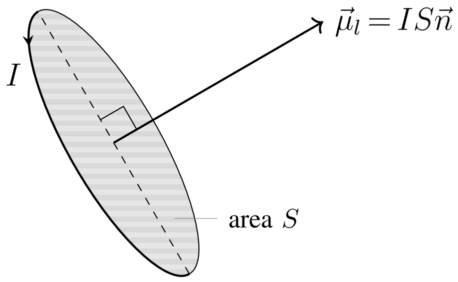


Figure 10.44: **Magnetic dipole**

We then have the prior-previous relation that can be written:

$$\vec{B}(M)\cong\frac{\mu_0}{4\pi r^2}\left[\frac{2\vec{\mu}_l}{r}-\frac{3\vec{u}}{r}\times(\vec{\mu}_l\times\vec{u})\right]\cong\frac{\mu_0}{4\pi}\left[\frac{2\vec{\mu}_l}{r^3}-\frac{3\vec{u}}{r^3}\times(\vec{\mu}_l\times\vec{u})\right] \tag{10.2.94}$$

By making use of the property of the double vector product (see section Vector Calculus page 1835):

$$\vec{u}\times(\vec{\mu}_l\times\vec{u})=(\vec{u}\circ\vec{u})\vec{\mu}_l-(\vec{u}\circ\vec{\mu}_l)\vec{u}=\vec{\mu}_l(\vec{u}\circ\vec{u})-\vec{u}(\vec{u}\circ\vec{\mu}_l) \tag{10.2.95}$$

We then get another form of expression of the approximate magnetic field generated by a dipole:

$$\begin{aligned}\vec{B}(M)&\cong\frac{\mu_0}{4\pi}\left[\frac{2\vec{\mu}_l}{r^3}-\frac{3\vec{u}}{r^3}\times(\vec{\mu}_l\times\vec{u})\right]\cong\frac{\mu_0}{4\pi r^3}\left[2\vec{\mu}_l-3(\vec{\mu}_l(\vec{u}\circ\vec{u})-\vec{u}(\vec{u}\circ\vec{\mu}_l))\right]\\&=\frac{\mu_0}{4\pi r^3}\left[2\vec{\mu}_l-3\vec{\mu}_l\underbrace{(\vec{u}\circ\vec{u})}_{=1}+3\vec{u}(\vec{u}\circ\vec{\mu}_l))\right]=\frac{\mu_0}{4\pi r^3}\left[3\vec{u}(\vec{u}\circ\vec{\mu}_l)-\vec{\mu}_l\right]\end{aligned} \tag{10.2.96}$$

It is in the latter form that we found the most often the expression of the magnetic moment in the literature.

This relation has to be compared (for fun) with the expression for the electric field for a rigid electric dipole:

$$\boxed{\begin{aligned}\vec{B}&\cong\frac{\mu_0}{4\pi r^3}\left[3\vec{u}(\vec{u}\circ\vec{\mu}_l))-\vec{\mu}_l\right]\\\vec{E}&\cong\frac{1}{4\pi\varepsilon r^3}\left[3\vec{u}_r(\vec{u}_r\circ\vec{p}))-\vec{p}\right]\end{aligned}} \tag{10.2.97}$$

and thus we see that there is perfect correspondence.

We still arrived to put this in a fairly identical and aesthetics and form after some approximations...

We have also:

$$\vec{\mu}_l\circ\vec{u}=\|\vec{\mu}_l\|\cdot\|\vec{u}\|\cos(\theta)=\mu_0\cos(\theta) \tag{10.2.98}$$

Therefore:

$$\vec{B}(M)\cong\frac{\mu_0}{4\pi r^3}\left[3\vec{u}\mu_l\cos(\theta)-\vec{\mu}_l\right] \tag{10.2.99}$$

Remember that the magnetic field due to a magnetic dipole moment, $\vec{\mu}_l$ at a point $\vec{R}$ relative to it may be written:

$$\vec{B}(r)\cong\frac{\mu_0}{4\pi r^3}\left[3\vec{u}\,(\vec{u}\circ\vec{\mu}_l)-\vec{\mu}_l\right] \tag{10.2.100}$$

where μ_0 is the vacuum permeability. In geomagnetism, it is usual to write the radial and angular components of $\vec{B}$ as:

$$\begin{aligned}B_r&=-2B_0\left(\frac{R_\mathrm{E}}{r}\right)^3\cos(\theta)\\B_\theta&=-B_0\left(\frac{R_\mathrm{E}}{r}\right)^3\sin(\theta)\\B_\phi&=0\end{aligned} \tag{10.2.101}$$

where θ is polar (colatitude) angle (relative to the magnetic North pole), ϕ is the azimuthal angle (longitude), and R_E is the Earth's radius, about $6,370$ [km].

Indeed consider the dipole to lie along the y-axis: $\vec{\mu}_l = \mu_l \vec{\jmath}$ and define $\vec{B}_0$ to be the field perpendicular to $\vec{\mu}_l$ at a distance R_E from it (i.e. on the Earth's magnetic equator):

$$\vec{B}_0 = -\frac{\mu_0 \mu_l}{4\pi R_\text{E}^3}\vec{\jmath} = B_0 \vec{\jmath} \tag{10.2.102}$$

Hence:

$$\vec{\mu}_l = \mu_l \vec{\jmath} = \frac{\vec{B}_0 4\pi R_\text{E}^2}{\mu_0} \tag{10.2.103}$$

Therefore:

$$\vec{B}(r) \cong \frac{\mu_0}{4\pi r^3}\left[3\vec{u}\left(\vec{u}\circ\vec{\mu}_l\right) - \vec{\mu}_l\right] = -B_0\left(\frac{R_\text{E}}{r}\right)^3 \left[3\cos(\theta)\vec{u} - \vec{\jmath}\right] \tag{10.2.104}$$

The radial component of the magnetic field is therefore:

$$B_r = \vec{B}(r)\circ\hat{\vec{r}} = -B_0\left(\frac{R_\text{E}}{r}\right)^3 \left[3\cos(\theta) - \cos(\theta)\right] = -2B_0\left(\frac{R_\text{E}}{r}\right)^3 \cos\theta \tag{10.2.105}$$

Its angular component is perpendicular to this:

$$B_\theta = \vec{B}\cdot\hat{\vec{\theta}} = -B_0\left(\frac{R_\text{E}}{r}\right)^3 \left[-\vec{\jmath}\circ\hat{\vec{\theta}}\right] = -B_0\left(\frac{R_\text{E}}{r}\right)^3 \sin(\theta) \tag{10.2.106}$$

Remark

As we can see it these are not the polar equations of a section of a (regular) torus! Therefore it is wrong to say that the field lines of a magnetic dipole are shaped like a "torus". The conventional terminology is to say that it is shaped liked a "dipolar geometry".

With this definition, B_0 denotes the magnitude of the mean value of the field at the magnetic equator on the Earth's surface, about $31.2\,[\mu\text{T}]$. With these definitions, we can plot the dipole component of the Earth's magnetic field as a Python™ Matplotlib streamplot. We first construct a meshgrid of (x, y) coordinates and convert them into polar coordinates with the relations:

$$\begin{aligned} r &= \sqrt{x^2+y^2} \\ \theta &= \text{atan}\,2\left(\frac{y}{x}\right) \end{aligned} \tag{10.2.107}$$

We can then use the relation above to calculate B_r and B_θ and convert back to Cartesian coordinates to plot the streamplot. The θ coordinate is offset by $\alpha = 9.6^\circ$ to account for the current tilt of the Earth's magnetic dipole with respect to its rotational axis.

Here is the Python™ code from Christian Hill (physicist and chemist) taken from `https://scipython.com/blog/visualizing-the-earths-magnetic-field/`:

```python
import sys
import numpy as np
import matplotlib.pyplot as plt
from matplotlib.patches import Circle

# Mean magnitude of the Earth's magnetic field at the equator in T
B0 = 3.12e-5
# Radius of Earth, Mm (10^6 m: mega-metres!)
RE = 6.370
# Deviation of magnetic pole from axis
alpha = np.radians(9.6)

def B(r, theta):
    """Return the magnetic field vector at (r, theta)."""
    fac = B0 * (RE / r)**3
    return -2 * fac * np.cos(theta + alpha), -fac * np.sin(theta + alpha)

# Grid of x, y points on a Cartesian grid
nx, ny = 64, 64
XMAX, YMAX = 40, 40
x = np.linspace(-XMAX, XMAX, nx)
y = np.linspace(-YMAX, YMAX, ny)
X, Y = np.meshgrid(x, y)
r, theta = np.hypot(X, Y), np.arctan2(Y, X)

# Magnetic field vector, B = (Ex, Ey), as separate components
Br, Btheta = B(r, theta)
# Transform to Cartesian coordinates: NB make North point up, not to the right.
c, s = np.cos(np.pi/2 + theta), np.sin(np.pi/2 + theta)
Bx = -Btheta * s + Br * c
By = Btheta * c + Br * s

fig, ax = plt.subplots()

# Plot the streamlines with an appropriate colormap and arrow style
color = 2 * np.log(np.hypot(Bx, By))
ax.streamplot(x, y, Bx, By, color=color, linewidth=1, cmap=plt.cm.inferno,
              density=2, arrowstyle='->', arrowsize=1.5)

# Add a filled circle for the Earth; make sure it's on top of the streamlines.
ax.add_patch(Circle((0,0), RE, color='b', zorder=100))

ax.set_xlabel('$x$')
ax.set_ylabel('$y$')
ax.set_xlim(-XMAX, XMAX)
ax.set_ylim(-YMAX, YMAX)
ax.set_aspect('equal')
```

```
plt.show()
```

That gives:

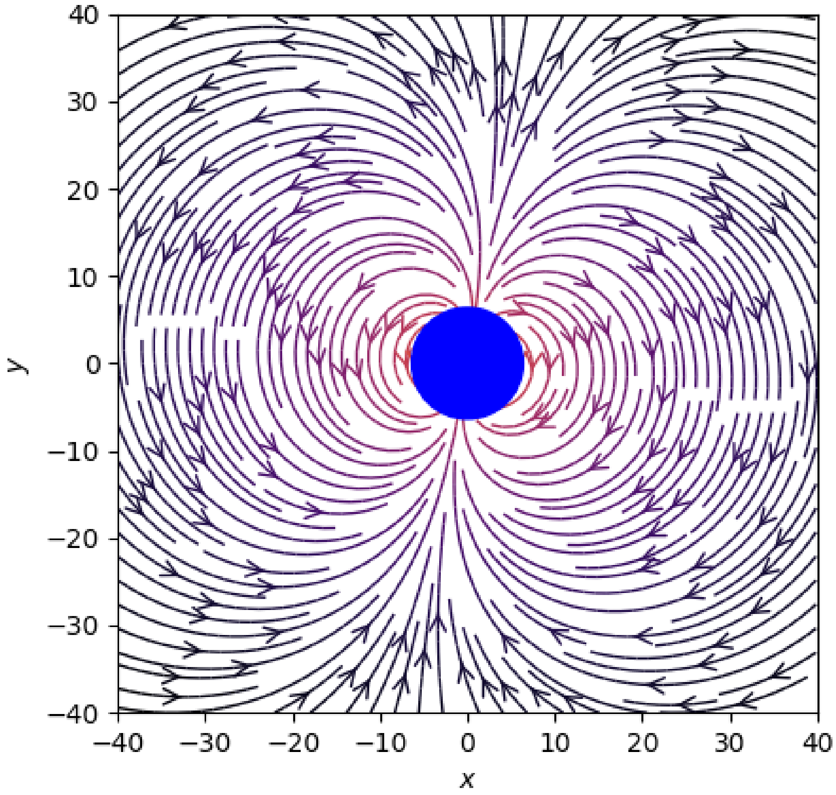


Figure 10.45: **Visualizing dipole magnetic field with Python™ (author: Christian Hill)**

The origin of the magnetic field of any material must be microscopic. Using the Bohr model of the atom (see section Corpuscular Quantum Physics page 4626) we can convince ourselves that atoms (at least some) have an intrinsic magnetic dipole moment. Indeed, the Bohr model of the Hydrogen atom consists of an electron of electric charge $q=-e$ in circular motion around a central nucleus (a proton with an electric charge $q=+e$) with a period $T=2\pi\omega$.

If we look over long time scales compared to T, it is as if there was an electric current:

$$I=\frac{q}{T}=\frac{q\omega}{2\pi} \tag{10.2.108}$$

So we have indeed a sort of circular loop, of mean radius equal to the average distance to the proton, i.e. the Bohr radius r_0 (see section Corpuscular Quantum Physics page 4630). The Hydrogen atom therefore have an intrinsic magnetic moment of:

$$\boxed{\vec{\mu}_l=IS\vec{n}=\frac{q\omega}{2\pi}\pi r_0^2\vec{n}=\frac{q\omega}{2}r_0^2\vec{n}=\frac{q}{2m}m\omega r_0^2\vec{n}=\frac{q}{2m}\vec{b}} \tag{10.2.109}$$

where $\vec{b}$ is the momentum of the electron and $q/2m$ the "gyromagnetic factor" (this result is very important for the Langevin model of diamagnetism!).

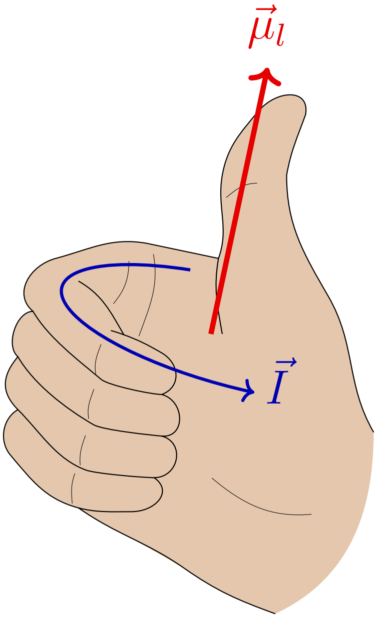


Figure 10.46: **Magnetic moment right-handed rule (author: Izaak Neutelings)**

This reasoning can be generalized to other atoms. Indeed, a set of electric charges in rotation about an axis will produce a magnetic moment proportional to the total angular momentum. This happens even if the total electric charge is zero (material or neutral atom): what matters is the (scalar) existence of a current.

Remark

We deduce then immediately that for a non-conductive ring of radius R having a linear charge density ρ, turning at a rate $\omega=2\pi/T$ we then have:

$$\vec{\mu}_l=IS\vec{n}=\frac{\rho 2\pi R}{T}\pi R^2\vec{n}=\frac{Q}{2\pi}\omega R^2\vec{n} \tag{10.2.110}$$

So this is the magnetic moment of a rotating non-conductive ring!

So in fact, we can qualitatively explain the magnetic properties of materials based on the orientation of the magnetic moments of the atoms that compose them:

- "Non-magnetic materials": These are materials where the magnetic moments are randomly distributed, there is no intrinsic magnetic field.
- "Diamagnetic materials": these are the materials that subjected to a magnetic field, have their magnetic moments that opposed to it and are (marginally) repelled by magnets. They therefore induce a magnetic moment opposite to the magnetic field direction.
- "Paramagnetic materials": it is the materials for which moments can be oriented in the direction of an external magnetic field and can therefore be magnetized so (attracted) momentarily. They therefore induce a magnetic moment in the direction of the magnetic field.

- "Ferromagnetic materials": are materials whose moments are already oriented in a particular direction, permanently (natural magnets).

Remark

Earth is known to have a dipolar magnetic field where the magnetic north pole corresponds to the geographic South Pole (at a given angle value). At the macroscopic level, the explanation for the existence of the magnetic field observed on the stars is still far from satisfactory. The theory of "dynamo effect" trying to explain the fields observed by the presence of currents, mainly azimuthal, in the core of stars. Several known facts remain partially unsolved:

- The magnetic cycles: the Sun (actually a G2V-type star, aka yellow dwarf) has a large-scale magnetic field similar to that of Earth, approximately dipolar. However, there is a polarity inversion every 11 years. For the Earth, geophysicists were able to show that there had been a reversal about $700,000$ years before.
- Non-alignment with the angular momentum of the celestial body: it is of the order of ten degrees for the Earth, it is perpendicular for Neptune!

Let us consider now a non-conducting disc of radius R, charge Q and uniform charge density σ is rotating about an axis passing through its center and perpendicular to its plane with an angular velocity ω. To determine it's magnetic moment (not a bad approximation of a non-relativistic Black Hole accretion disc), we divide the disk into many small rings of width $\mathrm{d}r$. The current produced by the ring is the total charge contained divided by the period for one rotation. Thus:

$$\mathrm{d}I=\frac{\sigma(2\pi r\mathrm{d}r)}{\omega/(2\pi)} \tag{10.2.111}$$

Factoring in the area of each ring gives

$$\mathrm{d}\mu=(\pi r^2)\frac{\sigma(2\pi r\mathrm{d}r)}{\omega/(2\pi)}=\sigma\omega\pi r^3\mathrm{d}r \tag{10.2.112}$$

Therefore:

$$\mu_l=\int\limits_0^R\sigma\omega\pi r^3\mathrm{d}r=\frac{1}{4}\sigma\omega\pi R^4 \tag{10.2.113}$$

Or in vector form we have then the magnetic moment of a rotating disc:

$$\boxed{\vec{\mu}_l=\frac{1}{4}\sigma\pi R^4\vec{\omega}} \tag{10.2.114}$$

Let us now consider a non-conductive sphere shell of radius R rotating a constant speed ω around the Z axis with a uniform surface electric charge σ

We split the sphere into thin segments (each corresponding to a ring), each one having for magnetic moment:

$$\mathrm{d}\mu_l=\mathrm{d}I\mathrm{d}S \tag{10.2.115}$$

where $\mathrm{d}S$ is the area enclosed by the segment (keep in mind that the projected radius is $r = R\sin(\theta)$):

$$\mathrm{d}S = 2\pi r\mathrm{d}r = 2\pi R\sin(\theta)\mathrm{d}r \tag{10.2.116}$$

The current $\mathrm{d}I$ on each segment is :

$$\mathrm{d}I = \frac{\mathrm{d}Q}{T} = \frac{\omega}{2\pi}\mathrm{d}Q \tag{10.2.117}$$

The charge on each segment is:

$$\mathrm{d}Q = \sigma\mathrm{d}s \tag{10.2.118}$$

Where $\mathrm{d}s$ is the surface area of the segment (see section Geometric Shapes page 2964):

$$\mathrm{d}s = r^2\sin(\theta)\mathrm{d}\theta\mathrm{d}\phi = R^2\sin^3(\theta)\mathrm{d}\theta\mathrm{d}\phi \tag{10.2.119}$$

This results in:

$$\mathrm{d}Q = \sigma R^2\sin^3(\theta)\mathrm{d}\theta\mathrm{d}\phi = \frac{q}{4\pi R^2}R^2\sin^3(\theta)\mathrm{d}\theta\mathrm{d}\phi \tag{10.2.120}$$

We now have an expression for $\mathrm{d}I$:

$$\mathrm{d}I = \frac{\omega QR^2\sin^3(\theta)}{8\pi^2R^2}\mathrm{d}\theta\mathrm{d}\phi \tag{10.2.121}$$

Now we can calculate the moment of each segment:

$$\mathrm{d}\mu_l = \frac{\omega QR^2\sin^3(\theta)}{8\pi^2R^2}2\pi r\mathrm{d}r\mathrm{d}\phi\mathrm{d}\theta = \frac{\omega Q}{4\pi}\sin^3(\theta)r\mathrm{d}r\mathrm{d}\phi\mathrm{d}\theta \tag{10.2.122}$$

Integrating this over the whole sphere gives:

$$\begin{aligned}\mu_l &= \frac{\omega Q}{4\pi}\int_0^R r\mathrm{d}r\int_0^\pi \sin^3(\theta)\mathrm{d}\theta\int_0^{2\pi}\mathrm{d}\phi = \frac{\omega Q}{4\pi}\frac{1}{2}R^2 2\pi\int_0^\pi \sin^3(\theta)\mathrm{d}\theta = \frac{\omega QR^2}{4}\int_0^\pi \sin^3(\theta)\mathrm{d}\theta\\ &= \frac{\omega QR^2}{4}\frac{4}{3} = \frac{\omega QR^2}{3}\end{aligned} \tag{10.2.123}$$

Hence the magnetic moment for a spherical shell:

$$\boxed{\mu_l = \frac{\omega QR^2}{3}} \tag{10.2.124}$$

Remark

It seems that the result is the same for a ball instead than for a spherical shell. If it's true, it's a useful example for some special exotic astronomical objects.

To see a last pretty practical case of the magnetic dipole, remember that in the section Electrostatics, we had shown that in spherical coordinates we had for the electric dipole:

$$E_r = \frac{p\cos(\theta)}{2\pi\varepsilon r^3} \quad E_\theta = \frac{p\sin(\theta)}{4\pi\varepsilon r^3} \quad E_\phi = 0 \tag{10.2.125}$$

And as we have shown there is perfect correspondence between the magnetic and electric dipole moment, then we can immediately write:

$$\boxed{B_r=\frac{\mu_0\mu_l\cos(\theta)}{2\pi r^3}\qquad B_\theta=\frac{\mu_0\mu_l\sin(\theta)}{4\pi r^3}\qquad B_\phi=0} \tag{10.2.126}$$

We can then deduce the norm of the magnetic field vector:

$$\begin{aligned}B=\|\vec{B}\|&=\sqrt{B_r^2+B_\theta^2+B_\phi^2}=\sqrt{\left(\frac{\mu_0\mu_l\cos(\theta)}{2\pi r^3}\right)^2+\left(\frac{\mu_0\mu_l\sin(\theta)}{4\pi r^3}\right)^2}\\&=\frac{\mu_0\mu_l}{2\pi r^3}\sqrt{\cos^2(\theta)+\frac{\sin^2(\theta)}{4}}=\frac{\mu_0\mu_l}{2\pi r^3}\sqrt{\frac{1}{4}(4\cos^2(\theta)+\sin^2(\theta))}=\frac{\mu_0\mu_l}{4\pi r^3}\sqrt{4\cos^2(\theta)+\sin^2(\theta)}\\&=\frac{\mu_0\mu_l}{4\pi r^3}\sqrt{4\cos^2(\theta)+(1-\cos^2(\theta))}=\frac{\mu_0\mu_l\sqrt{3\cos^2(\theta)+1}}{4\pi r^3}=\frac{\mu_0\mu_l\sqrt{1+3\cos^2(\theta)}}{4\pi r^3}\end{aligned} \tag{10.2.127}$$

Therefore at the North Pole we have:

$$\begin{aligned}B_r(r,0,\phi)&=\frac{\mu_0\mu_l\cos(\theta)}{2\pi r^3}=\frac{\mu_0\mu_l}{2\pi r^3}\\B_\theta(r,0,\phi)&=\frac{\mu_0\mu_l\sin(\theta)}{4\pi r^3}=0\end{aligned} \tag{10.2.128}$$

and at the equator:

$$\begin{aligned}B_r\left(r,\frac{\pi}{2},\phi\right)&=\frac{\mu_0\mu_l\cos(\theta)}{2\pi r^3}=0\\B_\theta\left(r,\frac{\pi}{2},\phi\right)&=\frac{\mu_0\mu_l\sin(\theta)}{4\pi r^3}=\frac{\mu_0\mu_l}{4\pi r^3}\end{aligned} \tag{10.2.129}$$

The magnitude of the total field at the pole is twice as strong as at the equator (most people think that intuitively it is the opposite)!

We can also write from the previous results:

$$\boxed{\mu_l=\frac{4\pi r^3 B}{\mu_0\sqrt{1+3\cos^2(\theta)}}} \tag{10.2.130}$$

also sometimes denoted by the letter m (yes... its obvious to refer the first letter of "magnetism").

✍**Example:**

Thus, we can have fun to calculate the magnetic moment of the Earth at the geomagnetic equator. What gives, knowing that the Earth's geomagnetic field at equator is in average equal to 32 [μT] (Earth's magnetic field intensity is roughly between $25,000$

- $65,000$ [nT] (i.e. 0.25 to 65 [Gauss]):

$$\mu_l=\frac{4\pi r^3 B}{\mu_0\sqrt{1+3\cos^2(\theta)}}=\frac{4\pi(6371)^3 32\cdot 10^{-6}}{\mu\sqrt{1+3\cos^2\left(\dfrac{\pi}{2}\right)}}\cong 78\cdot 10^{21}\,[\mathrm{A}\cdot\mathrm{m}^2]=78\cdot 10^{21}\,[\mathrm{C}\cdot\mathrm{s}^{-1}\cdot\mathrm{m}^2] \tag{10.2.131}$$

For the fun we can convert that easily to how much mass of simply ionized moving iron it corresponds:

$$\begin{aligned}M_{\mathrm{Fe}}&\cong\frac{78\cdot 10^{21}}{e\cdot N_{\mathbf{Av}}}\cdot M_{\text{mol,Fe}}=\frac{78\cdot 10^{21}}{1.609\cdot 10^{-19}\cdot 6.022\cdot 10^{23}}\cdot 55.847\cdot 10^{-3}\\&\cong 4.51\cdot 10^{16}\,[\mathrm{Kg}]\end{aligned} \tag{10.2.132}$$

As measurements (Geophysical Journal International, Volume 84, Issue 3, March 1986, Pages 561–579, https://doi.org/10.1111/j.1365-246X.1986.tb04371.x) give that the density of iron at the inner core-outer core boundary pressure is close to $1,300$ $[\mathrm{kg}\cdot\mathrm{m}^{-3}]$, the latter result give an Iron quantity corresponding to a moving volume of:

$$V_{\mathrm{Fe}}=\frac{4.51\cdot 10^{16}}{1,300}\cong 3.47\cdot 10^{13}\,[\mathrm{m}^3] \tag{10.2.133}$$

Volume of simply ionized moving iron corresponding to a sphere of radius:

$$R_{\mathrm{Fe}}\cong\sqrt[3]{\frac{4}{3}\frac{3.47\cdot 10^{13}}{\pi}}\cong 20\,[\mathrm{km}] \tag{10.2.134}$$

So it's not even that big considering the dimensions of Earth in comparison and especially in comparison to the whole inner core that has a radius estimated thanks to seismology techniques at $1,220$ kilometres...!

We can also estimated the magnetic field intensity at the surface of Earth's core. As the radius of Earth's core based on seismology is estimated to be equal at $3,484$ [km] and Earth's radius $6,371$ [km] we then have knowing (as proved above) that the magnetic field of dipole decrease as $1/r^3$, we then have:

$$B_{\text{Earth,core}}\cong 32\cdot 10^{-6}\frac{(6371\cdot 10^3)^3}{(3484\cdot 10^3)^3}\cong 196\,[\mu\mathrm{T}]=0.192\,[\mathrm{mT}] \tag{10.2.135}$$

Hence the magnetic field at Earth's core surface is approximately 6 times stronger at Earth's surface. Numerical models (see [29] and [100]) and theoretical consideration (see [510]) suggest an internal magnetic field of 1 to 4 [mT].

For the Earth (that is far from a perfect solid dipole at the opposite of Neutron Star), measurements give for the magnetic field amplitude:

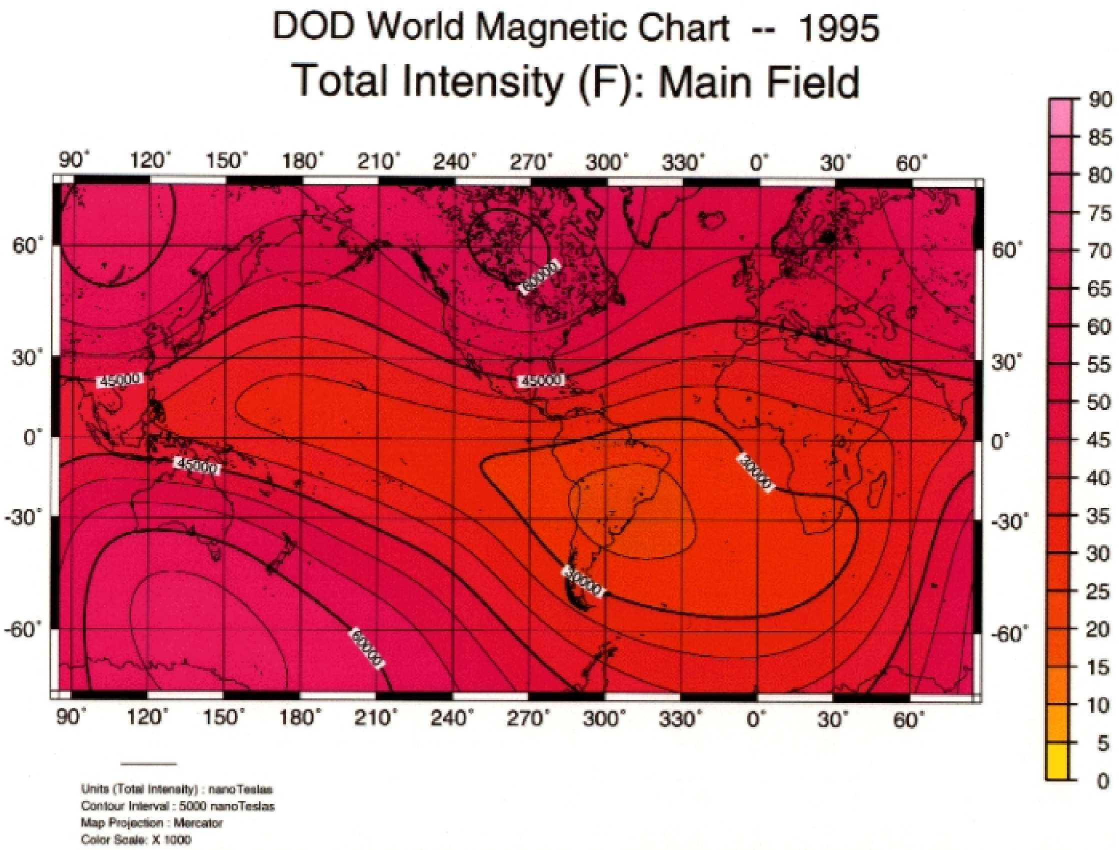
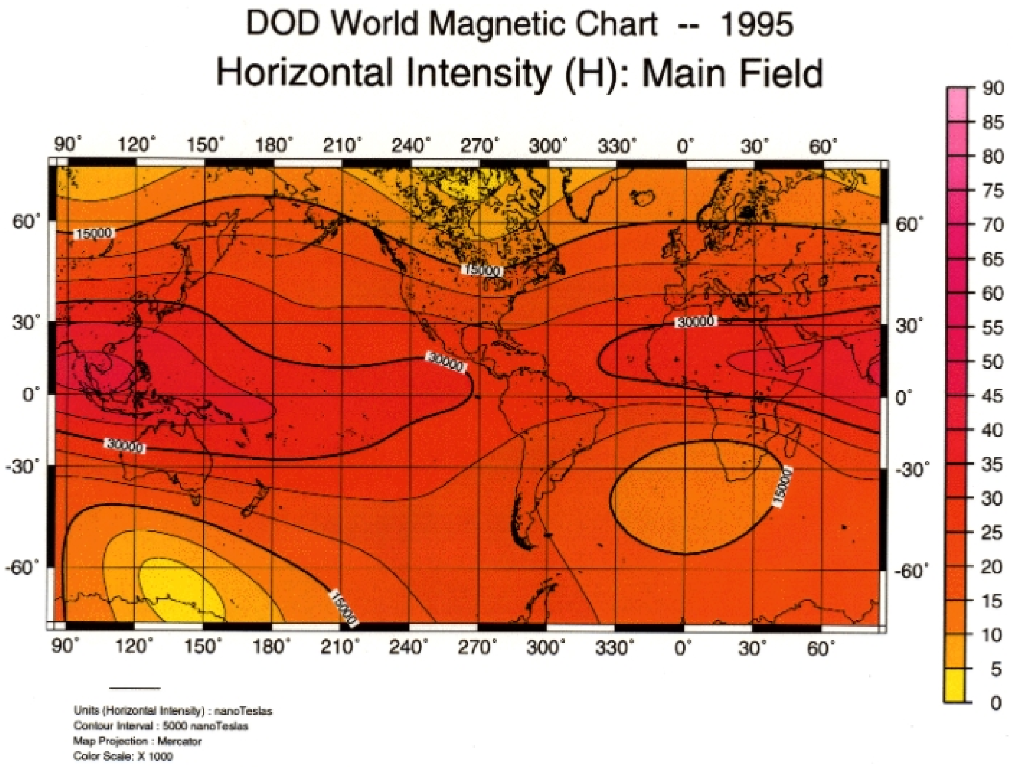
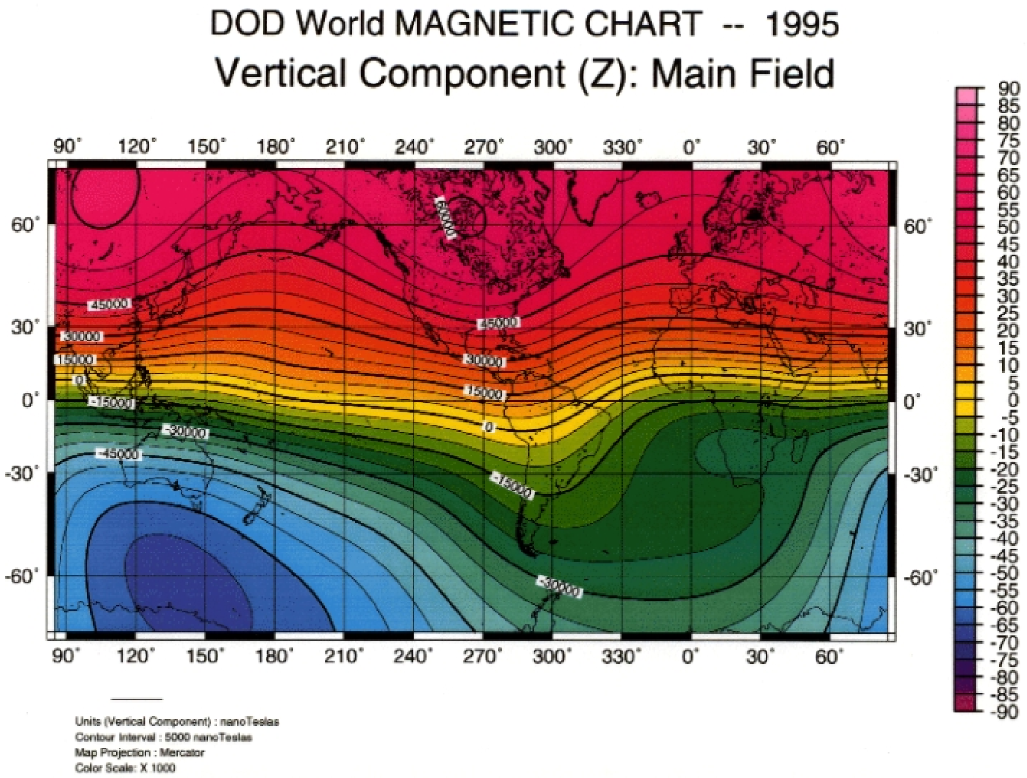

DOD World Magnetic Chart -- 1995
Total Intensity (F): Main Field
90° 120° 150° 180° 210° 240° 270° 300° 330° 0° 30° 60°
60° 30° 0° -30° -60°
45000
45000
45000
60000
30000
30000
60000
90 85 80 75 70 65 60 55 50 45 40 35 30 25 20 15 10 5 0
Units (Total Intensity) : nanoTeslas
Contour Interval : 5000 nanoTeslas
Map Projection : Mercator
Color Scale: X 1000

Figure 10.47: Earth's magnetic field (source: `http://geomag.usgs.gov`)

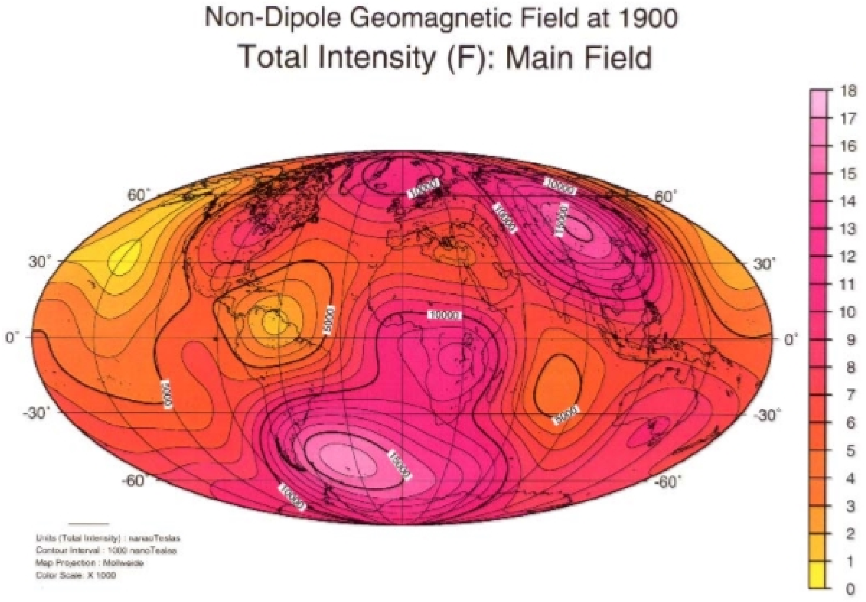
and for the purely horizontal component (tangential to the surface):


DOD World Magnetic Chart -- 1995
Horizontal Intensity (H): Main Field
90° 120° 150° 180° 210° 240° 270° 300° 330° 0° 30° 60°
60° 30° 0° -30° -60°
15000
15000
30000
30000
30000
15000
15000
90 85 80 75 70 65 60 55 50 45 40 35 30 25 20 15 10 5 0
Units (Horizontal Intensity) : nanoTeslas
Contour Interval : 5000 nanoTeslas
Map Projection : Mercator
Color Scale: X 1000

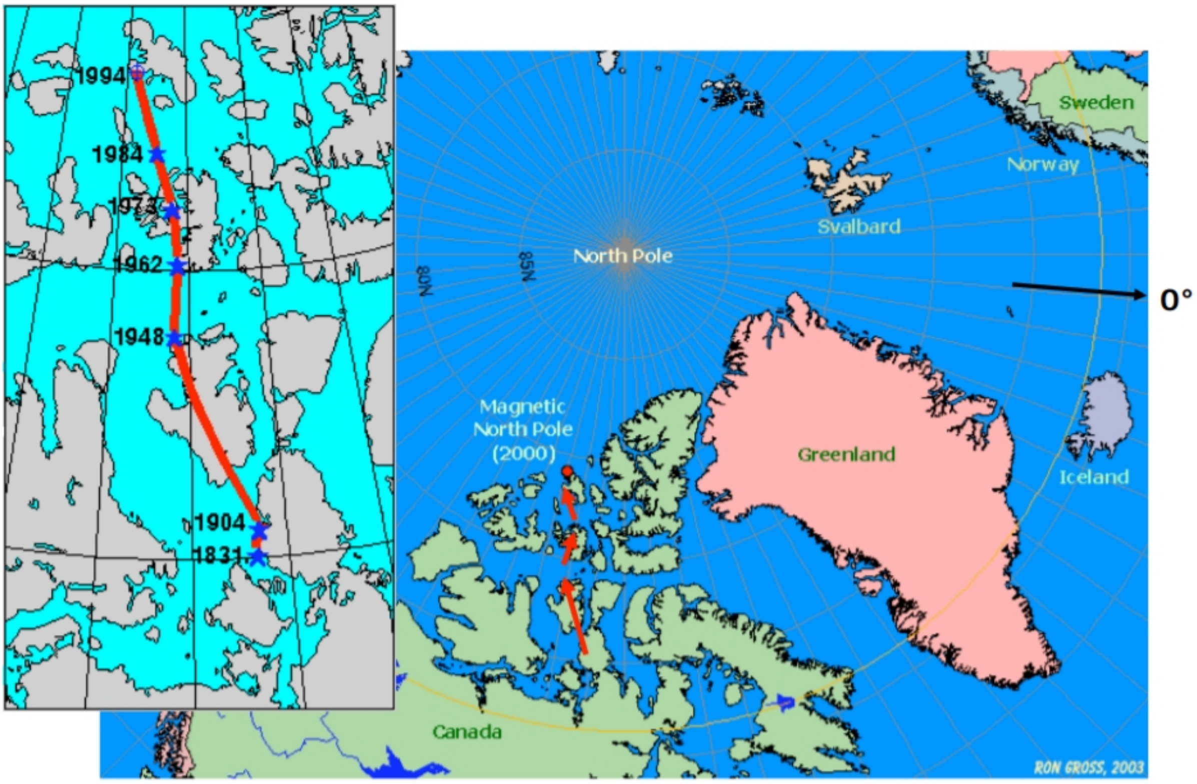
and for the purely vertical component:


DOD World MAGNETIC CHART -- 1995
Vertical Component (Z): Main Field
90° 120° 150° 180° 210° 240° 270° 300° 330° 0° 30° 60°
60°
30°
0°
-30°
-60°
45000
30000
15000
0
-15000
-30000
-45000
60000
90 85 80 75 70 65 60 55 50 45 40 35 30 25 20 15 10 5 0 -5 -10 -15 -20 -25 -30 -35 -40 -45 -50 -55 -60 -65 -70 -75 -80 -85 -90
Units (Vertical Component) : nanoTeslas
Contour Interval : 5000 nanoTeslas
Map Projection : Mercator
Color Scale: X 1000

But in fact, the Earth is far from being a perfect dipole as it is a complex non-homogeneous non-isotropic celestial object as shown in the figure below:


Non-Dipole Geomagnetic Field at 1900
Total Intensity (F): Main Field
60°
30°
0°
-30°
-60
10000
5000
18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0
Units (Total Intensity) : nanoTeslas
Contour Interval : 1000 nanoTeslas
Map Projection : Mollweide
Color Scale: X 1000

and furthermore the (main) magnetic north pole of Earth moves across the ages (this is named the "secular variation of Earth's magnetic north pole"):



A quite common question when dealing with such a topic is: Why does magnetism penetrate the huge majority of non-exotic (expected mu-metals and supra-conductors) so easily on thousand of kilometres like the magnetic field of Earth's inner core to Earth's surface!?

To answer this question we should reframe the question and ask: how do we prevent the penetration of magnetic fields in a solid in the first place? As we will see during our study of Maxwell equation, there are no magnetic monopoles (see page 4099) in nature, so the only possible way is to block the field is to run a current in the material which generates a field that counteracts the applied magnetic field.

If the only way for a material to to block an external magnetic field is through a current, we should think about how that current can be sustained in time. With the exception of superconductors, all materials have nonzero resistance. Therefore, the current needed to block penetration of magnetic fields will always decay in time. Thus, in general, there is no way to screen static magnetic fields in a material with non-zero resistance.

On the other hand, for alternating magnetic fields, most metals block magnetic fields just fine (provided they are thick compared to the skin depth) and the magnetic field does not penetrate except in an exponentially suppressed fashion. The key point here is Faraday's law (see page 4259) which tells us that we need a changing magnetic field to generate a current.

10.2.4.1 Magnetic torque

We have just introduced earlier the dipolar magnetic moment $\vec{\mu}_l$. For a magnet, the magnetic moment (denoted μ) is indirectly defined a quantity that determines the torque it will experience in an external magnetic field such that by definition:

$$\boxed{\vec{\tau}=\vec{\mu}\times\vec{B}} \quad (10.2.136)$$

Thus, if a magnetic dipole (or a bar magnet) is placed in a uniform magnetic field in oblique orientation, it experiences no force but experiences a torque. This torque tends to align the dipole moment along the direction of magnetic field!

We also have that a magnetic dipole moment may be defined as the torque acting on a magnetic dipole placed perpendicular to a uniform magnetic field of unit strength as in this case we have $\vec{\tau}=\vec{\mu}$.

In the magnetic pole representation of a material, a magnetic material can be identified as two magnetic poles some distance apart. The magnetic moment of a material bar is then given by the product of the "magnetic pole strength" denoted q_m and the distance between the poles L. Then for a magnetic bar where $\vec{n}$ is the unitary vector collinear to the magnet bar length, we have experimentally:

$$\vec{\tau}=\vec{\mu}\times\vec{B}=(q_m L\vec{n})\times\vec{B} \quad (10.2.137)$$

That is in scalar form:

$$\boxed{\tau=q_m LB\sin(\theta)} \quad (10.2.138)$$

This is the famous relation of the torque for a magnetic bar in a uniform constant magnetic field having two poles of magnetic strength q_m and $-q_m$:

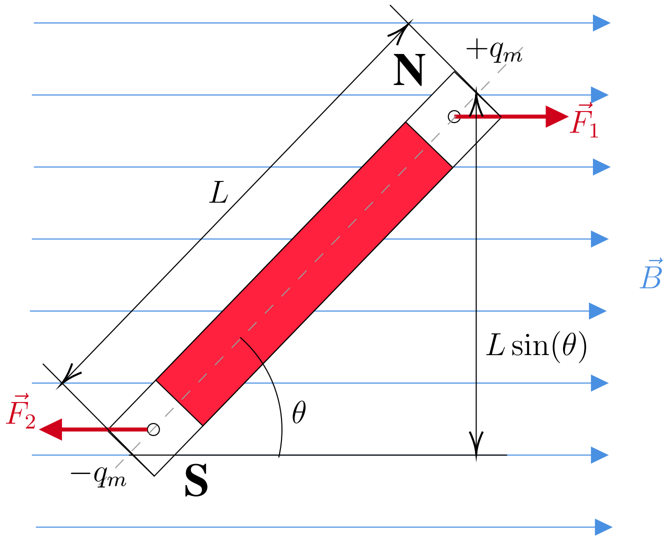
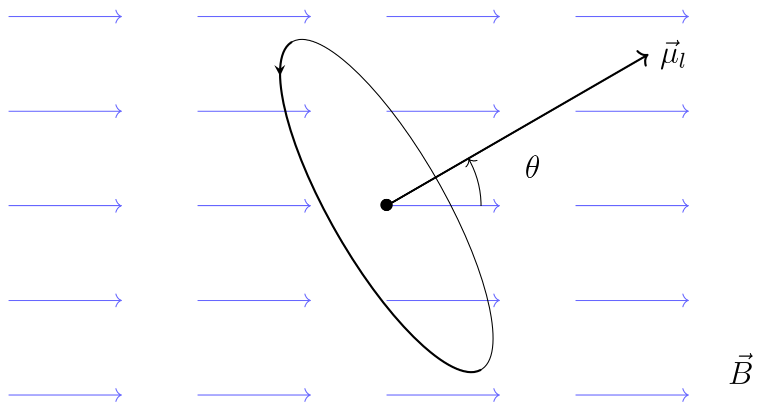


Figure 10.48: **Torque on a bar magnet in a uniform magnetic field**

The lines of action of both forces $\vec{F}_1$ and $\vec{F}_2$ are different, therefore, these two forces form a couple (or torque) which tends to rotate the magnet along the direction of magnetic field strength.

Thus, torque acting on the bar magnet is max, when it is placed perpendicular to the magnetic field.

Obviously the same applies for the dipolar magnetic moment $\vec{\mu}_l$:



The work done in rotating the dipole through a small angle $\mathrm{d}\theta$:

$$\mathrm{d}W = \tau \mathrm{d}\theta = \mu B \sin(\theta)\mathrm{d}\theta \tag{10.2.139}$$

If the dipole is rotated from $\theta = 0_1$ to $\theta = \theta_2$, total work done is given by:

$$W = \int\limits_{\theta_1}^{\theta_2} \mu B \sin(\theta)\mathrm{d}\theta = -\mu B \left(\cos(\theta_2) - \cos(\theta_1)\right) \tag{10.2.140}$$

If $\theta_1 = \pi/2$ and $\theta_2 = \theta$, then:

$$W = -\mu B \cos(\theta) \tag{10.2.141}$$

This work done is stored in the form of potential energy i.e:

$$E_p = -\mu B \cos(\theta) \tag{10.2.142}$$

We will see again this relation further below.

10.2.5 Lorentz law (Lorentz force)

In electrostatics, we calculated the force exerted by one or a collection of electric charges at rest (...) on a stationary or moving electric charge. The electric force was for recall written as following:

$$\vec{F}=q\vec{E}(\vec{r}) \tag{10.2.143}$$

In the most general case, where the acting electric charges are moving, the force they exert on a punctual electric charge q placed in a given point in space is the sum of two terms: one that is independent of the relative speed $\vec{v}$ of this electric charge, the other that depends on it. Here's how is written this relation:

$$\boxed{\vec{F}=q\left(\vec{E}+\vec{v}\times\vec{B}\right)} \tag{10.2.144}$$

that is the "Lorentz law" or "Lorentz force".

To prove this relation, we will put two assumptions, but first it is important to inform the reader that this proof requires not necessarily obvious mathematical tools (most reader will have to read the section of Analytical Mechanics and Wave Quantum Physics to understand it well):

H1. Given a non-relativistic point particle of mass m, of coordinates x_i with $i=1,2,3$ and velocity $\dot{x}_i$ still with $i=1,2,3$. We assume that it is subjected to a force $\vec{F}$ that satisfies Newton's equations:

$$m\ddot{\vec{x}}=\vec{F}(\vec{x},\vec{v},t) \tag{10.2.145}$$

with the following commutation relations (see sections of Analytical Mechanics page 3189 and Wave Quantum Physics page 4718):

$$\begin{gathered}[x_i,x_j]=0\\ m[x_i,v_j]=\mathrm{i}\hbar\delta_{ij}\end{gathered} \tag{10.2.146}$$

The reader must realize that the last relation is an assumption (hypothesis) and it is not equivalent to the commutation rules that we saw in the section of Wave Quantum Physics between positions and linear momentum!

H2. There exist fields $\vec{E}(\vec{x},t)$ and $\vec{B}(\vec{x},t)$, not depending on the speed, such that:

$$\vec{F}=q(\vec{E}+\vec{v}\times\vec{B}) \tag{10.2.147}$$

and which satisfy the Maxwell equations (see section Electrodynamics page 4095):

$$\begin{gathered}\vec{\nabla}\times\vec{E}=-\frac{\partial\vec{B}}{\partial t}\\ \vec{\nabla}\circ\vec{B}=0\end{gathered} \tag{10.2.148}$$

At a classical level, we express the commutation assumptions by using the Poisson correspondence commutation-brackets (see section Analytical Mechanics page 3189), that is:

$$\begin{gathered}\{x_i,x_j\}=0\\ m\{x_i,\dot{x}_j\}=\mathrm{i}\hbar\delta_{ij}\end{gathered} \tag{10.2.149}$$

with (recall):

$$\{A,B\}=\sum_{i=1}^{3}\left\{\frac{\partial A}{\partial x_i}\frac{\partial B}{\partial p_i}-\frac{\partial A}{\partial p_i}\frac{\partial B}{\partial x_i}\right\} \tag{10.2.150}$$

Now we define a vector potential $\vec{A}$ (see section Electrodynamics page 4057) such that:

$$p_j=m\dot{x}_j+qA_j \tag{10.2.151}$$

then the hypothesis of commutation ($\{x_i,p_j\}=\delta_{ij}$) can be written for $i\neq j$:

$$\{x_i,p_j\}=x_i(m\dot{x}_j+qA_j)-(m\dot{x}_j+qA_j)x_i=0=\{x_i,A_j\} \tag{10.2.152}$$

So we can say that $\vec{A}$ depends only on $\vec{x}$ and t and since it commutes identically to p_j.

Moreover, we know that classical mechanics admits a Lagrangian formulation (equivalent to Newton's equations) for which the equations of mechanics become (see section Analytical Mechanics page 3181):

$$\begin{aligned}p_j&=\frac{\partial L}{\partial \dot{x}_j}\\ \dot{p}_j&=\frac{\partial L}{\partial x_j}\end{aligned} \tag{10.2.153}$$

where L is the Lagrangian of the system. Therefore, with:

$$p_j=m\dot{x}_j+qA_j(\vec{x},t) \tag{10.2.154}$$

we can integrate the relation:

$$p_j=\frac{\partial L}{\partial \dot{x}_j} \tag{10.2.155}$$

and we get:

$$L=\frac{1}{2}m\dot{x}_j^2+q\left(\dot{x}_jA(\vec{x},t)-\phi(\vec{x},t)\right) \tag{10.2.156}$$

Just as the electric scalar potential ϕ is a potential energy per unit charge, the magnetic vector potential $\vec{A}$ is potential energy per unit length element of current (but current has a direction, so the magnetic vector potential has to have three components, that's why it's a vector too...) that is [J$\cdot$m^{-1}A^{-1}]. It is an old tradition sometimes to use a derivate unit name the "Weber" such that:

$$1\,[\mathrm{J}\cdot\mathrm{m}^{-1}\mathrm{A}^{-1}]=1\,[\mathrm{W}\cdot\mathrm{m}^{-1}] \tag{10.2.157}$$

The "$-$" sign of the constant of integration of the vector potential is justified to be consistent with what we will see during our study of gauge theory (see section Electrodynamics page 4115).

The second Lagrange equation $p_j = m\dot{x}_j + qA_j(\vec{x},t)$ gives us then:

$$\dot{p}_j = m\ddot{x}_j + q\dot{A}_j = \frac{\partial L}{\partial x_j} \tag{10.2.158}$$

By developing a little bit:

$$m\ddot{x}_j = \dot{p}_j - q\dot{A}_j \tag{10.2.159}$$

and:

$$\dot{p}_j = \frac{\partial L}{\partial x_j} = \frac{\partial}{\partial x_j}\left(\frac{1}{2}m\dot{x}_j^2 + q\left(\dot{x}_j A(\vec{x},t) - \phi(\vec{x},t)\right)\right) \tag{10.2.160}$$

For the set of all coordinates, this gives under condensed form and using the tools of vector analysis:

$$\boxed{m\ddot{\vec{x}} = q\left(-\vec{\nabla}\cdot\phi - \frac{\partial\vec{A}}{\partial t} + \dot{\vec{x}}\times\left(\vec{\nabla}\times\vec{A}\right)\right)} \tag{10.2.161}$$

So what must be absolutely noticed here is that it is the potential vector that appear in the Lagrangian and not the magnetic field!!! Therefore it's the potential vector that is at the origin of the variation of energy of a particle in an electromagnetic field and not the magnetic field!!!!!!!!

Therefore:

$$m\ddot{\vec{x}} = \vec{F} = q\left(\vec{E} + \dot{\vec{x}}\times\vec{B}\right) \tag{10.2.162}$$

or written differently:

$$\boxed{\vec{F} = q\left(\vec{E} + \vec{v}\times\vec{B}\right)} \tag{10.2.163}$$

We thus fall back well on the expression of the Lorentz force where $\vec{E}$ and $\vec{B}$ are given by:

$$\begin{aligned}\vec{E} &= -\vec{\nabla}\cdot\phi - \frac{\partial\vec{A}}{\partial t}\\ \vec{B} &= \vec{\nabla}\times\vec{A}\end{aligned} \tag{10.2.164}$$

as we will see it in gauges theory. Certainly the proof is far from being obvious, but is possible.

Many engineer have tested this result in school laboratory and know very well the following image (purple light is emitted along the electron path, due to the electrons colliding with gas molecules in the bulb and not because of bremsstrahlung that we will discuss later):

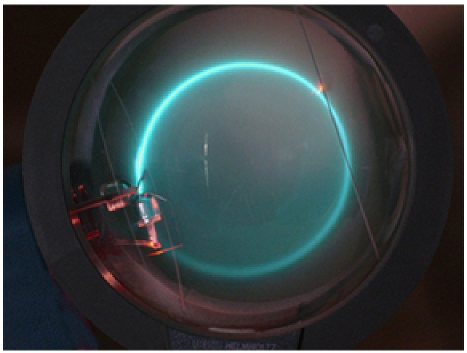


Figure 10.49: Beam of electrons moving in a circle, due to the presence of a magnetic field

You can also see this happening in photographs taken of charged particle tracks through a device known as the Wilson bubble chamber (we will come back on that one the section of Wave Quantum Physics as this detection procedure seems to be incompatible with Heisenberg's inequalities!):

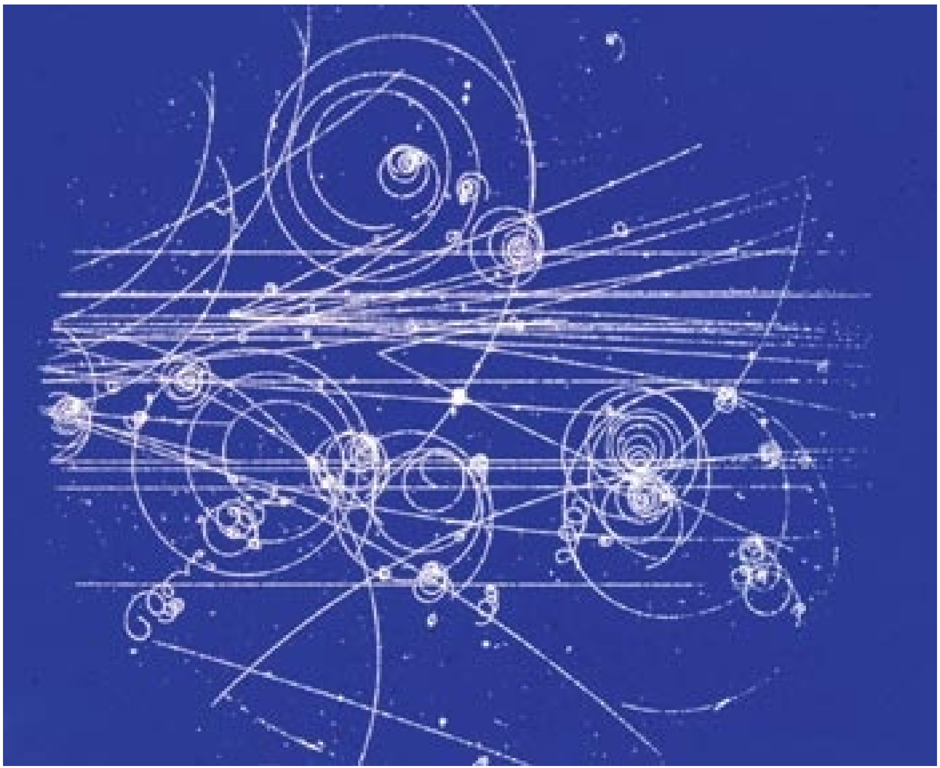
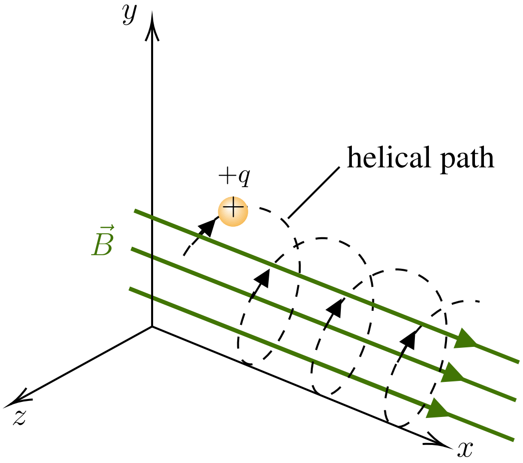


Figure 10.50: Bubble chamber (source: CERN)

In principle, this is still how physicists frequently determine charge to mass ratios of particles produced in particle physics experiments.

If a charged particle moves in a uniform magnetic field with its velocity at some arbitrary angle with respect to $\vec{B}$, its path is a helix as we will prove it further below (see page 4071):



Spiralling tracks as visible in the figure above are a common feature of bubble chamber pictures, and they are caused by electrons e^- (or positrons e^+, which - key point - have the same mass).

What a spiral tells us is that an electron loses energy at a considerable rate as it travels through a bubble chamber liquid. All other charged particles, unless they collide with a nucleus, very gradually slow down - get more curved - as they lose energy by ionisation (making bubbles in the bubble chamber).

Electrons are able to lose energy more quickly by another process known as "bremsstrahlung" (braking radiation). This process, which is a consequence of the fact that all accelerated charges radiate, is important for electrons because they have small masses. We will study mathematically this phenomenon in the section of Electrodynamics.

10.2.5.1 Magnetic Vector Potential

To get some experience with the magnetic vector potential, let us look first at what it is for a uniform magnetic field $\vec{B}_0$. This is calculation that will be useful during our study of Quantum Physics.

Taking our z-axis in the direction of $\vec{B}_0$, we must have:

$$\begin{aligned} B_x &= \frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z} = 0 \\ B_y &= \frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x} = 0 \\ B_z &= \frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y} = B_0 \end{aligned} \tag{10.2.165}$$

By inspection, we see that one possible solution of these equations is:

$$A_y = xB_0 \qquad A_x = 0 \qquad A_z = 0 \tag{10.2.166}$$

Or we could equally well take:

$$A_x = -yB_0 \qquad A_y = 0 \qquad A_z = 0 \tag{10.2.167}$$

Still another solution is a linear combination of the two:

$$A_x = -\frac{1}{2}yB_0 \qquad A_y = \frac{1}{2}xB_0 \qquad A_z = 0 \tag{10.2.168}$$

It is clear that for any particular field $\vec{B}$, the vector potential $\vec{A}$ is not unique; there are many possibilities.

The third solution, more general, shows that $\vec{A}$ is perpendicular to the z-axis in this special examples.

So there are actually (infinitely) many vector fields $\vec{A}$ whose curl will equal an arbitrary magnetic flux density $\vec{B}$. In other words, given some vector field $\vec{B}$, the solution $\vec{A}$ to the differential equation $\vec{\nabla} \times \vec{A} = \vec{B}$ is not unique!

To completely (i.e. uniquely) specify a vector field, we need to specify both its divergence and its curl and as we will see it in the section of Electrodynamics the divergence can satisfy the following gauge equation:

$$\vec{\nabla} \circ \vec{A} = 0 \tag{10.2.169}$$

Therefore finally we can write among other relations that we will also prove in the section of Electrodynamics that:

$$\boxed{\begin{aligned} &\vec{\nabla} \times \vec{A} = \vec{B} \\ &\vec{\nabla} \circ \vec{A} = 0 \end{aligned}} \tag{10.2.170}$$

Otherwise the vector potential for a uniform field can also be obtained in another way. The circulation of $\vec{A}$ on any closed loop Γ can be related to the surface integral of $\vec{\nabla}\times\vec{A}$ by Stokes' theorem (see section Vector Calculus page 1889):

$$\oint_{\Gamma}\vec{A}\circ\mathrm{d}\vec{r}=\int(\vec{\nabla}\times\vec{A})\circ\mathrm{d}\vec{S} \tag{10.2.171}$$

But the integral on the right is equal to the flux of $\vec{B}$ through the loop, so:

$$\oint_{\Gamma}\vec{A}\circ\mathrm{d}\vec{r}=\int\vec{B}\circ\mathrm{d}\vec{S} \tag{10.2.172}$$

So the circulation of $\vec{A}$ around any loop is equal to the flux of $\vec{B}$ through the loop. If we take a circular loop, of radius r in a plane perpendicular to a uniform field $\vec{B}$, the flux is just:

$$\Phi=\int\vec{B}\circ\mathrm{d}\vec{S}=\int\|\vec{B}\|\cdot\|\vec{n}\|\cos(\theta)\mathrm{d}\vec{S}=B\int\mathrm{d}S=B\pi r^2 \tag{10.2.173}$$

Let us see another way to derive that result. We want to try one of the possibilities for the vector potential to be:

$$\vec{A}=\frac{\vec{B}\times\vec{r}}{2} \tag{10.2.174}$$

It can be seen that:

$$\nabla\times\vec{A}=\frac{1}{2}\nabla\times(\vec{B}\times\vec{r})=\frac{1}{2}[\vec{B}(\vec{\nabla}\cdot\vec{r})-\vec{r}(\vec{\nabla}\cdot\vec{B})+(\vec{r}\cdot\vec{\nabla})\vec{B}-(\vec{B}\cdot\vec{\nabla})\vec{r}] \tag{10.2.175}$$

The first term gives $3\vec{B}$ since the $\vec{\nabla}\circ\vec{r}=3$. The second term is zero because $\vec{\nabla}\circ\vec{B}=0$. The third and the fourth terms are calculated as follows:

$$(\vec{r}\circ\vec{\nabla})\vec{B}=\left(x\frac{\partial}{\partial x}+y\frac{\partial}{\partial y}+z\frac{\partial}{\partial z}\right)\left(\vec{i}B_x+\vec{j}B_y+\vec{k}B_z\right)=\vec{0} \tag{10.2.176}$$

because the field is uniform.

$$\begin{aligned}(\vec{B}\circ\vec{\nabla})\vec{r}&=\left(B_x\frac{\partial}{\partial x}+B_y\frac{\partial}{\partial y}+B_z\frac{\partial}{\partial z}\right)(\vec{i}x+\vec{j}y+\vec{k}z)\\&=\vec{i}B_x+\vec{j}B_y+\vec{k}B_z=\vec{B}\end{aligned} \tag{10.2.177}$$

Adding all the four terms, result follows. It can be seen that the Coulomb gauge condition is satisfied. The divergence of the vector potential is given by:

$$\vec{\nabla}\circ\vec{A}=\frac{1}{2}\vec{\nabla}\circ(\vec{B}\times\vec{r})=\frac{1}{2}[\vec{r}\circ(\vec{\nabla}\times\vec{B})-\vec{B}\circ(\vec{\nabla}\times\vec{r})]=0 \tag{10.2.178}$$

because the first term vanishes as the field is uniform while the second term vanishes because the curl of position vector is zero. Thus the expression satisfies Coulomb gauge condition. If the magnetic field is along the z direction, we can take either of the following expressions for the vector potential components:

$$\begin{aligned}\vec{A}&=\left(-\frac{By}{2},\frac{Bx}{2},0\right)\\\vec{A}&=(-By,0,0)\end{aligned} \tag{10.2.179}$$

Both these expressions are valid expressions for the vector potential and they differ by gradient of a scalar field. It can be checked that if we add a term $\vec{\nabla}(\psi)$ to the first term where:

$$\psi = -\frac{Bxy}{2} \tag{10.2.180}$$

we get the second expression.

Now let us consider another interesting example. For an infinite wire along the z axis, the magnetic field, using cylindrical coordinates, is (see page 4026):

$$\vec{B} = \frac{I\mu_0}{2\pi r}\vec{u}_\phi \tag{10.2.181}$$

where r is the distance from the axis of the wire. We want to know the expression of the corresponding magnetic vector potential.

We know that the magnetic field has cylindrical symmetry and is directed along the circumferential direction and must satisfy:

$$\vec{B} = \frac{\mu_0 I}{2\pi r}\vec{u}_\phi = \nabla \times \vec{A} \tag{10.2.182}$$

Thus the curl of the vector potential only has ϕ component:

$$B_\phi = (\nabla \times \vec{A})_\phi = \frac{\partial A_r}{\partial z} - \frac{\partial A_z}{\partial r} = \frac{\mu_0 I}{2\pi r} \tag{10.2.183}$$

By symmetry, since the wire is infinite, the derivative with respect to z must be zero and we have:

$$-\frac{\partial A_z}{\partial r} = \frac{\mu_0 I}{2\pi r} \tag{10.2.184}$$

which gives:

$$\vec{A}(\vec{r}) = -\frac{\mu_0}{2\pi} I \ln(r)\vec{u}_z + \vec{\nabla}(\psi) \tag{10.2.185}$$

where we have explicitly added gradient of an arbitrary scalar field.

Now suppose we look inside the wire, taking the wire's radius to be R. At a distance r from the axis, only current within that radius contributes to the field, so we get from Ampère's law (see page 4011) using a circular loop within the wire (see page 4026):

$$\vec{B} = \frac{\mu_0 r I}{2\pi R^2}\vec{u}_\phi \tag{10.2.186}$$

Using exactly the same method as earlier we then get:

$$\vec{A}(\vec{r}) = -\frac{\mu_0 r^2 I}{4\pi R^2}\vec{u}_\phi + \vec{\nabla}(\psi) \tag{10.2.187}$$

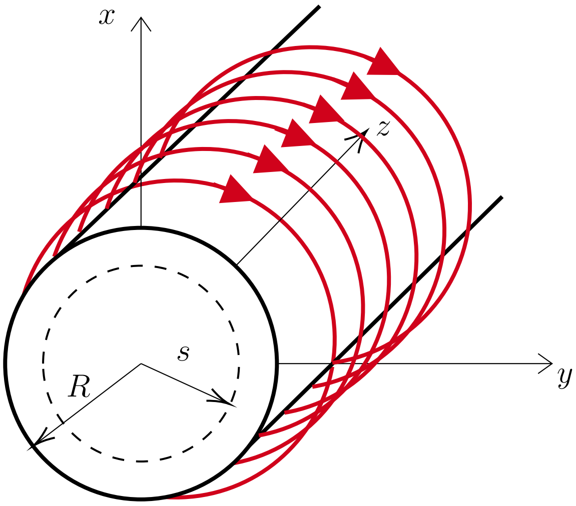
Again, we get $\vec{\nabla} \circ \vec{A} = 0$ and $\vec{\nabla} \times \vec{A} = \vec{B}$, so it looks like we're done.

Now let us move to another case! We know that the field inside a solenoid is along its axis (z direction) and is zero outside. If we take a circular path of radius s centered along its axis, the flux through the circular area is (see page 4014):

$$\Phi_B = \int \vec{B} \circ \mathrm{d}\vec{S} = \pi s^2 B = \pi \mu_0 n I s^2 \qquad (10.2.188)$$

The line integral of the vector potential is:

$$\oint\limits_{\Gamma} \vec{A} \circ \mathrm{d}\vec{s} = \int \vec{B} \circ \mathrm{d}\vec{S} = \pi \mu_0 n I s^2 \qquad (10.2.189)$$



The vector potential is thus given by for $s < R$:

$$\vec{A} = \mu_0 n I \frac{s}{2} \vec{u}_\phi \qquad (10.2.190)$$

Though the field outside is zero, the vector potential does not vanish outside the solenoid! This is because, if we take a circle of radius $s > R$, the flux through the circular area is:

$$\pi R^2 B = \pi R^2 \mu_0 n I \qquad (10.2.191)$$

flux being contributed only from inside the solenoid. Thus for $s > R$ as:

$$\oint\limits_{\Gamma} \vec{A} \circ \mathrm{d}\vec{s} = \int \vec{B} \circ \mathrm{d}\vec{S} = \pi \mu_0 n I R^2 \qquad (10.2.192)$$

We then get:

$$A \oint\limits_{\Gamma} \mathrm{d}s = A 2\pi s = \pi \mu_0 n I R^2 \qquad (10.2.193)$$

Hence:

$$\vec{A} = \mu_0 n I \frac{R^2}{2s} \vec{u}_\phi \qquad (10.2.194)$$

which falls off as inverse of distance from the axis.

Now an obvious question is if does vector potential have any physical significance?

It may be seen from the above example that the vector potential remains non-zero outside the solenoid even though the magnetic field itself has become zero. It turns out that the vector potential is not just a mathematical artefact but has physical reality!!!

If you are familiar with the Young's double slit experiment (see section Wave Optics page 4509) done with a beam of coherent light. The experiment is not really restricted to light wave but can be performed with matter wave such as a beam of electrons. We learn from quantum theory that like light exhibits dual behaviour, that of wave as well as particles known as photon, a wavelength is associated with material particles as well. This is known as "de Broglie wavelength" (see section Wave Quantum Physics page 4749) and is given by the ratio of the Planck's constant to the momentum of the particle.

The experiment is performed with an electron beam in place of light. What is done is to put a small solenoid just beyond the slits between the slits and the screen. Initially the solenoid does not carry any current and its dimensions are small enough so that it does not disturb the interference pattern produced on the screen because of the phase difference between electron waves arriving at the screen from the slits.

The current in the solenoid is switched on. The solenoid is of very small dimensions so that most of the electron beam passes outside the solenoid. Since the magnetic field outside is zero, the electron beams does not experience any force due to the magnetic field and should reach the screen undeflected. This should not then affect the interference pattern. However, what is found is that the pattern on the screen shifts suggesting a change in the phase relationship. It can be shown in quantum mechanics that the agent responsible for this phase change is the vector potential which is non-zero outside the solenoid though the magnetic field is zero.

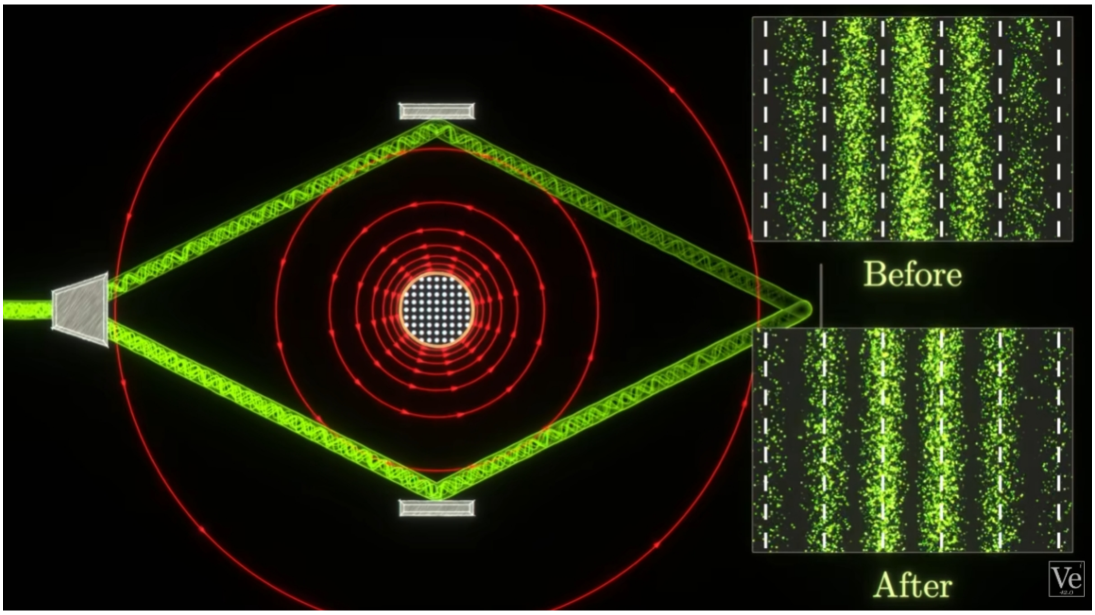


Figure 10.51: Illustrated concept of the Y. Aharonov and D. Bohm experiment (source: Vertisium)

The result is quite counter-intuitive because some people cannot accept that a particle can be influence by a potential in the absence of any force (that was the case of Niels Bohr)

For detailed experimental evidence see the papers [10] and [542].

10.2.5.2 Work of Magnetic Field

Let us stop a moment on the expression of the Lorentz force. We see with this relation that a stationary (or not) electric charge in an electric field will experience a force that will give him the necessary impulsion to make vary its kinetic energy (zero or non-zero at the start). This fact is however not valid for the magnetic field. Indeed, when we put a stationary electric charge in a magnetic field, this latter will not be influenced by magnetic field strength and therefore will not see its kinetic energy vary. If the electric charge has a non-zero initial velocity, it follows that the magnetic field will change the components of the velocity vector but not its norm!!! So, we used to say that "the magnetic field does not work" (in the sense that the magnetic field is not make move an electric charge that is initially at rest or change the norm of its speed). Let's see how we can prove mathematically that the magnetic field does not work.

Proof 10.2.2. We know that for an electric charged particle immersed in a magnetic field, we have:

$$\vec{F}=q\vec{v}\times\vec{B}=m\vec{a}=m\frac{\mathrm{d}\vec{v}}{\mathrm{d}t} \tag{10.2.195}$$

hence:

$$\frac{\mathrm{d}\vec{v}}{\mathrm{d}t}=\frac{q}{m}\vec{v}\times\vec{B} \tag{10.2.196}$$

and let us express the temporal variation of the kinetic energy:

$$\frac{\mathrm{d}}{\mathrm{d}t}E_c=\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}mv^2\right)=\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}m\vec{v}\circ\vec{v}\right)=m\vec{v}\circ\frac{\mathrm{d}\vec{v}}{\mathrm{d}t} \tag{10.2.197}$$

and substituting in it the derivative of the speed by prior-previous relation, we get:

$$\frac{\mathrm{d}}{\mathrm{d}t}E_c=m\vec{v}\circ\frac{\mathrm{d}\vec{v}}{\mathrm{d}t}=m\vec{v}\circ\left(\frac{q}{m}\vec{v}\times\vec{B}\right)=q\underbrace{\vec{v}\circ(\vec{v}\circ\vec{B})}_{\vec{v}\times\vec{v}\perp\vec{v}}=q\cdot 0=0 \tag{10.2.198}$$

The kinetic energy of the particle therefore does indeed not change because the magnetic field is conservative in the special case of stationary magnetic (and hence electric) fields.

This result also mean obvious that $\nabla\vec{F}=\vec{0}$. But if fact as we will prove it in the section of Electrodynamic this is a special case!! Indeed we will prove in the section of Electrokinetics the Faraday's induction law:

$$\vec{\nabla}\times\vec{E}=-\frac{\partial\vec{B}}{\partial t} \tag{10.2.199}$$

Therefore it is immediate that $\nabla\vec{F}$ will not be null anymore (and hence $\vec{F}$) if $\vec{B}$ is not static.

■ Q.E.D.

Now, if we are interested only in the second term of this relation, we can get to prove the "Laplace's force law"!

So we have considering only the second term of the Lorentz law:

$$\mathrm{d}\vec{F}=\mathrm{d}q\vec{v}\times\vec{B}=\rho\mathrm{d}V\vec{v}\times\vec{B}=\rho\mathrm{d}l\mathrm{d}S\vec{v}\times\vec{B} \tag{10.2.200}$$

where ρ is the volumetric electric charge density. If $\vec{v}$ and $\mathrm{d}\vec{l}$ are supposed parallels we can write that:

$$\mathrm{d}\vec{F}=\rho v\mathrm{d}S\mathrm{d}\vec{l}\times\vec{B} \tag{10.2.201}$$

A current density allows us to calculate the inertial speed of the electric charge carriers in a conductor. The number of conduction electrons in a wire is equal to:

$$nAl=\rho Al \tag{10.2.202}$$

where n is the number of conduction electrons per unit volume and Al the volume of the wire (and therefore A is the area of the section of the conductor).

A quantity of electric charges $q=\rho Al$ pass trough the surface section a wire in a time t given by:

$$t=\frac{l}{v} \tag{10.2.203}$$

The intensity of current I being defined by:

$$I=\frac{q}{t}=\frac{\rho Al}{\dfrac{l}{v}}=\rho Av \tag{10.2.204}$$

we get that:

$$v=\frac{J}{\rho} \tag{10.2.205}$$

From:

$$\mathrm{d}\vec{F}=\rho v\mathrm{d}S\mathrm{d}\vec{l}\times\vec{B} \tag{10.2.206}$$

We can now draw that:

$$\mathrm{d}\vec{F}=J\mathrm{d}S\mathrm{d}\vec{l}\times\vec{B} \tag{10.2.207}$$

Finally, we find that:

$$\boxed{\vec{F}=I\vec{l}\times\vec{B}=q\vec{v}\times\vec{B}} \tag{10.2.208}$$

which is the "Laplace's force law" or "Lorentz force" and therefore is derived from the Lorentz law.

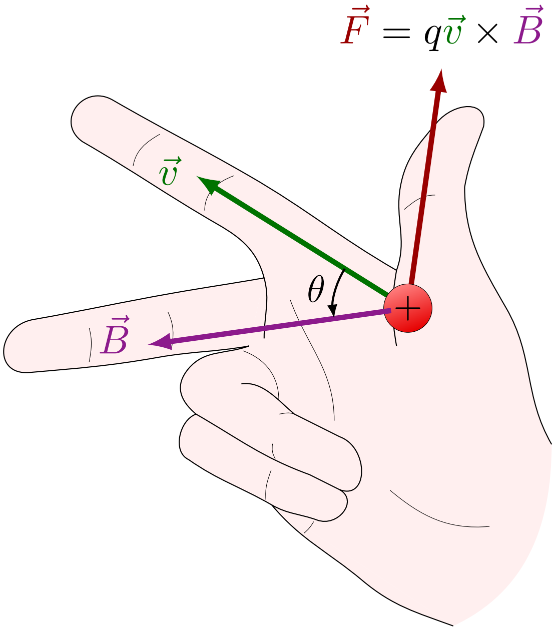


Figure 10.52: **Laplace's force right-handed rule (author: Izaadk Neutelings)**

We then deduce the units of the magnetic field (force per unit current and unit length of the source):

$$[\mathrm{N}\cdot\mathrm{m}^{-1}\cdot\mathrm{A}^{-1}]=[\mathrm{kg}\cdot\mathrm{s}^{-2}\cdot\mathrm{A}^{-1}]=[\mathrm{kg}\cdot\mathrm{s}^{-2}\cdot\mathrm{s}\cdot\mathrm{C}^{-1}]=[\mathrm{kg}\cdot\mathrm{s}^{-1}\cdot\mathrm{C}^{-1}]=[\mathrm{T}] \tag{10.2.209}$$

where T is a commonly tolerated unit in use named the "Tesla" in honour of Nikola Tesla one of the greatest spirit of all time with Albert Einstein. Therefore this unit implicitly contains the Coulomb unit which is then at the origin of the magnetic field! Knowing finally explicitly the units of the magnetic field, we can determine the units of the magnetic permeability constant by taking back the relation:

$$B=\frac{\mu_0 I}{2\pi r} \tag{10.2.210}$$

We have therefore for the units of the constant magnetic permeability:

$$[\mathrm{T}\cdot\mathrm{m}\cdot\mathrm{A}^{-1}]=[\mathrm{N}\cdot\mathrm{m}^{-1}\cdot\mathrm{A}^{-1}]=[\mathrm{kg}\cdot\mathrm{m}\cdot\mathrm{s}^{-2}\cdot\mathrm{A}^{-1}] \tag{10.2.211}$$

Remark

The magnitude of Earth's magnetic field at its surface ranges according to measurements from 25 to 65 $[\mu\mathrm{T}]$. Measurements done in Switzerland by the Biel Technical University the strength of the magnetic fields generated by electric trains changes constantly as the railroad car travels along, and is greatly dependent on the way the railroad car is advancing or braking, the magnetic fields were between 0.1 and 3 $[\mu\mathrm{T}]$.

So we see that it has for units a force by square current intensity. In SI units the permeability has (sadly) for units the "henry" per meter:

$$\frac{\mathrm{H}}{m}=\frac{\mathrm{kg}\cdot\mathrm{m}}{\mathrm{s}\cdot\mathrm{A}}$$

Having done this, let us see some important cases of application of the Lorentz law:

10.2.5.3 Classical Hall effect

Previously, we have studied the action of a magnetic induction on a wire-like circuit having for purpose to find the expression of the magnetic forces applied to the same area of this circuit. Now let us turn our attention to the conductivity electrons themselves, placing us in the case of the figure below:

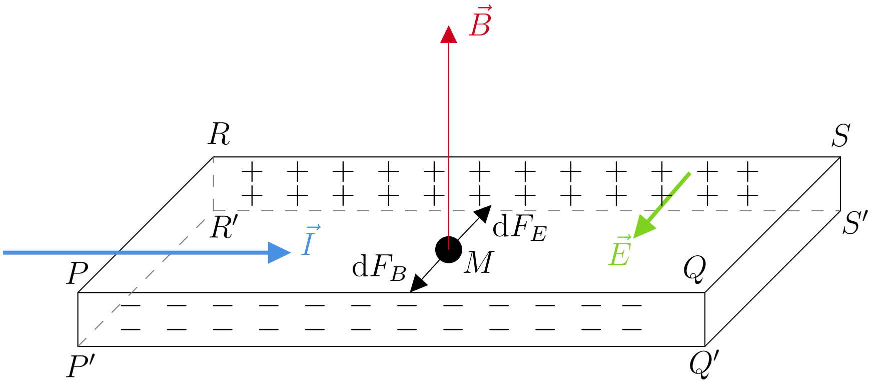


Figure 10.53: **DC current through a metallic ribbon for the study of classical Hall effect**

where a metallic ribbon is traversed by a continuous current $\vec{I}$. The vector of current density $\vec{J}$ is constant and parallel to the long sides $\overline{PQ}$ or $\overline{RS}$ of the ribbon.

Let us imagine then that the ribbon is immersed in a uniform magnetic field perpendicular to the planes PQ and RS (following the z-axis). The mobile electric charges of volumetric density ρ contained in a volume element $\mathrm{d}V$ is therefore subject to the magnetic force:

$$\mathrm{d}\vec{F}=\rho\mathrm{d}V\vec{v}\times\vec{B}=\vec{J}\times\vec{B}\mathrm{d}V \tag{10.2.212}$$

This force change the trajectories of mobile electrons and, during a transitional period, causing their accumulation on the front edge of the ribbon while an excess positive charge appears on the back edge.

This phenomenon produces an additional electric field parallel to $\overline{RP}$ which exerts on the mobile charges of volume $\mathrm{d}V$ an electric force:

$$\mathrm{d}\vec{F}=\rho\mathrm{d}V\vec{E} \tag{10.2.213}$$

The two forces are then opposing each other and the coulomb force tends to take back the electron paths in their initial position. A steady state is established gradually.

Remark

In fact, every time we talk about steady in physics, we lie a little bit. It is in fact just a stable equilibrium and in general, the system oscillates around its equilibrium position. After a while, a system like the conductor involved in our example shows negligible oscillations. Physics is sometimes also a matter of approximations...

When this state is reached, the current density is again parallel to $\overline{PQ}$ and the electric and magnetic forces above are vectorially opposed. So we have:

$$\vec{E}=-K(\vec{J}\times\vec{B}) \tag{10.2.214}$$

with:

$$K=\rho^{-1} \tag{10.2.215}$$

In some books, the cross product is made explicit in the form of its components such as:

$$E_y=\frac{J_x}{\rho}B_z \tag{10.2.216}$$

since the other components are zero (the current density is parallel to the ribbon and the magnetic field is perpendicular).

Now, as we have proved it in the section Electrokinetics we have:

$$\frac{J_x}{\rho}=v_x \tag{10.2.217}$$

therefore:

$$E_y=v_xB_z \tag{10.2.218}$$

We then define the "Hall coefficient" by:

$$R_H=\frac{E_y}{B_zJ_x} \tag{10.2.219}$$

R_H can be used at equilibrium for the measurement of J_x if we suppose $v_x=c^{te}$ and that the potential U (and therefore implicitly E_x) and external field B_z are known.

Notice that by construction we also have $R_H \propto \rho$ and this is used sometimes by engineer to determiner the density of electrical charge carriers in a new material sample.

Remark

We also tale of "Hall's resistance". It is simply the ratio of the Hall voltage on the current through the sample. It must not be confuse with the Hall resistance R_H.

In a two-dimensional semiconductor, the Hall effect is also measurable. By cons, at sufficiently low temperature, we observe a series of trays for the Hall resistance as a function of the magnetic field. These plates appear to specific resistance values, and this, independently of the sample used. This is the subject of the "quantum Hall effect" that we will not discuss in this chapter.

Under scalar form the relation of the "Hall effect", is written:

$$E=-KJB\sin(\alpha) \tag{10.2.220}$$

We can also express it by expliciting the potential difference that corresponds by definition to the electric field.

If l is the width of the ribbon, we have:

$$U=El \tag{10.2.221}$$

If e is its thickness, the current I that pass through it is equal to:

$$I = elJ \tag{10.2.222}$$

Given the relative positions of the various vectors, the relation expressing the Hall effect is equivalent to:

$$\frac{U}{l} = -K\frac{I}{el}B\sin(\alpha) \tag{10.2.223}$$

More aesthetically and in traditional form, the voltage of the Hall effect is given by:

$$U_H = k_H IB\sin(\alpha) \tag{10.2.224}$$

with:

$$\boxed{k_H = -\frac{1}{\rho e}} \tag{10.2.225}$$

which is the "Hall constant". It is inversely proportional to the density of the free electric charge carriers and in the context of metals, it is negative.

In other field of study such as the semiconductor, we write the Hall voltage in the following traditional form:

$$U_H = -\frac{1}{\rho e}IB\sin(\alpha) = -\frac{1}{qn}\frac{IB}{e}\sin(\alpha) = k_H\frac{IB}{e}\sin(\alpha) \tag{10.2.226}$$

where q is the electron charge, and n the traditional notation (sic!) of the carrier density within the study of semiconductors.

We then in the latter field the Hall constant which is defined by:

$$k_H = -\frac{1}{qn} \tag{10.2.227}$$

What has however made the reputation of the Hall effect, besides the fact that this result is enormously used to make magnetic fields sensors of all kinds (as Hall effect sensors operate without physical contact with the magnets), this is that for certain types of semiconductor this Hall constant is positive!!!! This would mean with the standard models we have available to us until now, there would be positive charges that would generate the current... and at the time of the establishment of this experience for semi-conductors this was inexplicable. At the time of Edwin Herbert Hall this experiment was used to check whether it was positive or negative charges that were moving and Hall concluded by testing it on conductive metals that only the negative electricity where moving in the conductors.

Remark

The most well known Hall probe by the population is the speedometer on bikes (odometer), which operates on the basis of attaching a small magnet on one of the spokes of a wheel and whose passage in front of the Hall sensor produces a signal processed by the electronic odometer.

We will prove later that by using quantum theory in the context of semiconductors (see section Electrokinetics page 4279) positive charges may yet appear under certain conditions and cause a current!

10.2.5.4 Larmor radius

A very interesting case of laboratory study is the movement of a charge in a uniform magnetic field. For this study, consider a particle of mass m and electric charge q placed in a uniform magnetic field with an initial velocity $\vec{v}_0$.

We have following Lorentz' law:

$$\frac{\mathrm{d}\vec{v}}{\mathrm{d}t}=\frac{q}{m}\vec{v}\times\vec{B} \tag{10.2.228}$$

We will take advantage of the fact that the magnetic force is zero in the direction of the magnetic field.

We'll break down the speed in two components, one parallel and one perpendicular to the magnetic field such as:

$$\vec{v}=\vec{v}_{\text{tg}}+\vec{v}_n \tag{10.2.229}$$

The equation of motion is then:

$$\frac{\mathrm{d}\vec{v}_{\text{tg}}}{\mathrm{d}t}=\vec{0} \qquad \frac{\mathrm{d}\vec{v}_n}{\mathrm{d}t}=\frac{q}{m}\vec{v}_n\times\vec{B}=\frac{q}{m}v_nB\sin(\alpha) \tag{10.2.230}$$

The trajectory remains rectilinear uniform in the direction of the magnetic field! In other words, if the speed of the charged particle was initially zero in the direction of the field (i.e. $\alpha=\pi/2$) then it will remain zero!

Let us now consider a Cartesian coordinate system with the Z axis is given by the direction of the magnetic field such that $\mathrm{d}v_z/\mathrm{d}t=0$. The equation of motion is therefore written only on two components, since:

$$\begin{pmatrix}\dfrac{\mathrm{d}v_x}{\mathrm{d}t}\\ \dfrac{\mathrm{d}v_y}{\mathrm{d}t}\\ \dfrac{\mathrm{d}v_z}{\mathrm{d}t}\end{pmatrix}=\begin{pmatrix}\dfrac{\mathrm{d}v_x}{\mathrm{d}t}\\ \dfrac{\mathrm{d}v_y}{\mathrm{d}t}\\ 0\end{pmatrix}=\frac{q}{m}\begin{pmatrix}v_x\\ v_y\\ v_z\end{pmatrix}\times\begin{pmatrix}B_x\\ B_y\\ B_z\end{pmatrix}=\frac{q}{m}\begin{pmatrix}v_x\\ v_y\\ v_z\end{pmatrix}\times\begin{pmatrix}0\\ 0\\ B\end{pmatrix}=\frac{q}{m}\begin{pmatrix}v_yB\\ -v_xB\\ 0\end{pmatrix} \tag{10.2.231}$$

hence:

$$\frac{\mathrm{d}v_x}{\mathrm{d}t}=\frac{qB}{m}v_y=\omega_{\text{gs}}v_y \qquad \frac{\mathrm{d}v_y}{\mathrm{d}t}=-\frac{qB}{m}v_x=-\omega_{\text{gs}}v_x \tag{10.2.232}$$

A very simple special solution to these two differential equations in non relativistic framework is obviously:

$$v_x=v_0\cos(\omega t) \qquad v_y=-v_0\sin(\omega t) \tag{10.2.233}$$

So where we choose an initial speed following the x-axis. By integrating, we get:

$$x=\frac{v_0}{\omega}\sin(\omega t) \qquad y=\frac{v_0}{\omega}\cos(\omega t) \tag{10.2.234}$$

where the integration constants were chosen as zero (arbitrary choice). The trajectory is thus a circle of radius:

$$\boxed{R_L=\frac{mv_0}{qB}} \tag{10.2.235}$$

perpendicular to the magnetic field and named "Larmor radius", described with the pulsation:

$$\omega_{\text{gs}}=\frac{qB}{m} \tag{10.2.236}$$

named "gyro-synchrotron pulsation". This circle is travelled in the conventional positive direction for negative charges.

Remark

The movement is circular only if the particle, initially, therefore has no speed in the direction of the magnetic field. If it has one, it will keep it (the magnetic field has no action in that direction).

The problem of such a configuration to build an accelerator is that if we increase the energy of the particle (by adding a synchronized electric field on the gyro-synchrotron pulsation and collinear to the movement), its speed increases and the radius Larmor also. But the "cyclotron" which is based on this system has a limited radius since it is difficult to maintain a constant magnetic field over a large area:

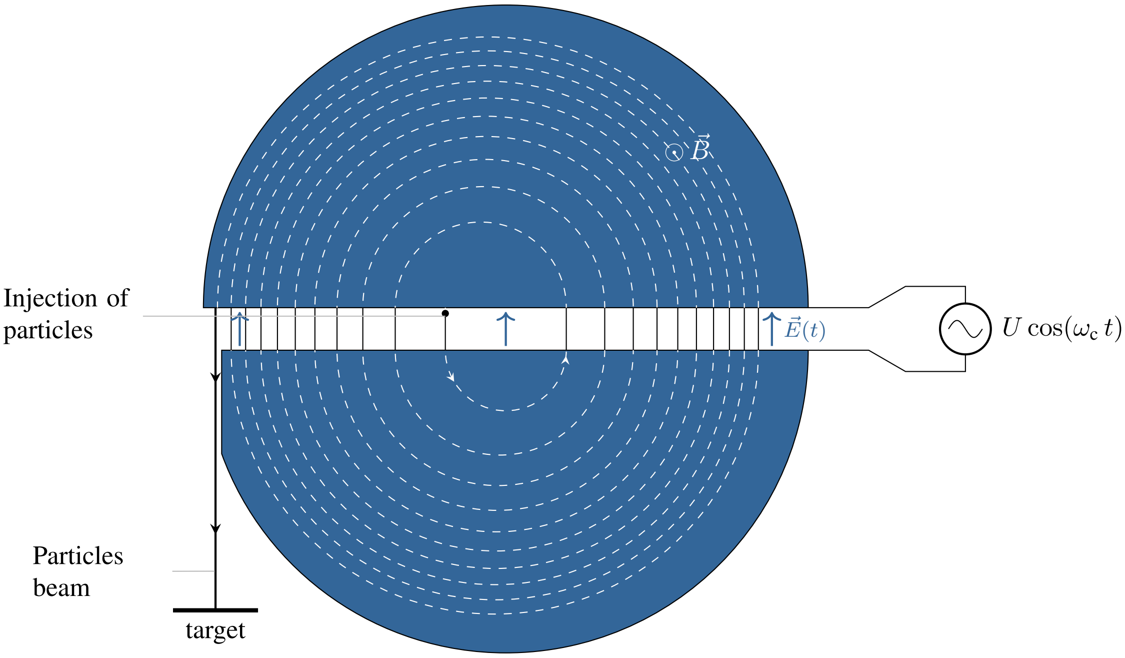


Figure 10.54: **Cyclotron principle (source: `http://femto-physique.fr`, author: Jimmy Roussel)**

Even more difficult, in the relativistic case, the pulse is then written with the Fitzgerald-Lorentz factor (see section Special Relativity page 5739):

$$\omega_{\text{gs}}=\frac{qB}{\gamma m} \tag{10.2.237}$$

We then see that we need to adjust the pulse of the electric field to the pulse of rotation when the sped increase: the accelerator is then now a "synchrotron".

To solve the problem of increasing radius, whereas we use a "Synchrotron" consisting of a single vacuum tube having straight sections containing the accelerating cavities and curved sections equipped with magnets creating at each instant the magnetic field adapted to the particle velocity. This technique, that it is easy to talk about but very difficult to put into practice, is the most used today. The CERN LHC is part of the family of synchrotrons.

From this relation, it is easy to have the kinetic energy of the particle:

$$E_c=\frac{1}{2}mv_0^2=\frac{1}{2}m\left(\frac{q}{m}BR_L\right)^2=\frac{1}{2}\frac{q^2}{m}(BR_L)^2 \tag{10.2.238}$$

As the kinetic energy if often acquired thanks to electric potential difference (when the technique is used with atoms or molecules they have to be ionized before), we have:

$$E_c=qU=\frac{1}{2}\frac{q^2}{m}(BR_L)^2 \tag{10.2.239}$$

Therefore after rearranging we have the famous relations named the "mass spectrometer equation" (also used for isotopic mass spectroscopy in nuclear physics):

$$\boxed{R_L=\frac{1}{B}\sqrt{\frac{2mU}{q}}\Leftrightarrow\frac{m}{q}=\frac{B^2R_L^2}{2U}} \tag{10.2.240}$$

It is based on this relation that works the "Dempster mass spectrometers". It is using this technique that researchers discovered in 11920 (holocene calendar) that atoms of the same chemical element does not necessarily have the same mass. The different varieties of atoms of the same chemical element, varieties that differ in mass, are the "isotopes" (see section Nuclear Physics page 5024):

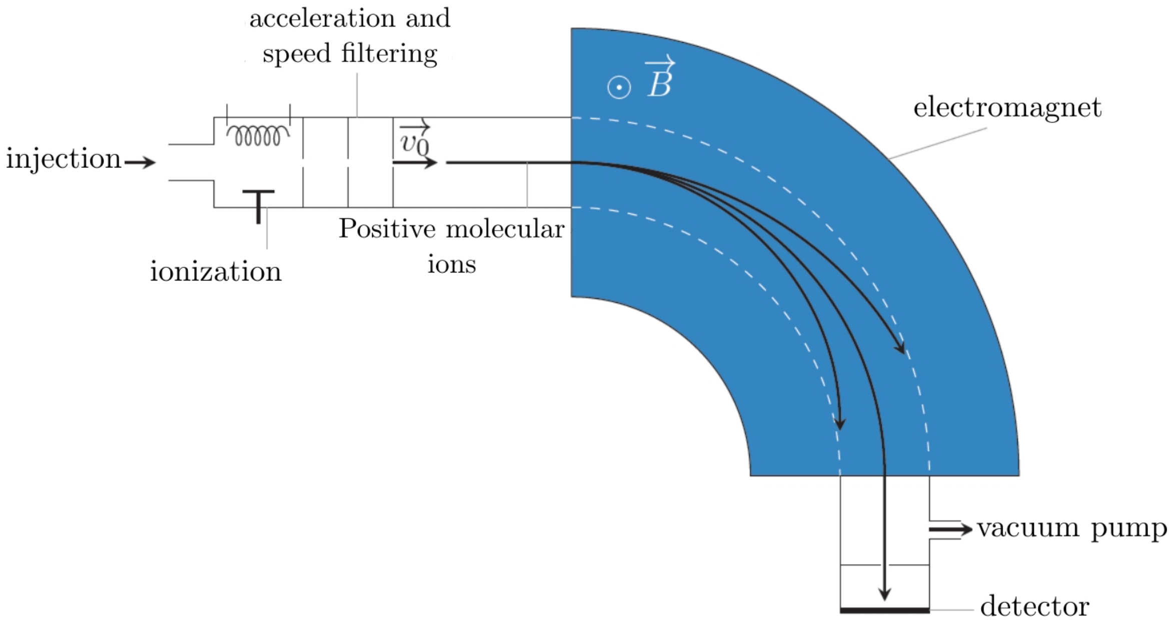


Figure 10.55: **Dempster mass spectrometers principle (source: `http://femto-physique.fr`, author: Jimmy Roussel)**

The Larmor radius is the greatest distance that a particle can travel in the transverse direction before being deflected from its path. This corresponds to a kind of trapping distance. Unless it receive additional kinetic energy, a charged particle is thus trapped in a magnetic field.

It is interesting to notice that more the kinetic energy of a particle is high (large mass or large transverse speed) and the Larmor radius is big. Conversely, the higher the magnetic field and the smaller is the Larmor radius (hence the fact we need hug magnetic field at CERN to keep high speed particles into a circular trajectory).

Let us consider now the case where $\alpha \neq \pi/2$, i.e. :

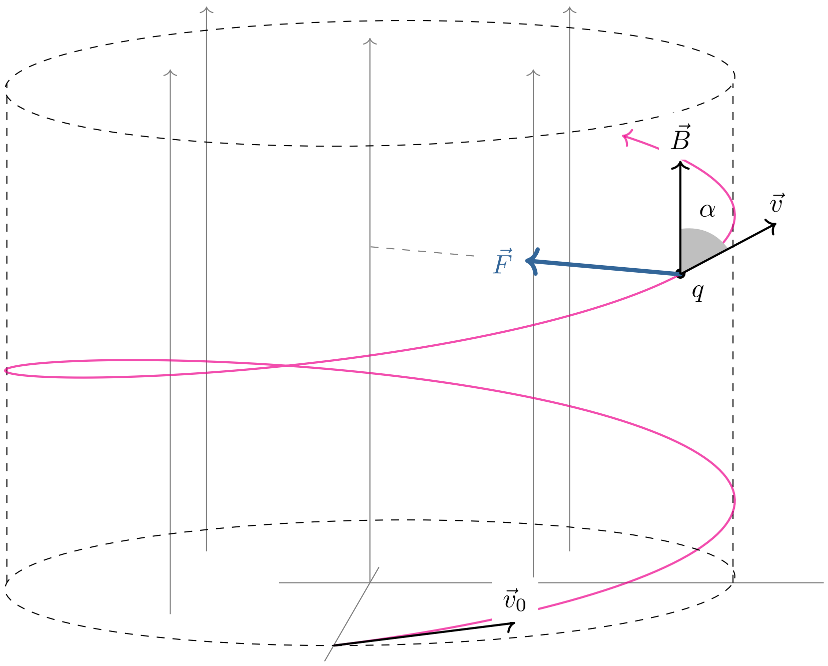


Figure 10.56: **Helical path of charge particle in a uniform constant magnetic field (author: Jimmy Roussel)**

Therefore is straightforward that Larmor radius we have obtained earlier is in fact special case of:

$$R_L = \frac{mv_n \sin(\alpha)}{qB} \tag{10.2.241}$$

The time needed to complete one revolution is obtained by definition of average velocity as:

$$\begin{aligned} v &= \frac{\Delta x}{\Delta t} \\ v_n &= \frac{2\pi R_L}{T} \\ \Rightarrow T &= \frac{2\pi R_L}{v_n} = \frac{2\pi\left(\dfrac{mv_n \sin(\alpha)}{qB}\right)}{v\sin(\theta)} = \frac{2\pi}{qB}m \end{aligned} \tag{10.2.242}$$

This period is also named the "cyclotron period" and its frequency is the reciprocal of the period with formula:

$$f = \frac{1}{T} = \frac{qB}{2\pi m} \tag{10.2.243}$$

The distance travelled parallel to the magnetic field B in one revolution is named the "pitch of the helical path" and is obtained as:

$$p = v_{\text{tg}} T = (v\cos(\theta)) \left(\frac{2\pi m}{qB}\right) \tag{10.2.244}$$

Thus, the formula for the pitch of the helix is:

$$p = \frac{2\pi m v \cos(\theta)}{qB} \tag{10.2.245}$$

To summarize so far:

- A helical path is formed when a charged particle enters with an angle of θ other than 90° into a uniform magnetic field. In the case of $\theta = 90^\circ$, a circular motion is generated.
- If the charged particle's velocity has components parallel and perpendicular to the uniform magnetic field then it moves in a helical path.
- In the parallel case, there is no force on the charged particle but in the perpendicular one, there is a centripetal acceleration toward the center.

We will come back on these concepts in the section Electrodynamics, where after studying the Maxwell equations, we will make some developments for the betatron type accelerator.

Remark

Confining the plasma in a tokamak is based on this property that charged particles have to describe a helical path around a magnetic field line. Hence the need to use a torus... in the simplest case.

To close this subject, and following the comment of a reader, let us develop a little more in detail the solution of the both differential equations seen previously:

$$\frac{\mathrm{d}v_x}{\mathrm{d}t} = \frac{qB}{m} v_y = \omega_{\text{gs}} v_y \qquad \frac{\mathrm{d}v_y}{\mathrm{d}t} = -\frac{qB}{m} v_x = -\omega_{\text{gs}} v_x \tag{10.2.246}$$

simplifying the notation, they become:

$$\begin{aligned} \ddot{x} &= \omega \dot{y} \\ \ddot{y} &= -\omega \dot{x} \\ \ddot{z} &= 0 \end{aligned} \tag{10.2.247}$$

and we will take as initial conditions:

$$\begin{aligned} x(0) &= 0 \quad & \dot{x}(0) &= v_{0x} \\ y(0) &= 0 \quad & \dot{y}(0) &= v_{0y} \\ z(0) &= 0 \quad & \dot{z}(0) &= v_{0z} \end{aligned} \tag{10.2.248}$$

The trick is to put:

$$\begin{aligned} X &= \dot{x} \\ Y &= \dot{y} \end{aligned} \tag{10.2.249}$$

The system of differential equations the becomes (we drop the Z component):

$$\begin{aligned} \dot{X} &= \omega Y \\ \dot{Y} &= -\omega X \end{aligned} \tag{10.2.250}$$

and therefore:

$$\ddot{X} = \omega\dot{Y} = \omega(-\omega X) = -\omega^2 X \Rightarrow \ddot{X} + \omega^2 X = 0 \tag{10.2.251}$$

And by doing the same for Y, we get:

$$\begin{aligned} \ddot{X} + \omega^2 X = 0 \\ \ddot{Y} + \omega^2 Y = 0 \end{aligned} \tag{10.2.252}$$

The both differential being identical it is sufficient to solve one to have the other solutions. The roots of the characteristic equation are (see section Differential and Integral Calculus page 1611):

$$K_{1,2} = \frac{-0 \pm \sqrt{0^2 - 4\omega^2}}{2} = \omega\sqrt{-1} = \mathrm{i}\omega \tag{10.2.253}$$

Since the discriminant is negative (the expression that is for recall under the root), then we have(see section Differential and Integral Calculus page 1609) the homogeneous solution is then:

$$X_h = e^{\alpha t}(A'\cos(\beta t) + B'\sin(\beta t)) = e^{0t}(A'\cos(\omega t) + B'\sin(\omega t)) \tag{10.2.254}$$

Hence:

$$X_h = A'\cos(\omega t) + B'\sin(\omega t) \tag{10.2.255}$$

And as there is no second member to our two differential equations, the homogeneous solution is then the general solution. Therefore:

$$X_G = X_h = A'\cos(\omega t) + B'\sin(\omega t) \tag{10.2.256}$$

and same:

$$Y_G = Y_h = C'\cos(\omega t) + D'\sin(\omega t) \tag{10.2.257}$$

Since we have put $X = \dot{x}$, $Y = \dot{y}$ and that we have $\ddot{x} = \omega\dot{y}$, then it comes:

$$\ddot{x} = \omega\dot{y} \Leftrightarrow \dot{X} = \omega Y \tag{10.2.258}$$

Thus explicitly:

$$-A'\omega\sin(\omega t) + B'\omega\cos(\omega t) = \omega(C'\cos(\omega t) + D'\sin(\omega t)) \tag{10.2.259}$$

Since this equation must be valid for all time t, let us consider the case where $t=0$, the equation is reduced to:

$$B'=C' \tag{10.2.260}$$

By doing the same for Y (always with $t=0$), we get:

$$D'=-A' \tag{10.2.261}$$

Taking into account the initial conditions for speed, we get:

$$\begin{aligned}\dot{x}(0)=X(0)=A'=-D'=v_{0x}\\ \dot{y}(0)=Y(0)=B'=C'=v_{0y}\end{aligned} \tag{10.2.262}$$

So:

$$\begin{aligned}X(t)=\dot{x}(t)=v_{0x}\cos(\omega t)+v_{0y}\sin(\omega t)\\ Y(t)=\dot{y}(t)=v_{0y}\cos(\omega t)-v_{0x}\sin(\omega t)\end{aligned} \tag{10.2.263}$$

Remark

By putting v_{0y} to be zero, we fall back on the simple special solution we proposed above with respect to speed.

To continue, we take the primitive to find the position coordinates. It comes then:

$$\begin{aligned}x(t)=\frac{v_{0x}}{\omega}\sin(\omega t)-\frac{v_{0y}}{\omega}\cos(\omega t)+c_x^{te}\\ y(t)=\frac{v_{0y}}{\omega}\sin(\omega t)+\frac{v_{0x}}{\omega}\cos(\omega t)+c_y^{te}\end{aligned} \tag{10.2.264}$$

For the initial conditions:

$$x(0)=0 \qquad y(0)=0 \tag{10.2.265}$$

to be satisfied, we see pretty quickly that we must have:

$$c_x^{te}=\frac{v_{0y}}{\omega} \qquad c_y^{te}=-\frac{v_{0x}}{\omega} \tag{10.2.266}$$

Thus finally:

$$\begin{aligned}x(t)=\frac{v_{0x}}{\omega}\sin(\omega t)-\frac{v_{0y}}{\omega}\cos(\omega t)+\frac{v_{0y}}{\omega}=\frac{1}{\omega}(v_{0x}\sin(\omega t)-v_{0y}\cos(\omega t)+v_{0y})\\ y(t)=\frac{v_{0y}}{\omega}\sin(\omega t)+\frac{v_{0x}}{\omega}\cos(\omega t)-\frac{v_{0x}}{\omega}=\frac{1}{\omega}(v_{0y}\sin(\omega t)+v_{0x}\cos(\omega t)-v_{0x})\end{aligned} \tag{10.2.267}$$

So if we take the Y component of the velocity to be zero, we have:

$$\begin{aligned}x(t)=\frac{v_{0x}}{\omega}\sin(\omega t)\\ y(t)=\frac{v_{0x}}{\omega}\cos(\omega t)-\frac{v_{0x}}{\omega}\end{aligned} \tag{10.2.268}$$

Which isn't anymore the path of a circle because of the initial conditions chosen for this development. To fall back on the circular path it require the initial condition on Y to be:

$$y(0)=\frac{v_{0x}}{\omega} \tag{10.2.269}$$

Trapped particles in magnetic fields are found in the Van Allen radiation belts around Earth, which are part of Earth's magnetic field. These belts were discovered by James Van Allen while trying to measure the flux of cosmic rays on Earth (high-energy particles that come from outside the solar system) to see whether this was similar to the flux measured on Earth. Van Allen found that due to the contribution of particles trapped in Earth's magnetic field, the flux was much higher on Earth than in outer space. Aurorae, like the famous aurora borealis (northern lights) in the Northern Hemisphere, are beautiful displays of light emitted as ions recombine with electrons entering the atmosphere as they spiral along magnetic field lines. Aurorae have also been observed on other planets, such as Jupiter and Saturn.

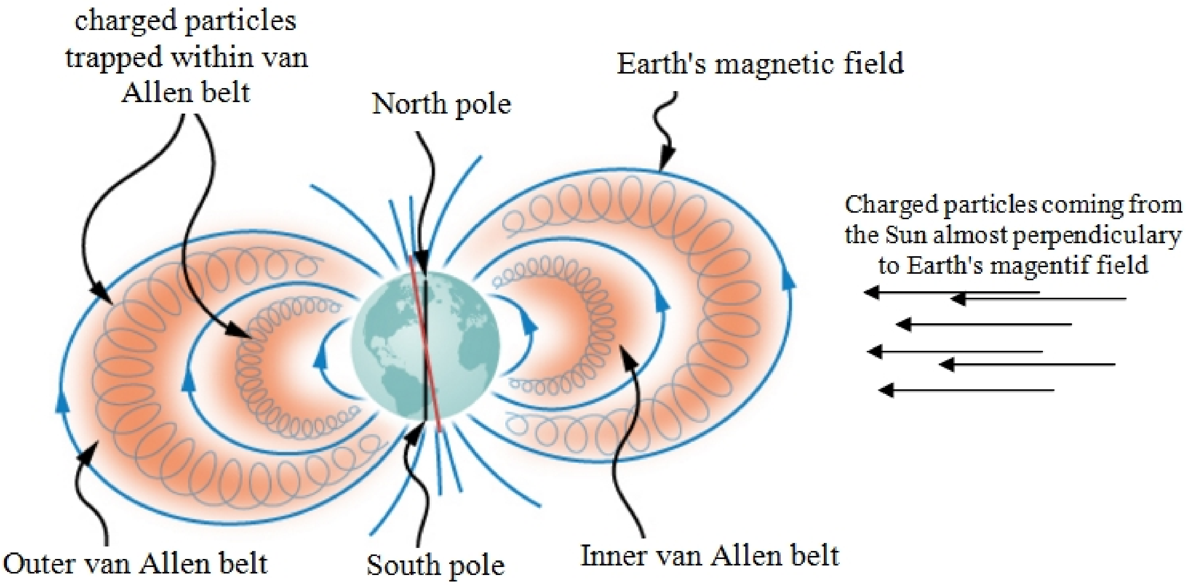

charged particles trapped within van Allen belt
North pole
Earth's magnetic field
Charged particles coming from the Sun almost perpendiculary to Earth's magentif field
Outer van Allen belt
South pole
Inner van Allen belt

Figure 10.57: Van Allen radiation belts around Earth (source: OpenStax)

Notice that if instead of the relation established earlier:

$$\begin{aligned}\ddot{x}&=\omega\dot{y}\\ \ddot{y}&=-\omega\dot{x}\\ \ddot{z}&=0\end{aligned}\qquad(10.2.270)$$

we add a proportional friction force term ($\vec{F}=-h\vec{v}$) like it is the case in a bubble chamber particle detector, we get:

$$\begin{aligned}\ddot{x}&=\omega\dot{y}-\frac{h}{m}\dot{x}\\ \ddot{y}&=-\omega\dot{x}-\frac{h}{m}\dot{y}\\ \ddot{z}&=0\end{aligned}\qquad(10.2.271)$$

We can write this system in matrix form (and we put $\tau=h/m$):

$$\begin{pmatrix}\ddot{x}\\ \ddot{y}\end{pmatrix}=\begin{bmatrix}-\frac{1}{\tau} & \omega\\ -\omega & -\frac{1}{\tau}\end{bmatrix}\begin{pmatrix}\dot{x}\\ \dot{y}\end{pmatrix}\qquad(10.2.272)$$

We are looking for the eigenvalues λ of the matrix. Its characteristic polynomial is written:

$$\begin{vmatrix} -\frac{1}{\tau}-\lambda & \omega \\ -\omega & -\frac{1}{\tau}-\lambda \end{vmatrix} = \left(-\frac{1}{\tau}-\lambda\right)^2+\omega^2=0 \tag{10.2.273}$$

The roots are: $\lambda_{1,2}=-\frac{1}{\tau}\pm\mathrm{i}\omega$. For $\lambda_1=-\frac{1}{\tau}+\mathrm{i}\omega$, the eigenvectors are:

$$\begin{bmatrix} -\mathrm{i}\omega & \omega \\ -\omega & -\mathrm{i}\omega \end{bmatrix} \cdot \begin{pmatrix} c_{11} \\ c_{12} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow \begin{pmatrix} c_{11} \\ c_{12} \end{pmatrix} = \begin{pmatrix} 1 \\ \mathrm{i} \end{pmatrix} \tag{10.2.274}$$

We put $\xi=\dot{X}$, hence:

$$\dot{\xi}=\begin{bmatrix} -\frac{1}{\tau} & \omega \\ -\omega & -\frac{1}{\tau} \end{bmatrix}\xi \tag{10.2.275}$$

Let us put :

$$\tilde{\xi}(t)=\mathrm{e}^{-\frac{t}{\tau}}\begin{pmatrix} \cos(\omega t)+\mathrm{i}\sin(\omega t) \\ -\sin(\omega t)+\mathrm{i}\cos(\omega t) \end{pmatrix} \tag{10.2.276}$$

A linear combination gives the solutions for ξ :

$$\xi(t)=C_1\mathrm{e}^{-\frac{t}{\tau}}\begin{pmatrix} \cos(\omega t) \\ -\sin(\omega t) \end{pmatrix}+C_2\mathrm{e}^{-\frac{t}{\tau}}\begin{pmatrix} \sin(\omega t) \\ \cos(\omega t) \end{pmatrix} \tag{10.2.277}$$

By integrating, with respect to t:

$$\begin{aligned} X &= \int \xi(t)\mathrm{d}t \\ &= C_1\frac{\mathrm{e}^{-\frac{1}{\tau}t}}{\frac{1}{\tau^2}+\omega^2}\begin{pmatrix} -\frac{1}{\tau}\cos\omega t+\omega\sin\omega t \\ \frac{1}{\tau}\sin\omega t+\omega\cos\omega t \end{pmatrix}+C_2\frac{\mathrm{e}^{-\frac{1}{\tau}t}}{\frac{1}{\tau^2}+\omega^2}\begin{pmatrix} -\frac{1}{\tau}\sin\omega t-\omega\cos\omega t \\ -\frac{1}{\tau}\cos\omega t+\omega\sin\omega t \end{pmatrix}+\begin{pmatrix} C_3 \\ C_4 \end{pmatrix} \end{aligned} \tag{10.2.278}$$

With the following initial conditions $x(0)=0, y(0)=0, \dot{x}(0)=v_{0,x}, \dot{y}(0)=v_{0,y}$, we get finally:

$$\begin{aligned} x(t) &= \frac{\tau}{1+\omega^2\tau^2}\left(-v_{0,x}\mathrm{e}^{-\frac{t}{\tau}}(\cos(\omega t)-\omega\tau\sin(\omega t))-v_{0,y}\mathrm{e}^{-\frac{t}{\tau}}(\sin(\omega t)+\omega\tau\cos(\omega t))+v_{0,x}+\omega\tau v_{0,y}\right) \\ y(t) &= \frac{\tau}{1+\omega^2\tau^2}\left(v_{0,x}\mathrm{e}^{-\frac{t}{\tau}}(\sin(\omega t)+\omega\tau\cos(\omega t))+v_{0,y}\mathrm{e}^{-\frac{t}{\tau}}(-\cos(\omega t)+\omega\tau\sin(\omega t))+v_{0,y}-\omega\tau v_{0,x}\right) \end{aligned} \tag{10.2.279}$$

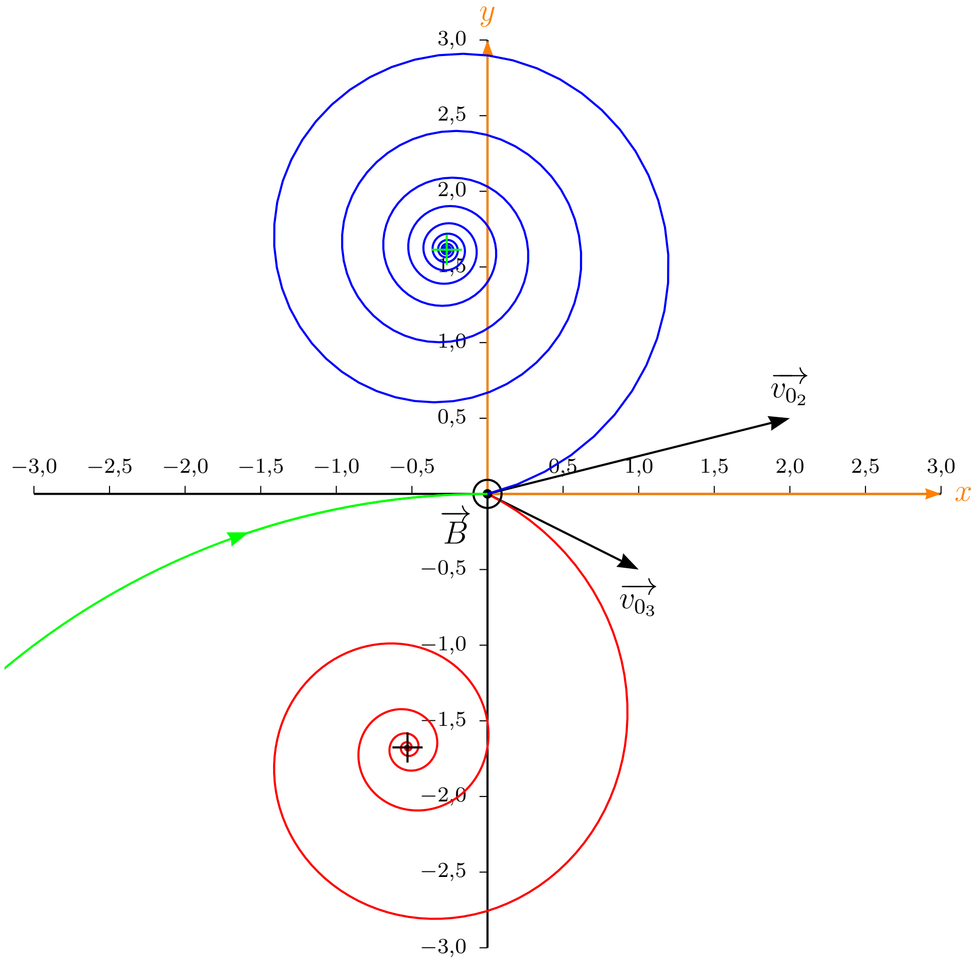


Figure 10.58: Bubble chamber helicoidal path

We can easily determine the asymptotic point! We put $t \to \infty$, therefore $\mathrm{e}^{-\frac{t}{\tau}} \to 0$:

$$\begin{aligned} x_{\infty} &= \frac{\tau}{1+\omega^2\tau^2}\left(v_{0,x}+\omega\tau v_{0,y}\right) \\ y_{\infty} &= \frac{\tau}{1+\omega^2\tau^2}\left(v_{0,y}-\omega\tau v_{0,x}\right) \end{aligned} \tag{10.2.280}$$

Now still in increasing difficulty, let us consider the case where we take into account also a movement in the z direction. We start from the general relation":

$$m\frac{\mathrm{d}\vec{v}}{\mathrm{d}t} = q\vec{v}\times\vec{B}-\lambda\vec{v} \tag{10.2.281}$$

Explicitly:

$$m\left(\ddot{x}\vec{e}_x+\ddot{y}\vec{e}_y+\ddot{z}\vec{e}_z\right) = q\left(\dot{x}\vec{e}_x+\dot{y}\vec{e}_y+\dot{z}\vec{e}_z\right)\wedge B\vec{e}_z-\lambda\left(\dot{x}\vec{e}_x+\dot{y}\vec{e}_y+\dot{z}\vec{e}_z\right) \tag{10.2.282}$$

Hence by projecting on the three axis:

$$\begin{aligned} m\ddot{x} &= qB\dot{y}-\lambda\dot{x} \\ m\ddot{y} &= -qB\dot{x}-\lambda\dot{y} \\ m\ddot{z} &= -\lambda\dot{z} \end{aligned} \tag{10.2.283}$$

Let us put $\omega_c = qB/m$ (the classic cyclotron pulsation) and $\tau = m/\lambda$ (time constant characteristic of energy losses), we arrive at:

$$\begin{gathered}\ddot{x} = \omega_c \dot{y} - \frac{1}{\tau}\dot{x}\\ \ddot{y} = -\omega_c \dot{x} - \frac{1}{\tau}\dot{y}\\ \ddot{z} = -\frac{1}{\tau}\dot{z}\end{gathered} \tag{10.2.284}$$

The third equation (in $\dot{z}$) is easily solved and gives with the initial conditions $\dot{z} = v_{0,z}$ and $z = 0$ at $t = 0$:

$$\dot{z} = v_{0,z} e^{-\frac{t}{T}} \quad \text{hence} \quad z = \tau v_{0,z}\left(1 - e^{-\frac{t}{\tau}}\right) \tag{10.2.285}$$

To solve the other two first equations we use a trick. We add the both of them but by multiplying the first one by $\cdot 1$ and the second one by $\cdot$i:

$$\begin{gathered}\ddot{x} = \omega_c \dot{y} - \frac{1}{\tau}\dot{x}\\ \mathrm{i}\ddot{y} = -\mathrm{i}\omega_c \dot{x} - \frac{\mathrm{i}}{\tau}\dot{y}\end{gathered} \tag{10.2.286}$$

The sum of the both gives:

$$(\ddot{x} + \mathrm{i}\ddot{y}) = \omega_c(\dot{y} - \mathrm{i}\dot{x}) - \frac{1}{\tau}(\dot{x} + \mathrm{i}\dot{y}) \tag{10.2.287}$$

If we rearrange this last relation a little bit, we get:

$$(\ddot{x} + \mathrm{i}\ddot{y}) = -\mathrm{i}\omega_c(\dot{x} + \mathrm{i}\dot{y}) - \frac{1}{\tau}(\dot{x} + \mathrm{i}\dot{y}) \tag{10.2.288}$$

Now we introduce a new variable $\xi = x + \mathrm{i}y$. Therefore:

$$\ddot{\xi} = -\mathrm{i}\omega_c \dot{\xi} - \frac{1}{\tau}\dot{\xi} \tag{10.2.289}$$

Or rearranging:

$$\ddot{\xi} = -\left(\frac{1}{\tau} + \mathrm{i}\omega_c\right)\dot{\xi} \tag{10.2.290}$$

For this equation, in fact in $\dot{\xi}$, the initial conditions at $t = 0$ are $x = y = 0$ therefore $\xi = 0$, $\dot{x} = v_0$, $\dot{y} = 0$ so $\dot{\xi} = v_0$, hence:

$$\dot{\xi} = v_0 e^{-\left(\frac{1}{\tau} + \mathrm{i}\omega_c\right)t} \tag{10.2.291}$$

Therefore we get immediately:

$$\xi = \frac{v_0}{\frac{1}{\tau} + \mathrm{i}\omega_c}\left(1 - e^{-\left(\frac{1}{\tau} + \mathrm{i}\omega_c\right)t}\right) \tag{10.2.292}$$

Using the Euler's formula (see section Numbers page 251), let us put:

$$\frac{1}{\tau}+\mathrm{i}\omega_c=\frac{1}{\tau^*}e^{\mathrm{i}\varphi} \tag{10.2.293}$$

with by construction:

$$\frac{1}{\tau^{*2}}=\frac{1}{\tau^2}+\omega_c^2 \tag{10.2.294}$$

and:

$$\tan(\varphi)=\tau\omega_c \tag{10.2.295}$$

Therefore:

$$\xi=\tau^*v_0e^{-\mathrm{i}\varphi}\left(1-e^{-\left(\frac{1}{\tau}+\mathrm{i}\omega_c\right)t}\right)=\tau^*v_0\left(e^{-\mathrm{i}\varphi}-e^{-\frac{t}{\tau}}e^{-\mathrm{i}(\omega_ct+\varphi)}\right) \tag{10.2.296}$$

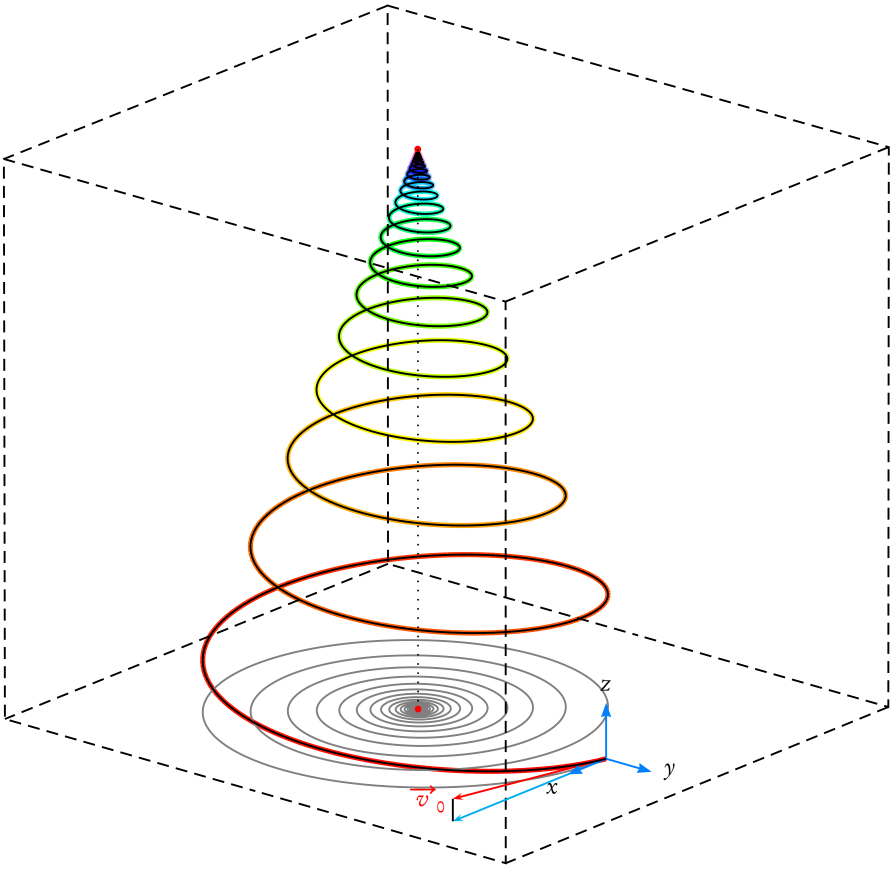
And finally:

$$\begin{aligned}x&=\Re(\xi)=\tau^*v_0\left[\cos(\varphi)-e^{-\frac{t}{\tau}}\cos\left(\omega_ct+\varphi\right)\right]\\y&=\Im(\xi)=\tau^*v_0\left[-\sin(\varphi)+e^{-\frac{t}{\tau}}\sin\left(\omega_ct+\varphi\right)\right]\end{aligned} \tag{10.2.297}$$

Therefore all three components are given by:

$$\begin{aligned}x&=\tau^*v_0\left[\cos(\varphi)-e^{-\frac{t}{\tau}}\cos\left(\omega_ct+\varphi\right)\right]\\y&=\tau^*v_0\left[-\sin(\varphi)+e^{-\frac{t}{\tau}}\sin\left(\omega_ct+\varphi\right)\right]\\z&=\tau v_{0,z}\left(1-e^{-\frac{t}{\tau}}\right)\end{aligned} \tag{10.2.298}$$

Here is the shape of the trajectory in perspective, for arbitrary values of the constants:



After about 7τ, the limit is reached and the charge has stopped at the coordinate point $x_\infty = \tau^* v_0 \cos(\varphi), y_\infty = -\tau^* v_0 \sin(\varphi), z_\infty = \tau v_{0,z}$.

10.2.5.5 Energy of a magnetic dipole

Thanks to the fact that we now have the units of the magnetic field and those of the magnetic permeability constant and the Lorentz law, we will now be able to determine by dimensional analysis and intuition the total energy of a static (oriented!) magnetic dipole which will be very useful for us for the theory of paramagnetism developed further below.

Let us consider for this a rigid magnet in the form of cylinder of length L and of negligible radius that can be regarded as a north / south dipole (a "bar magnet") immersed in a constant and homogeneous magnetic field in the plane perpendicular to the axis of rotation of the dipole:

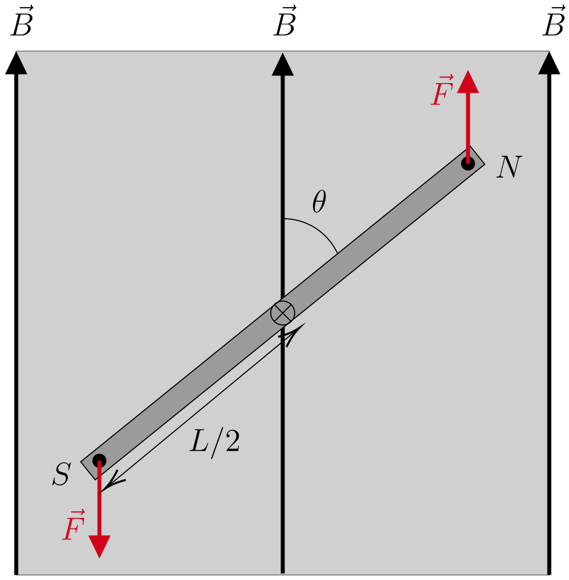


Figure 10.59: **Schematic diagram for the study of the static magnetic dipole**

The experience shows that when the dipole is collinear with the magnetic field, it does not move. It follows that the force on one end depends in proportion to the sine of the angle with the magnetic field such that:

$$\vec{F}\propto\vec{B}\sin(\theta) \tag{10.2.299}$$

At the dimension level we have so fare:

$$[\mathrm{kg}\cdot\mathrm{m}\cdot\mathrm{s}^{-2}]=[\mathrm{kg}\cdot\mathrm{s}^{-2}\cdot\mathrm{A}^{-1}] \tag{10.2.300}$$

We would have to get rid of Amps [A] by already making intervene at least obviously what characterizes a magnetic dipole... and we have already determined earlier above: its magnetic moment μ_l which units are for recall:

$$[\mathrm{A}\cdot\mathrm{m}^{-2}] \tag{10.2.301}$$

It seems quite natural enough to write to go a little further:

$$\vec{F}\propto\mu_l\vec{B}\sin(\theta) \tag{10.2.302}$$

This now gives at the dimensional level:

$$[\mathrm{kg}\cdot\mathrm{m}\cdot\mathrm{s}^{-2}]=[\mathrm{kg}\cdot\mathrm{s}^{-2}\cdot\mathrm{m}^{2}] \tag{10.2.303}$$

We then a unit of length in more. Then it seems quite natural to introduce the dipole length such that (the force should logically be equal at every point of the dipole so there is no reason to introduce here half of the dipole length!):

$$\vec{F}=\mu_l\vec{B}\sin(\theta)=\frac{\mu_l}{L}\vec{B}\sin(\theta) \quad (10.2.304)$$

Now to get back to the energy of the magnetic dipole, we consider that it is zero when the dipole is initially in a position perpendicular to the magnetic field. Using the common approximation as what infinitesimal movement of the two ends is given by for small angles:

$$\mathrm{d}l\cong\frac{L}{2}\mathrm{d}\theta \quad (10.2.305)$$

Thus, the elementary work (energy) to turn the dipole is (we multiply by two because we have to sum the two forces acting on each pole) by denoting by B the norm of the magnetic field:

$$\mathrm{d}E=2\left(2\frac{\mu_l}{L}B\sin(\theta)\right)\frac{L}{2}\mathrm{d}\theta=\mu_l B\sin(\theta)\mathrm{d}\theta \quad (10.2.306)$$

where we see that the dipole length is no longer involved. In fact we must not forget that it is in the magnetic dipole moment μ_l that we have the equivalent area of the dipole.

It comes by integration for a given final angle:

$$E=\mu_l B\int_{\pi/2}^{\theta}\sin(\theta)\mathrm{d}\theta=-\mu_l B\cos(\theta)\Big|_{\pi/2}^{\theta}=-\mu_l B\cos(\theta) \quad (10.2.307)$$

Therefore:

$$\boxed{E=-\vec{\mu}_l\circ\vec{B}} \quad (10.2.308)$$

Let us see now a classic and academic case that we can get from this result and that we will see again during our study of the spin in the section of Wave Quantum Physics.

So we have proved earlier above that an electric charged particle is deflected by a force given by the equation of the Lorentz' law:

$$\vec{F}=q\vec{v}\times\vec{B} \quad (10.2.309)$$

It follows that if the field has only a constant component in z and the speed only one component in x, this will cause a helical motion in the plane perpendicular to the field as we have already proved above in our study of the Larmor radius.

Let us consider now an electric charged particle launched at uniform speed along an x-axis between two poles of opposing magnets which generate a vertical and heterogeneous magnetic field and let us look only to the deflection along z of the trajectory of the particle.

From the perspective of the z-axis the particle can be considered in uniform acceleration (see section Classical Mechanics page 3224):

$$z=z_0+v_{0,z}t+\frac{1}{2}a_z t^2 \quad (10.2.310)$$

Since we assume that the initial position along z of the particle is midway between the two poles and that its initial velocity along the z-axis such as:

$$\begin{aligned}z_0&=0\\v_{0,z}&=0\end{aligned}\qquad(10.2.311)$$

Then we have:

$$z=\frac{1}{2}a_zt^2\qquad(10.2.312)$$

Since the magnetic field does not work (see proof earlier above) and that implies that the kinetic energy remains constant, we can write that the time is simply the ratio of the horizontal distance travelled by the particle by the speed module:

$$z=\frac{1}{2}a_z\left(\frac{L}{v}\right)^2=\frac{1}{2}a_zL^2\frac{1}{v^2}=\frac{1}{2}a_zL^2\left(\frac{m}{2E_c}\right)=\frac{mL^2}{4E_c}a_z\qquad(10.2.313)$$

Let us recall that we have proved just earlier above that the potential energy of a magnetic dipole was in a constant and homogeneous field given by:

$$E=-\vec{\mu}_l\circ\vec{B}\qquad(10.2.314)$$

Then in comes in our case:

$$E=-\mu_{lz}B_z\qquad(10.2.315)$$

And as we can associate a force to a potential energy, it comes by assuming that the dipole magnetic moment is constant along the z-axis and the magnetic field is still somewhat a little bit inhomogeneous (happy mix of assumptions ... but we still do engineering physics in this textbook!):

$$F_z=-\frac{\partial E}{\partial z}\cong\mu_{lz}\frac{\partial B_z}{\partial z}=ma_z\qquad(10.2.316)$$

And finally:

$$z=\frac{mL^2}{4E_c}a_z=\frac{mL^2}{4E_c}\left(\frac{\mu_{lz}}{m}\frac{\partial B_z}{\partial z}\right)=\frac{\mu_{lz}L^2}{4E_c}\frac{\partial B_z}{\partial z}\qquad(10.2.317)$$

So we see in any case that z can take a continuous spectrum of values that depends on the magnetic dipole moment of the particle. Now, as we shall see in the section of Wave Quantum Physics and Relativistic Quantum Physics during our study of quantum operators and especially that of the orbital angular momentum and of spin, an experiment named "Stern-Gerlach experiment" has shown that this was not the case for some particles or atoms for which the z are clearly discrete, what classical physics seemed unable to explain!!!!

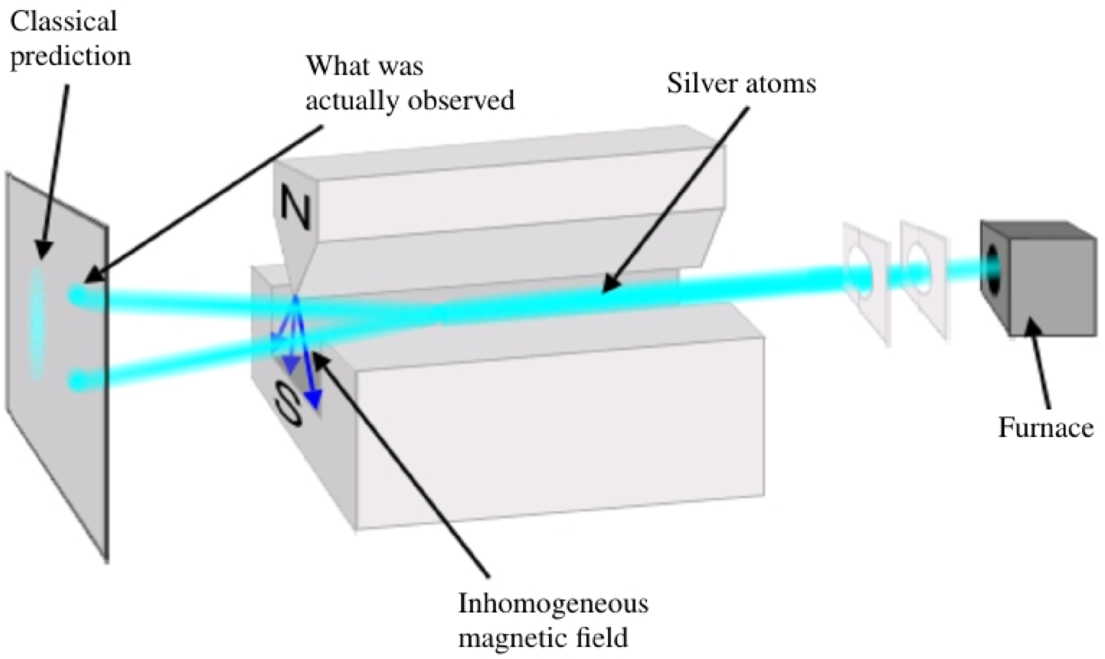

Classical prediction
What was actually observed
Silver atoms
N
S
Furnace
Inhomogeneous magnetic field

Figure 10.60: **Stern-Gerlarch experiment principle and result (source: Wikipedia, source: JohnDC†)**

Stern and Gerlach were astonished by what they saw. Not only were these "free" electrons deflected by the magnetic field, the pattern of their deflection was itself totally unexpected and mysterious. They found that the beam separated into two distinct parts. The basic experimental set-up is shown below.

This pointed to two things. First, the "free" electrons must have some kind of built-in magnetic moment because they are interacting with the magnetic field. Second, this magnetic moment is no ordinary dipole moment - the electrons are not acting like tiny little magnetic spheres shooting through the magnetic field.

10.2.6 Langevin treatment of Diamagnetism and Paramagnetism

We have define earlier what were diamagnetic and paramagnetic materials. In 11905 (holocene calendar) Langevin proposed after trial and error the theories of diamagnetism and paramagnetism (in 11906 - according to holocene calendar - Pierre Weiss proposed the theory of ferromagnetism) under some strong and simplified assumptions but that still work quite well to explain some material behaviours. We will see here what are the developments used to explained these behaviours.

10.2.6.1 Langevin model of diamagnetism

The purpose of this model is be able to explain the negative magnetism that is opposed to the magnetic excitation such that we can observe it in practice. This model is rough relatively to the quantum model but it is interesting for two main reasons: the first is that it already gives enough concepts to the reader to start the study of this subject if not familiar with quantum theory, the second being that it is a good training model (in the academic sense) because it shows how trials and successive approximations can lead to something relatively acceptable in practical perspective.

Remark

This model has its place in the section Magnetostatics because the excitation field used in the model is assumed to be constant!

For this, we consider the classical model of Langevin (the quantum model giving the same result as we will see in the corresponding section), where the electron is regarded as travelling a circular orbit of radius r and then can be assimilated to an electric current in a loop producing an electromotive force (see section Electrokinetics page 4253):

$$e=-\frac{\mathrm{d}\Phi_B}{\mathrm{d}t}=-\frac{\mathrm{d}(BS)}{\mathrm{d}t}=-\frac{\mathrm{d}(\mu_0 HS)}{\mathrm{d}t}=-\mu_0 S\frac{\mathrm{d}H}{\mathrm{d}t}=-\mu_0\pi r^2\frac{\mathrm{d}H}{\mathrm{d}t} \tag{10.2.318}$$

that we can assimilate to an electric field such that (see section Electrokinetics page 4256):

$$e=\oint\limits_{\text{orbit}} E_m\mathrm{d}l=E_m 2\pi r=-\mu_0\pi r^2\frac{\mathrm{d}H}{\mathrm{d}t} \tag{10.2.319}$$

Therefore it comes:

$$E_m=-\frac{\mu_0 r}{2}\frac{\mathrm{d}H}{\mathrm{d}t} \tag{10.2.320}$$

We then have by denoting m_e the rest mass of the electron and q_e its electric charge:

$$\frac{\mathrm{d}v}{\mathrm{d}t}=a=\frac{q_e E_m}{m_e}=-\frac{q_e}{m_e}\frac{\mu_0 r}{2}\frac{\mathrm{d}H}{\mathrm{d}t} \tag{10.2.321}$$

hence:

$$\Delta v=-\frac{q_e\mu_0 r}{2m_e}\Delta H=-\frac{q_e\mu_0 r}{2m_e}\left(H-\underbrace{H_0}_{=0}\right)=-\frac{-q_e\mu_0 r}{2m_e}H \tag{10.2.322}$$

The application of an external magnetic excitation will have for effect to change the magnetic dipole moment μ_l of quantity $\Delta\mu_l$. But, we have prove earlier above that the magnetic dipole moment was given by:

$$\mu_l=\frac{q\omega}{2}r^2=\frac{q\dfrac{v}{r}}{2}r^2=\frac{qr}{2}v \tag{10.2.323}$$

Then it comes for the electron:

$$\Delta\mu_l=\frac{q_e r}{2}\Delta v=\frac{q_e r}{2}\left(-\frac{q_e\mu_0 r}{2m_e}H\right)=-\frac{q_e^2\mu_0 r^2}{4m_e}H \tag{10.2.324}$$

Then, the very clever and rough trick (in the approximate sense of the term relatively to the experience and to the quantum model developed a few years later) is to take into consideration the fact that the electron in classical point of view can be considered as a point object that can move in a sphere of radius R given in the case of a mono-electronic atom and not just in a planar circular radius r perpendicular to the direction of the magnetic excitation field.

In this case, we then have of course:

$$x^2+y^2+z^2=R^2 \tag{10.2.325}$$

and we will consider that the three coordinates are independent and identically distributed random variables (so already the theoretical model is very rough but it's better than nothing...). Therefore, it follows that their expectation is equal such as:

$$\mathrm{E}(x^2+y^2+z^2)=\mathrm{E}(R^2) \tag{10.2.326}$$

Using the property of linearity of the expected mean (see section Statistics page 571) and the notation used by the physicist, it is then written:

$$\langle x^2\rangle+\langle y^2\rangle+\langle z^2\rangle=\langle R^2\rangle \tag{10.2.327}$$

and as the coordinates are considered like random variables identically distributed, we also have:

$$\langle x^2\rangle=\langle y^2\rangle=\langle z^2\rangle \tag{10.2.328}$$

It follows immediately that:

$$\langle x^2\rangle=\langle y^2\rangle=\langle z^2\rangle=\frac{\langle R^2\rangle}{3} \tag{10.2.329}$$

And so if we are interested only to the average disk radius containing all orbits perpendicular to the direction of the magnetic excitation field directed along the z-axis, then it comes:

$$\langle x^2\rangle+\langle y^2\rangle=\frac{2}{3}\langle R^2\rangle \tag{10.2.330}$$

Therefore, finally, for an electron in all possible orbits of a limited sphere:

$$\langle\Delta\mu_l\rangle=-\frac{q_e^2\mu_0\langle r^2\rangle}{4m_e}H=-\frac{q_e^2\mu_0\dfrac{2}{3}\langle R^2\rangle}{4m_e}H=-\frac{q_e^2\mu_0\langle R^2\rangle}{6m_e}H \tag{10.2.331}$$

where $\langle R^2\rangle$ can be explicitly calculated with the quantum model (with wave functions to be precise!).

For an atom containing Z electrons, we will do the rough assumption that a simple sum of the effects is valid ...:

$$\langle\Delta\mu_l\rangle=\sum_{i=1}^{Z}-\frac{q_e^2\mu\langle R_i^2\rangle}{6m_e}H=-\frac{q_e^2\mu_0H}{6m_e}\sum_{i=1}^{Z}\langle R_i^2\rangle \tag{10.2.332}$$

From a macroscopic point of view, the number of atoms in a unit volume will be the ratio of the density of the material divided by the atomic weight of the element in question multiplied by Avogadro's number:

$$n=N_{\text{A}}\frac{\rho}{m_A} \tag{10.2.333}$$

It comes then by unit volume:

$$\langle\Delta\mu_l\rangle=-n\frac{q_e^2\mu_0H}{6m_e}\sum_{i=1}^{Z}\langle R_i^2\rangle=-H\frac{nq_e^2\mu_0}{6m_e}\sum_{i=1}^{Z}\langle R_i^2\rangle=H\chi_d \tag{10.2.334}$$

and it is this result that is assimilated to the magnetic susceptibility as it's written in the form of the "Langevin's diamagnetic susceptibility relation":

$$\boxed{\chi_d=-\frac{q_e^2n\mu_0}{6m_e}\sum_{i=1}^{z}\langle R_i^2\rangle} \tag{10.2.335}$$

At unit level, we have well:

$$\chi_d=\frac{[\text{C}^2][\text{m}^{-3}][\text{kg}\cdot\text{m}\cdot\text{s}^{-2}\cdot\text{A}^{-2}]}{[\text{kg}]}[\text{m}^2]=[\text{C}^2][\text{s}^{-2}\cdot\text{A}^{-2}]=[\text{C}^2][\text{s}^{-2}\cdot\text{s}^2\cdot\text{C}^{-2}]=[-] \tag{10.2.336}$$

So it is a coefficient without units as expected.... and it is because this coefficient is negative that we assimilate this model to diamagnetism (by definition). The reader will have perhaps noticed that if the magnetic excitation is zero, the susceptibility is also zero... which is an expected minimum behaviour of the theoretical model. But by cons this latter does not depend on the temperature (the influence thereof is anyway almost negligible in standard laboratory temperature range).

The agreement experiment / theory is excellent for spherically symmetric elements in the order of $\pm 10\%$ error. For non-spherical elements the error often reaches 50% with respect to the experience.

10.2.6.2 Langevin model of paramagnetism

Langevin tried (with varying degrees of success here too) to explain the paramagnetism with the same underlying ideas, but however by opting for a completely different mathematical approach to ensure a positive outcome... (do it yourself physicist way of life...). What Langevin also knew it is that paramagnetism highly dependent on the temperature according to experimental studies of ferromagnetic materials, it was therefore necessary to choose an approach emphasizing the temperature and at the time there were no $10,000$ ways to do this! It follows that this model opens also the door to the theory of ferromagnetism!

As at the starting point at that time we naturally assumed the (Maxwell-)Boltzmann distribution proved in the section of Statistical Mechanics that describes for recall the distribution of detectable particles that do not interact with any constraints on the number of particles state... (at the time of Langevin there were only this model available...):

$$f_{\text{MB}}(E,T)=e^{-\frac{E}{kT}} \tag{10.2.337}$$

where as the reader maybe noticed it (see section Statistical Mechanics page 3554) the potential energy μ_i is taken to be zero, so that all energy is in the form of kinetic energy.

It is important to notice that the latter writing of the distribution function, as we have seen in the section Statistical Mechanics, ignores the normalization constant to make it a true probability density function (but we will obviously calculate that constant a little further below).

We have also proved just earlier that the magnetic potential energy of a dipole was given by:

$$E=-\vec{\mu}_l\circ\vec{B}=-\mu_l B\cos(\theta) \tag{10.2.338}$$

Now, let us recall that we have proved in the section of Geometric Shapes that the surface of a sphere element was given by:

$$\mathrm{d}S=R^2\sin(\theta)\mathrm{d}\theta\mathrm{d}\phi \tag{10.2.339}$$

It comes therefore (the reader may refer to the figure of the section of Geometric Shapes) that for a crown of the sphere defined by two parallel planes, which middle contains the origin of the sphere is then given by (if needed and on readers request we can draw another figure) :

$$\mathrm{d}S=R^2\sin(\theta)\frac{\mathrm{d}\theta}{2}2\pi=R^2\sin(\theta)\mathrm{d}\theta \tag{10.2.340}$$

Why are we talking about this? Well because a portion of the total number of magnetic dipoles included in an angle range $\mathrm{d}\theta$ is as we see above proportional to a surface element as:

$$\mathrm{d}\theta=\frac{\mathrm{d}S}{R^2\sin(\theta)} \tag{10.2.341}$$

We then have that this number is given by (to a given unknown constant factor K):

$$\mathrm{d}n\propto KR^2\sin(\theta)\mathrm{d}\theta e^{-E/kT}=KR^2\sin(\theta)\mathrm{d}\theta e^{\mu_l B\cos(\theta)/kT} \tag{10.2.342}$$

The corresponding proportion (so it is also a probability) over all angles, for a given angle, is then given after normalization obviously by:

$$P(\theta)=\frac{KR^2\sin(\theta)\mathrm{d}\theta e^{\mu_l B\cos(\theta)/kT}}{\int\limits_0^\pi KR^2\sin(\theta)\mathrm{d}\theta e^{\mu_l B\cos(\theta)/kT}\mathrm{d}\theta}=\frac{\sin(\theta)\mathrm{d}\theta e^{\mu_l B\cos(\theta)/kT}}{\int\limits_0^\pi \sin(\theta)\mathrm{d}\theta e^{\mu_l B\cos(\theta)/kT}\mathrm{d}\theta} \tag{10.2.343}$$

We saw during our study of the magnetic dipole moment that $\vec{\mu}_l$ contributes to the magnetic field. If we have a volumetric density of n magnetic dipoles, then we have a contribution of around $n\vec{\mu}_l$ if they are all oriented in the same direction. But if we project the vector of the magnetic moment on the direction of the magnetic field, the contribution of dipoles will then be written:

$$n\mu_l\cos(\theta) \tag{10.2.344}$$

but as there are many dipoles in different directions and that do many various angles with the field, then we have to take the average such that the contribution to the magnetic field is proportional to:

$$n\mu_l\langle\cos(\theta)\rangle \tag{10.2.345}$$

And using simply the properties of density statistical functions (see section Statistics page 633), the expected mean contribution is given by:

$$n\mu_l\langle\cos(\theta)\rangle=n\mu_l\int\limits_0^\pi\cos(\theta)P(\theta)\mathrm{d}\theta=n\mu_l\int\limits_0^\pi\cos(\theta)\frac{\sin(\theta)e^{\mu_l B\cos(\theta)/kT}}{\int\limits_0^\pi\sin(\theta)e^{-\mu_l B\cos(\theta)/kT}\mathrm{d}\theta} \tag{10.2.346}$$

and as the denominator is just a normalization constant, we can get it out of the first integral:

$$n\mu_l\langle\cos(\theta)\rangle=n\mu_l\frac{\int\limits_0^\pi\cos(\theta)\sin(\theta)e^{\mu_l B\cos(\theta)/kT}\mathrm{d}\theta}{\int\limits_0^\pi\sin(\theta)e^{\mu_l B\cos(\theta)/kT}\mathrm{d}\theta} \tag{10.2.347}$$

To integrate, let us do the little simplification of writing by putting:

$$\gamma=\frac{\mu_l B}{kT} \tag{10.2.348}$$

This gives us:

$$n\mu_l\langle\cos(\theta)\rangle=n\mu_l\frac{\int\limits_0^\pi\cos(\theta)\sin(\theta)e^{\gamma\cos(\theta)}\mathrm{d}\theta}{\int\limits_0^\pi\sin(\theta)e^{\gamma\cos(\theta)}\mathrm{d}\theta} \tag{10.2.349}$$

and let us put:

$$x=\cos(\theta) \tag{10.2.350}$$

Then we have:

$$\frac{\mathrm{d}x}{\mathrm{d}\theta}=-\sin(\theta)\Rightarrow \mathrm{d}\theta-\frac{\mathrm{d}x}{\sin(\theta)} \tag{10.2.351}$$

It comes then:

$$n\mu_{\langle}x\rangle=n\mu_l\frac{\displaystyle\int\limits_{+1}^{-1}xe^{\gamma x}\mathrm{d}x}{\displaystyle\int\limits_{+1}^{-1}e^{\gamma x}\mathrm{d}x} \tag{10.2.352}$$

The primitive numerator is known to us because it is part of the usual primitive proved in detail in the section of Differential and Integral Calculus! The integral in the denominator is trivial:

$$\begin{aligned}n\mu_l\langle x\rangle&=n\mu_l\frac{e^{\gamma x}\left(\dfrac{1}{\gamma}x-\dfrac{1}{\gamma^2}\right)\Bigg|_{+1}^{-1}}{\dfrac{1}{\gamma}e^{\gamma x}\Bigg|_{+1}^{-1}}=n\mu_l\frac{e^{-\gamma}\left(-1-\dfrac{1}{\gamma}\right)-e^{\gamma}\left(1-\dfrac{1}{\gamma}\right)}{e^{-\gamma}-e^{\gamma}}\\&=n\mu_l\frac{e^{\gamma}\left(1-\dfrac{1}{\gamma}\right)-e^{-\gamma}\left(-1-\dfrac{1}{\gamma}\right)}{e^{\gamma}-e^{-\gamma}}=n\mu_l\frac{e^{\gamma}+e^{-\gamma}-e^{\gamma}\dfrac{1}{\gamma}+e^{-\gamma}\dfrac{1}{\gamma}}{e^{\gamma}-e^{-\gamma}}\\&=n\mu_l\left[\frac{e^{\gamma}+e^{-\gamma}}{e^{\gamma}-e^{-\gamma}}-\frac{1}{\gamma}\frac{e^{\gamma}-e^{-\gamma}}{e^{\gamma}-e^{-\gamma}}\right]=n\mu_l\left(\frac{e^{\gamma}+e^{-\gamma}}{e^{\gamma}-e^{-\gamma}}-\frac{1}{\gamma}\right)=n\mu_l\left(\coth(\gamma)-\frac{1}{\gamma}\right)\end{aligned} \tag{10.2.353}$$

The function:

$$L(\gamma)=\coth(\gamma)-\frac{1}{\gamma} \tag{10.2.354}$$

is often named "Langevin's function" with for recall:

$$\gamma=\frac{\mu_l B}{kT} \tag{10.2.355}$$

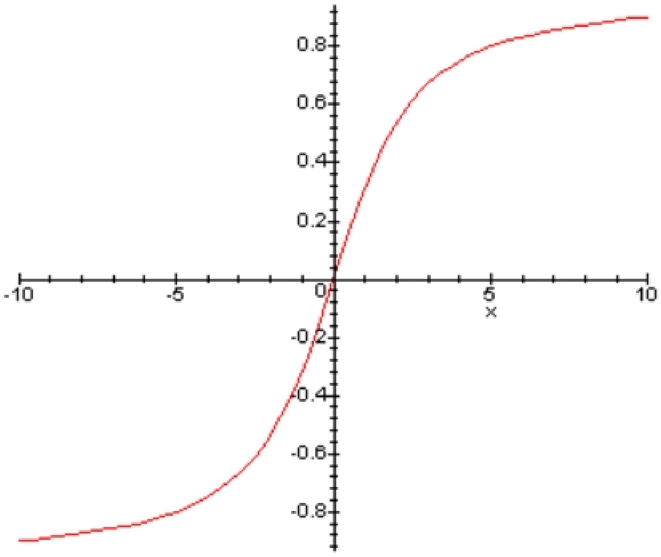


Figure 10.61: **Plot of Langevin's function in Maple™ 4.00b**

The Langevin function is equal 0 when its parameter is 0 and approaches 1 when its argument tends to infinity. So the system eventually saturate when the magnetic field increases, which corresponds to the experimental behaviour of paramagnetic materials. By cons the increase in temperature therefore decrease factor γ and makes the Langevin's function tend to 0 and has for effect to cancel the alignment of dipoles.

For small values of the parameter γ, the Langevin's function can be considered as linear as we see on the plot above. To simplify the expression, we'll calculate the Taylor approximation of the hyperbolic cotangent using detailed proved presented in the section of Sequences and Series and when the hyperbolic cotangent argument is to recall strictly less than 1 in absolute value:

$$\coth(\gamma) \cong \frac{1}{\gamma} + \frac{\gamma}{3} \tag{10.2.356}$$

Then we have:

$$L(\gamma) \cong \frac{\gamma}{3} = \frac{\mu_l B}{3kT} \tag{10.2.357}$$

We have therefore the magnetic field which is proportional to:

$$n\mu_0\langle x\rangle = n\mu_l L(\gamma) \cong n\mu_l \frac{\mu_l B}{3kT} = \frac{n\mu_l^3 B}{3kT} = \frac{n\mu_l^3 \mu_0 H}{3kT} = \frac{n\mu_l^2 \mu_0}{3kT} H \tag{10.2.358}$$

We see that the factor:

$$\frac{n\mu_l^2 \mu_0}{3kT} \tag{10.2.359}$$

is dimensionless. Indeed:

$$\frac{[\mathrm{m}^{-3}][\mathrm{A}\cdot\mathrm{m}^2]^2[\mathrm{kg}\cdot\mathrm{m}\cdot\mathrm{s}^{-2}\cdot\mathrm{A}^{-2}}{[\mathrm{m}^2\cdot\mathrm{kg}\cdot\mathrm{s}^{-2}\cdot\mathrm{K}^{-1}][\mathrm{K}]} = \frac{[\mathrm{m}^{-3}][\mathrm{A}^2\cdot\mathrm{m}^4][\mathrm{m}\cdot\mathrm{A}^{-2}]}{[\mathrm{m}^2]} = \frac{[\mathrm{m}^2]}{[\mathrm{m}^2]} = [-] \tag{10.2.360}$$

So we can consider it is the paramagnetic susceptibility and write the "Langevin's relation of the paramagnetic susceptibility":

$$\boxed{\chi_p \cong \frac{N\mu_l^2\mu_0}{3kT} = \frac{N\mu_l^2\mu_0}{3k}\frac{1}{T} = \frac{C}{T}} \tag{10.2.361}$$

better known under the name "Curie's law" and shows that the magnetic susceptibility is inversely proportional to the temperature (good but obviously this law becomes false at low temperatures and we must then derive empirically the Curie-Weiss law).

In physics and materials science, the "Curie temperature" T_C, or "Curie point", is the temperature above which certain materials lose their permanent magnetic properties, which can (in most cases) be replaced by induced magnetism.

Caution!! Keep in mind that's an oversimplified model! More realistic models based on quantum theory show that the magnetic susceptibility can increase or decrease in function of the temperature and pressure and therefore that the Curie temperature also varies quite strangely depending on the material property and the pressure following (see [472]):

$$\frac{T_\text{C}(p)}{T_\text{C}\left(P_0\right)} = \left[\exp\left(-\left(\frac{R\left(p, T_\text{C}(p)\right)}{R\left(p_0, T_\text{C}\left(p_0\right)\right)}-1\right)\right)\right]^{\frac{1}{n}} \tag{10.2.362}$$

As a little summary can always be useful here is one:

Property	**Diamagnetic**	**Paramagnetic**	**Ferromagnetic**
Magnetic induction	$B < B_0$	$B > B_0$	$B \gg B_0$
Susceptibility χ	$\chi < 0$	$\chi \geq 0$	$\chi \gg 0\,(10^2 - 10^3)$
Temperature dependence	No	$\chi_d = \dfrac{C}{T}$	$\chi_f = -\dfrac{C}{T - T_c}$
Relative permeability μ_r	$\mu_r < 1$	$\mu_r > 1$	$\mu_r \gg 1\,(10^2 - 10^3)$
Magnetising vector	Opposite to $\vec{H}$	In the direction of $\vec{H}$	In the direction of $\vec{H}$

Table 10.1: **Summary of common magnetic properties**

And here small numeric table of some Curie temperatures at room pressure:

Material	**Chemical Formula**	**Curie temp. [K]**	**Curie temp. [°C]**
Cobalt	Co	1388	1115
Iron	Fe	1043	770
Iron(III) oxide	Fe_2O_3	948	675
Nickel iron oxide	$NiOFe_2O_3$	858	585
Copper iron oxide	$CuOFe_2O_3$	728	455
Magnesium iron oxide	$MgOFe_2O_3$	713	440
Manganese bismuth	MnBi	630	357
Nickel	Ni	627	354
Neodymium-iron-boron	$Nd_2Fe_{14}B$	593	320
Manganese antimonide	MnSb	587	314
Manganese iron oxide	$MnOFe_2O_3$	573	300
Yttrium iron garnet	$Y_3Fe_5O_{12}$	560	287
Chromium(IV) oxide	CrO_2	386	113
Manganese arsenide	MnAs	318	45
Gadolinium	Gd	292	19
Terbium	Tb	219	-54
Dysprosium	Dy	88	-185
Europium(II) oxide	EuO	69	-204

Table 10.2: **Overview of the Curie temperature of various ferromagnetic and ferrimagnetic materials**

∼90 %

★★★★☆

48 votes, 79.17%

10.3. Electrodynamics

Classical electrodynamics is a branch of theoretical physics that studies the interactions between electric charges and currents using an extension of the classical Newtonian model. The theory provides an excellent description of electromagnetic phenomena whenever the relevant length scales and field strengths are large enough that quantum mechanical effects are negligible. For small distances and low field strengths, such interactions are better described by quantum electrodynamics.

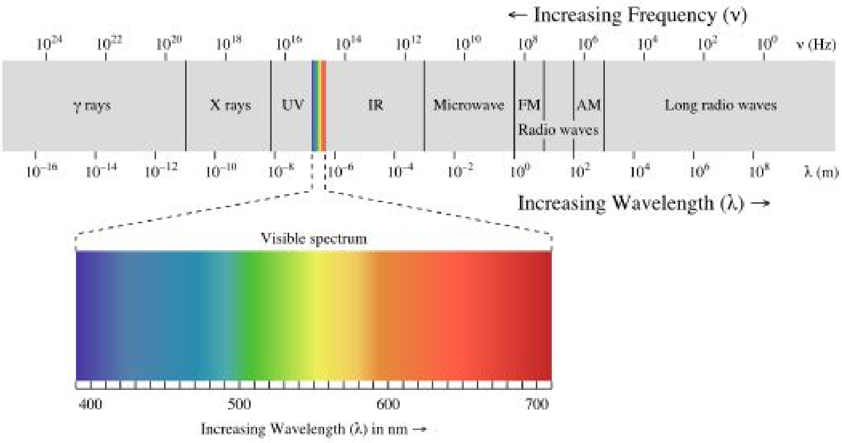
In this section we will build a set of equations that can sum up themselves our entire knowledge about electrostatics and magnetostatics. These equations, at the number of four, are named "Maxwell-Heaviside equations" (denomination that we will shorten as many other books under the name "Maxwell's equations") and will allow us to approach the branch of physics named "electrodynamics" and therefore electromagnetic waves (i.e.: light!).

Electrodynamics is a pillar of the electronic revolution! Without this theory: no radio, no phones or cell phones, no computers, no satellites, no satellites, no electric motor, we would still at the technological state of the late 119th century (holocene calendar) without it!

Remark

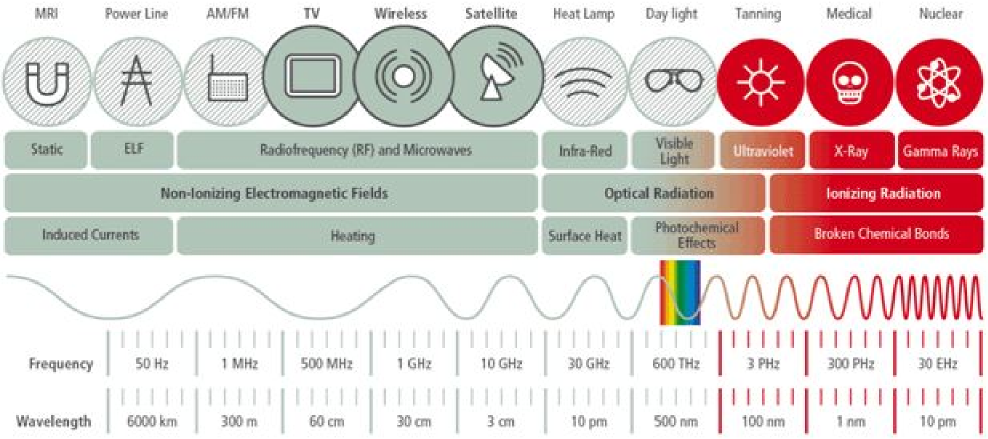
It is very important to understand what will follow! Some developments will be reused in the section of Special Relativity, Quantum Field Theory, etc. Furthermore, the reader should also read in parallel the section of Special Relativity to better understand the ins and outs of certain results and the provenance of some mathematical tools.

We assume before tackling mathematical models that everyone admits at the beginning of this third millennium than gamma rays, radio waves, microwaves, visible light (and not visible light) are electromagnetic waves (EM) of different frequencies:


← Increasing Frequency (ν)
10^{24}
10^{22}
10^{20}
10^{18}
10^{16}
10^{14}
10^{12}
10^{10}
10^{8}
10^{6}
10^{4}
10^{2}
10^{0}
ν (Hz)
γ rays
X rays
UV
IR
Microwave
FM
AM
Radio waves
Long radio waves
10^{-16}
10^{-14}
10^{-12}
10^{-10}
10^{-8}
10^{-6}
10^{-4}
10^{-2}
10^{0}
10^{2}
10^{4}
10^{6}
10^{8}
λ (m)
Increasing Wavelength (λ) →
Visible spectrum
400
500
600
700
Increasing Wavelength (λ) in nm →

With a non-exhaustive summary of the current applications of the wave frequencies in this early

121st century (holocene calendar):


MRI
Power Line
AM/FM
TV
Wireless
Satellite
Heat Lamp
Day light
Tanning
Medical
Nuclear
Static
ELF
Radiofrequency (RF) and Microwaves
Infra-Red
Visible Light
Ultraviolet
X-Ray
Gamma Rays
Non-Ionizing Electromagnetic Fields
Optical Radiation
Ionizing Radiation
Induced Currents
Heating
Surface Heat
Photochemical Effects
Broken Chemical Bonds
Frequency
50 Hz
1 MHz
500 MHz
1 GHz
10 GHz
30 GHz
600 THz
3 PHz
300 PHz
30 EHz
Wavelength
6000 km
300 m
60 cm
30 cm
3 cm
10 pm
500 nm
100 nm
1 nm
10 pm

Frequency allocation to the business economy, civil and military industry is the role of the International Telecommunication Union.

10.3.1 Maxwell Equations

As already mentioned Maxwell's equations are a set of partial differential equations that form the foundation of classical electrodynamics, classical optics, and electric circuits and encompass the knowledge of almost all 119th century (holocene calendar) about the subject. These fields in turn underlie modern electrical and communications technologies. Maxwell's equations describe how electric and magnetic fields are generated and altered by each other and by charges and currents.

If there is a some gods (concept that remains to be proved beyond reasonable evidence) many people say that probably god (we don't know the god of what religion...) wrote the Maxwell Equations first and let there be light: and there was light!

10.3.1.1 First Maxwell Equation (constant electric flow)

Let us define $\vec{\mathcal{E}}$ a vectors field in space. Consider a closed surface S in this field. Then at each point (x, y, z) belonging to the surface corresponds a vector of the field:

In this case the Ostrogradsky's theorem gives us (see section Vector Calculus page 1882):

$$\Phi_{\mathcal{E}} = \oiint_S (\vec{\mathcal{E}} \circ \vec{n})\mathrm{d}S = \iiint_V (\vec{\nabla} \circ \vec{\mathcal{E}})\mathrm{d}V \qquad (10.3.1)$$

with V being the volume delimited by the closed surface S (so-named for recall "Gauss surface").

Remark

The Ostrogradsky theorem is verified since there is no singularities in the volume V!

Before continuing let us also recall that in the case of the Ostrogradsky theorem the normal vector $\vec{n}$ is conventionally directed outwardly from the surface.

In the particular case of an electric field, we get some very interesting results. Indeed, given a charge Q located relatively to a frame by the vector $\vec{r}_Q$. Then, we already saw in the section of Electrostatic that in every point in space, there is a field $\vec{E}$ such that:

$$\vec{E}(\vec{r})=\frac{KQ}{\|\vec{r}-\vec{r}_Q\|^3}(\vec{r}-\vec{r}_Q) \tag{10.3.2}$$

therefore:

$$\vec{E}(x,y,z)=\frac{KQ}{\left[(x-r_{Q_x})^2+(y-r_{Q_y})^2+(z-r_{Q_z})^2\right]^{3/2}}\begin{pmatrix}x-r_{Q_x}\\ y-r_{Q_y}\\ z-r_{Q_z}\end{pmatrix} \tag{10.3.3}$$

As we can see it, the field $\vec{E}$ has a singularity on $(r_{Q_x},r_{Q_y},r_{Q_z})$. Let us consider a Gauss surface such as the charge Q lies outside of this surface! Within the volume V delimited by the surface S the field $\vec{E}$ then has no singularities. So we can calculate the divergence (see section Vector Calculus page 1879) of $\vec{E}$:

$$\begin{aligned}\frac{\partial E_x}{\partial x}(x,y,z)&=\frac{KQ}{\left[(x-r_{Q_x})^2+(y-r_{Q_y})^2+(z-r_{Q_z})^2\right]^{3/2}}-\frac{3KQ(r-r_{Q_x})^2}{\left[(x-r_{Q_x})^2+(y-r_{Q_y})^2+(z-r_{Q_z})^2\right]^{5/2}}\\ \frac{\partial E_y}{\partial y}(x,y,z)&=\frac{KQ}{\left[(x-r_{Q_x})^2+(y-r_{Q_y})^2+(z-r_{Q_z})^2\right]^{3/2}}-\frac{3KQ(r-r_{Q_x})^2}{\left[(y-r_{Q_y})^2+(y-r_{Q_y})^2+(z-r_{Q_z})^2\right]^{5/2}}\\ \frac{\partial E_z}{\partial z}(x,y,z)&=\frac{KQ}{\left[(x-r_{Q_x})^2+(y-r_{Q_y})^2+(z-r_{Q_z})^2\right]^{3/2}}-\frac{3KQ(z-r_{Q_z})^2}{\left[(x-r_{Q_x})^2+(y-r_{Q_y})^2+(z-r_{Q_z})^2\right]^{5/2}}\end{aligned} \tag{10.3.4}$$

Therefore we get:

$$\text{div}(\vec{E})=\vec{\nabla}\circ\vec{E}=\frac{\partial E_x}{\partial x}+\frac{\partial E_y}{\partial y}+\frac{\partial E_z}{\partial z}=\frac{(3KQ-3KQ)\left[(x-r_{Q_x})^2+(y-r_{Q_y})^2+(z-r_{Q_z})^2\right]}{\left[(x-r_{Q_x})^2+(y-r_{Q_y})^2+(z-r_{Q_z})^2\right]^{5/2}} \tag{10.3.5}$$

So if we calculate the flow through this surface we find (see the section of Vector Calculus for a detailed description of the "nabla" operator represented by the symbol "del"):

$$\Phi_E=\iiint_V(\vec{\nabla}\circ\vec{E})\mathrm{d}V=0 \tag{10.3.6}$$

The flow is equal to zero!

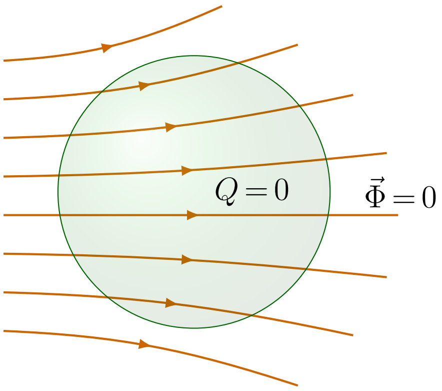
In the case where the charge Q is located inside the Gauss surface $\vec{\nabla}\circ\vec{E}$ is no longer defined on $(r_{Q_x},r_{Q_y},r_{Q_z})$ we have then:

$$\Phi_E=\iiint\limits_{V-B}(\vec{\nabla}\circ\vec{E})\mathrm{d}V+\Phi_B \tag{10.3.7}$$

With Φ_B being the flow $\vec{E}$ on a small ball B around the punctual charge Q.

In this case:

$$\Phi_E = \iiint\limits_{V-B} (\vec{\nabla} \circ \vec{E}) \mathrm{d}V = 0 \tag{10.3.8}$$



As the divergence is defined everywhere on $V-B$. So it remains:

$$\Phi_E = \Phi_B = \oiint_S \vec{E} \circ \mathrm{d}\vec{S} \tag{10.3.9}$$

But in the case of a sphere, it is quite easy to calculate:

$$\oiint_S \vec{E} \circ \mathrm{d}\vec{S} \tag{10.3.10}$$

We have:

$$\Phi_E = \oiint_S E \mathrm{d}S = E \oiint_S \mathrm{d}S = E \cdot S = 4\pi r^2 \left(\frac{1}{4\pi\varepsilon_0} \frac{Q}{r^2} \right) = \frac{Q}{\varepsilon_0} \tag{10.3.11}$$

hence the "first Maxwell equation" or "Gaussian law" (or "Gauss theorem") with a slightly condensed notation:

$$\boxed{\Phi_E = \oint_S \vec{E} \circ \mathrm{d}\vec{S} = \int\limits_V (\vec{\nabla} \circ \vec{E}) \mathrm{d}V = \frac{Q}{\varepsilon_0} \Leftrightarrow \vec{\nabla} \circ \vec{E} = \rho_Q \varepsilon_0^{-1}} \tag{10.3.12}$$

where ρ_Q is the charge density in Coulombs per m^3. On the left we have the integral form of the first Maxwell's (electro-magnetostatics) equation and its differential form on the right.

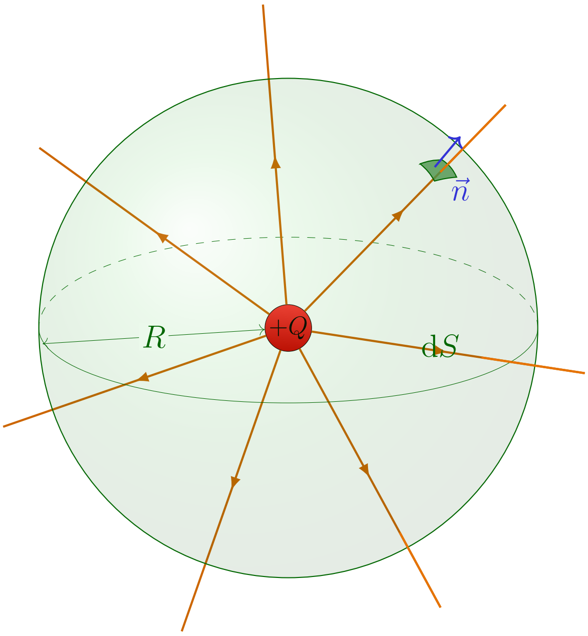


Figure 10.62: **Gauss law (author: Izaak Neutelings)**

This result suggests that the flow of the electric field through a closed surface (hence the circle on the integral) is equal to a given dimensional factor, to the total charge enclosed in the latter surface. Or in other words, the electrostatic field outside a charged sphere is identical to that of a point charge located at the spheres center.

Hence, the reader must keep in mind that the inverse square law applies to point sources. In Classical Electrodynamics there are not point sources (so quantum stuff is not the subject here!), so strictly speaking it never applies. In nature we have extended sources. If we are far away from the source, then it looks like a point source, so the inverse square law is a good approximation. Close to the source, the inverse square law doesn't apply unless the source happens to be perfectly spherical. Even in that case the law breaks down once we cross the surface and are inside the source.

Remark

The integral of the last relation is a line integral (thus evaluated on a curve)! In the field electrodynamics, line integrals apply very often on paths or closed surfaces hence the indication of a circle superimposed on the integral symbol and being named "circulation of the vector field".

If we now express this equation in terms of the electrical potential for which we have proved in the section Electrostatics that:

$$\vec{E}=-\vec{\nabla}\cdot U=-\overrightarrow{\text{grad}}(U) \tag{10.3.13}$$

We get:

$$\vec{\nabla}\circ(\vec{\nabla}\cdot U)=-\frac{\rho_Q}{\varepsilon_0} \tag{10.3.14}$$

We can write the above relation more aesthetically using the scalar Laplacian (see section Vector

Calculus page 1899) such that we get the relation:

$$\boxed{\Delta\cdot\varphi=-\frac{\rho_Q}{\varepsilon_0}} \tag{10.3.15}$$

named "Maxwell-Poisson equation".

10.3.1.2 Second Maxwell Equation (non-existence of magnetic monopole)

In the particular case of a magnetic field, we also get some very interesting results.

Indeed, given a current I located relatively by a reference frame by the vector $\vec{r}_I$. Then at each point $\vec{r}$ in space, we proved in the section Magnetostatics that there is a field $\vec{B}$ given by the Biot-Savart law (see page 4023) such that:

$$\vec{B}(\vec{r})=\underbrace{\frac{\mu_0}{4\pi}}_{=K}\frac{I\cdot\vec{l}\times(\vec{r}-\vec{r}_I)}{\|\vec{r}-\vec{r}_I\|^3} \tag{10.3.16}$$

Therefore:

$$\vec{B}(x,y,z)=\frac{KI}{[(x-r_{I_x})^2+(y-r_{I_y})^2+(z-r_{I_z})^2]^{3/2}}\begin{pmatrix}l_y(z-r_{I_z})-l_z(y-r_{I_y})\\ l_z(x-r_{I_x})-l_x(z-r_{I_z})\\ l_x(y-r_{I_y})-l_y(x-r_{I_x})\end{pmatrix} \tag{10.3.17}$$

As we can see it, the field $\vec{B}$ has a singularity on $(r_{I_x},r_{I_y},r_{I_z})$. Let us consider then a Gaussian surface such that the current I is outside this surface!

Within the volume V delimited by the surface S then field $\vec{B}$ has no singularities anymore. We can therefore calculate the divergence (see section Vector Calculus page 1879) of $\vec{B}$:

$$\begin{aligned}\frac{\partial B_x}{\partial x}(x,y,z)&=-\frac{3KI(x-r_{I_x})\left(l_y(z-r_{I_z})-l_z(y-r_{I_y})\right)}{[(x-r_{I_x})^2+(y-r_{I_y})^2+(z-r_{I_z})^2]^{3/2}}\\ \frac{\partial B_y}{\partial y}(x,y,z)&=-\frac{3KI(y-r_{I_y})\left(l_z(x-r_{I_x})-l_x(z-r_{I_z})\right)}{[(x-r_{I_x})^2+(y-r_{I_y})^2+(z-r_{I_z})^2]^{3/2}}\\ \frac{\partial B_z}{\partial z}(x,y,z)&=-\frac{3KI(z-r_{I_z})\left(l_x(y-r_{I_y})-l_y(x-r_{I_x})\right)}{[(x-r_{I_x})^2+(y-r_{I_y})^2+(z-r_{I_z})^2]^{3/2}}\end{aligned} \tag{10.3.18}$$

Therefore:

$$\mathrm{div}(\vec{B})=\vec{\nabla}\circ\vec{B}=\frac{\partial B_x}{\partial x}+\frac{\partial B_y}{\partial y}+\frac{\partial B_z}{\partial z}=0 \tag{10.3.19}$$

If we calculate the flow through this surface, we then find for the magnetic flux:

$$\Phi_B=\iiint_V(\vec{\nabla}\circ\vec{B})\mathrm{d}V=0 \tag{10.3.20}$$

The flow is equal to zero!

In the case where the current I is located within the Gauss surface, $\vec{\nabla}\circ\vec{B}$ is no longer defined in $(r_{I_x}, r_{I_y}, r_{I_z})$ we have then:

$$\Phi_B=\iiint\limits_{V-B'}(\vec{\nabla}\circ\vec{B})\mathrm{d}V+\Phi_{B'} \tag{10.3.21}$$

with $\Phi_{B'}$ being the flow of $\vec{B}$ on a small sphere B' surrounding partially the straight conductor carrying the current I. In this case:

$$\Phi_B=\iiint\limits_{V-B'}(\vec{\nabla}\circ\vec{B})\mathrm{d}V=0 \tag{10.3.22}$$

as the divergence is defined everywhere on $V-B'$. Therefore it remains only that:

$$\Phi_B=\Phi'_B=\oiint_S\vec{B}\circ\mathrm{d}\vec{S} \tag{10.3.23}$$

But in the case of a sphere, it is easy to calculate:

$$\oiint_S\vec{B}\circ\mathrm{d}\vec{S} \tag{10.3.24}$$

We then have the Gauss's law for the magnetic field:

$$\Phi_B=\oiint_S\vec{B}\circ\mathrm{d}\vec{S}=\oiint_S B\cdot\mathrm{d}S\cos(\phi)=0 \tag{10.3.25}$$

Indeed, in the case of the magnetic field, $\mathrm{d}\vec{S}$ and $\vec{B}$ are therefore perpendicular:

$$\varphi=\frac{\pi}{2}\Rightarrow\cos(\varphi)=0 \tag{10.3.26}$$

Remark

From which we can deduce that $\vec{B}\perp\vec{E}$!

Therefore, given a Gaussian surface in a magnetic field, then the magnetic flux through this surface is equal to:

$$\boxed{\Phi_B=\oiint_S\vec{B}\circ\mathrm{d}\vec{S}=\oiint_V(\vec{\nabla}\circ\vec{B})\mathrm{d}V=0\Leftrightarrow\vec{\nabla}\circ\vec{B}=0} \tag{10.3.27}$$

relation which is the "second Maxwell equation" or "Gauss law for magnetism" (also sometimes named "Gauss theorem for magnetism"). On the left we have the integral form of the fourth Maxwell's equation and its differential form on the right.

This second equation ($\vec{\nabla}\circ\vec{B}=0$) is equivalent to say that there is no "magnetic monopole" in nature, that is to say, at any positive pole, we have to find a negative one (from a magnet the field lines do not diverge). However, the second equation just add the idea (demonstrated by Dirac) that if it was possible to find a monopole in the Nature, it would be the point source of the magnetic field. We will see this a little further in detail below.

10.3.1.3 Third Maxwell Equation

We shall see in the section of Electrokinetics (because we need some concepts that we have not study yet), that the variation of the magnetic flux in time through a conductive loop induces a voltage in the loop given by the "Faraday's law" or "Lenz-Faraday's law":

$$\boxed{U(t)=-\frac{\mathrm{d}\Phi_B}{\mathrm{d}t}} \tag{10.3.28}$$

and we have already proved in the section of Electrostatics that:

$$\Delta U=-\int \vec{E}\circ\mathrm{d}\vec{r}=-\int E\mathrm{d}r \tag{10.3.29}$$

where the last equality is valid only in the special case where the path is collinear to the electric field.

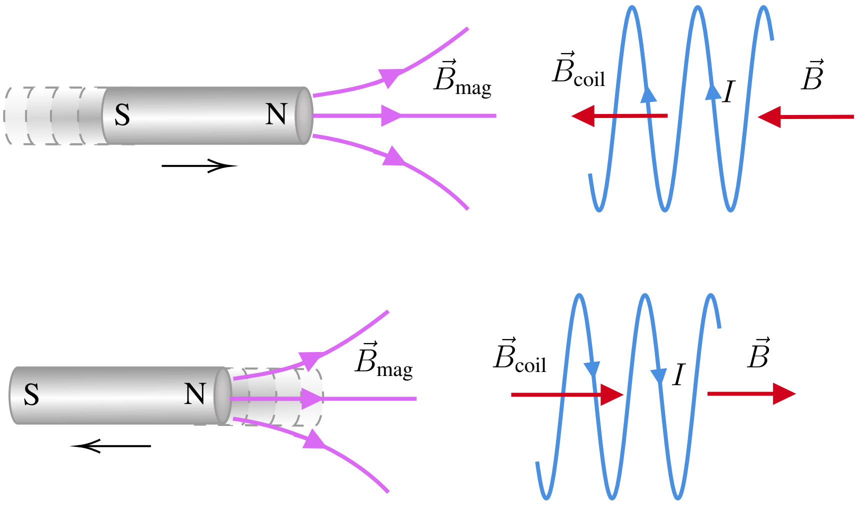


Figure 10.63: Lorenz law illustration

> **Remark**
>
> We will see in the section of Electrokinetics that it is not quite correct to denote the potential with U as above because in fact, Faraday's law expresses the electromotive force EMF (electromotive potential) denoted by e and this potential is not conservative unlike the electrostatic Coulomb potential (for which the integral over a closed path is zero as we have already proved it in the section of Electrostatics).

For an element $\mathrm{d}\vec{l}$ of a circuit Γ it comes:

$$U(t)=\int\limits_{\Gamma}\vec{E}\circ\mathrm{d}\vec{l} \tag{10.3.30}$$

the sign change is here justified by the Lenz law that said that the induced current (and magnetic flow associated with it) has such a direction that it opposes to the change in the flow through the circuit (see figure above).

If we develop this relation, using Stokes' theorem (see section Vector Calculus page 1889) which is for recall:

$$\oint_{\Gamma} \vec{f} \circ \mathrm{d}\vec{r} = \int_{S} (\vec{\nabla} \times \vec{f}) \circ \mathrm{d}\vec{S} \qquad (10.3.31)$$

Then we have:

$$U(t) = -\frac{\mathrm{d}\Phi_B}{\mathrm{d}t} = \oint_{\Gamma} \vec{E} \circ \mathrm{d}\vec{r} = \int_{S} (\vec{\nabla} \times \vec{E}) \circ \mathrm{d}\vec{S} \qquad (10.3.32)$$

Where, as we will see it in the section of Electrokinetics, the electric field above is not the simple Coulomb field but the sum of a Coulomb field and an electromotive field (implicitly generated by Biot-Savart force).

We then have:

$$-\frac{\mathrm{d}\Phi_B}{\mathrm{d}t} = -\frac{\mathrm{d}\displaystyle\int_{S} \vec{B} \circ \mathrm{d}\vec{S}}{\mathrm{d}t} = \int_{S} (\vec{\nabla} \times \vec{E}) \circ \mathrm{d}\vec{S} \qquad (10.3.33)$$

And if the surface element does not move in space and only the magnetic field varies with time, then we have:

$$-\frac{\mathrm{d}}{\mathrm{d}t} \int_{S} \vec{B} \circ \mathrm{d}\vec{S} = -\int_{S} \frac{\partial \vec{B}}{\partial t} \circ \mathrm{d}\vec{S} = \int_{S} (\vec{\nabla} \times \vec{E}) \circ \mathrm{d}\vec{S} \qquad (10.3.34)$$

Therefore:

$$\int_{S} \left(\vec{\nabla} \times \vec{E} + \frac{\partial B}{\partial t} \right) \circ \mathrm{d}\vec{S} = 0 \qquad (10.3.35)$$

A trivial solution is then to say that:

$$\vec{\nabla} \times \vec{E} + \frac{\partial B}{\partial t} = \vec{0} \qquad (10.3.36)$$

We then get finally:

$$\boxed{U(t) = -\frac{\mathrm{d}\Phi_B}{\mathrm{d}t} = \oint_{\Gamma} \vec{E} \circ \mathrm{d}\vec{l} \Leftrightarrow \vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t}} \qquad (10.3.37)$$

This is the "third Maxwell equation" or "Faraday-Maxwell law" sometimes still named "induction law". On the left we have therefore the integral form of the third Maxwell's equation and its differential form on the right.

The third equation therefore affirms that a variation of the magnetic field produces an electric field in a conductor loop. We then say that the term with the partial derivative of the magnetic field is the "magnetic coupling term". This equation is based on the Faraday's theory.

Remark

Often in the scientific literature, the potential $U(t)$ is simply denoted by a tiny u.

Faraday's law of induction is typically used by small portable devices such as below PEG (Personal Energy Generator) to charge portable electronic devices:

Figure 10.64: **Photo of a PEG (right) with a mobile phone (left)**

10.3.1.3.1 Betatron

Among the many examples of application of the third law that we will see in other sections of this textbook, there is one that is particularly nice because it is reminiscent of modern physics on a large scale (even if in reality we are very far).

Thanks to the Maxwell equations and the relations proved in the section Magnetostatics, we can make a small non-exhaustive theoretical study of the physical principle underlying one of the oldest non-linear particles accelerator.

One of the first non-linear methods that comes to mind is to accelerate a charged particle through magnetic induction. This type of accelerator is named a "betatron" (the idea is that it accelerates the electrons as fast as the one that appears in the beta decay...) and was conceptualized in the 11930s (holocene calendar).

The betatron is a particle accelerator that injects electrons into a vacuum torus (in white on the photo below) submitted to a magnetic field that will be considered here as homogeneous between the two magnets (in red in the photo below) to obtain intense X-rays or gamma rays useful in certain professional applications (medicine, analysis of structures, etc.). This accelerator is limited by the magnetic field that it can produce or support.

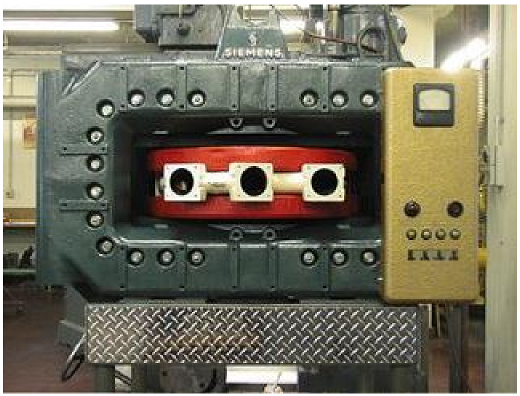


Figure 10.65: **Photo of a Betatron**

For this theoretical study, we will first use the result proved in the section Magnetostatics when we study the Larmor radius: an electron moving in a magnetic field will have a circular path which is perpendicular to the magnetic field.

Then we will also need the third Maxwell equation in the integral form:

$$U(t) = -\frac{\mathrm{d}\Phi_B}{\mathrm{d}t} = \int_\Gamma \vec{E} \circ \mathrm{d}\vec{l} \Leftrightarrow \vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \tag{10.3.38}$$

which - for recall - said that a change in the magnetic field produces an electric field in a conductor loop (or in a charged particle movement that can be assimilated to a conductor loop!!!!).

We have:

$$\int_\Gamma \vec{E} \circ \mathrm{d}\vec{L} = -\frac{\mathrm{d}\Phi_B}{\mathrm{d}t} = -\frac{\mathrm{d}}{\mathrm{d}t}\left(\oiint_S \vec{B} \circ \mathrm{d}\vec{S}\right) \tag{10.3.39}$$

and as the path is circular in the betatron as we have proved in our study of the Larmor radius in the section Magnetostatics, we have:

$$\int_\Gamma \vec{E} \circ \mathrm{d}\vec{l} = \int_\Gamma \|\vec{E}\| \cos(\alpha) \mathrm{d}l \tag{10.3.40}$$

Now, as the electric field is tangential to the circular path of the electrons and that they go in the opposite direction of that field (direction that is constant at any point of the path), we have since electrons travel in circles:

$$\int_\Gamma \vec{E} \circ \mathrm{d}\vec{l} = \int_\Gamma \|\vec{E}\| \cos(\alpha) \mathrm{d}l = \|\vec{E}\| \int_\Gamma \cos(\pi) \mathrm{d}l = -\|\vec{E}\| \int_\Gamma \mathrm{d}l = -2\pi R_L \|\vec{E}\| \tag{10.3.41}$$

But we also have:

$$-\frac{\mathrm{d}}{\mathrm{d}t}\left(\oiint_S \vec{B} \circ \mathrm{d}\vec{S}\right) = -\pi R_L^2 \frac{\mathrm{d}}{\mathrm{d}t}\|\vec{B}\| \tag{10.3.42}$$

Then it comes:

$$-2\pi R_L \|\vec{E}\| = -\pi R_L^2 \frac{\mathrm{d}}{\mathrm{d}t}\|\vec{B}\| \tag{10.3.43}$$

therefore:

$$\|\vec{E}\|=\frac{R_L}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\vec{B}\| \tag{10.3.44}$$

We would like to calculate the kinetic energy that the negatively charged particle acquires after several rounds. This latter is equal to the work done by the electric field to move the load on the circular path (remember that the magnetic field don't do "work"!!!).

As we have proved in the section of Electrostatics, we have along an electric (constant) field line:

$$\Delta W=qE\Delta x \tag{10.3.45}$$

Therefore it comes as the charge travels N times the circumference of the betatron:

$$\Delta E_c=\Delta W=qE\Delta x=N2\pi R_L q\|\vec{E}\|=N2\pi R_L q\frac{R_L}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\vec{B}\|=\pi NqR_L^2\frac{\mathrm{d}}{\mathrm{d}t}\|\vec{B}\| \tag{10.3.46}$$

✍**Example:**

Let us consider a sinusoidal magnetic field of magnitude $B_0=1\,[\text{T}]$ at a frequency of $f=50$ [Hz], therefore a period T of 20 [ms]. This means that in 5 [ms] the magnetic field passes from a maximum to a zero value.

Let us also consider that we have a betatron with a circular path of 1 [m] and the electron can remain approximately $480,000$ rounds on this circular trajectory with this specific radius without deviating too much (the equivalent of about $3,000$ [km] travelled). The electron is injected with an energy of 2 [MeV] in the vacuum torus (energy that is already very close to that of the speed of light!).

Let us first calculate the initial radius for the trajectory according to the relativistic Larmor radius. For this, first we need the speed corresponding to the energy of 2 [MeV]:

$$E_c=mc^2-m_0c^2=(\gamma-1)m_0c^2=\left(\frac{1}{\sqrt{1-\dfrac{v_0^2}{c^2}}}\right)m_0c^2 \tag{10.3.47}$$

After some elementary algebraic operations, we find:

$$v_0=299,407,509\cong 99.87\%\cdot c \tag{10.3.48}$$

Therefore the initial Larmor radius is equal to:

$$R_L=\frac{\gamma m_0v_0}{qB_0}=0.007\,[\text{m}] \tag{10.3.49}$$

We then for during all the time of the acceleration that will take the electron to a Larmor radius of 1 [m] a kinetic energy gain of:

$$\Delta E_c=\pi NqR_L^2\frac{\Delta\|\vec{B}\|}{\Delta t}\cong\pi 480,000\cdot 1.6\cdot 10^{-19}\cdot 1^2\cdot\frac{(1-0)}{(0.005-0)}=4.82548632\cdot 10^{-11}\,[\text{J}] \tag{10.3.50}$$

Which corresponds in electron volts to:

$$\Delta E_c \cong \frac{4.82548632 \cdot 10^{-11}}{1.6 \cdot 10^{-19}} \cong 300\,[\text{MeV}] \tag{10.3.51}$$

Hence the energy that was measured experimentally at that time. This energy also corresponds to a speed which is very close to that of light. Thus, with the same calculation as above, we get:

$$v_0 = 299,407,509 \cong 99.87\% \cdot c \tag{10.3.52}$$

speed reached only in a hundredths of a second!

Remark

Therefore in reality, the centrifugal (fictional) force increases gradually as the electron acquires kinetic energy (and therefore the speed). This force must be compensated by increasing the Lorentz force accordingly.

10.3.1.4 Fourth Maxwell Equation

The fourth Maxwell equation is in our point of view the most important one. It is a generalization of the Ampère's law which has already been presented in the section Magnetostatics and for which we got the circulation C_B of the magnetic field:

$$C_B = \int_{\Gamma} \vec{B} \circ \mathrm{d}\vec{l} = \mu_0 I \tag{10.3.53}$$

Remember that the third Maxwell's equation tells us that the variation of a magnetic field gives generated to an electric field. We can therefore assume that the converse may be probably true (need experimentation to be checked!).

A typical place where we can observe a variation of an electric field, for example, is the capacitor (see section Electrokinetics page 3997).

We know that:

$$I = \frac{\mathrm{d}Q}{\mathrm{d}t} \tag{10.3.54}$$

and that the electric field between two parallel planes, of surface S, bearing electric charges $Q_+ = -Q_-$, uniformly distributed is given by (see section Electrostatics page 3997):

$$E(x) = \frac{\sigma}{\varepsilon_0} = \frac{Q}{\varepsilon_0 S} \tag{10.3.55}$$

where σ is the surfacic density.

This result is independent of the distance l between the planes. The first Maxwell equation gives us:

$$\Delta U = -\int\limits_0^l E(x)\mathrm{d}x = -\frac{Q}{\varepsilon S}l \Rightarrow = |\Delta U| = \frac{Ql}{\varepsilon_0 S} \tag{10.3.56}$$

The capacity of a condensator is defined by (see section Electrostatics page 3997):

$$C = \frac{Q}{U} \tag{10.3.57}$$

we obtained in the particular case of plane and parallel capacitor that the capacity was equal to:

$$C = \frac{\varepsilon_0 S}{l} \tag{10.3.58}$$

Therefore it comes:

$$I = \frac{\mathrm{d}Q}{\mathrm{d}t} = \frac{\mathrm{d}Cu}{\mathrm{d}t} = C\frac{\mathrm{d}u}{\mathrm{d}t} \tag{10.3.59}$$

and using the fact that the electrostatic potential is the electric field multiplied by a distance that will be taken in this case as the distance D between the two planes of the capacitor, we have:

$$I = C\frac{\mathrm{d}(l\cdot E)}{\mathrm{d}t} = C\cdot E\frac{\mathrm{d}E}{\mathrm{d}t} = \varepsilon_0 S\frac{\mathrm{d}E}{\mathrm{d}t} \tag{10.3.60}$$

As the electric field may change with time, it is often customary to put a tiny i for the electric current variable (this is a tradition that we will see again the section Electrokinetics) and as between the two capacitor planes there only vacuum, then we speak of "displacement current", which is why the latter relation is often denoted as follows:

$$I_D = i_D = \varepsilon_0 S\frac{\mathrm{d}E}{\mathrm{d}t} \tag{10.3.61}$$

By expressing the above expression by using the surface current density, we get:

$$\vec{j}_D = \varepsilon_0\frac{\mathrm{d}\vec{E}}{\mathrm{d}t} \tag{10.3.62}$$

If the electric field is not uniform in space and thus depends on the spatial coordinates, we'll use the partial derivatives such as:

$$\vec{j}_D = \varepsilon_0\frac{\partial\vec{E}}{\partial t} \tag{10.3.63}$$

The displacement current generates a magnetic field calculable using Ampère's law:

$$C_B = \oint\limits_\Gamma \vec{B}\circ\mathrm{d}\vec{l} = \mu_0 I_D = \mu\iint\limits_\Sigma \vec{j}_D\circ\mathrm{d}\vec{S} = \mu_0\varepsilon_0\iint\limits_\Sigma \frac{\mathrm{d}\vec{E}}{\mathrm{d}t}\circ\mathrm{d}\vec{S} = \mu_0\varepsilon_0\frac{\mathrm{d}}{\mathrm{d}t}\iint\limits_\Sigma \vec{E}\circ\mathrm{d}\vec{S} \tag{10.3.64}$$

In all phenomena where we observe a charge displacement, we can assume that there is creation of a displacement current which is superimposed on the current flow because of capacitive effects in the material. We write therefore:

$$C_B = \oint\limits_\Gamma \vec{B}\circ\mathrm{d}\vec{l} = \mu_0 I_{\text{tot}} = \mu_0(I+I_D) = \mu_0\iint_\Sigma \vec{j}\circ\mathrm{d}\vec{S} + \mu_0\varepsilon_0\frac{\mathrm{d}}{\mathrm{d}t}\iint\limits_\Sigma \vec{E}\circ\mathrm{d}\vec{S} \tag{10.3.65}$$

where we have (recall of the section of Electrostatic and Magnetostatics):

$$\varepsilon=\varepsilon_0\varepsilon \qquad \text{and} \qquad \mu=\mu_0\mu_r \tag{10.3.66}$$

Furthermore, the Stokes theorem provides us that:

$$\int\limits_{\Gamma}\vec{H}\circ\mathrm{d}\vec{l}=\iint_{\Sigma}(\vec{\nabla}\times\vec{H})\circ\mathrm{d}\vec{S} \tag{10.3.67}$$

therefore:

$$\iint\limits_{\Sigma}\left(\vec{j}+\varepsilon\frac{\mathrm{d}\vec{E}}{\mathrm{d}t}\right)\circ\mathrm{d}\vec{S}=\iint_{\Sigma}(\vec{\nabla}\times\vec{H})\circ\mathrm{d}\vec{S} \tag{10.3.68}$$

and we get from this finally:

$$\boxed{\oint\limits_{\Gamma}\vec{B}\circ\mathrm{d}\vec{l}=\mu\iint\limits_{\Sigma}\vec{j}\circ\mathrm{d}\vec{S}+\mu\varepsilon\frac{\mathrm{d}}{\mathrm{d}t}\iint\limits_{\Sigma}\vec{E}\circ\mathrm{d}\vec{S}\Leftrightarrow\vec{\nabla}\times\vec{H}=\vec{j}+\varepsilon\frac{\partial\vec{E}}{\partial t}} \tag{10.3.69}$$

This is the "fourth Maxwell equation" or "Maxwell-Ampère equation". On the left we have the integral form of the fourth Maxwell's equation and its differential form on the right.

The fourth and last Maxwell equation combines the creation of a magnetic field to any variation of an electric field and / or the presence of an electric current (the presence of an electric current being a sufficient condition but not necessary in view of the second term). We then say that the term with the partial derivative of the electric field is the "term of electrical coupling".

To summarize a bit we have therefore the following four Maxwell equations in time domain[2] named "local forms of Maxwell's equations" in differential form (when the integrals are not indicated):

$$\boxed{\begin{array}{ll}\vec{\nabla}\circ\vec{E}=\dfrac{\rho_Q}{\varepsilon} & \vec{\nabla}\circ\vec{B}=0\\ \vec{\nabla}\times\vec{E}=-\dfrac{\partial\vec{B}}{\partial t} & \vec{\nabla}\times\vec{B}=\mu\vec{j}+\mu\varepsilon\dfrac{\partial\vec{E}}{\partial t}\end{array}} \tag{10.3.70}$$

In the case where $\mu_r,\varepsilon_r\neq 1$, that is to say if we do not work in vacuum but in the matter, we write the local Maxwell equations in the following form:

$$\boxed{\begin{array}{ll}\vec{\nabla}\circ\vec{D}=\rho_Q & \vec{\nabla}\circ\vec{B}=0\\ \vec{\nabla}\times\vec{E}=-\dfrac{\partial\vec{B}}{\partial t} & \vec{\nabla}\times\vec{H}=\vec{j}+\dfrac{\partial\vec{D}}{\partial t}\\ \vec{D}=\varepsilon\vec{E}=\varepsilon_0(1+\chi)\vec{E}=\varepsilon_0\varepsilon_r\vec{E} & \vec{B}=\mu\vec{H}=\mu_0(1+\chi_m)\vec{B}=\mu_0\mu_r\vec{B}\end{array}} \tag{10.3.71}$$

where $\vec{D}$ is (for recall) named the "displacement field" (also named "electric induction" or "electric flux density") and (for recall) $\vec{H}$ is the "magnetic excitation". We also have χ that is the "electric susceptibility" and χ_m the "magnetic susceptibility". Traditionally ε_r is named the "relative permittivity" and χ_r the "relative permeability".

---

[2] In some domains (like plasma physics) it is more useful to express the four equations in the "frequency domain" by taking the Fourier transform of both sides of each equality.

We often found the following notation:

$$\vec{D}=\varepsilon\vec{E}=\varepsilon_0(1+\chi)\vec{E}=\varepsilon_0\vec{E}+\varepsilon_0\chi\vec{E}=\varepsilon_0\vec{E}+\vec{P} \tag{10.3.72}$$

where $\vec{P}$ is named the "polarization density".

Caution!! $\vec{E}$ is a reaction of the vacuum to the field $\vec{D}$. This is due to the permittivity vacuum constant set in the integral (at least that's one way of seeing the thing...).

But in the vacuum and in the case where we consider an absence of charges, we get:

$$\boxed{\begin{array}{ll} \vec{\nabla}\circ\vec{E}=0 & \vec{\nabla}\circ\vec{B}=0 \\ \vec{\nabla}\times\vec{E}=-\dfrac{\partial\vec{B}}{\partial t} & \vec{\nabla}\times\vec{B}=\mu\varepsilon\dfrac{\partial\vec{E}}{\partial t} \end{array}} \tag{10.3.73}$$

This result is important because it expresses the possibility of the propagation of the electric and magnetic field even in the absence of sources!!! We will use these equations to determine the electromagnetic wave equations further below.

Remark

It is possible to express the Maxwell equations under relativistic form... but in reality, as we have already noticed, the equations are unchanged! Indeed, Maxwell's equations are already relativists. This is not surprising, because the vectors of electric and magnetic fields, the photons, travel at the speed of light. At this speed, relativity is the Queen and a correct theory could only be relativistic. However, we can express the latter equations using tensor mathematical notations (see below our demonstration of the tensor of the electromagnetic field). In this form the four equations become incredibly simple and compact (only one extremely short equation). Formulated in this way, the electric and magnetic fields are written as a single field named of course the "electromagnetic field". It is a tensor field as we shall see later.

In integral form the four Maxwell equations are written:

$$\boxed{\begin{array}{ll} \Phi_E=\oint\limits_{\Sigma}\vec{E}\circ\mathrm{d}\vec{S}=\oint\limits_{V}(\vec{\nabla}\circ\vec{E})\mathrm{d}V & \Phi_B=\oiint\limits_{\Sigma}\vec{B}\circ\mathrm{d}\vec{S}=\oiint\limits_{V}(\vec{\nabla}\circ\vec{B})\mathrm{d}V \\ U(t)=-\dfrac{\mathrm{d}\Phi_B}{\mathrm{d}t}=\oint\limits_{\Gamma}\vec{E}\circ\mathrm{d}\vec{l} & \oint\limits_{\Gamma}\vec{B}\circ\mathrm{d}\vec{l}=\mu\iint\limits_{\Sigma}\vec{j}\circ\mathrm{d}\vec{S}+\mu\varepsilon\dfrac{\mathrm{d}}{\mathrm{d}t}\vec{E}\circ\mathrm{d}\vec{S} \end{array}} \tag{10.3.74}$$

For information Maxwell himself didn't like electrodynamics the way he conceived it because he couldn't come up with an underlying mechanical model. Back then, the standard of beauty was a mechanical clockwork universe, but in Maxwell's theory electromagnetic fields just are - they are not made of anything else, no gears or notches, no fluids or valves. Maxwell was unhappy about his own theory because he thought that only « *when a physical phenomenon can* *be completely described as a change in the configuration and motion of a material system, the dynamical explanation of that phenomenon is said to be complete».* Maxwell tried for many years to explain electric and magnetic fields by something that would fit the mechanistic worldview. Alas, in vain.

> Caution!!! There is no principle of physics yet (in the beginning in the 121st century according to holocene calendar) that says physicals laws or constants have to be the same everywhere and always. Indeed, the reader must remember that the "laws" of physics are **descriptive**, not **imperative** and therefore they are some observed regularity in what we can observe experimentally with sufficiently wide occurrence or relevance! Asking whether the laws can change is no more and no less than asking whether regularities which have been observed up till now might not be quite as universal as was thought. The answer to that is: YES! This could drive the reader a bit mad. Indeed, how can we possibly do Science if the rules keep changing? But if you are a professional active scientist, you have to face up to that possibility - and you have a responsibility to check out if there is evidence beyond reasonable doubt about it! We will repeat this warning multiple times in this textbook as its seems that this is something necessary!

10.3.1.5 Magnetic Monopoles

Notice that opting for the natural measurement system where $c=1$, then we have to Maxwell's equations in vacuum:

$$\begin{aligned}\vec{\nabla}\circ\vec{E}=0 \qquad \vec{\nabla}\times\vec{B}-\frac{\partial\vec{E}}{\partial t}=\vec{0}\\ \vec{\nabla}\circ\vec{B}=0 \qquad \vec{\nabla}\times\vec{E}+\frac{\partial\vec{B}}{\partial t}=\vec{0}\end{aligned} \tag{10.3.75}$$

since as we prove it further below, in vacuum:

$$\varepsilon_0\mu_0=\frac{1}{c^2} \tag{10.3.76}$$

Then the transformation:

$$\vec{B}\mapsto\vec{E} \qquad \vec{E}\mapsto-\vec{B} \tag{10.3.77}$$

brings the second pair of preceding equations the first one!!!! This symmetry of Maxwell's equations is named "duality" and that's a clue that suggests that the electric and magnetic fields are only the unified parts of a whole.

Moreover, if we introduce the following complex field:

$$\vec{\mathcal{E}}=\vec{E}+\mathrm{i}\vec{B} \tag{10.3.78}$$

the duality (taking the real part only), then is written:

$$\left.\begin{aligned}\vec{B}\mapsto\vec{E}\\ \vec{E}\mapsto-\vec{B}\end{aligned}\right\}\vec{\mathcal{E}}\mapsto-\mathrm{i}\vec{\mathcal{E}} \tag{10.3.79}$$

the pair of Maxwell's equations indicated above is then reduced to (we use the property of linearity of the cross product) only one pair of equations which we must not forget to take only the real part:

$$\boxed{\vec{\nabla}\circ\vec{\mathcal{E}}=0 \qquad \vec{\nabla}\times\mathcal{E}=\mathrm{i}\frac{\vec{\mathcal{E}}}{\partial t}} \tag{10.3.80}$$

However, this symmetry does not extend to the Maxwell equations with sources expressed in the natural system by:

$$\begin{aligned}\vec{\nabla}\circ\vec{E}&=\rho_Q \qquad & \vec{\nabla}\times\vec{B}-\frac{\partial\vec{E}}{\partial t}&=\vec{j}\\ \vec{\nabla}\circ\vec{B}&=0 \qquad & \vec{\nabla}\times\vec{E}+\frac{\partial\vec{B}}{\partial t}&=\vec{0}\end{aligned} \tag{10.3.81}$$

but half of the time it does not work (make the substitution of $\vec{\mathcal{E}}$ you will see that you always get one of the equations on the pair that is consistent and the other not!). The trick then is to separate the two densities in their respective imaginary and real part (as we can see, generalize any physics theory to complex numbers always bring us to very interesting stuff as for maths!!!):

$$\begin{aligned}\rho_Q&=\rho_e+\mathrm{i}\rho_m\\ \vec{j}&=\vec{j}_e+\mathrm{i}\vec{j}_m\end{aligned} \tag{10.3.82}$$

We then get (always without forgetting to take the real parts and not forgetting that we are in natural units):

$$\begin{aligned}\vec{\nabla}\circ\vec{E}&=\rho_Q \qquad & \vec{\nabla}\times\vec{B}-\frac{\partial\vec{E}}{\partial t}&=\vec{j}_e\\ \vec{\nabla}\circ\vec{B}&=\rho_m \qquad & \vec{\nabla}\times\vec{E}+\frac{\partial\vec{B}}{\partial t}&=\vec{j}_m\end{aligned} \tag{10.3.83}$$

then we simply have to put $\rho_m=\vec{j}_m=0$. These equations are certainly charming but their generalization brings nothing new because no magnetic charge expressed as:

$$\vec{\nabla}\circ\vec{B}=\rho_m \tag{10.3.84}$$

and named "magnetic monopole" was observed to this date. In an experimental context, we say then that $\rho_Q,\vec{j}$ are real such that we have well:

$$\rho_m=\vec{j}_m=0 \tag{10.3.85}$$

The previous system rewritten as following is named "symmetrized Dirac-Maxwell equations" (in natural units):;

$$\begin{aligned}\vec{\nabla}\circ\vec{E}&=\rho_Q \qquad & \vec{\nabla}\circ\vec{B}&=\rho_m\\ \vec{\nabla}\times\vec{E}&=-\frac{\partial\vec{B}}{\partial t}-\vec{j} \qquad & \vec{\nabla}\times\vec{B}&=\vec{j}_e+\frac{\partial\vec{E}}{\partial t}\end{aligned} \tag{10.3.86}$$

To see if we have the right to write the relations above, remember first that if we take the divergence of the curl of the electric field, we know that this operation (see section Vector Calculus page 1911) is always equal to zero regardless of the function considered:

$$\vec{\nabla}\circ(\vec{\nabla}\times\vec{f})=0 \tag{10.3.87}$$

Then we have the following result:

$$\vec{\nabla}\circ(\vec{\nabla}\times\vec{E})=\vec{\nabla}\circ\left(-\frac{\partial\vec{B}}{\partial t}-\vec{j}_m\right)=0 \tag{10.3.88}$$

Hence:

$$-\frac{\partial(\vec{\nabla}\circ\vec{B})}{\partial t}-\vec{\nabla}\circ\vec{j}_m=0 \tag{10.3.89}$$

Therefore:

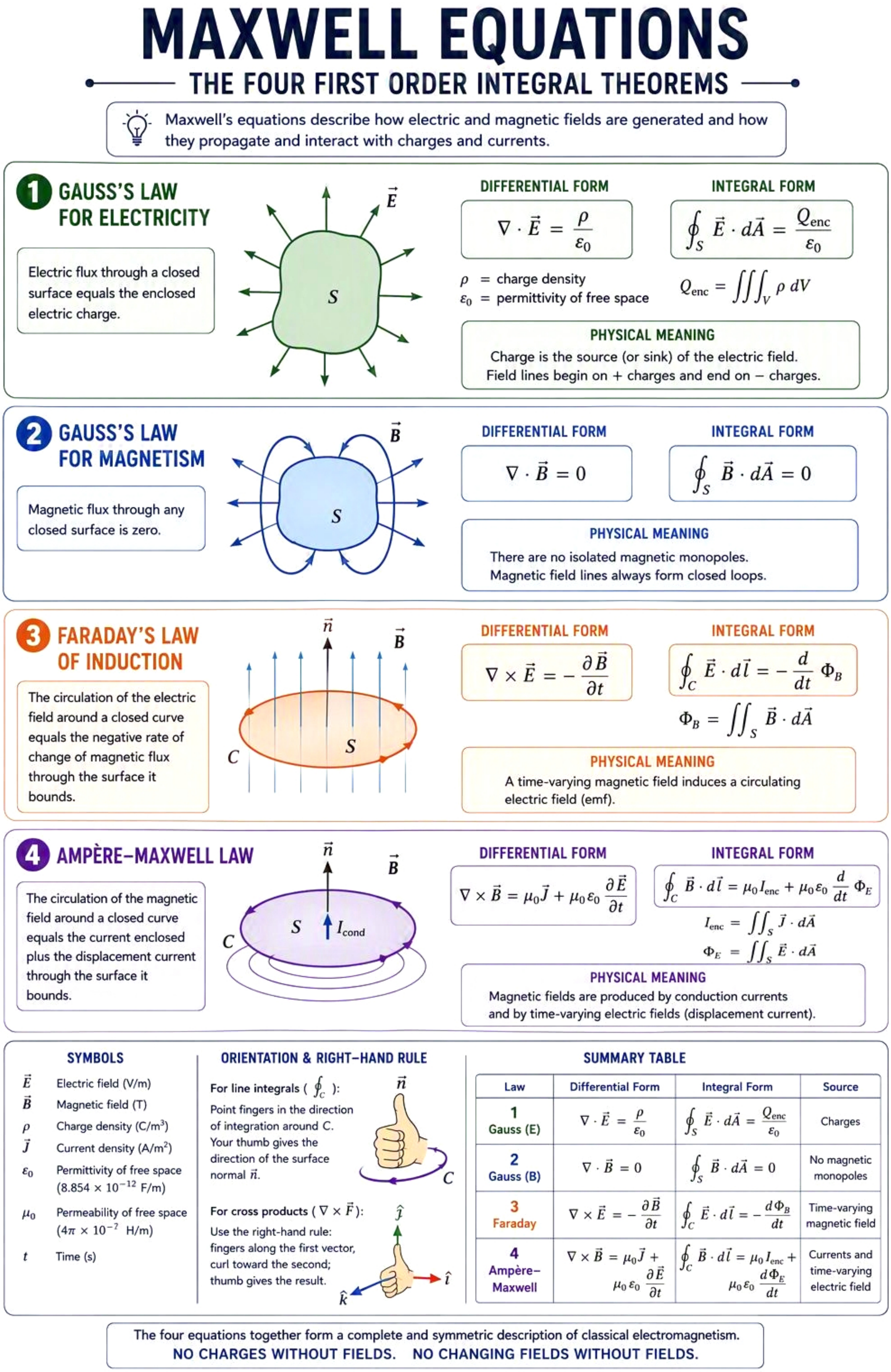
$$-\frac{\partial\rho_m}{\partial t}-\vec{\nabla}\circ\vec{j}_m=0 \tag{10.3.90}$$

Thus after rearrangement:

$$\vec{\nabla}\circ\vec{j}_m+\frac{\partial\rho_m}{\partial t}=0 \tag{10.3.91}$$

So is the fact of falling back on a continuity equation (identical in the form to the thermodynamics continuity equation, fluid mechanics continuity equation, electric field continuity equation or even that of the quantum probability continuity equation, etc.) that would have brought Dirac to complete the four Maxwell equations we have written just above.

Remember that current is the movement of charge. The continuity equation says that if charge is moving out of a differential volume (i.e. divergence of current density is positive) then the amount of charge within that volume is going to decrease, so the rate of change of charge density is negative. Therefore the continuity equation amounts to a conservation of charge.



MAXWELL EQUATIONS

THE FOUR FIRST ORDER INTEGRAL THEOREMS

Maxwell's equations describe how electric and magnetic fields are generated and how they propagate and interact with charges and currents.

1 GAUSS'S LAW FOR ELECTRICITY

Electric flux through a closed surface equals the enclosed electric charge.

DIFFERENTIAL FORM

$$\nabla \cdot \vec{E} = \frac{\rho}{\varepsilon_0}$$

ρ = charge density
ε_0 = permittivity of free space

INTEGRAL FORM

$$\oint_S \vec{E} \cdot d\vec{A} = \frac{Q_{\text{enc}}}{\varepsilon_0}$$

$$Q_{\text{enc}} = \iiint_V \rho \, dV$$

PHYSICAL MEANING

Charge is the source (or sink) of the electric field.
Field lines begin on + charges and end on − charges.

2 GAUSS'S LAW FOR MAGNETISM

Magnetic flux through any closed surface is zero.

DIFFERENTIAL FORM

$$\nabla \cdot \vec{B} = 0$$

INTEGRAL FORM

$$\oint_S \vec{B} \cdot d\vec{A} = 0$$

PHYSICAL MEANING

There are no isolated magnetic monopoles.
Magnetic field lines always form closed loops.

3 FARADAY'S LAW OF INDUCTION

The circulation of the electric field around a closed curve equals the negative rate of change of magnetic flux through the surface it bounds.

DIFFERENTIAL FORM

$$\nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t}$$

INTEGRAL FORM

$$\oint_C \vec{E} \cdot d\vec{l} = -\frac{d}{dt}\Phi_B$$

$$\Phi_B = \iint_S \vec{B} \cdot d\vec{A}$$

PHYSICAL MEANING

A time-varying magnetic field induces a circulating electric field (emf).

4 AMPÈRE–MAXWELL LAW

The circulation of the magnetic field around a closed curve equals the current enclosed plus the displacement current through the surface it bounds.

DIFFERENTIAL FORM

$$\nabla \times \vec{B} = \mu_0 \vec{J} + \mu_0 \varepsilon_0 \frac{\partial \vec{E}}{\partial t}$$

INTEGRAL FORM

$$\oint_C \vec{B} \cdot d\vec{l} = \mu_0 I_{\text{enc}} + \mu_0 \varepsilon_0 \frac{d}{dt}\Phi_E$$

$$I_{\text{enc}} = \iint_S \vec{J} \cdot d\vec{A}$$

$$\Phi_E = \iint_S \vec{E} \cdot d\vec{A}$$

PHYSICAL MEANING

Magnetic fields are produced by conduction currents and by time-varying electric fields (displacement current).

SYMBOLS

$\vec{E}$	Electric field (V/m)
$\vec{B}$	Magnetic field (T)
ρ	Charge density (C/m^3)
$\vec{J}$	Current density (A/m^2)
ε_0	Permittivity of free space (8.854×10^{-12} F/m)
μ_0	Permeability of free space ($4\pi \times 10^{-7}$ H/m)
t	Time (s)

ORIENTATION & RIGHT–HAND RULE

For line integrals ($\oint_C$):

Point fingers in the direction of integration around C. Your thumb gives the direction of the surface normal $\vec{n}$.

For cross products ($\nabla \times \vec{F}$):

Use the right-hand rule: fingers along the first vector, curl toward the second; thumb gives the result.

SUMMARY TABLE

Law	Differential Form	Integral Form	Source
1 Gauss (E)	$\nabla \cdot \vec{E} = \frac{\rho}{\varepsilon_0}$	$\oint_S \vec{E} \cdot d\vec{A} = \frac{Q_{\text{enc}}}{\varepsilon_0}$	Charges
2 Gauss (B)	$\nabla \cdot \vec{B} = 0$	$\oint_S \vec{B} \cdot d\vec{A} = 0$	No magnetic monopoles
3 Faraday	$\nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t}$	$\oint_C \vec{E} \cdot d\vec{l} = -\frac{d\Phi_B}{dt}$	Time-varying magnetic field
4 Ampère–Maxwell	$\nabla \times \vec{B} = \mu_0 \vec{J} + \mu_0 \varepsilon_0 \frac{\partial \vec{E}}{\partial t}$	$\oint_C \vec{B} \cdot d\vec{l} = \mu_0 I_{\text{enc}} + \mu_0 \varepsilon_0 \frac{d\Phi_E}{dt}$	Currents and time-varying electric field

The four equations together form a complete and symmetric description of classical electromagnetism.
NO CHARGES WITHOUT FIELDS. NO CHANGING FIELDS WITHOUT FIELDS.

10.3.2 Charge conservation equation

So we proved so far the four Maxwell's equations which are the foundations of classical electrodynamics.

The Maxwell's equations can be divided into two groups:

1. Equations without sources:

$$\vec{\nabla}\circ\vec{B}=0 \qquad \vec{\nabla}\times\vec{E}=-\frac{\partial\vec{B}}{\partial t} \tag{10.3.92}$$

2. Equations with sources[3] (in vacuum below):

$$\vec{\nabla}\circ\vec{E}=\frac{\rho_Q}{\varepsilon_0} \qquad \vec{\nabla}\times\vec{B}=\mu_0\vec{j}-\frac{1}{c^2}\frac{\partial\vec{E}}{\partial t} \tag{10.3.93}$$

Deriving the first equation with sources in respect to time:

$$\frac{\mathrm{d}(\vec{\nabla}\circ\vec{E})}{\mathrm{d}t}=\vec{\nabla}\circ\frac{\mathrm{d}\vec{E}}{\mathrm{d}t}=\frac{1}{\varepsilon_0}\frac{\mathrm{d}\rho_Q}{\mathrm{d}t} \tag{10.3.94}$$

by taking the divergence of the second we get:

$$\vec{\nabla}\circ(\vec{\nabla}\times\vec{B})=\mu_0\vec{\nabla}\circ\vec{j}+\frac{1}{c^2}\vec{\nabla}\circ\frac{\partial\vec{E}}{\partial t} \tag{10.3.95}$$

The divergence of a curl (rotational) is always zero as we have prove it in the section of Vector Calculus and therefore the last expression is zero. But since a reader asked us, we detail this result more explicitly simplifying a bit:

$$(\vec{\nabla}\circ\vec{\nabla})\times(\vec{\nabla}\circ\vec{B})=\mu_0\vec{\nabla}\circ\vec{j}+\frac{1}{c^2\varepsilon_0}\frac{\partial\rho_Q}{\partial_t} \tag{10.3.96}$$

but, $\vec{\nabla}\circ\vec{B}=0$ and therefore:

$$\mu_0\vec{\nabla}\circ\vec{j}+\frac{1}{c^2\varepsilon_0}\frac{\partial\rho_Q}{\partial t}=0 \tag{10.3.97}$$

After simplification and using natural units (see section Principia page 3124) we get:

$$\boxed{\vec{\nabla}\circ\vec{j}+\frac{\partial\rho_Q}{\partial t}=0} \tag{10.3.98}$$

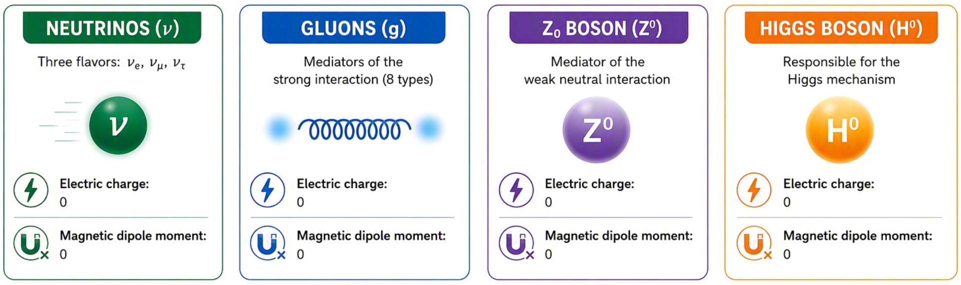
which is named the "charge conservation equation" or "equation of continuity" and said that in two near instants $t+\mathrm{d}t$, the variation $\mathrm{d}Q$ of the charge contained in a closed surface defining a system can only be exclusively attributed to a charge exchange with the outside.

This equation is very important because it involves in the study of relativity, the load is a translation invariant quantity.

---

[3] The fourth Maxwell equation as denoted here is sometimes named the "Oersted's law"

The reader has to know that not everything in the universe interacts with electric and magnetic fields. We have the neutrinos, the Z_0, the Higgs boson and the gluons that are electromagnetically inert particles:


NEUTRINOS (ν)
Three flavors: ν_e, ν_μ, ν_τ
ν
Electric charge: 0
Magnetic dipole moment: 0
GLUONS (g)
Mediators of the strong interaction (8 types)
Electric charge: 0
Magnetic dipole moment: 0
Z_0 BOSON (Z^0)
Mediator of the weak neutral interaction
Z^0
Electric charge: 0
Magnetic dipole moment: 0
HIGGS BOSON (H^0)
Responsible for the Higgs mechanism
H^0
Electric charge: 0
Magnetic dipole moment: 0

10.3.3 Gauge Theory

Before you start reading this subsection it is of primary importance for the reader to go for a ride in the Algebra chapter of this textbook, in which there is a section Vector Calculus where we prove some differential vector operators that are of first importance in physics and especially for gauge theory and where we also study their main properties.

What will follow is very important because besides the fact that we are going to naturally present a new vector field (the vector potential) which is essential in certain equations of relativistic quantum physics (see section of the same name page 4919) we will reuse this gauges approach in the section of Wave Quantum Physics where the consequences are much more important!

Given the known relation:

$$\vec{\nabla}\circ\vec{B}=0 \tag{10.3.99}$$

There exists by the properties of the divergence and curl (rotational) operators (see section Vector Calculus page 1866) a "vector potential $\vec{A}(\vec{r},t)$" such that:

$$\vec{B}=\vec{\nabla}\times\vec{A} \tag{10.3.100}$$

thus satisfying then (the divergence of the curl - rotational - of a vector field is always zero because it is a mathematical property proven in the section of Vector Calculus):

$$\vec{\nabla}\circ(\vec{\nabla}\times\vec{A})=0 \tag{10.3.101}$$

and in the context of magnetostatic named "magnetic potential".

Remark

The vector potential is therefore a potential... and a vector! Similarly as we have define a potential U which derivates from an electric field $\vec{E}$, we can thus define a potential $\vec{A}(\vec{r},t)$ for the field $\vec{B}$. But for technical reasons (coming from the expression of the curl - rotational - of $\vec{E}$ and of $\vec{B}$ in the Maxwell's equations), the potential vector $\vec{A}(\vec{r},t)$ is not that simple as U and can not be expressed as a simple scalar, we must use a vector potential.

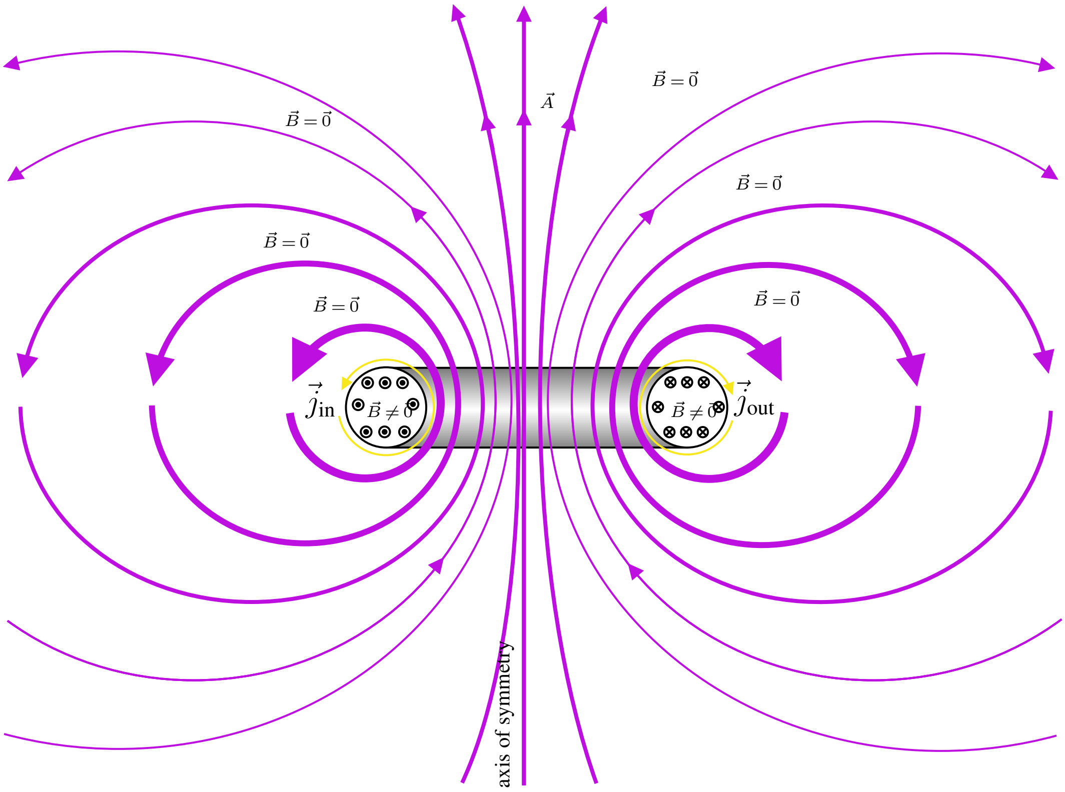

$\vec{A}$
$\vec{B}=\vec{0}$
$\vec{B}=\vec{0}$
$\vec{B}=\vec{0}$
$\vec{B}=\vec{0}$
$\vec{B}=\vec{0}$
$\vec{B}=\vec{0}$
$\vec{j}_{\text{in}}$
$\vec{B}\neq\vec{0}$
$\vec{B}\neq\vec{0}$
$\vec{j}_{\text{out}}$
axis of symmetry

Figure 10.66: **Vector Magnetic Potential $\vec{A}$**

Thus, a depiction of the $\vec{A}$ field around a loop of $\vec{B}$ flow (as would be produced in a toroidal inductor) is qualitatively the same as the $\vec{B}$ field around a loop of current.

The reader will have notice that at the opposite of the electric potential they can be no magnetic field but still a magnetic potential.

If we take the relation $\vec{B}=\vec{\nabla}\times\vec{A}$ in the Maxwell equation $\vec{\nabla}\times\vec{E}=-\partial\vec{B}/\partial t$ we get:

$$\boxed{\vec{\nabla}\times\left(\vec{E}+\frac{\partial\vec{A}}{\partial t}\right)=\vec{0}} \tag{10.3.102}$$

We put now (the notation $\vec{F}$ has no relation with the Newtonian force!):

$$\boxed{\vec{\nabla}\times\left(\vec{E}+\frac{\partial\vec{A}}{\partial t}\right)=\vec{\nabla}\times\vec{F}=\vec{0}} \tag{10.3.103}$$

and we use the mathematical properties of the curl (rotational) and gradient operators to write a new relation (the "$-$" sign is an anticipation of what will follow):

$$\boxed{\vec{F}=-\vec{\nabla}\phi} \tag{10.3.104}$$

hence:

$$\boxed{\vec{E}=-\vec{\nabla}\phi-\frac{\partial\vec{A}}{\partial t}} \tag{10.3.105}$$

where $\phi(\vec{r},t)$ is a "scalar potential" that is well known to us as being the electric potential U.

Remarks

R1. The field $\vec{F}$ seems to obey the same properties as that of the gravitational field (Newton-Poisson law), but it seem to be only a curiosity (units and other mathematical properties are not equivalent).

R2. The reader probably easily see that if the vector potential is zero, then we fall back on (see section Electrostatic page 3977):

$$\vec{E}=-\vec{\nabla}\cdot\phi\equiv\vec{\nabla}\cdot\phi \tag{10.3.106}$$

which reinforces the assumptions of previous developments (and that's not ...)

In addition, the fields $\vec{E}$ and $\vec{B}$ remain unchanged if we do in the previous relations the following replacements (the terms cancel trivially):

$$\phi\rightarrow\phi'=\phi-\frac{\partial\psi}{\partial t}\qquad\vec{A}\rightarrow\vec{A}'=\vec{A}+\vec{\nabla}\psi \tag{10.3.107}$$

where ψ is an arbitrary function of $\vec{r}$ and t.

We name such a transformation a "gauge change". The freedom on the choice of potentials allow us to impose a constraint that we name "gauge constraint".

There are several ways of forming this constraint among which we distinguish two very common in physics!

Thus we will use either the "Lorentz gauge" by imposing:

$$\boxed{\vec{\nabla}\circ\vec{A}+\frac{1}{c^2}\frac{\partial\phi}{\partial t}=0} \tag{10.3.108}$$

or the "Coulomb gauge" by imposing:

$$\boxed{\vec{\nabla}\circ\vec{A}=0} \tag{10.3.109}$$

holds. Thus, ψ should satisfy:

$$\begin{gathered}\vec{\nabla}\circ\vec{A}'=\vec{\nabla}\circ\vec{A}+\vec{\nabla}\circ\vec{\nabla}\psi\Rightarrow\vec{\nabla}\circ\vec{A}'-\vec{\nabla}\circ\vec{\nabla}\psi=\vec{\nabla}\circ\vec{A}=0\\ \Rightarrow\vec{\nabla}\circ\vec{A}'=\vec{\nabla}\circ\vec{\nabla}\psi=\Delta\psi=f(\vec{r},t)\end{gathered} \tag{10.3.110}$$

The relation:

$$\boxed{\vec{\nabla}\circ\vec{A}'=\Delta\psi} \tag{10.3.111}$$

which is named "Poisson equation of the potential vector".

Similarly, to show that it is always possible to impose the Lorenz condition, we simply need to find ψ in the above equations:

$$\phi\rightarrow\phi'=\phi-\frac{\partial\psi}{\partial t}\qquad\vec{A}'\rightarrow\vec{A}+\vec{\nabla}\psi \tag{10.3.112}$$

such that the relation (Lorenz gauge):

$$\boxed{\vec{\nabla}\circ\vec{A}+\frac{1}{c^2}\frac{\partial\phi}{\partial t}=0} \tag{10.3.113}$$

is satisfied. Therefore, ψ must satisfy:

$$\begin{gathered}\vec{\nabla}\circ\vec{A}'+\frac{1}{c^2}\frac{\partial\phi'}{\partial t}=\vec{\nabla}\circ\vec{A}+\vec{\nabla}\circ\vec{\nabla}\psi+\frac{1}{c^2}\frac{\partial\phi}{\partial t}-\frac{1}{c^2}\frac{\partial^2\psi}{\partial t^2}=\vec{\nabla}\circ\vec{A}+\Delta\psi+\frac{1}{c^2}\frac{\partial\phi}{\partial t}-\frac{1}{c^2}\frac{\partial^2\psi}{\partial t^2}\\ \Rightarrow\vec{\nabla}\circ\vec{A}+\frac{1}{c^2}\frac{\partial\phi'}{\partial t}=\Delta\psi-\frac{1}{c^2}\frac{\partial^2\psi}{\partial t^2}\end{gathered} \tag{10.3.114}$$

Or in other words and in a more condensed manner:

$$\Box\psi=\frac{1}{c^2}\frac{\partial\phi'}{\partial t}+\vec{\nabla}\circ\vec{A}'=g(\vec{r},t) \tag{10.3.115}$$

where the operator:

$$\boxed{\Box=\frac{1}{c^2}\frac{\partial^2}{\partial t^2}-\Delta} \tag{10.3.116}$$

is by definition named the "Alembertian" (we will often encounter this term from now as well in the section of Electrodynamics and Wave Quantum Physics) that is also invariant under Lorentz transformation as we will prove it in the section of Special Relativity.

Reporting the equations:

$$\vec{B}=\vec{\nabla}\times\vec{A}\quad\text{and}\quad\vec{E}=-\vec{\nabla}\phi-\frac{\partial\vec{A}}{\partial t} \tag{10.3.117}$$

in the other two Maxwell equations in vacuum:

$$\vec{\nabla}\circ\vec{E}=\frac{\rho_Q}{\varepsilon_0}\quad\text{and}\quad\vec{\nabla}\times\vec{B}=\mu_0\vec{j}+\vec{1}c^2\frac{\partial\vec{B}}{\partial t} \tag{10.3.118}$$

we get, by making appear the Laplacian of a vector field $\Delta\vec{A}$ by on of the properties of the curl, gradient and divergence vector operators (see section Vector Calculus page 1911):

$$\vec{\nabla}\times(\vec{\nabla}\times\vec{A})=\vec{\nabla}(\vec{\nabla}\circ\vec{A})-\Delta\vec{A} \tag{10.3.119}$$

the following relations:

$$\boxed{\begin{gathered}-\Delta\phi=\frac{\rho_q}{\varepsilon_0}+\vec{\nabla}\circ\frac{\partial\vec{A}}{\partial t}\\ \Box\vec{A}=\mu_0\vec{j}-\vec{\nabla}\left(\vec{\nabla}\circ\vec{A}+\frac{1}{c^2}\frac{\partial\phi}{\partial t}\right)\end{gathered}} \tag{10.3.120}$$

the latter relation being named the "arbitrary gauge".

For the Lorenz gauge, these last two equation simplify to (feel free to contact us if you do not see how):

$$\boxed{\begin{aligned}-\Box\phi&=\frac{\rho_q}{\varepsilon_0}\\ \Box\vec{A}&=\mu_0\vec{j}\end{aligned}}\tag{10.3.121}$$

that we name "wave equations for electromagnetic potentials" in analogy with the wave equations of electric and magnetic fields that we will determine further below. For the Coulomb gauge, the same equations simplify to:

$$\boxed{\begin{aligned}-\Box\phi&=-\frac{\rho_q}{\varepsilon_0}\\ \Box\vec{A}&=\mu_0\vec{j}-\frac{1}{c^2}\vec{\nabla}\frac{\partial\phi}{\partial t}\end{aligned}}\tag{10.3.122}$$

knowing that $c^2=1/(\varepsilon_0\mu_0)$ we can write the two wave equations of electromagnetic potentials in the form:

$$\begin{aligned}-\Delta\phi&=\frac{\rho_q}{\varepsilon_0}=\mu_0c^2\rho_Q\\ \Box\vec{A}&=\mu_0\vec{j}\end{aligned}\tag{10.3.123}$$

Let us now put $A^0=\phi/c$ (to homogenize the units) such that we define a "four-vector potential" which allows us to write vectorially the two above relations in unified manner:

$$\boxed{\Box A^\alpha=(\phi/c,\mu_0\vec{j})=(m_0c\rho_Q,\mu_0\vec{j})=\mu_0(c\rho_Q,\vec{j})}\tag{10.3.124}$$

Remark

The fact that the Alembertian of the four-vector potential is expressed from the four-vector current which is contravariant (see section Special Relativity page 5646) bring us to put that the four-vector potential is itself contravariant!

Relation that we will we denote is a more condense form as follows:

$$\boxed{\Box A^\alpha=\mu_0j^\alpha}\tag{10.3.125}$$

where j^α is name "four-vector current".

Remark

We will see again this four-vector during our determination of the electromagnetic field tensor further below (except that we will be in natural units but this does not change the idea...).

The four-vector potential as defined above leads us to be able to write the (quadrivergence) Lorenz gauge by making use of tensor notation:

$$\begin{aligned}\vec{\nabla}\circ\vec{A}+\frac{1}{c^2}\frac{\partial\phi}{\partial t}&=\frac{1}{c^2}\frac{\partial}{\partial t}\phi+\partial_xA_x+\partial_yA_y+\partial_zA_z\\ &=\frac{1}{c}\frac{\partial}{\partial t}\frac{\phi}{c}+\partial_1A^1+\partial_2A^2+\partial_3A^3=\partial_0A^0+\partial_1A^1+\partial_2A^2+\partial_3A^3=0\end{aligned}\tag{10.3.126}$$

Which ultimately allows you to write the Lorenz gauge in the covariant form below:

$$\boxed{\partial_\alpha A^\alpha = 0} \tag{10.3.127}$$

It is therefore an equation of the form of the Klein-Gordon equation for a massless particle (see section Relativistic Quantum Physics page 4921). So we can say in a sense that the invariance of the electromagnetic gauge is connected to the fact that the mass of the photon is zero!

Remark

It is useful to notice that the fact that writing $\dfrac{1}{c}\dfrac{\partial}{\partial t} = \partial_0$ (with or without natural units where $c = 1$) is a notation that will also be adopted in our study of the Dirac equation (see section Relativistic Quantum Physics page 4938) or also in quantum field theory (except that there will be an imaginary part).

These notations finally lead us to be able to write:

$$\begin{aligned}\partial_\alpha(\Box A^\alpha) &= \partial_\alpha(\mu_0 j^\alpha)\\ \partial_\alpha A^\alpha &= \mu_0 \partial_\alpha j^\alpha = 0\end{aligned} \tag{10.3.128}$$

We obtain the continuity equation:

$$\boxed{\partial_\alpha j^\alpha = 0} \tag{10.3.129}$$

that is tensor equivalent form of the following relation (see prof earlier above):

$$\vec{\nabla} \circ \vec{j} + \frac{\partial \rho_Q}{\partial t} = 0 \tag{10.3.130}$$

To summarize roughly:

A given number of physical effects are modelized, depending on the case, by fields that can be scalar, vector, tensor, or spinor fields that we therefore name "gauges". A given number of physical phenomena seems to comply with conditions of say of "symmetry", vis-à-vis these gauges. This symmetry is expressed by what we therefore name a "gauge invariance".

For example, the field that modelize well the electromagnetic field is as we have seen it, a four-vectors field consisting of a scalar potential ϕ (whose gradient is the electric field $\vec{E}$) and a vector-potential $\vec{A}$ (whose curl is the magnetic field $\vec{B}$). It is this quadrivectoriel field that helps us to modelize the electromagnetic field that is named "gauge".

It turns out that we were getting exactly the same physical effects on a system of electric charged particles if we replace this gauge by another gauge by adding it a gauge constraint (typical example between the Lorenz gauge or Coulomb gauge as seen above). The invariance of the laws of physics when switching from a gauge to another is a "gauge invariance". In the case of the electromagnetic field, this gauge invariance turns expressing the electric charge conservation (as we have just proved it).

Mathematically, such gauges changes appear to be the result of the action of a symmetry group of infinite dimensions (transforming these into each other gauges) that we name the "gauge group" of the considered interaction (here the electromagnetic interaction).

For the gravitational field, for example (see section General Relativity page 5720), the gravitational interaction is modelled by a symmetric tensor field of rank 2 with a given signature. This metric field is distributed on a 4D variety modelling the space-time (for a naive and technical definition of "space-time" see the section of Special Relativity at page 5708). This is the gauge of the gravitational interaction. According to General Relativity (the equivalence principle) we do not change anything to the gravitational interaction if we change the system of space-time coordinates in which we express the metric. The portion of an expression of the metric to another by changing coordinate system is also a gauge change. The invariance of General Relativity gauge then expressed the possibility to change to a gauge to another without changing the geodesic followed by the testing particles falling freely in the gravitational field modelled by the metric field.

The invariance of General Relativity gauge is what we named "diffeomorphism invariance" (bijective change of coordinate system with some degree of regularity) and the General Relativity gauge group is the "diffeomorphisms group" of $\mathbb{R}^2$ (named the "soft group").

It should be noted also that the potential-vector $\vec{A}$ is perhaps not so virtual as it may seem. Indeed, it is possible to modify the trajectories of charged particles passing outside a cylindrical volume where there is a magnetic field $\vec{B}$ induced by an electric current (travelling in the winding of a solenoid where this field $\vec{B}$ is "trapped"). It is therefore possible to influence the particle trajectory circulating in an area where the magnetic field $\vec{B}$ is zero but where its vector potential equation is not!!!

Furthermore, we will use the results we just get here for our study of the Yang-Mills theory in our way to the electroweak unification (see the Standard Model in the section of Field Quantum Theory).

Remark

The well-known experiment that involves the vector potential is the Aharonov-Bohm experiment. The "Aharonov–Bohm effect", sometimes named the "Ehrenberg–Siday–Aharonov–Bohm effect", is a quantum mechanical phenomenon in which an electrically charged particle is affected by an electromagnetic potential $(U, \vec{A})$, despite being confined to a region in which both the magnetic field $\vec{B}$ and electric field $\vec{E}$ are zero. The underlying mechanism is the coupling of the electromagnetic potential with the complex phase of a charged particle's wave function, and the Aharonov–Bohm effect is accordingly illustrated by interference experiments.

10.3.3.1 Electromagnetic field tensor

In electromagnetism, the "electromagnetic tensor" or "electromagnetic field tensor" (sometimes named the "field strength tensor", "Faraday tensor" or "Maxwell bivector") is a mathematical object that describes the electromagnetic field in space-time of a physical system. The field tensor was first used after the 4-dimensional tensor formulation of special relativity was introduced by Hermann Minkowski. The tensor allows some physical laws to be written in a very concise form.

To determine the electromagnetic field tensor let us assume at first that the action (see section

Analytical Mechanics page 3161) of an electric charged particle in an electromagnetic field is given by (a priori empirical choice... but you'll see a little further):

$$\boxed{S_1=\int\left(q\vec{A}\circ\vec{v}-q\phi\right)\mathrm{d}t=\int L_1\mathrm{d}t}\tag{10.3.131}$$

Remark

The notation S_0 is reserved to the action of a free particle (see section Special Relativity page 5681).

The Lagrangian for an electric charged particle in an electromagnetic field is therefore the sum of the Lagrangian of the particle interacting with the electromagnetic field L_1 added to the Lagrangian of the free particle L_0 (see section Special Relativity page 5681):

$$\boxed{L=L_0+L_1=-m_0c^2\sqrt{1-\frac{v^2}{c^2}}+q\vec{A}\circ\vec{v}-q\phi}\tag{10.3.132}$$

Remark

This is therefore the Lagrangian of the interaction of the particle with the field added to the Lagrangian of the mass of the particle. Thus we see that it still missing the Lagrangian of the electromagnetic field itself in the case of absence of electric charges (named: "Lagrangian of the free field") but we will see it further below.

This is therefore (a priori) the Lagrangian of a charged particle in an electromagnetic field having for source a given quantity of electric charges in relative movement.

We will now prove that this Lagrangian is correct (the previous relation could be stated as a theorem)!

For this remember first that the general momentum is (see section Analytical Mechanics page 3180 and Special Relativity page 5681):

$$\vec{p}=\frac{\partial L}{\partial\vec{v}}=\frac{m_0\vec{v}}{\sqrt{1+\dfrac{v^2}{c^2}}}+q\vec{A}\tag{10.3.133}$$

To verify that we made the right choice of Lagrangian initially, we will obtain the equations of motion and ensure that they coincide with the Lorentz force. The Lagrange equations are in this case:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial\vec{v}}-\frac{\partial L}{\partial\vec{r}}=0&=\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{m_0\vec{v}}{\sqrt{1+\dfrac{v^2}{c^2}}}+q\vec{A}\right)-\frac{\partial}{\partial\vec{r}}\left(-m_0c\sqrt{1-\frac{v^2}{c^2}}+q\vec{A}\circ\vec{v}-q\phi\right)\\&=\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{m_0\vec{v}}{\sqrt{1+\dfrac{v^2}{c^2}}}\right)+q\dot{\vec{A}}-q\vec{\nabla}\cdot(\vec{A}\circ\vec{v})+q\vec{\nabla}\cdot\phi\end{aligned}\tag{10.3.134}$$

But we have:

$$\vec{F}=\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{m_0\vec{v}}{\sqrt{1+\dfrac{v^2}{c^2}}}\right) \tag{10.3.135}$$

and therefore:

$$\vec{F}=-q\dot{\vec{A}}+q\vec{\nabla}\cdot\left(\vec{A}\circ\vec{v}\right)-q\vec{\nabla}\cdot\phi \tag{10.3.136}$$

But we pointed out in the definition of the scalar potential that $\vec{E}=-\vec{\nabla}\cdot\phi$ hence:

$$\vec{F}=q\vec{E}+q\left(\vec{\nabla}\left(\vec{A}\circ\vec{v}\right)-\dot{\vec{A}}\right) \tag{10.3.137}$$

We should necessarily have by analogy with the Lorentz force:

$$\vec{\nabla}\cdot(\vec{A}\circ\vec{v})-\dot{\vec{A}}=\vec{v}\times\vec{B}=\vec{v}\times\left(\vec{v}\times\vec{A}\right) \tag{10.3.138}$$

So we need before proceeding to check that:

$$\vec{\nabla}\cdot\left(\vec{A}\circ\vec{v}\right)-\frac{\mathrm{d}\vec{A}}{\mathrm{d}t}=\vec{v}\times\left(\vec{\nabla}\times\vec{A}\right) \tag{10.3.139}$$

With:

$$\mathrm{d}\vec{A}=\frac{\partial\vec{A}}{\partial x}\mathrm{d}x+\frac{\partial\vec{A}}{\partial y}\mathrm{d}y+\frac{\partial\vec{A}}{\partial z}\mathrm{d}z\Rightarrow\frac{\mathrm{d}\vec{A}}{\mathrm{d}t}=\frac{\partial\vec{A}}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t}+\frac{\partial\vec{A}}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t}+\frac{\partial\vec{A}}{\partial z}\frac{\mathrm{d}z}{\mathrm{d}t}=\frac{\partial\vec{A}}{\partial x}v_x+\frac{\partial\vec{A}}{\partial y}v_y+\frac{\partial\vec{A}}{\partial z}v_z \tag{10.3.140}$$

In components:

$$\frac{\mathrm{d}\vec{A}}{\mathrm{d}t}=\begin{pmatrix}v_x\partial_xA_x+v_y\partial_yA_x+v_z\partial_zA_x\\v_x\partial_xA_y+v_y\partial_yA_y+v_z\partial_zA_y\\v_x\partial_xA_z+v_y\partial_yA_z+v_z\partial_zA_z\end{pmatrix} \tag{10.3.141}$$

Therefore:

$$\begin{pmatrix}\dfrac{\partial}{\partial x}A_xv_x+\dfrac{\partial}{\partial x}A_yv_y+\dfrac{\partial}{\partial x}A_zv_z\\\dfrac{\partial}{\partial y}A_xv_x+\dfrac{\partial}{\partial y}A_yv_y+\dfrac{\partial}{\partial y}A_zv_z\\\dfrac{\partial}{\partial z}A_xv_x+\dfrac{\partial}{\partial z}A_yv_y+\dfrac{\partial}{\partial z}A_zv_z\end{pmatrix}-\dot{\vec{A}}=\vec{v}\times\begin{pmatrix}\partial_yA_z-\partial_zA_y\\\partial_zA_x-\partial_xA_z\\\partial_xA_y-\partial_yA_x\end{pmatrix} \tag{10.3.142}$$

That is:

$$\begin{pmatrix}v_x\partial_xA_x+v_y\partial_xA_y+v_z\partial_xA_z\\v_x\partial_yA_x+v_y\partial_yA_y+v_z\partial_yA_z\\v_x\partial_zA_x+v_y\partial_zA_y+v_z\partial_zA_z\end{pmatrix}=\begin{pmatrix}v_y(\partial_xA_y-\partial_xA_y)-v_z(\partial_zA_x-\partial_xA_z)\\v_z(\partial_yA_z-\partial_zA_y)-v_x(\partial_xA_y-\partial_yA_x)\\v_x(\partial_zA_x-\partial_xA_z)-v_y(\partial_yA_z-\partial_zA_y)\end{pmatrix}+\dot{\vec{A}} \tag{10.3.143}$$

Therefore, distributing the terms:

$$\begin{pmatrix}v_x\partial_xA_x+v_y\partial_xA_y+v_z\partial_xA_z\\v_x\partial_yA_x+v_y\partial_yA_y+v_z\partial_yA_z\\v_x\partial_zA_x+v_y\partial_zA_y+v_z\partial_zA_z\end{pmatrix}=\begin{pmatrix}v_y\partial_xA_y-v_y\partial_yA_x-v_z\partial_zA_x+v_z\partial_xA_z\\v_z\partial_yA_z-v_z\partial_zA_y-v_x\partial_xA_y+v_x\partial_yA_x\\v_x\partial_zA_x-v_x\partial_xA_z-v_y\partial_yA_z+v_y\partial_zA_y\end{pmatrix}+\dot{\vec{A}} \tag{10.3.144}$$

and as: And as:

$$\begin{aligned}&\begin{pmatrix}v_y\partial_x A_y - v_y\partial_y A_x - v_z\partial_z A_x + v_z\partial_x A_z\\ v_z\partial_y A_z - v_z\partial_z A_y - v_x\partial_x A_y + v_x\partial_y A_x\\ v_x\partial_z A_x - v_x\partial_x A_z - v_y\partial_y A_z + v_y\partial_z A_y\end{pmatrix} + \dot{\vec{A}}\\ &= \begin{pmatrix}v_y\partial_x A_y - \cancel{v_y\partial_y A_x} - \cancel{v_z\partial_z A_x} + v_z\partial_x A_z\\ v_z\partial_y A_z - \cancel{v_z\partial_z A_y} - \cancel{v_x\partial_x A_y} + v_x\partial_y A_x\\ v_x\partial_z A_x - \cancel{v_x\partial_x A_z} - \cancel{v_y\partial_y A_z} + v_y\partial_z A_y\end{pmatrix} + \begin{pmatrix}v_x\partial_x A_x + \cancel{v_y\partial_y A_x} + \cancel{v_z\partial_z A_x}\\ \cancel{v_x\partial_x A_y} + v_y\partial_y A_y + \cancel{v_z\partial_z A_y}\\ \cancel{v_x\partial_x A_z} + \cancel{v_y\partial_y A_z} + v_z\partial_z A_z\end{pmatrix}\\ &= \begin{pmatrix}v_y\partial_x A_y + v_z\partial_x A_z + v_x\partial_x A_x\\ v_z\partial_y A_z + v_x\partial_y A_x + v_y\partial_y A_y\\ v_x\partial_z A_x + v_y\partial_z A_y + v_z\partial_z A_z\end{pmatrix} = \begin{pmatrix}v_x\partial_x A_x + v_y\partial_x A_y + v_z\partial_x A_z\\ v_x\partial_y A_x + v_y\partial_y A_y + v_z\partial_y A_z\\ v_x\partial_z A_x + v_y\partial_z A_y + v_z\partial_z A_z\end{pmatrix}\end{aligned} \tag{10.3.145}$$

Therefore we have indeed the equality (yes i know written like this it looks stupid but don't forget the relation above we start from):

$$\begin{pmatrix}v_x\partial_x A_x + v_y\partial_x A_y + v_z\partial_x A_z\\ v_x\partial_y A_x + v_y\partial_y A_y + v_z\partial_y A_z\\ v_x\partial_z A_x + v_y\partial_z A_y + v_z\partial_z A_z\end{pmatrix} = \begin{pmatrix}v_x\partial_x A_x + v_y\partial_x A_y + v_z\partial_x A_z\\ v_x\partial_y A_x + v_y\partial_y A_y + v_z\partial_y A_z\\ v_x\partial_z A_x + v_y\partial_z A_y + v_z\partial_z A_z\end{pmatrix} \tag{10.3.146}$$

These developments thus confirm our initial hypothesis as what the action of the field can be written:

$$S_1 = \int \left(q\vec{A}\circ\vec{v} - q\phi\right)\mathrm{d}t = \int L_1 \mathrm{d}t \tag{10.3.147}$$

and that it expresses the interaction of a charged particle with a magnetic field (since it contains the Lorentz force!).

Here is a table with some interesting values based on Rayleigh criteria $\lambda = 550\,[nm]$:

Tool	**Diameter** D [m]	$\Delta\theta$ [rad]	$\Delta\theta$ [arcsec]	**Details on Moon**	**Details at** 200 [km]
Eye	0.0025	$2.7\cdot 10^{-4}$	55	103 [km]	53 [m]
	0.010	$6.7\cdot 10^{-5}$	13	25 [km]	13 [m]
Bincoular	0.050	$1.3\cdot 10^{-5}$	2.8	5 [km]	2.7 [m]
	0.10	$6.7\cdot 10^{-6}$	1.4	2.6 [km]	1.3 [m]
150 [mm] Telescope	0.15	$4.5\cdot 10^{-6}$	0.92	1.7 [km]	89 [cm]
	0.20	$3.4\cdot 10^{-6}$	0.69	1.3 [km]	67 [cm]
1 [m] Telescope	1.0	$6.7\cdot 10^{-7}$	0.14	260 [m]	13 [cm]
Hubble	2.4	$2.8\cdot 10^{-7}$	0.058	110 [m]	55 [mm]
James-Webb	6.5	$1.0\cdot 10^{-7}$	0.021	39 [m]	20 [mm]
VLT	8.0	$8.4\cdot 10^{-8}$	0.017	32 [m]	16 [mm]
Keck	10	$6.7\cdot 10^{-8}$	0.014	25 [m]	13 [mm]
ELT	40	$1.7\cdot 10^{-8}$	0.0035	6 [m]	3.3 [mm]

🟨	Discomfort	🟫	Danger
🟧	Extreme discomfort	🟥	Extreme Danger

Table 10.3: **Some typical power of resolution based on Rayleigh criteria**

So now we have proved that the "Lagrangian of the current-field interaction":

$$\boxed{L = -m_0c^2\sqrt{1-\frac{v^2}{c^2}} + q\vec{A}\circ\vec{v} - q\phi} \tag{10.3.148}$$

which we assumed empirically the shape earlier and that is finally well correct!

The action integral is then written:

$$S=\int_{t_1}^{t_2}\left(-m_0c^2\sqrt{1-\frac{v^2}{c^2}}+q\vec{A}\circ\vec{v}-q\phi\right)\mathrm{d}t \qquad (10.3.149)$$

Let us introduce the velocity $\vec{v}$ of the particle in the form $\vec{v}=\mathrm{d}\vec{r}/\mathrm{d}t$ and the integral is then written:

$$S=\int_{t_1}^{t_2}\left(-m_0c^2\sqrt{1-\frac{v^2}{c^2}}\mathrm{d}t+q\vec{A}\circ\mathrm{d}\vec{r}-q\phi\mathrm{d}t\right) \qquad (10.3.150)$$

Let us recall that we have proved in the section of Special Relativity the invariant:

$$\mathrm{d}s^2=c^2\mathrm{d}t^2-(\mathrm{d}x^2+\mathrm{d}y^2+\mathrm{d}z^2) \qquad (10.3.151)$$

and also:

$$\mathrm{d}s'^2=c^2\mathrm{d}t'^2-(\mathrm{d}x'^2+\mathrm{d}y'^2+\mathrm{d}z'^2) \qquad (10.3.152)$$

As the space-time intervals are invariants such as (see section Special Relativity page 5704):

$$c^2\mathrm{d}t'^2-(\mathrm{d}x'^2+\mathrm{d}y'^2+\mathrm{d}z'^2)=c^2\mathrm{d}t^2-(\mathrm{d}x^2+\mathrm{d}y^2+\mathrm{d}z^2) \qquad (10.3.153)$$

Let us introduce the velocity $\vec{v}$ of the particle in the form $\vec{v}=\mathrm{d}\vec{r}/\mathrm{d}t$ and the integral is then written:

$$S=\int_{t_1}^{t_2}\left(-m_0c^2\sqrt{1-\frac{v^2}{c^2}}\mathrm{d}t+q\vec{A}\circ\mathrm{d}\vec{r}-q\phi\mathrm{d}t\right) \qquad (10.3.154)$$

Let us recall that we have proved in the section of Special Relativity the invariant:

$$\mathrm{d}s^2=c^2\mathrm{d}t^2-(\mathrm{d}x^2+\mathrm{d}y^2+\mathrm{d}z^2) \qquad (10.3.155)$$

and also:

$$\mathrm{d}s'^2=c^2\mathrm{d}t'^2-(\mathrm{d}x'^2+\mathrm{d}y'^2+\mathrm{d}z'^2) \qquad (10.3.156)$$

As the space-time intervals are invariants such as (see section Special Relativity page 5704):

$$c^2\mathrm{d}t'^2-(\mathrm{d}x'^2+\mathrm{d}y'^2+\mathrm{d}z'^2)=c^2\mathrm{d}t^2-(\mathrm{d}x^2+\mathrm{d}y^2+\mathrm{d}z^2) \qquad (10.3.157)$$

If the repository O' is not in movement ($\mathrm{d}x'=\mathrm{d}y'=\mathrm{d}z'=0$), we have:

$$\mathrm{d}s^2=c^2\mathrm{d}t'^2=c^2\mathrm{d}t^2-(\mathrm{d}x^2+\mathrm{d}y^2+\mathrm{d}z^2) \qquad (10.3.158)$$

hence:

$$\mathrm{d}t'=\frac{\mathrm{d}s}{c}=\mathrm{d}t\sqrt{1-\frac{\mathrm{d}x^2+\mathrm{d}y^2+\mathrm{d}z^2}{c^2\mathrm{d}t^2}} \qquad (10.3.159)$$

which is also written as:

$$\frac{\mathrm{d}s}{c}=\mathrm{d}t\sqrt{1-\frac{v^2}{c^2}} \qquad (10.3.160)$$

Therefore:

$$S=\int_{t_1}^{t_2}\left(-m_c^2\sqrt{1-\frac{v^2}{c^2}}\mathrm{d}t+q\vec{A}\circ\mathrm{d}\vec{r}-q\phi\mathrm{d}t\right)=\int_{\alpha}^{\beta}(-m_0c\mathrm{d}s+q\vec{A}\circ\mathrm{d}\vec{r}-q\phi\mathrm{d}t) \qquad (10.3.161)$$

Now let us use of the contravariant four-vector potential (see above):

$$A^i=\left(\frac{\phi}{c},A_1,A_2,A_3\right)=\left(\frac{\phi}{c},\vec{A}\right) \qquad (10.3.162)$$

and the contravariant four-vector displacement (see section Special Relativity page 5639):

$$\mathrm{d}x^i=(c\mathrm{d}t,\mathrm{d}x_1,\mathrm{d}x_2,\mathrm{d}x_3)=(c\mathrm{d}t,\mathrm{d}\vec{x}) \qquad (10.3.163)$$

The expression of the action of an electric charged particle in an electromagnetic field and in a Minkowski metric η_{ij} (see sections of Special Relativity page 5705 and General Relativity page 5732) is finally reduced to the condensed expression:

$$\boxed{S=\int_{\alpha}^{\beta}-m_0c\mathrm{d}s-qA^i(\eta_{ij}\mathrm{d}x^j)=\int_{\alpha}^{\beta}-m_0c\mathrm{d}s-qA^i\mathrm{d}x_i} \qquad (10.3.164)$$

with therefore:

$$\mathrm{d}x_i=(c\mathrm{d}t,-\mathrm{d}x,-\mathrm{d}y,-\mathrm{d}z) \qquad (10.3.165)$$

without forgetting that here we use the metric $(+,-,-,-)$ (see sections of Special Relativity page 5705 and General Relativity page 5732).

Let us notice that the integral action in the absence of magnetic and electric field is then written:

$$S=-m_0c\int_{\alpha}^{\beta}\mathrm{d}s \qquad (10.3.166)$$

which corresponds well to what we have obtained in the section of Special Relativity for a free particle!

According to the principle of least action, the action integral has zero variation to the effective motion of the particle, therefore:

$$\delta S=-\left(\int_{\alpha}^{\beta}m_0c\mathrm{d}s+q\delta(A^i\mathrm{d}x_i)\right)=0 \qquad (10.3.167)$$

Remark

By the equality with zero, we can eliminate the minus sign before the integral.

Using the expression of the curvilinear abscissa (see section Special Relativity page 5705):

$$\mathrm{d}s^2=g_{ij}\mathrm{d}x^i\mathrm{d}x^j \tag{10.3.168}$$

for the Minkowski metric, we can write (remember that in the Euclidean metric only the terms of the diagonal where $i=j$ are non-zero):

$$\mathrm{d}s^2=\eta_{ij}\mathrm{d}x^j\mathrm{d}x^i=\mathrm{d}x_i\mathrm{d}x^i \tag{10.3.169}$$

Therefore:

$$\mathrm{d}s=\sqrt{\mathrm{d}x_i\mathrm{d}x^i}\Leftrightarrow\delta\mathrm{d}s=\frac{1}{2}\frac{\delta\mathrm{d}x_i\mathrm{d}x^i}{\sqrt{\mathrm{d}x_i\mathrm{d}x^i}}+\frac{1}{2}\frac{\mathrm{d}x_i\delta\mathrm{d}x^i}{\sqrt{\mathrm{d}x_i\mathrm{d}x^i}}=\frac{\mathrm{d}x^i\delta\mathrm{d}x_i}{\mathrm{d}s} \tag{10.3.170}$$

the preceding integral is then written:

$$\delta S=\int\limits_{\alpha}^{\beta}m_0c\delta\mathrm{d}s+q\delta(A^i\mathrm{d}x_i)=\int\limits_{\alpha}^{\beta}m_0c\frac{\mathrm{d}x^i\delta\mathrm{d}x_i}{\mathrm{d}x}+qA^i\delta\mathrm{d}x_i+q\delta A^i\mathrm{d}x_i \tag{10.3.171}$$

This gives using curvilinear components (see section Tensor Calculus page 1524):

$$\delta S=\int\limits_{\alpha}^{\beta}(m_0cu^i+qA^i)\delta\mathrm{d}x_i+\int\limits_{\alpha}^{\beta}q\delta A^i\mathrm{d}x_i \tag{10.3.172}$$

Let us integrate by parts (see section Differential and Integral Calculus page 1532) the first integral:

$$\delta S=(m_0cu^i+qA^i)\mathrm{d}x_i\Big|_{\alpha}^{\beta}-\int\limits_{\alpha}^{\beta}(m_0c\mathrm{d}u^i\delta x_i+q\mathrm{d}A^i\delta x_i)+\int\limits_{\alpha}^{\beta}q\delta A^i\mathrm{d}x_i \tag{10.3.173}$$

But as:

$$\delta x_i\Big|_{\alpha}=0\Rightarrow\mathrm{d}x_i\Big|_{\alpha}=0\qquad\text{and}\qquad\delta x_i\Big|_{\beta}=0\Rightarrow\mathrm{d}x_i\Big|_{\beta}=0 \tag{10.3.174}$$

Therefore:

$$\delta S=\int\limits_{\alpha}^{\beta}(-m_0c\mathrm{d}u^i\delta x_i-q\mathrm{d}A^i\delta x_i+q\delta A^i\mathrm{d}x_i) \tag{10.3.175}$$

with:

$$\delta A^i=\frac{\partial A^i}{\partial x^k}\delta x^k=\partial^k A^i\delta x^k\qquad\text{and}\qquad\mathrm{d}A^i=\frac{\partial A^i}{\partial x^k}\mathrm{d}x^k=\partial^k A^i\mathrm{d}x^k \tag{10.3.176}$$

can be written:

$$\begin{aligned}\delta S &= \int\limits_{\alpha}^{\beta} \left(-m_0 c \mathrm{d}u \delta x_i - q\partial^k A^i \mathrm{d}x^k \delta x_i + q\partial^k A^i \delta x^k \mathrm{d}x_i\right)\\ &= \int\limits_{\alpha}^{\beta} \left(-m_0 c\left(\frac{\mathrm{d}u^i}{\mathrm{d}s}\mathrm{d}s\right)\delta x_i - q\partial^k A^i (u^k \mathrm{d}s)\delta x_i + q\partial^k A^i \delta x^k (u_i \mathrm{d}s)\right)\\ &= \int\limits_{\alpha}^{\beta} \left(-m_0 c\left(\frac{\mathrm{d}u^i}{\mathrm{d}s}\mathrm{d}s\right)\delta x_i - q\partial^k A^i (u^k \mathrm{d}s)\delta x_i + q\partial^i A^k \delta x^i (u_k \mathrm{d}s)\right)\\ &= \int\limits_{\alpha}^{\beta} \left(-m_0 c\left(\frac{\mathrm{d}u^i}{\mathrm{d}s}\right) - q\partial^k A^i (u^k) + q\partial^i A^k (u_k)\right)\delta x_i \mathrm{d}s\\ &= \int\limits_{\alpha}^{\beta} \left(-m_0 c\left(\frac{\mathrm{d}u^i}{\mathrm{d}s}\right) + q(\partial^i A^k - \partial^k A^i)u^k\right)\delta x_i \mathrm{d}s\end{aligned} \tag{10.3.177}$$

The quantities δx_i being arbitrary, the expression in brackets is zero:

$$-m_0 c\left(\frac{\mathrm{d}u^i}{\mathrm{d}s}\right) + q(\partial^i A^k - \partial^k A^i)u^k = 0 \tag{10.3.178}$$

Let us put:

$$\boxed{F^{ik} = \partial^i A^k - \partial^k A^i} \tag{10.3.179}$$

The contravariant quantities F^{ik} form the contravariant components of what we named the "tensor of the electromagnetic field" or "Faraday's tensor" (hence the F...) or more commonly the "Maxwell tensor". We say then that F^{ik} the "curl of (the magnetic) potential".

The "equations of motion of a particle in an electromagnetic field" then take the form:

$$\boxed{m_0 c\frac{\mathrm{d}u^i}{\mathrm{d}s} - qF^{ik}u^k = 0} \tag{10.3.180}$$

that some physicists name the "geodesic corrected by a Lorentz force".

It is important not notice that the term:

$$F^{ik}u^k \tag{10.3.181}$$

corresponds obviously to the electric field seen by the observer!

Remark

The tensor of the electromagnetic field is invariant under the transformations:

$$A^i \rightarrow A^{i'} = A^i - \partial^i \alpha \tag{10.3.182}$$

Indeed:

$$F^{ik}=\partial^i(A^k-\partial^k\alpha)-\partial^k(A^i-\partial^i\alpha)=F^{ik}-\partial^i\partial^k\alpha+\partial^k\partial^i\alpha=F^{ik} \tag{10.3.183}$$

In Minkowski metric η_{ij} (we will need the tensor of the electromagnetic field in the section Special Relativity, hence the choice of this metric), we have however:

$$m_0c\frac{\eta_{ij}\mathrm{d}u^i}{\mathrm{d}s}-q\eta_{ij}(\partial^iA^k-\partial^kA^i)u^k=0 \tag{10.3.184}$$

Which gives:

$$m_0c\frac{\eta_{ij}\mathrm{d}u^i}{\mathrm{d}s}-q(\partial_jA^k-\partial^kA_j)u^k=0 \tag{10.3.185}$$

The term $\partial_jA^k-\partial^kA_j$ is often denoted F^{ik} (even if it is not more fully contravariant).

It remains to us to determine the contravariant components of the tensor F^{ik} (tensor which has the property of being antisymmetric such that $F^{ik}=-F^{ki}$ as we will see further below).

Let's start with the simplest. We assume as obvious that:

$$\begin{aligned}F^{00}&=\partial_0A^0-\partial^0A_0=0\\F^{11}&=\partial_1A^1-\partial^1A_1=0\\F^{22}&=\partial_2A^2-\partial^2A_2=0\\F^{33}&=\partial_3A^3-\partial^3A_3=0\end{aligned} \tag{10.3.186}$$

Then, remembering that $\vec{B}=\vec{\nabla}\times\vec{A}$:

$$\begin{pmatrix}B_x\\B_y\\B_y\end{pmatrix}=\begin{pmatrix}\partial_yA_z-\partial_zA_y\\\partial_zA_x-\partial_xA_z\\\partial_xA_y-\partial_yA_x\end{pmatrix}=\begin{pmatrix}\partial_2A_3-\partial_3A_2\\\partial_3A_1-\partial_1A_3\\\partial_1A_2-\partial_2A_1\end{pmatrix} \tag{10.3.187}$$

Hence (by choosing the Minkowski metric with signature $(+,-,-,-)$):

$$\begin{aligned}F^{21}=-F^{12}=\partial_2A^1-\partial^1A_2&=\frac{\partial}{-\partial y}A^1-\frac{\partial}{\partial x}(-A_2)=\frac{\partial}{\partial x}A_2-\frac{\partial}{\partial y}A^1=B_z\\F^{32}=-F^{23}=\partial_3A^2-\partial^2A_3&=\frac{\partial}{-\partial z}A^2-\frac{\partial}{\partial y}(-A_3)=\frac{\partial}{\partial y}A_3-\frac{\partial}{\partial z}A^2=B_x\\F^{13}=-F^{31}=\partial_1A^3-\partial^3A_1&=\frac{\partial}{-\partial x}A^3-\frac{\partial}{\partial z}(-A_1)=\frac{\partial}{\partial z}A_1-\frac{\partial}{\partial x}A^3=B_y\end{aligned} \tag{10.3.188}$$

Which gives us so far:

$$F^{ik}=\begin{bmatrix}F^{00}&F^{01}&F^{02}&F^{03}\\F^{10}&F^{11}&F^{12}&F^{13}\\F^{20}&F^{21}&F^{22}&F^{23}\\F^{30}&F^{31}&F^{32}&F^{33}\end{bmatrix} \tag{10.3.189}$$

Remark

Strictly speaking not to confuse the Faraday's tensor $F^{ik}=\partial^i A^k-\partial^k A^i$ with its matrix form, we should put the first term of the above equality between brackets $[F^{ik}]$ as we have already mention it in the section of Tensor Calculus!

Now, being known that $x_0=ct$ and $A_0=\phi/c$ the other components of the tensor F^{ki} are written taking into account that:

$$\vec{E}=-\vec{\nabla}\cdot\phi-\frac{\partial\vec{A}}{\partial t} \tag{10.3.190}$$

and therefore:

$$\frac{1}{c}\left(\frac{\partial\vec{A}}{\partial t}+\vec{\nabla}\cdot\phi\right)=\frac{\vec{E}}{c} \tag{10.3.191}$$

thus, with the partial contravariant derivatives according to the Minkowski metric:

$$\begin{aligned}F^{01}=-F^{10}=\partial_0 A^1-\partial^1 A_0=\frac{1}{c}\left(\frac{\partial A_1}{\partial t}-\frac{\partial\phi}{-\partial x}\right)=\frac{1}{c}\left(\frac{\partial A_1}{\partial t}+\frac{\partial\phi}{\partial x}\right)=-\frac{E_x}{c}\\F^{02}=-F^{20}=\partial_0 A^2-\partial^2 A_0=\frac{1}{c}\left(\frac{\partial A_2}{\partial t}-\frac{\partial\phi}{-\partial y}\right)=\frac{1}{c}\left(\frac{\partial A_2}{\partial t}+\frac{\partial\phi}{\partial y}\right)=-\frac{E_y}{c}\\F^{03}=-F^{30}=\partial_0 A^3-\partial^3 A_0=\frac{1}{c}\left(\frac{\partial A_3}{\partial t}-\frac{\partial\phi}{-\partial z}\right)=\frac{1}{c}\left(\frac{\partial A_3}{\partial t}+\frac{\partial\phi}{\partial z}\right)=-\frac{E_z}{c}\end{aligned} \tag{10.3.192}$$

Thus we have for the tensor of the electromagnetic field in contravariant components with and always with the Minkowski signature $(+,-,-,-)$:

$$F^{ik}=\begin{bmatrix}F^{00}&F^{01}&F^{02}&F^{03}\\F^{10}&F^{11}&F^{12}&F^{13}\\F^{20}&F^{21}&F^{22}&F^{23}\\F^{30}&F^{31}&F^{32}&F^{33}\end{bmatrix}=\begin{bmatrix}0&-\frac{E_x}{c}&-\frac{E_y}{c}&-\frac{E_z}{c}\\\frac{E_x}{c}&0&-B_z&B_y\\\frac{E_y}{c}&B_z&0&-B_x\\\frac{E_z}{c}&-B_y&B_x&0\end{bmatrix} \tag{10.3.193}$$

So that the equation of motion is finally:

$$\boxed{m_0c^2\frac{\mathrm{d}u^i}{\mathrm{d}s}-qF^{ik}u^k=0} \tag{10.3.194}$$

But as we will see it in the section of Special Relativity, the real tensor of the electromagnetic field is defined by (still in the metric $+,-,-,-$):

$$F^{\mu\nu}=cF^{ik}=\begin{bmatrix}0&-E_x&-E_y&-E_z\\E_x&0&-cB_z&cB_y\\E_y&cB_z&0&-cB_x\\E_z&-cB_y&B_x&0\end{bmatrix} \tag{10.3.195}$$

so that the Lorentz transformations are conform.

The expression in tensor form of the electromagnetic field clearly shows the unity of the electromagnetic field while generally the electric and magnetic fields are considered separately in classical theory.

But as in theoretical physics we often work in natural units (this is somewhat the "norm" ...), then we have in advanced graduate books:

$$\boxed{F^{\mu\nu}=\begin{bmatrix}0 & -E_x & -E_y & -E_z\\ E_x & 0 & -B_z & B_y\\ E_y & B_z & 0 & -B_x\\ E_z & -B_y & B_x & 0\end{bmatrix}} \tag{10.3.196}$$

Hence the equation of motion in natural units:

$$\boxed{m_0\frac{\mathrm{d}u^{\mu}}{\mathrm{d}s}-qF^{\mu\nu}u^{\nu}=0} \tag{10.3.197}$$

By denoting now the components of 1 to 4 instead of 0 to 3 (it is easier for students to find their way in the matrix) and without forgetting that the partial derivatives are covariant and adopting again the natural units such as $\mu=\varepsilon=1$ (verbatim $c=1$), the two Maxwell equations with source are written:

$$\begin{aligned}&\vec{\nabla}\circ\vec{E}=\frac{\rho_Q}{\varepsilon} && \xrightarrow{\text{natural units}}\vec{\nabla}\circ\vec{E}=\rho_Q\\ &\vec{\nabla}\times\vec{B}-\mu\varepsilon\frac{\mathrm{d}\vec{E}}{\mathrm{d}t}=\mu\vec{j} && \xrightarrow{\text{natural units}}\vec{\nabla}\times\vec{B}-\frac{\mathrm{d}\vec{E}}{\mathrm{d}t}=\vec{j}\end{aligned} \tag{10.3.198}$$

Using the tensor of the electromagnetic field, it remarkably appears that these two equations can be written as in the condensed tensor equation:

$$\boxed{\partial_{\mu}F^{\nu\mu}=j^{\nu}} \tag{10.3.199}$$

where j^{ν} is the "four-vector current" defined by (in natural units!):

$$j^{\nu}=(j^0,\vec{j})=(\rho_Q,\vec{j}) \tag{10.3.200}$$

Using the first definition of the Faraday's tensor (the one where the field components are divided by c) and taking as known (we will prove it later) that $c=1/\sqrt{\varepsilon_0\mu_0}$ we have in the SI system:

$$\partial_{\mu}F^{\nu\mu}=\mu_0 j^{\nu} \tag{10.3.201}$$

with:

$$j^{\nu}=(c\rho_Q,\vec{j}) \tag{10.3.202}$$

As we shall immediately see, the temporal part of this equation gives the divergence of the electric field and the spatial part the rotational magnetic field.

Indeed:

$$\begin{aligned}\partial_{\mu}F^{1\mu}&=\partial_1F^{11}+\partial_2F^{12}+\partial_3F^{13}+\partial_4F^{14}=\partial_tF^{11}+\partial_xF^{12}+\partial_yF^{13}+\partial_zF^{14}\\ &=\partial_t0+\partial_xE_x+\partial_yE_y+\partial_zE_z=\partial_xE_x+\partial_yE_y+\partial_zE_z\end{aligned} \tag{10.3.203}$$

This is latter equality is the equivalent of:

$$\partial_\mu F^{1\mu}=\vec{\nabla}\circ\vec{E} \tag{10.3.204}$$

And we also have:

$$\begin{aligned}\partial_\mu F^{2\mu}&=\partial_1 F^{21}+\partial_2 F^{22}+\partial_3 F^{23}+\partial_4 F^{24}=\partial_t F^{21}+\partial_x F^{22}+\partial_y F^{23}+\partial_z F^{24}\\&=\partial_t(-E_y)+\partial_x B_z+\partial_y 0+\partial_z(-B_x)=\partial_x B_z-\partial_z B_x-\partial_t E_y\\ \\ \partial_\mu F^{3\mu}&=\partial_1 F^{31}+\partial_2 F^{32}+\partial_3 F^{33}+\partial_4 F^{34}=\partial_t F^{31}+\partial_x F^{32}+\partial_y F^{33}+\partial_z F^{34}\\&=\partial_t(-E_y)+\partial_x B_z+\partial_y 0+\partial_z(-B_x)=\partial_x B_z-\partial_z B_x-\partial_t E_y\\ \\ \partial_\mu F^{4\mu}&=\partial_1 F^{41}+\partial_2 F^{42}+\partial_3 F^{43}+\partial_4 F^{44}=\partial_t F^{41}+\partial_x F^{42}+\partial_y F^{43}+\partial_z F^{44}\\&=\partial_t(-E_z)+\partial_x(-B_y)+\partial_y(B_x)+\partial_z 0=\partial_y B_x-\partial_x B_y-\partial_t E_z\end{aligned} \tag{10.3.205}$$

These three sets of equalities are the equivalent of:

$$\begin{pmatrix}\partial_\mu F^{2\mu}\\\partial_\mu F^{3\mu}\\\partial_\mu F^{4\mu}\end{pmatrix}=\vec{\nabla}\times\vec{B}-\partial_t\vec{E} \tag{10.3.206}$$

Also the both Maxwell equations:

$$\vec{\nabla}\circ\vec{B}=0 \qquad \vec{\nabla}\times\vec{E}+\frac{\partial\vec{B}}{\partial t}=\vec{0} \tag{10.3.207}$$

can be written in tensor form:

$$\boxed{\partial_\mu F^{\nu\lambda}+\partial_\nu F^{\lambda\mu}+\partial_\gamma F^{\mu\nu}=0} \tag{10.3.208}$$

Indeed we have first:

$$\begin{aligned}\partial_2 F^{43}+\partial_4 F^{32}+\partial_3 F^{24}&=\partial_x F^{43}+\partial_z F^{32}+\partial_y F^{24}=\partial_x B_x+\partial_z B_z+\partial_y B_y\\&=\partial_x B_x+\partial_y B_y+\partial_z B_z\end{aligned} \tag{10.3.209}$$

that corresponds obviously to:

$$\vec{B}\circ\vec{B} \tag{10.3.210}$$

and:

$$\begin{aligned}\partial_1 F^{43}+\partial_4 F^{31}+\partial_3 F^{14}&=\partial_t B_x+\partial_z(-E_y)+\partial_y(E_z)=\partial_y E_z-\partial_z E_y+\partial_t B_x\\\partial_1 F^{24}+\partial_2 F^{41}+\partial_4 F^{12}&=\partial_t B_y+\partial_x(-E_z)+\partial_z(E_x)=\partial_z E_x-\partial_x E_z+\partial_t B_y\\\partial_1 F^{32}+\partial_3 F^{21}+\partial_2 F^{13}&=\partial_t B_z+\partial_y(-E_x)+\partial_x(E_y)=\partial_x E_y-\partial_y E_x+\partial_t B_z\end{aligned} \tag{10.3.211}$$

These three sets of equalities are the equivalent of:

$$\begin{pmatrix}\partial_1 F^{43}+\partial_4 F^{31}+\partial_3 F^{14}\\\partial_1 F^{24}+\partial_2 F^{41}+\partial_4 F^{12}\\\partial_1 F^{32}+\partial_3 F^{21}+\partial_2 F^{13}\end{pmatrix}=\vec{\nabla}\times\vec{E}+\partial_t\vec{B} \tag{10.3.212}$$

Finally, all the Maxwell equations, adopting the natural units, are implicitly contained into the both relations:

$$\boxed{\begin{aligned}\partial_\mu F^{\nu\mu}=j^\nu\\ \partial_\mu F^{\nu\lambda}+\partial_v F^{\lambda\mu}+\partial_\lambda F^{\mu\nu}=0\end{aligned}} \tag{10.3.213}$$

We can also use the pseudo antisymmetric tensor of rank 4 that can be seen as a generalization of the Levi-Civita symbol (see section Tensor Calculus page 2045) such that we can write:

$$\boxed{\begin{aligned}\partial_\mu F^{\nu\mu}=j^\nu\\ \varepsilon^{\mu\nu\rho\sigma}F_{\rho\sigma}=0\end{aligned}} \tag{10.3.214}$$

with:

$$\varepsilon^{\mu\nu\rho\sigma}=\begin{cases}1 & \text{if } \mu\nu\rho\sigma \text{ is an even permuation of } (1234)\\ -1 & \text{if } \mu\nu\rho\sigma \text{ is an odd permuation of } (1234)\\ 0 & \text{otherwise}\end{cases} \tag{10.3.215}$$

The Lagrangian we have determined earlier above, however, is not complete. Indeed, when we apply the variational principle, we have already seen many times in the various section of this textbook (Classical Mechanics, Wave Mechanics, Magnetostatic, Special Relativity, General Relativity, etc.) that we could get the equations of motion (trajectories) of the subjects (bodies) of interest. The equations obtained also contain the parameters that explained the source of this movement (material properties, speed, field, etc.) as it was the case before!

Previously, we applied the variational principle on the Lagrangian of the charge-field interaction (electrostatic + magnetic) and got the equation of motion corrected by the Lorentz force.

When we determined the equations of motion of the charged particle from the principle of least action, we fixed the electromagnetic field (the field is known) and we varied the trajectory. The variational principle, must also allow us to get the field equations from the opposite approach: we fix the path of the particle (known of course) and we vary the electromagnetic field (potential and tensor field).

We should then obtain the Maxwell equations that, in the same way that we get what makes the motion of the particle when fixing the field in the variational principle, should gives us the information on what is the source of electric and magnetic field when the trajectory is fixed in the variational principle (I hope you have followed that explanation........).

The desire is then great to just simply start again from the action obtained earlier:

$$S=-\int m_0 c\mathrm{d}s-qA^i\mathrm{d}x_i \tag{10.3.216}$$

and apply to it a variation on the field.

As we have:

$$\rho_Q=\frac{\mathrm{dq}}{\mathrm{d}V} \tag{10.3.217}$$

we can write:

$$S=-\int m_0c\mathrm{d}s-A^i(\rho_Q\mathrm{d}V)\mathrm{d}x_i \qquad (10.3.218)$$

Let us consider electrical charges moving at the speed v and let us write the following quantity (do not forget that we continue to work in natural units such as $\mathrm{d}x_0=c\mathrm{d}t=\mathrm{d}t$!):

$$\mathrm{d}q\mathrm{d}x_i=(\rho_Q\mathrm{d}V)\mathrm{d}x_i=\rho_Q\mathrm{d}V\mathrm{d}x_0\frac{\mathrm{d}x_i}{\mathrm{d}x_0}=\rho_Q\mathrm{d}\Omega\frac{\mathrm{d}x_i}{\mathrm{d}t}=\left(\rho_Q\frac{\mathrm{d}x_i}{\mathrm{d}t}\right)\mathrm{d}\Omega=j_i\mathrm{d}\Omega \qquad (10.3.219)$$

with in natural units:

$$\mathrm{d}\Omega=\mathrm{d}x_0\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}x_3=\mathrm{d}t\mathrm{d}x\mathrm{d}y\mathrm{d}z \qquad (10.3.220)$$

Therefore we have:

$$S=\int -m_0\mathrm{d}s-A^ij_i\mathrm{d}\Omega \qquad (10.3.221)$$

If we apply the variational principle only on the field (constant in amplitude so that the source of the field is constant and therefore $\delta j_i=0$) and that we therefore consider the movement of electric charges known, it is immediate that the first term above is zero. Then we have:

$$\delta S=\int\delta(A^ij_i)\mathrm{d}\Omega=\int A^ij_i\mathrm{d}\Omega+A^i\delta j_i\mathrm{d}\Omega=\int\delta A^ij_i\mathrm{d}\Omega=\int j_i\delta A^i\mathrm{d}\Omega=0 \qquad (10.3.222)$$

that this integral to be zero it would require that j_i is zero ... which is pretty annoying if we wish to determine the characteristics of a source which then would not exist ... Therefore, we notice that something is missing in our Lagrangian!

The idea is the following: we know a tensor equation which involves the current density which is $\partial_\mu F^{\nu\mu}=j^\mu$ and that implicitly contains the only two Maxwell equations that give information on the source of the respective electric and magnetic fields (the other two giving properties of the field, not sources!) that is (always in natural units) for recall:

$$\begin{aligned}\vec{\nabla}\circ\vec{E}&=\rho_Q\\ \vec{\nabla}\times\vec{B}&=\frac{\mathrm{d}\vec{E}}{\mathrm{d}t}+\vec{j}\end{aligned} \qquad (10.3.223)$$

It is therefore sufficient to get these two equations (thus the corresponding tensor equation) following the variational principle to get the properties of the source of the field.

This simply means that ideally, we should (and expect) to have:

$$\delta S=\int\left(j^\nu-\partial_\mu F^{\nu\mu}\right)\delta A^\nu\mathrm{d}\Omega=0 \qquad (10.3.224)$$

where the integral equation vanishes exactly when $\partial_m uF^{\nu\mu}=j^\nu$!

It is then tempting to write something of the form (notice that we lowered the index of the potential A and raised that of the current density j in the second integral which does not change anything mathematically speaking to the result):

$$S=-\int m_0\mathrm{d}s-\int A_\nu j^\nu\mathrm{d}\Omega+\int ???\mathrm{d}\Omega \qquad (10.3.225)$$

We can use the following property of the Lagrangian quantities for to determine the missing expression "???": they all are invariant. In other words and for recall their pseudo-norm (scalar) is equal by Galilean basis change (see section Mechanics page 3239 and section Special Relativity page 5630) such that:

$$\begin{gathered}\mathrm{d}s=\mathrm{d}s'\\ A_\nu j^\nu = A'_\nu j'^\nu\end{gathered} \tag{10.3.226}$$

The first relation is quite obvious, we have already proved it many times. The second is perhaps less obvious therefore let us give a small indication (not general) to check whether it's correct: $A_\nu j^\nu$ is the dot product of j and A. If we apply the same (four) rotation to the two vectors, since Lorentz transformations are rotations (see section Special Relativity page 5642), the angle between j and A remains unchanged and thus the dot product.

But we must not forget that need to find the quantity "???" as a scalar invariant involving the Faraday tensor in one way or another.

We can then try directly with the following quantity (knowing in advance thanks to our precursors that this is the right hypothesis):

$$F_{\mu\nu}F^{\mu\nu}=(g_{\mu\sigma}g_{\nu\rho}F^{\sigma\rho})\,F^{\mu\nu}=(F^{\mu\nu})^2 \tag{10.3.227}$$

involving the covariant Faraday's tensor $F_{\mu\nu}$ and contravariant $F^{\mu\nu}$ because we know that:

1. It is a scalar invariant! Indeed, let us write $F_{\mu\nu}F^{\mu\nu}$ in terms of electric and magnetic fields to understand the physical meaning of that latter (still in natural units):

$$\begin{aligned}(g_{\mu\sigma}g_{\nu\rho}F^{\sigma\rho})F^{\mu\nu}=F_{\mu\nu}F^{\mu\nu}&=\begin{bmatrix}0 & E_x & E_y & E_z\\ -E_x & 0 & -B_z & B_y\\ -E_y & B_z & 0 & -B_x\\ -E_z & -B_y & B_x & 0\end{bmatrix}\begin{bmatrix}0 & -E_x & -E_y & -E_z\\ E_x & 0 & -B_z & B_y\\ E_y & B_z & 0 & -B_x\\ E_z & -B_y & B_x & 0\end{bmatrix}\\ &=F_{1\nu}F^{1\nu}+F_{2\nu}F^{2\nu}+F_{3\nu}F^{3\nu}+F_{4\nu}^{4\nu}\\ &=-E_x^2-E_y^2-E_z^2-E_x^2+B_z^2+B_y^2-E_y^2+B_z^2+B_x^2-E_z^2+B_y^2+B_x^2\\ &=2(\vec{B}\circ\vec{B}-\vec{E}\circ\vec{E})=2\left(\|\vec{B}\|^2-\|\vec{E}\|^2\right)\end{aligned} \tag{10.3.228}$$

Remark

If we were not in natural units, the calculation result would be of the form:

$$F_{\mu\nu}F^{\mu\nu}=\frac{2}{c^2}\left(c^2\|\vec{B}\|^2-\|\vec{E}\|^2\right) \tag{10.3.229}$$

The quantity:

$$c^2\|\vec{B}\|^2-\|\vec{E}\|^2 \tag{10.3.230}$$

(or $\|\vec{B}\|^2-\|\vec{E}\|^2$ in natural units) is therefore an invariant of the field.

✍**Example:**

In a repository O, let us consider a plane electromagnetic wave. The modules of the electric field and magnetic field are linked by the relation $E=cB$ (see below for proof). The invariant of the considered field is therefore zero. In another reference frame, with the same structure of the field, then we will also have $E'=cB'$.

2. Because a variational on this term gives:

$$\delta(F^{\mu\nu})^2=\delta(F_{\mu\nu}F^{\mu\nu})=F_{\mu\nu}\delta F^{\mu\nu}+\delta F_{\mu\nu}F^{\mu\nu}=2\delta F_{\mu\nu}F^{\mu\nu} \tag{10.3.231}$$

where we guess ... by digging a little bit, that $\delta F_{\mu\nu}$ implicitly contains the term δA^{ν}. We also see that a factor 2 appears such that we will have to introduce a normalization constant α, that anyway has to be introduced for a matter of homogeneity of the units of the expression of the action.

So eventually let try with something like:

$$S=-\int m_0\mathrm{d}s-\int A_{\mu}j^{\mu}\mathrm{d}\Omega+\alpha\int(F^{\mu\nu})^2\mathrm{d}\Omega \tag{10.3.232}$$

Now to find the equations of the electromagnetic field, we consider that the movement of electric charges are known and we use the principle of least action by varying only the components of the vector potential and those of the tensor of the electromagnetic field.

It follows that the variation of the first integral is zero and that it remains:

$$\begin{aligned}\delta S&=-\int\delta A_{\mu}j^{\mu}\mathrm{d}\Omega+\alpha\int\delta(F^{\mu\nu})^2\mathrm{d}\Omega\\&=-\int\delta A_{\mu}j^{\mu}\mathrm{d}\Omega+\alpha\int\delta(F_{\mu\nu}F^{\mu\nu})\mathrm{d}\Omega=-\int\delta A_{\mu}j^{\mu}\mathrm{d}\Omega+2\alpha\int F^{\mu\nu}\delta F_{\mu\nu}\mathrm{d}\Omega\end{aligned} \tag{10.3.233}$$

But we know that $F_{\mu\nu}$ is equal to $-F_{\nu\mu}$ since the Faraday tensor is antisymmetric:

$$F^{\mu\nu}\delta F_{\mu\nu}=-F^{\mu\nu}\partial_{\mu}\delta A_{\nu}-F^{\mu\nu}\partial_{\nu}\delta A_{\mu} \tag{10.3.234}$$

Nothing prevents us to swap the indices μ,ν in the first term at the right of the equality:

$$F^{\mu\nu}\delta F_{\mu\nu}=-F^{\mu\nu}\partial_{\mu}\delta A_{\nu}-F^{\mu\nu}\partial_{\nu}\delta A_{\mu}=-2F^{\mu\nu}\partial_{\nu}\delta A_{\mu} \tag{10.3.235}$$

But we know that $F_{\mu\nu}$ is equal to $-F_{\nu\mu}$ since the Faraday tensor is antisymmetric:

$$F^{\mu\nu}\delta F_{\mu\nu}=-F^{\mu\nu}\partial_{\mu}\delta A_{\nu}-F^{\mu\nu}\partial_{\nu}\delta A_{\mu} \tag{10.3.236}$$

Nothing prevents us to swap the indices μ,ν in the first term at the right of the equality:

$$F^{\mu\nu}\delta F_{\mu\nu}=-F^{\mu\nu}\partial_{\mu}\delta A_{\nu}-F^{\mu\nu}\partial_{\nu}\delta A_{\mu}=-2F^{\mu\nu}\partial_{\nu}\delta A_{\mu} \tag{10.3.237}$$

So finally:

$$\delta S=-\int\delta A_{\mu}j^{\mu}\mathrm{d}\Omega-4\alpha\int F^{\mu\nu}\partial_{\nu}\delta A_{\mu}\mathrm{d}\Omega=0 \tag{10.3.238}$$

Let us look at the second integral:

$$\int F^{\mu\nu}\partial_\nu \delta A_\mu \mathrm{d}\Omega = \int F^{\mu\nu}\frac{\partial}{\partial x^\nu}(\delta A_\mu)\mathrm{d}x_0\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}x_3 \qquad (10.3.239)$$

Applying the Fubini theorem (see section Differential and Integral Calculus page 1521) that says that we can integrate in any order the integration variables (under certain conditions) then we can apply the integration by parts (see section Differential and Integral Calculus page 1532) to write:

$$\int F^{\mu\nu}\partial_\nu \delta A_\mu \mathrm{d}\Omega = \int F^{\mu\nu}\delta A_\mu \mathrm{d}S - \int \partial_\nu F^{\mu\nu}\delta A_\mu \mathrm{d}\Omega \qquad (10.3.240)$$

where $\mathrm{d}S$ represents the boundary surface of the hyper-volume $\mathrm{d}\Omega$ which we were initially integrating and that omits the variable considered by the choice of the superscript ν.

Now according to the superscript ν concerned, the boundaries of the first term of the equality:

$$\int F^{\mu\nu}\delta A_\mu \mathrm{d}S \qquad (10.3.241)$$

will be on the components of time or components of space. If we focus on the temporal boundaries of integration, these are the initial and final moments of the action at which we apply this variational.

But at the time ends, the variational of the vector potential δA_μ is zero (by definition), therefore the integral over the time component will be zero.

Now on the spatial components, the (spatial) boundaries are those that permits to integrate the border surface of the hyper-volume at the final time. If this one is taken as being the infinity, the radius of the border area will be infinite and at every point of this surface, the energy carried by the field and the amplitude of the field components will be zero (see proof further below).

Therefore the variational of action is finally written:

$$\delta S = -\int \delta A_\mu j^\mu \mathrm{d}\Omega + 4\alpha\int \partial_v F^{\mu\nu}\delta A_\mu \mathrm{d}\Omega = \int (4\alpha\partial_\nu F^{\mu\nu} - j^\mu)\delta A_\mu \mathrm{d}\Omega = 0 \qquad (10.3.242)$$

The variations of the vector potential being arbitrary, the preceding integral will be zero if the integrand is also zero, hence the relations:

$$4\alpha\partial_v F^{\mu\nu} - j^\mu = 0 \qquad (10.3.243)$$

which brings us to:

$$4\alpha\partial_v F^{\mu\nu} = j^\mu \qquad (10.3.244)$$

Remark

We return on this Lagrangian with another (very interesting) approach in the section of Quantum Field Theory.

10.3.4 Electromagnetic wave equation

Maxwell supposed that the electromagnetic wave was a combination of phenomena that explain the third and fourth equations. If an electromagnetic wave is far from its source, then we can then neglect the surface density of current from the source as having no influence on the wave (we say then that it is the Maxwell equations without sources which we have already mentioned earlier above). So the third and fourth Maxwell equations are written:

$$\vec{\nabla}\times\vec{E}=-\frac{\partial\vec{B}}{\partial t}\qquad\text{and}\qquad\vec{\nabla}\times\vec{H}=\varepsilon\frac{\mathrm{d}\vec{E}}{\mathrm{d}t}\tag{10.3.245}$$

Let us first prove that the magnetic field and electric field are perpendicular!

Proof 10.2.3. A plane wave can be written:

$$\vec{E}=\hat{E}\cos(\omega t-\vec{k}\circ\vec{r}+\varphi_e)\tag{10.3.246}$$

$$\vec{B}=\hat{B}\cos(\omega t-\vec{k}\circ\vec{r}+\varphi_m)\tag{10.3.247}$$

with constant amplitude vectors $\hat{E}$, $\hat{B}$ and where the plane that is perpendicular to a given vector $\vec{k}$.

Now let's put the plane wave in the third Maxwell equation:

$$\begin{aligned}\vec{\nabla}\times\vec{E}&=\left(\frac{\partial E_z}{\partial y}-\frac{\partial E_y}{\partial z}\right)\vec{x}+\left(\frac{\partial E_x}{\partial z}-\frac{\partial E_z}{\partial x}\right)\vec{y}+\left(\frac{\partial E_y}{\partial x}-\frac{\partial E_x}{\partial y}\right)\vec{z}\\&=\left[\left(\hat{E}_zk_y-\hat{E}_yk_z\right)\vec{x}+\left(\hat{E}_xk_z-\hat{E}_zk_x\right)\vec{y}+\left(\hat{E}_yk_x-\hat{E}_xk_y\right)\vec{z}\right]\sin(\omega t-\vec{k}\circ\vec{r}+\varphi_e)\\&=-\frac{\partial\vec{B}}{\partial t}=\omega\hat{B}\sin(\omega t-\vec{k}\circ\vec{r}+\varphi_m)\end{aligned}\tag{10.3.248}$$

with the Cartesian coordinate vectors $\vec{x}$, $\vec{y}$, $\vec{z}$.

Here we see that the electric and magnetic components have to be in phase ($\varphi_e=\varphi_m$) and hence we can focus on the amplitudes $\hat{E}$, $\hat{B}$. To check for orthogonality, we evaluate the scalar product of the amplitudes and use the representation of $\hat{B}$ in terms of $\hat{E}$ we just found:

$$\begin{aligned}\hat{\vec{E}}\circ\hat{\vec{B}}/\omega&=[E_x\vec{x}+E_y\vec{y}+E_z\vec{z}]\circ\left[\left(\hat{E}_zk_y-\hat{E}_yk_z\right)\vec{x}+\left(\hat{E}_xk_z-\hat{E}_zk_x\right)\vec{y}+\left(\hat{E}_yk_x-\hat{E}_xk_y\right)\vec{z}\right]\\&=\hat{E}_x\hat{E}_zk_y-\hat{E}_x\hat{E}_yk_z+\hat{E}_y\hat{E}_xk_z-\hat{E}_y\hat{E}_zk_x+\hat{E}_z\hat{E}_yk_x-\hat{E}_z\hat{E}_xk_y=0\end{aligned}\tag{10.3.249}$$

■ Q.E.D.

The excitation magnetic field $\vec{H}$ and electric field $\vec{E}$ being perpendicular (in the special case of electromagnetic waves in vacuum!), let us place ourselves conveniently in a system of orthonormal and Euclidean axes $(\vec{i},\vec{j},\vec{k})$ belonging to $\mathbb{R}^3$ by choosing:

$$\vec{H}=(0,0,H)=\frac{B}{\mu}\vec{k}\qquad\text{and}\qquad\vec{E}=(0,E,0)\tag{10.3.250}$$

Remark

Warning! The reader must remember that in what follows, H is the z component of $\vec{H}$ and E the y component of $\vec{E}$.

The (simple) calculations of $\vec{\nabla}\times\vec{E}$ and $\vec{\nabla}\times\vec{H}$ gives, after simplification:

$$\vec{\nabla}\times\vec{E}=\frac{\partial\vec{E}}{\partial x}=-\frac{\partial\vec{B}}{\partial t}\qquad\text{and}\qquad\vec{\nabla}\times\vec{H}=-\frac{\partial H}{\partial x}=\varepsilon\frac{\partial\vec{E}}{\partial t}\tag{10.3.251}$$

Before going further, a reader asked us to develop the details that lead to the left equality. So we start from:

$$\vec{\nabla}\times\vec{E}=\begin{pmatrix}\partial_x\\\partial_y\\\partial_z\end{pmatrix}\times\begin{pmatrix}E_x\\E_y\\E_z\end{pmatrix}=\begin{pmatrix}\partial_y E_z-\partial_z E_y\\\partial_z E_x-\partial_x E_z\\\partial_x E_y-\partial_y E_x\end{pmatrix}=\begin{pmatrix}-\partial_z E_y\\0\\\partial_x E_y\end{pmatrix}=-\frac{\vec{B}}{\partial t}\tag{10.3.252}$$

But:

$$\partial_z E_y=0\tag{10.3.253}$$

because the wave is a plane wave and the component of the electric field being in y, it does not vary following z. Then we have:

$$\vec{\nabla}\times\vec{E}=\begin{pmatrix}-\partial_z E_y\\0\\\partial_x E_y\end{pmatrix}=\begin{pmatrix}0\\0\\\partial_x E_y\end{pmatrix}=\partial_x\vec{E}=-\frac{\partial\vec{B}}{\partial t}\tag{10.3.254}$$

This being done, if we continue, we have:

$$\frac{\partial E}{\partial x}=-\frac{\partial B}{\partial t}\Rightarrow\frac{\partial^2 E}{\partial x^2}=-\mu\frac{\partial^2 H}{\partial x\partial t}\qquad\text{and}\qquad-\frac{\partial H}{\partial x}=\varepsilon\frac{\partial E}{\partial t}\Rightarrow-\frac{\partial^2 H}{\partial x\partial t}=\varepsilon\frac{\partial^2 E}{\partial t^2}\tag{10.3.255}$$

Identifying similar terms, we get the "equation of propagation of the electric field":

$$\boxed{\frac{\partial^2 E}{\partial t^2}=\frac{1}{\mu\varepsilon}\frac{\partial^2 E}{\partial x^2}}\tag{10.3.256}$$

and proceeding exactly in a similar way:

$$\boxed{\frac{\partial^2 B}{\partial t^2}=\frac{1}{\mu\varepsilon}\frac{\partial^2 B}{\partial x^2}}\tag{10.3.257}$$

relations that are both in the form of a wave equation (see section Wave Mechanics page 3476) of the type (for recall) of a "Poisson equation" (specifically it is a "d'Alembert's equation"):

$$\frac{\partial^2 y}{\partial t^2}=v^2\frac{\partial^2 y}{\partial x^2}\tag{10.3.258}$$

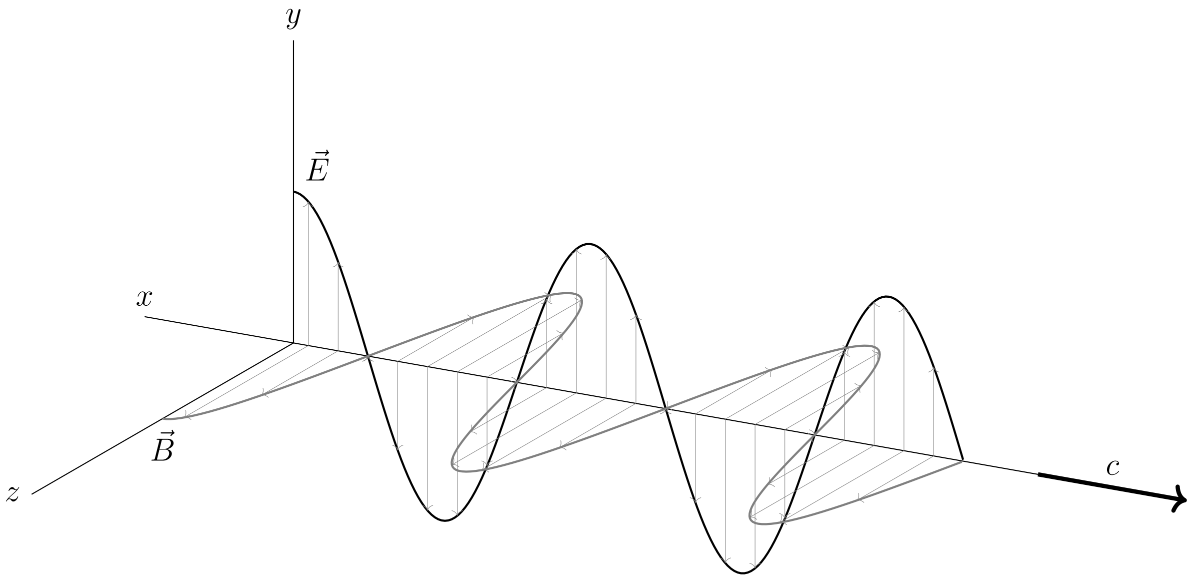
where we have:

$$\varepsilon=\varepsilon_0\varepsilon_r\,[\mathrm{A}\cdot\mathrm{s}\cdot\mathrm{V}^{-1}\cdot\mathrm{m}^{-1}]\qquad\text{and}\qquad\mu=\mu_0\mu_r\,[\mathrm{V}\cdot\mathrm{s}\cdot\mathrm{A}^{-1}\cdot\mathrm{m}^{-1}]\tag{10.3.259}$$

The propagation speed of an electromagnetic wave in vacuum is:

$$c=\frac{1}{\sqrt{\mu_0\varepsilon_0}}=\frac{1}{\sqrt{4\pi 10^{-7}\cdot 8.8542\cdot 10^{-12}}}=2.997292\cdot 10^8\,[\mathrm{m}\cdot\mathrm{s}^{-1}]\tag{10.3.260}$$

the units and numerical values agree...



$$c=\frac{E}{B}$$

$\vec{E}=$ electric field amplitude (instantaneous values)
$\vec{B}=$ magnetic field amplitude (instantaneous values)
$c=$ speed of light ($\sim 3\cdot 10^8\,[\text{m}\cdot\text{s}^{-1}]$)

$$c=\frac{1}{\sqrt{\mu_0\varepsilon_0}}$$

$\mu_0=$ magnetic permeability in vacuum, $\mu_0=\sim 1.3\cdot 10^{-6}\,[\text{N}\cdot\text{A}^{-2}]$
$\varepsilon_0=$ electric permeability in vacuum, $\varepsilon_0=\sim 8.9\cdot 10^{-12}\,[\text{C}^2\cdot\text{N}\cdot\text{m}^{-2}]$

Remark

The amplitudes of the two components of the electromagnetic waves are generally not comparable, and they are only drawn as equal for visual convenience! The reason for this is that the units of electric field and magnetic field don't generally match, so trying to compare an electric field amplitude with a magnetic field amplitude is completely meaningless. In certain unit systems this comparison can indeed be made, with the most prominent example being CGS (Centimeter, Gram, Second) units, where electric and magnetic fields have the same physical dimension and are therefore comparable; in this system the electric and magnetic field amplitudes for a plane wave are exactly equal! In other systems, such as SI units, no such comparison can be made.

The propagation velocity of the electromagnetic wave in the a material is then given by:

$$\boxed{v=\frac{1}{\sqrt{\mu_0\mu_r\varepsilon_0\varepsilon_r}}} \tag{10.3.261}$$

We have $v<c$ because experience has show so far that we can not exceed the speed of light, which is one of the postulates of Special and General Relativity.

In vacuum we have in S.I. units:

$$c=299,792.4562\pm 0.0011\,[\text{km}\cdot\text{s}^{-1}]$$

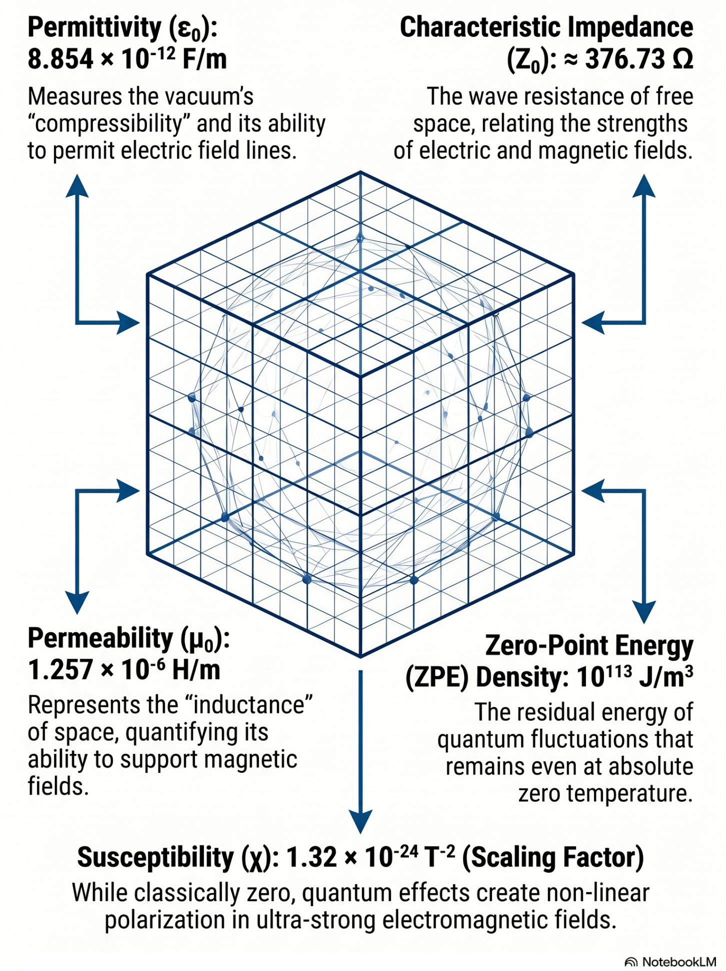
Or in imperial units (...):

$$c=186,282\,[\text{miles}],\ 698\,[\text{yards}],\ 2\,[\text{feet}],\ 5\frac{21}{127}\,[\text{inches}\cdot\text{s}^{-1}]$$

Or in natural units:

$$c = 1\,[\text{einstein}]$$

So as light propagates by self-sustaining electromagnetic wave in vacuum, we then have the obvious fact that space (or at least the electromagnetic quantum field) can be considered as a medium with at the following physical properties:



Permittivity (ε_0): 8.854 × 10^{-12} F/m

Measures the vacuum's "compressibility" and its ability to permit electric field lines.

Characteristic Impedance (Z_0): ≈ 376.73 Ω

The wave resistance of free space, relating the strengths of electric and magnetic fields.

Permeability (μ_0): 1.257 × 10^{-6} H/m

Represents the "inductance" of space, quantifying its ability to support magnetic fields.

Zero-Point Energy (ZPE) Density: 10^{113} J/m^3

The residual energy of quantum fluctuations that remains even at absolute zero temperature.

Susceptibility (χ): 1.32 × 10^{-24} T^{-2} (Scaling Factor)

While classically zero, quantum effects create non-linear polarization in ultra-strong electromagnetic fields.

Figure 10.67: Some properties of the medium "space" (vacuum) that we will see through this textbooks

Remarks

R1. In the 11970s (holocene calendar), interferometry was used to get the most accurate value for the speed of light that had been measured yet: $299,792.4562 \pm 0.0011\,[\text{km}\cdot\text{s}^{-1}]$. Then, in 11983, the meter was redefined in the International System of Units (SI) as the distance travelled by light in vacuum in $1/299,792,458$ of a second. As a result, the numerical value of the speed of light c in meters per second is now fixed exactly by the definition of the meter.

R2. The ancient Egyptian royal cubit (meh niswt) is the earliest attested standard measure of a length, and in that measurement system the speed of light is equal $571,033,253\,[\text{cubits}\cdot\text{s}^{-1}]$.

So we can finally write:

$$\begin{aligned}\frac{\partial^2 E}{\partial t^2}=c^2\frac{\partial^2 E}{\partial x^2}\Leftrightarrow\left(\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)E=0\\ \frac{\partial^2 B}{\partial t^2}=c^2\frac{\partial^2 B}{\partial x^2}\Leftrightarrow\left(\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)B=0\end{aligned}\tag{10.3.262}$$

Therefore using the d'Alembertian in one dimension:

$$\boxed{\begin{aligned}\Box E=0\\ \Box B=0\end{aligned}}\tag{10.3.263}$$

As we failed to get a direct expression of $E(x,t)$ and $B(x,t)$, we have just obtained differential equations containing only one of these fields. We name these equations respectively "wave equation for the electric field" and "wave equation for the magnetic field of induction".

The both equations have the same form and admit a solution of the same type. An obvious and particular solution (we leave it to the reader to make this check) of these differential equations is the sine trigonometric function:

$$\boxed{\begin{aligned}E(x,t)=\hat{E}\sin(kx-\omega t)\\ B(x,t)=\hat{B}\sin(kx-\omega t)\end{aligned}}\tag{10.3.264}$$

by not forgetting the relation between the pulsation ω, the propagation velocity c and the wave number k that we had proved in the section of Wave Mechanics!

A more general solution is the sum of the trivial solutions (see section Differential and Integral Calculus page 1607):

$$\begin{aligned}E(x,t)=\hat{E}\sin(kx-\omega t)+\hat{E}\cos(kx-\omega t)\\ B(x,t)=\hat{B}\sin(kx-\omega t)+\hat{B}\cos(kx-\omega t)\end{aligned}\tag{10.3.265}$$

But we have seen in our study of phasors (see section Wave Mechanics page 3510) that this real solution is only a particular case of a more general solution that is in the imaginary number field. So finally we can write:

$$E(x,t)=\hat{E}e^{\mathrm{i}(kx-\omega t)}\qquad B(x,t)=\hat{B}e^{\mathrm{i}(kx-\omega t)}\tag{10.3.266}$$

Which constitutes the "monochromatic plane wave" which is the simplest type of wave to manipulate in physics.

In three dimensions, the solution is by extension:

$$\vec{E}(\vec{r},t)=\vec{E}_0e^{\mathrm{i}(\vec{k}\circ\vec{r}-\omega(\vec{k})t)}\qquad \vec{B}(\vec{r},t)=\vec{B}_0e^{\mathrm{i}(\vec{k}\circ\vec{r}-\omega(\vec{k})t)}\tag{10.3.267}$$

where $\vec{k}$ is named the "propagation vector".

Remark

The monochromatic wave can has no real signification as physical reality. Indeed, if we calculate the electrical energy associated with in all space, we obtain for it an infinite energy (as it has neither beginning nor end!) Which is not realistic. We will see in the section of Wave Quantum Physics that in fact the light is enclosed in a quantum of energy.

But the wave equation is linear (solution is always the sum of other solutions). This implies that a superposition of waves of different frequencies (wave number and pulsation also!) is also a solution. Thus, by varying the wave vector $\vec{k}$ (and implicitly via its norm $\|\vec{k}\|$, pulsation ω, frequency f and period T) we also scan all the possible directions of propagation.

Written mathematically this gives, for the electric field:

$$\vec{E}(\vec{r},t)=\int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty}\vec{E}_0(\vec{k})e^{\mathrm{i}(\vec{k}\circ\vec{r}-\omega(\vec{k})t)}\mathrm{d}^3k=\int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty}\vec{E}_0(\vec{k})e^{\mathrm{i}(\vec{k}\circ\vec{r}-\omega(\vec{k})t)}\mathrm{d}k_x\mathrm{d}k_y\mathrm{d}k_z \tag{10.3.268}$$

And nothing prevents us from extracting a coefficient of the initial amplitude of the field such that:

$$\vec{E}(\vec{r},t)=\frac{1}{(2\pi)^{3/2}}\int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty}\vec{E}_0(\vec{k})e^{\mathrm{i}(\vec{k}\circ\vec{r}-\omega(\vec{k})t)}\mathrm{d}^3k \tag{10.3.269}$$

and we find here a relation very similar to that of an inverse Fourier transform (see section Sequences and Series page 1758) which is a remarkable fact! So the trick is now to put $t=0$ because the previous relation is not just a simple analogy with the Fourier transform, it is a Fourier transform!!!!

We can therefore relate the real field $\vec{E}(\vec{r},0)$ to the field $\vec{E}_0(\vec{k})$:

$$\begin{aligned}\vec{E}(\vec{r},0)&=\frac{1}{(2\pi)^{3/2}}\int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty}\vec{E}_0(\vec{k})e^{\mathrm{i}\vec{k}\circ\vec{r}}\mathrm{d}^3k\\ \vec{E}_0(\vec{r})&=\frac{1}{(2\pi)^{3/2}}\int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty}\vec{E}(\vec{r},0)e^{\mathrm{i}\vec{k}\circ\vec{r}}\mathrm{d}^3k\end{aligned} \tag{10.3.270}$$

These two relations are often condensed as following:

$$\begin{aligned}\vec{E}(\vec{r},0)&=\frac{1}{(2\pi)^{3/2}}\int\limits_{-\infty}^{+\infty}\vec{E}_0(\vec{k})e^{\mathrm{i}\vec{k}\circ\vec{r}}\mathrm{d}^3k\\ \vec{E}_0(\vec{r})&=\frac{1}{(2\pi)^{3/2}}\int\limits_{-\infty}^{+\infty}\vec{E}(\vec{r},0)e^{\mathrm{i}\vec{k}\circ\vec{r}}\mathrm{d}^3k\end{aligned} \tag{10.3.271}$$

The real field is thus at the initial instant the inverse Fourier transform of the field $\vec{E}_0(\vec{k})$. The term $\vec{E}_0(\vec{k})$ therefore represents the spectral component related to the particular wave vector $\vec{k}$ of the real field. This general solution of the wave equation is named "wave packet".

Remarks

R1. Identically to our study of Wave Mechanics the coefficients ω (pulsation) and $\vec{k}$ (wave number) coefficients are required to express the variation of the sine by radians and to give it a direction and a pulsation (see section Wave Mechanics page 3479).

R2. The periodicity of the sine function requires (see section Trigonometry page 2511):

$$\sin(kx-\omega(t+T))=\sin(kx-\omega t-2\pi) \qquad (10.3.272)$$

hence the definition of the period of a wave:

$$T=\frac{2\pi}{\omega} \qquad (10.3.273)$$

R3. The periodicity in space makes it possible to define identically the wavelength of the function as:

$$\lambda=\frac{2\pi}{k} \qquad (10.3.274)$$

We thus observe that the plane wave moves along x by travelling a distance λ in a time T. The velocity of the electromagnetic wave is then:

$$v=\frac{\lambda}{T}=c \qquad (10.3.275)$$

By introducing:

$$\begin{aligned}E(x,t)&=\hat{E}\sin(kx-\omega t)=\hat{E}\sin(k(x-ct))\\B(x,t)&=\hat{B}\sin(kx-\omega t)=\hat{B}\sin(k(x-ct))\end{aligned} \qquad (10.3.276)$$

Into the one dimensional version of:

$$\frac{\partial\vec{E}}{\partial t}=-\frac{\partial\vec{B}}{\partial t} \qquad (10.3.277)$$

we get:

$$\hat{E}k\cos(k(x-ct))=-\hat{B}(-kc)\cos(k(x-ct)) \qquad (10.3.278)$$

to finally obtain the famous result for the plane electromagnetic wave:

$$\boxed{\frac{\hat{E}}{\hat{B}}=c} \qquad (10.3.279)$$

Now let us generalize the above two relations as:

$$\begin{aligned}\vec{E}(x,t)&=\vec{E}_0e^{\mathrm{i}(\vec{k}\circ\vec{r}-\omega t)}\\\vec{B}(x,t)&=\vec{B}_0e^{\mathrm{i}(\vec{k}\circ\vec{r}-\omega t)}\end{aligned} \qquad (10.3.280)$$

We can see that the Maxwell relation impose other constraints in that case like:

$$\begin{aligned}\vec{\nabla}\circ\vec{E}=0 &\Longrightarrow \frac{\partial E_x}{\partial x}+\frac{\partial E_y}{\partial y}+\frac{\partial E_z}{\partial z}=0\\ &\Longrightarrow \mathrm{i}k_xE_x+\mathrm{i}k_yE_y+\mathrm{i}k_zE_z=0\\ &\Longrightarrow \vec{k}\circ\vec{E}=0\end{aligned} \tag{10.3.281}$$

This says that $\vec{E}$ is always perpendicular to $\vec{k}$.

Using $\vec{\nabla}\circ\vec{B}=0$ we get the analogous result for $\vec{B}$, namely:

$$\vec{k}\circ\vec{B}=0 \tag{10.3.282}$$

We also have then:

$$\begin{aligned}\nabla\times\vec{E}=-\frac{\partial\vec{B}}{\partial t} &\Longrightarrow \left(\frac{\partial}{\partial x},\frac{\partial}{\partial y},\frac{\partial}{\partial z}\right)\times\vec{E}=-\frac{\partial\vec{B}}{\partial t}\\ &\Longrightarrow (\mathrm{i}k_x,\mathrm{i}k_y,\mathrm{i}k_z)\times\vec{E}=-(-\mathrm{i}\omega)\vec{B}\\ &\Longrightarrow \vec{k}\times\vec{E}=\omega\vec{B}\end{aligned} \tag{10.3.283}$$

Therefore we get a quite famous relation used in optics and plasma physics:

$$\vec{B}=\frac{\vec{k}\times\vec{E}}{\omega}=\frac{\vec{k}\times\vec{E}}{ck} \tag{10.3.284}$$

10.3.4.1 Foucault light speed experiment

In 11850 (holocene calendar), Léon Foucault used a rotating mirror to perform a differential measurement of the speed of light in water versus its speed in air. In 11862 (holocene calendar), he used a similar apparatus to measure the speed of light in the air[4]. His published values in 11862 (holocene calendar) was $c=298,000\,[\mathrm{m}\cdot\mathrm{s}^{-1}]$ and was thought to be accurate to about $500\,[\mathrm{m}\cdot\mathrm{s}^{-1}]$

The Foucault experiment, conducted by Léon Foucault in 11850 (holocene calendar), is another clever method to measure the speed of light. Here's an intuitive explanation:

- Imagine a light source, a rotating mirror, and a fixed mirror placed some distance away. The light source sends a beam of light towards the rotating mirror.
- The rotating mirror reflects the light beam towards the fixed mirror. The fixed mirror then reflects the light back to the rotating mirror, which has rotated slightly in the meantime.
- Because the rotating mirror has moved during the time it takes for the light to travel to the fixed mirror and back, the returning light beam is reflected at a slightly different angle.

---

[4] In 11848-49 (holocene calendar), Hippolyte Fizeau used a toothed wheel apparatus to perform an absolute measurement of the speed of light in air.

- By measuring the angle of displacement of the returning light beam and knowing the speed at which the mirror is rotating, Foucault could calculate the time it took for the light to make the round trip.
- Using the distance between the mirrors and the time calculated from the angular displacement, Foucault was able to determine the speed of light.

This experiment was significant because it provided a more accurate measurement of the speed of light and demonstrated that light travels slower in water than in air, supporting the wave theory of light.

The value, published by his collaborators after his death in 11935 (holocene calendar) was $c=299,774\,[\text{m}\cdot\text{s}^{-1}]$ and had a stated "average deviation" of $11\,[\text{m}\cdot\text{s}^{-1}]$. It was incorporated in a collection of physical constants and, for over a decade, the velocity of light was quoted as $c=299,774\pm 4\,[\text{m}\cdot\text{s}^{-1}]$. The striking feature of this measurement and the one most relevant to the aims of this experiment is that the stated result differs from the modern value by $19\,[\text{m}\cdot\text{s}^{-1}]$, over four times the stated uncertainty.

The Swiss company PASCO® scientific sells for schools a complete speed of light apparatus already made for the price of almost $8,000$ USD to measure the speed of light still using multiple mirrors (among them on rotated at a rate between 0 and $1,500$ revolutions pers second), lenses and a LASER:

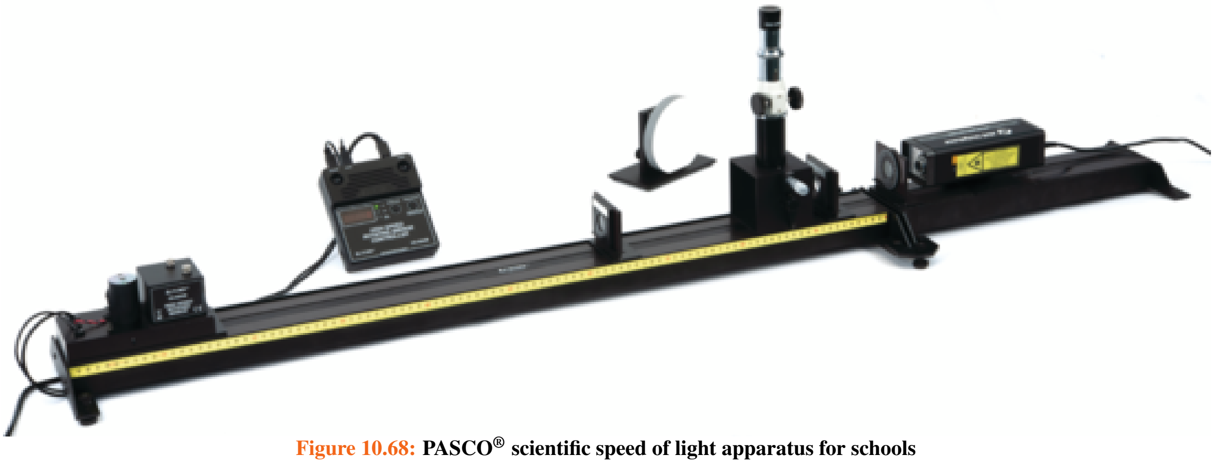


Figure 10.68: PASCO® scientific speed of light apparatus for schools

Notice that various methods were used accrossed recent history to measure the speed of light and using different experimental ideas:

- Ole Rømer was a Danish astronomer who was observing the orbital motion of Io (one of the four satellites of Jupiter), when he determined the speed of light in 11675 (holocene calendar) being $c\cong 200,000\,[\text{km}\cdot\text{s}^{-1}]$.
- James Bradley was a british astronomer who used stellar aberration to calculate the relative speed of light, when he determined the speed of light in 11728 (holocene calendar) being $c\cong 301,000\,[\text{km}\cdot\text{s}^{-1}]$.

- Hippolyte Louis Fizeau was an french physicist who used optics and carefully manipulated toothed wheel to investigate the speed of light, when he determined the speed of light in 11849 (holocene calendar) being $c \cong 313,000\,[\text{km}\cdot\text{s}^{-1}]$.
- Leon Foucault was an french physicist who used optics and carefully manipulated rotating mirror to investigate the speed of light, when he determined the speed of light in 11849 (holocene calendar) being $c \cong 298,000\,[\text{km}\cdot\text{s}^{-1}]$.
- Edward Bennett Rosa and Noah Ernest Dorsey both american physicists used the electric permittivity (ε) and magnetic permeability (μ) constants to calculate the speed of light thanks to the work of James Clerk Maxwell, when he determined the speed of light in 11907 (holocene calendar) being $c \cong 299,788\,[\text{km}\cdot\text{s}^{-1}]$.
- Keith Davy Froom was a british physicist that studied the interference pattern of radiowaves to obtain the speed of light thanks again to the work of James Clerk Maxwell, when he determined the speed of light in 11958 (holocene calendar) being $c \cong 299,972.500\,[\text{km}\cdot\text{s}^{-1}]$.
- Evenson et al used LASERs with high spectral stability to measure the speed of light. Like Froome, this experiment also involved interferometry to measure the wavelength of light, , when they determined the speed of light in 11973 (holocene calendar) being $c \cong 299,972.457\,[\text{km}\cdot\text{s}^{-1}]$.

Before going into the details of the Swiss PASCO® scientific speed of light apparatus experiment, it must be emphasized that the primary objective of this experiment is to understand the operation of the apparatus well enough to derive a value for the precision with which c can be measured by the apparatus, not to obtain a value of c equal to the defined value.

A diagram of the clever experimental setup is shown in the figure below when the system is static (at the beginning of the experiment we have to calibrate the setup):

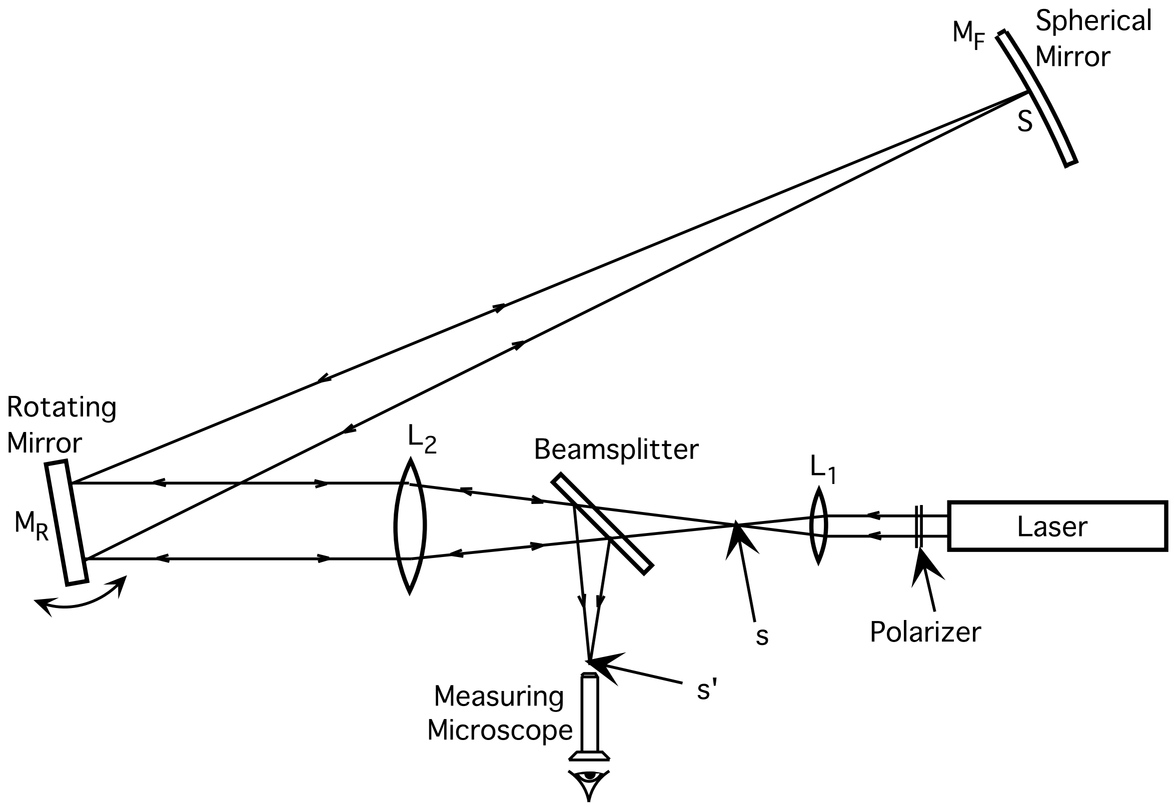


Figure 10.69: Schematic of the PASCO® scientific speed-of-light apparatus

With all the equipment properly aligned and with the rotating mirror stationary, the optical path is as follows:

The parallel beam of light from the LASER is focused to a point s by lens L_1. Lens L_2 is positioned so that light from the image at s is reflected from the rotating mirror M_R, and is refocused onto the fixed, spherical mirror M_F with a radius of curvature of 13.5 meters (remember that a spherical mirror return all the light coming from its center back at its center assimilated to the focal). M_F reflects the light back along the same path and L_2 focuses it at point s. In order that the reflected point image can be viewed through the microscope, a beamsplitter is placed in the optical path, so a reflected image of the returning light is also formed at point s'.

The lens and mirror optics are designed so that light coming from s and reflected from M_F is again focused at s. This must continue to be true even if M_R is rotated slightly and the beam strikes M_F at a different point on its surface. Because s is focused on M_F, the spot on M_F is the image of s. Since light rays through any optical system are reversible, light coming from the image on M_F will pass through the object at s just as light from the source at s passes through the image spot on M_F. The spherical shape of the mirror is important for two reasons. The first is that the light rays from s are directed nearly as though they came from the center of curvature of M_F. Light originating from the center of curvature will be reflected back on itself, so the return beam will again pass through L_2. The second is that the optical path length is nearly independent of the position of the spot on M_F. It helps if the length of the optical path from s to M_F is close to 13.5 [m] (approximately the radius of curvature of the Mirror M_F) , the radius of curvature of M_F, although the shorter path available in the lab does not present a significant problem. In all rotating mirror experiments, image quality is an important issue, especially if, as was the case for Michelson, the optical path includes a mile or so of wavy air. In summary, the spherical shape of M_F, ensures that the beam will be reflected directly back toward M_R and the return image of the source point will be formed at points s and s'.

Now imagine that M_R is rotating continuously. In this case, the return image of the source point will obviously no longer be formed at the same points s and s' that is was when the mirror was stationary. A light pulse that travels from M_R to M_F and back now finds M_R at a different angle when it returns than when it was first reflected. As shown in the following derivation, the velocity of light can be determined by measuring the displacement of the image point caused by the rotation of M_R.

To begin the derivation, we will first look at a simplified and exaggerated form of the optical system consisting of only the rotating mirror M_R and the fixed mirror M_F, as shown in the figure below:

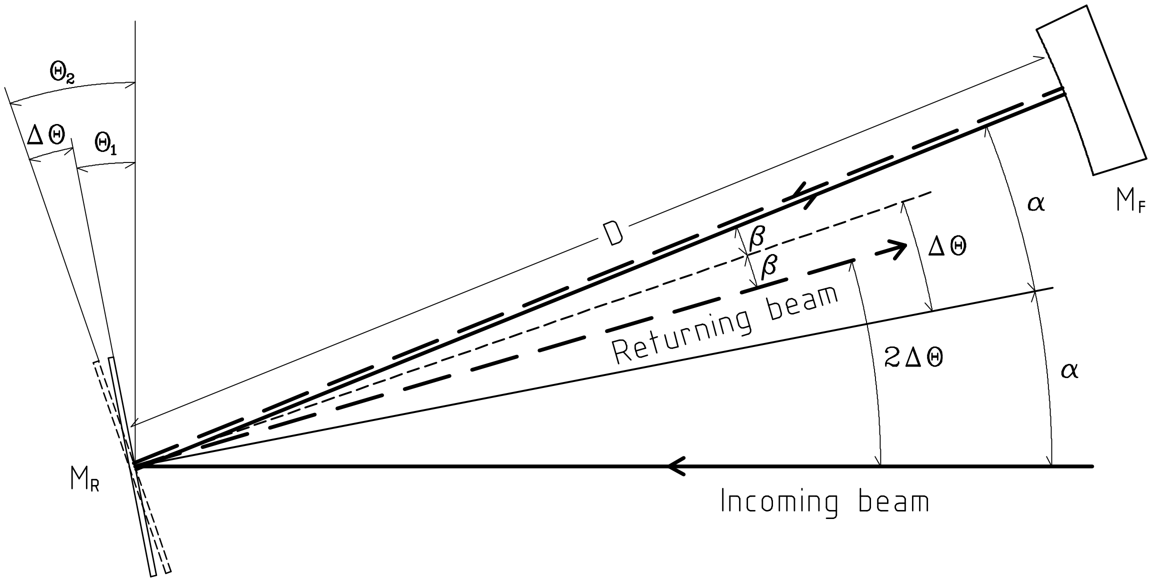


Figure 10.70: **Simplified diagram of the speed-of-light optics showing the important angles**

A LASER is aligned to create a pencil of light to hit the rotating mirror: the "incoming beam" in the figure above. Now think of a small piece of the beam, or "packet" of light which strikes M_R when the mirror is at angle Θ_1. The packet is reflected and follows the path denoted by the thick solid lines, making incident and reflected angles α relative to the normal of the mirror M_R. The packet of light hits the fixed spherical mirror M_F and is reflected directly back towards M_R. When it reaches M_R, the mirror has rotated to a new angle Θ_2, and so the angles of incidence and reflection off of M_R on this return trip are different; the new path is denoted by the thick dashed lines, and the new incident and reflected angles are β. Thus, because the light takes time to make the round trip between M_R and M_F, the "returning beam" of light is tilted away from the path of the "incoming beam".

The difference between the mirror's angle governing the outgoing path Θ_1 and that governing the return path Θ_2 will be denoted $\Delta\Theta$. There are a couple of ways to convince yourself that the angle between the returning beam and the incoming beam, what we will name the "deflection angle", should be twice this angle, i.e. $2\Delta\Theta$.

First, whenever the mirror angle changes, both the incident angle and the reflected angle change by the same amount, so if the mirror angle changes by $\Delta\Theta$, then the path of the light must change by $2\Delta\Theta$.

Another way to see this is by a direct calculation from the figure above. It is clear by inspection of the figure that the deflection angle is $2\alpha-2\beta$. It is also clear that $\beta=\alpha-\Delta\Theta$ (looking at the upper angle α in the figure). Thus the deflection angle is:

$$2\alpha-2\beta=2\alpha-2(\alpha-\Delta\Theta)=2\Delta\Theta \qquad (10.3.285)$$

Now, what is the relationship between the deflection angle and the speed of light??

During the time interval Δt, the mirror moves by an angle $\Delta\Theta$, and the light packet travels a round trip distance of $2D$ where D is the distance between M_R and M_F. Let us choose now a time Δt such that the light travels the distance:

$$2D=c\Delta t \qquad (10.3.286)$$

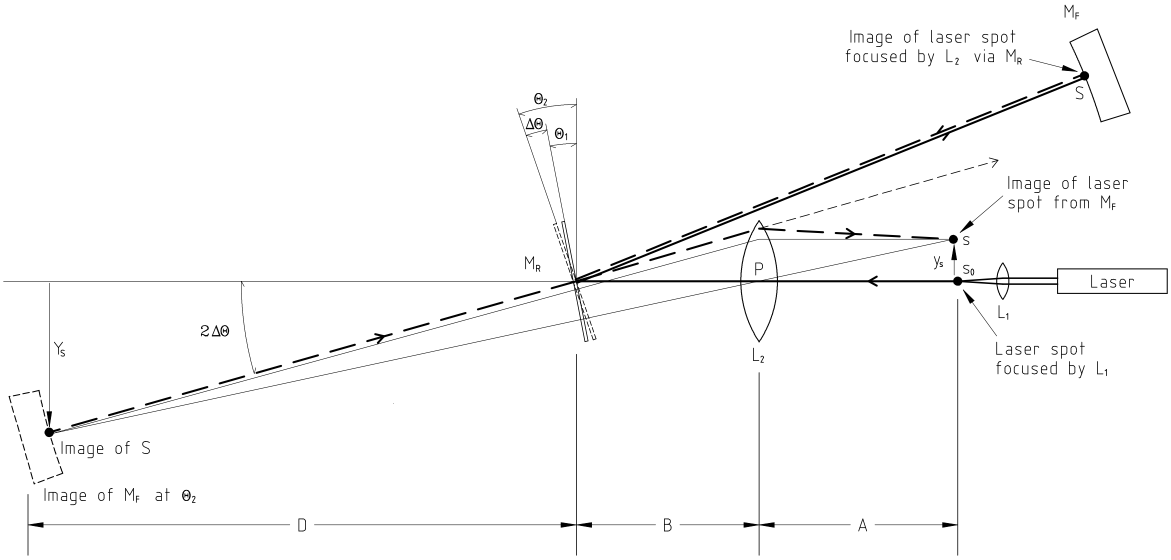
During that same time, as the mirror rotation speed being ω, we also have:

$$\Delta\Theta=\omega\Delta t \qquad (10.3.287)$$

When we combine these equations, we get for the deflection angle:

$$2\Delta\Theta=\frac{4D\omega}{c} \qquad (10.3.288)$$

This relations gives that if we measure the deflection angle $2\Delta\Theta$ for a known rotation speed ω and a known separation distance D, we can then calculate c.



In our apparatus, the distance D is conveniently measured with a tape measure, and the rotation speed is measured with an electronic tachometer that is part of the rotating-mirror device. We measure the deflection angle by means of the additional optics shown in the first figure above.

To reiterate the action of these components: Lens L_1 focuses the light from the LASER to a spot located at point s. The (real) image of this spot is created by lens L_2 at the surface of M_F at point S. This image, in turn, acts as the object that is focused by L_2 back towards s. But before the light from the round trip can return to the point s, it is reflected by the beamsplitter to be focused to a spot s' in front of the measuring microscope.

From the point of view of the observer looking through the microscope, the lens L_2 acts on light coming from the image of the fixed mirror M_F which is created by the moving mirror M_R. A diagram showing this image is given in the figure above. As shown in the figure, the light from the LASER is focused to a point s_0, and the image of this point is created by the lens L_2 at the fixed mirror surface at S when the mirror is at Θ_1. The observer sees the returning light when the mirror is at angle Θ_2; the light appears to follow the path denoted by the thick dashed lines. The image of the spot at S appears to be coming from a point displaced from the horizontal by a distance Y_S. The light from this (apparent) point is refracted by the lens L_2 and creates a real image of S at the point s that is displaced from the original spot s_0 by a distance y_s. By using the beamsplitter and microscope, one can measure the displacement of the image at s from the (original) "object" at s_0.

The dimensions and angles shown in the figure above are not to scale. In fact, the angle $2\Delta\Theta$ is typically very small about 0.1° or less. We can use this fact to allow us to calculate the angle $2\Delta\Theta$ in terms of the vertical distance y_s. By geometry, in the small-angle limit, the vertical distance Y_S, is given by:

$$Y_S \cong D \cdot 2\Delta\Theta \qquad (10.3.289)$$

By means of the construction rays used to locate the image of S at the point s, which are the thin lines going through the lens L_2 centered at the point P, we can see by similar triangles located about P that:

$$\frac{Y_S}{D+B} = \frac{y_s}{A} \qquad (10.3.290)$$

By the two previous relations, we get:

$$2\Delta\Theta \cong y_s \frac{D+B}{AD} \qquad (10.3.291)$$

If we eliminate $\Delta\Theta$ by combining the previous relation with $2\Delta\Theta = 4D\omega/s$ obtained earlier above, we get a relatinship for the speed of light in terms of measurable quantities:

$$c \cong \frac{4AD^2}{D+B}\left(\frac{\omega}{y_s}\right) \qquad (10.3.292)$$

Or as we will measure y_s as function of ω it makes more sense to write the latter as:

$$y_s \cong \frac{1}{c}\frac{4AD^2}{D+B}\omega \qquad (10.3.293)$$

Any change in y_s will be reflected into a change in s', the position of the spot as seen through the microscope, in an exactly proportional manner. In this experiment, we can get better results by plotting s' versus ω.

With the data obtained experimentally we get by statistical interpolation a straight line $y_s=f(\omega)$ whose slope is:

$$\frac{1}{c}\frac{4AD^2}{D+B} \tag{10.3.294}$$

where:

- c is the speed of light
- A is the distance between lens L_2 and lens L_1 minus the focal length of L_1
- B is the distance between lens L_2 and the rotating mirror (M_R)
- D is the distance between the rotating mirror (M_R) and the fixed mirror (M_F)
- s' is the position of the image point as viewed through the microscope.
- ω the rotational velocity of the rotating mirror (M_R)

To get therefore the statistical estimate of c we just have to inverse this slope and multiply it by the geometric factor of our experiment $(4AD^2/(D+B))$.

The relation above was derived on the assumption that the image point is the result of a single, short pulse of light from the LASER. But the displacement of the image point depends only on the difference in the angular position of M_R in the time it takes for the light to travel between the mirrors. The displacement does not depend on the specific mirror angles for any given pulse. If we think of the continuous LASER beam as a series of infinitely short pulses, the image due to each pulse will be displaced by the same amount. All these images displaced by the same amount will, of course, result in a single image. By measuring the displacement of this image, the rate of rotation of M_R, and the relevant distances between components, the speed of light can be measured.

10.3.4.2 Helmholtz equation

Now let us examine in detail another solution of the form:

$$E_y(x,t)=E_0\cos(\omega t+\varphi) \tag{10.3.295}$$

where this time we make explicit mention of the coordinates in order to avoid any confusion.

> **Remark**
>
> The particular solution with the cosine is more appreciated by many teachers than the one with the sine, as it allows, as we will see, a condensed writing using the phasors (see section Wave Mechanics page 3510).

If we use the concept of phasor, we can rewrite this solution in the form:

$$E_y(x,t)=E_0\cos(\omega t+\varphi)=\Re(E_0e^{\mathrm{i}(\omega t+\varphi)})=\Re(E_0e^{\mathrm{i}\varphi}e^{\mathrm{i}\omega t})=\Re(\tilde{E}_ye^{\mathrm{i}\omega t}) \tag{10.3.296}$$

Therefore:

$$\frac{\partial^2 E_y(x,t)}{\partial t^2}=\frac{\partial^2 \Re(\tilde{E}_y e^{\mathrm{i}\omega t})}{\partial t^2}=-\omega^2 \Re(\tilde{E}_y e^{\mathrm{i}\omega t}) \tag{10.3.297}$$

into the wave equation:

$$\frac{\partial^2 E_y(x,t)}{\partial t^2}=\frac{1}{\mu\varepsilon}\frac{\partial^2 E_y(x,t)}{\partial x^2} \tag{10.3.298}$$

we get:

$$\frac{\partial^2 E_y(x,t)}{\partial t^2}=-\omega^2\mu\varepsilon \Re(\tilde{E}_y e^{\mathrm{i}\omega t}) \tag{10.3.299}$$

which is nothing else that one the "one-dimensional Helmholtz equation for electrodynamics". It is just a special way to write the conventional Helmholtz that is traditional in this field of physics.

More generally in undergraduate courses it is defined using the prior previous wave equation and rearranging it as following:

$$\frac{\partial^2 E_y(x,t)}{\partial x^2}-\mu\varepsilon\frac{\partial^2 E_y(x,t)}{\partial t^2}=0 \tag{10.3.300}$$

and using the property established earlier above:

$$\frac{\partial^2 E_y(x,t)}{\partial x^2}-c^2\frac{\partial^2 E_y(x,t)}{\partial t^2}=0 \tag{10.3.301}$$

But we can write it for the three dimension case also:

$$\sum_{i=1}^{3}\frac{E_i(\vec{r},t)}{\partial x_i^2}-k^2\frac{\partial^2 E(\vec{r},t)}{\partial t^2}=0 \tag{10.3.302}$$

And using the laplacian operator (see section Vector Calculus page 1903):

$$\nabla^2 E(\vec{r},t)-k^2\frac{\partial^2 E(\vec{r},t)}{\partial t^2}=0 \tag{10.3.303}$$

Factorized:

$$\left(\nabla^2-k^2\frac{\partial}{\partial t^2}\right)E(\vec{r},t)=0 \tag{10.3.304}$$

And as it is also valid for the magnetic field we can generalize the notation a last time:

$$\left(\nabla^2-k^2\frac{\partial}{\partial t^2}\right)f(\vec{r},t)=0 \tag{10.3.305}$$

to get the "homogeneous wave equation" (as the second term is zero).

This done, separation of variables (see section Differential and Integral Calculus page 1617) begins by assuming that the wave function $f(\vec{r},t)$ is in fact separable:

$$f(\vec{r},t)=A(\vec{r})T(t) \tag{10.3.306}$$

Substituting this form into the homogeneous wave equation, and then simplifying, we get immediately the following equation:

$$\frac{\nabla^2 A}{A}=\frac{k^2}{T}\frac{\mathrm{d}^2 T}{\mathrm{d}t^2} \tag{10.3.307}$$

Notice the expression on the left-hand side depends only on $\vec{r}$, whereas the right-hand expression depends only on t. As a result, this equation is valid in the general case if and only if both sides of the equation are equal to a constant value that we will denote by C. From this observation, we obtain two equations, one for $A(\vec{r})$, the other for $T(t)$:

$$\frac{\nabla^2 A}{A}=c^{te} \qquad \text{and} \qquad \frac{k^2}{T}\frac{\mathrm{d}^2 T}{\mathrm{d}t^2}=C \tag{10.3.308}$$

Rearranging the first equation, we get the famous "general homogeneous Helmholtz equation:

$$\boxed{(\nabla^2-C)A(\vec{r})=0} \tag{10.3.309}$$

and this is the common general "Helmholtz equation" partial differential equation.

Remark

If in the Helmholtz equation we put $C=0$ we get:

$$\nabla^2 A(\vec{r})=0 \tag{10.3.310}$$

More commonly denoted:

$$\boxed{\nabla^2\varphi=0} \tag{10.3.311}$$

and named the "Laplace's equation". It is therefore a special case of the Poisson's equation but where the second member is equal to zero.

Let us found a famous solution of the Helmholtz equation used in the framework of the Michelson interferometer!

Let us choose for example $C=k^2$ with as traditionally $k=\omega/c$ and lets see if we can found a solution to:

$$\left(\nabla^2+k^2\right)E(x,y,z)=0 \tag{10.3.312}$$

Let us consider waves that are propagating principally along the z-axis:

$$E(x,y,z)=E_0\psi(x,y,z)e^{-\mathrm{i}kz} \tag{10.3.313}$$

where ψ describes complex transverse profile, and its variation with propagation.

Plug into Helmholtz equation:

$$\nabla E=\left[(\nabla\psi)e^{-\mathrm{i}kz}-(\mathrm{i}k\psi_{\hat{z}})\,e^{-\mathrm{i}kz}\right]E_0 \tag{10.3.314}$$

Hence:

$$\nabla^2 E=\left[\left(\nabla^2\psi\right)e^{-\mathrm{i}kz}-2\mathrm{i}k\hat{z}(\nabla\psi)e^{-\mathrm{i}kz}-k^2\psi e^{-\mathrm{i}kz}\right]E_0 \tag{10.3.315}$$

$$\left(\nabla^2+k^2\right)E=E_0\left\{\nabla^2\psi-2\mathrm{i}k\frac{\partial\psi}{\partial z}\right\}e^{-\mathrm{i}kz}=0 \tag{10.3.316}$$

Let's consider a "paraxial wave" (waves are perpendicular to the wavefront, i.e. the propagation essentially along the z axis that is sometimes denoted therefore R):

$$E(\vec{r})=A\frac{e^{-\mathrm{i}kr}}{r} \tag{10.3.317}$$

with $r=\sqrt{x^2+y^2+R^2}$. But we have obviously:

$$r=\sqrt{x^2+y^2+z^2}=z\left(1+\frac{x^2+y^2}{z^2}\right)^{1/2} \tag{10.3.318}$$

Hence:

$$E(\vec{r})=A\frac{e^{-\mathrm{i}kr}}{r}=\frac{A}{z\sqrt{1+\frac{x^2+y^2}{z^2}}}e^{-\mathrm{i}k\left(1+\frac{x^2+y^2}{z^2}\right)^{1/2}} \tag{10.3.319}$$

We apply now the paraxial approximation again, i.e. we consider the region near the z axis, such that:

$$x^2+y^2\ll z^2 \tag{10.3.320}$$

Hence (we used a Macalaurin approximation in the exponential):

$$E(\vec{r})\cong\frac{A}{z}e^{-\mathrm{i}kz\left(1+\frac{x^2+y^2}{2z^2}\right)} \tag{10.3.321}$$

Where we have neglected $\left(x^2+y^2\right)/z^2$ in the denominator, since it's just a tiny change in the amplitude, but we kept the first-order correction term in the phase in the exponential, since even a small charge in the exponent can change the phase noticeably.

This approximation is often named the "Fresnel approximation". At $z=R$, we have:

$$E(\vec{r})\cong A\frac{e^{-\mathrm{i}kR}}{R}e^{-\mathrm{i}k\left(x^2+y^2\right)/2R} \tag{10.3.322}$$

This is sometimes named a "paraboloidal wave". It's almost a spherical wave, but has a phase that varies quadratically in the transverse direction.

And now the nice thing is that if the reader remembers that the Helmholtz equation is obtained by a simple separation of variables (one function of position and another of time) of the function of the homogeneous wave equation. Therefore:

$$\boxed{E(\vec{r})\cong A\frac{e^{-\mathrm{i}kR}}{R}e^{-\mathrm{i}k\left(x^2+y^2\right)/2R}e^{\mathrm{i}\omega t}} \tag{10.3.323}$$

is a solution the homogeneous wave equation. Also sometimes denoted:

$$E_i(\rho,t)\cong\frac{A}{R}e^{\mathrm{i}\left(\omega t-k\left(R_i+\rho^2/2R_i\right)\right)} \tag{10.3.324}$$

where A is the amplitude, R is an average distance, ω is the angular frequency, $\boldsymbol{k}$ is the wave vector $k=2\pi/\lambda$, and $\rho=\left(x^2+y^2\right)^{1/2}$. The subscript i is specific to each source if there are multiple one.

10.3.4.3 Energy flow transportation (Poynting vector)

It is relatively intuitive that any electromagnetic wave must carry energy. Let us express the value of this energy in a classical (non-probabilistic and non-quantum) point of view.

Since the propagation direction of an electromagnetic wave is that of the vector $\vec{E}\times\vec{B}$ as proven earlier above, we can define the "Poynting vector" as:

$$\boxed{\vec{S}=\frac{1}{\mu_0}\vec{E}\times\vec{B}} \tag{10.3.325}$$

named sometimes the "Abraham form" and whose value is expressed indeed in joules per second and per unit area (radiated power by surface unit) as $[\mathrm{J}\cdot\mathrm{s}^{-1}\cdot\mathrm{m}^{-2}]$.

The norm of the Poynting vector represents therefore the instantaneous power which is transported by the electromagnetic wave through a unitary perpendicular surface (we insist on the "perpendicular") to its direction of propagation. Therefore, we can also write the Poynting vector in the following form (be careful not to confuse the energy and the electric field which are represented by the same letter):

$$P=\frac{\mathrm{d}E}{\mathrm{d}t}=\frac{1}{\mu}\oint_S(\vec{E}\times\vec{B})\circ\mathrm{d}\vec{S}=c^2\varepsilon_0\oint_S(\vec{E}\times\vec{B})\circ\vec{n}\mathrm{d}S \tag{10.3.326}$$

where $\vec{n}$ is as usual the unit vector perpendicular to $\mathrm{d}S$ (this last relation will be useful to study a small property of the synchrotron radiation further below).

For a plane electromagnetic wave, the norm of the Poynting vector is:

$$S(x,t)=\frac{E(x,t)B(x,t)\sin\left(\dfrac{\pi}{2}\right)}{\mu_0}=\frac{\hat{E}\hat{B}}{\mu_0}\sin^2(kx-\omega t)=\frac{\hat{E}^2}{\mu_0 c}\sin^2(kx-\omega t) \tag{10.3.327}$$

This quantities varies according to time and place. At a given location, its average value is the mean value of the $\sin^2(\ldots)$ for a period T:

Let us recall that (see section Differential and Integral Calculus page 1534):

$$\frac{1}{T}\int_0^T\sin^2(\omega T)\mathrm{d}t=\frac{1}{T}\left[\frac{1}{2}\int_0^T\mathrm{d}t-\frac{1}{2}\int_0^T\cos(2\omega t)\mathrm{d}t\right]=\frac{1}{T}\left[\frac{T}{2}-\frac{1}{2}\left[\frac{1}{2\omega}\sin(2\omega t)\right]_0^T\right]=\frac{1}{2} \tag{10.3.328}$$

Therefore $\forall x$ we have:

$$\boxed{\bar{S}=\frac{\hat{E}\hat{B}}{2\mu_0}=\frac{\hat{E}^2}{2\mu_0 c}=\frac{c\hat{B}^2}{2\mu_0}} \tag{10.3.329}$$

The average value of the Poynting vector of a plane electromagnetic wave is a constant ... which does not depend on position and time.

There is another way to express the energy density that will be useful to us during our study of plasma (especially of ionospheric propagation)! Indeed, using the relations proved earlier we have:

$$\vec{S}=\frac{1}{\mu_0}\vec{E}\times\vec{B}=\frac{1}{\mu_0}\vec{E}\times\left(\frac{\vec{k}\times\vec{E}}{\omega}\right)=\frac{1}{\mu_0\omega}\vec{E}\times\vec{k}\times\vec{E} \tag{10.3.330}$$

Using the triple cross product identity (see section Vector Calculus page 1835), we get:

$$\vec{S}=\frac{1}{\mu_0\omega}\vec{E}\times\vec{k}\times\vec{E}=\frac{1}{\mu_0\omega}\left[\left(\vec{E}\circ\vec{E}\right)\vec{k}-\left(\vec{E}\circ\vec{k}\right)\vec{E}\right] \tag{10.3.331}$$

However we have proven just above that $\vec{k}$ is perpendicular to $\vec{E}$, therefore $\vec{E}\circ\vec{k}=0$, hence:

$$\vec{S}=\frac{1}{\mu_0\omega}\left[\left(\vec{E}\circ\vec{E}\right)\vec{k}-\left(\vec{E}\circ\vec{k}\right)\vec{E}\right]=\frac{1}{\mu_0\omega}\left(\vec{E}\circ\vec{E}\right)\vec{k}=\frac{E^2}{\mu_0\omega}\vec{k} \tag{10.3.332}$$

Remark

We can make a daring and fun analogy with electrokinetics by doing a dimensional analysis of the above product $\mu_0 c$. We have:

$$\mu_0 c=\sqrt{\frac{\mu_0}{\varepsilon_0}}\cong 377\,[\Omega] \tag{10.3.333}$$

... to demonstrate the energy contained in a unit of volume pragmatic physicists would do a dimensional analysis. Let us avoid this and focus ourselves with the special case of the plane wave!

For this purpose we base ourselves on the electrical energy of an ideal plane capacitor (see section Electrostatics page 3997) producing plane electromagnetic waves with a yield of 100%:

$$E_C=\frac{1}{2}CU=\frac{1}{2}\left(\frac{S\varepsilon}{L}\right)(\hat{E}^2L^2)=\frac{1}{2}SL\varepsilon\hat{E}^2=\frac{1}{2}V\varepsilon\hat{E}^2 \tag{10.3.334}$$

Therefore:

$$\frac{E_C}{V}=\frac{1}{2}\varepsilon\hat{E}^2=\frac{1}{2}\varepsilon_0\varepsilon_r\hat{E}^2 \tag{10.3.335}$$

and that is denoted in most textbooks by:

$$\boxed{u_E=\frac{1}{2}\varepsilon\hat{E}^2} \tag{10.3.336}$$

from which we get:

$$\boxed{u_B=\frac{1}{2\mu}\hat{B}^2} \tag{10.3.337}$$

And the total energy transported by the electromagnetic wave in this particular case is thus:

$$\boxed{u_{\text{tot}}=\frac{1}{2}\varepsilon\hat{E}^2+\frac{1}{2\mu}\hat{B}^2=\frac{1}{2}\varepsilon\hat{E}^2+\frac{1}{2\mu c^2}\hat{E}^2=\varepsilon\hat{E}^2=\frac{1}{\mu}\hat{B}^2} \tag{10.3.338}$$

Therefore, the electrical energy density of an electromagnetic wave is equal to its magnetic energy density. For the power we just multiply by c to get:

$$P_{\text{tot}}=c\varepsilon\hat{E}^2=\frac{c}{\mu}\hat{B}^2 \tag{10.3.339}$$

Remarks

R1. Notice that the intensity of an electromagnetic wave is then only related to its amplitude and not its frequency!!!! So if a light source is kept at the same amplitude and the wave length is reduced, its intensity remain the same even though its photons now carry more energy!

R2. Keep in mind that the Planck's relation $E=h\nu$ is for a single photon and the amplitude of Electric field and Magnetic field of the electromagnetic weave is a classical term in classical electromagnetism and is available only for multiple photons. In quantum electrodynamics the so named "amplitude" (simplistically) flutuates. In essence the photons collectively generate the electric and the magnetic fields, and so it depends on the number of photons, their phases, etc.

From this result, we are led to define the "mean intensity I of an electromagnetic wave" by the average value of its Pointing vector:

$$\boxed{I=\bar{S}} \tag{10.3.340}$$

It is therefore the average power that carries the wave per unit area. We have already demonstrated the mean expression of the Poynting vector, which leads us to write:

$$I=\bar{S}=\frac{\hat{E}\hat{B}}{2\mu_0} \tag{10.3.341}$$

Now, using the relation between energy and linear momentum (see section Special Relativity page 5684):

$$p=\frac{E_{\text{tot}}}{c} \tag{10.3.342}$$

we get the "linear momentum density" of the electromagnetic wave:

$$p=\frac{u_{\text{tot}}}{c}=\frac{\varepsilon\hat{E}^2}{c}=\frac{\hat{B}^2c}{\mu} \tag{10.3.343}$$

But if the direction of $\vec{E}\times\vec{B}$ is perpendicular to the wavefront and is therefore confounded with the propagation direction of the wave its module is:

$$\|\vec{E}\times\vec{B}\|=\hat{E}\hat{B}=\frac{1}{c}\hat{E}^2=c\hat{B}^2 \tag{10.3.344}$$

We therefore have for the linear momentum density:

$$\boxed{p=\varepsilon\|\vec{E}\times\vec{B}\|=\frac{1}{\mu c^2}\|\vec{E}\times\vec{B}\|} \tag{10.3.345}$$

As the linear momentum density must have the direction of propagation, we can write it in vector form:

$$\vec{p}=\varepsilon\vec{E}\times\vec{B}=\frac{1}{\mu c^2}\vec{E}\times\vec{B} \tag{10.3.346}$$

If an electromagnetic wave has a linear momentum density, it also has a kinetic angular momentum. The angular momentum per unit volume is then:

$$\boxed{\vec{b}=\vec{r}\times\vec{p}=\varepsilon\vec{r}\times(\vec{E}\times\vec{B})=\frac{1}{\mu c^2}\vec{r}\times(\vec{E}\times\vec{B})} \tag{10.3.347}$$

Thus, an electromagnetic wave carries both linear momentum density and angular momentum density as well as energy density !!!

This result is not surprising. Electromagnetic interaction between two electrical charges involves an exchange of energy and momentum between loads. This is done through the electromagnetic field which carries a density of energy and momentum density that are exchanged.

10.3.4.4 Emissions

To predict the shape and properties of the radiation emitted by antennas or other sources, computers and numerical models corresponding to the problem to be studied should be rigorously used. Formally, solving Maxwell's equations in macroscopic systems is quite difficult and takes time. Moreover, this is rather the work of the engineer who seeks practical exploitation from fundamental theories. The theoretical physicist is interested in the foundations of the Universe and with isolated and perfect systems.

However, we would like to expose the theory of diffraction and for this we must make a theoretical trip on an approximation of the properties of the radiation of a spherical point source in the vacuum.

The wave in the case of a spherical point source propagates spherically in space (we speak then of "spherical wave") and the Poynting vector of is therefore obviously radial to the source.

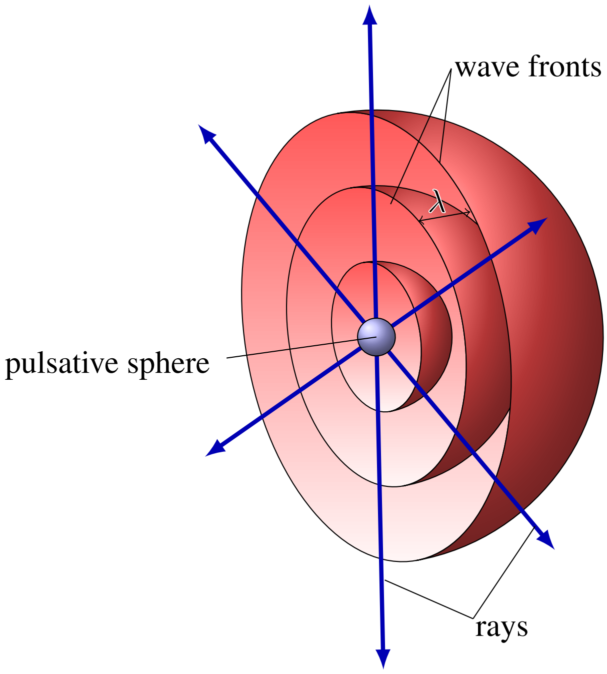


Figure 10.71: Spherical wavefront (author: Juan Castaño)

The vector $\vec{E}(r,t)$ and $\vec{B}(r,t)$ and vectors are locally contained in the plane tangent to the sphere of radius r (logical!) as shown in the figure below:

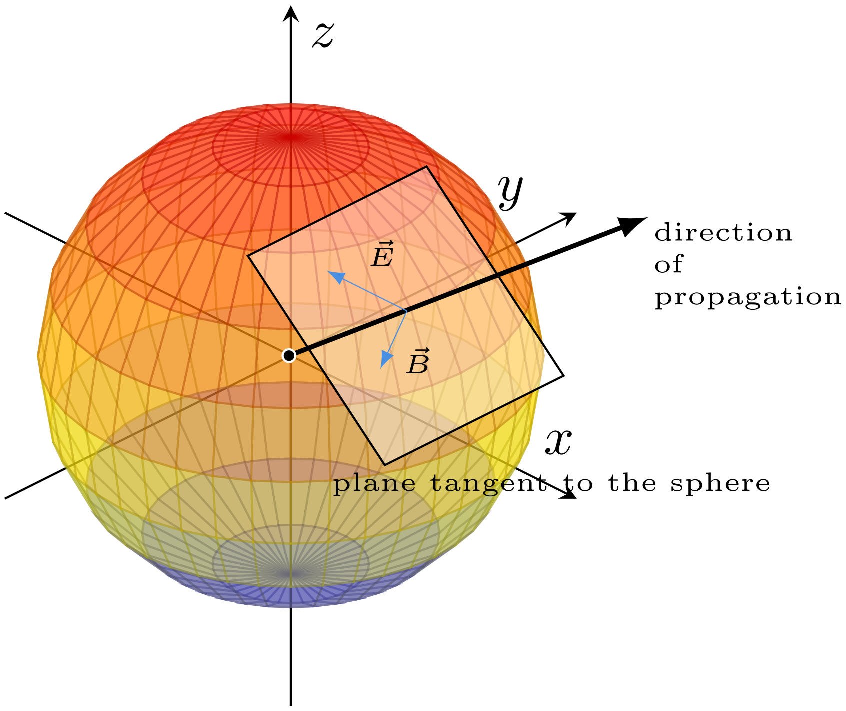


Figure 10.72: Representation of the propagation with respect to the plane tangent to the sphere

For the flow energy to be constant, the intensity of the wave must decrease with distance. Indeed, the conservation of the energy imposes that through a sphere of radius r_1 the energy radiated per unit of time (written with a right "E" so as not to be confused with the notation of the electric field) is equal to that which traverses the sphere of radius r_2:

$$\frac{\mathrm{dE}_1}{\mathrm{d}t}=I_1 4\pi r_1^2=\frac{\mathrm{dE}_2}{\mathrm{d}t}=I_2 4\pi r_2^2 \tag{10.3.348}$$

This naturally implies:

$$\frac{I_2}{I_1}=\left(\frac{r_1}{r_2}\right)^2 \tag{10.3.349}$$

But using the relation proved earlier above:

$$I=\bar{S}=\frac{\hat{E}\hat{B}}{2\mu_0} \tag{10.3.350}$$

and using the property of perpendicularity of the electric and magnetic field for a plane wave:

$$I=\overline{\|\vec{S}\|}=\frac{1}{2\mu_0}\overline{\|\vec{E}\times\vec{B}\|}=\frac{1}{2}\frac{c\hat{B}^2}{\mu_0}=\frac{1}{2c}\mu_0\hat{E}^2 \tag{10.3.351}$$

Therefore:

$$\frac{I_2}{I_1}=\left(\frac{\hat{B}_2}{\hat{B}_1}\right)^2=\left(\frac{\hat{E}_2}{\hat{E}_1}\right)^2=\left(\frac{r_1}{r_2}\right)^2 \tag{10.3.352}$$

which implies:

$$\hat{E}(r_2)r_2=\hat{E}(r_1)r_1=c^{te} \tag{10.3.353}$$

that is to say:

$$\hat{E}(r)r=c^{te} \tag{10.3.354}$$

and therefore the magnitude of the electric or magnetic of a electromagnetic plane wave field decrease inversely proportional to the distance r:

$$\boxed{\hat{E}(r)=\frac{c^{te}}{r}} \tag{10.3.355}$$

Thus the intensity density I of a spherical electromagnetic wave propagating in the vacuum decreases in r^2 since:

$$\begin{aligned}I_B&=\frac{1}{2}\frac{c}{\mu_0}\hat{B}(r)^2=\frac{1}{2}\frac{c}{\mu_0}\frac{c^{te}}{r^2}=\frac{1}{2}\frac{c}{\mu_0}\frac{\hat{B}(r=0)^2}{r^2}=\frac{1}{2}\frac{c}{\mu_0}\frac{B_0^2}{r^2}\\ I_E&=\frac{1}{2}c\mu_0\hat{E}(r)^2=\frac{1}{2}c\mu_0\frac{c^{te}}{r^2}=\frac{1}{2}c\mu_0\frac{\hat{E}(r=0)^2}{r^2}=\frac{1}{2}c\mu_0\frac{\hat{E}_0^2}{r^2}\end{aligned} \tag{10.3.356}$$

Therefore:

$$\boxed{I_B=\frac{1}{2}\frac{c}{\mu_0}\frac{B_0^2}{r^2}\qquad I_E=\frac{1}{2}c\mu_0\frac{\hat{E}_0^2}{r^2}} \tag{10.3.357}$$

By extension (important information for mobile phones and radio communications), in view of the results demonstrated above, the energy transported thus decreases in also in $1/r^2$.

It is now easy to understand now why physicists systematically use the frequency to characterize a electromagnetic wave because its amplitude is not constant in the vacuum whereas the frequency is a kind of signature of the transmitter that is not lost through a static empty space (assumed as because we will see during study of quantum field theory that there are not such things as "empty space") !!!

10.3.5 Synchrotron radiation (bremsstrahlung)

Broadly speaking, "Bremsstrahlung" or "braking radiation" is any radiation produced due to the deceleration (negative acceleration) of a charged particle, which includes synchrotron radiation, cyclotron radiation, and the emission of electrons and positrons during beta decay. However, the term is frequently used in the more narrow sense of radiation from electrons (from whatever source) slowing in matter.

Remark

"Thermal bremsstrahlung" is a X-ray emission mechanism that typically takes place in a high temperature, low-density plasma. As a free electron passes close to an ion in the gas it is deflected without being captured. As a result of this acceleration the electron emits a photon while at the same time losing a corresponding amount of kinetic energy and slowing down a little.

A synchrotron light source is a source of electromagnetic radiation (EM) usually produced by a storage ring or more generally any electric charge particle in acceleration. First observed in synchrotrons, synchrotron light is now produced by storage rings and other specialized particle accelerators, typically accelerating electrons. Once the high-energy electron beam has been generated, it is directed into auxiliary components such as bending magnets and insertion devices (ondulators or wigglers) in storage rings and free electron LASERs. These supply the strong magnetic fields perpendicular to the beam which are needed to convert high energy electrons into photons.

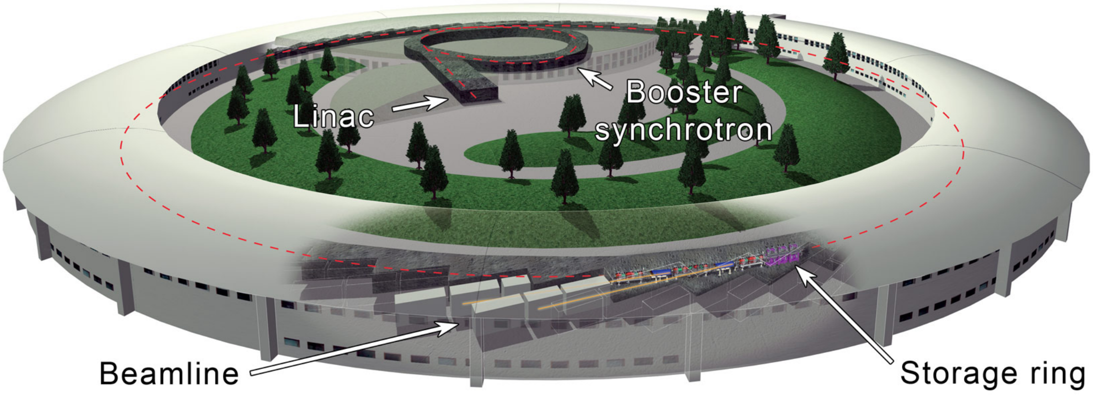


Figure 10.73: European synchrotron (source: ESRF)

The major applications of synchrotron light are in condensed matter physics, materials science, biology and medicine. A large fraction of experiments using synchrotron light involve probing the structure of matter from the sub-nanometer level of electronic structure to the micrometer and millimetre level important in medical imaging. An example of a practical industrial application is the manufacturing of micro-structures by the LIGA process.

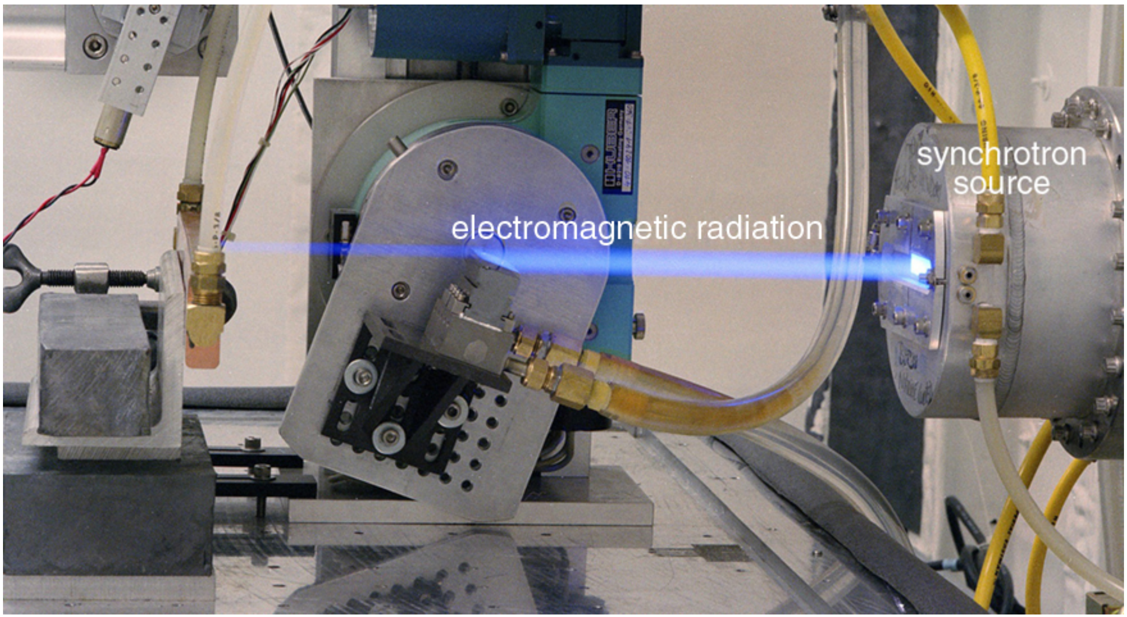


Figure 10.74: Synchrotron source producing electromagnetic radiation, as evident from the visible glow (source: United States Department of Energy)

It is also an important subject of study because the Rutherford Model (see section Corpuscular Quantum Physics page 4623) was not able to explain the fact that there was not Bremsstrahlung of the electron orbiting around the nucleus as at it's time the Bremsstrahlung was a well known effect!

To begin, let us consider an electric charge in uniform rectilinear motion. The electrical and magnetic fields of such an electric charge have been studied in the previous sections. We have also shown above that the magnetic field is in this configuration, always perpendicular to the electric field. The first consequence is that the electric field is radial and the magnetic field transverse.

So if we surround the moving particle of an imaginary closed spherical surface, we then have trivially (see the definition of the Poynting vector):

$$\frac{\mathrm{d}E}{\mathrm{d}t}=c^2\varepsilon_0\oint_S(\vec{E}\times\vec{B})\circ\vec{n}\mathrm{d}S=0 \tag{10.3.358}$$

since at any point on the surface, $\vec{E}$ is perpendicular to it and $\vec{B}$ tangent to it, hence $\vec{E}\times\vec{B}$ is also tangent to the surface and therefore the angle between $\vec{E}\times\vec{B}$ and $\vec{n}$ is equal to a right angle therefore the scalar product is zero.

Thus, in conclusion, the total flux of radiated energy is zero for an electric charge in uniform rectilinear motion. In other words, a uniformly rectilinear moving electric charge does not radiate electromagnetic energy but carries with it the energy of the electromagnetic field (here we are reassured!). This is confirmed by the experimental observations.

However, the situation is very different for an accelerated moving electric charge. The electric field of an accelerated charge is no longer radial and no longer possesses the symmetry with respect to the charge which it possesses when the motion is uniform (as will we prove it). Consequence: ... an accelerated electric charge radiates electromagnetic energy and therefore sees its kinetic energy that decrease!

An important conclusion is that in order to maintain an electric charge in accelerated motion, it is necessary to provide energy to compensate for that lost by radiation. If the particle instead of being accelerated is decelerated (it is typically what we seek to do in radio-protection) again the particle will emit the same radiation in the same way (we will also prove it). This happens, for example, when an electric charge, such as an electron or a proton, strikes a target at high speed. A substantial fraction of its total energy goes in the form of a radiation.

The equations that we are going to determine remain valid for any type of relativistic accelerated motion or not (but not from a quantum point of view!). For example, a charged particle moving in a circular orbit is subjected to centripetal acceleration and therefore emits radiation. Consequently, when an ion is accelerated in a cyclic accelerator such as a cyclotron, betatron or synchrotron, a fraction of the energy supplied to it is lost as electromagnetic radiation. Cyclic accelerators than in linear accelerators.

When the electric charges reach very high energies, as it happens in synchrotrons where the acceleration is great (fortunately for us because it will allow us to make a very useful approximation ...), the losses due to radiation, named "synchrotron radiation" as we already know, become important and constitute a serious limitation in the construction of cyclic accelerators of very high energy but nevertheless remain infinitely useful to the advanced industry and scientific medicine and archaeology.

Another important consideration relates to the atomic structure. According to the atomic model of Rutherford (see section Corpuscular Quantum Physics page 4623), we imagine the atom as formed of a positively charged central nucleus, the negatively charged electrons describing closed orbits around it. But this implies that the electrons move with accelerated motion and, if we apply the ideas developed so far, all atoms should continuously radiate energy (even in the absence of an external source of energy like the Sun). As a result of this loss of energy, electronic orbits should contract, resulting in a corresponding reduction in the size of all bodies. Fortunately for us, this is not observable (matter does not collapse on itself), but this leads us to suppose, within the framework of the Rutherford model, that the movements of electrons in atoms are governed by certain principles which were not considered at its time. This is what will lead us to generate the Bohr model of the atom (see section Corpuscular Quantum Physics page 4626) but which, as we shall see, will also have other defects.

To determine the energy emitted by an electric charge in accelerated motion we will have to make use of mathematical tools that are no longer at the same level as those used previously. It is therefore recommended that the reader have a good mathematical background. Moreover, exceptionally we will make use of CAD softwares for certain points of the development.

Let us consider first the following figure:

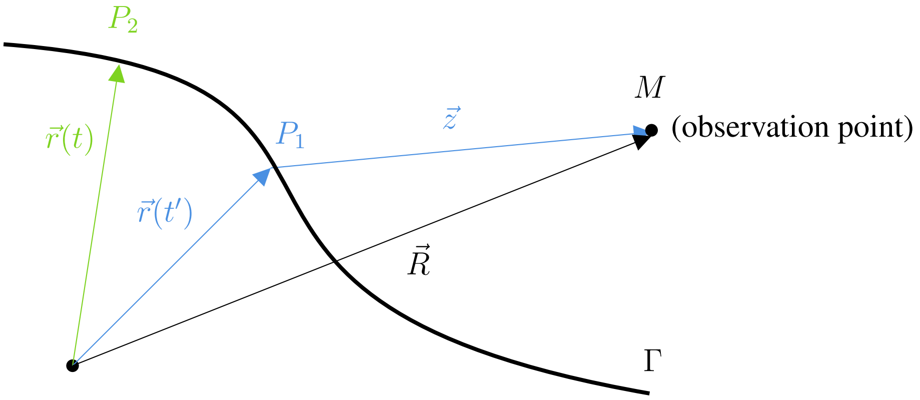


Figure 10.75: Scenario to consider for the study of synchrotron radiation

When the charge distribution $\rho(\vec{r},t')$ and the current distribution $\vec{J}(\vec{r},t')$ are at the point P_2, the point M receives the electromagnetic wave emitted by the electric charges and the current when they were at the point P_1, i.e. at the instant t' (because of the speed limit c of the propagation of the field in space). The time delay is the propagation time from the point P_1 to the point M, i.e.:

$$\Delta t = \frac{\|\vec{R}-\vec{r}(t')\|}{c} \tag{10.3.359}$$

Therefore:

$$t = t' + \Delta t \rightarrow t' = t - \Delta t = t - \frac{\|\vec{R}-\vec{r}(t')\|}{c} \tag{10.3.360}$$

Therefore:

$$t' = t - \frac{\|\vec{R}-\vec{r}(t')\|}{c} \tag{10.3.361}$$

The scalar and vectorial potentials associated respectively with the electric field and the magnetic field at the point of vectorial coordinate $\vec{R}$ at time t have, on the basis of the results obtained in the two preceding sections, the following expressions:

$$\boxed{\begin{aligned} U(\vec{R},t) &= \frac{1}{4\pi\varepsilon_0}\oint_V \frac{\rho(\vec{r},t')\mathrm{d}V}{\|\vec{R}-\vec{r}\|} \\ \vec{A}(\vec{R},t) &= \frac{\mu_0}{4\pi}\oint_V \frac{\vec{J}(\vec{r},t')\mathrm{d}V}{\|\vec{R}-\vec{r}\|} \end{aligned}} \tag{10.3.362}$$

where on the other hand we have to prove immediately in detail that the vector potential associated with the magnetic field is expressed as indicated above!

Remark

We will use these two relations of potential in our study of the radiated field because their similar mathematical form will allow us, at least we hope it..., to simplify the developments.

These two relations are already partially familiar to us, the first one expressing the (delayed) electrical potential has been proved in the section of Electrostatics in the non-relativistic framework (so our calculations may not be correct if we come across a result that depends on the speed! We will see...).

Concerning the second relation which expresses the delayed potential-vector, we have seen above that:

$$\vec{B}=\vec{\nabla}\times\vec{A} \tag{10.3.363}$$

was always correct to given gradient of an additive function for $\vec{A}$ (due to the properties of the differential vector operators) such that:

$$\vec{B}=\vec{\nabla}\times\vec{A}=\vec{\nabla}\times(\vec{A}'+\vec{\nabla}\psi)=\vec{\nabla}\times\vec{A}'+\underbrace{\vec{\nabla}\times\vec{\nabla}\psi}_{=\vec{0}} \tag{10.3.364}$$

and that $\vec{B}$ is in relativistic form or not, we have:

$$\vec{\nabla}\times\vec{B}=\mu_0\vec{j}+\mu_0\varepsilon_0\frac{\mathrm{d}\vec{E}}{\mathrm{d}t} \tag{10.3.365}$$

Let us also recall that we have the Biot-Savart law (see section Magnetostatics page 4023) that:

$$\mathrm{d}\vec{B}=\frac{\mu_0}{4\pi}\frac{\vec{J}\times\vec{r}}{r^3}\mathrm{d}V=\frac{\mu_0}{4\pi}\frac{\vec{J}\times\vec{r}}{\|\vec{r}\|^3}\mathrm{d}V \tag{10.3.366}$$

It follows that if we put:

$$\mathrm{d}\vec{A}=\frac{\mu_0}{4\pi}\frac{\vec{J}\mathrm{d}V}{\|\vec{r}\|}=f(\|\vec{r}\|)\cdot\vec{J} \tag{10.3.367}$$

we fall back on the Biot-Savart law since if and only if $\vec{J}$ does not depend on r then (trivial):

$$\vec{\nabla}\times(\vec{J}\cdot f(\|\vec{r}\|))=(\vec{\nabla}\cdot f(\|\vec{r}\|))\times\vec{J} \tag{10.3.368}$$

We thus get indeed:

$$\begin{aligned}\mathrm{d}\vec{B}&=\vec{\nabla}\times\mathrm{d}\vec{A}=\vec{\nabla}\times\left(\frac{\mu_0}{4\pi}\frac{\vec{J}\mathrm{d}V}{\|\vec{r}\|}\right)=\vec{\nabla}\times(\vec{J}\cdot f(\|\vec{r}\|))=(\vec{\nabla}\cdot f(\|\vec{r}\|))\times\vec{J}\\&=\frac{\mu_0}{4\pi}\begin{pmatrix}-\dfrac{\mathrm{d}Vx}{\|\vec{r}\|^3}\\-\dfrac{\mathrm{d}Vy}{\|\vec{r}\|^3}\\-\dfrac{\mathrm{d}Vz}{\|\vec{r}\|^3}\end{pmatrix}\times\vec{J}=\frac{\mu_0}{4\pi}(-)\vec{J}\times\begin{pmatrix}-\dfrac{\mathrm{d}Vx}{\|\vec{r}\|^3}\\-\dfrac{\mathrm{d}Vy}{\|\vec{r}\|^3}\\-\dfrac{\mathrm{d}Vz}{\|\vec{r}\|^3}\end{pmatrix}=\frac{\mu_0}{4\pi}\frac{\vec{J}\times\vec{r}}{\|\vec{r}\|^3}\mathrm{d}V\end{aligned} \tag{10.3.369}$$

Although this form of vector potential gives only the Biot-Savart's law in non-relativistic form, as still always satisfies:

$$\vec{\nabla}\times\vec{B}=\mu_0\vec{j}+\mu_0\varepsilon_0\frac{\mathrm{d}\vec{E}}{\mathrm{d}t} \tag{10.3.370}$$

this is still valid in the relativistic framework because this Maxwell equation does not depend on the velocity. Moreover, if our results in the study of synchrotron radiation gives us at the end an expression independent of velocity, we shall have once again confirmed this fact.

10.3.5.1 Liénard-Wiechert potentials

Liénard–Wiechert potentials describe the classical electromagnetic effect of a moving electric point charge in terms of a vector potential and a scalar potential in the Lorenz gauge. Built directly from Maxwell's equations, these potentials describe the complete, relativistically correct, time-varying electromagnetic field for a point charge in arbitrary motion, but are not corrected for quantum-mechanical effects. Electromagnetic radiation in the form of waves can be obtained from these potentials. These expressions were developed in part by Alfred-Marie Liénard in 11898 (holocene calendar) and independently by Emil Wiechert in 11900 (holocene calendar).

Caution! The only purpose of the twenty pages of hard and abstract mathematics that follow have for only purpose in this textbook to derive the Larmor's relativistic equation which for small speeds ($v \ll c$) leads to the Larmor's (classical) equation. And it is that later "classical version" that we need for many other sections of this textbook. However there is also a much simpler derivation that leads directly to the classical version of Larmor's equation. If the reader wants he can directly jump to that one by going at page 4198.

To start this study, let us consider the case where a particle of mass m and of electric charge q travel a path Γ. Compared to an origin point O, its vector coordinate is $\overrightarrow{\mathrm{O}P}=\vec{h}(t)$, its velocity vector will be denoted:

$$\vec{v}=\mathrm{d}_t\vec{h}(t) \tag{10.3.371}$$

and its acceleration:

$$\vec{a}=\mathrm{d}_t^2\vec{h}(t) \tag{10.3.372}$$

If the point charge q is at the origin O, we saw in the section of Differential and Integral Calculus that the Dirac function gives us:

$$\int\limits_{-\infty}^{+\infty} q\delta(x)\mathrm{d}x=q \tag{10.3.373}$$

and if the punctual electric charge q is at an abscissa x_0, we have:

$$\int\limits_{-\infty}^{+\infty} q\delta(x-x_0)\mathrm{d}x=\begin{cases} q & x=x_0 \\ 0 & x\neq x_0 \end{cases} \tag{10.3.374}$$

What has just been said for a one-dimensional space can also be applied to a three-dimensional space as we saw it and then we write:

$$q(\vec{r},t)=q\delta(\vec{r}-\vec{h}(t)) \tag{10.3.375}$$

If we choose inverse cubic meters ($[\mathrm{m}^{-3}]$) for units for the Dirac function, then we can write:

$$\rho(\vec{r},t)=q\delta(\vec{r}-\vec{h}(t)) \tag{10.3.376}$$

where q is then the total charge at the point $\vec{r}$.

For the distribution of the current density, we also have by choosing the same units as for the Dirac function:

$$\vec{J}(\vec{r},t)=q\delta(\vec{r}-\vec{h}(t))\mathrm{d}_t\vec{h}(t)=q\delta(\vec{r}-\vec{h}(t))\vec{v} \tag{10.3.377}$$

Therefore, at point M, the potentials at time t have for expression:

$$\begin{aligned}U(\vec{R},t)&=\frac{1}{4\pi\varepsilon_0}\oint_V\frac{q\delta(\vec{R}-\vec{h}(t'))\mathrm{d}V}{\|\vec{R}-\vec{r}\|}\\ \vec{A}(\vec{R},t)&=\frac{\mu_0}{4\pi}\oint_V\frac{q\delta(\vec{R}-\vec{h}(t'))\vec{v}\,'\mathrm{d}V}{\|\vec{R}-\vec{r}\|}\end{aligned} \tag{10.3.378}$$

This is a very useful formulation (a workaround) that will enable us to solve our problem.

For this purpose, when the electric charge is at the point P_1 at the time t', we put:

$$\vec{h}\,'=\vec{h}(t')\qquad\vec{v}\,'=\vec{v}(t')\qquad\vec{a}\,'=\vec{a}(t') \tag{10.3.379}$$

We will use a looooong artifice to solve the integral of the electrical potential (which is therefore a multiple integral in Cartesian coordinates)!

This begins by multiplying the factor under the integrand $U(\vec{R},t)$ by:

$$\delta(t'-t+\|\vec{R}-\vec{r}\|c^{-1}) \tag{10.3.380}$$

this does not modify the integral since:

$$t'=t-\|\vec{R}-\vec{r}\|c^{-1}\rightarrow t'-t+\|\vec{R}-\vec{r}\|c^{-1}=0 \tag{10.3.381}$$

and that (see section Functional Analysis page 1569):

$$\delta(0)=1 \tag{10.3.382}$$

We then have the following expression in which the time t' appears:

$$\begin{aligned}U(\vec{R},t)&=\frac{1}{4\pi\varepsilon_0}\int\delta\left(t'-t+\frac{\|\vec{R}-\vec{r}(t')\|}{c}\right)\mathrm{d}t'\int\frac{q\delta(\vec{R}-\vec{h}(t'))}{\|\vec{R}-\vec{r}(t')\|}\\ &=\frac{1}{4\pi\varepsilon_0}\int\frac{q\delta(\vec{R}-\vec{h}(t'))}{\|\vec{R}-\vec{r}(t')\|}\delta\left(t'-t+\frac{\|\vec{R}-\vec{r}(t')\|}{c}\right)\mathrm{d}x\mathrm{d}y\mathrm{d}\mathrm{d}t'\\ &=\frac{1}{4\pi\varepsilon_0}\int\frac{q}{\|\vec{R}-\vec{r}(t')\|}\delta(\vec{R}-\vec{h}(t'))\delta\left(t'-t+\frac{\|\vec{R}-\vec{r}(t')\|}{c}\right)\mathrm{d}x\mathrm{d}y\mathrm{d}z\mathrm{d}t'\end{aligned} \tag{10.3.383}$$

what we have the right to write because the second integral does not depend explicitly on t'.

Good now if we try to solve this integral, we will spend our lives on it... for nothing. We'll have to be more clever!

Before looking for a solution of this integral, we must first deal with the more general case of the following integral:

$$I=\int g[f_1,f_2,f_3,f_4]\delta[f_1(x,y,z,t')]\delta[f_2(x,y,z,t')]\delta[f_3(x,y,z,t')]\delta[f_4(x,y,z,t')]\mathrm{d}x\mathrm{d}y\mathrm{d}z\mathrm{d}t' \tag{10.3.384}$$

Or written in a more condensed form:

$$I=\int g[f_1,f_2,f_3,f_4]\delta[f_1]\delta[f_2]\delta[f_3]\delta[f_4]\mathrm{d}x\mathrm{d}y\mathrm{d}z\mathrm{d}t' \tag{10.3.385}$$

which it is easy to approximate with the prior-previous integral:

$$\begin{aligned}f_1&=\|\vec{R}-\vec{r}(t')\|\\ g[f_1,f_2,f_3,f_4]&=\frac{1}{4\pi\varepsilon_0}\frac{q}{\|\vec{R}-\vec{r}(t')\|}=\frac{1}{4\pi\varepsilon_0}\frac{q}{f_1}\\ \delta[f_1]&=\delta(R_x-h_x(t'))\\ \delta[f_2]&=\delta(R_y-h_y(t'))\\ \delta[f_3]&=\delta(R_z-h_z(t'))\\ \delta[f_4]&=\delta\left(t'-t+\frac{\|\vec{R}-\vec{r}(t')\|}{c}\right)\end{aligned} \tag{10.3.386}$$

where we have deal such that f_1,f_2,f_3,f_4 depend respectively only (explicitly) on x, y, z and t'.

We now want now to make the following change of variables:

$$x\rightarrow f_1\qquad y\rightarrow f_2\qquad z\rightarrow f_3\qquad t'\rightarrow f_4 \tag{10.3.387}$$

Let us recall that in changes of variables in multiple integrals (see the Jacobian in the section of Differential and Integral Calculus), we have, passing from Cartesian coordinates to curvilinear coordinates, the following relations:

$$I=\int_D F(x,y,z)\mathrm{d}x\mathrm{d}y\mathrm{d}z=\int_{D'}G(r_1,r_2,r_3)\left|\frac{\partial(x,y,z)}{\partial(r_1,r_2,r_3)}\right|\mathrm{d}r_1\mathrm{d}r_2\mathrm{d}r_3 \tag{10.3.388}$$

where for recall:

$$r_1=r_1(x,y,z)\qquad r_2=r_2(x,y,z)\qquad r_3=r_3(x,y,z) \tag{10.3.389}$$

and where:

$$\left|\frac{\partial(x,y,z)}{\partial(r_1,r_2,r_3)}\right| \tag{10.3.390}$$

is not an absolute value but the determinant of a matrix!

Now, in our case, let us recall that we have all the f_i which are null and therefore:

$$\delta[f_i]_{f_i=0}=1 \tag{10.3.391}$$

and in the case where during the developments one of the f_i would no longer be zero for reasons not yet determined, we would have:

$$\delta[f_i]_{f_i\neq 0}=0 \tag{10.3.392}$$

The multiple integral then becomes:

$$I=\int g[f_1,f_2,f_3,f_4]\left\{\delta[f_1]\delta[f_2]\delta[f_3]\delta[f_4]\left|\frac{\partial(x,y,z,t')}{\partial(f_1,f_2,f_3,f_4)}\right|\right\}_{f_i=0}\mathrm{d}f_1\mathrm{d}f_2\mathrm{d}f_3\mathrm{d}f_4 \tag{10.3.393}$$

Where the term between braces is taken at $f_i=0$ by necessity of the construction of previous developments preparing the mathematical artifice!

And let us recall once more (!!) the property of Dirac functions:

$$\int f(x)\delta(x)\mathrm{d}x=[f(x)]_{x=0} \tag{10.3.394}$$

We then have immediately the simplification:

$$I=g[f_1,f_2,f_3,f_4]\left|\frac{\partial(x,y,z,t')}{\partial(f_1,f_2,f_3,f_4)}\right|_{f_i=0} \tag{10.3.395}$$

where:

$$\left|\frac{\partial(x,y,z,t')}{\partial(f_1,f_2,f_3,f_4)}\right| \tag{10.3.396}$$

is therefore the Jacobian of the transformation of the artifice ...

It is evident that by construction of the Jacobian we have:

$$\left|\frac{\partial(x,y,z)}{\partial(r_1,r_2,r_3)}\right|\left|\frac{\partial(r_1,r_2,r_3)}{\partial(x,y,z)}\right|=|\det(J)||\det(J)|^{-1}=1 \tag{10.3.397}$$

Therefore it comes:

$$\left|\frac{\partial(x,y,z)}{\partial(r_1,r_2,r_3)}\right|=\left|\frac{\partial(r_1,r_2,r_3)}{\partial(x,y,z)}\right|^{-1} \tag{10.3.398}$$

For the integral I we then have:

$$I=g[f_1,f_2,f_3,f_4]\left|\frac{\partial(f_1,f_2,f_3,f_4)}{\partial(x,y,z,t')}\right|^{-1}_{f_i=0} \tag{10.3.399}$$

Let us now calculate our Jacobian ...:

$$|\det(J)|^{-1}=\left|\det\begin{bmatrix}\partial_x f_1 & \partial_x f_2 & \partial_x f_3 & \partial_x f_4\\ \partial_y f_1 & \partial_y f_2 & \partial_y f_3 & \partial_y f_4\\ \partial_z f_1 & \partial_z f_2 & \partial_z f_3 & \partial_z f_4\\ \partial_{t'} f_1 & \partial_{t'} f_2 & \partial_{t'} f_3 & \partial_{t'} f_4\end{bmatrix}\right|^{-1} \tag{10.3.400}$$

Returning to the our treated case, $\vec{R}-\vec{h}(t')$ therefore has for components:

$$\begin{aligned}
&f_1=R_x-h_x(t')=x-h'_x\\
&f_2=R_y-h_y(t')=y-h'_y\\
&f_3=R_z-h_z(t')=z-h'_z\\
&f_4=t'-t+\frac{\|\vec{R}-\vec{r}(t')\|}{c}=t'-t+\frac{\sqrt{(X-x)^2+(Y-y)^2+(Z-z)^2}}{c}
\end{aligned} \tag{10.3.401}$$

Therefore, we have the calculation of the elements of the inverse of the Jacobian:

$$|\det(J)|^{-1}=\left|\det\begin{bmatrix}\partial_x f_1=1 & \partial_x f_2=0 & \partial_x f_3=0 & \partial_x f_4=-\dfrac{X-x}{c\|\vec{R}-\vec{r}(t')\|}\\ \partial_y f_1=0 & \partial_y f_2=1 & \partial_y f_3=0 & \partial_y f_4=-\dfrac{Y-y}{c\|\vec{R}-\vec{r}(t')\|}\\ \partial_z f_1=0 & \partial_z f_2=0 & \partial_z f_3=1 & \partial_z f_4=-\dfrac{Z-z}{c\|\vec{R}-\vec{r}(t')\|}\\ \partial_{t'} f_1=-\dot{h}'_x & \partial_{t'} f_2=-\dot{h}'_y & \partial_{t'} f_3=-\dot{h}'_z & \partial_{t'} f_4=1\end{bmatrix}\right|^{-1} \tag{10.3.402}$$

Well... now that we have the components of the Jacobian matrix, we have only to calculate its determinant. So either we use the general relation of the calculus of determinant proved in the section of Linear Algebra, or we use Maple™ ... So to win a little bit time let's do it with Maple™ 4.00b:

```
>with(linalg):
>A:=matrix(4,4,[1,0,0,a,0,1,0,b,0,0,1,c,d,e,f,1]);
```

where (for people having difficulty to read Maple™ notations):

$$A=\begin{bmatrix}1&0&0&a\\0&1&0&b\\0&0&1&c\\d&e&f&1\end{bmatrix} \tag{10.3.403}$$

with:

$$\begin{gathered}
a=-\frac{X-x}{c\|\vec{R}-\vec{r}(t')\|}\qquad b=-\frac{Y-y}{c\|\vec{R}-\vec{r}(t')\|}\qquad c=-\frac{Z-z}{c\|\vec{R}-\vec{r}(t')\|}\\
d=\dot{h}'_x\qquad e=\dot{h}'_y\qquad f=\dot{h}'_z
\end{gathered} \tag{10.3.404}$$

Let us continue with Maple™ 4.00b:

```
>Det(A);
```

Which gives as output:

```
1-cf-eb-da=1-(fc+eb+da)
```

The inverse of the Jacobian has then for expression:

$$\begin{aligned}|\det(J)|^{-1}&=\left|1-\left\{(-\dot{h}'_x\left[-\frac{X-x}{c\|\vec{R}-\vec{r}(t')}\right]+(-\dot{h}'_z\left[-\frac{Z-z}{c\|\vec{R}-\vec{r}(t')||}\right])\right\}\right|^{-1}\\&=\left|1-\left\{\frac{\dot{h}'_x(X-x)+\dot{h}'_y(Y-y)+\dot{h}'_z(Z-z)}{c\|\vec{R}-\vec{r}(t')||}\right\}\right|^{-1}=\left|1-\left\{\frac{\dot{h}'\circ(\vec{R}-\vec{r}(t'))}{c\|\vec{R}-\vec{r}(t')||}\right\}\right|^{-1}\\&=\left|1-\left\{\frac{\vec{v}\,'\circ(\vec{R}-\vec{r}(t'))}{c\|\vec{R}-\vec{r}(t')||}\right\}\right|^{-1}\end{aligned} \tag{10.3.405}$$

Where we used the dot product in the prior-previous equality in order to condense the expression.

Therefore:

$$|\det(J)|^{-1}=\frac{1}{\left|1-\dfrac{\vec{v}\,'}{c}\dfrac{(\vec{R}-\vec{r}(t'))}{\|\vec{R}-\vec{r}(t')||}\right|} \tag{10.3.406}$$

The multiple integral:

$$I=g[f_1,f_2,f_3,f_4]\left|\frac{\partial(f_1,f_2,f_3,f_4)}{\partial(x,y,z,t')}\right|^{-1}_{f_i=0} \tag{10.3.407}$$

where for recall:

$$g[f_1,f_2,f_3,f_4]=\frac{1}{4\pi\varepsilon_0}\frac{q}{\|\vec{R}-\vec{r}(t')||} \tag{10.3.408}$$

or written differently:

$$g[f_1,f_2,f_3,f_4]=\frac{1}{4\pi\varepsilon_0}\frac{q}{\left\|\begin{bmatrix}X\\Y\\Z\end{bmatrix}-\begin{bmatrix}x\\y\\z\end{bmatrix}\right\|} \tag{10.3.409}$$

But as a result of our change of coordinate system we have for recall:

$$g[f_1,f_2,f_3,f_4]\rightarrow g\left[(f_1=x-h'_x),(f_2=y-h'_y),(f_3=z-h'_z),\left(f_4=t'-t+\frac{\|\vec{R}-\vec{r}(t')||}{c}\right)\right] \tag{10.3.410}$$

And let us recall once again that:

$$\int f(x)\delta(x)\mathrm{d}x=[f(x)]_{x=0} \tag{10.3.411}$$

Therefore we have to take g on $f_1=f_2=f_3=f_4=0$! It comes:

$$g\left[(x=h'_x),(y=h'_y),(h=h'_z),\left(t'=t-\frac{\|\vec{R}-\vec{r}(t')||}{c}\right)\right] \tag{10.3.412}$$

Which allows us to write:

$$g[f_1,f_2,f_3,f_4]=\frac{1}{4\pi\varepsilon_0}\frac{q}{\left\|\begin{bmatrix}X\\Y\\Z\end{bmatrix}-\begin{bmatrix}h'_x\\h'_y\\h'_z\end{bmatrix}\right\|}=\frac{1}{4\pi\varepsilon_0}\frac{q}{\|\vec{R}-\vec{h}\,'\|} \tag{10.3.413}$$

We have to apply the same process for:

$$|\det(J)|^{-1}=\frac{1}{\left|1-\dfrac{\vec{v}\,'}{c}\circ\dfrac{\left(\vec{R}-\vec{r}(t')\right)}{\|\vec{R}-\vec{r}(t')\|}\right|} \tag{10.3.414}$$

that can then be written as:

$$|\det(J)|^{-1}=\frac{1}{\left|1-\dfrac{\vec{v}\,'}{c}\circ\dfrac{\left(\vec{R}-\vec{h}\right)}{\|\vec{R}-\vec{h}\|}\right|} \tag{10.3.415}$$

Finally, the resolution of the integral I is written:

$$\begin{aligned}I&=g[f_1,f_2,f_3,f_4]\left|\frac{\partial(f_1,f_2,f_3,f_4)}{\partial(x,y,z,t')}\right|^{-1}_{f_i=0}=g[f_1,f_2,f_3,f_4]\frac{1}{\left|1-\dfrac{\vec{v}\,'}{c}\circ\dfrac{\left(\vec{R}-\vec{h}\right)}{\|\vec{R}-\vec{h}\|}\right|}\\&=\frac{1}{4\pi\varepsilon_0}\frac{q}{\|\vec{R}-\vec{h}\,'\|}\frac{1}{\left|1-\dfrac{\vec{v}\,'}{c}\circ\dfrac{\left(\vec{R}-\vec{h}\right)}{\|\vec{R}-\vec{h}'\|}\right|}=\frac{1}{4\pi\varepsilon_0}\frac{1}{\|\vec{R}-\vec{h}\|\left|1-\dfrac{\vec{v}\,'}{c}\circ\dfrac{\left(\vec{R}-\vec{h}\right)}{\|\vec{R}-\vec{h}\|}\right|}\end{aligned} \tag{10.3.416}$$

Finally, we get the expressions of the potentials.

- The scalar potential is then written:

$$\boxed{U(\vec{R},t)=\frac{1}{4\pi\varepsilon_0}\frac{q}{\|\vec{R}-\vec{h}\|\left|1-\dfrac{\vec{v}\,'}{c}\circ\dfrac{\left(\vec{R}-\vec{h}\right)}{\|\vec{R}-\vec{h}\|}\right|}} \tag{10.3.417}$$

- The vector potential is written:

$$A(\vec{R},t)=\frac{\mu_0}{4\pi}\oint_V\frac{q\delta(\vec{R}-\vec{h}(t'))\vec{v}\,'\mathrm{d}V}{\|\vec{R}-\vec{r}(t')\|}=\frac{\mu_0}{4\pi}\oint\frac{q\delta(\vec{R}-\vec{h})\vec{v}\,'}{\|\vec{R}-\vec{r}(t')\|}\mathrm{d}x\mathrm{d}y\mathrm{d}z \tag{10.3.418}$$

  Considering that the integral which is almost the same as for the scalar potential except the term $\vec{v}\,'$, we arrive by making the same developments as before with to the expression (please... don't ask us to put also the details for this one... plzzzzz!):

$$\boxed{\vec{A}(\vec{R},t)=\frac{\mu_0}{4\pi}\frac{q\vec{v}\,'}{\|\vec{R}-\vec{h}\|\left|1-\dfrac{\vec{v}\,'}{c}\circ\dfrac{(\vec{R}-\vec{h})}{\|\vec{R}-\vec{h}\|}\right|}} \tag{10.3.419}$$

In summary, the potentials taken at the moment (time delay of propagation):

$$t'=t-\frac{\|\vec{R}-\vec{r}(t')\|}{c} \tag{10.3.420}$$

have for expression in a very condensed form (K_e in the Coulomb constant for recall as introduced in the section of Electrostatics!):

$$\begin{aligned}U(\vec{R},t)&=K_e\frac{q}{s}\\ \vec{A}(\vec{R},t)&=\frac{\mu_0}{4\pi}\frac{q\vec{v}\,'}{s}\end{aligned} \tag{10.3.421}$$

These potentials are named "Liénard-Wiechert retarded potentials" with:

$$s=\left(\|\vec{R}-\vec{h}\|\left|1-\frac{\vec{v}\,'}{c}\circ\frac{\left(\vec{R}-\vec{h}\right)}{\|\vec{R}-\vec{h}\|}\right|\right)_{t'=t-\dfrac{\|\vec{R}-\vec{r}(t')\|}{c}} \tag{10.3.422}$$

10.3.5.2 Retarded Electric and Magnetic fields

Let us recall that we have proved earlier above (see page 4116) that:

$$\vec{E}(\vec{R},t)=-\vec{\nabla}U-\partial_t\vec{A} \tag{10.3.423}$$

and by definition and the construction of potential vector we also have:

$$\vec{B}(\vec{R},t)=\mu_0\vec{H}\rightarrow\vec{H}(\vec{R},t)=\frac{1}{\mu_0}\vec{\nabla}\times\vec{A} \tag{10.3.424}$$

Therefore a difficulty appears in the sense that the Liénard-Wiechert potentials are at the moment $t'=t-\frac{\|\vec{R}-\vec{r}\|}{c}$ and that the fields are given at the moment t. It is therefore advisable to review the expressions of the fields corresponding to the instant t'.

By taking back the two relations:

$$\begin{aligned}U(\vec{R},t)&=\frac{1}{4\pi\varepsilon_0}\frac{q}{\|\vec{R}-\vec{h}\|\left|1-\dfrac{\vec{v}\,'}{c}\circ\dfrac{\left(\vec{R}-\vec{h}\right)}{\|\vec{R}-\vec{h}\|}\right|}=\frac{q}{4\pi\varepsilon_0}\frac{1}{s}\\ \vec{A}(\vec{R},t)&=\frac{\mu_0}{4\pi}\frac{q\vec{v}\,'}{\|\vec{R}-\vec{h}\|\left|1-\dfrac{\vec{v}\,'}{c}\circ\dfrac{(\vec{R}-\vec{h})}{\|\vec{R}-\vec{h}\|}\right|}=\frac{\mu_0 q}{4\pi}\frac{\vec{v}}{s}\end{aligned} \tag{10.3.425}$$

with for recall:

$$\begin{aligned}&s=z\left(1-\frac{\vec{v}}{c}\frac{\vec{z}}{z}\right)=z-\frac{\vec{v}\circ\vec{z}}{c}\\ &t'=t-\frac{z}{c}\\ &\vec{h}=\vec{r}(t')\\ &\vec{z}=\vec{R}-\vec{r}(t')\quad\text{and}\quad z=\|\vec{R}-\vec{r}(t')\|\end{aligned} \tag{10.3.426}$$

The notation in the last line may be confusing, but it's just a change of notation.

Let us now determine the retarded electric field. For this, let us first rearrange t' given above:

$$t'=t-\frac{z}{c}\Leftrightarrow z=c(t-t') \quad (10.3.427)$$

As $z=z(t')$ and $t=t(t')$ (in other words, z is function of t' and t is function of t'), we have the following partial derivative:

$$\partial_t z=\partial_{t'}z\partial_t t' \quad (10.3.428)$$

We also have $z^2=\|\vec{z}\|^2$. Deriving relatively to t':

$$z\partial_{t'}z=\vec{z}\circ\partial_{t'}\vec{z} \quad (10.3.429)$$

By definition we put (!):

$$\partial_{t'}\vec{z}=-\vec{v} \quad (10.3.430)$$

By isolating the derivative with respect to t' in $z\partial_{t'}z=\vec{z}\circ\partial_{t'}\vec{z}$ and taking into account from $\partial_{t'}\vec{z}=-\vec{v}$, we get:

$$\partial_{t'}z=\frac{\vec{z}}{z}\circ\partial_{t'}\vec{z}=-\frac{\vec{z}\circ\vec{v}}{z} \quad (10.3.431)$$

If we derive $z=c(t-t')$ relatively to t:

$$\partial_t z=c(1-\partial_t t')\Leftrightarrow\partial_t t'=1-\frac{\partial_t z}{c} \quad (10.3.432)$$

and equating:

$$\partial_{t'}z\partial_t t'=-\frac{\vec{z}\circ\vec{v}}{z}\partial_t t' \quad \text{and} \quad \partial_t z=c(1-\partial_t t') \quad (10.3.433)$$

We get:

$$-\frac{\vec{z}\circ\vec{v}}{z}\partial_t t'=c(1-\partial_t t') \quad (10.3.434)$$

Isolating $\partial_t t'$:

$$\partial_t t'=\frac{1}{\left(1-\frac{\vec{z}\circ\vec{v}}{zc}\right)}=\frac{z}{\left(z-\frac{\vec{z}\circ\vec{v}}{c}\right)}=\frac{z}{s} \quad (10.3.435)$$

We multiply both side by $\partial_{t'}$:

$$\frac{\partial}{\partial t'}\frac{\partial t'}{\partial t}=\frac{\partial}{\partial t}=\frac{z}{s}\frac{\partial}{\partial t'} \quad (10.3.436)$$

At constant t we apply the nabla operator (i.e. gradient... and for reminder that gives then a vector!) to z:

$$\vec{\nabla}(z)=\vec{\nabla}\left(c(t-t')\right)=-c\vec{\nabla}(t') \quad (10.3.437)$$

But we also have (denoting by $\vec{\nabla}'$ the nabla operator only on the spatial components and remembering as seen above that $\partial_t z = \partial_{t'} z \partial_t t'$):

$$\vec{\nabla}(z) = \vec{\nabla}'(z) + \partial_t z = \vec{\nabla}'(z) + \partial_{t'} z \partial_t t' = \vec{\nabla}'(z) + \partial_{t'} z \vec{\nabla}(t') \tag{10.3.438}$$

where the last equality holds because anyway t' depends only on t.

Therefore we can write:

$$\vec{\nabla}(z) = c\vec{\nabla}(t') = \vec{\nabla}'(z) + \partial_{t'} z \vec{\nabla}(t') \tag{10.3.439}$$

But from:

$$\partial_{t'} z = -\frac{\vec{z} \circ \vec{v}}{z} \tag{10.3.440}$$

We can then write:

$$\vec{\nabla}(z) = c\vec{\nabla}(t') = \vec{\nabla}'(z) - \frac{\vec{z} \circ \vec{v}}{z} \vec{\nabla}(t') \tag{10.3.441}$$

Remark

Now let us denote $u = \|\vec{R} - \vec{r}(t')\|$ instead of z and imagine we explicit that relation in the following way:

$$u = \sqrt{f^2(x(t)) + g^2(y(t)) + h^2(z(t))} = \sqrt{k(x,y,z)} \tag{10.3.442}$$

Now let us recall that if f and g are both differentiable and $F(x)$ is the composite function defined by $F(x) = u(k(x))$ then F is differentiable and F' is given by the product chain rule:

$$F'(x) = u'(k(x))k'(x) \tag{10.3.443}$$

Then:

$$\begin{aligned}\vec{\nabla}(u) &= \begin{pmatrix} \dfrac{\partial u}{\partial x} \\ \dfrac{\partial u}{\partial y} \\ \dfrac{\partial u}{\partial z} \end{pmatrix} = \begin{pmatrix} \dfrac{\partial u}{\partial x}\dfrac{\partial k}{\partial x} \\ \dfrac{\partial u}{\partial y}\dfrac{\partial k}{\partial y} \\ \dfrac{\partial u}{\partial z}\dfrac{\partial k}{\partial z} \end{pmatrix} = \begin{pmatrix} \dfrac{1}{2}\dfrac{1}{\sqrt{f^2(x(t)) + g^2(y(t)) + h^2(z(t))}}2f(x) \\ \dfrac{1}{2}\dfrac{1}{\sqrt{f^2(x(t)) + g^2(y(t)) + h^2(z(t))}}2g(x) \\ \dfrac{1}{2}\dfrac{1}{\sqrt{f^2(x(t)) + g^2(y(t)) + h^2(z(t))}}2h(x) \end{pmatrix} \\ &= \frac{1}{\sqrt{f^2(x(t)) + g^2(y(t)) + h^2(z(t))}} \begin{pmatrix} f(x) \\ g(x) \\ h(x) \end{pmatrix} = \frac{\vec{u}}{u}\end{aligned} \tag{10.3.444}$$

So going back to the notation we have in this section:

$$\vec{\nabla}'(z) = \frac{\vec{z}}{z} \tag{10.3.445}$$

Taking into account what we just detailed in the remark above. We can then also write:

$$\vec{\nabla}(z)=c\vec{\nabla}(t')=\vec{\nabla}'(z)-\frac{\vec{z}\circ\vec{v}}{z}\vec{\nabla}(t')=\frac{\vec{z}}{z}-\frac{\vec{z}\circ\vec{v}}{z}\vec{\nabla}(t') \tag{10.3.446}$$

Rearranging the following par:

$$c\vec{\nabla}(t')=\vec{\nabla}'(z)-\frac{\vec{z}\circ\vec{v}}{z}\vec{\nabla}(t')=\frac{\vec{z}}{z}-\frac{\vec{z}\circ\vec{v}}{z}\vec{\nabla}(t') \tag{10.3.447}$$

we get:

$$\vec{\nabla}(t')=-\frac{\vec{z}}{c\left(z-\frac{\vec{z}\circ\vec{v}}{c}\right)}=-\frac{\vec{z}}{sc} \tag{10.3.448}$$

Therefore the relation we have derived earlier:

$$\vec{\nabla}(z)=\vec{\nabla}'(z)+\partial_{t'}z\vec{\nabla}(t') \tag{10.3.449}$$

can be written:

$$\vec{\nabla}(z)=\vec{\nabla}'(z)+\partial_{t'}z\left(-\frac{\vec{z}}{sc}\right)=\vec{\nabla}'(z)-\frac{\vec{z}}{sc}\partial_{t'}z \tag{10.3.450}$$

If we get rid of the z this gives the following useful relation for further calculations (nabla operator expressed form a retarded nabla operator):

$$\vec{\nabla}=\vec{\nabla}'-\frac{\vec{z}}{sc}\frac{\partial}{\partial t'} \tag{10.3.451}$$

Now remember that we have:

$$\vec{E}(\vec{R},t)=-\vec{\nabla}U-\partial_t\vec{A}\qquad\text{with}\qquad U(\vec{R},t)=\frac{q}{4\pi\varepsilon_0}\frac{1}{s}\qquad\text{and}\qquad\vec{A}(\vec{R},t)=\frac{\mu_0 q}{4\pi}\frac{\vec{v}}{s} \tag{10.3.452}$$

Therefore:

$$\vec{E}(\vec{R},t)=-\vec{\nabla}\left(\frac{q}{4\pi\varepsilon_0}\frac{1}{s}\right)-\partial_t\left(\frac{\mu_0 q}{4\pi}\frac{\vec{v}}{s}\right) \tag{10.3.453}$$

Rearranging (and remembering that $\varepsilon\mu_0c^2=1$) gives:

$$\frac{4\pi\varepsilon_0}{q}\vec{E}=-\vec{\nabla}\left(\frac{1}{s}\right)-\frac{1}{c^2}\partial_t\frac{\vec{v}}{s} \tag{10.3.454}$$

Now we can use the result we get before about the nabla operator because we will compute:

$$\vec{\nabla}\left(\frac{1}{s}\right)=\vec{\nabla}'\frac{1}{s}-\frac{z}{sc}\frac{\partial}{\partial t'}\frac{1}{s}=-\frac{\vec{\nabla}'(s)}{s^2}-\frac{\vec{z}}{sc}\left(-\frac{1}{s^2}\right)\frac{\partial s}{\partial t'}=-\frac{\vec{\nabla}'(s)}{s^2}+\frac{\vec{z}}{s^3c}\frac{\partial s}{\partial t'} \tag{10.3.455}$$

and we can also use the relation derived earlier above:

$$\frac{\partial}{\partial t}=\frac{z}{s}\frac{\partial}{\partial t'} \tag{10.3.456}$$

To calculate:

$$\frac{\partial}{\partial t}\frac{\vec{v}}{s}=\frac{z}{s}\frac{\partial}{\partial t'}\left(\frac{\vec{v}}{s}\right)=\frac{z}{s}\frac{s\frac{\partial\vec{v}}{\partial t'}-\vec{v}\frac{\partial s}{\partial t'}}{s^2}=\frac{z}{s^2}\dot{\vec{v}}-\frac{z\vec{v}}{s^3}\frac{\partial s}{\partial t'} \tag{10.3.457}$$

Therefore:

$$\frac{4\pi\varepsilon_0}{q}\vec{E}=\frac{\vec{\nabla}'s}{s^2}-\frac{\vec{z}}{s^3c}\frac{\partial s}{\partial t'}-\frac{z}{s^2c^2}\dot{\vec{v}}+\frac{z\vec{v}}{s^3c^2}\frac{\partial s}{\partial t'} \tag{10.3.458}$$

Now let us open a parenthesis! Remember that by definition we have put earlier above $\partial_{t'}\vec{z}=-\vec{v}$ and that we have just seen that:

$$\vec{\nabla}(z)=\frac{\vec{z}}{z} \tag{10.3.459}$$

So now let us consider the following rectilinear case:

$$\partial_{t'}\vec{z}=\partial_{t'}\begin{pmatrix}v_xt'\\v_zt'\\v_zt'\end{pmatrix}=\begin{pmatrix}v_x\\v_y\\v_z\end{pmatrix}=-\vec{v} \tag{10.3.460}$$

Then:

$$\begin{aligned}\vec{\nabla}(z)=\frac{\vec{z}}{z}&=\frac{1}{\sqrt{(v_xt')^2+(v_yt')^2+(v_zt')^2}}\begin{pmatrix}v_xt'\\v_zt'\\v_zt'\end{pmatrix}=\frac{1}{t'\sqrt{v_x^2+v_y^2+v_z^2}}\begin{pmatrix}v_xt'\\v_zt'\\v_zt'\end{pmatrix}\\&=\frac{1}{\sqrt{v_x^2+v_y^2+v_z^2}}\begin{pmatrix}v_x\\v_z\\v_z\end{pmatrix}\end{aligned} \tag{10.3.461}$$

Therefore:

$$\begin{aligned}\vec{\nabla}(z)v=\frac{\vec{z}}{z}\partial_{t'}\vec{z}&=\frac{1}{\sqrt{v_x^2+v_y^2+v_z^2}}\begin{pmatrix}v_x\\v_z\\v_z\end{pmatrix}v=\frac{1}{\cancel{\sqrt{v_x^2+v_y^2+v_z^2+}}}\begin{pmatrix}v_x\\v_z\\v_z\end{pmatrix}\cancel{\sqrt{v_x^2+v_y^2+v_z^2}}\\&=-\vec{v}\end{aligned} \tag{10.3.462}$$

But is:

$$\vec{\nabla}(z)v=-\vec{v} \tag{10.3.463}$$

valid for any type of movement? The answer is no! Indeed, consider the famous example of the circular motion in a plane (typical of many particles accelerators):

$$\vec{z}=\begin{pmatrix}R\cos(\omega t')\\R\sin(\omega t')\\0\end{pmatrix} \tag{10.3.464}$$

Then obviously:

$$\begin{aligned}\vec{\nabla}(z)=\frac{\vec{z}}{z}&=\frac{1}{\sqrt{(R\cos(\omega t'))^2+(R\sin(\omega t'))^2}}\begin{pmatrix}R\cos(\omega t')\\R\sin(\omega t')\\0\end{pmatrix}\\&=\frac{1}{R}\begin{pmatrix}R\cos(\omega t')\\R\sin(\omega t')\\0\end{pmatrix}=\begin{pmatrix}\cos(\omega t')\\\sin(\omega t')\\0\end{pmatrix}\end{aligned} \tag{10.3.465}$$

And therefore:

$$\begin{aligned}\vec{\nabla}(z)v=\frac{\vec{z}}{z}v=\frac{\vec{z}}{z}||\partial_{t'}\vec{z}||=\begin{pmatrix}\cos(\omega t')\\ \sin(\omega t')\\ 0\end{pmatrix}||\partial_{t'}\vec{z}||=\begin{pmatrix}\cos(\omega t')\\ \sin(\omega t')\\ 0\end{pmatrix}\left\|\begin{pmatrix}-R\omega\sin(\omega t')\\ R\omega\cos(\omega t')\\ 0\end{pmatrix}\right\|\\ =\begin{pmatrix}\cos(\omega t')\\ \sin(\omega t')\\ 0\end{pmatrix}\sqrt{(-R\omega\sin(\omega t'))^2+(R\omega\cos(\omega t'))^2}=\begin{pmatrix}\cos(\omega t')\\ \sin(\omega t')\\ 0\end{pmatrix}R\omega\\ =\begin{pmatrix}R\omega\cos(\omega t')\\ R\omega\sin(\omega t')\\ 0\end{pmatrix}\neq-\vec{v}\end{aligned}\tag{10.3.466}$$

So that means we can use:

$$\frac{\partial\vec{z}}{\partial t'}=\vec{\nabla}(z)v=\frac{\vec{z}}{z}v=-\vec{v}\tag{10.3.467}$$

only as a locally linear approximation of the particle path when the movement is collinear to the speed! Therefore all relations that will follow are approximations!

So according to what we have just say above, we can write locally in the particle path that:

$$\vec{\nabla}'(s)=\vec{\nabla}\left(z-\frac{\vec{z}\circ\vec{v}}{c}\right)=\vec{\nabla}(z)-\vec{\nabla}\left(\frac{\vec{z}\circ\vec{v}}{c}\right)=\frac{\vec{z}}{z}-\frac{\vec{z}}{z}\frac{v}{c}=\frac{\vec{z}}{z}-\frac{\vec{v}}{c}\tag{10.3.468}$$

Injecting that result in the earlier relation:

$$\frac{4\pi\varepsilon_0}{q}\vec{E}=\frac{\vec{\nabla}'s}{s^2}-\frac{\vec{z}}{s^3c}\frac{\partial s}{\partial t'}-\frac{z}{s^2c^2}\dot{\vec{v}}+\frac{z\vec{v}}{s^3c^2}\frac{\partial s}{\partial t'}\tag{10.3.469}$$

we get:

$$\frac{4\pi\varepsilon_0}{q}\vec{E}=\frac{1}{s^2}\left(\frac{\vec{z}}{z}-\frac{\vec{v}}{c}\right)+\frac{1}{s^3c}\left(\frac{z}{c}\vec{v}-\vec{z}\right)\frac{\partial s}{\partial t'}-\frac{z}{s^2c^2}\dot{\vec{v}}\tag{10.3.470}$$

But we also have:

$$\frac{\partial s}{\partial t'}=\frac{\partial}{\partial t'}\left(z-\frac{\vec{z}\circ\vec{v}}{c}\right)=\frac{\partial z}{\partial t'}-\frac{\vec{z}\circ\dot{\vec{v}}}{c}-\frac{\vec{v}}{c}\circ\frac{\partial\vec{z}}{\partial t'}=\frac{\partial z}{\partial t'}-\frac{\vec{z}\circ\dot{\vec{v}}}{c}+\frac{v^2}{c}\tag{10.3.471}$$

Injecting the result in the previous relation a remembering that we have proved earlier that:

$$\partial_{t'}z=\frac{\vec{z}}{z}\circ\partial_{t'}\vec{z}=-\frac{\vec{z}\circ\vec{v}}{z}\tag{10.3.472}$$

we get:

$$\frac{4\pi\varepsilon_0}{q}\vec{E}=\frac{1}{s^2z}\left(\vec{z}-\frac{z\vec{v}}{c}\right)+\frac{1}{s^3c}\left(\frac{z}{c}\vec{v}-\vec{z}\right)\left(-\frac{\vec{z}\circ\vec{v}}{z}-\frac{\vec{z}\circ\dot{\vec{v}}}{c}+\frac{v^2}{c}\right)-\frac{z}{s^2c^2}\dot{\vec{v}}\tag{10.3.473}$$

Rearranging a bit:

$$\frac{4\pi\varepsilon_0}{q}\vec{E}=\frac{1}{s^2z}\left(\vec{z}-\frac{z\vec{v}}{c}\right)+\frac{1}{s^3c}\left(\vec{z}-\frac{z}{c}\vec{v}\right)\left(\frac{\vec{z}\circ\vec{v}}{z}-\frac{v^2}{c}+\frac{\vec{z}\circ\dot{\vec{v}}}{c}\right)-\frac{z}{s^2c^2}\dot{\vec{v}}\tag{10.3.474}$$

Let us now focus on that term:

$$\frac{1}{s^2z}\left(\vec{z}-\frac{z\vec{v}}{c}\right)+\frac{1}{s^3c}\left(\vec{z}-\frac{z}{c}\vec{v}\right)\left(\frac{\vec{z}\circ\vec{v}}{z}-\frac{v^2}{c}+\frac{\vec{z}\circ\dot{\vec{v}}}{c}\right)-\frac{z}{s^2c^2}\dot{\vec{v}} \tag{10.3.475}$$

Rearranging we get:

$$\frac{1}{s^3}\left(\vec{z}-\frac{z\vec{v}}{c}\right)\left(\frac{s}{z}-\frac{v^2}{c^2}\right)+\frac{1}{s^3c}\left[\left(\vec{z}-\frac{z}{c}\vec{v}\right)\frac{\vec{z}\circ\vec{v}}{r}+\left(\vec{z}-\frac{z}{c}\vec{v}\right)\frac{\vec{z}\circ\dot{\vec{v}}}{c}\right]-\frac{z}{s^2c^2}\dot{\vec{v}} \tag{10.3.476}$$

Working on the first term we have:

$$\begin{aligned}\frac{1}{s^3}\left(\vec{z}-\frac{z\vec{v}}{c}\right)\left(\frac{s}{z}-\frac{v^2}{c^2}\right)&=\frac{1}{s^3}\left(\vec{z}-\frac{z\cdot\vec{v}}{c}\right)\left(\frac{z-\frac{\vec{z}\circ\vec{v}}{c}}{z}-\frac{v^2}{c^2}\right)\\&=\frac{1}{s^3}\left(\vec{z}-\frac{z\vec{v}}{c}\right)\left(1-\frac{\vec{z}\circ\vec{v}}{zc}-\frac{v^2}{c^2}\right)\\&=\frac{1}{s^3}\left(\vec{z}-\frac{z\vec{v}}{c}\right)\left(1-\frac{v^2}{c^2}\right)-\frac{1}{s^3}\left(\vec{z}-\frac{z\cdot\vec{v}}{c}\right)\frac{\vec{z}\circ\vec{v}}{zc}\end{aligned} \tag{10.3.477}$$

Injected into the previous relation and simplifying we get:

$$\begin{aligned}&\frac{1}{s^3}\left(\vec{z}-\frac{z\vec{v}}{c}\right)\left(1-\frac{v^2}{c^2}\right)-\frac{1}{s^3c}\left[-\cancel{\left(\vec{z}-\frac{z}{c}\vec{v}\right)\frac{\vec{z}\circ\vec{v}}{r}}+\cancel{\left(\vec{z}-\frac{z}{c}\vec{v}\right)\frac{\vec{z}\circ\vec{v}}{r}}-\left(\vec{z}-\frac{z}{c}\vec{v}\right)\frac{\vec{z}\circ\dot{\vec{v}}}{c}\right]-\frac{z}{s^2c^2}\dot{\vec{v}}\\&=\frac{1}{s^3}\left(\vec{z}-\frac{z\vec{v}}{c}\right)\left(1-\frac{v^2}{c^2}\right)+\frac{1}{s^3c}\left(\vec{z}-\frac{z}{c}\dot{\vec{v}}\right)\frac{\vec{z}\circ\dot{\vec{v}}}{c}-\frac{z}{s^2c^2}\dot{\vec{v}}\\&=\frac{1}{s^3}\left(\vec{z}-\frac{z\vec{v}}{c}\right)\left(1-\frac{v^2}{c^2}\right)+\frac{1}{s^3c}\left[\left(\vec{z}-\frac{z}{c}\vec{v}\right)\frac{\vec{z}\circ\dot{\vec{v}}}{c}-sz\dot{\vec{v}}\right]\\&=\frac{1}{s^3}\left(\vec{z}-\frac{z\vec{v}}{c}\right)\left(1-\frac{v^2}{c^2}\right)+\frac{1}{s^3c^2}\left[\left(\vec{z}-\frac{z}{c}\vec{v}\right)\vec{z}\circ\dot{\vec{v}}-z\left(z-\frac{\vec{z}\circ\vec{v}}{c}\right)\dot{\vec{v}}\right]\end{aligned} \tag{10.3.478}$$

The term between square bracket is a triple cross product and we have already proved in the section of Vector Calculus that this can be written (see page 1835):

$$\begin{aligned}&\left[\left(\vec{z}-\frac{z}{c}\vec{v}\right)\vec{z}\circ\dot{\vec{v}}-\dot{\vec{v}}\left(z-\frac{\vec{z}\circ\vec{v}}{c}\right)z\right]=\\&\vec{z}\times\left[\left(\vec{z}-\frac{z}{c}\vec{v}\right)\times\dot{\vec{v}}\right]=\left(\vec{z}-\frac{z}{c}\vec{v}\right)\left(\vec{z}\circ\dot{\vec{v}}\right)-\dot{\vec{v}}\left[\vec{z}\circ\left(\vec{z}-\frac{z\vec{v}}{c}\right)\right]\\&=\left(\vec{z}-\frac{z}{c}\vec{v}\right)\left(\vec{z}\circ\dot{\vec{v}}\right)-\dot{\vec{v}}\left[z\left(z-\frac{\vec{z}\circ\vec{v}}{c}\right)\right]\end{aligned} \tag{10.3.479}$$

Finally, we find for the retarded electric field, the following expression (we put now the approximate equal sign because we know that some assumptions made earlier make is an approximation):

$$\vec{E}\cong\frac{q}{4\pi\varepsilon_0 s^3}\left\{\left(1-\frac{v^2}{c^2}\right)\left(\vec{z}-\frac{z}{c}\vec{v}\right)+\frac{\vec{z}}{c^2}\times\left[\left(\vec{z}-\frac{z}{c}\vec{v}\right)\times\dot{\vec{v}}\right]\right\} \tag{10.3.480}$$

Now let us do the same work with the magnetic retarded field. We start from:

$$\vec{B}=\vec{\nabla}\times\vec{A} \quad \text{with} \quad \vec{A}=\frac{q\mu_0}{4\pi}\left(\frac{\vec{v}}{s}\right) \tag{10.3.481}$$

Therefore:

$$\vec{B}=\vec{\nabla}\times\left(\frac{q\mu_0}{4\pi}\frac{\vec{v}}{s}\right) \tag{10.3.482}$$

Rearranging and using $\mu_0\varepsilon_0c^2=1$ we get:

$$\frac{4\pi\varepsilon_0c^2}{q}\vec{B}=\vec{\nabla}\times\frac{\vec{v}}{s} \tag{10.3.483}$$

We have proved in the section of Vector Calculus that (see page 1916):

$$\vec{\nabla}\times\left(\frac{\vec{v}}{s}\right)=\frac{1}{s}(\vec{\nabla}\times\vec{v})+\vec{\nabla}\left(\frac{1}{s}\right)\times\vec{v} \tag{10.3.484}$$

Therefore:

$$\frac{4\pi\varepsilon_0c^2}{q}\vec{B}=\frac{1}{s}(\vec{\nabla}\times\vec{v})+\vec{\nabla}\left(\frac{1}{s}\right)\times\vec{v} \tag{10.3.485}$$

For the next step Larmor used a trick (still debated nowadays) to simplify the expression to be able to fall back later on a simplified expression of radiated electromagnetic field that was known before his theory. The trick (mathematically quite wrong but physically it works...) is to expression the nabla operator on the $\vec{\nabla}\times\vec{v}$ by it's retarded version derived earlier above for the position (...) and given for reminder by:

$$\vec{\nabla}=\vec{\nabla}'-\frac{\vec{z}}{sc}\frac{\partial}{\partial t'} \tag{10.3.486}$$

If we do that we get for the first parenthesis or the prior relation:

$$\vec{\nabla}\times\vec{v}_t=\vec{\nabla}'\times\vec{v}-\frac{\vec{z}}{sc}\times\frac{\partial\vec{v}}{\partial t'} \tag{10.3.487}$$

After without mathematical reasons or also without physical interpretation, but very likely to fall back on a known result as mentioned above, Larmor put:

$$\vec{\nabla}'\times\vec{v}=\vec{0} \tag{10.3.488}$$

claiming (we don't how this claim makes any sense...) that's because $\vec{v}=f(t')$....???

We may think that Larmor did that because if $\vec{v}$ is a vector field, then $\vec{\nabla}'\times\vec{v}=\vec{0}$ assume that the velocity field is irrotational and therefore the velocity field is conservative (such fields have for reminder the property that the line integral around any closed loop, often representing the work done in moving a particle, is zero). But if $\vec{v}$ is a conservative field, in our point of view, then we should also have $\vec{\nabla}\times\vec{v}=\vec{0}$... and then nothing makes sense anymore. That's why it is very likely to be a physicist trick just because we have to it to fall back later on a result that was already known at his time.

Then obviously we get for the first parenthesis:

$$\vec{\nabla}\times\vec{v}_t\cong-\frac{\vec{z}}{sc}\times\frac{\partial\vec{v}}{\partial t'}=-\frac{\vec{z}\times\dot{\vec{v}}}{sc} \qquad (10.3.489)$$

Now coming back to:

$$\frac{4\pi\varepsilon_0c^2}{q}\vec{B}=\vec{\nabla}\times\frac{\vec{v}}{s}=\frac{1}{s}(\vec{\nabla}\times\vec{v})+\vec{\nabla}\left(\frac{1}{s}\right)\times\vec{v}=-\frac{\vec{z}\times\dot{\vec{v}}}{s^2c}+\vec{\nabla}\left(\frac{1}{s}\right)\times\vec{v} \qquad (10.3.490)$$

Now let us focus on the second term that we can immediately write as (using again the relation between the nabla operator and retarded nabla operator):

$$\vec{\nabla}\left(\frac{1}{s}\right)\times\vec{v}=-\frac{1}{s^2}\vec{\nabla}(s)\times\vec{v}=-\frac{1}{s^2}\left[\vec{\nabla}'(s)-\frac{\vec{z}}{sc}\frac{\partial s}{\partial t'}\right]\times\vec{v} \qquad (10.3.491)$$

And using again the relation derived earlier above that was:

$$\vec{\nabla}'(s)=\vec{\nabla}\left(z-\frac{\vec{z}\circ\vec{v}}{c}\right)=\vec{\nabla}(z)-\vec{\nabla}\left(\frac{\vec{z}\circ\vec{v}}{c}\right)=\frac{\vec{z}}{z}-\frac{\vec{z}}{z}\frac{v}{c}=\frac{\vec{z}}{z}-\frac{\vec{v}}{c} \qquad (10.3.492)$$

We get:

$$\vec{\nabla}\left(\frac{1}{s}\right)\times\vec{v}=-\frac{1}{s^2}\vec{\nabla}(s)\times\vec{v}=-\frac{1}{s^2}\left[\frac{\vec{z}}{z}-\frac{\vec{v}}{c}-\frac{\vec{z}}{sc}\frac{\partial s}{\partial t'}\right]\times\vec{v} \qquad (10.3.493)$$

But we have proved earlier above that:

$$\frac{\partial s}{\partial t'}=\frac{\partial}{\partial t'}\left(z-\frac{\vec{z}\circ\vec{v}}{c}\right)=\frac{\partial z}{\partial t'}-\frac{\vec{z}\circ\dot{\vec{v}}}{c}-\frac{\vec{v}}{c}\circ\frac{\partial\vec{z}}{\partial t'}=-\frac{\vec{z}\circ\vec{v}}{z}-\frac{\vec{z}\circ\dot{\vec{v}}}{c}+\frac{v^2}{c} \qquad (10.3.494)$$

Therefore:

$$\vec{\nabla}\left(\frac{1}{s}\right)\times\vec{v}=-\frac{1}{s^2}\vec{\nabla}(s)\times\vec{v}=-\frac{1}{s^2}\left[\frac{\vec{z}}{z}-\frac{\vec{v}}{c}-\frac{\vec{z}}{sc}\left(-\frac{\vec{z}\circ\vec{v}}{z}-\frac{\vec{z}\circ\dot{\vec{v}}}{c}+\frac{v^2}{c}\right)\right]\times\vec{v} \qquad (10.3.495)$$

Hence:

$$\begin{aligned}\frac{4\pi\varepsilon_0c^2}{q}\vec{B}&=\vec{\nabla}\times\frac{\vec{v}}{s}=\frac{1}{s}(\vec{\nabla}\times\vec{v})+\vec{\nabla}\left(\frac{1}{s}\right)\times\vec{v}=-\frac{\vec{z}\times\dot{\vec{v}}}{s^2c}+\vec{\nabla}\left(\frac{1}{s}\right)\times\vec{v}\\&=-\frac{\vec{z}\times\dot{\vec{v}}}{s^2c}-\frac{1}{s^2}\left[\frac{\vec{z}}{z}-\frac{\vec{v}}{c}-\frac{\vec{z}}{sc}\left(-\frac{\vec{z}\circ\vec{v}}{z}-\frac{\vec{z}\circ\dot{\vec{v}}}{c}+\frac{v^2}{c}\right)\right]\times\vec{v}\end{aligned} \qquad (10.3.496)$$

Let us rearrange and simplify the right hand side:

$$
\begin{aligned}
\vec{\nabla}\times\frac{\vec{v}}{s} &= -\frac{\left(\vec{z}\times\dot{\vec{v}}\right)}{s^2c}+\frac{\vec{v}}{s^2}\times\left\{\frac{\vec{z}}{z}-\frac{\vec{v}}{c}-\frac{\vec{z}}{sc}\left[-\frac{\left(\vec{z}\circ\vec{v}\right)}{z}+\frac{v^2}{c}-\frac{\left(\vec{z}\circ\dot{\vec{v}}\right)}{c}\right]\right\}\\
&= -\frac{\left(\vec{z}\times\dot{\vec{v}}\right)}{s^2c}+\frac{\left(\vec{v}\times\vec{z}\right)}{s^2z}-\cancel{\frac{\left(\vec{v}\times\vec{v}\right)}{s^2c}}-\frac{\left(\vec{v}\times\vec{z}\right)}{s^3c}\left[-\frac{\left(\vec{z}\circ\vec{v}\right)}{z}+\frac{v^2}{c}-\frac{\left(\vec{z}\circ\dot{\vec{v}}\right)}{c}\right]\\
&= -\frac{\left(\vec{z}\times\dot{\vec{v}}\right)}{s^2c}+\frac{\left(\vec{v}\times\vec{z}\right)}{s^2z}+\frac{\left(\vec{v}\times\vec{z}\right)}{s^3c}\cdot\frac{\left(\vec{z}\circ\vec{v}\right)}{z}-\frac{\left(\vec{v}\times\vec{z}\right)}{s^3c}\cdot\frac{v^2}{c}+\frac{\left(\vec{v}\times\vec{z}\right)}{s^3c}\cdot\frac{\left(\vec{z}\circ\dot{\vec{v}}\right)}{c}\\
&= \frac{1}{s^3}\left[-\left(\vec{z}\times\dot{\vec{v}}\right)\left(\frac{s}{c}\right)+\left(\vec{v}\times\vec{z}\right)\left(\frac{s}{z}\right)+\left(\vec{v}\times\vec{z}\right)\frac{\vec{z}\circ\vec{v}}{zc}-\left(\vec{v}\times\vec{z}\right)\left(\frac{v^2}{c^2}\right)+\left(\vec{v}\times\vec{z}\right)\frac{\left(\vec{z}\circ\dot{\vec{v}}\right)}{c^2}\right]\\
&= \frac{1}{s^3}\left[-\left(\vec{z}\times\dot{\vec{v}}\right)\frac{z-\frac{\left(\vec{z}\circ\vec{v}\right)}{c}}{c}+\left(\vec{v}\times\vec{z}\right)\frac{z-\frac{\left(\vec{z}\circ\vec{v}\right)}{c}}{z}+\left(\vec{v}\times\vec{z}\right)\frac{\left(\vec{z}\circ\vec{v}\right)}{zc}-\left(\vec{v}\times\vec{z}\right)\left(\frac{v^2}{c^2}\right)+\left(\vec{v}\times\vec{z}\right)\frac{\left(\vec{z}\circ\dot{\vec{v}}\right)}{c^2}\right]\\
&= \frac{1}{s^3}\left[-\left(\vec{z}\times\dot{\vec{v}}\right)\frac{z}{c}+\left(\vec{z}\times\dot{\vec{v}}\right)\frac{\left(\vec{z}\circ\vec{v}\right)}{c^2}+\left(\vec{v}\times\vec{z}\right)-\cancel{\left(\vec{v}\times\vec{z}\right)\frac{\left(\vec{z}\circ\vec{v}\right)}{zc}}+\cancel{\left(\vec{v}\times\vec{z}\right)\frac{\left(\vec{z}\circ\vec{v}\right)}{zc}}-\left(\vec{v}\times\vec{z}\right)\left(\frac{v^2}{c^2}\right)+\left(\vec{v}\times\vec{z}\right)\frac{\left(\vec{z}\circ\dot{\vec{v}}\right)}{c^2}\right]\\
&= \frac{1}{s^3}\left[-\left(\vec{z}\times\dot{\vec{v}}\right)\frac{z}{c}+\left(\vec{z}\times\dot{\vec{v}}\right)\frac{\left(\vec{z}circ\vec{v}\right)}{c^2}+\left(\vec{v}\times\vec{z}\right)\left(1-\frac{v^2}{c^2}\right)+\left(\vec{v}\times\vec{z}\right)\frac{\left(\vec{z}\circ\dot{\vec{v}}\right)}{c^2}\right]\\
&= \frac{1}{s^3}\left[\left(1-\frac{v^2}{c^2}\right)\left(\vec{v}\times\vec{z}\right)+\left(\vec{v}\times\vec{z}\right)\frac{\left(\vec{z}\circ\dot{\vec{v}}\right)}{c^2}-\left(\vec{z}\times\dot{\vec{v}}\right)\frac{z-\frac{\left(\vec{z}\circ\vec{v}\right)}{c}}{c}\right]
\end{aligned}
\tag{10.3.497}
$$

Therefore:

$$
\frac{4\pi\varepsilon_0c^2}{q}\vec{B}=\vec{\nabla}\times\frac{\vec{v}}{s}=\frac{1}{s^3}\left[\left(1-\frac{v^2}{c^2}\right)\left(\vec{v}\times\vec{z}\right)+\left(\vec{v}\times\vec{z}\right)\frac{\left(\vec{z}\circ\dot{\vec{v}}\right)}{c^2}-\left(\vec{z}\times\dot{\vec{v}}\right)\frac{z-\frac{\left(\vec{z}\circ\vec{v}\right)}{c}}{c}\right]
\tag{10.3.498}
$$

We now propose to investigate the possibility of bringing together the last two terms inside the square brackets in the form of cross-products. That means:

$$
\left(\vec{v}\times\vec{z}\right)\frac{\left(\vec{z}\circ\dot{\vec{v}}\right)}{c^2}-\left(\vec{z}\times\dot{\vec{v}}\right)\frac{z-\frac{\left(\vec{z}\circ\vec{v}\right)}{c}}{c}
\tag{10.3.499}
$$

By inverting the factors of the first cross product and simplifying by reducing to the same denominator in the second term we get:

$$
-\frac{1}{c^2}\left(\vec{z}\times\vec{v}\right)\left(\vec{z}\cdot\dot{\vec{v}}\right)-\left(\vec{z}\times\dot{\vec{v}}\right)\left[\frac{z}{c}-\frac{\left(\vec{z}\circ\vec{v}\right)}{c^2}\right]
\tag{10.3.500}
$$

We multiply this expression by z/z and factorize $1/(zc)$:

$$
\frac{1}{zc}\left\{\left[-\frac{z}{c}\left(\vec{z}\times\vec{v}\right)\left(\vec{z}\circ\dot{\vec{v}}\right)\right]-\left(\vec{z}\times\dot{\vec{v}}\right)\left[z^2-\frac{z}{c}\left(\vec{z}\circ\vec{v}\right)\right]\right\}
\tag{10.3.501}
$$

We introduce the term $\vec{z}\times\vec{z}$ (that is obviously equal to $\vec{0}$):

$$
\begin{aligned}
&\frac{1}{zc}\left\{\left[\vec{z}\times\vec{z}-\frac{z}{c}\left(\vec{z}\times\vec{v}\right)\left(\vec{z}\circ\dot{\vec{v}}\right)\right]-\left(\vec{z}\times\dot{\vec{v}}\right)\left[z^2-\frac{z}{c}\left(\vec{z}\circ\vec{v}\right)\right]\right\}=\\
&\frac{1}{zc}\left\{\vec{z}\times\left[\left(\vec{z}-\frac{z}{c}\vec{v}\right)\left(\vec{z}\circ\dot{\vec{v}}\right)\right]-\left(\vec{z}\times\dot{\vec{v}}\right)\left[\vec{z}\left(\vec{z}-\frac{z}{c}\vec{v}\right)\right]\right\}=\\
&\frac{\vec{z}}{zc}\times\left\{\left[\left(\vec{z}-\frac{z}{c}\vec{v}\right)\left(\vec{z}\circ\dot{\vec{v}}\right)\right]-\dot{\vec{v}}\left[\vec{z}\circ\left(\vec{z}-\frac{z}{c}\vec{v}\right)\right]\right\}
\end{aligned}
\tag{10.3.502}
$$

But the term between curl braces:

$$\left[\left(\vec{z}-\frac{z}{c}\vec{v}\right)(\vec{z}\circ\dot{\vec{v}})\right]-\dot{\vec{v}}\left[\vec{z}\circ\left(\vec{z}-\frac{z}{c}\vec{v}\right)\right] \tag{10.3.503}$$

can be recognized as being a triple cross-product (see section Vector Calculus page 1835). Indeed, if we put:

$$\vec{A}=\vec{z} \qquad \vec{B}=\left(\vec{z}-\frac{z}{c}\vec{v}\right) \qquad \vec{C}=\dot{\vec{v}} \tag{10.3.504}$$

We know that then we have the following relation:

$$\vec{A}\times(\vec{B}\times\vec{C})=\vec{B}(\vec{A}\circ\vec{C})-\vec{C}(\vec{A}\circ\vec{B}) \tag{10.3.505}$$

and it's corresponding:

$$\left(\vec{z}-\frac{z}{c}\vec{v}\right)(\vec{z}\circ\dot{\vec{v}})-\vec{v}\left[\vec{z}\circ\left(\vec{z}-\frac{z}{c}\vec{v}\right)\right]=\vec{z}\times\left(\vec{z}-\frac{z}{c}\vec{v}\right)\times\dot{\vec{v}} \tag{10.3.506}$$

Therefore:

$$\frac{4\pi\varepsilon_0 c^2}{q}\vec{B}\cong\vec{\nabla}\times\frac{\vec{v}}{s}=\frac{1}{s^3}\left\{\left(1-\frac{v^2}{c^2}\right)(\vec{v}\times\vec{z})+\frac{\vec{z}}{zc}\times\left[\vec{z}\times\left(\vec{z}-\frac{z}{c}\vec{v}\right)\times\dot{\vec{v}}\right]\right\} \tag{10.3.507}$$

After rearranging:

$$\vec{B}\cong\frac{q}{4\pi\varepsilon_0 c^2 s^3}\left[\left(1-\frac{v^2}{c^2}\right)(\vec{z}\times\vec{v})+\frac{\vec{z}}{zc}\times\left[\vec{z}\times\left(\vec{z}-\frac{z}{c}\vec{v}\right)\times\dot{\vec{v}}\right]\right] \tag{10.3.508}$$

In summary we find for the delayed fields, the following expressions:

$$\begin{gathered}\vec{E}\cong\frac{q}{4\pi\varepsilon_0 s^3}\left\{\left(1-\frac{v^2}{c^2}\right)\left(\vec{z}-\frac{z}{c}\vec{v}\right)+\frac{\vec{z}}{c^2}\times\left[\left(\vec{z}-\frac{z}{c}\vec{v}\right)\times\dot{\vec{v}}\right]\right\}\\ \vec{B}\cong\frac{q}{4\pi\varepsilon_0 c^2 s^3}\left[\left(1-\frac{v^2}{c^2}\right)(\vec{v}\times\vec{z})+\frac{\vec{z}}{zc}\times\left[\vec{z}\times\left(\vec{z}-\frac{z}{c}\vec{v}\right)\times\dot{\vec{v}}\right]\right]\end{gathered} \tag{10.3.509}$$

Hence explicitly:

$$\boxed{\begin{gathered}\vec{E}\cong\frac{q}{4\pi\varepsilon_0\left(z\left(1-\dfrac{\vec{v}}{c}\circ\dfrac{\vec{z}}{z}\right)\right)^3}\left\{\left(1-\frac{v^2}{c^2}\right)\left(\vec{z}-\frac{z}{c}\vec{v}\right)+\frac{\vec{z}}{c^2}\times\left[\left(\vec{z}-\frac{z}{c}\vec{v}\right)\times\dot{\vec{v}}\right]\right\}\\ \vec{B}\cong\frac{q}{4\pi\varepsilon_0 c^2\left(z\left(1-\dfrac{\vec{v}}{c}\circ\dfrac{\vec{z}}{z}\right)\right)^3}\left[\left(1-\frac{v^2}{c^2}\right)(\vec{v}\times\vec{z})+\frac{\vec{z}}{zc}\times\left[\vec{z}\times\left(\vec{z}-\frac{z}{c}\vec{v}\right)\times\dot{\vec{v}}\right]\right]\end{gathered}} \tag{10.3.510}$$

Let us analyse these expressions of retarded electric fields and magnetic induction. For this analysis, we will modify somewhat the two expressions above.

For this purpose we put:

$$\frac{\vec{v}}{c}=\beta \qquad s=z-\frac{\vec{v}}{c}\circ\vec{z}=z-\vec{\beta}\circ\vec{z} \tag{10.3.511}$$

and we will denote by $\vec{u}_z$ the unitary vector collinear to $\vec{z}$.

Therefore:

$$\begin{aligned}\vec{E} &\cong \frac{q}{4\pi\varepsilon_0}\left[\frac{z(1-\beta^2)\left(\vec{u}_z-\vec{\beta}\right)}{z^3\left(1-\vec{\beta}\circ\vec{u}_z\right)^3}+\frac{z\cdot\vec{u}_z\times z\left(\vec{u}_z-\vec{\beta}\right)\times\dot{\vec{v}}}{c^2z^3\left(1-\vec{\beta}\circ\vec{u}_z\right)^3}\right]\\ \vec{B} &\cong \frac{q}{4\pi\varepsilon_0c^2}\left[\frac{z(1-\beta^2)\left(\dot{\vec{v}}\times\vec{u}_z\right)}{z^3\left(1-\vec{\beta}\circ\vec{u}_z\right)^3}+\frac{z\cdot\vec{u}_z\times[z\cdot\vec{u}_z\times z]\left(\vec{u}_z-\vec{\beta}\right)\times\dot{\vec{\beta}}}{z^4\left(1-\vec{\beta}\circ\vec{u}_z\right)^3}\right]\end{aligned} \tag{10.3.512}$$

Simplifying a bit more:

$$\begin{aligned}\vec{E} &\cong \frac{q}{4\pi\varepsilon_0}\left[\frac{(1-\beta^2)\left(\vec{u}_z-\vec{\beta}\right)}{z^2\left(1-\vec{\beta}\circ\vec{u}_z\right)^3}+\frac{\vec{u}_z\times\left(\vec{u}_z-\vec{\beta}\right)\times\dot{\vec{v}}}{c^2z\left(1-\vec{\beta}\circ\vec{u}_z\right)^3}\right]\\ \vec{B} &\cong \frac{q}{4\pi\varepsilon_0c^2}\left[\frac{(1-\beta^2)\left(\vec{v}\times\vec{u}_z\right)}{z^2\left(1-\vec{\beta}\circ\vec{u}_z\right)^3}+\frac{\vec{u}_z\times\left[\vec{u}_z\times\left(\vec{u}_z-\vec{\beta}\right)\right]\times\dot{\vec{\beta}}}{z\left(1-\vec{\beta}\circ\vec{u}_z\right)^3}\right]\end{aligned} \tag{10.3.513}$$

So we observe that the two retarded fields $\vec{E}$ and $\vec{B}$ are composed of two terms $\vec{E}_1+\vec{E}_2$ and $\vec{B}_1+\vec{B}_2$.

- The first terms $\vec{E}_1$ and $\vec{B}_1$ are:

  $$\begin{aligned}\vec{E}_1 &\cong \frac{q}{4\pi\varepsilon_0}\frac{(1-\beta^2)\left(\vec{u}_z-\vec{\beta}\right)}{z^2\left(1-\vec{\beta}\circ\vec{u}_z\right)^3}\\ \vec{B}_1 &\cong \frac{q}{4\pi\varepsilon_0c^2}\frac{(1-\beta^2)\left(\vec{v}\times\vec{u}_z\right)}{z^2\left(1-\vec{\beta}\circ\vec{u}_z\right)^3}\end{aligned} \tag{10.3.514}$$

  They correspond to a charge animated by a speed $\vec{v}$ and whose fields are functions of the speed of displacement of the so said charge and of the inverse of the square of the distance z.

  $$\vec{E}_1\cong f(\vec{v},z^{-2}) \qquad \vec{B}_1\cong f(\vec{v},z^{-2}) \tag{10.3.515}$$

  These fields decrease rapidly with the distance z.

  We notice that in the non-relativistic case $\beta\ll 1$, we fall back on the classical expression of the electric field and of magnetic induction:

  $$\begin{aligned}\vec{E}_1 &\cong \frac{q}{4\pi\varepsilon_0}\frac{\vec{u}_z}{z^2}\\ \vec{B}_1 &\cong \frac{q}{4\pi\varepsilon_0c^2}\frac{\vec{v}\times\vec{z}}{z^2}\end{aligned} \tag{10.3.516}$$

  Likewise, in the particular case where the speed is zero, only the electrostatic field remains as expected!

- The second terms $\vec{E}_2$ and $\vec{B}_2$ are:

$$\begin{aligned}\vec{E}_2 &\cong \frac{q}{4\pi\varepsilon_0 c}\frac{\vec{u}_z\times\left(\vec{u}_z-\vec{\beta}\right)\times\dot{\vec{\beta}}}{z\left(1-\vec{\beta}\circ\vec{u}_z\right)^3}\\ \vec{B}_2 &\cong \frac{q}{4\pi\varepsilon_0 c^2}\frac{\vec{u}_z\times\left[\vec{u}_z\times\left(\vec{u}_z-\vec{\beta}\right)\right]\times\dot{\vec{\beta}}}{z\left(1-\vec{\beta}\circ\vec{u}_z\right)^3}=\frac{\vec{u}_z\times\vec{E}_2}{c}\end{aligned}\tag{10.3.517}$$

They correspond to charges animated by an accelerated movement $\dot{\vec{\beta}}=\dot{\vec{v}}/c$ but which have a decreasing amplitude with z:

$$\vec{E}_2\cong f(\vec{v},z^{-1})\qquad \vec{B}_2\cong f(\vec{v},z^{-1})\tag{10.3.518}$$

These are radiating fields known as the "electromagnetic wave" (E.M).

These waves propagate from the charge in the form of a spherical wave but which at a great distance can be locally considered as plane.

In the non-relativistic $\beta\ll 1$ case we get:

$$\begin{aligned}\vec{E}_2 &\cong \frac{q}{4\pi\varepsilon_0 c}\frac{\vec{u}_z\times\left(\vec{u}_z\times\dot{\vec{\beta}}\right)}{z}\\ \vec{B}_2 &\cong \frac{q}{4\pi\varepsilon_0 c^2}\frac{\vec{u}_z\times\left[\vec{u}_z\times\left(\vec{u}_z\times\dot{\vec{\beta}}\right)\right]}{z}\end{aligned}\tag{10.3.519}$$

We can also write the fields as following (for the last equality see the explanations just afterwards!):

$$\begin{aligned}\vec{E}_2 &\cong \frac{q}{4\pi\varepsilon_0 c}\frac{z\cdot\vec{u}_z\times\left(z\cdot\vec{u}_z\times\dot{\vec{\beta}}\right)}{z\cdot z\cdot z}=\frac{q}{4\pi\varepsilon_0 c z^3}\vec{z}\times\left(\vec{z}\times\dot{\vec{\beta}}\right)\\ \vec{B}_2 &\cong \frac{q}{4\pi\varepsilon_0 c^2 z^3}z\left(\dot{\vec{\beta}}\times\vec{z}\right)\end{aligned}\tag{10.3.520}$$

Indeed, let us analyse how we can rewrite:

$$\vec{u}_z\times\left[\vec{u}_z\times\left(\vec{u}_z\times\dot{\vec{\beta}}\right)\right]\tag{10.3.521}$$

With a small change of variable we recognize again the triple cross-product (see section Vector Calculus page 1835):

$$\left.\begin{aligned}\vec{A}&=\vec{u}_z\\ \vec{B}&=\vec{u}_z\\ \vec{C}&=\vec{u}_z\times\dot{\vec{\beta}}\end{aligned}\right\}:\vec{A}\times\left(\vec{B}\times\vec{C}\right)=\vec{B}\left(\vec{A}\cdot\vec{C}\right)-\vec{C}\left(\vec{A}\circ\vec{B}\right)\tag{10.3.522}$$

Therefore:

$$\vec{u}_z\times\left[\vec{u}_z\times\left(\vec{u}_z\times\dot{\vec{\beta}}\right)\right]=\vec{u}_z\left[\vec{u}_z\circ\left(\vec{u}_z\times\dot{\vec{\beta}}\right)\right]-\left(\vec{u}_z\times\dot{\vec{\beta}}\right)\left(\vec{u}_z\circ\vec{u}_z\right)\tag{10.3.523}$$

It is immediate that $\vec{u}_z\circ\left(\vec{u}_z\times\dot{\vec{\beta}}\right)=0$, then it remains:

$$\vec{u}_z\times\left[\vec{u}_z\times\left(\vec{u}_z\times\dot{\vec{\beta}}\right)\right]=-\left(\vec{u}_z\times\dot{\vec{\beta}}\right)\left(\vec{u}_z\circ\vec{u}_z\right)=-\left(\vec{u}_z\circ\vec{u}_z\right)\left(\vec{u}_z\times\dot{\vec{\beta}}\right)=\left(\vec{u}_z\circ\vec{u}_z\right)\left(\dot{\vec{\beta}}\times\vec{u}_z\right)\tag{10.3.524}$$

Therefore:

$$\begin{aligned}\vec{B}_2 &\cong \frac{q}{4\pi\varepsilon_0 c^2}\frac{\vec{u}_z\times[\vec{u}_z\times(\vec{u}_z\times\dot{\vec{\beta}})]}{z}=\frac{q}{4\pi\varepsilon_0 c^2}\frac{zz(\vec{u}_z\circ\vec{u}_z)(\dot{\vec{\beta}}\times z\vec{u}_z)}{z^4}\\ &=\frac{q}{4\pi\varepsilon_0 c^2}\frac{(\vec{z}\circ\vec{z})(\dot{\vec{\beta}}\times\vec{z})}{z^4}=\frac{q}{4\pi\varepsilon_0 c^2 z^3}z(\dot{\vec{\beta}}\times z\vec{u}_z)=\frac{q}{4\pi\varepsilon_0 c^2 z}(\dot{\vec{\beta}}\times\vec{u}_z)\end{aligned} \quad (10.3.525)$$

In summary, keeping only the radiated part of the fields (!), we have:

- For the relativistic case:

$$\left.\begin{aligned}\vec{E}&\cong\frac{q}{4\pi\varepsilon_0 c}\frac{\vec{u}_z\times(\vec{u}_z-\vec{\beta})\times\dot{\vec{\beta}}}{z\left(1-\vec{\beta}\circ\vec{u}_z\right)^3}\\ \vec{B}&\cong\frac{q}{4\pi\varepsilon_0 c^2}\frac{\vec{u}_z\times[\vec{u}_z\times(\vec{u}_z\times\dot{\vec{\beta}})]}{z}\end{aligned}\right\}\vec{B}\cong\frac{\vec{u}_z\times\vec{E}}{c} \quad (10.3.526)$$

- And for the non-relativistic case:

$$\left.\begin{aligned}\vec{E}&=\frac{q}{4\pi\varepsilon_0 c z^3}\vec{z}\times(\vec{z}\times\dot{\vec{\beta}})\\ \vec{B}&=\frac{q}{4\pi\varepsilon_0 c^2 z^3}(\dot{\vec{\beta}}\times\vec{z})\end{aligned}\right\}\vec{B}\cong-\frac{\vec{z}\times\vec{E}}{c} \quad (10.3.527)$$

From now on, as the reader will have very likely noticed it, we will only work with the radiated fields and we will remove the index 2 from $\vec{E}_2$ and $\vec{B}_2$.

10.3.5.3 Power radiated through an elementary surface

Let us seek a relation for the power radiated through an elementary surface as function of $\vec{B}$.

Let us recall that we have(see page 4156):

$$\mathrm{d}P=\frac{1}{\mu_0}(\vec{E}\times\vec{B})\circ\mathrm{d}\vec{S} \quad (10.3.528)$$

And the relation above for the relativistic case:

$$\vec{B}\cong\frac{\vec{u}_z\times\vec{E}}{c} \quad (10.3.529)$$

We want to extract $\vec{E}$ as function of $\vec{B}$ from that latter relation to put it in the expression of the power.

For that purpose we use the following trick by calculating:

$$\vec{u}_z\times\vec{B}\cong\frac{\vec{u}_z\times(\vec{u}_z\times\vec{E})}{c} \quad (10.3.530)$$

We recognize here again the triple cross-product (see section Vector Calculus page 1835). Therefore if we put:

$$\left.\begin{aligned}\vec{A}&=\vec{u}_z\\ \vec{B}&=\vec{u}_z\\ \vec{C}&=\vec{E}\end{aligned}\right\}:\vec{A}\times\left(\vec{B}\times\vec{C}\right)=\vec{B}\left(\vec{A}\circ\vec{C}\right)-\vec{C}\left(\vec{A}\circ\vec{B}\right)\tag{10.3.531}$$

Therefore:

$$\vec{u}_z\times\left(\vec{u}_z\times\vec{E}\right)=\vec{u}_z\circ(\vec{u}_z\circ\vec{E})-\vec{E}(\vec{u}_z\circ\vec{u}_z)\tag{10.3.532}$$

For an electromagnetic wave far enough from its emission source, we have a plane wave as a first approximation and therefore: $\vec{E}\perp\vec{B}$ and $\{\vec{E},\vec{B}\}\perp\vec{z}$ which leads obviously to $\vec{u}_z\circ\vec{E}=0$, hence:

$$\vec{u}_z\times\left(\vec{u}_z\times\vec{E}\right)=-\vec{E}(\vec{u}_z\circ\vec{u}_z)\cong-\vec{E}\tag{10.3.533}$$

Therefore under these conditions:

$$\vec{u}_z\times\vec{B}\cong\frac{\vec{u}_z\times(\vec{u}_z\times\vec{E})}{c}\cong\frac{-\vec{E}}{c}\tag{10.3.534}$$

So:

$$\vec{E}=-c(\vec{u}_z\times\vec{B})\tag{10.3.535}$$

Injecting this into:

$$\mathrm{d}P=\frac{1}{\mu_0}(\vec{E}\times\vec{B})\circ\mathrm{d}\vec{S}\tag{10.3.536}$$

we get:

$$\mathrm{d}P=-\frac{c}{\mu_0}[(\vec{u}_z\times\vec{B})\times\vec{B}]\circ\mathrm{d}\vec{S}\tag{10.3.537}$$

We recognize here again the triple cross-product (see section Vector Calculus page 1835). So let us put:

$$\left.\begin{aligned}\vec{A}&=\vec{u}_z\\ \vec{B}&=\vec{B}\\ \vec{C}&=\vec{B}\end{aligned}\right\}:\vec{A}\times\left(\vec{B}\times\vec{C}\right)=\vec{B}\left(\vec{A}\circ\vec{C}\right)-\vec{C}\left(\vec{A}\circ\vec{B}\right)\tag{10.3.538}$$

Therefore:

$$(\vec{u}_z\times\vec{B})\times\vec{B}=\vec{B}\underbrace{(\vec{u}_z\circ\vec{B})}_{=0\text{ because }\perp}-\vec{u}_z\underbrace{(\vec{B}\circ\vec{B})}_{\|\vec{B}\|^2}=-\vec{u}_z\|\vec{B}\|^2\tag{10.3.539}$$

Hence:

$$\mathrm{d}P=\varepsilon_0c^3\|\vec{B}\|^2\vec{u}_z\circ\mathrm{d}\vec{S}\tag{10.3.540}$$

But we know that for the non-relativistic case:

$$\vec{B} \cong \frac{q}{4\pi\varepsilon_0 c^2 z}(\dot{\vec{\beta}} \times \vec{u}_z) \tag{10.3.541}$$

Injecting it in the relation above we get:

$$\begin{aligned}\mathrm{d}P &\cong \varepsilon_0 c^3 \left\| \frac{q}{4\pi\varepsilon_0 c^2 z} \underbrace{(\dot{\vec{\beta}} \times \vec{u}_z)}_{\|\dot{\vec{\beta}}\|\|\vec{u}_z\|\sin^2(\theta)} \right\|^2 \vec{u}_z \circ \mathrm{d}\vec{S} = \frac{q^2\|\dot{\vec{\beta}}\|^2}{16\pi^2\varepsilon_0 c}\frac{\sin^2(\theta)}{z^2}\vec{u}_z \circ \mathrm{d}\vec{S} \\ &= \frac{q^2\dot{\beta}^2}{16\pi^2\varepsilon_0 c}\frac{\vec{u}_z \circ \mathrm{d}\vec{S}}{z^2}\sin^2(\theta)\end{aligned} \tag{10.3.542}$$

Assuming that the following fraction can be interpreted as the solid angle that is the cone of vision of the surface $\mathrm{d}\vec{S}$ whose vertex coincides with the center of the charge and its axis coincides with the direction of $\vec{u}_z$:

$$\mathrm{d}\Omega = \frac{\vec{u}_z \circ \mathrm{d}\vec{S}}{z^2} \tag{10.3.543}$$

So finally we get:

$$\frac{\mathrm{d}P}{\mathrm{d}\Omega} \cong \frac{q^2\dot{\vec{\beta}}^2}{16\pi^2\varepsilon_0 c} = \frac{q^2\vec{a}^2}{16\pi^2\varepsilon_0 c^3}\sin^2(\theta) \tag{10.3.544}$$

where for recall, $\vec{a}$ is the acceleration of the electric charge $t' = t - z/c$.

So what we can say so far is that:

- The radiated power per steradian emitted by the accelerated charge is proportional to the square of its acceleration and to the square of its electric charge
- It is zero in the direction of the acceleration and maximum in the direction perpendicular $\theta = \pm\pi/2$.
- Importantly, it does not depend on the measuring distance (this is an approximation obviously!). This allows to detect it at a great distance.

We will now calculate the total power radiated by the accelerated electric charge on a sphere having for center the point O (the electric charge itself!) and for radius z (the distance between the electric charge and the observation point). In this case, it is useful to work with spherical coordinates

In spherical coordinates, a surface element $\mathrm{d}\vec{S}$ traced on the surface of the sphere of origin O and of radius z has the expression (see section Geometric Shapes page 2968):

$$\mathrm{d}\vec{S} = z^2\sin(\theta)\mathrm{d}\theta\mathrm{d}\phi\vec{u}_z \tag{10.3.545}$$

Injecting it in:

$$\mathrm{d}P \cong \frac{q^2\dot{\beta}^2}{16\pi^2\varepsilon_0 c}\frac{\vec{u}_z \circ \mathrm{d}\vec{S}}{z^2}\sin^2(\theta) \tag{10.3.546}$$

we get (see the usual primitives in the section of Differential and Integral Calculus):

$$\begin{gathered}P_T\cong\frac{q^2a^2}{16\pi^2\varepsilon_0c^3}\int\limits_0^{\pi}\sin^3\theta\mathrm{d}\theta\int\limits_0^{2\pi}\mathrm{d}\phi=\frac{q^2a^2}{16\pi^2\varepsilon_0c^3}\left[-\cos(\theta)+\frac{\cos^3(\theta)}{3}\right]_0^{\pi}\cdot 2\pi\\=\frac{q^2a^2}{8\pi\varepsilon_0c^3}\left[\underbrace{\left(1-\frac{1}{3}\right)-\left(-1+\frac{1}{3}\right)}_{=2-\frac{2}{3}=\frac{4}{3}}\right]=\frac{q^2a^2}{6\pi\varepsilon_0c^3}\end{gathered}\tag{10.3.547}$$

Therefore the expression of the total power radiated by an accelerated charge in the non-relativistic case is given by:

$$P_T\cong\frac{q^2a^2}{6\pi\varepsilon_0c^3}\tag{10.3.548}$$

And as we can notice it, the distance z has vanished as we would expect it!

The total power emitted in the form of radiation from a charge moving in an accelerated motion is proportional to the square of the acceleration a and the square of its charge q. It is independent of the distance z.

Now we want to calculate the power emitted by the electric charge in the relativistic case!

In the application of special relativity, the following two elements will be taken into account:

- there are invariants in all inertial frames of reference (here: q, ε_0, c).
- in the relativistic expression of power, we must be able to fall back on the previous relation if we consider the case of low speeds compared to the speed of light!

If we multiply the norm of the four-vector acceleration derived in the section of Special Relativity (see page 5649):

$$a^va_v=-\gamma^6\left[\vec{a}^2-(\vec{a}\times\vec{\beta})^2\right]\tag{10.3.549}$$

by the mass of the electric charge, we have then what is conventionally named the "Abraham-Becker radiation damping force".

The reader must remember that this result is performed in an instantaneous rest frame where the charged particle velocity and acceleration are collinear, and is independent of the internal structure of the particle and shows that the radiation damping force is essentially a relativistic effect.

$$P_T\cong-\frac{q^2}{6\pi\varepsilon_0c^3}\gamma^6\left[\vec{a}^2-(\vec{a}\times\vec{\beta})^2\right]\tag{10.3.550}$$

That is the "Liénard's generalization" to the Larmor's equation[5] we will just see afterwards:

$$\boxed{P_T\cong-\frac{q^2}{6\pi\varepsilon_0c^3(1-\beta^2)^3}\left[\vec{a}^2-(\vec{a}\times\vec{\beta})^2\right]}\tag{10.3.551}$$

---

[5] It was obtained by A. Liénard in 11898 (holocene calendar), only one year after the publication of Larmor's result, and seven years prior to the invention of special relativity!

We see obviously that for small speed $v \ll c$, it leads to:

$$\boxed{P_T \cong -\frac{q^2\vec{a}^2}{6\pi\varepsilon_0 c^3}} \tag{10.3.552}$$

And this is the result named the "Larmor's equation". It implies that any charged particle radiates when accelerated and that the total radiated power is proportional to the square of the acceleration. The largest astrophysical accelerations are usually produced by electromagnetic forces, so the acceleration is proportional to the charge/mass ratio of the particle.

This result is also an interesting and very difficult paradox that a number of physicists have addressed over the last century, including Pauli. If the charge radiates under uniform acceleration, it will violate the principle of equivalence. The famous Richard Feynman was one of those rare physics celebrities who did insist, based on his line of argumentation, that this is false paradigm that an uniformly accelerated charge actually radiates, and that actually it does not, and only oscillating and other non-uniform accelerations can lead to radiation (people interested in a detailed analysis of this point of view can read [496] and [495]). But with this remarkable exception, the overwhelming consensus of the physicists is that a uniformly accelerated charge does radiate! There is a large amount of literature and experience (worldwide) since many decades on synchrotron radiation (leptons but also hadrons for very high energies) and no further fundamental discussion required. Also keep in mind that accelerated charged particles in LINACs radiate a lot.

Remark

We will also see during our study or Special Relativity that the covariance of Maxwell's equations under general coordinate transformations, that, despite inertial observers can indeed detect electromagnetic radiation emitted from an accelerated charge, comoving observers will see only a static electric field. Therefore when the charge radiates it doesn't lost momentum!

Also the ready should keep in mind that a charge moving along a geodesic doesn't accelerate by definition (!), then it doesn't emit any electromagnetic radiation. Furthermore let us remind that we will prove during our study of General Relativity that gravitation isn't a force!

Larmor's extremely useful equation will be the basis for our derivations of radiation from a short dipole antenna as well as for synchrotron emission from astrophysical sources. Beware that Larmor's equation is non-relativistic; it is valid only in frames moving at velocities $v \ll c$ with respect to the radiating particle. Also the reader should keep in mind that an electric charge in free fall in a uniform and homogeneous gravitational field does not radiate because it is following a geodesic, i.e. it is moving by inertia!

Also, Larmor's equation does not incorporate the constraints of quantum mechanics, so it should be applied only with great caution to microscopic systems such as atoms. For example, Larmor's equation incorrectly predicts that the electron orbiting in the lowest energy level of a hydrogen atom will quickly radiate away all of its kinetic energy and fall into the nucleus. On the other hand, it correctly predicts the radio power emitted by an electron orbiting in a very high energy level of a hydrogen atom.

$$P_T \cong \frac{q^2}{6\pi\varepsilon_0 c^3(1-\beta^2)^3}\left[\vec{a}^2-(\vec{a}\times\vec{\beta})^2\right] \tag{10.3.553}$$

Anyway, that's a lot of words and not much to look at, so how about some soothing animations?

Below is a Flash animation illustrating the effect, made with MATLAB®, that will play in a PDF Reader having Adobe Flash installed and activated (otherwise see here: https://vimeo.com/575746945):

Figure 10.76: **Synchrotron radiation of a circle moving particle with $\gamma = 1.2$ (author: Jason Cole)**

The top left represents the Poynting flux S, or the radiation which would be observed. The retarded time is plotted beneath, and the two components of the electric field are plotted on the right. You can see how the retarded time is trailing the particle, and spreads out in a circle at the speed of light.

Now remember that by definition the Lorentz factor is given by:

$$\gamma=\frac{1}{\sqrt{1-\frac{v^2}{c^2}}}=\frac{1}{\sqrt{1-\beta^2}} \tag{10.3.554}$$

We can turn up the energy a bit, now $\gamma=5$ (the animation will play in a PDF Reader having Adobe Flash installed and activated otherwise see here: https://vimeo.com/575747231):

Figure 10.77: Synchrotron radiation of a circle moving particle with $\gamma=5$ (author: Jason Cole)

Here the electron is much closer to the speed of light, and is consequently chasing it's own light-cone – notice how there is a much larger jump in retarded time (from yellow to green). This causes the Poynting flux to be squeezed into a smaller time window, it's much more compressed, intense, and energetic. The temporal shortening of the radiation pulse corresponds to a broad bandwidth, and it is precisely this pulse of radiation that corresponds to broadband synchrotron radiation.

What about something less extreme, an electron travelling at constant velocity in a straight line, again at $\gamma = 1.2$ (the animation will play in a PDF Reader having Adobe Flash installed and activated otherwise see here: `https://vimeo.com/575745050`):

Figure 10.78: **Synchrotron radiation of a uniform moving particle with $\gamma = 1.2$ (author: Jason Cole)**

We now see that the field is being dragged along with the particle, but otherwise not radiating. Because the electron is travelling significantly slower than the speed of light, the field looks relatively undistorted. If you squint you'll notice that the transverse field E_y is stronger than the longitudinal field, which is a manifestation of special relativity (Lorentz transforming a uniform field back into the lab frame where the charge is moving).

Back to something more interesting, a simple dipole motion at $\gamma = 1.2$ (the animation will play in a PDF Reader having Adobe Flash installed and activated otherwise see here: https://vimeo.com/575745668):

Figure 10.79: Synchrotron simple dipole motion with $\gamma = 1.2$ (author: Jason Cole)

We see what we hoped for, oscillating radiation perpendicular to the dipole, and zero far-field radiation parallel to the dipole. In the near-field (close to the dipole), there are electric field components everywhere, but it is clear that these must correspond to non-radiative terms and they die off rapidly.

Finally, let us turn to a use for this radiation: ondulators/wigglers. These are machines which purposefully oscillate an electron bunch to force synchrotron radiation emission, where the electron travels with high velocity in one direction and oscillates in the other (the animation will play in a PDF Reader having Adobe Flash installed and activated otherwise see here: https://vimeo.com/575747509):

Figure 10.80: Synchrotron uniform oscillating moving particle (author: Jason Cole)

Notice that:

$$P_T \cong \frac{q^2\gamma^6}{6\pi\varepsilon_0 c^3}\left[\vec{a}^2-(\vec{a}\times\vec{\beta})^2\right] \tag{10.3.555}$$

as:

$$E=\gamma m_0 c^2 \tag{10.3.556}$$

can be written:

$$P_T \cong \frac{q^2E^6}{6\pi m^6\varepsilon_0 c^{15}}\left[\vec{a}^2-(\vec{a}\times\vec{\beta})^2\right] \tag{10.3.557}$$

The killer for CERN is that E^6 factor because if we double the energy of a particle, it radiates much more times the equivalent received energy. In the case of electrons, this very quickly becomes such a high power that the accelerator can't accelerate any more – the electrons radiate away their energy quicker than it can be pumped in. For protons of the same energy, because of the m^6 factor at the denominator the E^6 factor is 10^{19} times smaller (as $m_p/m_e \cong 1836$), so radiate much less. This is why the LHC is so big, and uses protons rather than electrons. There are plans to build an even bigger LHC, unimaginatively named the "Very Large Hadron Collider", or to accelerate electrons in a straight line so that they radiate much less energy at the proposed International Linear Collider. Both of these propositions are very expensive though, and very large, so there is much interest in smaller accelerators like plasma accelerators.

We can also make appear the influence of the radius of the accelerator. Indeed, starting from:

$$P_T \cong \frac{q^2E^6}{6\pi m^6\varepsilon_0 c^{15}}\left[\|\vec{a}\|^2-\|\vec{a}\times\vec{\beta}\|^2\right] \tag{10.3.558}$$

Remember that we have seen in the section of Classical Mechanics that for a uniform circular motion:

$$\vec{a}=-\omega^2\begin{pmatrix}R\sin(\omega t)\\ R\cos(\omega t)\end{pmatrix}=-\omega^2\vec{r} \tag{10.3.559}$$

Therefore it follows:

$$\|\vec{a}\|^2=\omega^4R^2=\left(\frac{v}{R}\right)^4R^2=\frac{v^2}{R^2} \tag{10.3.560}$$

Hence:

$$P_T \propto \frac{q^2E^6v^2}{6\pi m^6\varepsilon_0 c^{15}R^2} \tag{10.3.561}$$

We therefore understand that the bigger is the radius, the smaller is the radiated power. This is why linear accelerator (considers as infinite curvature radius[6]) are much more interesting than circular accelerator (but more difficult to build as we need to find and large enough area to build them...).

10.3.5.3.1 Larmor's equation (non-relativistic derivation)

Maxwell's equations imply that all classical electromagnetic radiation is generated by accelerating electrical charges. It is possible to derive the intensity and angular distribution of the radiation from a point charge (a charged particle) subject to an arbitrary but small acceleration $\Delta v/\Delta t$ via Maxwell's equations, but the complicated math we have seen above obscures the physical interpretation that remains clear in J. J. Thomson's illuminating derivation below for the non-relativistic case (see [**<empty citation>**]).

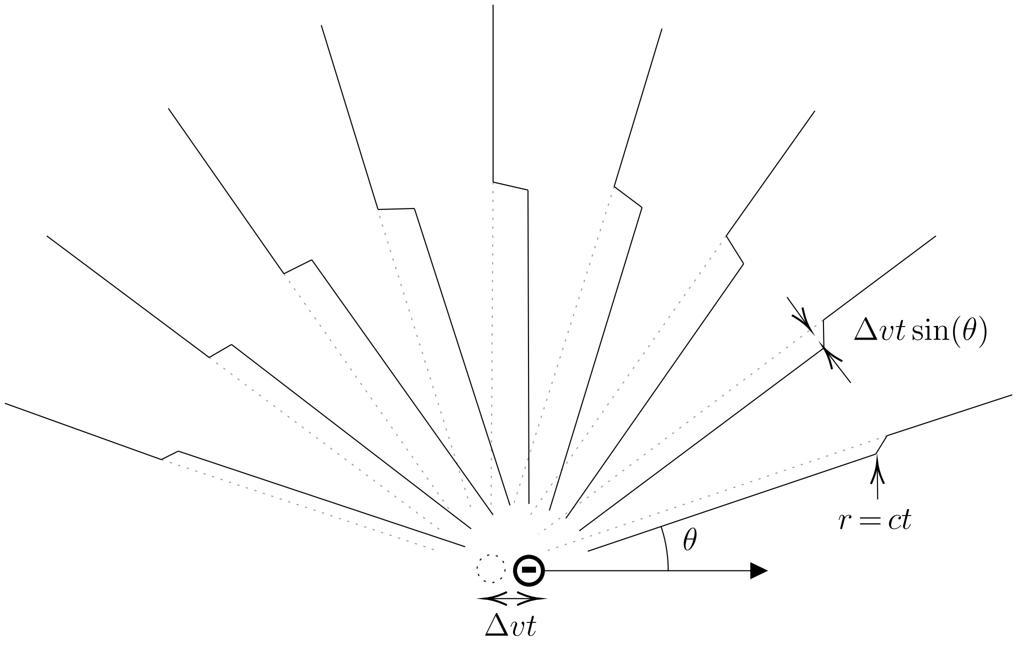
---

[6] And obviously for linear accelerator: $\|\vec{a}\times\vec{\beta}\|^2=0$ as the acceleration is collinear to the speed.

If a particle with electrical charge q is at rest or moving with a constant velocity, its electric field lines are purely radial: $|\vec{E}| = E_r$. Suppose a charged particle initially at rest is accelerated to a small velocity $\Delta v \ll c$ in a short time Δt. This disturbs the lines of force, and the disturbance travels outward at the speed of light c. The figure below shows that at time t after the acceleration, the disturbance will have propagated to $r = ct$ and the perpendicular component of the electric field will have magnitude:

$$\frac{E_\perp}{E_r} = \frac{\Delta vt\sin(\theta)}{c\Delta t} \qquad (10.3.562)$$

where θ is the angle between the acceleration vector and the line of sight connecting the charge to the observer.



In the above figure the electric field lines from an accelerated electron. The dotted circle shows the initial position of the electron, and the dotted lines are the radial lines of force emanating from that position. At time t after a small acceleration $\Delta v/\Delta t$, the electron position has moved by Δvt and its lines of force have shifted transversely by $\Delta vt\sin(\theta)$. Coulomb's law for the radial component E_r of the electric field (electric force per unit charge) a distance r from a stationary charge q is:

$$E_r = \frac{1}{4\pi\varepsilon_0}\frac{q}{r^2} \qquad (10.3.563)$$

Substituting r/c for t in the prior-previous relation and using Coulomb's law to eliminate E_r gives:

$$E_\perp = \frac{1}{4\pi\varepsilon_0}\frac{q}{r^2}\left(\frac{\Delta v}{\Delta t}\right)\frac{r\sin(\theta)}{c^2} \qquad (10.3.564)$$

and:

$$E_\perp = \frac{1}{4\pi\varepsilon_0}\frac{q\dot{v}\sin(\theta)}{rc^2} \qquad (10.3.565)$$

where:

$$\dot{v} \equiv \lim_{\Delta t \to 0} \left(\frac{\Delta v}{\Delta t} \right) \tag{10.3.566}$$

The above relation of $E_\perp$ valid for any small acceleration, not just a sinusoidal acceleration at a single frequency. The transverse field $E_\perp$ at $r = ct$ mimics the acceleration at $t = 0$. Thus a sinusoidal acceleration would result in a sinusoidal variation of $E_\perp$ with the same frequency.

Notice that $E_\perp \propto r^{-1}$ falls more slowly with distance r than $E_r \propto r^{-2}$. Far from the charged particle, only $E_\perp$ will contribute significantly to the observed electric field. From a distant observer's point of view, only the "visible" component of acceleration perpendicular to the line of sight $(\dot{v}\sin(\theta))$ contributes to the radiated electric field; the "invisible" component of acceleration parallel to the line of sight $\dot{v}\cos(\theta))$ does not appear to radiate! Likewise, the radiated electric field is linearly polarized in the direction parallel to the component of acceleration perpendicular to the line of sight.

How much power is radiated in each direction? The Poynting flux, or power per unit area is given by (see page 4156):

$$\bar{S} = \frac{\hat{E}^2}{2\mu_0 c} = \frac{c\varepsilon_0 \hat{E}^2}{2} \tag{10.3.567}$$

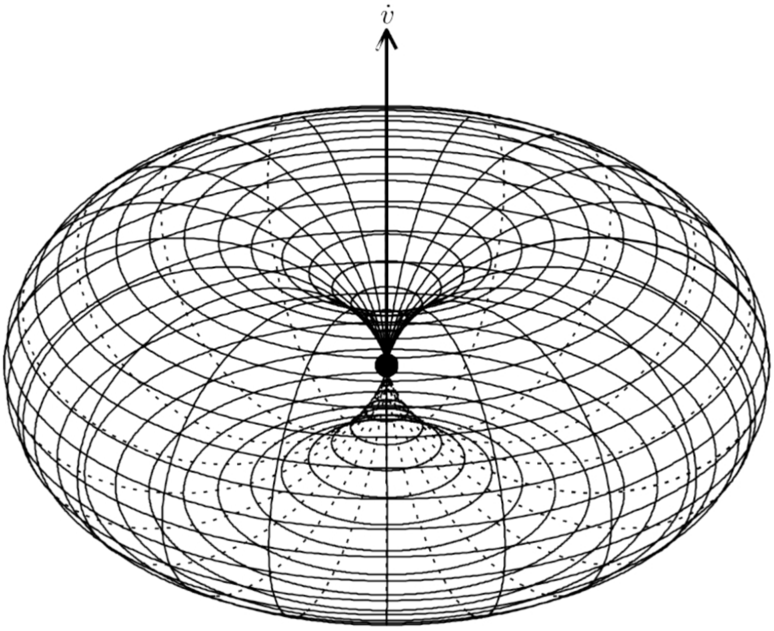
Inserting:

$$E_\perp = \frac{1}{4\pi\varepsilon_0} \frac{q\dot{v}\sin(\theta)}{rc^2} \tag{10.3.568}$$

into $\bar{S}$ gives:

$$\bar{S} = \frac{c\varepsilon_0}{2} \left(\frac{1}{4\pi\varepsilon_0} \frac{q\dot{v}\sin(\theta)}{rc^2} \right)^2 = \left(\frac{q^2\dot{v}^2}{32\varepsilon_0\pi^2 c^3} \right) \frac{\sin^2(\theta)}{r^2} \tag{10.3.569}$$

The accelerated charge radiates then with a dipolar power pattern $\propto \sin^2(\theta)$ shaped like a doughnut whose axis is parallel to the acceleration $\dot{v}$:



In the figure above the power pattern of Larmor radiation from a charged particle shown for an acceleration vector $\dot{v}$ tilted 60 degrees from the line of sight. The power received in any direction is proportional to the component of $\dot{v}$ perpendicular to the line of sight.

The total power emitted is the integral of $\bar{S}$ on a spherical surface of radius r:

$$P=\int\limits_{\text{sphere}}\bar{S}\mathrm{d}A=\frac{q^2\dot{v}^2}{32\varepsilon_0\pi^2c^3}2\int\limits_{\phi=0}^{2\pi}\int\limits_{\theta=0}^{\pi}\frac{\sin^2(\theta)}{r^2}r^2\sin(\theta)\mathrm{d}\theta\mathrm{d}\phi=\frac{q^2\dot{v}^2}{8\varepsilon_0\pi c^3}\int\limits_{\theta=0}^{\pi}\sin^3(\theta)\mathrm{d}\theta \quad (10.3.570)$$

The integral $\int_{\theta=0}^{\pi}\sin^3(\theta)\mathrm{d}\theta=4/3$, so the total power emitted by the accelerated charged particle is:

$$P=\frac{q^2\dot{v}^2}{8\varepsilon_0\pi c^3}\frac{4}{3}=\frac{q^2\dot{v}^2}{6\varepsilon_0\pi c^3}=\frac{q^2a^2}{6\varepsilon_0\pi c^3} \quad (10.3.571)$$

So we fall back on the classical Larmor's equation but without the complicated maths!

10.3.5.3.2 Abraham-Lorentz force

In the physics of electromagnetism, the "Abraham–Lorentz force" is an approximated recoil force on an accelerating charged particle caused by the particle emitting[7] electromagnetic radiation outside of a Quantum Mechanics framework. It is also named the "radiation reaction force" or "radiation damping force" or even the "self-force" or... "radiation resistance".

The formula - then we will derive further below in a special case - predates the theory of Special Relativity and is not valid at velocities on the order of the speed of light. Its relativistic generalization is named the "Abraham–Lorentz–Dirac force". Both of these are in the domain of classical physics, not quantum physics, and therefore may not be valid at distances of roughly the Compton wavelength or below. There is, however, an analogue of the formula that is both fully quantum and relativistic, named the "Abraham–Lorentz–Dirac–Langevin equation".

As we will see it, force is proportional to the square of the object's charge, times the jerk (rate of change of acceleration) that it is experiencing. The force points in the direction of the jerk. For example, in a cyclotron, where the jerk points opposite to the velocity, the radiation reaction is directed opposite to the velocity of the particle, providing a braking action. The Abraham–Lorentz force is also the source of the radiation resistance of a radio antenna radiating radio waves.

Remark

There seem to be pathological solutions of the Abraham–Lorentz–Dirac equation in which a particle accelerates in advance of the application of a force, so-named "pre-acceleration solutions". Since this would represent an effect occurring before its cause (retro-causality), some theories have speculated that the equation allows signals to travel backward in time, thus challenging the physical principle of causality. One resolution of this problem was discussed by Arthur D. Yaghjian and is further discussed by Fritz Rohrlich and Rodrigo Medina.

---

[7] The formula may also give an approximation when photons are absorbed.

The simplest non-rigorous derivation (in the engineer way of life...) for the self-force is found for periodic motion from the Larmor equation (see section Electrodynamics page 4192) for the power radiated from a point charge:

$$P_T=-\frac{q^2}{6\pi\varepsilon_0 c^3}\vec{a}^2 \tag{10.3.572}$$

If we assume the motion of a charged particle is periodic, then the average work done on the particle by the Abraham–Lorentz force is the Larmor power integrated over one period from τ_1 to τ_2:

$$\int\limits_{\tau_1}^{\tau_2}\vec{F}_{\text{rad}}\circ\vec{v}\mathrm{d}t=\int\limits_{\tau_1}^{\tau_2}P\mathrm{d}t=-\int\limits_{\tau_1}^{\tau_2}\frac{q^2}{6\pi\varepsilon_0 c^3}\vec{a}^2\mathrm{d}t=-\int\limits_{\tau_1}^{\tau_2}\frac{q^2}{6\pi\varepsilon_0 c^3}\frac{\mathrm{d}\vec{v}}{\mathrm{d}t}\circ\frac{\mathrm{d}\vec{v}}{\mathrm{d}t}\mathrm{d}t \tag{10.3.573}$$

The above expression can be integrated by parts. If we assume that there is periodic motion, the boundary term in the integral by parts disappears:

$$\int\limits_{\tau_1}^{\tau_2}\vec{F}_{\text{rad}}\circ\vec{v}\mathrm{d}t=-\frac{q^2}{6\pi\varepsilon_0 c^3}\frac{\mathrm{d}\vec{v}}{\mathrm{d}t}\circ\vec{v}\bigg|_{\tau_1}^{\tau_2}+\int\limits_{\tau_1}^{\tau_2}\frac{q^2}{6\pi\varepsilon_0 c^3}\frac{\mathrm{d}^2\vec{v}}{\mathrm{d}t^2}\circ\vec{v}\mathrm{d}t=-0+\int\limits_{\tau_1}^{\tau_2}\frac{q^2}{6\pi\varepsilon_0 c^3}\dot{\vec{a}}\circ\vec{v}\mathrm{d}t \tag{10.3.574}$$

Clearly, we can identify:

$$\boxed{\vec{F}_{\text{rad}}=\frac{q^2}{6\pi\varepsilon_0 c^3}\dot{\vec{a}}} \tag{10.3.575}$$

It should be noted that already in 11905 (holocene calendar), Abraham clearly stated that this result cannot be always correct! A more rigorous derivation, which does not require periodic motion, was found using an Effective Field Theory formulation. An alternative derivation, finding the fully relativistic expression, was found by Dirac.

Here $\vec{F}_{\text{rad}}$ is the force, $\dot{\vec{a}}$ is the derivative of acceleration, or the third derivative of displacement, also named "jerk".

Physically, an accelerating charge emits radiation (according to the Larmor formula), which carries momentum away from the charge. Since momentum is conserved, the charge is pushed in the direction opposite to the direction of the emitted radiation.

10.3.6 Dipole antenna radiation pattern derivation (Hertz Electric Dipole)

Now that we have acquire a quite good knowledge of Maxwell's equations, potential vector, power emission and retarded potential, we can finally study a simple antenna configuration.

Given a current distribution on an antenna, the problem is one of determining the $\vec{E}$ and $\vec{H}$ fields due to this current distribution which satisfies all four of Maxwell's equations along with the boundary conditions, if any.

In the vector potential approach (a very clever one and sadly we don't know who was the first who found that very clever trick) we carry out the solution to this problem in second steps by defining intermediate potential functions. In the first step, we determine the potential function due to the current distribution in sinusoidal regime and in the second step we solve the geometric problem under some approximations.

In the analysis that follows, the relationships between the vector potential and the current distribution as well as the $\vec{E}$ and H fields are derived. All four of Maxwell's equations are embedded in these relationships.

Let us start with the second Maxwell's equations:

$$\vec{\nabla}\circ\vec{B}=0 \tag{10.3.576}$$

Since the curl of a vector field is divergence-free (vector identity: $\vec{\nabla}\circ(\vec{\nabla}\times\vec{A})=0$), the magnetic intensity $\vec{B}$ can be expressed as a curl of an arbitrary vector field, $\vec{A}$ as we already now it! Thus:

$$\mu\vec{H}=\vec{B}=\vec{\nabla}\times\vec{A} \tag{10.3.577}$$

or:

$$\vec{H}=\frac{1}{\mu}\vec{\nabla}\times\vec{A} \tag{10.3.578}$$

Substituting this into the third Mawell's equation:

$$\vec{\nabla}\times\vec{E}=-\frac{\partial\vec{B}}{\partial t} \tag{10.3.579}$$

in sinsuoidal regime and phasor notation:

$$\vec{\nabla}\times\vec{E}=-\frac{\partial\vec{B}}{\partial t}=-\mathrm{i}\omega\vec{B}=-\mathrm{i}\omega\mu\vec{H}=-\mathrm{i}\omega(\vec{\nabla}\times\vec{A}) \tag{10.3.580}$$

Or:

$$\vec{\nabla}\times(\vec{E}+\mathrm{i}\omega\vec{A})=\vec{0} \tag{10.3.581}$$

Since as we know that the curl of a gradient function is zero (vector identity: $\vec{\nabla}\times\vec{\nabla}\cdot U=0$ as proved in the section of Vector Calculus page 1912), the previous relation suggests that the quantity in brackets can be replaced by the gradient of a scalar function. Specifically, a scalar potential function U is defined such that:

$$\vec{E}+\mathrm{i}\omega\vec{A}=-\vec{\nabla}\cdot U \tag{10.3.582}$$

Using this we relate the $\vec{E}$ field to the potential functions as:

$$\vec{E}=-(\vec{\nabla}\cdot U+\mathrm{i}\omega\vec{A}) \tag{10.3.583}$$

The following relations:

$$\vec{H}=\frac{1}{\mu}\vec{\nabla}\times\vec{A} \qquad \vec{E}=-(\vec{\nabla}\cdot U+\mathrm{i}\omega\vec{A}) \tag{10.3.584}$$

relate the $\vec{H}$ and $\vec{E}$ fields to the potential functions $\vec{A}$ and $\vec{U}$.

Now, to satisfy the fours Maxwell's equation:

$$\vec{\nabla}\times\vec{H}=\vec{j}+\varepsilon\frac{\partial\vec{E}}{\partial t} \tag{10.3.585}$$

in sinusoidal (harmonic) regime in phasor notation:

$$\vec{\nabla}\times\vec{H}=\vec{j}+\mathrm{i}\varepsilon\omega\vec{E} \tag{10.3.586}$$

we substitute the expression for the $\vec{E}$ and $\vec{H}$ fields in therms of the potentials functions:

$$\frac{1}{\mu}\vec{\nabla}\times(\vec{\nabla}\times\vec{A})=-\mathrm{i}\varepsilon\omega(\vec{\nabla}\cdot U+\mathrm{i}\omega\vec{A})+\vec{j} \tag{10.3.587}$$

which is valid for a homogeneous medium. Expanding the left hand side using the vector identity proved in the section of Vector Calculus:

$$\vec{\nabla}\times(\vec{\nabla}\times\vec{A})=\vec{\nabla}\cdot(\vec{\nabla}\circ\vec{A})-\vec{\nabla}^2\vec{A} \tag{10.3.588}$$

We have (after rearranging):

$$\vec{\nabla}^2\vec{A}+\omega^2\mu\varepsilon\vec{A}=-\mu\vec{j}+\vec{\nabla}\cdot(\vec{\nabla}\circ\vec{A}+\mathrm{i}\omega\mu\varepsilon U) \tag{10.3.589}$$

So far we have satisfied three of Maxwell's four equations. Note that only the curl of $\vec{A}$ is defined so far. Since the curl and divergence are two independent parts of any vector field, we can now define the divergence of $\vec{A}$. We define $\vec{\nabla}\circ\vec{A}$ so as to relate $\vec{A}$ and U as well as simplify the previous relation by eliminating the second term on the right hand side of the equality. We relate $\vec{A}$ and U by the relation:

$$\vec{\nabla}\circ\vec{A}=-\mathrm{i}\omega\mu\varepsilon U \tag{10.3.590}$$

This empirical choice (relation) is as we already known, named the "Lorentz gauge". With this the magnetic vector potential $\vec{A}$ satisfies the vector wave equation (Helmoltz inhomogenous equation):

$$\vec{\nabla}^2\vec{A}+k^2\vec{A}=-\mu\vec{j} \tag{10.3.591}$$

where $k=\omega\sqrt{\mu\varepsilon}$ is the "propagation constant" (in radian per meter) in the medium.

Now, to satisfy the last remaining Maxwell's equation:

$$\vec{\nabla}\circ\vec{E}=\frac{\rho_Q}{\varepsilon} \tag{10.3.592}$$

For this purpose we substitute:

$$\vec{E}=-(\vec{\nabla}\cdot U+\mathrm{i}\omega\vec{A})=-\vec{\nabla}\cdot U-\mathrm{i}\omega\vec{A} \tag{10.3.593}$$

in it to get:

$$\vec{\nabla}\circ(-\vec{\nabla}\cdot U-\mathrm{i}\omega\vec{A})=\frac{\rho_Q}{\varepsilon} \tag{10.3.594}$$

Or:

$$\vec{\nabla}^2U+\mathrm{i}\omega(\vec{\nabla}\circ\vec{A})=-\frac{\rho_Q}{\varepsilon} \tag{10.3.595}$$

Eliminating $\vec{A}$ from this equation using the Lorentz gauge:

$$\vec{\nabla}^2U+k^2U=-\frac{\rho_Q}{\varepsilon} \tag{10.3.596}$$

Thus, both $\vec{A}$ and U must satisfy the wave equation, the source function being the current density for the magnetic vector potential, $\vec{A}$, and the charge density for the electric scalar potential function U.

Let us seek now a solution to:

$$\vec{\nabla}^2U+k^2U=-\frac{\rho_Q}{\varepsilon} \tag{10.3.597}$$

Consider a spherically symmetric charge distribution of finite volume V, centered on the origin. Our goal is to compute the scalar potential $U(x,y,z)$ (or $U(r,\theta,\phi)$) due to this source, which is the solution of the in-homogeneous wave equation as given by above. Since the charge is spherically symmetric, we will solve the wave equation in the spherical coordinate system.The Laplacian $\vec{\nabla}^2U$ in the spherical coordinate system is written as proved in the section of Vector Calculus (see page 1899) by:

$$\vec{\nabla}^2U=\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial U}{\partial r}\right)+\frac{1}{r^2\sin(\theta)}\frac{\partial}{\partial U}\left(\sin(\theta)\frac{\partial\theta}{\partial\theta}\right)+\frac{1}{r^2\sin^2(\theta)}\frac{\partial^2U}{\partial\phi^2} \tag{10.3.598}$$

The scalar potential $U(r,\theta,\phi)$, produced by a spherically symmetric charge distribution is independent of θ and ϕ. Therefore, the wave equation reduces to:

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial U}{\partial r}\right)+k^2U=-\frac{\rho_Q}{\varepsilon} \tag{10.3.599}$$

The right hand side of this equation is zero everywhere except at the source itself. Therefore, in the source-free region, U satisfies the homogeneous wave equation:

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial U}{\partial r}\right)+k^2U=0 \tag{10.3.600}$$

The solutions for U are the scalar spherical waves given by (or the sum of them!):

$$\boxed{U(r)=U_0^{\pm}\frac{e^{\mp ikr}}{r}} \tag{10.3.601}$$

where U_0^+ is a complex amplitude constant and $e^{-\text{i}kr}/r$ is a spherical wave travelling in the $+r$-direction. U_0^- is the complex amplitude of the scalar spherical wave $e^{+\text{i}kr}/r$ travelling in the $-r$-direction. By substituting this in the wave equation, it can be shown that it satisfies the homogeneous wave equation!

Consider a static point charge q kept at a point with position vector $\vec{r}\,'$. The electric potential, U, at a point (r,θ,ϕ) (or in cartesian coordinates $P(x,y,z)$), with the position vector $\vec{r}$ is given by:

$$U(r,\theta,\phi)=\frac{q}{4\pi\varepsilon R}\;[\text{J}\cdot\text{C}^{-1}] \tag{10.3.602}$$

(equivalently given in Volts [V]) where R is the distance between the charge and the observation point:

$$R=\|\vec{R}\|=\|\vec{r}-\vec{r}\,'\| \tag{10.3.603}$$

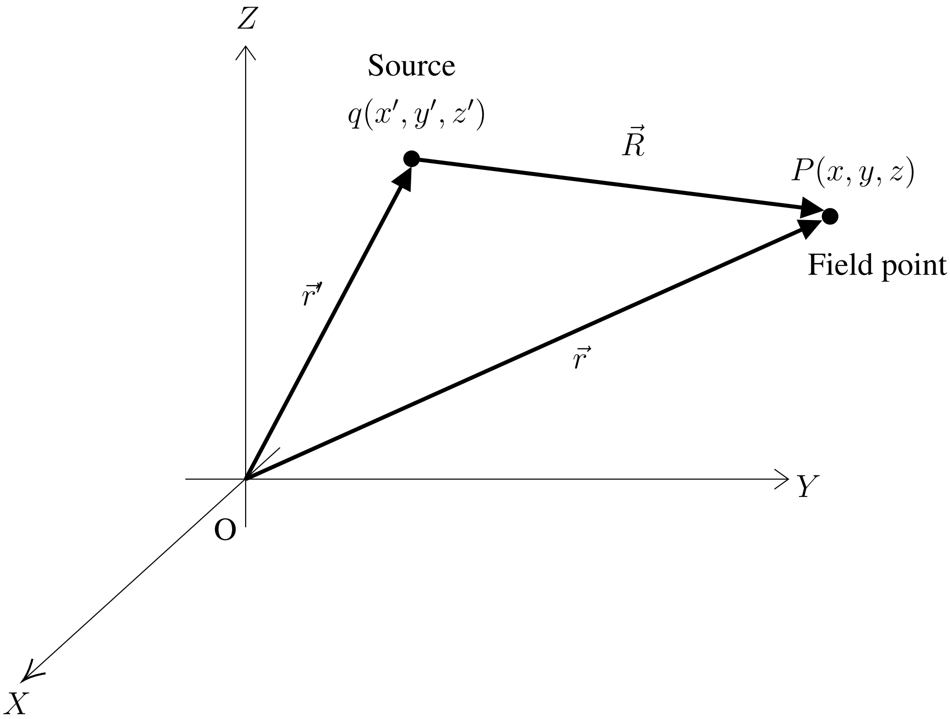


Figure 10.81: Position vectors of source and field points

If there are more than one point charges, the potential is obtained by the superposition principle, i.e., summing the contributions of all the point charges. If the source is specified as a charge density distribution over a volume, the potential at any field point is obtained by integration over the source volume. To do this, we first consider a small volume $\mathrm{d}V$ centered on $\vec{r}\,'$.

The charge contained in this volume is then $\rho(\vec{r}^{'})\mathrm{d}V$, where $\rho(\vec{r}^{'})$ is the volume charge density distribution function. So we can compute the potential at any field point $\vec{r}$ due to charge contained in the volume $\mathrm{d}V$ using:

$$\mathrm{d}U(r,\theta,\phi)=\frac{\rho \mathrm{d}V}{4\pi\varepsilon R} \tag{10.3.604}$$

Now let us recall the solution we get in a sinusoidal regime for a spherical symmetry configuration:

$$U(r)=U_0^{\pm}\frac{e^{\mp \mathrm{i}kr}}{r} \tag{10.3.605}$$

So by analogy we may write:

$$\mathrm{d}U(r,\theta,\phi)=\frac{\rho e^{-\mathrm{i}kr}\mathrm{d}V}{4\pi\varepsilon R} \tag{10.3.606}$$

The potential at point (r,θ,ϕ) due to a charge distribution $\rho(x',y',z')$ is obtained by integrating the previous relation over the source distribution:

$$U(r,\theta,\phi)=\frac{1}{4\pi\varepsilon}\oint_V\frac{\rho e^{-\mathrm{i}kR}\mathrm{d}V}{4\pi R} \tag{10.3.607}$$

This must be compared with the couple of relations we had proved earlier during our study of the Bremsstrahlung (see section Electrodynamics page 4163):

$$\begin{aligned}U(\vec{R},t)&=\frac{1}{4\pi\varepsilon_0}\oint_V\frac{\rho(\vec{r},t')\mathrm{d}V}{\|\vec{R}-\vec{r}\|}\\ \vec{A}(\vec{R},t)&=\frac{\mu_0}{4\pi}\oint_V\frac{\vec{J}(\vec{r},t')\mathrm{d}V}{\|\vec{R}-\vec{r}\|}\end{aligned} \tag{10.3.608}$$

Then the same developments, lead us to write:

$$\vec{A}(\vec{R},t)=\frac{\mu}{4\pi}\oint_V\frac{\vec{J}(\vec{r},t')e^{-\mathrm{i}kR}\mathrm{d}V}{\|\vec{R}-\vec{r}\|} \tag{10.3.609}$$

Now, let us consider a dipole antenna centered on the origin and oriented along the Z-axis with length L. We want to find the radiation pattern equation in polar coordinates (as the configuration is quite symmetrical!).

The current on the antenna will be approximately sinusoidal, with zeros at the ends $(L/2)$ of the antenna (hence a "harmonic" oscillator), represented by:

$$\vec{I}(z')=I_0\sin\left(k\left(\frac{L}{2}-|z'|\right)\right)\vec{a}_z \tag{10.3.610}$$

where:

$$k=\frac{2\pi}{\lambda}\qquad \lambda=\frac{c}{f}\qquad |z'|\leq\frac{L}{2} \tag{10.3.611}$$

The results we will derive starting from now only in the situation where the antenna is short, i.e., $L\ll\lambda$. That assumption allow us to say that the current in the antenna is independent of position along the antenna, depending only on time!

We use the potential vector expression found earlier above for a sinusoidal regime in a spherical symmetry configuration to find the Z-component of the vector potential:

$$\vec{A}_z=\frac{\mu}{4\pi}\int\limits_{-L/2}^{+L/2}\frac{\vec{I}(z')e^{-\mathrm{i}k\|\vec{R}\|}\mathrm{d}z'}{\|\vec{R}\|} \tag{10.3.612}$$

where:

$$\vec{R}=\vec{r}-z'\vec{a}_z \tag{10.3.613}$$

and the norm of $\vec{R}$ can be approximated as (the last row is a Maclaurin series):

$$\begin{aligned}\|\vec{R}\|&=\sqrt{(\vec{r}-z'\vec{a}_z)\circ(\vec{r}-z'\vec{a}_z)}=\sqrt{\vec{r}^2+(z'\vec{a}_z)^2-2\vec{r}\circ(z'\vec{a}_z)}\\&=r\sqrt{1-\frac{2\vec{r}\circ(z'\vec{a}_z)}{r}-\left(\frac{z'a_z}{r}\right)^2}\cong r\sqrt{1-\frac{2\vec{r}\circ(z'\vec{a}_z)}{r}}\\&\cong r\left(1-\frac{\vec{r}\circ(z'\vec{a}_z)}{r}\right)=r-z'\cos(\theta)\end{aligned} \tag{10.3.614}$$

Keep in mind that θ is the angle between the radius vector of the point of interest and the longitudinal axes of the antenna!

Substituting yields:

$$A_z\cong\mu\int\limits_{-L/2}^{+L/2}\frac{I_0\sin\left(k\left(\frac{L}{2}-|z'|\right)\right)e^{-\mathrm{i}k(r-z'\cos(\theta))}}{4\pi(r-z'\cos(\theta))}\mathrm{d}z' \tag{10.3.615}$$

As $r\gg z'\cos(\theta)$, the term $z'\cos(\theta)$ can be neglected in the denominator. However, it cannot be neglected in the exponential as it is a phase offset:

$$A_z\cong\frac{\mu I_0}{4\pi}\int\limits_{-L/2}^{+L/2}\frac{\sin\left(k\left(\dfrac{L}{2}-|z'|\right)\right)e^{-\mathrm{i}kr}e^{\mathrm{i}kz'\cos(\theta)}}{r}\mathrm{d}z' \tag{10.3.616}$$

Rearranging yields:

$$A_z\cong\frac{\mu I_0e^{-\mathrm{i}kr}}{4\pi r}\int\limits_{-L/2}^{+L/2}\sin\left(k\left(\frac{L}{2}-|z'|\right)\right)e^{\mathrm{i}kz'\cos(\theta)}\mathrm{d}z' \tag{10.3.617}$$

We eliminate the absolute value by splitting into two integrals:

$$A_z\cong\frac{\mu I_0e^{-\mathrm{i}kr}}{4\pi r}\left[\int\limits_{-L/2}^{0}\sin\left(k\left(\frac{L}{2}+z'\right)\right)e^{\mathrm{i}kz'\cos(\theta)}\mathrm{d}z'+\int\limits_{0}^{+L/2}\sin\left(k\left(\frac{L}{2}-z'\right)\right)e^{\mathrm{i}kz'\cos(\theta)}\mathrm{d}z'\right] \tag{10.3.618}$$

Flip limits on the first integral and combine:

$$A_z \cong \frac{\mu I_0 e^{-\mathrm{i}kr}}{4\pi r}\int\limits_0^{+L/2} \sin\left(k\left(\frac{L}{2}-z'\right)\right)\left(e^{-\mathrm{i}kz'\cos(\theta)}+e^{\mathrm{i}kz'\cos(\theta)}\right)\mathrm{d}z' \tag{10.3.619}$$

Substitute $e^{\mathrm{i}x}+e^{-\mathrm{i}x}=2\cos(x)$:

$$A_z \cong \frac{\mu I_0 e^{-\mathrm{i}kr}}{2\pi r}\int\limits_0^{+L/2} \sin\left(k\left(\frac{L}{2}-z'\right)\right)\cos(kz'\cos(\theta))\mathrm{d}z' \tag{10.3.620}$$

Now, integrate it, setting aside the constants and limits temporarily:

$$\alpha = \int \sin\left(k\left(\frac{L}{2}-z'\right)\right)\cos(kz'\cos(\theta))\mathrm{d}z' \tag{10.3.621}$$

Integrate by parts:

$$\int u\mathrm{d}v = uv - \int v\mathrm{d}u \tag{10.3.622}$$

with:

$$\begin{aligned}
u &= \sin\left(k\left(\frac{L}{2}-z'\right)\right)\\
\mathrm{d}u &= -k\cos\left(k\left(\frac{L}{2}-z'\right)\right)\\
v &= \frac{\sin(kz'\cos(\theta))}{k\cos(\theta)}\\
\mathrm{d}v &= \cos(kz'\cos(\theta))
\end{aligned} \tag{10.3.623}$$

We have then:

$$\alpha = \sin\left(k\left(\frac{L}{2}-z'\right)\right)\frac{\sin(kz'\cos(\theta))}{k\cos(\theta)} + \int \cos\left(k\left(\frac{L}{2}-z'\right)\right)\frac{\sin(kz'\cos(\theta))}{\cos(\theta)}\mathrm{d}z' \tag{10.3.624}$$

And again integration by part for the second term above:

$$\begin{aligned}
u &= \cos\left(k\left(\frac{L}{2}-z'\right)\right)\\
\mathrm{d}u &= k\sin\left(k\left(\frac{L}{2}-z'\right)\right)\\
v &= -\frac{\cos(kz'\cos(\theta))}{k\cos^2(\theta)}\\
\mathrm{d}v &= \frac{\sin(kz'\cos(\theta))}{\cos(\theta)}
\end{aligned} \tag{10.3.625}$$

Let us simplify the result:

$$\begin{aligned}\alpha = &\overbrace{\sin\left(k\left(\frac{L}{2}-z'\right)\right)\frac{\sin(kz'\cos(\theta))}{k\cos(\theta)} - \cos\left(k\left(\frac{L}{2}-z'\right)\right)\frac{\cos(kz'\cos(\theta))}{k\cos^2(\theta)}}^{\beta} + \\ &\underbrace{\int \sin\left(k\left(\frac{L}{2}-z'\right)\right)\frac{\cos(kz'\cos(\theta))}{\cos^2(\theta)}\mathrm{d}z'}_{=\alpha/\cos^2(\theta)}\end{aligned} \tag{10.3.626}$$

Time for some algebra:

$$\begin{aligned}\alpha &= \int \sin\left(k\left(\frac{L}{2}-z'\right)\right)\cos(kz'\cos(\theta))\mathrm{d}z' \\ \beta &= \sin\left(k\left(\frac{L}{2}-z'\right)\right)\frac{\sin(kz'\cos(\theta))}{k\cos(\theta)} - \cos\left(k\left(\frac{L}{2}-z'\right)\right)\frac{\cos(kz'\cos(\theta))}{k\cos^2(\theta)} \\ \alpha &= \beta + \frac{\alpha}{\cos^2(\theta)} \\ \alpha\cos^2(\theta) &= \beta\cos^2(\theta) + \alpha \\ \alpha &= \beta\frac{\cos^2(\theta)}{\cos^2(\theta)-1}\end{aligned} \tag{10.3.627}$$

So:

$$\alpha = \left[\sin\left(k\left(\frac{L}{2}-z'\right)\right)\frac{\sin(kz'\cos(\theta))}{k\cos(\theta)} - \cos\left(k\left(\frac{L}{2}-z'\right)\right)\frac{\cos(kz'\cos(\theta))}{k\cos^2(\theta)}\right]\frac{\cos^2(\theta)}{\cos^2(\theta)-1} \tag{10.3.628}$$

Now substitute we substitute the famous trigonometric identity $\cos^2(\theta)-1=-\sin^2(\theta)$ to get:

$$\alpha = \frac{1}{k\sin^2\theta}\left[\cos\left(k\left(\frac{L}{2}-z'\right)\right)\cos(kz'\cos(\theta)) - \sin\left(k\left(\frac{L}{2}-z'\right)\right)\sin(kz'\cos(\theta))\cos(\theta)\right] \tag{10.3.629}$$

and plug in the limits $z'=0$ and $z'=L/2$ and bring back the constants:

$$\begin{aligned}A_z \cong \frac{\mu I_0 e^{-\mathrm{i}kr}}{2\pi rk\sin^2(\theta)}\Bigg\{\Bigg[&\cos\left(k\left(\frac{L}{2}-\frac{L}{2}\right)\right)\cos\left(k\frac{L}{2}\cos(\theta)\right) - \\ &\sin\left(k\left(\frac{L}{2}-\frac{L}{2}\right)\right)\sin\left(k\frac{L}{2}\cos(\theta)\right)\cos(\theta)\Bigg] - \Bigg[\cos\left(k\left(\frac{L}{2}-0\right)\right)\cos(k_0\cos(\theta)) - \\ &\sin\left(k\left(\frac{L}{2}-0\right)\right)\sin(k_0\cos(\theta))\cos(\theta)\Bigg]\Bigg\}\end{aligned} \tag{10.3.630}$$

After simplification we get:

$$A_z \cong \frac{\mu I_0 e^{-\mathrm{i}kr}}{2\pi rk\sin^2(\theta)}\left[\cos\left(\frac{kL}{2}\cos(\theta)\right) - \cos\left(\frac{kL}{2}\right)\right] \tag{10.3.631}$$

In the section of Vector Calculus (see page 1849) we have proved that the change of coordinates between rectangular and spherical vector components is given by:

$$\begin{pmatrix} A_r \\ A_\theta \\ A_\phi \end{pmatrix} = \begin{bmatrix} \sin(\theta)\cos(\phi) & \sin(\theta)\sin(\phi) & \cos(\theta) \\ \cos(\theta)\cos(\phi) & \cos(\theta)\sin(\phi) & -\sin(\theta) \\ -\sin(\phi) & \cos(\phi) & 0 \end{bmatrix} \begin{pmatrix} A_x \\ A_y \\ A_z \end{pmatrix} \tag{10.3.632}$$

As A_x and A_y are zero (vector potential has the same orientation as current!), then we have:

$$A_\theta = -A_z \sin(\theta) \tag{10.3.633}$$

Therefore:

$$A_\theta \cong -\frac{\mu I_0 e^{-\mathrm{i}kr}}{2\pi r k \sin(\theta)}\left[\cos\left(\frac{kL}{2}\cos(\theta)\right) - \cos\left(\frac{kL}{2}\right)\right] \tag{10.3.634}$$

Now we want to focus on the electric field only (as it is curiously the main point of interest of people scared of mobile antennas). For this purpose we know that:

$$\vec{\nabla} \times \vec{H} = \vec{j} + \mathrm{i}\varepsilon\omega\vec{E} \tag{10.3.635}$$

And in the region outside of where lies the electric current, this reduces to:

$$\vec{\nabla} \times \vec{H} = \mathrm{i}\varepsilon\omega\vec{E} \tag{10.3.636}$$

Hence:

$$\vec{E} = -\frac{\mathrm{i}}{\omega\varepsilon}\vec{\nabla} \times \vec{H} = -\frac{\mathrm{i}}{\omega\varepsilon}\vec{\nabla} \times \left(\frac{1}{\mu}\vec{\nabla} \times \vec{A}\right) = -\frac{\mathrm{i}}{\omega\mu\varepsilon}\vec{\nabla} \times \left(\vec{\nabla} \times \vec{A}\right) \tag{10.3.637}$$

As we have proved in the section of Vector Calculus (see 1898), the curl in spherical coordinates is given by:

$$\vec{\nabla} \times \vec{f} = \begin{pmatrix} \dfrac{1}{r\sin(\theta)}\dfrac{\partial}{\partial\theta}(\sin(\theta)f_\phi) - \dfrac{1}{r\sin(\theta)}\dfrac{\partial f_\theta}{\partial\phi} \\ \dfrac{1}{r\sin(\theta)}\dfrac{\partial f_r}{\partial\phi} - \dfrac{1}{r}\dfrac{\partial}{\partial r}(rf_\phi) \\ \dfrac{1}{r}\dfrac{\partial}{\partial r}(rf_\theta) - \dfrac{1}{r}\dfrac{\partial f_r}{\partial\theta} \end{pmatrix} \tag{10.3.638}$$

As in our case as $A_\phi = A_r = 0$ and A_θ doesn't depends on ϕ we get:

$$\vec{\nabla} \times \vec{A} = \begin{pmatrix} 0 \\ 0 \\ \dfrac{1}{r}\dfrac{\partial}{\partial r}(rA_\theta) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \dfrac{\mathrm{i}\mu I_0 e^{\mathrm{i}kr}}{2\pi r\sin(\theta)}\left[\cos\left(\dfrac{kL}{2}\cos(\theta)\right) - \cos\left(\dfrac{kL}{2}\right)\right] \end{pmatrix} \tag{10.3.639}$$

And again:

$$\vec{\nabla} \times (\vec{\nabla} \times \vec{A}) = \begin{pmatrix} 0 \\ 0 \\ \dfrac{1}{r}\dfrac{\partial}{\partial r}(r(\vec{\nabla} \times \vec{A})) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \dfrac{\mu I_0 k e^{-\mathrm{i}kr}}{2\pi r\sin(\theta)}\left[\cos\left(\dfrac{kL}{2}\cos(\theta)\right) - \cos\left(\dfrac{kL}{2}\right)\right] \end{pmatrix} \tag{10.3.640}$$

Therefore focusing only on the remaining non-null component:

$$\begin{aligned}E_\theta &\cong -\frac{\mathrm{i}}{\omega\mu\varepsilon}\vec{\nabla}\times\left(\vec{\nabla}\times\vec{A}\right)=-\frac{\mathrm{i}\mu I_0 k e^{-\mathrm{i}kr}}{2\pi\omega\mu\varepsilon r\sin(\theta)}\left[\cos\left(\frac{kL}{2}\cos(\theta)\right)-\cos\left(\frac{kL}{2}\right)\right]\\ &=-\frac{\mathrm{i}I_0 k e^{-\mathrm{i}kr}}{2\pi\omega\varepsilon r\sin(\theta)}\left[\cos\left(\frac{kL}{2}\cos(\theta)\right)-\cos\left(\frac{kL}{2}\right)\right]\end{aligned} \tag{10.3.641}$$

As $k=2\pi/\lambda$ and $\omega=2\pi f=2\pi c/\lambda$:

$$E_\theta\cong-\frac{\mathrm{i}I_0 k e^{-\mathrm{i}kr}}{2\pi\varepsilon c r\sin(\theta)}\left[\cos\left(\frac{kL}{2}\cos(\theta)\right)-\cos\left(\frac{kL}{2}\right)\right] \tag{10.3.642}$$

Therefore we get the "far-fields for a dipole antenna in harmonic regime of length L":

$$\boxed{E_\theta\cong\frac{-\mathrm{i}I_0 e^{-\mathrm{i}kr}}{2\pi\varepsilon c r\sin(\theta)}\left[\cos\left(\frac{kL}{2}\cos(\theta)\right)-\cos\left(\frac{kL}{2}\right)\right]} \tag{10.3.643}$$

This equation for E_θ (that should be denoted $\hat{E}_\theta$ rigorously!) is the general form for the θ component in spherical coordinates of the far-field $\vec{E}$ field of a dipole antenna of any length oriented along the z-axis. The r component is zero due to the far-field assumption and the ϕ component is zero due to the electric field's orientation along the z-axis!

For the Hertz dipole (or "Hertzian dipole"), the magnetic field has only ϕ-component and the electric field does not have the ϕ-component. The electric field lies in the (r,θ) plane.

The electrostatic field is inversely proportional to the frequency. As the frequency of the current approaches zero, this field diverges to infinity. This field is essentially due to the accumulation of charges on the tip of the antenna. When the current flows in the dipole, the opposite charges get accumulated on the tips of the antenna giving a dipole. With the reversal of the current (every half cycle) dipole reverses its polarity giving an oscillating dipole. The electrostatic field is due to this oscillating dipole. As the frequency decreases, the accumulated charge for a given current increases and therefore the electrostatic field increases.

Here is a MATLAB® script to reproduce the 2D electric field radiation pattern of the Hertz dipole antenna:

Listing 10.1: **MATLAB® code for 2D Hertz dipole antenna**

```
n=377;
Io=1;
r=10;
lambda=0.3;
k=(2*pi)/lambda;
L=lambda/2;

theta=0:0.01:2*pi;
E=j*n*Io*exp(-j*k*r)*(1/(2*pi*r))*((cos(k*L*cos(theta)/2)
-cos(k*L/2))./sin(theta));
polar(theta, abs(E))
```

Here are the three corresponding plots[8] for $L=\lambda/2$ (left), $L=3\lambda$ (right) and $L=6\lambda$ (bottom):

---

[8] Notice the interesting fact that the maximum of the electric field radiation is not always at $\pi/2$...!

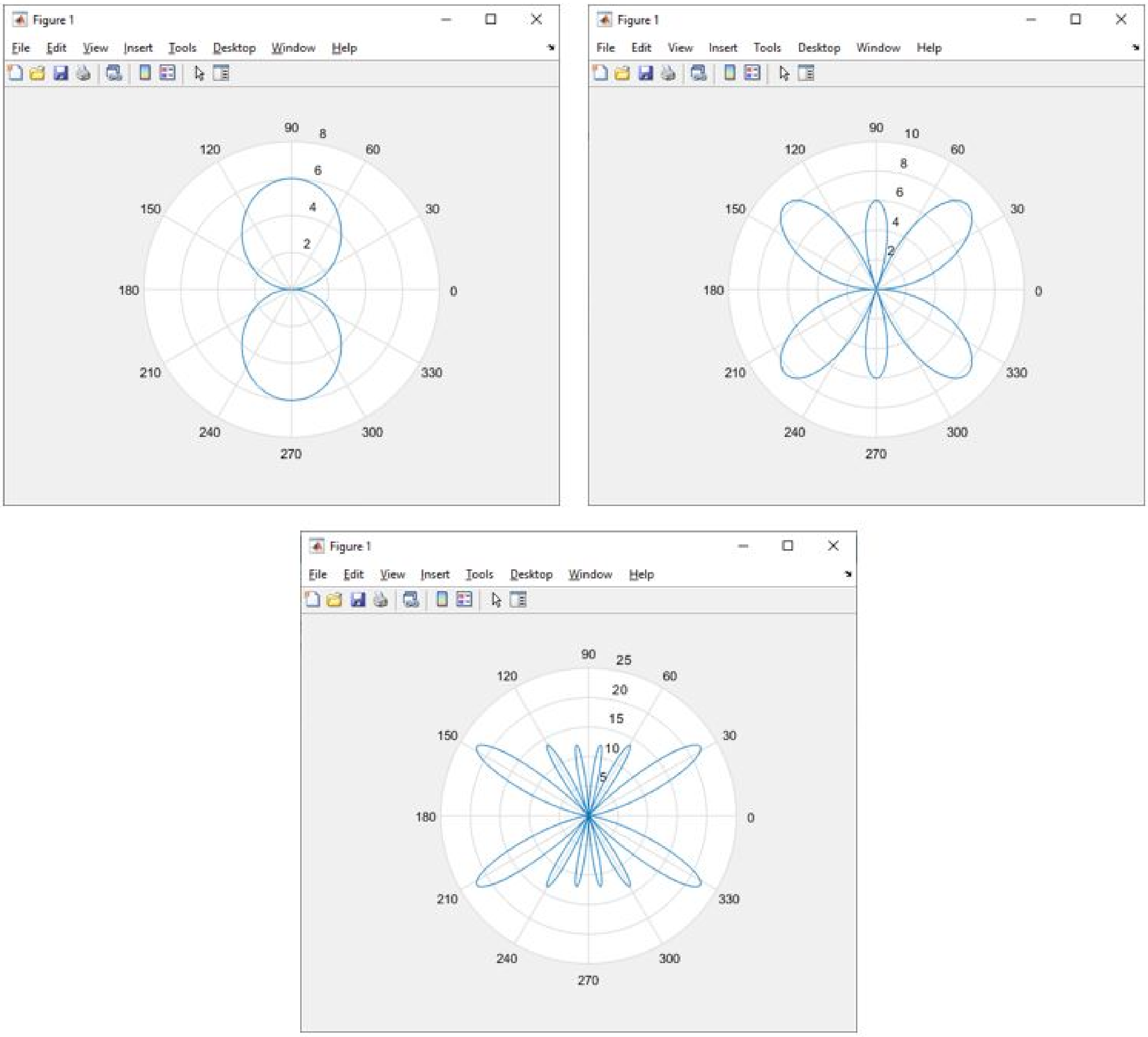


Figure 1
File Edit View Insert Tools Desktop Window Help
90 8 120 60 6 150 4 30 2 180 0 210 330 240 300 270
Figure 1
File Edit View Insert Tools Desktop Window Help
90 10 120 60 8 6 150 4 30 2 180 0 210 330 240 300 270
Figure 1
File Edit View Insert Tools Desktop Window Help
90 25 120 60 20 15 150 10 30 5 180 0 210 330 240 300 270

Figure 10.82: **Plot of some 2D Hertz dipole radiation pattern with MATLAB® 2019b**

Note that this picture is only a 2D slice of a 3D pattern. So here is the corresponding MATLAB® script to get a 3D view of the electric field radiation pattern:

Listing 10.2: **MATLAB® code for 3D Hertz dipole antenna**

```
theta=[0:0.1:2*pi];
phi=[0:0.1:2*pi];
n=377;
Io=1;
r=10;
lambda=20;
k=(2*pi)/lambda;
L=lambda*1.5;
E=j*n*Io*exp(-j*k*r)*(1/(2*pi*r))*((cos(k*L*cos(theta)/2)-cos(k*L/2
   ))./sin(theta));
udb=10*log10(abs(E));

%normalizing in order to make U vector positive
minu=min(udb);
u=udb-minu;
```

```

%expanding theta,phi,u to span entire space
u(1,1)=0;
for n=1:length(phi)
    theta(n,:)=theta(1,:);
end
phi=phi';
for m=1:length(phi)
    phi(:,m)=phi(:,1);
end
for k=1:length(u);
    u(k,:)=u(1,:);
end
[x,y,z]=sph2cart(phi,theta,u);
surf(x,y,z);
title('3D radiation pattern of dipole antenna');
```

Therefore we get for $L=\lambda/2$ (left) and $L=1.5\lambda$ (right):

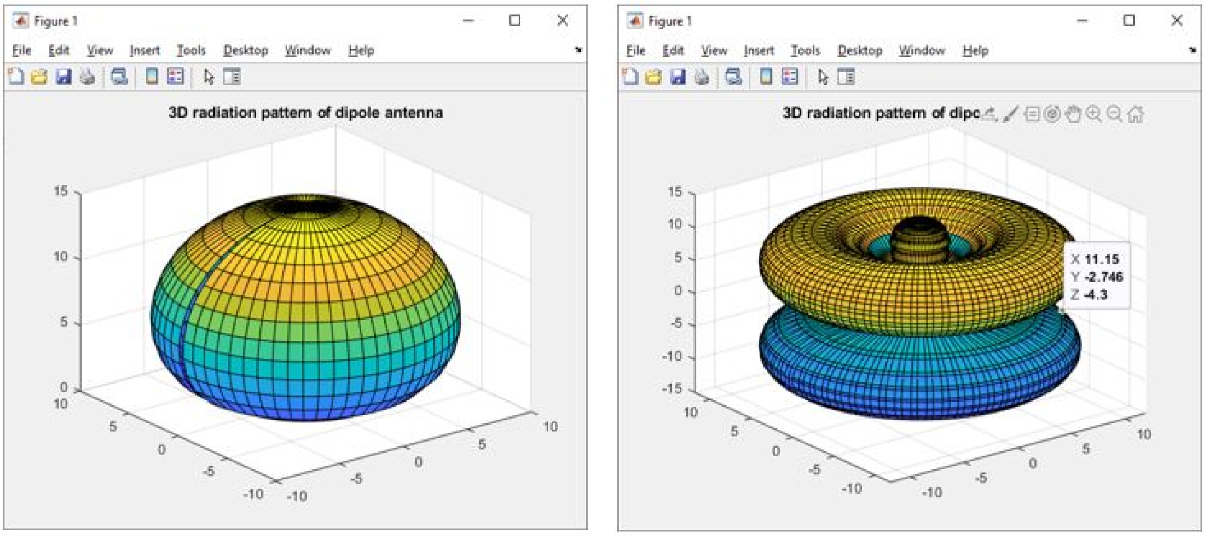


Figure 1
File Edit View Insert Tools Desktop Window Help
3D radiation pattern of dipole antenna
15
10
5
0
10
5
0
-5
-10
-10
-5
0
5
10
Figure 1
File Edit View Insert Tools Desktop Window Help
3D radiation pattern of dipo
15
10
5
0
-5
-10
-15
10
5
0
-5
-10
-10
-5
0
5
10
X 11.15
Y -2.746
Z -4.3

Figure 10.83: Plot of some 3D Hertz dipole radiation pattern with MATLAB® 2019b

10.3.6.1 The $L=\lambda/2$ (half-wave) far-field resonant antenna

As $k=2\pi/\lambda$ and $\omega=2\pi f=2\pi c/\lambda$ and that many textbooks like to take $L=\lambda/2$ (therefore $\cos(kL/2)=\cos(2\pi L/(2\cdot 2L)=0)$, and that also $\mu=1/(c^2\varepsilon)$, you can find the last relation:

$$E_\theta \cong \frac{-\mathrm{i}I_0e^{-ikr}}{2\pi\varepsilon cr\sin(\theta)}\left[\cos\left(\frac{kL}{2}\cos(\theta)\right)-\cos\left(\frac{kL}{2}\right)\right] \tag{10.3.644}$$

sometimes under the following form:

$$E_\theta \cong \frac{-\mathrm{i}I_0e^{-ikr}}{2\pi\varepsilon cr}\frac{\cos\left(\frac{\pi}{2}\cos(\theta)\right)}{\sin(\theta)} \tag{10.3.645}$$

Numerically it can be shown that:

$$\frac{\cos\left(\dfrac{\pi}{2}\cos(\theta)\right)}{\sin(\theta)} \cong \sin(\theta) \tag{10.3.646}$$

such that we fall back on a famous relation that we can find in many textbooks:

$$\boxed{E_\theta \cong \frac{-\mathrm{i}I_0 e^{-\mathrm{i}kr}}{2\pi\varepsilon cr}\sin(\theta)} \tag{10.3.647}$$

The radiating pattern of the $\lambda/2$ resonant far-field antenna with this last approximation looks very similar to a normal dipole. As we can illustrate it again with the previous MATLAB® script.

Let us calculate the total emitted power by integrating P/S on the surface of sphere of radius r. We have then:

$$\frac{P}{S} = \frac{\varepsilon c}{2}\hat{E}^2 = \frac{\varepsilon c}{2}\left(\frac{-\mathrm{i}I_0}{2\pi\varepsilon cr}\sin(\theta)\right)^2 = \frac{\varepsilon c}{2}\frac{I_0^2}{4\pi^2\varepsilon^2c^2r^2}\sin^2(\theta) \tag{10.3.648}$$

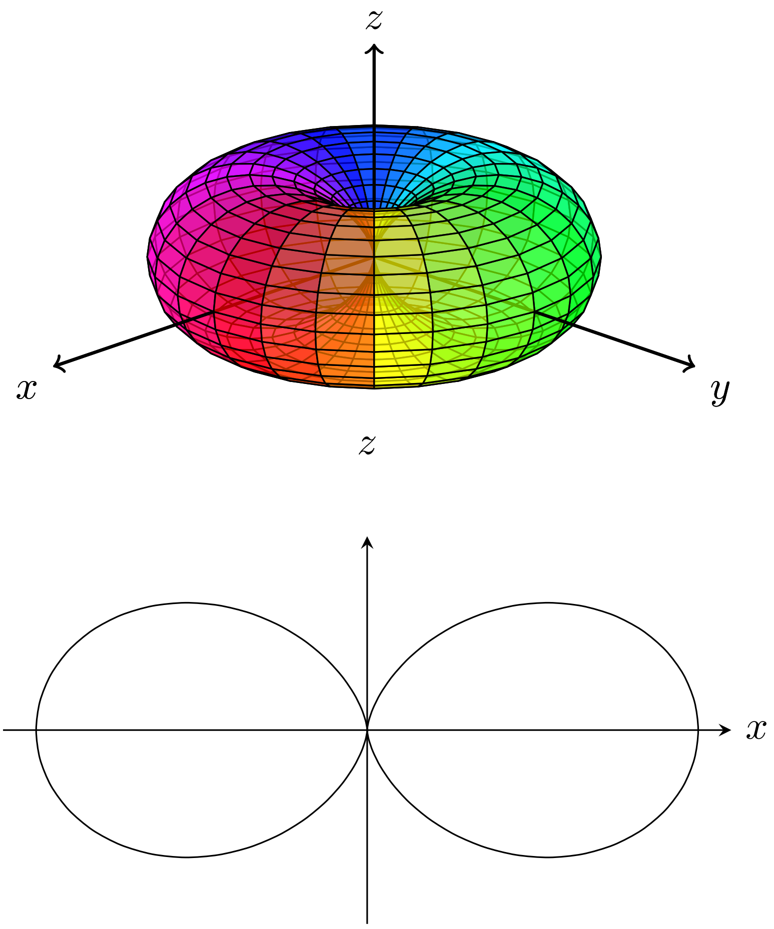


Figure 10.84: **Power radiation Hertzian dipole 2d and 3d**

And as we have already proved it, the surface element of the sphere is:

$$\mathrm{d}S=2\pi r^2\sin(\theta)\mathrm{d}\theta \tag{10.3.649}$$

Therefore the total emitted power is given by:

$$\begin{gathered}P\cong\oint_\Sigma\frac{P}{S}\mathrm{d}S=\int\limits_0^\pi\frac{\varepsilon c}{2}\frac{I_0^2\sin^2(\theta)}{4\pi^2\varepsilon^2c^2r^2}2\pi r^2\sin(\theta)\mathrm{d}\theta=\frac{I_0^2}{4\pi\varepsilon c}\int\limits_0^\pi\sin^3(\theta)\mathrm{d}\theta\\=\frac{I_0^2}{2\pi\varepsilon c}\int\limits_0^{\pi/2}\sin^3(\theta)\mathrm{d}\theta=\frac{I_0^2}{2\pi\varepsilon c}\frac{2}{3}=\frac{I_0^2}{3\pi\varepsilon c}\end{gathered} \tag{10.3.650}$$

Now let's see something for the fun. As we have proved it in the section of Electrical Engineering (see page 8451) we have:

$$\bar{p}=U_{\text{eff}}I_{\text{eff}}=\frac{U_0}{\sqrt{2}}\frac{I_0}{\sqrt{2}} \tag{10.3.651}$$

Therefore:

$$\bar{p}=RI_{\text{eff}}^2=R\left(\frac{I_0}{\sqrt{2}}\right)^2=\frac{RI_0^2}{2} \tag{10.3.652}$$

Therefore for our $L=\lambda/2$ far-field resonant antenna we get the value of a theoretical resistance that summarize the antenna Joule effect:

$$R\cong\frac{2\bar{p}}{I_0^2}=\frac{2\dfrac{I_0^2}{3\pi\varepsilon c}}{I_0^2}=\frac{2}{3\pi\varepsilon c}\cong 80\;[\Omega] \tag{10.3.653}$$

Remark

The natural electric field of the Earth refers to the planet Earth having a natural direct current (DC) electric field or potential gradient from the ground upwards to the ionosphere. The static fair-weather electric field in the atmosphere is $\sim 150\,[\mathrm{V}\cdot\mathrm{m}^{-1}]$ near the Earth's surface, but it drops exponentially with height to under $1\,[\mathrm{V}\cdot\mathrm{m}^{-1}]$ at 30 [km] altitude, as the conductivity of the atmosphere increases. Keep also in mind that at the Earth's surface, the energy surfacic density is approximately $1,000\,[\mathrm{W}\cdot\mathrm{m}^{-1}]$ for a surface perpendicular to the Sun's rays at sea level on a clear day.

Some readers may ask themselves if we could build antenna that emit in the range of the visible light 400 to 800 [THz]? The answer is that with traditional circuitry then answer is likely NO. At such high frequencies there are no good conductors so it would be very hard to drive such an antenna. We probably would need some waveguides rather than wires. And antenna size should be comparable to (or even slightly less than) wavelength, which further complicates things.

But such oscillators do exist. Consider LED diode, or still better - LASER. Both rely on electronic transitions. So we probably can't build oscillator in question, but we can use atoms, quantum wells and dots and many others acting as such oscillator.

Of the total power P_t (transmitted power) supplied to the antenna, a part, P_{rad}, is radiated out into space, and the remainder, P_{loss} is dissipated as heat in the antenna structure. The "radiation efficiency" ξ is defined as:

$$\xi=\frac{P_{\text{rad}}}{P_t} \tag{10.3.654}$$

The "gain of an antenna", in a given direction, is defined as the ratio between the power per unit area radiated in the chosen direction and the power per unit area that would radiate an isotropic (hypothetical) antenna that would radiate the same total power :

$$G=\frac{\left(\dfrac{P}{S}\right)_{\text{ant}}}{\left(\dfrac{P}{S}\right)_{\text{iso}}} \tag{10.3.655}$$

We put all these terms conditional because the isotropic antenna does not exist (or at least is quite difficult to develop at the day we write these lines). We can, if necessary, build a network of antennas whose radiation is almost isotropic!

When we talk about the gain of an antenna without specifying in which direction, it is the gain in the direction where it is maximum.

Let's calculate the gain of an elementary dipole of antenna note. The best direction is for $\theta=\pi/2$ when $L=\lambda/2$. Then:

$$\left(\frac{P}{S}\right)_{\text{ant}}=\frac{\varepsilon c}{2}\frac{I_0^2}{4\pi^2\varepsilon^2c^2r^2}\sin^2(\theta)=\frac{I_0^2}{8\pi^2\varepsilon c r^2} \tag{10.3.656}$$

And for the isotrope:

$$\left(\frac{P}{S}\right)_{\text{iso}}=\frac{\dfrac{RI_0^2}{2}}{4\pi r^2}=\frac{RI_0^2}{8\pi r^2} \tag{10.3.657}$$

Then:

$$G=\frac{\left(\dfrac{P}{S}\right)_{\text{ant}}}{\left(\dfrac{P}{S}\right)_{\text{iso}}}=\frac{\dfrac{I_0^2}{8\pi^2\varepsilon c r^2}}{\dfrac{RI_0^2}{8\pi r^2}}=\frac{1}{R\pi\varepsilon c} \tag{10.3.658}$$

But as:

$$R\cong\frac{2}{3\pi\varepsilon c}\cong 80\;[\Omega] \tag{10.3.659}$$

We then have:

$$G=\frac{1}{R\pi\varepsilon c}=\frac{3}{2}=1.5=1.76\;[\text{dB}] \tag{10.3.660}$$

Remark

For anyone wishing to deepen the subject of antennas, we strongly recommend the lecture of the master piece *Electromagnetic Waves and Antennas* ([428]) that also contains numerous MATLAB® scripts!

~70 %

★★★★☆

29 votes, 75.86%

10.4. Electrokinetics

THE DEVELOPMENT of electrodynamics has enabled a part of the humanity to significantly change its life quality. We know almost all today what we have thanks to it: fridge, radio, TV, computers, scanners, cars, trams, trains, planes, robots, smartphones, tablets, and other wonderful things and also sometimes less wonderful...

Before you start studying electrokinetics (engineers speak of "electronics" or "electrotechniques"), that is to say the structures involving electric charges in movement, we will define the two fundamental laws (the term "law" is a misnomer, as the first one was proved in the section Electrostatics and the second one in the section of Electrodynamics but anyway...) of the study of electrokinetic and the basic terminology/jargon of electrical circuits or equipment (practical cases being studied in the section Electrical Engineering at page 8437). Even if some elements in the beginning of this section will perhaps not immediately be understood by the reader (especially industrial applications), they will become trivial as the progress of reading of the section.

Definitions (#10.4.479):

D1. An electrical circuit is constituted by a set of devices named "dipole", interconnected by a conductive wire. If the dipole is a source of energy we speak of "active dipole" other in the opposite case its a "passive dipole".

D2. A "node" of a circuit is an interconnection where three or more conductor wire are connected together.

D3. A "branch" is a circuit section between two nodes.

D4. A "mesh" is a set of branches forming a closed loop[9].

D5. A "floating ground[10]" is a type of ground in which the ground doesn't have a physical connection to the earth; it simply serves as a type of 0 [V] reference line that serves as a return path for current back to the negative side of the power supply.

D6. An "earth ground" is when a circuit has a physical connection to the earth (ground), in order to sink electrons through to a ground connected to the earth for safety reasons.

---

[9] A mesh is a closed path which further can't be divided into any other closed path whereas loop is a closed path which can or can not be divided further into closed path a mesh is a loop but it is not necessary that a loop will be a mesh it will only be a mesh when further division is not possible.

[10] I think in automobile DC circuits, floating ground is also referred to as "chassis ground". The chassis is connected to the negative battery terminal and, thus, provides the infrastructure for closing the DC circuit.

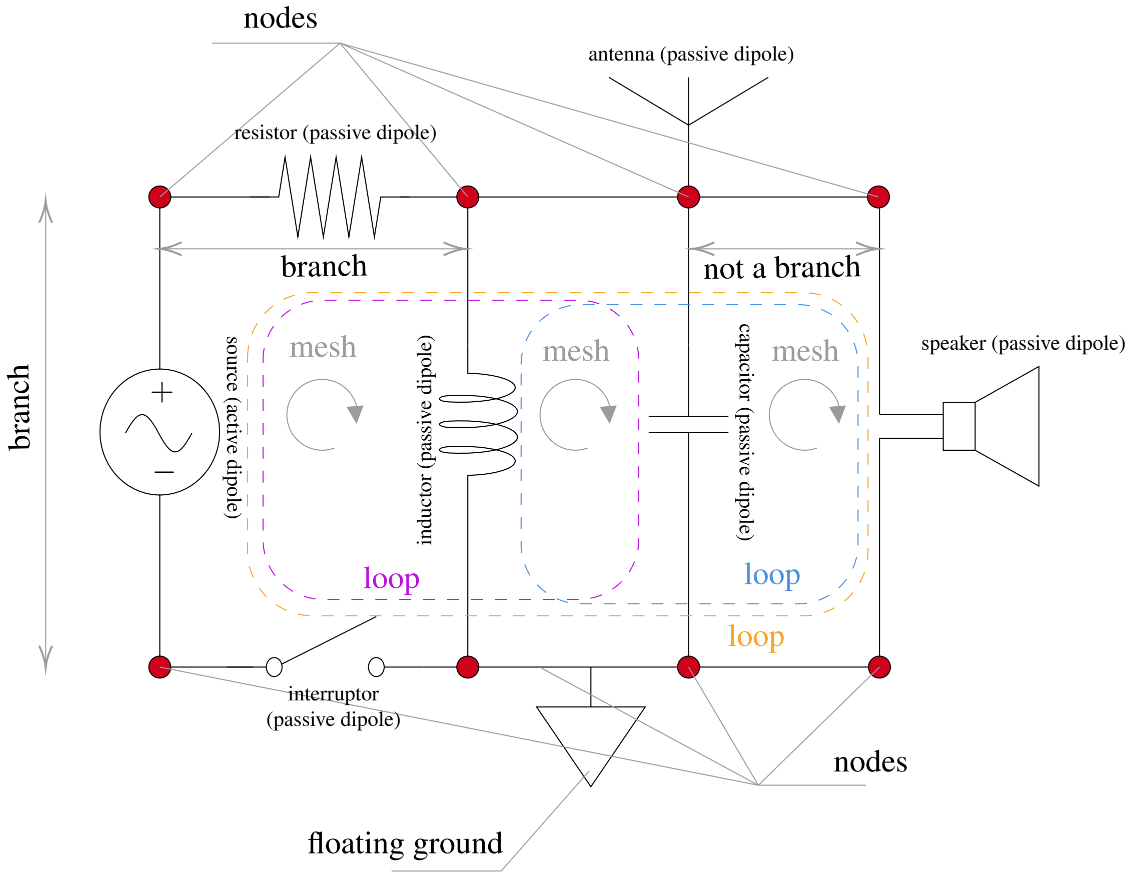

nodes
antenna (passive dipole)
resistor (passive dipole)
branch
not a branch
branch
source (active dipole)
mesh
inductor (passive dipole)
mesh
capacitor (passive dipole)
mesh
speaker (passive dipole)
loop
loop
loop
interruptor
(passive dipole)
nodes
floating ground

Figure 10.85: **Circuit basic vocabulary**

Remark

It is very important to understand what will follow! Some developments will be reused in the section of Special Relativity, Quantum Field Theory, etc. Furthermore, the reader should also read in parallel the section of Special Relativity to better understand the ins and outs of certain results and the provenance of some mathematical tools.

The dipole is characterized by the response in a current I by (most of time) a difference of electric potential U between its terminals. That is to say i.e. by the characteristic curve:

$$I = f(U) \tag{10.4.1}$$

Obviously some dipoles reacts on radiations difference, or on rotation difference, on pressure, on humidity and son on...

We shall see that in any conductor, the presence of a resistivity (see below the concept) causes a voltage drop and, strictly speaking, it is the same for the wires (made of conductor material). But these latter being connected in series with other dipoles, we neglect usually in small circuits the resistance of wires relatively to those of the dipoles. Therefore, the wires located between two dipoles of a small circuit are supposed to be equipotentials (the potential is the same on the two terminals).

10.4.1 Kirchhoff's laws

If you connect lots of passive or active elements together in a complicated network, then currents will flow through all the various elements so as to insure that charge is conserved, energy is conserved, and Ohm's Law is satisfied for each resistor.

Simultaneously satisfying all these conditions will give you exactly one solution. The method for writing down equations to represent these conservation laws is named "Kirchhoff's Laws" (not to be confused with those of the thermodynamics and optics bearing the same name!) and express the physical properties of the charge and the electric field and are at the number of 2 (one law for each property).

Briefly:

1. The total current flowing towards a node is equal to the total current flowing from that node
2. In a closed circuit, the algebraic sum of the products of the current and the resistance of each part of the circuit is equal to the resultant electromotive force in the circuit.

They will enable us without using the heavy mathematical artillery implicitly hidden behind just to get highly relevant results.

10.4.1.1 Mesh law (Kirchhoff's Loop Law)

The "mesh law" (implicitly it is simply the conservation of energy), also named "Kirchoff's loop law" or "Kirchoff's voltage law" (KVL), expresses the fact that when a charge browse a closed circuit (closed path), the energy it loses in by browsing a part of the circuit is equal to the energy it gains in the other. Thus, the algebraic sum of the potentials along a mesh is zero such that:

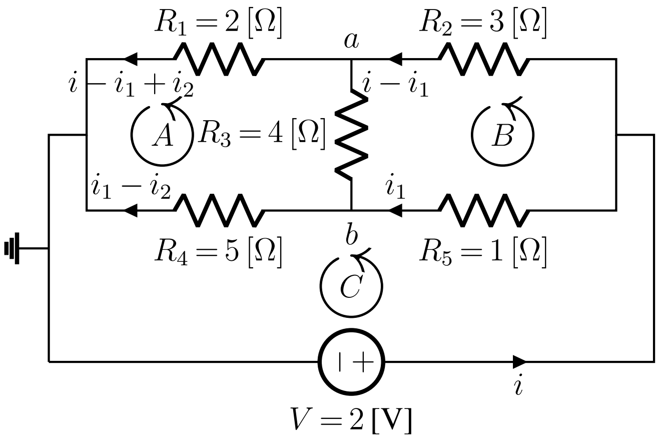
$$\boxed{\sum_{\text{mesh}} U_k = 0} \tag{10.4.2}$$

For this, we must arbitrarily choose a direction of travel of the mesh and agree that the tensions whose arrow points in the direction of travel are counted as positive and others as negative.

Remark

This law simply expresses the fact that the electric field (Coulomb) is a conservative vector field as we have seen in the section of Electrostatic.

In mesh analysis, Kirchhoff's voltage law $\sum V = 0$ (using the LaTeX Circuitikz package notation) is written for each independent loop, and a system of equation is solved. The number of independent loops can be determined by finding the minimum number of wire cuts needed such that there are no loops.



We have three independent loops, labelled as A, B, C, which gives the system of three equations:

$$0=-i_2R_3-(i-i_1+i_2)R_1+(i_1-i_2)R_4 \tag{10.4.3}$$

$$0=-(i-i_1)R_2+i_2R_3+i_1R_5 \tag{10.4.4}$$

$$0=V-i_1R_5-(i_1-i_2)R_4 \tag{10.4.5}$$

Solving this system gives:

$$(i,i_1,i_2)=\left(-\frac{144}{181},\frac{82}{181},\frac{26}{181}\right) \tag{10.4.6}$$

which can be used to determine the currents in all resistors and the associated voltages.

10.4.1.2 Nodes law (Kirchhoff's Point Law)

The "nodes law" (implicitly it is simply the current conservation law), also named "Kirchoff's point law" or "Kirchoff's current law" (KCL), expresses the conservation of charge, which means that the sum of currents leaving a node (a node can be seen as a separator of field lines - in extenso volumes connected by a same surface) is equal to the sum of currents entering the node. In other words, the algebraic sum of the currents is zero at every node of a circuit such that:

$$\boxed{\sum_{\text{node}} I_k=0} \tag{10.4.7}$$

This is also often denoted (if we denote by a the charges arriving in the node and by d those departing):

$$\sum_a\left(n_a\frac{\Delta q}{\Delta t}\right)=\sum_d\left(n_d\frac{\Delta q}{\Delta t}\right)\Leftrightarrow\sum_a i_a=\sum_d i_d \tag{10.4.8}$$

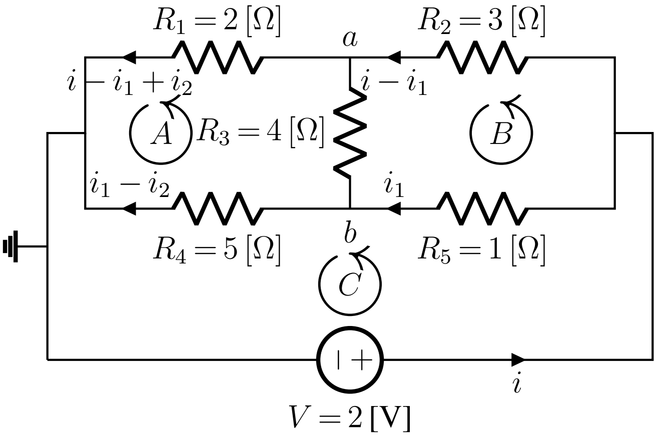
For this, we must choose a sign for incoming currents and the opposite sign for outgoing currents (as we do in Thermodynamics with the mass).

Remark

This law simply expresses the charge conservation equation (or equivalently the law continuity of the charge) that we have proved in the section Electrodynamics.

Figure 10.86: **Try to apply Kirchhoff's law...**

In nodal analysis, we deal with the potentials at each node and write out Kirchhoff's current law $\sum I = 0$ for the current leaving (or alternatively, entering) each node for each node where the potential is unknown. Let the nodes labelled a and b have potentials V_a and V_b in the figure below:



$R_1 = 2\,[\Omega]$
a
$R_2 = 3\,[\Omega]$
$i - i_1 + i_2$
$i - i_1$
A
$R_3 = 4\,[\Omega]$
B
$i_1 - i_2$
i_1
$R_4 = 5\,[\Omega]$
b
$R_5 = 1\,[\Omega]$
C
i
$V = 2\,[\text{V}]$

Then we have:

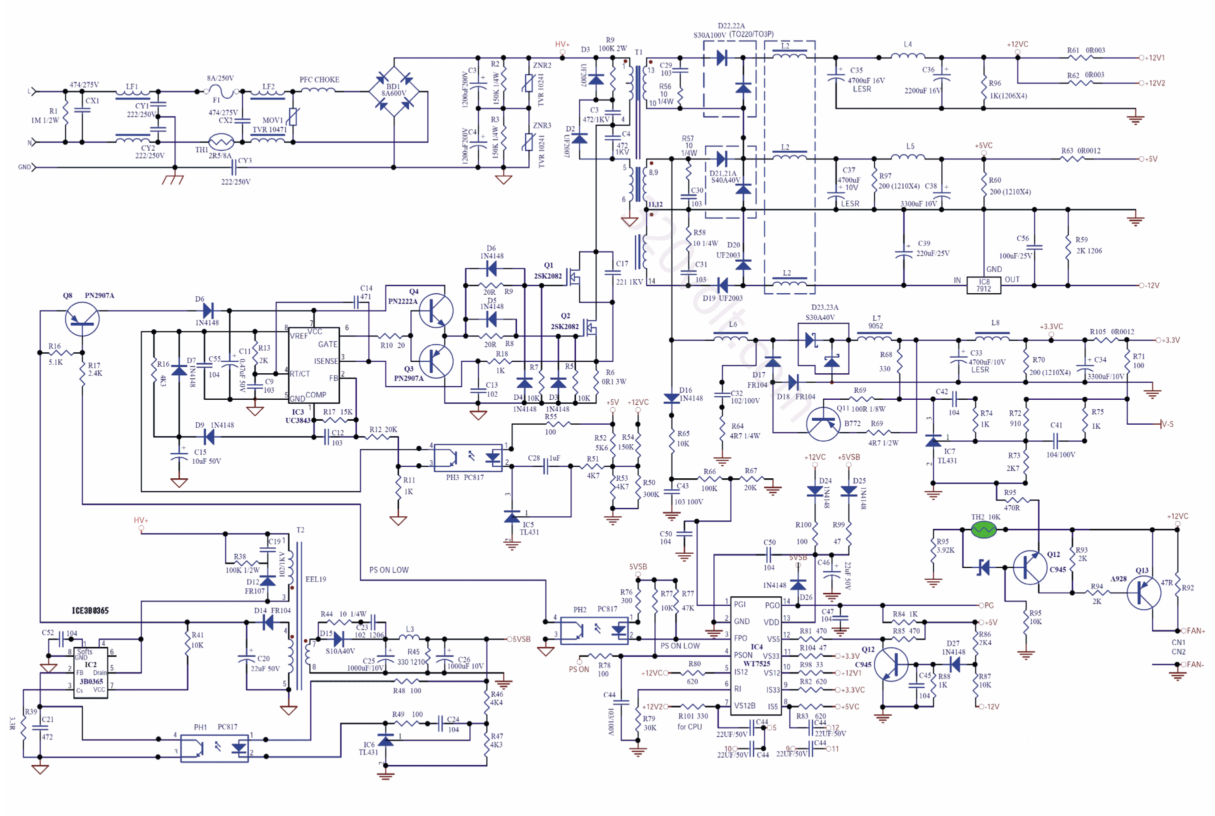
$$0 = \frac{V_a - V}{R_2} + \frac{V_a - V_b}{R_3} + \frac{V_a - 0}{R_1} \tag{10.4.9}$$

$$0=\frac{V_b-V}{R_5}+\frac{V_b-V_a}{R_3}+\frac{V_b-0}{R_4} \tag{10.4.10}$$

which after solving, gives:

$$(V_a,V_b)=\left(\frac{176}{181},\frac{280}{181}\right) \tag{10.4.11}$$

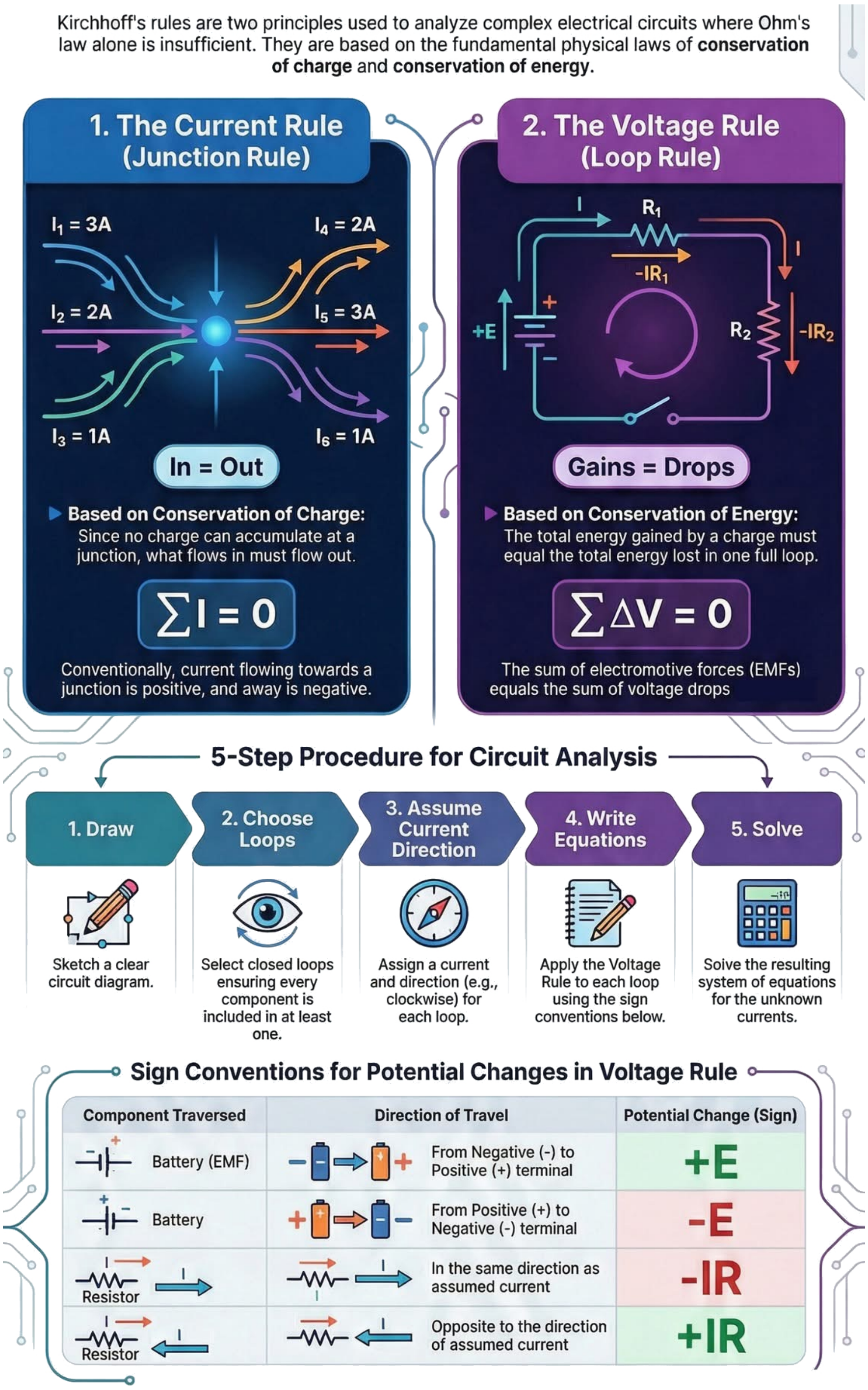
which can be easily confirmed via the mesh analysis done earlier above.



PFC CHOKE
BD1 8A/600V
F1 8A/250V
LF1
LF2
CX1 474/275V
CY1 222/250V
CY2 222/250V
CY3 222/250V
R1 1M 1/2W
MOV1 TVR 10471
TH1 2R5/8A
GND
HV+
T1
T2
D3
R9 100K 2W
D22,22A S30A100V (TO220/TO3P)
L1
L2
L3
L4
L5
L6
L7 9052
L8
C35 4700uF 16V LESR
C36 2200uF 16V
R96 1K(1206X4)
R61 0R003
R62 0R003
+12V1
+12V2
+12VC
+5VC
R63 0R0012
+5V
C37 4700uF 10V LESR
R97 200 (1210X4)
C38 3300uF 10V
R60 200 (1210X4)
C39 220uF/25V
C56 100uF/25V
R59 2K 1206
IC8 7912
IN
GND
OUT
-12V
D23,23A S30A40V
+3.3VC
R105 0R0012
+3.3V
C33 4700uF/10V LESR
R70 200 (1210X4)
C34 3300uF/10V
R71 100
Q8 PN2907A
Q4 PN2222A
Q3 PN2907A
Q1 2SK2082
Q2 2SK2082
IC3 UC3843
VREF
VCC
GATE
ISENSE
RT/CT
FB
GND
COMP
PH3 PC817
IC5 TL431
IC7 TL431
Q11 B772
TH2 10K
Q12 C945
Q13 A928
IC4 WT7525
PGI
PGO
VDD
FPO
VS5
PSON
VS33
IS12
VS12
RI
IS33
VS12B
IS5
PS ON LOW
PS ON
PH2 PC817
PH1 PC817
ICE3B0365
IC2 3B0365
EEL19
IC6 TL431
5VSB
+5V
+12VC
-12V
PG
CN1
CN2
FAN+
FAN-
V-S

Figure 10.87: Example of a real non-trivial electric circuit (source: `https://320volt.com`)

Circuits such as the above one can also be analysed with the KVL and KCL. It just take more than a day for an average human to do the whole analysis on a sheet of paper...



Kirchhoff's rules are two principles used to analyze complex electrical circuits where Ohm's law alone is insufficient. They are based on the fundamental physical laws of **conservation of charge** and **conservation of energy.**

1. The Current Rule (Junction Rule)

In = Out

Based on Conservation of Charge: Since no charge can accumulate at a junction, what flows in must flow out.

$$\sum I = 0$$

Conventionally, current flowing towards a junction is positive, and away is negative.

2. The Voltage Rule (Loop Rule)

Gains = Drops

Based on Conservation of Energy: The total energy gained by a charge must equal the total energy lost in one full loop.

$$\sum \Delta V = 0$$

The sum of electromotive forces (EMFs) equals the sum of voltage drops

5-Step Procedure for Circuit Analysis

1. **Draw** Sketch a clear circuit diagram.
2. **Choose Loops** Select closed loops ensuring every component is included in at least one.
3. **Assume Current Direction** Assign a current and direction (e.g., clockwise) for each loop.
4. **Write Equations** Apply the Voltage Rule to each loop using the sign conventions below.
5. **Solve** Soive the resulting system of equations for the unknown currents.

Sign Conventions for Potential Changes in Voltage Rule

Component Traversed	Direction of Travel	Potential Change (Sign)
Battery (EMF)	From Negative (-) to Positive (+) terminal	+E
Battery	From Positive (+) to Negative (-) terminal	-E
I Resistor I	In the same direction as assumed current	-IR
I Resistor I	Opposite to the direction of assumed current	+IR

Figure 10.88: **Kirchoff's rules summary**

10.4.2 Drude model

The Drude model of electrical conduction will allow us to introduce the basic concepts of the electrokinetic. First, let us define inf what will follow the concept of "Electric Current", "electric current density", and "electric resistance".

An electrical conductor (we are not talking about semiconductors and superconductors at this level of our discussion) can be seen in a very simplified way as a pipe section $\vec{S}$ containing an electron gas of n elementary electric charges q per unit volume.

In the absence of electric field, each electron has a zero vector average speed because it remains in the vicinity of the atom.

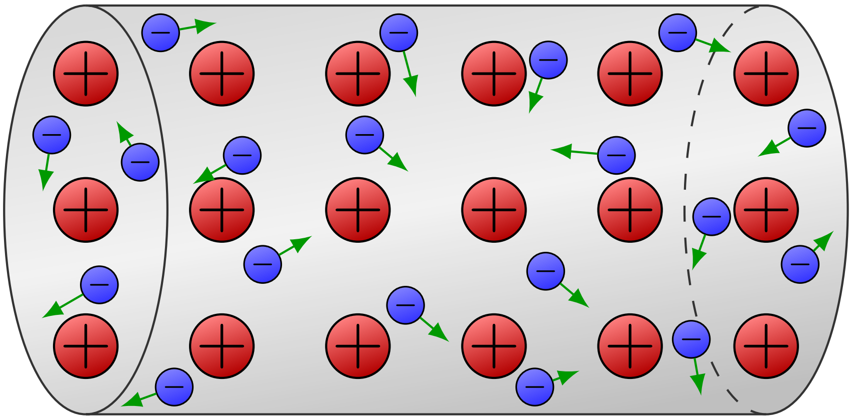


Figure 10.89: Typical conductor electrons behaviour without voltage (author: Izaak Neutelings)

Under the action of a constant and homogeneous electric field $\vec{E}$ (the case of direct current therefore!), some electrons are moved in a particular direction, until they collide with another atom (traditional aspect... not quantum one obviously!!!) where they take again an average vectorial zero speed drift and so on.

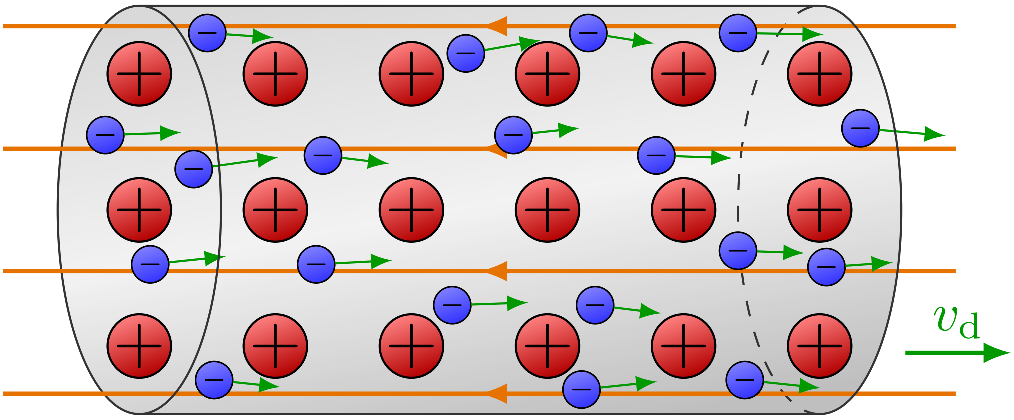


Figure 10.90: Typical conductor electrons behaviour with voltage (author: Izaak Neutelings)

This is the oldest model and the most basic one of the electric current. The bases were laid by Paul Drude in 11902 (holocene calendar), shortly after the discovery of the electron by Joseph John Thomson in 11897 (holocene calendar). Hence the name "Drude model".

Insufficient to conceive and develop most of the active electrical components that exists since the late 120th century (holocene calendar), the billiard balls model has nevertheless considerable interests:

- This is a useful tool to give to our limited mind a picture of phenomena we do not get any direct perception (at least at this day), since they take place in the infinitesimal world of atoms an elementary particles.
- The results, for the engineer, of more accurate modern theories, especially such as energy band theory, allow themselves to be formulated using the same concepts as those appearing in the "Billiard Balls" Drude model. Let us quote among them the "density number" and the "electron mobility" concepts (that we will introduce later rigorously further below in our study of energy band theory).
- Even if this approach is quite primitive, this model leads to a phenomenological interesting interpretation of the fundamental laws such as Ohm's law or the Joule's law. It binds together certain microscopic phenomena and observable quantities.

As it's friendly name suggests it, this model treats the electrons as tiny billiard balls. These particles are therefore considered as classical objects, simply governed by Newton's and Maxwell's laws proved in previous sections of this textbook. This corpuscular conception of the electron is also not totally opposed to the results of quantum physics (study in the next sections of this textbook), in which a wave packet can always be interpreted as a particle with its mass and speed (see the Ehrenfest's theorem in the section of Wave Quantum Physics at page 4870).

In a millimetre copper cube, we assume that the number of electrons is so high that it therefore does not matter then treat individually, which would also be irrelevant. It is the average behaviour of electrons that should be studied! Two types of interactions that determine behaviour are:

- The interaction of electrons with the material in which they operate, and to which they belong;
- The interaction of electrons with the electromagnetic field applied from the outside (all other interaction being neglected).

The distance λ travelled by an electron distance is named "electron mean free conduction path" and if τ is the time interval between two successive collisions then we have trivially:

$$\lambda = v\tau \qquad (10.4.12)$$

where v is obviously the mean electron velocity of the material.

The collision time is a random variable. All physical parameters maintained as constant, this random variable is stationary, its average value is named "mean collision time".

We suppose that:

$$v = a\tau \qquad (10.4.13)$$

the mean velocity, is therefore generated by the acceleration of the electric field (see section Electrostatic page 3969):

$$a = \frac{F}{m} = \frac{qE}{m} \tag{10.4.14}$$

We then get the "mean drift velocity" or simply electrons "drift velocity" given by:

$$v = \frac{qE}{m}\tau \tag{10.4.15}$$

This relation is so named because their initial speed is maintained due to the thermal excitation of the external environment and corresponds to the thermal velocity for which we have determined the expression in our study of the Maxwell-Boltzmann distribution in the section of Statistical Mechanics (we will calculate their further below in this section).

We admit, in the context of the billiard balls model, that electrons behave like atoms of an ideal gas. This is a gross approximation but enough satisfying for now!

The mean velocity is assumed to be identical for all the free electrons when a constant homogeneous electric field is applied, stationary, and directed along a single axis. It gives the possibility to define the "current intensity I" of electric current in the conductor.

Definition (#10.4.480): The "electric current" or "electric intensity", denoted by I, measures the charge $\mathrm{d}Q = nq$ that passes through the cross section S of a conductor per unit of time $\mathrm{d}t$ and is given according to what has been shown just before by:

$$\boxed{I = \frac{\mathrm{d}Q}{\mathrm{d}t}} \tag{10.4.16}$$

A slice of conductor, of volume $\mathrm{d}V = S\mathrm{d}L$ contains therefore the electric charge:

$$\mathrm{d}Q = \frac{nq}{V}\mathrm{d}V = \rho_q S\mathrm{d}L \tag{10.4.17}$$

It passes through the section S in a time $\mathrm{d}t$, such as:

$$\mathrm{d}L = v\mathrm{d}t \tag{10.4.18}$$

The current is therefore written:

$$I = \frac{\mathrm{d}Q}{\mathrm{d}t} = \rho_q S v \tag{10.4.19}$$

Remark

Attention to the usage of the latter relation in practice! If we consider a theoretical wire, the potential difference across the wire is zero. So there will be no gain / loss of kinetic energy of the electron and therefore no change in speed. If we now consider a real wire, resistive therefore, the potential difference across its terminals will be low but not zero, and the electron's potential drop will not be won in kinetic energy but dissipated as heat in the wire.

If I is seen as the flow of a "current density J" through the surface S, then we have:

$$I=\oint\limits_S J\mathrm{d}S=JS \tag{10.4.20}$$

the current density being assumed constant at each point of the surface.

We have therefore:

$$I=\rho_q Sv=JS \tag{10.4.21}$$

and after simplification:

$$J=\rho_q v \tag{10.4.22}$$

which is therefore the expression of the "current density" in the conductor.

As we know the expression of the velocity, this lead us to the famous "drift current" relation:

$$J=\frac{\rho_q q\tau}{m}E \tag{10.4.23}$$

Remark

That latter relation is often denoted:

$$J_{\text{drift}}=q\rho\mu E \tag{10.4.24}$$

where $\rho=\rho_q/m$ is the density of charges (for recall ρ_q was the massic density of charges) and μ is another notation of the mobility τ. Or this relation is also sometimes denoted:

$$J_{\text{drift}}=qn\mu E \tag{10.4.25}$$

And we define the "conductivity" by:

$$\sigma=\frac{\rho_q q\tau}{m}=\frac{nq^2\tau}{m}\;[\mathrm{kg}^{-1}\cdot\mathrm{m}^{-3}\cdot\mathrm{s}^3\cdot\mathrm{A}^2] \tag{10.4.26}$$

where this time n is not the number of electrons, but the number density of electrons! Notice that $[\mathrm{kg}^{-1}\cdot\mathrm{m}^{-3}\cdot\mathrm{s}^3\cdot\mathrm{A}^2]$ is the equivalent of $[\mathrm{kg}^{-1}\cdot\mathrm{m}^{-3}\cdot\mathrm{s}\cdot\mathrm{C}^2]$. Conductivity has in SI units (sadly) of "siemens" per metre $[\mathrm{S}\cdot\mathrm{m}^{-1}]$.

By definition, the "resistivity" (not to be confused with the resistance denoted R!!), is the inverse of the conductivity:

$$\boxed{\rho=\frac{1}{\sigma}=\frac{m}{\rho_q q\tau}\;[\mathrm{kg}\cdot\mathrm{m}^3\cdot\mathrm{s}^{-3}\cdot\mathrm{A}^{-2}]} \tag{10.4.27}$$

More commonly denoted:

$$\rho=\frac{1}{\sigma}\;[\Omega\cdot\mathrm{m}] \tag{10.4.28}$$

where Ω is the unit named "Ohm".

We notice that the conductivity contains the product of the number density of electron by their mobility. Therefore it is necessary that at least one of these variables has a high value for a material has a high conductivity.

The mobility is greater in the semiconductors than in metals. This characteristic however is completely masked by the ratio of the volumetric numbers of electrons: n is between $1,000,000$ to $100,000,000$ times lower in semiconductors than in metals, which explains the higher conductivity of these.

Following the relation:

$$\sigma=\frac{\rho_q q\tau}{m}=\frac{nq^2\tau}{m} \tag{10.4.29}$$

proved just earlier above, the conductivity depends on the electric field, through the collision time. Indeed, the more the electric field grows, the more the speed of the electrons increases. The distance between the points of possible shocks remaining the same, the collision time, and hence the conductivity should decrease (and thus the resistance increase!).

However, the independence of the conductivity (and respectively the resistance) with the electric field is an experimental fact established precisely with all usual conductors in normal civilian uses.

The origin of this contradiction lies in the large difference in the magnitudes of the thermal velocity given by the Maxwell-Boltzmann distribution (see section Statistical Mechanics page 3554):

$$\langle v_{\text{th}}\rangle=\sqrt{\langle v^2\rangle}=\sqrt{\frac{3kT}{m}} \tag{10.4.30}$$

and of the average speed drift seen above:

$$v=\frac{qE}{m}\tau \tag{10.4.31}$$

with the mean free path time that will be obtained using the expression:

$$\sigma=\frac{\rho_q q\tau}{m}=\frac{nq^2\tau}{m} \tag{10.4.32}$$

We have proved in the section of Statistical Mechanics that for an electron at room temperature:

$$\langle v_{\text{th}}\rangle\cong 1.2\cdot 10^5\;[\text{m}\cdot\text{s}^{-1}] \tag{10.4.33}$$

And let us calculate the drift velocity for copper with for this particular metal the following values:

$$\sigma=5.81\cdot 10^7\;[\Omega^{-1}]\qquad\text{and}\qquad n=1.16\cdot 10^{29}\;[\text{m}^{-3}] \tag{10.4.34}$$

This allows us to get the value:

$$\tau=\frac{\sigma m}{nq^2}=1.79\cdot 10^{-14}\;[\text{s}] \tag{10.4.35}$$

and therefore:

$$v=\frac{qE}{m}\tau=\frac{qE}{m}\frac{\sigma m}{nq^2}=\frac{qE}{m}\cdot(1.79\cdot 10^{-14}) \qquad (10.4.36)$$

Taking $E=0.32\,[\text{V}\cdot\text{m}^{-1}]$, which is considered as a high value since this field produces a current density of:

$$J=\frac{nq^2\tau}{m}E=1.86\cdot 10^7\,[\text{A}\cdot\text{m}^{-2}]=18.6\,[\text{A}\cdot\text{mm}^{-2}] \qquad (10.4.37)$$

we finally have:

$$v=\frac{qE}{m}\cdot(1.79\cdot 10^{-14})=1\cdot 10^{-3}\,[\text{m}\cdot\text{s}^{-1}] \qquad (10.4.38)$$

Therefore, even in a strong industrial electric field, the drift velocity is negligible compared to the thermal velocity!

As the thermal velocity depends only a little bit of the electric field, it turns out that in practice the electron velocity is independent of the electric field. In other words, establishing a current, even intense, has only a negligible effect on the speed of electrons!

It is therefore not the electrons that carry the energy, contrary to a falsely held belief, due to bad analogies (which are always bad equivalences as we already know...!) - typically the analogy of electrons that behaves like water flow in cables (good luck to explain with water flow basic AC circuits with capacitance, inductance or semiconductors) - as taught in the non-graduate degrees of the school education system!!!!

Remark

In the vast majority of cases, the conductor sizes are large, compared to the average distance that an electron travel between two consecutive shocks. The behaviour of the surface of the conductive electrons then are of secondary importance. This is why the conductive medium is often implicitly considered as infinite. The FET (Field-Effect Transistor) and MOST (Metal Oxide Semiconductor Transistor) are an important exception to this. The current circulates in a layer thin enough that the electron mobility is affected by electron scattering on surfaces defining this layer.

However, an important point to notice is the calculation of the mean free path of electrons in the classical Drude model. We have indeed:

$$l\cong v_{\text{th.}}\tau\cong 2\,[\text{nm}] \qquad (10.4.39)$$

which is much higher, at least an order of magnitude (factor of 10), to the inter-atomic distances. It results from this that successive collisions with atoms of the network is not responsible for Ohm's law (which we will see now) contrary to one of the initial assumptions of the Drude model but that it are the impurities and defects of the material involved in it that generates the collisions! We will see a further below in the theoretical model of energy bands that the mean free path is in fact still much greater!

Warning!!! This relation may suggest that since the mean free path is proportional to the thermal velocity and therefore proportional to the square root of the temperature, that the resistance decreases with temperature. But in fact it is not so! The Drude model is too simplistic because in reality it is the opposite that occurs for conductors (resistance increases with temperature because the time interval τ between collisions decreases faster than the speed increases). And then there is also an opposite problem ... almost at a temperature equal to the zero Kelvin (and over for some materials) the mean free path should be almost zero but superconductors show us that it is not the case! In short, without explicit relation depending on the temperature we are in total darkness!

The only thing we know how to do is to admit that to a given constant factor α (positive or negative), a temperature change requires a relative change in resistance by (first order Taylor development):

$$\frac{\Delta\rho}{\rho_0}=\alpha\Delta T \qquad (10.4.40)$$

hence:

$$\frac{\rho-\rho_0}{\rho_0}=\alpha(T-T_0) \qquad (10.4.41)$$

Then we get the relation know in high-schools (the same relation applies for resistance if the reader replaces ρ by R and ρ_0 by R_0):

$$\rho=\rho_0[1+\alpha(T-T_0)] \qquad (10.4.42)$$

Finally, let us notice that the fourth Maxwell equation (see section Electrodynamics page 4106) can then be written by the results obtained just previously:

$$\vec{\nabla}\times\vec{H}=\vec{j}+\frac{\mathrm{d}\vec{D}}{\mathrm{d}t}=\sigma\vec{E}+\frac{\mathrm{d}\vec{D}}{\mathrm{d}t} \qquad (10.4.43)$$

which then explicitly makes appear the conductivity coefficient.

10.4.3 Ohm's law

Ohm's law states that the current through a conductor between two points is directly proportional to the voltage across the two points. Introducing the constant of proportionality, the resistance.

From the relation proved just above:

$$J=\frac{\rho_q q\tau}{m}E \tag{10.4.44}$$

and taking the definition of "conductivity" by:

$$\sigma=\frac{\rho_q q\tau}{m}=\frac{nq^2\tau}{m} \tag{10.4.45}$$

We have finally:

$$\boxed{J=\sigma E} \tag{10.4.46}$$

which is the "local Ohm's law". We have already see it in differential form in the section of Statistical Mechanics and we already know therefore that it belongs in fact to the family of diffusion laws!

Type of Materials	**Electrical conductivity** σ $[\Omega\cdot\mathrm{m}^{-1}]$
Silver	$6.3\cdot 10^7$
Copper	$5.8\cdot 10^7$
Gold	$4.2\cdot 10^7$
Aluminum	$3.4\cdot 10^7$
Germanium	2.2
Silicon	$4.3\cdot 10^{-4}$
Polyethylene	10^{-14}
Diamond	10^{-14}

Table 10.4: **Table of some electrical conductivities at room temperature**

Remark

Since the conductivity is necessarily a scalar, the vector notation of Ohm's law implies that the electrostatic field lines also indicate the path taken by the electrical charges. Moreover, as the conductivity is a scalar necessarily positive in the traditional model, this implies that the current has the same direction as the electric field.

If we multiply the previous equality under scalar left and right by a length L we get:

$$J\cdot L=\sigma E\cdot L\Leftrightarrow I=\sigma\frac{S}{L}U=GU \tag{10.4.47}$$

Then we have:

$$I=G\cdot U \tag{10.4.48}$$

or:

$$U=\frac{1}{G}I \tag{10.4.49}$$

We define the inverse of conductivity as the "electric resistance", or more commonly simply named "resistance", defined by:

$$\boxed{R=\frac{1}{G}=\frac{1}{\sigma\dfrac{S}{L}}\;[\Omega]} \tag{10.4.50}$$

where in S.I. units $[\Omega]=[\mathrm{kg}\cdot\mathrm{m}^2\cdot\mathrm{s}^{-3}\cdot\mathrm{A}^{-2}]$.

Remark

It is important to notice that the electrical resistance is proportional to the length of the resistive element and inversely proportional to its sectional area. For example in high voltage cables, the resistance is given in ohms per kilometre, which permits then to calculates the power lost by kilometre and therefore the money lost by Joule loss.

Therefore, we can write Ohm's local law in its most commonly known form:

$$\boxed{U=RI} \tag{10.4.51}$$

whence (beware !!!) the potential U is the potential difference over the length of the resistive element (also named "resistive dipole") as we see it in the previous developments and not the total outside potential!

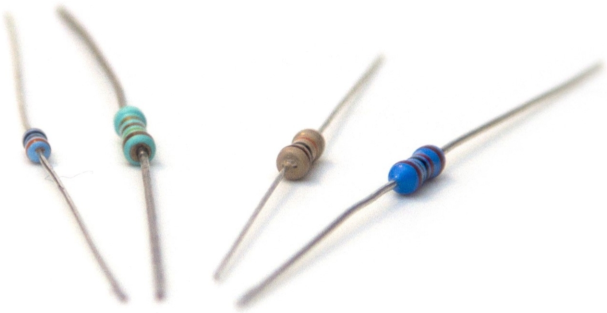


Figure 10.91: Some resistive dipoles (source: `http://www.e-style.ch`, author: Martin Bircher)

As for the capacitors seen in the section Electrostatics, resistors have also color codes that have as far as we know until know the same definitions:

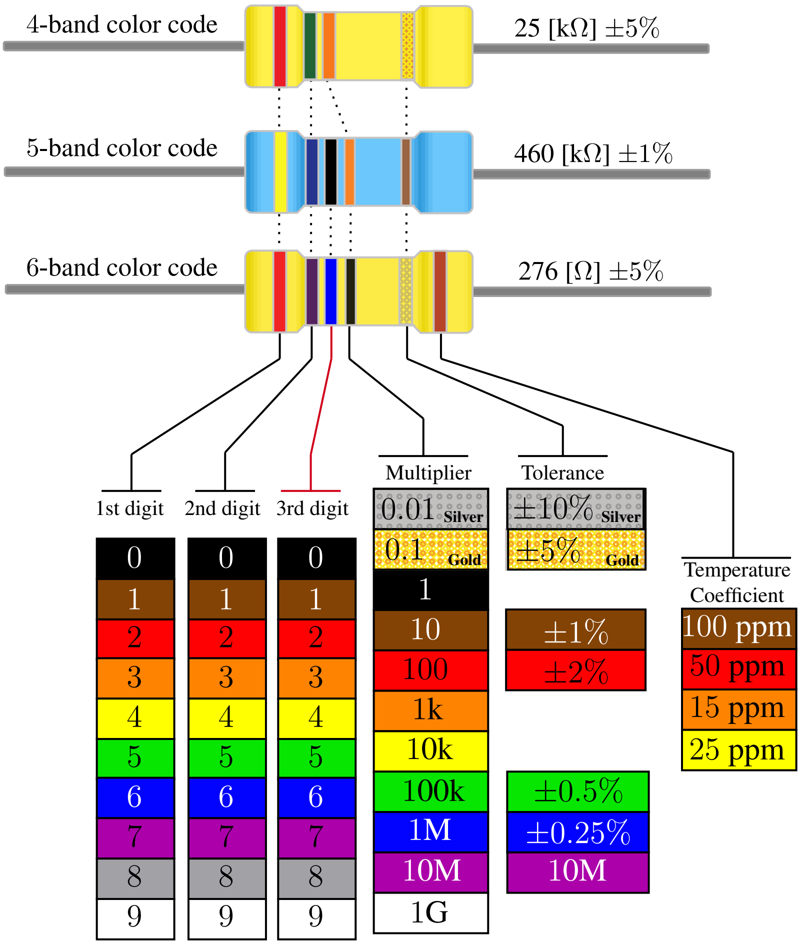

4-band color code
25 [kΩ] ±5%
5-band color code
460 [kΩ] ±1%
6-band color code
276 [Ω] ±5%
1st digit
2nd digit
3rd digit
Multiplier
Tolerance
Temperature Coefficient
0 1 2 3 4 5 6 7 8 9
0.01 Silver
0.1 Gold
1
10
100
1k
10k
100k
1M
10M
1G
±10% Silver
±5% Gold
±1%
±2%
±0.5%
±0.25%
10M
100 ppm
50 ppm
15 ppm
25 ppm

Figure 10.92: Resistor color codes

Remark

This relation is valid only for ideal conductors under normal conditions of temperature and pressure and for which the Drude model applies. So semiconductors and superconductors are excluded.

Since U is the potential of the resistive element, then we often do reference in the field of electrical engineering to the "voltage drop" (indeed, beyond the resistive element the potential is not the same that at the point above this same resistive element).

For copper cables in typical non-industrial use, there is a very useful American table in practice giving with a relatively good accuracy the resistivity according to the diameter and the maximum permissible current. Here is a sample of this table:

AWG	Wire ∅ in [mm] with isolation	Resistance in $[\Omega \cdot \text{m}^{-1}]$	Max. theoretically eligible current in [A]	Max. theoretically eligible current in [A] outdoors
1	7.35	0.0040	211	119
2	6.54	0.0051	181	94
...	...	...	...	...
12	2.05	0.00521	41	9.3
13	1.83	0.00657	35	7.4
14	1.63	0.00829	32	5.9
15	1.45	0.0104	28	4.7
16	1.29	0.0132	22	3.7
...	...	...	...	...

Table 10.5: AWG Codes (source: Wikipedia)

where AWG stands for "American Wire Gauge" and corresponds to a small gauge that can be buy easily on Internet to determine the diameter of a cable without having to use a calibrator:

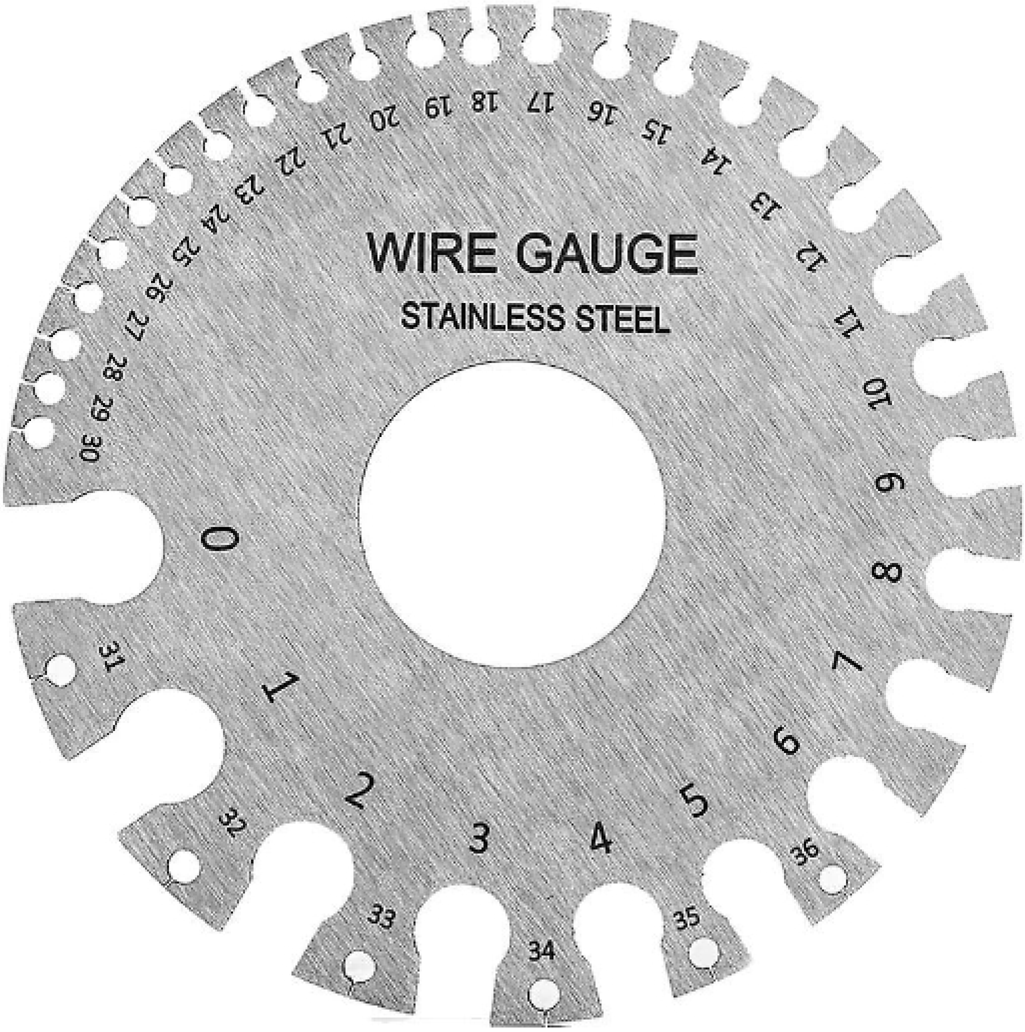
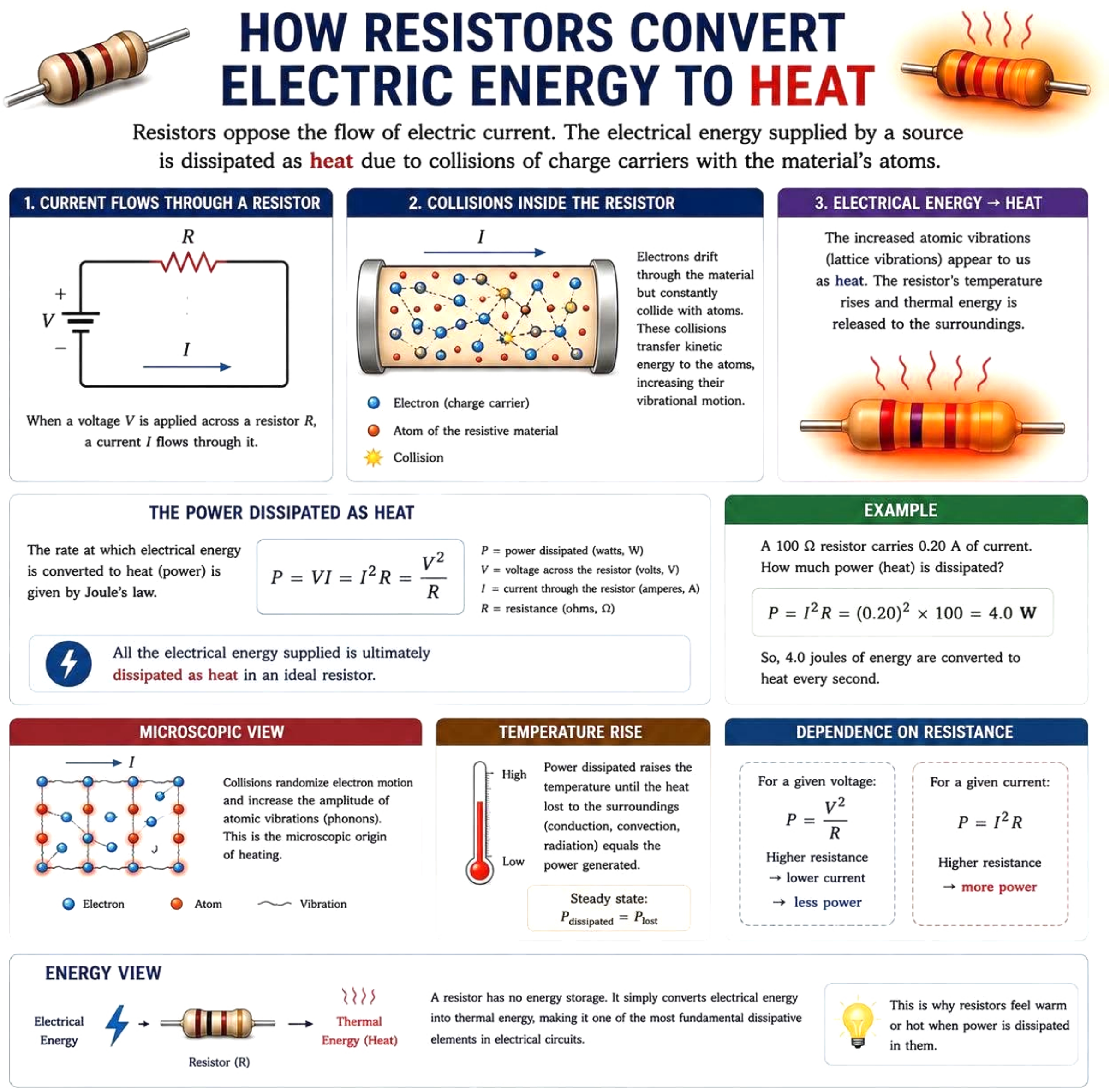

WIRE GAUGE
STAINLESS STEEL
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36

Figure 10.93: AWG Gauge (source: Wikipedia)



HOW RESISTORS CONVERT ELECTRIC ENERGY TO HEAT

Resistors oppose the flow of electric current. The electrical energy supplied by a source is dissipated as **heat** due to collisions of charge carriers with the material's atoms.

1. CURRENT FLOWS THROUGH A RESISTOR

When a voltage V is applied across a resistor R, a current I flows through it.

2. COLLISIONS INSIDE THE RESISTOR

Electrons drift through the material but constantly collide with atoms. These collisions transfer kinetic energy to the atoms, increasing their vibrational motion.

- ● Electron (charge carrier)
- ● Atom of the resistive material
- ✸ Collision

3. ELECTRICAL ENERGY → HEAT

The increased atomic vibrations (lattice vibrations) appear to us as **heat**. The resistor's temperature rises and thermal energy is released to the surroundings.

THE POWER DISSIPATED AS HEAT

The rate at which electrical energy is converted to heat (power) is given by Joule's law.

$$P = VI = I^2 R = \frac{V^2}{R}$$

P = power dissipated (watts, W)
V = voltage across the resistor (volts, V)
I = current through the resistor (amperes, A)
R = resistance (ohms, Ω)

All the electrical energy supplied is ultimately **dissipated as heat** in an ideal resistor.

EXAMPLE

A 100 Ω resistor carries 0.20 A of current. How much power (heat) is dissipated?

$$P = I^2 R = (0.20)^2 \times 100 = 4.0 \ \mathbf{W}$$

So, 4.0 joules of energy are converted to heat every second.

MICROSCOPIC VIEW

Collisions randomize electron motion and increase the amplitude of atomic vibrations (phonons). This is the microscopic origin of heating.

● Electron ● Atom ~ Vibration

TEMPERATURE RISE

Power dissipated raises the temperature until the heat lost to the surroundings (conduction, convection, radiation) equals the power generated.

Steady state:
$P_{\text{dissipated}} = P_{\text{lost}}$

DEPENDENCE ON RESISTANCE

For a given voltage:

$$P = \frac{V^2}{R}$$

Higher resistance
→ lower current
→ **less power**

For a given current:

$$P = I^2 R$$

Higher resistance
→ **more power**

ENERGY VIEW

A resistor has no energy storage. It simply converts electrical energy into thermal energy, making it one of the most fundamental dissipative elements in electrical circuits.

This is why resistors feel warm or hot when power is dissipated in them.

10.4.3.1 Complex Conductivity

The "complex conductivity" is a simple result of a simple extension of the classical Drude model for the material conduction. In the Drude model it is assumed that in the conduction material there is a concentration n of electrons and that an electron has a probability $1/\tau$, where τ is named "scattering time" or "relaxation time", for scattering with the lattice for losing its momentum. The motion of the electrons can be described by Newton's law:

$$m\frac{\mathrm{d}v}{\mathrm{d}t}=-qE-\frac{mv}{\tau} \tag{10.4.52}$$

where m is the (effective) electron mass, v the mean electron velocity in field direction, and E is the electric field strength.

This relation is easy to derived. Indeed, consider the motion of an electron in an electric field of strength $E\,[\mathrm{V}\cdot\mathrm{m}^{-1}]$. The force on the electron is $-qE$. In the event of a collision between the electron and a gas molecule, energy is lost in the form of heat. The frictional or retarding force due to the collision is mv/τ, or $mv\nu$, where τ is the mean time between collision, ν is the frequency of collision and m is the effective mass of electron. Therefore, the net force on the electron is:

$$F=-qE-\frac{mv}{\tau} \tag{10.4.53}$$

Since the electron is moving with a velocity v, the force on the electron is equal to the rate of change of the momentum:

$$F=-qE-\frac{mv}{\tau}=m\frac{\mathrm{d}v}{\mathrm{d}t} \tag{10.4.54}$$

The term $-mv/\tau$ resembles a "friction term" in the equation of motion. Assuming sinusoid variation of the electric field, i.e.:

$$v\propto E\propto e^{-\mathrm{i}\omega t} \tag{10.4.55}$$

Thus from the differential equation above it follows:

$$-q\hat{E}e^{-\mathrm{i}\omega t}-\frac{m\hat{v}e^{-\mathrm{i}\omega t}}{\tau}=m\frac{\mathrm{d}(\hat{v}e^{-\mathrm{i}\omega t})}{\mathrm{d}t} \tag{10.4.56}$$

Therefore:

$$-q\hat{E}e^{-\mathrm{i}\omega t}-\frac{m\hat{v}e^{-\mathrm{i}\omega t}}{\tau}=-m\hat{v}\mathrm{i}\omega e^{-\mathrm{i}\omega t} \tag{10.4.57}$$

Simplifying:

$$-q\hat{E}-\frac{m\hat{v}}{\tau}=-m\hat{v}\mathrm{i}\omega \tag{10.4.58}$$

Hence:

$$\hat{v}=-\frac{q\hat{E}}{m}\frac{1}{\tau^{-1}-\mathrm{i}\omega} \tag{10.4.59}$$

This gives the current density:

$$\hat{J}=-nq\hat{v}=\frac{nq^2\hat{E}}{m}\frac{1}{1/\tau-\mathrm{i}\omega}=\frac{\rho_q q}{m}\frac{1}{\tau^{-1}-\mathrm{i}\omega}\hat{E} \tag{10.4.60}$$

and thus the complex material conductivity is defined as:

$$\boxed{\hat{\sigma}=\frac{nq_e^2}{m_e}\frac{1}{\tau^{-1}-\mathrm{i}\omega}} \tag{10.4.61}$$

The complex conductivity relates both the magnitude and the phase of the current density to the magnitude and phase of the applied electric field similar to the current and voltage in an alternative current circuit are related by the complex conductance. This can be most easily seen when you remember that only real parts of the complex values of current and electric field are used and you express the complex conductivity by the complex exponential $\sigma=|\hat{\sigma}|\exp(-\mathrm{i}\phi)$. Then you see that $|\hat{\sigma}|$ is the magnitude of the proportionality constant and φ is the current-field phase shift.

The real part of $\hat{\sigma}$ thus describes the convective current and the imaginary part corresponds to a displacement current (imaginary part indicates that the current lags behind the electrical field, which happens because the electrons need roughly a time τ to accelerate in response to a change in the electrical field).

Like radio waves, the electric field of the optical light acts on the electrons in a material and can cause them to move. But visible light has a much shorter wavelength than radio waves (about 10^6 shorter) and so it doesn't drive macroscopic currents. Moreover, the conductivity decreases, which we can see in the Drude Model:

$$\hat{\sigma}(\omega)=\frac{\sigma_0}{1+\mathrm{i}\omega\tau}=\frac{\sigma_0}{1+\omega^2\tau^2}-\mathrm{i}\omega\tau\frac{\sigma_0}{1+\omega^2\tau^2} \tag{10.4.62}$$

where for recall ω is the light frequency, τ is the relaxation time (due to scattering and related to the carrier density, temperature, etc.), and $\sigma_0=\frac{\rho_q q\tau}{m}$ is the "effective DC conductivity" (complex conductivity when $\omega=0$) also named "nominal conductivity" (n is the number density and m is an effective mass which can be different for electrons and holes depending on the material).

We can see that when the frequency ω is large (like for optical waves) the conductivity goes to zero. This happens because at really high frequencies the electrons do not have time to respond to the field and so they no longer (appreciably) move back and forth. Above this frequency the material becomes transparent (identical to the behaviour at the critical frequency in plasmas).

10.4.3.2 Simple (harmonic oscillator) model of dielectric and conductors (and plasmas)

As we have seen during our study of Rayleigh scattering (see section Nuclear Physics page 5102), a simple model for the dielectric properties of a material may be obtained by considering the motion of a bound electron in the presence of an applied electric field. As the electric field tries to separate the electron from the positively charged nucleus, it generates an electric dipole moment. Averaging this dipole moment over the volume of the material gives rise to a macroscopic dipole moment per unit volume.

A simple model for the dynamics of the displacement x of the bound electron is as follows (with $\dot{x}=\mathrm{d}x/\mathrm{d}t$):

$$m_e\ddot{x}=q_eE-kx-m_e\gamma\dot{x} \tag{10.4.63}$$

where we assumed that the electric field is acting in the x-direction and that there is a spring-like restoring force due to the binding of the electron to the nucleus, and a friction-type force proportional to the velocity of the electron.

The spring constant k is related to the resonance frequency of the spring via the relationship $\omega_0=\sqrt{k/m_e}$, or, $k=m_e\omega_0^2$. Therefore, we may rewrite the above relation as:

$$\ddot{x}+\gamma\dot{x}+\omega_0^2x=\frac{q_e}{m_e}E \tag{10.4.64}$$

The limit $\omega_0=0$ corresponds to unbound electrons and describes the case of good conductors. The frictional term $\gamma\dot{x}$ arises from collisions that tend to slow down the electron. The parameter y is a measure of the rate of collisions per unit time, and therefore, $\tau=1/\gamma$ will represent the mean-time between collisions.

In a typical conductor, τ is of the order of 10^{-14} seconds, for example, for copper, $\tau=2.4\cdot 10^{-14}$ [s] and $\gamma=4.1\cdot 10^{13}$ [s^{-1}]. The case of a tenuous, collisionless, plasma can be obtained in the limit $\gamma=0$. Thus, the above simple model can describe the following cases:

a. Dielectrics, $\omega_0\neq 0,\gamma\neq 0$

b. Conductors, $\omega_0=0,\gamma\neq 0$

c. Collisionless Plasmas, $\omega_0=0,\gamma=0$

The basic idea of this model is that the applied electric field tends to separate positive from negative charges, thus, creating an electric dipole moment. In this sense, the model contains the basic features of other types of polarization in materials, such as Ionic/molecular polarization arising from the separation of positive and negative ions by the applied field, or polar materials that have a permanent dipole moment.

10.4.3.2.1 Dielectrics

The applied electric field $E(t)$ above can have any time dependence. In particular, if we will assume it is sinusoidal with frequency ω, $E(t)=E_0e^{\mathrm{i}\omega t}$, then the above differential equation will have for solution as proved during our study of Rayleigh scattering (see section Nuclear Physics page 5102):

$$x=\frac{\frac{q_e}{m_e}E}{\omega_0^2-\omega^2+\mathrm{i}\omega\gamma} \tag{10.4.65}$$

The corresponding velocity of the electron will also be sinusoidal $v(t)=ve^{\mathrm{i}\omega t}$, where $v=\dot{x}=\mathrm{i}\omega x$. Thus, we have:

$$v=\mathrm{i}\omega x=\frac{\mathrm{i}\omega\frac{q_e}{m_e}E}{\omega_0^2-\omega^2+\mathrm{i}\omega\gamma} \tag{10.4.66}$$

From the both relations above we can find the polarization per unit volume P. We assume that there are N such elementary dipoles per unit volume. The individual electric dipole moment is $p=q_ex$ (see section Electrostatics page 3984). Therefore, the polarization per unit volume will be:

$$P=Np=Nq_ex=\frac{\frac{Nq_e^2}{m_e}E}{\omega_0^2-\omega^2+\mathrm{i}\omega\gamma}:=\varepsilon_0\chi(\omega)E \tag{10.4.67}$$

The electric flux density (see page 4108) will be then:

$$D=\varepsilon_0E+P=\varepsilon_0(1+\chi(\omega))E:=\varepsilon(\omega)E \tag{10.4.68}$$

where the "effective permittivity" $\varepsilon(\omega)$ is then defined by:

$$\varepsilon(\omega)=\varepsilon_0+\frac{\frac{Nq_e^2}{m_e}}{\omega_0^2-\omega^2+\mathrm{i}\omega\gamma} \tag{10.4.69}$$

This can be written in a more convenient form, as follows:

$$\boxed{\varepsilon(\omega)=\varepsilon_0+\frac{\varepsilon_0\omega_p^2}{\omega_0^2-\omega^2+\mathrm{i}\omega\gamma}} \tag{10.4.70}$$

where ω_p^2 is the so-named "plasma frequency" of the material defined by (we have already met this relation during our study of plasmas at page 3947):

$$\boxed{\omega_p^2=\frac{Nq_e^2}{\varepsilon_0m_e}} \tag{10.4.71}$$

The model defined by:

$$\varepsilon(\omega)=\varepsilon_0+\frac{\varepsilon_0\omega_p^2}{\omega_0^2-\omega^2+\mathrm{i}\omega\gamma} \tag{10.4.72}$$

is known as a "Lorentz dielectric". The corresponding susceptibility, defined through $\varepsilon(\omega)=\varepsilon_0(1+x(\omega))$, is:

$$\chi(\omega)=\frac{\omega_p^2}{\omega_0^2-\omega^2+\mathrm{i}\omega\gamma} \tag{10.4.73}$$

For a dielectric, we may assume $\omega_0 \neq 0$. Then, the low-frequency limit ($\omega = 0$) gives the "nominal dielectric constant":

$$\varepsilon(\omega) = \varepsilon_0 + \frac{\frac{Nq_e^2}{m_e}}{\omega_0^2 - \omega^2 + \mathrm{i}\omega\gamma} \Rightarrow \varepsilon(0) = \varepsilon_0 + \varepsilon_0\frac{\omega_p^2}{\omega_0^2} = \varepsilon_0 + \frac{Nq_e^2}{m_e\omega_0^2} \tag{10.4.74}$$

The real and imaginary parts of $\varepsilon(\omega)$ characterize by convention the refractive and absorptive properties of the material. By convention, we define the imaginary part with the negative sign (because we use $e^{\mathrm{i}\omega t}$ time dependence):

$$\varepsilon(\omega) = \varepsilon'(\omega) - \mathrm{i}\varepsilon''(\omega) \tag{10.4.75}$$

It follows from:

$$\varepsilon(\omega) = \varepsilon_0 + \frac{\varepsilon_0\omega_p^2}{\omega_0^2 - \omega^2 + \mathrm{i}\omega\gamma} \tag{10.4.76}$$

that:

$$\varepsilon'(\omega) = \varepsilon_0 + \frac{\varepsilon_0\omega_p^2\left(\omega_0^2 - \omega^2\right)}{\left(\omega^2 - \omega_0^2\right)^2 + \gamma^2\omega^2} \qquad \varepsilon''(\omega) = \frac{\varepsilon_0\omega_p^2\omega\gamma}{\left(\omega^2 - \omega_0^2\right)^2 + \gamma^2\omega^2} \tag{10.4.77}$$

The figure below shows a plot of $\varepsilon'(\omega)$ and $\varepsilon''(\omega)$. Around the resonant frequency ω_0, the real part $\varepsilon'(\omega)$ behaves in an anomalous manner, that is, it drops rapidly with frequency to values less than ε_0 and the material exhibits strong absorption. The term "normal dispersion" refers to an $\varepsilon'(\omega)$ that is an increasing function of ω, as is the case to the far left and right of the resonant frequency:

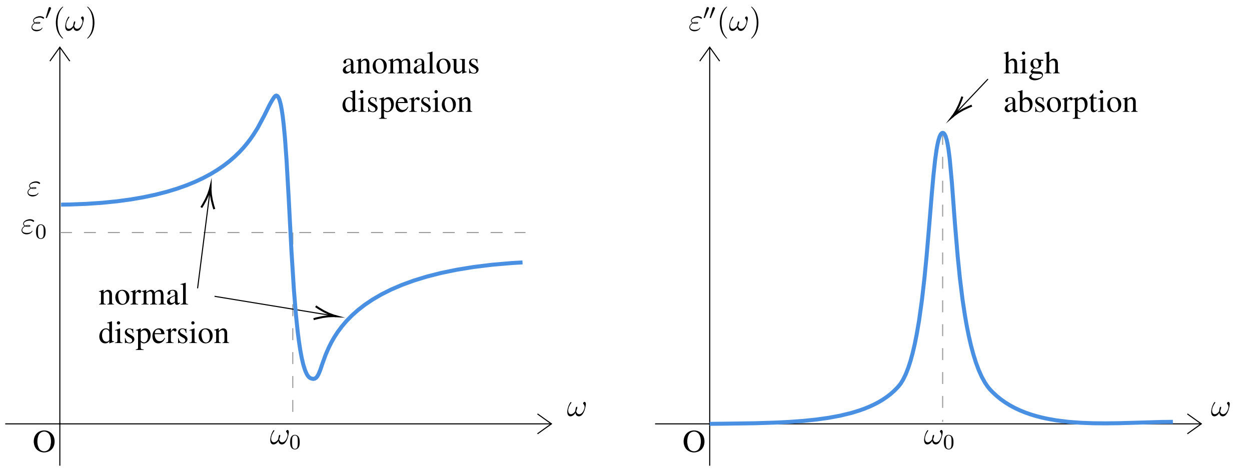


Figure 10.94: **Real and imaginary parts of the effective permittivity $\varepsilon(\omega)$**

Real dielectric materials exhibit, of course, several such resonant frequencies corresponding to various vibrational modes and polarization mechanisms (e.g., electronic, ionic, etc.). The permittivity becomes the sum of such terms:

$$\varepsilon(\omega) = \varepsilon_0 + \varepsilon_0\sum_i \frac{N_i q_{e,i}^2/m_{e,i}\varepsilon_0}{\omega_i^2 - \omega^2 + \mathrm{i}\omega\gamma_i} \tag{10.4.78}$$

Remark

A more correct quantum-mechanical treatment leads essentially to the same relation:

$$\varepsilon(\omega)=\varepsilon_0+\varepsilon_0\sum_{j>i}\frac{f_{ji}\left(N_i-N_j\right)q_e^2/m_e\varepsilon_0}{\omega_{ji}^2-\omega^2+\mathrm{i}\omega\gamma_{ji}} \tag{10.4.79}$$

where ω_{ji} are transition frequencies between energy levels, that is, $\omega_{ji}=\left(E_j-E_i\right)/h$ and N_i, N_j are the populations of the lower, E_i, and upper, E_j, energy levels. The quantities f_{ji} are named "oscillator strengths." For example, for a two-level atom we have:

$$\varepsilon(\omega)=\varepsilon_0+\varepsilon_0\frac{f\omega_p^2}{\omega_0^2-\omega^2+\mathrm{i}\omega\gamma} \tag{10.4.80}$$

where we defined:

$$\omega_0=\omega_{21}, \qquad f=f_{21}\frac{N_1-N_2}{N_1+N_2}, \qquad \omega_p^2=\frac{\left(N_1+N_2\right)e^2}{m\varepsilon_0} \tag{10.4.81}$$

Normally, lower energy states are more populated, $N_i>N_j$, and the material behaves as a classical absorbing dielectric. However, if there is population inversion, $N_i<N_j$ then the corresponding permittivity term changes sign. This leads to a negative imaginary part, $\varepsilon''(\omega)$, representing a gain.

As:

$$n^2(\omega)=\frac{c}{v_n}=\sqrt{\varepsilon_r\mu_r}=\sqrt{\varepsilon(\omega)/\varepsilon_0} \tag{10.4.82}$$

For a material have and relative permeability equal to 1 (vacuum) then:

$$n^2(\omega)=\sqrt{\varepsilon_r}=\sqrt{\varepsilon(\omega)/\varepsilon_0} \tag{10.4.83}$$

Therefore:

$$\varepsilon(\omega)=\varepsilon_0+\varepsilon_0\sum_i\frac{N_iq_{e,i}^2/m_{e,i}\varepsilon_0}{\omega_i^2-\omega^2+\mathrm{i}\omega\gamma_i} \tag{10.4.84}$$

can be written in the following form, known as the "Sellmeier equation" (where the B_i are constants):

$$\boxed{n^2(\omega)=1+\sum_i\frac{B_i\omega_i^2}{\omega_i^2-\omega^2+\mathrm{i}\omega\gamma_i}} \tag{10.4.85}$$

In practice the relation above is applied in frequency ranges that are far from any resonance so that one can effectively set $\gamma_i=0$:

$$n^2(\omega)=1+\sum_i\frac{B_i\omega_i^2}{\omega_i^2-\omega^2}=1+\sum_i\frac{B_i\lambda^2}{\lambda^2-\lambda_i^2} \tag{10.4.86}$$

or:

$$n_\lambda=\sqrt{1+\sum_i\frac{B_i\lambda^2}{\lambda^2-\lambda_i^2}} \tag{10.4.87}$$

where λ,λ_i denote the corresponding free-space wavelengths (e.g., $\lambda=2\pi c/\omega$). In practice, refractive index data are fitted to the relation above using $2-4$ terms over a desired frequency range. For example, fused silica (SiO_2) is very accurately represented over the range $0.2\leq\lambda\leq 3.7\;[\mu\text{m}]$ by the following formula where λ and λ_i are in units of μ:

$$n_\lambda^2=1+\frac{0.6961663\lambda^2}{\lambda^2-(0.0684043)^2}+\frac{0.4079426\lambda^2}{\lambda^2-(0.1162414)^2}+\frac{0.8974794\lambda^2}{\lambda^2-(9.896161)^2} \tag{10.4.88}$$

Remark

The previous relation:

$$n^2(\omega)=1+\sum_i\frac{B_i\omega_i^2}{\omega_i^2-\omega^2} \tag{10.4.89}$$

Can be found in some textbook for the first term ($i=0$) under the following form:

$$n^2\approx 1+\chi_e(0)\frac{\omega_{0,e}^2}{\omega_{0,e}^2-\omega^2}=1+\chi_e(0)\frac{1}{1-\omega^2/\omega_{0,e}^2}\approx 1+\chi_e(0)\left(1+\frac{\omega^2}{\omega_{0,e}^2}\right) \tag{10.4.90}$$

Therefore:

$$n^2\approx 1+\chi_e(0)+\frac{\chi_e(0)}{\omega_{0,e}^2}\left(\frac{2\pi c}{\lambda_0}\right)^2=A+\frac{B}{\lambda_0^2}\quad\text{and}\quad n\approx\left(A+\frac{B}{\lambda_0^2}\right)^{1/2} \tag{10.4.91}$$

A et B being two constants.When the dispersion is low, $B/\left(\lambda_0^2A\right)\ll 1$, we fall back on the well-known "Cauchy's formula":

$$n\approx n_0+\frac{C}{\lambda_0^2} \tag{10.4.92}$$

This relationship is well verified for dielectrics used in optics. However the constants n_0 and C are determined experimentally because the simple model adopted here does not take into account the complexity of the electronic properties of materials.

10.4.3.2.2 Conductors

We will see now that using the first results above we can fall back on relation of the complex conductivity we get at page 4238.

The conductivity properties of a material are described by Ohm's law. To derive this law from our simple model, we use the relationship $J=\rho v$, where the volume density of the conduction charges is $\rho=Nq_e$. It follows from above that:

$$J=\rho v=Nq_ev=\frac{\mathrm{i}\omega\frac{Nq_e^2}{m}E}{\omega_0^2-\omega^2+\mathrm{i}\omega\gamma}:=\sigma(\omega)E \tag{10.4.93}$$

and therefore, we identify the conductivity $\sigma(\omega)$:

$$\sigma(\omega)=\frac{\mathrm{i}\omega\frac{Nq_e^2}{m}}{\omega_0^2-\omega^2+\mathrm{i}\omega\gamma}=\frac{\mathrm{i}\omega\varepsilon_0\omega_p^2}{\omega_0^2-\omega^2+\mathrm{i}\omega\gamma} \tag{10.4.94}$$

We note that $\sigma(\omega)/\mathrm{i}\omega$ is essentially the electric susceptibility considered above. Indeed, we have:

$$J = Nq_e v = Nq_e \mathrm{i}\omega x = \mathrm{i}\omega P \tag{10.4.95}$$

and thus, $P = J/\mathrm{i}\omega = (\sigma(\omega)/\mathrm{i}\omega)E$. It follows that $\varepsilon(\omega) - \varepsilon_0 = \sigma(\omega)/\mathrm{i}\omega$, and:

$$\varepsilon(\omega) = \varepsilon_0 + \frac{\varepsilon_0\omega_p^2}{\omega_0^2 - \omega^2 + \mathrm{i}\omega\gamma} = \varepsilon_0 + \frac{\sigma(\omega)}{\mathrm{i}\omega} \tag{10.4.96}$$

since in a metal the conduction charges are unbound, we may take $\omega_0 = 0$. After cancelling a common factor of $\mathrm{i}\omega$, we get:

$$\boxed{\sigma(\omega) = \frac{\varepsilon_0\omega_p^2}{\gamma + \mathrm{i}\omega} = \frac{\varepsilon_0\omega_p^2}{\tau^{-1} + \mathrm{i}\omega}} \tag{10.4.97}$$

So we fall back on the expected result. The nominal conductivity is obtained at the low-frequency limit as we know, $\omega = 0$:

$$\sigma = \frac{\varepsilon_0\omega_p^2}{\gamma} = \frac{Nq_e^2}{m_e\gamma} \tag{10.4.98}$$

✍**Example:**

Copper has a mass density of $8.9\cdot 10^6\;[\text{gr}\cdot\text{m}^{-3}]$ and atomic weight of $63.5\;[\text{gr}\cdot\text{mole}^{-1}]$. Using Avogadro's number of $6\cdot 10^{23}$ atoms per mole, and assuming one conduction electron per atom, we find for the volume density N :

$$N = \frac{6\cdot 10^{23}\frac{\text{atoms}}{\text{mole}}}{63.54\frac{\text{gr}}{\text{mole}}}\left(8.9\cdot 10^6\,\frac{\text{gr}}{\text{m}^3}\right)\left(1\frac{\text{electron}}{\text{atom}}\right) = 8.4\cdot 10^{28}\;[\text{electrons}\cdot\text{m}^{-3}]$$

It follows that:

$$\sigma = \frac{Nq_e^2}{m_e\gamma} = \frac{(8.4\cdot 10^{28})\,(1.6\cdot 10^{-19})^2}{(9.1\cdot 10^{-31})\,(4.1\cdot 10^{13})} = 5.8\cdot 10^7\;[\Omega^{-1}\cdot\text{m}^{-1}] = 5.8\cdot 10^7\;[\text{Siemens}\cdot\text{m}^{-1}]$$

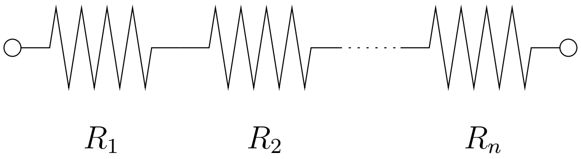
where we used $q_e = 1.6\cdot 10^{-19}$ [C], $m_e = 9.1\cdot 10^{-31}$ [kg], $\gamma = 4.1\cdot 10^{13}\;[\text{s}^{-1}]$. The plasma frequency of copper can be calculated by:

$$f_P = \frac{\omega_\text{P}}{2\pi} = \frac{1}{2\pi}\sqrt{\frac{Nq_e^2}{m_e\varepsilon_0}} = 2.6\cdot 10^{15}\;[\text{Hz}]$$

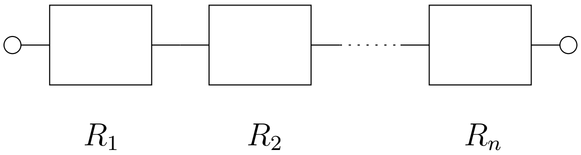
which lies in the ultraviolet range. For frequencies such that $\omega \ll \gamma$, the conductivity may be considered to be independent of frequency and equal to the de value of $\sigma = Nq_e^2/m_e\gamma$. This frequency range covers most present-day RF applications. For example, assuming $\omega \leq 0.1\gamma$, we find $f \leq 0.1\gamma/2\pi = 653$ [GHz].

10.4.3.3 Equivalent Resistance

We can now study the entire length of an electric field line collinear with a constant current I supposed to be constant at any point (this is an approximation obviously...) to obtain total resistance if n resistive elements are put next to each other linearly:



Or using another typical symbol in electronics (the both are accepted and documented in *IEC 60617: Standard Graphic Symbols for Circuit Diagrams*):



The answer is relatively simple since if we denote by U_{n-1} the potential in the first extremity of the resistive element and U_n that of the other extremity. We then have (the reader will have notice that the use of the mesh law in the following relation logically without even necessarily be aware of it existence):

$$U=(U_0-U_1)+(U_1-U_2)+\ldots+(U_{n-1}-U_n)=R_1I+R_2I+\ldots+R_nI \tag{10.4.99}$$

that is to say, a result similar to that obtained by a single resistance whose value is approximately given by (if the electric current is constant throughout the wire) the "equivalent resistance of resistors in series":

$$\boxed{R_e\cong\sum_{i=1}^{n}R_i} \tag{10.4.100}$$

which is the arithmetic sum of the individual resistances.

Let us now consider n resistance in parallel all at a given voltage U (by law of mesh!) and alimented by an electric current I. The current then splits into n streams such that:

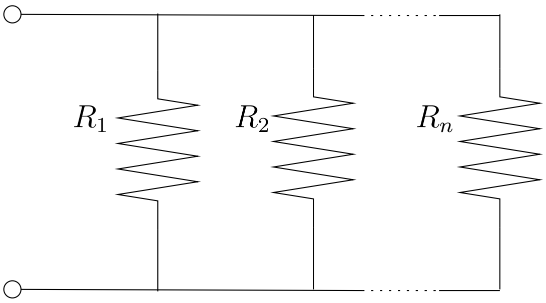
$$I_i=\frac{U}{R_i} \tag{10.4.101}$$

in each of the n mesh. Applying the node law, we have:

$$I_e=\sum_{i=1}^{n}I_i=\sum_{i=1}^{n}\frac{U}{R_i}=\frac{U}{R_e} \tag{10.4.102}$$

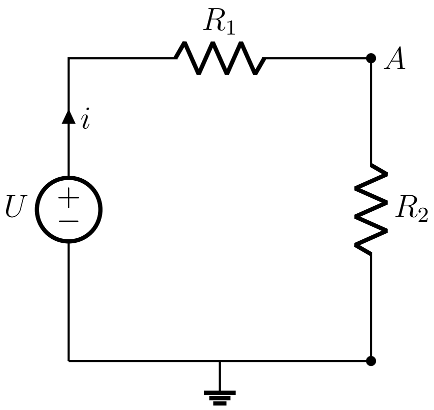
that is to say that the set of all the resistance put in parallel is analogous to an "equivalent resistance of resistors in parallel":

$$\boxed{\frac{1}{R_e}=\sum_{i=1}^{n}\frac{1}{R_i}\Leftrightarrow R_e=\frac{1}{\displaystyle\sum_{i=1}^{n}\frac{1}{R_i}}} \tag{10.4.103}$$



Plugging devices in parallel allows to always have the same voltage across them (neglecting the voltage drops). This is the way that the electrical outlets are installed in a domestic installation!

Therefore for a series circuit, we can calculate the equivalent resistance by adding them. Using the notation of the LaTeX Circuitikz package we have:



For this circuit, we have:

$$R_{\text{eq}}=R_1+R_2 \tag{10.4.104}$$

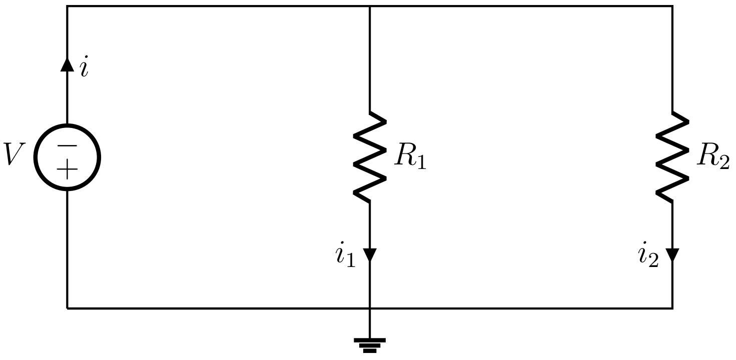
and the voltage:

$$U_A=\frac{R_2}{R_1+R_2}U \tag{10.4.105}$$

The current is given obviously by $I=\frac{U}{R_1+R_2}$ and the voltage at A is given by:

$$U_A=U-IR_1=U\left(1-\frac{R_1}{R_1+R_2}\right)=\frac{R_2}{R_1+R_2}U \tag{10.4.106}$$

For a parallel circuit, the effective resistance is the harmonic sum as proven earlier:



In this circuit, we have therefore:

$$R_{\text{eq}}=\left(\frac{1}{R_1}+\frac{1}{R_2}\right)^{-1}=\frac{R_1R_2}{R_1+R_2} \tag{10.4.107}$$

and the current in the branches is given by:

$$I_1=\frac{R_2}{R_1+R_2}I \tag{10.4.108}$$

Indeed, as the voltage drop across R_1 and R_2 must be the same, we have:

$$I_1R_1=I_2R_2 \tag{10.4.109}$$

We also have $I_2=I-I_1$, which gives:

$$I_1R_1=(I-I_1)R_2 \Longrightarrow I_1=\frac{R_2}{R_1+R_2}I \tag{10.4.110}$$

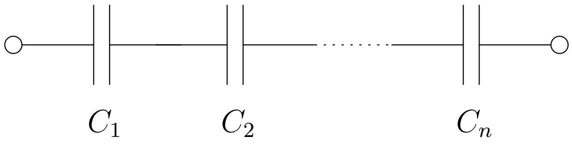
10.4.3.4 Equivalent Capacities

Even if this has nothing to do with the Ohm's law we can apply the previous reasoning to capacitor.

Let us recall that we defined in the section Electrostatics, the capacity given by:

$$C=\frac{Q}{U} \tag{10.4.111}$$

Let us consider, as well as resistors, n capacitors of capacity C_i in series set behind each other:



We put at potential U_0 and U_n the two extremities of the chain and we bring the charge Q on the whole system. The potential (voltage) across the total capacitor chain is then written simply:

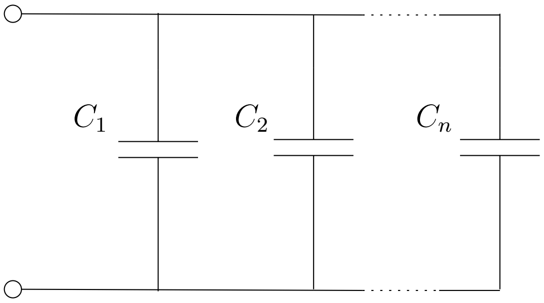
$$U=U_0-U_n=(U_0-U_1)+(U_1-U_2)+\ldots+(U_{n-1}-U_n)=\frac{Q_1}{C_1}+\frac{Q_2}{C_2}+\ldots+\frac{Q_n}{C_n}=Q\sum_{i=1}^{n}\frac{1}{C_i} \tag{10.4.112}$$

and therefore corresponds to that of a single capacitor C_e that is the "equivalent capacitance of capacitors in series":

$$\boxed{\frac{1}{C_{\text{eq}}}=\sum_{i=1}^{n}\frac{1}{C_i}\Leftrightarrow C_e=\frac{1}{\displaystyle\sum_{i=1}^{n}\frac{1}{C_i}}} \tag{10.4.113}$$

where we also find here a harmonic mean.

Now let us consider n capacitor of capacity C_i in parallel with the same potential U:



The electric charge of each is then imposed (by the mesh law) by the relation:

$$Q_i=C_iU \tag{10.4.114}$$

The total electrical load is simply:

$$Q=\sum_{i=1}^{n}Q_i=U\sum_{i=1}^{n}C_i \tag{10.4.115}$$

which corresponds to an "equivalent capacitance of capacitors in parallel":

$$\boxed{C_{\text{eq}}=\sum_{i=1}^{n}C_i} \tag{10.4.116}$$

which is the arithmetic sum of the individual capacities.

Finally to close this subject, let us recall that we have proved in the section of Electrostatics that:

$$i=C\frac{\mathrm{d}u}{\mathrm{d}t} \tag{10.4.117}$$

In the case where a capacity is alone in series with an AC sine generator (fairly typical case in the industrial world of the 119th and 120th century), then we have:

$$i=C\frac{\mathrm{d}}{\mathrm{d}t}U_0\sin(\omega t)=\omega CU_0\cos(\omega t)=\omega CU_0\sin\left(\omega t+\frac{\pi}{2}\right) \tag{10.4.118}$$

And therefore we get:

$$I_0=\omega CU_0 \tag{10.4.119}$$

That we will write in analogy with Ohm's law in the form:

$$U_0 = \frac{1}{\omega C} I_0 = X_C I_0 \tag{10.4.120}$$

hence:

$$X_C = \frac{1}{\omega C} = \frac{1}{2\pi f C} \tag{10.4.121}$$

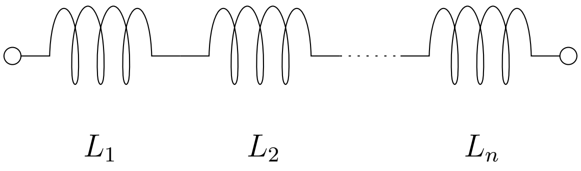
is named the "capacitive reactance". We notice that in the continuous case where the pulsation is zero, the capacitive reactance becomes infinite and that we find the known situation where the capacity does not pass current (at least in the ideal case ...).

Therefore, a capacitor connected to a circuit that changes over a given range of frequencies can be said to be "Frequency dependent".

10.4.3.5 Equivalent Inductance

Even if this has nothing to do with the Ohm's law we can apply the previous reasoning to inductance (see page 4261).

Let us consider the following circuit:



Since it is a series circuit, the flowing current will be the same through all inductors, while each of them may have different voltages across them.

The current in the circuit is:

$$i = i_1 = i_2 = i_3 = i_n \tag{10.4.122}$$

Using Kirchhoff's Voltage Law, the total voltage is:

$$u = u_1 + u_2 + u_3 + \ldots + u_n \tag{10.4.123}$$

Substituting:

$$u_i = L_i \frac{\mathrm{d}i}{\mathrm{d}t} \tag{10.4.124}$$

Results in:

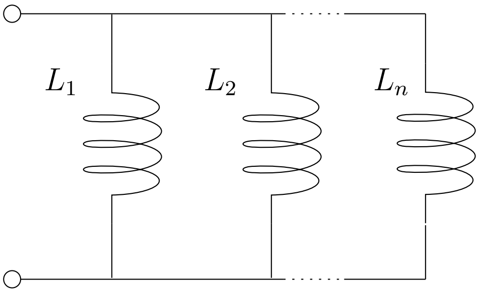
$$\begin{aligned} u &= L_1 \frac{\mathrm{d}i}{\mathrm{d}t} + L_2 \frac{\mathrm{d}i}{\mathrm{d}t} + L_3 \frac{\mathrm{d}i}{\mathrm{d}t} + \ldots + L_n \frac{\mathrm{d}i}{\mathrm{d}t} \\ &= (L_1 + L_2 + L_3 + \ldots + L_n) \frac{\mathrm{d}i}{\mathrm{d}t} \\ &= \left(\sum_{i=1}^{n} L_i \right) = L_{\text{eq}} \frac{\mathrm{d}i}{\mathrm{d}t} \end{aligned} \tag{10.4.125}$$

Therefore we have the "equivalent inductance of inductors in series":

$$\boxed{L_{\text{eq}}=\sum_{i=1}^{n}L_i} \tag{10.4.126}$$

In conclusion, the equivalent inductance of n inductors in series is (approximately because of heat dissipation) the sum of all the individual inductances.

For the parallel case, consider the circuit below consisting again of n inductors:



Since it is a parallel circuit, the voltage will be the same across all inductors, while each of them may have different current through them.

The voltage across the inductors are:

$$u=u_1=u_2=u_3=u_n \tag{10.4.127}$$

The total current in the circuit is:

$$i=i_1+i_2+i_3+\ldots+i_n \tag{10.4.128}$$

Substituting:

$$i_i=\frac{1}{L_i}\int_{t_0}^{t}v\mathrm{d}t+i_i\left(t_0\right) \tag{10.4.129}$$

Results in:

$$\begin{aligned}i&=\frac{1}{L_1}\int_{t_0}^{t}v\mathrm{d}t+i_1\left(t_0\right)+\frac{1}{L_2}\int_{t_0}^{t}v\mathrm{d}t+i_2\left(t_0\right)+\ldots+\frac{1}{L_n}\int_{t_0}^{t}v\mathrm{d}t+i_n\left(t_0\right)\\&=\left(\frac{1}{L_1}+\frac{1}{L_2}+\ldots+\frac{1}{L_n}\right)\int_{t_0}^{t}v\mathrm{d}t+i_1\left(t_0\right)+i_2\left(t_0\right)+\ldots+i_n\left(t_0\right)\\&=\sum_{i=1}^{n}\frac{1}{L_i}\int_{t_0}^{t}v\mathrm{d}t+\sum_{i=1}^{n}i_i(t_0)\\&=\frac{1}{L_{\text{eq}}}\int_{t_0}^{t}v\mathrm{d}t+i\left(t_0\right)\end{aligned} \tag{10.4.130}$$

Hence, the "equivalent inductance of inductors in parallel":

$$\boxed{\frac{1}{L_{\text{eq}}}=\sum_{i=1}^{n}\frac{1}{L_i}\Leftrightarrow L_{\text{eq}}=\frac{1}{\sum_{i=1}^{n}\frac{1}{L_i}}} \tag{10.4.131}$$

In conclusion, the equivalent inductance of n inductors in parallel is the reciprocal of the sum of all the individual inductances.

The total inductance of parallel inductors is always smaller than the smallest individual inductance in the calculation.

From the inductance formula in series and parallel above, we can conclude that the way we find equivalent inductance is the same as when we find equivalent resistance.

To finish on that topic, the reader must however keep in mind that as always... reality is a little bit more complex than idealized mathematical models. Relatively to inductances we can consider the following three cases:

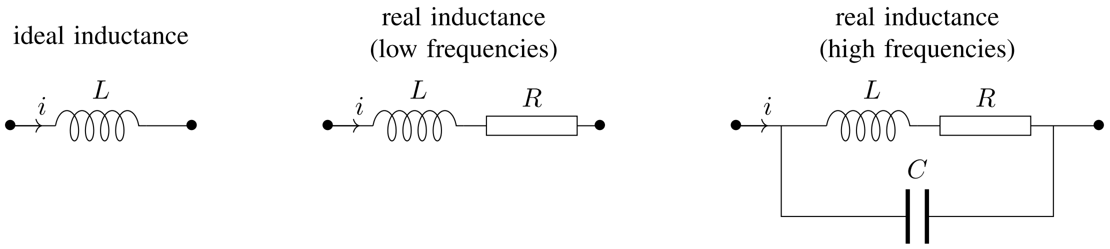

ideal inductance
i
L
real inductance
(low frequencies)
i
L
R
real inductance
(high frequencies)
i
L
R
C

Figure 10.95: Different inductance models (author: Jimmy Roussel)

10.4.4 Electromotive Force

Given AB a portion of an electric circuit travelled by a constant current I from A to B. The existence of this current implies that the potential on A is greater (different) in absolute value than in B (in absolute value). This potential difference is reflected in the existence of the electrostatic field $\vec{E}$ producing a Coulomb force:

$$\vec{F}=q\vec{E} \tag{10.4.132}$$

capable of accelerating a charge q.

Then, given:

$$P_q=Fv \tag{10.4.133}$$

the power required to give a speed v to a any particle of charge q. Knowing that this in the conductor there is ρ_q electric charge per unit volume, the total power P in the circuit AB travelled by a current I is:

$$\begin{aligned}P&=\int\limits_V n_qP_q\mathrm{d}V=\int\limits_A^B\mathrm{d}l\int\limits_S n_qP_q\mathrm{d}S=\int\limits_A^B\mathrm{d}l\int\limits_S n_qFv\mathrm{d}S=\int\limits_A^B\mathrm{d}l\int_S(n_qqEv)\mathrm{d}S\\&=\int\limits_A^B(\rho_qv\mathrm{d}S)E\mathrm{d}l=\int\limits_S J\mathrm{d}S\int\limits_A^B E\mathrm{d}l=I\int\limits_A^B E\mathrm{d}l=I(U_A-U_B)=I\Delta U:=IU\end{aligned} \tag{10.4.134}$$

That is to say:

$$\boxed{P=UI} \tag{10.4.135}$$

where:

$$U=U_A-U_B>0 \tag{10.4.136}$$

This power is therefore the "electric power" available between A and B, simply because there flows a current I.

If we consider in this electric circuit $\overline{AB}$ a resistive for which we measure a potential difference:

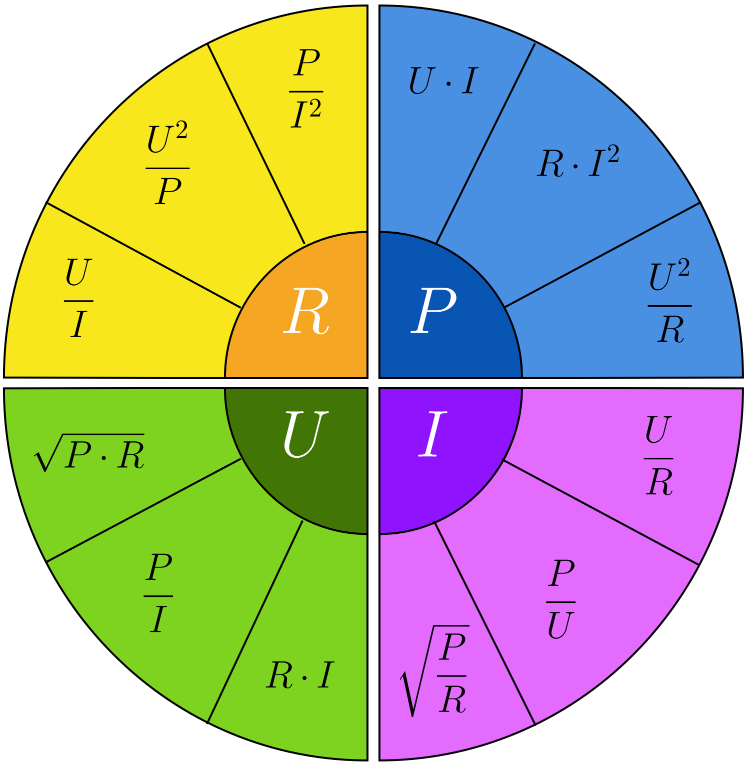
$$U=RI \tag{10.4.137}$$

then the power available inside thereof is given by the "Joule power":

$$\boxed{P_J=RI^2} \tag{10.4.138}$$

Thus, among this available power, a certain part is dissipated as heat (Joule effect) in a passive dipole such as a resistance. Obviously it is this power that charge (invoice) us our electricity provider and to know the corresponding energy consumed, we simply multiply the power of the device which is used by the functioning duration.

Now that we have the latter relation we can finally introduce the famous Ohm Circle that related the most important DC relations:



However..., something is wrong in our previous developments if we look more closely. Indeed, if we apply the reasoning in a closed circuit, that is to say, if we look at the total power supplied between A and A by the Coulomb force, we get (obviously because the Coulomb electrostatic field is conservative):

$$P=I\int_A^A E\mathrm{d}l=I(U_A-U_A)=I\cdot 0=0 \qquad (10.4.139)$$

that is to say null power ?! Eh yes! This means they can not be a steady state current in a closed loop and when there is a current, then this implies that the Coulomb force is not responsible for the global movement of charge carriers in a conductor!!

This is exactly what happens in an electric circuit. If we want that a permanent current flows, there must be a force other than the electrostatic force that enables the electric charges to close the path (this is a purely mathematical reasoning)! It is for this reason that we must involve an "artificial" external energy source such as an "electrical generator" which is then equivalent to the hydraulic pump for water.

The generator involve then as physical property that only when its circuit is open (the current I then being equal to zero) a "potential difference" is maintained between its terminals necessarily involving the presence of another force compensating the Coulomb attraction of the conductor. Thus, the total force acting on an electric charge q is written thus:

$$\vec{F}=q(\vec{E}_S+\vec{E}_M) \qquad (10.4.140)$$

with being $\vec{E}_S$ the electrostatic field and $\vec{E}_M$ the "electromotive field". In equilibrium and in the absence of current, we must have:

$$\vec{E}_S+\vec{E}_M=\vec{0} \qquad (10.4.141)$$

This means that the difference of potential across an open generator is then:

$$\vec{E}_S+\vec{E}_M=\vec{0}\Leftrightarrow\vec{E}_S=-\vec{E}_M\Leftrightarrow\int_A^B\vec{E}_S\circ\mathrm{d}\vec{l}=-\int_A^B\vec{E}_M\circ\mathrm{d}\vec{l}=U_A-U_B \quad (10.4.142)$$

We name and denote by:

$$e=\int_A^B\vec{E}_M\circ\mathrm{d}\vec{l}=U_B-U_A>0 \quad (10.4.143)$$

(somewhat abusively) the proper "electromotive force EMF" of the generator.

Since, inside the generator, we have:

$$\vec{E}_S=-\vec{E}_M\neq\vec{0} \quad (10.4.144)$$

at open circuit, this means that a generator is a non-conductive equipotential (or with "non-conservative field").

At equilibrium, but in the presence of a current I (generator set in a closed circuit), the charge carriers responsible for the current undergo additional force due to collisions occurring inside the conductor. For an ideal generator, these collisions are negligible and we get:

$$U_A-U_B=-e \quad (10.4.145)$$

However, for a non-ideal generator, such collisions occur and result in the existence of an internal resistance r (very small for generators that just go out of the factory!). Thus, the true electromotive force is given by:

$$\boxed{(U_A-U_B)+e=rI\Rightarrow U_A-U_B=rI-e} \quad (10.4.146)$$

The internal resistance of the generator introduces a voltage drop proportional to the current supplied, so it outputs a potential lower than that given by it electromotive force.

The latter relation is sometimes noted as follows:

$$U_B-U_A=e-rI \quad (10.4.147)$$

and often with the following writing:

$$U=\mathcal{E}-rI \quad (10.4.148)$$

What is measured with a voltmeter is however the generator electromotive force (GEF) since the generators have an internal resistance admit as infinite and therefore involve a current I almost equal to zero:

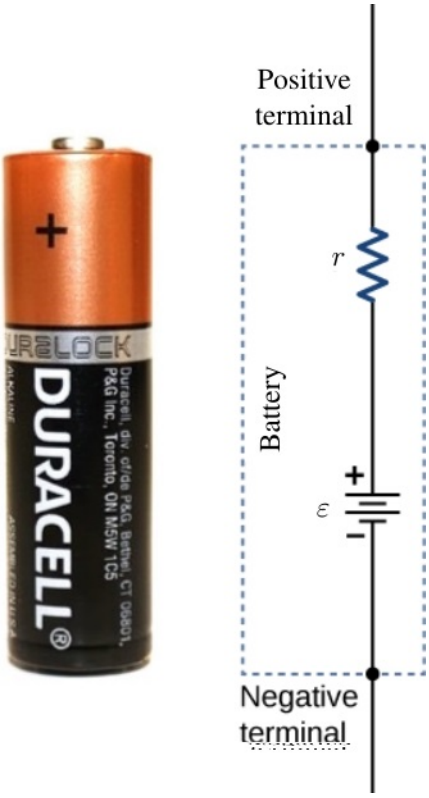


Figure 10.96: **Electromotive force schematic concept (source: OpenStax)**

Generators vary depending on the energy source used and the method of conversion of the latter into electrical energy (i.e., the nature of $\vec{E}_m$). We can then produce electrical energy from a battery (chemical energy), an electrostatic generator (mechanical energy), a dynamo (mechanical energy) a solar cell (radiation energy) or thermocouple (heat energy).

The total power P to be provided in steady state is therefore:

$$\begin{aligned}P&=\int_V n_q P_q \mathrm{d}V=\oint_{\text{circuit}}\mathrm{d}l\int_S n_q P_q \mathrm{d}S=\oint\mathrm{d}l\int_S n_q F v \mathrm{d}S=\oint_{\text{circuit}}\mathrm{d}l\int_S (n_q q v \mathrm{d}S)\frac{F}{q}\\&=\oint_{\text{circuit}}\frac{F\mathrm{d}l}{q}\int_S J\mathrm{d}S=I\oint_{\text{circuit}}\frac{F\mathrm{d}l}{q}=Ie\end{aligned}\qquad(10.4.149)$$

where:

$$e=\oint_{\text{circuit}}\frac{F}{q}\mathrm{d}l=\oint_{\text{circuit}}E_M\mathrm{d}l\qquad(10.4.150)$$

is the total EMF of the circuit. The integral being on the whole circuit, the total EMF is the sum of the EMF present along the circuit (if any). If they are located in dipoles, the expression becomes:

$$e=\sum_k e_k\qquad(10.4.151)$$

where the e_k are the algebraic values of the different EMF:

- $e_k > 0$ corresponds to "generators" devices (production of electrical energy)
- $e_k < 0$ corresponds to "receptors" devices (consumption of electrical energy)

We also have for the electric power:

$$P = Ie = \frac{e}{R+r}e = \frac{e^2}{R+r} \tag{10.4.152}$$

and for the Joule (resistive) Power:

$$P_J = RI^2 = R\left(\frac{e}{R+r}\right)^2 \tag{10.4.153}$$

A motor converts electrical energy into mechanical energy and therefore corresponds to a EMF receptor, we also say that it has a "counter-electromotive force" or CEMF.

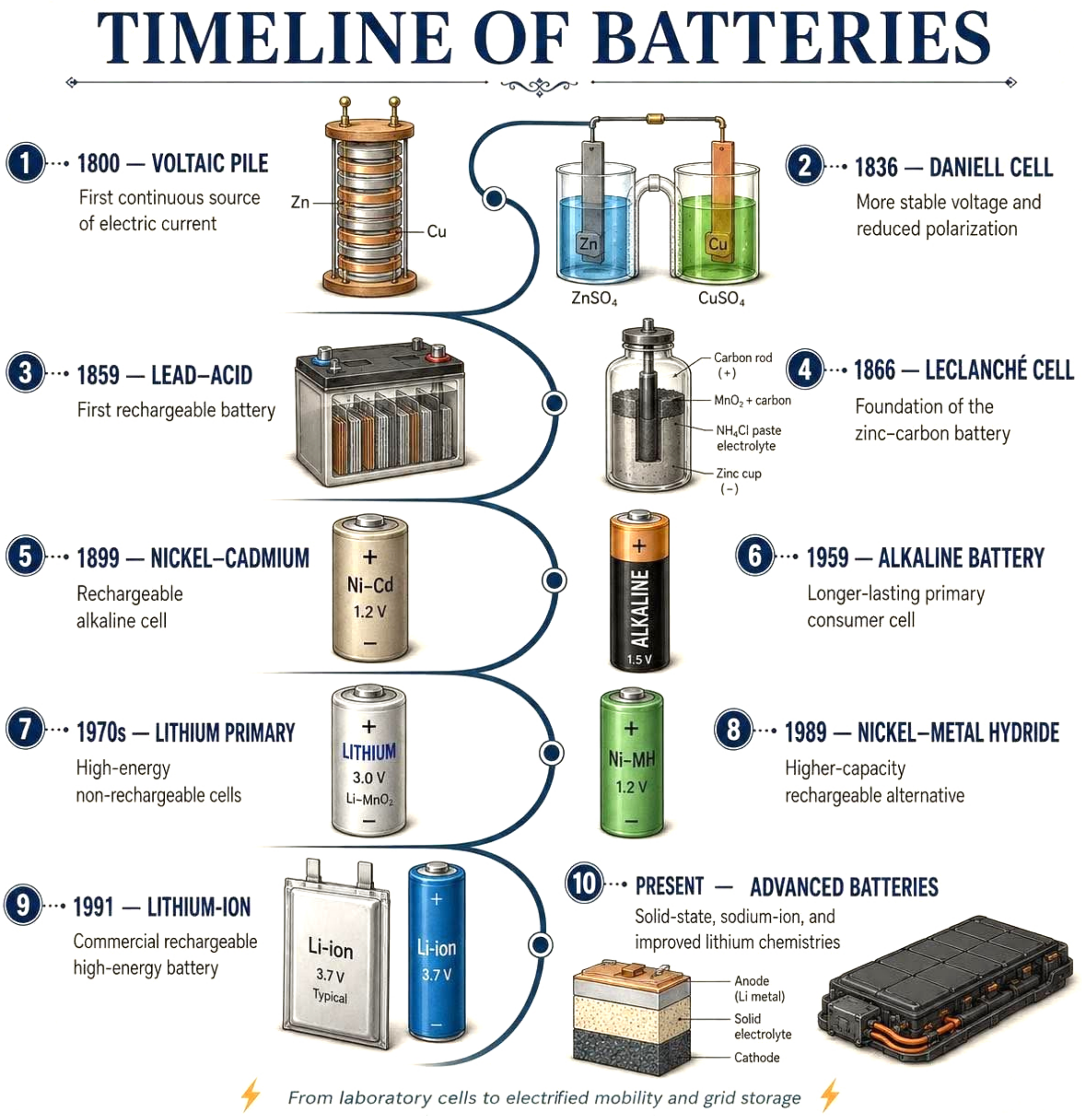

TIMELINE OF BATTERIES
1 1800 — VOLTAIC PILE
First continuous source of electric current
Zn
Cu
2 1836 — DANIELL CELL
More stable voltage and reduced polarization
Zn
Cu
$ZnSO_4$
$CuSO_4$
3 1859 — LEAD–ACID
First rechargeable battery
Carbon rod (+)
MnO_2 + carbon
NH_4Cl paste electrolyte
Zinc cup (−)
4 1866 — LECLANCHÉ CELL
Foundation of the zinc–carbon battery
5 1899 — NICKEL–CADMIUM
Rechargeable alkaline cell
+
Ni-Cd
1.2 V
−
+
ALKALINE
1.5 V
6 1959 — ALKALINE BATTERY
Longer-lasting primary consumer cell
7 1970s — LITHIUM PRIMARY
High-energy non-rechargeable cells
+
LITHIUM
3.0 V
$Li-MnO_2$
−
+
Ni-MH
1.2 V
−
8 1989 — NICKEL–METAL HYDRIDE
Higher-capacity rechargeable alternative
9 1991 — LITHIUM-ION
Commercial rechargeable high-energy battery
Li-ion
3.7 V
Typical
+
Li-ion
3.7 V
−
10 PRESENT — ADVANCED BATTERIES
Solid-state, sodium-ion, and improved lithium chemistries
Anode (Li metal)
Solid electrolyte
Cathode
From laboratory cells to electrified mobility and grid storage

Figure 10.97: **Timeline of batteries (years based on Gregorian calendar)**

10.4.4.1 Faraday's law of induction

Now that we have proved the necessity of the electromotive force, we will be able to prove the origin of the "Faraday's law" and also of the "Lenz's law" that we had used in the section of Electrodynamics to prove the third Maxwell equation.

The Determination of Faraday's law will also allow us to define the concept of inductance and study its properties.

Let us do the same approach as did Faraday and ask ourselves the following question: How do we generate an electric current?

An electric current is a moving set of electric charges in a conductive material. These electric charges are moved thanks to a difference of potential that is maintained by an electromotive force. Thus, a battery by converting the chemical energy during a time $\mathrm{d}t$ provides a power P modifying the kinetic energy of the $\mathrm{d}Q$ charge carriers producing then an electric current I.

Given P_q the power required to communicate a speed $\vec{v}$ to an electric charge particle q. Knowing that in a conductor, there is n charge carriers per unit volume, the total power P_q to be provided by the (ideal) generator is then (see above):

$$P=Ie \tag{10.4.154}$$

We therefore put that the ideal electromotive force of a circuit is:

$$e=\frac{P}{I}=\oint\limits_{\text{circuit}}\frac{F\mathrm{d}l}{q}=\oint\limits_{\text{circuit}}\frac{\vec{F}\circ\mathrm{d}\vec{l}}{q} \tag{10.4.155}$$

However, the Coulomb force is unable to produce an electromotive force as we have proved earlier. To generate a direct (continuous) electric current in a closed circuit, we therefore need an electromotive field which circulation along the circuit is not zero. The Faraday's experiment thus shows that it is the existence of the magnetic field that allows the creation of a current (!!!!). This means that the Lorentz force must be responsible for the creation of an electromotive force, that is to say:

$$e=\oint\limits_{\text{circuit}}\frac{q\left(\vec{E}+\vec{v}\times\vec{B}\circ\mathrm{d}\vec{l}\right)}{q}=\oint\limits_{\text{circuit}}\left(\vec{E}+\vec{v}\times\vec{B}\circ\mathrm{d}l\right)\circ\mathrm{d}\vec{l} \tag{10.4.156}$$

Therefore:

$$e=\oint\limits_{\text{circuit}}\left(\vec{E}+\vec{v}\times\vec{B}\circ\mathrm{d}l\right)\circ\mathrm{d}\vec{l}=\oint\limits_{\text{circuit}}\vec{E}\circ\mathrm{d}\vec{l}+\oint\limits_{\text{circuit}}(\vec{v}\times\vec{B})\circ\mathrm{d}\vec{l}=0+\oint\limits_{\text{circuit}}(\vec{v}\times\vec{B})\circ\mathrm{d}\vec{l} \tag{10.4.157}$$

The properties of the vector cross product (see section Vector Calculus 1830) giving us:

$$(\vec{x}\times\vec{y})\circ\vec{z}=-(\vec{x}\times\vec{z})\circ\vec{y} \tag{10.4.158}$$

We can then write:

$$e=\oint\limits_{\text{circuit}}\left(\vec{v}\times\vec{B}\right)\circ\mathrm{d}\vec{l}=-\oint\limits_{\text{circuit}}\left(\vec{v}\times\mathrm{d}\vec{l}\right)\circ\vec{B}=-\frac{\mathrm{d}t}{\mathrm{d}t}\oint\limits_{\text{circuit}}\left(\vec{v}\times\mathrm{d}\vec{l}\right)\circ\vec{B}=-\frac{1}{\mathrm{d}t}\oint\limits_{\text{circuit}}\left(\vec{v}\mathrm{d}t\times\mathrm{d}\vec{l}\right)\circ\vec{B} \tag{10.4.159}$$

A small remark is essential at this level. If $\vec{v}$ is the velocity vector of the charges q it cannot be the one which is collinear with $\mathrm{d}\vec{l}$ because otherwise we would have:

$$\vec{v}\mathrm{d}t\times\mathrm{d}\vec{l}=\vec{0} \tag{10.4.160}$$

and therefore e would be zero and this is not possible because it would contradict all the developments made so far! In fact, $\vec{v}$ is the speed of the whole circuit which brings with it all the charges at the same speed!

Thus, during a time $\mathrm{d}t$, the circuit moves by a distance:

$$\mathrm{d}\vec{r}=\vec{v}\mathrm{d}t \tag{10.4.161}$$

vector that is perpendicular to $\mathrm{d}\vec{l}$. Since then:

$$\mathrm{d}\vec{r}\times\mathrm{d}\vec{l} \tag{10.4.162}$$

is the surface (see the properties of the cross product in the section of Vector Calculus) described by the displacement of the element $\mathrm{d}\vec{l}$ on the distance $\mathrm{d}\vec{r}$ such that:

$$\mathrm{d}\vec{r}\times\mathrm{d}\vec{l}=\mathrm{d}^2\vec{S} \tag{10.4.163}$$

We then have:

$$e=-\frac{1}{\mathrm{d}t}\oint\limits_{\text{circuit}}\left(\vec{v}\mathrm{d}t\times\mathrm{d}\vec{l}\right)\circ\vec{B}=-\frac{1}{\mathrm{d}t}\oint\limits_{\text{circuit}}\vec{B}\circ\mathrm{d}^2\vec{S} \tag{10.4.164}$$

We recognize here the expression of the flux (named sometimes "cutted flux") through the elementary surface $\mathrm{d}^2\vec{S}$. Which brings us to write (there is a little bit of intuition - common sense - with the manipulation of differentials but hey that's also physics ...):

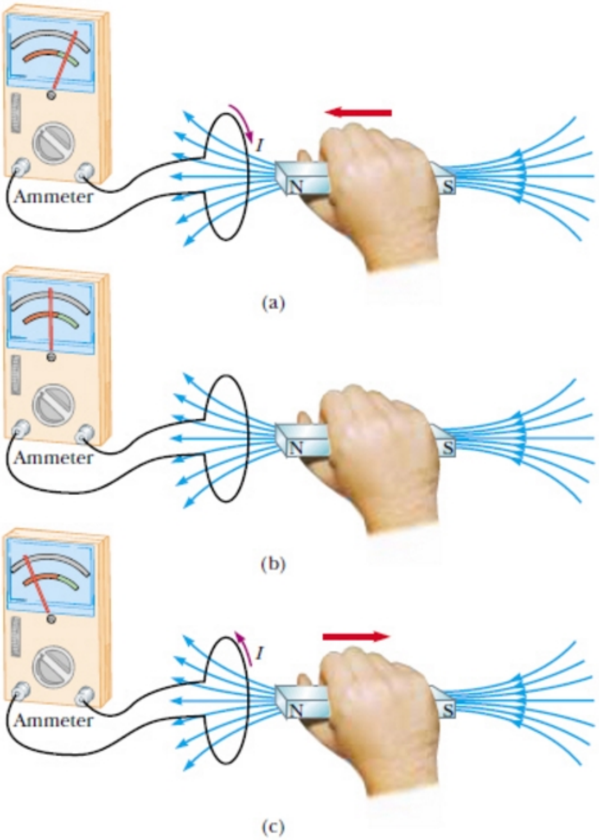
$$e=-\frac{1}{\mathrm{d}t}\oint\limits_{\text{circuit}}\vec{B}\circ\mathrm{d}^2\vec{S}=-\frac{1}{\mathrm{d}t}\oint\limits_{\text{circuit}}\mathrm{d}^2\Phi_c=-\frac{1}{\mathrm{d}t}\mathrm{d}\oint\limits_{\text{circuit}}\mathrm{d}\Phi_c=-\frac{\mathrm{d}\Phi_B}{\mathrm{d}t} \tag{10.4.165}$$

We have just proved the "Faraday's law of induction" in the case of a rigid circuit immersed in a variable magnetic field. We have seen the expression of the cut flux appears naturally. In fact, the only thing that matters is the existence of movement of the overall or a part of the circuit (review the proof to be convinced). Thus, the expression of the induced EMF:

$$\boxed{e=-\frac{\mathrm{d}\Phi_B}{\mathrm{d}t}} \tag{10.4.166}$$

remains valid for a circuit deformed and / or displaced in a static magnetic field. This proof was made from the Lorentz force and is therefore a priori independent of the chosen reference system!

Phenomenologically this can be summarized into the following figure:



(a) When a magnet is moved toward a loop of wire connected to a sensitive ammeter, the ammeter deflects as shown, indicating that a current is induced in the loop.

(b) When the magnet is held stationary, there is no induced current in the loop, even when the magnet is inside the loop

(c) When the magnet is moved away from the loop, the ammeter deflects in the opposite direction, indicating that the induced current is opposite that shown in part (a). Changing the direction of the magnet's motion changes the direction of the current induced by that motion.

10.4.4.1.1 Lenz's law

The statement of "Lenz's law" is as follows: Induction produces effects which oppose the causes which gave rise to it.

This law is, like the rule of maximum flux, already contained in the equations and brings nothing more, except an intuition of the physical phenomena. In this case, Lenz's law is only the expression of the sign "$-$" contained in Faraday's law.

✍**Example:**

If we approach a circuit of the north pole of a magnet, the flux increases and therefore the induced EMF is negative. The induced current will then be negative and will itself produce an induced magnetic field opposite to that of the magnet. Two consequences:

- The increase in flux through the circuit is lessened.
- A negative Laplace force appears (see section Magnetostatics page 4063) $\vec{F}=I\mathrm{d}\vec{l}\times\vec{B}$, opposing to the approach of the magnet.

This sign "$-$" in Faraday's law (Lenz's law) describes the fact that under normal conditions, there is no possible runaway (example: current only increasing).

This is the reason why the Lenz's law is often named the "Lenz-Faraday law".

10.4.4.2 Inductance

So we have:

$$\Phi_B=\int\vec{B}\circ\mathrm{d}\vec{S} \tag{10.4.167}$$

But the Biot-Savart's law gives us (see section Magnetostatics page 4023):

$$\mathrm{d}\vec{B}=\frac{\mu_0}{4\pi}\frac{\mathrm{d}\vec{l}\times\vec{r}}{r^3}\Leftrightarrow\oint\limits_{\Gamma}\mathrm{d}\vec{B}=\frac{\mu_0}{4\pi}I\oint\limits_{\Gamma}\frac{\mathrm{d}\vec{l}\times\vec{r}}{r^3}=\vec{B} \tag{10.4.168}$$

Therefore:

$$\Phi_B=\int_S\left(\frac{\mu_0}{4\pi}I\oint\limits_{\Gamma}\frac{\mathrm{d}\vec{l}\times\vec{r}}{r^3}\right)\circ\mathrm{d}\vec{S}=I\left[\frac{\mu_0}{4\pi}\int_S\left(\oint\limits_{\Gamma}\frac{\mathrm{d}\vec{l}\times\vec{r}}{r^3}\right)\circ\mathrm{d}\vec{S}\right] \tag{10.4.169}$$

which we historically write in condensed form as follows:

$$\boxed{\Phi_B=I\cdot L} \tag{10.4.170}$$

where:

$$L=\frac{\mu_0}{4\pi}\int_S\left(\oint\limits_{\Gamma}\frac{\mathrm{d}\vec{l}\times\vec{r}}{r^3}\right)\circ\mathrm{d}\vec{S} \tag{10.4.171}$$

is named the "self-induction coefficient" or "self-inductance", or simply "self" or commonly "inductance", expressed in "Henry" [H]. It only depends on the geometric properties of the circuit and is necessarily positive.

Explicitly this unit can be written in various forms but the most common one are:

$$\mathrm{H}=\frac{\mathrm{kg}\cdot\mathrm{m}^2}{\mathrm{s}^2\cdot\mathrm{A}^2}=\frac{\mathrm{kg}\cdot\mathrm{m}^2}{\mathrm{C}^2}=\frac{\mathrm{J}}{\mathrm{A}^2}=\frac{\mathrm{T}\cdot\mathrm{m}^2}{\mathrm{A}}=\frac{\mathrm{V}\cdot\mathrm{s}}{\mathrm{A}} \tag{10.4.172}$$

With the laws that we have stated so far, we are able to study some variable regimes (also named "transcient regimes"). Indeed, all reasoning based on the notion of a field (electric or magnetic) constant over time can easily be applied to variable physical systems (time-dependent fields), provided that this variability takes place on long times scales compared to the characteristic adjustment time of the field in the corresponding materials.

Here is a concrete case:

Most laws of magnetostatics assume a permanent current, that is, the same throughout the circuit. When we close a switch, an electromagnetic signal propagates throughout the circuit and this is how a permanent current can be established: it takes a time of the order of l/c where l is the length of the circuit and c the speed of light. If we now have a sinusoidal voltage generator with period T (this is just an example ... taken randomly ...), then we can still use the relations deduced from magnetostatic if:

$$T\gg l/c \tag{10.4.173}$$

Thus, although the current is variable, the creation of a magnetic field will obey the Biot-Savart law as long as the above criterion remains satisfied. This type of variable regime is named "quasi static regime" in the sense that it is transient.

So, since we have:

$$e=-\frac{\mathrm{d}\Phi_B}{\mathrm{d}t}\quad\text{and}\quad\Phi_B=LI \tag{10.4.174}$$

we then have if and only if the current is variable in the circuit:

$$\boxed{e=-\frac{\mathrm{d}\Phi_B}{\mathrm{d}t}=-L\frac{\mathrm{d}i}{\mathrm{d}t}} \tag{10.4.175}$$

L being constant for a rigid circuit. The self therefore generates an electromotive force opposite to that generated by the current at its terminals. This electromotive force therefore has an opposite direction to that of the electric generator.

In the case where an inductor is alone in series with a sinusoidal alternating current generator (fairly typical case in the industrial world of the 119th and 120th century), then the conservation of the field will give:

$$u+e=0 \tag{10.4.176}$$

And therefore explicitly in the typical case of a sine regime:

$$U_0\sin(\omega t)-L\frac{\mathrm{d}i}{\mathrm{d}t}=0 \tag{10.4.177}$$

And therefore:

$$\mathrm{d}i=\frac{U_0}{L}\sin(\omega t)\mathrm{d}t \tag{10.4.178}$$

To get the current, we integrate:

$$i(t)=\frac{U_0}{L}\int\sin(\omega t)=-\frac{U_0}{\omega L}\sin\left(\omega t-\frac{\pi}{2}\right) \tag{10.4.179}$$

So in this special case, the current is behind the voltage (or the voltage is ahead of the current) by a quarter of a period. It is also customary to write:

$$I_0=\frac{U_0}{\omega L} \tag{10.4.180}$$

Therefore:

$$U_0=\omega L I_0=X_L I_0 \tag{10.4.181}$$

where by analogy with pure resistance, the expression:

$$X_L=\omega L \tag{10.4.182}$$

is named "inductive reactance".

Remark

We see well in the relation obtained above that in stationary regime, if the current is constant, then the electromotive force is zero and the self then behaves like a simple equipotential!

We must now give an important and simple example of Lenz's law at the same time by applying it to the calculation of the inductance of a solenoid of radius r (the inductance of a toric solenoid with circular section having already been treated in the section of Magnetostatics). We proved in the section of Magnetostatics that the magnetic field in a solenoid was given by:

$$B=\mu\frac{IN}{\ell} \tag{10.4.183}$$

where as a reminder N is the number of turns and l is the length of the solenoid. We saw above that Faraday's law of induction was given by:

$$e=-\frac{1}{\mathrm{d}t}\mathrm{d}\oint\limits_{\text{circuit}}\mathrm{d}\Phi_c=-\frac{\mathrm{d}\Phi_B}{\mathrm{d}t}=-L\frac{\mathrm{d}i}{\mathrm{d}t} \tag{10.4.184}$$

and in the case of a turn we will travel N times the path of the integral. It then comes:

$$e=-N\frac{\mathrm{d}\Phi}{\mathrm{d}t}=-L\frac{\mathrm{d}i}{\mathrm{d}t} \tag{10.4.185}$$

We saw above that the flux of the magnetic field was given by (if the field is perpendicular to the crossed surface):

$$\Phi=B\cdot S \tag{10.4.186}$$

Therefore:

$$\Phi=B\cdot S=\left(\mu\frac{IN}{\ell}\right)S=\frac{\mu SN}{\ell}I \tag{10.4.187}$$

Remark

Caution!! The flux in a solenoid is not equal to the flux in a turn multiplied by the number of turns of the solenoid!

The rate of change of the magnetic flux is found by derivation, that is:

$$N\frac{\mathrm{d}\Phi_B}{\mathrm{d}t}=N\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\mu SN}{\ell}I\right)=\frac{\mu SN^2}{I}\frac{\mathrm{d}I(t)}{\mathrm{d}t}=\frac{\mu SN^2}{\ell}\frac{\mathrm{d}i}{\mathrm{d}t} \tag{10.4.188}$$

Or in the case of circular turns:

$$N\frac{\mathrm{d}\Phi_B}{\mathrm{d}t}=\frac{\mu SN^2}{l}\frac{\mathrm{d}i}{\mathrm{d}t}=\frac{\mu(2\pi R^2)N^2}{\ell}\frac{\mathrm{d}i}{\mathrm{d}t} \tag{10.4.189}$$

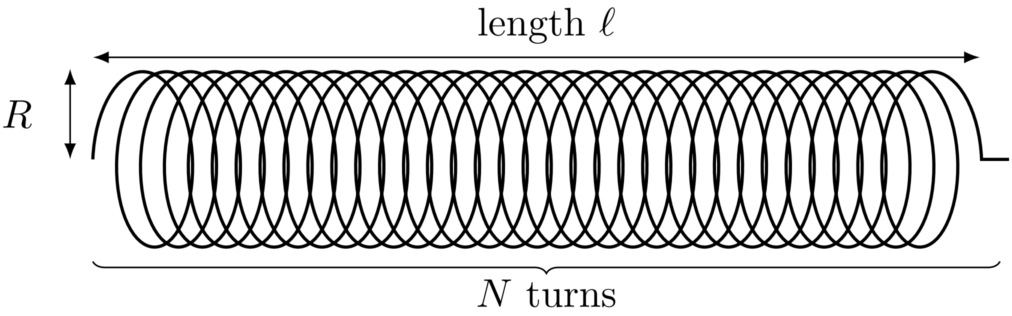


Figure 10.98: **Simple solenoid (author: Izaak Neutelings)**

The electromotive force generated is then:

$$e=-N\frac{\mathrm{d}\Phi_B}{\mathrm{d}t}=-L\frac{\mathrm{d}i}{\mathrm{d}t}=-\left(\frac{\mu(\pi R^2)N^2}{\ell}\right)\frac{\mathrm{d}i}{\mathrm{d}t} \tag{10.4.190}$$

and therefore by correspondence in vacuum:

$$L=\frac{\mu(\pi r^2)N^2}{l}\cong 4\pi^2\cdot 10^{-7}\frac{N^2r^2}{l} \tag{10.4.191}$$

Now calculate the instantaneous power received by a coil. We have proved above that we still have in our case study and if we modelize the inductance as a non-ideal dipole:

$$p=\Delta u\cdot i=u\cdot i=(r\cdot i-e)i=r\cdot i^2-e\cdot i \tag{10.4.192}$$

where the lowercase letters indicate as always that we are in a non-constant regime:

$$p=r\cdot i^2-e\cdot i=r\cdot i^2-\left(-L\frac{\mathrm{d}i}{\mathrm{d}t}\right)i=r\cdot i^2+L\frac{\mathrm{d}i}{\mathrm{d}t}i=p_j+\frac{\mathrm{d}\left(\dfrac{1}{2}Li^2\right)}{\mathrm{d}t} \tag{10.4.193}$$

At the contrary to the development which we had made in the section of Electrostatics for the same calculation with regard to the capacitor, we have not neglected here the dissipation of energy by Joule effect. But the reader must be aware that in most cases this term is also neglected!

So by integration in a given time interval from 0 to t we have for the second term:

$$E_L=\frac{1}{2}Li^2 \tag{10.4.194}$$

This energy is therefore always positive and is stored in electrostatic form in the inductance.

Or explicitly in the case of a circular solenoid:

$$E_L=\frac{1}{2}Li^2\cong\frac{1}{2}\left(4\pi^2\cdot 10^{-7}\frac{N^2R^2}{\ell}\right)i^2 \tag{10.4.195}$$

And since for a Solenoid $B=\mu NI/\ell$, we have in vacuum:

$$E_L=\frac{1}{2}\left(\frac{\mu_0 N^2}{\ell}\pi R^2\right)\left(\frac{B\ell}{\mu_0 N}\right)^2=\frac{1}{2}\frac{1}{\mu_0}\left(\pi R^2\ell\right)B^2 \tag{10.4.196}$$

When i decreases, the coil restores this energy. So we cannot store energy in an isolated coil unlike a capacitor.

Under a sinusoidal regime, the average power will be zero. We can generalize this by admitting that a perfect inductance does not dissipate any power by the Joule effect.

Let us now determine the inductance of a coaxial structure (similarly as we did of the capacitance of such a cable at page 3997) as shown in the figure below:

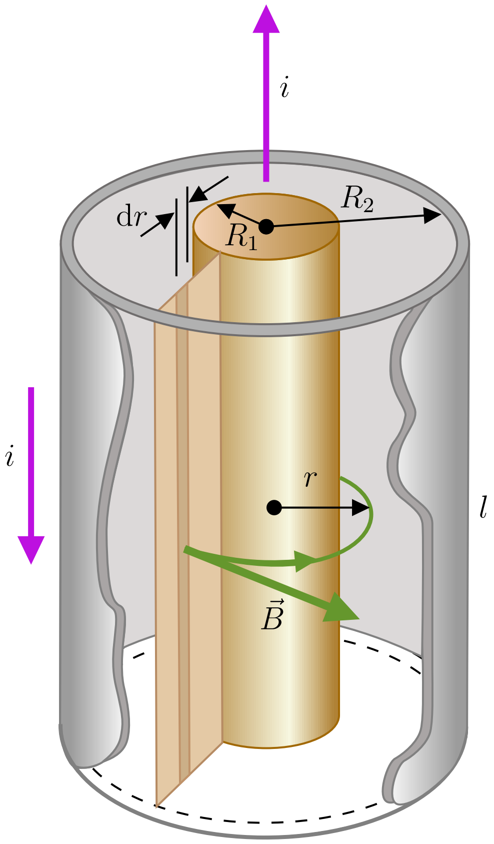


Figure 10.99: **Section of a long coaxial cable**

where the inner and outer conductors carry equal currents in opposite directions.

The inductance of this structure is of interest for a number of reasons – in particular, for determining the characteristic impedance of coaxial transmission line.

We must find the magnetic flux through the light gold rectangle in the figure above. Ampère's law tells us that the magnetic field in the region between the conductors is due to the inner conductor alone and that its magnitude is:

$$B=\frac{\mu i}{2\pi r} \tag{10.4.197}$$

where r is measured from the common center of the cylinders. A sample circular field line is shown in the figure above, along with a field vector tangent to the field line.

The magnetic field is perpendicular to the light gold rectangle of length l, and width R_2-R_1, the cross section of interest.

Because the magnetic field varies with radial position across this rectangle, we must use calculus to find the total magnetic flux.

Divide the light gold rectangle into strips of width $\mathrm{d}r$ such as the darker strip in the figure above. We evaluate the magnetic flux through such a strip:

$$\mathrm{d}\Phi=B\mathrm{d}S=Bl\mathrm{d}r \tag{10.4.198}$$

Substitute for the magnetic field and integrate over the entire light gold rectangle gives us:

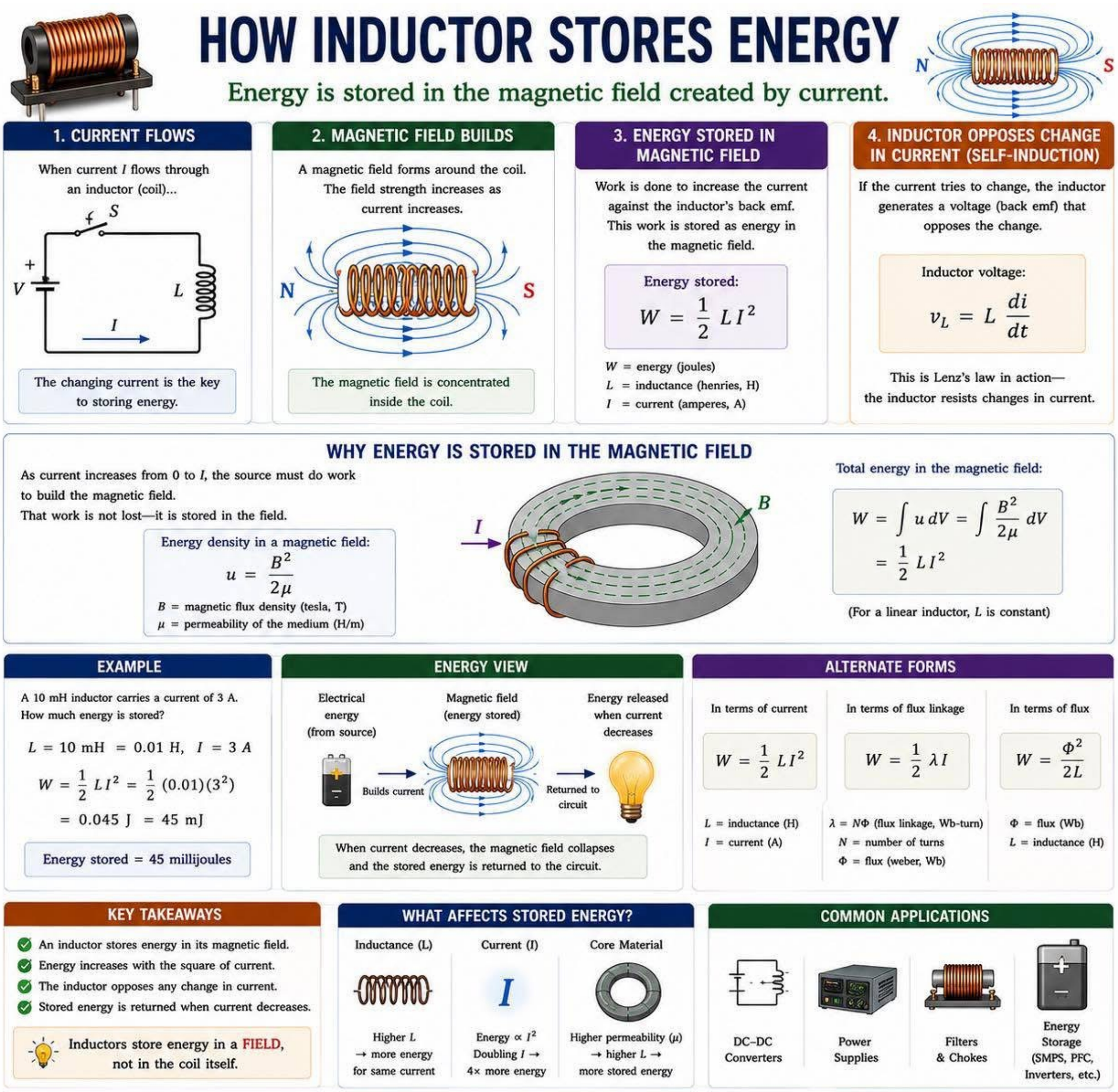
$$\Phi=\int\limits_{R_1}^{R_2}\frac{\mu i}{2\pi r}l\mathrm{d}r=\frac{\mu il}{2\pi r}\int\limits_{R_1}^{R_2}\frac{1}{r}\mathrm{d}r=\frac{\mu il}{2\pi r}\ln\left(\frac{R_2}{R_1}\right) \tag{10.4.199}$$

Using the definition of inductance we then get:

$$\boxed{L=\frac{\Phi_B}{i}=\frac{\mu l}{2\pi r}\ln\left(\frac{R_2}{R_1}\right)} \tag{10.4.200}$$

and therefore we can also define the "coaxial inductance per unit length":

$$L=\frac{\Phi_B}{I}=\frac{\mu}{2\pi r}\ln\left(\frac{R_2}{R_1}\right) \tag{10.4.201}$$



HOW INDUCTOR STORES ENERGY

Energy is stored in the magnetic field created by current.

1. CURRENT FLOWS

When current I flows through an inductor (coil)...

The changing current is the key to storing energy.

2. MAGNETIC FIELD BUILDS

A magnetic field forms around the coil. The field strength increases as current increases.

The magnetic field is concentrated inside the coil.

3. ENERGY STORED IN MAGNETIC FIELD

Work is done to increase the current against the inductor's back emf. This work is stored as energy in the magnetic field.

Energy stored:

$$W = \frac{1}{2} L I^2$$

W = energy (joules)
L = inductance (henries, H)
I = current (amperes, A)

4. INDUCTOR OPPOSES CHANGE IN CURRENT (SELF-INDUCTION)

If the current tries to change, the inductor generates a voltage (back emf) that opposes the change.

Inductor voltage:

$$v_L = L \frac{di}{dt}$$

This is Lenz's law in action—the inductor resists changes in current.

WHY ENERGY IS STORED IN THE MAGNETIC FIELD

As current increases from 0 to I, the source must do work to build the magnetic field.
That work is not lost—it is stored in the field.

Energy density in a magnetic field:

$$u = \frac{B^2}{2\mu}$$

B = magnetic flux density (tesla, T)
μ = permeability of the medium (H/m)

Total energy in the magnetic field:

$$W = \int u \, dV = \int \frac{B^2}{2\mu} \, dV = \frac{1}{2} L I^2$$

(For a linear inductor, L is constant)

EXAMPLE

A 10 mH inductor carries a current of 3 A. How much energy is stored?

$$L = 10 \text{ mH} = 0.01 \text{ H}, \quad I = 3 \ A$$

$$W = \frac{1}{2} L I^2 = \frac{1}{2} (0.01)(3^2)$$

$$= 0.045 \text{ J} = 45 \text{ mJ}$$

Energy stored = 45 millijoules

ENERGY VIEW

Electrical energy (from source) → Builds current → Magnetic field (energy stored) → Returned to circuit → Energy released when current decreases

When current decreases, the magnetic field collapses and the stored energy is returned to the circuit.

ALTERNATE FORMS

In terms of current	In terms of flux linkage	In terms of flux
$W = \frac{1}{2} L I^2$	$W = \frac{1}{2} \lambda I$	$W = \frac{\Phi^2}{2L}$
L = inductance (H) I = current (A)	$\lambda = N\Phi$ (flux linkage, Wb-turn) N = number of turns Φ = flux (weber, Wb)	Φ = flux (Wb) L = inductance (H)

KEY TAKEAWAYS

- ✓ An inductor stores energy in its magnetic field.
- ✓ Energy increases with the square of current.
- ✓ The inductor opposes any change in current.
- ✓ Stored energy is returned when current decreases.

Inductors store energy in a FIELD, not in the coil itself.

WHAT AFFECTS STORED ENERGY?

Inductance (L)	Current (I)	Core Material
Higher L → more energy for same current	Energy ∝ I^2 Doubling I → 4× more energy	Higher permeability (μ) → higher L → more stored energy

COMMON APPLICATIONS

- DC–DC Converters
- Power Supplies
- Filters & Chokes
- Energy Storage (SMPS, PFC, Inverters, etc.)

10.4.5 Electromagnetic Waves in Conductors

The study of electromagnetic waves in conductors is very important as it helps to understand and model some effects of current propagation in real life common conductors that are a priori not intuitive like:

1. The fact that the current and the propagation of an electric field along a standard conductor line are the two faces of a same coin
2. Understand why the current is concentrated mainly near the surface of the (standard) conductor and not at its center
3. The fact the amplitude of an incident electromagnetic wave entering in a standard conductor is quickly attenuated.
4. The very low phase and group velocity of an electromagnetic wave inside a standard conductor that carry the electrons of the current.

OK let's start with the maths!

10.4.5.1 Skin effect

The "skin effect" or "pellicular effect" (or, more rarely, "Kelvin effect") is an electromagnetic phenomenon which makes that at high frequency a current tends to circulate only at the surface of the conductors. This phenomenon of electromagnetic origin exists for all the conductors traversed by alternating currents. It causes the current density to decrease as one moves away from the periphery of the conductor. This results in an increase in the resistance of the conductor.

This effect can be used to lighten the weight of high frequency transmission lines by using tubular conductors, or even pipes, without (too much) current loss. It is also used in the electromagnetic shielding of the coaxial wires by surrounding them with a thin metal case which keeps the currents induced by the high ambient frequencies on the outside of the cable.

What we would now like to determine is the attenuation of the electric field (or an attenuation coefficient) in the material of a solid cylindrical conductive cable of conductivity σ and permeability ε:

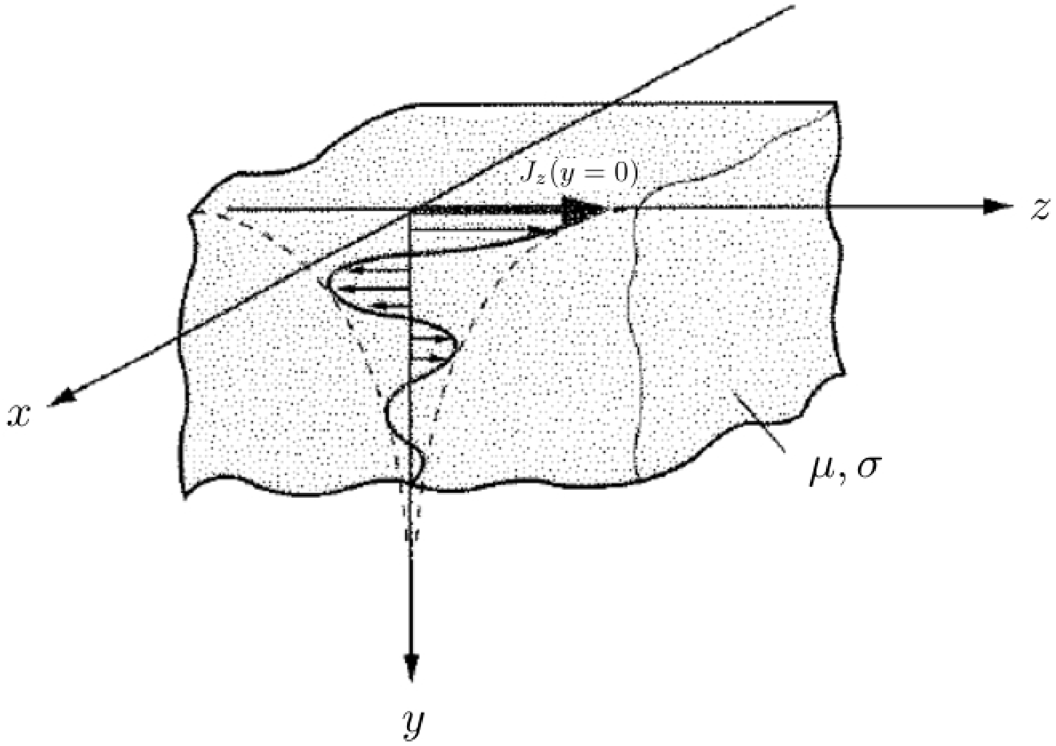


Figure 10.100: **Configuration study for skin effect (source: Introductory Electromagnetics, Z. and B.D. Popovic**

and assuming that the current density vector is parallel to the boundary surface, and that is has a single component, for example, $\vec{J} = J_z \vec{e}_z$, depending on the coordinate y (the distance from the interface) only. We wish to determine the distribution of current in the conducting half-space.

To do this, we take the fourth Maxwell's equation in the form given previously:

$$\vec{\nabla} \times \vec{H} = \vec{j} + \frac{\mathrm{d}\vec{D}}{\mathrm{d}t} = \sigma \vec{E} + \frac{\mathrm{d}\vec{D}}{\mathrm{d}t} \qquad (10.4.202)$$

and we assume to work with a conductor which has no capacitive effect (i.e. no displacement current), contrary to the general case demonstrated in the Electrodynamics section, the latter relation then reduces to:

$$\vec{\nabla} \times \vec{H} = \frac{1}{\mu} \vec{\nabla} \times \vec{B} = \sigma \vec{E} \qquad (10.4.203)$$

and if we associate it with Maxwell's third equation (see section Electrodynamics page 4101), and assuming a harmonic excitation, we have for recall:

$$\vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} = -\mathrm{i}\omega \vec{B} \qquad (10.4.204)$$

Therefore, from that latter:

$$\vec{\nabla} \times \vec{j} = -\mathrm{i}\omega\sigma \vec{B} \qquad (10.4.205)$$

So we have so far:

$$\vec{\nabla} \times \vec{j} = -\mathrm{i}\omega\sigma \vec{B} \qquad \vec{\nabla} \times \vec{B} = \mu \vec{j} \qquad (10.4.206)$$

We assumed the current density vector has only a z component, which depends only on y. From the Biot-Savart law and symmetry it therefore follows that there is only an x component of the vector $\vec{B}$ (the reader can request we put the details of the section of Magnetism here again if necessary). According to the expression for the curl in a rectangular coordinates system, the both previous relations become:

$$\frac{\mathrm{d}j_z}{\mathrm{d}y}=-\mathrm{i}\omega\sigma B_x \qquad -\frac{\mathrm{d}B_x}{\mathrm{d}y}=\mu j_z \tag{10.4.207}$$

where we use ordinary derivatives (not partial) because j_z and B_x depend only on y. From the both relations above we can eliminate B_x to get an equation in j_z:

$$\frac{\mathrm{d}^2 j_z}{\mathrm{d}y^2}=\mathrm{i}\omega\mu\sigma j_z \tag{10.4.208}$$

Let us place ourselves in the important case of a harmonic regime:

$$j_z(y,t)=j_0(y)\cos(ky-\omega t) \tag{10.4.209}$$

and let us use temporarily the phasors notation:

$$j_z(y,t)=j_0e^{\mathrm{i}(ky-\omega t)}=j_0e^{\mathrm{i}ky}e^{-\mathrm{i}\omega t} \tag{10.4.210}$$

We then have:

$$\begin{aligned}\frac{\partial^2 j_z}{\partial y^2}&=\frac{\partial\left(j_0e^{\mathrm{i}ky}\right)}{\partial y^2}e^{-\mathrm{i}\omega t}\\ \frac{\partial j_z}{\partial t}&=-\mathrm{i}\omega j_0e^{\mathrm{i}ky}e^{-\mathrm{i}\omega t}\end{aligned} \tag{10.4.211}$$

By injecting this into the previous differential equation and simplifying, we then get:

$$\frac{\partial^2 e^{\mathrm{i}ky}}{\partial y^2}=-\mathrm{i}\omega\sigma\mu e^{\mathrm{i}ky} \tag{10.4.212}$$

Thus:

$$-k^2=-\mathrm{i}\omega\sigma\mu \tag{10.4.213}$$

and therefore:

$$k=\pm\sqrt{\mathrm{i}\omega\sigma\mu} \tag{10.4.214}$$

remembering (see section Numbers page 243) that:

$$\sqrt{\mathrm{i}}=\sqrt{e^{\mathrm{i}\frac{\pi}{2}}}=\left(e^{\mathrm{i}\frac{\pi}{2}}\right)^{\frac{1}{2}}=e^{\mathrm{i}\frac{\pi}{4}}=\cos\left(\frac{\pi}{4}\right)-\mathrm{i}\sin\left(\frac{\pi}{4}\right)=\frac{1-\mathrm{i}}{\sqrt{2}} \tag{10.4.215}$$

it comes:

$$k=\pm\frac{1-\mathrm{i}}{\sqrt{2}}\sqrt{\omega\sigma\mu}=\pm(1-\mathrm{i})\sqrt{\frac{\omega\sigma\mu}{2}} \tag{10.4.216}$$

Therefore:

$$j_z(y,t)=j_0(y)e^{\mathrm{i}ky}e^{-\mathrm{i}\omega t}=j_0e^{\pm(1-\mathrm{i})\sqrt{\frac{\omega\sigma\mu}{2}}y}e^{-\mathrm{i}\omega t}=j_0e^{\pm(\mathrm{i}+1)\sqrt{\frac{\omega\sigma\mu}{2}}y}e^{-\mathrm{i}\omega t} \tag{10.4.217}$$

Thus:

$$j_z(y,t)=j_0e^{\pm\sqrt{\frac{\omega\sigma\mu}{2}}y}e^{\mathrm{i}\left(\pm\sqrt{\frac{\omega\sigma\mu}{2}}y-\omega t\right)} \tag{10.4.218}$$

We must reject for physical reasons (conservation of energy) the solution:

$$k=+\sqrt{\mathrm{i}\omega\sigma\mu} \tag{10.4.219}$$

Therefore it remains:

$$\boxed{j_z(y,t)=j_0e^{-\sqrt{\frac{\omega\sigma\mu}{2}}z}e^{-\mathrm{i}\left(\sqrt{\frac{\omega\sigma\mu}{2}}y+\omega t\right)}} \tag{10.4.220}$$

that physicists write:

$$j_z(y,t)=j_0e^{-\frac{y}{\delta}}e^{\mathrm{i}\left(-\frac{y}{\delta}-\omega t\right)} \tag{10.4.221}$$

because the units of δ are meters (parameter that is zero if the electric field is constant) and is assimilated to the "attenuation coefficient" that we had set ourselves to determine at the beginning:

$$\delta=\sqrt{\frac{2}{\omega\sigma\mu}}=\sqrt{\frac{2\sigma}{\omega\mu}} \tag{10.4.222}$$

For a copper conductor, we have according to Wikipedia the following values:

Frequency	**δ [mm]**
50 [Hz]	9.38
60 [Hz]	8.57
10 [kHz]	0.66
100 [kHz]	0.21
1 [MHz]	0.066
10 [MHz]	0.021

Table 10.6: **Somme attenuation skin effect factor for Copper (source: Wikipedia)**

10.4.5.2 Phase and group velocity of EM waves inside conductors (modified Maxwell equations)

Let us remind that three of the Maxwell equations are:

$$\vec{\nabla}\circ\vec{D}=\rho \quad \vec{\nabla}\circ\vec{B}=0$$
$$\vec{\nabla}\times\vec{E}=-\frac{\partial\vec{B}}{\partial t} \tag{10.4.223}$$

And in conductors the fourth Maxwell equation is written:

$$\vec{\nabla}\times\vec{B}=\mu\vec{j}+\mu\varepsilon\frac{\partial\vec{E}}{\partial t} \tag{10.4.224}$$

If we explicit the first term on the right of the quality using Ohm's law:

$$\vec{\nabla}\times\vec{B}=\mu\sigma\vec{E}+\mu\varepsilon\frac{\partial\vec{E}}{\partial t} \tag{10.4.225}$$

Or even more explicit assuming $\mu_r=1$:

$$\vec{\nabla}\times\vec{B}=\mu_0\sigma\vec{E}+\mu_0\varepsilon_r\varepsilon_0\frac{\partial\vec{E}}{\partial t} \tag{10.4.226}$$

Let us take the curl of:

$$\vec{\nabla}\times\vec{E}=-\frac{\partial\vec{B}}{\partial t} \tag{10.4.227}$$

Therefore (we use on the right term the identity proved in the section of Vector Calculus page 1911):

$$\vec{\nabla}\times(\vec{\nabla}\times\vec{E})=-\frac{\partial}{\partial t}(\vec{\nabla}\times\vec{B})=\vec{\nabla}(\vec{\nabla}\circ\vec{E})-\vec{\nabla}^2\vec{E} \tag{10.4.228}$$

Taking the gradient of :

$$\vec{\nabla}\circ\vec{D}=\rho\Rightarrow\vec{\nabla}\rho_c=\varepsilon_r\varepsilon_0\vec{\nabla}(\vec{\nabla}\circ\vec{E})=\vec{0} \tag{10.4.229}$$

Hence under this assumption:

$$\vec{\nabla}(\vec{\nabla}\circ\vec{E})=\vec{0} \tag{10.4.230}$$

Therefore:

$$\vec{\nabla}\times(\vec{\nabla}\times\vec{E})=-\frac{\partial}{\partial t}(\vec{\nabla}\times\vec{B})=\cancel{\vec{\nabla}(\vec{\nabla}\circ\vec{E})}-\vec{\nabla}^2\vec{E} \tag{10.4.231}$$

Hence:

$$\vec{\nabla}^2\vec{E}=\frac{\partial}{\partial t}(\vec{\nabla}\times\vec{B}) \tag{10.4.232}$$

Injecting in the latter:

$$\vec{\nabla}\times\vec{B}=\mu\sigma\vec{E}+\mu\varepsilon\frac{\partial\vec{E}}{\partial t} \tag{10.4.233}$$

we get :

$$\vec{\nabla}^2\vec{E}=\frac{\partial}{\partial t}(\vec{\nabla}\times\vec{B})=\mu_0\varepsilon_r\varepsilon_0\frac{\partial^2\vec{E}}{\partial t^2}+\mu_0\sigma\frac{\partial\vec{E}}{\partial t} \tag{10.4.234}$$

Hence we have the "modified wave equation":

$$\boxed{\vec{\nabla}^2\vec{E}=\mu_0\varepsilon_r\varepsilon_0\frac{\partial^2\vec{E}}{\partial t^2}+\mu_0\sigma\frac{\partial\vec{E}}{\partial t}} \tag{10.4.235}$$

As $\sigma\rightarrow 0$, we recover the wave equation for linear media as we would expect.

Remark

Respectively from $\vec{\nabla}\times(\vec{\nabla}\times\vec{B})$ we get:

$$\vec{\nabla}^2\vec{B}=\mu\varepsilon\frac{\partial^2\vec{B}}{\partial t^2}+\mu\sigma\frac{\partial\vec{B}}{\partial t} \tag{10.4.236}$$

The solution can be written as an attenuated plane wave:

$$E_x=E_0e^{\mathrm{i}(\beta z-\omega t)}e^{-\alpha z}=E_0e^{\mathrm{i}((\beta+\mathrm{i}\alpha)z-\omega t)}=E_0e^{\mathrm{i}(kz-\omega t)} \tag{10.4.237}$$

Remark

If an electromagnetic wave hits a conductor (starting in air or vacuum, say), the wave will attenuate as it penetrates the conductor with a characteristic distance. The attenuation length in a conductor is known as the "skin depth" and defined by:

$$\delta=\frac{1}{\alpha}$$

The skin depth refers therefore to the point at which the field reaches approximately 37% (i.e. $1/e$) of its nominal value. Where did the missing corresponding energy go? Part of it was reflected (high conductivity materials reflect most of the energy) and part of it was absorbed and turned into heat. (Heating occurs because the electromagnetic field accelerates electrons, and those electrons lose energy through resistance, that is, collisions with other particles.)

In geophysics, skin depth is primarily applied in two techniques: electromagnetic conductivity surveys and ground-penetrating radar (GPR).

Where therefore:

$$k:=\beta+\mathrm{i}\alpha \tag{10.4.238}$$

Substituting $E_0e^{\mathrm{i}(\beta z-\omega t)}e^{-\alpha z}$ back into the modified wave equation:

$$(\mathrm{i}\beta-\alpha)^2E_0=\mu_0\varepsilon_r\varepsilon_0(-\mathrm{i}\omega)^2E_0+\mu_0\sigma(-\mathrm{i}\omega)E_0 \tag{10.4.239}$$

Expanded, rearranged and simplified:

$$(-\beta^2-2\mathrm{i}\beta\alpha+\alpha^2)=-\mu_0\varepsilon_r\varepsilon_0\omega^2-\mathrm{i}\mu_0\sigma\omega \tag{10.4.240}$$

Equating the real and imaginary parts:

$$-\beta^2+\alpha^2=-\mu_0\varepsilon_r\varepsilon_0\omega^2 \qquad -2\beta\alpha=-\mu_0\sigma\omega \tag{10.4.241}$$

Often denoted:

$$-\beta^2+\alpha^2=-\mu\varepsilon\omega^2 \qquad 2\beta\alpha=\mu\sigma\omega \tag{10.4.242}$$

Hence:

$$\beta=\frac{\mu\sigma\omega}{2\alpha} \qquad -\left(\frac{\mu\sigma\omega}{2\alpha}\right)^2+\alpha^2=-\mu\varepsilon\omega^2 \tag{10.4.243}$$

Let us focus on the right equality, i.e.:

$$-\left(\frac{\mu\sigma\omega}{2\alpha}\right)^2+\alpha^2=-\mu\varepsilon\omega^2 \tag{10.4.244}$$

We rearrange:

$$-\left(\frac{\mu\sigma\omega}{2\alpha}\right)^2+\alpha^2+\mu\varepsilon\omega^2=0 \tag{10.4.245}$$

Let us multiply both sides by α^2 and rearrange to get a quadratic equation>:

$$\alpha^4+\mu\varepsilon\omega^2\alpha^2-\left(\frac{\mu\sigma\omega}{2}\right)^2=0 \tag{10.4.246}$$

We do a change of variable $\alpha^2=x$ and we have therefore a simple square polynomial:

$$x^2+\mu\varepsilon\omega^2 x-\left(\frac{\mu\sigma\omega}{2}\right)^2=0 \tag{10.4.247}$$

And the roots of the previous quadratic equation are immediate:

$$\alpha=\sqrt{\frac{-\mu\varepsilon\omega^2\pm\sqrt{\left(\mu\varepsilon\omega^2\right)^2+(\mu\sigma\omega)^2}}{2}} \tag{10.4.248}$$

Because we enforce that α is real and positive, and all of the parameters within the square root are positive, we need to choose the + sign:

$$\begin{aligned}
&\alpha=\sqrt{\frac{\sqrt{\left(\mu\varepsilon\omega^2\right)^2+(\mu\sigma\omega)^2}+\mu\varepsilon\omega^2}{2}}\\
&\rightarrow\alpha=\sqrt{\frac{\mu\varepsilon\omega^2}{2}}\sqrt{\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}-1}\\
&\rightarrow\alpha=\omega\sqrt{\frac{\varepsilon\mu}{2}}\left[\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}-1\right]^{1/2}
\end{aligned} \tag{10.4.249}$$

This is the imaginary part of k:

$$k:=\beta+\mathrm{i}\alpha=\beta+\mathrm{i}\omega\sqrt{\frac{\varepsilon\mu}{2}}\left[\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}-1\right]^{1/2} \tag{10.4.250}$$

Now remember that we have:

$$-\beta^2+\alpha^2=-\mu\varepsilon\omega^2 \tag{10.4.251}$$

Hence:

$$\begin{aligned}\beta=\sqrt{\mu\varepsilon\omega^2+\alpha^2}&=\sqrt{\mu\varepsilon\omega^2+\omega^2\frac{\varepsilon\mu}{2}\left[\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}-1\right]}\\&=\sqrt{\frac{\mu\varepsilon\omega^2}{2}+\omega^2\frac{\varepsilon\mu}{2}\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}}\\&=\omega\sqrt{\frac{\varepsilon\mu}{2}}\left[\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}+1\right]^{1/2}\end{aligned} \tag{10.4.252}$$

As a check, if (insulator):

$$\sigma\rightarrow 0,\alpha\rightarrow 0,k\rightarrow\omega\sqrt{\frac{\varepsilon\mu}{2}}\sqrt{2}=\omega\sqrt{\varepsilon\mu}=\frac{\omega}{v} \tag{10.4.253}$$

and we are back to the linear medium!

For a perfect conductor:

$$\sigma\rightarrow+\infty,\alpha\rightarrow+\infty \tag{10.4.254}$$

and therefore there are no electromagnetic waves and $E,B=0$.

The skin depth is therefore given by:

$$\delta=\frac{1}{\alpha}=\frac{1}{\omega\sqrt{\frac{\varepsilon\mu}{2}}\left[\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}+1\right]^{1/2}} \tag{10.4.255}$$

Let us remind the expression of the phase velocity (see section Mechanics page 3480):

$$v_\phi=\frac{\omega}{k}=\frac{\omega}{\Re(k)} \tag{10.4.256}$$

and of the group velocity (see section Mechanics page 3481):

$$v_g=\frac{\Delta\omega}{\Delta k}=\frac{1}{\mathrm{d}\Re(k)/\mathrm{d}\omega}=\left(\frac{\mathrm{d}\Re(k)}{\mathrm{d}\omega}\right)^{-1} \tag{10.4.257}$$

Therefore in our case:

$$v_\phi=\frac{\omega}{\Re(k)}=\frac{1}{\beta}=\frac{1}{\sqrt{\frac{\varepsilon\mu}{2}}\left[\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}+1\right]^{1/2}} \tag{10.4.258}$$

and:

$$v_g=\left(\frac{\mathrm{d}\Re(k)}{\mathrm{d}\omega}\right)^{-1}=\left(\frac{\mathrm{d}}{\mathrm{d}\omega}\omega\sqrt{\frac{\varepsilon\mu}{2}}\left[\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}+1\right]^{1/2}\right) \tag{10.4.259}$$

So let's work out what the group speed v_g:

$$\begin{aligned}\frac{\mathrm{d}k}{\mathrm{d}\omega}&=\sqrt{\frac{\varepsilon\mu}{2}}\left[\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}+1\right]^{1/2}+\omega\sqrt{\frac{\varepsilon\mu}{2}}\frac{\mathrm{d}}{\mathrm{d}\omega}\left[\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}+1\right]^{1/2}\\&=\sqrt{\frac{\varepsilon\mu}{2}}\left[\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}+1\right]^{1/2}+\omega\sqrt{\frac{\varepsilon\mu}{2}}\cdot\frac{\frac{1}{2}\cdot\frac{\cancel{1}}{\cancel{2}}\cdot\cancel{2}\cdot\left(\frac{\sigma}{\varepsilon}\right)^2\left(-\frac{1}{\omega^3}\right)}{\left[\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}+1\right]^{1/2}\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}}\\&=\sqrt{\frac{\varepsilon\mu}{2}}\left[\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}+1\right]^{1/2}-\frac{1}{2}\sqrt{\frac{\varepsilon\mu}{2}}\cdot\frac{\left(\frac{\sigma}{\varepsilon\omega}\right)^2}{\left[\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}+1\right]^{1/2}\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}}\\&=\sqrt{\frac{\varepsilon\mu}{2}}\left[\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}+1\right]^{1/2}\cdot\left\{1-\frac{1}{2}\frac{\left(\frac{\sigma}{\varepsilon\omega}\right)^2}{\left[\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}+1\right]\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}}\right\}\end{aligned}\tag{10.4.260}$$

The group speed of an electromagnetic plane wave propagating in a conductor is (eeek!!!):

$$v_g(\omega)=\left[\frac{\mathrm{d}k(\omega)}{\mathrm{d}\omega}\right]^{-1}=\frac{1}{\sqrt{\frac{\varepsilon\mu}{2}}\left[\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}+1\right]^{1/2}}\cdot\frac{1}{\left\{1-\frac{1}{2}\frac{\left(\frac{\sigma}{\varepsilon\omega}\right)^2}{\left[\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}+1\right]\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}}\right\}}\tag{10.4.261}$$

That can be rewritten:

$$v_g(\omega)=v_\phi\frac{1}{\left\{1-\frac{1}{2}\frac{\left(\frac{\sigma}{\varepsilon\omega}\right)^2}{\left[\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}+1\right]\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}}\right\}}\tag{10.4.262}$$

Now let us focus on a good conductor that has for property:

$$\sigma\gg\varepsilon\omega\Leftrightarrow\frac{\sigma}{\varepsilon\omega}\gg 1\tag{10.4.263}$$

Then for a good conductor:

$$\beta=\omega\sqrt{\frac{\varepsilon\mu}{2}}\left[\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}+1\right]^{1/2}\cong\omega\sqrt{\frac{\varepsilon\mu}{2}}\left[\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}\right]^{1/2}\cong\omega\sqrt{\frac{\varepsilon\mu}{2}}\left[\sqrt{\frac{\sigma}{\varepsilon\omega}}\right]^{1/2}=\omega\sqrt{\frac{\cancel{\varepsilon}\mu\sigma}{2\cancel{\varepsilon}\omega}}=\sqrt{\frac{\omega\mu\sigma}{2}}\tag{10.4.264}$$

and:

$$\alpha=\omega\sqrt{\frac{\varepsilon\mu}{2}}\left[\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}-1\right]^{1/2}\cong\omega\sqrt{\frac{\varepsilon\mu}{2}}\left[\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}\right]^{1/2}\cong\omega\sqrt{\frac{\varepsilon\mu}{2}}\left[\sqrt{\frac{\sigma}{\varepsilon\omega}}\right]^{1/2}=\omega\sqrt{\frac{\cancel{\varepsilon}\mu\sigma}{2\cancel{\varepsilon}\omega}}=\sqrt{\frac{\omega\mu\sigma}{2}}\tag{10.4.265}$$

Therefore for a good conductor:

$$\beta\cong\alpha\cong\sqrt{\frac{\omega\mu\sigma}{2}}\tag{10.4.266}$$

Therefore still for a good conductor:

$$\delta=\frac{1}{\alpha}\simeq\sqrt{\frac{2}{\omega\mu\sigma}} \tag{10.4.267}$$

and also:

$$\lambda=\frac{2\pi}{\Re(k)}=\frac{2\pi}{\beta}\cong 2\pi\sqrt{\frac{2}{\omega\mu\sigma}} \tag{10.4.268}$$

The phase speed and group speed in a good conductor are respectively for the phase velocity:

$$v_\phi\cong\frac{\omega}{\Re(k)}=\frac{\omega}{\beta}\cong\frac{\omega}{\sqrt{\frac{\omega\mu\sigma}{2}}}=\sqrt{\frac{2\omega}{\mu\sigma}} \tag{10.4.269}$$

and for the group velocity:

$$v_g=\left[\frac{\mathrm{d}\Re(k)}{\mathrm{d}\omega}\right]^{-1}=\left[\frac{\mathrm{d}\beta}{\mathrm{d}\omega}\right]^{-1}\cong\left[\frac{\mathrm{d}}{\mathrm{d}\omega}\sqrt{\frac{\omega\mu\sigma}{2}}\right]^{-1}=\frac{2}{\sqrt{\frac{\omega\mu\sigma}{2}}}=\sqrt{\frac{8\omega}{\mu\sigma}}=2v_\phi(\omega) \tag{10.4.270}$$

So we the famous relation:

$$v_g\cong\sqrt{\frac{8\omega}{\mu\sigma}}=\sqrt{\frac{16\pi f}{\mu\sigma}}=2v_\phi(\omega) \tag{10.4.271}$$

For copper, with $\sigma=5.96\cdot 10^7\,[\mathrm{S}\cdot\mathrm{m}^{-1}]$ and $\mu\approx\mu_0=4\pi\cdot 10^{-7}\,[\mathrm{H}\cdot\mathrm{m}^{-1}]$, and for a plane wave with frequency $50\,[\mathrm{Hz}]$, the group velocity is roughly $5.79\,[\mathrm{m}\cdot\mathrm{s}^{-1}]$! This group velocity and corresponding phase velocity is obviously very slow for electromagnetic radiation!

Remark

Permittivity is typically associated with dielectric materials, however metals are described as having an effective permittivity, with real relative permittivity equal to one. In the high-frequency region, which extends from radio frequencies to the far infrared and terahertz region, the plasma frequency of the electron gas is much greater than the electromagnetic propagation frequency, so the refractive index n of a metal is very nearly a purely imaginary number. In the low frequency regime, the effective relative permittivity is also almost purely imaginary: It has a very large imaginary value related to the conductivity and a comparatively insignificant real-value.

Without doing the approximations we did above thanks to Maple™ and using the same values for f,μ,σ we can that changing the range value of ε_r from $1\cdot 10^{-50}$ to $1\cdot 10^{13}$ doesn't change the result (we still get $5.79\,[\mathrm{m}\cdot\mathrm{s}^{-1}]$). But when $\varepsilon_r>10^{13}$ the changes for start to be significant for v_g.

Notice that for copper the skin depth is therefore equal to:

$$\delta=\frac{1}{\Im(k)}=\frac{1}{\alpha}=\sqrt{\frac{2}{\omega\mu\sigma}}\sqrt{\frac{1}{\pi f\mu\sigma}}\cong 0.92\,[\mathrm{cm}] \tag{10.4.272}$$

We see something interesting here! The expression of the skin depth is exactly the same of that of the skin effect derived at page 4268. Are they they the two sides of the same coin???

Since $\vec{J} = \sigma\vec{E}$, the current density in the conductor and the electric field have the same profile. In particular, the characteristic depth for the current in the conductor is the same as the characteristic fall-off depth for the penetration of the wave; both are proportional to $e^{-z/\delta}$ (if the incident wave is in the positive z.

To clarify: the current and the electromagnetic field interact and sustain each other intimately. You cannot have the one without the other! Analysing the one also informs you about the other.

Instead of thinking of a transmission line as two wires with special behavior, the reader need to think of a transmission line as a guiding structure for electromagnetic waves. When we talk about currents and voltages and capacitance and inductance in a transmission line, that's a useful simplification for many cases. But to understand skin effect the reader should instead think about an electromagnetic wave that is guided by conductive structures. The currents and voltages that we measure on the transmission line are caused by the electromagnetic wave travelling along it, not the other way around.

So then, when an currents wave encounters a conductive material, what happens? The EM wave is reflected. It doesn't penetrate deeply into the conductor. But if the conductivity is not infinite, it does penetrate slightly. And up to the depth that it penetrates, it causes currents to flow in the conductor. That is the skin effect!

10.4.6 Semiconductors

The main defect of the Drude model seen above is to consider the electron as a classical particle. A set of such particles is obviously not subject to quantum distributions and therefore to an explicit relations of temperature.

Moreover, if we observe our Drude model, it is difficult to say anything about resistivity as a function of temperature.

In fact, we generally consider four dates at the source of the development of the semiconductor theory:

- In 11833 (holocene calendar), Michael Faraday reported the conductivity of a material that increases with temperature.
- In 11839 (holocene calendar), Antoine Becquerel discovers that under illumination an electrical tension appears at the junction of certain materials (and liquids). It is the photovoltaic effect, which will give birth much later (around 11950 according to holocene calendar) to the solar cells.
- In 11873 (holocene calendar), Willoughby Smith shows that the conductivity of certain substances increases when illuminated. This is photo-conductivity.
- Finally, in 11874 (holocene calendar), Karl Ferdinand Braun discovers the phenomenon of electrical straightening when a metallic tip is deposited on certain conductors, that is to say that the electric current passes in one direction when the electric potential applied to the point is positive but not when it is negative!

Although these discoveries were totally misunderstood and especially not recognized as the different expressions of the same physical phenomenon (semi-conductivity), practical applications were immediate and led to the second industrial revolution, that of microelectronics!

This type of difficulty (among many others...) largely disappears with the model of the free electron in a potential well, as imagined by Arnold Sommerfeld in 11928 according to holocene calendar (see section Wave Quantum Physics page 4796). In this model the electrons, subjected to the Pauli exclusion principle, follow the Fermi-Dirac energy distribution (see section Statistical Mechanics page 3571), whereas in the Drude model they followed the Maxwell-Boltzmann energy distribution.

There are two important results:

- Only a fraction of the electrons is likely to see its energy vary under the effect of an external action (temperature, electric field, etc.)
- Even at absolute zero, the kinetic energy of the electrons is not zero.

The Sommerfeld model provides a basis for the construction of more specific theories and is the basis of the field of "solid physics" according to some sources. It is therefore not a completed model dealing with a specific problem such as electrical conduction or thermoelectronic emission. This base is the energy distribution of the electrons, obtained by the product of two functions: the density of the states and the Fermi-Dirac distribution.

As the years go by, we will complete the developments that will follow to finally try to have the whole detailed approach. Until then, the reader will have to be patient or to go search the information in other sources on the Internet...

We will make abstraction of the concepts that are not absolutely necessary for the introduction of the model to present here only the essential which is sufficient for the engineer in his daily work.

To begin the mathematical part of the semiconductor study, we will consider a crystal subjected to a potential difference. A conduction electron of the crystal will therefore be subjected, on the one hand, to an internal force $\vec{F}_i$ resulting from the crystalline field and on the other hand to a force of external origin $\vec{F}_e$ resulting from the electric field applied to the crystal.

The model assumptions (hypothesis) of the model are:

H1. There is a high enough potential barrier on the metal surface that prevents electrons from leaving the material.

H2. Inside the material, electrons are subjected to constant potential!

H3. The electrons are independent (no interactions between them).

H4. The electrons obey the laws of Quantum Physics and Classical Mechanics.

H5. The electrons obey to the Maxwell's electrodynamics laws.

H6. The energy bands form a continuous spectrum of energy levels.

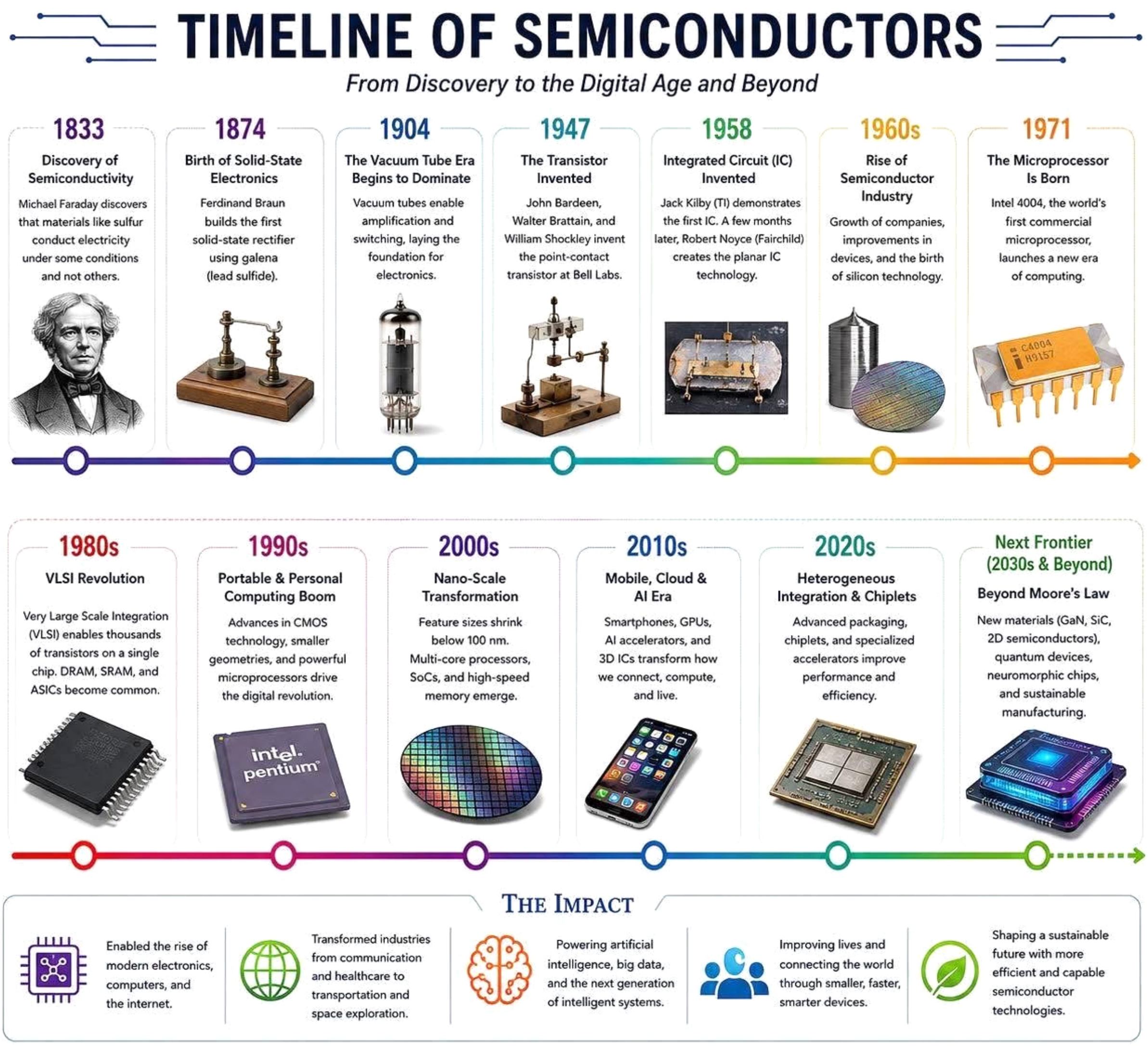
The first hypothesis is based on the following observation: electrons travelling in a metal do not cross, at room temperature at least, the surfaces limiting the sample.

The second hypothesis appears rather brutal. It banish from the model the notion of the "structure of matter". It will be replaced in the model of the energy bands by a periodic potential that accounts for the influence of the positively charged nuclei. This hypothesis reflects the fact that electrons are considered as free in the potential well defined by the sample.

The potential barrier has a finite width but infinite height, that is to say that the passage of the potential inside the material to the potential outside it is done on some inter-atomic distances. However, since the dimensions of the sample are in practice always very great with respect to an inter-atomic distance, the potential barrier can be considered as infinitely abrupt, which simplifies the calculations.

Remark

In order to simplify the calculations, we will assume that the electrons move in only one direction (that of the electric field), thus avoid to work with vectors.



TIMELINE OF SEMICONDUCTORS

From Discovery to the Digital Age and Beyond

1833 Discovery of Semiconductivity
Michael Faraday discovers that materials like sulfur conduct electricity under some conditions and not others.

1874 Birth of Solid-State Electronics
Ferdinand Braun builds the first solid-state rectifier using galena (lead sulfide).

1904 The Vacuum Tube Era Begins to Dominate
Vacuum tubes enable amplification and switching, laying the foundation for electronics.

1947 The Transistor Invented
John Bardeen, Walter Brattain, and William Shockley invent the point-contact transistor at Bell Labs.

1958 Integrated Circuit (IC) Invented
Jack Kilby (TI) demonstrates the first IC. A few months later, Robert Noyce (Fairchild) creates the planar IC technology.

1960s Rise of Semiconductor Industry
Growth of companies, improvements in devices, and the birth of silicon technology.

1971 The Microprocessor Is Born
Intel 4004, the world's first commercial microprocessor, launches a new era of computing.

1980s VLSI Revolution
Very Large Scale Integration (VLSI) enables thousands of transistors on a single chip. DRAM, SRAM, and ASICs become common.

1990s Portable & Personal Computing Boom
Advances in CMOS technology, smaller geometries, and powerful microprocessors drive the digital revolution.

2000s Nano-Scale Transformation
Feature sizes shrink below 100 nm. Multi-core processors, SoCs, and high-speed memory emerge.

2010s Mobile, Cloud & AI Era
Smartphones, GPUs, AI accelerators, and 3D ICs transform how we connect, compute, and live.

2020s Heterogeneous Integration & Chiplets
Advanced packaging, chiplets, and specialized accelerators improve performance and efficiency.

Next Frontier (2030s & Beyond) Beyond Moore's Law
New materials (GaN, SiC, 2D semiconductors), quantum devices, neuromorphic chips, and sustainable manufacturing.

THE IMPACT

Enabled the rise of modern electronics, computers, and the internet.

Transformed industries from communication and healthcare to transportation and space exploration.

Powering artificial intelligence, big data, and the next generation of intelligent systems.

Improving lives and connecting the world through smaller, faster, smarter devices.

Shaping a sustainable future with more efficient and capable semiconductor technologies.

The equation of the dynamics is then written naturally for this electron:

$$m_0\frac{\mathrm{d}\vec{v}}{\mathrm{d}t}=\vec{F}_i+\vec{F}_e \qquad (10.4.273)$$

We then write (nothing prevents us to do so) that the electron in the crystal responds to the stress of the external force $\vec{F}_e$ as a quasi-particle of mass m_e in a vacuum:

$$m_0\frac{\mathrm{d}\vec{v}}{\mathrm{d}t}=\vec{F}_i+\vec{F}_e=\vec{F}_i+m_e\frac{\mathrm{d}\vec{v}}{\mathrm{d}t} \qquad (10.4.274)$$

It is the study of the latter term which will interest us. For this purpose let us recall that in the detailed study of the propagation of the free electron in a vacuum, where we neglect the effects of its spin, we have shown that it must be described according to the Schrödinger equation by a wave packet (see section Wave Quantum Physics page 4786) centered on a state k_0, otherwise its energy would be infinite.

However, we can ask ourselves ... what leads us to consider it as free...? Well it is experience that shows that when we apply a certain threshold potential, a current begins to appear in the semiconductors.

We have proved (always in the context of the propagation of the free spin-free particle in the section of Wave Quantum Physics) that the wave packet can then be seen in its mathematical solution as a (free) plane wave moving at phase speed:

$$v_\varphi = \frac{\omega_0}{k_0} \tag{10.4.275}$$

which we will write for as following in order to simplify the notations of the next developments:

$$v = \frac{\omega}{k} \tag{10.4.276}$$

However, in the crystal lattice, the phase velocity can vary, depending on the location of the electron in the lattice due to the geometrical shape of the potential in the crystal. We must therefore use the instantaneous phase velocity:

$$\vec{v} = \frac{\mathrm{d}\omega}{\mathrm{d}\vec{k}} = \begin{pmatrix} \dfrac{\mathrm{d}\omega}{\mathrm{d}\vec{k}_x} \\ \dfrac{\mathrm{d}\omega}{\mathrm{d}\vec{k}_y} \\ \dfrac{\mathrm{d}\omega}{\mathrm{d}\vec{k}_z} \end{pmatrix} \tag{10.4.277}$$

Let us recall (see section Corpuscular Quantum Physics page 4627) that we also have in general the total energy given by:

$$E = h\nu = h\frac{\omega}{2\pi} = \hbar\omega \tag{10.4.278}$$

Therefore we get:

$$v = \frac{\mathrm{d}\omega}{\mathrm{d}\vec{k}} = \frac{1}{\hbar}\frac{\mathrm{d}E}{\mathrm{d}\vec{k}} \tag{10.4.279}$$

The term:

$$\frac{\mathrm{d}E}{\mathrm{d}\vec{k}} \tag{10.4.280}$$

is by no means simple in the case of a crystal (it is even a nightmare ...).

Obviously for a free particle (see section Wave Quantum Physics page 4788), let us recall that it is equal to:

$$\frac{\mathrm{d}E}{\mathrm{d}k} = \frac{\mathrm{d}}{\mathrm{d}k}\left(\frac{\hbar^2 k^2}{2m}\right) = \frac{\hbar^2 k}{m} \tag{10.4.281}$$

But for a particle in a potential field having a complex geometry, the energy E begins to have an expression dependent of k as a function of the zones which can become very complex (see the examples of the section Wave Quantum Physics). Hence the justification for the use of the derivative.

The acceleration in the classical sense of this electron is then given by:

$$\vec{a} = \frac{\mathrm{d}\vec{v}}{\mathrm{d}t} = \frac{1}{\hbar}\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\mathrm{d}E}{\mathrm{d}\vec{k}}\right) = \frac{1}{\hbar}\frac{\mathrm{d}}{\mathrm{d}\vec{k}}\frac{\mathrm{d}E}{\mathrm{d}t} \tag{10.4.282}$$

We also have (see section Classical Mechanics page 3293):

$$\mathrm{d}E=\vec{F}\circ\mathrm{d}\vec{x}=\vec{F}\circ\vec{v}\mathrm{d}t \tag{10.4.283}$$

Therefore:

$$\frac{\mathrm{d}E}{\mathrm{d}t}=\vec{F}\circ\vec{v} \tag{10.4.284}$$

Hence:

$$\vec{a}=\frac{\mathrm{d}\vec{v}}{\mathrm{d}t}=\frac{1}{\hbar}\frac{\mathrm{d}}{\mathrm{d}\vec{k}}\frac{\mathrm{d}E}{\mathrm{d}t}=\frac{1}{\hbar}\frac{\mathrm{d}}{\mathrm{d}\vec{k}}\left(\vec{F}\circ\vec{v}\right)=\frac{1}{\hbar}\frac{\mathrm{d}}{\mathrm{d}\vec{k}}\left(\vec{F}\circ\frac{1}{\hbar}\frac{\mathrm{d}E}{\mathrm{d}\vec{k}}\right) \tag{10.4.285}$$

The derivative of $\vec{F}$ with respect to $\vec{k}$ in the preceding relation will be cancelled because the force is derived from the potential applied to the semiconductor only and not from the wave vector of the electron itself! We then have:

$$\vec{a}=\vec{F}\frac{1}{\hbar^2}\frac{\mathrm{d}^2E}{\mathrm{d}\vec{k}^2} \tag{10.4.286}$$

Since here E is only the total energy coming from the externally submitted potential, then the force $\vec{F}$ is the external force $\vec{F}_e$ generated by the application of this same potential. We then have:

$$\vec{F}_e=m_e\frac{\mathrm{d}\vec{v}}{\mathrm{d}t}=m_e\cdot\vec{a} \tag{10.4.287}$$

and:

$$\vec{F}_e=\frac{1}{\dfrac{1}{\hbar^2}\dfrac{\mathrm{d}^2E}{\mathrm{d}k^2}}\vec{a} \tag{10.4.288}$$

It then comes by equalization:

$$m_e=\frac{\hbar^2}{\dfrac{\mathrm{d}^2E}{\mathrm{d}k^2}} \tag{10.4.289}$$

Since the energy of the electron can have a complicated mathematical form according to the practical applications cases presented in the section of Wave Quantum Physics, let us express $E(\vec{k})$ in the form of a limited Taylor development (see section Sequences and Series page 1703) of a function of three variables at the second order by dropping the terms of interactions and by not taking the terms of first degree:

$$\begin{aligned}E(\vec{k})&=E(k_x,k_y,k_z)\\&\cong E(k_{x0},k_{y0},k_{z0})+\frac{1}{2!}\frac{\partial^2E}{\partial k_x^2}(k_x-k_{x0})^2+\frac{1}{2!}\frac{\partial^2E}{\partial k_y^2}(k_y-k_{y0})^2+\frac{1}{2!}\frac{\partial^2E}{\partial k_z^2}(k_z-k_{z0})^2\end{aligned} \tag{10.4.290}$$

In fact, this rough but still acceptable approximation in many practical cases is due to the fact that experience shows that the energy surfaces as a function of k approximate a parabolic shape in certain semiconductor crystals.

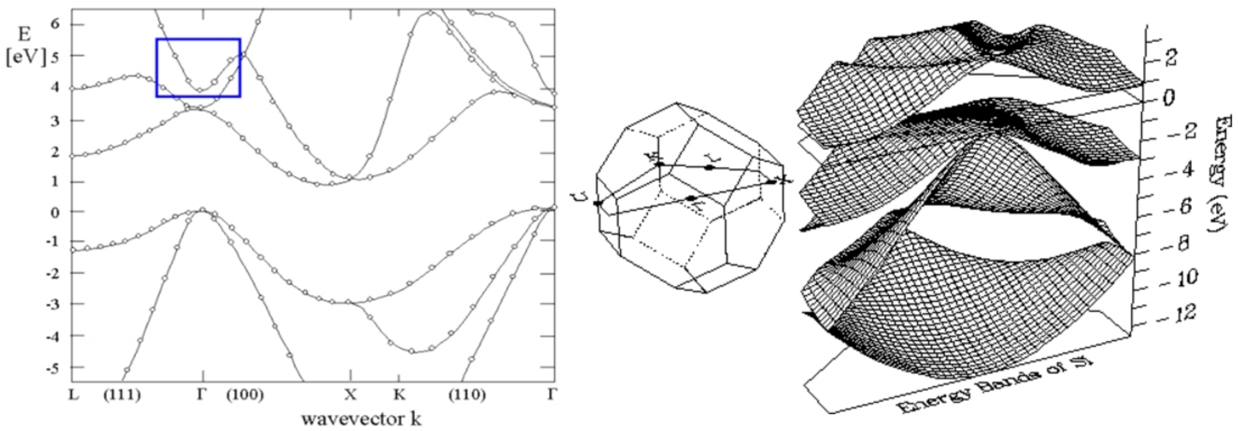


Figure 10.101: Band structure of Silicium following different plane in the crystal. The blue rectangle shows a region of one of the conduction bands which may be described by a parabola approximately

In the conductors, the approximation of the preceding relation is taken only at the first term.

Another way of seeing it is that for a free electron, for recall, in one dimension, the dispersion curve (see section Wave Quantum Physics page 4785):

$$E(k)=\frac{\hbar^2}{2m}k^2 \tag{10.4.291}$$

which is indeed a parabola as a function of k. Indeed, if we take our Taylor development in one dimension it remains:

$$E(k)\cong E(k_0)+\frac{1}{2!}\frac{\partial^2 E}{\partial k^2}(k-k_0)^2 \tag{10.4.292}$$

and as we determined before that:

$$m_e=\frac{\hbar^2}{\dfrac{\mathrm{d}^2E}{dk^2}} \tag{10.4.293}$$

It comes:

$$E(k)\cong E(k_0)+\frac{1}{2!}\frac{\partial^2 E}{\partial k^2}(k-k_0)^2=\frac{1}{2!}\frac{\hbar^2}{m_e}(k-k_0)^2 \tag{10.4.294}$$

If the electron is free, the dispersion curve imposes us to have (without the presence of a potential):

$$E(k_0)=0 \tag{10.4.295}$$

which is then considered as the "energy of the minimum" E_{min}. It then remains:

$$E(k)=\frac{\hbar^2}{2m_e}(k_x-k_{x0})^2 \tag{10.4.296}$$

and by taking $k_{x0}=0$ we fall back on the dispersion curve of a free particle (which therefore justifies the fact of having to choose $E(k_0)=0$ for a free electron):

$$E(k)=\frac{\hbar^2}{2m_e}k_x^2 \tag{10.4.297}$$

This shows that the approximation is not too false ... and justifies the fact that in some textbooks the previous relation (Taylor series) describes a so-named "quasi-free particle".

But let us come back to:

$$E(k)=E(k_0)+\frac{\hbar^2}{2m_e}(k-k_0)^2 \qquad (10.4.298)$$

And since the wave packet is centered around k_0, let us normalize it as being equal to 0 (which is equivalent to centering the wave vector values). We then have:

$$E(k)=E(k_0)+\frac{\hbar^2}{2m_e}k^2=E_0+\frac{\hbar^2}{2m_e}k^2 \qquad (10.4.299)$$

What is interesting with these developments is that we started from a free electron in the form of a wave packet and thanks to the Taylor development we find ourselves with an extremely simple expression of the energy of a quasi-free electron.

It emerges that for a quasi-free electron, without interactions and without taking into account the effects of spin we have:

$$k\cong\sqrt{(E-E_0)\frac{2m_e}{\hbar}} \qquad (10.4.300)$$

We then notice a very sympathetic thing! This is that our quasi-free electron has a wave number which resembles in all points tot particle stuck in a potential well with rectilinear walls (see the proof in the section of Wave Quantum Physics).

We now wish to calculate the density of states (in extenso of electrons) in the volume given by the corresponding rectangular well, by means of the expression of k (not having directly that of E as being too complex).

We have proved in the section of Wave Quantum Physics that for the potential well with rectangular barriers that:

$$k=n\frac{\pi}{L} \qquad (10.4.301)$$

if we imposed an integer half wavelength. If we impose an integer wavelength ("Born-von Karman conditions" so that after a translation of the periodic lattice of the crystal we fall back on the same properties) so that the solution is physically acceptable, then we have:

$$k=n\frac{2\pi}{L} \qquad (10.4.302)$$

Which obviously implies two times fewer states.

By extension, for space, we then have in the three-dimensional case:

$$k_x=n_x\frac{2\pi}{L_x},k_y=n_y\frac{2\pi}{L_y},k_z=n_z\frac{2\pi}{L_z} \qquad (10.4.303)$$

with:

$$E_{n_x,n_y,n_z}=\frac{\pi^2\hbar^2}{2mL^2}\left(n_x^2+n_y^2+n_z^2\right)=E_0\left({n_x}^2+{n_y}^2+{n_z}^2\right) \qquad (10.4.304)$$

and where $n_{x,y,z}=1,2,3,\ldots$.

The result is very similar to that of the one-dimensional infinite rectangular potential well but now we have special boundary conditions in the purpose to have a correspondence with the experiment and three main quantum numbers instead of just one. Moreover, each combination of these three numbers corresponds to a different wave function (state). Moreover, these numbers are independent (no condition imposed).

✍**Example:**

Consider a solid metal cube of edge length 2 [cm]. We want to know the value of the lowest energy level for an electron within the metal and what is the spacing between this level and the next energy level!

The lowest energy level corresponds to the quantum numbers $n_x=n_y=n_z=1$. From the above relation, the energy of this level is:

$$E_{1,1,1}=\frac{\pi h^2}{2m_eL^2}(1^2+1^2+1^2)\cong\frac{3\pi^2(1.05\cdot 10^{-34})^2}{2(9.11\cdot 10^{-31})(2\cdot 10^{-2})^2}\cong 4.48\cdot 10^{-34}\,[\text{J}] \qquad (10.4.305)$$

The next-higher energy level is reached by increasing any one of the three quantum numbers by 1. Hence, there are actually three quantum states with the same energy. Suppose we increase n_x by 1. Then the energy becomes:

$$E_{2,1,1}=\frac{\pi^2h^2}{2m_eL^2}(2^2+1^2+1^2)\cong\frac{6\pi^2(1.05\cdot 10^{-34})^2}{2(9.11\cdot 10^{-31})(2\cdot 10^{-2})^2}\cong 8.96\cdot 10^{-34}\,[\text{J}] \qquad (10.4.306)$$

The energy spacing between the lowest energy state and the next-highest energy state is therefore:

$$E_{2,1,1}-E_{1,1,1}=2.8\cdot 10^{-15}\,[\text{eV}] \qquad (10.4.307)$$

his is a very small energy difference. Compare this value to the average kinetic energy of a particle, as the product kT is in comparison $1,000$ times greater than the previously calculated energy spacing.

We then have for the first level where all n are unitary:

$$k_xk_yk_z=\frac{(2\pi)^3}{L_xL_yL_z} \qquad (10.4.308)$$

If we accept to simplify that the well has edges of equal length (cubic crystal lattice semiconductor), we then have:

$$k_xk_yk_z=\frac{(2\pi)^3}{L^3}=\left(\frac{2\pi}{L}\right)^3 \qquad (10.4.309)$$

Let us represent the space k for such a cubic lattice and for different multiples of n_x, n_y, n_z:

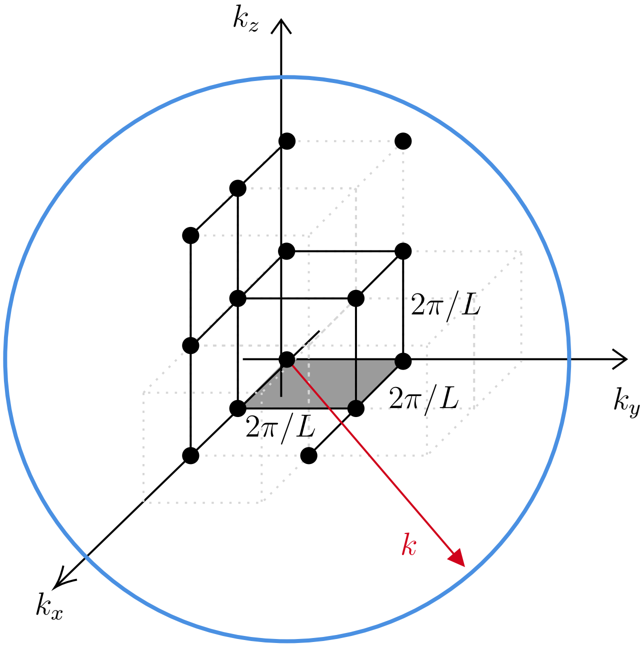
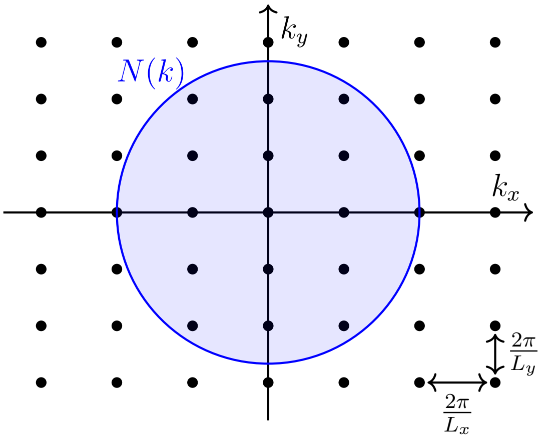


Figure 10.102: k-space for a cubic crystal lattice

Or in two dimensions:



Therefore, all quantized states can take only values space from $2\pi/L$ in the space of k, which means that by elementary volume there is only one possible wave vector and therefore only one associated state. Indeed, the reader can draw a picture over the figure above if he wants and he will see (!) but do not rely on the big blackheads that are there only to show the ends of the elementary volumes and which do not all correspond to possible states!

Remark

The sphere of radius k containing the levels with one electron is sometimes referred to as the "Fermi sphere". The radius value is then denoted k_F and named the "Fermi wave vector". The surface of the Fermi sphere, which separates the occupied levels from those which are not occupied as we shall see later, is named the "Fermi surface".

Thus, in a spherical volume with radius k of the k-space. We have a precise number (upper limit) of elementary volumes (states), remembering that $V = L^3$, given by:

$$\boxed{N(k) = \frac{\frac{4}{3}\pi k^3}{k_x k_y k_z} = \frac{\frac{4}{3}\pi k^3}{(2\pi/L)^3} = \frac{k^3}{6\pi^2}V} \tag{10.4.310}$$

where in literature it is customary (tradition) to retain only the form of the second equality in the developments. This relation has been a useful to us for recall in the section of Thermodynamics to determine the Debye-Einstein model of the constant-volume (isochoric) thermal capacity of crystalline solids (see page 3622)!

The density of modes in a volume V will then be given by (relation used in the section of Thermodynamics to express the calorific capacity at constant volume of solids):

$$D(\omega) = \frac{\mathrm{d}N}{\mathrm{d}k} = \frac{\mathrm{d}k}{\mathrm{d}\omega} = \frac{Vk^2}{2\pi^2}\frac{\mathrm{d}k}{\mathrm{d}\omega} \tag{10.4.311}$$

In facts, due to the presence of potential bonds between the atoms of the crystal, the sphere is not perfect and has like giant pockmarks/photholes on what is otherwise a smooth surface:

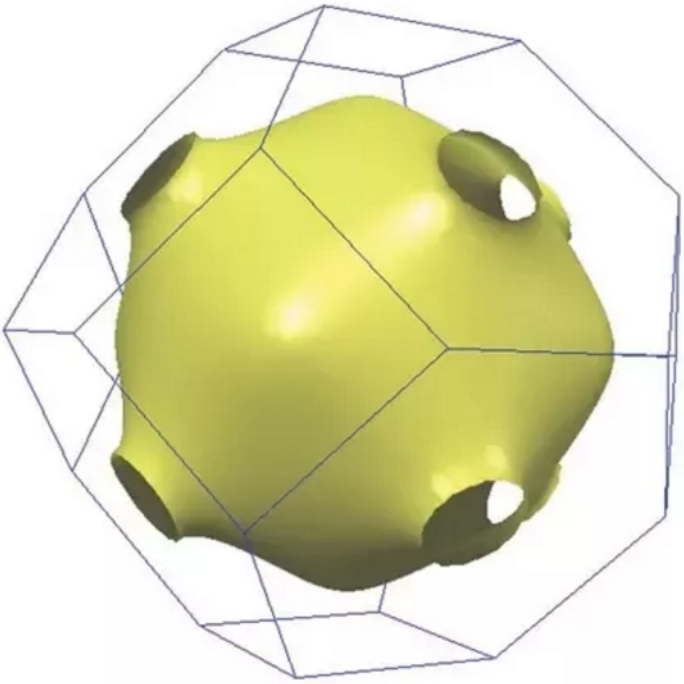


Figure 10.103: **Fermi surface of copper**

or for real, here is a (supposed) Fermi surface and electron momentum density of Copper in the reduced zone schema measured with 2D Angular Correlation of Electron Positron Annihilation Radiation method ("supposed" as I was not able to find the original article with the picture):

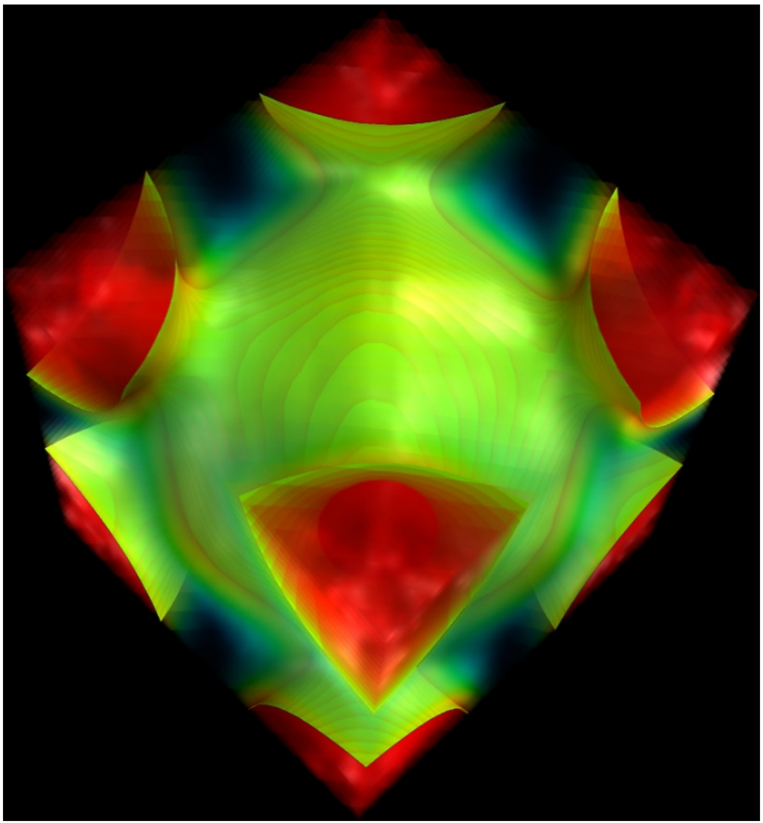


Figure 10.104: **Fermi surface of copper with 2D ACAR method (source: Wikipedia)**

Now considering the spin (yes indeed why not do it...?!) we multiply by 2 since there are two spin states possible by state for the electron (remembering that $V=L^3$):

$$N(k)=2\cdot\frac{4}{3}\frac{\pi k^3}{(2\pi)^3}L^3=\frac{k^3}{3\pi^2}V \tag{10.4.312}$$

(relation that we will see again in the section of Nuclear Physics chapter during our study of the liquid drop nucleus model at page 5141) and by injecting in it:

$$k\cong\sqrt{(E-E_0)\frac{2m_e}{\hbar^2}} \tag{10.4.313}$$

we then have:

$$\begin{aligned}N(E)&\cong 2\cdot\frac{\frac{4}{3}\pi}{(2\pi)^3}\left((E-E_0)\frac{2m_e}{\hbar^2}\right)^{3/2}L^3=\frac{\frac{4}{3}}{(2\pi)^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}(E-E_0)^{3/2}L^3\\&=\frac{\frac{4}{3}}{(2\pi)^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}(E-E_0)^{3/2}\cdot V\end{aligned} \tag{10.4.314}$$

The volume density of (quasi-free) states will be obtained by deriving this last relation by the volume:

$$\rho(E)=\frac{\mathrm{d}N(E)}{\mathrm{d}V}=\frac{\frac{4}{3}}{(2\pi)^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}(E-E_0)^{3/2} \tag{10.4.315}$$

And if we want the density of (quasi-)free states (of vibrations) per unit of energy and volume, we will have to derive also with respect to the energy:

$$\frac{\mathrm{d}\rho(E)}{\mathrm{d}E}\cong\frac{\frac{4}{3}}{(2\pi)^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}\frac{\mathrm{d}}{\mathrm{d}E}(E-E_0)^{3/2}=\frac{\frac{4}{3}}{(2\pi)^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}\frac{3}{2}(E-E_0)^{1/2} \tag{10.4.316}$$

which gives:

$$\boxed{D(E)\cong\frac{1}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}(E-E_0)^{1/2}} \tag{10.4.317}$$

This result does not depend on the volume, it is unchanged when the latter tends towards infinity! So it is valid for any point of the semiconductor crystal if this one is perfect and without bonds between atoms...

What we also find sometimes in the following (somewhat unfortunate ...) forms in some textbooks:

$$\begin{aligned}D(E)&\cong\frac{1}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}\sqrt{E}\\D(E)&\cong\frac{4}{h^3\sqrt{\pi}}(2\pi m_e)^{3/2}(E-E_0)^{1/2}\end{aligned} \tag{10.4.318}$$

and there are also those who take into account the spin into account only much more later ... which gives a form identical to that of the previous three relations but divided by a factor 2.and there are also those who take into account the spin into account only much more later ... which gives a form identical to that of the previous three relations but divided by a factor 2.

The difference $E-E_0$ as we will see further below is denoted E_F, and for obvious reasons related to Statistical Mechanics that we will see further below named the "Fermi Energy". Therefore notice that we have using the previous results:

$$N(E)\cong\frac{4/3}{(2\pi)^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}(E-E_0)^{3/2}V=\frac{1}{3\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}E_F^{3/2}V \tag{10.4.319}$$

After rearranging we get immediately:

$$\begin{aligned}E_F&=\frac{\hbar^2}{2m_e}\left(\frac{N}{V}\right)^{2/3}(3\pi^2)^{2/3}=\frac{h^2}{8\pi^2m_e}\left(\frac{3N}{V}\right)^{2/3}\pi^{4/3}=h^28m_e\left(\frac{3N}{V}\right)^{2/3}\pi^{-2/3}\\&=h^28m_e\left(\frac{3N}{\pi V}\right)^{2/3}\end{aligned} \tag{10.4.320}$$

Therefore:

$$\boxed{E_F=\frac{h^2}{8m_e}\left(\frac{3N}{\pi V}\right)^{2/3}} \tag{10.4.321}$$

Fermi energies for selected materials are listed in the following table:

Element	Conduction band electron density	Free-Electron Model Fermi Energy
Al	$18.1\cdot10^{28}$ [m^{-3}]	11.7 [eV]
Ba	$3.15\cdot10^{28}$ [m^{-3}]	3.64 [eV]
Cu	$8.47\cdot10^{28}$ [m^{-3}]	7.00 [eV]
Au	$5.90\cdot10^{28}$ [m^{-3}]	5.53 [eV]
Fe	$17.0\cdot10^{28}$ [m^{-3}]	11.1 [eV]
Ag	$5.86\cdot10^{28}$ [m^{-3}]	5.49 [eV]

Table 10.7: **Fermi energy for some materials**

We can obviously associate the Fermi temperature (see section Statistical Mechanics page 3576) with the Fermi Energy as $E=kT$. Therefore:

$$T_F=\frac{E_F}{k}=\frac{h^2}{8m_ek}\left(\frac{3N}{\pi V}\right)^{2/3} \quad (10.4.322)$$

✍**Example:**

The Fermi temperature associated to the Fermi energy of silver is:

$$T_F\cong\frac{8.8\cdot 10^{-19}}{1.38\cdot 10^{-23}}\cong 6.4\cdot 10^4\,[\text{K}] \quad (10.4.323)$$

which is much higher than room temperature and also the typical melting point ($\sim 10^3$ [K]) of a metal.

10.4.6.1 Non-degenerated statistic density of negative electric charge carriers

In short, however this relation has an issue (one more...)! Indeed, we have seen in the section of Statistical Mechanics in our study of Quantum Statistics that in a system where even the energy spectrum is considered continuous it is impossible not to take into account the degeneration of the different levels of energy. We have then demonstrated that for a population of fermions, at a given energy (or temperature) the percentage of degenerate levels occupied is given by the Fermi-Dirac function:

$$f_{\text{FD}}(E,T,\mu)=\frac{1}{\exp\left(\dfrac{E-\mu}{kT}\right)+1} \quad (10.4.324)$$

and that function therefore returns a value between 0 and 1.

This function therefore gives for a fixed temperature T the probability that an electron occupies a state of energy E.

Hence our relation $D(E)$ overestimates the real density value of (quasi-)free occupied states for a given energy (or temperature). In order to obtain a better approximation, we logically write the volumetric density of (quasi-free) states per unit of energy by:

$$D(E)\cong\frac{1}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}(E-E_0)^{1/2}f_{\text{FD}}(E,T,\mu)=\frac{1}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}(E-E_0)^{1/2}\cdot\frac{1}{\exp\left(\dfrac{E-\mu}{kT}\right)+1} \quad (10.4.325)$$

However, in practice, we will try to calculate the volume density of (quasi-free) states in a spectrum (interval) of energy. It then comes with the correction added previously:

$$\frac{\mathrm{d}\rho(E)}{\mathrm{d}E}\cong\frac{1}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}\frac{(E-E_0)^{1/2}}{\exp\left(\dfrac{E-\mu}{kT}\right)+1} \quad (10.4.326)$$

Thus:

$$\mathrm{d}\rho(E)\cong\frac{1}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}\frac{(E-E_0)^{1/2}}{\exp\left(\dfrac{E-\mu}{kT}\right)+1}\mathrm{d}E \tag{10.4.327}$$

It then immediately follows that the volumetric density of (quasi-free) states at a given temperature (normal temperature conditions for civilian applications) taking into account all possible states (continuous levels) of energy is then given by:

$$\rho(E)\cong\frac{1}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}\int\limits_{E_0}^{+\infty}\frac{(E-E_0)^{1/2}}{\exp\left(\dfrac{E-\mu}{kT}\right)+1}\mathrm{d}E \tag{10.4.328}$$

Take E_0 as a lower boundary avoids us, as we shall see explicitly a further below, to find ourselves with a negative root ... which would be very a priori troublesome!

Moreover, we can, without any appreciable error, postpone the limit of the integral to the infinity because $f_{\text{FD}}\rightarrow 0$ when E is large.

Unfortunately, this (improper) integral is generally not analytically soluble. We will have to use approximations.

We will start by making the assumption that we are in the classical regime of electron gas. That is, we have:

$$\exp\left(\frac{E-\mu}{kT}\right)\gg 1 \tag{10.4.329}$$

which implies:

$$E-\mu\gg kT \tag{10.4.330}$$

Therefore, we also have the approximation:

$$\frac{1}{\exp\left(\dfrac{E-\mu}{kT}\right)+1}\cong\frac{1}{\exp\left(\dfrac{E-\mu}{kT}\right)+1}=\exp\left(-\frac{E-\mu}{kT}\right) \tag{10.4.331}$$

In other words, the energy E must be much higher than the chemical potential μ (assimilated often unfortunately to my knowledge wrongly in the literature on semiconductors to the Fermi level E_F). Physicists then note this energy E_C to distinguish it and name it "minimal energy of the conduction band" (which corresponds to the minimum energy of a quasi-free electron to satisfy this condition).

Therefore, we also change the notation for the charge density:

$$\rho(E)\longrightarrow\rho_C(E) \tag{10.4.332}$$

Remark

Unfortunately, as mentioned in the previous paragraph (!) in many quality textbooks on semiconductors, the chemical potential μ, which is a purely thermodynamic notion implying a hypothesis of interactions, is replaced by the concept of Fermi energy E_F and however this is not the same thing! The two energies coincide only in the case where the temperature T is equal to zero!

So we must consider the term "Fermi level" as being nothing but a synonym for "chemical potential" in the context of semiconductors.

We then have:

$$\begin{aligned}\rho_C(E) &\cong \frac{1}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}\int\limits_{E_C}^{+\infty}(E-E_C)^{1/2}\exp\left(-\frac{E-\mu}{kT}\right)\mathrm{d}E\\ &=\frac{1}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}\int\limits_{E_C}^{+\infty}(E-E_C)^{1/2}f_{\text{MB}}\mathrm{d}E\\ &=\frac{1}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}\exp\left(\frac{\mu}{kT}\right)\int\limits_{E_C}^{+\infty}(E-E_C)^{1/2}\left(-\frac{E}{kT}\right)\mathrm{d}E\end{aligned} \tag{10.4.333}$$

where f_{MB} is the Maxwell-Boltzmann distribution (see section Statistical Mechanics page 3554) given for recall by:

$$f_{\text{MB}}(E,T,\mu)=\exp\left(-\frac{E-\mu}{kT}\right) \tag{10.4.334}$$

and thus corresponds well to a non-quantum behaviour (i.e. a non-degenerate electron gas!) because when:

$$E-\mu\gg kT \tag{10.4.335}$$

we have:

$$f_{\text{MB}}(E,T,\mu)\ll 1 \tag{10.4.336}$$

and therefore the energy states of are far from being all occupied by electrons (there is therefore no degeneration).

We are therefore well in a situation where Classical Physics predominates over Quantum Physics. This is why in this approximation (of Maxwell-Boltzmann type) we say that we are dealing with a "non-degenerate semiconductor" because the electrons are not all stacked into the lowest available levels.

To continue, we make a change of variable by putting:

$$x=\frac{E-E_C}{kT} \tag{10.4.337}$$

hence:

$$\mathrm{d}x=\frac{1}{kT}\mathrm{d}E \quad \text{and} \quad E=x\cdot kT+E_C \tag{10.4.338}$$

Therefore it comes:

$$\begin{aligned}\int\limits_{E_C}^{+\infty}\sqrt{E-E_C}\left(-\frac{E}{kT}\right)\mathrm{d}E&=\sqrt{kT}\int\limits_0^{+\infty}x^{1/2}\exp\left(-\frac{x\cdot kT+E_C}{kT}\right)\mathrm{d}xkT\\&=(kT)^{3/2}\exp\left(-\frac{E_C}{kT}\right)\int\limits_0^{+\infty}x^{1/2}e^{-x}\mathrm{d}x\end{aligned} \tag{10.4.339}$$

We do an integration by parts:

$$\int\limits_0^{+\infty}x^{1/2}e^{-x}\mathrm{d}x=-\left[x^{1/2}e^{-x}\right]_0^{+\infty}+\frac{1}{2}\int\limits_0^{+\infty}\frac{e^{-x}}{\sqrt{x}}\mathrm{d}x=\frac{1}{2}\int\limits_0^{+\infty}\frac{e^{-x}}{\sqrt{x}}\mathrm{d}x \tag{10.4.340}$$

We then make a change of variable by putting:

$$u=\sqrt{x}\Rightarrow\frac{\mathrm{d}u}{\mathrm{d}x}=\frac{1}{2\sqrt{x}}\Rightarrow\mathrm{d}x=2\sqrt{x}\mathrm{d}u=2u\mathrm{d}u \tag{10.4.341}$$

which gives:

$$\frac{1}{2}\int\limits_0^{+\infty}\frac{e^{-x}}{\sqrt{x}}\mathrm{d}x=\frac{1}{2}\int\limits_0^{+\infty}\frac{e^{-u^2}}{u}2u\mathrm{d}u=\int\limits_0^{+\infty}e^{-u^2}\mathrm{d}u \tag{10.4.342}$$

We have already calculated this integral in the Statistics section. It comes:

$$\int\limits_0^{+\infty}x^{1/2}e^{-x}\mathrm{d}x=\int\limits_0^{+\infty}e^{-u^2}\mathrm{d}u=\frac{\sqrt{\pi}}{2} \tag{10.4.343}$$

We then have finally:

$$\begin{aligned}\rho_C(E)&\cong\frac{1}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}\exp\left(\frac{\mu}{kT}\right)(kT)^{3/2}\exp\left(-\frac{E_C}{kT}\right)\frac{\sqrt{\pi}}{2}\\&=\frac{1}{2\pi^2}\left(\frac{2m_ekT}{\hbar^2}\right)^{3/2}\exp\left(\frac{\mu-E_C}{kT}\right)\frac{\sqrt{\pi}}{2}=\frac{1}{2^2\pi^{3/2}}\left(\frac{2m_ekT}{\hbar^2}\right)^{3/2}\exp\left(\frac{\mu-E_C}{kT}\right)\\&=2\left(\frac{m_ekT}{2\pi\hbar^2}\right)^{3/2}\exp\left(\frac{\mu-E_C}{kT}\right)\end{aligned} \tag{10.4.344}$$

Where, for recall, m_e is the mass of the quasi-particle (and not the mass of the electron for recall!). So after integration everything happens as if all the electrons were concentrated on the level of energy E_C with a number of available places corresponding to:

$$2\left(\frac{m_ekT}{2\pi\hbar^2}\right)^{3/2} \tag{10.4.345}$$

The relation $\rho_C(E)$ is traditionally written (and in a somewhat unfortunate way ... because it is not easy to remember that it is a density):

$$N_n = N_{n,T}\exp\left(\frac{\mu - E_C}{kT}\right) \tag{10.4.346}$$

or also:

$$\boxed{N_n = N_C\exp\left(\frac{\mu - E_C}{kT}\right)} \tag{10.4.347}$$

where we have approximately at ambient temperature the following values of (quasi-)free states for Silicium:

$$N_C \cong 2\left(\frac{m_e kT}{2\pi\hbar^2}\right)^{3/2} \cong 2.7\cdot 10^{19}\ \left[\mathrm{cm}^{-3}\right] \tag{10.4.348}$$

and for the Germanium:

$$N_C \cong 2\left(\frac{m_e kT}{2\pi\hbar^2}\right)^{3/2} \cong 1\cdot 10^{19}\ \left[\mathrm{cm}^{-3}\right] \tag{10.4.349}$$

while there is a density of about $4.5\cdot 10^{22}$ atoms by cubic centimetres and about 10^{24} electron by cubic centimetre for these two elements.

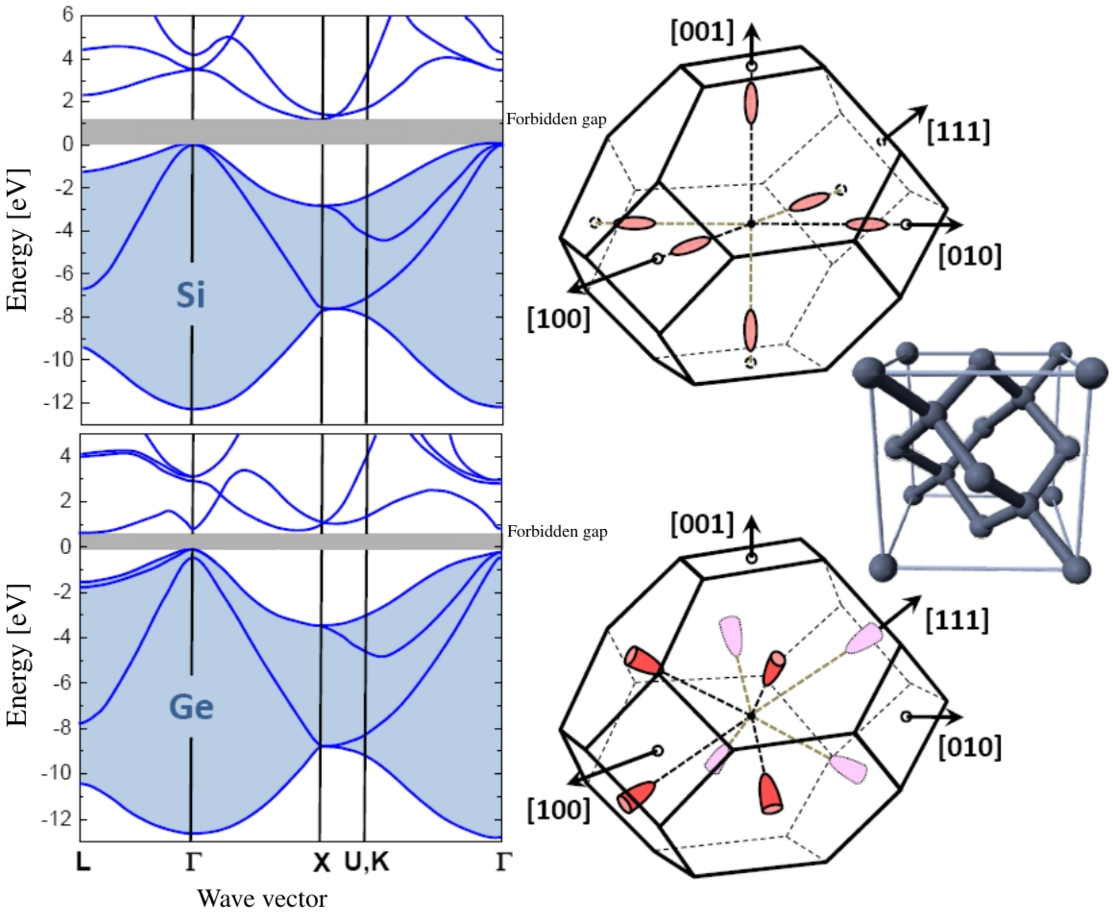


Figure 10.105: Silicium vs Germanium band and crystal structure (source: ?)

This means that there is thus a ratio of: $10^{19}/10^{24}=10^{-5}$ between the total electron density and the number of quasi-free electrons.

We also notice that this theoretical model does not take into account the electronic structure (atomic number) of the studied material.

Thus, we see that the variations of quasi-free electron densities as a function of temperature (in the temperature validity range ...) are essentially of increasing or decreasing exponential type.

From the density of the free electrons (caution! it must be remembered that these are only the quasi-free electrons that are wandering in our mathematical equations so far...) in the semi-conductor crystal, we can deduce the Fermi energy level (more rigorously it is the chemical potential!):

$$N_n=N_C\exp\left(\frac{\mu-E_C}{kT}\right)\Rightarrow\frac{N_n}{N_C}=\exp\left(\frac{\mu-E_C}{kT}\right)\Rightarrow\ln\left(\frac{N_n}{N_C}\right)=\frac{\mu-E_C}{kT}\qquad(10.4.350)$$

hence:

$$\mu=E_C+kT\ln\left(\frac{N_n}{N_C}\right)\qquad(10.4.351)$$

and since $N_{n,T}\geq N_n$ we always have because of the logarithm which is negative, the energy of Fermi (more rigorously it is the chemical potential!) which is less than or equal to the energy of the quasi-free electrons:

$$\mu\leq E_C\qquad(10.4.352)$$

or in other words, (quasi-)free electrons have an energy higher than the Fermi energy (chemical potential...), which is in line with the non-degenerate gas approximation made earlier above. This gives a condition of major importance for the negative carriers to be the generators of the conduction in the material.

Thus, when we place ourselves at a temperature different from absolute zero, the electronic states are not all degenerate: there is spreading of occupied states in the neighbourhood of what constitutes by definition the Fermi energy level (see sections of Statistical Mechanics page 3576 and Wave Quantum Physics page 4805), effect that is as increase as the temperature is high.

10.4.6.2 Non-degenerated statistic density of positive electric charge carriers

First of all, we must know that in the present state of our knowledge, "holes" do not emerge from equations, but are an empirical construction that makes it possible to match theory and experience (positive charges of the Hall effect for example). It is therefore an artifice to make a simple theory of a question that seems rigorously nowadays not manageable by Quantum Physics.

Personally, I consider the holes in the same way as the Lagrange points in astronomy: Even if there are no bodies at these points of Lagrange this does not avoid a satellite from orbiting around them (possibility that we have not proved in the section of Astronomy) as if there was a mass even if the orbit is quasi-stable! Moreover, experiments would have shown in the early

12000s (holocene calendar) that Lagrange points appear at the level of the atom under certain ideal and simplified conditions!

That said, it must be remembered that a hole is not a missing electron! It is an idiocy (in my opinion ...) that we see in some specialized works.

At the risk of being repeated a little often, let us recall that for a fixed temperature T the probability that an electron occupies a state of energy E is given by:

$$f_{\text{FD}}(E,T,\mu)=\frac{1}{\exp\left(\dfrac{E-\mu}{kT}\right)+1} \tag{10.4.353}$$

What bring us in order to get a better approximation, we logically had written the volume density of occupied states per unit of energy:

$$D(E)\cong\frac{1}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}(E-E_0)^{1/2}\cdot\frac{1}{\exp\left(\dfrac{E-\mu}{kT}\right)+1} \tag{10.4.354}$$

This finally led to the following relation of the volumetric density of negative charge states where the presence of a mass in the relation indicates that the occupied states are by quasi-particles such as:

$$N_n=2\left(\frac{m_ekT}{2\pi\hbar^2}\right)^{3/2}\exp\left(\frac{\mu-E_C}{kT}\right) \tag{10.4.355}$$

But what about the probability that an electron does not occupy for a fixed temperature T a state of energy E and trivially given by the difference:

$$f_p(E,T,\mu)=1-f_n(E,T,\mu)=1-f_{\text{FD}}(E,T,\mu) \tag{10.4.356}$$

where the n in index is there to indicate that the distribution concerns the "negative" carriers (distribution given as we have rightly proved earlier by the Maxwell-Boltzmann distribution which follows from an approximation of the Fermi-Dirac law).

Well! We shall see that the equations lead us to the possibility to associate also to these unoccupied states a density of states with a given effective mass. We shall also see later that it will be possible to associate to these unoccupied states a positive and equal electric charge to that of the electron, hence the p in index in the preceding relation and signifying "positive".

We therefore have for these positive carriers:

$$\begin{aligned}f_p(E,T,\mu)&=1-\frac{1}{\exp\left(\dfrac{E-\mu}{kT}\right)+1}=\frac{\exp\left(\dfrac{E-\mu}{kT}\right)+1}{\exp\left(\dfrac{E-\mu}{kT}\right)+1}-\frac{1}{\exp\left(\dfrac{E-\mu}{kT}\right)+1}\\&=\frac{\exp\left(\dfrac{E-\mu}{kT}\right)}{\exp\left(\dfrac{E-\mu}{kT}\right)+1}=\frac{\exp\left(\dfrac{E-\mu}{kT}\right)}{\exp\left(\dfrac{E-\mu}{kT}\right)+1}\frac{\exp\left(\dfrac{\mu-E}{kT}\right)}{\exp\left(\dfrac{\mu-E}{kT}\right)}=\frac{1}{\exp\left(\dfrac{\mu-E}{kT}\right)+1}\end{aligned} \tag{10.4.357}$$

let us now make an approximation similar to that used for the negative carriers, that is to say:

$$f_p(E,T,\mu)\ll 1 \tag{10.4.358}$$

to impose a semi-classical regime and therefore the states of energy are by far not all occupied by the holes (there is therefore no degeneration).

This restriction requires:

$$\exp\left(\frac{\mu-E}{kT}\right)\gg 1 \tag{10.4.359}$$

Either written in the same way as for negative carriers:

$$\exp\left(-\frac{E-\mu}{kT}\right)\gg 1 \tag{10.4.360}$$

Unlike negative carriers, this imposes:

$$E-\mu\ll kT \tag{10.4.361}$$

in other words the energy must be much lower than the Fermi level (chemical potential). Physicists then write this energy E_V to distinguish it and name it "maximum energy of the valence band" (which corresponds to the maximum energy of a quasi-free hole to satisfy this condition).

Remark

The reader may observe that the conditions mentioned earlier above also impose that E is either very small in absolute value or negative. Which gives us already a track for the integration terminals to come...

We then have:

$$f_p(E,T,\mu)=\frac{1}{\exp\left(\dfrac{\mu-E}{kT}\right)+1}\cong\frac{1}{\exp\left(\dfrac{\mu-E}{kT}\right)}=\exp\left(-\frac{\mu-E}{kT}\right) \tag{10.4.362}$$

Therefore, we also have the approximation:

$$f_p(E,T,\mu)\cong\exp\left(-\frac{\mu-E}{kT}\right) \tag{10.4.363}$$

We are therefore well in a situation where classical physics predominates over quantum physics. This is why in this approximation we say that we are dealing with a "non-degenerate semiconductor" because the holes are not crammed into the highest levels available.

We have then:

$$\begin{aligned}\rho_V(E)&\cong\frac{1}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}\int\limits_{-\infty}^{E_V}(E_V-E)^{1/2}\exp\left(-\frac{\mu-E}{kT}\right)\mathrm{d}E\\&=\frac{1}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}\exp\left(\frac{-\mu}{kT}\right)\int\limits_{-\infty}^{E_V}(E_V-E)^{1/2}\exp\left(\frac{E}{kT}\right)\mathrm{d}E\end{aligned} \tag{10.4.364}$$

where the reader will have observed that the integration terminals have been chosen according to the remark we had just made earlier and that the terms in the square root have been swapped so as not to have any negative value.

To continue, we make a change of variable by putting:

$$x=\frac{E_V-E}{kT} \tag{10.4.365}$$

hence:

$$\mathrm{d}x=-\frac{1}{kT}\mathrm{d}E \qquad \text{and} \qquad E=-x\cdot kT+E_V \tag{10.4.366}$$

Therefore it comes:

$$\begin{aligned}\int_{-\infty}^{E_V}\sqrt{E_V-E}\exp\left(\frac{E}{kT}\right)\mathrm{d}E&=-\sqrt{kT}\int_{+\infty}^{0}x^{1/2}\exp\left(\frac{-x\cdot kT+E_V}{kT}\right)\mathrm{d}xkT\\&=-(kT)^{3/2}\exp\left(\frac{E_V}{kT}\right)\int_{+\infty}^{0}x^{1/2}e^{-x}\mathrm{d}x\end{aligned} \tag{10.4.367}$$

We do an integration by parts:

$$\int_{+\infty}^{0}x^{1/2}e^{-x}\mathrm{d}x=-\left[x^{1/2}e^{-x}\right]\Big|_{+\infty}^{0}+\frac{1}{2}\int_{+\infty}^{0}\frac{e^{-x}}{\sqrt{x}}\mathrm{d}x=\frac{1}{2}\int_{+\infty}^{0}\frac{e^{-x}}{\sqrt{x}}\mathrm{d}x \tag{10.4.368}$$

we then make a change of variable by putting:

$$u=\sqrt{x}\Rightarrow\frac{\mathrm{d}u}{\mathrm{d}x}=\frac{1}{2\sqrt{x}}\Rightarrow\mathrm{d}x=2\sqrt{x}\mathrm{d}u=2u\mathrm{d}u \tag{10.4.369}$$

which gives:

$$\frac{1}{2}\int_{+\infty}^{0}\frac{e^{-x}}{\sqrt{x}}\mathrm{d}x=\frac{1}{2}\int_{+\infty}^{0}\frac{e^{-u^2}}{u}2u\mathrm{d}u=\int_{+\infty}^{0}e^{-u^2}\mathrm{d}u \tag{10.4.370}$$

We have already calculated this integral in the section Statistics. It comes (since the function is even we use the property demonstrated in the section of Differential and Integral Calculus):

$$\int_{+\infty}^{0}x^{1/2}e^{-x}\mathrm{d}x=-\int_{0}^{+\infty}e^{-u^2}\mathrm{d}u=-\frac{\sqrt{\pi}}{2} \tag{10.4.371}$$

We then have finally:

$$\begin{aligned}\rho_V(E)&\cong\frac{1}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}\exp\left(-\frac{\mu}{kT}\right)(kT)^{3/2}\exp\left(\frac{E_V}{kT}\right)\frac{\sqrt{\pi}}{2}\\&=\frac{1}{2\pi^2}\left(\frac{2m_ekT}{\hbar^2}\right)^{3/2}\exp\left(\frac{E_V-\mu}{kT}\right)\frac{\sqrt{\pi}}{2}=\frac{1}{2^2\pi^{3/2}}\left(\frac{2m_ekT}{\hbar^2}\right)^{3/2}\exp\left(\frac{E_V-\mu}{kT}\right)\\&=2\left(\frac{m_ekT}{2\pi\hbar^2}\right)^{3/2}\exp\left(\frac{E_V-\mu}{kT}\right)\end{aligned} \tag{10.4.372}$$

Where, for recall, m_e is the mass of the quasi-particle (and not the mass of the hole for recall!). So after integration everything happens as if all the holes were concentrated on the level of energy E_V with a number of available places corresponding to:

$$2\left(\frac{m_e kT}{2\pi\hbar^2}\right)^{3/2} \tag{10.4.373}$$

What we write traditionally (and in a somewhat unfortunate way ... because it is not easy to remember that it is a density):

$$N_p = N_{p,T}\exp\left(\frac{E_V-\mu}{kT}\right) \tag{10.4.374}$$

or:

$$\boxed{N_p = N_V\exp\left(\frac{E_V-\mu}{kT}\right)} \tag{10.4.375}$$

where we have approximately at ambient temperature the following values of (quasi-)free states for Silicon:

$$N_V = 2\left(\frac{m_e kT}{2\pi\hbar^2}\right)^{3/2} \cong 1.1\cdot 10^{19}\;\left[\text{cm}^{-3}\right] \tag{10.4.376}$$

and for Germanium:

$$N_V = 2\left(\frac{m_e kT}{2\pi\hbar^2}\right)^{3/2} \cong 0.5\cdot 10^{19}\;\left[\text{cm}^{-3}\right] \tag{10.4.377}$$

10.4.6.3 Energy bands (conduction band, band gap, valence band)

The previous developments for negative and positive carriers have shown us that, in the approximation of a non-degenerate fermion gas, the energy of the negative carriers must be well above the Fermi level (chemical potential) and positive carrier energy well below.

It is therefore as if there was a forbidden energy interval or neither electrons nor holes have the right to be located! This energy interval is traditionally referred to as "forbidden energy band" or simply "band gap" and abbreviated B.G. The forbidden energy interval is often named simple "gap" and is denoted E_g.

Let us see this in a rough schematic form, taking care that this scheme is somewhat misleading because it gives the impression that the conduction or valence band occupies an entire block (furthermore "flat"...), whereas in reality the valence band is constituted by the last layer completely filled, the band of permitted energy which follows being named "conduction band".

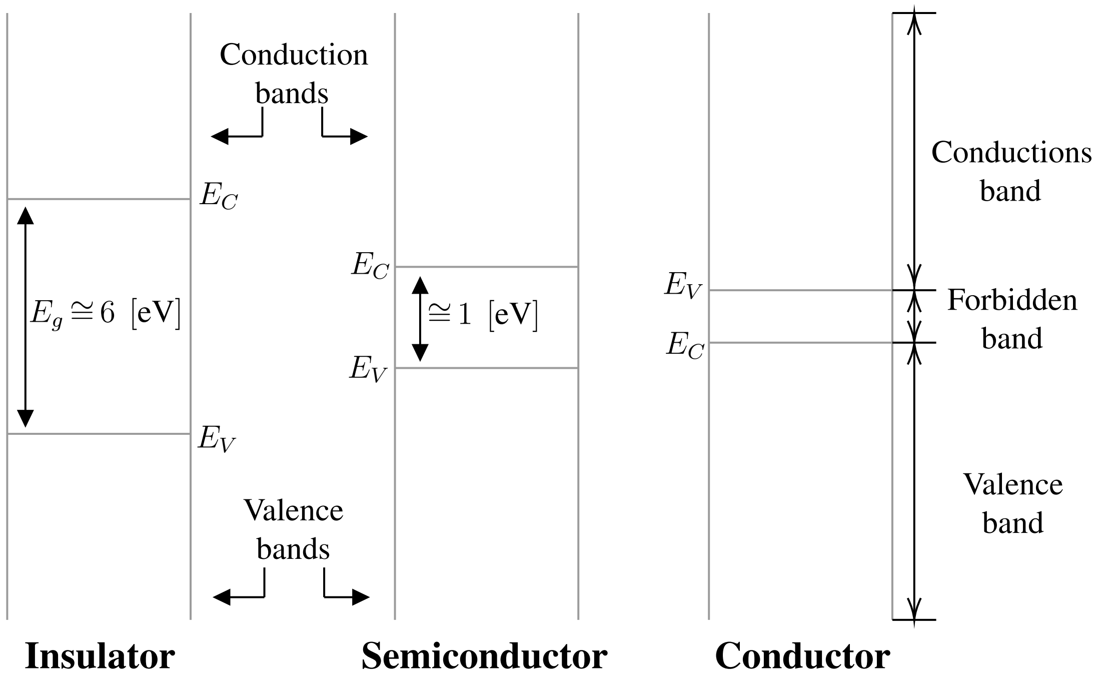

Conduction bands
E_C
$E_g \cong 6$ [eV]
E_V
E_C
$\cong 1$ [eV]
E_V
Valence bands
E_V
E_C
Conductions band
Forbidden band
Valence band
Insulator
Semiconductor
Conductor

Figure 10.106: **Representation of band structures in different materials**

Moreover, knowing that molecular chemistry makes it possible to demonstrate that structures are composed of multiple bands (as a function of the first and second quantum number), the following rigorous definitions are obtained:

D1. The "conducting band" (denoted sometimes "CB") of a solid structure is the lowest energy partially occupied or empty band (other bands are above in energy terms but will only fill at high temperatures And exist only by a theoretical description when they are empty).

D2. The "valence band" (denoted sometimes "BV") of a solid structure is the band of highest energy that is saturated, that is to say where all the states of which are occupied (knowing that there may be below the valence band other multiple bands in energy terms and all saturated).

We also have the traditional schematic association of the conduction and valence bands with the Fermi-Dirac function (which as already mentioned in all rigour should be the chemical potential at non-zero temperature!) represented in simplified form by:

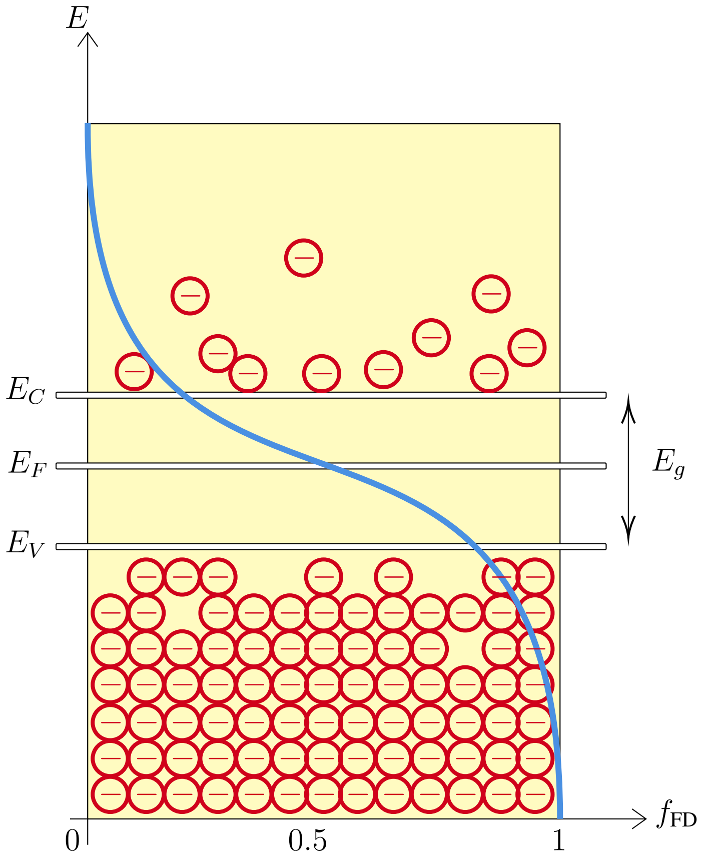


Figure 10.107: **Association of band structure with Fermi-Dirac function**

But in fact this representation, which we find almost everywhere in certain textbooks, is relatively erroneous ... since by making a semi-classical approximation by the Maxwell-Boltzmann distribution there is no longer question of rigorously representing the distribution in the form of a Fermi-Dirac distribution as the attentive reader will have noticed! This shows that we must always be useful with figures since the traditional representation of f_{FD} in the semi-classical model would indicate that there would be occupied states in the forbidden band, whereas if we represented the Maxwell-Boltzmann distribution we see two distinct functions, above and below the forbidden band!

And it must be remembered (!!!) that the above figure (even if it is quite false) represents conceptually a non-degenerate semiconductor following semi-classical approximations that we have made in our developments using the model of a non-degenerate gas (Maxwell-Boltzmann approximation) and that theoretically imposed:

$$E_C \gg \mu \gg E_V \qquad (10.4.378)$$

that many authors write (again it's unfortunate but it is so...):

$$E_C \gg E_F \gg E_V \qquad (10.4.379)$$

So there is another possible definition of the non-degenerate semiconductor: this is where the Fermi level (the chemical potential!) lies in the forbidden band and this case corresponds to how the majority of the microelectronic components work.

Remark

Let us recall (see section Statistical Mechanics page 3554) that the Maxwell-Boltzmann statistic was constructed on the assumption of the absence of interaction between the particles involved. Moreover, this statistic is constructed in the framework of Classical Mechanics and therefore applies only when the quantum effects are negligible, for example at high enough temperatures!

Here are some experimental values for common semiconductors:

E_g [eV]	300 [K]	0 [K]
C	5.47	5.51
Ge	0.66	0.75
Si	1.12	1.16
GeAs	1.43	1.53

Table 10.8: **Values of a few semiconductors gaps**

and visually and in 3D here is another figure representing de various idealized bands:

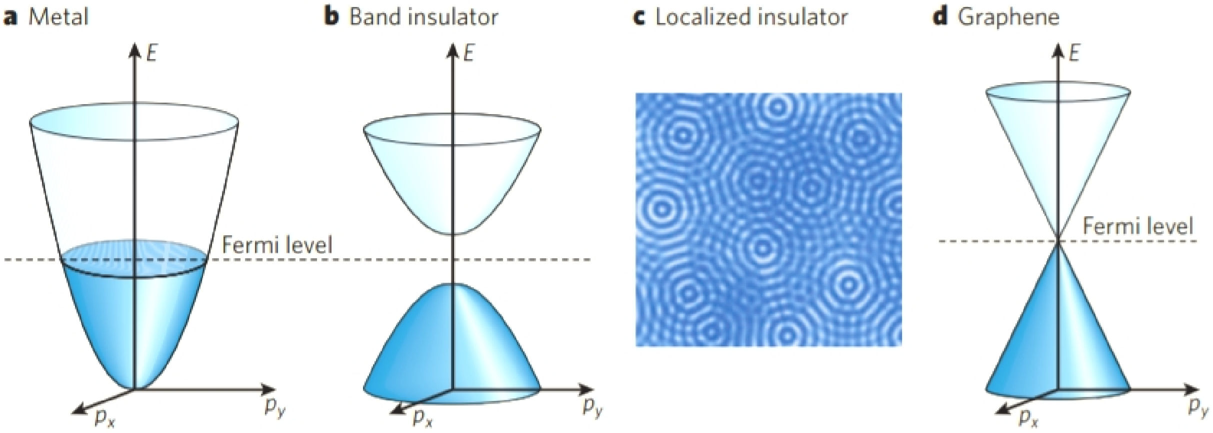


Figure 10.108: **Metal-insulator transitions (source: Carbon conductor corrupted, authors: Michael S. Fuhrer and Shaffique Adam)**

We understand then on the basis of these figures why the diamond, with a quasi-identical crystalline and atomic structure, is insulating while the silicon becomes conductive!

What is interesting for researchers is to combine materials in order to play with the width of the equation according to the needs!

Moreover, we can also conclude hastily ... that what differentiates insulators and semiconductors is the width of their forbidden band.

It should also be noted that the energy required for an electron to pass from the valence band to the conduction band can be supplied by radiation. In the case of light absorption, the energy equation of a photon may be sufficient for this as long as:

$$h\nu \geq E_g \tag{10.4.380}$$

At low temperatures, such a process is capable of making the material conductive (low-temperature space telescope technology). This property is named "photo-conductivity".

Finally, let us recall the two relations obtained above:

$$\begin{aligned}\bar{n}=\rho_C(E)=2\left(\frac{m_e kT}{2\pi\hbar^2}\right)^{3/2}\exp\left(\frac{\mu-E_C}{kT}\right)\\ \bar{p}=\rho_V(E)=2\left(\frac{m_e kT}{2\pi\hbar^2}\right)^{3/2}\exp\left(\frac{E_V-\mu}{kT}\right)\end{aligned} \tag{10.4.381}$$

The product of these two densities possesses a very interesting property. We can observe that it is independent of the position of the Fermi level and named "intrinsic density":

$$\begin{aligned}\bar{n}\cdot\bar{p}&=2\left(\frac{m_e kT}{2\pi\hbar^2}\right)^{3/2}\exp\left(\frac{\mu-E_C}{kT}\right)\cdot 2\left(\frac{m_e kT}{2\pi\hbar^2}\right)^{3/2}\exp\left(\frac{E_V-\mu}{kT}\right)\\ &=\left[2\left(\frac{m_e kT}{2\pi\hbar^2}\right)^{3/2}\right]^2\exp\left(\frac{E_V-E_C}{kT}\right)=N^2\exp\left(\frac{E_V-E_C}{kT}\right)=N^2\exp\left(\frac{E_g}{kT}\right)\end{aligned} \tag{10.4.382}$$

For example, some values of the square root of the intrinsic density at 300 [K] are given in the table below:

Material	$\bar{n}\bar{p}$
Ge	$2.5\cdot 10^{13}$ [cm^{-3}]
Si	$1.6\cdot 10^{10}$ [cm^{-3}]
GaAs	$1.1\cdot 10^{7}$ [cm^{-3}]

Table 10.9: **Some square root intrinsic density values**

We also notice that the critical density is strongly temperature dependent. These values of densities are obviously idealized, in reality these values are much lower due to imperfections (residual impurities, defects in crystallization, etc.) which locally disrupt the periodicity of the potential and, therefore, introduce energy levels which can be accessible to electrons. In contrast to the levels corresponding to the pure material, we will speak of "extrinsic levels".

10.4.6.4 Ohm's law of Semiconductors

We have proved in the framework of the Drude model that the conductivity was given by:

$$\sigma=\frac{nq^2\tau}{m} \tag{10.4.383}$$

where n is for recall the density of carriers in the material. We have also proved that the current is inversely proportional to the conductivity according to the relation:

$$U=RI=\left(\frac{1}{\sigma\dfrac{S}{L}}\right)I \tag{10.4.384}$$

In the framework of the developments made above we have seen that the density n of the carriers was given respectively by the following relations for a constant potential (model hypothesis):

$$N_n=2\left(\frac{m_e kT}{2\pi\hbar^2}\right)^{3/2}\exp\left(\frac{\mu-E_C}{kT}\right)\qquad\text{and}\qquad N_p=2\left(\frac{m_e kT}{2\pi\hbar^2}\right)^{3/2}\exp\left(\frac{E_V-\mu}{kT}\right) \tag{10.4.385}$$

where the relative masses m_e of the quasi-particle (negative carrier or positive carrier) are not necessarily equal! Thus, we have the resistance which can be approximated by a relation of the form:

$$\boxed{R=A\exp\left(\frac{B}{T}\right)} \tag{10.4.386}$$

and we easily verify this dependence by graphically representing:

$$\ln(R)=\ln(A)+\frac{B}{T} \tag{10.4.387}$$

that is $\ln(R)$ as a function of $1/T$ (the resistance therefore depends only on the theoretical temperature at constant voltage).

The real complexity lies in the fact that many terms are dependent on temperature (Fermi level, mean free travel time, etc.) and the applied potential, which means that in reality the curves obtained are by no means according to the theory....!

A numerical application shows that the carrier densities N_n and N_p therefore increase very quickly already from the ambient temperature! What is consistent with the experience with non-degenerate semiconductors because we will then have the conductivity which increases just as strongly which implies a rapid decrease in resistance!

The high sensitivity of the conductivity of certain solids to temperature variations is the cause of many applications, both for conductive metals and semiconductors. This is what we name "thermistors".

Finally, let us notice that in the case of Silicium, we have $E_g=1.12$ [eV] whereas the kinetic energy due to the thermal agitation (see section Continuum Mechanics page 3897) is given at ambient temperature by:

$$E_{c,\text{th}}=\frac{3kT}{2}=\frac{1}{2}m_e v_{th}^2\cong 0.037\ [\text{eV}] \tag{10.4.388}$$

But, we have already seen that only the electrons whose energy was close to that of the Fermi level could participate in the conduction. Their kinetic energy then given by:

$$E_{c,F}=\frac{1}{2}m_e v_F^2\cong 1.12\,[\text{eV}] \tag{10.4.389}$$

where v_F is the "Fermi velocity" (velocity of electron-wave in a conductor).

By equalizing the last two relations:

$$\frac{1}{2}m_e v_F^2=\frac{1}{2}m_e v_{\text{th}}^2\cong 0.037\,[\text{eV}] \tag{10.4.390}$$

There is therefore a ratio of a factor of 30 between the two energies, either by taking the square root, a ratio of 5 between the velocities. Therefore we have:

$$v_F \gg v_{\text{th}} \tag{10.4.391}$$

But we have already seen, in our study of the Drude model, that the thermal velocity led us to a superior average distance of one order of magnitude (factor 10) of the inter-atomic distances. And here we have a factor 5 more !!!! Thus more than 50 inter-atomic distances! The mean free path of an electron of conduction is therefore much greater than that which we had determined from the classical Drude model. Thus, the mean free path does not seem to be due to collisions with the ions of the network but it is due to the network's imperfections: structural defects, foreign atoms ...

A perfect (pure) semiconductor, without imperfections, as we have theoretically treated it so far, is named an "intrinsic semiconductor": it therefore has no impurity and its electrical behaviour depends only on the Structure of the material. This behaviour corresponds to a perfect semiconductor, that is to say without structural defects or chemical impurities. A real semiconductor is never perfectly intrinsic, but can sometimes be as close as pure mono-crystalline silicon.

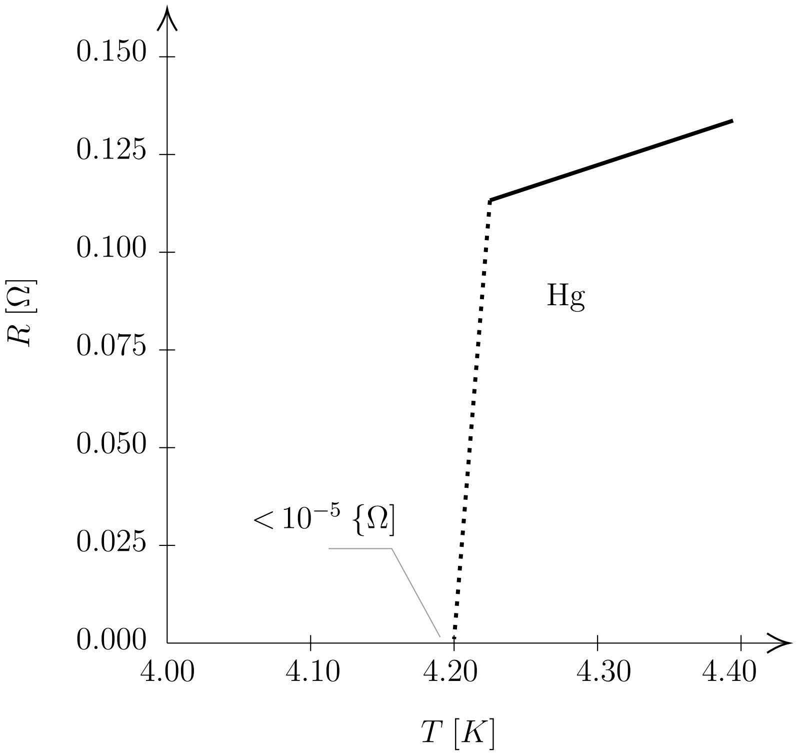
In an intrinsic semiconductor, the charge carriers are generated only by thermal excitation. The number of electrons in the conduction band is then equal to the number of holes in the valence band as shown by our theoretical model.

It must be realized that these semiconductors actually do not, or very little, conduct the current, except if we carry them at high temperature.

10.4.7 Superconductivity

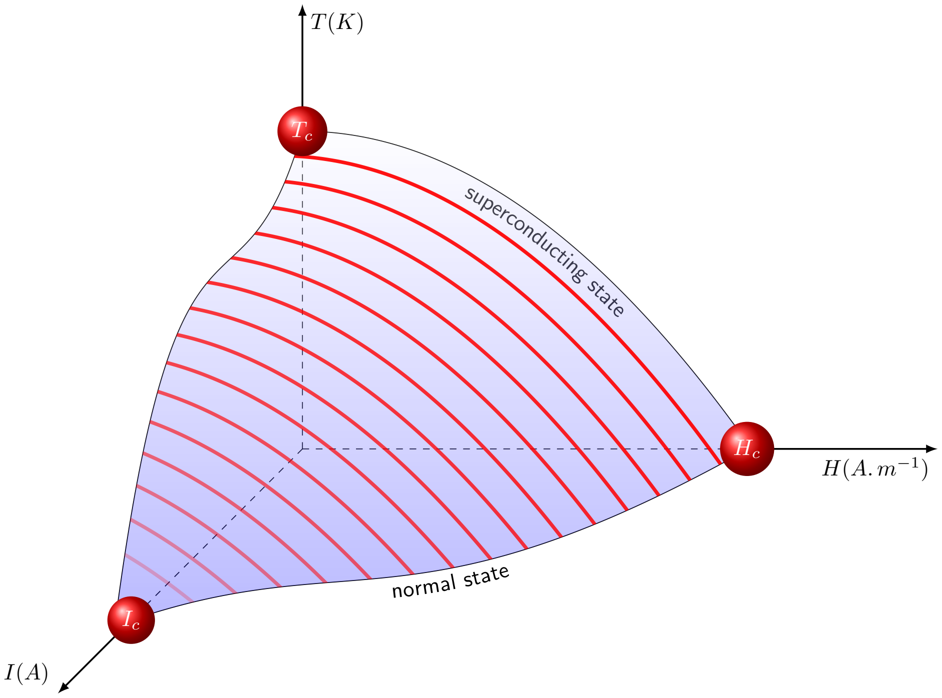
"Superconductivity" is a phenomenon of exactly zero electrical resistance and expulsion of magnetic flux fields occurring in certain materials, named "superconductors", when cooled below a characteristic critical temperature. It was discovered by Dutch physicist Heike Kamerlingh Onnes on April 8, 11911 (holocene calendar), in Leiden (he get the Nobel Prize for this discovery on with the element Mercury Hg).

Like ferromagnetism and atomic spectral lines, superconductivity is a quantum mechanical phenomenon. It is characterized by the Meissner effect, the complete ejection of magnetic field lines from the interior of the superconductor as it transitions into the superconducting state. The occurrence of the Meissner effect indicates that superconductivity cannot be understood simply as the idealization of perfect conductivity in classical physics.



Superconductors are also able to maintain a current with no applied voltage whatsoever, a property exploited in superconducting electromagnets such as those found in MRI machines. Experiments have demonstrated that currents in superconducting coils can persist for years without any measurable degradation. Experimental evidence points to a current lifetime of at least 100,000 years. Theoretical estimates for the lifetime of a persistent current can exceed the estimated lifetime of the universe, depending on the wire geometry and the temperature.

In reality the superconductivity depend of different variables as illustrated below:



10.4.7.1 Meissner effect (second London's equation)

The "Meissner effect" is the expulsion of a magnetic field from a superconductor during its transition to the superconducting state. The German physicists Walther Meissner and Robert Ochsenfeld discovered this phenomenon in 11933 (holocene calendar) by measuring the magnetic field distribution outside superconducting tin and lead samples. The samples, in the presence of an applied magnetic field, were cooled below their superconducting transition temperature. Below the transition temperature the samples cancelled nearly all interior magnetic fields. They detected this effect only indirectly because the magnetic flux is conserved by a superconductor: when the interior field decreases, the exterior field increases. The experiment demonstrated for the first time that superconductors were more than just perfect conductors and provided a uniquely defining property of the superconductor state.

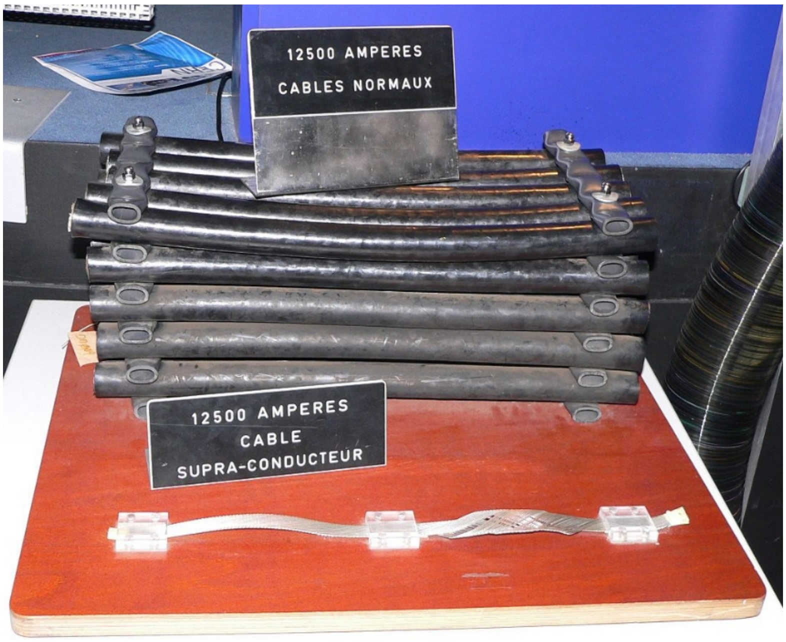


Figure 10.109: **Electric cables for accelerators at CERN (source: Wikipedia)**

A superconductor with little or no magnetic field within it is said to be in the "Meissner state". The Meissner state breaks down when the applied magnetic field is too large. Superconductors can be divided into two classes according to how this breakdown occurs:

- In "Type I superconductors", superconductivity (perfect diamagnetism behaviour) is abruptly destroyed when the strength of the applied field rises above a critical value H_c. Depending on the geometry of the sample, one may obtain an intermediate state consisting of a baroque pattern of regions of normal material carrying a magnetic field mixed with regions of superconducting material containing no field.

- In "Type II superconductors", raising the applied field past a critical value H_{c1} leads to a mixed state (also known as the "vortex state") in which an increasing amount of magnetic flux penetrates the material, but there remains no resistance to the flow of electric current as long as the current is not too large. At a second critical field strength H_{c2}, superconductivity is destroyed. The mixed state is actually caused by vortices in the electronic superfluid, sometimes named "fluxons" because the flux carried by these vortices is quantized. Most pure elemental superconductors, except niobium and carbon nanotubes, are Type I, while almost all impure and compound superconductors are Type II.

We will now determine London's second equation. It relate the external magnetic field applied on a sample to the value of this same magnetic field inside the sample.

We consider for this purpose that in the superconductor phase, the electric charge carrier of mass m, of electric charge q, of density ρ, animated of a speed v such that their kinetic energy is written:

$$E_c = \iiint_V \frac{1}{2} m\vec{v}^2 \rho \circ \mathrm{d}^3\vec{r} \qquad (10.4.392)$$

where the integral is obviously extended to the whole sample volume of the superconductor.

We have proved earlier that we also:

$$\vec{j}=\rho q\vec{v} \tag{10.4.393}$$

and in static regime, it is linked to the magnetic field $\vec{B}$ by the four Maxwell equation and if the conductor is perfect it has no capacity and we can omit the second term such that it remain:

$$\vec{\nabla}\times\vec{B}=\mu\vec{j} \tag{10.4.394}$$

Combining these relations we have:

$$\begin{aligned}E_c&=\iiint_V \frac{1}{2}m\vec{v}^2\rho\circ\mathrm{d}\vec{V}=\iiint_V \frac{1}{2}m\left(\frac{\vec{j}}{\rho q}\right)^2\rho\circ\mathrm{d}\vec{V}=\iiint_V \frac{1}{2}m\left(\frac{\vec{\nabla}\times\vec{B}}{\mu\rho q}\right)^2\rho\circ\mathrm{d}^3\vec{r}\\&=\frac{m}{2\mu^2\rho q^2}\iiint_V(\vec{\nabla}\times\vec{B})^2\circ\mathrm{d}\vec{V}=\frac{1}{2\mu}\left(\frac{m}{\mu\rho q^2}\right)\iiint_V(\vec{\nabla}\times\vec{B})^2\mathrm{d}V\\&=\frac{1}{2\mu}\left(\sqrt{\frac{m}{\mu\rho q^2}}\right)^2\iiint_V(\vec{\nabla}\times\vec{B})^2\circ\mathrm{d}\vec{V}\end{aligned} \tag{10.4.395}$$

It is important at this step to notice that we assume that in the superconductor that since $\vec{v}\neq 0$ there is like as permanent current! So if we observe the Meissner effect it means that implicitly that there is something like a current inside the superconductor sample and this without that we need to provide any external potential (voltage)!

The term in parenthesis has the dimension of a length and is named "London penetration depth" and denoted λ_L such that the latter equality is written:

$$E_c=\frac{1}{2\mu}\lambda_L^2\iiint_V\vec{\nabla}\times\vec{B}\circ\mathrm{d}\vec{V} \tag{10.4.396}$$

Furthermore, we have prove in the section of Electrodynamics that the energy density of a monochromatic wave was given by:

$$u_B=\frac{1}{2\mu}\hat{B}^2 \tag{10.4.397}$$

Therefore we assume that to the external magnetic field energy is given by:

$$E_B=\iiint_V\frac{1}{2\mu}\vec{B}^2\circ\mathrm{d}\vec{V} \tag{10.4.398}$$

The total energy is the written:

$$E_{\text{tot}}=\frac{1}{2\mu}\iiint_V\left(\vec{B}^2+\lambda_L^2(\vec{\nabla}\times\vec{B})^2\right)\circ\mathrm{d}\vec{V} \tag{10.4.399}$$

The total energy should be, by the Maupertuis principle, minimal for the real magnetic field profile that we will denote $B^*(\vec{r})$ and that we are looking for.

The variational δE_{tot} is then equal to zero if the profile $B(r)$ departs only in an infinitesimal way from the real profile from a quantity $\delta B(\vec{r})$. Therefore:

$$\begin{aligned}\delta E_{\text{tot}}&=\frac{1}{2\mu}\iiint_V\left(2\vec{B}\delta\vec{B}+2\lambda_L^2(\vec{\nabla}\times\vec{B})\delta(\vec{\nabla}\times\vec{B})\right)\circ\mathrm{d}\vec{V}=\frac{1}{\mu}\iiint_V\left(\vec{B}\delta\vec{B}+\lambda_L^2(\vec{\nabla}\times\vec{B})\delta(\vec{\nabla}\times\vec{B})\right)\circ\mathrm{d}\vec{V}\\&=\frac{1}{\mu}\iiint_V\left(\vec{B}\delta\vec{B}+\lambda_L^2(\vec{\nabla}\times\vec{B})(\vec{\nabla}\times\delta\vec{B})\right)\circ\mathrm{d}\vec{V}=\frac{1}{\mu}\iiint_V\left(\vec{B}\delta\vec{B}+\lambda_L^2\mathrm{rot}(\vec{B})\mathrm{rot}(\delta\vec{B})\right)\circ\mathrm{d}\vec{V}\end{aligned} \tag{10.4.400}$$

But we have introduced in the section of vector Calculus the following identity:

$$\vec{\nabla}(\vec{x}\times\vec{y})=\vec{y}\cdot(\text{rot}(\vec{x}))-\vec{x}\cdot(\text{rot}(\vec{y})) \tag{10.4.401}$$

Therefore with $\vec{x}=\text{rot}(\vec{B})$ and $\vec{y}=\delta\vec{B}$ we have:

$$\vec{\nabla}(\text{rot}(\vec{B})\times\delta\vec{B})=\delta\vec{B}\cdot(\text{rot}(\text{rot}(\vec{B})))-\text{rot}(\vec{B})\cdot(\text{rot}(\delta\vec{B})) \tag{10.4.402}$$

Therefore:

$$\text{rot}(\vec{B})\cdot(\text{rot}(\delta\vec{B}))=\delta\vec{B}\cdot(\text{rot}(\text{rot}(\vec{B})))-\vec{\nabla}(\text{rot}(\vec{B})\times\delta\vec{B}) \tag{10.4.403}$$

Therefore our variational can be written:

$$\begin{aligned}\delta E_{\text{tot}}&=\frac{1}{\mu}\iiint_V\left(\vec{B}\delta\vec{B}+\lambda_L^2\text{rot}(\vec{B})\text{rot}(\delta\vec{B})\right)\circ\text{d}\vec{V}\\&=\frac{1}{\mu}\iiint_V\left[\vec{B}\delta\vec{B}+\lambda_L^2\left(\delta\vec{B}\cdot(\text{rot}(\text{rot}(\vec{B})))-\vec{\nabla}(\text{rot}(\vec{B})\times\delta\vec{B})\right)\right]\circ\text{d}\vec{V}\\&=\frac{1}{\mu}\left\{\iiint_V\left[\vec{B}+\lambda_L^2\text{rot}(\text{rot}(\vec{B}))\right]\cdot\delta\vec{B}\circ\text{d}\vec{V}-\lambda_L^2\iiint_V\vec{\nabla}(\text{rot}(\vec{B})\times\delta\vec{B})\circ\text{d}\vec{V}\right\}\end{aligned} \tag{10.4.404}$$

Now let us recall the Ostrogradsky theorem proved in the section of Vector Calculus:

$$\iiint_V(\vec{\nabla}\circ\vec{f})\text{d}V=\iiint_V\vec{\nabla}(\vec{f})\circ\text{d}\vec{V}=\oiint_S\vec{f}\circ\text{d}\vec{S} \tag{10.4.405}$$

Therefore:

$$\begin{aligned}\delta E_{\text{tot}}&=\frac{1}{\mu}\left\{\iiint_V\left[\vec{B}+\lambda_L^2\text{rot}(\text{rot}(\vec{B}))\right]\cdot\delta\vec{B}\circ\text{d}\vec{V}-\lambda_L^2\iiint_V\vec{\nabla}(\text{rot}(\vec{B})\times\delta\vec{B})\circ\text{d}\vec{V}\right\}\\&=\frac{1}{\mu}\left\{\iiint_V\left[\vec{B}+\lambda_L^2\text{rot}(\text{rot}(\vec{B}))\right]\cdot\delta\vec{B}\circ\text{d}\vec{V}-\lambda_L^2\oiint_S(\text{rot}(\vec{B})\times\delta\vec{B})\circ\text{d}\vec{S}\right\}\end{aligned} \tag{10.4.406}$$

The surface integral is always zero since, by hypothesis, $\vec{B}$ is fixed on the sample boundary ($\delta\vec{B}(r_{\text{surface}})=0$), therefore it remains:

$$\delta E_{\text{tot}}=\frac{1}{\mu}\iiint_V\left[\vec{B}+\lambda_L^2\text{rot}(\text{rot}(\vec{B}))\right]\cdot\delta\vec{B}\circ\text{d}\vec{V} \tag{10.4.407}$$

The real $B^*(\vec{r})$ we are looking for is such that:

$$\delta E_{\text{tot}}=\frac{1}{\mu}\iiint_V\left[B^*(\vec{r})+\lambda_L^2\text{rot}(\text{rot}(B^*(\vec{r})))\right]\cdot\delta B^*(\vec{r})\circ\text{d}\vec{V}=0 \tag{10.4.408}$$

Therefore we can try the trivial solution:

$$B^*(\vec{r})+\lambda_L^2\text{rot}(\text{rot}(B^*(\vec{r})))=\vec{0} \tag{10.4.409}$$

But we have seen in the section of Vector Calculus that:

$$\text{rot}(\text{rot}(\vec{f}))=\overrightarrow{\text{grad}}(\text{div}(\vec{f}))-\Delta\vec{f} \tag{10.4.410}$$

Therefore (don't forget that the divergence of the magnetic field is always equal to zero!):

$$\text{rot}(\text{rot}(B^*(\vec{r}))) = \overrightarrow{\text{grad}}(\text{div}(B^*(\vec{r}))) - \Delta B^*(\vec{r}) = -\Delta B^*(\vec{r}) \tag{10.4.411}$$

Therefore:

$$B^*(\vec{r}) - \lambda_L^2 \Delta B^*(\vec{r}) = \vec{0} \tag{10.4.412}$$

Hence:

$$\Delta B^*(\vec{r}) - \frac{1}{\lambda_L^2} B^*(\vec{r}) = \vec{0} \tag{10.4.413}$$

Now it must be notice that we cannot have B_x that of x if the sample is put in a magnetic field that is static and uniform otherwise we would not have $\text{div}(B^*(\vec{r})) = 0$. And as we need $\text{rot}(B^*(\vec{r})) = \mu\vec{j}$ then as simple solution is to suppose that each component B_i depends only on a coordinate that is perpendicular to itself. Therefore we will choose B_z as depending on x such that the latter differential equation will be written for this component:

$$\frac{\partial^2 B_z(x)}{\partial x^2} - \frac{1}{\lambda_L^2} B_z(x) = 0 \tag{10.4.414}$$

A special simple solution is obviously:

$$\boxed{B_z(x) = B_0 e^{-x/\lambda_L}} \tag{10.4.415}$$

This relation is knows as the "second London's equation". The fact that the magnetic field decrease exponential inside the material is interpreted (...) in many textbooks and by many physicists as the fact the superconductor rejects the magnetic field since in laboratories we have the following observation:

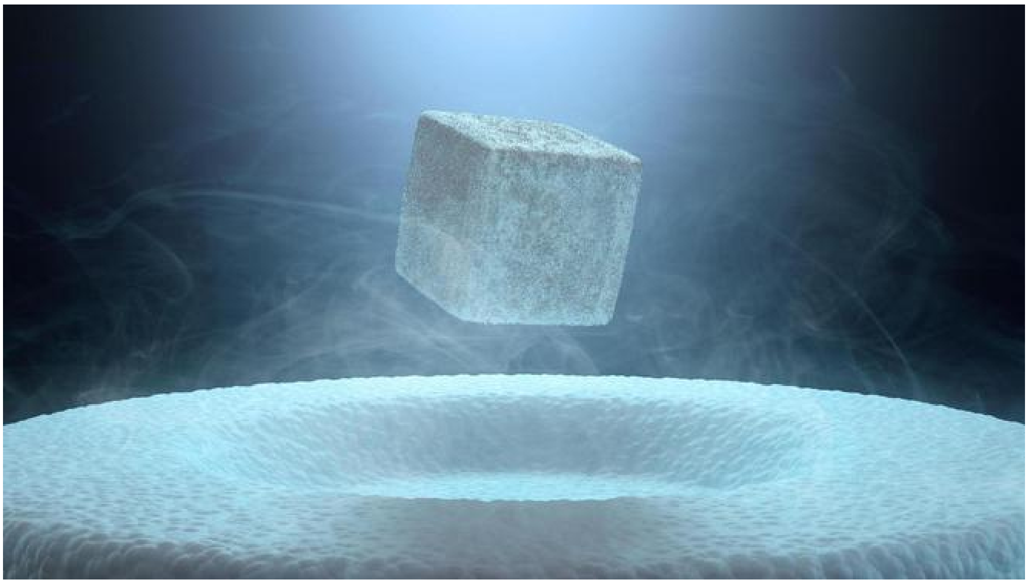
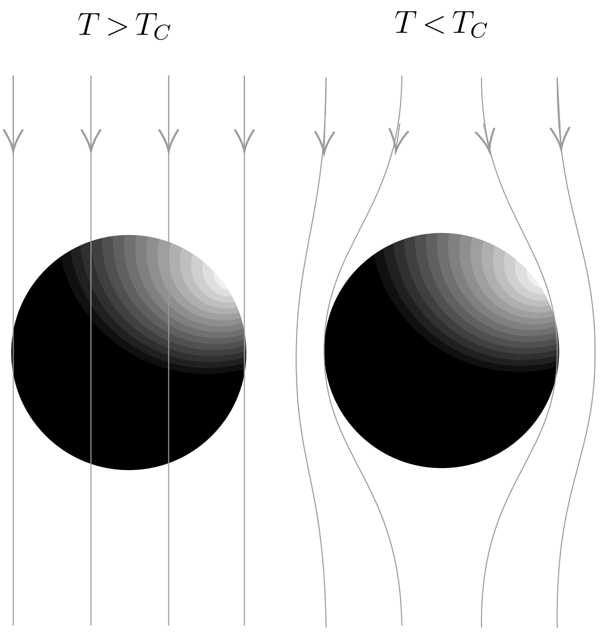


Figure 10.110: **Observation of a superconductor in levitation in magnetic field due to Meissner effect**

This bring us to the following well known misleading figure below that we can see on many Internet sites and in many textbooks:



The derivation above proposed by Pierre-Gilles de Gennes for undergraduate class that don't use Quantum Physics is however a coincidence (as there exists many others in physics)! Obviously we see that this derivation of the London equation is very not appropriate. In this derivation the fact that the metal is in superconducting state is not used at all for example... Therefore, if this derivation is correct, any metal should show the Meissner effect.

∼40 %

★★★★☆

23 votes, 66.09%

10.5. Optics (ray optics)

OPTICS is the study of the fraction of radiant energy sensitive to the retina, that is to say the "light" or said more generally: the "electromagnetic waves" (see section Electrodynamics page 4138) and over a wide frequency band which is not limited (depending on the case studies) to visible light!

We chose in this textbook to split the study of "geometrical optics" or simply "optics" into three parts: the photometry (see below), geometrical optics (this section) and wave optics (next section).

Definitions (#10.5.481):

D1. "Photometry" is responsible for the study of some of the variables definitions of energy properties of electromagnetic waves with respect to the visual sensitivity.

D2. "Geometrical optics" is responsible for the study that describes the propagation of light in transparent media without involving the nature of light. This is a part of physics has the advantage not to require complicated mathematical tools, but a lot of good geometrical sense...

D3. "Wave optics" is responsible for the study where the luminous phenomena are interpreted taking into account of the nature of light. This is considered as an electromagnetic wave of a given wavelength defining its color (subjective concept as we will discussed below). This use much more complicate tools as Fourier transforms and Green theorem for example.

In some experiments, however, we must consider light as a corpuscular phenomenon (see section Wave Quantum Physics). We assume then that it is formed by particles, "photons" whose energy is proportional to the light frequency according to the Planck's law (not the Planck's law of thermodynamics ... the other one!).

For consistency reasons, as we have already mentioned, we have chosen to put photometry in the section of Geometric Optics (here so...).

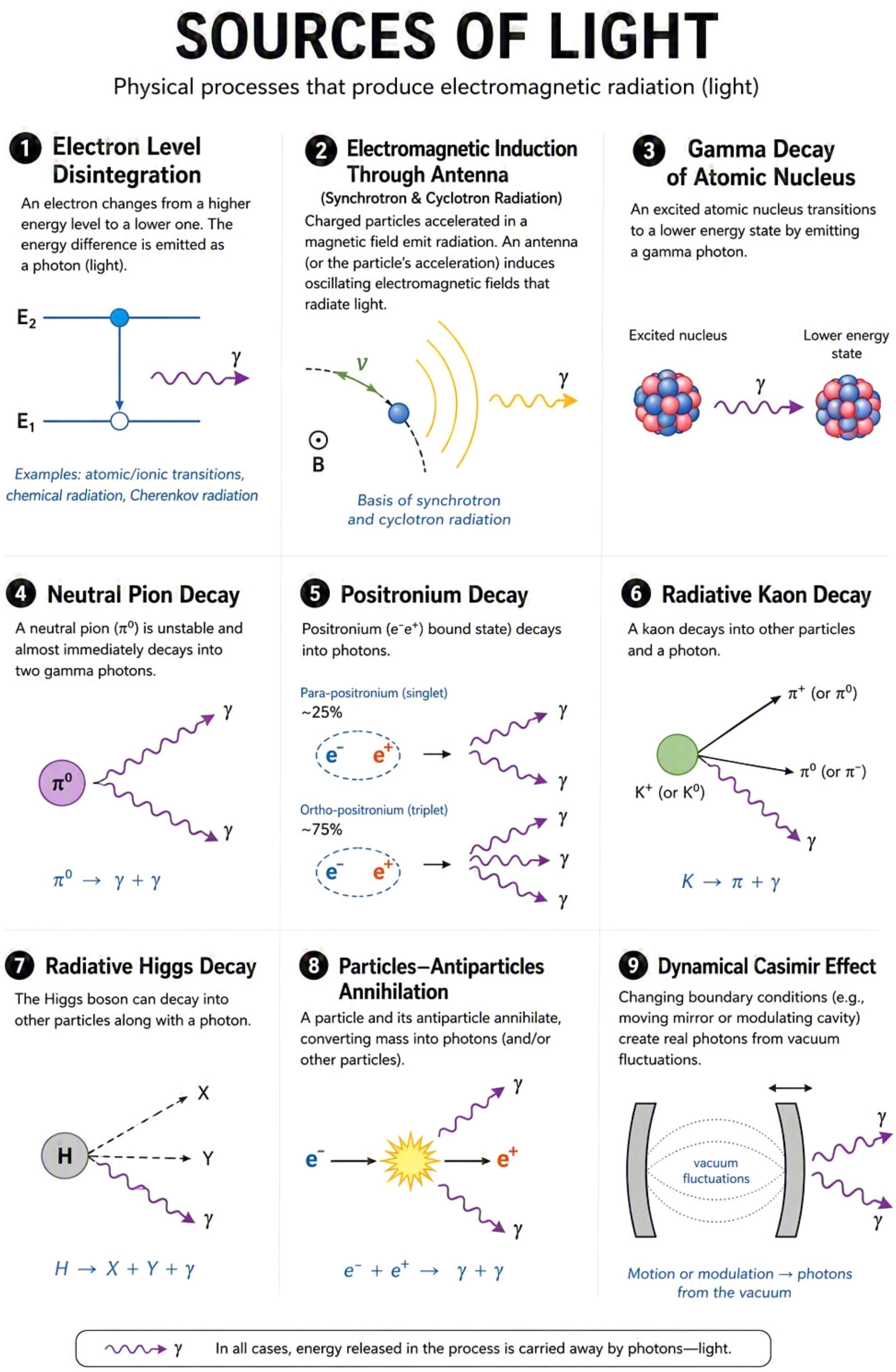
Before we start studying the mathematics of geometrical optics, it seemed good to us clarify some blurred areas of optics that are rarely clearly defined or not even treated at all in most physics books. Thus, we will first present what is a source of light or an absence of light and then how colors are seen and treated by humans (and most similar animals).

Remark

It would be pretentious to claim to do with this section as well and also complete as the *Optics* textbook of Eugene Hecht that is a priori unrivalled in content and quality to this date. It is therefore strongly recommended to refer to it if the reader wants to drive full information about Optics ([239])!

10.5.1 Sources and Shadows

Experience teaches us that in a homogeneous and transparent medium light travels in a straight line (at least in a flat space) and that the latter always comes from "light sources".



SOURCES OF LIGHT

Physical processes that produce electromagnetic radiation (light)

1 Electron Level Disintegration

An electron changes from a higher energy level to a lower one. The energy difference is emitted as a photon (light).

Examples: atomic/ionic transitions, chemical radiation, Cherenkov radiation

2 Electromagnetic Induction Through Antenna

(Synchrotron & Cyclotron Radiation)

Charged particles accelerated in a magnetic field emit radiation. An antenna (or the particle's acceleration) induces oscillating electromagnetic fields that radiate light.

Basis of synchrotron and cyclotron radiation

3 Gamma Decay of Atomic Nucleus

An excited atomic nucleus transitions to a lower energy state by emitting a gamma photon.

Excited nucleus
γ
Lower energy state

4 Neutral Pion Decay

A neutral pion (π^0) is unstable and almost immediately decays into two gamma photons.

π^0
γ
γ

$\pi^0 \rightarrow \gamma + \gamma$

5 Positronium Decay

Positronium (e^-e^+) bound state) decays into photons.

Para-positronium (singlet)
~25%
e^- e^+
γ
γ
Ortho-positronium (triplet)
~75%
e^- e^+
γ
γ
γ

6 Radiative Kaon Decay

A kaon decays into other particles and a photon.

π^+ (or π^0)
π^0 (or π^-)
K^+ (or K^0)
γ

$K \rightarrow \pi + \gamma$

7 Radiative Higgs Decay

The Higgs boson can decay into other particles along with a photon.

$H \rightarrow X + Y + \gamma$

8 Particles–Antiparticles Annihilation

A particle and its antiparticle annihilate, converting mass into photons (and/or other particles).

e^-
e^+
γ
γ

$e^- + e^+ \rightarrow \gamma + \gamma$

9 Dynamical Casimir Effect

Changing boundary conditions (e.g., moving mirror or modulating cavity) create real photons from vacuum fluctuations.

vacuum fluctuations
γ
γ

Motion or modulation → photons from the vacuum

γ In all cases, energy released in the process is carried away by photons—light.

Some objects are illuminated by themselves (Sun, Flames, etc.). Some other objects are not generally visible in the dark in the visible bandwidth (when there is absolutely no source of light) but if they are illuminated they send back all or part of light in some or all directions and therefore behave like light sources but with less intensity as the original one.

Definitions (#10.5.482):

D1. A "punctual source" is a unique "bright spot".

D2. An "extended source" is a set being a sum of punctual sources.

D3. A "light ray" is any straight line (or "geodesic"...) that light will follow.

D4. A "reversible light ray" is a light ray that will follow the exactly same path backwards if we reverse its direction of propagation.

D5. A "light beam" is a set of light rays.

D6. A "ray diagram" is used to determine the image location, size, orientation and type of image formed by objects when place at a given location from a mirror or lens. They proved useful information about object-image relations and focal-geometry properties.

D7. An "apparent diameter" is the angle, usually small, under which we see one of the dimensions of an object (angle in radians as seen in the section of Trigonometry).

Light passes through empty space (assumed as because we will see during study of quantum field theory that there are not such things as "empty space") without undergoing alteration (on small distance at least...). Thus is how Sunlight, before reaching the limit of the Earth's atmosphere, travel through huge empty spaces without undergoing major transformations.

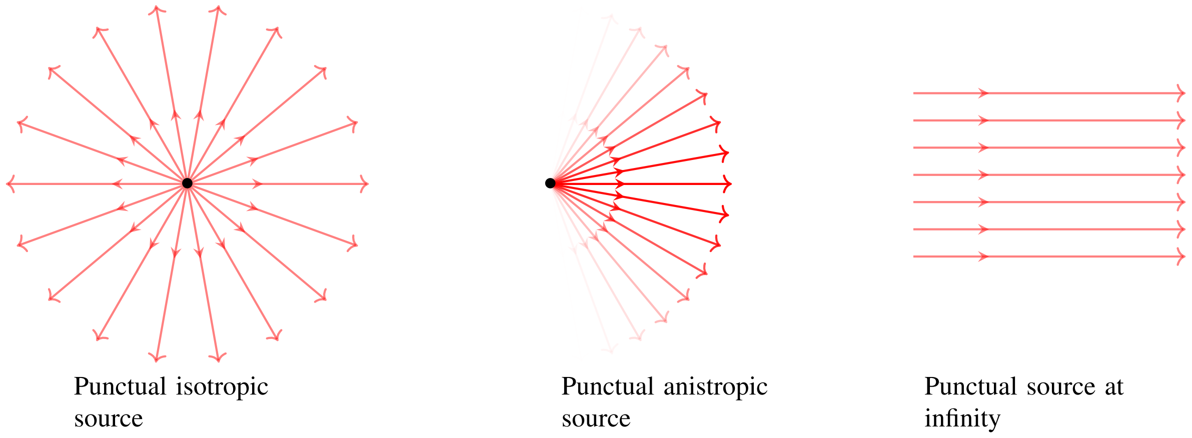

Punctual isotropic source
Punctual anistropic source
Punctual source at infinity

Figure 10.111: **Various types of light punctual sources (author: Jimmy Roussel)**

On Earth, between a luminous object and the seeing eye of the object, the light passes through a certain thickness of air. The object remains visible in other gases, or through a sheet of glass, mica, cellophane ..., or even through a layer of water, alcohol, glycerine ... such bodies are "transparent media" (even if they alter the quality of the image).

Most bodies do not let light pass through. Placed between the eye and a luminous object, they suppress the vision of the object: then we say that they are "opaque bodies".

In fact, no substance is perfectly transparent and propagation in a transparent medium is always accompanied by a weakening. This phenomenon of absorption depends on the nature of the medium and increases with the thickness of material traversed. Thus the water, even very pure, is opaque at a thickness of a few hundred meters. Also the deep seabed never receive sunlight.

Sometimes some bodies, named "translucent bodies", let light filter without allowing the eye to identify the object that emits light. These are frosted glass, striped glass, thin porcelain, oiled paper...

In a dark space, the eye located outside the path of light, sees this path thanks to fine solid particles (dust, tobacco smoke, fog, etc.) suspended in the air. These illuminated particles scatter the light they receive, becoming as many bright spots which materialize the volume traversed by the light. The familiar observation shows that these light volumes still seem limited by straight lines of light.

We can therefore apply the theorem of Thales to some light phenomena. So imagine the following experiment:

We create quite small dimensions light sources that we can consider as point sources (i.e. bright spots).

Consider S is such a point source of light. Consider the volume that the source S illuminates through an aperture in a diaphragm located in the path of the light at a distance d. If we denote by AB the circular diameter of the opening diaphragm K and that we cut the light path by a screen E, parallel to K and at a distance D from the source, we would observe that the irradiated region is limited by a disc of diameter $A'B'$.

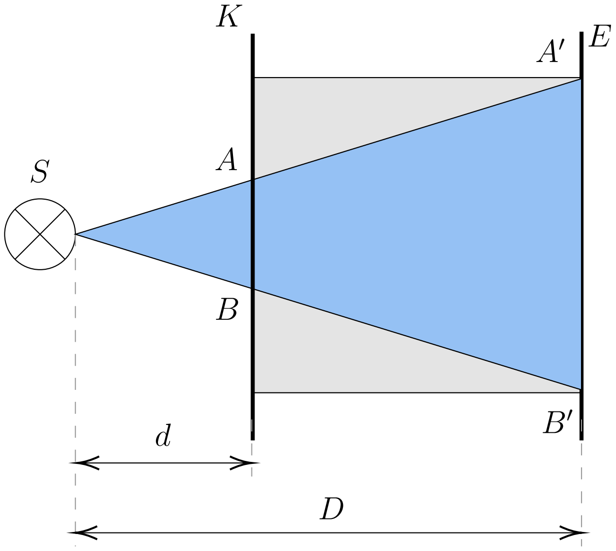


Figure 10.112: Application of Thales' theorem on punctual sources

If we can measure the diameters AB and $A'B'$ of the two discs and their distances d and D to the source, we would find that they satisfy the Thales' theorem and therefore that:

$$\frac{A'B'}{AB}=\frac{D}{d} \tag{10.5.1}$$

It is also proof that the light volume is effectively limited by straight lines issues from S and leaning on the edge of the aperture.

These observational facts and basic experience suggest the following hypothesis:

In a homogeneous transparent medium (remember that a medium is homogeneous when all is infinitesimal volumes have the same physical properties) light coming from a light spot propagates along straight lines from this point. These lines are named as we already know "light rays".

If we return to the previous above figure, all the light rays contained in the cone defined by the source S and the diaphragm K is also as we already know a "light beam".

Definitions (#10.5.483):

D1. Light travelling here from S we say then that the ray "diverges" or that the beam is a "divergent beam" (at the opposite of the LASER for those that know what is a LASER).

D2. When a point source is at infinity (as is almost a star in the night sky for example), the rays that come from it are considered as parallel and the beams that they form are named "parallel beam" or "cylindrical beams".

D3. Thanks to a convergent lens (magnifying glass for example), we will see that it is possible to change the direction of rays from a punctual source and make the concentrate on a point S'. Such a set of rays then constitutes a "convergent beam".

A very narrow light beam takes the name of "light pen". For example, the rays from a light spot on the eye always form a very slender pencil of light, because the distance from the observed eye is necessarily large compared to the diameter of the pupil.

If we return to our experience with the diaphragm: If we decrease the opening of the diaphragm that will limit a light beam to a light pen, we observe (when the diameter is reduced to less than a few tenths of a millimetre) that the trace of the brush on a screen E, rather than diminish, expands (!) and this is an evidence that the light reaches now points outside of the cone $SA'B'$.

It is as if the very small aperture AB itself became a punctual source: we say that the light is "diffracted". We will return later to this property of light because it is a fairly elaborate mathematical study (see section Wave Optics page 4492) and therefore complex to handle but still very very interesting (having huge implications because related to quantum physics!).

Now consider a punctual source of light. Between the source and a screen E, let us interpose an opaque body of any shape. Under the assumption of rectilinear propagation, we observe a "shadow cone" limited by the rays that are based on the outline of the interposed body.

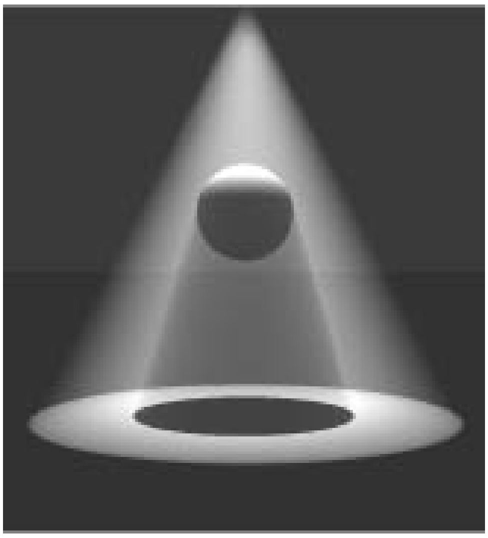


Figure 10.113: Shadow cone with proper shadow example

Definitions (#10.5.484):

D1. The non-illuminated region of an opaque body is the "proper shadow" or "umbra".

D2. The non-illuminated region of an opaque body that matches the screen is the "projected shadow".

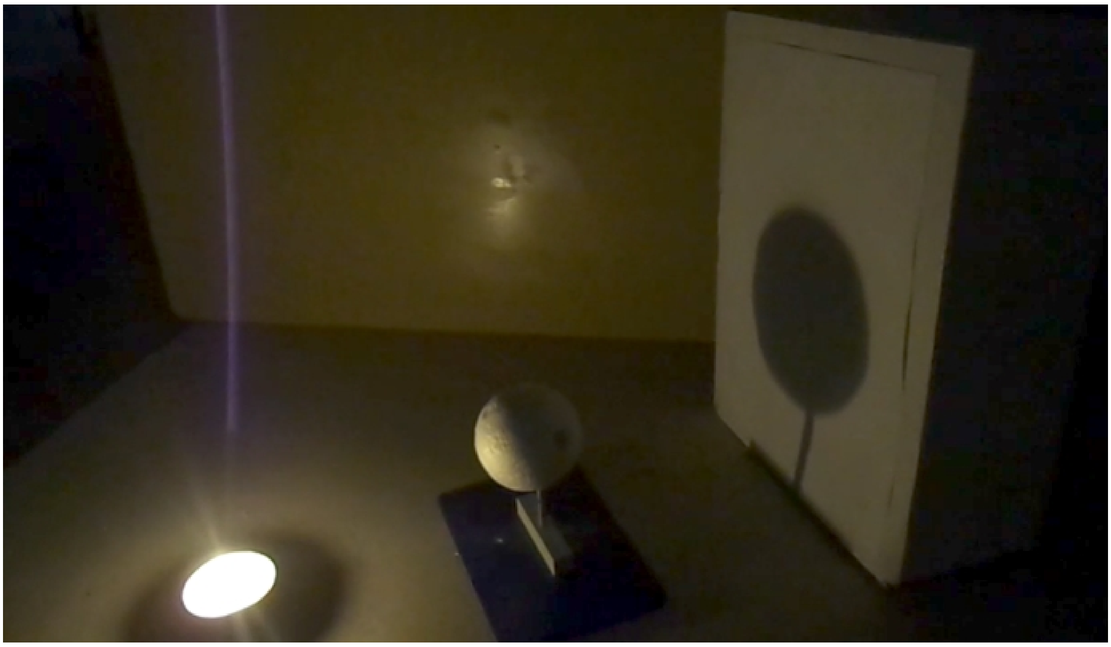


Figure 10.114: Shadow of a ball with a point like source of light

If the light source is extended, the proper shadow and projected shadow have no longer clearly defined contours. Their edges are surrounded by an intermediate area which is named the "penumbra". A well known example is given by the figure below (obviously not to scale[11]...!):

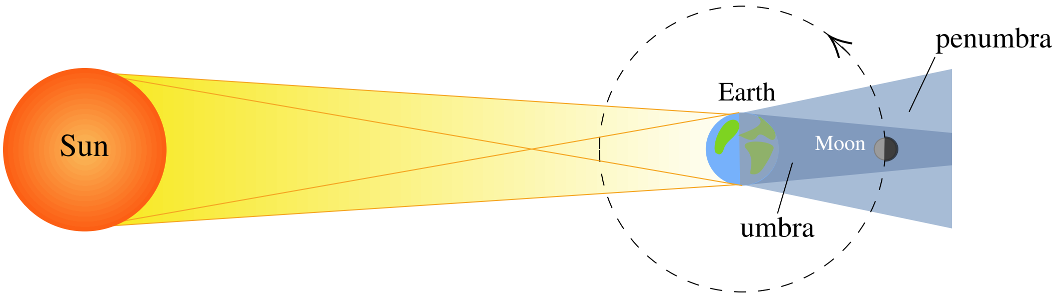

penumbra
Earth
Sun
Moon
umbra

Figure 10.115: Shadow cone with penumbra example

Below we can see an experimental configuration with an extended source of light to see the penumbra of a ball:

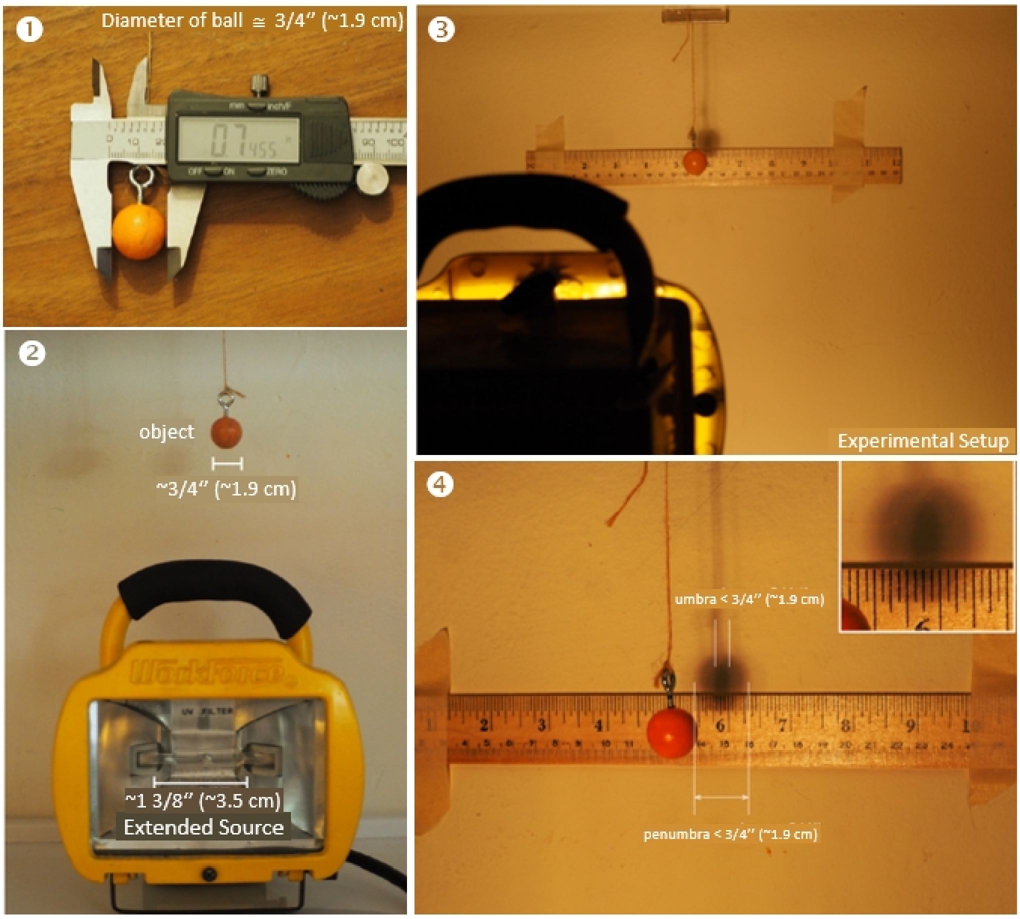

1
Diameter of ball ≅ 3/4" (~1.9 cm)
2
object
~3/4" (~1.9 cm)
~1 3/8" (~3.5 cm)
Extended Source
3
Experimental Setup
4
umbra < 3/4" (~1.9 cm)
penumbra < 3/4" (~1.9 cm)

Figure 10.116: Penumbra and umbra of a ball with an extended like source of light (credit: L. Hedberg City College New-York Planetarium)

Definition (#10.5.485): As the "shadow" (whatever the type of shadow, "proper shadow/umbra" or "projected shadow") is an area defined by the light that surrounds it, then obviously it is quite common to say that «*the shadow is refracted*» or that «*the shadow is reflected*» when in reality it is the light surrounding the shadow that is reflected or refracted.

---

[11] Don't forget that the Sun has actually a diameter of approximately 109 times that of the Earth and that the angular size of the Sun is of 0.53° (or 32 arcminutes). Hence even if in the figure above - that is not to scale for reminder... - the light rays seem significantly not parallel, in reality there are almost parallel!!

As we can see in the photo below, the mirrors seems to reflect the various shadows of the surrounding room when in reality it is the light surrounding that shadows that is reflected!:

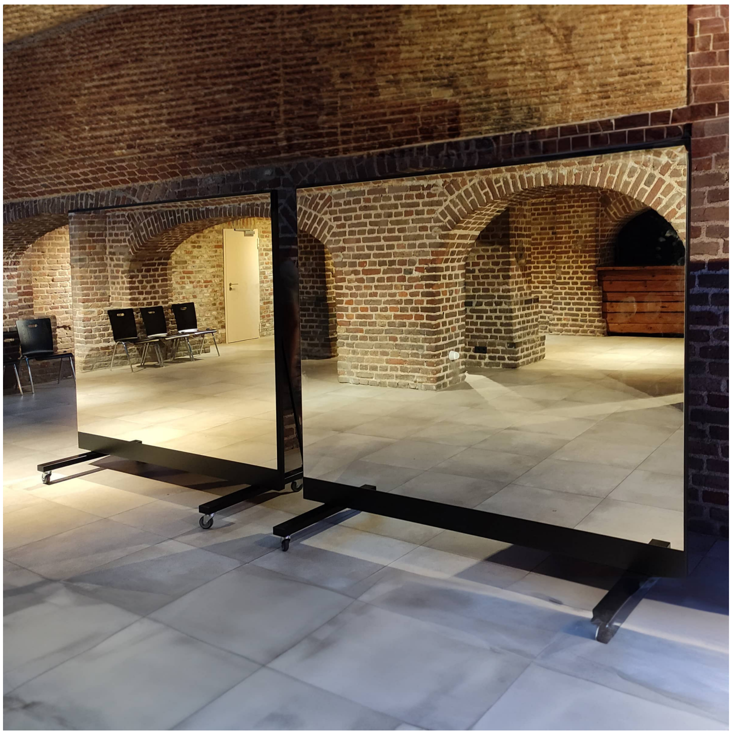


Figure 10.117: **Reflection of shadows (light surrounding shadows) by mirrors**

And here below an example of the shadow of a pen that is refracted by a recipient of glass with water:

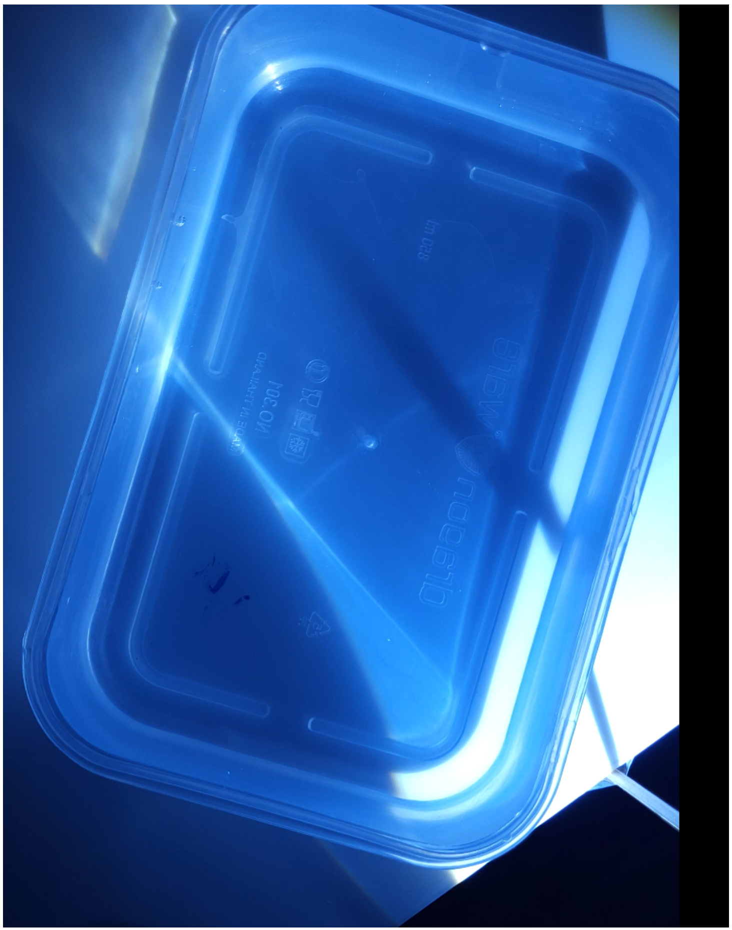


Figure 10.118: **Refraction of shadows (light surrounding shadows) by mirrors**

We may even define also a third type of shadow!

Definition (#10.5.486): A "caustic shadow" is an optical phenomenon where light, after being reflected or refracted by a curved or transparent surface (like glass or water), creates intricate patterns of brightness and shadow on another surface by interference.

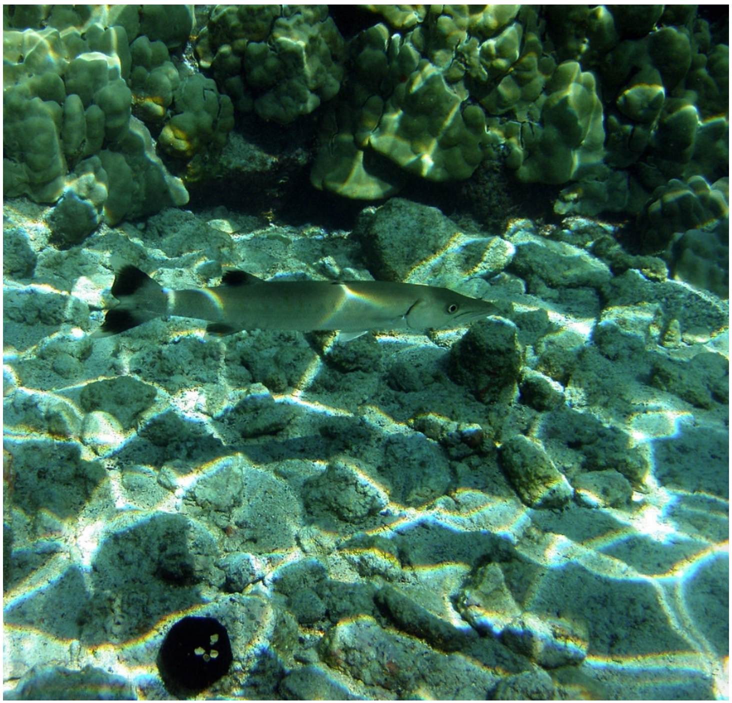


Figure 10.119: **Caustic on water surface**

To close this topic on shadows, let us mention the "magnetic shadow effect" that refers to a perceptual illusion where overlapping, blurry shadows (bokeh) in out-of-focus images appear to attract and merge, creating "blisters". For example, the sequence below from show such an effect:

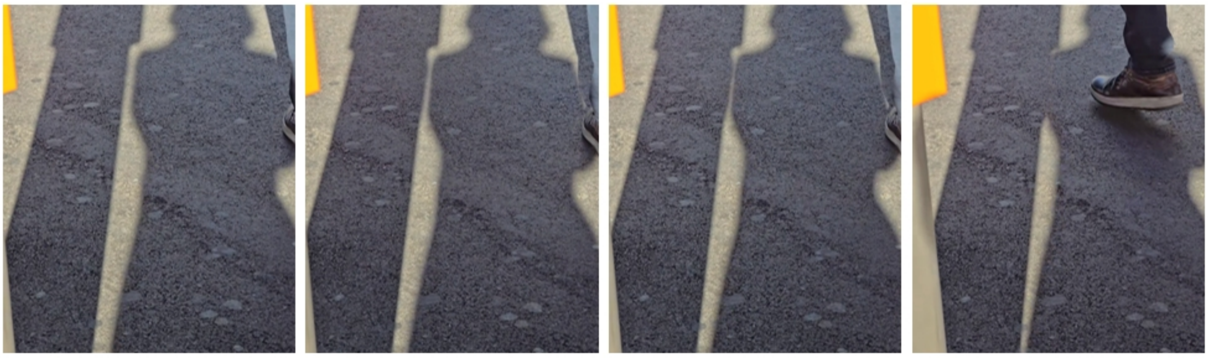


Figure 10.120: **Magnetic shadow effect**

The shadow the reader can see above is the really dark part (umbra). Imagine various lights behind you (the Sun rays). The part that blocks most rays is the darkest. Other parts block some rays but not all, so they are only slightly dark (penumbra). When slightly dark parts intersect, they get darker, like the figure below

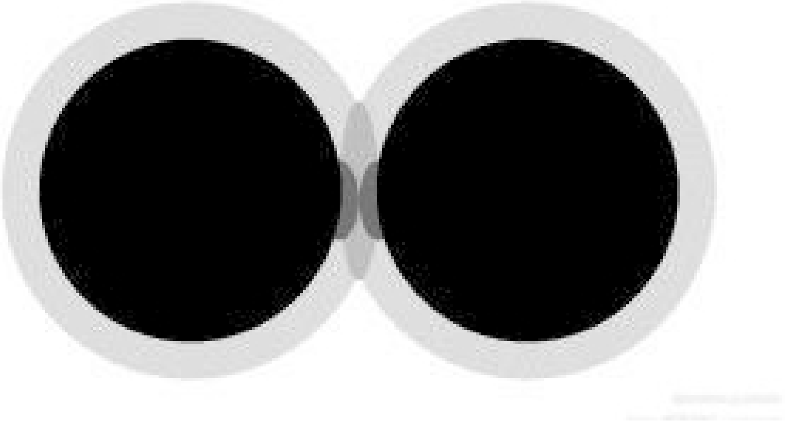


Figure 10.121: **Magnetic shadow effect explained**

By the way... Shadow patterns are sometimes difficult for the human brain to grasp (particularly for the brains of an unfortunate yet significant segment of the population that fails spatial cognitive tests like the quite famous one made by Ramful, A., Lowrie, T., and Logan, T.):

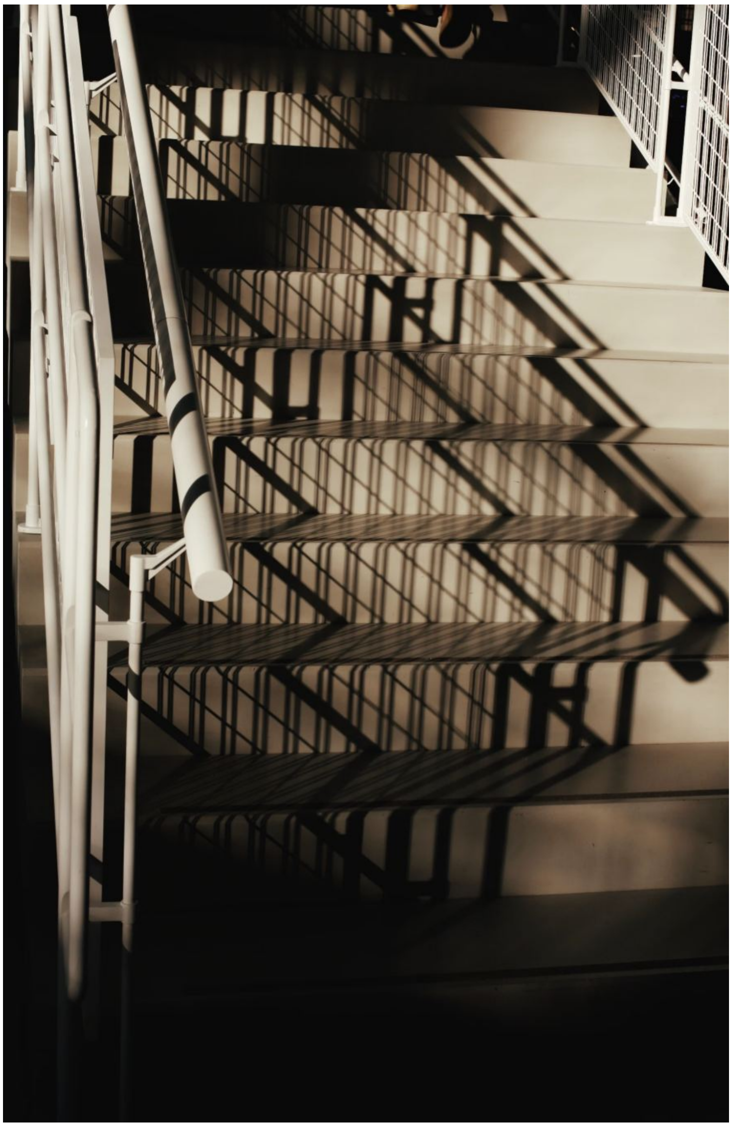


Figure 10.122: **Stairs with sunlight (photograph: Sofia Guzeva)**

10.5.2 Colors

Definitions (#10.5.487):

D1. We name "color" the perception of light excitation following neurophotochemical reaction process in the eye following of one or more light wave frequencies with one or more given amplitudes. This term is generally used to identify a frank or pure color, such as the primary and secondary colors of Itten, that is to say the yellow, the red (magenta), the blue, the orange, the purple and the green (therefore obviously "black" isn't a color as it correspond to the absence of radiation nor "white" as it sum of all colors of the visible spectrum!).

D2. We name "hue" the three primary colours (red, blue, and yellow) as well as the three secondary colours (orange, green, and purple). These colours are pure, meaning that white or black have not been mixed in with them.

D3. We name "tint" that is lighter than a pure hue, this is achieved by adding white. However, the term "tint" can also refer to a mixture of two colors. For example, turquoise is a shade resulting from the mixture of blue and green.

D4. We name "tone" a color to which white or black has been added (i.e. grey). Depending on how much white or black is added, a lighter or darker tone results. When you add white, you get a gradient tone, while when you add black, you get a broken tone.

D5. We name "shade" the result we get when just black is added to a pure hue. This makes the original colour darker. In a technical sense, black and white are not colors as we have already mention it, they're shades!

It is important never to confuse "color" that is a perceptive concept (it is not a physical property!!!) and "wavelength" that is a physical concept. Thus, the human eye is most often unable to distinguish a theoretical yellow monochromatic (single wavelength) of a corresponding composition of green and red. This illusion will display yellow on our computer screens, and more generally any color.

By the fact that the sensitive part of the retina of the human eye is made up of elements named "cones" each sensitive to a small bandwidth of wavelength corresponding respectively[12] to Red (via the erythrolabe molecule sensible to wavelength $\sim 700\,[\text{nm}]$ and denoted by the letter L) , Green (via the chlorolabe molecule sensible to wavelength $\sim 545\,[\text{nm}]$ and denoted by the letter M) and Blue (via the molecule cyanolabe sensible to wavelength $\sim 440\,[\text{nm}]$ and denoted by the letter S):

---

[12] According to some rare studies at this day (year 12019 according to holocene calendar), it seems that a few humans are able to see in some very special conditions (impulse LASER) low infra-red around $\sim 1000\,[\text{nm}]$

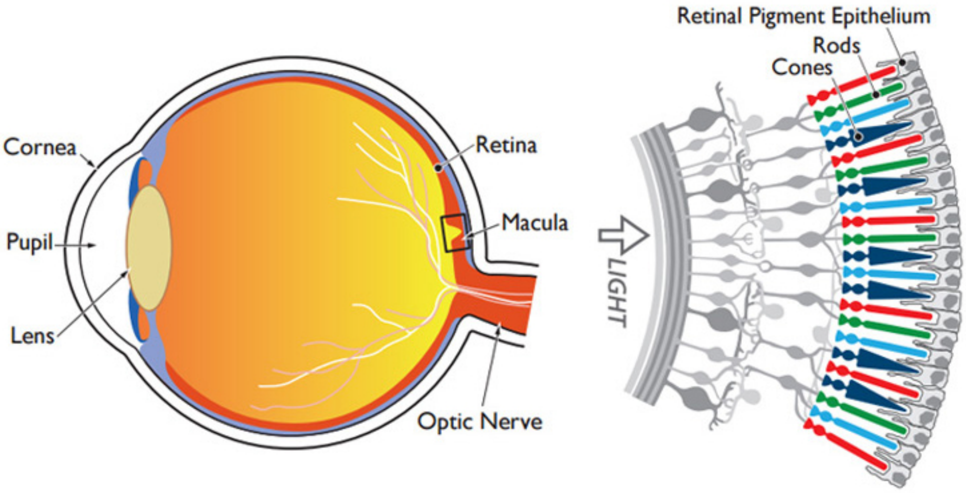


Figure 10.123: **Naive eye structure summary**

We can generate any color by adding these three basic colors named "additive primary colors" (or "additive primary colors"). This is named the "additive synthesis of colors".

Remark

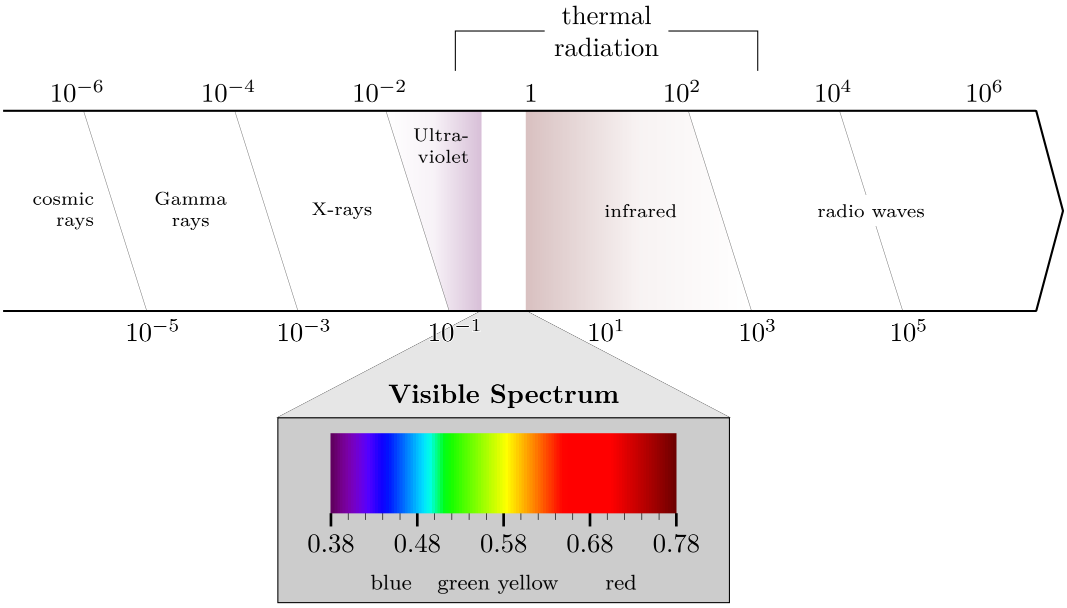
Some scientists think (without proof and without being able to reproduce this fact) that the sensitivity to each of theses colors comes from: for the Red for the time where evolution was in water and animals were looking in infra-red, for the Green at the time where animals were on Earth with a luxurious and dense vegetation and finally for the Blue because of the blue sky!

In what follows, we will denote the red by (R), the green by (G), the blue by (B), the white by (W) and the black by (N):

Color	Wavelength interval	Frequency interval
red	$\sim 625-740$ [nm]	$\sim 480-405$ [THz]
orange	$\sim 590-625$ [nm]	$\sim 510-480$ [THz]
yellow	$\sim 565-590$ [nm]	$\sim 530-510$ [THz]
green	$\sim 500-565$ [nm]	$\sim 600-530$ [THz]
cyan	$\sim 485-500$ [nm]	$\sim 620-600$ [THz]
blue	$\sim 440-485$ [nm]	$\sim 680-620$ [THz]
violet	$\sim 380-440$ [nm]	$\sim 790-680$ [THz]

Table 10.10: **Visible spectrum summary**

As suggested by a unique scientific paper on a very small sample of people grouped as females/males (see [4]) there is a relatively small (of the order of a few percent only!) but clear and statistically significant, difference between males and females in the hue sensations elicited by almost the entire color spectrum. Generally, males required a slightly longer wavelength (a shift of 3 to 10 nanometres) to experience the same hue as did females. Furthermore there is a difference in pupil diameter between individuals that can sometimes double between to individuals of the same age (a smaller diameter means that less light enters the eye, so scotopic vision may be prevalent) that's why when people go chase the northern lights (or southern lights) there are sometimes important differences in the perception of the phenomena (see [593]).


thermal radiation
10^{-6}
10^{-4}
10^{-2}
1
10^{2}
10^{4}
10^{6}
cosmic rays
Gamma rays
X-rays
Ultra-violet
infrared
radio waves
10^{-5}
10^{-3}
10^{-1}
10^{1}
10^{3}
10^{5}
Visible Spectrum
0.38
0.48
0.58
0.68
0.78
blue
green
yellow
red

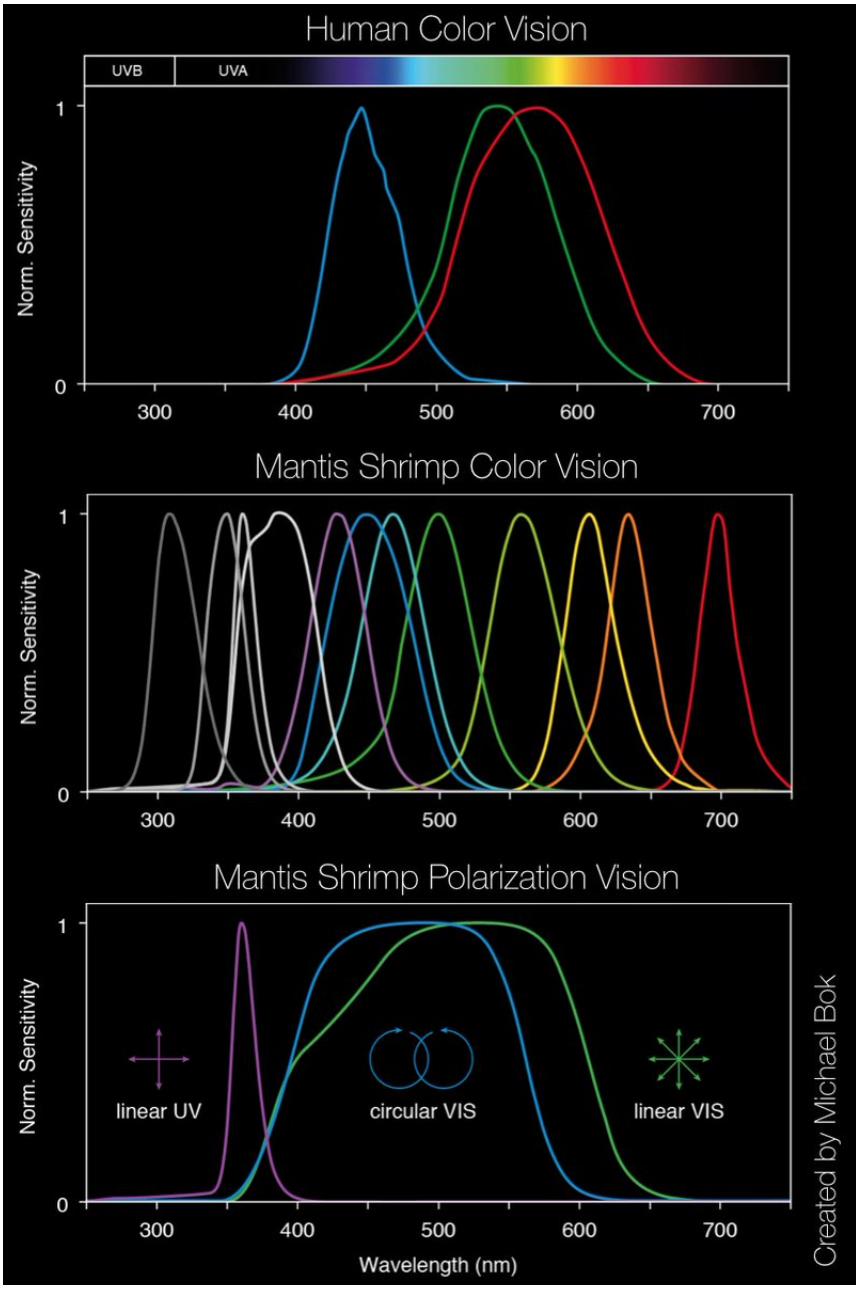
The entire rainbow of radiation observable to the human eye only makes up a tiny portion of the electromagnetic spectrum – about 0.0035%.

However, if you want to calculate the percentage represented by visible light at a given temperature of a black body, you can therefore use Planck's law (see section Statistical Mechanics page 3719):

$$p_{\text{vis.}}=\frac{2\pi\frac{h}{c^2}\int_{405\cdot 10^{12}}^{790\cdot 10^{12}}\frac{\nu^3}{e^{\frac{h\nu}{kT}}-1}\,\mathrm{d}\nu}{2\pi\frac{h}{c^2}\int_{0}^{+\infty}\frac{\nu^3}{e^{\frac{h\nu}{kT}}-1}\,\mathrm{d}\nu}=\frac{2\pi\frac{h}{c^2}\int_{405\cdot 10^{12}}^{790\cdot 10^{12}}\frac{\nu^3}{e^{\frac{h\nu}{kT}}-1}\,\mathrm{d}\nu}{\frac{2\pi^5k^4}{15c^2h^3}T^4}\cong\frac{2\pi\frac{h}{c^2}\int_{405\cdot 10^{12}}^{790\cdot 10^{12}}\frac{\nu^3}{e^{\frac{h\nu}{kT}}-1}\,\mathrm{d}\nu}{0.5188\cdot 10^{8}} \tag{10.5.2}$$

and for a temperature of $T=5,500$ [K] this ratio (computed with Maple™ 4.00b) is equal to 42.1% and for room temperature ($T=293.15$ [K]) it is equal to $0.752\cdot 10^{-22}\%$...

It's interesting to compare the human color vision with that of another animal like the mantis shrimp:



The French Association for Standardization (AFNOR) has defined in the 120th century (holocene calendar) the visual trivariance principle as follows: « *A radiation of any color can be produced visually identically by algebraic mixture, in a uniquely defined proportions, of luminous beam of three rays that can be arbitrarily selected, provided that none of them can be reproduced by a combination of the other two*».

Clearly when we see the frequencies of the visible spectrum it will not be tomorrow with the known material of this early 121st century (holocene calendar) that we are going to build antennas or parabola able to transmit at such frequencies! Already $120\;[\text{GHz}]$ is a challenge by the year 12010 (holocene calendar) then $500\;[\text{THz}]$ will not be for tomorrow...

The reader should know that until the years 11800 (holocene calendar) it was not known if the colors were limited or not those visible to the human eye. It was with the advent of thermometers with mercury that were sufficiently sensitive and specific that the astronomer Herschel placed one in front of a light spectrum and found that moving it from a color strip to the other, from violet to red, the temperature raised. To his surprise, it continued to rise when he accidentally let the thermometer to one or two centimetres beyond the red light visible bandwidth. Herschel detected invisible light to the human eye, described later as being infra-red radiation.

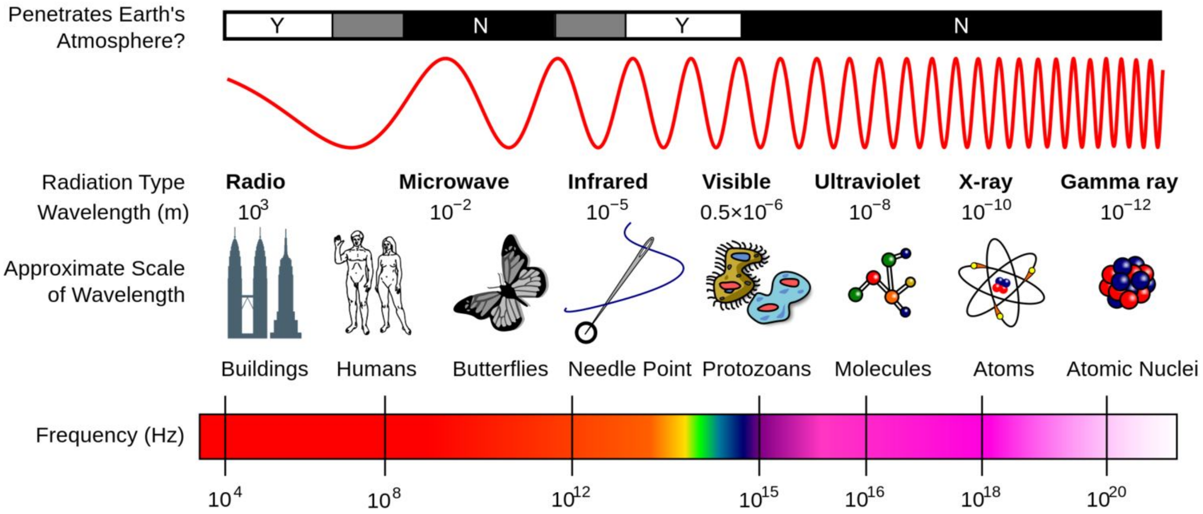

Penetrates Earth's Atmosphere?
Y
N
Y
N
Radiation Type
Wavelength (m)
Radio
10^3
Microwave
10^{-2}
Infrared
10^{-5}
Visible
0.5×10^{-6}
Ultraviolet
10^{-8}
X-ray
10^{-10}
Gamma ray
10^{-12}
Approximate Scale of Wavelength
Buildings
Humans
Butterflies
Needle Point
Protozoans
Molecules
Atoms
Atomic Nuclei
Frequency (Hz)
10^4
10^8
10^{12}
10^{15}
10^{16}
10^{18}
10^{20}

Figure 10.124: Whole standard electromagnetic spectrum (source: Wikipedia)

Or corresponding temperatures:

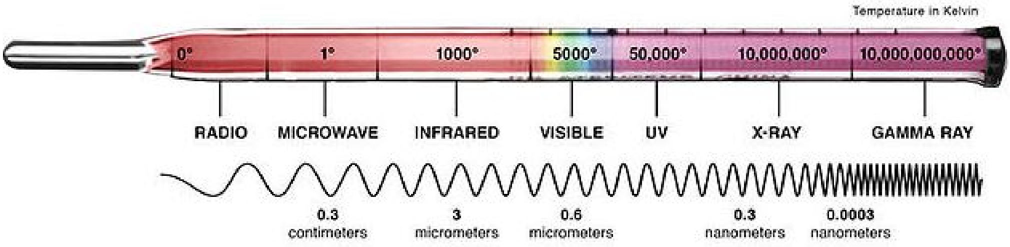

Temperature in Kelvin
0°
1°
1000°
5000°
50,000°
10,000,000°
10,000,000,000°
RADIO
MICROWAVE
INFRARED
VISIBLE
UV
X-RAY
GAMMA RAY
0.3 centimeters
3 micrometers
0.6 micrometers
0.3 nanometers
0.0003 nanometers

Figure 10.125: Temperature spectrum

Remark

Experimental measurements of the early 121st century (holocene calendar) provide us strong evidence that hydrogen bonds ionization ranges between $1000\,[\text{kJ}\cdot\text{mol}^{-1}]$ to $2000\,[\text{kJ}\cdot\text{mol}^{-1}]$ (i.e. 10 [eV] to 20 [keV]) and hydrogen bonds energy ranges between $6\,[\text{kJ}\cdot\text{mol}^{-1}]$ to $1000\,[\text{kJ}\cdot\text{mol}^{-1}]$ (i.e. 0.06 [eV] to 10 [keV]). Keep in mind that to ionize or break such bonds, photons must have at least the same energy (the energy of multiple photons can't be summed easily as it is quantum physics under the hood). For the 0.06 [eV] ionization bond we need then photons with a frequency of at least 1 [THz]...

A magistral and pedagogical example of what can be seen in the various spectra is the Crab Nebula that is a supernova remnant resulting from the explosion of a historical supernova (SN 1054) observed by several Chinese astronomers of the dynasty Song from July 1054 to April 1056:

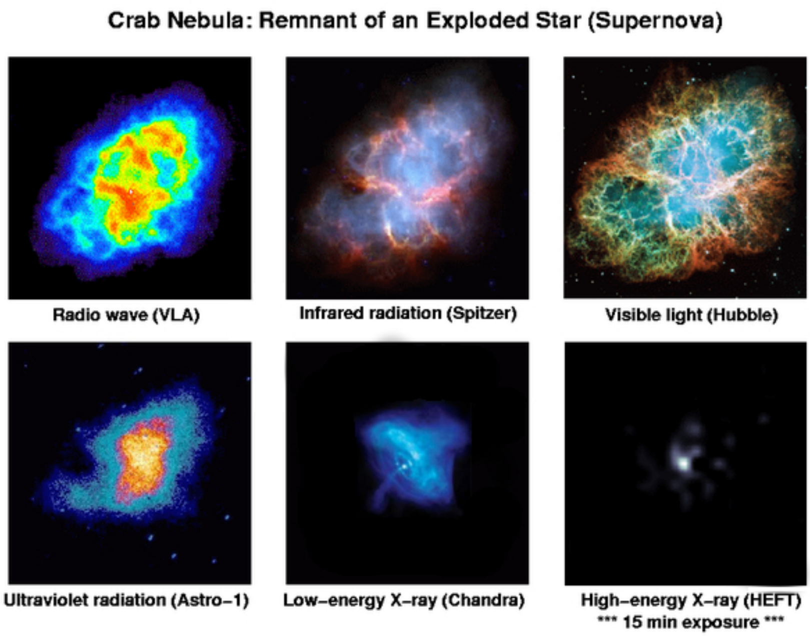

Crab Nebula: Remnant of an Exploded Star (Supernova)
Radio wave (VLA)
Infrared radiation (Spitzer)
Visible light (Hubble)
Ultraviolet radiation (Astro-1)
Low-energy X-ray (Chandra)
High-energy X-ray (HEFT)
*** 15 min exposure ***

Figure 10.126: **SN 1054 observed under different wavelengths and different telescopes**

Or also these different photos of the Earth under different wavelength:

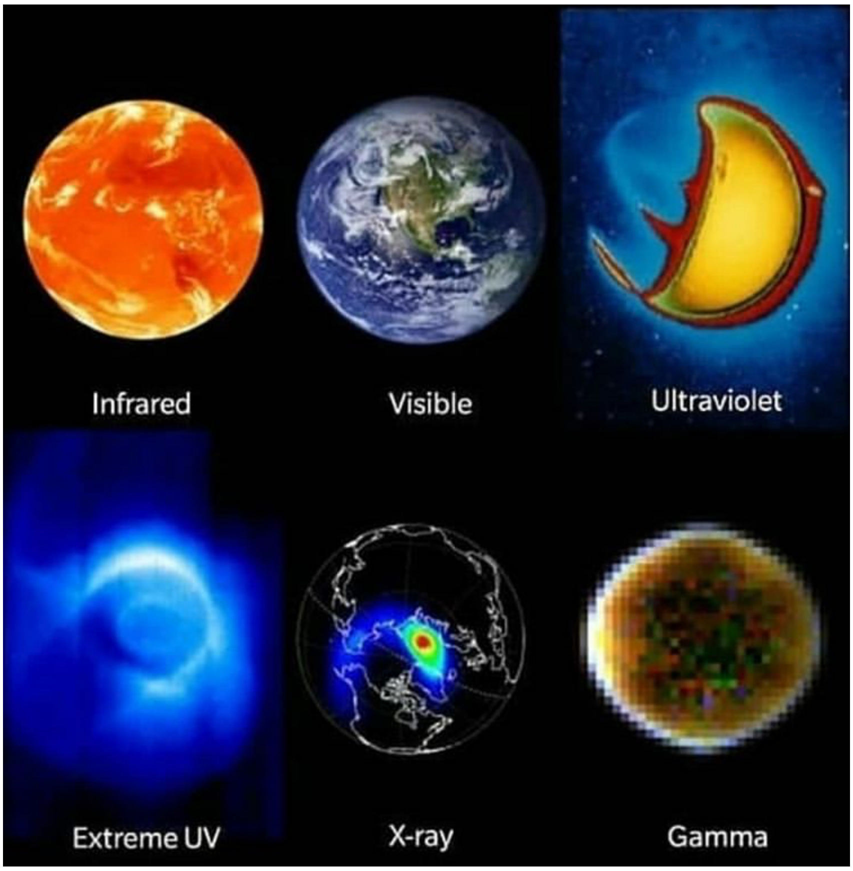

Infrared
Visible
Ultraviolet
Extreme UV
X-ray
Gamma

Figure 10.127: **Earth viewed under different wavelengths**

Also for health prevention reasons, let us inform that Ultraviolet[13] is commonly separated in three well known groups (UVC in space is a killer factor...):

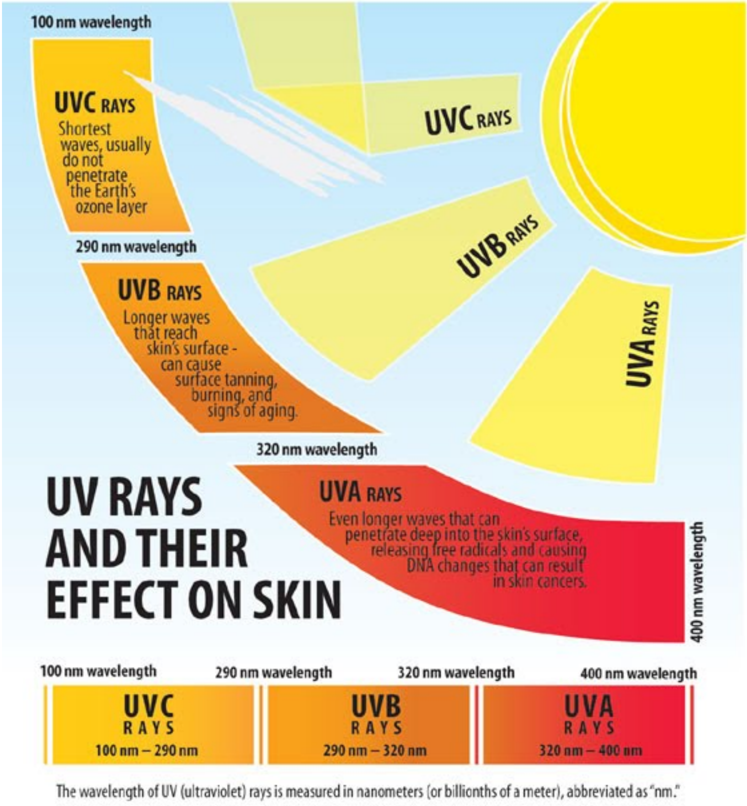

100 nm wavelength
UVC RAYS
Shortest waves, usually do not penetrate the Earth's ozone layer
290 nm wavelength
UVB RAYS
Longer waves that reach skin's surface - can cause surface tanning, burning, and signs of aging.
320 nm wavelength
UVA RAYS
Even longer waves that can penetrate deep into the skin's surface, releasing free radicals and causing DNA changes that can result in skin cancers.
400 nm wavelength
UVC RAYS
UVB RAYS
UVA RAYS
UV RAYS AND THEIR EFFECT ON SKIN
100 nm wavelength
290 nm wavelength
320 nm wavelength
400 nm wavelength
UVC RAYS 100 nm – 290 nm
UVB RAYS 290 nm – 320 nm
UVA RAYS 320 nm – 400 nm
The wavelength of UV (ultraviolet) rays is measured in nanometers (or billionths of a meter), abbreviated as "nm."

Figure 10.128: **Various types of Ultraviolet rays**

And also let us give the following figure that is quite instructive (representing the state of knowledge and Earth's atmosphere of the early 121st century):

---

[13] The penetration depth of ultraviolet radiation in various materials depends on factors like wavelength and material type. In polymers the UV penetration depth is about 0.025 [mm], for glass the penetration depth may approach 0.5 [mm], for skin the UV penetration is about 2 [mm].

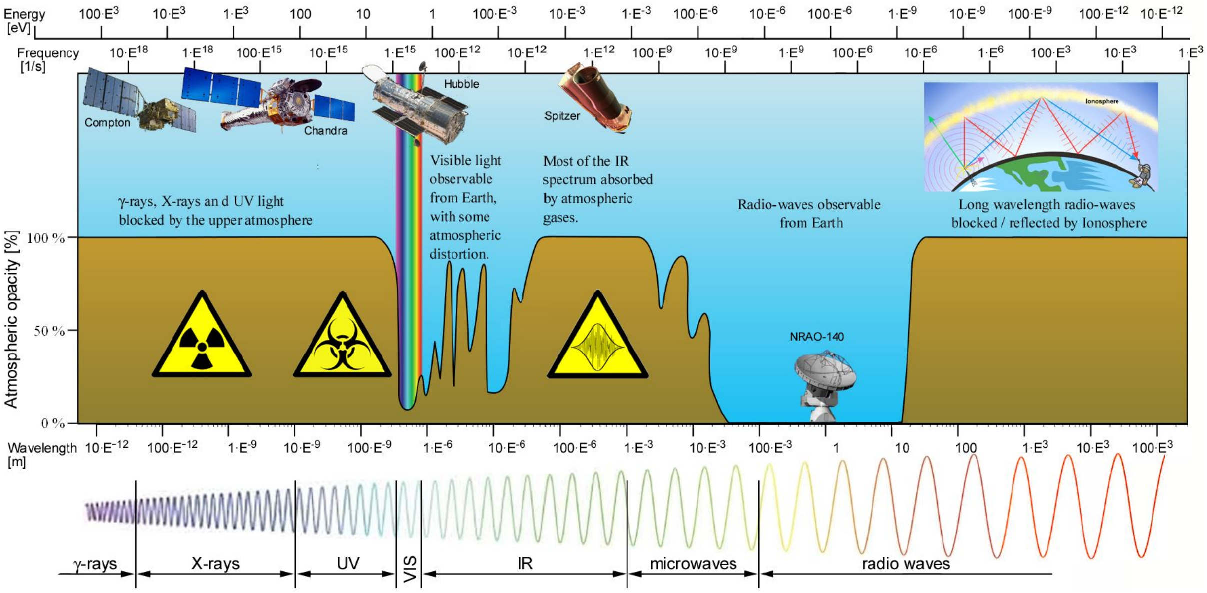


Figure 10.129: **Electromagnetic spectrum energy and atmosphere absorption (source: [177])**

Pointing three light beams (R, G and B) in the same place, we can get (in fact it would be more rigorous to say "perceived" because this is specific only to certain trichromats mammals): white light. We say then that white (W) (in the human sense) is the sum of the three additive primary colors (remember that white is rigorously the sum of all the colors of the spectrum - so that white is made of a continuous light spectrum). All imaginable colors are obtained by varying the intensity of each of the three beams. Black is obtained when we do send no light at all.

For example, if we add (in the theoretical sense: with infinitely small and transparent colors component ...) just red (R) and green (G), we get the yellow (Y), if we add the red and blue we get the magenta (M), if we add green (G) and blue (B), we get the Cyan (C). So we can sum this up by the following equations:

$$\begin{aligned} RG&=R+G=Y\\ RB&=R+B=M\\ GB&=G+B=C \end{aligned} \tag{10.5.3}$$

These three colors (YMC) obtained by adding two additive primary colors are named "secondary additive colors". The notion that blue and yellow form green is only a perception in our eyes!!! If the reader shine blue and yellow light on a chromograph it will show blue and yellow independently. In fact light doesn't interfere with light in our regular experience set at the opposite of sound. And this because as our eye only measure three things at the opposite of ears that measure every sound on the spectrum we can hear (as it do a indirectly a Fourier Transform).

Diagram of additive synthesis:

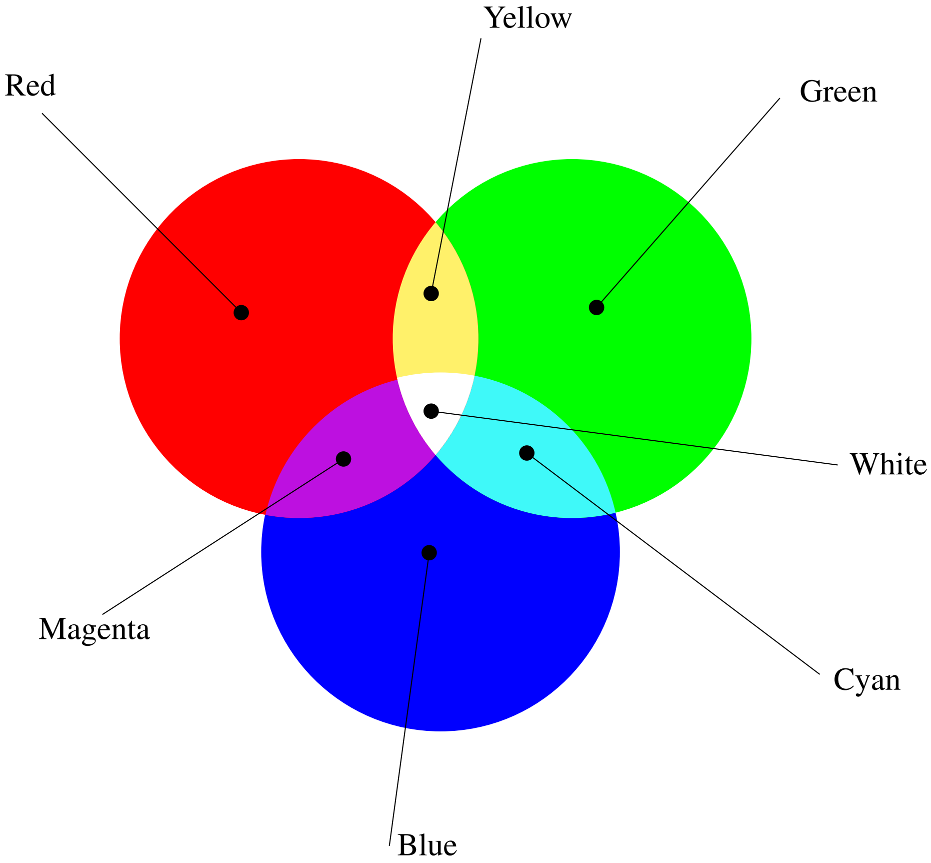


Figure 10.130: **Additive synthesis (RGB)**

The existence of these three types of pigments in our photoreceptor cones serves as physiological basis to the "trichromatic model" or of "visual trivariance".

Definition (#10.5.488): A color is named "complementary color" of another if they give white[14] when you add them. For example, yellow (Y) is the complementary color of blue (B):

$$Y+B=(R+G)+B=W \tag{10.5.4}$$

Remark

These chromatic concepts are quite useful or even important to understand as quantum chromodynamics use a special quantum number named the "color" to describe the state of quark. And the reasons that this quantum number is named "color" and that subject of study is named "quantum chromodynamics" is because it behaves similarly as the RGB system!!!

In contrast to additive synthesis, there are the "subtractive colors": it is that we are talking about when we take away the color to a base color. This is for example the case of the ink or of color filters (in the sense that there is a base support where color must be treated).

---

[14] Keep in mind that strictly speaking, white nor black are colors. White is the sum of all colors (electromagnetic waves radiation) from all the visible spectrum and black is the absence of colors (the absence of electromagnetic waves in the visible spectrum)

To understand what this is, let us put a red filter on a projector. The projected light will be red. We note that the filter has removed color to white light: W became R but as $W = RGB$, this means that the red color filter removed the GB to the white light of the overhead projector. With the same reasoning, we understand that a G filter subtracts the colors RB and B filter subtracts RG.

If we overlay two filters of different primary colors: for example, R and G filter filters, we will not get anything at all, that is, to say N. Indeed, the R filter let pass only red light and G filter subtracts this color and nothing remains. This leaves no more color, i.e. N.

We notice that filters R, G and B fail to synthesize different colors by subtraction as we get the black (N) as soon as we superimpose two different. What is very annoying when the concerned support is paper and the purpose is to print something colourful.

It is therefore more useful to use the yellow, magenta and cyan filters (Y, M and C) that is to say secondary additive colors. Indeed, a filter Y let pass yellow, that is to say RG. It therefore subtracts only the blue B to the original white W light. By the same principle, a magenta M filter subtracts green G and a filter cyan C subtracts red R.

We note therefore that the superposition of two filters of these secondary colors gives a new color to an existing support. We can synthesize any color by varying the intensity of each of the three filters Y, M and C that we overlay (on the projector or paper for example). We name these three colors "basic subtractive colors".

Scheme for subtractive synthesis:

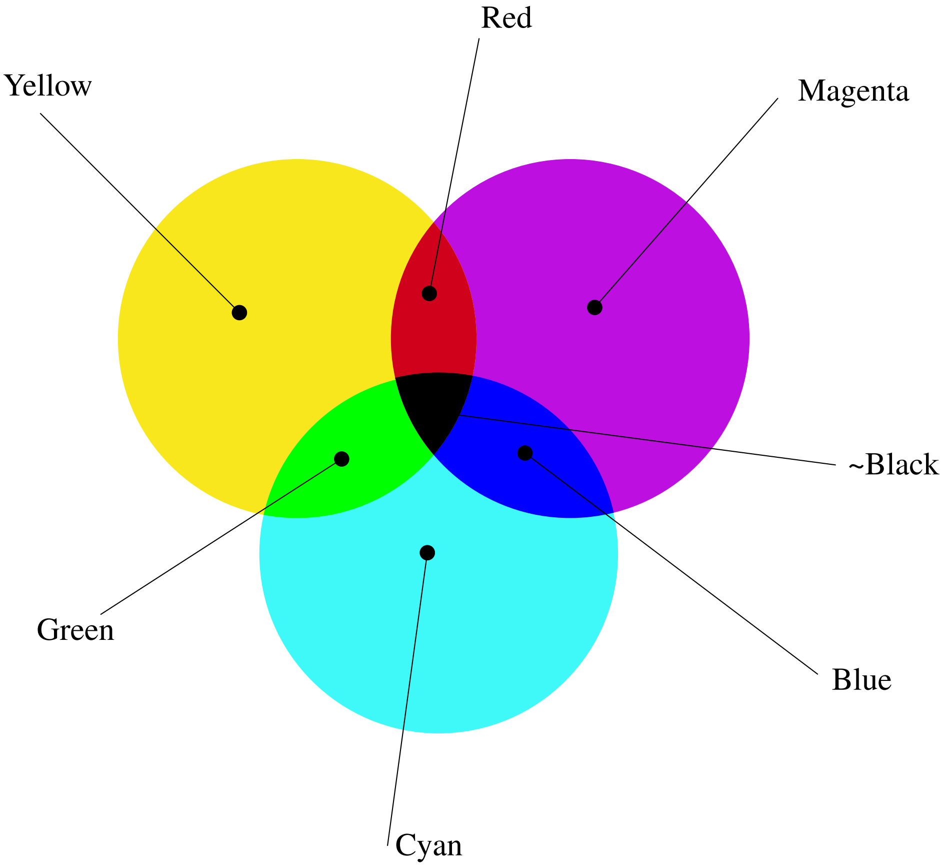


Figure 10.131: **Subtractive synthesis (CMYK)**

In reality we do not get a perfect black N this is why printers have a black color ink.

✍**Examples:**

E1. A television or computer screen works on the principle of additive color mixing. Indeed, looking at the screen through a magnifying glass, one can realize that it is filled with small groups (pixels) of three phosphors (bright area when excited by electricity) R, G and B. These phosphors are so close that when they light up together, they seem to get confused (at least on screen have a good "definition": pixels per inch) and most human therefore only sees the additive synthesis of the three phosphors. For example, on a television screen entirely red, only red phosphors glow. By cons, if the screen turns yellow, this means that green phosphors glow along with the red one.

E2. At the opposite to television or computer screens, we find the printing processes that operate in subtractive synthesis. Indeed, the sheet is white and you have to remove from it to get the color we are looking for. The technique is the same as that of filter: inks contain pigments which filter certain colors. Using Y, M and C inks we can get all colors of the visible spectrum. However, the pigments are not perfect and black N is very difficult to obtain (ink and color overload rather dark brown). So we use black as a fourth color. This system is named "quadrichromy printing". It is used for example in most color printers and in rotary newspapers.

It is interesting now to focus on the phenomena that overlap the two concepts (if we can say ...).

Thus, a system that projects color following a RGB concepts by additive or subtractive system itself may be illuminated by an equivalent system. This results in an effect overlay.

So when we talk about the color of objects, we normally refer to the appearance they have when they are illuminated with white light!

✍**Example:**

A red R tomato, absorbs some of the white light W (GB part) and distributes the rest (red R) back around. That is why it appears Red R to us when under white light. A yellow lemon, appears yellow to us because it absorbs the Blue B color distributes the rest around ($RG=R+G=Y$) .... But what about a tomato lit by a blue light? What looks like a yellow lemon if we illuminate it by red light?

We can answer by reasoning as follows: as tomato absorbs GB and therefore intrinsically Blue (B), there remains nothing. It then appears black N. For the lemon, as it absorbs the blue B and diffuses back $R+V=Y$ light so if we only illuminate it with red R it will only diffuse the red and therefore appear will appear Red R to us.

Because of how additive and subtractive light interference work, shadows aren't always black as illustrated below:

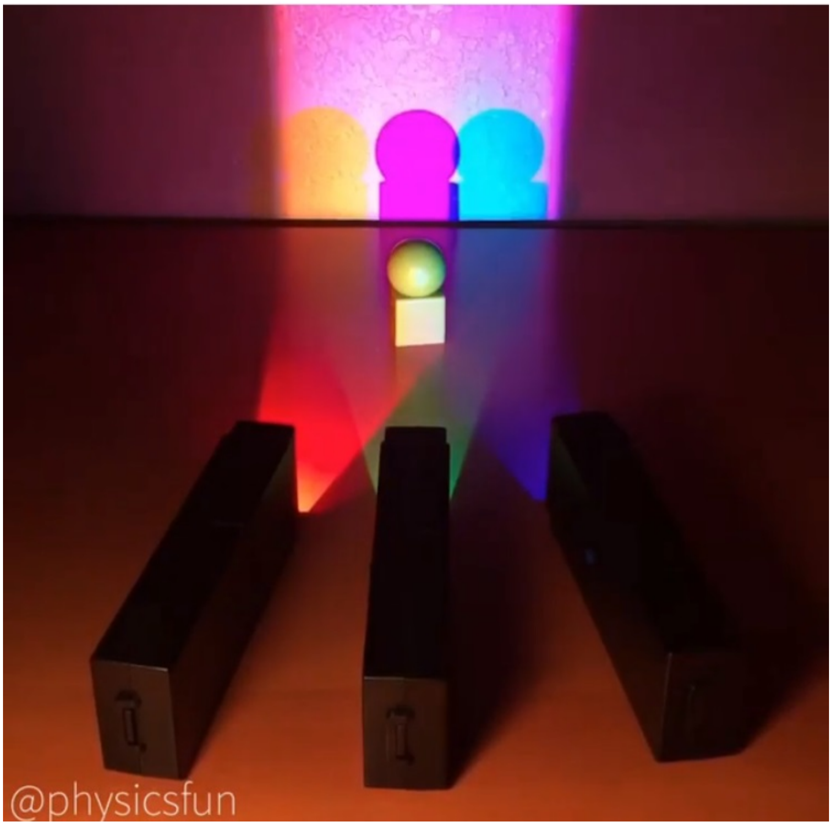


Figure 10.132: **Additive and Subtractive experiment (source: physicsfun instagram)**

10.5.3 Radiometry/Photometry

Material are capable of emitting, transmitting and / or absorb electromagnetic energy. Several factors characterize the radiation such as its spectral range, intensity, direction, and some intrinsic properties to the material. Photometry proposes to seek the magnitudes that are specific to the material and the laws governing them.

We recognize two types of photometry: "energetic photometry" and "visual photometry". In what follows, we will stick mainly to energetic photometry which generally relate to the energy carried by electromagnetic radiation, whatever its wavelength.

Beforehand, we have to specify the conditions under which we will define the new variables/quantities. We will assume the following assumptions:

H1. The radiation propagates in a transparent medium for all intensities, wavelength and polarization.

H2. The propagation takes place along the solid angles (see section Trigonometry page 2549). We therefore omit the radiation (approximation) in parallel rays.

H3. The elementary surfaces $\mathrm{d}S$ that are study are sufficiently small so that the radiation of their points are identical but not too small to avoid phenomena such as diffraction.

10.5.3.1 Energy flow

Definition (#10.5.489): The "energy flow" or "radiant flow" of a radiation source is the power it radiates. The flow equation is measured in Watts [W] (or joules per second $[\mathrm{J}\cdot\mathrm{s}^{-1}]$), and it follows therefore that for a source that radiates energy (not necessarily constant), we have:

$$\Phi=\frac{\mathrm{dE}_r(t)}{\mathrm{d}t} \tag{10.5.5}$$

In some professional fields the energy flow is expressed in photometric units as the "Lumen" denoted [lm] or in photonic units as a number of photons per second: $[\mathrm{s}^{-1}]$. This is why, when buying lamps or displays in some stores, the units are not the same from one brand to another... (ISO norms are not respected...!).

10.5.3.1.1 Beer–Lambert law

The "Beer–Lambert law", also known as "Beer's law", the "Lambert–Beer law", or the "Beer–Lambert–Bouguer law" relates the attenuation of light to the properties of the material through which the light is travelling. The law is commonly applied to chemical analysis measurements and used in understanding attenuation in physical optics, for photons, neutrons or rarefied gases.

If the absorption and diffusion of a medium can be considered proportional to the thickness $\mathrm{d}z$ of the crossing matters, the energy flow variation can be written:

$$\mathrm{d}\Phi=-\mu\Phi_0\mathrm{d}z \tag{10.5.6}$$

in this expression Φ_0 is the incident energy flow and $\mu\,[\mathrm{m}^{-1}]$ is the "linear attenuation coefficient", which is a function of the radiation frequency depending of the material the radiation pass through.

So we have a simple differential equation (see section Differential and Integral Calculus page 1609):

$$\boxed{\frac{\mathrm{d}\Phi}{\Phi_0}=-\mu\mathrm{d}z \Rightarrow \Phi(z)=\Phi_0 e^{-\mu z}} \tag{10.5.7}$$

which is the "Beer-Lambert law" (which can also be expressed from the light intensity that we will define further below).

We have typically the following order of amplitude:

$$\mu_{\text{yellow}}=10^{-5}\,[\mathrm{m}^{-1}] \quad \mu_{\text{purple}}=4\cdot 10^{-5}\,[\mathrm{m}^{-1}] \quad \mu_{\text{Glas (BKZ)},0.55[\mu\mathrm{m}]}=0.2\,[\mathrm{m}^{-1}]$$

Remark

The variation of the atmospheric absorption coefficient with the wavelength allows in particular to explain the blue color of the sky and because of water light reflection why our oceans looks like blue seen from space (yes don't forget that water don't have color... it's transparent!).

There are many other formulations of the Beer-Lambert law with one fairly used in nuclear physics (see section of the same name in this textbook page 5019) in the framework of radiation protection. Let's see now what it is:

Let us consider a flow Φ of particles striking perpendicularly the surface of a material of thickness $\mathrm{d}z$ and of atomic density ρ_N (in $[\text{atoms}\cdot\mathrm{m}^3]$). If we consider the particles striking a surface S, the latter can theoretically meet $\rho_N S\mathrm{d}z$ targets atoms in this layer. The number of interacting particles will be proportional to the intensity times this number, and we have:

$$\mathrm{d}\Phi S=-\sigma\Phi_0 NS\mathrm{d}z \tag{10.5.8}$$

where σ is a proportionality constant named "microscopic cross section." Its units are often expressed as "barn" ($1\,[\text{barn}]=10^{-26}\,[m]$).

Remark

The atomic density ρ_N is equal to:

$$\rho_N=\frac{\rho N_{\text{A}}}{M_{\text{Mol.}}}$$

where ρ is the density in $\mathrm{kg}\cdot\mathrm{m}^{-3}$, N_{A} is the Avogadro's number $6.022\cdot 10^{23}\,[\text{atoms}\cdot\text{mole}^{-1}]$ and $M_{\text{mol.}}$ the molar mass of the target expressed in $[\mathrm{kg}\cdot\text{mole}^{-1}]$.

If we now assume that the scattering centers are the electrons and not the target atoms, then we must replace ρ_N by ρ_{N_e} where $N_e=\rho_N Z$ with Z being the number of electrons interacting by target atom. Therefore:

If we now admit that the scattering centers are the electrons, not the target atoms, then we must replace N by N_e where $N_e = NZ$ where Z being the number of electrons interacting with the target atom. From where:

$$\Phi = \Phi_e^{-\sigma N_e z} = \Phi_0 e^{-\sigma Z \frac{\rho N_{\text{Av}}}{M_{\text{mol}}} z} \tag{10.5.9}$$

By identifying with the first formulation of the Beer-Lambert law, we see that μ plays the same role as:

$$\mu = \sigma Z \frac{\rho N_{\text{Av}}}{M_{\text{mol}}} \qquad \text{and} \qquad \sigma = \frac{\mu}{\rho} \frac{M_{\text{mol}}}{Z N_{\text{Av}}} \tag{10.5.10}$$

And the hypothesis that the electron is a "sphere of action" having a front surface πr_e^2, where r_e is the radius of this sphere, then:

$$\sigma = \pi r_e^2 = \frac{\mu}{\rho} \frac{M_{\text{mol}}}{Z N_{\text{Av}}} \tag{10.5.11}$$

and we have for the radius of the sphere of action of the electron:

$$r_e = \sqrt{\frac{\mu}{\rho} \frac{M_{\text{mol}}}{Z N_{\text{Av}} \phi}} \tag{10.5.12}$$

10.5.3.2 Light Intensity (Radiant Intensity)

To describe the energy flow Φ of a source, you must first measure it. The used sensor (thermocouple, bolometer, photocell, eye or others) may receive only one part: that which happens in the solid angle $\mathrm{d}\Sigma$ defined by its section.

Definition (#10.5.490): The "light intensity", "energy intensity" or "radiant intensity" denoted I of a punctual source is radiated flow Φ in the unit solid angle Ω centered around a direction Δ of emission:

$$I = \frac{\mathrm{d}\Phi}{\mathrm{d}\Omega} \tag{10.5.13}$$

The light intensity is expressed in certain professional fields in the photometric units (see summary table further below) "Candela" denoted [Cd] or in photonic units in $[\mathrm{W} \cdot \mathrm{s}^{-1}]$ (remember that the steradians don't have any units as well as for radians). This is why, when you buy screens or lamps in some stores, the units are not the same from one brand to another.

Remark

A source is named an "anisotropic source" or "directional source" if its intensity varies with the direction of observation.

By comparison (as it may help), a Candela unit is equivalent to the intensity of a source in a given direction, which emits a monochromatic radiation of frequency 540.1012 [Hz] (which roughly corresponds to the frequency at which the eye is the most sensitive), and whose luminous flow(or intensity) in that direction of $1/683$ [W] per steradian.

10.5.3.3 Energy Emittance (Radiant Emittance)

Definition (#10.5.491): The "emittance", "exitance" or "illumination" M of a source is the radiated energy flow (power) per unit surface $\mathrm{d}S$ in $[\mathrm{W}\cdot\mathrm{m}^{-2}]$ in all directions of outer space the source and depends on the physico-chemical properties of the emitting surface:

$$M=\frac{\mathrm{d}\Phi_{\text{ext}}}{\mathrm{d}S} \tag{10.5.14}$$

A common mistake is to make a confusion between the emittance and the energy!!!

Remark

"Radiant emittance" is an old term for this quantity. Radiant exitance is often named "intensity" in branches of physics other than radiometry, but in radiometry this usage leads to confusion with radiant intensity.

The emittance is often assimilated in the contemporary vocabulary to the "luminosity" of a light source which sometimes leads to confusion with the concept of light intensity.

The emittance is expressed by many professional of the field in photometric units named "Lux" denoted [lx] or in photonic units $[\mathrm{W}\cdot\mathrm{m}^{-2}]$ or even worst ... in $[\mathrm{lm}\cdot\mathrm{m}^{-2}]$ or even in $[\mathrm{cd}\cdot\mathrm{srm}^{-2}]$. For example when you buy a car, the headlights are indicated as being about ~ 20 [lx].

If the source is punctual (or considered as!) and its radiation isotropic , its direction has not to be taken into consideration. In the case of the related sphere of radius r, the emittance is then expressed as:

$$M=\frac{\Phi}{4\pi r^2} \tag{10.5.15}$$

In the case of the sphere, one element $\mathrm{d}S$ of the spherical surface receives perpendicularly the radiation. Very generally, an elementary surface can be inclined relative to the direction of radiation with an angle θ. So we have to project the surface perpendicular to the radiation using the basic reasoning of trigonometry (see section Trigonometry page 2503):

$$\boxed{\mathrm{d}\Phi=M\mathrm{d}S\cos(\theta)} \tag{10.5.16}$$

It is this projection that explains (among other factors!) the seasons on Earth: the area swept by the emittance more or less constant and isotropic of the Sun (considered as a punctual source) is maximum at the equator (perpendicular surface) and implies a radiation higher compared to that receives at higher or lower latitude for which the perpendicular projection of the surface in question is smaller than at the equator for the same emittance.

Remarks

R1. The energy emittance is calculated only in the outer half space before (the point from which we look at the source), because only half of the energy exchanged by the points of the surface dS is emitted as radiation. The other half is exchanged with the atoms located in the body.

R2. The emittance is usually also sometimes denoted F or even E. The practitioner will however take care not to confuse the emittance M with the magnitude (denoted by the same letter) that we define in the section of Astrophysics.

Now that we have some basics on radiant emittance let us deal with a quite recurrent question on social networks related to the fact that on most actual videos of the NASA we can't see the stars. This question is related to the calculation of mow many visible photons hit a pixel on a CCD sensor? With a few simplifying assumptions let us see that it not too difficult to calculate an approximation result to this question and get a simple answer to why we can't see distant stars on most videos made by CCD cameras.

Let us recall for this purpose that we have proved in the section of Thermodynamics (see page 3719) that the emittance of the black-body was given by (remember that $\nu=c/\lambda$):

$$\begin{aligned}M(T)&=\frac{c}{4}\int\limits_0^{+\infty}R(\nu,T)\mathrm{d}\nu=\frac{c}{4}\int\limits_0^{+\infty}\frac{h\nu^3}{c^3}\frac{1}{e^{\frac{h\nu}{kT}}-1}\mathrm{d}\nu=2\pi\frac{h}{c^2}\int\limits_0^{+\infty}\frac{\nu^3}{e^{\frac{h\nu}{kT}}-1}\mathrm{d}\nu\\&=2\pi\frac{h}{c^2}\int\limits_0^{+\infty}\frac{\left(\frac{c}{\lambda}\right)^3}{e^{\frac{hc}{kT\lambda}}-1}\left(-\frac{c}{\lambda^2}\mathrm{d}\lambda\right)=2\pi hc^2\int\limits_0^{+\infty}\frac{1}{\lambda^5\left(e^{\frac{hc}{kT\lambda}}-1\right)}\mathrm{d}\lambda\end{aligned}\tag{10.5.17}$$

The power of spectral radiant exitance contained in light within a specified wavelength range is then obviously obtained by integrating:

$$M(\lambda,T)=\frac{2\pi hc^2}{\lambda^5\left(e^{\frac{hc}{kT\lambda}}-1\right)}\tag{10.5.18}$$

over a given range. The result is named "irradiance" E_e and given by:

$$E_e=\int\limits_{\lambda_1}^{\lambda_2}M(\lambda,T)\mathrm{d}\lambda\;[\mathrm{W}\cdot\mathrm{m}^{-2}]\tag{10.5.19}$$

this can be expressed in terms of photons as follows using the Planck-Einstein relation ($E=hc/\lambda$);

$$\boxed{N_{\text{ph}}=\int\limits_{\lambda_1}^{\lambda_2}\frac{M(\lambda,T)\lambda\mathrm{d}\lambda}{hc}=2\pi c\int\limits_{\lambda_1}^{\lambda_2}\frac{1}{\lambda^4\left(e^{\frac{hc}{kT\lambda}}-1\right)}\mathrm{d}\lambda\;[\text{photons}\cdot\mathrm{s}^{-1}\cdot\mathrm{m}^{-2}]}\tag{10.5.20}$$

Such calculation is easily done numerically. For instance the daylight Sun (actually a G2V-type star, aka yellow dwarf) is considered to be a black-body radiator of around temperature $5,300$ [K]. In units better suited to a photographic sensor context it therefore emits about $4.549\cdot10^{13}$ [photons $\cdot\,\mathrm{s}^{-1}\cdot\mu\mathrm{m}^{-2}$] in the standard visible range (380-780) in all directions at its surface (multiplied by the surface of the Sun it gives a number of approximately $1.5\cdot10^{41}$ [photons $\cdot\,\mathrm{s}^{-1}$]).

As we have proved in the section of Electromagnetism, the intensity of the electromagnetic field decrease in the inverse square of the distance ($4\pi R^{-2}$). Then in reality we receive form the Sun:

$$N_{\text{ph}}\cong\frac{1.5\cdot10^{41}}{4\pi R^2}\cong\frac{1.5\cdot10^{41}}{4\pi(149.6\cdot10^9)^2}\cong1.33\cdot10^8\;[\text{photons}\cdot\mathrm{s}^{-1}\cdot\mathrm{m}^{-2}]\tag{10.5.21}$$

Or in units more suited to CCD (in 12019 according to the holocene calendar the typical civil CCD pitch size is a square of $1.51\ [\mu\text{m}]$):

$$N_{\text{ph}} \cong 1.33 \cdot 10^{-4}\ [\text{photons} \cdot \text{s}^{-1} \cdot \mu\text{m}^{-2}] \tag{10.5.22}$$

Keeping in mind that most CCD HD (high definition) civil cameras in the beginning of the 120th century (holocene calendar) have a capture rate of 20 to 30 frames per second! So the reader can easily understand now why cameras having a too fast frame rate, cannot captures stars (that you are in space or on Earth...).

Remark

Notice that for a distant star (the nearest one being at Proxima Centauri at almost $4.37 \cdot 10^{9}$ [km]), the same calculation gives:

$$N_{\text{ph}} \cong 4.97 \cdot 10^{-7}\ [\text{photons} \cdot \text{s}^{-1} \cdot \mu\text{m}^{-2}] \tag{10.5.23}$$

All of these photons do not necessarily make it through the atmosphere which absorbs varying amounts of light energy depending on wavelength. On the other hand a household incandescent bulb at around $3,000$ [K] emits about $1.267 \cdot 10^{12}\ [\text{photons} \cdot \text{s}^{-1} \cdot \mu\text{m}^{-2}]$ most of which make it to our sensors.

To determine what exposure the equivalent quantities above correspond to we simply convert the given irradiance to photometric units by taking into consideration its visible wavelengths only.

The photometric version of irradiance E_e is named "illuminance" and defined by:

$$E_\nu = K_m \int\limits_{380}^{780} M(\lambda, T)V(\lambda)\mathrm{d}\lambda\ [\text{lumens} \cdot \text{m}^{-2}] \tag{10.5.24}$$

where $V(\lambda)$ is the photonic eye response function and K_m the luminous efficiency conversion constant $683\ [\text{lm} \cdot \text{W}^{-1}]$. Remember that $[\text{lumens} \cdot \text{m}^{-2}]$ are named lux [lx]...

Now that we know how to determine how many photons impinge on a sensor we can estimate its Effective Quantum Efficiency (EQE), that is the efficiency with which it turns such a photon flux into photoelectrons (e^-), which will then be converted to raw data to be stored in the captured raw file:

$$\text{EQE} = \frac{N_{e^-}\ \text{produced by average pixel}}{N_{\text{ph}}\ \text{incident on average pixel}} \tag{10.5.25}$$

It represents the probability that a photon arriving on the sensing plane from the scene will be converted to a photo-electron by a typical digital camera sensor. It therefore includes the effect of microlenses, fill factor, CFA and other filters on top of silicon in the pixel. It is usually expressed as a percentage. For instance if an average of 100 photons per pixel within the sensor's passband were incident on a uniformly lit spot of the sensor and on average each pixel produced a signal of 20 photoelectrons we would say that the Effective Quantum Efficiency of the sensor is 20%.

Remark

It seems that the eye can have a some wavelength an EQE of 100% same as some expensive CCD...

10.5.3.4 Radiance and Luminance

Given a non-punctual source which emittance M is known at any point. An element $\mathrm{d}S$ of a surface of this kind of source will by definition not necessarily of isotropic intensity and therefore brighter (more powerful) when we observe it col-linearly to the vector $\mathrm{d}\vec{S}$.

The intensity I that radiates in a direction forming an angle θ with the normal to the emission surface S is always less than that radiated in the direction of the vector $\mathrm{d}\vec{S}$. So by simple application of trigonometry, we get the definition of "luminance" (or "radiance"):

$$L(\theta)=\frac{\mathrm{d}I}{\mathrm{d}S\cos(\theta)}=\frac{\mathrm{d}^2\Phi}{\mathrm{d}S\mathrm{d}\Omega\cos(\theta)} \tag{10.5.26}$$

more often written (to avoid a unit confusion because of the presence of the square on the differential element):

$$L(\theta)=\frac{\mathrm{d}\Phi}{\mathrm{d}S\mathrm{d}\Omega\cos(\theta)} \tag{10.5.27}$$

expressed in certain areas, in "Nits"... photometric units $[\mathrm{Cd}\cdot\mathrm{m}^{-2}]$ or photonic units $[\mathrm{W}\cdot\mathrm{m}^{-2}]$ (without specifying explicitly the steradian).

Radiance is useful because it indicates how much of the power emitted, reflected, transmitted or received by a surface will be received by an optical system looking at that surface from some angle of view. In this case, the solid angle of interest is the solid angle subtended by the optical system's entrance pupil. Since the eye is an optical system, radiance and its cousin luminance are good indicators of how bright an object will appear. For this reason, radiance and luminance are both sometimes named "brightness", even if this usage is now discouraged.

Remark

When we are concerned only with the human visible part of light, the luminance and radiance of a source is named by practitioners "brightness" (note this is not the case when we deal of shining as we will see it in the section of Astrophysics).

Luminance is a measure of how much power (energy over time) is emitted in a particular direction (basically how much light your TV will produce), Radiance broadly speaking, measures how much of the luminance you'll actually see. Finally brightness doesn't really measure anything it's broadly a catch-all term describing both luminance and radiance.

We can obviously write by making some elementary algebra that:

$$I(\theta)=\int\limits_{\Sigma} L(\theta)\cos(\theta)\mathrm{d}S \tag{10.5.28}$$

that gives us the energy intensity that radiates a source of luminance L in a given direction.

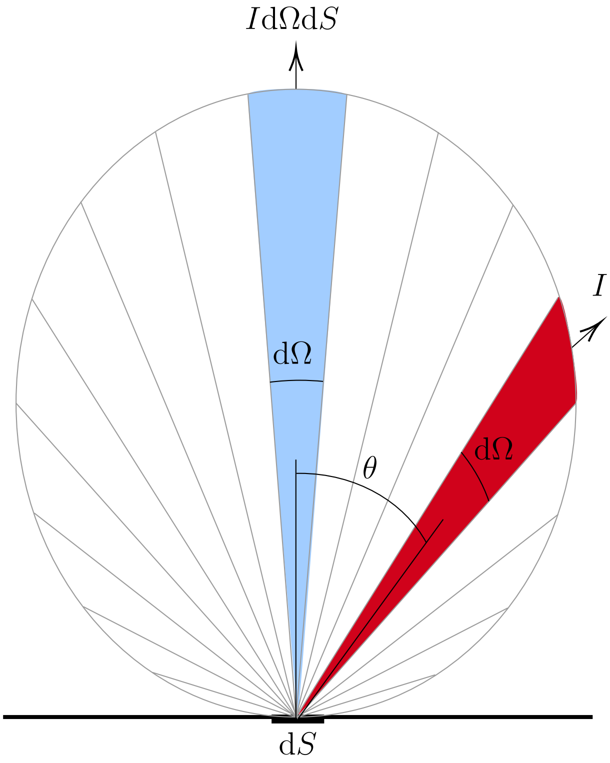
Johann Heinrich Lambert (11728-11777 according to holocene calendar) has observed that the energy intensity of some anisotropic sources (among all imaginable types of sources...) decreases as the cosine of the angle θ, around the direction perpendicular to the surface source, such that:

$$\boxed{I(\theta)=I_0\cos(\theta)} \tag{10.5.29}$$

and that we name "Lambert's cosine law". A source that respect this law is named "Lambert's source" (the Sun is considered as a near-perfect black-body and Lambert's source in broad stellar modelling).

Caution! There is often a confusion from students that make them think that there is a contradiction because the black-body emitter (that is "Lambertian") has a radiation that is independent of direction. A Lambertiam emitter appears equally bright from all viewing angles too! That is, the power per unit perceived area stays the same! It is both the power received and the solid angle it arrives from decrease as $\cos(\theta)$ leaving their ratio unchanged!

So the radiant intensity or luminous intensity observed from an ideal diffusely reflecting surface or ideal diffuse radiator is directly proportional to the cosine of the angle θ between the direction of the incident light and the surface normal:



$I\mathrm{d}\Omega\mathrm{d}S$
$I\cos(\theta)\mathrm{d}\Omega\mathrm{d}S$
$\mathrm{d}\Omega$
θ
$\mathrm{d}\Omega$
$\mathrm{d}S$

Figure 10.133: **Lambertian emittor**

A surface which obeys Lambert's law is said to be "Lambertian surface". The apparent brightness of a Lambertian surface to an observer is the same regardless of the observer's angle of view. More technically, the surface's luminance is isotropic, and the luminous intensity obeys Lambert's cosine law. Unfinished wood exhibits roughly Lambertian reflectance, but wood finished with a glossy coat of polyurethane does not, since the glossy coating generates specular highlights. Not all rough surfaces are Lambertian reflectors, but this is often a good approximation when the characteristics of the surface are unknown.

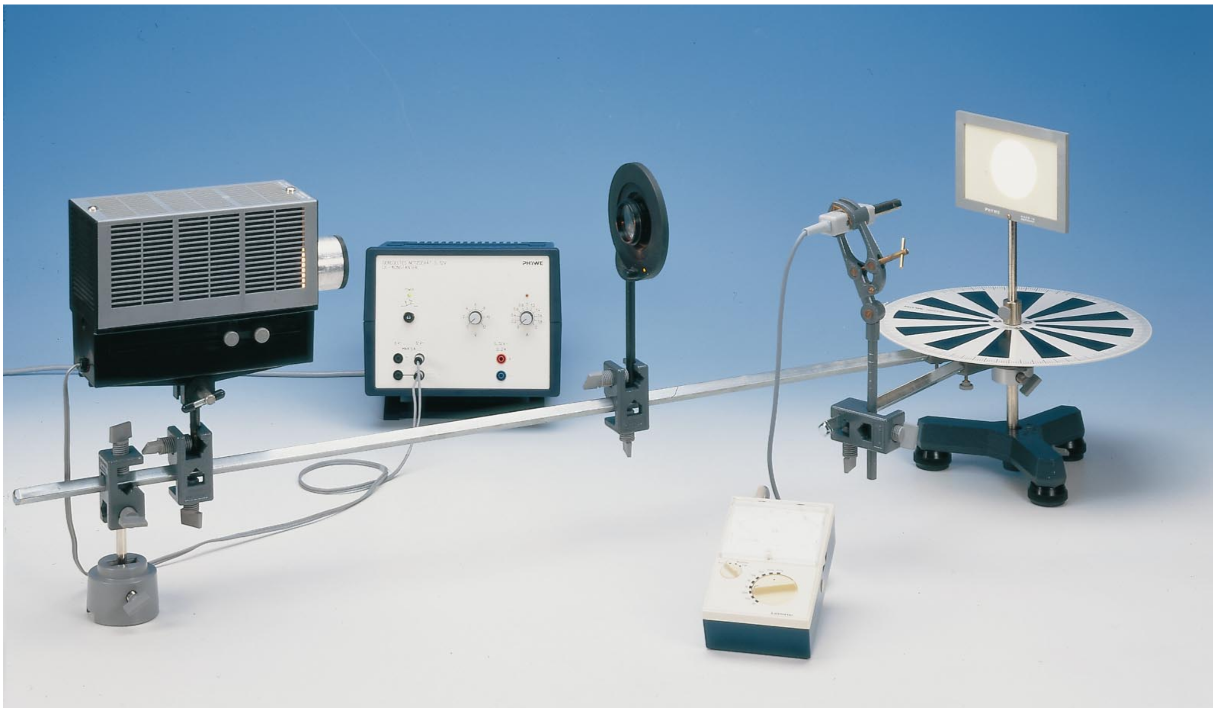
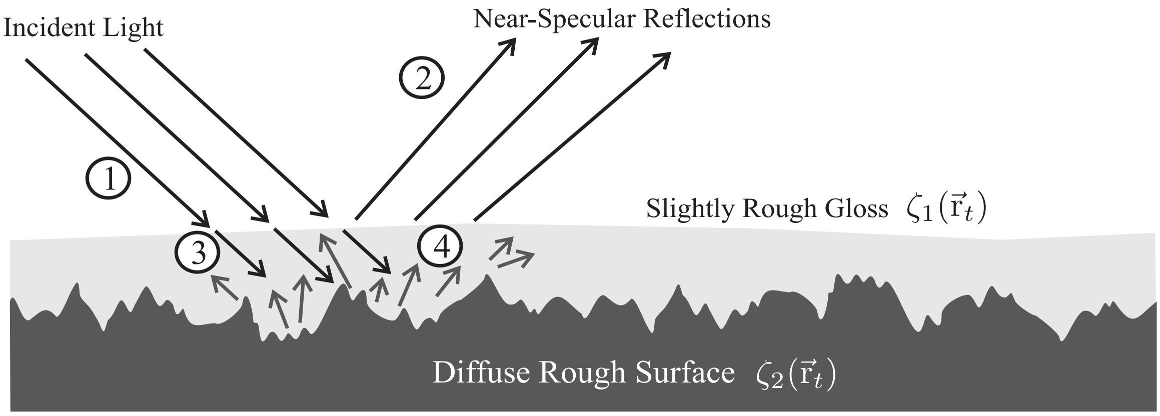


Figure 10.134: **Typical experimental setup used in universities to check Lambert's cosine law**

Remark

We speak of "luminance" of a source of "illumination" of an object (by a source).

Consider the physical diagram of the figure, which illustrates the composition of a typical object surface at the microscopic level. Many objects possess a glossy veneer that undulates very slightly across the surfaces. Incident light - whether from a coherent or incoherent source - would partially reflect and partially transmit through this first interface. The transmitted portion of the light then travels through the dielectric surface to the much rougher material interface below. This surface, usually regarded as Lambertian, then scatters the light in nearly every direction (as well as coloring the light with frequency-selective absorption). The diffusely scattered light then partially transmits back into the scene from the surface. Thus, there are two potential components to the light emanating from a surface at optical frequencies: a near-specular component and a diffuse, Lambertian component.


Incident Light
Near-Specular Reflections
①
②
③
④
Slightly Rough Gloss $\zeta_1(\vec{r}_t)$
Diffuse Rough Surface $\zeta_2(\vec{r}_t)$

In the figure above whe have 1) incident light partially reflects off a 2) smooth, slightly undulating surface. Then 3) transmitted light enters the second medium and strikes the Lambertian surface which re-radiates diffuse power away from the surface in all directions.

The basic physical model of the figure above helps to explain why objects look the way they do to our eyes. First of all, the diffuse scattering of light allows us to see surfaces even when angle-of-observation is not equal to angle-of-incidence to a light source. However, there can be a strong near-specular component present, which provides a glossy sheen and acts as an imperfect mirror for any image that falls upon the surface. Capturing this behavior is the key part of realistic computer graphic rendering for digital art, web illustrations, and video games!

In light of our knowledge of rough surface scattering, we will parse one of the most famous empirical models in computer graphics - the heurestic "Phong reflection model" (see [442]). This reflection model estimates the light intensity contribution I (in Watts) at observation point $(\vec{v})$ due to a surface point $\vec{r}_s$ in a scene may be computed with the following summation over the light sources present:

$$I(\vec{r})=\sum_{m=1}^{\substack{\#\text{ of}\\ \text{sources}}}\left(\underbrace{\kappa_a}_{\text{ambient}}+\underbrace{\kappa_d\cos(\theta_{nm})}_{\text{diffuse}}+\underbrace{\kappa_s\cos^{\alpha}(\theta_{rm})}_{\text{specular}}\right)I_m \qquad (10.5.30)$$

where:

$$\cos(\theta_{rm})=\hat{k}_{rm}\circ\frac{(\vec{r}-\vec{r}_s)}{\|\vec{r}-\vec{r}_s\|}\qquad\cos(\theta_{nm})=\hat{n}\circ\frac{(\vec{r}_m-\vec{r}_s)}{\|\vec{r}_m-\vec{r}_s\|} \qquad (10.5.31)$$

with (see the figure below for a geometrical representation):

- I_m is the intensity of the m th light source
- κ_a is the ambient coefficient due to multiple diffusely scattered waves
- κ_d is the diffuse scattering coefficient due to Lambertian roughness
- κ_s is the specular reflection coefficient from glossy layer
- α is the shininess factor (0 is pure Lambertian, ∞ is a perfect mirror)

- $\vec{r}$ is the point of observation
- $\vec{r}_s$ is the surface point under observation
- $\vec{r}_m$ is the location vector of the m th source
- $\hat{n}$ is the surface normal vector
- $\hat{k}_{rm}$ is the direction of specular reflection due to m th source

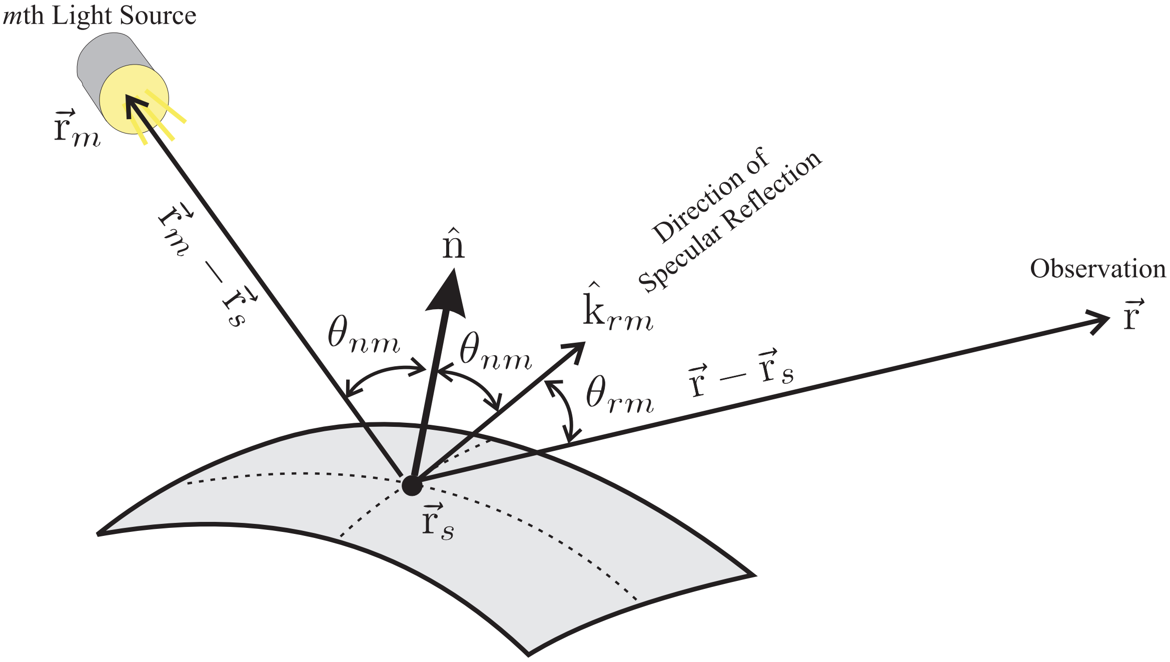


Figure 10.135: **Geometry used in the Phong reflection model**

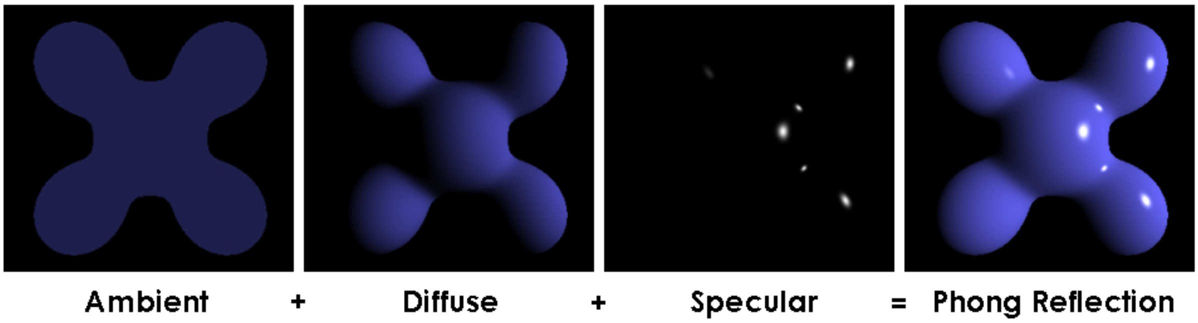
The model described by the heurestic equation above (Phong equation) breaks down the optical contribution of a point on a surface into specular, diffuse, and ambient components:

- *Ambient Term*: This term is physically linked to the "ambient" light in the environment which is mostly due to light that has already diffusely scattered from a source. This component provides a dull, soft illumination of a surface point from nearly every direction and to every direction. Though we would expect the reflection of ambient illumination to vary across a realistic scene, the omnipresence of diffuse scattering at optical frequencies allows us to approximate ambient illumination as uniform. Captured by a single coefficient κ_a, this model of ambient illumination - the most empirical part of the Phong reflection model - provides enough realism for most graphics applications. In the Phong equation, the ambient contribution will scale with the total intensity of the sources in the scene.
- *Specular Term*: This component imparts a rough, mirror-like sheen to points on the surface of an object. In addition to the material-dependent coefficient κ_s, which measures overall intensity of the specular reflection, there is a shininess term α. When $\alpha = 0$, the specular reflection takes on a pseudo-Lambertian pattern where light reflects equally in all

directions. When α becomes very large, this reflection contribution only occurs in specular directions. In our rough surface analysis, we would describe the glossy layer surface displacement function $\zeta_1\left(\vec{r}_t\right)$ as having increasing wavenumber content as α decreases.

- *Diffuse Term*: This component is due to the underlying, highly rough reflective layer under any glossy veneer. Its surface displacement function, $\zeta_2\left(\vec{r}_t\right)$ would have near-uniform wavenumber content over the interval $\|\vec{k}\|<4\pi/\lambda$. The scattering in this case is near-Lambertian and radiates as a function of observation angle with respect to the surface normal.

Visual illustration of the Phong equation: here the light is white, the ambient and diffuse colors are both blue, and the specular color is white, reflecting a small part of the light hitting the surface, but only in very narrow highlights. The intensity of the diffuse component varies with the direction of the surface, and the ambient component is uniform (independent of direction):



These terms are often applied for each color of light in an illumination spectrum. Although the heuresticc Phong reflection model is highly empirical, it is easy to see how it approximately reproduces the physical behavior in the electromagnetic interactions of a realistic rough surfaces. Because the eye is very forgiving, the approximate behavior is sufficient for many types of realistic scene renderings.

For more advanced models see [392], or [417] or [142].

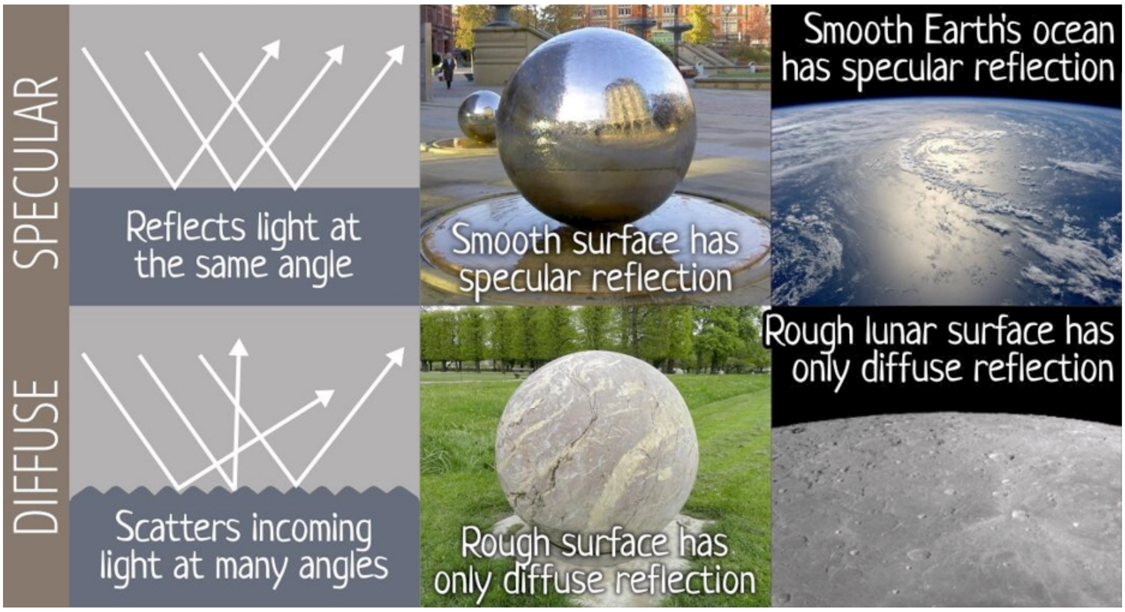
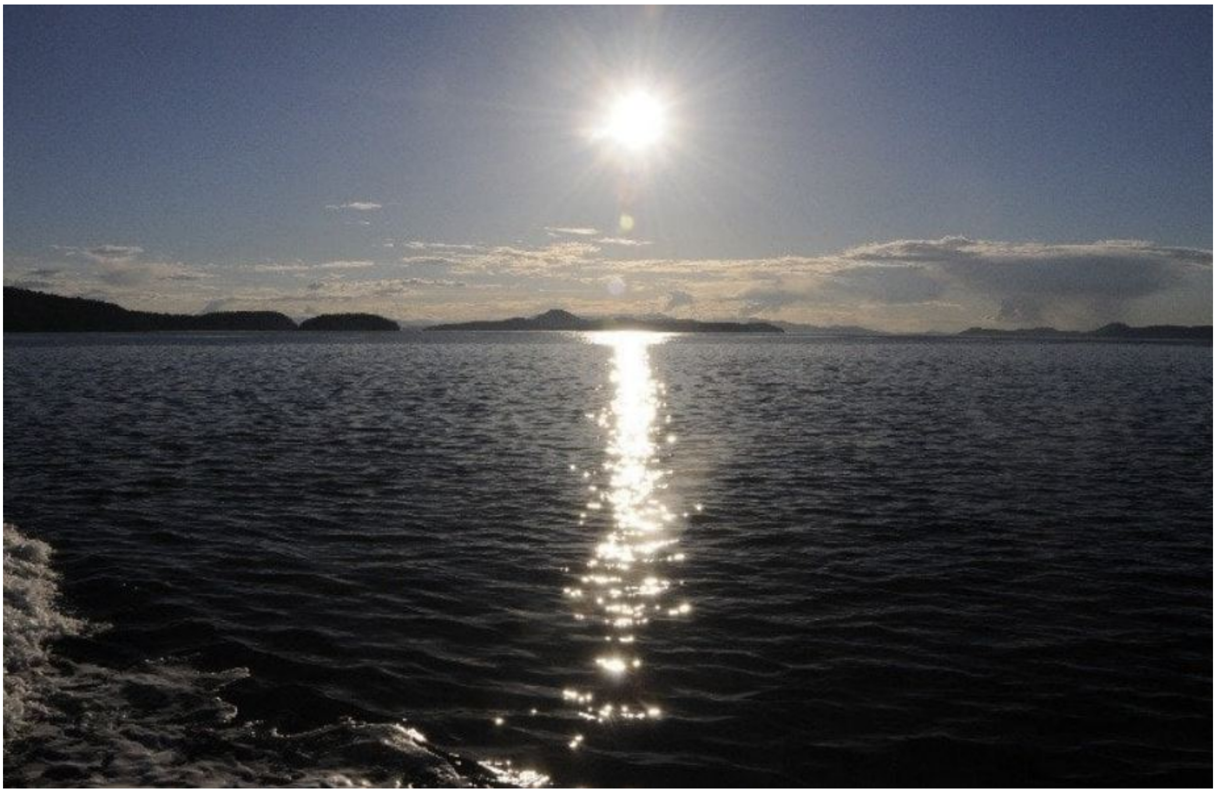

SPECULAR
Reflects light at the same angle
Smooth surface has specular reflection
Smooth Earth's ocean has specular reflection
DIFFUSE
Scatters incoming light at many angles
Rough surface has only diffuse reflection
Rough lunar surface has only diffuse reflection

Figure 10.136: **Specular vs Diffuse reflection**

And the Phong reflection model reproduces among everything we see in real life the well known Sun glitter pattern (collection of specular reflections of the Sun - or "glints" - on the surface of water) on the surface of a liquid that has a rough surface (otherwise on a calm surface of liquid this pattern doesn't appear and we just have a mirror image of the Sun's disc):



10.5.3.4.1 Lambert's law

A light source follows the Lambert's cosine law if its luminance (or radiance) is the same in all direction. Indeed, if we put first:

$$I=\frac{\mathrm{d}\Phi}{\mathrm{d}\Omega}=I_0 \tag{10.5.32}$$

we have:

$$L(\theta)=\frac{I}{S\cos(\theta)}=\frac{I_0\cos(\theta)}{S\cos(\theta)}=\frac{I_0}{S} \tag{10.5.33}$$

and we have:

$$\Phi=\int\limits_{2\pi} I\mathrm{d}\Omega=\int\limits_{2\pi} I_0\cos\theta)\mathrm{d}\Omega \tag{10.5.34}$$

But we have proved in the section of trigonometry that:

$$\mathrm{d}\Omega=2\pi\sin(\theta)\mathrm{d}\theta \tag{10.5.35}$$

Which brings us to write:

$$\begin{aligned}\Phi&=\int\limits_{2\pi} I_0\cos(\theta)\mathrm{d}\Omega=\int\limits_{2\pi} I_0\cos(\theta)2\pi\sin(\theta)\mathrm{d}\theta=I_02\pi\int\limits_0^{2\pi}\cos(\theta)\sin(\theta)\mathrm{d}\theta=I_02\pi\int\limits_0^{2\pi}\frac{\sin(2\theta)}{2}\mathrm{d}\theta\\&=I_0\pi\int\limits_0^{\pi/2}\sin(2\theta)\mathrm{d}\theta=I_0\pi\left(-\frac{1}{2}\cos(2\theta)\right)\Bigg|_0^{\pi/2}=I_0\pi\left(\frac{1}{2}-\left(-\frac{1}{2}\right)\right)=I_0\phi\end{aligned} \tag{10.5.36}$$

The energy emittance being equal to:

$$M=\frac{\Phi}{S}\Rightarrow M=\frac{I_0\pi}{S}=\frac{I_0}{S}\pi=L\pi \tag{10.5.37}$$

This result is important for the study of black-body radiation, since the luminance value measured by a sensor gives the possibility to deduce the emittance M, so the energy flow from the source is equal to:

$$\Phi=SL\pi \tag{10.5.38}$$

Here is a small summary of what we have seen so far as sometimes it can be quite confusing:

Radiometry			**Photometry**		
Quantity	**Symbol**	**Units**	**Units**	**Symbol**	**Quantity**
Radiant Energy	E_r	J	lm s	E_r	Luminous Energy
Radiant Flow	Φ	W	lm s	Φ	Luminous Radiant
Light Intensity	I	W/sr	Cd	I	Radiant Intensity
Emittance	M	W/m^2	lm/m^2	M	Radiant Emittance
Radiant Intensity	I	W/sr	lm/sr	I	Luminous Intensity
Radiance	L	W/(m^{2}sr)	lm/(m^{2}sr)	L	Luminance

Table 10.11: **Radiometric and Photometric quantities**

10.5.3.5 Kirchhoff's law of Radiation

Any body irradiated by an energy source sees the incident energy flow incident divided into three intuitive terms:

$$\Phi=\Phi_r+\Phi_t+\Phi_a \tag{10.5.39}$$

where:

- $\Phi_r=\rho\Phi$: is the energy flow reflected or diffused.
- $\Phi_t=\tau\Phi$: is the energy that flow through the body without interactions (full transparency).
- $\Phi_a=\alpha\Phi$: is the energy flow transformed into other forms of energy.

The three coefficients respectively named "reflectance factor ρ" or more simply "albedo", "transmittance factor τ" and "absorption factor α" depend obviously on the wavelength λ (and therefore the frequency) of the incident light and the of the temperature of receiver body.

For each object we have obviously:

$$\rho+\tau+\alpha=1 \tag{10.5.40}$$

which is the expression of the "simple Kirchhoff law" in photometry (at the opposite of the differential one).

The albedo is an important concept in climatology as we will use it in the corresponding section of the book (in the legend of the image below, the years are given based on the holocene calendar).

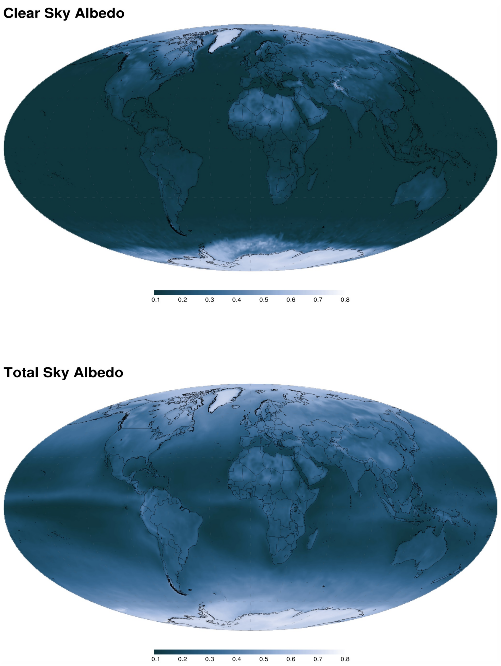


Figure 10.137: 12003–12004 mean annual clear-sky and total-sky albedo (source: Wikipedia)

A interesting (funny) but also very useful modern material for space telescopes imagery quality is Vantablack whose albedo is of 0.00035...! The result is quite mind blowing as you can see in the photo below:

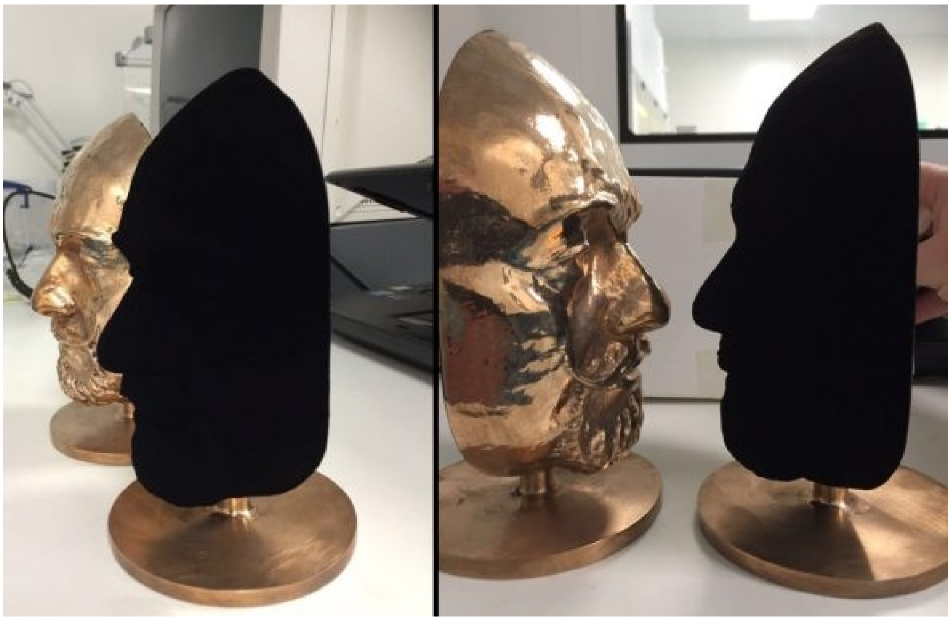


Figure 10.138: **Vantablack painted face...**

Looks like quite unreal isn't it? ;-)

10.5.3.6 Spectral Decomposition

From what has been said above, it follows that all the previously defined quantities can be related to their spectral decomposition in wavelength λ. This results comes from the superposition principle: any radiation can be treated as the superposition of monochromatic radiation (even if we will prove during our study of quantum physics later that a monochromatic radiation does not really exists).

Thus, we define:

$$\Phi=\int_{\lambda}\Phi(\lambda)\mathrm{d}\lambda \quad I=\int_{\lambda}I(\lambda)\mathrm{d}\lambda \quad L=\int_{\lambda}L(\lambda)\mathrm{d}\lambda \quad M=\int_{\lambda}M(\lambda)\mathrm{d}\lambda \tag{10.5.41}$$

Where we also write to simplify the notations:

$$\Phi(\lambda)=\Phi_{\lambda} \quad I(\lambda)=I_{\lambda} \quad L(\lambda)=L_{\lambda} \quad M(\lambda)=M_{\lambda} \tag{10.5.42}$$

Remark

The units AND values of spectral decomposed energy flow, spectral intensity, spectral luminance/radiance, spectral emittance and also the spectral absorption factor, spectral reflectance factor and spectral transmission factor are of course not equivalent to their integrated expression.

We will need for the density of the emittance in the study of black-body in the section of Thermodynamics of the chapter Mechanics. Remember only that we have in S.I. units based on the spectral decomposition principle (and vice versa superposition):

$$M\,[\mathrm{W}\cdot\mathrm{m}^{-2}] \qquad M(\lambda)\,[\mathrm{W}\cdot\mathrm{m}^{-3}] \tag{10.5.43}$$

Remark

We have seen in the section of Thermodynamic that parameters defined above, being dependent on wavelength, they are also dependent on the temperature of the source which emits these waves.

10.5.4 Law of Refraction and Reflection

The law of refraction, which is generally known as "Snell's law", governs the behaviour of light-rays as they propagate across a sharp interface between two transparent dielectric media.

Pierre de Fermat proposed that light rays (electromagnetic waves) met a very general principle that the path taken by the light to travel from one point to another was the one for which the travel time was minimum (actually an extremum which may be a minimum or maximum). This proposal, known as "Fermat's principle of least time", a fundamental principle of geometrical optics is based on the principle of least action (principle we have already introduced in the section of Analytical Mechanics) as we prove it later below.

Before starting the developments let us give some important definitions and introduce some concepts. some of which are related to the figure below:

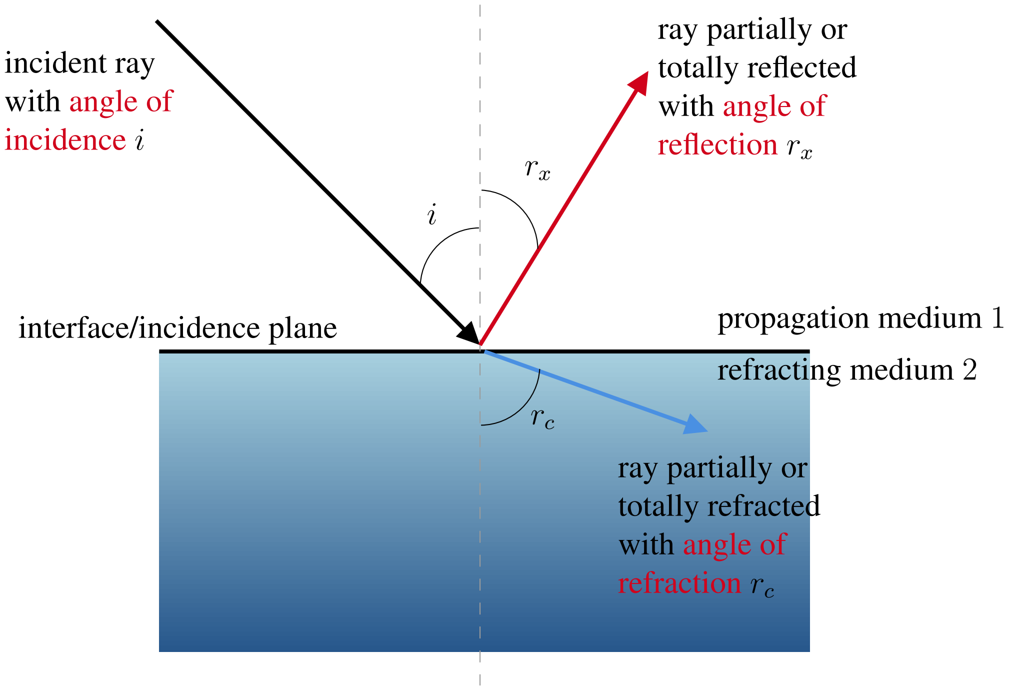
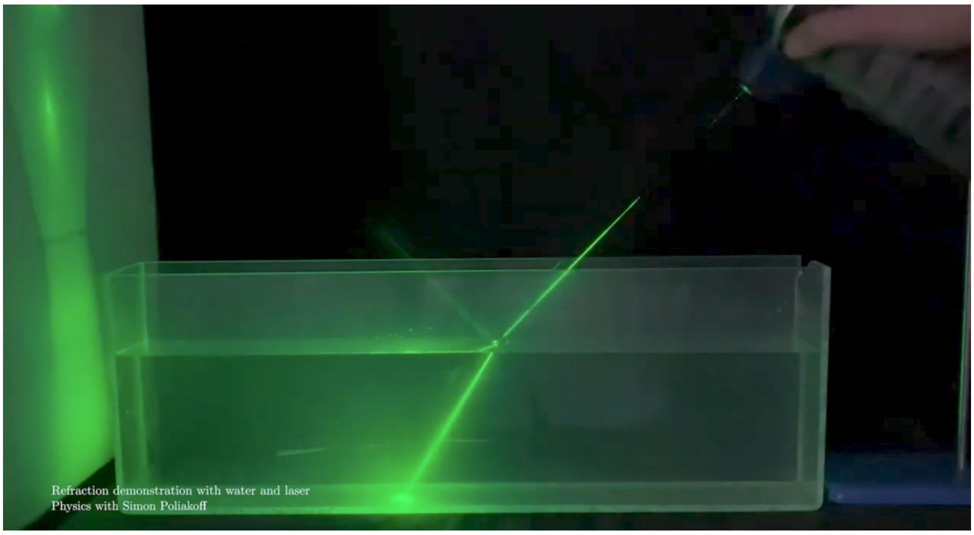

incident ray
with angle of
incidence i
ray partially or
totally reflected
with angle of
reflection r_x
r_x
i
interface/incidence plane
propagation medium 1
refracting medium 2
r_c
ray partially or
totally refracted
with angle of
refraction r_c

Figure 10.139: **Vocabulary for the study of geometrical optics**

Or in real life a demonstration of refraction and reflection with a LASER, a recipient of water (and a spray to highlight the reflected light):


Refraction demonstration with water and laser
Physics with Simon Poliakoff

Definitions (#10.5.492):

D1. A "refracting medium" is a medium that causes the deflection of an incident light ray.

D2. The "incident ray" is the ray of light propagating in a medium 1, pass wholly or partially in a refracting medium 2, the rest being absorbed or partially reflected.

D3. The "angle of incidence", sometimes denoted i is the angle by which the incident beam enters the refracting medium.

D4. The "totally or reflected ray" is the part of the light ray that having met the interface between the propagation medium and the refractive medium, continues his path in the propagation medium.

D5. The "reflection angle", sometimes denoted r_x or simply r if there is no possible confusion, is the angle by which the ray is reflected from the plane representing the interface between the propagation medium and itself. We will prove that the incident and reflected angles are equal in absolute values.

D6. The "partially or totally refracted ray" is the portion of the light ray having met the interface between the propagation medium and the refractive medium, continues its path in the refracting medium.

D7. The "angle of refraction", sometimes denoted r_c or sometimes simply r if there is no confusion, is the angle by which the ray is refracted by the plane representing the interface between the propagation medium and refracting medium. The incident and refracted angles are linked by a relation we prove further below.

D8. "Dioptrics" is the branch of optics dealing with refraction, especially by lenses. In contrast, the branch dealing with mirrors, reflection, is known as "catoptrics" (that's why telescopes that create their image with an objective that is a convex lens - refractors - are said to be "dioptric" telescopes).

An illustration that could help to understand some of these definitions better is the following:

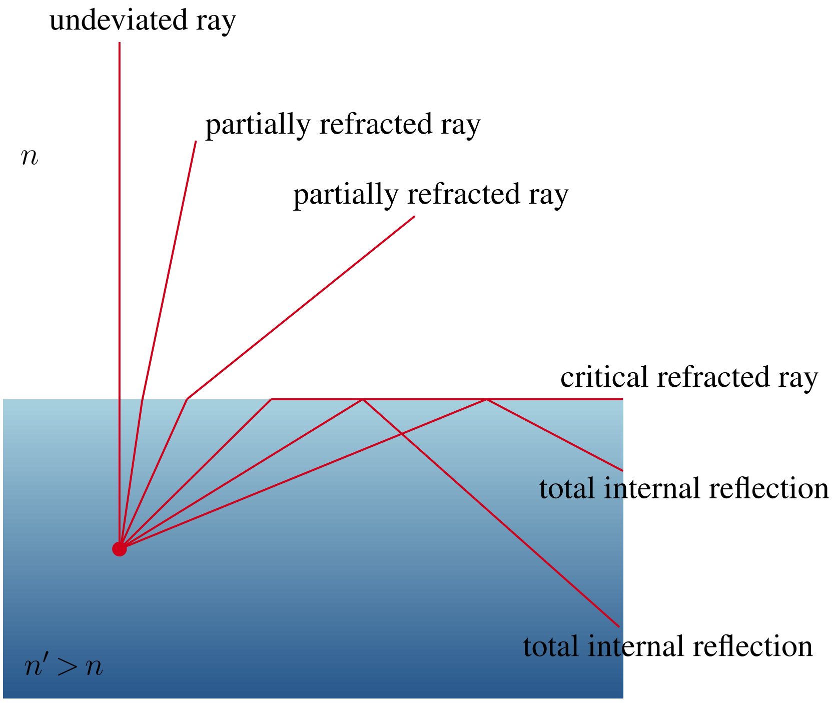

undeviated ray
partially refracted ray
n
partially refracted ray
critical refracted ray
total internal reflection
total internal reflection
$n' > n$

Figure 10.140: **Types of refractions and reflections**

For example with water this can give quite funny observations:

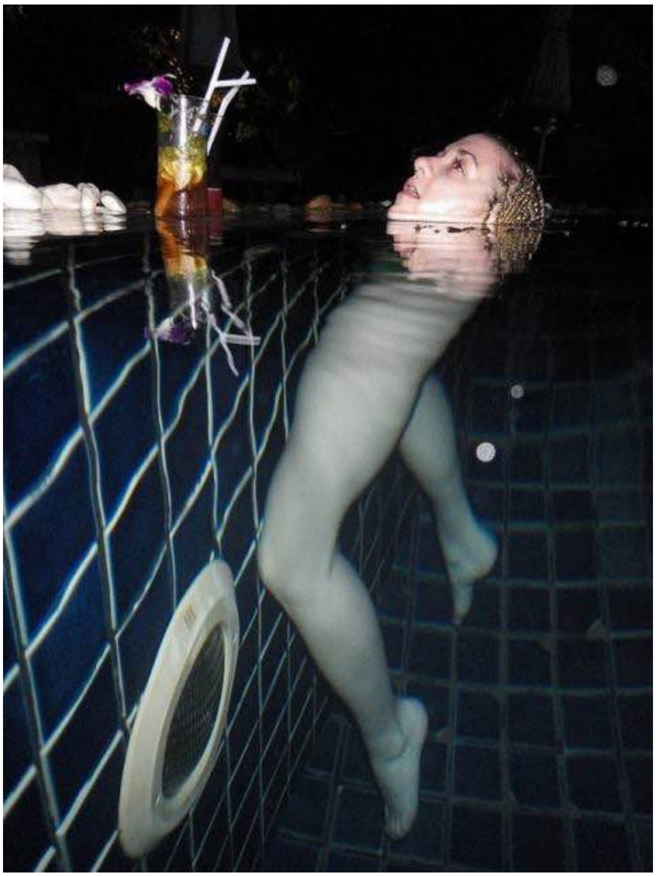


Figure 10.141: **Funny water reflection effect**

with its corresponding schematic explanation:

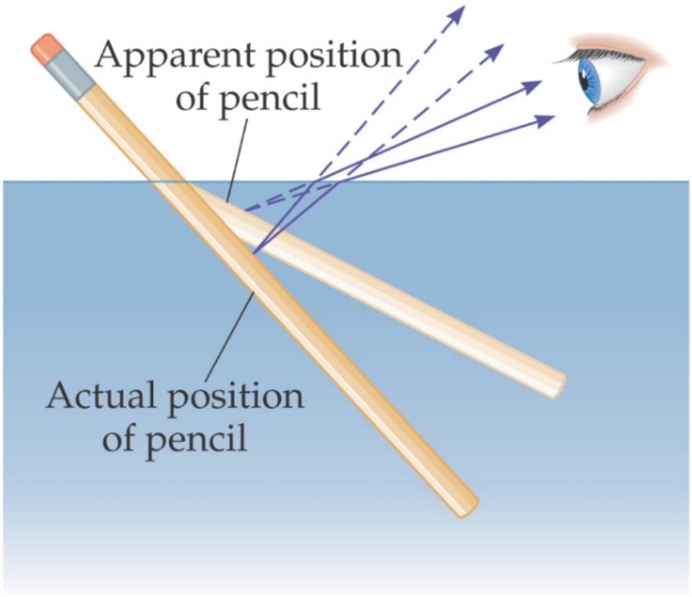


Figure 10.142: Water reflection ray diagram

Or a more tricky example that can be reproduced at home:

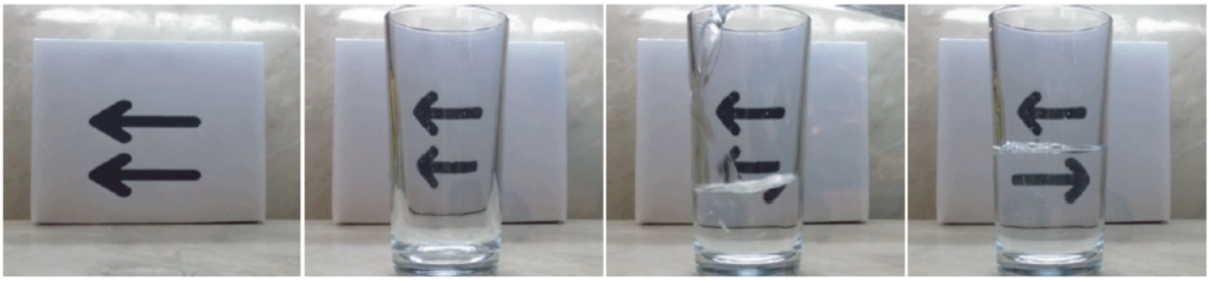


Figure 10.143: Source: Maric Vladimir

Before we continue, something important must be understand (or recall): "ideal refraction" must be discerned form "diffused refraction"! Since the light strikes different parts of the surface at different angles, it is reflected in many different directions, or diffused. So when we study refraction we in physics, we consider an "ideal refraction" with a "directional source of light"!

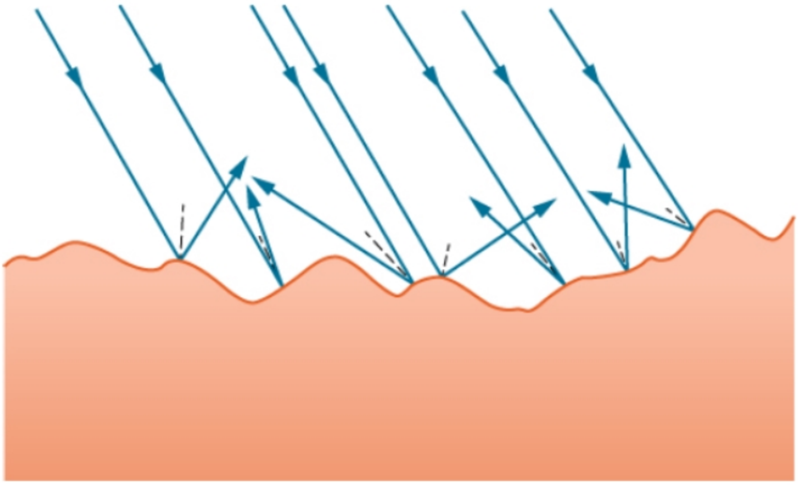
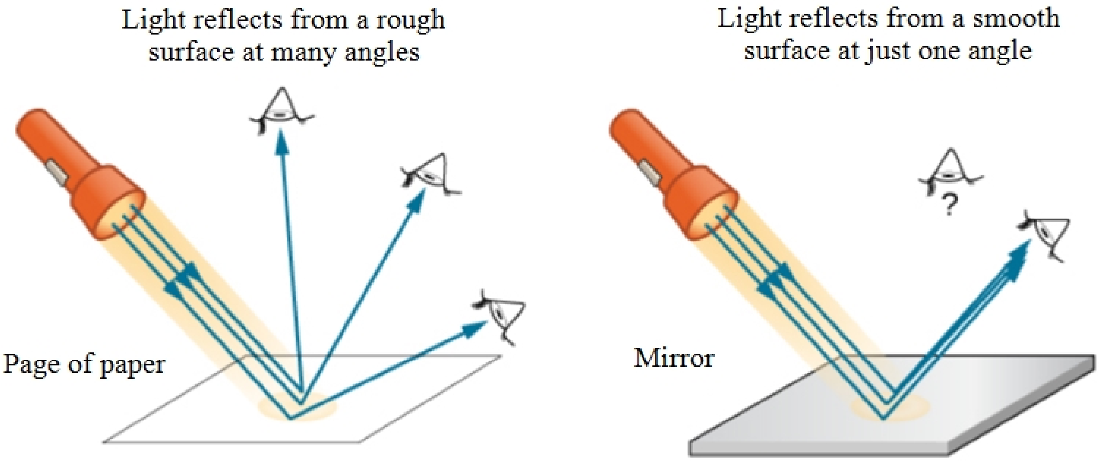


Figure 10.144: **Light is diffused when it reflects from a rough surface (source: OpenStax)**

Diffused light is what allows us to see a sheet of paper from any angle:


Light reflects from a rough surface at many angles
Light reflects from a smooth surface at just one angle
Page of paper
Mirror
?

10.5.4.1 Refractive index

The "absolute refractive index n_λ" of a medium at a given wavelength λ (and hence at a given frequency ν) measure the reduction factor of the phase velocity of light, denoted v_n in Geometric Optics, in the considered medium relatively to the speed of light in vacuum and is given quite generally by the "Cauchy's formula" (see section Electrokinetics page 4244):

$$n_\lambda=\frac{c}{v_n}=A+\frac{B}{\lambda^2}+\frac{C}{\lambda^4}+\ldots\cong A+\frac{B\nu^2}{c^2} \tag{10.5.44}$$

where A and B are experimentally determined constants. We can notice through the Cauchy's formula that the absolute refractive index decreases as the wavelength increases (verbatim when the frequency decrease).

Anyway, the theory of light-matter interaction on which Cauchy based this equation was later found to be incorrect. The "Sellmeier equation" is a later development of Cauchy's work that handles anomalously dispersive regions, and more accurately models a material's refractive index across the ultraviolet, visible, and infra-red spectrum and that is given by (see proof earlier above page 4243):

$$n_\lambda=\frac{c}{v_n}\cong\sqrt{1+\frac{B_1\lambda^2}{\lambda^2-C_1}+\frac{B_2\lambda^2}{\lambda^2-C_2}+\frac{B_3\lambda^2}{\lambda^2-C_3}} \tag{10.5.45}$$

Following what we have seen in the section of Electrodynamics it comes immediately:

$$n=\frac{c}{v_n}=cv_n^{-1}=\frac{1}{\sqrt{\varepsilon_0\mu_0}}\left(\frac{1}{\sqrt{\varepsilon_r\varepsilon_0\mu_r\mu_0}}\right)^{-1}=\frac{\sqrt{\varepsilon_r\varepsilon_0\mu_r\mu_0}}{\sqrt{\varepsilon_0\mu_0}}=\sqrt{\varepsilon_r\mu_r} \tag{10.5.46}$$

Therefore:

$$\boxed{n_\lambda=\frac{c}{v_n}\cong\sqrt{\varepsilon_r\mu_r}\cong\sqrt{1+\frac{B_1\lambda^2}{\lambda^2-C_1}+\frac{B_2\lambda^2}{\lambda^2-C_2}+\frac{B_3\lambda^2}{\lambda^2-C_3}}} \tag{10.5.47}$$

Remark

Obviously in reality all that stuff is also a function of the temperature and... not only...

All pure materials have absolute refractive index of a positive value greater than 1. More a medium is dense, more the phase velocity of light is slowed down, more the absolute refractive index is high.

Remark

There exist however material with smaller than 1 refractive index and even negative-index meta-material.

The process of describing light transport via the quantum mechanical description isn't trivial to explain the "slow down" of the phase velocity. The use of photons to explain such process involves the understanding of not just the properties of photons, but also the quantum mechanical properties of the material itself (something one learns in Solid State Physics courses).

A common explanation that has been provided is that a photon moving through the material still moves at the speed of c, but when it encounters the atom of the material, it is absorbed by the atom via an atomic transition. After a very slight delay, a photon is then re-emitted. This explanation is incorrect and inconsistent with empirical observations. If this is what actually occurs, then the absorption spectrum will be discrete because atoms have only discrete energy states. Yet, in glass for example, we see almost the whole visible spectrum being transmitted with no discrete disruption in the measured speed. In fact, the index of refraction (which reflects the speed of light through that medium) varies continuously, rather than abruptly, with the frequency of light.

Secondly, if that assertion is true, then the index of refraction would ONLY depend on the type of atom in the material, and nothing else, since the atom is responsible for the absorption of the photon. Again, if this is true, then we see a problem when we apply this to carbon, let's say. The index of refraction of graphite and diamond are different from each other. Yet, both are made up of carbon atoms. In fact, if we look at graphite alone, the index of refraction is different along different crystal directions. Obviously, materials with identical atoms can have different index of refraction. So it points to the evidence that it may have nothing to do with an "atomic transition".

When atoms and molecules form a solid, they start to lose most of their individual identity and form a "collective behaviour" with other atoms. It is as the result of this collective behaviour that one obtains a metal, insulator, semiconductor, etc. Almost all of the properties of solids that we are familiar with are the results of the collective properties of the solid as a whole, not the properties of the individual atoms. The same applies to how a photon moves through a solid.

A solid has a network of ions and electrons fixed in a "lattice". Think of this as a network of balls connected to each other by springs. Because of this, they have what is known as "collective vibrational modes", often named "phonons" as we will see it during our study of semi-conductors in the section Electrokinetics. These are quanta of lattice vibrations, similar to photons being the quanta of EM radiation. It is these vibrational modes that can absorb a photon. So when a photon encounters a solid, and it can interact with an available phonon mode (i.e. something similar to a resonance condition), this photon can be absorbed by the solid and then converted to heat (it is the energy of these vibrations or phonons that we commonly refer to as heat). The solid is then opaque to this particular photon (i.e. at that frequency). Now, unlike the atomic orbitals, the phonon spectrum can be broad and continuous over a large frequency range. That is why all materials have a "bandwidth" of transmission or absorption. The width here depends on how wide the phonon spectrum is.

On the other hand, if a photon has an energy beyond the phonon spectrum, then while it can still cause a disturbance of the lattice ions, the solid cannot sustain this vibration, because the phonon mode isn't available. This is similar to trying to oscillate something at a different frequency than the resonance frequency. So the lattice does not absorb this photon and it is re-emitted but with a very slight delay. This, naively, is the origin of the apparent slowdown of the light speed in the material. The emitted photon may encounter other lattice ions as it makes its way through the material and this accumulate the delay.

Moral of the story: the properties of a solid that we are familiar with have more to do with the "collective" behaviour of a large number of atoms interacting with each other. In most cases, these do not reflect the properties of the individual, isolated atoms.

Finally, the "relative refractive index n_{12}" specifically refers to comparing (ratio) of the absolute refractive index of an optically dense media to another background media such that:

$$n_{12}=\frac{n_1}{n_2}=\frac{\dfrac{c}{v_{p_1}}}{\dfrac{c}{v_{p_2}}}=\frac{v_{p2}}{v_{p1}} \tag{10.5.48}$$

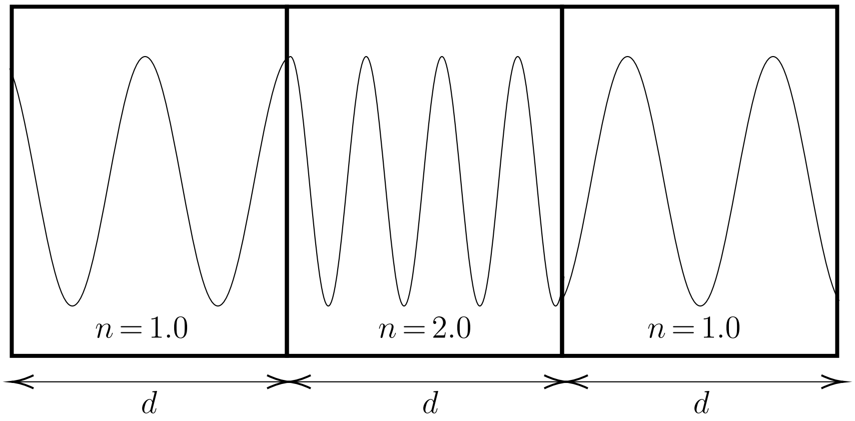
We can also use wavelengths to calculate refractive index. By substituting (notice that by conservation of energy the frequency is assumed to be the constant as $E=h\nu$):

$$v_{p_1}=\nu\lambda_{p_1} \qquad v_{p_2}=\nu\lambda_{p_2} \tag{10.5.49}$$

then:

$$\boxed{n_{12}=\frac{n_1}{n_2}=\frac{\dfrac{c}{v_{p_1}}}{\dfrac{c}{v_{p_2}}}=\frac{v_{p2}}{v_{p1}}=\frac{\lambda_{p2}}{\lambda_{p1}}} \tag{10.5.50}$$

This can be depicted by the following figure:



10.5.4.1.1 Index of refraction rate of change with altitude

The index of refraction obviously doesn't only depend on the wavelength as proven by the Sellmeier equation introduced earlier above. It depends also on temperature, pressure, medium density and gravitational acceleration when studied in the context of gases in Earth atmosphere. This point is especially important to understand some visual mirages (see further below at page 4397).

To determine the most famous relation of the index of refraction rate of change with altitude let us remind the Ideal Gas law (see section Continuum Mechanics page 3903)

$$PV=nRT \tag{10.5.51}$$

Now we have knowing the moral mass M_m:

$$PV=\frac{m}{M_m}RT \tag{10.5.52}$$

Hence:

$$\rho=\frac{m}{V}=\frac{PM_m}{mRT} \tag{10.5.53}$$

Assuming pressure and temperature are functions of altitude z:

$$\rho(z):=\frac{M_mP(z)}{RT(z)} \tag{10.5.54}$$

Let us take the derivative of density with respect to altitude:

$$\frac{\mathrm{d}}{\mathrm{d}z}\rho(z)\rightarrow\frac{M_m\frac{\mathrm{d}}{\mathrm{d}z}P(z)}{RT(z)}-\frac{M_mP(z)\frac{\mathrm{d}}{\mathrm{d}z}T(z)}{RT(z)^2} \tag{10.5.55}$$

Let us reorganize that:

$$\frac{\mathrm{d}p}{\mathrm{d}z}(z)=\frac{M_m}{R}\left(\frac{\frac{\mathrm{d}P(z)}{\mathrm{d}z}T(z)-\frac{\mathrm{d}T}{\mathrm{d}z}P(z)}{T^2(z)}\right) \tag{10.5.56}$$

Now remember that:

$$\rho=\frac{M_mP}{RT} \tag{10.5.57}$$

and also that[15] (see page):

$$\frac{\mathrm{d}P}{\mathrm{d}z}=-\rho(z)g \tag{10.5.58}$$

Let us define the "lapse rate":

$$\gamma:=-\frac{\mathrm{d}T}{\mathrm{d}z} \tag{10.5.59}$$

Injecting these three relations into:

$$\frac{\mathrm{d}p}{\mathrm{d}z}(z)=\frac{M_m}{R}\left(\frac{\frac{\mathrm{d}P(z)}{\mathrm{d}z}T(z)-\frac{\mathrm{d}T}{\mathrm{d}z}P(z)}{T^2(z)}\right) \tag{10.5.60}$$

we get:

$$\frac{\mathrm{d}\rho}{\mathrm{d}z}(z)\cong\frac{M_m}{R}\left(-\frac{g\rho(z)}{T(z)}+\frac{\gamma P(z)}{T^2(z)}\right) \tag{10.5.61}$$

Let us choose a simple model is that the index of refraction varies linearly with atmospheric density ρ, which is a function of altitude z:

$$n(z)\cong\alpha\rho(z)+1\Leftrightarrow\rho(z)\cong\frac{n(z)-1}{\alpha} \tag{10.5.62}$$

---

[15] We know that g is not constant in height but that a good approximation for the first few kilometres of altitude to assume it constant!

where α is a given dimensional constant.

Therefore we also have:

$$\alpha\frac{\mathrm{d}\rho(z)}{\mathrm{d}z}\cong\frac{\mathrm{d}n(z)}{\mathrm{d}z}\Leftrightarrow\frac{\mathrm{d}\rho(z)}{\mathrm{d}z}\cong\frac{1}{\alpha}\frac{\mathrm{d}n(z)}{\mathrm{d}z}\qquad(10.5.63)$$

Hence:

$$\begin{aligned}\frac{\mathrm{d}\rho}{\mathrm{d}z}(z)&\cong\frac{M_m}{R}\left(-\frac{g\rho(z)}{T(z)}+\frac{\gamma P(z)}{T^2(z)}\right)\\&=-\frac{M_m g}{TR}\left(\frac{n(z)-1}{\alpha}\right)+\frac{M_m}{R}\frac{\gamma P(z)}{T^2(z)}\\&=-\frac{\rho(z)g}{P(z)}\left(\frac{n(z)-1}{\alpha}\right)+\frac{\gamma}{T(z)}\left(\frac{n-1}{\alpha}\right)\end{aligned}\qquad(10.5.64)$$

In the last equality:

$$\frac{\mathrm{d}\rho}{\mathrm{d}z}(z)=-\frac{\rho(z)g}{P(z)}\left(\frac{n(z)-1}{\alpha}\right)+\frac{\gamma}{T(z)}\left(\frac{n-1}{\alpha}\right)\qquad(10.5.65)$$

We can substitute:

$$\frac{\mathrm{d}\rho(z)}{\mathrm{d}z}\cong\frac{1}{\alpha}\frac{\mathrm{d}n(z)}{\mathrm{d}z}\qquad(10.5.66)$$

So we get:

$$\frac{1}{\alpha}\frac{\mathrm{d}n(z)}{\mathrm{d}z}=-\frac{\rho(z)g}{P(z)}\left(\frac{n(z)-1}{\alpha}\right)+\frac{\gamma}{T(z)}\left(\frac{n-1}{\alpha}\right)\qquad(10.5.67)$$

So:

$$\frac{\mathrm{d}n}{\mathrm{d}z}\cong-(n(z)-1)\left(\frac{\rho(z)g}{P(z)}-\frac{\gamma}{T(z)}\right)=-\frac{(n(z)-1)}{T(z)}\left(\frac{\rho(z)gT(z)}{P(z)}-\gamma\right)\qquad(10.5.68)$$

Finally:

$$\boxed{\frac{\mathrm{d}n}{\mathrm{d}z}\cong-\frac{(n(z)-1)}{T(z)}\left(\frac{\rho(z)gT(z)}{P(z)}+\frac{\mathrm{d}T}{\mathrm{d}z}\right):=-\frac{(n(z)-1)}{T(z)}(\gamma_g-\gamma)}\qquad(10.5.69)$$

✍**Example:**

Let us take the following quite well known common lapse rate assumption:

$$\gamma=0.0113\,[\Delta°\mathrm{C}\cdot\mathrm{m}^{-1}]$$

And a typical air temperature:

$$T=295.15\,[\mathrm{K}]$$

But also a typical air pressure at sea level:

$$P=101,325\,[\mathrm{Pa}]$$

We take for the average acceleration on Earth at sea level gravity:

$$g = 9.80665\,[\text{m}\cdot\text{s}^{-1}]$$

and the index of refraction of air at $0\,[^\circ\text{C}]$ and at the sea level:

$$n = 1.000293$$

For the density of air at standard ambient temperature and pressure:

$$\rho_{\text{SATP}} = 1.168\,[\text{kg}\cdot\text{m}^{-3}]$$

Then we get:

$$\begin{aligned}
n' &\cong -\frac{(n(z)-1)}{T(z)}\left(\frac{\rho(z)gT(z)}{P(z)}-\gamma\right)\\
&\cong -\frac{(1.000293-1)}{295.15}\left(\frac{1.168\cdot 9.80665\cdot 295.15}{101,325}-0.0113\right)\\
&\cong -2.1904\cdot 10^{-8}\,[\text{m}^{-1}]
\end{aligned}$$

10.5.4.2 Snell's law

Let us consider (see figure below) two respective medium[16] M_n and M_m of refractive indices n and m (implicitly dependent on the wavelength as we will consider constant temperature) and whose contact surface is flat. Consider two points A and B respectively in the medium of index n (point A) and in the medium of index m (point B).

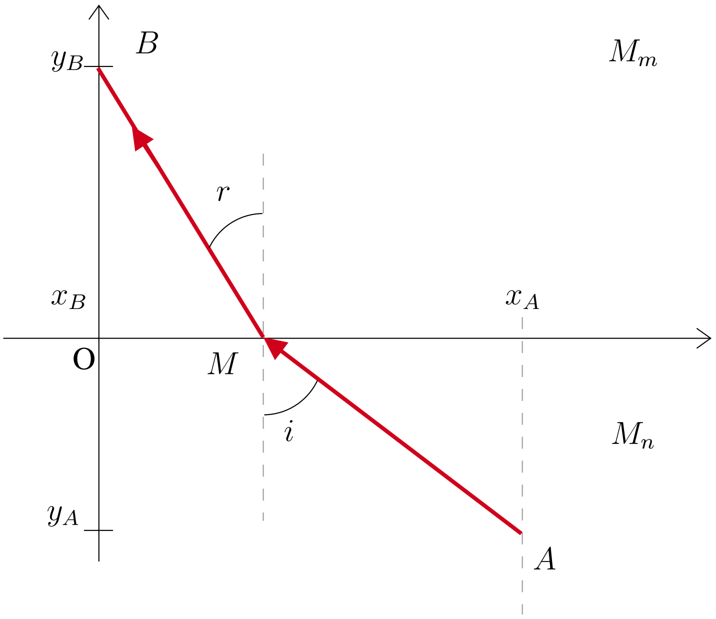
Consider the path of light going from A to B. Fermat's principle teaches us that the path taken by the light is such that the time taken to travel is minimum.

Therefore we propose in a first time to apply a standard method to calculate the path of the light ray and secondly, we will show that Fermat's principle can be stated as a variational principle!!!

Let us choose a reference frame that simplifies the problem: let us make coincide the x-axis with the contact plane of both medium and the y-axis through point B. In such a reference frame, points A and B have the following coordinates: $A(x_A, y_A)$, $B(0, y_B)$.

Let us denote by $M(x, 0)$, the point where the light beam passes through the contact surface between the two medium. The time t taken by light to get from A to B is by applying simple kinematics relations:

$$t=\frac{\overline{AM}}{v_n}+\frac{\overline{MB}}{v_m}=\frac{n\overline{AM}}{c}+\frac{m\overline{MB}}{c} \tag{10.5.70}$$



where:

$$v_n=\frac{c}{n} \quad \text{and} \quad v_m=\frac{c}{m} \tag{10.5.71}$$

are the phase velocity of light in the medium M_n and M_m.

---

[16] It can also be one medium but with two different physical properties values like different temperature, different density and different pressure.

We can observe on the figure above that the incident rays are refracted on the other side of the axis perpendicular to the interface. This is a typically characteristic of material having a positive refraction index. But it is physically possible to build since the 11990s (holocene calendar) artificial composites "meta-materials" having negative refraction index:

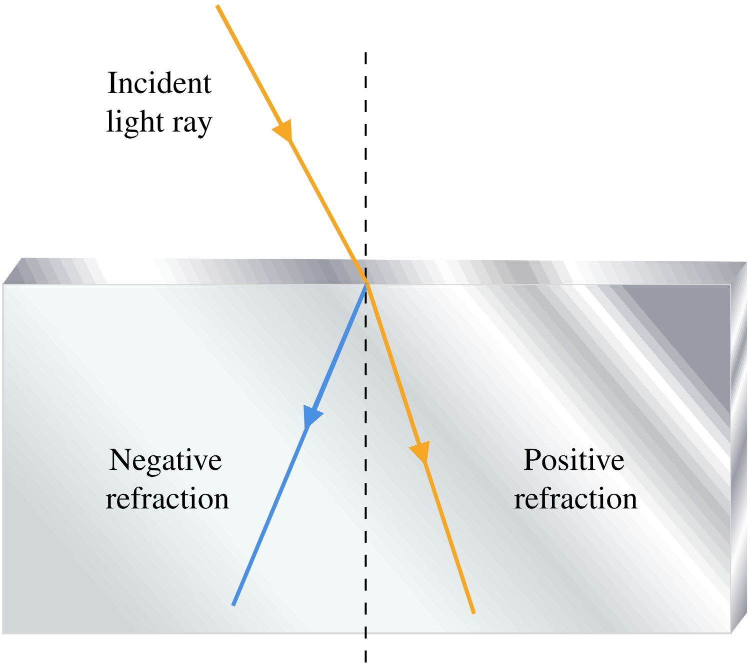


Figure 10.145: **Positive and negative refraction index material behaviour**

The writing of the two previous relations:

$$v_n=\frac{c}{n}\quad \text{and} \quad v_m=\frac{c}{m} \tag{10.5.72}$$

Developing the values of $\overline{AM}$ and $\overline{MB}$ we get the following dependence of t as a function of the position x of M:

$$t(x)=\frac{n\sqrt{(x_A-x)^2+y_A^2}}{c}+\frac{m\sqrt{x^2+y_B^2}}{c} \tag{10.5.73}$$

According to Fermat's principle, the path taken by the light is the one for which t is minimum. The extremum of $t(x)$ is reached when the derivative with respect to x is zero. Therefore:

$$\frac{\mathrm{d}t}{\mathrm{d}x}=\frac{n}{c}\frac{x-x_A}{\sqrt{(x_A-x)^2+y_A^2}}+\frac{m}{c}\frac{x}{\sqrt{x^2+y_B^2}} \tag{10.5.74}$$

Notice that:

$$\frac{x-x_A}{\sqrt{(x_A-x)^2+y_A^2}}=\frac{x-x_A}{\overline{AM}}=-\sin(i)\quad \text{and} \quad \frac{x}{\sqrt{x^2+y_B^2}}=\frac{x}{\overline{MB}}=\sin(r) \tag{10.5.75}$$

The condition of an extremum time brought by the light is then expressed by:

$$\frac{\mathrm{d}t}{\mathrm{d}x}=0\Leftrightarrow -\frac{n}{c}\sin(i)+\frac{m}{c}\sin(r)=0 \tag{10.5.76}$$

It is sufficient that the angles of incidence and refraction meet that condition so that the path travelled by light is actually the one that takes the least time.

Hence we get the relation, known as the "Snell's law" (which is not anymore a law as we just proved it):

$$\boxed{n\sin(i)=m\sin(r)} \tag{10.5.77}$$

We write more frequently the Snell-Descartes law in physics as following:

$$\boxed{n_{12}=\frac{n}{m}=\frac{n_1}{n_2}=\frac{\sin(r)}{\sin(i)}} \tag{10.5.78}$$

where the i stands for "incident" (often denoted θ_1) and the r for "refracted" (often denoted θ_2).

Remarks

R1. We will see during our study of wave optics that we can fall back (prove) the same relation but without the assumptions bases of geometrical optics. Therefore, the latter relation is named "Descartes-Snell relation " or just "Snell's law".

R2. When we speak of the relative refractive index of a given medium without reference to another medium, the default environment is the vacuum.

R3. Some materials do not have an absolute isotropic index of refraction : it then depends on the direction of propagation and the polarization state of light. This property is named "birefringence".

Let us now study the relation between the relative refractive index and the phase velocity of light in the various medium through which it passes.

A light ray connects two points A_1 and A_2 located on either side of an interface S between the two mediums. This light ray is not shown in the figure below. Only are drawn the path located on either side of the light ray that satisfy the extremum (we rely now on the study of the maximum path length). By hypothesis, they are extremely close, so the distance $\overline{PQ}$ is very small:

$$\overline{PQ}\ll\overline{A_1A_2} \tag{10.5.79}$$

We will admit that they correspond to the same travel time.

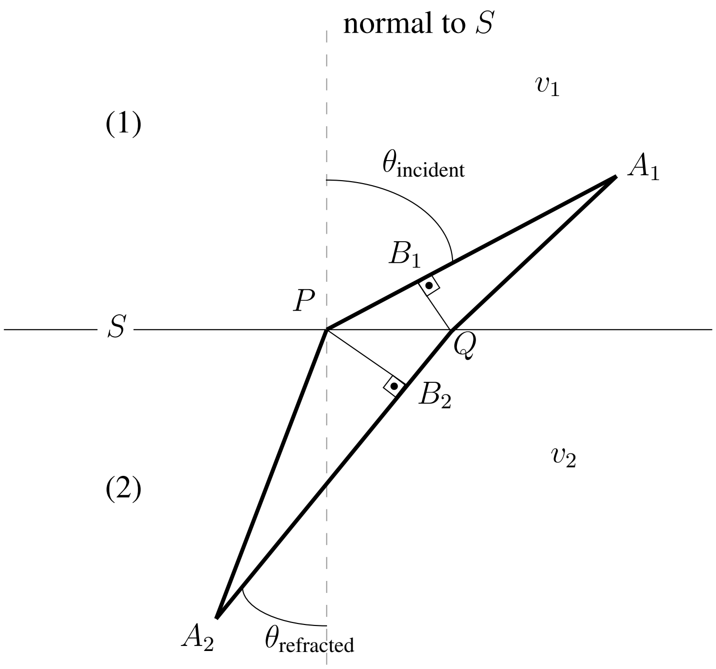


Figure 10.146: **Figure allowing to connect phase velocity and refractive index**

Since both paths are very close, we can assume the equality of the distances $\overline{A_1B_1}$ and $\overline{A_1Q}$ one hand, and of $\overline{PA_2}$ and $\overline{B_2A_2}$ of the other one. Thus, by hypothesis:

$$\frac{\overline{PB_1}}{v_1}=\frac{\overline{QB_2}}{v_2} \tag{10.5.80}$$

But, under the same hypothesis:

$$\begin{aligned}\overline{PB_1}=\overline{PQ}\cos\left(\frac{\pi}{2}-\theta_i\right)=\overline{PQ}\sin(\theta_i)\\ \overline{PB_2}=\overline{PQ}\cos\left(\frac{\pi}{2}-\theta_r\right)=\overline{PQ}\sin(\theta_r)\end{aligned} \tag{10.5.81}$$

such that:

$$\frac{\sin(\theta_r)}{\sin(\theta_i)}=\frac{\sin(r)}{\sin(i)}=\frac{v_2}{v_1} \tag{10.5.82}$$

The "refraction law" is finally stated in general as following:

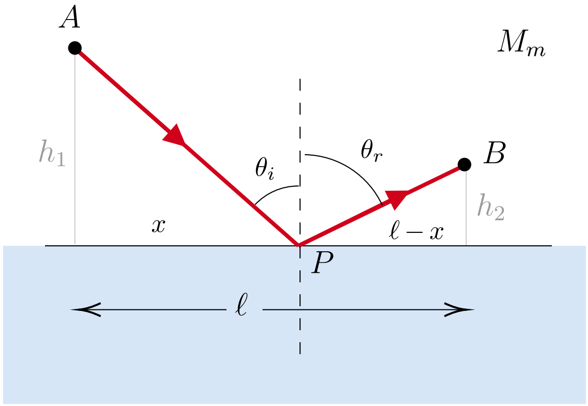
$$\boxed{n_{12}=\frac{v_2}{v_1}=\frac{n_1}{n_2}=\frac{\sin(r)}{\sin(i)}} \tag{10.5.83}$$

Remark

The angle $\pi/2-i$ (or often denoted $\pi/2-\theta_1$) is commonly named the "grazing angle".

And about the reflection angle, as we have already stated, it remains equal to the angle of incidence if the reflective surface is perfectly smooth and flat. This can also be proven using Fermat principle!

Indeed, consider the light ray shown in the figure below:



A ray of light starting at point A reflects off the surface at point P before arriving at point B, a horizontal distance ℓ from point A. We calculate the length of each path and divide the length by the speed of light to determine the time required for the light to travel between the two points:

$$t(x)=\frac{\sqrt{x^2+h_1^2}}{c}+\frac{\sqrt{(\ell-x)^2+h_2^2}}{c} \tag{10.5.84}$$

To minimize the time we set the derivative of the time with respect to x equal to zero. We also use the definition of the sine as opposite side over hypotenuse to relate the lengths to the angles of incidence and reflection. Hence:

$$0=\frac{\mathrm{d}t}{\mathrm{d}x}=\frac{x}{c\sqrt{x^2+h_1^2}}+\frac{-(\ell-x)}{c\sqrt{(\ell-x)^2+h_2^2}} \tag{10.5.85}$$

Rearranging and simplifying we get:

$$\frac{x}{\sqrt{x^2+h_1^2}}=\frac{(1\ell-x)}{\sqrt{(\ell-x)^2+h_2^2}} \tag{10.5.86}$$

Using the definition of sine this lead us to write:

$$\sin(\theta_i)=\sin(\theta_r) \tag{10.5.87}$$

So finally we get the "law of reflection":

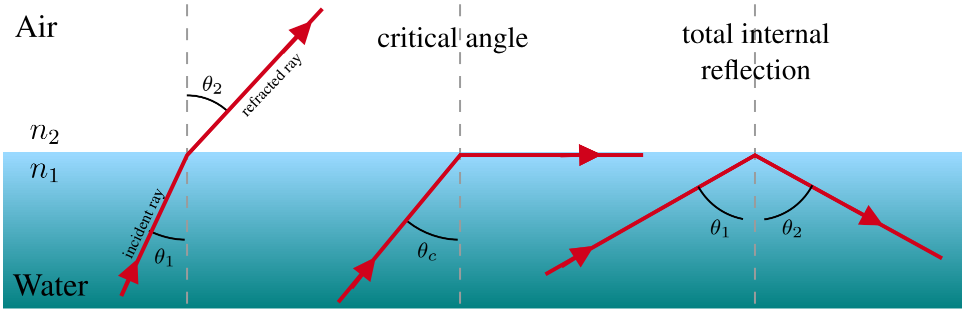
$$\boxed{\theta_i=\theta_r} \tag{10.5.88}$$

The law of reflection states then that on reflection from a smooth surface, the angle of the reflected ray is always equal to the angle of the incident ray!

Now let us come back to the law of refraction and rearranging it as following:

$$\sin(i)=\frac{n_2}{n_1}\sin(r) \tag{10.5.89}$$

and the case where $n_1>n_2$ (for example passage from water to air). Then, for values close to 1, that is to say for grazing angles of incidence (incident beam close to the surface), the Snell's law gives a value greater than 1. We then get out of the validity domain of the law. This corresponds to situations where there is no refraction but only for reflection, then we speak of "total reflection":


Air
n_2
n_1
Water
θ_2
refracted ray
incident ray
θ_1
critical angle
θ_c
total internal reflection
θ_1
θ_2

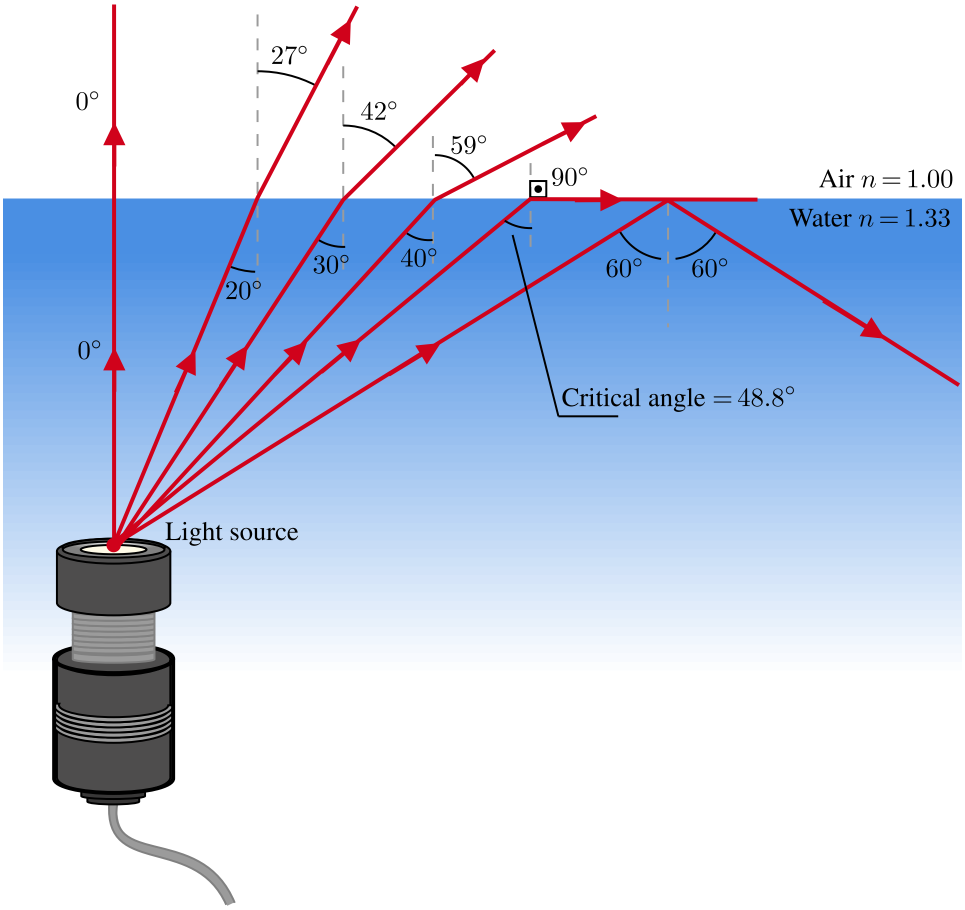

0°
27°
42°
59°
90°
Air $n = 1.00$
Water $n = 1.33$
20°
30°
40°
60°
60°
0°
Critical angle = 48.8°
Light source

Figure 10.147: **Critical angle schematic illustration (raytracing)**

To find the "critical angle", we find the value for θ_i when θ_r is equal to $\pi/2$ and thus $\sin(\theta_r) = 1$. The resulting value of θ_i is equal to the critical angle θ_c.

Now, we can solve for θ_i, and we get the equation for the critical angle:

$$\boxed{\theta_c = \theta_i = \arcsin\left(\frac{n_2 \sin(90^\circ)}{n_1}\right) = \arcsin\left(\frac{n_2}{n_1}\right)} \tag{10.5.90}$$

In the case of a water-air boundary where $n_1 = 1.33$ and $n_2 = 1$ then the critical angle is $\theta_c =$

$48.7°$. Note this only happens when transitioning from more dense (high index of refraction) to less dense (lower index of refraction) medium!

For angles greater than the critical angle, the incident light is totally reflected, obeying the law of reflection.

So in the real life we can get such things:

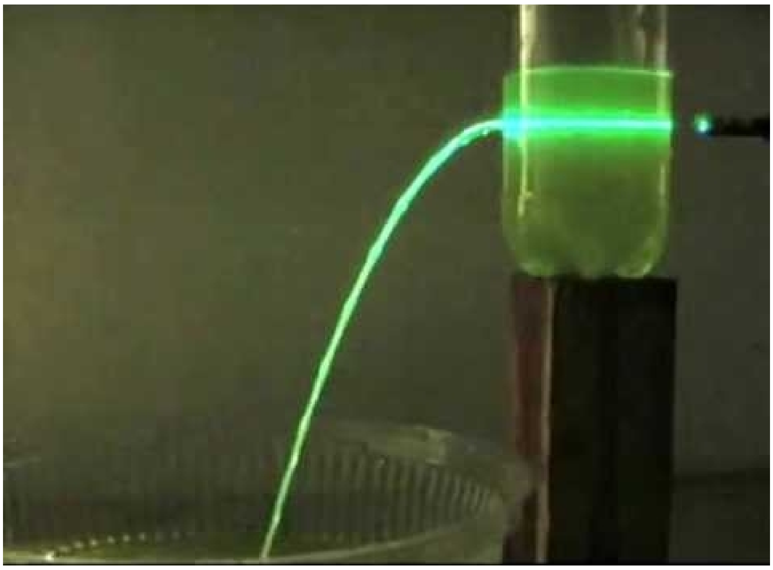


Figure 10.148: Air-Water total reflection example

Remark

Total internal reflection is used to carry light in fiber optics as we will see later below when introducing the mathematics of these information transmitters. The critical angle for a diamond-to-air surface is only $24.4°$, and so when light enters a diamond, it has trouble getting back out. Although light freely enters the diamond, it can exit only if it makes an angle less than $24.4°$ when he enters it.

Fermat's principle therefore has obvious similarities with the principle of least action in that it consists of a minimum principle. Even if a rigorous description of light requires the introduction of Quantum Mechanics, it is however possible to apprehend it with the tools of Analytical Mechanics and to apply on it, under certain assumptions, the principle of least action. We will prove just now that we then fall back on the Fermat's principle.

The calculations that we will present, introduce many hazardous assumptions, but this process should be considered as an approximation. It must be known to the reader that Fermat's principle also carries an approximation that we can qualify of "classical limit".

Now let us Imagine for the proof that light is composed of material "grains". We must admit that these grains obey to rather unusual physical properties: its mass is zero since according to the classical description, the light rays are not deflected by the gravitational field. This lack of mass thus makes them insensitive to the Earth's gravitational field (be careful and don't forget: we are in a "classical" description of the light as it was done in 119th century and before!!!!).

Let us write the action for one of these grains of light (see section Analytical Mechanics page

3162):

$$S=\int(T-V)\mathrm{d}t \tag{10.5.91}$$

Now by assuming that the only existing potential field V is that one that derives from the gravitational field and that we assume that light as it is insensitive to it (we know in General Relativity that this is wrong but we said just now that we would stay in classical point of view!!!), it follows that the action of light can be written:

$$S=\int T\mathrm{d}t \tag{10.5.92}$$

But no force applied to the light, so the kinetic energy T is a constant of motion. Applying the variational principle of least action we then have:

$$\delta S=0=\delta\left(\int T\mathrm{d}t\right)=T\delta\int\mathrm{d}t \tag{10.5.93}$$

Hence we get:

$$\delta\int\mathrm{d}t=0 \tag{10.5.94}$$

This equation means that the time taken by the light along its path is minimum (or more generally, is an extremum). We then fall back indeed on Fermat's principle. So we have proved that the under the classical limit Fermat's principle follows directly from the principle of least action.

Caution! The Fermat's principle doesn't explain anything, even if it's accurate that light follow the shortest path! That's why most textbooks introducing refraction as we did above geometrically (named sometimes the "marching soldier analogy", but wrong as nothing explain why the solders don't continue you walk straight) or using the variational principal (i.e. "Fermat's principle") are quite misleading. Using Huygens principle as we will see further below (see page 4470) is also wrong! In fact we should use Maxwell equation to understand better what happens physically!

10.5.4.3 Rainbow

A rainbow is an optical and meteorological phenomenon that makes the spectrum of light visible when the Sun (actually a G2V-type star, aka yellow dwarf) shines through the rain and the observer looks at the sky in a direction opposite to that of the Sun. It is a coloured arc with red on the outside and purple inside (excepted for higher order rainbows...).

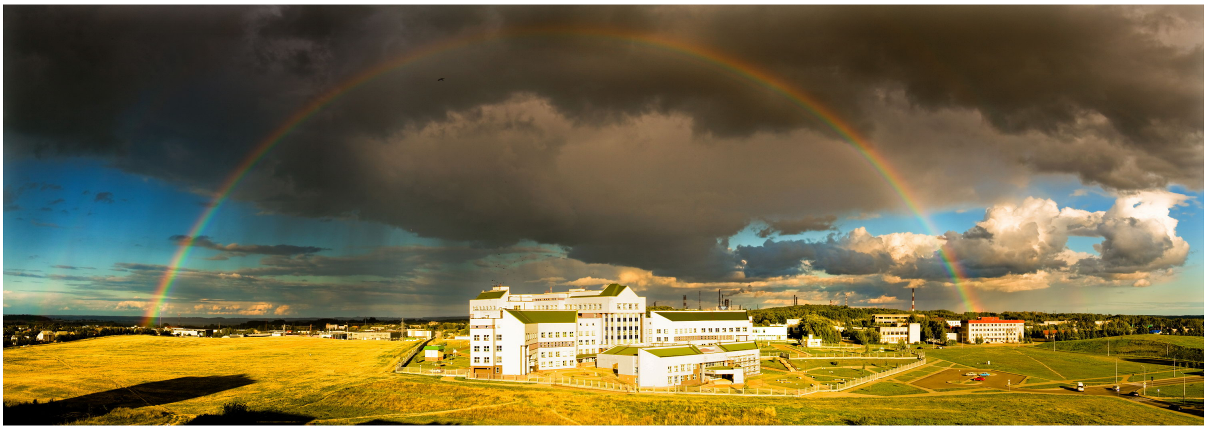


Figure 10.149: **Sun Rainbow after the rain, Grodno, Belarus (source: Wikipedia)**

The rainbow is caused by the scattering of sunlight by raindrops approximately spherical. The light is first refracted by penetrating the surface of the drop, then undergoes a partial reflection at the back of the drop and is refracted again on leaving (see figure below). The overall effect is that the incoming light is mainly refracted back at an angle of approximately $40-42^\circ$ (see proof below), regardless of the size of the drop. The precise value of the refraction angle depends on the wavelength (color) of light components as we know.

In the case of entry in a refracting medium, the angle of refraction of the blue light is lower than that of red light (phenomenon highlighted in the study of triangular prisms). Thus, after reflection at the water-air interface, the blue light comes out of a drop above the red light (see again figure below). The observer being fixed (at rest), he sees light from different drops of water with different angles with respect to the sunlight. Red therefore appears higher in the sky than the blue.

Sometimes a less bright second rainbow can be seen above the primary rainbow. It is caused by a double reflection of sunlight inside the raindrops and appears in a $50-53^\circ$ angle still in the opposite direction to the Sun. Due to the additional reflection, the colors of the second arc are reversed relative to the primary arc, with the blue outside and red inside, and the arc is less bright. That is why it is more difficult to observe.

quote
10.5:4372.#64

> *Understanding how a rainbow works doesn't make it less beautiful, it makes it more beautiful. That is Science!*
>
> Richard Dawkins

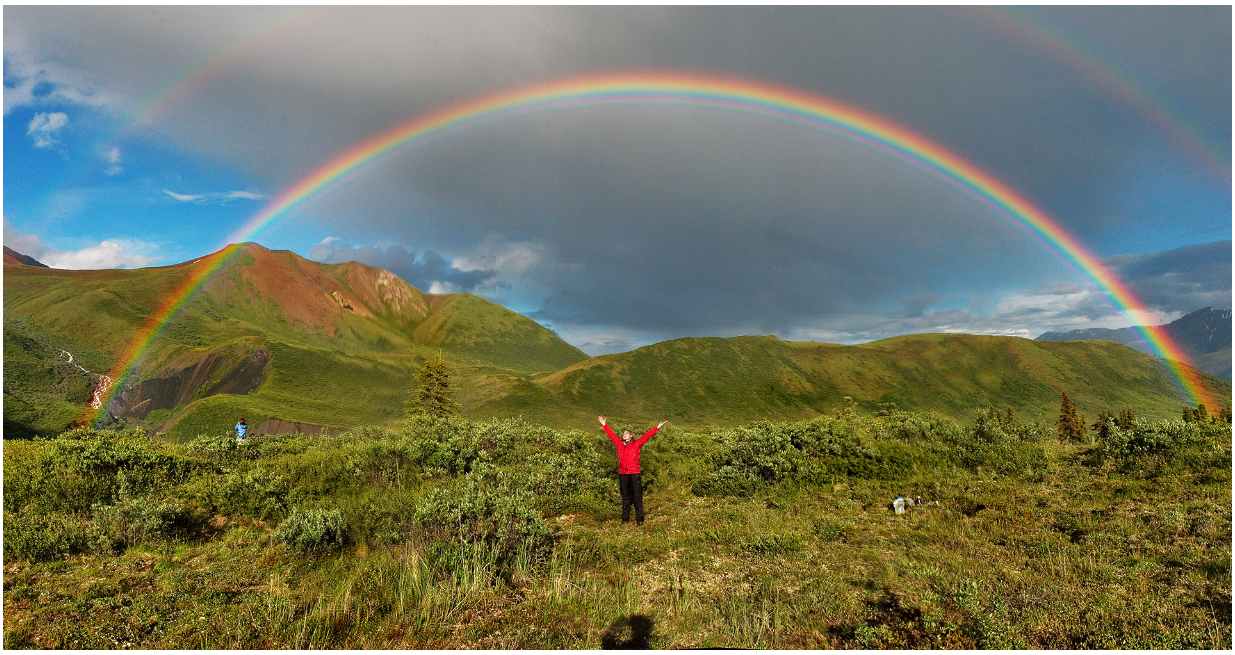


Figure 10.150: **Double Sun rainbow and supernumerary rainbows on the inside of the primary arc. The shadow of the photographer's head on the bottom marks the antisolar point (source: Wikipedia, Eric Rolph)**

A third rainbow sky may be present near the second, and inverted with respect thereto (i.e., identical to the first)...

However, third rainbow is much less bright and observable only under exceptional conditions. In practice it is not very easy to distinguish supernumerary arcs associated with the secondary arc. It corresponds to the light rays having undergone five reflections in the water drops.

Two rainbows inverted with respect to each other can also be observed in the opposite direction, about $45°$ from the Sun (therefore in the direction of it), but this is particularly difficult because of the proximity to the Sun. The few observations of these two arcs mention pieces of rainbows visible intermittently. These two rainbows correspond to light rays having undergone three and four reflections in the water drops. As they are located facing the Sun, it is not the same drops of water that contribute to it. In practice, favourable configurations to their observation are much fewer than those who favour the observation of the secondary arc, especially because of their proximity to the Sun.

To study the phenomenon let first consider the spherical water drop below with an incident beam of light (one refraction, two reflections) we represented the red and purple component (the angles are approximate) and the refractive indices water and air:

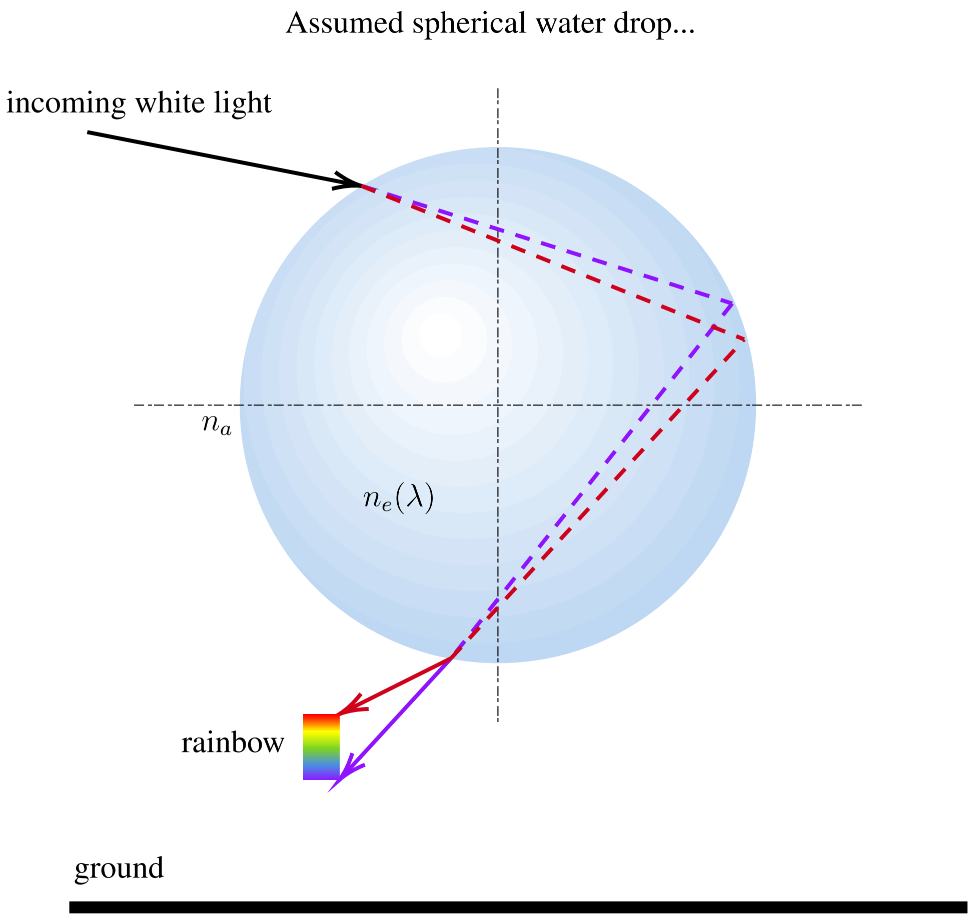


Figure 10.151: **Spherical water drop generating the rainbow sky**

We seek to determine the angle between the incoming light ray (beam considered to contain all components of visible light) and the outgoing light beam (opposite the Sun: antisolar). Thus, the angle difference for two colors gives us the angle by which we must change our view to observe two different colors in the rainbow sky.

Remark

There is no sense to me to calculate the angle we must look at relatively with the ground (assumed plane) to observe a rainbow as some books mention it. Indeed, anyway if we turn our eyes to a rainbow sky will seen anyway over a large angle from the ground. The only thing that really makes sense, is the angle difference between two "monochromatic" colors.

For this study, we will consider the following approximate figure:

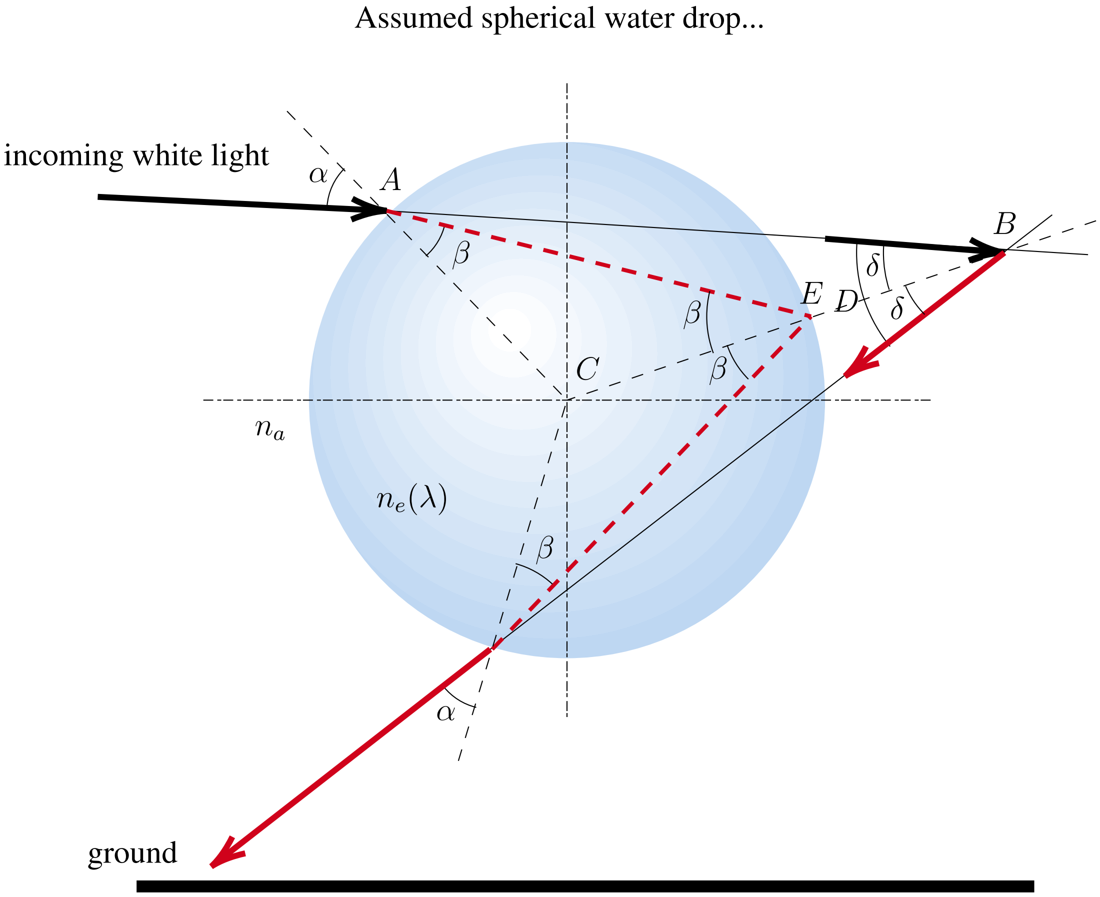

Assumed spherical water drop...
incoming white light
α
A
β
B
δ
E
D
δ
β
C
β
n_a
$n_e(\lambda)$
β
α
ground

Figure 10.152: Path travelled by light in a spherical water drop

With the Snell-Descartes law we have initially:

$$\frac{\sin(\alpha)}{\sin(\beta)}=\frac{n_e}{n_a}=n(\lambda) \quad (10.5.95)$$

What interests us here is the angle of apparent reflection 2δ, which we will denote D (careful! Some teachers choose the convention $D=\pi-2\delta$). To determine this, we start from the following relation of the triangle ABE:

$$(\alpha-\beta)+\delta+(\pi-\beta)=\pi \quad (10.5.96)$$

Therefore it comes:

$$\delta=2\beta-\alpha \quad (10.5.97)$$

Hence:

$$2\delta=4\beta-2\alpha=D \quad (10.5.98)$$

What we will write finally:

$$D=4\beta-2\alpha \quad (10.5.99)$$

and as:

$$\sin(\beta)=\sin(\alpha)\frac{n_a}{n_e}\rightarrow\beta=\arcsin\left(\frac{n_a}{n_e}\sin(\alpha)\right) \quad (10.5.100)$$

Therefore it comes:

$$\boxed{D(n_a,n_e,\alpha)=4\arcsin\left(\frac{n_a}{n_e}\sin(\alpha)\right)-2\alpha} \tag{10.5.101}$$

which is sometimes abusively written:

$$D(n_a,n_e,\alpha)=4\arcsin\left(n\sin(\alpha)\right)-2\alpha \tag{10.5.102}$$

If we make a practical application we have for red 750 [nm] with for example an incidence angle of 30° in the spherical water drop:

$$D(n_a,n_e,\alpha)=D(1.00,1.32,\pi/6)=4\arcsin\left(\frac{1.00}{1.32}\sin\left(\frac{\pi}{6}\right)\right)-2\frac{\pi}{6}\cong 0.506=29.03^\circ \tag{10.5.103}$$

and for purple 400 [nm] with the same angle of incidence of 30° in the spherical water drop:

$$D(n_a,n_e,\alpha)=D(1.00,1.34,\pi/6)=4\arcsin\left(\frac{1.00}{1.34}\sin\left(\frac{\pi}{6}\right)\right)-2\frac{\pi}{6}\cong 0.482=27.63^\circ \tag{10.5.104}$$

Thus an angle difference of about 2.4°.

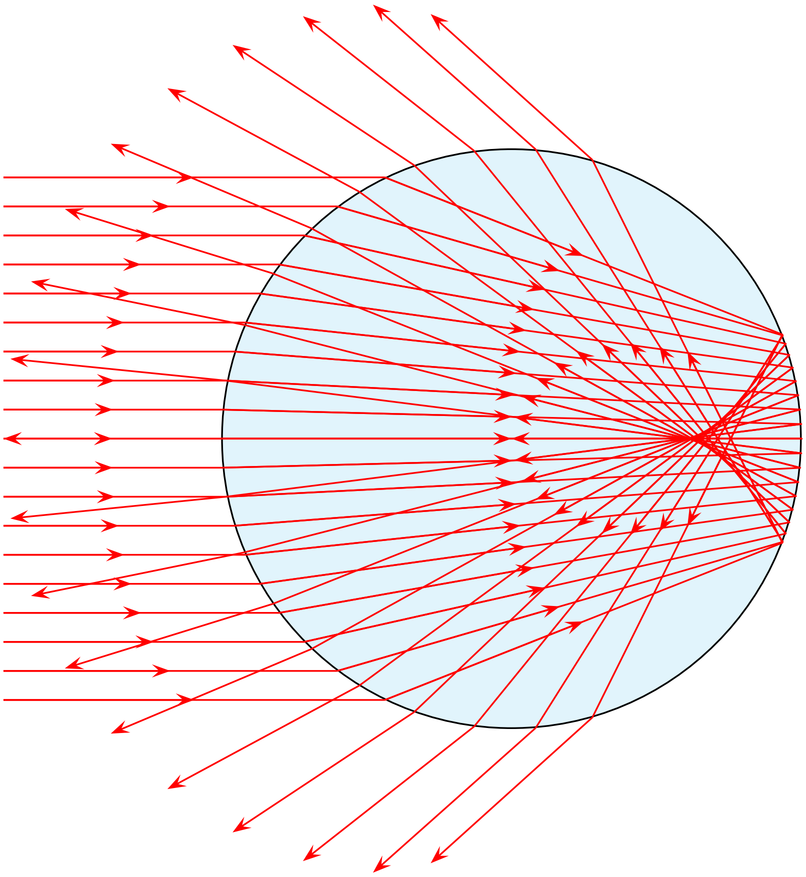
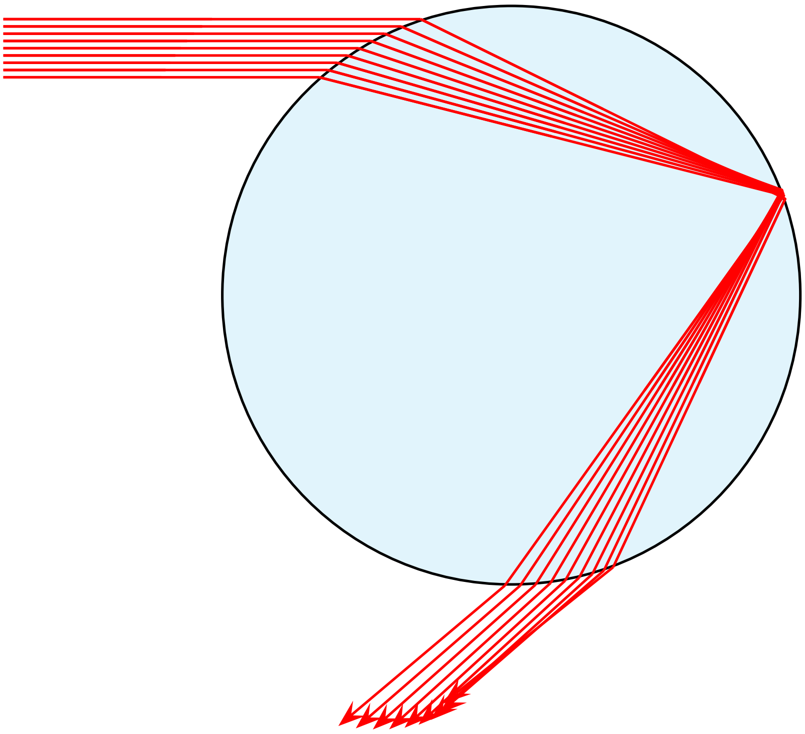


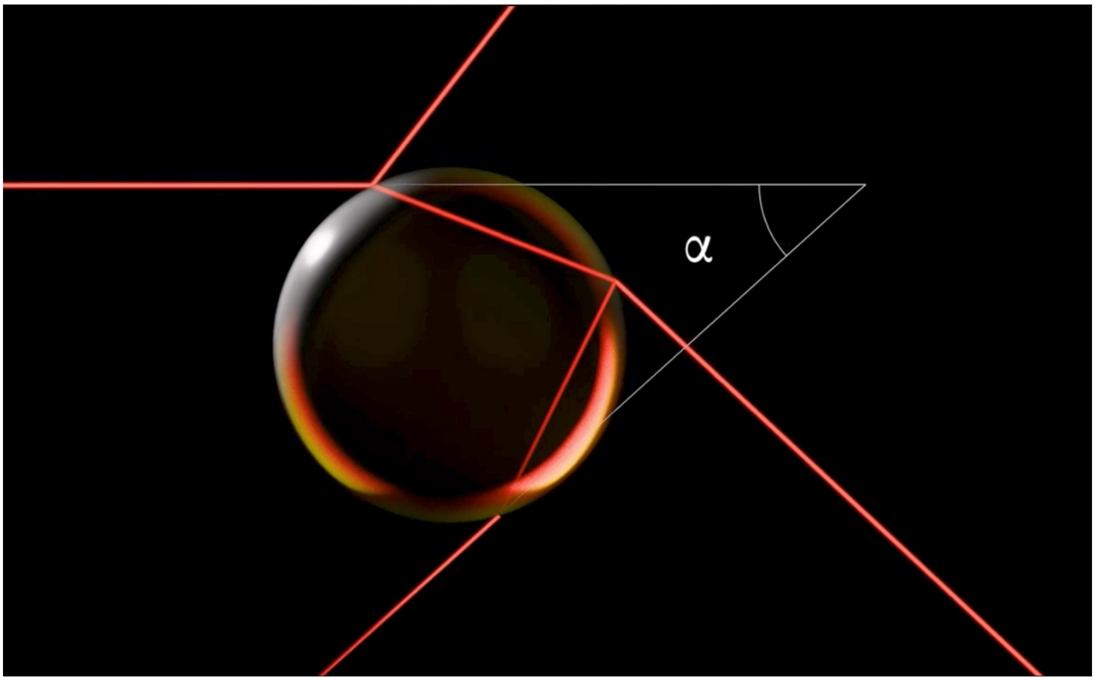
Figure 10.153: **Case of a bundle of parallel rays of identical color in a spherical rain drop**

It is obvious that only the incident rays which strike the drop in its upper hemisphere will reach the eye of the terrestrial observer, after an internal reflection.

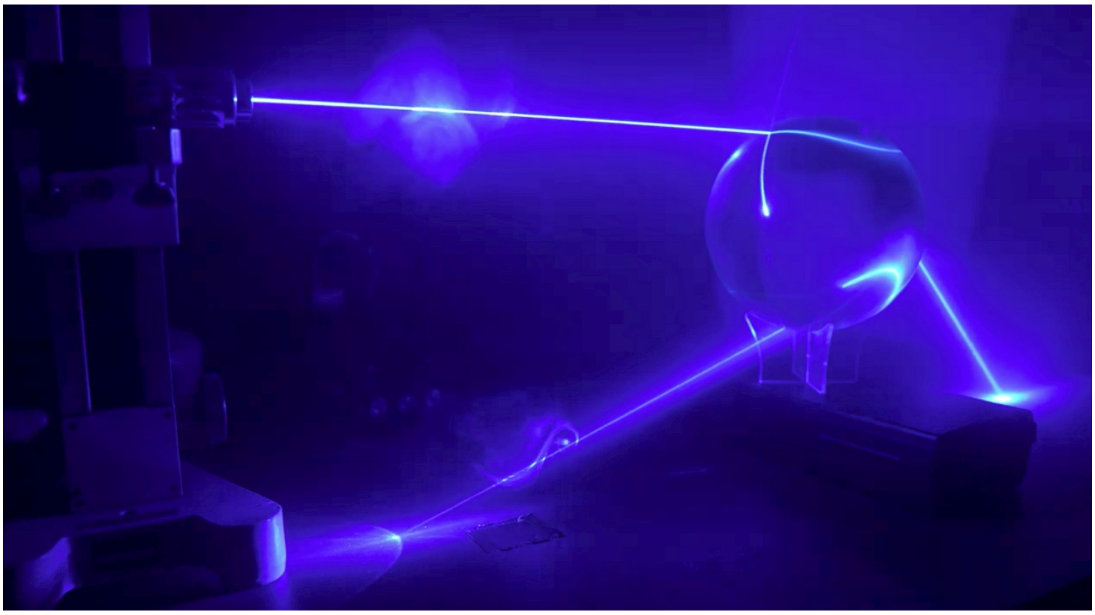
We redo the drawing, limiting ourselves to the upper incident rays:



If we do the illustration by including refracted and reflect rays the reader must not forget that the reality is much closer to the following illustration:



Or in real life with a LASER (whatever the color) and a spherical ball made of glass:



Finally, we could be interested in the angle α at which the angle D is maximum (which corresponds to the angle of the most visible rainbow sky in term of size in reality). We start then of:

$$f(\alpha)=-2\alpha+4\arcsin(n\sin(\alpha)) \tag{10.5.105}$$

and we seek the solutions of:

$$f'(\alpha)=0 \tag{10.5.106}$$

with $\alpha\in[0,\pi/2]$ and $n\in]0,1[$.

Let us recall that we have proved in the section of Differential and Integral Calculus that:

$$\arcsin'(x)=\frac{1}{\sqrt{1-x^2}} \tag{10.5.107}$$

Then we have:

$$f'(\alpha)=-2+\frac{4n\cos(\alpha)}{\sqrt{1-n^2\sin^2(\alpha)}}=0 \tag{10.5.108}$$

Hence:

$$\begin{gathered}\frac{2n\cos(\alpha)}{\sqrt{1-n^2\sin^2(\alpha)}}=1\Leftrightarrow\frac{4n^2\cos^2(\alpha)}{1-n^2\sin^2(\alpha)}=\Leftrightarrow\frac{4n^2\cos^2(\alpha)}{1-n^2(1-\cos^2(\alpha))}=1\\ \Leftrightarrow 4n^2\cos^2(\alpha)=1-n^2+n^2\cos^2(\alpha)\Leftrightarrow\cos^2(\alpha)=\frac{1-n^2}{3n^2}\end{gathered} \tag{10.5.109}$$

From there we get:

$$|\cos(\alpha)|=\frac{\sqrt{1-n^2}}{\sqrt{3}n} \tag{10.5.110}$$

that is to say:

$$\cos(\alpha)=\frac{\sqrt{1-n^2}}{\sqrt{3}n} \tag{10.5.111}$$

as we seek for solutions with $\alpha\in[0,\pi/2]$ and that the cosine is positive on this interval.

Therefore:

$$\alpha_{D_{\max}}=\arccos\left(\frac{\sqrt{1-n^2}}{\sqrt{3}n}\right) \tag{10.5.112}$$

He then comes for $n\cong 0.746$ (purple):

$$\alpha_{D_{\max}}=\arccos\left(\frac{\sqrt{1-n^2}}{\sqrt{3}n}\right)\cong 1.029\cong 60^\circ=\frac{\pi}{3} \tag{10.5.113}$$

This corresponds fairly well to the reality (the angle at which we raise our heads to see the most visible / larger part of the rainbow).

The corresponding deviation is then:

$$D=4\arcsin\left(n\sin(\alpha)\right)-2\alpha\cong 4\arcsin\left(0.746\sin\left(\frac{\pi}{3}\right)\right)-2\frac{\pi}{3}\cong 0.715\cong 40.97^\circ \tag{10.5.114}$$

and is named "rainbow angle" for purple:

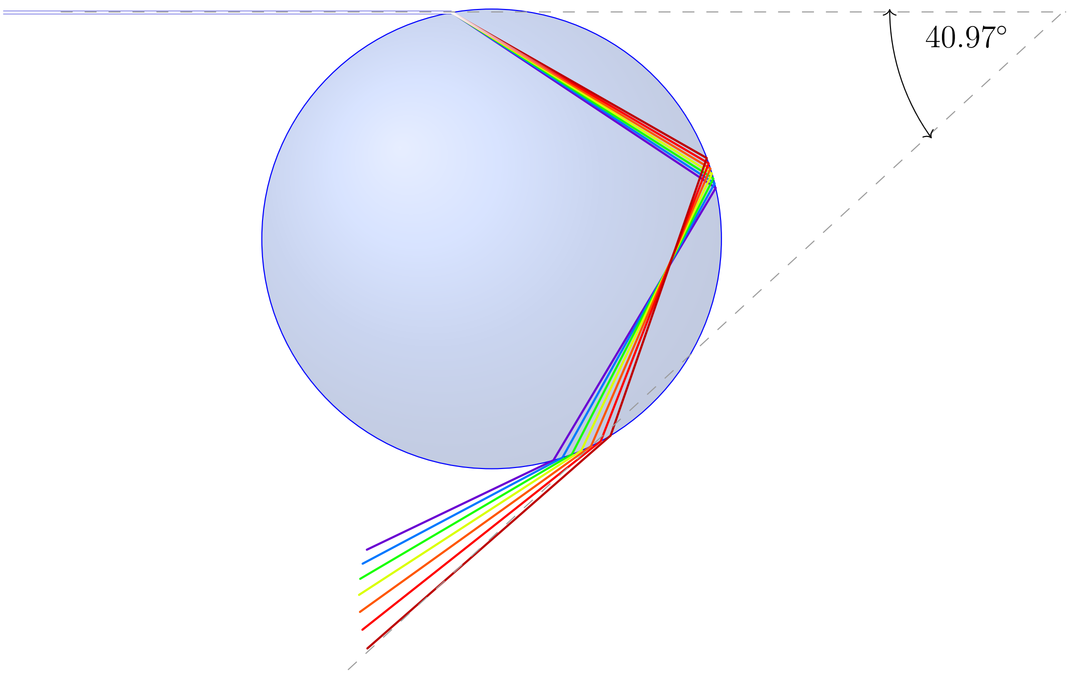
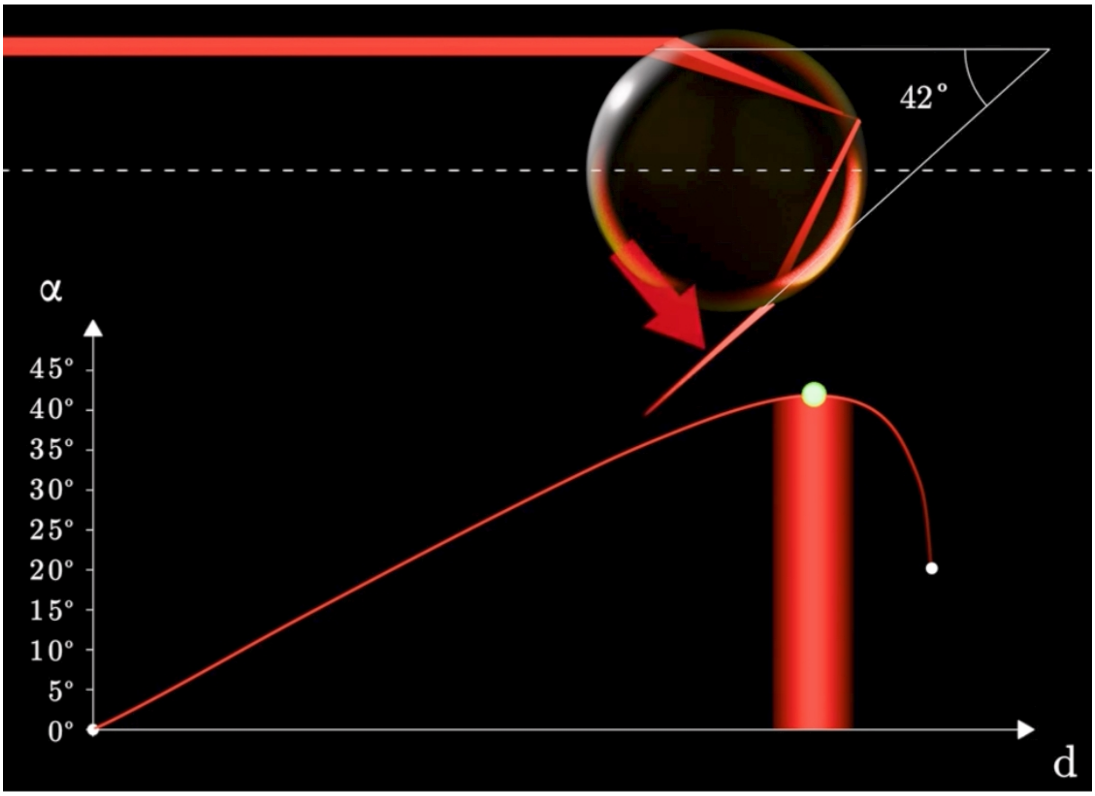


Figure 10.154: **Rainbow angle**

For the blue we get:

$$D=4\arcsin\left(n\sin(\alpha)\right)-2\alpha\cong 4\arcsin\left(0.757\sin\left(\frac{\pi}{3}\right)\right)-2\frac{\pi}{3}\cong 44^\circ \tag{10.5.115}$$

So the average gives the famous 42° that we can read in many textbooks (average corresponding to approximately to the red color).



The reader must keep in mind that we have proved the Cauchy formula (see page 4244) that gives:

$$n \approx n_0 + \frac{C}{\lambda_0^2} \qquad (10.5.116)$$

with for water $n_0 = 1.3246$ and $C = 3092$, if the wavelength of λ_0 is expresse in nanometers. For the two extreme colors of the spectrum we have for $\lambda = 800$ [nm] (red) the value $n = 1.332$ and for $\lambda = 400$ [nm] (purple) the value $n = 1.345$.

Each color will therefore be deflected differently and this is what gives the phenomenon of the rainbow as depicted in the figure above. The angle between the extreme colors is very small, but the observer receiving this coloured beam being situated several kilometres away will see an arc of greater thickness the farther he is from it.

Thus, the light rays received by the observer and in which the red (outer edge of the rainbow) dominates correspond to all the rays coming from the wall of rain and making angle of about 40° with the direction of the sunlight (see figure below). The light rays forming each color of the rainbow sky then form the top of the cone observer's eyes and for axis the solar ray passing through the eyes of the observer:

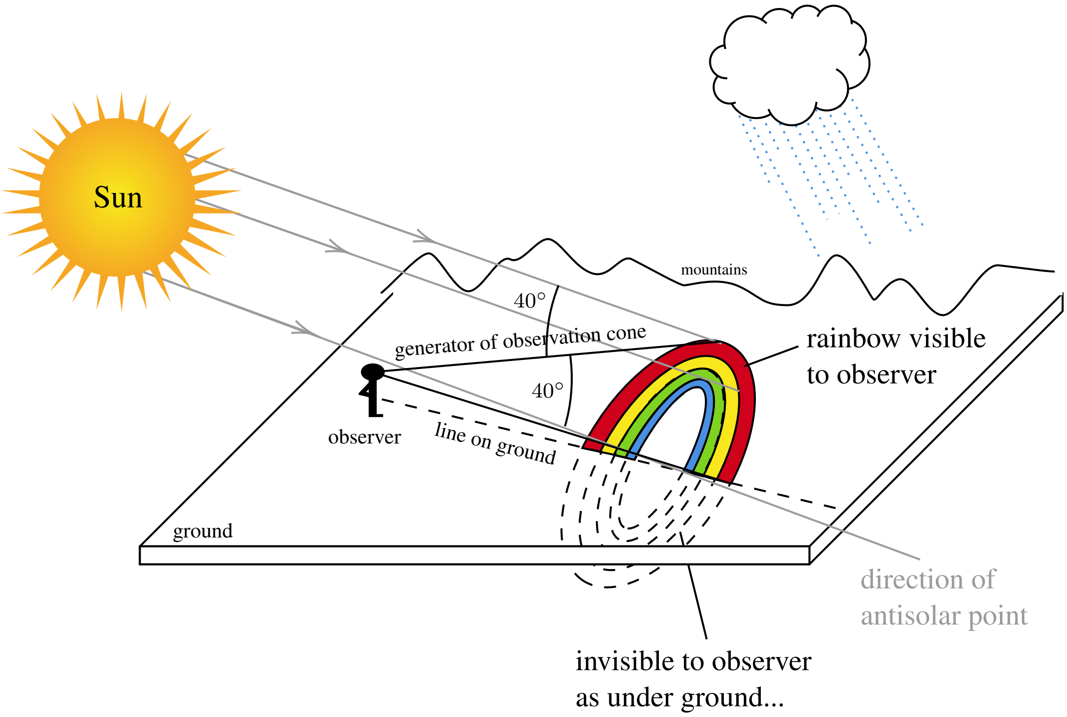

Sun
mountains
40°
generator of observation cone
40°
observer
line on ground
rainbow visible
to observer
ground
direction of
antisolar point
invisible to observer
as under ground...

Figure 10.155: Naive illustration of the generation of the primary rainbow (objects are not to scale!)

That gives in real life:

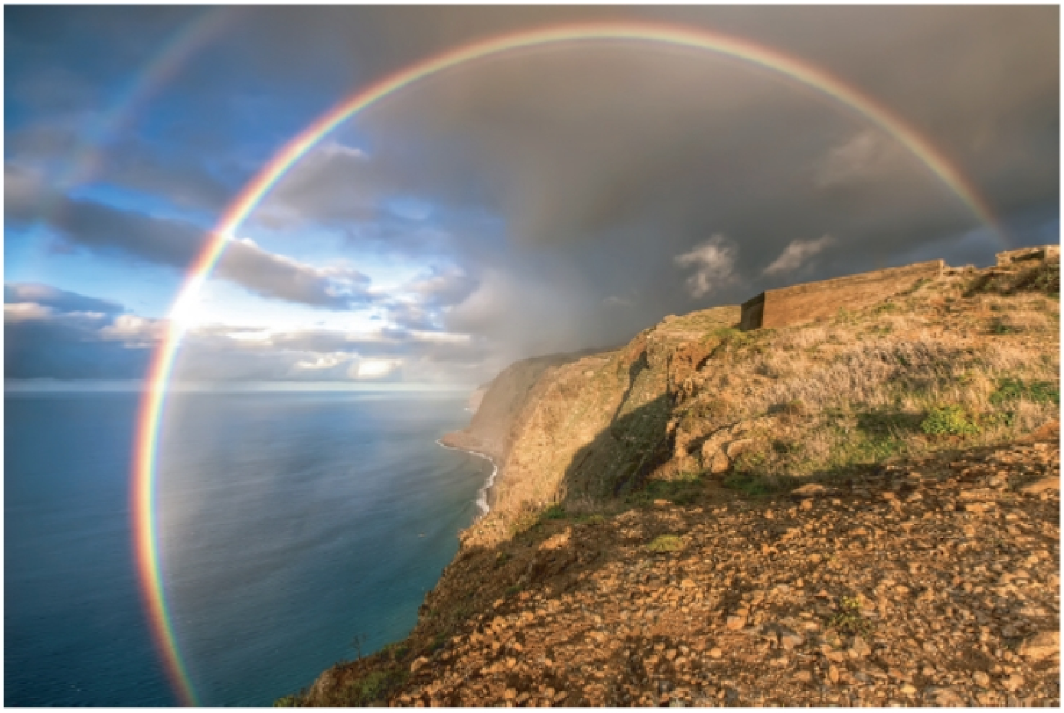


Figure 10.156: Full-circle rainbow (source: ed g2s, Christophe Afonso)

To observe a secondary rainbow (see the photo above in the upper left corner), it is necessary to consider the rays which strike the lower hemisphere of the drop of water and which after two successive reflections leave again downwards, towards the possible observers:

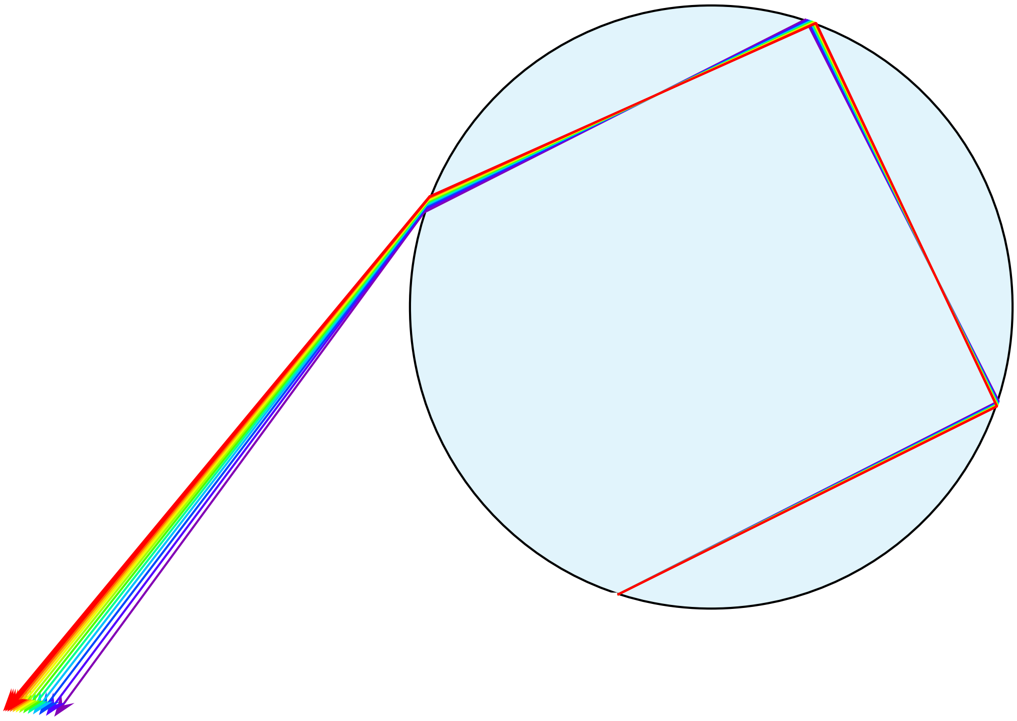


Figure 10.157: **Double rainbow mechanism**

We can see in the figure above that the order of colors is reversed (as we can also see it in the previous photo) and the spectrum is wider. However, due to the additional reflection, the secondary rainbow is much less luminous!

Remark

Even if in reality the rainbow has almost an infinity of colors a lot of people like to associate to the latter seven colors that can be remember with the following mnemonic technique: "*Richard (red) Of (orange) York (yellow) Gave (green) Battle (blue) In (indigo) Vain (violet)*"

The Sun rainbow is easy to reproduce experimentally at the macroscopic level as illustrated in the well-known case below in a photo:

However we have not been able to find a laboratory photo to this date (in a scientific paper) of a single drop of water which scatters the light into a rainbow (if anyone knows of such a photo, please do not hesitate to contact us!).

A "moonbow" (also known as a "moon rainbow" or "lunar rainbow") is a rainbow produced by moonlight rather than direct sunlight. Other than the difference in the light source, its formation is the same as for a solar rainbow: It is caused by the refraction of light in many water droplets, such as a rain shower or a waterfall, and is always positioned in the opposite part of the sky from the Moon relative to the observer.

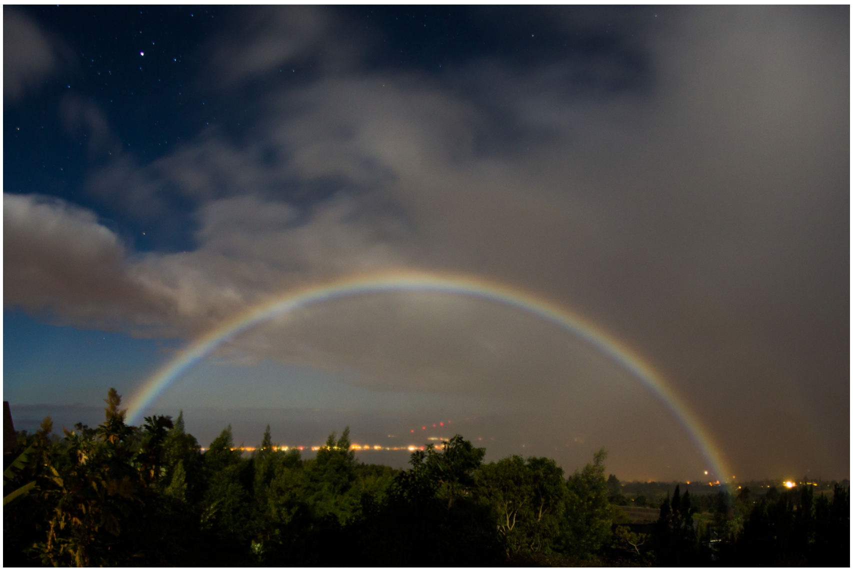


Figure 10.158: Moonbow over the town of Kihei, seen from Kula, on Maui, Hawaii, US on the 28th February 12016 (source: Wikipedia, author: Arne-kaiser)

A "Brocken spectre", well known and badly missinterpreted by religious people, is the magnified (and apparently enormous) shadow of an observer (individual, airplane, bird, crowd, etc.) cast upon clouds opposite the Sun's direction. The figure's head is often surrounded by the halo-rainbow like rings of coloured light forming a glory, which appears opposite to the Sun's direction when uniformly-sized water droplets in clouds refract and backscatter sunlight:

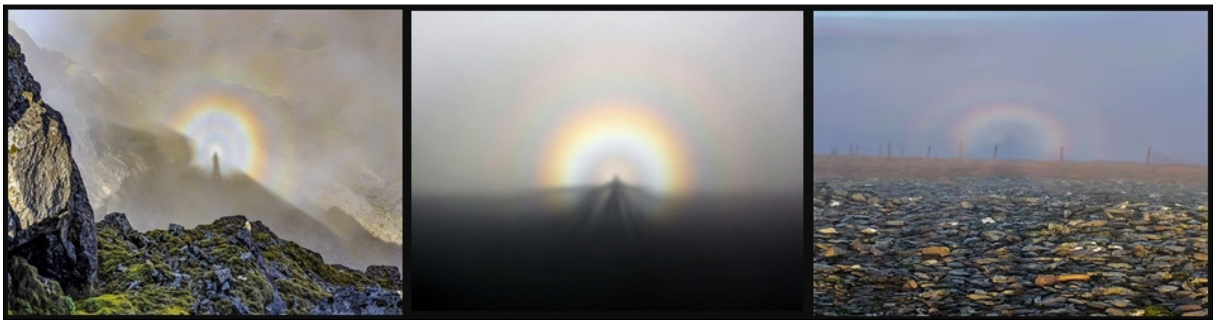


Figure 10.159: **Different photos of Brocken spectrums**

10.5.4.4 The 22° Halo

Under certain atmospheric conditions, a number of different halo phenomena occur, producing rings, arcs, columns, crosses or false suns from reflected or refracted sunlight. These are distinct from other atmospheric optical phenomena such as rainbows in that the surfaces of reflection and refraction are those of ice crystals, and not of liquid water droplets! This immediately suggests an explanation for the plethora of halos as compared to the limited number and form of bows and coronas, since the shape of ice crystals is much more varied than that of water droplets. Their reduced symmetry means that a given ice crystal can produce different halos depending upon orientation, and that a given halo can change shape as the Sun (actually a G2V-type star, aka yellow dwarf) changes height.

A "22° halo" is an atmospheric optical phenomenon that consists of a halo with an apparent radius of approximately 22° around the Sun or Moon. Around the Sun, it may also be named a "Sun halo". Around the Moon, it is also known as a "moon ring", "storm ring", or "winter halo".

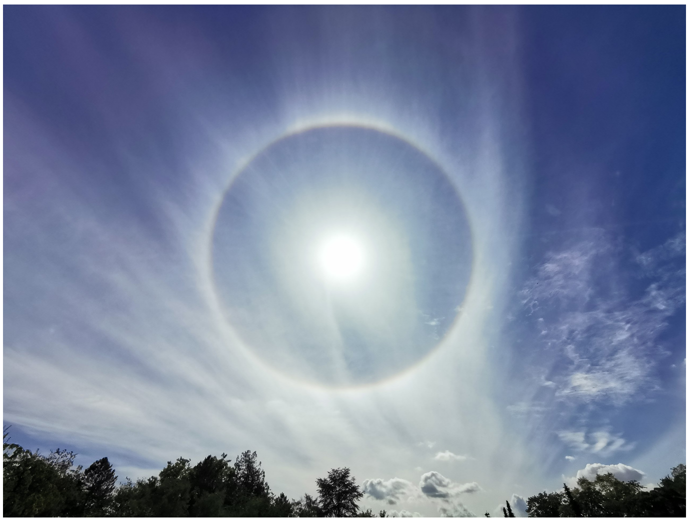


Figure 10.160: 22° **Sun Halo**

It forms as sunlight or moonlight is refracted by millions of hexagonal ice crystals (60° faces) suspended in the atmosphere in thin cirro-stratus clouds:

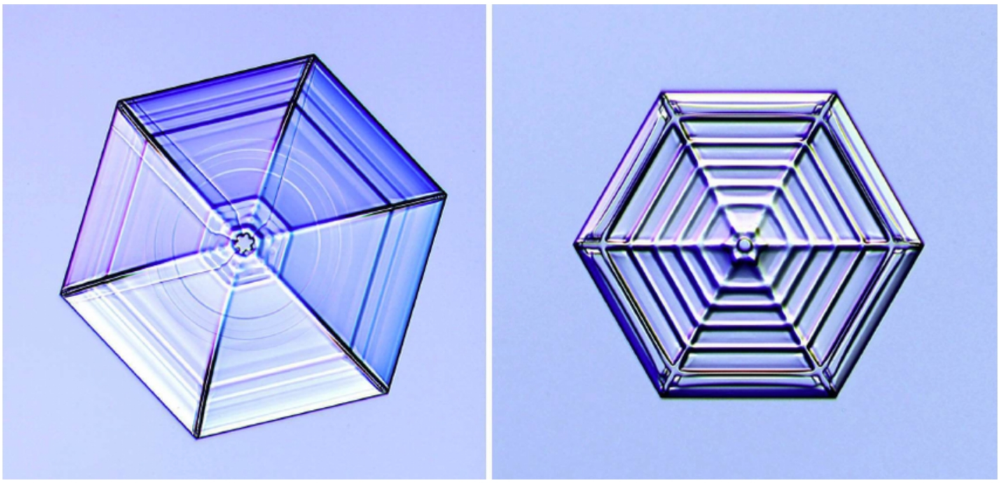


Figure 10.161: Photographs showing two hexagonal POP ice crystals with several commonly seen morphological

Atmospheric ice crystals form as hexagonal prisms, referred to as "plates" if the axis is much shorter than the diameter, and as "needles" if it is much longer. The ends of the prisms can be either flat, or have pyramidal caps.

Remark

Even though it is one of the most common types of halo, the shape and orientation of the ice crystals responsible for the 22° halo are the topic of debate. Hexagonal, randomly oriented columns are usually put forward as the most likely candidate, but this explanation presents problems, such as the fact that the aerodynamic properties of such crystals leads them to be oriented horizontally rather than randomly. Alternative explanations include the involvement of clusters of bullet-shaped ice columns.

As light passes through the 60° apex angle of the hexagonal ice prisms, it is deflected twice, resulting in deviation angles ranging from 22° to 50°. Given the angle of incidence onto the hexagonal ice prism $\theta_{\text{incidence}}$ and the refractive index inside the prism n, then the angle of deviation $\theta_{\text{deviation}}$ can be derived from Snell's law:

$$\theta_{\text{deviation}}=\theta_{\text{incidence}}+\sin^{-1}\left[n\sin\left(\dfrac{\pi}{3}-\sin^{-1}\dfrac{\sin(\theta_{\text{incidence}})}{n}\right)\right]-\dfrac{\pi}{3} \tag{10.5.117}$$

The ice crystals in the clouds all deviate the light similarly, but only the ones from the specific ring at 22° contribute to the effect for an observer at a set distance. As no light is refracted at angles smaller than 22°, the sky is darker inside the halo.

Let us prove the relation above! First let us remind that all the interior angles of a hexagon are equal to $\pi/3$ (see section Geometric Shapes page 2909):

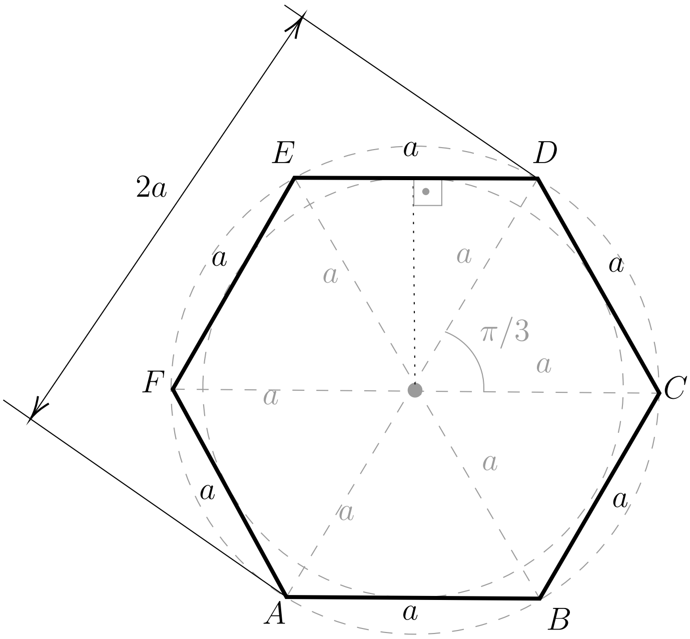
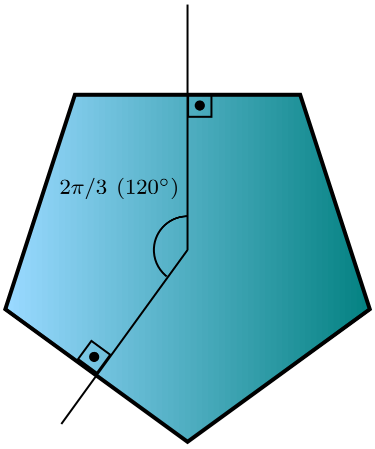
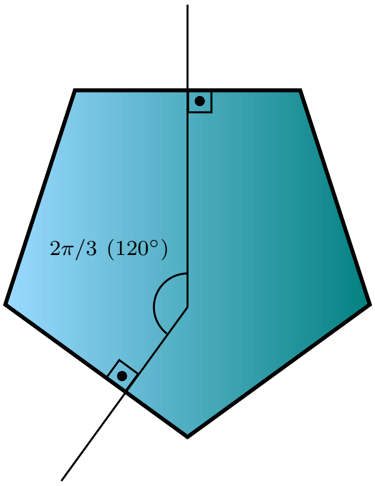

E
a
D
2a
a
a
a
a
π/3
F
a
a
C
a
a
a
a
A
a
B

Figure 10.162: **Regular hexagon details (source: Wikipedia)**

Therefore it's quite obvious that we have:



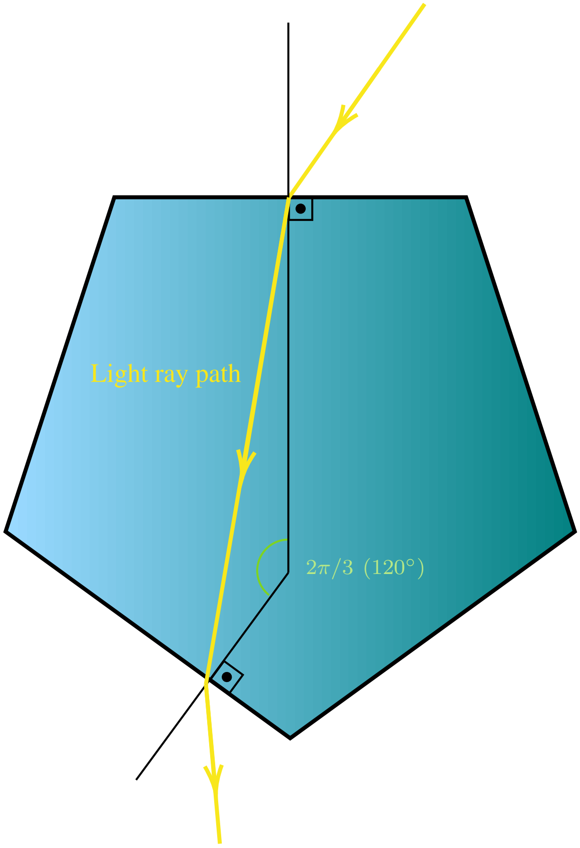
Notice that the angle visible above remain the same if we translate the hexagon as following:



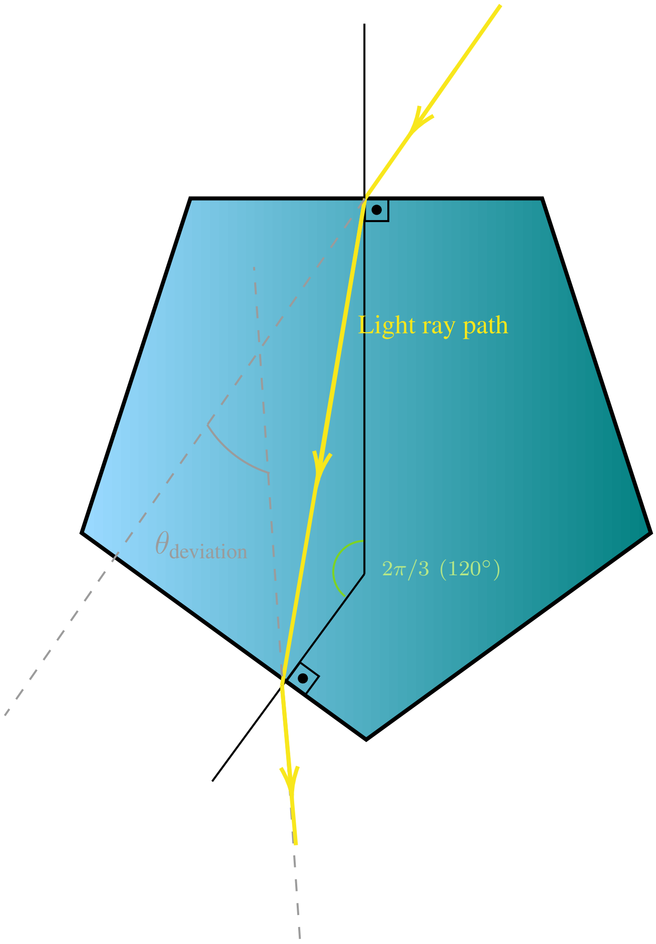
Remark

In atmospheric optics, a "photometeor" is a bright object or other optical phenomenon appearing in the Earth's atmosphere when sunlight or moonlight creates a reflection, refraction, diffraction or interference under particular circumstances. The most common examples include halos, rainbows, fogbows, cloud iridescences (or irisation), glories, Bishop's rings, coronas, crepuscular rays, sun dogs, light pillars, mirages, scintillations, and green flashes.

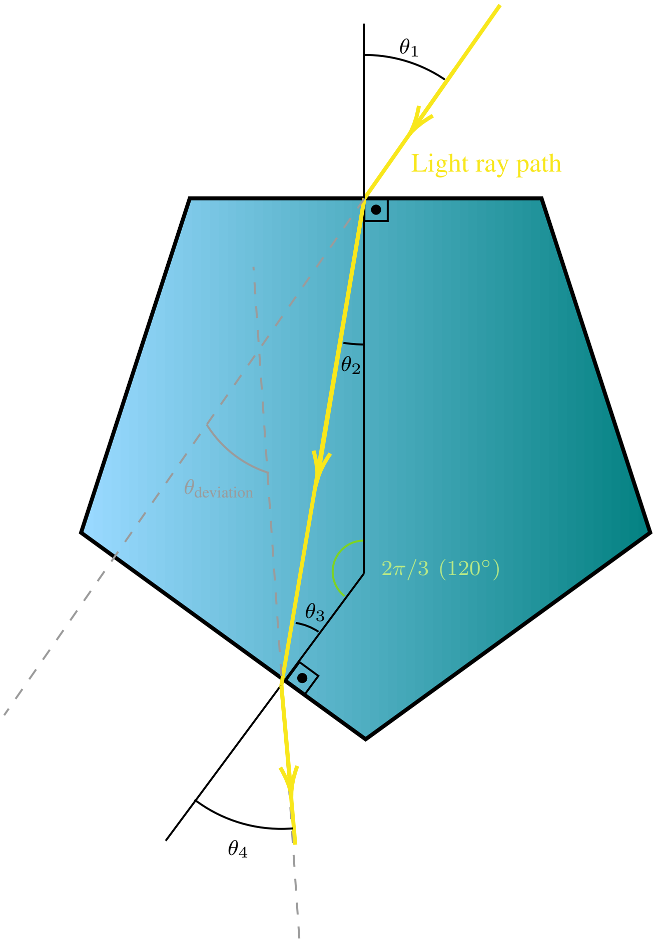
Now consider a ray path as illustrated below in the special case that lights when entering the ice crystal makes only two refractions:



What is of interest for us is the $\theta_{\text{deviation}}$ visible in gray below:


Light ray path
$\theta_{\text{deviation}}$
$2\pi/3$ (120°)

To determine $\theta_{\text{deviation}}$ let us complete the figure above as following:



From Snell's law we have first trivially:

$$\frac{\sin(\theta_1)}{\sin(\theta_2)}=n \qquad \frac{\sin(\theta_4)}{\sin(\theta_3)}=n \tag{10.5.118}$$

Secondly we have we have (still trivially) for the central triangle, applying the Triangle Sum Theorem:

$$120+\theta_2+\theta_3=180 \tag{10.5.119}$$

And finally (still by looking the figure it should be quite trivial):

$$\theta_{\text{deviation}}=(\theta_1-\theta_2)+(\theta_4-\theta_3) \tag{10.5.120}$$

Now we will mix together these three relations. We start first by rearranging them (easy work!):

$$\begin{gathered}\theta_2=\sin^{-1}\left(\frac{\sin(\theta_1)}{n}\right)\\ \theta_3=60-\theta_2=\frac{\pi}{3}-\theta_2\\ \theta_4=\sin^{-1}\left(n\sin(\theta_3)\right)\\ \theta_{\text{deviation}}=\theta_1+\theta_4-60=\theta_1+\theta_4-\frac{\pi}{3}\end{gathered} \tag{10.5.121}$$

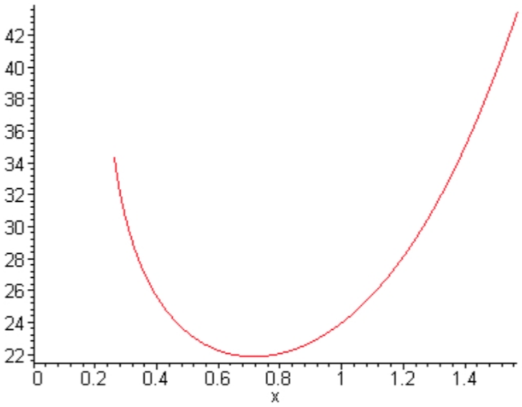
Therefore:

$$\begin{aligned}\theta_{\text{deviation}} &= \theta_1 + \theta_4 - \frac{\pi}{3}\\ &= \theta_1 - \frac{\pi}{3} + \sin^{-1}\left(n\sin\left(\frac{\pi}{3} - \sin^{-1}\left(\frac{\sin(\theta_1)}{n}\right)\right)\right)\\ &= \theta_1 + \sin^{-1}\left(n\sin\left(\frac{\pi}{3} - \sin^{-1}\left(\frac{\sin(\theta_1)}{n}\right)\right)\right) - \frac{\pi}{3}\end{aligned} \tag{10.5.122}$$

Hence we fall back on relation presented at the beginning:

$$\boxed{\theta_{\text{deviation}} = \theta_{\text{incidence}} + \sin^{-1}\left[n\sin\left(\frac{\pi}{3} - \sin^{-1}\frac{\sin(\theta_{\text{incidence}})}{n}\right)\right] - \frac{\pi}{3}} \tag{10.5.123}$$

For $n = 1.309$, the angle of minimum deviation is almost $22°$ ($21.76°$, when $\theta_{\text{incidence}} = 40.88° \cong 0.71$ [rad]):



More specifically, the angle of minimum deviation is $21.84°$ on average ($n = 1.31$); $21.54°$ for red light ($n = 1.306$) and $22.37°$ for blue light ($n = 1.317$). This wavelength-dependent variation in refraction causes the inner edge of the circle to be reddish while the outer edge is bluish (see photo below).

Sadly as far as we know there are no closed formal solution to find the optimum analytically of:

$$\frac{d\theta_{\text{deviation}}}{d\theta_{\text{incidence}}} = 0 \tag{10.5.124}$$

Even if we can get (we don't provide the details here):

$$\frac{d\theta_{\text{deviation}}}{d\theta_{\text{incidence}}} = 1 - \frac{\cos(\theta_{\text{incidence}})\cos\left(\arcsin\left(\frac{\sin(\theta_{\text{incidence}})}{n}\right) - \frac{\pi}{3}\right)}{\sqrt{1 - \frac{\sin^2(\theta_{\text{incidence}})}{n^2}}\sqrt{1 - n^2\sin^2\left(\arcsin\left(\frac{\sin(\theta_{\text{incidence}})}{n}\right) - \frac{\pi}{3}\right)}} \tag{10.5.125}$$

The problem is to solve when we put the latter equal to zero...

The figure below depicts most common Halos in the sky, for a solar elevation of $26°$. This view is looking straight up toward the zenith, as what could be photographed using an all-sky camera (with a fisheye lens). The outermost thick green circle represents the horizon, and the zenith point is marked as a dot in the center of this circle. Some of these halos are extremely rare. Most would not be seen at the same time, because they require ice crystals of different shapes and having perfect orientation. This fish-eye view distorts the 9°, 22°, and 46° halos, which would appear as circles to the observer. The halos of other diameters (including the 28° Scheiner's halo) are not plotted.

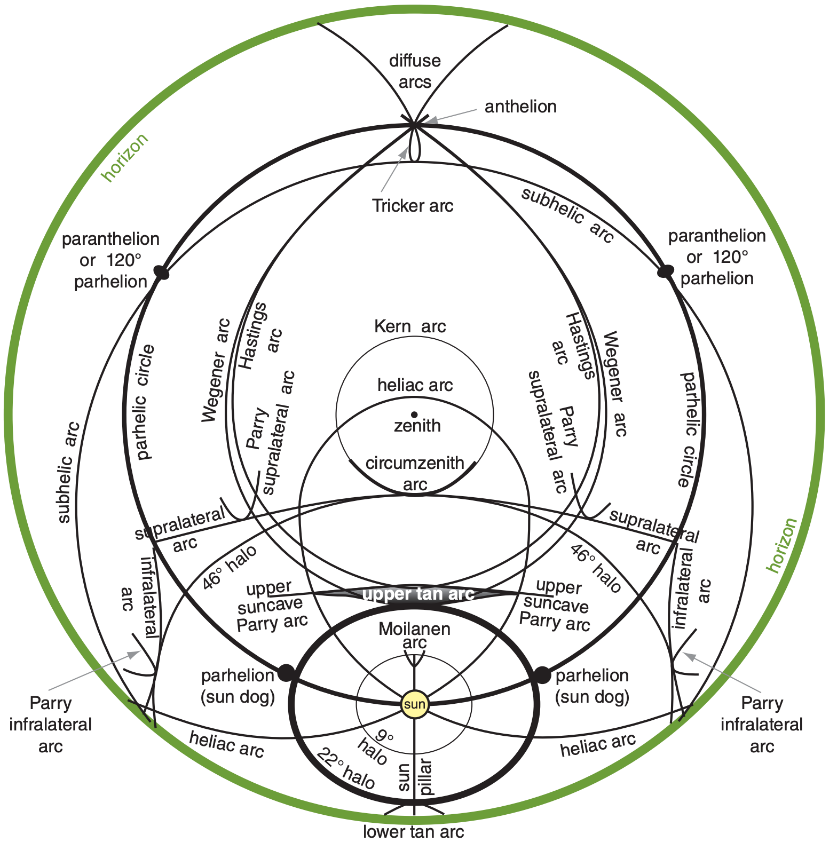

diffuse arcs
anthelion
horizon
Tricker arc
subhelic arc
paranthelion or 120° parhelion
paranthelion or 120° parhelion
Kern arc
Wegener arc
Hastings arc
Hastings arc
Wegener arc
heliac arc
zenith
Parry supralateral arc
Parry supralateral arc
parhelic circle
parhelic circle
subhelic arc
circumzenith arc
horizon
supralateral arc
supralateral arc
infralateral arc
infralateral arc
46° halo
46° halo
upper suncave Parry arc
upper tan arc
upper suncave Parry arc
Moilanen arc
parhelion (sun dog)
parhelion (sun dog)
Parry infralateral arc
Parry infralateral arc
sun
heliac arc
heliac arc
9° halo
22° halo
sun pillar
lower tan arc

Figure 10.163: **Most common types of Sun Halos and Arcs**

Halos constitute a vast subject of atmospheric optics, coupled with meteorology and ice crystallography. The variety of halos and their properties depend on the geometry of the crystals and their degree of orientation in space. Some remain of uncertain or even unknown explanation; this is the case, for example, of elliptical halos and Moilanen's "V"-shaped arc which lies halfway between the Sun and the upper part of the halo at 22°. What information about crystals do these halos give us? Conversely, choosing a geometry, what halos can we form? Finally, what about on other planets? Numerical simulation and observation are essential to answer these questions. Hopefully this article on some aspects of terrestrial halos will have aroused the reader's curiosity about these beautiful photometeors and will encourage them to look up to the sky. May everyone be a careful observer and sharpen their physical mind. Because we are all potential halo discoverers!

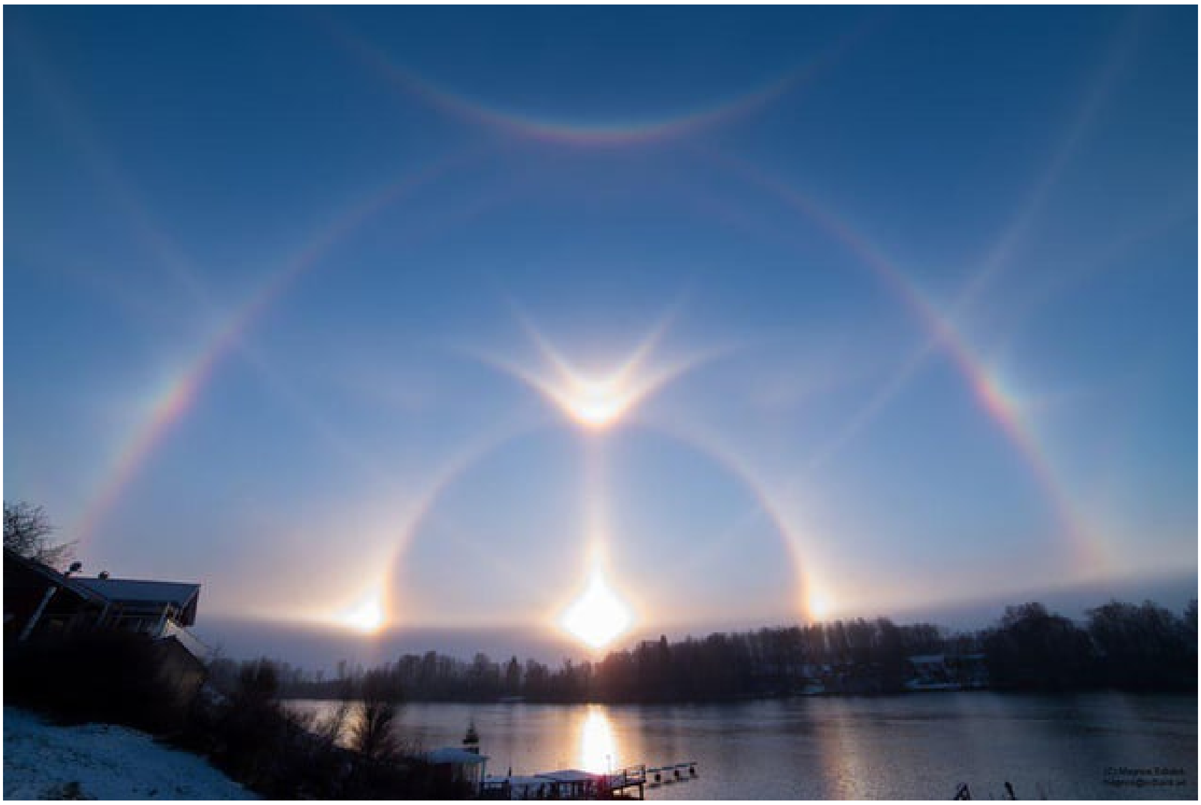
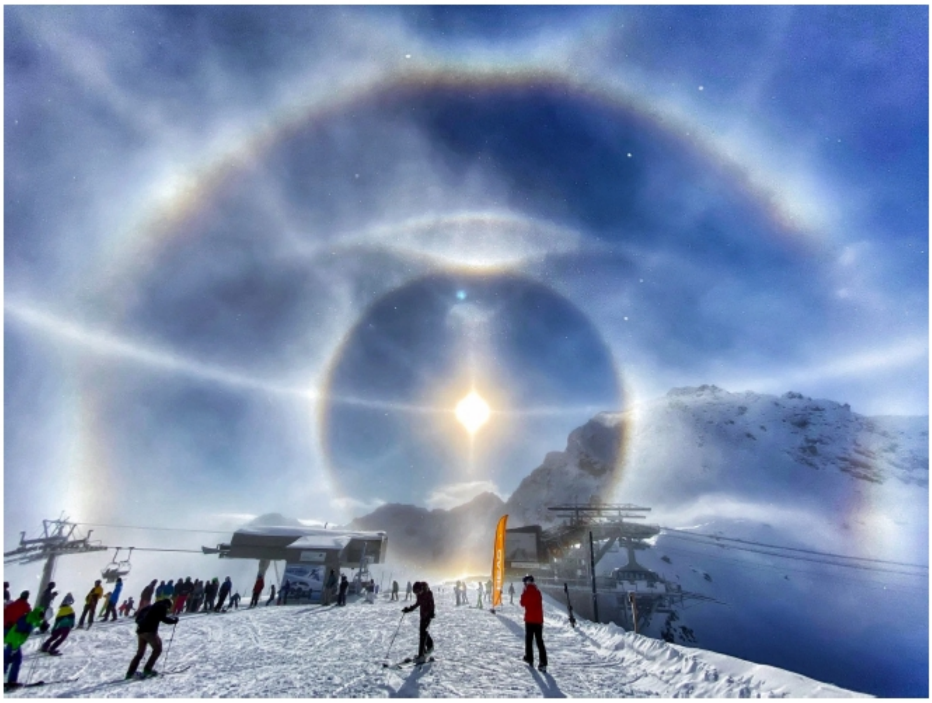


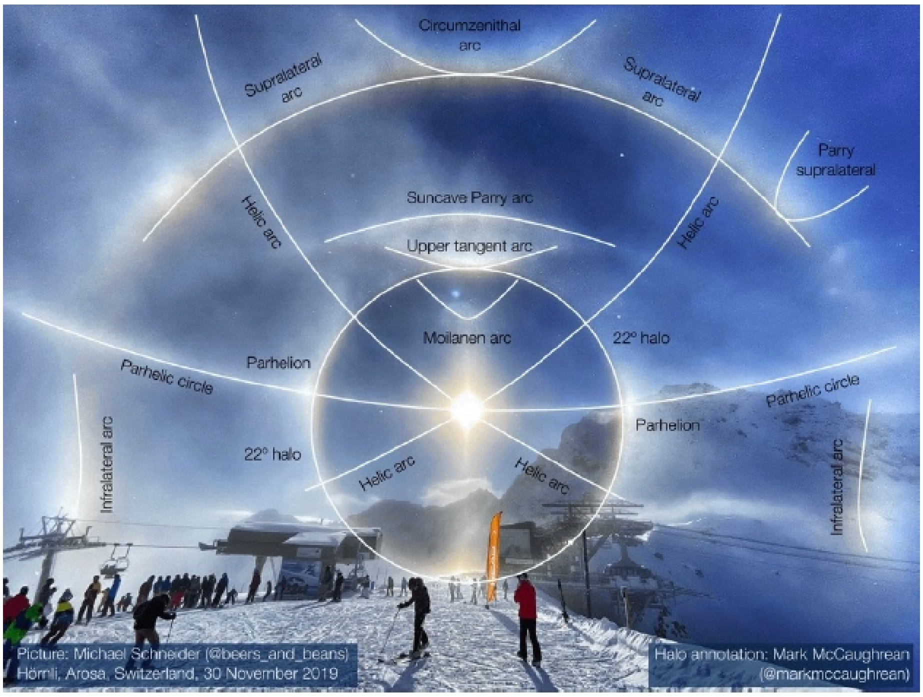
Figure 10.164: **Various types of Halos in one shot in real life**

For people who wants more detailed and general derivations but also laboratory tests we recommend the lecture (in increasing order of complexity) of [140], [107] and [493].

This stunning picture taken by Michael Schneider in Switzerland shows a multitude of other optical phenomena. Note that the 22° halo crosses in front of the mountains behind. It follows that these effects were due to ice crystals between the photographer and the mountains. In fact, they are probably due to minute crystals in the air immediately surrounding the photographer:



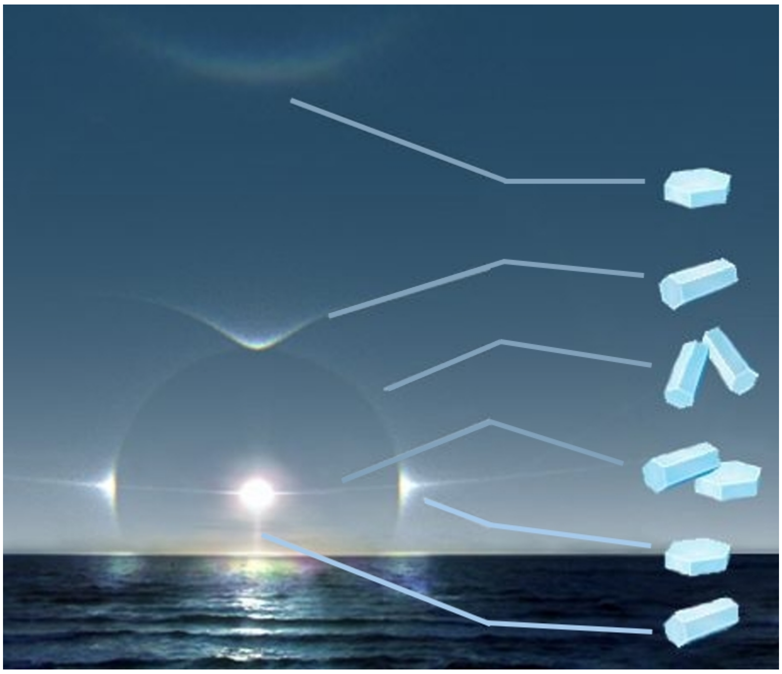
A key to all the effects is shown below:


Circumzenithal arc
Supralateral arc
Supralateral arc
Parry supralateral
Suncave Parry arc
Helic arc
Helic arc
Upper tangent arc
Moilanen arc
22° halo
Parhelion
Parhelic circle
Parhelic circle
Parhelion
Infralateral arc
22° halo
Helic arc
Helic arc
Infralateral arc

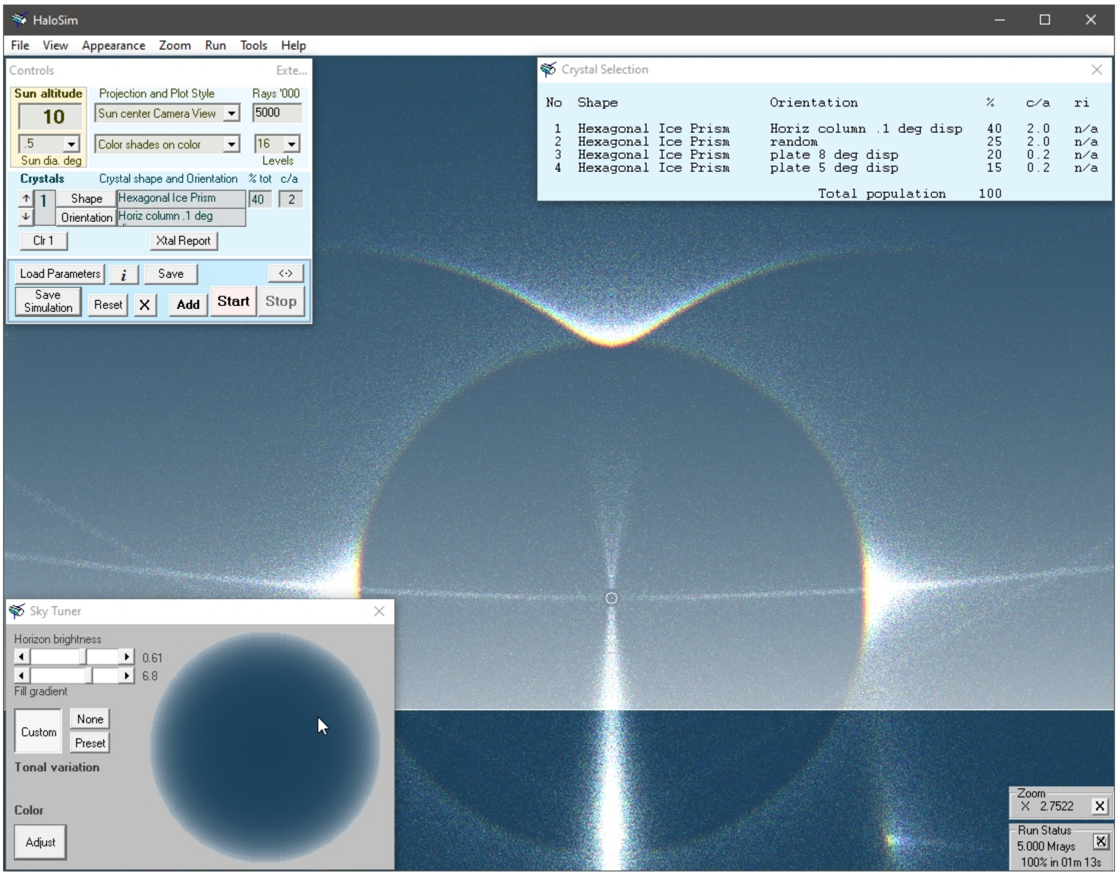
Picture: Michael Schneider (@beers_and_beans)
Hörnli, Arosa, Switzerland, 30 November 2019

Halo annotation: Mark McCaughrean
(@markmccaughrean)

It is supposed (because as far as we know such atmospheric optical structure has never been reproduced yet in a laboratory controlled environment) that the different curves are made by different crystals in different orientations as illustrated below:



Even if such a thing was never reproduced in laboratory and is much harder to reproduce that a rainbow for any human there are specialized university free softwares that implement millions (or billions) ray tracing with the possibility to configure various crystal mixtures, densities and orientations like HaloSim 3.6 developed by Les Cowley and Michael Schroeder. Let us see a few screenshots of HaloSim 3.6 with various settings:


HaloSim
File View Appearance Zoom Run Tools Help
Controls
Exte...
Clr 1
Xtal Report
Load Parameters
i
Save
<->
Save Simulation
Reset
X
Add
Start
Stop
Sky Tuner
Custom
None
Preset
Tonal variation
Color
Adjust

Sun altitude: 10
Sun dia. deg: .5
Projection and Plot Style: Sun center Camera View
Color shades on color
Rays '000: 5000
Levels: 16
Crystals: 1
Crystal shape and Orientation: Shape Hexagonal Ice Prism; Orientation Horiz column .1 deg
% tot: 40
c/a: 2

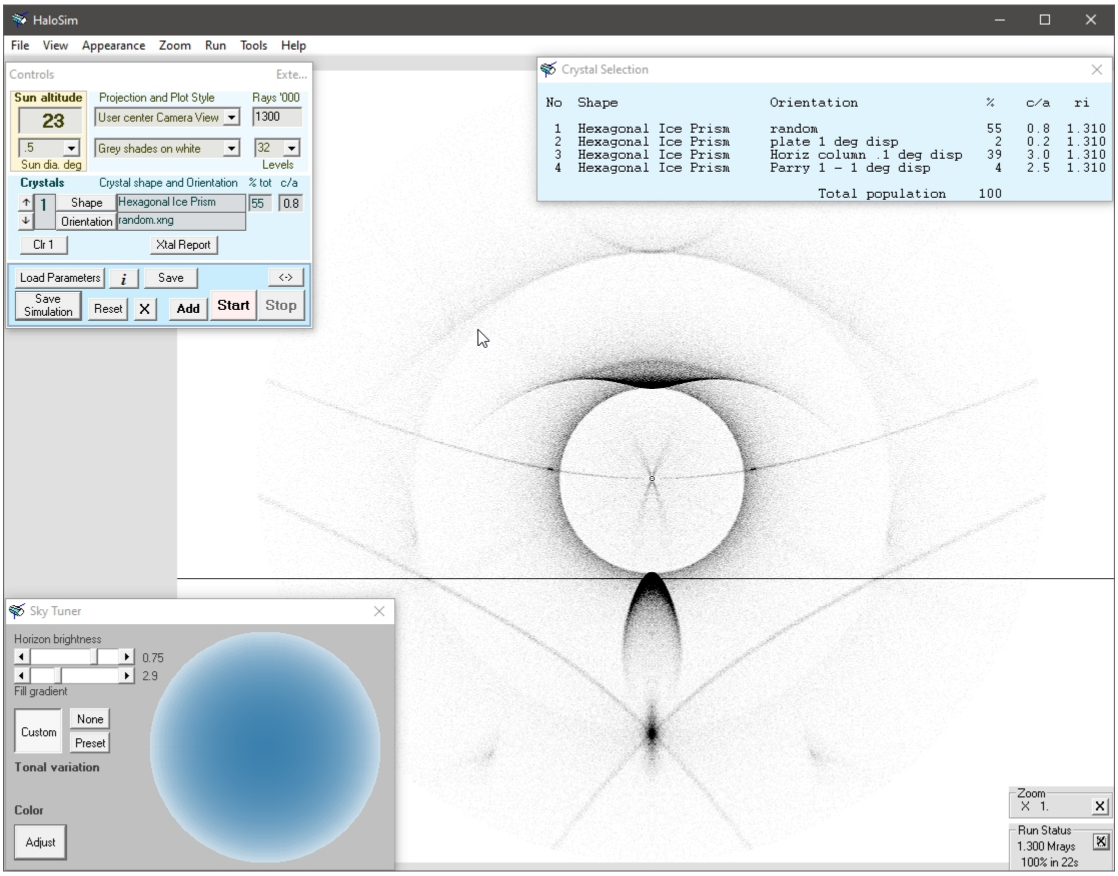
Crystal Selection

```
No  Shape                  Orientation             %    c/a   ri

 1  Hexagonal Ice Prism    Horiz column .1 deg disp 40   2.0   n/a
 2  Hexagonal Ice Prism    random                   25   2.0   n/a
 3  Hexagonal Ice Prism    plate 8 deg disp         20   0.2   n/a
 4  Hexagonal Ice Prism    plate 5 deg disp         15   0.2   n/a

                              Total population     100
```

Horizon brightness: 0.61
Fill gradient: 6.8

Zoom: X 2.7522

Run Status: 5.000 Mrays
100% in 01m 13s


HaloSim
File
View
Appearance
Zoom
Run
Tools
Help
Controls
Exte...
Clr 1
Xtal Report
Load Parameters
i
Save
<·>
Save Simulation
Reset
X
Add
Start
Stop
Sky Tuner
Fill gradient
Custom
None
Preset
Tonal variation
Color
Adjust
Zoom

Sun altitude 23
Sun dia. deg .5
Projection and Plot Style: User center Camera View
Grey shades on white
Rays '000 1300
Levels 32
Crystals — Crystal shape and Orientation — % tot — c/a
1 Shape: Hexagonal Ice Prism — 55 — 0.8
Orientation: random.xng

Crystal Selection

```
No  Shape                   Orientation              %    c/a   ri

 1  Hexagonal Ice Prism     random                  55    0.8  1.310
 2  Hexagonal Ice Prism     plate 1 deg disp         2    0.2  1.310
 3  Hexagonal Ice Prism     Horiz column .1 deg disp 39   3.0  1.310
 4  Hexagonal Ice Prism     Parry 1 - 1 deg disp     4    2.5  1.310

                               Total population    100
```

Horizon brightness 0.75
2.9
Zoom X 1.
Run Status 1.300 Mrays 100% in 22s

Let us mention also "cloud iridescence" or "irisation" that is a colorful optical phenomenon that occurs in a cloud (that's why it is also named "rainbow cloud") and appears in the general proximity of the Sun or Moon produce by the same mechanisms of the Sun halos.

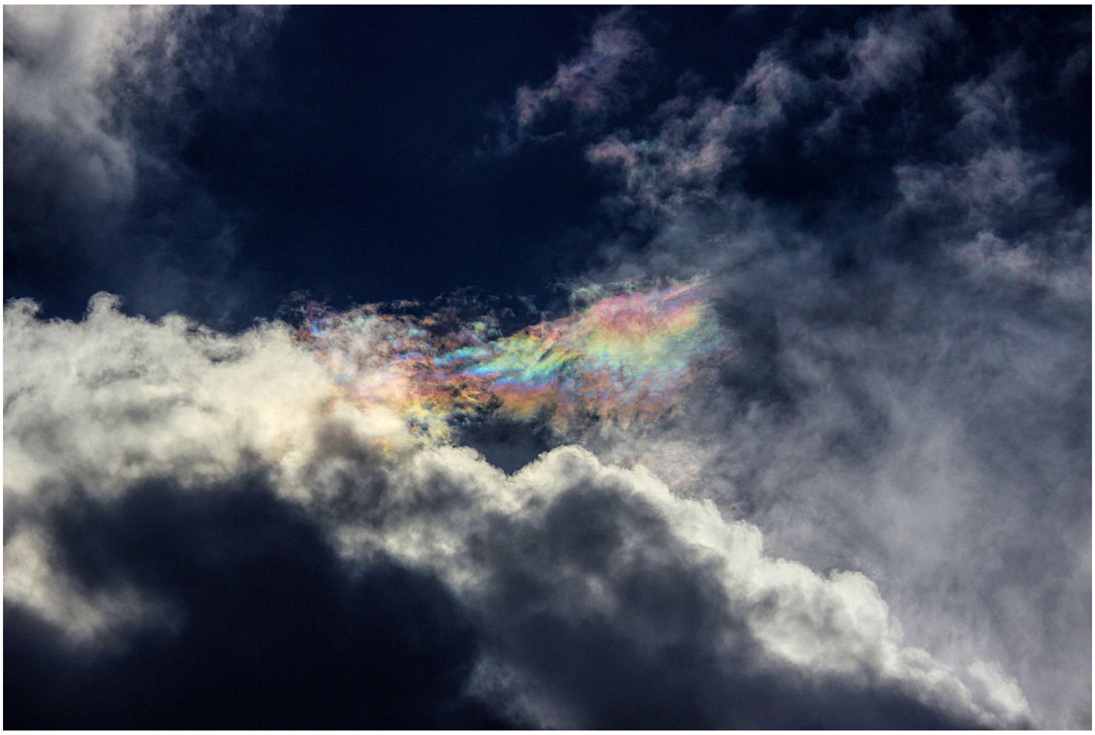


Figure 10.165: Cloud iridescence, seen above the clouds covered with grey clouds (source: Wikipedia, place: Pondicherry (India), date: 29th November 12015 by Karthik Easvur)

10.5.4.5 Optical Mirages

Mirages are generated when light passes through air of different temperatures. There are commonly two types of mirages:

- An "inferior mirage" occurs when you have a dense layer of cold air sitting on above of your line of sight, with a layer of less dense warmer air below your line of sight. This happens frequently on hot summer days when the Sun (actually a G2V-type star, aka yellow dwarf) shines on a black asphalt highway, and the hot ground heats the bottom centimeters of air. When rays of light coming down from the sky get close to the hot air near the ground, the light rays are bent (refracted) back up toward your eyes.

  This is your typical 'oasis in the desert' mirage, where you look down toward the highway or desert sands and see what appears as a pool of blue water. In reality, it is light the blue sky bending back up toward your eyes. Since warm air rises and cool air sinks, the air layers are likely to mix and inferior mirages will not last long unless there is continued heating by the sun.

  Inferior mirages happen less frequently over the ocean because the ocean surface usually doesn't get extremely hot. However, it can occur over the ocean when very cold arctic blows over unfrozen water.

  Note in the image below, an inferior mirage makes it look like there is water on the pavement, when in fact there is not:

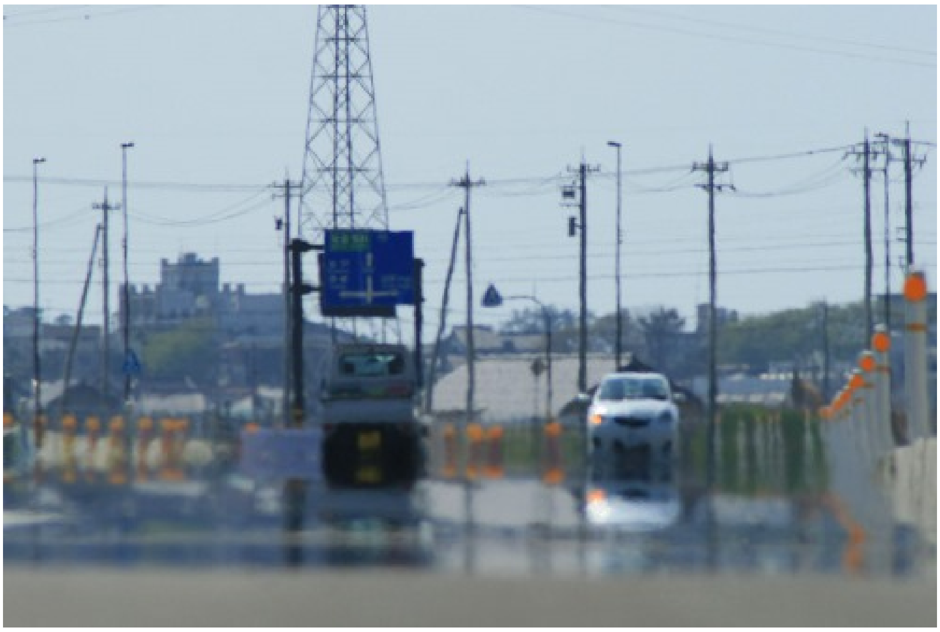
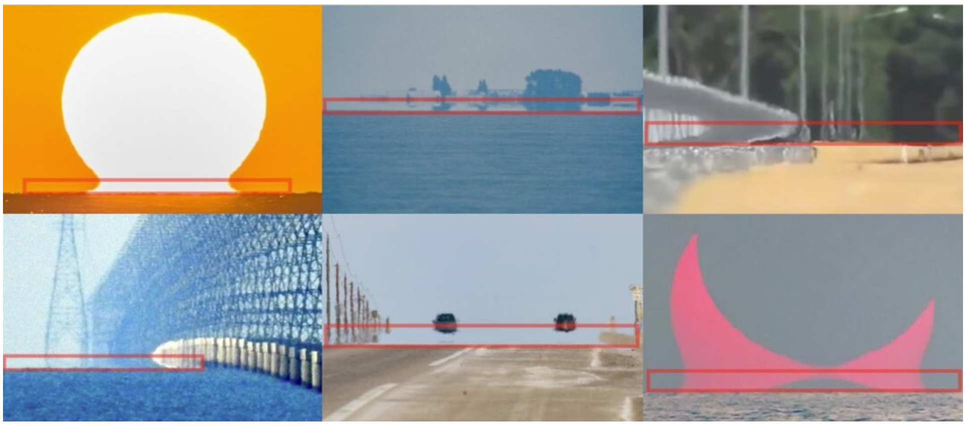
  

  Figure 10.166: **Sky inferior mirage (credit: U-ichiro Murakami - Own work, Public Domain)**

  and some other well known inferior images:



- A "superior mirage" occurs when a layer of warm air sits above your line of sight with a cool layer beneath it. The mirage above is caused by a meteorological phenomenon named a "temperature inversion". Normally, the air temperature drops with increasing altitude, making mountaintops colder than the foothills. But in a temperature inversion, warm air sits on top of a band of colder air, playing havoc with our visual perception. In case of temperature inversion light bends down towards the denser air, but because our eyes assume the light we see travels straight, the object we are seeing appears higher than it actually is. Therefore we see the image, or inversions of it, above where it actually is.

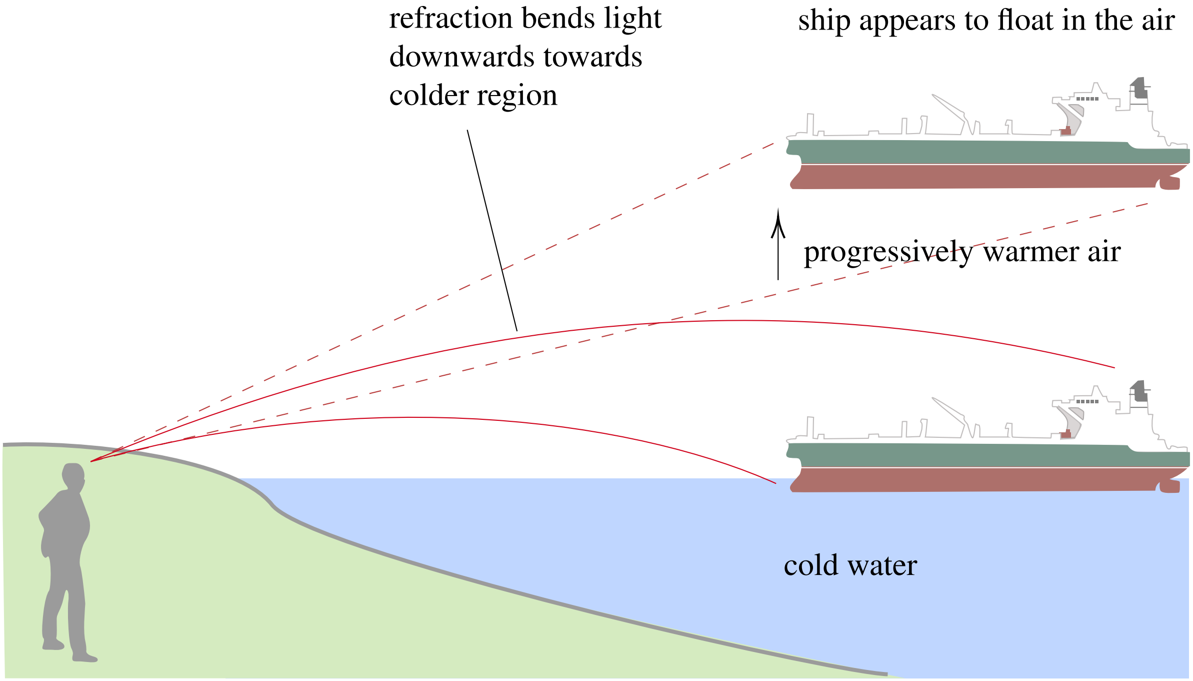

refraction bends light downwards towards colder region
ship appears to float in the air
progressively warmer air
cold water

Figure 10.167: **Guardian graphic. Source: HyperPhysics, Georgia State University (vertical scale exaggerated)**

Or for a real photo[17] (notice that the mirage obviously reproduce the colors and the details of the original object as it is just deflected light!):

---

[17] Respectively in this type of thermal configuration, objects lying below the horizon due to the curvature of the Earth give the impression of being at the level of the horizon and therefore that the Earth is flat (reason for which some flat earthers, observing with binoculars the other shore of a large body of water, are convinced that the Earth is flat because of the fact that they slept during Optics lessons...

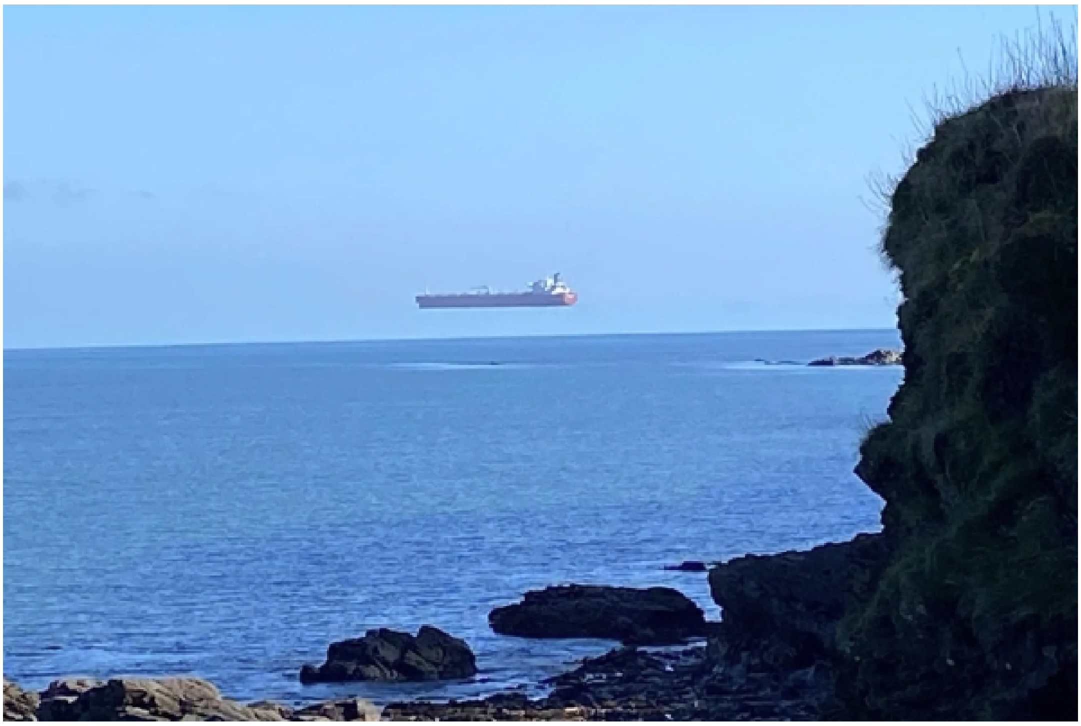
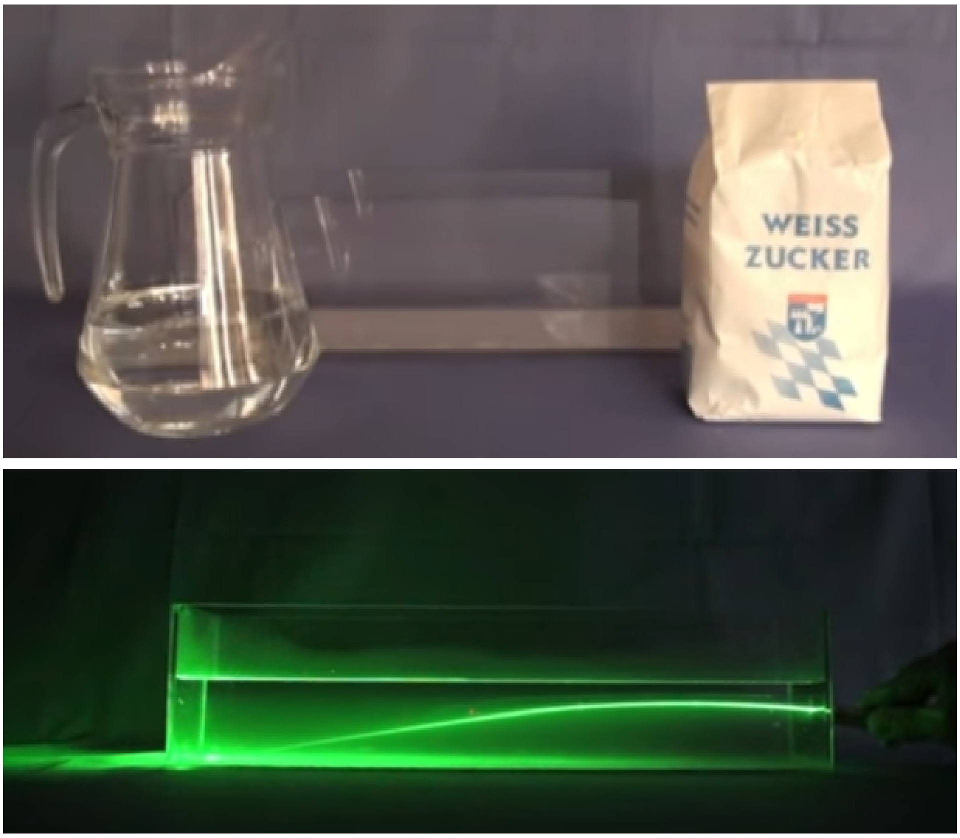


Figure 10.168: **Ship hovering above the surface of the sea off southwestern England (credit: David Morris/Apex)**

> In the case of the virtual image of a superior mirage is completely distorted or inversed, we speak then of a "Fata Morgana".

More run-of-the-mill optical illusions are the “inferior mirages” that give rise to apparent oases in the desert and puddles on hot summer roads, caused by cooler air sitting on a layer of hot air, directly above a road, for example. When sunlight coming down from the sky approaches the air near the hot surface, it is bent back upwards to the observer’s eye, making the blue sky appear to be reflected on the road like water.

As atmosphere is a fluid, and water is also a fluid (just with different characteristics), the bending of light can be easily tested and observed easily with a domestic LASER, hot water and a parallelepiped recipient and waiting a few hours after the mixing the both. Then we can observe the bending of water:



10.5.4.5.1 Lagrangian approach to optical mirage

The approach of using the Lagrangian to determine the path equation of light for an inferior mirage is the most technical and mathematically the most beautiful one to read. However, as we will see it, the result requires to know three constants that are very hard to measure in practice. Once we will be finished with the derivation of this formal Lagrangian way, we will use other techniques that lead to much more simple result to compute numerically with much less constants!

So as already mentioned, in an inferior mirage the rays bend as shown in the figure below where the observer sees a virtual image of the sky superimposed on the ground:

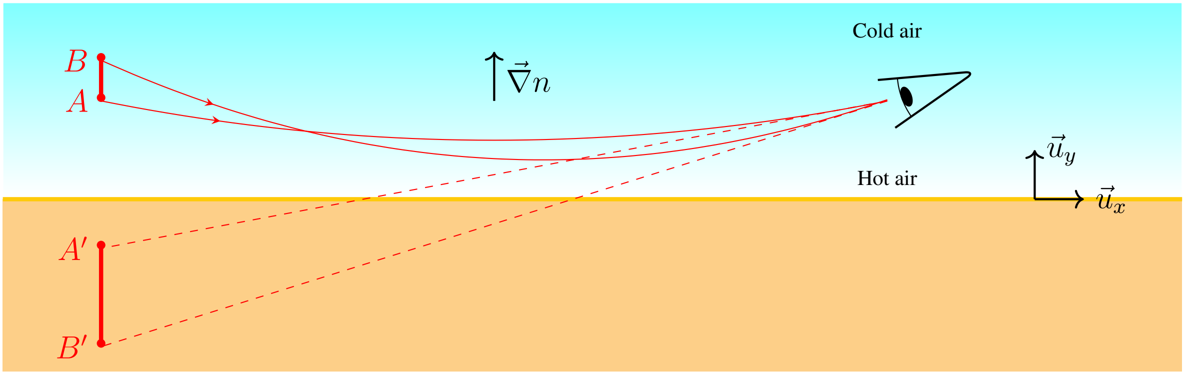


Figure 10.169: **Curvature of the rays near an overheated ground (source: Femto-Physique.fr, creator: Jimmy Roussel)**

Let us establish the differential equation of the radius by supposing that the index only varies according to the altitude y of the place. The optical path is written, in the plane $x\text{O}y$:

$$L=\int n(y)\mathrm{d}s=\int_{\text{A}}^{\text{B}} n(y)\sqrt{\mathrm{d}x^2+\mathrm{d}y^2}=\int n(y)\sqrt{1+y'(x)^2}\mathrm{d}x \tag{10.5.126}$$

Let's assume that the index of refraction varies linearly with y such that:

$$n(y)\cong(1+\alpha y)n_0 \tag{10.5.127}$$

where n_0 is the value of the index of refraction at the ground.

Therefore:

$$L\cong\int(1+\alpha y)n_0\sqrt{1+y'^2}\mathrm{d}x \tag{10.5.128}$$

Let us put:

$$f=(1+\alpha y)n_0\sqrt{1+y'^2} \tag{10.5.129}$$

According to the principle of least action, the path of light is the one that satisfies (see section Analytical Mechanics page 3169):

$$\frac{\partial L}{\partial q_i}-\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial\dot{q}_i}=0 \tag{10.5.130}$$

And let us remind that this has let us to the Beltrami identity (see section Analytical Mechanics page 3177) given by:

$$L-\dot{q}_i\frac{\partial L}{\partial\dot{q}_i}=c^{te} \tag{10.5.131}$$

In our case we have to replace L by f and q_i by y and therefore:

$$f-y'\frac{\partial f}{\partial y'}=(1+\alpha y)n_0\sqrt{1+y'^2}-y'^2(1+\alpha y)n_0\sqrt{\frac{1}{1+y'^2}}=\frac{(1+\alpha y)n_0}{\sqrt{1+y'^2}} \tag{10.5.132}$$

is therefore a constant, that we will denote as c_1. We can solve y' from the equation:

$$\frac{(1+\alpha y)n_0}{\sqrt{1+y'^2}}=c_1 \tag{10.5.133}$$

that is:

$$y'=\frac{\mathrm{d}y}{\mathrm{d}x}=\pm\sqrt{\frac{(1+\alpha y)^2n_0^2}{c_1^2}-1} \tag{10.5.134}$$

Explicited and rearranged:

$$x=\pm\int\frac{\mathrm{d}y}{\sqrt{\frac{(1+\alpha y)^2n_0^2}{c_1^2}-1}} \tag{10.5.135}$$

We can find the integral by a change of variable:

$$\frac{(1+\alpha y)n_0}{c_1}=\cosh(\theta) \tag{10.5.136}$$

Therefore (form the list of all usual derivatives we have detailed):

$$\sinh(\theta)\mathrm{d}\theta=\frac{\alpha n_0}{c_1}\mathrm{d}y \tag{10.5.137}$$

and:

$$\sqrt{\frac{(1+\alpha y)^2n_0^2}{c_1^2}-1}=\sqrt{\cosh^2(\theta)-1}=\sinh(\theta) \tag{10.5.138}$$

Hence:

$$x=\pm\int\frac{\mathrm{d}y}{\sqrt{\frac{(1+\alpha y)^2n_0^2}{c_1^2}-1}}=\pm\frac{c_1}{\alpha n_0}\int\frac{\sinh(\theta)\mathrm{d}\theta}{\sinh(\theta)} \tag{10.5.139}$$

The solution can therefore be written as:

$$x=\pm\frac{c_1}{\alpha n_0}\int\mathrm{d}\theta=\pm\frac{c_1}{\alpha n_0}\theta+c_2 \tag{10.5.140}$$

Where c_2 in the constant of integration. Les us put $c_2:=x_0$. Therefore we can write the latter as:

$$\theta=\pm\frac{\alpha n_0}{c_1}(x-x_0) \tag{10.5.141}$$

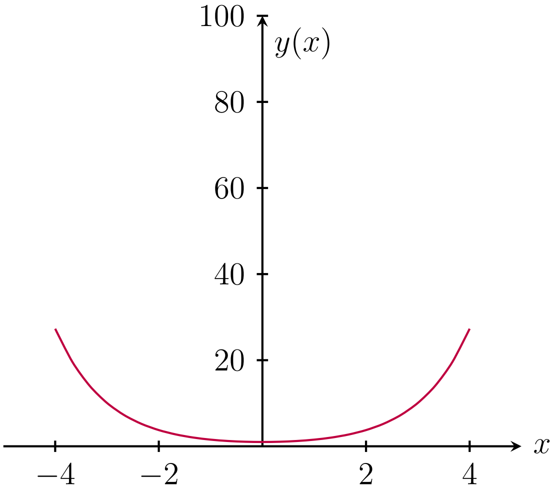
Substituting this θ into the relation we got earlier above, we have :

$$\frac{(1+\alpha y)n_0}{c_1}=\cosh(\theta) \tag{10.5.142}$$

we get after rearranging (and we can keep only the positive sign because $\cosh$ is an even function):

$$\boxed{y\cong\frac{c_1}{\alpha n_0}\cosh\left(\frac{\alpha n_0}{c_1}(x-x_0)\right)-\frac{1}{\alpha}} \tag{10.5.143}$$

The most important information from the solution is that y is always convex as seen in the section of trigonometry page 2538 (or concave up). Below we have the plot of $\cosh(x)$:



Remark

A less technical derivation based on Snell's law exists but it assumes that we already know the trajectory of the light path is a parabola. So the derivation doesn't make too much sense in our opinion...

Let us see now an other approach much less formal but that leads to practical usable result!

10.5.4.5.2 Snell's law differential approach to air mirroring

Let us remind the Snell-Descartes law (i.e. refraction law) given by (see section Electromagnetism page 4364):

$$n_{12}=\frac{n_1}{n_2}=\frac{\sin(r)}{\sin(i)} \tag{10.5.144}$$

where the i stands for "incident" (often denoted θ_1) and the r for "refracted" (often denoted θ_2).

And also remember the following figure:

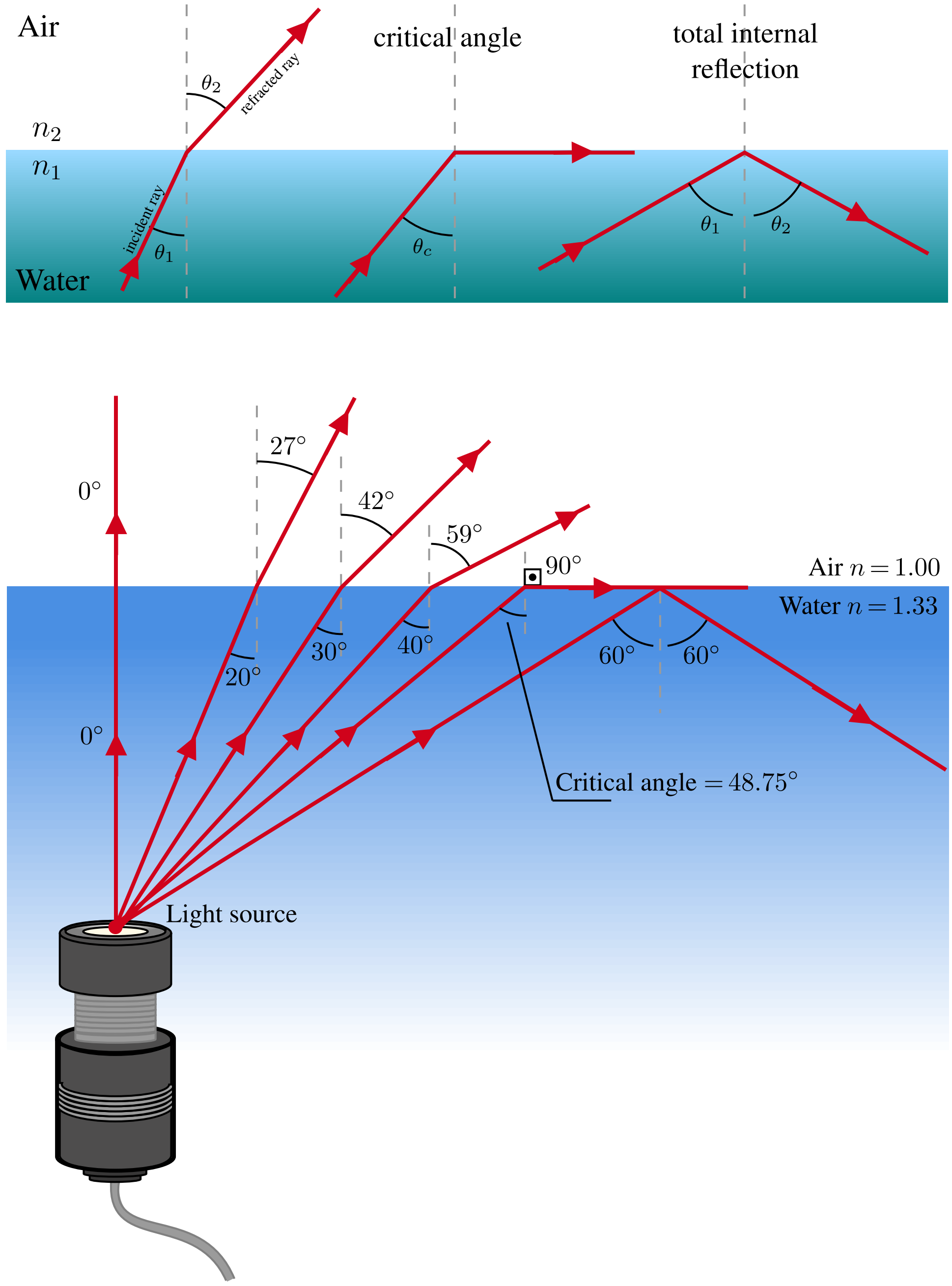

Air
θ_2
refracted ray
critical angle
total internal reflection
n_2
n_1
incident ray
θ_1
θ_c
θ_1
θ_2
Water
0°
27°
42°
59°
90°
Air $n = 1.00$
Water $n = 1.33$
0°
20°
30°
40°
60°
60°
Critical angle = 48.75°
Light source

Figure 10.170: **Critical angle schematic illustration (raytracing)**

Remark

Remember two things when looking at the figure above:

1. The light ray-tracing is symmetric. The path of light remains the same wherever on which side of the interface separation is the source!
2. Remember also that that below (incident) critical angle the refraction law applies:

$$n_{12} = \frac{n_1}{n_2} = \frac{\sin(\text{refraction})}{\sin(\text{incident})}$$

and beyond the critical angle it's the reflection law that applies:

$$\theta_{\text{incident}}=\theta_{\text{reflection}}$$

and that happens only if light move from a dense medium to a less dense medium!

We shall now generalize this law to the case, where the index of refraction changes continuously in the y-direction, that is, along the normal to the surface of the materials. We shall calculate the infinitesimal deviation, corresponding to an infinitesimal change of the index of refraction $n=n(y)$. The angle that the light ray forms with the normal is denoted θ. So the law of refraction writes:

$$\frac{\sin(\theta)}{\sin(\theta+\mathrm{d}\theta)}=\frac{n(y)}{n(y+\mathrm{d}y)}\quad\Leftrightarrow n(y+\mathrm{d}y)\sin(\theta)=n(y)\sin(\theta+\mathrm{d}\theta) \tag{10.5.145}$$

By subtracting $n(y)\sin(\theta)$ on both sides and putting outside a parenthesis, we have:

$$\sin(\theta)(n(y+\mathrm{d}y)-n(y))=n(y)(\sin(\theta+\mathrm{d}\theta)-\sin(\theta)) \tag{10.5.146}$$

Hence written in a differential form:

$$\sin(\theta)\mathrm{d}n=n(y)\mathrm{d}\sin(\theta) \tag{10.5.147}$$

Therefore:

$$\frac{\mathrm{d}\sin(\theta)}{\sin(\theta)}=\frac{\mathrm{d}n}{n(y)} \tag{10.5.148}$$

This equation can immediately be integrated, i.e.:

$$\int_{\theta_1}^{\theta_2}\frac{\mathrm{d}\sin(\theta)}{\sin(\theta)}=\int_{n_1}^{n_2}\frac{\mathrm{d}n}{n(y)} \tag{10.5.149}$$

We know that that this is equal to:

$$\ln\left(\sin(\theta_2)\right)-\ln\left(\sin(\theta_1)\right)=\ln\left(n(y_2)\right)-\ln\left(n(y_1)\right) \tag{10.5.150}$$

Hence rearranging and taking the exponential, this leads to:

$$\boxed{\frac{\sin(\theta_2)}{\sin(\theta_1)}=\frac{n(y_2)}{n(y_1)}} \tag{10.5.151}$$

Thus, we recover the law of refraction unchanged. The refraction angle depends only on the index of refraction in the upper layer and the lower layer. This means that refraction through a layer with a thickness of hundreds of meters can be treated as refraction in a single borderline!

This has the important consequence that total reflection is possible in layers of air many meters thick, and this is the key to explain the phenomena of air mirroring and mirages.

Now let us remind that we have introduce in the section of Optics the concept of total reflection, (see page 4369 and we have shown that for the transition water/air, the relative index of refraction was about $1/1.33$, which gave us a limit critical angle of:

$$\theta_c = \theta_i = \arcsin\left(\frac{n_2 \sin(90^\circ)}{n_1}\right) = \arcsin\left(\frac{1}{1.33}\right) = 48.7^\circ \qquad (10.5.152)$$

And as we have already mentioned it, the refractive index depends not only on the temperature but also on pressure, composition, and strictly speaking also the wavelength of the light...

In air the index of refraction is much less. At a relative index of refraction $1/1.001$, we find a limiting angle about 87°, which corresponds to a grazing angle of 3°.

Before continuing we can using the empirical Ciddor equation or a modified version of the empirical Edlén equation get the refractive index of the air for different temperatures. For example at a wavelength 633 [nm] and at a pressure of 101.325 [kPa] a 50% relative humidity and for a carbon dioxide content of 450 Micromole per Mole[18] [parts per million, ppm]:

Temperature	n **(air)**
-40	1.000342056
-10	1.000302844
-8	1.000300541
-6	1.000298271
-4	1.000296032
-2	1.000293825
0	1.00029165
2	1.0002895
4	1.00028739
6	1.0002853
8	1.00028323
10	1.0002812
12	1.000279185
14	1.0002772
16	1.000275234
18	1.00027329
20	1.00027137
22	1.000269473
24	1.00026759
26	1.00026573
28	1.00026389
30	1.000262061
32	1.00026025
34	1.00025845
36	1.00025667
38	1.000254898

---

[18] The values are obtained with the online calculator available at: `https://emtoolbox.nist.gov/Wavelength/Ciddor.asp`

40	1.00025314
42	1.00025139

Table 10.12: **Various index of refraction of air**

The absolute index of refraction for air as we can see in the table above at zero degree Celsius is approximately equal to 1.000292, so the quick conclusion would be that the limiting angle for total reflection in air can at most be $\theta_i = 88.6^\circ$, which corresponds to a grazing angle of 1.4°.

The lower air mirroring example is the most common mirage, since it can be experienced on sunny days, where it appears that there is water on the asphalt road ahead a vision which disappears as soon as we get nearer. We generally consider that when it is really hot in Switzerland, that is to say when the air reaches 30 $[^\circ\text{C}]$ to 40 $[^\circ\text{C}]$ in the Sun (101.325 [kPa]) hence a refractive index of approximately $n = 1.000260748$, the road surface heats up to approximately 60 $[^\circ\text{C}]$ to 70 $[^\circ\text{C}]$ (290 [kPa]) hence a refractive index of approximately $n = 1.000328216$ (remember that hot air is less dense than cold air!). Hence:

$$\theta_c = \theta_i = \arcsin\left(\frac{n(y_2)\sin(90^\circ)}{n(y_1)}\right) = \arcsin\left(\frac{1.000260748}{1.000328216}\right) \cong 89.33^\circ \tag{10.5.153}$$

The explanation is straightforward! Because the air gets heated near the black asphalt lane, this gives it a smaller density, and consequently a less index of refraction (see table above).

As emphasized above, the phenomena of total reflection is independent of whether the reflection takes place at a borderline or it takes place in a gradual change of the index of refraction through an arbitrary thick layer. Hence the puddles that you can see on a dark road in sunny weather is in reality an (inverse) lower mirroring of the sky!

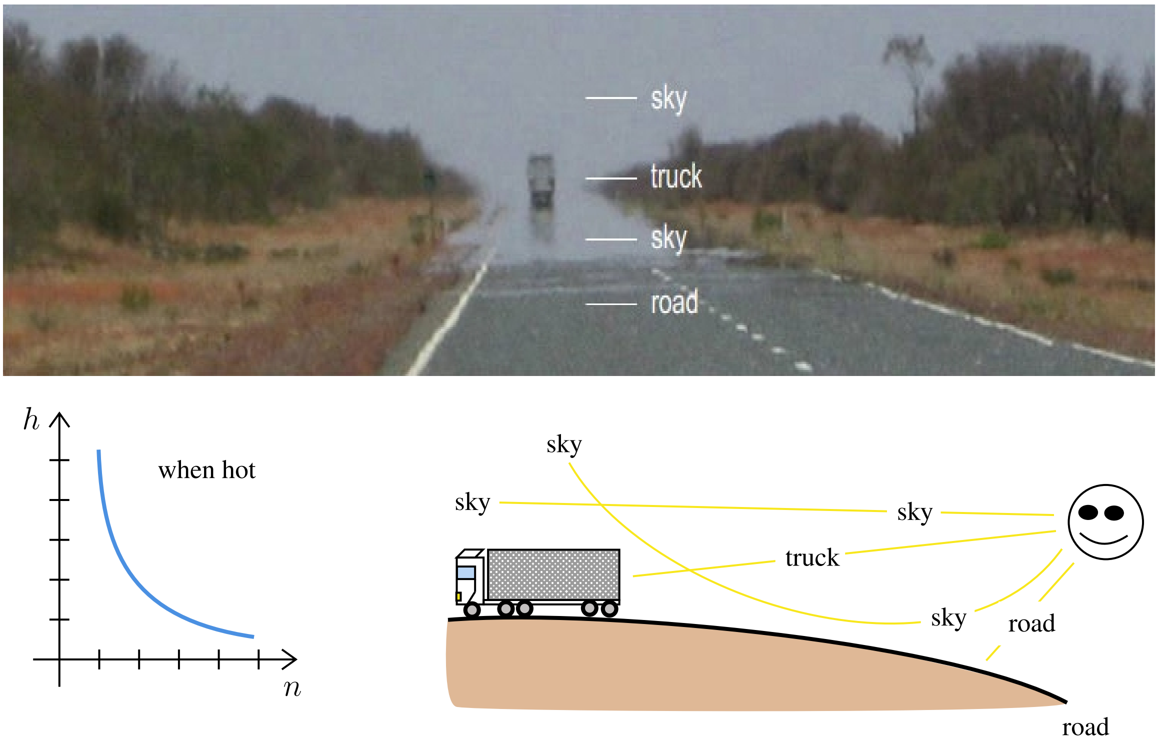


Figure 10.171: **Floating road train on the Stuart Highway, Northern Territory, Australia**

The upper mirroring is the same phenomena, the only difference being that the air near the ground is cooler than the air higher up. The upper mirrorings are much more rare at our latitudes, and surprisingly enough they appear mostly in polar areas.

Common for the two kinds of mirroring is that they produce an inverse image of the object. Concerning the mirages (fata morgana), it is not a mirroring, but instead it is caused by light that is refracted in the same direction as the curvature of the earth. For this reason it is possible to see objects that otherwise are below the horizon. The figures below

Before we move to the next topic, let us consider the figure below:

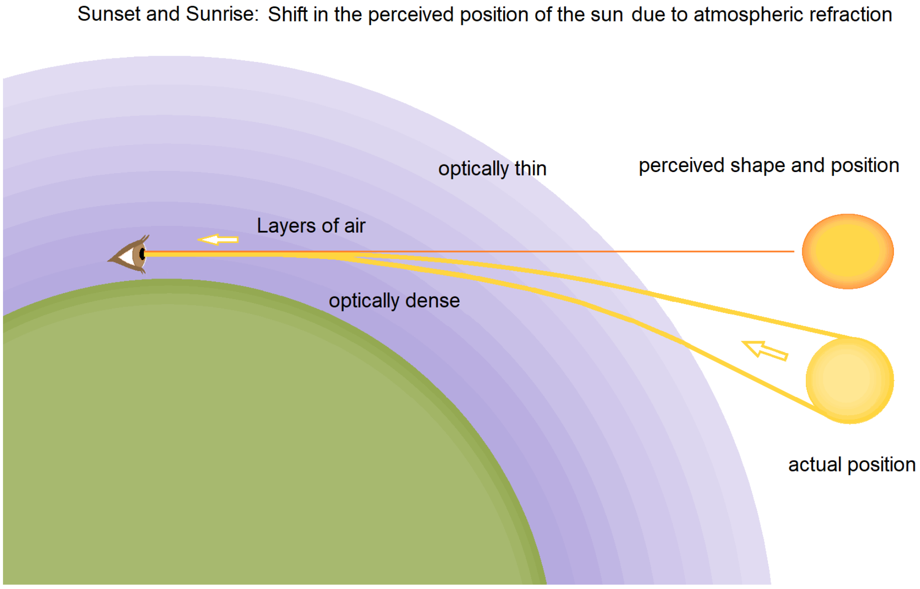

Sunset and Sunrise: Shift in the perceived position of the sun due to atmospheric refraction
optically thin
perceived shape and position
Layers of air
optically dense
actual position

Figure 10.172: **Shift of the perceived form and position of the Sun due to the atmospheric refraction (source: Wikipedia, author: Sciencia58)**

And remember the relation:

$$\frac{\sin(\theta_2)}{\sin(\theta_1)}=\frac{n(y_2)}{n(y_1)} \qquad (10.5.154)$$

Let us take for $n(y_2)$ the index refraction of air at a temperature of $40\,[^\circ\mathrm{C}]$ (hot ground) and at a pressure of 101.325 [kPa] and for $n(y_1)$ the index refraction of air at a temperature of $-40\,[^\circ\mathrm{C}]$ (high altitude) and at a pressure of 10 [kPa]. Then we have with these two extrema the following limiting angle for sunset/sunrise (as commonly named... a better denomination being nightfall/daybreak) as the refraction angle depends only on the index of refraction in the upper layer and lower layer (under the assumption of a linear change!):

$$\theta_c=\arcsin\left(\frac{n(T=-40,P=10)}{n(T=40,P=101.325)}\right)=\arcsin\left(\frac{1.000033717}{1.000271373}\right)\cong 88.751^\circ \qquad (10.5.155)$$

Hence we can see the sunset/sunrise is some weather conditions when the Sun (actually a G2V-type star, aka yellow dwarf) is in reality already $90^\circ-88.751^\circ\cong 1.24^\circ$ below horizon!

Below we can see an astronomical size mirage diagram (obviously not to scale and oversimplified) due to atmospheric refraction named "selenelion". It is a partial lunar eclipse phenomenon at dusk or dawn. When a partial lunar eclipse occurs near dawn or dusk, you can often see the red Sun in one direction and a part of the red Moon in the opposite direction, with the Sun's light passing over your head:

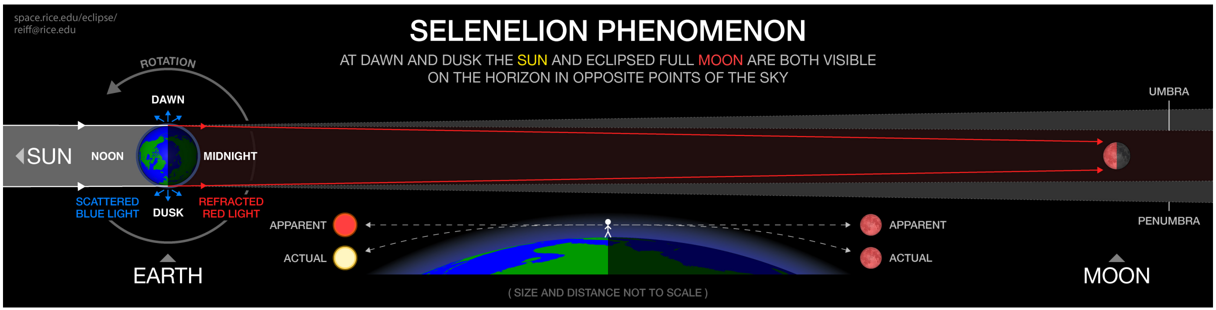

space.rice.edu/eclipse/
reiff@rice.edu
SELENELION PHENOMENON
AT DAWN AND DUSK THE SUN AND ECLIPSED FULL MOON ARE BOTH VISIBLE
ON THE HORIZON IN OPPOSITE POINTS OF THE SKY
ROTATION
DAWN
SUN
NOON
MIDNIGHT
SCATTERED BLUE LIGHT
DUSK
REFRACTED RED LIGHT
EARTH
APPARENT
ACTUAL
APPARENT
ACTUAL
UMBRA
PENUMBRA
MOON
(SIZE AND DISTANCE NOT TO SCALE)

Figure 10.173: **Selenelion diagram (author: Patricia Reiff and Colin Law, Rice University)**

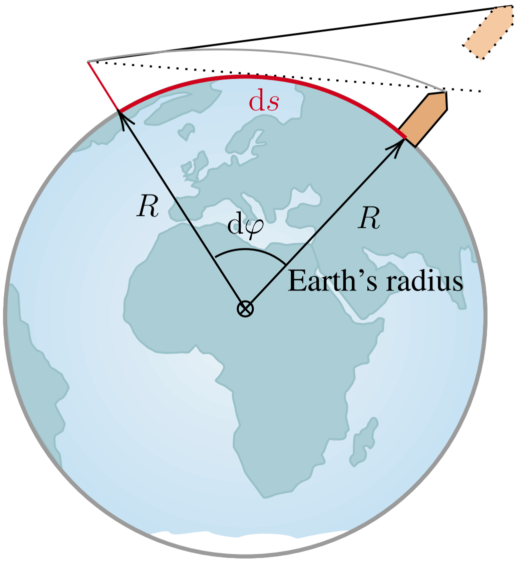
For a scientific treatment of the selenelion eclipse see [475].

10.5.4.5.3 Differential form of long range mirages

We shall discuss in somewhat more detail the phenomena of mirages (fata morgana), which occurs when a light ray is refracted along the curvature of the Earth.

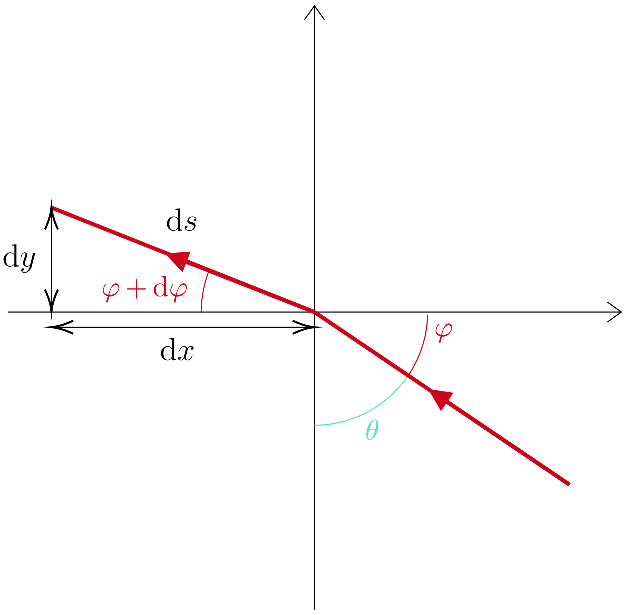
By the way the phenomenology of the analysis that will follow will be useful to us to calculate the visual horizon in the section of Engineering at page 8247.

The figure below shows a kind of tower, which lie below the horizon, which means that it cannot be seen, if the light followed a straight line. If the light ray is bended when passing the atmosphere, however, the tower will be seen in a direction following the tangent to the ray at the point of observation, as "floating" in the air. The refraction is due to an abnormal temperature gradient, that is, the temperature rises with the height from the ground. The cooler air has a higher density, and therefore a higher index of refraction.



The figure above to the left depicts a section of the surface of the earth, where the connection between the $\mathrm{d}s$, the radius of the Earth R and the angle $\mathrm{d}\varphi$ is $\mathrm{d}s = R \cdot \mathrm{d}\varphi$

A light ray which moves slightly upwards will move towards a slightly less (relative) index of refraction, and consequently be refracted away from the normal to the layer, that is, downwards. The next step is to establish a relation between the differential bending $\mathrm{d}\varphi$ of the beam and the change $\mathrm{d}n/\mathrm{d}y$ in the index of refraction per unit length along the normal to the layer:



The figure above shows a light ray on its way up through the atmosphere getting an infinitesimal deflection $\mathrm{d}\varphi$ at the distance $\mathrm{d}s$. Where φ is the grazing angle between the borderlines, and $\varphi=90^\circ-\theta$, where θ is the incoming angle with the normal to the layer.

We have proven just earlier above during our study of mirages (see page 4397) that:

$$\frac{\mathrm{d}\sin(\theta)}{\sin(\theta)}=\frac{\mathrm{d}n}{n(y)} \tag{10.5.156}$$

and we have:

$$\mathrm{d}\sin(\theta)=\cos(\theta)\mathrm{d}\theta \tag{10.5.157}$$

Hence:

$$\frac{\mathrm{d}\sin(\theta)}{\sin(\theta)}=\frac{\mathrm{d}n}{n}\Leftrightarrow\cos(\theta)\mathrm{d}\theta=\frac{\sin(\theta)}{n}\mathrm{d}n \tag{10.5.158}$$

And now we use $\varphi=90^\circ-\theta$ hence $\mathrm{d}\varphi=-\mathrm{d}\theta$, but also $\cos(\varphi)=\cos(90^\circ-\theta)=\sin(\theta)$ and $\sin(\varphi)=\sin(90^\circ-\theta)=\cos(\theta)$, therefore we car rewrite:

$$\cos(\theta)\mathrm{d}\theta=\frac{\sin(\theta)}{n}\mathrm{d}n\Leftrightarrow\mathrm{d}\varphi=-\frac{\cos(\varphi)}{n\sin(\varphi)}\mathrm{d}n \tag{10.5.159}$$

We then use:

$$\mathrm{d}n=\frac{\mathrm{d}n}{\mathrm{d}y}\mathrm{d}y \tag{10.5.160}$$

and since (see figure above):

$$\mathrm{d}y=\sin(\varphi)\mathrm{d}s \tag{10.5.161}$$

Then:

$$\mathrm{d}n=\frac{\mathrm{d}n}{\mathrm{d}y}\sin(\varphi)\mathrm{d}s \tag{10.5.162}$$

We finally have:

$$\mathrm{d}\varphi=-\frac{\cos(\varphi)}{n\sin(\varphi)}\sin(\varphi)\frac{\mathrm{d}n}{\mathrm{d}y}\mathrm{d}s \quad\Leftrightarrow\quad \mathrm{d}\varphi=-\frac{\cos(\varphi)}{n}\frac{\mathrm{d}n}{\mathrm{d}y}\mathrm{d}s \tag{10.5.163}$$

If a light ray should follow the curvature of the earth, we may determine the gradient of the index of refraction. Comparing this with $\mathrm{d}s=R\cdot\mathrm{d}\varphi\Leftrightarrow\mathrm{d}\varphi=1/R\mathrm{d}s$ we have:

$$\mathrm{d}\varphi=-\frac{\cos(\varphi)}{n}\frac{\mathrm{d}n}{\mathrm{d}y}\mathrm{d}s \qquad \text{and} \qquad \mathrm{d}\varphi=\frac{1}{R}\mathrm{d}s \tag{10.5.164}$$

That gives:

$$-\frac{\cos(\varphi)}{n}\frac{\mathrm{d}n}{\mathrm{d}y}=\frac{1}{R} \quad\Leftrightarrow\quad \frac{\mathrm{d}n}{\mathrm{d}y}=-\frac{n}{R\cos(\varphi)} \tag{10.5.165}$$

Ok we have here the main relations. The most important one in reality is:

$$\boxed{\Delta\varphi=-\frac{\cos(\varphi)}{n}\frac{\mathrm{d}n}{\mathrm{d}y}\Delta s} \tag{10.5.166}$$

where $\mathrm{d}n/\mathrm{d}y$ is named the "refraction rate".

✍**Example:**

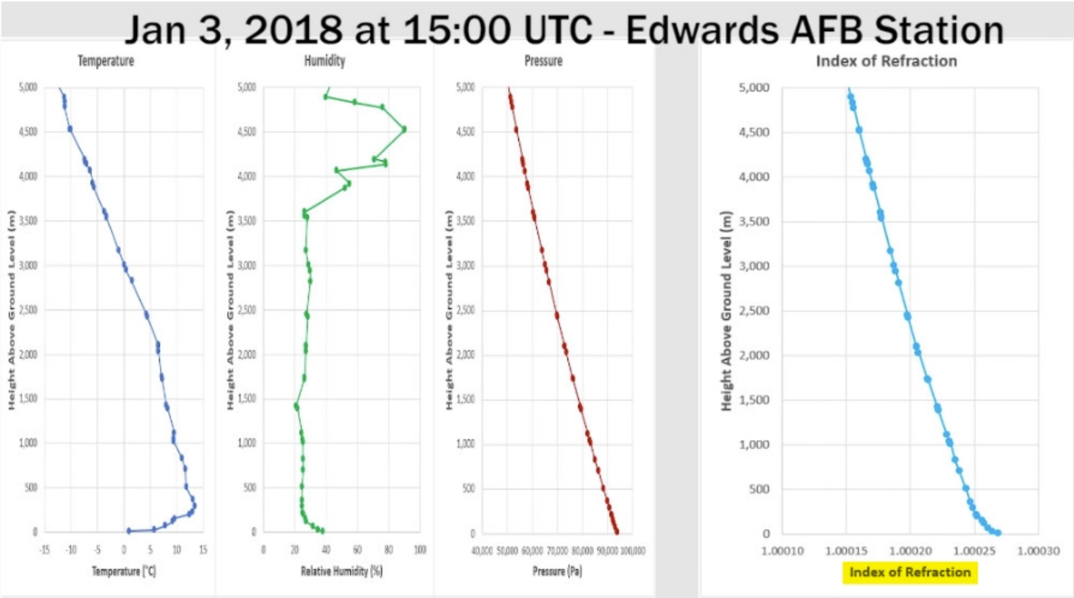
Let us consider the Monte Cinto ($2,710$ [m]) in Corsica and Mont Chauve d'Aspremont (854 [m]) near Nicea. It is measured between both of them that $\Delta s\cong 210$ [km] and we have assume to have for grazing angle $\varphi\cong 1.88^\circ$ and for Earth's radius $R\cong 6,370$ [km].

Hence, taking $n=1$ we have for the gradient of index of refraction:

$$\frac{\mathrm{d}n}{\mathrm{d}y}=\frac{n}{R\cos(\varphi)}=\frac{1}{6,370\cdot 10^3\cdot\cos(1.88^\circ)}\cong 1.57\cdot 10^{-7}\,[\mathrm{m}^{-1}]$$

According to the references the gradient will seldom have such a large value!

Indeed, see the measurement in Winter below where the index of refraction (right field below) has been calculated thanks to the Ciddor equation (source: cinnamoncontrol):



Using the rightmost chart we get (for the theoretical derivation of $\mathrm{d}n/\mathrm{d}y$ see page 4360):

$$\frac{\mathrm{d}n}{\mathrm{d}y} \cong \frac{1.00027-1.00015}{5,000} \cong 2.4\cdot 10^{-8}$$

Assuming therefore a refraction index gradient of $2.4\cdot 10^{-8}$ $[\mathrm{m}^{-1}]$, that is, approximately one tenth of the value above. Then the deviation on the distance 210 [km] from Monte Cinto to Mont Chauve d'Aspremont (near Nicea) can be calculated from:

$$\Delta\varphi = -\frac{\cos(\varphi)}{n}\frac{\mathrm{d}n}{\mathrm{d}y}\Delta s \cong -\frac{\cos(1.88^\circ)}{1}\cdot(2.4\cdot 10^{-8})\cdot(210\cdot 10^{3}) \cong -0.288^\circ$$

That means that the abnormal temperature gradient virtually brings together the to location of $(0.288^\circ = -0.00504$ [rad]) :

$$\Delta(\Delta s) = \Delta\varphi R = -0.00504\cdot 6,370\cdot 10^{3} \cong -32.1\ [\mathrm{km}]$$

This calculation thus delivers an (also quantitative) explanation of the phenomena that Monte Cinto in clear weather and winter can be viewed from Nicea!

We can therefore with the example above see that in the case of a gradient typical of that of winter, it's kinda similar as if the radius of the Earth is shrinked. That's why many of the horizon distance formulas replace the radius of the Earth R, by the "effective radius of Earth" :

$$R' = \frac{R}{1+k} \qquad (10.5.167)$$

where k is named "refraction coefficient" (not to be confused with the "refraction index" n).

10.5.4.6 Cherenkov radiation

Cherenkov radiation, also known as Vavilov–Cherenkov radiation, is electromagnetic radiation emitted when a charged particle (such as an electron) passes through a dielectric medium at a speed greater than the phase velocity of light in that medium. The characteristic blue glow of an underwater nuclear reactor is due to Cherenkov radiation. It is named after Soviet scientist Pavel Alekseyevich Cherenkov, the 11958 (holocene calendar) Nobel Prize winner who was the first to detect it experimentally. A theory of emitted corresponding spectrum was later developed within the framework of Einstein's special relativity theory by Igor Tamm and Ilya Frank, who also shared the Nobel Prize. Cherenkov radiation had been qualitatively predicted by the English polymath Oliver Heaviside in papers published in 11888–89 (holocene calendar).

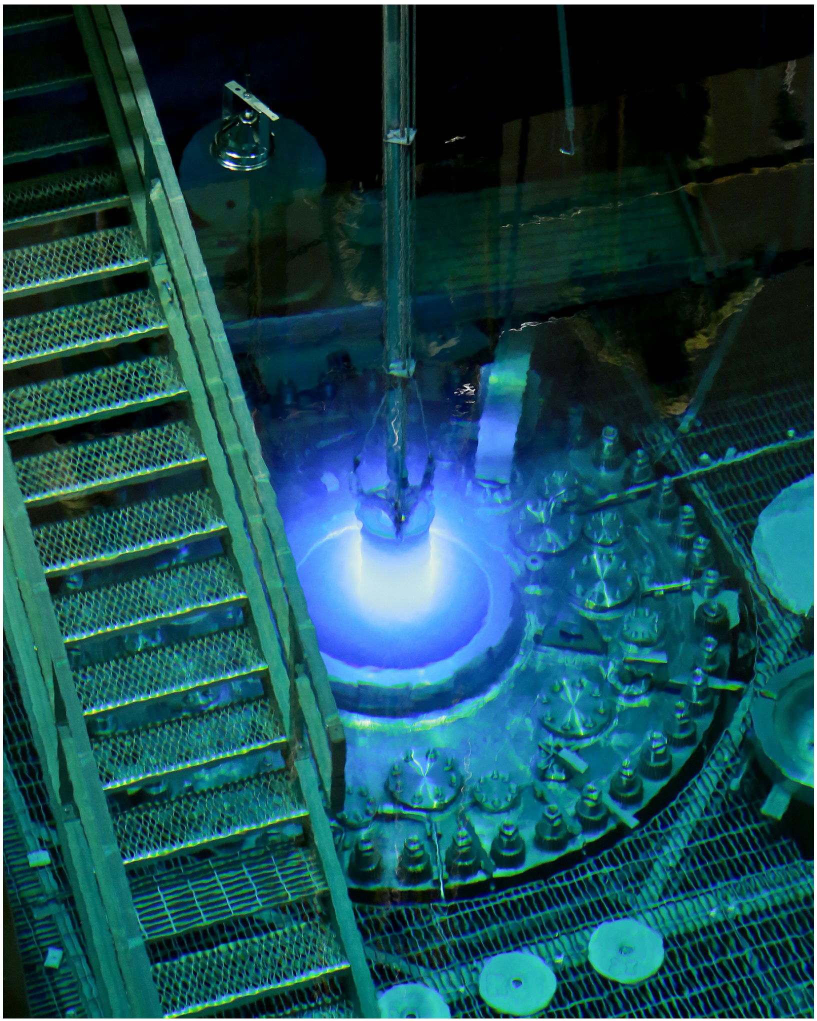


Figure 10.174: Oak Ridge National Laboratory Nuclear Reactor

Remark

Sometimes some wonder why charged particles can go faster than light in a medium other than a vacuum. It is simple to the point: even if the two particles meet roughly the same obstacles and difficulties to spread in a medium the photon cannot be accelerated by a pulse when a charged particle can be accelerated by a given phenomenon in a given medium.

We saw in the preceding paragraphs the hypothesis (relatively intuitive) than the phase propagation speed of light in an absolute refractive index medium n was not equal to c but still less by writing this:

$$v_n = \frac{c}{n} \tag{10.5.168}$$

The Cherenkov effect is (basically) a phenomenon similar to that of an (acoustic) shock wave, but producing a flash of light instead of sound, which takes place on the path of a charged particle moving in a medium with a higher phase velocity that of the speed of light in that medium (rigorous explanation is beyond the scope of study of this textbook because of its complexity treatment!).

Indeed, let us first recall that we proved in the section Electrodynamics that any moving charged particle emits electromagnetic radiation. Then we have proved in the preceding paragraphs that the speed of light in a given medium depends on the absolute refractive index n of the medium (hypothesis which is verified by the experimental accuracy of the theoretical developments arising therefrom).

So we have two basic informations:

1. The speed of the charged particle that can be written as follows with the traditional relativistic notations:

$$v = \beta c \tag{10.5.169}$$

2. The phase velocity of light in a medium with absolute refractive index given:

$$v_c = \frac{c}{n} \tag{10.5.170}$$

It is easy to see that to get $v > v_c$ we must have:

$$v > v_c \Leftrightarrow \beta c > \frac{c}{n} \tag{10.5.171}$$

Therefore:

$$\beta > \frac{1}{n} \tag{10.5.172}$$

Some authors prefer to compare the distance travelled by the light in relation to the distance travelled by the particle. It comes that as:

$$\boxed{\frac{d_{\text{light}}}{d_{\text{part}}} = \frac{\left(\dfrac{c}{n}\right)\Delta t}{\beta c \Delta t} = \frac{1}{\beta n}} \tag{10.5.173}$$

And therefore for the particle traverses distances equal to those of light at in the same time interval we must have that $\beta = 1/n$. Beyond appears the Cherenkov effect.

10.5.5 Descartes' Formulas

We have previously discussed some phenomena that occur when a wavefront passes from one medium to another in which the propagation is different. Not only have we analysed what becomes the wavefront, but we have introduced the concept of "radius" which is particularly useful for geometric constructions. We now propose to deepen the phenomena of refraction and reflection of a geometric point of view using the concept of radius as the tool to describe the processes taking place in the discontinuity surfaces of the propagation. We also assume that the process is limited to reflections and refractions, no other changes affecting the wave surfaces.

This geometry treatment is correct until surfaces and discontinuities encountered by the wave during its propagation are very large relatively to the wavelength. As long as this condition is met, the treatment applies both to light waves, sound (in particular ultrasound - very high frequency), seismic waves, etc.

We begin by considering the wave reflection on a spherical surface. For this purpose we must first given some obvious definitions: The center of curvature C (see section Differential Geometry page 2866) is the center of the spherical surface of the figure below and the top O is the pole of the spherical cap.

Definition (#10.5.493): The line passing through the pole cap O and the center of curvature C is named the "optical axis".

If we take O as coordinate origin, all measured quantities to the right of O will be taken as positive, all those on the left as negative !!!

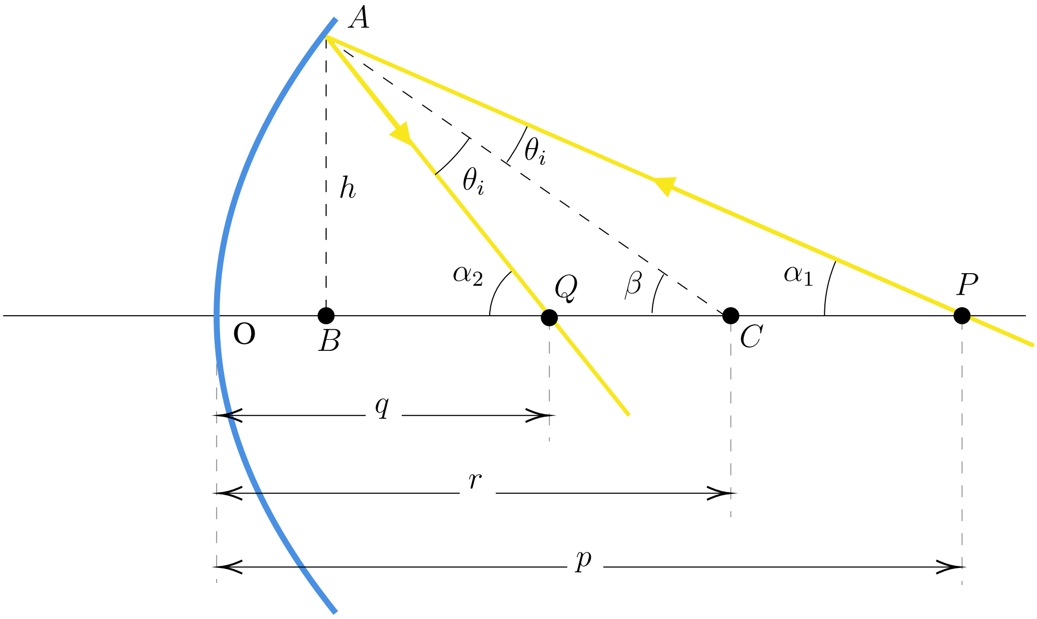


Figure 10.175: **Representation of the optical axis concept**

Let us suppose that the point P is a source of spherical waves. The radius $\overline{PA}$ gives by reflection the radius $\overline{AQ}$ and, as the angles of incidence and reflection are equal with respect to the perpendicular $\overline{AC}$ to the surface (as we have already observed it in our study of refraction), we see in the figure that:

$$\beta=\theta_i+\alpha_1 \quad \text{and} \quad \alpha_2=\beta+\theta_i \tag{10.5.174}$$

hence:

$$\alpha_1+\alpha_2=2\beta \tag{10.5.175}$$

Assuming that the angles α_1,α_2 and β are very small, that is to say, the rays are "para-axial" and the source is very far or that the sensor is very small compared to the source, we can write with a good approximation with a Maclaurin development (see section Sequences and Series page 1703) for small angles:

$$\begin{aligned}\alpha_1\cong\tan(\alpha_1) &= \frac{\overline{AB}}{\overline{BP}}\cong\frac{h}{p}\\ \alpha_2\cong\tan(\alpha_2) &= \frac{\overline{AB}}{\overline{BQ}}\cong\frac{h}{q}\\ \beta\cong\tan(\beta) &= \frac{\overline{AB}}{\overline{BC}}\cong\frac{h}{r}\end{aligned} \tag{10.5.176}$$

Substituting these approximations of α_1,α_2 and β into $\alpha_1+\alpha_2=2\beta$, we get:

$$\boxed{\frac{1}{p}+\frac{1}{q}\cong\frac{2}{r}} \tag{10.5.177}$$

which is the "Descartes formula for reflection on a concave spherical surface". It involves, in the approximation used to establish it, that for all incident rays passing through P will pass through Q after reflection on the surface. We can then say that Q is "the image of the object" P.

In the special case where the incident beam is parallel to the optical axis, that is equivalent to place the object at a great distance of the lens, we have $p=+\infty$. The Descartes formula for reflection on a concave spherical surface becomes therefore:

$$\frac{1}{q}=\frac{2}{r} \tag{10.5.178}$$

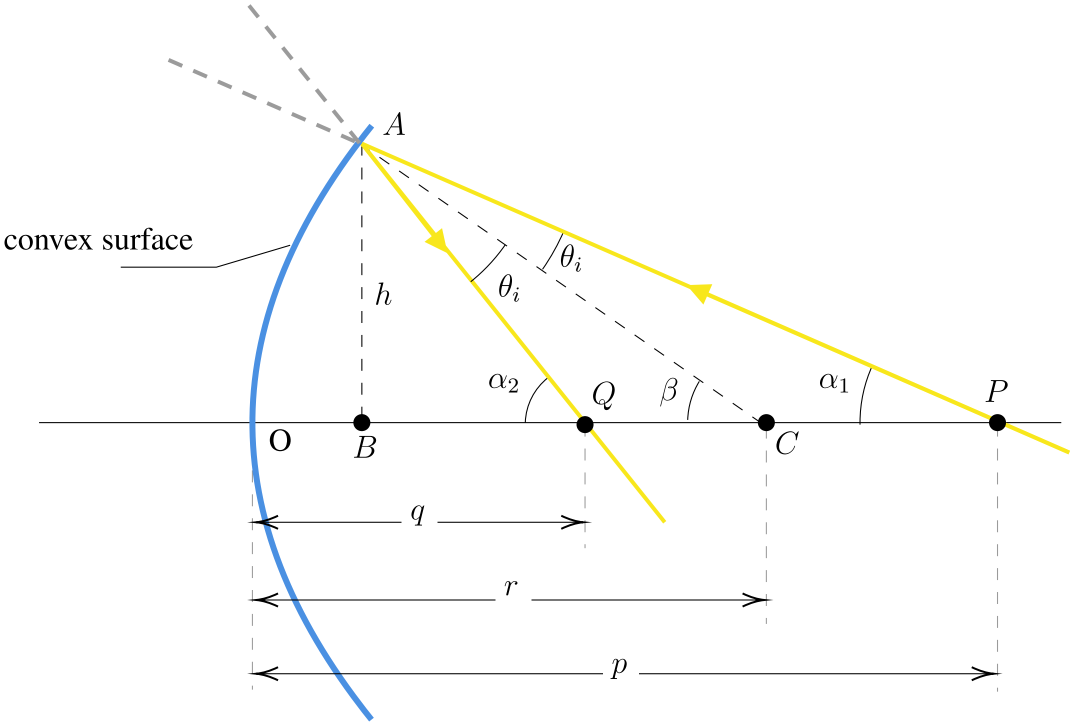
and the image is formed at the point f named the "focal" and its distance from the spherical lens is then given by:

$$\boxed{f=q\cong\frac{r}{2}} \tag{10.5.179}$$

and is named "focal length". We also get the ratio $r/2$ if we make q tend to infinity (that means the curvature radius is put to an infinite value).

So the bigger is r, the bigger is C (smaller curvature = wider angle photo) and therefore the bigger is q.

The relation obtained above is also valid for a convex surface. Indeed, we simply need pull the lines representing the light rays beyond the concave surface to see that the object of study is the same to a given symmetry:



The only difference between the concave and convex surface is that in the case of the convex surface, the reflected image of the object appears as if it is behind the surface (at the equivalent of to the point P). This leads us to define the following terminology:

Definitions (#10.5.494):

D1. A "virtual image" is a term used in optics to refer to any image formed before the exit face of an optical instrument (in the direction of travel of light) and therefore will not be displayed on the projection screen. For a thin converging spherical lens an object placed between the object focus and the optical center of the lens will give a right (not inversed) virtual image. This is particularly the case of an optical system used as a magnifying glass, which provides a magnified image of the object observed through the lens.

D2. A "real picture" is a term used in optics to refer to any image formed after the exit face of an optical instrument (in the direction of travel of light). For a thin spherical converging lens an object placed against the object focus of the lens will give a real (inversed) image.

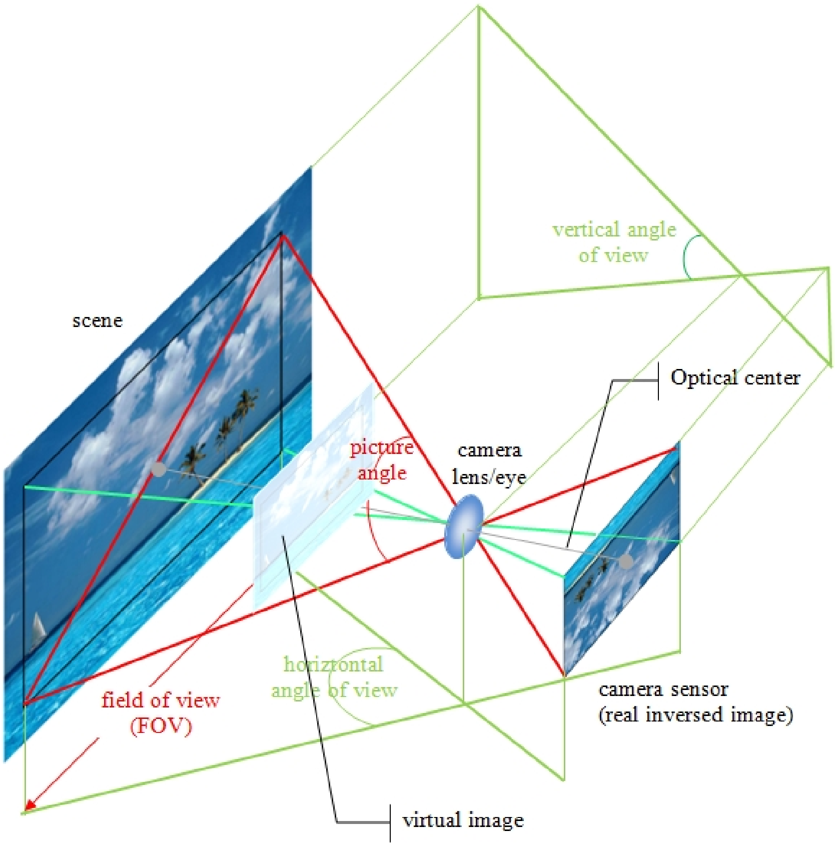
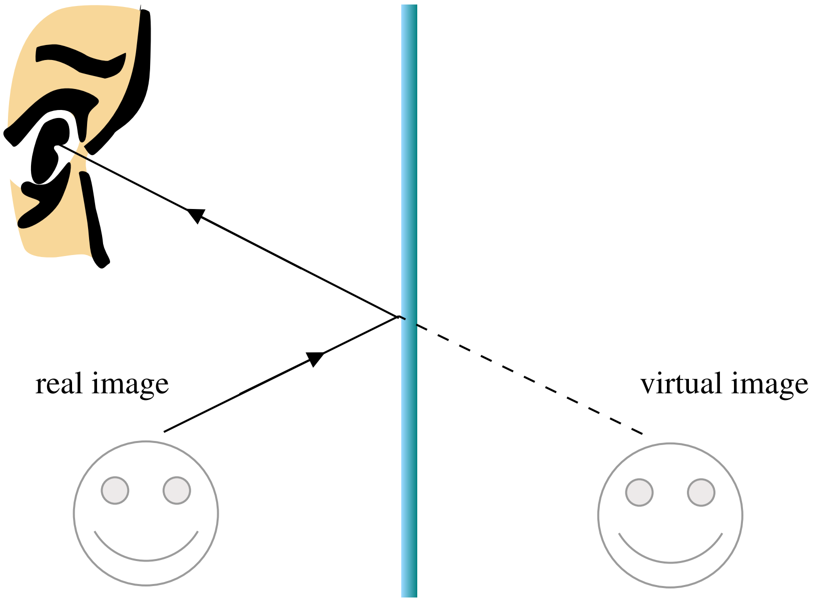

vertical angle of view
scene
Optical center
picture angle
camera lens/eye
horizontal angle of view
field of view (FOV)
camera sensor (real inversed image)
virtual image

Figure 10.176: **Why images are reversed when passing through a thin spherical lens (for simplification purposes the focal in the drawing above is supposed to be inside the lens...)**

In the context of flat mirrors[19] the meaning is quite different:


real image
virtual image

---

[19] When we stand in front of a mirror, our mirror image flips right with left, left with right. But never top with bottom, bottom with top. The reason why is explained by the fact that this is an act of our brain! Our brain always generates a frame of reference with a vertical axis. Hence, we see things flipping vertically. This is because our brain is conditioned and trained to always flip images in the direction which demands less distortion.

obviously for flat mirrors the distance p of the real object to the mirror will have virtually the same distance q that has the "virtual object" to this same mirror so that virtually we could write $p=q$. For flat mirror the image is obviously upright and it produces and apparent left-right reversal!

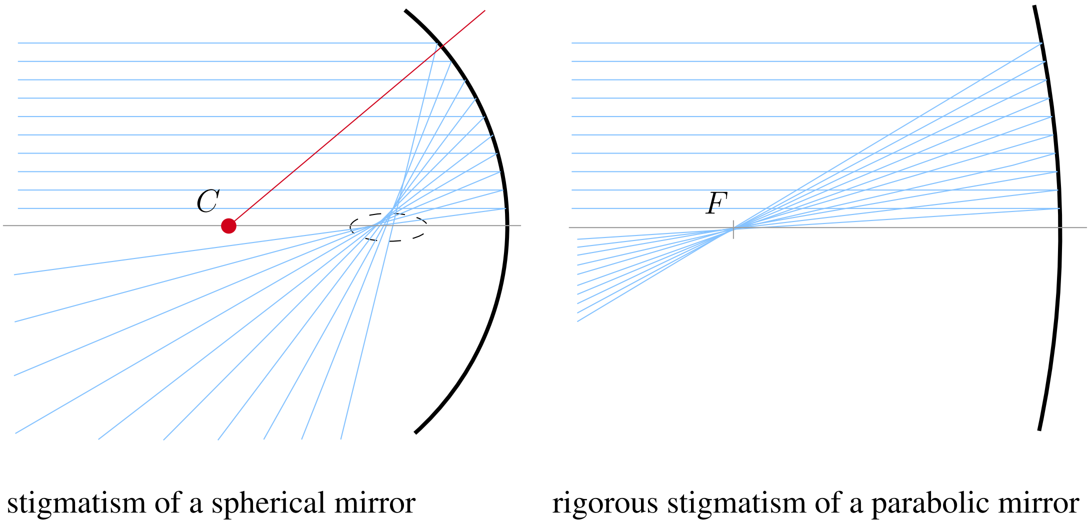
10.5.5.1 Stigmatism

If the opening (wide angle) of the mirror is large, so it receives steeply inclined rays, the above Descartes formula isn't anymore, we know it by construction, a good approximation. There is in this case no more well defined punctual image of a "point object" but an infinite number of them: consequently the image of a large object appears blurred since the images are superimposed. This effect is named "spherical aberration" and part of the optical axis which contains all the reflected images is then named the "caustic reflection". The spherical aberration can not be eliminated, but a suitable design of the surface allows to remove it for certain positions on the optical axis named "stigmatic". For example, in our previous study case, it is clear (by geometrical construction) that if we put P on C then the point C becomes the "stigmatic point". We say then that the point is "strictly stigmatic".

Remark

If we limit ourselves to the rays close to the optical axis and with a slight inclination relatively to it (these rays are said to be "paraxial"), the rays almost all intersect at the same image point: there is "approached stigmatism". This case is named the "Gaussian approximation".

By cons, for the parabolic mirror all the rays converge on the focus of the mirror where is concentrated the light energy received by the mirror. Conversely, we place the filament of a lamp at the focus of a parabolic mirror to get far-reaching headlights (typically lighthouse having no Fresnel lens). We also give a parabolic shape to antennas needing to receive radio waves. For television broadcast by satellite as it works at the centimetre waves (frequency of several GHz) a focal distance of one meter is suitable for the antenna (i.e. this applies to telescopes and radio telescopes).

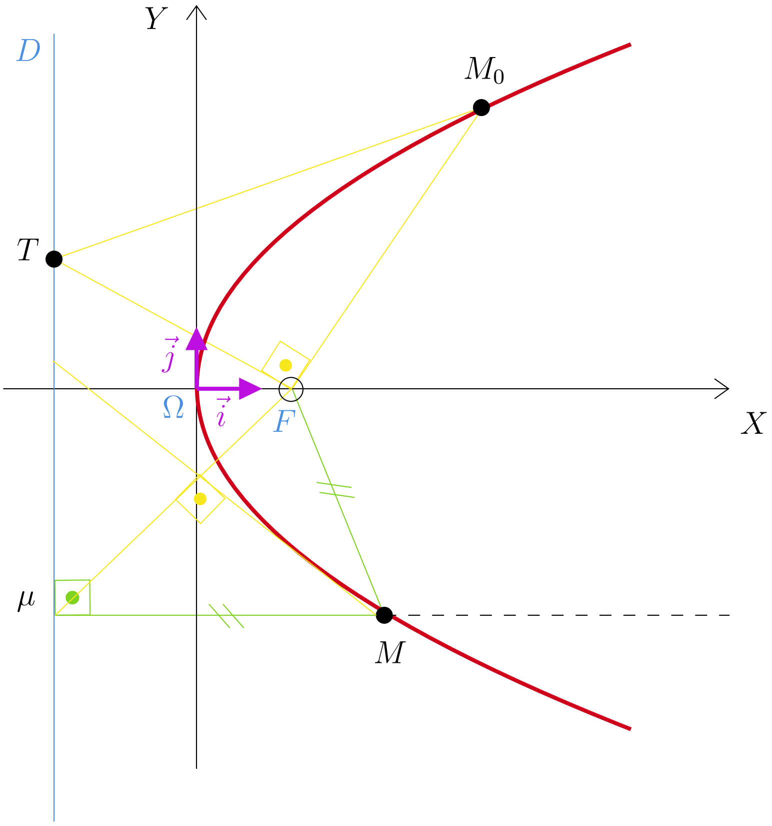


stigmatism of a spherical mirror rigorous stigmatism of a parabolic mirror

Figure 10.177: Representation of the concept of stigmatism

The idea to prove mathematically that the focus of a parabola is the rigorous stigmatic point is the following[20]:

Let us recall the following figure we used in our study of conical in the section of Analytical Geometry:



We have added on the figure the point μ that is the orthogonal projection of the point M (point of incidence of the light beam) and also the tangent to the parabola at the point M. If we can prove that the tangent to M is the mediator of segment $\overline{F\mu}$, then we also prove that the angle of incidence and reflection are equal. Therefore as we already know that incidence and reflection angles are equal, then we prove indirectly that on any point M of the parabola the incident ray comes on F.

Proof 10.2.4. Let us consider the equation of the parabola relatively to the focal as proved in the section of Analytical Geometry:

$$y^2-2hx=0 \qquad (10.5.180)$$

We have also proved in the section of Analytical Geometry that relatively to Ω the Focal is at the position $F(h/2,0)$ and the direction had for equation:

$$x+\frac{h}{2}=0 \qquad (10.5.181)$$

---

[20] That the shape of the caustic of a half-circular mirror is a "nephroid" is of minor interest so we won't do the proof in this textbook for now.

We get the equation of the tangent at $M(x_0,y_0)$ by the derivative at the same point (caution! ... remember the particular orientation of the parabola!):

$$x'_{y_0}=\frac{(x-x_0)}{(y-y_0)}=\left(\frac{y'}{2h}\right)'_{y_0}=\frac{2y_0}{2h}=\frac{y_0}{h}\Rightarrow(x-x_0)-(y-y_0)\frac{y_0}{h}=0 \tag{10.5.182}$$

Which can also be written:

$$hx-hx_0-y_0y+y_0^2=0\Leftrightarrow hx_0-hx-y_0^2+y_0y=0 \tag{10.5.183}$$

and knowing that:

$$y^2-2hx=0\Leftrightarrow y_0^2=2hx_0 \tag{10.5.184}$$

So we get the equation of the tangent:

$$hx_0-hx-y_0^2+y_0y=hx_0-hx-2hx_0+y_0y=y_0y-h(x+x_0)=0 \tag{10.5.185}$$

On of the vector of the tangent is therefore as proved in the section of Analytical Geometry during our study of the parametric equation of a the line:

$$\vec{v}=y_0\vec{i}+p\vec{j} \tag{10.5.186}$$

where in our case of the parabola, p is equal to h.

On the other hand, we have (this is easily verified by taking $x_0=0$):

$$\overrightarrow{MF}=\left(\frac{h}{2}-x_0\right)\vec{i}-y_0\vec{j}\quad\text{and}\quad\overrightarrow{M\mu}=-\left(\frac{h}{2}+x_0\right)\vec{i} \tag{10.5.187}$$

So we have the scalar product:

$$\vec{v}\circ\overrightarrow{MF}=-y_0\left(x_0+\frac{h}{2}\right)=\vec{v}\circ\overrightarrow{M\mu} \tag{10.5.188}$$

as the vectors $\overrightarrow{MF}$ and $\overrightarrow{M\mu}$ have the same norm by definition of the parabola in Analytical Geometry[21], we deduce that the vector $\vec{v}$ (giving the direction of the tangent) leads the line bisecting the angle of the $\overrightarrow{MF}$ and $\overrightarrow{M\mu}$ and therefore by extension that the tangent M is the mediator of $\overline{M\mu}$.

■ Q.E.D.

---

[21] Yes, for recall a "parabola" is the set of all points which are equidistant from a point, called the focus, and a line, named the "directrix"

Figure 10.178: **Parabolic antenna for reception or emission (source: OpenStax)**

Let us also calculate the focus position of a parabolic reflector whose equation is given by (see section Analytical Geometry page 2983):

$$\frac{x^2}{100}+\frac{y^2}{100}=\frac{z}{4} \tag{10.5.189}$$

We fix for example x to 0 then it remains:

$$\frac{y^2}{100}=\frac{z}{4} \tag{10.5.190}$$

After rearranging we get:

$$y^2=25z \tag{10.5.191}$$

As:

$$25=2h \tag{10.5.192}$$

we get immediately:

$$h=12.5 \tag{10.5.193}$$

And we have prove in the section of Analytical Geometry that:

$$F=\left(\frac{h}{2},0\right)=(6.25,0) \tag{10.5.194}$$

A circular paraboloid is theoretically unlimited in size. Any practical reflector uses just a segment of it. Often, the segment includes the vertex of the paraboloid, where its curvature is greatest, and where the axis of symmetry intersects the paraboloid. However, if the reflector is used to focus incoming energy onto a receiver, the shadow of the receiver falls onto the vertex of the paraboloid, which is part of the reflector, so part of the reflector is wasted. This can be avoided by making the reflector from a segment of the paraboloid which is offset from the vertex and the axis of symmetry:

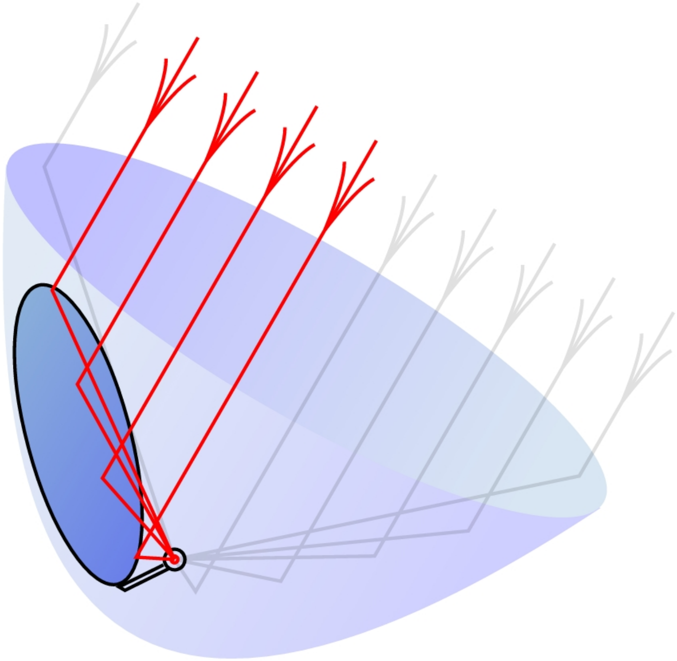


Figure 10.179: Off-axis receiver dish construction principle (source: Wikipedia)

The receiver is still placed at the focus of the paraboloid, but it does not cast a shadow onto the reflector. The whole reflector receives energy, which is then focused onto the receiver. This is frequently done, for example, in satellite-TV receiving dishes, and also in some types of astronomical telescope:

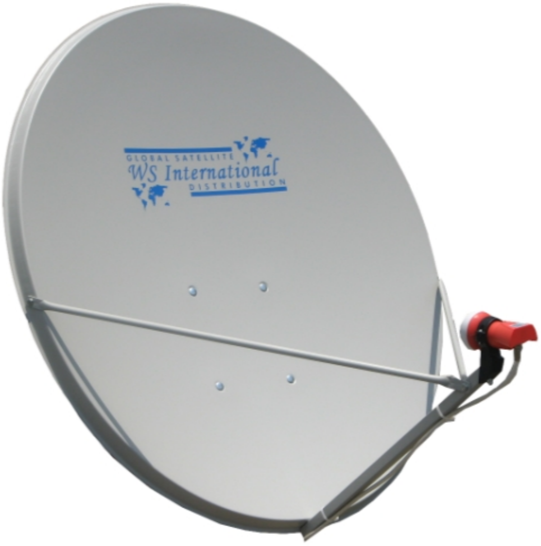

GLOBAL SATELLITE
WS International
DISTRIBUTION

Figure 10.180: Off-axis satellite dish

10.5.5.2 Lenses

Let us now do a similar study to that carried out earlier, with the same properties of symmetry and defects, but the "spherical dioptres boundaries" that is a spherical boundary between to mediums (interesting results regarding the study of the eye). The results will be helpful to us before tackling the study of spherical lens (the traditional magnifying glass).

So we will now consider the refraction of the light during the passage through a spherical surface separating two medium with absolute refractive indices n_1 and n_2 (see figure below):

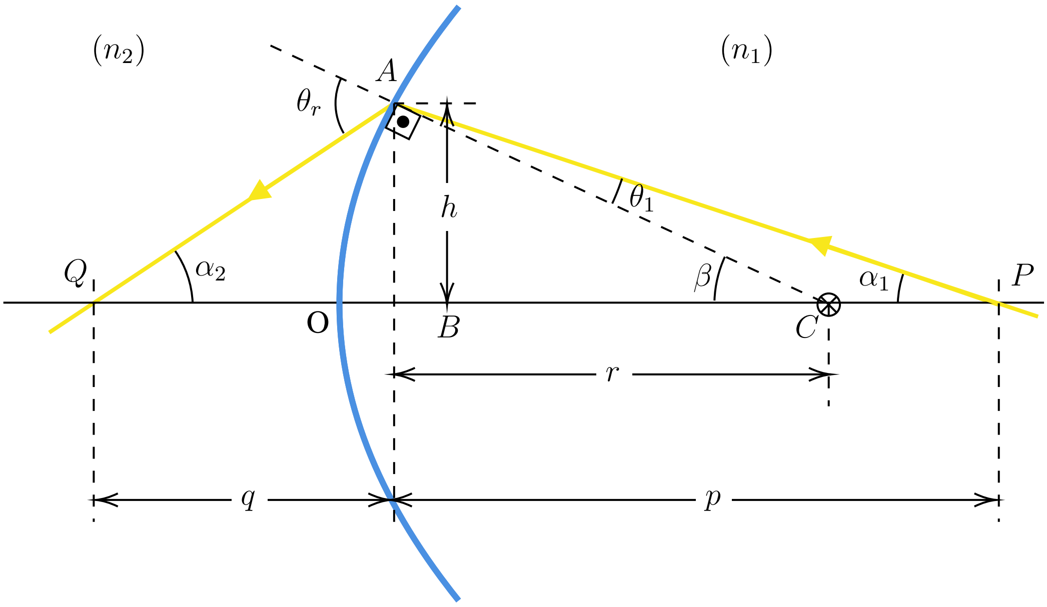


Figure 10.181: A biconvex lens in real life... (source: Wikipedia)

where for recall the center of curvature C (see section Differential Geometry page 2866) is the center of the spherical surface of the figure below and the top O is the pole of the spherical cap.

The basic geometric elements are the same as those specified for spherical surfaces. We therefore consider initially a concave refractive surface (concave dioptre) and observing that the "object distance" is located at the opposite of the other points, we must adopt a sign convention to put this observation in evidence in the equations. Thus, q will be defined as a negative value.

An incident ray such as $\overline{PA}$ is refracted as following $\overline{AQ}$ and thus cross the optical axis Q. We see from the figure that:

$$\beta=\theta_i+\alpha_i \qquad \text{and} \qquad \beta=\theta_r-\alpha_2 \tag{10.5.195}$$

We have the Snell-Descartes' law:

$$\frac{n_1}{\sin(\theta_r)}=\frac{n_2}{\sin(\theta_i)} \tag{10.5.196}$$

and we will admit as for spherical surfaces that the ray a only little bit inclined rays. Under these conditions the angles $\theta_i,\theta_r,\alpha_1,\alpha_2$ and β are very small and we can write using the Maclaurin series expansion (see section Sequences and Series page 1703):

$$\sin(\theta_i)\cong\theta_i \qquad \text{and} \qquad \sin(\theta_r)\cong\theta_r \tag{10.5.197}$$

so that the Snell-Descartes' law is now written:

$$n_1(\beta-\alpha_1)=n_2(\beta+\alpha_2) \tag{10.5.198}$$

From the figure, we can do the approximations:

$$\begin{aligned}\tan(\alpha_1)&\cong\alpha_1\cong\frac{h}{p}\\ \tan(\alpha_2)&\cong\alpha_2\cong\frac{h}{q}\\ \tan(\beta)&\cong\beta\cong\frac{h}{r}\end{aligned} \tag{10.5.199}$$

so that by substituting in the approximation of the Snell-Descartes' law we find elementary simplification:

$$n_1\left(\frac{h}{r}-\frac{h}{p}\right)=n_2\left(\frac{h}{r}+\frac{h}{q}\right)\Leftrightarrow\left(\frac{1}{r}-\frac{1}{p}\right)=n_2\left(\frac{1}{r}+\frac{1}{q}\right) \tag{10.5.200}$$

hence for a concave surface:

$$\boxed{\frac{n_1}{p}+\frac{n_2}{q}=\frac{n_1-n_2}{r}} \tag{10.5.201}$$

The "object focal F_0" also named "first focal point" of a refractive spherical surface is the position of a punctual object on the optical axis such that the refracted rays are parallel to the optical axis, which is equivalent to form the image of the point at infinity, where $q=+\infty$.

The distance of the object to the spherical surface is then named "focal distance object" and we denote it by f_0. By putting $p=f_0$ and $q=+\infty$, then we have for the concave case:

$$\frac{n_1}{f_0}=\frac{n_1-n_2}{r}\Rightarrow f_0=\left(\frac{n_1}{n_1-n_2}\right)r \tag{10.5.202}$$

The focal length f_0 is positive and the system is the convergent when the object focus is real, is placed in front of the spherical surface. When the object focus is virtual, the focal length f_0 is negative and the system is then divergent.

Similarly, if the incident rays are parallel to the optical axis, which is equivalent as to have a very distant object form the spherical surface $p=+\infty$, the refracted rays pass through a point F_i of the optical axis named "image focus" or "second focal point" (again with the same problems of stigmatims).

In this case the distance of the spherical surface to the image is named "focal image distance" and we denote it by f_i. By putting $p=+\infty$ and $q=f_i$ then we have for the concave case:

$$\frac{n_2}{f_i}=\frac{n_1-n_2}{r}\Rightarrow f_i=\left(\frac{n_1}{n_1-n_2}\right)r \tag{10.5.203}$$

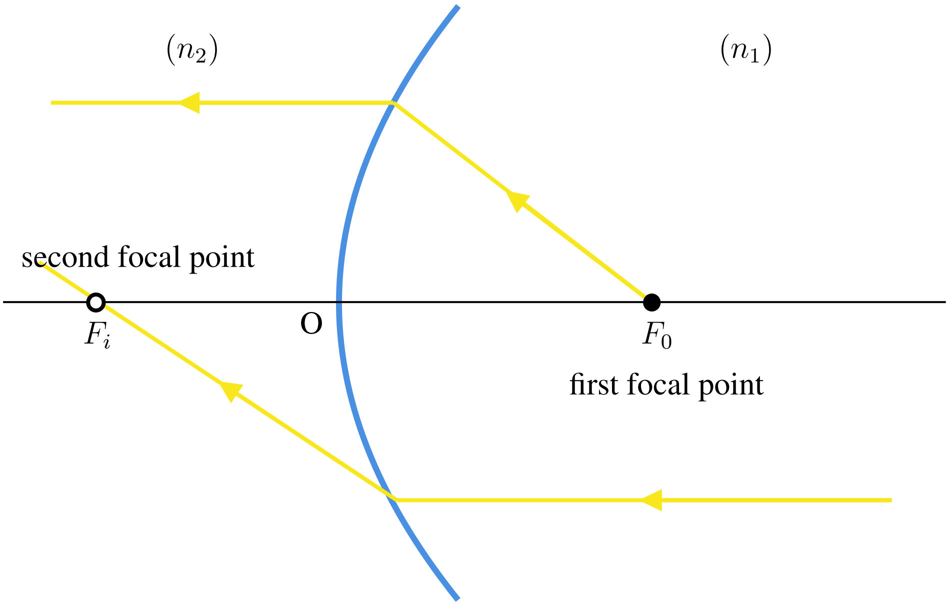


Figure 10.182: Position of first and second focal points in a dioptre

By mixing the two previous relations, we have a useful result in practice:

$$\boxed{\left.\begin{array}{l} f_0=\dfrac{n_1}{n_1-n_2}r \\ f_i=\dfrac{n_2}{n_1-n_2}r \end{array}\right\} \Rightarrow \dfrac{f_0}{f_i}=\dfrac{n_1}{n_2}=n} \tag{10.5.204}$$

So if we know the relative refractive index and one of the focal lengths, we can infer the other. Or (and we can of course make also other combinations):

$$\left.\begin{array}{l} f_0=\dfrac{n_1}{n_1-n_2}r \\ f_i=\dfrac{n_2}{n_1-n_2}r \end{array}\right\} \Rightarrow f_0-f_i=-r \tag{10.5.205}$$

Now let us take look at the type of reflective surfaces and refractions we were waiting for: the lenses!

Definition (#10.5.495): A "lens" is a transmissive optical device that focuses or disperses a light beam by means of refraction. A simple lens consists of a single piece of transparent material (generally spherical), while a compound lens consists of several simple lenses (elements), usually arranged along a common axis. Lenses are made from materials such as glass or plastic, and are ground and polished or moulded to a desired shape. A lens can focus light to form an image, unlike a prism, which refracts light without focusing. Devices that similarly focus or disperse radiation other than visible light are also named "lenses", such as microwave lenses, electron lenses or acoustic lenses.

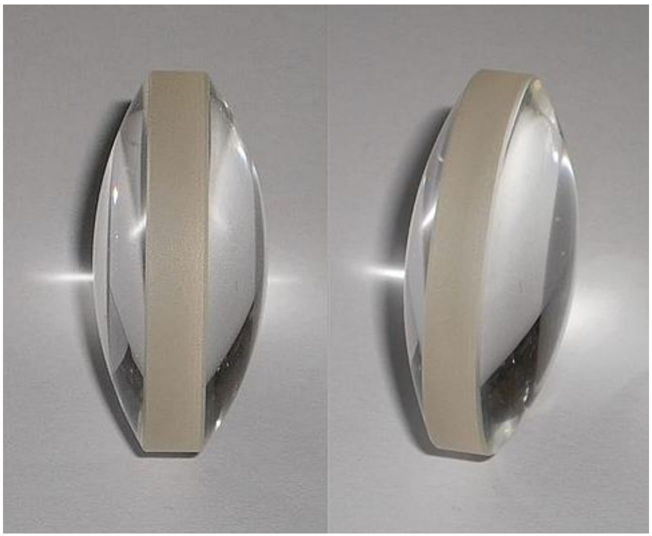


Figure 10.183: **Concept of spherical dioptre**

The study of optical lenses is a huge and subtle subject and our only purpose in this textbook is to give the mathematical basic properties of some elementary lenses shapes. As we will show it through some illustration, even traditional camera have already some huge complications to get qualitative results for the professional photographer as illustrated in the figure below:

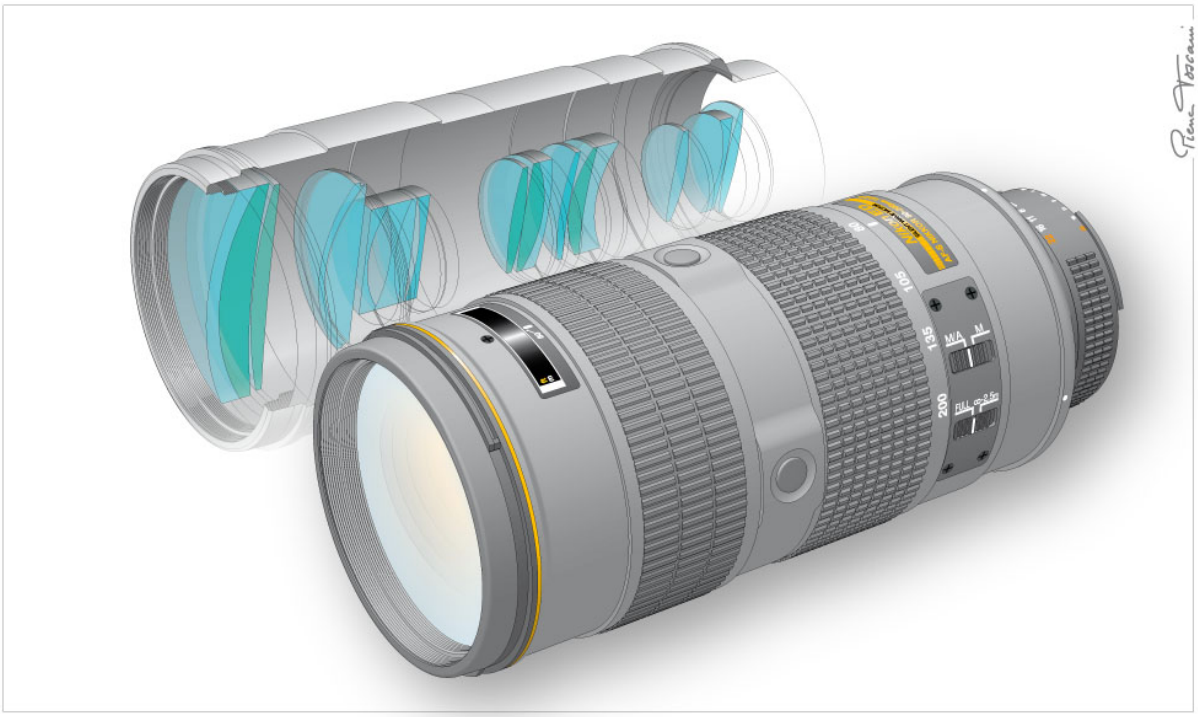


Figure 10.184: **Zoom-Nikkor AF-S 80-200 mm f/2.8D IF-ED (source: http://www.pierretoscani.com, author: Pierre Toscani)**

or even better (optical system also existing for smartphones!):

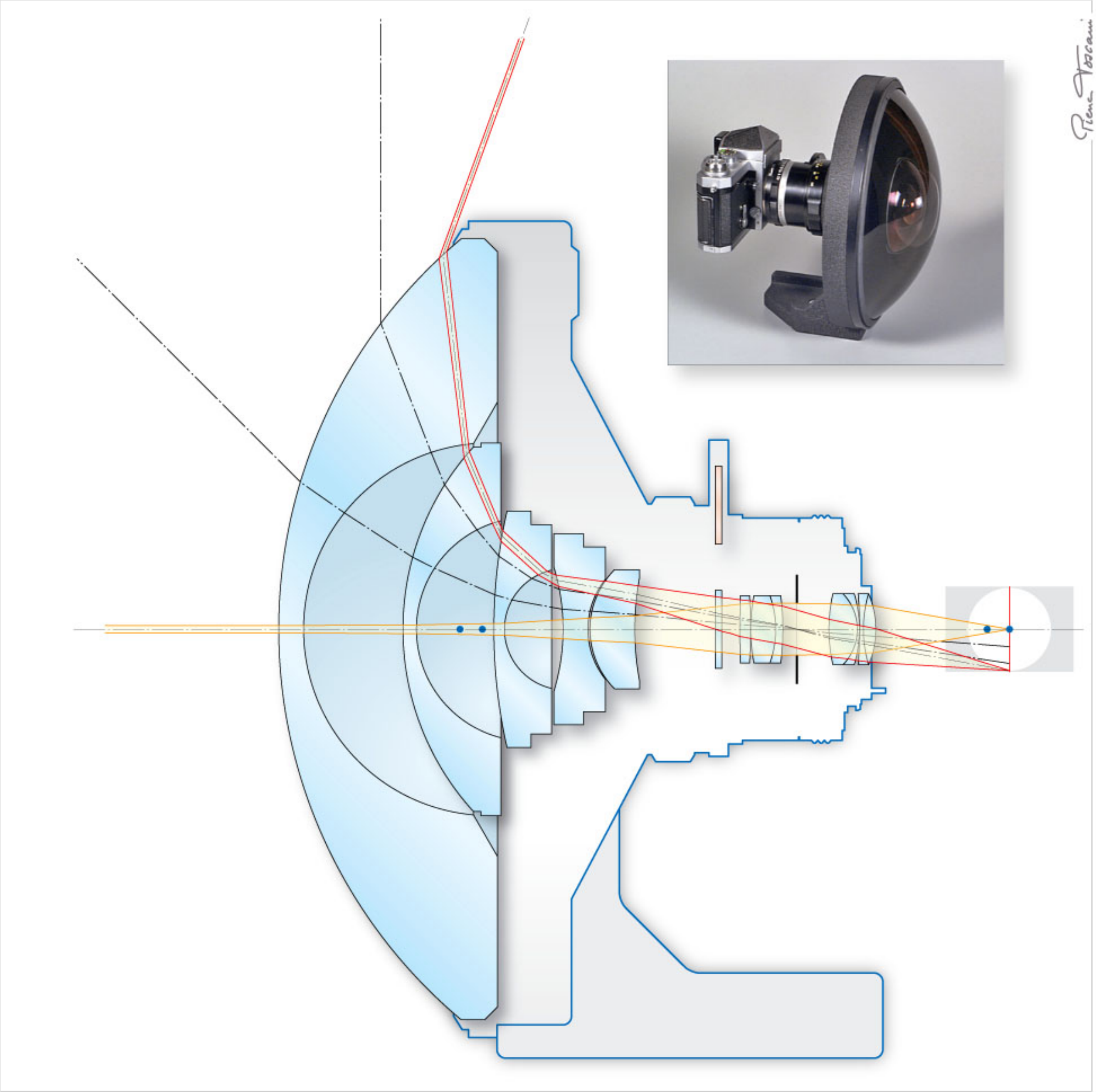


Figure 10.185: Fisheye-Nikkon 6 mm f/2.8 simplified cut (source: http://www.pierretoscani.com, author: Pierre Toscani,)

An incident ray thus undergoes two refractions at the crossing of the lens. Assume for simplicity that the media of both sides of the lens are identical and their absolute refraction index equal to 1 (air or vacuum for example). We will also consider only thin lenses in this textbook, that is to say whose thickness is very small compared to the radii of curvature:

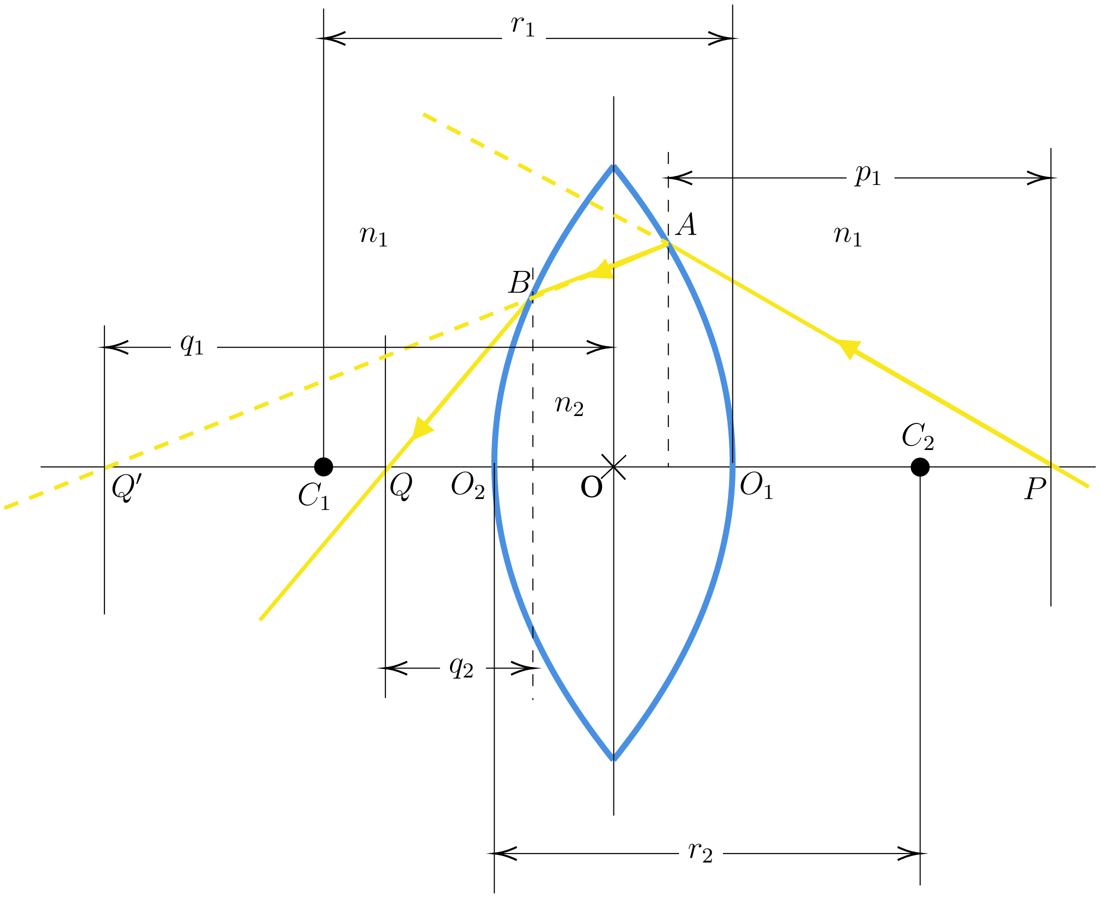


Figure 10.186: Technical representation of a lens

The optical axis is the line determined by the two centers $\overline{C_1C_2}$. We seek to establish a relation between the position of P and Q from easily measured physical parameters!

We will consider for the analysis that the image formed after the refraction on the first surface is the object for the refraction on the second surface.

Let us consider the incident ray $\overline{PA}$ through P. When passing the first surface, the incident ray is refracted following the radius $\overline{AB}$ and continues virtually to Q' according to the behaviour of a convex spherical surface. So we apply the relation proved just earlier above for the convex diopter:

$$\frac{n_1}{p_1}+\frac{n_2}{q_1}=\frac{n_2-n_1}{r_1} \qquad (10.5.206)$$

In B the ray undergoes a second refraction and becomes the radius $\overline{BQ}$ according to the behaviour of a concave spherical surface. We can imagine that the incident beam in B come from a virtual point P' (not show in the figure above) that is on the optical axis and immersed in the material virtually extended to the right of the lens (that is the difficult side of this proof ... you have to imagine that!).

So we have to apply the relation proved earlier above for the concave refractive surface (dioptre) but by being careful this time to the order of absolute refractive indices (subtle trap!) and by imagining that the point:

$$\frac{n_2}{p_2}+\frac{n_1}{q_2}=\frac{n_1-n_2}{r_2} \tag{10.5.207}$$

As we consider the lens is surrounded by air (unitary refractive index), then we have:

$$\frac{1}{p_1}+\frac{n}{q_1}=\frac{n-1}{r_1} \qquad \frac{n}{p_2}+\frac{1}{q_2}=\frac{1-n}{r_2} \tag{10.5.208}$$

We will now assume that the thickness of the lens tends to zero. In other words his two radii tend to infinity. Then we have:

$$\left.\begin{array}{l} \dfrac{1}{p_1}+\dfrac{n}{q_1}=\dfrac{n-1}{r_1}=_{r_1\rightarrow+\infty}0 \\ \dfrac{n}{p_2}+\dfrac{1}{q_2}=\dfrac{1-n}{r_2}=_{r_1\rightarrow+\infty}0 \end{array}\right\} \Rightarrow \frac{1}{p_1}+\frac{1}{q_1}=\frac{n}{p_2}+\frac{1}{q_2} \tag{10.5.209}$$

By identifying term to term and remembering that what is left from the origin O is negative, then we have:

$$p_2\cong -q_1 \tag{10.5.210}$$

while subtly ignoring the other two terms ... (we understand now easily why this proof is often omitted in the literature ...).

Therefore, both prior previous relations become:

$$\frac{1}{p_1}+\frac{n}{q_1}=\frac{n_1}{r_1} \qquad \frac{n}{-q_1}+\frac{1}{q_2}\cong\frac{1-n}{r_2} \tag{10.5.211}$$

Summing it comes finally:

$$\frac{1}{p_1}+\frac{1}{q_2}\cong(n-1)\left(\frac{1}{r_1}-\frac{1}{r_2}\right) \tag{10.5.212}$$

and that is often denoted in the following form, name "Descartes's first formula for thin lenses" or simply "thin lens equation":

$$\boxed{\frac{1}{p}+\frac{1}{q}\cong(n-1)\left(\frac{1}{r_1}-\frac{1}{r_2}\right)} \tag{10.5.213}$$

While being careful to rehabilitate the figure above such that it becomes:

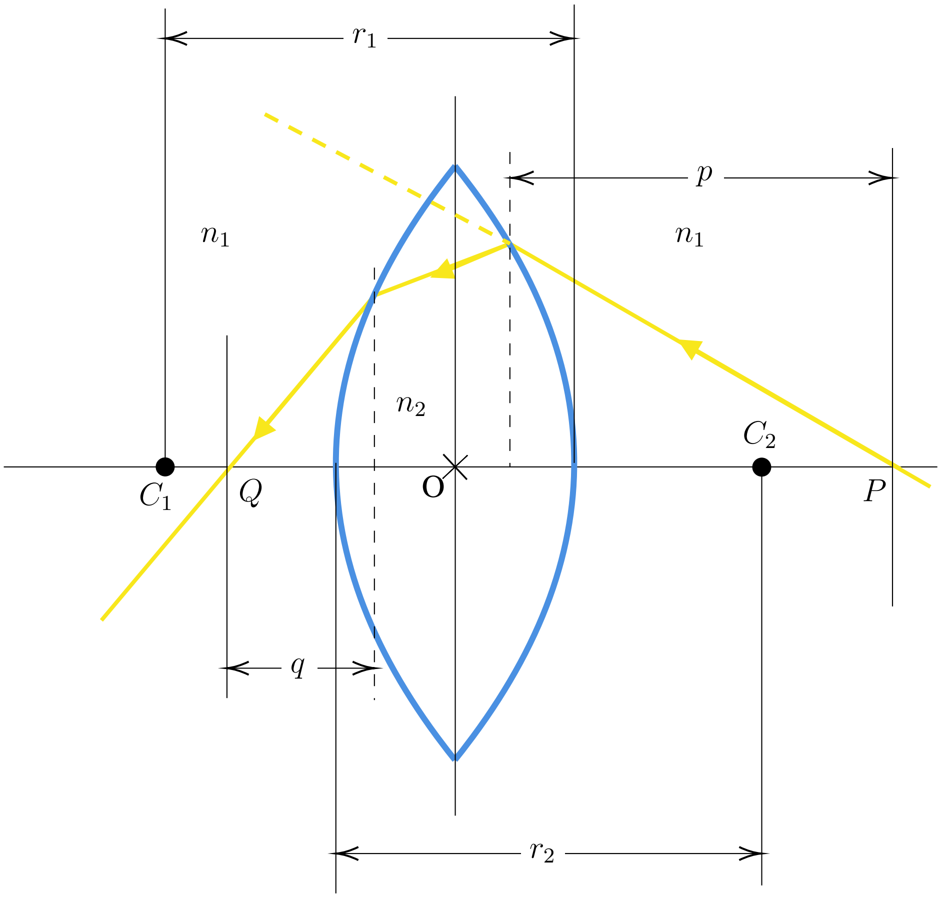


Figure 10.187: **Thin lens equation corresponding technical figure**

By writing this formula it should be applied to r_1, r_2 the sign convention we have set, that is to say, the rays are positive for a concave surface and negative for a convex surface, as seen from the side on which the light hits the lens. Thus, if the two radii are the same, we have:

$$\frac{1}{p}+\frac{1}{q}\cong(n-1)\left(\frac{1}{r}-\frac{1}{(-r)}\right)=(n-1)\frac{2}{r} \tag{10.5.214}$$

The right term of the thin lens equation is only a constant depending on the physical characteristics of the lens and it customary to name it "refractive power" or "dioptric power" and whose unity is the "diopter" and denoted:

$$\frac{1}{p}+\frac{1}{q}\cong\mathcal{D} \tag{10.5.215}$$

Converging lenses have positive optical power, while diverging lenses have negative power. When a lens is immersed in a refractive medium, its optical power and focal length change.

Notice that if we put the source at infinity such that $p=+\infty$ then all light rays will focus on the same point q, therefore denoted f_0 as we have:

$$\frac{1}{f_0}\cong(n-1)\left(\frac{1}{r}-\frac{1}{(-r)}\right)=(n-1)\frac{2}{r} \tag{10.5.216}$$

Hence:

$$\boxed{f_0=\frac{r}{2}(n-1)} \tag{10.5.217}$$

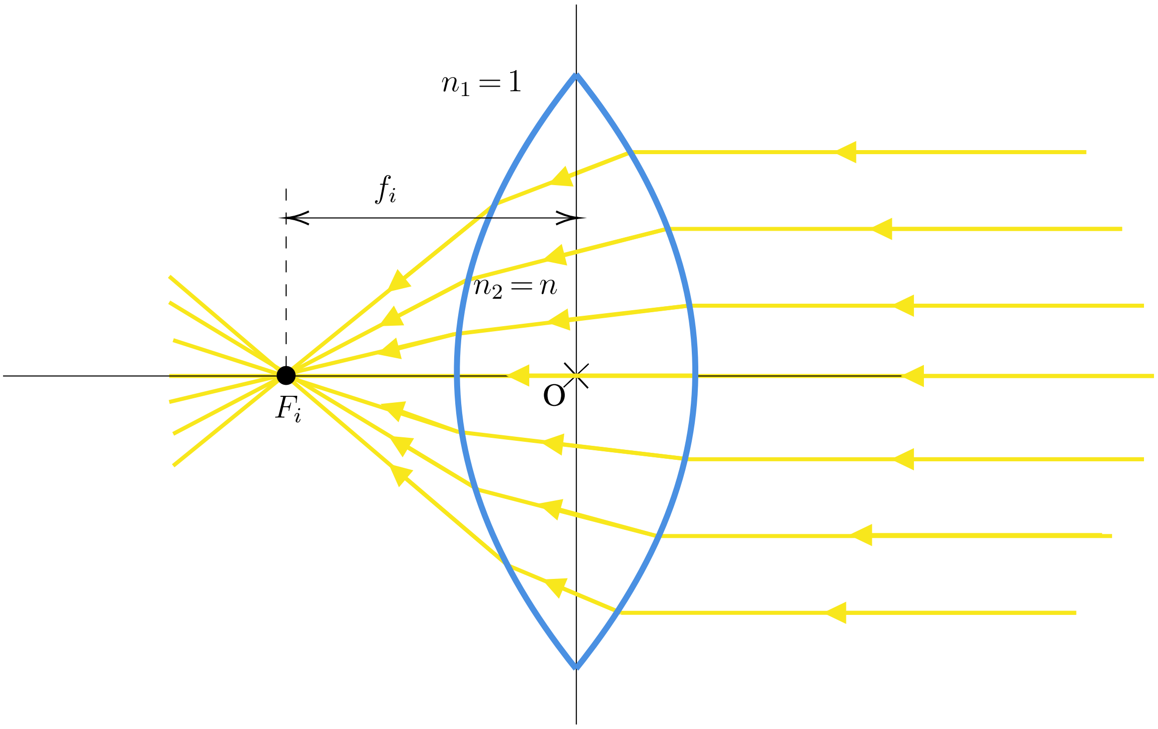


Figure 10.188: **Focal of a thin lens with parallel entering light rays**

Now it's interesting to see that the concept is reversible. That means if you consider the source as being the focus, i.e. such that (by symmetry) $q = f_0$, the by using the previous relation, we get:

$$\frac{1}{p}+\frac{1}{q}=(n-1)\frac{2}{r}\Rightarrow\frac{1}{f_0}+\frac{1}{q}=(n-1)\frac{2}{r} \tag{10.5.218}$$

Therefore:

$$\frac{1}{q}=(n-1)\frac{2}{r}-\frac{1}{f_0}=(n-1)\frac{2}{r}-(n-1)\frac{2}{r}=0 \tag{10.5.219}$$

Hence $q = +\infty$ and the target is at infinity. So ray tracing is invertible because if the source is a the focus, the all rays go out of the lens parallel at infinity!

The following figure can help to understand (play with the values of q, p, r in the previous relations to understand it completely):

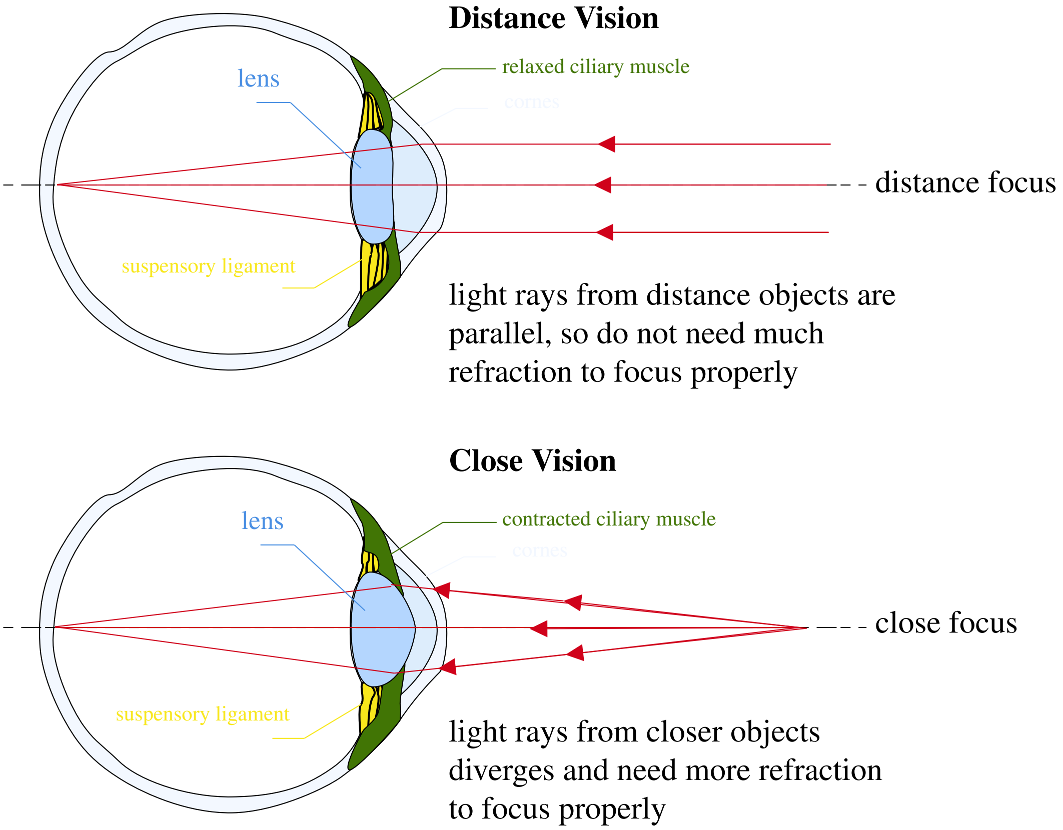

Distance Vision
lens
relaxed ciliary muscle
cornea
distance focus
suspensory ligament
light rays from distance objects are parallel, so do not need much refraction to focus properly
Close Vision
lens
contracted ciliary muscle
cornea
close focus
suspensory ligament
light rays from closer objects diverges and need more refraction to focus properly

Figure 10.189: Human eye adaptative refractive power mechanism

Remark

An eye that has too much or too little refractive power to focus light onto the retina has a refractive error. A myopic eye has too much power so light is focused in front of the retina. Conversely, a hyperopic eye has too little power so when the eye is relaxed, light is focused behind the retina. An eye with a refractive power in one meridian that is different from the refractive power of the other meridians has astigmatism. Anisometropia is the condition in which one eye has a different refractive power than the other eye.

The point O in the previous figure, is selected to coincide with the "optical center" of the lens. As we know, the optical center has the property of being such that any ray passing through it exit parallel to the direction of the incident ray !! This is an important property because any point of an object lying on one side of a lens (whatever which one because of the symmetry of the lens) will emit light from which some rays will pass through the optical center. Thus allowing to have similar triangles on the left and right of the symmetry axis of the lens and apply Thales' theorem (see section Euclidean Geometry page 2610) to calculate the concept of "magnification" (see further below).

To prove that such a point exists let us consider in the lens below (with horizontal and vertical symmetry):

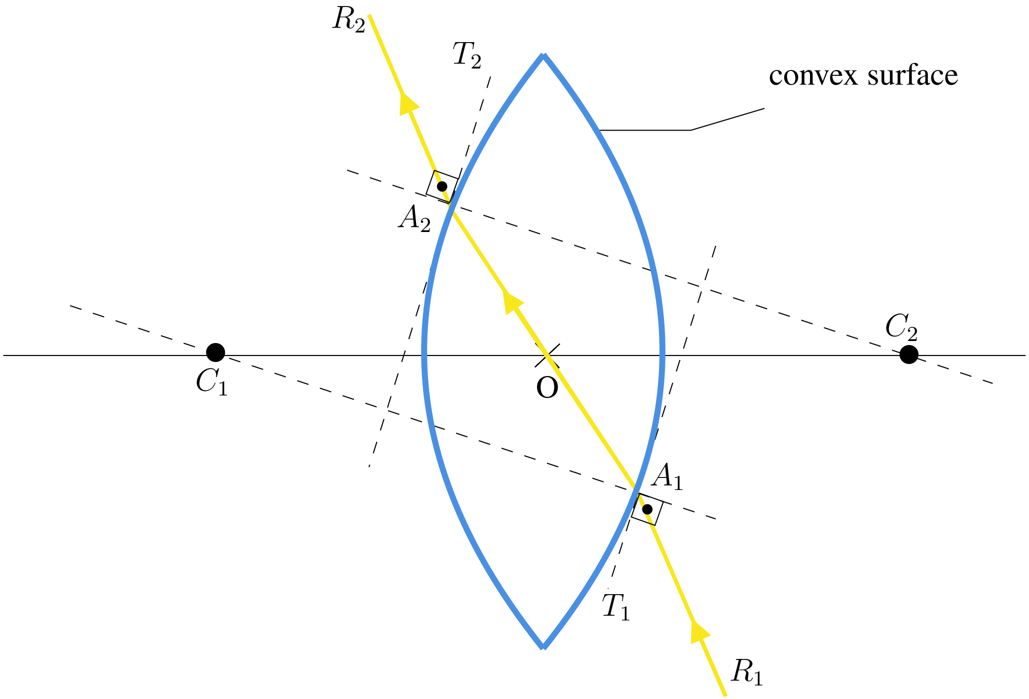


Figure 10.190: **Representation of the property of the optical center of a lens**

Let us consider two parallel curvatures radius:

$$\overline{C_1A_1}=r_1 \qquad \overline{A_2C_2}=-r \tag{10.5.220}$$

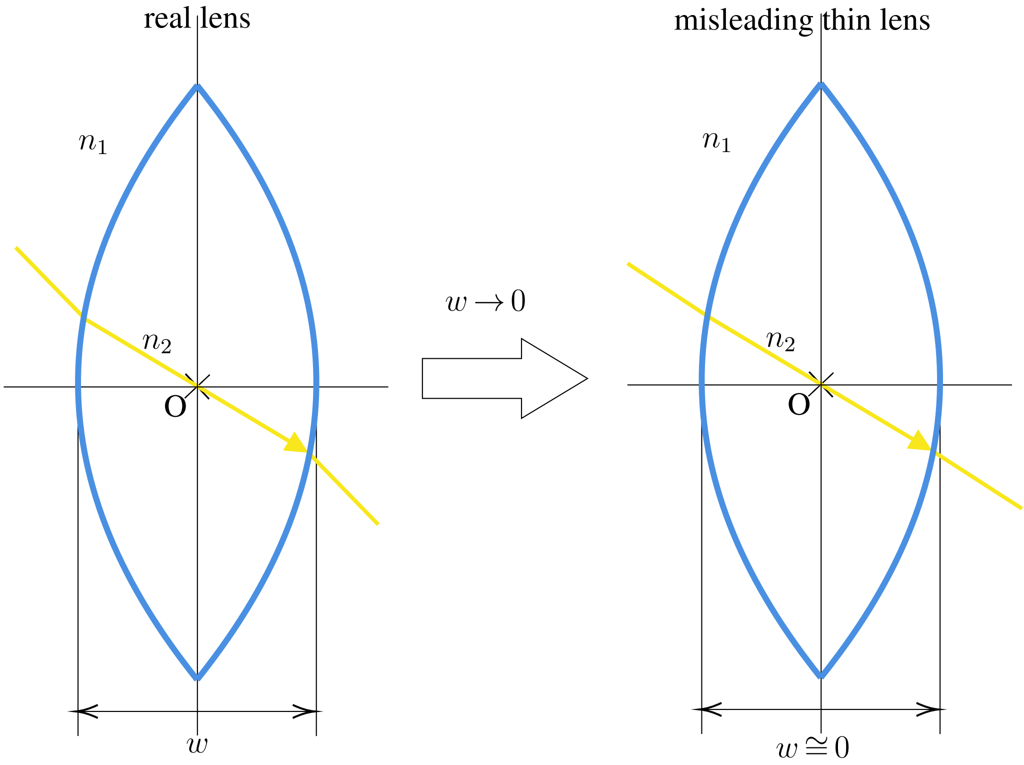
generators of the diopters (elements of the thin spherical lens for recall...) chosen such that the corresponding tangent planes T_1 and T_2 are also parallel.

For the radius $\overline{R_1A_1}$, which direction is such that it is refracted following $\overline{A_1A_2}$, the emerging ray $\overline{A_2R_2}$ is parallel to $\overline{A_1R_1}$ by the horizontal symmetry of the lens. Thus the triangles $C_1A_1\text{O}$ and $C_2A_2\text{O}$ are similar regardless of the "generators radius", thus we see that the position of the optical center O is satisfied by the relation:

$$\frac{\overline{C_1\text{O}}}{\overline{\text{O}C_2}}=\frac{\overline{C_1A_1}}{\overline{A_2C_2}}=-\frac{r_1}{r_2} \tag{10.5.221}$$

and therefore exists independently of the generating radii.

Notice that in the "thin lens" assumption the width is assumed to tend to zero. This is why in many textbook the reader can see the following quite abusive and misleading representation on the right of the figure below:



As in the case of a single diopter, the "focal object" F_0, or "first focal point of a lens" is the position of the object for which the rays emerge parallel to the optical axis ($q = +\infty$) after have passed through the lens. The distance from the lens to the object focus is named the "focal distance object" and we designate it in practice often by the simple letter f for thin lenses.

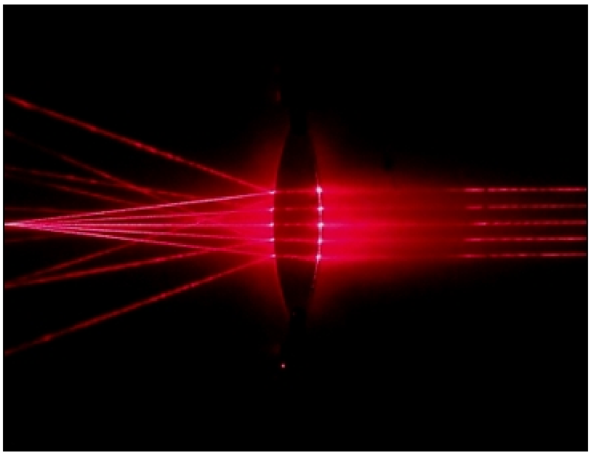


Figure 10.191: **First focal point (source) is here on the left of the thin lens**

Or for those who cannot see well the figure above, there is another more explicit illustration of what happen with the first focal point of a bi-convex lens:

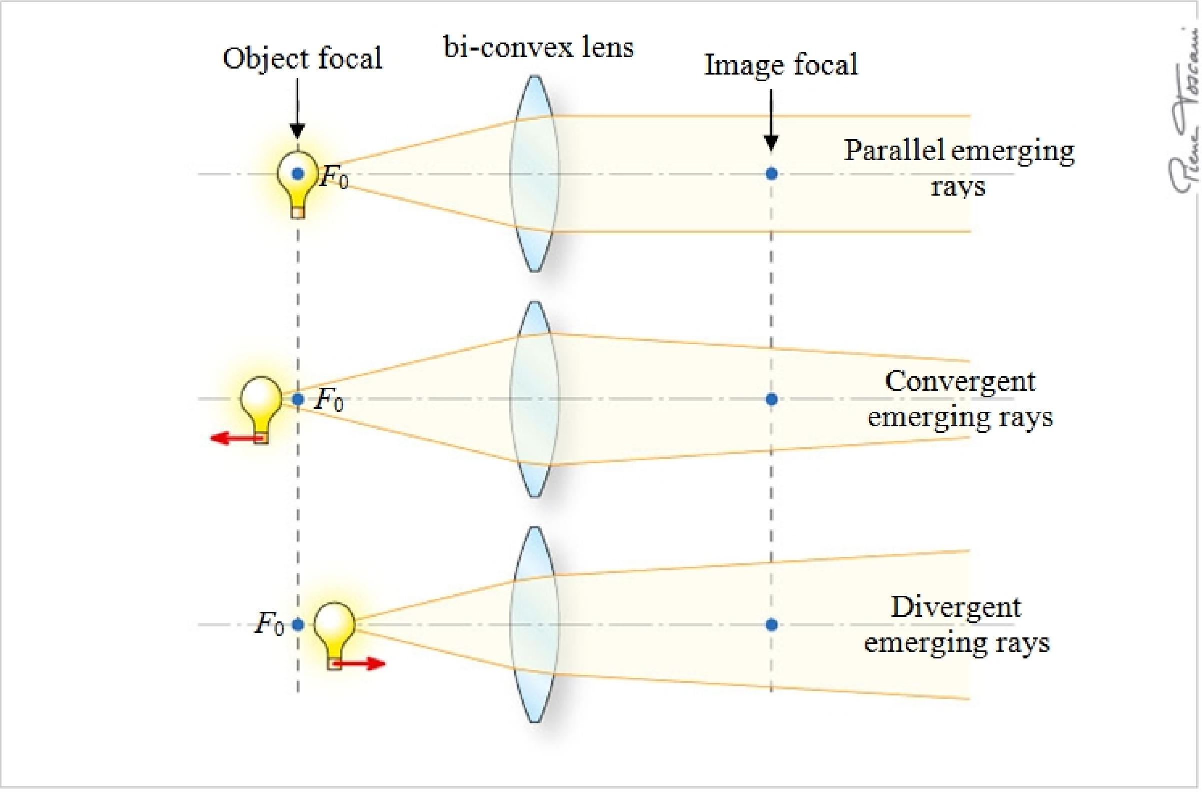

Object focal
bi-convex lens
Image focal
F_0
Parallel emerging rays
F_0
Convergent emerging rays
F_0
Divergent emerging rays
Pierre Toscani

Figure 10.192: **First focal point with bi-convex lens behaviour (source: http://www.pierretoscani.com, author: Pierre Toscani)**

Then by putting $p = f$ and $q = +\infty$ in the equation of spherical thin lenses we get:

$$\frac{1}{p}+\frac{1}{q}=(n-1)\left(\frac{1}{r_1}-\frac{1}{r_2}\right) \tag{10.5.222}$$

We then get the following focal distance object in the form named "focal length of the lens":

$$\boxed{\frac{1}{f}=(n-1)\left(\frac{1}{r_1}-\frac{1}{r_2}\right)} \tag{10.5.223}$$

Similarly in the case of a simple diopter, the "focal image" F_i, or "second focal point of a lens" is where converge the light rays after passing through the lens, but that were before the lens parallel between them and to the optical axis ($p = +\infty$). Thus, given the central spherical symmetry of thin lenses it is simple enough to reverse in our imagination the previous picture to get the concept:

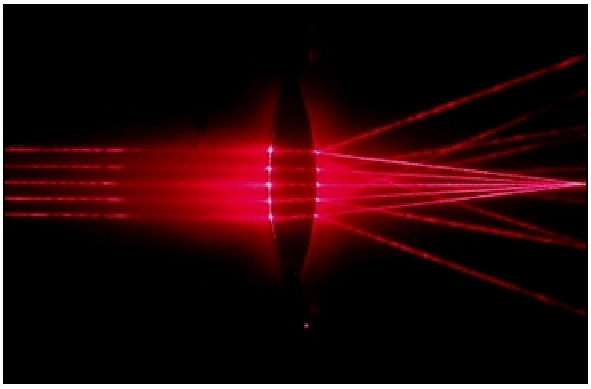


Figure 10.193: **Second focal point (target) is here on the right of the thin lens**

Or for those who cannot see well the figure above, there is another more explicit illustration of what happen with the second focal point of a bi-convex lens:

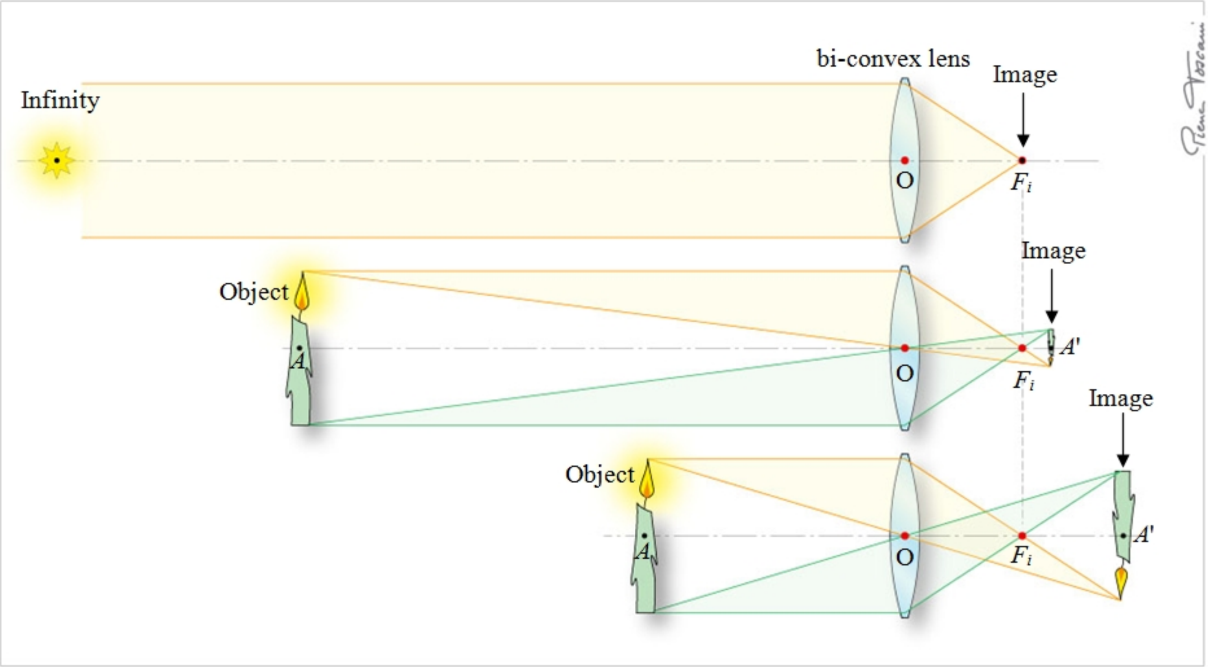


Figure 10.194: **Second focal point with bi-convex lens behaviour (source: `http://www.pierretoscani.com`, author: Pierre Toscani)**

The distance from the lens to the image focus is then named the "focal image distance" and we denote it in practice with the same letter f as by symmetry of the thin lens, by putting $q=f$ and $p=+\infty$ we have:

$$\frac{1}{f}=(n-1)\left(\frac{1}{r_1}-\frac{1}{r_2}\right) \tag{10.5.224}$$

Therefore, in a thin lens the two foci are located symmetrically on each side of the vertical axes of symmetry of the lens.

So as in both the inverse of the focal length is equal to the refractive power nothing prevents us to write the equation of thin lenses as we often found in textbooks:

$$\boxed{\frac{1}{p}+\frac{1}{q}=\frac{1}{f}=(n-1)\left(\frac{1}{r_1}-\frac{1}{r_2}\right)=\mathcal{D}} \tag{10.5.225}$$

In this form it is then named the "equation of Opticians" or "eyewear equation" or "Lens maker equation"... In the following simple form that does not make appear the physical properties of the lens (which is often the relation shown in high-school classes and could be the reason why it seems that it has a different name):

$$\frac{1}{p}+\frac{1}{q}=\frac{1}{f} \tag{10.5.226}$$

or more explicitly:

$$\frac{1}{\text{object distance}}+\frac{1}{\text{image distance}}=\frac{1}{\text{focal length}} \tag{10.5.227}$$

we name it "conjugation equation" or "second Descartes formula for thin lenses" or "spherical mirror equation".

Furthermore, if the focal distance is positive, and thus respectively the dioptric power as well, then the lens is say to be (obviously) a "convergent lens" and if the focal length is negative, and thus the refractive power respectively also, the lens is say to be a "diverging lens" (notice also the traditional symbolic representations of lenses for example in the E-draw software):

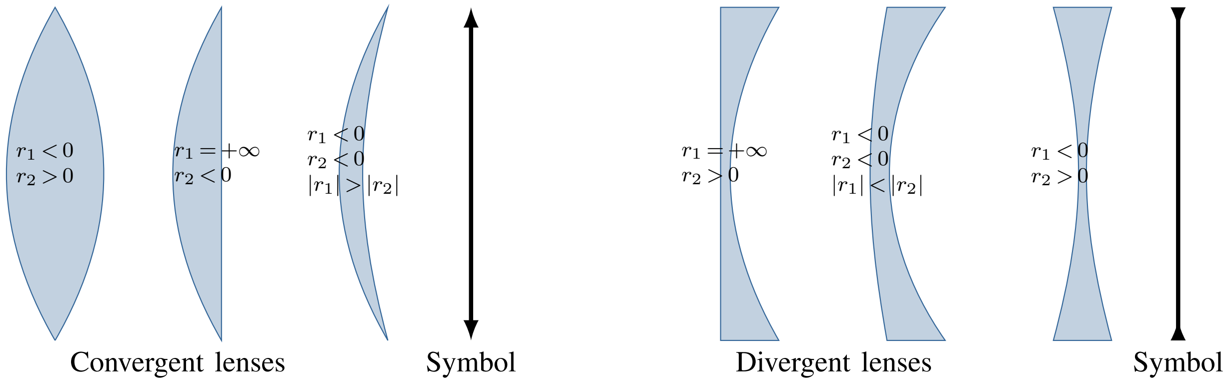

$r_1 < 0$
$r_2 > 0$
$r_1 = +\infty$
$r_2 < 0$
$r_1 < 0$
$r_2 < 0$
$|r_1| > |r_2|$
$r_1 = +\infty$
$r_2 > 0$
$r_1 < 0$
$r_2 < 0$
$|r_1| < |r_2|$
$r_1 < 0$
$r_2 > 0$
Convergent lenses
Symbol
Divergent lenses
Symbol

Figure 10.195: Examples of diverging and converging lenses:
bi-convex, plano-convex, convex meniscus, bi-concav, plano-concav, concav meniscus

If necessary here is figure of why they are respectively named convergent and divergent lenses:

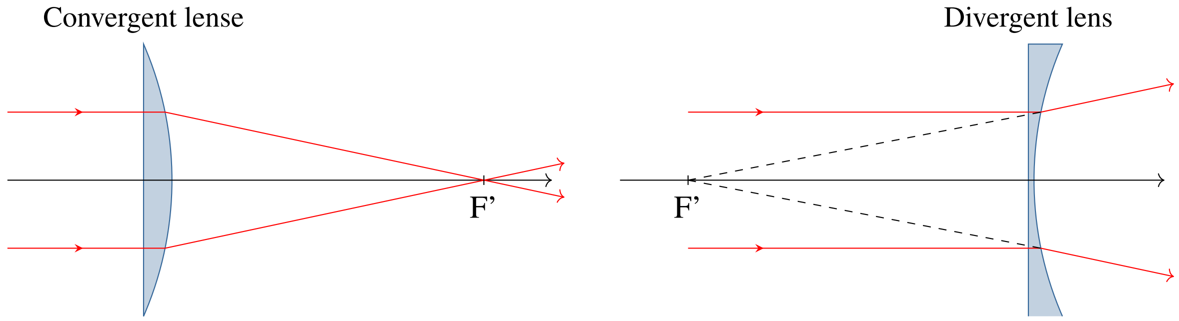

Convergent lense
Divergent lens
F'
F'

Figure 10.196: Convergent and divergent lens concepts (author: Jimmy Roussel)

✍**Example:**

We want to find the radius of a biconcave lens symmetrically ground from a glass with index of refractive 1.55 so that its focal length in air is 20 [cm] (for a biconcave lens, both surfaces have the same radius of curvature). To found the answer we use the thin-lens form of the lens maker's equation:

$$\frac{1}{f} = (n-1)\left(\frac{1}{r_1} - \frac{1}{r_2}\right) = \left(\frac{n_2}{n_1} - 1\right)\left(\frac{1}{r_1} - \frac{1}{r_2}\right) \tag{10.5.228}$$

where $r_1 < 0$ and $r_2 > 0$. Since we are making a symmetric biconcave lens, we have

$|r_1| = |r_2| = r$. Therefore:

$$\frac{1}{f} = \left(\frac{n_2}{n_1} - 1\right)\left(-\frac{2}{R}\right) \tag{10.5.229}$$

Solving for R and inserting $f = -20$ [cm], $n_2 = 1.55$, and $n_1 = 1.00$ gives:

$$R = -2f\left(\frac{n_2}{n_1} - 1\right) = -2(-20)\left(\frac{1.55}{1.00}\right) = 22\,[\text{cm}] \tag{10.5.230}$$

10.5.5.2.1 Optical Magnification (zoom)

Magnification is the process of enlarging something only in appearance, not in physical size. This enlargement is quantified by a calculated number also named "magnification". When this number is less than one, it refers to a reduction in size, sometimes named "minification" or "de-magnification".

Typically, magnification is related to scaling up visuals or images to be able to see more detail, increasing resolution, using microscope, printing techniques, or digital processing. In all cases, the magnification of the image does not change the perspective of the image.

Figure 10.197: **The stamp appears larger with the use of a magnifying glass (source: Wikipedia)**

Before studying the magnification of spherical convex lenses, let us focus on the general definition of a magnification. For this purpose, consider the following figure:

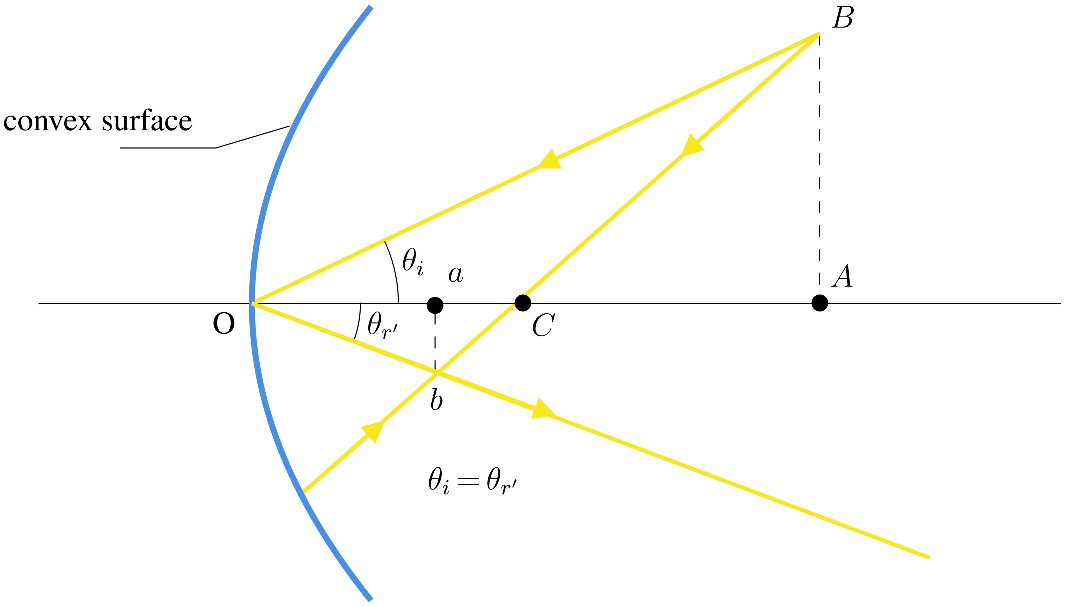


Figure 10.198: **Magnification basic principle**

where for recall the center of curvature C (see section Differential Geometry page 2866) is the center of curvature of the spherical surface of the figure above and the top O is the pole of the spherical cap.

Thus, the "magnification M" of any optical system is defined as the ratio of the size of the image $\overline{ab}$ to that of the real object $\overline{AB}$, that is to say:

$$\boxed{M = \frac{\overline{ab}}{\overline{AB}}} \qquad (10.5.231)$$

We see from the figure above that:

$$\begin{aligned} \tan(\theta_i) &= \frac{\overline{AB}}{\overline{\mathrm{O}A}} = \frac{\overline{AB}}{p} \\ \tan(\theta'_r) &= \frac{\overline{ab}}{\overline{\mathrm{O}a}} = \frac{\overline{ab}}{q} \end{aligned} \qquad (10.5.232)$$

We therefore get, taking into account that $\theta_i = \theta'_r$:

$$\frac{\tan(\theta'_r)}{\tan(\theta_i)} = 1 = \frac{\overline{ab}}{q}\frac{p}{\overline{AB}} \qquad (10.5.233)$$

Hence the "magnification equation":

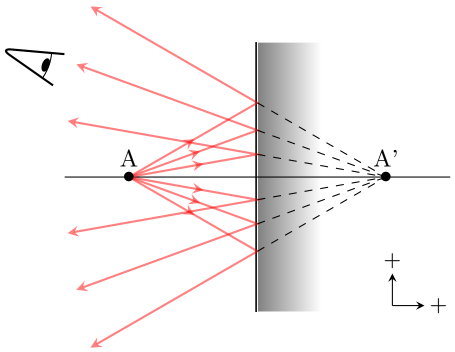
$$\boxed{M = \frac{\overline{ab}}{\overline{AB}} = \frac{q}{p}} \qquad (10.5.234)$$

So in the case of a thin symmetrical spherical lens whose rays pass through the optical center we obtain rays which are describe similar triangles on each side of the axis of symmetry of the spherical thin lens then applying Thales' theorem we also get:

$$M = \frac{\overline{ab}}{\overline{AB}} = \frac{q}{p} \qquad (10.5.235)$$

That's the same result as that obtained already above.

Notice that for a flat mirror we have $\overline{ab}=\overline{AB}$, hence for $p\neq 0$, we always have $M=1$. So whatever the distance $d\neq 0$ from a flat mirror, an object looks like always the same size!!! The reason why many people think that the reflection gets smaller is that most often they see mirrors sitting above the bathroom sink; when they back away the see more of themselves (but the know the mirror is of the same size!), so they think the reflection has shrunk. It's hard to remember that half the mirror start out full of the reflected image of the sink...



Example:

Two sides of a convex lens have a radius of 3 [cm]. The index of the lens material is 1.52. An object of 1.80 [m] height is set to 14 [m] distance of the lens (no matter that it is left or right of the lens because it is assumed as biconvex and therefore symmetrical). Then the dioptric power of the lens is first:

$$\mathcal{D}=(n-1)\left(\frac{1}{r_1}-\frac{1}{r_2}\right)=\frac{1}{f}=(1.52-1)\frac{2}{r}=0.52\frac{2}{0.3}\cong 3.47\,[\text{m}^{-1}] \tag{10.5.236}$$

Which is indeed a positive value and thus gives a focal length (image or object focal regardless because of symmetry of the spherical thin lens!) of $f\cong 28.84\,[\text{cm}]$ so it is better to have a camera with a telephoto lens in the present case... We notice by the value of the focal length, that the subject is obviously beyond the focus. The position of the image is given by:

$$\frac{1}{p}+\frac{1}{q}=\frac{1}{14}+\frac{1}{q}\cong 3.47 \tag{10.5.237}$$

Therefore:

$$q\cong 0.2942\,[\text{m}] \tag{10.5.238}$$

So the picture is about 29.42 centimetres beyond the focus and will by definition be named the "real (reversed) image". The magnification value is then:

$$M=\frac{q}{p}=\frac{0.2942}{14}\cong 0.021 \tag{10.5.239}$$

The size of the inverted real image will be in q:

$$h\cong 0.021\cdot 1.8=0.0378=3.78\,[\text{cm}] \tag{10.5.240}$$

Therefore we have:

$$\boxed{M=\frac{q}{p}=\frac{\left(\mathcal{D}-\dfrac{1}{p}\right)^{-1}}{p}=\frac{\left(\dfrac{1}{f}-\dfrac{1}{p}\right)^{-1}}{p}=\frac{\left(\dfrac{2}{r}-\dfrac{1}{p}\right)^{-1}}{p}} \tag{10.5.241}$$

If we want a magnification (zoom), we must have:

$$M=\frac{1}{p\left(\dfrac{2}{r}-\dfrac{1}{p}\right)}=\frac{1}{\dfrac{2p}{r}-1}\geq 1 \tag{10.5.242}$$

Therefore:

$$\frac{2p}{r}-1\leq 1 \tag{10.5.243}$$

Is trivially:

$$\frac{2p}{r}\leq 2 \tag{10.5.244}$$

Finally:

$$p\leq r \tag{10.5.245}$$

So to do that there is magnification (zoom) it is necessary that the real object is between the center of curvature and the vertical axis of symmetry of the lens.

Remark

The study of this model will allow us to understand partially how the prism works and also of goniometer in astronomy for spectrum analysis as well as X-ray diffraction by a network of atoms (the importance of the latter being quite significant!).

We must be careful with how the human brain process the world! Many configurations of objects are misleading for the way our internal machine process images. For example in the image below both Moons have the same size but our internal processing machine (brain) has the tendency to interpret the above Moon as being bigger than the lower Moon... :

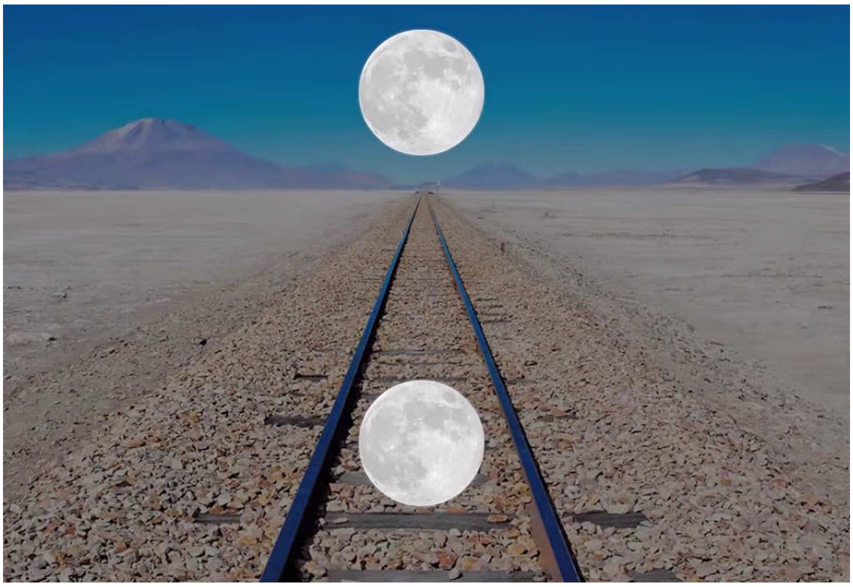


Figure 10.199: **Magnification optical illusion (source: NASA)**

We haven't so far studied all possibles configuration of biconvex lenses (mainly because I'm not a fan of the subject...) but here is very good summary founded on Instagram of all main configurations:

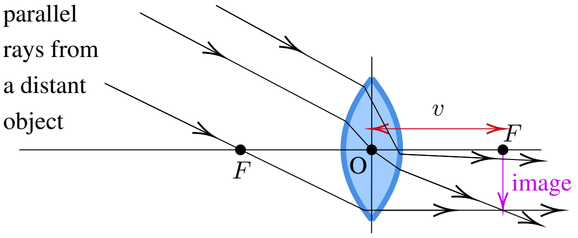
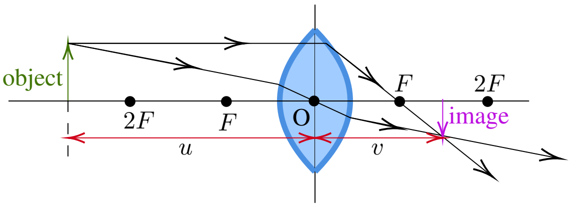
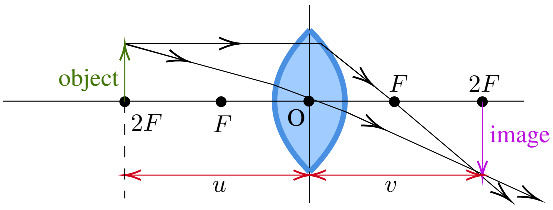
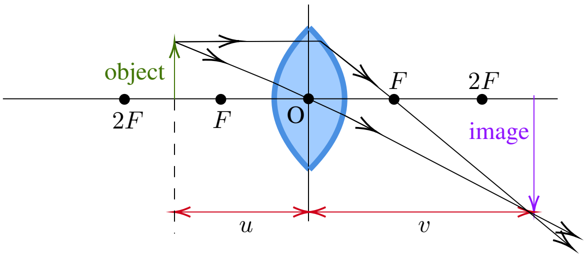
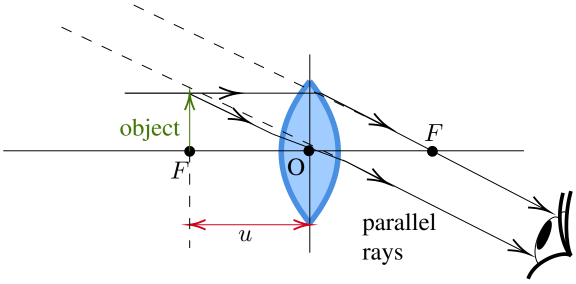
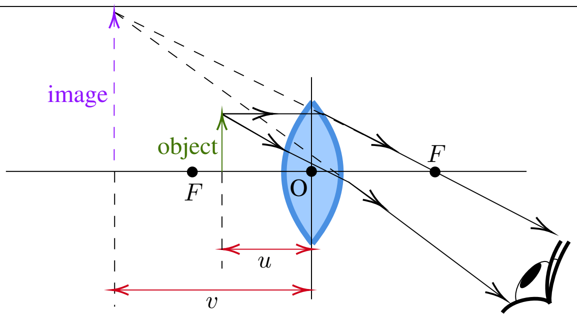
Object distance (u)	**Ray diagram**	**Type of image**	**Image distance** (v)	**Uses**
$u = +\infty$	parallel rays from a distant object 	- inverted - real - diminished	$v = F$ - oposite side of the lens	- object lens of a telescope
$u > 2F$	 object $2F$ F O F $2F$ image u v	- inverted - real - diminished	$F < v < 2F$ - oposite side of the lens	- camera - eye
$u = 2F$	 object $2F$ F O F $2F$ image u v	- inverted - real - same size	$v = 2F$ - oposite side of the lens	- photocopier making same-sized copy
$F < u < 2F$	 object $2F$ F O F $2F$ image u v	- inverted - real - magnified	$v > 2F$ - oposite side of the lens	- projector - photograph enlarger
$u = F$	image at infinity  object F O F u parallel rays	- upright - virtual - magnified	- image at infinity - same side of the lens	- to produce a parallel beam of light, e.g. a spotlight
$u < F$	 image object F O F u v	- upright - virtual - magnified	- image is behind the object - same side of the lens	- magnifying glass

Figure 10.200: **Biconvex lenses typical configurations and applications**

The concave mirror bends light inwards, towards ourselves. If we imagine the mirror as a part of a larger sphere, then if we view the mirror from a point closer than the center of this sphere, we see an enlarged (magnified) upright image. If look further away fro the center we will see reversed image as illustrated by the photo below:

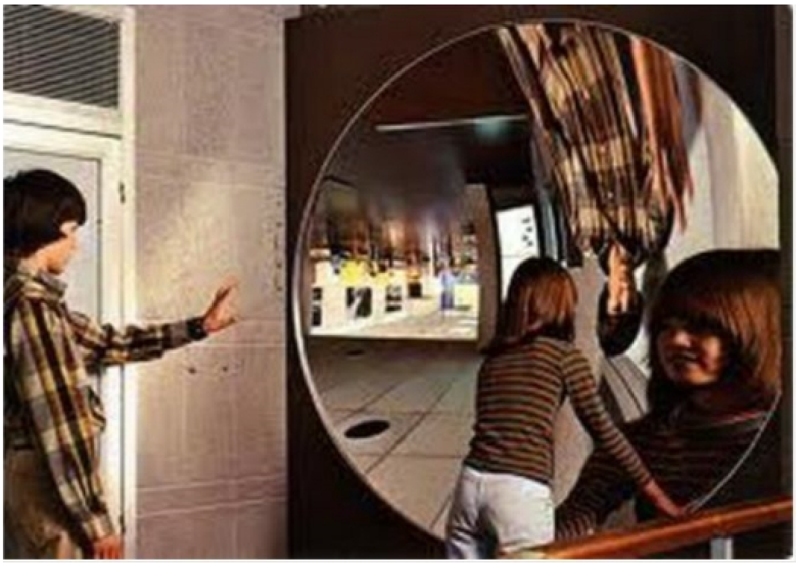


Figure 10.201: **Concave mirror (source: ?)**

Notice above now the image of the girl with the striped shirt is right side up and magnified. She is inside the focal point. The person in the plaid shirt is outside the focal point, so it is upside down and about the same size. But the image from the back of the room is both upside down and smaller than normal, because it is much further away from the focal point!

The same rules applies with a famous case know by all young children:

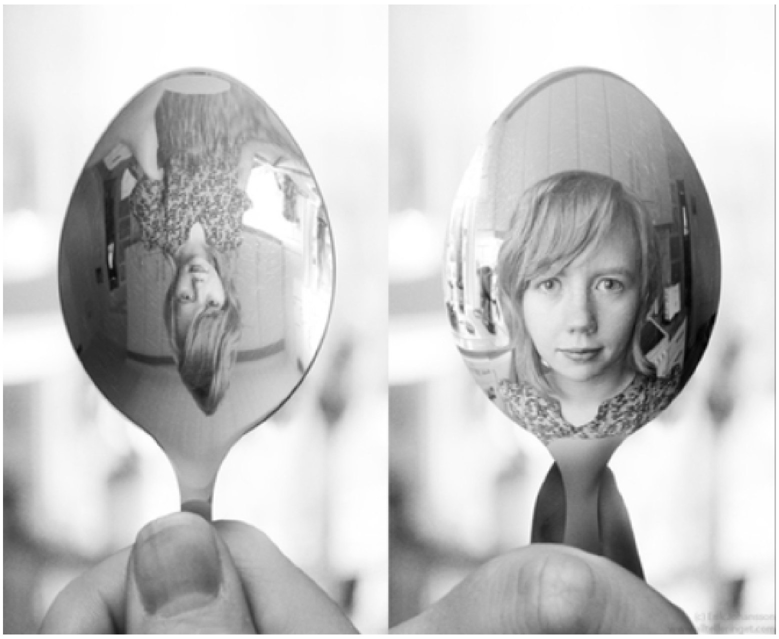


Figure 10.202: **Left concave spoon-mirror, Right convex spoon-mirror (source: ?)**

Remark

Depending on the manufacturer a makeup mirror is concave or convex as they have the same structure as a spoon but at the difference that they have a very polish surface. But most of time the visible part is a concave mirror build in such a way that a person at 25 [cm] away from the surface see its image doubled.

For practical purposes, let us indicate the following figure (left global view, right zoom on the curve part):

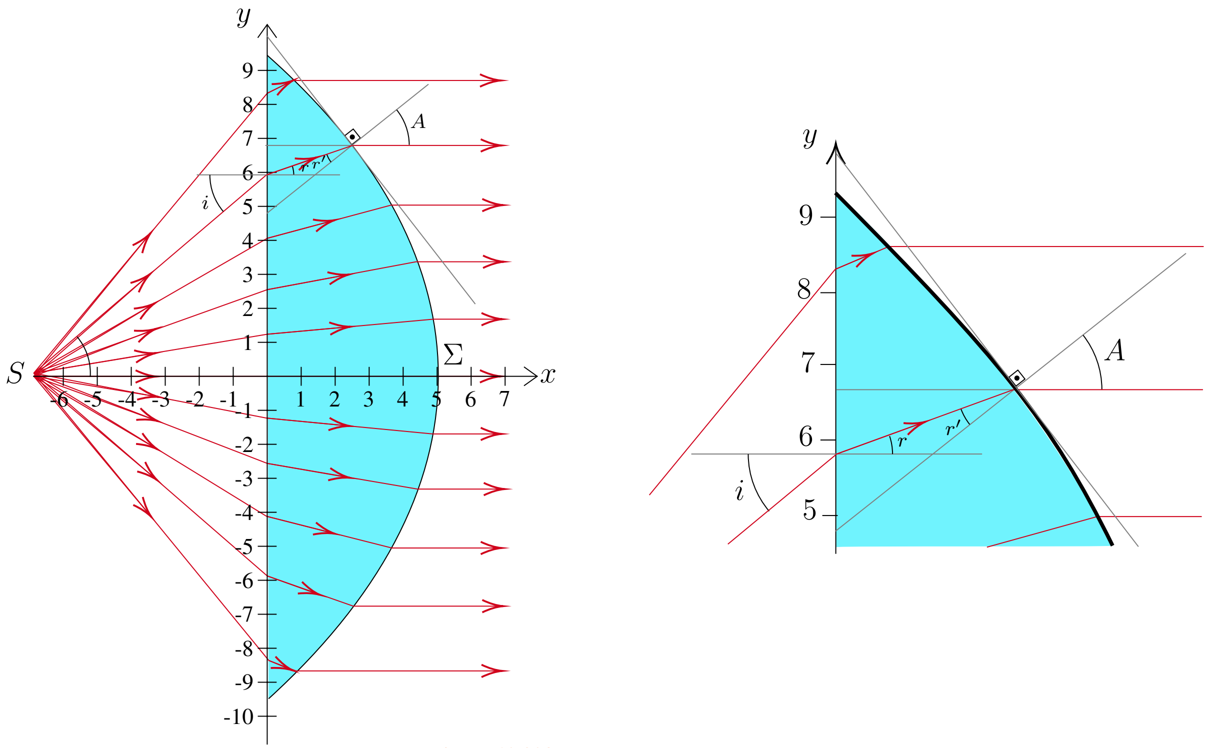


Figure 10.203: **Full plane-curved lens**

and therefore a full plane-curved lens, as a parabolic mirror, has the property of making parallel the rays having for source its focal. It produced by refraction the effect that the parabolic mirror produced by reflection...!

Fresnel invented a quite clever lens that is seen in many lighthouses and that achieves the same result with less material and therefore with less intensity loss:

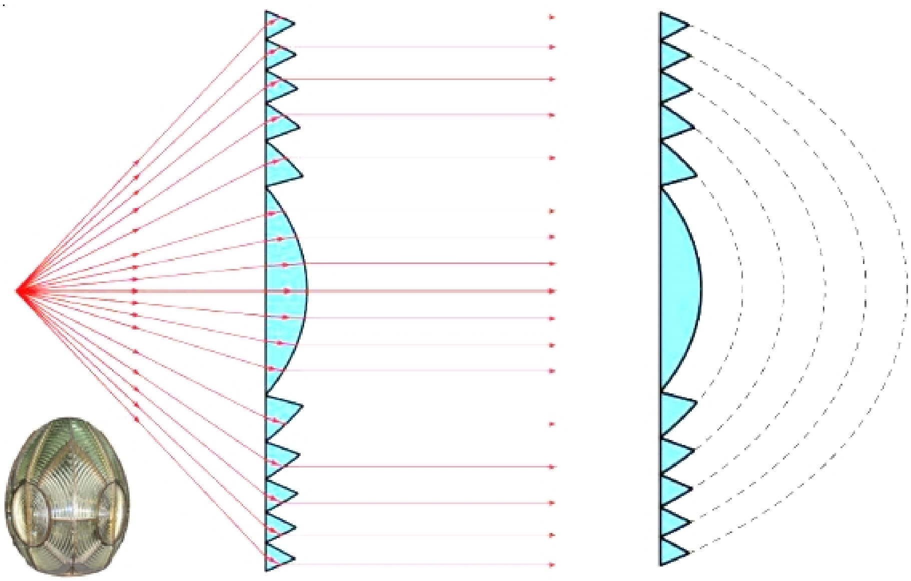


Figure 10.204: Fresnel lens with lighthouse example in the left bottom corner (source: ?)

So far it could be important to make a summary of the sign convention for the equation of opticians (or "mirror equation") and the magnification equation:

$$\begin{gathered}\frac{1}{p}+\frac{1}{q}=\frac{1}{f}\cong\frac{2}{r}\\ M=\frac{q}{p}\end{gathered} \tag{10.5.246}$$

therefore:

- Focal length f:
  - Positive for concave mirrors or concave lenses
  - Negative for convex mirrors or convex lenses
- Magnification M:
  - Positive for upright images
  - Negative for inverted images
  - Enlarged when $|M|>1$
  - Reduced when $|M|<1$
- Image distance q:
  - Positive for real images
  - Negative for virtual images

Remark

Fun fact! A mirror doubles the perceived distance of any object! If you are at 50 [cm] away from the mirror, your reflection will appear to be at 100 [cm] away (since you reflection appears to be 50 [cm] behind the mirror). At 20 [cm] you will appear 40 [cm] away. At 10 [cm] you will appear 20 [cm] away. At 0.5 [cm] you will appear 1 [cm] away. At 0 [cm] you will appear to be 0 [cm] away. The relation is according to the mirror equation $r=2d$, where r is an object's reflection perceived distance and d is the object's actual distance from the mirror.

✍**Example:**

Consider that the Sun (actually a G2V-type star, aka yellow dwarf) is the object, so the object distance is essentially infinity: $q=+\infty$. The desired image distance is $q=40.0$ [cm]. We use the mirror equation to find the focal length of the mirror:

$$f=\left(\frac{1}{p}+\frac{1}{q}\right)^{-1}=\left(\frac{1}{+\infty}+\frac{1}{0.4}\right)^{-1}=0.4\,[\text{m}] \tag{10.5.247}$$

But we have proved earlier that in the approximation of small angles:

$$\frac{1}{p}+\frac{1}{q}\cong\frac{2}{r} \tag{10.5.248}$$

Thus, the radius of the mirror if considered as spheric, will be equal to:

$$r\cong 2f=0.8\,[\text{m}] \tag{10.5.249}$$

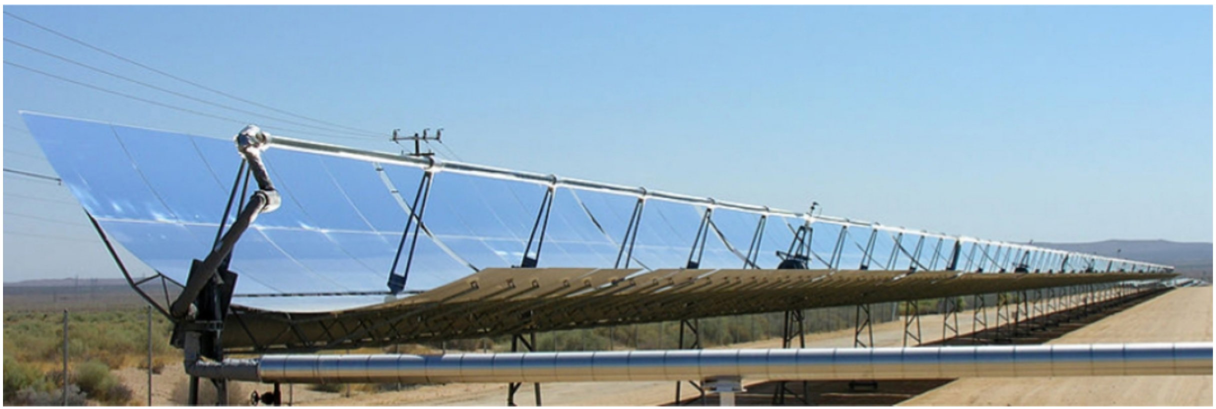


Figure 10.205: Parabolic trough collectors are used to generate electricity in southern California (source: Wikipedia)

10.5.5.2.2 Angle of View

In photography, the "angle of view" (AOV) describes the angular extent of a given scene that is imaged by a camera. It is used interchangeably with the more general term "field of view" (FOV).

What we will see further below should help the reader to understand what happens in both following figures made of real photos:

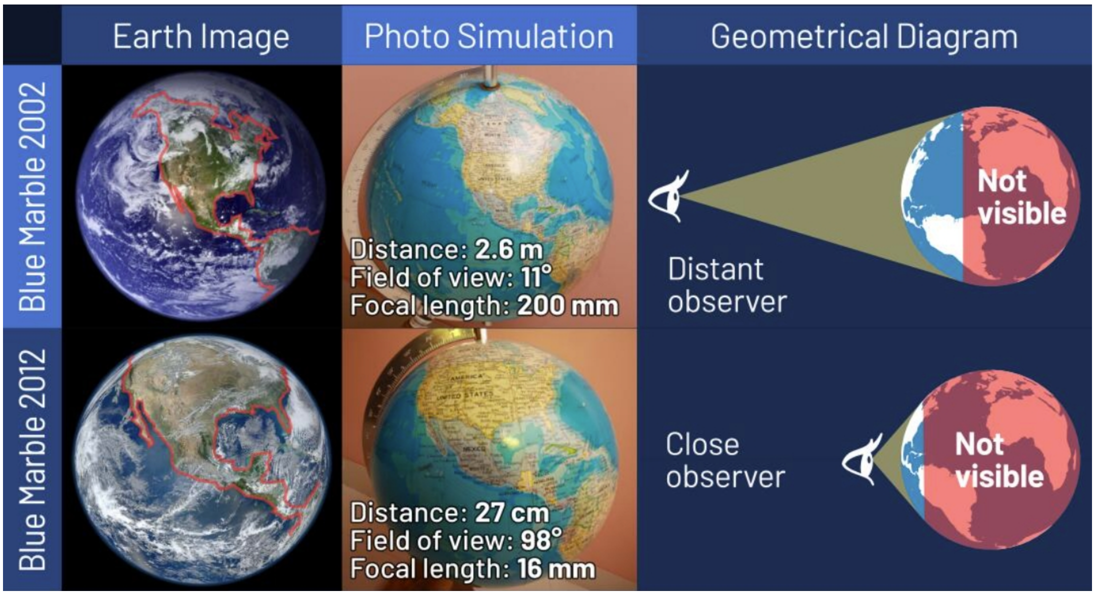

Earth Image
Photo Simulation
Geometrical Diagram
Blue Marble 2002
Distance: 2.6 m
Field of view: 11°
Focal length: 200 mm
Not visible
Distant observer
Blue Marble 2012
Distance: 27 cm
Field of view: 98°
Focal length: 16 mm
Close observer
Not visible

Figure 10.206: Field of view real case with Earth's photo (source: FleatEarth.ws)

The narrower the field of view, the larger they appear in a photo.

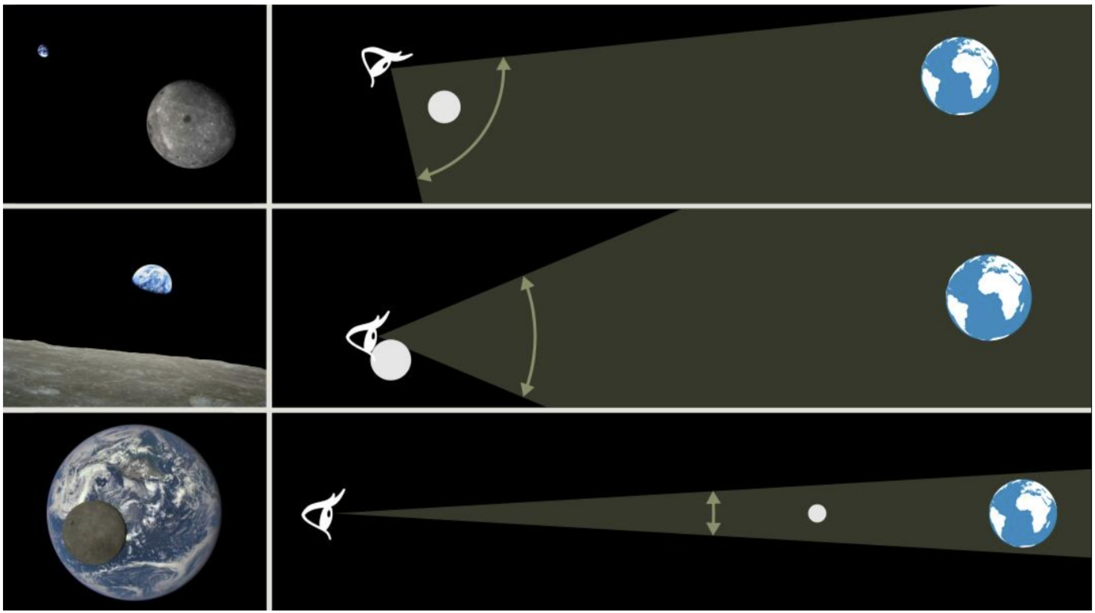
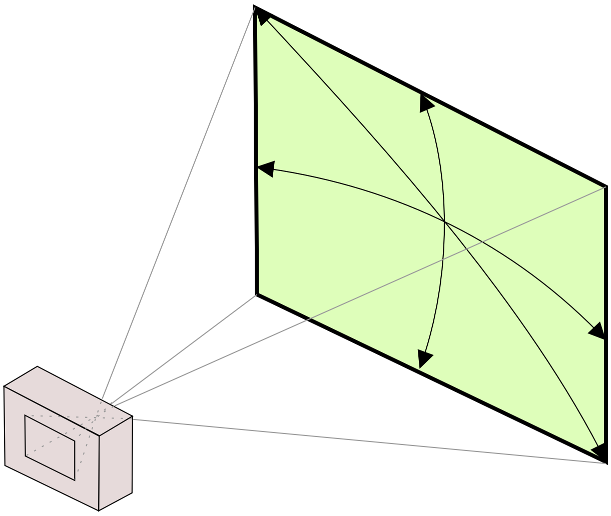


Figure 10.207: Field of view real case Earth and Moon (source: FleatEarth.ws)

A camera's angle of view can be measured horizontally, vertically, or diagonally:



As mentioned above, a camera's angle of view depends not only on the lens, but also on the image sensor or film. Digital sensors are usually smaller than 35 [mm] film, and this causes the lens to have a narrower angle of view than with 35 [mm] film, by a constant factor for each sensor (named the "crop factor"). In everyday digital cameras, the crop factor can range from around 1, called full frame (professional digital SLRs where the sensor size is similar to the 35 [mm] film), to 1.6 (consumer SLR), to 2 (Micro Four Thirds ILC), and to 6 (most compact cameras).

For lenses projecting rectilinear (non-spatially distorted) images of distant objects, the effective focal length and the image format dimensions completely define the angle of view.

For a lens projecting a rectilinear image (focused at infinity, see proof below), the angle of view (α) can be calculated from the chosen dimension (d), and effective focal length (f) (f will be defined here as the distance of the lens with respect to the image plane. For a thick lens, it is the distance of the rear principal plane of the lens with respect to the image plane) as follows:

$$\alpha = 2\arctan\left(\frac{d}{2f}\right) \tag{10.5.250}$$

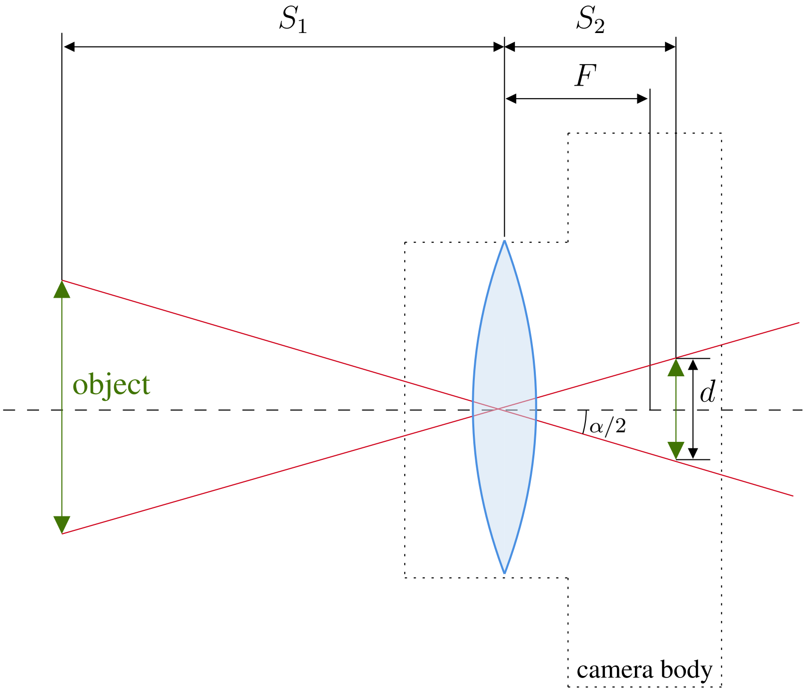
d represents the size of the film (or sensor) in the direction measured. For example, for 35 [mm] film which is 36 [mm] wide and 24 [mm] high, $d = 36$ [mm] would be used to obtain the horizontal angle of view and $d = 24$ [mm] for the vertical angle.

Because this is a trigonometric function, the angle of view does not vary quite linearly with the reciprocal of the focal length. However, except for wide-angle lenses, it is reasonable to approximate:

$$\alpha \cong \frac{d}{f} \tag{10.5.251}$$

or $180d/(\pi F)$ degrees.

Proof 10.2.5. Consider a rectilinear lens in a camera used to photograph an object at a distance S_1, and forming an image that just barely fits in the dimension, d, of the frame (the film or image sensor). Treat the lens as if it were a pinhole at distance S_2 from the image plane (technically, the center of perspective of a rectilinear lens is at the center of its entrance pupil where chief rays meet):



Now $\alpha/2$ is the angle between the optical axis of the lens and the ray joining its optical center to the edge of the film. Here α is defined to be the angle-of-view, since it is the angle enclosing the largest object whose image can fit on the film. We want to find the relationship between:

- the angle α
- the "opposite" side of the right triangle, $d/2$ (half the film-format dimension)
- the "adjacent" side, S_2 (distance from the lens to the image plane)

Using basic trigonometry, we find:

$$\tan(\alpha/2)=\frac{d/2}{S_2} \tag{10.5.252}$$

which we can solve for α, giving:

$$\alpha=2\arctan\left(\frac{d}{2S_2}\right) \tag{10.5.253}$$

To project a sharp image of distant objects, S_2 needs to be equal to the focal length, that we will denote here F, which is attained by setting the lens for infinity focus. Then the angle of view is given by:

$$\boxed{\text{FOV}=\alpha=2\arctan\left(\frac{d}{2F}\right)} \tag{10.5.254}$$

■ Q.E.D.

For macro photography, we cannot neglect the difference between S_2 and F. From the conjugation equation (see page 4440):

$$\frac{1}{F}=\frac{1}{S_1}+\frac{1}{S_2} \tag{10.5.255}$$

The absolute transverse magnification (the absolute ratio of the image height to the object height) can be expressed as we have seen earlier above by the magnification equation (see 4442):

$$M=\frac{S_2}{S_1} \tag{10.5.256}$$

we can substitute S_1 and with some algebra find:

$$S_2=F\cdot(1+M) \tag{10.5.257}$$

Therefore:

$$\alpha=2\arctan\left(\frac{d}{2S_2}\right) \tag{10.5.258}$$

can be rewritten:

$$\boxed{\text{FOV}_{\text{macro}}=\alpha=2\arctan\left(\frac{d}{2F\cdot(1+M)}\right)} \tag{10.5.259}$$

✍**Example:**

Consider a 35 [mm] camera with a lens having a focal length of $F=50$ [mm]. The dimensions of the 35 [mm] image format are 24 [mm] (vertically) $\times 36$ [mm] (horizontal), giving a diagonal of about 43.3 [mm].

At infinity focus, $f=F$, the angles of view are:

- Horizontally:

$$\alpha_h=2\arctan\left(\frac{h}{2f}\right)=2\arctan\left(\frac{36}{2\cdot 50}\right)\cong 39.6^\circ \tag{10.5.260}$$

- Vertically:

$$\alpha_v=2\arctan\left(\frac{v}{2f}\right)=2\arctan\left(\frac{24}{2\cdot 50}\right)\cong 27.0^\circ \tag{10.5.261}$$

- Diagonally:

$$\alpha_d=2\arctan\left(\frac{d}{2f}\right)=2\arctan\left(\frac{43.3}{2\cdot 50}\right)\cong 46.8^\circ \tag{10.5.262}$$

In the mosaic below the angle of view and distance of the camera while keeping the object in frame results in vastly differing images. At distances approaching infinity, the light rays are nearly parallel to each other, resulting in a "flattened" image. At low distances and high angles of view objects appear "foreshortened"

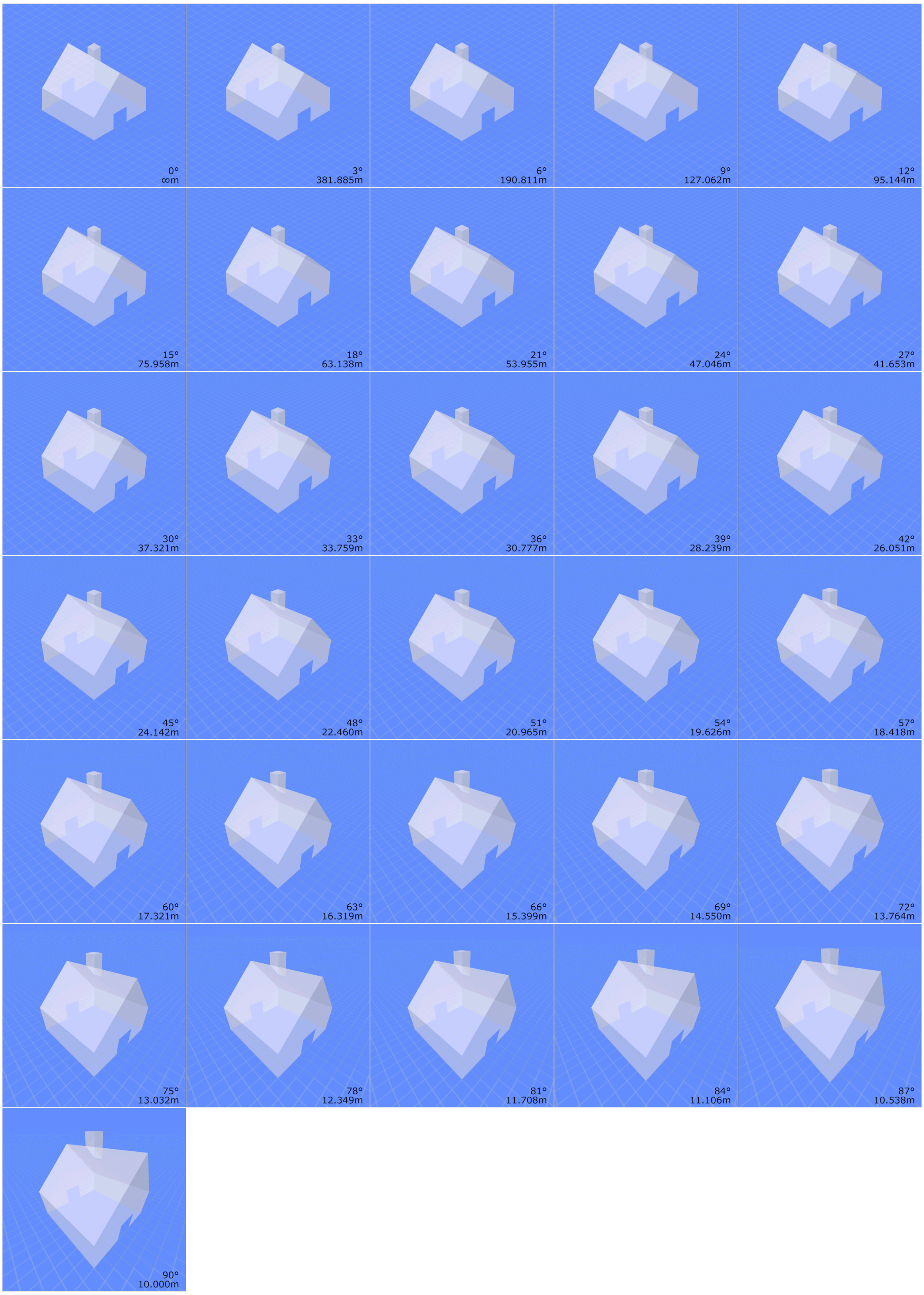
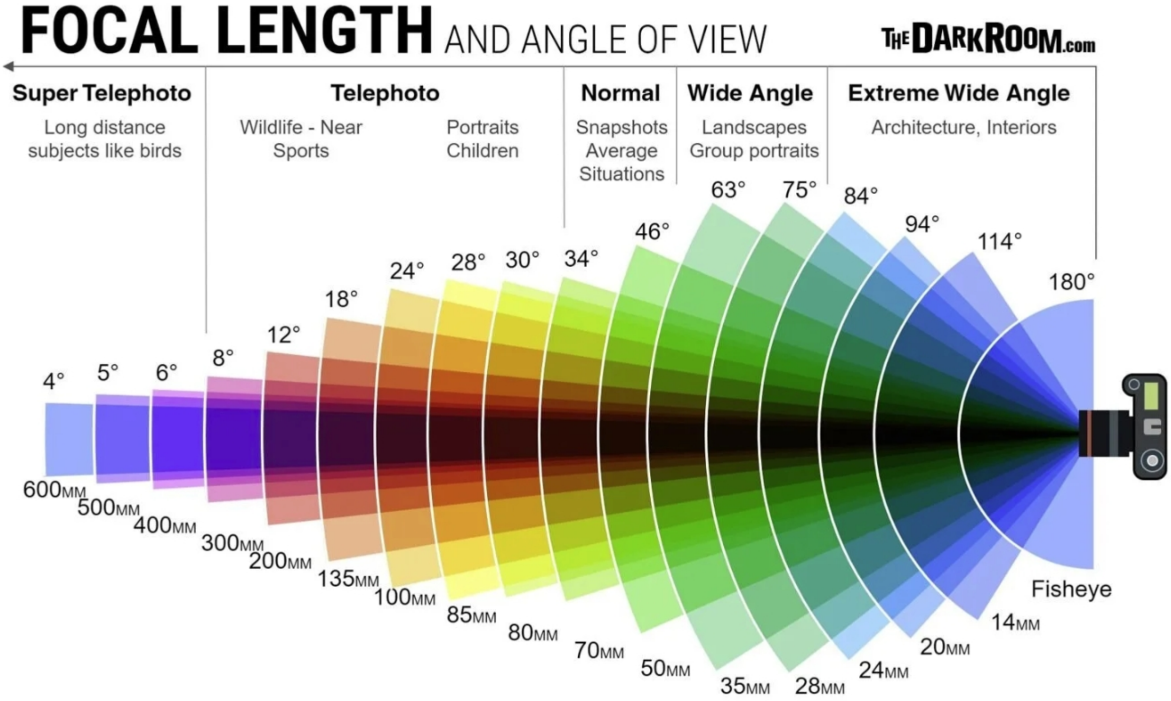

0°
∞m
3°
381.885m
6°
190.811m
9°
127.062m
12°
95.144m
15°
75.958m
18°
63.138m
21°
53.955m
24°
47.046m
27°
41.653m
30°
37.321m
33°
33.759m
36°
30.777m
39°
28.239m
42°
26.051m
45°
24.142m
48°
22.460m
51°
20.965m
54°
19.626m
57°
18.418m
60°
17.321m
63°
16.319m
66°
15.399m
69°
14.550m
72°
13.764m
75°
13.032m
78°
12.349m
81°
11.708m
84°
11.106m
87°
10.538m
90°
10.000m

Figure 10.208: **Mosaic of the impact of angle of view on deformation of objects (source: Wikipedia, author: Michael Horvath)**

Given the behaviour of the function $\arctan(1/F)$, we can represent the angle of view of a camera length like the famous kind of infography below tha profesionnals photographers know very well:


FOCAL LENGTH AND ANGLE OF VIEW
THE DARKROOM.com
Super Telephoto
Long distance subjects like birds
Telephoto
Wildlife - Near Sports
Portraits Children
Normal
Snapshots Average Situations
Wide Angle
Landscapes Group portraits
Extreme Wide Angle
Architecture, Interiors
4°
5°
6°
8°
12°
18°
24°
28°
30°
34°
46°
63°
75°
84°
94°
114°
180°
600MM
500MM
400MM
300MM
200MM
135MM
100MM
85MM
80MM
70MM
50MM
35MM
28MM
24MM
20MM
14MM
Fisheye

This explains the famous effect of "background compression" also well represented below by Gavin Hoey (the model is Fern Meave for information...):

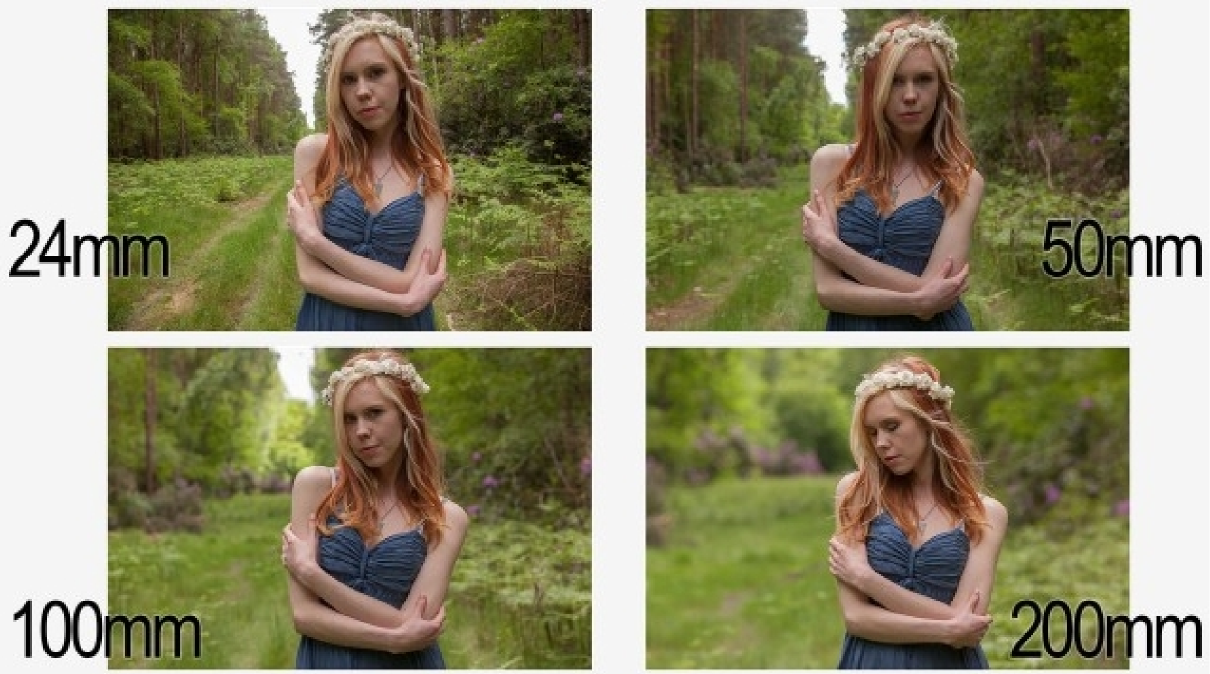

24mm
50mm
100mm
200mm

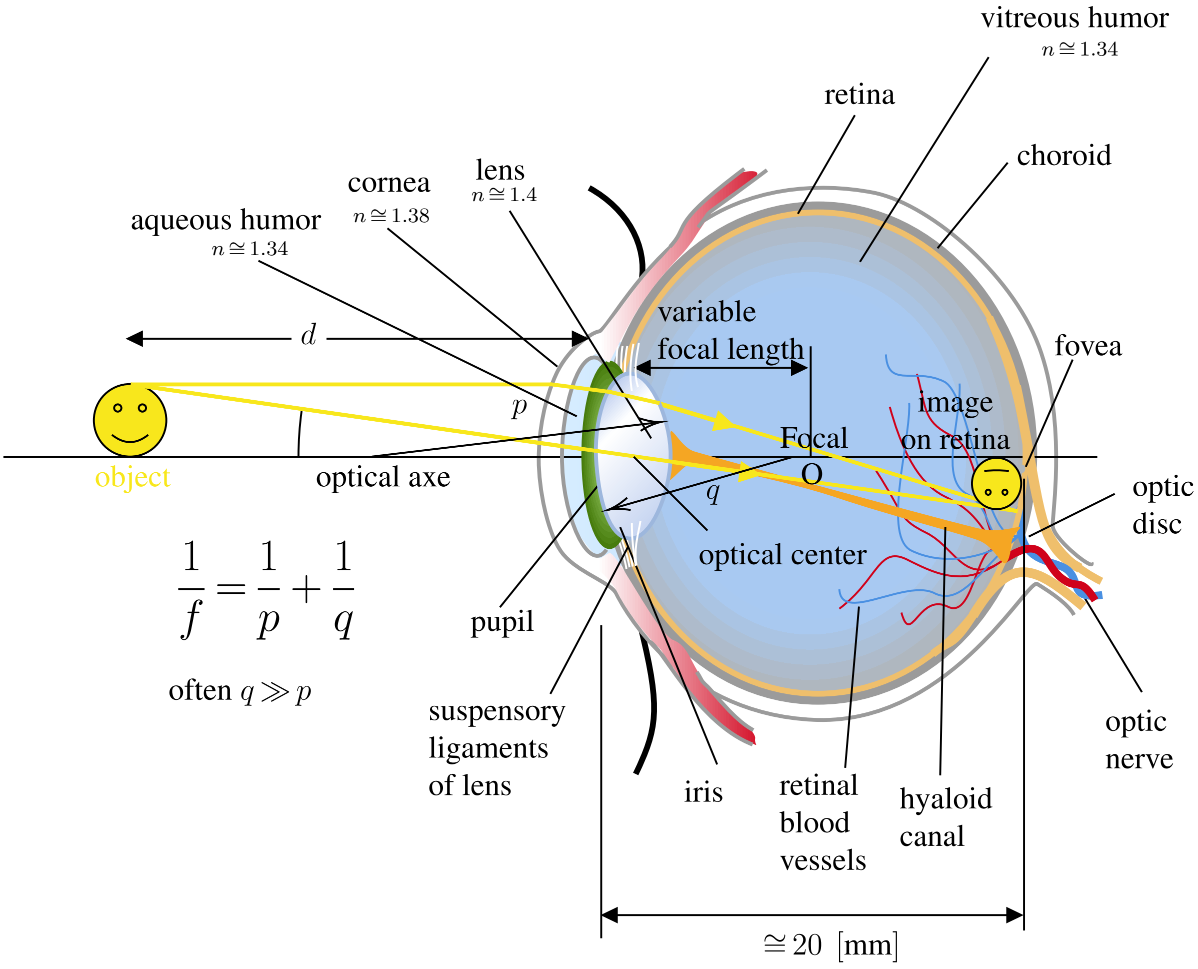
Figure 10.209: **Illustration of background compression at difference focal lengths for different distance by keeping the model size on the screen (author: Gavin Hoey, source:** `https://www.gavtrain.com/`**)**

The reader must really keep in mind that in order to keep the model the same size, the photogrpaher had to increase his shooting distance as then zoomed in and that increase in distance that causes the background to compress. Not the focal length change!

10.5.5.2.3 Human eyes

Let us do some human biology to close this topic about lenses...

The lens of the eye which can be deformed under the effect of certain muscles, constitutes a variable focus lens to accommodate variable distance vision of objects. The distance from the optical center to the retina r being fixed (see figure below), the only way to clearly see objects at different distances d is to change the focal length f. In its normal state, the lens has a relatively flat configuration, with a large radius of curvature.



vitreous humor
$n \cong 1.34$
retina
choroid
lens
$n \cong 1.4$
cornea
$n \cong 1.38$
aqueous humor
$n \cong 1.34$
d
variable
focal length
fovea
image
on retina
p
Focal
O
object
optical axe
q
optic
disc
optical center
$\frac{1}{f}=\frac{1}{p}+\frac{1}{q}$
pupil
often $q \gg p$
suspensory
ligaments
of lens
optic
nerve
iris
retinal
blood
vessels
hyaloid
canal
$\cong 20$ [mm]

Figure 10.210: Human eye optics (simplified, not to scale and in a special configuration case)

The human eye has the function of focusing light from an object at infinity on the retina. But all eyes are not doing this correctly and the far point (maximum distance of distinct vision without accommodation) or "punctum remotum" is sometimes at a finite distance, sometimes even less than 5 meters (probably causing eyestrain) as practitioners consider 6 meters already as infinity relatively to the retina modification over this distance being most of time not significant.

If the object approaches the eye, the muscles of the retina contract, the lens swells and its focal length decreases so that the image is always formed on the retina. The nearest point which can be seen clearly with maximum accommodation is named the "near point" or "punctum proximum". This distance is changing significantly with age: it is ten centimetres for a ten years old child, a hundred centimetres for a person of sixty years old (that is "presbyopia").

The reader must keep in mind that the image projected onto the back of our eyes is upside down. Our brain software decodes this image so that we perceive it the right way up (at least by the huge majority of people as a few case well known by the scientific community are born with a defective brain visual software). Those little ray diagrams as above for non-optical specialists are very misleading and deceptively simple. In a real system, even a simple lens, there a rays going everywhere. In most textbooks - as the present one - they sometimes only draw two. Feynman liked to say that these textbooks cheated you because they never really explained the ray diagrams or how they switched from one kind to another when they talked about different things.

Remark

Cones in the retina are sensitive to single photons. However, such tiny signals are filtered out during the image processing in the human optic nerve and brain. It takes 7 to 30 photons to "see" an event. In comparison the company Photon Force™ manufactures single-photon sensitive cameras! For standard mainstream cameras the sensitivity depends on many factors: aperture, focal, sensitivity of the sensor/tape, signal to noise ratio of the sensor/tape. So there is no general answer to the sensitiveness of a camera as the technologies are veeery diversified.

10.5.5.2.3.1 Eye magnification

When we want to examine the details of an object, we approach it to our eye. By doing so, we get a bigger picture on the retina. Indeed, the distance from the focal point O of the eye to the retina can be considered to be 15 [mm], $\overline{ab}$ image of $\overline{AB}$ is the intersection of the rays $\overline{AO}$ and $\overline{AB}$ with the retina (the rays being considered as a first approximation in a straight line):

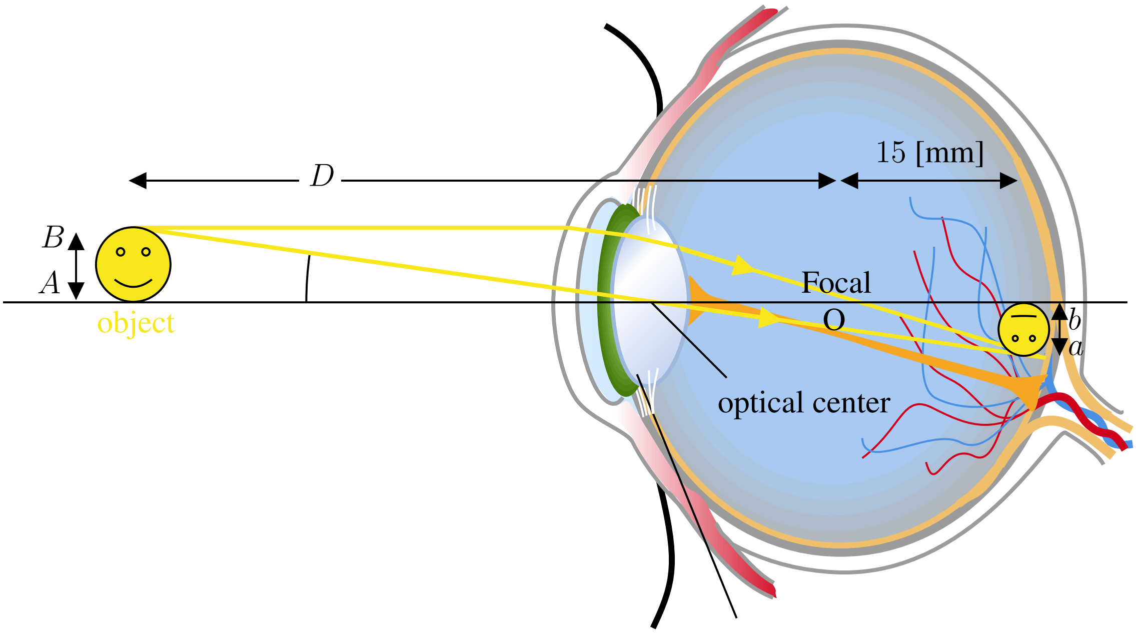


Figure 10.211: **Eye as magnifier tool**

We then use Thales theorem (i.e. intercept theorem) once again and assuming that everything is measured in millimetres (and not forgetting that we consider the rays as a first approximation in a straight line from the focus to the original object):

$$\frac{\overline{ab}}{\overline{AB}}\cong\frac{15}{D} \tag{10.5.263}$$

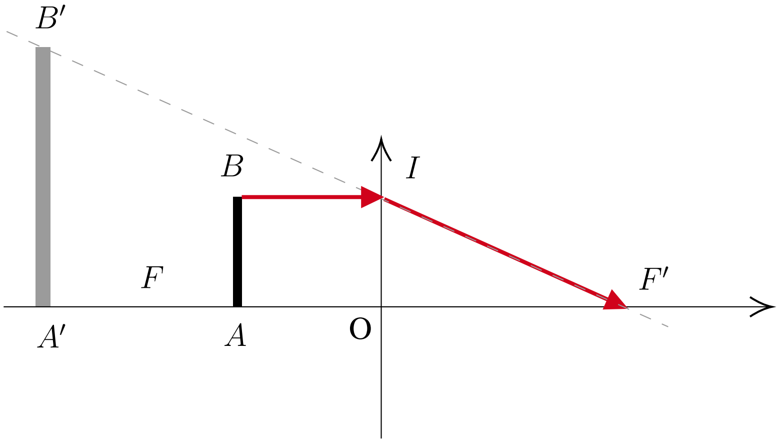
or:

$$\overline{ab}\cong 15\frac{\overline{AB}}{D} \tag{10.5.264}$$

However, the distance D cannot become smaller than that for which the accommodation is maximum. This minimum distinct vision distance can be considered to be 250 [mm]; it allows the fairly prolonged examination of objects without too much "exhaustion" of the eye.

We have shown that a converging lens gives an enlarged virtual image of a real object provided that the object is placed between the center of curvature and the vertical axis of symmetry of the converging lens. So let's look under what configuration this device is really advantageous for human vision (considering that I is a symbolic 1D representation of converging lens).

Let us place the eye at the focal point F'; instead of the object $\overline{AB}$, it sees the image $\overline{A'B'}$ which is at distance $\overline{A'F'}$:



The size of the corresponding retinal image is then according to the previous image:

$$\overline{a'b'}\cong 15\frac{\overline{A'B'}}{\overline{A'F'}} \tag{10.5.265}$$

Then from the similarity of the triangles, we can write (application of Thales theorem again):

$$\frac{\overline{A'B'}}{\overline{A'F'}}=\frac{\overline{\mathrm{O}I}}{\overline{\mathrm{O}F'}}=\frac{\overline{AB}}{f} \tag{10.5.266}$$

since the distance $\overline{\mathrm{O}F'}$ corresponds to the focal length f of the converging lens. We then have:

$$\overline{a'b'}\cong 15\frac{\overline{A'B'}}{\overline{A'F'}}=15\frac{\overline{AB}}{f} \tag{10.5.267}$$

However, as we mentioned, if the object were seen with the naked eye, the retinal image would have under the aforementioned hypotheses an optimum of:

$$\overline{ab} \cong 15\frac{\overline{AB}}{250} \tag{10.5.268}$$

So that we have an advantage to observe through a covering lens, which in this case we name a "magnifying glass", $\overline{a'b'}$ must be larger than $\overline{ab}$. Either mathematically:

$$\overline{a'b'} > \overline{ab} \Leftrightarrow 15\frac{\overline{AB}}{f} > 15\frac{\overline{AB}}{250} \tag{10.5.269}$$

which give the fundamental condition of the converging lens:

$$f < 250\,[\text{mm}] \tag{10.5.270}$$

Thus, a converging lens can only be used as a magnifying glass if its focal length is less than 250 [mm] under the aforementioned assumptions. By definition, we named "magnification" of the magnifying glass the ratio:

$$M = \frac{\overline{a'b'}}{\overline{ab}} = \frac{250}{f} \tag{10.5.271}$$

of the retinal images obtained through the magnifying glass and with the naked eye in the best conditions.

Remark

The horizontal field of view for a human is approximately 200 degrees, while the vertical field is about 135 degrees.

10.5.5.3 Triangular Prism

a prism is a transparent optical element with flat, polished surfaces that refract light. At least two of the flat surfaces must have an angle between them. The exact angles between the surfaces depend on the application. The traditional geometrical shape is that of a triangular prism with a triangular base and rectangular sides, and in colloquial use "prism" usually refers to this type. Some types of optical prism are not in fact in the shape of geometric prisms. Prisms can be made from any material that is transparent to the wavelengths for which they are designed. Typical materials include glass, plastic and fluorite.

A dispersive prism can be used to break light up into its constituent spectral colors (the colors of the rainbow). Furthermore, prisms can be used to reflect light, or to split light into components with different polarizations.

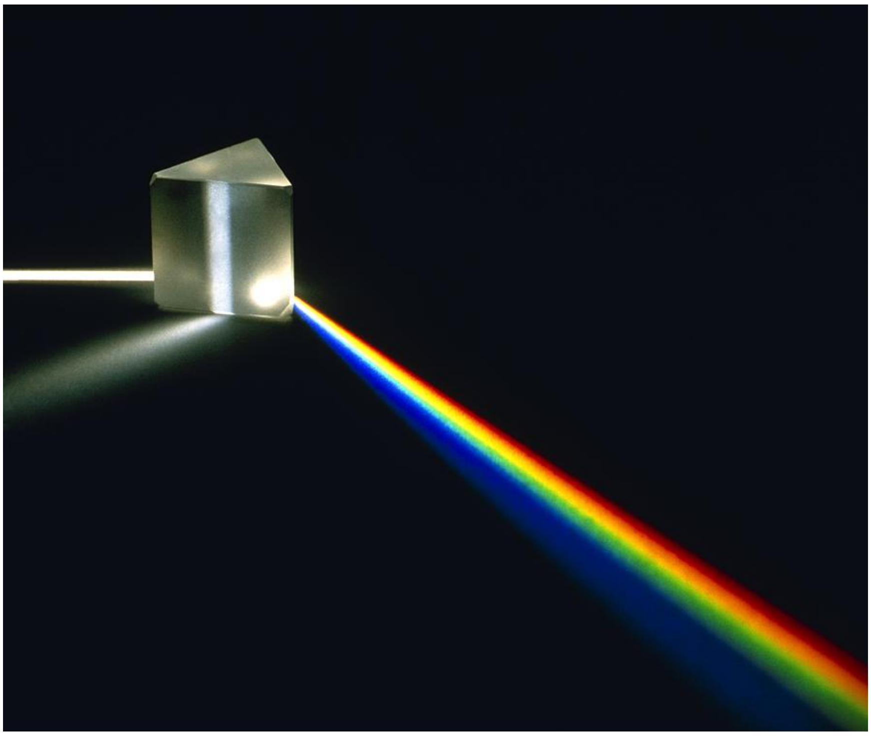


Figure 10.212: **Triangular prism dispersing white light ray (source: Wikipedia, David Parker)**

In optics, the triangular prism is one of the most important components of optics. It can be found in chemistry, condensed matter physics, astrophysics, optoelectronics and still many other popular devices of everyday life (such as lentils). This is probably the first tool shaped by human to make "spectroscopy" (spectrum analysis) after the rainbow sky obviously... which is a natural phenomenon spectroscopy.

We will in the following paragraphs identify the most important relations in relation to know about triangular prisms and useful to the engineer and physicist.

We will focus on light rays entering through one side and exiting by another having undergone two refractions (we do not study the internal reflections).

Here is the typical schematic representation of a triangular prism in optical geometry with the incident light ray S and outgoing one S' and the two normals N, N', to the edges of the top of opening angle α. More the different angles of incidence and refraction:

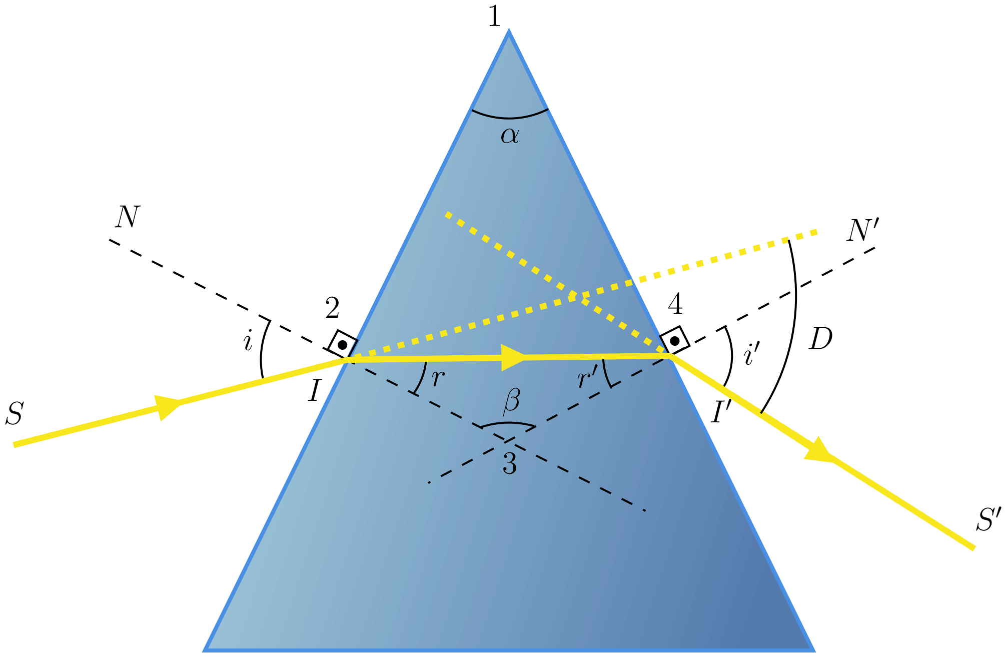


Figure 10.213: **Schematic triangular prism**

We know that the sum of the angles of a quadrilateral (always decomposable into two triangles whose the sum of the angles is π) is equals 2π (see section Geometric Shapes page 2593). So in the quadrilateral defined by the nodes 1234. We have the sum of:

$$\alpha+2\frac{\pi}{2}+\beta=2\pi \Rightarrow \beta=\pi-\alpha \quad (10.5.272)$$

Now that the situation has know let us focus on the optical part ...

We have 4 fundamental relations to prove for to the triangular prism.

First, we have at the point of incidence I and I' the Descartes' law that allows us to write:

$$\sin(i)=\frac{m}{n}\sin(r)=n_{21}\sin(r) \quad (10.5.273)$$

As the absolute refractive index of air is 1 we simply have in I:

$$\sin(i)=m\sin(r) \quad (10.5.274)$$

In the same idea in I' we have:

$$\sin(r')=\frac{n}{m}\sin(i')=\frac{1}{m}\sin(i') \quad (10.5.275)$$

and therefore:

$$\sin(i')=m\sin(r') \quad (10.5.276)$$

We also have the relation:

$$\pi=r+r'+\beta=r+r'+\pi-\alpha \quad (10.5.277)$$

Therefore:

$$\alpha = r + r' \tag{10.5.278}$$

The deflection angle D is easy to determine. We just have to take the central quadrilateral:

$$2\pi = i + i' + \beta + (\pi - D) \Rightarrow D = -\pi + i + i' + \beta = -\pi + i + i' + (\pi - \alpha) \tag{10.5.279}$$

Therefore:

$$D = i + i' - \alpha \tag{10.5.280}$$

So we have the 4 fundamental relations of the triangular prism:

$$\boxed{\begin{aligned} &\sin(i) = m\sin(r) \\ &\sin(i') = m\sin(r') \\ &\alpha = r + r' \\ &D = i + i' - \alpha \end{aligned}} \tag{10.5.281}$$

Knowing i and i' and the relative refractive index m then we can determine all parameters.

The ideal would still be able to get rid of the experimental knowledge of i'.

So we have:

$$i' = \arcsin(m\sin(r')) \tag{10.5.282}$$

But as:

$$r' = \pi - \beta - r = \pi - (\pi - \alpha) - r = \alpha - r = \alpha - \arcsin\left(\frac{\sin(i)}{m}\right) \tag{10.5.283}$$

Therefore it comes:

$$i' = \arcsin(m\sin(r')) = \arcsin\left\{m\sin\left[\alpha - \arcsin\left(\frac{\sin(i)}{m}\right)\right]\right\} \tag{10.5.284}$$

Therefore:

$$\boxed{D = i + \arcsin\left\{m\sin\left[\alpha - \arcsin\left(\frac{\sin(i)}{m}\right)\right]\right\}} \tag{10.5.285}$$

Since it is obvious that the index n of a material varies with the wavelength according to Cauchy law introduce at the beginning of this section, it is easy to understand that the triangular prism is capable of dispersing white light.

Finally if i is small:

$$D \cong \arcsin\left\{m\sin\left[\alpha - \arcsin\left(\frac{\sin(i)}{m}\right)\right]\right\} - \alpha \tag{10.5.286}$$

and if i and α are small, we have to first order in Maclaurin development (see section Sequences and Series 1703):

$$D\cong\arcsin\left\{m\sin\left[\alpha-\frac{i}{m}\right]\right\}-\alpha\cong\arcsin(m\alpha-i)-\alpha\cong\arcsin(m\alpha)-\alpha\cong m\alpha-\alpha=\alpha(m-1) \tag{10.5.287}$$

Therefore:

$$\boxed{D\cong\alpha(m-1)} \tag{10.5.288}$$

either by explicitly introducing Cauchy's law as introduced earlier in this section:

$$\boxed{D\cong\alpha(m_\lambda-1)=\alpha\left(\left(A+\frac{B}{\lambda^2}\right)-1\right)} \tag{10.5.289}$$

The fact that the index of refraction in glass depends on wavelength (Cauchy's law) is then the reason that prisms can spread the colors of the rainbow!

10.5.5.4 Pentaprism

The pentaprism roof that is enclosed in most viewfinder of single-lens reflex cameras is an adaptation of the pentaprism, or "optical square", invented in 11864 (holocene calendar) by Charles Moÿse Goulier.

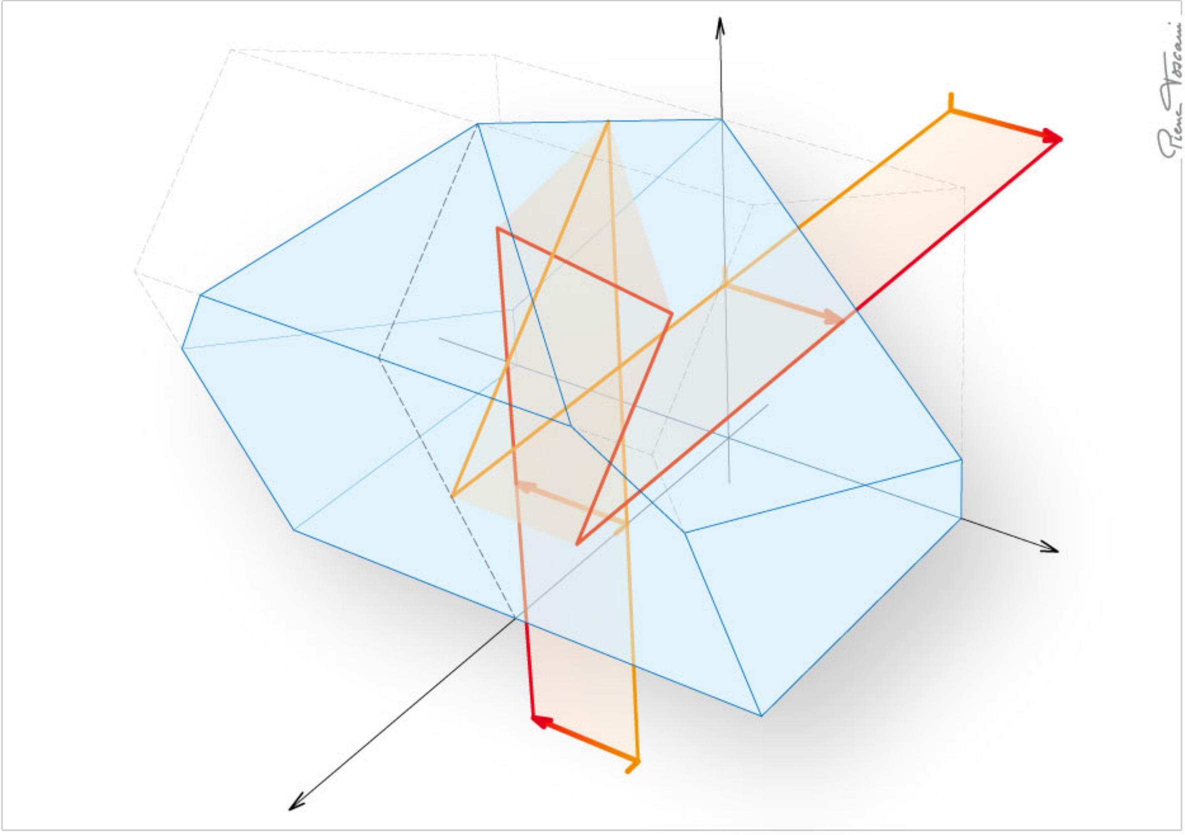


Figure 10.214: **Pentaprism roof (source: http://www.pierretoscani.com, author: Pierre Toscani)**

A pentaprism is a five-sided reflecting prism used to deviate a beam of light by a constant 90°, even if the entry beam is not at 90° to the prism (see proof below). The beam reflects inside the prism twice, allowing the transmission of an image through a right angle without inverting it vertically (that is, without changing the image's handedness) as an ordinary right-angle prism or mirror would.

The reflections inside the prism are not caused by total internal reflection, since the beams are incident at an angle less than the critical angle (the minimum angle for total internal reflection). Instead, the two faces are coated to provide mirror surfaces. The two opposite transmitting faces are often coated with an anti-reflection coating to reduce spurious reflections. The fifth face of the prism is not used optically but truncates what would otherwise be an awkward angle joining the two mirrored faces.

As we have show it, the camera lens renders an image that is both vertically and laterally reversed, and the reflex mirror re-inverts it leaving an image laterally reversed. In this case, the image needs to be reflected left-to-right as the prism transmits the image formed on the camera's focusing screen. This lateral inversion is done by replacing one of the reflective faces of a normal pentaprism with a "roof" section, with two additional surfaces angled towards each other and meeting at 90°, which laterally reverses the image back to normal.

In a precision optical device, when the path of a light beam must be deflected at 90°, the simplest solution is to use a plane mirror oriented at 45° relatively the direction of the incident beam (see figure below). The mirror must be accurately positioned and stable because any angular positioning error induces a double error in the direction of the reflected beam: the reflected beam rotates twice as fast as the mirror (this is obvious as when the mirror in oriented at $\alpha = 45^\circ$ the output beam is at an output angle of $2\alpha = 90^\circ$ and therefore in general any modification δ implies a deflection of 2δ):

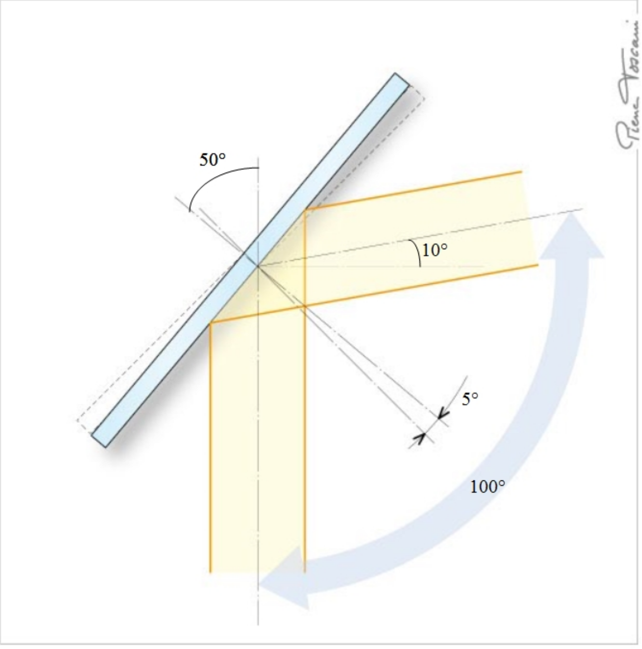
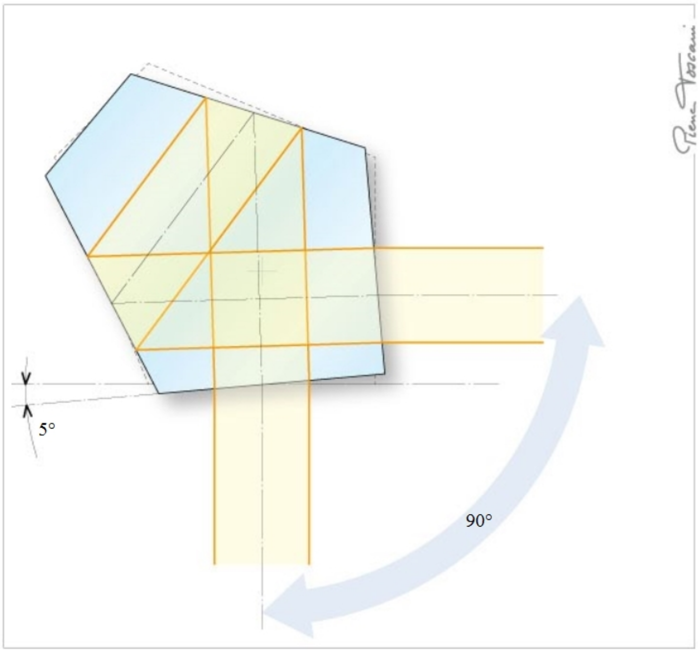


Figure 10.215: **Mirror double deflection property**
(source: http://www.pierretoscani.com, author: Pierre Toscani)

But for the pentaprism it is different... It behaves like a set square:



Before we deal with the maths, the reader should be aware that this pentaprism has the following properties:

- Two transparent faces (input or output) forming an angle of $90°$
- Two reflecting faces forming an angle of $45°$, separated by a chamfer (fifth face, no function)
- The both reflective surfaces have to be metallized to provide good reflection coefficient.

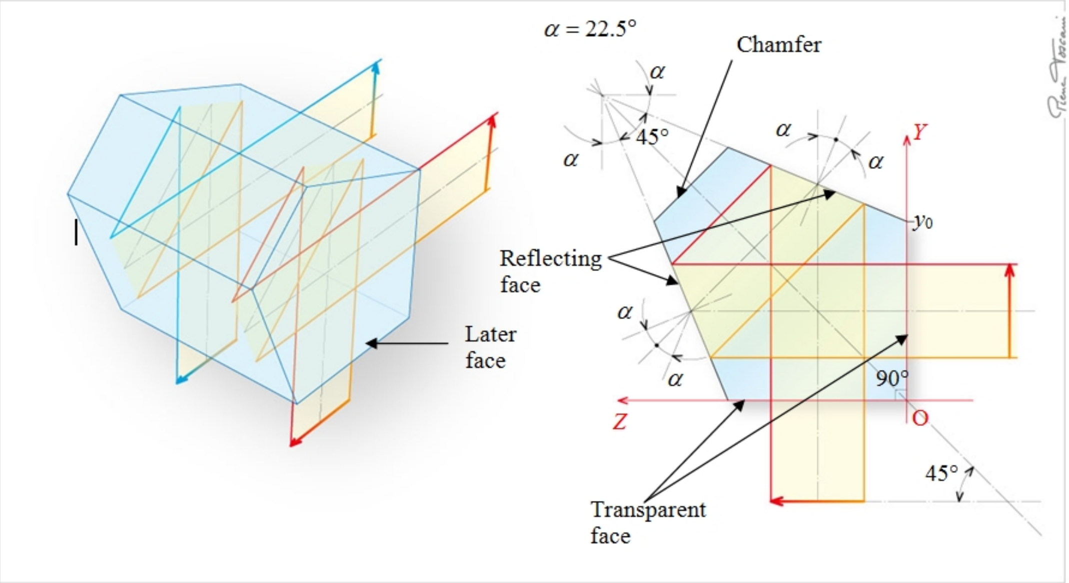
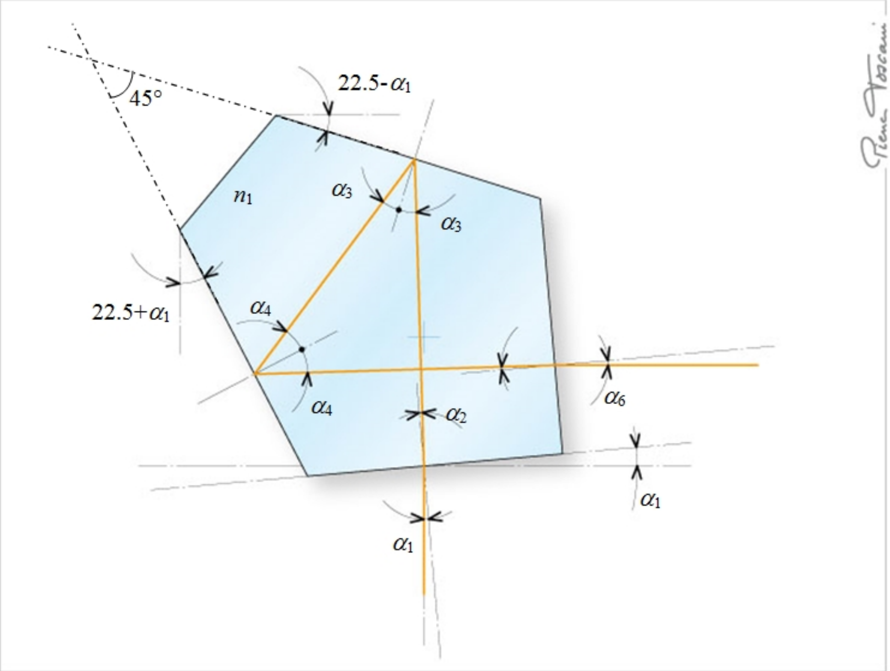

$\alpha = 22.5°$
Chamfer
α
45°
α
α
α
Y
y_0
Reflecting face
α
Later face
α
90°
Z
O
45°
Transparent face

Figure 10.216: **Pentaprism detailed design (source: http://www.pierretoscani.com, author: Pierre Toscani)**

Now to prove that property consider the following figure:


45°
22.5-α_1
n_1
α_3
α_3
22.5+α_1
α_4
α_4
α_2
α_6
α_1
α_1

We have then looking this figure:

$$n\sin(\alpha_1)=m\sin(\alpha_2) \tag{10.5.290}$$

That we will rearrange and write:

$$\sin(\alpha_1)=\frac{m}{n}\sin(\alpha_2)=n_r\sin(\alpha_2) \tag{10.5.291}$$

Now in the above triangle we have:

$$\alpha_3=(22.5^\circ-\alpha_1)+\alpha_1-\alpha_2 \tag{10.5.292}$$

Therefore:

$$\alpha_3=22.5^\circ-\alpha_2 \tag{10.5.293}$$

On the left triangle we have:

$$\alpha_4=90^\circ-(22.5^\circ+\alpha_1)-(22.5^\circ-\alpha_1)-\alpha_3 \tag{10.5.294}$$

Therefore:

$$\alpha_4=45^\circ-\alpha_3 \tag{10.5.295}$$

That is:

$$\alpha_4=22.5^\circ+\alpha_2 \tag{10.5.296}$$

Finally in the last triangle we have:

$$\alpha_5=\alpha_1-(22.5^\circ+\alpha_1)+\alpha_4 \tag{10.5.297}$$

And therefore:

$$\alpha_5=\alpha_4-22.5^\circ \tag{10.5.298}$$

That is:

$$\alpha_5=\alpha_4-22.5^\circ \tag{10.5.299}$$

Finally:

$$\alpha_5=\alpha_2 \tag{10.5.300}$$

And now at the exit of the pentaprism we have:

$$\sin(\alpha_6)=n_r\sin(\alpha_5) \tag{10.5.301}$$

But as $\alpha_5=\alpha_2$ we have:

$$\sin(\alpha_6)=n_r\sin(\alpha_5)=n_r\sin(\alpha_2) \tag{10.5.302}$$

Therefore it follows that:

$$\alpha_6=\alpha_1 \tag{10.5.303}$$

And this finish our proof!

~90 %

★★★★☆

48 votes, 76.25%

10.6. Wave Optics

In this section will study some elements that led to the development of quantum mechanics. Indeed, quantum mechanics was born, first, by a careful study of the nature of light. Although this new science was developed in the early 120th century (holocene calendar), the considerations which have guided to it are undoubtedly the result of 25 centuries of maturation. Basically, it is a long history of the light full of controversies to which quantum mechanics in the 120th century finally brings a masterful conclusion.

Electron microscopes can make images of individual atoms, but why will a visible-light microscope never be able to? Stereo speakers create the illusion of music that comes from a band arranged in your living room, but why doesn't the stereo illusion work with bass notes? Why are computer chip manufacturers investing billions of dollars in equipment to etch chips with X-rays instead of visible light?

The answers to all of these questions have to do with the subject of wave optics. So far this textbook has discussed the interaction of light waves with matter, and its practical applications to optical devices like mirrors, but we have used the ray model of light almost exclusively. Hardly ever have we explicitly made use of the fact that light is an electromagnetic wave. We were able to get away with the simple ray model because the chunks of matter we were discussing, such as lenses and mirrors, were thousands of times larger than a wavelength of light. We now turn to phenomena and devices that can only be understood using the wave model of light.

10.6.1 Huygens' principle

Huygens visualized the propagation of light as a result of the process of a generation of spherical wavelets in each point reached by a wavefront, wavelet whose sum gave the propagation field. By drawing the tangent to the wavefronts of the wavelet at a given time, one obtained the wavefront of the total wave at the same moment.

Caution! As we have already mentioned it earlier, the Fermat's principle doesn't explain anything, even if it's accurate that light follow the shortest path! That's why most textbooks introducing refraction as we did above geometrically (named sometimes the "marching soldier analogy", but wrong as nothing explain why the solders don't continue you walk straight) or using the variational principal (i.e. "Fermat's principle") are quite misleading. Using Huygens principle as we will do it now isn't also accurate as we will explain it (even if we fall back on the same result as using Fermat's principle or the marching soldier analogy)! In fact we should use Maxwell equation to understand better what happens physically!

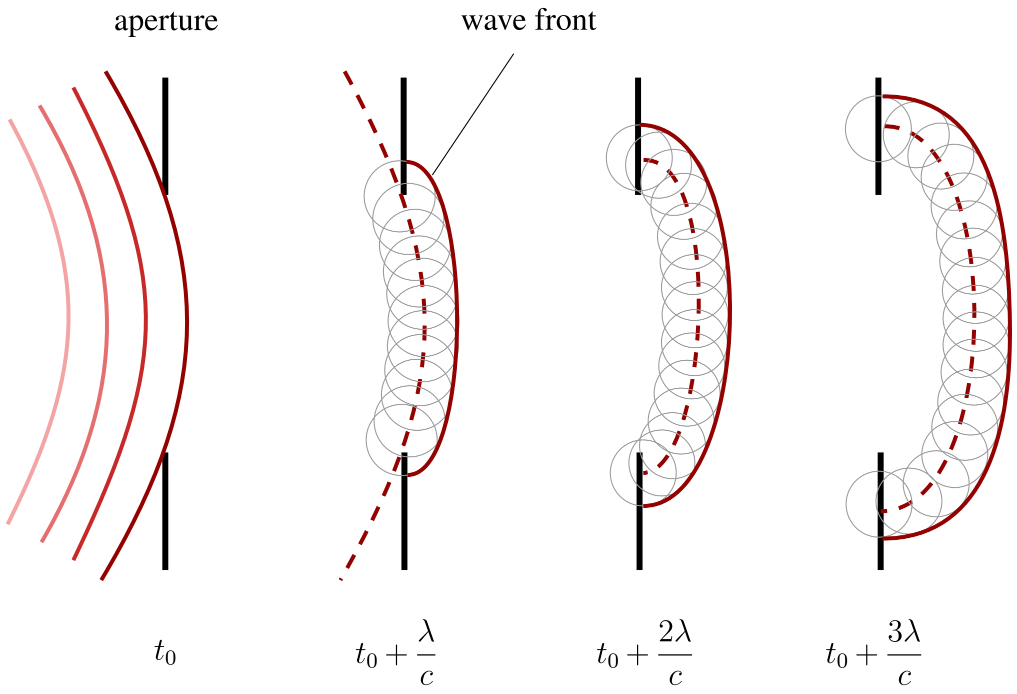


Figure 10.217: **Illustration of Huygen's construction of a propagating wave**

Definition (#10.6.496): A "wavefront surface" is the localization of points of the medium reach by the wave motion at the same instant. The disturbance has therefore the same phase at any point of a wavefront. For a plane wave, for example, the disturbance is expressed by (we have prove this in the section of Wave Mechanics):

$$f(kx-vt) \tag{10.6.1}$$

or in a more general way:

$$f(\vec{k}\circ\vec{r}-vt) \tag{10.6.2}$$

which therefore gives the expression of the propagation of the disturbance for which the "wavefront" is the locus of points where the phase $\vec{k}\circ\vec{r}-vt$ has the same value at a given instant. The wave surface is given accordingly by the equation:

$$\vec{k}\circ\vec{r}-vt=c^{te} \tag{10.6.3}$$

Huygens gave a pictorial representation method of the passage of a wave surface to another in the case where the wave is assumed to result from the movement of the particles constituting the material medium. Thus, if we consider the wave surface S below:

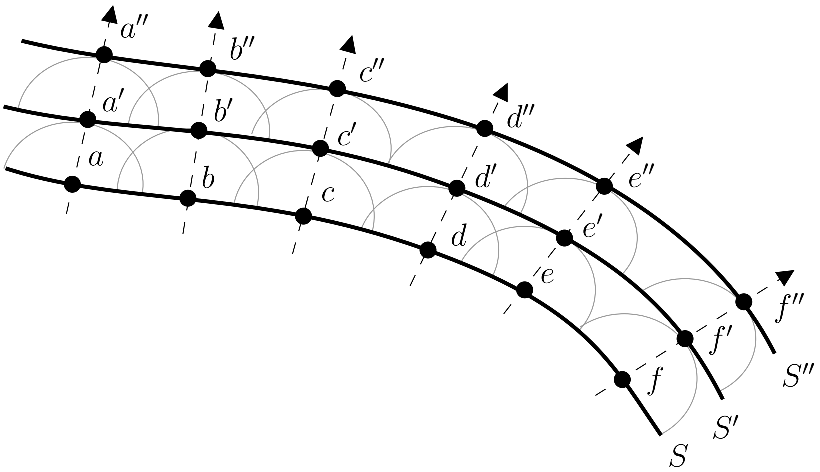


Figure 10.218: **Representation of a wavefront according to Huygens**

When the wave motion reaches this surface, each particle $a, b, c, \ldots$ of the surface becomes in turn a wave source, emitting secondary waves (indicated by the small half-circles in the figure above) that reach the next layer of particles of the medium. These particles are put in motion and form the new wave surface S' and so on... So, Huygens had a wave conception of light, but he did not consider the periodic nature of the wave, which not allow him to introduce the concept of light of color (frequency); moreover, according to its principle, a wave propagating in the opposite direction to that of the incident wave should also occur, which is not the case in a homogeneous material...

Figure 10.219: **Crepuscular rays as a real life example of Huygens-Fresnel secondary source effect**

The intuition of Huygens is however close to reality as will show if Fresnel in its diffraction theory (see further below). However, it was Kirchhoff, which will introduce a tilt factor (obliquity) in the theory, to gen an explanation of the absence of wave propagating backward (when the time will come we will write related mathematical developments).

As on the previous figure all the "corresponding points" $(a,b,c,\ldots),(a',b',c',\ldots)$ are equidistant, by Huygens' principle, the time interval between corresponding points of two wavefronts is the same for any two corresponding points!

The consequences are (refer simultaneously to the figure below):

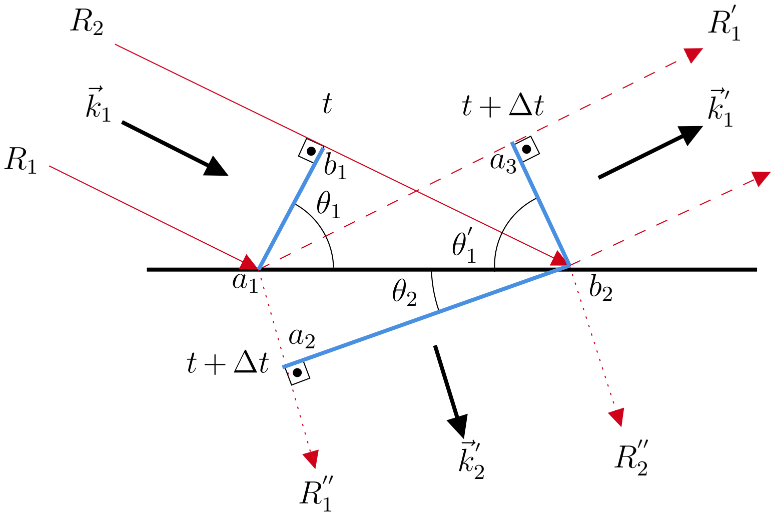


Figure 10.220: **Schematic diagram**

If we denote the propagation velocities of the incident rays R_1, R_2 by v_1 and v_2 respectively, we have:

$$\overline{b_1b_2}=v_1\Delta t \qquad \overline{a_1a_3}=v_1\Delta t \qquad \overline{a_1a_2}=v_2\Delta t \tag{10.6.4}$$

- When the wave propagates in a homogeneous medium, the light rays must be straight and the wave surfaces remain parallel.
- When the wave change of medium, the distance between two pairs of corresponding points vary from one medium to another, if the propagation velocities are different.

Theorem 10.3. Let us prove that this principle allows to find the Descartes-Snellius law that we have already proven in the section of Geometric Optics, which ensures a priori that the Huygens principle is still valid in the context of Geometric Optics.

Proof 10.3.1. Following the figure above, we have:

$$\Delta t=\frac{\overline{b_1b_2}}{v_1}=\frac{\overline{a_1a_3}}{v_1}=\frac{\overline{a_1a_2}}{v_2} \tag{10.6.5}$$

by dividing each term by $\overline{a_1b_2}$, we get:

$$\frac{\sin(\theta_1)}{v_1}=\frac{\sin(\theta'_1)}{v_1}=\frac{\sin(\theta_2)}{v_2} \tag{10.6.6}$$

so we fall back well on the Descartes-Snellius law as we had obtained it in the section of Geometric Optics:

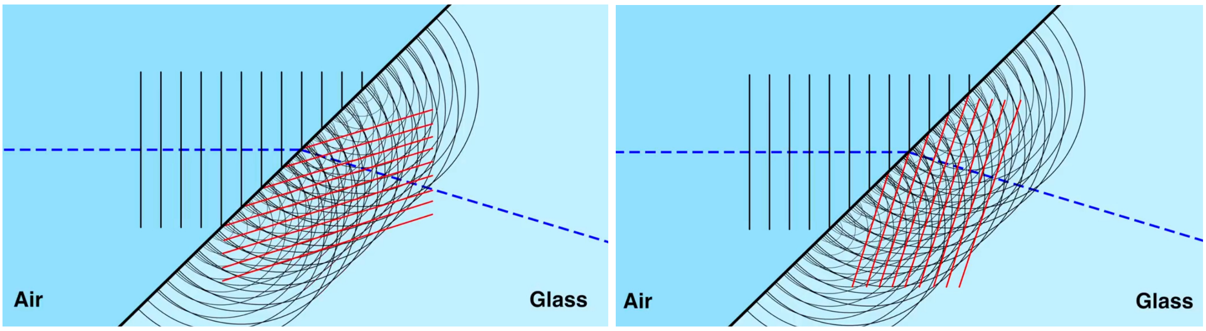
$$\frac{\sin(\theta_1)}{\sin(\theta_2)}=\frac{v_1}{v_2}=n_{12} \tag{10.6.7}$$

noting on the way that on the diagram, we also have $\theta_1=\theta'_1$.

■ Q.E.D.

So to derive Snell's law we have seen so far three commonly suggested explanations that are wrong, or at least incomplete. They are for Fermat's principle, the analogy of soldiers marching and Huygens' principle. Fermat's principle says that light will travel from one point to the other, taking the minimum amount of time. It's often explained in terms of a lifeguard in a drowning swimmer in order to maximize the chances that the swimmer will be saved, the lifeguard needs to get to them as soon as possible. But while that's true for light, it doesn't explain anything. It just says what it does, not why it does it. So that's not an explanation and we need to come up with a proper reason. Let's look at the second option on the list. A very common explanation used to teach why light bends involves soldiers marching over firm ground, but then who encounter a wall, although not all soldiers hit mud at the same time. The idea is simple. The first soldier who encounters the mud cannot move as quickly as the rest of the soldiers, so he slows down. Each soldier hits the mud and turn, resulting in a direction change. Eventually the lines of soldiers are moving in a different direction at a different speed. This is how it's often taught. The problem is that this doesn't work. In fact, the soldiers will slow down, but their direction doesn't change. They will continue to walk in the same direction, but with marching lines that are slanted compared to how they started. This explanation gets the angle of the lines of soldiers right but not the direction of travel. In fact, the only way that this explanation can be right is if the lines of soldiers are rigid, then the first soldier hitting the mud puts a torque on the entire line and that would work. The problem is that this torque means that the soldiers on the top of the screen move faster, which is to say that if they were light, this part of the beam of light would have to move faster than light, so this explanation doesn't work either.

What about Huygens' principle? Remember in the figure on the left below that the lines denote the top of the peaks of the waves. We can see that there are many places where the peaks line up. We can overlay a couple of lines that show the direction that the wave is travelling and then remove the circular waves and we see that this explanation seems to be doing a good job of predicting what happens when light goes from air to glass. Looks good, right? But not so fast! What happens when we look at all of the waves coming through the glass? Things get a lot more complicated. We do see that the waves add together to predict the direction that light moves in glass. On the other hand, that's not the only alignment! For instance, we see other geometry's where the waves line up a visible in the figure on the right below.



(a) What most textbooks show (b) What only a few textbooks admit...

Figure 10.221: **Failure of Huygens' principle**

So it seems that this approach doesn't give a unique prediction. OK, we've ruled out Fermat's principle, the soldier analogy, and Huygens' principle!

These are all often seen as explanations of the cause of refraction given by people who really do know some physics. So what's the real answer? It turns out that the only way to really answer the question of why light bends when it goes from air to glass is to get serious about the nature of light and to embrace the fact that it is made of oscillating electromagnetic fields, and that means we need Maxwell's equations!

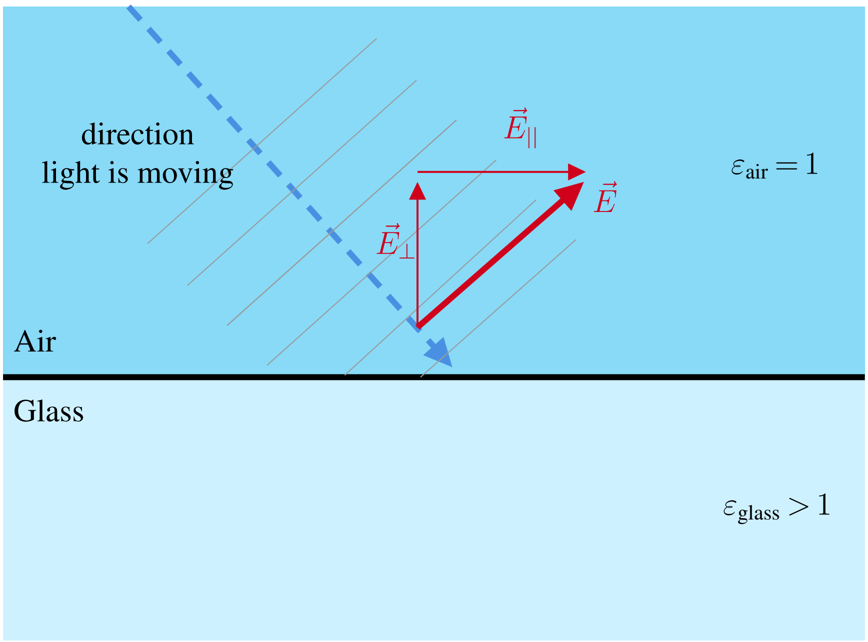
We will focus here only on the electric fields. So we start with Maxwell's equations:

$$\begin{aligned} \vec{\nabla}\circ\vec{E} &= \frac{\rho_Q}{\varepsilon} & \vec{\nabla}\circ\vec{B} &= 0 \\ \vec{\nabla}\times\vec{E} &= -\frac{\partial\vec{B}}{\partial t} & \vec{\nabla}\times\vec{B} &= \mu\vec{j}+\mu\varepsilon\frac{\partial\vec{E}}{\partial t} \end{aligned} \tag{10.6.8}$$

In fact, we're going to need only the bottom two equations:

$$\varepsilon\vec{\nabla}\circ\vec{E} = \rho_Q \qquad \vec{\nabla}\times\vec{E} = -\frac{\partial\vec{B}}{\partial t} \tag{10.6.9}$$

So let's see what's going on. We start with light going from air to glass, hitting the surface at an angle. In our figure, we can replace the waves with the direction of motion. Now it turns out that the electric field of light is perpendicular to the direction that light is travelling, and we can add that field direction to the diagram. And it's very important to remember that this field has components. One parallel to the surface of the glass and one that is perpendicular to the glass:


direction
light is moving
$\vec{E}_{||}$
$\vec{E}$
$\vec{E}_{\perp}$
$\varepsilon_{\text{air}} = 1$
Air
Glass
$\varepsilon_{\text{glass}} > 1$

And here is where Maxwell's equations come into play!

Two copies of the equations are written here, one that covers when light is travelling in air:

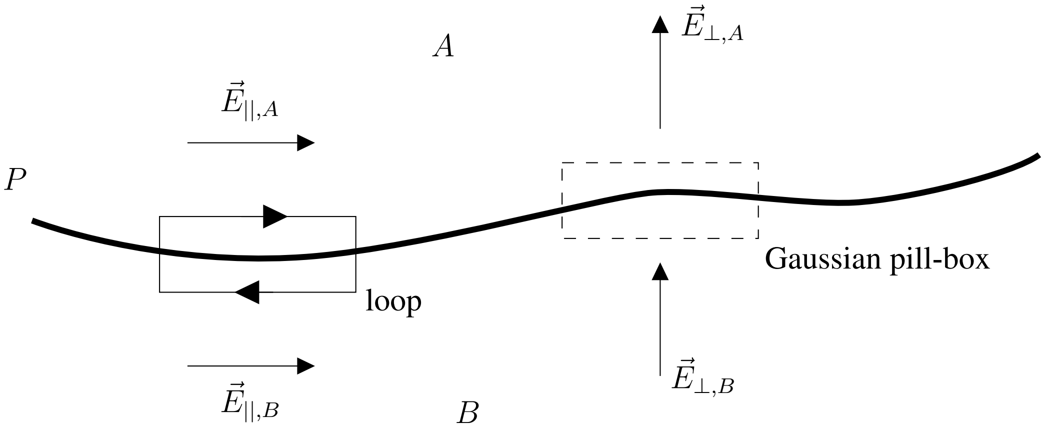
$$\varepsilon_{\text{air}}\vec{\nabla}\circ\vec{E}_{\text{air}} = \rho_Q \qquad \vec{\nabla}\times\vec{E}_{\text{air}} = -\frac{\partial\vec{B}}{\partial t} \tag{10.6.10}$$

and one where it is travelling in glass:

$$\varepsilon_{\text{glass}}\vec{\nabla}\circ\vec{E}_{\text{glass}} = \rho_Q \qquad \vec{\nabla}\times\vec{E}_{\text{glass}} = -\frac{\partial\vec{B}}{\partial t} \tag{10.6.11}$$

So here's the key point! The surface belongs to both the air region and the glass region. This means that at the surface, the equations on the top and the equations on the bottom have to apply, and with that little bit of calculus we can find two important restrictions. Let us see which are these restrictions?

For that purpose let us see what are the most general boundary conditions satisfied by the electric field at the interface between two media: e.g., the interface between a vacuum and a conductor? Consider an interface P between two media A (for example "air") and B (for example "glass"):


A
$\vec{E}_{\perp,A}$
$\vec{E}_{||,A}$
P
Gaussian pill-box
loop
$\vec{E}_{||,B}$
$\vec{E}_{\perp,B}$
B

Let us, first of all, apply Gauss' law (see section Electrodynamics page 4095):

$$\oint_S \vec{E}\circ\mathrm{d}\vec{S}=\frac{1}{\varepsilon_0}\int_V \rho\mathrm{d}V \tag{10.6.12}$$

to a Gaussian pill-box S of cross-sectional area A whose two ends are locally parallel to the interface (see figure above). The ends of the box can be made arbitrarily close together. In this limit, the flux of the electric field out of the sides of the box is obviously negligible. The only contribution to the flux comes from the two ends. In fact:

$$\oint_S \vec{E}\circ\mathrm{d}\vec{S}=(E_{\perp A}-E_{\perp B})A \tag{10.6.13}$$

where $E_{\perp A}$ is the perpendicular (to the interface) electric field in medium A at the interface, etc. The charge enclosed by the pill-box is simply σ_A, where σ is the sheet charge density on the interface. Note that any volume distribution of charge gives rise to a negligible contribution to the right-hand side of the above equation, in the limit where the two ends of the pill-box are very closely spaced. Thus, Gauss' law yields:

$$E_{\perp A}-E_{\perp B}=\frac{\sigma}{\varepsilon_0} \tag{10.6.14}$$

at the interface: i.e., the presence of a charge sheet on an interface causes a discontinuity in the perpendicular component of the electric field.

The way electric permittivity works is that it lets us use Gauss' law as if there is no charge, provided we integrate $\vec{D}=\varepsilon\vec{E}$ over the surface rather than $\vec{E}$. Therefore the boundary condition is $D_{\perp A}=D_{\perp B}$, which implies:

$$\varepsilon_A E_{\perp A}=\varepsilon_B E_{\perp B} \tag{10.6.15}$$

What about the parallel electric field? Let us apply Faraday's law (see section Electrokinetics page 4259) to a rectangular loop C whose long sides, length l, run parallel to the interface:

$$\oint_C \vec{E}\circ\mathrm{d}\vec{l}=-\frac{\partial}{\partial t}\int_S \vec{B}\circ\mathrm{d}\vec{S} \tag{10.6.16}$$

(see figure above). The length of the short sides is assumed to be arbitrarily small. The dominant contribution to the loop integral comes from the long sides:

$$\oint_C \vec{E}\circ\mathrm{d}\vec{l}=(E_{\parallel A}-E_{\parallel B})l \tag{10.6.17}$$

where $E_{\parallel A}$ is the parallel (to the interface) electric field in medium A at the interface, etc. The flux of the magnetic field through the loop is approximately $B_{\perp A}$, where $B_\perp$ is the component of the magnetic field which is normal to the loop, and A is the area of the loop. But, $A\rightarrow 0$ as the short sides of the loop are shrunk to zero. So, unless the magnetic field becomes infinite (we shall assume that it does not), the flux also tends to zero. Thus:

$$E_{\parallel A}-E_{\parallel B}=0 \tag{10.6.18}$$

i.e., there can be no discontinuity in the parallel component of the electric field across an interface.

So, coming back to our glass and air case, this leads us to:

$$E_\parallel^{\text{glass}}=E_\parallel^{\text{air}}\quad\text{and}\quad\varepsilon_{\text{glass}}E_\perp^{\text{glass}}=\varepsilon_{\text{air}}E_\perp^{\text{air}} \tag{10.6.19}$$

Remark

For the magnetic field, the boundary conditions will be affected due to the magnetic moment of the charges in the medium, as encoded in the permeability μ. Since a magnetic field induces in a material a current perpendicular to the field, only $B_{\parallel}$ will be affected (because of the change of material perpendicular to the surface in the plane where the induced current should happen), not $B_{\perp}$.

Thus:

$$B_{\perp A} = B_{\perp B} \tag{10.6.20}$$

Since the part of the magnetic field which is sensitive to an accumulated current in a material is treated using $\vec{H} = \vec{B}/\mu$ the condition is then that:

$$\frac{1}{\mu_A} B_{\parallel A} = \frac{1}{\mu_B} B_{\parallel B} \tag{10.6.21}$$

In transparent materials, the electric permittivity can vary significantly in comparison to air, but so does not the magnetic permeability. So some textbooks assume that $\mu_A \cong \mu_B$. In this case, the four boundary conditions:

$$E_{\parallel}^{\text{glass}} = E_{\parallel}^{\text{air}} \quad \text{and} \quad \varepsilon_{\text{glass}} E_{\perp}^{\text{glass}} = \varepsilon_{\text{air}} E_{\perp}^{\text{air}} \tag{10.6.22}$$

$$\frac{1}{\mu_{\text{glass}}} B_{\parallel}^{\text{glass}} = \frac{1}{\mu_{\text{air}}} B_{\parallel}^{\text{air}} \quad \text{and} \quad B_{\perp}^{\text{glass}} = B_{\perp}^{\text{air}} \tag{10.6.23}$$

reduce to:

$$E_{\parallel}^{\text{glass}} = E_{\parallel}^{\text{air}} \quad \text{and} \quad \varepsilon_{\text{glass}} E_{\perp}^{\text{glass}} = \varepsilon_{\text{air}} E_{\perp}^{\text{air}} \tag{10.6.24}$$

$$\vec{B}_{\text{glass}} = \vec{B}_{\text{air}} \tag{10.6.25}$$

The first is that the electric field parallel to the surface of the glass has to be the same in the air in the glass and similarly perpendicular to the surface:

$$\varepsilon_{\text{glass}} \vec{\nabla} \circ \vec{E}_{\text{air}} = \rho_Q = \varepsilon_{\text{air}} \vec{\nabla} \circ \vec{E}_{\text{air}} \tag{10.6.26}$$

$$\vec{\nabla} \times \vec{E}_{\text{air}} = -\frac{\partial \vec{B}}{\partial t} = \vec{\nabla} \times \vec{E}_{\text{glass}} \tag{10.6.27}$$

Therefore:

$$\varepsilon_{\text{glass}} \vec{\nabla} \circ \vec{E}_{\text{air}} = \varepsilon_{\text{air}} \vec{\nabla} \circ \vec{E}_{\text{air}} \tag{10.6.28}$$

$$\vec{\nabla} \times \vec{E}_{\text{air}} = \vec{\nabla} \times \vec{E}_{\text{glass}} \tag{10.6.29}$$

Because ε is bigger in glass than in air, that means that the perpendicular electric field in glass has to be smaller than it is in air.

Now we remember that the direction the late travels is perpendicular to the electric field, so we can put in an arrow to show the direction light must travel in the glass. And finally, we can see what light does when it enters glass or water or any transparent medium, it bends. And the reason that it bends is because the epsilon in glass is bigger than an air.

OK, what we just seen is an equation thing. The reader may probably asking himself what that epsilon actually physically means. It's there because how the electric fields from the light interacts with the matter in the glass. We start out with glass with no electric field in it. The glasses charges in it, but they're arranged in a random way. But when we send light in, we impose an electric field on it. That field makes the charges move around, which sets up a counterbalancing electric field from the charges. The result is that the electric field in the glass is lower than it is in air because of how the electric field from the glass is in the opposite direction. And this is the reason that the perpendicular electric field is lower in glass, and that is why light bends when it goes from here to glass. It's not because of many of the ordinary explanations, it's because of how light interacts with glass and changes the glasses properties, and it's because the electric field inside the glass is affected by the arrangements of atoms and molecules in the glass when light hits it. It's the same reason why. Light slows down in matter inside matter. The electric field due to the light and the electric field due to the matter both have to be taken into account. Outside they don't. And now you know some serious physics! We imagine that some of you will may ask about the quantum explanation, which is similar to the one we just told you here using Maxwell equations except using energy and momentum of the photons instead.

10.6.1.1 Normal refraction

Let us start with the case of normal incidence (perpendicular to the interface). To find out how much light is reflected, we need to work out the impedance, which determines the reflection and transmission coefficients.

For normal incidence the incident angle to the normal is $\theta_1=0$ (see the figure 10.223 below). Thus $E_\perp=B_\perp=0$. That is, the electric and magnetic fields are both polarized in the plane of the interface so there simply is no $\perp$ component.

Without loss of generality, let's take a plane wave of frequency ω moving in the z direction with $\vec{E}$ in the y direction. Then since (see page 4145):

$$\vec{B}=\frac{1}{ck}\vec{k}\times\vec{E} \tag{10.6.30}$$

the magnetic field points in the x direction. So the incident fields are:

$$\vec{E}_I=E_I\vec{u}_y e^{\mathrm{i}\omega\left(t-\frac{1}{v_1}z\right)} \qquad \vec{B}_I=B_I\vec{u}_x e^{\mathrm{i}\omega\left(t-\frac{1}{v_1}z\right)} \tag{10.6.31}$$

The transmitted and reflected waves are also moving normal to the surface (by Snell's law), so we can write:

$$\begin{aligned}\vec{E}_T&=E_T\vec{u}_y e^{\mathrm{i}\omega\left(t-\frac{1}{v_1}z\right)} & \vec{B}_T&=B_T\vec{u}_x e^{\mathrm{i}\omega\left(t-\frac{1}{v_1}z\right)}\\ \vec{E}_R&=E_R\vec{u}_y e^{\mathrm{i}\omega\left(t+\frac{1}{v_1}z\right)} & \vec{B}_R&=-B_R\vec{u}_x e^{\mathrm{i}\omega\left(t+\frac{1}{v_1}z\right)}\end{aligned} \tag{10.6.32}$$

with E_T,E_R,B_T and B_R the transmission and reflection coefficients to be determined. Note that we have flipped the sign on the z term in the phase since the reflected wave is moving in the $-\vec{u}_z$ direction (that is, $\vec{k}$ flips).

Since $\omega\vec{B}=\vec{k}\times\vec{E}$, this means that $\vec{B}$ flips sign too, which explains the minus sign in $\vec{B}_R$.

Since there is no $\perp$ component and that we have the boundary condition $E_{\parallel}^{(1)}=E_{\parallel}^{(2)}$, we will assume that this implies:

$$E_I+E_R=E_T \tag{10.6.33}$$

Similarly we will assume that:

$$\frac{1}{\mu_1}B_{\parallel}^{(1)}=\frac{1}{\mu_2}B_{\parallel}^{(2)} \tag{10.6.34}$$

implies:

$$\frac{1}{\mu_1}B_I-\frac{1}{\mu_1}B_R=\frac{1}{\mu_2}B_T \tag{10.6.35}$$

Since $||\vec{B}||=\frac{1}{v}||\vec{E}||$ for any plane wave, we have obviously:

$$B_I=\frac{1}{v_1}E_I \qquad B_T=\frac{1}{v_2}E_T \tag{10.6.36}$$

and the above relation becomes:

$$\frac{1}{\mu_1 v_1}(E_I-E_R)=\frac{1}{\mu_2 v_2}E_T \tag{10.6.37}$$

Combining:

$$\begin{gathered}E_I+E_R=E_T\\ \frac{1}{\mu_1 v_1}(E_I-E_R)=\frac{1}{\mu_2 v_2}E_T\end{gathered} \tag{10.6.38}$$

and solving gives:

$$E_R=E_I\frac{Z_2-Z_1}{Z_2+Z_1} \qquad E_T=E_I\frac{2Z_2}{Z_2+Z_1} \tag{10.6.39}$$

where:

$$Z_i=\mu_i v_i=\mu_i\frac{c}{n_i}=\sqrt{\frac{\mu_i}{\varepsilon_i}} \tag{10.6.40}$$

For most materials, μ_i is pretty constant, so the form $Z=\mu_i c/n_i$, which says $Z\propto n_i^{-1}$, is most useful.

What is the power reflected and transmitted assuming the above assumptions correct? The incident power in a medium is given by the Poynting vector (see page 4156):

$$\vec{S}=\frac{1}{\mu}\vec{E}\times\vec{B} \tag{10.6.41}$$

times area A. So:

$$P_I=\frac{1}{\mu v}||\vec{E}_I||^2A=\frac{1}{Z}||\vec{E}_I||^2A \tag{10.6.42}$$

Thus the reflected power is:

$$P_R = \frac{E_R^2}{Z_1}A = \left(\frac{Z_2 - Z_1}{Z_2 + Z_1}\right)^2 \frac{E_I^2}{Z_1}A = \left(\frac{Z_1 - Z_2}{Z_1 + Z_2}\right)^2 P_I \qquad (10.6.43)$$

and the transmitted power is:

$$P_T = \frac{E_T^2}{Z_2}A = \left(\frac{2Z_2}{Z_2 + Z_1}\right)^2 \frac{E_I^2}{Z_2}A = \frac{4Z_1Z_2}{(Z_1+Z_2)^2}\frac{E_I^2}{Z_1}A = \frac{4Z_1Z_2}{(Z_1+Z_2)^2}P_I \qquad (10.6.44)$$

These satisfy $P_T + P_R = P_I$ as expected. Note that these equations hold for normal incidence only and the others assumptions!

10.6.1.2 Thin film interference

Using the relations we get from the above developments:

$$E_R = E_I\frac{Z_2 - Z_1}{Z_2 + Z_1} \qquad E_T = E_I\frac{2Z_2}{Z_2 + Z_1} \qquad (10.6.45)$$

and that since $Z \propto n^{-1}$, the first relation becomes:

$$\frac{E_R}{E_I} \propto \frac{n_1 - n_2}{n_1 + n_2} \qquad (10.6.46)$$

Thus for $n_2 > n_1$ the reflected electric field has a phase flip compared to the incoming field, and for $n_1 < n_2$ there is no phase flip.

So what happens when light hits a thin film?

The film will generally have $n_2 = n_{\text{film}} > n_{\text{air}} = n_1$. Thus we get a phase flip at the top surface, but no phase flip at the bottom surface. The picture looks like this:

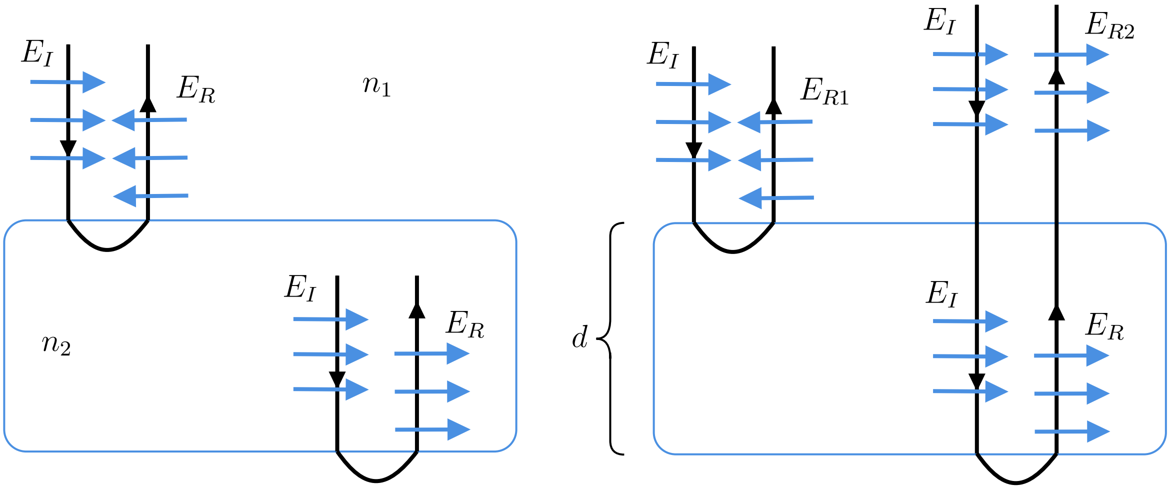


Figure 10.222: **There is a π phase flip from reflections off the top surface, where $n_2 > n_1$, but not off the bottom surface, where $n_2 < n_1$.**

So what happens if light of wavelength λ hits a film of thickness d?

There will be two reflected waves, one off the top surface (A) of the film and one off the bottom (B) of the film which will interfere!

Say the incoming electric field on (A) is:

$$E_{I,(A\to A)} = E_{I,(A)}\cos(kz-\omega t) \tag{10.6.47}$$

pointing to the right. Then the wave which reflects off the top surface $(z=0)$ will be:

$$E_{R1,(A\to A)} = RE_{I,(A)}\cos(-\omega t-\pi) = -RE_{I,(A)}\cos(-\omega t) \tag{10.6.48}$$

with R being the reflection coefficient. This now points to the left!

The wave which gets through the surface (A) from the film will be:

$$E_{I,(A\to B)}T\cos\left(kz-\omega t\right) \tag{10.6.49}$$

with T the transmission coefficient. This wave then goes to the bottom surface, reflects with no phase flip, and comes back and exits with no more phase flips. So when it exits, it is back to $z=0$ after having traversed a distance $\Delta z = 2d$. Thus it is:

$$E_{T,(B\to A)} = E_{I,(A)}T^2R\cos\left(k\Delta z-\omega t+\right) = E_{I,(A)}T^2R\cos\left(2\pi\frac{2d}{\lambda}-\omega t\right) \tag{10.6.50}$$

The total wave at $t=0$ and $z=0$ is therefore:

$$E_{\text{tot}} = E_{I,(A\to A)} + E_{R1,(A\to A)} + E_{T,(B\to A)} = E_{I,(A)}RT^2\cos\left(\frac{4\pi d}{\lambda}\right) \tag{10.6.51}$$

where λ is the wavelength in the second medium (inside the film).

In the limit that $d \ll \lambda$ the two reflections (taken at $t=0$ and $z=0$):

$$E_{R1,(A\to A)} = -RE_{I,(A)} \qquad \text{and} \qquad E_{T,(B\to A)} = E_{I,(A)}T^2R \tag{10.6.52}$$

are exactly out of phase. For $T\approx 1$, there will be complete destructive interference and no reflection. But there will also be complete destructive interference whenever $\cos\left(\frac{4\pi d}{\lambda}\right)=1$, which is when:

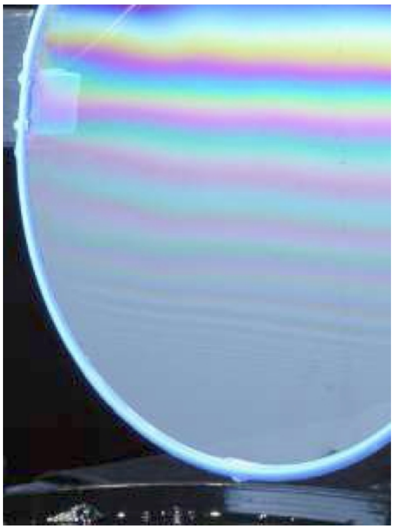
$$d = \frac{\lambda}{2},\frac{2\lambda}{2},\frac{3\lambda}{2},\frac{4\lambda}{2}\cdots \qquad \text{(complete destructive interference)} \tag{10.6.53}$$

On the other hand if the two waves are completely in phase there will be constructive interference. This happens when $\cos\left(\frac{4\pi d}{\lambda}\right)=-1$ which is when:

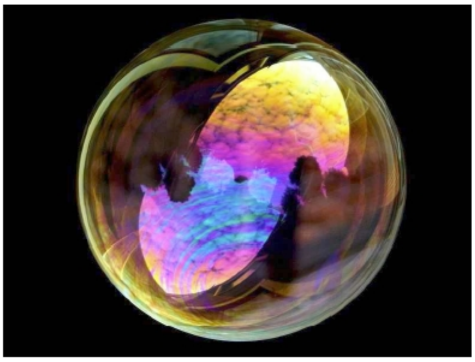
$$d = \frac{\lambda}{4},\frac{3\lambda}{4},\frac{5\lambda}{4},\cdots \qquad \text{(complete constructive interference)} \tag{10.6.54}$$

If the material is much thicker than the wavelength of light, and not of completely uniform thickness, then there will be some constructive and some destructive interference and we won't see much interesting. However, if the material has a well-defined thickness which is of the same order of magnitude as the wavelength of visible light, we will see different wavelengths with different intensities. This happens in a soap film.

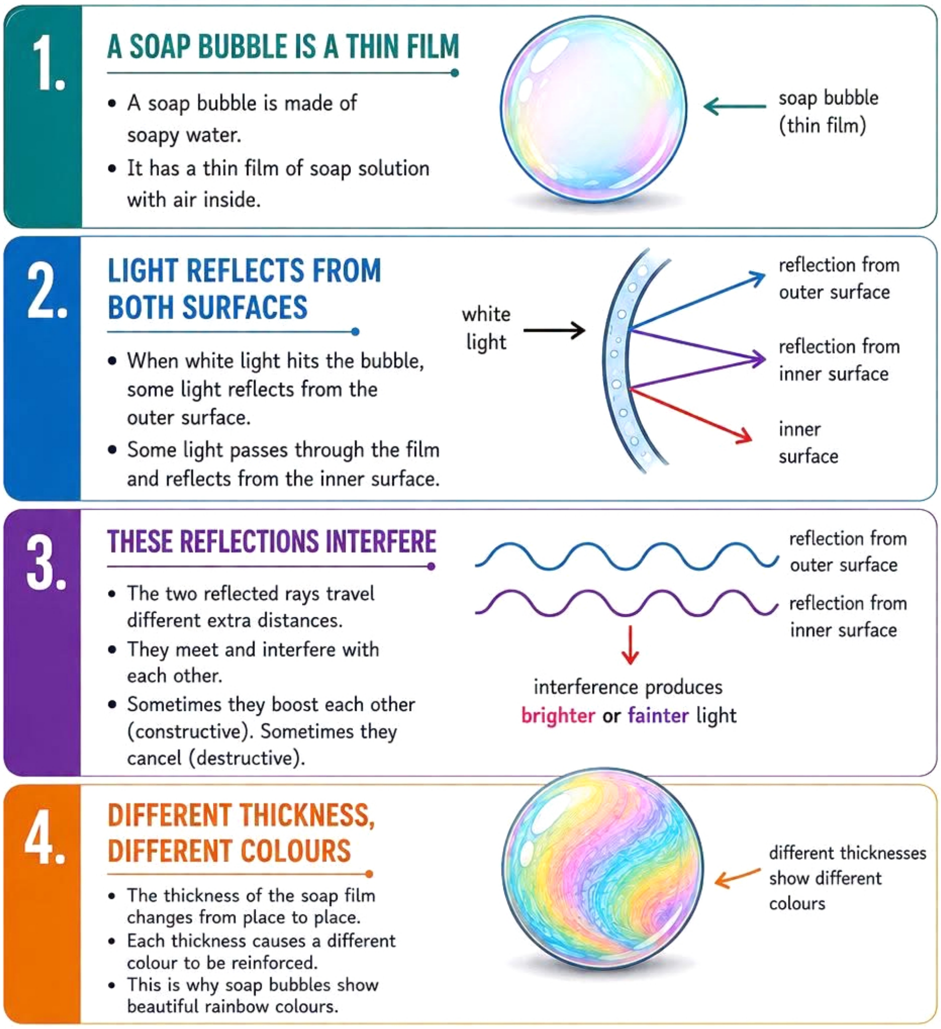
If we put a soap film vertically, then gravity will make it denser at the bottom. The result is the following:



Note at the very top, the film is black because there is complete destructive interference for all wavelengths. Similar patterns can be seen in soap bubbles:



Explanation:



1. A SOAP BUBBLE IS A THIN FILM

- A soap bubble is made of soapy water.
- It has a thin film of soap solution with air inside.

soap bubble
(thin film)

2. LIGHT REFLECTS FROM BOTH SURFACES

- When white light hits the bubble, some light reflects from the outer surface.
- Some light passes through the film and reflects from the inner surface.

white
light
reflection from
outer surface
reflection from
inner surface
inner
surface

3. THESE REFLECTIONS INTERFERE

- The two reflected rays travel different extra distances.
- They meet and interfere with each other.
- Sometimes they boost each other (constructive). Sometimes they cancel (destructive).

reflection from
outer surface
reflection from
inner surface
interference produces
brighter or fainter light

4. DIFFERENT THICKNESS, DIFFERENT COLOURS

- The thickness of the soap film changes from place to place.
- Each thickness causes a different colour to be reinforced.
- This is why soap bubbles show beautiful rainbow colours.

different thicknesses
show different
colours

Colour due to thin-film interference is known as "iridescence". The color of many butterflies and the gorgets of hummingbirds is due to iridescence:

10.6.1.3 Fresnel coefficients

Now let's consider the more general case. Suppose we have a plane wave moving in the $\vec{k}$ direction towards a surface with normal vector $\vec{n}$. Thus the angle θ_1 that the wave is coming in at satisfies:

$$\vec{n}\circ\vec{k}=\|\vec{k}\|\|\vec{n}\|\cos(\theta_1) \tag{10.6.55}$$

The two vectors $\vec{k}$ and $\vec{n}$ form a plane.

There are two linearly independent possibilities for the polarization: the electric field $\vec{E}$ can be polarized in the $\vec{k}-\vec{n}$ plane of the boundary, or transverse to that plane.

We know that these are named "vertical polarizations" and "horizontal polarization" respectively. Snell's law then tells us the directions of the transmitted and reflected electric fields. The magnetic field is always determined by the electric field through (see page 4145):

$$\omega\vec{B}=\vec{k}\times\vec{E} \tag{10.6.56}$$

What we need to solve for is the amplitude of the transmitted and reflected fields! The two cases look like (crosses indicate vectors point into the page, and dots that the vectors come out):

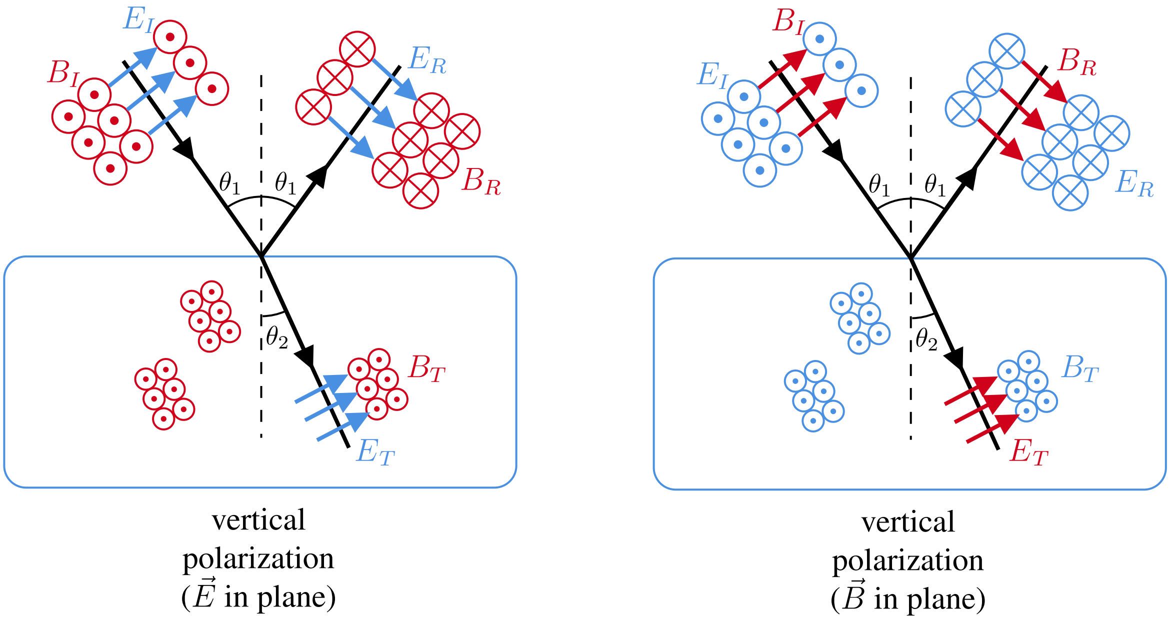


Figure 10.223: Two linearly independent linear polarizations are vertical and horizontal

Let's start with vertical polarization. From the left figure above we see that $\vec{B}$ is always parallel to the surface, so $B_\perp=0$. The electric field has for components project on the surface of the interface between the two medium (i.e. the parallel component):

$$E_\parallel^I=E^I\cos(\theta_1)\qquad E_\parallel^R=E^R\cos(\theta_1)\qquad E_\parallel^T=E_T\cos(\theta_2) \tag{10.6.57}$$

as well as:

$$E_\perp^I=E^I\sin(\theta_1)\qquad E_\perp^R=-E^R\sin(\theta_1)\qquad E_\perp^T=E_T\sin(\theta_2) \tag{10.6.58}$$

The reader can check that when $\theta_1=\theta_2=0$, this reduces to the normal incidence case see earlier above, so we have the sines and cosines right.

Note the relative sign between $E_\perp^I$ and $E_\perp^R$. You can check this sign because when $\theta=\pi/2$, then E_I points up and E_R points down, i.e. they have opposite signs. There is similar sign flip between B_I and B_R, as can be seen in the figure, since $\vec{B}$ always lags behind $\vec{E}$ by $\pi/2$.

Now we simply plug these relations into our boundary conditions and solve as seen earlier above (see page 4477):

$$\varepsilon_1 E_\perp^{(1)}=\varepsilon_2 E_\perp^{(2)} \qquad E_{||}^{(1)}=E_{||}^{(2)} \tag{10.6.59}$$

imply:

$$\begin{gathered}\varepsilon_1 E_I\sin(\theta_1)-\varepsilon_1 E_R\sin(\theta_1)=\varepsilon_2 E_T\sin(\theta_2)\\ E_I\cos(\theta_1)+E_R\cos(\theta_1)=E_T\cos(\theta_2)\end{gathered} \tag{10.6.60}$$

Hence:

$$\begin{gathered}\varepsilon_1\sin(\theta_1)(E_I-E_R)=\varepsilon_2 E_T\sin(\theta_2)\\ \cos(\theta_1)(E_I+E_R)=E_T\cos(\theta_2)\end{gathered} \tag{10.6.61}$$

The relation:

$$B_\perp^{(1)}=B_\perp^{(2)} \tag{10.6.62}$$

is trivially satisfied since $B_\perp=0$. Finally, the relation:

$$\frac{1}{\mu_1}B_{||}^{(1)}=\frac{1}{\mu_2}B_{||}^{(2)} \tag{10.6.63}$$

implies:

$$\frac{1}{\mu_1}(B_I-B_R)=\frac{1}{\mu_2}B_T \tag{10.6.64}$$

Using $B=E/v$ this last relation becomes:

$$\frac{1}{v_1\mu_1}(E_I-E_R)=\frac{1}{v_2\mu_2}E_T \tag{10.6.65}$$

Injecting it in:

$$\varepsilon_1\sin(\theta_1)(E_I-E_R)=\varepsilon_2 E_T\sin(\theta_2) \tag{10.6.66}$$

we get:

$$\varepsilon_1 v_1\mu_1\sin(\theta_1)=\varepsilon_2 v_2\mu_2\sin(\theta_2) \tag{10.6.67}$$

Using:

$$v=\frac{1}{\sqrt{\mu\varepsilon}} \qquad \text{and} \qquad n=\frac{c}{v}=\sqrt{\frac{\mu\varepsilon}{\mu_0\varepsilon_0}} \tag{10.6.68}$$

We have for example:

$$\varepsilon_1 v_1 \mu_1 = \frac{\varepsilon_1 \mu_1}{\sqrt{\varepsilon_1 \mu_1}} = \frac{\sqrt{\varepsilon_1 \mu_1}}{\sqrt{\varepsilon_1 \mu_1}} \frac{\varepsilon_1 \mu_1}{\sqrt{\varepsilon_1 \mu_1}} = \sqrt{\varepsilon_1 \mu_1} \tag{10.6.69}$$

Hence:

$$\sqrt{\varepsilon_1 \mu_1} \sin(\theta_1) = \sqrt{\varepsilon_2 \mu_2} \sin(\theta_2) \tag{10.6.70}$$

we then get by dividing both sides by $\sqrt{\varepsilon_0 \mu_0}$:

$$n_1 \sin(\theta_1) = n_2 \sin(\theta_2) \tag{10.6.71}$$

which is Snell's law. It is reassuring that our derivation here reproduces Snell's law...

To simplify the solution it is helpful to define one parameter which depends on the angles but not on the materials:

$$\alpha \equiv \frac{\cos(\theta_2)}{\cos(\theta_1)} \tag{10.6.72}$$

and another which depends only on the materials:

$$\beta = \frac{Z_1}{Z_2} = \frac{n_2 \mu_1}{n_1 \mu_2} = \frac{n_1 \varepsilon_2}{n_2 \varepsilon_1} = \frac{\varepsilon_2 v_2}{\varepsilon_1 v_1} = \sqrt{\frac{\varepsilon_2 \mu_1}{\varepsilon_1 \mu_2}} \tag{10.6.73}$$

Then the equations:

$$\begin{gathered}\frac{1}{v_1 \mu_1}(E_I - E_R) = \frac{1}{v_2 \mu_2} E_T \\ \cos(\theta_1)(E_I + E_R) = E_T \cos(\theta_2)\end{gathered} \tag{10.6.74}$$

reduce to:

$$\begin{gathered}E_I - E_R = \beta E_T \\ E_I + E_R = \alpha E_T\end{gathered} \tag{10.6.75}$$

with solutions (keep in mind that's for the vertical polarization!):

$$E_R^{\text{vert}} = \frac{\alpha - \beta}{\alpha + \beta} E_I^{\text{vert}}, \quad E_T^{\text{vert}} = \frac{2}{\alpha + \beta} E_I^{\text{vert}}, \quad n_1 \sin(\theta_1) = n_2 \sin(\theta_2) \tag{10.6.76}$$

The solution for horizontal polarization would lead to a similar result:

$$E_R^{\text{horiz}} = \frac{\alpha\beta - 1}{\alpha\beta + 1} E_I, \quad E_T^{\text{horiz}} = \frac{2}{\alpha\beta + 1} E_I, \quad n_1 \sin(\theta_1) = n_2 \sin(\theta_2) \tag{10.6.77}$$

These are known as "Fresnel coefficients". The Fresnel coefficients tell us how much of each polarization is reflected and transmitted. Since any polarization vector can be written as a linear combination of vertical and horizontal polarizations, we can use the Fresnel coefficients to understand how any polarizations reflect.

To interpret the Fresnel coefficients, we would like to know not just the amplitude of the wave transmitted, but the intensity of the light that passes through, or equivalently the power. Recall that the power in a plane wave in the vacuum is (see section Electrodynamics page 4156):

$$P_{\text{tot}} = c\varepsilon_0 \|\vec{E}\|^2 \tag{10.6.78}$$

For a plane wave in a medium, ε changes and only the component of the velocity moving into the medium is relevant, so this becomes (if we don't multiply the $\cos(\theta)$ we wouldn't have conservation of energy as we will check it further below!):

$$P=v\cos(\theta)\varepsilon\|\vec{E}\|^2=\cos(\theta)\sqrt{\frac{\varepsilon}{\mu}}\|\vec{E}\|^2=\frac{\cos(\theta)}{Z}\|\vec{E}\|^2=\frac{\cos(\theta)}{Z}E^2 \tag{10.6.79}$$

Thus the fraction of power reflected for vertical and horizontal polarizations is by using the Fresnel coefficients relations:

$$\frac{P_R^{\text{vert}}}{P_I^{\text{vert}}}=\left(\frac{\alpha-\beta}{\alpha+\beta}\right)^2 \qquad \frac{P_R^{\text{horiz}}}{P_I^{\text{horiz}}}=\left(\frac{\alpha\beta-1}{\alpha\beta+1}\right)^2 \tag{10.6.80}$$

For the fraction of power transmitted, and:

$$\frac{P_T^{\text{vert}}}{P_I^{\text{vert}}}=\frac{\cos(\theta_2)}{\cos(\theta_1)}\frac{Z_1}{Z_2}\left(\frac{2}{\alpha+\beta}\right)^2=\alpha\beta\left(\frac{2}{\alpha+\beta}\right)^2 \tag{10.6.81}$$

and:

$$\frac{P_T^{\text{horiz}}}{P_I^{\text{horiz}}}=\frac{\cos(\theta_2)}{\cos(\theta_1)}\frac{Z_1}{Z_2}\left(\frac{2}{\alpha\beta+1}\right)^2=\alpha\beta\left(\frac{2}{\alpha\beta+1}\right)^2 \tag{10.6.82}$$

Remark

If $\theta_1=\theta_2=0$ we have $\alpha=1$ and if $\mu_1=\mu_2$ (no polarization), then:

$$\frac{P_R^{\text{vert}}}{P_I^{\text{vert}}}=\left(\frac{\alpha-\beta}{\alpha+\beta}\right)^2=\left(\frac{n_1-n_2}{n_1+n_2}\right)^2=R \tag{10.6.83}$$

and is named the "reflectance" or "reflectivity" or "power reflection coefficient".

For an air-water interface R is equal to approximately 4%. By depositing a thin metallic layer on the interface, we can bring back the reflective power very close to 100%: we then speak of "mirror". It is also possible to deposit a thin layer of MgF_2 to reduce the reflective power; we then speak of "anti-reflective layer".

One can check that for either polarization (just put the above relation inside this identity) energy conservation is respected:

$$P_T+P_R=P_I\Leftrightarrow\frac{P_T}{P_I}+\frac{P_R}{P_I}=1 \tag{10.6.84}$$

If the reader remember, for the normal case we get earlier for the reflected power:

$$P_R=\frac{E_R^2}{Z_1}A=\left(\frac{Z_2-Z_1}{Z_2+Z_1}\right)^2\frac{E_I^2}{Z_1}A=\left(\frac{Z_1-Z_2}{Z_1+Z_2}\right)^2P_I \tag{10.6.85}$$

and the for transmitted power :

$$P_T=\frac{E_T^2}{Z_2}A=\left(\frac{2Z_2}{Z_2+Z_1}\right)^2\frac{E_I^2}{Z_2}A=\frac{4Z_1Z_2}{(Z_1+Z_2)^2}\frac{E_I^2}{Z_1}A=\frac{4Z_1Z_2}{(Z_1+Z_2)^2}P_I \tag{10.6.86}$$

And obviously we get the equality with the previous expression only if the conditions of the normal incidence are met. That is:

$$\alpha = 1, \beta = \frac{Z_1}{Z_2} \qquad (10.6.87)$$

And that makes completely sense! This is another evidence that our previous results are not so wrong...

For example, the air-glass interface has $\beta = 1.5$. Then the transmitted and reflected power in vertical and horizontal polarizations are:

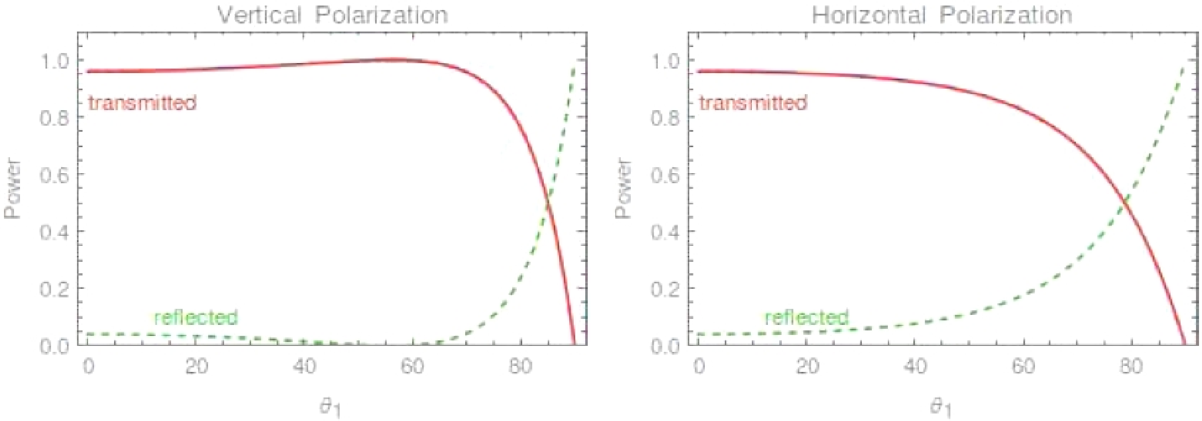


Figure 10.224: **Transmitted and reflected power as a function of incident angle for the two polarizations**

We can also plot the ratio of power reflected in vertical to horizontal polarizations:

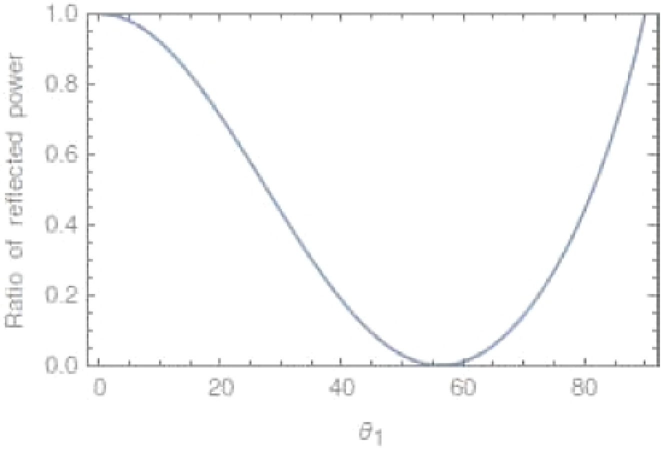


Figure 10.225: **Ratio $P_R^{\text{vert}}/P_R^{\text{horiz}}$ of power reflected vertically polarized horizontally**

From these plots we see there is an angle where the vertical polarization exactly vanishes. This angle is named the "Brewster's angle", θ_B. We can see from the plots that for the glass-air interface $\theta_B \approx 56°$. At this angle, the reflected light is completely (uniquely) polarized.

We can also see that Brewster's angle is an angle of incidence at which light with a particular polarization is perfectly transmitted through a transparent dielectric surface, with no reflection.

What is the general formula for θ_B? From the fraction of power reflected for vertical and horizontal polarizations is:

$$\frac{P_R^{\text{vert}}}{P_I^{\text{vert}}} = \left(\frac{\alpha - \beta}{\alpha + \beta}\right)^2 \qquad \frac{P_R^{\text{horiz}}}{P_I^{\text{horiz}}} = \left(\frac{\alpha\beta - 1}{\alpha\beta + 1}\right)^2 \tag{10.6.88}$$

we see that $P_R^{\text{vert}} = 0$ when $\alpha = \beta$. That is:

$$\frac{\cos(\theta_2)}{\cos(\theta_B)} = \sqrt{\frac{\varepsilon_2\mu_1}{\varepsilon_1\mu_2}} \tag{10.6.89}$$

For most materials $\mu_1 \sim \mu_2 \approx \mu_0$. Then we can use $n = \sqrt{\mu\varepsilon}$ to get:

$$\frac{\cos(\theta_2)}{\cos(\theta_B)} = \frac{n_2}{n_1} \tag{10.6.90}$$

Using also:

$$n_1 \sin(\theta_B) = n_2 \sin(\theta_2) \tag{10.6.91}$$

we can solve for θ_B giving:

$$\tan(\theta_B) = \frac{n_2}{n_1} \tag{10.6.92}$$

Hence the Brewster's angle:

$$\boxed{\theta_B = \tan^{-1}\left(\frac{n_2}{n_1}\right)} \tag{10.6.93}$$

For the air-water interface, $\theta_B = \tan^{-1}(1.5) = 56.3°$. One can see from the previous figure that one does not have to be exactly at this angle to have little reflected vertical polarization. Angles close to θ_B work almost as well.

What is going on physically at Brewster's angle? We know that Snell's law:

$$n_1 \sin(\theta_1) = n_2 \sin(\theta_2) \tag{10.6.94}$$

is always satisfied. At Brewster's angle, when $\theta_1 = \theta_B$, we found also $n_1 \cos(\theta_2) = n_2 \cos(\theta_1)$.

Dividing these two relations gives:

$$\cos(\theta_1)\sin(\theta_1) = \cos(\theta_2)\sin(\theta_2) \tag{10.6.95}$$

or equivalently:

$$\sin(2\theta_1) = \sin(2\theta_2) \qquad \text{when } \theta_1 = \theta_B \tag{10.6.96}$$

Now, by definition θ_1 and θ_2 are both between 0 and $\frac{\pi}{2}$, so $2\theta_1$ and $2\theta_2$ are between 0 and π. Thus:

$$\sin(2\theta_1) = \sin(2\theta_2) \tag{10.6.97}$$

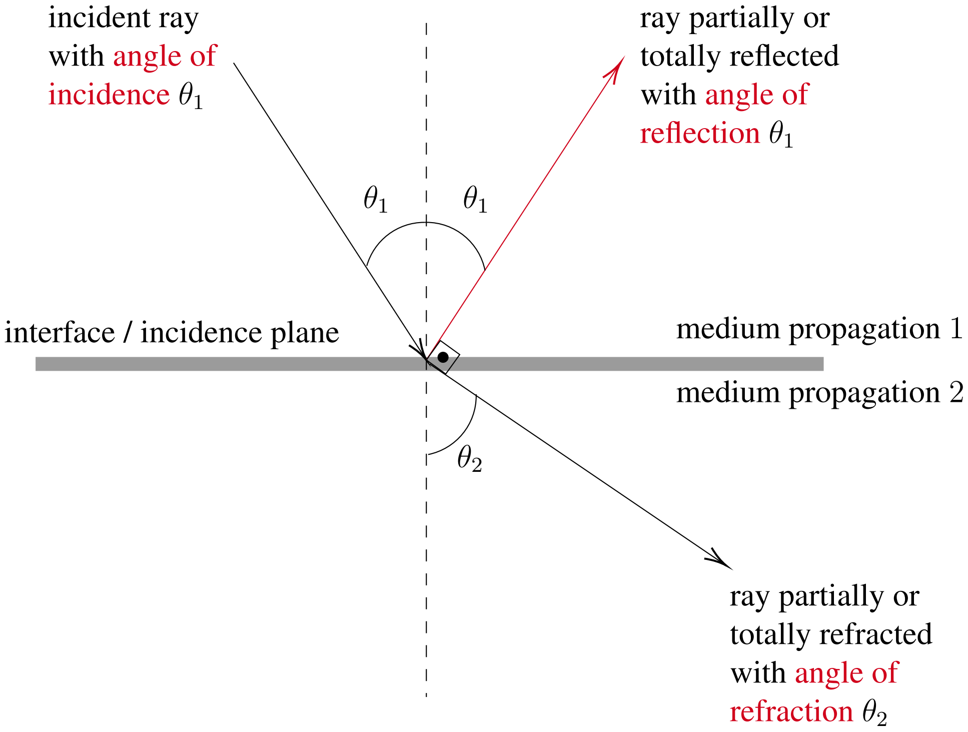
has two solutions. Either $\theta_1 = \theta_2$, which corresponds to $n_1 = n_2$ so the light just passes through, or:

$$\frac{\pi}{2} - 2\theta_1 = 2\theta_2 - \frac{\pi}{2} \qquad (10.6.98)$$

which simplifies to:

$$\theta_1 + \theta_2 = \frac{\pi}{2} \qquad (10.6.99)$$

Using simple geometry we see that the transmitted (i.e. refracted) and reflected waves are perpendicular at Brewster's angle:


incident ray
with angle of
incidence θ_1
ray partially or
totally reflected
with angle of
reflection θ_1
θ_1
θ_1
interface / incidence plane
medium propagation 1
medium propagation 2
θ_2
ray partially or
totally refracted
with angle of
refraction θ_2

Since the reflected wave has to be produced by the motion of particles in the surface we can interpret why there may be no reflection of the vertical polarization component at Brewster's angle: at this angle particles moving in the surface can only produce light polarized transverse to their direction of motion and a Brewster's angle the vertical motion is hard or even impossible.

10.6.2 Fraunhofer Diffraction

From the point of view of geometrical optics, a light beam is a cylinder of section Σ which brings together a large number of parallel rays. It is therefore assumed to be straight when it is set in a homogeneous medium.

The "energy emittance" M $[\mathrm{W}\cdot\mathrm{m}^{-2}]$ of the beam varies only if a lens (or another device) make vary its section Σ or if the medium absorbs energy.

The light beam "bursts" when an obstacle allows only a portion of its section Σ.

Huygens' principle shows that it is the edges of the obstacle that generate this diffraction.

The phenomenon is general but is well observed only if the ratio L/Σ is very large. L being the length of the edges. This condition is necessary so that the intensity of the undiffracted portion of the beam does not mask the effect.

D1. We speak of "Fraunhofer diffraction" when, as assumed above, the incident light rays are parallel and the phenomenon observed at relatively large distance from the screen.

D2. We speak of "Fresnel diffraction" when the incident rays form a divergent beam from a point source or if we observe the phenomenon at close range.

Let us consider a generic case and in fact the most widespread case in the physics laboratories which is the diffraction by a narrow rectangular aperture!

For this, we consider that the incident beam, perpendicular to the aperture, has a plane electromagnetic periodic wavefront and is given by (see section Wave Mechanics page 3479):

$$f(x,t)=\hat{f}\sin(kx-\omega t) \tag{10.6.100}$$

where for recall, its wavelength is given by:

$$\lambda=\frac{2\pi}{k} \tag{10.6.101}$$

10.6.2.1 Case of a small rectangular aperture (slit)

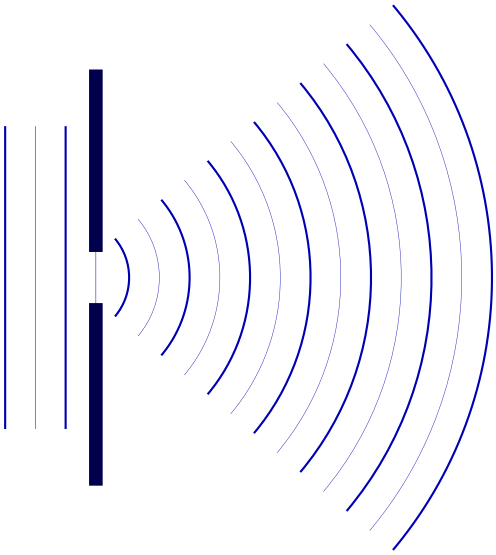
The span width e of the aperture is oriented along the y axis, the height h (parameter that can not be represented in the figure below because it is a view from above) is assumed as very big so we can neglect the border effects.

Following the Huygens principle (see page 4470), the front of the plane wave, delimited by the aperture, is a multitude of sources $\mathrm{d}f(x,t)$, of width $\mathrm{d}y$, which emit in phase, spherical wavelet described by their associated field vector:

$$\mathrm{d}f(x,t)=\frac{\hat{f}}{e}\mathrm{d}y\sin(kx-\omega t) \tag{10.6.102}$$

Now consider an observation point P, at a distance R from the source (assimilated to an aperture). We have proved during our study of the emission sources of spherical type (see section Electrodynamics page 4160) that their amplitude decreases inversely proportional to the distance as:

$$\mathrm{d}f(x,t)=\frac{\hat{f}}{eR}\mathrm{d}y\sin(kx-\omega t) \tag{10.6.103}$$



Remark

Using modern phasor notation the above relation can be found in many textbooks in the following form:

$$\mathrm{d}f(x,t)=\frac{\hat{f}}{eR}e^{\mathrm{i}(kx-\omega t)}\mathrm{d}y$$

or some textbooks replace the source strength per unit length $\hat{f}/e$ (assumed uniform!) by a new term like:

$$\mathrm{d}f(x,t)=\frac{\hat{\rho}_L}{R}e^{\mathrm{i}(kx-\omega t)}\mathrm{d}y$$

However, the wavelets, each depending on the point of the aperture to which it is assimilated, will not all travel the same distance R but a distance proper distance r. However, if R is sufficiently distant from the aperture, we will allow ourselves to approximate:

$$\frac{\hat{f}}{eR}\cong\frac{\hat{f}}{er} \tag{10.6.104}$$

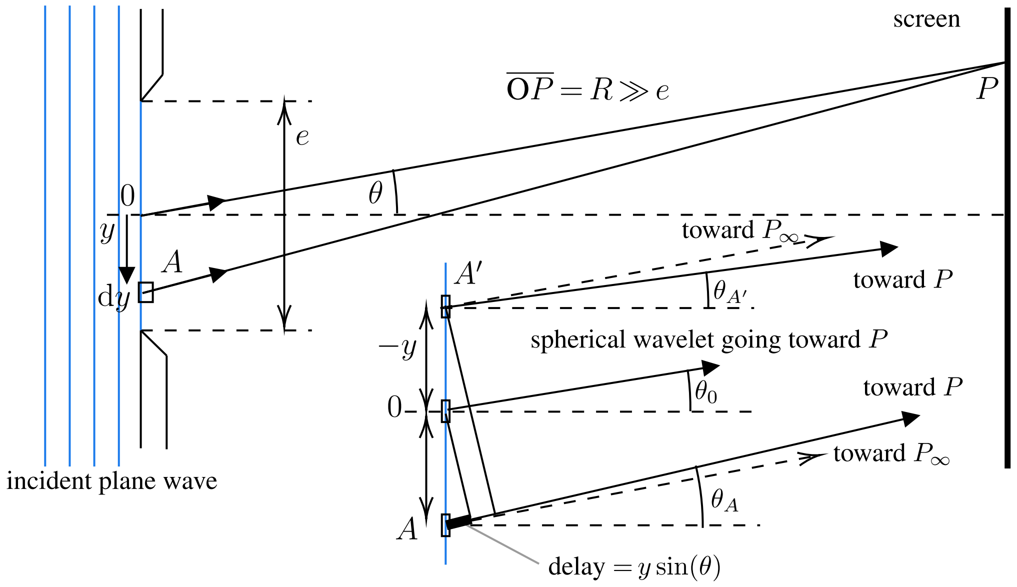
remains still the periodic term $\sin(kx-\omega t)$ where we put $x=r$. But we have for extremal values:

$$R-\frac{e}{2}\sin(\theta)\leq r\leq R+\frac{e}{2}\sin(\theta) \tag{10.6.105}$$

These extremal values correspond respectively, to the advance and the delay of the wave function describing the propagation of the wavelet at the borders of the aperture.

Indeed, it is enough to see the figure below, considering $R \gg e$ and therefore:

$$\theta := \theta_{A'} \cong \theta_0 \cong \theta_A \tag{10.6.106}$$


screen
$\overline{\mathrm{O}P} = R \gg e$
P
e
θ
0
y
A
dy
toward P_∞
A'
toward P
$\theta_{A'}$
spherical wavelet going toward P
$-y$
θ_0
toward P
0
toward P_∞
θ_A
A
delay $= y\sin(\theta)$
incident plane wave

Thus, by placing the origin of the y coordinate in the middle of the aperture, we have:

$$\mathrm{d}f(x,t) = \frac{\hat{f}}{eR}\mathrm{d}y\sin(kr-\omega t) = \frac{\hat{f}}{eR}\mathrm{d}y\sin(k(R+y\sin(\theta))-\omega t) \tag{10.6.107}$$

So the different wavelets are out of phase and produce interferences.

Definition (#10.6.497): In wave mechanics, we speak of "interference" to describe a phenomenon in which two waves superpose to form a resultant wave of greater, lower, or the same amplitude. Interference usually refers to the interaction of waves that are correlated or coherent with each other, either because they come from the same source or because they have the same or nearly the same frequency. Interference effects can be observed with all types of waves, for example, light, radio, acoustic, surface water waves or matter waves.

The diffracted wave in the direction of θ, is given by the sum of all contributions:

$$f(x,t) = \frac{\hat{f}}{eR}\int\limits_{-e/2}^{+e/2}\sin(k(R+y\sin(\theta))-\omega t)\mathrm{d}y = \frac{\hat{f}}{eR}\int\limits_{-e/2}^{+e/2}\sin((kR-\omega t)+ky\sin(\theta))\mathrm{d}y \tag{10.6.108}$$

Knowing that (trigonometric relation are explicited on request of a reader):

$$\begin{aligned}\int\limits_{-e/2}^{+e/2}\sin(a+by)\mathrm{d}y&=\int\limits_{-e/2}^{+e/2}(\cos(a)\sin(by)+\sin(a)\cos(by))\mathrm{d}y\\&=\cos(a)\int\limits_{-e/2}^{+e/2}\sin(by)\mathrm{d}y+\sin(a)\int\limits_{-e/2}^{+e/2}\cos(by)\mathrm{d}y\\&=-\frac{\cos(a)\cos(by)}{b}+\frac{\sin(a)\sin(by)}{b}\Bigg|_{-e/2}^{+e/2}\\&=\frac{2\sin(a)\sin\left(b\dfrac{e}{2}\right)}{b}\end{aligned}\tag{10.6.109}$$

We have therefore:

$$\begin{aligned}f(x,t)&=\frac{\hat{f}}{eR}\int\limits_{-e/2}^{e/2}\sin((kR-\omega t)+ky\sin(\theta))\mathrm{d}y=\frac{\hat{f}}{eR}\frac{2\sin(kR-\omega t)\sin\left(k\sin(\theta)\dfrac{e}{2}\right)}{k\sin(\theta)}\\&=\frac{\hat{f}}{R}\frac{\sin\left(k\sin(\theta)\dfrac{e}{2}\right)}{\dfrac{ek\sin(\theta)}{2}}\sin(kR-\omega t)=\frac{\hat{f}}{R}\frac{\sin\left(\dfrac{\pi e}{\lambda}\sin(\theta)\right)}{\dfrac{\pi e}{\lambda}\sin(\theta)}\sin(kR-\omega t)\end{aligned}\tag{10.6.110}$$

We have proved in the section Electrodynamics that the energy (i.e. the intensity) of an electromagnetic wave was given (in vacuum) by the average scalar value of the Poynting vector:

$$\bar{S}=\frac{\hat{E}\hat{B}}{2\mu_0}\tag{10.6.111}$$

We therefore have by considering that the magnetic and electric field are proportional to the term:

$$\frac{\sin\left(\dfrac{\pi e}{\lambda}\sin(\theta)\right)}{\dfrac{\pi e}{\lambda}\sin(\theta)}\tag{10.6.112}$$

the following result:

$$\bar{S}\propto\overline{S_0}\left(\frac{\sin\left(\dfrac{\pi e}{\lambda}\sin(\theta)\right)}{\dfrac{\pi e}{\lambda}\sin(\theta)}\right)^2=\overline{S_0}\left(\frac{\sin\left(\pi\Gamma\sin(\theta)\right)}{\pi\Gamma\sin(\theta)}\right)^2\tag{10.6.113}$$

which is the luminous emittance emitted in the direction θ and where have put:

$$\Gamma=\frac{e}{\lambda}\tag{10.6.114}$$

If we introduce the sinus cardinal sinc that we have meet in our study of Fourier transforms in the section of Sequences and Series and that was defined in the section Trigonometry then we can write the previous relation in a more clearly condensed way:

$$\bar{S}\propto\overline{S_0}\mathrm{sinc}^2(\pi\Gamma\sin(\theta))\tag{10.6.115}$$

For which can have a generic form plotted with Maple™ 4.00b using:

```
>Gamma:=3;plot((sin(Gamma*x)/(Gamma*x))^2, x=-Pi..Pi);
```

So we can get the same result by taking the squared modulus of the Fourier transform of a monochromatic signal through a rectangular window. Thus, it seems possible to study the phenomena of diffraction by using the Fourier transform, and this area is named "Fourier optics" that we will study later.

Here is a graphical representation of the relation $\bar{S}/\overline{S_0}$ for different values of the ratio $\Gamma = e/\lambda$:

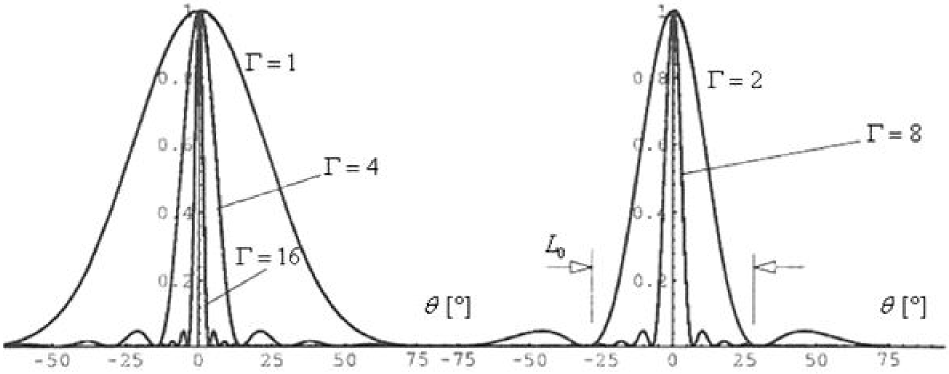
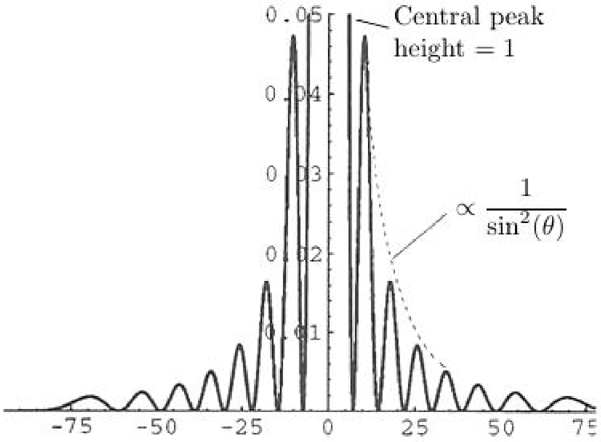


Figure 10.226: **Representation of diffraction fringes**

On either side of the central fringe, there are other fringes, more narrow and arranged symmetrically. Their intensity decreases very rapidly as the preponderant term at the denominator:

$$\frac{\overline{S_{\max}}}{\overline{S_0}} \propto \frac{1}{(\pi\Gamma\sin(\theta))^2} = \frac{\lambda^2}{\pi^2 e^2}\frac{1}{\sin^2(\theta)} \qquad (10.6.116)$$



from which we have a real image here:

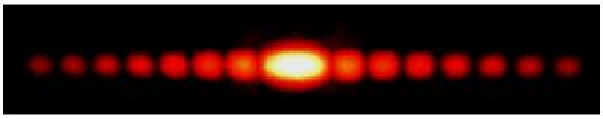


Figure 10.227: **Photo of a real Fraunhofer diffraction fringe of a rectangular aperture**

Between the fringes, there are dark areas that are the seat of destructive interference. Their position is given by the condition:

$$\sin\left(\frac{\pi e}{\lambda}\sin(\theta)\right)=0\Rightarrow\frac{\pi e}{\lambda}\sin(\theta)=n\pi \tag{10.6.117}$$

excepted for $n=0\Rightarrow\theta=0$ where we observe a maximum!

So we see black bands in the directions:

$$\theta_n=\arcsin\left(\frac{n\lambda}{e}\right)\qquad\forall n\in\mathbb{N}^* \tag{10.6.118}$$

Thus the angular width of the central band is twice the angular value obtained for the first minimum:

$$L_0=2R\arcsin\left(\frac{\lambda}{e}\right) \tag{10.6.119}$$

We get the width of the following peaks as follows:

Two successive minima satisfy the conditions:

$$\sin(\theta_n)=\frac{n\lambda}{e}\qquad\sin(\theta_{n+1})=\frac{(n+1)\lambda}{e} \tag{10.6.120}$$

Therefore:

$$\begin{aligned}\sin(\theta_n)-\sin(\theta_{n+1})=\frac{\lambda}{e}&=2\sin\left(\frac{\theta_{n+1}-\theta_n}{2}\right)\cos\left(\frac{\theta_{n+1}-\theta_n}{2}\right)\\&=2\sin\left(\frac{\Delta\theta}{2}\right)\cos\left(\frac{\theta_{n+1}-\theta_n}{2}\right)\end{aligned} \tag{10.6.121}$$

Put putting:

$$\frac{\theta_{n+1}+\theta_n}{2}\cong\theta \tag{10.6.122}$$

Therefore it comes:

$$\Delta\theta=2\arcsin\left(\frac{\lambda}{2e\cos(\theta)}\right) \tag{10.6.123}$$

Since the emittance decreases very rapidly, only the first bands (for which $\cos(\theta)\cong 1$) are observable. Therefore it remains:

$$L_n=2R\arcsin\left(\frac{\lambda}{2e}\right)\cong\frac{L_0}{2} \tag{10.6.124}$$

The positions of the maxima are then given by the condition:

$$\begin{aligned}\frac{\mathrm{d}M}{\mathrm{d}\theta}&=0=\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\frac{\sin\left(\frac{\pi e}{\lambda}\sin(\theta)\right)}{\frac{\pi e}{\lambda}\sin(\theta)}\right)=\frac{\lambda}{\pi e}\frac{\sin\left(\frac{\pi e}{\lambda}\sin(\theta)\right)}{\sin(\theta)}\\&=\frac{\lambda}{\pi e}\frac{1}{\sin^2(\theta)}\left[\sin(\theta)\frac{\mathrm{d}}{\mathrm{d}\theta}\sin\left(\frac{\pi e}{\lambda}\sin(\theta)\right)-\sin\left(\frac{\pi e}{\lambda}\sin(\theta)\right)\frac{\mathrm{d}}{\mathrm{d}\theta}\sin(\theta)\right]\\&=\frac{\lambda}{\pi e}\frac{1}{\sin^2(\theta)}\left[\sin(\theta)\cos\left(\frac{\pi e}{\lambda}\sin(\theta)\right)\frac{\pi e}{\lambda}\cos(\theta)-\sin\left(\frac{\pi e}{\lambda}\sin(\theta)\right)\cos(\theta)\right]\\&=\cos\left(\frac{\pi e}{\lambda}\sin(\theta)\right)\cot(\theta)-\frac{\lambda}{\pi e}\sin\left(\frac{\pi e}{\lambda}\sin(\theta)\right)\frac{\cot(\theta)}{\sin(\theta)}\\&\Rightarrow\cos\left(\frac{\pi e}{\lambda}\sin(\theta)\right)\cot(\theta)=\frac{\lambda}{\pi e}\sin\left(\frac{\pi e}{\lambda}\sin(\theta)\right)\frac{\cot(\theta)}{\sin(\theta)}\Rightarrow\tan\left(\frac{\pi e}{\lambda}\sin(\theta)\right)\\&=\frac{\pi e}{\lambda}\sin(\theta)\end{aligned}\tag{10.6.125}$$

Let us put:

$$\phi=\frac{\pi e}{\lambda}\sin(\theta)\tag{10.6.126}$$

The numerical resolution of:

$$\tan(\phi)=\phi\tag{10.6.127}$$

gives:

$$\phi=\pm\{0,4.49,7.72,10.9,14.1\ldots\}\tag{10.6.128}$$

The positions of the successive maxima are then:

$$\begin{aligned}&\phi=0\Rightarrow\theta=0\\&\phi=4.49\Rightarrow\theta=\arcsin\left(\frac{4.49}{\pi}\frac{\lambda}{e}\right)\cong\arcsin\left(1.5\frac{\lambda}{e}\right)\end{aligned}\tag{10.6.129}$$

etc. We could have also easily get a decent approximation of this result, considering that the intensity is maximum when:

$$\left.\frac{\sin\left(\frac{\pi e}{\lambda}\sin(\theta)\right)}{\frac{\pi e}{\lambda}\sin(\theta)}\right|_{\max}\cong\left.\sin\left(\frac{\pi e}{\lambda}\sin(\theta)\right)\right|_{\max}\Rightarrow\sin\left(\frac{\pi e}{\lambda}\sin(\theta)\right)=1\Rightarrow\frac{\pi e}{\lambda}\sin(\theta)=(2n+1)\frac{\pi}{2}\tag{10.6.130}$$

Which bring us to write:

$$\theta_{\max,n}=\arcsin\left(\frac{(2n+1)\lambda}{2e}\right)\tag{10.6.131}$$

with $n\in\mathbb{N}^*$.

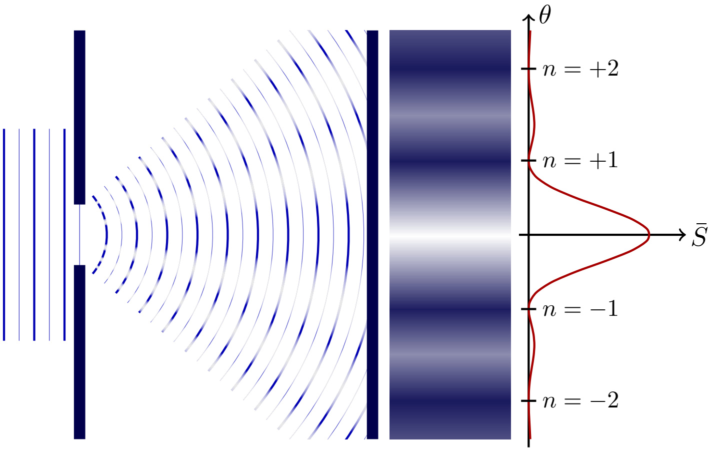


Figure 10.228: **The big picture of the diffraction single slit projection (author: Izaak Neutelings)**

A remarkable result of the Fraunhofer experience is that it challenges the corpuscular view of light as we had in the 119th century (holocene calendar).

Indeed, many experiments such at the projection of the shadow of an object on a wall seemed to show that the light was such a particle that does not pass through materials and being net stopped by any obstacle either in its center or by its edges (you must draw your attention to the "edges" in particular).

But, Fraunhofer experience and in particular that of Fresnel regarding the edges (we will see further because it is mathematically more difficult to study), show that the light appears to behave not as a single particle but as a wave (from the Huygens principle that we used for our developments) as we showed to us the earlier developments that perfectly explain the experimental results of Fraunhofer diffraction.

But then why keep the corpuscular model of light? Simply because of other experimental and theoretical results of which the best known are the photoelectric effect or Compton scattering (see section Nuclear Physics page 5096), which can be explained theoretically quite well if it is not perfectly with a corpuscular model light (and also with some other particle of different size, charge, spin, etc.).

By the early 11800s (holocene calendar), there were two theories about the nature of light. One of them, going back to Newton, is that light is a ray. But there was another idea – going back even farther (to Christiaan Huygens) – that light might also be a wave. And this gained a lot of support in 11799 (holocene calendar), when Thomas Young first passed light through two thin, nearby slits. So if you're a good theorist, and you're interested in studying light, what do you do?

Well, if you're famed French mathematician and physicist Simeon Poisson, you would think of the most ridiculous configuration you could imagine in the hopes of disproving the light-is-a-wave theory. And that's exactly what he did in 11818 (holocene calendar).

He imagined that you took a wave source of light, and had it shine on and around a completely black, spherical obstacle, setting up a screen behind it. Obviously, he reasoned, you would see some light on the screen indicating the outside of the sphere, and darkness, or a shadow, on the inside.

But, he calculated, if the wave theory of light were correct, you would get something completely absurd!

Sure, you'd get light on the outside, and shadow on the inside, but what's that at the very center? Poisson predicted, using the wave theory of light, that you'd actually get a bright spot of light at the very center of this shadow! How absurd was that! And therefore, he reasoned, the wave theory of light was absolutely crazy, and had to be wrong.

Shortly after Poisson's prediction, Francois Arago decided to put the theory to the test, and actually performed the experiment to look for the "theoretically absurd" spot.

And what happens if, in fact, you perform this experiment yourself? :

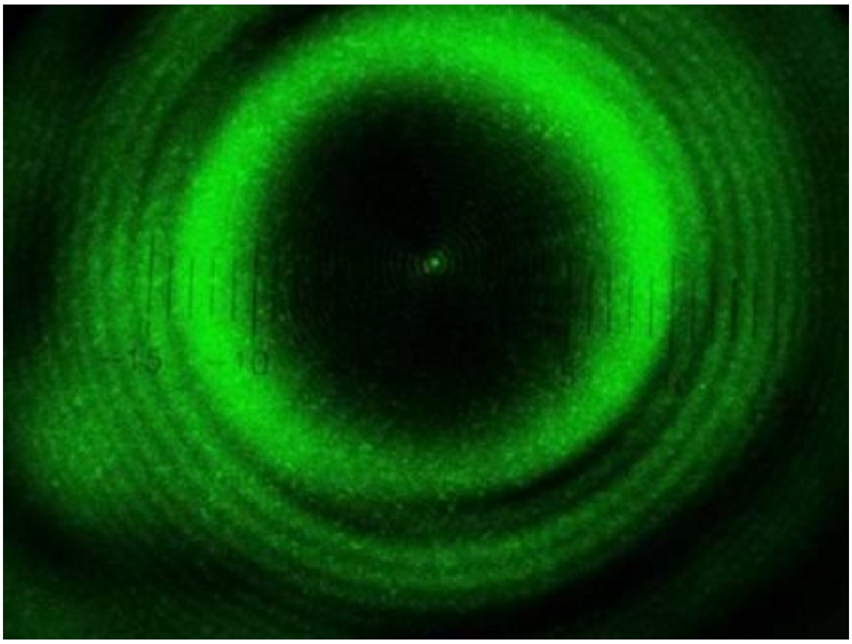


Figure 10.229: Arago spot (source: Thomas Bauer at Wellesley)

Amazingly, the spot is real! If your theory is any good, scientifically, this is exactly what it will do. It will not only explain what's already been observed, it will allow you to apply it to new situations, and make testable predictions about what you can expect to find. The crazier the prediction, and the more successful the experiment, the more compelling the theory becomes.

10.6.2.1.1 Optical resolution in the far field region for small rectangular aperture

According to the criterion of the English physicist Lord Rayleigh: the "optical resolution" (or "power angular resolution") of a circular aperture, is the angle α_{min} between two light rays of wavelength λ, emitted from two point sources S_1,S_2, far away, whose diffraction figures are separated such that the first zero of the diffraction pattern is found at the place of the maximum of the other as illustrated below:

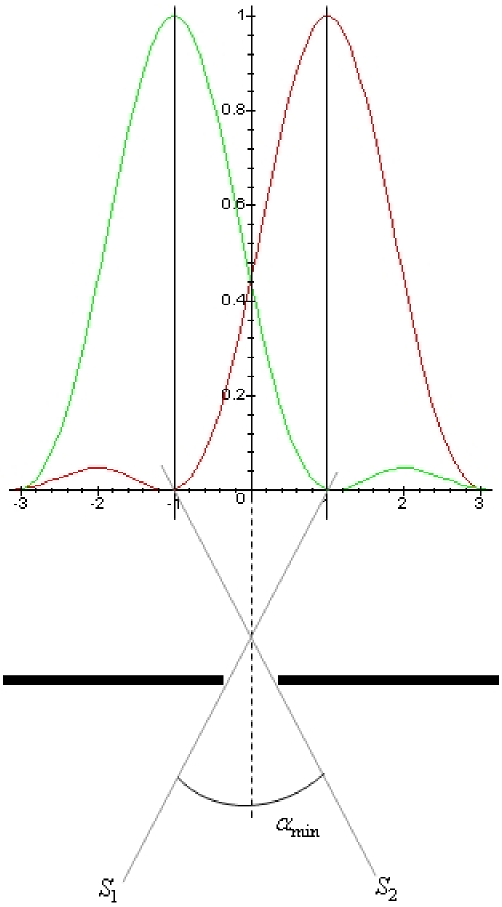
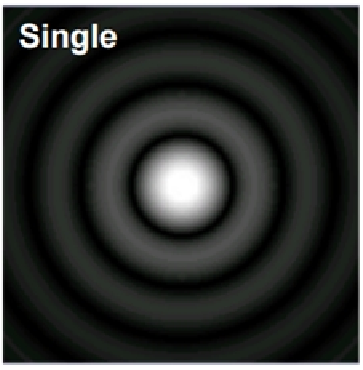
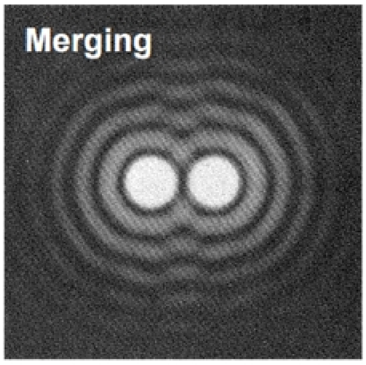
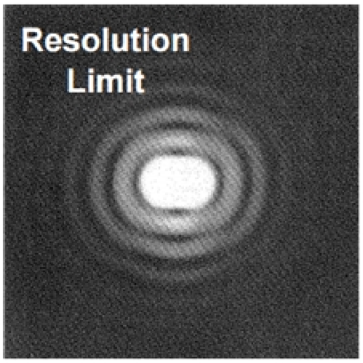

1
0.8
0.6
0.4
0.2
-3
-2
-1
0
2
3
$\alpha_{\min}$
S_1
S_2

Figure 10.230: Optical resolution as defined by Lord Rayleigh

Or with a photo of experiment perhaps it's better:






Resolution
Limit

This concept is used extensively in photography, astronomy, radio astronomy, etc. So the reader should pay a special attention to it!

Now we have proved that the minima for a rectangular aperture were given by:

$$\sin(\theta_n) = \frac{n\lambda}{e} \tag{10.6.132}$$

And if we take the case where $n=1$, we fall back on the relation available in many books without proof:

$$\boxed{\theta \cong \sin(\theta_n) = \frac{\lambda}{e} = \alpha_{\text{min}}} \tag{10.6.133}$$

named "Rayleigh criterion of a small rectangular aperture (slit)".

Therefore the angular resolution of a rectangular aperture is proportional to the ratio of the wavelength λ to the thickness e of the slot. Obviously in practice the goal is to have either:

1. A large as possible value of the angle α_{min} so that the objects (both source in this case) does not coincide and that their images are therefore distinct.
2. A small as possible value so that the sole observed object (only one source) do not have a blurry image. This is one reason why many popular books say you have to have a smaller wavelength than the observed object so that it is observable (so not too be blurred!).

To increase the resolving power to decrease the blur effect, so you have to be working with shorter wavelengths or increase the thickness of the slot of the instrument, as the wavelength is often imposed by the observed object, it is natural to make vary e.

Therefore if the light that passes through a slot (rectangular or circular!) forms an image on a screen, and that this image is observed under a microscope, for example, it is impossible, whatever the magnification of the microscope ,to observe more details in the image than it is allowed by the resolving power of the slot. We must take these considerations into account when we design of optical instruments.

10.6.2.2 Case of a network of small rectangular apertures (slits)

Let us now consider an array of N apertures of width $e\cong\lambda$, of height $H\gg e$ and distant each other by d. A single incident beam illuminates all apertures.

Remark

The study of this model will allow us to understand partially how the prism works and also the goniometer in astronomy for spectrum analysis as well as X-ray diffraction by a network of atoms (the importance of the latter being quite significant!).

Given the following figure:

to the screen at $R \to +\infty$

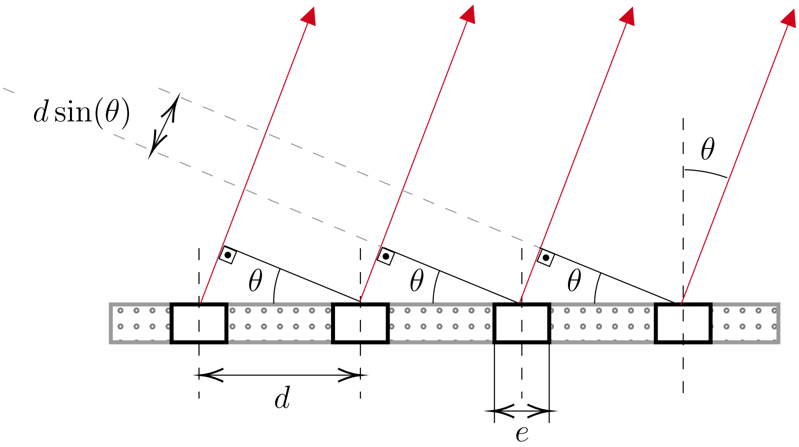


Figure 10.231: **Basic principle of the rectangular aperture network**

We see in the diagram above that for some directions θ, the distance $d\sin(\theta)$ is such that constructive or destructive interference are occurring. In optics, these devices are therefore wave splitters, that is to say that a wave is separated into two or more to make them interfere. We name these kind of configurations "wavefront splitting systems".

Let us consider that this network is placed in the YZ plane and the beam direction is along the X axis. We put ourselves in an observation point P in the XY plane. Following the properties of electromagnetic waves (see section Electrodynamics page 4138), the electric field vector $\vec{E}_i$ of the wave emitted by the i-th slot is perpendicular to the direction of observation and can be expressed as:

$$E_i(\theta,R,t)=\frac{\hat{E}_i(\theta)}{R}\sin(kR-\omega t) \tag{10.6.134}$$

and we also proved earlier above that:

$$f(R,t)=\frac{\hat{f}}{R}\frac{\sin\left(\dfrac{\pi e}{\lambda}\sin(\theta)\right)}{\dfrac{\pi e}{\lambda}\sin(\theta)}\sin(kR-\omega t) \tag{10.6.135}$$

where by analogy between these two relations, we recognize that:

$$\hat{E}_i(\theta)=\hat{f}\frac{\sin\left(\dfrac{\pi e}{\lambda}\sin(\theta)\right)}{\dfrac{\pi e}{\lambda}\sin(\theta)} \tag{10.6.136}$$

and therefore it comes:

$$\begin{gathered}\frac{\hat{E}_i(\theta)}{R}\sin(kR-\omega t)=\frac{\hat{E}_{i,\theta=0}}{R}\frac{\sin\left(\dfrac{\pi e}{\lambda}\sin(\theta)\right)}{\dfrac{\pi e}{\lambda}\sin(\theta)}\sin(kR-\omega t)\\ \Rightarrow\frac{\hat{E}_i(\theta)}{R}=\frac{\hat{E}_{i,\theta=0}}{R}\frac{\sin\left(\dfrac{\pi e}{\lambda}\sin(\theta)\right)}{\dfrac{\pi e}{\lambda}\sin(\theta)}\Rightarrow\hat{E}_i(\theta)=\hat{E}_{i,\theta=0}\frac{\sin\left(\dfrac{\pi e}{\lambda}\sin(\theta)\right)}{\dfrac{\pi e}{\lambda}\sin(\theta)}\end{gathered}\tag{10.6.137}$$

In any direction θ, the waves issued from the two adjacent apertures are out of phase of $d\sin(\theta)$ (where d is the step value between two apertures) and at the point P of observation, the resulting electric field is given by the sum of the contributions of each aperture with its own i-th shift. Therefore:

$$\begin{aligned}E(\theta,R,t)&=\sum_{i=1}^{N}E_i(\theta,R,t)=\sum_{i=1}^{N}\hat{E}_i(\theta)\sin(k(R+(i-1)d\sin(\theta))-\omega t)\\&=\hat{E}\sum_{i=1}^{N}\sin((kR-\omega t)+k(i-1)d\sin(\theta))\end{aligned}\tag{10.6.138}$$

So we see that each wave is out of phase of:

$$\Delta\varphi=kd\sin(\theta)=\frac{2\pi}{\lambda}d\sin(\theta)=c^{te}\tag{10.6.139}$$

We can now represent $E(\theta,R,t)$ using phasors (see section Wave Mechanics page 3510) in the phase space such that:

$$E(\theta,R,t)=\hat{E}\sum_{j=1}^{N}e^{\mathrm{i}[(kR-\omega t)+k(j-1)d\sin(\theta)]}=\hat{E}\sum_{j=1}^{N}e^{\mathrm{i}(kR-\omega t)}e^{\mathrm{i}k(j-1)d\sin(\theta)}\tag{10.6.140}$$

Which gives graphically for the second term containing the summation variable j for a fixed distance R:

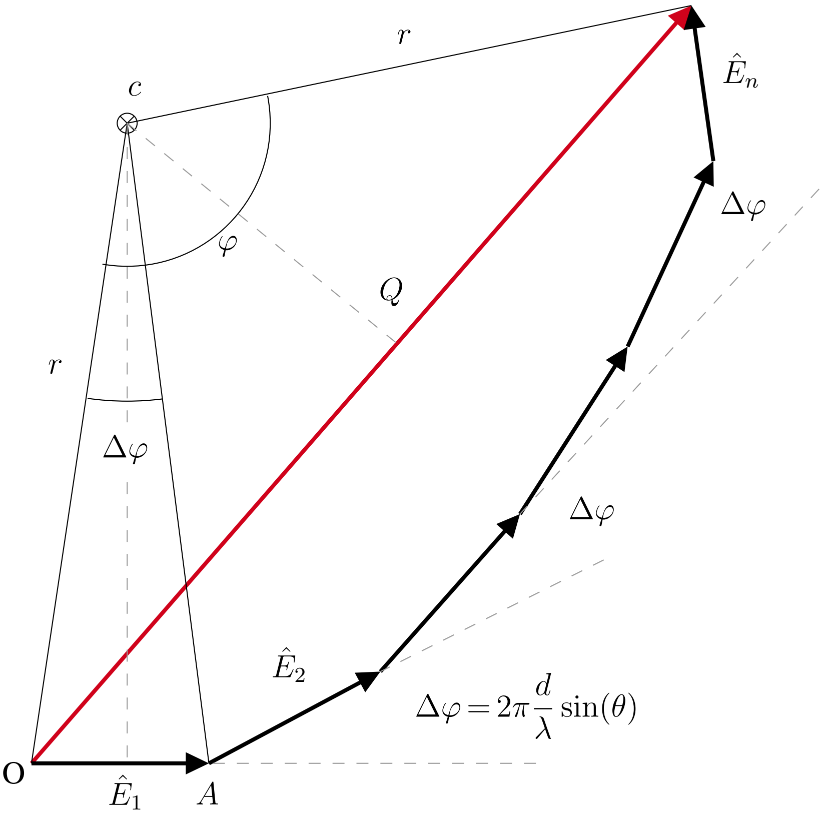


Figure 10.232: **Representation of the summation of terms**

We see that the $\vec{E}_i$ put together form a regular polygon inscribed in a circle of radius:

$$r = \frac{\frac{\hat{E}_i}{2}}{\sin\left(\frac{\Delta\varphi}{2}\right)} \tag{10.6.141}$$

The norm of the resulting electric field being equal to the chord defined by the angle:

$$\varphi = \sum_{i=1}^{N} \Delta\varphi \tag{10.6.142}$$

we will have:

$$\hat{E}_{\text{resultant}} = 2\rho\sin\left(\frac{\varphi}{2}\right) = 2\frac{\hat{E}_i}{2}\frac{\sin\left(\frac{N\Delta\varphi}{2}\right)}{\sin\left(\frac{\Delta\varphi}{2}\right)} = \hat{E}_i\frac{\sin\left(N\frac{\pi}{\lambda}d\sin(\theta)\right)}{\sin\left(\frac{\pi}{\lambda}d\sin(\theta)\right)} \tag{10.6.143}$$

The light energy (i.e. intensity) emitted in the direction θ is proportional to the square of the electric field (see section Electrodynamics page 4156), then we have for destructive or constructive interference:

$$\overline{S}(\theta, N, d) = \overline{S}_i(\theta)\left(\frac{\sin\left(N\frac{\pi}{\lambda}d\sin(\theta)\right)}{\sin\left(\frac{\pi}{\lambda}d\sin(\theta)\right)}\right)^2 \tag{10.6.144}$$

We now substitute $\overline{S}_i(\theta)$ by the expression found in our earlier study of diffraction by a single aperture:

$$\overline{S}\propto\overline{S}_0\left(\frac{\sin\left(\dfrac{\pi e}{\lambda}\sin(\theta)\right)}{\dfrac{\pi e}{\lambda}\sin(\theta)}\right)^2=\overline{S}_0\left(\frac{\sin(\pi\Gamma\sin(\theta))}{\pi\Gamma\sin(\theta)}\right)^2 \tag{10.6.145}$$

Thus we get to the addition of interference and diffraction effects:

$$\boxed{\overline{S}(\theta,N,d)=\overline{S}_i(\theta)\left(\frac{\sin\left(N\dfrac{\pi}{\lambda}d\sin(\theta)\right)}{\sin\left(\dfrac{\pi}{\lambda}d\sin(\theta)\right)}\right)^2=\overline{S}_0\left(\frac{\sin\left(\dfrac{\pi e}{\lambda}\sin(\theta)\right)}{\dfrac{\pi e}{\lambda}\sin(\theta)}\right)^2\left(\frac{\sin\left(N\dfrac{\pi}{\lambda}d\sin(\theta)\right)}{\sin\left(\dfrac{\pi}{\lambda}d\sin(\theta)\right)}\right)^2} \tag{10.6.146}$$

Although this relation seems complicated, its parameters do not have the same practical significance. Indeed, consider the function:

$$\psi=\left(\frac{\sin(Nx)}{\sin(x)}\right)^2=\frac{A}{B} \tag{10.6.147}$$

The term A has maxima when:

$$Nx_{\text{max}}=\frac{\pi}{2}+m\pi\Leftrightarrow=\frac{\pi}{2}N+m\frac{m\pi}{N} \tag{10.6.148}$$

and is equal to zero if:

$$Nx_{\text{null}}=m\pi\Leftrightarrow x_{\text{null}}=m\frac{\pi}{N} \tag{10.6.149}$$

with $m\in\mathbb{N}$.

Although the term B makes the relation diverge for $x=m\pi$, l'Hôpital's rule (see section Differential and Integral Calculus page 1461) provides that:

$$\lim_{x\to m\pi}\left(\frac{\sin(Nx)}{\sin(x)}\right)^2=\left(\lim_{x\to m\pi}\frac{\sin(Nx)}{\sin(x)}\right)^2=\left(\lim_{x\to m\pi}\frac{\mathrm{d}_x\sin(Nx)}{\mathrm{d}_x\sin(x)}\right)^2=\left(\lim_{x\to m\pi}\frac{N\cos(Nx)}{\cos(x)}\right)^2=N^2 \tag{10.6.150}$$

It follows that for $x=m\pi$ and therefore for zero values of A and B, the function ψ presents enormous peak height:

$$\psi=N^2 \tag{10.6.151}$$

Since the huge amplitude, the main peaks are those we observe experimentally the more easy. Thus, the angular position of the maxima of the function ψ is given by:

$$B=\sin(x_{\text{main peaks}})=0\Rightarrow x_{\text{main peaks}}=m\pi \tag{10.6.152}$$

with $m\in\mathbb{Z}$.

The value m, designate the "order number of the maximum's interference".

Let us apply these results to the interference relation:

$$\overline{S}(\theta,N,d)=\overline{S}_i(\theta)\left(\frac{\sin\left(N\dfrac{\pi}{\lambda}d\sin(\theta)\right)}{\sin\left(\dfrac{\pi}{\lambda}d\sin(\theta)\right)}\right)^2 \tag{10.6.153}$$

The peak of order m is center on the equivalent value θ that cancel the numerator and the denominator of this fraction such that:

$$\frac{\pi}{\lambda}d\sin(\theta)=m\pi\Rightarrow\sin(\theta)=m\frac{\lambda}{d} \tag{10.6.154}$$

with $m\in\mathbb{Z}$, hence:

$$\boxed{\sin(\theta_m)=m\frac{\lambda}{d}} \tag{10.6.155}$$

Thus an aperture network that we know the step value d can be used to measure the wavelength λ of an incident unknown light beam.

However, if the incident light is polychromatic (typically for astronomical observations), the above relation gives us for a given wavelength given the position of the interference fringes. Thus, an astronomer passing polychromatic light of its telescope through a diffraction network can make a spectroscopic analysis of the light.

This relation also gives us that for fixed values of m and d that the greater is λ the greater is the angle θ is in a range $[0,2\pi]$. Thus, the spectral lines arising from the impact of a polychromatic beam show a spectrum from violet (shorter wavelength so small angle) to red (high angle so large wavelength).

Using a goniometer, we measure the angles θ_m of the main peaks of order m for the greatest possible number of m values. We deduce from these measurement the slope λ of the function plot:

$$\lambda=d\frac{\Delta\sin(\theta_m)}{\Delta n} \tag{10.6.156}$$

The peak bottom is located at an angle $\theta_m+\Delta\theta_m$ where the numerator:

$$\sin\left(N\frac{\pi}{\lambda}d\sin(\theta)\right) \tag{10.6.157}$$

cancel for the first time after passage of the peak.

Since the argument of this function increases of $m\pi$ between two successive peaks (among all main and secondary peaks), it is equal to $N(m\pi)$ at the place of the peak of order m (main peak therefore) and has to travel π radians further to reach the bottom of the peak.

The numerator is then:

$$N\frac{\pi}{\lambda}d\sin(\theta_m+\Delta\theta_m)=mn\pi+\pi=(mn+1)\pi\Rightarrow\sin(\theta_m+\Delta\theta_m)=\left(m+\frac{1}{N}\right)\frac{\lambda}{d} \tag{10.6.158}$$

The angular distance $\Delta\theta_m$ between the top and the bottom of the main peak is the given by:

$$\sin(\theta_m+\Delta\theta_m)-\sin(\theta_m)=\left(m+\frac{1}{N}\right)\frac{\lambda}{d}-m\frac{\lambda}{d}=\frac{1}{N}\frac{\lambda}{d} \qquad (10.6.159)$$

But from the first order, we have $\theta_m \ll \Delta\theta_m$. The difference of two sinus gives (see section Trigonometry page 2519):

$$\begin{aligned}\sin(\theta_m+\Delta\theta_m)-\sin(\theta_m)&=2\sin\left(\frac{\theta_m+\Delta\theta_m-\theta_m}{2}\right)\cos\left(\frac{\theta_m+\Delta\theta_m+\theta_m}{2}\right)\\&=2\sin\left(\frac{\Delta\theta_m}{2}\right)\cos\left(\frac{2\theta_m+\Delta_m}{2}\right)\cong 2\sin\left(\frac{\Delta\theta_m}{2}\right)\cos(\theta_m)\end{aligned} \qquad (10.6.160)$$

A Maclaurin development (see section Sequences and Series page 1703) gives when taking the first term of the development $\sin(\alpha)\cong\alpha$:

$$2\sin\left(\frac{\Delta\theta_m}{2}\right)\cos(\theta_m)\cong\Delta\theta_m\cos(\theta_m) \qquad (10.6.161)$$

But we also have the identity (see section Trigonometry page 2519):

$$\cos(\theta_m)=\sqrt{1-\sin^2(\theta_m)} \qquad (10.6.162)$$

Hence the angular width of a peak of order m:

$$\Theta_m=2\Delta\theta_m=2\frac{1}{N}\frac{\lambda}{d}\frac{1}{\cos(\theta)}\frac{1}{Nd}\frac{2\lambda}{\sqrt{1-\sin^2(\theta_m)}}=\frac{1}{N}\frac{2\lambda}{\sqrt{d^2-d^2\sin^2(\theta_m)}} \qquad (10.6.163)$$

But as:

$$\sin(\theta_m)=m\frac{\lambda}{d}\Leftrightarrow d\sin(\theta_m)=m\lambda \qquad (10.6.164)$$

Therefore:

$$\boxed{\Theta_m=\frac{1}{N}\frac{2\lambda}{d^2-(m\lambda)^2}} \qquad (10.6.165)$$

It is clear that two superposed light lines will be seen as distinct if they are separated by an angular distance equal to their angular width. The expression:

$$\sin(\theta_m)=m\frac{\lambda}{d} \qquad (10.6.166)$$

establishes that to two angular positions correspond two wavelengths. So we can give the separation of two light lines by $\Delta\lambda=\lambda_2-\lambda_1$ instead of $\Delta\theta=\theta_2-\theta_1$.

Therefore from:

$$\Theta_m=2\Delta\theta_m=\frac{2}{N}\frac{\lambda}{d}\frac{1}{\cos(\theta_m)} \qquad (10.6.167)$$

we get:

$$\cos(\theta_m)\Delta\theta_m=\frac{1}{N}\frac{\lambda}{d} \tag{10.6.168}$$

But:

$$\sin(\theta_m)=m\frac{\lambda}{d}\Rightarrow m\frac{\delta\lambda}{d}=\Delta\sin(\theta_m) \tag{10.6.169}$$

When $\Delta\theta_m$ and θ_m are small, we have:

$$\Delta\sin(\theta_m)\cong\cos(\theta_m)\Delta\theta_m \tag{10.6.170}$$

Which brings us to write by substitution:

$$m\frac{\Delta\lambda}{d}\cong\frac{1}{N}\frac{\lambda}{d}\Rightarrow\frac{\Delta\lambda}{\lambda}\cong\frac{1}{mN} \tag{10.6.171}$$

The resolving power R of a network is its ability to separate two spectral lines of neighbouring wavelengths λ and $\lambda+\Delta\lambda$ such as:

$$\boxed{R=\frac{\lambda}{\Delta\lambda}\cong mN} \tag{10.6.172}$$

We see that the resolving power increases proportionally with the diffraction order. This relation is very important in spectroscopy for astronomy. We can also calculate what should be the minimum number N of lines that must have a network capable of separating two wavelengths in a spectrum of given order m.

10.6.2.3 Young's interference experiment

According to the wave-particle duality principle, light behaves both as a wave and as a particle (material particle). It is the solving of problems like those of the black-body (see section Thermodynamics page 3704), of the photoelectric effect (see section Nuclear Physics page 5106) or that of the Compton effect (see section Nuclear Physics page 5096) , which revealed the existence of this duality.

But we will now study the most mind blowing proof of the duality aspect of the matter at the atomic scale using the Young's experiment. We will address this in a simplified manner as a special case of the rectangular slots network that has the advantage of being able to be easy to put in practice in an experiment showing the dual and probabilistic behaviour of matter at the atomic scale.

So to study this subject, let us consider a light source S, which radiates monochromatic wave Ψ of wavelength λ through two slots F_1 and F_2 perforated in an opaque barrier to light, as shown in the figure below:

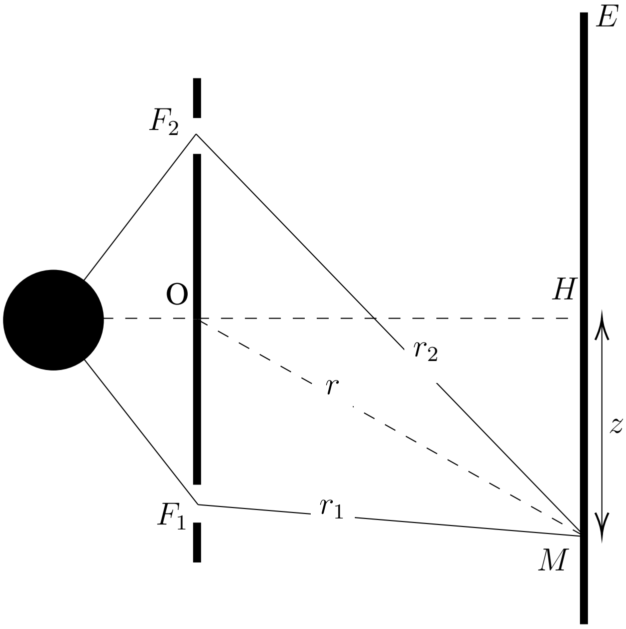


Figure 10.233: **Implementation of Young's experience-slit**

> **Remark**
>
> The advantage of that device is that it can produce two sources of coherent light. That is to say two sources whose phase difference is constant throughout the experiment.

We have on a viewing (projection) screen E at a point H such that the distance is:

$$\overline{\text{O}H} = D \gg \overline{F_1F_2} = a \qquad (10.6.173)$$

where a would typically be of the order of the millimetres and D of the meter.

The wave Ψ will after passing through the slots F_1 and F_2, as we have already seen it just previously, generates sub-waves $\Psi_1(r_1,t)$ and $\Psi_2(r_1,t)$ of same pulsation ω that will travel respectively the paths r_1 and r_2 and that will go interfere at the point M of the screen E that we are interested for.

If the interference is constructive in M, this point will then be located on a bright fringe and if the interference is destructive M, it will be on a dark fringe. To observe this, let us first write the resulting wave at M:

$$\Psi_M = \Psi_1(r_1,t) + \Psi(r_2,t) \qquad (10.6.174)$$

in which we have in terms of phasor (see section Wave Mechanics page 3510):

$$\Psi_1(r_1,t) = Ae^{\mathrm{i}(kr_1-\omega t)} \quad \text{and} \quad \Psi_2(r_2,t) = Ae^{\mathrm{i}(kr_2-\omega t)} \qquad (10.6.175)$$

where A is the amplitude, k is the wave vector and t is the time variable as we have already discussed in detail in the section of Wave Mechanics.

Now let make a change of variable (just to not have to manage long exponential functions):

$$\Psi_1(r_1,t) = Ae^{\mathrm{i}kD_1} \quad \text{and} \quad \Psi_2(r_2,t) = Ae^{\mathrm{i}kD_2} \qquad (10.6.176)$$

Remark

We will see later that in fact $D_1 = r_1$ and $D_2 = r_2$.

To calculate the intensity at the point M, we will take the complex norm (module) of Ψ_M that is thus written as the product of the complex variable and its conjugate (see section Numbers page 244):

$$I = |\Psi_M| = \Psi_M \bar{\Psi}_M = (\Psi_1 + \Psi_2)(\overline{\Psi_1 + \Psi_2}) \tag{10.6.177}$$

Remark

This calculation is very important because the analogy with the Wave Quantum Physics is very strong at this level and similar to the calculation of the probability amplitude (see section Wave Quantum Physics page 4693).

Therefore:

$$\begin{aligned} I &= A^2\left(e^{\mathrm{i}kD_1} + e^{\mathrm{i}kD_2}\right)\left(e^{-\mathrm{i}kD_1} + e^{-\mathrm{i}kD_2}\right) = A^2\left[1 + e^{\mathrm{i}k(D_1 - D_2)} + e^{\mathrm{i}k(D_2 - D_1)} + 1\right] \\ &= A^2\left[2 + e^{\mathrm{i}k(D_1 - D_2)} + e^{\mathrm{i}k(D_2 - D_1)}\right] = A^2\left[2 + \cos(k(D_1 - D_2)) + \cos(k(D_2 - D_2))\right] \\ &= A^2[2 + 2\cos(k(D_1 - D_2))] \end{aligned} \tag{10.6.178}$$

The intensity is then maximum if and only if:

$$\cos(k(D_1 - D_2)) = 1 \tag{10.6.179}$$

So that:

$$k(D_1 - D_2) = 2n\pi \Leftrightarrow D_1 - D_2 = n\lambda \tag{10.6.180}$$

with $n \in \mathbb{N}$. Which gives:

$$D_1 - D_2 = (r_1 - \omega t) - (r_2 - \omega t) = r_1 - r_2 = n\lambda \tag{10.6.181}$$

Remark

It is here that we see obviously that $D_1 = r_1$ and $D_2 = r_2$

The intensity is zero if and only if:

$$\cos(k(D_1 - D_2)) = -1 \tag{10.6.182}$$

So that:

$$k(D_1 - D_2) = 2\left(n + \frac{1}{2}\right)\pi \Leftrightarrow D_1 - D_2 = \left(n + \frac{1}{2}\right)\lambda \tag{10.6.183}$$

with $n \in \mathbb{N}$. Which gives:

$$r_1 - r_2 = \left(n + \frac{1}{2}\right)\lambda \tag{10.6.184}$$

Now we must calculate $r_1 - r_2$ in function of z to find out what we see on the projection screen E.

For this let us consider the following figure:

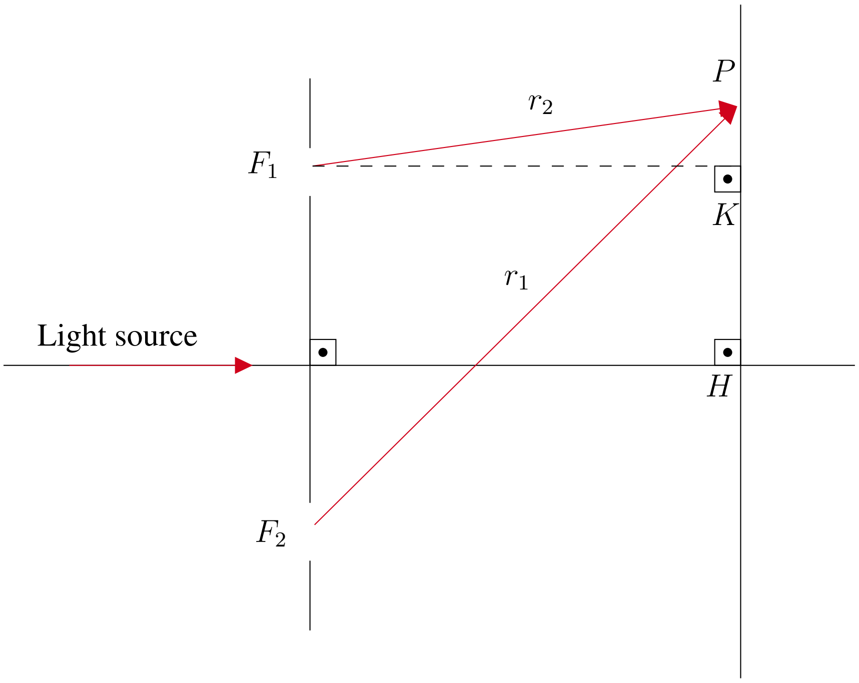


Figure 10.234: **Magnification and special case of Young experiment**

where $\overline{F_1P}$ and $\overline{F_2P} = r_2$.

We have on our figure:

$$r_1^2 = D^2 + \overline{PK}^2 \tag{10.6.185}$$

But:

$$\overline{PK} = z - \frac{a}{2} \tag{10.6.186}$$

So we have:

$$r_1^2 = D^2 + \left(z - \frac{a}{2}\right)^2 \Rightarrow r_1 = D\sqrt{1 + \frac{1}{D^2}\left(z - \frac{a}{2}\right)^2} \tag{10.6.187}$$

As z and a are small relatively to D and using the approximation:

$$\sqrt{1+\varepsilon} \cong 1 + \frac{\varepsilon}{2} \tag{10.6.188}$$

if ε is small compared to 1. Then we have:

$$r_1 = D\sqrt{1 + \frac{1}{D^2}\left(z - \frac{a}{2}\right)^2} \cong D\left(1 + \frac{1}{2D^2}\left(z - \frac{a}{2}\right)^2\right) = D + \frac{1}{2D}\left(z - \frac{a}{2}\right)^2 \tag{10.6.189}$$

Also:

$$r_2=D\sqrt{1+\frac{1}{D^2}\left(z+\frac{a}{2}\right)^2}\cong D\left(1+\frac{1}{2D^2}\left(z+\frac{a}{2}\right)^2\right)=D+\frac{1}{2D}\left(z+\frac{a}{2}\right)^2 \tag{10.6.190}$$

So subtracting those two relations we get:

$$r_1-r_2=-z\frac{a}{D} \tag{10.6.191}$$

So finally using the relation:

$$r_1-r_2=n\lambda \tag{10.6.192}$$

It comes:

$$z=-\frac{n\lambda D}{a} \tag{10.6.193}$$

Thus, the distance between two consecutive maxima is:

$$i_{\max}=\frac{\lambda D}{a} \tag{10.6.194}$$

and is named "interfringe".

For fringes with zero intensity it comes immediately:

$$z=-\left(n+\frac{1}{2}\right)\frac{\lambda D}{a} \tag{10.6.195}$$

This relation indicates that the intensity I has minima (dark fringes) and maxima (bright fringes) distributed along the z direction periodically. This does not surprise us more than that for now because it derives from the more general case studied above.

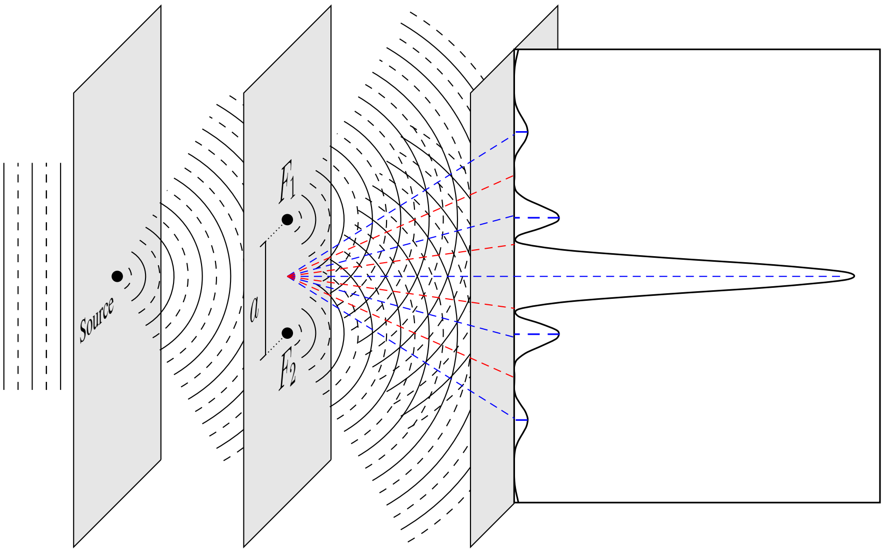


Figure 10.235: **Interfinge simplified schema for the Young experiment**

It should be noticed that the above calculations show that the intensity of the fringes is equal everywhere. But we observe experimentally (see figure above) that their intensity decreases with distance from the projection screen center. As we have already seen it, two are at the origin of this observation:

1. The slots have a given width, which implies a diffraction phenomenon. Indeed, a light sent on a small hole does not go out isotropically. This results in that the light is predominantly directed forward. This effect is reflected in the figure observed after Young's slots: the intensity of fringes decreases gradually as we moves away from the center of the projection screen.

2. The fact that the waves emitted in F_1 and F_2 are spherical waves, that is to say, their amplitude decreases gradually as we go away from the screen projection center. Thus the amplitude of F_1 and F_2 will not be the same at the point M.

So our calculations are remains approximate relatively to the study we made of the rectangular slots network but that is how Young's slots experiment is presented in most schools and this is enough to highlight the main result.

The original experiment of Thomas Young can be interpreted using the simple Fresnel's laws as we have done with the network of rectangular of slots. Which highlights the wave nature of light. But this experience has subsequently be refined, particularly by ensuring that the source S emits a quantum at a time. For example, it is possible in the third quarter of the 120th century (holocene calendar) to emit photons or electrons or atoms one by one. These are detected one by one placed on the screen after the double-slit. We see then that these impacts form gradually the interference pattern. By conventional laws concerning the trajectories of these particles, it is impossible to interpret this phenomenon!!! Hence the importance of the theoretical study of the Young's experiment!!!

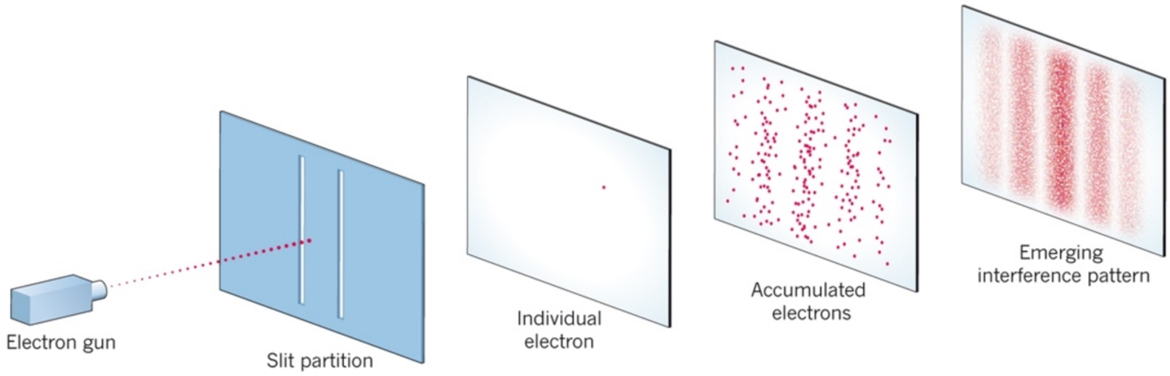


Figure 10.236: **Young's interference experiment (source: ?)**

From left to right, top to bottom, here are the patterns obtained by accumulating 10, 300, $2,000$ and $6,000$ electrons with a flow of 10 electrons / second (the same experiment with the same result has been reproduced with neutrons and atoms!). The accumulation of electrons eventually form interference fringes which is quite confusing a priori!

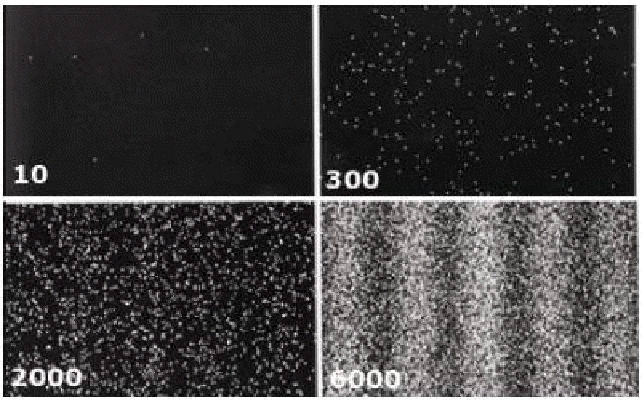


Figure 10.237: **Photos of patterns obtained in a real Young's experience (source: ?)**

We will come back on this crucial phenomenon in the section of Wave Quantum Physics to say a little bit more about it.

For summary here is another more detailed schema of what we haven seen so far:

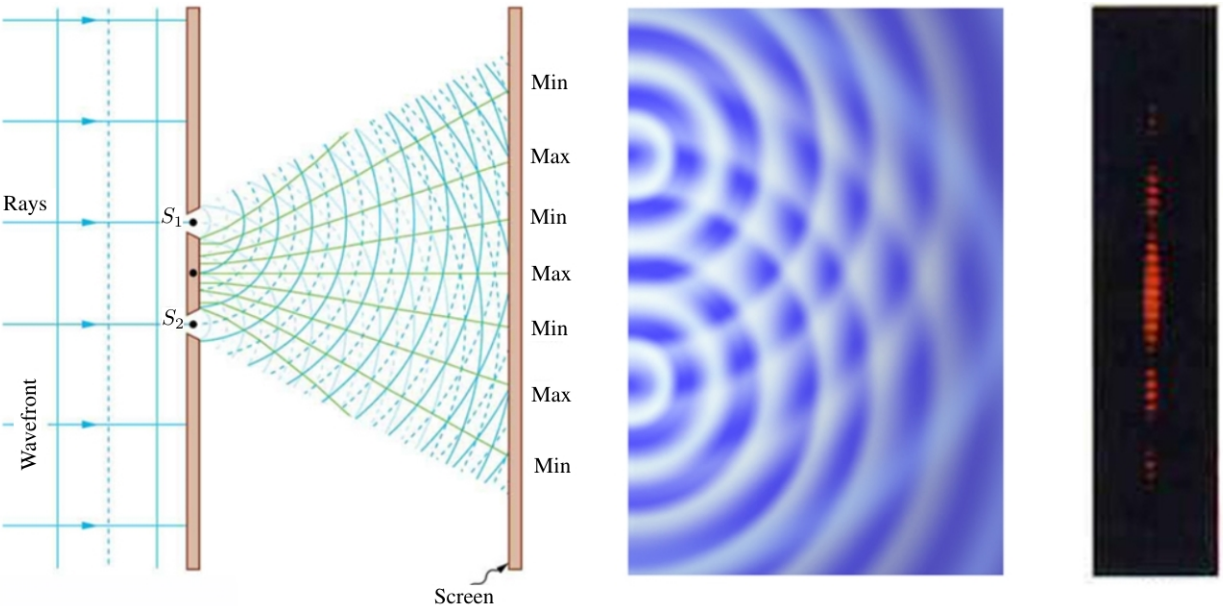


Figure 10.238: **Young experiment summary (source: OpenStax)**

MIT's experiment of 12025 (holocene calendar) on *Coherent and Incoherent Light Scattering by Single-Atom Wave Packets* (see [172]) offers a groundbreaking look into the quantum behavior of light and atoms. By using ultracold atoms released from an optical lattice, researchers created a highly controlled version of the double-slit experiment - this time with single atoms acting as slits and single photons as probes. They observed how light scattered off these atoms, revealing whether the scattering was coherent (wave-like) or incoherent (particle-like). Crucially, they showed that the coherence of scattered light is independent of whether the atom is trapped or free, challenging assumptions about the role of atomic motion and entanglement in quantum measurements.

This experiment directly addresses a century-old debate between Albert Einstein and Niels Bohr. Einstein believed that photons should behave like particles, passing through one slit and exerting a measurable force - thus preserving classical realism. But MIT's results confirm that light's wave–particle duality is fundamentally incompatible with such realism: observing which slit a photon passes through destroys the interference pattern, proving that both wave and particle aspects cannot be seen simultaneously. In doing so, the experiment validates Bohr's interpretation of quantum mechanics and shows that Einstein's view, while brilliant, doesn't hold up under quantum scrutiny!

10.6.3 Fresnel diffraction

In optics, the "Fresnel diffraction" is diffraction mathematical analysis method that is applied to the propagation of waves in the near field. In other words it is used to calculate the diffraction pattern generated by waves passing through an aperture or around an object, when viewed from relatively close to the object. In contrast the diffraction pattern in the far field region is given by the "Fraunhofer diffraction" methods used earlier above.

Remarks

R1. The formula we used earlier for the Fraunhofer diffraction and the one we will use now for the Fresnel diffraction are in fact both an approximation of a formula named the "Fresnel-Kirchhoff diffraction" who takes into account secondary waves that are propagating in various directions as per Huygens' principle (for a nice derivation see [568]).

R2. Also the reader should notice that our previous study of Fraunhofer diffraction and the one that will follow on Fresnel diffraction belongs to the field of "scalar diffraction theory", which ignores polarization effects. In some situations, ignoring polarization is benign, but in other situations ignoring polarization effects produces significant errors!

Disclaimer! The huge majority of what will follow, including the figures is inspired and copied from the very excellent textbooks on Optics from Eugene Hecht [239].

10.6.3.1 Case of a rectangular aperture

Consider the configuration depicted in the figure below:

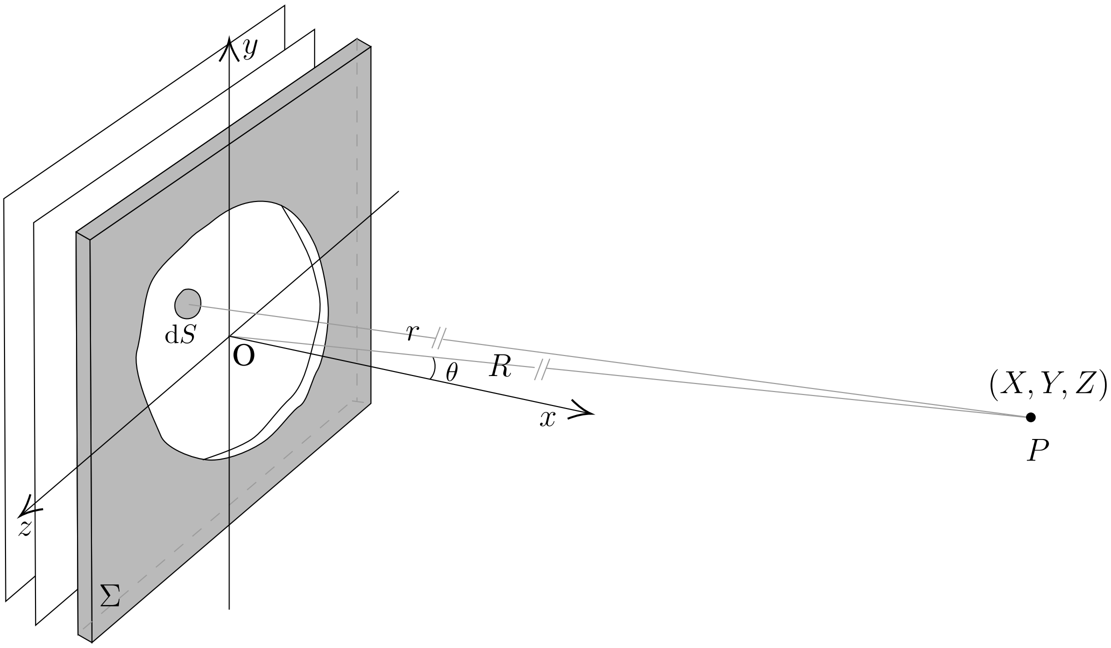


Figure 10.239: Fraunhofer diffraction from an arbitrary aperture, where r and R are very large compared to the size of the hole (source: [239])

A monochromatic plane wave propagating in the x-direction is incident on the opaque diffracting screen Σ with an aperture of area A. We wish to find the consequent (far-field) flux-density distribution in space or equivalently at some arbitrary distant point P. According to the Huygens principle, a differential area $\mathrm{d}S$, within the aperture, may be envisioned as being covered with coherent secondary point sources. But $\mathrm{d}S$ is much smaller in extent than is λ, so that all the contributions at P remain in-phase and interfere constructively. This is true regardless of θ; that is, $\mathrm{d}S$ emits a spherical wave. If ρ_A is the source strength per unit area, assumed to be constant over the entire aperture, then the optical disturbance at P due to $\mathrm{d}S$ is either the real or imaginary part of:

$$\mathrm{d}f=\frac{\hat{f}}{Ar}e^{\mathrm{i}(\omega t-kr)}\mathrm{d}S=\frac{\rho_A}{r}e^{\mathrm{i}(\omega t-kr)}\mathrm{d}S \tag{10.6.196}$$

The choice is yours and depends only on whether you like sine or cosine waves, there being no difference except for a phase shift. The distance from $\mathrm{d}S$ (of coordinate $(0,y,z)$) to P (of coordinate (X,Y,Z)) is (see figure above):

$$r=\left[X^2+(Y-y)^2+(Z-z)^2\right]^{1/2} \tag{10.6.197}$$

and as we have seen, the Fraunhofer condition occurs when this distance approaches infinity. As before, it will suffice to replace r by the distance $\overline{OP}$, that is, R, in the amplitude term, as long as the aperture is relatively small. But the approximation for the phase needs to be treated a bit more carefully; $k=2\pi/\lambda$ is a large number. To that end we expand out the above relation, by making use of:

$$R=\left[X^2+Y^2+Z^2\right]^{1/2} \tag{10.6.198}$$

we get:

$$r=(X^2+Y^2-2Yy+y^2+Z^2+2Zz+z^2)^{1/2}=R\left[1+\left(y^2+z^2\right)/R^2-2(Yy+Zz)/R^2\right]^{1/2} \tag{10.6.199}$$

In the far-field case R is very large in comparison to the dimensions of the aperture, and the $\left(y^2+z^2\right)/R^2$ term is certainly negligible. Since P is very far from Σ,θ can still be kept small, even though Y and Z are fairly large, and this mitigates any concern about the directionality of the emitters (the obliquity factor). Now:

$$r\cong R\left[1-2(Yy+Zz)/R^2\right]^{1/2} \tag{10.6.200}$$

and dropping all but the first two terms in the binomial expansion (see section Algebra page 1366), we have:

$$r\cong R\left[1-(Yy+Zz)/R^2\right] \tag{10.6.201}$$

The total disturbance arriving at P is:

$$\tilde{f}\cong\frac{\rho_A e^{\mathrm{i}(\omega t-kR)}}{R}\iint\limits_{\text{Aperture}} e^{\mathrm{i}k(Yy+Zz)/R}\mathrm{d}S \tag{10.6.202}$$

Consider now the specific configuration shown in the figure below:

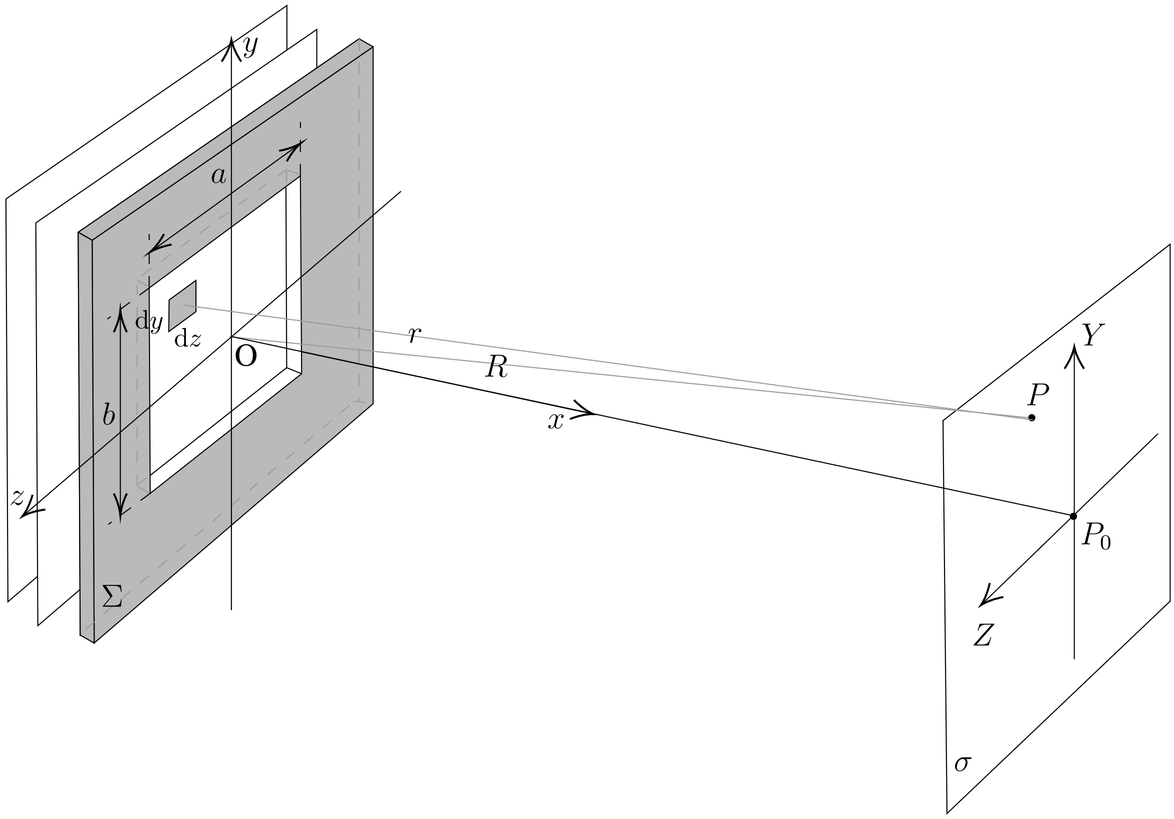


Figure 10.240: **A rectangular aperture for near-field analysis (source: [239])**

The previous relation can now be written as:

$$f \cong \frac{\rho_A e^{\mathrm{i}(\omega t - kR)}}{R} \int\limits_{-b/2}^{b/2} e^{\mathrm{i}kYy/R} \mathrm{d}y \int\limits_{-a/2}^{a/2} e^{\mathrm{i}kZz/R} \mathrm{d}z \tag{10.6.203}$$

where $\mathrm{d}S = \mathrm{d}y\mathrm{d}z$. With $\beta' \equiv kbY/2R$ and $\alpha' \equiv kaZ/2R$, we have (we used Euler formula as proved at page 251):

$$\int\limits_{-b/2}^{+b/2} e^{\mathrm{i}kYy/R} \mathrm{d}y = b\left(\frac{e^{\mathrm{i}\beta'} - e^{-\mathrm{i}\beta'}}{2\mathrm{i}\beta'}\right) = b\left(\frac{\sin(\beta')}{\beta'}\right) \tag{10.6.204}$$

and similarly:

$$\int\limits_{-a/2}^{+a/2} e^{\mathrm{i}kZz/R} \mathrm{d}z = a\left(\frac{e^{\mathrm{i}\alpha'} - e^{-\mathrm{i}\alpha'}}{2\mathrm{i}\alpha'}\right) = a\left(\frac{\sin(\alpha')}{\alpha'}\right) \tag{10.6.205}$$

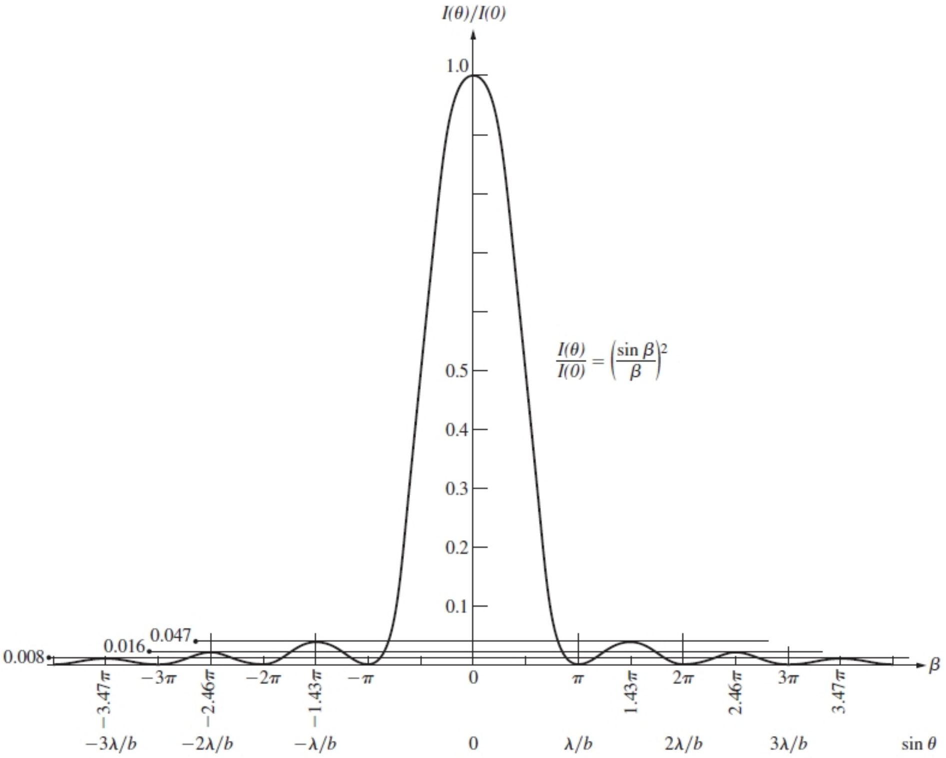
so that:

$$\boxed{f \cong \frac{A\rho_A e^{\mathrm{i}(\omega t - kR)}}{R} \left(\frac{\sin(\alpha')}{\alpha'}\right)\left(\frac{\sin(\beta')}{\beta'}\right) = \frac{A\rho_A e^{\mathrm{i}(\omega t - kR)}}{R} \mathrm{sinc}^2(\alpha')\,\mathrm{sinc}^2(\beta')} \tag{10.6.206}$$

where A is the area of the aperture. Let us denote that relation as following:

$$I(Y,Z) \cong I(0)\mathrm{sinc}^2(\alpha')\mathrm{sinc}^2(\beta') \tag{10.6.207}$$

where $I(0)$ is the irradiance at P_0; that is, at $Y=0, Z=0$. At values of Y and Z such that $\alpha'=0$ or $\beta'=0, I(Y,Z)$ assumes the following familiar shape (source: [239]):



When β' and α' are non-zero integer multiples of π or, equivalently, when Y and Z are non-zero integer multiples of $\lambda R/b$ and $\lambda R/a$, respectively, $I(Y,Z)=0$, and we have a rectangular grid of nodal lines. Notice that the pattern in the Y-Z directions varies inversely with the y-,z-aperture dimensions. A horizontal, rectangular opening will produce a pattern with a vertical rectangle at its center and vice versa:

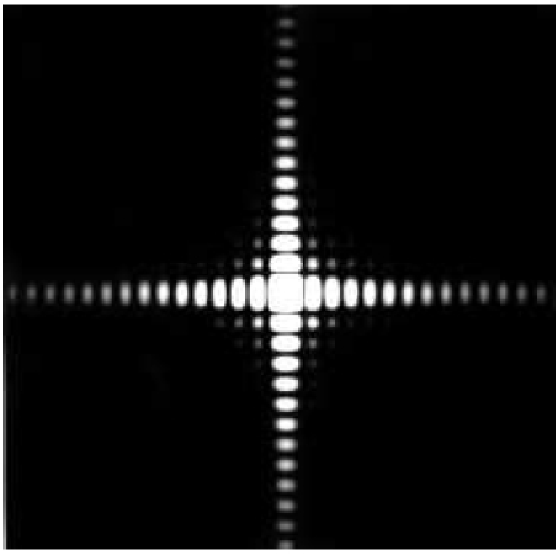


Figure 10.241: **Fraunhofer pattern of a square aperture (source: [239])**

An illustration or 3D computer simulation would give the following figures (source: [239]):

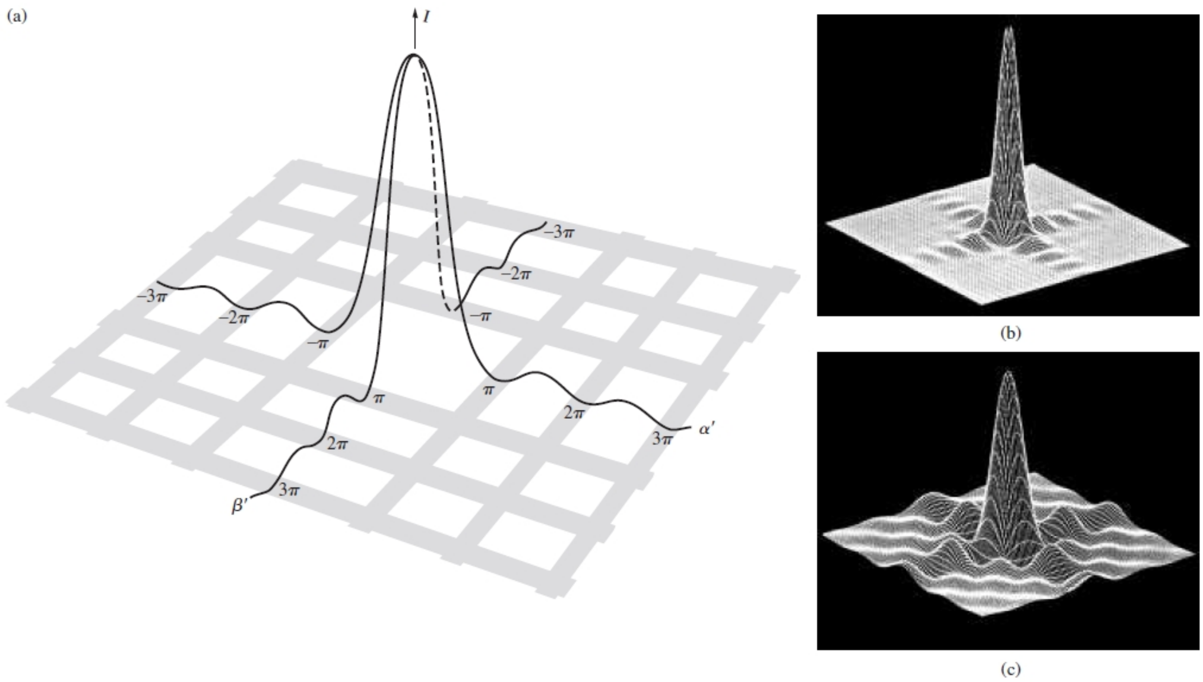
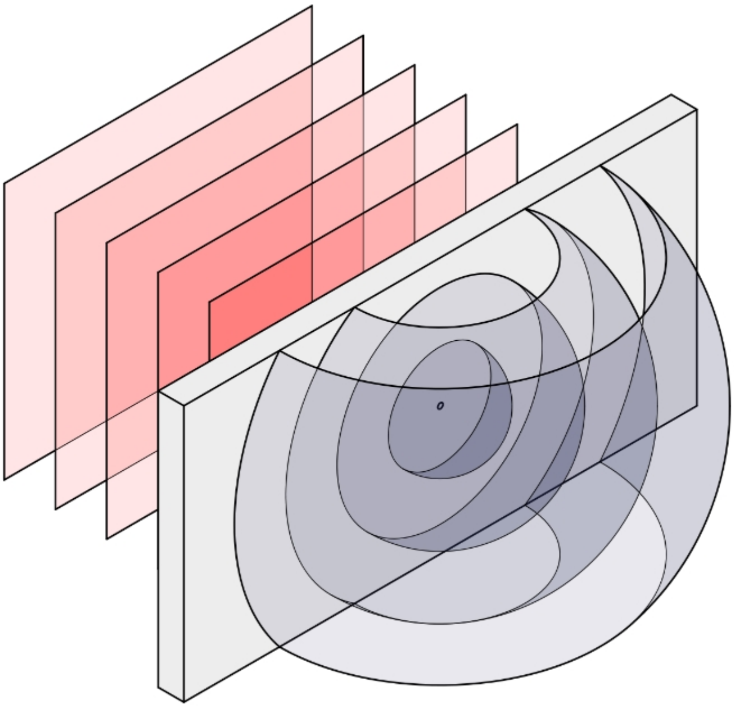


Figure 10.242: **(a) The irradiance distribution for a square aperture. (b) The irradiance produced by Fraunhofer diffraction at a square aperture. (c) The electric-field distribution produced by Fraunhofer diffraction via a square aperture. (source: R.G. Wilson, Illinois Wesleyan University)**

10.6.3.2 Case of a circular aperture

Most experimental situations in optics (e.g. telescopes, microscopes, LASER) have circular apertures, so the study of diffractions from such apertures is of particular interest. It is interesting to know that in 11818 (holocene calendar) Augustin Fresnel entered a competition sponsored by the French Academy. His paper on the theory of diffraction ultimately won first prize, but not until it had provided the basis for a rather interesting story. The judging committee consisted of Pierre Laplace, Jean. B. Biot, Siméon D. Poisson, Dominique F. Arago and Joseph L. Gay-Lussac... a quite formidable group indeed. Poisson, who was an ardent critic of the wave description of light, deduced a remarkable and seemingly untenable conclusion from Fresnel's theory. He showed that a bright spot would be visible at the center of the shadow of a circular opaque obstacle, a result that then proved the absurdity of Fresnel's treatment. This surprising prediction, fashioned by Poisson as the death blow to the wave theory, was almost immediately verified experimentally by Arago; the spot actually existed! Amusingly enough, "Poisson's spot" or "Arago spot", as it is now called, had been observed many years earlier (11723 according to holocene calendar), but this work had long gone unnoticed.

We start with a plane wave incident normally on a circular hole with radius a:

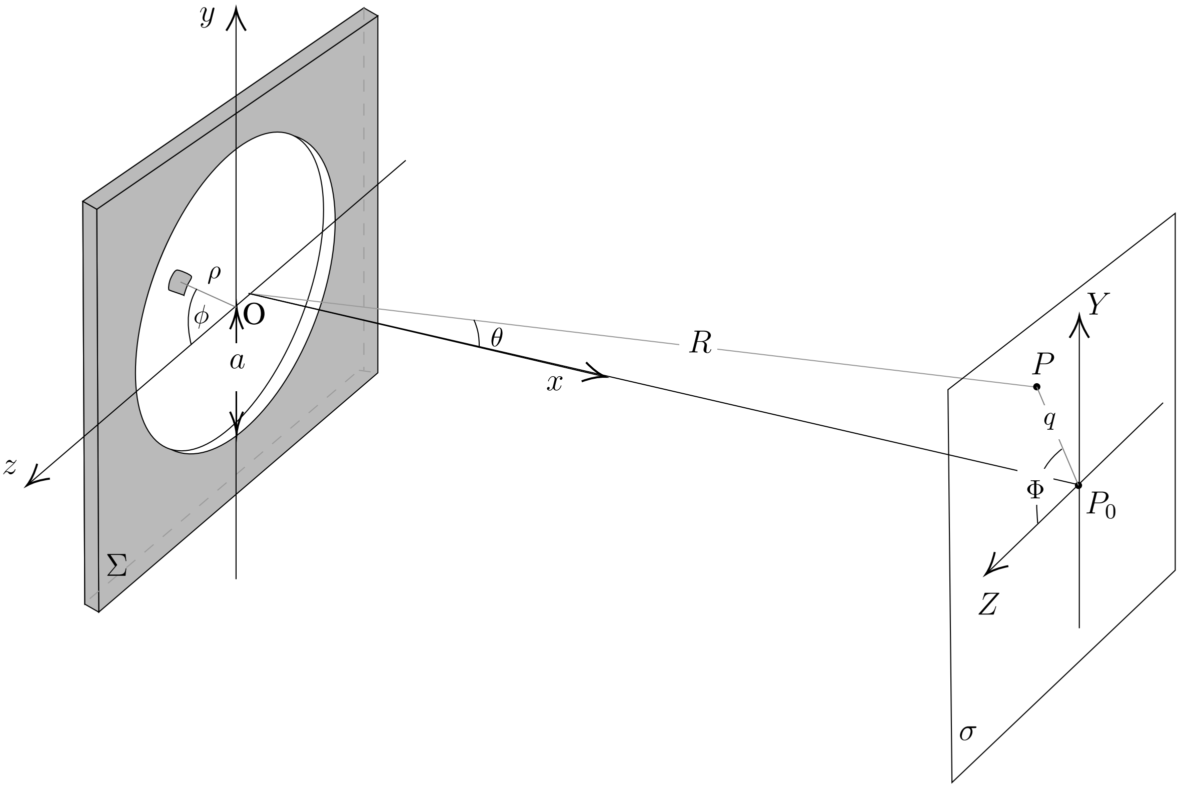


in an otherwise opaque screen, and ask ourselves what is the distribution of the intensity of the light on a screen on distance $R \gg a$ away?

We can use the same expression for the f field that we had for the rectangular aperture as for any possible aperture, as long as the limits of integration are appropriate. So we can write for the optical disturbance at point P in the far-field case:

$$f \cong \frac{\rho_A}{R} e^{\mathrm{i}(kR-\omega t)} \iint\limits_{\text{aperture}} e^{-\mathrm{i}k(Yy+Zz)/R} \mathrm{d}y\mathrm{d}z \tag{10.6.208}$$

For a circular aperture this integration is most easily done with cylindrical coordinates. Look at the figure (source: [239]):



Then we have:

$$\begin{aligned}z&=\rho\cos(\phi)\\y&=\rho\sin(\phi)\\Z&=q\cos(\Phi)\\Y&=q\sin(\Phi)\end{aligned}\tag{10.6.209}$$

The differential area is now:

$$\mathrm{d}S=\mathrm{d}x\mathrm{d}y=\rho\mathrm{d}\rho\mathrm{d}\phi\tag{10.6.210}$$

Then:

$$Yy+Zz=\rho q\cos(\varphi)\cos(\Phi)+\rho q\sin(\varphi)\sin(\Phi)\tag{10.6.211}$$

or:

$$Yy+Zz=\rho q\cos(\varphi-\Phi)\tag{10.6.212}$$

and the integral becomes:

$$f\cong\frac{\rho_A}{R}e^{\mathrm{i}(kR-\omega t)}\int\limits_0^a\int\limits_0^{2\pi}e^{\mathrm{i}K\rho q\cos(\varphi-\Phi)/R}\rho\mathrm{d}\rho\mathrm{d}\varphi\tag{10.6.213}$$

In order to do this integral we need to learn a little about Bessel functions (see section Sequences and Series page 1768) and especially the following first kind of order 0 in its integral representation (see section Differential and Integral Calculus page 1564):

$$J_0(u)=\frac{1}{2\pi}\int\limits_0^{2\pi}e^{\mathrm{i}u\cos(v)}\mathrm{d}v\tag{10.6.214}$$

that is a special case (for reminder...) of a more general definition of Bessel functions :

$$J_m(u)=\frac{1}{2\pi}\int\limits_0^{2\pi} e^{\mathrm{i}(mv+u\cos(v))}\mathrm{d}v \tag{10.6.215}$$

They have a number of interesting properties such as the recurrence relations (see section Sequences and Series page 1770):

$$\frac{\mathrm{d}}{\mathrm{d}u}[u^m J_m(u))]=u^m J_{m-1}(u) \tag{10.6.216}$$

so that for example when $m=1$:

$$\int\limits_0^u u' J_0(u')\mathrm{d}u'=uJ_1(u) \tag{10.6.217}$$

In order to numerically calculate the value of a Bessel function one uses the expansion (see section Sequences and Series page 1770):

$$J_m(x)=\sum_{s=0}^{+\infty}\frac{(-1)^2}{s!(m+s)!}\left(\frac{x}{2}\right)^{m+2s} \tag{10.6.218}$$

Now we want to evaluate the integral:

$$f\cong\frac{\rho_A}{R}e^{\mathrm{i}(kR-\omega t)}\int\limits_0^a\int\limits_0^{2\pi} e^{-\mathrm{i}k\rho q\cos(\varphi-\Phi)/R}\rho\mathrm{d}\rho\mathrm{d}\varphi \tag{10.6.219}$$

which we can do at any value of Φ since the problem is symmetric about Φ. So we can simplify things greatly if we do the integral at $\Phi=0$:

$$f\cong\frac{\rho_A}{R}e^{\mathrm{i}(kR-\omega t)}\int\limits_0^a\int\limits_0^{2\pi} e^{-\mathrm{i}k\rho q\cos(\varphi)/R}\rho\mathrm{d}\rho\mathrm{d}\varphi \tag{10.6.220}$$

which becomes:

$$f\cong\frac{\rho_A}{R}e^{\mathrm{i}(kR-\omega t)}2\pi\int\limits_0^a J_0\left(-\frac{k\rho q}{R}\right)\rho\mathrm{d}\rho \tag{10.6.221}$$

Now J_0 is an even function so we can drop the minus sign and rewrite the expression as:

$$f\cong\frac{\rho_A}{R}e^{\mathrm{i}(kR-\omega t)}2\pi\int\limits_0^a J_0\left(\frac{k\rho q}{R}\right)\rho\mathrm{d}\rho \tag{10.6.222}$$

To do this integral we change variables:

$$\begin{aligned}w&=\frac{k\rho q}{R}\\ \rho&=\frac{wR}{kq}\\ \mathrm{d}\rho&=\frac{R}{kq}\mathrm{d}w\end{aligned} \tag{10.6.223}$$

So that:

$$\begin{aligned}\int\limits_0^a J_0\left(\frac{k\rho q}{R}\right)\rho\mathrm{d}\rho &= \int\limits_0^{kaq/R}\left(\frac{R}{kq}\right)^2 J_0(w)w\mathrm{d}w = \left(\frac{R}{kq}\right)^2\left(\frac{kaq}{R}\right)J_1\left(\frac{kaq}{R}\right)\\ &= a^2\left(\frac{R}{kaq}\right)J_1\left(\frac{kaq}{R}\right) = a^2\frac{J_1\left(\dfrac{kaq}{R}\right)}{\dfrac{kaq}{R}}\end{aligned} \tag{10.6.224}$$

So finally we have the result:

$$\boxed{f \cong \rho_A\frac{e^{\mathrm{i}(kR-\omega t)}}{R}2\pi a^2\frac{J_1\left(\dfrac{kaq}{R}\right)}{\dfrac{kaq}{R}}} \tag{10.6.225}$$

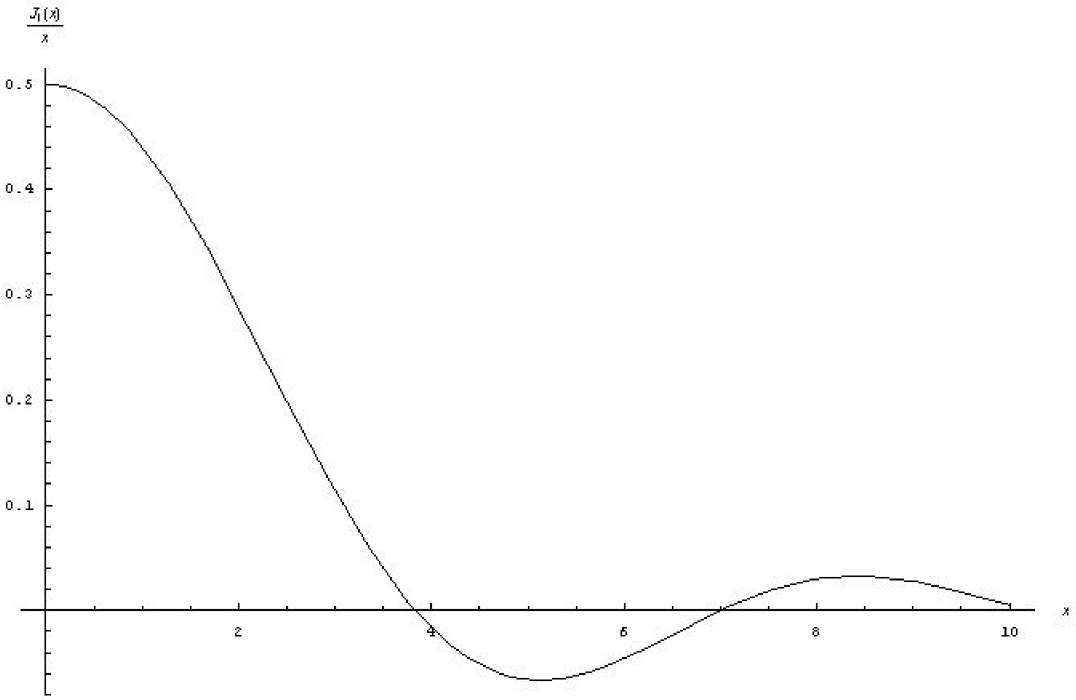
Or recognizing that πa^2 is the area of the aperture A and squaring to get the intensity we write:

$$I \cong I_0\left[\frac{2J_1\left(\dfrac{kaq}{R}\right)}{\dfrac{kaq}{R}}\right]^2 \tag{10.6.226}$$

If we want to write this in terms of the angle θ then one uses the fact that $q/R = \sin(\theta)$:

$$I(\theta) \cong I_0\left[\frac{2J_1(ka\sin(\theta))}{ka\sin(\theta)}\right]^2 \tag{10.6.227}$$

Below is a plot made with Mathematica of the function $J_1(x)/x$. Notice how it peaks at $1/2$ which is why there is the factor of two in the expression for the irradiance:



We easily do the plot of $(J_1(x)/x)^2$ with Maple™ 4.00b for example:

```
>with(plots):
>i:=(BesselJ(1,Pi*theta)/(Pi*theta))^2;
>plot(i,theta=-4..4);
```

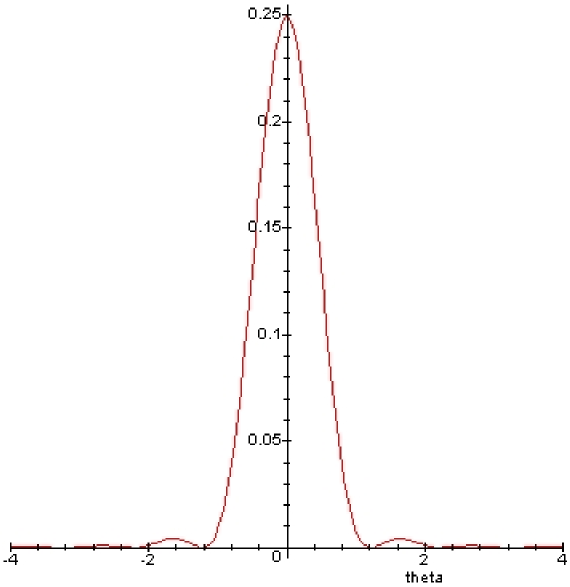
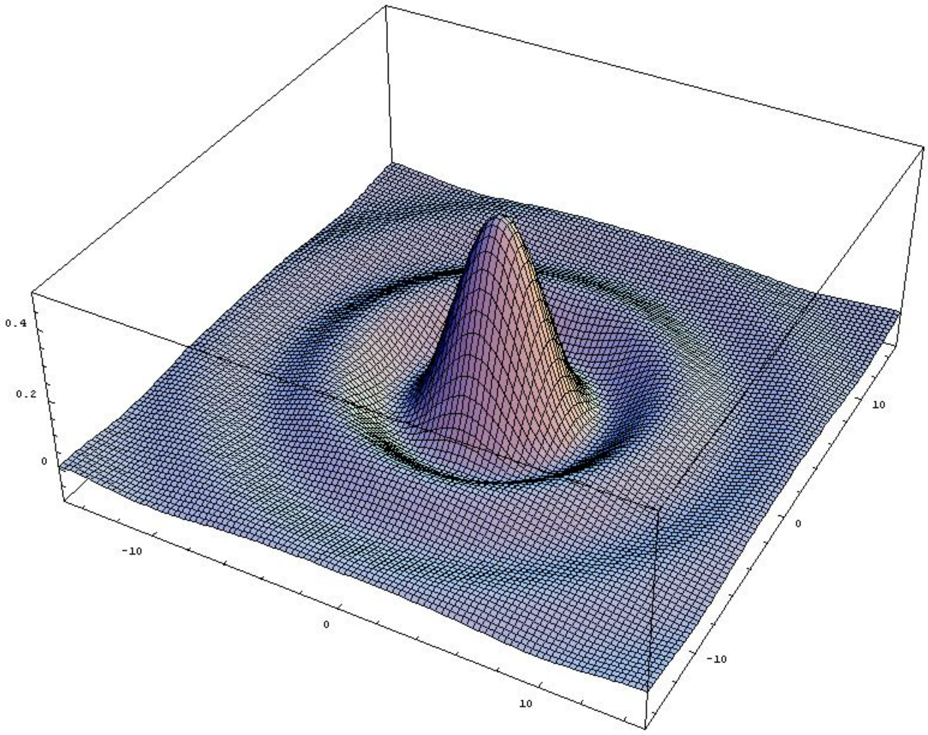


Figure 10.243: **Fresnel circular aperture diffraction pattern 2D plot in Maple™ 4.00b**

Below is a 3D plot also made with Mathematica of the same thing (i.e. $J_1(r)/r$). Notice the rings:



Below is a 3D plot of $(J_1(r)/r)^2$ which corresponds to the irradiance one sees. The central peak out to the first ring of zero is named the "Airy disk" (which contains 84% of the total light intensity) - named after Sir George Biddell Airy, an English astronomer who described this pattern mathematically in 11834 (holocene calendar). We can plot that with Maple™ 4.00b:

```
>with(plots):
>i:=(BesselJ(1,Pi*sqrt(x^2+y^2))/(Pi*sqrt(x^2+y^2)))^2;
>contourplot(i,x=-3..3,y=-3..3,filled=true,coloring=[white,blue],
contours=[0.001,0.003,0.008,0.012,0.016,0.02],grid=[100,100]);
```

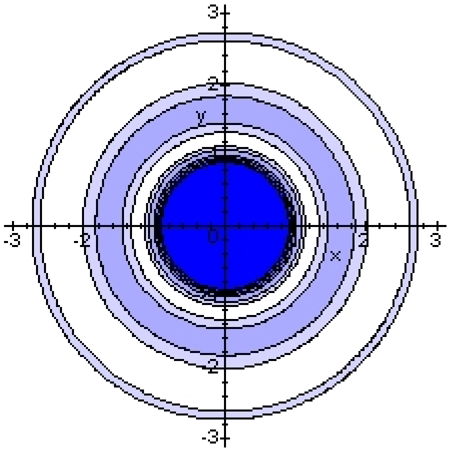


Figure 10.244: **Fresnel Airy disk 2D plot in Maple™ 4.00b**

or the same thing in 3D still with Maple™ 4.00b:

```
>with(plots):
>plot3d((BesselJ(1,Pi*sqrt(x^2+y^2))/(Pi*sqrt(x^2+y^2)))^2,x=-4..4,
y=-4..4,orientation=[30,70],axes=box,style=patchcontour,contours=40,
view=0..0.01,grid=[100,100]);
```

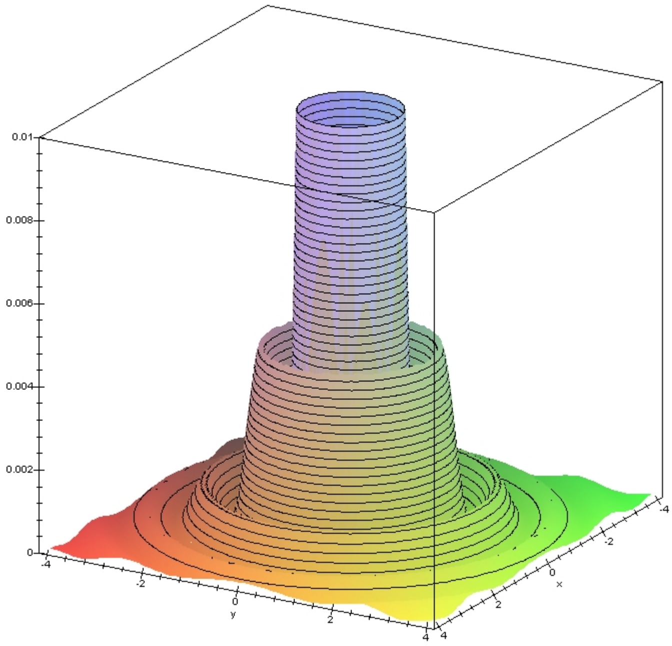


Figure 10.245: Fresnel circular aperture diffraction pattern 3D plot in Maple™ 4.00b

Remark

For information a Poisson's sport doesn't appear during a total solar eclipse because the Sun (actually a G2V-type star, aka yellow dwarf) is an extended, non-coherenct light source. Poisson's spot (or Arago spot) results from Fresnel diffraction of a coherent plane wavefront.

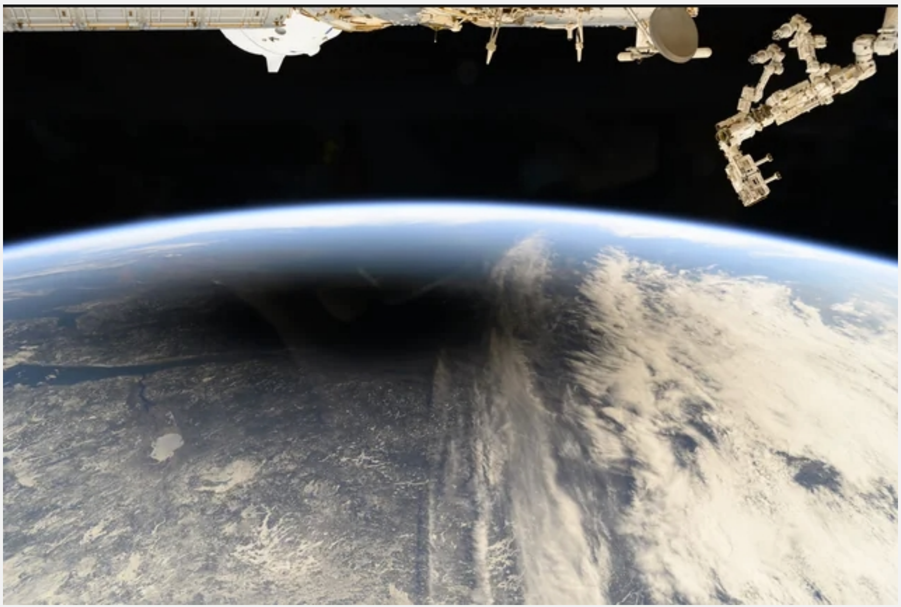


Figure 10.246: Photo of the total solar eclipse from 12024 (holocene calendar) as seen from the International Space Station

10.6.3.2.1 Optical resolution in the near field region for circular aperture

Notice that the destructive part of the pattern where $J_1(r)/r=0$ can be numerically evaluated to give $r\cong 3.83$ for the first ring.

For our circular aperture above this means the first zero occurs at:

$$\frac{kaq_1}{R}\cong 3.83 \tag{10.6.228}$$

or:

$$\frac{2\pi}{\lambda}\frac{aq_1}{R}\cong 3.83 \tag{10.6.229}$$

Therefore:

$$q_1\cong\frac{1.22R\lambda}{2a} \tag{10.6.230}$$

In our case a is the radius of the aperture and we can rewrite the expression using the diameter $D=2a$:

$$q_1\cong\frac{1.22R\lambda}{D} \tag{10.6.231}$$

The factor 1.22 is more precisely $1.21966989\ldots$

Light passing through any circular aperture is going to be diffracted in this manner and this sets the limit of resolution on an optical device such as a telescope. Say one is trying resolve two sources, we can say the limit of resolution is when the central spot of one Airy disk is on the zero of the other Airy pattern (say differently: the central maximum of the Airy pattern of one point emitter is directly overlapping with the first minimum of the Airy pattern of the other). This is known as the "Rayleigh criterion[22]" or "Rayleigh resolution criterion".

While it is possible to define other criteria, this is the most commonly used. See:

---

[22] There are other existing criterion in practice like the "Sparrow limit", the "Abbe Limit", the "FWHM", etc. Astronomical studies of equal brightness stars have shown that Sparrow's criterion is by far more useful.

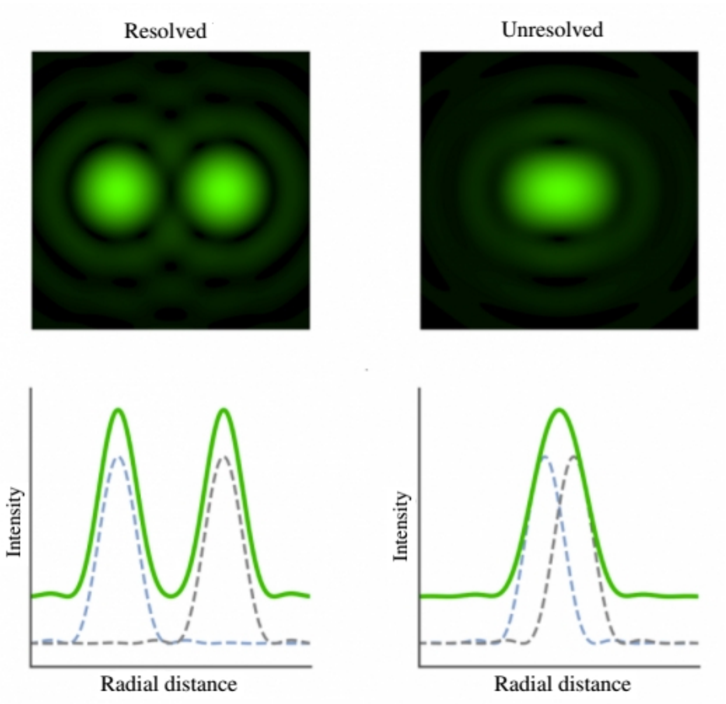


Figure 10.247: Overlap of PSFs (point spread function) and resolution

In the above figure the two sources can clearly be resolved (according to Rayleigh criterion). On the right the two sources are going to be difficult to resolve...

So we say that the limit of our resolution occurs when the distance Δq between two sources is:

$$\Delta q \cong \frac{1.22R\lambda}{D} \tag{10.6.232}$$

Or in the small angle limit $\Delta\theta = \Delta q/R$, we get the famous "spatial coherence":

$$\boxed{\Delta\theta \cong \frac{1.22\lambda}{D}} \tag{10.6.233}$$

By convention (as an approximate condition) we declare that two particles are too close together to resolve (identify as two separate particles rather than one) if their central diffraction peaks overlap. since the angle spanned by the central diffraction peak is $1.22\lambda/D$,this means that we say we can't resolve objects unless the angle between them (as measured from the microscope lens) is greater than $1.22\lambda/D$.

✍**Example:**

A positive lens having a 40 [mm] diameter is used to form the image of two stars on a CCD in a camera. If the stars are $1,000$ light-years from Earth, how far apart are they if they are just resolvable according to the Rayleigh criterion assuming $\lambda = 550\,[\text{nm}]$?

The solution is obviously given by:

$$\Delta\theta \cong \frac{1.22\lambda}{D} = \frac{1.22\,(550\cdot 10^{-9})}{40\cdot 10^{-3}} \cong 1.6775\cdot 10^{-5}\,[\text{rad}]$$

The stars' separation, L, is then:

$$L \cong R\Delta\theta \cong 1000\left(1.6775\cdot 10^{-5}\right)$$

and therefore $L \cong 0.0168$ light-year.

If the smallest resolvable separation between images is to be reduced (i.e., if the resolving power is to be increased), the wavelength, for instance, might be made smaller. Using ultraviolet rather than visible light in microscopy allows for the perception of finer detail (that's also why microchips are made with ultraviolet LASER, indeed it is the highest non-destructive light frequency that permits small angular separation of integrated transistors without being destructive for silicon wafers). The electron microscope utilizes equivalent wavelengths of about 10^{-4} to 10^{-5} that of light. This makes it possible to examine objects that would otherwise be completely obscured by diffraction effects in the visible spectrum. On the other hand, the resolving power of a telescope can be increased by increasing the diameter of the objective lens or mirror. Besides collecting more of the incident radiation, this will also result in a smaller Airy disk and therefore a sharper, brighter image.

So now in the case of a telescope the circular opening of the telescope generates a circular interference pattern.

Because of this interference pattern, when we make an image of a star, it does NOT focus to a perfect point. Rather, it focuses to a disk, and if we set our telescope for high magnification and examine the image carefully, we can see that there is a disk with faint rings around it - this is the interference pattern that is caused by the circular aperture of our telescope:

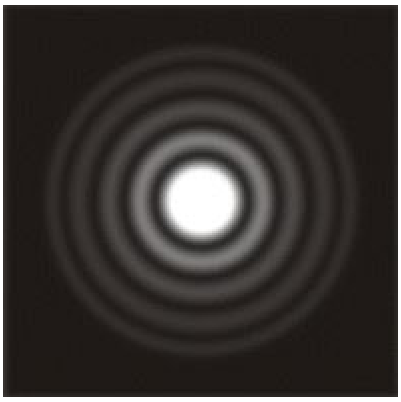


Figure 10.248: **Idealized picture of Airy disk**

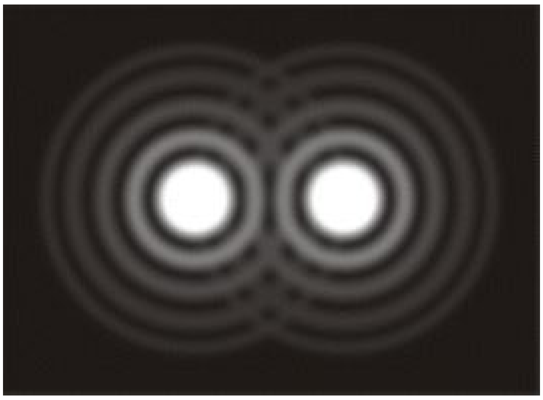


Figure 10.249: **When close stars appear as two Airy disks**

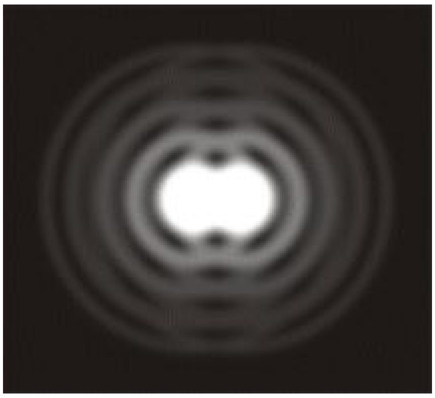


Figure 10.250: **Too close Airy disk that starts to make distinction difficult**

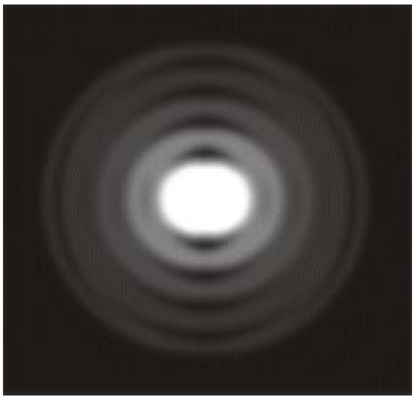


Figure 10.251: **When closer together than the radius of the Airy disk, we can no longer tell them apart**

The angular resolution may be converted into a spatial resolution, ΔL, by multiplication of the angle (in radians) with the distance to the object. For a microscope, that distance is close to the focal length f of the objective. For this case, the Rayleigh criterion reads:

$$\Delta L \cong 1.22 \frac{f\lambda}{D_{\text{scope}}} \qquad (10.6.234)$$

This is the size, in the imaging plane, of smallest object that the lens can resolve, and also the radius of the smallest spot to which a collimated beam of light can be focused.

With regard to telescopes, what is important in this equation is that the size of the Airy disk depends on the diameter of the objective ONLY, and as the diameter of the objective gets bigger, the Airy disk gets smaller.

This means as the diameter of our scope gets bigger, we can see smaller and smaller detail –or equivalently, we can split stars that are closer together. Conversely, this means that there is a limit to the detail we can see with our scope. That limit is dictated ENTIRELY by the diameter of the scope, and it is due to the diffraction effects caused by the scope opening.

For this reason, the radius of the Airy disk, as calculated above, is known as the "diffraction limit". Note then that when we can see the rings of the Airy Disk, this signals to us that we are operating at the limit of the scope's power to resolve detail.

The diffraction limit equation gives an answer in radians, whereas most dimensions in astronomy are given in degrees, minutes, or seconds of arc. Also because of various other factors, the actual limit of detail that an observer can see does not match exactly to the diffraction limit. Interestingly, a skilled observer can do better than the diffraction formula would suggest.

In 11867 (holocene calendar), William Rutter Dawes determined the practical limit on resolving power for a telescope, known as the "Dawes limit". Dawes expressed this as the closest that two stars could be together in the sky and still be seen as two stars. The Dawes Limit is 4.56 seconds of arc, divided by the telescope aperture in inches. We can multiply the Dawes limit by 25.4 to convert to the metric system (in [mm]), which gets us to 115.8, and then round to a number that is more convenient when doing the math in our head, and we get the resolving power of the scope, P_R in seconds of arc:

$$P_R = \frac{115.8}{D_0} \cong \frac{120}{D_0} \qquad (10.6.235)$$

In meter we can often found that later rounder to 11.6 such that:

$$P_R \cong \frac{11.6}{D_0} \cong \frac{12}{D_0} \qquad (10.6.236)$$

✍**Example:**

Hubble's mirror is $2.4\,[\text{m}] = 240\,[\text{cm}]$ across. Plugging that into the above relation, we see that Hubble's resolution is:

$$P_R \cong \frac{11.6}{240} = 0.05'' \qquad (10.6.237)$$

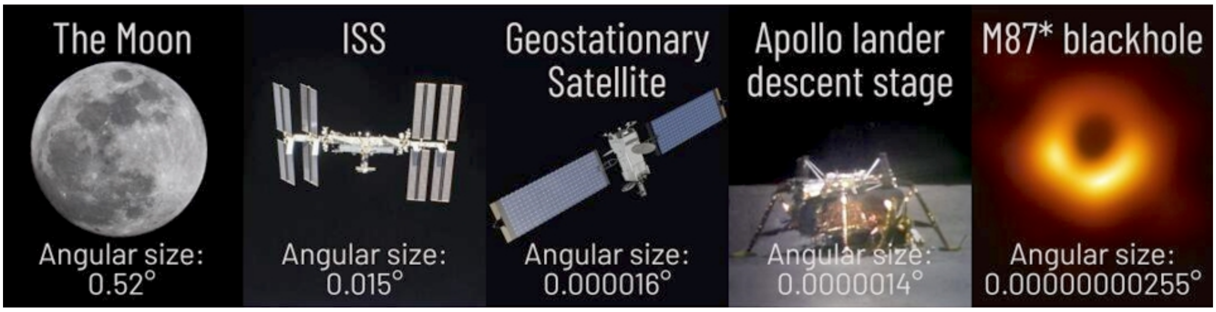
That's an incredibly small size; a human would have to be nearly 8,000 kilometres away to be 0.05 arcseconds in size!

Remark

With the example above it is quit easy to understand why the Hubble space telescope cannot resolve object size of the Apollo mission on the Moon as they're are not more much bigger than human size and especially that they are (at nearest point) 356,700 [km] away! In fact Hubble's space telescope resolution on the Moon is about 200 meters! In other words, even a football stadium on the Moon would look like a dot to Hubble (that's also point out the difference between "seeing" a candle on the Moon, and "resolve" it into a disc!).

That's a pretty big surprise to most people. They're used to seeing magnificent detail in Hubble images, stars in galaxies and wisps of gas in beautiful nebulae. But those objects are far, far larger than the Moon lander. Hubble's resolution is 0.1" no matter how far away an object is. Those wisps of gas appear to be finely resolved, but they're billions of kilometres across. That's a bit roomier than the lunar landers were.

An angle is a fraction of a circle. We can consider the angular size of an object to be equal to its actual (linear) size d, which in turn is roughly the fraction of a circle with a radius equal to the distance of the object, D (actually, d is the "chord" of an arc, and the arc is the fraction of the circumference, thus the cosine factor mentioned in other comments. But we can ignore that for small angles).



So: Fraction of circle (our angle) $= \alpha/360°$ (because there are 360 degrees in a circle).

Fraction of circle (size of object) $= d/(2\pi D)$ (because the circumference of the circle is 2π times the radius, and in this case we're using a circle of radius D).

So:

$$\frac{\alpha}{360^\circ}=\frac{d}{2\pi D} \tag{10.6.238}$$

Rearrange:

$$\frac{d}{D}\cdot\frac{360^\circ}{2\pi}=\alpha \tag{10.6.239}$$

But remember, we want our number to be in arcseconds, not degrees, so

Rearrange (60 arcminutes per degree, 60 arcseconds per arcminute):

$$\frac{d}{D}\cdot\frac{360^\circ\cdot 60'\cdot 60''}{2\pi}=\alpha \tag{10.6.240}$$

Calculating all that out, we get:

$$\frac{d}{D}\cdot 206,264.81''=\alpha \tag{10.6.241}$$

Round the constant up to $206,265$ and we're done.

So let's look at our lunar descent stage. It's 4 meters across, but $350,000,000$ meters away. That gives it an angular size of:

$$\frac{4}{350,000,000}\cdot 206,265\cong 0.002'' \tag{10.6.242}$$

Hey, wait a sec! This is far below Hubble's resolution! So the lander is way too small to be seen as anything more than a dot, even by Hubble.

Remark

At then end of the 120th century (holocene calendar), it seems that the KH-11 reconnaissance satellite series, first launched by the American National Reconnaissance Office, with their supposed 2.4 meter mirror (quite similar to Hubble!) observing in the visual (i.e. at a wavelength of 500 [nm]) have the best resolution power in the World with a diffraction limited resolution of around 0.05 arcsec, which from an orbital altitude of 250 [km] corresponds to a ground sample distance of 0.06 [m]...

10.6.3.3 Case of a circular obstacle

We can handle the case of diffraction by a circular obstacle quite easily now, using the result just obtained. For the field,

$$\begin{aligned}
f &\cong \frac{2\pi\rho_A e^{\mathrm{i}(kR-\omega t)}}{R}\int\limits_{a}^{+\infty}\rho J_0\left(\frac{kq\rho}{R}\right)\mathrm{d}\rho\\
&=\frac{2\pi\rho_A e^{\mathrm{i}(kR-\omega t)}}{R}\int\limits_{0}^{+\infty}\rho J_0\left(\frac{kq\rho}{R}\right)\mathrm{d}\rho-\frac{2\pi\rho_A e^{\mathrm{i}(kR-\omega t)}}{R}\int\limits_{0}^{a}\rho J_0\left(\frac{kq\rho}{R}\right)\mathrm{d}\rho\\
&=\rho_A e^{\mathrm{i}(kR-\omega t)}-\frac{\rho_A\pi a^2 e^{\mathrm{i}(kR-\omega t)}}{R}\frac{R}{kaq}2J_1\left(\frac{kaq}{R}\right)\\
&=\rho_A e^{\mathrm{i}(kR-\omega t)}-\frac{\rho_A\pi a^2 e^{\mathrm{i}(kR-\omega t)}}{R}\frac{2J_1\left(\dfrac{kaq}{R}\right)}{\dfrac{kaq}{R}}
\end{aligned}\tag{10.6.243}$$

Therefore the intensity is proportional to:

$$\begin{aligned}
I\propto\hat{f}^2&=\left(\rho_A-\frac{\rho_A\pi a^2}{R}\frac{2J_1\left(\dfrac{kaq}{R}\right)}{\dfrac{kaq}{R}}\right)^2\\
&=\rho_A^2\left[1-\frac{\pi a^2}{R}\frac{4J_1\left(\dfrac{kaq}{R}\right)}{\dfrac{kaq}{R}}+\left(\frac{\pi a^2}{R}\frac{2J_1\left(\dfrac{kaq}{R}\right)}{\dfrac{kaq}{R}}\right)^2\right]
\end{aligned}\tag{10.6.244}$$

This result has the curious property of not being zero inside the shadow of the obstacle (when $q\rightarrow 0$). To prove that fact let see first a special limit of the first order Bessel function.

Take the recurrence relation at $m=1$ and use the chain rule:

$$uJ_0(u)=\frac{\mathrm{d}}{\mathrm{d}u}\left[uJ_1(u)\right]=J_1(u)+u\frac{\mathrm{d}J_1}{\mathrm{d}u}(u)\tag{10.6.245}$$

Therefore:

$$J_0(u)=\frac{\mathrm{d}J_1}{\mathrm{d}u}(u)+\frac{J_1(u)}{u}\tag{10.6.246}$$

Note that $J_0(0)=1$ and $J_1(0)=0$:

$$1=\frac{\mathrm{d}J_1}{\mathrm{d}u}(0)+\lim_{u\rightarrow 0}\frac{J_1(u)}{u}=\frac{\mathrm{d}J_1}{\mathrm{d}u}(0)+\lim_{u\rightarrow 0}\frac{\dfrac{\mathrm{d}J_1}{\mathrm{d}u}(u)}{1}=2\frac{\mathrm{d}J_1}{\mathrm{d}u}(0)\tag{10.6.247}$$

or:

$$\lim_{u\rightarrow 0}\frac{J_1(u)}{u}=\frac{\mathrm{d}J_1}{\mathrm{d}u}(0)=\frac{1}{2}\tag{10.6.248}$$

In fact, there's a sharp peak exactly in the center, with peak intensity:

$$\begin{aligned}I_{\text{center},q=0} &\propto \lim_{q\rightarrow 0}\left[1-\frac{\pi a^2}{R}\frac{4J_1\left(\dfrac{kaq}{R}\right)}{\dfrac{kaq}{R}}+\left(\frac{\pi a^2}{R}\frac{2J_1\left(\dfrac{kaq}{R}\right)}{\dfrac{kaq}{R}}\right)^2\right]\\ &=1-\frac{2\pi a^2}{R}+\frac{\pi^2 a^4}{R^2}>0\end{aligned} \tag{10.6.249}$$

as $a/R \ll 1$.

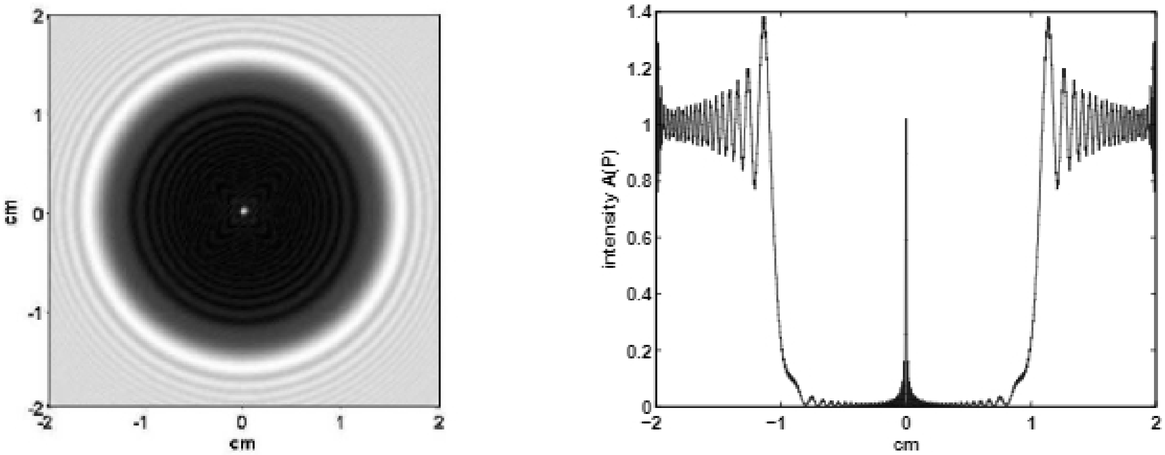


Figure 10.252: **Spot of Poisson-Arago Intensity distribution behind a coin ($a = 1$ [cm]) on a detector placed at $R = 5$ [m].**

10.6.4 Light polarization

It was not before the 119th century (holocene calendar) that it was discovered the polarization of light (we will immediately explain what it is). However, at the time of Newton, we already knew a phenomenon due to polarization: the existence of crystals named "birefringent crystals" (as the "Iceland spar") which have the property of refracting a single beam into two separate beams[23].

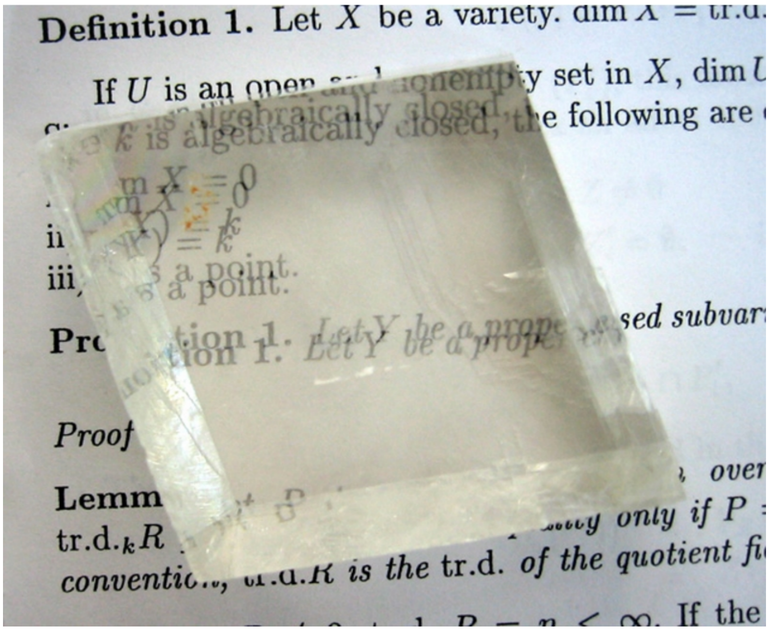
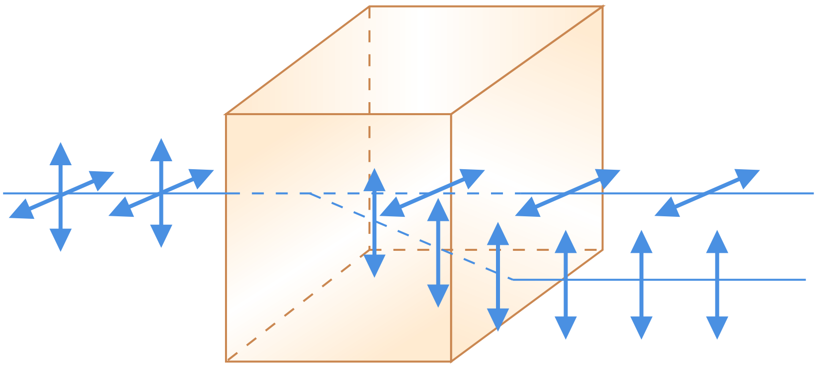


Figure 10.253: **Iceland spar (crystallized calcite)**

Now we know that the two rays refracted by such a crystal are polarized:



To understand the "polarization of light" let us come back to the case of a wave propagating on a string (see section Wave Mechanics page 3475). Such a wave can do it in a vertical plane (right) as well as in a horizontal plane (left) or in all intermediate levels:

---

[23] In the mid-12026 (holocene calendar) a peer-review scientific paper has been published about a potential observation of quantum vacuum birefringence ("vacuum crystallization") at the vicinity of the surface of a Magnetar-type star in accordance with the predictions of Quantum Field Theory in the context of ultra-intense magnetic fields (see [515]).

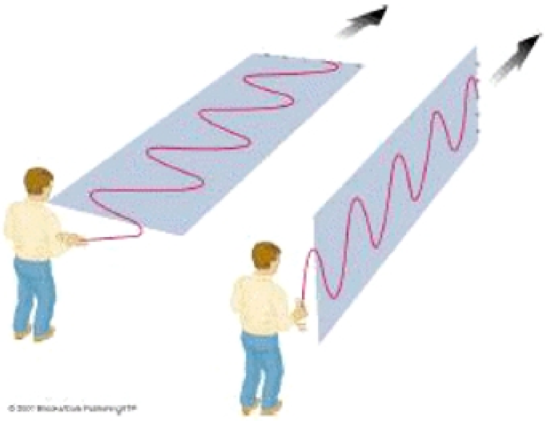


Figure 10.254: **Pictorial representation of the light polarization concept with a string (source: ?)**

In both cases, we say that the wave is "linearly polarized", which means that the oscillations are only and always in the same plane, named "polarization plane". Such a wave can pass through a vertical slot if it is polarized vertically, a horizontally polarized wave can not.

Let us recall that we have seen in the section of Electrodynamics that for electromagnetic waves, the electric $\vec{E}$ field oscillates (at least for the standard solution of Maxwell's equations) and is orthogonal to the direction of propagation.

The electric field vector $\vec{E}$ of a wave can be decomposed into two perpendicular components to each other, $(\vec{E}_x,\vec{E}_y)$ if the wave propagates in the z direction and each carrying half of the intensity of the wave. These two components change at any time when $\vec{E}$ varies. The result at any time is a total horizontal field and a total vertical field.

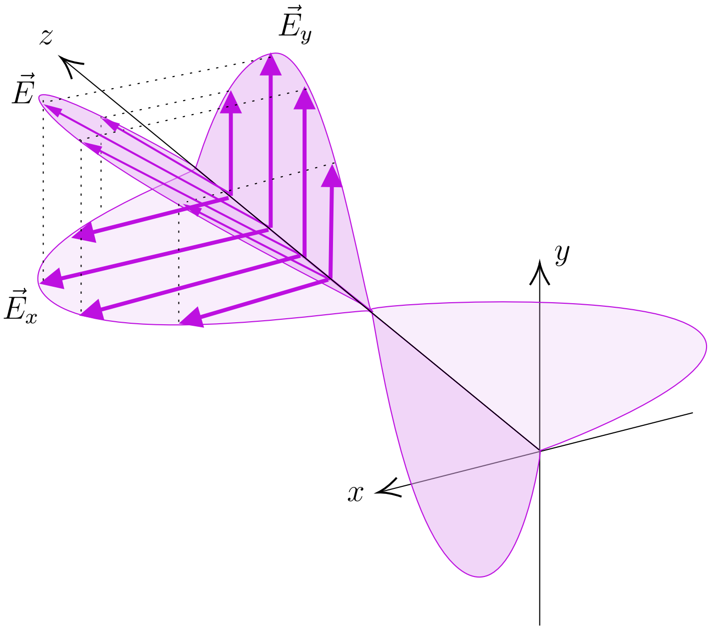


Figure 10.255: **Illustration of the decomposition of the electric field of a propagating wave (source: ?)**

If $\vec{E}$ revolves around the propagation direction with its end describing circle, then we say that the wave is "circularly polarized":

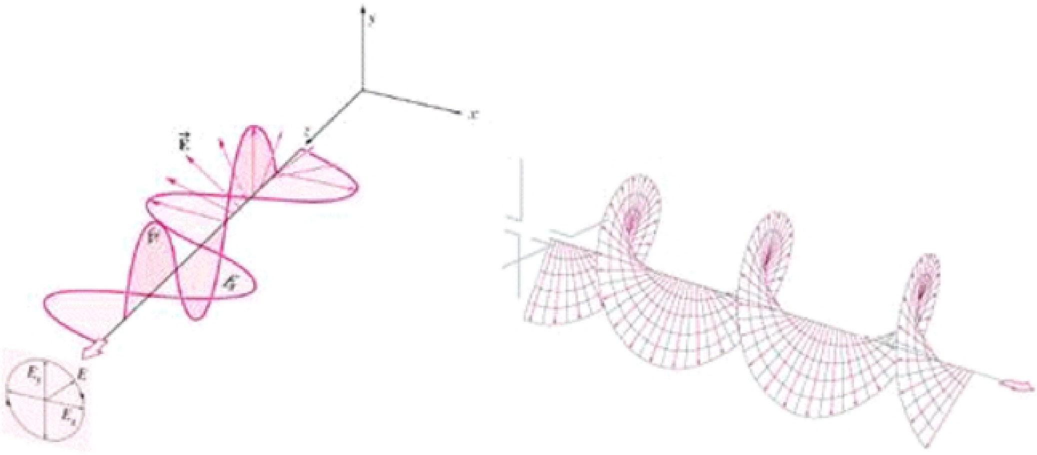


Figure 10.256: **Pictorial representation of a circular polarized wave (source: ?)**

Then $\vec{E}$ remains of constant modulus but turns while moving, making one complete turn for each distance equal to one wavelength λ.

Remark

The light is not necessarily polarized! Every atom emits a wave train that lasts less than one hundred-millionth of a second (these wave trains are perfectly explained by the spread of a the free particle in quantum physics with the Fourier transforms as proved in the section of Wave Quantum Physics) and all these waves have no phase correlation or orientation. The resultant field in a given position in space, is the geometric sum of all these wave trains: it is constantly changing.

Thus, the natural light is a random and quickly variable mixture of linearly polarized waves in all directions. Looking to the source we observe a resulting field $\vec{E}$ oscillating in a certain direction for a fraction of time and then suddenly jumps to a new random direction while remaining perpendicular to the direction of propagation:

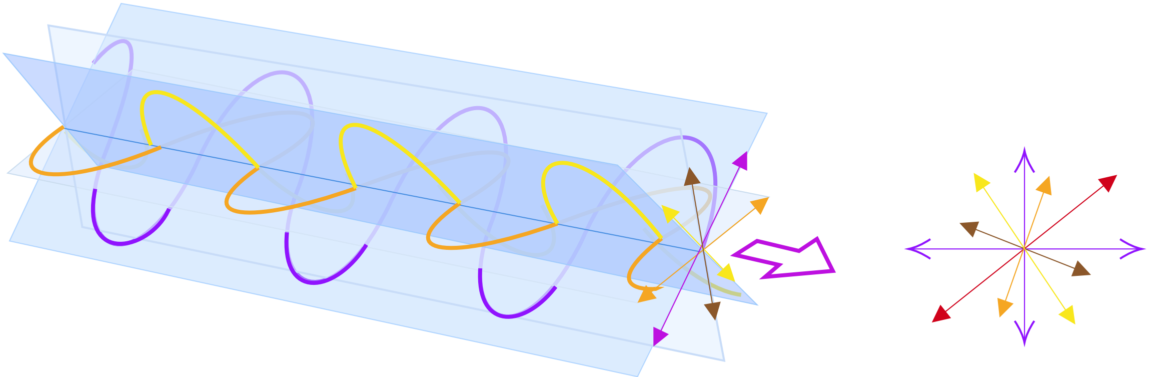
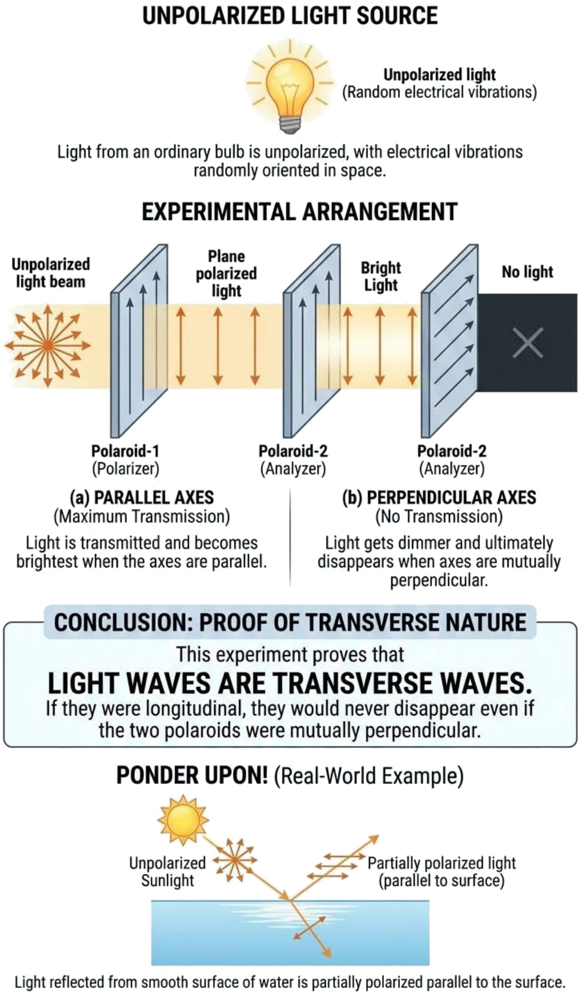


Figure 10.257: **Pictorial representation of "natural" light wave (source: ?)**


UNPOLARIZED LIGHT SOURCE
Unpolarized light
(Random electrical vibrations)
Light from an ordinary bulb is unpolarized, with electrical vibrations randomly oriented in space.
EXPERIMENTAL ARRANGEMENT
Unpolarized light beam
Plane polarized light
Bright Light
No light
Polaroid-1
(Polarizer)
Polaroid-2
(Analyzer)
Polaroid-2
(Analyzer)
(a) PARALLEL AXES
(Maximum Transmission)
Light is transmitted and becomes brightest when the axes are parallel.
(b) PERPENDICULAR AXES
(No Transmission)
Light gets dimmer and ultimately disappears when axes are mutually perpendicular.
CONCLUSION: PROOF OF TRANSVERSE NATURE
This experiment proves that
LIGHT WAVES ARE TRANSVERSE WAVES.
If they were longitudinal, they would never disappear even if the two polaroids were mutually perpendicular.
PONDER UPON! (Real-World Example)
Unpolarized Sunlight
Partially polarized light
(parallel to surface)
Light reflected from smooth surface of water is partially polarized parallel to the surface.

This introduction done, let us move now to something a bit more formal:

We therefore saw in the section Electrodynamics that a hypothetical progressive monochromatic plane wave (even if physically it cannot exist ...) propagating in the vacuum consisted of an electric field $\vec{E}$ and a magnetic field $\vec{E}$ and was characterized by its pulsation ω (or corresponding wavelength λ), its electric field amplitude $\hat{E}$ and magnetic field amplitude $\hat{B}$ and its propagation direction given by a unit vector $\vec{u}_x$, $\vec{u}_y$, $\vec{u}_z$ at choice depending on the orientation of the selected reference frame.

We have also proved that these waves have remarkable structural properties, in particular:

- $\vec{E}$ and $\vec{B}$ are transverse, that is to say, their direction is at any point and at any time orthogonal to the direction of propagation (Malus theorem). This, allow us to define a wave plane, plane generated by the two directions of $\vec{E}$ and $\vec{B}$.
- The norms of these two vectors are connected by the relations $\hat{B}=\hat{E}/c$ where c is the speed of light in vacuum (this is this huge ratio between the magnetic field and electric field of an electromagnetic wave that makes developments presented further below are done preferably compared with the electric field component of the wave).
- Finally, these two vectors are orthogonal to each other, and the trihedron $(\vec{E},\vec{B},\vec{u}_i)$ is an

orthogonal trihedron .

These three properties can be summarized as we have prove it by the relation:

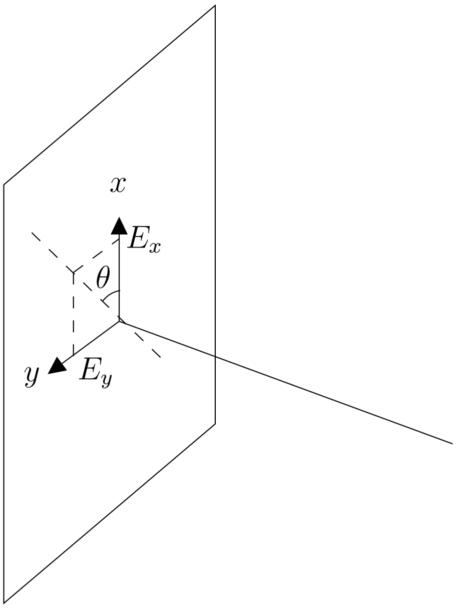
$$\vec{B}=\frac{\vec{u}_z\times\vec{E}}{c} \tag{10.6.250}$$

where we have chose the reference frame such as the wave propagates in the direction $\vec{u}_z$. Furthermore, we proved that the electric field is a trigonometric wave function given by:

$$\|\vec{E}\|=\hat{E}\cos(\omega t-kz+\varphi_1) \qquad \text{or} \qquad \|\vec{E}\|=\hat{E}\sin(\omega t-kz+\varphi_2) \tag{10.6.251}$$

Let us now position ourselves in a base (x,y,z). The most general expression of the electric field of a progressive monochromatic plane wave progressing in the $\vec{u}_z$ direction can be decomposed into two components:

$$\vec{E}=\begin{pmatrix}E_x\\E_y\\E_z\end{pmatrix}=\begin{pmatrix}\hat{E}\cos(\theta)\cos(\omega t-kz+\varphi_x)\\\hat{E}\sin(\theta)\cos(\omega t-kz+\varphi_y)\\0\end{pmatrix}=\begin{pmatrix}\hat{E}_{0x}\cos(\omega t-kz+\varphi_x)\\\hat{E}_{0y}\cos(\omega t-kz+\varphi_y)\\0\end{pmatrix} \tag{10.6.252}$$



The norm of the field is therefore given by:

$$\|\vec{E}(t)\|=\sqrt{\hat{E}\cos^2(\theta)\cos^2(\omega t-kz+\varphi_x)+\hat{E}^2\sin^2(\theta)\cos^2(\omega t-kz+\varphi_y)} \tag{10.6.253}$$

If $\varphi_x=\varphi_y$ (which is most often the case) we have then:

$$\|\vec{E}(t)\|=\hat{E}\cos(\omega t-kz+\varphi)\leq\hat{E} \tag{10.6.254}$$

By choosing another beginning of time, we are lead to write:

$$\vec{E}=\begin{pmatrix}\hat{E}_{0x}\cos(\omega t-kz)\\\hat{E}_{0y}\cos(\omega t-kz-\varphi)\\0\end{pmatrix} \tag{10.6.255}$$

with:

$$\varphi=\varphi_x-\varphi_y \tag{10.6.256}$$

Remark

The choice of writing $-\varphi$ instead of $+\varphi$ will be useful later for the use of remarkable trigonometric relations and will allow us to find the equation of an ellipse (...patience the proof is not far anymore).

Using the phasors (see section Wave Mechanics page 3510) the last relations can be reduced to:

$$\vec{E}=\Re\begin{pmatrix}\hat{E}_{0x}e^{\mathrm{i}(\omega t-kz)}\\ \hat{E}_{0y}e^{\mathrm{i}(\omega t-kz-\varphi)}\\ 0\end{pmatrix}=\Re\begin{pmatrix}\hat{E}\cos(\theta)e^{\mathrm{i}(\omega t-kz)}\\ \hat{E}\sin(\theta)e^{\mathrm{i}(\omega t-kz-\varphi)}\\ 0\end{pmatrix} \quad (10.6.257)$$

However, to describe this field, and therefore the whole wave, it is convenient to put ourselves in the plane $z=0$ and describe the evolution of the vector in $\vec{E}$ in this plane. That's what we'll do next. This is equivalent as to choose a coordinate origin following the z-axis. In this case, we can write:

$$\vec{E}=\Re\begin{pmatrix}\hat{E}_{0x}e^{\mathrm{i}(\omega t)}\\ \hat{E}_{0y}e^{\mathrm{i}(\omega t-\varphi)}\\ 0\end{pmatrix}=\Re\begin{pmatrix}\hat{E}\cos(\theta)e^{\mathrm{i}(\omega t)}\\ \hat{E}\sin(\theta)e^{\mathrm{i}\omega t}e^{-\varphi}\\ 0\end{pmatrix}=\hat{E}\Re\begin{pmatrix}\cos(\theta)\\ \sin(\theta)e^{-\varphi}\\ 0\end{pmatrix}e^{\mathrm{i}\omega t} \quad (10.6.258)$$

But the most general polarization is described by a complex vector normalized to unity in a two-dimensional space of components:

$$\lambda=\cos(\theta)e^{\mathrm{i}\varphi_x} \qquad \mu=\sin(\theta)e^{\mathrm{i}\varphi_y} \quad (10.6.259)$$

with:

$$|\lambda|^2+|\mu|^2=1 \quad (10.6.260)$$

such that:

$$\vec{E}=\hat{E}\cdot\Re\begin{pmatrix}\cos(\theta)e^{\mathrm{i}\varphi_x}\\ \sin(\theta)e^{\mathrm{i}\varphi_y}\\ 0\end{pmatrix}e^{\mathrm{i}\omega t} \quad (10.6.261)$$

10.6.4.1 Linear polarization

Definition (#10.6.498): We say that a wave is "linearly polarized" when $\varphi=0$ or $\varphi=\pi$.

In the first case ($\varphi=0\Rightarrow\varphi_x=\varphi_y$), we have:

$$\vec{E}=\Re\begin{pmatrix}\hat{E}_{0x}e^{\mathrm{i}\omega t}\\ \hat{E}_{0y}e^{\mathrm{i}\omega t}\end{pmatrix} \quad (10.6.262)$$

Therefore, we have $\hat{E}_{0x}$,$\vec{E}_{0y}$ which have respectively values between:

$$[-\hat{E}_{0x},+\hat{E}_{0x}],[-\hat{E}_{0y},+\hat{E}_{0y}] \quad (10.6.263)$$

With respect to a diagram that we will see below it should be taken into account that when a component is positive the other is also positive, and vice versa.

We then have at every instant:

$$\left.\begin{aligned} E_x &= \hat{E}_{0x}\cos(\omega t) \\ E_y &= \hat{E}_{0y}\cos(\omega t) \end{aligned}\right\} : \frac{E_x}{\hat{E}_{0x}} = \frac{E_y}{\hat{E}_{0y}} \tag{10.6.264}$$

which means that the field keeps a fixed direction. Hence the fact that we speak of linearly polarized wave.

If $\varphi = \pi$ we then have:

$$\vec{E} = \begin{pmatrix} \hat{E}_{0x}\cos(\omega t - kz) \\ \hat{E}_{0y}\cos(\omega t - kz - \pi) \\ 0 \end{pmatrix} = \begin{pmatrix} \hat{E}_{0x}\cos(\omega t - kz) \\ -\hat{E}_{0y}\cos(\omega t - kz) \\ 0 \end{pmatrix} \tag{10.6.265}$$

and that can therefore be written:

$$\vec{E} = \Re\begin{pmatrix} \hat{E}_{0x}e^{\mathrm{i}\omega t} \\ -\hat{E}_{0y}e^{\mathrm{i}\omega t} \end{pmatrix} \tag{10.6.266}$$

Therefore, we have $\hat{E}_{0x},\vec{E}_{0y}$ which have respectively values between:

$$[-\hat{E}_{0x}, +\hat{E}_{0x}], [-\hat{E}_{0y}, +\hat{E}_{0y}] \tag{10.6.267}$$

We then have at every instant:

$$\left.\begin{aligned} E_x &= \hat{E}_{0x}\cos(\omega t) \\ E_y &= -\hat{E}_{0y}\cos(\omega t) \end{aligned}\right\} : \frac{E_x}{\hat{E}_{0x}} = -\frac{E_y}{\hat{E}_{0y}} \tag{10.6.268}$$

which means that the field keeps a fixed direction. Hence the fact that we speak of linearly polarized wave.

10.6.4.2 Elliptical polarization

If φ is arbitrary, and placing ourselves at $h = 0$, we always have by starting from:

$$\vec{E} = \begin{pmatrix} \hat{E}_{0x}\cos(\omega t - kz) \\ \hat{E}_{0y}\cos(\omega t - kz - \varphi) \\ 0 \end{pmatrix} \tag{10.6.269}$$

The first relation:

$$\frac{E_x}{\hat{E}_{0x}} = \cos(\omega t) \tag{10.6.270}$$

as well as:

$$\frac{E_y}{\hat{E}_{0y}} = \cos(\omega t)\cos(\varphi) + \sin(\omega t)\sin(\varphi) \tag{10.6.271}$$

hence:

$$\sin(\omega t)\sin(\varphi)=\frac{E_y}{\hat{E}_{0y}}-\cos(\omega t)\cos(\varphi)=\frac{E_y}{\hat{E}_{0y}}-\frac{E_x}{\hat{E}_{0x}}\cos(\varphi) \tag{10.6.272}$$

Moreover, we can write:

$$\frac{E_x}{\hat{E}_{0x}}=\cos(\omega t)\Leftrightarrow\frac{E_x}{\hat{E}_{0x}}\sin(\varphi)=\cos(\omega t)\sin(\varphi) \tag{10.6.273}$$

In squaring the two previous relations:

$$\begin{gathered}\sin^2(\omega t)\sin^2(\varphi)=\left(\frac{E_y}{\hat{E}_{0y}}\right)^2-2\frac{E_y}{\hat{E}_{0y}}\frac{E_x}{\hat{E}_{0x}}\cos(\varphi)+\left(\frac{E_x}{\hat{E}_{0x}}\right)^2\cos^2(\varphi)\\ \left(\frac{E_x}{\hat{E}_{0x}}\right)^2\sin^2(\varphi)=\cos^2(\omega t)\sin^2(\varphi)\end{gathered} \tag{10.6.274}$$

and by summing, we eliminate the time and get:

$$\sin^2(\varphi)=\left(\frac{E_y}{\hat{E}_{0y}}\right)^2-2\frac{E_y}{\hat{E}_{0y}}\frac{E_x}{\hat{E}_{0x}}\cos(\varphi)+\left(\frac{E_x}{\hat{E}_{0x}}\right)^2 \tag{10.6.275}$$

$$\sin^2(\varphi)=\left(\frac{E_y}{\hat{E}_{0y}}\right)^2-2\frac{E_y}{\hat{E}_{0y}}\frac{E_x}{\hat{E}_{0x}}\cos(\varphi)+\left(\frac{E_x}{\hat{E}_{0x}}\right)^2 \tag{10.6.276}$$

We notice that if $\varphi=0$, $\varphi=\pi$ we fall back on:

$$\left(\frac{E_y}{\hat{E}_{0y}}\right)^2-2\frac{E_y}{\hat{E}_{0y}}\frac{E_x}{\hat{E}_{0x}}+\left(\frac{E_x}{\hat{E}_{0x}}\right)^2=\left(\frac{E_y}{\hat{E}_{0y}}-\frac{E_x}{\hat{E}_{0x}}\right)^2=0\Rightarrow\frac{E_y}{\hat{E}_{0y}}=\frac{E_x}{\hat{E}_{0x}} \tag{10.6.277}$$

By the way, in the prior previous relation we recognize the equation of an ellipse (see section Analytical Geometry page 2768):

$$y^2+x^2+(-2\cos(\varphi))xy+(-\sin^2(\varphi))=0 \tag{10.6.278}$$

similar in all respects to the general equation of the conical we have proved in the section of Analytical Geometry that was for recall:

$$\alpha y^2+\beta x^2+\gamma xy+\delta x+\varepsilon y+\xi=0 \tag{10.6.279}$$

In this case, the end of the vector $\vec{E}$ thus describes an ellipse and we talk therefore naturally of "elliptical polarization".

According to the value of φ, the ellipse may be traversed in one direction or the other. To determine this direction, let us derivate the expression of the field:

$$\vec{E}=\begin{pmatrix}\hat{E}_{0x}\cos(\omega t-kz)\\ \hat{E}_{0y}\cos(\omega t-kz-\varphi)\\ 0\end{pmatrix} \tag{10.6.280}$$

with respect to time and let us put ourselves at $t=0$ still in the same wave plane in $z=0$:

$$\mathrm{d}_t\vec{E}=\begin{pmatrix}-\hat{E}_{0x}\omega\sin(\omega t-kz)\\ \hat{E}_{0y}\omega\sin(\omega t-kz-\varphi)\\ 0\end{pmatrix}=\begin{pmatrix}0\\ -\hat{E}_{0y}\omega\sin(\varphi)\\ 0\end{pmatrix}_{t=0} \qquad (10.6.281)$$

Therefore:

- If $0<\varphi<\pi/2$ the ellipse is travelled in the forward direction (counter clock wise) as shown in the figure below. We say then that the polarization is "elliptical direct left".
- If $\pi/2<\varphi<\pi$ the ellipse is travelled also in the forward direction (counter clock wise) as shown in the figure below. We say then that the polarization is "elliptical direct right".
- If $\pi<\varphi<3\pi/2$ the ellipse is travelled clockwise. We say then that the polarization is "elliptical indirect right".
- If $3\pi/2<\varphi<2\pi$ the ellipse is travelled clockwise as shown in the figure below. We say when the polarization is "elliptical indirect left".

10.6.4.3 Circular polarization

If:

$$\varphi=\frac{\pi}{2}\qquad\varphi=\frac{3\pi}{2} \qquad (10.6.282)$$

and:

$$\hat{E}_{0x}=\hat{E}_{0y}=E_0 \qquad (10.6.283)$$

we have then the equation of the ellipse which reduces to:

$$E_x^2+E_y^2=E_{0x,y}^2 \qquad (10.6.284)$$

which is the equation of a circle of radius $E_{0x,y}$, the sign being always given by the sign of the sine:

- If $\varphi=\pi/2$ it is a "left circular polarization"
- If $\varphi=3\pi/2$it is a "right circular polarization

....see figure below for a visual summary.

10.6.4.4 Natural polarization

We can consider the emission of a source as a succession of progressive theoretical monochromatic plane waves whose expression will be:

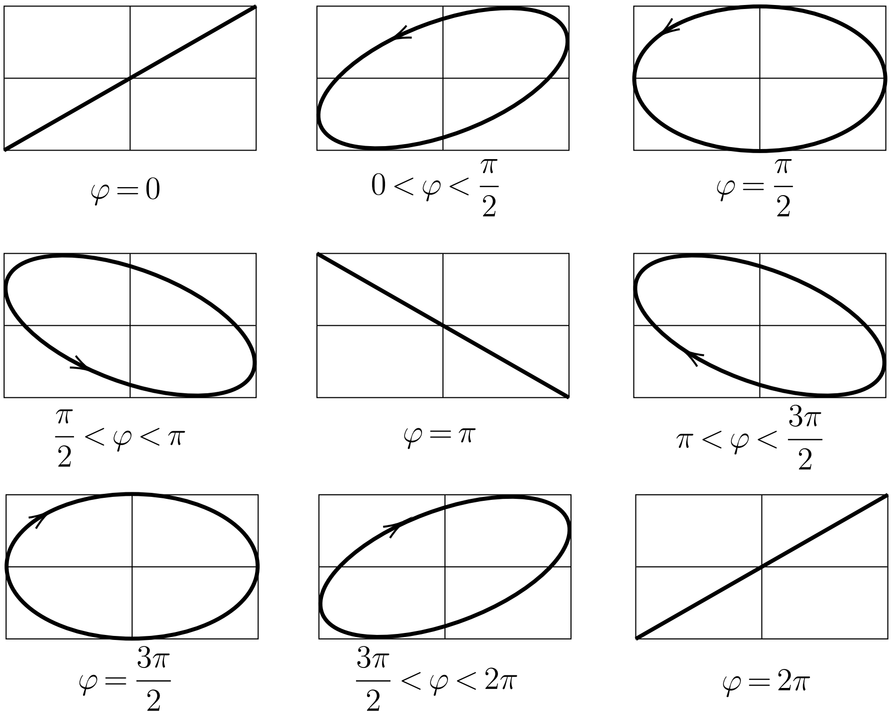
$$\vec{E}=\begin{pmatrix} \hat{E}_{0x}\cos(\omega t-kz) \\ \hat{E}_{0y}\cos(\omega t-kz-\varphi_l) \\ 0 \end{pmatrix} \tag{10.6.285}$$

These wave trains are in a particular polarization state. However, this states varies randomly from one wave train to another, and this in a very short time relative to the integration time of detectors. They therefore will see no particular bias and the field $\vec{E}$ will have no particular direction.

We speak therefore of "unpolarized light". If we overlay this light to a polarized wave, we get what we name a "partial polarization".

Finally, we can summarize what we have seen so far in the following figure where we have:

- The linear polarization: $\varphi=0,\varphi=\pi,\varphi=2\pi$
- The linear partial polarization (not represented below)
- The direct elliptical polarization left $0<\varphi<\pi/2$ or right $\pi/2<\varphi<\pi$
- The indirect elliptical polarization left $\pi<\varphi<3\pi/2$ or $3\pi/2<\varphi<2\pi$
- The partial elliptical polarization (not represented below)
- The circular polarization left $\varphi=\pi/2$ or right $\varphi=3\pi/2$
- The partial circular polarization (not represented below)


$\varphi=0$ $0<\varphi<\dfrac{\pi}{2}$ $\varphi=\dfrac{\pi}{2}$

$\dfrac{\pi}{2}<\varphi<\pi$ $\varphi=\pi$ $\pi<\varphi<\dfrac{3\pi}{2}$

$\varphi=\dfrac{3\pi}{2}$ $\dfrac{3\pi}{2}<\varphi<2\pi$ $\varphi=2\pi$

Figure 10.258: **Representations of different polarizations**

We can represent this with an animated plot in Maple™ 4.00b (we have not put the *.gif below as we don't wanted to overload the PDF too much...) using the following commands:

```
> with (plots):
> Ex:=1;Ey:=1;phi:=Pi/4;k:=1;omega:=1;
> animate3d([x,a*Ex*cos(omega*t-k*x),a*Ey*cos(omega*t-k*x-phi)],
a=0..1,x=-10..10,t=0..2*Pi,frames=15,grid=[35,35],style=patchnogrid,axes=boxed);
```

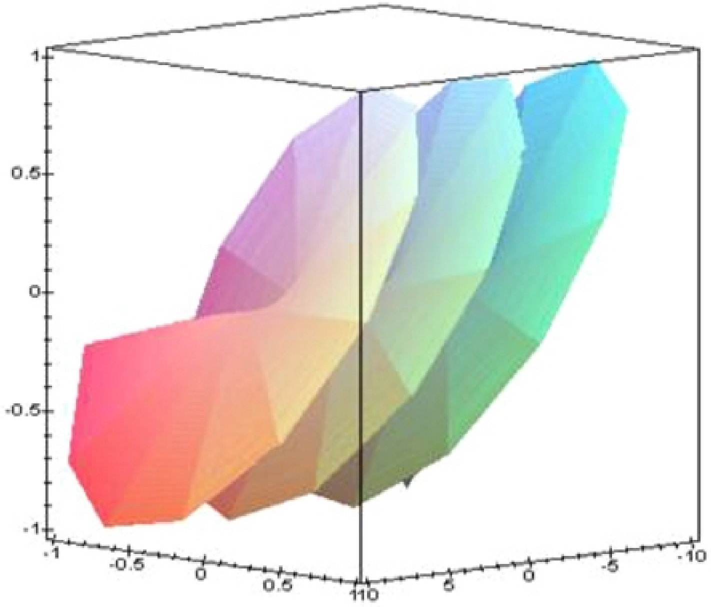


Figure 10.259: **Animation of a polarized wave in Maple™ 4.00b**

It is of course possible to change the parameters. For example, $\phi=\pi/2$ gives circular polarization, $\pi=\pi$ gives a linear polarization as we have prove above.

We can also put the figure of Wikipedia which also summarizes very well the subject:

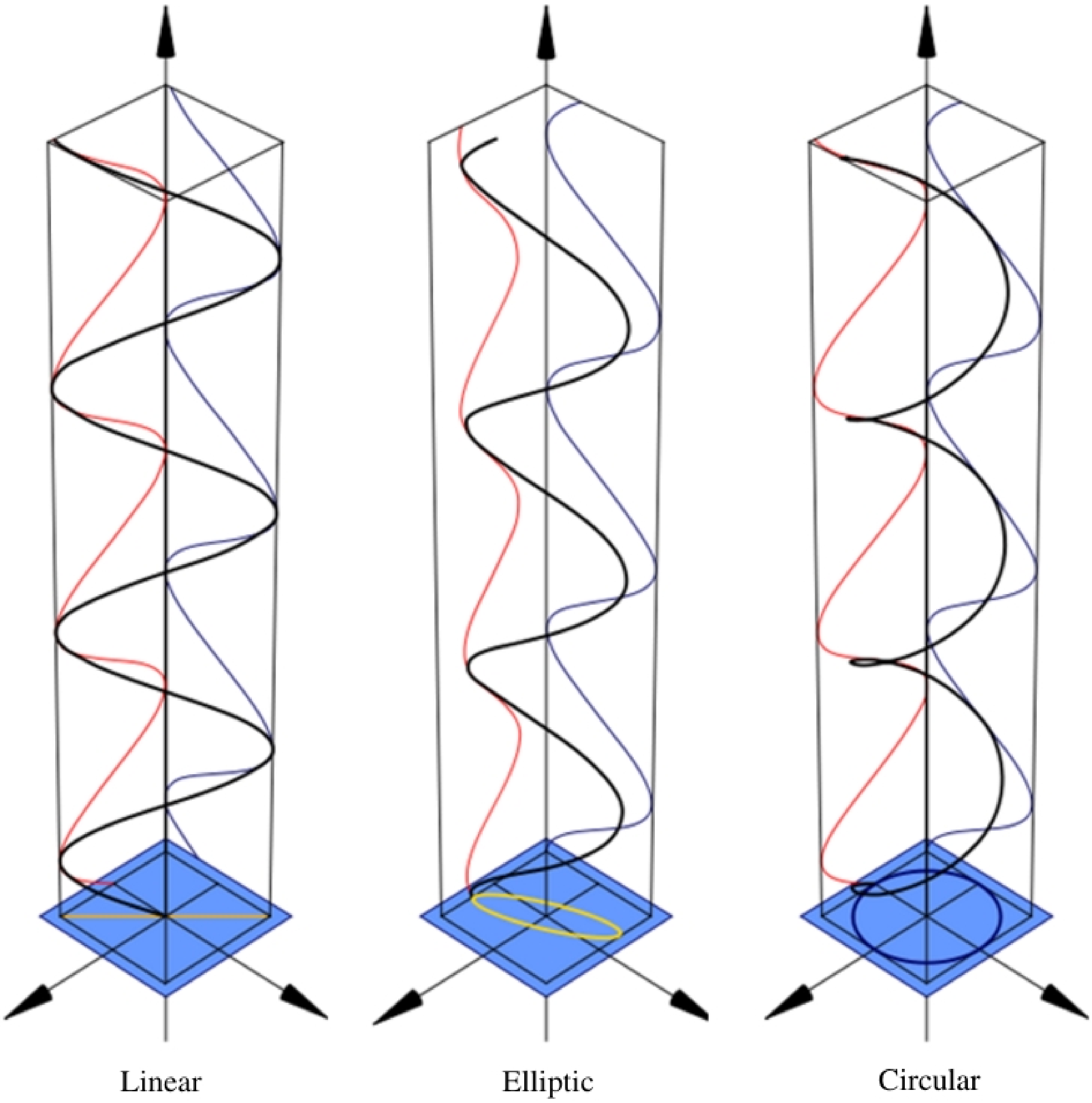


Figure 10.260: **Visual polarized wave principle summary as proposed by Wikipedia**

10.6.4.5 Malus' law

To polarize light, the physicist will use polarisers. We will not enter (because it is not part of the field of Wave Optics) in the details of atomic or molecular properties of matter that are the cause of the polarization of the transmitted light.

For our purposes, we will restrict ourselves to a polariser that polarizes incident light linearly along the x axis (the component E_y being therefore zero). We have therefore:

$$\vec{E}=\begin{pmatrix}E_x\\E_y\\E_z\end{pmatrix}=\Re\begin{pmatrix}\hat{E}\cos(\theta)e^{\mathrm{i}(\omega t-kz)}\\ \hat{E}\sin(\theta)e^{\mathrm{i}(\omega t-kz-\varphi)}\\ 0\end{pmatrix}=\Re\begin{pmatrix}\hat{E}\cos(\theta)e^{\mathrm{i}(\omega t-kz)}\\ 0\\ 0\end{pmatrix} \tag{10.6.286}$$

But, we have proved in the section of Electrodynamics during our study of Maxwell equations that:

$$I=\bar{S}=\frac{\hat{E}\hat{B}}{2\mu_0}=\frac{\hat{E}^2}{2\mu_0 c} \tag{10.6.287}$$

Therefore it comes for the maximum intensity (such that $e^{\mathrm{i}(\omega t-kz)}=1$):

$$\boxed{I=\frac{\hat{E}^2\cos^2(\theta)}{2\mu_0 c}=\frac{\hat{E}}{2\mu_0 c}\cos^2(\theta)=I_0\cos^2(\theta)} \tag{10.6.288}$$

relation that is the famous "Malus' law".

Here is a table with some famous values:

Angle θ	**Malus Law** $\cos^2(\theta)$ (I/I_0)
$0°$	100%
$22.5°$	$\sim 85.35\%$
$45°$	50%
$67.5°$	$\sim 14.64\%$
$95°$	0%

Some people are surprised in the above table that as $22.5°$ is half-way between $0°$ and $45°$ we have 83% rather than 75%.

We can rotate the second polarizer by intermediate angles to get more data points, and we can then rotate it the other way to get a set of data points for negative angles. Then you could plot a curve like this:

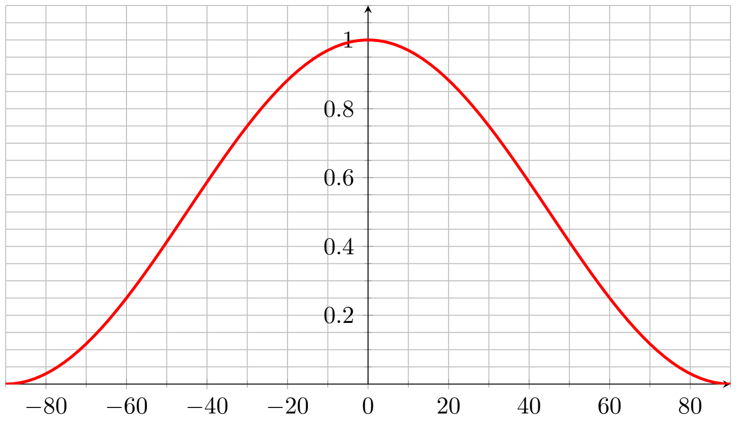


Figure 10.261: Malus' law plot for various angles

It's not a straight line, it's a cosine-like curve! As to why, it's because a polarizing filter isn't just a filter! It's simple physics. A polarizing filter doesn't just filter, it also rotates the light passing through it. When you have two polarizers at $90°$ to one another, none of the light gets through. But when you put a third polarizer between them at $45°$ to the first, it lets half the light through and rotates it. Then the third polarizer is at $45°$ to the second, and that lets half the light through again, and rotates it again.

To investigate quantitatively the polarization, we will use a polariser / analyser set a described below:

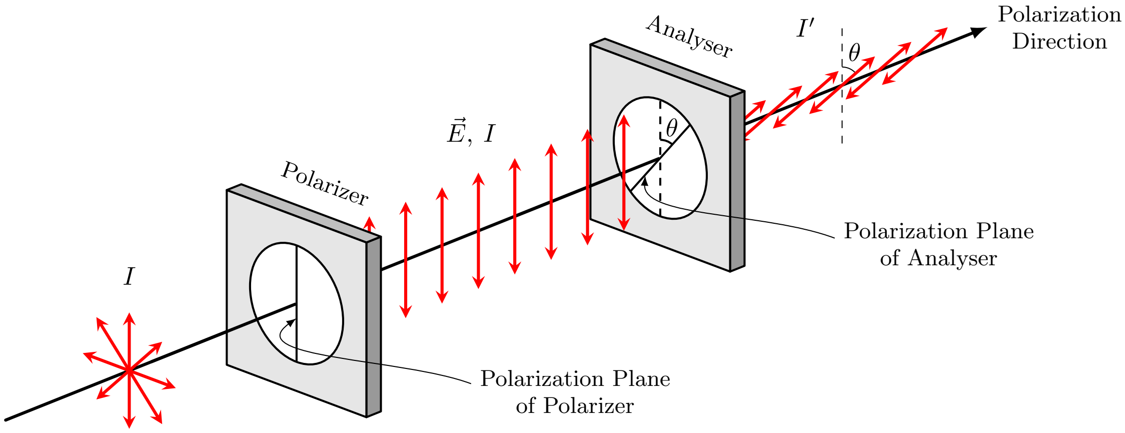

Analyser
I'
θ
Polarization
Direction
$\vec{E},\ I$
θ
Polarizer
Polarization Plane
of Analyser
I
Polarization Plane
of Polarizer

Figure 10.262: Simplified example of a polariser / analyser set (author: Alexandros Tsagkaropoulos)

We first let light pass through a polariser whose axis makes an angle θ with the x-axis, then pass through a second polariser, referred to as "analyser", whose axis makes an angle α with the same axis (see figure below) with:

$$\vec{n} = (\cos(\alpha), \sin(\alpha)) \qquad (10.6.289)$$

whose norm is equal to unity!

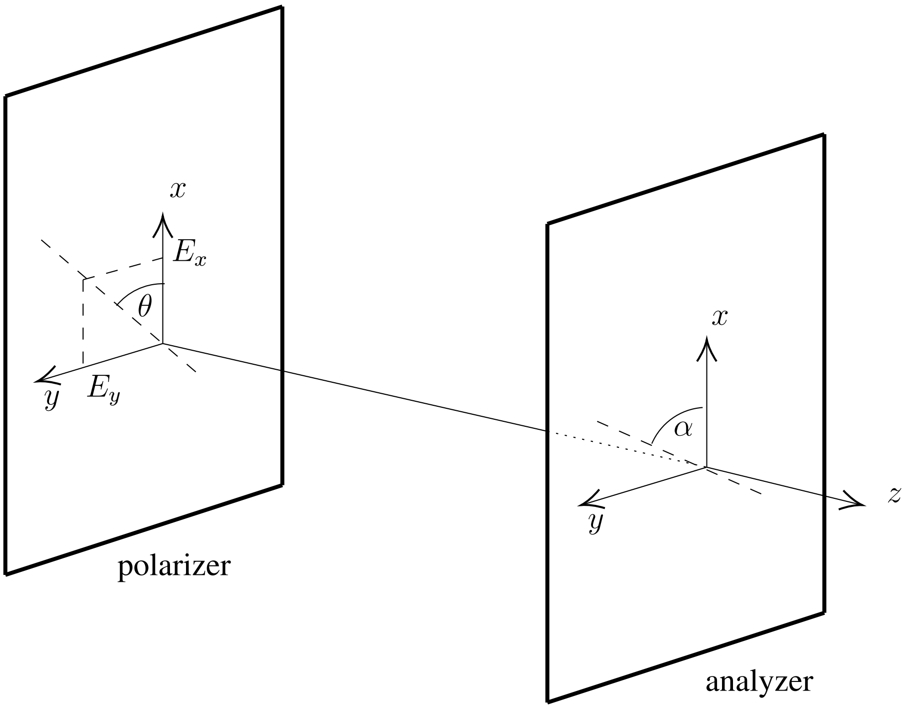


Figure 10.263: **Coordinates system used for our study**

At the exit of the analyser, the electric field $\vec{E}\,'$ is obtained by projecting the linear polarized light ($\varphi = 0$) obtained at the exit of polariser:

$$\vec{E} = \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix} = \begin{pmatrix} \hat{E}\cos(\theta)\cos(\omega t) \\ \hat{E}\sin(\theta)\cos(\omega t) \\ 0 \end{pmatrix} = \hat{E}\vec{p}\cos(\omega t) \qquad (10.6.290)$$

with:

$$\vec{p} = (\cos(\theta), \sin(\theta)) \qquad (10.6.291)$$

on $\vec{n}$ (which means: projection = dot product, to get a vector we multiply for recall by the vector on which we do the projection):

$$\begin{aligned}\vec{E}' &= (\vec{E}\circ\vec{n})\vec{n} = (E_0\cos(\omega t)(\vec{p}\circ\vec{n}))\vec{n} = E_0\cos(\omega t)(\cos(\theta)\cos(\alpha)+\sin(\theta)\sin(\alpha))\vec{n} \\ &= [E_0\cos(\omega t)\cos(\theta-\alpha)]\vec{n}\end{aligned} \qquad (10.6.292)$$

We the deduce the Malus law for the intensity:

$$\boxed{I' = I\cos^2(\theta-\alpha)} \qquad (10.6.293)$$

for the case of a linear polarization of course. We will reuse that result for our study of quantum cryptography (see section Cryptography page 7263).

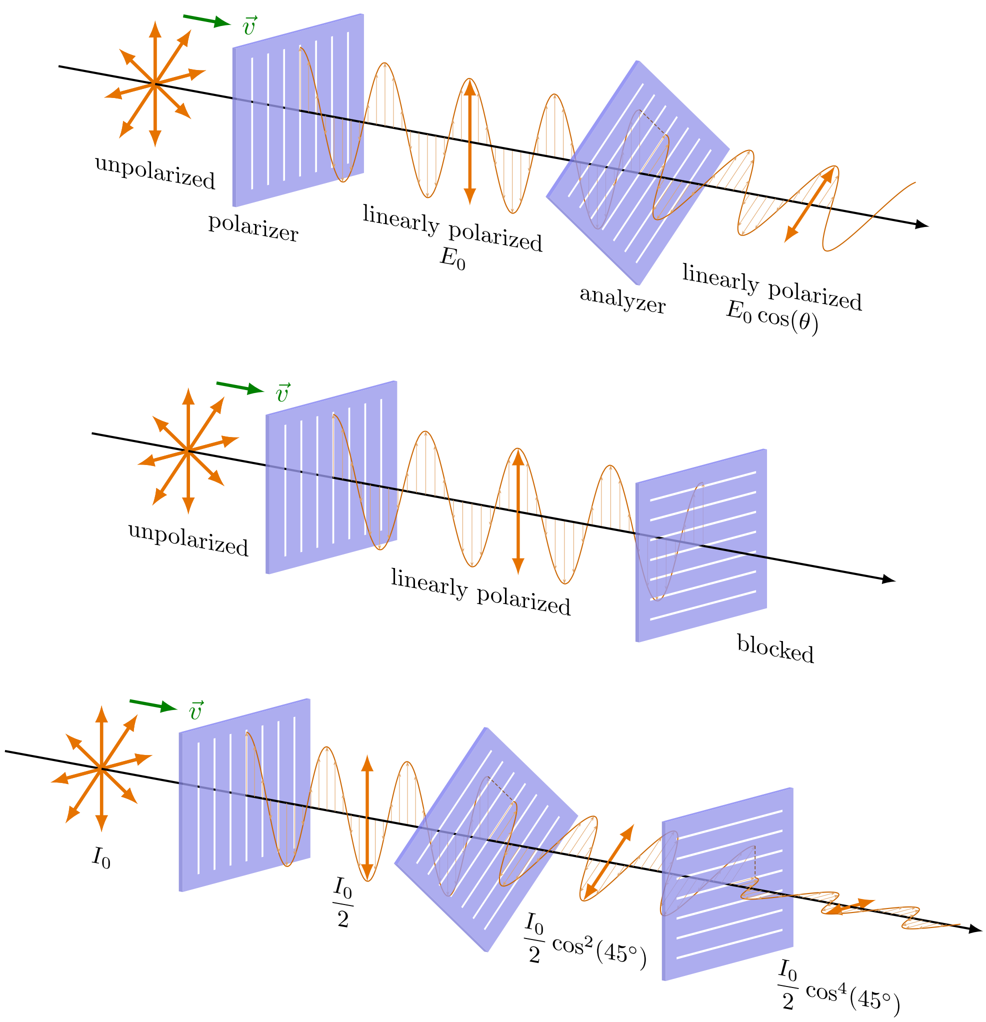

$\vec{v}$
unpolarized
polarizer
linearly polarized
E_0
analyzer
linearly polarized
$E_0\cos(\theta)$
$\vec{v}$
unpolarized
linearly polarized
blocked
$\vec{v}$
I_0
$\frac{I_0}{2}$
$\frac{I_0}{2}\cos^2(45°)$
$\frac{I_0}{2}\cos^4(45°)$

Figure 10.264: **Different configuration of linear polarizers (author: Izaak Neutelings)**

A polarizer does not work by transmitting one very specific polarization angle and blocking all other. The reason for this is that any incident polarization can always be seen as a linear superposition of the polarizer's polarization and the one orthogonal to it. The orthogonal polarization is completely blocked and the parallel polarization is completely transmitted - this is the definition of what a perfect polarizer does. Now, an incident electric field vector $\vec{E}$ at angle θ is:

$$\vec{E}_0 = \cos(\theta)E_{\parallel} + \sin(\theta)E_{\perp} \tag{10.6.294}$$

in terms of the field in the parallel direction $E_{\parallel}$and the field in the orthogonal direction $E_{\perp}$. After the polarizer, all that is left is $\vec{E}_1 = \cos(\theta)E_{\parallel\parallel}$. Therefore, the ratio of the intensities is:

$$\frac{I_1}{I_0} \propto \frac{\left(\vec{E}_1\right)^2}{\left(\vec{E}_0\right)^2} = \cos^2(\theta) \tag{10.6.295}$$

which shows Malus' law.

If the polarization angle of the the incident radiation is uniformly randomly distributed among the possible angles, then the average intensity will be:

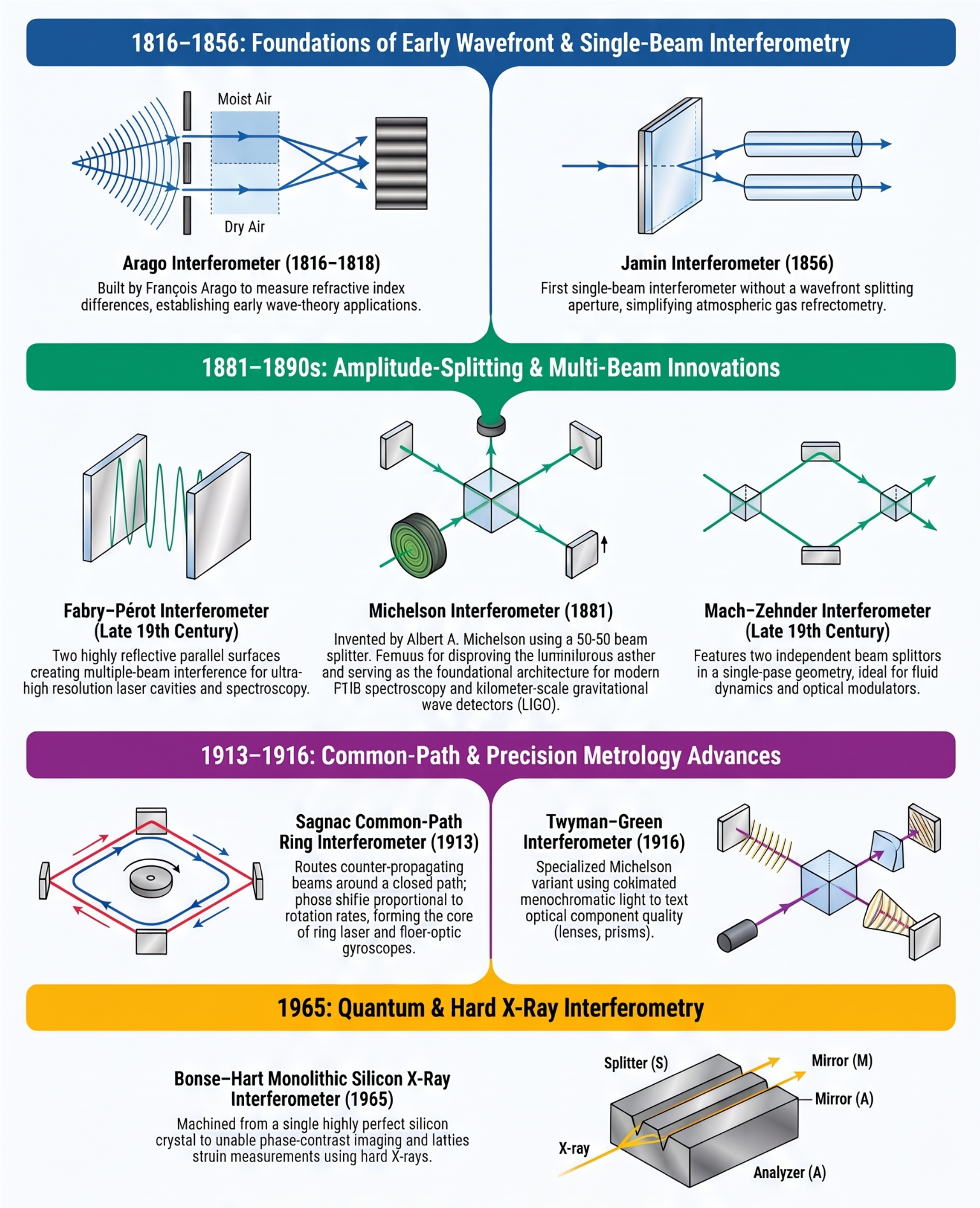
$$\frac{1}{2\pi}\int_{0}^{2\pi}\cos^2(\theta)\mathrm{d}\theta=\frac{1}{2} \qquad (10.6.296)$$

i.e. "totally unpolarized" photons have, on average, an equal chance to be absorbed by or transmitted through the polarizer.

10.6.5 Coherence and interference (interferometry)

Interferometry is a technique which uses the interference of superimposed waves to extract information for various purposes. Interferometry typically uses electromagnetic waves and is an important investigative technique in the fields of astronomy, fiber optics, engineering, metrology, optical metrology, oceanography, seismology, spectroscopy (and its applications to chemistry), aeronautics, quantum mechanics, nuclear and particle physics, plasma physics, biomolecular interactions, surface profiling, microfluidics, mechanical stress/strain measurement, velocimetry, optometry, radiology, and making holograms.

Interferometers are devices that extract information from interference. They are widely used in science and industry for the measurement of microscopic displacements, refractive index changes and surface irregularities. In the case with most interferometers, light from a single source is split into two beams that travel in different optical paths, which are then combined again to produce interference; two incoherent sources can also be made to interfere under some circumstances. The resulting interference fringes give information about the difference in optical path lengths. In analytical science, interferometers are used to measure lengths and the shape of optical components with many times nanometer precision or even below (like for gravitational interferometers); they are the highest-precision length measuring instruments in existence. In Fourier transform spectroscopy they are used to analyze light containing features of absorption or emission associated with a substance or mixture. An astronomical interferometer consists of two or more separate dishes that correlates their signals, offering a resolution equivalent to that of a single-dish telescope of diameter equal to the largest separation between its individual elements.


1816–1856: Foundations of Early Wavefront & Single-Beam Interferometry
Moist Air
Dry Air
Arago Interferometer (1816–1818)
Jamin Interferometer (1856)
1881–1890s: Amplitude-Splitting & Multi-Beam Innovations
Fabry–Pérot Interferometer (Late 19th Century)
Michelson Interferometer (1881)
Mach–Zehnder Interferometer (Late 19th Century)
1913–1916: Common-Path & Precision Metrology Advances
Sagnac Common-Path Ring Interferometer (1913)
Twyman–Green Interferometer (1916)
1965: Quantum & Hard X-Ray Interferometry
Bonse–Hart Monolithic Silicon X-Ray Interferometer (1965)
Splitter (S)
Mirror (M)
Mirror (A)
X-ray
Analyzer (A)

Built by François Arago to measure refractive index differences, establishing early wave-theory applications.

First single-beam interferometer without a wavefront splitting aperture, simplifying atmospheric gas refractometry.

Two highly reflective parallel surfaces creating multiple-beam interference for ultra-high resolution laser cavities and spectroscopy.

Invented by Albert A. Michelson using a 50-50 beam splitter. Femuus for disproving the luminilurous asther and serving as the foundational architecture for modern FTIB spectroscopy and kilometer-scale gravitational wave detectors (LIGO).

Features two independent beam splittors in a single-pase geometry, ideal for fluid dynamics and optical modulators.

Routes counter-propagating beams around a closed path; phose shifie proportional to rotation rates, forming the core of ring laser and floer-optic gyroscopes.

Specialized Michelson variant using cokimated menochromatic light to text optical component quality (lenses, prisms).

Machined from a single highly perfect silicon crystal to unable phase-contrast imaging and latties struin measurements using hard X-rays.

Figure 10.265: Timeline of the most well-known interferometers

We will now see what are the conditions such that a plane waves interfere with each other. These developments help to explain well much about the vision of the world around us through our eye (especially why all the waves received by our retinas do not mix and so the colors either!).

Let us consider two plane waves Σ_1 and Σ_2 of pulsation ω_1 and ω_2 and of wave vectors $\vec{k}_1$ and $\vec{k}_2$ propagating both parallel to the z-axis.

We denote by Ψ_1 and $\Psi_2$the complex amplitudes of the two waves and we are interested in the average intensity observed at a point O taken as the coordinate origin:

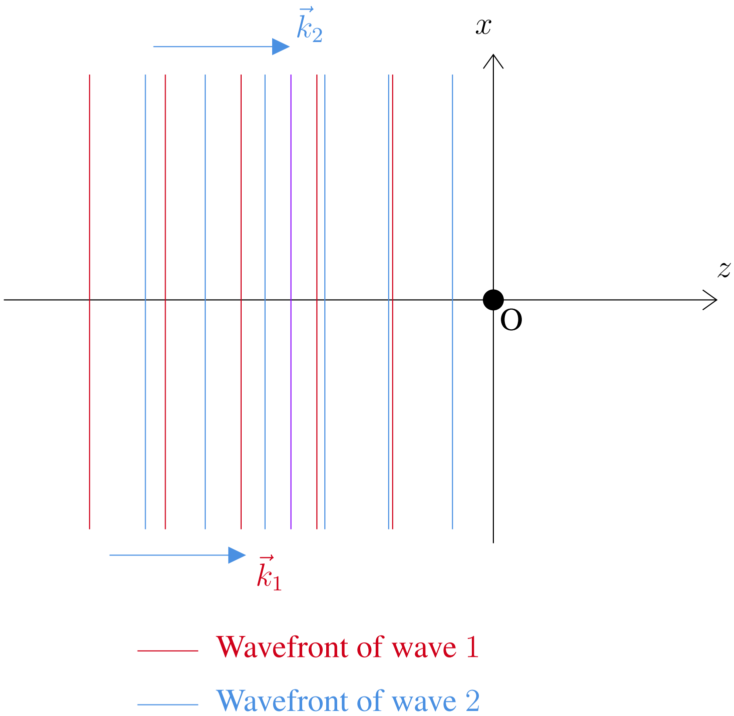


Figure 10.266: Representation of plane waves

We put:

$$\omega_0=\frac{\omega_1+\omega_2}{2} \tag{10.6.297}$$

and we will assume:

$$\delta\omega=\omega_2-\omega_1\ll 2\omega_0 \tag{10.6.298}$$

At the point O the complex amplitudes are written:

$$\begin{gathered}\Psi_1=\Psi_{01}e^{-\mathrm{i}\omega_1 t}=|\Psi_1|e^{\mathrm{i}\theta_1}e^{-\mathrm{i}\omega_1 t}\\ \Psi_2=\Psi_{02}e^{-\mathrm{i}\omega_2 t}=|\Psi_2|e^{\mathrm{i}\theta_2}e^{-\mathrm{i}\omega_2 t}\end{gathered} \tag{10.6.299}$$

where θ_1 and θ_2 represent the phases of Ψ_1 and Ψ_2.

Let us now calculate the instantaneous intensity at the point O will be denoted $J(t)$. As the average intensity I is proportional to the square of the amplitude, we assume it will be the same for the instantaneous intensity. Which brings us to calculate the sum of the real parts of the amplitudes of the two waves:

$$J(t)=\left(\Re(\Psi_1)+\Re(\Psi)\right)^2=\Re(\Psi_1)^2+\Re(\Psi_2)^2+2\Re(\Psi_1)\Re(\Psi_2) \tag{10.6.300}$$

What is written keeping in mind that (see section Numbers page 243):

$$\Re(\Psi)=\frac{1}{2}(\Psi+\bar{\Psi}) \tag{10.6.301}$$

Therefore:

$$\begin{aligned}\Re(\Psi_1)\Re(\Psi_2)&=\frac{1}{2}(\Psi_1+\bar{\Psi}_1)\frac{1}{2}(\Psi_2+\bar{\Psi}_2)=\frac{1}{4}(\Psi_1+\bar{\Psi}_1)(\Psi_2+\bar{\Psi}_2)\\&=\frac{1}{4}(\Psi_1\Psi_2+\Psi_1\bar{\Psi}_2+\bar{\Psi}_1\Psi_2+\bar{\Psi}_1\bar{\Psi}_2)=\frac{1}{4}(\Psi_1\Psi_1+\bar{\Psi}_1\bar{\Psi}_2)+\frac{1}{4}(\bar{\Psi}_1\Psi_2+\Psi_1\bar{\Psi}_2)\\&=\frac{1}{2}\Re(\Psi_1\Psi_2)+\frac{1}{2}\Re(\Psi_1\bar{\Psi}_2)\end{aligned} \tag{10.6.302}$$

And then we have:

$$J(t)=\Re(\Psi_1)^2+\Re(\Psi_2)^2+2\Re(\Psi_1)\Re(\Psi_2)=\Re(\Psi_1)^2+\Re(\Psi_2)^2+\Re(\Psi_1\Psi_1)+\Re(\Psi_1\bar{\Psi}_2) \tag{10.6.303}$$

It follows the sum of four terms:

$$\begin{aligned}J(t)=&|\Psi_1|^2\cos^2(\omega_1 t-\theta_1)+|\Psi_2|^2\cos^2(\omega_2 t-\theta_2)+|\Psi_1\Psi_2|\cos\left[(\omega_2+\omega_1)t-(\theta_1+\theta_2)\right]\\&+|\Psi_1\Psi_2|\cos\left[(\omega_2-\omega_1)t+(\theta_1-\theta_2)\right]\end{aligned} \tag{10.6.304}$$

To calculate the average intensity, we will choose an experimental approach:

The average intensity of the exposure time τ of the detector (electronic or biological) will be given by:

$$I=\frac{1}{\tau}\int\limits_0^{\tau}J(t)\mathrm{d}t=\langle J(t)\rangle_{\tau} \tag{10.6.305}$$

where I is then the sum of the averages of the four terms intervening in $J(t)$ given just above. For visible light (in the case of our eye), the frequencies are of the order of 10^{15} [Hz] and the exposure time of the detectors vary between the millisecond and the second. Then τ typically contains 10^{12} periods of Ψ_1 and Ψ_2!!

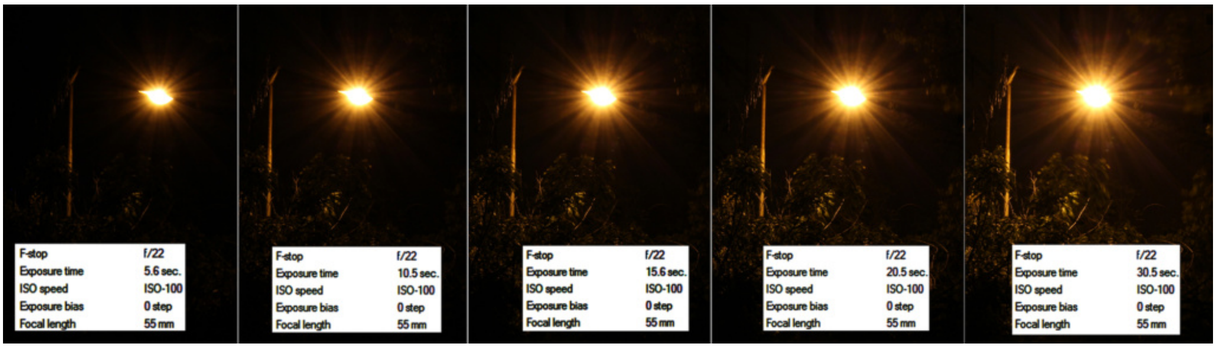

F-stop f/22
Exposure time 5.6 sec.
ISO speed ISO-100
Exposure bias 0 step
Focal length 55 mm
F-stop f/22
Exposure time 10.5 sec.
ISO speed ISO-100
Exposure bias 0 step
Focal length 55 mm
F-stop f/22
Exposure time 15.6 sec.
ISO speed ISO-100
Exposure bias 0 step
Focal length 55 mm
F-stop f/22
Exposure time 20.5 sec.
ISO speed ISO-100
Exposure bias 0 step
Focal length 55 mm
F-stop f/22
Exposure time 30.5 sec.
ISO speed ISO-100
Exposure bias 0 step
Focal length 55 mm

Figure 10.267: Example of the influence of various exposure times with a camera

Let us examinate the effect on the average value of each of the term of $J(t)$ by first recalling the relation (see section Wave Mechanics page 3480):

$$\omega=2\pi f=\frac{2\pi}{T} \tag{10.6.306}$$

1. We have using the usual integral proved in the section of Differential and Integral Calculus, the following valuation of the integrand on an important number of periods:

$$\begin{aligned}
I_1&=\frac{1}{nT_1}\int\limits_0^{nT_1}|\Psi_1|^2\cos^2(\omega_1 t-\theta)\mathrm{d}t=\frac{|\Psi_1|^2}{nT_1\omega_1}\int\limits_{-\theta_i}^{n2\pi-\theta_i}\cos^2(x)\mathrm{d}x\\
&=\frac{|\Psi_1|^2}{nT_1 2\pi f_1}\int\limits_{-\theta_1}^{n2\pi-\theta_i}\cos^2(x)\mathrm{d}x=\frac{|\Psi_1|^2}{n2\pi}\int\limits_{-\theta_i}^{n2\pi-\theta_i}\cos^2(x)\mathrm{d}x\\
&=\frac{|\Psi_1|^2}{2\pi n}\left[\frac{1}{2}\left(\int\limits_{-\theta_i}^{n2\pi-\theta_i}\mathrm{d}x+\sin(x)\cos(x)\Big|_{-\theta_i}^{n2\pi-\theta_i}\right)\right]\\
&=\frac{|\Psi_1|^2}{2\pi n}\left[\frac{1}{2}\left([(n2\pi-\theta_1)-(-\theta_1)]+(\sin(2n\pi-\theta_1)\cos(n2\pi-\theta_1)-\sin(-\theta_1)\cos(-\theta_1))\right)\right]\\
&=\frac{|\Psi_1|^2}{2\pi n}\left[\frac{1}{2}\left(n2\pi+(\sin(-\theta_1)\cos(-\theta_1)-\sin(-\theta_1)\cos(-\theta_1))\right)\right]\\
&=\frac{|\Psi_1|^2}{2\pi n}\left[\frac{1}{2}(n2\pi)\right]=\frac{|\Psi_1|^2}{2}
\end{aligned}$$

(10.6.307)

Calculation that is traditionally written and very abusively under the following condensed form:

$$I_1=\frac{1}{2\pi}\int\limits_0^{2\pi}|\Psi_1|^2\cos^2(\omega_1 t-\theta_1)\mathrm{d}t=\frac{|\Psi_1|^2}{2\pi}\int\limits_0^{2\pi}\cos^2(\omega_1 t-\theta_1)\mathrm{d}t=\frac{|\Psi_1|^2}{2\pi}\pi=\frac{|\Psi_1|^2}{2}$$

(10.6.308)

We can estimate that a large number of periods (opening time of the detector), it is this average that will be measured (in fact its really that latter!).

2. We have identically:

$$I_2=\frac{1}{2\pi}\int\limits_0^{2\pi}|\Psi_2|^2\cos^2(\omega_2 t-\theta_2)\mathrm{d}t=\frac{|\Psi|^2}{2\pi}\int\limits_0^{2\pi}\cos^2(\omega_2 t-\theta_2)\mathrm{d}t=\frac{|\Psi_2|^2}{2\pi}\pi=\frac{|\Psi_2|^2}{2}$$

(10.6.309)

with the same remarks as above regarding to the detector!

3. For the third term it is a bit different:

$$\begin{aligned}I_3&=\frac{1}{2\pi}\int\limits_0^{2\pi}|\Psi_1\Psi_2|\cos\left[(\omega_2+\omega_1)t-(\theta_1+\theta_2)\right]\mathrm{d}t=\frac{|\Psi_1\Psi_2|}{2\pi}\int\limits_0^{2\pi}\cos\left[(\omega_2+\omega_1)t-(\theta_1+\theta_2)\right]\mathrm{d}t\\&=\frac{|\Psi_1\Psi_2|}{2\pi}\left[\int\limits_0^{2\pi}\cos((\omega_2+\omega_1)t)\cos(\theta_1+\theta_2)\mathrm{d}t+\int\limits_0^{2\pi}\sin((\omega_2+\omega_1)t)\sin(\theta_1+\theta_2)\mathrm{d}t\right]\\&=\frac{|\Psi_1\Psi_2|}{2\pi}\left[\cos(\theta_1+\theta_2)\int\limits_0^{2\pi}\cos((\omega_2+\omega_1)t)\mathrm{d}t+\sin(\theta_1+\theta_2)\int\limits_0^{2\pi}\sin((\omega_2+\omega_1)t)\mathrm{d}t\right]\\&=\frac{|\Psi_1\Psi_2|}{2\pi}\left[\cos(\theta_1+\theta_2)\int\limits_0^{2\pi}\cos(2\omega_0t)\mathrm{d}t+\sin(\theta_1+\theta_2)\int\limits_0^{2\pi}\sin(2\omega_0t)\mathrm{d}t\right]\end{aligned}\tag{10.6.310}$$

However, the average of a cosine and a sine on a period is zero. So if the detector makes a measure on an exposure time above $1/2\omega_0$, that is on a large number of periods, we will have:

$$\begin{aligned}I_3&=\frac{|\Psi_1\Psi_2|}{2\pi}\left[\cos(\theta_1+\theta_2)\int\limits_0^{2\pi}\cos(2\omega_0t)\mathrm{d}t+\sin(\theta_1+\theta_2)\int\limits_0^{2\pi}\sin(2\omega_0t)\mathrm{d}t\right]\\&=\frac{|\Psi_1\Psi_2|}{2\pi}\left[\cos(\theta_1+\theta_2)\cdot0+\sin(\theta_1+\theta_2)\cdot0\right]=0\end{aligned}\tag{10.6.311}$$

4. For the fourth term it is still different in the experimental approximation. Indeed:

$$\begin{aligned}I_4&=\frac{1}{2\pi}\int\limits_0^{2\pi}|\Psi_1\Psi_2|\cos\left[\cos(\omega_2-\omega_1)t+(\theta_1-\theta_2)\right]\mathrm{d}t=\frac{|\Psi_1\Psi_2|}{2\pi}\int\limits_0^{2\pi}\int\cos\left[(\omega_2-\omega_1)t+(\theta_1-\theta_2)\right]\mathrm{d}t\\&=\frac{|\Psi_1\Psi_2|}{2\pi}\left[\int\limits_0^{2\pi}\cos((\omega_2-\omega_1)t)\cos(\theta_1-\theta_2)\mathrm{d}t-\int\limits_0^{2\pi}\sin((\omega_2-\omega_1)t)\sin(\theta_1-\theta_2)\mathrm{d}t\right]\\&=\frac{|\Psi_1\Psi_2|}{2\pi}\left[\cos(\theta_1-\theta_2)\int\limits_0^{2\pi}\cos((\omega_2-\omega_1)t)\mathrm{d}t-\sin(\theta_1-\theta_2)\int\limits_0^{2\pi}\sin((\omega_2-\omega_1)t)\mathrm{d}t\right]\\&=\frac{|\Psi_1\Psi_2|}{2\pi}\left[\cos(\theta_1+\theta_2)\int\limits_0^{2\pi}\cos(\delta\omega t)\mathrm{d}t-\sin(\theta_1+\theta_2)\int\limits_0^{2\pi}\sin(\delta\omega t)\mathrm{d}t\right]\end{aligned}\tag{10.6.312}$$

But, $\delta\omega\ll2\omega_0$. Therefore the detector does not have time to measure the average intensity over a whole period in first approximation since:

$$T=\frac{2\pi}{\delta\omega}\tag{10.6.313}$$

and that this value is much much greater in the visible spectrum that the opening / sampling time of the eye that it is 0.1 [s].

Thus, we will notice the average of the fourth term by:

$$I_4=\frac{|\Psi_1\Psi_2}{2\pi}\int\limits_0^{2\pi}\cos\left[(\omega_2-\omega_1)t+(\theta_1-\theta_2)\right]\mathrm{d}t=|\Psi_1\Psi_2|\langle\cos\left[\delta\omega t+(\theta_1-\theta_2)\right]\rangle_\tau=f(\tau)|\Psi_1\Psi_2| \tag{10.6.314}$$

The average intensity is therefore in an experimental context:

$$I=\frac{|\Psi_1|^2}{2}+\frac{|\Psi_2|^2}{2}+f(\tau)|\Psi_1\Psi_2| \tag{10.6.315}$$

or:

$$\boxed{I=I_1+I_2+f(\tau)\sqrt{I_1I_2}} \tag{10.6.316}$$

If the pulsations ω_1, ω_2 are equal (or almost equal), then it is the interference between two monochromatic plane waves. The average intensity is then written:

$$\begin{aligned}I&=I_1+I_2+2\sqrt{I_1I_2}f(\tau)=I_1+I_2+2\sqrt{I_1I_2}\langle\cos(\theta_1-\theta_2)\rangle_\tau\\&=I_1+I_2+2\sqrt{I_1I_2}\frac{1}{\tau}\int\limits_0^{\tau}\cos(\theta_1-\theta_2)\mathrm{d}t=I_1+I_2+2\sqrt{I_1I_2}\cos(\theta_1-\theta_2)\end{aligned} \tag{10.6.317}$$

When we know that the eye interprets the intensity to form perceptions of the objects we understand why two objects of two different colors are not a perception corresponding to a mixture of the two colors because even if in the visible spectrum, the pulses are almost equal, their phase is rarely at a given point in space zero such that:

$$I=I_1+I_2+2\sqrt{I_1I_2}\cos(0)=I_1+I_2+2\sqrt{I_1I_2} \tag{10.6.318}$$

There is therefore no interference and we have in reality:

$$I=I_1+I_2+2\sqrt{I_1I_2}\cos(\theta_2-\theta_1)\cong I_1+I_2 \tag{10.6.319}$$

Remark

During the composition of several waves, we can always consider that there is interference. However, we name "interference conditions" the conditions of observation of these interferences, i.e. conditions for the result of their composition to be stable enough to be observed. It is customary to speak of "visibility", which restricts to the single observation by the (human) eye.

We have seen for the eye that the sampling time frequency is $10\,[\mathrm{s}^{-1}]$. Knowing that the visible light has a frequency of $f\cong 10^{14}\,[s^{-1}]$, the frequency must then be stabilized by the source during:

$$\frac{10}{10^{14}}=10^{-13} \tag{10.6.320}$$

which is materially impossible except that the source is the same. We conclude that for interference to be visible to the eye, the sources must be synchronized at best at 10^{-13} which in practice leads to consider only sources absolutely synchronized on a single source.

In the previous model, we also have neglected the fact that a real wave is limited in time. A photon is represented by a limited wave packet. Given T, it will have a length $l_c = cT$ in a vacuum or in the air that we name "temporal coherence length".

A given radiation is thus a superposition of succession of wave trains whose average length is l_c, the successive wave trains have no phase relation between them: they can not interfere!

10.6.5.1 Michelson Interferometer

The "Michelson interferometer" is a very common and important configuration for optical interferometry[24] and was invented by the American physicist Albert Abraham Michelson in 11887 (holocene calendar) and at its time the setup was mounted on a stone slab that was floating in an annular trough of mercury (yellow light from a sodium flame (for alignment), or white light (for the actual observations).

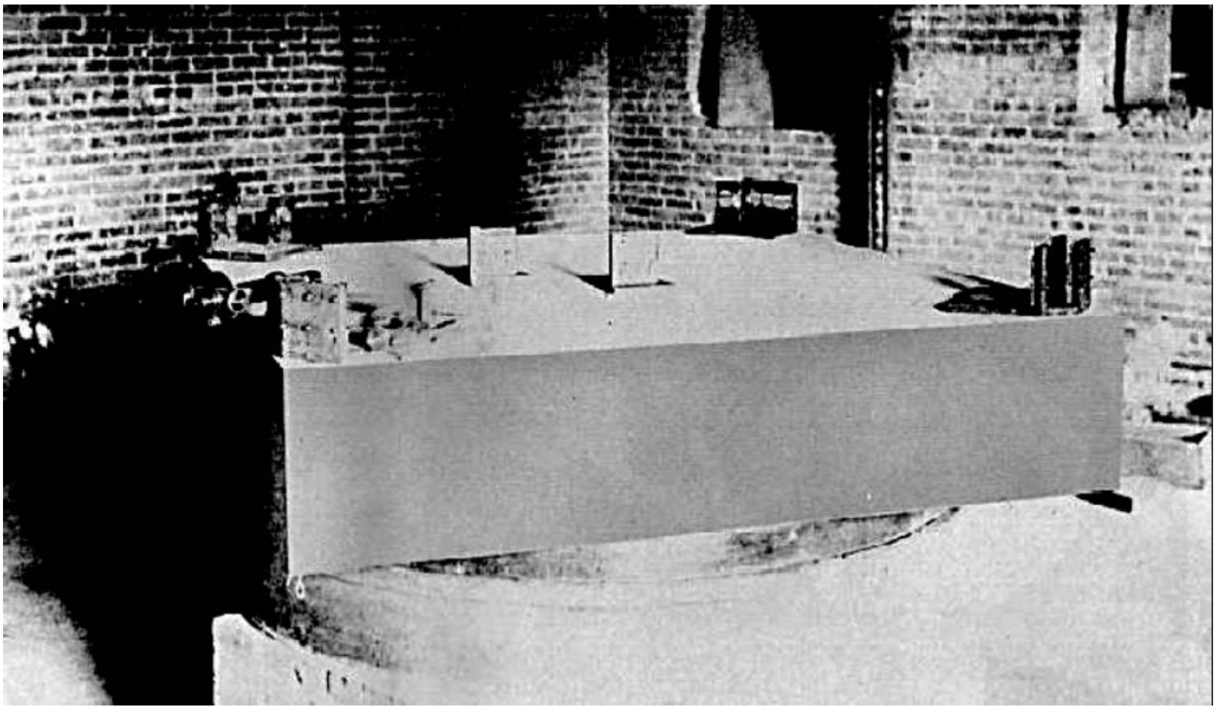


Figure 10.268: Setup of Michelson Morley experiment in 11887 at what is now Case Western Reserve University

It is considered as the single most important design in the history of precision measurement.

Using a beam splitter, a light source is split into two arms. Each of those light beams is reflected back toward the beamsplitter which then combines their amplitudes using the superposition principle. The resulting interference pattern that is not directed back toward the source is typically directed to some type of photoelectric detector or camera. For different applications of the interferometer, the two light paths can be with different lengths or incorporate optical elements or even materials under test.

The Michelson interferometer is employed in many scientific experiments and became well known for its use by Albert Abraham Michelson and Edward Morley in the famous Michelson-Morley experiment (11887 according to holocene calendar) in a configuration which would have detected the Earth's motion through the supposed luminiferous aether that most physicists at the time believed was the medium in which light waves propagated. The null result of that experiment essentially gave evidence against the existence of such an aether, leading eventually to the Special Relativity and the revolution in physics at the beginning of the twentieth century (see [390]). In 12015 (holocene calendar), another application of the Michelson interferometer, LIGO, made the first direct observation of gravitational waves. That observation confirmed an important prediction of General Relativity, validating the theory's prediction of space-time distortion in the context of large scale cosmic events (known as strong field tests).

---

[24] The underlying idea of this interferometer is use for the gravitational waves detectors.

Remark

The idea was that if Earth moves through the ether, light should travel faster in one direction than another (like swimming with vs. against a current). The interferometer would show this as a fringe shift when rotated. The result ? No fringe shift! None. The speed of light was the same in every direction, regardless of Earth's motion. This "null result" was one of the most important failed experiments in history.

A Michelson interferometer consists minimally of two mirrors M_1 and M_2 and a beam splitter M (although a diffraction grating is also used). In the figure below:

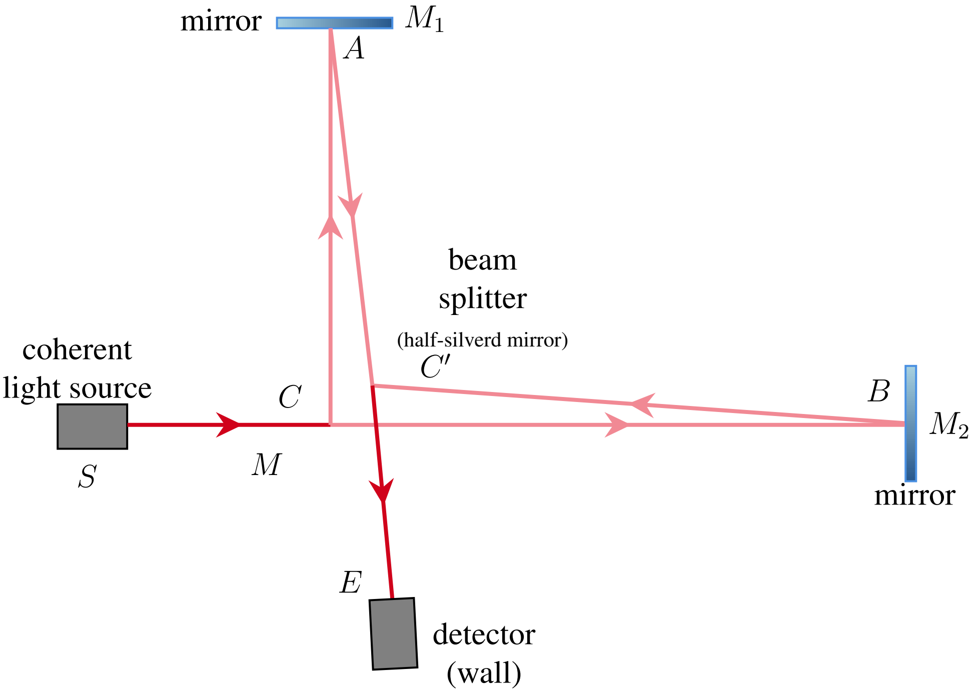


Figure 10.269: **Path of light in a simple Michelson interferometer**

A source S emits light that hits the beam splitter (in this case, a plate beamsplitter) surface M at point C. M is partially reflective, so part of the light is transmitted through to point B while some is reflected in the direction of A. Both beams recombine at point C' to produce an interference pattern incident on the detector at point E (or on the retina of a person's eye). If there is a slight angle between the two returning beams, for instance, then an imaging detector will record a sinusoidal fringe pattern - named "Haidinger fringes" - as shown in the figure below:

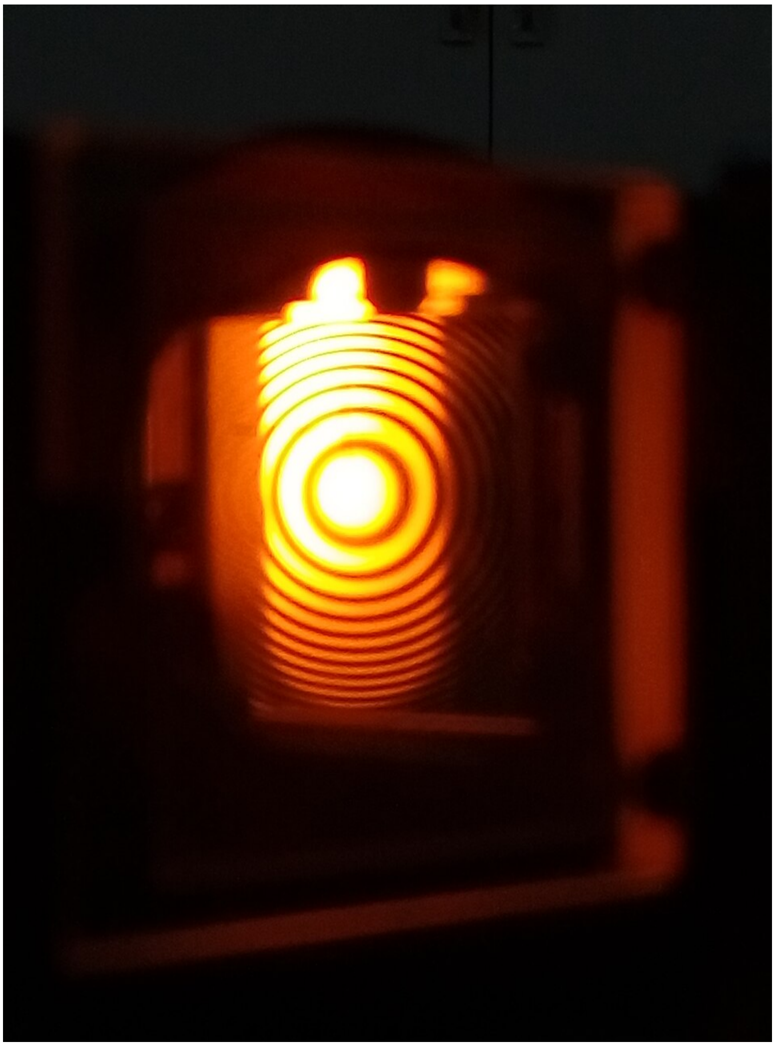


Figure 10.270: **Fringe pattern formed by the Michelson interferometer using monochromatic light - sodium D lines (source: Wikipedia, author: Warren Leywon)**

A typical Swiss PASCO® Scientific OS-925AA setup with a Helium/Neon $4\,[\text{mW}]$ LASER looks like this:

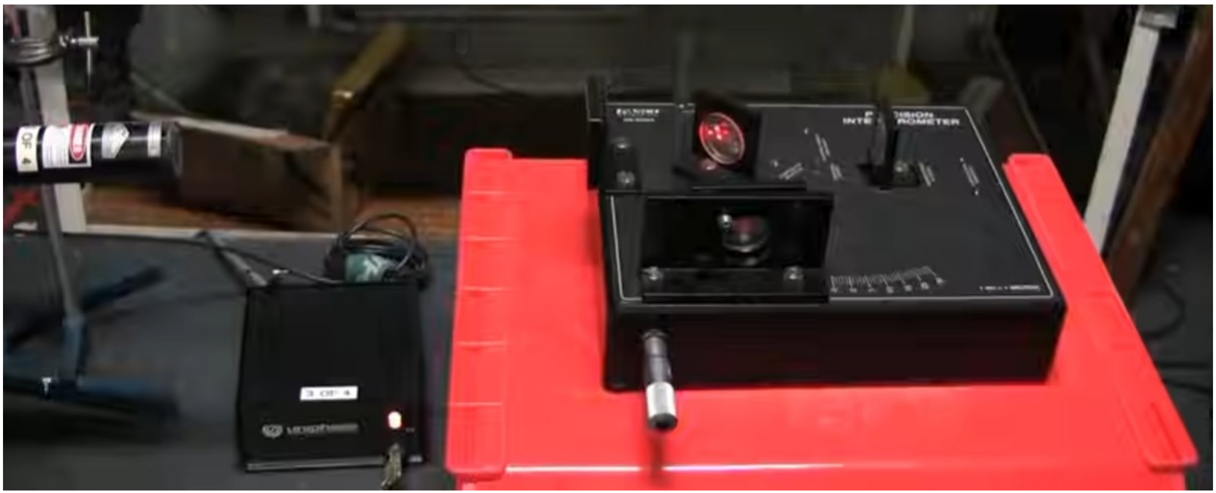


Figure 10.271: **Michelson interferometer typical setup (source: MIT)**

where we can see the beam splitter (the lens the multiple LASER dots) and at the bottom with the graduation a mirror ajustable with a micrometer knob.

From the top this is how it looks like if we draw in a childish way the LASER light wave (the LASER light enters from the left of the photo and the interference runs away from the top of the photo)

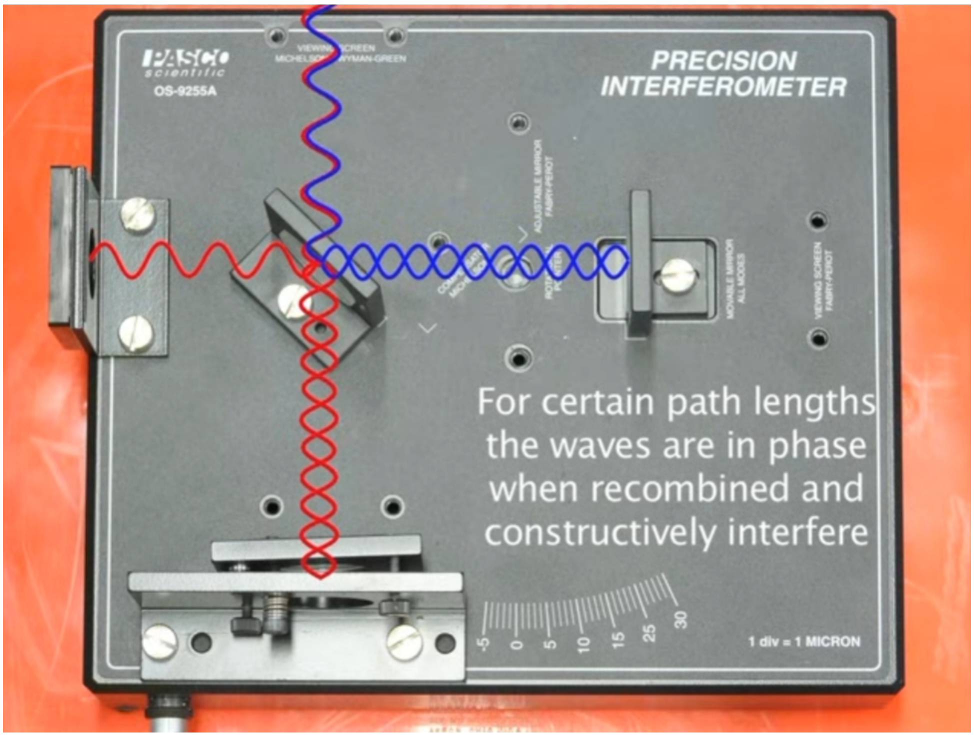
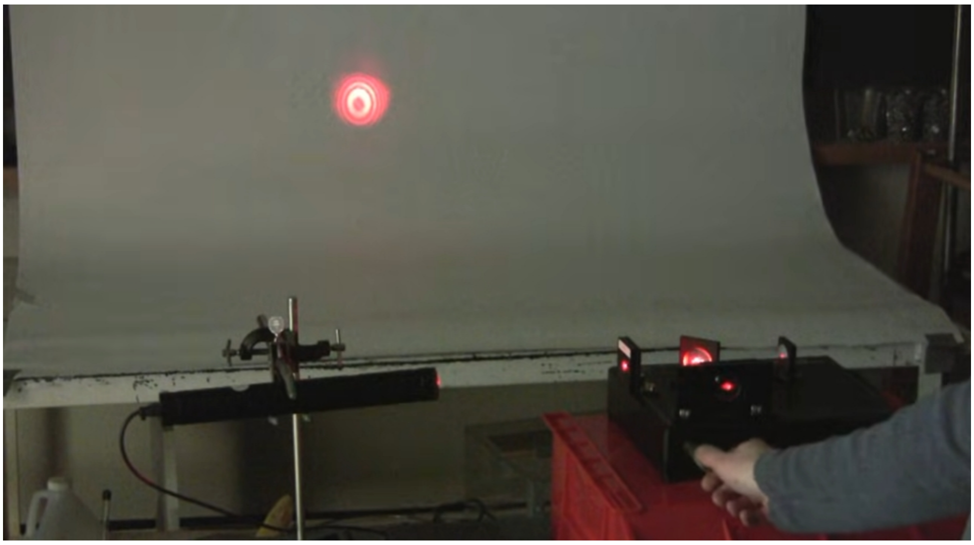

PASCO scientific
OS-9255A
VIEWING SCREEN
MICHELSON TWYMAN-GREEN
PRECISION INTERFEROMETER
ADJUSTABLE MIRROR FABRY-PEROT
MOVABLE MIRROR ALL MODES
VIEWING SCREEN FABRY-PEROT
For certain path lengths the waves are in phase when recombined and constructively interfere
-5
0
5
10
15
25
30
1 div = 1 MICRON

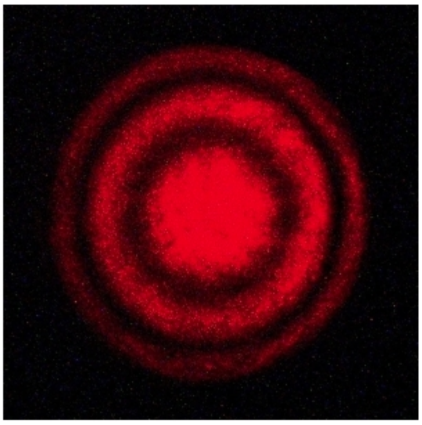
Figure 10.272: Michelson interferometer schematic constructive interference (source: MIT)

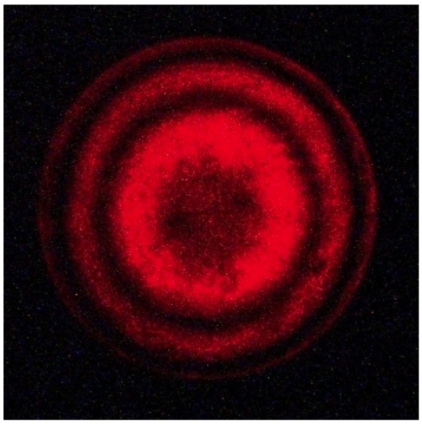
We have on the projection wall something typically that will look like this:



If the path lengths differ by exactly an integral number of wavelengths plus (or minus) a half wavelength, they undergo destructive interference to give an intensity minimum. One thus observes a pattern of fringes that appears to move as one adjusts the length of the variable arm.

When the two mirrors are "coplanar," that is, they are exactly perpendicular to the beam paths, one observes a symmetrical pattern such as those shown below:





The image on the left shows an interference maximum at the center, and the image at right shows an interference minimum at the center. In each photo you can see two fringes (and perhaps part of the next) surrounding the center spot. When the two arms are exactly the same length (i.e., they differ by exactly zero wavelengths!) and the mirrors are coplanar, the central maximum grows to fill the entire field.

When the mirrors are not coplanar, one can obtain so-called "hyperbolic" fringes (see proof further below), which appear as straight vertical lines. The fringes in the photograph at the top of this page indicate a mirror alignment somewhere between that which would give circular fringes and that which would give hyperbolic fringes.

After with stickers we mark the interference fringes to make them more easy to follow with our eyes when we change the micrometer knob:

Interference fringes are marked so that
cycles may be counted more easily

Remark

If there is perfect spatial alignment between the returning beams, then there will not be any such pattern but rather a constant intensity over the beam dependent on the differential pathlength; this is difficult, requiring very precise control of the beam paths.

Now we move the micrometer knob a at least for one cycle of constructive interference (one is enough but if we do it multiple times our calculation will give a smaller confidence interval according[25] to what we have seen in the section of Statistics page 855):

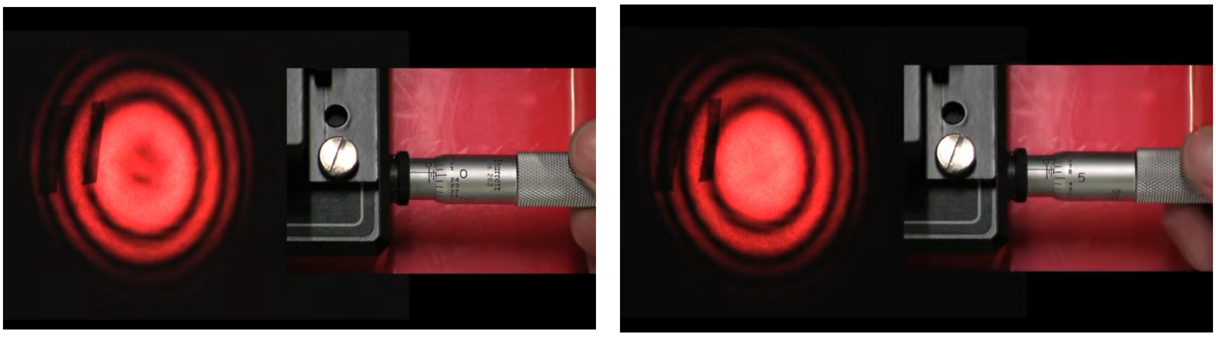


Figure 10.273: **Counting the interference fringe pattern when turning the micrometer**

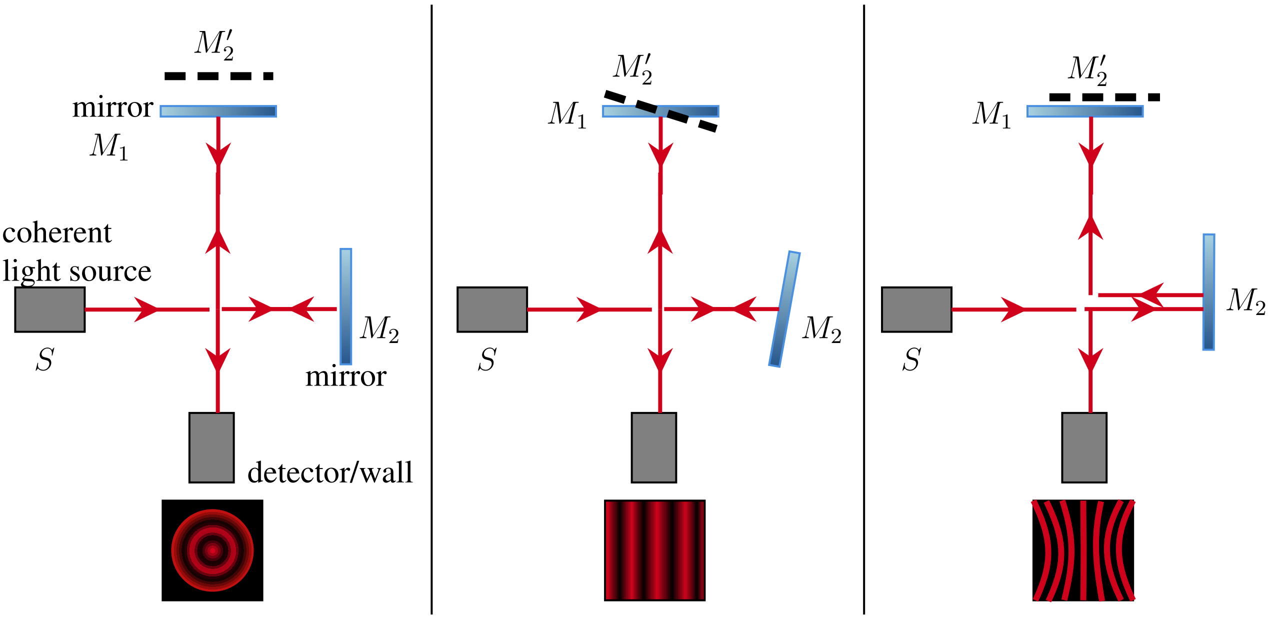
In the experiment above consider we have counted 20 times (it should be enough to have a quite small confidence interval) the crossing of the interference pattern between our stickers and that the micrometer knob shows that we have turned it from its initial position by $d \cong 6.5\,[\mu\text{m}]$.

Therefore:

$$\lambda = \frac{2d}{N} = \frac{2 \cdot 6.5\,[\mu\text{m}]}{20} \cong 0.65\,[\mu\text{m}] \tag{10.6.321}$$

and red has indeed a wavelength ranging from about 620 to 750 nanometers.

Now let us deal with the mathematical explanation. But keep in mind that in reality there can be three types of Haidinger fringes as illustrated below:



---

[25] Remember that the standard deviation does not become lower when the number of measurement grows but so do the confidence interval as it depends on $\sigma/\sqrt{n}$.

10.6.5.1.1 Case of circular Haidinger fringes

A point source produces a spherical wave. If the light from two point sources overlaps, the interference pattern maps out the way in which the phase difference between the two waves varies in space. This depends on the wavelength and on the separation of the point sources. The figure to the right shows interference between two spherical waves. The wavelength increases from top to bottom, and the distance between the sources increases from left to right.

Now remember that we have proven during our introduction of the homogeneous wave equation (see page **??**) that one possible solution in the Fresnel approximation (see page 4155) was:

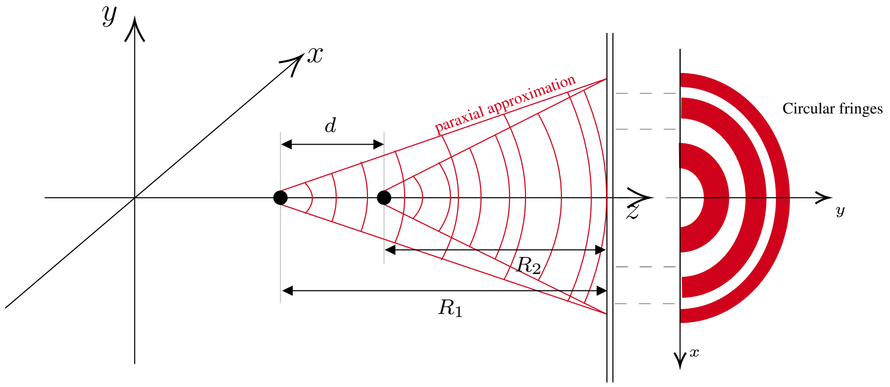
$$E_i(\rho,t)\cong\frac{A}{R}e^{\mathrm{i}\left(\omega t-k\left(R_i+\rho^2/2R_i\right)\right)} \tag{10.6.322}$$

here A is the amplitude, R is an average distance, ω is the angular frequency, $\boldsymbol{k}$ is the wave vector $k=2\pi/\lambda$, and $\rho=(x^2+y^2)^{1/2}$. The subscript i is specific to each source if there are multiple one.

Hence:

$$E_i(\rho,t)\cong\frac{A}{R}e^{\mathrm{i}\omega t}e^{-\mathrm{i}k\left(R_i+\rho^2/2R_i\right)} \tag{10.6.323}$$

Now let us consider two sources with different R_i aligned on the same axis as illustrated below (this is an equivalent to the Michelson interferometer):



Therefore the phase difference is:

$$\begin{aligned}\Delta\phi&=k\left(R_2+\frac{\rho^2}{2R_2}\right)-k\left(R_1+\frac{\rho^2}{2R_1}\right)=k(R_2-R_1)+\frac{\rho^2k}{2}\left(\frac{1}{R_2}-\frac{1}{R_1}\right)\\&=\frac{2\pi}{\lambda}(R_2-R_1)+\frac{\rho^22\pi}{2\lambda}\left(\frac{1}{R_2}-\frac{1}{R_1}\right)=\frac{2\pi}{\lambda}d+\frac{\rho^2\pi}{\lambda}\left(\frac{1}{R_2}-\frac{1}{R_1}\right)\\&=\frac{2\pi}{\lambda}d+\frac{\rho^2\pi}{\lambda}\frac{R_1-R_2}{R_1R_2}=\frac{2\pi}{\lambda}d-\frac{\rho^2\pi}{\lambda}\frac{d}{R_1R_2}\end{aligned} \tag{10.6.324}$$

Now let us consider common approximation that is $R_1R_2\cong R^2$, hence:

$$\Delta\phi=\frac{2\pi}{\lambda}d-\frac{\rho^2\pi}{\lambda}\frac{d}{R^2} \tag{10.6.325}$$

from the relation above where for reminder ρ is the distance from the optical axis in the observation plane, we see immediately that $\Delta\phi$ depends only on ρ!

Let the phase difference at the center ($\rho = 0$) be a reference state. The phase shift relative to the center, or the change in phase as you move outward radially by distance ρ, is governed by the second term:

$$\delta\phi(\rho) = \frac{\rho^2\pi}{\lambda}\frac{d}{R^2} \tag{10.6.326}$$

The fact that the phase difference above depends only on the radial distance ρ from the optical axis - not on the azimuthal angle, immediately tells you that curves of constant phase (i.e., fringes) must be circles centered on the axis, since any two points with the same ρ have identical $\delta\phi$ regardless of direction!

If we suppose can constructive or destructive interference may happen (suppose we don't already know the experimental result) then bright (or dark) rings occur wherever the total phase difference is an integer multiple of 2π :

$$\delta\phi = 2\pi m, \quad m \in \mathbb{Z} \tag{10.6.327}$$

Hence we can write:

$$2\pi m = \frac{\rho_m^2\pi}{\lambda}\frac{d}{R^2} \tag{10.6.328}$$

Rearranging:

$$\rho_m = \sqrt{m\left(\frac{2\lambda R^2}{d}\right)} \tag{10.6.329}$$

What is commonly denoted:

$$r_n = \sqrt{\frac{2\lambda n R^2}{d}}, \quad n \in \mathbb{Z}_+ \tag{10.6.330}$$

Remark

If both spherical source are not aligned the setup is therefore equivalent to the one of the double slit experiment (see section Wave Optics page 4509) and then we fall back - if the projection is made on a small finite horizontal screen - on a range of linear interference fringes.

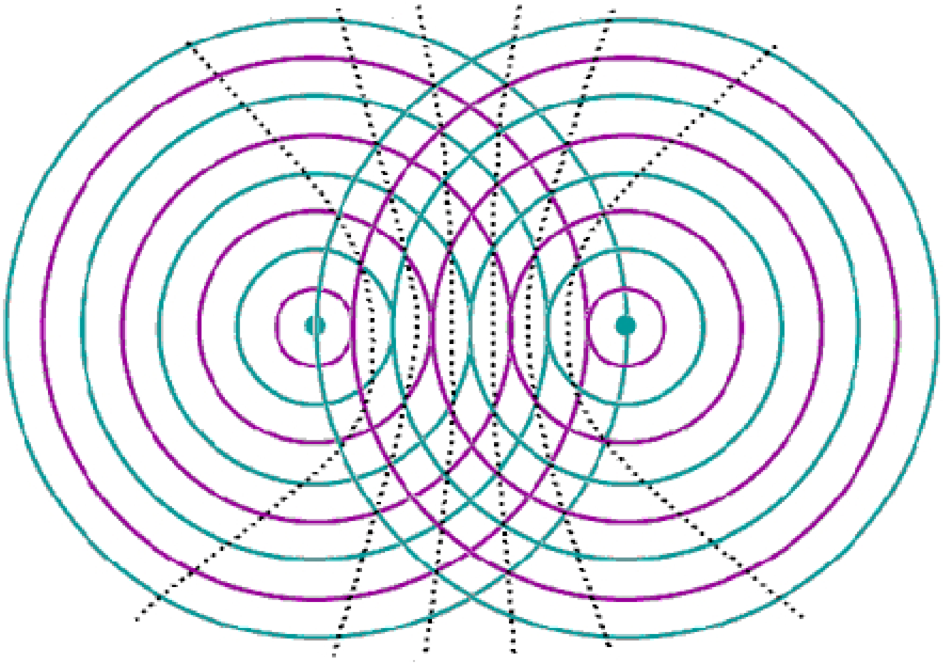
10.6.5.1.2 Case of hyperbolic Haidinger fringes

Haidinger fringes are interference fringes formed by the interference of monochromatic and coherent light to form visible dark and bright fringes. Fringe localization is the region of space where fringes with reasonably good contrast are observed.

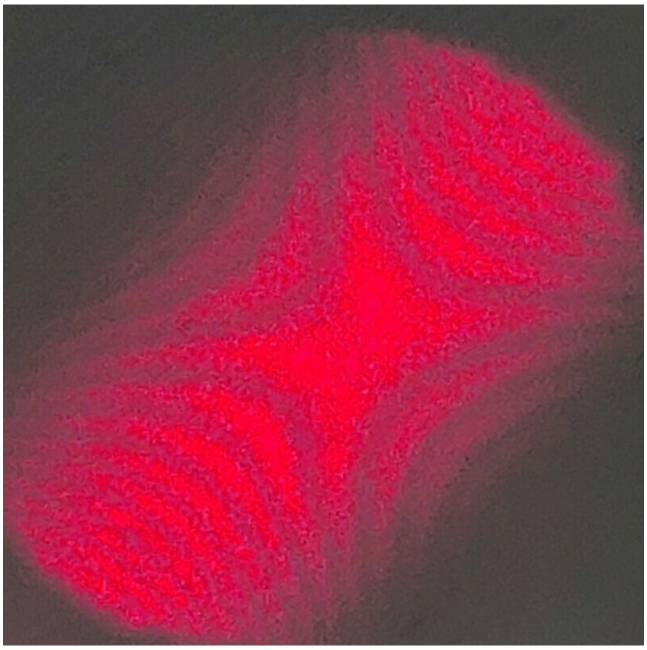
Haidinger fringes are fringes localized at infinity. Also known as fringes of equal inclination, these fringes result when light from an extended source falls on a thin film made of an optically denser medium. These fringes indicate the positions where light interferes, emerging from the medium at an equal angle.

There is almost no chance to apply two or even many independent sources to get interference patterns which are stable in time! Therefore, the interfering beams must have their origin in the same source.

Introducing the visibility or contrast (an interactive animation is available here: `https://www.walter-fendt.de/html5/phen/interference_en.htm`):



The hyperbolic interference pattern shown is typical of interferometers where there is a slight tilt of a mirror or a small angular difference between the interfering beams. Such patterns can be used to adjust interferometers, to analyze optical path length changes and to precisely measure surfaces or refractive index changes:



Consider two point sources, S_1 and S_2, separated by a distance d emitting spherical waves of wavelength λ.

At any arbitrary observation point $P(x, y, z)$ in 3D space, let r_1 and r_2 be the distances from S_1 and S_2 to P. Constructive interference (a bright fringe of order m) occurs when the path difference Δs is an integer multiple of the wavelength:

$$|r_1 - r_2| = k\lambda = C \quad (\text{ where } k \in \mathbb{Z} \text{ and } C \leq d) \tag{10.6.331}$$

For a given fringe order k, the path length difference C is a constant.

By definition, the locus of points P in space whose absolute difference in distances to two fixed points (foci) is a constant C forms a hyperboloid of revolution (see section Analytical Geometry page 2789).

Since the two sources S_1 and S_2 act as the foci:

- 3D Space: A fringe of constant phase difference forms a 3D hyperboloid of revolution around the axis containing S_1 and S_2.
- 2D Cross-section: The intersection of this 3D hyperboloid with a planar screen produces 2D hyperbolic curves.

Indeed, let us place S_1 and S_2 along the z-axis, centered symmetrically at the origin:

$$S_1 = \left(0, 0, -\frac{d}{2}\right) \qquad S_2 = \left(0, 0, \frac{d}{2}\right) \tag{10.6.332}$$

The distances from S_1 and S_2 to $P(x,y,z)$ are:

$$\begin{aligned}r_1&=\sqrt{x^2+y^2+\left(z+\frac{d}{2}\right)^2}\\r_2&=\sqrt{x^2+y^2+\left(z-\frac{d}{2}\right)^2}\end{aligned}\tag{10.6.333}$$

Setting $r_1-r_2=C$:

$$r_1=C+r_2\tag{10.6.334}$$

We square both sides:

$$r_1^2=C^2+2Cr_2+r_2^2\Longrightarrow r_1^2-r_2^2=C^2+2Cr_2\tag{10.6.335}$$

We calculate $r_1^2-r_2^2$:

$$r_1^2-r_2^2=\left[x^2+y^2+\left(z+\frac{d}{2}\right)^2\right]-\left[x^2+y^2+\left(z-\frac{d}{2}\right)^2\right]=2zd\tag{10.6.336}$$

We substitute this back into the squared equation:

$$2zd=C^2+2Cr_2\Longrightarrow r_2=\frac{zd}{C}-\frac{C}{2}\tag{10.6.337}$$

We square both sides again:

$$r_2^2=\left(\frac{zd}{C}-\frac{C}{2}\right)^2\tag{10.6.338}$$

But as :

$$r_2=\sqrt{x^2+y^2+\left(z-\frac{d}{2}\right)^2}\tag{10.6.339}$$

We have therefore:

$$x^2+y^2+\left(z-\frac{d}{2}\right)^2=\frac{z^2d^2}{C^2}-zd+\frac{C^2}{4}\tag{10.6.340}$$

Expand the left side and cancel $-zd$ from both sides:

$$x^2+y^2+z^2+\frac{d^2}{4}=\frac{z^2d^2}{C^2}+\frac{C^2}{4}\tag{10.6.341}$$

Rearrange terms to group z^2 :

$$\left(\frac{d^2}{C^2}-1\right)z^2-(x^2+y^2)=\frac{d^2-C^2}{4}\tag{10.6.342}$$

Hence:

$$\left(\frac{d^2-C^2}{C^2}\right)z^2-(x^2+y^2)=\frac{d^2-C^2}{4}\tag{10.6.343}$$

We divide through by $(d^2-C^2)/4$:

$$\frac{z^2}{\left(\frac{C}{2}\right)^2}-\frac{x^2+y^2}{\left(\frac{\sqrt{d^2-C^2}}{2}\right)^2}=1 \tag{10.6.344}$$

Let us put:

$$x=\frac{C}{2}=\frac{k\lambda}{2} \qquad b=\frac{\sqrt{d^2-C^2}}{2} \tag{10.6.345}$$

Therefore:

$$\frac{z^2}{a^2}-\frac{x^2+y^2}{b^2}=1 \tag{10.6.346}$$

This is the canonical equation of a two-sheeted circular hyperboloid of revolution symmetric about the z-axis (see section Analytical Geometry page 2841).

10.6.5.1.3 Case of straight Haidinger fringes

That's the original case of what Albert Abraham Michelson observed in his setup!

The key observation is that tilting one of the effective mirrors in a Michelson interferometer, produces two returning waves travelling in slightly different directions. Their phase difference varies linearly across the observation plane, giving straight interference fringes.

In a homogeneous medium, the scalar electric field $E(\vec{r},t)$ satisfies:

$$\nabla^2 E-\frac{1}{v^2}\frac{\partial^2 E}{\partial t^2}=0 \tag{10.6.347}$$

where $v=c/n$ is the speed of light in the medium.

For monochromatic light, let us write:

$$E(\vec{r},t)=\mathcal{R}\left\{\psi(\vec{r})e^{-\mathrm{i}\omega t}\right\} \tag{10.6.348}$$

Substitution gives the Helmholtz equation:

$$\nabla^2\psi+k^2\psi=0 \tag{10.6.349}$$

where $k=\omega/v=2\pi/\lambda$.

Its plane-wave solutions are:

$$\psi(\vec{r})=Ae^{\mathrm{i}\vec{k}\circ\vec{r}}, \quad |\vec{k}|=k \tag{10.6.350}$$

We will now derive the two returning wavevectors from the mirror geometry.

Consider, the two beams after recombination at the beam splitter. Choose z along the nominal direction of propagation toward the detector.

Suppose, the first effective mirror is perpendicular to the optical axis, while the second is tilted through a small angle α.

Reflection from a plane mirror transforms the incident wavevector according to:

$$\vec{k}_r = \vec{k}_i - 2\left(\vec{k}_i \circ \hat{\vec{n}}\right)\hat{\vec{n}} \tag{10.6.351}$$

where $\hat{\vec{n}}$ is the mirror's unit normal.

A mirror tilt of α therefore changes the reflected beam direction by 2α.

Indeed, let $\hat{\vec{n}}$ be the unit normal vector pointing outward from the surface of the mirror. Any vector, including the incident wavevector $\vec{k}_i$, can be resolved into components parallel ($\parallel$) and perpendicular $(\perp)$ to the surface normal:

$$\vec{k}_i = \vec{k}_{i,\parallel} + \vec{k}_{i,\perp} \tag{10.6.352}$$

The perpendicular component is given by the projection of $\vec{k}_i$ onto $\hat{\vec{n}}$:

$$\vec{k}_{i,\perp} = \left(\vec{k}_i \circ \hat{\vec{n}}\right)\hat{\vec{n}} \tag{10.6.353}$$

Consequently, the parallel component is whatever remains:

$$\vec{k}_{i,\parallel} = \vec{k}_i - \left(\vec{k}_i \circ \hat{\vec{n}}\right)\hat{\vec{n}} \tag{10.6.354}$$

For a wave to satisfy boundary conditions continuously across an infinite plane interface at all points along the surface, the spatial phase of the incident wave must match the spatial phase of the reflected wave along the boundary. This requires the parallel component of the wavevector to be invariant under reflection:

$$\vec{k}_{r,\parallel} = \vec{k}_{i,\parallel} = \left(\vec{k}_i \circ \hat{\vec{n}}\right)\hat{\vec{n}} \tag{10.6.355}$$

Because the reflection is elastic in a uniform medium, the magnitude of the wavevector is conserved:

$$||\vec{k}_r|| = ||\vec{k}_i|| \tag{10.6.356}$$

Since the total magnitude squared is the sum of the squares of its orthogonal components ($||\vec{k}||^2 = ||\vec{k}_\parallel||^2 + ||\vec{k}_\perp||^2$), and the parallel component $\vec{k}_\parallel$ is identical for both incident and reflected waves, the magnitude of the perpendicular component must also be identical:

$$||\vec{k}_{r,\perp}|| = ||\vec{k}_{i,\perp}|| \tag{10.6.357}$$

Physically, the normal component must reverse its direction (pointing into the opposite half-space away from the mirror):

$$\vec{k}_{r,\perp} = -\vec{k}_{i,\perp} = -\left(\vec{k}_i \circ \hat{\vec{n}}\right)\hat{\vec{n}} \tag{10.6.358}$$

The total reflected wavevector k_r is the sum of its parallel and perpendicular components:

$$\vec{k}_r = \vec{k}_{r,\parallel} + \vec{k}_{r,\perp} \tag{10.6.359}$$

Substituting from our previous expressions:

$$\vec{k}_r = \vec{k}_{r,\parallel} + \vec{k}_{r,\perp} = \left[\vec{k}_i - \left(\vec{k}_i \circ \hat{\vec{n}}\right)\hat{\vec{n}}\right] - \left(\vec{k}_i \circ \hat{\vec{n}}\right)\hat{\vec{n}} = \vec{k}_i - 2\left(\vec{k}_i \circ \hat{\vec{n}}\right)\hat{\vec{n}} \tag{10.6.360}$$

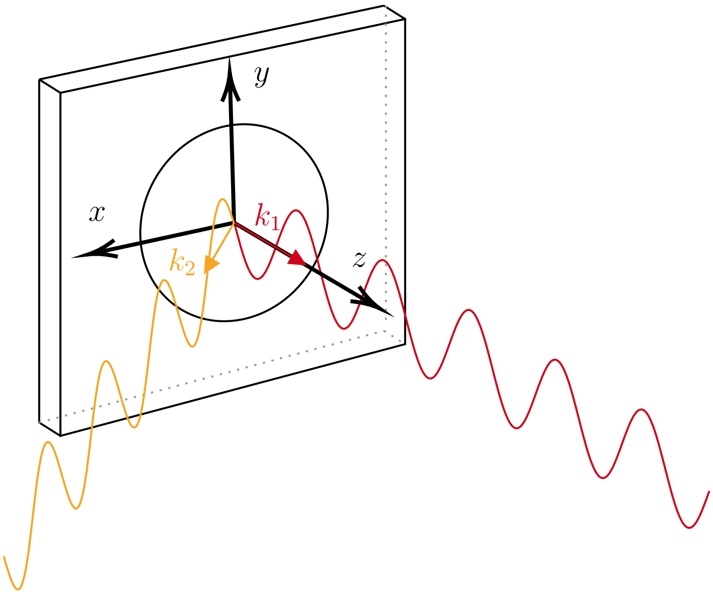
This reflection law follows by matching the phase of the incident and reflected solutions of the Helmholtz equation along the mirror surface.

After accounting for the beam splitter, we can choose coordinates such that the two recombined wavevectors are:

$$\vec{k}_1 = k(0,0,1) \tag{10.6.361}$$

and:

$$\vec{k}_2 = k(\sin(2\alpha), 0, \cos(2\alpha)) \tag{10.6.362}$$



Notice that we have not assumed anything about the shape of the fringes. We have obtained the wavevectors from the optical geometry.

The complex fields are:

$$\psi_1 = A_1 e^{\mathrm{i}kz} \tag{10.6.363}$$

and:

$$\psi_2 = A_2 e^{\mathrm{i}[kx\sin(2\alpha) + kz\cos(2\alpha) + \phi_0]} \tag{10.6.364}$$

Here ϕ_0 represents a constant phase difference arising from the optical path lengths and reflections.

The resultant field is:

$$\psi = \psi_1 + \psi_2 \tag{10.6.365}$$

The time-averaged intensity is proportional to:

$$\begin{aligned}I &= |\psi_1+\psi_2|^2 = |\psi_1+\psi_2|^2 = (\psi_1+\psi_2)\left(\bar{\psi}_1+\bar{\psi}_2\right) = \psi_1\bar{\psi}_1+\psi_2\bar{\psi}_2+\psi_1\bar{\psi}_2+\bar{\psi}_1\psi_2\\ &= \left(A_1e^{\mathrm{i}kz}\right)\left(A_1e^{-\mathrm{i}kz}\right)+\left(A_2e^{\mathrm{i}[kx\sin(2\alpha)+kz\cos(2\alpha)+\phi_0]}\right)\left(A_2e^{-\mathrm{i}[kx\sin(2\alpha)+kz\cos(2\alpha)+\phi_0]}\right)\\ &+\left(A_1e^{\mathrm{i}kz}\right)\left(A_2e^{-\mathrm{i}[kx\sin(2\alpha)+kz\cos(2\alpha)+\phi_0]}\right)+\left(A_1e^{-\mathrm{i}kz}\right)\left(A_2e^{\mathrm{i}[kx\sin(2\alpha)+kz\cos(2\alpha)+\phi_0]}\right)\\ &= A_1^2+A_2^2+A_1A_2\left(e^{\mathrm{i}kz-\mathrm{i}[kx\sin(2\alpha)+kz\cos(2\alpha)+\phi_0]}\right)\\ &+A_1A_2\left(e^{\mathrm{i}[kx\sin(2\alpha)+kz\cos(2\alpha)+\phi_0]--\mathrm{i}kz}\right)\\ &= A_1^2+A_2^2+A_1A_2\left(e^{\mathrm{i}kz-\mathrm{i}[kx\sin(2\alpha)+kz\cos(2\alpha)+\phi_0]}+e^{\mathrm{i}[kx\sin(2\alpha)+kz\cos(2\alpha)+\phi_0]-\mathrm{i}kz}\right)\end{aligned} \tag{10.6.366}$$

Using Euler's formula (see section Number page 251):

$$\cos(\phi) = \frac{e^{\mathrm{i}\phi}+e^{-\mathrm{i}\phi}}{2} \tag{10.6.367}$$

we get:

$$I = A_1^2+A_2^2+2A_1A_2\cos(\Delta\phi) \tag{10.6.368}$$

where:

$$\Delta\phi = kx\sin(2\alpha)+kz[\cos(2\alpha)-1]+\phi_0 \tag{10.6.369}$$

Now consider a detector in the plane $z=z_0$.

All terms involving z_0 are constants, so that we can write:

$$I(x,y) = A_1+A_2+2A_1A_2\cos[kx\sin(2\alpha)+\phi] \tag{10.6.370}$$

where ϕ is a constant.

This is the required intensity distribution, derived from the wave equation and mirror geometry.

Bright fringes occur when the phase difference is an integer multiple of 2π :

$$kx\sin(2\alpha)+\phi = 2\pi m \tag{10.6.371}$$

where m is an integer.

Therefore:

$$x_m = \frac{2\pi m-\phi}{k\sin(2\alpha)} \tag{10.6.372}$$

Since x_m is independent of y, every bright fringe is a straight line parallel to the y-axis.

The fringe spacing is:

$$\Delta x = \frac{2\pi}{k|\sin(2\alpha)|} \tag{10.6.373}$$

Using $k=2\pi/\lambda$:

$$\Delta x = \frac{\lambda}{|\sin(2\alpha)|} \tag{10.6.374}$$

For small angles, $\sin(2\alpha)\simeq 2\alpha$, so:

$$\Delta x \cong \frac{\lambda}{2|\alpha|} \tag{10.6.375}$$

10.6.5.1.4 Michelson interferometer and Special Relativity

At the origin this experiment was devised by Albert Abraham Michelson to measure the speed of light within its hypothesized medium (the ether), based on the classical law of the addition of velocities. It appeared that the Earth - moving in its orbit at a speed of approximately 30 $[\text{km}\cdot\text{s}^{-1}]$ relative to the Sun-provided an ideal laboratory for detecting variations in the speed of light over paths of equal length; such paths, however, were expected to require different amounts of time to traverse, depending on whether they were aligned with the direction of motion or oriented perpendicular to the "ether wind".

The apparatus is constructed in such a way that mirrors M_1 and M_2 are at an equal distance, d, from the beam splitter.

So, if the Earth is stationary relative to the ether, the two paths - in the two perpendicular directions - are equal (the same distance is traveled, and the travel time is the same).

If, on the other hand, the Earth is in motion relative to the ether - at a speed v in a certain direction (toward M_1, for example) - then the two journeys are not made at the same speed, and the time required is not the same in both directions:

- A round trip in the direction of the Earth's motion (toward M_1) requires a certain amount of time:

  $$t_1=\frac{d}{c-v}+\frac{d}{c+v}=\frac{2d\cdot c}{c^2-v^2}=\frac{2d}{c}\frac{1}{1-\frac{v^2}{c^2}}\cong\frac{2L}{c}\left(1+\frac{v^2}{c^2}\right) \tag{10.6.376}$$

  Michelson obtained this expression correctly in 11881 (holocene calendar), however, in transverse direction that we will derive now he obtained the incorrect expression (this was corrected by Alfred Potier on year later and Hendrik Lorentz five yers later).

- A round trip perpendicular to the Earth's motion (towards M_2) requires a certain amount of time t_2. This inclined travel path follows from the transformation from the interferometer rest frame to the aether rest frame. One direction path will have length using Pythagoras theorem:

  $$ct=\sqrt{d^2+(vt)^2} \tag{10.6.377}$$

  and consequently the one direction travel time is after rearranging:

  $$t=\frac{d}{\sqrt{c^2-v^2}} \tag{10.6.378}$$

  which is the same for the backward journey. The total travel time is therefore equal to:

  $$t_2=\frac{2d}{\sqrt{c^2-v^2}}=\frac{2d}{c}\frac{1}{\sqrt{1-\frac{v^2}{c^2}}}\cong\frac{2d}{c}\left(1+\frac{v^2}{2c^2}\right) \tag{10.6.379}$$

The time difference between t_1 and t_2 is given by:

$$t_1 - t_2 = \frac{2L}{c}\left(\frac{1}{1-\frac{v^2}{c^2}} - \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}\right) \tag{10.6.380}$$

The expected time difference is too small to be measured directly: the speed of light was already known to be on the order of $3\cdot 10^8\,[\mathrm{m}\cdot\mathrm{s}^{-1}]$, the speed of the Earth relative to the Sun - and thus relative to the ether - on the order of $3\cdot 10^4\,[\mathrm{m}\cdot\mathrm{s}^{-1}]$, and the dimensions of the apparatus on the order of 10 [m]. This results in a measurement on the order of $10\cdot 10^9/10^{25}$, or 10^{-15} seconds.

Son instead we use the the path difference, simply by multiplying by c:

$$\Delta\lambda = 2L\left(\frac{1}{1-\frac{v^2}{c^2}} - \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}\right) \tag{10.6.381}$$

The path difference is denoted by $\Delta\lambda$ because the beams are out of phase by a some number of wavelengths (λ). To visualise this, consider taking the two beam paths along the longitudinal and transverse plane, and lying them straight. One path will be longer than the other, this distance is $\Delta\lambda$.

If the relation $v^2/c^2 \ll 1$ is true (if the velocity of the aether is small relative to the speed of light), then the expression can be simplified using a first order binomial expansion:

$$(1-x)^n \cong 1 - nx \tag{10.6.382}$$

So, rewriting the above in terms of powers:

$$\Delta\lambda = 2d\left(\left(1-\frac{v^2}{c^2}\right)^{-1} - \left(1-\frac{v^2}{c^2}\right)^{-1/2}\right) \tag{10.6.383}$$

Applying binomial simplification:

$$\Delta\lambda \cong 2d\left(\left(1+\frac{v^2}{c^2}\right) - \left(1+\frac{v^2}{2c^2}\right)\right) = 2d\frac{v^2}{2c^2} = d\frac{v^2}{c^2} \tag{10.6.384}$$

The derivation above shows that the presence of an aether wind would produce a difference in optical path lengths between the two arms of the interferometer. This path difference depends on the orientation of the interferometer relative to the aether wind. Specifically, the derivation assumes that the longitudinal arm is aligned parallel to the presumed direction of the aether wind. If instead the longitudinal arm is oriented perpendicular to the aether wind, the resulting path difference would have the opposite sign.

The magnitude of the path difference can vary continuously and may represent any fraction of the wavelength, depending on both the angle between the apparatus and the aether wind and the wind's speed.

In fact, this is related to a whole series of experiments conducted between 11881 (by Michelson alone) and 11887 (by both of them together). By the latter date, it was definitively accepted - a conclusion that has been consistently reaffirmed whenever the experiment is repeated with the benefit of technical advancements allowing for greater precision - that the speed of light is constant.

The negative result led Michelson to the conclusion that there is no measurable aether drift. However, he never accepted this on a personal level, and the negative result haunted him for the rest of his life.

The interpretation of this result led physicists to question the existence of the ether (hypothesized to be the material medium for the vibrations of electromagnetic waves, such as light) - or, at the very least, to question its motion. It also demonstrated that the speed of light is the same in all directions, up to the second order in (v/c), which was the level of precision of the experiment.

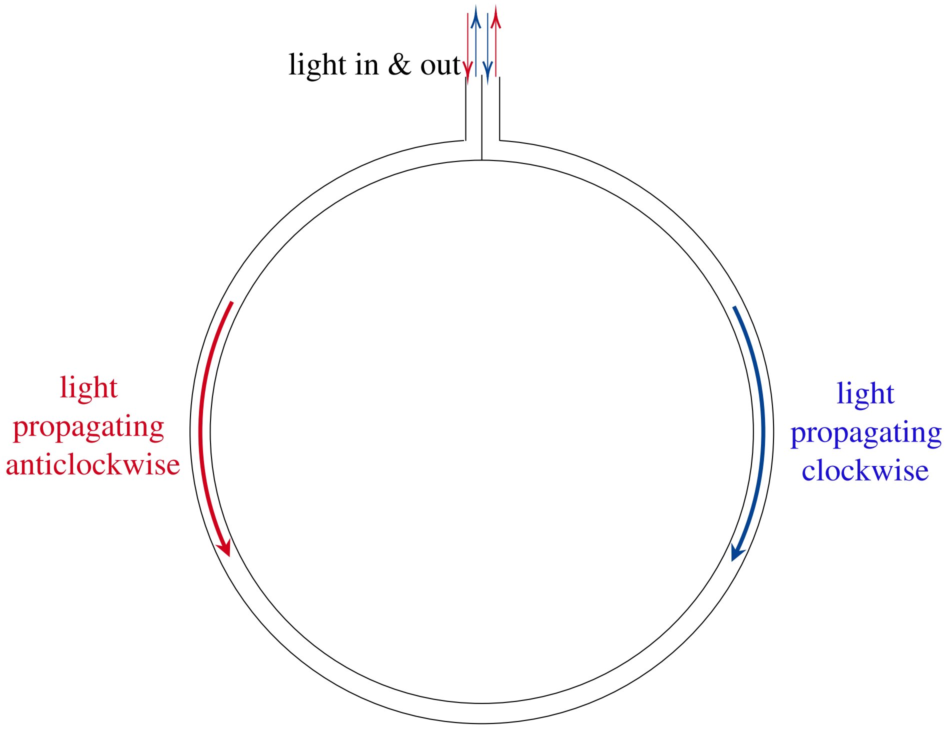
Remark

The extent to which the null result of the Michelson-Morley experiment influenced Albert Einstein is disputed. Alluding to some statements of Albert Einstein, many historians argue that it played no significant role in his path to special relativity, while other statements of Einstein probably suggest that he was influenced by it. In any case, the null result of the Michelson-Morley experiment helped the notion of the constancy of the speed of light gain widespread and rapid acceptance.

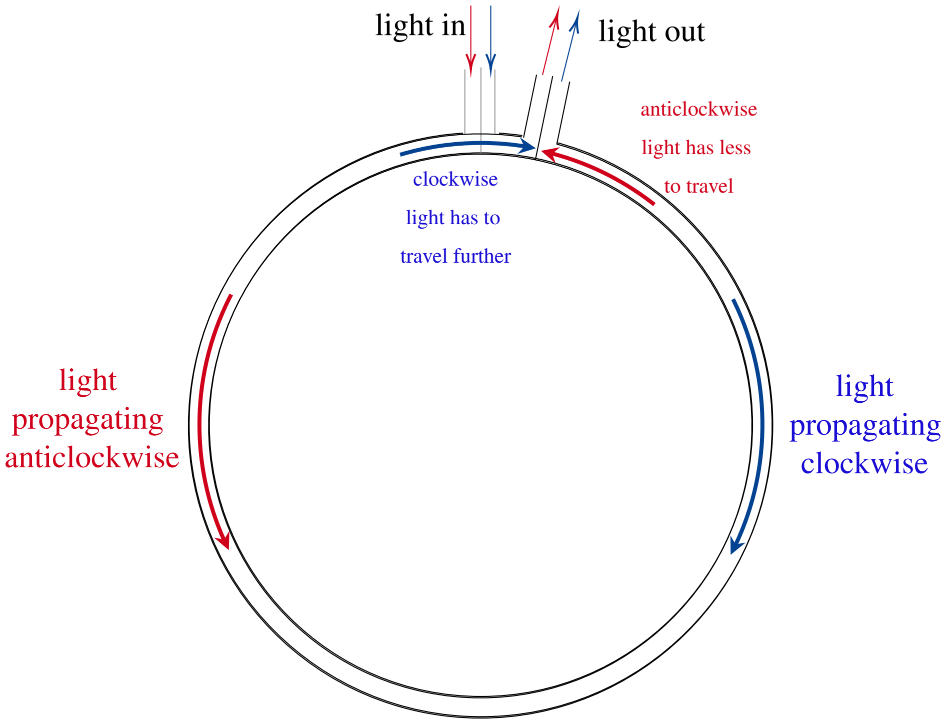
10.6.5.2 Sagnac interferometer (Sagnac effect)

The "Sagnac effect", also named "Sagnac interference", named after French physicist Georges Sagnac, is a phenomenon encountered in interferometry that is elicited by rotation. The Sagnac effect manifests itself in a setup called a ring interferometer or Sagnac interferometer. A beam of light is split and the two beams are made to follow the same path but in opposite directions. On return to the point of entry the two light beams are allowed to exit the ring and undergo interference. The relative phases of the two exiting beams, and thus the position of the interference fringes, are shifted according to the angular velocity of the apparatus. In other words, when the interferometer is at rest with respect to a nonrotating frame, the light takes the same amount of time to traverse the ring in either direction. However, when the interferometer system is spun, one beam of light has a longer path to travel than the other in order to complete one circuit of the mechanical frame, and so takes longer, resulting in a phase difference between the two beams. Georges Sagnac set up this experiment in 11913 (holocene calendar) in an attempt to prove the existence of the aether that Einstein's theory of special relativity makes superfluous.

Imagine beams of light coming from an emitter and travelling around a circular path in both directions, until they arrive back at the source. Such an arrangement can be constructed by using an optic fibre in a circular loop, injecting light at both ends. The distance travelled by the clockwise beam is the same as the distance travelled by the anticlockwise beam, and the speed of light in both directions is the same, so the time taken for each beam to travel from the source back to the origin is the same. So far, so good:



Now imagine the whole thing is rotating - let's say clockwise. For reference we will use the numbers on a clock face and the finer divisions into 60 minutes. The optic fibre ring runs around the edge of the clock, with the light source and a detector at 12. Now imagine that the clock rotates fast enough that by the time the clockwise-going light reaches the original 12 position, the clock has rotated so that 12 is now located at the original 1 minute past 12 position. The light has to travel an extra 60th of the circle to reach its starting position (actually a tiny bit more than that because the clock is still rotating and will have gone a tiny bit further by the time the light beam catches up). But the light going anticlockwise reaches the source early, only needing to travel a tiny bit more than 59/60 of the circle. The travel times of the two beams of light around the circle are different!



This is a very simplified explanation, and figuring out the mathematics of exactly what happens involves using special relativity, since the speed of light is involved, but it can be shown that there is indeed a time difference between the travel times of beams of light heading in opposite directions around a rotating loop. The time difference is proportional to the speed of rotation and to the area of the loop (and to the cosine of the angle between the rotation axis and the perpendicular to the loop, for those who enjoy vector mathematics).

The shift in interference fringes in a ring interferometer can be viewed intuitively and naively as a consequence of the different distances that light travels due to the rotation of the ring. The simplest derivation is for a circular ring of radius R, with a refractive $n = 1$, rotating at an angular velocity of ω, but the result is general for loop geometries with other shapes. If a light source emits in both directions from one point on the rotating ring, light travelling in the same direction as the rotation direction needs to travel more than one circumference around the ring before it catches up with the light source from behind. The time t_1 that it takes to catch up with the light source is given by:

$$t_1 = \frac{2\pi R + \Delta L}{c} \qquad (10.6.385)$$

where ΔL is the distance that the mirror has moved in that same time:

$$\Delta L = R\omega t_1 \qquad (10.6.386)$$

Eliminating ΔL from the two equations above we get:

$$t_1 = \frac{2\pi R}{c - R\omega} \qquad (10.6.387)$$

Likewise, the light travelling in the opposite direction of the rotation will travel less than one circumference before hitting the light source on the front side. So the time for this direction of light to reach the moving source again is:

$$t_2=\frac{2\pi R}{c+R\omega} \tag{10.6.388}$$

The time difference is therefore:

$$\Delta T=t_1-t_2=\frac{4\pi R^2\omega}{c^2-R^2\omega^2} \tag{10.6.389}$$

For $R\omega=v\ll c$, this reduces to the "Sagnac Newtonian formula":

$$\Delta T\cong\frac{4\pi R^2\omega}{c^2}=\frac{4A\omega}{c^2} \tag{10.6.390}$$

where A is the area of the ring. Hence the phase shift[26]:

$$\boxed{\Delta\phi=c\Delta T\frac{1}{\lambda}\cdot 2\pi\cong\frac{8\pi\omega}{c\lambda}} \tag{10.6.391}$$

Although this simple derivation is for a circular ring with an index of refraction of one, the result holds true for any shape of rotating loop with area A.

We see that this result has been achieved at the cost of introducing an anisotropy in the velocity of light and of actually allowing velocities higher than the velocity of light (the $c+R\omega$). This anisotropy contradicts the relativistic law for adding velocities (even in first order), and it contradicts the constancy of the velocity of light. In other words, although these ideas do generate a prediction which agrees (in a first approximation) with the correct prediction, it is internally untenable.

Remark

Since emitter and detector are travelling at the same speeds, Doppler effects cancel out, so the Sagnac effect does not involve the Doppler effect. In the case of ring LASER interferometry, it is important to be aware of this. When the ring LASER setup is rotating, the counterpropagating beams undergo frequency shifts in opposite directions.

Let us see now the relativist approach version!

Consider a ring interferometer where two counter-propagating light beams share a common optical path determined by a loop of an optical fiber. The loop may have an arbitrary shape, and can move arbitrarily in space. The only restriction is that it is not allowed to stretch.

Consider a small segment of the fiber, whose length in its rest frame is $d\ell'$. The time intervals, $dt'_\pm$, it takes the left and right moving light rays to traverse the segment in the rest frame coincide and are given by (n is the index of refraction):

$$\mathrm{d}t'_\pm=\frac{n}{c}\mathrm{d}\ell' \tag{10.6.392}$$

---

[26] For reminder the $c\Delta T$ give obviously the distance shift, but as the phase shift is by definition in radian in fraction of the wavelength that why divide by λ and to get the result in radian we multiply by 2π.

Let $\mathrm{d}\ell=|\mathrm{d}\vec{x}|$ be the length of this small segment in the lab frame. By the relativistic length contraction formula (see section Special Relativity page 5654)), $\mathrm{d}\ell'=\gamma\mathrm{d}\ell\cong\mathrm{d}\ell$ correct to first order in the velocity $\vec{v}$ of the segment. The time intervals $\mathrm{d}t_{\pm}$for traversing the segment in the lab frame are given by Lorentz transformation as:

$$\mathrm{d}t_{\pm}=\gamma\left(\mathrm{d}t'\pm\frac{\vec{v}\circ d\vec{x}'}{c^2}\right)\cong\frac{n}{c}\mathrm{d}\ell\pm\frac{\vec{v}}{c^2}\circ\mathrm{d}\vec{x}\tag{10.6.393}$$

correct to first order in the velocity $\vec{v}$. In general, the two beams will visit a given segment at slightly different times, but, in the absence of stretching, the length $\mathrm{d}\ell$ is the same for both beams.

It follows that the time difference for completing a cycle for the two beams is:

$$\Delta T=\int(\mathrm{d}t_{+}-\mathrm{d}t_{-})\cong\frac{2}{c^2}\oint\vec{v}\circ\mathrm{d}\vec{x}\tag{10.6.394}$$

Imagine a screen for viewing fringes placed at the light source (alternatively, use a beamsplitter to send light from the source point to the screen). Given a steady light source, interference fringes will form on the screen with a fringe displacement given by $\Delta\phi\approx\frac{2\pi c}{\lambda}\Delta T$ where the first factor is the frequency of light. This gives the generalized Sagnac formula:

$$\Delta\phi\cong\frac{4\pi}{\lambda c}\oint\vec{v}\circ\mathrm{d}\vec{x}\tag{10.6.395}$$

In the special case that the fiber moves like a rigid body with angular frequency $\vec{\omega}$, the velocity is $\vec{v}=\vec{\omega}\times\vec{x}$ and the line integral can be computed in terms of the area of the loop (we have us the property of the dot and cross product as give at page 1836):

$$\oint\vec{v}\circ\mathrm{d}\vec{x}=\oint\vec{\omega}\times\vec{x}\circ\mathrm{d}\vec{x}=\oint\vec{\omega}\circ\vec{x}\times\mathrm{d}\vec{x}=2\oint\vec{\omega}\circ\mathrm{d}\vec{A}=2\vec{\omega}\circ\vec{A}\tag{10.6.396}$$

This gives the "Sagnac formula for ring interferometers" of arbitrary shape and geometry:

$$\boxed{\Delta\phi\cong\frac{8\pi}{\lambda c}\vec{\omega}\circ\vec{A}}\tag{10.6.397}$$

Caution! According to the ether hypothesis, light is a disturbance similar to sound. It takes place in a substance called luminiferous ether. When we assume the interferometer is placed in an inertial system that is stationary with respect to the ether, then the result given by (see proof above):

$$\Delta\phi\cong\frac{8\pi\omega}{\lambda c}\tag{10.6.398}$$

is correct if the rotating platform does not drag the ether along with its motion.

On the other hand, it was already noticed by Max von Laue that relativity predicts, to first order approximation, the same kind of result as an ether theory for a Sagnac type experiment (see proof also just above):

$$\Delta\phi \cong \frac{8\pi}{\lambda c}\vec{\omega}\circ\vec{A} \qquad (10.6.399)$$

Since the Sagnac effect is compatible either with relativity or the ether hypothesis, the conclusion is that the Sagnac experiment cannot discriminate between the two models.

However, the famous 11887 (holocene calendar) Michelson-Morley experiment (see page 4562) can be explained only if the ether is fully dragged with the earth's motion. On the other hand, in 11925 (holocene calendar), Michelson and Gale confirmed the Sagnac effect using the earth as a rotating platform. This experiment can only be explained if the ether is not dragged with the earth's rotation.

Thus, neither the Michelson-Morley nor the Michelson-Gale experiments, taken separately, are evidence against the ether. However, when considered jointly, they are incompatible with the existence of the luminiferous ether since it cannot be consistently assumed the ether is fully dragged with the earth translation and, at the same time, remains unperturbed by its rotation. Nonetheless, both experiments can be explained by relativity.

The first relativistic proofs of the effect were presented by Max von Laue in 11920 (holocene calendar) and with a different approach by Paul Langevin in 11921 (holocene calendar). Presumably, it took so long to find relativistic explanations because, at the time, the experiment was not widely known.

In this early 121st century (holocene calendar), the relativistic nature of the Sagnac effect is, of course, beyond reasonable doubt. There exists a vast literature with an exhaustive analysis of the phenomenon. Its applications range from GPS, fiber optic and ring LASER gyroscopes, navigation systems to geodesic and seismology. There are even derivations based on the mathematics of General Relativity that leads to the same result (see [**lambare2024sagnac**] or [**logunov1988special**]).

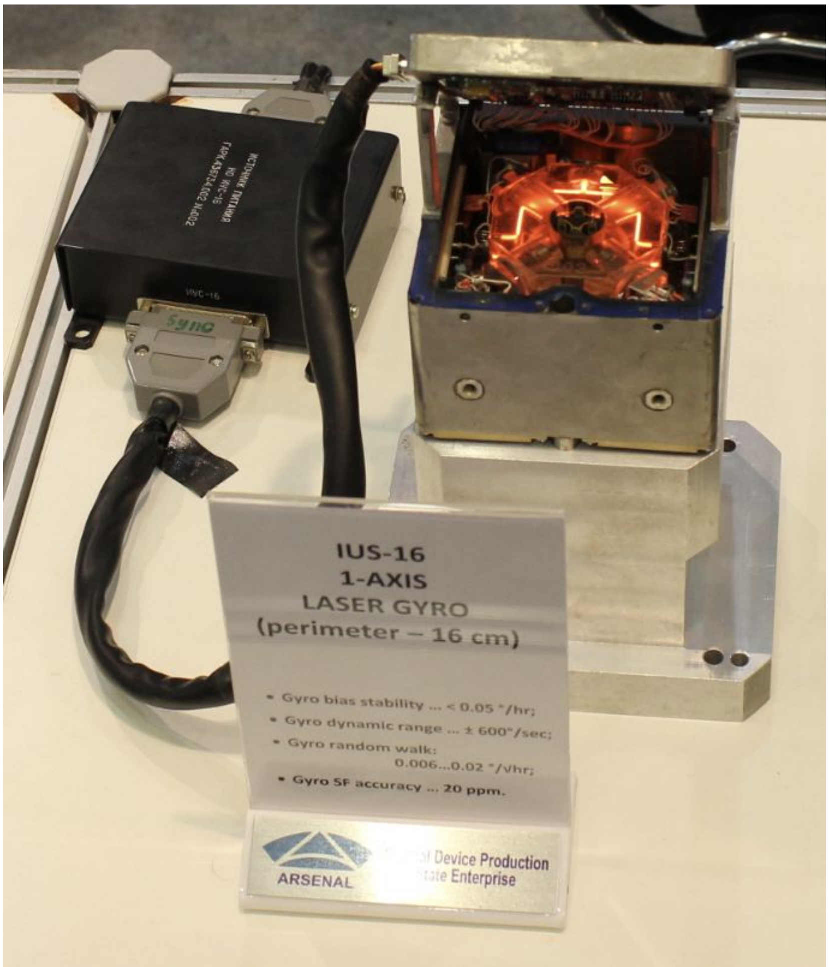

IUS-16
1-AXIS
LASER GYRO
(perimeter – 16 cm)
• Gyro bias stability ... < 0.05 °/hr;
• Gyro dynamic range ... ± 600°/sec;
• Gyro random walk: 0.006...0.02 °/√hr;
• Gyro SF accuracy ... 20 ppm.
ARSENAL
Device Production
State Enterprise
Sync

Figure 10.274: **IUS-16 - 1-axis LASER gyroscope produced by "Arsenal" company, Ukraine (author: VictorAnyakin, source: Wikipedia)**

Measuring the minuscule time difference between the propagation of the light beams is not difficult, due to the wave nature of light itself. The wavelength of visible light is just a few hundred nanometres, so even a time difference of the order of 10^{-16} seconds can be observed because it moves the wave crests and troughs of the two beams relative to one another, causing visible interference patterns as they shift out of synchronisation. This makes the device an interferometer that is very sensitive to rotational speed.

Remark

A gimbal mounted mechanical gyroscope remains pointing in the same direction after spinning up, and thus can be used as a rotational reference for an inertial navigation system. With the development of so-called LASER gyroscopes and fiber optic gyroscopes based on the Sagnac effect, bulky mechanical gyroscopes can be replaced by those with no moving parts in many modern inertial navigation systems. A conventional gyroscope relies on the principle of conservation of angular momentum whereas the sensitivity of the ring interferometer to rotation arises from the invariance of the speed of light for all inertial frames of reference.

As the Earth moves through space, it wobbles. Researchers in Germany have now directly observed this wobble with the highest precision yet thanks to a large ring LASER gyroscope they developed for this purpose (other have also measured the rotation of Earth long time before but with less precision: see [391] and [18]). The instrument, which is located in southern Germany and operates continuously, represents an important advance in the development of super-sensitive rotation sensors. If further improved, such sensors could help us better understand the interior of our planet and test predictions of relativistic effects, including the distortion of space-time due to Earth's rotation. The instrument can sense the Earth's rotation to within an accuracy of 48 parts per billion (ppb), which corresponds to picoradians per second.

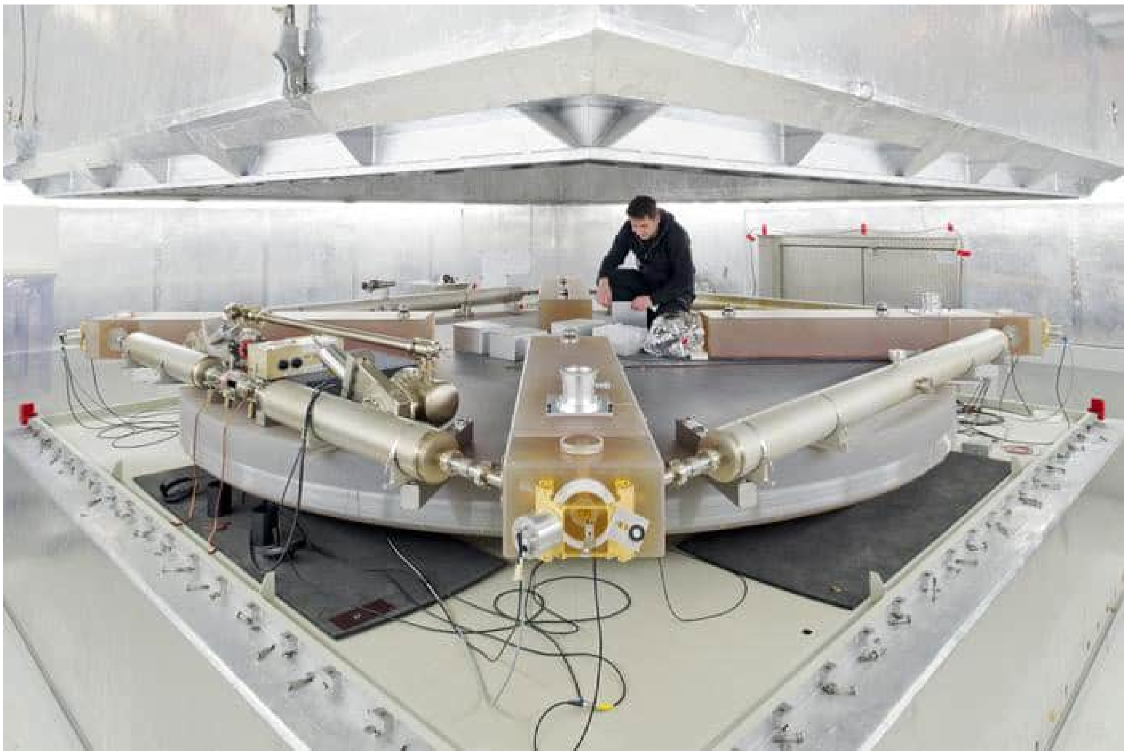


Figure 10.275: Ring LASER at the Geodetic Observatory Wettzell in Germandy (courtesy: Astrid Eckert / TUM)

10.6.6 LASER

The conceptual contribution of Albert Einstein to light-matter interaction is essential. His approach, which was to introduce the concept of absorption and stimulated emission of radiation, is the origin of the LASER emission process (Light Amplification by Stimulated Emission of Radiation) used in many areas such as: read / write optical discs, satellite data transmission, conferences pointers, precision measuring devices (telemetry, anemometry), surgery (eye, hair removal, gyrometry, cutting or industrial welding, micromachining, priming nuclear reactions (megajoule LASER) prototyping (3D printers), cleaning surfaces (ablation / LASER etching), structure analysis, cooling, etc.

By analysing the thermal equilibrium conditions in the interaction of electromagnetic radiation with matter, Albert Einstein understood that taking into account the spontaneous emission permits only to find the Wien's distribution law (see section Thermodynamics page 3727). However Planck's law (see section Thermodynamics page 3719) can not be obtained if we assume the existence of a stimulated emission process. This is where Einstein's work germ contain the development of coherent electromagnetic radiation sources.

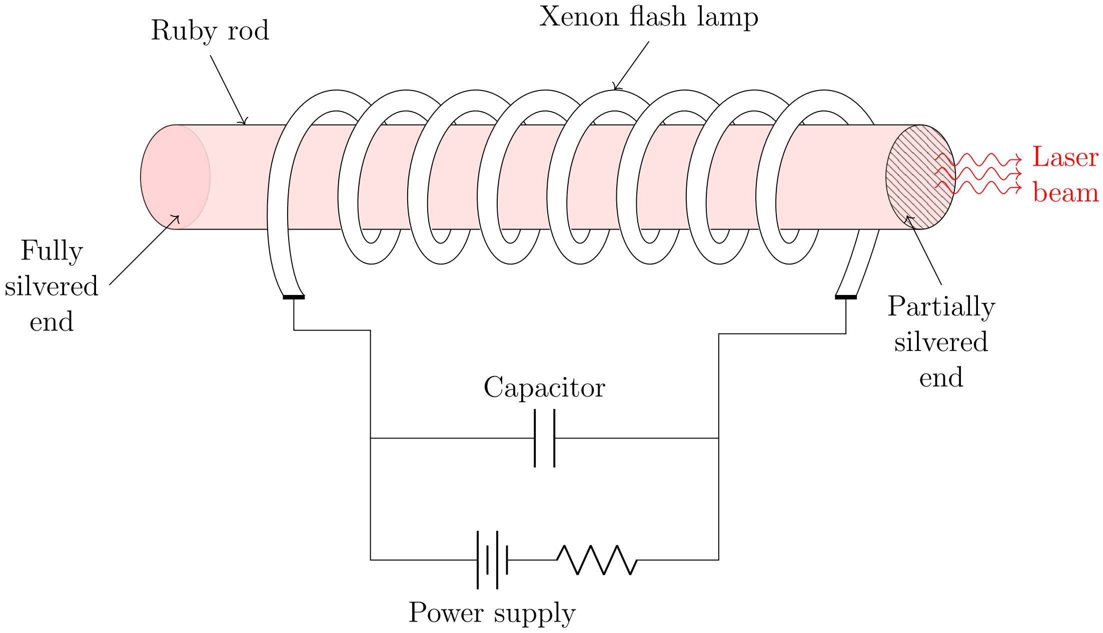


Figure 10.276: **Schematic principle of a ruby LASER (solid-state LASER)**

Indeed, the first proofs of a coherent emission of radiation have involved specialists in atomic physics to those of electromagnetism. This gave birth to the MASER (Microwave Amplification by Stimulated Emission of Radiation).

It is the evolution of this multifaceted tool and the response that it has be able to bring to problems that contributed to the development of optoelectronics, optronics and biophotonics.

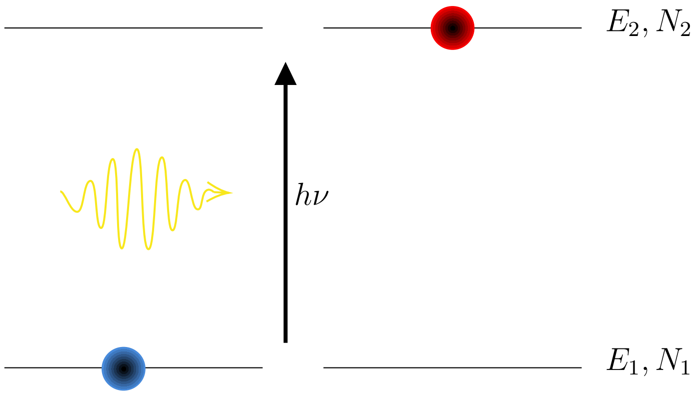
I have hesitated during a long time to put the basic developments of the LASER in the section of Statistical Mechanics, or Thermodynamics or Electrodynamics. As in my private circle the majority of people (not scientific) associated the LASER with optics, I thought it was more appropriate to present the below developments in this section of Wave Optics.

We will, as did Albert Einstein, show by contradiction that consider only stimulated absorption and spontaneous emission are insufficient to fall back on the law Planck proved in the section Thermodynamics (law which describes the radiation density at equilibrium in a cavity of finite dimension):

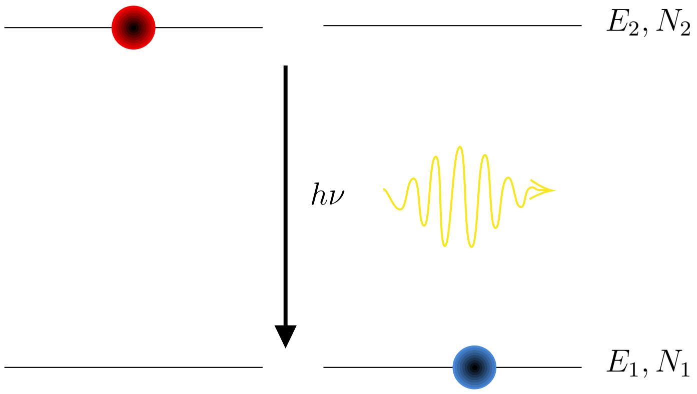
$$R(\nu,T)=\frac{8\pi h\nu^3}{c^3}\frac{1}{e^{\frac{h\nu}{kT}}-1} \tag{10.6.400}$$

Let us represent schematically what are stimulated absorption and spontaneous emission disregarding the notations of energy levels as used in the sections of Corpuscular Quantum Physics, Wave Quantum Physics and Quantum Chemistry:

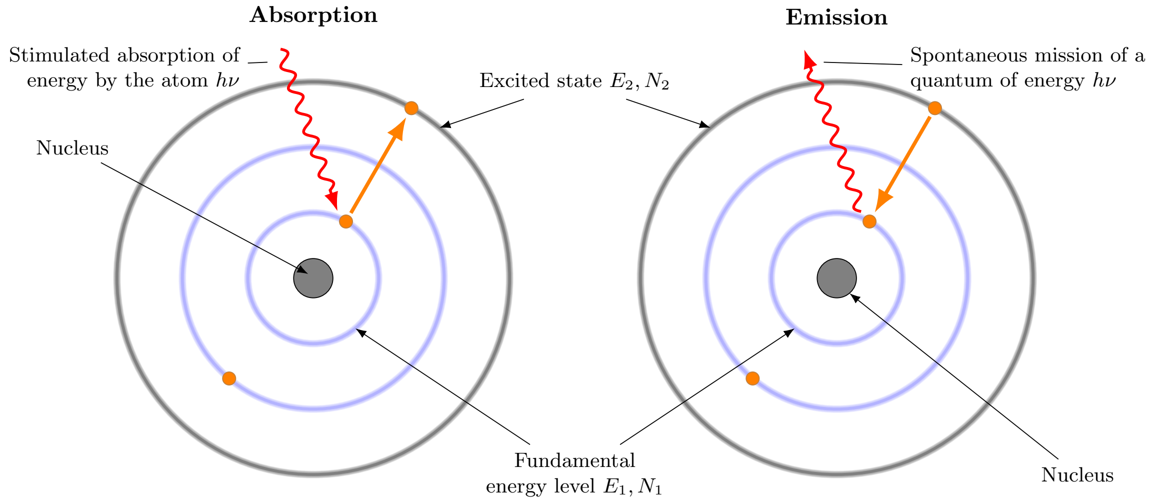
stimulated absorption



spontaneous emmission



Or seen from a naive atomistic level point of view (the old Bohr model...):


Absorption
Stimulated absorption of energy by the atom $h\nu$
Nucleus
Excited state E_2, N_2
Emission
Spontaneous mission of a quantum of energy $h\nu$
Fundamental energy level E_1, N_1
Nucleus

In our gedankenexperiment, N_{tot} atoms are divided into the populations N_1 and N_2 respectively on two energy levels E_1 and E_2.

We observe a resonant absorption of electromagnetic radiation when the frequency ν of the radiation is equal to the energy difference between the two considered levels. The energy emitted or absorbed is then linked to the relation:

$$E_2 - E_1 = h\nu \tag{10.6.401}$$

The absorption rate obviously depends on the spectral energy density of the incident electromagnetic field $R(\nu, T)$, of the population N_1 of the lower level and finally it is probably proportional to a factor B_{12} that would give the properties of the atomic system. Under these assumptions, we are led to write naturally:

$$\left.\frac{\mathrm{d}N_2}{\mathrm{d}t}\right|_{\text{absorption}} = N_1 R(\nu, T) B_{12} = -\frac{\mathrm{d}N_1}{\mathrm{d}t}_{\text{absorption}} \tag{10.6.402}$$

The phenomenon of stimulated resonant absorption is superimposed the fact that an excited system returns to its initial state with a characteristic time: the lifetime of the excited state. The de-excitation rate is proportional to the population of the upper level of the transition such as this brings us to write:

$$\left.\frac{\mathrm{d}N_2}{\mathrm{d}t}\right|_{\text{spontaneous}} = -N_2 A_{21} = \frac{\mathrm{d}N_1}{\mathrm{d}t}_{\text{spontaneous}} \tag{10.6.403}$$

The conservation of the total number of atoms:

$$N_1(t) + N_2(t) = N_{\text{tot}} \tag{10.6.404}$$

interacting with the radiation is reflected by the kinetic condition (it is just the derivative with respect to time of the preceding expression):

$$\frac{\mathrm{d}N_1(t)}{\mathrm{d}t} + \frac{\mathrm{d}N_2(t)}{\mathrm{d}t} = N_{\text{tot}} \tag{10.6.405}$$

By injecting previous relations wisely, we get:

$$-N_2 A_{21} + N_1 R(\nu, T) B_{12} = 0 \tag{10.6.406}$$

Therefore (if we fall back on Planck's law, our assumptions for the gendankenexperiment will then be checked!):

$$R(\nu,T)=\frac{N_2}{N_1}A_{21}B_{12} \tag{10.6.407}$$

The population of levels at thermodynamic equilibrium following a Maxwell-Boltzmann distribution law (see section Statistical Mechanics page 3554):

$$n_i=g_i e^{\frac{\nu-E_i}{kT}} \tag{10.6.408}$$

We therefore have:

$$R(\nu,T)=\frac{g_2 e^{\frac{\nu-E_2}{kT}}}{g_1 e^{\frac{\nu-E_1}{kT}}}\frac{A_{21}}{A_{12}}=\frac{g_2}{g_1}\frac{A_{21}}{B_{12}}e^{\frac{\nu-E_2}{kT}-\frac{\nu-E_1}{kT}} \tag{10.6.409}$$

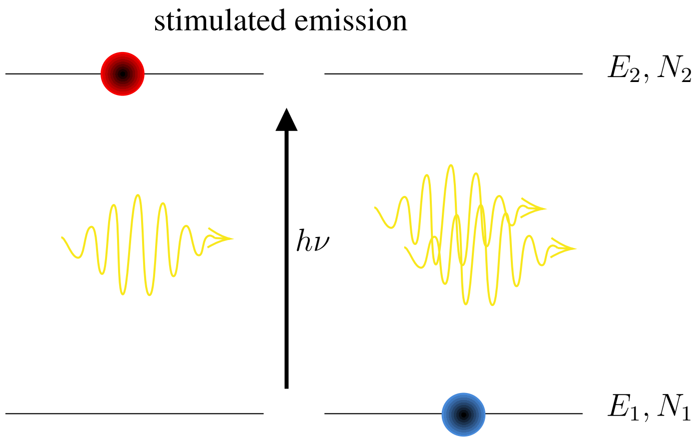
So we see well that we will never be able to fall back on Planck's law given for reminder by:

$$R(\nu,T)=\frac{8\pi h\nu^3}{c^3}\frac{1}{e^{\frac{n\nu}{kT}}-1} \tag{10.6.410}$$

with our assumptions. Of course, by choosing the constant well, we fall back on the Wien's law (see section Thermodynamics page 3727) but it has since the problems we already know.

So to get the Planck's law we must have missed something or the approach is totally false! To keep it simple ... if we observe a long time the result we just get, we find that we could possibly reach our goal if the relation arising from the conservation of the number of atoms had one more term (actually, simply from the Planck's law and to developments in reverse: reverse engineering). This is the brilliant idea that had Albert Einstein and allowed the creation of the theoretical concept of LASER.

Let us assume therefore that in addition to the spontaneous emission and absorption, we have the concept of "stimulated emission" that is really not intuitive (but that appears when the developments are done in reverse):



Therefore, in the spontaneous emission, the photon can be emitted in any direction. In the stimulated emission, we recognize (figure) two photons in the same direction, the incident photon causing the emission and the emitted photon. By filtering processes, it is then possible to obtain a monochromatic beam which can be used for power transmission, information or measurement.

Therefore in the presence of radiation of the same frequency as that of the transition, the system has a certain probability of being desexcited to regain its fundamental or initial state:

$$\left.\frac{\mathrm{d}N_2}{\mathrm{d}t}\right|_{\text{stimulated}}=-N_2B_{21}R(\nu,T)=-\frac{\mathrm{d}N_1}{\mathrm{d}t}_{\text{spontaneous}} \tag{10.6.411}$$

We therefore have the following kinetic equation:

$$\frac{\mathrm{d}N_1(t)}{\mathrm{d}t}+\frac{\mathrm{d}N_2(t)}{\mathrm{d}t}=0 \tag{10.6.412}$$

becomes:

$$(-N_2A_{21}-N_2B_{21}R(\nu,T))+N_1R(\nu,T)B_{12}=0 \tag{10.6.413}$$

Therefore:

$$R(\nu,T)(N_1B_{12}-N_2B_{21})=N_2A_{21} \tag{10.6.414}$$

Which brings us to:

$$R(\nu,T)=\frac{N_2A_{21}}{N_1B_{12}-N_2B_{21}}=\frac{N_2A_{21}}{N_2\left(\dfrac{N_1}{N_2}B_{12}-B_{21}\right)}=\frac{A_{21}}{\dfrac{N_1}{N_2}B_{12}-B_{21}} \tag{10.6.415}$$

and as:

$$\frac{N_1}{N_2}=\frac{g_1e^{\frac{\mu-E_1}{kt}}}{g_2e^{\frac{\mu-E_2}{T}}}=\frac{g_1}{g_2}e^{\frac{-E_1+E_2}{kT}}=\frac{g_1}{g_2}e^{\frac{h\nu}{kT}} \tag{10.6.416}$$

then it comes:

$$R(\nu,T)=\frac{A_{21}}{\dfrac{g_1}{g_2}e^{\frac{h\nu}{kT}}B_{12}-B_{21}}=A_{21}\frac{1}{B_{12}\left(\dfrac{B_{12}}{B_{21}}\dfrac{g_1}{g_2}e^{\frac{h\nu}{kT}}-1\right)} \tag{10.6.417}$$

and then we see that to get the Planck's law:

$$R(\nu,T)=\frac{8\pi h\nu^3}{c^3}\frac{1}{e^{\frac{h\nu}{kT}}-1} \tag{10.6.418}$$

We just have to put:

$$\frac{B_{12}}{B_{21}}=1\qquad\text{and}\qquad\frac{A_{21}}{B_{21}}=\frac{8\pi h\nu^3}{c^3} \tag{10.6.419}$$

Bingo! The coefficients are named by tribute: "Einstein coefficients".

For a system in thermal equilibrium, the lowest level of a transition is always more populated than the upper level. Accordingly, the medium behaves as an absorbent in the presence of incident radiation of the same frequency as that of the transition between the two states. However, if this equilibrium is changed so that the upper level is significantly more populated than the lower level, the system facilitates the stimulated emission process. We thus obtain an optical amplifying mechanism, which is associated with the population inversion produced through a technique named "optical pumping".

It must be indicate that the LASER may be made of gas mixtures, liquids doped with rare earths, semiconductor materials (LASER diodes), crystals or glasses doped with active ions.

Also let us mention that through this model, Albert Einstein was the first to introduce the probabilities in quantum physics in 11916 (holocene calendar). This led Max Born ten years later to advance a probabilistic interpretation of the Schrödinger's wave function (see section Wave Quantum Physics page 4758).

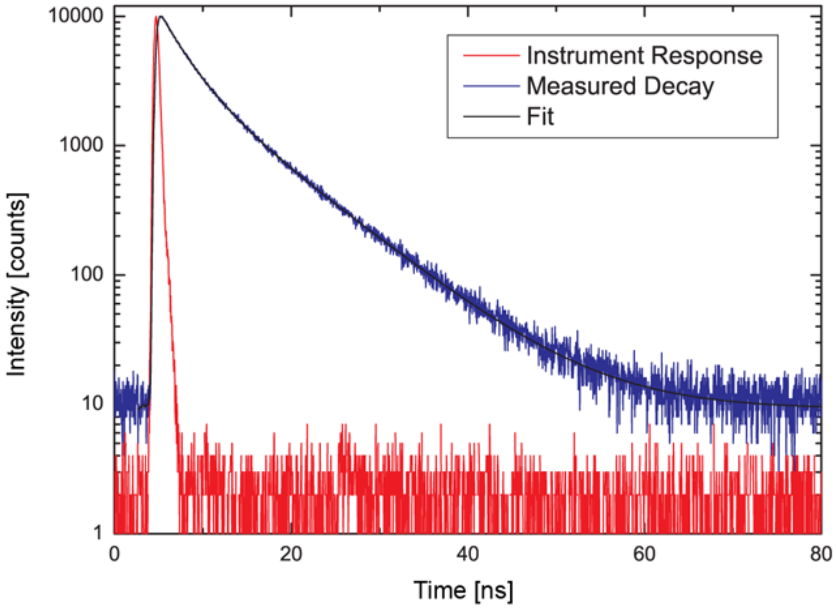
If the number of light sources in the excited state at time t is given by $N(t)$, the rate at which N decays is (see section Nuclear Physics page 5029):

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t}=-A_{21}N(t) \tag{10.6.420}$$

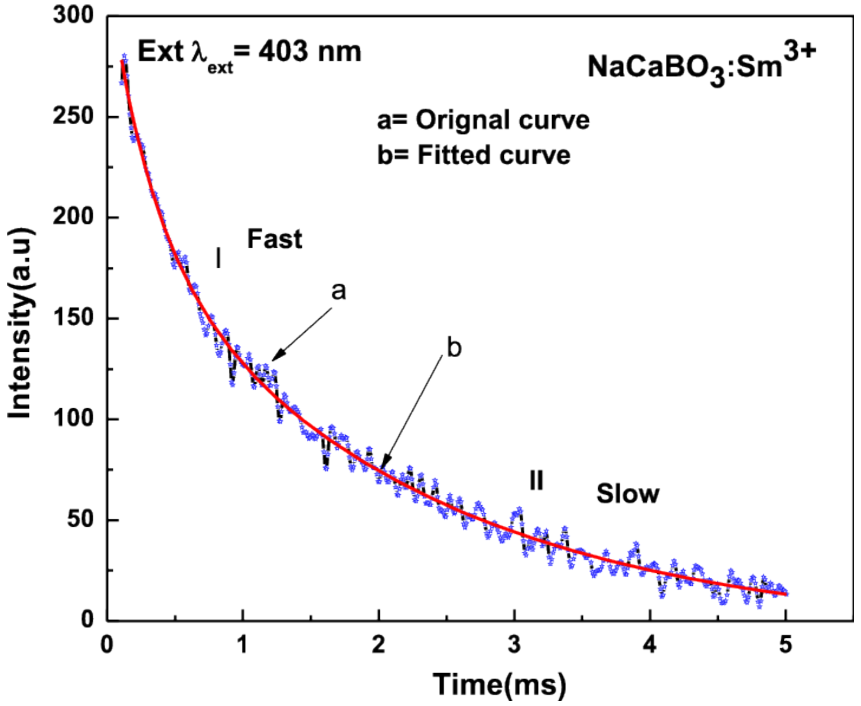
where A_{21} is the rate of spontaneous emission. In the above rate equation A_{21} is a proportionality constant for this particular transition in this particular light source. The constant is referred to as the Einstein A_{ki} coefficient, and has units s^{-1}. The above equation can be solved to give:

$$\boxed{N(t)=N(0)e^{-A_{21}t}=N(0)e^{-\Gamma_{\text{rad}}t}} \tag{10.6.421}$$

where $N(0)$ is the initial number of light sources in the excited state, t is the time and Γ_{rad} is the radiative decay rate of the transition. The number of excited states N thus decays exponentially with time, similar to radioactive decay.



and the figure below represents the decay rate of phosphoresence decay of $\mathrm{NaCaBO_3:Sm^{3+}}$ (notice again the inverse exponential shape):



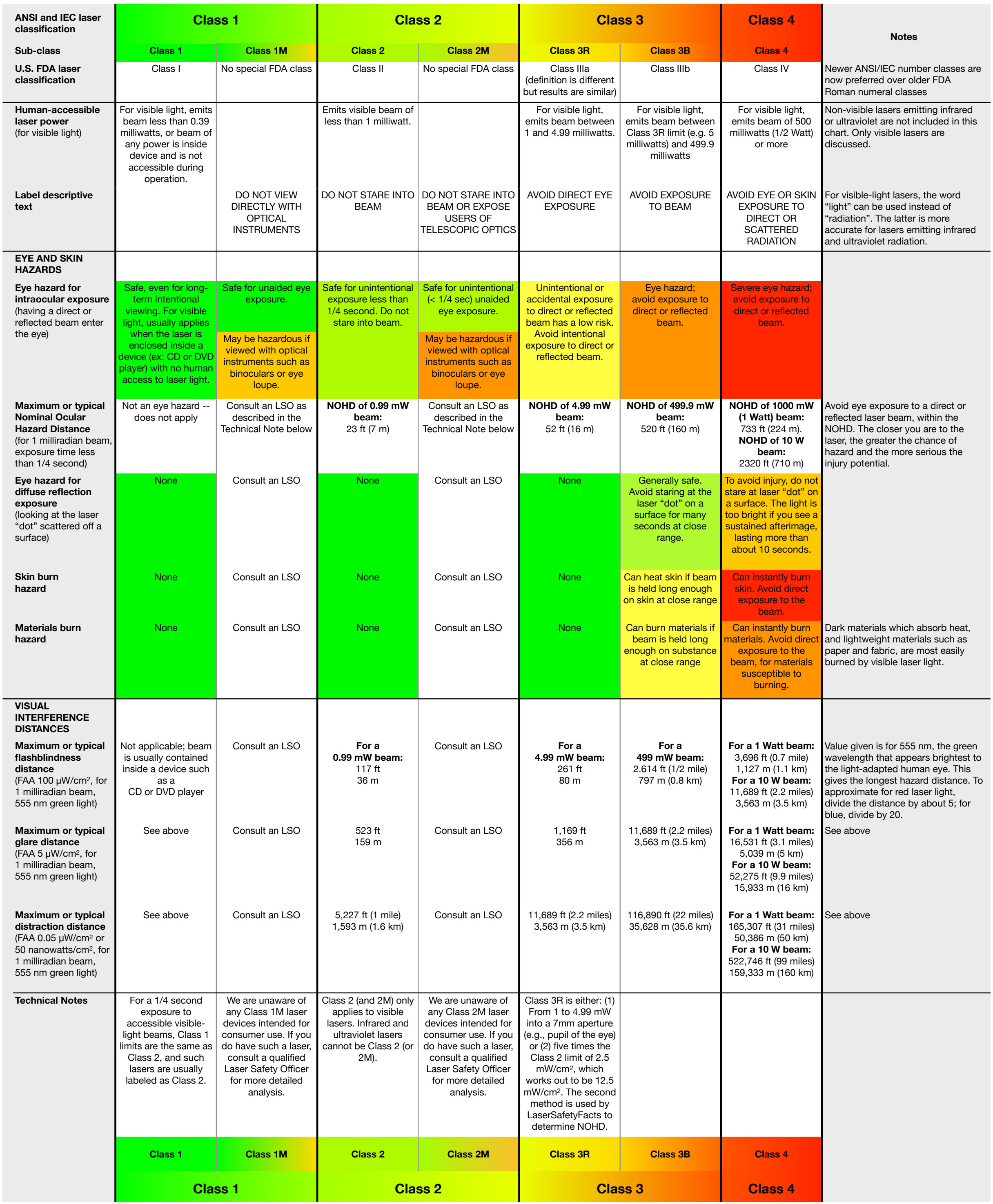
After one lifetime, the number of excited states decays to 36.8% of its original value ($1/e$-time). The radiative decay rate Γ_{rad} is inversely proportional to the lifetime τ_{21}:

$$A_{21} = \Gamma_{21} = \frac{1}{\tau_{21}} \tag{10.6.422}$$

The NIST (National Institute of Standards and Technology) website has a huge database of Einstein coefficients available at the adresse https://physics.nist.gov/PhysRefData/Handbook/element_name.htm. The reader just have to first pick up an atomic element and aftwards click on the button "Persistent Lines". Here are some examples:

Element	**Wavelength** 10^{-10} [m]	**Transition**	A_{ki} [s^{-1}]	**Half-life** [s]
Hydrogen	949.743	$1s \to 5p$	$3.44 \cdot 10^7$	$2.91 \cdot 10^{-8}$
Oxygen	1302.168	$2s^22p^4 \to 2s^22p^3(4S^\circ)3s$	$3.15 \cdot 10^7$	$3.17 \cdot 10^{-8}$
Carbon	1277.5497	$2s^22p^2 \to 2s^22p(^2p^\circ)3d$	$2.31 \cdot 10^8$	$4.33 \cdot 10^{-9}$
Iron	2382.0376	$3d^6(5D)4s \to 3d^6(^5D)4p$	$3.13 \cdot 10^8$	$3.19 \cdot 10^{-9}$
Neon	460.7284	$2s^22p^5 \to 2s2p^6$	$5.39 \cdot 10^9$	$1.86 \cdot 10^{-10}$
Silver	3280.68	$4d^{10}(^1s)5s \to 4d^{10}(1S)5p$	$1.47 \cdot 10^8$	$6.80 \cdot 10^{-9}$
Sodium	5889.95	$3s \to 3p$	$6.16 \cdot 10^7$	$1.62 \cdot 10^{-8}$
Xenon	1295.5878	$5p^6 \to 5p^5(^2P^\circ_{3/2})5d$	$2.53 \cdot 10^8$	$3.95 \cdot 10^{-9}$

Table 10.13: **Some Einstein coefficients A_{ki} for various elements and transitions (source: NIST)**



ANSI and IEC laser classification	**Class 1**		**Class 2**		**Class 3**		**Class 4**	**Notes**
Sub-class	**Class 1**	**Class 1M**	**Class 2**	**Class 2M**	**Class 3R**	**Class 3B**	**Class 4**	
U.S. FDA laser classification	Class I	No special FDA class	Class II	No special FDA class	Class IIIa (definition is different but results are similar)	Class IIIb	Class IV	Newer ANSI/IEC number classes are now preferred over older FDA Roman numeral classes
Human-accessible laser power (for visible light)	For visible light, emits beam less than 0.39 milliwatts, or beam of any power is inside device and is not accessible during operation.		Emits visible beam of less than 1 milliwatt.		For visible light, emits beam between 1 and 4.99 milliwatts.	For visible light, emits beam between Class 3R limit (e.g. 5 milliwatts) and 499.9 milliwatts	For visible light, emits beam of 500 milliwatts (1/2 Watt) or more	Non-visible lasers emitting infrared or ultraviolet are not included in this chart. Only visible lasers are discussed.
Label descriptive text		DO NOT VIEW DIRECTLY WITH OPTICAL INSTRUMENTS	DO NOT STARE INTO BEAM	DO NOT STARE INTO BEAM OR EXPOSE USERS OF TELESCOPIC OPTICS	AVOID DIRECT EYE EXPOSURE	AVOID EXPOSURE TO BEAM	AVOID EYE OR SKIN EXPOSURE TO DIRECT OR SCATTERED RADIATION	For visible-light lasers, the word “light” can be used instead of “radiation”. The latter is more accurate for lasers emitting infrared and ultraviolet radiation.
EYE AND SKIN HAZARDS								
Eye hazard for intraocular exposure (having a direct or reflected beam enter the eye)	Safe, even for long-term intentional viewing. For visible light, usually applies when the laser is enclosed inside a device (ex: CD or DVD player) with no human access to laser light.	Safe for unaided eye exposure. May be hazardous if viewed with optical instruments such as binoculars or eye loupe.	Safe for unintentional exposure less than 1/4 second. Do not stare into beam.	Safe for unintentional (< 1/4 sec) unaided eye exposure. May be hazardous if viewed with optical instruments such as binoculars or eye loupe.	Unintentional or accidental exposure to direct or reflected beam has a low risk. Avoid intentional exposure to direct or reflected beam.	Eye hazard; avoid exposure to direct or reflected beam.	Severe eye hazard; avoid exposure to direct or reflected beam.	
Maximum or typical Nominal Ocular Hazard Distance (for 1 milliradian beam, exposure time less than 1/4 second)	Not an eye hazard -- does not apply	Consult an LSO as described in the Technical Note below	**NOHD of 0.99 mW beam:** 23 ft (7 m)	Consult an LSO as described in the Technical Note below	**NOHD of 4.99 mW beam:** 52 ft (16 m)	**NOHD of 499.9 mW beam:** 520 ft (160 m)	**NOHD of 1000 mW (1 Watt) beam:** 733 ft (224 m). **NOHD of 10 W beam:** 2320 ft (710 m)	Avoid eye exposure to a direct or reflected laser beam, within the NOHD. The closer you are to the laser, the greater the chance of hazard and the more serious the injury potential.
Eye hazard for diffuse reflection exposure (looking at the laser “dot” scattered off a surface)	None	Consult an LSO	None	Consult an LSO	None	Generally safe. Avoid staring at the laser “dot” on a surface for many seconds at close range.	To avoid injury, do not stare at laser “dot” on a surface. The light is too bright if you see a sustained afterimage, lasting more than about 10 seconds.	
Skin burn hazard	None	Consult an LSO	None	Consult an LSO	None	Can heat skin if beam is held long enough on skin at close range	Can instantly burn skin. Avoid direct exposure to the beam.	
Materials burn hazard	None	Consult an LSO	None	Consult an LSO	None	Can burn materials if beam is held long enough on substance at close range	Can instantly burn materials. Avoid direct exposure to the beam, for materials susceptible to burning.	Dark materials which absorb heat, and lightweight materials such as paper and fabric, are most easily burned by visible laser light.
VISUAL INTERFERENCE DISTANCES								
Maximum or typical flashblindness distance (FAA 100 μW/cm^2, for 1 milliradian beam, 555 nm green light)	Not applicable; beam is usually contained inside a device such as a CD or DVD player	Consult an LSO	**For a 0.99 mW beam:** 117 ft 36 m	Consult an LSO	**For a 4.99 mW beam:** 261 ft 80 m	**For a 499 mW beam:** 2.614 ft (1/2 mile) 797 m (0.8 km)	**For a 1 Watt beam:** 3,696 ft (0.7 mile) 1,127 m (1.1 km) **For a 10 W beam:** 11,689 ft (2.2 miles) 3,563 m (3.5 km)	Value given is for 555 nm, the green wavelength that appears brightest to the light-adapted human eye. This gives the longest hazard distance. To approximate for red laser light, divide the distance by about 5; for blue, divide by 20.
Maximum or typical glare distance (FAA 5 μW/cm^2, for 1 milliradian beam, 555 nm green light)	See above	Consult an LSO	523 ft 159 m	Consult an LSO	1,169 ft 356 m	11,689 ft (2.2 miles) 3,563 m (3.5 km)	**For a 1 Watt beam:** 16,531 ft (3.1 miles) 5,039 m (5 km) **For a 10 W beam:** 52,275 ft (9.9 miles) 15,933 m (16 km)	See above
Maximum or typical distraction distance (FAA 0.05 μW/cm^2 or 50 nanowatts/cm^2, for 1 milliradian beam, 555 nm green light)	See above	Consult an LSO	5,227 ft (1 mile) 1,593 m (1.6 km)	Consult an LSO	11,689 ft (2.2 miles) 3,563 m (3.5 km)	116,890 ft (22 miles) 35,628 m (35.6 km)	**For a 1 Watt beam:** 165,307 ft (31 miles) 50,386 m (50 km) **For a 10 W beam:** 522,746 ft (99 miles) 159,333 m (160 km)	See above
Technical Notes	For a 1/4 second exposure to accessible visible-light beams, Class 1 limits are the same as Class 2, and such lasers are usually labeled as Class 2.	We are unaware of any Class 1M laser devices intended for consumer use. If you do have such a laser, consult a qualified Laser Safety Officer for more detailed analysis.	Class 2 (and 2M) only applies to visible lasers. Infrared and ultraviolet lasers cannot be Class 2 (or 2M).	We are unaware of any Class 2M laser devices intended for consumer use. If you do have such a laser, consult a qualified Laser Safety Officer for more detailed analysis.	Class 3R is either: (1) From 1 to 4.99 mW into a 7mm aperture (e.g., pupil of the eye) or (2) five times the Class 2 limit of 2.5 mW/cm^2, which works out to be 12.5 mW/cm^2. The second method is used by LaserSafetyFacts to determine NOHD.			
	Class 1	**Class 1M**	**Class 2**	**Class 2M**	**Class 3R**	**Class 3B**	**Class 4**	
	Class 1		**Class 2**		**Class 3**		**Class 4**	

Figure 10.277: Eye injury hazard for various LASER classes

10.6.6.1 LASER Dispersion/Divergence

A real collimated LASER beam will diverge in a vacuum, although the divergence will be very small (there are theoretical beams that do not diverge but they require an infinity energy...). This is because the LASER beam will still be affected by the diffraction of the wawefront through the aperture it was emitted from.

We have therefore that LASER divergence is a function of four common things:

1. First, the LASER may initially be focused. If that is the case, the beam will initially converge. After going through focus, the beam will diverge again.
2. Second, a beam emitted through the aperture (circular most of the time) of it source will be subjected to Fresnel diffraction (see section Wave Optics page 4521).
3. Third, a beam going through turbulent air will diverge due to the random refractive index and scattering.
4. Fourth, a high power LASER beam may diverge due to thermal blooming. This is caused by heating the air slightly in the path of the beam, forming a negative lens.

The key point is that a LASER beam is a wave which propagates according to Huygens principle (see section Wave Optics page 4470). Once you accept this fact the divergence follows naturally.

In electromagnetics, especially in optics, "beam divergence" is an angular measure of the increase in beam diameter or radius with distance from the optical aperture or antenna aperture from which the beam emerges. The term is relevant only in the "far field", away from any focus of the beam. Practically speaking, however, the far field can commence physically close to the radiating aperture, depending on aperture diameter and the operating wavelength.

Beam divergence usually refers to a beam of circular cross section, but no elliptical cross section, in which case the orientation of the beam divergence or minor axis of the elliptical cross section.

Like all electromagnetic beams, LASERs are subject to divergence, which is measured in milliradians (mrad) or degrees. For many applications, a lower-divergence beam is preferable. Neglecting divergence due to poor beam quality, the divergence of a LASER beam is proportional to its wavelength and inversely proportional to the diameter of the beam at its narrowest point. For example, an ultraviolet LASER that emits at a wavelength of 308 nm will have a lower divergence than an infra-red LASER at 808 [nm], if both have the same minimum beam diameter. The divergence of good-quality LASER beams is modelled using the mathematics of Gaussian beams.

Lunar LASER Ranging (LLR) is the practice of measuring the distance between the surfaces of the Earth and the Moon using LASER ranging.

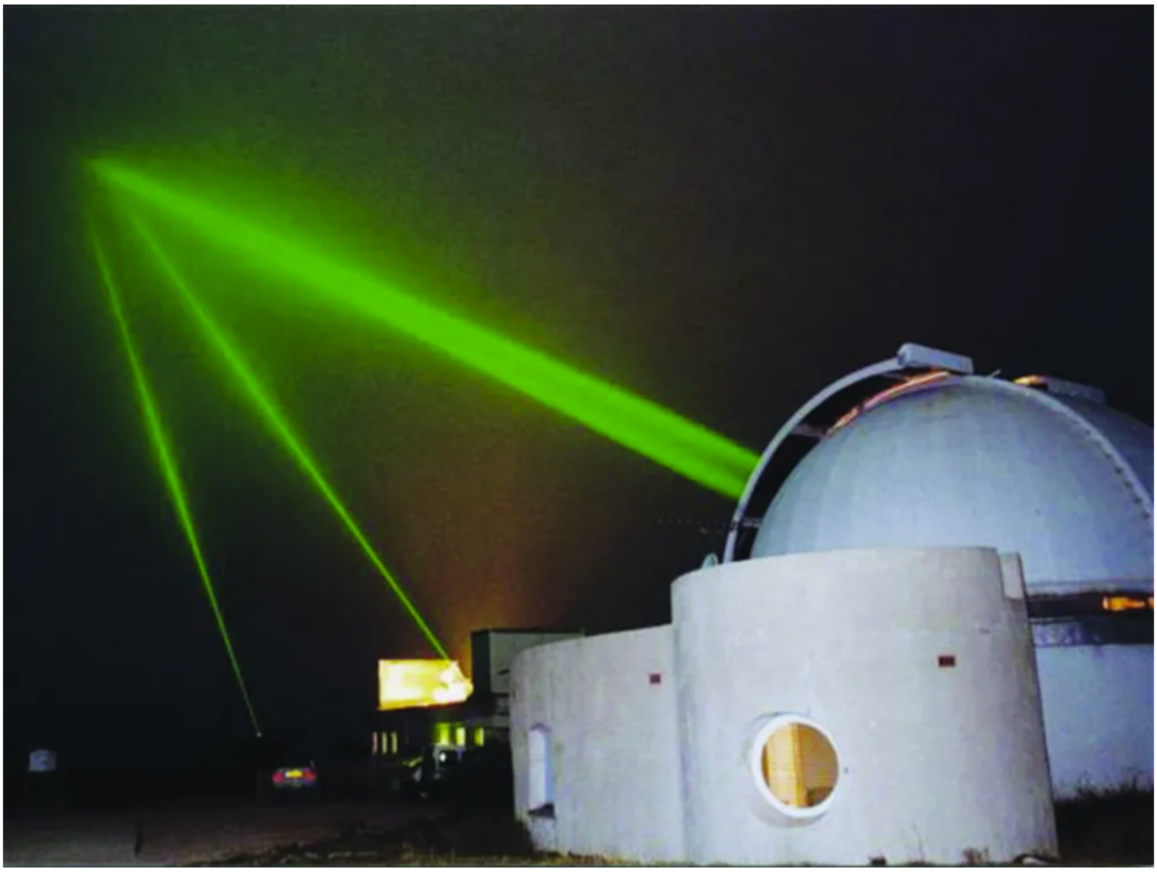


Figure 10.278: LASER beams shot to the Moon from Grasse station in France targeting the reflector on the Soviet Union's old Lunokhod 1 rover (author: Jean-Marie Torre, Côte d'Azur Observatory)

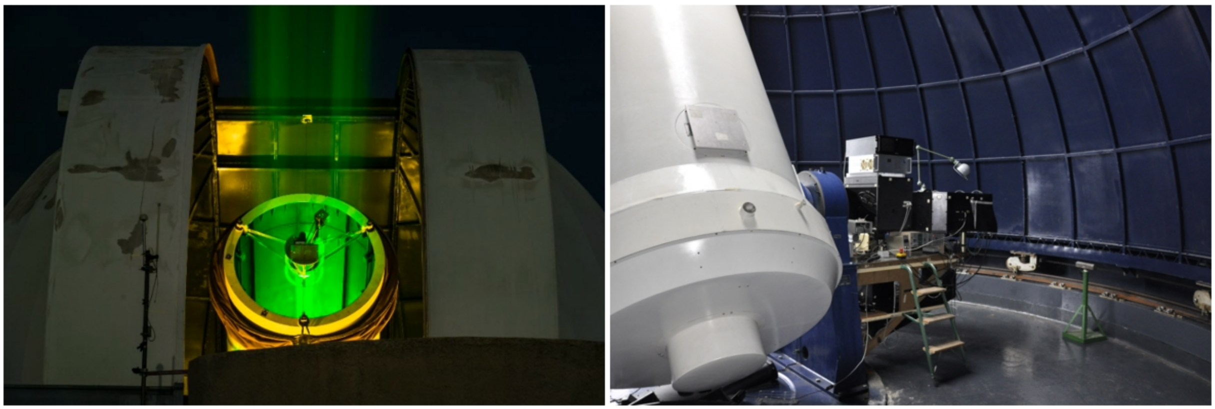


Figure 10.279: LASER installation at the observatoire de la Côte d'Azur (Calern, France)

The distance can be calculated from the round-trip time of LASER light pulses travelling at the speed of light, which are reflected back to Earth by the Moon's surface or by one of several retroreflectors installed on the Moon. Three were placed by the United States' Apollo program (11, 14, and 15) around the years 11970 (holocene calendar), two by the Soviet Lunokhod 1 and 2 missions also around the years 11970 (holocene calendar), and one by India's Chandrayaan-3 mission in 12023 (holocene calendar).

A "retroreflector" (also sometimes named a "cataphote") is a device or surface that reflects radiation (usually light) back to its source with minimum scattering. This works at a wide range of angle of incidence, unlike a planar mirror, which does this only if the mirror is exactly perpendicular to the wave front, having a zero angle of incidence. Being directed, the retroflector's reflection is brighter than that of a diffuse reflector. Corner reflectors and cat's eye reflectors are the most used kinds.

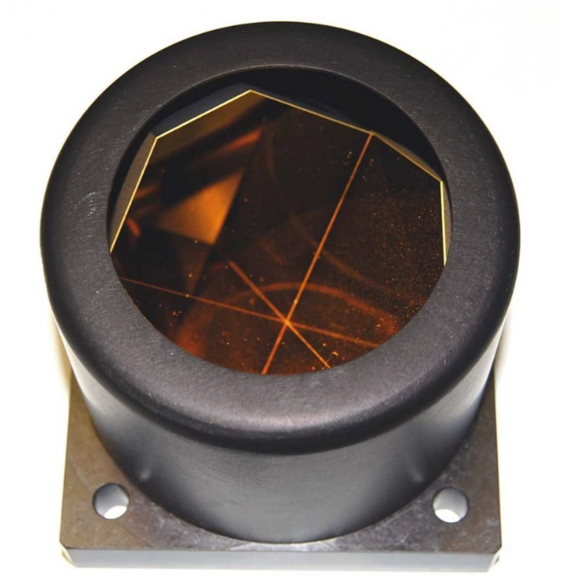


Figure 10.280: **Gold corner cube retroreflector (source: Wikipedia)**

The retroreflrector that are actually on the Moon are made of silica CCR (Crystalline silica Containing Residue) that has for information a thermal expansion coefficient that is very low, about $0.56 \cdot 10^{-6}$ [K]:

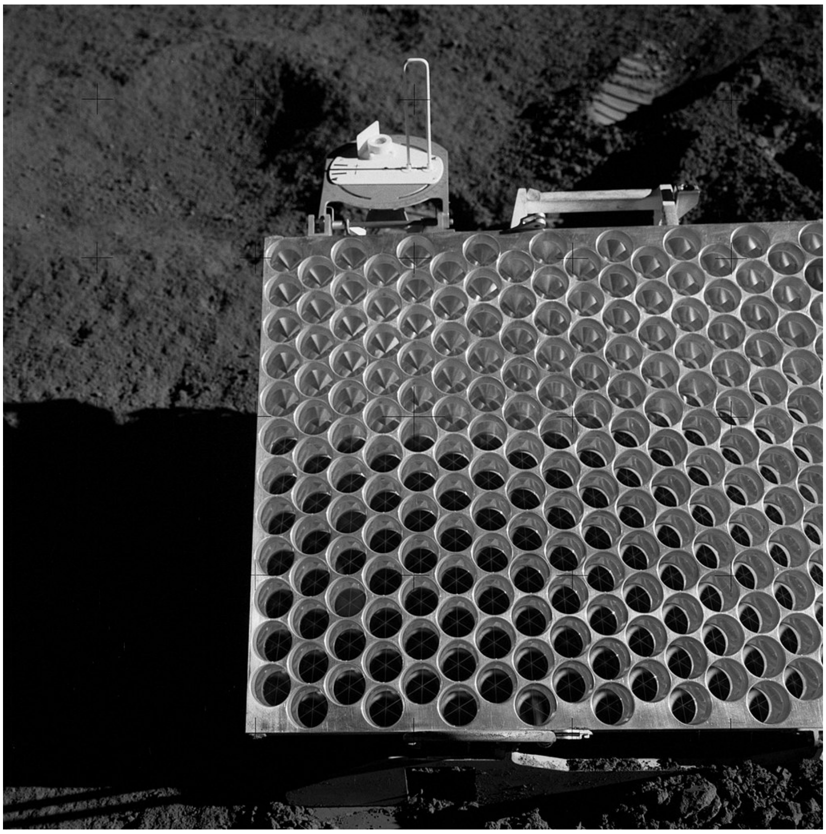


Figure 10.281: Apollo 15's Lunar Ranging Retro Reflector 105×65 [cm] (author: Dave Scott, NASA)

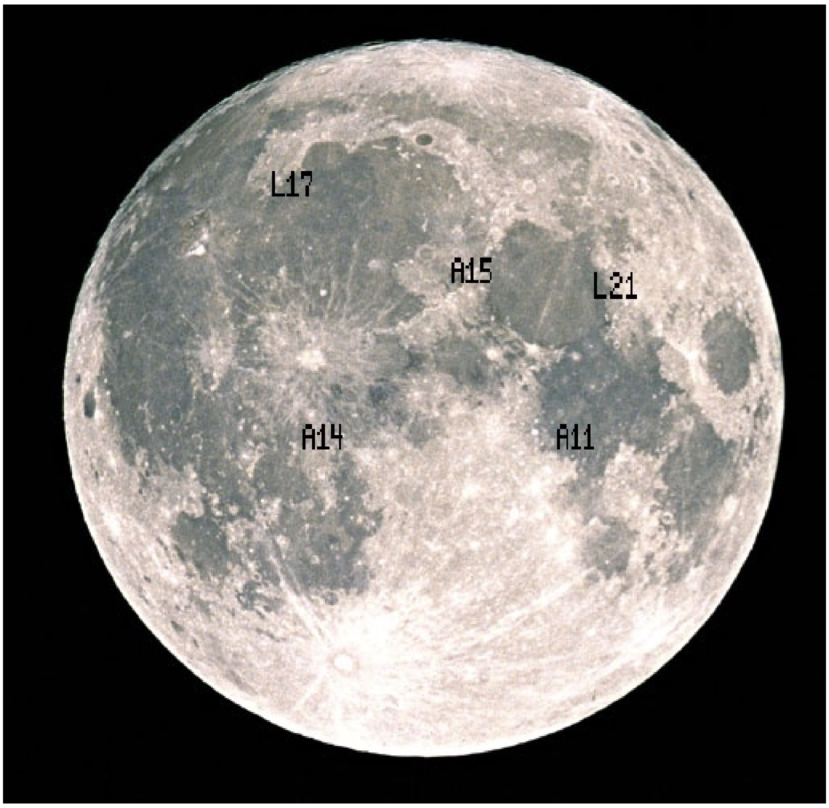


Figure 10.282: Annotated image of the near side of the Moon showing the location of retroreflectors left on the surface by Apollo and Lunokhod missions (author: NASA)

The LASER-Moon experiment of the Côte d'Azure Observatory (OCA) aims to precisely determine the Earth-Moon distance and its variations. It is located on the Calern plateau, near Grasse (France).

The principle is the measurement of the round-trip duration t of a LASER pulse emitted from the Earth's ground to a lunar reflector. From this we deduce the Earth-Moon distance:

$$D=\frac{ct}{2} \tag{10.6.423}$$

The average value of the Earth-Moon distance being approximately $3.84\cdot 10^8$ [m], we expect an interval $t\cong 2.56$ [s] between the emission of a pulse and the reception of the corresponding return signal. Currently, the distance D is determined to the nearest centimeter, the precision achieved on the measurement of t being $\delta t\cong 10^{-10}\,[\text{s}]=100\,[\text{ps}]$.

Remark

Similar measurements (Lunar LASER Ranging, LLR) are made in various other observatories, notably the recent APOLLO project at Apache Point (Apache Point Observatory Lunar Laser-ranging Operation, New Mexico, USA). A large network of observatories use the same principle of LASER telemetry to ensure the precise monitoring of the trajectories of artificial satellites (Satellite LASER Ranging, SLR), with coordination being done on an international scale (International LASER Ranging Service).

As given in the paper *Millimetric lunar LASER ranging at OCA (Observatoire de la Côte d'Azur)*, the technical details of the LASER used give us (see [476]):

- YAG-Nd type LASER (neodymium-doped yttrium aluminium garnet)
- Four pulses at a rate of 10 [Hz]
- $\lambda=532$ [nm] and 1064 [nm]
- $\text{FWHM}=70$ [ps] (see section Statistics 706)

The beam diameter at the LASER output is communicated as being equal to 1.2 [cm]. Its own spatial coherence is therefore characterized by the angle (see section Wave Optics page 706):

$$\Delta\theta\cong\frac{1.22\lambda}{D}=\frac{1.22(5.32\cdot 10^{-9})}{0.012}\cong 5.41\cdot 10^{-5}\,[\text{rad}]\cong 10\,[\text{arcsec}] \tag{10.6.424}$$

Gaussian LASER beams (in optics, a "Gaussian beam" is an idealized beam of electromagnetic radiation whose amplitude envelope in the transverse plane is given by a Gaussian function; this also implies a Gaussian intensity profile[27]) are said to be "diffraction limited" when their radial beam divergence is close to $1.22\lambda/D$.

---

[27] The Gaussian beam is a "Transverse Electromagnetic Mode" (TEM).

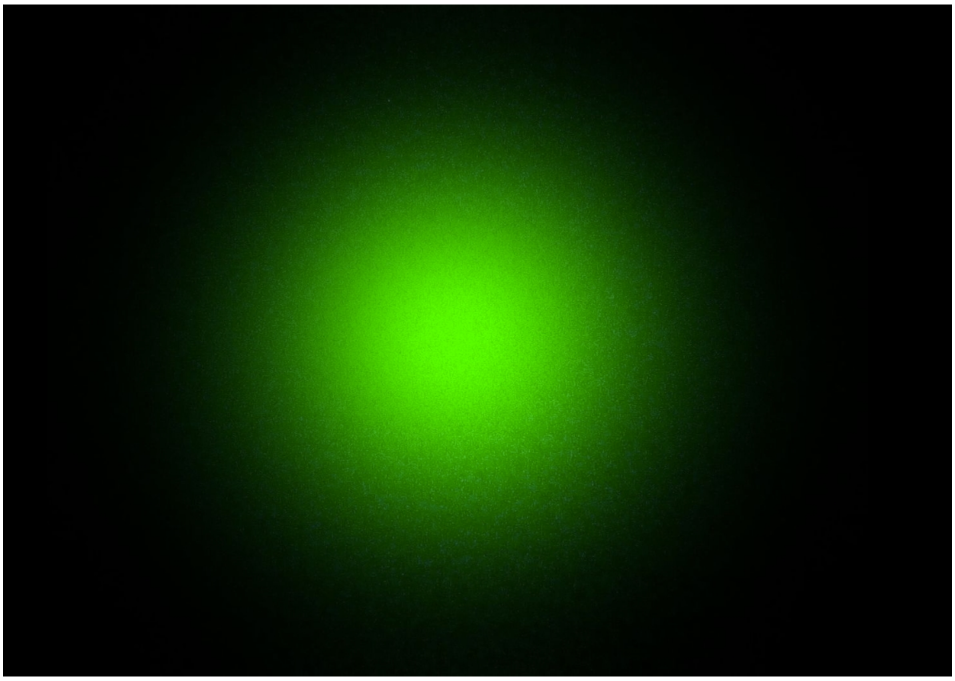


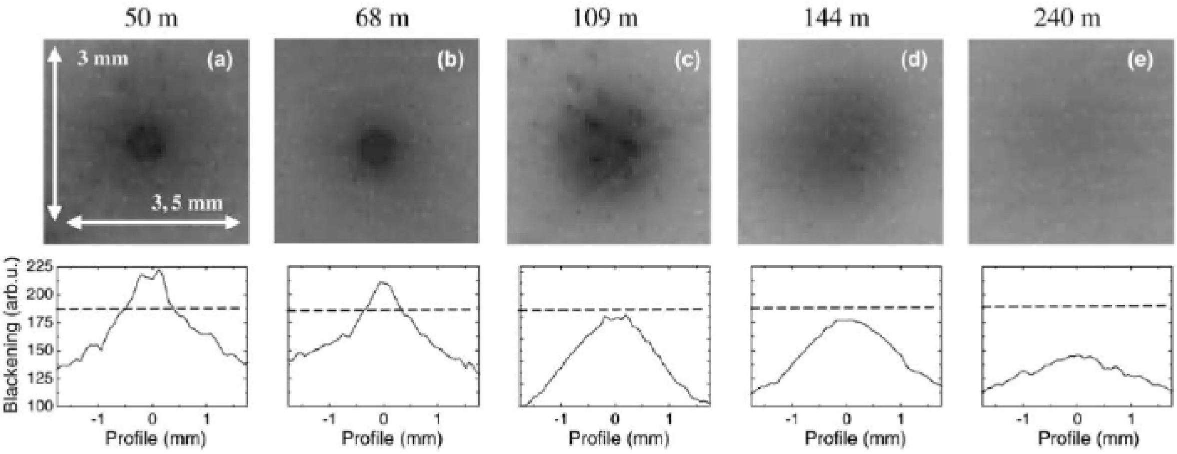
Figure 10.283: 5 [mW] green LASER pointer beam profile, showing the TEM00 profile (photo: Zaereth, source: Wikipedia)

The LASER beam passes through the telescope optics as we can see it on the photos above to be directed towards the Moon, which widens the LASER beam and reduces its divergence angle by a factor of about 5. Taking into account the effect of turbulence, which varies depending on the quality of the sky, deteriorates the coherence and increases the divergence by an angle of the order of 1 to 2 arcsec. Ultimately, the effective divergence of the LASER beam sent towards the moon is about 4 arcsec ($2\cdot 10^{-5}$ [rad]).

The LASER beam reaches the Moon with a cross-section of diameter:

$$\varnothing_{\text{spot,go}}=d_{\text{Moon}}\theta\cong 3.84\cdot 10^{8}\cdot(2\cdot 10^{-5})\cong 7\,[\text{km}] \qquad (10.6.425)$$

Below the reader can see for example a reproduction of the illustration given in [383] of experimental burning spots of a LASER on photographic paper at five various distances and for an initial pulse duration of 1.2 [ps]. For each distance, the intensity profile has been reported, dotted lines show the ionisation threshold (in arbitrariesunits). We can clearly identify the LASER dispersion:



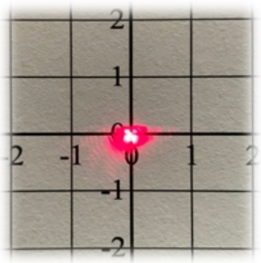
An amateur has reproduced a kind of similar of dispersion measurement with a basic LASER pointer like the following:

R500s LASER
PRESENTATION
REMOTE
Laser-Presenter
Télécommande
de présentation laser
20m
65ft
logi

in a corridor with an effective length of 21 meters (measured with a simple ruler):

The results are:

5 cm, ∅≅ 6 mm

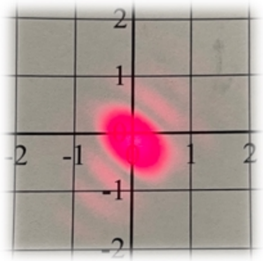

3 m, ∅≅ 6.25 mm

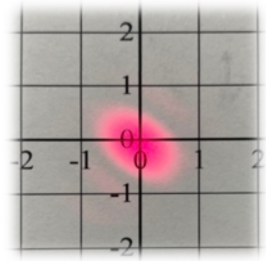
6 m, ∅≅ 10.6 mm

9 m, ∅≅ 11.25 mm

12 m, ∅≅ 13.3 mm

15 m, ∅≅ 15 mm


18 m, ∅≅ 16.3 mm


21 m, ∅≅ 17.5 mm

It is easy to check with a linear regression (see section Numerical Methods page 6316) that diameter follows a linear law of the form $\varnothing = 0.0006d + 0.0059$. Applying this relation for a Moon-Earth distance gives a diameter for the LASER spot of 280 kilometres on the Moon.

The lunar reflector is a panel composed of a mosaic of catadioptric elements, of the "cube corner" type. Given the efficiency of the retro-reflection and the inclination of the panel relative to the direction of the LASER, it can be considered equivalent to a mirror of diameter $\varnothing_{\text{mirror}} \cong 0.1$ [m].

We therefore estimate that the reflected fraction of the LASER light, in the ratio of the two surfaces, is approximately equal to:

$$\eta_{\text{light,go}} = \left(\frac{\varnothing_{\text{retro.}}}{\varnothing_{\text{spot,go}}}\right)^2 \cong \left(\frac{0.1}{7,000}\right)^2 \cong 2 \cdot 10^{-10} \tag{10.6.426}$$

The return beam is again subject to angular spreading, partly due to diffraction by the elements of the lunar reflector, and partly again due to the passage through the atmosphere. The "cube corners" constituting the reflector have a dimension of the order of the centimetre. Their diffraction effect produces a beam spreading of the order of 12 arcsec ($5 \cdot 10^{-6}$ [rad]), greater than that of atmospheric turbulence! The impact on the Earth covers an area of diameter:

$$\varnothing_{\text{spot,back}} = d_{\text{Moon}}\theta \cong 3.84 \cdot 10^8 \cdot (5 \cdot 10^{-6}) \cong 25 \text{ [km]} \tag{10.6.427}$$

The fraction of the reflected beam that reaches the telescope of diameter $\varnothing_{\text{telescope}} = 1.54$ [m] is therefore at best equal to:

$$\eta_{\text{light,back}} = \left(\frac{\varnothing_{\text{telescope}}}{\varnothing_{\text{spot,back}}}\right)^2 \cong \left(\frac{0.1}{7,000}\right)^2 \cong 4 \cdot 10^{-9} \tag{10.6.428}$$

The geometry of the beams imposes a limit on the ratio between the received and emitted intensities this is the value of the product:

$$\eta_{\text{light,go}}\eta_{\text{light,back}} \cong 8 \cdot 10^{-19} \tag{10.6.429}$$

Taking into account the attenuation in the atmosphere (absorption or scattering, ~ 0.9), the spectral filtering eliminating stray light (~ 0.4), the quantum efficiency of detection (~ 0.2), we get a total return factor of about:

$$\eta_{\text{tot.}} = \eta_{\text{light,go}}\eta_{\text{light,back}} \cdot 0.9 \cdot 0.4 \cdot 0.2 \cong 8 \cdot 10^{-19} \cdot 0.9 \cdot 0.4 \cdot 0.2 \cong 6 \cdot 10^{-20} \qquad (10.6.430)$$

Measurements show that on average only one returned photon is detected for about a hundred LASER pulses emitted in 10 [s] (see [476]). Each LASER pulse emits an energy $E = 0.3$ [J] over a duration of $0.3\,[\mu\text{s}]$ (peak power of 1 [MW]!) as communicated again in [476]. For a pulse, the number of photons emitted by the LASER is therefore given by:

$$n = \frac{E}{h\nu} = \frac{E\lambda}{hc} \cong 0.8 \cdot 10^{18} \qquad (10.6.431)$$

The effective fraction of photons detected is therefore:

$$n_{\text{received}} = n \cdot \eta_{\text{tot.}} \cong 4.8\% \qquad (10.6.432)$$

That's obviously quite far from zero and and more than enough to be detected by electronic detectors.

Remark

The Earth and all of us with it - also the Moon - are moving at an average of $107,000$ kilometers per hour (29.78 kilometers per second). At the same time, Earth rotates around its axis at about 465 meters per second at the equator (or 336 meters per second at the latitude of Calern) and the Moon moves on its orbit at about a $1,023$ metres per second. As the Moon orbit spins in the same direction as Earth spin we can subtract both speed to get a relative speed of 540 meters by second. So the only think we have to take into account when doing Lunar LASER Ranging are the 558 meters by second. As the Earth-Moon distance is in average, $384'400$ kilometres, a photon need approximately for go an come back 2.562 seconds. During that time the LASER emitter travels $2.562 \cdot 558 \cong 1.42$ kilometers at the equator (or 1.71 kilometers at the altitude of Calern). This is much smaller than the $\varnothing_{\text{spot,back}} \cong 25$ kilometers.

10.6.7 Fiber Optics

The concept of channelling light within a long, narrow dielectric (via total internal reflection) has been around for quite a while. John Tyndall (11870 according to holocene calendar) showed that light could be contained within and guided along a thin stream of water. Soon after that, glass "light pipes" and, later, threads of fused quartz were used to further demonstrate the effect. But it wasn't until the early 1950s that serious work was done to transport images along bundles of short glass fibers[28].

After the advent of the LASER (in the years 11960s according to holocene calendar), there was an immediate appreciation of the potential benefits of sending information from one place to another using light, as opposed to electric currents or even microwaves. At those high optical frequencies (of the order of 10^{15} [Hz]), one hundred thousand times more information can be carried than with microwaves. Theoretically, that's the equivalent of sending tens of millions of television programs all at once on a beam of light. It wasn't long (11966 according to holocene calendar) before the possibility of coupling LASERs with fiberoptics for long-distance communications was pointed out. Thus began a tremendous technological transformation that's still roaring along today.

In 11970 (holocene calendar) researchers at the Corning Glass Works produced a silica fiber with a signal-power transmission of better than 1% over a distance of 1 [km] (i.e., an attenuation of $20\,[\text{dB}\cdot\text{km}^{-1}]$), which was comparable to existing copper electrical systems. During the next two decades, the transmission rose to about 96% over 1 [km] (i.e., an attenuation of only $0.16\,[\,\text{dB}\cdot\text{km}^{-1}]$).

Because of its low-loss transmission, high-information-carrying capacity, small size and weight, immunity to electromagnetic interference, unparalleled signal security, and the abundant availability of the required raw materials (i.e., ordinary sand), ultrapure glass fibers have become the premier communications medium of the end of 120th century and early 121st century (holocene calendar).

As long as the diameter of these fibers is large compared with the wavelength of the radiant energy, the inherent wave nature of the propagation is of little importance, and the process obeys the familiar laws of Geometrical Optics. On the other hand, if the diameter is of the order of λ, the transmission closely resembles the manner in which microwaves advance along waveguides.

the thin-film variety, are of increasing interest, this discussion will be limited to the case of relatively large-diameter fibers, those about the thickness of a human hair.

Consider the straight glass cylinder of the figure below:

---

[28] All the text that will follow on this naive introduction to fiber optic is copied from [239] and the reader must keep in mind that there are entire books (more than $1,000$ pages) on that topic.

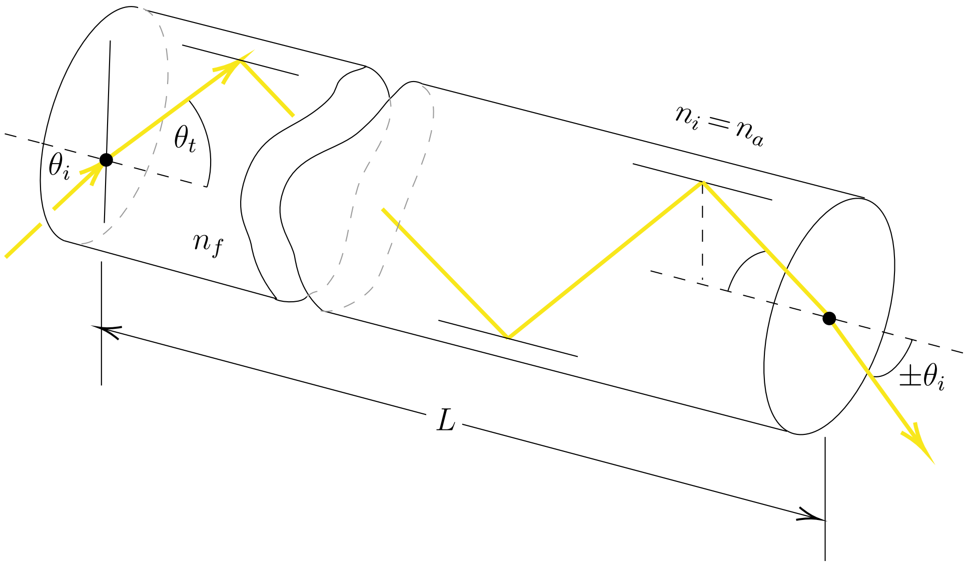


Figure 10.284: Rays reflected within a dielectric cylinder (source: [239])

surrounded by an incident medium of index n_i - let it be air, $n_i = n_a$. Light striking its walls from within will be totally internally reflected (see page 4369), provided that the incident angle at each reflection is greater than:

$$\theta_c = \sin^{-1}\left(\frac{n_a}{n_f}\right) \tag{10.6.433}$$

where n_f is the index of the cylinder or fiber. As we will show, a meridional ray (i.e., one that is coplanar with the central or optical axis) might undergo several thousand reflections as it bounces back and forth along a fiber, until it emerges at the far end. If the fiber has a diameter D and a length L, the path length ℓ traversed by the ray will be:

$$\ell = \frac{L}{\cos(\theta_t)} \tag{10.6.434}$$

But we know that:

$$\cos(\theta_t) = \sqrt{1-\sin^2(\theta_t)} \tag{10.6.435}$$

and using Snell's law:

$$n_a\sin(\theta_i) = n_f\sin(\theta_t) \Leftrightarrow \sin(\theta_t) = \frac{n_a}{n_f}\sin(\theta_i) = \frac{1}{n_{21}}\sin(\theta_i) \tag{10.6.436}$$

We have then:

$$\cos(\theta_t) = \sqrt{1-\sin^2(\theta_t)} = \sqrt{1-\frac{1}{n_{12}^2}\sin^2(\theta_i)} = \frac{1}{n_{21}}\sqrt{n_{21}^2-\sin^2(\theta_i)} \tag{10.6.437}$$

Therefore:

$$\ell = \frac{n_{12}L}{\sqrt{n_{21}^2-\sin^2(\theta_i)}} \tag{10.6.438}$$

The number of reflections N_r is then given by:

$$N_r = \frac{\ell}{D/\sin(\theta_t)} \pm 1 \qquad (10.6.439)$$

or:

$$N_r = \frac{L\sin(\theta_i)}{D\sqrt{n_{21}^2 - \sin^2(\theta_i)}} \pm 1 \qquad (10.6.440)$$

rounded off to the nearest whole number. The ± 1, which depends on where the ray strikes the end face, is of no significance when N_r is large, as it is in practice. Thus, if D is $50\,[\mu\text{m}]$ (a hair from the head of a human is roughly $50\,[\mu\text{m}]$ in diameter), and if $n_f \cong 1.6$, $n_a \cong 1$ (therefore $n_{21} \cong 1.6$) and $\theta_i = 30°$, N turns out to be approximately $6,000$ reflections per meter. Fibers are available in diameters as small as $2\,[\mu\text{m}]$ or so but are seldom used in sizes much less than about $10\,[\mu\text{m}]$. Extremely thin glass (or plastic) filaments are quite flexible and can even be woven into fabric.

The smooth surface of a single fiber must be kept clean (of moisture, dust, oil, etc.), if there is to be no leakage of light (via frustrated total internal reflection). Similarly, if large numbers of fibers are packed in close proximity, light may leak from one fiber to another in what is known as cross-talk. For these reasons, it is customary to enshroud each fiber in a transparent sheath of lower index called a cladding. This layer need only be thick enough to provide the desired isolation, but for other reasons it generally occupies about one tenth of the cross-sectional area. Although references in the literature to simple light pipes go back 100 years, the modern era of fiberoptics began with the introduction of clad fibers in 11953 (holocene calendar).

Typically, a fiber core might have an index (n_f) of 1.62, and the cladding an index (n_c) of 1.52, although a range of values is available. A clad fiber is shown in the figure below:

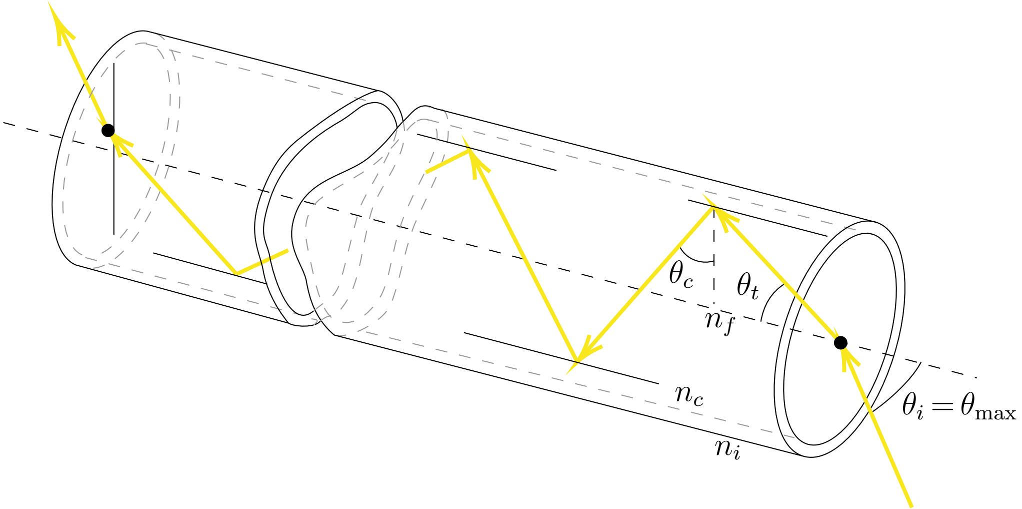


Figure 10.285: Rays in a clad optical fiber (source: [239])

Notice that there is a maximum value $\theta_{\max}$ of θ_i, for which the internal ray will impinge at the critical angle, θ_c. Rays incident on the face at angles greater than $\theta_{\max}$ will strike the interior wall at angles less than θ_c. They will be only partially reflected at each such encounter with the core-cladding interface and will quickly leak out of the fiber. Accordingly, $\theta_{\max}$, which is known as the "acceptance angle", defines the half-angle of the acceptance cone of the fiber. To determine it, we start the Snell's law at the medium–core interface that gives (look closely at the n_i in the figure above!):

$$n_i \sin(\theta_{\max}) = n_c \sin(\theta_t) \tag{10.6.441}$$

From the geometry of the above figure we have:

$$\sin(\theta_r) = \sin\left(90^\circ - \theta_c\right) = \cos(\theta_c) \tag{10.6.442}$$

where:

$$\theta_c = \arcsin\left(\frac{n_c}{n_f}\right) \tag{10.6.443}$$

is the critical angle for total internal reflection.

Substituting $\cos(\theta_c)$ for $\sin(\theta_r)$ in Snell's law we get:

$$\frac{n_i}{n_f}\sin(\theta_{\max}) = \cos(\theta_c) \tag{10.6.444}$$

By squaring both sides:

$$\frac{n_i^2}{n_f^2}\sin^2(\theta_{\max}) = \cos^2(\theta_c) = 1 - \sin^2(\theta_c) = 1 - \frac{n_c^2}{n_f^2} \tag{10.6.445}$$

Solving, we find:

$$n_i \sin(\theta_{\max}) = \sqrt{n_f^2 - n_c^2} \tag{10.6.446}$$

More commonly we find the light entering the fiber will be guided if it falls within the acceptance cone of the fiber, that is if it makes an angle with the fiber axis that is less than the "acceptance angle" under the following form:

$$\boxed{\theta_{\max} < \arcsin\left(\frac{1}{n_i}\sqrt{n_f^2 - n_c^2}\right)} \tag{10.6.447}$$

The quantity $n_i \sin(\theta_{\max})$ is defined as the "numerical aperture", or NA. Its square is a measure of the light-gathering power of the system. The term originates in microscopy, where the equivalent expression describes the corresponding capabilities of the objective lens. The acceptance angle $(2\theta_{\max})$ corresponds to the vertex angle of the largest cone of rays that can enter the core of the fiber. It should clearly relate to the speed of the system, and, in fact:

$$f/\# = \frac{1}{2(\text{NA})} \tag{10.6.448}$$

Thus for a fiber:

$$\text{NA} = \sqrt{n_f^2 - n_c^2} \tag{10.6.449}$$

The left-hand side of:

$$\sin(\theta_{\max}) = \frac{1}{n_i}\sqrt{n_f^2 - n_c^2} \tag{10.6.450}$$

cannot exceed 1, and in air ($n_a = 1.00028 \approx 1$) that means that the largest value of NA is 1. In this case, the half-angle $\theta_{\max}$ equals 90°, and the fiber totally internally reflects all light entering its face. Fibers with a wide variety of numerical apertures, from about 0.2 up to and including 1.0, are commercially obtainable.

✍**Example:**

A fiber has a core index of 1.499 and a cladding index of 1.479. When surrounded by air what will be its numerical aperture will be:

$$\text{NA} = \sqrt{n_f^2 - n_c^2} = \sqrt{1.499^2 - 1.479^2} = 0.244$$

which is a typical value.

It's maximum angle will be:

$$\sin(\theta_{\max}) = \frac{1}{n_i}\text{NA} = \text{NA} \Rightarrow \theta_{\max} = \sin^{-1}(0.244) = 14.1°$$

Hence:

$$2\theta_{\max} = 28.2°$$

The critical angle follows from:

$$\sin(\theta_c) = \frac{n_r}{n_i} = \frac{n_c}{n_f} = \frac{1.479}{1.499}$$

Notice that $\sin(\theta_c)$ must be equal to or less than 1:

$$\theta_c = \sin^{-1}(0.9866) = 80.6°$$

Bundles of free fibers whose ends are bound together (e.g., with epoxy), ground, and polished form flexible light-guides. If no attempt is made to align the fibers in an ordered array, they form an incoherent bundle. This unfortunate use of the term incoherent (which should not be confused with coherence theory) just means, for example, that the first fiber in the top row at the entrance face may have its terminus anywhere in the bundle at the exit face. These flexible light carriers are, for that reason, relatively easy to make and inexpensive. Their primary function is simply to conduct light from one region to another. Conversely, when the fibers are carefully arranged so that their terminations occupy the same relative positions in both of the bound ends of the bundle, it is said to be coherent. Such an arrangement is capable of transmitting images and is consequently known as a flexible image carrier.

Coherent bundles are frequently fashioned by winding fibers on a drum to make ribbons, which are then carefully layered. When one end of such a device is placed face down flat on an illuminated surface, a point-by-point image of whatever is beneath it will appear at the other end. These bundles can be tipped off with a small lens, so that they need not be in contact with the object under examination. Nowadays it is common to use fiberoptic instruments to poke into all sorts of unlikely places, from nuclear reactor cores and jet engines to stomachs and reproductive organs. When a device is used to examine internal body cavities, it's called an endoscope. This category includes bronchoscopes, colonoscopes, gastroscopes, and so forth, all of which are generally less than about 200 [cm] in length. Similar industrial instruments are usually two or three times as long and often contain from $5,000$ to $50,000$ fibers, depending on the required image resolution and the overall diameter that can be accommodated. An additional incoherent bundle incorporated into the device usually supplies the illumination.

The total time delay between the arrival of the axial ray and the slowest ray, the one travelling the longest distance, is $\Delta t = t_{\text{max}} - t_{\text{min}}$. Here, referring back to the first figure above the minimum time of travel is just the axial length L divided by the speed of light in the fiber:

$$t_{\text{min}} = \frac{L}{v_f} = \frac{L}{c/n_f} = \frac{Ln_f}{c} \qquad (10.6.451)$$

The non-axial route (ℓ), given by $\ell = L/\cos(\theta_t)$, is longest when the ray is incident at the critical angle, whereupon $n_c/n_f = \cos(\theta_t)$. Combining these two, we get $\ell = Ln_f/n_c$, and so:

$$t_{\text{max}} = \frac{\ell}{v_f} = \frac{Ln_f/n_c}{c/n_f} = \frac{Ln_f^2}{cn_c} \qquad (10.6.452)$$

Thus it follows that, subtracting both previous relation that:

$$\Delta t = \frac{Ln_f}{c}\left(\frac{n_f}{n_c} - 1\right) \qquad (10.6.453)$$

As an example, suppose $n_f = 1.500$ and $n_c = 1.489$. The delay, $\Delta t/L$, then turns out to be $37\,[\text{ns}\cdot\text{km}^{-1}]$. In other words, a sharp pulse of light entering the system may be spread out in time some 37 [ns] for each kilometre of fiber traversed. Moreover, travelling at a speed $v_f = c/n_f = 2.0\cdot 10^8\,[\text{m}\cdot\text{s}^{-1}]$, it may spread in space over a length of $7.4\,[\text{m}\cdot\text{km}^{-1}]$. To make sure that the transmitted signal will still be easily readable, we might require that the spatial (or temporal) separation be at least twice the spreadout width. Now imagine a line to be 1.0 [km] long.

In that case, the output pulses are 7.4 [m] wide on emerging from the fiber and so should be separated by 14.8 [m]. This means that the input pulses must be at least 14.8 [m] apart; they must then be separated in time by 74 [ns] and so cannot come any faster than one every 74 [ns], which is a rate of 13.5 million pulses per second. In this way the inter-modal dispersion (which is typically 15 to $30\,[\text{ns}\cdot\text{km}^{-1}]$) limits the frequency of the input signal, thereby dictating the rate at which information can be fed through the system. Stepped-index multimode fibers are used for low-speed, short-distance lines.

These large-core fibers are used mainly in image transmission and illumination bundles. They're also useful for carrying high-power LASER beams where the energy is distributed over a larger volume, thereby avoiding damage to the fiber.

For more details the reader can refer again to [239] where twelve pages are dedicate to the subject. Otherwise there are plenty of textbooks or handbooks of a few hundreds or even thousands of pages dedicated only to that topic.

Optical fiber are used in a wide range of domestic, industrial and scientific applications but they are especially well known for the Internet infrastructure as illustrated below for Europe:

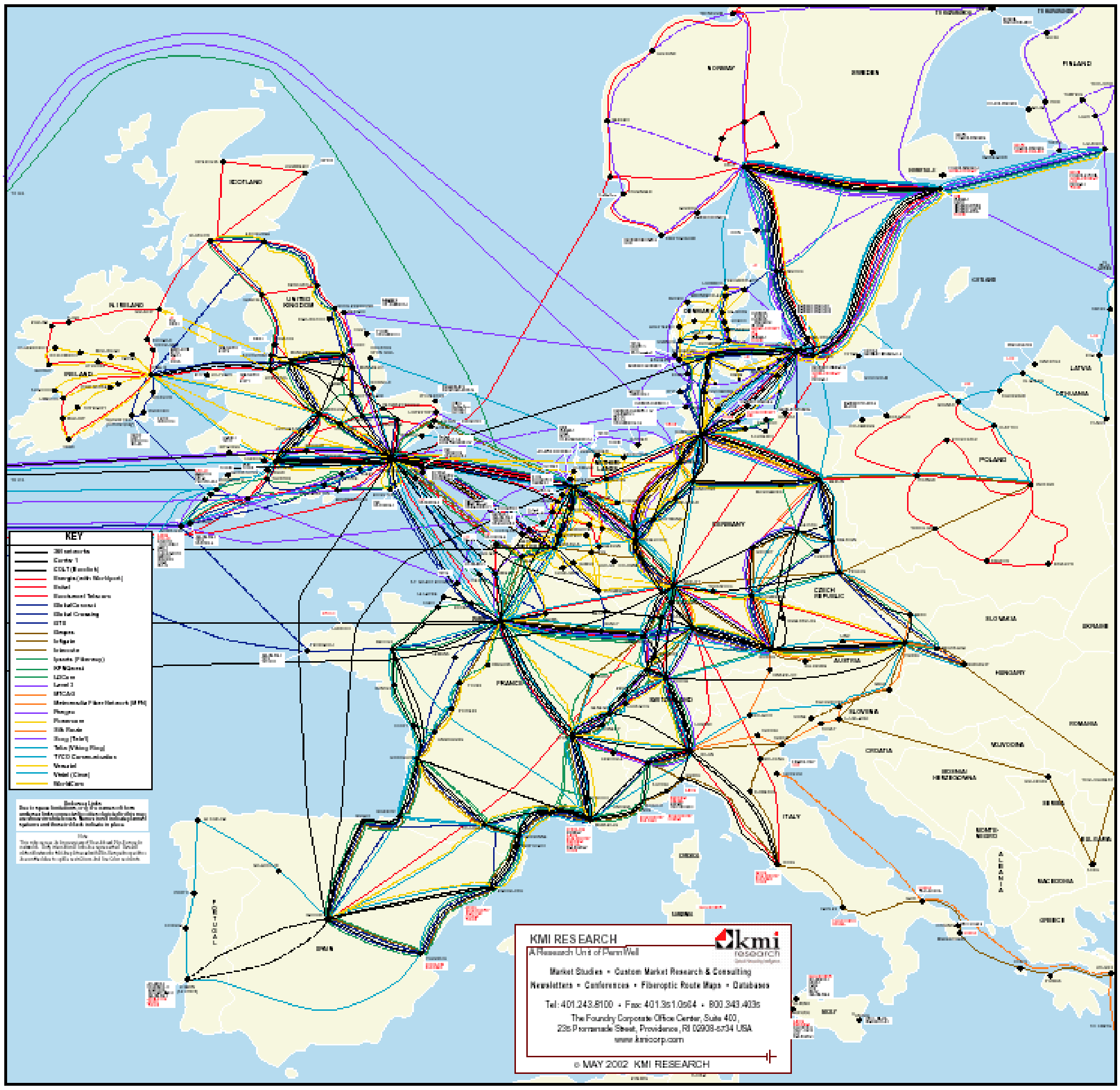

NORWAY
SWEDEN
FINLAND
SCOTLAND
N. IRELAND
IRELAND
UNITED KINGDOM
DENMARK
LATVIA
LITHUANIA
POLAND
GERMANY
CZECH REPUBLIC
SLOVAKIA
UKRAINE
AUSTRIA
HUNGARY
SLOVENIA
CROATIA
ROMANIA
FRANCE
SWITZERLAND
ITALY
SPAIN
PORTUGAL
BOSNIA HERZEGOVINA
SERBIA
BULGARIA
MACEDONIA
ALBANIA
GREECE
KEY
KMI RESEARCH
A Research Unit of PennWell
kmi research
Market Studies • Custom Market Research & Consulting
Newsletters • Conferences • Fiberoptic Route Maps • Databases
Tel: 401.243.8100 • Fax: 401.351.0564 • 800.343.4035
The Foundry Corporate Office Center, Suite 400,
235 Promenade Street, Providence, RI 02908-5734 USA
www.kmicorp.com
© MAY 2002 KMI RESEARCH

Figure 10.286: **Pan-European fiber optic networks (planned or in place) (source: Rutherford et al. (2004, Figure 1), copyright KMI Research)**

Or for the World:

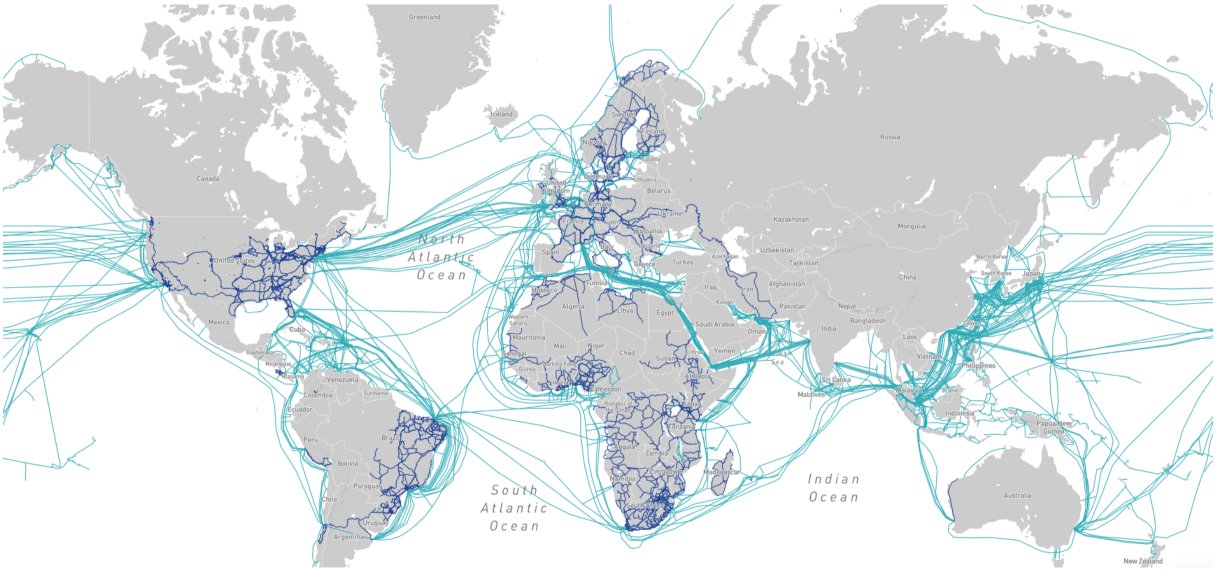

Greenland
Iceland
Canada
Russia
Kazakhstan
Mongolia
China
United States
Mexico
North Atlantic Ocean
South Atlantic Ocean
Indian Ocean
Brazil
Peru
Bolivia
Colombia
Venezuela
Ecuador
Chile
Paraguay
Uruguay
Argentina
Algeria
Libya
Egypt
Mali
Niger
Chad
Sudan
Mauritania
Turkey
Iraq
Saudi Arabia
Yemen
Oman
Iran
Afghanistan
Pakistan
India
Nepal
Bangladesh
Uzbekistan
Tajikistan
Laos
Vietnam
Philippines
Japan
South Korea
North Korea
Indonesia
Papua New Guinea
Australia
New Zealand
Madagascar
Zambia
Namibia
Angola
Tanzania
Sri Lanka
Maldives
Cuba
Spain
Belarus
Ukraine
Finland
Sweden

Figure 10.287: **Infrapedia global Internet infrastructure fiber**

$\sim$90 %

★★★☆☆

30 votes, 64.00%

11

Atomistic

The atomic physics is the part of physics that deals with quantified energy states of corpuscular and wave particles and of exchange of energies within the atom. (Larousse)

Contents

11.1 Corpuscular Quantum Physics . . . 4616
  11.1.1 Dalton's Model . . . 4620
  11.1.2 Thomson's Model . . . 4622
  11.1.3 Rutherford's Model . . . 4623
  11.1.4 Bohr's Model . . . 4626
    11.1.4.1 Bohr's Postulates . . . 4626
    11.1.4.2 Quantification . . . 4627
    11.1.4.3 Hydrogen Type Atoms Model without dragging . . . 4629
    11.1.4.4 Hydrogen Type Atoms Model with dragging . . . 4638
    11.1.4.5 Neutron Assumption . . . 4641
  11.1.5 Wilson and Sommerfeld's Model . . . 4644
  11.1.6 Relativistic Sommerfeld Model . . . 4648
    11.1.6.1 Magnetic dipole moment . . . 4666
    11.1.6.2 Spin . . . 4669
    11.1.6.3 Pauli exclusion principle . . . 4672
  11.1.7 Electron configuration (atomic orbital) . . . 4675
11.2 Wave Quantum Physics . . . 4688
  11.2.1 Postulates . . . 4690
    11.2.1.1 1st Postulate: Quantum State . . . 4691
    11.2.1.2 2nd postulate: Time evolution of a quantum state . . . 4694
    11.2.1.3 3rd postulate: Observables and operators . . . 4695
    11.2.1.4 4th postulate: Measure of a property (collapse postulate) . . . 4697

11.2.1.5 5th postulate: Average of a property . . . 4699

11.2.2 Classical principles of uncertainty . . . 4704

11.2.2.1 First classical uncertainty relation . . . 4705

11.2.2.2 Second classical uncertainty relation . . . 4706

11.2.2.3 Third classical uncertainty relation . . . 4708

11.2.3 Quantum algebra . . . 4711

11.2.3.1 Linear functional operators . . . 4711

11.2.3.2 Robertson uncertainty relations . . . 4733

11.2.3.3 Fourier uncertainty principle . . . 4744

11.2.4 Schrödinger Model . . . 4749

11.2.4.1 de Broglie associated wave . . . 4749

11.2.4.2 Classical (time independent) Schrödinger Wave Equation . . . 4758

11.2.4.3 Classical Shrödinger equation of evolution . . . 4769

11.2.4.4 Implications and Applications . . . 4784

11.2.4.5 Superposition principle . . . 4857

11.2.4.6 Ehrenfest theorems . . . 4870

11.2.5 Angular momentum and Spin . . . 4877

11.2.5.1 Spin–orbit interaction (LS coupling) . . . 4899

11.2.5.2 Spin operator for spin $1/2$ particles . . . 4903

11.2.6 Planck’s dimensions . . . 4905

11.2.7 Wave Quantum Physics Interpretation . . . 4911

11.2.7.1 Copenhagen interpretation . . . 4911

11.2.7.2 De Broglie-Bohm (pilot-wave) interpretation . . . 4915

11.3 Relativistic Quantum Physics . . . 4919

11.3.1 Relativistic Schrödinger evolution equation . . . 4920

11.3.1.1 Antimatter . . . 4922

11.3.2 Generalized Klein-Gordon Equation . . . 4924

11.3.2.1 Generalized Klein-Gordon Equation continuity equation . . . 4927

11.3.3 Classical free Dirac equation . . . 4931

11.3.3.1 Gamma, alpha, beta matrices . . . 4941

11.3.3.2 Free particle solution . . . 4946

11.3.3.3 Free particle solution at rest . . . 4952

11.3.3.4 Majorana equation . . . 4954

11.3.4 Linearised Dirac Equation . . . 4957

11.3.4.1 Linearised Dirac Equation continuity equation . . . 4970

11.3.4.2 Linearised Dirac Equation Angular Momentum + Helicity Conservation . . . 4973

11.3.5 Generalized Dirac Equation . . . 4978

11.3.6 Pauli Equation 4979
11.3.6.1 Landé g-factor 4989
11.3.6.2 Lamb Shift 4996
11.3.7 CPT Symmetries 5002
11.3.7.1 Charge symmetry 5002
11.3.7.2 Parity symmetry 5003
11.3.7.3 CP symmetry (from Matter to Antimatter) 5005
11.3.7.4 Time Reversal symmetry 5005
11.3.7.5 CP and T symmetries 5006
11.3.7.6 CPT violations 5007
11.3.7.7 CPT classical analogy (with Maxwell equations) 5011
11.3.7.8 CPT with Dirac equation 5013
11.4 Nuclear Physics 5019
11.4.1 Nuclear Weapon 5019
11.4.2 Radioactivity 5023
11.4.2.1 Disintegration 5029
11.4.2.2 Activity 5033
11.4.2.3 Two level radioactive cascade 5051
11.4.2.4 Radioactive phenomena 5056
11.4.3 Radiation protection 5092
11.4.3.1 Bethe formula 5092
11.4.3.2 Compton scattering 5096
11.4.3.3 Rayleigh scattering 5102
11.4.3.4 Photoelectric effect 5106
11.4.3.5 Rutherford scattering 5111
11.4.3.6 X-rays and Gamma rays 5120
11.4.3.7 Radiation dosimetry 5129
11.4.4 Liquid Drop Model of Nucleus 5141
11.4.4.1 Volumetric binding energy 5143
11.4.4.2 Superficial binding energy 5143
11.4.4.3 Coulomb electrostatic repulsive energy 5144
11.4.4.4 Energy of asymmetry (Pauli energy) 5145
11.4.4.5 Pairing Energy 5149
11.5 Quantum Field Theory 5158
11.5.1 Yukawa potential 5163
11.5.1.1 Mass fields 5166
11.5.1.2 Non-mass fields 5167

11.5.2 Euler-Lagrange equation for Fields 5169
11.5.2.1 Field interactions ($\lambda\varphi^4$ theory) 5182
11.5.3 Gauge Theories 5185
11.5.3.1 Global Gauge invariance 5186
11.5.3.2 Local Gauge invariance 5188
11.6 Elementary Particle Physics 5194
11.6.1 Coupling Constants 5196
11.6.2 Spin magnetic resonance 5206

11.1. Corpuscular Quantum Physics

Now it is time to plunge into the dark and impenetrable waters of atomic physics. It goes without saying that we will cover the theories of atomic physics only in the outline. In facts, we will limit ourselves only to the theoretical developments made between the years 11910 and about 11935 according to holocene calendar (beyond the complexity of theories requires too many pages for a general textbook like this one). We will also pass on many mathematical details that have already been proven and checked in other sections of this textbook.

> Caution!!!! The word "quantum" has become a favorite buzzword, tossed around by marketers, self-proclaimed gurus, and charlatans. If you want to spot quantum quackery then dive into real science and read the next sections of Corpuscular Quantum Physics, Wave Quantum Physics, Quantum Field Theory and Atomistic. Then you will be able to understand the principles and arm yourself against manipulations. Remember, quantum isn't a shortcut to enlightenment; it's a rigorous field of study! So, next time someone peddles quantum-infused nonsense, raise an eyebrow, reach for that textbook, and separate the genuine from the gibberish.

Atomic physics as you probably already know is the world of the infinitely small (points of zero dimension). This is a world, you'll see, almost special where fairly classical laws, those that govern our everyday macroscopic, does not apply in an intuitive way.

Thus, in the early 120th century (holocene calendar) we knew only that atoms were formed by a simple core and electrons in orbit.

The electron, the first subatomic particle (smaller than the atom) to be revealed, was showed to us by experiments on electrical currents in solids, liquids and gases. In the 119th century (holocene calendar), physicists had no idea what was the charge, if it was continuous or particulate. Today we know that the charge seems to be a property of matter and that the total load in a system seems to be a multiple of an elementary charge corresponding to the charge of an electron (or proton).

Definition (#11.1.499): A "particle" in the field of Atomistic is a concept useful to describe through mathematical tools what happens at the atomic level. Particles can be fundamental (not having any constituents) or at the opposite be composite particles (typically made of other particles name "quarks"). Composite particles can very often be considered point-like, elementary particles are truly punctual in the Standard Model! Particles are all made of energy but professional physicists like better to define them as: irreducible representation of the symmetries of the lagrangian[1] (it's a pretty wordy definition, enough to end any conversation on the subject, but it's nonetheless a completely mathematical thing!).

---

[1] As in particle physics particles can in fact be identify by how they transform under the group of symmetries, we think of particles as irreducible unitary representations of the Poincare group.

Michael Faraday suggested by electrolysis experiments that electricity consisted of particles of elementary charge e and that a mole of such charges (see Chemistry section for the definition of the mole) was equivalent to a load of 1 Faraday that is to say $96,485\,[\text{C}]$. As Avogadro's number was not known at the time, it was not possible to determine e. However, one mole of a monovalent substance that can carry $1\,[\text{F}]$ charge, it had to follow only half a mole of the same substance would carry $1/2\,[\text{F}]$ and so on until the most smallest unit of charge e, which was to be carried by the smallest unit of mass m, corresponding to the mass of a single atom of the substance. In 11881 (holocene calendar), Helmholtz stated that if one accepts the hypothesis that the elementary substances were composed of atoms, we must logically conclude that electricity, both positive and negative, should be divided into finite portions that should behave like electricity atoms. George Stoney named this fundamental unit charge "electron". The elementary charge value is named prosaically today "quantum charge".

All subatomic charge known today whether positive or negative, carry a net charge that is an integer multiple of e. Quarks have a fractional charge but they do not appear as isolated entities. There are also fractional charges in the quantum Hall effect, but this is another story...

Even today, some of the best physicists say that they don't really know what are an electron or even an atom. In fact, we still do not know what really what matter is (matter is energy but what is energy??)... The only thing that interest the scientist anyway is to have a mathematical tool that explains thinks quite well its field of interest and the interpretation in comparison with the sensible reality is not the main priority...

Scientists have tried the development of several models to explain the experimental observations of the microscopic world. Thus, there has been in the order for the must well known one: the models of Dalton, Thomson, Rutherford, Bohr, Schrödinger, Sommerfeld (the latter including the major contributions of Heisenberg, de Broglie, Pauli, Dirac and Einstein for the most famous) and the Standard Model.

We can situate the birth of the corpuscular quantum physics or simply "quantum physics" ("quantum" meaning "fixed amount") in 11900 (holocene calendar), when Max Planck presented his famous paper on black-body radiation (see section Thermodynamics page 3704) at a meeting of the German physical Society and the inability of classical physics (mechanics, thermodynamics, electromagnetism) to explain certain behaviour of matter at the microscopic level, that is to say at the level of phenomena where particles of small masses located in very small regions of space are the main object for study.

To arrive to a coherent interpretation of these experiments, it was necessary to introduce concepts radically different from those of classical physics. For example, we had to abandon the notion of classical trajectory, adopt the quantification of energy (Planck's law) and consider that microscopic particles sometimes have a behaviour similar to a wave. All these new concepts gave birth to a new physics, "quantum physics", which has developed rapidly since in 11927 (holocene calendar), already, the foundations of the theory were completed. By abandonment of the key concepts of classical mechanics, we can say that quantum physics is a revolution (named also the "second revolution", the first being the theory of relativity) in the way of interpreting experimental measurements. With relativity introduced by Einstein, quantum physics is one of the pillars of the theoretical edifice of modern physics in the 121st century (holocene calendar).

Just as relativity contains classical mechanics as a limiting case (relativistic laws approaching classical laws when the velocity of a particle is sufficiently low compared to that of light), the new quantum physics contains also in the limit case the classical laws of statistical mechanics or even electromagnetism.

Remark

We will see that the fundamental constant characterizing quantum physics (like the speed of light characterizes relativity) is the Planck constant and deeper the fine-structure constant.

First photo ever of a single positively-charged strontium atom emitting light (illuminated by a blue-violet light) in the visible spectrum through a window of a ultra-high vacuum chamber, held near motionless by an electric field trap (strontium has 38 protons and is approximately 215 billionths of a millimetre across) :

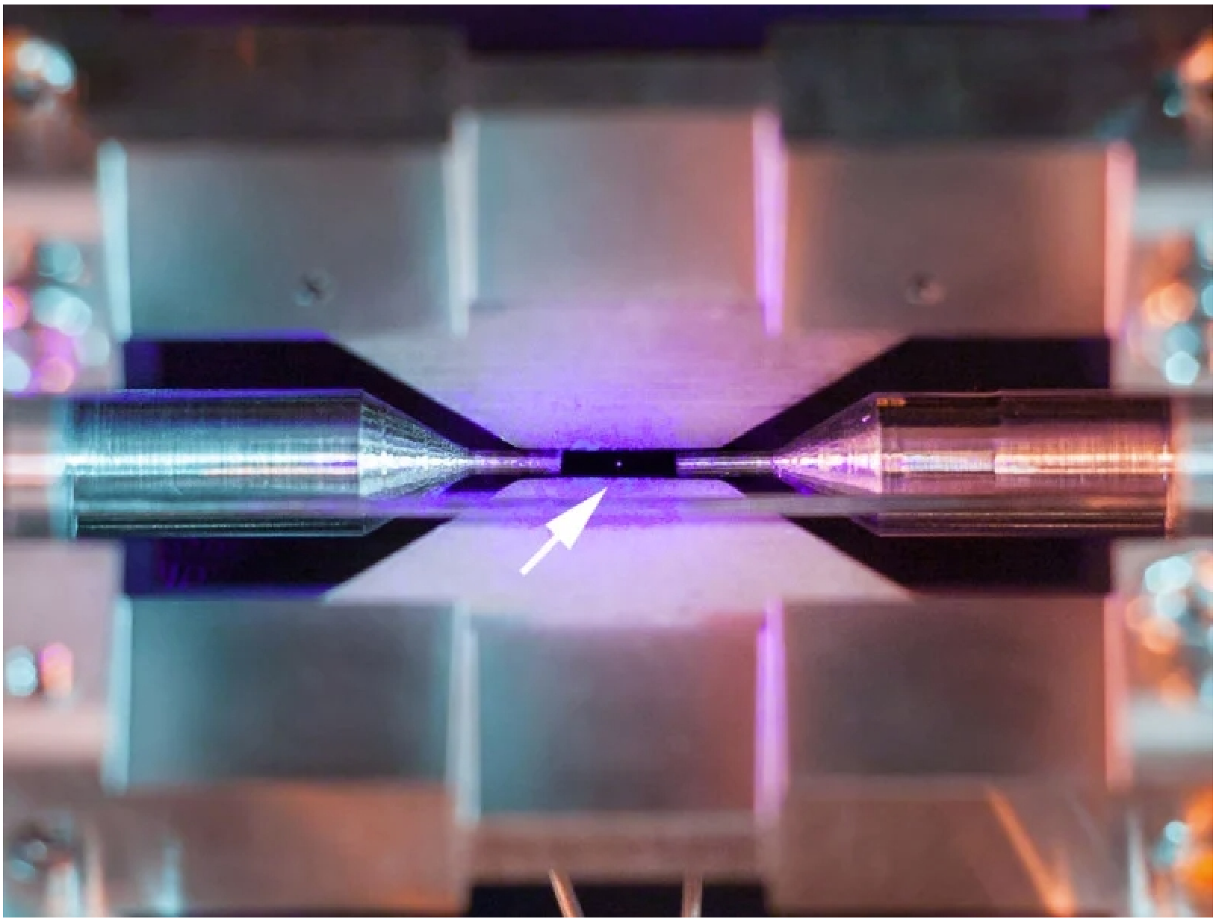


Figure 11.1: **First photo ever of a single atom emitting light (credit: David Nadlinger, University Of Oxford)**

The photo was captured on August 7th, 12017 (holocene calendar(, using a Canon 5D Mark II DSLR, a Canon EF 50mm f/1.8 lens, extension tubes, and two flash units with color gels. The distance between the ion trap's small needle tips is less than 0.2 [cm].

Oviously the only reason why we can see the atom in the photo is because it absorbed and then re-emitted LASER light at a speed capturable by a long camera exposure. So, the photo is actually of the LASER light being re-emitted, rather than the outline of an atom. Without the long exposure effect, the atom wouldn't be visible to the naked eye.

11.1.1 Dalton's Model

In 11803 (holocene calendar), John Dalton made the assumption that matter is composed of atoms of different masses and combine respecting simple mass proportions (though the idea of atom was really not new, it dates from thousands years before!). It is this theory that Dalton proposed that is the cornerstone of modern physical science. In 11808 (holocene calendar), Dalton's work entitled *A new system of chemical philosophy* was published. In this textbook he listed the atomic weights of a number of known elements relative to the mass of hydrogen. Its masses "AMU" (see further below page 5027) were not entirely correct, but they are the basis for the modern periodic table of elements. Dalton arrived at his atomic theory through a study of the physical properties of atmospheric air and other gases[2].

Dalton assumed that the atom was a sphere:

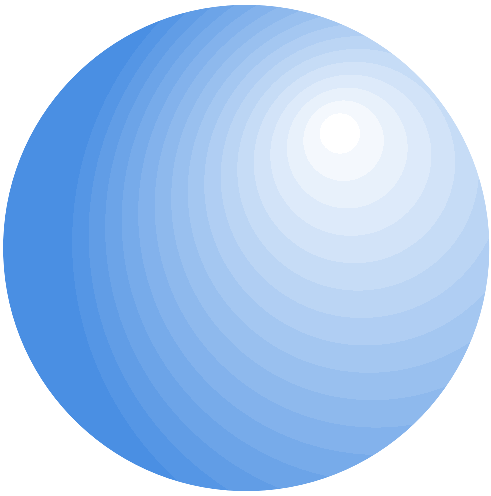


Figure 11.2: The ideal approach of Dalton

Thus, he could make a first estimate of the size of atoms:

Indeed, either ρ the typical density, m_A the atomic mass and R the radius (unknown value) of an element which we seek to determine the size of the atom. We then have very simply:

$$\rho\frac{4}{3}\pi R^3=m_A\Leftrightarrow R=\sqrt[3]{\frac{3}{4}\frac{m_A}{\rho\pi}} \tag{11.1.1}$$

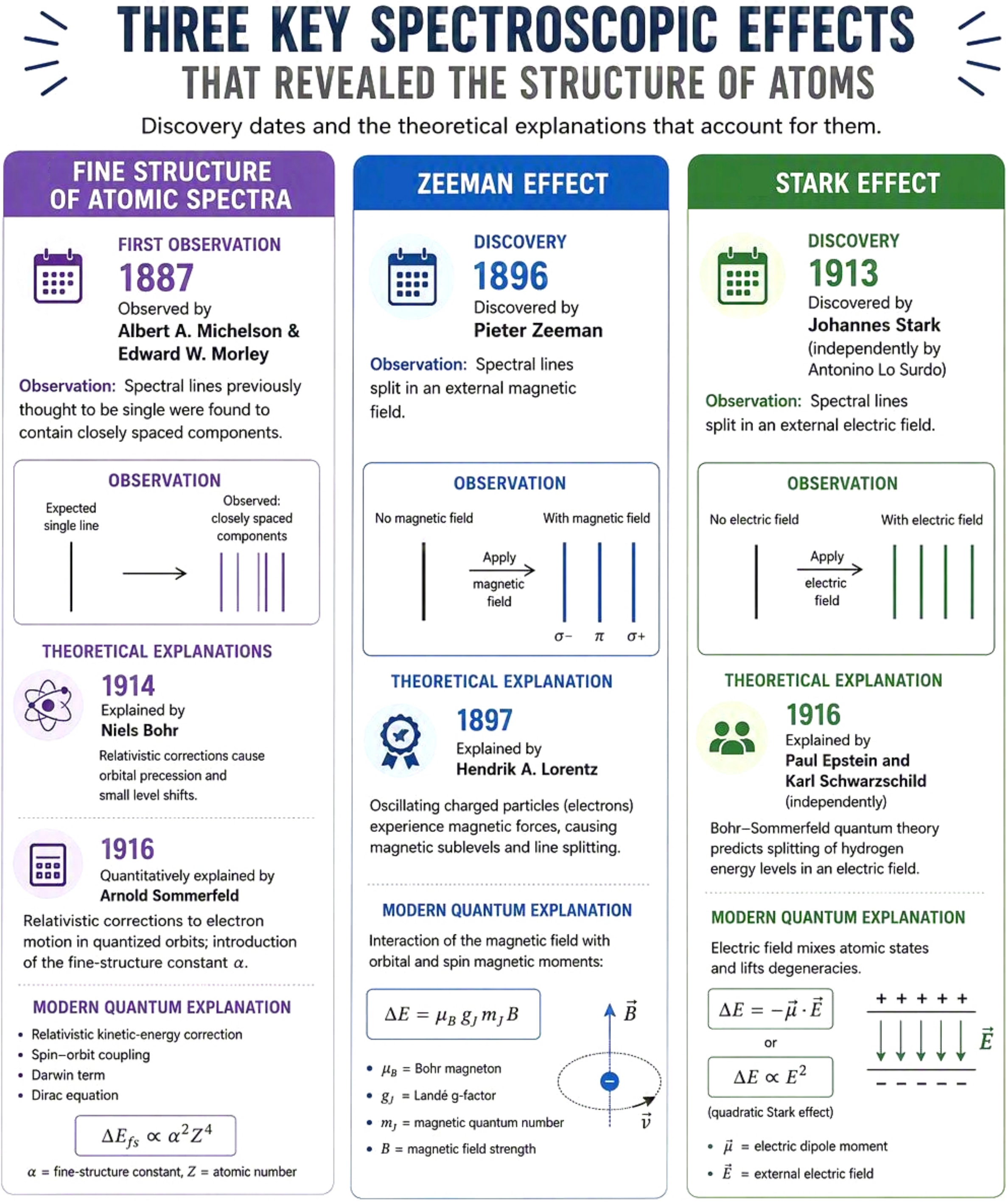
Dalton knowing ρ and m_A thanks to experimentation, he get:

$$R\cong 1\cdot 10^{-10}\;[\text{m}] \tag{11.1.2}$$

After this model three important spectroscopic observations of the hydrogen atom have boosted the field experimentally and theoretically! Here is an infography summarizing these three observations (the dates are based on the Gregorian calendar):

---

[2] References to the concept of atomism and its atoms appeared in both ancient Greek and ancient Indian philosophical traditions. Leucippus (year 9500 according to holocene calendar) is the earliest figure whose commitment to atomism is well attested and he is usually credited with inventing atomism. He and other ancient Greek atomists theorized that nature consists of two fundamental principles: atom and void. Clusters of different shapes, arrangements, and positions give rise to the various macroscopic substances in the world.



THREE KEY SPECTROSCOPIC EFFECTS
THAT REVEALED THE STRUCTURE OF ATOMS

Discovery dates and the theoretical explanations that account for them.

FINE STRUCTURE OF ATOMIC SPECTRA

FIRST OBSERVATION

1887

Observed by
Albert A. Michelson & Edward W. Morley

Observation: Spectral lines previously thought to be single were found to contain closely spaced components.

OBSERVATION

THEORETICAL EXPLANATIONS

1914
Explained by
Niels Bohr
Relativistic corrections cause orbital precession and small level shifts.

1916
Quantitatively explained by
Arnold Sommerfeld
Relativistic corrections to electron motion in quantized orbits; introduction of the fine-structure constant α.

MODERN QUANTUM EXPLANATION

- Relativistic kinetic-energy correction
- Spin–orbit coupling
- Darwin term
- Dirac equation

$$\Delta E_{fs} \propto \alpha^2 Z^4$$

α = fine-structure constant, Z = atomic number

ZEEMAN EFFECT

DISCOVERY

1896

Discovered by
Pieter Zeeman

Observation: Spectral lines split in an external magnetic field.

OBSERVATION

THEORETICAL EXPLANATION

1897
Explained by
Hendrik A. Lorentz

Oscillating charged particles (electrons) experience magnetic forces, causing magnetic sublevels and line splitting.

MODERN QUANTUM EXPLANATION

Interaction of the magnetic field with orbital and spin magnetic moments:

$$\Delta E = \mu_B \, g_J \, m_J \, B$$

- μ_B = Bohr magneton
- g_J = Landé g-factor
- m_J = magnetic quantum number
- B = magnetic field strength

STARK EFFECT

DISCOVERY

1913

Discovered by
Johannes Stark
(independently by Antonino Lo Surdo)

Observation: Spectral lines split in an external electric field.

OBSERVATION

THEORETICAL EXPLANATION

1916
Explained by
Paul Epstein and Karl Schwarzschild
(independently)

Bohr–Sommerfeld quantum theory predicts splitting of hydrogen energy levels in an electric field.

MODERN QUANTUM EXPLANATION

Electric field mixes atomic states and lifts degeneracies.

$$\Delta E = -\vec{\mu}\cdot\vec{E}$$

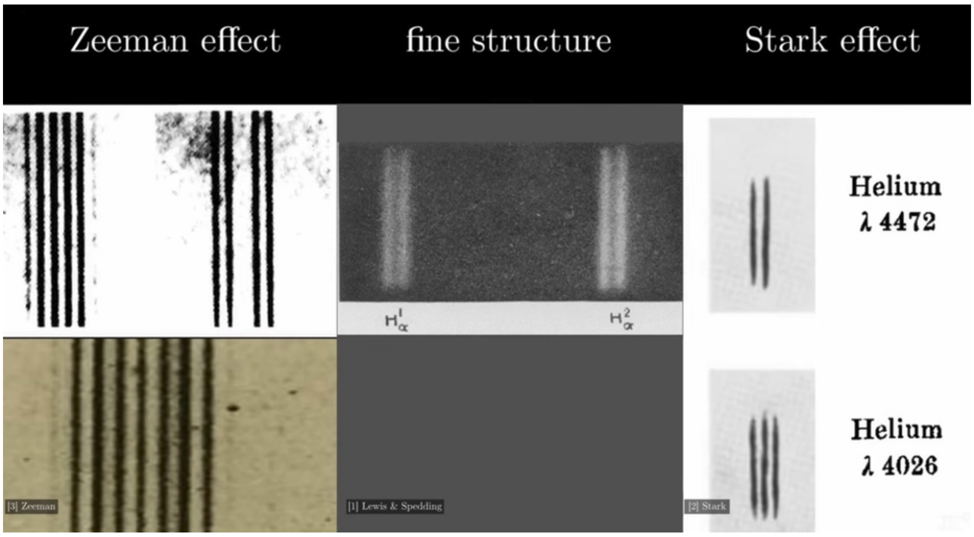
or

$$\Delta E \propto E^2$$

(quadratic Stark effect)

- $\vec{\mu}$ = electric dipole moment
- $\vec{E}$ = external electric field

And here are some of the early observations for each of the effect:


Zeeman effect
fine structure
Stark effect
Helium
λ 4472
Helium
λ 4026
H$_\alpha^1$
H$_\alpha^2$
[3] Zeeman
[1] Lewis & Spedding
[2] Stark

11.1.2 Thomson's Model

Thomson is at the origin of the discovery of the electron by his experiments on the flow of particles (electrons) generated by cathode rays. Theoretician and experimenter, Thomson advanced in 11898 (holocene calendar) the "raisin bread theory" (see [539]) of atomic structure, in which the electrons are considered negative grapes pressed into a loaf of positive material. His model of the atom is represented by the figure below:

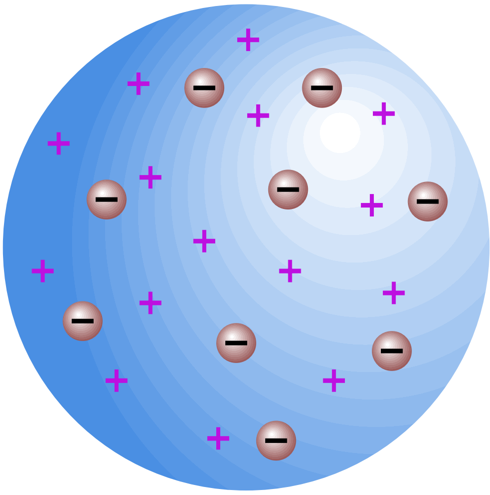


Figure 11.3: **Thomson's Gourmet approach...**

But, we know (the physicists of the 119th century also knew it) that no arrangement of static charges is stable if these charges are under the influence of the Coulomb force:

$$\vec{F}=q\left(\frac{1}{4\pi\varepsilon}\sum_i\frac{Q_i}{\|\vec{r}-\vec{r}_i\|^3}(\vec{r}-\vec{r}_i)\right) \tag{11.1.3}$$

that we had studied in detail in the section of Electrostatic. Thus that this requires that the particles that make up the atom are in motion which leads us to develop another model: the following "Rutherford model".

11.1.3 Rutherford's Model

Thus Ernest Rutherford assimilated intuitively by this theoretical fact, a few years after the discovery of Thomson, the atom to a planetary system whose center was occupied by a nucleus of positive charge, which contained almost the entire mass of the atom. The core was a hundred thousand times smaller than the atom and thus occupied a very small volume. The assumptions was more adapted to experimental results (11909 according to holocene calendar, see [471]):

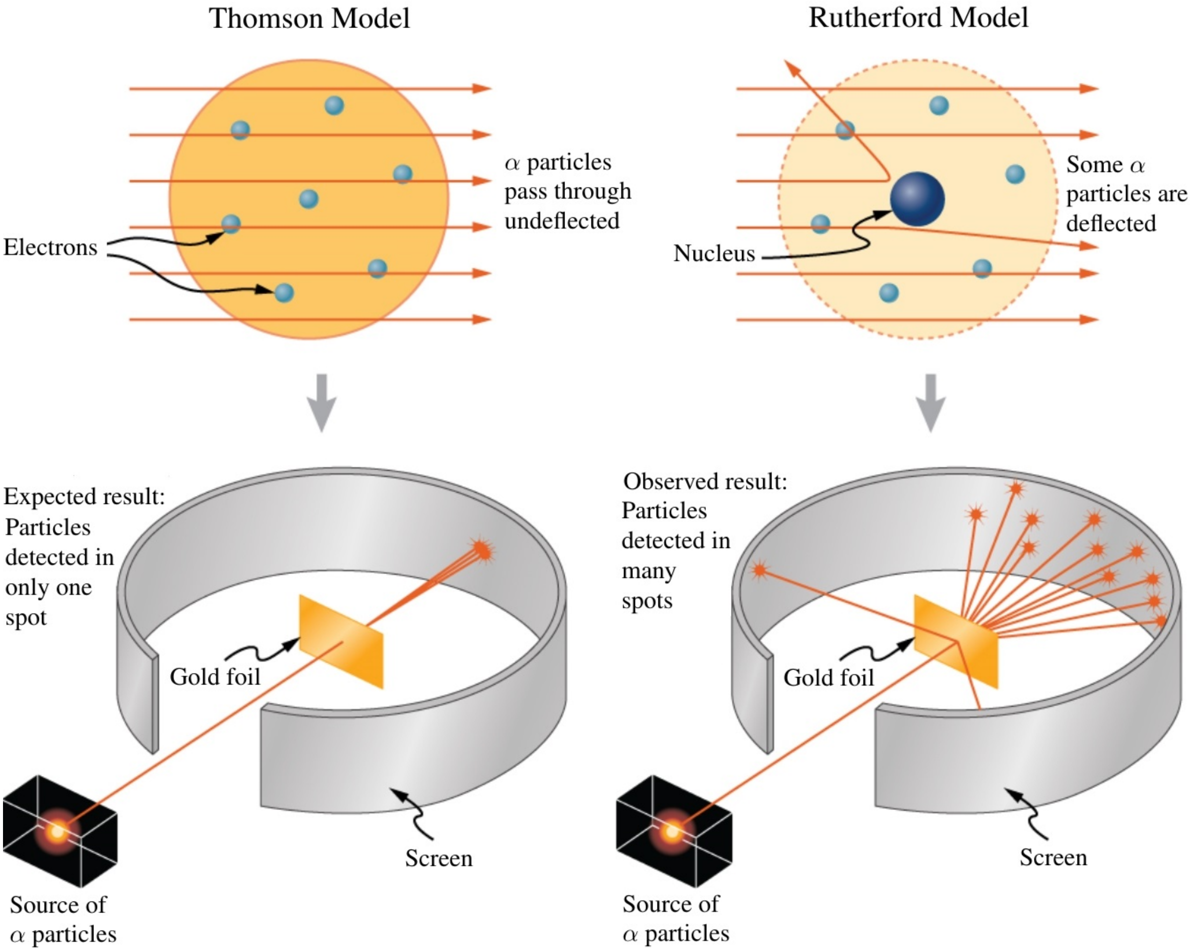

Thomson Model
Rutherford Model
α particles pass through undeflected
Electrons
Nucleus
Some α particles are deflected
Expected result: Particles detected in only one spot
Observed result: Particles detected in many spots
Gold foil
Gold foil
Screen
Screen
Source of α particles
Source of α particles

Figure 11.4: Rutherford experiment (source: OpenStax)

Therefore here is a pictorial representation of how should looks like an atom in this model (the distances are not at scale for obvious reasons...):

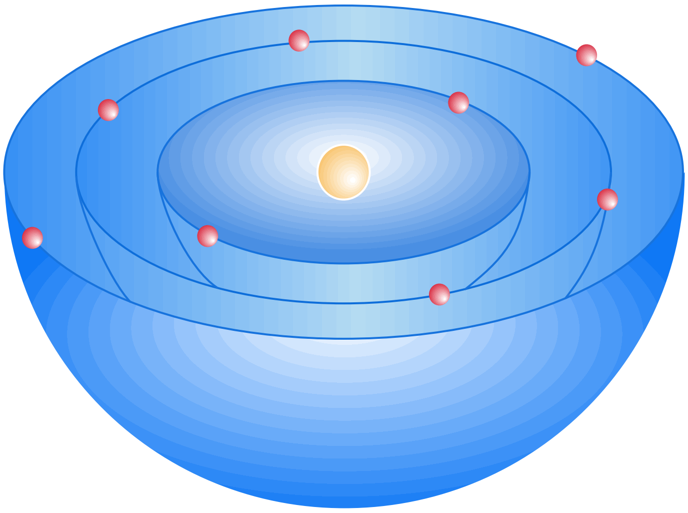


Figure 11.5: **The Rutherford's planetary approach...**

Rutherford applied the results we get in astronomy (see section Astronomy page 5234) during our study of Keplerian orbits in the atom and therefore obtained conical trajectories for the rotation of the electron around the nucleus such as:

$$r = \frac{p}{1+\cos\theta} \tag{11.1.4}$$

where e is the eccentricity ($e = c/a < 1$) and p the focal parameter ($p = b^2/a$) of an ellipse (see section Analytical Geometry page 2779), and where:

$$p = \frac{K^2}{\mu} \qquad \text{and} \qquad e = \sqrt{1 + \frac{2EK^2}{\mu^2}} \tag{11.1.5}$$

Remarks

R1. You will have to remember that when we will discuss later on Bohr's model that in the Rutherford model, r can take theoretically any value!

R2. We will see in our study of the Rutherford scattering (see section Nuclear Physics page 5111) that Rutherford determined that the size of the gold atom as being worth $R \cong 3\cdot 10^{-14}\,[\text{m}]$. So we have a model with a factor 10,000 in comparison of the Dalton's model (that is to say...).

But we have proven that during our study of electromagnetism that Maxwell displacement equations (third Maxwell equation) were given by (see section Electrodynamics page 4101):

$$U(t) = -\frac{\mathrm{d}\Phi_B}{\mathrm{d}t} = \oint_\Gamma \vec{E}\circ \mathrm{d}\vec{l} \Leftrightarrow \vec{\nabla}\times\vec{E} = -\frac{\partial \vec{B}}{\partial t} \tag{11.1.6}$$

and:

$$\oint_\Gamma \vec{B}\circ \mathrm{d}\vec{l} = \mu \iint_\Sigma \vec{j}\circ \mathrm{d}\vec{S} + \mu\varepsilon \iint_\Sigma \vec{E}\circ \mathrm{d}\vec{S} \Leftrightarrow \vec{\nabla}\times\vec{H} = \vec{j} + \varepsilon\frac{\partial \vec{E}}{\partial t} \tag{11.1.7}$$

describe that an electron in motion (accelerated) emits energy in the form of electromagnetic radiation that we name in physics "Bremsstrahlung" explained by the Liénard-Wiechert's potentials (see section Electrodynamics page 4168).

Rutherford and Thomson were therefore faced with the following dilemma:

If the electron emits energy in the form of electromagnetic radiation, it loses kinetic energy (speed) and thus necessarily sooner or later (except external intervention) will fall on the core of the atom (illustration of the phenomenon in the figure below). But we see that the matter surrounding us is however in fact quite stable...

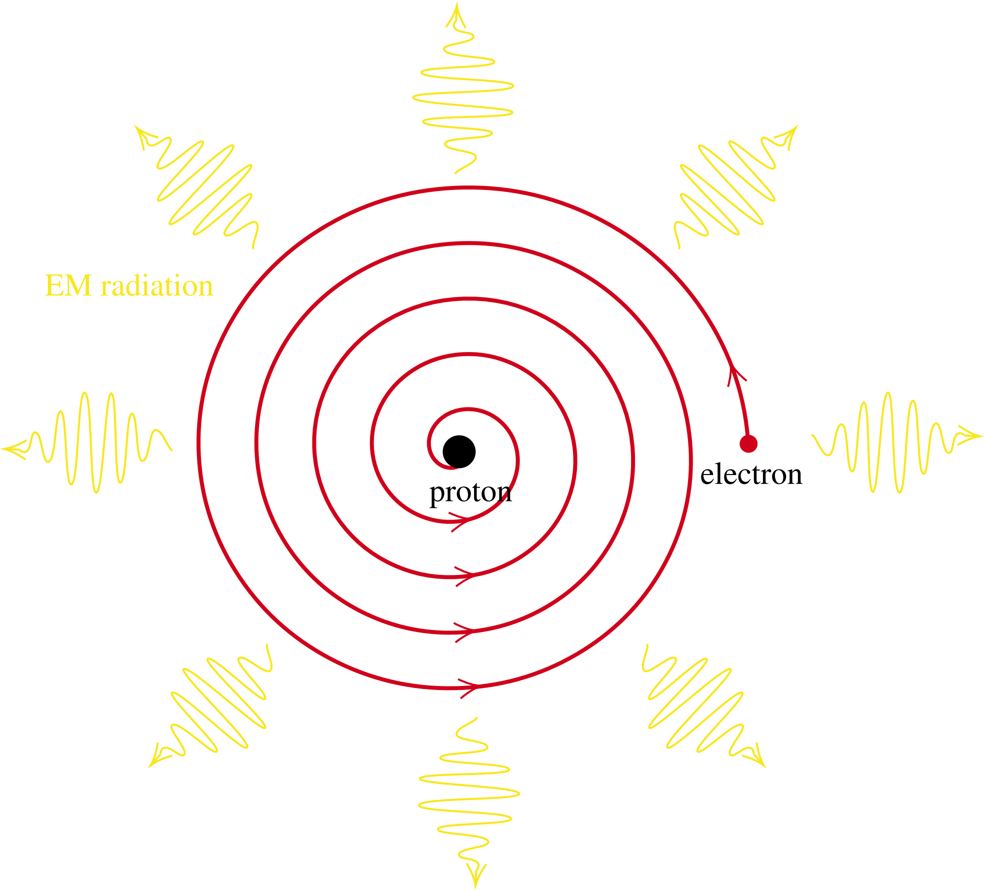


Figure 11.6: **Simplistic illustration of Bremsstrahlung**

So they rejected their model and Bohr intervened then with a revolutionary hypothesis but that had to explain facts and not necessarily true reality... (Bohr was a genius experimentalist physicist and probably the best debate partner of Einstein).

11.1.4 Bohr's Model

In 11913 (holocene calendar), Niels Bohr, who participated to the work of Rutherford on the diffusion α of particles (nuclei of two proton and two neutrons free of electrons), takes the model of Rutherford but includes in it three fundamental assumptions (see [62]):

11.1.4.1 Bohr's Postulates

P1. The electron does not emit radiation when on certain orbits named "stationary orbits". This assertion is contrary to the theories of classical electrodynamics. So this implies that all the orbits are not allowed and this is a revolution in the approach of theoretical physics (the prohibited orbits are named "non-stationary orbits").

P2. On any stable orbit the linear momentum p integrated on the path r of the orbit is an integer multiple of the Planck's constant h (assumption arising from the first one) according to the quantification of energy exchanges established by Planck's relation (see section Thermodynamics page 3719). This assumption is sometimes named "Planck's quantum hypothesis".

P3. The assumed (only experimentally verified and theoretically assumed as unprovable) "Planck-Einstein relation":

$$\boxed{E = h\nu} \tag{11.1.8}$$

applies to the emission[3] or absorption of radiation during the transition of an electron from one energy state E_1 to a state E_2 (which solidifies the first postulate).

In fact, we find here a revolutionary and unprovable concept (today and to our knowledge) consisting in the quantification of certain physical properties.

Remark

The German physicists James Franck and Gustav Hertz performed a collisional excitation experiment in 11914 (holocene calendar) that provided early dramatic evidence for discrete energy levels in atoms. Electrons with varying amounts of initial kinetic energy were sent through a gas of mercury vapor, and the number of electrons per second making it through the gas was measured. When the initial kinetic energy of the electrons was less than $4.9\,[\text{eV}]$, many electrons reached the detector. However, when the initial kinetic energy reached $4.9\,[\text{eV}]$, few electrons reached the detector. The Franck-Hertz experiment was correctly interpreted to mean that the first excited state of a mercury atom is $4.9\,[\text{eV}]$ above the ground state, and an electron with kinetic energy of $4.9\,[\text{eV}]$ can excite the atom, leaving the electron with almost no kinetic energy, so the number of electrons that made it through the gas became very small.

Let us continue our analysis:

---

[3] The time of emission (after absorption) is not fixed and follows also follows a statistical law! This "radiation lifetime" is described by the Wigner-Weisskopf theory of spontaneous decay that we will maybe introduce in this textbook in the future...

11.1.4.2 Quantification

Let M be the mass of the core of an atom with an electric charge $+e$ and m the mass of the electron in "orbit" around the core. We assume that $M \gg m$ and that the central mass is motionless (which is obviously false in reality).

We assimilate the circular motion of the electron around the nucleus to that of a harmonic oscillator (mass connected to a spring exerting a force opposed to a proportional constant k_r to retain the electron linked).

If the oscillations occurs in the plane, its differential equation is (see section Classical Mechanics page 3417):

$$m\frac{\mathrm{d}^2x}{\mathrm{d}t^2}+k_rx=0 \Leftrightarrow \frac{\mathrm{d}^2x}{\mathrm{d}t^2}+\omega_0x=0 \tag{11.1.9}$$

A (special) simple solution of this equation is (see section Differential and Integral Calculus page 1610):

$$x=A_0\sin(\omega_0t) \tag{11.1.10}$$

The kinetic energy of the system is therefore given by (see section Classical Mechanics page 3297):

$$E_c=\frac{1}{2}mv^2=\frac{1}{2}m\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2=\frac{1}{2}mA^2\omega_0^2\cos^2(\omega_0t) \tag{11.1.11}$$

and the potential energy of the system (see section Wave Mechanics page 3418):

$$E_p=\frac{1}{2}k_rx^2=\frac{1}{2}m\omega_0^2A^2\sin^2(\omega_0t) \tag{11.1.12}$$

If we denote by ν the frequency of oscillation of the oscillatory motion, we then have of course (see section Wave Mechanics page 3480):

$$\omega_0=2\pi\nu \tag{11.1.13}$$

The total energy of the system is finally written after summation and simplification (elementary trigonometry):

$$E_t=E_c+E_p=2\pi^2mA^2\nu^2 \tag{11.1.14}$$

We now assume that the bounded electron can only occupy certain energy levels (first hypothesis) according to the "Planck-Einstein relation":

$$\boxed{E=n\varepsilon=nh\nu=\frac{h}{2\pi}=\hbar\omega=h\frac{c}{\lambda}} \tag{11.1.15}$$

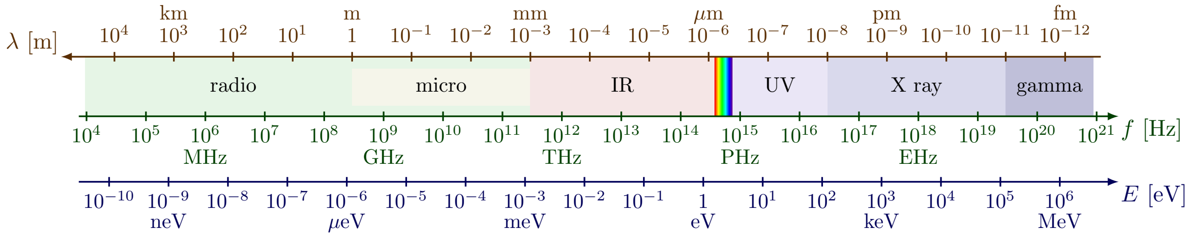

λ [m]
km
m
mm
μm
pm
fm
10^4 10^3 10^2 10^1 1 10^−1 10^−2 10^−3 10^−4 10^−5 10^−6 10^−7 10^−8 10^−9 10^−10 10^−11 10^−12
radio
micro
IR
UV
X ray
gamma
10^4 10^5 10^6 10^7 10^8 10^9 10^10 10^11 10^12 10^13 10^14 10^15 10^16 10^17 10^18 10^19 10^20 10^21 f [Hz]
MHz
GHz
THz
PHz
EHz
10^−10 10^−9 10^−8 10^−7 10^−6 10^−5 10^−4 10^−3 10^−2 10^−1 1 10^1 10^2 10^3 10^4 10^5 10^6 E [eV]
neV
μeV
meV
eV
keV
MeV

Figure 11.7: Energy-Wavelength-Frequency electromagnetic values (author: Izaak Neutelings)

Remark

A nice application of this relation is that since we know the radiation power of a light bulb, a star or an antenna for a given surface and a given wavelength, we can calculate the number of photons (having the corresponding wavelength) per unit surface! Typically at a distance of one meter of a 100 [W] light bulb we get 10^{15} visible photons per square centimetre and per second! For an isotropic FM antenna with a power of 100 [W] at a distance of 100 [km] we have a flux of 10^{12} photons per square centimetre and per second.

This gives us when we include the Planck-Einstein relation in the penultimate equation:

$$A^2=n\frac{h}{2\pi^2 m\nu} \tag{11.1.16}$$

We note here that since the energy of the electron is quantified the amplitude of its movement is also quantified.

Consider now the following integral path also named "action integral" (this is in fact angular momentum):

$$\oint p\mathrm{d}r=\oint mv\mathrm{d}r=\oint mv\frac{\mathrm{d}r}{\mathrm{d}t}\mathrm{d}t=\oint mv^2\mathrm{d}t \tag{11.1.17}$$

and considering the expression for velocity previously obtained:

$$\oint p\mathrm{d}r=\oint mA^2\omega_0^2\cos^2(\omega_0 t)\mathrm{d}t=mA^2\omega_0^2\oint\cos^2(\omega_0 t)\mathrm{d}t \tag{11.1.18}$$

Over a period of revolution, we have:

$$\oint\cos^2(\omega_0 t)\mathrm{d}t=\int_0^T\cos^2(\omega_0 t)\mathrm{d}t \tag{11.1.19}$$

Since (see section Trigonometry page 2519):

$$\cos^2(\omega_0 t)=\frac{1+\cos\left(2\omega_0 t\right)}{2} \tag{11.1.20}$$

The integration becomes:

$$\int_0^T\cos^2(\omega_0 t)\mathrm{d}t=\frac{1}{2}\int_0^T\left[1+\cos(2\omega_0 t)\right]\mathrm{d}t=\frac{1}{2}\left(T+\frac{1}{2\omega_0}\sin(2\omega_0 t)\Big|_0^T\right) \tag{11.1.21}$$

Because $\omega=\dfrac{2\pi}{T}$ (see section Wave Mechanics page 3480) we have:

$$\int_0^T\cos^2(\omega_0 t)\mathrm{d}t=\frac{1}{2}\left(T+\frac{1}{2\omega_0}\left[\sin(2\frac{2\pi}{T}T)-\sin(0)\right]\right)=\frac{1}{2}T \tag{11.1.22}$$

So finally we get:

$$\oint \text{pdr} = \frac{1}{2} m A^2 \omega_0^2 T \tag{11.1.23}$$

Given that $A^2 = n\dfrac{h}{2\pi^2 m\nu}$ and $\omega_0 = 2\pi\nu$ and also $\nu = \dfrac{1}{T}$ we get:

$$\oint \text{pdr} = \frac{1}{2} m A^2 \omega_0^2 T = \frac{1}{2} m \left(\frac{nh}{2\pi^2 m\nu}\right) 4\pi^2\nu^2 \frac{1}{\nu} = nh \tag{11.1.24}$$

Finally:

$$\oint \text{pdr} = nh \tag{11.1.25}$$

This condition imposed by Bohr (second postulate) results from the quantization of energy exchanges (Planck-Einstein relation). This has the effect of imposing stationary energy levels that the electron around the nucleus can occupy.

For a circular orbit (remember though for now that we consider a circular orbit!) or radius r the angular momentum (yes in fact the integral action is just the angular momentum) along the length of the orbital is:

$$\oint \text{pd}l = mv(2\pi r) = nh \Rightarrow mvr = n\frac{h}{2\pi} = n\hbar \tag{11.1.26}$$

or by using one of the traditional notation of angular momentum in quantum physics:

$$\boxed{b = mvr = n\hbar} \tag{11.1.27}$$

The angular momentum is quantized and non-zero in the Bohr's model (since n is not zero). We will see that this is no longer the case in the wave model where the angular momentum can be zero.

According to this result, and further results we will get, it is commonly accepted in science that action values or changes smaller than $\hbar \cong 1.06 \cdot 10^{-32}$ [J · s] are not observed.

11.1.4.3 Hydrogen Type Atoms Model without dragging

We define the study of "hydrogen type atom without dragging" as the fact of considering the atoms with a single electron of mass m rotating around a central core with a charge of $Z \cdot e$ and mass M such that $M \gg m$ (so the kernel is supposed fixed).

Let us now calculate the stationary orbits radius!

On its stationary orbit, the electron is in equilibrium because there is a true antagonism between the Coulomb force and centrifugal (fictional) force. This should result in the following equality of strengths:

$$\frac{Ze^2}{4\pi\varepsilon_0 r^2} = m\frac{v^2}{r} = m\frac{(\omega r)^2}{r} = m\omega^2 r \tag{11.1.28}$$

We will write starting from now (to reduce the notations):

$$k=\frac{1}{4\pi\varepsilon_0} \tag{11.1.29}$$

This allows us to write the relation:

$$\omega^2 r^3=k\frac{Ze^2}{m} \tag{11.1.30}$$

By using the quantization condition of Bohr and by squaring:

$$m\omega r^2=n\hbar=n\frac{h}{2\pi}\Leftrightarrow\omega^2 r^4=n^2\frac{h^2}{4\pi^2 m^2} \tag{11.1.31}$$

By dividing the last two relations to one another by:

$$\frac{\omega^2 r^2}{\omega^2 r^3}=\frac{\dfrac{n^2h^2}{4\pi^2m^2}}{k\dfrac{Ze^2}{m}} \tag{11.1.32}$$

we get:

$$r=\frac{1}{k}\frac{n^2h^2}{4\pi^2 mZe^2}=n^2\frac{\varepsilon_0 h^2}{\pi mZe^2} \tag{11.1.33}$$

given the expression of k.

The radius of the electron orbits allowed is then:

$$\boxed{r_n=n^2\frac{\varepsilon_0 h^2}{\pi mZe^2}} \tag{11.1.34}$$

with $n\in\mathbb{N}^*$. This relation is commonly named the "Bohr radius" for $n=Z=1$ and denoted r_0 (or a_0 in some textbooks).

The orbits of an atom in this model therefore looks like:

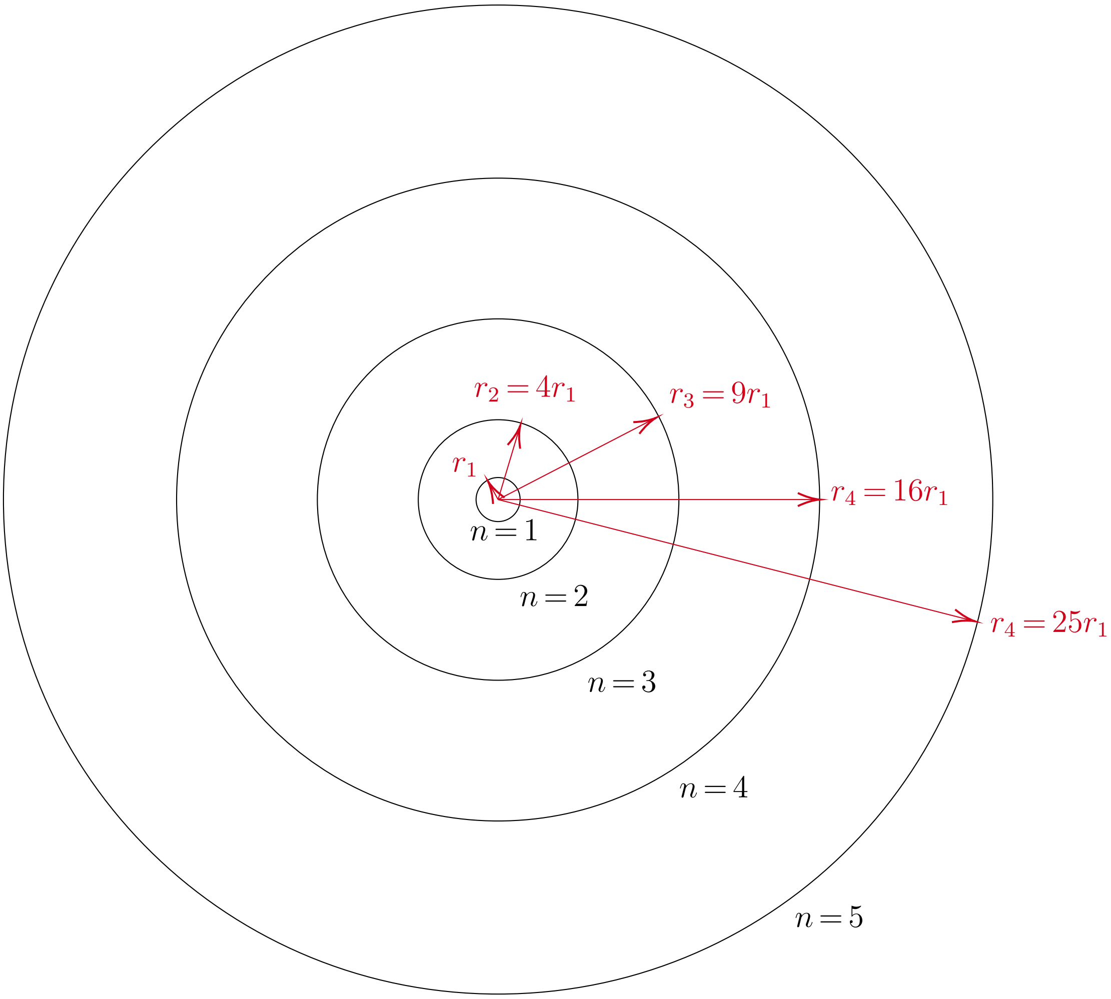


Figure 11.8: Bohr planar planetary model

The energy of the hydrogenoid-like atom without dragging is given by classical mechanics (case of a central force), the sum of kinetic and potential electrostatic energy:

$$E_{\text{tot}} = E_c + E_p = \frac{1}{2}m\omega^2 r^2 - k\frac{Ze^2}{r} \tag{11.1.35}$$

with:

$$\frac{Ze^2}{4\pi\varepsilon_0 r^2} = m\omega^2 r \tag{11.1.36}$$

we get:

$$E_{\text{tot}} = E_c + E_p = \frac{1}{2}m\omega^2 r^2 - k\frac{Ze^2}{r} = \frac{1}{2}k\frac{Ze^2}{r} - k\frac{Ze^2}{r} = -\frac{1}{2}k\frac{Ze^2}{r} \tag{11.1.37}$$

By introducing the expression of quantified radius obtained previously:

$$E_{\text{tot}} = -\frac{1}{2}k\frac{Ze^2}{r_n} = -\frac{1}{2}k\frac{Ze^2}{n^2\dfrac{\varepsilon_0 h^2}{\pi m Z e^2}} = -\frac{mZ^2e^4}{8\varepsilon_0^2 n^2 h^2} \tag{11.1.38}$$

We therefore find that the total energy of the atom is quantified and considered as negative (corresponding to stable states because it takes a supply of energy to undo the link) such that:

$$\boxed{E_{\text{tot}}=-\frac{1}{n^2}\frac{mZ^2e^4}{8\varepsilon_0^2h^2}} \tag{11.1.39}$$

Remark

This relation is often expressed in many textbooks for $Z=1$ and in eV (electron-volts) units as following:

$$E_{\text{tot}}\cong-\frac{13.606}{n^2}\,[\text{eV}] \tag{11.1.40}$$

Between two levels, the transition of an electron of the level n_2 to the level n_1 (we will specify how during the study of the photoelectric effect later) results in the emission of a frequency line given by the expression of Planck's quantization hypothesis:

$$E_{\text{tot}}^{n_2}-E_{\text{tot}}^{n_1}=h\nu_{n_2,n_1} \tag{11.1.41}$$

Thus, quantum physics explains the photon energy emission at a transition between two levels!

In fact, if we assume with Bohr that the energies of an electron on its orbit are given by the inverse of the square of the integer, the energy difference between two orbits characterized by high values of these integers approaches zero when the integers tend to infinity. We find then a semblance of continuous change for the energies exchanged by an atom with the electromagnetic field and the concept of trajectory of an electron takes then again sense!

By using the full expression of the total energy, we then find the appropriate frequency to the line emitted:

$$\nu_{n_2,n_1}=Z^2\frac{me^4}{8\varepsilon_0^2h^3}\left(\frac{1}{n_1^2}-\frac{1}{n_2^2}\right) \tag{11.1.42}$$

The wavelength emitted is thus easily deduced:

$$\boxed{\frac{1}{\lambda}=Z^2\frac{me^4}{8\varepsilon_0^2h^3}\left(\frac{1}{n_1^2}-\frac{1}{n_2^2}\right)=Z^2R_H\left(\frac{1}{n_1^2}-\frac{1}{n_2^2}\right)} \tag{11.1.43}$$

The constant R_H (denoted as R_y depending on the context) is named the "Rydberg constant".

An electron occupying an orbit n is in a "steady state" if its energy does not change. If it's orbit change, we speak then of "quantum jump" or "quantum leap".

Remark

Since the very beginning of quantum mechanics, a debate has raged about how to interpret its bizarre predictions, and at the heart and origin of that debate is the quantum jump or quantum leap. It comes from the idea that electrons in atoms jump instantaneously from one orbit or energy level to another without ever occupying the intervening space. Erwin Schrödinger himself never accepted the idea of the quantum jump, but could also never prove it wrong. That proof required precision measurements that didn't exist in Schrödinger's time. However, in the beginning of the 121st century (holocene calendar) the reality of the quantum jump has finally been tested. Schroedinger claimed that both epicycles and quantum jumps were ingenious constructs of the human mind that nevertheless were not true descriptions of nature. By the 11980s (holocene calendar), we had learn how to trap and cool a single atom with LASERs, and in 11986 (holocene calendar) almost simultaneously, three different teams observed quantum jumps in such an atom. In this way physicists were able to gain fairly direct evidence of a single quantum jump and the downward jumps when the electron decade appeared to occur at completely random times, just as Bohr had predicted. Although the jump appeared random, there was no way to tell whether it was instantaneous or whether the electron passed through some intermediate states during the jump. Fast forward another 30 plus years technology has advanced to the point that we can not only see individual quantum jumps, we can monitor their progress and even interrupt them mid jump. Now this isn't with an actual atom, but rather a sort of artificial atom made of two superconducting circuits. The scientists could actually zoom in on a quantum jump and finally figure out whether it was truly an instantaneous transition, and they found that no, it was not instantaneous after all, but rather was a continuous transition over intermediate states that took a few microseconds, and that transition appeared to be perfectly described by theory. In this case, "quantum trajectory theory" (QTT). QTT addresses the measurement problem in quantum mechanics by providing a detailed description of what happens during the so-called "collapse of the wave function". The theory suggests that "quantum jumps" are not instantaneous but happen in a coherently driven system as a smooth transition through a series of superposition states. This prediction was tested experimentally in 12019 (holocene calendar) by a team at Yale University led by Michel Devoret and Zlatko Minev, in collaboration with Carmichael and others at Yale University and the University of Auckland. But what about the randomness of the event? Well, the theorists showed how quantum states can transition very predictably through a series of superposition states, much as Schrodinger proposed. So nearly a century after Bohr and Schrodinger's started the argument, we may be on the verge of the next quantum leap to learning whether quantum jumps are, instantaneous or continuous, and perhaps whether the quantum world is built upon fundamentally random processes or is driven by a rigidly deterministic mechanic of space-time (for a naive and technical definition of "space-time" see the section of Special Relativity at page 5708).

However, a direct transition $n_2 \rightarrow n_1 (n_1 < n_2)$ is accompanied by the emission of a photon whose energy is given by the frequency as we proved it before. When an electron in an excited energy level of an atom drops down to a lower level, it emits a photon that provides the common source of light or, more generally, electromagnetic radiation, over a spectral range from X-rays to ultraviolet light. This is a quantum phenomenon that, in most interpretations of quantum mechanics, occurs spontaneously, that is, without cause.

The "ionization energy" is the energy it takes to give away an electron to infinity of its orbit. Thus for the ground state of hydrogen, it would be necessary to write $n_1 = 1$ and $n_2 = +\infty$.

The result obtained by Bohr for expressing the frequency depending on the electron energy levels is a great result (which surprised Bohr himself) because by involving only major fundamental constants he theoretically founded the spectrum law that hydrogen lines followed.

Law that Johann Jakob Balmer had discovered experimentally in 11885 according to holocene calendar (28 years before).

Balmer had noticed that the spectral lines were extremely thin. This implied that the energy is not emitted by the atoms continuously but only at certain specific frequencies. Moreover, this fine stripes explains the precision with which he could determine the Rydberg constant.

Chemists had also seen that each element had its own atomic spectrum. It was therefore clear that any atomic theory should reflect these two characteristics and this is what made brilliantly the Bohr's model using the assumptions of energy levels.

We define, moreover, the following spectrum series of the hydrogen atom:

- For the series starting from $n_1=1$ and going to $n_2=2,3,4,\ldots$ we get the result of measurements (spectrum) by Lyman in 11906 (holocene calendar) in the UV.
- For the series starting from $n_1=2$ and going to $n_2=3,4,5,6,\ldots$ we get the result of measurements (spectrum) by Balmer in 11885 (holocene calendar) in the visible spectrum (respectively red, bluegreen, violet, violet).
- For the series starting from $n_1=3$ and going to $n_2=4,5,6\ldots$ we get the result of measurements (spectrum) by Paschen in 11908 (holocene calendar) in the infra-red.
- For the series starting from $n_1=4$ and going to $n_2=5,6,7\ldots$ we get the result of measurements (spectrum) by Brackett in 11928 (holocene calendar) in the infra-red.
- For the series starting from $n_1=5$ and going to $n_2=6,7,8\ldots$ we get the result of measurements (spectrum) by Pfund in 11924 (holocene calendar) in the infra-red.

Remark

If a photon hits let's say a hydrogen atom i.e. with a single electron) with a given energy, four cases can be considered: a) it can scatter elastically, which means it retains its frequency but changes angle b) it can scatter inelastically, which means it changes frequency givin some of its energy to the entire atom c) it can be absorbed raising the energy level of an electron (in a lattice, in a molecule, in an atom) and a different photon is emitted and phases are lost.

Here is a picture representation summarizing this:

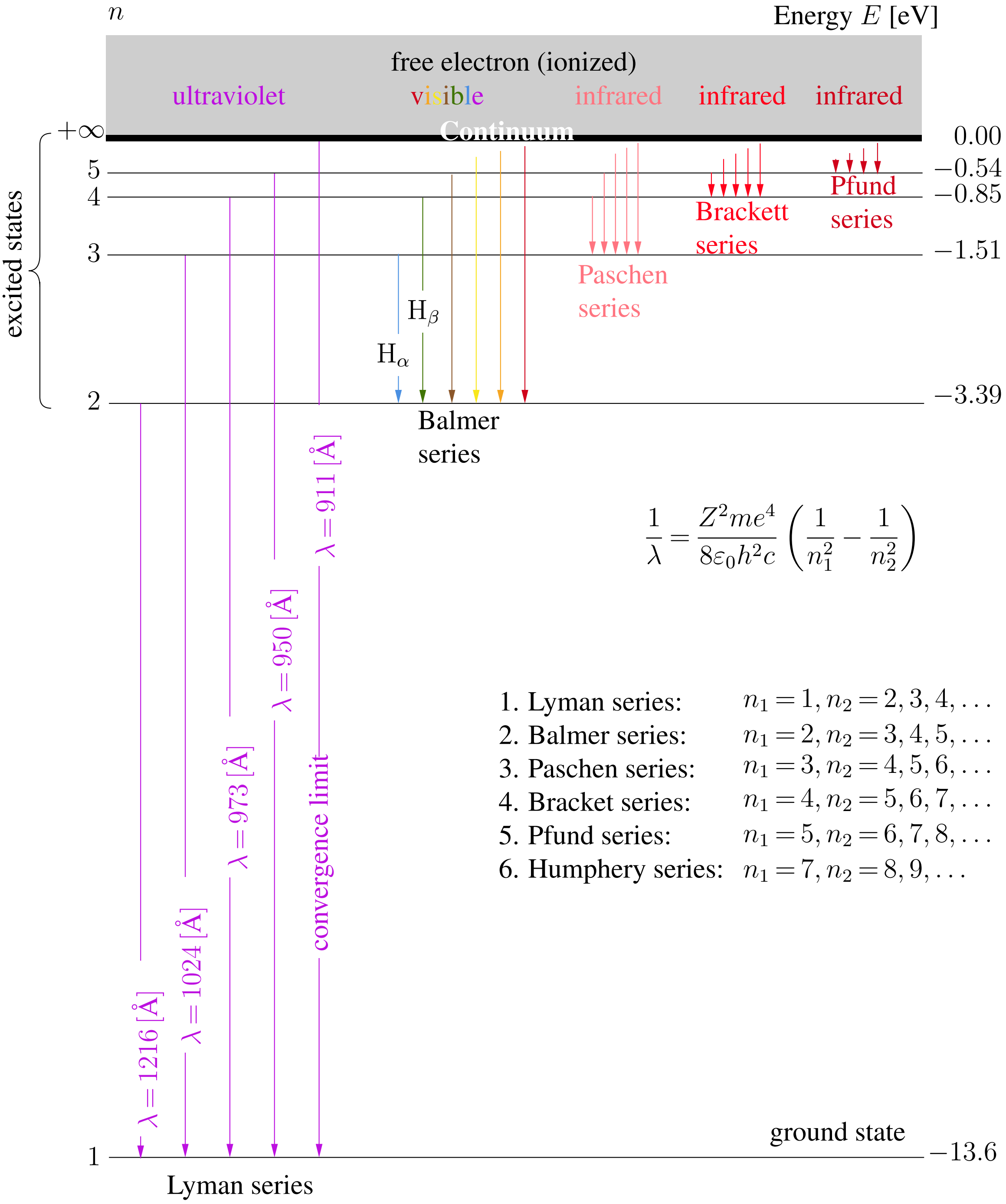

n
Energy E [eV]
free electron (ionized)
ultraviolet
visible
infrared
infrared
infrared
Continuum
$+\infty$
0.00
5
-0.54
4
-0.85
3
-1.51
2
-3.39
1
-13.6
excited states
Brackett series
Pfund series
Paschen series
H_β
H_α
Balmer series
$\lambda = 911$ [Å]
$\lambda = 950$ [Å]
$\lambda = 973$ [Å]
$\lambda = 1024$ [Å]
$\lambda = 1216$ [Å]
convergence limit
ground state
Lyman series
$\frac{1}{\lambda} = \frac{Z^2 m e^4}{8\varepsilon_0 h^2 c}\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$
1. Lyman series: $n_1 = 1, n_2 = 2, 3, 4, \ldots$
2. Balmer series: $n_1 = 2, n_2 = 3, 4, 5, \ldots$
3. Paschen series: $n_1 = 3, n_2 = 4, 5, 6, \ldots$
4. Bracket series: $n_1 = 4, n_2 = 5, 6, 7, \ldots$
5. Pfund series: $n_1 = 5, n_2 = 6, 7, 8, \ldots$
6. Humphery series: $n_1 = 7, n_2 = 8, 9, \ldots$

Figure 11.9: Some spectral series of the hydrogen atom

or another representation for the three most known series:

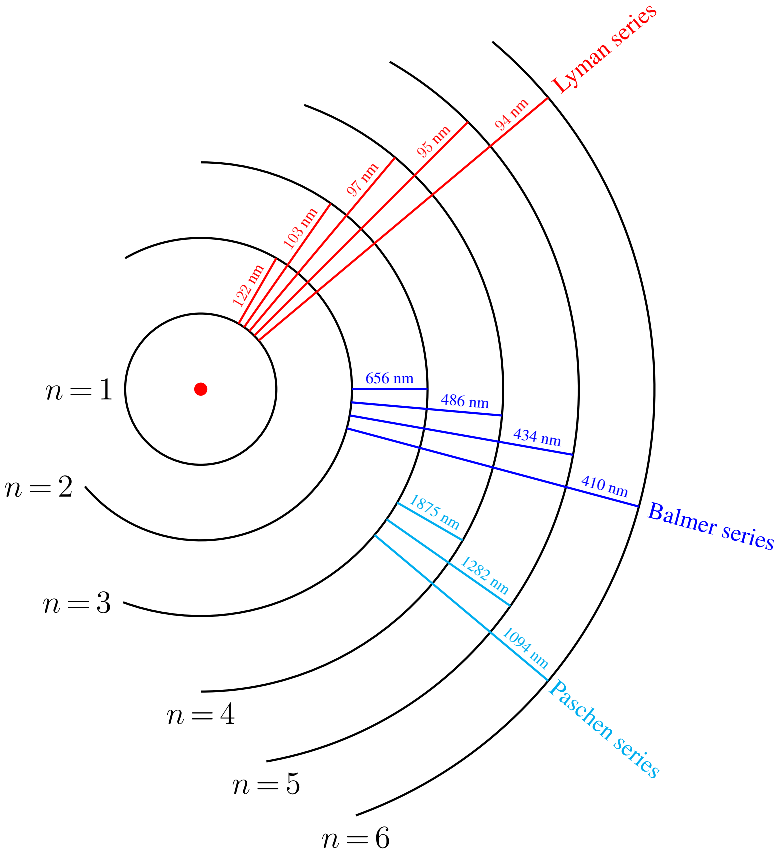

Lyman series
94 nm
95 nm
97 nm
103 nm
122 nm
$n=1$
656 nm
486 nm
434 nm
410 nm
Balmer series
$n=2$
1875 nm
1282 nm
1094 nm
$n=3$
Paschen series
$n=4$
$n=5$
$n=6$

Figure 11.10: **Three main series of the hydrogen atom (author: M. El Founani)**

It follows, from what we just seen, something very important in chemistry (to recognize easily with what element we are dealing with) and astrophysics (to analyse the elements in a cloud of gas surrounding a star or another planet turning around a star): the fact that materials (elements) all have a specific Line Spectra (emission spectrum) as shown in the figure below!

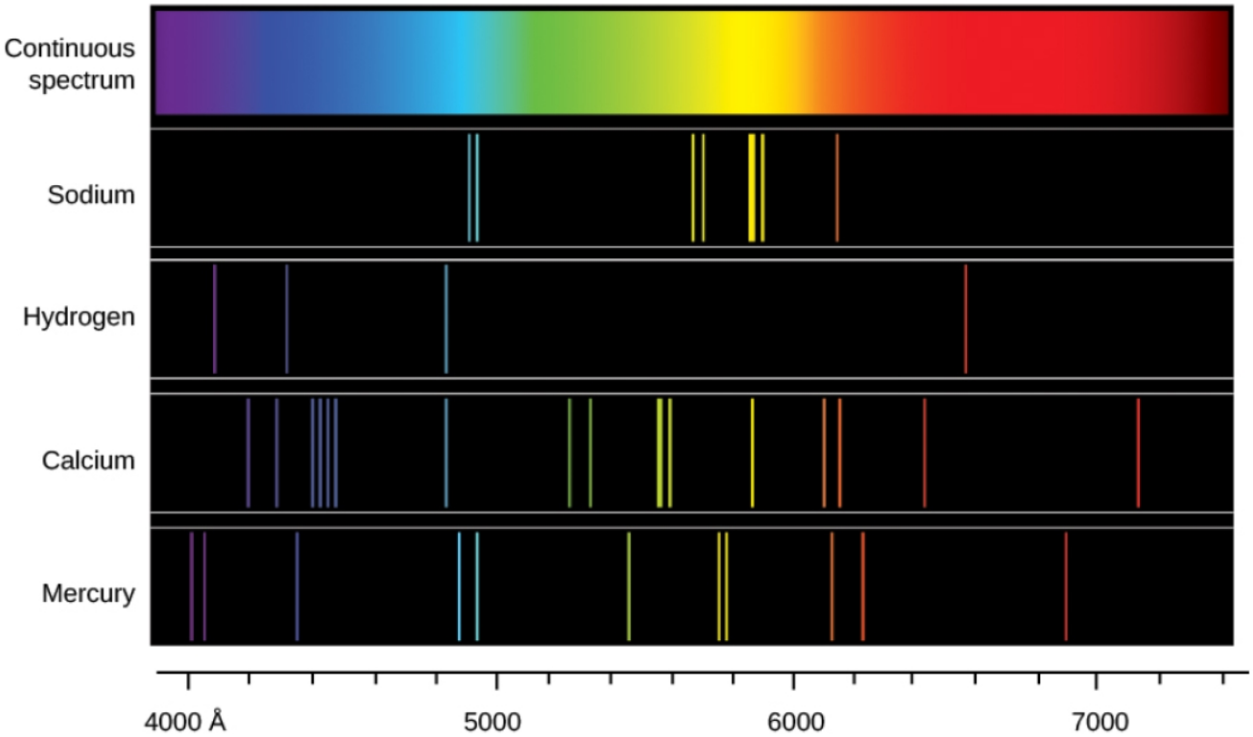
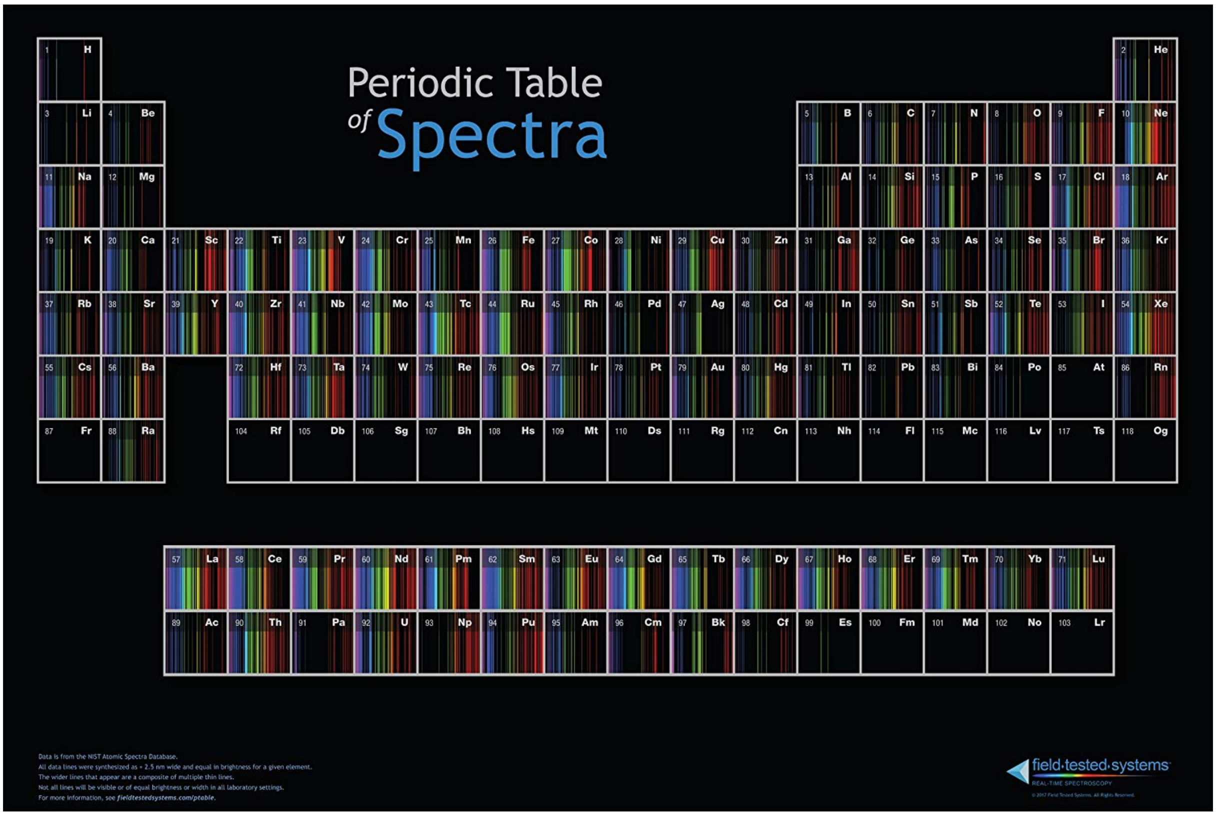


Figure 11.11: Visible line spectra of various elements (source: OpenStax)

Or a more complete table of elements line spectra (still only in the visible):



Data is from the NIST Atomic Spectra Database.
All data lines were synthesized as ~ 2.5 nm wide and equal in brightness for a given element.
The wider lines that appear are a composite of multiple thin lines.
Not all lines will be visible or of equal brightness or width in all laboratory settings.
For more information, see *fieldtestedsystems.com/ptable*.

Figure 11.12: Visible line spectra of elements (source: NIST atomic spectra database)

Keep in mind that the opposite of an emission spectrum is an absorption spectrum!

Whatever the model of atom we take (Bohr, Sommerfeld, Schrödinger) we will prove that the energy level are so that the structure of the atoms are independent of any fundamental parameters (mass of electron, proton or of fine-structure constant!). All universes with a wide range of values of these parameters will have the same chemistry. This will change only when we can no longer assume the electrons reside outside the nucleus or when the motion of protons can no longer be neglected. So it is very likely that all these constants are not fine-tuned to give chemistry and life based on it.

11.1.4.4 Hydrogen Type Atoms Model with dragging

The atomic nucleus has a mass M that we have assumed stationary for simplification. In reality, the assembly core (M) and electron (m) is rotating about a common center of mass (obviously!).

Hypothesis (assumptions):

H1. The hydrogen-like atom is considered an isolated system.

H2. The nucleus and the electron orbit each in a circular orbit around a common center: the "center of mass CM" (see section Classical Mechanics page 3213).

H3. They have the same angular velocity.

The hydrogenoid atom being an isolated system, the movement of the center of mass is either in uniform rectilinear motion or at rest. It is therefore acceptable to place there an inertial reference system.

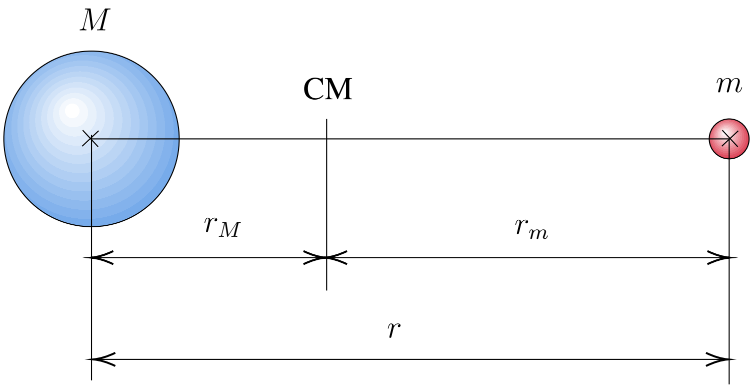


Figure 11.13: **Defining hydrogenoid atom configuration**

The definition of the center of mass in a laboratory system is given by the theorem of the center of mass (see section Classical Mechanics page 3213):

$$\vec{r}=\frac{m_1\vec{r}_1+m_2\vec{r}_2}{m_1+m_2} \tag{11.1.44}$$

The present study will be made relatively to the center of mass! Then, the above relation becomes (see section Classical Mechanics page 3213):

$$M\vec{r}_M + m\vec{r_m} = \vec{0} \tag{11.1.45}$$

From the above relation, taking the norm and absolute value, it follows that:

$$r_M = \frac{m}{M} r_m \tag{11.1.46}$$

The distance between the nucleus and the electron being constant (hypothesis) such that $r = r_M + r_m$ we write:

$$r = r_M + r_m = \frac{m}{M} r_m + r_m = \left(1 + \frac{m}{M}\right) r_m = \left(\frac{M+m}{M}\right) r_m \tag{11.1.47}$$

We conclude that trivially:

$$r_m = \frac{M}{m+M} r \quad \text{and} \quad r_M = \frac{m}{M+m} r \tag{11.1.48}$$

Applying the law of dynamics, we write that the sum of forces (electrostatic and centrifugal) on the electron (only!) is balanced such that:

$$k\frac{Ze^2}{r^2} = m\omega^2 r_m \tag{11.1.49}$$

that we can rewrite by isolating ω^2:

$$\omega^2 = \frac{kZe^2}{mr^2\left(\dfrac{M}{m+M}\right) r} = \frac{kZe^2}{r^3}\frac{1}{\dfrac{mM}{m+M}} = \frac{kZe^3}{\mu r^3} \tag{11.1.50}$$

We find again the well-knows expression of the reduced mass for a two-body system:

$$\mu = \frac{mM}{m+M} = \frac{m}{1 + \dfrac{m}{M}} \tag{11.1.51}$$

The both relations:

$$r = \frac{M+m}{M} r_m \qquad \text{and} \qquad \mu = \frac{mM}{m+M} \tag{11.1.52}$$

will also be useful to us in the sections of Quantum Chemistry and Molecular Chemistry!

Let us now determine the total energy of the atom!

The kinetic energy of the atom is the sum of the kinetic energies of the core (N) and the electron (e) such that:

$$E_c = \frac{1}{2} m v_e^2 + \frac{1}{2} M v_N^2 \tag{11.1.53}$$

As equation with the assumption that the angular velocity is identical for the nucleus and the electron:

$$E_c = \frac{\omega^2}{2}\left(m r_m^2 + M r_N^2\right) \tag{11.1.54}$$

Using the different radii determined previously:

$$\begin{aligned}E_c=&\frac{\omega^2r^2}{2}\left[m\frac{M^2}{(m+M)^2}+M\frac{m^2}{(m+M)^2}\right]=\frac{\omega^2r^2}{2}\left[\frac{mM^2+Mm^2}{(m+M)^2}\right]\\=&\frac{\omega^2r^2}{2}\left[\frac{mM}{(m+M)^2}(m+M)\right]=\mu\frac{\omega^2r^2}{2}\end{aligned}\tag{11.1.55}$$

The potential energy of the electron relatively to the center of mass is given by (see section Electrostatics page 4005):

$$E_p=-\frac{kZe^2}{r}\tag{11.1.56}$$

The total energy of the hydrogen-like atom is then:

$$E_{\text{tot}}=\frac{kZe^2}{2r}-\frac{kZe^2}{r}=-\frac{k}{2}\frac{Ze^2}{r}\tag{11.1.57}$$

Relatively to the center of mass, the total angular momentum is the sum of the angular momentum of the electron b_m and the kernel b_M (remember that the angular momentum is also often denoted by the letter L):

$$b_{\text{tot}}=b_m+b_M=mv_mr_m+Mv_mr_M=\omega(mr_m^2+Mr_m^2)\tag{11.1.58}$$

The parenthesis of the last equality has already been the subject of a calculation previously so we have:

$$b_{\text{tot}}=\mu\omega r^2\tag{11.1.59}$$

It is here that Bohr introduced its quantization condition:

$$b_{\text{tot}}=\mu\omega r^2=n\hbar\tag{11.1.60}$$

But we know the detailed expression of the square of the pulsation:

$$b_{\text{tot}}^2=\mu^2\omega^2r^4=\mu^2\left(\frac{kZe^2}{\mu r^3}\right)r^4=n^2\hbar^2\tag{11.1.61}$$

The quantified radius has therefore for expression:

$$\boxed{r_n=n^2\frac{\varepsilon_0h^2}{\pi\mu Ze^2}}\tag{11.1.62}$$

The total energy of the atom finally becomes:

$$E_{\text{tot}}^n=-\frac{k}{2}\frac{Ze^2}{\dfrac{n^2\varepsilon_0h^2}{\pi\mu Ze^2}}=-\frac{1}{n^2}\frac{\mu Z^2e^4}{8\varepsilon_0^2h^2}=-\frac{1}{n^2}\frac{mZ^2e^4}{8\varepsilon_0^2h^2}\left(\frac{1}{1+\dfrac{m}{M}}\right)\tag{11.1.63}$$

Either with a condensed notation:

$$\boxed{E_{\text{tot}}^n=-\frac{1}{n^2}hcZ^2R_H\left(\frac{1}{1+\dfrac{m}{M}}\right)}\tag{11.1.64}$$

From this last relation, we can easily determine the expression (as we have already done) for wavelengths emitted by a de-excitation of the electron from one orbit n_2 to n_1. Let us calculate with the same way as we did for the free training model, the expression of the wavelength emitted during the transition from one level to another. We then identical developments:

$$E_{\text{tot}}^{n_2}-E_{\text{tot}}^{n_1}=h\nu_{n_2,n_1} \tag{11.1.65}$$

It comes then:

$$\nu_{n_2,n_1}=\left(\frac{1}{n_1^2}-\frac{1}{n_2^2}\right)cZ^2R_H\left(\frac{1}{1+\dfrac{m}{M}}\right) \tag{11.1.66}$$

The wavelength emitted is then easily deduced:

$$\frac{1}{\lambda}=Z^2R_H\left(\frac{1}{1+\dfrac{m}{M}}\right)\left(\frac{1}{n_1^2}-\frac{1}{n_2^2}\right) \tag{11.1.67}$$

Remark

It obviously to realize that this model is more accurate than the previous one.

11.1.4.5 Neutron Assumption

Spectroscopy results are known with high precision, therefore the Rydberg's constants also (because dependent on the mass of the atomic element studied).

The two blue stripes of the Balmer series for the hydrogen atom denoted by the letter H (composed of one proton and one electron) and of deuterium which atom denoted by the letter D (isotope of hydrogen consisting in one neutron more) have a difference of wavelength of $\Delta\lambda\cong 1.32\mathring{A}$.

The wavelength belonging to the Balmer series ($n_1=2, n_2=3,4,5,\ldots$ is therefore expressed by (with the correction of the mass center as seen above):

$$\nu_n=Z^2R_H\left(\frac{1}{2^2}-\frac{1}{n^2}\right)\Rightarrow\lambda=\frac{1}{Z^2R_H\left(\dfrac{1}{2^2}-\dfrac{1}{n^2}\right)}\left(1+\frac{m}{M}\right) \tag{11.1.68}$$

This last expression written successively for hydrogen and deuterium leads to:

$$\lambda_{\text{H}}=\frac{1}{Z^2R_H\left(\dfrac{1}{2^2}-\dfrac{1}{n^2}\right)}\left(1+\frac{m}{M_{\text{H}}}\right)\quad\text{and}\quad\lambda_{\text{D}}=\frac{1}{Z^2R_H\left(\dfrac{1}{2^2}-\dfrac{1}{n^2}\right)}\left(1+\frac{m}{M_{\text{D}}}\right) \tag{11.1.69}$$

where we recall that the mass of the electron m is known to us!

What is interesting here is that these two elements have identical chemical properties (hydrogen and deuterium) but different spectrum lines. Scientists of this time wondered why and after that the Bohr hydrogen atom with dragging get available to them they were able to conclude that this difference came from the difference in the mass of the atomic nucleus.

It is still necessary to determine the difference in mass and explain its origin!

We therefore have:

$$\begin{aligned}\frac{\Delta\lambda}{\lambda_{\mathrm{H}}}=\frac{\lambda_{\mathrm{H}}-\lambda_{\mathrm{D}}}{\lambda_{\mathrm{H}}}&=\frac{\dfrac{1}{Z^2R_H\left(\dfrac{1}{2^2}-\dfrac{1}{n^2}\right)}\left[\left(1+\dfrac{m}{M_{\mathrm{H}}}\right)-\left(1+\dfrac{m}{M_{\mathrm{D}}}\right)\right]}{\dfrac{1}{Z^2R_H\left(\dfrac{1}{2^2}-\dfrac{1}{n^2}\right)}\left(1+\dfrac{m}{M_{\mathrm{H}}}\right)}\\&=\frac{\left(1+\dfrac{m}{M_{\mathrm{H}}}\right)-\left(1+\dfrac{m}{M_{\mathrm{D}}}\right)}{1+\dfrac{m}{M_{\mathrm{H}}}}=\frac{\dfrac{m}{M_{\mathrm{H}}}-\dfrac{m}{M_{\mathrm{D}}}}{1+\dfrac{m}{M_{\mathrm{H}}}}=\frac{\dfrac{M_{\mathrm{H}}}{m}}{\dfrac{M_{\mathrm{H}}}{m}}\frac{\dfrac{m}{M_{\mathrm{H}}}-\dfrac{m}{M_{\mathrm{D}}}}{1+\dfrac{m}{M_{\mathrm{H}}}}\end{aligned}\tag{11.1.70}$$

This simplifies to:

$$\frac{\Delta\lambda}{\lambda_{\mathrm{H}}}=\frac{1-\dfrac{M_{\mathrm{H}}}{M_{\mathrm{D}}}}{\dfrac{M_{\mathrm{H}}}{m}+1}\cong\frac{1-\dfrac{M_{\mathrm{H}}}{M_{\mathrm{D}}}}{1836.5+1}=\frac{1-\dfrac{M_{\mathrm{H}}}{M_{\mathrm{D}}}}{1837.5}\tag{11.1.71}$$

Therefore:

$$\frac{M_{\mathrm{H}}}{M_{\mathrm{D}}}=1-1837.5\frac{\Delta\lambda}{\lambda_{\mathrm{H}}}=1-1837.5\frac{1.32}{4861.3}\cong 0.501\tag{11.1.72}$$

Which showed the scientists of that time that the deuterium nucleus consists of two particle mass equivalent to that of the proton. Thus by logical deduction, this core must be made of a proton (which is obviously knows!) and a neutral particle!!!

This hypothesis is this of the "neutron", which was later discovered experimentally in 11932 (holcoene calendar) by Chadwick.

Today we know that:

$$m_p\cong 1.67262\cdot 10^{-27}\,[\mathrm{kg}]\quad\text{and}\quad m_n\cong 1.675\cdot 10^{-27}\,[\mathrm{kg}]\tag{11.1.73}$$

The reader must know that the experimental and theoretical discovery of the neutron is very important! Indeed, many famous scientists (including Einstein and Rutherford) before that discovery were sure that supply humanity with atom power was an utopia (impossible!) since make an external proton collide with the nucleus of an atom to break that latter require huge amount of energy because both were positively charged (especially the nucleus obviously!). But the discovery of the neutron made this plausible as it is... neutral! Hence the only point remaining was to discover an element that when the nucleus collide with a neutron break itself, emits energy, and afterwards emits a neutron again (stable nuclear reaction useful for power-plants) or emits more than one neutron in order to great a divergent reaction that would lead to an explosion (massive weapon).

Remark

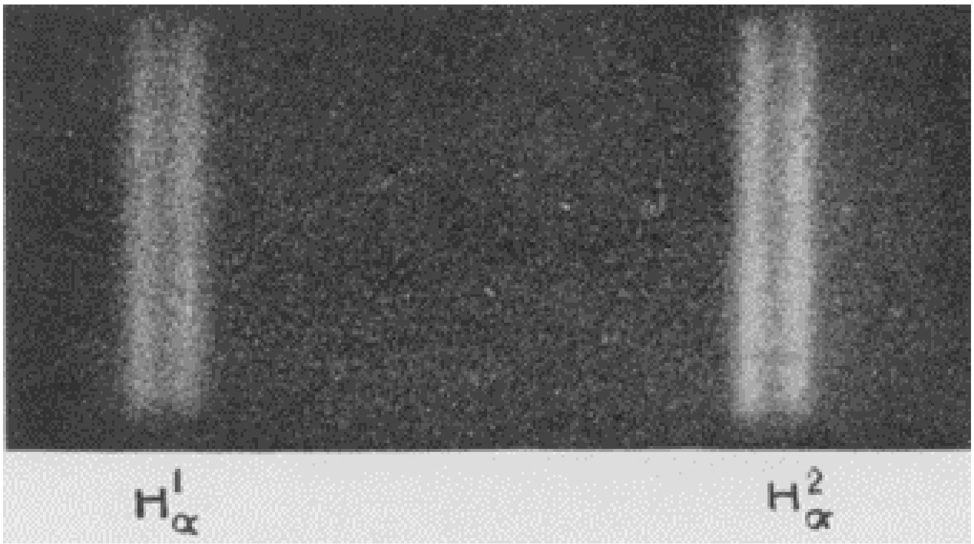
The neutron out of the nucleus will decay! With an average lifetime of about 10 minutes. More precisely, 885 seconds plus or minus 10 (even with measurement made in 12015 according to holocene calendar!). The curious part is such a huge incertitude with technologies of 12015! This is a puzzling situation and might be a harbinger of new physics for the time we write these lines.

For more information about Nuclear Physics the reader should report to the corresponding section on page 5019.

11.1.5 Wilson and Sommerfeld's Model

In developing their model, Sommerfeld and Wilson used classical dynamics to generalize Bohr's model to Keplerian orbits type (thus not only circular but elliptical in the general case!) because the Bohr's model did not explain the splitting of certain spectrum lines (without presence of any electric or magnetic field) that at the time was named "fine structure" (or "multiplet").

Around 11930 (holocene calendar), several spectroscopists using high resolution instruments found that lines in the hydrogen atom spectrum actually are not single lines but they are multiplets as shown for an isotopic mixture of hydrogen, (H_α^1) and deuterium, $(\mathrm{H}_\alpha^2\ )$, in the figure below (such fine structure also was found in spectra of one-electron ions such as He^+):



Above we have a photo of the first line in the Balmer series for atomic hydrogen and deuterium. These lines are identified as H_α^1 (hydrogen) and H_α^2 (deuterium), respectively. From H.E. White, *Introduction to Atomic Spectra* (McGraw-Hill, New York, 1934) p. 132 and G.N. Lewis and F.H. Spedding, A *Spectroscopic Search for* H^3 *in Concentrated* H^2 Phys. Rev. 43, 964 (11933).

We should recall that the H_α^1 line in the Balmer series at 656.279 [nm] was understood as resulting from a single transition of an electron from the $n=3$ energy level to the $n=2$ level. The observation of fine structure revealed that an orbital energy level diagram does not completely describe the energy levels of atoms. This fine structure also provided key evidence at the time for the existence of electron spin, which was used not only to give a qualitative explanation for the multiplets but also to furnish highly accurate calculations of the multiplet splittings.

As we have seen above, in the case of a system with two bodies solicited by a central force, the total energy of the system (we neglect gravitational potential energy...):

$$E_{\text{tot}}=\frac{1}{2}mv^2-k\frac{Ze^2}{r}=m\left(\frac{1}{2}v^2-k\frac{Ze^2}{mr}\right):=mE \tag{11.1.74}$$

To find the expression of the trajectory of the mass m, we will proceed in exactly the same manner as the one used in astronomy (see section Astronomy page 5340) to determine the Keplerian orbits.

Thus, we have proved in the section Astronomy (see page 5234) that:

$$\boxed{r(\theta)=\frac{p}{1+e\cos(\theta)}} \tag{11.1.75}$$

With:

$$p=\frac{K^2}{\mu}=\frac{\left(r^2\dfrac{\mathrm{d}\theta}{\mathrm{d}t}\right)^2}{Gm} \quad \text{and} \quad e=\sqrt{1+\frac{2EK^2}{\mu^2}}=\sqrt{1+\frac{2E_{\text{tot}}\left(r^2\dfrac{\theta}{\mathrm{d}t}\right)}{m(Gm)^2}} \tag{11.1.76}$$

We don't need to insist that in our case, it's not anymore a gravitational potential but an electric potential. Which brings us to write to our problem as following:

$$\mu=\frac{kZe^2}{m} \tag{11.1.77}$$

We still yet have to find the expression of K in a quantified form (according to the Bohr's postulate).

We fill focus first to determine the expression of the focal parameter p of the trajectory:

In our current problem, kinetic and potential energy expressed in polar coordinates give (see section Vector Calculus page 1863):

$$E_c=\frac{1}{2}m(\dot{r}+r^2\dot{\theta}^2) \quad \text{and} \quad E_p=-k\frac{Ze^2}{r} \tag{11.1.78}$$

The total energy of the atom is therefore given by:

$$E_{\text{tot}}=E_c+E_p=\frac{1}{2}m(\dot{r}+r^2\dot{\theta}^2)-k\frac{Ze^2}{r} \tag{11.1.79}$$

In an identical way of that of Bohr, Sommerfeld and Wilson applied the same form of quantification for the vector-radius and extended it to the quantification for the azimuthal angle.

Given the angular momentum:

$$b_r=\oint p_r\mathrm{d}r=n_rh \text{ and } b_\theta=\oint p_\theta\mathrm{d}\theta=n_\theta h \tag{11.1.80}$$

Momentums are obtained by derivation of the Lagrangian relatively to the generalized coordinates (see section Analytical Mechanics page 3181):

$$\begin{aligned}p_r&=\frac{\partial L}{\partial\dot{r}}=\frac{\partial}{\partial\dot{r}}\left[\frac{1}{2}m(\dot{r}+r^2\dot{\theta}^2)+k\frac{Ze^2}{r}\right]=m\dot{r}\\p_\theta&=\frac{\partial L}{\partial\dot{\theta}}=\frac{\partial}{\partial\dot{\theta}}\left[\frac{1}{2}m(\dot{r}+r^2\dot{\theta}^2)+k\frac{Ze^2}{r}\right]=mr^2\dot{\theta}\end{aligned} \tag{11.1.81}$$

The quantification on the angle is immediate, since p_θ is a constant of the motion. Indeed, the Lagrangian L being independent of θ (but no of $\dot{\theta}$), the invariance of the angular momentum results in the following Lagrange equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial\dot{\theta}}-0=0\Leftrightarrow\dot{p}_\theta=0\Leftrightarrow p_\theta=c^{te} \tag{11.1.82}$$

This gives us:

$$\oint p_\theta \mathrm{d}\theta = p_\theta \oint \mathrm{d}\theta = p_\theta 2\pi = n_\theta h \Rightarrow p_\theta = n_\theta \hbar \tag{11.1.83}$$

with $n_\theta \in \mathbb{N}^*$ being the "azimuthal quantum number" (i.e. the second quantum number), to remember that it is related to the quantization of the polar angle.

From this last relation we also get:

$$p_\theta = n_\theta \hbar = mr^2\dot{\theta} \Leftrightarrow \frac{\mathrm{d}\theta}{\mathrm{d}t} = \frac{n_\theta \hbar}{mr^2} \tag{11.1.84}$$

Let us come back to:

$$p = \frac{K^2}{\mu} = \frac{\left(r^2\dfrac{\mathrm{d}\theta}{\mathrm{d}t}\right)^2}{\dfrac{kZe^2}{m}} \tag{11.1.85}$$

That give us:

$$\boxed{p = \frac{\left(\dfrac{n_\theta \hbar}{m}\right)^2}{\dfrac{kZe^2}{m}} = n_\theta^2 \frac{\hbar^2}{mkZe^2}} \tag{11.1.86}$$

Let us focus now to determine the eccentricity e of the path (notation not to be confused with the electrical charge if possible!).

This give us:

$$e = \sqrt{1 + \frac{2}{\mu^2}EK^2} = \sqrt{1 + \frac{2}{\left(\dfrac{k^2Z^2e^4}{m^2}\right)} \frac{E_{\text{tot}}}{m} \frac{n_\theta^2 \hbar^2}{m^2}} = \sqrt{1 + \frac{2E_{\text{tot}} n_\theta^2 \hbar^2}{mk^2Z^2e^4}} \tag{11.1.87}$$

To determine the quantization of the angular momentum with respect to the radial variable, we will use a substitution:

$$\oint p_r \mathrm{d}r = \oint m\dot{r}\mathrm{d}r = n_r h \tag{11.1.88}$$

By simply denoting r' the derivative $\mathrm{d}r/\mathrm{d}\theta$, the integral will be written:

$$\oint m\dot{r}\mathrm{d}r = \int\limits_0^{2\pi} m{r'}_2\dot{\theta}\mathrm{d}\theta = p_\theta \int\limits_0^{2\pi} \frac{r'^2}{r^2}\mathrm{d}\theta \tag{11.1.89}$$

where we used $\dot{\theta} = p_\theta/(mr^2)$ as we have already prove it.

Reporting it:

$$r(\theta) = \frac{p}{1 + e\cos(\theta)} \tag{11.1.90}$$

in the integral of the radial momentum, we get (easy to obtain but you can ask for the details if needed as always):

$$b_r = n_r h = 2\pi\left(\frac{1}{\sqrt{1-e^2}} - 1\right)p_\theta \tag{11.1.91}$$

from which we deduce given $p_\theta = n_\theta \hbar$ that:

$$b_r = n_r h = 2\pi\left(\frac{1}{\sqrt{1-e^2}} - 1\right)n_\theta \hbar \tag{11.1.92}$$

which brings us to:

$$e = \sqrt{1 - \frac{1}{\left(\dfrac{n_\theta + n_r}{n_\theta}\right)^2}} \tag{11.1.93}$$

and therefore:

$$1 - \frac{1}{\left(\dfrac{n_\theta + n_r}{n_\theta}\right)^2} = e^2 = 1 + \frac{2E_{\text{tot}} n_\theta^2 \hbar^2}{mk^2 Z^2 e^4} \tag{11.1.94}$$

After some elementary algebraic simplifications we finally get:

$$\boxed{E_{\text{tot}}^n = \frac{-mZ^2 e^4}{8\varepsilon_0 h^2 (n_\theta + n_r)^2}} \tag{11.1.95}$$

where r, also named "radial quantum number" can be null! Because this is the case when $\dot{r} = 0$, that is to say, if the trajectory is a circle (special case of Bohr model!!!).

We introduce then an integer n so named "principal quantum number" as:

$$\boxed{n = n_r + n_\theta} \tag{11.1.96}$$

with $n \in \mathbb{N}^*$.

Sommerfeld and Wilson the proved that the orbital Bohr model must be determined by these two new quantum numbers:

✍**Example:**

For $n = 2$ we have two possible sub-orbital:

$$n = 2 = n_r + n_\theta \begin{cases} 2 = 1 + 1 \\ 2 = 2 + 0 \end{cases} \tag{11.1.97}$$

The value $n_\theta = 0$ is impossible by definition because it would mean that the minor axis is zero (ellipse degenerated into a line) and the electron can not pass through the nucleus (in the classical model at least...). So the smallest possible integer value for n_θ is 1.

There is therefore n orbits giving the same spectral term. In other words, there are n times the same quantization energy. We also say that the energy level (total) E_n is "n times degenerated".

The idea of Sommerfeld was to account the richness of the observed spectra. From this point of view, the results are disappointing: the quantification of all degrees of freedom clearly shows more states (now we need two quantum numbers to completely specify the state, while the Bohr model only considers one) but the additional degree only introduces an energy degeneration (sad!).

To summarize this model, so there is exactly the same number of energy levels and therefore the same number of transitions of possible energy states than the Bohr model. At the spectral point of view, the Sommerfeld-Wilson theory does nothing more than that of Bohr except that the orbits are elliptical and therefore does not explain the extent of the observed spectra.

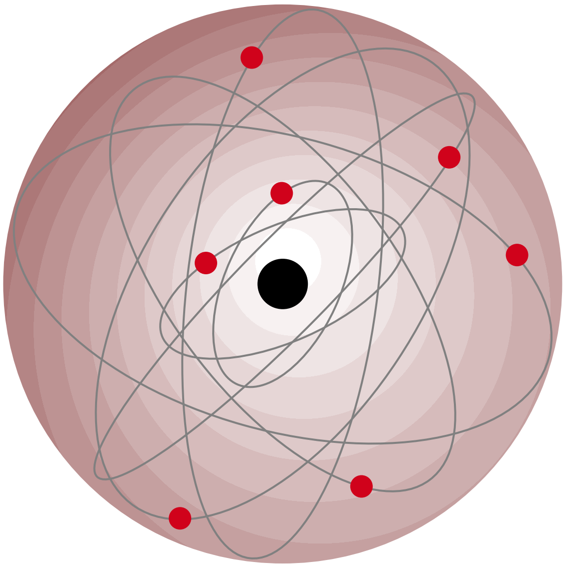


Figure 11.14: The eccentric... Sommerfeld approach...

In fact, the idea is going to be from now to take again the same model but by adding relativistic corrections. The work will necessarily be longer but oh so successful!

11.1.6 Relativistic Sommerfeld Model

However, the Sommerfeld-Wilson's model can be considered incomplete if we do not take into account changes in parameters engendered by the results of the theory of Special Relativity (see section Special Relativity page 5627) as did Sommerfeld a few years after (see [506]).

By assuming that the relativity of time would cause the mass of an electron moving in an elliptical orbit to increase as it approaches perihelion (or perinucleon) where the tangential velocity is maximum and decrease as it recedes from perihelion, Sommerfeld introduced a perturbation into the equation of motion that produced a precession of the perihelion that accounted for the fine structure in the spectral lines. The fine structure in the spectral lines was first identified in 11887 (holocene calendar) by Michelson and Morley yet it went unexplained by the Bohr model. Sommerfeld's correction to the hydrogen atom was based on an analogy with Einstein's relativistic correction to Newton's model of the solar system. The anomalous precession of the perihelion of Mercury amounted to a fine structure of the planetary orbits that could not be explained by Newton's law of gravitation. Einstein's relativistic correction accounted for the fine structure of the solar system; Sommerfeld's relativistic correction accounted for the fine structure of the spectral lines of hydrogen!

Remark

The section of Quantum Relativistic Quantum Physics is reserved solely to the study of probabilistic quantum physics based on the relativistic version of the Schrödinger equation (so strictly speaking the Relativistic Quantum Physics section should be named "Relativistic Wave Quantum Physics"). This is why it seemed to us more appropriate to put a corpuscular model, deterministic and relativistic (opposite of a wave model, probabilistic and relativistic) such as we will now study in the section of Corpuscular Quantum Physics.

Indeed, as we have proved in the development of the Bohr model, the kinetic energy of the electron is given by:

$$E_c^n=\frac{1}{2}mv^2=\frac{Z^2e^4m}{8\varepsilon_0^2n^2h^2} \tag{11.1.98}$$

That give us:

$$v=\frac{Ze^2}{2\varepsilon_0 nh} \tag{11.1.99}$$

For hydrogen and the level $n=1$, we find $v=2.19\cdot 10^{19}\;[\text{ms}^{-1}]$ and as Michelson-Morley factor (see section Special Relativity page 5636):

$$\frac{1}{\sqrt{1-\beta^2}}=\frac{1}{\sqrt{1-\left(\dfrac{v}{c}\right)^2}}\cong 1.00002667 \tag{11.1.100}$$

What the reader can quickly check with the then English version of Microsoft® Excel®:

```
=1/ROOT(1-(1*(1.60217656E-19)^2/(2*8.854187E-12*1*6.62068E-34))^2/(299792458)^2)
```

For sure the change is small but spectrometry values were so accurate that it was necessary to introduce Special Relativity to accommodate these tiny variations and thus validate the theory by experiment.

Remark

As we can easily see it, the relation shows that the particle is more away from the nucleus (n large) more its speed is low. This result has been confirmed experimentally by replacing the electron artificially by a muon and scientists have found that the lifetime of the latter slightly increased with the value of n.

Let us determine in the order of things, the expression of quantification conditions with the relativistic factors. Before we begin, it is important to understand that we consider the nucleus as fixed and as a repository of our system. Thus, compared to the referential, the mass of the electron undergoes a relativistic variation but not the electric potential (we should take into account the variation of the latter - as it is energy - if and only if the repository was the electron itself).

In relativistic dynamics (see section Special Relativity page 5679), we proved that the kinetic energy (in the form of Lagrangian notation Lagrangian with "T" instead of E_c) is expressed in the form:

$$T=mc^2-m_0c^2=mc^2\left(1-\sqrt{1-\beta^2}\right) \tag{11.1.101}$$

The potential energy (in Lagrangian notation with "V" instead of E_p) not undergoing relativistic variation, we always have:

$$V=-k\frac{Ze^2}{r} \tag{11.1.102}$$

The Lagrangian is therefore:

$$L=T-V=mc^2\left(1-\sqrt{1-\beta^2}\right)+k\frac{Ze^2}{r} \tag{11.1.103}$$

By working in polar coordinates, where the velocity has for expression (see section Vector Calculus page 1863):

$$v^2=\dot{r}^2+r^2\dot{\theta}^2 \tag{11.1.104}$$

Therefore:

$$\beta=\left(\frac{v}{c}\right)^2=\frac{\dot{r}^2+r^2\dot{\theta}^2}{c^2} \tag{11.1.105}$$

The Sommerfeld quantization conditions being:

$$\begin{aligned}b_r&=\oint p_r\mathrm{d}r=n_rh\\ b_\theta&=\oint p_\theta\mathrm{d}\theta=n_\theta h\end{aligned} \tag{11.1.106}$$

Now we must seek relativistic expressions for b_r and b_θ.

Let us begin p_r:

$$p_r=\frac{\partial L}{\partial\dot{r}}=\frac{\partial}{\partial\dot{r}}\left[mc^2(1-\sqrt{1-\beta^2})+k\frac{Ze^2}{r}\right] \tag{11.1.107}$$

with:

$$\beta^2=\frac{1}{c^2}\left(\dot{r}^2+r^2\dot{\theta}^2\right) \tag{11.1.108}$$

Therefore:

$$\begin{aligned}m_ec^2\frac{\partial}{\partial\dot{r}}\left(mc^2(1-\sqrt{1-\beta^2})\right)^{1/2}&=-m_ec^2\left[\frac{1}{2}(1-\beta^2)^{-1/2}\right]\frac{\partial}{\partial\dot{r}}(1-\beta^2)\\ \frac{\partial}{\partial\dot{r}}(1-\beta^2)=-\frac{\partial}{\partial\dot{r}}\beta^2=-\frac{1}{c^2}\frac{\partial}{\partial\dot{r}}v^2&=-\frac{1}{c^2}\frac{\partial}{\partial\dot{r}}\left(\dot{r}^2+r^2\dot{\theta}^2\right)=-\frac{2\dot{r}}{c^2}\end{aligned} \tag{11.1.109}$$

This gives:

$$mc^2\left[\frac{1}{2}(1-\beta^2)^{-1/2}\right]\left(-\frac{2\dot{r}}{c^2}\right)=\frac{mc^22\dot{r}}{2(1-\beta^2)^{1/2}c^2}=\frac{m\dot{r}}{(1-\beta^2)^{1/2}} \tag{11.1.110}$$

As:

$$\frac{\partial}{\partial\dot{r}}k\frac{Ze^2}{r}=0 \tag{11.1.111}$$

We finally have:

$$p_r=\frac{m\dot{r}}{(1-\beta^2)^{1/2}} \tag{11.1.112}$$

The first quantification condition gives therefore:

$$\boxed{b_r=\oint p_r\mathrm{d}r=\oint\frac{m\dot{r}}{(1-\beta^2)^{1/2}}\mathrm{d}r=n_rh} \tag{11.1.113}$$

For p_θ:

$$p_\theta=\frac{\partial L}{\partial\dot{\theta}}=\frac{\partial}{\partial\dot{\theta}}\left[mc^2(1-\sqrt{1-\beta^2})+k\frac{Ze^2}{r}\right] \tag{11.1.114}$$

always with:

$$\beta^2=\frac{1}{c^2}(\dot{r}^2+r^2\dot{\theta}^2) \tag{11.1.115}$$

Therefore:

$$\begin{gathered}mc^2\frac{\partial}{\partial\dot{\theta}}(1-\sqrt{1-\beta^2})=-mc^2\left[\frac{1}{2}(1-\beta^2)^{-1/2}\right]\frac{\partial}{\partial\dot{\theta}}(1-\beta^2)\\ \frac{\partial}{\partial\dot{\theta}}(1-\beta^2)=-\frac{\partial}{\partial\dot{\theta}}\beta^2=-\frac{1}{c^2}\frac{\partial}{\partial\dot{\theta}}(\dot{r}^2+r^2\dot{\theta}^2)=-\frac{2r^2\dot{\theta}}{c^2}\end{gathered} \tag{11.1.116}$$

Which gives:

$$-mc^2\left[\frac{1}{2}(1-\beta^2)^{-1/2}\left(-\frac{2r^2\dot{\theta}}{c^2}\right)\right]=\frac{mr^2\dot{\theta}}{(1-\beta)^{1/2}} \tag{11.1.117}$$

As:

$$\frac{\partial}{\partial\dot{\theta}}k\frac{Ze^2}{r}=0 \tag{11.1.118}$$

we finally have:

$$p_\theta=\frac{mr^2\dot{\theta}}{(1-\beta^2)^{1/2}} \tag{11.1.119}$$

The second quantification condition is therefore written:

$$\boxed{b_\theta=\oint p_\theta\mathrm{d}\theta=\oint\frac{mr^2\dot{\theta}}{(1-\beta^2)^{1/2}}\mathrm{d}\theta=n_\theta h} \tag{11.1.120}$$

In summary, the conditions of quantification for the relativistic atom of Sommerfeld are:

$$b_r=\oint\frac{m\dot{r}}{(1-\beta^2)^{1/2}}\mathrm{d}r=n_rh \quad \text{and} \quad b_\theta=\oint\frac{mr^2\dot{\theta}}{(1-\beta^2)^{1/2}}\mathrm{d}\theta=n_\theta h \tag{11.1.121}$$

We could, seeing the bot above results, conclude a little bit too quickly by thinking it would suffice to multiply the two conditions of quantification by the Michelson-Morley factor relative to the relativistic mass transformation. But, such a shortcut is completely wrong and all but rigorous! Indeed, if you apply this reasoning, it would be sufficient to take the expression of the total energy of the non-relativistic Sommerfeld-Wilson model and introduce everywhere where the mass appears the Michelson-Morley factor. However the result we get following this way has absolutely nothing in common with the result that we will get further below. So we must always be careful and work as a mathematician: without leapfrogging!

The total relativistic energy of the atom (sum of kinetic energy, mass energy and the potential energy of the electric field for the whole atom) is given by:

$$E_{\text{tot}}=\frac{mc^2}{(1-\beta^2)^{1/2}}-k\frac{Ze^2}{r} \tag{11.1.122}$$

Indeed, using the notation of the section of Special Relativity:

$$E_{\text{tot}}=E_c+E_{\text{m}}+E_p=\underbrace{E_c+m_0c^2}_{=mc^2}+\left(-k\frac{Ze^2}{r}\right)=mc^2+\left(-k\frac{Ze^2}{r}\right)=\frac{m_0c^2}{\sqrt{1-\beta^2}}+\left(-k\frac{Ze^2}{r}\right) \tag{11.1.123}$$

Now, as part of the study of particle quantum physics, it is (unfortunately) usage to write the rest mass with the symbol of the relativistic mass as:

$$m_0=m \tag{11.1.124}$$

Therefore:

$$E_{\text{tot}}=\frac{mc^2}{\sqrt{1-\beta^2}}+\left(-k\frac{Ze^2}{r}\right) \tag{11.1.125}$$

We must express this total energy in function of the quantization conditions. There's a long mathematical work to do but necessary to achieve the result of our study.

Given the calculation of the expression:

$$\frac{p_\theta^2}{r^2}+p_r^2 \tag{11.1.126}$$

with:

$$p_\theta=\frac{mr^2\dot{\theta}}{\sqrt{1-\beta^2}} \quad \text{and} \quad p_r=\frac{m\dot{r}}{\sqrt{1-\beta^2}} \tag{11.1.127}$$

Squaring:

$$p_\theta=\frac{mr^2\dot{\theta}}{\sqrt{1-\beta^2}}\rightarrow\frac{p_\theta^2}{r^2}=\frac{m^2r^2\dot{\theta}^2}{1-\beta^2} \quad \text{and} \quad p_r^2=\frac{m^2\dot{r}^2}{1-\beta^2} \tag{11.1.128}$$

Therefore:

$$\frac{p_\theta^2}{r^2}+p_r^2=\frac{m^2r^2\dot{\theta}^2}{1-\beta^2}+\frac{m^2\dot{r}^2}{1-\beta^2}=\frac{m^2(\dot{r}^2+r^2\dot{\theta}^2)}{1-\beta^2}=\frac{m^2v^2}{1-\beta^2} \tag{11.1.129}$$

We add of both sides of the equality m^2c^2 (in the idea to include the mass energy as you'll see some lines below), which gives:

$$\begin{aligned}m^2c^2+\frac{p_\theta^2}{r^2}+p_r^2=m^2c^2+\frac{m^2v^2}{1-\beta^2}&=m^2c^2+\frac{m^2\dfrac{v^2}{c^2}c^2}{1-\beta^2}\\&=\frac{m^2c^2(1-\beta^2)+m^2\beta^2c^2}{1-\beta^2}=\frac{m^2c^2}{1-\beta^2}\end{aligned} \tag{11.1.130}$$

By multiplying both sides by c^2 we get:

$$c^2\left(m^2c^2+\frac{p_\theta^2}{r^2}+p_r^2\right)=\frac{m^2c^4}{1-\beta^2} \tag{11.1.131}$$

Extracting the square root:

$$\sqrt{c^2\left(m^2c^2+\frac{p_\theta^2}{r^2}+p_r^2\right)}=\frac{mc^2}{\sqrt{1-\beta^2}} \tag{11.1.132}$$

If we introduce this last relation in the expression of the total energy we get:

$$\boxed{E_{\text{tot}}=c\sqrt{m^2c^2+\frac{p_\theta^2}{r^2}+p_r^2}-k\frac{Ze^2}{r}} \tag{11.1.133}$$

Now it remains to us to determine the expressions of $N-r$ and n_θ according to p_r and p_θ.

The integral of quantification of the azimuthal angle is immediate:

$$b_\theta=\oint\frac{mr^2\dot{\theta}}{(1-\beta^2)^{1/2}}\mathrm{d}\theta=\int\limits_0^{2\pi}\frac{mr^2\dot{\theta}}{(1-\beta^2)^{1/2}}\mathrm{d}\theta=\frac{mr^2\dot{\theta}}{(1-\beta^2)^{1/2}}\theta\bigg|_0^{2\pi}=\frac{2\pi mr^2\dot{\theta}}{(1-\beta^2)^{1/2}}=n_\theta h \tag{11.1.134}$$

Therefore:

$$\boxed{\frac{mr^2\dot{\theta}}{(1-\beta^2)^{1/2}}=p_\theta=n_\theta\hbar} \tag{11.1.135}$$

The integral of quantification of the radius-vector requires a more significant development:

$$b_r=\oint\frac{m\dot{r}}{(1-\beta^2)^{1/2}}\mathrm{d}r=n_rh \tag{11.1.136}$$

Then come long and joyful mathematical developments...

Reusing the expression of the total energy:

$$E_{\text{tot}}=c\sqrt{m^2c^2+\frac{p_\theta^2}{r^2}+p_r^2}-k\frac{Ze^2}{r} \tag{11.1.137}$$

We get:

$$\left(E_{\text{tot}}+k\frac{Ze^2}{r}\right)=c\sqrt{m^2c^2+\frac{p_\theta^2}{r^2}+p_r^2} \tag{11.1.138}$$

By squaring and by making a few changes:

$$\left(E_{\text{tot}}+k\frac{Ze^2}{r}\right)^2=c^2\left(m^2c^2+\frac{p_\theta^2}{r^2}+p_r^2\right) \tag{11.1.139}$$

$$\frac{1}{c^2}\left(E_{\text{tot}}+k\frac{Ze^2}{r}\right)^2=m^2c^2+p_r^2+\frac{p_\theta^2}{r^2} \tag{11.1.140}$$

$$p_r^2=\frac{1}{c^2}\left(E_{\text{tot}}+k\frac{Ze^2}{r}\right)^2-m^2c^2-\frac{p_\theta^2}{r^2} \tag{11.1.141}$$

Therefore:

$$\begin{aligned}p_r^2&=\frac{1}{c^2}\left(E_{\text{tot}}^2+\frac{k^2Z^2e^4}{r^2}+2E_{\text{tot}}\frac{kZe^2}{r}\right)-m^2c^2-\frac{p_\theta^2}{r^2}\\&=\frac{E_{\text{tot}}^2}{c^2}+\frac{k^2Z^2e^4}{c^2r^2}+2\frac{E_{\text{tot}}}{c^2}\frac{kZe^2}{r}-m^2c^2-\frac{p_\theta^2}{r^2}\\&=\left(\frac{E_{\text{tot}}^2}{c^2}-m^2c^2\right)+\frac{k^2Z^2e^4}{c^2r^2}+2\frac{E_{\text{tot}}}{c^2}\frac{kZe^2}{r}-\frac{p_\theta^2}{r^2}\end{aligned} \tag{11.1.142}$$

Working on the bracketed term, we will put it as being equal to A as:

$$A=\frac{E_{\text{tot}}^2}{c^2}-m^2c^2 \tag{11.1.143}$$

By adding and subtracting $2E_{\text{tot}}m$ and decomposing the term $-m^2c^2$ in $m^2c^2-2m^2c^2$ and then by regrouping:

$$\begin{aligned}A&=\left[\frac{E_{\text{tot}}^2}{c^2}+2E_{\text{tot}}m-2E_{\text{tot}}m\right]+[m^2c^2-2m^2c^2]\\&=\left[\frac{E_{\text{tot}}^2}{c^2}-2E_{\text{tot}}m+m^2c^2\right]+[2E_{\text{tot}}m-2m^2c^2]\\&=\frac{2E_{\text{tot}}^2-2E_{\text{tot}}mc^2+m^2c^4}{c^2}+[2E_{\text{tot}}m-2m^2c^2]=\frac{(2E_{\text{tot}}-mc^2)^2}{c^2}+2m(2E_{\text{tot}}-mc^2)\end{aligned} \tag{11.1.144}$$

We will put for the simplification of the calculations (to reduce the number of terms to handle):

$$E'=E_{\text{tot}}-mc^2 \tag{11.1.145}$$

Therefore we get:

$$A=\frac{E'^2}{c^2}=E_{\text{tot}}+2mE' \tag{11.1.146}$$

By taking m^2c^2 in evidence, we have:

$$A=m^2c^2\left(\frac{E'^2}{m^2c^4}+2\frac{E'}{mc^2}\right) \tag{11.1.147}$$

By adding and subtracting 1 in the parenthesis:

$$A=m^2c^2\left[\left(\frac{E'^2}{m^2c^4}+2\frac{E'}{mc^2}+1\right)-1\right]=m^2c^2\left[\left(1+\frac{E'}{mc^2}\right)^2-1\right] \tag{11.1.148}$$

$$\frac{k^2Z^2e^4}{c^2r^2}+2\frac{E_{\text{tot}}}{c^2}\frac{kZe^2}{r}-\frac{p_\theta^2}{r^2} \tag{11.1.149}$$

As $E_{\text{tot}}=E'+mc^2$ we have:

$$2\frac{kZe^2(E'+mc^2)}{rc^2}+\frac{k^2Z^2e^4}{c^2r^2}-\frac{p_\theta^2}{r^2}=\frac{2}{r}\frac{kZe^2(E'+mc^2)}{c^2}+\frac{k^2Z^2e^4}{c^2r^2}-\frac{p_\theta^2}{r^2} \tag{11.1.150}$$

By putting:

$$B=k\frac{Ze^2E'}{c^2}+kmZe^2 \tag{11.1.151}$$

By putting also:

$$C=\frac{k^2Z^2e^4}{c^2}-n_\theta^2\hbar^2 \tag{11.1.152}$$

as $p_\theta=n_\theta\hbar$

Sommerfeld then introduced what he named the "fine-structure constant α" defined by the relation:

$$\boxed{\alpha=\frac{e^2}{\hbar c}\cong\frac{1}{137.0359911}=7.29735\cdot 10^{-3}} \tag{11.1.153}$$

We will meet this constant again in the section of Particle Physics.

Remark

The fine-structure constant is one of the most important physical constants. First, because it is dimensionless, and secondly because it is so far the best known (in terms of accuracy) of all constants and thirdly, because it depends only on terms that seem to be fundamental constants of the atomic world. So physicists and astrophysicists are looking to see if the value of this constant varies over time, which would imply immediately that at least one of the implicit constant is not timeless.

Caution!!!! In quantum electrodynamics the renormalization group dictates how the strength of the electromagnetic interaction grows logarithmically as the relevant energy scale increases. The value of the fine-structure constant α is linked to the observed value of this coupling associated with the energy scale of the electron mass: the electron's mass gives a lower bound for this energy scale, because it (and the positron) is the lightest charged object whose quantum loops can contribute to the running. Therefore, $1/137.03600$ is the "asymptotic value of the fine-structure constant at zero energy". At higher energies, such as the scale of the Z boson, about 90 $[\text{GeV}]$, one instead measures an effective $\alpha\approx 1/127$ (see [194] for more information).

Given the fine-structure constant, we write:

$$C=k^2Z^2\left(\frac{e^4}{c^2\hbar^2}\right)-n_\theta^2\hbar^2=k^2Z^2\alpha^2\hbar^2-n_\theta^2\hbar^2=-n_\theta^2\hbar^2\left(1-\frac{k^2Z^2\alpha^2}{n_\theta^2}\right) \quad (11.1.154)$$

To summarize:

$$p_r=\pm\sqrt{A+\frac{2B}{r}+\frac{C}{r^2}} \quad (11.1.155)$$

With:

$$\begin{aligned}A&=\frac{E'^2}{c^2}+2mE'=m^2c^2\left[\left(1+\frac{E'}{mc^2}\right)^2-1\right]\\ B&=\frac{kZe^2E'}{c^2}+kmZe^2\\ C&=-n_\theta^2\hbar^2\left(1-\frac{k^2Z^2\alpha^2}{n_\theta^2}\right)\end{aligned} \quad (11.1.156)$$

So we arrive at the following integral:

$$\oint p_r\mathrm{d}r=\oint\frac{m\dot{r}}{\sqrt{1-\beta^2}}\dot{r}\mathrm{d}r=\oint\sqrt{A+\frac{2B}{r}+\frac{C}{r^2}}\mathrm{d}r=n_rh \quad (11.1.157)$$

The residue theorem (see section Complex Analysis page 2440) applied to the preceding integral gives for expression:

$$\oint\sqrt{A+\frac{2B}{r}+\frac{C}{r^2}}\mathrm{d}r=2\pi\mathrm{i}\sum r_i \quad (11.1.158)$$

We see that there is a trivially a pole at origin $r=0$.

We will calculate the residue at this point going to the limit $r=0$. We put for this:

$$f(r)=\sqrt{A+\frac{2B}{r}+\frac{C}{r^2}} \quad (11.1.159)$$

Passing to the limit built on the basis of the residue theorem:

$$\lim_{r\to0}rf(r)=\lim_{r\to0}\sqrt{Ar^2+2Br+C}=\sqrt{C} \quad (11.1.160)$$

The residue corresponding to the pole $r=0$ is then:

$$r_1=\sqrt{C} \quad (11.1.161)$$

We also see that there is a second residue at infinity $r=+\infty$ and to calculate it, we make again a change of variable. We put (according to the method that we saw in the section of Complex Analysis):

$$z=\frac{1}{r}\Rightarrow\mathrm{d}z=-\frac{\mathrm{d}r}{r^2}=-z^2\mathrm{d}r\Rightarrow\mathrm{d}r=-\frac{\mathrm{d}r}{z^2} \quad (11.1.162)$$

The integral is then written:

$$\oint\sqrt{A+\frac{2B}{r}+\frac{C}{r^2}}\mathrm{d}r=\oint\sqrt{A+2Bz+Cz^2}\left(-\frac{\mathrm{d}z}{z^2}\right)=\oint-\frac{\sqrt{A}}{z^2}\sqrt{1+\frac{2B}{A}z+\frac{C}{A}z^2}\mathrm{d}z \tag{11.1.163}$$

To find the residue, we will make a Laurent series expansion of:

$$-\frac{\sqrt{A}}{z^2}\sqrt{1+\frac{2B}{A}z+\frac{C}{A}z^2} \tag{11.1.164}$$

around this pole of zero value. To do this, we put:

$$u=\frac{2B}{A}z+\frac{C}{A}z^2 \tag{11.1.165}$$

We know the Taylor expansion (see section Sequences and Series page 1703) of the resulting expression of this variable change:

$$\sqrt{1+u}=1+\frac{1}{2}u-\frac{1}{8}u^2+\ldots \tag{11.1.166}$$

Applied to the radical, we get:

$$\sqrt{1+\frac{2B}{A}z+\frac{C}{A}z^2}=1+\frac{1}{2}\left(\frac{2B}{A}z+\frac{C}{A}z^2\right)-\frac{1}{8}\left(\frac{2B}{A}z+\frac{C}{A}z^2\right)^2+\ldots \tag{11.1.167}$$

It comes then automatically the Laurent series (nice!):

$$\begin{aligned}-\frac{\sqrt{A}}{z^2}\sqrt{1+\frac{2B}{A}z+\frac{C}{A}z^2}&=\frac{\sqrt{A}}{z^2}-\frac{\sqrt{A}}{2z^2}\left(\frac{2B}{A}z+\frac{C}{A}z^2\right)+\frac{\sqrt{A}}{8z^2}\left(\frac{2B}{A}z+\frac{C}{A}z^2\right)^2-\ldots\\&=-\frac{\sqrt{A}}{z^2}-\underbrace{\frac{\sqrt{A}B}{A}}_{\text{residue}}\frac{1}{z}-\frac{\sqrt{A}C}{2A}+\frac{\sqrt{A}}{8z^2}\left(\frac{4B^2z^2}{A^2}+\frac{C^2}{A}z^4+\frac{2BC}{A^2}z^3\right)-\ldots\end{aligned} \tag{11.1.168}$$

where we see quite quickly that the pole is of order 2.

The second residue is the coefficient on $1/z$:

$$r_2=-\frac{B}{\sqrt{A}} \tag{11.1.169}$$

Indeed, we have simply applied the relation proved in the section of Complex Analysis:

$$\text{Res}[g(z);0]=-\lim_{t\to 0}\frac{1}{(k-1)!}\frac{\mathrm{d}^{k-1}}{\mathrm{d}z^{k-1}}\left[z^k\frac{1}{z^2}f\left(\frac{1}{z}\right)\right] \tag{11.1.170}$$

$$\oint\sqrt{A+\frac{2B}{r}+\frac{C}{r^2}}\mathrm{d}r=2\pi\mathrm{i}\sum r_i=2\pi\mathrm{i}\left(\sqrt{C}-\frac{B}{\sqrt{A}}\right)=n_rh \tag{11.1.171}$$

With:

$$\sqrt{C}=\sqrt{-n_\theta^2\hbar^2\left(1-\frac{k^2Z^2\alpha^2}{n_\theta^2}\right)}=\mathrm{i}n_\theta\hbar\sqrt{1-\frac{k^2Z^2\alpha^2}{n_\theta^2}} \tag{11.1.172}$$

For the calculation of $\dfrac{B}{\sqrt{A}}$ we have:

$$\begin{aligned}\frac{B}{\sqrt{A}}&=\frac{\dfrac{kZe^2E'}{c^2}+kmZe^2}{\sqrt{m^2c^2\left[\left(1+\dfrac{E'}{mc^2}\right)^2-1\right]}}=\frac{\dfrac{kZe^2E'}{mc^3}+\dfrac{kZe^2}{c}}{\sqrt{\left(1+\dfrac{E'}{mc^2}\right)^2-1}}=\frac{\dfrac{kZe^2E'}{mc^3}+\dfrac{kZe^2}{c}}{\mathrm{i}\sqrt{1-\left(1+\dfrac{E'}{mc^2}\right)^2}}\\&=-\mathrm{i}\frac{\dfrac{kZe^2E'}{mc^3}+\dfrac{kZe^2}{c}}{\sqrt{1-\left(1+\dfrac{E'}{mc^2}\right)^2}}\end{aligned} \tag{11.1.173}$$

Therefore the curvilinear integral has for expression:

$$\boxed{2\pi\mathrm{i}\left(\mathrm{i}n_\theta\hbar\sqrt{1-\frac{k^2Z^2\alpha^2}{n_\theta^2}}\right)-2\pi\mathrm{i}\left[-\mathrm{i}\frac{\dfrac{kZe^2E'}{mc^3}+\dfrac{kZe^2}{c}}{\sqrt{1-\left(1+\dfrac{E'}{mc^2}\right)^2}}\right]=n_rh} \tag{11.1.174}$$

After simplification:

$$-n_\theta h\sqrt{1-\frac{k^2Z^2\alpha^2}{n_\theta^2}}-\frac{2\pi\left(\frac{kZe^2E'}{mc^3}+\frac{kZe^2}{c}\right)}{\sqrt{1-\left(1+\frac{E'}{mc^2}\right)^2}}=n_rh$$

$$-\frac{\frac{2\pi kZe^2E'}{mc^3}+\frac{2\pi kZe^2}{c}}{\sqrt{1-\left(1+\frac{E'}{mc^2}\right)^2}}=n_\theta h\sqrt{1-\frac{k^2Z^2\alpha^2}{n_\theta^2}}+n_rh$$

$$-\frac{\frac{2\pi kZe^2E'}{mc^3h}+\frac{2\pi kZe^2}{ch}}{\sqrt{1-\left(1+\frac{E'}{mc^2}\right)^2}}=n_\theta\sqrt{1-\frac{k^2Z^2\alpha^2}{n_\theta^2}}+n_r=n_r+\sqrt{n_\theta^2-k^2Z^2\alpha^2}\tag{11.1.175}$$

$$-\frac{\sqrt{1-\left(1+\frac{E'}{mc^2}\right)^2}}{\frac{2\pi kZe^2E'}{mc^3h}+\frac{2\pi kZe^2}{ch}}=\frac{1}{n_r+\sqrt{n_\theta^2-k^2Z^2\alpha^2}}$$

$$-\frac{\sqrt{1-\left(1+\frac{E'}{mc^2}\right)^2}}{kZ\underbrace{\frac{2\pi e^2}{ch}}_{=\alpha}\left(\frac{E'}{mc^2}+1\right)}=\frac{1}{n_r+\sqrt{n_\theta^2-k^2Z^2\alpha^2}}$$

Finally:

$$-\frac{\sqrt{1-\left(1+\frac{E'}{mc^2}\right)^2}}{kZ\left(\frac{E'}{mc^2}+1\right)}=\frac{\alpha}{n_r+\sqrt{n_\theta^2-k^2Z^2\alpha^2}}\tag{11.1.176}$$

We raise squared and at each line below we simplify and rearrange:

$$\begin{aligned}
-\frac{1-\left(1+\dfrac{E'}{mc^2}\right)^2}{k^2Z^2\left(\dfrac{E'}{mc^2}+1\right)}&=\frac{\alpha^2}{\left(n_r+\sqrt{n_\theta^2-k^2Z^2\alpha^2}\right)^2}\\
-\frac{1-\left(1+\dfrac{E'}{mc^2}\right)^2}{\left(\dfrac{E'}{mc^2}+1\right)}&=\frac{k^2Z^2\alpha^2}{\left(n_r+\sqrt{n_\theta^2-k^2Z^2\alpha^2}\right)^2}\\
\frac{1}{\left(1+\dfrac{E'}{mc^2}\right)^2}-1&=\frac{k^2Z^2\alpha^2}{\left(n_r+\sqrt{n_\theta^2-k^2Z^2\alpha^2}\right)^2}\\
\frac{1}{\left(1+\dfrac{E'}{mc^2}\right)^2}&=1+\frac{k^2Z^2\alpha^2}{\left(n_r+\sqrt{n_\theta^2-k^2Z^2\alpha^2}\right)^2}\\
1+\frac{E'}{mc^2}&=\left[1+\frac{k^2Z^2\alpha^2}{\left(n_r+\sqrt{n_\theta^2-k^2Z^2\alpha^2}\right)^2}\right]^{-1/2}
\end{aligned}\tag{11.1.177}$$

Therefore:

$$\frac{E'}{mc^2}=-1+\left[1+\frac{k^2Z^2\alpha^2}{\left(n_r+\sqrt{n_\theta^2-k^2Z^2\alpha^2}\right)^2}\right]^{-1/2}\tag{11.1.178}$$

hence:

$$\begin{aligned}
E'&=mc^2\left\{-1+\left[1+\frac{k^2Z^2\alpha^2}{\left(n_r+\sqrt{n_\theta^2-k^2Z^2\alpha^2}\right)^2}\right]^{-1/2}\right\}\\
&=-mc^2\left\{1-\left[1+\frac{k^2Z^2\alpha^2}{\left(n_r+\sqrt{n_\theta^2-k^2Z^2\alpha^2}\right)^2}\right]^{-1/2}\right\}
\end{aligned}\tag{11.1.179}$$

We put $n=n_r+n_\theta$. By working on the denominator:

$$n_r+\sqrt{n_\theta^2-k^2Z^2\alpha^2}\tag{11.1.180}$$

we can transform:

$$n_r+n_\theta\sqrt{1-\left(\frac{kZ\alpha}{n_\theta}\right)^2}\tag{11.1.181}$$

By adding and subtracting n_θ:

$$(n_r+n_\theta)+n_\theta\left(\sqrt{1-\left(\frac{kZ\alpha}{n_\theta}\right)^2}-1\right)=n+n_\theta\left(\sqrt{1-\left(\frac{kZ\alpha}{n_\theta}\right)^2}-1\right) \tag{11.1.182}$$

Therefore:

$$E'=-mc^2\left\{1-\left[1+\frac{k^2Z^2\alpha^2}{\left[n+n_\theta\left(\sqrt{1-\left(\frac{kZ\alpha}{n_\theta}\right)^2}-1\right)\right]^2}\right]^{-1/2}\right\} \tag{11.1.183}$$

or also:

$$E'=-mc^2\left\{1-\left[1+\frac{k^2Z^2\left(\frac{\alpha}{n}\right)^2}{\left[1+\frac{n_\theta}{n}\left(\sqrt{1-\left(\frac{kZ\alpha}{n_\theta}\right)^2}-1\right)\right]^2}\right]^{-1/2}\right\} \tag{11.1.184}$$

or also:

$$E'=-mc^2\left(1-\left\{1+k^2Z^2\left(\frac{\alpha}{n}\right)^2\left[1+\frac{n_\theta}{n}\left(\sqrt{1-\left(\frac{kZ\alpha}{n_\theta}\right)^2}-1\right)\right]^{-2}\right\}^{-1/2}\right) \tag{11.1.185}$$

We consider in the term:

$$\left(\sqrt{1-\left(\frac{kZ\alpha}{n_\theta}\right)^2}-1\right) \tag{11.1.186}$$

the radical that can be written:

$$\sqrt{1-\left(\frac{kZ\alpha}{n_\theta}\right)^2}\left\{1+\left[-\left(\frac{kZ\alpha}{n_\theta}\right)^2\right]\right\}^{1/2} \tag{11.1.187}$$

Given the series development (see section Sequences and Series page 1703) in analogy with the previous relation:

$$(1+x)^{1/2}=1+\frac{1}{2}x-\frac{1}{8}x^2+\frac{3}{48}x^3-\ldots \tag{11.1.188}$$

then:

$$\sqrt{1-\left(\frac{kZ\alpha}{n_\theta}\right)^2}=\left\{1+\left[-\left(\frac{kZ\alpha}{n_\theta}\right)^2\right]\right\}^{1/2}=1+\frac{1}{2}\left[-\left(\frac{kZ\alpha}{n_\theta}\right)^2\right]-\frac{1}{8}\left[-\left(\frac{kZ\alpha}{n_\theta}\right)^2\right]+\ldots \tag{11.1.189}$$

Therefore:

$$\left\{1+\left[-\left(\frac{kZ\alpha}{n_\theta}\right)^2\right]\right\}^{1/2}=1-\frac{1}{2}\left(\frac{kZ\alpha}{n_\theta}\right)^2-\frac{1}{8}\left(\frac{kZ\alpha}{n_\theta}\right)^4+\ldots \tag{11.1.190}$$

As $\dfrac{\alpha}{n_\theta}\ll 1$, we can neglect the terms above the order 2 such that:

$$\left\{1+\left[-\left(\frac{kZ\alpha}{n_\theta}\right)^2\right]\right\}^{1/2}\cong 1-\frac{1}{2}\left(\frac{kZ\alpha}{n_\theta}\right)^2 \tag{11.1.191}$$

Finally we can the write:

$$\left(\sqrt{1-\left(\frac{kZ\alpha}{n_\theta}\right)^2}-1\right)\cong 1-\frac{1}{2}\left(\frac{kZ\alpha}{n_\theta}\right)^2-1=-\frac{1}{2}\left(\frac{kZ\alpha}{n_\theta}\right)^2 \tag{11.1.192}$$

Working on the term in the brackets we get taking into account our previous result:

$$\begin{aligned}E'&=\left[1+\frac{n_\theta}{n}\left(\sqrt{1-\left(\frac{kZ\alpha}{n_\theta}\right)^2}-\right)-1\right]^{-2}\cong\left[1+\frac{n_\theta}{n}\left(-\frac{1}{2}\left(\frac{kZ\alpha}{n_\theta}\right)^2\right)\right]^{-2}\\&=\left[1-\frac{1}{2}\frac{n_\theta}{n}\left(\frac{kZ\alpha}{n_\theta}\right)^2\right]^{-2}\end{aligned} \tag{11.1.193}$$

Given the following Taylor series development (see section Sequences and Series page 1703):

$$(1+x)^{-2}=1-2x+3x^2-\ldots \tag{11.1.194}$$

then:

$$\left[1+\left(-\frac{1}{2}\frac{n_\theta}{n}\left(\frac{kZ\alpha}{n_\theta}\right)^2\right)\right]^{-2}=1-2\left[-\frac{1}{2}\frac{n_\theta}{n}\left(\frac{kZ\alpha}{n_\theta}\right)^2\right]+3\left[-\frac{1}{2}\frac{n_\theta}{n}\left(\frac{kZ\alpha}{n_\theta}\right)^2\right]^2-\ldots \tag{11.1.195}$$

Neglecting terms over order 2 we have finally:

$$\left[1+\frac{n_\theta}{n}\sqrt{1-\left(\frac{kZ\alpha}{n_\theta}\right)^2}-1\right]^{-2}\cong 1+\frac{n_\theta}{n}\left(\frac{kZ\alpha}{n_\theta}\right)^2 \tag{11.1.196}$$

Therefore the term between braces can be written:

$$\left\{1\left(\frac{kZ\alpha}{n}\right)^2\left[1+\frac{n_\theta}{n}\left(\sqrt{1-\left(\frac{kZ\alpha}{n_\theta}\right)^2}-1\right)\right]^{-2}\right\}^{-1/2}\cong\left\{1+\left(\frac{kZ\alpha}{n}\right)^2\left[1+\frac{n_\theta}{n}\left(\frac{kZ\alpha}{n_\theta}\right)^2\right]\right\} \tag{11.1.197}$$

We undertake the Taylor series development of the right term between braces using again:

$$(1+x)^{-1/2}=1-\frac{1}{2}x+\frac{3}{8}x^2-\frac{15}{48}x^3+\ldots \tag{11.1.198}$$

Neglecting terms over order 2 we have finally (an approximation of an approximation that is itself an approximation...):

$$\begin{gathered}\left\{1+\underbrace{\left(\frac{kZ\alpha}{n}\right)^2\left[1+\frac{n_\theta}{n}\left(\frac{kZ\alpha}{n_\theta}\right)^2\right]}_{:=x}\right\}^{-1/2}\\ \cong 1-\frac{1}{2}\left\{\left(\frac{kZ\alpha}{n}\right)^2\left[1+\frac{n_\theta}{n}\left(\frac{kZ\alpha}{n_\theta}\right)^2\right]\right\}+\frac{3}{8}\left\{\left(\frac{kZ\alpha}{n}\right)^2\left[1+\frac{n_\theta}{n}\left(\frac{kZ\alpha}{n_\theta}\right)^2\right]\right\}^2\end{gathered} \tag{11.1.199}$$

By developing the square of the third term, we get:

$$\left\{\left(\frac{kZ\alpha}{n}\right)^2\left[1+\frac{n_\theta}{n}\left(\frac{kZ\alpha}{n_\theta}\right)^2\right]\right\}^2=\left\{\left(\frac{kZ\alpha}{n}\right)^4\left[1+\underbrace{\left(\frac{n_\theta}{n}\right)^2\left(\frac{kZ\alpha}{n}\right)^4}_{\ll 1}+2\underbrace{\frac{n_\theta}{n}\left(\frac{kZ\alpha}{n}\right)^2}_{\ll 1}\right]\right\} \tag{11.1.200}$$

Therefore:

$$\left\{1+\underbrace{\left(\frac{kZ\alpha}{n}\right)^2\left[1+\frac{n_\theta}{n}\left(\frac{kZ\alpha}{n_\theta}\right)^2\right]}_{:=x}\right\}^{-1/2}\cong 1-\frac{1}{2}\left\{\left(\frac{kZ\alpha}{n}\right)^2\left[1+\frac{n_\theta}{n}\left(\frac{kZ\alpha}{n_\theta}\right)^2\right]\right\}+\frac{3}{8}\left[\left(\frac{kZ\alpha}{n}\right)^2\right]^2 \tag{11.1.201}$$

The total energy of this atom the becomes:

$$\begin{aligned}
{E'}_{n,n_\theta}&=-mc^2\left(1-\left\{1+k^2Z^2\left(\frac{\alpha}{n}\right)^2\left[1+\frac{n_\theta}{n}\left(\sqrt{1-\left(\frac{kZ\alpha}{n_\theta}\right)^2}-1\right)\right]^{-2}\right\}^{-1/2}\right)\\
&\cong-mc^2\left(1-\left\{1-\frac{1}{2}\left\{\left(\frac{kZ\alpha}{n}\right)^2\left[1+\frac{n_\theta}{n}\left(\frac{kZ\alpha}{n_\theta}\right)^2\right]\right\}+\frac{3}{8}\left[\left(\frac{kZ\alpha}{n}\right)^2\right]^2\right\}\right)\\
&\cong-mc^2\left(1-\left\{1-\frac{k^2Z^2}{2}\left\{\left(\frac{\alpha}{n}\right)^2\left[1+\frac{n_\theta}{n}\left(\frac{kZ\alpha}{n_\theta}\right)^2\right]\right\}+\frac{3}{8}\left[\left(\frac{kZ\alpha}{n}\right)^2\right]^2\right\}\right)\\
&=-mc^2\left(\frac{k^2Z^2}{2}\left\{\left(\frac{\alpha}{n}\right)^2\left[1+\frac{n_\theta}{n}\left(\frac{kZ\alpha}{n_\theta}\right)^2\right]\right\}-\frac{3}{8}\left[\left(\frac{kZ\alpha}{n}\right)^2\right]^2\right)\\
&=-\frac{mc^2k^2Z^2\alpha^2}{2n^2}\left[1+\frac{n_\theta}{n}\left(\frac{kZ\alpha}{n_\theta}\right)^2-\frac{3}{4}\left(\frac{kZ\alpha}{n}\right)^2\right]\\
&=-\frac{mc^2k^2Z^2\alpha^2}{2n^2}\left[1+\frac{n_\theta n}{n^2}\left(\frac{kZ\alpha}{n_\theta}\right)^2-\frac{3}{4}\left(\frac{kZ\alpha}{n}\right)^2\right]\\
&=-\frac{mc^2k^2Z^2\alpha^2}{2n^2}\left[1+\frac{n}{n_\theta}\left(\frac{kZ\alpha}{n}\right)^2-\frac{3}{4}\left(\frac{kZ\alpha}{n}\right)^2\right]
\end{aligned}\tag{11.1.202}$$

Finally we get of the expression of the energy:

$${E'}_{n,n_\theta}\cong-\frac{mc^2k^2Z^2\alpha^2}{2n^2}\left\{1+\left(\frac{kZ\alpha}{n}\right)^2\left[\frac{n}{n_\theta}-\frac{3}{4}\right]\right\}\tag{11.1.203}$$

We can give another expression for the energy of the hydrogenoid atom as:

$$k=\frac{1}{4\pi\varepsilon_0}\quad\text{and}\quad\alpha=\frac{e^2}{\hbar c}=\frac{2\pi e^2}{hc}\tag{11.1.204}$$

The expression of the total energy of the hydrogen-like atom becomes then:

$${E'}_{n,n_\theta}\cong-\frac{mc^2\left(\dfrac{1}{16\pi^2\varepsilon_0^2}Z^2\left(\dfrac{4\pi^2e^4}{h^2c^2}\right)\right)}{2n^2}\left\{1+\left[\frac{\dfrac{1}{4\pi\varepsilon_0}Z\dfrac{2\pi e^2}{hc}}{n}\right]^2\left[\frac{n}{n_\theta}-\frac{3}{4}\right]\right\}\tag{11.1.205}$$

Therefore:

$${E'}_{n,n_\theta}\cong-\frac{mZ^2e^4}{8\varepsilon_0^2h^2n^2}\left\{1+\frac{Z^2e^4}{4\varepsilon_0^2h^2c^2n^2}\left[\frac{n}{n_\theta}-\frac{3}{4}\right]\right\}\tag{11.1.206}$$

In the literature, we find other expressions for the total energy that are more interesting than the previous one (because more traditional). So, considering that $n=n_r+n_\theta$, we get:

$${E'}_{n,n_\theta}\cong-\frac{mc^2k^2Z^2\alpha^2}{2n^2}\left\{1+\left(\frac{kZ\alpha}{n}\right)^2\left[\frac{n_r+n_\theta}{n_\theta}-\frac{3}{4}\right]\right\}=-\frac{mc^2k^2Z^2\alpha^2}{2n^2}\left\{1+\left(\frac{kZ\alpha}{n}\right)^2\left[\frac{1}{4}+\frac{n_r}{n_\theta}\right]\right\}\tag{11.1.207}$$

If we seek for an expression using to the Rydberg constant R_H (see above) we get:

$$E'_{n,n_\theta} \cong \underbrace{\frac{me^4}{8\varepsilon_0^2 h^3 c}}_{R_H} Z^2 \frac{hc}{n^2}\left\{1+\frac{Z^2e^2}{4\varepsilon_0^2h^2c^2n^2}\left[\frac{n}{n_\theta}-\frac{3}{4}\right]\right\}=-\frac{hc}{n^2}Z^2R_H\left\{1+\frac{Z^2e^2}{4\varepsilon_0^2h^2c^2n^2}\left[\frac{n}{n_\theta}-\frac{3}{4}\right]\right\} \tag{11.1.208}$$

So the most condensed expression of the total relativistic energy of the hydrogenoid that we can find in the literature and we adopt in this textbook is:

$$\boxed{E'_{n,n_\theta} \cong -\frac{hc}{n^2}Z^2R_H\left\{1+\frac{Z^2e^2}{4\varepsilon_0^2h^2c^2n^2}\left[\frac{n}{n_\theta}-\frac{3}{4}\right]\right\}} \tag{11.1.209}$$

The above relation shows well the existence of a fine structure as the characteristics n_r and $n_\theta$of the electron orbit appear separately in a ratio and not just as a sum as in the first model of Sommerfeld and Wilson.

But strictly speaking, we should write because of the nucleus movement:

$$E'_{n_r,n_\theta} \cong -\frac{\mu Z^2e^4}{8\varepsilon_0^2h^2n^2}\left\{1+\frac{Z^2e^4}{4\varepsilon_0^2h^2c^2n^2}\left[\frac{n}{n_\theta}-\frac{3}{4}\right]\right\} \tag{11.1.210}$$

or:

$$E'_{n_r,n_\theta} \cong -\frac{hc}{n^2}Z^2R'_H\left\{1+\frac{Z^2e^2}{4\varepsilon_0^2h^2c^2n^2}\left[\frac{1}{4}+\frac{n_r}{n_\theta}\right]\right\} \tag{11.1.211}$$

In which the Rydberg constant R'_H is defined as:

$$R'_H=\frac{\mu e^4}{8\varepsilon_0^2h^3c} \tag{11.1.212}$$

However, as the mass of the core is $1,840$ times heavier than that of the electron, we can admit that as a first approximation in the case of the hydrogenous atom:

$$\mu=\frac{m_em_N}{m_e+m_N}=\frac{m_em_p}{m_e+m_p}\cong m_e \tag{11.1.213}$$

11.1.6.1 Magnetic dipole moment

At the time of the development of the Sommerfeld model, some physicists strive to study another property of the atom. They observed that under the application of the magnetic field, the spectroscopic lines are doubled. To explain this, they had the brilliant and very simple idea to explain this phenomenon by the magnetic moment of the electron.

Remark

We will see during our study of Wave Quantum Physics, that even in the absence of a magnetic field a very thing measuring of the rays shows that they are all double and this because of the "spin-orbit coupling". Therefore, a correct interpretation is that the magnetic field doubles the duplication of rays.

Thus, given the expression of the norm of the magnetic dipole moment (see section Magnetostatics page 4037):

$$\mu_l = S \cdot I \tag{11.1.214}$$

The magnetic moment is equal, in the corpuscular point of view...., to the area enclosed by the orbit of the electron multiplied by the current of the electron (perpendicular to the unit vector of the surface) on its orbit path thus:

$$\mu_l = \pi r_n^2 \frac{e}{T_n} \tag{11.1.215}$$

where:

$$T_n = \frac{2\pi r_n}{v_n} \tag{11.1.216}$$

is the period of the movement.

We have seen that the total angular momentum being equal to:

$$b = n\hbar = m r_n v_n \tag{11.1.217}$$

thus the ratio magnetic moment/angular momentum gives:

$$\frac{\mu_l}{b} = \frac{\pi r_n^2 \dfrac{e}{T_n}}{m r_n v_n} = \frac{e}{2m} \Leftrightarrow \mu_l = \frac{e}{2m} b = \frac{e\hbar}{2m} n = \mu_B n \tag{11.1.218}$$

The ratio $e/2m$ equation is named the "orbital gyromagnetic ratio (for a classical rotating body)" and is equal to:

$$\gamma_e = \frac{e}{2m} \cong -8.792 \cdot 10^{10}\,[\mathrm{C} \cdot \mathrm{kg}^{-1}] \tag{11.1.219}$$

and the quantity:

$$\boxed{\mu_B \cong \frac{\hbar}{2}\gamma_e = -4.636 \cdot 10^{-24}\,[\mathrm{A} \cdot \mathrm{m}^2] = -4.636 \cdot 10^{-24}\,[\mathrm{J} \cdot \mathrm{T}^{-1}]} \tag{11.1.220}$$

is the "Bohr Magneton".

Remark

It is important to remember the few developments and definitions just be made above for when we will develop the Pauli equation in the section of Relativistic Quantum Physics.

Frequently we write the above relation as well:

$$\boxed{\mu_l=\frac{en}{2m}\hbar=m_l\hbar} \tag{11.1.221}$$

where m_l is named "magnetic quantum number".

Knowing that the principal quantum number n is divided by the radial n_r and azimuthal n_θ quantum numbers, there are so many magnetic moments as there are different orbital geometries for a given value of the principal quantum number. In fact, there are even the double if we consider that the electron can rotate in clockwise or counter-clockwise (the magnetic moment is a vector quantity!).

Now let us take two examples using the previous proven rules:

$$n=1=n_\theta+n_r=1+0 \tag{11.1.222}$$

and:

$$n=2=n_\theta+n_r=\begin{cases}2=1+1\\2=2+0\end{cases} \tag{11.1.223}$$

for which we put now:

$$\boxed{l:=n_r} \tag{11.1.224}$$

that is the number that we name "quantum number of orbital angular momentum" and having values between (as shown in the both previous examples):

$$\boxed{0\geq l\geq n-1} \tag{11.1.225}$$

What have we finally so far?

1. When $n=1$, we can only have $l=0$ and as $n=1$, there is only one sub-layer, so when applying a magnetic field we still always have one and only one visible line:

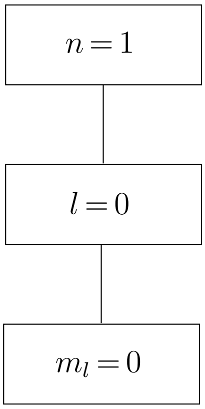


Figure 11.15: Decomposition of a very low orbital...

2. When $n=2$, we have $l=0$ and $l=1$ and as $n=2$ has two sub-layers. When no magnetic field is applied, the spectrum lines of the two sub-layers are superimposed so indistinguishable (we see only one). But when a magnetic field is applied to the two sub-layers are distinguished by their magnetic moment and therefore we-have two lines but in total there are 3 theoretically (without a field, and two with field):

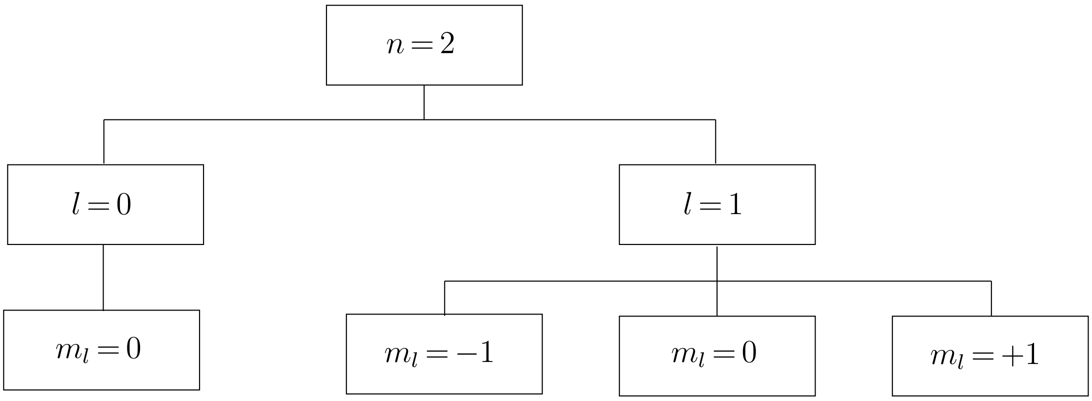


Figure 11.16: **Decomposition of an upper orbital...**

Thus we have:

$$\boxed{-l\leq m_l\leq l}\tag{11.1.226}$$

where for recall:

$$\boxed{0\leq l\leq n-1}\tag{11.1.227}$$

The potential energy of a magnetic moment μ_l placed in a magnetic field B is (see section Magnetostatics page 4050):

$$E=\mu_l B\cos(\alpha)=m_l\hbar B\cos(\alpha)\tag{11.1.228}$$

So finally for each orbital electron subjected to a magnetic field, we have:

$$\boxed{E\cong E'_{n_r,n_\theta}+m_l\hbar B\cos(\alpha)\cong-\frac{hc}{n^2}Z^2R'_H\left\{1+\frac{Z^2e^2}{4\varepsilon_0^2h^2c^2n^2}\left[\frac{1}{4}+\frac{n_r}{n_\theta}\right]\right\}+m_l\hbar B\cos(\alpha)}\tag{11.1.229}$$

always with:

$$-l\leq m_l\leq l\tag{11.1.230}$$

The observation of the spectrum of an atom in a magnetic field has the effect of adding spectrum lines by the potential energy of the magnetic moment. This is what we name the "Zeeman effect" because it is the latter who measured these spectrum lines for the first time (before the theory that we will see later).

11.1.6.2 Spin

The concept of electron spin in particle physics was introduced in 11925 (holocene calendar(by George Uhlenbeck and Samuel Goudsmit, who proposed that the electron possesses an intrinsic angular momentum. This idea explained puzzling features of various experimental facts that led to attribute to the electron an intrinsic magnetic moment and especially the experiment of the duplication of spectrum lines named "abnormal Zeeman effect".

It has indeed been experimentally measured that the magnetic moment was just equal to the value of the Bohr magneton. It is then tempting to attribute this to the electron magnetic moment and speculate that it might come from the fact that it turns on itself (intrinsic angular momentum[4]), i.e. it would have a "spin" equal to the Bohr magneton and this latter could take negative or positive values. We speak then of "spin quantum number" or historically "Pauli quantum number" and it gives the number of different values that can take the spin. Therefore the spin is one of the two types of angular momentum in quantum mechanics, the other being the orbital angular momentum

However, this classic vision of a intrinsic rotation (intrinsic angular momentum) of the particle is actually too naive and therefore erroneous! As we will see later Paul Dirac's relativistic equation for the electron naturally incorporated spin (and antiparticles) as a fundamental property.

Indeed, at first, if the particle is punctual, the notion of intrinsic rotation around its axis is simply devoid of physical sense. Recall that since by definition, the axis of rotation of an object is the locus of points of the object which remain stationary, so if the particle is punctual, its own axis is on the particle, thus the latter is stationary...

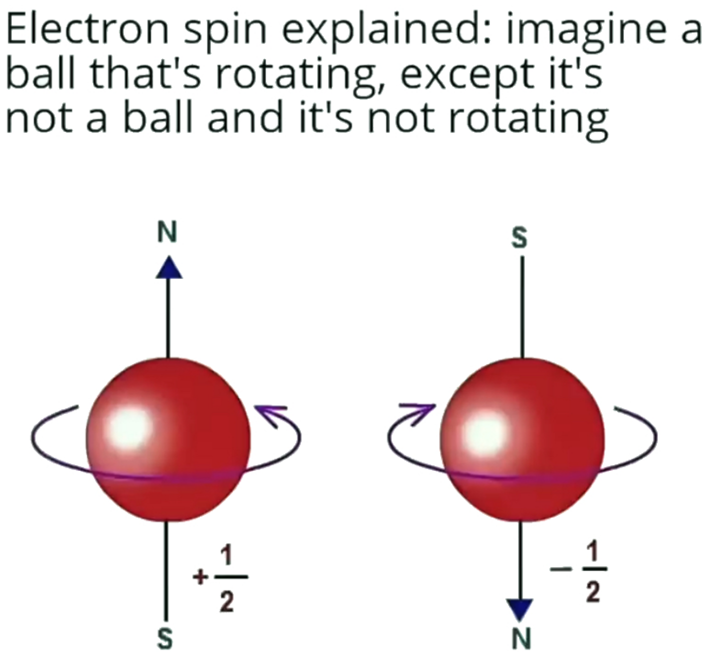

Electron spin explained: imagine a ball that's rotating, except it's not a ball and it's not rotating
N
S
$+\frac{1}{2}$
S
N
$-\frac{1}{2}$

Figure 11.17: **A meme (joke) about spin...**

---

[4] This expression is used for non-rotating objects that obey the law of conservation of angular momentum as if they were rotating.

Secondly, if the particle is not punctual, then the rotation has a meaning, but it bring in this case to another difficulty. For example, suppose that the particle is an electron, modelled as a radius of a spherical body. We obtain an estimate of the radius by writing that the mass energy of the electron is of the order of magnitude of its electrostatic potential energy (see section Electrostatic page 4005), that is:

$$mc^2 \cong \frac{e^2}{4\pi\varepsilon_0 r_c} \Rightarrow r_c \cong \frac{e^2}{4\pi\varepsilon mc^2} \tag{11.1.231}$$

The numerical value of this "classical electron radius" is $r_c \cong 10^{-15}$ [m] taking its mass at rest.

If we attribute then this electron an angular momentum equal to $\hbar/2$ (which has units of angular momentum), we get for a point on the equator of the electron a speed v satisfying:

$$mr_c v = \frac{\hbar}{2} \Rightarrow v = \frac{\hbar}{2mr_c} \cong \frac{2\pi\varepsilon_0 \hbar c^2}{e^2} \tag{11.1.232}$$

The numerical value of the speed is therefore roughly $v \cong 6 \cdot 10^{10}\,[\text{m}\cdot\text{s}^{-1}]$... so the speed is higher than the speed of light in vacuum... (as we will see during our study of Quantum Field Theory, vacuum isn't really empty), which obviously causes problems with the theory of relativity (see section Special Relativity page 5629).

We can not with the mathematical tools of particle quantum physics rigorously formalize the notion of spin, but we will come back on this subject in the section of Relativistic Quantum Physics (especially during our study of the Pauli equation) and show that the spin is in fact (sadly) something much more subtle than a simple rotation...

But let us come back to our classical point of view. So when we see a duplication of spectrum lines (abnormal Zeeman effect), we assume that this is due to the electron spin s which can take two different directions (vector direction!).

It has been measured that the intrinsic magnetic moment of the electron is equal to the value of the Bohr magneton, that is:

$$\mu_l = \mu_B = \frac{e}{2m}\hbar \tag{11.1.233}$$

If we put $e = 1, m = 1$ (what physicists like to do...) then we have:

$$\boxed{\mu_l = \mu_B = \frac{e}{2m}\hbar \underset{\substack{e=1\\m=1}}{=} \frac{1}{2}\hbar = s\hbar} \tag{11.1.234}$$

(this just to get a similarity with $b = n\hbar$...)

This value is constant but may be negative or positive depending on the intrinsic rotation direction of the electron relatively to the observer (the angular momentum having a vector direction). So:

$$\boxed{s = \pm\frac{1}{2}} \tag{11.1.235}$$

and therefore in practice, spin is given as a dimensionless spin quantum number

This result, of the utmost importance, also leads us to the conclusion that each magnetic quantum number is degenerate twice by the spin quantum number! Thus, as we shall see a little further into concrete examples (with supporting diagrams), each principal quantum number n is say to be "degenerated" a number $2n^2$ of times.

The sceptical reader should know that Spin has important theoretical implications and practical applications. Well-established direct applications of spin include (we will come back on some elements of this list in further section but in a very technical point of view):

- Nuclear magnetic resonance (NMR) spectroscopy in chemistry
- Electron spin resonance spectroscopy in chemistry and physics
- Magnetic resonance imaging (MRI) in medicine, a type of applied NMR, which relies on proton spin density

SIEMENS Healthineers
MAGNETOM Altea

Figure 11.18: **Siemens MRI scan Healthineers model**

- Giant magnetoresistive (GMR) drive head technology in modern hard disks.

Electron spin plays an important role in magnetism, with applications for instance in computer memories. The manipulation of nuclear spin by radiofrequency waves (nuclear magnetic resonance) is important in chemical spectroscopy and medical imaging.

11.1.6.3 Pauli exclusion principle

The Pauli exclusion principle is the quantum mechanical principle that states that two identical fermions (particles with half-integer spin) cannot occupy the same quantum state simultaneously. In the case of electrons, it can be stated as follows: it is impossible for two electrons of a poly-electron atom to have the same values of the four quantum numbers: n, the principal quantum number, l the angular momentum quantum number, m_l the magnetic quantum number, and m_s the spin quantum number. For example, if two electrons reside in the same orbital, and if their n, l, and m_l values are the same, then their m_s must be different, and thus the electrons must have opposite half-integer spins of $1/2$ and $-1/2$. This principle was formulated by Austrian physicist Wolfgang Pauli in 11925 (holocene calendar) for electrons, and later extended to all fermions with his spin-statistics theorem of 11940 (holocene calendar).

Particles with an integer spin, or bosons, are not subject to the Pauli exclusion principle: any number of identical bosons can occupy the same quantum state, as with, for instance, photons produced by a LASER and Bose–Einstein condensate (see section Statistical Mechanics page 3576).

Therefore to summarize, due to the fact that the state of an atomic electron can be characterized with at least the following 4 quantum numbers (the first quantum number coming from Bohr, the two others from Sommerfeld and the last one by Pauli) then we have just proved the origin:

$$\boxed{n,l,m_l,m_s} \tag{11.1.236}$$

or under the following extended form:

$$\boxed{n=n_\theta+n_r,l,m_l,m_s} \tag{11.1.237}$$

Wolfgang Pauli has then postulate to explain some regularities in the atomic properties (specifically: chemistry) a principle of exclusion today named "Pauli exclusion principle" (PEP) and stated as follows:

In an atom, two electrons can't have the same ordered quadruplet n,l,m_l,m_s of quantum numbers.

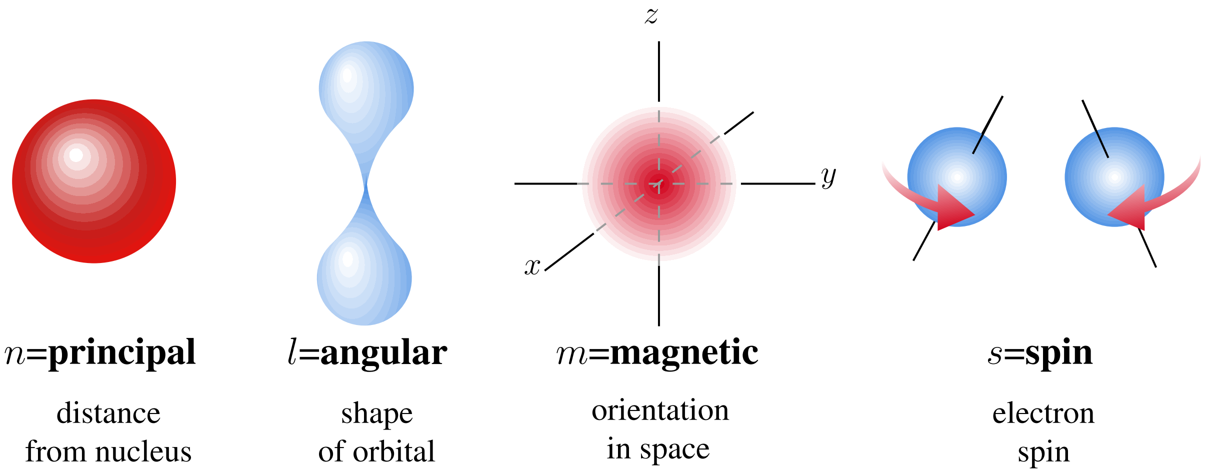


Figure 11.19: Summary of quantum numbers

A more rigorous statement is that concerning the exchange of two identical particles: the total (many-particle) wave function is antisymmetric for fermions, and symmetric for bosons. This means that if the space and spin coordinates of two identical particles are interchanged, then the total wave function changes its sign for fermions and does not change for bosons.

> Caution!!!! The term "wave function" in Quantum Physics is somewhat misleading because it suggests that quantum states behave exactly like classical waves, which isn't entirely accurate. While the wave function Ψ does exhibit wave-like properties - such as interference and diffraction - it is actually a mathematical representation of a probabilistic quantum system's state rather than a physical wave. Sadly this confusing terminology still remains in this early 121st century (holocene calendar) and we will also sadly use it.

Formally, let us take for example a two particle fermion system that are not interacting. Because they are not interacting we can assume the two-particle wavefunction can be written as a product of the single particle wavefunctions (by the principle of probability independence!). Let us label the two single particles with a_1 and a_2, then we have:

$$\psi\left(\vec{r}_1,\vec{r}_2\right)=\psi_{a_1}\left(\vec{r}_1\right)\psi_{a_2}\left(\vec{r}_2\right) \tag{11.1.238}$$

Since we can't distinguish between the two particles, we can also write the above wavefunction as:

$$\psi'\left(\vec{r}_1,\vec{r}_2\right)=\psi_{a_1}\left(\vec{r}_2\right)\psi_{a_2}\left(\vec{r}_1\right) \tag{11.1.239}$$

All we can say is that the system must be in a linear superposition of ψ and ψ'. Mathematically we can only combine the two in only two correctly normalised way: First the symmetric case (bosons, e.g. photons):

$$\Psi_s\left(\vec{r}_1,\vec{r}_2\right)=\frac{1}{\sqrt{2}}\left[\psi_{a_1}\left(\vec{r}_1\right)\psi_{a_2}\left(\vec{r}_2\right)+\psi_{a_1}\left(\vec{r}_2\right)\psi_{a_2}\left(\vec{r}_1\right)\right] \tag{11.1.240}$$

And the second case being the antisymmetric combination (fermions, e.g. electrons):

$$\Psi_{\text{anti}}\left(\vec{r}_1,\vec{r}_2\right)=\frac{1}{\sqrt{2}}\left[\psi_{a_1}\left(\vec{r}_1\right)\psi_{a_2}\left(\vec{r}_2\right)-\psi_{a_1}\left(\vec{r}_2\right)\psi_{a_2}\left(\vec{r}_1\right)\right] \tag{11.1.241}$$

Now since we're considering fermions, for identical single particles i.e. $a_1=a_2$, then $\Psi_{\text{anti}}=0$, which means the probability amplitude of two fermions occupying the same state is 0. So we see that just by considering the form of the wavefunction for a system of identical elements we managed to arrive at Pauli's exclusion principle. Finally, as we can tell from $\Psi_s\neq 0$ for identical particles, simply implies that bosons do not follow such exclusion principle and nothing prevents them from occupying the same state.

Remarks

R1. Sometimes, depending on the situations, we write s rather than m_s (but this is not very important...).

R2. Once again we know from the Wave Quantum Physics that the exclusion principle applies to particles that are "fermions". These are the particles (elementary or composite) which have half-integer spin, such as the proton, neutron and neutrino. This principle does not apply to the group of particles named "bosons" which have zero or integer spin.

It is possible from this principle, to establish a kind of catalogue of atomic elements from the orbital filling possibilities, assumed layered, improving Mendeleev classification (see just further below the corresponding table).

Students often see them for the first time in school when they study chemistry. They use it most of the time, without knowing what they really represent (and also sometimes the teachers...).

Notice that Pauli himself couldn't put a finger why the principle was the way it was. From his Nobel lecture (11945 according to holocene calendar): «*Already in my original paper I stressed the circumstance that I was unable to give a logical reason for the exclusion principle or to deduce it from more general assumptions. I had always the feeling and I still have it today, that this is a deficiency. ... The impression that the shadow of some incompleteness [falls] here on the bright light of success of the new quantum mechanics seems to me unavoidable.*»

It wasn't that the principle's ontology was sorted over time. In 11963 (holocene calendar), Richard Feynman said: «.... *Why is it that particles with half-integral spin are Fermi particles (...) whereas particles with integral spin are Bose particles (...)? We apologize for the fact that we can not give you an elementary explanation. An explanation has been worked out by Pauli from complicated arguments from Quantum Field Theory and Special Relativity. He has shown that the two must necessarily go together, but we have not been able to find a way to reproduce his arguments on an elementary level. It appears to be one of the few places in physics where there is a rule which can be stated very simply, but for which no one has found a simple and easy explanation. (...) This probably means that we do not have a complete understanding of the fundamental principle involved. For the moment, you will just have to take it as one of the rules of the world.*»

Remark

The Ramberg and Snow experiment (see the literature on the subject) tried to falsify the PEP (Pauli Exclusion Principle) in 11990 (holocene calendar). Ramberg and Snow did not detect any violation but were able to put an upper bound on the violation probability $<1.7\cdot 10^{-27}$. Following their concept, a much improved version of the experiment, named "VIP" (violation of the Pauli principle), was set up in the LNGS underground laboratory in Gran Sasso, Italy, in 12006 (holocene calendar). After four years of data taking, VIP set a new limit on the EP (Exclusion Principle) violation for electrons at $<4.7\cdot 10^{-29}$. After improvements in 12017 (holocene calendar) the upper bound was moved to $<10^{-31}$.

To conclude, we note that the EP has many associated philosophical issues, as Pauli himself was well aware of, and these are being studied within a dedicated project involving VIP collaborators, and supported by the John Templeton Foundation. One such issue is the notion of "identicalness", which does not seem to have an analogue outside quantum mechanics because there are no two fundamentally identical classical objects.

11.1.7 Electron configuration (atomic orbital)

In the years 11920 (holocene calendar), Niels Bohr, Edmund Clifton Stoner and others designed a model of the electronic structure of atoms that gives the possibility to understand quite well the periodic table of elements. The work of Henry Moseley was used to determine the number of protons in the nucleus and, as the atom is globally neutral, then also the number of orbital electrons. It is not simple to determine the atomic structure and in this analysis, physicists have been helped by the experiments conducted by chemists (as for highly complexed atoms, electrons interacts together!).

Thus, according to chemists, electrons occupy layers and sub-layers around the core by (average) increasing energy order based on rules associated with their quantum numbers we have previously determined. Thus, the "electronic configuration" is the arrangement of electrons in an atom, molecule or another body. Specifically, it is the position of the electrons in an atomic orbital, molecular or other forms of electronic orbitals.

Knowledge of the electron configuration of different atoms is useful in understanding the structure of the periodic table of elements. The concept is also useful for describing the chemical bonds that hold atoms together. In bulk materials, this same idea helps explain the peculiar properties of LASERs and semiconductors.

Remark

Rigorously the concept of "layer" as we can imagine it has absolutely no sense if we refer to the results of quantum mechanics (see section Wave Quantum Physics page 4688)!!! This is why the debate on how to fill the layers is sterile because it does not strictly exist without a rough approximation of the general rule

Each layer corresponds to a specific value of the "principal quantum number n" and traditionally these layers are designated (this tradition should be abandoned ... but as all the traditions it has a long life...) by the capital letters:

$$K, L, M, N, O, \ldots \qquad (11.1.242)$$

corresponding to the principal numbers 1, 2, 3, 4, 5 ... that can take the principal quantum number n. The "second/azimuthal quantum number" conventionally denoted by the letter l corresponds to the degenerate states that can take the layers (shells) for a given value of n such that:

For the layer K $(n=1)$ we have as we know a unique underlayer (sub-shell):

$$[n_\theta, n_r = 0] := s \qquad (11.1.243)$$

and therefore for the layer L $(n=2)$ we have two sub-layers:

$$\begin{gathered}[n_\theta = 1, n_r = 1] := p\\ [n_\theta = 2, n_r = 0] := s\end{gathered} \qquad (11.1.244)$$

and so for M $(n=3)$ we have three sub-layers:

$$\begin{gathered}[n_\theta = 1, n_r = 2] := d\\ [n_\theta = 2, n_r = 1] := p \quad [n_\theta = 3, n_r = 0] := s\end{gathered} \qquad (11.1.245)$$

and so on...

Chemists are accustomed to note the first sub-layers with the Latin letters:

$$s\,\text{(sharp)}, p\,\text{(main)}, d\,\text{(diffuse)}, f\,\text{(fundamental)}\ldots \tag{11.1.246}$$

that are the alphabetical equivalent of the azimuthal quantum number l. To summarize this in a table we have:

Main quantum number n	**Number of sublayers**	**Notation of sublayers**	**Maximum number of electrons $2n^2$**
1 (K)	1	$1s$	2
2 (L)	2	$2s2p$	8
3 (M)	3	$3s3p3d$	18
4 (N)	4	$4s4p4d4f$	32
5 (O)	5	$5s5p5d5f5g$	50
6 (P)	6	$6s6p6d6f6g6h$	72
7 (Q)	7	$7s7p7d7f7g7h$	98

Table 11.1: **Sub-levels of electrons in the main layers**

Often represented as in the figure below:

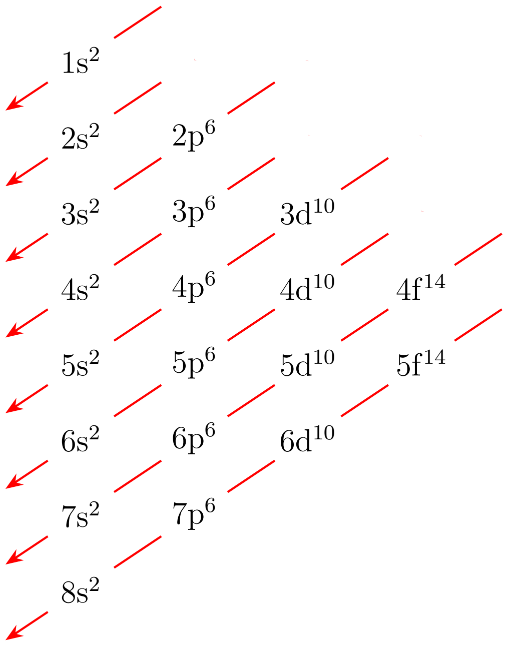


Figure 11.20: **Möller diagram electronic structure**

Definitions (#11.1.500):

D1. An "electronic layer" is a group of states that have the same principal quantum number n.

D2. A "sub-layer" or "sub-shell" is a smaller group of states which are characterized by the quantum numbers n and l.

D3. An "orbital" is specified by the three numbers quantum n, l, m_l and it can contain two electrons one with spin up and the other with spin down.

D4. A "state" is defined by the four quantum numbers n, l, m_l, m_s and contains a single electron as required by the principle of exclusion.

D5. In chemistry, a "valence electron" is an outer shell electron that is associated with an atom, and that can participate in the formation of a chemical bond if the outer shell is not

The number of electrons in an atom's outermost valence shell governs its bonding behaviour. Therefore, elements whose atoms can have the same number of valence electrons are grouped together in the periodic table of the elements. Valence electrons are also responsible for the electrical conductivity of an element; as a result, an element may be classified as a metal, a non-metal, or a semiconductor (or metalloid).

Let us summarize in the form of diagrams that we have seen so far only for the first two main layers beginning with K $(n=1)$:

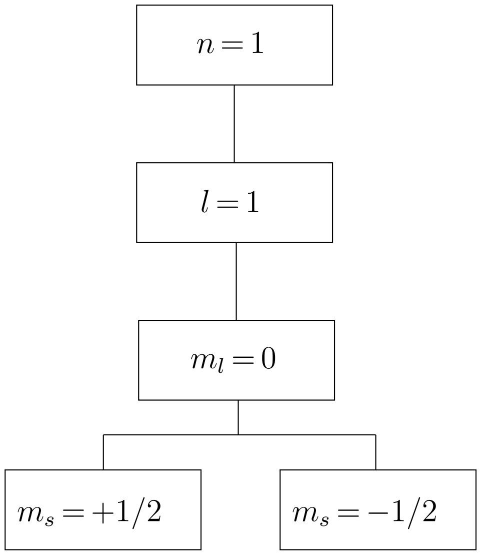


Figure 11.21: Decomposition of a low orbit with spin

Thus, the Pauli exclusion principle provides that 2 electrons in the layer K are authorized.

And for K $(n=2)$:

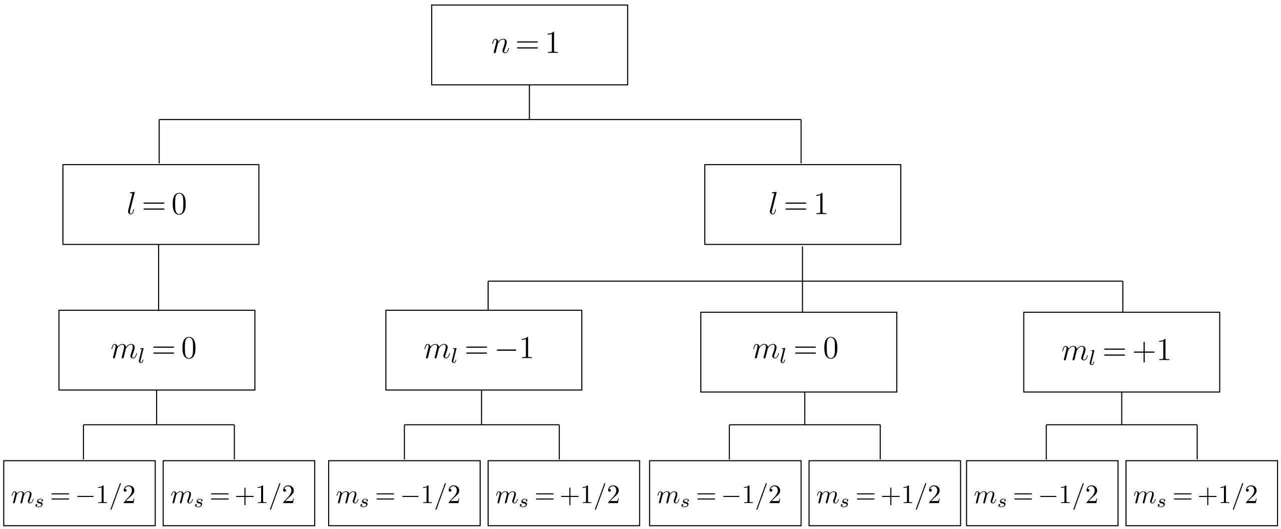


Figure 11.22: Decomposition of a less lower orbit with spin

Thus, the Pauli exclusion principle provides that 8 electrons in the layer L are authorized.

And so on ... We notice that we have in adequation with what we have proved earlier that every layer can contain a number of electrons equal to $2n^2$.

Under the form of a table this gives:

Principal quantum number	Period notation	Radial quantum number	Azimuthal quantum number	Orbital quantum number	Orbital notation	Magnetic quantum number	Spin	Maximum number of electrons in	
$n{=}n_r{+}n_\theta$	symbol	$n_r{=}n{-}1,\ldots,0$	$n_\varphi{=}k{=}1,\ldots,n$	$l{=}k{-}1$	symbol	m_l	m_s	shell	period
1	K	0	1	0	1s	0	$+,-$	2	2
2	L	1	1	0	2s	0	$+,-$	2	8
		0	2	1	2p	$-1,0+1$	$+,-$	6	
3	M	2	1	0	3s	0	$+,-$	2	18
		1	2	0	3p	$-1,0,+1$	$+,-$	6	
		0	3	1	3d	$-2,-1,0,+1,+2$	$+,-$	10	
4	N	3	1	0	4s	0	$+,-$	2	32
		2	2	1	4p	$-1,0,+1$	$+,-$	6	
		1	3	2	4d	$-2,-1,0,+1,+2$	$+,-$	10	
		0	4	3	4f	$-3,-2,-1,0,+1,+2,+3$	$+,-$	14	
5	O	4	1	0	5s	0	$+,-$	2	50
		3	2	1	5p	$-1,0,+1$	$+,-$	6	
		2	3	2	5d	$-2,-1,0,+1,+2$	$+,-$	10	
		1	4	3	5f	$-3,-2,-1,0,+1,+2,+3$	$+,-$	14	
		0	5	4	5g	$-4,-3,-2,-1,0,+1,+2,+3,+4$	$+,-$	18	

Table 11.2: Modern quantum number notation (source: Aco Z. Muradjan)

Under the form of Bohr diagram according to the naive Bohr model, this gives (the "groups" refers to columns of the periodic table of elements we will see further below):

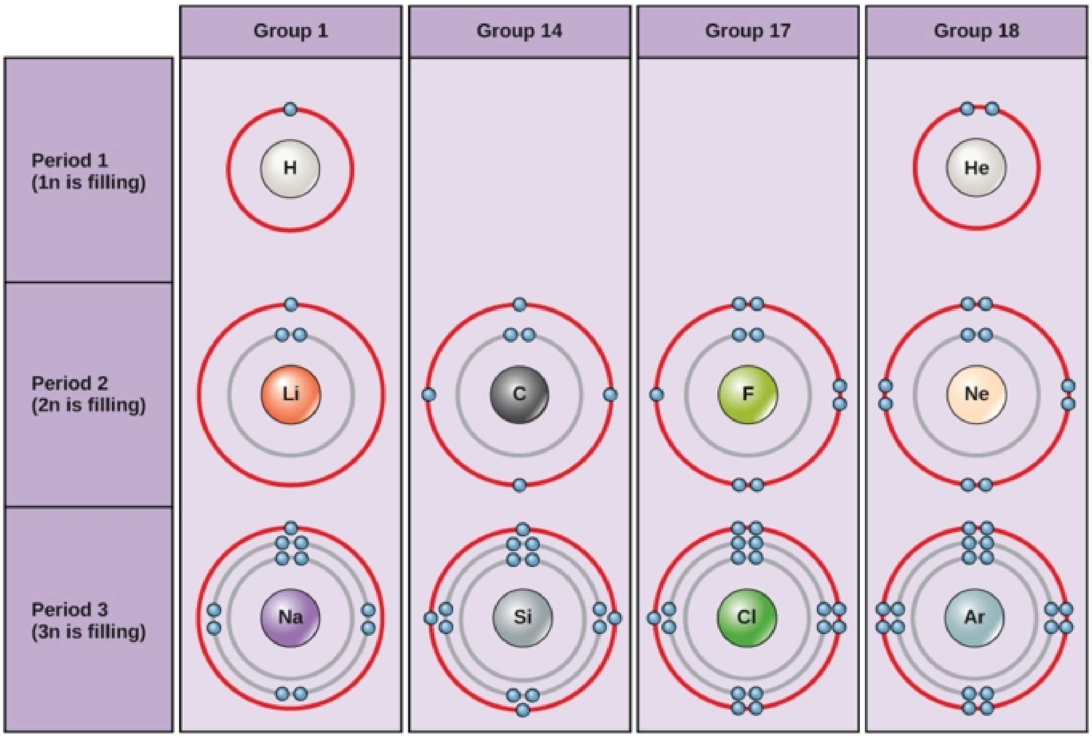

Group 1
Group 14
Group 17
Group 18
Period 1 (1n is filling)
Period 2 (2n is filling)
Period 3 (3n is filling)
H
He
Li
C
F
Ne
Na
Si
Cl
Ar

Figure 11.23: **Example of Bohr Diagram (source: ?)**

With a notation consistent with that of chemists, the basic configurations of some elements are then written:

Element	Symbol	Atomic Number	Electronic configuration
Hydrogen	H	1	$1s$
Helium	He	2	$1s^2$
Lithium	Li	3	$1s^2 2s$
Beryllium	Be	4	$1s^2 2s^2$
Boron	B	5	$1s^2 2s^2 2p$
Carbon	C	6	$1s^2 2s^2 2p^2$
Nitrogen	N	7	$1s^2 2s^2 2p^3$
Oxygen	O	8	$1s^2 2s^2 2p^4$
Fluorine	F	9	$1s^2 2s^2 2p^5$
Neon	Ne	10	$1s^2 2s^2 2p^6$
Sodium	Na	11	$1s^2 2s^2 2p^6 3s$
Magnesium	Mg	12	$1s^2 2s^2 2p^6 3s^2$
Aluminium	Al	13	$1s^2 2s^2 2p^6 3s^2 3p$
Silicon	Si	14	$1s^2 2s^2 2p^6 3s^2 3p^2$
Phosphorus	P	15	$1s^2 2s^2 2p^6 3s^2 3p^3$
Sulphur	S	16	$1s^2 2s^2 2p^6 3s^2 3p^4$
Chlorine	Cl	17	$1s^2 2s^2 2p^6 3s^2 3p^5$
Argon	Ar	18	$1s^2 2s^2 2p^6 3s^2 3p^6$
...	...	...	...

Table 11.3: **Examples of some atomic elements electronic configuration**

Which is a condensed form of the following corresponding table:

				K	**L**		**M**			**N**			
			n**:**	1	2		3			4			
Z	**Element**		l**:**	s	s	p	s	p	d	s	p	d	f
1	Hydrogen	H		1									
2	Helium	He		2									
3	Lithium	Li		2	1								
4	Beryllium	Be		2	2								
5	Boron	B		2	2	1							
6	Carbon	C		2	2	2							
7	Nitrogen	N		2	2	3							
8	Oxygen	O		2	2	4							
9	Fluorine	F		2	2	5							
10	Neon	Ne		2	2	6							
11	Sodium	Na		2	2	6	1						
12	Magnesium	Mg		2	2	6	2						
13	Aluminium	Al		2	2	6	2	1					
14	Silicon	Si		2	2	6	2	2					
15	Phosphorus	P		2	2	6	2	3					
16	Sulphur	S		2	2	6	2	4					
17	Chlorine	Cl		2	2	6	2	5					
18	Argon	Ar		2	2	6	2	6					
19	Potassium	K		2	2	6	2	6	·	1			
20	Calcium	Ca		2	2	6	2	6	·	2			
21	Scandium	Sc		2	2	6	2	6	1	2			
22	Titanium	Ti		2	2	6	2	6	2	2			
23	Vanadium	V		2	2	6	2	6	3	2			
24	Chromium	Cr		2	2	6	2	6	4	1			
25	Manganese	Mn		2	2	6	2	6	5	2			
26	Iron	Fe		2	2	6	2	6	6	2			
27	Cobalt	Co		2	2	6	2	6	7	2			
28	Nickel	Ni		2	2	6	2	6	8	2			
29	Copper	Cu		2	2	6	2	6	10	1			
30	Zinc	Zn		2	2	6	2	6	10	2			
31	Gallium	Ga		2	2	6	2	6	10	2	1		
32	Germanium	Ge		2	2	6	2	6	10	2	2		
33	Arsenic	As		2	2	6	2	6	10	2	3		
...	...			...	...		...			...			

Table 11.4: **Examples of some atomic elements electronic configuration (extended version)**

As we said earlier this is not an accurate way as the layers anywas overlap. This is why there exist another way to fill-in the layer (this other way works very well for the ground states of the atoms for the first 18 elements, then decreasingly well for the following 100 elements).

Definition (#11.1.501): The "Aufbau principle" (from the German Aufbau, "building up, construction") states that a maximum of 2 electrons are put into orbitals in the order of increasing orbital energy: the lowest-energy orbitals are filled before electrons are placed in higher-energy orbitals such that:

1. Orbitals are filled in the order of increasing $n+l$
2. Where two orbitals have the same value of $n+l$, they are filled in order of increasing n

This gives the following order for filling the orbitals:

$$1s, 2s, 2p, 3s, 3p, 4s, 3d, 4p, 5s, 4d, 5p, 6s, 4f, 5d, 6p, 7s, 5f, 6d, 7p, (8s, 5g, 6f, 7d, 8p, \text{ and } 9s) \tag{11.1.247}$$

and can be illustrated as following:

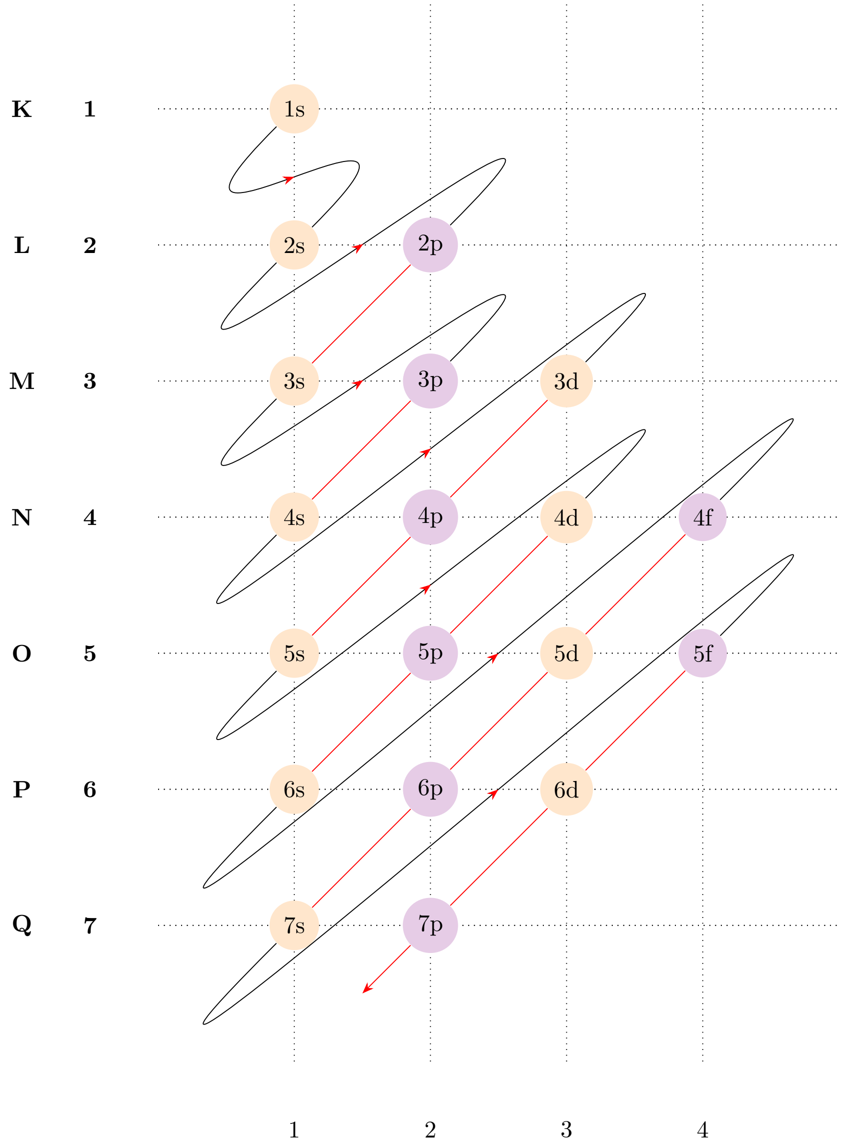


Figure 11.24: **Aufbau Principle**

In this list the orbitals in parentheses are not occupied in the ground state of the heaviest atom known so far (Uuo, $Z = 118$).

The Aufbau principle can be applied, in a modified form, to the protons and neutrons in the atomic nucleus, as in the shell model of nuclear physics and nuclear chemistry.

And finally this is the only image of a nice periodic table that we have been able to find for this textbook (there exist much more complete and technical one but their screen rendering quality is awful...):

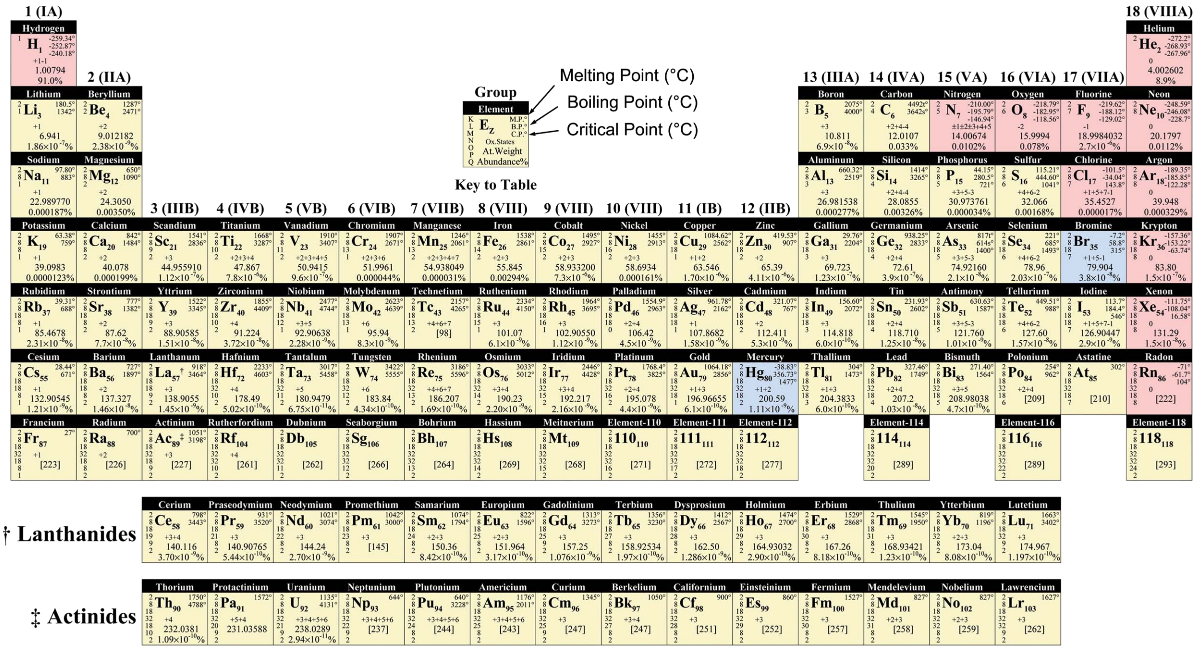
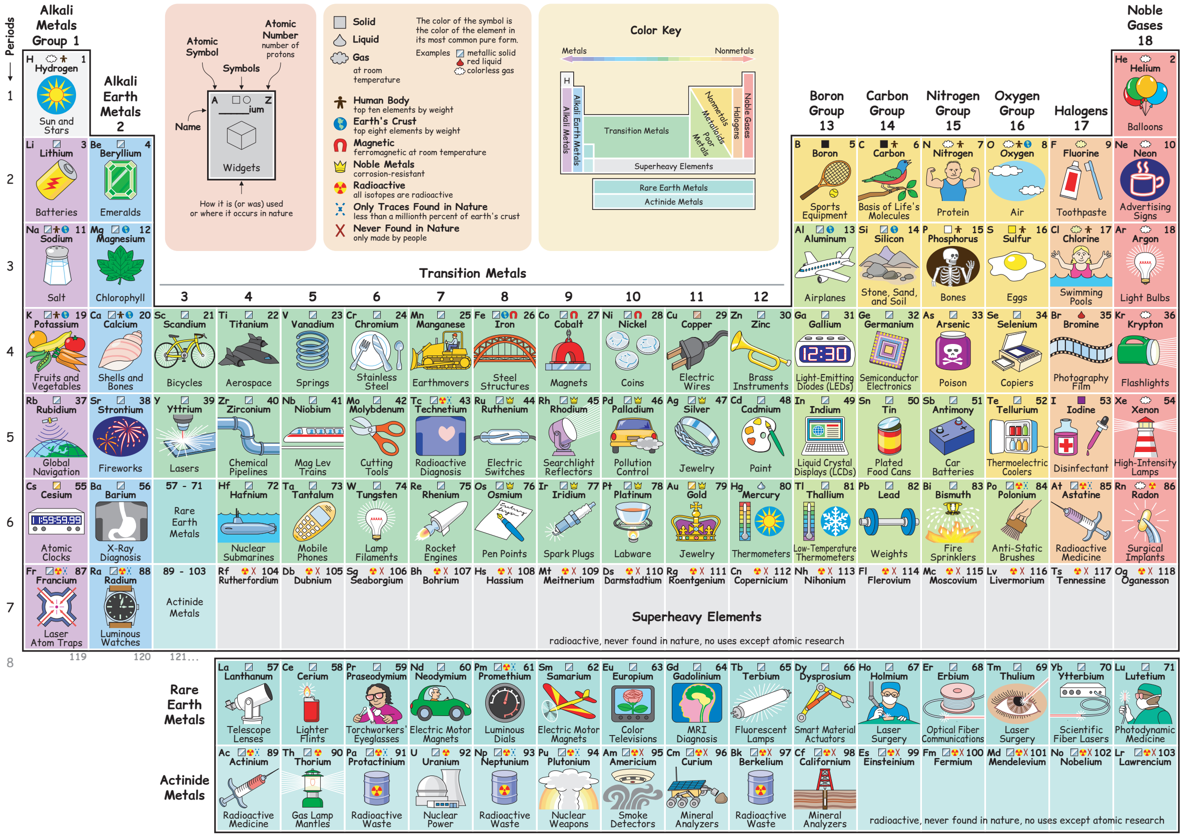

1 (IA)
18 (VIIIA)
2 (IIA)
13 (IIIA)
14 (IVA)
15 (VA)
16 (VIA)
17 (VIIA)
3 (IIIB)
4 (IVB)
5 (VB)
6 (VIB)
7 (VIIB)
8 (VIII)
9 (VIII)
10 (VIII)
11 (IB)
12 (IIB)
Group
Element
Melting Point (°C)
Boiling Point (°C)
Critical Point (°C)
Key to Table
† Lanthanides
‡ Actinides

Figure 11.25: Periodic (Mendeleev) table (source: European Synchrotron)

Or a much more pedagogical one:


Periods
Alkali Metals Group 1
Alkali Earth Metals 2
Atomic Symbol
Atomic Number number of protons
Symbols
Name
Widgets
How it is (or was) used or where it occurs in nature
Solid
Liquid
Gas at room temperature
The color of the symbol is the color of the element in its most common pure form.
Examples metallic solid red liquid colorless gas
Human Body top ten elements by weight
Earth's Crust top eight elements by weight
Magnetic ferromagnetic at room temperature
Noble Metals corrosion-resistant
Radioactive all isotopes are radioactive
Only Traces Found in Nature less than a millionth percent of earth's crust
Never Found in Nature only made by people
Color Key
Metals
Nonmetals
Alkali Metals
Alkali Earth Metals
Transition Metals
Poor Metals
Metalloids
Nonmetals
Halogens
Noble Gases
Superheavy Elements
Rare Earth Metals
Actinide Metals
Noble Gases 18
Boron Group 13
Carbon Group 14
Nitrogen Group 15
Oxygen Group 16
Halogens 17
Transition Metals
3 4 5 6 7 8 9 10 11 12
1 2 3 4 5 6 7 8
H 1 Hydrogen Sun and Stars
He 2 Helium Balloons
Li 3 Lithium Batteries
Be 4 Beryllium Emeralds
B 5 Boron Sports Equipment
C 6 Carbon Basis of Life's Molecules
N 7 Nitrogen Protein
O 8 Oxygen Air
F 9 Fluorine Toothpaste
Ne 10 Neon Advertising Signs
Na 11 Sodium Salt
Mg 12 Magnesium Chlorophyll
Al 13 Aluminum Airplanes
Si 14 Silicon Stone, Sand, and Soil
P 15 Phosphorus Bones
S 16 Sulfur Eggs
Cl 17 Chlorine Swimming Pools
Ar 18 Argon Light Bulbs
K 19 Potassium Fruits and Vegetables
Ca 20 Calcium Shells and Bones
Sc 21 Scandium Bicycles
Ti 22 Titanium Aerospace
V 23 Vanadium Springs
Cr 24 Chromium Stainless Steel
Mn 25 Manganese Earthmovers
Fe 26 Iron Steel Structures
Co 27 Cobalt Magnets
Ni 28 Nickel Coins
Cu 29 Copper Electric Wires
Zn 30 Zinc Brass Instruments
Ga 31 Gallium Light-Emitting Diodes (LEDs)
12:30
Ge 32 Germanium Semiconductor Electronics
As 33 Arsenic Poison
Se 34 Selenium Copiers
Br 35 Bromine Photography Film
Kr 36 Krypton Flashlights
Rb 37 Rubidium Global Navigation
Sr 38 Strontium Fireworks
Y 39 Yttrium Lasers
Zr 40 Zirconium Chemical Pipelines
Nb 41 Niobium Mag Lev Trains
Mo 42 Molybdenum Cutting Tools
Tc 43 Technetium Radioactive Diagnosis
Ru 44 Ruthenium Electric Switches
Rh 45 Rhodium Searchlight Reflectors
Pd 46 Palladium Pollution Control
Ag 47 Silver Jewelry
Cd 48 Cadmium Paint
In 49 Indium Liquid Crystal Displays (LCDs)
Sn 50 Tin Plated Food Cans
Sb 51 Antimony Car Batteries
Te 52 Tellurium Thermoelectric Coolers
I 53 Iodine Disinfectant
Xe 54 Xenon High-Intensity Lamps
Cs 55 Cesium Atomic Clocks
11:59:59.99
Ba 56 Barium X-Ray Diagnosis
57 - 71 Rare Earth Metals
Hf 72 Hafnium Nuclear Submarines
Ta 73 Tantalum Mobile Phones
W 74 Tungsten Lamp Filaments
Re 75 Rhenium Rocket Engines
Os 76 Osmium Pen Points
Ir 77 Iridium Spark Plugs
Pt 78 Platinum Labware
Au 79 Gold Jewelry
Hg 80 Mercury Thermometers
Tl 81 Thallium Low-Temperature Thermometers
Pb 82 Lead Weights
Bi 83 Bismuth Fire Sprinklers
Po 84 Polonium Anti-Static Brushes
At 85 Astatine Radioactive Medicine
Rn 86 Radon Surgical Implants
Fr 87 Francium Laser Atom Traps
Ra 88 Radium Luminous Watches
89 - 103 Actinide Metals
Rf 104 Rutherfordium
Db 105 Dubnium
Sg 106 Seaborgium
Bh 107 Bohrium
Hs 108 Hassium
Mt 109 Meitnerium
Ds 110 Darmstadtium
Rg 111 Roentgenium
Cn 112 Copernicium
Nh 113 Nihonium
Fl 114 Flerovium
Mc 115 Moscovium
Lv 116 Livermorium
Ts 117 Tennessine
Og 118 Oganesson
Superheavy Elements
radioactive, never found in nature, no uses except atomic research
119 120 121…
Rare Earth Metals
La 57 Lanthanum Telescope Lenses
Ce 58 Cerium Lighter Flints
Pr 59 Praseodymium Torchworkers' Eyeglasses
Nd 60 Neodymium Electric Motor Magnets
Pm 61 Promethium Luminous Dials
Sm 62 Samarium Electric Motor Magnets
Eu 63 Europium Color Televisions
Gd 64 Gadolinium MRI Diagnosis
Tb 65 Terbium Fluorescent Lamps
Dy 66 Dysprosium Smart Material Actuators
Ho 67 Holmium Laser Surgery
Er 68 Erbium Optical Fiber Communications
Tm 69 Thulium Laser Surgery
Yb 70 Ytterbium Scientific Fiber Lasers
Lu 71 Lutetium Photodynamic Medicine
Actinide Metals
Ac 89 Actinium Radioactive Medicine
Th 90 Thorium Gas Lamp Mantles
Pa 91 Protactinium Radioactive Waste
U 92 Uranium Nuclear Power
Np 93 Neptunium Radioactive Waste
Pu 94 Plutonium Nuclear Weapons
Am 95 Americium Smoke Detectors
Cm 96 Curium Mineral Analyzers
Bk 97 Berkelium Radioactive Waste
Cf 98 Californium Mineral Analyzers
Es 99 Einsteinium
Fm 100 Fermium
Md 101 Mendelevium
No 102 Nobelium
Lr 103 Lawrencium
radioactive, never found in nature, no uses except atomic research

© 2005–2016 Keith Enevoldsen elements.wlonk.com Creative Commons Attribution-ShareAlike 4.0 International License

The structure of the periodic table is closely related to the electron configuration of the atoms of the elements:

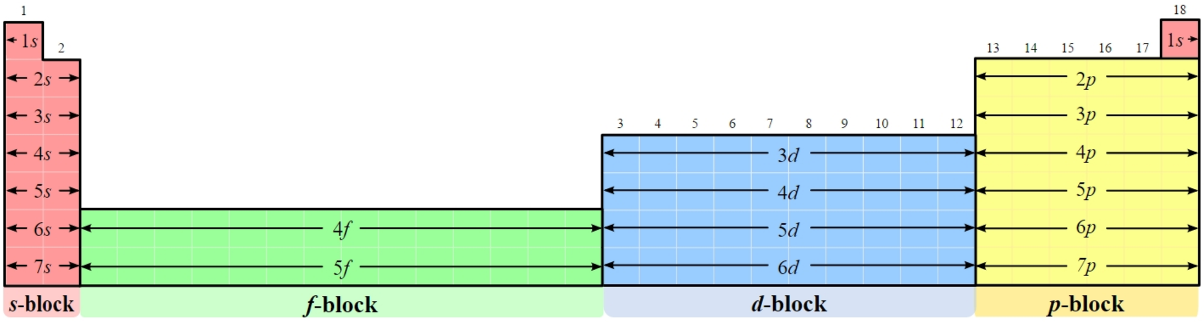


Figure 11.26: **Periodic (Mendeleev) table structure (source: Wikipedia)**

A comparison of ionic radii with atomic radii shows that a cation is always smaller than its parent neutral atom, and an anion is always larger than the parent neutral atom. When one or more electrons is removed from a neutral atom, two things happen: (1) repulsions between electrons in the same principal shell decrease because fewer electrons are present, and (2) the effective nuclear charge felt by the remaining electrons increases because there are fewer electrons to shield one another from the nucleus. Conversely, adding one or more electrons to a neutral atom causes electron–electron repulsions to increase and the effective nuclear charge to decrease, so the size of the probability region increases:

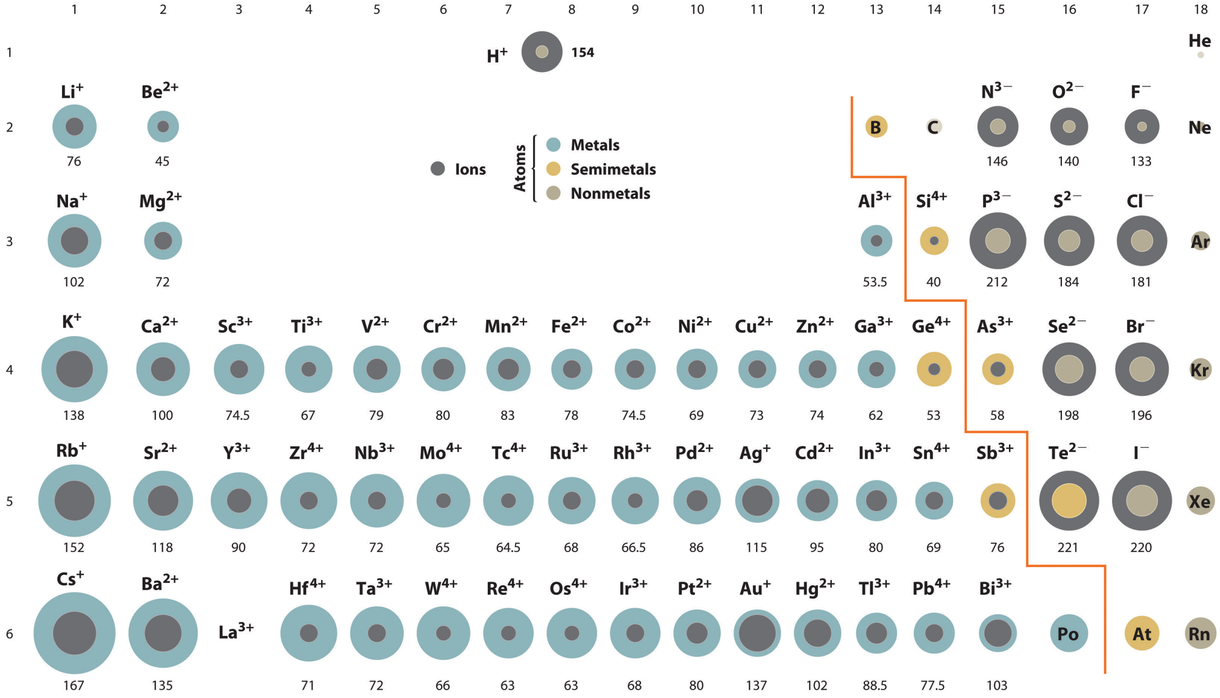


Figure 11.27: **Ionic Radii (in picometres) of the most common oxidation states of the s-, p-, and d-Block Elements**

However, although the relativistic model of Sommerfeld is an amazing accurate and consistent model with experimental observations, it does not explain some important phenomena that we observe at the scale of the atom. Thus, this model is unable to explain the disintegration of the elements, the dual (complementary) behaviour of matter between wave and particle, the annihilation of matter and antimatter, the split of spectral lines when in presence of a magnetic field and many more.

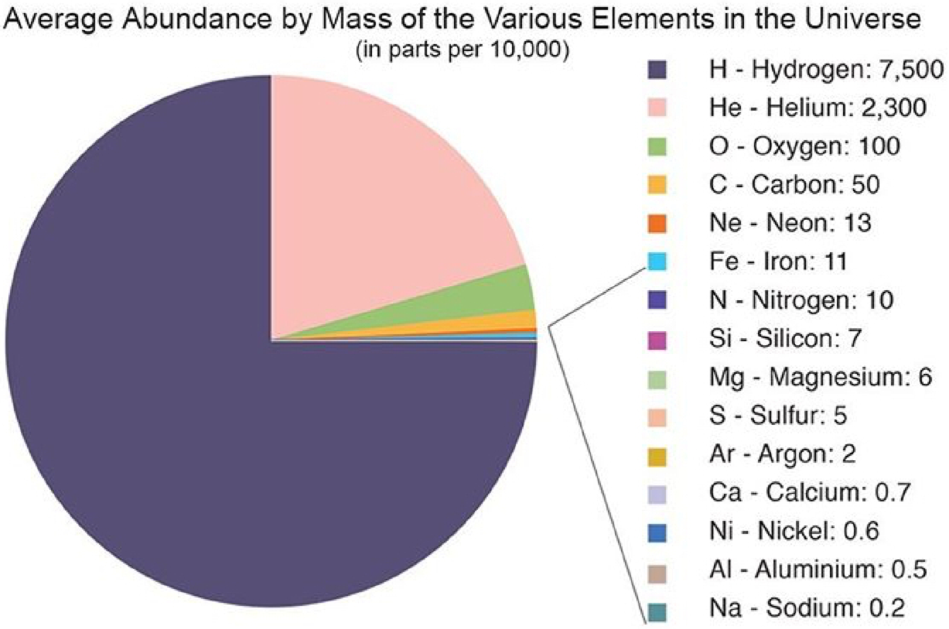
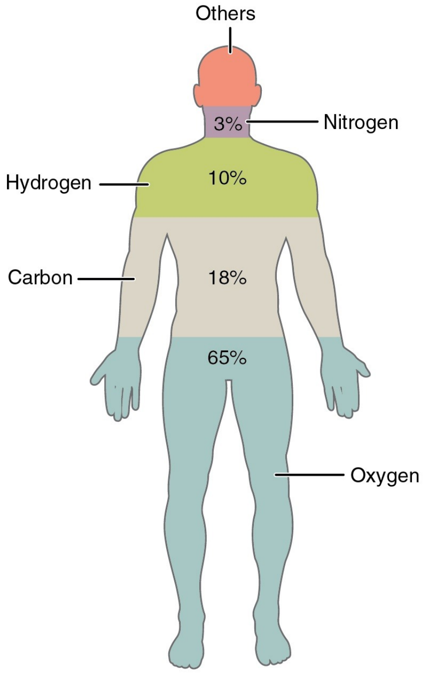


Figure 11.28: Relative abundance of elements in the observable Universe as measured in the beginning of the 121st century

The human relative abundance elements can be compared to those of various kind of Clay (as some people still believe we were made from Clay and some magic...):

Composition	Purified Clay (%)	Na-Clay (%)	Cu-Clay (%)
Na_2O	0.01	2.66	–
CuO	–	–	3.15
MgO	3.37	2.60	1.80
Al_2O_3	24.15	17.72	16.35
SiO_2	67.66	73.76	74.45
P_2O_5	0.013	0.01	0.012
SO_3	0.306	–	0.098
K_2O	1.27	1.48	1.63
CaO	0.013	0.011	0.012
TiO_2	0.177	0.175	0.353
MnO	0.098	–	–
Fe_2O_3	2.80	1.95	2.059
ZnO	0.023	-	-
Rb_2O	0.034	0.013	0.0029
SrO	0.024	-	-
Y_2O_3	0.013	-	-
ZrO_2	0.013	0.010	0.0176
Nb_2O_5	0.0122	-	0.011

Table 11.5: Relative abundance of elements if various types of Clay



Element	Symbol	Percentage in Body
Oxygen	O	65.0
Carbon	C	18.5
Hydrogen	H	9.5
Nitrogen	N	3.2
Calcium	Ca	1.5
Phosphorus	P	1.0
Potassium	K	0.4
Sulfur	S	0.3
Sodium	Na	0.2
Chlorine	Cl	0.2
Magnesium	Mg	0.1
Trace elements include boron (B), chromium (Cr), cobalt (Co), copper (Cu), fluorine (F), iodine (I), iron (Fe), manganese (Mn), molybdenum (Mo), selenium (Se), silicon (Si), tin (Sn), vanadium (V), and zinc (Zn).		less than 1.0

Figure 11.29: **Relative abundance of elements in the human body**

There are more complex developments (i.e. models) and at the same time (hopefully!) consistent with what we saw earlier which will be developed in the next section dealing with Wave Quantum Physics that give the opportunity to explain in a very satisfactory way a large number of phenomena that were unexplained at the atomic size (but sill not all as we will see after that we have studied Relativistic Quantum Physics).

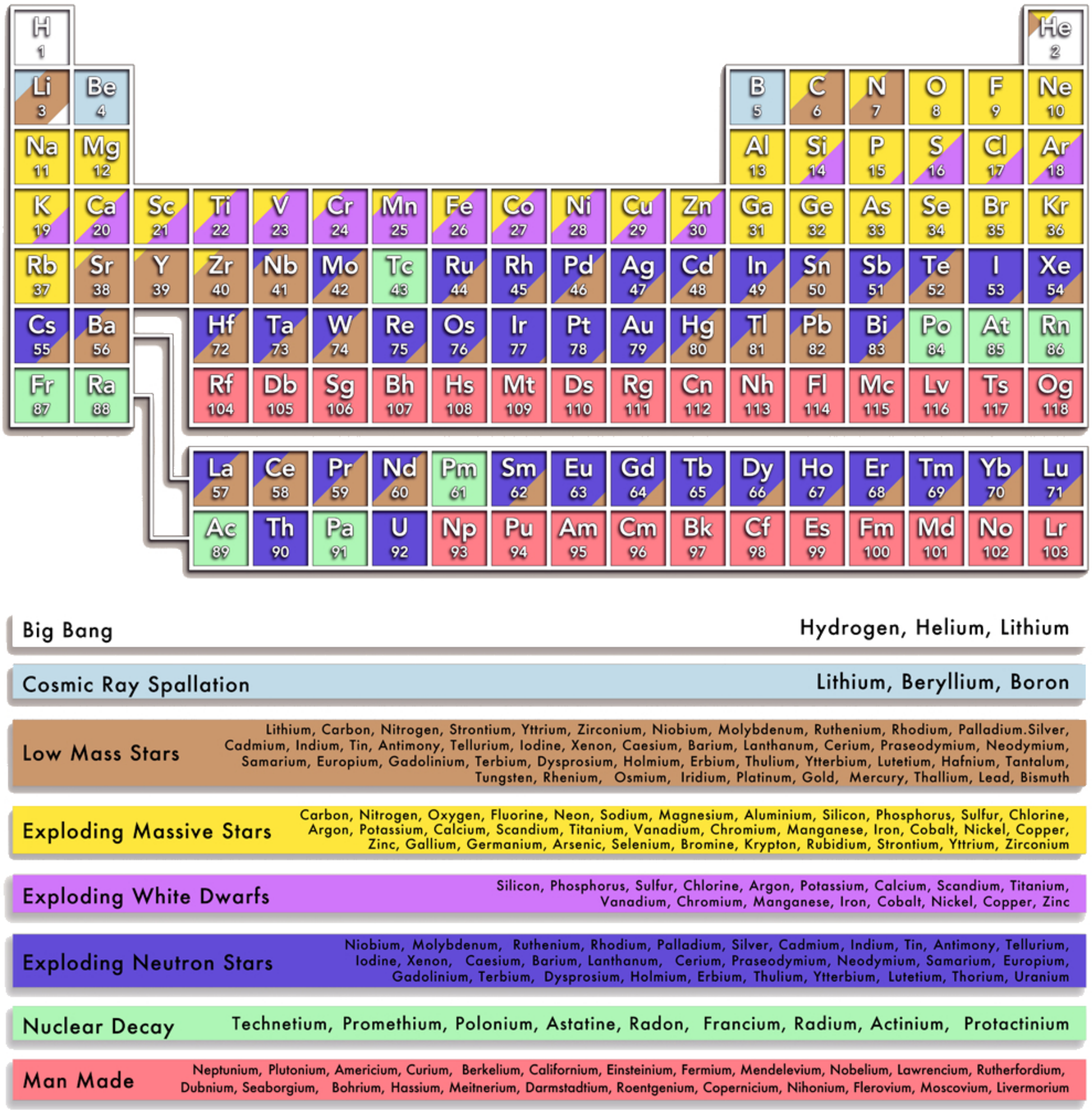

H 1
He 2
Li 3
Be 4
B 5
C 6
N 7
O 8
F 9
Ne 10
Na 11
Mg 12
Al 13
Si 14
P 15
S 16
Cl 17
Ar 18
K 19
Ca 20
Sc 21
Ti 22
V 23
Cr 24
Mn 25
Fe 26
Co 27
Ni 28
Cu 29
Zn 30
Ga 31
Ge 32
As 33
Se 34
Br 35
Kr 36
Rb 37
Sr 38
Y 39
Zr 40
Nb 41
Mo 42
Tc 43
Ru 44
Rh 45
Pd 46
Ag 47
Cd 48
In 49
Sn 50
Sb 51
Te 52
I 53
Xe 54
Cs 55
Ba 56
Hf 72
Ta 73
W 74
Re 75
Os 76
Ir 77
Pt 78
Au 79
Hg 80
Tl 81
Pb 82
Bi 83
Po 84
At 85
Rn 86
Fr 87
Ra 88
Rf 104
Db 105
Sg 106
Bh 107
Hs 108
Mt 109
Ds 110
Rg 111
Cn 112
Nh 113
Fl 114
Mc 115
Lv 116
Ts 117
Og 118
La 57
Ce 58
Pr 59
Nd 60
Pm 61
Sm 62
Eu 63
Gd 64
Tb 65
Dy 66
Ho 67
Er 68
Tm 69
Yb 70
Lu 71
Ac 89
Th 90
Pa 91
U 92
Np 93
Pu 94
Am 95
Cm 96
Bk 97
Cf 98
Es 99
Fm 100
Md 101
No 102
Lr 103
Big Bang
Hydrogen, Helium, Lithium
Cosmic Ray Spallation
Lithium, Beryllium, Boron
Low Mass Stars
Lithium, Carbon, Nitrogen, Strontium, Yttrium, Zirconium, Niobium, Molybdenum, Ruthenium, Rhodium, Palladium.Silver, Cadmium, Indium, Tin, Antimony, Tellurium, Iodine, Xenon, Caesium, Barium, Lanthanum, Cerium, Praseodymium, Neodymium, Samarium, Europium, Gadolinium, Terbium, Dysprosium, Holmium, Erbium, Thulium, Ytterbium, Lutetium, Hafnium, Tantalum, Tungsten, Rhenium, Osmium, Iridium, Platinum, Gold, Mercury, Thallium, Lead, Bismuth
Exploding Massive Stars
Carbon, Nitrogen, Oxygen, Fluorine, Neon, Sodium, Magnesium, Aluminium, Silicon, Phosphorus, Sulfur, Chlorine, Argon, Potassium, Calcium, Scandium, Titanium, Vanadium, Chromium, Manganese, Iron, Cobalt, Nickel, Copper, Zinc, Gallium, Germanium, Arsenic, Selenium, Bromine, Krypton, Rubidium, Strontium, Yttrium, Zirconium
Exploding White Dwarfs
Silicon, Phosphorus, Sulfur, Chlorine, Argon, Potassium, Calcium, Scandium, Titanium, Vanadium, Chromium, Manganese, Iron, Cobalt, Nickel, Copper, Zinc
Exploding Neutron Stars
Niobium, Molybdenum, Ruthenium, Rhodium, Palladium, Silver, Cadmium, Indium, Tin, Antimony, Tellurium, Iodine, Xenon, Caesium, Barium, Lanthanum, Cerium, Praseodymium, Neodymium, Samarium, Europium, Gadolinium, Terbium, Dysprosium, Holmium, Erbium, Thulium, Ytterbium, Lutetium, Thorium, Uranium
Nuclear Decay
Technetium, Promethium, Polonium, Astatine, Radon, Francium, Radium, Actinium, Protactinium
Man Made
Neptunium, Plutonium, Americium, Curium, Berkelium, Californium, Einsteinium, Fermium, Mendelevium, Nobelium, Lawrencium, Rutherfordium, Dubnium, Seaborgium, Bohrium, Hassium, Meitnerium, Darmstadtium, Roentgenium, Copernicium, Nihonium, Flerovium, Moscovium, Livermorium

Figure 11.30: **Origin of chemical elements (author: ?)**

~95 %

★★★★★

81 votes, 100.00%

11.2. Wave Quantum Physics

DAUGHTER of the old quantum theory (see section Corpuscular Quantum Physics page 4616), the wave quantum physics known also simply as "quantum mechanics" is the mainstay of a set of physical theories that we group under the general heading of "quantum physics".

This denomination is opposed to that of Classical Physics, this latter failing in its description of the microscopic world (atoms and particles) and in that of certain properties of electromagnetic radiation (typically see the experiences of Young slits in the section of Wave Optics at page 4509) or semiconductor (typically see the Hall Effect in the section of Electrokinetics at page 4065).

Before we start let us quote a Quantum Physics professor: « *Today we gonna start Quantum Mechanics. Now I got bad news and good news. The bad news is that's it's a subject thats kind of hard to follow intuitevely. And the good news is that nobody can follow it intuitively. Richard Feynman one of the big figures in Physics used to say that no one understand Quantum Mechanics. So in some sense the pressure is off for you guys because I don't get it, you don't get it, Feynman doesn't get it. The point is here is my goal: Right now i'm the only one to not understand Quantum Mechanics and in about 7 days all of you will be unable to understand Quantum Mechanics and you go back and spread your ignorance everywhere else.* »

Remark

The relativistic extension of (probabilistic) Wave Quantum Physics is Relativistic Quantum Physics (see section of the same name page 4919).

Quantum Physics took over and developed the idea of wave-particle duality introduced by de Broglie considering all material particles (even the atoms) not only as point particles, but as waves, with some spatial extent. These two aspects wave/corpuscle of particles ("quanton" or "wavicles"), mutually exclusive, can not be observed simultaneously as far as we know. If we see a wave property, corpuscular appearance disappears and vice versa in this this old perspective (we know today that these both concepts, particle/wave, are replaced by the the fields excitation in the framework of Quantum Field Theory as introduced at page 5158).

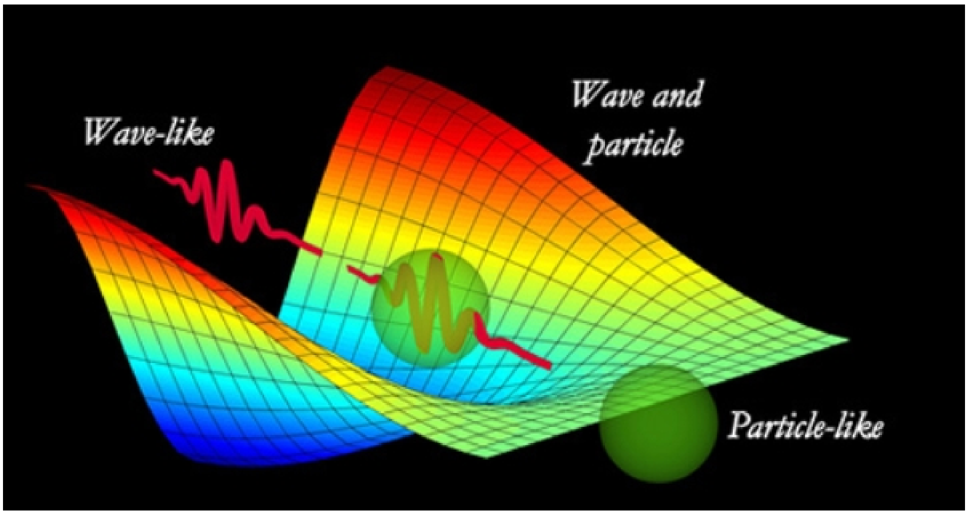


Figure 11.31: **Dual nature of light (source: S. Tanzilli, CNRS)**

To this date, no contradiction could be found between the predictions of quantum physics and associated experimental tests. This success has unfortunately a price: the theory is based on a rather abstract mathematical formalism, which makes it quite difficult at first. Hopefully until we not deal with the spin we can be dispensed Heisenberg's matrix formalism to focus on formalism of Schrödinger's wave equation that is much simpler and which to a certain level can get a mental picture of what is happening.

An even greater challenge is that this field of physics is extremely difficult—if not impossible - to introduce in a linear pedagogical manner. As a result, many textbooks on the subject (including this one), whether intended for specialists or a broader audience, often see their explanations subject to various interpretations, critiques, revisions, and additions.

To find a plausible solution it is favourable to take as basis the "principle of objectivity" due to Werner Heisenberg, which is the basis of "standard quantum physics": exists only what is experimentally observable and reproducible. We could also align with the conviction of Max Born to rebuild corpuscular quantum physics that is objectively a pot-pourri of quantum and classical physics rules to give the place to a new coherent theory based only on a few postulates (Max Born was inspired by the elegance of Albert Einstein's theories that made use of a few well-defined assumptions). Wolfgang Pauli also felt it was essential to cease to sate arbitrary ad hoc hypotheses each time experiments produced data that were in disagreement with the theory.

The duality principle is accepted by the majority of physicists, but not all. Is an electron in several places at the same time? For this to be admissible, we must have an experience that found them several places at the same time, which is not possible (problem of clock sync) then we are not required to answer the question! Say that it is in several places before we observe is not admissible in physics: principle of objectivity. In general, we'll also give up the notion of trajectory and movement, which will allow to remove the contradiction of radiation by breaking/Bremsstrahlung (see section Electrodynamics page 4163): because there is no more movement in the classical sense. The concepts of speed and acceleration lose all sense at this scale!

Remark

So quantum stuff aren't really waves and there aren't really particles. Instead, everything in the Universe (including, for all we known, the Universe itself as we will see it during our study of Quantum Cosmology) is described by a probabilistic wave function that has properties of both particles and waves. Sometimes this wave function appears more like a wave, sometimes more like a particle. But fundamentally it is neither - it's a new category in its own right. Strictly speaking, therefore, we shouldn't refer to "elementary particles" at all, which is why one of my profs suggested we instead named them "elementary things."

A minority of physicists denies this principle and founded a non-standard quantum physics with classical quantities this is why we can find especially in popular scientific papers presentations that deviate from the standard quantum physics (that of the most physicists). This non-standard version gives the same predictions for any feasible experience, so this is a possible model.

Finally quantum physics is a theory considered by the current majority of physicists as unfinished and in which many points still remain rather obscure.

Before we tackle the mathematical part, we would like to indicate to the reader that we will limit ourselves only to theoretical developments made between about 11910 and 11935 according to holocene calendar (beyond the complexity of theories requires too many pages to a general textbook as ours). Indeed, we very much think that Wave Quantum Physics, as Relativistic Quantum Physics and Nuclear Physics are not finished "products", but rather a work in progress. They have developed historically, they continue to be simplified, clarified, expanded and applied through the hard work of physicists who see these theories from different angles. While we present in this textbook all these theories in a more or less linear fashion, we attempt to provide multiple viewpoints whenever possible.

11.2.1 Postulates

Unlike most textbooks on the subject, we are pedagogically (not technically!) very unconvinced about the impact of the presentation of the postulates of quantum physics at the beginning of its study in classrooms. We allow ourselves to present our reasons (experiment done):

1. They can be deduced from simple mathematical and logical reasoning (elementary algebra and probabilities) based on the postulates of corpuscular quantum physics and the principle of complementarity and therefore can be deduce from an evolution of this latter. Even if the process is rigorously false at least it is pedagogical!

2. These assumptions are indigestible or even incomprehensible if quantum physics (its formalism and vocabulary) was not initially apprehended by a given number of exercises or a regular use.

We can then consider that the only non-provable elements theoretically (to our knowledge) that have their place as assumption would be: de Broglie complementarity principle (we'll talk about it later), the Planck-Einstein relation (already seen in a previous section) and the measurement of an observable.

Nevertheless..., in the objective of the respect of traditions (traditions are not always good things!), and especially in the purpose to respect the scientific methodology, we have chosen to present these assumptions at the beginning of this section but without overemphasize them! However, we strongly recommend the non-initiated to read these assumptions (axioms) without spending too much time trying to understand but just by thinking about them and coming back to them regularly while reading this section when necessary. Afterwards, everything will maybe be clear and light will be...

Remark

We will see practical cases, in this section itself, of quantum theory for later application in the sections of Quantum Field Theory (see page 5158) and Nuclear Physics (see page 5019). However, we advise the reader to consult at the same time the sections of Quantum Computing (see page 7263) and of Quantum Chemistry (see page 6123) and Molecular Chemistry (see page 6190) which it seems helps greatly in the understanding of some little too theoretical topics presented further below.

11.2.1.1 1st Postulate: Quantum State

The state of a classical quantum system is specified by the generalized coordinates $(q_1, q_2, \ldots)$ (see section Analytical Mechanics page 3156) and is fully described by a finite function differentiable everywhere and denoted in all generality:

$$\Psi(q_1, q_2, \ldots, t) \tag{11.2.1}$$

named "state function" or "(probabilistic) wave function", whose squared modulus (multiplication of the function by its complex conjugate) must give the probability density to find instantly the system in the configuration $(q_1, q_2, \ldots)$ at time t (if the system is time dependent):

$$P(q_1, q_2, \ldots, t) = |\Psi(q_1, q_2, \ldots, t)|^2 \tag{11.2.2}$$

that we will justify later!

The above relation is named the "Born rule" or "Born interpretation" (Max Born introduced the statistical interpretation of the quantum mechanical wave function in 11926 - holocene calendar - this interpretation, which earned him the Nobel Prize in Physics, states that the square of the magnitude $|\Psi|^2$, give the probability density of finding a particle in a specific location and was introduced in his paper *Zur Quantenmechanik der Stossvorgänge*). The Born rule is one of the key principles of quantum mechanics. There have been many attempts to derive the Born rule from the other assumptions of quantum mechanics, with so far inconclusive results...

The units of the probabilistic wave function depend on the dimensionality of the system. For a single particle in one dimension (1D), the units are typically $[\Psi] = [\mathrm{L}^{-1/2}]$ (length to the power of negative one half). In 3D, the units would be $[\Psi] = [\mathrm{L}^{-3/2}]$ and the integral over all space of the square function has therefore units of probability is therefore unit less $[-]$.

This is because the integral of $|\Psi|^2$ over all space must equal 1 (normalization condition), and $|\Psi|^2$ has units of probability per unit length $[\mathrm{L}^{-1/2}]$

Remarks

R1. The fact that we speak of "wave" instead of "particle" comes from the brilliant postulate and in fact quite logical of de Broglie that we name "complementarity principle" (which we will detail later too) and which associates with any material particle, a wave and vice versa (for reminder that's because of the results of the Young, Photoelectric and Airy disk experiments as we have already know!).

R2. The fact that we deal with probabilities and that they are proportional to the square of the wave function module comes from the Heisenberg's uncertainty principle (HUP) that we will prove further below (see 4704) later and mainly from the experience of Young slits experiment with electrons (see section Wave Optics page 4509).

Caution!!!! There is at the day we write these lines an experiment (only one... see [584]) providing quite strong evidence that we have to consider the (probabilistic) wave function ontic view, i.e. that a pure quantum state directly corresponds to physical reality (in contrast, the epistemic view treats the state as representing knowledge about some underlying physical state)!

As a corollary, the particle being necessarily located somewhere in the entire space, we have the normalization condition that the integral over all space is:

$$\int |\Psi|^2 \mathrm{d}V = 1 \tag{11.2.3}$$

to a given phase factor. In other words Ψ must be normalized, what we traditionally name the "de Broglie normalization condition" (even if a posteriori the concept seems to come from Max Born).

Indeed the Born rule was formulated by Born in a 11926 (holocene calendar) paper. In this paper, Born solves the Schrödinger equation for a scattering problem and, inspired by Einstein's work on the photoelectric effect, concluded, in a footnote, that the Born rule gives the only possible interpretation of the solution. In 11954 (holocene calendar), together with Walther Bothe, Born was awarded the Nobel Prize in Physics for this and other work.

Remarks

R1. Note that even normalized, Ψ is determined to a given phase factor. In addition, it is preferable for Ψ to be differentiable, because differential operators act on it to get the theoretical predictions on measurable properties, and also finished to be normalizable...

R2. When the integral given above provides a finite amount, we say it is "square integrable". Otherwise, you have to normalize it so that the theoretical model corresponds to reality! We will also discussed it in more details later (with proofs!).

Let us recall that a "phase factor" is a complex constant factor having a unitary module. We can write it (depending on what we have studied in the section on Numbers during our study of complex numbers) $e^{i\delta}$, where δ is any angle, named the "phase" (see section Wave Mechanics page 251). We will also prove further below rigorously why it has no influence!

We can express this postulate in a little more formal way, because as we will see it in several examples, the wave function is often a complex polynomial which can then be expressed in the Hilbert space of polynomials. This gives therefore in the language of Dirac bra-ket formalism (see further below for details) the following definition:

Definition (#11.2.502): The state vector "ket" represented by $|\Psi(t)\rangle$ belonging to the vector space $\mathcal{H}$ (Hilbert space) defines the state of the quantum system at time t. This state vector has all the mathematical properties required by quantum physics and especially the dot product of the vector $|\Psi(t)\rangle$ by its dual vector (complex conjugate) "bra" $\langle\Psi(t)|$ must satisfy the functional scalar product:

$$\boxed{\langle\Psi(t)|\Psi(t)\rangle = 1} \tag{11.2.4}$$

Remark

The bra-ket notation was introduced by Paul Adrien Maurice Dirac to facilitate (at least it's supposed to...) the notation of quantum physics equations, but also to highlight the potential vector aspect of the object representing a quantum state. What is specific to quantum physics is that the vectors are not drawn with arrows but with ket and bra (not the bra you can think about...), but it is only a matter of notations and brings nothing mathematically new. Moreover, you should not imagine that we wrote explicitly in the calculations these vectors as column vectors (think to complex numbers.... it is rare that we write them in vector form in fact)!

To summarize these last paragraphs, the two relations:

$$\langle\Psi(t)|\Psi(t)\rangle = 1 \tag{11.2.5}$$

and:

$$\int |\Psi|^2 \mathrm{d}V = 1 \tag{11.2.6}$$

are therefore equivalent!

A generalization of the Born rule also exist (still as a postulate!). Indeed, if $|\varphi\rangle$ is a function (or vector) representing a state of a system and if $|\chi\rangle$ represents another state of this same system, then there exists and amplitude of probability $A(\varphi \rightarrow \chi)$ to find $|\varphi\rangle$ in the state $|\chi\rangle$, which is given by the dot product on $\mathcal{H}$:

$$A(\varphi \rightarrow \chi) = \langle\chi|\varphi\rangle \tag{11.2.7}$$

The probability $P(\varphi \rightarrow \chi)$ for the state $|\varphi\rangle$ to pass the test $|\chi\rangle$ is then obtained by taking the square module of this amplitude:

$$\boxed{P(\varphi \rightarrow \chi) = \frac{|\langle\chi|\varphi\rangle|^2}{|||\chi\rangle||^2|||\varphi\rangle||^2}} \tag{11.2.8}$$

That is the probability equivalent to the quantum system that is the normalized projection P_χ of the original system ket onto the ket (or kets) corresponding to the result of the measurement, i.e.:

$$|\psi'\rangle = \frac{P_\chi|\psi\rangle}{\sqrt{\langle\psi|P_\chi|\psi\rangle}} = \frac{|\chi\rangle\,\langle\chi\,|\,\psi\rangle}{\sqrt{\langle\psi\,|\,\chi\rangle\,\langle\chi\,|\,\psi\rangle}} \qquad (11.2.9)$$

and this is related to the 4th postulate of the measure of a property that we will introduce just further below.

11.2.1.2 2nd postulate: Time evolution of a quantum state

If the system is not disturbed, the evolution (supposed non-relativistic!) of its state is governed by the Schrödinger's equation of evolution (so time dependent!):

$$\mathrm{i}\hbar\frac{\partial H}{\partial t} = H\Psi \qquad (11.2.10)$$

also denoted in ket-bra notation:

$$\mathrm{i}\hbar\frac{d}{dt}|\psi(t)\rangle = H(t)|\psi(t)\rangle \qquad (11.2.11)$$

This relation simply means that it is the operator "total energy" or "Hamiltonian" H of the system, which is responsible for the system to evolve over time. Indeed, the form of the equation shows that, in applying the Hamiltonian to the wave function Ψ of the system, we obtain the derivative with respect to time, that is to say: how it varies over time!

The term on the left side represents the time evolution of the wave function. The units of the imaginary unit i are dimensionless. The reduced Planck constant $\hbar$ has units of $[\mathrm{J}\cdot\mathrm{s}]$. The time derivative $\partial/\partial t$ has units of $[\mathrm{s}^{-1}]$. The wave function in the the 3D case has unit of $[\mathrm{m}^{-3/2}]$. Therefore, the entire left side and right side have units of $[\mathrm{J}\cdot\mathrm{m}^{-3/2}]$.

Caution!!!! A state vector and a wave function are not the same object, even though in many simple quantum-mechanics courses they appear to be interchangeable. A state vector is the abstract, basis-independent description of a quantum state; a wave function is just its coordinate representation in a chosen basis (often the position basis). This distinction matters because quantum states can be represented in many different ways, and the wave function is only one of them.

Remark

We will prove further below this relation in details (it will not be trivial unfortunately... but it is possible (!) and so it eliminates the need to define it as a postulate).

In the latter relation, H is the Hamiltonian operator (total energy) of the system that we will prove further below as having for value in a particular and simple case:

$$H = E_c + E_p = -\frac{\hbar^2}{2m}\cdot\vec{\nabla} + V(q_1, q_2, \ldots, t) = \sum_{i=1}^{n}\left(-\frac{\hbar^2}{2m}\frac{\partial}{\partial q_k^2}\right) + V(q_1, q_2, \ldots, t) \qquad (11.2.12)$$

In the case where the potential $\vec{V}(q_1, q_2, \ldots, q_n)$ is independent of time (corresponding to a conservative system in classical mechanics), there are (we will see it in various examples later) a set of independent particular solutions time independent and satisfying (relation which we will prove the origin later):

$$H\psi_k(q_1, q_2, \ldots) = E_k\psi_k(q_1, q_2, \ldots) \tag{11.2.13}$$

where the $\psi_k(q_1, q_2, \ldots)$ are the "eigenfunctions" (in analogy with the eigenvectors seen in the section Linear Algebra at page 1997) of the Hamiltonian / operator H with eigenvalue / observable E_k.

These particular solutions then describe special states named "stationary states" (as independent of time...), which we will show later the properties and origin of the name, and which form an orthogonal basis.

The previous eigenvalue equation is often named "time independent Schrödinger equation". It defines the stationary states and has a sense of course only that if the system is conservative.

This is especially the time independent Schrödinger equation that concerns quantum chemistry and molecular chemistry (topics we cover in the Chemistry chapter of the textbook). Indeed, we seek in these fields to get the wave functions describing the stationary states, and especially the state of lowest energy, "fundamental state" of atoms and molecules. The transitions observed in spectroscopy are done between these stationary states (as we will prove it further below), their determination is thus a prerequisite for the study of spectroscopy. However, we must remember that it is the Schrödinger evolution equation, which is (at first ...) the fundamental equation of non-relativistic wave quantum physics: it plays the same role as the Newton equation in Classical Mechanics, or that of an equation of motion (see the proof of the Ehrenfest's theorem further below).

Remark

In fact, we will prove (see section Relativistic Quantum Physics page 4921) that the Schrödinger evolution equation is only a special case of what we name the "free Klein-Gordon equation" which itself is a special case of the "generalized Klein-Gordon equation", which itself is a limited model relatively to the "linearised Dirac equation" ... in short we have not finished to do maths....

11.2.1.3 3rd postulate: Observables and operators

For each measurable physical property (an "observable") of a system denoted for example by:

$$\mathcal{O}(q_k, p_k) \tag{11.2.14}$$

where q_k are the generalized coordinates and p_k the generalized momentum according to notations adopted in the section of Analytical Mechanics, corresponds a linear operator (so it can also be a matrix!) named "Hermitian operator", frequently denoted with a hat such that for the chosen example above it will be denoted by:

$$\hat{\mathcal{O}}(q_k, p_k) \tag{11.2.15}$$

which always intervene in the theoretical calculation of a physically measurable property (see the section Linear Algebra page 1953 for a recall of what is an Hermitian matrix).

In other words (but this is more related to the postulates that will follow): An observable value is represented by a quantum operator which operates on the wave function to predict the measurement value of the latter observable (the operator $\hat{\mathcal{O}}$ fix the mathematical representation of $\mathcal{O}$).

To make ... a Hermitian operator in quantum physics is a mathematical expression as such that if we take its complex conjugate (or its adjoint matrix if the mathematical expression is a matrix!) then the theoretical calculation of the measurable value is always given by the same expression.

Here are the best known examples for which we will prove in details the origin in this section and of this of Relativistic Quantum Physics:

✍**Examples:**

E1. Coordinates:

$$q_k \rightarrow \hat{q}_k \tag{11.2.16}$$

for which we will see a practical example with the Ehrenfest theorems in this section.

E2. Momentum:

$$p_k \rightarrow \hat{p}_k \equiv -\mathrm{i}\hbar\partial_k \tag{11.2.17}$$

for which we will also see several practical examples (including one with the Ehrenfest theorems).

These two examples are sometimes referred to under the name "correspondence principle".

E3. Kinetic momentum:

$$b_z \rightarrow \hat{b}_z \equiv x\hat{p}_y - \hat{p}_x y \tag{11.2.18}$$

E4. Pauli matrices (that as we will see later correspond to the spin operators):

$$\sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \sigma_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \sigma_2 = \begin{bmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{bmatrix}, \sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \tag{11.2.19}$$

E5. The operator of evolution of energy of a quantum state:

$$E \rightarrow \hat{E} \equiv \mathrm{i}\hbar\partial_t \tag{11.2.20}$$

Remarks

R1. This may seem to fall from the sky ... but we will see that it just comes naturally when we will later do the development of some very concrete examples or when reading the section of Quantum Computing.

R2. As part of this textbook, we write indifferently, the operators and observable without the hat symbol (that's the reader to know what we are working with... without being confused...).

We will see also that some operators are not commutative and obey to what we name "anticommutation relations" (which are the source of the Heisenberg' principles of uncertainty).

✍**Example:**

Here is an example of such a relation that we will prove later:

$$[q_k,p_l]=[q_k,p_l]_-=q_kp_l-p_lq_k=\mathrm{i}\hbar\delta_{kl} \tag{11.2.21}$$

then we say that the two components of the two observable position and momentum are "canonically conjugate".

We will also see using a practical case that two observables A, B whose respective operators commute such as:

$$[\hat{A},\hat{B}]=0 \tag{11.2.22}$$

possess a common eigenvectors base. We then say that they are simultaneously measurable (the precise determination of one does not preclude the other's) with accuracy (otherwise we have an uncertainty... of Heisenberg). The two quantities A, B can then be named "compatible observable C.O." This then means that the two operators should have the same specific eigenfunctions (so they are compatible only if the associated operators admit such common functions).

The set of all attached C.O. to a physical system is a "complete set of compatible observables C.S.C.O.".

11.2.1.4 4th postulate: Measure of a property (collapse postulate)

The consequence of the previous postulate is that the measurement of $\mathcal{O}$ therefore always gives an eigenvalue of the associated Hermitian operator $\hat{\mathcal{O}}$. In other words, the only observable values of the property $\mathcal{O}$ are the eigenvalues (denoted for example: o) of the operator $\hat{\mathcal{O}}$!

$$\underbrace{\hat{\mathcal{O}}}_{\text{operator}}\cdot\underbrace{\phi_k}_{k\text{th eigenfunction}}=\underbrace{o_k}_{k\text{th eigenvalue}}\cdot\underbrace{\phi_k}_{k\text{th eigenfunction}} \tag{11.2.23}$$

The eigenvectors and eigenvalues of an operator values have special meaning: the eigenvalues are the values that can result from an ideal measure of this property, the eigenvectors are the quantum states of the system during the measurement.

It is because of this assumption that it is important to ensure that any physical property is represented by a Hermitian operator.

Theorem 11.1. In other words, the hermiticity of $\hat{\mathcal{O}}$ ensures that its eigenvalues (so denoted by example: o) are real.

Let's us first do the proof with the usual algebraic notation and then with another approach using the Dirac notation.

Proof 11.1.1. Since an operator is associated with a specific eigenfunction and an eigenvalue by (see section Linear Algebra page 1997):

$$\hat{\mathcal{O}}\varphi_k = o_k\varphi_k \tag{11.2.24}$$

Using the property proved during in our study of complex numbers in the section Numbers like that the complex conjugate product of two complex numbers is the product of two conjugate numbers we have:

$$\overline{\hat{\mathcal{O}}}\bar{\varphi}_k = \bar{o}_k\bar{\varphi}_k \tag{11.2.25}$$

If we multiply the first relation by the complex conjugate of φ_k and integrate over all space, we have:

$$\int \bar{\varphi}_k\hat{\mathcal{O}}\varphi_k \mathrm{d}V = o_k\int \bar{\varphi}_k\varphi_k \mathrm{d}V \tag{11.2.26}$$

and same with the prior previous relation:

$$\int \varphi_k\hat{\mathcal{O}}\bar{\varphi}_k \mathrm{d}V = \bar{o}_k\int \varphi_k\bar{\varphi}_k \mathrm{d}V \tag{11.2.27}$$

As by definition $\hat{\mathcal{O}}$ is Hermitian (therefore equal to its own conjugate complex), both left sides of the previous two relations are equal. Then we have:

$$o_k\int \bar{\varphi}_k\varphi_k \mathrm{d}V = \bar{o}_k\int \varphi_k\bar{\varphi}_k \mathrm{d}V \tag{11.2.28}$$

Therefore it comes:

$$(o_k-\bar{o}_k)\int \bar{\varphi}_k\varphi_k \mathrm{d}V = 0 \tag{11.2.29}$$

and as the integral is physically of square integrable and therefore not zero, this requires that:

$$o_k-\bar{o}_k = 0 \tag{11.2.30}$$

and therefore:

$$o_k = \bar{o}_k \tag{11.2.31}$$

and this is only possible if the eigenvalue is real.

Now, with the Dirac notation and with an approach a little bit different:

Since an operator is associated with a specific eigenfunction and eigenvalue by (see section Linear Algebra page 1997):

$$\hat{\mathcal{O}}|\varphi_k\rangle = o_k|\varphi_k\rangle \tag{11.2.32}$$

We then have if the eigenfunction is normalized in Dirac notation:

$$o_k = \langle\varphi_k|\hat{\mathcal{O}}|\varphi_k\rangle \tag{11.2.33}$$

And if the operator $\hat{\mathcal{O}}$ is a Hermitian operator (therefore equal to its complex conjugate in the case of a function and equal to the transposed complex in the case of a matrix), we have:

$$\bar{o}_k = \overline{\langle\varphi_k|\hat{\mathcal{O}}|\varphi_k\rangle} = \langle\varphi_k|\hat{\mathcal{O}}^\dagger|\varphi_k\rangle = \langle\varphi_k|\hat{\mathcal{O}}|\varphi_k\rangle = o_k \tag{11.2.34}$$

So if the operator is indeed Hermitian we have:

$$\bar{o}_k = o_k \tag{11.2.35}$$

and this can only be satisfy if the eigenvalues are real.

■ Q.E.D.

Thus, own hermitic operator values are always real numbers (hopefully ...).

Now we can introduce the final form of the postulate: After a measurement of A that yields the result a_n, the quantum system is in a new state that is the normalized projection P_n of the original system ket onto the ket (or kets) corresponding to the result of the measurement:

$$|\psi'\rangle = \frac{P_n|\psi\rangle}{\sqrt{\langle\psi|P_n|\psi\rangle}} = \frac{|a_n\rangle\langle a_n\,|\,\psi\rangle}{\sqrt{\langle\psi\,|\,a_n\rangle\langle a_n\,|\,\psi\rangle}} = \frac{\langle a_n\,|\,\psi\rangle}{|\langle a_n\,|\,\psi\rangle|}\,|a_n\rangle = e^{i\varphi}\,|a_n\rangle \tag{11.2.36}$$

and hence:

$$\langle\psi'\,|\,\psi'\rangle = \langle a_n|\,e^{-i\varphi}e^{i\varphi}\,|a_n\rangle = \langle a_n\,|\,a_n\rangle = 1 \tag{11.2.37}$$

11.2.1.5 5th postulate: Average of a property

This postulate is the less intuitive and more difficult to prove (we will prove it with an example when studying the Ehrenfest theorems at page 4870). Its statement is as follows:

The mean value (expected) of a physical property $\mathcal{O}$, when the system is in the state described by the normalized function Ψ is given by (do not confuse the notation of the average of an operator with that of the complex conjugate of an eigenvalue !!!):

$$\boxed{\langle\mathcal{O}\rangle = \bar{\mathcal{O}} = \int\bar{\Psi}\hat{\mathcal{O}}\Psi\mathrm{d}V = \langle\Psi|\hat{\mathcal{O}}|\Psi\rangle} \tag{11.2.38}$$

An equivalent expression quite difficult to read is the following: the probability of finding the eigenvalue o_k (of the operator $\hat{\mathcal{O}}$ predominantly Hermitian), when measuring the property $\mathcal{O}$

at time t on a quantum system prepared in state described by the function Ψ is given by the square of module of the projection of the function Ψ on the eigenfunction φ_k associated to the eigenvalue o_k (and its operator):

$$P(o_k,t|\Psi)=|\langle\varphi_k|\Psi\rangle|^2 \tag{11.2.39}$$

where the "projection" (or "representative") is defined by:

$$\langle\varphi_k|\Psi\rangle=\int\bar{\varphi}_k\Psi\mathrm{d}V \tag{11.2.40}$$

the index k being here to indicate that there may be for a given quantum system multiple eigenvalues with their corresponding eigenvectors.

The point of eigenstates and the entire linear algebra of quantum mechanics is that the projections $\langle\phi\,|\,\psi\rangle$ of the state $|\psi\rangle$ onto each eigenstate $|\phi\rangle$ represent the probability amplitudes of each eigenstate. In particular, this means:

$$\sum|\langle\phi\,|\,\psi\rangle|^2=1=|\langle\psi\,|\,\psi\rangle|^2 \tag{11.2.41}$$

Where the summation is taken over all the eigenstates of an operator. As this must be true for all states $|\psi\rangle$, the thing on the left must be a Pythagoran sum, so the $|\phi\rangle$ s must form an orthogonal basis. Alternatively, one may just note that we must have $\langle\phi_1\,|\,\phi_2\rangle=0$ if the eigenvalues corresponding are distinct, as two distinct observations must be mutually exclusive.

Remark

We will come back on this formalism and relations further below. They are also several practical examples provided in the section of Quantum Computing and a single and very nice example at the end of the section of Quantum Chemistry (for the mean radius and the mean angular momentum).

For example, in one dimension and for a system that is time dependent, we use the operators presented in the previous postulate (but on which we will also return in detail):

$$\begin{aligned}\langle x\rangle&=\int\bar{\Psi}\hat{x}\Psi\mathrm{d}x=\int\bar{\Psi}x\Psi\mathrm{d}x=\int x\bar{\Psi}\Psi\mathrm{d}x\\ \langle p\rangle&=\int\bar{\Psi}\hat{p}\Psi\mathrm{d}x=\int\bar{\Psi}(-\mathrm{i}\hbar\partial_x)\Psi\mathrm{d}x=-\mathrm{i}\hbar\int x\bar{\Psi}\frac{\partial\Psi}{\partial x}\mathrm{d}x\\ \langle E\rangle&=\int\bar{\Psi}\hat{E}\Psi\mathrm{d}x=\int\bar{\Psi}(\mathrm{i}\hbar\partial_t)\Psi\mathrm{d}x=\mathrm{i}\hbar\int x\bar{\Psi}\frac{\partial\Psi}{\partial t}\mathrm{d}x\end{aligned} \tag{11.2.42}$$

For which we will see practical examples in the sections already mentioned!

Notice that therefore quantum mechanics is also an interesting example of an experimentally verifiable theory that doesn't require all mathematical structures to be directly observables. Indeed, wave functions are not themselves measurable (at least today!); what is measurable is merely the probability distribution that is derived from the wave function! Now, not everybody is happy with this state of affairs. But however quantum mechanics is highly successful regardless. So there isn't a priori anything unscientific about a theory that contains unobservable elements at the moment that the predictions can be reproduced and are consistent (i.e. lead to the same conclusions) around the World.

Remark

Some physicists have tried to rule out the imaginary number i that appears in various equations of Quantum Physics. Even if its possible in the majority of situations it can't be done for some special cases as proven theoretically in [97] and this may be experimentally refutable as some further developments show that complex quantum theory should outperforms statistically real quantum theory (see [572]).

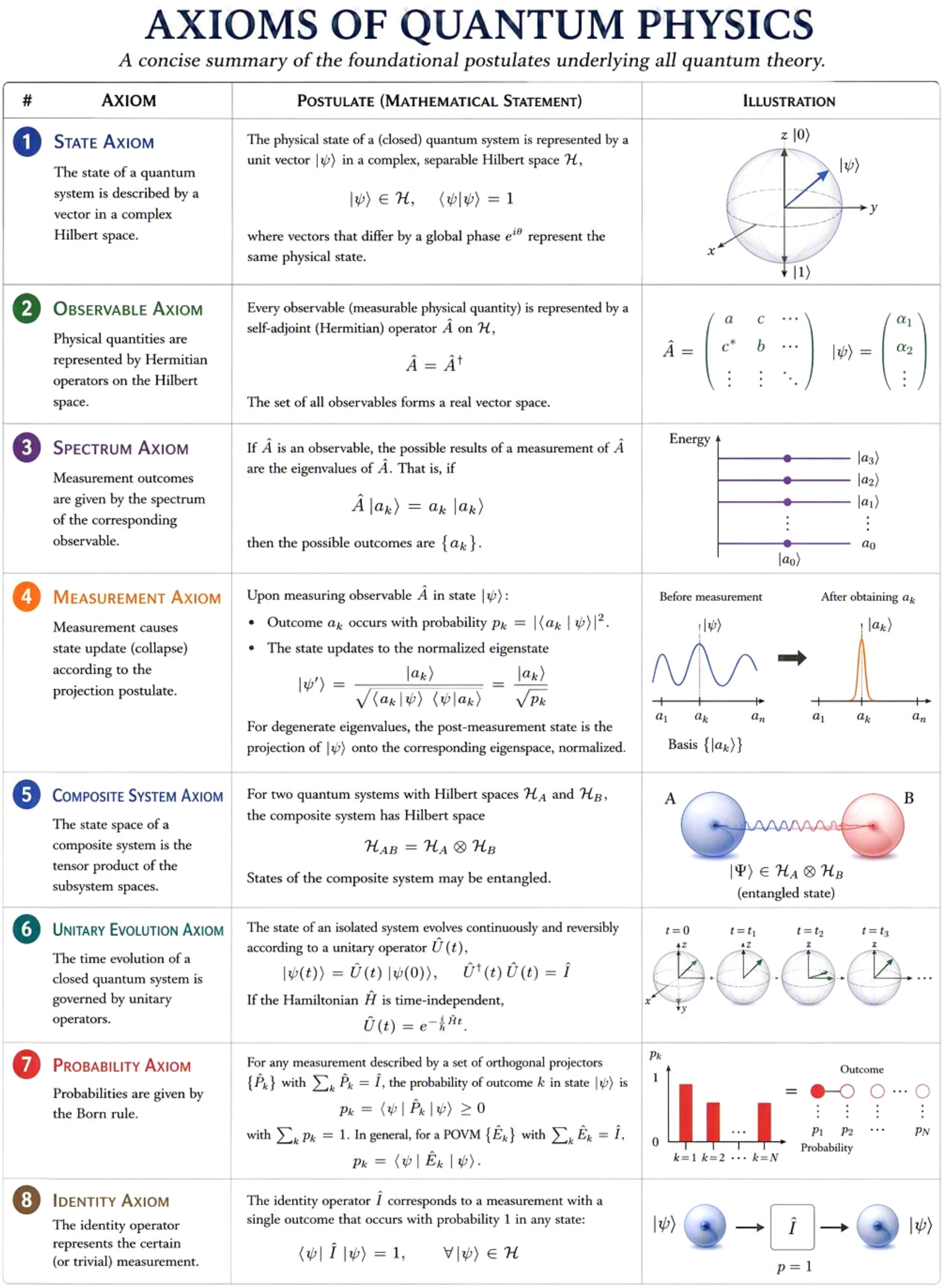
Notice that some textbooks or lectures merge together some of the postulates of the table below and others introduce the Pauli exclusion principle as a postulate (for pedagogical reasons to undergraduate students) even if the latter can be proved with the Spin-Statistics theorem. Also some textbooks and lectures add the "expansion theorem" as a postulate (the eigenfunctions of a linear and hermitian operator form a complete basis set) even if the latter is a property of Hilbert spaces (this is subject to debates...). So a summary of the most common postulates can be given by (the order of the postulates are important for some people so we tried to respect that point):

Number	Postulate	Formalism
P1	**The state vector:** An isolated physical system is represented, at a given time t, by a state vector $\vert\Psi\rangle$ belonging to a Hilbert state space $\mathcal{H}$	$\vert\Psi(\vec{r},t)\rangle \in \mathcal{H}$
P2	**Observable and operators:** Every measurable physical quantity O_k corresponds to a mathematical linear hermitian operator $\hat{O}$	$O_k \rightarrow \hat{O}$ (e.g. $\hat{p} \rightarrow -\mathrm{i}\hbar\nabla$)
P3	**Quantized outcomes:** Every measurable physical quantity $\mathcal{A}$ is described by a Hermitian operator $\mathcal{O}$ acting in the state space $\mathcal{H}$. This operator is an observable, meaning that its eigenvectors form a basis for $\mathcal{H}$. The result of measuring a physical quantity $\mathcal{A}$ must be one of the eigenvalues O_k of the corresponding observable	$\underbrace{\hat{O}}_{\text{operator}} \cdot \underbrace{\psi_k}_{\text{eigenfunction}} = \underbrace{O_k}_{\text{eigenvalue}} \cdot \underbrace{\psi_k}_{\text{eigenfunction}}$
P4	**Collapse rule:** If the measurement of a physical quantity $\mathcal{A}$ of a system in the state $\vert\psi\rangle$ gives the result a_n, then the state of the system immediately after the measurement is the normalized projection of $\vert\psi\rangle$ onto the eigensubspace associated with a_n	$\vert\psi'\rangle = \dfrac{P_n\vert\psi\rangle}{\sqrt{\langle\psi\vert P_n\vert\psi\rangle}} = \dfrac{\vert a_n\rangle\langle a_n\mid\psi\rangle}{\sqrt{\langle\psi\mid a_n\rangle\langle a_n\mid\psi\rangle}}$

P5	**Average of a property:** The expectation value of any property can be found by performing the expectation value integral with respect to its associated wavefunction, a special case being the Born rule when $\varphi_k := \Psi$	$P\left(o_k, t \mid \Psi\right) = \left\lvert\langle\varphi_k \mid \Psi\rangle\right\rvert^2 = \left\lvert\int \bar{\varphi}_k \Psi \mathrm{d}V\right\rvert^2$
P6	**Time evolution:** The time evolution of a closed system (the change of a system's state over time) is a linear transformation via a unitary operator	$\lvert\Psi(t)\rangle = \underbrace{U\left(t, t_0\right)}_{\text{time operator}} \lvert\Psi\left(t_0\right)\rangle$ $\equiv$ $\mathrm{i}\hbar\frac{\mathrm{d}}{\mathrm{d}t}\lvert\psi(t)\rangle = H(t)\lvert\psi(t)\rangle$
P7	**Density operator:** The Hilbert space of a composite system is the Hilbert space tensor product of the state spaces associated with the component systems	$\hat{\rho} = \lvert\phi\rangle \otimes \langle\psi\rvert = \underbrace{\begin{bmatrix} \phi_1\psi_1^* & \phi_1\psi_2^* & \ldots \\ \phi_2\psi_1^* & \phi_1\psi_2^* & \ldots \\ \ldots & \ldots & \phi_n\psi_n^* \end{bmatrix}}_{\text{density matrix}}$

Table 11.6: **Summary of common wave (probabilistic) quantum physics in the orthodox interpretation**

The first five postulates above can be generalized with the density operator!



AXIOMS OF QUANTUM PHYSICS

A concise summary of the foundational postulates underlying all quantum theory.

#	Axiom	Postulate (Mathematical Statement)	Illustration
1	**State Axiom** The state of a quantum system is described by a vector in a complex Hilbert space.	The physical state of a (closed) quantum system is represented by a unit vector $\lvert\psi\rangle$ in a complex, separable Hilbert space $\mathcal{H}$, $$\lvert\psi\rangle \in \mathcal{H}, \quad \langle\psi\vert\psi\rangle = 1$$ where vectors that differ by a global phase $e^{i\theta}$ represent the same physical state.	z $\lvert 0\rangle$ $\lvert\psi\rangle$ y x $\lvert 1\rangle$
2	**Observable Axiom** Physical quantities are represented by Hermitian operators on the Hilbert space.	Every observable (measurable physical quantity) is represented by a self-adjoint (Hermitian) operator $\hat{A}$ on $\mathcal{H}$, $$\hat{A} = \hat{A}^\dagger$$ The set of all observables forms a real vector space.	$\hat{A} = \begin{pmatrix} a & c & \cdots \\ c^* & b & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \quad \lvert\psi\rangle = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \end{pmatrix}$
3	**Spectrum Axiom** Measurement outcomes are given by the spectrum of the corresponding observable.	If $\hat{A}$ is an observable, the possible results of a measurement of $\hat{A}$ are the eigenvalues of $\hat{A}$. That is, if $$\hat{A}\,\lvert a_k\rangle = a_k\,\lvert a_k\rangle$$ then the possible outcomes are $\{a_k\}$.	Energy $\lvert a_3\rangle$ $\lvert a_2\rangle$ $\lvert a_1\rangle$ $\vdots$ a_0 $\lvert a_0\rangle$
4	**Measurement Axiom** Measurement causes state update (collapse) according to the projection postulate.	Upon measuring observable $\hat{A}$ in state $\lvert\psi\rangle$: • Outcome a_k occurs with probability $p_k = \lvert\langle a_k \mid \psi\rangle\rvert^2$. • The state updates to the normalized eigenstate $$\lvert\psi'\rangle = \frac{\lvert a_k\rangle}{\sqrt{\langle a_k\vert\psi\rangle\ \langle\psi\vert a_k\rangle}} = \frac{\lvert a_k\rangle}{\sqrt{p_k}}$$ For degenerate eigenvalues, the post-measurement state is the projection of $\lvert\psi\rangle$ onto the corresponding eigenspace, normalized.	Before measurement After obtaining a_k $\lvert\psi\rangle$ $\lvert a_k\rangle$ a_1 a_k a_n a_1 a_k a_n Basis $\{\lvert a_k\rangle\}$
5	**Composite System Axiom** The state space of a composite system is the tensor product of the subsystem spaces.	For two quantum systems with Hilbert spaces $\mathcal{H}_A$ and $\mathcal{H}_B$, the composite system has Hilbert space $$\mathcal{H}_{AB} = \mathcal{H}_A \otimes \mathcal{H}_B$$ States of the composite system may be entangled.	A B $\lvert\Psi\rangle \in \mathcal{H}_A \otimes \mathcal{H}_B$ (entangled state)
6	**Unitary Evolution Axiom** The time evolution of a closed quantum system is governed by unitary operators.	The state of an isolated system evolves continuously and reversibly according to a unitary operator $\hat{U}(t)$, $$\lvert\psi(t)\rangle = \hat{U}(t)\,\lvert\psi(0)\rangle, \qquad \hat{U}^\dagger(t)\,\hat{U}(t) = \hat{I}$$ If the Hamiltonian $\hat{H}$ is time-independent, $$\hat{U}(t) = e^{-\frac{i}{\hbar}\hat{H}t}.$$	$t=0$ $t=t_1$ $t=t_2$ $t=t_3$ z z z z x y $\cdots$
7	**Probability Axiom** Probabilities are given by the Born rule.	For any measurement described by a set of orthogonal projectors $\{\hat{P}_k\}$ with $\sum_k \hat{P}_k = \hat{I}$, the probability of outcome k in state $\lvert\psi\rangle$ is $$p_k = \langle\psi \mid \hat{P}_k \mid \psi\rangle \geq 0$$ with $\sum_k p_k = 1$. In general, for a POVM $\{\hat{E}_k\}$ with $\sum_k \hat{E}_k = \hat{I}$, $$p_k = \langle\psi \mid \hat{E}_k \mid \psi\rangle.$$	p_k 1 0 $k=1$ $k=2$ $\cdots$ $k=N$ = Outcome p_1 p_2 $\cdots$ p_N Probability
8	**Identity Axiom** The identity operator represents the certain (or trivial) measurement.	The identity operator $\hat{I}$ corresponds to a measurement with a single outcome that occurs with probability 1 in any state: $$\langle\psi\vert\ \hat{I}\ \vert\psi\rangle = 1, \qquad \forall\,\lvert\psi\rangle \in \mathcal{H}$$	$\lvert\psi\rangle \rightarrow \boxed{\hat{I}} \rightarrow \lvert\psi\rangle$ $p = 1$

11.2.2 Classical principles of uncertainty

Before attacking directly quantum physics in front and the corresponding mathematical tools in-deep (and the pseudo-proofs of the five postulates), we must first introduce a simple classic example in which appears a special type of phenomena: the intrinsic presence of uncertainty in any measurement (as it is intrinsic we don't speak of "statistical uncertainty" as we have already study in the section Statistics!).

This study in classical form is fundamentally not very rigorous (for a more formal and rigorous approach, see below the "Robertson uncertainty relations" at page 4733), it will however normally help the average reader to better understand the quantum uncertainty (at least we hope so...) that we will study and determine further below and that will need no previous experimental considerations! However it seems that it was really one of the approaches used by Werner Heisenberg himself[5]!

Imagine that we would like to measure through a microscope the abscissa x of a particle and the components of its momentum $\vec{p}$. For the measurement of x to be possible, there must be a monochromatic light beam (to simplify...) parallel to the x-axis comes illuminate the particle, and that at least one photon collides the particle and reaches the eye of the observer:

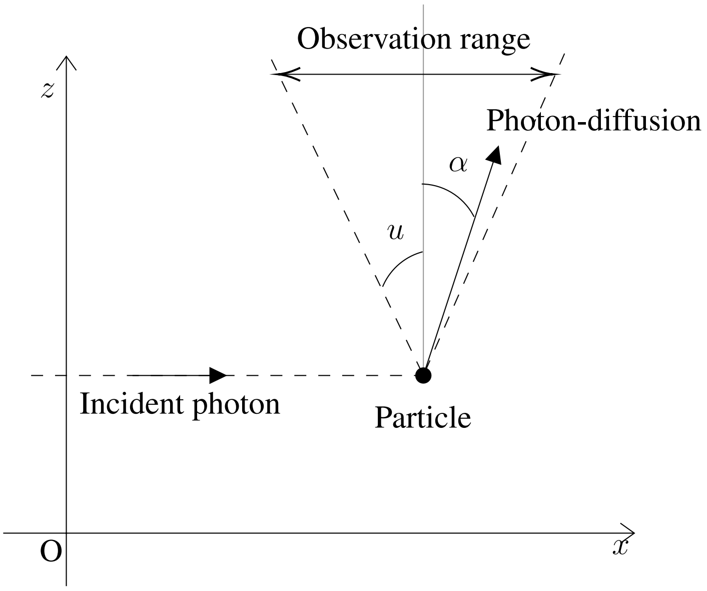


Figure 11.32: Heisenberg's microscope

Remark

Heisenberg's microscope exists only as a thought experiment, one that was proposed by Werner Heisenberg, criticized by his mentor Niels Bohr, and subsequently served as the nucleus of some commonly held ideas, and misunderstandings, about Quantum Mechanics. In particular, it provided an argument for the uncertainty principle on the basis of the principles of classical optics. While the act of measurement does lead to uncertainty, the loss of precision is less than that predicted by Heisenberg's argument when measured at the level of an individual state. The formal mathematical result remains valid.

---

[5] This isn't the way Werner Heisenberg derived the uncertainty relationships in his original book *Die physikalischen Prinzipien der Quantentheorie* (11929 according to holocene calendar). How he did involves the probabilistic wave function and a quite ugly assumption that's why we won't introduce here the original historical way of the proof by Heisenberg.

Once x measured, we can imagine any method for measuring the linear momentum.

Let us put α as the angle that do the direction of the photon after collision with the z-axis. Let us suppose to simplify the calculations that the particle has a relatively high mass so that we can neglect the energy change of the photon. We see that after the collision, the components of the linear momentum of the scattered photon following respectively the x-axis and z-axis are:

$$\frac{h\nu}{c}\sin(\alpha) \qquad \frac{h\nu}{c}\cos(\alpha) \tag{11.2.43}$$

Indeed, let us recall that the relations between the electromagnetic waves, the mass-energy equivalence and the linear momentum (see section Special Relativity page 5679) are:

$$E=mc^2=h\nu\Rightarrow m=\frac{h\nu}{c^2}\Rightarrow p=mv=mc=\frac{h\nu}{c^2}c=\frac{h\nu}{c} \tag{11.2.44}$$

It follows that the particle will see its amount of linear momentum altered. The components of the variation are then (remember that initially it was zero following the z-axes) those of the variation of the photon following:

$$\frac{h\nu}{c}(1-\sin(\alpha)) \qquad -\frac{h\nu}{c}\cos(\alpha) \tag{11.2.45}$$

between it initial and final linear momentum.

The only information we have about the angle α is that it is, in absolute value less than or equal to the opening angle u of the microscope objective (technical restrictions).

This implies that:

$$|\sin(\alpha)|\leq\sin(u) \qquad \cos(u)\leq|\cos(\alpha)|\leq\sin(u) \tag{11.2.46}$$

11.2.2.1 First classical uncertainty relation

When we will have measured the linear momentum $\vec{p}$ at the end of the experiment, we just saw that we will have to make the corrections:

$$\frac{h\nu}{c}(1-\sin(\alpha)),-\frac{h\nu}{c}\cos(\alpha) \tag{11.2.47}$$

of the linear momentum of the photon to know the true value of $\vec{p}$ of the particle just before the start of measurement.

In these corrections, there is an unknown portion corresponding to measurement errors on p_x and p_z of the photon. It is possible to establish that the maximum error Δp_x and Δp_z on the initial linear momentum is given by the x component of the "first classical Heisenberg's uncertainty relation":

$$\boxed{\Delta p_x = \frac{h\nu}{c}\sin(u)} \tag{11.2.48}$$

as we have $|\sin(\alpha)| < \sin(u)$. So it is like a worst error!

11.2.2.2 Second classical uncertainty relation

Let us now see what we can say about the measurement of the position of the particle.

Let us recall now that (see section Wave Optics page 4497) for a rectangular aperture we have by putting $n = 1$:

$$\sin(\theta) = \frac{\lambda}{e} \tag{11.2.49}$$

where θ (in Wave Optics) is the angle to clearly distinguish two diffraction minima (and therefore clearly an object emitting radiation between the same two points). Conversely, from the viewpoint of diffraction, the opening width e is given by:

$$e = \frac{\lambda}{\sin(\theta)} \tag{11.2.50}$$

The value of e may also be seen as the vision area (orthogonal projection of the rectangular aperture on the x axis) of width $x = e$ of the particle. Therefore:

$$x = \frac{\lambda}{\sin(\theta)} \tag{11.2.51}$$

Just as the maximum error of the linear momentum is given by the condition $|\sin(\alpha)| < \sin(u)$, we can also write $|\sin(\theta)| < \sin(u)$, which leads us to write:

$$\Delta x = \frac{\lambda}{\sin(u)} \tag{11.2.52}$$

If we multiply the following relation we just proved before:

$$\Delta p_x = \frac{h\nu}{c}\sin(u) = \frac{h}{\lambda}\sin(u) \tag{11.2.53}$$

with the prior-previous one, we get:

$$\Delta x \Delta p_x = \frac{\lambda}{\sin(u)}\frac{h\nu}{c}\sin(u) = \frac{h}{\lambda}\sin(u) \tag{11.2.54}$$

After simplification we get the "second classical Heisenberg's uncertainty relation" also named "classical spatial uncertainty":

$$\boxed{\Delta x \Delta p_x = h} \tag{11.2.55}$$

which thus represents the maximum experimental error of a small rectangular aperture opening of a microscope. Many quantum physics textbooks show that we fall back on exactly the same expression in many situations.

In his celebrated 11927 (holocene calendar) paper, *Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik*), Werner Heisenberg established this expression as the minimum amount of unavoidable momentum disturbance caused by any position measurement, but he did not give a precise definition for the uncertainties Δx and Δp. Instead, he gave some plausible estimates in each case separately.

Remark

The reader will easily verify that this relation applied to for a macroscopic object (size of the order of the centimetre), which position would be measurable with a precision of the micrometer gives a ridiculously low uncertainty of linear momentum and thus for the speed. But (!) the same relation applied to the mass of a particle such as that of an electron with a position measurement accuracy assumed of a tenth of a nanometre give an uncertainty about the speed of about $1,000\,[\text{m}\cdot\text{s}^{-1}]$!!

Thus, if we try to locate a particle with an accuracy always bigger, its momentum reaches extreme values. At some point, the magnitude of the linear momentum can be so big that the corresponding energy is sufficient to produce a particle-antiparticle pair. In other words, if we try to confine a particle in an increasingly small box on the one hand, we know less and less its linear momentum and from a certain threshold, we do not even know how many particles there are in the box!

However (!), we will see during our study of commutators applied to quantum physics theory that the true uncertainty relation (whose value differs from the one above as proved at page 4733) only appears naturally from mathematical properties and the definition of the linear momentum.

More generally, for a particle in a volume of dimensions x,y,z, a vector state is characterized by six quantities (x,y,z,p_x,p_y,p_z) in the phase space (phase space which is therefore of 6 dimensions) and the quantum state occupies the "cube" of volume:

$$(\Delta x\Delta p_x)(\Delta y\Delta p_y)(\Delta z\Delta p_z)=h^3 \qquad (11.2.56)$$

What is remarkable in this simplistic approach is that the Planck constant naturally emerges as the unit of minimum universal measurement of uncertainty of experimental physics through the wave-particle dualism! Heisenberg wrote that this result establishes the ultimate failure of causality in quantum physics.

11.2.2.3 Third classical uncertainty relation

In special relativity, we saw that (x, y, z, ct) are the components of a space-time four-vector and also (p_x, p_y, p_z, Ec^{-1}) those of a vector of a four energy-momentum vector.

It is therefore natural to complete the three spatial relation of the type $\Delta x\Delta p_x = h$ by extension:

$$\underbrace{\begin{gathered}x \Rightarrow p_x \rightarrow \Delta x\Delta p_x = h\\ y \Rightarrow p_y \rightarrow \Delta y\Delta p_y = h\\ z \Rightarrow p_z \rightarrow \Delta z\Delta p_z = h\end{gathered}}_{ct\Rightarrow\frac{E}{c}\rightarrow\Delta(ct)\Delta\left(\frac{E}{c}\right)=h} \tag{11.2.57}$$

We obtain thus roughly the "third classical uncertainty relation" also named "classical temporal uncertainty" or "energy-time uncertainty":

$$\boxed{\Delta t\Delta E = h} \tag{11.2.58}$$

However (!), we will see during our study of commutators applied to the theory of quantum physics, that this uncertainty relation (whose value differs from the one above) also appears naturally from only the mathematical properties and the definition of the linear momentum.

Remark

We will come back later on the implications of this temporal uncertainty whose implications are the basis of Quantum Cosmology (and therefore of the creation of our Universe... as seen at page 6083!) and also of Quantum Field Theory and in particular regarding to the Yukawa potential (see section Quantum Field Theory page 5164).

The established classical uncertainties relation will allow us to better understand the "real" uncertainties relations in their modern quantum form. For this, among others, we will have to make use of the necessary mathematical artillery. However, for the sake of clarity, we wanted to present the wave quantum physics in the simplest and least formal way possible as always in this textbook. This presentation may perhaps take the reader to many misinterpretations, he must therefore remain cautious as long as it has not seen the rigorous proof!

Let us now come back as promise in the section of Magnetostatics on the Wilson bubble chamber that obviously seems to be incompatible with Heisenberg' inequalities:

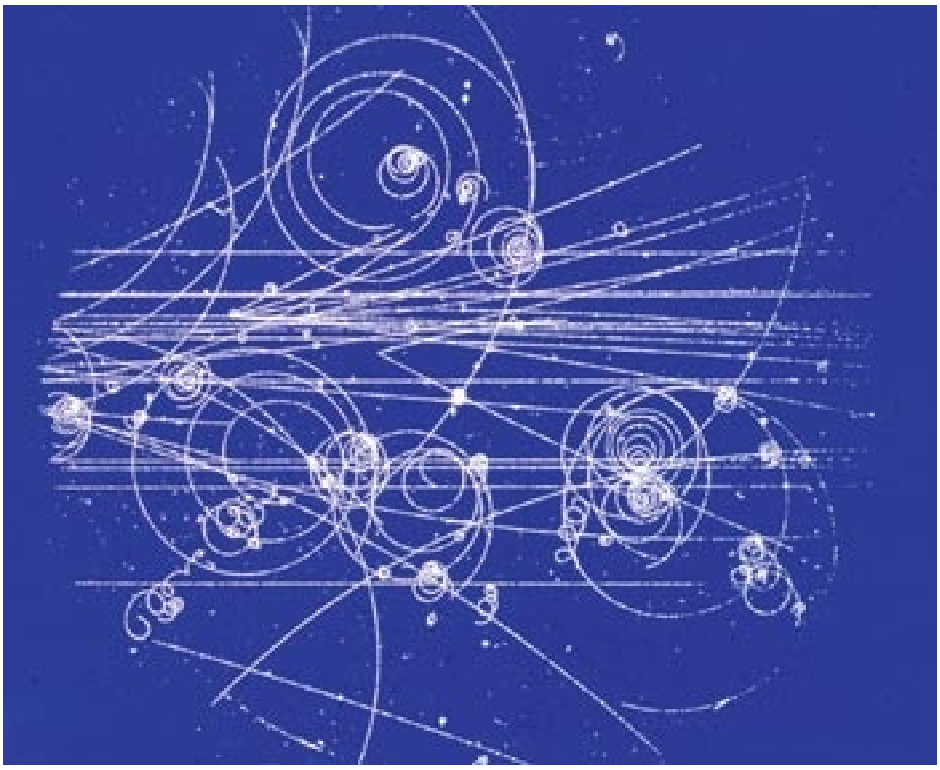


Figure 11.33: Wilson bubble chamber (source: CERN)

The issue of having the possibility to track a particle with such precision and on such a long distance that seems to violate both: Heisenberg's incertitude principle and Wave-Particle duality is named in general the "Mott problem[6]".

The problem was first formulated in 11929 (holocene calendar) by Sir Nevill Francis Mott and Werner Heisenberg, illustrating the paradox of the collapse of a spherically symmetric wave function into the linear tracks seen in a bubble chamber.

In practice, virtually all high energy physics experiments, such as those conducted at particle colliders, involve wave functions which are inherently spherical. Yet, when the results of a particle collision are detected, they are invariably in the form of linear tracks (see, for example, the illustrations accompanying the article on bubble chambers). It is somewhat strange to think that a spherically symmetric wave function should be observed as a straight track, and yet, this occurs on a daily basis in all particle collider experiments.

A related variant formulation was given in 11953 (holocene calendar) by Mauritius Renninger, and is now known as "Renninger's negative-result gedanken experiment". In this formulation, it is noted that the absence of a particle detection can also constitute a quantum measurement; namely, that a measurement can be performed even if no particle whatsoever is detected.

In the original 11929 (holocene calendar) formulation by Mott and Heisenberg, the spherical wave function of an alpha ray emitted from the decay of a radioactive atomic nucleus was considered. It was noted that the result of such a decay is always observed as linear tracks seen in Wilson's cloud chamber (i.e. bubble chamber). Intuitively, one might think that such a wave function should randomly ionize atoms throughout the cloud chamber, but this is not the case. Mott demonstrated that by considering the interaction in configuration space, where all of the atoms of the cloud chamber play a role, it is overwhelmingly probable that all of the condensed droplets in the cloud chamber will lie close to the same straight line. What is uncertain is which straight line the wave packet will reduce to; the probability distribution of straight tracks is spherically symmetric.

---

[6] So a Mott problem is a paradox that illustrates some of the difficulties of understanding the nature of wave function collapse and measurement in quantum mechanics

Remark

Rigorously the "Heisenberg's uncertainty principles" (HUP) should be named "Heisenberg's indetermination theorem" as they can be proven (therefore it is not a principle!) and they are not a limit of the instrumentation of a limit of precision of Nature itself (hence the "indetermination" in place of the "uncertainty").

But this is a detail as many statements in physics are provable and keep the name "principle" because at the time they were discovered they were not provable!

11.2.3 Quantum algebra

Under uncommon and unofficial term of "quantum algebra" (so do not to abuse it!) we wish to introduce and recall to the reader some mathematical tools that will be very useful to solve certain equations of Quantum Physics. It is therefore of prime importance to understand (or to have understand regarding to the reminders) at best what will follow!

Remark

Purists may climb the walls when reading what follows ...

11.2.3.1 Linear functional operators

Definition (#11.2.503): The "linear operators" are mathematical entities acting on functions (see section Functional Analysis page 2209) or vectors (see section Vector Calculus page 1797).

The functions on which can operate on these operators can be functions of a single variable x, such as $f(x)$, or of three coordinates of a point x,y,z such as $f(x,y,z)$ or written briefly $f(\vec{x})$.

We will have to write integrals of these functions, which are often extended to the entire space. In the case of a function of the three-dimensional coordinates of a point, we will adopt the following notation that we already know:

$$\iiint f(x,y,z)\mathrm{d}x\mathrm{d}y\mathrm{d}z=\int f(\vec{x})\mathrm{d}^3\vec{x} \qquad (11.2.59)$$

These notations, that are quite essential to simplify the expressions that we will meet in quantum physics being established, we can go back to our operators.

Starting from a function f, if we associate to it a function g of the same kind, that is to say dependent on the same variables, we can say that g is the result of the action of an operator α on f and write this symbolically as a single product:

$$g=\alpha f \qquad (11.2.60)$$

So α multiply, or better "acts on", f.

But we introduce immediately a fundamental restriction: only interest us linear operators (as in Linear Algebra...), that is to say, such as:

$$\alpha(\lambda_1 f_1+\lambda_2 f_2)=\lambda_1(\alpha f_1)+\lambda_2(\alpha f_2) \qquad (11.2.61)$$

whatever the coefficients λ_1 and λ_2

A very simple category of operator consists simply of the numbers (real or complex scalars). Thus, in the relation:

$$\chi=\lambda f \qquad (11.2.62)$$

χ is a function that depends linearly of f, trough a linear operator that we write λ, λ being a number (typically the position operator in quantum physics).

There are two important special cases to consider:

1. Zero operator: $\lambda=0$ where $\chi=0\cdot f=0$ will be obviously a function equal to zero everywhere...

2. Unit operator (or identity operator): $\lambda=1$ where $\chi=1\cdot f=1$ (which is just as simple...)

Remark

The "nabla" operator ∇ is also a functional linear operator (we will see that a bit later) that in quantum physics is found in the energy operator.

We also check easily for functional operators that they are (see the sections of Set Theory and Linear Algebra if necessary):

- Commutative with respect to the addition
- Associations in relation to the addition and multiplication
- Distributive over addition left and right

So far, nothing distinguishes algebra operators of that of numbers operators. But there are however two properties that we must always keep in mind to not do mistakes when we make calculations with operators:

1. Two operators do not commute in general compared to the multiplication (as in linear algebra ...), that is to say that in general given two functional operators α and β:

$$\alpha\cdot\beta\neq\beta\cdot\alpha \tag{11.2.63}$$

2. If we encounter an expression such as $\alpha_1\beta+\beta\alpha_2$, so we do not have the right to do in general a factorization (there is therefore a particular structure group that is non-commutative)!

✍**Example:**

A simple and important example, as useful for what will follow (very close to a practical case that we will see later), of two operators that do not commute with a function of a single variable is as follows (where f is any).

Let us consider the operator $\mathrm{d}/\mathrm{d}x$ acting on $xf(x)$:

$$\frac{\mathrm{d}}{\mathrm{d}x}(x\cdot f)=1\cdot f+x\frac{\mathrm{d}}{\mathrm{d}x}f \tag{11.2.64}$$

simplifying by f:

$$\frac{\mathrm{d}}{\mathrm{d}x}x=1+x\frac{\mathrm{d}}{\mathrm{d}x} \tag{11.2.65}$$

So we have above an example of two operators that do not commute since:

$$\frac{\mathrm{d}}{\mathrm{d}x}x \neq x\frac{\mathrm{d}}{\mathrm{d}x} \tag{11.2.66}$$

Remarks

R1. If an operator can switch anyhow with another operator is that it is a number (it joined with the concept of measurement which we have mentioned in the postulates).

R2. When a state (a mathematical function in the formal sense) is unchanged by an operator, the sate is then named "eigenstate" or "eigenvector" of the system (we will see practical examples below). The state is then perfectly measurable and is assimilated to the classical observable.

Before going to the next subject, to close this introduction on linear functional operators, consider that we know the three-dimensional Schrödinger equation (which we will prove further below):

$$\frac{\partial^2\Psi}{\partial x^2}+\frac{\partial^2\Psi}{\partial y^2}+\frac{\partial^2\Psi}{\partial z^2}+\frac{8\pi^2 m}{h^2}(E_{\text{tot}}-E_p)\Psi=0 \tag{11.2.67}$$

or written differently (that is more aesthetic ...) with the Laplacian of $\vec{\nabla}^2=\Delta$ of a scalar field (see section Vector Calculus page 1899):

$$\Delta\Psi+\frac{2m}{\hbar^2}(E_{\text{tot}}-E_p)\Psi=\Delta\Psi+\frac{2m}{\hbar^2}(H-V)\Psi=0 \tag{11.2.68}$$

or after rearranging:

$$\frac{2m}{\hbar^2}(E_{\text{tot}}-E_p)\Psi=-\Delta\Psi \tag{11.2.69}$$

or after simplification:

$$\frac{2m}{\hbar^2}(E_{\text{tot}}-E_p)=-\Delta \tag{11.2.70}$$

Then the total energy operator (the Hamiltonian H in other words ...) is expressed as:

$$E_{\text{tot}}=-\frac{\hbar^2}{2m}\Delta+E_p \tag{11.2.71}$$

or in Hamiltonian notation:

$$\boxed{H=-\frac{\hbar^2}{2m}\Delta+V} \tag{11.2.72}$$

Remark

Here we fall back naturally on the second expression given in the second postulate but the notation V for the potential energy can be sometimes confusing with the electrical potential for beginners that have never deal too much with Analytical Mechanics.

On the other hand, we know that:

$$E_{\text{tot}}=E_c+E_p=\frac{p^2}{2m}+E_p \tag{11.2.73}$$

The last two expressions must be identical. The only way to meet these conditions is to put:

$$\boxed{p_x=\frac{1}{\mathrm{i}}\hbar\frac{\partial}{\partial x}=-\mathrm{i}\hbar\frac{\partial}{\partial x}\qquad p_y=\frac{1}{\mathrm{i}}\hbar\frac{\partial}{\partial y}=-\mathrm{i}\hbar\frac{\partial}{\partial y}\qquad p_z=\frac{1}{\mathrm{i}}\hbar\frac{\partial}{\partial z}=-\mathrm{i}\hbar\frac{\partial}{\partial z}} \tag{11.2.74}$$

which are the "Hermitian operators of linear momentum" in wave quantum physics (expressed here in Cartesian coordinates) and that need to be remembered throughout this section of the textbook!

Remark

Here we fall back naturally on one of the operators listed in the third postulate.

We can check the legitimacy of these operators by re-injecting them into the expression of the kinetic energy:

$$\begin{aligned}E_c&=\frac{p^2}{2m}=\frac{p_x^2+p_y^2+p_z^2}{2m}\overset{?}{=}\frac{1}{2m}\left[\left(-\mathrm{i}\hbar\frac{\partial}{\partial x}\right)^2+\left(-\mathrm{i}\hbar\frac{\partial}{\partial y}\right)^2+\left(-\mathrm{i}\hbar\frac{\partial}{\partial z}\right)^2\right]\\&\overset{?}{=}\frac{1}{2m}\left[+\mathrm{i}^2\hbar^2\frac{\partial^2}{\partial x^2}+\mathrm{i}^2\hbar^2\frac{\partial^2}{\partial y^2}+\mathrm{i}^2\hbar^2\frac{\partial^2}{\partial z^2}\right]=-\frac{1}{2m}\hbar^2\left[\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right]\\&=-\frac{1}{2m}\hbar^2\Delta\end{aligned} \tag{11.2.75}$$

Moreover, it is easy to verify that this development is remain correct if we take the complex conjugate of the linear momentum operator!!!

Thus, the total energy operator H (Hamiltonian) is, also, well Hermitian! This result is very important to check such calculations using the orthogonality property of eigenfunctions that we will see further below.

11.2.3.1.1 Hermitian and Self-adjoint operators

Caution!!!! Reading the following lines could be quite abstract... However, if you do not understand much this is not very important because often everything will become evident during the study and the developments of concrete examples that will be give later, after which you will be able to reread the following and understand it at the same time.

Let us consider the two following integrals extended to the whole space (inside the integral it is a multiplication of functions and operators) without trying to understand their utility for the moment:

$$Q=\int\bar{f}(\alpha\cdot\chi)\mathrm{d}\vec{x}\qquad\text{and}\qquad\bar{Q}=\int\bar{\chi}(\beta\cdot f)\mathrm{d}\vec{x} \tag{11.2.76}$$

where let us we recall that the notation $\bar{f}$ is the complex conjugate of f. It should be noted that in these two integrals, α and β represent operators.

We observe in the developments of quantum physics that these two integrals are equal (and for Q and $\bar{Q}$ to be equal we must therefore have $Q \in \mathbb{R}$) and that there is a one-to-one correspondence between the operators α and β, we say then that β is the "adjoint operator" of α (the transpose of the conjugate) or that it is the "hermitian" of α (both terms are customary) and we write:

$$\beta = \alpha^\dagger \tag{11.2.77}$$

if the two previous integrals are satisfied.

From this definition, we deduce the following important identity:

$$\boxed{\int \bar{f}(\alpha \cdot \chi)\mathrm{d}\vec{x} = \overline{\left[\int \bar{\chi}(\alpha^\dagger \cdot f)\mathrm{d}\vec{x}\right]}} \tag{11.2.78}$$

✍**Example:**

Let us consider the operator:

$$\alpha = \frac{\mathrm{d}}{\mathrm{d}x} \tag{11.2.79}$$

Then by integration by parts:

$$\begin{aligned}\int_{-\infty}^{+\infty} \bar{f}(\alpha\chi)\mathrm{d}x &= \int_{-\infty}^{+\infty} \bar{f}\left(\frac{\mathrm{d}\chi}{\mathrm{d}x}\right)\mathrm{d}x = (\bar{f}\chi(+\infty) - \bar{f}\chi(-\infty)) - \int_{-\infty}^{+\infty} \left(\frac{\mathrm{d}}{\mathrm{d}x}\bar{f}\right)\chi\mathrm{d}x \\ &= -\int_{-\infty}^{+\infty} \chi\left(\frac{\mathrm{d}}{\mathrm{d}x}\bar{f}\right)\mathrm{d}x = \int_{-\infty}^{+\infty} \chi\overline{\left(-\frac{d}{\mathrm{d}x}\bar{f}\right)}\mathrm{d}x = \int_{-\infty}^{+\infty} \chi\overline{\left(\alpha^\dagger f\right)}\mathrm{d}x \\ &= \int_{-\infty}^{+\infty} \overline{\bar{\chi}\left(\alpha^\dagger f\right)}\mathrm{d}x = \overline{\int_{-\infty}^{+\infty} \bar{\chi}\left(\alpha^\dagger f\right)\mathrm{d}x}\end{aligned} \tag{11.2.80}$$

Remark

We will prove the relation above in a concrete but particular example in our study of Quantum Field Theory (see the next section) and we will return to this in a more rigorous way in our presentation of the Dirac formalism in the section Relativistic Quantum Physics.

The adjoint operator has several properties, the only ones that will interest us in this section are:

P1. $(\alpha^\dagger)^\dagger$ which it is unnecessary to prove, since this relation comes from the very definition of the adjoint operator.

P2. α being considered as a complex number (special case of operator such as the linear momentum one that we have seen before) then we have $\alpha^{\dagger}=\bar{\alpha}$.

An extremely important category of operators is thus constituted by the "self-adjoint Hermitian operators", or simply "self-adjoint operators" equal by definition to their adjoint:

$$\alpha=\alpha^{\dagger} \tag{11.2.81}$$

It must be known to the reader that sadly some teachers and physicists say that "self-adjoint" and "hermitian" are equivalent concepts. This is an abusive language used by physicists that you will never found by mathematicians.

Remark

Remember that these operators can also be matrices!

We also notice that if we take a Hermitian operator (such as that of the linear momentum to make a simple example) and that we multiply that latter by the unitary imaginary number i then it becomes non-Hermitian (being not in $\mathbb{C}$ anymore) but at the same time it becomes self-adjoint as it will be a real operator (belonging to $\mathbb{R}$).

Theorem 11.2. Any operator, that we will denote by ξ, can be decomposed in a unique way into Hermitian (self-ajdoint) α and non-Hermitian (non-self-adjoint) $\mathrm{i}\beta$ parts, that is, we can write:

$$\xi=\alpha+\mathrm{i}\beta \tag{11.2.82}$$

where α and β are Hermitian (self-adjoints).

Proof 11.2.1. If:

$$\xi=\alpha+\mathrm{i}\beta \tag{11.2.83}$$

then:

$$\xi^{\dagger}=\alpha-\mathrm{i}\beta \tag{11.2.84}$$

as it is a simple complex number, then:

$$\alpha=\frac{1}{2}(\xi+\xi^{\dagger}) \qquad \beta=\frac{1}{2}(\xi-\xi^{\dagger}) \tag{11.2.85}$$

The sum or the difference of the operator and its adjoint is therefore a self-adjoint operator (the sum or subtraction between self-adjoints operators, hence remains self-adjoints).

■ Q.E.D.

In general, it is quite obvious that the product of two self-adjoints operators α and β is not necessarily itself self-adjoint operator, that is:

$$\alpha\beta=(\alpha\beta)^{\dagger} \tag{11.2.86}$$

since we can verify that the condition for which the product of two self-adjoint operators is itself self-adjoint is that the two operators "commute" (see below):

Theorem 11.3. The product of two self-adjoint operators is self-adjoint if and only if the two operators commute.

Proof 11.3.1. First remember that by definition and explicitly:

$$\xi^\dagger = \bar{\xi}^T \tag{11.2.87}$$

Therefore (see section Linear Algebra page 1951):

$$(\alpha\beta)^\dagger = \left(\overline{\alpha\beta}\right)^T = \bar{\beta}^T\bar{\alpha}^T = \beta^\dagger\alpha^\dagger \tag{11.2.88}$$

Therefore we see that:

$$(\alpha\beta)^\dagger = \alpha^\dagger\beta^\dagger \tag{11.2.89}$$

if and only if $\alpha^\dagger$ and $\beta^\dagger$ commute or equivalently, if and only if:

$$(\alpha\beta)^\dagger = (\beta\alpha)^\dagger \tag{11.2.90}$$

That is to say if and only if:

$$\alpha\beta = \beta\alpha \tag{11.2.91}$$

■ Q.E.D.

✍**Example:**

In the Schrödinger picture we will see later that we have:

$$\mathrm{i}\partial_t|\psi(t)\rangle = \hat{H}|\psi(t)\rangle \tag{11.2.92}$$

where $\hat{H}$ (which may be explicitly time dependent or not) is the Hamilton operator, which must be selfadjoint.

Taking the adjoint of this equation, you get:

$$-\mathrm{i}\partial_t\langle\psi(t)| = \langle\psi(t)|\hat{H}^\dagger = \langle\psi(t)|\hat{H} = \langle\hat{H}\psi(t)| \tag{11.2.93}$$

11.2.3.1.2 Commutators and Anticommutators

In mathematics, a "commutator" gives an indication of the extent to which a certain binary operation fails to be commutative. There are different definitions used in group theory and ring theory. **Definitions (#11.2.504):**

D1. The "commutator" of two operators α and β, is written:

$$\boxed{[\alpha,\beta]=\alpha\beta-\beta\alpha=[\alpha,\beta]_-} \tag{11.2.94}$$

D2. The "anticommutator" of two operators α and β, is written:

$$\boxed{[\alpha,\beta]_+=\alpha\beta+\beta\alpha} \tag{11.2.95}$$

Which is generally written in the following form:

$$\{\alpha,\beta\}=\alpha\beta+\beta\alpha \tag{11.2.96}$$

Remarks

R1. As the commutator is much more common in developments than the anticommutator, if there is no possible confusion, we denote it simply $[\alpha,\beta]$.

R2. Concrete and trivial examples of these commutators in our study of Quantum Physics will be presented in the text that follows.

Let us mention some obvious properties of commutators (those that we will use the most in this textbook):

P1. Anticommutativity:

$$[\alpha,\beta]=-[\beta,\alpha] \tag{11.2.97}$$

P2. Left-linearity:

$$[\lambda_1\alpha_1+\lambda_2\alpha_2,\beta]=\lambda_1[\alpha_1,\beta]+\lambda_2[\alpha_2,\beta] \tag{11.2.98}$$

P3. Right-linearity:

$$[\alpha,\lambda_1\beta_1+\lambda_2\beta_2]=\lambda_1[\alpha,\beta_1]+\lambda_2[\alpha,\beta_2] \tag{11.2.99}$$

where λ_1, λ_2 are any number (proofs are made - if necessary - during the development of practical examples further below).

Let us look for the adjoint of $[\alpha,\beta]$:

$$[\alpha,\beta]^\dagger=(\alpha\beta-\beta\alpha)^\dagger=(\alpha\beta)^\dagger-(\beta\alpha)^\dagger=(\beta^\dagger\alpha^\dagger-\alpha^\dagger\beta^\dagger) \tag{11.2.100}$$

Hence a very simple result:

$$[\alpha,\beta]^{\dagger}=[\alpha^{\dagger},\beta^{\dagger}] \tag{11.2.101}$$

The following relation is very useful in practice (trivial, but as usual if necessary we can add the detailed proof on request):

$$[\alpha_1\alpha_2,\beta]=\alpha_1[\alpha_2,\beta]+[\alpha_1,\beta]\alpha_2 \tag{11.2.102}$$

We also have:

$$[\alpha,\beta_1\beta_2]=\beta_1[\alpha,\beta_2]+[\alpha,\beta_1]\beta_2 \tag{11.2.103}$$

We will prove later with a concrete case that if two operators do not commute, then it is impossible to have a state having a precise and unique value for both operators at the same time (in quantum physics there is an experimental configuration where the first operator represents the linear momentum and the second the spatial coordinate), this is why the operators are often referred to as "observables".

Let us take a moment to look at a concrete example of commutators and whose one of the results is absolutely fundamental!

We have proved earlier above the relations:

$$p_x=\frac{1}{\mathrm{i}}\hbar\frac{\partial}{\partial x}\qquad p_y=\frac{1}{\mathrm{i}}\hbar\frac{\partial}{\partial y}\qquad p_z=\frac{1}{\mathrm{i}}\hbar\frac{\partial}{\partial z} \tag{11.2.104}$$

Consider the relation (simple usual mathematical differential):

$$\frac{\partial}{\partial x}[x\psi(x)]=x\frac{\partial}{\partial x}\psi(x)+\psi(x) \tag{11.2.105}$$

If we divide by ψ on both sides of the equality and then multiply by $\hbar/\mathrm{i}$, then we get:

$$\frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial x}x=x\frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial x}+\frac{\hbar}{\mathrm{i}} \tag{11.2.106}$$

what give us:

$$p_x x=xp_x-\mathrm{i}\hbar\Leftrightarrow xp_x-p_x x=\mathrm{i}\hbar \tag{11.2.107}$$

therefore it comes that the commutator of x and p_x is equal to $\mathrm{i}\hbar$ and therefore that the quantities do not commute. We therefore have the following non-commuting relation:

$$\boxed{[x,p_x]_{\text{cycl.}}=\mathrm{i}\hbar} \tag{11.2.108}$$

that we often find in literature in the form (with the symbol of Kronecker):

$$[x_\mu,p_\nu]=\mathrm{i}\hbar\delta_{\mu\nu} \tag{11.2.109}$$

Thus (in the context of the second postulate), the two observables x and p_x, whose operators do not commute, do not have a common eigenvector basis. They are therefore not simultaneously measurable with precision and therefore constitute an Heisenberg uncertainty (see proof after the remark)!

Remarks

R1. The abbreviation (cycl.) means that the letters (x, y, z) can be circularly changed and the result remains the same.

R2. Although this result may seem surprising, it is none the less extremely correct as it results from mathematical reasoning. We can not make it more simpler and rigorous.

Theorem 11.4. Operators that do not commute, do not have a common eigenvector (eigenfunction) basis or conversely operators that do commute, have a common eigenvector (eigenfunction) basis.

In other words, given A and B two hermitian operators (i.e. matrices $M_{nn}(\mathbb{C})$) such that $[A,B]=0$. We can the found a basis $\mathcal{H}$ made by common eigenvectors of A and B.

Proof 11.4.1. Let us consider two operators A and B which represent observables and are therefore Hermitian. If there is a ψ such that:

$$A\psi=a\psi \qquad \text{and} \qquad B\psi=b\psi \tag{11.2.110}$$

then ψ is a simultaneous (common) eigenfunction (eigenvector) of A and B, belonging to eigenvalues a and b, respectively. Hence:

$$\begin{aligned}BA\psi=B(a\psi)=aB\psi=ab\psi\\AB\psi=A(b\psi)=bA\psi=ba\psi\end{aligned} \tag{11.2.111}$$

and therefore:

$$(AB-BA)\psi=0 \tag{11.2.112}$$

■ Q.E.D.

Let us now also consider the relation:

$$\frac{\partial}{\partial x}[y\psi(x)]=y\frac{\partial}{\partial x}\psi(x) \tag{11.2.113}$$

and proceeding in the same way as before, we get:

$$\boxed{[y,p_x]_{\text{cycl.}}=0} \tag{11.2.114}$$

The two relations:

$$[x,p_x]_{\text{cycl.}}=\mathrm{i}\hbar \qquad \text{and} \qquad [y,p_x]_{\text{cycl.}}=0 \tag{11.2.115}$$

can be summarized as follows:

$$\boxed{[q_k,p_l]=[q_k,p_l]_{\_}=q_kp_l-p_lq_k=\mathrm{i}\hbar\delta_{kl}} \tag{11.2.116}$$

using the generalized coordinates and generalized moments (see section Analytical Mechanics page 3156 and page 3180) and are remarkable from several point of views:

1. First, because from purely theoretical and mathematical considerations we fall back on an equivalent (but not equal) uncertainty to that obtained during our study of the classical Heisenberg's uncertainty principles seen earlier above (which we recall was equal to h rather than $\hbar$).

   Indeed, if we take the module of the left commutator, then we get the "spatial uncertainty relation of Heisenberg":

   $$\boxed{|[x,p_x]|_{\text{cycl.}}=\hbar} \tag{11.2.117}$$

   which for recall, can also be written in the form:

   $$|[x,v_x]|_{\text{cycl.}}=\frac{\hbar}{m} \tag{11.2.118}$$

   The Planck constant being extremely small, this explains why this effect is impossible to detect on our macroscopic scale. By cons, the mass of the electrons being extremely small too, the above fraction becomes noticeable for an electron and the effect of this uncertainty is then important!

   Finally, by the commutation of the components of the linear momentum quadrivector (see section Special Relativity page 5684), we get the "temporal uncertainty relation of Heisenberg" or "energy-time uncertainty":

   $$\boxed{|[t,E]|=\hbar} \tag{11.2.119}$$

   Caution!!!! The time-energy uncertainty relation (and other time-"observable" uncertainty relations that can be constructed) is (considered) not to have same meaning as canonical uncertainty relations. Meaning uncertainty relations constructed from canonical dynamical variables/observables (in the Hamiltonian sense), like position and momentum, since time parameter is not an observable and also not an operator in QM/QFT formalisms.

   Remark

   With the use of more advanced mathematics we can prove that in fact a factor $1/2$ is missing in front of the $\hbar$ (see further below page 4733).

   There is a quite fantastic consequence from the uncertainty about time and energy and relativity. Imagine the most complete vacuum (quantum vacuum[7]) and suppose that we look at what happens at a given point in space for a very short time. Then the principle of temporal uncertainty tells us that the energy of this state (the vacuum!) is very imprecise. But Special Relativity says that energy is also mass (and therefore also a field), and therefore particles. So, during this very short time particles can spontaneously appear from the vacuum (see [382])! We name them "virtual particles" (even if they are real because we can measure their consequences in particles accelerators) because they disappear very quickly and are generated by the "vacuum quantum fluctuations". We will come back on virtual particles later in the section of Quantum Field Theory.

---

[7] As we already know it from our study of Probabilities and as we will see now from the Quantum Physics point of view the "nothingness" (see page 4742) in the philosophical sense is a physical impossibility. So, a professional scientist who talks about "nothingness" is actually talking most of the time about a quantum vacuum!

Definition (#11.2.505): A "quantum fluctuation" refers to the fact that when a physical system is prepared in a given state and some property or quantity is measured, and this procedure is repeated many times, then the measured outcomes may fluctuate even though the system was prepared each time in the same state.

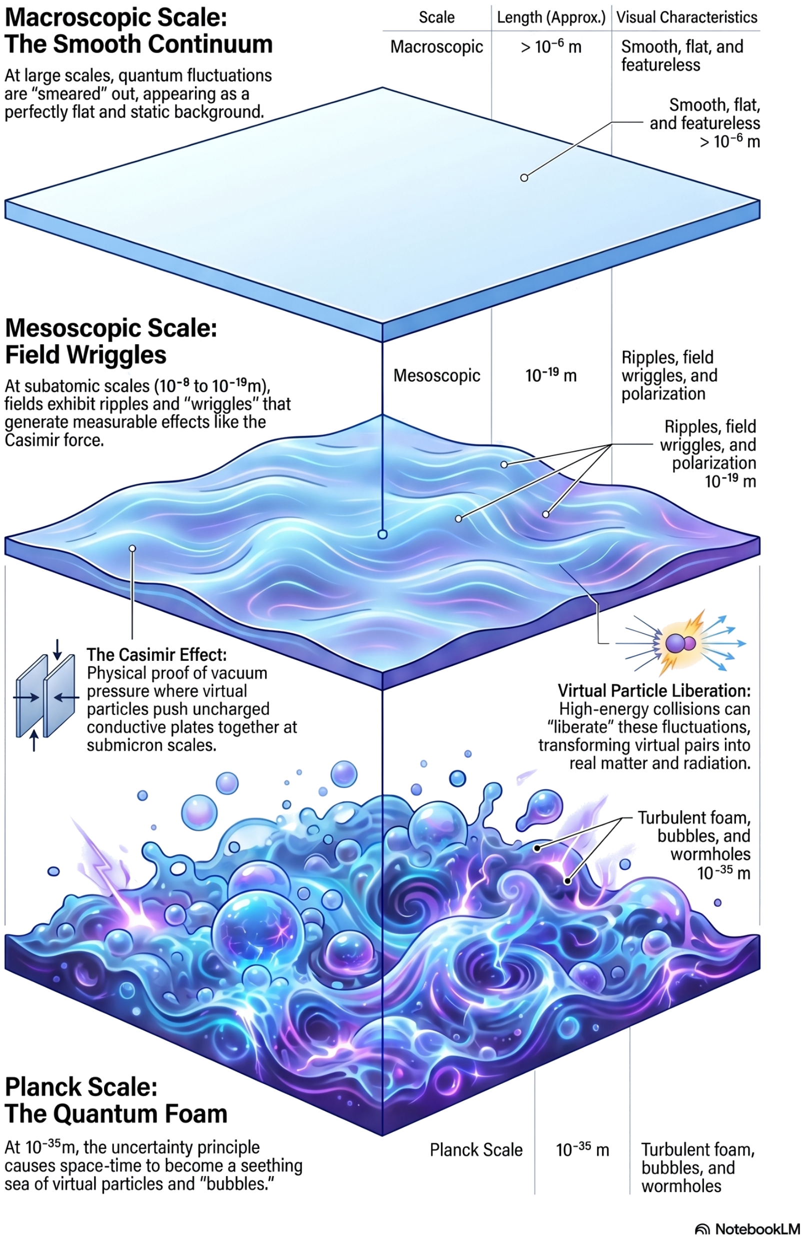

Macroscopic Scale:
The Smooth Continuum
At large scales, quantum fluctuations are "smeared" out, appearing as a perfectly flat and static background.
Scale
Length (Approx.)
Visual Characteristics
Macroscopic
> 10^{-6} m
Smooth, flat, and featureless
Smooth, flat, and featureless > 10^{-6} m
Mesoscopic Scale:
Field Wriggles
At subatomic scales (10^{-8} to 10^{-19}m), fields exhibit ripples and "wriggles" that generate measurable effects like the Casimir force.
Mesoscopic
10^{-19} m
Ripples, field wriggles, and polarization
Ripples, field wriggles, and polarization 10^{-19} m
The Casimir Effect:
Physical proof of vacuum pressure where virtual particles push uncharged conductive plates together at submicron scales.
Virtual Particle Liberation:
High-energy collisions can "liberate" these fluctuations, transforming virtual pairs into real matter and radiation.
Turbulent foam, bubbles, and wormholes 10^{-35} m
Planck Scale:
The Quantum Foam
At 10^{-35}m, the uncertainty principle causes space-time to become a seething sea of virtual particles and "bubbles."
Planck Scale
10^{-35} m
Turbulent foam, bubbles, and wormholes
NotebookLM

Figure 11.34: **Conceptual (artistic) representation of vacuum quantum fluctuations (source: ?)**

However (!!!) the above "time-energy uncertainty relation" (and other time-"observable" uncertainty relations that can be constructed) is (considered) not to have same meaning as previous "canonical uncertainty relations". Meaning uncertainty relations constructed from canonical dynamical variables/observables (in the Hamiltonian sense), like position and momentum, since time parameter is not an observable and also not an operator in

Quantum Mechanics or Quantum Field Theory formalisms.

In fact, there are various approaches and interpretations of time-energy uncertainty. For example:

- Energy dispersion (ΔE) of a state and lifetime (Δt) of the state itself
- Energy exchange (ΔE) and time-frame (Δt) during which this can happen
- Energy measurement (ΔE) and time (Δt) it needs for accuracy (although this is rigorously disputed)

...other similar or specialized formulations of the above.

2. Secondly, these relations are remarkable because uncertainty is a complex value. This leads us to consider that the set $\mathbb{C}$ of complexes number is inherent to the real structure of our Universe (space-time) at the level of the quantum world (as confirmed by the Casimir effect where the sum to infinity of integers prolongation with the Riemann Zeta function is equal to $-1/12$ as seen in the section of Sequences and Series at page 1683). The quantum world is therefore a world of complex uncertainty. And this probability does not seem to be a consequence of our vagueness or ignorance but seems to be an intrinsic property of Nature!

Caution!!!! We can say that the quantum fluctuations are caused by the quantum zero-point energy, although that's somewhat tautological. Quantum fluctuations exist on the paper because quantum theory is a wave theory, from which the Heisenberg uncertainty principle is a natural consequence. There is no further understanding of how this came to be! Because quantum theory was developed to describe certain phenomena, and thus is a descriptive framework, not an explanatory framework!

It is also interesting to notice that time plays an unusual and subtle role in quantum physics. Unlike position, time is not usually treated as an operator; rather it is a simple parameter!

Remark

The relations and properties of commutators and anti-commutators will be indispensable to develop the quantified theory of angular momentum and spin further below.

11.2.3.1.3 Representatives

Let us now introduce contemporary quantum notations, which we will consider for now as abbreviations of integrals relating to wave functions, we will write (for the future purpose of calculating probability densities):

$$\langle\psi\,|\,\chi\rangle=\int\bar{\psi}\cdot\chi\mathrm{d}\vec{x}\qquad\text{and}\qquad\langle\psi\,|\,\alpha\,|\,\chi\rangle=\int\bar{\psi}(\alpha\cdot\chi)\mathrm{d}\vec{x}\tag{11.2.120}$$

since it is a complex functional dot product (see section Functional Analysis page 2379 and Vector Calculus page 1821).

With this notation, the relation that we presented earlier in our study of operators:

$$\int \bar{\psi}\cdot(\alpha\cdot\chi)\mathrm{d}\vec{x}=\overline{\left[\int \bar{\chi}\cdot(\alpha^{\dagger}\cdot\psi)\mathrm{d}\vec{x}\right]} \tag{11.2.121}$$

becomes (it is lighter already ... but less pedagogical):

$$\langle\psi\,|\,\alpha\,|\,\chi\rangle=\overline{\langle\chi\,|\,\alpha^{\dagger}\,|\,\psi\rangle} \tag{11.2.122}$$

That said, the set E of the functions ψ that interest us in Wave Quantum Physics constitutes a linear functional space. Indeed, in Quantum Physics, the differential equations that we have to solve (Schrödinger's equation) to describe the behaviour of a particle are such that the general solution can be very often decomposed into the sum of the particular solutions (we will prove this later!). In mathematics, we say that "the states are linear", that is, every combination of states is still a state!

Thus, the state of a particle is, as will be proved later, represented by a "quantum state" or a "state vector" denoted $|\Psi\rangle$ which also corresponds to a mathematical function describing it completely.

For example, if $|\Psi_1\rangle$ and $|\Psi_2\rangle$ are two possible states, then:

$$|\,\Psi_1\rangle+|\,\Psi_2\rangle \tag{11.2.123}$$

is also a possible state for the system (due to the property of linear functional spaces).

Let us now return to our linear functional space (or "linear space of states"). The fact that the set E of the functions ψ that interest us constitutes a "linear functional space" means that if $\psi\in E$, $\chi\in E$ we have:

$$(\lambda\psi+\mu\chi)\in E \tag{11.2.124}$$

regardless of the coefficients λ and μ, and also that:

$$\langle\varphi_i\,|\,\varphi_j\rangle=\delta_{ij} \tag{11.2.125}$$

where for recall, δ_{ij} is the Kronecker symbol (see section Tensor Calculus page 2044).

Definition (#11.2.506): The base is name a "complete base" if of course any $\psi\in E$ can be develop in series of eigenfunctions such as:

$$\psi=\sum_i c_i\Phi_i \tag{11.2.126}$$

where c_i is an arbitrary number (it is in part here that we have to return back to the 4th and 5th postulates of Wave Quantum Physics).

Let us now calculate the functional scalar product (see section Functional Analysis page 2379):

$$\int\bar{\varphi}_j\cdot\psi\mathrm{d}\vec{x}=\langle\varphi_i\,|\,\psi\rangle=\int\bar{\varphi}_j\cdot\sum_i c_i\varphi_i\mathrm{d}\vec{x}=\sum_i c_i\int\bar{\varphi}_j\cdot\varphi_i\mathrm{d}\vec{x}=\sum_i c_i\delta_{ij}=c_j \tag{11.2.127}$$

This last relation shows that we have identically (we change the notation of the indices):

$$\langle\varphi_j\,|\,\psi\rangle=\sum_i c_i\delta_{ij} \tag{11.2.128}$$

Thus, in a complete orthonormal basis $(\varphi_i$, an function ψ will be well described by the data of the coefficients c_i. It will often be useful to put them in the format of the representative matrix of ψ in the basis (φ_i):

$$\begin{bmatrix} c_1\\ c_2\\ \ldots\\ \ldots\\ c_i \end{bmatrix} \tag{11.2.129}$$

Let us consider now an operator α such that:

$$\chi=\alpha\cdot\psi=\alpha\cdot\sum_i c_i\varphi_i \tag{11.2.130}$$

But we can also write (notice the apostrophe in the relation!):

$$\chi=\sum_i\alpha\cdot c_i\cdot\varphi_i=\sum_i c_i'\cdot\varphi_i \tag{11.2.131}$$

Let us multiply this last relation by $\bar{\varphi}_j$ and calculate the functional scalar product:

$$\int\bar{\varphi}_j\cdot\chi\mathrm{d}\vec{x}=\int\bar{\varphi}_j\cdot\sum_i c_i'\cdot\varphi_i\mathrm{d}\vec{x}=\sum_i c_i'\int\bar{\varphi}_j\cdot\alpha\cdot\varphi_i\mathrm{d}\vec{x}=\sum_i c_i'\langle\varphi_j\,|\,\alpha\,|\,\varphi_i\rangle=\sum_i c_i'A_{ij} \tag{11.2.132}$$

To be compared with (obtained above):

$$\langle\varphi_j\,|\,\psi\rangle=\sum_i c_i\delta_{ij} \tag{11.2.133}$$

By noting A_{ij}, the "representative matrix" of α in the basis φ_i, we can thanks to the relation:

$$\langle\psi\,|\,\alpha\,|\,\chi\rangle=\overline{\langle\chi\,|\,\alpha^\dagger\,|\,\psi\rangle} \tag{11.2.134}$$

write finally:

$$\langle\varphi_j\,|\,\alpha\,|\,\varphi_i\rangle=A_{ij}=\overline{\langle\varphi_i\,|\,\alpha^\dagger\,|\,\varphi_j\rangle}=\bar{A}_{ij} \tag{11.2.135}$$

11.2.3.1.4 Eigenvalues and Eigenfunctions

Given an operator α (hermitian or not). The number a is named the "eigenvalue of the operator" of α, if there exists a function φ not identically zero such that (for a reminder of similar notions see the section of Linear Algebra):

$$\alpha\cdot\varphi=a\cdot\varphi \tag{11.2.136}$$

$\varphi$is then say to be the "eigenfunction" (in analogy with the "eigenvectors") of α, associated with the eigenvalue of a. Notice that a may very well be equal to zero (you will understand this better when we move on to the study of concrete cases).

In more physical terms, this means that when a state (a mathematical function in the formal sense such as φ) is unchanged by an operator, then the state is named an "eigenstate" or "eigenvector" of the system.

Let us consider the set E_a of eigenfunctions associated to a and a functional linear space which we will name "sub-eigenspace" associated to a. The number of dimensions of E_a is named the "multiplicity" or "degeneracy order" of the eigenvalue a, and we denote it g.

Let us consider now a a being a simple eigenvalue, or non degenerated: $g=1$. This means that there is only one eigenfunction associated with a, with a non-zero multiplicative coefficient.

If $g=2$ (double eigenvalue), we can find two non-proportional eigenfunctions (unbounded) associated with a, etc.

✍**Example:**

Let us see a particular example of an eigenfunction with an eigenvalue other than the classical case using the Energy.

Given:

$$\alpha\cdot\Psi=a\cdot\Psi \tag{11.2.137}$$

with $\alpha=-\mathrm{i}\hbar\dfrac{\partial}{\partial x}$ (operator that we have already seen previously) and a an eigenvalue.

The equation becomes:

$$-\mathrm{i}\hbar\frac{\partial\Psi}{\partial x}=a\cdot\Psi \tag{11.2.138}$$

which is easily verified if:

$$\boxed{\Psi=e^{\frac{\mathrm{i}ax}{k}}} \tag{11.2.139}$$

which is indeed an eigenfunction of the previously mentioned operator and which will be useful to us in the following developments!

11.2.3.1.4.1 Orthogonality of eigenfunctions

Two eigenfunctions (eigenvectors) φ and φ' associated with two different eigenvalues of the same Hermitian operator are orthogonal, that is to say:

$$\langle\varphi\,|\,\varphi'\rangle=0 \tag{11.2.140}$$

Proof 11.4.2. Let us start this time first with the Dirac notation with two eigenfunctions and two associated eigenvalues:

$$\begin{gathered}\alpha\cdot\varphi=a\cdot\varphi\\ \alpha\varphi'=a'\cdot\varphi'\end{gathered} \tag{11.2.141}$$

with $a\neq a'$.

We multiply the two previous relations by $\bar{\varphi}'$ and $\bar{\varphi}$ and we integrate to get the functional scalar product:

$$\begin{aligned}\langle\varphi'\,|\,\alpha\,|\,\varphi\rangle &= a\langle\varphi'\,|\,\varphi\rangle \qquad (a)\\ \langle\varphi\,|\,\alpha\,|\,\varphi'\rangle &= a'\langle\varphi\,|\,\varphi'\rangle \qquad (b)\end{aligned} \tag{11.2.142}$$

Let us recall to continue that we have proved earlier above that:

$$\langle\psi\,|\,\alpha\,|\,\chi\rangle=\overline{\langle\chi\,|\,\alpha^\dagger\,|\,\psi\rangle} \tag{11.2.143}$$

So if the operator α is self-adjoint (which is the case of the Hamiltonian as we have shown it), that is to say that $\alpha=\alpha^\dagger$, we have:

$$\langle\psi\,|\,\alpha\,|\,\chi\rangle=\overline{\langle\chi\,|\,\alpha\,|\,\psi\rangle} \tag{11.2.144}$$

Hence, by subtracting from the relation (11.2.142a) the conjugate complex of relation (11.2.142b), the eigenvalue a being assumed real (or an integer...), we have:

$$\begin{gathered}\langle\varphi'\,|\,\alpha\,|\,\varphi\rangle-\overline{\langle\varphi\,|\,\alpha\,|\,\varphi'\rangle}=a\langle\varphi'\,|\,\varphi\rangle-a'\overline{\langle\varphi\,|\,\varphi'\rangle}\\ \langle\varphi'\,|\,\alpha\,|\,\varphi\rangle-\langle\varphi'\,|\,\alpha\,|\,\varphi\rangle=0=a\langle\varphi'\,|\,\varphi\rangle-a'\langle\varphi'\,|\,\varphi\rangle=(a-a')\langle\varphi'\,|\,\varphi\rangle\end{gathered} \tag{11.2.145}$$

which proves indeed that:

$$\langle\varphi'\,|\,\varphi\rangle=0 \tag{11.2.146}$$

since $a\neq a'$.

Let us now make the same proof with but using the classical algebraic notation and a different approach. We start from:

$$\begin{aligned}\int\bar{\varphi}'\alpha\varphi dV &= a\int\bar{\varphi}'\varphi\mathrm{d}V \qquad (a)\\ \int\bar{\varphi}\alpha\varphi' dV &= a'\int\bar{\varphi}\varphi'\mathrm{d}V \qquad (b)\end{aligned} \tag{11.2.147}$$

and as we have proved earlier above that:

$$\overline{\int\bar{\varphi}'\alpha\varphi\mathrm{d}V}=\int\bar{\varphi}\alpha^\dagger\varphi'\mathrm{d}V \tag{11.2.148}$$

So if the operator α is self-adjoint (which is the case of the Hamiltonian as we have shown it), that is to say that $\alpha=\alpha^{\dagger}$, we have:

$$\overline{\int \bar{\varphi}'\alpha\varphi\mathrm{d}V}=\int \bar{\varphi}\alpha\varphi'\mathrm{d}V \tag{11.2.149}$$

It follows that the two relations (11.2.147a) and (11.2.147b) are equal (do not forget that the eigenvalue is assumed to be real!). As we can write:

$$a\int \bar{\varphi}'\varphi\mathrm{d}V=a'\int \bar{\varphi}\varphi'\mathrm{d}V \tag{11.2.150}$$

We can then write:

$$(a-a')\int \bar{\varphi}'\varphi\mathrm{d}V=0 \tag{11.2.151}$$

So if the eigenvalues are not equal it is forced that it is the integral that is zero (the functional scalar product) and therefore that the two eigenfunctions are orthogonal!

■ Q.E.D.

The same proof, but with the traditional and more pedagogical notation gives:

$$\begin{gathered}H\Psi_1=E_1\Psi_1\\ H\Psi_2=E_2\Psi_2\end{gathered} \tag{11.2.152}$$

If we multiply the first equation on the left by $\bar{\Psi}_2$, and the second equation by $\bar{\Psi}_1$, and that we integrate over the totality of space, we get the following two expressions (corresponding to mean values):

$$\begin{gathered}\int\limits_{-\infty}^{+\infty}\bar{\Psi}_2 H\Psi_1\mathrm{d}x\mathrm{d}y\mathrm{d}z=\int\limits_{-\infty}^{+\infty}\bar{\Psi}_2 E_1\Psi_1\mathrm{d}x\mathrm{d}y\mathrm{d}z=E_1\int\limits_{-\infty}^{+\infty}\bar{\Psi}_2\Psi_1\mathrm{d}x\mathrm{d}y\mathrm{d}z\\ \int\limits_{-\infty}^{+\infty}\bar{\Psi}_1 H\Psi_2\mathrm{d}x\mathrm{d}y\mathrm{d}z=\int\limits_{-\infty}^{+\infty}\bar{\Psi}_1 E_1\Psi_2\mathrm{d}x\mathrm{d}y\mathrm{d}z=E_2\int\limits_{-\infty}^{+\infty}\bar{\Psi}_1\Psi_2\mathrm{d}x\mathrm{d}y\mathrm{d}z\end{gathered} \tag{11.2.153}$$

If we take the case of real functions, we can write:

$$\begin{gathered}\int\limits_{-\infty}^{+\infty}\Psi_2 H\Psi_1\mathrm{d}x\mathrm{d}y\mathrm{d}z=E_1\int\limits_{-\infty}^{+\infty}\Psi_2\Psi_1\mathrm{d}x\mathrm{d}y\mathrm{d}z\\ \int\limits_{-\infty}^{+\infty}\Psi_1 H\Psi_2\mathrm{d}x\mathrm{d}y\mathrm{d}z=E_2\int\limits_{-\infty}^{+\infty}\Psi_1\Psi_2\mathrm{d}x\mathrm{d}y\mathrm{d}z\end{gathered} \tag{11.2.154}$$

Since the operator H is Hermitian (self-adjoint) as we have proved it above, we have:

$$\overline{\int\limits_{-\infty}^{+\infty}\Psi_1 H\Psi_2\mathrm{d}x\mathrm{d}y\mathrm{d}z}=\int\limits_{-\infty}^{+\infty}\Psi_2 H\Psi_1\mathrm{d}x\mathrm{d}y\mathrm{d}z \tag{11.2.155}$$

and as E_2, Ψ_1, Ψ_2 are admitted as real functions, we also have:

$$\overline{E_2\int_{-\infty}^{+\infty}\Psi_1\Psi_2 \mathrm{d}x\mathrm{d}y\mathrm{d}z}=E_2\int_{-\infty}^{+\infty}\Psi_1\Psi_2 \mathrm{d}x\mathrm{d}y\mathrm{d}z=E_2\int_{-\infty}^{+\infty}\Psi_2\Psi_1 \mathrm{d}x\mathrm{d}y\mathrm{d}z \tag{11.2.156}$$

Therefore:

$$\int_{-\infty}^{+\infty}\Psi_2 H\Psi_1 \mathrm{d}x\mathrm{d}y\mathrm{d}z-\overline{\int_{-\infty}^{+\infty}\Psi_1 H\Psi_2 \mathrm{d}x\mathrm{d}y\mathrm{d}z}=E_1\int_{-\infty}^{+\infty}\Psi_2\Psi_1 \mathrm{d}x\mathrm{d}y\mathrm{d}z-\overline{E_2\int_{-\infty}^{+\infty}\Psi_1\Psi_2 \mathrm{d}x\mathrm{d}y\mathrm{d}z} \tag{11.2.157}$$

can be written:

$$\int_{-\infty}^{+\infty}\Psi_2 H\Psi_1 \mathrm{d}x\mathrm{d}y\mathrm{d}z-\int_{-\infty}^{+\infty}\Psi_2 H\Psi_1 \mathrm{d}x\mathrm{d}y\mathrm{d}z=E_1\int_{-\infty}^{+\infty}\Psi_2\Psi_1 \mathrm{d}x\mathrm{d}y\mathrm{d}z-E_2\int_{-\infty}^{+\infty}\Psi_2\Psi_1 \mathrm{d}x\mathrm{d}y\mathrm{d}z \tag{11.2.158}$$

Therefore it comes:

$$(E_1-E_2)\int_{-\infty}^{+\infty}\Psi_2\Psi_1 \mathrm{d}x\mathrm{d}y\mathrm{d}z=0 \tag{11.2.159}$$

which shows well that Ψ_1, Ψ_2 are orthogonal according to the definition of the functional scalar product.

11.2.3.1.5 Dirac formalism

Dirac has conceived a very practical general formalism, world-wide used by physicists, of which we will give the essential elements. The notations used have already been partially introduced in the preceding paragraphs.

We will use the Dirac formalism for two points, the first being to better understand what has been seen so far when introducing to functional operators, the second being to introduce a notation and a method of solving that we can find in some textbooks. Moreover, in this textbook by simplification of writing, we will sometimes use this formalism.

11.2.3.1.5.1 Kets and Bras

We consider a vector space E_n with n dimensions where n can very well be infinite (Hilbert space). A vector is defined by n components x_i with $i=1,2,\ldots,n$ which we can store in column to form a column-matrix:

$$\begin{vmatrix} x_1 \\ x_2 \\ \ldots \\ x_n \end{vmatrix} \tag{11.2.160}$$

We will say that this matrix describes the "right vector" or the "ket $|x\rangle$" (this must remind the "representatives"). It is possible to associate with the column-matrix the adjoint matrix (conjugated transpose for recall...):

$$\overline{\bar{x}_1\,\bar{x}_2\,\ldots\,\bar{x}_n} \tag{11.2.161}$$

where the $\bar{x}_i$ are the conjugate complexes of the x_i. We will say that the adjoint line matrix describes the "left vector" or the "bra $\langle x|$" (this must also remind you the "representatives").

The addition and multiplication by a number λ are evident. Notice that if:

$$|y\rangle=\lambda|x\rangle \tag{11.2.162}$$

we have trivially:

$$\langle y|=\bar{\lambda}\langle x| \tag{11.2.163}$$

With two vectors of components x_i and y_i, we can then form the following quantity, named "Hermitian scalar product" or "Hermitian inner product":

$$\overline{\bar{x}_1\,\bar{x}_2\,\ldots\,\bar{x}_n}\begin{vmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{vmatrix}=\sum_i \bar{x}_i y_i \tag{11.2.164}$$

We agree to write it $\langle x|y\rangle$. Notice that:

$$\langle x\,|\,y\rangle=\overline{\langle y\,|\,x\rangle} \tag{11.2.165}$$

the Hermitian scalar product is therefore not simply commutative!

The product $\langle x|y\rangle$ depends linearly on $\langle x|$ and $|y\rangle$. Conversely, if a number Q depends linearly on a ket $|x\rangle$, there exists a bra $\langle a|$ such that:

$$Q=\langle a\,|\,x\rangle \tag{11.2.166}$$

In quantum physics, $\langle x|y\rangle$ is associated to the "amplitude" of being in the state x if the system is in the state y. This Hermitian scalar product will be interpreted as the probability that the physical system is projected into the state x if it is in the state y.

An orthonormal basis of the studied space is constituted by n vectors $|i\rangle$ such that (see section Vector Calculus page 1833):

$$\langle i\,|\,j\rangle=\delta_{ij} \tag{11.2.167}$$

where for recall, δ_{ij} is the Kronecker symbol (see section Tensor Calculus page 2044).

Any vector (or function if we generalize vector space to function space) $|x\rangle$ of E_n can be developed on this basis according to (see section Vector Calculus page 1806):

$$|x\rangle=\sum_i x_i|i\rangle \tag{11.2.168}$$

where the x_i are the components of $|x\rangle$ in the chosen base. We can easily verify that (already seen many times in the section of Vector Calculus):

$$x_i = \langle i|x\rangle \tag{11.2.169}$$

If a ket $|y\rangle$ depends linearly on a ket $|x\rangle$, we write symbolically:

$$|\,y\rangle = \alpha\,|\,x\rangle \tag{11.2.170}$$

where α is a linear operator.

Given a linear operator defined by the preceding relation and a bra $\langle u|$, the Hermitian scalar product:

$$\langle u\,|\,y\rangle = \langle u\,|\,\big[\alpha\,|\,x\rangle\big]\rangle = Q \tag{11.2.171}$$

is a number Q which depends linearly on $|x\rangle$. From what has been seen above, there is a bra $\langle v|$ such as $Q = \langle v|x\rangle$. Then $\langle v|$ obviously depends of $\langle u|$ linearly. We agree to put:

$$\langle v\,| = \langle u\,|\,\alpha \tag{11.2.172}$$

Using this convention we can write:

$$\langle u\,|\,\big[\alpha\,|\,x\rangle\big]\rangle = \big[\langle u\,|\,\alpha\big]\,|\,x\rangle = \langle u\,|\,\alpha\,|\,x\rangle \tag{11.2.173}$$

If $|y\rangle = \alpha|x\rangle$, depends linearly on $\langle x|$. By definition, we will write:

$$\langle y\,| = \langle x\,|\,\alpha^\dagger \tag{11.2.174}$$

where for recall $\alpha^\dagger$ is the adjoint of α.

Let us form with a bra $\langle u|$ the Hermitian scalar product:

$$\langle u\,|\,y\rangle = \langle u\,|\,\alpha\,|\,x\rangle \tag{11.2.175}$$

and we can write (as we have proved it previously):

$$\overline{\langle u\,|\,y\rangle} = \langle y\,|\,u\rangle = \langle x\,|\,\alpha^\dagger\,|\,u\rangle \tag{11.2.176}$$

Hence the relation of the first importance which we have already met several times without really explaining its origin:

$$\boxed{\langle x\,|\,\alpha^\dagger\,|\,u\rangle = \overline{\langle u\,|\,\alpha\,|\,x\rangle}} \tag{11.2.177}$$

We simply recall with this relation that a Hermitian (self-adjoint) operator is an operator equal to its adjoint.

Thanks to Dirac's formalism, what was before abstract definitions have now become more obvious!

To summarize:

Classic formalism	Dirac formalism
$\int \bar{\varphi}_i\varphi_j \mathrm{d}V$	$\langle\varphi_i\vert\varphi_j\rangle$
$\int \bar{\varphi}_i\alpha\varphi_j \mathrm{d}V$	$\langle\varphi_i\vert\alpha\vert\varphi_j\rangle$
$\overline{\int \bar{\varphi}_i\varphi_j \mathrm{d}V} = \int \bar{\varphi}_j\varphi_i \mathrm{d}V$	$\overline{\langle\varphi_i\vert\varphi_j\rangle} = \langle\varphi_j\vert\varphi_i\rangle$
$\overline{\int \bar{\varphi}_i\alpha\varphi_j \mathrm{d}V} = \int \bar{\varphi}_j\alpha^\dagger\varphi_i \mathrm{d}V = \int \bar{\varphi}_j\alpha\varphi_i \mathrm{d}V$	$\overline{\langle\varphi_i\vert\alpha\vert\varphi_j\rangle} = \langle\varphi_j\vert\alpha\vert\varphi_i\rangle$
$\Psi = \sum\limits_i c_i\varphi_i$	$\vert\Psi\rangle = \sum_i c_i\vert\varphi_i\rangle$
$\sum\limits_i c_i' \int \bar{\varphi}_j\alpha\varphi_i \mathrm{d}\vec{x}$	$\sum\limits_i c_i'\langle\varphi_i\vert\psi\rangle$
$c_i = \int \bar{\varphi}_j\psi \mathrm{d}V$	$c_i = \langle\varphi_i\vert\psi\rangle$

Table 11.7: **Comparison between Classic and Dirac formalism**

So the reader has probably noticed that Dirac bra-kets and their algebra are very much like the Linear Algebra.

A ket is like a vector, a bra is like the conjugate transpose of a vector, a bra-ket is like a complex inner product, a ket-bra is like an outer product, and operator is like a matrix, and operator acting on a state is like matrix product. Even we speak about eigenstates which are very much like eigenvectors.

So how did it happened that we invented a new notation for Linear Algebra instead of sticking to the usual notation of vectors and matrices as usual?

In fact it is not that it's "like Linear Algebra" but it is exactly Linear Algebra!!!! At the time of Dirac Linear Algebra was not taught to all undergraduates[8]. So he invented his own notation. Physicists found them convenient for the questions they consider. The clear evidence of this is that Heisenberg had to invent matrix multiplication when he created quantum mechanics... Actually there is probably a causal relation between the invention of quantum mechanics and the spread of linear algebra as a part of the standard curriculum.

Remark

Again, an excellent practical example of the application of the Dirac formalism for engineers is proposed in the section on Quantum Computing (see page 7263).

---

[8] In the first decade (when they were all students) of 120th century (holocene calendar) matrix multiplication was not taught to students on a regular basis in the best European universities.

11.2.3.2 Robertson uncertainty relations

The reader remember maybe that we have derived earlier (see page 4704) the classical Heisenberg's uncertainty relations (and also based on the wave-packet approach at page 4791). However, now that we have a better mathematical quantum toolbox, we can go a bit further! Indeed, the most common general form of the classical Heisenberg's uncertainty principle is the Robertson uncertainty relation established in 11929 (holocene calendar) by... Howard P. Robertson that we will derive now.

For an arbitrary Hermitian operator $\hat{\mathcal{O}}$ we can associate a standard deviation:

$$\sigma_{\hat{\mathcal{O}}}^2=\left\langle\left(\hat{\mathcal{O}}-\langle\hat{\mathcal{O}}\rangle\right)^2\right\rangle \tag{11.2.178}$$

where as we know it already, the brackets $\langle\hat{\mathcal{O}}\rangle$ indicate an expectation value. For recall, the expected mean of $\hat{\mathcal{O}}$ in the state ψ is defined as:

$$\langle\hat{\mathcal{O}}\rangle_\Psi=\langle\Psi|\hat{\mathcal{O}}|\Psi\rangle=\int\limits_{-\infty}^{+\infty} x\,|\Psi(x)|^2\mathrm{d}x \tag{11.2.179}$$

So we have therefore:

$$\langle\sigma_{\hat{\mathcal{O}}}^2\rangle_\Psi=\langle\Psi|\hat{\mathcal{O}}|\Psi\rangle=\int\limits_{-\infty}^{\infty} x\,|\Psi(x)|^2\mathrm{d}x \tag{11.2.180}$$

For any Hermitian operator $\hat{A}$, based upon the definition of variance, we have:

$$\sigma_{\hat{A}}^2=\langle(\hat{A}-\langle\hat{A}\rangle)\Psi|(\hat{A}-\langle\hat{A}\rangle)\Psi\rangle \tag{11.2.181}$$

we let $|f\rangle=|(\hat{A}-\langle\hat{A}\rangle)\Psi\rangle$ and thus:

$$\sigma_{\hat{A}}^2=\langle f\,|\,f\rangle \tag{11.2.182}$$

Similarly, for any other Hermitian operator $\hat{B}$ in the same state:

$$\sigma_{\hat{B}}^2=\langle(\hat{B}-\langle\hat{B}\rangle)\Psi|(\hat{B}-\langle\hat{B}\rangle)\Psi\rangle=\langle g\,|\,g\rangle \tag{11.2.183}$$

for $|g\rangle=|(\hat{B}-\langle\hat{B}\rangle)\Psi\rangle$.

The product of the two deviations can thus be expressed as:

$$\sigma_A^2\sigma_B^2=\langle f\,|\,f\rangle\langle g\,|\,g\rangle \tag{11.2.184}$$

In order to relate the two vectors $|f\rangle$ and $|g\rangle$, we use the Cauchy–Schwarz integral inequality (see section Functional Analysis page 2388) which is defined for recall as:

$$\langle f\,|\,f\rangle\langle g\,|\,g\rangle\geq|\langle f\,|\,g\rangle|^2 \tag{11.2.185}$$

and thus the prior-previous can be written as:

$$\sigma_{\hat{A}}^2\sigma_{\hat{B}}^2\geq|\langle f\,|\,g\rangle|^2 \tag{11.2.186}$$

Since $\langle f\,|\,g\rangle$ is in general a complex number, we use the fact that the modulus squared of any complex number z is defined as $|z|^2=z\bar{z}$, where $\bar{z}$ is the complex conjugate of z. The modulus squared can also be expressed as:

$$|z|^2=(\Re(z))^2+(\Im(z))^2=\left(\frac{z+\bar{z}}{2}\right)^2+\left(\frac{z-\bar{z}}{2\mathrm{i}}\right)^2 \tag{11.2.187}$$

we let $z=\langle f\,|\,g\rangle$ and $\bar{z}=\langle g\,|\,f\rangle$ and substitute these into the equation above to get:

$$|\langle f\,|\,g\rangle|^2=\left(\frac{\langle f\,|\,g\rangle+\langle g\,|\,f\rangle}{2}\right)^2+\left(\frac{\langle f\,|\,g\rangle-\langle g\,|\,f\rangle}{2\mathrm{i}}\right)^2 \tag{11.2.188}$$

The inner product $\langle f\,|\,g\rangle$ is written out explicitly as:

$$\langle f\,|\,g\rangle=\langle(\hat{A}-\langle\hat{A}\rangle)\Psi|(\hat{B}-\langle\hat{B}\rangle)\Psi\rangle \tag{11.2.189}$$

and using the fact that $\hat{A}$ and $\hat{B}$ are Hermitian operators, we find:

$$\begin{aligned}\langle f\,|\,g\rangle&=\langle\Psi|(\hat{A}-\langle\hat{A}\rangle)(\hat{B}-\langle\hat{B}\rangle)\Psi\rangle\\&=\langle\Psi\,|\,(\hat{A}\hat{B}-\hat{A}\langle\hat{B}\rangle-\hat{B}\langle\hat{A}\rangle+\langle\hat{A}\rangle\langle\hat{B}\rangle)\Psi\rangle\\&=\langle\Psi\,|\,\hat{A}\hat{B}\Psi\rangle-\langle\Psi\,|\,\hat{A}\langle\hat{B}\rangle\Psi\rangle-\langle\Psi\,|\,\hat{B}\langle\hat{A}\rangle\Psi\rangle+\langle\Psi\,|\,\langle\hat{A}\rangle\langle\hat{B}\rangle\Psi\rangle\\&=\langle\hat{A}\hat{B}\rangle-\langle\hat{A}\rangle\langle\hat{B}\rangle-\langle\hat{A}\rangle\langle\hat{B}\rangle+\langle\hat{A}\rangle\langle\hat{B}\rangle\\&=\langle\hat{A}\hat{B}\rangle-\langle\hat{A}\rangle\langle\hat{B}\rangle\end{aligned} \tag{11.2.190}$$

Similarly it can be shown that $\langle g\,|\,f\rangle=\langle\hat{B}\hat{A}\rangle-\langle\hat{A}\rangle\langle\hat{B}\rangle$.

Thus we have:

$$\langle f\,|\,g\rangle-\langle g\,|\,f\rangle=\langle\hat{A}\hat{B}\rangle-\langle\hat{A}\rangle\langle\hat{B}\rangle-\langle\hat{B}\hat{A}\rangle+\langle\hat{A}\rangle\langle\hat{B}\rangle=\langle[\hat{A},\hat{B}]\rangle \tag{11.2.191}$$

and:

$$\langle f\,|\,g\rangle+\langle g\,|\,f\rangle=\langle\hat{A}\hat{B}\rangle-\langle\hat{A}\rangle\langle\hat{B}\rangle+\langle\hat{B}\hat{A}\rangle-\langle\hat{A}\rangle\langle\hat{B}\rangle=\langle\{\hat{A},\hat{B}\}\rangle-2\langle\hat{A}\rangle\langle\hat{B}\rangle \tag{11.2.192}$$

where we have introduced the anticommutator:

$$\{\hat{A},\hat{B}\}=\hat{A}\hat{B}+\hat{B}\hat{A} \tag{11.2.193}$$

We now substitute the above two relations above back into:

$$|\langle f\,|\,g\rangle|^2=\left(\frac{\langle f\,|\,g\rangle+\langle g\,|\,f\rangle}{2}\right)^2+\left(\frac{\langle f\,|\,g\rangle-\langle g\,|\,f\rangle}{2\mathrm{i}}\right)^2 \tag{11.2.194}$$

and get:

$$|\langle f\,|\,g\rangle|^2=\left(\frac{1}{2}\langle\{\hat{A},\hat{B}\}\rangle-\langle\hat{A}\rangle\langle\hat{B}\rangle\right)^2+\left(\frac{1}{2\mathrm{i}}\langle[\hat{A},\hat{B}]\rangle\right)^2 \tag{11.2.195}$$

Substituting the above into:

$$\sigma_{\hat{A}}^2\sigma_{\hat{B}}^2\geq|\langle f\,|\,g\rangle|^2 \tag{11.2.196}$$

we get the "Roberston-Schrödinger uncertainty relation":

$$\boxed{\sigma_{\hat{A}}\sigma_{\hat{B}} \geq \sqrt{\left(\frac{1}{2}\langle\{\hat{A},\hat{B}\}\rangle - \langle\hat{A}\rangle\langle\hat{B}\rangle\right)^2 + \left(\frac{1}{2\mathrm{i}}\langle[\hat{A},\hat{B}]\rangle\right)^2}} \tag{11.2.197}$$

Notice that in some textbooks, the relation:

$$|z|^2 = (\Re(z))^2 + (\Im(z))^2 = \left(\frac{z+\bar{z}}{2}\right)^2 + \left(\frac{z-\bar{z}}{2\mathrm{i}}\right)^2 \tag{11.2.198}$$

is simplified as we may have $\Re(z) \geq 0$:

$$|z|^2 \geq (\Im(z))^2 = \left(\frac{z-\bar{z}}{2\mathrm{i}}\right)^2 \tag{11.2.199}$$

Therefore the Roberston-Schrödinger uncertainty relation is simplified to:

$$\sigma_{\hat{A}}\sigma_{\hat{B}} \geq \sqrt{\left(\frac{1}{2\mathrm{i}}\langle[\hat{A},\hat{B}]\rangle\right)^2} \tag{11.2.200}$$

Hence (the vertical bars are the absolute value, not the modulus!):

$$\sigma_{\hat{A}}\sigma_{\hat{B}} \geq \left|\frac{1}{2\mathrm{i}}\langle[\hat{A},\hat{B}]\rangle\right| \tag{11.2.201}$$

✍**Example:**

For position and linear momentum, the canonical commutation relation:

$$[\hat{x},\hat{p}] = \mathrm{i}\hbar \tag{11.2.202}$$

we get:

$$\left|\frac{1}{2\mathrm{i}}\langle[\hat{A},\hat{B}]\rangle\right| = \left|\frac{1}{2\mathrm{i}}\langle\mathrm{i}\hbar\rangle\right| = \left|\frac{1}{2\mathrm{i}}\mathrm{i}\hbar\right| = \frac{1}{2}\hbar \tag{11.2.203}$$

Therefore:

$$\sigma_{\hat{p}}\sigma_{\hat{x}} \geq \frac{1}{2}\hbar \tag{11.2.204}$$

And we then get the same for energy and time:

$$\sigma_{\hat{E}}\sigma_t \geq \frac{1}{2}\hbar \tag{11.2.205}$$

Notice that this inequality is quite smaller (one order of magnitude) than the classical Heisenberg's uncertainty relations that we get before:

$$(\Delta x\Delta p = h) \gg \left(\sigma_{\hat{p}}\sigma_{\hat{x}} \geq \frac{1}{2}\hbar\right) \tag{11.2.206}$$

For energy and time, the relation:

$$\sigma_{\hat{E}}\sigma_t \geq \frac{1}{2}\hbar \tag{11.2.207}$$

is often found in various textbooks and videos under the following form (for an alternative derivation of his inequality see further below):

$$\boxed{\Delta t \Delta E \geq \frac{\hbar}{2} \cong 5.27286 \cdot 10^{-35}\,[\mathrm{J}\cdot\mathrm{s}]} \tag{11.2.208}$$

Caution!!!! Remember that the time-energy uncertainty relation (and other time-"observable" uncertainty relations that can be constructed) is (considered) not to have same meaning as canonical uncertainty relations. Meaning uncertainty relations constructed from canonical dynamical variables/observables (in the Hamiltonian sense), like position and momentum, since time parameter is not an observable and also not an operator in QM/QFT formalisms.

This is the relation that leads to the concept of "quantum vacuum fluctuation" also simply named "quantum fluctuation". The quantum fluctuations - i.e. zero-point fluctuations - have no cause[9] (not even with hidden variables) and are experimentally verified with:

- Speed of light (electric permittivity)
- Spontaneous light emission (Wigner-Weisskopf spontaneous decay)
- Radioactivity
- Casimir effect
- Lamb shift
- Anomalous electron g-factor
- Anomalous muon g-factor
- Josephson tunnel junctions
- False vacuum decay
- Vacuum polarization
- Vacuum birefringence
- Quantum critical point (quantum phase transition)
- Fine-structure constant variation
- LIGO Interferometer squeezed quantum vacuum
- Neutrino oscillation in vacuum
- Fulling–Davies–Unruh effect
- Spontaneous Parametric Down-Conversion (SPDC)

Let's us quote a two physicists on the topic (on of them is a Nobel Price in Physics):

---

[9] See the definition of "cause" in the section Mechanics at page 3206.

> ❞ *What causes the universe to pop out of nothing? No cause is needed. If you have a radioactive atom, it will decay, and quantum mechanics gives the decay probability in a given interval of time, say, a minute. There is no reason why the atom decayed at this particular moment and not another. The process is completely random. No cause is needed for the quantum creation of the universe.* ❞
>
> Alexander Vilenkin

quote 11.2:4736.#65

> ❞ *The discovery that individual events are irreducibly random is probably one of the most significant findings of the twentieth century. ... for the individual event in quantum physics, not only do we not know the cause, there is no cause! The instant when a radioactive atom decays, or the path taken by a photon behind a half-silvered beamsplitter are objectively random.* ❞
>
> Anton Zeilinger (Nobel Price)

quote 11.2:4737.#66

Or as stated by the philosopher S. Quentin in [499]:

quote
11.2:4738.#67

> *There is sufficient evidence at present to justify the belief that the universe began to exist without being caused to do so. This evidence includes the Hawking-Penrose singularity theorems that are based on Einstein's General Theory of Relativity, and the recently introduced Quantum Cosmological Models of the early universe.*
>
> Smith Quentin

The Heisenberg position-momentum uncertainty principle has obviously been tested and confirmed experimentally like in *On the experimental verification of the uncertainty principle of position and momentum* ([485]) where we have the following conclusive figure:

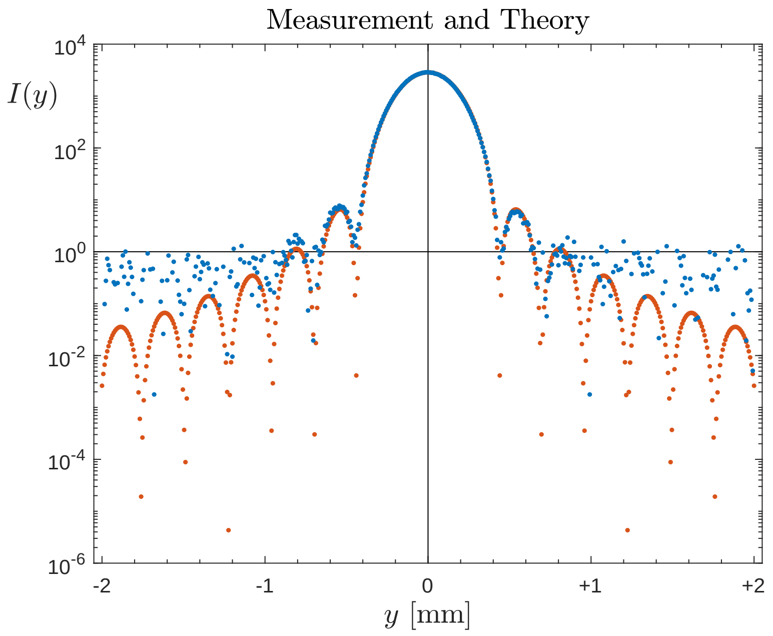


Figure 11.35: Measured (blue) versus theoretical (orange) density in semi-logarithmic representation. The signal-to-noise ratio outside the range of $y_c = \pm 1$ mm is almost zero and therefore of minor relevance for the computation of expectation values under consideration.

11.2.3.2.1 Energy-Time Heisenberg's uncertainty principle

The text that follows is almost a copy/paste of the content of the video on YouTube made by Sam Dietterich!

Under normal circumstances, when we say uncertainty in quantum mechanics, we mean the standard deviation. This can be calculated for a given quantum state and observable from the quantum state's wave function and the observable's operator. This is done using the following formula:

$$\sigma_A = \sqrt{\langle A^2 \rangle - \langle A \rangle^2} \qquad (11.2.209)$$

When considering the energy-time uncertainty principle, the situation is a little more complicated than this. While there is an energy operator (the Hamiltonian) that can be used to evaluate the above formula, there is no such thing as a time operator. When making a measurement of some observable, the naturally relevant concept of time uncertainty is the time it takes for the average (expectation value) $\langle Q \rangle$ of a given observable Q to change by one standard deviation σ_Q, given that the standard deviation is the uncertainty for that observable. So, we can make the following definition:

$$\Delta t := \frac{\sigma_Q}{|\mathrm{d}\langle Q \rangle / \mathrm{d}t|} \qquad (11.2.210)$$

Or put another way: the temporal uncertainty Δt is the number of times that we can put the variation of the temporal magnitude of the mean of the observable $\mathrm{d}\langle Q \rangle / \mathrm{d}t$ in a standard deviation of the latter σ_Q.

The reader that is careful will notice that this definition is made by anticipation of the result we want. Therefore it is a circular reasoning!

If we then take the definition of the energy uncertainty to be the usual one:

$$\Delta E = \sigma_H = \sqrt{\langle H^2 \rangle - \langle H \rangle^2} \qquad (11.2.211)$$

We can actually use the generalized Heisenberg uncertainty relation to construct an energy-time uncertainty relation. The generalized Heisenberg uncertainty relation is as proven earlier above:

$$\sigma_A \sigma_B \geq \left| \frac{1}{2\mathrm{i}} \langle [A, B] \rangle \right| \qquad (11.2.212)$$

Doing this requires the assistance of the generalized Ehrenfest theorem (see page 4874), which is for recall:

$$\frac{\mathrm{d}}{\mathrm{d}t} \langle Q \rangle = \frac{\mathrm{i}}{\hbar} \langle [H, Q] \rangle + \left\langle \frac{\partial Q}{\partial t} \right\rangle \qquad (11.2.213)$$

The energy time uncertainty principle is usually considered for cases where Q is not explicitly time dependent, so the Ehrenfest Theorem reduces to:

$$\frac{\mathrm{d}}{\mathrm{d}t} \langle Q \rangle = \frac{\mathrm{i}}{\hbar} \langle [H, Q] \rangle \qquad (11.2.214)$$

Rearranging:

$$\langle[H,Q]\rangle=\frac{\hbar}{\mathrm{i}}\frac{\mathrm{d}}{\mathrm{d}t}\langle Q\rangle \tag{11.2.215}$$

We can now insert this into the generalized uncertainty relation for the case of the Hamiltonian and the observable whose operator is Q :

$$\sigma_H\sigma_Q\geq\left|\frac{1}{2i}\frac{\hbar}{\mathrm{i}}\frac{\mathrm{d}}{\mathrm{d}t}\langle Q\rangle\right|=\frac{\hbar}{2}\left|\frac{\mathrm{d}}{\mathrm{d}t}\langle Q\rangle\right| \tag{11.2.216}$$

Hence:

$$\sigma_H\sigma_Q\geq\frac{\hbar}{2}\left|\frac{\mathrm{d}}{\mathrm{d}t}\langle Q\rangle\right| \tag{11.2.217}$$

We can then divide by the rate of change of the average:

$$\sigma_H\frac{\sigma_Q}{|\mathrm{d}\langle Q\rangle/\mathrm{d}t|}\geq\frac{\hbar}{2} \tag{11.2.218}$$

We thus have the famous energy-time uncertainty relation:

$$\boxed{\Delta E\Delta t\geq\frac{\hbar}{2}} \tag{11.2.219}$$

11.2.3.2.2 Zero-Point Energy from the uncertainty principles

It is a well-known deduction from quantum mechanics, confirmed by experimental data, that a minimum vibrational energy exists for an atomic oscillator undergoing simple harmonic vibratory motion (see further below page 4807). This minimum energy, or the zero-point vibrational energy is as we have already seen given by:

$$E_0=\frac{1}{2}\hbar\omega_0=\frac{1}{2}h\nu_0 \tag{11.2.220}$$

The value above which cannot so far be deduced from classical theory arises as we have already proved it from the solution of the Schrödinger equation given for reminder by:

$$\frac{\mathrm{d}^2\Psi}{\mathrm{d}x^2}+\frac{8\pi^2m}{h^2}\left(E_{\text{tot}}-E_{\text{p}}\right)\Psi=\frac{\mathrm{d}^2\Psi}{\mathrm{d}x^2}+\frac{8\pi^2m}{h^2}\left(E_{\text{tot}}-\frac{1}{2}kx^2\right)\Psi=0\,[\mathrm{m}^{-5/2}] \tag{11.2.221}$$

in which m is the mass of the oscillator, E_{tot} its total vibrational energy; and its potential energy is $\frac{1}{2}kx^2$, where k is the force constant and x the displacement from the equilibrium position x_0.

Some researchers who discuss this problem then go on to state that the zero-point energy, $\frac{1}{2}h\nu_0$, when $n=0$, is a necessary consequence of the uncertainty principle, because the position, x, and momentum, $p=mv$, cannot both be known exactly at the same instant; and therefore the particle in question cannot be merely standing at rest at $x=0$ and $p=0$.

We will now show that the zero-point vibrational energy, like the ground-energy and size of the H atom can be deduced quite simply, and without solving the Schrödinger equation, if we assume the uncertainty principle, in its most rigorous form, as a fundamental postulate!

The total energy of the oscillator is the sum of its potential and kinetic energies:

$$E_{\text{tot}} = E_p + E_c = \frac{1}{2}kx^2 + \frac{1}{2}mv^2 = \frac{1}{2}kx^2 + \frac{p^2}{2m} \tag{11.2.222}$$

It is seen that the total energy will be at its minimum value when x and p are as small as possible, or when the derivation of the (Robertson) uncertainty principle states that:

$$\Delta x \Delta p \geq \frac{h}{4\pi} \tag{11.2.223}$$

where Δx and Δp are, respectively, the uncertainty in position and momentum of the particle.

But the average values of the distance, $\bar{x}$, from equilibrium, and the momentum, $\bar{p}$, cannot be less than the uncertainty, Δx and Δp, in these quantities:

$$\begin{aligned} \bar{x} &\geq \Delta x \\ \bar{p} &\geq \Delta p \end{aligned} \tag{11.2.224}$$

The smallest average value of the product $\bar{x}\bar{p}$ is therefore:

$$\bar{x}\bar{p} = \Delta x \Delta p = \frac{h}{4\pi} \tag{11.2.225}$$

and:

$$\bar{p} = \frac{h}{4\pi\bar{x}} \tag{11.2.226}$$

If we substitute this value of $\bar{p}$ in the equation for the total energy we get:

$$E_{\text{tot}} = E_p + E_c = \frac{1}{2}kx^2 + \frac{p^2}{2m} = \frac{1}{2}k\bar{x}^2 + \frac{h^2}{32\pi^2 m\bar{x}^2} \tag{11.2.227}$$

Differentiating the above relation with respect to $\bar{x}$, and setting the derivative equal to zero, we get the value of $\bar{x}$ for which E_{tot} is a minimum:

$$\frac{\mathrm{d}E_{\text{tot}}}{\mathrm{d}\bar{x}} = 0 = k\bar{x} - \frac{h^2}{16\pi^2 m\bar{x}^3} \tag{11.2.228}$$

Therefore:

$$\bar{x}^2_{\text{min}} = \frac{h}{4\pi\sqrt{mk}} \tag{11.2.229}$$

The minimum E_{tot} is obtained by substituting this value of $\bar{x}^2_{\text{min}}$ in the total energy relation:

$$\begin{aligned} E_{\text{min}} &= \frac{1}{2}kx^2 + \frac{p^2}{2m} = \frac{1}{2}k\frac{h}{4\pi\sqrt{mk}} + \frac{h^2}{32\pi^2 m}\frac{4\pi\sqrt{mK}}{h} \\ &= \frac{h}{8\pi}\sqrt{\frac{k}{m}} + \frac{h}{8\pi}\sqrt{\frac{k}{m}} = \frac{h}{4\pi}\sqrt{\frac{k}{m}} \end{aligned} \tag{11.2.230}$$

For the harmonic oscillator the force constant is related to the vibration frequency by (see section Social Sciences page 8224):

$$2\pi\nu = \sqrt{\frac{k}{m}} \tag{11.2.231}$$

and therefore:

$$E_{\min}=\left(\frac{h}{4\pi}\right)\cdot(2\pi\nu)=\frac{1}{2}h\nu \tag{11.2.232}$$

which is the zero-point energy! For the anharmonic oscillator Heisenberg also derived a more complicated expression, also involving a zero-point energy.

If a harmonic oscillator were to have zero energy, both its potential and kinetic energy would have to be zero. The case $E_p=\frac{1}{2}kx^2=0$ would correspond to a precise knowledge of the position of the particle $(x=0)$ and thus imply that the momentum p is completely uncertain. However, then the mean value of $E_c=p^2/2m$ would be infinite. Conversely, for E_c to be zero, E_p would have to be infinite.

Definition (#11.2.507): As absolute zero is the point at which a quantum system has minimal vibrational motion, retaining only quantum mechanical zero-point energy (ZPE)-induced particle motion, we define usually in quantum physics "nothing" as the smallest relative energy to the zero-point energy that satisfies $\Delta E_0\Delta t\geq {}^1\!/\!{}_2\hbar$ for a given quantum system (remember that energy is defined by a "quantity which is conversed or not depending on invariance under time translations" by Noether theorem and that time multiplied by a given speed is equal to a space distance!). Formerly in quantum physics:

$$\varnothing_{\text{Q.Ph.}}:=E_0\geq\frac{\hbar}{2\Delta t} \tag{11.2.233}$$

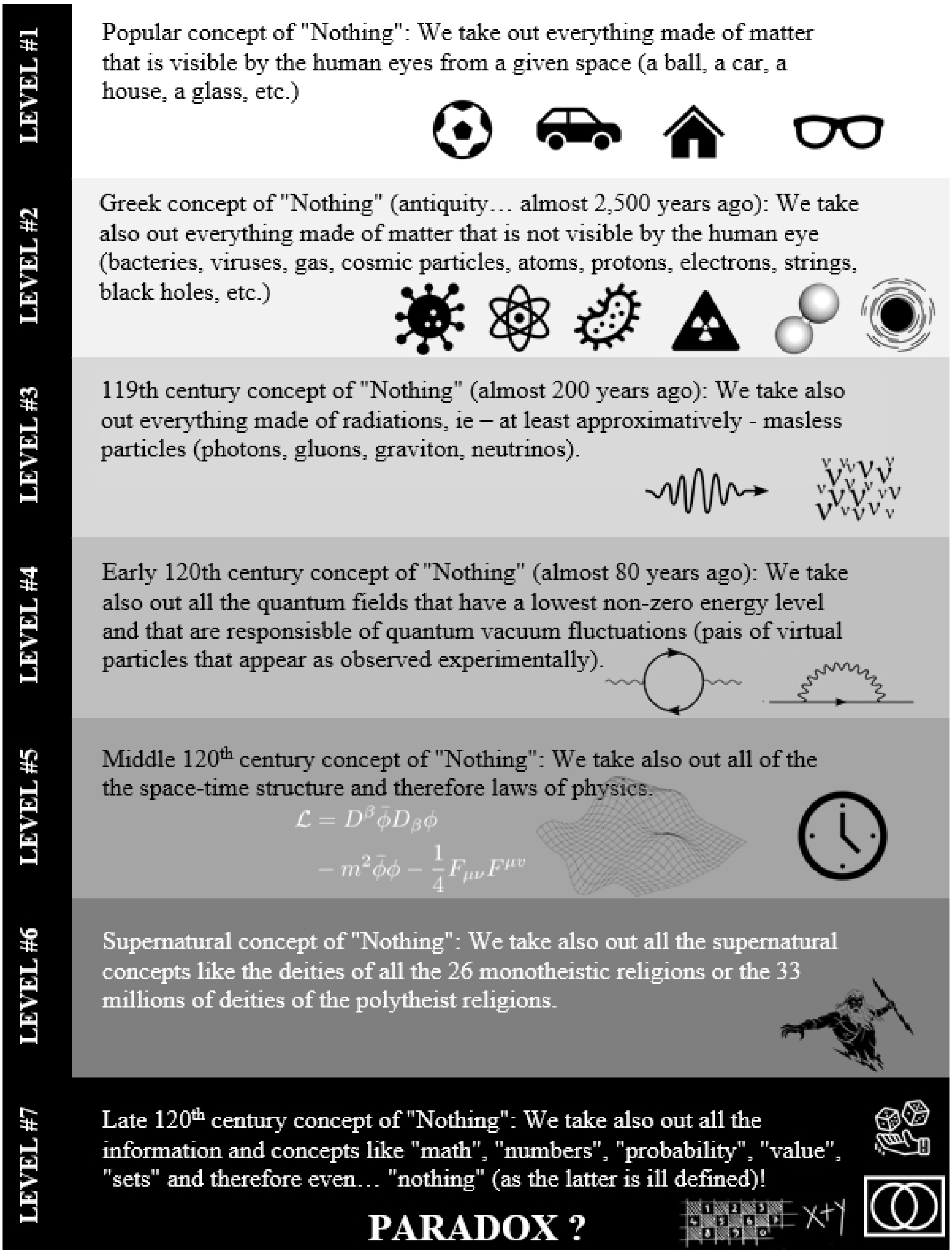
This can be found sometimes in a more technical notation in some textbooks as:

$$\varnothing_{\text{Q.Ph.}}:=\sigma_H\geq\frac{\hbar}{2\sigma_Q}|\mathrm{d}\langle Q\rangle/\mathrm{d}t| \tag{11.2.234}$$

As illustrated in the figure below, the concept of nothingness is (as for everything) more complex than it seems[10]:

---

[10] Notice that the levels are empirical and are not based on any international scientific consensus


LEVEL #1
LEVEL #2
LEVEL #3
LEVEL #4
LEVEL #5
$\mathcal{L} = D^{\beta}\bar{\phi}D_{\beta}\phi - m^{2}\bar{\phi}\phi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}$
LEVEL #6
LEVEL #7
PARADOX ?
X+Y

Popular concept of "Nothing": We take out everything made of matter that is visible by the human eyes from a given space (a ball, a car, a house, a glass, etc.)

Greek concept of "Nothing" (antiquity... almost 2,500 years ago): We take also out everything made of matter that is not visible by the human eye (bacteries, viruses, gas, cosmic particles, atoms, protons, electrons, strings, black holes, etc.)

119th century concept of "Nothing" (almost 200 years ago): We take also out everything made of radiations, ie – at least approximatively - masless particles (photons, gluons, graviton, neutrinos).

Early 120th century concept of "Nothing" (almost 80 years ago): We take also out all the quantum fields that have a lowest non-zero energy level and that are responsisble of quantum vacuum fluctuations (pais of virtual particles that appear as observed experimentally).

Middle 120th century concept of "Nothing": We take also out all of the the space-time structure and therefore laws of physics.

Supernatural concept of "Nothing": We take also out all the supernatural concepts like the deities of all the 26 monotheistic religions or the 33 millions of deities of the polytheist religions.

Late 120th century concept of "Nothing": We take also out all the information and concepts like "math", "numbers", "probability", "value", "sets" and therefore even... "nothing" (as the latter is ill defined)!

Figure 11.36: The seven levels of nothingness

If we go for the etymology of "nothingness" as the "negation of everything", we essentially construct the set:

$$\{P \mid P = \neg Q, \forall Q\} \tag{11.2.235}$$

where P, Q are statements.

Interestingly, via Russell's paradox (see section Set Theory page 138) the inconsistency of this set is swiftly established via the statement "*The above set exists*". Then it must contain its negation and its double negation. Hence nothingness as by the definition of the word itself leads to an impossibility. Or in the form a syllogism:

- Premise 1: Something exists.
- Premise 2: Something cannot come from nothing.
- Conclusion: Therefore, a state of "absolute nothingness" never existed.

For people who want to know more on the topic we recommend the following paper *Is zero-point energy physical? A toy model for Casimir-like effect* (see [416]).

> *If something doesn't come from nothing and something exists, therefore there was never nothing!*

11.2.3.3 Fourier uncertainty principle

Let us see that there is another way quite similar to the above one but slightly different and interesting that leads to the same result!

Let us rewrite the Fourier transform pair as following (Fourier transform first and inverse Fourier transform in second as seen in the section of Sequences and Series page 1758):

$$\begin{aligned} g(k)&=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} f(x)e^{-\mathrm{i}kx}\mathrm{d}x \\ f(x)&=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} g(k)e^{\mathrm{i}kx}\mathrm{d}k \end{aligned} \tag{11.2.236}$$

Let us assume that the both functions above are normalized such that:

$$\int\limits_{-\infty}^{+\infty}|f(x)|^2\mathrm{d}x=1 \qquad \int\limits_{-\infty}^{+\infty}|g(k)|^2\mathrm{d}k=1 \tag{11.2.237}$$

As f and g are complex functions, let us define the following mean inspired from our knowledge of the definition of mean during our study of Statistics for a continuous variable:

$$\begin{aligned} \mu_x=\langle x\rangle &=\int\limits_{-\infty}^{+\infty}\bar{f}(x)xf(x)\mathrm{d}x \\ \mu_k=\langle k\rangle &=\int\limits_{-\infty}^{+\infty}\bar{g}(k)kg(k)\mathrm{d}k \end{aligned} \tag{11.2.238}$$

Without loss in generality, let us put that:

$$\begin{aligned}\langle x\rangle &= \int\limits_{-\infty}^{+\infty} \bar{f}(x)xf(x)\mathrm{d}x := 0\\ \langle k\rangle &= \int\limits_{-\infty}^{+\infty} \bar{g}(k)kg(k)\mathrm{d}k := 0\end{aligned} \tag{11.2.239}$$

Now again as f and g are complex functions, let us define the following variance inspired from our knowledge of the definition of the variance during our study of Statistics for a continuous variable:

$$\sigma_x^2 = \left\langle (x-\mu_x)^2\right\rangle = \int\limits_{-\infty}^{+\infty} f^*(x)(x-\mu_x)^2 f(x)\mathrm{d}x\sigma_k^2 = \left\langle (k-\mu_k)^2\right\rangle \quad = \int\limits_{-\infty}^{+\infty} \bar{g}(k)(k-\mu_k)^2 g(k)\mathrm{d}k \tag{11.2.240}$$

However as we have assumed without loss of generality that $\mu_x = \mu_k = 0$, we have (remember that $z\cdot\bar{z} = a^2+b^2 = |z|^2$):

$$\begin{aligned}\sigma_x^2 &= \left\langle (x-\mu_x)^2\right\rangle = \int\limits_{-\infty}^{+\infty} \bar{f}(x)(x-\mu_x)^2 f(x)\mathrm{d}x = \int\limits_{-\infty}^{+\infty} |xf(x)|^2\mathrm{d}x\\ \sigma_k^2 &= \left\langle (k-\mu_k)^2\right\rangle = \int\limits_{-\infty}^{+\infty} \bar{g}(k)(k-\mu_k)^2 g(k)\mathrm{d}k = \int_{-\infty}^{+\infty} |kg(k)|^2\mathrm{d}k\end{aligned} \tag{11.2.241}$$

Now let us computer the standard deviation product:

$$\sigma_x^2\sigma_k^2 = \int\limits_{-\infty}^{+\infty} |xf(x)|^2\mathrm{d}x \int\limits_{-\infty}^{+\infty} |kg(k)|^2\mathrm{d}k \tag{11.2.242}$$

Remember the propery of time derivative (see page 2316):

$$\mathcal{F}\left(\frac{\mathrm{d}}{\mathrm{d}t}x(t)\right) = \mathrm{i}\omega X(\mathrm{i}\omega) \tag{11.2.243}$$

Let us adapt the notation and rearrange to our study case that we have here:

$$-\mathrm{i}\frac{\mathrm{d}f(x)}{\mathrm{d}x} \leftrightarrow kg(k) \tag{11.2.244}$$

Hence:

$$\sigma_x^2\sigma_k^2 = \int\limits_{-\infty}^{+\infty} |xf(x)|^2\mathrm{d}tx \int\limits_{-\infty}^{+\infty} \left|-\mathrm{i}\frac{\mathrm{d}f(x)}{\mathrm{d}x}\right|^2 \mathrm{d}x \tag{11.2.245}$$

Using the Cauchy-Schwarz hermitian inequality for integrals (see section Numerical Methods page 2388):

$$\sigma_x^2\sigma_k^2=\int\limits_{-\infty}^{+\infty}|xf(x)|^2\mathrm{d}tx\int\limits_{-\infty}^{+\infty}\left|-\mathrm{i}\frac{\mathrm{d}f(x)}{\mathrm{d}x}\right|^2\mathrm{d}x\geq\left|\int\limits_{-\infty}^{+\infty}-\mathrm{i}\frac{\mathrm{d}f(x)}{\mathrm{d}x}x\bar{f}(x)\mathrm{d}x\right|^2 \tag{11.2.246}$$

Developing the last term (as $z\bar{z}=|z|^2$):

$$\left|\int\limits_{-\infty}^{+\infty}-\mathrm{i}\frac{\mathrm{d}f(x)}{\mathrm{d}x}x\bar{f}(x)\mathrm{d}x\right|^2=\underbrace{\left(\int\limits_{-\infty}^{+\infty}\mathrm{i}\frac{\mathrm{d}\bar{f}(x)}{\mathrm{d}x}xf(x)\mathrm{d}x\right)}_{:=\bar{z}}\underbrace{\left(\int\limits_{-\infty}^{+\infty}-\mathrm{i}\frac{\mathrm{d}f(x)}{\mathrm{d}x}x\bar{f}(x)\mathrm{d}x\right)}_{:=z} \tag{11.2.247}$$

And as we we have proven in our study of complex numbers (see page 245) that:

$$|\Im(z)|\leq|z|\Leftrightarrow\bar{z}z=|z|^2\geq|\Im(z)|^2 \tag{11.2.248}$$

Therefore:

$$\sigma_x^2\sigma_k^2=\underbrace{\left(\int\limits_{-\infty}^{+\infty}\mathrm{i}\frac{\mathrm{d}\bar{f}(x)}{\mathrm{d}x}xf(x)\mathrm{d}x\right)}_{\bar{z}}\underbrace{\left(\int\limits_{-\infty}^{+\infty}-\mathrm{i}\frac{\mathrm{d}f(x)}{\mathrm{d}x}x\bar{f}(x)\mathrm{d}x\right)}_{z}\geq|\Im(z)|^2 \tag{11.2.249}$$

We will now simplify this relation. First remember that we have just put:

$$z=\int\limits_{-\infty}^{+\infty}-\mathrm{i}\frac{\mathrm{d}f(x)}{\mathrm{d}x}x\bar{f}(x)\mathrm{d}x \tag{11.2.250}$$

Using integration by parts and our normalization assumption we have:

$$\begin{aligned}\mathrm{i}z=\int\limits_{-\infty}^{+\infty}x\bar{f}(x)\frac{\mathrm{d}f(x)}{\mathrm{d}x}\mathrm{d}x&=x\bar{f}(x)f(x)\Big|_{-\infty}^{+\infty}-\int\limits_{-\infty}^{\infty}\frac{\mathrm{d}}{\mathrm{d}x}\left[x\bar{f}(x)\right]f(x)\mathrm{d}x\\&=0-\int\limits_{-\infty}^{+\infty}\left[\bar{f}(x)+x\frac{\mathrm{d}\bar{f}x)}{\mathrm{d}x}\right]f(x)\mathrm{d}x\\&=-\int\limits_{-\infty}^{+\infty}|f(x)|^2\mathrm{d}x-\int\limits_{-\infty}^{+\infty}xf(x)\frac{\mathrm{d}\bar{f}(x)}{\mathrm{d}x}\mathrm{d}x\qquad=-1+\mathrm{i}\bar{z}\end{aligned} \tag{11.2.251}$$

Therefore:

$$z-\bar{z}=\mathrm{i} \tag{11.2.252}$$

As we have :

$$\sigma_x^2\sigma_k^2\geq|\Im(z)|^2 \tag{11.2.253}$$

We know from our study of complex numbers that:

$$|\Im(z)|=\left[\frac{1}{2\mathrm{i}}(z-\bar{z})\right]^2 \tag{11.2.254}$$

Injecting into it the result $z-\bar{z}=\mathrm{i}$, this leads us to:

$$\sigma_x^2\sigma_k^2\geq\frac{1}{4} \tag{11.2.255}$$

So finally we get the "Fourier uncertainty principle":

$$\sigma_x\sigma_k\geq\frac{1}{2} \tag{11.2.256}$$

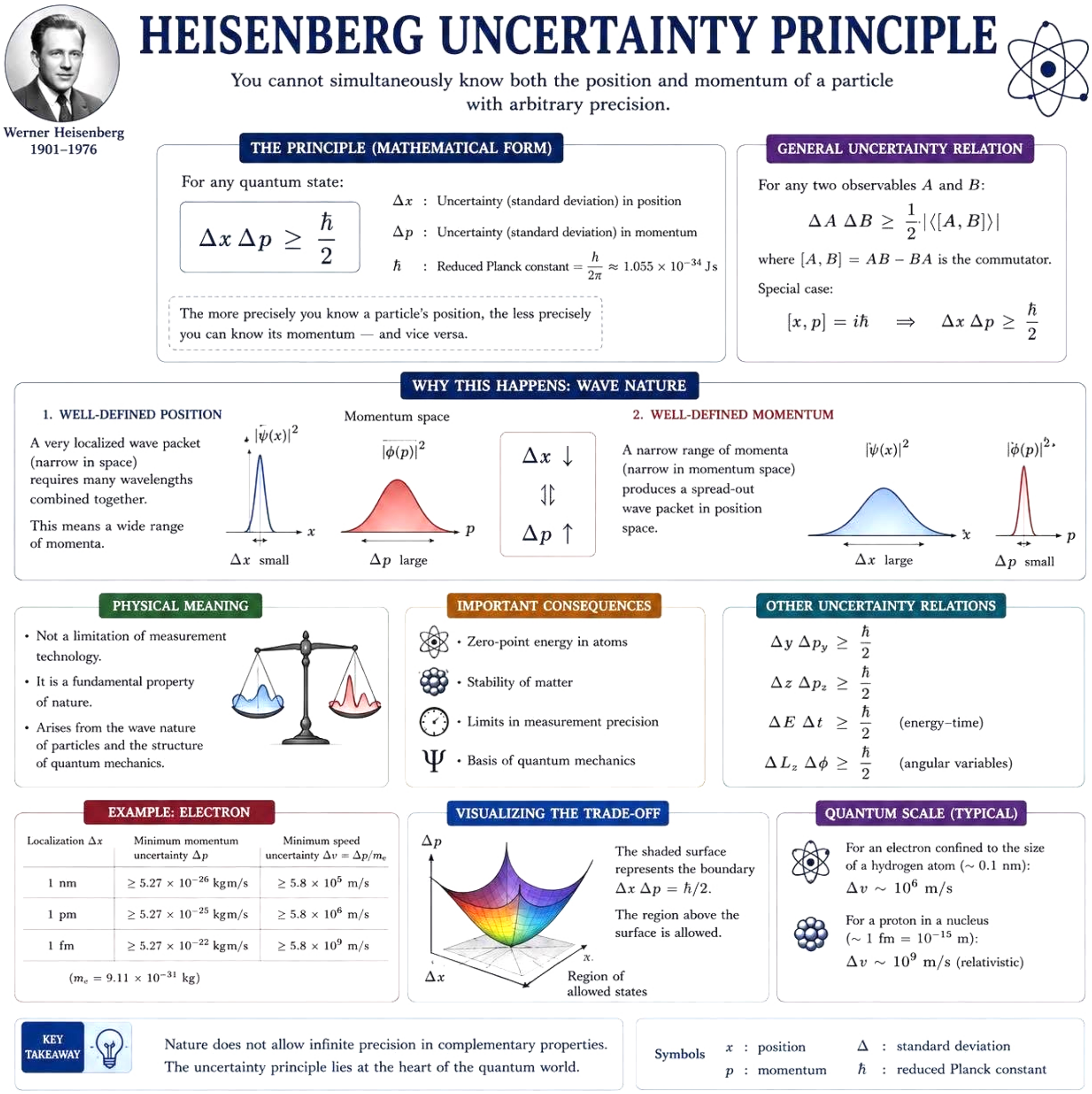
We see that the latter has almost the same form as the Heisenberg's uncertainty principle given for recall by:

$$\sigma_x\sigma_k\geq\frac{\hbar}{2} \tag{11.2.257}$$

This similarity is not a coincidence! Indeed, if we do in our previous developments the change of variable $p=\hbar k$, the we fall back on Heisenberg's uncertainty principle!

So to summarize the big picture:

1. Wave function tells you location of object, with uncertainty in position Δx.
2. Wave functions can be made out of sines and cosines.
3. Fourier Transforms connect uncertainty in position and uncertainty in the wavenumber (i.e. the number of waves or cycles per unit distance)
4. Momentum is related to wavelength


$|\psi(x)|^2$
x
Δx small
Momentum space
$|\phi(p)|^2$
p
Δp large
$\Delta x \downarrow$
$\Delta p \uparrow$
$|\psi(x)|^2$
x
Δx large
$|\phi(p)|^2$
p
Δp small
Δp
Δx
x
Region of allowed states

HEISENBERG UNCERTAINTY PRINCIPLE

You cannot simultaneously know both the position and momentum of a particle with arbitrary precision.

Werner Heisenberg
1901–1976

THE PRINCIPLE (MATHEMATICAL FORM)

For any quantum state:

$$\boxed{\Delta x\, \Delta p \geq \frac{\hbar}{2}}$$

Δx : Uncertainty (standard deviation) in position

Δp : Uncertainty (standard deviation) in momentum

$\hbar$: Reduced Planck constant $= \frac{h}{2\pi} \approx 1.055 \times 10^{-34}$ Js

The more precisely you know a particle's position, the less precisely you can know its momentum — and vice versa.

GENERAL UNCERTAINTY RELATION

For any two observables A and B:

$$\Delta A\, \Delta B \geq \frac{1}{2} |\langle [A, B] \rangle|$$

where $[A, B] = AB - BA$ is the commutator.

Special case:

$$[x, p] = i\hbar \quad \Rightarrow \quad \Delta x\, \Delta p \geq \frac{\hbar}{2}$$

WHY THIS HAPPENS: WAVE NATURE

1. WELL-DEFINED POSITION

A very localized wave packet (narrow in space) requires many wavelengths combined together.

This means a wide range of momenta.

2. WELL-DEFINED MOMENTUM

A narrow range of momenta (narrow in momentum space) produces a spread-out wave packet in position space.

PHYSICAL MEANING

- Not a limitation of measurement technology.
- It is a fundamental property of nature.
- Arises from the wave nature of particles and the structure of quantum mechanics.

IMPORTANT CONSEQUENCES

- Zero-point energy in atoms
- Stability of matter
- Limits in measurement precision
- Basis of quantum mechanics

OTHER UNCERTAINTY RELATIONS

$$\Delta y\, \Delta p_y \geq \frac{\hbar}{2}$$

$$\Delta z\, \Delta p_z \geq \frac{\hbar}{2}$$

$$\Delta E\, \Delta t \geq \frac{\hbar}{2} \quad \text{(energy–time)}$$

$$\Delta L_z\, \Delta \phi \geq \frac{\hbar}{2} \quad \text{(angular variables)}$$

EXAMPLE: ELECTRON

Localization Δx	Minimum momentum uncertainty Δp	Minimum speed uncertainty $\Delta v = \Delta p / m_e$
1 nm	$\geq 5.27 \times 10^{-26}$ kgm/s	$\geq 5.8 \times 10^{5}$ m/s
1 pm	$\geq 5.27 \times 10^{-25}$ kgm/s	$\geq 5.8 \times 10^{6}$ m/s
1 fm	$\geq 5.27 \times 10^{-22}$ kgm/s	$\geq 5.8 \times 10^{9}$ m/s

($m_e = 9.11 \times 10^{-31}$ kg)

VISUALIZING THE TRADE-OFF

The shaded surface represents the boundary $\Delta x\, \Delta p = \hbar/2$.

The region above the surface is allowed.

QUANTUM SCALE (TYPICAL)

For an electron confined to the size of a hydrogen atom (~ 0.1 nm):
$\Delta v \sim 10^{6}$ m/s

For a proton in a nucleus (~ 1 fm $= 10^{-15}$ m):
$\Delta v \sim 10^{9}$ m/s (relativistic)

KEY TAKEAWAY

Nature does not allow infinite precision in complementary properties. The uncertainty principle lies at the heart of the quantum world.

Symbols

x : position

p : momentum

Δ : standard deviation

$\hbar$: reduced Planck constant

11.2.4 Schrödinger Model

Experiments (Compton effect, photoelectric effect, Young's slits, geometric / wave optics, etc.) have shown that waves can be treated as corpuscles (and vice versa). It is these observations which led Niels Bohr to state his "principle of complementarity" which says that, according to the experiments carried out, we must consider matter either as a wave or as corpuscles. These two aspects complement each other.

11.2.4.1 de Broglie associated wave

In 11924 (holocene calendar), the French physicist Louis Victor de Broglie suggested that particles (electrons, protons, and others, and even atoms and molecules) could also, in some cases, exhibit wave properties in the same way as the light! De Broglie then expressed the idea that there existed between the associated fictitious wavelength of a particle of matter and its momentum a relation similar to that of a photon, that is to say (ν is the notation for the frequency for recall...):

$$E_{\text{tot}}=h\nu=h\frac{\omega}{2\pi}=\hbar\omega=mc^2 \tag{11.2.258}$$

so we can write using the relation established in the section of Wave Mechanics (see page 3479):

$$p=mc=\frac{E_{\text{tot}}}{c}=\frac{h\nu}{c}=\frac{\hbar\omega}{c}=\frac{\hbar kc}{c}=\hbar k=\frac{h}{\lambda} \tag{11.2.259}$$

Where:

$$\boxed{p=\hbar k} \tag{11.2.260}$$

is named the "de Broglie relation".

Hence the ratio:

$$\lambda=\frac{h}{p} \tag{11.2.261}$$

De Broglie then made the hypothesis that for a corpuscle of mass m and velocity v we have:

$$\boxed{\lambda=\frac{h}{m\nu}} \tag{11.2.262}$$

where λ is named the "associated de Broglie wavelength" and also denoted sometimes λ_{db}.

The moving matter would therefore have an associated wavelength!? It is an extremely small wavelength for masses of the order of a kilogram. If the speed is for example of the order of $1\;[\text{m}\cdot\text{s}^{-1}]$ then $\lambda\cong 10^{-34}$ [m].

As we have seen it, the phenomena of interference and diffraction are important only when the size of the objects or slits is not much greater than the wavelength. It is therefore impossible to detect the wave properties of everyday objects. It is not the same for elementary particles, the electrons in particular.

The electrons can therefore have wavelengths of the order of 10^{-10} [m] which corresponds to the spacing of the atoms of a crystal. C. J. Davisson and L. H. Germer carried out a crucial experiment: they scattered electrons on the surface of a crystal and in early 11927 (holocene calendar) observed that the ejected electrons were distributed in regular peaks. When they interpreted these peaks as diffraction peaks, they found that the wavelength of the diffracted electron was exactly that predicted by de Broglie!

But then what is an electron?? The illustrations that show an electron as a minuscule negatively charged sphere are only convenient but inaccurate images. In fact, we have to use the corpuscular or wave model, the one that works best according to the situation so that we can understand what is happening. But we must not conclude that an electron is a wave or a particle. Rather, we should say that an electron is "all of its measurable properties". Some physicists still use the term "quanton" to describe any system behaving either as a wave or as a particle.

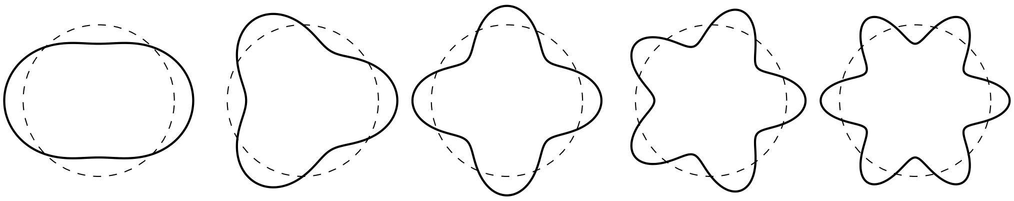
De Broglie was then able to suggest that each quantified electron orbit (according to Bohr's quantization postulate) is then a stationary wave and that an electron can only occupy orbits that can accept an integer number of wavelengths of its associated fictitious wave (today we assimilated this to the fact that the electron is trapped in a potential wells). If there was no exact coincidence, there could be no stationary wave, and therefore no stationary orbit either.

Consequently, as for the resonant modes of a string (standing waves), only the waves whose circumference of the circular orbit contains an integer number of λ exist, or (the amplitude will be calculated by particular techniques which will be seen further):

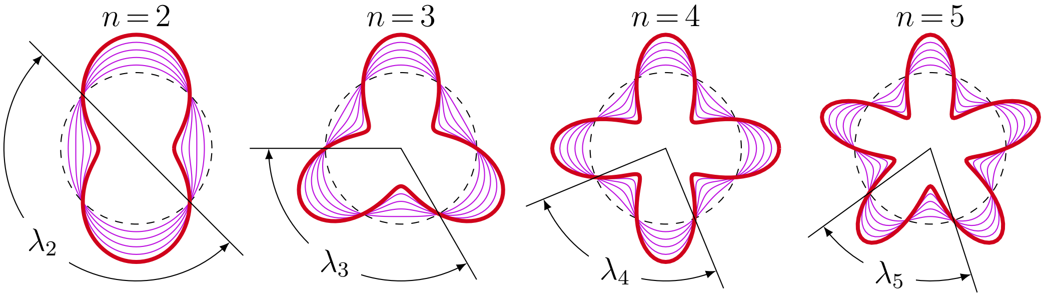
$$2\pi r_n = n\lambda \tag{11.2.263}$$

with $n \in \mathbb{N}^*$.

Visually for $n = 2, \ldots, 5$, we have for example:



And a more complete version by Juan Castaño:



and respecting the idea of growing scale of orbital radius (for $n = 2, \ldots, 4$):

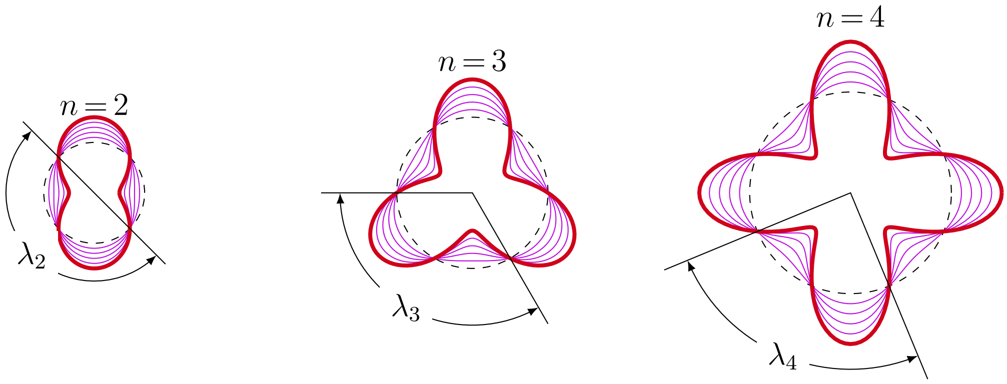


Figure 11.37: **Approach to the wave aspect of orbits by de Broglie**

By replacing λ by h/mv, we get:

$$2\pi r_n=\frac{nh}{mv}\Leftrightarrow mvr_n=\frac{nh}{2\pi} \tag{11.2.264}$$

That is indeed the quantum condition proposed by Bohr (see section Corpuscular Quantum Physics page 4626). The orbits and the quantized energy states of the Bohr model are due to the wave nature of the electron and to the fact that only resonant stationary waves persist. This supposes that the wave-particle duality is the basis of the structure of the atom.

The wave notion of the particle then allowed the physicist Erwin Schrödinger to develop a "wave equation" to describe the wave properties of particles.

Let us do a small sympathetic interlude... since the associated wave of de Broglie known and given the result seen during our study of the virial theorem in the section of Continuum Mechanics, we can relate together:

$$\lambda=\frac{h}{p}=\frac{h}{mv}\qquad E_c=\frac{1}{2}mv^2=\frac{1}{2}pv=\frac{3}{2}kT\Rightarrow v=\sqrt{\frac{3kT}{m}} \tag{11.2.265}$$

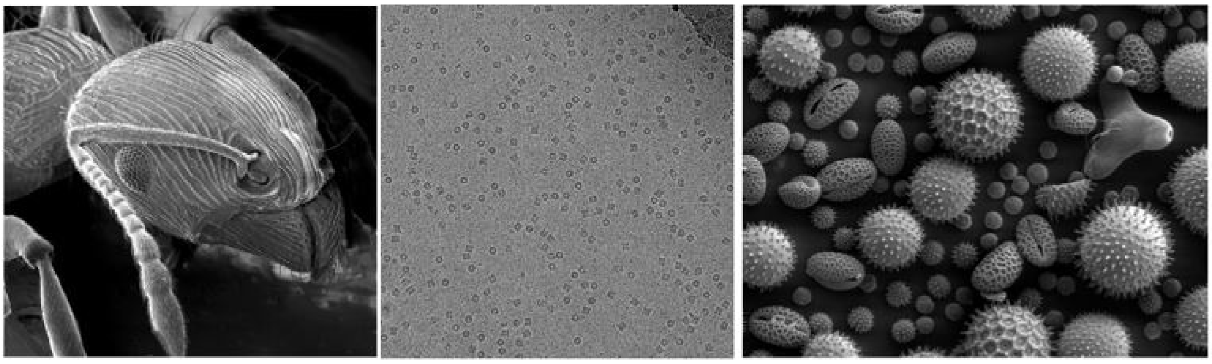
Thus, for a fluid (liquid), we can get the value of the associated "de Broglie thermal wave". Which gives us:

$$\lambda=\frac{h}{mv}=\frac{h}{m\sqrt{\frac{3kT}{m}}}=\frac{h}{\sqrt{3mkT}} \tag{11.2.266}$$

We will come back to this relation during our study of superfluids in the section of Continuum Mechanics.

11.2.4.1.1 Electronic microscopy

Usually, we see the world using photons in certain wavelengths[11]. Our eyes can see only a very limited width of the electromagnetic spectrum. The quantum revolution taught us that we can use other particles (and other wavelengths) to see the world and the Universe in ways we could have never ever imagined. This fact is even more general and can be thought valid even for gravitational waves (bunches of gravitons!).



There are several types of electron microscopes:

1. Transmission electron microscopy. The original form of the electron microscope, the transmission electron microscope (TEM), uses a high voltage electron beam to illuminate the specimen and generate an image. From Wikipedia, The resolution of TEMs is limited primarily by spherical aberration, but a new generation of hardware correctors can reduce spherical aberration to increase the resolution in high-resolution transmission electron microscopy (HRTEM) to below 0.5 Angstrom, enabling magnifications above 50 million times. The ability of HRTEM to determine the positions of atoms within materials is useful for nano-technologies research and development. TEM consists of an emission source or cathode, which may be a tungsten filament or needle, or a lanthanum hexaboride. Cryo-TEM is the cryoscopic modification of TEM in order to do EM for biology and precision TEM imaging. Samples cooled to cryogenic temperatures and embedded in an environment of vitreous water allows useful biological studies, and it deserved a Nobel Prize in 12017 (holocene calendar), to Jacques Dubochet, Joachim Frank, and Richard Henderson for developing cryo-electron microscopy for the high-resolution structure determination of biomolecules in solution.

2. Scanning electron microscopy (SEM). The scanning electron microscope (SEM) is a type of electron microscope that produces images of a sample by scanning the surface with a focused beam of electrons. The electrons interact with atoms in the sample, producing various signals that contain information about the surface topography and composition of the sample. The electron beam is scanned in a raster scan pattern, and the position of the beam is combined with the intensity of the detected signal to produce an image. In the most common SEM mode, secondary electrons emitted by atoms excited by the electron beam are detected using an Everhart-Thornley detector. The number of secondary electrons that can be detected, and thus the signal intensity, depends, among other things, on

---

[11] Thanks to J.F. G. for having authorized us to reproduce here the following blog entry: `http://www.thespectrumofriemannium.com/2019/09/19/log233-electron-microscopes/`

specimen topography. SEM can achieve resolution better than 1 nanometer. It can also made cryoscopic, as Wikipedia says: «(...)Scanning electron cryomicroscopy (CryoSEM) is a form of electron microscopy where a hydrated but cryogenically fixed sample is imaged on a scanning electron microscope's cold stage in a cryogenic chamber. The cooling is usually achieved with liquid nitrogen. CryoSEM of biological samples with a high moisture content can be done faster with fewer sample preparation steps than conventional SEM. In addition, the dehydration processes needed to prepare a biological sample for a conventional SEM chamber generate numerous distortions in the tissue leading to structural artefacts during imaging(...)».

3. Reflection electron microscopy (REM). In the reflection electron microscope (REM) as in the TEM, an electron beam is incident on a surface but instead of using the transmission (TEM) or secondary electrons (SEM), the reflected beam of elastically scattered electrons is detected. This technique is typically coupled with reflection high energy electron diffraction (RHEED) and reflection high-energy loss spectroscopy (RHELS). Another variation is spin-polarized low-energy electron microscopy (SPLEEM), which is used for looking at the microstructure of magnetic domains.

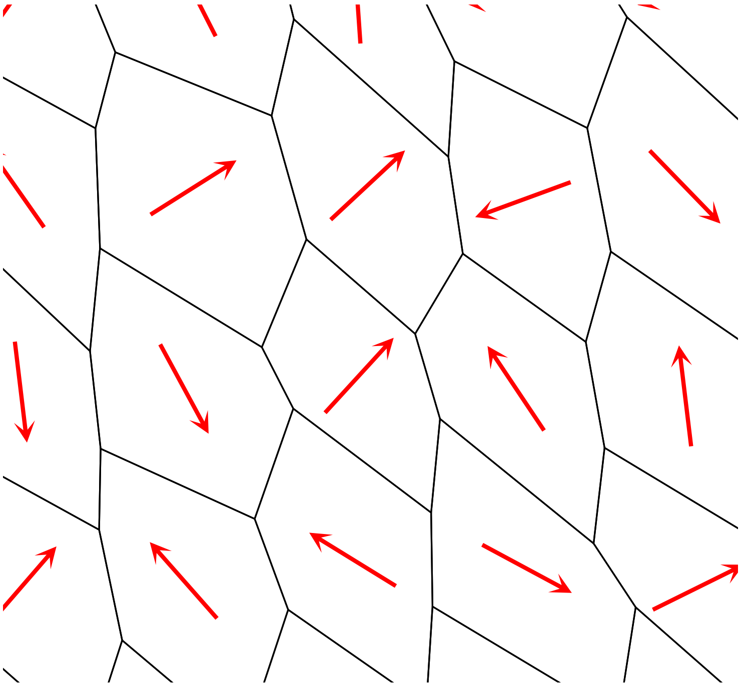


Figure 11.38: Schematic magnetic domains randomly assigned

4. ...

First, let us recall that simple non-relativistic electrons have a kinetic energy of:

$$E_c = \frac{1}{2} m_e v^2 \tag{11.2.267}$$

Any electron that is accelerated by a voltage ΔU change its kinetic energy in a conservative way, so:

$$\Delta E_c = -\Delta E_p = -q_e \Delta U \tag{11.2.268}$$

Suppose that initially $v_0=0\,[\text{m}\cdot\text{s}^{-1}]$ and $U_0=0$ [V], $U_f=U$. Then, the final kinetic energy reads:

$$E_{c,f}=\frac{1}{2}mv_f^2=\frac{p^2}{2m_e}=q_eU \tag{11.2.269}$$

where $p=mv$ is the non-relativistic linear momentum. In the quantum realm, any particle like the electron has an associated wave and wavelength. It is the de Broglie wavelength. And it reads:

$$\lambda_{\text{db}}=\frac{h}{mv} \tag{11.2.270}$$

Using the energy equation above, we can derive that:

$$p=\sqrt{2m_eq_eU} \tag{11.2.271}$$

and thus, we can derive the wavelength of the (non-relativistic) electron microscopy:

$$\lambda_e=\frac{h}{\sqrt{2m_eq_eU}} \tag{11.2.272}$$

We can generalize this relation to microscope of X-particles, where X-particles are particles with mass m_X and electric charger $q_X=Zq_e$ as follows:

$$\boxed{\lambda_X=\frac{h}{\sqrt{2m_XZq_eU}}} \tag{11.2.273}$$

Good!!!! Now, some numerology. We can use the value of the Planck constant as roughly $h\cong 6.63\cdot 10^{-34}\,[\text{J}\cdot\text{s}^{-1}]$, and then we can write:

$$\lambda_e=\frac{12.25\cdot 10^{-10}\,[\text{m}]}{\sqrt{U}}=\frac{1.225\,[\text{nm}]}{\sqrt{U}} \tag{11.2.274}$$

Or more generally:

$$\lambda_X=\frac{12.25\cdot 10^{-10}\,[\text{m}]}{\sqrt{N_XZ_XV}}=\frac{1.225\,[\text{nm}]}{\sqrt{N_XZ_XU}} \tag{11.2.275}$$

and where we wrote $m_X=N_Xm_e$ and $q_X=Z_Xq_e$. If, in particular, the energy is given in [eV] (electron-volts), then we get:

$$\begin{aligned}\lambda_e&=3.88\,[\text{pm}]\text{ at }E=100\,[\text{keV}]\\ \lambda_e&=2.74\,[\text{pm}]\text{ at }E=200\,[\text{keV}]\\ \lambda_e&=2.24\,[\text{pm}]\text{ at }E=300\,[\text{keV}]\\ \lambda_e&=1.23\,[\text{pm}]\text{ at }E=1\,[\text{keV}]\end{aligned} \tag{11.2.276}$$

Imagine muon microscopy, tau particle microscopy or W boson microscopy...

Now, we can enter the special relativistic electron microscope realm. Just as we have for photons (or any massless particle)

$$E=pc\quad\text{and}\quad\lambda=\frac{hc}{E} \tag{11.2.277}$$

for any relativistic particle with relativistic mass $M = m_0\gamma$ and energy:

$$E = Mc^2 = m_0\gamma c^2 \tag{11.2.278}$$

(rest mass $E_0 = m_0c^2$), the kinetic energy reads:

$$E_c = E - m_0c^2 = m_0c^2(\gamma - 1) \tag{11.2.279}$$

since:

$$E = E_c + m_0c^2 \tag{11.2.280}$$

Again, for a conservative force set-up:

$$\Delta E_c = -\Delta E_p = -q_e\Delta U = +q_eU \tag{11.2.281}$$

Taking into account that:

$$\gamma = \frac{1}{\sqrt{1-\dfrac{v^2}{c^2}}} \tag{11.2.282}$$

the special relativity theory generalizes the de Broglie relationship (indeed, de Broglie himself used SR in his wave-particle duality!):

$$\lambda = \frac{h}{m_0\gamma v} = \frac{h}{m_0 v}\sqrt{1-\frac{v^2}{c^2}} \tag{11.2.283}$$

From $E = m_0c^2\gamma$ we get:

$$1-\frac{v^2}{c^2} = \left(\frac{m_0c^2}{E}\right)^2 \tag{11.2.284}$$

From Einstein's relation we get:

$$p^2 = \frac{1}{c^2}\left(E^2-(m_0c^2)^2\right) \tag{11.2.285}$$

Therefore:

$$\begin{aligned} p &= \sqrt{\frac{1}{c^2}\left(E^2-(m_0c^2)^2\right)} = \sqrt{\frac{1}{c^2}\left((E_c+m_0c^2)^2-(m_0c^2)^2\right)} = \sqrt{\frac{1}{c^2}\left(E_c^2+2E_cm_0c^2\right)} \\ &= \sqrt{2E_cm\left(1+\frac{E_c}{2m_0c^2}\right)} \end{aligned} \tag{11.2.286}$$

Inserting this momentum into the relativistic de Broglie wavelength we finally derive the full relativistic electron microscope equation:

$$\lambda = \frac{h}{p} = \frac{h}{\sqrt{2E_cm\left(1+\dfrac{E_c}{2m_0c^2}\right)}} \tag{11.2.287}$$

or inserting E_c as q_eU units, we also get equivalently:

$$\boxed{\lambda=\frac{h}{p}=\frac{h}{\sqrt{2q_eUm\left(1+\dfrac{q_eU}{2m_0c^2}\right)}}} \tag{11.2.288}$$

Some numbers:

Voltage U [kV]	λ_{nr} [nm]	λ_r [nm]	mass $(\cdot m_e)$	$v\ (\cdot 10^8\ [\mathrm{m}\cdot\mathrm{s}^{-1}])$
100	0.00386	0.00370	1.196	1.644
200	0.00274	0.00251	1.391	2.086
400	0.00193	0.00164	1.783	2.484
1000	0.00122	0.00087	2.957	2.823

(11.2.289)

and where λ_{nr} is the non-relativistic wavelength and λ_r is the full relativistic wavelength. They are linked through the following expressions:

$$\begin{gathered}\lambda_{nr}=\frac{h}{\sqrt{2m_eq_eU}}\\ \Xi=\frac{1}{\sqrt{\left(1+\dfrac{E_c}{2m_0c^2}\right)}}\\ \lambda_r=\lambda_{nr}\Xi\end{gathered} \tag{11.2.290}$$

With simple scaling rules, we can extended the relativistic electron microscopes to relativistic X-particle microscopes as follows:

$$\boxed{\lambda=\frac{h}{p}=h\left[2m_XZ_Xq_eU\left(1+\frac{Z_Xq_eU}{2m_Xc^2}\right)\right]^{-1/2}} \tag{11.2.291}$$

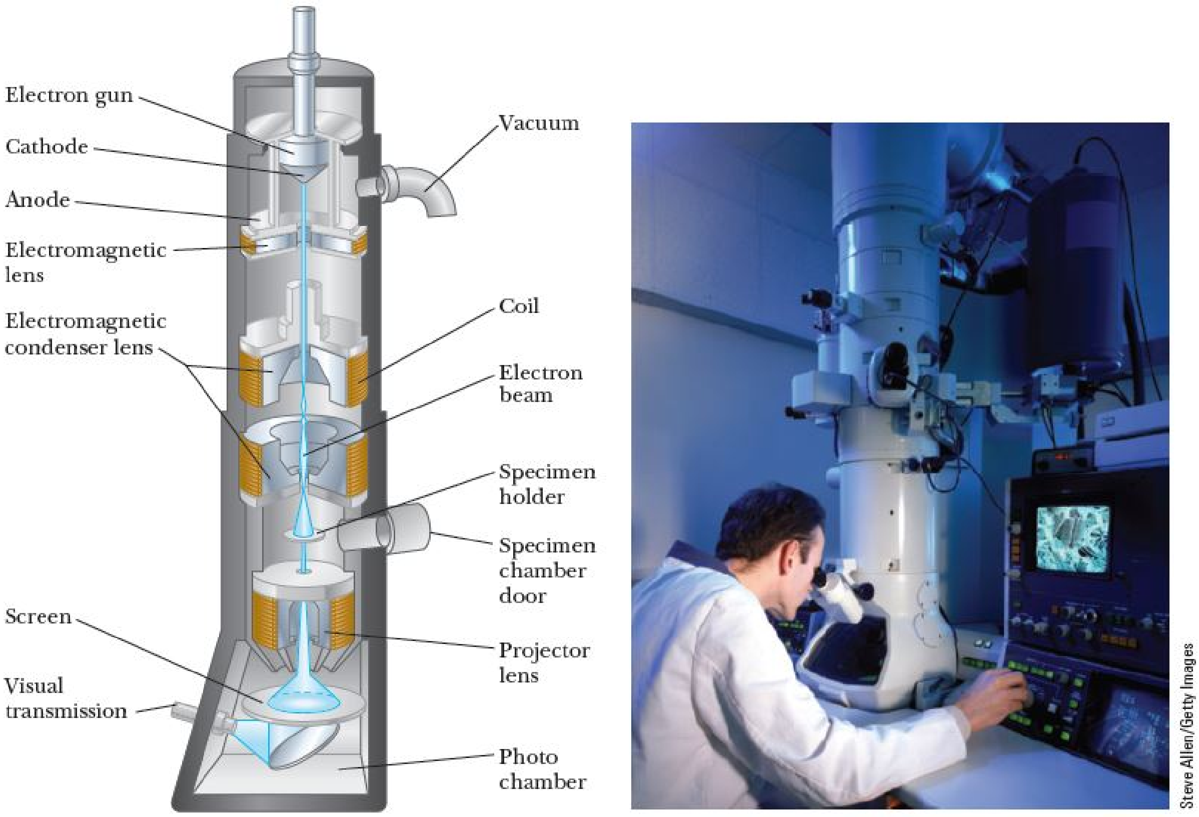

Electron gun
Cathode
Anode
Electromagnetic lens
Electromagnetic condenser lens
Screen
Visual transmission
Vacuum
Coil
Electron beam
Specimen holder
Specimen chamber door
Projector lens
Photo chamber
Steve Allen/Getty Images

Figure 11.39: Naive diagram of a transmission electron microscopy (source: [491])

In 2021 scientists achieved a record level of visual detail with an electron microscope imaging technique. The image below shows an electron ptychographic reconstruction of a praseodymium orthoscandate ($\mathrm{PrScO_3}$) crystal, zoomed in 100 million times (the blurry dots are atoms!):

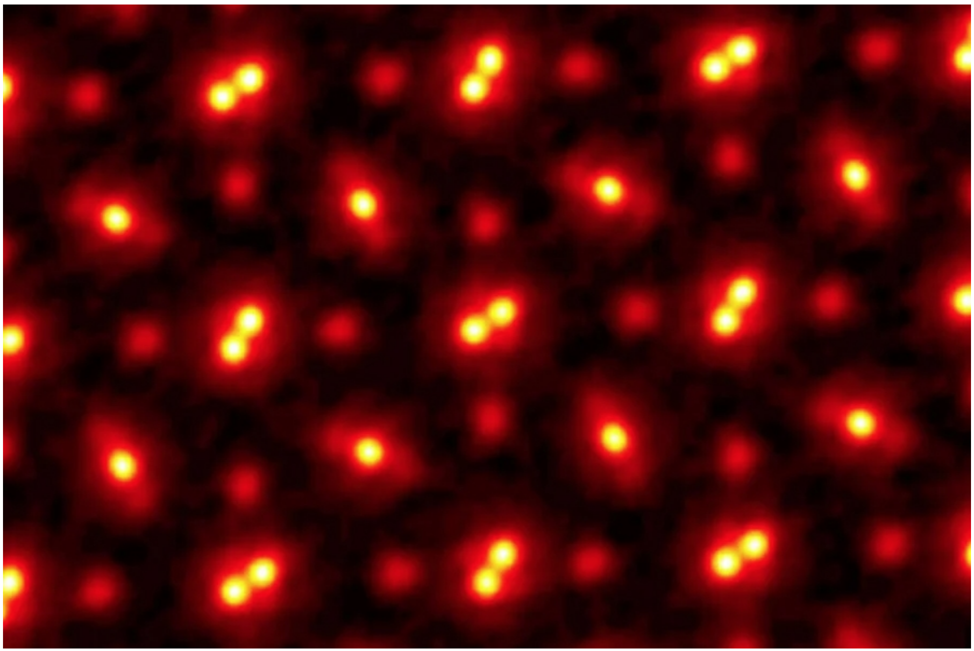


Figure 11.40: Electron ptychographic reconstruction of a praseodymium orthoscandate (source: Cornell University)

11.2.4.2 Classical (time independent) Schrödinger Wave Equation

The physicist Peter Debye found the de Broglie model a bit far fetched. He argued that the physics of waves, from the sound waves to the electromagnetic waves, even the waves propagating on a string requires an equation that describes them. There was no wave equation for de Broglie's atomic model (for the simple reason that the latter had never tried and Albert Einstein either). The physicist Erwin Schrödinger then took it upon himself to find the missing equation and did it brilliantly.

Let us recall the one-dimensional form of the wave equation (see section Wave Mechanics page 3476):

$$\frac{\partial^2 f(x,t)}{\partial x^2}=\frac{1}{v^2}\frac{\partial^2 f(x,t)}{\partial t^2} \tag{11.2.292}$$

To simplify, let us look for a particular solution of the form (see the section on Wave mechanics or the section on Electrodynamics for the analogy):

$$f(x,t)=\Psi(x)\sin(\omega t) \tag{11.2.293}$$

$\Psi(x)$ is the amplitude of the field associated with the particle. It is important to notice that the periodic part does not contain any displacement parameter k (as is the case in Electrodynamics, for example) because the function must describe "static" solutions (be careful not to take this literally).

For historical reasons this amplitude is as we know commonly named "wave function" although this name is misleading. It might be better to name it simply "amplitude field associated with matter".

It is the search for the expression of this function that will lead us during the study of a special case (see much further below in text) to the well-known expression of the ionisation energy of an electron of given main quantum number n and for its atom of given atomic number N.

If we introduce 11.2.293 into 11.2.292, we get:

$$\begin{aligned}&\cancel{\sin(\omega t)}\frac{\partial^2\Psi(x)}{\partial x^2}=-\frac{\omega^2}{v^2}\Psi(x)\cancel{\sin(\omega t)}\\&\Rightarrow\frac{\partial^2\Psi(x)}{\partial x^2}=-\frac{\omega^2}{c^2}\Psi(x)\Rightarrow\frac{\partial^2\Psi(x)}{\partial x^2}=-\frac{(2\pi v)^2}{c^2}\Psi(x)\\&\Rightarrow\frac{\partial^2\Psi(x)}{\partial x^2}=-\frac{\left(2\pi\dfrac{c}{\lambda}\right)^2}{c^2}\Psi(x)\Rightarrow\frac{\partial^2\Psi(x)}{\partial x^2}=-\frac{4\pi^2}{\lambda^2}\Psi(x)\\&\Rightarrow\frac{\partial^2\Psi(x)}{\partial x^2}+\frac{4\pi^2}{\lambda^2}\Psi(x)=0\end{aligned} \tag{11.2.294}$$

We have also:

$$E_c=\frac{1}{2}mv^2=\frac{1}{2m}p^2=\frac{1}{2m}\left(\frac{h}{\lambda}\right)^2=E_{\text{tot}}-E_p \tag{11.2.295}$$

Hence:

$$\frac{1}{\lambda^2}=\frac{2mE_c}{h^2}=\frac{2m(E_{\text{tot}}-E_p)}{h^2} \tag{11.2.296}$$

if we introduce 11.2.296 into 11.2.294 then we get the (time-independent) "classical one-dimensional Schrödinger equation" (in the absence of magnetic field ...):

$$\frac{\partial^2\Psi}{\partial x^2}+\frac{8\pi^2 m(E_{\text{tot}}-E_p)}{h^2}\Psi(x)=0\,[\text{m}^{-5/2}] \tag{11.2.297}$$

Remark

The potential energy could be gravitational as well as electric or both combined (therefore of any kind). But gravity is assumed to be so small at such scales compared to electrostatic forces that it is neglected.

We can rewrite the preceding equation by generalizing it to a three-dimensional system and using the reduced Planck constant. What ultimately gives us:

$$\boxed{\Delta\Psi+\frac{2m}{\hbar}(E_{\text{tot}}-E_p)\Psi=0\,[\text{m}^{-7/2}]} \tag{11.2.298}$$

where for recall Δ is the Laplacian of a scalar field (see section Vector Calculus page 1899):

$$\Delta=\frac{\partial^2}{\partial^2 x}+\frac{\partial^2}{\partial^2 y}+\frac{\partial^2}{\partial^2 z} \tag{11.2.299}$$

The above relation is also often denoted after expliciting the Laplacian and rearranging the terms:

$$\left(-\frac{\hbar}{2m}\nabla^2+E_p\right)\Psi=E_{\text{tot}}\Psi \tag{11.2.300}$$

Let $\vec{r}_i$ be the position vector of electron i and $\vec{R}_j$ be the position vector of proton j. Then, taking into account the electric force between all charged particles and neglecting the motions of the protons and neutrons in the nucleus, which are very slow compared to the electrons, Schrödinger's equation becomes:

$$\left\{-\frac{\hbar^2}{2m_e}\sum_{i=1}^{z}\nabla_i^2-\sum_{i=1}^{z-1}\sum_{j=1+1}^{z}\frac{k_E e^2}{||\vec{r}_i-\vec{R}_j||^2}+\sum_{i}^{z-1}\sum_{k=i+1}^{z}\frac{k_E e^2}{||\vec{r}_i-\vec{r}_k||^2}\right\}\Psi=E_{\text{tot}}\Psi \tag{11.2.301}$$

This is complicated, but it is completely explicit. It describes every atom in the periodic table! Note that we have taken care of not only the electrical force between each electron and each proton in the nucleus but also the forces between each electron pair. The neutron has a small magnetic interaction with the electron and protons, but this can be neglected for most chemical purposes.

Remarks

R1. This equation is not a Lorentz invariant given that it was established from the classical expression of energy (and not the relativistic one).

R2. The plane wave function that we have taken at the beginning does not have a physical significance since it carries an infinite energy. A better solution is to consider a wave packet. However, the wave packets generally employed consist of a superposition of plane waves (see proof further below). Hence, by studying its effects on one of the plane waves, we can still accept the physical conclusions which we can deduce from them...

However as simple as it may seem, it took all the skill and experience of Schrödinger to be the first to write this wave equation and it was the foundation upon which he built the mathematical edifice of Wave Quantum Physics in the following months. But first it had to prove that it was indeed the wave equation sought by applying it to the hydrogen atom for which some results were well known thanks to Sommerfeld results (see section Corpuscular Quantum Physics page 4648). His model eliminated all the successive and empirical tinkering of Corpuscular Quantum Physics.

Some physicists judged Schrödinger's model to be pure madness (in particular Sommerfeld), but they changed mind shortly afterwards... The model was effective and replaced advantageously the horrible abstract matrix model of Heisenberg on which even Pauli had failed to use to explain the experimental results of the hydrogen atom. Soon after Max Born described the wave model as the deepest form of quantum laws, which obviously did not please his close friend Heisenberg...

However, the Heisenberg model describes particles as Schrodinger describes waves. The advantage of having two different formalisms - but equivalent - in quantum physics quickly became evident. For most of the problems physicists encounter, wave mechanics offers the easiest way to the solution. However, for others, such as those involving spin, Heisenberg's matrix approach proves (or rather "defines") its value (see section Relativistic Quantum Physics page 4985).

11.2.4.2.1 Schrödinger Hamiltonian

The Schrödinger equation can also be written as (after a few small elementary factorizations) as follows:

$$\left(E_p-\frac{\hbar}{2m}\Delta\right)\Psi=E_{\text{tot}}\Psi \tag{11.2.302}$$

We write this in quantum physics in the form (the units are given for the three dimensional case):

$$\boxed{H\cdot\Psi=E_{\text{tot}}\cdot\Psi\;[\mathrm{J}\cdot\mathrm{m}^{-3/2}]} \tag{11.2.303}$$

Where H is therefore the Hamiltonian of the system (or total energy) and constitutes a functional operator and where the total energy E_{tot} is its eigenvalue.

The Schrödinger equation is therefore an equation with partial derivatives of the second order, homogeneous linear. Whatever the total energy, it admits solutions (phew!), but we show that in general these solutions grow very rapidly (exponential growth) when we go to infinity in certain directions and are therefore physically unacceptable. There are only particular values of the total energy which give rise to physically acceptable solutions and in general all of these values include discrete values (trigonometric functions at the source with integer parameters) which are the "linked levels" of the system (because their eigenfunction decreases rapidly to infinity) and a continuum of values which are the "unbound levels" (their eigenfunction remaining finite at infinity). More precisely, if W is the lower limit of the potential energy values at infinity, the bounded levels lie below W, while the values greater than W constitute the continuum of the unbound levels.

For example, in the study of the harmonic oscillator (one of the most difficult practical cases in terms of formalism) that we will do later, we will prove that we have:

$$E_p=\frac{Cx^2}{2} \tag{11.2.304}$$

with $W=+\infty$. So there are only linked levels.

For the hydrogen atom:

$$E_p=\frac{-e^2}{4\pi\varepsilon_0 r} \tag{11.2.305}$$

with $W=0$. Linked levels are negative, therefore all positive energy values will be unbound levels.

This having been said, let us also consider as an example (very important) how to determine the Hamiltonian H of the Schrödinger equation of a non-relativistic charged particle in an electromagnetic field!

We have seen in the section of Analytical Mechanics (see page 3155) that the (classical) Lagrangian was defined by the subtraction of kinetic and potential energy according to the relation:

$$L=T-V \tag{11.2.306}$$

We have proved in the section of Electrodynamics that the Lagrangian of the relativistic field-current interaction was given by (see page 4124):

$$L=L_0+L_1=-m_0c^2\sqrt{1-\frac{v^2}{c^2}}+q\vec{A}\circ\vec{v}-q\phi \tag{11.2.307}$$

where for recall ϕ is the vector potential of the electric field (whose gradient is the electric field $\vec{E}$) and $\vec{A}$ is the vector potential of the magnetic field (whose curl is the magnetic field $\vec{B}$).

If we add an electric field (and therefore an electrostatic potential U) in addition to the electromagnetic field, the Lagrangian is then written (since the potential is subtracted according to the Lagrangian definition):

$$L=L_0+L_1=-m_0c^2\sqrt{1-\frac{v^2}{c^2}}-qU+q\vec{A}\circ\vec{v}-q\phi \tag{11.2.308}$$

In the classical (non-relativistic) approximation we know that we have $v/c\ll 1$ and using the Maclaurin series of (see section Sequences and Series page 1703):

$$\sqrt{1+x}\cong 1+\frac{1}{2}x \tag{11.2.309}$$

we can then write:

$$L=-m_0c^2+\frac{1}{2}m_0v^2-qU+q\vec{A}\circ\vec{v}-q\phi \tag{11.2.310}$$

As we restrict ourselves to the non-relativistic case, we can eliminate the constant term of energy of the mass at rest such that:

$$L=\frac{1}{2}m_0v^2-qU+q\vec{A}\circ\vec{v}-q\phi \tag{11.2.311}$$

Still in the section of Analytical Mechanics, we showed that the Hamiltonian was given by:

$$H=\sum_i \dot{x}_ip_i-L \tag{11.2.312}$$

Therefore we have

$$H=\sum_i \dot{x}_ip_i-\frac{1}{2}m_0v^2+qU-q\vec{A}\circ\vec{v}+q\phi \tag{11.2.313}$$

Moreover, we have seen in the section of Analytical Mechanics (see page 3155) that:

$$p_i=\frac{\partial L}{\partial \dot{x}_i} \tag{11.2.314}$$

It therefore comes that:

$$p_i=\frac{\partial}{\partial \dot{x}_i}\left(\frac{1}{2}m_0v_i^2-qU+q\vec{A}\circ\vec{v}-q\phi\right)=m_0v_i+qA_i \tag{11.2.315}$$

Finally:

$$H=m_0v^2+q\vec{A}\circ\vec{v}-\frac{1}{2}m_0v^2+qU-q\vec{A}\circ\vec{v}+q\phi \tag{11.2.316}$$

Either after simplification:

$$H=\frac{1}{2}m_0v^2+qU+q\phi=\frac{1}{2}m_0v^2+qU+V \tag{11.2.317}$$

H therefore contains the kinetic energy and total potential energy. There is no magnetic term because the Laplace force, as we proved it in the section of Magnetostatic, does not work (there are some who are lucky...). H is indeed the total energy of the classical system, but the preceding relation is not really adapted to Hamilton's formalism because the conjugate moments do not appear. But it is very simple to introduce them from the result obtained previously which was:

$$p_i=m_0v_i+qA_i\Rightarrow v_i=\frac{p_i-qA_i}{m_0} \tag{11.2.318}$$

Therefore:

$$H=\frac{1}{2}m_0v^2+q\cdot U+V=\frac{1}{2}m_0\left(\frac{p_i-qA_i}{m_0}\right)^2+qU+V=\frac{(p_i-qA_i)^2}{2m_0}+qU+V \tag{11.2.319}$$

If we go into Quantum Physics, we have to replace the p_i by their respective operators:

$$p_i=-\mathrm{i}\hbar\partial_i \tag{11.2.320}$$

of which we have already proved the origin earlier above. Thus, we have that:

$$H=\frac{(p_i-qA_i)^2}{2m_0}+qU+V \tag{11.2.321}$$

which must be written in the general case (as we do not know if the vector potential commutes with the linear momentum operator, we will assume that it does not commute):

$$H_i = \frac{p_i^2}{2m_0} - \frac{q(p_iA_i + A_ip_i)}{2m_0} + \frac{q^2A_i^2}{2m_0} + qU + V \tag{11.2.322}$$

What is traditionally written in the form (sic!):

$$\boxed{H = \frac{p^2}{2m_0} - q\frac{(pA + Ap)}{2m_0} + \frac{q^2A^2}{2m_0} + qU + V} \tag{11.2.323}$$

This last relation is very often found in the following simplified form in textbooks in the absence of potential U and by making V explicit:

$$H = \frac{1}{2m_0}(p - qA)^2 + q\phi \tag{11.2.324}$$

Remark

In the section of Relativistic Quantum Physics we will prove the relativistic form of this Hamiltonian associated with the generalized Klein-Gordon equation (see page 4924) or that of Dirac (see page 4978) which includes the $1/2$ spin!

Fortunately, we will not discuss examples where we will have to find solutions to the Schrödinger equation with such a Hamiltonian in this textbook...

In the academic case, however, when application premises are addressed ..., we cancel either the vector potential (i.e. the particle is not immersed in a magnetic field) or the scalar potential (i.e. the particle is not immersed in an electric field). Moreover, when one of the fields is chosen as being non-zero, we take the case where it is constant and in a single dimension... Thus, if we want a constant electric field along a single axis (for example the x-axis), we take as scalar potential:

$$\phi(x,t) = E_0 \cdot x \tag{11.2.325}$$

since its gradient will give a constant according to x. In the case of a constant magnetic field along a single axis (for example the z-axis), we will arrange to take the vector potential:

$$A = \frac{B_0}{2}\begin{pmatrix} -y \\ x \\ 0 \end{pmatrix} \tag{11.2.326}$$

whose rotational (curl) gives a constant magnetic field according to z.

11.2.4.2.2 De Broglie normalization condition

In general, in a given dynamic state, the particle (if it is a one-particle system) described by the resolution of the Schrödinger equation for well-defined parameters is badly localized because x, y and z are not well defined by Heisenberg's principle of uncertainty. It is therefore necessary to define a probability $\mathrm{d}P$ of finding the particle in the volume element $\mathrm{d}x\mathrm{d}y\mathrm{d}z$ surrounding a point (x,y,z), hence the existence of a distribution function of the coordinates such that:

$$\mathrm{d}P=\rho(x,y,z)\mathrm{d}x\mathrm{d}y\mathrm{d}z \tag{11.2.327}$$

where $\rho(x,y,z)$ is therefore an essentially positive or zero quantity (probabilities obliged!) which must be expressed by means of the Schrödinger equation function $\Psi(x,y,z)$. We have, moreover, such very detailed examples in this section and at the end of that of Quantum Chemistry.

Analogies with Classical Wave Physics, more precisely with Electrodynamics, have led us to admit that since the density of energy of an electromagnetic wave is proportional to the square of its amplitude (see section Electrodynamics page 4156), the volumetric density probability must be proportional to the square of the intensity of the associated field such that:

$$\rho(x,y,z)=c^{te}\cdot|\Psi(x,y,z)|^2 \tag{11.2.328}$$

where we use the Schrödinger function module as an amplitude analogy and where the constant is a real number. In the framework of Quantum Physics, it is much more common to find this last relation in the following obvious form:

$$\rho(x,y,z)=\frac{|\Psi(x,y,z)|^2}{\displaystyle\int|\Psi(x,y,z)|^2\mathrm{d}x\mathrm{d}y\mathrm{d}z}=\left|\frac{\Psi(x,y,z)}{\sqrt{\displaystyle\int|\Psi(x,y,z)|^2\mathrm{d}x\mathrm{d}y\mathrm{d}z}}\right|^2=c^{te}|\Psi(x,y,z)|^2 \tag{11.2.329}$$

which highlights the necessary normalization of the Schrödinger equation.

Finally, it is important to know that during mathematical developments, the huge majority of physicists have the habit of keeping the same notation for the non-normalized Schrödinger function as for the normalized one (which can lead to some confusion) such as:

$$|\Psi(x,y,z)|^2=\left|\frac{\Psi(x,y,z)}{\sqrt{\displaystyle\int|\Psi(x,y,z)|^2\mathrm{d}x\mathrm{d}y\mathrm{d}z}}\right|^2 \tag{11.2.330}$$

where $|\Psi(x,y,z)|^2$ then represents the density probability of finding the particle at a certain point in space.

It is evident then that with this way of writing things we have then over all space:

$$\boxed{\iiint_V\rho(x,y,z)\mathrm{d}x\mathrm{d}y\mathrm{d}z=\iiint_V|\Psi(x,y,z)|^2\mathrm{d}x\mathrm{d}y\mathrm{d}z=1} \tag{11.2.331}$$

We can now consider the physical significance which may be attached to the intensity of the field associated with matter. Since this field describes the motion of a particle, we can say that the regions of space in which the particle is most likely to be found are those in which the $|\Psi(x,y,z)|^2$ is maximum.

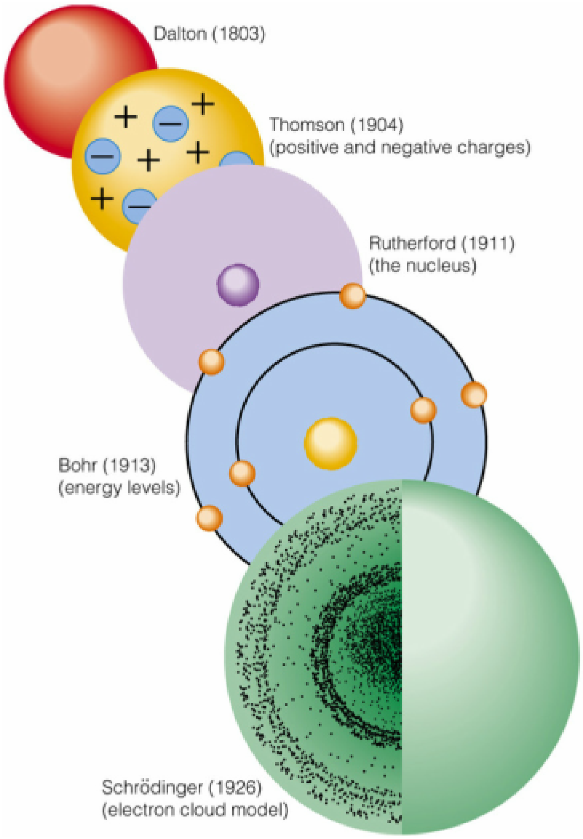

Dalton (1803)
Thomson (1904)
(positive and negative charges)
Rutherford (1911)
(the nucleus)
Bohr (1913)
(energy levels)
Schrödinger (1926)
(electron cloud model)

Figure 11.41: **Atomic model evolution (source: ?)**

Let us also inform you that the prior-previous relation is often written using the Dirac ket-bra notation in the very refined (and very common...) way:

$$\| \Psi\rangle |^2 = \left| \frac{|\Psi\rangle}{\sqrt{\langle\Psi \mid \Psi\rangle}} \right|^2 = \frac{\| \Psi\rangle |^2}{\langle\Psi \mid \Psi\rangle} \tag{11.2.332}$$

where the module in the denominator disappears, since, for recall, the integral is a real number. Or in the following way (it is simply the square root of the previous one):

$$|\Psi\rangle = \frac{|\Psi\rangle}{\sqrt{\langle\Psi \mid \Psi\rangle}} \tag{11.2.333}$$

It should never be forgotten that physicists, for the vast majority, note in a similar way the non-normalized and normalized Schrödinger function as this last relation reminds us.

As we have already said, we will see many detailed examples of this normalization in this section with one-dimensional spaces and in the context of volumes in the section on Quantum Chemistry.

✍**Example:**

A ball is constrained to move along a line inside a tube of length L. The ball is equally likely to be found anywhere in the tube at some time t. What is the probability of finding the ball in the left half of the tube at that time? (the answer is 50%, of course, but how do we get this answer by using the probabilistic interpretation of the quantum mechanical wave function?)

The first step to the answer is to write down the wave function. The ball is equally like to be found anywhere in the box, so one way to describe the ball with a constant wave function:

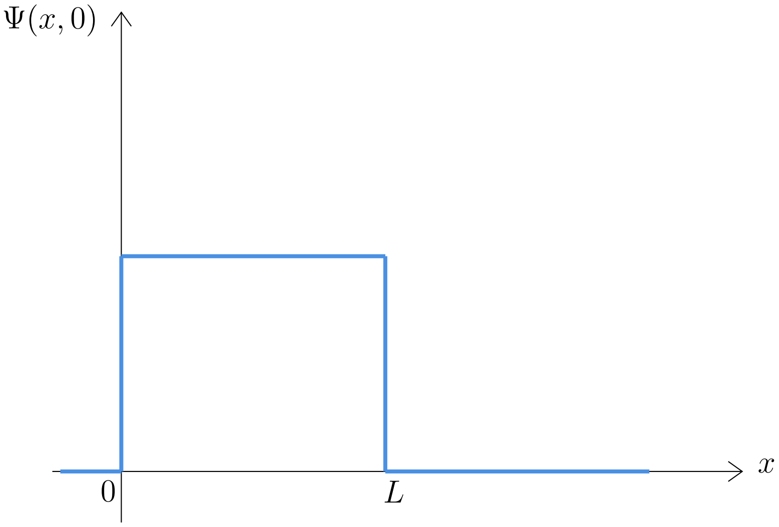


Figure 11.42: **Wave function for a ball in a tube of length L**

The normalization condition can be used to find the value of the function and a simple integration over half of the box yields the final answer.

The wave function of the ball can be written as:

$$\Psi(x,t)=C^{te}(0\leq x\leq L) \tag{11.2.334}$$

where c^{te} is a constant, and $\Psi(x,t)=0$ otherwise. We can determine the constant C^{te} by applying the normalization condition (we set $t=0$ to simplify the notation):

$$P(-\infty\leq x\leq +\infty)=\int\limits_{-\infty}^{+\infty}\Psi\bar{\Psi}\mathrm{d}x=\int\limits_{-\infty}^{+\infty}|C^{te}|^2\mathrm{d}x=1 \tag{11.2.335}$$

This integral can be broken into three parts: (1) negative infinity to zero, (2) zero toL, and (3) L to infinity. The particle is constrained to be in the tube, so $C^{te}=0$ outside the tube and the first and last integrations are zero. The above equation can therefore be written:

$$P(x=0,L)=\int\limits_{0}^{L}|C^{te}|^2\mathrm{d}x=1 \tag{11.2.336}$$

The constant C^{te} does not depend on x and can be taken out of the integral, so we obtain:

$$P(x=0,L)=|C^{te}|^2\int_0^L \mathrm{d}x=1 \tag{11.2.337}$$

Integration gives obviously:

$$C^{te}=\sqrt{\frac{1}{L}} \tag{11.2.338}$$

To determine the probability of finding the ball in the first half of the box $(0\leq x\leq L)$, we have:

$$P(x=0,L/2)=\int_0^{L/2}\left|\sqrt{\frac{1}{L}}\right|^2\mathrm{d}x=\left(\frac{1}{L}\right)\frac{L}{2}=0.5 \tag{11.2.339}$$

The probability of finding the ball in the first half of the tube is 50%, as expected. Two observations are noteworthy. First, this result corresponds to the area under the constant function from $x=0$ to $L/2$ (the area of a square left of $L/2$). Second, this calculation requires an integration of the square of the wave function. A common mistake in performing such calculations is to forget to square the wave function before integration.

Let us say one more thing about normalization. If you observe the Hamiltonian expression of the Schrödinger equations seen so far, then if λ is a real or complex constant we always have:

$$H(\lambda\Psi)=\lambda H(\Psi) \tag{11.2.340}$$

If we assume that Ψ is a solution of the Schrödinger equation, we see that $\lambda\Psi$ is also a solution of the equation. Indeed, we get:

$$H(\lambda\Psi)=\lambda H(\Psi)=\lambda E\Psi=E(\lambda\Psi) \tag{11.2.341}$$

Taking into account the fact that the function $\lambda\Psi$ is normalized, we then get:

$$\int|\lambda\Psi|^2\,\mathrm{d}x=1\Rightarrow|\lambda|^2\int|\Psi|^2\,\mathrm{d}x=1\Rightarrow|\lambda|^2=1 \tag{11.2.342}$$

hence (some textbooks restrict themselves to these solutions for pedagogical reasons):

$$\lambda=\pm 1\qquad\text{or}\qquad\lambda=\pm\mathrm{i} \tag{11.2.343}$$

Or rigorously we have more generally:

$$\lambda=e^{\mathrm{i}\theta} \tag{11.2.344}$$

where θ is a real number. This is what we name the "phase arbitrariness" which we had already mentioned at the beginning of this section without proof. We will also come back on this in the section of Quantum Field Theory for our study of Gauge theories applied to Quantum Physics.

These solutions are normalized and correspond to the same energy value E as the same probability density. This shows that it is not useful to look for the meaning of a negative or complex value of Ψ (if we take the particular pedagogical case of $\lambda=-1$), because $|\Psi|^2$ is real and is not negative (but nothing prevent us that the fact we require a result in $\mathbb{R}$ is wrong...). Only the square of a wave function, which corresponds to the probability density, seems to be interesting from a physical point of view.

11.2.4.2.3 Bound and unbound states

Let us suppose that $\Psi(x,y,z)$ decreases rapidly to infinity, so that the integral:

$$\int\limits_{-\infty}^{+\infty} |\,\Psi(x,y,z)\,|^2 \, \mathrm{d}x\mathrm{d}y\mathrm{d}z \tag{11.2.345}$$

converge. It is then possible to take advantage of the arbitrariness prevailing on the wave function (the fact that Ψ and $\lambda\Psi$ describe the same state) to make this integral equal to unity. We then say that Ψ is a "standardized field state function":

$$\int\limits_{-\infty}^{+\infty} |\,\Psi(x,y,z)\,|^2 \, \mathrm{d}x\mathrm{d}y\mathrm{d}z = 1 \tag{11.2.346}$$

We notice that there is still an arbitrary on Ψ by a complex number of module 1, $e^{\mathrm{i}\theta}$, without the condition of normalization being altered. We know that this is the "phase arbitrariness" and we have already introduced it just before!

Such a dynamic state is named "bound state" or "bound level", because the particle manifests itself in a limited region of space because of a potential. When, for example, the hydrogen atom is located on a fundamental level, it is in a bound state. We know that there is no chance of finding the electron at more than a few Angstroms of the proton, treated as infinitely heavy and originally placed as we have seen in the study of the Bohr model. Here is a good schematic view of the thing (bound state):

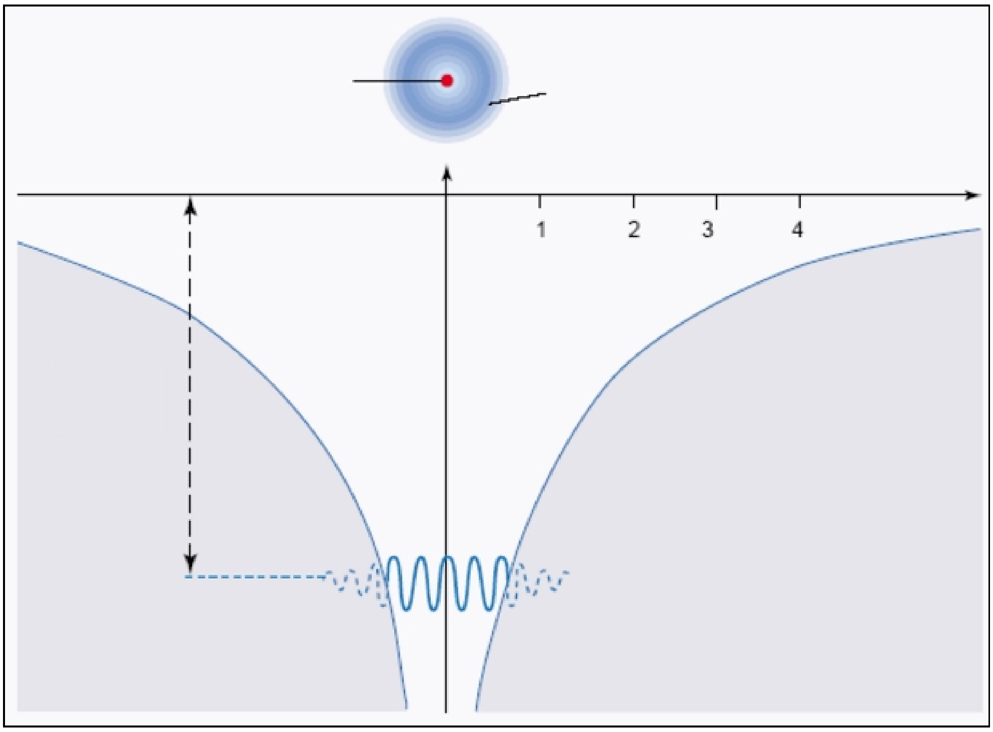


Figure 11.43: **Pictorial representation of a bound state (source: Pour la science)**

An example of an "unbounded (default) state" is the free particle that can propagate indefinitely in all directions of space (by the way, for the latter example it is a little more complicated ... but we will deal with it later).

Remark

It is perhaps good to know that these bound state concepts have classical analogues. Thus, the bound levels of the hydrogen atom correspond to the elliptic orbits, the unbound (positive energy) levels correspond to the hyperbolic orbits.

11.2.4.3 Classical Shrödinger equation of evolution

We know that in Classical Mechanics the dynamic state of a system evolves, in general, in time. This means that the position and the linear momentum (for example) are a function of time. For a given Hamiltonian system, knowledge of the initial dynamic state makes it possible to predict exactly the future evolution of this system due to the well-known properties of Hamilton's equations.

In Quantum Physics, dynamic states will generally evolve over time. The wave function describing a dynamic state will then not only depend on the coordinates of the particles constituting the system, but it will also depend on time and be written:

$$\Psi(x,y,z,t) \tag{11.2.347}$$

It is quite natural to assume, by analogy with Classical Mechanics, that for a given Hamiltonian system the knowledge of the initial dynamic state at the instant t_0 allows us to predict what will be the dynamic state of the system at a later time $t>t_0$.

On the way let us notice that this is equivalent to say that an initially "pure set" remains a pure set in the during the subsequent evolution of the systems which constitute it without external action. It would therefore cease to be true if all the systems of the whole had not exactly the same Hamiltonian.

Let us indicate that there are at least two simple possible approaches to determine the time-dependent functions:

- The first one, common in many fields of application of Quantum Physics, consists in using an "evolution operator" and makes it possible to make explicit the Schrödinger equation of evolution. We will begin with this one even if it is the most complicated or abstract approach in our point of view.
- The second, which is widely used for educational purposes, allows the time-dependent functions to be obtained by means of the technique of separation of the variables of differential equations (see section Differential and Integral Calculus page 1617) but requires to admit the Schrödinger equation of evolution as a postulate.

11.2.4.3.1 Operator of evolution

Given $\Psi(t)$ the normalized wave function describing the dynamic state of the system at time t (we do not write the other variables on which equation depends for simplicity, i.e. the spatial coordinates of the system's particles!). From the above, if $\Psi(t_0)$ is known, $\Psi(t)$ is also known. Then we have the correspondence:

$$\Psi(t) \rightarrow \Psi(t_0) \qquad (11.2.348)$$

and we will admit that this dependence is linear! There is therefore an operator $T(t,t_0)$, named "evolution operator", such that:

$$\Psi(t) = T(t,t_0)\Psi(t_0) \qquad (11.2.349)$$

The function $\Psi(t+\mathrm{d}t)$ depends also linearly on $\Psi(t)$. Then the same applies to:

$$\frac{\Psi(t+\mathrm{d}t)-\Psi(t)}{\mathrm{d}t} \qquad (11.2.350)$$

Therefore we assume that it exist a linear operator K, such that:

$$\frac{\partial\Psi}{\partial t} = \mathrm{i}K\Psi \qquad (11.2.351)$$

where the pure imaginary complex number i simply comes from the fact that we intuitively guess that the result will be a complex wave function. What also led the physicists to write the latter equality thus were the known results of the wave equation describing a dynamic state according to the idea of de Broglie. We will then show why writing this equality is justified.

We have to determine K. Since the knowledge of the Hamiltonian H controls the evolution of the system, so K must probably depend on H. To define the law that binds K to H, we will examine a particular case, that of the free particle (for which e will make a detailed study later). In this case, H is identifies with kinetic energy only.

According to de Broglie's ideas it is natural to assume that the wave function describing a dynamic state in which the linear momentum is well-defined, i.e. $p=\hbar k$ (relation proved during the study of the free particle), and where the total energy is thus also well defined, i.e. equal $E=\hbar\omega$, is a plane wave of the classical form:

$$\Psi(x,y,z,t)=\Psi_0 e^{\mathrm{i}(\vec{k}\circ\vec{r}+\omega\cdot t)} \tag{11.2.352}$$

where as we already know, and for recall, $\vec{k}$ is the wave vector of the wave and $\vec{r}=(x,y,z)$ its spatial coordinates.

It is common in Quantum Physics to take a wave that propagates in the $+X$ direction. Therefore (see section Wave Mechanics 3476) using the De Broglie relation ($E=\hbar\omega$):

$$\frac{\partial\Psi}{\partial t}=\frac{\partial\Psi_0 e^{\mathrm{i}(\vec{k}\circ\vec{r}-\omega\cdot t)}}{\partial t}=-\mathrm{i}\omega\Psi=\frac{E}{\hbar\mathrm{i}}\Psi \tag{11.2.353}$$

But we have the relation between the Hamiltonian operator and the corresponding eigenvalue:

$$H\Psi=E\Psi \tag{11.2.354}$$

The two preceding equations lead us to write:

$$\frac{\partial\Psi}{\partial t}=\frac{H}{\hbar\mathrm{i}}\Psi \tag{11.2.355}$$

By comparing this last relation with:

$$\frac{\partial\Psi}{\partial t}=\mathrm{i}K\Psi \tag{11.2.356}$$

We are led to put:

$$K=-\frac{H}{\hbar} \tag{11.2.357}$$

Physicists assume that this relation between K and H is general. Then the equation:

$$\frac{\partial\Psi}{\partial t}=\mathrm{i}K\Psi \tag{11.2.358}$$

In which K is replaced by its expression:

$$K=-\frac{H}{\hbar} \tag{11.2.359}$$

Therefore becomes:

$$\boxed{\mathrm{i}\hbar\frac{\partial\Psi}{\partial t}=H\Psi} \tag{11.2.360}$$

This equation constitutes the "Schrödinger's classical evolution equation" or "time-dependant Schrödinger's equation" which allows to study systems evolving over time, in particular the processes involving the absorption and the emission of radiation and the diffusion of the radiation by the atoms.

We recognize also in the previous relation the Energy operator:

$$\hat{E}=\mathrm{i}\hbar\frac{\partial}{\partial t} \tag{11.2.361}$$

In particular, for a spinless particle subjected to a potential energy E_p, by always maintaining that the relation between K and H is general, the equation of evolution is then written:

$$\mathrm{i}\hbar\frac{\partial\Psi}{\partial t}=\left(E_p-\frac{\hbar^2}{2m}\Delta\right)\Psi \tag{11.2.362}$$

where the terms in parentheses correspond to the expression of the Hamiltonian. Notice something quite important...! Time is derived once but space twice. This asymmetry of treatment between time and space is incompatible with what we know from Special Relativity where space and time and treated equally. This is an issue that we will solve much later with the Dirac equation.

The figure below illustrates how we can move from the Special Relativity Einstein's relation to the most famous equations in Quantum Physics:

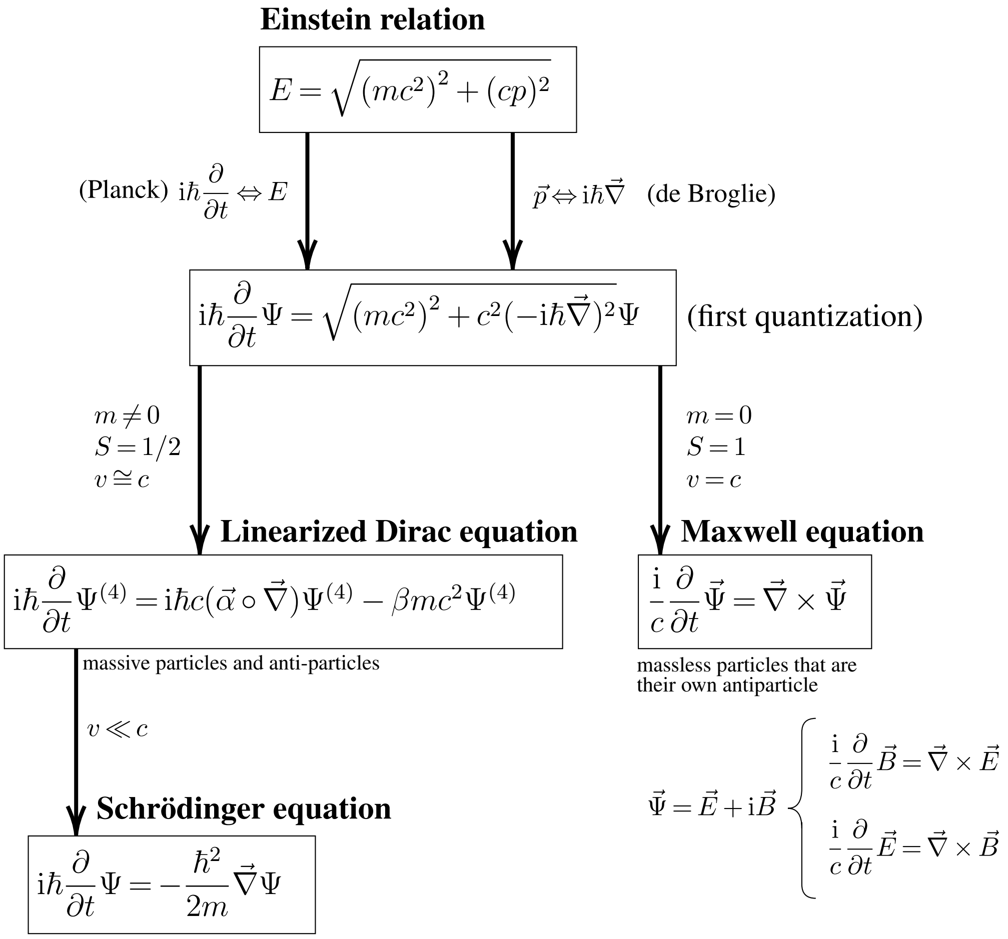

Einstein relation
$E=\sqrt{(mc^2)^2+(cp)^2}$
(Planck) $\mathrm{i}\hbar\frac{\partial}{\partial t}\Leftrightarrow E$
$\vec{p}\Leftrightarrow\mathrm{i}\hbar\vec{\nabla}$ (de Broglie)
$\mathrm{i}\hbar\frac{\partial}{\partial t}\Psi=\sqrt{(mc^2)^2+c^2(-\mathrm{i}\hbar\vec{\nabla})^2\Psi}$
(first quantization)
$m\neq 0$
$S=1/2$
$v\cong c$
$m=0$
$S=1$
$v=c$
Linearized Dirac equation
$\mathrm{i}\hbar\frac{\partial}{\partial t}\Psi^{(4)}=\mathrm{i}\hbar c(\vec{\alpha}\circ\vec{\nabla})\Psi^{(4)}-\beta mc^2\Psi^{(4)}$
massive particles and anti-particles
Maxwell equation
$\frac{\mathrm{i}}{c}\frac{\partial}{\partial t}\vec{\Psi}=\vec{\nabla}\times\vec{\Psi}$
massless particles that are their own antiparticle
$v\ll c$
$\vec{\Psi}=\vec{E}+\mathrm{i}\vec{B}$
$\frac{\mathrm{i}}{c}\frac{\partial}{\partial t}\vec{B}=\vec{\nabla}\times\vec{E}$
$\frac{\mathrm{i}}{c}\frac{\partial}{\partial t}\vec{E}=\vec{\nabla}\times\vec{B}$
Schrödinger equation
$\mathrm{i}\hbar\frac{\partial}{\partial t}\Psi=-\frac{\hbar^2}{2m}\vec{\nabla}\Psi$

Figure 11.44: **Relationships between most famous quantum physics equations**

It is now necessary to solve the Schrödinger differential equation of evolution. For this, we will use the condition of normalization of de Broglie.

Let us recall that this condition is written:

$$\int\limits_{-\infty}^{+\infty} |\,\Psi(x,y,z)\,|^2\,\mathrm{d}x\mathrm{d}y\mathrm{d}z=1 \tag{11.2.363}$$

and let us generalize this condition to a multidimensional and temporal study such that (according to the properties of complex numbers):

$$\int |\,\Psi(t,q)\,|^2\,\mathrm{d}q=\int \bar{\Psi}(t,q)\Psi(t,q)\mathrm{d}q=1 \tag{11.2.364}$$

Let us now perform the derivative with respect to time of the previous integral. We therefore necessarily have:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int \bar{\Psi}(t,q)\Psi(t,q)\mathrm{d}q&=\int\left[\frac{\mathrm{d}}{\mathrm{d}t}\bar{\Psi}\right]\Psi(t,q)\mathrm{d}q+\int\bar{\Psi}(t,q)\left[\frac{\mathrm{d}}{\mathrm{d}t}\Psi\right]\mathrm{d}q\\&=\int\dot{\bar{\Psi}}(t,q)\Psi(t,q)\mathrm{d}q+\int\bar{\Psi}(t,q)\dot{\Psi}(t,q)\mathrm{d}q=0\end{aligned} \tag{11.2.365}$$

and let use the Schrödinger equation of evolution:

$$\mathrm{i}\hbar\dot{\Psi}=H\Psi\Rightarrow\begin{cases}\dot{\Psi}=-\dfrac{\mathrm{i}}{\hbar}H\Psi\\ \dot{\bar{\Psi}}=+\dfrac{\mathrm{i}}{\hbar}\bar{H}\bar{\Psi}\end{cases} \tag{11.2.366}$$

What gives us for our integral after substitution:

$$\begin{aligned}&\int\left[-\frac{\mathrm{i}}{\hbar}H\bar{\Psi}(t,q)\right]\Psi(t,q)\mathrm{d}q+\int\bar{\Psi}(t,q)\left[\frac{\mathrm{i}}{\hbar}\bar{H}\Psi(t,q)\right]\mathrm{d}q\\&=\frac{\mathrm{i}}{\hbar}\left[\int\bar{\Psi}(t,q)\bar{H}\Psi(t,q)\mathrm{d}q-\int H\bar{\Psi}\Psi(t,q)\mathrm{d}q\right]=\int\bar{\Psi}(t,q)\bar{H}\Psi(t,q)\mathrm{d}q-\int H\bar{\Psi}(t,q)\Psi(t,q)\mathrm{d}q=0\end{aligned} \tag{11.2.367}$$

Let us now prove that we can write:

$$\int\bar{\Psi}(t,q)\bar{H}\Psi(t,q)\mathrm{d}q-\int H\bar{\Psi}(t,q)\Psi(t,q)\mathrm{d}q=\int\bar{\Psi}(t,q)\bar{H}\Psi(t,q)\mathrm{d}q-\int\bar{\Psi}(t,q)H\Psi(t,q)\mathrm{d}q \tag{11.2.368}$$

This would then be equivalent to the proof that H can act identically "backwards" such that:

$$\int H\bar{\Psi}(t,q)\Psi(t,q)\mathrm{d}q=\int\bar{\Psi}(t,q)H\Psi(t,q)\mathrm{d}q \tag{11.2.369}$$

where H can be (or "contain" if you prefer) an operator (differential one for example).

This equality can be proved if and only if Ψ is a decreasing function towards infinity and whose derivative tends towards zero towards the infinite (typical of bounded states)!

Let us prove this on a particular case (but frequent in physics) and to see how this can be done, let us consider in H, a particular term of the following form (we don't care of the constant corresponding the potential energy as it doesn't change anything in the result below):

$$H=f(q)\cdot\frac{\partial}{\partial q}\cdot g(q)=f\frac{\partial}{\partial q}g \tag{11.2.370}$$

Which leads us to write:

$$\int \bar{\Psi}(t,q)H\Psi(t,q)\mathrm{d}q=\int \bar{\Psi}(t,q)\left[f\frac{\partial}{\partial q}g\Psi(t,q)\right]\mathrm{d}q \tag{11.2.371}$$

By integration by parts (see section Differential and Integral Calculus page 1532) on the differential operator term $\partial_q(g\Psi(t,q))$, we get:

$$\int\limits_{-\infty}^{+\infty}\bar{\Psi}(t,q)f\cdot\frac{\partial}{\partial q}\big(g\Psi(t,q)\big)\mathrm{d}q=\bar{\Psi}(t,q)f\cdot g\Psi(t,q)\mid_{-\infty}^{+\infty}-\int\limits_{-\infty}^{+\infty}\frac{\partial}{\partial q}\big(\bar{\Psi}(t,q)f\big)\cdot g\Psi(t,q)\mathrm{d}q \tag{11.2.372}$$

But, since Ψ is a function decreasing towards infinity by hypothesis (physically necessary for bound states!), we will have the first term which will always be zero. It remains to us therefore:

$$\int\limits_{-\infty}^{+\infty}\bar{\Psi}f\cdot\frac{\partial}{\partial q}(g\Psi)\mathrm{d}q=-\int\limits_{-\infty}^{+\infty}\frac{\partial}{\partial q}(\bar{\Psi}f)\cdot g\Psi\mathrm{d}q=\int\limits_{-\infty}^{+\infty}\left(-\frac{\partial}{\partial q}f\bar{\Psi}\right)\cdot g\Psi\mathrm{d}q \tag{11.2.373}$$

So it makes no difference to consider that the operator differentiates everything that is on the right or whatever what is on the left, so far as it must be accepted that the latter case involves a change of sign and that the considerate states are bounded. It is customary to name this result sometimes the "condition of hermicity" ....

So we are indeed authorized to write:

$$\int \bar{\Psi}(t,q)\bar{H}\Psi(t,q)\mathrm{d}q-\int \bar{\Psi}(t,q)H\Psi(t,q)\mathrm{d}q=0 \tag{11.2.374}$$

Which also leads us to write (this highlight why the constant related to the potential energy was ignored in the proof above as it is obvious that the constants cancels each other):

$$\int \bar{\Psi}(t,q)(\bar{H}-H)\Psi(t,q)\mathrm{d}q=0 \tag{11.2.375}$$

This can only be satisfied if:

$$\bar{H}=H \tag{11.2.376}$$

And in the mathematical field dealing with operators we saw that we should note this equality:

$$H^{\dagger}=H \tag{11.2.377}$$

Which brings us to write:

$$\int \bar{\Psi}(t,q)(H^{\dagger}-H)\Psi(t,q)dq \tag{11.2.378}$$

Or using the representative (ket-bra) notation:

$$\langle\Psi\mid H^{\dagger}-H\mid\Psi\rangle=0 \tag{11.2.379}$$

To return to the resolution of:

$$\mathrm{i}\hbar\frac{\partial\Psi}{\partial t}=H\Psi \tag{11.2.380}$$

It is obvious that a possible solution is then:

$$\boxed{\Psi(t,q)=\Psi(0,q)e^{-\frac{\mathrm{i}t}{\hbar}H}} \tag{11.2.381}$$

which is therefore made of a purely spatial (independent of time) component and a time-dependent complex exponential. Let us check this:

$$\begin{aligned}\mathrm{i}\hbar\frac{\partial}{\partial t}\Psi(t,q)&=\mathrm{i}\hbar\left(\Psi(0,q)e^{-\frac{\mathrm{i}t}{\hbar}H}\right)=\mathrm{i}\hbar\left(\frac{\partial\Psi(0,q)}{\partial t}e^{-\frac{\mathrm{i}t}{\hbar}H}+\Psi(0,q)\frac{\partial e^{-\frac{\mathrm{i}t}{\hbar}H}}{\partial t}\right)\\&=\mathrm{i}\hbar\Psi(0,q)\left(\frac{\mathrm{i}}{\hbar}\right)He^{-\frac{\mathrm{i}t}{\hbar}H}=\Psi(0,q)He^{-\frac{\mathrm{i}t}{\hbar}H}=H\Psi(t,q)\end{aligned} \tag{11.2.382}$$

This is what had to be proved (...).

Let us also notice that once the purely spatial solutions are determined, the time- and space-dependent solutions are easily obtained.

Similarly, thanks to the relation $H^\dagger=H$ that we have demonstrated before, we can write:

$$\bar{\Psi}(t,q)=\bar{\Psi}(0,q)e^{\frac{\mathrm{i}t}{\hbar}H} \tag{11.2.383}$$

Finally, the relation:

$$\int\bar{\Psi}(t,q)X\Psi(t,q)\mathrm{d}q=1 \tag{11.2.384}$$

becomes:

$$\int\bar{\Psi}(0,q)e^{-\frac{\mathrm{i}t}{\hbar}H}X\Psi(0,q)e^{\frac{\mathrm{i}t}{\hbar}H}\mathrm{d}q=\int\bar{\Psi}(0,q)X(t)\Psi(0,q)\mathrm{d}q=1 \tag{11.2.385}$$

with the "Heisenberg operator (of evolution)" defined by:

$$\boxed{X(t)=e^{-\frac{\mathrm{i}t}{\hbar}H}Xe^{\frac{\mathrm{i}t}{\hbar}H}=e^{\frac{\mathrm{i}t}{\hbar}H}Xe^{-\frac{\mathrm{i}t}{\hbar}H}} \tag{11.2.386}$$

and with the "time evolution operator" given by:

$$\boxed{U(t,t_0)=e^{-iH(t-t_0)/\hbar}} \tag{11.2.387}$$

that is unitary since:

$$U^\dagger U=1 \tag{11.2.388}$$

Remark

It may very well happen that X is sometimes a simple constant (we will see an example further below).

One way is to think of the operator X as time-independent and to consider all of the time dependence of its expectation value as coming from the state vector. This point of view is known as the "Schrödinger picture". Alternatively, we may think of the operator as evolving in time, while the state vector stays constant. This is known as the "Heisenberg picture".

To sum up, the time evolution of a state vector following Schrödinger picture is given by:

$$\Psi(t,q)=\Psi(0,q)e^{-\frac{\mathrm{i}t}{\hbar}H} \tag{11.2.389}$$

or written more generally:

$$\Psi(t,q)=\Psi(t_0,q)e^{-\frac{\mathrm{i}(t-t_0)}{\hbar}H} \tag{11.2.390}$$

or in representative notation:

$$|\Psi(t,q)\rangle=e^{-\frac{\mathrm{i}(t-t_0)}{\hbar}H}|\Psi(t_0,q)\rangle \tag{11.2.391}$$

and that of an operator in the Heisenberg picture is:

$$X(t)=e^{-\frac{\mathrm{i}t}{\hbar}H}Xe^{\frac{\mathrm{i}t}{\hbar}H}=e^{\frac{\mathrm{i}t}{\hbar}H}Xe^{-\frac{\mathrm{i}t}{\hbar}H} \tag{11.2.392}$$

The Heisenberg picture is useful because we can see a closer connection to classical physics than with the Schrödinger picture. In classical physics, we describe the evolution of a system in terms of the time evolution of the observables, such as position or angular momentum, as dictated by the classical equations of motion. Classical mechanics does not include the concept of state vectors, as quantum mechanics does.

11.2.4.3.2 Conserved Quantities and Commutators

Let us recall the time-dependent Schrödinger equation obtained just earlier:

$$H\Psi=\mathrm{i}\hbar\frac{\mathrm{d}\Psi}{\mathrm{d}t} \tag{11.2.393}$$

We saw earlier that the expectation value of an observable physical quantity $\mathcal{O}$ is given by:

$$\langle\mathcal{O}\rangle=\int\limits_{-\infty}^{+\infty}\bar{\Psi}\hat{\mathcal{O}}\Psi\mathrm{d}x \tag{11.2.394}$$

where for recall $\hat{\mathbb{O}}$ is the operator representing the quantity (observable) $\mathcal{O}$. For $\mathcal{O}$ to be conserved, its value $\langle\mathcal{O}\rangle$ must not change in time, so the question is, when is $\langle\mathcal{O}\rangle$ independent of time? To answer that question, we assume that $\langle\hat{\mathcal{O}}\rangle$ is independent of time and compute $\mathrm{d}\langle\mathcal{O}\rangle/\mathrm{d}t$ as follows:

$$\frac{\mathrm{d}\langle\mathcal{O}\rangle}{\mathrm{d}t}=\frac{\mathrm{d}}{\mathrm{d}t}\int\limits_{-\infty}^{+\infty}\bar{\Psi}\hat{\mathcal{O}}\Psi\mathrm{d}x=\int\limits_{-\infty}^{+\infty}\frac{\mathrm{d}\bar{\Psi}}{\mathrm{d}t}\hat{\mathcal{O}}\Psi\mathrm{d}x+\int\limits_{-\infty}^{+\infty}\bar{\Psi}\hat{\mathcal{O}}\frac{\mathrm{d}\Psi}{\mathrm{d}t}\mathrm{d}x=\int\limits_{-\infty}^{+\infty}\left(\frac{\mathrm{d}\bar{\Psi}}{\mathrm{d}t}\hat{\mathcal{O}}\Psi+\bar{\Psi}\hat{\mathcal{O}}\frac{\mathrm{d}\Psi}{\mathrm{d}t}\right)\mathrm{d}x \tag{11.2.395}$$

The complex conjugate Schrödinger equation is:

$$\overline{(H\Psi)}=-\mathrm{i}\hbar\frac{\mathrm{d}\bar{\Psi}}{\mathrm{d}t} \tag{11.2.396}$$

Using the fact that H is real or hermitian, this can be written (in the case where the Hamiltonian is a matrix we use here the property $(AB)^T=B^TA^T$ proved in the section of Linear Algebra page 1951):

$$\bar{\Psi}H=\overline{(H\Psi)}=-\mathrm{i}\hbar\frac{\mathrm{d}\bar{\Psi}}{\mathrm{d}t} \tag{11.2.397}$$

Therefore combining the two previous time-dependent Schrödinger equation into the previous integral we get:

$$\begin{aligned}\frac{\mathrm{d}\langle\mathcal{O}\rangle}{\mathrm{d}t}&=\int\limits_{-\infty}^{+\infty}\left(\frac{\mathrm{d}\bar{\Psi}}{\mathrm{d}t}\hat{\mathcal{O}}\Psi+\bar{\Psi}\hat{\mathcal{O}}\frac{\mathrm{d}\Psi}{\mathrm{d}t}\right)\mathrm{d}x=\int\limits_{-\infty}^{+\infty}\left(-\frac{1}{\mathrm{i}\hbar}\bar{\Psi}H\hat{\mathcal{O}}\Psi+\bar{\Psi}\hat{\mathcal{O}}\frac{1}{\mathrm{i}\hbar}H\Psi\right)\mathrm{d}x\\&=\frac{\mathrm{i}}{\hbar}\int\limits_{-\infty}^{+\infty}\left(\bar{\Psi}H\hat{\mathcal{O}}\Psi-\bar{\Psi}\hat{\mathcal{O}}H\Psi\right)\mathrm{d}x=\frac{\mathrm{i}}{\hbar}\int\limits_{-\infty}^{+\infty}\bar{\Psi}\left(H\hat{\mathcal{O}}-\hat{\mathcal{O}}H\right)\Psi\mathrm{d}x\end{aligned} \tag{11.2.398}$$

So if the observable $\mathcal{O}$ is conserved then we must have:

$$\frac{\mathrm{d}\langle\mathcal{O}\rangle}{\mathrm{d}t}=\frac{\mathrm{i}}{\hbar}\int\limits_{-\infty}^{+\infty}\bar{\Psi}\left(H\hat{\mathcal{O}}-\hat{\mathcal{O}}H\right)\Psi\mathrm{d}x=0 \tag{11.2.399}$$

That is:

$$H\hat{\mathcal{O}}-\hat{\mathcal{O}}H=0 \tag{11.2.400}$$

Hence a conserved quantity in quantum mechanics can be defined as one for which the probabilities of measuring the various eigenvalues for that quantity are independent of time in all states. Mathematically conserved quantities correspond to Hermitian operators $\hat{\mathcal{O}}$ satisfying:

$$\boxed{[H,\hat{\mathcal{O}}]=0} \tag{11.2.401}$$

This is a very important result for our study of Wave Quantum Physics and also for the next section of Relativistic Quantum Physics!

11.2.4.3.3 Schrödinger picture by separation of variables

Let us also see an interesting mathematical manipulation and somewhat similar to the previous one of Schrödinger's equation of evolution. This manipulation will allow us to see that the separation of variables works very well with the equation of evolution and that it will allow us to fall back on a result obtained previously (it is always pedagogically to see several approaches).

We have therefore in a particular case:

$$\mathrm{i}\hbar\frac{\partial\Psi}{\partial t}=\left(E_p-\frac{\hbar^2}{2m}\Delta\right)\Psi \tag{11.2.402}$$

Rewritten in traditional form (according to the literature) and in one dimension, for a constant potential in time, this relation is then written:

$$-\frac{\hbar}{\mathrm{i}}\frac{\partial\Psi(x,t)}{\partial t}=-\frac{\hbar}{2m}\frac{\partial^2\Psi(x,t)}{\partial^2 x}+U(x)\Psi(x,t) \qquad (11.2.403)$$

Let us suppose now that the wave function can separate into two functions of which it is the product such that:

$$\Psi(x,t)=f(t)\Psi(x) \qquad (11.2.404)$$

We would then have:

$$\frac{\partial\Psi(x,t)}{\partial t}=\frac{\partial f(t)}{\partial t}\Psi(x) \qquad \text{and} \qquad \frac{\partial^2\Psi(x,t)}{\partial x^2}=f(t)\frac{\partial^2\Psi(x)}{\partial x^2} \qquad (11.2.405)$$

What injected into the one-dimensional evolution equation yields:

$$-\frac{\hbar}{\mathrm{i}}\frac{\partial f(t)}{\partial t}\Psi(x)=-\frac{\hbar}{2m}\frac{\partial^2\Psi(x)}{\partial^2 x}f(t)+U(x)\Psi(x)f(t) \qquad (11.2.406)$$

Which gives after simplification:

$$-\frac{\hbar}{\mathrm{i}}\frac{1}{f(t)}\frac{\partial f(t)}{\partial t}=-\frac{\hbar^2}{2m}\frac{1}{\Psi(x)}\frac{\partial^2\Psi(x)}{\partial^2 x}+U(x) \qquad (11.2.407)$$

The left term depends only on t, the right one on x. Since they are equal, they are necessarily also equal to a constant which has the dimension of an energy ($U(x)$ is a potential energy for recall).

So for the left term:

$$-\frac{\hbar}{\mathrm{i}}\frac{1}{f(t)}\frac{\partial f(t)}{\partial t}=H\Rightarrow \ln(f(t))=\frac{\mathrm{i}Ht}{\hbar}+c^{te} \qquad (11.2.408)$$

then:

$$f(t)=e^{c^{te}}e^{-\mathrm{i}\frac{Ht}{\hbar}}=Ae^{-\mathrm{i}\frac{Ht}{\hbar}} \qquad (11.2.409)$$

And for the right term:

$$-\frac{\hbar^2}{2m}\frac{1}{\Psi(x)}\frac{\partial^2\Psi(x)}{\partial^2 x}+U(x) \qquad (11.2.410)$$

Which can be written:

$$-\frac{\hbar^2}{2m}\frac{\partial^2\Psi(x)}{\partial^2 x}+U(x)\Psi(x)=H\Psi(x) \qquad (11.2.411)$$

After factorization:

$$\frac{\partial^2\Psi(x)}{\partial^2 x}+\frac{2m}{\hbar^2}[H-U(x)]\Psi(x)=0 \qquad (11.2.412)$$

Either with the notations of this textbook (after having rearranged a bit):

$$\left(E_p-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial^2 x}\right)\Psi=E_{\text{tot}}\Psi \qquad (11.2.413)$$

we thus fall back on the classical one-dimensional Schrödinger equation which is not bad at all as result!

Now, since we have put:

$$\Psi(x,t)=f(x)\Psi(x) \tag{11.2.414}$$

then we have finally:

$$\Psi(x,t)=Ae^{-\mathrm{i}\frac{Ht}{\hbar}}\Psi(x) \tag{11.2.415}$$

Which we can write under the notations of the preceding paragraphs:

$$\Psi(t,q)=\Psi(0,q)e^{-\mathrm{i}\frac{Ht}{\hbar}} \tag{11.2.416}$$

We find also this last relation in several different forms in the literature of which few samples:

$$\begin{gathered}\Psi(x,t)=\Psi(x)e^{-\mathrm{i}\frac{Ht}{\hbar}}\\ \Psi(x,t)=\varphi(x)e^{-\mathrm{i}\frac{Ht}{\hbar}}\\ \Psi(x,t)=\varphi(x)e^{-\mathrm{i}\frac{Et}{\hbar}}\\ |\Psi(x,t)\rangle=e^{-\mathrm{i}\frac{Ht}{\hbar}}|\Psi(x,t_0)\rangle\\ |\Psi(x,t-t_0)\rangle=\Psi(x,t_0)e^{-\mathrm{i}\frac{H(t-t_0)}{\hbar}}\end{gathered} \tag{11.2.417}$$

11.2.4.3.4 Linear combination of eigen-states (quantum superposition)

We must notice before we move on to another subject something very important that we had just mentioned in the second postulate!

Indeed, any equation of the following form seen previously:

$$\Psi(x)e^{-\mathrm{i}\frac{Ht}{\hbar}} \tag{11.2.418}$$

Is therefore a solution of Schrödinger's evolutionary equation and as in quantum systems the Hamiltonian can take (or be associated with) several discrete eigenvalues traditionally denoted by E_n, we then have, as mentioned at the beginning of this section, by the principle of linear combination of differential equations (see section Differential and Integral Calculus page 1605) the following more general solution:

$$\Psi(x,t)=\sum_{n=1}^{+\infty}\Psi_n(x)e^{-\mathrm{i}\frac{E_nt}{\hbar}} \tag{11.2.419}$$

of which we shall have several practical examples (of the discretization of the states of energy and that these are in infinite number) in this section and in that of Quantum Chemistry.

If we write the normalization constant of $\Psi_n(x)$ of the previous relation, then we have:

$$\Psi(x,t)=\sum_{n=1}^{+\infty}A_n\Psi_n(x)e^{-\mathrm{i}\frac{E_nt}{\hbar}} \tag{11.2.420}$$

This last relation would be written in the following traditional ket-bra form:

$$\boxed{|\,\Psi\rangle=\sum_{n=1}^{+\infty}A_n\Psi_n(x)e^{-\mathrm{i}\frac{E_nt}{\hbar}}=\sum_{n=1}^{+\infty}c_n\,|\,\psi_n\rangle} \tag{11.2.421}$$

where the constant coefficient c_n is obviously assimilated to A_n (admit that it notation is much simpler isn't it?). We can also found it with the phi notation instead than the psi notation in the literature:

$$|\,\Phi\rangle=\sum_{n=1}^{+\infty}c_n\,|\,\varphi_n\rangle \tag{11.2.422}$$

We then say that the state $|\Psi\rangle$ (or $|\,\Phi\rangle$) is a linear combination of elementary (eigen-)states $|\psi\rangle$ (or $|\,\varphi_n\rangle$). Therefore $|\Psi\rangle$ $|\Psi\rangle$ (or $|\,\Phi\rangle$) also represents a wave particle as being simultaneously in several different sub-states (i.e. eigenstates) $|\psi\rangle$ (or $|\,\varphi_n\rangle$).

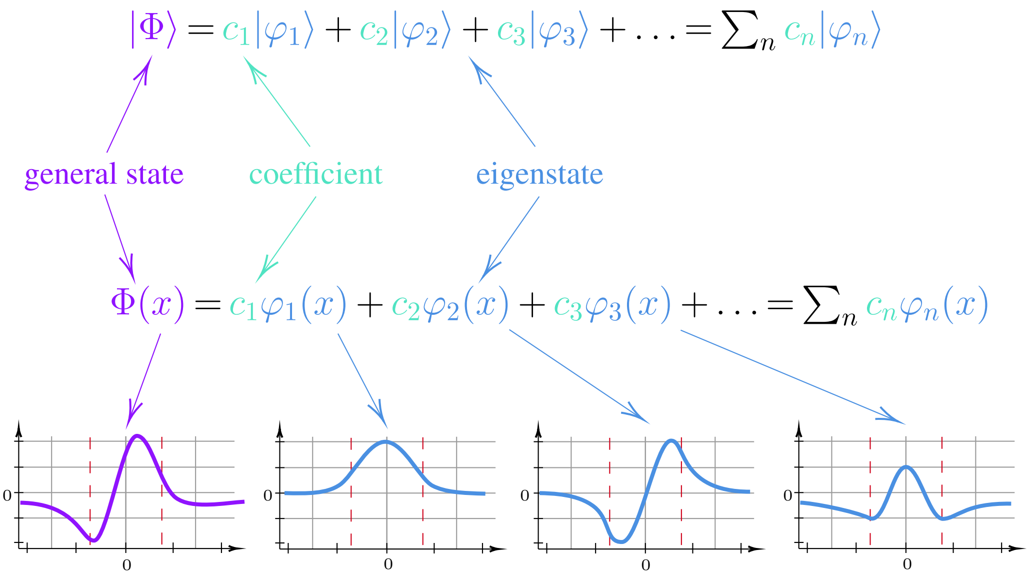

$|\Phi\rangle=c_1|\varphi_1\rangle+c_2|\varphi_2\rangle+c_3|\varphi_3\rangle+\ldots=\sum_n c_n|\varphi_n\rangle$
general state
coefficient
eigenstate
$\Phi(x)=c_1\varphi_1(x)+c_2\varphi_2(x)+c_3\varphi_3(x)+\ldots=\sum_n c_n\varphi_n(x)$
0
0
0
0
0
0
0
0

Figure 11.45: **Quantum state as a superposition of eigenstates**

✍**Example:**

Given $|\varphi\rangle$ and $|\chi\rangle$ that are vectors of $\mathcal{H}$ representing physical states, the unitary vector of $\mathcal{H}$:

$$|\psi\rangle=\frac{\lambda|\varphi\rangle+\mu|\chi\rangle}{||\lambda|\varphi\rangle+\mu|\chi\rangle||} \tag{11.2.423}$$

where $\lambda,\mu\in\mathbb{C}$, also represent a physical state.

It is interesting to notice that each solution:

$$A_n\Psi_n(x)e^{-\mathrm{i}\frac{E_nt}{\hbar}} \tag{11.2.424}$$

describes a "steady state" or "stationnary state". Let us see (at last!) rigorously what it is.

Indeed, we have:

$$\left|A_n\Psi_n(x)e^{-\mathrm{i}\frac{E_n t}{\hbar}}\right|^2 = |\ A_n\Psi_n(x)\ |^2 \left|e^{-\mathrm{i}\frac{E_n t}{\hbar}}\right|^2 = |\ A_n\Psi_n(x)\ |^2 \quad (11.2.425)$$

which is therefore independent of the time from which the origin of the name "stationary state" (we promised to define its origin at the beginning of section... so that is done!).

The functions being normalized we therefore have:

$$\int |\ \Psi\ |^2\,\mathrm{d}x = \int \bar{\Psi}\Psi\mathrm{d}x = \int \left(\sum_k \bar{A}_k\bar{\Psi}_k\right)\left(\sum_n A_n\Psi_n\right)\mathrm{d}x = \sum_k\sum_n \bar{A}_k A_n \int \bar{\Psi}_k\Psi_n\mathrm{d}x \quad (11.2.426)$$

The calculations above have showed us (we had proved it in two different ways) that the eigenfunctions have the following properties:

$$\int \bar{\Psi}_k\Psi_n\mathrm{d}x = 1 \quad (11.2.427)$$

When $k = n$ and:

$$\int \bar{\Psi}_k\Psi_n\mathrm{d}x = 0 \quad (11.2.428)$$

when $k \neq n$. That is to say with a more general notation:

$$\int \bar{\Psi}_k\Psi_n\mathrm{d}x = \delta_{ij} \quad (11.2.429)$$

It is this property that has led us in the third postulate to speak of "orthogonal basis of the stationary eigenfunctions".

Let us continue our calculation which can be written using the Kronecker symbol (see section Tensor Calculus page 2044):

$$\int |\ \Psi\ |^2\,\mathrm{d}x = \sum_k\sum_n \bar{A}_k A_n \int \bar{\Psi}_k\Psi_n\mathrm{d}x = \sum_k\sum_n \bar{A}_k A_n\delta_{kn} = \sum_k \bar{A}_k A_k = \sum_k |\ A_k\ |^2 = 1 \quad (11.2.430)$$

We can then interpret the term $|A_k|^2$ as the weight of the eigenfunction Ψ_k in the quantum state k, the probability of actually being in the eigenstate Ψ_k is then equal to $|A_k|^2$ the normalization then imposes:

$$\sum_k |\ A_k\ |^2 = 1 \quad (11.2.431)$$

We must therefore remember that any quantum state can always (as far as we know) be interpreted as a linear combination of eigenstates. The coefficient A_k of an eigenfunction/eigenstate Ψ_k is then associated with a probability $|A_k|^2$.

It is this mathematical result, very important, which is at the origin of the paradox of "Schrödinger's cat" (among others ...) and of many debates.

To close this small subject, let us notice one thing:

If the coefficients A_n are not already normalized, then the physicists denote their normalization as follows:

$$|A_n|^2 = \frac{|A_n|^2}{\sum\limits_n |A_n|^2} \tag{11.2.432}$$

because very often they use the same notation for the normalized coefficient and the non normalized one in their developments ...

The writing of the last relation is easily justified because let us recall that we must have:

$$\sum_n |A_n|^2 = 1 \tag{11.2.433}$$

and we have indeed after rearranging:

$$|A_n|^2 = \frac{|A_n|^2}{\sum\limits_n |A_n|^2} \Rightarrow \sum_n |A_n|^2 = \frac{|A_n|^2}{|A_n|^2} = 1 \tag{11.2.434}$$

Let us notice that with the traditional ket-bra notation, the relation:

$$|A_n|^2 = \frac{|A_n|^2}{\sum\limits_n |A_n|^2} \tag{11.2.435}$$

is often written in some specialized textbooks:

$$p(x_n) = \frac{|c_n|^2}{\sum\limits_n |c_n|^2} \tag{11.2.436}$$

which therefore always gives the probability of finding the state n at the position x.

With these definitions it is easy to describe the process of collapse. For any observable, the wave function is initially some linear combination of the eigenbasis $\{|\phi_i\rangle\}$ of that observable. When an external agency (an observer, experimenter) measures the observable associated with the eigenbasis $\{|\phi_i\rangle\}$, the wave function collapses from the full $|\psi\rangle$ to just one of the basis eigenstates, $|\phi_i\rangle$, that is:

$$|\psi\rangle \rightarrow |\phi_i\rangle\,.$$

The probability of collapsing to a given eigenstate $|\phi_k\rangle$ is the Born probability, $P_k = |c_k|^2$. Immediately post-measurement, other elements of the wave function vector, $c_{i\neq k}$, have "collapsed" to zero, and $|c_i|^2 = 1$.

More generally, collapse is defined for an operator $\hat{Q}$ with eigenbasis $\{|\phi_i\rangle\}$. If the system is in state $|\psi\rangle$, and $\hat{Q}$ is measured, the probability of collapsing the system to eigenstate $|\phi_i\rangle$ and measuring the eigenvalue q_i of $|\phi_i\rangle$ with respect to $\hat{Q}$ would be $|\langle\psi\,|\,\phi_i\rangle|^2$. Note that this is not the probability that the particle is in state $|\phi_i\rangle$; it is in state $|\psi\rangle$ until cast to an eigenstate of $\hat{Q}$!

11.2.4.3.5 Schrödinger continuity equation

Let us consider now the important example of the equation of evolution for a free particle, that is to say with $E_p=0$. Therefore we have:

$$-\frac{\hbar}{\mathrm{i}}\frac{\partial\Psi}{\partial t}+\frac{\hbar^2}{2m}\Delta\Psi=0 \tag{11.2.437}$$

The probability of finding the particle in a volume V is, as we have seen, given by:

$$\int_V \rho(\vec{r},t)\mathrm{d}\vec{r}=\int_V \mid\Psi(\vec{r},t)\mid^2\mathrm{d}\vec{r} \tag{11.2.438}$$

Hence:

$$\frac{\partial}{\partial t}\int_V \rho(\vec{r},t)\mathrm{d}\vec{r}=\int_V\left(\bar{\Psi}(\vec{r},t)\frac{\partial\Psi(\vec{r},t)}{\partial t}+\frac{\partial\bar{\Psi}(\vec{r},t)}{\partial t}\Psi(\vec{r},t)\right)\mathrm{d}\vec{r} \tag{11.2.439}$$

Taking into account the equation of evolution of the free particle, the second term of equality is written:

$$\frac{\partial}{\partial t}\int_V \rho(\vec{r},t)\mathrm{d}\vec{r}=\frac{\mathrm{i}\hbar^2}{2m}\int_V\left(\bar{\Psi}\vec{\nabla}^2\Psi-\vec{\nabla}^2\bar{\Psi}\Psi\right)\mathrm{d}\vec{r}=-\int_V\vec{\nabla}\circ\vec{j}\mathrm{d}\vec{r} \tag{11.2.440}$$

Where we have put:

$$\vec{j}=\frac{\hbar}{2m\mathrm{i}}\left(\bar{\Psi}(\vec{\nabla}\cdot\Psi)-(\vec{\nabla}\cdot\bar{\Psi})\Psi\right) \tag{11.2.441}$$

According to the Ostrogradsky theorem (see section Vector Calculus page 1882), it therefore comes:

$$\frac{\partial}{\partial t}\int_V \rho(\vec{r},t)\mathrm{d}\vec{r}=-\int_V\vec{j}\circ\mathrm{d}\vec{S} \tag{11.2.442}$$

where the integral of the right is carried out on the surface S which limits the volume V. The preceding relation therefore expresses indeed that the variation per unit of time of the probability of finding the particle in V is equal to the flux passing through the surface S and the vector $\vec{j}$ can be interpreted as a probability current density which satisfies the continuity equation as we have determined it in the section of Thermodynamics:

$$\frac{\partial}{\partial t}\int_V \rho(\vec{r},t)\mathrm{d}\vec{r}=-\int_V\vec{\nabla}\circ\vec{j}\mathrm{d}\vec{r}\Leftrightarrow\int_V\left[\frac{\partial}{\partial t}\rho(\vec{r},t)\right]\mathrm{d}\vec{r}=\int_V\left[-\vec{\nabla}\circ\vec{j}\right]\mathrm{d}\vec{r} \tag{11.2.443}$$

Hence:

$$\frac{\partial\rho}{\partial t}+\vec{\nabla}\circ\vec{j}=0 \tag{11.2.444}$$

In Quantum physics, it seems therefore that there is conservation of the particles flow: there is no creation, neither destruction of particle, but in the Nature (experimental observations) we observe such phenomenons of destruction and creation (but in fact the particle must be in a non-zero potential field)... In other words it means that a particle can not appear or disappear in a given volume V, there must be a particle flow in the walls of V for particles entering or leaving V.

That latter relation is more often written:

$$\boxed{\frac{\partial|\Psi|^2}{\partial t}+\vec{\nabla}\circ\vec{j}=0} \tag{11.2.445}$$

Intuitively, the above quantities indicate this represents the flow of probability. The chance of finding the particle at some position r and time t flows like a fluid; hence the term probability current, a vector field. The particle itself does not flow deterministically in this vector field.

11.2.4.4 Implications and Applications

The different definitions and tools that have been seen before, sometimes very abstract, will allow us to study some fundamental cases that lead to quite splendid results.

In a first step, we will see how to treat the case of the free particle (unbound state) and what are the problems of this simple configuration.

Then we solve the Schrödinger equation with a spin-free particle in a potential well with rectilinear infinite walls and show that with the formalism of quantum physics we will find the same results as the Bohr model (even more generalized!).

After this, we will introduce the study of the harmonic oscillator by revisiting briefly the resolution of the Schrödinger equation of a free particle. This example constitutes a form of introduction to the theoretical study of quantum atomic systems. It is in this example that we will use all the power of functional linear operators. It will therefore be important not to read it in diagonal (if possible and if the reader is still motivated).

We will also have to study another famous phenomenon, the "tunnel effect"! Obviously, we decided to introduce a particular case so that the reader can see the reasoning that led to the discovery of this astonishing (but logical) phenomenon. Again, this example will support the validity of quantum theory and demonstrate the value of the decay constants of many nuclear isotopes!

With respect to relativistic cases, with or without spin, we refer the reader to the sections of Relativistic Quantum Physics, and as regards the simple atomic model, we refer it to the section on Quantum Chemistry.

Enjoy!

11.2.4.4.1 Free particles (zero potential)

Curiously, the resolution of the Schrödinger equation for a free particle (where the potential is zero) is the simple ... most complex case ... mathematically speaking because the integration bounds of normalization are infinite. Let us see this!

Let us recall first that we have proved in a simplified way in the section of Sequences and Series that the Fourier transform (page 1758) of a function f and its inverse were given by:

$$\mathcal{F}(\vec{k})=\frac{1}{(2\pi)^{3/2}}\iiint f(\vec{r})e^{-\mathrm{i}\vec{k}\circ\vec{r}}\mathrm{d}^3r \quad f(\vec{r})=\frac{1}{(2\pi)^{3/2}}\iiint F(\vec{k})e^{\mathrm{i}\vec{k}\circ\vec{r}}\mathrm{d}^3k \tag{11.2.446}$$

Either in a one-dimensional form:

$$\mathcal{F}(k)=\frac{1}{\sqrt{2\pi}}\int f(x)e^{-\mathrm{i}kx}\mathrm{d}x \qquad f(x)=\frac{1}{\sqrt{2\pi}}\int F(k)e^{\mathrm{i}kx}\mathrm{d}k \tag{11.2.447}$$

Let us now proceed to the change of variable which connects the wave number k to the linear momentum (relation introduced at the beginning of this section):

$$k=\frac{P}{\hbar} \tag{11.2.448}$$

Which gives us:

$$\mathcal{F}(k)=\frac{1}{\sqrt{2\pi}}\int f(x)e^{-\mathrm{i}\frac{P}{\hbar}x}\mathrm{d}x \qquad f(x)=\frac{1}{\sqrt{2\pi}}\int F(k)e^{\mathrm{i}\frac{P}{\hbar}x}\mathrm{d}p \tag{11.2.449}$$

Let us now return to the Schrödinger's equation of evolution:

$$\mathrm{i}\hbar\frac{\partial\Psi}{\partial t}=\left(E_p-\frac{\hbar^2}{2m}\Delta\right)\Psi \tag{11.2.450}$$

If the particle is free, there is no potential and at one dimension we then have:

$$\mathrm{i}\frac{\partial\Psi}{\partial t}=-\frac{\hbar}{2m}\frac{\partial^2}{\partial x^2}\Psi \tag{11.2.451}$$

This differential equation admits monochromatic plane wave solutions of the type (see section Electrodynamics page 4142):

$$\Psi(x,t)=A^{\mathrm{i}(kx-\omega t)} \tag{11.2.452}$$

with obviously the small nuance that we have to use the relation (otherwise it does not work by cons!):

$$\omega=\frac{\hbar k^2}{2m}=\frac{p^2}{2m\hbar} \tag{11.2.453}$$

without forgetting that (this will be useful afterwards):

$$E=\frac{\hbar p^2}{2m}=\frac{\hbar^2k^2}{2m} \tag{11.2.454}$$

The curve of the energy E as a function of the wave vector k is sometimes named a "dispersion curve" and it is a parabola (since k is squared) for a free particle!

Obviously, the probability density of this solution is equal to:

$$\langle\Psi\mid\Psi\rangle=\bar{\Psi}\Psi=A^{\mathrm{i}(kx-\omega t)}\bar{A}^{-\mathrm{i}(kx-\omega t)}=\mid A\mid^2 \tag{11.2.455}$$

but this can not correspond to reality because! Indeed, a monochromatic planar wave of constant norm will have to carry an infinite energy and this is not possible!

In fact the solution comes from the fact that the true solution uses the principle of superposition of all the monochromatic waves of all the frequencies such as:

$$\Psi(x,t)=\int\limits_{-\infty}^{+\infty}A(k)e^{\mathrm{i}(kx-\omega(k)t)}\mathrm{d}k \tag{11.2.456}$$

and we fall back here a relation very similar to an inverse Fourier transform (see section Sequences and Series page 1758). Such a superposition of plane waves is named "one-dimensional wave packet".

What we can rewrite:

$$\Psi(x,t)=\int\limits_{-\infty}^{+\infty}A(k)e^{\mathrm{i}kx}e^{-\mathrm{i}\frac{\hbar k^2}{2m}t}\mathrm{d}k=\int\limits_{-\infty}^{+\infty}A(p)e^{\mathrm{i}\frac{P}{\hbar}x}e^{-\mathrm{i}\frac{p^2}{2m\hbar}t}\mathrm{d}p=\int\limits_{-\infty}^{+\infty}A(p)e^{\mathrm{i}\frac{P}{\hbar}x}e^{-\mathrm{i}\frac{E}{\hbar}t}\mathrm{d}p \tag{11.2.457}$$

But, we see immediately that we will not be able to normalized following:

$$\int\limits_{-\infty}^{+\infty}\mid\Psi\mid^2\mathrm{d}x=\int\limits_{-\infty}^{+\infty}\bar{\Psi}(x,t)\Psi(x,t)\mathrm{d}x=1 \tag{11.2.458}$$

Therefore there is no longer a general solution. It is necessary to give a carrier envelope to the waves imposing a possible normalization. This carrier envelope can be a Dirac or a Gaussian (see below this special case) or other functions of more or less complex distributions. Then physicists must use a property of the Fourier transforms which naturally reveal the true Heisenberg's uncertainties. Thus, the latter are a condition for the normalization of free particles using Fourier transforms.

To date, as already mentioned during our study of classical Heisenberg's uncertainties, we have no pedagogical and simple proof to propose on this last point. It will come in a near of far future.

On the other hand, we can take as a trivial solution the eigenmodes of the particle such that:

$$\Psi_x=e^{\pm\mathrm{i}(kx-\omega t)} \tag{11.2.459}$$

Indeed:

$$\begin{aligned}&\mathrm{i}\frac{\partial\Psi}{\partial t}=-\frac{\hbar}{2m}\frac{\partial^2}{\partial x^2}\Psi\Rightarrow\mathrm{i}\frac{\partial e^{\mathrm{i}(kx-\omega t)}}{\partial t}=-\frac{\hbar}{2m}\frac{\partial^2 e^{\mathrm{i}(kx-\omega t)}}{\partial x^2}\\&\Rightarrow-\mathrm{i}^2\omega e^{\mathrm{i}(kx-\omega t)}=-\frac{\hbar}{2m}\mathrm{i}^2k^2e^{\mathrm{i}(kx-\omega t)}\Rightarrow-\omega=-\frac{\hbar}{2m}k^2\\&\Rightarrow\omega=\frac{\hbar}{2m}k^2=\frac{p^2}{2m\hbar}\end{aligned} \tag{11.2.460}$$

This is what we will use as a relation for our study of the harmonic oscillator further below.

Before studying the particular case of the quasi-monochromatic wave packet, we will recall some results concerning the sum of two plane waves.

Let us start by summing two monochromatic plane waves of neighbouring frequencies:

$$\Psi_1(x,t)=Ae^{\mathrm{i}(k_1x-\omega_1t)} \qquad \text{and} \qquad \Psi_2(x,t)=Ae^{\mathrm{i}(k_2x-\omega_2t)} \tag{11.2.461}$$

with:

$$\omega_1=\omega_0-\Delta\omega \qquad \text{and} \qquad \omega_2=\omega_0-\Delta\omega \tag{11.2.462}$$

and:

$$k_1=k_0-\Delta k \qquad \text{and} \qquad k_2=k_0-\Delta k \tag{11.2.463}$$

Notice that we therefore impose:

$$\frac{\Delta\omega}{\omega_0}\ll 1 \qquad \text{and} \qquad \frac{\Delta k}{k_0}\ll 1 \tag{11.2.464}$$

The resultant wave is the expressed as:

$$\begin{aligned}\Psi(x,t)&=\Psi_1(x,t)+\Psi_2(x,t)=A\left(e^{\mathrm{i}(k_1x-\omega_1t)}+e^{\mathrm{i}(k_2x-\omega_2t)}\right)\\&=A(\cos(k_1x-\omega_1t)+\mathrm{i}\sin(k_1x-\omega_1t)+\cos(k_2x-\omega_2t)+\mathrm{i}\sin(k_2x-\omega_2t))\\&=A\begin{pmatrix}\cos((k_0-\Delta k)x-(\omega_0-\Delta\omega)t)+\mathrm{i}\sin((k_0-\Delta k)x-(\omega_0-\Delta\omega)t)+\\ \cos((k_0+\Delta k)x-(\omega_0+\Delta\omega)t)+\mathrm{i}\sin((k_0+\Delta k)x-(\omega_0+\Delta\omega)t)\end{pmatrix}\\&=A\begin{pmatrix}\cos((k_0-\Delta k)x-(\omega_0-\Delta\omega)t)+\cos((k_0+\Delta k)x-(\omega_0+\Delta\omega)t)+\\ \mathrm{i}\sin((k_0-\Delta k)x-(\omega_0-\Delta\omega)t)+\mathrm{i}\sin((k_0+\Delta k)x-(\omega_0+\Delta\omega)t)\end{pmatrix}\end{aligned} \tag{11.2.465}$$

Either by using the remarkable trigonometric relations (see section Trigonometry page 2519):

$$\begin{aligned}\Psi(x,t)&=A(2\cos(k_0x-\omega_0t)\cos(\Delta kx-\Delta\omega t)+2\mathrm{i}\sin(k_0x-\omega_0t)\cos(\Delta kx-\Delta\omega t))\\&=2A\cos(\Delta kx-\Delta\omega t)(\cos(k_0x-\omega_0t)+\mathrm{i}\sin(k_0x-\omega_0t))\\&=2A\cos(\Delta kx-\Delta\omega t)e^{\mathrm{i}(k_0x-\omega_0t)}=\Psi_0e^{\mathrm{i}(k_0x-\omega_0t)}\end{aligned} \tag{11.2.466}$$

which is a plane wave propagating along the positive x (see section Wave Mechanics page 3476) with the pulsation ω_0 and the mean wave vector k_0, and therefore at the phase velocity:

$$v_\varphi=\frac{\omega_0}{k_0}=\frac{\frac{\hbar k_0^2}{2m}}{k_0}=\frac{\hbar k_0}{2m}=\frac{\hbar\frac{\sqrt{E2m}}{\hbar}}{2m}=\frac{\sqrt{E2m}}{2m}=\sqrt{\frac{E}{2m}} \tag{11.2.467}$$

The cosine term is then interpreted as the slowly variable amplitude of this plane wave ("amplitude modulation" as we have seen in the section of Wave Mechanics).

Notice a rather important point !: The phase velocity is not consistent with the velocity we get by using the kinetic energy of a free particle. Indeed:

$$E_c=\frac{1}{2}mv^2\Leftrightarrow v=\sqrt{\frac{2E_c}{m}} \tag{11.2.468}$$

Therefore the phase velocity does not represent the velocity in the usual conventional sense but that of the wave travelling at the group velocity:

$$v_{\text{group}}=\frac{\Delta\omega}{\Delta k}\underset{\Delta\to 0}{=}\frac{\mathrm{d}\omega}{\mathrm{d}k}=\frac{\mathrm{d}}{\mathrm{d}k}\left(\frac{\hbar k^2}{2m}\right)=\frac{\hbar 2k}{2m}=\frac{\hbar k}{m}=\frac{\hbar\dfrac{\sqrt{E2m}}{\hbar}}{m}=\sqrt{\frac{2E}{m}} \tag{11.2.469}$$

where we thus fall back on the classic formulation of the velocity from kinetic energy (not bad...)!

We can easily represent all this with Maple™ 4.00b:

```
>restart:with(plots):
>lambda[0]:=1; T[0]:=1; k[0]:=2*Pi/lambda[0]; w[0]:=2*Pi/T[0];
>delta_k:=k[0]/8:  k[1]:=k[0]-delta_k; k[2]:=k[0]+delta_k;
delta_w:=w[0]/10:  w[1]:=w[0]-delta_w; w[2]:=w[0]+delta_w;
>P1:=animate(cos(k[1]*x-w[1]*t)+cos(k[2]*x-w[2]*t), x=0..1*2*Pi/delta_k,
t=0..2*Pi/delta_w, numpoints=200, frames=15, color=red):
>P2:=animate(2*cos(-1/2*k[1]*x+1/2*w[1]*t+1/2*k[2]*x-1/2*w[2]*t),
-2*cos(-1/2*k[1]*x+1/2*w[1]*t+1/2*k[2]*x-1/2*w[2]*t),
x=0..1*2*Pi/delta_k,
t=0..2*Pi/delta_w, numpoints=100, frames=15, color=blue):
>display(P1,P2);
```

Which gives:

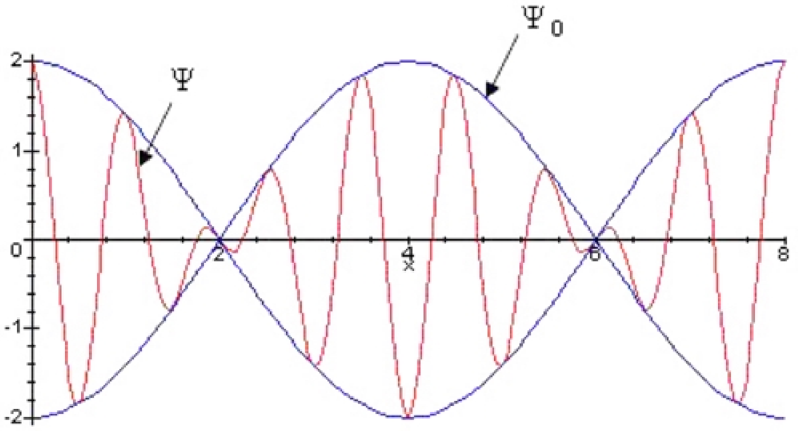


Figure 11.46: Representation of the concept of phase velocity and group velocity

Unlike the harmonic plane wave, this wave does not have a constant norm: its norm is zero in certain zones. On the other hand, it always extends over an infinite distance, and therefore has a infinite norm (the sum of the probability over all space). It therefore has no physical meaning (at least as far as we know).

The preceding study can be extended by summing an increasing number N of plane waves in the neighbourhood of ω_0 and k_0. Such a superposition leads to an increasingly localized function in certain areas of space (in particular near $x=0$, for example, for $t=0$), the distance between these zones increasing proportionally with N. At the limit $N\rightarrow+\infty$, then only the zone near $x=0$ remains, the others being rejected to infinity. The transition to this limit $N\rightarrow+\infty$ is done by replacing the discrete sum on plane waves by a continuous summation i.e. by an integral of the form:

$$\Psi(x,t)=\int\limits_{k_0-\delta k/2}^{k_0+\delta k/2}F(k)e^{\mathrm{i}(kx-\omega t)}\mathrm{d}k \tag{11.2.470}$$

with:

$$k\in\left[k_0-\frac{\delta k}{2},k_0+\frac{\delta k}{2}\right] \tag{11.2.471}$$

with therefore:

$$F(k)=A \qquad \text{and} \qquad \frac{\delta k}{k_0}\ll 1 \tag{11.2.472}$$

Such a packet is therefore named a "quasi-monochromatic wave packet".

This expression can be rewritten:

$$\Psi(x,t)=A\int\limits_{k_0-\delta k/2}^{k_0+\delta k/2}e^{\mathrm{i}(kx-\omega t)}\mathrm{d}k \tag{11.2.473}$$

It is important to understand that ω is a function of k given by the dispersion equation. We will compute this expression using the fact that $\delta k/k_0\ll 1$. This condition implies that $\delta\omega/\omega_0\ll 1$. It is possible to carry out a limited Taylor development in the neighbourhood of ω_0:

$$\omega=\omega_0+(k-k_0)\left(\frac{\partial\omega}{\partial k}\right)_{k=k_0}+\cdots=\omega_0+v_{\text{group}}(k-k_0) \tag{11.2.474}$$

Therefore:

$$kx-\omega t=(k-k_0+k_0)x-[\omega_0+v_{\text{group}}(k-k_0)]t=k_0x-\omega t+(x-v_{\text{group}}t)(k-k_0) \tag{11.2.475}$$

Let us put $y=k-k_0$:

$$\Psi(x,t)=Ae^{\mathrm{i}(k_0x-\omega_0t)}\int\limits_{-\delta k/2}^{+\delta k/2}e^{\mathrm{i}(x-v_{\text{group}}t)y}\mathrm{d}y \tag{11.2.476}$$

Let us calculate the integral:

$$\Psi(x,t)=Ae^{\mathrm{i}(k_0x-\omega_0t)}\int\limits_{-\delta k/2}^{+\delta k/2}e^{\mathrm{i}(x-v_{\text{group}}t)y}\mathrm{d}y=\frac{Ae^{\mathrm{i}(k_0x-\omega_0t)}}{\mathrm{i}(x-v_{\text{group}}t)}e^{\mathrm{i}\alpha y}\Bigg|_{-\delta k/2}^{+\delta k/2} \tag{11.2.477}$$

with:

$$\alpha = x - v_{\text{group}}t \tag{11.2.478}$$

Therefore:

$$\begin{aligned}
\Psi(x,t) &= \frac{A}{\mathrm{i}(x-v_{\text{group}}t)}\left(e^{\mathrm{i}\alpha\frac{\delta k}{2}} - e^{-\mathrm{i}\alpha\frac{\delta k}{2}}\right)e^{\mathrm{i}(k_0x-\omega_0t)}\\
&= \frac{A}{\mathrm{i}(x-v_{\text{group}}t)}\left(\cos\left(\alpha\frac{\delta k}{2}\right)+\mathrm{i}\sin\left(\alpha\frac{\delta k}{2}\right)-\cos\left(\alpha\frac{\delta k}{2}\right)+\mathrm{i}\sin\left(\alpha\frac{\delta k}{2}\right)\right)e^{\mathrm{i}(k_0x-\omega_0t)}\\
&= \frac{A}{\mathrm{i}(x-v_{\text{group}}t)}\left(\mathrm{i}\sin\left(\alpha\frac{\delta k}{2}\right)+\mathrm{i}\sin\left(\alpha\frac{\delta k}{2}\right)\right)e^{\mathrm{i}(k_0x-\omega_0t)}\\
&= 2A\frac{\sin\left((x-v_{\text{group}}t)\dfrac{\delta k}{2}\right)}{(x-v_{\text{group}}t)}e^{\mathrm{i}(k_0x-\omega_0t)} = 2A\frac{\delta k}{2}\frac{\sin\left((x-v_{\text{group}}t)\dfrac{\delta k}{2}\right)}{(x-v_{\text{group}}t)\dfrac{\delta k}{2}}e^{\mathrm{i}(k_0x-\omega_0t)}\\
&= A\delta k\frac{\sin\left((x-v_{\text{group}}t)\frac{\delta k}{2}\right)}{(x-v_{\text{group}}t)\frac{\delta k}{2}}e^{\mathrm{i}(k_0x-\omega_0t)} = A\delta k\,\mathrm{sinc}\left((x-v_{\text{group}}t)\frac{\delta k}{2}\right)e^{\mathrm{i}(k_0x-\omega_0t)}\\
&= \Psi_0e^{\mathrm{i}(k_0x-\omega_0t)}
\end{aligned} \tag{11.2.479}$$

The last term is again interpreted as a plane wave moving at the phase velocity:

$$v_\varphi = \frac{\omega_0}{k_0} \tag{11.2.480}$$

The amplitude of this plane wave is given by a sinus cardinal function. At $t=0$, this sinc function has important values only in the zone where:

$$\delta x = \frac{2\pi}{\delta k} \tag{11.2.481}$$

It is therefore a well-localized function. Consequently, Ψ is a summable square function. The calculation gives:

$$\int\limits_{-\infty}^{+\infty} |\,\Psi\,|^2\,\mathrm{d}x = |\,A\,|^2\,\delta k \tag{11.2.482}$$

The function can therefore be normalized by putting:

$$A = \frac{1}{\sqrt{\delta k}} \tag{11.2.483}$$

We have therefore succeeded in obtaining a function satisfying both the Schrödinger equation and the normalization condition, by using an infinite sum of harmonic waves. The example we have dealt with is only one particular case. Other types of wave packets can be obtained by taking other distributions for the amplitudes of the plane waves that make up the packet (we assumed here that they all had the same amplitude). Consequently, the group velocity is conventionally associated with the velocity of the particle of mass m and linear momentum p.

Thus, the wave packet moves globally at the group velocity, which is identified with the velocity given by Classical Mechanics.

Uncertainty relationships have already been introduced at the beginning of this section in two different ways. But in the example of the wave packet studied in the previous paragraph, we have seen that the function is located in a zone of range (half-height width):

$$\delta x=\frac{2\pi}{\delta k} \tag{11.2.484}$$

We therefore have the relation:

$$\delta x\cdot\delta k=2\pi \tag{11.2.485}$$

We fall back here on an expression of the uncertainty. The numerical coefficient could be slightly different according to the definition chosen for δx and δk, or the type of packet. In particular, it could be significantly larger in some cases. We thus have in fact an inequality of the type:

$$\delta x\cdot\delta k\geq 2\pi \tag{11.2.486}$$

In quantum physics, these inequalities are expressed as a function of the linear momentum p, connected to k as we have prove earlier by $p=\hbar k$. We therefore have:

$$\delta x\cdot\delta p_x\geq\hbar \tag{11.2.487}$$

Thus, the more precisely the linear momentum of a wave packet is defined (implicitly the wavelength), the less it has components and the more it is spread, which increases the uncertainty relative to its position and respectively the more its position is known, the less its wavelength will be precise.

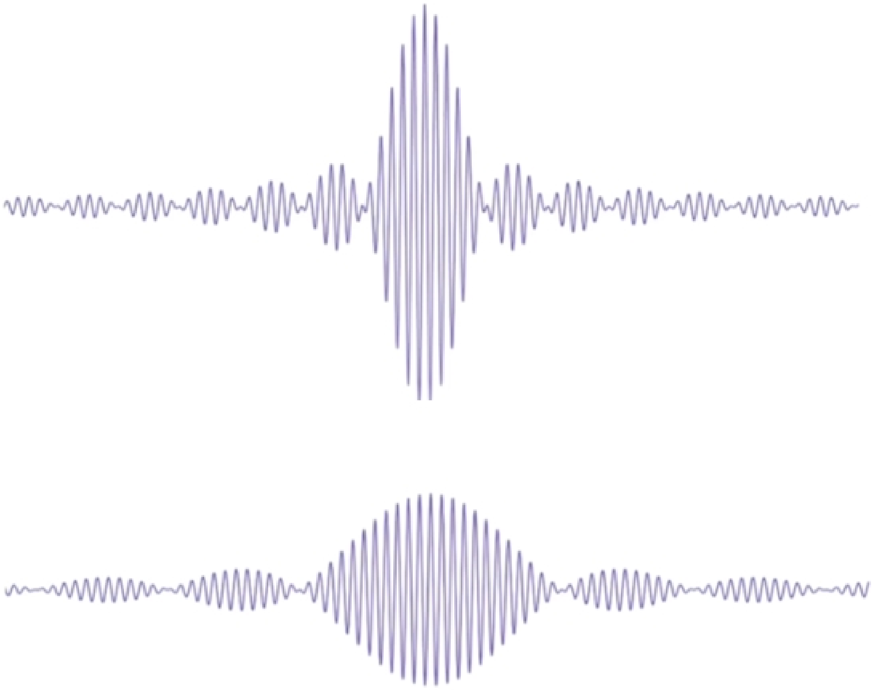


Figure 11.47: Spreading plane wave with decreasing δk from top top bottom

It is therefore not a question of uncertainties in the sense of measurement, and which would be limited by the measuring devices, but of an intrinsic fundamental property, linked to the quantum representation of a particle according to the proposed mathematical model. The model of the Bohr atom have therefore to be rejected for the energy levels which are close to this equality.

11.2.4.4.1.1 Gaussian Wavepacket

Now, a probability is a real number between 0 and 1. An outcome of a measurement which has a probability 0 is an impossible outcome, whereas an outcome which has a probability 1 is a certain outcome. According to what we have seen above, the probability of a measurement of x yielding a result between $-\infty$ and $+\infty$ is:

$$P_{x\in -\infty:+\infty}(t)=\int\limits_{-\infty}^{+\infty}|\Psi(x,t)|^2\mathrm{d}x \tag{11.2.488}$$

However, a measurement of x must yield a value between $-\infty$ and $+\infty$, since the particle has to be located somewhere. It follows that $P_{x\in -\infty:+\infty}=1$, or:

$$\int\limits_{-\infty}^{+\infty}|\Psi(x,t)|^2\mathrm{d}x=1 \tag{11.2.489}$$

which is generally known as the normalization condition for the wave function.

For example, suppose that we wish to normalize the wave function of a Gaussian wave packet, centered on $x=x_0$, and of characteristic width σ i.e.:

$$\Psi(x)=\Psi_0 e^{-(x-x_0)^2/\left(2\sigma^2\right)} \tag{11.2.490}$$

In order to determine the normalization constant Ψ_0, we simply substitute the corresponding relation to get:

$$|\Psi_0|^2\int\limits_{-\infty}^{+\infty}e^{-(x-x_0)^2/\left(2\sigma^2\right)}\mathrm{d}x=1 \tag{11.2.491}$$

Changing the variable of integration to $y=(x-x_0)/(\sqrt{2}\sigma)$, we get:

$$|\Psi_0|^2\sqrt{2}\sigma\int\limits_{-\infty}^{+\infty}e^{-y^2}\mathrm{d}y=1 \tag{11.2.492}$$

However (as proved in the section of Statistics at page 695):

$$\int\limits_{-\infty}^{+\infty}e^{-y^2}\mathrm{d}y=\sqrt{\pi} \tag{11.2.493}$$

which implies that:

$$|\Psi_0|^2 = \frac{1}{(2\pi\sigma^2)^{1/2}} \tag{11.2.494}$$

Hence, a general normalized Gaussian wavefunction takes the form:

$$\Psi(x) = \frac{e^{\mathrm{i}\varphi}}{(2\pi\sigma^2)^{1/4}} e^{-(x-x_0)^2/(2\sigma^2)} := A(x)e^{\mathrm{i}\varphi} \tag{11.2.495}$$

where φ is an arbitrary real phase-angle. The latter gauss wave-packet is often denoted:

$$\boxed{\Psi(x) = \frac{e^{\mathrm{i}k_0 x}}{(2\pi\sigma^2)^{1/4}} e^{-(x-x_0)^2/(2\sigma^2)} := A(x)e^{\mathrm{i}k_0 x}} \tag{11.2.496}$$

and still satisfies:

$$\int\limits_{-\infty}^{+\infty} |\Psi(x,t)|^2 \mathrm{d}x = 1 \tag{11.2.497}$$

It is straightforward that we have (we take $x_0 = 0$):

$$|\Psi|^2 = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \tag{11.2.498}$$

$|\Psi|^2$ is the probability density of the particle position, x, described by a gaussian function with standard deviation σ.

Remark

In many textbooks authors plot the most simplest case of the above relation by taking $\sigma = 1, k_0 = 1$ and the imaginary part, that is:

$$\Im(\Psi(x)) = \frac{\sin(x)}{(2\pi)^{1/4}} e^{-x^2/2}$$

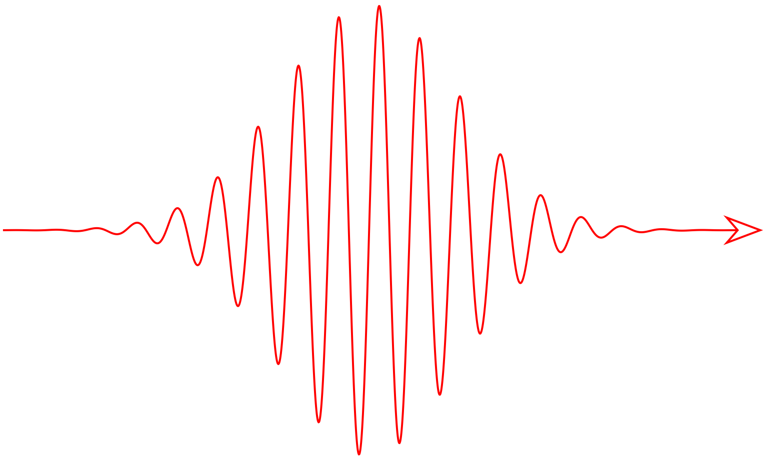


Figure 11.48: **Plot of the Gaussian wavepacket in 2D (projection of the 3D wavepacket on the real or imaginay plane)**

And the corresponding one in 3D (Gaussian-modulated complex exponential):

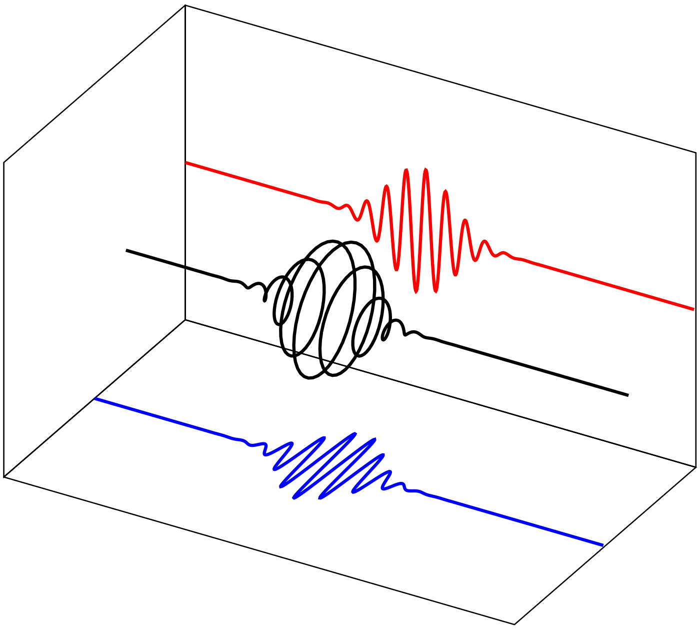


Figure 11.49: **Plot of the Gaussian wavepacket in 3D (author: (author: H. Zeljko)**

Now, it is important to demonstrate that if a wavefunction is initially normalized then it stays normalized as it evolves in time according to Schrödinger's equation. If this is not the case then the probability interpretation of the wavefunction is untenable, since it does not make sense for the probability that a measurement of x yields any possible outcome (which is, manifestly, unity) to change in time. Hence, we require that:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int\limits_{-\infty}^{+\infty}|\Psi(x,t)|^2\mathrm{d}x=0 \tag{11.2.499}$$

for wavefunctions satisfying Schrödinger's equation. The above relation gives:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int\limits_{-\infty}^{+\infty}\bar{\Psi}\Psi\mathrm{d}x=\int\limits_{-\infty}^{+\infty}\left(\frac{\partial\bar{\Psi}}{\partial t}\Psi+\bar{\Psi}\frac{\partial\Psi}{\partial t}\right)\mathrm{d}x=0 \tag{11.2.500}$$

Now, multiplying Schrödinger's equation by $\bar{\psi}/(\mathrm{i}\hbar)$, we get:

$$\bar{\Psi}\frac{\partial\Psi}{\partial t}=\frac{\mathrm{i}\hbar}{2m}\bar{\Psi}\frac{\partial^2\Psi}{\partial x^2}-\frac{\mathrm{i}}{\hbar}U|\Psi|^2 \tag{11.2.501}$$

The complex conjugate of this expression yields:

$$\Psi\frac{\partial\bar{\Psi}}{\partial t}=-\frac{\mathrm{i}\hbar}{2m}\Psi\frac{\partial^2\bar{\Psi}}{\partial x^2}+\frac{\mathrm{i}}{\hbar}U|\Psi|^2 \tag{11.2.502}$$

since $(\overline{AB})=\bar{A}\bar{B}$, $\bar{\bar{A}}=A$, and $\bar{\mathrm{i}}=-\mathrm{i}$. Summing the previous two equations, we get:

$$\frac{\partial\bar{\Psi}}{\partial t}\psi+\bar{\Psi}\frac{\partial\Psi}{\partial t}=\frac{\mathrm{i}\hbar}{2m}\left(\bar{\Psi}\frac{\partial^2\Psi}{\partial x^2}-\Psi\frac{\partial^2\bar{\Psi}}{\partial x^2}\right)=\frac{\mathrm{i}\hbar}{2m}\frac{\partial}{\partial x}\left(\bar{\Psi}\frac{\partial\Psi}{\partial x}-\Psi\frac{\partial\bar{\Psi}}{\partial x}\right) \tag{11.2.503}$$

The relation:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int\limits_{-\infty}^{+\infty}\bar{\Psi}\Psi\mathrm{d}x=\int\limits_{-\infty}^{+\infty}\left(\frac{\partial\bar{\Psi}}{\partial t}\Psi+\bar{\Psi}\frac{\partial\Psi}{\partial t}\right)\mathrm{d}x=0 \tag{11.2.504}$$

and:

$$\frac{\partial\bar{\Psi}}{\partial t}\Psi+\bar{\Psi}\frac{\partial\Psi}{\partial t}=\frac{\mathrm{i}\hbar}{2m}\left(\bar{\Psi}\frac{\partial^2\Psi}{\partial x^2}-\Psi\frac{\partial^2\bar{\Psi}}{\partial x^2}\right)=\frac{\mathrm{i}\hbar}{2m}\frac{\partial}{\partial x}\left(\bar{\Psi}\frac{\partial\Psi}{\partial x}-\Psi\frac{\partial\bar{\Psi}}{\partial x}\right) \tag{11.2.505}$$

can be combined to produce:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int\limits_{-\infty}^{+\infty}|\Psi|^2\mathrm{d}x=\frac{\mathrm{i}\hbar}{2m}\left[\bar{\Psi}\frac{\partial\Psi}{\partial x}-\Psi\frac{\partial\bar{\Psi}}{\partial x}\right]_{-\infty}^{+\infty}=0 \tag{11.2.506}$$

The above equation is satisfied provided:

$$|\Psi|\rightarrow 0 \quad \text{as} \quad |x|\rightarrow\pm\infty$$

However, this is a necessary condition for the integral on the left-hand side of:

$$\int\limits_{-\infty}^{+\infty}|\Psi(x,t)|^2\mathrm{d}x=1 \tag{11.2.507}$$

to converge. Hence, we conclude that all wavefunctions which are square-integrable have the property that if the normalization condition is satisfied at one instant in time then it is satisfied at all subsequent times.

In the following, all wavefunctions are assumed to be square-integrable and normalized, unless otherwise stated.

11.2.4.4.2 Infinite (rectangular) potential well

Let us take as a first example, very important for the Nuclear Physics section and for semiconductor specialists, the resolution in the classical form of the infinite potential well with straight walls, also named "infinite rectangular well" (this example is really very important, take your time to understand it and to master it at best) or "particle in a box problem".

This is the simplest example of a potential function $E_p(x)$, zero within the well and infinitely big on the walls, distant by a length L:

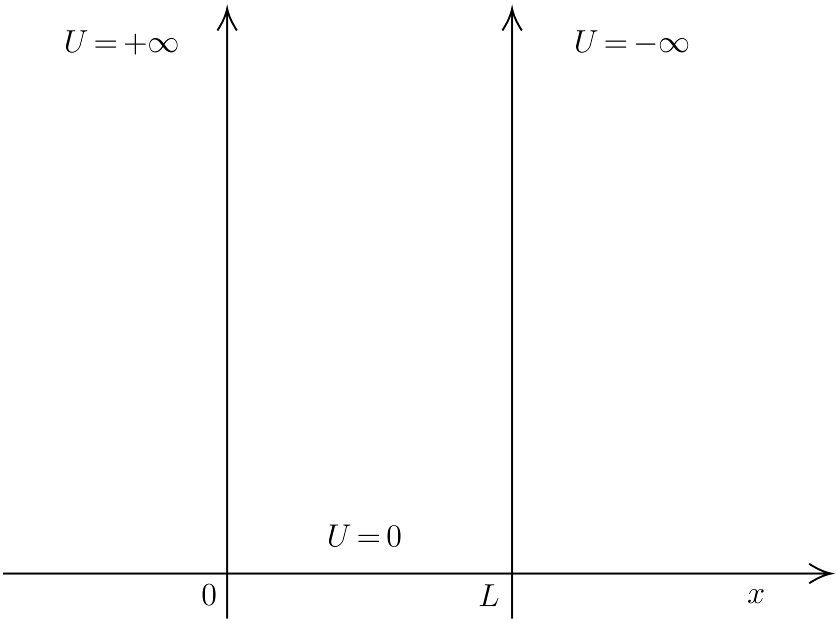


Figure 11.50: **Infinite potential well illustration**

> **Remark**
>
> When $U=+\infty$ we say sometimes that the walls are "perfectly reflective".

We assume a particle trapped in this well. It can not escape because the walls (i.e. the potential U) have an infinite height. But inside, it is free to move without interacting with the walls.

This configuration is reflected by the boundary conditions where the potential electrostatic energy is denoted U:

$$U=0 \quad \text{if} \quad 0\leq x\leq L \tag{11.2.508}$$

and:

$$U=+\infty \quad \text{if} \quad x<0 \quad \text{or} \quad x>L \tag{11.2.509}$$

There exist at least two ways to approach this problem. Let us consider the two types of treatments because the first one allows to have a simplistic approach whereas the second one allows to have a more general approach which will be useful to us later on in our study of the tunnel effect:

11.2.4.4.3 First approach

The (classical) Schrödinger equation:

$$\frac{\partial^2\Psi}{\partial x^2}+\frac{\partial^2\Psi}{\partial y^2}+\frac{\partial^2\Psi}{\partial z^2}+\frac{8\pi^2 m}{h^2}\cdot\left(E_{\text{tot}}-E_p\right)\Psi=0 \quad (11.2.510)$$

has therefore a simple particular solution respecting the initial conditions in one dimension, of the type:

$$\Psi=\Psi_0\sin(kx) \quad (11.2.511)$$

whose second derivative is:

$$\frac{\partial^2\Psi}{\partial x^2}=-k^2\Psi \quad (11.2.512)$$

Introduced in the Schrödinger equation, we get, after some elementary algebra simplifications:

$$k^2=\frac{8\pi^2 m}{h^2}\left(E_{\text{tot}}-E_p\right)=\frac{2m}{\hbar^2}\left(E_{\text{tot}}-E_p\right) \quad (11.2.513)$$

So finally the solution is written:

$$\Psi=\Psi_0\sin\left(\frac{\sqrt{2m\left(E_{\text{tot}}-E_p\right)}}{\hbar}\cdot x\right) \quad (11.2.514)$$

to which the boundary conditions must be applied (the cosine solution is in every aspect similar).

If we want to be able, later, to make a parallel with one or more electron(s) trapped in the potential well of an atom (which is him not rectangular and not infinite), we are taken to the following considerations:

The stability of the atoms suggests the existence of an electronic standing wave in the well. In addition, experimental observation shows that only certain energy levels appear to be allowed in the latter.

If we make a similarity with the vibrating strings, the wave function of the electron must be such that:

1. For $x=0$ and $L=0$ there must be a vibration node, i.e. $\Psi(0)=\Psi(L)=0$
2. The wave function Ψ must have an integer value of half wave length along the length L
3. In the box $U=0$ therefore $E_p=0$
4. If at the ends ($x=0$ and $x=L$) $\Psi=0$ then the argument of the sine equals $n\pi\;\forall n\in\mathbb{N}^*$

Therefore we must have:

$$x=L\Rightarrow\sqrt{\frac{8\pi^2 m}{h^2}\left(E_{\text{tot}}-E_p\right)}\cdot L=n\pi \quad (11.2.515)$$

hence since the potential energy is zero in the well:

$$\boxed{E_n=\frac{\pi^2h^2}{2mL^2\pi^24}n^2=\frac{\pi^2\hbar^2}{2mL^2}n^2} \quad (11.2.516)$$

The total energy of the particle thus ranges over a discrete sequence of values, the only ones permitted. The value of L is determined using the Bohr or Sommerfeld model as depending on the situations (see section Corpuscular Quantum Physics page 4626 and 4648).

The total energy of the above particle is composed of the eigenvalues of the energy in the potential well.

Thus the Schrödinger's equation allows us to ignore Bohr's third postulate in the sense that it directly explains the notion of quantification of the levels by integer (discrete) values of the boundary conditions of a potential well considered as perfect.

The corresponding wave functions in the well where $U=0\equiv E_p=0$ are therefore:

$$\Psi=\Psi_0\sin\left[\sqrt{\frac{8\pi^2m}{h^2}E_{\text{tot}}}\cdot x\right]=\Psi_0\sin\left[\sqrt{\frac{8\pi^2m}{h^2}\frac{h^2}{8mL^2}n^2}\cdot x\right] \quad (11.2.517)$$

Either after simplification:

$$\Psi_n=\Psi_0\sin\left(\frac{\pi n}{L}x\right) \quad (11.2.518)$$

This is the expression of one of the solutions of the differential equation for the ideal rectangular potential well. Thus, there exists a discrete sequence of solutions wave functions. These are the obviously the eigenfunctions of the particle.

The constant Ψ_0 in this expression is determined by the de Broglie normalization condition (about which we spoke at the beginning of this section), that is to say by the condition:

$$\int_0^L|\Psi_n|^2\,\mathrm{d}x=\Psi_0^2\int_0^L\sin^2\left(\frac{n\pi x}{L}\right)\mathrm{d}x=1 \quad (11.2.519)$$

We then find (normally integration calculus that should be obvious to the reader but on request we can as always give the details):

$$\Psi_0=\sqrt{\frac{2}{L}} \quad (11.2.520)$$

and the final expression of the wave function associated with the eigenvalue E_n thus reads:

$$\boxed{\Psi_n(x)=\sqrt{\frac{2}{L}}\sin\left(\frac{n\pi x}{L}\right)} \quad (11.2.521)$$

Some physicists typically write this in a complex form by obviously considering only the real part of the following expression (we use the "Euler formula" seen when introducing the complex numbers in the section Numbers):

$$\Psi_n(x)=\sqrt{\frac{2}{L}}e^{ikx} \quad (11.2.522)$$

with:

$$k_x = \frac{n\pi}{L_x} \quad (11.2.523)$$

We then say that we have "quantification conditions" on k imposed by the boundary conditions.

This notation is sometimes useful and we will use it when studying the tunnel effect in the section of Nuclear Physics.

We can deduce from the expression obtained the main properties of the wave functions describing the stationary states of the particle in a box:

P1. The figure below shows the Ψ_n (left) and the probability density functions $|\Psi_n|^2$ (right) for the first energy levels E_n:

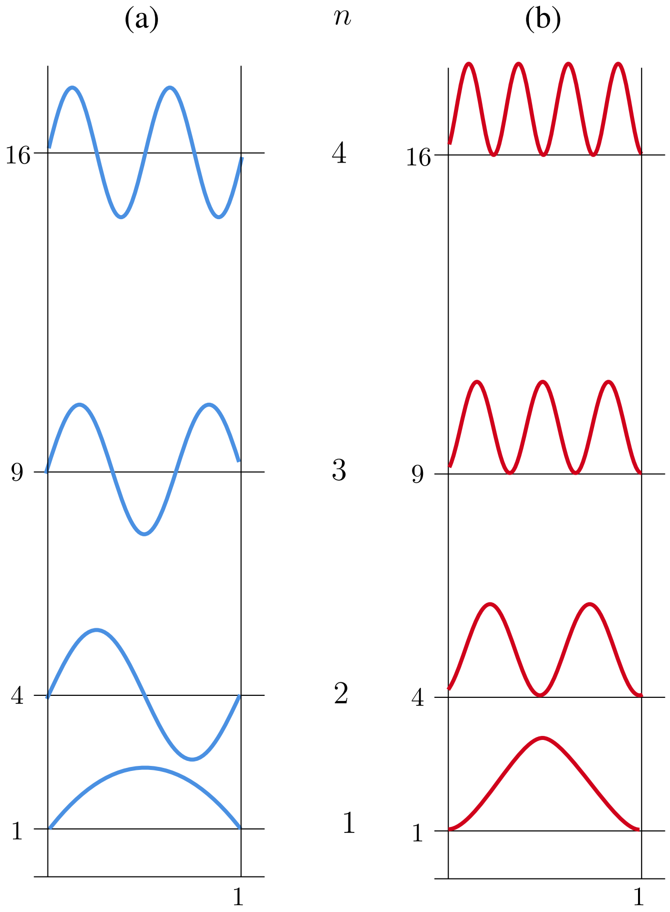


Figure 11.51: Representation of the wave functions and density for some energy levels

We notice that (obviously we could analyse this analytically and not graphically if we wanted to), in addition to the points $x=0$ and $x=L$, Ψ_n has $n-1$ zeros located in:

$$x = x_m = \frac{mL}{n} \quad \text{with} \quad m = 1, 2, \ldots, n-1 \quad (11.2.524)$$

These points, where the wave function and the probability density are zero, are named "nodal points" or simply "nodes" of the wave function. The number of nodes increases when n increases, that is to say when one moves to more and more excited states.

The wave function Ψ_1 of the ground state with $n=1$ and therefore with:

$$E=E_1=\frac{h^2}{8mL^2} \tag{11.2.525}$$

has no node, that of the first excited state Ψ_2 of energy:

$$E=E_2=4E_1 \tag{11.2.526}$$

has one nodal point, that of the second excited state Ψ_3 has two nodal points, etc.

The variation of the nodal properties of the wave functions when n varies translates the orthogonality of the stationary states of different energy. Indeed, we easily check that $\langle\Psi_n|\Psi_m\rangle$ is zero when $m\neq n$:

$$\begin{aligned}\langle\Psi_n\,|\,\Psi_m\rangle&=\int\limits_0^L\Psi_n(x)\Psi_m(x)\mathrm{d}x=\frac{2}{L}\int_0^L\sin\left(\frac{n\pi x}{L}\right)\sin\left(\frac{m\pi x}{L}\right)\mathrm{d}x\\&=\frac{1}{L}\int\limits_0^L\left\{\cos\left[\frac{(n-m)\pi x}{L}\right]-\cos\left[\frac{(n+m)\pi x}{L}\right]\right\}\mathrm{d}x\\&=\frac{1}{L}\left\{\frac{L}{(n-m)\pi}\sin\left[\frac{(n-m)\pi x}{L}\right]+\frac{L}{(n+m)\pi}\sin\left[\frac{(n-m)\pi x}{L}\right]\right\}\Bigg|_0^L=0\end{aligned} \tag{11.2.527}$$

where we used one of the remarkable trigonometric identity demonstrated in the section of Trigonometry.

P2. As we can see it in the previous figure, the probability density associated with any stationary state of the particle is symmetric with respect to the midpoint $x=L/2$.

We therefore expect that the mean value of x will be exactly equal to $L/2$ in such a state. Indeed, we have seen in the section of Statistics that the expected mean (average) of a probability event $P(x)$ is defined by:

$$\mathrm{E}(x)=\langle x\rangle=\int\limits_{-\infty}^{+\infty}xP(x)\mathrm{d}x \tag{11.2.528}$$

where x, $\mathrm{E}(x)$ and $P(x)$ have no units (caution! as we will do a dimensional analysis!).

Now, in quantum physics $\mathrm{E}(x)$ and x are identical dimensional quantities. This means that the dimensions of $P(x)$ must cancel those of $\mathrm{d}x$. Thus, we guess following the study of the de Broglie normalization conditions that:

$$P(x)\equiv|\,\Psi_n(x)\,|^2 \tag{11.2.529}$$

is a linear probability of the presence of the particle in $[\mathrm{m}^{-1}]$.

The integration domain being $[0,L]$ we have finally:

$$\begin{aligned}
\langle x\rangle &= \int_0^L x\,|\,\Psi_n(x)\,|^2\,\mathrm{d}x=\Psi_0^2\int_0^L x\sin^2\left(\frac{n\pi x}{L}\right)\mathrm{d}x=\frac{2}{L}\int_0^L x\sin^2\left(\frac{n\pi x}{L}\right)\mathrm{d}x\\
&=\frac{1}{L}\int_0^L\left\{1-\cos\left(\frac{2n\pi x}{L}\right)\right\}x\mathrm{d}x=\frac{1}{L}\left\{\int_0^L x\mathrm{d}x-\int_0^L\cos\left(\frac{2n\pi x}{L}\right)x\mathrm{d}x\right\}\\
&=\frac{1}{L}\left\{\frac{1}{2}x^2\Big|_0^L-\left(\frac{L}{2n\pi}\right)\left[\sin\left(\frac{2n\pi x}{L}\right)x\Big|_0^L-\int_0^L\sin\left(\frac{2n\pi x}{L}\right)\mathrm{d}x\right]\right\}\\
&=\frac{1}{L}\left\{\frac{L^2}{2}-\left(\frac{L}{2n\pi}\right)\left[0+\frac{L}{2n\pi}\cos\left(\frac{2n\pi x}{L}\right)\Big|_0^L\right]\right\}\\
&=\frac{1}{L}\left\{\frac{L^2}{2}-\left(\frac{L}{2n\pi}\right)\left[0+\frac{L}{2n\pi}(1-1)\right]\right\}=\frac{1}{L}\frac{L^2}{2}=\frac{L}{2}
\end{aligned}\tag{11.2.530}$$

P3. Also we have for the linear momentum:

$$\begin{aligned}
\langle p_x\rangle &=\int\limits_0^L\bar{\Psi}_n p\Psi_n\mathrm{d}x=\int\limits_0^L\bar{\Psi}_n\left(-\mathrm{i}\frac{\mathrm{d}}{\mathrm{d}x}\right)\Psi_n\mathrm{d}x\\
&=-\mathrm{i}\hbar\int\limits_0^L\sqrt{\frac{2}{L}}\sin\left(\frac{n\pi x}{L}\right)\frac{\mathrm{d}}{\mathrm{d}x}\left(\sqrt{\frac{2}{L}}\sin\left(\frac{n\pi x}{L}\right)\right)\mathrm{d}x\\
&=-\mathrm{i}\frac{2\hbar}{L}\int\limits_0^L\sin\left(\frac{n\pi x}{L}\right)\left(\frac{n\pi}{L}\cos\left(\frac{n\pi x}{L}\right)\right)\mathrm{d}x\\
&=-\mathrm{i}\frac{2n\pi\hbar}{L^2}\int\limits_0^L\sin\left(\frac{n\pi x}{L}\right)\cos\left(\frac{n\pi x}{L}\right)\mathrm{d}x=-\mathrm{i}\frac{2n\pi\hbar}{L^2}\int\limits_0^L\frac{1}{2}\sin\left(\frac{2n\pi x}{L}\right)\mathrm{d}x\\
&=-\mathrm{i}\frac{n\pi\hbar}{L^2}\frac{L}{2n\pi}\int\limits_0^{2\pi n}\sin(\phi)\mathrm{d}\phi=-\mathrm{i}\frac{\hbar}{2L}\cdot 0=0
\end{aligned}\tag{11.2.531}$$

P4. We can also verify without too much difficulty that what we saw in the statement of the second postulate of Quantum Physics is well verified in this example. That is to say, the eigenfunctions of the wave are connected to the Hamiltonian operator via the eigenvalues of the energy:

$$H\Psi=E_{\text{tot}}\Psi\tag{11.2.532}$$

Indeed, in our example, this gives:

$$\begin{aligned}&\left(\underbrace{E_p}_{=0}-\frac{\hbar^2}{2m}\Delta\right)\Psi=E_{\text{tot}}\Psi\\&\Rightarrow -\frac{\hbar^2}{2m}\Delta\Psi=\frac{\pi^2\hbar^2n^2}{2mL^2}\Psi\\&\Rightarrow -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}\left(\sqrt{\frac{2}{L}}\sin\left(\frac{n\pi x}{L}\right)\right)=\frac{\pi^2\hbar^2n^2}{2mL^2}\left(\sqrt{\frac{2}{L}}\sin\left(\frac{n\pi x}{L}\right)\right)\\&\Rightarrow -\frac{\partial^2}{\partial x^2}\left(\sin\left(\frac{n\pi x}{L}\right)\right)=\frac{\pi^2n^2}{L^2}\left(\sin\left(\frac{n\pi x}{L}\right)\right)\\&\Rightarrow -\frac{\pi^2n^2}{L^2}\left(-\sin\left(\frac{n\pi x}{L}\right)\right)=\frac{\pi^2n^2}{L^2}\left(\sin\left(\frac{n\pi x}{L}\right)\right)\\&\Rightarrow 1=1\end{aligned} \qquad (11.2.533)$$

P5. Our analysis of the quantum particle in a box would not be complete without discussing "Bohr's correspondence principle". This principle states that for large quantum numbers, the laws of quantum physics must give identical results as the laws of classical physics. To illustrate how this principle works for a quantum particle in a box, we see that in the plot given earlier above of $|\Psi|^2$ when $n\rightarrow+\infty$ that the a number of "peaks" will increase to infinite and therefore result in a uniform (constant) probability to find the particle between 0 and L.

That's it... for the first approach of the problem. Let us now look at the second:

11.2.4.4.4 Second approach

We thus have the Schrödinger equation in the one-dimensional case:

$$\frac{\partial^2\Psi}{\partial x^2}+\frac{8\pi^2m}{h^2}\left(E_{\text{tot}}-E_p\right)\Psi=0 \qquad (11.2.534)$$

In regions outside the box where the potential is infinite, we have:

$$\frac{\partial^2\Psi}{\partial x^2}+\frac{8\pi^2m}{h^2}\left(E_{\text{tot}}-\infty\right)\Psi=0 \qquad (11.2.535)$$

Therefore:

$$\frac{\partial^2\Psi}{\partial x^2}=\infty\Psi \qquad (11.2.536)$$

which gives:

$$\Psi=\frac{1}{\infty}\frac{\partial^2\Psi}{\partial x^2}=0 \qquad (11.2.537)$$

Let us consider now the case of the well where, since the electrostatic potential is zero, the Schrödinger equation reduces to:

$$\frac{\partial^2\Psi}{\partial x^2}+\frac{8\pi^2 m}{h^2}E_{\text{tot}}\Psi=0 \qquad (11.2.538)$$

It is therefore a linear differential equation of order 2 with constant coefficients, equation which is relatively easy to solve in the general case (see section Differential and Integral Calculus page 1610). Given the equation:

$$y''+py'+qy=0 \qquad (11.2.539)$$

Using the results we get during the treatment of the particular solution, let us suppose that the function y that satisfy this differential equation is of the form $y=e^{Kx}$. We have then:

$$K^2e^{Kx}+pKe^{Kx}+qe^{Kx}=0 \qquad \text{or} \qquad K^2+pK+q=0 \qquad (11.2.540)$$

provided, of course, that $e^{Kx}\neq 0$. This last relation is therefore the auxiliary quadratic equation of the differential equation (characteristic polynomial). It has two solutions / roots (it is a simple resolution of a polynomial of the second degree) that we will denote in the general case K_1, K_2. Which means that:

$$K_1^2e^{K_1x}+pK_1e^{K_1x}+qe^{K_1x}=0 \qquad \text{or} \qquad K_2^2e^{K_2x}+pK_2e^{K_2x}+qe^{K_2x}=0 \qquad (11.2.541)$$

is satisfied for both roots. If we make the sum since the both are equal to the same constant:

$$\left(K_1^2e^{K_1x}+K_2^2e^{K_2x}\right)+p\left(K_1e^{K_1x}+K_2e^{K_2x}\right)+q\left(e^{K_1x}+e^{K_2x}\right)=0 \qquad (11.2.542)$$

Thus, it is immediate that the general solution of y is of the type:

$$y=Ae^{K_1x}+Be^{K_2x} \qquad (11.2.543)$$

where the reader should normally be able to verify that the addition of the constants A and B does not change the developments of the preceding paragraphs.

In the present case that interest us:

$$\Psi''+\frac{8\pi^2 m}{h^2}E_{\text{tot}}=0 \qquad (11.2.544)$$

The quadratic equation is:

$$k^2+\frac{8\pi^2 m}{h^2}E_{\text{tot}}=0\Rightarrow k^2=-2mE_{\text{tot}}\hbar^{-2} \qquad (11.2.545)$$

thus:

$$k=\pm\frac{\sqrt{-2mE_{\text{tot}}}}{\hbar}=\pm\mathrm{i}\frac{\sqrt{2mE_{\text{tot}}}}{\hbar} \qquad (11.2.546)$$

So finally the general solution is of the form:

$$\Psi=A\exp\left(\mathrm{i}\frac{\sqrt{2mE_{\text{tot}}}}{\hbar}x\right)+B\exp\left(-\mathrm{i}\frac{\sqrt{2mE_{\text{tot}}}}{\hbar}x\right) \qquad (11.2.547)$$

Let us put now:

$$\theta = \frac{\sqrt{2mE_{\text{tot}}}}{\hbar}x \tag{11.2.548}$$

We then have:

$$\begin{aligned}\Psi &= Ae^{\mathrm{i}\theta} + Be^{-\mathrm{i}\theta} = A\cos(\theta) + \mathrm{i}A\sin(\theta) + B\cos(\theta) - \mathrm{i}B\sin(\theta)\\ &= A'\cos(\theta) + B'\sin(\theta)\end{aligned} \tag{11.2.549}$$

with:

$$A' = A + B \qquad \text{and} \qquad B' = \mathrm{i}(A - B) \tag{11.2.550}$$

It is now necessary to determine A' and B' using the boundary conditions. Thus, at $x = 0$ and $x = L$ we should have $\Psi = 0$ and we have for $x = 0$:

$$\Psi = A'\cos\left(\frac{\sqrt{2mE_{\text{tot}}}}{\hbar}x\right) + B'\sin\left(\frac{\sqrt{2mE_{\text{tot}}}}{\hbar}x\right) = A' \tag{11.2.551}$$

The coefficient A' must therefore be equal to zero. And in $x = L$ we should have:

$$\begin{aligned}\Psi &= A'\cos\left(\frac{\sqrt{2mE_{\text{tot}}}}{\hbar}x\right) + B'\sin\left(\frac{\sqrt{2mE_{\text{tot}}}}{\hbar}x\right)\\ &= 0\cdot\cos\left(\frac{\sqrt{2mE_{\text{tot}}}}{\hbar}x\right) + B'\sin\left(\frac{\sqrt{2mE_{\text{tot}}}}{\hbar}L\right) = B'\sin\left(\frac{\sqrt{2mE_{\text{tot}}}}{\hbar}L\right) = 0\end{aligned} \tag{11.2.552}$$

But in this case, B' must be different from zero. Indeed, if it were zero, the wave function would be zero throughout the whole well, which is contrary to the physical reality of the problem. It is therefore necessary that the sinus to be zero, or that its argument be equal to a multiple of a non-zero integer number of angle π such that:

$$\frac{\sqrt{2mE_{\text{tot}}}}{\hbar}L = \pm n\pi \tag{11.2.553}$$

Therefore:

$$E_{\text{tot}} = \frac{\pi^2\hbar^2}{2mL^2}n^2 \tag{11.2.554}$$

We thus find exactly the same result as the preceding method!

It remains to determine B and the method is exactly identical to the first method of resolution which we have seen above. Thus, we have:

$$\Psi = \sqrt{\frac{2}{L}}\sin\left(\frac{n\pi x}{L}\right) \tag{11.2.555}$$

What is especially important in this method is to remember for a moment the general form of the solution:

$$\Psi = Ae^{\mathrm{i}kx} + Be^{-\mathrm{i}kx} \tag{11.2.556}$$

11.2.4.4.5 Fermi Energy

The "Fermi energy", E_F is a concept in Quantum Physics that designates the energy of the highest quantum state occupied in an idealized system in which all the layers are filled successively and without discontinuities, that is to say in practice when a system is at absolute zero: 0 [K] (see section Statistical Mechanics page 3571). It is an important concept in Solid states physics as we have seen it in our study of semiconductors in the section of Electrokinetics.

The rectilinear infinite potential well is an excellent teaching tool to introduce practically the Fermi energy level and then to extend it to other particular cases.

Let us recall that we have just obtained in two different ways:

$$E_n = \frac{\pi^2\hbar^2}{2mL^2\pi^2 4}n^2 = \frac{\pi^2\hbar^2}{2mL^2}n^2 \tag{11.2.557}$$

Then the total energy of an entire system composed of N particles can all have the same fundamental state $n=1$ and thus that violate the principle of Pauli will be:

$$E_{\text{tot}} = \left(\frac{\pi^2\hbar^2}{2mL^2}\right)N \tag{11.2.558}$$

But if we apply the Pauli's exclusion principle of electrons (fermions), each level can then take only two states (spin oppositions) in this model (which does not contain sublayers or other subtleties). Henceforth, each fundamental level can be occupied by only two states, and so it is for each level. The total energy is then in reality:

$$E_{\text{tot}} = 2\sum_{n=1}^{N/2} E_n = 2\sum_{n=1}^{N/2}\frac{\pi^2\hbar^2}{2mL^2}n^2 = \frac{\pi^2\hbar^2}{mL^2}\sum_{n=1}^{N/2}n^2 \tag{11.2.559}$$

Using the sum of the squares demonstrated in the section of Sequences and Series and writing as it is usual:

$$n_{\text{max}} = \frac{N}{2} \tag{11.2.560}$$

we have then:

$$E_{\text{tot}} = \frac{\pi^2\hbar^2}{mL^2}\sum_{n=1}^{N/2}n^2 = \frac{\pi^2\hbar^2}{mL^2}\left(\frac{n_{max}(n_{\text{max}}+1)(2n_{\text{max}}+1)}{6}\right) \cong \frac{\pi^2\hbar^2}{mL^2}\frac{n_{\text{max}}^3}{3} = \frac{\pi^2\hbar^2}{24mL^2}N^3 \tag{11.2.561}$$

The average energy per particle is then:

$$\bar{E} = \frac{E_{\text{tot}}}{N} \cong \frac{\pi^2\hbar^2}{24mL^2}N^2 \tag{11.2.562}$$

By definition, the energy of the last level to be occupied is that of Fermi and therefore given by:

$$E_n = \frac{E_{\text{tot}}}{N} \cong \frac{\pi^2\hbar^2}{2mL^2}N^2 \Rightarrow E_{N/2} = \frac{\pi^2\hbar^2}{8mL^2}\left(\frac{N}{2}\right)^2 = \frac{\pi^2\hbar^2}{8mL^2}N^2 \tag{11.2.563}$$

Therefore:

$$\boxed{E_F=\frac{\pi^2\hbar^2}{8mL^2}N^2} \tag{11.2.564}$$

We also notice that we have whatever the value of N:

$$\bar{E}=\frac{1}{3}E_F \tag{11.2.565}$$

The three-dimensional isotropic case is known as the "Fermi sphere" as already introduced in the section of Electrokinetics during our study of semiconductors.

Let us now consider a three-dimensional cubical box that has a side length L. This turns out to be a very good approximation for describing electrons in a metal. The states are now labelled by three quantum numbers n_x, n_y, and n_z. The single particle energies are (where m is the mass of fermion (electron in this case)) as we have seen in the section of Electrokinetics during our study of semiconductor:

$$E_{n_x,n_y,n_z}=\frac{\pi^2\hbar^2}{2mL^2}\left(n_x^2+n_y^2+n_z^2\right) \tag{11.2.566}$$

where n_x, n_y, and n_z are positive integers. There are multiple states with the same energy, for example $E_{211}=E_{121}=E_{112}$. Now let's put N non-interacting fermions of spin $1/2$ into this box. To calculate the Fermi energy, we look at the case where N is large.

If we introduce a vector $\vec{n}=\{n_x,n_y,n_z\}$ then each quantum state corresponds to a point in "n-space" with energy:

$$E_{\vec{n}}=\frac{\pi^2\hbar^2}{2mL^2}\|\vec{n}\|^2 \tag{11.2.567}$$

with $\|\vec{n}\|^2$ denoting the square of the usual Euclidean length (norm). The number of states with energy less than $E_{\vec{n}}$, defined as the "Fermi energy", is equal to the number of states that lie within a sphere of radius $R=\|\vec{n}_F\|^2$ in the region of n-space where n_x, n_y and n_z are positive. In the ground state this number equals the number of fermions in the system:

$$N=2\cdot\frac{1}{8}\cdot\frac{4}{3}\pi R^3=2\cdot\frac{1}{8}\cdot\frac{4}{3}\pi n_F^3 \tag{11.2.568}$$

the factor 2 is once again because there are two spin states, the factor of $1/8$ is because only $1/8$ of the sphere lies in the region where all n are positive. Therefore we get:

$$n_F=\left(\frac{3N}{\pi}\right)^{1/3} \tag{11.2.569}$$

so the Fermi energy is given by:

$$E_F=\frac{\hbar^2\pi^2}{2mL^2}n_F^2=\frac{\hbar^2\pi^2}{2mL^2}\left(\frac{3N}{\pi}\right)^{2/3} \tag{11.2.570}$$

Which results in a relationship between the Fermi energy and the number of particles per volume (when we replace L^2 with $V^{2/3}$):

$$\boxed{E_F=\frac{\hbar^2}{2m}\left(\frac{3\pi^2N}{V}\right)^{2/3}} \tag{11.2.571}$$

Confusingly, the term "Fermi energy" is often used to refer to a different but closely related concept, the "Fermi level" (also named "electrochemical potential"). There are a few key differences between the Fermi level and Fermi energy, at least as they are used in this textbook:

- The Fermi energy is only defined at absolute zero, while the Fermi level is defined for any temperature.
- The Fermi energy is an energy difference (usually corresponding to a kinetic energy), whereas the Fermi level is a total energy level including kinetic energy and potential energy.
- The Fermi energy can only be defined for non-interacting fermions (where the potential energy or band edge is a static, well defined quantity), whereas the Fermi level (the electrochemical potential of an electron) remains well defined even in complex interacting systems, at thermodynamic equilibrium.

Since the Fermi level in a metal at absolute zero is the energy of the highest occupied single particle state, then the Fermi energy in a metal is the energy difference between the Fermi level and lowest occupied single-particle state, at zero-temperature.

11.2.4.4.6 1D-Harmonic oscillator

The study of the harmonic oscillator corresponding to that of a wave function stuck in a well of parabolic potential. This is roughly equivalent to the atoms where the walls of the potential well are naturally not rectangular and infinite ... The study that follows is therefore what is closest to what is available in Nature at the atomic level and this is why it is used a lot in Quantum Chemistry and also in Nuclear Physics.

In the case of a free particle in rectilinear displacement, we have seen that the potential energy is zero ($E_p = 0$) and the Schrödinger equation then becomes:

$$\frac{\partial^2 \Psi}{\partial x^2} + \frac{2mE_{\text{tot}}}{\hbar^2}\Psi = 0 \qquad (11.2.572)$$

However, for a free particle (in the absence of a potential field) the total energy is therefore equal to the kinetic energy:

$$E_{\text{tot}} = \frac{1}{2}mv^2 = \frac{p^2}{2m} \qquad (11.2.573)$$

But we have:

$$E_{\text{tot}} = h\nu = h\frac{\omega}{2\pi} = \hbar\omega = mc^2 \Leftrightarrow p = mc = \frac{E_{\text{tot}}}{c}\frac{h\nu}{c} = h\lambda \qquad (11.2.574)$$

The ratio:

$$\lambda = \frac{h}{p} \qquad (11.2.575)$$

being the associated de Broglie wavelength. By introducing the wave number $k=2\pi/\lambda$ (see section Wave Mechanics page 3480), we have:

$$\boxed{p=\hbar k} \tag{11.2.576}$$

named "de Broglie's relation". Finally:

$$E_{\text{tot}}=\frac{p^2}{2m}=\frac{(\hbar k)^2}{2m} \tag{11.2.577}$$

Therefore, Schrödinger's equation can be written:

$$\frac{\partial^2\Psi}{\partial x^2}+k^2\Psi=0 \tag{11.2.578}$$

We see by direct substitution that this differential equation admits for solutions the wave functions:

$$\Psi_1(x)=Ae^{\text{i}kx} \qquad \text{and} \qquad \Psi_2(x)=Be^{-\text{i}kx} \tag{11.2.579}$$

These two different solutions represent the displacement of the same particle once in the $+x$ direction and the other in $-x$. If $A=B=1$ we have:

$$|\Psi(x)|^2=\bar{\Psi}(x)\cdot\Psi(x)=e^{-\text{i}kx}\cdot e^{\text{i}kx}=1 \tag{11.2.580}$$

The fact that this result is equal to unity means that the probability of finding the particle is the same at all points as we don't need any special normalization. In other words, $\Psi(x)=e^{\pm\text{i}kx}$ describes a situation in which the uncertainty on the position is total. This result is in agreement with the principle of uncertainty since $\Psi(x)=e^{\pm\text{i}kx}$ describes a particle of which we know with precision the linear momentum $p=\hbar k$: that is to say that $\Delta p=0$, which implies $\Delta x\rightarrow+\infty$.

In the section of Differential and Integral Calculus we showed that the most general solution of a differential equation is the sum of these solutions. In other words, in our example:

$$\Psi(x)=Ae^{\text{i}kx}+Be^{-\text{i}kx} \tag{11.2.581}$$

with:

$$k=\frac{\sqrt{2mE_{\text{tot}}}}{\hbar} \tag{11.2.582}$$

By the way, we can notice that if $E_p\neq 0$ then the result is the same with the difference that we will have:

$$k=\frac{\sqrt{2m(E_{\text{tot}}-E_p)}}{\hbar} \tag{11.2.583}$$

When the particle of interest is in a well of potential described by the function (parabola):

$$E_p(x)=\frac{1}{2}c^{te}x^2 \tag{11.2.584}$$

we then speak of "1D-harmonic oscillator".

This system is very important because the Hamiltonian of the equation intervenes in all the problems involving oscillations such as molecular and crystalline vibrations (see section Quantum Chemistry page 6211).

Let us first take as an example the classic harmonic oscillator which consists of a body subjected to move along an axis and subjected to a return force proportional to the distance to a point situated on this axis.

The equation of this body is governed by the equation of dynamics:

$$m\frac{\mathrm{d}^2x}{\mathrm{d}t^2}+kx=0 \tag{11.2.585}$$

We have seen in the section of Classical Mechanics that the general solution of this equation is:

$$x=x_0\cos(\omega t+\varphi) \tag{11.2.586}$$

with for pulsation:

$$\omega=\sqrt{\frac{k}{m}} \tag{11.2.587}$$

The total energy of the system being the classical Hamiltonian, we write:

$$H=E_c+E_p=\frac{1}{2}m\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2+\frac{1}{2}kx(t)^2=\frac{p^2}{2m}+\frac{1}{2}kx(t)^2 \tag{11.2.588}$$

It follows from the expression of the potential energy that the constant we had in the initial relation which defined the well of parabolic potential:

$$E_p(x)=\frac{1}{2}c^{te}x^2 \tag{11.2.589}$$

is simply k and therefore:

$$E_p(x)=\frac{1}{2}\omega^2mx^2 \tag{11.2.590}$$

but which we will write for the next development and by tradition in the following form:

$$E_p(x)=\frac{1}{2}\omega_0^2mx^2 \tag{11.2.591}$$

Now let us come back to our quantum framework. From this point of view we have for Hamiltonian (or total energy):

$$H=E_c+E_p=\frac{p^2}{2m}+\frac{c^{te}}{2}x^2 \tag{11.2.592}$$

where for recall, following what we have just seen:

$$c^{te}=\omega_0^2m \tag{11.2.593}$$

Using what we define as a "reduced writing", we write:

$$H=\frac{\hbar\omega_0}{2}(P^2+Q^2) \tag{11.2.594}$$

where the "reduced operators" of linear momentum and positions are therefore respectively:

$$P=\frac{p}{\sqrt{m\hbar\omega_0}} \qquad \text{and} \qquad Q=\sqrt{\frac{c^{te}}{\hbar\omega_0}}x \tag{11.2.595}$$

where by a stupid tradition we keep the constant instead of replacing by $\omega_0^2 m$.

It is more or less easy to obtain the relation of commutation:

$$[Q,P]=\sqrt{\frac{c^{te}}{m}}\frac{1}{\hbar\omega_0}[x,p] \tag{11.2.596}$$

Proof 11.4.3. Remember the relations below that we saw in our study of functional linear operators at the beginning of this section:

$$p_x=\frac{1}{\mathrm{i}}\hbar\frac{\partial}{\partial x},p_y=\frac{1}{\mathrm{i}}\hbar\frac{\partial}{\partial y},p_z=\frac{1}{\mathrm{i}}\hbar\frac{\partial}{\partial z} \tag{11.2.597}$$

Let us consider the properties of the commutators with the linear momentum. We have also proved above earlier above the relation:

$$\frac{\mathrm{d}}{\mathrm{d}x}x=1+x\frac{\mathrm{d}}{\mathrm{d}x} \tag{11.2.598}$$

By multiplying the latter by $\hbar/\mathrm{i}$, it comes:

$$\frac{\partial}{\partial x}\frac{\hbar}{\mathrm{i}}x=x\frac{\partial}{\partial x}\frac{\hbar}{\mathrm{i}}+1\frac{\hbar}{\mathrm{i}} \tag{11.2.599}$$

That we can also write:

$$p_x x=xp_x-\hbar\mathrm{i} \tag{11.2.600}$$

If you remember the definition of the of commutators ($[\alpha,\beta]=\alpha\beta-\beta\alpha$), we have:

$$[x,p_x]=p_x x-xp_x=\hbar\mathrm{i} \tag{11.2.601}$$

We therefore have for our oscillator:

$$P=\frac{p}{\sqrt{m\hbar\omega_0}}=\frac{p}{\sqrt{m\hbar\omega_0}}\frac{\partial}{\partial x}\mathrm{i}\hbar \qquad \text{and} \qquad Q=\sqrt{\frac{c^{te}}{\hbar\omega_0}}x \tag{11.2.602}$$

Let's write the definition of the commutator:

$$[Q,P]=QP-PQ \tag{11.2.603}$$

Therefore:

$$\begin{aligned}[Q,P]&=\sqrt{\frac{c^{te}}{\hbar\omega_0}}x\frac{p}{\sqrt{m\hbar\omega_0}}\frac{\partial}{\partial x}\mathrm{i}\hbar-\frac{p}{\sqrt{m\hbar\omega_0}}\frac{\partial}{\partial x}\mathrm{i}\hbar\sqrt{\frac{c^{te}}{\hbar\omega_0}}x\\&=\sqrt{\frac{c^{te}}{\hbar\omega_0}}\frac{1}{\sqrt{m\hbar\omega_0}}\left(x\frac{\partial}{\partial x}\mathrm{i}\hbar-\frac{\partial}{\partial x}\mathrm{i}\hbar x\right)=\sqrt{\frac{c^{te}}{m}}\frac{1}{\hbar\omega_0}\left(x\frac{\partial}{\partial x}\mathrm{i}\hbar-\frac{\partial}{\partial x}\mathrm{i}\hbar x\right)\\&=\sqrt{\frac{c^{te}}{m}}\frac{1}{\hbar\omega_0}\left(xp_x-p_x x\right)=\sqrt{\frac{c^{te}}{m}}\frac{1}{\hbar\omega_0}[x,p_x]=\omega_0\frac{1}{\hbar\omega_0}[x,p_x]=\frac{1}{\hbar}[x,p_x]=\mathrm{i}\end{aligned} \tag{11.2.604}$$

that's was what has to be proved...

■ Q.E.D.

We now have an interest for solving the differential equation to use the non-Hermitian operators equation defined by[12] (this is a definition so do not try to understand!):

$$\begin{gathered}\sqrt{2}a=Q+\mathrm{i}P\\ \sqrt{2}a^{\dagger}=Q-\mathrm{i}P\end{gathered} \tag{11.2.605}$$

What thus defines us the operators (by temporarily putting $m=1$):

$$a=\dfrac{Q+\mathrm{i}P}{\sqrt{2}}=\dfrac{\omega x+\mathrm{i}p}{\sqrt{2\hbar\omega_0}}\qquad\qquad a^{\dagger}=\dfrac{Q-\mathrm{i}P}{\sqrt{2}}=\dfrac{\omega x-\mathrm{i}p}{\sqrt{2\hbar\omega_0}} \tag{11.2.606}$$

which are linear combinations of the position and linear momentum operators and are trivially not Hermitian (not equal to its own conjugate for recall!).

We see these two operators very frequently in Quantum Physics, especially in Quantum Field physics, and physicists then speak of "creation operator $a^{\dagger}$" or "raising operator a_+" and of the "destruction operator a" or "lowering operator a_-" (also named "annihilation operator"). The reasons of these names will be more clear later (it seems that it is Dirac himself that introduce them for the first time in the framework of the harmonic oscillator)!

Given the commutators relation, we check (we need that for the next developments):

$$2aa^{\dagger}=P^2+Q^2+1\qquad\text{and}\qquad 2a^{\dagger}a=P^2+Q^2-1 \tag{11.2.607}$$

Proof 11.4.4. First:

$$\begin{aligned}2aa^{\dagger}&=(Q+\mathrm{i}P)(Q-\mathrm{i}P)=Q^2+\mathrm{i}PQ-\mathrm{i}QP+P^2\\&=Q^2-\mathrm{i}[QP-PQ]+P^2=Q^2-\mathrm{i}[Q,P]+P^2\\&=P^2+Q^2-\mathrm{i}\cdot\mathrm{i}=P^2+Q^2+1\end{aligned} \tag{11.2.608}$$

and:

$$\begin{aligned}2a^{\dagger}a&=(Q-\mathrm{i}P)(Q+\mathrm{i}P)=Q^2-\mathrm{i}PQ+\mathrm{i}QP+P^2\\&=Q^2+\mathrm{i}[QP-PQ]+P^2=Q^2+\mathrm{i}[Q,P]+P^2\\&=P^2+Q^2+\mathrm{i}\cdot\mathrm{i}=P^2+Q^2+1\end{aligned} \tag{11.2.609}$$

■ Q.E.D.

And on the other hand we also need:

$$[a,a^{\dagger}]_{\_}=aa^{\dagger}-a^{\dagger}a=1 \tag{11.2.610}$$

Proof 11.4.5.

$$2aa^{\dagger}-2a^{\dagger}a=(P^2+Q^2)-(P^2+Q^2)+1-(-1)=2 \tag{11.2.611}$$

and therefore by dividing by 2 on both sides of the equality, we get:

$$aa^{\dagger}-a^{\dagger}a=1 \tag{11.2.612}$$

---

[12] This approach of solving this problem is sometimes named "Ladder operators for the quantum harmonic oscillator"

■ Q.E.D.

Now let us come back to:

$$H=\frac{\hbar\omega_0}{2}(P^2+Q^2) \quad (11.2.613)$$

and using:

$$a^\dagger a=\frac{P^2+Q^2-1}{2}\Rightarrow\frac{P^2+Q^2}{2}=a^\dagger a+\frac{1}{2}=N+\frac{1}{2} \quad (11.2.614)$$

where it appears that N can be seen as an operator, named "counter operator" (hermitic) since:

$$N=a^\dagger a \quad (11.2.615)$$

The same reasoning and assumption bring us to:

$$aa^\dagger-\frac{1}{2}=N'-\frac{1}{2} \quad (11.2.616)$$

with:

$$N'=aa^\dagger \quad (11.2.617)$$

It is then sufficient to quantify the counter operator to know the eigenvalues and the eigenstates of H since these two elements are then linked by:

$$H=\hbar\omega_0\left(\frac{P^2+Q^2}{2}\right)=\hbar\omega_0\left(a^\dagger a+\frac{1}{2}\right)=\hbar\omega_0\left(N+\frac{1}{2}\right) \quad (11.2.618)$$

Therefore in general using the other notation:

$$H=\hbar\omega_0\left(a_\pm a_\mp\pm\frac{1}{2}\right) \quad (11.2.619)$$

therefore the Schrödinger equation can be written:

$$\boxed{H\Psi=E\Psi\Leftrightarrow\hbar\omega_0\left(a_\pm a_\mp\pm\frac{1}{2}\right)=E\Psi} \quad (11.2.620)$$

We now make the assumption that Ψ is an eigenfunction of N associated with the eigenvalue n such that:

$$\begin{aligned}&H\Psi=E\Psi\\&\Rightarrow\hbar\omega_0\left(N+\frac{1}{2}\right)\Psi_n=\hbar\omega_0\left(n+\frac{1}{2}\right)\Psi_n\\&\Rightarrow N\Psi_n=n\Psi_n\end{aligned} \quad (11.2.621)$$

This assumption is very important because we will use it as an induction principle to find all the eigenfunctions starting from the fundamental one!

Let us now establish commutator relations between N and N' and the operators a or $a^{\dagger}$. For this, let us first multiply $aa^{\dagger}-a^{\dagger}a=1$ by $a^{\dagger}$, then we get:

$$(aa^{\dagger}-a^{\dagger}a)a^{\dagger}=(aa^{\dagger})a^{\dagger}-a^{\dagger}(aa^{\dagger})=N'a^{\dagger}-a^{\dagger}N'=a^{\dagger} \qquad (11.2.622)$$

Similarly, by multiplying the $aa^{\dagger}-a^{\dagger}a=1$ by a, we get:

$$aa^{\dagger}a-a^{\dagger}aa=a(a^{\dagger}a)-(a^{\dagger}a)a=aN-Na=a \qquad (11.2.623)$$

Since according to our assumption Ψ and n are respectively eigenfunction and eigenvalue of N, we can write:

$$N\Psi=n\Psi \qquad (11.2.624)$$

But as we have:

$$Na=aN-a \qquad (11.2.625)$$

which multiplied on both side of the equality by the wave function Ψ gives the relation:

$$Na\Psi=(n-1)a\Psi=na\Psi-a\Psi \qquad (11.2.626)$$

This equation has the following consequences:

- We have $a\Psi=0$ such that $Na\Psi=na\Psi$
- Or we have $a\Psi$ that is an eigenfunction of N for the eigenvalue $(n-1)$

The same reasoning would establish that $a^{\dagger}\Psi$ is an eigenfunction of N' for the eigenvalue $n'+1$, if it is not zero (we shall see later that $a^{\dagger}\Psi$ is never zero):

$$N\left(a^{\dagger}\Psi\right)=\left(n+1\right)a^{\dagger}\Psi \qquad (11.2.627)$$

With the notations above it was still not clear for one of our read with the operator were named creation and annihilation operators. Let us see this with another approach before we continue! The following development was more explicit to explain their name (we keep the other notation of these operators for obvious reasons):

$$\begin{aligned}H\left(a_{+}\psi\right)&=\hbar\omega_0\left[a_{+}a_{-}+\frac{1}{2}\right]\left(a_{+}\psi\right)=\hbar\omega_0\left[a_{+}a_{-}a_{+}+\frac{1}{2}a_{+}\right]\psi\\&=\hbar\omega_0 a_{+}\left[a_{-}a_{+}+\frac{1}{2}\right]\psi=\hbar\omega_0 a_{+}\left[a_{+}a_{-}+\frac{1}{2}+1\right]\psi\\&=a_{+}\left[\hbar\omega_0\left(a_{+}a_{-}+\frac{1}{2}\right)+\hbar\omega_0\right]\psi=a_{+}\left(H+\hbar\omega_0\right)\psi\\&=a_{+}\left(E+\hbar\omega_0\right)\psi=\left(E+\hbar\omega_0\right)a_{+}\psi\end{aligned} \qquad (11.2.628)$$

By the same argument, we can show that $a_{-}\Psi$ is a wave function with energy $E-\hbar\omega_0$. For this reason a_{+} and a_{-} are called the raising and lowering operators, respectively:

$$\boxed{H(a_{\pm}\Psi)=(E\pm\hbar\omega_0)a_{\pm}\Psi} \qquad (11.2.629)$$

Now, since the harmonic oscillator potential is parabolic, we would expect no upper limit to the energy levels, but we would expect a lower limit, so that applying the lowering operator to the ground state should give us zero.

Therefore we know that there exists a lower eigenvalue n_0 smaller all the others corresponding to the fundamental level (according to the Bohr-Sommerfeld model this eigenvalue always exists).

Necessarily, its eigenfunction equation obeys the relation (the reader will be able to check with the results further):

$$a\Psi_0 = 0 \tag{11.2.630}$$

as they can not be , as we have already mention it, a level lower that the lower limit by applying the lowering operator!

Otherwise $n_0 - 1$ would a eigenvalue and there would a physical contradiction.

By multiplying this last relation by $a^\dagger$ we get:

$$a^\dagger a\Psi_0 = N\Psi_0 = 0 \tag{11.2.631}$$

Which shows that the minimum eigenvalue n_0 is zero. We therefore know the fundamental level of the oscillator:

$$E_0 = \hbar\omega_0\left(n_0 + \frac{1}{2}\right) = \frac{1}{2}\hbar\omega_0 \tag{11.2.632}$$

named "zero-point energy".

Remark

It should be noticed that the oscillator is never in a state of rest (put $n = 0$ in the expression of the energy) which also means that the absolute zero can not be accessible since the temperature "fix" the atomic agitation and that a rest state does not exist!

To get the corresponding eigenfunction, we need the explicit expression of a. According to:

$$P = \frac{p}{\sqrt{m\hbar\omega_0}} \qquad \text{and} \qquad Q = \sqrt{\frac{c^{te}}{\hbar\omega_0}}x = \sqrt{\frac{\omega_0 m}{\hbar}}x \tag{11.2.633}$$

Then we have:

$$P = \frac{1}{\sqrt{m\hbar\omega_0}}\frac{\hbar}{\mathrm{i}}\frac{\mathrm{d}}{\mathrm{d}x} \qquad \text{and} \qquad \mathrm{d}Q = \sqrt{\omega_0 m\hbar}\mathrm{d}x \Rightarrow \frac{1}{\mathrm{d}x} = \sqrt{\frac{\omega_0 m}{\hbar}}\frac{1}{\mathrm{d}Q} \tag{11.2.634}$$

which gives us:

$$P = \frac{1}{\sqrt{m\hbar\omega_0}}\frac{\hbar}{\mathrm{i}}\sqrt{\frac{\omega_0 m}{\hbar}}\frac{\mathrm{d}}{\mathrm{d}Q} = \frac{1}{\mathrm{i}}\frac{\mathrm{d}}{\mathrm{d}Q} \tag{11.2.635}$$

Hence:

$$\begin{aligned}\sqrt{2}a &= Q + \frac{\mathrm{d}}{\mathrm{d}Q}\\ \sqrt{2}a^\dagger &= Q - \frac{\mathrm{d}}{\mathrm{d}Q}\end{aligned} \tag{11.2.636}$$

But according to $a\Psi_0=0$:

$$\left(Q+\dfrac{\mathrm{d}}{\mathrm{d}Q}\right)\Psi_0=0 \tag{11.2.637}$$

hence:

$$\dfrac{1}{\Psi_0}\dfrac{\mathrm{d}}{\mathrm{d}Q}\Psi_0=-Q \tag{11.2.638}$$

thus (resolution of a simple differential equation):

$$\Psi_0(Q)=Ae^{-Q^2/2} \tag{11.2.639}$$

where A is the constant of integration. Thus, explicitly:

$$\Psi_0(x)=Ae^{-m\omega_0 x^2/\hbar 2} \tag{11.2.640}$$

We can normalize this function using the standard Gauss integral (see section Statistics page 695):

$$\int\limits_{-\infty}^{+\infty} e^{-ax^2}\mathrm{d}x=\sqrt{\dfrac{\pi}{a}} \tag{11.2.641}$$

so we get:

$$1=\int\limits_{-\infty}^{+\infty}|\Psi_0|^2\mathrm{d}x=|A|^2\int\limits_{-\infty}^{+\infty}e^{-m\omega x^2/\hbar}\mathrm{d}x=|A|^2\sqrt{\dfrac{\hbar\pi}{m\omega_0}} \tag{11.2.642}$$

Therefore:

$$A=\left(\dfrac{m\omega_0}{\pi\hbar}\right)^{1/4} \tag{11.2.643}$$

So finally:

$$\Psi_0=\left(\dfrac{m\omega_0}{\pi\hbar}\right)^{1/4}e^{-m\omega_0 x^2/\hbar 2} \tag{11.2.644}$$

It remains to build the other eigenfunctions and to normalize them. Indeed, if Ψ_n is a normalized function associated with the level E_n, we have seen above that $a^\dagger\Psi_n$ is an eigenfunction associated with the level $n+1$ (the operator of creation $a^\dagger$ increase the energy as we have show it...), but there is no reason to normalize it again since it is precisely associated with an eigenfunction already normalized!

Therefore we can write:

$$a^\dagger\Psi_n=\alpha_n\Psi_{n+1} \tag{11.2.645}$$

Where α_n is a coefficient to be determined. Let us express the fact that Ψ_n is already normed:

$$|\alpha_n|^2=\int\limits_{-\infty}^{+\infty}\overline{(a^\dagger\Psi_n)}\cdot(a^\dagger\Psi_n)\mathrm{d}x=\int\limits_{-\infty}^{+\infty}\bar{\Psi}_n\cdot(aa^\dagger\Psi_n)\mathrm{d}x \tag{11.2.646}$$

Considering the relation $aa^\dagger - a^\dagger a = 1$ we have:

$$|\alpha_n|^2 = \int_{-\infty}^{+\infty} \bar{\Psi}_n[(1+a^\dagger a)\Psi_n]\mathrm{d}x \tag{11.2.647}$$

Let us recall that $N(a^\dagger\Psi) = (n+1)(a^\dagger\Psi)$ therefore:

$$|\alpha_n|^2 = \int\limits_{-\infty}^{+\infty} \bar{\Psi}_n[(1+a^\dagger a)\Psi_n]\mathrm{d}x = \int\limits_{-\infty}^{+\infty} \bar{\Psi}_n[(1+N)\Psi_n]\mathrm{d}x = n+1 \tag{11.2.648}$$

We have just verified on the way that $a^\dagger\Psi_0$ is never zero (fact that we had supposed above).

All the functions Ψ_n (except the already Ψ_0 already fixed) have an arbitrary phase factor (concept that we have seen during the definition of bounded and unbound states), independently of each other, the argument α_n therefore remains available to us and we will choose α_n as being a fixed positive real number. This fixes all the Ψ_n according to the previous development:

$$a^\dagger\Psi_n = \sqrt{n+1}\Psi_{n+1} \tag{11.2.649}$$

By iterating this relation on the wave function, we obtain easily (elementary algebra but that we can detail on request):

$$\Psi_n = \frac{1}{\sqrt{n!}}(a^\dagger)^n\Psi_0 \tag{11.2.650}$$

by taking into account the following relations (which we have already proved just earlier above):

$$\begin{gathered}\sqrt{2}a = Q + \frac{\mathrm{d}}{\mathrm{d}Q}\\ \text{and} \qquad \Psi_0 = A^{-1/2}\pi^{-1/4}e^{-Q^2/2} = \left(\frac{m\omega_0}{\pi\hbar}\right)^{1/4} e^{-Q^2/2}\\ \sqrt{2}a^\dagger = Q - \frac{\mathrm{d}}{\mathrm{d}Q}\end{gathered} \tag{11.2.651}$$

We then have:

$$\Psi_n = \left(\frac{m\omega_0}{\pi\hbar}\right)^{1/4}(n!2^n)^{-1/2}\left(Q-\frac{\mathrm{d}}{\mathrm{d}Q}\right)^n e^{-Q^2/2} \tag{11.2.652}$$

This relation takes a simpler form, relying on the relation:

$$\left(Q-\frac{\mathrm{d}}{\mathrm{d}Q}\right)f(Q) = -e^{Q^2/2}\frac{\mathrm{d}}{\mathrm{d}Q}e^{-Q^2/2}f(Q) \tag{11.2.653}$$

Let us check this:

$$\begin{aligned}-e^{Q^2/2}\frac{\mathrm{d}}{\mathrm{d}Q}e^{-Q^2/2}f(Q) &= -e^{Q^2/2}\left(-Qe^{-Q^2/2}f(Q) + e^{-Q^2/2}\frac{\mathrm{d}}{\mathrm{d}Q}f(Q)\right)\\ &= -e^{Q^2/2}e^{-Q^2/2}\left(-Qf(Q)+\frac{\mathrm{d}}{\mathrm{d}Q}f(Q)\right) = \left(Q-\frac{\mathrm{d}}{\mathrm{d}Q}\right)f(Q)\end{aligned} \tag{11.2.654}$$

hence, in the language of operators:

$$Q-\frac{\mathrm{d}}{\mathrm{d}Q}=-e^{Q^2/2}\frac{\mathrm{d}}{\mathrm{d}Q}e^{-Q^2/2} \tag{11.2.655}$$

Therefore:

$$\left(Q-\frac{\mathrm{d}}{\mathrm{d}Q}\right)^n=(-1)^n e^{Q^2/2}\left(\frac{\mathrm{d}}{\mathrm{d}Q}\right)^n e^{-Q^2/2} \tag{11.2.656}$$

We thus obtain the expression of Ψ_n:

$$\Psi_n=(-1)^n\sqrt{\frac{1}{2^n n!}}\left(\frac{m\omega_0}{\pi\hbar}\right)^{1/4}\left[e^{Q^2/2}\frac{\mathrm{d}^n}{\mathrm{d}Q^n}e^{-Q^2/2}\right]e^{-Q^2/2} \tag{11.2.657}$$

Moreover, in the mathematical theory of families of orthogonal polynomials with the weight (with respect to the weight function[13] $w(x)=e^{-x}/2$), we encounter the "Hermite polynomials" defined by (physicist version!):

$$\boxed{H_n(Q)=(-1)^n e^{Q^2}\frac{\mathrm{d}^n}{\mathrm{d}Q^n}e^{-Q^2}} \tag{11.2.658}$$

They are polynomials of degree n, even or odd ($H_0=1$, $H_1=2Q$, $H_2=4Q^2-2$, $H_3=8Q^3-12Q$,...):

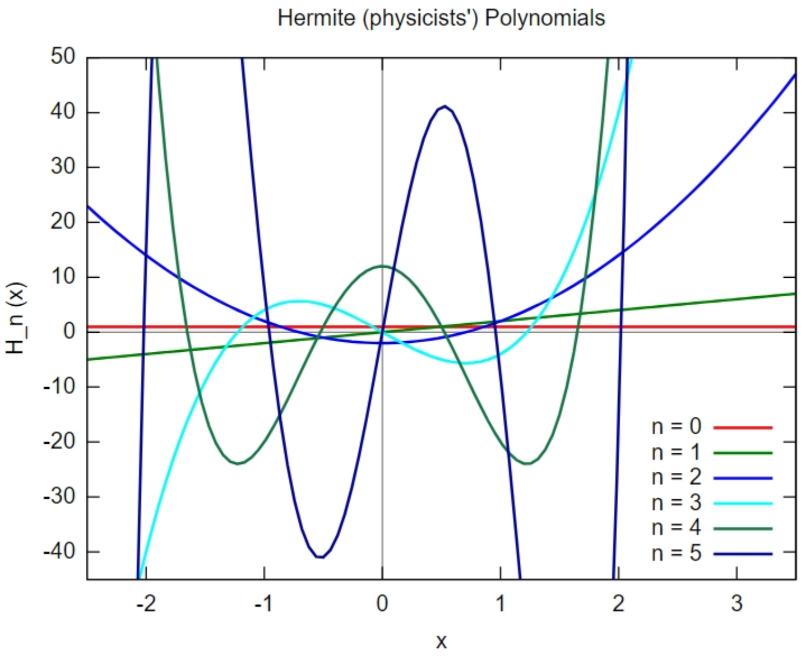


Figure 11.52: First five (physicists) Hermite polynomials (source: Wikipedia)

By using these physicists' Hermite polynomials, the notation of:

$$\Psi_n=(-1)^n\sqrt{\frac{1}{2^n n!}}\left(\frac{m\omega_0}{\pi\hbar}\right)^{1/4}\left[e^{Q^2/2}\frac{\mathrm{d}^n}{\mathrm{d}Q^n}e^{-Q^2/2}\right]e^{-Q^2/2} \tag{11.2.659}$$

---

[13] If you don't remember what is the weight function, you can refer of the section of Function Analysis

is simplified and becomes:

$$\boxed{\Psi_n=\sqrt{\frac{1}{2^n n!}}\left(\frac{m\omega_0}{\pi\hbar}\right)^{1/4}e^{-Q^2/2}H_n(Q)=N_n e^{-Q^2/2}H_n(Q)} \qquad (11.2.660)$$

where N_n is the normalization constant. Thus fully explicitly:

$$\Psi_n=(-1)^n\sqrt{\frac{1}{2^n n!}}\left(\frac{m\omega_0}{\pi\hbar}\right)^{1/4}e^{-\frac{\left(\sqrt{\frac{\omega_0 m}{\hbar}}x\right)^2}{2}}\left[e^{\frac{\left(\sqrt{\frac{\omega_0 m}{\hbar}}x\right)^2}{2}}\frac{\mathrm{d}^n}{\mathrm{d}\left(\sqrt{\frac{\omega_0 m}{\hbar}}x\right)^n}e^{-\left(\sqrt{\frac{\omega_0 m}{\hbar}}x\right)^2}\right] \qquad (11.2.661)$$

These polynomials therefore constitute an orthonormal basis of the global quantum state and thus appear naturally in the general expression of the functions / eigenstates.

Finally we have (keeping in mind that $H_0=1$, $H_1=2Q$, $H_2=4Q^2-2$, $H_3=8Q^3-12Q$, ...):

$\boldsymbol{n}$	$\boldsymbol{E_n}$	$\boldsymbol{\Psi_n}$
0	$\frac{1}{2}\hbar\omega_0$	$\left(\frac{m\omega_0}{\pi\hbar}\right)^{1/4}e^{-\frac{\omega_0 m}{\hbar}\frac{x^2}{2}}$
1	$\frac{3}{2}\hbar\omega_0$	$-\sqrt{\frac{1}{2}}\left(\frac{m\omega_0}{\pi\hbar}\right)^{1/4}2\sqrt{\frac{\omega_0 m}{\hbar}}xe^{-\frac{\omega_0 m}{\hbar}\frac{x^2}{2}}$
2	$\frac{5}{2}\hbar\omega_0$	$\sqrt{\frac{1}{8}}\left(\frac{m\omega_0}{\pi\hbar}\right)^{1/4}\left(4\frac{\omega_0 m}{\hbar}x^2-2\right)e^{-\frac{\omega_0 m}{\hbar}\frac{x^2}{2}}$
3	$\frac{7}{2}\hbar\omega_0$	$-\sqrt{\frac{1}{48}}\left(\frac{m\omega_0}{\pi\hbar}\right)^{1/4}\left(8\left(\frac{\omega_0 m}{\hbar}\right)^{3/2}x^3-12\sqrt{\frac{\omega_0 m}{\hbar}}x\right)e^{-\frac{\omega_0 m}{\hbar}\frac{x^2}{2}}$
...	...	...

Table 11.8: **First eigenfunctions and eigenvalues for the 1D harmonic oscillator**

Remark

The reader may notice that if we put $\sigma:=\sqrt{\hbar/(2\omega_0 m)}$ the fundamental mode density probability has a bell (Normal) shape:

$$\Psi_0=\sqrt{\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{x^2}{4\sigma^2}}}\Rightarrow|\Psi_0|^2=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}}=\mathcal{N}(0,\sigma^2) \qquad (11.2.662)$$

With the no less famous graphic representation with on the left the eigenfunctions associated Ψ_n and on the right the probability of presence $|\Psi_n|^2$:

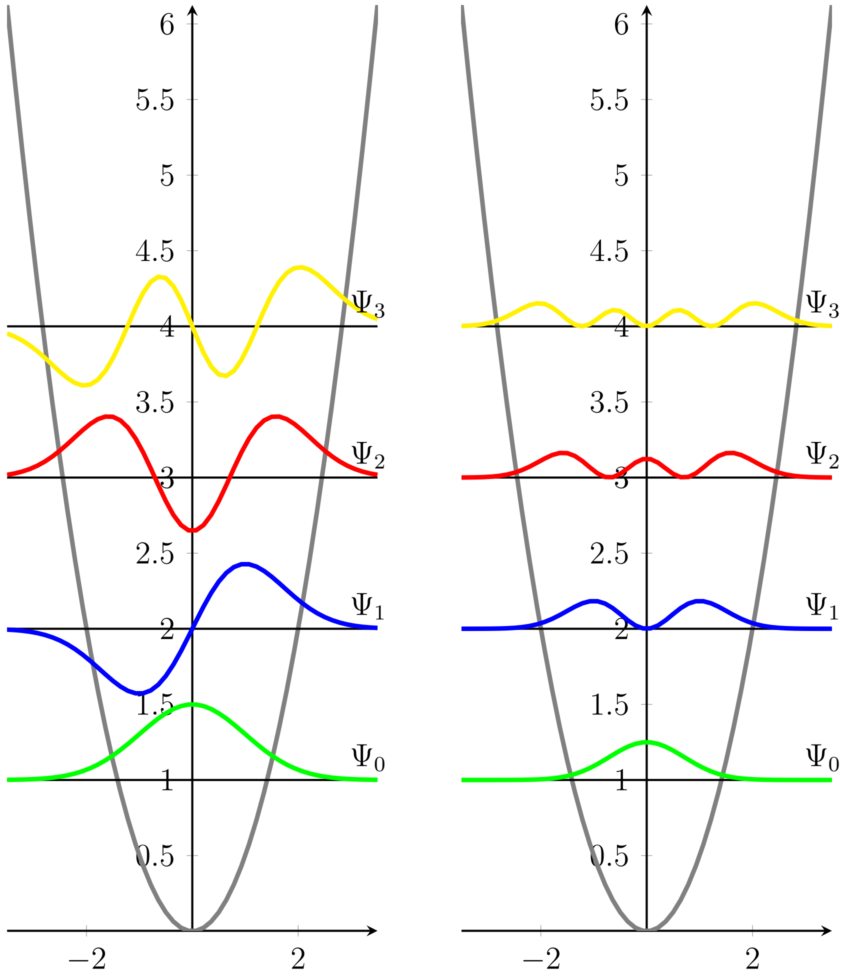


Figure 11.53: Plot of first eigenfunctions and density functions of some energy levels of the 1D harmonic oscillator

We can easily get the left part with Maple™ 4.00b (the example below is not normalized from hence the abscissa which is a bit special compared to the figure above!):

```
>m:=1;omega:=1;h:=1;
>plot([(m*omega/(Pi*h))^(-1/8)*exp(-m*omega*x^2/(2*h)),-((m*omega/(Pi*h))
^(1/4)*2)^(-1/2)*2*sqrt(m*omega/h)*x*exp(-m*omega*x^2/(2*h)),((m*omega/(Pi*h))
^(1/4)*8)^(-1/2)*(4*m*omega/h*x^2-2)*exp(-m*omega*x^2/(2*h)),-((m*omega
/(Pi*h))^(1/4)*48)^(-1/2)*(8*(m*omega/h)^(2/3)*x^3-12*sqrt(m*omega/h)*x)
*exp(-m*omega*x^2/(2*h))],x=-6..6);
```

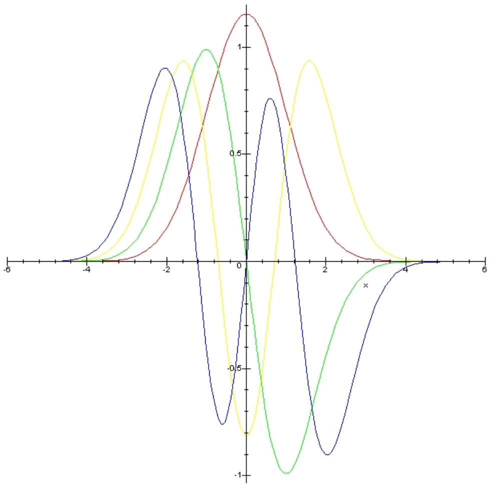
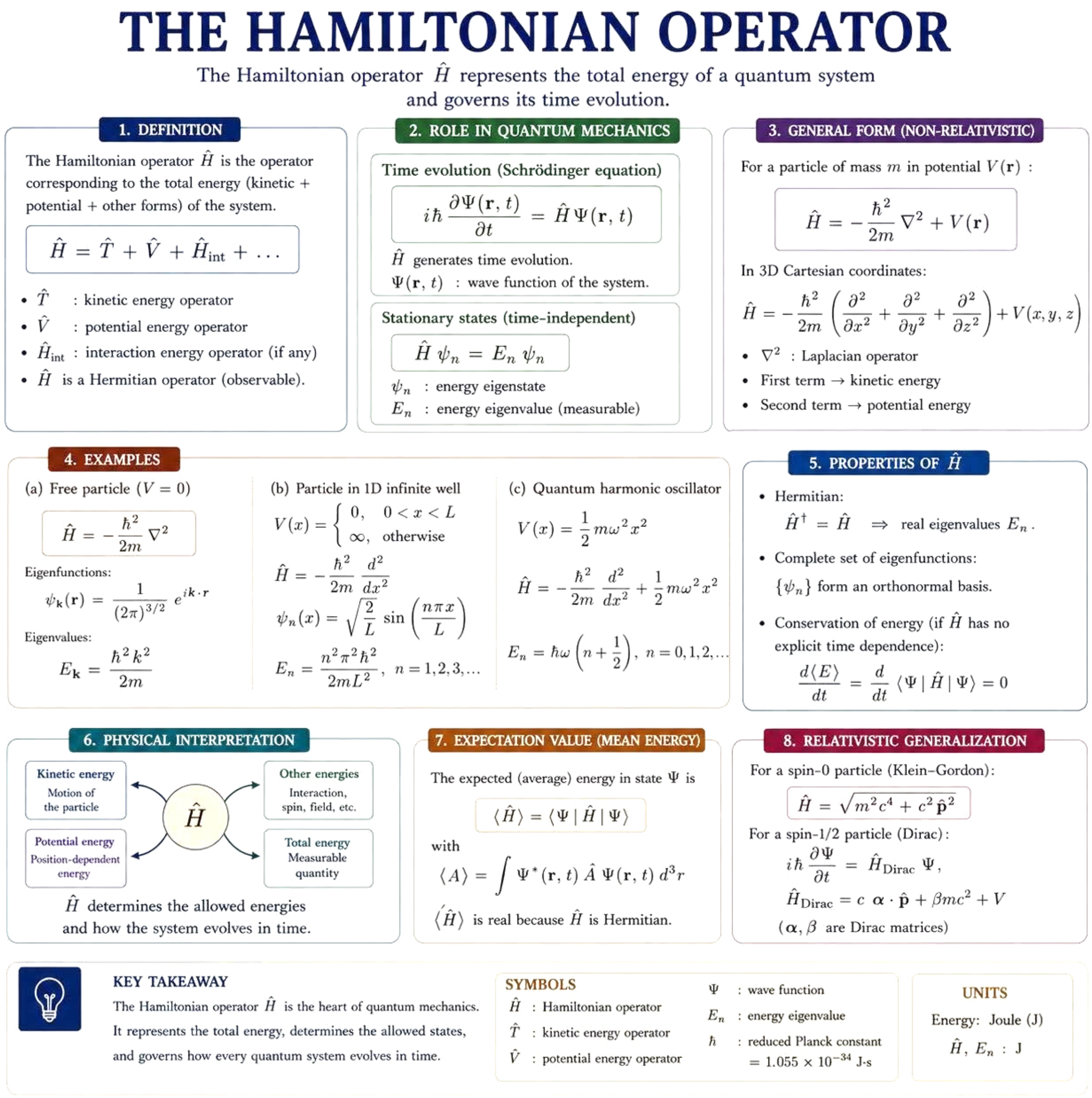


Figure 11.54: Plot of first eigenfunctions of the 1D harmonic oscillator in Maple™ 4.00b

By analysing these wave functions, we fall back on many classical results: the particle in the potential well has a wider probability of presence if it has a higher energy (a ball at the bottom of a well will rise higher on The edges if it has more energy), the particle is more likely to find itself on the positions distant from the center of the well when its energy is high.

For all calculations where particles are in a potential well, the harmonic approximation is very interesting. For example, if we wish to study a two-dimensional "harmonic trap", i.e. 2D Bose-Einstein condensate (see section Statistical Mechanics page 3576), we can set the following Hamiltonian to begin the study (in analogy with the one at 1 dimension used above):

$$H=\frac{p_x^2+p_y^2}{2m}+\frac{1}{2}m\omega_0^2\left(x^2+y^2\right) \tag{11.2.663}$$



THE HAMILTONIAN OPERATOR

The Hamiltonian operator $\hat{H}$ represents the total energy of a quantum system and governs its time evolution.

1. DEFINITION

The Hamiltonian operator $\hat{H}$ is the operator corresponding to the total energy (kinetic + potential + other forms) of the system.

$$\boxed{\hat{H} = \hat{T} + \hat{V} + \hat{H}_{\text{int}} + \dots}$$

- $\hat{T}$: kinetic energy operator
- $\hat{V}$: potential energy operator
- $\hat{H}_{\text{int}}$: interaction energy operator (if any)
- $\hat{H}$ is a Hermitian operator (observable).

2. ROLE IN QUANTUM MECHANICS

Time evolution (Schrödinger equation)

$$\boxed{i\hbar\,\frac{\partial \Psi(\mathbf{r},t)}{\partial t} = \hat{H}\,\Psi(\mathbf{r},t)}$$

$\hat{H}$ generates time evolution.

$\Psi(\mathbf{r},t)$: wave function of the system.

Stationary states (time-independent)

$$\boxed{\hat{H}\,\psi_n = E_n\,\psi_n}$$

ψ_n : energy eigenstate

E_n : energy eigenvalue (measurable)

3. GENERAL FORM (NON-RELATIVISTIC)

For a particle of mass m in potential $V(\mathbf{r})$:

$$\boxed{\hat{H} = -\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{r})}$$

In 3D Cartesian coordinates:

$$\hat{H} = -\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right) + V(x,y,z)$$

- ∇^2 : Laplacian operator
- First term → kinetic energy
- Second term → potential energy

4. EXAMPLES

(a) Free particle ($V = 0$)

$$\boxed{\hat{H} = -\frac{\hbar^2}{2m}\nabla^2}$$

Eigenfunctions:

$$\psi_{\mathbf{k}}(\mathbf{r}) = \frac{1}{(2\pi)^{3/2}}\,e^{i\mathbf{k}\cdot\mathbf{r}}$$

Eigenvalues:

$$E_{\mathbf{k}} = \frac{\hbar^2 k^2}{2m}$$

(b) Particle in 1D infinite well

$$V(x) = \begin{cases} 0, & 0 < x < L \\ \infty, & \text{otherwise} \end{cases}$$

$$\hat{H} = -\frac{\hbar^2}{2m}\frac{d^2}{dx^2}$$

$$\psi_n(x) = \sqrt{\frac{2}{L}}\,\sin\left(\frac{n\pi x}{L}\right)$$

$$E_n = \frac{n^2\pi^2\hbar^2}{2mL^2}, \quad n = 1, 2, 3, \dots$$

(c) Quantum harmonic oscillator

$$V(x) = \frac{1}{2}m\omega^2 x^2$$

$$\hat{H} = -\frac{\hbar^2}{2m}\frac{d^2}{dx^2} + \frac{1}{2}m\omega^2 x^2$$

$$E_n = \hbar\omega\left(n + \frac{1}{2}\right), \quad n = 0, 1, 2, \dots$$

5. PROPERTIES OF $\hat{H}$

- Hermitian:

  $\hat{H}^\dagger = \hat{H} \quad \Rightarrow \quad$ real eigenvalues E_n .
- Complete set of eigenfunctions:

  $\{\psi_n\}$ form an orthonormal basis.
- Conservation of energy (if $\hat{H}$ has no explicit time dependence):

$$\frac{d\langle E\rangle}{dt} = \frac{d}{dt}\langle\Psi\,|\,\hat{H}\,|\,\Psi\rangle = 0$$

6. PHYSICAL INTERPRETATION

Kinetic energy
Motion of the particle

Other energies
Interaction, spin, field, etc.

$\hat{H}$

Potential energy
Position-dependent energy

Total energy
Measurable quantity

$\hat{H}$ determines the allowed energies and how the system evolves in time.

7. EXPECTATION VALUE (MEAN ENERGY)

The expected (average) energy in state Ψ is

$$\boxed{\langle\hat{H}\rangle = \langle\Psi\,|\,\hat{H}\,|\,\Psi\rangle}$$

with

$$\langle A\rangle = \int \Psi^*(\mathbf{r},t)\,\hat{A}\,\Psi(\mathbf{r},t)\,d^3r$$

$\langle\hat{H}\rangle$ is real because $\hat{H}$ is Hermitian.

8. RELATIVISTIC GENERALIZATION

For a spin-0 particle (Klein–Gordon):

$$\boxed{\hat{H} = \sqrt{m^2c^4 + c^2\hat{\mathbf{p}}^2}}$$

For a spin-1/2 particle (Dirac):

$$i\hbar\,\frac{\partial\Psi}{\partial t} = \hat{H}_{\text{Dirac}}\,\Psi,$$

$$\hat{H}_{\text{Dirac}} = c\;\boldsymbol{\alpha}\cdot\hat{\mathbf{p}} + \beta mc^2 + V$$

($\boldsymbol{\alpha}, \beta$ are Dirac matrices)

KEY TAKEAWAY

The Hamiltonian operator $\hat{H}$ is the heart of quantum mechanics. It represents the total energy, determines the allowed states, and governs how every quantum system evolves in time.

SYMBOLS

$\hat{H}$: Hamiltonian operator

$\hat{T}$: kinetic energy operator

$\hat{V}$: potential energy operator

Ψ : wave function

E_n : energy eigenvalue

$\hbar$: reduced Planck constant $= 1.055 \times 10^{-34}$ J·s

UNITS

Energy: Joule (J)

$\hat{H}$, E_n : J

11.2.4.4.7 Planck's Zero-Point Energy

The assumption of discrete radiation energy in Planck's 11901 (holocene calendar) radiation law, conflicted with Planck's intuition in radiation of continuous waves. To reconcile his quantum hypothesis with his conception of wave radiation, he avoided the conclusion that radiation energy must be made of particles, and postulated that radiation is a transition between the energy levels of an oscillator. Furthermore, ignoring the symmetry between emission and absorption, he maintained that the absorption of radiation energy is continuous.

Under these assumptions, Planck derived in 11912 (holocene calendar), a second radiation law in which zero-point energy also appears. We will show that Planck's derivation of his 11912 (holocene calendar) radiation law only recovers the zero-point energy that he unknowingly assumed in his model from the start. Furthermore, the distribution of 11912 (holocene calendar) Planck's law radiation law is, in fact, the approximated 11901 (holocene calendar Boson statistics of Planck's radiation law.

Remark

Our study earlier of the Quantum oscillator and also the following study of Planck's 11912 (holocene calendar) radiation law are two of the many derivation of zero-point energy that all lead to the same result. There are other famous derivations like the one based on the Heisenberg uncertainty relations or based on Quantum Field Theory.

Our main result is that Planck's ZPE radiation law is equivalent to the combined three assumptions of zero-point energy hypothesis, the Quantum Law, and the approximated Boson Statistics distribution law. The validity of Planck's 11912 (holocene calendar) radiation law, and the existence of Planck's zero-point energy are doubtful.

Planck's 11912 (holocene calendar) oscillator model assumes[14] a probability p_ω for the oscillator to radiate, and a probability $q_\omega = 1 - p_\omega$ to not radiate. He assumes afterwards in a quite non-intuitive way (see above page 4814):

- Energy of $\frac{1}{2}\hbar\omega$ with probability $p_{\omega,1}$
- Energy of $\frac{3}{2}\hbar\omega$ with probability $p_{\omega,2} = p_{\omega,1}q_\omega$
- ...
- Energy of $\frac{2n-1}{2}\hbar\omega$ with probability $p_{\omega,n} = p_{\omega,1}q_\omega^{n-1}$

Since (using the result of geometric series as proved page 1679):

$$1 = p_{\omega,1} + p_{\omega,2} + \ldots = p_{\omega,1}\left(1 + q_\omega + q_\omega^2 + \ldots\right) = \frac{p_{\omega,1}}{1 - q_\omega} = \frac{p_{\omega,1}}{p_\omega} \tag{11.2.664}$$

We put:

$$p_{\omega,1} = p_\omega, \quad \text{therefore} \quad p_{\omega,n} = p_\omega q_\omega^{n-1} \tag{11.2.665}$$

---

[14] The texts and developments that follow are copied from [124].

Therefore, the oscillator average entropy is:

$$\begin{aligned}
\bar{S}_\omega &= -k\left(p_{\omega,1}\ln(p_{\omega,1})+p_{\omega,2}\ln(p_{\omega,2})+\ldots\right)\\
&= -kp_\omega\left(\ln(p_\omega)+q_\omega\ln(p_\omega q_\omega)+q_\omega^2\ln(p_\omega q_\omega^2)+\ldots\right)\\
&= -kp_\omega\left(1+q_\omega+q_\omega^2+\ldots\right)\ln(p_\omega)-kp_\omega q_\omega\left(1+2q_\omega+3q_0^2+\ldots\right)\ln(q_\omega)\\
&= -kp_\omega\ln(p_\omega)\frac{1}{1-q_\omega}-kp_\omega q_\omega\ln(q_\omega)\frac{\mathrm{d}}{\mathrm{d}q_\omega}\left(1+q_\omega+q_\omega^2+\ldots\right)\\
&= -kp_\omega\ln(p_\omega)\frac{1}{1-q_\omega}-kp_\omega q_\omega\ln(q_\omega)\frac{\mathrm{d}}{\mathrm{d}q_\omega}\frac{1}{1-q_\omega}\\
&= -kp_\omega\ln(p_\omega)\frac{1}{1-q_\omega}-kp_\omega q_\omega\ln(q_\omega)\frac{1}{(1-q_\omega)^2}\\
&= -k(1-q_\omega)\ln(p_\omega)\frac{1}{1-q_\omega}-k(1-q_\omega)(1-p_\omega)\ln(1-p_\omega)\frac{1}{(1-q_\omega)^2}\\
&= -k\ln(p_\omega)-k\frac{1-p_\omega}{p_\omega}\ln\left(1-p_\omega\right)\\
&= k\left\{-\ln(p_\omega)-\frac{1-p_\omega}{p_\omega}\ln\left(\frac{1-p_\omega}{p_\omega}\right)-\frac{1-p_\omega}{p_\omega}\ln(p_\omega)\right\}\\
&= k\left\{\frac{1}{p_\omega}\ln\left(\frac{1}{p_\omega}\right)-\left(\frac{1}{p_\omega}-1\right)\ln\left(\frac{1}{p_\omega}-1\right)\right\}
\end{aligned} \tag{11.2.666}$$

The expected mean radiation energy of an oscillator is:

$$\begin{aligned}
\bar{E}_\omega &= \frac{1}{2}\hbar\omega p_{\omega,1}+\frac{3}{2}\hbar\omega p_{\omega,2}+\frac{5}{2}\hbar\omega p_{\omega,3}+\frac{7}{2}\hbar\omega p_{\omega,4}+\ldots\\
&= \frac{1}{2}\hbar\omega p_\omega\left\{1+3q_\omega+5q_\omega^2+7q_\omega^3+\ldots\right\}\\
&= \frac{1}{2}\hbar\omega p_\omega\left\{\left(1+q_\omega+q_\omega^2+q_\omega^3\ldots\right)+2q_\omega\left(1+2q_\omega+3q_\omega^2\cdots\right)\right\}\\
&= \frac{1}{2}\hbar\omega p_\omega\left\{\frac{1}{1-q_\omega}+2q_\omega\frac{\mathrm{d}}{\mathrm{d}q_\omega}\left(1+q_\omega+q_\omega^2+q_\omega^3+\ldots\right)\right\}\\
&= \frac{1}{2}\hbar\omega p_\omega\left\{\frac{1}{p_\omega}+2q_\omega\frac{\mathrm{d}}{\mathrm{d}q}\frac{1}{1-q_\omega}\right\}=\frac{1}{2}\hbar\omega p_\omega\left\{\frac{1}{p_\omega}+2q_\omega\frac{1}{(1-q_\omega)^2}\right\}\\
&= \hbar\omega\left\{\frac{1}{2}+(1-p_\omega)\frac{1}{p_\omega}\right\}=\hbar\omega\left(\frac{1}{p_\omega}-\frac{1}{2}\right)
\end{aligned} \tag{11.2.667}$$

Therefore $\bar{S}_\omega$ in terms of $\bar{E}_\omega$ is:

$$\bar{S}_\omega=k\left\{\left(\frac{\bar{E}_\omega}{\hbar\omega}+\frac{1}{2}\right)\ln\left(\frac{\bar{E}_\omega}{\hbar\omega}+\frac{1}{2}\right)-\left(\frac{\bar{E}_\omega}{\hbar\omega}-\frac{1}{2}\right)\ln\left(\frac{\bar{E}_\omega}{\hbar\omega}-\frac{1}{2}\right)\right\} \tag{11.2.668}$$

Using $\Delta S=\Delta Q/T$ (see section Thermodynamics page 3658) that leads us to:

$$\frac{\partial\bar{S}_\omega}{\partial\bar{E}_\omega}=\frac{1}{T} \tag{11.2.669}$$

Hence:

$$\frac{\partial\bar{S}_\omega}{\partial\bar{E}_\omega}=\frac{1}{T}=k\left(\frac{\ln\left(\frac{\bar{E}_\omega}{\hbar\omega}+\frac{1}{2}\right)}{\hbar\omega}-\frac{\ln\left(\frac{\bar{E}_\omega}{\hbar\omega}-sols\frac{1}{2}\right)}{\hbar\omega}\right) \tag{11.2.670}$$

Solving for $\bar{E}_\omega$ gives:

$$\bar{E}_\omega = \frac{1}{2}\frac{\hbar\omega\,(1+\exp(\hbar\omega/kT))}{\exp(\hbar\omega/kT)-1} \qquad (11.2.671)$$

that after rearranging gives us the "Planck's 11912 (holocene calendar) radiation law":

$$\boxed{\bar{E}_\omega = \frac{\hbar\omega}{\exp(\hbar\omega/kT)-1}+\frac{1}{2}\hbar\omega} \qquad (11.2.672)$$

The zero-point energy of $\frac{1}{2}\hbar\omega$ is the mid-energy that was assumed with probability p_ω. Planck's derivation only recovers the ZPE that he assumed at the start.

As already mentioned, the Casimir force (see page 4825) between two plates is related to the zero-point energy of the electromagnetic field. The zero-point energy appeared for the first time in Planck's "second theory" of black-body radiation, more than a decade before it was actually derived from Heisenberg's matrix mechanics in 11925 (holocene calendar) Planck obtained the expression:

$$\bar{E}_\omega = \frac{\hbar\omega}{\exp(\hbar\omega/kT)-1}+\frac{1}{2}\hbar\omega \qquad (11.2.673)$$

for the average energy of an oscillator of frequency ω in equilibrium with radiation at temperature T. When $T\rightarrow 0$, $\bar{E}_\omega \rightarrow \frac{1}{2}\hbar\omega$. In other words, Planck's work implied the existence of a zero-point energy persisting even at the absolute zero of temperature where, according to conventional classical theory, all motion ceases and given by:

$$\boxed{E_{\text{ZPE}} = \frac{1}{2}\hbar\omega} \qquad (11.2.674)$$

Planck soon became convinced that this constant additive zero-point energy is of no physical consequence. Albert Einstein, however, was intrigued by the idea. In a paper with Otto Stern in 11913 (holocene calendar), he noted that a zero-point energy seems necessary in order for the relation:

$$\bar{E}_\omega = U = \frac{\hbar\omega}{\exp(\hbar\omega/kT)-1}+\frac{1}{2}\hbar\omega \qquad (11.2.675)$$

to give the classical relation $U=kT$ in the classical limit $kT \gg \hbar\omega$. In this limit, retaining terms up to first order in $\hbar\omega/kT$:

$$\begin{aligned} U &= \frac{\hbar\omega}{\exp(\hbar\omega/kT)-1}+\frac{1}{2}\hbar\omega \cong \frac{\hbar\omega}{1+\hbar\omega/kT+\frac{1}{2}(\hbar\omega/kT)^2-1}+\frac{1}{2}\hbar\omega \\ &= \frac{kT}{1+\frac{1}{2}\hbar\omega/kT}+\frac{1}{2}\hbar\omega \cong kT-\frac{1}{2}\hbar\omega+\frac{1}{2}\hbar\omega = kT \end{aligned} \qquad (11.2.676)$$

The zero-point energy $\frac{1}{2}\hbar\omega$ is thus necessary in order that U does not have first-order quantum corrections to kT in the classical limit $kT \gg \hbar\omega$. For this and other reasons, Einstein and Stern concluded, The existence of a zero-point energy $\left[\frac{1}{2}\hbar\omega\right]$ is probable.

Zero-point energy was a "hot topic" during the decade preceding Heisenberg's matrix mechanics and Schrödinger's wave mechanics, mainly because of the great interest at the time in low-temperature phenomena such as superconductivity. Debye, for instance, showed that, if Planck's zero-point energy were real, there should be a reduction in the X-ray scattering from crystals even as $T \rightarrow 0$ because of zero-point fluctuations about the lattice sites. Mulliken's spectroscopic studies indicates that a zero-point energy of molecular vibrations was necessary to fit the data accurately, and Stern found a zero-point energy to be necessary in calculating the vapour pressure of solids. Early ideas about the zero-point energy have been reviewed elsewhere (see [394]).

It is typical in more recent times for textbooks to regard the zero-point energy as a curious prediction of quantum theory that can be discarded by simply redefining the zero of energy. This can be misleading if it implies that the zero-point energy is of no physical consequence whatsoever, for it turns out that changes in the zero-point energy are observable. The most precise observations have been made in connection with changes in the electromagnetic zero-point energy.

Many physical effects attributed to zero-point energy have been experimentally verified, such as (according to the scientific consensus in the early 121st century) spontaneous emission, Casimir force, Lamb shift, magnetic moment of the electron and Delbrück scattering. These effects are usually named "radiative corrections". In more complex non-linear theories (e.g. QCD) zero-point energy can give rise to a variety of complex phenomena such as multiple stable states, symmetry breaking, chaos and emergence. Many physicists think that the vacuum may hold one of the main key to a full understanding of Nature and that studying it is critical in the search for the theory of everything. Active areas of research include the effects of virtual particles, quantum entanglement, the difference (if any) between inertial and gravitational mass, variation in the speed of light, a reason for the observed value of the cosmological constant and the nature of dark energy.

The vacuum has, implicitly, all of the properties that a particle may have: spin, or polarization in the case of light, energy, and so on. On average, most of these properties cancel out: the vacuum is, after all, "empty" in this sense. One important exception is the vacuum energy or the vacuum expectation value of the energy.

11.2.4.4.8 Electromagnetic Casimir Force

What happens if we take two (almost) perfect mirrors (conductors or insulator) and arrange them so that they are facing each other a few nanometres apart in empty space (assumed as because we will see during study of quantum field theory that there are not such things as "empty space")? Your first reaction might be “nothing at all”. In fact, both mirrors are mutually attracted to each other by the simple presence of the vacuum[15].

This startling phenomenon was first predicted in 11948 (holocene calendar) by the Dutch theoretical physicist Hendrik Casimir while he was working at Philips Research Laboratories in Eindhoven on – of all things – colloidal solutions. The phenomenon is now named the "Casimir effect", while the force between the mirrors is known as the "Casimir force".

The Casimir effect is important because it gives strong evidence that the vacuum is not empty but a dynamical medium with properties (like $\varepsilon_0, \mu_0, \Lambda ...$) that led to an emergence point of view of elementary particles: they are collective motions of the vacuum medium!

---

[15] Some configurations may produce repulsive Casimir forces and these latter seems to have been confirmed experimentally (see [404]).

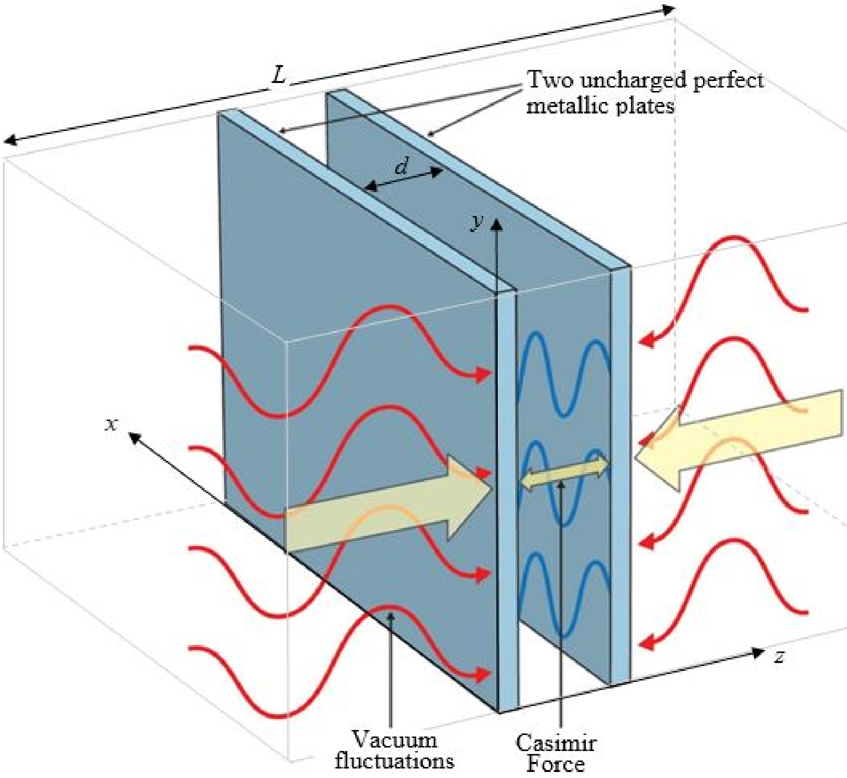
The experimental configuration of the Casimir force involving electromagnetic radiation is named "Electromagnetic Casimir force" because there are also theories involving the same type of force but for gravitation named "Gravitational Casimir force".

Remark

Some physicists disagree with this point of view (i.e. vacuum fluctuation cause) and argue - some with poor mathematical proofs and some others with not bad one - that the phenomenon is due to pure Van der Waals forces. However the Van der Waals forces seems not to be able yet to explain the case where the experimental configuration generates a repulsive force. However careful and always more precise experiments seems still in year 12020 (holocene calendar) to favour the quantum fluctuation option and therefore be a plausible explanation of dark energy too (see [377] for more details). More investigations are needed to get to a conclusive fact!

11.2.4.4.8.1 One dimensional case (naive approach)

Consider for our introduction to the study of that phenomenon, the following configuration:



It is an ideal condenser with infinite perfectly conducting plates in the x and y directions separated by a distance d along the z direction (the origin of the z axis is located on the left plate).

All electromagnetic fields have a characteristic "spectrum" containing many different frequencies. In a free vacuum all of the frequencies are of equal importance. But inside a cavity, where the field is reflected back and forth between the mirrors, the situation is different. The field is amplified if integer multiples of half a wavelength can fit exactly inside the cavity. This wavelength corresponds to a "cavity resonance". At other wavelengths, in contrast, the field is suppressed. Vacuum fluctuations are suppressed or enhanced depending on whether their frequency corresponds to a cavity resonance or not.

And to begin, for pedagogical reasons, we will consider the side view (profile) of the above configuration to deal with the one-dimensional case (along the z axis).

When two plates are at a distance d from each other, because of the boundary conditions the field modes which may exist between them are restricted to wavelengths which are whole fractions of d (otherwise particle from vacuum would be absorbed by the plates and it may violated conservation energy).

Let us recall that the electromagnetic zero-point energy is given by (see page 4822):

$$E_{\text{ZPE}}=\frac{1}{2}\hbar\omega=\frac{1}{2}\hbar 2\pi f=\frac{1}{2}\hbar 2\pi\frac{c}{\lambda}=\frac{\hbar c\pi}{\lambda} \tag{11.2.677}$$

Then if we consider only one dimension of space, the zero-point energy of mode n, using (see section Mechanics page 3485):

$$\lambda=\frac{2d}{n} \tag{11.2.678}$$

is then equal to:

$$E_{\text{ZPE},n}=\frac{n\hbar c\pi}{2d} \tag{11.2.679}$$

If we sum over all the modes, we therefore have the total zero-point energy as a function of the distance between the plates d which is equal to:

$$E_{\text{ZPE},n}(d)=\frac{\hbar c\pi}{2d}\left(\sum_{n=1}^{+\infty}n\right) \tag{11.2.680}$$

We can use the Riemann zeta regularization result (see section Sequences and Series page 1683):

$$\zeta(-1)=\sum_{n=1}^{+\infty}n=-\frac{1}{12} \tag{11.2.681}$$

and then we get:

$$E_{\text{ZPE}}(d)=-\frac{\hbar c\pi}{24d} \tag{11.2.682}$$

We can then derive the 1D electromagnetic vacuum force exerted on the plates by deriving the energy with respect to d:

$$\boxed{F_{\text{ZPE}}^{\text{Casimir}}(d)=\frac{\mathrm{d}E_{\text{ZPE}}(d)}{\mathrm{d}d}=\frac{\hbar c\pi}{24d^2}} \tag{11.2.683}$$

But if the reader has a problem with using the equality $\sum_n n=-1/12$, we can look for a more physical justification. When the plates are at a distance d the energy just calculated seems infinite. But if we move the plates away at a great distance, there is always a density energy associated with the modes of the electromagnetic field. The difference is that in this situation the modes are no longer quantized. So we basically have to replace $\sum_n n$ by $\int x\mathrm{d}x$. Here too we get something infinite, but a little bigger. Now if we take this situation as a point of reference and we have take fun subtracting between the two situations (plates at distance d and plates "very far") we find:

$$\Delta E_{\text{ZPE}}(d)=\frac{\hbar c\pi}{2L}\left(\sum_{n=1}^{+\infty}n-\int\limits_0^{+\infty}x\,\mathrm{d}x\right) \tag{11.2.684}$$

So we subtract two infinities... how to do we deal with that?

We simply use the Euler-Maclaurin expansion (see section Sequences and Series page 1674) that gives us:

$$\sum_{n=1}^{+\infty}f(n)-\int\limits_0^{+\infty}f(x)\mathrm{d}x=\frac{1}{2}f(0)-\frac{1}{6\cdot 2!}f'(0)+\ldots=\frac{1}{2}f(0)-\frac{1}{12}f'(0)+\ldots \tag{11.2.685}$$

The funny thing is that if we (illegally) apply this formula to the function $f(x)=x$, we find:

$$\left(\sum_{n=1}^{+\infty}n-\int x\,\mathrm{d}x\right)=-\frac{1}{12} \tag{11.2.686}$$

And we fall back on the Casimir electromagnetic energy (and therefore also the corresponding force).

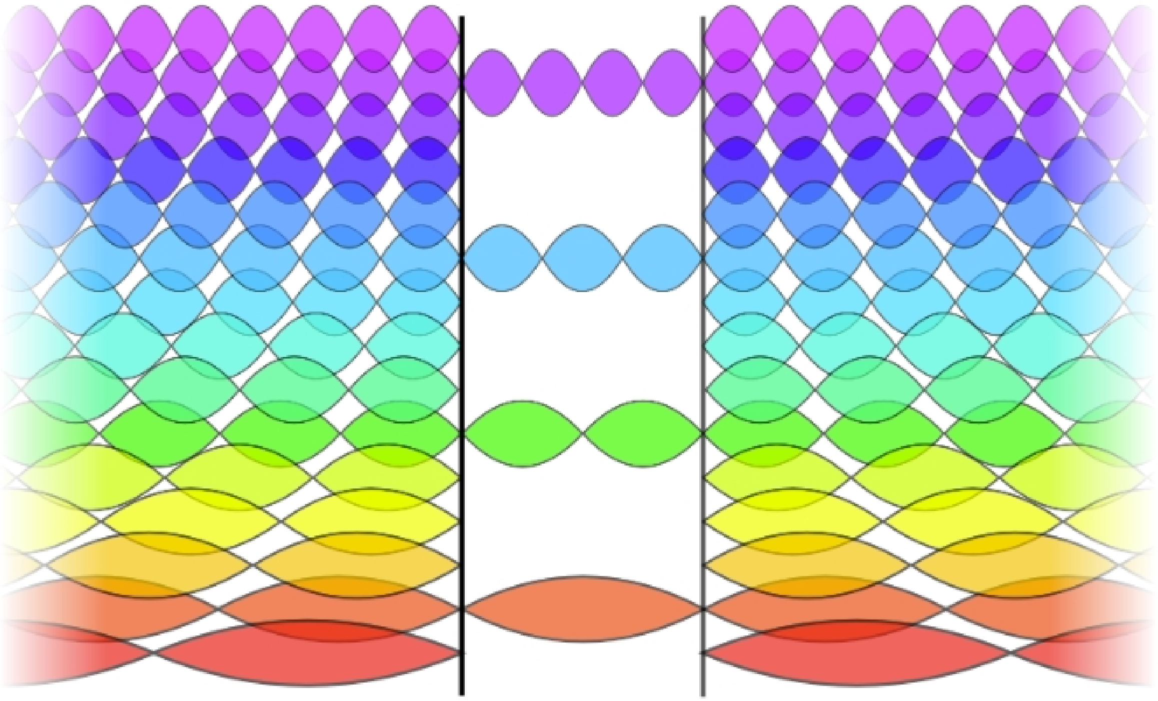


Figure 11.55: Outside the cavity formed by the plates, all vacuum frequencies are allowed. Within the cavity, however, the vacuum modes take on discrete frequencies. Changing the width of the cavity changes the density of modes relative to free space, which yields an energy difference.

On the other hand we could not do:

$$\left(\sum_{n=1}^{+\infty} n - \int x \mathrm{d}x\right) = \frac{n(n+1)}{2} - \int x \mathrm{d}x = -\frac{1}{12} \tag{11.2.687}$$

That latter approach sheds light on the deep origin of $-1/12$. In fact the sum $1+2+3+4+\ldots$ is indeed infinite, but $-1/12$ is what separates it from $\int x\mathrm{d}x$ which is also infinite, is that we can see as a base that we subtract. In Casimir's case, this is indeed the "based level" energy level, when the plates are very far apart.

11.2.4.4.8.2 Three dimensional case

Let us again consider an ideal condenser with infinite perfectly conducting plates in the x and y directions separated by a distance d along the z direction (the origin of the z axis is located on the left plate) as shown earlier above in the figure above.

The interaction energy between the two plates can be defined as the difference between the zero-point energies contained in the space between the planes, in the two respective configurations. This supposes that the field outside the cavity is not changed, which, of course, holds classically. For a discussion of this point, as well as for corrections due the finiteness of the plates. In general, the zero-point energy of the electromagnetic field is given:

$$E_{\text{cav}} = \sum_k g_s(k)\frac{\hbar\omega_k}{2} \tag{11.2.688}$$

where the sum runs over the normal modes of the field, where the ω_k are the frequencies of these modes and where $g_s(k)$ is the degeneracy of the mode k, due to polarization.

Let us start with the case of the cavity. In order to identify the modes more easily, let us consider, as usual, field configurations which are periodic in the x and y directions with wavelengths L_x and L_y respectively. The normal modes are determined by the boundary conditions on the surface of the conductors.

We remind that the tangential electric field and the normal component of the magnetic field should vanish on these boundaries. Let us recall that during our study of Maxwell's equations we have proved that plane wave solutions are given by:

$$\vec{E} = \vec{E}_0 e^{\mathrm{i}(\vec{k}\cdot\vec{x}-\omega t)} \qquad \vec{B} = \vec{B}_0 e^{\mathrm{i}(\vec{k}\cdot\vec{x}-\omega t)} \tag{11.2.689}$$

Here, $\vec{E}_0$ and $\vec{B}_0$ are called the "polarization vectors" for the electric and magnetic fields and often denoted $\vec{\mathcal{E}}$. These are complex 3 dimensional vectors. The wavevector $\vec{k}$ and angular frequency ω are real and in the vacuum are related by $\omega = c|\vec{k}|$.

And remember that:

$$\vec{k}\circ\vec{E}_0 = 0 \qquad \vec{k}\circ\vec{B}_0 = 0 \quad \vec{E}_0\circ\vec{B}_0 = 0 \tag{11.2.690}$$

Let us remind that:

$$k_x = n_x\frac{2\pi}{L_x} \quad \text{and} \quad k_y = n_y\frac{2\pi}{L_y} \tag{11.2.691}$$

where n_x and n_y are integer numbers and with:

$$k_z = n_z \frac{\pi}{d} \tag{11.2.692}$$

where n_z is a positive integer, the vector $\vec{\mathcal{E}}$ is the polarization vector and where $\vec{k} = (k_x, k_y, k_z)$.

Remark

The first case, $k = 2\pi n/L$, corresponds to the solutions of the Schrödinger's equation with zero potential, in the interval $[0, L]$, provided periodic boundary conditions are used, i.e. the wavefunction must satisfy the conditions $\Psi(0) = \Psi(L)$ and $\Psi'(0) = \Psi'(L)$.

The second case, $k = \pi n/d$, corresponds to the solutions of the Schrödinger's equation with zero potential, in the interval $[0, d]$, provided hard wall boundary conditions are used, i.e. the wavefunction must vanish at the boundary points 0 and d.

We consider a conducting box of volume L^3, with an $L \times L$ plate contained within, parallel to the x and y axes of the box.

The conducting walls of the box impose Dirichlet boundary conditions on the possible electromagnetic modes of the cavity: the tangential component of the electromagnetic field must vanish there. Solving the wave equation with these conditions reveals that the boundaries quantize the x, y, and z components of the wavevectors $\vec{k}$, such that they take on the discrete values:

$$k_x = n_x \frac{2\pi}{L} \qquad k_y = n_y \frac{2\pi}{L} \qquad k_z = n_z \frac{\pi}{d} \tag{11.2.693}$$

where:

$$k^2 = k_x^2 + k_y^2 + k_z^2 \tag{11.2.694}$$

In a vacuum, angular frequency ω is related to the wave number k by the speed of light c:

$$\omega = ck \tag{11.2.695}$$

This allows us to express the zero-point energy of the field:

$$E_{\text{ZPE}} = \frac{1}{2} \sum_{\vec{k},\lambda} \hbar\omega \tag{11.2.696}$$

as a sum over quantized modes, identified by k_x, k_y, and k_z:

$$E_{\text{ZPE}} = \hbar c \sum \frac{1}{2} \sqrt{k_x^2 + k_y^2 + k_z^2} \tag{11.2.697}$$

Polarization allows for two waves corresponding to each wave number except when one k_i is 0[16] (the k_z in our present situation!). Therefore, we multiply all $k > 0$ terms by 2, and expand the summation in x, y, and z:

$$E_{\text{ZPE}} = \hbar c \sum_{n_x=0}^{+\infty} \sum_{n_y=0}^{+\infty} \left[\frac{1}{2} \sqrt{\left(\frac{n_x 2\pi}{L}\right)^2 + \left(\frac{n_y 2\pi}{L}\right)^2} + \sum_{n_z=1}^{+\infty} \sqrt{\left(\frac{n_x 2\pi}{L}\right)^2 + \left(\frac{n_y 2\pi}{L}\right)^2 + \left(\frac{n_z \pi}{d}\right)^2} \right] \tag{11.2.698}$$

---

[16] This apparent special case arises from the "folding over" of the sum over both positive and negative $\vec{k}$ (from $-\infty$ to $+\infty$) into a sum from 0 to $+\infty$ - which simplifies evaluation later.

since n_x and n_y are presumed to be very large compared to L, we treat them as continuous variables and convert their sums to integrals, writing:

$$E_{\text{ZPE}}=\hbar c\int\limits_0^{+\infty}\int\limits_0^{+\infty}\frac{1}{2}\sqrt{\left(\frac{n_x2\pi}{L}\right)^2+\left(\frac{n_y2\pi}{L}\right)^2}+\sum_{n_z=1}^{+\infty}\sqrt{\left(\frac{n_x2\pi}{L}\right)^2+\left(\frac{n_y2\pi}{L}\right)^2+\left(\frac{n_z\pi}{a}\right)^2}\mathrm{d}n_x\mathrm{d}n_y \tag{11.2.699}$$

The zero point energy of the cavity for the $k_z\neq 0$ modes (indicated by the prime) is given by:

$$E'_{\text{ZPE}}=\hbar c\sum_{n_x=-\infty}^{+\infty}\sum_{n_y=-\infty}^{+\infty}\sum_{n_z=1}^{+\infty}\sqrt{\left(\frac{n_x2\pi}{L_x}\right)^2+\left(\frac{n_y2\pi}{L_x}\right)^2+\left(\frac{n_z\pi}{d}\right)^2} \tag{11.2.700}$$

As usual, one replaces the summation on n_x and n_y by integration on continuous variables:

$$E'_{\text{ZPE}}=\hbar cL_xL_y\int\limits_{-\infty}^{+\infty}\mathrm{d}n_x\int\limits_{-\infty}^{+\infty}\sum_{n_z=1}^{+\infty}\sqrt{\left(\frac{n_x2\pi}{L_x}\right)^2+\left(\frac{n_y2\pi}{L_x}\right)^2+\left(\frac{n_z\pi}{d}\right)^2}y\mathrm{d}n_y \tag{11.2.701}$$

which causes no difference when the limit $L_x,L_y\rightarrow+\infty$ is taken. Changing variables from n_x,n_y to k_x,k_y, dividing the expression by L_xL_y and taking the limit of L_x and L_y tending to infinity, yields for the energy per unit area:

$$\frac{E'_{\text{ZPE}}}{S}=\hbar c\sum_{n=1}^{+\infty}\int\limits_{-\infty}^{+\infty}\frac{\mathrm{d}k_x}{2\pi}\int\limits_{-\infty}^{+\infty}\sqrt{k_x^2+k_y^2+\left(\frac{n\pi}{d}\right)^2}\frac{\mathrm{d}k_y}{2\pi} \tag{11.2.702}$$

Finally, one integrates over the angle of the wave vector in the $x-y$ plane. To do this we put (polar coordinates transformation):

$$k_x=k_\perp\cos(\theta)\qquad k_y=k_\perp\sin(\theta) \tag{11.2.703}$$

and let us recall that the Jacobian (see section Differential and Integral Calculus page 1528) is $k_\perp$.

Then:

$$\frac{E'_{\text{ZPE}}}{S}=\hbar c\sum_{n=1}^{+\infty}\int\limits_0^{+\infty}k_\perp\frac{\mathrm{d}k_\perp}{2\pi}\int\limits_0^{2\pi}\left[k_\perp^2\sin^2(\theta)+k_\perp^2\cos^2(\theta)+\left(\frac{n\pi}{d}\right)^2\right]^{1/2}\frac{\mathrm{d}\theta}{2\pi} \tag{11.2.704}$$

As there are no θ in the integrand. This simplifies to:

$$\frac{E'_{\text{ZPE}}}{S}=\frac{\hbar c}{2\pi}\sum_{n=1}^{+\infty}\int\limits_0^{+\infty}\sqrt{k_\perp^2+\left(\frac{n\pi}{d}\right)^2}k_\perp\mathrm{d}k_\perp \tag{11.2.705}$$

We introduce now the auxiliary variable:

$$u=\frac{\left(k_x^2+k_y^2\right)d^2}{\pi^2}=\frac{k_\perp^2d^2}{\pi^2}\Rightarrow\mathrm{d}k_\perp=\frac{\pi^2}{2k_\perp d^2}\mathrm{d}u \tag{11.2.706}$$

and get:

$$\frac{E'_{\text{ZPE}}}{S}=\frac{\hbar c}{2\pi}\sum_{n=1}^{+\infty}\int\limits_{0}^{+\infty}\sqrt{\frac{u\pi^2}{d^2}+\left(\frac{n\pi}{d}\right)^2}\frac{\pi^2}{2d^2}\mathrm{d}u=\frac{\hbar c\pi^2}{4d^3}\sum_{n=1}^{+\infty}\int\limits_{0}^{+\infty}\left(u+n^2\right)^{1/2}\mathrm{d}u \tag{11.2.707}$$

We still have to add the contribution of the $k_z=0$ (or $n=0$) modes. One has:

$$\frac{E_{\text{ZPE}}}{S}=\frac{\hbar c\pi^2}{4d^3}\left[\sum_{n=1}^{+\infty}\int\limits_{0}^{+\infty}\left(u+n^2\right)^{1/2}\mathrm{d}u+\frac{1}{2}\int\limits_{0}^{+\infty}\sqrt{u}\mathrm{d}u\right] \tag{11.2.708}$$

This expression is divergent, as is the similar expression for the energy of the free field. It may be hoped that the difference between the two expressions is finite.

We need the value of the energy of the free field in the volume of the cavity. In general, the energy per unit volume is given by[17] (remember that $k=E/\hbar c$ and that the units of k are $[\mathrm{m}^{-1}]$):

$$\begin{aligned}\frac{E_{\text{free}}}{V}&=\frac{1}{L^2d}\sum_{n_x=1}^{+\infty}\sum_{n_y=1}^{+\infty}\sum_{n_z=1}^{+\infty}E(n_x,n_y,n_z)=\frac{1}{L^2d}\int\limits_{0}^{+\infty}\int\limits_{0}^{+\infty}\int\limits_{0}^{+\infty}E(n_x,n_y,n_z)\mathrm{d}n_x\mathrm{d}n_y\mathrm{d}n_z\\&=\frac{1}{L^2d}\int\limits_{0}^{+\infty}\int\limits_{0}^{+\infty}\int\limits_{0}^{+\infty}\hbar ck\mathrm{d}\left(\frac{k_xL}{2\pi}\right)\mathrm{d}\left(\frac{k_yL}{2\pi}\right)\mathrm{d}\left(\frac{k_zL}{2\pi}\right)=\hbar c\int\limits_{0}^{+\infty}\int\limits_{0}^{+\infty}\int\limits_{0}^{+\infty}\frac{k}{(2\pi)^3}\mathrm{d}k_x\mathrm{d}k_y\mathrm{d}k_z\\&=\hbar c\int\frac{k\mathrm{d}^3\vec{k}}{(2\pi)^3}\end{aligned} \tag{11.2.709}$$

It is advantageous to rewrite this expression as:

$$\frac{E_{\text{free}}}{V}=\frac{\hbar c}{(2\pi)^3}\int\mathrm{d}^2\vec{k}_{\perp}\int\sqrt{k_{\perp}^2+k_z^2}\mathrm{d}k_z \tag{11.2.710}$$

or as:

$$\frac{E_{\text{free}}}{V}=\frac{\hbar c}{(2\pi)^2}\int\limits_{0}^{+\infty}k_{\perp}\mathrm{d}k_{\perp}\int\limits_{-\infty}^{+\infty}\sqrt{k_{\perp}^2+k_z^2}\mathrm{d}k_z \tag{11.2.711}$$

Taking account of the fact that the integrand is an even function of k_z and introducing the auxiliary variables $u=k_{\perp}^2d^2/\pi^2$ and $x=k_z/d$ lead to:

$$\frac{E_{\text{free}}}{V}=\frac{\hbar c\pi^2}{4d^4}\int\limits_{0}^{+\infty}\mathrm{d}x\int\limits_{0}^{+\infty}\sqrt{u+x^2}\mathrm{d}u \tag{11.2.712}$$

The energy of the free field in the volume of the cavity is thus obtained by multiplying this expression by the volume V (equal to $S\cdot d$). One then obtains for the change in the zero-point energy per unit surface:

$$\frac{\Delta E_{\text{ZPE}}}{S}=\frac{E_{\text{cav}}}{S}-\frac{E_{\text{free}}}{S}=\frac{\hbar c\pi^2}{4d^3}\left\{\frac{1}{2}\int\limits_{0}^{+\infty}\sqrt{u}\mathrm{d}u+\sum_{n=1}^{+\infty}\int\limits_{0}^{+\infty}\left(u+n^2\right)^{1/2}\mathrm{d}u-\int\limits_{0}^{+\infty}\mathrm{d}x\int\limits_{0}^{+\infty}\left(u+x^2\right)^{1/2}\mathrm{d}u\right\} \tag{11.2.713}$$

---

[17] The $\vec{k}=\vec{0}$ mode does not pose any worry, since its contribution is vanishing.

All terms in the right hand side are divergent! We will come to this problem soon more in details.

Notice that it is remarkable that this expression involves the difference between the integral from zero to infinity of the function $f(x)=\int\limits_0^{+\infty}(u+x^2)^{1/2}\,\mathrm{d}u$ and the sum of the values of this function on the positive integers. As we know, there is a famous theorem by Euler and Maclaurin connecting these quantities, in general (see section Sequences and Series page 1674). It states that:

$$\begin{aligned}\sum_{n=1}^{+\infty} f(n)&=\int\limits_0^{+\infty} f(x)\mathrm{d}x-\frac{1}{2}[f(0)+f(\infty)]+\frac{1}{6\cdot 2!}\left[f'(\infty)-f'(0)\right]-\frac{1}{30\cdot 4!}\left[f'''(\infty)-f'''(0)\right]+\ldots\\&=\int\limits_0^{+\infty} f(x)\mathrm{d}x-\frac{1}{2}[f(0)+f(\infty)]+\frac{1}{12}\left[f'(\infty)-f'(0)\right]-\frac{1}{720}\left[f'''(\infty)-f'''(0)\right]+\ldots\end{aligned}\tag{11.2.714}$$

where the dots indicates similar terms for higher order odd derivatives. The coefficients in front of the brackets are related to the Bernoulli numbers B_i : they are equal to $-B_{2k}/(2k)!$ for the term involving the derivatives of order $2k-1$.

We can write the function $f(x)$ mentioned above as (replacing $u+x^2=t$):

$$f(x)=\int\limits_0^{+\infty}\left(u+x^2\right)^{1/2}\mathrm{d}u=\int\limits_{x^2}^{+\infty}\sqrt{t}\mathrm{d}t=\frac{2}{3}t^{3/2}\Big|_{x^2}^{+\infty}\tag{11.2.715}$$

Formally, considering the dependence upon x through the lower bound of the integral only (!), one has:

$$f'(x)=-2x^2,\ f''(x)=-4x,\ f'''(x)=-4\tag{11.2.716}$$

and all higher derivatives are vanishing. So, assuming (for the moment without arguing why!) retaining the single nondivergent term (which corresponds to $f'''(0)$), so that:

$$\sum_{n=1}^{+\infty} f(n)=\int\limits_0^{+\infty} f(x)\mathrm{d}x+\frac{1}{720}f'''(0)+\ldots=\int\limits_0^{+\infty} f(x)\mathrm{d}x-\frac{4}{720}\tag{11.2.717}$$

We finally get (ignoring again without arguing for now the divergent terms!):

$$\frac{\Delta E_{\text{ZPE}}}{S}=\frac{E_{\text{cav}}}{S}-\frac{E_{\text{free}}}{S}=\frac{\hbar c\pi^2}{4d^3}\left(-\frac{4}{720}\right)\tag{11.2.718}$$

Therefore:

$$\frac{\Delta E_{\text{ZPE}}(d)}{S}=-\frac{\hbar c\pi^2}{720d^3}\tag{11.2.719}$$

The force (per unit surface) acting between the plates is then given by:

$$\boxed{\frac{F_{\text{ZPE}}^{\text{Casimir}}(d)}{S}=\frac{\mathrm{d}(\Delta E_{\text{ZPE}/S}(d))}{\mathrm{d}d}=-\frac{\hbar c\pi^2}{240d^4}}\tag{11.2.720}$$

This is the expression of the 3D electromagnetic Casimir force.

Let us comment now the divergence problems. Of course, the expression:

$$\frac{\Delta E_{\text{ZPE}}}{S}=\frac{E_{\text{cav}}}{S}-\frac{E_{\text{free}}}{S}=\frac{\hbar c\pi^2}{4d^3}\left\{\frac{1}{2}\int\limits_0^{+\infty}\sqrt{u}\,\mathrm{d}u+\sum_{n=1}^{+\infty}\int\limits_0^{+\infty}\left(u+n^2\right)^{1/2}\mathrm{d}u-\int\limits_0^{+\infty}\mathrm{d}x\int\limits_0^{+\infty}\left(u+x^2\right)^{1/2}\mathrm{d}u\right\}\tag{11.2.721}$$

and the Euler-Maclaurin theorem apply when the quantities are convergent. However, one can make the final result meaningful by regularizing the integral and the sum. The regularisation at infinity is not a problem. It is easy to introduce a suitable integration factor. For instance, it is easy to see that all terms at infinity vanish owing to the substitution:

$$f(x)\rightarrow\int\limits_0^{+\infty}\left(u+x^2\right)^{1/2}e^{-\left(u+x^2\right)\alpha}\mathrm{d}u\tag{11.2.722}$$

The terms corresponding to higher order derivatives (at $x=0$ as well as at $x=+\infty$) are finite and vanish as $\alpha\rightarrow 0$. Similarly, the first and third derivatives at $x=0$ are incremented by quantities that vanishes as $\alpha\rightarrow 0$. Actually, the regularisation can be achieved by any cutoff function which decreases sufficiently rapidly at large x (non necessarily exponentially), which goes to unity as x goes to zero with vanishing derivatives, like the functions $g(x)=1-\exp(-a/x)$.

Remark

As an example, we chose the regularisation provided by substitution of the above or:

$$f(x)=\int\limits_{x^2}^{+\infty}\sqrt{t}e^{-\alpha t}\mathrm{d}t\tag{11.2.723}$$

The first three derivatives are given by:

$$f'(x)=-2x^2e^{-\alpha x^2}\tag{11.2.724}$$

and by:

$$f''(x)=\left(-4x+4\alpha x^3\right)e^{-\alpha x^2}\tag{11.2.725}$$

and:

$$f'''(x)=\left(-4+20\alpha x^2-8\alpha^2x^4\right)e^{-\alpha x^2}\tag{11.2.726}$$

respectively. All derivatives at infinity are vanishing. It is also easy to verify that $f'(x)$ vanishes at $x=0$ and that $f'''(0)=-4$ and that higher order derivatives at $x=0$ are either vanishing or are proportional to positive powers of α. The first term in the curly bracket of:

$$E'_{\text{cav}}=\hbar c\sum_{n_x=-\infty}^{+\infty}\sum_{n_y=-\infty}^{+\infty}\sum_{n_z=1}^{+\infty}\left[\left(\frac{n_x2\pi}{L_x}\right)^2+\left(\frac{n_y2\pi}{L_x}\right)^2+\left(\frac{n_z\pi}{d}\right)^2\right]^{1/2}\tag{11.2.727}$$

being equal to $f(0)$, formula:

$$\frac{\Delta E_{\text{ZPE}}}{S}=-\frac{\hbar c\pi^2}{720d^3} \qquad (11.2.728)$$

results directly from the application of the Euler-Maclaurin formula to the function $f(x)=\int\limits_{x^2}^{+\infty}\sqrt{t}e^{-\alpha t}\mathrm{d}t$ at the limit of vanishing α.

For many years the Casimir effect was little more than a theoretical curiosity. But interest in the phenomenon has blossomed in the early 121st century (holocene calendar). Experimental physicists have realized that the Casimir force affects the workings of micromachined devices, while advances in instrumentation have enabled the force to be measured with ever-greater accuracy.

The new enthusiasm has also been fired by fundamental physics. Many theorists have predicted the existence of “large” extra dimensions in 10- and 11-dimensional unified field theories of the fundamental forces. These dimensions, they say, could modify classical Newtonian gravitation at sub-millimetre distances. Measuring the Casimir effect could therefore help physicists to test the validity of such radical ideas.

Remark

The fact that an attractive force exists between two conducting metal plates was first predicted in 11948 (holocene calendar) by Hendrik Casimir of Philips Research Laboratories in the Netherlands. At the time, however, Casimir was studying the properties of "colloidal solutions". These are viscous materials, such as paint and mayonnaise, that contain micron-sized particles in a liquid matrix. The properties of such solutions are determined by van der Waals forces – long-range, attractive forces that exist between neutral atoms and molecules.

One of Casimir’s colleagues, Theo Overbeek, realized that the theory that was used at the time to explain van der Waals forces, which had been developed by Fritz London in 11932 (holocene calendar), did not properly explain the experimental measurements on colloids. Overbeek therefore asked Casimir to investigate the problem. Working with Dirk Polder, Casimir discovered that the interaction between two neutral molecules could be correctly described only if the fact that light travels at a finite speed was taken into account.

Soon afterwards, Casimir noticed that this result could be interpreted in terms of vacuum fluctuations. He then asked himself what would happen if there were two mirrors – rather than two molecules – facing each other in a vacuum. It was this work that led to his famous prediction of an attractive force between reflecting plates.

Although the Casimir force seems completely counter-intuitive, it is actually well understood. In the old days of classical mechanics the idea of a vacuum was simple. The vacuum was what remained if you emptied a container of all its particles and lowered the temperature down to absolute zero. The arrival of quantum mechanics, however, completely changed our notion of a vacuum. All fields – in particular electromagnetic fields – have fluctuations. In other words at any given moment their actual value varies around a constant, mean value. Even a perfect vacuum at absolute zero has fluctuating fields known as "vacuum fluctuations", the mean energy of which corresponds to half the energy of a photon.

However, vacuum fluctuations are not some abstraction of a physicist's mind. They have observable consequences that can be directly visualized in experiments on a microscopic scale. For example, an atom in an excited state will not remain there infinitely long, but will return to its ground state by spontaneously emitting a photon. This phenomenon is a consequence of vacuum fluctuations. Imagine trying to hold a pencil upright on the end of your finger. It will stay there if your hand is perfectly stable and nothing perturbs the equilibrium. But the slightest perturbation will make the pencil fall into a more stable equilibrium position. Similarly, vacuum fluctuations cause an excited atom to fall into its ground state.

An important physical quantity when discussing the Casimir force is the "field radiation pressure". Every field – even the vacuum field – carries energy. As all electromagnetic fields can propagate in space they also exert pressure on surfaces, just as a flowing river pushes on a floodgate. This radiation pressure increases with the energy – and hence the frequency – of the electromagnetic field. At a cavity-resonance frequency the radiation pressure inside the cavity is stronger than outside and the mirrors are therefore pushed apart. Out of resonance, in contrast, the radiation pressure inside the cavity is smaller than outside and the mirrors are drawn towards each other.

While the Casimir force is too small to be observed for mirrors that are several metres apart, it can be measured if the mirrors are within microns of each other. For example, two mirrors with an area of $1\;[\text{cm}^2]$ separated by a distance of $1\;[\mu\text{m}]$ have an attractive Casimir force of about 10^{-7} [N] – roughly the weight of a water droplet that is half a millimetre in diameter. Although this force might appear small, at distances below a micrometre the Casimir force becomes the strongest force between two neutral objects. Indeed at separations of $10\;[\text{nm}]$ – about a hundred times the typical size of an atom – the Casimir effect produces the equivalent of 1 atmosphere of pressure!

Although we do not deal directly with such small distances in our everyday lives, they are important in nanoscale structures and microelectromechanical systems (MEMS). These are "intelligent" micron-sized devices in which mechanical elements and moving parts, such as tiny sensors and actuators, are carved into a silicon substrate. Electronic components are then wired on to the device to process information that it senses or to drive the movement of its mechanical parts. MEMS have many possible applications in science and engineering, and are already used as car air-bag pressure sensors.

As MEMS devices are fabricated on the micron and submicron scale, the Casimir force can cause the tiny elements in a device to stick together – as reported in 12001 (holocene calendar) by Michael Roukes and co-workers at the California Institute of Technology (see [80]).

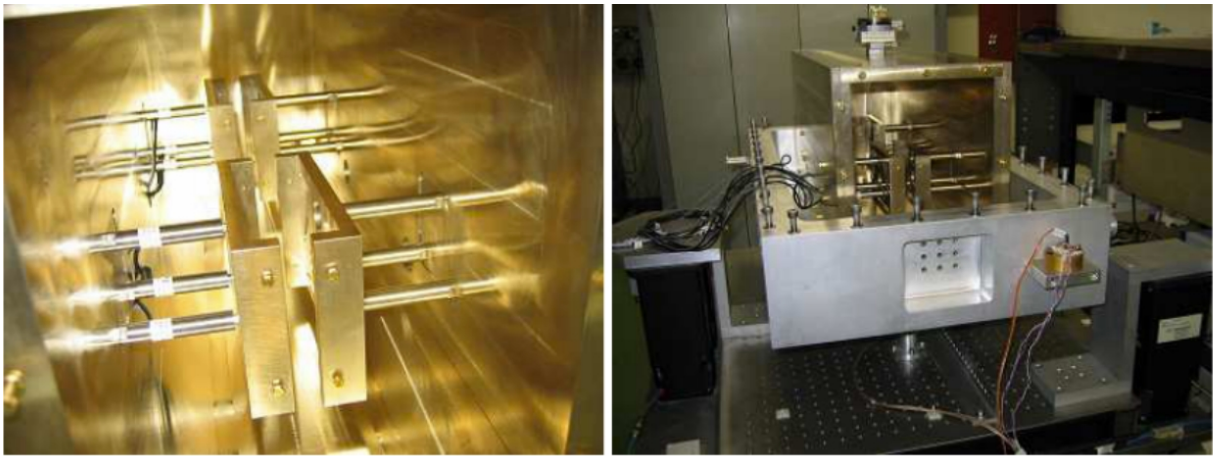
When the Casimir effect was first predicted in 11948 (holocene calendar) it was very difficult to measure using the equipment of the time. One of the first experiments was carried out in 11958 (holocene calendar) by Marcus Spaarnay at Philips in Eindhoven, who investigated the Casimir force between two flat, metallic mirrors made from either aluminium, chromium or steel. Spaarnay measured the force using a spring balance, the extension of which was determined by the capacitance of the two plates. To prevent the Casimir force from being swamped by the electrostatic force, the mirrors had to be kept neutral by first touching them together before each measurement was made. Spaarnay also had to ensure that the plane mirrors were exactly parallel to each other, as the Casimir force is very sensitive to changes in distance. Spaarnay managed to overcome these difficulties and concluded that his results « *did not contradict Casimir's theoretical prediction*».

Since those early days, however, sophisticated equipment has made it much easier to study the Casimir effect (see [63]). A new generation of measurements began in 11997 (holocene calendar). Steve Lamoreaux, who was then at the University of Washington in Seattle, measured the Casimir force between a 4 [cm] diameter spherical lens and an optical quartz plate about 2.5 [cm] across, both of which were coated with copper and gold. The lens and plate were connected to a torsion pendulum – a twisting horizontal bar suspended by a tungsten wire – placed in a cylindrical vessel under vacuum. When Lamoreaux brought the lens and plate together to within several microns of each other, the Casimir force pulled the two objects together and caused the pendulum to twist. He found that his experimental measurements agreed with theory to an accuracy of 5%.

Inspired by Lamoreaux's breakthrough, many other researchers tried new Casimir measurements. Umar Mohideen and co-workers at the University of California at Riverside, for example, attached a polystyrene sphere 200 [μm] in diameter to the tip of atomic force microscope. In a series of experiments they brought the sphere, which was coated with either aluminium or gold, to within 0.1 [μm] of a flat disk, which was also coated with these metals. The resulting attraction between the sphere and the disk was monitored by the deviation of a LASER beam. The researchers were able to measure the Casimir force to within 1% of the expected theoretical value.

Thomas Ederth at the Royal Institute of Technology in Stockholm, Sweden, has also used an atomic force microscope to study the Casimir effect. He measured the force between two gold-coated cylinders that were arranged at $90°$ to each other and that were as little as 20 [nm] apart. His results agreed to within 1% of theory.

In the photo below we can see the experimental set-up for the measurement of the Casimir force in a parallel plane configuration. On the left, the plateholders are connected to the external frame through three precision pieoelectric positioners each plate. On the right is anoverall view of the vacuum chamber with the fiber holder on the top (see [78]):



However, very few recent experiments have measured the Casimir force using the original configuration of two plane, parallel mirrors. The reason is that the mirrors have to be kept perfectly parallel during the experiment, which is difficult. It is much easier to bring a sphere close up to a mirror because the separation between the two objects is simply the distance of closest approach. The only drawback of using a sphere and a plane mirror is that the calculations of the Casimir force are not as accurate as those between two plane mirrors. In particular it has to be assumed that the contributions to the force between the sphere and plate are completely independent at each point. This is true only if the radius of the sphere is much larger than the distance between it and the plate.

The only recent experiment to replicate Casimir's original set-up of two plane, parallel mirrors was carried out by Gianni Carugno, Roberto Onofrio and co-workers at the University of Padova in Italy. They measured the force between a rigid chromium-coated plate and the flat surface of a cantilever made from the same material that were separated by distances ranging from 0.5-3 [μm] (see [77]). The researchers found that the measured Casimir force agreed to within 15% of the expected theoretical value. This relatively poor fit reflected the technical difficulties involved in the experiment.

The problem with studying the Casimir effect is that real mirrors are not like the perfectly smooth plane mirrors that Hendrik Casimir originally considered. In particular, real mirrors do not reflect all frequencies perfectly. They reflect some frequencies well – or even nearly perfectly – while others are reflected badly. In addition, all mirrors become transparent at very high frequencies. When calculating the Casimir force the frequency-dependent reflection coefficients of the mirrors have to be taken into account – a problem first tackled by Evgeny Lifshitz in the mid-11950s (holocene calendar), and then by Julian Schwinger and many others.

It turns out that the measured Casimir force between real metallic mirrors that are 0.1 [μm] apart is only half the theoretical value predicted for perfect mirrors.

Another problem with calculating the expected Casimir force for a real system is the fact that experiments are never carried out at absolute zero – as originally envisaged in Casimir's calculations – but at room temperature. This causes thermal – as well as vacuum – fluctuations to come into play. These thermal fluctuations can produce their own radiation pressure and generate a bigger Casimir force than expected. For example, the Casimir force between two plane mirrors 7 [μm] apart is twice as large at room temperature than at absolute zero. Fortunately, thermal fluctuations at room temperature are only important at distances above 1 [μm], below which the wavelength of the fluctuations is too big to fit inside the cavity.

Although the temperature dependence of the Casimir force has not yet been studied in detail experimentally at the time we write those lines, it must be included in calculations of the force at separations above 1 [μm]. Many researchers have tackled this problem for perfectly reflecting mirrors, including Lifshitz and Schwinger back in the 11950s (holocene calendar). It has also been examined more recently by Michael Bordag at Leipzig University, Bo Sernelius at Linköping University in Sweden, Galina Klimchitskaya and Vladimir Mostepanenko at the University of Paraiba in Brazil, and by our group in Paris. Indeed the temperature dependence of the Casimir force was for some time a matter of hot debate in the community. The various contradictions, however, now seem to have been resolved, and this has given an additional motivation to an experimental observation of the influence of temperature on the Casimir force.

A third and final problem in calculating the Casimir force is that real mirrors are not perfectly smooth. Most mirrors are made by coating a substrate with a thin metal film using the technique of "sputtering". However, this produces films with a roughness of about 50 [nm]. While such roughness is invisible to the naked eye, it does affect measurements of the Casimir force, which is very sensitive to small changes in distance.

Mohideen and his group in California have used surface deformations to show that two surfaces can also have a lateral Casimir force that acts in a parallel – rather than a perpendicular – direction to the surface of the mirrors. In the experiments they prepared specially corrugated mirrors the surfaces of which were sinusoidally curved. They then moved the mirrors parallel to one another so that a peak of one mirror passed successively over the peaks and troughs of the other mirror. The researchers found that the lateral Casimir force varied sinusoidally with the phase difference between the two corrugations. The size of the force was about ten times smaller than the ordinary Casimir force between two mirrors the same distance apart. The lateral force is also due to vacuum fluctuations.

To summarize there are plenty of independent experiments that confirms the (static) Casimir effect (SCE) theoretical prediction - emergence of particles from zero point vacuum energy state - using different methods with accuracies between 1% and 15% and at the day we write these lines there are even two experiments that were able to detect photons directly from the dynamic Casimir effect (DCE):

Team	**Method**	**Place**	**Year** (holocene calendar)	**Type**
Marcus Spaarnay	Metallic mirrors	Eindhoven (Netherlands)	11958	SCE
Steve Lamoreaux	Spherical lens	Washington University (Seattle/USA)	11997	SCE
Michael Roukes and Al.	MEMS	California Institute of Technology (USA)	12001	SCE
Frederico Capasso	MEMS	New Jersey (USA)	12001	SCE
Umar Mohideen and Al.	Polystyrene sphere	University of California (USA)	11998	SCE
Thomas Ederth	Gold-coated-cylinders	Royal Institute of Technology (Stockholm)	12000	SCE
G. Carguno, R. Onofrio and Al.	Chromium-coated plates	University of Padova (Italy)	12002	SCE
Jens Gundlach and Al.	Torsion pendulum	Washington (USA)	12000	SCE
G. Bressi and Al.	Fractalized metallic plates	Instituto Nazionale di Fisica Nucleara (Italy)	12002	SCE
Riccardo Decca and Al.	Torsional oscillator	University Purdure (Indinia/USA)	12006	SCE
G. L. Klimichitskaya	Graphene systems	St. Petersburg (Russia)	12022	SCE
Benjamin Spreng and Al.	Biased seminconductors	University of California (USA)	12024	SCE
Pasi Lähteenmäki and Al.	Josephson metamaterial	Finland	12011	DCE
C.M. Wilson and Al.	Superconducting circuit	University of Technology (Göteborg, Sweden)	12011	DCE
...	...	...	...	

Table 11.9: **Most well known variations of the Casimir effect experiment and others**

For people interested to go further into the subject (multiple ways to resolve the problem, various configurations and assumptions, taking into account more realistic configurations, etc.), we recommend the lecture of [395].

11.2.4.4.9 Tunnel Effect (barrier potential / radioactivity)

The "tunnel effect" or "quantum tunneling" refers to the property of a quantum object to cross a potential barrier, which can not be explained by Classical Mechanics. Generally, the wave function of a particle, whose square of the module represents the amplitude of its probability of presence, does not cancel at the level of the barrier but attenuates within the barrier (see the previous plot of the harmonic oscillator!), practically exponentially for a fairly wide barrier as we will prove it soon. If at the output of the potential barrier the particle possesses a non-zero probability of presence, it can thus cross this barrier.

The theoretical study of this phenomenon is of crucial importance in the theory of semiconductors (see section Electrokinetics), disintegration/fission in nuclear physics (see section Nuclear Physics page 5075) and also nuclear fusion (see section Nuclear Physics page 5064). It is therefore necessary to pay special attention to it!

The quantum barrier of width L separates in simple cases the space in three (in the 1 dimensional case), the left and right parts of which are considered to have constant potentials up to infinity. The intermediate part constitutes the barrier, which may be complicated, revealing a soft profile, or on the contrary formed by rectangular or possibly other type of barriers/wells.

Let us now consider the case of systems where the potential energy E_p (implicitly the relative potential) tends to finite boundaries, not necessarily equal when $x \rightarrow \pm\infty$. It is therefore a problem of unbounded states.

First, we define a region `I` far to the left where $E_p(x)$ will be denoted:

$$E_p(-\infty)=E_p^{\mathtt{I}} \qquad (11.2.729)$$

A region `III` far to the right where $E_p(x)$ will be denoted:

$$E_p(+\infty)=E_p^{\mathtt{III}} \qquad (11.2.730)$$

By limiting ourselves to the simplest situations, there are three possibilities relative to the relations given above:

- potential well (a)
- potential step (b)
- potential barrier (c)

as represented in the same order in the figure below:

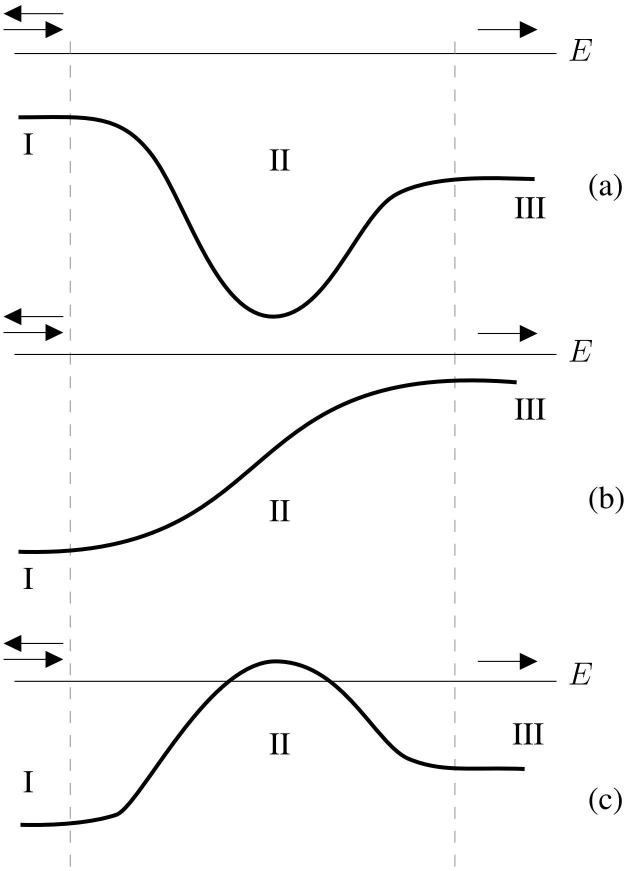


Figure 11.56: **Classical types of potential variations for Tunnel Effect study**

Now, let us write Schrödinger's equation:

$$\Delta\Psi+\frac{2m}{\hbar^2}(E_{\text{tot}}-E_p)\Psi=0 \tag{11.2.731}$$

In the regions $\mathtt{I}$ and $\mathtt{III}$ of the potential barrier, the idea is that $E_{\text{tot}}-E_p$ is constant and positive so the differential equation can be written in one dimension:

$$\frac{\partial^2\Psi}{\partial x^2}+\frac{2m}{\hbar^2}(E_{\text{tot}}-E_p)\Psi=\frac{\partial^2\Psi}{\partial x^2}+k^2\Psi=0 \tag{11.2.732}$$

We thus get very simply the analytic expression of Ψ in these regions in general form:

$$\begin{aligned}\Psi_{\mathtt{I}}&=Ae^{\mathrm{i}k_{\mathtt{I}}x}+Be^{-\mathrm{i}k_{\mathtt{I}}x}\\ \Psi_{\mathtt{III}}&=A'e^{\mathrm{i}k_{\mathtt{III}}x}+B'e^{-\mathrm{i}k_{\mathtt{III}}x}\end{aligned} \tag{11.2.733}$$

We fall back on these two expressions that are identical to those of our study of the potential well with rectangular walls, with the difference that we have written more general solutions of the differential equation (see section Differential and Integral Calculus page 1610) without having determine the coefficients (because we are interested here in a generalization).

Thus, in our study of the well with rectangular walls earlier above we had already determined that:

$$k_{\mathtt{I}}=\sqrt{\frac{8\pi^2m}{h^2}(E_{\text{tot}}-E_p^{\mathtt{I}})}\quad\text{and}\quad k_{\mathtt{III}}=\sqrt{\frac{8\pi^2m}{h^2}(E_{\text{tot}}-E_p^{\mathtt{III}})} \tag{11.2.734}$$

Remarks

R1. We see that the wave numbers k are therefore proportional to the root of the kinetic energy. And since the kinetic energy is proportional to the square velocity of the particles, then the velocity is proportional to the wave number (and vice versa)!

R2. In some textbooks, in order to simplify the notations, the potential in regions `I` and `III` is assumed as a reference and therefore equalized to 0. The term E_p therefore disappears from the two preceding expressions and this has the effect of equalizing the both wave numbers $k_{\mathtt{I}}$, $k_{\mathtt{III}}$ which are then simply denoted k.

In region `II`, the idea is that $E_{\text{tot}}-E_p$ is negative and constant (yes that means indeed we assume the corresponding moment to be mathematically an imaginary number[18]!) so the differential equation can be written in one dimension:

$$\frac{\partial^2\Psi}{\partial x^2}+\frac{2m}{\hbar^2}(E_{\text{tot}}-E_p)\Psi=\frac{\partial^2\Psi}{\partial x^2}-\frac{2m}{\hbar^2}(E_p-E_{\text{tot}})\Psi=\frac{\partial^2\Psi}{\partial x^2}-K^2\Psi=0 \qquad (11.2.735)$$

and as we have seen it in our study of the infinite rectangular potential well according to the second approach, the solution is then of the form:

$$\Psi_{\mathtt{II}}=Ce^{Kx}+De^{-Kx} \qquad (11.2.736)$$

with:

$$K=\frac{\sqrt{2m(E_p-E_{\text{tot}})}}{\hbar} \qquad (11.2.737)$$

Remark

The parenthesis under the root of the preceding relation must therefore be positive. This would mean that the kinetic energy of the particle is negative $E_c<0$ (and therefore the square of the momentum of the particle $p^2=2mE_c$ is also negative, then the momentum $p=\sqrt{p^2}$ is mathematically an imaginary number). To overcome this problem within the framework of this simplified model, it is said that the particle has no right to exist in the barrier and that the particle borrow of energy from vacuum (the particle is thus termed "unphysical" because it does not have the normal mathematics of a measurable particle; but we still can describe it mathematically). But there are other more complex models that do not require this kind of "fantasies".

We thus get very simply the analytic expression of Ψ in the three regions in general form:

$$\begin{aligned}\Psi_{\mathtt{I}}&=Ae^{\mathrm{i}k_{\mathtt{I}}x}+Be^{-\mathrm{i}k_{\mathtt{I}}x}\\ \Psi_{\mathtt{II}}&=Ce^{Kx}+De^{-Kx}\\ \Psi_{\mathtt{III}}&=A'e^{\mathrm{i}k_{\mathtt{III}}x}+B'e^{-\mathrm{i}k_{\mathtt{III}}x}\end{aligned} \qquad (11.2.738)$$

Let us suppose now that we have at $-\infty$ (region `I`), a source of particles (which sends them to the right), with a kinetic energy obviously equal to $E_{\text{tot}}-E_p^{\mathtt{I}}$.

---

[18] Inside the barrier, the kinetic energy of the particle, $E_c=E_{\text{tot}}-E_p$, is negative and therefore the square of the momentum of the particle, $p^2=2mE_c$, is also negative. The momentum $p=\sqrt{p^2}$ is therefore mathematically an imaginary number, which is a number containing a factor of $\sqrt{-1}$.

Thus, these particles have an energy E_{tot} and the wave function that describes them obeys the Schrödinger equation. In Region `III`, it will be assumed that there can be only particles going to the right (no source at $+\infty$ by hypothesis).

The region `III`, as well as region `I`, is of infinite extension, so the principle of uncertainty allows us to speak in theory of a perfectly determined linear momentum which we denote by p'.

We know that (this is Classical Mechanics!) in the region `III` we then have:

$$\frac{p'^2}{2m}=E_{\text{tot}}-E_p^{\texttt{III}} \tag{11.2.739}$$

If $E_{\text{tot}}>E_p^{\texttt{III}}$ then p' is positive, so thanks to the preceding relation and to the de Broglie relation we have:

$$p'=\sqrt{2m\left(E_{\text{tot}}-E_p^{\texttt{III}}\right)}=\frac{hk_{\texttt{III}}}{2\pi} \tag{11.2.740}$$

Therefore:

$$k_{\texttt{III}}=\frac{\sqrt{2m\left(E_{\text{tot}}-E_p^{\texttt{III}}\right)}}{\hbar}\qquad k_{\texttt{I}}=\frac{\sqrt{2m\left(E_{\text{tot}}-E_p^{\texttt{I}}\right)}}{\hbar} \tag{11.2.741}$$

Since the wave numbers are now known formally, let us return to the interpretation of solution `III`:

$$\Psi_{\texttt{III}}=A'e^{\text{i}k_{\texttt{III}}x}+B'e^{-\text{i}k_{\texttt{III}}x} \tag{11.2.742}$$

The assumption that the particles come from the left imposes us that $B'=0$ so that the solution only describes particles that go to the right. Then, it is possible, for those coming from the left, to take $A'=1$. The region `III` is therefore relatively simple to analyse...

Remark

The conditions and assumptions used previously are sometimes referred to as "scattering conditions".

The constants A and B of the region `I` will be completely determined by carrying out the connection of the solutions from one region to the other.

Let us now turn to the interpretation of the equation in Region `I`:

$$\Psi_{\texttt{I}}=A'e^{\text{i}k_{\texttt{I}}x}+B'e^{-\text{i}k_{\texttt{I}}x} \tag{11.2.743}$$

It is evident that $Ae^{\text{i}k_{\texttt{I}}}$ describes particles which, in region `I` are directed to the right, when $Be^{-\text{i}k_{\texttt{I}}}$ describes particles which in this same region are directed to the left. As we know, the first are the incident particles, the second are the reflected particles.

What we ask of Quantum Physics now appears in a clear way: a particle arriving from the left (incident) can either:

1. Continue to the right, that is, cross the region `II` and become a transmitted particle

2. Go back to the left and become a reflected particle.

We are led to define a "transmission coefficient T" assimilated to the probability that the incident particle has to cross the region `II` and a "reflection coefficient R", the probability that the incident particle will be reflected. We must obviously have:

$$R+T=1 \tag{11.2.744}$$

In the case of a potential barrier, T is also named the "barrier transparency".

To calculate R and T, we will define the current fluxes of the various categories of particles (incident, transmitted, reflected).

For example, since the incident particles are described by $A^{\mathrm{i}k_{\mathtt{I}}x}$, the average number of these particles per unit length in region `I` must certainly be proportional to a given factor with $|A|^2$.

Let $v_{\mathtt{I}}$ be their velocity, we see that the flow of the incident particles denoted j_i, is then proportional to a given factor to $|A|^2 v_{\mathtt{I}}$ (through dimensional analysis). Thus, the coefficient of proportionality being of the same nature for the three categories of particles (incident i, reflected j, transmitted t) and from the fact that $v_{\mathtt{I}}$ and $v_{\mathtt{III}}$ are proportional to $k_{\mathtt{I}}$ and $k_{\mathtt{III}}$, it follows that j_i, j_r (incident flow and reflected flow) and j_t (transmitted flow) are respectively proportional (hence always to a given dimensional factor) to $k_{\mathtt{I}}|A|^2$, $k_{\mathtt{I}}|B|^2$, and $k_{\mathtt{III}}$ (since we recall that for region `III` we have found $A'=1$ and $B'=0$).

We deduce from here simply, by a simple ratio, the expressions of the coefficients of reflection R and transmission T:

$$R=\frac{|B|^2}{|A|^2}=\left|\frac{B}{A}\right|^2 \qquad T=\frac{|A'|^2}{|A|^2}=\frac{k_{\mathtt{III}}}{k_{\mathtt{I}}}\frac{1}{|A|^2} \tag{11.2.745}$$

and as in our special case $k_{\mathtt{I}}=kk_{\mathtt{III}}$ it comes:

$$T=\frac{1}{|A|^2} \tag{11.2.746}$$

Another way of writing things is to say that since the incident wave is summarized to:

$$\Psi^{\mathtt{I}}=A^{\mathrm{i}kx} \tag{11.2.747}$$

and the transmitted wave to:

$$\Psi^{\mathtt{T}}=A'^{\mathrm{i}kx} \tag{11.2.748}$$

then:

$$\boxed{T=\frac{|\Psi^{\mathtt{T}}|^2}{|\Psi^{I}|^2}} \tag{11.2.749}$$

In all these situations, quantum theory leads in general to small but not zero values of R and T!

✍**Example:**

Let us determine the explicit expression of transparency for the example of our rectangular barrier.

For this we know that we must impose the continuity of Ψ in $x=0$ and $x=L$, as well as the continuity of $\mathrm{d}\Psi/\mathrm{d}x$ in $x=0$ and $x=L$.

Let us first recall that we have the three relations (by putting the reference of the potential at 0):

$$\begin{gathered}\Psi_{\mathrm{I}}=Ae^{\mathrm{i}k_{\mathrm{I}}x}+Be^{-\mathrm{i}k_{\mathrm{I}}x}\\ \Psi_{\mathrm{II}}=Ce^{Kx}+De^{-Kx}\\ \Psi_{\mathrm{III}}=A'e^{\mathrm{i}k_{\mathrm{III}}x}+B'e^{-\mathrm{i}k_{\mathrm{III}}x}\end{gathered}\tag{11.2.750}$$

with therefore:

$$k=\frac{\sqrt{2mE_{\text{tot}}}}{\hbar}\qquad\text{and}\qquad K=\frac{\sqrt{2m(E_{\text{tot}}-E_p)}}{\hbar}\tag{11.2.751}$$

We then have for the continuity of Ψ in $x=0$ and $x=L$:

$$\begin{gathered}A+B=C+D\\ Ce^{KL}+De^{-KL}=A'e^{\mathrm{i}kl}+B'e^{-\mathrm{i}kl}\end{gathered}\tag{11.2.752}$$

as well as the continuity of $\mathrm{d}\Psi/\mathrm{d}x$ in $x=0$ and $x=L$:

$$\begin{gathered}\mathrm{i}kA-\mathrm{i}kB=KC-KD\\ KCe^{KL}-KDe^{-KL}=ikA'e^{\mathrm{i}kL}+\mathrm{i}kB'e^{-\mathrm{i}kL}\end{gathered}\tag{11.2.753}$$

Since B' is zero we have a system of 4 equations with 5 unknowns:

$$\begin{cases}A+B=C+D\\ \mathrm{i}kA-\mathrm{i}kB=KC-KD\\ Ce^{KL}+De^{-KL}=A'e^{\mathrm{i}kl}\\ KCe^{KL}-KDe^{-KL}=\mathrm{i}kA'e^{\mathrm{i}kL}\end{cases}\Rightarrow\begin{cases}-B+C+D=A\\ \mathrm{i}kB+KC-KD=\mathrm{i}kA\\ Ce^{KL}+De^{-KL}-A'e^{\mathrm{i}kl}=0\\ KCe^{KL}-KDe^{-KL}-\mathrm{i}kA'e^{\mathrm{i}kL}=0\end{cases}\tag{11.2.754}$$

We will choose to express all the constants from A. To do this we multiply the first line by $\mathrm{i}l$ and we sum it to the second line. We then get:

$$\begin{array}{r}-\mathrm{i}kB+\mathrm{i}kC+\mathrm{i}kD=\mathrm{i}kA\\ +\quad \mathrm{i}kB+KC-KD=\mathrm{i}kA\\ \hline (K+\mathrm{i}k)C-(K-\mathrm{i}k)D=2\mathrm{i}kA\end{array}\tag{11.2.755}$$

and then we multiply the third line by $-\mathrm{i}k$ and the sum it to the fourth line. We have then:

$$\begin{array}{r}-\mathrm{i}kCe^{KL}-\mathrm{i}kDe^{-KL}+\mathrm{i}kA'e^{\mathrm{i}kl}=0\\ +\quad KCe^{KL}-KDe^{-KL}-\mathrm{i}kA'e^{\mathrm{i}kl}=0\\ \hline (K-\mathrm{i}k)e^{KL}C-(K+\mathrm{i}k)e^{-KL}D=0\end{array}\tag{11.2.756}$$

Therefore we have the two relations:

$$\begin{gathered}(K+\mathrm{i}k)C-(K-\mathrm{i}k)D=2\mathrm{i}kA\\(K-\mathrm{i}k)e^{KL}C-(K+\mathrm{i}k)e^{-KL}D=0\end{gathered}\tag{11.2.757}$$

or by putting $\alpha=k/K$:

$$\begin{gathered}(1+\mathrm{i}\alpha)C-(1-\mathrm{i}\alpha)D=2\mathrm{i}\alpha A\\(1-\mathrm{i}\alpha)e^{KL}C-(1+\mathrm{i}\alpha)e^{-KL}D=0\end{gathered}\tag{11.2.758}$$

From the second relation, it comes:

$$D=\frac{(1-\mathrm{i}\alpha)e^{KL}C}{(1+\mathrm{i}\alpha)e^{-KL}}\tag{11.2.759}$$

And injected into the first one:

$$(1+\mathrm{i}\alpha)C-(1-\mathrm{i}\alpha)\frac{(1-\mathrm{i}\alpha)e^{KL}C}{(1+\mathrm{i}\alpha)e^{-KL}}=2\mathrm{i}\alpha A\tag{11.2.760}$$

Therefore:

$$C\frac{(1+\mathrm{i}\alpha)^2e^{-KL}-(1-\mathrm{i}\alpha)^2e^{KL}}{(1+\mathrm{i}\alpha)e^{-KL}}=2\mathrm{i}\alpha A\tag{11.2.761}$$

Then we have:

$$C=\frac{2\mathrm{i}\alpha(1+\mathrm{i}\alpha)e^{-KL}}{(1+\mathrm{i}\alpha)^2e^{-KL}-(1-\mathrm{i}\alpha)^2e^{KL}}A\tag{11.2.762}$$

and if we put:

$$F=\frac{1}{2}\left((1+\mathrm{i}\alpha)^2e^{-KL}-(1-\mathrm{i}\alpha)^2e^{KL}\right)\tag{11.2.763}$$

Therefore it comes:

$$C=\frac{\mathrm{i}\alpha(1+\mathrm{i}\alpha)e^{-KL}}{F}A\tag{11.2.764}$$

Similarly, re-starting from:

$$\begin{gathered}(1+\mathrm{i}\alpha)C-(1-\mathrm{i}\alpha)D=2\mathrm{i}\alpha A\\(1-\mathrm{i}\alpha)e^{KL}C-(1+\mathrm{i}\alpha)e^{-KL}D=0\end{gathered}\tag{11.2.765}$$

From the second relation, it comes:

$$C=\frac{(1+\mathrm{i}\alpha)e^{-KL}D}{(1-\mathrm{i}\alpha)e^{KL}}\tag{11.2.766}$$

And injected into the first one:

$$(1+\mathrm{i}\alpha)\frac{(1+\mathrm{i}\alpha)e^{-KL}D}{(1-\mathrm{i}\alpha)e^{KL}}-(1-\mathrm{i}\alpha)D=2\mathrm{i}\alpha A\tag{11.2.767}$$

Therefore:

$$D=\frac{(1+\mathrm{i}\alpha)^2e^{-KL}-(1-\mathrm{i}\alpha)^2e^{KL}}{(1-\mathrm{i}\alpha)e^{KL}}=2\mathrm{i}\alpha A \tag{11.2.768}$$

We then get:

$$D=\frac{2\mathrm{i}\alpha(1-\mathrm{i}\alpha)e^{KL}}{(1+\mathrm{i}\alpha)^2e^{-KL}-(1-\mathrm{i}\alpha)^2e^{KL}}A \tag{11.2.769}$$

and let us still put:

$$F=\frac{1}{2}\left((1+\mathrm{i}\alpha)^2e^{-KL}-(1-\mathrm{i}\alpha)^2e^{KL}\right) \tag{11.2.770}$$

Therefore it comes:

$$D=\frac{\mathrm{i}\alpha(1-\mathrm{i}\alpha)e^{KL}}{F}A \tag{11.2.771}$$

Notice that we also have:

$$\begin{aligned}
F&=2\mathrm{i}\alpha\cosh(KL)-(1-\alpha)^2\sinh(KL)=2\mathrm{i}\alpha\frac{e^{KL}+e^{-KL}}{2}-(1-\alpha^2)\frac{e^{KL}-e^{-KL}}{2}\\
&=\frac{2\mathrm{i}\alpha e^{KL}+2\mathrm{i}\alpha e^{-KL}-(1-\alpha^2)\left(e^{KL}-e^{-KL}\right)}{2}\\
&=\frac{1}{2}\left(2\mathrm{i}\alpha e^{KL}+2\mathrm{i}\alpha e^{-KL}-e^{KL}+e^{-KL}+\alpha^2e^{KL}-\alpha^2e^{-KL}\right)\\
&=\frac{1}{2}\left(\alpha^2e^{KL}+2\mathrm{i}\alpha e^{KL}-e^{KL}+2\mathrm{i}\alpha e^{-KL}-\alpha^2e^{-KL}+e^{-KL}\right)\\
&=\frac{1}{2}\left[\left(\alpha^2+2\mathrm{i}\alpha-1\right)e^{KL}+\left(2\mathrm{i}\alpha-\alpha^2+1\right)e^{-KL}\right]\\
&=\frac{1}{2}\left[\left(2\mathrm{i}\alpha-\alpha^2+1\right)e^{-KL}+\left(\alpha^2+2\mathrm{i}\alpha-1\right)e^{KL}\right]\\
&=\frac{1}{2}\left[\left(2\mathrm{i}\alpha-\alpha^2+1\right)e^{-KL}-\left(-\alpha^2-2\mathrm{i}\alpha+1\right)e^{KL}\right]\\
&=\frac{1}{2}\left[\left(1+2\mathrm{i}\alpha-\alpha^2\right)e^{-KL}-\left(1-2\mathrm{i}\alpha-\alpha^2\right)e^{KL}\right]=\frac{1}{2}\left[(1+\mathrm{i}\alpha)^2e^{-KL}-(1-\mathrm{i}\alpha)^2e^{KL}\right]
\end{aligned} \tag{11.2.772}$$

We can now express the constants A' and B as a function of A with the help of the preceding relations. Let us first begin with B:

$$\begin{aligned}
B &= C+D-A=\frac{\mathrm{i}\alpha(1+\mathrm{i}\alpha)e^{-KL}}{F}A+\frac{\mathrm{i}\alpha(1-\mathrm{i}\alpha)e^{KL}}{F}A-A\\
&=A\frac{\mathrm{i}\alpha(1+\mathrm{i}\alpha)e^{-KL}+\mathrm{i}\alpha(1-\mathrm{i}\alpha)e^{KL}-F}{F}\\
&=A\frac{\mathrm{i}\alpha e^{-KL}-\alpha^2e^{-KL}+\mathrm{i}\alpha e^{KL}+\alpha^2e^{KL}-F}{F}\\
&=A\frac{\mathrm{i}\alpha(e^{KL}+e^{-KL})+\alpha^2(e^{KL}-e^{-KL})-F}{F}=A\frac{\mathrm{i}\alpha 2\cosh(KL)+\alpha^2 2\sinh(KL)-F}{F}\\
&=A\frac{\mathrm{i}\alpha 2\cosh(KL)+\alpha^2 2\sinh(KL)-2\mathrm{i}\alpha\cosh(KL)+(1-\alpha^2)\sinh(KL)}{F}\\
&=A\frac{\alpha^2 2\sinh(KL)+(1-\alpha^2)\sinh(KL)}{F}=A\frac{(1+\alpha^2)\sinh(KL)}{F}
\end{aligned} \tag{11.2.773}$$

and:

$$Ce^{KL}+De^{-KL}-A'e^{\mathrm{i}kl}=0 \tag{11.2.774}$$

Therefore:

$$\begin{aligned}
A' &= e^{-\mathrm{i}kl}\left(Ce^{KL}+De^{-KL}\right)\\
&=e^{-\mathrm{i}kl}\left(\frac{\mathrm{i}\alpha(1+\mathrm{i}\alpha)e^{-KL}}{F}Ae^{KL}+\frac{\mathrm{i}\alpha(1-\mathrm{i}\alpha)e^{KL}}{F}Ae^{-KL}\right)\\
&=\frac{\mathrm{i}\alpha Ae^{-\mathrm{i}kl}}{F}\left((1+\mathrm{i}\alpha)+(1-\mathrm{i}\alpha)\right)=\frac{\mathrm{i}\alpha Ae^{-\mathrm{i}kl}}{F}2=\frac{2\mathrm{i}\alpha Ae^{-\mathrm{i}kl}}{F}
\end{aligned} \tag{11.2.775}$$

So finally we have:

$$\begin{aligned}
\Psi_{\mathrm{I}} &= Ae^{\mathrm{i}kx}+Be^{-\mathrm{i}kx}=Ae^{\mathrm{i}kx}+A\frac{(1+\alpha^2)\sinh(KL)}{F}e^{-\mathrm{i}kx}\\
&=A\left(e^{\mathrm{i}kx}+\frac{(1+\alpha^2)\sinh(KL)}{F}e^{-\mathrm{i}kx}\right)\\
\\
\Psi_{\mathrm{II}} &= Ce^{kx}+De^{-kx}=\frac{\mathrm{i}\alpha(1+\mathrm{i}\alpha)e^{-KL}}{F}Ae^{Kx}+\frac{\mathrm{i}\alpha(1-\mathrm{i}\alpha)e^{KL}}{F}Ae^{-Kx}\\
&=A\left(\frac{\mathrm{i}\alpha(1+\mathrm{i}\alpha)e^{K(x-L)}+\mathrm{i}\alpha(1-\mathrm{i}\alpha)e^{-K(x-L)}}{F}\right)\\
\\
\Psi_{\mathrm{III}} &= A'e^{\mathrm{i}kx}+B'e^{-\mathrm{i}kx}=A'e^{\mathrm{i}kx}=\frac{2\mathrm{i}\alpha Ae^{-\mathrm{i}kl}}{F}e^{\mathrm{i}kx}=\frac{2\mathrm{i}\alpha Ae^{\mathrm{i}k(x-L)}}{F}
\end{aligned} \tag{11.2.776}$$

and therefore:

$$T=\frac{\left|\Psi^T\right|^2}{\left|\Psi^{\mathrm{I}}\right|^2}=\frac{\left|\frac{2\mathrm{i}\alpha Ae^{\mathrm{i}k(x-L)}}{F}\right|^2}{|Ae^{\mathrm{i}kx}|^2} \tag{11.2.777}$$

using the properties of the complex module (see section Numbers page 245):

$$T=\frac{\left|\frac{2\mathrm{i}\alpha Ae^{-\mathrm{i}kL}}{F}\right|^2\left|Ae^{\mathrm{i}kx}\right|^2}{\left|Ae^{\mathrm{i}kx}\right|^2}=\left|\frac{2\mathrm{i}\alpha Ae^{-\mathrm{i}kL}}{F}\right|^2=\left|2\mathrm{i}\alpha\right|^2\left|\frac{e^{-\mathrm{i}kL}}{F}\right|^2=4\alpha^2\frac{\left|e^{-\mathrm{i}kL}\right|^2}{|F|^2}=4\alpha^2\frac{1}{|F|^2} \tag{11.2.778}$$

It remains for us to calculate:

$$\begin{aligned}|F|^2=F\cdot\bar{F}&=|2\mathrm{i}\alpha\cosh(KL)-(1-\alpha^2)\sinh(KL)|\\&=\big(2\mathrm{i}\alpha\cosh(KL)-(1-\alpha^2)\sinh(KL)\big)\big(-2\mathrm{i}\alpha\cosh(KL)-(1-\alpha^2)\sinh(KL)\big)\\&=4\alpha^2\cosh^2(KL)-2\mathrm{i}\alpha(1-\alpha^2)\sinh(KL)\cosh(KL)+2\mathrm{i}\alpha\cosh(KL)(1-\alpha^2)\sinh(KL)\\&+(1-\alpha^2)^2\sinh^2(KL)\\&=4\alpha^2\cosh^2(KL)+(1-\alpha^2)^2\sinh^2(KL)\\&=4\alpha^2\cosh^2(KL)+(1-2\alpha^2+\alpha^4)\sinh^2(KL)\\&=4\alpha^2\cosh^2(KL)+\sinh^2(KL)-2\alpha^2\sinh^2(KL)+\alpha^4\sinh^2(KL)\\&=2\alpha^2\cosh^2(KL)+\sinh^2(KL)+2\alpha^2+\alpha^4\sinh^2(KL)\\&=2\alpha^2(1+\cosh^2(KL))+(1+\alpha^4)\sinh^2(KL)=2\alpha^2(2+\sinh^2(KL))+(1+\alpha^4)\sinh^2(KL)\\&=4\alpha^2+2\alpha^2\sinh^2(KL)+\sinh^2(KL)+\alpha^4\sinh^2(KL)\\&=4\alpha^2+(2\alpha^2+1+\alpha^4)\sinh^2(KL)=4\alpha^2+(1+\alpha^2)^2\sinh^2(KL)\end{aligned} \tag{11.2.779}$$

Therefore:

$$\begin{aligned}T=4\alpha^2\frac{1}{4\alpha^2+(1+\alpha^2)^2\sinh^2(KL)}&=\frac{1}{1+\dfrac{(1+\alpha^2)^2}{4\alpha^2}\sinh^2(KL)}\\&=\frac{1}{1+\dfrac{\left(1+\dfrac{k^2}{K^2}\right)^2}{4\dfrac{k^2}{K^2}}\sinh^2(KL)}=\frac{1}{1+\dfrac{\left(1+\dfrac{E_{\text{tot}}}{E_p-E_{\text{tot}}}\right)^2}{4\dfrac{E_{\text{tot}}}{E_p-E_{\text{tot}}}}\sinh^2(KL)}\\&=\frac{1}{1+\dfrac{E_p-E_{\text{tot}}}{4E_{\text{tot}}}\left(1+\dfrac{E_{\text{tot}}}{E_p-E_{\text{tot}}}\right)^2\sinh^2(KL)}\\&=\frac{1}{1+\dfrac{E_p-E_{\text{tot}}}{4E_{\text{tot}}}\left(1+2\dfrac{E_{\text{tot}}}{E_p-E_{\text{tot}}}+\dfrac{{E_{\text{tot}}}^2}{(E_p-E_{\text{tot}})^2}\right)\sinh^2(KL)}\end{aligned} \tag{11.2.780}$$

$$\begin{aligned}
&=\frac{1}{1+\dfrac{1}{4}\left(\dfrac{E_p-E_{\text{tot}}}{E_{\text{tot}}}+2+\dfrac{E_{\text{tot}}}{E_p-E_{\text{tot}}}\right)\sinh^2(KL)}\\
&=\frac{1}{1+\dfrac{1}{4}\dfrac{(E_p-E_{\text{tot}})^2+2E_{\text{tot}}(E_p-E_{\text{tot}})+E_{\text{tot}}^2}{E_{\text{tot}}(E_p-E_{\text{tot}})}\sinh^2(KL)}\\
&=\frac{1}{1+\dfrac{E_p^2}{4E_{\text{tot}}(E_p-E_{\text{tot}})}\sinh^2(KL)}
\end{aligned}\tag{11.2.781}$$

We have therefore:

$$T=\frac{1}{1+\dfrac{E_p^2}{4E_{\text{tot}}(E_p-E_{\text{tot}})}\sinh^2(KL)}\tag{11.2.782}$$

But as (see section Trigonometry page 2532):

$$\sinh(KL)=\frac{e^{KL}-e^{-KL}}{2}\tag{11.2.783}$$

If $KL\gg 1$ (so at the atomic scale it is rather K which is huge relatively to L) we have:

$$\sinh(KL)\cong\frac{e^{KL}}{2}\tag{11.2.784}$$

Therefore:

$$T\cong\frac{1}{1+\dfrac{E_p^2e^{2KL}}{16E_{\text{tot}}(E_p-E_{\text{tot}})}}\cong\frac{1}{\dfrac{E_p^2e^{2KL}}{16E_{\text{tot}}(E_p-E_{\text{tot}})}}=\frac{16E_{\text{tot}}(E_p-E_{\text{tot}})}{E_p^2}e^{-2KL}\tag{11.2.785}$$

relation that we can find in many textbooks (but without detailed proof as above). Below we have plotted T:

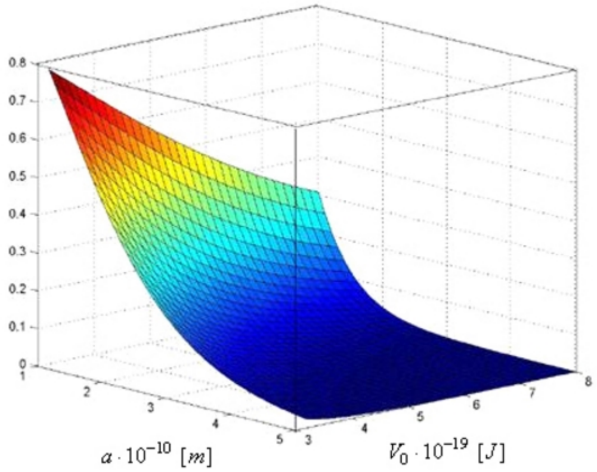


Figure 11.57: **Plot of the transmission T coefficient with MATLAB® 5.0.0.473 for tunneling effect**

following the relation:

$$\boxed{T \cong \frac{16E_{\text{tot}}(V_0 - E_{\text{tot}})}{V_0^2} e^{-2Ka}} \tag{11.2.786}$$

The above relation is often denoted:

$$T \propto e^{-G} \tag{11.2.787}$$

where G is named the "Gamow factor" and defined when we model the Coulomb barrier as a succession of thin barries of varying height by:

$$G = \frac{2}{\hbar}\int_{R}^{r_{\text{C}}} \left[2m\left|V_{\text{Coul.}}(r) - E_{\alpha}\right|\right]^{1/2} \mathrm{d}r \tag{11.2.788}$$

We thus find in the (special) above example that the coefficient T is very sensitive (exponentially) to a small variation of the width of the barrier, a, when the potential of this barrier is small. We can thus visualize atomic sites, for example in Silicium, using a point very close to the material (or "scanning tunneling microscope (STM)") to be observed. This is the principle of the tunneling microscope where, by approaching a very finely cut conductive point (a few atoms) to a proximity of about 5 Angstroms of a conductive surface, and imposing a potential difference of a few [mV], we measures a current of a few nanoamperes. The number of electrons passing through the potential barrier (here the vacuum between the two conductive electrodes) decreases exponentially with the width of the barrier.

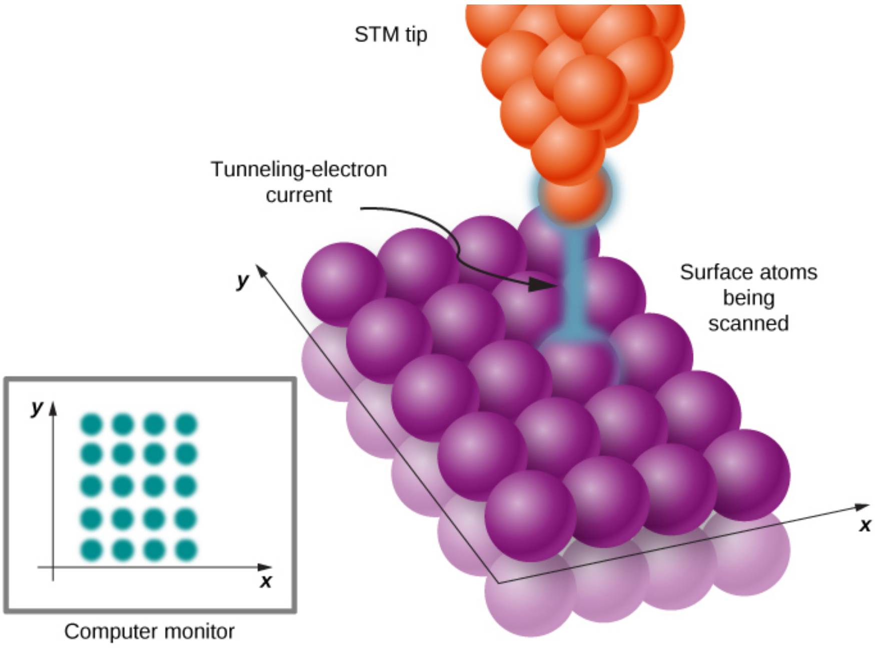


Figure 11.58: **Scanning tunneling microscope concept (source: OpenStax)**

By analysing the signal of the current passing through the circuit, one can access a very precise mapping of the measured surface of the order of 0.1 Angstroms in vertical.

We also notice according to the relation obtained that the not "heavy" particles like the electrons have a greater probability of making a tunnel effect than heavier particles because of the term of mass involved.

STM images are in therefore gray scale, and colouring is added to bring up details to the human eye:

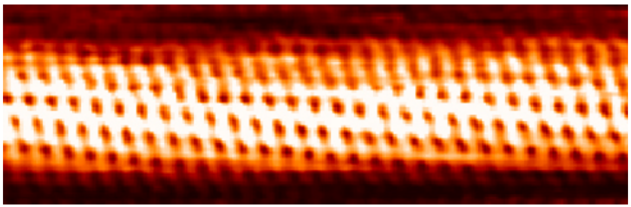


Figure 11.59: **An STM image of a carbon nanotube (source: OpenStax)**

Using the relation obtained previously, one can quite simply calculate the probability that a human being of mass m crosses a wall with a height h (it is then easy to calculate the potential energy) and a thickness a. The probability is of the order of $10^{-4\cdot 10^{30}}$....

However, the most famous example of Quantum tunnelling that can be treated is that of the emission of α particles by radioactive heavy nuclei, the explanation of which was given by the Russian physicist G. Gamov in 11928 (holocene calendar).

The proof is relatively simple, but since it constitutes a particular practical case, we have not detailed it in this section, but in that of Nuclear Physics. However, to solve this problem, we need to use an approximation method known as the "W.K.B. method" Named after the physicists Wentzel, Kramers and Brillouin.

The results therefore give a transmission factor T for the α "particle" of:

$$T \cong 2\cdot 10^{-39} \tag{11.2.789}$$

For the Uranium atom ${}^{238}_{92}\text{U}$. Moreover, in the semi-classical approximation, the particle α has, in the well, a velocity of the order of 10^7 $[\text{m}\cdot\text{s}^{-1}]$, and it goes back and forth in a nucleus whose radius is of the order of 10^{-14} [m]. It thus performs approximately 10^{21} oscillations per second where each time it has a probability T to cross the potential barrier. This probability per unit time is thus determined by:

$$\lambda \cong \frac{v}{r}\cdot T \cong 10^{21}\cdot T \cong 2\cdot 10^{-18}\ [\text{s}^{-1}] \tag{11.2.790}$$

Experimentally, we find:

$$\lambda \cong 5\cdot 10^{-18}\ [\text{s}^{-1}] \tag{11.2.791}$$

Caution!!! Although a radioactive decay event seems spontaneous and is unpredictable (i.e. without cause), it is indeed triggered by a physical agent. That physical agent is a vacuum fluctuation (the latter having no cause!). Due to the quantum nature of our universe, a vacuum always contains vacuum fluctuations!!!

Remark

As we will see in the section of Cosmology, looking far in space is equivalent at looking in the past (because light takes time to travel from one place to another, we see objects not as they are now but as they were at the time when they released the light that has travelled across the universe to us). Therefore modern telescopes are like time machines as we can go back as 13 billion years in the past by spectral observation. Thanks to spectroscopy we can check hydrogen spectral series that directly depends on the fine-structure constant (see page 4655) - as for the relation of activity decay that we have derived earlier above that depends only in constant parameters involved in the fine-structure constant - and see if the latter has changed these last 13 billion years. Experimental evidence as presented in [574] find a spatial variation of the order of $\sim 10^{-5}$ is preferred over a no-variation model at the 3.9σ level:

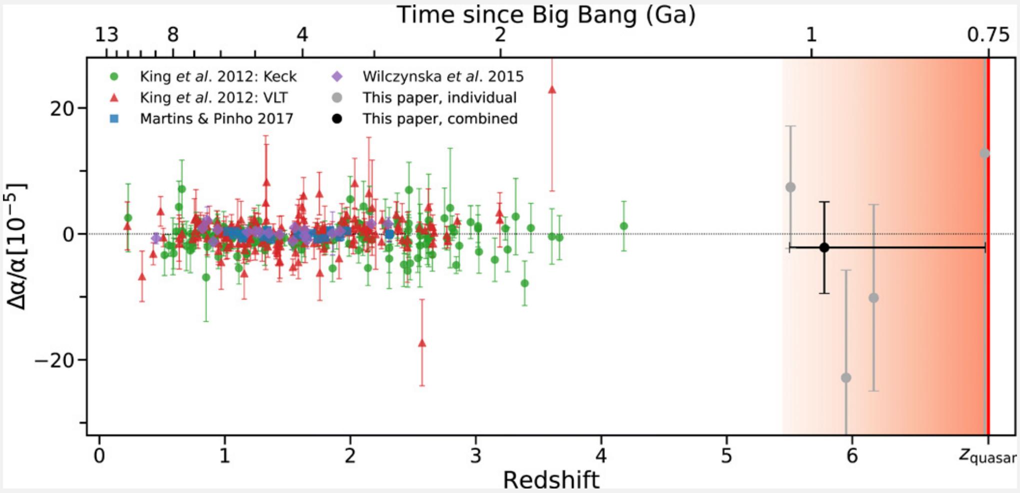


Figure 11.60: Direct measurements of $\Delta\alpha/\alpha$, taken from quasar absorption measurements

This variation of $\sim 10^{-5}$ of the fine-structure constant would lead in a relative variation in the radioactivity decay of the order of $\sim 10^{-4}$.

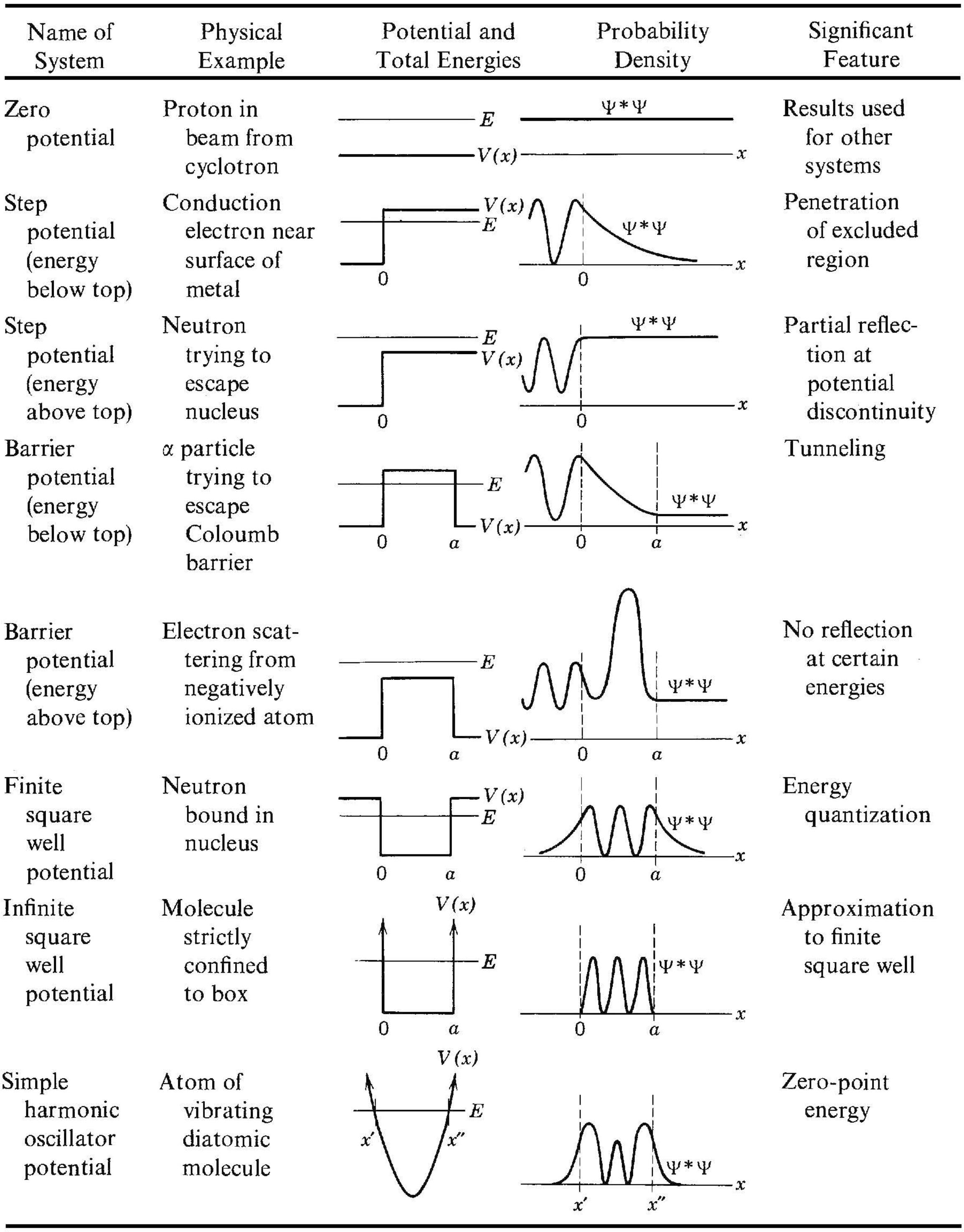
The model that we will see in detail in the section on Nuclear Physics gives therefore quite satisfactory results.

And here is how looks like a "typical" STM in this beginning of the 121st century (holocene calendar):

Figure 11.61: **Omicron Scanning tunneling microscope (source: Omicron)**

Besides these technical examples, we encounter the phenomenon of quantum tunneling in a much more accessible and very pedagogical case: Thus, when under the condition of total reflection of a beam of light, we approach another prism (on the face of the prism where no ray of light comes out or re-enters) so as to produce a sufficiently thin air space, a small transmitted beam is observed. This is sometimes named "optical tunneling".

In the table we summarize some of the properties of the systems studied in this so far. The table gives an abbreviated name for each idealized system, and an example of a physical system whose potential and total energies are approximated by the idealization. It also gives sketches of the forms of the potential and total energies, and corresponding probability density functions, for each system. If the particle is not bound, it is incident from the left. We have chosen one significant feature of each system to list in the table, but there are many other significant features that we have discussed, which are not listed. In fact, in this chapter we have obtained most of the important predictions of quantum mechanics for systems involving one particle moving in a one-dimensional potential. In the following chapters we shall obtain predictions from the theory for systems involving three dimensions and several particles.



Name of System	Physical Example	Potential and Total Energies	Probability Density	Significant Feature
Zero potential	Proton in beam from cyclotron			Results used for other systems
Step potential (energy below top)	Conduction electron near surface of metal			Penetration of excluded region
Step potential (energy above top)	Neutron trying to escape nucleus			Partial reflection at potential discontinuity
Barrier potential (energy below top)	α particle trying to escape Coloumb barrier			Tunneling
Barrier potential (energy above top)	Electron scattering from negatively ionized atom			No reflection at certain energies
Finite square well potential	Neutron bound in nucleus			Energy quantization
Infinite square well potential	Molecule strictly confined to box			Approximation to finite square well
Simple harmonic oscillator potential	Atom of vibrating diatomic molecule			Zero-point energy

Figure 11.62: A summary of the systems studied so far (source: [162])

We saw above that a particle inside a barrier has imaginary momentum but still can be described by a mathematically proper wave function. Furthermore, and more important, the prediction of quantum theory that the particle can tunnel though the barrier to the other side, a process not allowed in classical physics, has been thoroughly confirmed by observations for over fifty years. While it is true that we cannot measure the particle's momentum inside the barrier, we can still treat it mathematically. This demonstrates that physicists are not limited to speaking only about that which they can directly observe and measure. The mathematics that describes observable events can also be used to describe unobservable events and provide useful predictions of other events that are observable.

11.2.4.4.9.1 Quantum time tunneling

Recent experiments on quantum tunneling have re-ignited the longstanding debate over how long a particle takes to tunnel through a barrier. Naïve calculations suggest that faster-than-light tunneling is possible, while some experiments have superficially suggested that such a phenomenon might have been observed. However, it has been argued that information does not exceed the speed of light in these experiments, so that relativistic causality remains intact.

The analysis of tunneling time is complicated because time plays an unusual and subtle role in quantum mechanics. Unlike position, time is not usually treated as an operator; rather it is a parameter. Consequently, the energy-time uncertainty principle does not enjoy the unassailable central position in the theory as does the position-momentum uncertainty principle. This leads to considerable ambiguity when it comes to the measurement of the duration between quantum events.

Attempts to define tunneling time have led to an extensive and confused literature so far even in the early 121st century (holocene calendar). Most theoretical treatments focus on the behaviour of a wave packet as it traverses a square barrier. However, such a barrier is dispersive, so the packet is disrupted by the experience. Also, interference between parts of the wave packet reflected from the barrier and parts still approaching further complicates matters.

A simple heuristic argument to estimate the tunnelling time goes as follows. To surmount a square barrier of height E_p, a particle with energy E must borrow an amount of energy $E_p - E$. According to the energy Heisenberg's uncertainty principle, this must be repaid after a time:

$$\Delta t = \frac{\hbar}{\Delta E} = \frac{\hbar}{E_p - E} \tag{11.2.792}$$

Or according to Howard P. Robertson derivation seen earlier:

$$\Delta t \geq \frac{\hbar}{2\Delta E} = \frac{\hbar}{2(E_p - E)} \tag{11.2.793}$$

This provides a crude upper bound for the tunnelling time.

If the width of the barrier is L, then the effective speed of the particle during the tunnelling can be:

$$\Delta v = \frac{L}{\Delta t} \leq \frac{2L(E_p - E)}{\hbar} \tag{11.2.794}$$

As L can be made as large as we please, there is no upper bound on this effective velocity. In particular, it may exceed the speed of light, in apparent violation of relativistic causality. Moreover, the above expression has the odd feature that as the height E_p of the potential hill is increased, so the tunnelling time decreases, i.e. the more repulsive the potential, the faster the particle moves in the forward direction!

11.2.4.5 Superposition principle

At the beginning of the section we have already introduce the basics of quantum superposition during our study of the linear combination of quantum states (see page 4779). We would like now to deepen the subject a bit more to get some very important results.

Let us recall that the notion of the dynamic state of a classical system plays a crucial role in Classical Mechanics.

Is it possible to fall back on this concept when we are dealing with a quantum system, that is to say a system such as an atom, a nucleus or a molecule, in short a microphysics system?

At first glance no! Because we know that one defines the dynamic state of a classical system by the data of the generalized coordinates q_i and conjugate moments p_i at a given instant (see section Analytical Mechanics). However, the principle of uncertainty is opposed to this procedure as soon as we are in the field of microphysics, given the impossibility of accurately measuring the q_i and p_i. This is particularly clear when the system is reduced to a single particle which we describe by its cartesian coordinates $q_i(x,y,z)$ and the components of its linear momentum $\vec{p}(p_x,p_y,p_z)$.

Fortunately, there is another definition of the dynamic state of a system which applies indifferently to Classical and Quantum systems and which, in the case of the former, is identified with the usual definition. We will give this definition on the basis of a brief theory of the sets of identical systems.

If we have a set (E) of a very large number of identical systems, we will make a statistical survey to characterize this set as follows: we take a system of the set, we measure a dynamic variable (coordinate, linear momentum, kinetic energy, etc.) and we reject the system (which disturbed by the measure, must not be reincorporated to the initial set). We thus draw up a balance sheet which is translated by distribution functions of all the possible dynamic variables. This makes it possible to define unambiguously the notion of "identity":

Definition (#11.2.508): Two physical sets are say to be "identical physical systems", if the results of the measurements are the same for the both.

Let us now consider a unique set (E). Is it possible to realize it by juxtaposition of two sets non-identical sets (E_1) and (E_2)? This would allow us to write:

$$(E)=(E_1)+(E_2) \tag{11.2.795}$$

If Yes, we will say that (E) is a "mixture of physical systems". Conversely, by suitable sampling, a mixture can be broken down into two different sub-systems. If not, we will say that (E) is a "pure set" or "pure physical system". Any sampling will decompose the pure set into two subsets identical to each other and necessarily also identical to (E)! We then agree to say that all systems of a pure set are in the same dynamic state and that two different pure sets give rise to different dynamic states. It goes without saying that the systems constituting a mixture will be in different dynamic states.

Suppose now that the studied systems obey the laws of Classical Mechanics. If the systems of a set have different pairs (q_i, p_i), we sort them by grouping them by systems all having a same pairs (q_i, p_i). We check then that the new definition of the dynamic state coincides with the usual definition. Let us notice this obvious but important fact (as opposed to quantum systems): in a pure set of classical systems, that is to say for a given dynamic state, every dynamic variable is well determined. Indeed, in Analytical Mechanics, such a variable is a function of the q_i and p_i and, therefore, has a unique value.

Let us pass to quantum systems. It is now possible to define for them a dynamic state, but immediately we see a fundamental distinction with Classical Mechanics. Indeed, in a pure set of quantum systems, that is to say for a given dynamic state, a dynamic variable is not, in general, well determined as already mentioned. When we measure it on systems extracted from the pure set, we usually do not find a single value, but a distribution of values as a result.

The uncertainty which reigns over the value of a dynamic variable in a given dynamic state is therefore purely quantum in nature and it is well to distinguish it from the indeterminacy of statistical origin manifested in a mixture, whatever they are classical or quantum systems.

The formalism of quantum physics can only be edified if we know how to describe mathematically dynamic states and dynamic variables. We have seen that we can not expect from this formalism a precise prediction as in Classical Mechanics, but simply the probabilities of obtaining this or that value when we measure a dynamic variable on a system whose dynamic state is given.

The whole theory we have seen so far allows us to conclude that the dynamic states of a system of a spin-free particle are described by complex, wave functions non-zero everywhere.

If we apply this condition to dynamic systems, then we have the following assumption:

$$\Psi=\lambda_1\Psi_1+\lambda_2\Psi_2 \tag{11.2.796}$$

Ψ is then a wave function describing a possible dynamic state of the system. What is often written with the Dirac ket-bra notation in the following form:

$$|\,\Psi\rangle=\lambda_1\,|\,\Psi_1\rangle+\lambda_2\,|\,\Psi_2\rangle \tag{11.2.797}$$

This postulate seems quite natural because of the wave aspect of the physics of microsystems. Indeed, in the wave phenomena of classical physics, wave equations are, in most cases, linear homogeneous, and it follows that the waves can be superimposed (see section Wave Mechanics page **??**). Now, the great interest of this postulate is that it contains in germ the explanation of this fundamental fact which is quantum indeterminacy (also sometimes named "quantum coherence").

Let us see it on a very simple case where we assume that a dynamic variable A has a well-defined value a_1 in the dynamic state Ψ_1 and a well-defined value a_2 in the dynamic state Ψ_2 with $a_1 \neq a_2$. This means that if we repeat the measurement of A on systems all in the dynamic state described by Ψ_1, we will find each time a_1, same for Ψ_2 with a_2.

A question comes naturally to mind: if we measure A on systems all in the dynamic state Ψ what are we going to get? A naive idea would be to think that A will take a well defined intermediate value between a_1 and a_2.

These two assumptions are false and we know that! First, A is not well defined in quantum physics (uncertainty) and second it is not necessarily mathematically located between a_1 and a_2. The correct interpretation is as follows:

If we measure A on the system in the dynamic state Ψ, we find as a result of measurement, sometimes a_1, with a probability p_1, sometimes a_2, with a probability $p_2=1-p_1$. Of course, p_1 and p_2 and will have to be calculated according to λ_1 and λ_2.

Remark

The reader must especially not confuse the pure set of systems described by Ψ with the mixture that we would obtain by juxtaposing two pure sets of systems respectively Ψ_1 and Ψ_2.

✍**Example:**

Let us consider the following superposed academic case:

$$|\,\Psi\rangle=\frac{\sqrt{3}}{3}\,|\,\Psi_1\rangle+\frac{2}{3}\,|\,\Psi_2\rangle+\frac{\sqrt{2}}{3}\,|\,\Psi_3\rangle \tag{11.2.798}$$

We can already easily verify that the system is normalized:

$$|\langle\Psi\,|\,\Psi\rangle|^2=\left|\frac{\sqrt{3}}{3}\langle\Psi_1\,|\,\Psi_1\rangle+\frac{2}{3}\langle\Psi_2\,|\,\Psi_2\rangle+\frac{\sqrt{2}}{3}\langle\Psi_3\,|\,\Psi_3\rangle\right|^2=\left|\frac{1}{3}+\frac{4}{9}+\frac{2}{9}\right|^2=|1|^2=1 \tag{11.2.799}$$

Since the superposition is normalized, then it comes the probability of finding the system in for example in the state Ψ_1 which is then under the assumption of the orthonormality $\langle\Psi_i|\Psi_j\rangle=\delta_{ij}$:

$$|\langle\Psi_1\,|\,\Psi\rangle|^2=\left|\frac{\sqrt{3}}{3}\underbrace{\langle\Psi_1\,|\,\Psi_1\rangle}_{=1}+\frac{2}{3}\underbrace{\langle\Psi_1\,|\,\Psi_2\rangle}_{=0}+\frac{\sqrt{2}}{3}\underbrace{\langle\Psi_1\,|\,\Psi_3\rangle}_{=0}\right|^2=\left|\frac{\sqrt{3}}{3}\right|^2=\frac{1}{3} \tag{11.2.800}$$

and if we do the same for each of the other two states the sum of the probabilities will always be equal to 1. It should be pointed out that this probability is also the proportion of states that will be measured in the state Ψ_1 if the system is constituted of N identical components.

The superposition principle applies not only to particles but also to atoms and even some objects that have the size of molecules. A famous well known case is the nuclear (radioactive) decay. Over a single half life the atom changes as follows:

$$|U\rangle\rightarrow\frac{1}{\sqrt{2}}(|U\rangle+|D\rangle) \tag{11.2.801}$$

where $|U\rangle$ is the undecayed state and $|D\rangle$ is the decayed state. The probability is given by the square of the number in front of the states: $\left(\frac{1}{\sqrt{2}}\right)^2=\frac{1}{2}$. After the half life, there are two versions of the atom. Before the decay it looks like there is just one instance of the atom, but this is misleading!

What is important to understand here is that even before we observe (measure) the decay, it has already happened with a certain probability. So the effect (the decay) happens before the cause which is the observation (the collapse of the wave function). So there is a violation of causality in the classical macroscopic sense of the term!

In fact, the interpretation given by de Broglie's theory (associate a wave function to a particle) with the principles of uncertainty is the most striking and well known example of superposition principle of states in quantum physics (Schrödinger's Cat case putted aside)! To see why, let us consider a de Broglie wave propagating in the direction of the x-axis, but limited to an interval $(-a,+a)$ at a given instant ($t=0$ if we want). Hence at $t=0$ the wave is written, dropping the multiplicative constant:

$$\begin{aligned}\Psi(x)&=e^{\mathrm{i}k_0x} \quad \text{for} \quad |x|<a\\ \Psi(x)&=0 \quad \text{for} \quad |x|>a\end{aligned} \tag{11.2.802}$$

If we measure the coordinate of the particle, we must find it there necessarily where Ψ is not zero (otherwise we could not measure anything). We can say that $x=0$ with obviously an uncertainty $\Delta x=a$ (the interval where we are sure to find the particle with respect to the origin divided by two: $d((-a,+a))/2$.

If we measure p, what do we find? We must not find $\hbar k_0$ (a relation which we have already proved earlier above), as this would be true for an indefinite plane wave, which is not the case here. Then we will decompose the wave into plane waves by means of the Fourier transform (see section Sequences and Series page 1758):

$$\Psi(x)=\int\limits_{-\infty}^{+\infty}a(k)e^{\mathrm{i}kx}\mathrm{d}k \tag{11.2.803}$$

How to interpret this relation? One of the elementary plane waves (which we can also interpret as a state):

$$a(k)e^{\mathrm{i}kx}\mathrm{d}k \tag{11.2.804}$$

whose sum gives back $\Psi(x)$, leads to a value $p=\hbar k$ of the linear momentum. But, the values of k form a continuum. We are led to say that the possible values of p then form also continuum and that there is therefore an uncertainty about the value of p. To go further, one has to evaluate $a(k)$ (which must be considered as variable of the probability of presence of each plane wave coming from the decomposition of $\Psi(x)$) by means of the relation (according to the properties of the Fourier transform proved in the section Sequences ans Series):

$$a(k)=\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty}\Psi(x)e^{-\mathrm{i}kx}\mathrm{d}x \tag{11.2.805}$$

which gives (reduced to) here:

$$a(k)=\frac{1}{2\pi}\int\limits_{-a}^{+a}e^{-\mathrm{i}(k_0-k)x}\mathrm{d}x \tag{11.2.806}$$

Let us put $k_0 - k = \Delta k$, then the integral becomes:

$$a(k)=\frac{1}{2\pi}\int_{-a}^{+a}\Psi(x)e^{-\mathrm{i}\Delta kx}\mathrm{d}x=\frac{1}{2\pi}\left[\frac{1}{-\mathrm{i}\Delta k}e^{-\mathrm{i}\Delta kx}\right]_{-a}^{+a}=\frac{1}{-\pi\Delta k}\left[\frac{e^{\mathrm{i}(-a\Delta k)}-e^{\mathrm{i}(a\Delta k)}}{2\mathrm{i}}\right]=\frac{a}{\pi}\frac{\sin(a\Delta k)}{a\Delta k} \tag{11.2.807}$$

Let us recall the plot of $\sin(u)/u$ (that is the "cardinal sine" as seen in the section of Trigonometry) that show that $a(k)$ takes values that can be considered negligible for $|u|>\pi$:

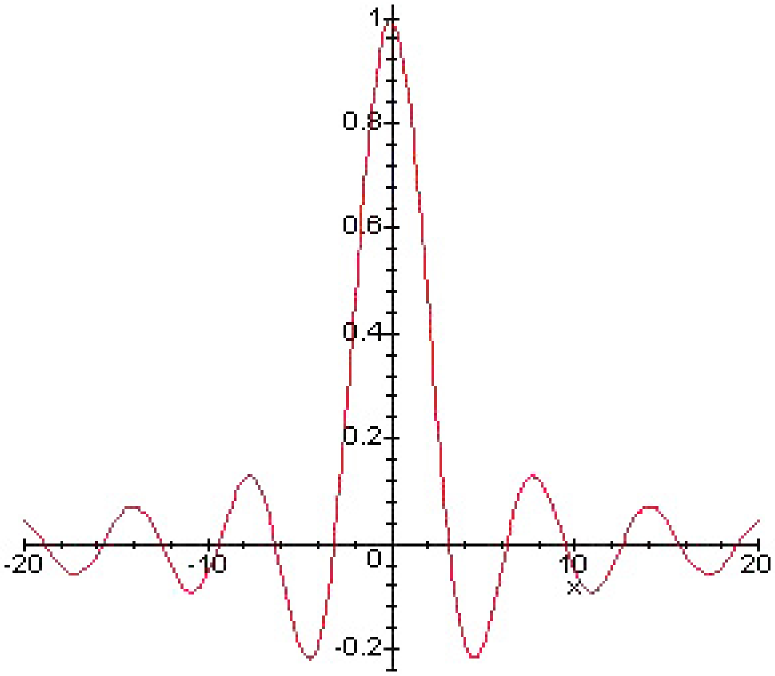


Figure 11.63: 2D plot of the sinc function in Maple™ 4.00b

It follows that in the integral:

$$\Psi(x)=\int_{-\infty}^{+\infty}a(k)e^{\mathrm{i}kx}\mathrm{d}k \tag{11.2.808}$$

it is the k at the neighbourhood of k_0 that are effective, and more precisely the k such that:

$$|k-k_0|a<\pi \tag{11.2.809}$$

since:

$$a(k)=\frac{a}{\pi}\frac{\sin(a\Delta k)}{a\Delta k}=\frac{a}{\pi}\frac{\sin(a(k-k_0))}{a\Delta k} \tag{11.2.810}$$

It follows that the values to be retained of p are those close to $\hbar k_0$, more precisely we have if we multiply the prior previous relation by $\hbar$:

$$|k\hbar-k_0\hbar|a<\pi \tag{11.2.811}$$

Thus:

$$|p-p_0|a<\pi \tag{11.2.812}$$

That is:

$$|p-p_0|\frac{(a-(-a))}{2}<\pi\hbar \tag{11.2.813}$$

Therefore:

$$|p-p_0|(a-(-a))<2\pi\hbar \tag{11.2.814}$$

Finally (it is usage to replace a by the letter x):

$$\boxed{\Delta p\Delta x<h} \tag{11.2.815}$$

So we fall back on a well known result already introduced at the beginning of this section but now with an inequality "$<$".

Similarly, if we propose to determine the x-coordinate of an electron by passing it through a slit of width $2b$ pierced in a screen:

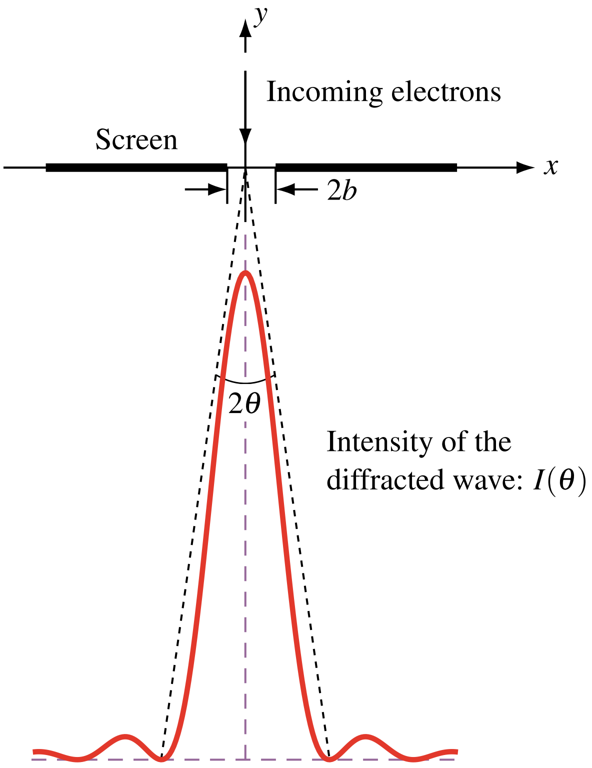


Figure 11.64: **Study configuration of the electron passing through a rectangular slot (author: H. Zeljko)**

The precision with which we know the position of this electron is limited by the size of the slit, i.e. $\Delta x=b$. On the other hand, the slit disturbs the associated wave. The result is a modification of the electron's motion, which is reflected by the diffraction pattern of the wave (which is in fact a representation of the linear superposition of its intrinsic states).

The uncertainty on the dynamic component p_x of the electron linear momentum is determined by the angle θ corresponding to the central maximum of the diffraction pattern. According to the theory of diffraction (see section Wave Optics page 4492) produced by a rectangular slot, we have $\sin(\theta)=\lambda/2b$ since the intensity $I(\theta)$ is written:

$$I(\theta)=I(0)\sin^2\left(\frac{1}{\left[\dfrac{2\pi b}{\lambda}\sin\theta\right]}\right) \tag{11.2.816}$$

Thus p_x is comprised between $p\sin(\theta)$ and $-p\sin(\theta)$, p being the linear momentum of the incident electron. Thus the uncertainty Δp_x is:

$$\Delta p_x=p\sin=\frac{h}{\lambda}\frac{\lambda}{2b}=\frac{h}{2}\cdot b \tag{11.2.817}$$

This simple result is quite extraordinary if we put it in relation, in order of magnitude, with the result we had obtained just above and that was for recall:

$$\Delta p\Delta x<h \tag{11.2.818}$$

We can draw several conclusions of the first importance:

1. The associated de Broglie wave is closely related to the principle of uncertainty and quantum physics must take simultaneously in account these two properties.
2. If we take into account that the distribution of the intensity is obtained from the counting of the electrons (or particles) as a function of the angle and that we get the same distribution regardless of the intensity of the monokinetic electron beam that arrives on the slot, even if the electrons are sent one by one. We observe then that the motion of particles is no longer deterministic but probabilistic. Thus, the wave equation of the electron can be considered as a linear superposition of the states defined each as we have done previously, that is to say by its possible spectral decomposition by the Fourier transform.

The observation of interference patterns in double-slit experiments with massive particles is generally regarded as the ultimate demonstration of the quantum nature of these objects (see section Wave Optics page 4509). Such matter–wave interference has been observed for electrons, neutrons, atoms, and molecules quite huge molecules (phthalocyanine) and, in contrast to classical physics, quantum interference can be observed when single particles arrive at the detector one by one. The build-up of such patterns in experiments with electrons has been described as the most beautiful experiment in physics!

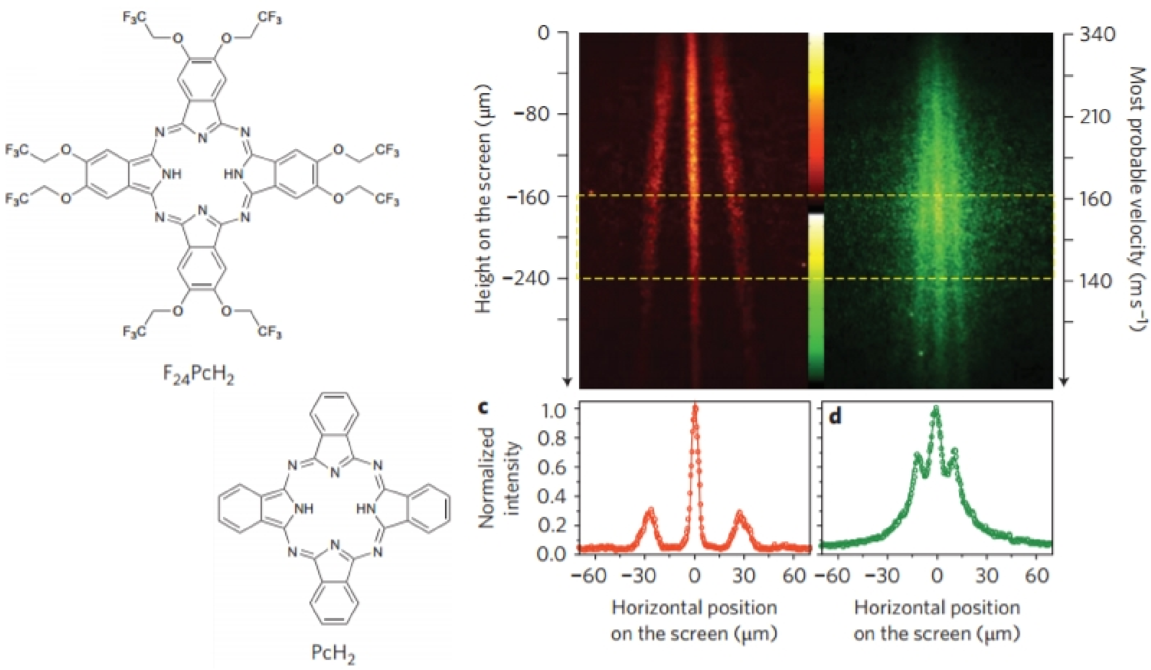


Figure 11.65: Comparison of interference patterns for $\mathrm{PcH_2}$ and $\mathrm{F_{24}PcH_2}$ (source: [298])

What can we conclude from all that we have seen so far:

1. The equations of Wave Quantum Physics give us a probability density to find a particle in a certain volume of space-time.
2. The linear superposition of states can be interpreted as the fact that it is possible to find a particle at several points in space-time at a given instant, and with for each of these points a certain probability of finding it there (by possible decomposition of the wave equation). This has been confirmed experimentally with a single neutron in an experiment described in the following paper [346].

If point (1) has been widely studied so far on this site, point (2) is new and arises from a simple mathematical operation of decomposition or superposition.

It seems that Richard Feynman called the double-split experiment «*impossible, absolutely impossible, to explain in any classical way*» and said it «*has in it the heart of quantum mechanics. In reality, it contains the only mystery.*».

But what happens if we try to measure the energy of an atom that is in a superposition of states of energy? We will never detect this superposition, but only one of the energies that constitute it, the action of measuring eliminates the superposition of the states in favour of a single one - we speak then of "quantum decoherence" (this is the Copenhagen interpretation which we mentioned implicitly at the very beginning of this section and to which we shall return later). But which one? Quantum physics can not simply answer this question for now.... It seems that this choice is made randomly! On the other hand, in the absence of predicting the precise state which will be measured among all those who constitute the superposition, quantum theory can give the probability that one has to measure each state (what we have already repeated many times, right here). If many measurements are made, we finally find the proportions predicted by the theory (even if each measure is unpredictable).

Erwin Schrödinger, had emphasized the absurdity (according to him) of these superpositions by resorting to a thought experiment become famous: Imagine a cat enclosed in an airtight box. In the box is also a radioactive atom and a device capable of spreading poison. When the radioactive atom disintegrates, it triggers the deadly device: the poison spreads into the box and the cat dies.

The absurdity of this experience (that we will in-deep in the section of Quantum Computing) is obvious ... but difficult to prove, at least until we understand what distinguishes a Cat from a Particle. Always the problem of the quantum-classical boundary ...

It was not until the 11980s (holocene calendar) that the situation finally progressed, both on the front of experience and on that of theory. In 11982 (holocene calendar), Wojciech Zurek, a researcher at the Los Alamos National Laboratory in New Mexico, takes up a very simple but brilliant idea: in a measurement, what produces decoherence, is the interaction of the system with its environment. More generally, quantum objects are never completely isolated from their environment - we mean everything that interacts with the system: an apparatus, air molecules, light photons, etc. So that in reality the quantum laws must be applied to the whole consisting of the object and all that surrounds it! However, Zurek shows that the multiple interactions with the environment lead to a very rapid destruction of the quantum coherence of the superpositions of states (also called "quantum interference" since mathematically one deals with wave functions). By destroying interferences, the environment suppresses state superpositions and quantum behaviour, so that only simple states remain and classic behaviour returns.

In a macroscopic object - a cat for example ... - each of the atoms is surrounded by many other atoms that interact with it. All these interactions spontaneously cause a jamming of the quantum interferences which disappear very quickly. This is why quantum physics does not apply to our scale (the probability that it happens is too small): systems are never isolated!

The speed of the decoherence increases with the size of the system: a cat that has 10^{27} particles, "decoheres" in 10^{-23} seconds, which explains why we have never seen zombie living-dead cat until today...!

Quantum physics is therefore a theory:

- Non-deterministic (probabilistic) hence the fact that it is considered as a theory of information
- Non-local: quantum objects may simultaneously have several positions
- non-separable: several quantum objects can be superimposed so that they can not be considered separately.

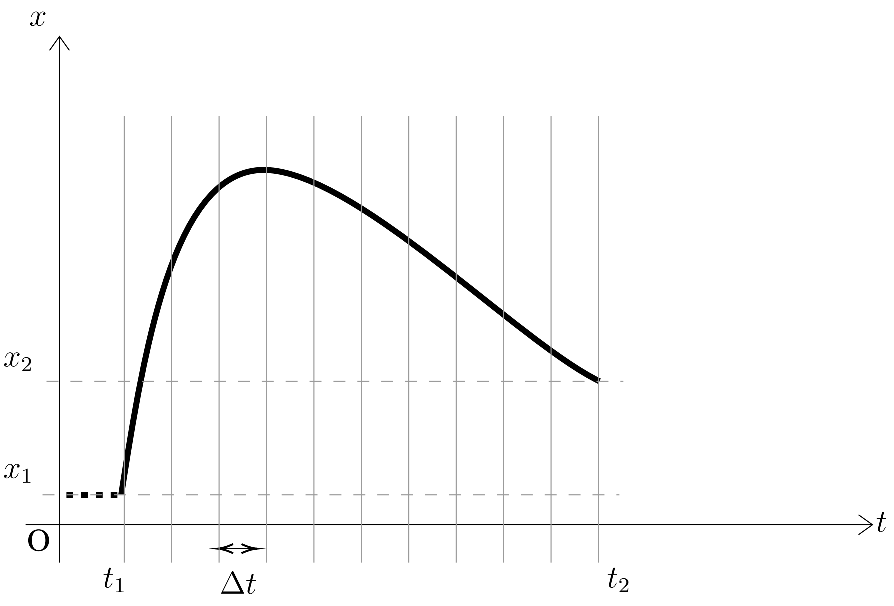
Another excellent example of the linear superposition of states is a remarkable application to the principle of least action!

To see this let us consider a quantum particle going from point M_1 at a moment t_1 to a point M_2 at a moment t_2. We know that the probability of finding a particle at a given point and at a given instant is related to the square of the module of the wave function associated with it. Let us put in the simplest case where the wave function of the particle is a undimensional plane wave $\Psi(x,t)$ given by the solution function of the Schrödinger equation of evolution:

$$\Psi(x,t)=e^{2\pi\mathrm{i}\left(\frac{x}{\lambda}-\nu t\right)} \tag{11.2.819}$$

where λ and ν are respectively the wavelength and frequency of the wave associated with the particle.

The particle can take an infinite number of paths to get from (M_1,t_1) to (M_2,t_2). Let us choose any one of these paths which we shall denote by the letter C. We can cut the path C into an integer number of lengths $\mathrm{d}t$.



After having travelled the first part of the path, the wave function has the following value:

$$\Psi(x_1+\Delta x,t_1+\Delta t)=e^{2\pi\mathrm{i}\left(\frac{x_1+\Delta x}{\lambda}-\nu(t_1+\Delta t)\right)}=e^{2\pi\mathrm{i}\left(\frac{x_1}{\lambda}-\nu t_1\right)}e^{2\pi\mathrm{i}\left(\frac{\Delta x}{\lambda}-\nu\Delta t\right)} \tag{11.2.820}$$

From this we get that:

$$\Psi(x_1+\Delta x,t_1+\Delta t)=\Psi(x_1,t_1)e^{2\pi\mathrm{i}\left(\frac{\Delta x}{\lambda}-\nu\Delta t\right)} \tag{11.2.821}$$

But, Planck and de Broglie have established the following relations as we have showed it earlier above:

$$E=h\nu \qquad \text{and} \qquad \lambda=\frac{h}{p} \tag{11.2.822}$$

hence, by replacing λ and ν in the preceding relation, we get:

$$\Psi(x_1+\Delta x,t_1+\Delta t)=\Psi(x_1,t_1)e^{\frac{2\pi\mathrm{i}}{h}\left(p\Delta x-E\Delta t\right)}=\Psi(x_1,t_1)\cdot e^{\frac{2\pi\mathrm{i}}{h}\left[\left(p\frac{\Delta x}{\Delta t}-E\right)\Delta t\right]} \tag{11.2.823}$$

Applying the same technique for the following part we get:

$$\begin{aligned}\Psi(x_1+2\Delta x,t_1+2\Delta t)&=\Psi(x_1+\Delta x,t_1+\Delta t)e^{\frac{2\pi\mathrm{i}}{h}\left[\left(p\frac{\Delta x}{\Delta t}-E\right)\Delta t\right]}\\&=\Psi(x_1,t_1)e^{\frac{2\pi\mathrm{i}}{h}\left[\left(p\left(\frac{\Delta x}{\Delta t}\right)_{\text{part 1}}-E\right)\Delta t\right]}\cdot e^{\frac{2\pi\mathrm{i}}{h}\left[\left(p\left(\frac{\Delta x}{\Delta t}\right)_{\text{part 2}}-E\right)\Delta t\right]}\end{aligned} \tag{11.2.824}$$

Proceeding like this from part to part along the path C then we get the value of the wave function on (M_2,t_2) for the particle coming from (M_1,t_1) by following the path C:

$$\Psi_C(x_2,t_2)=\Psi(x_1,t_1)\prod_{t_1}^{t_2}e^{\frac{2\pi\mathrm{i}}{h}\left[\left(p\frac{\Delta x}{\Delta t}-E\right)\Delta t\right]}=\Psi(x_1,t_1)e^{\frac{2\pi\mathrm{i}}{h}\left[\sum\limits_{t_1}^{t_2}\left(p\frac{\Delta x}{\Delta t}-E\right)\Delta t\right]} \tag{11.2.825}$$

Now, let the take the limit of the duration Δt of each trajectory part to zero such that $\Delta t\rightarrow \mathrm{d}t$. The quantity $\Delta x/\Delta t$ then tends to the instantaneous velocity $\mathrm{d}x/\mathrm{d}t$ of the particle that we will denote by $\dot{x}$. The preceding relation then becomes:

$$\Psi_C(x_2,t_2)=\Psi(x_1,t_1)\cdot e^{\frac{2\pi\mathrm{i}}{h}\left[\int\limits_{t_1}^{t_2}\left(p\dot{x}-E\right)\mathrm{d}t\right]} \tag{11.2.826}$$

In the section of Analytical Mechanics, we have shown that the quantity $p\dot{x}-E$ is equal to the Lagrangian L. By substituting the Lagrangian in the preceding relation, we get:

$$\boxed{\Psi_C(x_2,t_2)=\Psi(x_1,t_1)\cdot e^{\frac{2\pi\mathrm{i}}{h}\left[\int\limits_{t_1}^{t_2}L(t)\mathrm{d}t\right]}=\Psi(x_1,t_1)e^{\frac{2\pi\mathrm{i}}{h}\cdot S_C}} \tag{11.2.827}$$

where S_C is the action of the particle that has travelled the path C. Notice (without proof but we can detail on demand) that the module $\Psi_C(x_2,t_2)$ takes the same value for:

$$S_C=2n\pi\hbar=nh \tag{11.2.828}$$

for all $n\in\mathbb{N}$. The Planck constant then finds a physical significance directly related to the action of the particle!

Let us recall the de Broglie condition of normalization:

$$|\Psi_c(x_2,t_2)|^2 \tag{11.2.829}$$

which therefore gives the probability that the particle, starting from x_1 at the instant t_1, is in x_2 at the instant t_2, having travelled the path C.

The total probability is thus:

$$|\Psi_{\text{total}}(x_1,t_1\rightarrow x_2,t_2)|^2 \tag{11.2.830}$$

To calculate the probability that the particle going from x_1 at the instant t_1 to x_2 at the instant t_2 requires to makes the sum of the contributions of each path either (by applying the principle of linear superposition since we perform a sum of the wave functions):

$$\Psi_{\text{total}}(x_1,t_1\rightarrow x_2,t_2)|=\int\limits_{\text{all paths}}\Psi_C(x_2,t_2)=\int\limits_{\text{all paths}}\Psi_c(x_1,t_1)e^{\frac{2\pi\mathrm{i}}{h}S_C}=\int\limits_{\text{all paths}}\Psi_c(x_1,t_1)\cdot e^{\mathrm{i}\frac{S_C}{\hbar}} \tag{11.2.831}$$

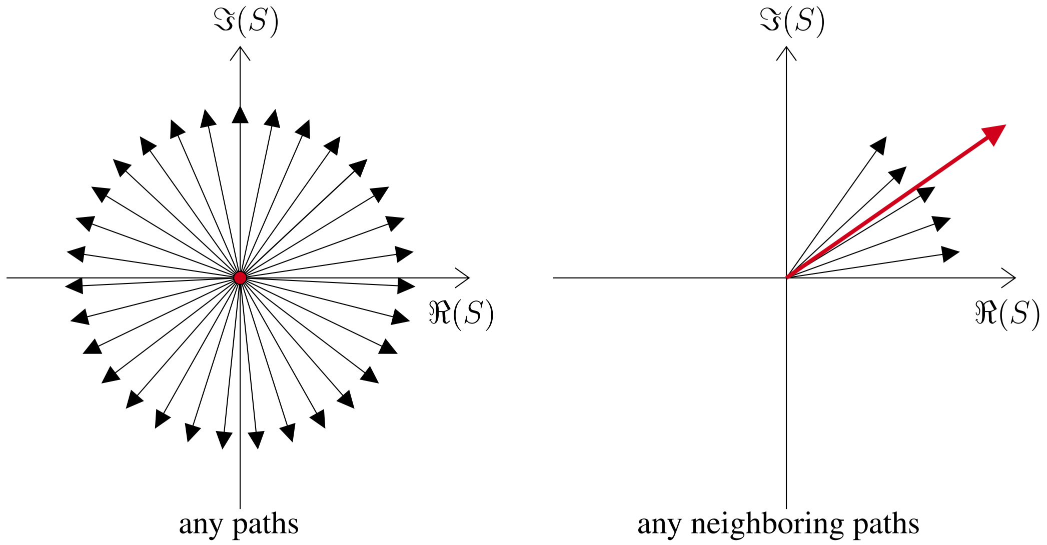
That latter equality is most of written as:

$$\boxed{\Psi(a,b)=\int\limits_a^b e^{\frac{\mathrm{i}S_C}{\hbar}}\mathcal{D}x=\int\limits_a^b e^{\frac{\mathrm{i}\int\limits_0^T L\mathrm{d}t}{\hbar}}\mathcal{D}x} \qquad (11.2.832)$$

where $\mathcal{D}x$ denotes integration over all paths.

This integral was discovered by Richard Feynman and is named "Feynman path integral" or sometimes "sum over histories method". In the first analysis it seems to diverge as there exist an infinity of possible paths between two points. Let's take a closer look at what's going on. Let us place ourselves in the case where the trajectory is macroscopic. The value of the action S_c is then much greater than $\hbar$ and varies greatly from one path to another, except for the paths close to the classical physical path for which the variation is almost zero (application of the variational statement of the principle Of lesser action).

Since the actions of paths intervene as a phase in the path integral above, their contributions are destructive and therefore tend to cancel out, except in the case of paths close to the classical physical path where contributions are added. It follows that the integral of path takes the value of the classical action, indicating that quantum physics makes it possible to find the laws of classical mechanics on a macroscopic scale.



any paths

any neighboring paths

Figure 11.66: Representation of the phases according to the type of path

The situation becomes very different at the quantum scale, that is, for values of the action S_C whose order of magnitude is that of the constant $\hbar$. An infinity of paths then brings non-destructive contributions. Feynman was able to prove that the path integral converged but on the other hand, it is no longer possible to predict which path the particle will take to the point that the concept of path itself vanishes. Thus on the quantum scale the particle seems to seek its way among all those that are possible but on a macroscopic scale, this quantum trial and error seems to have allowed the particle to find the "right path".

The path integral formalism is a very original way of approaching and interpreting quantum physics that has been added to those developed by Schrödinger.

Remark

A common question on Internet is: *Are quantum entanglement and superposition principle the same*? In fact, both are different concepts. Superposition of two state means a quantum system is in two state at a time. But, entanglement says the correlation of two or more system in a ensemble. Which means even if two system are spatially separated the measurement of any observable will be effected by the other.

Quantum fluctuation is inherent in the path integral formulation of quantum mechanics, where the path integral calculates a particle's quantum amplitude by summing over all possible paths it could take, not just the classical one. These "fluctuations" around the classical trajectory are the infinite, quantum-mechanically allowed paths that contribute to the final probability amplitude, which is dominated by paths near the classical one but still includes the probabilistic effects of quantum fluctuations.

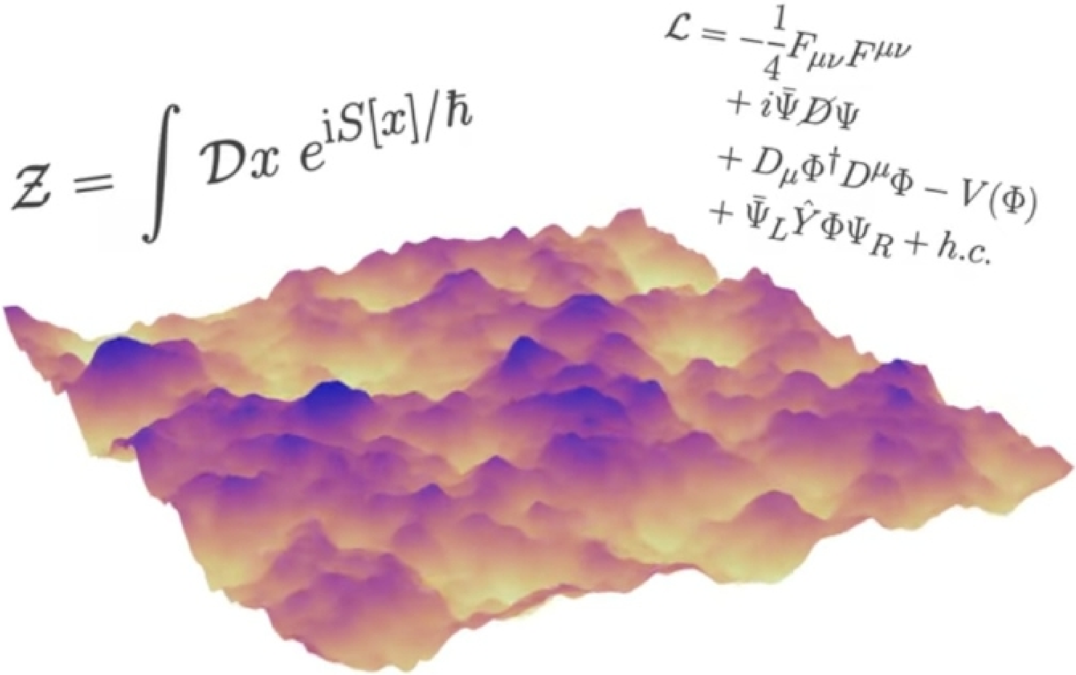

$\mathcal{Z} = \int \mathcal{D}x\, e^{iS[x]/\hbar}$
$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu}$
$+ i\bar{\Psi}\not{D}\Psi$
$+ D_\mu\Phi^\dagger D^\mu\Phi - V(\Phi)$
$+ \bar{\Psi}_L\hat{Y}\Phi\Psi_R + h.c.$

Figure 11.67: **Conceptual (artistic) representation of quantum fluctuations (source: ScienceClic, author: Alessandro Roussel)**

11.2.4.6 Ehrenfest theorems

This theorem makes it possible to connect Newton's Classical Mechanics to Quantum Physics by establishing similar relations with respect to the linear momentum and the force.

For this, we start from the special case of a massive particle moving at a non-relativistic velocity in a potential. We then have the one-dimensional Schrödinger equation of evolution:

$$\mathrm{i}\hbar\frac{\partial\Psi}{\partial t}=-\frac{\hbar^2}{2m}\frac{\partial^2\Psi}{\partial x^2}+E_p(x)\Psi \tag{11.2.833}$$

from which we get (useful for further below):

$$\frac{\partial\Psi}{\partial t}=\frac{\mathrm{i}}{\hbar}\left(-\frac{\hbar^2}{2m}\frac{\partial^2\Psi}{\partial x^2}+E_p(x)\Psi\right) \tag{11.2.834}$$

If we take the complex conjugate on both sides of the equality and multiply the two members by $\hbar$:

$$\mathrm{i}\hbar\frac{\partial\bar{\Psi}}{\partial t}=-\frac{\hbar^2}{2m}\frac{\partial^2\bar{\Psi}}{\partial x^2}+E_p(x)\bar{\Psi} \tag{11.2.835}$$

from which we get (useful also for further below):

$$\frac{\partial\bar{\Psi}}{\partial t}=\frac{\mathrm{i}}{\hbar}\left(-\frac{\hbar^2}{2m}\frac{\partial^2\bar{\Psi}}{\partial x^2}+E_p(x)\bar{\Psi}\right) \tag{11.2.836}$$

Let us consider the temporal variation of the mean position of the particle (5th postulate):

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\langle x\rangle=\frac{\mathrm{d}}{\mathrm{d}t}\int_{-\infty}^{+\infty}\bar{\Psi}x\Psi\mathrm{d}x=\int_{-\infty}^{+\infty}\frac{\mathrm{d}}{\mathrm{d}t}\left(\bar{\Psi}x\Psi\right)\mathrm{d}x=\int_{-\infty}^{+\infty}\left(\frac{\partial\bar{\Psi}}{\partial t}x\Psi+\bar{\Psi}x\frac{\partial\Psi}{\partial t}\right)\mathrm{d}x\\
&=\int_{-\infty}^{+\infty}\frac{\partial\bar{\Psi}}{\partial t}x\Psi\mathrm{d}x+\int_{-\infty}^{+\infty}\bar{\Psi}x\frac{\partial\Psi}{\partial t}\mathrm{d}x=\int_{-\infty}^{+\infty}\left(\frac{\partial\bar{\Psi}}{\partial t}\right)x\Psi\mathrm{d}x+\int_{-\infty}^{+\infty}\bar{\Psi}x\left(\frac{\partial\Psi}{\partial t}\right)\mathrm{d}x\\
&=\frac{\mathrm{i}}{\hbar}\left[\int_{-\infty}^{+\infty}\left(-\frac{\hbar^2}{2m}\frac{\partial^2\bar{\Psi}}{\partial x^2}+E_p(x)\bar{\Psi}\right)x\Psi\mathrm{d}x-\int_{-\infty}^{+\infty}\bar{\Psi}x\left(-\frac{\hbar^2}{2m}\frac{\partial^2\Psi}{\partial x^2}+E_p(x)\Psi\right)\mathrm{d}x\right]\\
&=\frac{\mathrm{i}\hbar}{2m}\left[\int_{-\infty}^{+\infty}\left(-\frac{\partial^2\bar{\Psi}}{\partial x^2}\right)x\Psi\mathrm{d}x-\int_{-\infty}^{+\infty}\bar{\Psi}x\left(-\frac{\partial^2\Psi}{\partial x^2}\right)\mathrm{d}x\right]=\frac{\mathrm{i}\hbar}{2m}\int_{-\infty}^{+\infty}\left(\bar{\Psi}x\frac{\partial^2\Psi}{\partial x^2}-\frac{\partial^2\bar{\Psi}}{\partial x^2}x\Psi\right)\mathrm{d}x
\end{aligned} \tag{11.2.837}$$

We have:

$$\frac{\partial^2\left(x\Psi\right)}{\partial x^2}=\frac{\partial}{\partial x}\left(\frac{\partial x}{\partial x}\Psi+x\frac{\partial\Psi}{\partial x}\right)=\left(\frac{\partial^2 x}{\partial x^2}\Psi+\frac{\partial x}{\partial x}\frac{\partial\Psi}{\partial x}+\frac{\partial x}{\partial x}\frac{\partial\Psi}{\partial x}+x\frac{\partial^2\Psi}{\partial x^2}\right) \tag{11.2.838}$$

hence:

$$\frac{\partial^2\left(x\Psi\right)}{\partial x^2}-2\frac{\partial\Psi}{\partial x}=x\frac{\partial^2\Psi}{\partial x^2} \tag{11.2.839}$$

Let us use this last relation:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\langle x\rangle&=\frac{\mathrm{i}\hbar}{2m}\int\limits_{-\infty}^{+\infty}\left(\bar{\Psi}x\frac{\partial^2\Psi}{\partial x^2}-\frac{\partial^2\bar{\Psi}}{\partial x^2}x\Psi\right)\mathrm{d}x=\frac{\mathrm{i}\hbar}{2m}\int\limits_{-\infty}^{+\infty}\left(\bar{\Psi}\left(\frac{\partial^2(x\Psi)}{\partial x^2}-2\frac{\partial\Psi}{\partial x}\right)-\frac{\partial^2\bar{\Psi}}{\partial x^2}x\Psi\right)\mathrm{d}x\\&=\frac{\mathrm{i}\hbar}{2m}\int\limits_{-\infty}^{+\infty}\left(\bar{\Psi}\frac{\partial^2(x\Psi)}{\partial x^2}-\frac{\partial^2\bar{\Psi}}{\partial x^2}(x\Psi)\right)\mathrm{d}x-\frac{\mathrm{i}\hbar}{2m}\int\limits_{-\infty}^{+\infty}\bar{\Psi}\frac{\partial\Psi}{\partial x}\mathrm{d}x\end{aligned}\tag{11.2.840}$$

Let us now use the relation:

$$\bar{\Psi}\frac{\partial^2(x\Psi)}{\partial x^2}-\frac{\partial^2\bar{\Psi}}{\partial x^2}(x\Psi)=\frac{\partial}{\partial x}\left(\bar{\Psi}\frac{\partial(x\Psi)}{\partial x}-\frac{\partial\bar{\Psi}}{\partial x}(x\Psi)\right)\tag{11.2.841}$$

and let us inject it into the prior-previous relation:

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle x\rangle=\frac{\mathrm{i}\hbar}{2m}\int\limits_{-\infty}^{+\infty}\frac{\partial}{\partial x}\left(\bar{\Psi}\frac{\partial(x\Psi)}{\partial x}-\frac{\partial\bar{\Psi}}{\partial x}(x\Psi)\right)\mathrm{d}x+\frac{1}{m}\int\limits_{-\infty}^{+\infty}\bar{\Psi}\left(-i\hbar\frac{\partial\Psi}{\partial x}\right)\mathrm{d}x\tag{11.2.842}$$

The first term to the right of the equality is easy to integrate ... (since there is no need to integrate it...):

$$\int\limits_{-\infty}^{+\infty}\frac{\partial}{\partial x}\left(\bar{\Psi}\frac{\partial(x\Psi)}{\partial x}-\frac{\partial\bar{\Psi}}{\partial x}(x\Psi)\right)\mathrm{d}x=\left[\bar{\Psi}\frac{\partial(x\Psi)}{\partial x}-\frac{\partial\bar{\Psi}}{\partial x}(x\Psi)\right]\Bigg|_{-\infty}^{+\infty}\tag{11.2.843}$$

and since the wave function must be 0 at $x=\pm\infty$ (otherwise the energy is infinite) then this last relation is equal zero. We then have:

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle x\rangle=\frac{1}{m}\int\limits_{-\infty}^{+\infty}\bar{\Psi}\left(-\mathrm{i}\hbar\frac{\partial\Psi}{\partial x}\right)\mathrm{d}x\tag{11.2.844}$$

Thus:

$$m\frac{\mathrm{d}\langle x\rangle}{\mathrm{d}t}=\int\limits_{-\infty}^{+\infty}\bar{\Psi}\left(-\mathrm{i}\hbar\frac{\partial}{\partial x}\right)\Psi\mathrm{d}x\tag{11.2.845}$$

and finally:

$$\boxed{m\frac{\mathrm{d}\langle x\rangle}{\mathrm{d}t}=\int\limits_{-\infty}^{+\infty}\bar{\Psi}p_x\Psi\mathrm{d}x=\langle p_x\rangle}\tag{11.2.846}$$

Which is the equivalent in Classical Mechanics of:

$$m\frac{\mathrm{d}x}{\mathrm{d}t}=mv=p\tag{11.2.847}$$

and which reconfirms the existence of the mathematical entity:

$$p_x=-\mathrm{i}\hbar\frac{\partial}{\partial x}=\frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial x} \tag{11.2.848}$$

as being the linear momentum operator, and which we determined earlier by falling back on Newton's second law!

But we can do a little better at the level of the classic/quantum analogy by derivating:

$$\langle p_x\rangle=\int\limits_{-\infty}^{+\infty}\bar{\Psi}p_x\Psi\mathrm{d}x \tag{11.2.849}$$

relatively to time, which gives:

$$\frac{\mathrm{d}\langle p_x\rangle}{\mathrm{d}t}=\frac{\mathrm{d}}{\mathrm{d}t}\int\limits_{-\infty}^{+\infty}\bar{\Psi}p_x\Psi\mathrm{d}x=-\mathrm{i}\hbar\int\limits_{-\infty}^{+\infty}\frac{\mathrm{d}}{\mathrm{d}t}\left(\bar{\Psi}\frac{\partial\Psi}{\partial x}\right)\mathrm{d}x=-\mathrm{i}\hbar\int\limits_{-\infty}^{+\infty}\left(\frac{\partial\bar{\Psi}}{\partial t}\frac{\partial\Psi}{\partial x}+\bar{\Psi}\frac{\partial}{\partial t}\frac{\partial\Psi}{\partial x}\right)\mathrm{d}x \tag{11.2.850}$$

hence:

$$-\frac{1}{\mathrm{i}\hbar}\frac{\mathrm{d}\langle p_x\rangle}{\mathrm{d}t}=\int\limits_{-\infty}^{+\infty}\frac{\partial\bar{\Psi}}{\partial t}\frac{\partial\Psi}{\partial x}\mathrm{d}x+\int\limits_{-\infty}^{+\infty}\bar{\Psi}\frac{\partial}{\partial t}\frac{\partial\Psi}{\partial x}\mathrm{d}x=\int\limits_{-\infty}^{+\infty}\frac{\partial\bar{\Psi}}{\partial t}\frac{\partial\Psi}{\partial x}\mathrm{d}x+\int\limits_{-\infty}^{+\infty}\bar{\Psi}\frac{\partial}{\partial x}\frac{\partial\Psi}{\partial t}\mathrm{d}x \tag{11.2.851}$$

Using:

$$\mathrm{i}\hbar\frac{\partial\Psi}{\partial t}=H\Psi\Leftrightarrow\mathrm{i}\hbar\frac{\partial\Psi}{\partial t}=\left(-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}+E_p(x)\right)\Psi \tag{11.2.852}$$

It comes:

$$\begin{aligned}-\frac{1}{\mathrm{i}\hbar}\frac{\mathrm{d}\langle p_x\rangle}{\mathrm{d}t}&=-\frac{1}{\mathrm{i}\hbar}\int\limits_{-\infty}^{+\infty}H\bar{\Psi}\frac{\partial\Psi}{\partial x}\mathrm{d}x+\frac{1}{\mathrm{i}\hbar}\int\limits_{-\infty}^{+\infty}\bar{\Psi}\frac{\partial(H\Psi)}{\partial x}\mathrm{d}x\\&=-\frac{1}{\mathrm{i}\hbar}\int\limits_{-\infty}^{+\infty}\left(-\frac{\hbar^2}{2m}\frac{\partial^2\bar{\Psi}}{\partial x^2}+E_p(x)\bar{\Psi}\right)\frac{\partial\Psi}{\partial x}\mathrm{d}x+\frac{1}{\mathrm{i}\hbar}\int\limits_{-\infty}^{+\infty}\bar{\Psi}\frac{\partial\left(-\frac{\hbar^2}{2m}\frac{\partial^2\Psi}{\partial x^2}+E_p(x)\Psi\right)}{\partial x}\mathrm{d}x\\&=\frac{1}{\mathrm{i}\hbar}\left[-\int\limits_{-\infty}^{+\infty}\left(-\frac{\hbar^2}{2m}\frac{\partial^2\bar{\Psi}}{\partial x^2}+E_p(x)\bar{\Psi}\right)\frac{\partial\Psi}{\partial x}\mathrm{d}x+\int\limits_{-\infty}^{+\infty}\bar{\Psi}\frac{\partial\left(-\frac{\hbar^2}{2m}\frac{\partial^2\Psi}{\partial x^2}+E_p(x)\Psi\right)}{\partial x}\mathrm{d}x\right]\\&=\frac{1}{\mathrm{i}\hbar}\left[\frac{\hbar^2}{2m}\int\limits_{-\infty}^{+\infty}\left(\frac{\partial^2\bar{\Psi}}{\partial x^2}\frac{\partial\Psi}{\partial x}-\bar{\Psi}\frac{\partial}{\partial x}\frac{\partial^2\Psi}{\partial x^2}\right)\mathrm{d}x+\int\limits_{-\infty}^{+\infty}\left(\bar{\Psi}\frac{\partial\big(E_p(x)\Psi\big)}{\partial x}-E_p(x)\bar{\Psi}\frac{\partial\Psi}{\partial x}\right)\mathrm{d}x\right]\end{aligned} \tag{11.2.853}$$

Let us focus on:

$$\int\limits_{-\infty}^{+\infty}\left(\frac{\partial^2\bar{\Psi}}{\partial x^2}\frac{\partial\Psi}{\partial x}-\bar{\Psi}\frac{\partial}{\partial x}\frac{\partial^2\Psi}{\partial x^2}\right)\mathrm{d}x \tag{11.2.854}$$

Let us integrate the first term twice according to the relation proved in the section of Differential and Integral Calculus:

$$\int\limits_a^b f(t)g'(t)\mathrm{d}t = f(t)g(t)\Big|_a^b - \int\limits_a^b f'(t)g(t)\mathrm{d}t \tag{11.2.855}$$

We then have (always considering Ψ as a decreasing function to infinity):

$$\int\limits_{-\infty}^{+\infty} \frac{\partial^2\bar{\Psi}}{\partial x^2}\frac{\partial\Psi}{\partial x}\mathrm{d}x = \int\limits_{-\infty}^{+\infty} \frac{\partial\Psi}{\partial x}\frac{\partial^2\bar{\Psi}}{\partial x^2}\mathrm{d}x = \frac{\partial\Psi}{\partial x}\frac{\partial\bar{\Psi}}{\partial x}\Bigg|_{-\infty}^{+\infty} - \int\limits_{-\infty}^{+\infty} \frac{\partial}{\partial x}\frac{\partial\Psi}{\partial x}\frac{\partial\bar{\Psi}}{\partial x}\mathrm{d}x = -\int\limits_{-\infty}^{+\infty} \frac{\partial^2\Psi}{\partial^2 x}\frac{\partial\bar{\Psi}}{\partial x}\mathrm{d}x \tag{11.2.856}$$

and once again:

$$-\int\limits_{-\infty}^{+\infty} \frac{\partial^2\Psi}{\partial^2 x}\frac{\partial\bar{\Psi}}{\partial x}\mathrm{d}x = -\frac{\partial^2\Psi}{\partial^2 x}\bar{\Psi}\Bigg|_{-\infty}^{+\infty} + \int\limits_{-\infty}^{+\infty} \frac{\partial}{\partial x}\frac{\partial^2\Psi}{\partial^2 x}\bar{\Psi}\mathrm{d}x = \int\limits_{-\infty}^{+\infty} \frac{\partial}{\partial x}\frac{\partial^2\Psi}{\partial^2 x}\bar{\Psi}\mathrm{d}x \tag{11.2.857}$$

So finally:

$$\int\limits_{-\infty}^{+\infty} \left(\frac{\partial^2\bar{\Psi}}{\partial x^2}\frac{\partial\Psi}{\partial x} - \bar{\Psi}\frac{\partial}{\partial x}\frac{\partial^2\Psi}{\partial x^2}\right)\mathrm{d}x = \int\limits_{-\infty}^{+\infty} \left(\frac{\partial}{\partial x}\frac{\partial^2\Psi}{\partial^2 x}\bar{\Psi} - \bar{\Psi}\frac{\partial}{\partial x}\frac{\partial^2\Psi}{\partial x^2}\right)\mathrm{d}x = 0 \tag{11.2.858}$$

We then have:

$$\begin{aligned}
\frac{\mathrm{d}\langle p_x\rangle}{\mathrm{d}t} &= -\left[\int\limits_{-\infty}^{+\infty} \left(\bar{\Psi}\frac{\partial\big(E_p(x)\Psi\big)}{\partial x} - E_p(x)\bar{\Psi}\frac{\partial\Psi}{\partial x}\right)\mathrm{d}x\right] \\
&= -\left[\int\limits_{-\infty}^{+\infty} \left(\bar{\Psi}\frac{\partial\big(E_p(x)\big)}{\partial x}\Psi + \cancel{\bar{\Psi}E_p(x)\frac{\partial\Psi}{\partial x}} - \cancel{E_p(x)\bar{\Psi}\frac{\partial\Psi}{\partial x}}\right)\mathrm{d}x\right] = -\int\limits_{-\infty}^{+\infty} \bar{\Psi}\frac{\partial E_p(x)}{\partial x}\Psi\mathrm{d}x \\
&= -\int\limits_{-\infty}^{+\infty} \bar{\Psi}\left(-\frac{dE_p}{\mathrm{d}x}\right)\Psi\mathrm{d}x
\end{aligned} \tag{11.2.859}$$

However, we have demonstrated in the section of Classical Mechanics that:

$$\vec{F} = -\vec{\nabla}E_p \tag{11.2.860}$$

It therefore comes that:

$$\boxed{\frac{\mathrm{d}\langle p_x\rangle}{\mathrm{d}t} = \int\limits_{-\infty}^{+\infty} \bar{\Psi}F_x\Psi\mathrm{d}x} = \langle F_x\rangle \tag{11.2.861}$$

This extraordinarily simple result constitutes the "Ehrenfest theorem". We thus find the fundamental law of Classical Dynamics in the sense of mean values of position and force, calculated using the probability of presence!

11.2.4.6.1 Generalized Ehrenfest Theorem

The text that follows is almost a copy/paste of the content of the video on YouTube made by Sam Dietterich!

When we think of the Ehrenfest Theorem, we usually think of these two equations:

$$m\frac{\mathrm{d}}{\mathrm{d}t}\langle x\rangle=\langle p\rangle \quad \frac{\mathrm{d}}{\mathrm{d}t}\langle p\rangle=\left\langle -\frac{\partial}{\partial x}V(x)\right\rangle \tag{11.2.862}$$

The reader has probably seen the long derivations for each of them earlier above. This may leave the reader thinking two things:

- First, are there analogous equations for other observables, or even a general one for an arbitrary observable?
- And second, we would expect that the derivation of such a general equation would have to be at least as long as the derivation of these two equations.

It turns out that there is a Generalized Ehrenfest Theorem equation, but its derivation is actually shorter then either of the above! It turns out that when one does the general derivation with abstract notation instead of appealing to the position representation, then the derivation is extremely short. It is simply the following:

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle Q\rangle=\frac{\mathrm{d}}{\mathrm{d}t}\langle\psi|Q|\psi\rangle=\left[\frac{\partial}{\partial t}\langle\psi|\right]Q|\psi\rangle+\langle\psi|\frac{\partial Q}{\partial t}|\psi\rangle+\langle\psi|Q\frac{\partial}{\partial t}|\psi\rangle \tag{11.2.863}$$

But we have from Schrödinger equation:

$$\begin{matrix}\mathrm{i}\hbar\frac{\partial}{\partial t}|\psi\rangle=H|\psi\rangle & \rightarrow & \frac{\partial}{\partial t}|\psi\rangle=-\frac{\mathrm{i}}{\hbar}H|\psi\rangle\\ -\mathrm{i}\hbar\frac{\partial}{\partial t}\langle\psi|=\langle\psi|H & \rightarrow & \frac{\partial}{\partial t}\langle\psi|=\frac{\mathrm{i}}{\hbar}\langle\psi|H\end{matrix} \tag{11.2.864}$$

So substitute that and rearranging give us:

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle Q\rangle=\frac{\mathrm{d}}{\mathrm{d}t}\langle\psi|Q|\psi\rangle=\left[\frac{\mathrm{i}}{\hbar}\langle\psi|H\right]Q|\psi\rangle-\frac{\mathrm{i}}{\hbar}\langle\psi|QH|\psi\rangle+\langle\psi|\frac{\partial Q}{\partial t}|\psi\rangle \tag{11.2.865}$$

The Hamiltonian is Hermitian as it is given by:

$$H:=-\frac{\hbar^2}{2m}\Delta+V(\vec{x}) \tag{11.2.866}$$

and can therefore act equivalently to the left or to the right. Therefore we can drop the brackets:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\langle Q\rangle&=\frac{\mathrm{i}}{\hbar}\langle\psi|HQ|\psi\rangle-\frac{\mathrm{i}}{\hbar}\langle\psi|QH|\psi\rangle+\langle\psi|\frac{\partial Q}{\partial t}|\psi\rangle\\&=\frac{\mathrm{i}}{\hbar}\langle\psi|(HQ-QH)|\psi\rangle+\langle\psi|\frac{\partial Q}{\partial t}|\psi\rangle\\&=\frac{\mathrm{i}}{\hbar}\langle\psi|[H,Q]|\psi\rangle+\langle\psi|\frac{\partial Q}{\partial t}|\psi\rangle\end{aligned} \tag{11.2.867}$$

Therefore we get the "general Ehrenfest theorem":

$$\boxed{\frac{\mathrm{d}}{\mathrm{d}t}\langle Q\rangle=\frac{\mathrm{i}}{\hbar}\langle[H,Q]\rangle+\left\langle\frac{\partial Q}{\partial t}\right\rangle} \tag{11.2.868}$$

Then all we have to do to derive ehrenfest theorem for any special case is evaluate the commutator of the hamiltonian with that operator and the time derivative if it's explicitly time dependent and this does easily reproduce the results we had for the position and momentum Ehrenfest theorem that we derived earlier above!

This result will be useful to us to derive rigorously the Energy-Time Heisenberg's uncertainty principle (see page 4739)!

But let us check if we can fall back, on the previous results!

We start with:

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle\hat{x}\rangle=\frac{1}{i\hbar}\langle[\hat{x},\hat{H}]\rangle \tag{11.2.869}$$

And remember that:

$$H=\frac{p^2}{2m}+V(x) \tag{11.2.870}$$

Therefore:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\langle\hat{x}\rangle&=\frac{1}{\mathrm{i}\hbar}\left\langle\left[\hat{x},\frac{\hat{p}^2}{2m}\right]\right\rangle=\frac{1}{\mathrm{i}\hbar 2m}\langle[\hat{x},\hat{p}\hat{p}]\rangle\\&=\frac{1}{\mathrm{i}\hbar 2m}\langle[\hat{x},\hat{p}]\hat{p}+\hat{b}[\hat{x},\hat{p}]\rangle=\frac{1}{\mathrm{i}\hbar 2m}\langle[\hat{x},\hat{p}]\hat{p}+\hat{p}[\hat{x},\hat{p}]\rangle\\&=\frac{1}{\mathrm{i}\hbar 2m}\langle\mathrm{i}\hbar\hat{p}+\mathrm{i}\hbar\hat{p}\rangle=\frac{1}{m}\langle\hat{p}\rangle\end{aligned} \tag{11.2.871}$$

And now:

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle\hat{p}\rangle=\frac{1}{\mathrm{i}\hbar}\langle[\hat{p},\hat{H}]\rangle \tag{11.2.872}$$

always keeping in mind that:

$$H=\frac{p^2}{2m}+V(x) \tag{11.2.873}$$

Therefore:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\langle\hat{p}\rangle\varphi&=\frac{1}{\mathrm{i}\hbar}\left\langle\left[-\mathrm{i}\hbar\frac{\partial}{\partial x},V(x)\right]\varphi\right\rangle=-\left\langle\left[\frac{\partial}{\partial x},V(x)\right]\varphi\right\rangle\\&=-\left\langle\frac{\partial(V\varphi)}{\partial x}-V\frac{\partial\varphi}{\partial x}\right\rangle=-\left\langle V\frac{\partial\varphi}{\partial x}+\varphi\frac{\partial}{\partial x}V-V\frac{\partial\varphi}{\partial x}\right\rangle\\&=-\varphi\left\langle\frac{\partial V}{\partial x}\right\rangle\end{aligned} \tag{11.2.874}$$

Hence:

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle\hat{p}\rangle=-\left\langle\frac{\partial V}{\partial x}\right\rangle=\langle F\rangle \qquad (11.2.875)$$

There is no direct "empirical evidence" for the Ehrenfest theorem itself, as it is a mathematical proof showing the link between quantum and classical mechanics, not a prediction that can be directly tested with a single experiment. Instead, its "evidence" is the successful and consistent way quantum mechanics (and thus the Ehrenfest theorem) can be used to predict the behavior of quantum systems, which then transitions to the predictions of classical mechanics in the macroscopic limit, as explained by the correspondence principle. Experiments that confirm the predictions of quantum mechanics for microscopic particles serve as indirect validation for the theorem's derivation.

11.2.5 Angular momentum and Spin

In classical physics, angular momentum is commonly divided into two types: orbital and rotational. In the case of the Earth, for example, the orbital angular momentum is associated with the revolution of the Earth about the Sun, while rotational angular momentum is associated with the Earth rotating on its axis. Rotational angular momentum is often known as spin, for fairly obvious reasons.

In classical physics, of course, there is really only one type of angular momentum, since both the orbital and rotational types are associated with the rotation of masses about some axis. The division into the two types is purely a computational convenience.

In quantum physics, the situation is quite different is quite different as we will see (it is important that the reader refers to the Corpuscular Quantum Physics section where we have proved the it is seems not acceptable ton consider the electron spinning on itself).

Like the harmonic oscillator, the notion of angular momentum is of main importance in quantum theory and has many applications in all fields of physics: atomic and molecular physics, nuclear and subnuclear physics, physics of condensed state, etc. Thus, it plays an essential role in the study of the motion of a particle in a potential with spherical symmetry, as we shall see in the section of Quantum Chemistry (which is an excellent practical example). The angular momentum is also at the base of the group of rotations (see section Set Algebra page 1424) which satisfies the algebra of the angular momentum operators. Thus, it not only allows to construct the wave function of a given quantum system of symmetry, but also to predict if an optical transition is allowed and to determine its intensity (for example, during the study optical transitions between impurity states (in solid state), molecular states (quantum chemistry), nuclear physics, etc.).

Finally, we will see that the algebraic method applied to the study of angular momentum will allow us to introduce quite naturally the notion of intrinsic angular momentum of a particle, i.e. the "spin", which has in fact no classical equivalent (so it's not an angular momentum!).

The following developments may seem rather disconcerting in the sense that it is no longer necessary to rely on intuition but only on the properties and results of mathematics. As usual, if the reader need additional information, he must not hesitate to contact us.

Remark

"Spintronics" (a portmanteau meaning "spin transport electronics"), also known as "spinelectronics" or "fluxtronics", is the study of the intrinsic spin of the electron and its associated magnetic moment, in addition to its fundamental electronic charge, in solid-state devices or in Quantum computers. The simplest method of generating a spin-polarised current in a metal is to pass the current through a ferromagnetic material. The most common applications of this effect involve giant magnetoresistance (GMR) devices. For more, see the excellent introductory textbook [40].

To begin, let us recall that the angular momentum of a particle relatively to their origin is given by (see section Classical Mechanics page 3244):

$$\vec{b}=m\vec{r}\times\vec{v}=\vec{r}\times\vec{p} \tag{11.2.876}$$

Since the linear momentum is quantified (it is an eigenvalue related to energy in one way or another), the angular momentum is necessarily also quantified (the angular momentum is therefore an eigenvalue) and the experience has supported this result (Stern-Gerlach experiment).

Given z the component of the resulting vector product:

$$\boxed{b_z = xp_y - yp_x \quad \text{cycl.}} \tag{11.2.877}$$

This relationship being cyclic, we can change the indices to get the other coordinates.

Since x and y commute (in the sense that their commutator is equal to zero) and that we have proved earlier above that:

$$p_x = \mathrm{i}\hbar\frac{\partial}{\partial x} \qquad p_y = \mathrm{i}\hbar\frac{\partial}{\partial y} \tag{11.2.878}$$

We then have:

$$b_z = xp_y - yp_x = \frac{\hbar}{\mathrm{i}}x\frac{\partial}{\partial y} - \frac{\hbar}{\mathrm{i}}y\frac{\partial}{\partial x} = \frac{\hbar}{\mathrm{i}}\left(x\frac{\partial}{\partial y} - y\frac{\partial}{\partial x}\right) \tag{11.2.879}$$

Which gives:

$$b_z = \frac{\hbar}{\mathrm{i}}\left(x\frac{\partial}{\partial y} - y\frac{\partial}{\partial x}\right) \quad \text{cycl.} \tag{11.2.880}$$

Using the gradient (we will see again this relation in the section of Relativistic Quantum Physics during our study of the Pauli equation!!):

$$\boxed{\vec{b} = \frac{\hbar}{\mathrm{i}}\vec{r}\times\vec{\nabla} = -\mathrm{i}\hbar\vec{r}\times\vec{\nabla}} \tag{11.2.881}$$

and putting for the "operator of the orbital angular momentum":

$$\frac{1}{\mathrm{i}}\vec{r}\times\vec{\nabla} = \frac{1}{\hbar}\vec{r}\times\vec{p} = \vec{l} \tag{11.2.882}$$

Which lead us to write:

$$\boxed{\vec{b} = \hbar\vec{l}} \tag{11.2.883}$$

with:

$$\begin{pmatrix} l_x \\ l_y \\ l_z \end{pmatrix} = \frac{1}{\hbar}\begin{pmatrix} yp_z - zp_y \\ zp_x - xp_z \\ xp_y - yp_x \end{pmatrix} \tag{11.2.884}$$

Remark

Most often in the literature the orbital angular momentum is denoted $\vec{L}$ (we have already made this remark in the section of Classical Mechanics) but we avoided this notation here in order to differentiate the orbital angular momentum and the total orbital angular momentum.

We will establish some commutation relations concerning $\vec{l}$ which will play an essential role in our study of the spin. Using the following commutation relations (proved during our study of the Heisenberg's classical uncertainty principles page 4704):

$$[x,p_x]=\mathrm{i}\hbar \quad \text{cycl.} \tag{11.2.885}$$

and:

$$[y,p_x]=0 \quad \text{cycl.} \tag{11.2.886}$$

We have the relation (it is traditional to do the analysis on the component z of the projection of $\vec{l}$):

$$[\hbar l_z,x]=[xp_y,x]-[yp_x,x]=-y[p_x,x]=\mathrm{i}\hbar y \tag{11.2.887}$$

Therefore:

$$[l_z,x]=\mathrm{i}y \quad \text{cycl.} \tag{11.2.888}$$

and proceeding in the same way:

$$\boxed{[l_z,y]=-\mathrm{i}x \quad \text{cycl.}} \qquad \text{and} \qquad \boxed{[l_z,z]=0 \quad \text{cycl.}} \tag{11.2.889}$$

Remark

We find analogous relations with the linear momentum:

$$\begin{aligned}
&[l_z,p_x]=\mathrm{i}p_y\\
&[l_z,p_y]=-\mathrm{i}p_x \qquad \text{cycl.}\\
&[l_z,p_z]=0
\end{aligned} \tag{11.2.890}$$

Let ow now evaluate the quantity (following the request of a reader, we put all the details):

$$\begin{aligned}
[l_x,\hbar l_y]&=[l_x,zp_x-xp_z]=[l_x,zp_x]-[l_x,xp_z]=\left[\frac{1}{\hbar}\left(yp_z-zp_y\right),zp_x\right]-\left[\frac{1}{\hbar}\left(yp_z-zp_y\right),xp_z\right]\\
&=\left[\frac{1}{\hbar}yp_z,zp_x\right]-\left[\frac{1}{\hbar}zp_y,zp_x\right]-\left[\frac{1}{\hbar}yp_z,xp_z\right]+\left[\frac{1}{\hbar}zp_y,xp_z\right]\\
&=\left(\frac{1}{\hbar}yp_zzp_x-\frac{1}{\hbar}z\underbrace{p_xy}_{=-\mathrm{i}\hbar\frac{\partial}{\partial x}y=0}p_z\right)-\left(\frac{1}{\hbar}z\underbrace{p_yz}_{=-\mathrm{i}\hbar\frac{\partial}{\partial y}z=0}p_x-\frac{1}{\hbar}z\underbrace{p_xz}_{=-\mathrm{i}\hbar\frac{\partial}{\partial x}z=0}p_y\right)\\
&-\left(\frac{1}{\hbar}y\underbrace{p_zx}_{=-\mathrm{i}\hbar\frac{\partial}{\partial z}x=0}p_z-\frac{1}{\hbar}x\underbrace{p_zy}_{=-\mathrm{i}\hbar\frac{\partial}{\partial z}y=0}p_z\right)+\left(\frac{1}{\hbar}z\underbrace{p_yx}_{=-\mathrm{i}\hbar\frac{\partial}{\partial y}x=0}p_z-\frac{1}{\hbar}xp_zzp_y\right)\\
&=\frac{1}{\hbar}yp_zzp_x-\frac{1}{\hbar}xp_zzp_y=\frac{1}{\hbar}y\left(-\mathrm{i}\hbar\frac{\partial}{\partial z}z\right)p_x-\frac{1}{\hbar}x\left(-\mathrm{i}\hbar\frac{\partial}{\partial z}z\right)p_y=\frac{1}{\hbar}y(-\mathrm{i}\hbar)p_x-\frac{1}{\hbar}x(-\mathrm{i}\hbar)p_y\\
&=-\mathrm{i}yp_x+\mathrm{i}xp_y=\mathrm{i}\left(xp_y-yp_x\right)=\mathrm{i}\hbar l_z
\end{aligned} \tag{11.2.891}$$

Either after simplification (it is rather annoying for the experience that this does not commute):

$$\boxed{[l_x,l_y]=\mathrm{i}l_z} \tag{11.2.892}$$

By the way, at this stage, the reader that has already covered the Spinor Calculus section beforehand (if not already done we strongly recommend the reader to go take a look at it!), he will notice that the Pauli matrices satisfy the preceding relations if we put ourselves in natural units (the reduced Planck constant then being equal to 1):

$$\begin{aligned}\sigma_x\sigma_y-\sigma_y\sigma_x&=2\mathrm{i}\sigma_z\\ \sigma_y\sigma_z-\sigma_z\sigma_y&=2\mathrm{i}\sigma_x\\ \sigma_z\sigma_x-\sigma_x\sigma_z&=2\mathrm{i}\sigma_y\end{aligned} \tag{11.2.893}$$

This observation will be useful for our study of Relativistic Quantum Physics (see section of the same name page 4919). Indeed, we know from our study of Spinor Calculus that the 2×2 matrices in $\mathbb{C}$ of determinant 1 form the group of rotations in the space $\mathrm{SU}(2)$, whose Pauli matrices are the generators. Basically, the origin of the spin comes from the link between $\mathrm{SU}(2)$ and the group of rotations of our ordinary space, $\mathrm{SO}(3)$ (see section Set Algebra page 1424). Now, let us consider the norm:

$$\|\vec{l}\|^2=l_x^2+l_y^2+l_z^2 \tag{11.2.894}$$

where we must consider the square of one of these operators in the following form:

$$l_z^2=\left(\frac{1}{\mathrm{i}}(x\partial y-y\partial_x)\right)^2=-\left(x\partial y-y\partial_x\right)^2 \tag{11.2.895}$$

Let us study its commutation with a component (without having to explicit the thing!):

$$\begin{aligned}[][l_z,\|\vec{l}\|^2]&=[l_z,l_x^2]+[l_z,l_y^2]+\underbrace{[l_z,l_z^2]}_{0}=l_x[l_z,l_x]+[l_z,l_x]l_x+l_y[l_z,l_y]+[l_z,l_y]l_y\\ &=l_x[l_z,l_x]+[l_z,l_x]l_x-l_y[l_y,l_x]-[l_y,l_z]l_y\end{aligned} \tag{11.2.896}$$

Using the cyclic relation $[l_x,l_y]=\mathrm{i}l_z$ it comes:

$$[l_z,\|\vec{l}\|^2]=l_x[l_z,l_x]+[l_z,l_x]l_x-l_y[l_y,l_x]-[l_y,l_z]l_y=\mathrm{i}\big(l_xl_y+l_yl_x\big)-\mathrm{i}\big(l_xl_y+l_yl_x\big)=0 \tag{11.2.897}$$

Therefore the norm of the orbital angular momentum commutes with its components:

$$\boxed{[l_z,\|\vec{l}\|^2]=0\quad\text{cycl.}} \tag{11.2.898}$$

Conclusions of the results obtained so far: Since the commutator is equal to zero (the quantities commutes), it is therefore possible to simultaneously measure with precision a component as well as the square of the angular momentum (its squared norm), but it is impossible to do the same for two components!

Let us also notice that finally the fact that we have explicitly:

$$[l_x,l_y]=l_xl_y-l_yl_x=\mathrm{i}l_z\quad\text{cycl.} \tag{11.2.899}$$

we can write (its quite an interesting relation):

$$\boxed{\vec{l}\times\vec{l}=\mathrm{i}\vec{l}} \tag{11.2.900}$$

If we have a system of particles numbered by an index k, each has an individual kinetic moment $\hbar\vec{l}^{(k)}$ and the total orbital angular momentum of the system $\hbar\vec{L}$ (not to confuse the notation of the L with that of the Lagrangian that latter being anyway as scalar!!!) is then obviously given by:

$$\hbar\vec{L}=\sum_k \hbar\vec{l}^{(k)} \tag{11.2.901}$$

and if we simplify by $\hbar$ (or if we work in natural units):

$$\vec{L}=\sum_k \vec{l}^{(k)} \tag{11.2.902}$$

But $\vec{L}$ is not really the total angular momentum of the system! Indeed, a particle may possess an intrinsic angular momentum, or "spin". We can give a simple picture of the spin by saying that it translates an infinitesimal rotation of the particle on itself (caution! it is only an image because in fact the particle does not turn on itself!). As we saw in the section of Spinor Calculus, this corresponds mathematically to the limited development of the matrix of rotations in the vicinity of the identity matrix.

Therefore we see that the usual approach to quantum spin is just to postulate that elementary particles have an intrinsic spin, and that this spin isn't due to any physical motion of the particle (again refer to the section of Corpuscular Quantum Physics); it just ***is***, in much the same way that particles have mass, charge and (in the case of the more esoteric particles) several other quantities such as strangeness. This is not terribly satisfactory from an intellectual point of view, but since I'm not qualified to write about the details of particle physics, we have to start somewhere.

We will denote by $\vec{s}^{(k)}$ the intrinsic angular momentum of the k-th particle (in natural units) and the relation:

$$\vec{S}=\sum_k \vec{s}^{(k)} \tag{11.2.903}$$

will be the "total spin" and finally (still in natural units or after simplification of $\hbar$):

$$\boxed{\vec{J}=\vec{L}+\vec{S}=\sum_k \vec{l}^{(k)}+\sum_k \vec{s}^{(k)}} \tag{11.2.904}$$

will be the "total angular momentum" of the system (do not confuse the notation J with the orbital angular momentum or the current density....!) and we will prove during our study of the spin-orbit coupling that this angular momentum is a constant of movement in the presence of this coupling.

Remark

In the weak field case this vector model ($\vec{J}=\vec{L}+\vec{S}$) at left implies that the coupling of the orbital angular momentum $\vec{L}$ to the spin angular momentum $\vec{S}$ is stronger than their coupling to the external field. In this case where spin-orbit coupling is dominant, they can be visualized as combining to form a total angular momentum $\vec{J}$ which then precesses about the magnetic field direction (see further below for an illustration).

In the strong-field case, $\vec{S}$ and $\vec{L}$ couple more strongly to the external magnetic field than to each other, and can be visualized as independently precessing about the external field direction such that $\vec{J}\neq\vec{L}+\vec{S}$. This is what we name the Paschen-Back effect". The fields required to generate Paschen-Back conditions for sodium are unrealistically high. Lithium, on the other hand, has a spin-orbit splitting of only 0.00004 [eV] compared to 0.0021 [eV] for sodium. The Paschen-Back conditions are met in some lithium spectra observed on the Sun, so this effect does have astronomical significance.

We will assume (but this is relatively easy to prove once, among others, spinors are know[19]) that each $\vec{s}^{(k)}$ and $\vec{l}^{(k)}$ also obeys the commutation rules seen previously:

$$[s_x^{(k)},s_y^{(k)}]=\mathrm{i}s_z^{(k)}\quad\text{cycl.}\qquad\text{and}\qquad[l_x^{(k)},l_y^{(k)}]=\mathrm{i}l_z^{k}\quad\text{cycl.}\tag{11.2.905}$$

Therefore to sum up we have:

$$\boxed{[L_i,L_j]=\mathrm{i}\varepsilon_{ijk}L_k\qquad[S_i,S_j]=\mathrm{i}\varepsilon_{ijk}S_k\qquad[J_i,J_j]=\mathrm{i}\varepsilon_{ijk}J_k}\tag{11.2.906}$$

where for recall ε_{ijk} is the Levi-Civita symbol (see section Tensor Calculus page 2045).

We notice that in the matrix representation of Heisenberg there are also other matrices than the 2×2 Pauli matrices that satisfy these relations (we can give the detail of the proof on reader request). For example, the following zero trace Hermitian matrices (of which the conjugate transpose is equal to itself for recall ....):

$$\underbrace{J_z=\begin{bmatrix}0&-\mathrm{i}&0\\ \mathrm{ii}&0&0\\ 0&0&0\end{bmatrix},J_x=\begin{bmatrix}0&0&0\\ 0&0&-\mathrm{i}\\ 0&\mathrm{i}&0\end{bmatrix},J_y=\begin{bmatrix}0&0&\mathrm{i}\\ 0&0&0\\ -\mathrm{i}&0&0\end{bmatrix}}_{(J_i)_{mn}=-\mathrm{i}\varepsilon_{imn}}\tag{11.2.907}$$

Remark

We will see in the section of Relativistic Quantum Physics that in fact the three matrices above are the spin operator of the massless spin-1 particles (photons).

The prior-previous relations implies (also) in the same way as for the orbital angular momentum:

$$[J_z,\|\vec{J}\|^2]=0\quad\text{cycl.}\tag{11.2.908}$$

with obviously the relation:

$$\|\vec{J}\|^2=J_x^2+J_y^2+J_z^2\tag{11.2.909}$$

---

[19] As always we can detail on reader request

named by the mathematicians "Casimir element" or "Casimir operator" (a simple development perfectly similar to that obtained above is sufficient to prove it and we can write it on reader request).

Let us now define in a purely formal way the two non-Hermitian operators named "scale operators" (Pauli's matrices and not only (!) still satisfy these relations!):

$$J_{\pm}=J_x\pm \mathrm{i}J_y \qquad \left(J_+\right)^{\dagger}=J_- \tag{11.2.910}$$

where respectively J_+ is named "elevator operator" and J_- "step-down operator".

The $J_{\pm}$ commutes with $\|\vec{J}\|^2$, since it commutes with J_x and J_y. This allows us to write the product:

$$J_+\cdot\|\vec{J}\|^2=\|\vec{J}\|^2\cdot J_- \tag{11.2.911}$$

Furthermore:

$$\begin{aligned}[J_z,J_+]&=[J_z,J_x+\mathrm{i}J_y]=J_z(J_x+\mathrm{i}J_y)-(J_x+\mathrm{i}Jy)J_z\\&=J_zJ_x+\mathrm{i}J_zJ_y-J_xJ_z+\mathrm{i}J_yJ_z=[J_z,J_x]+\mathrm{i}[J_z,J_y]=\mathrm{i}J_y-\mathrm{i}[J_y,J_z]\\&=\mathrm{i}J_y-\mathrm{ii}J_x=\mathrm{i}J_y+J_x=J_x+\mathrm{i}J_y\end{aligned} \tag{11.2.912}$$

Therefore:

$$[J_z,J_+]=J_zJ_+-J_+J_z=J_+ \tag{11.2.913}$$

Identically (we can detail on reader request):

$$J_zJ_--J_-J_z=J_- \tag{11.2.914}$$

Finally, let us evaluate the products J_+J_- and J_-J_+:

$$J_-J_+=(J_x-i\mathrm{i}J_y)(J_x+\mathrm{i}J_y)=J_x^2+J_y^2+\mathrm{i}(J_xJ_y-J_yJ_x)=\|\vec{J}\|^2-J_z^2-J_z \tag{11.2.915}$$

Identically (we can also detail on reader request):

$$J_+J_-=\|\vec{J}\|^2-J_z^2+J_z \tag{11.2.916}$$

Since the two Hermitian operators $\|\vec{J}\|^2$ and J_z commute they therefore have common states and eigenvalues and, more precisely, they have a common complete common basis. When observables commuted and have their own common base, let us recall that we usually speak of a "CSCO" (Complete Set of Commuting Operators).

To study their eigenvalues let us put:

$$\begin{aligned}\|\vec{J}\|^2\left|\Psi\right\rangle&=K\left|\Psi\right\rangle\\J_z\left|\Psi\right\rangle&=m\left|\Psi\right\rangle\end{aligned} \tag{11.2.917}$$

System which is sometimes denoted in the following form in the specialized literature (...):

$$\boxed{\begin{aligned}\|\vec{J}\|^2\left|K,m\right\rangle&=K\left|K,m\right\rangle\\J_z\left|K,m\right\rangle&=m\left|K,m\right\rangle\end{aligned}} \tag{11.2.918}$$

as it highlights the fact that the associated eigenstates will be defined at least in part by the parameters K, m themselves.

Many times in textbooks the last couple of relations is written:

$$\begin{gathered}\hat{J}^2\big|K,m\rangle = K\big|K,m\rangle\\ \hat{J}_z\big|K,m\rangle = m\big|K,m\rangle\end{gathered}\tag{11.2.919}$$

or even worst...:

$$\begin{gathered}J^2\big|Km\rangle = K\big|Km\rangle\\ J_z\big|Km\rangle = m\big|Km\rangle\end{gathered}\tag{11.2.920}$$

To begin, we know that the eigenvalues K and m are not independent since we have:

$$\|\vec{J}\|^2 - J_z^2 = J_x^2 + J_y^2\tag{11.2.921}$$

The mean being denoted by the brackets $\langle\rangle$, we have by linearity of the mean (see section Statistics page 571):

$$\langle\|\vec{J}\|^2\rangle - \langle J_z^2\rangle = \langle J_x^2\rangle + \langle J_y^2\rangle\tag{11.2.922}$$

What can be written:

$$\langle\|\vec{J}\|^2 - J_z^2\rangle = \langle J_x^2\rangle + \langle J_y^2\rangle\tag{11.2.923}$$

We see that the left-hand side of the above relation is therefore equal by definition to:

$$K - m^2\tag{11.2.924}$$

Since the square of the total orbital angular momentum is anyway Hermitian (it has no complex component in $\mathbb{C}$), we then construct by the postulates of quantum physics:

$$\langle J_x^2\rangle = \langle\Psi\big|J_x^2\big|\Psi\rangle \geq 0\tag{11.2.925}$$

It comes then that:

$$\left.\begin{matrix}\langle J_x^2\rangle \geq 0\\ \langle J_y^2\rangle \geq 0\end{matrix}\right\}\langle\|\vec{J}\|^2 - J_z^2\rangle \geq 0\tag{11.2.926}$$

The latter relation therefore implies that:

$$K \geq m^2\tag{11.2.927}$$

This brings so far to the following informations:

$$\boxed{K \geq 0 \qquad |m| \leq \sqrt{K}}\tag{11.2.928}$$

From $|\Psi\rangle$, we build the state $J_+|\Psi\rangle$, we will show that if this state is not identically zero, it is an eigenvalue of $\|\vec{J}\|^2$ and of J_z. From the relation:

$$J_+\cdot\|\vec{J}\|^2 = \|\vec{J}\|^2\cdot J_-\tag{11.2.929}$$

already introduced previously, we put:

$$\| \vec{J} \|^2 J_+ \big|\Psi\rangle = J_- \| \vec{J} \|^2 \big|\Psi\rangle = K J_+ \big|\Psi\rangle \tag{11.2.930}$$

The $J_{\pm}$ commutate with $\|\vec{J}\|^2$, since it commutes with J_x and J_y. This gives us that the previous relation is zero such that:

$$K J_+ \big|\Psi\rangle = 0 \Leftrightarrow J_+ \big|\Psi\rangle = 0 \tag{11.2.931}$$

From the relation $J_z J_+ - J_+ J_z = J_+$ we put in and identical way (we multiply both side of equality left by $|\Psi\rangle$) and rearrange a bit:

$$J_z J_+ \big|\Psi\rangle = J_+ J_z \big|\Psi\rangle + J_+ \big|\Psi\rangle = (m+1) J_+ \big|\Psi\rangle \tag{11.2.932}$$

Still without forgetting that:

$$J_+ \big|\Psi\rangle = 0 \tag{11.2.933}$$

We finally have the following package of relations so far:

$$\boxed{J_{\pm} \big|\Psi\rangle = 0 \qquad \| \vec{J} \|^2 J_{\pm} \big|\Psi\rangle = K J_{\pm} \big|\Psi\rangle \qquad J_z J_{\pm} \big|\Psi\rangle = (m \pm 1) J_{\pm} \big|\Psi\rangle} \tag{11.2.934}$$

Therefore in the corresponding order, we have that:

1. $J_+|\Psi\rangle$ and $J_-|\Psi\rangle$ are identically zero
2. $J_+|\Psi\rangle$ and $J_-|\Psi\rangle$ are eigenstates of the operator $\|\vec{J}\|^2$ for the eigenvalue K
3. $J_{\pm}|\Psi\rangle$ are eigenstates for the operator J_z and the eigenvalue $m \pm 1$

Since the angular momentum is quantified as we know from our study of Corpuscular Quantum Physics and also from experiments, its eigenvalues must therefore have a minimum and a maximum with for each of these values and associated eigenfunction.

Let us put for what will follow that m' and $|\Psi'\rangle$ are the maximum eigenvalue and eigenstate and m'' and $|\Psi''\rangle$ the minimum eigenvalue and eigenstate.

Given the following three relations proved earlier above:

$$\begin{aligned} J_+ J_- &= \| \vec{J} \|^2 - J_z^2 + J_z \\ J_- J_+ &= \| \vec{J} \|^2 - J_z^2 - J_z \\ J_{\pm} \big|\Psi\rangle &= 0 \end{aligned} \tag{11.2.935}$$

We write using also the properties of eigenvalues (see section Linear Algebra page 1997):

$$\begin{aligned} J_+ J_- \big|\Psi'\rangle &= \| \vec{J} \|^2 |\Psi'\rangle - J_z^2 |\Psi'\rangle + J_z |\Psi'\rangle = (K - m'^2 - m') \big|\Psi'\rangle = 0 \\ J_- J_+ \big|\Psi''\rangle &= \| \vec{J} \|^2 |\Psi''\rangle - J_z^2 |\Psi''\rangle - J_z |\Psi''\rangle = (K - m''^2 - m'') \big|\Psi''\rangle = 0 \end{aligned} \tag{11.2.936}$$

What intuitively is not obvious to put but which mathematically is quite justifiable.

From the last two relations above, we can write by subtracting the first one to the second one (use of the superposition of sate principle), that gives us immediately for the eigenvalue that:

$$m'^2-m''^2+m'+m''=0 \tag{11.2.937}$$

thus:

$$(m'+m'')(m'-m''+1)=0 \tag{11.2.938}$$

As we have m' being the maximum, m'' being the minimum of the same set, we then have:

$$m'-m''\geq 0 \tag{11.2.939}$$

What gives us after simplification of the second parenthesis that is therefore not zero (because if it could be zero we would not be able to simplify this parenthesis):

$$(m'+m'')=0\Leftrightarrow m''=-m' \tag{11.2.940}$$

Let us denote by J the value m' (which corresponds to the maximum eigenvalue of the quantity J_z) since $m''=-m'$ we have:

$$m''=-J \tag{11.2.941}$$

(where often in the literature we find a lowercase j to avoid confusion with the associated operator) therefore:

$$m'-m''=J-(-J)=2J \tag{11.2.942}$$

Since the difference on the left of the equality is necessarily an integer number (based on the known results that we get in the section of Corpuscular Quantum Physics), this imposes that $2J$ is a positive integer or zero, but it also implies directly that J can only be an integer (!), half-integer or zero such that:

$$\boxed{J=0,\frac{1}{2},1,\frac{3}{2},2,\frac{5}{2},\ldots} \tag{11.2.943}$$

Now as for recall: $m'=J$ and $m''=-J$ it follows that if we denote by m the values that can take m'' and m' respectively, then for each value of J we have obviously:

$$-J\leq m\leq J \tag{11.2.944}$$

Finally, as:

$$(K-m'^2-m')=0\Leftrightarrow K=m'^2+m' \qquad \text{and} \qquad (K-m''^2-m'')=0\Leftrightarrow K=m''^2+m'' \tag{11.2.945}$$

Then we have:

$$K=J^2+J=J(J+1) \qquad \text{and} \qquad K=(-J)^2-(-J)=J(J+1) \tag{11.2.946}$$

And finally this gives us the eigenvalue:

$$K=J(J+1) \tag{11.2.947}$$

And since we have put that m' is equal to J and that we have the relation:

$$|m| \leq \sqrt{K} \tag{11.2.948}$$

therefore:

$$J \leq \sqrt{J(J+1)} \tag{11.2.949}$$

In a more explicit and less confused form (be careful not to confuse the eigenvalues with the operators!):

$$J \leq \sqrt{J_z(J_z+1)} \tag{11.2.950}$$

So we have finally:

$$\boxed{b \leq \sqrt{J_z(J_z+1)}\hbar} \tag{11.2.951}$$

As we have introduced earlier above the spin-orbit interaction:

$$\boxed{\|\vec{J}\|^2\,|\Psi\rangle = \hbar^2 J(J+1)|\Psi\rangle} \tag{11.2.952}$$

As we introduced earlier above:

$$J = L + S \tag{11.2.953}$$

(component by component of their respective vector/matrix) and if the particle has no spin ($S=0$) then we have the eigenvalue of the total orbital angular momentum which reduces to the eigenvalue of the angular momentum:

$$b \leq \sqrt{J_z(J_z+1)}\hbar = \sqrt{L(L+1)}\hbar \tag{11.2.954}$$

where we no longer indicate the indices of the components or of the matrix indices (useless!).

If we have only one particle then:

$$L = \sum_{i=1}^{n} l_i = l_1 = l \tag{11.2.955}$$

Therefore, the orbital angular momentum is written by remembering that l is also quantified (see section Corpuscular Quantum Physics page 4666):

$$\boxed{b \leq \sqrt{l(l+1)}\hbar} \tag{11.2.956}$$

If we have $l \gg 1$, then in this case:

$$b \cong l\hbar \tag{11.2.957}$$

We thus fall back on the result obtained at the beginning of our study of the angular momentum.

Roughly, if we now put $l \cong n$, we thus that we fall back on the angular momentum relation postulated by Bohr as seen in the section of Corpuscular Quantum Physics. This is why it is customary to take only the integer values of l!

Remark

Let us recall that in reality $0 \leq l \leq n-1$ and therefore that unlike the Bohr corpuscular model the angular momentum can be zero wit the above approach...! Another way of accepting the values taken by n, putting apart the fact to refer to the Bohr model seen in the section of Corpuscular Quantum Physics, is to look at the values that can be taken by l in the quantum model of the hydrogen atom of the section of Quantum Chemistry and that if the latter is equal to 0 then the associated Legendre polynomials are no longer defined!

This finding now justifies physically the use of the quantum number l in the periodic table of elements as we saw and defined it (without any real justification) in the previous section of Corpuscular Quantum Physics.

Finally, let us mention that exactly the same reasoning and development leads to the following possible values of the pseudo angular momentum of spin:

$$\boxed{S=0,\frac{1}{2},1,\frac{3}{2},2,\frac{5}{2},\ldots} \qquad (11.2.958)$$

where experiments show us (to list only the most known examples) that spin $S=0$ is a characteristic of the Higgs boson or of certain atoms, spin $S=1/2$ is a characteristic of the electron / positron, spin $S=1$ is a characteristic of the photon, spin $S=2$ would be a still theoretical characteristic of the graviton. At the time of writing, no particles of spin $S=3/2$ or $S=5/2$ are known (some supposed it could be dark energy particles spin).

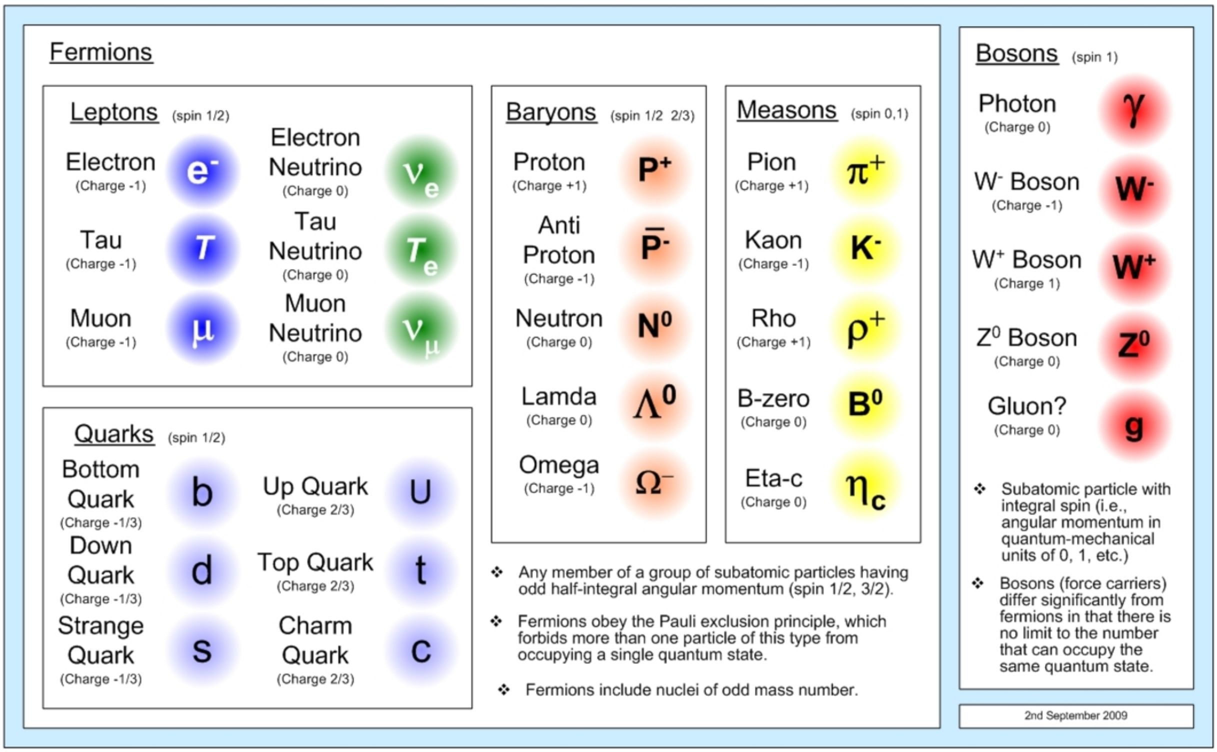

Fermions
Leptons (spin 1/2)
Electron (Charge -1) e-
Electron Neutrino (Charge 0) νe
Tau (Charge -1) τ
Tau Neutrino (Charge 0) τe
Muon (Charge -1) μ
Muon Neutrino (Charge 0) νμ
Quarks (spin 1/2)
Bottom Quark (Charge -1/3) b
Up Quark (Charge 2/3) U
Down Quark (Charge -1/3) d
Top Quark (Charge 2/3) t
Strange Quark (Charge -1/3) s
Charm Quark (Charge 2/3) c
Baryons (spin 1/2 2/3)
Proton (Charge +1) P+
Anti Proton (Charge -1) P̄-
Neutron (Charge 0) N0
Lamda (Charge 0) Λ0
Omega (Charge -1) Ω-
Measons (spin 0,1)
Pion (Charge +1) π+
Kaon (Charge -1) K-
Rho (Charge +1) ρ+
B-zero (Charge 0) B0
Eta-c (Charge 0) ηc
Any member of a group of subatomic particles having odd half-integral angular momentum (spin 1/2, 3/2).
Fermions obey the Pauli exclusion principle, which forbids more than one particle of this type from occupying a single quantum state.
Fermions include nuclei of odd mass number.
Bosons (spin 1)
Photon (Charge 0) γ
W- Boson (Charge -1) W-
W+ Boson (Charge 1) W+
Z0 Boson (Charge 0) Z0
Gluon? (Charge 0) g
Subatomic particle with integral spin (i.e., angular momentum in quantum-mechanical units of 0, 1, etc.)
Bosons (force carriers) differ significantly from fermions in that there is no limit to the number that can occupy the same quantum state.
2nd September 2009

Figure 11.68: Some particles classification by spin (Author: James Childs)

We will see another diagram of the same type including the Higgs particle in the section of Particle Physics.

We see therefore that the integer or semi-integral value of the spin seems to determine a crucial property of the particle: if its spin is integer, it is a "boson", if its spin is half-integer, it is a "fermion" (this can be proven with the "spin-statistics theorem" that as already mentioned we will perhaps in the future give the proof in this textbook).

The total angular momentum is thus approximately given by:

$$\parallel \vec{J} \parallel = \parallel \vec{L} + \vec{S} \parallel \cong (l+s)\hbar = \sqrt{J(J+1)}\hbar \tag{11.2.959}$$

By analogy (it is really a dubious analogy ...), we write for J sufficiently large...:

$$\parallel \vec{J} \parallel \cong (l+s)\hbar = j\hbar \tag{11.2.960}$$

But since the spin of the electron may have only two possible orientations[20], the values of j will be in the case of a spin particle $1/2$:

$$\begin{aligned} j &\cong l+\frac{1}{2} \qquad \text{parallel spins} \\ j &\cong l-\frac{1}{2} \qquad \text{anti-parallel spins} \end{aligned} \tag{11.2.961}$$

Hence a possible classification of the atomic electrons taking into account their spin:

Orbital type	***s***	***p***	***d***	***f***	...
l	0	1	2	3	...
j	$\frac{1}{2}$	$\frac{1}{2},\frac{3}{2}$	$\frac{3}{2},\frac{5}{2}$	$\frac{5}{2},\frac{7}{2}$	...
Notation:	$s_{1/2}$	$p_{1/2},p_{3/2}$	$d_{3/2},d_{5/2}$	$f_{5/2},f_{7/2}$	...

Table 11.10: **Types of orbital and spin for electrons**

We see above that for a multiplicity of $2l+1$ states. Indeed, for $l=0$ we have 1 states, for $l=1$ we have a total of $1+2=3$ states, for $l=2$ we have a total of $1+2+2=5$ states, and so on....

Either in schematic form with the corresponding energy levels:

---

[20] Electrons spin in all directions, but in the Stern-Gerlach experiment they deflect only in the z direction because there was only a field in the z direction (see section Relativistic Quantum Physics page 4979). This is why we speak of two "privileged" possible orientations.

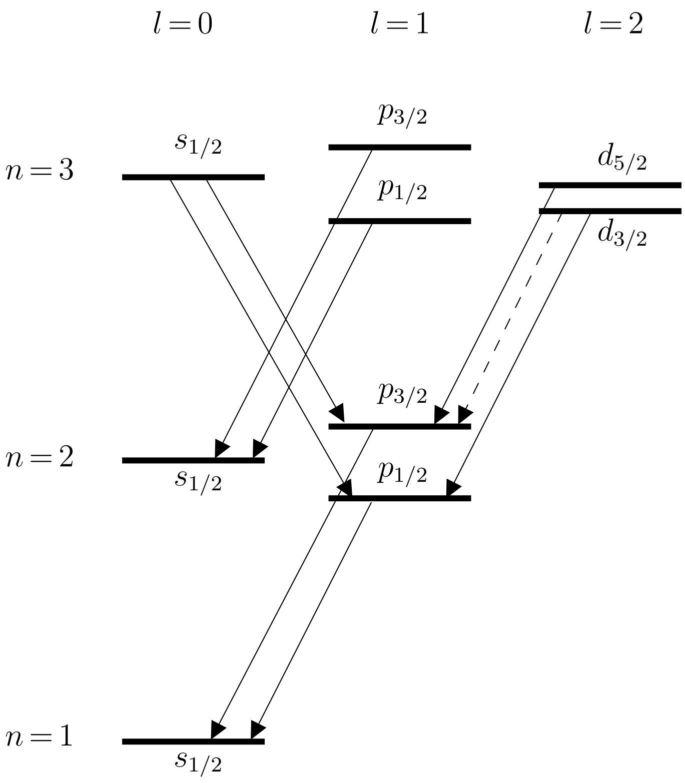


Figure 11.69: Schematic form of associated orbitals and spin energy levels

This table leads us to conclude that we finally:

$$\boxed{|l-s| \leq j \leq |l+s|} \tag{11.2.962}$$

Also we therefore introduce a common nomenclature for naming energy terms when we deal with the details of spin-orbit coupling is[21]:

$$\boxed{n^{2s+1}l_j} \tag{11.2.963}$$

where for recall n is the principal quantum number, s is the spin therm, l the orbital angular momentum and j the total angular momentum.

To return to more practical considerations ... we finally obtained for the norm of total angular momentum (in the case of a single particle and without spin):

$$J \leq \sqrt{l(l+1)}\hbar \tag{11.2.964}$$

where l is an integer (at least in the case where we don't do the approximation done earlier above). We also know from the section of Corpuscular Quantum Physics (see page 4666) that the magnetic moment is given by:

$$\mu_l = m_l\hbar \tag{11.2.965}$$

and that the secondary quantum number l and the magnetic quantum number m_l are in some way indissociable.

---

[21] the spin term is sometimes omitted because it's the same for all the levels

In the same way we get:

$$S \leq \sqrt{s(s+1)}\hbar \tag{11.2.966}$$

where can take for a particle like the electron only the values (don't forget that is doesn't mean it can take only two directions! It just means it can take two states and the famous "two directions" legend is due to the Stern-Gerlach experiment type configuration):

$$s = \pm\frac{1}{2} \tag{11.2.967}$$

Which simply correspond (in natural units) to the two eigenvalues of the eigenmatrices:

$$\frac{1}{2}\sigma_i \tag{11.2.968}$$

which links the spin operator to the Pauli matrices for the electron by the Dirac equation as we will prove it in the section of Relativistic Quantum Physics (see page 4931):

$$S = \frac{\hbar}{2}s_i = \frac{\hbar}{2}\sigma_i \tag{11.2.969}$$

Remark

It is more accurate to say that in reality that the eigenvalues $\pm 1/2\hbar$ are linked to eigenmatrices $\frac{\hbar}{2}\sigma_i$ but as the $\hbar$ simplifies...

Now, what we know of our results obtained in the section of Corpuscular Quantum Physics is that when l is equal to 1 we have the magnetic moment which can take three different values according to whether a magnetic field is applied or not:

$$m_l = -1, 0, +1 \tag{11.2.970}$$

At this time, although the norm of the total angular momentum remains constant (as it is conservative quantity), its components must necessarily change. Since we know only one of the components of the angular momentum by knowing its norm (commutating operators), we choose to focus for pedagogical conventions on J_z.

We choose a reference frame such that one of the spatial components is always equal to zero (this is always possible as we know it!). It is then sufficient, for example, in the chosen planar referential X, Z (i.e. the component Y will be zero) to have the norm of J that is equal to $l = 1$:

$$J \leq \sqrt{1(1+1)}\hbar = \sqrt{2}\hbar \tag{11.2.971}$$

and same with S by imposing that the norm for $s = 1/2$ is equal to:

$$S \leq \sqrt{\frac{1}{2}\left(\frac{1}{2}+1\right)}\hbar = \frac{\sqrt{3}}{2}\hbar \tag{11.2.972}$$

There are then three possibilities to arrive at the same result by simply applying the Euclidean norm if one of the components is always imposed as zero! It is because we have:

$$\left.\begin{array}{l} J_x = \hbar,\ J_z = \hbar \\ J_x = \sqrt{2}\hbar,\ J_z = 0 \\ J_x = \hbar,\ J_z = -\hbar \end{array}\right\} \parallel \vec{J} \parallel = \sqrt{2}\hbar \tag{11.2.973}$$

What we write so (because in fact there is an infinity of possibility) because we want to find a way to introduce the quantum number of orbital projection (which therefore quantifies the projection of the orbital angular momentum along Z and is in multiplicity $2l+1$):

$$\left.\begin{array}{ll} J_x=\hbar, J_z=m_l\hbar & m_l=+1 \\ J_x=\sqrt{2}\hbar, J_z=m_l\hbar & m_l=0 \\ J_x=\hbar, J_z=m_l\hbar & m_l=-1 \end{array}\right\} \|\vec{J}\|=\sqrt{2}\hbar \tag{11.2.974}$$

What some physicists like to represent in a very simplified way by the following diagram:

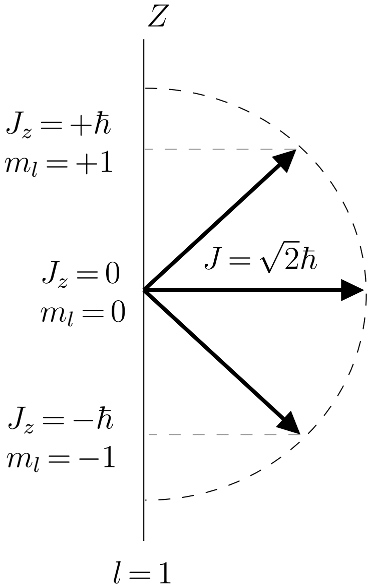


Figure 11.70: **Biased schematic representation of the quantification of total angular momentum in 2D**

But which in reality (by the square of the components of the norm) should be drawn in the following form in 2D (as in 3D this is represented by cones as we will see below):

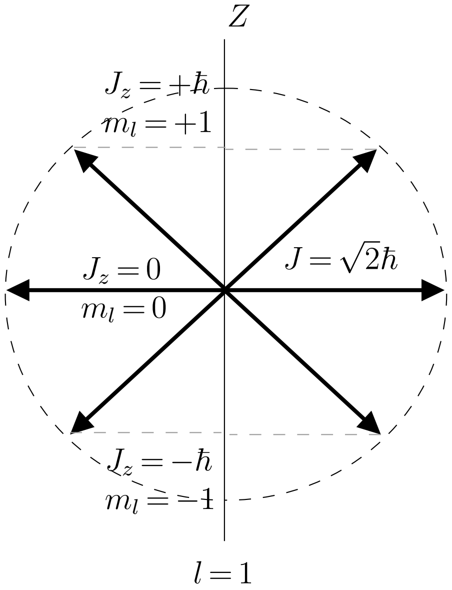


Figure 11.71: **Improved schematic representation of the quantification of total angular momentum in 2D**

This allows us to observe on the way that as we have mention earlier above, we can always rotate the system of axes such that one of the components is zero and earlier we did the choice:

$$\langle J_y \rangle = 0 \tag{11.2.975}$$

Notice that with elementary trigonometry we get:

$$\theta = \arccos\left(\frac{J_z}{\| \vec{J} \|}\right) = \arccos\left(\frac{m_l \hbar}{\hbar\sqrt{J_x^2 + J_z^2}}\right) = \arccos\left(\frac{m_l}{\sqrt{l(l+1)}}\right) \tag{11.2.976}$$

and therefore the angles take the following values:

$$\arccos\left(\frac{-l}{\sqrt{l(l+1)}}\right), \arccos\left(\frac{-l+1}{\sqrt{l(l+1)}}\right), \ldots, \frac{\pi}{2}, \ldots, \arccos\left(\frac{l-1}{\sqrt{l(l+1)}}\right), \arccos\left(\frac{l}{\sqrt{l(l+1)}}\right) \tag{11.2.977}$$

Finally, let us indicate that we then have in the particular case of the figure above where $l = 1$, $m_l = \pm 1$, that the (quite abstract) system introduced earlier above:

$$\begin{aligned} \| \vec{J} \|^2 \big|K, m\rangle &= K \big|K, m\rangle \\ J_z \big|K, m\rangle &= m \big|K, m\rangle \end{aligned} \tag{11.2.978}$$

that will be written as:

$$\begin{aligned} &\| \vec{L} \|^2 \big|l, m_l\rangle = \hbar^2 l(l+1) \big|l, m_l\rangle \Rightarrow \| \vec{L} \|^2 \big|1, +1\rangle = 2\hbar^2 \big|1, +1\rangle \\ &L_z \big|l, m_l\rangle = \hbar l \big|l, m_l\rangle \Rightarrow L_z \big|1, +1\rangle = \hbar \big|1, +1\rangle \end{aligned} \tag{11.2.979}$$

Which is often written in condensed form as following:

$$\begin{aligned} \hat{L}^2 \big|1, +1\rangle &= 2\hbar^2 \big|1, +1\rangle \\ \hat{L}_z \big|1, +1\rangle &= \hbar \big|1, +1\rangle \end{aligned} \tag{11.2.980}$$

In the same way with the spin $1/2$ of electron we have by introducing the quantum number of spin (which therefore quantifies the projection of the spin moment of spin along Z and can take as many values as there are between $-s$ and $+s$ but by increments of 1 as required by the experimental results, for which reason there is no zero Z component below):

$$\left.\begin{aligned} S_x = \frac{\hbar}{\sqrt{2}}, S_z = m_s \hbar \qquad m_s = +\frac{1}{2} \\ S_x = \frac{\hbar}{\sqrt{2}}, S_z = m_s \hbar \qquad m_s = -\frac{1}{2} \end{aligned}\right\} \| \vec{S} \| = \frac{\sqrt{3}}{2}\hbar \tag{11.2.981}$$

What some physicists like to represent in a very simplified way by the following diagram:

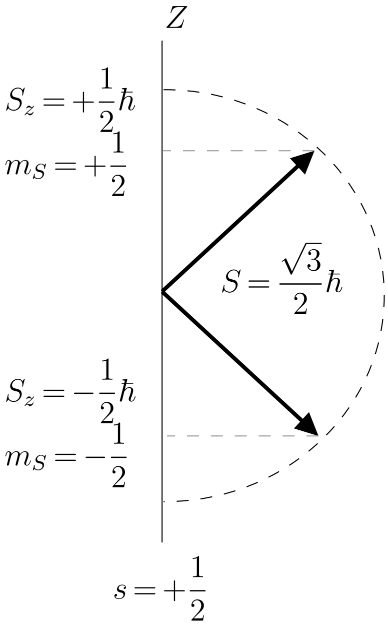


Figure 11.72: **Biased schematic representation of the quantification of spin in 2D**

with for angle:

$$\begin{aligned}\arccos\left(\frac{m_s}{\sqrt{s(s+1)}}\right)&=\arccos\left(\frac{\pm\frac{1}{2}}{\sqrt{\frac{1}{2}\left(\frac{1}{2}+1\right)}}\right)=\arccos\left(\frac{\pm\frac{1}{2}}{\frac{\sqrt{3}}{2}}\right)\\&=\arccos\left(\pm\frac{1}{\sqrt{3}}\right)\cong\pm54.73^\circ\end{aligned}\tag{11.2.982}$$

As before, we have in the particular case where $s=\frac{1}{2}$ and $m_s=\pm\frac{1}{2}$:

$$\begin{aligned}&\|\vec{S}\|^2\left|s,m_s\right\rangle=\hbar^2 s(s+1)\left|s,m_s\right\rangle\Rightarrow\|\vec{S}\|^2\left|\frac{1}{2},\pm\frac{1}{2}\right\rangle=\frac{3}{4}\hbar^2\left|\frac{1}{2},\pm\frac{1}{2}\right\rangle\\&S_z\left|s,m_s\right\rangle=\hbar s\left|s,m_s\right\rangle\Rightarrow S_z\left|\frac{1}{2},\pm\frac{1}{2}\right\rangle=\frac{\hbar}{2}\left|\frac{1}{2},\pm\frac{1}{2}\right\rangle\end{aligned}\tag{11.2.983}$$

This can be written in a more condensed form as following:

$$\begin{aligned}\hat{S}^2\left|\frac{1}{2},\pm\frac{1}{2}\right\rangle&=\frac{3}{4}\hbar^2\left|\frac{1}{2},\pm\frac{1}{2}\right\rangle\\\hat{S}_z\left|\frac{1}{2},\pm\frac{1}{2}\right\rangle&=\frac{\hbar}{2}\left|\frac{1}{2},\pm\frac{1}{2}\right\rangle\end{aligned}\tag{11.2.984}$$

We therefore have the only experimentally measurable variables that are:

$$\boxed{L_z=m_l\hbar}\qquad\text{and}\qquad\boxed{S_z=s\hbar}\tag{11.2.985}$$

Which are therefore discrete observables (bivalent for the spin).

With an artist's view of the concept for the pleasure of the eyes:

Figure 11.73: **Schematic representation of various quantifications of angular momentum by the physicist and sculptor Julian Voss-Andreae**

Therefore, by applying a constant uniform magnetic field, the Pauli Hamiltonian (see section Relativistic Quantum Physics page 4985) will make jumps equivalent to the relation:

$$\boxed{E_p = -\frac{q\hbar}{2m_0}B(m_l + 2s)} \quad (11.2.986)$$

This result means that the energy levels for a given energy (n-layer) are separated into several levels distant of $\dfrac{q\hbar}{2m_0}B$ when the atom is placed in a magnetic field. This result is the Zeeman effect which we have already discussed several times.

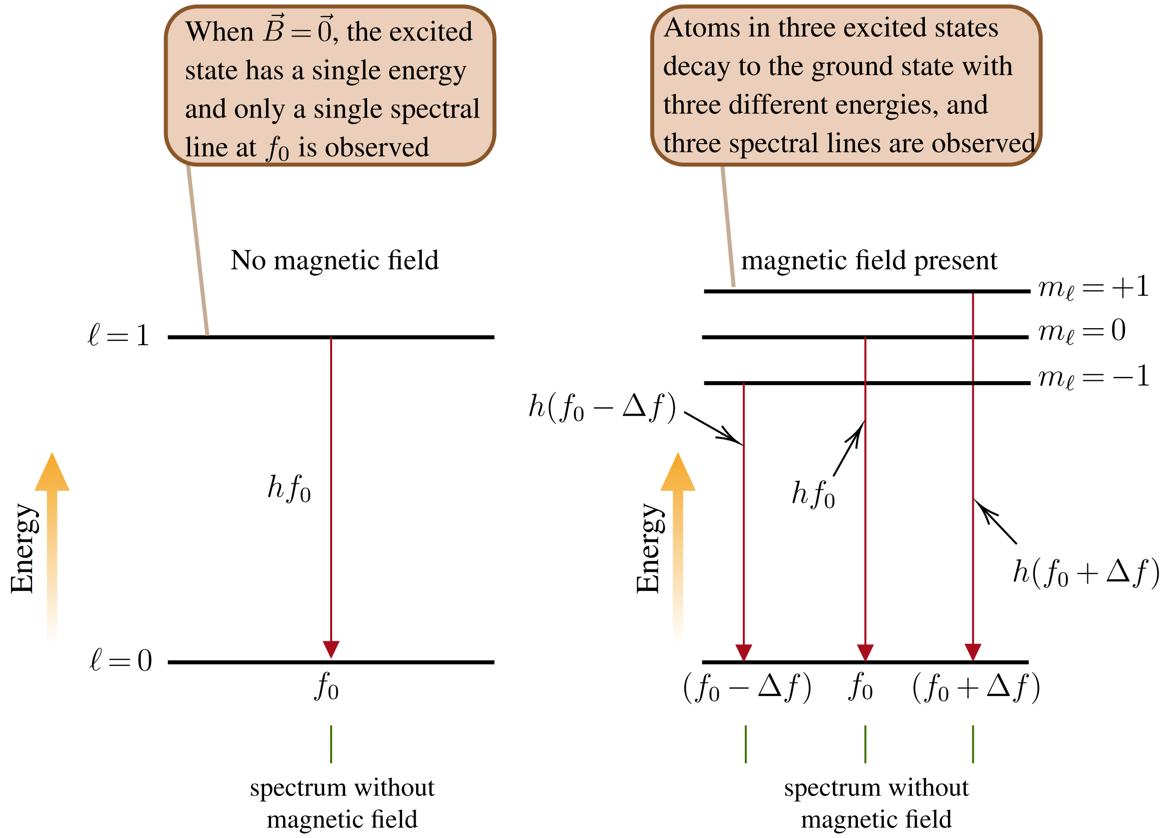

When $\vec{B}=\vec{0}$, the excited state has a single energy and only a single spectral line at f_0 is observed
Atoms in three excited states decay to the ground state with three different energies, and three spectral lines are observed
No magnetic field
magnetic field present
$\ell=1$
$\ell=0$
hf_0
Energy
f_0
spectrum without magnetic field
$m_\ell=+1$
$m_\ell=0$
$m_\ell=-1$
$h(f_0-\Delta f)$
hf_0
$h(f_0+\Delta f)$
Energy
$(f_0-\Delta f)$
f_0
$(f_0+\Delta f)$
spectrum without magnetic field

Figure 11.74: Zeeman effect (source: [491])

All this makes it possible to better understand the mathematical origin of the 4 quantum numbers (principal quantum number, secondary or azimuthal quantum number, magnetic quantum number, spin):

$$n,l,m_l,m_s \tag{11.2.987}$$

also written (since in the particular case of the particles studied in this textbook the magnetic quantum number of spin projection has the same value as the spin since we are dealing mainly with the electron):

$$n,l,m_l,s \tag{11.2.988}$$

To sum up a bit all this...:

$$\begin{aligned}
&J=L+S\\
&n=1,2,3,4\ldots\\
&0\leq l\leq n-1\\
&j=0,\frac{1}{2},1,\frac{3}{2},2,\frac{5}{2},\ldots\\
&s=0,\frac{1}{2},1,\frac{3}{2},2,\frac{5}{2},\ldots\\
&|l-s|\leq j\leq |l-s|\\
&m_l=-l,-l+1,-l+2,\ldots,0,\ldots,l-2,l-1,l\\
&-l\leq m_l\leq +l\\
&m_s=-s,-s+1,-s+2,\ldots,0,\ldots,s-2,s-1,l\\
&-s\leq m_s\leq +s
\end{aligned} \tag{11.2.989}$$

The reader must not think that what we have seen so far was developed in one day or even in one year by physicist. Furthermore this abstract approach was highly speculative and needed to be check with experiments. So far it works very good to explain and predict experimental observations but still in this beginning of the 121st century (holocene calendar) the spin matrices theory is confronted and tested by new experimentation!

✍**Example:**

E1. Consider we have an imaginary a system of two electrons whose orbital quantum numbers are $l_1=2$ and $l_2=4$ respectively. What are the possible values of l (total orbital angular momentum quantum number), s (total spin angular momentum quantum number), j (total angular momentum quantum number).

The quantum numbers associated with the total orbital angular momentum will range from a maximum value found by adding the individual quantum numbers together to a minimum value found by taking the absolute value of the difference of the two numbers in integer steps:

$$\begin{aligned} l_{\max} &= l_1+l_2=2+4=6 \\ l_{\min} &= |l_1-l_2|=|2-4|=2 \end{aligned} \tag{11.2.990}$$

Therefore the possible values for L of the system are $2,3,4,5$ and 6.

The quantum numbers associated with the total spin angular momentum will range from a maximum value found by adding the individual quantum numbers together to a minimum value found by taking the absolute value of the difference of the two numbers in integer steps:

$$\begin{aligned} s_{\max} &= s_1+s_2={}^1\!/\!_2+{}^1\!/\!_2=1 \\ s_{\min} &= |s_1-s_2|=|{}^1\!/\!_2-{}^1\!/\!_2|=0 \end{aligned} \tag{11.2.991}$$

The possible values for s of the system are therefore 0 and 1.

The largest possible value for j will be found by adding together the maximum values for both l and s. The minimum value for j will be found by subtracting the largest possible value of s from the smallest possible value for l:

$$\begin{aligned} j_{\max} &= l_{\max}+s_{\max}=6+1=7 \\ j_{\min} &= |l_{\min}-s_{\max}|=|2-1|=1 \end{aligned} \tag{11.2.992}$$

Therefore the possible values for J of the system are $1,2,3,4,5$ and 6.

E2. A hydrogen atom is known to be in a state characterized the quantum numbers $n=3, l=2$. What are the values of the total angular momentum and their corresponding magnitude?

Therefore the values of the total angular momentum are:

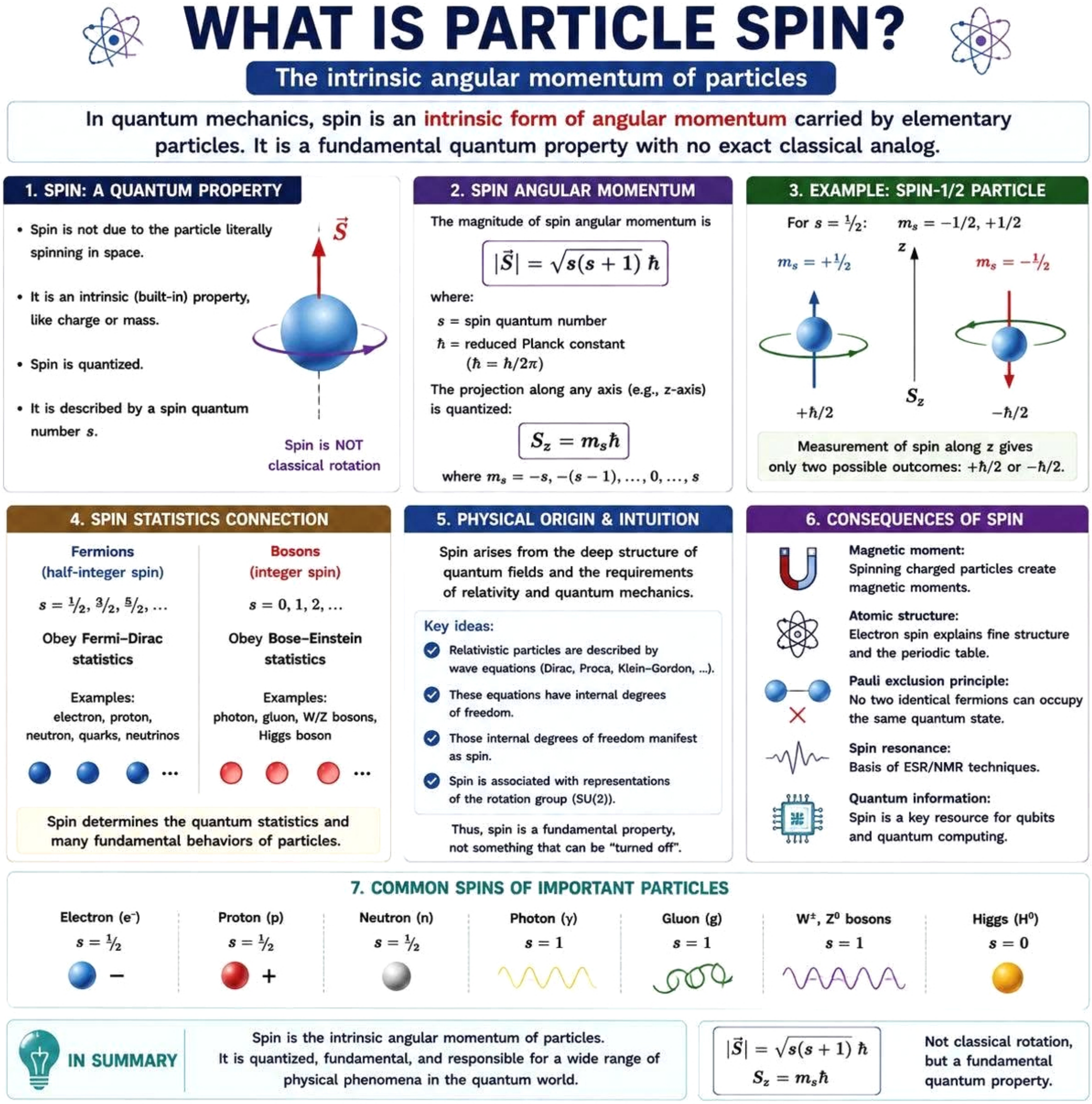
$$\begin{aligned} j_{\max} &= l_{\max}+s_{\max}=2+{}^1\!/\!_2={}^5\!/\!_2 \\ j_{\min} &= l_{\min}+s_{\min}=2-{}^1\!/\!_2={}^3\!/\!_2 \end{aligned} \tag{11.2.993}$$

We have as we have proved above:

$$\|\vec{J}\|^2|j,m\rangle = j(j+1)\hbar^2|j,m\rangle \qquad (11.2.994)$$

Therefore we have for $j_{\text{max}} = {}^5/_2$ and $j_{\text{min}} = {}^3/_2$ the following mangitude:

$$\begin{gathered}\|\vec{J}\|^2|{}^5/_2,m\rangle : j(j+1)\hbar^2 = {}^{35}/_4\hbar^2\\ \|\vec{J}\|^2|{}^3/_2,m\rangle : j(j+1)\hbar^2 = {}^{15}/_4\hbar^2\end{gathered} \qquad (11.2.995)$$



WHAT IS PARTICLE SPIN?

The intrinsic angular momentum of particles

In quantum mechanics, spin is an **intrinsic form of angular momentum** carried by elementary particles. It is a fundamental quantum property with no exact classical analog.

1. SPIN: A QUANTUM PROPERTY

- Spin is not due to the particle literally spinning in space.
- It is an intrinsic (built-in) property, like charge or mass.
- Spin is quantized.
- It is described by a spin quantum number s.

$\vec{S}$

Spin is NOT classical rotation

2. SPIN ANGULAR MOMENTUM

The magnitude of spin angular momentum is

$$\boxed{|\vec{S}| = \sqrt{s(s+1)}\,\hbar}$$

where:

s = spin quantum number
$\hbar$ = reduced Planck constant
($\hbar = h/2\pi$)

The projection along any axis (e.g., z-axis) is quantized:

$$\boxed{S_z = m_s\hbar}$$

where $m_s = -s, -(s-1), \ldots, 0, \ldots, s$

3. EXAMPLE: SPIN-1/2 PARTICLE

For $s = {}^1/_2$: $m_s = -1/2, +1/2$

$m_s = +{}^1/_2$ z $m_s = -{}^1/_2$

$+\hbar/2$ S_z $-\hbar/2$

Measurement of spin along z gives only two possible outcomes: $+\hbar/2$ or $-\hbar/2$.

4. SPIN STATISTICS CONNECTION

Fermions (half-integer spin)	Bosons (integer spin)
$s = {}^1/_2, {}^3/_2, {}^5/_2, \ldots$	$s = 0, 1, 2, \ldots$
Obey Fermi–Dirac statistics	Obey Bose–Einstein statistics
Examples: electron, proton, neutron, quarks, neutrinos	Examples: photon, gluon, W/Z bosons, Higgs boson
● ● ● …	● ● ● …

Spin determines the quantum statistics and many fundamental behaviors of particles.

5. PHYSICAL ORIGIN & INTUITION

Spin arises from the deep structure of quantum fields and the requirements of relativity and quantum mechanics.

Key ideas:

- ✓ Relativistic particles are described by wave equations (Dirac, Proca, Klein–Gordon, ...).
- ✓ These equations have internal degrees of freedom.
- ✓ Those internal degrees of freedom manifest as spin.
- ✓ Spin is associated with representations of the rotation group (SU(2)).

Thus, spin is a fundamental property, not something that can be "turned off".

6. CONSEQUENCES OF SPIN

Magnetic moment:
Spinning charged particles create magnetic moments.

Atomic structure:
Electron spin explains fine structure and the periodic table.

Pauli exclusion principle:
No two identical fermions can occupy the same quantum state.

Spin resonance:
Basis of ESR/NMR techniques.

Quantum information:
Spin is a key resource for qubits and quantum computing.

7. COMMON SPINS OF IMPORTANT PARTICLES

Electron (e^-)	Proton (p)	Neutron (n)	Photon (γ)	Gluon (g)	$W^{\pm}$, Z^0 bosons	Higgs (H^0)
$s = {}^1/_2$	$s = {}^1/_2$	$s = {}^1/_2$	$s = 1$	$s = 1$	$s = 1$	$s = 0$
−	+					

IN SUMMARY

Spin is the intrinsic angular momentum of particles. It is quantized, fundamental, and responsible for a wide range of physical phenomena in the quantum world.

$$\begin{gathered}|\vec{S}| = \sqrt{s(s+1)}\,\hbar\\ S_z = m_s\hbar\end{gathered}$$

Not classical rotation, but a fundamental quantum property.

11.2.5.1 Spin–orbit interaction (LS coupling)

In quantum physics, the "spin–orbit interaction" (also named "spin–orbit effect" or "spin–orbit coupling") is an interaction of a particle's spin with its motion. The first and best known example of this is that spin–orbit interaction causes shifts in an electron's atomic energy levels due to electromagnetic interaction between the electron's spin and the magnetic field generated by the electron's orbit around the nucleus. This is detectable as a splitting of spectral lines, which can be thought of as a Zeeman Effect due to the internal field. A similar effect, due to the relationship between angular momentum and the strong nuclear force, occurs for protons and neutrons moving inside the nucleus, leading to a shift in their energy levels in the nucleus shell model. In the field of spintronics, spin–orbit effects for electrons in semiconductors and other materials are explored for technological applications. The spin–orbit interaction is one cause of magneto-crystalline anisotropy.

Let us now recall that we have pointed out in the section of Corpuscular Quantum Physics that when we analyse the spectral lines of hydrogen in the absence of any external field at high resolution, we see that they are in fact made up of very close doublets, separated of 0.016 [nm] in terms of wavelength. This phenomenon is due to a so-called spin-orbit coupling. It is now time to see where it comes from. So let us also recall that we previously obtained:

$$\vec{J}=\vec{L}+\vec{S} \tag{11.2.996}$$

Hence, the squared norm (what is measured) leads us to write:

$$\|\vec{J}\|^2=\|\vec{L}+\vec{S}\|^2=\left(L_x+S_x\right)^2+\left(L_y+S_y\right)^2+\left(L_z+S_z\right)^2 \tag{11.2.997}$$

Which gives us after grouping:

$$\|\vec{J}\|^2=\|\vec{L}+\vec{S}\|^2=\|\vec{L}\|^2+\|\vec{S}\|^2+2\vec{L}\circ\vec{S} \tag{11.2.998}$$

The term:

$$\boxed{\vec{L}\circ\vec{S}} \tag{11.2.999}$$

is named "spin-orbit coupling". It is the term that during the very precise measurements of spectral lines reveals a splitting of the lines due to the coupling between the electron spin and the orbital angular momentum (this is not an $\|\vec{S}\|^2$ because this term is always positive).

Remark

Let us recall that when we have two bodies in interaction the total angular momentum is a constant of motion as we have proved it in the section of Classical Mechanics. There may thus be a transfer of angular momentum between these two bodies (this is the spin-orbit coupling!). One loses angular momentum and the other gain angular momentum.

The measured deviation is thus attributed to the interaction of the electron spin with its orbital angular momentum. The electron revolves around the nucleus, but if we place ourselves on the electron, we see the nucleus turning (on Earth the Sun revolves around the Earth!). Everything happens as if the nucleus generate a magnetic field at the level of the electron, and this field interacts with the magnetic moment of the electron, the spin, and this differs according to whether the spin is in the direction of the field or opposite, It is this difference that adds or subtracts some energy from the level.

Here is a first simplified diagram that summarizes the whole:

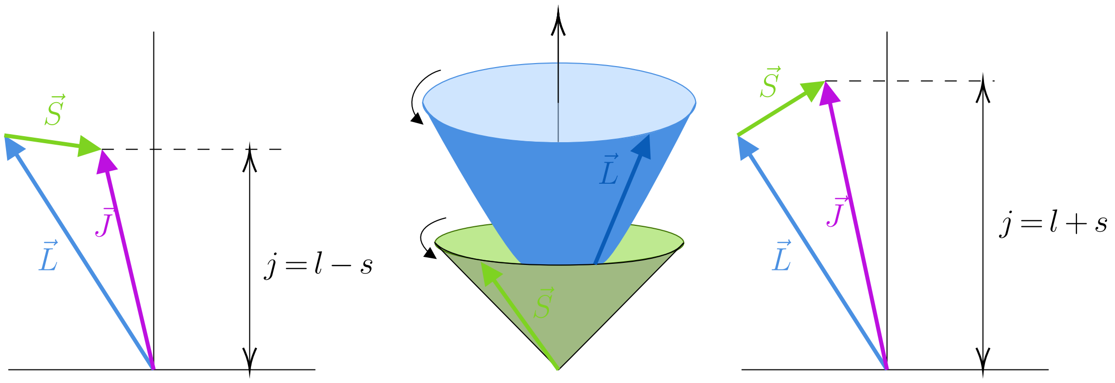
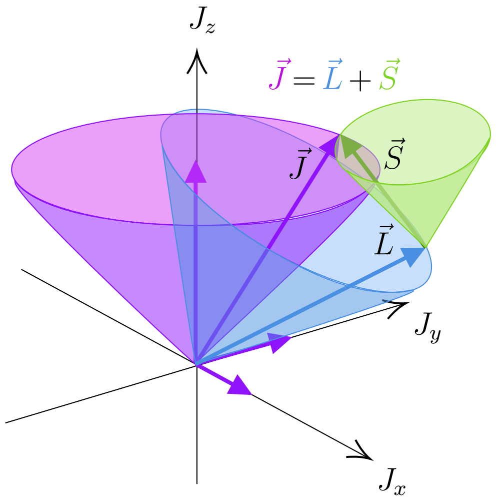


Figure 11.75: **Pictorial representation of the spin-orbit interaction**

or in 3D perspective and that is a bit more accurate:



Let us prove that $\vec{J}$ as defined is a constant of motion. We have (needless to say that by squaring, these are the components of the vector that we put square and not the vector itself!):

$$\vec{L}\circ\vec{S}=\frac{1}{2}\left[\left(\vec{L}+\vec{S}\right)^2-\vec{L}^2-\vec{S}^2\right] \quad (11.2.1000)$$

Hence:

$$\left[\vec{J},\vec{L}\circ\vec{S}\right]=\frac{1}{2}\left[\vec{J},\vec{J}^2-\vec{L}^2-\vec{S}^2\right] \quad (11.2.1001)$$

Let's make the development for one component:

$$\begin{aligned}\frac{1}{2}\left[J_x,J_x^2-L_x^2-S_x^2\right]=&\frac{1}{2}\left[J_x\left(J_x^2-L_x^2-S_x^2\right)-\left(J_x^2-L_x^2-S_x^2\right)J_x\right]\\&\frac{1}{2}\left[\cancel{J_x^3}-J_xL_x^2-J_xS_x^2-\cancel{J_x^3}+L_x^2J_x+S_x^2J_x\right]\\&=-\frac{1}{2}\left[L_x^2J_x+S_x^2J_x-J_xL_x^2-J_xS_x^2\right]\\&=-\frac{1}{2}\left[J_x,L_x^2+S_x^2\right]\end{aligned}\tag{11.2.1002}$$

But, by definition (of notation) $J_x=L_x+S_x$ therefore:

$$\begin{aligned}\left[\vec{J},\vec{L}\circ\vec{S}\right]_x&=-\frac{1}{2}\left[J_x,L_x^2+S_x^2\right]=-\frac{1}{2}\left[L_x+J_x,L_x^2+S_x^2\right]\\&=-\frac{1}{2}\left[\left(L_xL_x^2+L_xS_x^2+S_xL_x^2+S_xS_x^2\right)-\left(L_x^2L_x+S_x^2L_x+L_x^2S_x+S_x^2S_x\right)\right]\\&=-\frac{1}{2}\left(\left[L_x,L_x^2\right]+\left[S_x,S_x^2\right]+\left[L_x,S_x^2\right]+\left[S_x,L_x^2\right]\right)\end{aligned}\tag{11.2.1003}$$

Now we know that $[L_x,L_x^2]=0$ (because an operator always commutes with itself by construction) and concerning $[S_x,S_x^2,$ we have mentioned it in the section of Spinor Calculus and we will prove it in the framework of the study of the free classical Dirac equation (see section Relativistic Quantum Physics page 4931), that the electron spin is fully described by the Pauli matrices which are linear operators. Let us then write to a constant factor:

$$S_x=c^{te}\cdot\sigma_x=c^{te}\cdot\begin{bmatrix}0&1\\1&0\end{bmatrix}\tag{11.2.1004}$$

and we will see that this is in conformity with the Pauli equation that we introduce and study in the section of Relativistic Quantum Physics (and vice versa) !!!

Therefore, disregarding the multiplicative constant:

$$[S_x,S_x^2]=\begin{bmatrix}0&1\\1&0\end{bmatrix}\left(\begin{bmatrix}0&1\\1&0\end{bmatrix}\right)^2-\left(\begin{bmatrix}0&1\\1&0\end{bmatrix}\right)^2\begin{bmatrix}0&1\\1&0\end{bmatrix}=0\tag{11.2.1005}$$

Which was in any case 100% predictable since in any case, once again, a same operator always commutes with himself.

So finally:

$$\left[\vec{J},\vec{L}\circ\vec{S}\right]_x=-\frac{1}{2}\left(\left[L_x,S_x^2\right]+\left[S_x,L_x^2\right]\right)\tag{11.2.1006}$$

Therefore:

$$\begin{aligned}&[L_x,S_x^2]=L_x\begin{bmatrix}0&1\\1&0\end{bmatrix}^2-\begin{bmatrix}0&1\\1&0\end{bmatrix}^2L_x=L_x\begin{bmatrix}1&0\\0&1\end{bmatrix}-\begin{bmatrix}1&0\\0&1\end{bmatrix}L_x=0\\&[S_x,L_x^2]=\begin{bmatrix}0&1\\1&0\end{bmatrix}\cdot L_x^2-L_x^2\begin{bmatrix}0&1\\1&0\end{bmatrix}=0\end{aligned}\tag{11.2.1007}$$

Hence finally:

$$\left[\vec{J},\vec{L}\circ\vec{S}\right]=0 \tag{11.2.1008}$$

$\vec{J}$ is the indeed total angular momentum which, even in the presence of spin-orbit interaction, is a constant of motion (an obligation for an isolated system!).

Remark

Another way of reading that thing is to say that the measurement on one of the elements of the preceding commutator adapts the other one immediately so that their commutation is zero therefore by extension the total angular momentum is a constant of the motion.

One notable atomic spectral line of sodium vapour is the so-called D-line, which may be observed directly as the sodium flame-test line and also the major light output of low-pressure sodium lamps (these produce pressure sodium lamps (these produce an unnatural yellow. The D-line is one of the classified Fraunhofer lines Sodium vapour in the upper layers of lines. Sodium vapour in the upper layers of the Sun generates a dark line in the emitted spectrum of electromagnetic radiation by absorbing visible light in a band of wavelengths around 589.5 [nm]. This wavelength corresponds to transitions in atomic sodium in which valence-electron transitions from $3s$ to $3p$ electronic state.

The splitting of the sodium doublet in the presence of an external magnetic field was observed by Pieter Zeeman in 11896 (holocene calendar), and the effect was subsequently named the "Zeeman effect". It is remarkable that so much detailed spectroscopy was done long before the Bohr theory, and perhaps even more remarkable that Zeeman's first study of the sodium Zeeman splitting was done the year before J. J.Thomson's discovery of the electron in 11897 (holocene calendar).

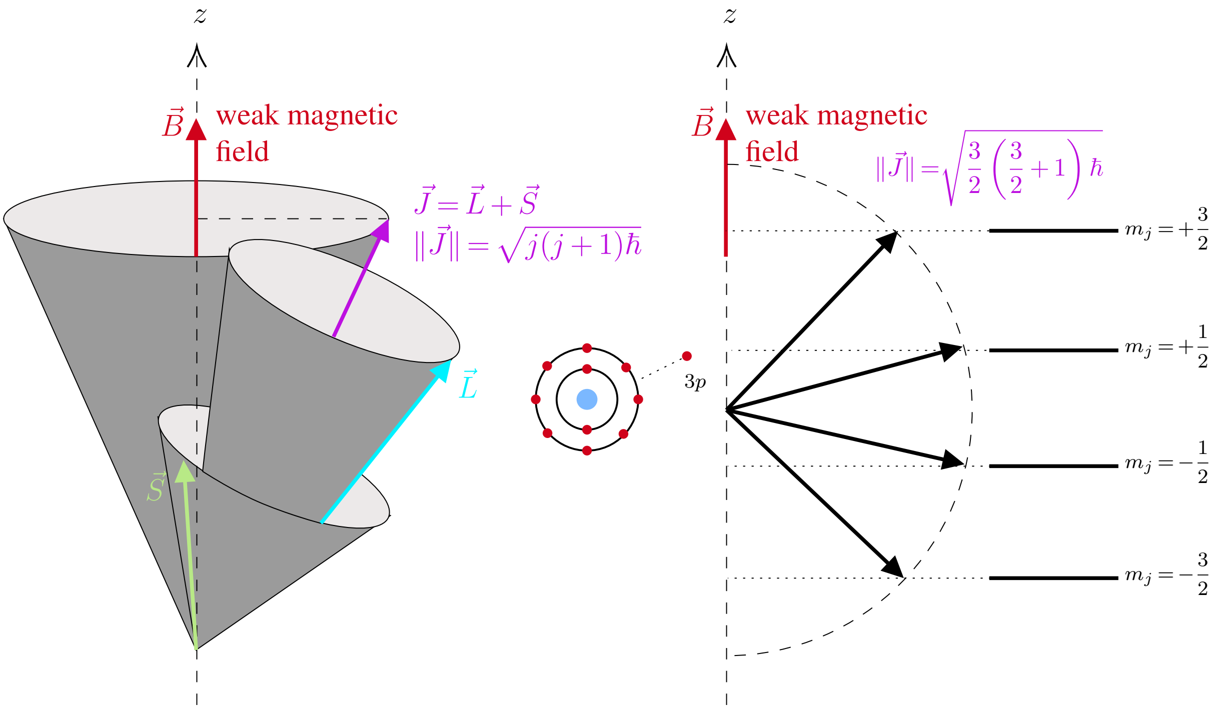


Figure 11.76: **Focus on $3P_{3/2}$ Sodium splitting (source: Hyperphysics)**

Or more detailed:

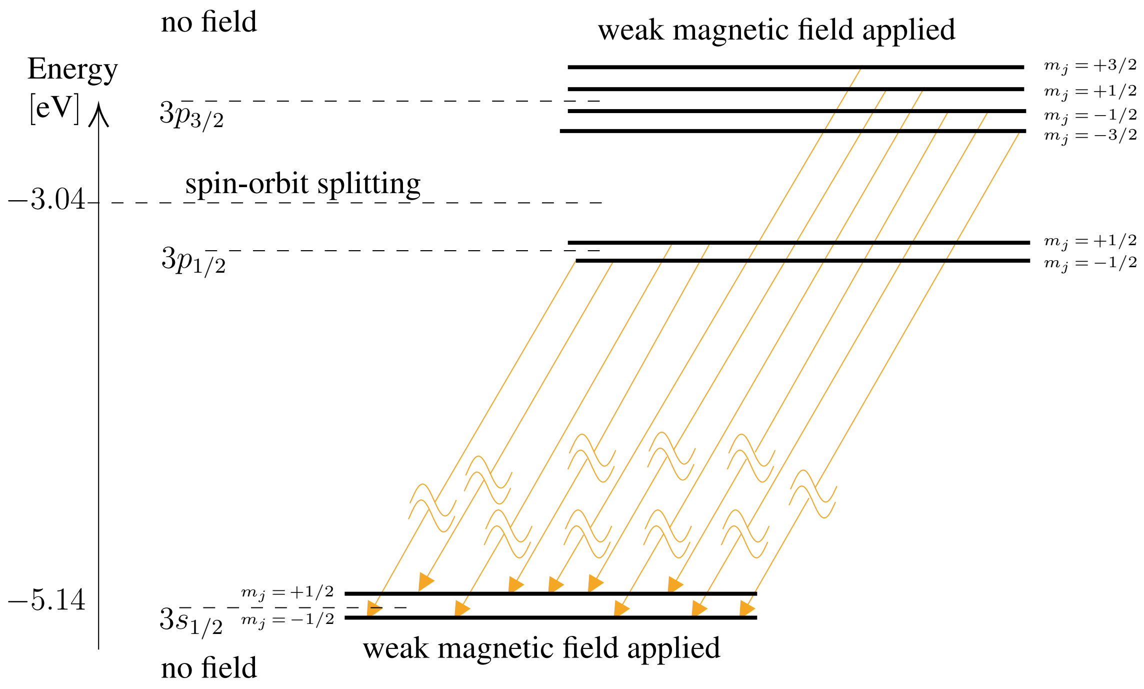

no field
weak magnetic field applied
Energy
[eV]
$3p_{3/2}$
$m_j = +3/2$
$m_j = +1/2$
$m_j = -1/2$
$m_j = -3/2$
spin-orbit splitting
-3.04
$3p_{1/2}$
$m_j = +1/2$
$m_j = -1/2$
-5.14
$3s_{1/2}$
$m_j = +1/2$
$m_j = -1/2$
weak magnetic field applied
no field

Figure 11.77: **General overview of Sodium splitting (source: Hyperphysics)**

After Thomson's work, Zeeman and Lorentz did further study of the influence of magnetic fields on the spectral emissions from atoms. By analysis of the splitting of the sodium doublet, they were able to demonstrate that the charge to mass ratio of the charge responsible for the splitting was the same as Thomson's electron. This was the first direct demonstration that electrons were involved in the production of the spectral line emissions.

11.2.5.2 Spin operator for spin $1/2$ particles

Let us now return to the relation demonstrated earlier above for $1/2$ spin particles:

$$\begin{aligned}\hat{S}^2\left|\frac{1}{2},\pm\frac{1}{2}\right\rangle &= \frac{3}{4}\hbar^2\left|\frac{1}{2},\pm\frac{1}{2}\right\rangle\\ \hat{S}_z\left|\frac{1}{2},\pm\frac{1}{2}\right\rangle &= \frac{\hbar}{2}\left|\frac{1}{2},\pm\frac{1}{2}\right\rangle\end{aligned} \tag{11.2.1009}$$

We have for sure obtained the eigenvalues, but it would be wise to determine the expression of the corresponding Spin operators $\hat{S}^2$ and $\hat{S}_z$. We have seen in the section of Spinor Calculus that:

$$\sigma_z = \begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix} \tag{11.2.1010}$$

and we proved that for the eigenvalue $+1$ the associated eigenvectors were:

$$\pm\begin{pmatrix}1\\0\end{pmatrix} \tag{11.2.1011}$$

We then have:

$$\begin{bmatrix}1 & 0\\0 & -1\end{bmatrix}\begin{pmatrix}1\\0\end{pmatrix}=1\begin{pmatrix}1\\0\end{pmatrix} \tag{11.2.1012}$$

By multiplying to the left and right by a familiar term:

$$\frac{\hbar}{2}\begin{bmatrix}1 & 0\\0 & -1\end{bmatrix}\begin{pmatrix}1\\0\end{pmatrix}=\frac{\hbar}{2}\begin{pmatrix}1\\0\end{pmatrix} \tag{11.2.1013}$$

And by analogy of the fact that the matrix of Pauli are special case of rotations matrices $\text{SU}(2)$ (see section Set Algebra page 1425), we put:

$$\hat{S}_z=\frac{\hbar}{2}\begin{bmatrix}1 & 0\\0 & -1\end{bmatrix} \tag{11.2.1014}$$

And of course we deduce:

$$\hat{S}^2=\left\|\vec{S}\right\|^2=\left(\frac{\hbar}{2}\sigma_x\right)^2+\left(\frac{\hbar}{2}\sigma_y\right)^2+\left(\frac{\hbar}{2}\sigma_z\right)^2=\frac{\hbar^2}{4}\left(\sigma_x^2+\sigma_y^2+\sigma_z^2\right)=\frac{3\hbar^2}{4}\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix} \tag{11.2.1015}$$

Therefore for spin-$1/2$ particles we have:

$$\boxed{\hat{S}^2=\frac{3\hbar^2}{4}\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\qquad \hat{S}_z=\frac{\hbar}{2}\begin{bmatrix}1 & 0\\0 & -1\end{bmatrix}} \tag{11.2.1016}$$

and we know that it is not the most explicit form as we should write the one we have determine in the section of Spinor Calculus for the Pauli matrices.

11.2.6 Planck's dimensions

It is time now to open a small parenthesis to finish about the Planck's constant that we promised in the section Principles of the chapter Mechanics to in-deep during our study of Wave Quantum Physics (because many textbooks mention what we will see without the precautions of rigour in our point of view). We have just seen that the measurement of objects depends on the Heisenberg's uncertainty principles. This precision affects both the time measurements and the particle trajectory or the energy density of the Universe. Let's see that by extension ... there are other possible speculative implications that lead to the following table:

Derived Unit	**Expression**	**SI Equivalent (approx.)**
Planck's length	$l_P=\sqrt{\dfrac{G\hbar}{c^3}}$	$1.616\cdot 10^{-35}$ [m]
Planck's surface	$l_P^2=\dfrac{G]\hbar}{c^3}$	$2.6121\cdot 10^{-70}$ [m^2]
Planck's volume	$l_P^3=\left(\dfrac{\hbar G}{c^3}\right)^{3/2}$	$4.2217\cdot 10^{-105}$ [m^3]
Planck's mass	$m_P=\dfrac{\hbar c}{G}$	$2.175\cdot 10^{-8}$ [kg]
Planck's time	$t_P=\dfrac{\hbar G}{c^5}$	$5.391\cdot 10^{-44}$ [s]
Planck's temperature	$T_P=\dfrac{\hbar c^5}{Gk^2}$	$1.416\cdot 10^{32}$ [T]
Planck's energy	$E_P=\sqrt{\dfrac{\hbar c^5}{G}}$	$1.956\cdot 10^{9}$ [J]
Planck's density	$\rho_P=\dfrac{c^5}{G^2\hbar}$	$5.155\cdot 10^{96}$ [kg $\cdot$ m^{-3}]
Planck's power	$P_P=\dfrac{c^5}{G}$	$3.629\cdot 10^{52}$ [W]
Planck's pulsation	$\omega_P=\sqrt{\dfrac{c^5}{G}}$	$1.855\cdot 10^{43}$ [rad $\cdot$ s^{-1}]
Planck's charge	$q_P=\sqrt{c\hbar 4\pi\varepsilon_0}$	$1.875\cdot 10^{-18}$ [C]
Planck's current	$I_P=\sqrt{\dfrac{c^6 4\pi\varepsilon_0}{G}}$	$3.479\cdot 10^{25}$ [A]
Planck's voltage	$U_P=\sqrt{\dfrac{c^4}{G4\pi\varepsilon_0}}$	$1.043\cdot 10^{27}$ [V]
Planck's impedance	$Z_P=\dfrac{1}{4\pi\varepsilon_0 c}$	29.98 [Ω]

Table 11.11: **Various common Planck dimensions**

We have proved previously at the beginning of this section that one of the uncertainties relations is given, when taking the module, by (from the order of Planck's constant to one given factor):

$$\big|\,[x,p_x]\,\big|\propto\hbar \tag{11.2.1017}$$

So, we can say roughly that at a fluctuation Δx of space, we can associate the linear momentum:

$$p_x=\frac{\hbar}{\Delta x} \tag{11.2.1018}$$

To that latter, according to our results of the section of Special Relativity, corresponds the relation of the energy $E=Mc^2=pc$, or the equivalent mass (by dividing by c^2) p/c. By denoting by M this mass associated with the perturbation Δx, we thus have:

$$M=\frac{\hbar}{c\cdot\Delta x} \tag{11.2.1019}$$

The gravitation due to this mass is characterized by a length R which we will determine in order of magnitude by writing that the potential energy associated with it (this assumes that Classical and Quantum gravitation are governed by the same laws...) GM^2/R (see section Astronomy page 5255), is equal to the mass-energy Mc^2. This gives:

$$R=\frac{GM}{c^2} \tag{11.2.1020}$$

or, by replacing M by its previous expression:

$$R=\frac{G\hbar}{c^3\Delta x} \tag{11.2.1021}$$

So that there is no self-amplification (and thus divergence) of the phenomenon of vacuum quantum fluctuation, we must preferably have $R=\Delta x$. By writing the equality between these two quantities, we arrive at a quantity which represents the minimum dimension (in order of magnitude) that physics can conceive according to the interpretation of some physicists. This is the famous "Planck's length" where we replace R by l_p:

$$\boxed{l_P=\sqrt{\frac{G\hbar}{c^3}}\cong 1.616\cdot 10^{-35}\;[\text{m}]} \tag{11.2.1022}$$

For which it corresponds the period or "Planck time", also named "Planck-Wheeler time", $t_P=l_p/c$ hence:

$$\boxed{t_P=\sqrt{\frac{G\hbar}{c^5}}\cong 5.391\cdot 10^{-44}\;[\text{s}]} \tag{11.2.1023}$$

We can now return to another more interesting expression of the fluctuating mass. Since:

$$M=\frac{\hbar}{c\cdot\Delta x}\Rightarrow l_P=\frac{\hbar}{c\cdot M}\quad\text{and}\quad l_P=\sqrt{\frac{G\hbar}{c^3}} \tag{11.2.1024}$$

We have therefore after rearranging the "Planck mass":

$$\boxed{m_P=\sqrt{\frac{\hbar c}{G}}\cong 2.176\cdot 10^{-8}\;[\text{Kg}]} \tag{11.2.1025}$$

Caution!!! The Planck mass is a good an easy example that we must not take all the different Planck values (mass, length, charge, density, energy, etc) as impossible values to reach (or to overtake in one direction or the other...)!

In 11973 (holocene calendar), Jacob Bekenstein published a paper ([47]) where he showed that the surface area of a Black Hole increases by $1A_p$ (i.e. the Planck Area - equal to l_p^2) for every bit of information which crosses the event horizon. Therefore via Bekenstein's result and basic calculus, the change in radius of a Black Hole is much smaller than than the 1 Planck length!

We will see also further below that the same conclusion applies for the Planck energy, the Planck impedance that have no extraordinary values and that can be easily reached and overtaken in laboratories and that Planck charge has a value larger than the electric charge of the electron... Indeed, let us recall that only evidence based experience can validate a Planck value as a limit (like the Planck quantum of action equal to h that is a fact beyond reasonable evidence).

Another way to get the Planck's mass but that gives a slightly different result is first to suppose that the main three universal constant are:

$$h\,[\mathrm{kg}\cdot\mathrm{m}^2\cdot\mathrm{s}] \qquad c\,[\mathrm{m}\cdot\mathrm{s}^{-1}] \qquad G\,[\mathrm{kg}^{-1}\cdot\mathrm{m}^3\cdot\mathrm{s}^{-2}] \tag{11.2.1026}$$

and that we seek to write a mass by combining h, c and G. So this is a dimensional analysis as:

$$m_p = h^{\alpha}c^{\beta}G^{\gamma} \tag{11.2.1027}$$

where we need to find the values of α, β and γ to get a mass...

We see quickly that to for the mass (kg) we must have:

$$1 = \alpha - \gamma \tag{11.2.1028}$$

for the distance (m), we must have:

$$0 = 2\alpha + \beta + 3\gamma \tag{11.2.1029}$$

and for the time (s):

$$0 = -\alpha - \beta - 2\gamma \tag{11.2.1030}$$

So we just need to solve this linear system:

$$\begin{aligned} 0 &= 2\alpha + \beta + 3\gamma \\ 1 &= \alpha - \gamma \\ 0 &= -\alpha - \beta - 2\gamma \end{aligned} \tag{11.2.1031}$$

So first iteration:

$$\begin{aligned} 0 &= 2\alpha + \beta + 3\gamma \\ 1 &= \alpha - \gamma \\ 0 &= \alpha + \gamma \end{aligned} \tag{11.2.1032}$$

Second iteration:

$$\begin{aligned} 0 &= 2\alpha+\beta+3\gamma \\ 2 &= \alpha \\ 1 &= -2\gamma \end{aligned} \quad (11.2.1033)$$

Third iteration:

$$\begin{aligned} \alpha &= \frac{1}{2} \\ \gamma &= -\frac{1}{2} \\ 0 &= 1+\beta-\frac{3}{2} \end{aligned} \quad (11.2.1034)$$

so finally:

$$\alpha=\frac{1}{2} \qquad \beta=\frac{1}{2} \qquad \gamma=-\frac{1}{2} \quad (11.2.1035)$$

Therefore:

$$m_P=h^{1/2}c^{1/2}G^{-1/2}=\sqrt{\frac{hc}{G}} \quad (11.2.1036)$$

to a given multiplicative dimensionless constant... This missing multiplicative constant is known to us as we have determined the Planck's mass earlier before. So with dimensional analysis we should consider $\hbar$ as being more fundamental than h.

Dimensional analysis gives to a given constant factor and according to the virial theorem (see section Continuum Mechanics page 3897):

$$E=kT \quad (11.2.1037)$$

and therefore:

$$E=m_Pc^2=pc=kT\Rightarrow\frac{p}{c}=\frac{kT}{c^2}=m_p\Rightarrow\sqrt{\frac{\hbar c}{G}}=\frac{kT}{c^2} \quad (11.2.1038)$$

Hence the "Planck's temperature":

$$\boxed{T_P=\sqrt{\frac{\hbar c^5}{Gk}}\cong 1.42\cdot 10^{32}\;[\mathrm{K}]} \quad (11.2.1039)$$

and "Planck's energy":

$$\boxed{E_P=m_Pc^2=\sqrt{\frac{\hbar c}{G}}c^2=\sqrt{\frac{\hbar c^5}{G}}\cong 1.956\cdot 10^{9}\;[\mathrm{J}]} \quad (11.2.1040)$$

After all this, we easily get the "Planck's density":

$$\rho_P=\frac{M_P}{\lambda_P^3}=\frac{\sqrt{\dfrac{\hbar c}{G}}}{\sqrt{\dfrac{G\hbar}{c^3}}^3}=\frac{\sqrt{\dfrac{\hbar c}{G}}\sqrt{\dfrac{c^3}{G\hbar}}}{\dfrac{G\hbar}{c^3}}=\sqrt{\frac{c^4}{G^2}\frac{c^3}{G\hbar}}=\sqrt{\frac{c^{10}}{G^4\hbar^2}}=\frac{c^5}{G^2\hbar} \quad (11.2.1041)$$

That is:

$$\boxed{\rho_P \cong 5.1 \cdot 10^{96}\,[\mathrm{kg} \cdot \mathrm{m}^{-3}]} \tag{11.2.1042}$$

We can have fun getting other Planck values, but that have no physical experimental justification excepted for most of them invalid empirical and subjective interpretations (and we could go on for a long time with so many other magnitudes as we have give a quite exhaustive list in the section Principles of the Mechanics chapter). For example:

- The "Planck's force":

$$F_P = m_P \frac{\lambda_P}{t_P^2} = \sqrt{\frac{\hbar c}{G}} \frac{\sqrt{\dfrac{G\hbar}{c^3}}}{\left(\sqrt{\dfrac{G\hbar}{c^5}}\right)^2} = \frac{c^4}{G} \tag{11.2.1043}$$

  Therefore:

$$\boxed{F_P \cong 1.21 \cdot 10^{44}\,[\mathrm{N}]} \tag{11.2.1044}$$

- The "Planck's power":

$$\boxed{P_P = \frac{E_P}{t_P} = \frac{c^5}{G} \cong 3.629 \cdot 10^{52}\,[\mathrm{W}]} \tag{11.2.1045}$$

- The "Planck's pulsation":

$$\boxed{\omega_P = 2\pi f_P = \frac{1}{t_P} = \sqrt{\frac{c^5}{\hbar G}} \cong 1.855 \cdot 10^{43}\,[\mathrm{rad} \cdot \mathrm{s}^{-1}]} \tag{11.2.1046}$$

- By proceeding with the same initial reasoning that the one done with mass, but using the potential electrostatic energy instead of gravitational potential energy we can get the "Planck's electric charge":

$$\boxed{q_P = \sqrt{c\hbar 4\pi\varepsilon_0} \cong 1.875 \cdot 10^{-18}\,[\mathrm{C}]} \tag{11.2.1047}$$

- Then we can calculate a "Planck's current":

$$\boxed{I_P = \frac{Q_P}{T_P} = \sqrt{\frac{c^6 4\pi\varepsilon_0}{G}} \cong 3.479 \cdot 10^{25}\,[\mathrm{A}]} \tag{11.2.1048}$$

- And also the "Planck's voltage":

$$\boxed{U_P = \frac{E_P}{q_P} = \sqrt{\frac{c^4}{G 4\pi\varepsilon_0}} \cong 1.0432 \cdot 10^{27}\,[\mathrm{V}]} \tag{11.2.1049}$$

- ...and the "Planck's impedance" (...):

$$\boxed{Z_P=\frac{V_P}{I_P}=\frac{1}{4\pi\varepsilon_0 c}=\frac{Z_0}{4\pi}=29.98\,[\Omega]} \tag{11.2.1050}$$

- and so on[22]...

Remark

Some physicists have used (and still use) the above results for wacky and dangerous reasonings that are only hazardous interpretations! It is therefore advisable to take with care all the information concerning the Planck dimensions that you could find in textbooks and Internet (even if these latter seem very sympathetic!). The best known example is given by the Compton wavelength λ_C (see section Nuclear Physics page 5098) which depends on the mass-energy of the photon. If this wavelength is equal to the classical Schwarzschild radius for the same mass-energy (see section General Relativity page 5793), then in this case its value is that of the Planck's length and its mass is equal to the Planck's mass. It is then tempting to say that the particle then forms a Black Hole. But this is an analogy because in this case, nothing tells us that the expression of the Schwarzschild radius applies to quantum physics... !

---

[22] see section Principia of the chapter Mechanics page 3122 to see a most exhaustive and structured list

11.2.7 Wave Quantum Physics Interpretation

Ok now that we have some mathematical basics on Quantum Physics, before we continue with Relativistic Quantum Physics and Quantum Field theory we think that it is the good time to speak about the main existing interpretations of the results and mechanisms of the different models and results we get so far[23]. Indeed, we have seen that Quantum mechanics cannot, for most people, easily be reconciled with everyday language and observation. Its interpretation has often seemed counter-intuitive to physicists, including some of its inventors.

11.2.7.1 Copenhagen interpretation

In 11930 (holocene calendar), the probabilistic interpretation of the wave amplitude of a particle and the Heisenberg's uncertainty principle constitute the elements of the "non-deterministic" standard interpretation of quantum physics as we have already mention it at the beginning of this section. This interpretation is often referred to as the "Copenhagen interpretation", as Niels Bohr who contributed largely to it, headed a renowned institute of physics at that time in this city. However many physicists, such as Albert Einstein and Erwin Schrödinger, who accepted the mathematical formulation of quantum physics, were not at ease with the Copenhagen interpretation of and criticized it. And up to the present day (still in year 12017 according to holocene calendar), the question of the correct interpretation of mathematical formulation remains a problem.

Indeed, we can ask ourselves the following question: Where is the reality? Is there a reality? Niels Bohr answers «*No*»: there is nothing at the quantum level, reality exists or appears only during a measurement. This view shared by most physicists (Copenhagen interpretation) implies that the measure "creates" the position of the electron. In other words: No elementary phenomenon is a real phenomenon before being an observed phenomenon!

Albert Einstein thought that quantum physics, although very effective and very impressive, is not complete and gives only an imperfect picture of the quantum world. For him, there would be something else beyond that which would clarify and refine our present vision (in the same way as the theory of gases for which we had to wait for statistical models, Albert Einstein thought that hidden variables had yet to be discovered).

Thus, in the Copenhagen interpretation of quantum mechanics, the principle of uncertainty means that, at an elementary level, the physical universe no longer exists in a deterministic way, but rather as a series of probabilities or potentials. For example, the pattern produced by millions of photons passing through a diffraction slot can be calculated using quantum mechanics, but the path of each photon can not be predicted by any known method following the Copenhagen interpretation. It is this interpretation that Albert Einstein doubted when he said: «*I can not believe that God is playing dice with the Universe*». From a physical as well as a philosophical point of view, the principle of uncertainty implies the refutation of the universal determinism defended by Laplace at the beginning of the 119th century (holocene calendar).

An instantaneous reduction of all the possible states occurs from the observation of the system according to the Copenhagen interpretation (wave function collapse). This random decision of the observed state respects the probabilities, corresponding to the square of the amplitudes of the states as we have seen it. In addition, the Copenhagen interpretation states that during a measurement, a reduction process, originating from the macroscopic object, eliminates superposition of quantum states.

---

[23] For some physicists, interpret mathematics is a non-sense ans anyway human brain is limited. But as many people like to interpret stuff we still dedicate a subsection about this

The Copenhagen interpretation therefore leads to the problem of measurement, since the Schrödinger cat's thought experiment states that when we measure a quantity, such as position or momentum, we intervene in the process of measurement by causing a radical change of the quantum state of the wave function. We change the quantities measured in an unpredictable manner and this (transition) state can not be described by the Schrödinger equation. Physicists and philosophers have reacted in several ways to this interpretation:

- Either we consider as Niels Bohr and Werner Heisenberg that this principle is law and that it is better not to seek the ultimate interpretation as our brain capacity is anyway limited in comparison to the powerful investigation tool that are mathematics. This attitude is accepted by most physicists.

- Either we consider quantum physics to be an incomplete theory and some, such as Albert Einstein, Eugene Wigner or David Bohm (de Broglie-Bohm pilot-wave interpretation that we will introduce very briefly further below), have not hesitated to seek other solutions, that seems sterile until now.

- Finally, Hugh Everett and many others take Schrödinger's equation very seriously, considering it a representation of reality. They consider that the Copenhagen interpretation really represents the evolution of the wave function. The different terms of the equation would correspond to the different levels of energy in which the system is located. The reduction of the wave packet would be interpreted as a total division of the object and the measuring instrument into parallel universes (many-worlds interpretation).

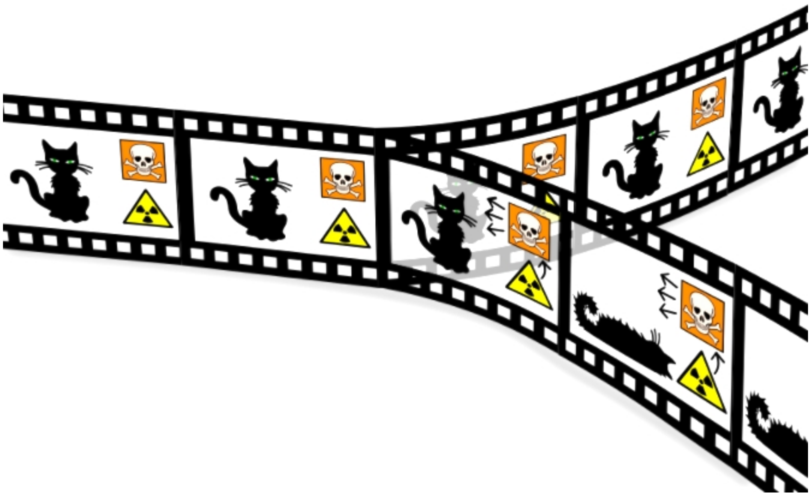


Figure 11.78: Schrödinger's cat paradox is resolved by assuming that different branches of the Universe are created with each possible outcome (source: Wikipedia, author: Christian Schirm)

Today the debate remains open, but several experiments carried out since the 11930s (holocene calendar) allow us, step by step, to dispel the thick fog that covers the bottom of reality and answer a few questions. However, all these experiments confirm that the period of certainties is long gone. The most famous experiment remaining the "EPR paradox" following the publication of an article by Einstein, Podolsky and Rosen whose sole purpose is to undermine the Copenhagen interpretation.

11.2.7.1.1 EPR paradox

The original article being a bit difficult (in a future version of this textbook we will perhaps write the underlying mathematics in details), we will take the simplified version of use in small schools but that is used in laboratories, originally proposed by David Bohm. Whereas originally the paradox was presented with the pair (position, linear momentum), Bohm proposed to use the fermion spin which is a purely binary quantum property much more easy to deal with.

As an electron can have only two states of spin "top" $|+\rangle$ or "bottom" $|-\rangle$ (in the famous z orient magnetic field experiment), so the EPR experiment proposed by David Bohm consists in taking a null spin particle which disintegrates, thus producing two electrons A and B each being in the superposition of states $|-+\rangle$ or $|+-\rangle$. Since their combined spin must remain zero, one of the electrons must have its spin at the top and the other at the bottom when their are measured. The electrons are the driven in opposite directions until the distance separating them is large enough to eliminate any physical interaction between them, and the spin of each electron is measured exactly at the same instant using a spin detector.

According to Niels Bohr, as long as no measurements have been made, neither electron A nor electron B possess a spin pre-existing in any direction. Instead, before being observed, electrons exist in a superposition of states, so that they are up and down at the same time. Since the two electrons are entangled, the information concerning the state of their spin is given by a wave function of the type:

$$\Psi=(A\text{ spin }\uparrow\text{ and }B\text{ spin }\downarrow)\text{ or }(A\text{ spin }\downarrow\text{ and }B\text{ spin }\uparrow)=|+-\rangle+|-+\rangle \tag{11.2.1051}$$

The electron A has no determined spin component before a measurement is performed. The measurement collapses the wave function of the system A and B, after which it is either "top" or "bottom". At that moment, its partner B acquires the opposite spin in the same direction, even though it is at the other side of the Universe violating therefore Special Relativity postulates (no information can go faster than the speed of light c). The Copenhagen interpretation is then named "non-local interpretation" whereas Albert Einstein believed in local realism: that is, a particle can not be instantly influenced by a distant event, and its properties exist independently of any measure.

However, Bohm's approach has a flaw that Albert Einstein would probably have used as an argument: correlations could be explained by arguing that the two electrons each have spin values defined on each of the three axes x, y, z. Whether measured or not. So again, according to Albert Einstein, the fact that the pre-existing spin states of the electron couple can not be taken into account by wave quantum physics would still have been a proof of its incompleteness (hidden variables).

A physicist, John Bell, however, had the idea of an experimental and theoretical means to break the impasse of the EPR paradox by changing the relative orientation of the two spin detectors.

Thus, if the detectors measuring the electron spin A and B are aligned so as to be parallel, then there is a 100% correlation between the two sets of measurements each time the spin is measured at "top" by the first detector, the "bottom" spin is measured by the other detector, and vice versa. If one of the detectors is turned slightly, they are no longer aligned. Now, if we measure the spin state of many pairs of entangled electrons, when we find "top" for electron A, the corresponding measure for B will sometimes give "top" too. Increasing the angle between the axes of the two detectors therefore leads to a reduction in the degree of correlation. If the detectors are at right angles to one another and the experiment is repeated a large number of times, it is only in half of the cases that a spin is detected at the "bottom" for B when we detect a "top" spin for A following the z-axis. If the detectors are oriented 180 degrees apart, the electron couple will be completely anti-correlated. If the measurement gives "top" for the spin state of A, then the spin state of B will be "bottom".

Although it is an imaginary experiment, it is possible to calculate the exact degree of spin correlation for a given orientation of the detectors, as predicted by quantum theory. However, it is not possible to perform a similar calculation by using an archetypal hidden variable theory and conserving the locality. The only thing that such a theory could predict would be an imperfect coupling between the spin states of A and B. However, rigorously, it is insufficient to choose between quantum theory and a local theory with hidden variables. John Bell made an astonishing discovery. It was possible to decide between the predictions of quantum mechanics and those of any hidden-variable theory by measuring the correlations of electron pairs for a given configuration of the detectors and then repeating the experiment with a different orientation. This allowed John Bell to calculate the total correlation for the two orientation configurations in terms of individual results predicted by any local theory with hidden variables. Since, in any such theory, the result of a measurement performed by a detector can not be affected by what is measured with the other, it is possible to distinguish between hidden variables and quantum mechanics.

John Bell succeeded in calculating the limits of the degree of spin correlation between pairs of entangled electrons in a Bohm-modified EPR experiment (if we have the time we will provide the mathematical details in the future). He found that in the ethereal kingdom of quanta there is a greater degree of correlation if quantum mechanics reigns as an absolute rule than in any universe that depends on hidden variables and locality. Bell's theorem stated that no local theory with hidden variables could reproduce the same set of correlations as quantum mechanics. Any local theory with hidden variables would lead to spin correlations generating numbers, name obviously "correlation coefficients", between -2 and $+2$. However, for some orientations of spin detectors, quantum mechanics predicts correlation coefficients that lie outside the range, named "Bell inequalities", ranging from -2 to $+2$.

The Bell theorem thus makes it possible to test face to face the local reality advocated by Albert Einstein with Niels Bohr Copenhagen interpretation, namely that the quantum universe exists independently of observation and that physical effects can not be transmitted to a speed higher than that of light. Bell had carried the Einstein-Bohr debate into a new arena: experimental philosophy. If Bell's inequality resisted, then Einstein's assertion that quantum mechanics was incomplete would be accurate. If this inequality were to be violated, Bohr would triumph. No more thought experiences! It would now be Einstein vs Bohr in the lab.

The first experiment that tested the Bell's inequalities used pairs of photons instead of pairs of electrons. This change was possible because the photons possess the polarization property, which for the needs of the test played the role of the quantum spin (moreover, the photons are easier to manipulate). This is certainly a simplification, but a photon can be considered to be polarized as "top" or "bottom". Like the spin of the electron, if the polarization of one of the photons on the z-axis is measured as "top", then the measurement of the other will give "bottom", since the combined polarizations Of the two photons must result in a zero spin (polarization state).

The results violated Bell's inequalities, which was in favour of Bohr's non-local Copenhagen interpretation and against the local reality supported by Albert Einstein.

Bell derived this inequality from two assumptions (hypothesis):

H1. First, there is a reality independent of the observer. This is reflected in the fact that a particle has a well-defined property such as spin before it is measured.

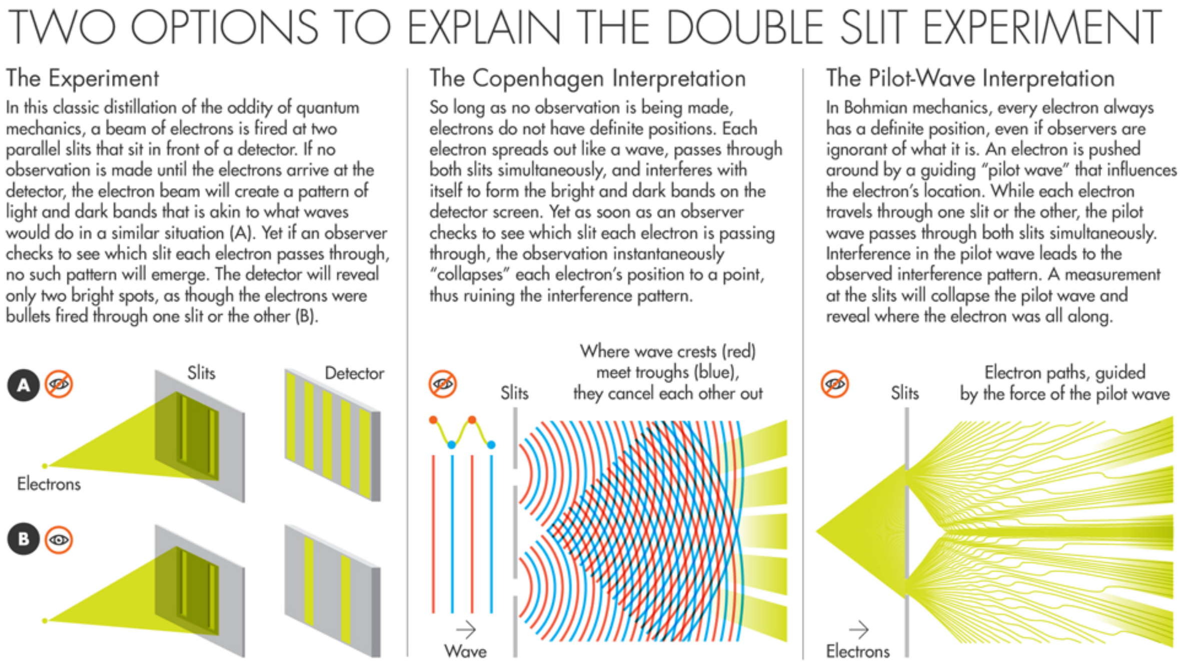
H2. Second, the locality is conserved. There is no supra-luminous influence, so that what happens here can not instantly affect what happens elsewhere.

The experimental results mean that one of these two assumptions must be abandoned, but which one? Bell was willing to abandon the locality.

11.2.7.2 De Broglie-Bohm (pilot-wave) interpretation

The "de Broglie–Bohm theory", also known as the "pilot-wave theory" or "Bohm's interpretation" is another interpretation of quantum theory. The theory is deterministic and explicitly non-local: the velocity of any one particle depends on the value of the guiding equation, which depends on the configuration of the system given by its wavefunction; the latter depends on the boundary conditions of the system, which in principle may be the entire universe...

The theory was historically developed in the 11920s (according to holocene calendar) by de Broglie, who in 11927 (according to holocene calendar) was persuaded to abandon it in favour of the then-mainstream Copenhagen interpretation. David Bohm, dissatisfied with the prevailing orthodoxy, rediscovered de Broglie's pilot-wave theory in 11952 (according to holocene calendar). Bohm's suggestions were not widely received then, partly due to reasons unrelated to their content, connected to Bohm's youthful communist affiliations. De Broglie–Bohm theory was widely deemed unacceptable by mainstream theorists, mostly because of its explicit non-locality. Bell's theorem (11964 according to holocene calendar) was inspired by Bell's discovery of the work of David Bohm and his subsequent wondering whether the obvious non-locality of the theory could be eliminated. Since the 11990s (holocene calendar), there has been renewed interest in formulating extensions to de Broglie–Bohm theory, attempting to reconcile it with Special Relativity and Quantum Field Theory, besides other features such as spin or curved spatial geometries.



TWO OPTIONS TO EXPLAIN THE DOUBLE SLIT EXPERIMENT

The Experiment

In this classic distillation of the oddity of quantum mechanics, a beam of electrons is fired at two parallel slits that sit in front of a detector. If no observation is made until the electrons arrive at the detector, the electron beam will create a pattern of light and dark bands that is akin to what waves would do in a similar situation (A). Yet if an observer checks to see which slit each electron passes through, no such pattern will emerge. The detector will reveal only two bright spots, as though the electrons were bullets fired through one slit or the other (B).

The Copenhagen Interpretation

So long as no observation is being made, electrons do not have definite positions. Each electron spreads out like a wave, passes through both slits simultaneously, and interferes with itself to form the bright and dark bands on the detector screen. Yet as soon as an observer checks to see which slit each electron is passing through, the observation instantaneously "collapses" each electron's position to a point, thus ruining the interference pattern.

The Pilot-Wave Interpretation

In Bohmian mechanics, every electron always has a definite position, even if observers are ignorant of what it is. An electron is pushed around by a guiding "pilot wave" that influences the electron's location. While each electron travels through one slit or the other, the pilot wave passes through both slits simultaneously. Interference in the pilot wave leads to the observed interference pattern. A measurement at the slits will collapse the pilot wave and reveal where the electron was all along.

A
Slits
Detector
Electrons
B
Where wave crests (red) meet troughs (blue), they cancel each other out
Slits
Wave
Electron paths, guided by the force of the pilot wave
Slits
Electrons

We strongly recommend the reader to go see on Internet videos about the "walking droplets[24]" also named "walker" to understand better the idea behind the concept of pilot wave.

In the below gallery (read from left to right and up to bottom) thanks to a little layer of air, droplets of silicone oil can continuously bounce on the surface of the pool beneath them without being absorbed. But even though the droplets don't directly connect with the pool beneath them, they still make waves, and then interact with those waves, resulting in a forward momentum.

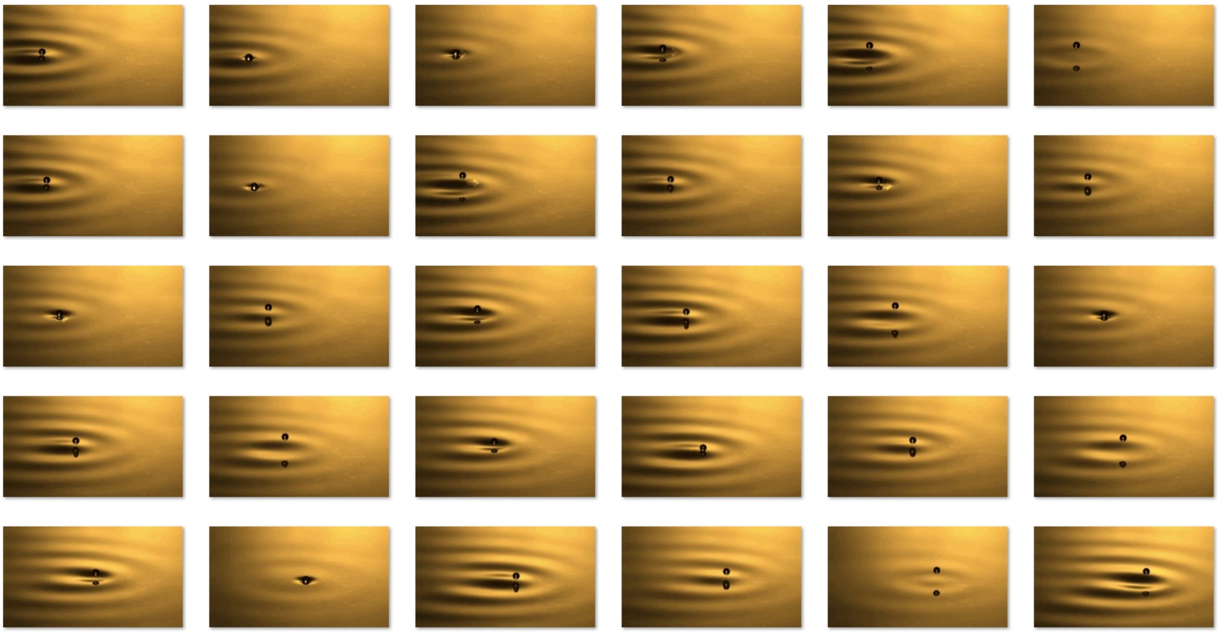


Figure 11.79: **Thirty successive pictures of a walker**

---

[24] hydrodynamic quantum analogue involving bouncing fluid droplets over a vibrating fluid bath

All observed phenomena are related to the particle-wave interaction. The droplet moves in a medium modified by previously generated waves.As the droplet moves, the points of the interface it visits keep emitting waves.The wavefield has thus a complex structure that contains a "memory"of the path.

While this experiment isn't on the quantum scale, it does help to demonstrate the way quantum-scale particles may operate according to the pilot wave theory. And for any lay people who have struggled with grasping why things are so strange on the quantum scale according to the standard interpretation, this pilot wave theory provides a far more palatable framework for understanding quantum mechanics.

And what about walkers facing a double split? Here you are (still read from left to right and up to bottom):

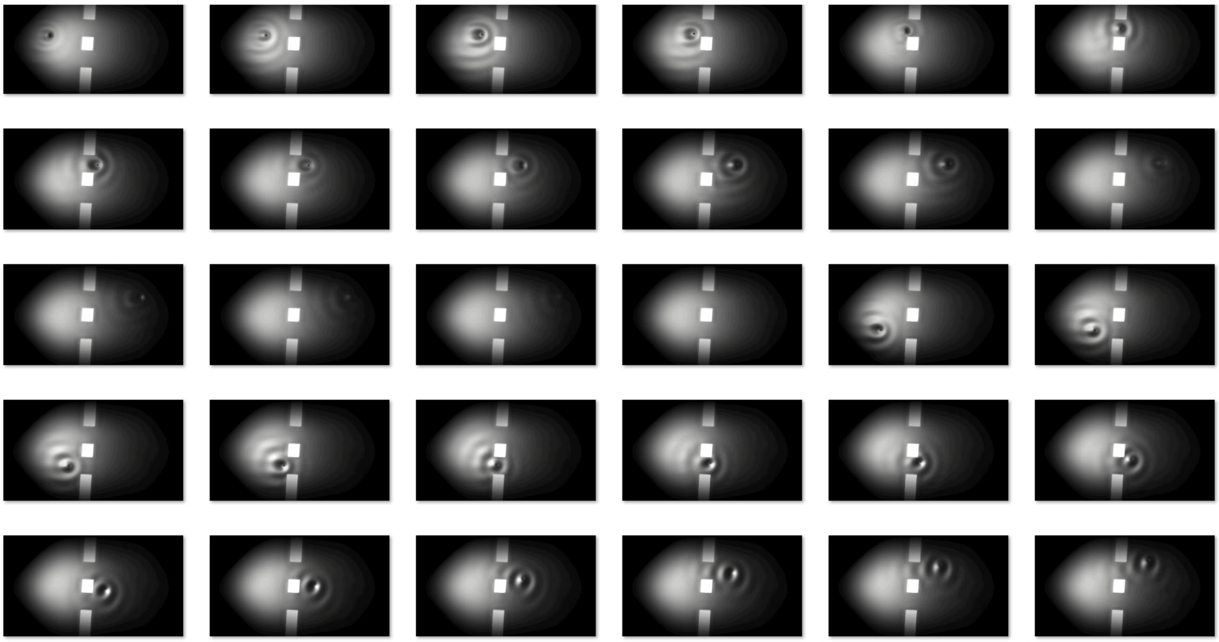


Figure 11.80: **Double split experiment with two walkers**

The diffraction pattern obtained with the walker experiment is highly similar tot that of the diffraction pattern of photon or electron with the same experimental configuration.

In a survey made on 81 professional physicists (see [96]) in 12024 (holocene calendar) we get the following result for the fate of the matter that falls into a Black Hole:

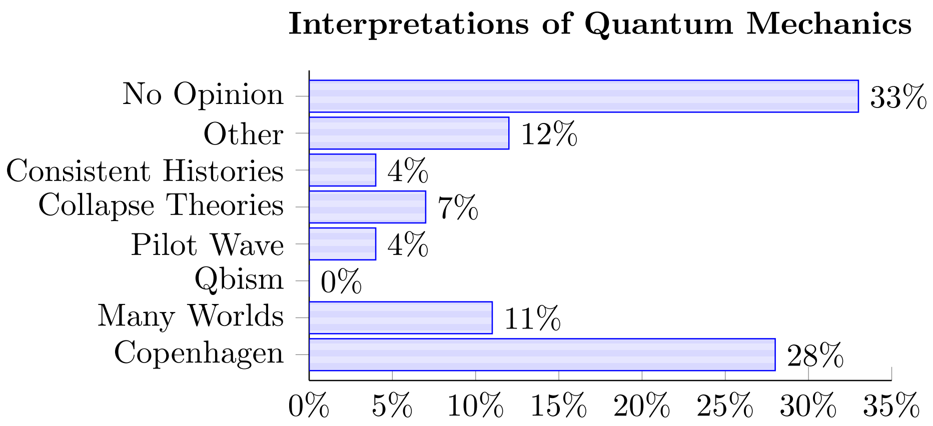


Figure 11.81: **Copenhagen survey on interpretations of Quantum Physics**

If we have the time in the future we will perhaps write a formal introduction (i.e. mathematical introduction) to this theory.

~90 %

★★★★☆

61 votes, 74.43%

11.3. Relativistic Quantum Physics

THE attentive reader will have noticed that quantum mechanics (wave quantum physics) is a non-relativistic theory: it does not incorporate the principles of Einstein's Special Relativity (see section Relativity page 5628). We will strive to fill this gap by studying now the relativistic version of the wave quantum physics.

Relativistic quantum mechanics (RQM) is any Poincaré covariant formulation of quantum mechanics. This theory is applicable to massive particles propagating at all velocities up to those comparable to the speed of light c, and can accommodate massless particles. The theory has application in high energy physics,particle physics and accelerator physics, as well as atomic physics, chemistry and condensed matter physics. Non-relativistic quantum mechanics refers to the mathematical formulation of quantum mechanics applied in the context of Galilean relativity, more specifically quantizing the equations of classical mechanics by replacing dynamical variables by operators. Relativistic quantum mechanics is quantum mechanics applied with Special Relativity, but not General Relativity. An attempt to incorporate General Relativity into quantum theory is the subject of quantum gravity, yet an unsolved problem in physics, although some theories, such as the Kaluza-Klein, have been proposed but are unfounded and without proof. Although the earlier formulations, like the Schrödinger picture and Heisenberg picture were originally formulated in a non-relativistic background, these pictures of quantum mechanics also apply with Special Relativity.

The relativistic formulation is more successful than the original quantum mechanics in some contexts, in particular: the prediction of antimatter, electron spin, spin magnetic moments of elementary spin-$1/2$ particles, fine structure, and quantum dynamics of charged particles in electromagnetic fields. The key result is the Dirac equation, from which these predictions emerge automatically. By contrast, in quantum mechanics, terms have to be introduced artificially into the Hamiltonian operator to achieve agreement with experimental observations.

Nevertheless, RQM is only an approximation to a fully self-consistent relativistic theory of known particle interactions because it does not describe cases where the number of particles changes; for example in matter creation and annihilation. By yet another theoretical advance, a more accurate theory that allows for these occurrences and other predictions is Relativistic Quantum Field Theory in which particles are interpreted as field quanta (see article for details).

Before we tackle the mathematical part, we emphasize and reiterate that we will limit ourselves only to theoretical developments made between 11910 and about 11935 according to holocene calendar (beyond the complexity of theories requires too many pages to a general textbook like this one).

11.3.1 Relativistic Schrödinger evolution equation

Particle physics can not be properly and completely described in the context of quantum mechanics as described by the Schrödinger equation as the later is constructed on the incomplete version of the Einstein's relation ($E=mc^2$ instead of $E^2=c^2p^2+m_0^2c^4$). As the energies are generally superior to the masses of the particles, it is necessary, in addition, to work in the context of the full theory of Special Relativity. Let's see how to include it in a first basic approach.

The energy-momentum of a free particle of mass m, satisfied as we have proved in the section of Special Relativity the relation:

$$E^2=c^2p^2+m_0^2c^4 \tag{11.3.1}$$

We seek to quantify this equation. For this, we will return to the relations we have proved in the study of functional linear operators and of the evolutionary Schrödinger equation.

Recall that the momentum is described by the relation (using the divergence operator):

$$p=-\mathrm{i}\hbar\vec{\nabla} \tag{11.3.2}$$

and the total energy by:

$$E=\mathrm{i}\hbar\frac{\partial}{\partial t} \tag{11.3.3}$$

Both relations have been proved in details in the section of Wave Quantum Physics!

The substitutions of the two previous relations applied to the equality:

$$E^2=c^2p^2+m_0^2c^4 \tag{11.3.4}$$

leads to:

$$(\mathrm{i}\hbar\partial_t)^2=c^2(-\mathrm{i}\hbar\vec{\nabla})^2+m_0^2c^4 \tag{11.3.5}$$

Applying the squares:

$$-\hbar^2\partial_t^2=-c^2\hbar^2\vec{\nabla}^2+m_0^2c^4 \tag{11.3.6}$$

and rearranging and using the notation $\vec{\nabla}^2=\Delta$:

$$-\hbar^2\partial_t^2+c^2\hbar^2\Delta=m_0^2c^4 \tag{11.3.7}$$

And multiplied by the wave function Ψ on both sides of equality lead to development:

$$(-\hbar^2\partial_t^2+\hbar^2\Delta c^2)\Psi=m_0^2c^4\Psi \tag{11.3.8}$$

Hence:

$$(-\hbar^2\partial_t^2+\hbar^2\Delta c^2)\Psi\Leftrightarrow-\hbar^2\left(\frac{1}{c^2}\partial_t^2-\triangle\right)\Psi=m_0^2c^2\Psi \tag{11.3.9}$$

And rearranged:

$$\left(\hbar^2\left(\frac{1}{c^2}\partial_t^2-\triangle\right)+m_0^2c^2\right)\Psi=0 \tag{11.3.10}$$

Using the d'Alembertian (see section Electrodynamics page 4118), we can write this latter relation in a more condensed form named "relativistic Schrödinger evolution equation" or more frequently "free Klein-Gordon equation" or also named "free Klein-Gordon-Fock equation" (in the absence of a magnetic field!):

$$\boxed{(\hbar^2\Box+m_0^2c^2)\Psi=0} \tag{11.3.11}$$

with the signature $(+,-,-,-)$ for the partial derivative:

$$\Box=\frac{\partial^2}{c^2\partial t^2}-\frac{\partial^2}{\partial x^2}-\frac{\partial^2}{\partial y^2}-\frac{\partial^2}{\partial z^2} \tag{11.3.12}$$

Some authors prefer the signature $(-,+,+,+)$ therefore we have (this is also the choices we will make in the section of Quantum Field Theory as it is the tradition...):

$$(\hbar^2\Box-m_0^2c^2)\Psi=0 \tag{11.3.13}$$

with:

$$\Box=-\frac{\partial^2}{c^2\partial t^2}+\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2} \tag{11.3.14}$$

Remark

In particle physics, the equation $(\hbar^2\Box+m_0^2c^2)\Psi=0$ is sometimes named the "bosons relativistic covariant equation" whatever the chosen signature.

The free Klein-Gordon equation is often given as follows (more aesthetic):

$$\left(\Box+\left(\frac{m_0c}{\hbar}\right)^2\right)\Psi=0 \tag{11.3.15}$$

It is important to note that the Klein-Gordon equation involves scalar and therefore characterizes zero spin particles.

Remarks

R1. We can verify that the plane waves of the form:

$$\Psi=\Psi_0e^{\frac{\mathrm{i}}{\hbar}(\vec{p}\circ\vec{r}-E_{\text{tot}}t)} \tag{11.3.16}$$

are solutions of the free Klein-Gordon equation (we'll go more into details in the section Elementary Particles Physics).

R2. We will come back during our study of the Dirac equation and fermions spins on the free Klein-Gordon equation (to generalize it).

11.3.1.1 Antimatter

When we have proved the free Klein-Gordon equation, we purposely left aside a very interesting case of development that we have made.

Maybe you did not notice it, but the equation:

$$E^2=c^2p^2+m_0^2c^4 \tag{11.3.17}$$

can take two values for a given pulsation:

$$E=\pm\sqrt{c^2p^2+m_0^2c^4} \tag{11.3.18}$$

one positive and the other negative. The value of energy could therefore take all values in $[-\infty,+\infty]$...!!!

So far we have implicitly assumed in Classical Mechanics that negative solutions were not natural and should therefore simply be discarded. This cannot be done in quantified theory without lead to serious inconsistencies. Rather than ignore these negative energy solutions, we should find them a physical interpretation.

We notice first that all the negative energies are authorized by the above equation (as well as for positive energy). We say then that negative energy states are occupied but not observable; the electrons are named "virtual electrons".

Theorem 11.5. Let us imagine a wave packet formed by a superposition of plane waves at a narrow interval in pulse. This packet moves in space. In the one-dimensional case, it propagates at the speed:

$$v=\frac{\mathrm{d}E}{\mathrm{d}p}=\frac{c^2p}{E} \tag{11.3.19}$$

Proof 11.5.1. Always based on the assumption that the potential field is zero, we have:

$$E=\frac{1}{2}mv^2\Rightarrow\frac{\mathrm{d}E}{\mathrm{d}v}=mv\Rightarrow\frac{\mathrm{d}E}{m\mathrm{d}v}=v\Rightarrow\frac{\mathrm{d}E}{\mathrm{d}p}=v \tag{11.3.20}$$

and:

$$\begin{gathered}E^2=c^2p^2+m_0^2c^4\Leftrightarrow\mathrm{d}(E^2)=E\mathrm{d}E+\mathrm{d}EE=2E\mathrm{d}E\\ \Rightarrow\frac{\mathrm{d}E^2}{\mathrm{d}p}=\frac{2E\mathrm{d}E}{\mathrm{d}p}=\frac{\mathrm{d}(c^2p^2+m_0^2c^4)}{\mathrm{d}p}=2c^2p\\ \Rightarrow\frac{2E\mathrm{d}E}{\mathrm{d}p}=2c^2p\Leftrightarrow\frac{E\mathrm{d}E}{\mathrm{d}p}=c^2p\Rightarrow\frac{\mathrm{d}E}{\mathrm{d}p}=\frac{c^2p}{E}\end{gathered} \tag{11.3.21}$$

So to summarize:

$$v=\frac{\mathrm{d}E}{\mathrm{d}p}=\frac{c^2p}{E} \tag{11.3.22}$$

■ Q.E.D.

Let us consider first a particle of positive energy E^+. Its position in function of time is given by:

$$E^+ > 0: \quad x(t) = x_0 + \frac{c^2 p}{E^+} t \tag{11.3.23}$$

A particle of negative energy E^- moves following:

$$E^- < 0: \quad x(t) = x_0 + \frac{c^2 p}{E^-} t = x_0 + \frac{c^2 p}{|E^-|}(-t) \tag{11.3.24}$$

In other words, and this will be our first conclusion, we can say that a particle of negative energy $E^- < 0$ is equivalent to a particle of positive energy $|E^-| > 0$ moving backwards in time and this is what we name an "antiparticle".

Now we have to see what is the interpretation of a particle moving backwards in time:

To simplify, we consider a non-relativistic particle of electric charge $(-q)$ immersed in a static electric field $\vec{E}$ and a static magnetic field $\vec{B}$. It satisfies to the equation of motion:

$$m\frac{\mathrm{d}^2\vec{x}}{\mathrm{d}t^2} = (-q)\vec{E} + (-q)\frac{\mathrm{d}\vec{x}}{\mathrm{d}t} \times \vec{B} \tag{11.3.25}$$

We have studied in the section of Electrodynamics that the fields $\vec{E}$ and $\vec{B}$ could be constructed from a quadripotential $\vec{A}$. So we can rewrite the above equation from the two relation proven in the section of Electrodynamics:

$$\vec{B} = \vec{\nabla} \times \vec{A} \quad \text{and} \quad \vec{E} = -\vec{\nabla}\phi - \frac{\partial \vec{A}}{\partial t} \tag{11.3.26}$$

However, it is always possible to impose the following gauge (we leave it to do reader to do the check by using exactly the same methodology as that used in the section of Electrodynamics):

$$\vec{\nabla}\phi = \vec{0} \tag{11.3.27}$$

The equation of motion becomes:

$$m\frac{\mathrm{d}^2\vec{x}}{\mathrm{d}t^2} = -(-q)\frac{\partial \vec{A}}{\partial t} + (-q)\frac{\mathrm{d}\vec{x}}{\mathrm{d}t} \times (\vec{\nabla} \times \vec{A}) \tag{11.3.28}$$

or also:

$$m\frac{\mathrm{d}^2\vec{x}}{\mathrm{d}(-t)^2} = -(-q)\frac{\partial \vec{A}}{\partial(-t)} + (-q)\frac{\mathrm{d}\vec{x}}{\mathrm{d}(-t)} \times (\vec{\nabla} \times \vec{A}) \tag{11.3.29}$$

Comparing the last two equations, we come to our second conclusion: a particle of charge q moving backwards in time obeys to the same equations of motion that an oppositely charged particle $-q$ moving forward in the time. The physical interpretation of the second particle is obvious ("Feynman-Stückelberg interpretation").

Relativistic quantum physics implies the existence of anti particles, which are actually observed in laboratories (accelerators and not only...!).

All this to get where exactly?

1. First, the theoretical discovery of antimatter provides a possible explanation for the existence of the Universe which previously violated the principle of conservation of energy. The theory we have just seen, therefore predicted that our Universe should contain as much matter as antimatter. Scientists are still looking for the presence of this antimatter.
2. Secondly, if we consider in vacuum a photon of energy $h\nu > 2m_ec^2$, it is capable of carrying a virtual electron to a state of positive energy, where it becomes real. It appears then locally a gap, or "hole" in the region of negative energies. According to the principle of charge conservation, one sees a positive electron, or positron, symmetrical and antimatter equivalent particle of the electron.

Thus, the photon materializes in the form of a pair e^+, e^-, with:

$$h\nu = 2m_ec^2 + E_c \tag{11.3.30}$$

Remark

Some experimental results suggest that antiparticles are not the perfect mirrors of the particles we know. Indeed, the right/left and time symmetries does not appear to be respected (there is symmetry breaking). We do not have anything written about in this textbook at this date but we will do so as soon as we can.

11.3.2 Generalized Klein-Gordon Equation

The free Klein-Gordon equation that we initially presented and proved earlier does not take into account the influence of magnetic field on the observation of the splitting of spectral lines of atoms (experimental observation!). That's why Klein and Gordon have integrated it into their equation the magnetic field. However, they did not take into account the spin of the electron. Only after their work Pauli developed his equation (named "Pauli equation") which then led to the Dirac equation (see further below).

To determine the expression of the Klein-Gordon generalized equation of a charged particle in a magnetic field and an electrostatic potential, we will use the power of the Lagrangian formalism!

The classical equation of motion accepted (see section Analytical Mechanics page 3169), as valid also in Special Relativity, is given as we proved it by (Euler-Lagrange equation):

$$\mathrm{d}_t(\partial_{\dot{q}_i}L) - \partial_{q_i}L \tag{11.3.31}$$

In the section of Special Relativity, we have proved that the Lagrangian of a free particle is expressed as:

$$L_0 = \frac{m_0c^2}{\gamma} \tag{11.3.32}$$

with for recall:

$$\gamma = (1-\beta^2)^{-1/2} \qquad \beta = \frac{v}{c} \tag{11.3.33}$$

and in the section of Electrodynamics that the total Lagrangian was given by:

$$L=\frac{m_0c^2}{\gamma}-qU+q\vec{v}\circ\vec{A} \tag{11.3.34}$$

For future needs, start by calculating:

$$\partial_{\dot{x}}L=\partial_{\dot{x}}\left(\frac{m_0c^2}{\gamma}-qU+q\vec{v}\circ\vec{A}\right)_x \tag{11.3.35}$$

Let us calculate the first term:

$$\partial_{\dot{x}}\frac{1}{\gamma_x}=\partial_{\dot{x}}\left(1-\frac{v_x^2}{c^2}\right)^{1/2}=\frac{1}{2}\left(1-\frac{v_x^2}{c^2}\right)^{-1/2}x-2\frac{v_x}{c^2}=-\gamma\frac{v_x}{c^2} \tag{11.3.36}$$

As the potential does not dependent on the speed, the term:

$$\partial_{\dot{x}}(-qU) \tag{11.3.37}$$

is equal zero.

The vector potential does not depend on the velocity of the particle also, then:

$$\partial_{\dot{x}}qv_xA_x=qA_x \tag{11.3.38}$$

Therefore it comes:

$$\partial_{\dot{x}}L=-m_0c^2\left(-\gamma\frac{v_x}{c^2}\right)+qA_x=\gamma m_0v_x+qA_x=p_x+qAx \tag{11.3.39}$$

The classical Hamiltonian is written (see section Analytical Mechanics page 3181):

$$H=\sum_i p_ix_i-L=\vec{v}\circ\partial_{\vec{v}}L-L \tag{11.3.40}$$

So we have previously proved that:

$$\partial_{\dot{x}}L=p_x+qA_x \tag{11.3.41}$$

We can write the Hamiltonian with this relation in the form:

$$\begin{aligned}H&=\vec{v}\circ(\vec{p}+q\vec{A})+\frac{m_0c^2}{\gamma}+qU-q\vec{v}\circ\vec{A}=\vec{v}\circ\vec{p}+q\vec{v}\circ\vec{A}+\frac{m_0c^2}{\gamma}+qU-q\vec{v}\circ\vec{A}\\&=\vec{v}\circ\vec{p}+\frac{m_0c^2}{\gamma}+qU\end{aligned} \tag{11.3.42}$$

The dot product $\vec{v}\circ\vec{p}$ has for expressed (since $\vec{p},\vec{v}$ are collinear)

$$\vec{v}\circ\vec{p}=v\gamma m_0v=\gamma m_0v^2 \tag{11.3.43}$$

The Hamiltonian is therefore written:

$$H=\gamma m_0v^2+\frac{m_0c^2}{\gamma}+qU \tag{11.3.44}$$

By working on the two first terms:

$$\gamma m_0 v^2 + \frac{m_0 c^2}{\gamma} = \gamma m_0 c^2 \left(\frac{v^2}{c^2} + \frac{1}{\gamma^2}\right) \tag{11.3.45}$$

But:

$$\gamma = \left(1 - \frac{v^2}{c^2}\right)^{-1/2} \Rightarrow \frac{1}{\gamma^2} = 1 - \frac{v^2}{c^2} \tag{11.3.46}$$

Therefore:

$$\gamma m_0 v^2 + \frac{m_0 c^2}{\gamma} = \gamma m_0 c^2 \left(\frac{v^2}{c^2} + 1 - \frac{v^2}{c^2}\right) = \gamma m_0 c^2 \tag{11.3.47}$$

Finally, we get (for a conservative system):

$$H = \gamma m_0 c^2 + qU = E_{\text{tot}} \tag{11.3.48}$$

Always in the case of a particle moving in an electromagnetic field, the relation between the energy and impulsion (which is different from that of momentum by the presence of a term containing the vector potential) is calculated as follows:

As (we have just prove it):

$$E_{\text{tot}} = \gamma m_0 c^2 = H - qU \tag{11.3.49}$$

and as we have prove it in the section of Wave Quantum Physics:

$$p^2 = (m_0 v)^2 2 = m_0^2 (v)^2 = m_0^2 \left(\frac{\vec{p} - q\vec{A}}{m_0}\right)^2 = (\vec{p} - q\vec{A})^2 \tag{11.3.50}$$

where it is necessary to see the last term as an abusive notation for the square of the norm.

Then by substituting into:

$$E^2 = p^2 c^2 + m_0^2 c^4 \tag{11.3.51}$$

we get (we change of notation for the Hamiltonian):

$$(E_{\text{tot}} - qU)^2 - (\vec{p} - q\vec{A})^2 c^2 = m_0^2 c^4 \tag{11.3.52}$$

If we rewrite this relation by making use of the corresponding operators (see section Wave Quantum Physics page 4695) of energy and momentum (canonical quantization):

$$E = \mathrm{i}\hbar\partial_t \quad \text{and} \quad \vec{p} = -\mathrm{i}\hbar\vec{\nabla} \tag{11.3.53}$$

then finally we can write in analogy with the free Klein-Gordon equation (in the absence of field) the "generalized Klein-Gordon equation":

$$\boxed{\left[(\mathrm{i}\hbar\partial_t - qU)^2 - (-\mathrm{i}\hbar\vec{\nabla} - q\vec{A})^2 c^2\right] \Psi = m_0^2 c^4 \Psi} \tag{11.3.54}$$

This equation is that of Klein-Gordon that applies to a particle of charge q without spin moving in an electromagnetic field.

If $U=0$, $\vec{A}=\vec{0}$ the previous relation simplifies in multiple stages:

$$\left[(\mathrm{i}\hbar\partial_t)^2-(-\mathrm{i}\hbar\vec{\nabla})^2c^2\right]\Psi=m_0^2c^4\Psi \tag{11.3.55}$$

$$(-\hbar^2\partial_t^2+\hbar^2\Delta c^2)\Psi=m_0c^4\Psi \tag{11.3.56}$$

$$-\hbar^2\left(\frac{1}{c^2}\partial_t^2+\Delta\right)\Psi=m_0^2c^2\Psi \tag{11.3.57}$$

$$\hbar^2\left(\frac{1}{c^2}\partial_t^2+\Delta\right)\Psi+m_0^2c^2\Psi=0 \tag{11.3.58}$$

$$(\hbar^2\Box+m_0c^2)\Psi=0 \tag{11.3.59}$$

We thus fall back on the Klein-Gordon equation for a free particle without spin!

11.3.2.1 Generalized Klein-Gordon Equation continuity equation

It would be interesting now to look at the expression of the continuity equation (let us recall that it expresses the conservation of energy!!) with the inclusion of the magnetic field (because in fact the continuity equation still makes problem... and even a very big problem!!!). For this, consider the case of a free particle moving with a momentum $\vec{p}$ and having an energy E. We saw that we could associate to it a plane wave of the form:

$$\psi=Ae^{-\mathrm{i}(Et-\vec{p}\circ\vec{r})} \tag{11.3.60}$$

Given the free Klein-Gordon equation and its complex conjugate (we work with natural units):

$$-\partial_t^2\Psi+\Delta\Psi-m_0^2\Psi=0 \tag{11.3.61a}$$

$$-\partial_t^2\bar{\Psi}+\Delta\bar{\Psi}-m_0^2\bar{\Psi}=0 \tag{11.3.61b}$$

We multiply (a) by $-\mathrm{i}\bar{\Psi}$ and (b) by $-\mathrm{i}\Psi$:

$$-\mathrm{i}\bar{\Psi}\left(-\partial_t^2\Psi\right)-\mathrm{i}\bar{\Psi}\Delta\Psi+\mathrm{i}\bar{\Psi}m_0^2\Psi=0 \tag{11.3.62a}$$

$$-\mathrm{i}\Psi\left(-\partial_t^2\bar{\Psi}\right)-\mathrm{i}\Psi\Delta\bar{\Psi}+\mathrm{i}\Psi m_0^2\bar{\Psi}=0 \tag{11.3.62b}$$

Therefore:

$$\mathrm{i}\bar{\Psi}\partial_t^2\Psi-\mathrm{i}\bar{\Psi}\Delta\Psi+\mathrm{i}\bar{\Psi}m_0^2\Psi=0 \tag{11.3.63a}$$

$$\mathrm{i}\Psi\partial_t^2\bar{\Psi}-\mathrm{i}\Psi\Delta\bar{\Psi}+\mathrm{i}\Psi m_0^2\bar{\Psi}=0 \tag{11.3.63b}$$

By difference (a)-(b):

$$\mathrm{i}\boxed{\left(\bar{\Psi}\partial_t^2\Psi-\Psi\partial_t^2\bar{\Psi}\right)}-\mathrm{i}\left(\bar{\Psi}\Delta\Psi-\Psi\Delta\bar{\Psi}\right)=0 \tag{11.3.64}$$

Computing the derivatives with respect to t of the following functions:

$$\partial_t(\bar{\Psi}\partial_t\Psi)=\partial_t\bar{\Psi}\partial_t\Psi+\bar{\Psi}\partial_t^2\Psi \tag{11.3.65a}$$

$$\partial_t(\Psi\partial_t\bar{\Psi})=\partial_t\Psi\partial_t\bar{\Psi}+\Psi\partial_t^2\bar{\Psi} \tag{11.3.65b}$$

By difference (a)-(b):

$$\partial_t(\bar{\Psi}\partial_t\Psi-\Psi\partial_t\bar{\Psi})=\boxed{\bar{\Psi}\partial_t^2\Psi-\Psi\partial_t^2\bar{\Psi}} \tag{11.3.66}$$

Which gives us finally:

$$\partial_t\left[\mathrm{i}(\bar{\Psi}\partial_t\Psi-\Psi\partial_t\bar{\Psi})\right]-\mathrm{i}\boxed{\bar{\Psi}\Delta\Psi-\Psi\Delta\bar{\Psi}}=0 \tag{11.3.67}$$

Let f be a scalar field and $\vec{g}$ and a vector field. The vector analysis gives:

$$\vec{\nabla}\circ(f\cdot\vec{g})=(\vec{\nabla}f)\circ\vec{g}+f(\vec{\nabla}\circ\vec{g})\Rightarrow f(\vec{\nabla}\circ\vec{g})=\vec{\nabla}\circ(f\cdot\vec{g})-(\vec{\nabla}f)\circ\vec{g} \tag{11.3.68}$$

Let us put:

$$f=\Psi \qquad \vec{g}=\vec{\nabla}\cdot\bar{\Psi} \tag{11.3.69}$$

Therefore:

$$\vec{\nabla}\circ(f\cdot\vec{g})=\vec{\nabla}\circ(f\cdot\vec{g})-(\vec{\nabla}f)\circ\vec{g}\Rightarrow\Psi(\vec{\nabla}\circ\vec{\nabla}\bar{\Psi})=\vec{\nabla}\circ(\Psi\vec{\nabla}\bar{\Psi})-(\vec{\nabla}\Psi)\circ\vec{\nabla}\bar{\Psi} \tag{11.3.70}$$

Let us put now:

$$f=\bar{\Psi},\,g=\vec{\nabla}\Psi \tag{11.3.71}$$

Therefore:

$$f(\vec{\nabla}\circ\vec{g})=\vec{\nabla}\circ(f\cdot\vec{g})-(\vec{\nabla})\circ\vec{g}\Rightarrow\bar{\Psi}(\vec{\nabla}\circ\vec{\nabla}\Psi)=\vec{\nabla}\circ(\bar{\Psi}\vec{\nabla}\Psi)-(\vec{\nabla}\bar{\Psi})\circ\vec{\nabla}\Psi \tag{11.3.72}$$

So we have finally two relations:

$$\vec{\nabla}\circ(f\cdot\vec{g})=\vec{\nabla}\circ(f\cdot\vec{g})-(\vec{\nabla}f)\circ\vec{g}\Rightarrow\Psi(\vec{\nabla}\circ\vec{\nabla}\bar{\Psi})=\vec{\nabla}\circ(\Psi\vec{\nabla}\bar{\Psi})-(\vec{\nabla}\Psi)\circ\vec{\nabla}\bar{\Psi} \tag{11.3.73a}$$

$$f(\vec{\nabla}\circ\vec{g})=\vec{\nabla}\circ(f\cdot\vec{g})-(\vec{\nabla})\circ\vec{g}\Rightarrow\bar{\Psi}(\vec{\nabla}\circ\vec{\nabla}\Psi)=\vec{\nabla}\circ(\bar{\Psi}\vec{\nabla}\Psi)-(\vec{\nabla}\bar{\Psi})\circ\vec{\nabla}\Psi \tag{11.3.73b}$$

We subtract (a)-(b) to get:

$$\Psi(\vec{\nabla}\circ\vec{\nabla}\bar{\Psi})-\bar{\Psi}(\vec{\nabla}\circ\vec{\nabla}\Psi)=\vec{\nabla}\circ(\Psi\vec{\nabla}\bar{\Psi})-\vec{\nabla}\circ(\bar{\Psi}\vec{\nabla}\Psi) \tag{11.3.74}$$

As $\vec{\nabla}\circ\vec{\nabla}=\vec{\nabla}^2=\Delta$:

$$\Psi\Delta\bar{\Psi}-\bar{\Psi}\Delta\Psi=\vec{\nabla}\circ(\Psi\vec{\nabla}\bar{\Psi}-\bar{\Psi}\vec{\nabla}\Psi) \tag{11.3.75}$$

By changing the signs:

$$\bar{\Psi}\Delta\Psi-\Psi\Delta\bar{\Psi}=\vec{\nabla}\circ(\bar{\Psi}\vec{\nabla}\Psi-\Psi\vec{\nabla}\bar{\Psi}) \tag{11.3.76}$$

This latter relation and:

$$\partial_t\left[\mathrm{i}(\bar{\Psi}\partial_t\Psi-\Psi\partial_t\bar{\Psi})\right]-\mathrm{i}(\bar{\Psi}\Delta\Psi-\Psi\Delta\bar{\Psi})=0 \tag{11.3.77}$$

gives together:

$$\boxed{\partial_t\left[\mathrm{i}(\bar{\Psi}\partial_t\Psi-\Psi\partial_t\bar{\Psi})\right]-\vec{\nabla}\circ\left[\mathrm{i}(\bar{\Psi}\vec{\nabla}\Psi-\Psi\vec{\nabla}\bar{\Psi})\right]=0} \tag{11.3.78}$$

Again, let us compare this relation with the continuity equation:

$$\partial_t\rho+\vec{\nabla}\circ\vec{j}=0 \tag{11.3.79}$$

Let us recall that during our first study of the Klein-Gordon equation we have seen that in quantum physics equivalent is given by the same equation but with the following meanings: ρ is the probability density, $\vec{j}$ is the flux density of particles. We then have:

$$\boxed{\begin{aligned}&\rho=\mathrm{i}(\bar{\Psi}\partial_t\Psi-\Psi\partial_t\bar{\Psi})\\&\vec{j}=-\mathrm{i}(\bar{\Psi}\vec{\nabla}\Psi-\Psi\vec{\nabla}\bar{\Psi})\end{aligned}} \tag{11.3.80}$$

If the associated wave function Ψ and its conjugate complex $\bar{\Psi}$ are given in a special case by:

$$\Psi=A^{-\mathrm{i}(Et-\vec{p}\circ\vec{r})}\qquad\bar{\Psi}=A^{\mathrm{i}(Et-\vec{p}\circ\vec{r})} \tag{11.3.81}$$

The derivatives with respect to time of these functions are obviously:

$$\partial_t\Psi=-\mathrm{i}EA^{-\mathrm{i}(Et-\vec{p}\circ\vec{r})}=-\mathrm{i}E\Psi\qquad\partial_t\bar{\Psi}=\mathrm{i}EA^{\mathrm{i}(Et-\vec{p}\circ\vec{r})}=-\mathrm{i}E\bar{\Psi} \tag{11.3.82}$$

And their gradients are calculated as follows:

$$\begin{aligned}\vec{\nabla}\Psi&=\partial_{\vec{r}}A^{-\mathrm{i}(Et-\vec{p}\circ\vec{r})}=\mathrm{i}\vec{p}A^{-\mathrm{i}(Et-\vec{p}\circ\vec{r})}=\mathrm{i}\vec{p}\Psi\\\vec{\nabla}\bar{\Psi}&=\partial_{\vec{r}}A^{\mathrm{i}(Et-\vec{p}\circ\vec{r})}=-\mathrm{i}\vec{p}A^{-\mathrm{i}(Et-\vec{p}\circ\vec{r})}=-\mathrm{i}\vec{p}\bar{\Psi}\end{aligned} \tag{11.3.83}$$

By taking again the expression of the probability density and given previous differential, we get:

$$\begin{aligned}\rho&=\mathrm{i}\left(\bar{\Psi}\partial_t\Psi-\Psi\partial_t\bar{\Psi}\right)\\&=\mathrm{i}\left[A\cancel{e^{\mathrm{i}(Et-\vec{p}\circ\vec{r})}}\cdot\left(-\mathrm{i}EA\cancel{e^{-\mathrm{i}(Et-\vec{p}\circ\vec{r})}}\right)\right]-\left[A\cancel{e^{-\mathrm{i}(Et-\vec{p}\circ\vec{r})}}\cdot\left(\mathrm{i}EA\cancel{e^{\mathrm{i}(Et-\vec{p}\circ\vec{r})}}\right)\right]\\&=\mathrm{i}\left(-\mathrm{i}EA^2-\mathrm{i}EA^2\right)=\mathrm{i}\left(-2\mathrm{i}EA^2\right)\end{aligned} \tag{11.3.84}$$

The density of probability has therefore for expression:

$$\boxed{\rho=2EA^2} \tag{11.3.85}$$

Taking again the expression of probabilistic current density and considering the differential, we get:

$$\begin{aligned}\vec{J}&=-\mathrm{i}\left(\bar{\Psi}\vec{\nabla}\Psi-\Psi\vec{\nabla}\bar{\Psi}\right)\\&=-\mathrm{i}\left(\left[A\cancel{e^{\mathrm{i}(Et-\vec{p}\circ\vec{r})}}\cdot\left(\mathrm{i}\vec{p}A\cancel{e^{-\mathrm{i}(Et-\vec{p}\circ\vec{r})}}\right)\right]-\left[A\cancel{e^{-\mathrm{i}(Et-\vec{p}\circ\vec{r})}}\cdot\left(-\mathrm{i}\vec{p}A\cancel{e^{\mathrm{i}(Et-\vec{p}\circ\vec{r})}}\right)\right]\right)\\&=-\mathrm{i}\left[\mathrm{i}\vec{p}A^2-\left(-\mathrm{i}\vec{p}A^2\right)\right]=-\mathrm{i}\left(2\mathrm{i}\vec{p}A^2\right)\end{aligned} \tag{11.3.86}$$

The probabilistic current density has therefore for expression:

$$\boxed{\vec{j}=2\vec{p}A^2} \tag{11.3.87}$$

Putting ourselves in the situation of that time of knowledge of quantum physics, the Klein-Gordon equation has several pathologies and disadvantages:

- The probability density $\rho = 2EA^2$ may become negative (since as we have seen, the energy can also be negative), which is inexplicable. Such a situation does not exist with the Schrödinger equation.

- The Klein-Gordon equation has the disadvantage of being of the second order in t (the Schrödinger equation is of the first order). The temporal evolution requires therefore the knowledge of not only of $\Psi(t_0)$ but also of its derivative $(\partial_t \Psi)_{t=t}$

- If we applied this equation to the hydrogen atom, we would not find the same fine structure[25] energy levels as observed.

All this led at the time preceding the works of Dirac, to reject this equation that, in addition, did not include the spin.

---

[25] "Fine-structure" refers to splitting of levels with the same l but different j

11.3.3 Classical free Dirac equation

So far, any particle was regarded as punctual and without any structure or internal degree freedom. In this context, all the information on the system state at time t is then reputed as entirely contained in the wave function $\Psi(x,y,z,t)$.

However, such a description is not sufficient, as we shall see. This insufficiency comes from the experimental evidence demonstrating that a particle such as an electron or also neutron has a magnetic intrinsic moment, independently of any rotation in space around a center. The existence of this magnetic moment in turn leads to the existence of an intrinsic angular momentum which was named "spin" because it was thought at first that this degree of freedom was linked to a rotation of the particle on itself. This degree of freedom is "internal" - although the electron continues to be regarded as a point particle; that is, along with the electric charge or mass, an intrinsic attribute, given once and for all. It seems so far impossible to physicist to give to the spin a classic analogy! Imagine the electron as a small non-zero radius ball that turns on itself leads to absurdities (for example we found that a point located on the surface of the electron has a much higher speed than c). It remains that the spin of a massic particle is its angular momentum in the repository where it is at rest. It is commonly accepted that hypothesis of the electron spin was formulated first by George Uhlenbeck and Samuel Goudsmit in 11925 (holocene calendar).

The spin of a particle is always half full or unitary, it is a fact of experience until now. The half or unitary characteristic nature of the spin defines two broad families of particles:

- bosons (integer spin)
- fermions (half-integer spin)

obeying very different statistics such as those we have discussed and study in the section of Statistical Mechanics (hence the existence of a relation named "spin-statistics theorem").

Let us return to the case of the electron. The two possible values revealed by measuring S (the μ_l we had in the section of Corpuscular Quantum Physics page 4666) are the eigenvalues $\pm\hbar/2$ (see section Wave Quantum Physics page 4877) associated with the two possible values of a quantum number $m_s=\pm 1/2$ itself associated at the free state ($\vec{L}=\vec{0}$) to the angular momentum:

$$b\leq\sqrt{J(J+1)}\hbar=\sqrt{(L+S)((L+s)+1)}\hbar=\sqrt{S(S+1)}\hbar \tag{11.3.88}$$

Therefore:

$$b\leq\sqrt{S(S+1)}\hbar=\sqrt{\pm\frac{1}{2}\left(\pm\frac{1}{2}+1\right)}\hbar=\sqrt{\frac{3}{4}}\hbar=\frac{\sqrt{3}\hbar}{2} \tag{11.3.89}$$

A complete description of the state of the electron therefore necessarily contains a wave function as usual giving the probability density of presence, but also taking into account the degree of freedom of the spin, hence the notation $\Psi(x,y,z,m_s,t)$. If the coordinates of space are continuous real values, however the spin variable is essentially discrete!

Maintaining the usual interpretation, the quantity $|\Psi(x,y,z,m_s,t)|^2\mathrm{d}r^2$ is the probability of presence around the point selected with the value $m_s\hbar$ for the spin. The condition of normalization of the probabilities introduced as always a sum, which covers not only the orbital degrees (continuous summation, that is to say: integration) but also the degrees of spin (discrete summation):

$$\int\left|\Psi\left(\vec{r},+\frac{1}{2}\right)\right|^2\mathrm{d}r^3+\int\left|\Psi\left(\vec{r},-\frac{1}{2}\right)\right|^2\mathrm{d}r^3=1 \qquad (11.3.90)$$

expressing in particular the fact that we exhaust all the possibilities of the spin by summing on the two possible values. In any case, the electron no longer has one but two wave functions, one for each value of m_s.

The previous notation is not necessarily the best for spin-free particles greater than $1/2$ as we have seen in our study of angular momentum. About the fact of having a variable taking discrete values, it is just as legitimate to put m_s in index of the Ψ and to put: $\Phi_{m_S}(\vec{r},t)$. Finally, it is convenient to use a matrix notation, storing in columns the various functions corresponding to the possible values of the discrete variable m_s. Thus, for the electron, we will now assume that all information in the sense of Wave Quantum Physics is contained in a two-line column vector named "spinor" (see section Spinor Calculus page 2145) and denoted:

$$\begin{bmatrix}\Psi_{+\frac{1}{2}}(\vec{r},t)\\ \Psi_{-\frac{1}{2}}(\vec{r},t)\end{bmatrix} \quad \text{or} \quad \begin{bmatrix}\Psi_{+}(\vec{r},t)\\ \Psi_{-}(\vec{r},t)\end{bmatrix} \qquad (11.3.91)$$

Let us now come back to the free Klein-Gordon equation (more general than the Schrödinger equation, of course, but less than the one including the magnetic field):

$$\hbar^2\frac{\partial^2\Psi}{\partial t^2}=\hbar^2c^2\Delta\Psi-m_0^2c^4\Psi \qquad (11.3.92)$$

This equation is as we know unfortunately incomplete because it contains no information on the spin of the electron (if we focus still only on this particle).

We can, however, in order to find a solution to this problem, make a parallel with the electromagnetic field. That latter includes also a spin, residing in the polarization of the field (see section Electrodynamics page 4538). This polarization is closely related to the vector nature of the electromagnetic field and is reflected in the Maxwell equations, which are as we know first-order derivatives. However, by combining Maxwell's equations, we saw in the Electrodynamics section that we could obtain the wave equations:

$$\nabla^2\vec{E}-\frac{1}{c^2}\frac{\partial^2\vec{E}}{\partial t^2}=\Delta\vec{E}-\frac{1}{c^2}\frac{\partial^2\vec{E}}{\partial t^2}=\vec{0} \quad \text{and} \quad \nabla^2\vec{B}-\frac{1}{c^2}\frac{\partial^2\vec{B}}{\partial t^2}=\Delta\vec{B}-\frac{1}{c^2}\frac{\partial^2\vec{B}}{\partial t^2}=\vec{0} \qquad (11.3.93)$$

which are (very relevant coincidence!) a special case of the Klein-Gordon equation when $m_0=0$:

$$\Delta\Psi-\frac{1}{c^2}\frac{\partial^2\Psi}{\partial t^2}=0 \qquad (11.3.94)$$

However, the wave equations contain less information than the original Maxwell equations: they do not explicitly contain any relation between the various components of the fields $\vec{E}$ and $\vec{B}$, such as the fact that in an electromagnetic wave of wave vector $\vec{k}$ the fields $\vec{E}$ and $\vec{B}$ are mutually perpendicular and both are perpendicular to the wave vector $\vec{k}$ (i.e. they are bot perpendicular to the direction of propagation of the wave). To establish these constraints, we must return to Maxwell's equations and thus to equations with first-order derivatives.

The same is true for fermions (electrons are part of it). Klein-Gordon's equation, even it is not false, is incomplete. We must try here to establish a first-order equation in derivatives which describes well the spin $1/2$ of the electrons of the fermions. This last condition means that this equation must therefore involve the two components of a spinor (in analogy with the one we have determined above):

$$\chi = \begin{bmatrix}\chi_1\\ \chi_2\end{bmatrix} \tag{11.3.95}$$

We will write this equation which we seek as:

$$D\chi = 0 \tag{11.3.96}$$

where D is an 2×2 matrix involving first order derivatives (a first-order differential operator) that we have to determine (and that we will!!!).

To give an example before going further, let us look at how Klein-Gordon's equation can be expressed in such a form.

We have first (free Klein-Gordon equation):

$$\hbar^2\frac{\partial^2\Psi}{\partial t^2} = \hbar^2c^2\Delta\Psi - m_0^2c^4\Psi \tag{11.3.97}$$

or (generalized Klein-Gordon equation):

$$\left[(\mathrm{i}\hbar\partial_t - qU)^2 - \left(-\mathrm{i}\hbar\vec{\nabla} - q\vec{A}\right)^2 c^2\right]\Psi = m_0^2c^4\Psi \tag{11.3.98}$$

What can also be written for the free Klein-Gordon equation:

$$\left(1 + \frac{\mathrm{i}\hbar}{m_0c^2}\frac{\partial}{\partial t}\right)\left(1 - \frac{\mathrm{i}\hbar}{m_0c^2}\frac{\partial}{\partial t}\right)\Psi = \frac{\hbar^2}{m_0^2c^2}\Delta\Psi \tag{11.3.99}$$

or for the generalized Klein-Gordon equation:

$$\begin{gathered}
\left[-\hbar^2\partial_t^2 - 2\mathrm{i}\hbar\partial_t qU + q^2U^2 + \hbar^2\vec{\nabla}^2c^2 - 2\mathrm{i}\hbar\vec{\nabla}q\vec{A}c^2 - q^2\vec{A}^2c^2\right]\Psi = m_0^2c^4\Psi\\
\left[1 + \frac{\hbar^2\partial_t^2}{m_0^2c^4}\right]\Psi = \frac{\left(-q^2\vec{A}^2c^2 - 2\mathrm{i}\hbar\vec{\nabla}q\vec{A}c^2 - \hbar^2\Delta c^2 + q^2U^2 - 2\mathrm{i}\hbar\partial_t qU\right)}{m_0^2c^4}\Psi\\
\left(1 + \frac{\mathrm{i}\hbar}{m_0c^2}\frac{\partial}{\partial t}\right)\left(1 - \frac{\mathrm{i}\hbar}{m_0c^2}\frac{\partial}{\partial t}\right)\Psi = \frac{\left(-q^2\vec{A}^2c^2 - 2\mathrm{i}\hbar\vec{\nabla}q\vec{A}c^2 - \hbar^2\Delta c^2 + q^2U^2 - 2\mathrm{i}\hbar\partial_t qU\right)}{m_0^2c^4}\Psi
\end{gathered} \tag{11.3.100}$$

Let us now restrict ourselves to the case of the free Klein-Gordon equation (the reasoning being similar but ... longer for the generalized version).

The last expression of the free Klein-Gordon equation suggests introducing the two combinations found after (it seems...) numerous trial and errors by our predecessors:

$$\begin{aligned}\Psi_a&=\Psi+\frac{\mathrm{i}\hbar}{m_0c^2}\frac{\partial\Psi}{\partial t}=\left(1+\frac{\mathrm{i}\hbar}{m_0c^2}\frac{\partial}{\partial t}\right)\Psi\\ \Psi_b&=\Psi-\frac{\mathrm{i}\hbar}{m_0c^2}\frac{\partial\Psi}{\partial t}=\left(1-\frac{\mathrm{i}\hbar}{m_0c^2}\frac{\partial}{\partial t}\right)\Psi\end{aligned}\qquad(11.3.101)$$

from which results:

$$\begin{aligned}\Psi&=\frac{1}{2}\left(\Psi_a+\Psi_b\right)\\ \frac{\partial\Psi}{\partial t}&=\frac{m_0c^2}{2\mathrm{i}\hbar}\left(\Psi_a-\Psi_b\right)\end{aligned}\qquad(11.3.102)$$

Therefore:

$$\left(1+\frac{\mathrm{i}\hbar}{m_0c^2}\frac{\partial}{\partial t}\right)\left(1-\frac{\mathrm{i}\hbar}{m_0c^2}\frac{\partial}{\partial t}\right)\Psi=\frac{\hbar^2}{m_0^2c^2}\Delta\Psi\qquad(11.3.103)$$

Can be written in two ways after respective substitutions:

$$\begin{aligned}\left(1-\frac{\mathrm{i}\hbar}{m_0c^2}\frac{\partial}{\partial t}\right)\Psi_a&=\frac{\hbar^2}{2m_0^2c^2}\Delta\left(\Psi_a+\Psi_b\right)\\ \left(1+\frac{\mathrm{i}\hbar}{m_0c^2}\frac{\partial}{\partial t}\right)\Psi_b&=\frac{\hbar^2}{2m_0^2c^2}\Delta\left(\Psi_a+\Psi_b\right)\end{aligned}\qquad(11.3.104)$$

Either in matrix form:

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\begin{bmatrix}\Psi_a\\ \Psi_b\end{bmatrix}=\begin{bmatrix}m_0c^2-\dfrac{\hbar^2}{2m_0}\Delta & -\dfrac{\hbar^2}{2m_0}\Delta\\ \dfrac{\hbar^2}{2m_0}\Delta & -m_0c^2+\dfrac{\hbar^2}{2m_0}\Delta\end{bmatrix}\begin{bmatrix}\Psi_a\\ \Psi_b\end{bmatrix}\qquad(11.3.105)$$

or even:

$$\left(\mathrm{i}\hbar\frac{\partial}{\partial t}\begin{bmatrix}1&0\\0&1\end{bmatrix}-\begin{bmatrix}m_0c^2-\dfrac{\hbar^2}{2m_0}\Delta & -\dfrac{\hbar^2}{2m_0}\Delta\\ \dfrac{\hbar^2}{2m_0}\Delta & -m_0c^2+\dfrac{\hbar^2}{2m_0}\Delta\end{bmatrix}\right)\begin{bmatrix}\Psi_a\\ \Psi_b\end{bmatrix}=\vec{0}\qquad(11.3.106)$$

What we can write:

$$\begin{bmatrix}\mathrm{i}\hbar\dfrac{\partial}{\partial t}-m_0c^2+\dfrac{\hbar^2}{2m_0}\Delta & \dfrac{\hbar^2}{2m_0}\Delta\\ \dfrac{\hbar^2}{2m_0}\Delta & -\mathrm{i}\hbar\dfrac{\partial}{\partial t}-m_0c^2+\dfrac{\hbar^2}{2m_0}\Delta\end{bmatrix}\begin{bmatrix}\Psi_a\\ \Psi_b\end{bmatrix}=\vec{0}\qquad(11.3.107)$$

Therefore, relatively to our initial idea of having a relation in the form:

$$D\chi=0\qquad(11.3.108)$$

We can make the similarity (correspondence term by term) with the prior-previous equation:

$$\chi=\begin{bmatrix}\Psi_a\\ \Psi_b\end{bmatrix}\quad\text{and}\quad D=\begin{bmatrix}\mathrm{i}\hbar\dfrac{\partial}{\partial t}-m_0c^2+\dfrac{\hbar^2}{2m_0}\Delta & \dfrac{\hbar^2}{2m_0}\Delta\\ \dfrac{\hbar^2}{2m_0}\Delta & -\mathrm{i}\hbar\dfrac{\partial}{\partial t}-m_0c^2+\dfrac{\hbar^2}{2m_0}\Delta\end{bmatrix}\qquad(11.3.109)$$

where D is indeed a 2×2 matrix.

$$D=A+\vec{B}\circ\vec{\sigma} \tag{11.3.110}$$

where A is a vector (but represented by tradition as a scalar...), $\vec{B}$ is a vector (and denoted as such and has nothing to do with the notation of the magnetic field) and σ a vector composed of 2×2 symmetric matrices (by reading the rest you will see that posing this makes it possible to find what we are looking for...).

Let us recall that the multiplication between $\vec{B}$ and $\vec{\sigma}$ constitutes a dot product such as that defined in our study of the Spinor Calculus section.

Remark

We reader must be very cautious in the developments that follow, because the traditional notations in the field make it very difficult to distinguish between products, dot products, and products of vector components forming a vector (element-wise multiplication).

Let us put (by the way our predecessors also made numerous trials and errors before putting this ...):

$$\mathrm{D}=\frac{\partial}{\partial t}+\lambda\vec{\sigma}\circ\vec{\nabla} \tag{11.3.111}$$

What we were initially seeking, that is, $\mathrm{D}\xi=0$, then becomes:

$$\frac{\partial}{\partial t}\chi+\lambda\vec{\sigma}\circ\vec{\nabla}\chi=\vec{0} \tag{11.3.112}$$

Thus $A=\partial/\partial t$, $\vec{B}=\lambda\vec{\nabla}$ and for $\vec{\sigma}$ that latter remains (let us imagine...) unknown. We must also determine λ.

Always by analogy with the example above, let us try to find the wave equation to determine the constant λ:

$$\left(\frac{\partial}{\partial t}-\lambda\vec{\sigma}\circ\vec{\nabla}\right)\left(\frac{\partial}{\partial t}+\lambda\vec{\sigma}\circ\vec{\nabla}\right)\chi=\vec{0}\Rightarrow\left(\frac{\partial^2}{\partial t^2}-\lambda^2\left(\vec{\sigma}\circ\vec{\nabla}\right)\left(\vec{\sigma}\circ\vec{\nabla}\right)\right)\chi=\vec{0} \tag{11.3.113}$$

So that we fall back on the wave equation, we must therefore have:

1. First that $\left(\vec{\sigma}\circ\vec{\nabla}\right)\left(\vec{\sigma}\circ\vec{\nabla}\right)=\vec{\nabla}^2$. Indeed (caution! the reader must have ideally at least studied the section of Spinor Calculus to know in detail how we get at this development!):

$$\begin{aligned}\left(\vec{\sigma}\circ\vec{\nabla}\right)\left(\vec{\sigma}\circ\vec{\nabla}\right)&=\begin{bmatrix}\partial_z & \partial_x-\mathrm{i}\partial_y\\ \partial_x+\mathrm{i}\partial_y & -\partial_z\end{bmatrix}\begin{bmatrix}\partial_z & \partial_x-\mathrm{i}\partial_y\\ \partial_x+\mathrm{i}\partial_y & -\partial_z\end{bmatrix}\\ &=\begin{bmatrix}\partial_z^2+\left(\partial_x+\mathrm{i}\partial_y\right)\left(\partial_x-\mathrm{i}\partial_y\right) & \partial_z\left(\partial_x-\mathrm{i}\partial_y\right)-\partial_z\left(\partial_x-\mathrm{i}\partial_y\right)\\ \partial_z\left(\partial_x+\mathrm{i}\partial_y\right)-\partial_z\left(\partial_x+\mathrm{i}\partial_y\right) & \left(\partial_x-\mathrm{i}\partial_y\right)\left(\partial_x+\mathrm{i}\partial_y\right)+\partial_z^2\end{bmatrix}\\ &=\begin{bmatrix}\partial_x^2+\partial_y^2+\partial_z^2 & 0\\ 0 & \partial_x^2+\partial_y^2+\partial_z^2\end{bmatrix}=\mathbb{1}\vec{\nabla}^2=\vec{\nabla}^2\end{aligned} \tag{11.3.114}$$

2. And that $\lambda=c$:

$$\left(\frac{\partial}{\partial t}-\lambda\vec{\sigma}\circ\vec{\nabla}\right)\left(\frac{\partial}{\partial t}+\lambda\vec{\sigma}\circ\vec{\nabla}\right)\chi=\left(\frac{\partial^2}{\partial t^2}-c^2\vec{\nabla}^2\right)\chi=\vec{0} \tag{11.3.115}$$

There are therefore, always in analogy with the previous developments, two possibilities which can be applied to different fields that we will be denoted $\chi^{(\pm)}$ and which are named "Weyl spinors". We thus have a sort of double spinor (as in the parenthesis we have a two components vector formed by 2×2 matrices) such that:

$$\left(\frac{\partial}{\partial t}\pm c\vec{\sigma}\circ\vec{\nabla}\right)\chi^{\pm}=\vec{0} \tag{11.3.116}$$

These equations are named "Weyl equations" as the above relation represents in reality 4 equations (don't forget that the parenthesis is a 2 components vector made of 2×2 matrices and therefore 4 rows). Let us recall that a two dimensional spinor is a 2 components vector that are both in $\mathbb{C}$.

We now have to generalize the Weyl equations to the case of a half-integer spin fermion with mass. This new equation must respect the following constraints:

C1. It must be reduced to the Weyl equations when the mass tends to zero

C2. It must lead to the free Klein-Gordon equation

C3. It must describe particles with a spin

The very clever solution (extremely difficult to guess) then consists in coupling the two Weyl equations by using different factor such as:

$$\begin{aligned}\mathrm{i}\hbar\left(\frac{\partial}{\partial t}+c\vec{\sigma}\circ\vec{\nabla}\right)\chi^{(+)}=m_0c^2\chi^{(-)}\\ \mathrm{i}\hbar\left(\frac{\partial}{\partial t}-c\vec{\sigma}\circ\vec{\nabla}\right)\chi^{(-)}=m_0c^2\chi^{(+)}\end{aligned} \tag{11.3.117}$$

To verify that these factors have been correctly chosen, we apply:

$$\hbar\left(\frac{\partial}{\partial t}+c\vec{\sigma}\circ\vec{\nabla}\right) \tag{11.3.118}$$

to the first equation on both sides such that we get:

$$\mathrm{i}\hbar\left(\frac{\partial}{\partial t}-c\vec{\sigma}\circ\vec{\nabla}\right)\hbar\left(\frac{\partial}{\partial t}+c\vec{\sigma}\circ\vec{\nabla}\right)\chi^{(+)}=m_0c^2\hbar\left(\frac{\partial}{\partial t}-c\vec{\sigma}\circ\vec{\nabla}\right)\chi^{(-)} \tag{11.3.119}$$

As the second equation give us:

$$\mathrm{i}\hbar\left(\frac{\partial}{\partial t}-c\vec{\sigma}\circ\vec{\nabla}\right)\chi^{(-)}=m_0c^2\chi^{(+)} \tag{11.3.120}$$

After rearranging in the latter the complex term a put it as numerator on the right hand side by multiplying by i/i :

$$\hbar\left(\frac{\partial}{\partial t}-c\vec{\sigma}\circ\vec{\nabla}\right)\chi^{(-)}=-\mathrm{i}m_0c^2\chi^{(+)} \tag{11.3.121}$$

and substitute to the previous in the right hand side, we thus find:

$$\mathrm{i}\hbar\left(\frac{\partial}{\partial t}-c\vec{\sigma}\circ\vec{\nabla}\right)\hbar\left(\frac{\partial}{\partial t}+c\vec{\sigma}\circ\vec{\nabla}\right)\chi^{(+)}=-\mathrm{i}m_0^2c^4\chi^{(+)} \tag{11.3.122}$$

or developing the left extreme previous term and keeping the right extreme previous term and dividing by i (remember that the square of any of the three Pauli spin matrices is equal to the identity!):

$$\hbar^2\left(\frac{\partial^2}{\partial t^2}-c^2\nabla^2\right)\chi^{(+)}=-m_0^2c^4\chi^{(+)} \tag{11.3.123}$$

to compare with:

$$\hbar^2\left(\frac{\partial^2}{\partial t^2}-c^2\nabla^2\right)\Psi=-m_0^2c^4\Psi \tag{11.3.124}$$

that is the free Klein-Gordon equation (we prove the same correspondence for the component χ^-) and thus reinforces the validity of the assumptions and developments made so far.

It is usual to combine the two spinors in a single spinor (this becomes a "bispinor") of 4 components (a spinor with 4 components, two of which are in fact associated with the particles and two with the antiparticles as we shall see):

$$\Psi=\begin{bmatrix}\chi^{(+)}\\\chi^{(-)}\end{bmatrix}=\begin{bmatrix}\Psi_1\\\Psi_2\\\Psi_3\\\Psi_4\end{bmatrix} \tag{11.3.125}$$

and to define the following two matrices (commonly named "Gamma matrices") in a so-named "standard form":

$$\gamma^0=\begin{bmatrix}0 & \mathbb{1}\\\mathbb{1} & 0\end{bmatrix} \tag{11.3.126}$$

where $\mathbb{1}$ is the 2×2 unit matrix traditionally defined by:

$$\mathbb{1}=\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix} \tag{11.3.127}$$

and:

$$\gamma^i=\begin{bmatrix}0 & -\sigma_i\\\sigma_i & 0\end{bmatrix}=\alpha^i\cdot\beta=\begin{bmatrix}0 & \sigma_i\\\sigma_i & 0\end{bmatrix}\begin{bmatrix}\mathbb{1} & 0\\0 & -\mathbb{1}\end{bmatrix} \tag{11.3.128}$$

where the σ_i are the "Pauli matrices" given by (see section Spinor Calculus page 2158):

$$\sigma_0=\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix},\sigma_1=\begin{bmatrix}0 & 1\\1 & 0\end{bmatrix},\sigma_2=\begin{bmatrix}0 & -\mathrm{i}\\\mathrm{i} & 0\end{bmatrix},\sigma_3=\begin{bmatrix}1 & 0\\0 & -1\end{bmatrix} \tag{11.3.129}$$

which must satisfy for recall (proved earlier above):

$$\left(\vec{\sigma}\circ\vec{\nabla}\right)\left(\vec{\sigma}\circ\vec{\nabla}\right)=\vec{\nabla}^2 \tag{11.3.130}$$

which is for recall 2×2 matrix. Pauli's matrices are therefore good candidates to solve our problem!

We can also found in the literature the following notation based on the tensor product (see section Tensor Calculus page 2073):

$$\gamma^0=\sigma_1\otimes\mathbb{1}=\begin{bmatrix}0&1\\1&0\end{bmatrix}\otimes\begin{bmatrix}1&0\\0&1\end{bmatrix}=\begin{bmatrix}0&0&1&0\\0&0&0&1\\1&0&0&0\\0&1&0&0\end{bmatrix}=\begin{bmatrix}0&\mathbb{1}\\\mathbb{1}&0\end{bmatrix}\tag{11.3.131}$$

and:

$$\gamma^k=-\mathrm{i}\sigma_2\otimes\sigma_k\tag{11.3.132}$$

Remarks

R1. As we saw in the section of Spinor Calculus (Algebra Chapter), σ_0 is not really a Pauli matrix in itself. However, in some textbooks, it is indicated as being one (this is also our choice here).

R2. As we have also seen in it in the section of Spinor Calculus, let us recall that the Pauli matrices implicitly represent infinitesimal spatial rotations of a spinor.

Caution!!! In the gamma matrices, the traditional notation 0 means in fact that it is a 2×2 matrices having all components equal to zero. It should be instead denoted as seen in the section of Linear Algebra.

This allows us, finally, to combine the equations:

$$\begin{aligned}\mathrm{i}\hbar\left(\frac{\partial}{\partial t}+c\vec{\sigma}\circ\vec{\nabla}\right)\chi^{(+)}=m_0c^2\chi^{(-)}\\\mathrm{i}\hbar\left(\frac{\partial}{\partial t}-c\vec{\sigma}\circ\vec{\nabla}\right)\chi^{(-)}=m_0c^2\chi^{(+)}\end{aligned}\tag{11.3.133}$$

in only one (do not forget the association of the operators $E\rightarrow\mathrm{i}\hbar\partial/\partial t,p\rightarrow-\mathrm{i}\hbar\vec{\nabla}$):

$$\left(\gamma_0\mathrm{i}\frac{\partial}{c\partial t}+\gamma_i\frac{\partial}{\partial x_i}-\frac{m_0c}{\hbar}\right)\Psi=\vec{0}\tag{11.3.134}$$

Using the traditional notation in Tensor Calculus and choosing the natural units ($c=1,\hbar=1$) we have:

$$\boxed{\left(\mathrm{i}\gamma^\mu\partial_\mu-m_0\right)\Psi=\vec{0}}\tag{11.3.135}$$

which is the usual form of the "Dirac equation" (implicitly a system of four coupled differential equations) or "relativistic equation of the electron" with the "covariant derivative":

$$\partial_\mu=\left(\frac{\partial}{c\partial t},\frac{\partial}{\partial x},\frac{\partial}{\partial y},\frac{\partial}{\partial z}\right)\tag{11.3.136}$$

and where the null vector is (for recall...) a vector with four components all equal to zero!

We define the "covariant form of the Dirac operator" as:

$$\hat{D}=\mathrm{i}\gamma^{\mu}\partial_{\mu}-m_0 \tag{11.3.137}$$

Notice something important...! Time is derived once and also space at the opposite of Schrödinger equation. This symmetry of treatment between time and space is now compatible with what we know from Special Relativity where space and time and treated equally. So we have solved the issue we had earlier Schrödinger equation.

Or explicitly:

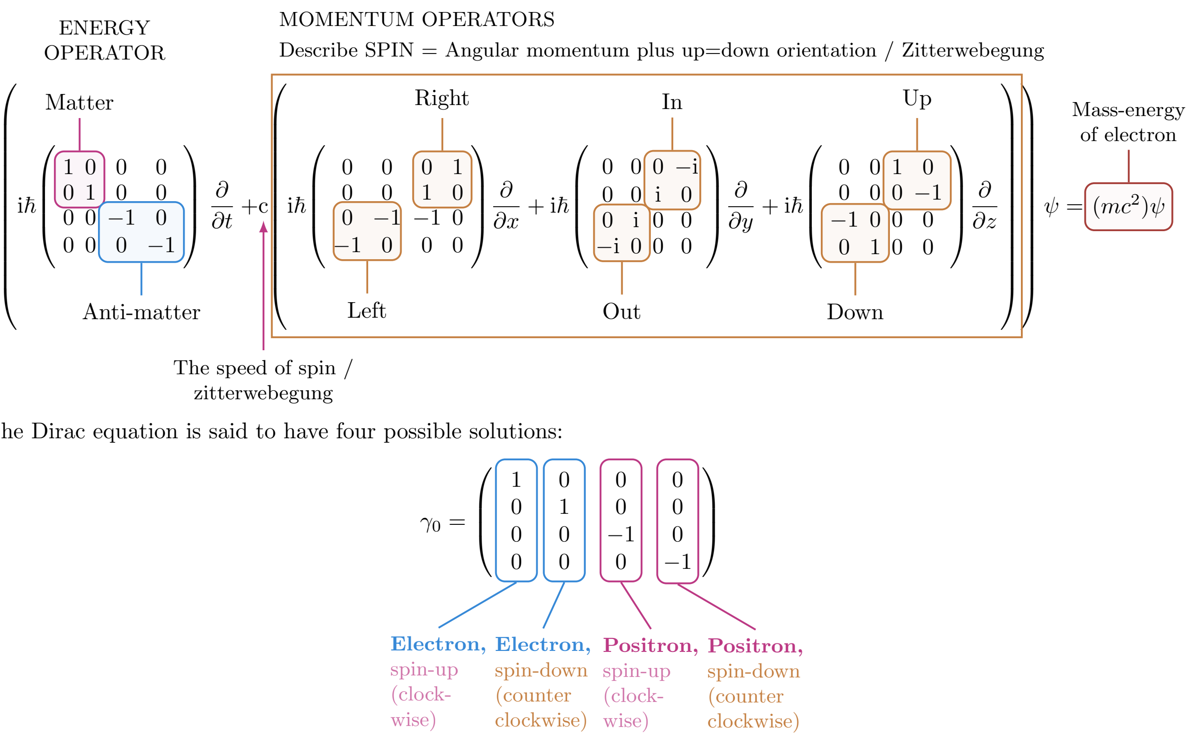

ENERGY OPERATOR
MOMENTUM OPERATORS
Describe SPIN = Angular momentum plus up=down orientation / Zitterwebegung
Matter
Right
In
Up
Mass-energy of electron
Anti-matter
Left
Out
Down
The speed of spin / zitterwebegung
The Dirac equation is said to have four possible solutions:
$\gamma_0 =$
Electron, spin-up (clock-wise)
Electron, spin-down (counter clockwise)
Positron, spin-up (clock-wise)
Positron, spin-down (counter clockwise)

Figure 11.82: Explicit free Dirac equation with explanations

Remark

In elementary particle physics, the antecedent relation is named the "covariant relativistic equation of fermions" because it describes particles with spin $1/2$.

The matrices γ^μ are named commonly "Dirac matrices". In an even more condensed form (using the "Feynman slash") the Dirac equation is sometimes written:

$$\boxed{(\mathrm{i}\not{\partial}-m_0)\,\Psi=\vec{0}} \tag{11.3.138}$$

or more explicitly in natural units (rare form but still correct):

$$\left(\mathrm{i}\gamma^0\frac{\partial}{\partial t}+\mathrm{i}\vec{\gamma}\circ\vec{\nabla}-m_0\right)\Psi=\vec{0} \tag{11.3.139}$$

Or with a minimum of abbreviations (all terms explicited):

$$\begin{bmatrix} \mathrm{i}\dfrac{\partial}{\partial t}-m_0 & 0 & \mathrm{i}\dfrac{\partial}{\partial z} & \mathrm{i}\dfrac{\partial}{\partial x}+\dfrac{\partial}{\partial y} \\ 0 & \mathrm{i}\dfrac{\partial}{\partial t}-m_0 & \mathrm{i}\dfrac{\partial}{\partial t}-\dfrac{\partial}{\partial y} & -\mathrm{i}\dfrac{\partial}{\partial z} \\ -\mathrm{i}\dfrac{\partial}{\partial z} & -\mathrm{i}\dfrac{\partial}{\partial x}-\dfrac{\partial}{\partial y} & -\mathrm{i}\dfrac{\partial}{\partial t}-m_0 & 0 \\ -\mathrm{i}\dfrac{\partial}{\partial x}+\dfrac{\partial}{\partial y} & \mathrm{i}\dfrac{\partial}{\partial z} & 0 & -\mathrm{i}\dfrac{\partial}{\partial t}-m_0 \end{bmatrix} \begin{pmatrix} \Phi_1 \\ \Phi_2 \\ \Phi_3 \\ \Phi_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \qquad (11.3.140)$$

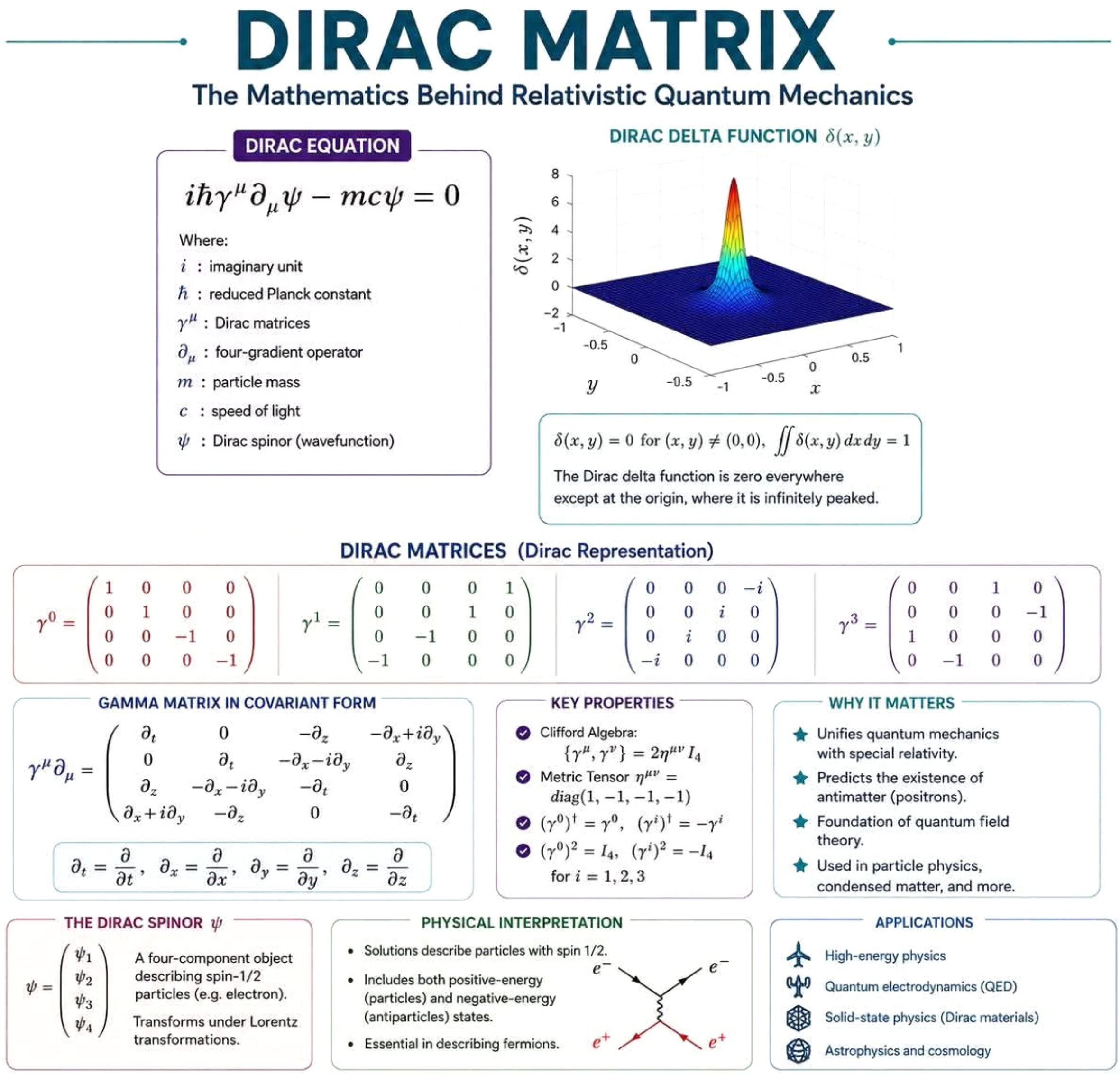
We thus have, as in analogy with Maxwell's equations, a differential equations of the first order which have for properties:

P1. To allow to fall back on the Klein-Gordon equation, in extenso on Wave equation (as for the Maxwell's equations)

P2. To take into account (explicitly) describe the spinor character of wave functions as we shall see by looking more closely at the Pauli matrices.

Remark

Since the Dirac equation applies to the spin $1/2$ particles, it also applies to neutrinos whose mass at rest is almost zero (thus the resolution of the Dirac equation is largely simplified).



DIRAC MATRIX

The Mathematics Behind Relativistic Quantum Mechanics

DIRAC EQUATION

$$i\hbar\gamma^\mu\partial_\mu\psi - mc\psi = 0$$

Where:

i : imaginary unit

$\hbar$: reduced Planck constant

γ^μ : Dirac matrices

∂_μ : four-gradient operator

m : particle mass

c : speed of light

ψ : Dirac spinor (wavefunction)

DIRAC DELTA FUNCTION $\delta(x,y)$

$\delta(x,y)=0$ for $(x,y)\neq(0,0)$, $\iint \delta(x,y)\,dx\,dy = 1$

The Dirac delta function is zero everywhere except at the origin, where it is infinitely peaked.

DIRAC MATRICES (Dirac Representation)

$$\gamma^0 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \quad \gamma^1 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix} \quad \gamma^2 = \begin{pmatrix} 0 & 0 & 0 & -i \\ 0 & 0 & i & 0 \\ 0 & i & 0 & 0 \\ -i & 0 & 0 & 0 \end{pmatrix} \quad \gamma^3 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}$$

GAMMA MATRIX IN COVARIANT FORM

$$\gamma^\mu\partial_\mu = \begin{pmatrix} \partial_t & 0 & -\partial_z & -\partial_x+i\partial_y \\ 0 & \partial_t & -\partial_x-i\partial_y & \partial_z \\ \partial_z & -\partial_x-i\partial_y & -\partial_t & 0 \\ \partial_x+i\partial_y & -\partial_z & 0 & -\partial_t \end{pmatrix}$$

$$\partial_t = \frac{\partial}{\partial t},\ \partial_x = \frac{\partial}{\partial x},\ \partial_y = \frac{\partial}{\partial y},\ \partial_z = \frac{\partial}{\partial z}$$

KEY PROPERTIES

- ✓ Clifford Algebra: $\{\gamma^\mu,\gamma^\nu\} = 2\eta^{\mu\nu}I_4$
- ✓ Metric Tensor $\eta^{\mu\nu} = diag(1,-1,-1,-1)$
- ✓ $(\gamma^0)^\dagger = \gamma^0$, $(\gamma^i)^\dagger = -\gamma^i$
- ✓ $(\gamma^0)^2 = I_4$, $(\gamma^i)^2 = -I_4$ for $i = 1,2,3$

WHY IT MATTERS

- ★ Unifies quantum mechanics with special relativity.
- ★ Predicts the existence of antimatter (positrons).
- ★ Foundation of quantum field theory.
- ★ Used in particle physics, condensed matter, and more.

THE DIRAC SPINOR ψ

$$\psi = \begin{pmatrix} \psi_1 \\ \psi_2 \\ \psi_3 \\ \psi_4 \end{pmatrix}$$

A four-component object describing spin-1/2 particles (e.g. electron).

Transforms under Lorentz transformations.

PHYSICAL INTERPRETATION

- Solutions describe particles with spin 1/2.
- Includes both positive-energy (particles) and negative-energy (antiparticles) states.
- Essential in describing fermions.

APPLICATIONS

- High-energy physics
- Quantum electrodynamics (QED)
- Solid-state physics (Dirac materials)
- Astrophysics and cosmology

11.3.3.1 Gamma, alpha, beta matrices

In the purpose now of interpreting the physical content of the Dirac equation, we will use a different representation of the gamma matrices. We have introduced earlier the representation by the following matrices 4×4:

$$\gamma^0=\beta=\begin{bmatrix} 0 & \mathbb{1} \\ \mathbb{1} & 0 \end{bmatrix},\gamma^i=\begin{bmatrix} 0 & -\sigma_i \\ \sigma_i & 0 \end{bmatrix} \tag{11.3.141}$$

named "standard representation" and denoted rigorously:

$$\gamma_S^0=\beta_S=\begin{bmatrix} 0 & \mathbb{1} \\ \mathbb{1} & 0 \end{bmatrix},\gamma_S^i=\begin{bmatrix} 0 & -\sigma_i \\ \sigma_i & 0 \end{bmatrix} \tag{11.3.142}$$

with for recall:

$$\sigma_1=\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},\sigma_2=\begin{bmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{bmatrix},\sigma_3=\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \tag{11.3.143}$$

But in fact these matrices are by far not the only 4×4 matrices that satisfy the following set of equalities that we will justify later during our study of the linearised Dirac equation (page 4957):

$$\begin{aligned}&\beta^2 = \alpha_1^2 = \alpha_2^2 = \alpha_3^2 = 1\\&[\alpha_1, \alpha_2]_+ = [\alpha_1, \alpha_3]_+ = [\alpha_2, \alpha_3]_+ = 0\\&[\alpha_1, \beta]_+ = [\alpha_2, \beta]_+ = [\alpha_3, \beta]_+ = 0\end{aligned} \tag{11.3.144}$$

For example, we will use many times the "Dirac representation" defined by:

$$\gamma_D^0 = \beta_D = \begin{bmatrix} \mathbb{1} & 0 \\ 0 & -\mathbb{1} \end{bmatrix}, \gamma_D^i = -\gamma_D^0 \alpha_D^i = \begin{bmatrix} 0 & \sigma_i \\ -\sigma_i & 0 \end{bmatrix} \tag{11.3.145}$$

Or more explicitly:

$$\gamma_D^0 = \beta_D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}, \gamma_D^1 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{bmatrix}, \gamma_D^2 = \begin{bmatrix} 0 & 0 & 0 & -\mathrm{i} \\ 0 & 0 & \mathrm{i} & 0 \\ 0 & \mathrm{i} & 0 & 0 \\ -\mathrm{i} & 0 & 0 & 0 \end{bmatrix}, \gamma_D^3 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \tag{11.3.146}$$

With the corresponding Dirac alpha and beta matrices:

$$\beta_D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}, \alpha_D^1 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \alpha_D^2 = \begin{bmatrix} 0 & 0 & 0 & -\mathrm{i} \\ 0 & 0 & \mathrm{i} & 0 \\ 0 & -\mathrm{i} & 0 & 0 \\ \mathrm{i} & 0 & 0 & 0 \end{bmatrix}, \alpha_D^3 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix} \tag{11.3.147}$$

This representation (also not unique!) is particularly useful because it highlights the spinor character (due to the half-integer spin) of the electron wave function and separates the positive and negative energy components.

Notice that $\gamma^1, \gamma^2, \gamma^3$ are anti-Hermitian:

$$\gamma^{0\dagger} = \gamma^0, \gamma^{1\dagger} = -\gamma^1, \gamma^{2\dagger} = -\gamma^2, \gamma^{3\dagger} = -\gamma^3 \tag{11.3.148}$$

With Maple™ 4.00b it can easily be checked that they satisfy the set of three equalities introduced just earlier (see the Maple™ companion book):

```
>with(linalg):
>beta:=array([[1,0,0,0],[0,1,0,0],[0,0,-1,0],[0,0,0,-1]]);
>alpha1:=array([[0,0,0,1],[0,0,1,0],[0,1,0,0],[1,0,0,0]]);
>alpha2:=array([[0,0,0,-I],[0,0,I,0],[0,-I,0,0],[I,0,0,0]]);
>alpha3:=array([[0,0,1,0],[0,0,0,-1],[1,0,0,0],[0,-1,0,0]]);
>evalm(beta^2);
>evalm(alpha1^2);
>evalm(alpha2^2);
>evalm(alpha3^2);
>readlib(commutat):
```

```
>evalm(c(alpha1,alpha2));
>evalm(c(alpha1,alpha3));
>evalf(evalm(c(alpha2,alpha3)));
>evalf(evalm(c(alpha1,beta)));
>evalf(evalm(c(alpha2,beta)));
>evalf(evalm(c(alpha3,beta)));
```

We can also easily check (elementary linear algebra) that the Dirac representation is obtained from the standard representation by the transformation:

$$\gamma_D^\mu \rightarrow U\gamma_S^\mu U^\dagger \tag{11.3.149}$$

where:

$$\gamma_D^\mu \rightarrow U\gamma_S^\mu U^\dagger \quad \text{with} \quad U=\frac{1}{\sqrt{2}}\begin{bmatrix}\mathbb{1} & \mathbb{1}\\ \mathbb{1} & -\mathbb{1}\end{bmatrix} \tag{11.3.150}$$

Let us recall that $U^\dagger$ is the adjoint matrix (the conjugate of the transposed matrix) of U (see section Linear Algebra page 1953). Now, when all the elements are real as it is the case above and the matrix is a square one symmetric relatively to all its off-diagonal elements then we have obviously that $U=U^\dagger$.

Proof 11.5.2.

$$\begin{aligned}\gamma_D^0=U\gamma_S^0U^\dagger&=\frac{1}{\sqrt{2}}\begin{bmatrix}\mathbb{1} & \mathbb{1}\\ \mathbb{1} & -\mathbb{1}\end{bmatrix}\begin{bmatrix}0 & \mathbb{1}\\ \mathbb{1} & 0\end{bmatrix}\frac{1}{\sqrt{2}}\begin{bmatrix}\mathbb{1} & \mathbb{1}\\ \mathbb{1} & -\mathbb{1}\end{bmatrix}=\frac{1}{2}\begin{bmatrix}\mathbb{1} & \mathbb{1}\\ \mathbb{1} & -\mathbb{1}\end{bmatrix}\begin{bmatrix}0 & \mathbb{1}\\ \mathbb{1} & 0\end{bmatrix}\begin{bmatrix}\mathbb{1} & \mathbb{1}\\ \mathbb{1} & -\mathbb{1}\end{bmatrix}\\ &=\frac{1}{2}\begin{bmatrix}\mathbb{1} & \mathbb{1}\\ \mathbb{1} & -\mathbb{1}\end{bmatrix}\begin{bmatrix}\mathbb{1} & -\mathbb{1}\\ \mathbb{1} & \mathbb{1}\end{bmatrix}=\frac{1}{2}\begin{bmatrix}\mathbb{1}+\mathbb{1} & -\mathbb{1}+\mathbb{1}\\ \mathbb{1}-\mathbb{1} & -\mathbb{1}-\mathbb{1}\end{bmatrix}=\frac{1}{2}\begin{bmatrix}2\mathbb{1} & 0\\ 0 & -2\mathbb{1}\end{bmatrix}=\begin{bmatrix}\mathbb{1} & 0\\ 0 & -\mathbb{1}\end{bmatrix}\end{aligned} \tag{11.3.151}$$

and:

$$\begin{aligned}\gamma_D^i=U\gamma_S^iU^\dagger&=\frac{1}{\sqrt{2}}\begin{bmatrix}\mathbb{1} & \mathbb{1}\\ \mathbb{1} & -\mathbb{1}\end{bmatrix}\begin{bmatrix}0 & -\sigma_i\\ \sigma_i & 0\end{bmatrix}\frac{1}{\sqrt{2}}\begin{bmatrix}\mathbb{1} & \mathbb{1}\\ \mathbb{1} & -\mathbb{1}\end{bmatrix}=\frac{1}{2}\begin{bmatrix}\mathbb{1} & \mathbb{1}\\ \mathbb{1} & -\mathbb{1}\end{bmatrix}\begin{bmatrix}0 & -\sigma_i\\ \sigma_i & 0\end{bmatrix}\begin{bmatrix}\mathbb{1} & \mathbb{1}\\ \mathbb{1} & -\mathbb{1}\end{bmatrix}\\ &=\frac{1}{2}\begin{bmatrix}\mathbb{1} & \mathbb{1}\\ \mathbb{1} & -\mathbb{1}\end{bmatrix}\begin{bmatrix}-\sigma_i & \sigma_i\\ \sigma_i & \sigma_i\end{bmatrix}=\frac{1}{2}\begin{bmatrix}0 & 2\sigma_i\\ -2\sigma_i & 0\end{bmatrix}=\begin{bmatrix}0 & \sigma_i\\ -\sigma_i & 0\end{bmatrix}\end{aligned} \tag{11.3.152}$$

■ Q.E.D.

But notice that there is also another possible choice, named the "chiral representation" defined by:

$$\gamma_C^0=\beta_C=\begin{bmatrix}0 & -\mathbb{1}\\ -\mathbb{1} & 0\end{bmatrix},\gamma_C^i=-\alpha_D^i\gamma_C^0=\begin{bmatrix}0 & \sigma_i\\ -\sigma_i & 0\end{bmatrix} \tag{11.3.153}$$

Or more explicitly:

$$\gamma_C^0=\beta_C=\begin{bmatrix}0 & 0 & -1 & 0\\ 0 & 0 & 0 & -1\\ -1 & 0 & 0 & 0\\ 0 & -1 & 0 & 0\end{bmatrix},\gamma_C^1=\begin{bmatrix}0 & 0 & 0 & 1\\ 0 & 0 & 1 & 0\\ 0 & -1 & 0 & 0\\ -1 & 0 & 0 & 0\end{bmatrix},\gamma_C^2=\begin{bmatrix}0 & 0 & 0 & -\mathrm{i}\\ 0 & 0 & \mathrm{i} & 0\\ 0 & \mathrm{i} & 0 & 0\\ -\mathrm{i} & 0 & 0 & 0\end{bmatrix},\gamma_C^3=\begin{bmatrix}0 & 0 & 1 & 0\\ 0 & 0 & 0 & -1\\ -1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\end{bmatrix} \tag{11.3.154}$$

With the corresponding chiral alpha and beta matrices:

$$\beta_C=\begin{bmatrix}0 & -\mathbb{1}\\ -\mathbb{1} & 0\end{bmatrix},\alpha_C^i=\begin{bmatrix}\sigma_i & 0\\ 0 & -\sigma_i\end{bmatrix} \tag{11.3.155}$$

Thus more explicitly:

$$\beta_C=\begin{bmatrix}0 & 0 & -1 & 0\\ 0 & 0 & 0 & -1\\ -1 & 0 & 0 & 0\\ 0 & -1 & 0 & 0\end{bmatrix},\alpha_C^1=\begin{bmatrix}0 & 1 & 0 & 0\\ 1 & 0 & 0 & 0\\ 0 & 0 & 0 & -1\\ 0 & 0 & -1 & 0\end{bmatrix},\alpha_C^2=\begin{bmatrix}0 & -\mathrm{i} & 0 & 0\\ \mathrm{i} & 0 & 0 & 0\\ 0 & 0 & 0 & \mathrm{i}\\ 0 & 0 & -\mathrm{i} & 0\end{bmatrix},\alpha_C^3=\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & -1 & 0 & 0\\ 0 & 0 & -1 & 0\\ 0 & 0 & 0 & 1\end{bmatrix} \tag{11.3.156}$$

Its advantage is that both spinors transform independently under rotations and translations. It is particularly useful for massless particles, the equations being considerably simplified. It has been used for the neutrino although neutrino oscillations show that their mass is non-zero[26] (the electron-neutrino has for example a mass of the order of $1.4\cdot 10^{-36}$ [kg]).

But notice that there is again also another possible choice, named the "Weyl representation" defined by:

$$\gamma_W^0=\beta_W=\begin{bmatrix}0 & \mathbb{1}\\ \mathbb{1} & 0\end{bmatrix},\gamma_W^i=\alpha_W^i\gamma_W^0=\begin{bmatrix}0 & \sigma_i\\ -\sigma_i & 0\end{bmatrix} \tag{11.3.157}$$

Or more explicitly:

$$\gamma_W^0=\beta_C=\begin{bmatrix}0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\\ 1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\end{bmatrix},\gamma_W^1=\begin{bmatrix}0 & 0 & 0 & 1\\ 0 & 0 & 1 & 0\\ 0 & -1 & 0 & 0\\ -1 & 0 & 0 & 0\end{bmatrix},\gamma_W^2=\begin{bmatrix}0 & 0 & 0 & -\mathrm{i}\\ 0 & 0 & \mathrm{i} & 0\\ 0 & \mathrm{i} & 0 & 0\\ -\mathrm{i} & 0 & 0 & 0\end{bmatrix},\gamma_W^3=\begin{bmatrix}0 & 0 & 1 & 0\\ 0 & 0 & 0 & -1\\ -1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\end{bmatrix} \tag{11.3.158}$$

With the corresponding Weyl alpha and beta matrices:

$$\beta_W=\begin{bmatrix}0 & \mathbb{1}\\ \mathbb{1} & 0\end{bmatrix},\alpha_W^i=\begin{bmatrix}\sigma_i & 0\\ 0 & -\sigma_i\end{bmatrix} \tag{11.3.159}$$

Thus more explicitly:

$$\beta_W=\begin{bmatrix}0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\\ 1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\end{bmatrix},\alpha_W^1=\begin{bmatrix}0 & 1 & 0 & 0\\ 1 & 0 & 0 & 0\\ 0 & 0 & 0 & -1\\ 0 & 0 & -1 & 0\end{bmatrix},\alpha_W^2=\begin{bmatrix}0 & -\mathrm{i} & 0 & 0\\ \mathrm{i} & 0 & 0 & 0\\ 0 & 0 & 0 & \mathrm{i}\\ 0 & 0 & -\mathrm{i} & 0\end{bmatrix},\alpha_W^3=\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & -1 & 0 & 0\\ 0 & 0 & -1 & 0\\ 0 & 0 & 0 & 1\end{bmatrix} \tag{11.3.160}$$

This representation is useful when one seeks to derive the Dirac equation using the irreducible representations of the Lorentz group.

---

[26] The 12015 (holocene calendar) Nobel Prize in Physics has been awarded to Takaaki Kajita and Arthur B. McDonald for the discovery of neutrino oscillations, which shows that neutrinos have mass. There are three flavors of neutrinos: electron, muon, and tau neutrinos. But there are also three different mass neutrinos: ν_1, ν_2, and ν_3. If each flavor of neutrino matched up with one specific mass, then life would be simple... but boring. Neutrinos are tricky, and it turns out that the mass and flavor neutrinos do not overlap perfectly. Instead, Quantum Mechanics comes into play. Each neutrino of a specific flavor is actually a superposition of neutrinos of different masses!

But notice that there is again also another possible choice, named the "Majorana representation" defined by:

$$\gamma_M^0=\beta_M=\begin{bmatrix}0&0&0&-\mathrm{i}\\0&0&-\mathrm{i}&0\\0&\mathrm{i}&0&0\\\mathrm{i}&0&0&0\end{bmatrix},\gamma_M^1=\begin{bmatrix}0&-\mathrm{i}&0&0\\-\mathrm{i}&0&0&0\\0&0&0&\mathrm{i}\\0&0&\mathrm{i}&0\end{bmatrix},\gamma_M^2=\begin{bmatrix}-\mathrm{i}&0&0&0\\0&\mathrm{i}&0&0\\0&0&-\mathrm{i}&0\\0&0&0&\mathrm{i}\end{bmatrix},\gamma_M^3=\begin{bmatrix}0&0&0&\mathrm{i}\\0&0&\mathrm{i}&0\\0&\mathrm{i}&0&0\\\mathrm{i}&0&0&0\end{bmatrix}\tag{11.3.161}$$

Majorana matrices in the representation above obey also the following mandatory gamma anticommutation relations (as so far we don't know the β and α_i corresponding Majorana matrices) as seen at page 4963 (and that can easily be verified using the same type of Maple™ code as given just earlier!):

$$[\gamma_i,\gamma_j]_+=\gamma_i\gamma_j+\gamma_j\gamma_i=2\eta_{ij}\mathbb{1}_4\tag{11.3.162}$$

Thus, for $i\neq j$:

$$[\gamma_i,\gamma_j]_+=\gamma_i\gamma_j+\gamma_j\gamma_i=0\tag{11.3.163}$$

and it is also immediate using this anticommutation relation that:

$$(\gamma^0)^2=\mathbb{1},(\gamma^i)^2=-\mathbb{1}\tag{11.3.164}$$

A further, purely imaginary representation of the Majorana matrices spread in the literature is given by:

$$\begin{gathered}\tilde{\gamma}_M^0=\begin{bmatrix}0&0&0&-\mathrm{i}\\0&0&\mathrm{i}&0\\0&-\mathrm{i}&0&0\\\mathrm{i}&0&0&0\end{bmatrix}=\begin{bmatrix}0&\sigma_2\\\sigma_2&0\end{bmatrix},\tilde{\gamma}_M^1=\begin{bmatrix}\mathrm{i}&0&0&0\\0&-\mathrm{i}&0&0\\0&0&\mathrm{i}&0\\0&0&0&-\mathrm{i}\end{bmatrix}=\mathrm{i}\begin{bmatrix}\sigma_3&0\\0&\sigma_3\end{bmatrix}\\\tilde{\gamma}_M^2=\begin{bmatrix}0&0&0&\mathrm{i}\\0&0&-\mathrm{i}&0\\0&-\mathrm{i}&0&0\\\mathrm{i}&0&0&0\end{bmatrix}=\begin{bmatrix}0&-\sigma_2\\\sigma_2&0\end{bmatrix},\tilde{\gamma}_M^3=\begin{bmatrix}0&-\mathrm{i}&0&0\\-\mathrm{i}&0&0&0\\0&0&0&-\mathrm{i}\\0&0&-\mathrm{i}&0\end{bmatrix}=-\mathrm{i}\begin{bmatrix}\sigma_1&0\\0&\sigma_1\end{bmatrix}\end{gathered}\tag{11.3.165}$$

That latter Majorana representation can be obtained from the original representation by the unitary transformation:

$$\tilde{\gamma}_M^\mu=\tilde{U}\gamma_M^\mu\tilde{U}^{-1}\quad\text{and}\quad\tilde{U}=\frac{1}{2}=\begin{bmatrix}1&1&1&-1\\-1&1&1&1\\1&1&-1&1\\1&-1&1&1\end{bmatrix}\tag{11.3.166}$$

There is alsoooooo another Majorana representation (sorry...) also denoted identically as previously (sic!) and defined by:

$$\tilde{\gamma}_M^0=\sigma_2\otimes\sigma_1,\tilde{\gamma}_M^1=\mathrm{i}\sigma_1\otimes\mathbb{1},\tilde{\gamma}_M^2=\mathrm{i}\sigma_3\otimes\mathbb{1},\tilde{\gamma}_M^3=\mathrm{i}\sigma_2\otimes\sigma_2\tag{11.3.167}$$

That is explicitly:

$$\tilde{\gamma}_M^0=\begin{bmatrix}0&0&0&-\mathrm{i}\\0&0&-\mathrm{i}&0\\0&\mathrm{i}&0&0\\\mathrm{i}&0&0&0\end{bmatrix},\tilde{\gamma}_M^1=\begin{bmatrix}0&0&\mathrm{i}&0\\0&0&0&\mathrm{i}\\\mathrm{i}&0&0&0\\0&\mathrm{i}&0&0\end{bmatrix},\tilde{\gamma}_M^2=\begin{bmatrix}\mathrm{i}&0&0&0\\0&\mathrm{i}&0&0\\0&0&-\mathrm{i}&0\\0&0&0&-\mathrm{i}\end{bmatrix},\tilde{\gamma}_M^3=\begin{bmatrix}0&0&0&-\mathrm{i}\\0&0&\mathrm{i}&0\\0&\mathrm{i}&0&0\\-\mathrm{i}&0&0&0\end{bmatrix}\tag{11.3.168}$$

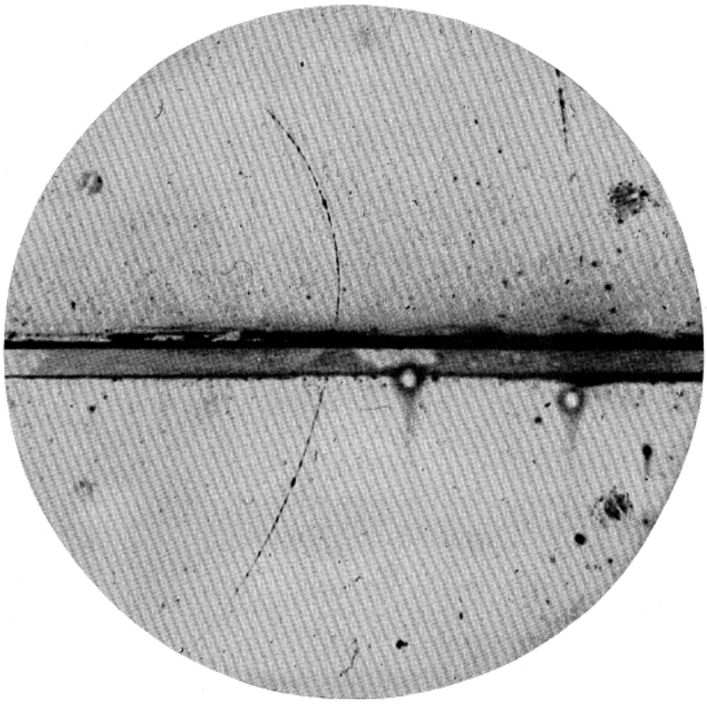
This representation has the interesting property that all matrices γ^μ are purely imaginary! What makes the Dirac field of the Dirac equation a non-complex field where the particles are therefore their own antiparticles!

Remarks

R1. All these possibilities explains why physicists speaks sometimes of Dirac, Majorana or Weyl fermions.

R2. Sadly it seems that there is no international convention on the definition and names of the above representations. It is a bit a mess and depending on the textbook you will not get the same definitions. They're also the leading causes of stress disorders amongst physicists.

Below a photo of what is assumed to be the first official detection and measurement in a Wilson bubble chamber of a positron (see [17]) by Carl David Anderson in 11933 (Holocene calendar) who shared six years later the Nobel Prize in Physics for having observed this particle with Victor Hess and that confirmed the existence of antimatter:



11.3.3.2 Free particle solution

Let us now look for solutions specific to the Dirac equation in the form of a free particle[27]:

$$\Psi=\Psi_{\vec{p}}e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}=\Psi_{\vec{p}}e^{-\mathrm{i}(Et-\vec{p}\circ\vec{r})} \tag{11.3.169}$$

---

[27] If the developments may bee to hard for the reader, he can jump straight to the case of the free particle at rest solution at page 4952

By substituting in the Dirac equation and after simplification by $e^{\mathrm{i}\vec{p}\circ\vec{r}-Et}$ we find easily:

$$\boxed{\left(\mathrm{i}\gamma^0(-\mathrm{i}E)+\mathrm{i}\sum_j\gamma^j(\mathrm{i}p_j)-m_0\right)\Psi_{\vec{p}}=\vec{0}} \tag{11.3.170}$$

Indeed in natural units:

$$\begin{aligned}
&(\mathrm{i}\gamma^\mu\partial_\mu-m_0)\,\Psi_{\vec{p}}e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}=\mathrm{i}\gamma^\mu\partial_\mu\Psi_{\vec{p}}e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}-m_0\Psi_{\vec{p}}e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}\\
&=\mathrm{i}\gamma^0\partial_0\Psi_{\vec{p}}e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}+\mathrm{i}\gamma^1\partial_1\Psi_{\vec{p}}e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}+\mathrm{i}\gamma^2\partial_2\Psi_{\vec{p}}e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}+\mathrm{i}\gamma^3\partial_3\Psi_{\vec{p}}e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}-m_0\Psi_{\vec{p}}e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}\\
&=\mathrm{i}\gamma^0\frac{\partial}{\partial t}\Psi_{\vec{p}}e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}+\mathrm{i}\gamma^1\frac{\partial}{\partial x}\Psi_{\vec{p}}e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}+\mathrm{i}\gamma^2\frac{\partial}{\partial y}\Psi_{\vec{p}}e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}+\mathrm{i}\gamma^3\frac{\partial}{\partial z}\Psi_{\vec{p}}e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}-m_0\Psi_{\vec{p}}e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}\\
&=(-\mathrm{i}E)\mathrm{i}\gamma^0\Psi_{\vec{p}}e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}+\mathrm{i}\gamma^0e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}\frac{\partial}{\partial t}\Psi_{\vec{p}}\\
&+\mathrm{i}\gamma^1\frac{\partial}{\partial x}\Psi_{\vec{p}}e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}+\mathrm{i}\gamma^2\frac{\partial}{\partial y}\Psi_{\vec{p}}e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}+\mathrm{i}\gamma^3\frac{\partial}{\partial z}\Psi_{\vec{p}}e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}-m_0\Psi_{\vec{p}}e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}\\
&=(-\mathrm{i}E)\mathrm{i}\gamma^0\Psi_{\vec{p}}e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}\\
&+\mathrm{i}\gamma^1\mathrm{i}p_x\Psi_{\vec{p}}e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}+\mathrm{i}\gamma^2\mathrm{i}p_y\Psi_{\vec{p}}e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}+\mathrm{i}\gamma^3\mathrm{i}p_z\Psi_{\vec{p}}e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}-m_0\Psi_{\vec{p}}e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}\\
&+\mathrm{i}\gamma^1e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}\frac{\partial}{\partial x}\Psi_{\vec{p}}+\mathrm{i}\gamma^2e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}\frac{\partial}{\partial y}\Psi_{\vec{p}}+\mathrm{i}\gamma^3e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}\frac{\partial}{\partial z}\Psi_{\vec{p}}\\
&=(-\mathrm{i}E)\mathrm{i}\gamma^0\Psi_{\vec{p}}e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}\\
&+\mathrm{i}\gamma^1\mathrm{i}p_x\Psi_{\vec{p}}e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}+\mathrm{i}\gamma^2\mathrm{i}p_y\Psi_{\vec{p}}e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}+\mathrm{i}\gamma^3\mathrm{i}p_z\Psi_{\vec{p}}e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}-m_0\Psi_{\vec{p}}e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}
\end{aligned} \tag{11.3.171}$$

With the representation of Dirac, we get after development (trivial calculation that we can detail on request as always):

$$\begin{pmatrix}E-m_0 & -\vec{p}\circ\vec{\sigma}\\ \vec{p}\circ\vec{\sigma} & -E-m_0\end{pmatrix}\Psi_{\vec{p}}=\vec{0} \tag{11.3.172}$$

Or more explicitly:

$$\begin{pmatrix}(E-m_0)\mathbb{1} & -\vec{p}\circ\vec{\sigma}\\ \vec{p}\circ\vec{\sigma} & (-E-m_0)\mathbb{1}\end{pmatrix}\Psi_{\vec{p}}=\vec{0} \tag{11.3.173}$$

Indeed, with the Einstein summation convention (see section Tensor Calculus page 2044):

$$\begin{aligned}
&\left(\mathrm{i}\gamma^0(-\mathrm{i}E)+\mathrm{i}\sum_j\gamma^j(\mathrm{i}p_j)-m_0\right)\Psi_{\vec{p}}=\left(\gamma^0E-\gamma^jp_j-m_0\right)\Psi_{\vec{p}}\\
&=\left(\begin{bmatrix}\mathbb{1}&0\\0&-\mathbb{1}\end{bmatrix}E-\begin{bmatrix}0&\sigma_1\\-\sigma_1&0\end{bmatrix}p_1-\begin{bmatrix}0&\sigma_2\\-\sigma_2&0\end{bmatrix}p_2-\begin{bmatrix}0&\sigma_3\\-\sigma_3&0\end{bmatrix}p_3-m_0\right)\Psi_{\vec{p}}\\
&=\left(\begin{bmatrix}E&0\\0&-E\end{bmatrix}-\begin{bmatrix}0&\sigma_1p_1\\-\sigma_1p_1&0\end{bmatrix}-\begin{bmatrix}0&\sigma_2p_2\\-\sigma_2p_2&0\end{bmatrix}-\begin{bmatrix}0&\sigma_3p_3\\-\sigma_3p_3&0\end{bmatrix}-m_0\right)\Psi_{\vec{p}}\\
&=\left(\begin{bmatrix}E&0\\0&-E\end{bmatrix}-\begin{bmatrix}0&\sigma_1p_1+\sigma_2p_2+\sigma_3p_3\\-\sigma_1p_1-\sigma_2p_2-\sigma_3p_3&0\end{bmatrix}-m_0\right)\Psi_{\vec{p}}\\
&=\left(\begin{bmatrix}E&0\\0&-E\end{bmatrix}-\begin{bmatrix}0&\vec{p}\circ\vec{\sigma}\\-\vec{p}\circ\vec{\sigma}&0\end{bmatrix}-m_0\right)\Psi_{\vec{p}}=\left(\begin{bmatrix}E&0\\0&-E\end{bmatrix}+\begin{bmatrix}0&-\vec{p}\circ\vec{\sigma}\\\vec{p}\circ\vec{\sigma}&0\end{bmatrix}-m_0\right)\Psi_{\vec{p}}\\
&=\left(\begin{bmatrix}E&-\vec{p}\circ\vec{\sigma}\\\vec{p}\circ\vec{\sigma}&-E\end{bmatrix}-m_0\right)\Psi_{\vec{p}}=\left(\begin{bmatrix}E&-\vec{p}\circ\vec{\sigma}\\\vec{p}\circ\vec{\sigma}&-E\end{bmatrix}\Psi_{\vec{p}}-m_0\Psi_{\vec{p}}\right)\\
&=\begin{pmatrix}E\phi_a-\vec{p}\circ\vec{\sigma}\phi_b-m_0\phi_a\\\vec{p}\circ\vec{\sigma}\phi_a-E\phi_b-m_0\phi_b\end{pmatrix}=\begin{pmatrix}\phi_a(E-m_0)-\vec{p}\circ\vec{\sigma}\phi_b\\\vec{p}\circ\vec{\sigma}\phi_a+\phi_b(-E-m_0)\end{pmatrix}=\begin{pmatrix}E-m_0&-\vec{p}\circ\vec{\sigma}\\\vec{p}\circ\vec{\sigma}&-E-m_0\end{pmatrix}\begin{pmatrix}\phi_a\\\phi_b\end{pmatrix}\\
&=\begin{pmatrix}E-m_0&-\vec{p}\circ\vec{\sigma}\\\vec{p}\circ\vec{\sigma}&-E-m_0\end{pmatrix}\Psi_{\vec{p}}=\vec{0}
\end{aligned}\tag{11.3.174}$$

In order for this matrix equation to have non-zero solutions, as usual, the determinant of the matrix must be zero (see section Linear Algebra page 1980). We easily check that:

$$\begin{aligned}
&\det\begin{pmatrix}E-m_0&-\vec{p}\circ\vec{\sigma}\\\vec{p}\circ\vec{\sigma}&-E-m_0\end{pmatrix}=(E-m_0)(-E-m_0)-(\vec{p}\circ\vec{\sigma})(-\vec{p}\circ\vec{\sigma})\\
&=-(E-m_0)(E+m_0)+\|\vec{p}\|^2=-(E^2-m_0^2)+\|\vec{p}\|^2=-E^2+m_0^2+\|\vec{p}\|^2=\vec{0}
\end{aligned}\tag{11.3.175}$$

Which implies (do not forget that we are in natural units!):

$$\boxed{E=\pm\sqrt{\|\vec{p}\|^2+m_0^2}}\tag{11.3.176}$$

Notice that coming back to our initial guess:

$$\Psi=\Psi_{\vec{p}}e^{-\mathrm{i}(Et-\vec{p}\circ\vec{r})}\tag{11.3.177}$$

The same solution with negative energy (positive charge) exist if we consider a particle going backward in time and with a negative motion such that:

$$\Psi=\Psi_{\vec{p}}e^{-\mathrm{i}(Et-\vec{p}\circ\vec{r})}=\Psi_{\vec{p}}e^{-\mathrm{i}((-E)(-t)-(-\vec{p})\circ(-\vec{r}))}\tag{11.3.178}$$

This is therefore a system of equations very funny (...) to solve....... (do not forget that each term of the above relation is implicitly a 2 × 2 matrix).

With Chirale representation we would have obtained (always by adopting the traditional notation of this textbook for the scalar product):

$$\begin{aligned}
&\left(\mathrm{i}\gamma^0(-\mathrm{i}E)+\mathrm{i}\sum_j\gamma^j(\mathrm{i}p_j)-m_0\right)\Psi_{\vec{p}}=\left(\gamma^0E-\gamma^jp_j-m_0\right)\Psi_{\vec{p}}\\
&=\left(\begin{bmatrix}0&\mathbb{1}\\\mathbb{1}&0\end{bmatrix}E-\begin{bmatrix}0&-\sigma_1\\\sigma_1&0\end{bmatrix}p_1-\begin{bmatrix}0&-\sigma_2\\\sigma_2&0\end{bmatrix}p_2-\begin{bmatrix}0&-\sigma_3\\\sigma_3&0\end{bmatrix}p_3-m_0\right)\Psi_{\vec{p}}\\
&=\left(\begin{bmatrix}0&E\\E&0\end{bmatrix}-\begin{bmatrix}0&-\sigma_1p_1\\\sigma_1p_1&0\end{bmatrix}-\begin{bmatrix}0&-\sigma_2p_2\\\sigma_2p_2&0\end{bmatrix}-\begin{bmatrix}0&-\sigma_3p_3\\\sigma_3p_3&0\end{bmatrix}-m_0\right)\Psi_{\vec{p}}\\
&=\left(\begin{bmatrix}0&E\\E&0\end{bmatrix}-\begin{bmatrix}0&-\sigma_1p_1-\sigma_2p_2-\sigma_3p_3\\\sigma_1p_1+\sigma_2p_2+\sigma_3p_3&0\end{bmatrix}-m_0\right)\Psi_{\vec{p}}\\
&=\left(\begin{bmatrix}0&E\\E&0\end{bmatrix}-\begin{bmatrix}0&-\vec{p}\circ\vec{\sigma}\\\vec{p}\circ\vec{\sigma}&0\end{bmatrix}-m_0\right)\Psi_{\vec{p}}=\left(\begin{bmatrix}0&E\\E&0\end{bmatrix}+\begin{bmatrix}0&\vec{p}\circ\vec{\sigma}\\-\vec{p}\circ\vec{\sigma}&0\end{bmatrix}-m_0\right)\Psi_{\vec{p}}\\
&=\left(\begin{bmatrix}0&\vec{p}\circ\vec{\sigma}+E\\-\vec{p}\circ\vec{\sigma}+E&0\end{bmatrix}-m_0\right)\Psi_{\vec{p}}=\left(\begin{bmatrix}0&\vec{p}\circ\vec{\sigma}+E\\-\vec{p}\circ\vec{\sigma}+E&0\end{bmatrix}\Psi_{\vec{p}}-m_0\Psi_{\vec{p}}\right)\\
&=\begin{pmatrix}0\phi_a+(\vec{p}\circ\vec{\sigma}+E)\phi_b-m_0\phi_a\\(-\vec{p}\circ\vec{\sigma}+E)\phi_a+0\phi_b-m_0\phi_b\end{pmatrix}=\begin{pmatrix}-m_0\phi_a+(\vec{p}\circ\vec{\sigma}+E)\phi_b\\(-\vec{p}\circ\vec{\sigma}+E)\phi_a-m_0\phi_b\end{pmatrix}=\begin{pmatrix}-m_0&\vec{p}\circ\vec{\sigma}+E\\-\vec{p}\circ\vec{\sigma}+E&-m_0\end{pmatrix}\begin{pmatrix}\phi_a\\\phi_b\end{pmatrix}\\
&=\begin{pmatrix}-m_0&\vec{p}\circ\vec{\sigma}+E\\-\vec{p}\circ\vec{\sigma}+E&-m_0\end{pmatrix}\Psi_{\vec{p}}=\vec{0}
\end{aligned}\tag{11.3.179}$$

Since the mass seems always positive, the Dirac equation thus has four linearly independent solutions, two of which have positive energy:

$$E=E_{\vec{p}}=\sqrt{\|\vec{p}\|^2+m_0^2}\tag{11.3.180}$$

and two with negative energy:

$$E=-E_{\vec{p}}\tag{11.3.181}$$

It is therefore indeed the antiparticles that we had determined during our study of the free Klein-Gordon equation but with the spin in addition, hence the doubling of the additional solutions (two possible spin orientations per particle and by antiparticle). With the Chirale representation, we would not have fallen back on this result. Hence the necessity of the use of the representation of Dirac of the Pauli matrices.

We therefore know that there are solutions to the Dirac equation. Let us now determine these. Let us put:

$$\Psi_{\vec{p}}=\begin{pmatrix}\phi_a\\\phi_b\end{pmatrix}\tag{11.3.182}$$

where ϕ_a, ϕ_b are for recall the two double components of the spinor. We thus write the system of equations:

$$\begin{aligned}(E-m_0)\phi_a-\vec{p}\circ\vec{\sigma}\phi_b&=\vec{0}\\\vec{p}\circ\vec{\sigma}\phi_a-(E+m_0)\phi_b&=\vec{0}\end{aligned}\tag{11.3.183}$$

which gives us (do not forget that the term to the denominator is actually a 2×2 matrix...):

$$\boxed{\phi_b=\frac{\vec{p}\circ\vec{\sigma}}{(E+m_0)}\phi_a}\quad\text{and}\quad\boxed{\phi_a=\frac{\vec{p}\circ\vec{\sigma}}{(E-m_0)}\phi_b}\tag{11.3.184}$$

Thus, we have:

$$\Psi_{\vec{p}}=\begin{pmatrix}\phi_a\\ \phi_b\end{pmatrix}=\begin{pmatrix}\phi_a\\ \dfrac{\vec{p}\circ\vec{\sigma}}{(E+m_0)}\phi_a\end{pmatrix}=\begin{pmatrix}\dfrac{\vec{p}\circ\vec{\sigma}}{(E-m_0)}\phi_b\\ \phi_b\end{pmatrix} \tag{11.3.185}$$

We know that solutions exist and Quantum Physics imposes that these solutions are linearly independent. Thus, let us choose the solutions for ϕ_a, ϕ_b as being proportional to:

$$\phi_a,\phi_b=\begin{pmatrix}1\\0\end{pmatrix} \quad \text{or} \quad \phi_a,\phi_b=\begin{pmatrix}0\\1\end{pmatrix} \tag{11.3.186}$$

and as (see section Spinor Calculus page 2158):

$$\vec{p}\circ\vec{\sigma}=\begin{bmatrix}p_z & p_x-\mathrm{i}p_y\\ p_x+\mathrm{i}p_y & -p_z\end{bmatrix} \tag{11.3.187}$$

We then have the following possibilities:

$$\begin{aligned}
&(1)\quad \phi_a=\begin{pmatrix}1\\0\end{pmatrix}\Rightarrow\phi_b=\frac{\vec{p}\circ\vec{\sigma}}{E+m_0}\phi_a=\frac{\vec{p}\circ\vec{\sigma}}{E+m_0}\begin{pmatrix}1\\0\end{pmatrix}=\frac{1}{E+m_0}\begin{pmatrix}p_z\\p_+\end{pmatrix}\qquad p_\pm=p_x\pm\mathrm{i}p_y\\
&(2)\quad \phi_a=\begin{pmatrix}0\\1\end{pmatrix}\Rightarrow\phi_b=\frac{\vec{p}\circ\vec{\sigma}}{E+m_0}\phi_a=\frac{\vec{p}\circ\vec{\sigma}}{E+m_0}\begin{pmatrix}0\\1\end{pmatrix}=\frac{1}{E+m_0}\begin{pmatrix}p_-\\-p_z\end{pmatrix}\\
&(3)\quad \phi_b=\begin{pmatrix}1\\0\end{pmatrix}\Rightarrow\phi_a=\frac{\vec{p}\circ\vec{\sigma}}{E-m_0}\phi_b=\frac{\vec{p}\circ\vec{\sigma}}{E-m_0}\begin{pmatrix}1\\0\end{pmatrix}=\frac{1}{E-m_0}\begin{pmatrix}p_z\\p_+\end{pmatrix}\\
&(4)\quad \phi_b=\begin{pmatrix}0\\1\end{pmatrix}\Rightarrow\phi_a=\frac{\vec{p}\circ\vec{\sigma}}{E-m_0}\phi_b=\frac{\vec{p}\circ\vec{\sigma}}{E-m_0}\begin{pmatrix}0\\1\end{pmatrix}=\frac{1}{E-m_0}\begin{pmatrix}p_-\\-p_z\end{pmatrix}
\end{aligned} \tag{11.3.188}$$

The question is now ... should we use:

$$E=+\sqrt{\|\vec{p}\|^2+m_0^2} \qquad \text{or} \qquad E=-\sqrt{\|\vec{p}\|^2+m_0^2} \tag{11.3.189}$$

Well, for (1) and (2) above we must use $E=+\sqrt{\ldots}$ otherwise $1/(E+m_0)$ becomes a singularity for $\vec{p}=\vec{0}$. For (3) and (4) we must use $E=-\sqrt{\ldots}$ otherwise $1/(E-m_0)$ becomes a singularity for $\vec{p}=\vec{0}$.

Remark

The term $E=+\sqrt{\ldots}$ is often referred to as the "particle solution" in the literature and $E=-\sqrt{\ldots}$ as the "antiparticle solution". The interpretation of this result as an antiparticle would sometimes be named as we already mentioned the "Feynman-Stückelberg inter-operation".

Going back to:

$$\Psi_{\vec{p}}=\begin{pmatrix}\phi_a\\ \phi_b\end{pmatrix} \tag{11.3.190}$$

and noting the spinors by (we change the notation to come back to the original one):

$$\chi^{(+)}=\begin{pmatrix}1\\0\end{pmatrix},\chi^{(-)}=\begin{pmatrix}0\\1\end{pmatrix} \tag{11.3.191}$$

We finally have using (1) and (2) and by denoting $N()$ the part of the solution we should normalize (see page 4972), the following possible solutions and which are independent:

$$\Psi_{\vec{p}}^{(1)}=N\begin{pmatrix}\phi_a\\ \phi_b\end{pmatrix}=N\begin{pmatrix}\begin{pmatrix}1\\0\end{pmatrix}\\ \begin{pmatrix}\dfrac{p_z}{E+m_0}\\ \dfrac{p_+}{E+m_0}\end{pmatrix}\end{pmatrix}=N\begin{pmatrix}1\\ \dfrac{\vec{p}\circ\vec{\sigma}}{E+m_0}\end{pmatrix}\chi^{(+)} \tag{11.3.192}$$

$$\Psi_{\vec{p}}^{(2)}=N\begin{pmatrix}\phi_a\\ \phi_b\end{pmatrix}=N\begin{pmatrix}\begin{pmatrix}0\\1\end{pmatrix}\\ \begin{pmatrix}\dfrac{p_-}{E+m_0}\\ \dfrac{-p_z}{E+m_0}\end{pmatrix}\end{pmatrix}=N\begin{pmatrix}1\\ \dfrac{\vec{p}\circ\vec{\sigma}}{E+m_0}\end{pmatrix}\chi^{(-)} \tag{11.3.193}$$

with $E=+\sqrt{\ldots}$ and also:

$$\Psi_{\vec{p}}^{(3)}=N\begin{pmatrix}\phi_a\\ \phi_b\end{pmatrix}=N\begin{pmatrix}\begin{pmatrix}\dfrac{p_z}{E-m_0}\\ \dfrac{p_+}{E-m_0}\end{pmatrix}\\ \begin{pmatrix}1\\0\end{pmatrix}\end{pmatrix}=N\begin{pmatrix}\dfrac{\vec{p}\circ\vec{\sigma}}{E-m_0}\\ 1\end{pmatrix}\chi^{(+)} \tag{11.3.194}$$

$$\Psi_{\vec{p}}^{(4)}=N\begin{pmatrix}\phi_a\\ \phi_b\end{pmatrix}=N\begin{pmatrix}\begin{pmatrix}\dfrac{p_-}{E-m_0}\\ \dfrac{-p_z}{E-m_0}\end{pmatrix}\\ \begin{pmatrix}0\\1\end{pmatrix}\end{pmatrix}=N\begin{pmatrix}\dfrac{\vec{p}\circ\vec{\sigma}}{E-m_0}\\ 1\end{pmatrix}\chi^{(-)} \tag{11.3.195}$$

with $E=-\sqrt{\ldots}$. It is from usage to conventionally associate the two above functions to the "positron" (the positron is used in Hospitals in PET scans - Positron Emission Topography - that is an imaging technique that can help reveal the metabolic or biochemical function of tissues and organs and hence check for signs of cancer, heart disease and brain disorders).

All this can be abbreviated globally:

$$\boxed{\Psi_{E_+}^{(\pm)}(\vec{p})=N\begin{pmatrix}1\\ \dfrac{\vec{p}\circ\vec{\sigma}}{E+m_0}\end{pmatrix}\chi^{(\pm)},\ \Psi_{E_-}^{(\pm)}(\vec{p})=N\begin{pmatrix}\dfrac{\vec{p}\circ\vec{\sigma}}{E-m_0}\\ 1\end{pmatrix}\chi^{(\pm)}} \tag{11.3.196}$$

Let us indicate that the term $\vec{p}\circ\vec{\sigma}$ is therefore the projection of the momentum (momentum) on a mathematical entity directly related to the spin is named "helicity" and defined by:

$$\boxed{\hat{h}=\vec{p}\circ\vec{\sigma}} \tag{11.3.197}$$

Antiparticles have therefore opposite charge and opposite chirality/helicity!

11.3.3.3 Free particle solution at rest

We start again from:

$$\Psi=\Psi_{\vec{p}}e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)}=\Phi_0(\vec{p},E)e^{\mathrm{i}(\vec{p}\circ\vec{r}-Et)} \tag{11.3.198}$$

Let us consider the derivatives of the free particle solution above:

$$\partial_0\Psi=\frac{\partial\Psi}{\partial t}=-\mathrm{i}E\Psi \qquad \partial_x\Psi=\frac{\partial\Psi}{\partial x}=\mathrm{i}p_x\Psi \qquad \ldots \tag{11.3.199}$$

Substituting these into the Dirac equation gives:

$$(\gamma^0E-\gamma^1p_x-\gamma^2p_y-\gamma_3p_z-m)\Psi_0=\vec{0} \tag{11.3.200}$$

which can be written:

$$(\gamma^\mu p_\mu-m)\Psi_0=\vec{0} \tag{11.3.201}$$

This is the Dirac equation in "momentum" – note it contains no derivatives.

For a particle at rest $\vec{p}=\vec{0}$ then:

$$\Psi=\Psi_0(E,\vec{0})e^{-\mathrm{i}Et} \tag{11.3.202}$$

Hence the Dirac equation reduces to:

$$E\gamma^0\Psi_0-m\Psi_0=\vec{0} \tag{11.3.203}$$

Explicitly:

$$E\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&-1&0\\0&0&0&-1\end{bmatrix}\begin{pmatrix}\phi_1\\\phi_2\\\phi_3\\\phi_4\end{pmatrix}=m\begin{pmatrix}\phi_1\\\phi_2\\\phi_3\\\phi_4\end{pmatrix} \tag{11.3.204}$$

We see quite immediately that this equation has for independent orthogonal solutions:

$$\Psi_{0,1}(m,0)=\begin{pmatrix}1\\0\\0\\0\end{pmatrix},\Psi_{0,2}(m,0)=\begin{pmatrix}0\\1\\0\\0\end{pmatrix},\Psi_{0,3}(m,0)=\begin{pmatrix}0\\0\\1\\0\end{pmatrix},\Psi_{0,4}(m,0)=\begin{pmatrix}0\\0\\0\\1\end{pmatrix} \tag{11.3.205}$$

The two first solutions leads us to:

$$\left.\begin{aligned}\Psi_{0,1}(m,0)&=\begin{pmatrix}1\\0\\0\\0\end{pmatrix}\\\Psi_{0,2}(m,0)&=\begin{pmatrix}0\\1\\0\\0\end{pmatrix}\end{aligned}\right\}\Rightarrow E=m \tag{11.3.206}$$

And the last two lead us to:

$$\left.\begin{aligned}\Psi_{0,3}(m,0)&=\begin{pmatrix}0\\0\\1\\0\end{pmatrix}\\ \Psi_{0,4}(m,0)&=\begin{pmatrix}0\\0\\0\\1\end{pmatrix}\end{aligned}\right\}\Rightarrow E=-m \tag{11.3.207}$$

So we still have negative energy solutions (antiparticles)!

Including the time dependence from $\Psi=\Psi_0(E,\vec{0})e^{-\mathrm{i}Et}$ gives:

$$\Psi_1=\begin{pmatrix}1\\0\\0\\0\end{pmatrix}e^{-\mathrm{i}mt},\Psi_2=\begin{pmatrix}0\\1\\0\\0\end{pmatrix}e^{-\mathrm{i}mt},\Psi_3=\begin{pmatrix}0\\0\\1\\0\end{pmatrix}e^{+\mathrm{i}mt},\Psi_4=\begin{pmatrix}0\\0\\0\\1\end{pmatrix}e^{+\mathrm{i}mt} \tag{11.3.208}$$

We therefore have two spin states with $E>0$ and two spin states with $E<0$. More explicitly, Φ_1 then describes an $S=\frac{1}{2}$ (in Planck units) fermion of mass m with spin ↑, Φ_2 describes an $S=\frac{1}{2}$ (in Planck units) fermion of mass m with spin ↓, Φ_3 describes an $S=\frac{1}{2}$ (in Planck units) antifermion of mass m with spin ↑, Φ_4 describes an $S=\frac{1}{2}$ (in Planck units) antifermion of mass m with spin ↓.

In quantum physics we can't just discard the $E<0$ solutions as unphysical as we require a complete set of states, i.e. four solutions!

We can also see something important! The antiparticle solution is the complex conjugate of the particle solution! Keep this in mind for when we will study the Majorana equation later!

For information the reader can check that if we take back the relations proved earlier:

$$\Psi_{\vec{p}}^{(1)}=N\begin{pmatrix}\phi_a\\ \phi_b\end{pmatrix}=N\begin{pmatrix}\begin{pmatrix}1\\0\end{pmatrix}\\ \begin{pmatrix}\dfrac{p_z}{E+m_0}\\ \dfrac{p_+}{E+m_0}\end{pmatrix}\end{pmatrix}=N\begin{pmatrix}1\\ \dfrac{\vec{p}\circ\vec{\sigma}}{E+m_0}\end{pmatrix}\chi^{(+)} \tag{11.3.209}$$

$$\Psi_{\vec{p}}^{(2)}=N\begin{pmatrix}\phi_a\\ \phi_b\end{pmatrix}=N\begin{pmatrix}\begin{pmatrix}0\\1\end{pmatrix}\\ \begin{pmatrix}\dfrac{p_-}{E+m_0}\\ \dfrac{-p_z}{E+m_0}\end{pmatrix}\end{pmatrix}=N\begin{pmatrix}1\\ \dfrac{\vec{p}\circ\vec{\sigma}}{E+m_0}\end{pmatrix}\chi^{(-)} \tag{11.3.210}$$

$$\Psi_{\vec{p}}^{(3)}=N\begin{pmatrix}\phi_a\\ \phi_b\end{pmatrix}=N\begin{pmatrix}\begin{pmatrix}\dfrac{p_z}{E-m_0}\\ \dfrac{p_+}{E-m_0}\end{pmatrix}\\ \begin{pmatrix}1\\0\end{pmatrix}\end{pmatrix}=N\begin{pmatrix}\dfrac{\vec{p}\circ\vec{\sigma}}{E-m_0}\\ 1\end{pmatrix}\chi^{(+)} \tag{11.3.211}$$

$$\Psi_{\bar{p}}^{(4)}=N\begin{pmatrix}\phi_a\\ \phi_b\end{pmatrix}=N\begin{pmatrix}\begin{pmatrix}\dfrac{p_-}{E-m_0}\\ \dfrac{-p_z}{E-m_0}\end{pmatrix}\\ \begin{pmatrix}0\\ 1\end{pmatrix}\end{pmatrix}=N\begin{pmatrix}\dfrac{\vec{p}\circ\vec{\sigma}}{E-m_0}\\ 1\end{pmatrix}\chi^{(-)} \tag{11.3.212}$$

and we put $\vec{p}=\vec{0}$ we fall back on the same type of spinors as we just get above!

11.3.3.4 Majorana equation

An electrically charged particle is different from its antiparticles as it has the opposite electric charge, and electric charge is a measurable, stable property. It is possible, however, for an electrically neutral particle to be its own antiparticle. Photons, which have spin 1 (in units of the rationalized Planck's constant $\hbar$) are a familiar case; neutral ions (spin 0) are a further example, and gravitons (spin 2) another speculated one! Particles that are their own antiparticles must be created by field Ψ that obey:

$$\Psi=\Psi^{\dagger} \tag{11.3.213}$$

that is... real fields (!), because the complex-conjugate field $\Psi^{\dagger}$ create their antiparticles.

The equations for particles with spin 0, spin 1 and spin 2 - the Klein-Gordon, Maxwell (electromagnetism) and Einstein field equations (General Relativity) respectively - readily accommodate real fields, as these equations are formulated using real numbers.

Indeed, when in 11928 (holocene calendar), Paul Dirac discovered the theoretical framework for describing spin-$1/2$ particles, it seemed that complex numbers were unavoidable. Dirac's original equation contained both real and imaginary numbers, and therefore it can only pertain to complex fields. For Dirac, who was concerned with describing electrons, this feature posed no problem, and even came to seem an advantage because it explained why positrons, the antiparticles of electrons, exist.

Enters Ettore Majorana! In his 11937 (holocene calendar) paper, Majorana posed, and answered, the question of whether equations for spin-$1/2$ fields must necessarily, like Dirac's original equation, involve complex numbers.

Considerations of mathematical elegance and symmetry both motivated and guided his investigation. Majorana discovered that, to the contrary, there is a simple, clever modification of Dirac's equation that involves only real numbers. With this discovery, Majorana made the idea that spin-$1/2$ particles could be their own antiparticles theoretically respectable, that is, consistent with the general principles of relativity and quantum theory.

Majorana inquired whether it might be possible for a spin-$1/2$ particle to be its own antiparticle, by attempting to find the equation that such an object would satisfy. To get an equation of Dirac's type (that is, suitable for spin-$1/2$) but capable of governing a real field, required γ matrices that are purely imaginary!

These matrices that we have already introduced earlier and named the "Majorana representation" are defined by:

$$\gamma_M^0=\beta_M=\begin{bmatrix}0&0&0&-\mathrm{i}\\0&0&-\mathrm{i}&0\\0&\mathrm{i}&0&0\\\mathrm{i}&0&0&0\end{bmatrix},\gamma_M^1=\begin{bmatrix}0&-\mathrm{i}&0&0\\-\mathrm{i}&0&0&0\\0&0&0&\mathrm{i}\\0&0&\mathrm{i}&0\end{bmatrix},\gamma_M^2=\begin{bmatrix}-\mathrm{i}&0&0&0\\0&\mathrm{i}&0&0\\0&0&-\mathrm{i}&0\\0&0&0&\mathrm{i}\end{bmatrix},\gamma_M^3=\begin{bmatrix}0&0&0&\mathrm{i}\\0&0&\mathrm{i}&0\\0&\mathrm{i}&0&0\\\mathrm{i}&0&0&0\end{bmatrix}\tag{11.3.214}$$

And we know that the Majorana matrices in the representation above obey also the following mandatory gamma anticommutation relations (as so far we don't know the β and α_i corresponding Majorana matrices for recall!) as seen at page 4963:

$$[\gamma_i,\gamma_j]_+=\gamma_i\gamma_j+\gamma_j\gamma_i=2\eta_{ij}\mathbb{1}_4\tag{11.3.215}$$

Thus, for $i\neq j$:

$$[\gamma_i,\gamma_j]_+=\gamma_i\gamma_j+\gamma_j\gamma_i=0\tag{11.3.216}$$

and it is also immediate using this anticommutation relation that:

$$(\gamma^0)^2=1,(\gamma^i)^2=-1\tag{11.3.217}$$

Another purely imaginary representation of the Majorana matrices quite common in the literature is also as we know given by:

$$\begin{gathered}\tilde{\gamma}_M^0=\begin{bmatrix}0&0&0&-\mathrm{i}\\0&0&\mathrm{i}&0\\0&-\mathrm{i}&0&0\\\mathrm{i}&0&0&0\end{bmatrix}=\begin{bmatrix}0&\sigma_2\\\sigma_2&0\end{bmatrix},\tilde{\gamma}_M^1=\begin{bmatrix}\mathrm{i}&0&0&0\\0&-\mathrm{i}&0&0\\0&0&\mathrm{i}&0\\0&0&0&-\mathrm{i}\end{bmatrix}=\mathrm{i}\begin{bmatrix}\sigma_3&0\\0&\sigma_3\end{bmatrix}\\\tilde{\gamma}_M^2=\begin{bmatrix}0&0&0&\mathrm{i}\\0&0&-\mathrm{i}&0\\0&-\mathrm{i}&0&0\\\mathrm{i}&0&0&0\end{bmatrix}=\begin{bmatrix}0&-\sigma_2\\\sigma_2&0\end{bmatrix},\tilde{\gamma}_M^3=\begin{bmatrix}0&-\mathrm{i}&0&0\\-\mathrm{i}&0&0&0\\0&0&0&-\mathrm{i}\\0&0&-\mathrm{i}&0\end{bmatrix}=-\mathrm{i}\begin{bmatrix}\sigma_1&0\\0&\sigma_1\end{bmatrix}\end{gathered}\tag{11.3.218}$$

That latter Majorana representation can be obtained from the original representation by the unitary transformation:

$$\tilde{\gamma}_M^\mu=\tilde{U}\gamma_M^\mu\tilde{U}^{-1}\quad\text{and}\quad\tilde{U}=\frac{1}{2}=\begin{bmatrix}1&1&1&-1\\-1&1&1&1\\1&1&-1&1\\1&-1&1&1\end{bmatrix}\tag{11.3.219}$$

There is alsoooooo again as we already know another Majorana representation also denoted identically as previously (sic!) and defined by:

$$\tilde{\gamma}_M^0=\sigma_2\otimes\sigma_1,\tilde{\gamma}_M^1=\mathrm{i}\sigma_1\otimes\mathbb{1},\tilde{\gamma}_M^2=\mathrm{i}\sigma_3\otimes\mathbb{1},\tilde{\gamma}_M^3=\mathrm{i}\sigma_2\otimes\sigma_2\tag{11.3.220}$$

That is explicitly:

$$\tilde{\gamma}_M^0=\begin{bmatrix}0&0&0&-\mathrm{i}\\0&0&-\mathrm{i}&0\\0&\mathrm{i}&0&0\\\mathrm{i}&0&0&0\end{bmatrix},\tilde{\gamma}_M^1=\begin{bmatrix}0&0&\mathrm{i}&0\\0&0&0&\mathrm{i}\\\mathrm{i}&0&0&0\\0&\mathrm{i}&0&0\end{bmatrix},\tilde{\gamma}_M^2=\begin{bmatrix}\mathrm{i}&0&0&0\\0&\mathrm{i}&0&0\\0&0&-\mathrm{i}&0\\0&0&0&-\mathrm{i}\end{bmatrix},\tilde{\gamma}_M^3=\begin{bmatrix}0&0&0&-\mathrm{i}\\0&0&\mathrm{i}&0\\0&\mathrm{i}&0&0\\-\mathrm{i}&0&0&0\end{bmatrix}\tag{11.3.221}$$

This representation has the interesting property that all matrices γ^μ are pure imaginary what make the Dirac fields of the Dirac equation a non-complex field where the particles are therefore their own antiparticles.

Indeed, the "Majorana's equation", then, is simply:

$$\boxed{(\mathrm{i}\tilde{\gamma}^\mu\partial_\mu - m_0)\Psi^\dagger = \vec{0}} \tag{11.3.222}$$

Because the $\tilde{\gamma}^\mu$ matrices are purely imaginary, the matrices $\mathrm{i}\tilde{\gamma}^\mu$ are real, and consequently this equation can govern a real field $\tilde{\Psi}$.

Particles corresponding to Majorana spinors are known as "Majorana particles", due to the above self-conjugacy constraint. All the fermions included in the Standard Model have been excluded as Majorana fermions (since they have non-zero electric charge they cannot be antiparticles of themselves) with the exception of the neutrino (which is neutral).

Theoretically, the neutrino is a possible exception to this pattern. If so, neutrinoless double-beta decay, as well as a range of lepton-number violating meson and charged lepton decays, are possible. A number of experiments probing whether the neutrino is a Majorana particle are currently (year 12017 according to holocene calendar) underway.

11.3.4 Linearised Dirac Equation

We saw at the beginning of our study of Wave Quantum Physics that the classic Schrödinger equation of evolution was:

$$\mathrm{i}\hbar\partial_t\Psi = \left(E_\mathrm{p} - \frac{\hbar}{2m}\Delta\right)\Psi \tag{11.3.223}$$

ie a differential equation of the first order with respect to time and the second with respect to the spatial coordinates.

We had then determined the relativistic Schrödinger equation of evolution (free Klein-Gordon equation) given by:

$$\hbar^2\partial_t^2\Psi = \left(\hbar^2c^2\Delta - m_0^2c^4\Delta\right)\Psi \tag{11.3.224}$$

We notice that by passing to a relativistic form, we now have a differential equation of the second order in time AND space.

Then through the generalized Klein-Gordon equation which also contained a second-order differential equation in time and space:

$$\left[(\mathrm{i}\hbar\partial_t - qU)^2 - (-\mathrm{i}\hbar\vec{\nabla} - q\vec{A})^2c^2\right]\Psi = m_0^2c^4\Psi \tag{11.3.225}$$

and with the free Dirac equation we have just proved that we get in the same way a matrix differential equation of first order in time and second order in space:

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\begin{bmatrix}\Psi_a\\ \Psi_b\end{bmatrix} = \begin{bmatrix} m_0c^2 - \dfrac{\hbar^2}{2m_0}\Delta & -\dfrac{\hbar^2}{2m_0}\Delta \\ \dfrac{\hbar^2}{2m_0}\Delta & -m_0c^2 + \dfrac{\hbar^2}{2m_0}\Delta \end{bmatrix}\begin{bmatrix}\Psi_a\\ \Psi_b\end{bmatrix} \tag{11.3.226}$$

These changes of order of the differentials of a relativistic model or not impose of course in the case of a first order to know the initial conditions in time and space of the wave equation, which is feasible. However, when a second order appears, it is necessary to know more about the initial conditions of the derivatives of the wave functions (see section Differential and Integral Calculus page 1605). Moreover, even if mathematically the (relative) rigour naturally led us to the different orders obtained, it is strange that changing to a relativist model that we change order of the differentials. Why?: For the quite simple reason that by approximating the relativistic equations we are not able because of the presence of the Planck constant factor to make approximations (development in series of v/c) that would bring us back to the first order. The relativistic and non-relativistic equations seems then a priori incompatible within the non relativistic limits!

The Dirac method to solve this problem will have been the following:

The orders of the Klein-Gordon differential equation coming from the relation of the total energy (see the beginnings of our developments of the free Klein-Gordon equation) in the absence of any field:

$$\begin{aligned}E^2 &= m_0^2c^4 + c^2\|\vec{p}\|^2 = m_0^2c^4 + c^2p_x^2 + c^2p_y^2 + c^2p_z^2 = m_0^2c^4 + c^2p_1^2 + c^2p_2^2 + c^2p_3^2\\ &= m_0^2c^4 + c^2(p_1^2 + p_2^2 + p_3^2) = H^2\end{aligned} \tag{11.3.227}$$

It seems that Dirac would have had the brilliant idea of linearising this Hamiltonian by putting (sadly we don't know how he get this idea):

$$\boxed{H_D = \beta m_0 c^2 + c\alpha_1 p_1 + c\alpha_2 p_2 + c\alpha_3 p_3 = \beta m_0 c^2 + c\vec{\alpha} \circ \vec{p}} \tag{11.3.228}$$

which we will have to determine the parameters β, $\alpha(\alpha_1, \alpha_2, \alpha_3)$ which can be scalars, vectors or matrices (wait a little ... the answer will come). It follows that the Hamiltonian is also a scalar, a vector, or a matrix.

Thus, the simplest relativistic evolution wave equation that we can construct will be:

$$\mathrm{i}\hbar\partial_t \Psi = H\Psi \Rightarrow \mathrm{i}\hbar\partial_t \Psi = H_D \Psi \Rightarrow \mathrm{i}\hbar\partial_t \Psi = (\beta m_0 c^2 + c\vec{\alpha} \circ \vec{p})\Psi \tag{11.3.229}$$

In a much more common form in the literature:

$$\boxed{(\mathrm{i}\hbar\partial_t - \beta m_0 c^2 - c\vec{\alpha} \circ \vec{p})\Psi = 0} \tag{11.3.230}$$

Or also:

$$\boxed{(c\vec{\alpha} \circ \vec{p} + \beta m_0 c^2)\Psi = \mathrm{i}\hbar\partial_t \Psi} \tag{11.3.231}$$

Or in natural units:

$$\boxed{(\vec{\alpha} \circ \vec{p} + \beta m_0)\Psi = \mathrm{i}\hbar\partial_t \Psi} \tag{11.3.232}$$

As here:

$$\vec{p} = -\mathrm{i}\hbar\vec{\nabla} \tag{11.3.233}$$

we find then the prior-previous relation also in the form:

$$\boxed{(\mathrm{i}\hbar\partial_t - \mathrm{i}\hbar c\vec{\alpha} \circ \vec{\nabla} + \beta m_0 c^2)\Psi = 0} \tag{11.3.234}$$

So now we have complete the full path of relationships visible in the figure below:

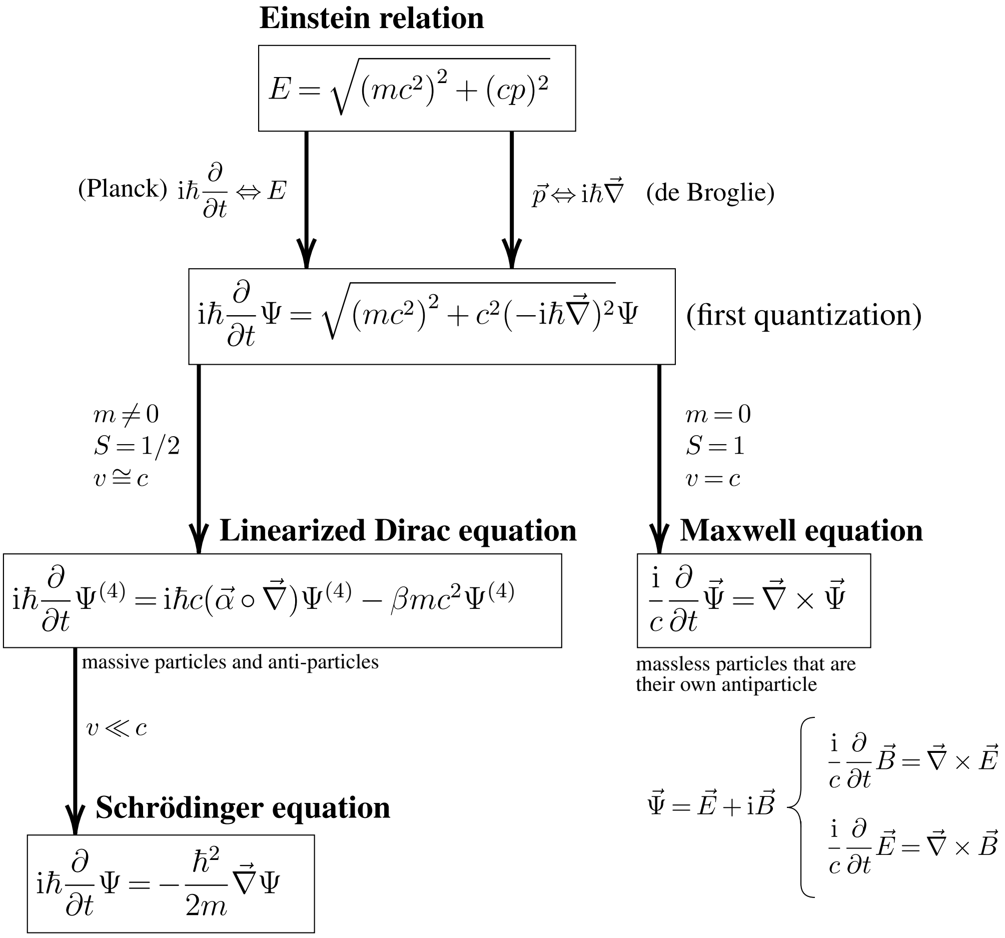

Einstein relation
$E=\sqrt{(mc^2)^2+(cp)^2}$
(Planck) $\mathrm{i}\hbar\frac{\partial}{\partial t}\Leftrightarrow E$
$\vec{p}\Leftrightarrow \mathrm{i}\hbar\vec{\nabla}$ (de Broglie)
$\mathrm{i}\hbar\frac{\partial}{\partial t}\Psi=\sqrt{(mc^2)^2+c^2(-\mathrm{i}\hbar\vec{\nabla})^2}\Psi$
(first quantization)
$m\neq 0$
$S=1/2$
$v\cong c$
$m=0$
$S=1$
$v=c$
Linearized Dirac equation
$\mathrm{i}\hbar\frac{\partial}{\partial t}\Psi^{(4)}=\mathrm{i}\hbar c(\vec{\alpha}\circ\vec{\nabla})\Psi^{(4)}-\beta mc^2\Psi^{(4)}$
massive particles and anti-particles
$v\ll c$
Maxwell equation
$\frac{\mathrm{i}}{c}\frac{\partial}{\partial t}\vec{\Psi}=\vec{\nabla}\times\vec{\Psi}$
massless particles that are their own antiparticle
$\vec{\Psi}=\vec{E}+\mathrm{i}\vec{B}$
$\frac{\mathrm{i}}{c}\frac{\partial}{\partial t}\vec{B}=\vec{\nabla}\times\vec{E}$
$\frac{\mathrm{i}}{c}\frac{\partial}{\partial t}\vec{E}=\vec{\nabla}\times\vec{B}$
Schrödinger equation
$\mathrm{i}\hbar\frac{\partial}{\partial t}\Psi=-\frac{\hbar^2}{2m}\vec{\nabla}\Psi$

Figure 11.83: **Relationships between most famous quantum physics equations**

If the linear momentum were to be zero, we would thus fall back on the energy at rest for the Hamiltonian:

$$\mathrm{i}\hbar\partial_t\Psi=\beta m_0c^2\Psi \tag{11.3.235}$$

where as we shall see later $\beta=\pm 1$.

The validity of this linearisation must be verified by falling back on the results obtained during our previous study of the Dirac equation.

Let us now raise the operator to the square:

$$H_D^2=(\beta m_0c^2+c\alpha_1p_1+c\alpha_2p_2+c\alpha_3p_3)^2 \tag{11.3.236}$$

and let us put:

$$A=(\beta m_0c^2+c\alpha_1p_1)\quad\text{and}\quad B=(c\alpha_2p_2+c\alpha_3p_3) \tag{11.3.237}$$

At this point, it is important to notice that we may be working with operators (typically matrices) that might not commute because the α, β are unknown. Therefore, the squaring will be done as follows:

$$H_D^2=(A+B)^2=A^2+B^2+AB+BA \tag{11.3.238}$$

Thus simply developed but without simplifying the sum $AB+BA$ into a $2AB$ or a $2BA$ since we are not sure for now if they will be commutativity or not.

We therefore develop the Dirac Hamiltonian:

$$\begin{aligned}H_D^2=&(\beta^2m_0^2c^4+c^2\alpha_1^2p_1^2+\beta m_0c^2\cdot c\alpha_1p_1+c\alpha_1p_1\beta m_0c^2)\\&+(c^2\alpha_2^2p_2^2+c^2\alpha_3^2p_3^2+c^2\alpha_2p_2\cdot c^2\alpha_3p_3+c^2\alpha_3p_3\cdot c^2\alpha_2p_2)\\&+(\beta m_0c^2+c\alpha_1p_1)(c\alpha_2p_2+c\alpha_3p_3)+(c\alpha_2p_2+c\alpha_3p_3)(\beta m_0c^2+c\alpha_1p_1)\end{aligned}\tag{11.3.239}$$

By making the products of the terms in brackets and respecting the order of the operators, it comes:

$$\begin{aligned}H_D^2=&\underbrace{\beta^2m_0^2c^4}_{1}+\underbrace{c^2\alpha_1^2p_1^2}_{2}+\underbrace{m_0c^3\beta\alpha_1p_1}_{11}+\underbrace{m_0c^3\alpha_1p_1\beta}_{12}+\underbrace{c^2\alpha_2^2p_2^2}_{3}+\underbrace{c^2\alpha_3^2p_3^2}_{4}\\&+\underbrace{c^2\alpha_2p_2\alpha_3p_3}_{9}+\underbrace{c^2\alpha_3p_3\alpha_2p_2}_{10}+\underbrace{m_0c^3\beta\alpha_2p_2}_{13}+\underbrace{m_0c^3\beta\alpha_3p_3}_{15}+\underbrace{c^2\alpha_1p_1\alpha_2p_2}_{5}\\&+\underbrace{c^2\alpha_1p_1\alpha_3p_3}_{7}+\underbrace{m_0c^3\alpha_2p_2\beta}_{14}+\underbrace{c^2\alpha_2p_2\alpha_1p_1}_{6}+\underbrace{m_0c^3\alpha_3p_3\beta}_{16}+\underbrace{c^2\alpha_3p_3\alpha_1p_1}_{8}\end{aligned}\tag{11.3.240}$$

By grouping certain terms:

$$\begin{aligned}H_D^2=&\underbrace{\beta^2m_0^2c^4}_{1}+c^2\left(\underbrace{\alpha_1^2p_1^2}_{2}+\underbrace{\alpha_2^2p_2^2}_{3}+\underbrace{\alpha_3^2p_3^2}_{4}\right)+c^2\left(\underbrace{\alpha_1\alpha_2}_{5}+\underbrace{\alpha_2\alpha_1}_{6}\right)p_1p_2\\&+c^2\left(\underbrace{\alpha_1\alpha_3}_{7}+\underbrace{\alpha_3\alpha_1}_{8}\right)p_1p_3+c^2\left(\underbrace{\alpha_2\alpha_3}_{9}+\underbrace{\alpha_3\alpha_2}_{10}\right)p_2p_3+m_0c^3\left(\underbrace{\beta\alpha_1}_{11}+\underbrace{\alpha_1\beta}_{12}\right)p_1+\\&+m_0c^3\left(\underbrace{\beta\alpha_2}_{13}\ \underbrace{\alpha_2\beta}_{14}\right)p_2+m_0c^3\left(\underbrace{\beta\alpha_3}_{15}+\underbrace{\alpha_3\beta}_{16}\right)p_3\end{aligned}\tag{11.3.241}$$

To be consistent with our linearisation assumptions, we must have:

$$\begin{aligned}&1\rightarrow\beta^2m_0^2c^4\equiv m_0^2c^4\Rightarrow\beta^2=1\\&2,3,4\rightarrow c^2(\alpha_1^2p_1^2+\alpha_2^2p_2^2+\alpha_3^2p_3^2)\equiv c^2(p_1^2+p_2^2+p_3^2)\Rightarrow\alpha_1^2=\alpha_2^2=\alpha_3^2=1\\&5,6\rightarrow c^2(\alpha_1\alpha_2+\alpha_2\alpha_1)p_1p_2=0\Rightarrow\alpha_1\alpha_2+\alpha_2\alpha_1=0\\&7,8\rightarrow c^2(\alpha_1\alpha_3+\alpha_3\alpha_1)p_1p_3=0\Rightarrow\alpha_1\alpha_3+\alpha_3\alpha_1=0\\&9,10\rightarrow c^2(\alpha_2\alpha_3+\alpha_3\alpha_2)p_2p_3=0\Rightarrow\alpha_2\alpha_3+\alpha_3\alpha_2=0\\&11,12\rightarrow m_0c^3(\beta\alpha_1+\alpha_1\beta)p_1=0\Rightarrow\beta\alpha_1+\alpha_1\beta=0\\&13,14\rightarrow m_0c^3(\beta\alpha_2+\alpha_2\beta)p_2=0\Rightarrow\beta\alpha_2+\alpha_2\beta=0\\&15,16\rightarrow m_0c^3(\beta\alpha_3+\alpha_3\beta)p_3=0\Rightarrow\beta\alpha_3+\alpha_3\beta=0\end{aligned}\tag{11.3.242}$$

and we will see immediately that to satisfy these conditions the α, β will have to be 4×4 matrices.

Written in the form of commutators, we therefore have the following three conditions to satisfy:

$$\boxed{\begin{aligned}&\beta^2=\alpha_1^2=\alpha_2^2=\alpha_3^2=1\\&[\alpha_1,\alpha_2]_+=[\alpha_1,\alpha_3]_+=[\alpha_2,\alpha_3]_+=0\\&[\alpha_1,\beta]_+=[\alpha_2,\beta]_+=[\alpha_3,\beta]_+=0\end{aligned}}\tag{11.3.243}$$

We observe the following:

- The square of each operator β and α_j is equal to 1 (or to the unit matrix if it is a matrix...).
- $[\alpha_i,\alpha_j]_+ = 0, \forall i,j = \{1,2,3\}, i \neq j$ is an anticommutator
- $[\alpha_i,\beta]_+ = 0$ is an anticommutator

These three relations can be summarized as follows:

$$\beta^2 = 1 \qquad [\alpha_i,\alpha_j]_+ = 2\delta_{ij} \qquad [\alpha_i,\beta]_+ = 0 \tag{11.3.244}$$

At this point, we need to look for mathematical objects that meet the three conditions above. We could show that a square matrix of dimension 2 or 3 does not satisfy the three conditions and a scalar even less!

Dirac then adopted by analogy with the previous developments, square matrices of dimension 4 including Pauli matrices and admitted for β a unit matrix (this choice made by Dirac is particular, there are others possible choices hence the necessity to confront the models with experimental observations!).

So what we denoted "1" before is actually a square unit matrix of dimension 4! And so the wave functions must have 4 components and mathematical beings that have specificity are the spinors!

The matrices considered by Dirac (which are a particular choice!) are therefore for α_i, β:

$$\alpha_i = \begin{bmatrix} 0 & \sigma_i \\ \sigma_i & 0 \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} \mathbb{1} & 0 \\ 0 & -\mathbb{1} \end{bmatrix} \tag{11.3.245}$$

and for example in certain domains of quantum physics we use Weyl's choice:

$$\alpha_i = \begin{bmatrix} \sigma_i & 0 \\ 0 & -\sigma_i \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} 0 & \mathbb{1} \\ \mathbb{1} & 0 \end{bmatrix} \tag{11.3.246}$$

In the choice considered by Dirac or Weyl, we have the following Pauli matrices and unit matrix:

$$\sigma_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \sigma_2 = \begin{bmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{bmatrix} \quad \sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad \mathbb{1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{11.3.247}$$

This leads to the matrices α_i, β (we recognize that the α_i are in fact the gamma matrices γ_i in their Chiral representation as introduced earlier above at the difference of a $-$ sign on one component but that anyway leads to exactly the same results!!!!!):

$$\alpha_1 = \begin{bmatrix} 0 & \sigma_1 \\ \sigma_1 & 0 \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \\ \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \end{bmatrix} \quad \alpha_2 = \begin{bmatrix} 0 & \sigma_2 \\ \sigma_2 & 0 \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} & \begin{bmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{bmatrix} \\ \begin{bmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \end{bmatrix}$$

$$\alpha_3 = \begin{bmatrix} 0 & \sigma_3 \\ \sigma_3 & 0 \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} & \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \\ \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \end{bmatrix} \quad \beta = \begin{bmatrix} \mathbb{1} & 0 \\ 0 & -\mathbb{1} \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \\ \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} & -\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{bmatrix} = \begin{bmatrix} \mathbb{1} & 0 \\ 0 & -\mathbb{1} \end{bmatrix} \tag{11.3.248}$$

Remark

Notice that all matrices above are Hermitians!

We can check that the linearisation conditions are satisfied by the preceding matrices:

- First condition:

$$\beta^2=1\rightarrow\left\{\begin{array}{l}\beta_+\cdot\beta_+=\begin{bmatrix}\mathbb{1}&0\\0&-\mathbb{1}\end{bmatrix}\begin{bmatrix}\mathbb{1}&0\\0&-\mathbb{1}\end{bmatrix}=\begin{bmatrix}\mathbb{1}&0\\0&\mathbb{1}\end{bmatrix}=1\\ \beta_-\cdot\beta_-=-\begin{bmatrix}\mathbb{1}&0\\0&-\mathbb{1}\end{bmatrix}\left\{-\begin{bmatrix}\mathbb{1}&0\\0&-\mathbb{1}\end{bmatrix}\right\}=\begin{bmatrix}\mathbb{1}&0\\0&\mathbb{1}\end{bmatrix}=1\end{array}\right\}\tag{11.3.249}$$

  Similarly for the α_j:

$$\begin{gathered}\alpha_1^2=\begin{bmatrix}0&\sigma_1\\\sigma_1&0\end{bmatrix}^2=\begin{bmatrix}0&0&0&1\\0&0&1&0\\0&1&0&0\\1&0&0&0\end{bmatrix}\begin{bmatrix}0&0&0&1\\0&0&1&0\\0&1&0&0\\1&0&0&0\end{bmatrix}=\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix}=1\\ \alpha_2^2=\begin{bmatrix}0&\sigma_2\\\sigma_2&0\end{bmatrix}^2=\begin{bmatrix}0&0&0&-\mathrm{i}\\0&0&\mathrm{i}&0\\0&-\mathrm{i}&0&0\\\mathrm{i}&0&0&0\end{bmatrix}\begin{bmatrix}0&0&0&-\mathrm{i}\\0&0&\mathrm{i}&0\\0&-\mathrm{i}&0&0\\\mathrm{i}&0&0&0\end{bmatrix}=\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix}=1\\ \alpha_3^2=\begin{bmatrix}0&\sigma_3\\\sigma_3&0\end{bmatrix}^2=\begin{bmatrix}0&0&1&0\\0&0&0&-1\\1&0&0&0\\0&-1&0&0\end{bmatrix}\begin{bmatrix}0&0&1&0\\0&0&0&-1\\1&0&0&0\\0&-1&0&0\end{bmatrix}=\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix}=1\end{gathered}\tag{11.3.250}$$

  The first condition is therefore fulfilled!

- Second condition (caution with notations that sadly slip a bit by tradition between matrices and scalars!):

$$\begin{aligned}&[\alpha_i,\alpha_j]_+=\begin{bmatrix}0&\sigma_i\\\sigma_i&0\end{bmatrix}\begin{bmatrix}0&\sigma_j\\\sigma_j&0\end{bmatrix}+\begin{bmatrix}0&\sigma_j\\\sigma_j&0\end{bmatrix}\begin{bmatrix}0&\sigma_i\\\sigma_i&0\end{bmatrix}=\\&\begin{bmatrix}\sigma_i\sigma_j&0\\0&\sigma_i\sigma_j\end{bmatrix}+\begin{bmatrix}\sigma_j\sigma_i&0\\0&\sigma_j\sigma_i\end{bmatrix}=\begin{bmatrix}\sigma_i\sigma_j+\sigma_j\sigma_i&0\\0&\sigma_i\sigma_j+\sigma_j\sigma_i\end{bmatrix}\\&=(\sigma_i\sigma_j+\sigma_j\sigma_i)\begin{bmatrix}1&0\\0&1\end{bmatrix}=\sigma_i\sigma_j+\sigma_j\sigma_i\\&\text{If}\quad i=j\rightarrow 2(\sigma_i\sigma_i)=2(\sigma_i)^2\\&\text{If}\quad i=1\rightarrow 2(\sigma_1)^2=2\begin{bmatrix}0&1\\1&0\end{bmatrix}^2=2\\&\text{If}\quad i=2\rightarrow 2(\sigma_2)^2=2\begin{bmatrix}0&-\mathrm{i}\\\mathrm{i}&0\end{bmatrix}^2=2\\&\text{If}\quad i=3\rightarrow 2(\sigma_3)^2=2\begin{bmatrix}1&0\\0&-1\end{bmatrix}^2=2\end{aligned}\tag{11.3.251}$$

and:

If $i=1$ and $j=2$

$$\sigma_1\sigma_2+\sigma_2\sigma_1=\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}0 & -\mathrm{i}\\ \mathrm{i} & 0\end{bmatrix}+\begin{bmatrix}0 & -\mathrm{i}\\ \mathrm{i} & 0\end{bmatrix}\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}=\begin{bmatrix}\mathrm{i} & 0\\ 0 & -\mathrm{i}\end{bmatrix}+\begin{bmatrix}-\mathrm{i} & 0\\ 0 & \mathrm{i}\end{bmatrix}=0$$

If $i=1$ and $j=3$

$$\sigma_1\sigma_3+\sigma_3\sigma_1=\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}+\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}=\begin{bmatrix}0 & -1\\ 1 & 0\end{bmatrix}+\begin{bmatrix}0 & 1\\ -1 & 0\end{bmatrix}=0 \tag{11.3.252}$$

Therefore (caution! this is in fact a 4×4 matrix!!):

$$[\alpha_i,\alpha_j]_+=\sigma_i\sigma_j+\sigma_j\sigma_i=2\delta_{ij} \tag{11.3.253}$$

The second condition is well fulfilled!

Remark

The last relation should be written:

$$[\alpha_i,\alpha_j]_+=\alpha_i\alpha_j+\alpha_j\alpha_i=2\delta_{ij}\mathbb{1}_4 \tag{11.3.254}$$

and if we take the gamma matrices with the Dirac or any other representation introduce earlier above, the same is satisfied but generalized if we include γ_0 to:

$$\boxed{[\gamma_i,\gamma_j]_+=\gamma_i\gamma_j+\gamma_j\gamma_i=2\eta_{ij}\mathbb{1}_4} \tag{11.3.255}$$

where η_{ij} the Minkowski metric with signature $(+,-,-,-)$. This relation is important because it is sometimes used as the definition of the gamma matrices (!!!) and it makes appears the metric!!!

From the above relation we have immediately for $i\neq j$:

$$[\gamma_i,\gamma_j]_+=\gamma_i\gamma_j+\gamma_j\gamma_i=0 \tag{11.3.256}$$

and it is also immediate using the latter anticommutation relation that:

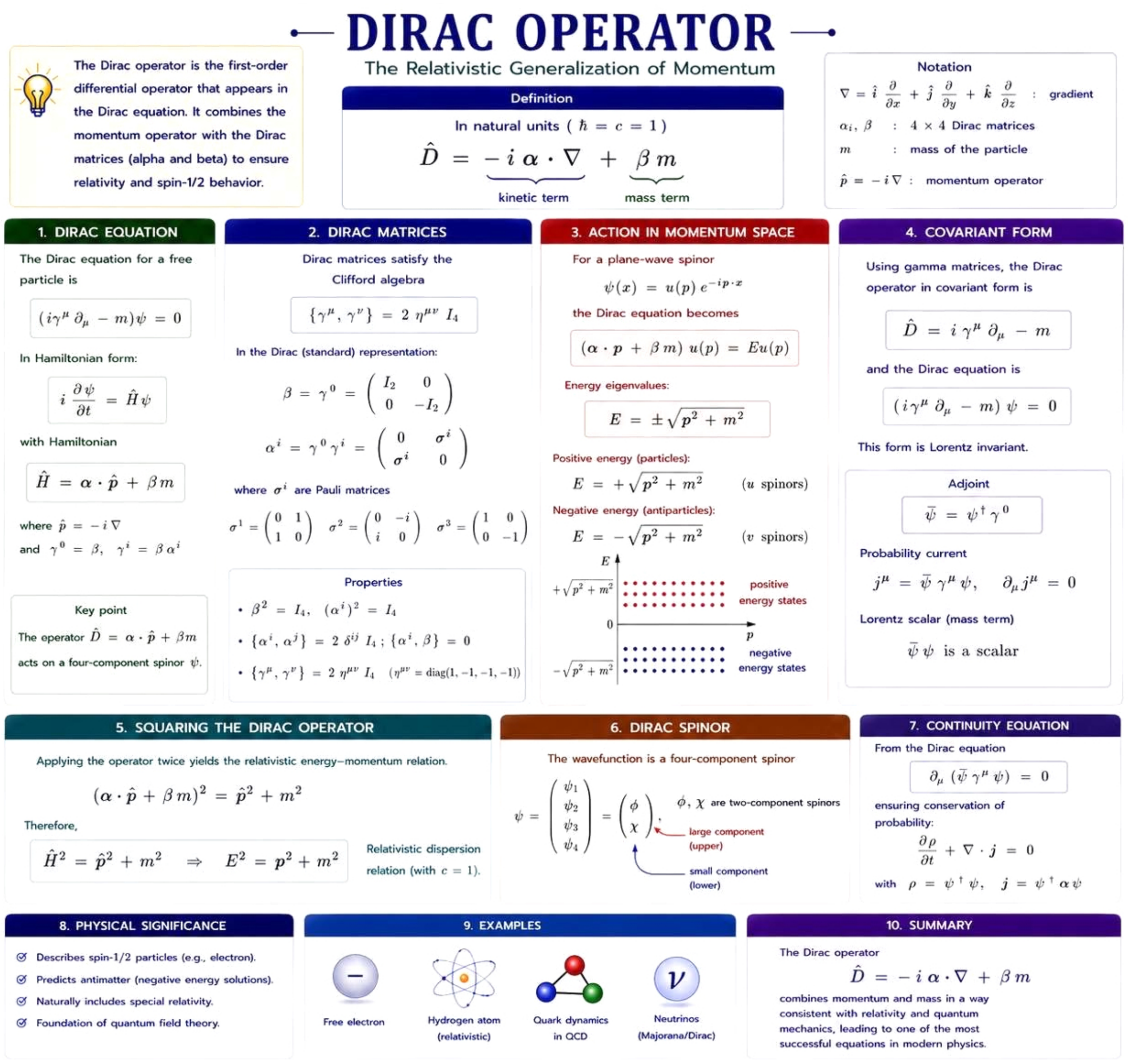
$$(\gamma^0)^2=1,(\gamma^i)^2=-1 \tag{11.3.257}$$

The reader should also know that physicists have derived from the above relation a small dozens of remarkable identities involving the γ-matrices and that also can be used to check that the chosen matrices can have indeed the status of γ-matrices. We will not present and prove these identities actually in this textbook as they are useless relatively to the actual content!

- Third condition:

$$[\alpha_i,\beta]_+=\begin{bmatrix}0 & \sigma_i\\ \sigma_i & 0\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}+\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}\begin{bmatrix}0 & \sigma_i\\ \sigma_i & 0\end{bmatrix}=\begin{bmatrix}0 & -\sigma_i\\ \sigma_i & 0\end{bmatrix}+\begin{bmatrix}0 & \sigma_i\\ -\sigma_i & 0\end{bmatrix}=\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix} \tag{11.3.258}$$

The third and last condition is thus fulfilled.



DIRAC OPERATOR

The Relativistic Generalization of Momentum

The Dirac operator is the first-order differential operator that appears in the Dirac equation. It combines the momentum operator with the Dirac matrices (alpha and beta) to ensure relativity and spin-1/2 behavior.

Definition

In natural units ($\hbar = c = 1$)

$$\hat{D} = \underbrace{-i\,\boldsymbol{\alpha}\cdot\nabla}_{\text{kinetic term}} + \underbrace{\beta\,m}_{\text{mass term}}$$

Notation

$\nabla = \hat{\imath}\,\frac{\partial}{\partial x} + \hat{\jmath}\,\frac{\partial}{\partial y} + \hat{k}\,\frac{\partial}{\partial z}$: gradient

$\alpha_i,\ \beta$: 4×4 Dirac matrices

m : mass of the particle

$\hat{p} = -i\nabla$: momentum operator

1. DIRAC EQUATION

The Dirac equation for a free particle is

$$\boxed{(i\gamma^\mu\,\partial_\mu - m)\psi = 0}$$

In Hamiltonian form:

$$\boxed{i\,\frac{\partial\psi}{\partial t} = \hat{H}\psi}$$

with Hamiltonian

$$\boxed{\hat{H} = \boldsymbol{\alpha}\cdot\hat{\mathbf{p}} + \beta m}$$

where $\hat{p} = -i\nabla$

and $\gamma^0 = \beta,\quad \gamma^i = \beta\,\alpha^i$

Key point

The operator $\hat{D} = \boldsymbol{\alpha}\cdot\hat{\mathbf{p}} + \beta m$ acts on a four-component spinor ψ.

2. DIRAC MATRICES

Dirac matrices satisfy the Clifford algebra

$$\boxed{\{\gamma^\mu, \gamma^\nu\} = 2\,\eta^{\mu\nu}\,I_4}$$

In the Dirac (standard) representation:

$$\beta = \gamma^0 = \begin{pmatrix} I_2 & 0 \\ 0 & -I_2 \end{pmatrix}$$

$$\alpha^i = \gamma^0\gamma^i = \begin{pmatrix} 0 & \sigma^i \\ \sigma^i & 0 \end{pmatrix}$$

where σ^i are Pauli matrices

$$\sigma^1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \sigma^2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \quad \sigma^3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

Properties

- $\beta^2 = I_4,\quad (\alpha^i)^2 = I_4$
- $\{\alpha^i, \alpha^j\} = 2\,\delta^{ij}\,I_4\ ;\ \{\alpha^i, \beta\} = 0$
- $\{\gamma^\mu, \gamma^\nu\} = 2\,\eta^{\mu\nu}\,I_4 \quad (\eta^{\mu\nu} = \mathrm{diag}(1, -1, -1, -1))$

3. ACTION IN MOMENTUM SPACE

For a plane-wave spinor

$$\psi(x) = u(p)\,e^{-ip\cdot x}$$

the Dirac equation becomes

$$\boxed{(\boldsymbol{\alpha}\cdot\mathbf{p} + \beta\,m)\,u(p) = Eu(p)}$$

Energy eigenvalues:

$$\boxed{E = \pm\sqrt{\mathbf{p}^2 + m^2}}$$

Positive energy (particles):

$$E = +\sqrt{\mathbf{p}^2 + m^2} \qquad (u\ \text{spinors})$$

Negative energy (antiparticles):

$$E = -\sqrt{\mathbf{p}^2 + m^2} \qquad (v\ \text{spinors})$$

4. COVARIANT FORM

Using gamma matrices, the Dirac operator in covariant form is

$$\boxed{\hat{D} = i\,\gamma^\mu\,\partial_\mu - m}$$

and the Dirac equation is

$$\boxed{(i\gamma^\mu\,\partial_\mu - m)\,\psi = 0}$$

This form is Lorentz invariant.

Adjoint

$$\boxed{\bar{\psi} = \psi^\dagger\gamma^0}$$

Probability current

$$j^\mu = \bar{\psi}\,\gamma^\mu\,\psi, \qquad \partial_\mu j^\mu = 0$$

Lorentz scalar (mass term)

$\bar{\psi}\,\psi$ is a scalar

5. SQUARING THE DIRAC OPERATOR

Applying the operator twice yields the relativistic energy–momentum relation.

$$(\boldsymbol{\alpha}\cdot\hat{\mathbf{p}} + \beta\,m)^2 = \hat{\mathbf{p}}^2 + m^2$$

Therefore,

$$\boxed{\hat{H}^2 = \hat{\mathbf{p}}^2 + m^2 \quad\Rightarrow\quad E^2 = \mathbf{p}^2 + m^2}$$

Relativistic dispersion relation (with $c = 1$).

6. DIRAC SPINOR

The wavefunction is a four-component spinor

$$\psi = \begin{pmatrix} \psi_1 \\ \psi_2 \\ \psi_3 \\ \psi_4 \end{pmatrix} = \begin{pmatrix} \phi \\ \chi \end{pmatrix},$$

$\phi,\ \chi$ are two-component spinors

large component (upper)

small component (lower)

7. CONTINUITY EQUATION

From the Dirac equation

$$\boxed{\partial_\mu\,(\bar{\psi}\,\gamma^\mu\,\psi) = 0}$$

ensuring conservation of probability:

$$\frac{\partial\rho}{\partial t} + \nabla\cdot\mathbf{j} = 0$$

with $\rho = \psi^\dagger\,\psi,\quad \mathbf{j} = \psi^\dagger\,\boldsymbol{\alpha}\,\psi$

8. PHYSICAL SIGNIFICANCE

- ✓ Describes spin-1/2 particles (e.g., electron).
- ✓ Predicts antimatter (negative energy solutions).
- ✓ Naturally includes special relativity.
- ✓ Foundation of quantum field theory.

9. EXAMPLES

10. SUMMARY

The Dirac operator

$$\hat{D} = -i\,\boldsymbol{\alpha}\cdot\nabla + \beta\,m$$

combines momentum and mass in a way consistent with relativity and quantum mechanics, leading to one of the most successful equations in modern physics.

Referring to the beginning equation written with the Dirac formalism:

$$\mathrm{i}\hbar\partial_t\Psi = H_D\Psi \Rightarrow \mathrm{i}\hbar\partial_t\Psi = \beta m_0 c^2\Psi + c\alpha_1 p_1\Psi + c\alpha_2 p_2\Psi + c\alpha_3 p_3\Psi \tag{11.3.259}$$

With:

$$\beta = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}, \alpha_1 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \alpha_2 = \begin{bmatrix} 0 & 0 & 0 & -\mathrm{i} \\ 0 & 0 & \mathrm{i} & 0 \\ 0 & -\mathrm{i} & 0 & 0 \\ \mathrm{i} & 0 & 0 & 0 \end{bmatrix}, \alpha_3 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix} \tag{11.3.260}$$

Which gives finally:

$$\mathrm{i}\hbar\partial_t=\begin{bmatrix}\Psi_1\\\Psi_2\\\Psi_3\\\Psi_4\end{bmatrix}=m_0c^2\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&-1&0\\0&0&0&-1\end{bmatrix}\begin{bmatrix}\Psi_1\\\Psi_2\\\Psi_3\\\Psi_4\end{bmatrix}+c\begin{bmatrix}0&0&0&1\\0&0&1&0\\0&1&0&0\\1&0&0&0\end{bmatrix}p_1\begin{bmatrix}\Psi_1\\\Psi_2\\\Psi_3\\\Psi_4\end{bmatrix}+$$
$$c\begin{bmatrix}0&0&0&-\mathrm{i}\\0&0&\mathrm{i}&0\\0&-\mathrm{i}&0&0\\\mathrm{i}&0&0&0\end{bmatrix}p_2\begin{bmatrix}\Psi_1\\\Psi_2\\\Psi_3\\\Psi_4\end{bmatrix}+c\begin{bmatrix}0&0&1&0\\0&0&0&-1\\1&0&0&0\\0&-1&0&0\end{bmatrix}p_3\begin{bmatrix}\Psi_1\\\Psi_2\\\Psi_3\\\Psi_4\end{bmatrix}\tag{11.3.261}$$

We find ourselves in front of a state function having 4 components in which:

$$\begin{bmatrix}\Psi_1\\\Psi_2\end{bmatrix}\quad\text{and}\quad\begin{bmatrix}\Psi_3\\\Psi_4\end{bmatrix}\tag{11.3.262}$$

are spinners and the set:

$$\begin{bmatrix}\Psi_1\\\Psi_2\\\Psi_3\\\Psi_4\end{bmatrix}\tag{11.3.263}$$

is therefore "Dirac bispinor" and we denote by:

$$\Psi=\begin{bmatrix}\Psi_1\\\Psi_2\\\Psi_3\\\Psi_4\end{bmatrix}\tag{11.3.264}$$

the "Dirac state function". The reader will have maybe noticed that we fall back on the same concepts as in our study of the free non-linearised Dirac equation. This is quite a good sign!

By developing, it comes:

$$\begin{aligned}\mathrm{i}\hbar\partial_t\Psi_1&=m_0c^2\Psi_1+cp_1\Psi_4-\mathrm{i}cp_2\Psi_4+cp_3\Psi_3\\&=m_0c^2\Psi_1+c(p_1-\mathrm{i}p_2)\Psi_4+cp_3\Psi_3\\i\hbar\partial_t\Psi_2&=m_0c^2\Psi_2+cp_1\Psi_3+\mathrm{i}cp_2\Psi_3-cp_3\Psi_4\\&=m_0c^2\Psi_2+c(p_1+\mathrm{i}p_2)\Psi_4-cp_3\Psi_4\\\mathrm{i}\hbar\partial_t\Psi_3&=-m_0c^2\Psi_3+cp_1\Psi_2-\mathrm{i}cp_2\Psi_2+cp_3\Psi_4\\&=-m_0c^2\Psi_3+c(p_1-\mathrm{i}p_2)\Psi_2+cp_3\Psi_1\\\mathrm{i}\hbar\partial_t\Psi_4&=-m_0c^2\Psi_4+cp_1\Psi_1+icp_2\Psi_1-cp_3\Psi_2\\&=-m_0c^2\Psi_4+c(p_1+\mathrm{i}p_2)\Psi_1-cp_3\Psi_2\end{aligned}\tag{11.3.265}$$

For a free electron, we know that the solution is:

$$\Psi(\vec{r},t)=\Psi_0\exp\left(-\mathrm{i}\frac{Et-\vec{p}\circ\vec{r}}{\hbar}\right)\tag{11.3.266}$$

With Dirac bispinor we have:

$$\Psi(\vec{r},t)=\Psi_\alpha(\vec{r})\exp\left(-\mathrm{i}\frac{Et}{\hbar}\right)\tag{11.3.267}$$

with:

$$\Psi_\alpha(\vec{r})=\begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} \exp\left(\mathrm{i}\frac{\vec{p}\circ\vec{r}}{\hbar}\right) \tag{11.3.268}$$

where b_1 to b_2 are the components of the Dirac bispinor.

We will write:

$$\left.\begin{array}{l} \Psi_1=b_1[\quad] \\ \Psi_2=b_2[\quad] \\ \Psi_3=b_3[\quad] \\ \Psi_4=b_4[\quad] \end{array}\right\} \quad \text{with} \quad [\quad]=\exp\left(\mathrm{i}\frac{\vec{p}\circ\vec{r}}{\hbar}\right)\exp\left(-\mathrm{i}\frac{Et}{\hbar}\right) \tag{11.3.269}$$

By calculating their derivatives with respect to t:

$$\begin{aligned} &\mathrm{i}\hbar\partial_t\Psi_1=\mathrm{i}\hbar\partial_t\left[b_1\exp\left(\mathrm{i}\frac{\vec{p}\circ\vec{r}}{\hbar}\right)\right]\exp\left(-\mathrm{i}\frac{Et}{\hbar}\right)=\mathrm{i}\hbar\left(-v\mathrm{i}\frac{E}{\hbar}\right)b_1[\quad]=Eb_1[\quad] \\ &\mathrm{i}\hbar\partial_t\Psi_2=Eb_2[\quad] \\ &\mathrm{i}\hbar\partial_t\Psi_3=Eb_3[\quad] \\ &\mathrm{i}\hbar\partial_t\Psi_4=Eb_4[\quad] \end{aligned} \tag{11.3.270}$$

With (11.3.269) and (11.3.270) in (11.3.265), it comes:

$$\begin{aligned} Eb_1[\quad]&=m_0c^2b_1[\quad]+c(p_1-\mathrm{i}p_2)b_4[\quad]+cp_3b_3[\quad] \\ Eb_2[\quad]&=m_0c^2b_2[\quad]+c(p_1+\mathrm{i}p_2)b_3[\quad]-cp_3b_4[\quad] \\ Eb_3[\quad]&=-m_0c^2b_3[\quad]+c(p_1-\mathrm{i}p_2)b_2[\quad]+cp_3b_1[\quad] \\ Eb_4[\quad]&=-m_0c^2b_4[\quad]+c(p_1+\mathrm{i}p_2)b_1[\quad]-cp_3b_2[\quad] \end{aligned} \tag{11.3.271}$$

Hence a system of equations whose unknowns are b_1,b_2,b_3,b_4:

$$\begin{cases} (E-m_0c^2)b_1+0b_2-cp_3b_3-c(p_1-\mathrm{i}p_2)b_4=0 & (a) \\ 0b_1+(E-m_0c^2)b_2-c(p_1+\mathrm{i}p_2)b_3+cp_3b_4=0 & (b) \\ -cp_3b_1-c(p_1-\mathrm{i}p_2)b_2+(E+m_0c^2)b_3+0b_4=0 & (c) \\ -c(p_1+\mathrm{i}p_2)b_1+cp_3b_2+0b_3+(E+m_0c^2)b_4=0 & (d) \end{cases} \tag{11.3.272}$$

We have solutions that are not all zero if and only if the determinant of the coefficients is zero (to know the reasons, refer to the section of Linear Algebra at page 1926) and therefore an infinite number of solutions (for the components of the Dirac spinor). Therefore:

$$\begin{vmatrix} E-m_0c^2 & 0 & -cp_3 & -c(p_1-\mathrm{i}p_2) \\ 0 & E-m_0c^2 & -c(p_1+\mathrm{i}p_2) & cp_3 \\ -cp_3 & -c(p_1-\mathrm{i}p_2) & E+m_0c^2 & 0 \\ -c(p_1+\mathrm{i}p_2) & cp_3 & 0 & E+m_0c^2 \end{vmatrix}=0 \tag{11.3.273}$$

By simplifying by c:

$$\begin{vmatrix} \dfrac{E}{c}-m_0c & 0 & -p_3 & -(p_1-\mathrm{i}p_2) \\ 0 & \dfrac{E}{c}-m_0c & -(p_1+\mathrm{i}p_2) & p_3 \\ -p_3 & -(p_1-\mathrm{i}p_2) & \dfrac{E}{c}+m_0c & 0 \\ -(p_1+\mathrm{i}p_2) & p_3 & 0 & \dfrac{E}{c}+m_0c \end{vmatrix}=0 \tag{11.3.274}$$

The division in the preceding determinant allows the calculation of the partial determinants (see section Linear Algebra page 1971):

$$\begin{vmatrix} \left(\dfrac{E}{c}-m_0c\right)^2 & \begin{array}{c}[-p_3^2-(p_1-\mathrm{i}p_2)(p_1+\mathrm{i}p_2)] \\ -[p_3^2+(p_1-\mathrm{i}p_2)(p_1+\mathrm{i}p_2)]\end{array} \\ \begin{array}{c}[-p_3^2-(p_1-\mathrm{i}p_2)(p_1+\mathrm{i}p_2)] \\ -[p_3^2+(p_1-\mathrm{i}p_2)(p_1+\mathrm{i}p_2)]\end{array} & \left(\dfrac{E}{c}+m_0c\right)^2 \end{vmatrix}=0 \tag{11.3.275}$$

In solving the preceding determinant, it comes:

$$\begin{aligned} &\left[\left(\frac{E}{c}-m_0c\right)\left(\frac{E}{c}+m_0c\right)\right]^2-\left[p_3^2+(p_1-\mathrm{i}p_2)(p_1+\mathrm{i}p_2)\right]^2=0 \\ &\left(\frac{E^2}{c^2}-m_0^2c^2\right)^2-\left\{p_3^2+\left[p_1^2-(\mathrm{i}p_2)^2\right]\right\}=0 \\ &\left(\frac{E^2}{c^2}-m_0^2c^2\right)^2-\underbrace{(p_3^2+p_1^2+p_2^2)}_{(p^2)^2}=0 \end{aligned} \tag{11.3.276}$$

Hence the following relation:

$$\frac{E^2}{c^2}-m_0^2c^2=p^2\Leftrightarrow E^2-m_0^2c^4=p^2c^2 \tag{11.3.277}$$

The values of the energy given by the Dirac equation are thus (we recognize here a famous relation determined in the section of Special Relativity!):

$$E=\pm\sqrt{m_0^2c^4+p^2c^2} \tag{11.3.278}$$

Therefore:

$$E_+=+\sqrt{m_0^2c^4+p^2c^2} \tag{11.3.279}$$

$$E_-=-\sqrt{m_0^2c^4+p^2c^2} \tag{11.3.280}$$

If we adopt for E_+ two constant values for b_1 and b_2 we have two relations to calculate b_3 and b_4, thus:

- With (11.3.272 c):

  $$-cp_3b_1-c(p_1-\mathrm{i}p_2)b_2+(E_++m_0c^2)b_3=0 \tag{11.3.281}$$

  Thus:

  $$b_3=\frac{c\left[p_3b_1+(p_1-\mathrm{i}p_2)b_2\right]}{E_++m_0c^2} \tag{11.3.282}$$

- With (11.3.272 d):

  $$-c(p_1+\mathrm{i}p_2)b_1-cp_3b_2+(E_++m_0c^2)b_4=0 \tag{11.3.283}$$

Thus:

$$b_4 = \frac{c\left[(p_1 + \mathrm{i}p_2)b_1 - p_3 b_2\right]}{E_+ + m_0 c^2} \tag{11.3.284}$$

By adopting $b_1 = 1$, $b_2 = 0$, it comes:

$$b_3 = \frac{cp_3}{E_+ + m_0 c^2} = \frac{cp_z}{E_+ + m_0 c^2} \qquad b_4 = \frac{c(p_1 + \mathrm{i}p_2)}{E_+ + m_0 c^2} = \frac{cp_+}{E_+ + m_0 c^2} \tag{11.3.285}$$

Taking the natural units:

$$b_3 = \frac{p_z}{E_+ + m_0} \qquad b_4 = \frac{p_+}{E_+ + m_0} \tag{11.3.286}$$

If we adopt for E_- two constant values for b_3, b_4 we have two relations to calculate b_1, b_2, thus:

- With (11.3.272 a):

$$(E_- - m_0 c^2)b_1 - cp_3 b_3 - c(p_1 - \mathrm{i}p_2)b_4 = 0 \tag{11.3.287}$$

Thus:

$$b_1 = \frac{c\left[p_3 b_3 + (p_1 - \mathrm{i}p_2)b_4\right]}{E_- - m_0 c^2} \tag{11.3.288}$$

- With (11.3.272 b):

$$(E_- - m_0 c^2)b_2 - c(p_1 + \mathrm{i}p_2)b_3 - cp_3 b_4 = 0 \tag{11.3.289}$$

Thus:

$$b_2 = \frac{c\left[-p_3 b_4 + (p_1 + \mathrm{i}p_2)b_3\right]}{E_- - m_0 c^2} \tag{11.3.290}$$

Notice that by adopting $b_3 = 1$, $b_4 = 0$ it comes:

$$b_1 = \frac{cp_3}{E_- - m_0 c^2} = \frac{cp_z}{E_- - m_0 c^2} \qquad b_2 = \frac{c(p_1 + \mathrm{i}p_2)}{E_- - m_0 c^2} = \frac{cp_+}{E_- - m_0 c^2} \tag{11.3.291}$$

With the natural units:

$$b_1 = \frac{p_3}{E_- - m_0} = \frac{p_z}{E_- - m_0} \qquad b_2 = \frac{p_1 + \mathrm{i}p_2}{E_- - m_0} = \frac{p_+}{E_- - m_0} \tag{11.3.292}$$

By adopting $b_3 = 0$, $b_4 = 1$ it comes:

$$b_1 = \frac{c(p_1 - \mathrm{i}p_2)}{E_- - m_0 c^2} = \frac{cp_-}{E_- - m_0 c^2} \qquad b_2 = \frac{-cp_3}{E_- - m_0 c^2} = \frac{-cp_z}{E_- - m_0 c^2} \tag{11.3.293}$$

Either with the natural units:

$$b_1 = \frac{p_-}{E_- - m_0} \qquad b_2 = \frac{-p_z}{E_- - m_0} \tag{11.3.294}$$

Even if the method is different, we fall back on the coefficients of the spinors obtained in our study of the classical free Dirac equation. This comfort us in the assumptions made at the beginning of this linearisation and validates these results. Moreover, the preceding relations also indicate a degeneracy of order two of the energy for each value of the pulsation. In the absence of an external field, the free electron is therefore not influenced by the orientation of its spin. We thus find the same results for either the classical or linearised free Dirac equation.

However, Dirac's explanation for the positive and negative energies is that his equation applies not only to the state of a positive energy particle (in this case the electron) but also to the state of a particle with negative energy (its antiparticle is the "positron"). The absolute value of these two energies being strictly equal.

The presence of the negative sign affecting the energy posed a problem at the time for its interpretation (in the context where we omit the time variable since we had seen when studying the free Klein-Gordon equation, a particle with negative energy can be seen as a particle that goes back in time).

If we reason in the case where the term p^2c^2 is small compared to m_0c^2, we ask ourselves: how and what are the consequences of a transition between a state of energy m_0c^2 and that of the state of energy $-m_0c^2$ with a gap of $-m_0c^2$ (we will find again this value in our study of materialization in the section of Nuclear Physics).

Dirac uses the image of a sea of negative energy named "Dirac sea" (since, remember, the number of solutions to our matrix system is infinite, hence the analogy with a sea than a discrete context) in which all negative energy states are occupied by the electrons and the positive energy states would be empty. If an electron is subjected to a transition (via, for example, a photon of energy greater than $2m_0c^2$), it leaves this sea leaving behind a gap (the famous "hole" of positive charge to which the electronics sometimes refers to...). This gap becomes a positive charge, of energy E_+. The appearance of this gap is assimilated to the appearance of a particle having a positive charge. Obviously, we can imagine the opposite case, it is only a matter of conventions.

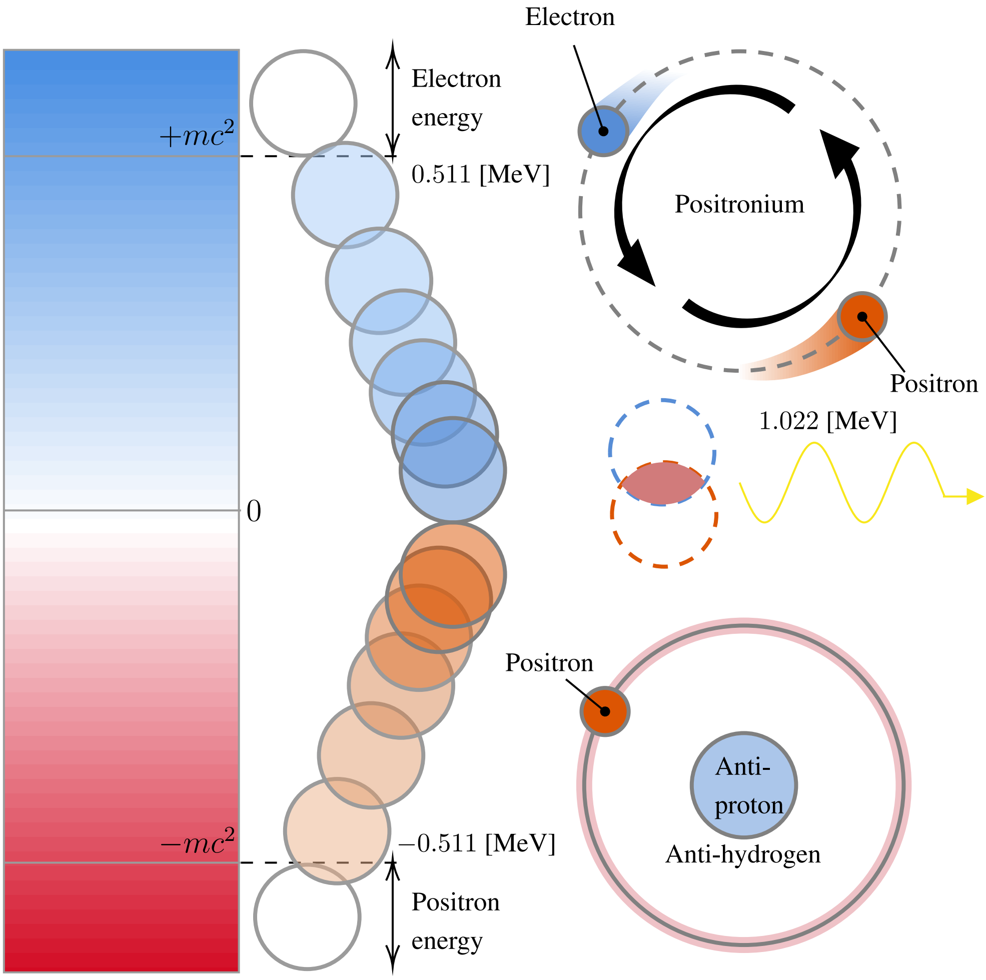

Electron
Electron energy
$+mc^2$
0.511 [MeV]
Positronium
Positron
1.022 [MeV]
0
Positron
Anti-proton
Anti-hydrogen
$-mc^2$
−0.511 [MeV]
Positron energy

Figure 11.84: **Dirac sea**

11.3.4.1 Linearised Dirac Equation continuity equation

Now let us consider the probability density/current. This is where the perceived problems with the Klein-Gordon equation arose! For this, identically as we did for the Generalized Klein-Gordon equation, we start with the linearised Dirac equation (in natural units to simplify):

$$(\vec{\alpha}\circ\vec{p}+\beta m_0)\Psi=\mathrm{i}\partial_t\Psi \qquad (11.3.295)$$

or:

$$(\mathrm{i}\vec{\alpha}\vec{\nabla}+\beta m_0)\Psi=\mathrm{i}\partial_t\Psi \qquad (11.3.296)$$

Where we know now obviously that:

1. $\vec{\alpha}$ are Pauli matrices
2. $\vec{\alpha}$ and $\vec{\beta}$ are Hermitian

Writing the above relation explicitly gives us:

$$-\mathrm{i}\alpha_x\frac{\partial\Psi}{\partial x}-\mathrm{i}\alpha_y\frac{\partial\Psi}{\partial y}-\mathrm{i}\alpha_z\frac{\partial\Psi}{\partial z}+m_0\beta\Psi=\mathrm{i}\frac{\partial\Psi}{\partial t} \qquad (11.3.297)$$

and its Hermitian conjugate:

$$+\mathrm{i}\alpha_x^\dagger\frac{\partial\Psi^\dagger}{\partial x}+\mathrm{i}\alpha_y^\dagger\frac{\partial\Psi^\dagger}{\partial y}+\mathrm{i}\alpha_z^\dagger\frac{\partial\Psi^\dagger}{\partial z}+m_0\beta^\dagger\Psi^\dagger=-\mathrm{i}\frac{\partial\Psi^\dagger}{\partial t} \tag{11.3.298}$$

Now, exactly as we did with the Generalized Klein-Gordon equation, we take the first above expansion that we multiply by $\Phi^\dagger$ and we subtract by the second Hermitian expansion that we multiply by Φ:

$$\Psi^\dagger\left(-\mathrm{i}\alpha_x\frac{\partial\Psi}{\partial x}-\mathrm{i}\alpha_y\frac{\partial\Psi}{\partial y}-\mathrm{i}\alpha_z\frac{\partial\Psi}{\partial z}+m_0\beta\Psi\right)-\Psi\left(+\mathrm{i}\alpha_x^\dagger\frac{\partial\Psi^\dagger}{\partial x}+\mathrm{i}\alpha_y^\dagger\frac{\partial\Psi^\dagger}{\partial y}+\mathrm{i}\alpha_z^\dagger\frac{\partial\Psi^\dagger}{\partial z}+m_0\beta^\dagger\Psi^\dagger\right)=\mathrm{i}\Psi^\dagger\frac{\partial\Psi}{\partial t}+\mathrm{i}\Psi\frac{\partial\Psi^\dagger}{\partial t} \tag{11.3.299}$$

Remembering that α and β are Hermitian, we can write:

$$\Psi^\dagger\left(-\mathrm{i}\alpha_x\frac{\partial\Psi}{\partial x}-\mathrm{i}\alpha_y\frac{\partial\Psi}{\partial y}-\mathrm{i}\alpha_z\frac{\partial\Psi}{\partial z}+m_0\beta\Psi\right)-\Psi\left(+\mathrm{i}\alpha_x^\dagger\frac{\partial\Psi^\dagger}{\partial x}+\mathrm{i}\alpha_y\frac{\partial\Psi^\dagger}{\partial y}+\mathrm{i}\alpha_z\frac{\partial\Psi^\dagger}{\partial z}+m_0\beta\Psi^\dagger\right)=\mathrm{i}\Psi^\dagger\frac{\partial\Psi}{\partial t}+\mathrm{i}\Psi\frac{\partial\Psi^\dagger}{\partial t} \tag{11.3.300}$$

The right hand side can easily be simplified by chain derivation rule:

$$\Psi^\dagger\left(-\mathrm{i}\alpha_x\frac{\partial\Psi}{\partial x}-\mathrm{i}\alpha_y\frac{\partial\Psi}{\partial y}-\mathrm{i}\alpha_z\frac{\partial\Psi}{\partial z}+m_0\beta\Psi\right)-\Psi\left(+\mathrm{i}\alpha_x\frac{\partial\Psi^\dagger}{\partial x}+\mathrm{i}\alpha_y\frac{\partial\Psi^\dagger}{\partial y}+\mathrm{i}\alpha_z\frac{\partial\Psi^\dagger}{\partial z}+m_0\beta\Psi^\dagger\right)=\mathrm{i}\frac{\partial\Psi^\dagger\Psi}{\partial t} \tag{11.3.301}$$

Se see also that the terms $m_0\beta\Psi$ cancels each other, then it remains:

$$\Psi^\dagger\left(-\mathrm{i}\alpha_x\frac{\partial\Psi}{\partial x}-\mathrm{i}\alpha_y\frac{\partial\Psi}{\partial y}-\mathrm{i}\alpha_z\frac{\partial\Psi}{\partial z}\right)-\Psi\left(+\mathrm{i}\alpha_x\frac{\partial\Psi^\dagger}{\partial x}+\mathrm{i}\alpha_y\frac{\partial\Psi^\dagger}{\partial y}+\mathrm{i}\alpha_z\frac{\partial\Psi^\dagger}{\partial z}\right)=\mathrm{i}\frac{\partial\Psi^\dagger\Psi}{\partial t} \tag{11.3.302}$$

We multiply both sides by i and rearrange a bit:

$$\Psi^\dagger\left(\alpha_x\frac{\partial\Psi}{\partial x}+\alpha_y\frac{\partial\Psi}{\partial y}+\alpha_z\frac{\partial\Psi}{\partial z}\right)+\Psi\left(\alpha_x\frac{\partial\Psi^\dagger}{\partial x}+\alpha_y\frac{\partial\Psi^\dagger}{\partial y}+\alpha_z\frac{\partial\Psi^\dagger}{\partial z}\right)+\frac{\partial\Psi^\dagger\Psi}{\partial t}=0 \tag{11.3.303}$$

Now using the identity:

$$\Psi^\dagger\alpha_x\frac{\partial\Psi}{\partial x}+\frac{\partial\Psi^\dagger}{\partial x}\alpha_x\Psi=\frac{\partial(\Psi^\dagger\alpha_x\Psi)}{\partial x} \tag{11.3.304}$$

Gives the linearised continuity Dirac equation:

$$\boxed{\vec{\nabla}(\Psi^\dagger\vec{\alpha}\Psi)+\frac{(\Psi^\dagger\Psi)}{\partial t}=0} \tag{11.3.305}$$

The probability density and current can be identified as:

$$\boxed{\rho=\Psi^\dagger\Psi\qquad\text{and}\qquad\vec{j}=\Psi^\dagger\vec{\alpha}\Psi} \tag{11.3.306}$$

where:

$$\rho=\Psi^\dagger\Psi=|\Psi_1|^2+|\Psi_2|^2+|\Psi_3|^2+|\Psi_4|^2>0 \tag{11.3.307}$$

Then, unlike the generalized Klein-Gordon equation, the Dirac equation has probability densities which are always positive!

In regard of what we have seen earlier, we can introduce the "four-vector fermion current":

$$j^{\mu}=\Psi^{\dagger}\gamma^{\mu}\Psi=(\psi^{\dagger}\gamma^{0}\Psi,\Psi^{\dagger}\vec{\gamma}\Psi)=(\rho,\vec{j}) \tag{11.3.308}$$

and ρ is the probability density as we know:

$$\rho=j^{0}=\Psi^{\dagger}\gamma^{0}\Psi=\Psi^{\dagger}\Psi \tag{11.3.309}$$

Now let us come back to the normalization of N that appeared during our study of the free Dirac particle solution page 4951. If we proceed by analogy (...), remember first that we get with the Klein-Gordon equation that:

$$\rho=2EA^{2}=2E \tag{11.3.310}$$

And just now we get that:

$$\rho=\Psi^{\dagger}\Psi \tag{11.3.311}$$

So now knowing this, let us consider the following solution that we have obtained:

$$\Psi_{\vec{p}}^{(1)}=N\begin{pmatrix}\phi_a\\ \phi_b\end{pmatrix}=N\begin{pmatrix}\begin{pmatrix}1\\0\end{pmatrix}\\ \begin{pmatrix}\dfrac{p_z}{E+m_0}\\ \dfrac{p_x+\mathrm{i}p_y}{E+m_0}\end{pmatrix}\end{pmatrix}=N\begin{pmatrix}\begin{pmatrix}1\\0\end{pmatrix}\\ \begin{pmatrix}\dfrac{p_z}{E+m_0}\\ \dfrac{p_+}{E+m_0}\end{pmatrix}\end{pmatrix}=N\begin{pmatrix}1\\ \dfrac{\vec{p}\circ\vec{\sigma}}{E+m_0}\end{pmatrix}\chi^{(+)} \tag{11.3.312}$$

that to simplify we will denote as Ψ. Then (still in natural units):

$$\begin{aligned}\rho=\Psi^{\dagger}\Psi&=|N|^2\left(1+\frac{p_z^2}{E+m}+\frac{p_x^2+p_y^2}{E+m}\right)=|N|^2\left(\frac{(E+m)^2+|\vec{p}|^2}{(E+m)^2}\right)\\&=|N|^2\left(\frac{(E+m)^2+(\sqrt{E^2-m^2})^2}{(E+m)^2}\right)=|N|^2\left(\frac{(E+m)^2+E^2-m^2}{(E+m)^2}\right)\\&=|N|^2\frac{2E^2+2Em}{(E+m)^2}=|N|^2\frac{2E(E+m)}{(E+m)^2}=|N|^2\frac{2E}{E+m}\end{aligned} \tag{11.3.313}$$

Now to get the same result as for the Klein-Gordon equation, we must have:

$$\rho=2EA^2=|N|^2\frac{2E}{E+m} \tag{11.3.314}$$

Hence:

$$N=\pm A\sqrt{E+M} \tag{11.3.315}$$

Or more differently written:

$$N=\pm\Psi_0\sqrt{E+M} \tag{11.3.316}$$

11.3.4.2 Linearised Dirac Equation Angular Momentum + Helicity Conservation

For the Dirac linearised equation model is orbital angular momentum a good quantum number? I.e. does:

$$\vec{L}=\vec{r}\times\vec{p}=\begin{pmatrix} yp_z-zp_y \\ zp_x-xp_z \\ xp_y-yp_x \end{pmatrix} \tag{11.3.317}$$

commute with the Hamiltonian:

$$[H,\vec{L}]=[\vec{\alpha}\circ\vec{p}+\beta m_0,\vec{r}\times\vec{p}] \tag{11.3.318}$$

where for recall the corresponding Dirac alpha and beta matrices are:

$$\beta_D=\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&-1&0\\0&0&0&-1\end{bmatrix},\alpha_D^1=\begin{bmatrix}0&0&0&1\\0&0&1&0\\0&1&0&0\\1&0&0&0\end{bmatrix},\alpha_D^2=\begin{bmatrix}0&0&0&-\mathrm{i}\\0&0&\mathrm{i}&0\\0&-\mathrm{i}&0&0\\\mathrm{i}&0&0&0\end{bmatrix},\alpha_D^3=\begin{bmatrix}0&0&1&0\\0&0&0&-1\\1&0&0&0\\0&-1&0&0\end{bmatrix} \tag{11.3.319}$$

To check this, let us consider the x component of $\vec{L}$ (we implicitly assume that we work with the Dirac alpha and beta matrices so we don't denote the lower index D anymore):

$$[H,L_x]=[\vec{\alpha}\circ\vec{p},(\vec{r}\times\vec{p})_x]=[\alpha_1p_x+\alpha_2p_y+\alpha_3p_z+\beta m_0,yp_z-zp_y] \tag{11.3.320}$$

So we may think that:

$$[H,L_x]=[\alpha_1p_x+\alpha_2p_y+\alpha_3p_z+\beta m_0,yp_z-zp_y]=\alpha_1p_x+\alpha_2p_y+\alpha_3p_z,0]=0 \tag{11.3.321}$$

even if the shortcut is nice and may look correct... this is a wrong shortcut in fact especially when we deal with differential operators as both arguments of the commutator (typical well known trap in maths when we do things too quickly)!!!

Remark

The term βm_0 disappears as it is not a differential operator (only a real matrix with a constant factor m_0). By cons, the first term has the $\vec{p}$ in it, and this latter is a vector of differential operators!

We already know that for the linear momentum, the only non-zero contributions come form:

$$[x,p_x]=[y,p_y]=[z,p_z]=\mathrm{i}\hbar \tag{11.3.322}$$

Hence (we use a lot here the Poisson brackets properties seen at page 3189):

$$\begin{aligned}[H,L_x]&=[\vec{\alpha}\circ\vec{p},yp_z-zp_y]\\&=[\vec{\alpha}\circ\vec{p},yp_z]-[\vec{\alpha}\circ\vec{p},zp_y]\\&=[\alpha_1p_x+\alpha_2p_y+\alpha_3p_z,yp_z]-[\alpha_1p_x+\alpha_2p_y+\alpha_3p_z,zp_y]\\&=[\alpha_1p_x,yp_z]+[\alpha_2p_y,yp_z]+[\alpha_3p_z,yp_z]-[\alpha_1p_x,zp_y]-[\alpha_2p_y,zp_y]-[\alpha_3p_z,zp_y]\\&=[\alpha_1p_x,y]p_z+y[\alpha_1p_x,p_z]+[\alpha_2p_y,y]p_z+y[\alpha_2p_y,p_z]+[\alpha_3p_z,y]p_z+z[\alpha_3p_z,p_z]\\&-[\alpha_1p_x,z]p_y-z[\alpha_1p_x,p_y]-[\alpha_2p_y,z]p_y-z[\alpha_2p_y,p_y]-[\alpha_3p_z,z]p_y-z[\alpha_3p_z,p_y]\\&=0+0-\mathrm{i}\alpha_2p_z+0+0+0-0-0-0-0+\mathrm{i}\alpha_3p_x-0\\&=-\mathrm{i}(\alpha_2p_z-\alpha_3p_y)\end{aligned} \tag{11.3.323}$$

Two techniques were employed along the way: between the 4th and the 5th line of the above development, we used (see page 3189):

$$[A,BC]=[A,B]C+B[A,C] \tag{11.3.324}$$

and between the 7th and 8th we used:

$$[p_k,x_l]=-[x_l,p_k]=-\mathrm{i}\delta_{kl} \tag{11.3.325}$$

The same strategy can be adopted to solve for $[H,L_y]$ and $[H,L_z]$:

$$\begin{aligned}[H,L_y]&=\mathrm{i}(\alpha_1 p_z-\alpha_3 p_x)\\ [H,L_z]&=-\mathrm{i}(\alpha_1 p_x-\alpha_2 p_x)\end{aligned} \tag{11.3.326}$$

Thus $[H,\vec{L}]$ can be determined by:

$$\begin{aligned}[H,\vec{L}]&=[H,L_x]\vec{e}_x+[H,L_y]\vec{e}_y+[H,L_z]\vec{e}_z\\ &=-\mathrm{i}(\alpha_2 p_z-\alpha_3 p_y)\vec{e}_x+\mathrm{i}(\alpha_1 p_z-\alpha_3 p_x)\vec{e}_y+(-\mathrm{i}(\alpha p_y-\alpha_2 p_x)\,\vec{e}_z\end{aligned} \tag{11.3.327}$$

Therefore:

$$\boxed{[H,\vec{L}]=-\mathrm{i}\vec{\alpha}\times\vec{p}\neq\vec{0}} \tag{11.3.328}$$

Hence the angular momentum does not commute with the Hamiltonian and is not a constant of motion!

Introduce a new 4×4 operator:

$$\boxed{\vec{S}=\frac{1}{2}\vec{\Sigma}=\frac{1}{2}\begin{bmatrix}\vec{\sigma} & 0\\ 0 & \vec{\sigma}\end{bmatrix}} \tag{11.3.329}$$

where as we know, the $\vec{\sigma}$ are the Pauli spin matrices i.e.:

$$\Sigma_1=\begin{bmatrix}0&1&0&0\\1&0&0&0\\0&0&0&1\\0&0&1&0\end{bmatrix},\Sigma_2=\begin{bmatrix}0&-\mathrm{i}&0&0\\\mathrm{i}&0&0&0\\0&0&0&-\mathrm{i}\\0&0&\mathrm{i}&0\end{bmatrix},\Sigma_3=\begin{bmatrix}1&0&0&0\\0&-1&0&0\\0&0&1&0\\0&0&0&-1\end{bmatrix} \tag{11.3.330}$$

Now consider the commutator:

$$[H,\vec{\Sigma}]=[\vec{\alpha}\circ\vec{p}+\beta m_0,\vec{\Sigma}] \tag{11.3.331}$$

here (obvious!):

$$[\beta,\vec{\Sigma}]=\begin{bmatrix}\mathbb{1}&0\\0&-\mathbb{1}\end{bmatrix}\begin{bmatrix}\vec{\sigma}&0\\0&\vec{\sigma}\end{bmatrix}-\begin{bmatrix}\vec{\sigma}&0\\0&\vec{\sigma}\end{bmatrix}\begin{bmatrix}\mathbb{1}&0\\0&-\mathbb{1}\end{bmatrix}=0 \tag{11.3.332}$$

and hence:

$$[H,\vec{\Sigma}]=[\vec{\alpha}\circ\vec{p},\vec{\Sigma}] \tag{11.3.333}$$

Consider the x component:

$$[H,\Sigma_1]=[\alpha_1 p_x+\alpha_2 p_y+\alpha_3 p_z,\Sigma_1]=p_x[\alpha_1,\Sigma_1]+p_y[\alpha_2,\Sigma_1]+p_z[\alpha_3,\Sigma_1] \tag{11.3.334}$$

Taking each of the commutators in turn:

$$\begin{aligned}
[\alpha_1,\Sigma_1]&=\begin{bmatrix}0&\sigma_x\\ \sigma_x&0\end{bmatrix}\begin{bmatrix}\sigma_x&0\\0&\sigma_x\end{bmatrix}-\begin{bmatrix}\sigma_x&0\\0&\sigma_x\end{bmatrix}\begin{bmatrix}0&\sigma_x\\ \sigma_x&0\end{bmatrix}=0\\
[\alpha_2,\Sigma_1]&=\begin{bmatrix}0&\sigma_y\\ \sigma_y&0\end{bmatrix}\begin{bmatrix}\sigma_x&0\\0&\sigma_x\end{bmatrix}-\begin{bmatrix}\sigma_x&0\\0&\sigma_x\end{bmatrix}\begin{bmatrix}0&\sigma_y\\ \sigma_y&0\end{bmatrix}\\
&=\begin{bmatrix}0&\sigma_y\sigma_y-\sigma_y\sigma_x\\ \sigma_y\sigma_x-\sigma_x\sigma_y&0\end{bmatrix}=\begin{bmatrix}0&-2\mathrm{i}\sigma_z\\-2\mathrm{i}\sigma_z&0\end{bmatrix}=-2\mathrm{i}\alpha_3\\
[\alpha_3,\Sigma_1]&=2\mathrm{i}\alpha_2
\end{aligned} \tag{11.3.335}$$

Hence:

$$[H,\Sigma_1]=p_x[\alpha_1,\Sigma_1]+p_y[\alpha_2,\Sigma_1]+p_z[\alpha_3,\Sigma_1]=-2\mathrm{i}p_y\alpha_3+2\mathrm{i}p_z\alpha_2=2\mathrm{i}(\vec{\alpha}\times\vec{p})_x \tag{11.3.336}$$

The same strategy can be used to solve for $[H,\Sigma_2]$ and $[H,\Sigma_3]$:

$$\begin{aligned}
[H,\Sigma_2]&=-\mathrm{i}(\alpha_1 p_z-\alpha_3 p_x)\\
[H,\Sigma_3]&=\mathrm{i}(\alpha_1 p_y-\alpha_2 p_x)
\end{aligned} \tag{11.3.337}$$

Therefore:

$$\boxed{[H,\vec{\Sigma}]=2\mathrm{i}\vec{\alpha}\times\vec{p}\neq\vec{0}} \tag{11.3.338}$$

Hence the observable corresponding to the operator $\vec{\Sigma}$ is also not a constant of motion. However, referring back to our relation:

$$[H,\vec{L}]=-\mathrm{i}\vec{\alpha}\times\vec{p} \tag{11.3.339}$$

We have:

$$[H,\vec{S}]=\frac{1}{2}[H,\vec{\Sigma}]=\mathrm{i}\vec{\alpha}\times\vec{p}=-[H,\vec{L}] \tag{11.3.340}$$

Therefore:

$$\boxed{[H,\vec{S}+\vec{L}]=[H,\vec{L}]+[H,\vec{S}]=[H,\vec{L}]-[H,\vec{L}]=\vec{0}} \tag{11.3.341}$$

Because:

$$\vec{S}=\frac{1}{2}\begin{bmatrix}\vec{\sigma}&0\\0&\vec{\sigma}\end{bmatrix} \tag{11.3.342}$$

Thus, Dirac equation proved a description of "intrinsic" angular moment (i.e. spin).

How let us show that helicity $\hat{h}$ (projection of the spin along the direction of motion for recall), does commute with the Hamiltonian, which means that helicity is a conserved quantity. Let us recall the definition of helicity:

$$\hat{h}=\frac{\vec{p}\circ\vec{\sigma}}{2\|\vec{p}\|} \tag{11.3.343}$$

So we want to calculate:

$$[H,\hat{h}]=\left[\vec{\alpha}\circ\vec{p}+\beta m_0,\frac{\vec{p}\circ\vec{\sigma}}{2\|\vec{p}\|}\right] \tag{11.3.344}$$

We see here a problem... Indeed, on the left term we have 4×4 matrices, and on the right term a 2×2 matrices. So we try by moving to then generalized total spin matrix introduced earlier above:

$$[H,\hat{h}]=\left[\vec{\alpha}\circ\vec{p}+\beta m_0,\frac{\vec{p}\circ\vec{\Sigma}}{2\|\vec{p}\|}\right] \tag{11.3.345}$$

Therefore if we focus on a only a component given that the direction is only along the z-axis, we can write:

$$\begin{aligned}[H,\hat{h}_z]&=\left[\alpha_3p_z+\beta m_0,\frac{p_z\Sigma_3}{2\|\sqrt{p_z^2}\|}\right]=\left[\alpha_3p_z+\beta m_0,\frac{\Sigma_3}{2}\right]=\frac{1}{2}\left[\alpha_3p_z+\beta m_0,\begin{bmatrix}\sigma_1 & 0\\ 0 & \sigma_1\end{bmatrix}\right]\\ &=\frac{1}{2}\left\{(\alpha_3p_z+\beta m_0)\begin{bmatrix}\sigma_3 & 0\\ 0 & \sigma_3\end{bmatrix}-\begin{bmatrix}\sigma_3 & 0\\ 0 & \sigma_3\end{bmatrix}(\alpha_3p_z+\beta m_0)\right\}\end{aligned} \tag{11.3.346}$$

And, for recall, as:

$$\beta_D=\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&-1&0\\0&0&0&-1\end{bmatrix}=\begin{bmatrix}\mathbb{1}&0\\0&-\mathbb{1}\end{bmatrix},\alpha_D^3=\begin{bmatrix}0&0&1&0\\0&0&0&-1\\1&0&0&0\\0&-1&0&0\end{bmatrix}=\begin{bmatrix}0&\sigma_3\\\sigma_3&0\end{bmatrix} \tag{11.3.347}$$

Hence:

$$\begin{aligned}[H,\hat{h}_z]=&\frac{1}{2}\left\{(\alpha_3p_z+\beta m_0)\begin{bmatrix}\sigma_3&0\\0&\sigma_3\end{bmatrix}-\begin{bmatrix}\sigma_3&0\\0&\sigma_3\end{bmatrix}(\alpha_3p_z+\beta m_0)\right\}\\ =&\frac{1}{2}\left\{\left(\begin{bmatrix}0&\sigma_3\\\sigma_3&0\end{bmatrix}p_z+\begin{bmatrix}\mathbb{1}&0\\0&-\mathbb{1}\end{bmatrix}m_0\right)\begin{bmatrix}\sigma_3&0\\0&\sigma_3\end{bmatrix}-\begin{bmatrix}\sigma_3&0\\0&\sigma_3\end{bmatrix}\left(\begin{bmatrix}0&\sigma_3\\\sigma_3&0\end{bmatrix}p_z+\begin{bmatrix}\mathbb{1}&0\\0&-\mathbb{1}\end{bmatrix}m_0\right)\right\}\\ =&\frac{1}{2}\begin{bmatrix}0&\sigma_3p_z\\\sigma_3p_z&0\end{bmatrix}\begin{bmatrix}\sigma_3&0\\0&\sigma_3\end{bmatrix}+\frac{1}{2}m_0\begin{bmatrix}\mathbb{1}&0\\0&-\mathbb{1}\end{bmatrix}\begin{bmatrix}\sigma_3&0\\0&\sigma_3\end{bmatrix}-\frac{1}{2}\begin{bmatrix}\sigma_3&0\\0&\sigma_3\end{bmatrix}\begin{bmatrix}0&\sigma_3p_z\\\sigma_3p_z&0\end{bmatrix}\\ &-\frac{m_0}{2}\begin{bmatrix}\sigma_3&0\\0&\sigma_3\end{bmatrix}\begin{bmatrix}\mathbb{1}&0\\0&-\mathbb{1}\end{bmatrix}\\ =&\frac{1}{2}\begin{bmatrix}0&\sigma_3p_z\sigma_3\\\sigma_3p_z\sigma_3&\end{bmatrix}-\frac{1}{2}\begin{bmatrix}0&\sigma_3\sigma_3p_z\\\sigma_3\sigma_3p_z&\end{bmatrix}\end{aligned} \tag{11.3.348}$$

If:

$$\sigma_3p_z\sigma_3=\sigma_3\sigma_3p_z \tag{11.3.349}$$

then helicity commutes with the Hamiltonian, and therefore helicity is a conserved quantity. Let us verify the previous equality:

$$\sigma_3p_z\sigma_3=\begin{bmatrix}1&0\\0&-1\end{bmatrix}p_z\begin{bmatrix}1&0\\0&-1\end{bmatrix}=\begin{bmatrix}p_z&0\\0&-p_z\end{bmatrix}\begin{bmatrix}1&0\\0&-1\end{bmatrix}=\begin{bmatrix}p_z&0\\0&p_z\end{bmatrix} \tag{11.3.350}$$

and:

$$\sigma_3\sigma_3 p_z=\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}p_z=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}p_z=\begin{bmatrix}p_z & 0\\ 0 & p_z\end{bmatrix} \tag{11.3.351}$$

Therefore:

$$[H,\hat{h}_z]=\frac{1}{2}\begin{bmatrix}0 & \sigma_3 p_z\sigma_3\\ \sigma_3 p_z\sigma_3 & \end{bmatrix}-\frac{1}{2}\begin{bmatrix}0 & \sigma_3\sigma_3 p_z\\ \sigma_3\sigma_3 p_z & \end{bmatrix}=0 \tag{11.3.352}$$

The same strategy can be adopted to solve for $[H,h_1]$ and $[H,h_2]$:

$$[H,h_1]=0[H,h_2]\quad=0 \tag{11.3.353}$$

Therefore:

$$\boxed{[H,\hat{h}]=0} \tag{11.3.354}$$

Helicity is therefore a conserved quantity!

11.3.5 Generalized Dirac Equation

In the case of the free electron, we have thus often seen and proved that the Hamiltonian has as expression:

$$E=(p^2c^2+m_0^2c^4)^{1/2} \tag{11.3.355}$$

In the case of an electron moving in an electromagnetic field, we have also proved in our study the Klein-Gordon equation at the beginning of this section that:

$$(E-qU)^2=\left(\vec{p}-q\vec{A}\right)^2c^2+m_0^2c^4 \tag{11.3.356}$$

Therefore:

$$E=\left[\left(\vec{p}-q\vec{A}\right)^2c^2+m_0^2c^4\right]^{\frac{1}{2}}+qU \tag{11.3.357}$$

OK finish for the recall!

If we now take back the Dirac Hamiltonian for the free electron proved earlier above (which for recall is a 4×4 matrix):

$$H_D=c\vec{\alpha}\circ\vec{p}+\beta m_0c^2 \tag{11.3.358}$$

and from the fact that we must add to the Hamiltonian the term of potential electrostatic energy via the potential vector, we get:

$$H_D=c\vec{\alpha}\circ\left(\vec{p}-q\vec{A}\right)+qU+\beta m_0c^2 \tag{11.3.359}$$

where the potential U will also have to be expressed in the form of a diagonal 4×4 matrix.

We then get the generalized Hamiltonian of Dirac in the following traditional form:

$$\boxed{(H_D-qU)=c\vec{\alpha}\circ\left(\vec{p}-q\vec{A}\right)+\beta m_0c^2} \tag{11.3.360}$$

Therefore we have in another well known form:

$$\mathrm{i}\hbar\frac{\partial\Psi}{\partial t}=\left[c\vec{\alpha}\circ\left(\vec{p}-q\vec{A}\right)+qU+\beta m_0c^2\right]\Psi \tag{11.3.361}$$

Or in natural units:

$$\mathrm{i}\hbar\frac{\partial\Psi}{\partial t}=\left[\vec{\alpha}\circ\left(\vec{p}-q\vec{A}\right)+qU+\beta m_0\right]\Psi \tag{11.3.362}$$

And as we know $\vec{p}=\mathrm{i}\vec{\nabla}=\mathrm{i}\partial_\mu$ and $\vec{\alpha}=\vec{\gamma}$, hence the well known form (using $q=e$ for the electron charge) of the generalized Dirac equation:

$$\boxed{\mathrm{i}\hbar\frac{\partial\Psi}{\partial t}=\left[\gamma^\mu\left(\mathrm{i}\partial_\mu-eA_\mu\right)+eU+\beta m_0\right]\Psi} \tag{11.3.363}$$

Caution!!! There is no principle of physics yet (in the beginning in the 121st century according to the holocene calendar) that says physicals laws or constants have to be the same everywhere and always. Indeed, the reader must remember that the "laws" of physics are **descriptive**, not **imperative** and therefore they are some observed regularity in what we can observe experimentally with sufficiently wide occurrence or relevance! Asking whether the laws can change is no more and no less than asking whether regularities which have been observed up till now might not be quite as universal as was thought. The answer to that is: YES! This could drive the reader a bit mad. Indeed, how can we possibly do Science if the rules keep changing? But if you are a professional active scientist, you have to face up to that possibility - and you have a responsibility to check out if there is evidence beyond reasonable doubt about it! We will repeat this warning multiple times in this textbook as its seems that this is something necessary!

11.3.6 Pauli Equation

In quantum mechanics, the "Pauli equation" or "Schrödinger–Pauli equation" is the formulation of the Schrödinger equation for spin-$1/2$ particles, which takes into account the interaction of the particle's spin with an external electromagnetic field. It is the non-relativistic limit of the Dirac equation and can be used where particles are moving at speeds much less than the speed of light, so that relativistic effects can be neglected. It was formulated by Wolfgang Pauli in 11927 (holocene calendar).

Let us now consider a two-component representation of the spinor:

$$\Psi(\vec{r},t)=\begin{bmatrix}\phi_a(\vec{r},t)\\ \phi_b(\vec{r},t)\end{bmatrix} \tag{11.3.364}$$

and let us recall that:

$$\alpha=\begin{bmatrix}0 & \vec{\sigma}\\ \vec{\sigma} & 0\end{bmatrix} \quad \text{and} \quad \beta=\begin{bmatrix}\mathbb{1} & 0\\ 0 & -\mathbb{1}\end{bmatrix} \tag{11.3.365}$$

Therefore it comes:

$$\mathrm{i}\hbar\frac{\partial\Psi}{\partial t}=c\begin{bmatrix}0 & \vec{\sigma}\\ \vec{\sigma} & 0\end{bmatrix}\circ\left(\vec{p}-q\vec{A}\right)\Psi+qU\Psi+\begin{bmatrix}\mathbb{1} & 0\\ 0 & -\mathbb{1}\end{bmatrix}m_0c^2\Psi \tag{11.3.366}$$

Hence:

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\begin{bmatrix}\phi_a\\ \phi_b\end{bmatrix}=c\begin{bmatrix}0 & \vec{\sigma}\\ \vec{\sigma} & 0\end{bmatrix}\circ\left(\vec{p}-q\vec{A}\right)\begin{bmatrix}\phi_a\\ \phi_b\end{bmatrix}+qU\begin{bmatrix}\phi_a\\ \phi_b\end{bmatrix}+\begin{bmatrix}\mathbb{1} & 0\\ 0 & -\mathbb{1}\end{bmatrix}m_0c^2\begin{bmatrix}\phi_a\\ \phi_b\end{bmatrix} \tag{11.3.367}$$

What after simplification gives:

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\begin{bmatrix}\phi_a\\ \phi_b\end{bmatrix}=\begin{pmatrix}c\vec{\sigma}\circ\left(\vec{p}-q\vec{A}\right)\phi_b\\ c\vec{\sigma}\circ\left(\vec{p}-q\vec{A}\right)\phi_a\end{pmatrix}+qU\begin{bmatrix}\phi_a\\ \phi_b\end{bmatrix}+m_0c^2\begin{bmatrix}\phi_a\\ -\phi_b\end{bmatrix} \tag{11.3.368}$$

Before continuing, let us open an important parenthesis, otherwise we will not be able to find a solution to these two equations.

Let us recall that one of the spinors solution of the free Dirac equation was given by (as we have proved it earlier above):

$$\phi_b=\frac{\vec{p}\circ\vec{\sigma}}{E+m_0}\phi_a \qquad (11.3.369)$$

Either in International Standard units:

$$\phi_b=\frac{c\vec{p}\circ\vec{\sigma}}{E+m_0c^2}\phi_a \qquad (11.3.370)$$

In order to simplify the calculation of the prior-previous equations we will reduce the situation to a non-relativistic case, that is to say when the mass energy is much greater than the kinetic energy. Therefore the previous solution becomes (we forget the second which would cause problems ...):

$$\phi_b\cong\frac{\vec{p}\circ\vec{\sigma}}{2m_0c}\phi_a \qquad (11.3.371)$$

The idea is simple but someone had to think about it! It should be noticed that depending on the authors and professors it is the following term which is named "helicity":

$$\frac{\vec{p}\circ\vec{\sigma}}{2m_0c}=\frac{\vec{p}\circ\vec{\sigma}}{2\|\vec{p}\|} \qquad (11.3.372)$$

as it is the projection (as we will see later) of the spin operator on the linear momentum.

The idea is then to find a solution to:

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\begin{bmatrix}\phi_a\\ \phi_b\end{bmatrix}=\begin{pmatrix}c\vec{\sigma}\circ\left(\vec{p}-q\vec{A}\right)\phi_b\\ c\vec{\sigma}\circ\left(\vec{p}-q\vec{A}\right)\phi_a\end{pmatrix}+qU\begin{bmatrix}\phi_a\\ \phi_b\end{bmatrix}+m_0c^2\begin{bmatrix}\phi_a\\ -\phi_b\end{bmatrix} \qquad (11.3.373)$$

such that we make a non-relativistic approximation and we cancel the magnetic field (in extenso the vector potential), we fall back on:

$$\phi_b=\frac{\vec{p}\circ\vec{\sigma}}{2m_0c}\phi_a \qquad (11.3.374)$$

After many trials and errors (yes Quantum Physics was not made in one day... not even if one year...) we find that a particular solution satisfying our previous idea is:

$$\begin{bmatrix}\phi_a\\ \phi_b\end{bmatrix}=\begin{bmatrix}\phi_{a0}\\ \phi_{b0}\end{bmatrix}e^{-\mathrm{i}\frac{mc^2t}{\hbar}} \qquad (11.3.375)$$

Indeed:

$$\begin{gathered}
\mathrm{i}\hbar\frac{\partial}{\partial t}\begin{bmatrix}\phi_{a0}\\ \phi_{b0}\end{bmatrix}e^{-\mathrm{i}\frac{mc^2t}{\hbar}}=\begin{pmatrix}c\vec{\sigma}\circ\left(\vec{p}-q\vec{A}\right)\phi_{b0}e^{-\mathrm{i}\dfrac{mc^2t}{\hbar}}\\ c\vec{\sigma}\circ\left(\vec{p}-q\vec{A}\right)\phi_{a0}e^{-\mathrm{i}\frac{mc^2t}{\hbar}}\end{pmatrix}+\left(qU\begin{bmatrix}\phi_{a0}\\ \phi_{b0}\end{bmatrix}+m_0c^2\begin{bmatrix}\phi_{a0}\\ -\phi_{b0}\end{bmatrix}\right)e^{-\mathrm{i}\frac{mc^2t}{\hbar}}\\
\mathrm{i}\hbar\begin{bmatrix}\dfrac{\partial}{\partial t}\phi_{a0}\\ \dfrac{\partial}{\partial t}\phi_{b0}\end{bmatrix}e^{-\mathrm{i}\frac{mc^2t}{\hbar}}+\mathrm{i}\hbar\begin{bmatrix}\phi_{a0}\\ \phi_{b0}\end{bmatrix}\frac{\partial}{\partial t}e^{-\mathrm{i}\frac{mc^2t}{\hbar}}=\begin{pmatrix}c\vec{\sigma}\left(\vec{p}-q\vec{A}\right)\phi_{b0}\\ c\vec{\sigma}\left(\vec{p}-q\vec{A}\right)\phi_{a0}\end{pmatrix}e^{-\mathrm{i}\dfrac{mc^2t}{\hbar}}+\left(qU\begin{bmatrix}\phi_{a0}\\ \phi_{b0}\end{bmatrix}+m_0c^2\begin{bmatrix}\phi_{a0}\\ -\phi_{b0}\end{bmatrix}\right)e^{-\mathrm{i}\dfrac{mc^2t}{\hbar}}\\
\mathrm{i}\hbar\begin{bmatrix}\dfrac{\partial}{\partial t}\phi_{a0}\\ \dfrac{\partial}{\partial t}\phi_{b0}\end{bmatrix}e^{-\mathrm{i}\frac{mc^2t}{\hbar}}+mc^2\begin{bmatrix}\phi_{a0}\\ \phi_{b0}\end{bmatrix}e^{-\mathrm{i}\dfrac{mc^2t}{\hbar}}=\begin{pmatrix}c\vec{\sigma}\left(\vec{p}-q\vec{A}\right)\phi_{b0}\\ c\vec{\sigma}\left(\vec{p}-q\vec{A}\right)\phi_{a0}\end{pmatrix}e^{-\mathrm{i}\frac{mc^2t}{\hbar}}+\left(qU\begin{bmatrix}\phi_{a0}\\ \phi_{b0}\end{bmatrix}+m_0c^2\begin{bmatrix}\phi_{a0}\\ -\phi_{b0}\end{bmatrix}\right)e^{-\mathrm{i}\frac{mc^2t}{\hbar}}\\
\mathrm{i}\hbar\frac{\partial}{\partial t}\begin{bmatrix}\phi_{a0}\\ \phi_{b0}\end{bmatrix}=\begin{pmatrix}c\vec{\sigma}\circ\left(\vec{p}-q\vec{A}\right)\phi_{b0}\\ c\vec{\sigma}\circ\left(\vec{p}-q\vec{A}\right)\phi_{a0}\end{pmatrix}+qU\begin{bmatrix}\phi_{a0}\\ \phi_{b0}\end{bmatrix}-2m_0c^2\begin{bmatrix}0\\ \phi_{b0}\end{bmatrix}
\end{gathered}\tag{11.3.376}$$

We finally have 2 equations (yes yes... we know that in fact as spinor has two components and therefore in reality we don't have 2 equations but 4 equations...!):

$$\begin{gathered}
\mathrm{i}\hbar\frac{\partial}{\partial t}\phi_{a0}=c\vec{\sigma}\circ\left(\vec{p}-q\vec{A}\right)\phi_{b0}+qU\phi_{a0}\\
\mathrm{i}\hbar\frac{\partial}{\partial t}\phi_{b0}=c\vec{\sigma}\circ\left(\vec{p}-q\vec{A}\right)\phi_{a0}+qU\phi_{b0}-2m_0c^2\phi_{b0}
\end{gathered}\tag{11.3.377}$$

Now let us consider only the second equation:

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\phi_{b0}=c\vec{\sigma}\circ\left(\vec{p}-q\vec{A}\right)\phi_{a0}+qU\phi_{b0}-2m_0c^2\phi_{b0}\tag{11.3.378}$$

Assuming (for free, after which it will be necessary to compare that assumption to experimental results) that the term $|\partial_t\phi_b|$ is much smaller than $|2m_0c^2\phi_{b0}|$ we can write:

$$0\cong c\vec{\sigma}\circ\left(\vec{p}-q\vec{A}\right)\phi_{a0}+qU\phi_{b0}-2m_0c^2\phi_{b0}\tag{11.3.379}$$

By making the same assumption with $|qU\phi_{b0}|$ we have:

$$0\cong c\vec{\sigma}\circ\left(\vec{p}-q\vec{A}\right)\phi_{a0}-2m_0c^2\phi_{b0}\tag{11.3.380}$$

We then have (do not forget that the denominator is a diagonal matrix in reality ...):

$$\phi_{b0}\cong\frac{\vec{\sigma}\circ\left(\vec{p}-q\vec{A}\right)}{2m_0c}\phi_{a0}\tag{11.3.381}$$

Now, we see that if the magnetic field (in extenso the vector potential) vanishes, we fall back well on our starting idea! So the bet was good!

Because of all these downward approximations, the ϕ_b component is often taken as the "small" component of the wave function Ψ, relatively to the large component ϕ_a.

The first equation:

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\phi_{a0}=c\vec{\sigma}\circ\left(\vec{p}-q\vec{A}\right)\phi_{b0}+qU\phi_{a0}\tag{11.3.382}$$

can now be simplified easily by taking the previous solution such that:

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\phi_{a0} \cong c\vec{\sigma}\circ\left(\vec{p}-q\vec{A}\right)\left(\frac{\vec{\sigma}\circ\left(\vec{p}-q\vec{A}\right)}{2m_0c}\right)\phi_{a0}+qU\phi_{a0} \tag{11.3.383}$$

Therefore:

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\phi_{a0} \cong \frac{c\vec{\sigma}\circ\left(\vec{p}-q\vec{A}\right)\vec{\sigma}\circ\left(\vec{p}-q\vec{A}\right)}{2m_0c}\phi_{a0}+qU\phi_{a0} \tag{11.3.384}$$

Using the remarkable identity proved in the section of Spinor Calculus:

$$(\vec{\sigma}\circ\vec{a})(\vec{\sigma}\circ\vec{b})=(\vec{a}\circ\vec{b})\mathbb{1}+\mathrm{i}\circ(\vec{a}\times\vec{b}) \tag{11.3.385}$$

we get:

$$\begin{aligned}\vec{\sigma}\circ(\vec{p}-q\vec{A})\vec{\sigma}\circ(\vec{p}-q\vec{A}) &= \left(\vec{p}-q\vec{A}\right)^2+\mathrm{i}\vec{\sigma}\circ\left[\left(\vec{p}-q\vec{A}\right)\times\left(\vec{p}-q\vec{A}\right)\right]\\ &= \left(\vec{p}-q\vec{A}\right)^2+\mathrm{i}\vec{\sigma}\circ\left[\left(-\mathrm{i}\hbar\vec{\nabla}-q\vec{A}\right)\times\left(-\mathrm{i}\hbar\vec{\nabla}-q\vec{A}\right)\right]\end{aligned} \tag{11.3.386}$$

Let us detail the cross product by remembering that it will act as an operator on ϕ_{a0}:

$$\begin{aligned}&\vec{\sigma}\circ\left(-\mathrm{i}\hbar\vec{\nabla}-q\vec{A}\right)\times\left(-\mathrm{i}\hbar\vec{\nabla}-q\vec{A}\right)\phi_{a0}\\ &=\underbrace{\vec{\sigma}\circ\left(-\mathrm{i}\hbar\vec{\nabla}\times-\mathrm{i}\hbar\vec{\nabla}\right)\phi_{a0}}_{=0}+\vec{\sigma}\circ\left(-\mathrm{i}\hbar\vec{\nabla}\times-q\vec{A}\right)\phi_{a0}+\vec{\sigma}\circ\left(-q\vec{A}\times-\mathrm{i}\hbar\vec{\nabla}\right)\phi_{a0}+\underbrace{\vec{\sigma}\circ\left(-q\vec{A}\times-q\vec{A}\right)\phi_{a0}}_{=0}\\ &=\mathrm{i}\hbar q\left[\vec{\sigma}\circ\left(\vec{\nabla}\times\vec{A}\right)\phi_{a0}+\vec{\sigma}\circ\left(\vec{A}\times\vec{\nabla}\right)\phi_{a0}\right]\end{aligned} \tag{11.3.387}$$

But, we have (see section Spinor Calculus page 2158):

$$\begin{aligned}&\vec{\sigma}\circ\left(\vec{\nabla}\times\vec{A}\right)\phi_{a0}+\vec{\sigma}\circ\left(\vec{A}\times\vec{\nabla}\right)\phi_{a0}=\\ &=\begin{bmatrix}\partial_xA_y-\partial_yA_x & (\partial_yA_z-\partial_zA_y)-\mathrm{i}(\partial_zA_x-\partial_xA_z)\\ (\partial_yA_z-\partial_zA_y)+\mathrm{i}(\partial_zA_x-\partial_xA_z) & -(\partial_xA_y-\partial_yA_x)\end{bmatrix}\\ &+\begin{bmatrix}A_x\partial_y-A_y\partial_x & (A_y\partial_z-A_z\partial_y)-\mathrm{i}(A_z\partial_x-A_x\partial_z)\\ (A_y\partial_z-A_z\partial_y)+\mathrm{i}(A_z\partial_x-A_x\partial_z) & -(A_x\partial_y-A_y\partial_x)\end{bmatrix}\end{aligned} \tag{11.3.388}$$

Let us just look at the component in the upper left corner (if not the calculations are too long and boring to do) of this sum of matrices. We must not forget that this component of the matrix will act on the first component as an operator on ϕ_{a0} (written with the same notation as the spinor itself to avoid to many indices):

$$\begin{aligned}&(\partial_xA_y\phi_{a0}-\partial_yA_x\phi_{a0})+(A_x\partial_y\phi_{a0}-A_y\partial_x\phi_{a0})\\ &=(\partial_x(A_y\phi_{a0})-\partial_y(A_x\phi_{a0}))+(A_x\partial_y\phi_{a0}-A_y\partial_x\phi_{a0})\\ &=(\partial_x(A_y\phi_{a0})-A_y\partial_x\phi_{a0}))+(A_x\partial_y\phi_{a0}-\partial_y(A_x\phi_{a0}))\end{aligned} \tag{11.3.389}$$

But:

$$\partial_x(A_y\phi_{a0})=A_y\partial_x\phi_{a0}+\phi_{a0}\partial_xA_y\Rightarrow\partial_x(A_y\phi_{a0})-A_y\partial_x\phi_{a0}=\phi_{a0}\partial_xA_y \tag{11.3.390}$$

$$\partial_y(A_x\phi_{a0})=A_x\partial_y\phi_{a0}+\phi_{a0}\partial_yA_x\Rightarrow A_x\partial_y\phi_{a0}-\partial_y(A_x\phi_{a0})=-\phi_{a0}\partial_yA_x \tag{11.3.391}$$

Therefore:

$$(\partial_x(A_y\phi_{a0})-A_y\partial_x\phi_{a0}))+(A_x\partial_y\phi_{a0}-\partial_y(A_x\phi_{a0}))=\phi_{a0}\partial_xA_y-\phi_{a0}\partial_yA_x \tag{11.3.392}$$

But, we recognize here the third component of a cross product that does not act as an operator. Finally, it comes:

$$\vec{\sigma}\circ\left(-\mathrm{i}\hbar\vec{\nabla}-q\vec{A}\right)\times\left(-\mathrm{i}\hbar\vec{\nabla}-q\vec{A}\right)\phi_{a0}=\vec{\sigma}\circ\left(-\mathrm{i}\hbar\vec{\nabla}\times-q\vec{A}\right)\phi_{a0}=\mathrm{i}\hbar q\vec{\sigma}\circ\left(\vec{\nabla}\times\vec{A}\right)\phi_{a0} \tag{11.3.393}$$

Therefore:

$$\begin{gathered}\left(\vec{p}-q\vec{A}\right)^2+\mathrm{i}\vec{\sigma}\circ\left[\left(-\mathrm{i}\hbar\vec{\nabla}-q\vec{A}\right)\times\left(-\mathrm{i}\hbar\vec{\nabla}-q\vec{A}\right)\right]=\left(\vec{p}-q\vec{A}\right)^2-\hbar q\vec{\sigma}\circ\left(\vec{\nabla}\times\vec{A}\right)\\=\left(\vec{p}-q\vec{A}\right)^2-\hbar q\vec{\sigma}\circ\vec{B}\end{gathered} \tag{11.3.394}$$

Thus, the relation of the main component:

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\phi_{a0}\cong\frac{c\vec{\sigma}\circ\left(\vec{p}-q\vec{A}\right)\vec{\sigma}\circ\left(\vec{p}-q\vec{A}\right)}{2m_0c}\phi_{a0}+qU\phi_{a0} \tag{11.3.395}$$

becomes:

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\phi_{a0}\cong\frac{\left(\vec{p}-q\vec{A}\right)^2-\hbar q\vec{\sigma}\circ\vec{B}}{2m_0}\phi_{a0}+qU\phi_{a0} \tag{11.3.396}$$

The reader will notice that the notations are not the most enjoyable (between vectors, matrices and constants it is necessary to follow well to know what is in there ...).

After rearrangement:

$$\boxed{\mathrm{i}\hbar\frac{\partial}{\partial t}\phi_{a0}\cong\left[\frac{\left(\vec{p}-q\vec{A}\right)^2}{2m_0}-\frac{\hbar q}{2m_0}\vec{\sigma}\circ\vec{B}+qU\right]\phi_{a0}} \tag{11.3.397}$$

which constitutes the "Pauli equation", or Schrödinger-Pauli equation", and thus describes in a relativistic way the two components of ϕ_{a0} of the spin of the electron (this is really a two-equation system).

The expression:

$$\boxed{\frac{\hbar q}{2m_0}\vec{\sigma}\circ\vec{B}} \tag{11.3.398}$$

is named the "Stern-Gerlach term" and represents the interaction energy of the magnetic field with the intrinsic moment of the electron (this is a 2×2 matrix for recall). And the term:

$$\frac{\left(\vec{p}-q\vec{A}\right)^2}{2m_0}-\frac{\hbar q}{2m_0}\vec{\sigma}\circ\vec{B}+qU \tag{11.3.399}$$

is named the "Pauli Hamiltonian" and denoted H_P.

The Pauli equation, and therefore that of Dirac (since the latter is more general), give the correct gyromagnetic factor for a free electron of $g=2$. To verify this, let us take as it has been done experimentally, a constant magnetic field.

We can easily verify that the choice of a vector potential corresponding to a constant magnetic field is then:

$$\vec{A}=\frac{1}{2}\vec{B}\times\vec{x} \tag{11.3.400}$$

This choice will have the effect of make disappear the vector potential in favour of the magnetic field in the Pauli equation, which will reveal the interaction between the orbital angular momentum and the magnetic field as we shall see:

Indeed, we have:

$$\vec{B}=\vec{\nabla}\times\vec{A}=\vec{\nabla}\times\left(\frac{1}{2}\vec{B}\times\vec{r}\right) \tag{11.3.401}$$

As requested by a reader, here is the proof of that latter relation:

$$\begin{aligned}&\vec{\nabla}\times\left(\frac{1}{2}\vec{B}\times\vec{r}\right)=\vec{\nabla}\times\left(\begin{pmatrix}B_yz-B_zy\\B_zx-B_xz\\B_xy-B_yx\end{pmatrix}\right)=\frac{1}{2}\begin{bmatrix}\partial_y(B_xy-B_yx)-\partial_z(B_zx-B_xz)\\\partial_z(B_yz-B_zy)-\partial_x(B_xy-B_yx)\\\partial_x(B_zx-B_xz)-\partial_y(B_yz-B_zy)\end{bmatrix}\\&=\frac{1}{2}\begin{pmatrix}B_x+B_x\\B_y+B_y\\B_z+B_z\end{pmatrix}=\begin{pmatrix}B_x\\B_y\\B_z\end{pmatrix}=\vec{B}\end{aligned} \tag{11.3.402}$$

We then have in Pauli equation:

$$\begin{aligned}\left(\vec{p}-q\vec{A}\right)^2=\left(\vec{p}-\frac{q}{2}\vec{B}\times\vec{x}\right)^2=\|\vec{p}\|^2-q(\vec{B}\times\vec{x})\circ\vec{p}+\underbrace{\frac{q^2}{4}\|\vec{B}\times\vec{x}\|^2}_{\cong 0}=\\=\|\vec{p}\|^2-q(\vec{B}\times\vec{x})\circ\vec{p}\end{aligned} \tag{11.3.403}$$

But, let us recall that we have seen in the section of Vector Calculus page 1830 that:

$$(\vec{x}\times\vec{y})\circ\vec{z}=\vec{x}\circ(\vec{y}\times\vec{z}) \tag{11.3.404}$$

This gives us therefore:

$$\|\vec{p}\|^2-q(\vec{B}\times\vec{x})\circ\vec{p}=\|\vec{p}\|^2-q\vec{B}\circ(\vec{x}\times\vec{p})=\|\vec{p}\|^2-q\vec{B}\circ\vec{L} \tag{11.3.405}$$

where:

$$\vec{L}=\vec{x}\times\vec{p} \tag{11.3.406}$$

denoted also $\vec{b}$ (see sections of Classical Mechanics page 3244 and Corpuscular Quantum Physics page 4629), is therefore an operator representing the angular momentum.

Therefore we have:

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\phi_{a0}\cong\left[\frac{\|\vec{p}\|^2}{2m_0}-\frac{q\vec{B}\circ\vec{L}}{2m_0}-\frac{\hbar q}{2m_0}\vec{\sigma}\circ\vec{B}+qU\right]\phi_{a0} \tag{11.3.407}$$

By proceeding as Dirac did, that is, by defining the spin operator as the matrix (oh! we fall back on something known and seen in the section of Wave Quantum Physics!! it's beautiful no!?):

$$\boxed{\vec{S}=\frac{1}{2}\hbar\vec{\sigma}} \tag{11.3.408}$$

This relationship will be very useful to us in the section Quantum Computing. Let us indicate the following notation (quite logic) in the literature for the intrinsic magnetic moment:

$$\vec{\mathcal{M}}=\frac{\hbar q}{2m_0}\vec{\sigma} \tag{11.3.409}$$

The Pauli equation is then written:

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\phi_{a0}\cong\left[\frac{\|\vec{p}\|^2}{2m_0}-\frac{q\vec{B}\circ\vec{L}}{2m_0}-\frac{q}{m_0}\vec{S}\circ\vec{B}+qU\right]\phi_{a0} \tag{11.3.410}$$

or after factorization (the Hamiltonian will then be named "Pauli Hamiltonian for a constant direction magnetic field"):

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\phi_{a0}\cong\left[\frac{\|\vec{p}\|^2}{2m_0}-\frac{q}{2m_0}(\vec{L}+2\vec{S})\circ\vec{B}+qU\right]\phi_{a0} \tag{11.3.411}$$

or even more condensed by being careful to differentiate what is an operator, what is a vector and what is a simple multiplication from an inner product and what is a function of a spinor ... (only happiness...) and by putting:

$$g=2 \tag{11.3.412}$$

as being the "Landé g-factor[28]" (see the quantum origin further below at page 4989) or "gyromagnetic factor", we have:

$$\boxed{\mathrm{i}\hbar\frac{\partial}{\partial t}\phi_{a0}\cong\left[\frac{\|\vec{p}\|^2}{2m_0}-(\vec{\mu}_L+g\vec{\mu}_S)\circ\vec{B}+qU\right]\phi_{a0}} \tag{11.3.413}$$

with $\vec{\mu}_L$ being the "orbital magnetic moment" (by comparing with the expression of the magnetic moment proved in the Magnetostatic section you will see that it actually has the same form and also the same units), $\vec{\mu}_S$ the "magnetic moment of spin".

Sometimes we find the expression in parentheses in the above-mentioned relationship as follows:

$$\vec{\mu}=\vec{\mu}_L+g\vec{\mu}_S \tag{11.3.414}$$

---

[28] It's important to specify the "Landé" in front of the g because experiments have provided evidence that $g=2$ is inaccurate! More precise calculations done by Julian Schwinger (see [438] sections 6.2 and 6.3 pages 184 to 196), and named the "Schwinger g-factor", gives a correction of $+\alpha/2\pi$ (anomalous magnetic dipole moment) that let to $g\cong 2+\alpha/2\pi\cong 2.0011614$. Even more precise calculation give $g\cong 2.0023193804386$.

So we have just seen that we can derive the electron's spin g-factor by starting with the Dirac equation for an electron in a weak electromagnetic field and looking at its non-relativistic limit. We then find that there is a supplementary interaction energy term having a factor 2.

With all this the Stern-Gerlach term (magnetic moment) becomes explicitly:

$$\boxed{\vec{\mathcal{M}}_S=g\frac{q}{2m_0}\frac{1}{2}\hbar\vec{\sigma}=\frac{g}{2}\mu_B\vec{\sigma}=g\frac{q}{2m_0}S} \tag{11.3.415}$$

Recalling that:

$$\mu_B=\frac{q\hbar}{2m_0} \tag{11.3.416}$$

is the Bohr magneton which we had introduced rigorously in the section of Corpuscular Quantum Physics.

Interestingly ... the neutron has (experimental observation!) a spin magnetic moment that is non-zero yet its charge is zero ... so it must be composed of charged particles (if we remain within the framework of a model of the explanation of nature by particles ...). It is interesting to know that in practice we have for the electron, the proton and the neutron:

$$\begin{aligned}\mu_{\text{electron}}&\cong\mu_g\\ \mu_{\text{proton}}&\cong 2.79\mu_B\\ \mu_{\text{neutron}}&\cong -1.91\mu_B\end{aligned} \tag{11.3.417}$$

and the following values were measured for spin $1/2$ particles such as the electron, proton and neutron (caution! the sign may change depending on how the Dirac equation is written):

$$\begin{aligned}g_e&\cong 2\\ g_p&\cong -5.585\\ g_n&=3.826\end{aligned} \tag{11.3.418}$$

Therefore the Dirac theory in the non-relativistic framework predicts in a good approximation that the elementary particles of spin $1/2$ have a gyromagnetic factor of 2, and this prediction consistent with the experiment for the electron is the greatest triumph of the Dirac equation. The deviations of the theoretical value (consequent deviations!) for the proton and the neutron are perfectly explained in the framework of quantum electrodynamics. These deviations show that the structure of the proton and that of the neutron are more complex (subparticle composition) than a spin $1/2$ point particle whereas in the case of the electron, it seems that it has no underlying substructure.

Remark

The gyromagnetic factor is sometimes taken as negative but it is only a matter of convention.

It is by the way the interaction term between the magnetic field and the addition of the orbital angular momentum and intrinsic spin:

$$\boxed{E_p=-\frac{q}{2m_0}(\vec{L}+2\vec{S})\circ\vec{B}} \tag{11.3.419}$$

of the Pauli Hamiltonian which gives the values measured by the Zeeman effect! Reason why this expression is sometimes named "Zeeman energy" and sometimes written in the form of a Hamiltonian operator:

$$\hat{H}_Z = -\frac{q}{2m_0}(\vec{L}+2\vec{S})\circ\vec{B} \tag{11.3.420}$$

Remark

The above treatment of the Zeeman effect describes the phenomenon when the magnetic field is small enough and that the orbital and spin angular momentum can be considered to be coupled (see section Wave Quantum Physics page 4899). For extremely strong magnetic fields this coupling is broken and another approach must be taken. The strong field effect is named the "Paschen-Back effect".

We know that this last relation can also be written for any single particle (see section Wave Quantum Physics page 4877) immersed in a magnetic field collinear to the z-axis and zero in the other directions:

$$\begin{aligned} E_p &= -(\vec{\mu}_L + g\vec{\mu}_S)\circ\vec{B} = -\frac{q}{2m_0}(\vec{L}+2\vec{S})\circ\vec{B} = -\frac{q}{2m_0}(L_z+2S_z)B_z \\ &= -\frac{q}{2m_0}(m_l\hbar+2s\hbar)B = -\frac{q}{2m_0}\hbar(m_l+2s)B \\ &= -\gamma\hbar(m_l+2s)B = -\mu_B\hbar(m_l+2s)B \end{aligned} \tag{11.3.421}$$

where γ is the "gyromagnetic ratio" and the μ_B the "Bohr magnetron" (and we saw in the proof of this relation that the Bohr magnetron is only a special case with the Landé g factor equal to that of the electron).

Let us now focus only on the smallest variation in energy between two states of the orbital angular momentum (since the smallest variation is 1 by the quantification of l). It will then always be equal to the form:

$$\Delta E_{p,l} = -\gamma\hbar B \tag{11.3.422}$$

Remark

As $\gamma\hbar \cong 5.8\cdot 10^{-5}$ [eV $\cdot$ T^{-1}] and that largest field which can regularly be produced in a lab is around 10 [T]. Thus, $\Delta E_{p,l}\cong 10^{-3}$ [eV]. This is a very small energy. Thus magnetic effects are wiped out by thermal fluctuation at about 10 [K]! Thus experiments which measure the magnetism of non-interacting systems must be carried out at low temperatures. These experiments typically measure the susceptibility of the system with a Faraday balance or a magnetometer (SQUID: Superconducting QUantum Interference Device).

This variation of energy will be restored in the form of electromagnetic waves corresponding to:

$$h\nu = h\frac{\omega}{2\pi} = -\gamma\hbar B = -\gamma\frac{h}{2\pi}B \tag{11.3.423}$$

hence:

$$\omega = -\gamma B \tag{11.3.424}$$

which is the "Larmor relation" (not to be confused with the "Larmor radius" seen in the Magnetostatic section). But in practice we mainly use the relation:

$$\nu = -\gamma \frac{B}{2\pi} \tag{11.3.425}$$

which gives what we name the "resonance frequency" of the orbital angular momentum.

Similarly, for the intrinsic kinetic moment, we get for a constant orbital kinetic moment (during the spin transition from $+1/2$ to $-1/2$ or vice versa):

$$\Delta E_{p,s} = -\Delta(\gamma\hbar 2s)B = -\gamma\hbar B\Delta(2s) = -2\gamma\hbar B\Delta s = -2\gamma\hbar B \tag{11.3.426}$$

and putting $g = 2$ it comes:

$$\Delta E_{p,s} = -g\gamma\hbar B \tag{11.3.427}$$

hence:

$$\omega = -g\gamma B \tag{11.3.428}$$

and obviously:

$$\boxed{\nu = -g\gamma \frac{B}{2\pi}} \tag{11.3.429}$$

This last study of energy variation due to the application of a magnetic field and the resulting energy emission frequencies is the basis of "nuclear magnetic resonance (NMR)", which therefore works only for the particles having the magnetic moment of spin μ_S (by construction of the Pauli Hamiltonian).

NMR consists in modifying the nuclear magnetic moment, in other words, to pass the nucleus from one level of energy to another by applying magnetic fields to the sample to be studied. When the energy of the photons constituting these magnetic fields corresponds to the energy of transition from one level of energy to the other, these photons can be absorbed by the nucleus: we then say that there is "nuclear resonance".

We can characterize the transition energy of the nuclear magnetic spin moment by giving the frequency of the electromagnetic wave that allows resonance. For the usual fields (of the order of the Tesla), the proton resonance takes place in the range of radio waves (100 [MHz] approximately): 42 [MHz] in a field of 1.0 [T] and 63 [MHz] In a field of 1.5 [T].

Let us point out that on the way that in some textbooks, since we have:

$$\Delta E_{p,s} = -g\gamma\hbar B \tag{11.3.430}$$

Then for a particle with two possible states of spin the energy of each of the level is then obviously half the above difference, which is often denoted by:

$$E_{\pm,s} = \hbar\omega_{\pm} = -\frac{g\gamma\hbar B}{2} = -\frac{g\mu_B B}{2} \tag{11.3.431}$$

11.3.6.1 Landé g-factor

We have seen previously that the Landé g-factor is a multiplicative term appearing in the expression for the energy levels of an atom in a weak magnetic field. We seek now for a more detailed expression and you will see that all the abstract stuff we have introduced during our study of orbital-spin coupling in the section of Wave Quantum Physics will be useful to us.

In the weak field limit, we assume that the magnetic dipole moment of the atom is proportional to the total angular momentum $\vec{J}$:

$$\vec{\mu}_J = g_J m_B \vec{J} \tag{11.3.432}$$

where m_B is still the Bohr magneton and g_J is the Landé g factor for $\vec{J}$. To find a value of g_J, we relate $\vec{J}$ to the known values of g_L and g_S and we use (see section Wave Quantum Physics page 4881):

$$\vec{J} = \vec{L} + \vec{S} \tag{11.3.433}$$

and therefore:

$$\vec{\mu}_J = \vec{\mu}_L + \vec{\mu}_S \tag{11.3.434}$$

with the assumption that:

$$\vec{\mu}_L = g_L m_B \vec{L} \qquad \text{and} \qquad \vec{\mu}_S = g_S m_B \vec{S} \tag{11.3.435}$$

Hence:

$$\vec{\mu}_J = g_J m_B \vec{J} = g_L m_B \vec{L} + g_S m_B \vec{S} \tag{11.3.436}$$

Therefore after simplification:

$$g_J \vec{J} = g_L \vec{L} + g_S \vec{S} \tag{11.3.437}$$

We then take the dot product with $\vec{J}$ and get:

$$g_J \vec{J} \circ \vec{J} = g_L \vec{L} \circ \vec{J} + g_S \vec{S} \circ \vec{J} \tag{11.3.438}$$

Therefore:

$$g_J \|\vec{J}\|^2 = g_L \vec{L} \circ (\vec{L} + \vec{S}) + g_S \vec{S} \circ (\vec{L} + \vec{S}) = g_L (\|\vec{L}\|^2 + \vec{L} \circ \vec{S}) + g_S \circ (\vec{S} \circ \vec{L} + \|\vec{S}\|^2) \tag{11.3.439}$$

That is often written for simplification purposes:

$$g_J J^2 = g_L (L^2 + \vec{L} \circ \vec{S}) + g_S (\vec{S} \circ \vec{L} + S^2) \tag{11.3.440}$$

Now let us recall that in the section of Wave Quantum Physics during the study of spin-orbit coupling (see page 4900) we have proved that (still simplifying the notation of the norm of a vector):

$$\vec{L} \circ \vec{S} = \frac{1}{2} \left(J^2 - L^2 - S^2 \right) \tag{11.3.441}$$

Therefore:

$$\begin{aligned} g_J J^2 &= g_L(L^2+0.5\left(J^2-L^2-S^2\right))+g_S(0.5\left(J^2-L^2-S^2\right)+S^2) \\ &= 0.5g_L(J^2-L^2-S^2)+0.5g_S(J^2-L^2+S^2) \end{aligned} \tag{11.3.442}$$

Now we introduce the eigenvalues of J^2, L^2 and J^2 as we have determined the during our study of Angular Momentum and Spin (see page 4877) and we simplify by $\hbar^2$ directly and that gives:

$$g_J J(J+1)=0.5g_L(J(J+1)-L(L+1)-S(S+1))+0.5g_S(J(J+1)-L(L+1)+S(S+1)) \tag{11.3.443}$$

Therefore:

$$\boxed{g_J=\frac{g_L(J(J+1)-L(L+1)-S(S+1))+g_S(J(J+1)-L(L+1)+S(S+1))}{2J(J+1)}} \tag{11.3.444}$$

If we put $g_L=1$ and $g_S\cong 2$ (implying that the spin angular momentum is twice as effective in producing a magnetic moment) then the above expression simplifies to:

$$\begin{aligned} g_J &\cong \frac{(J(J+1)-L(L+1)-S(S+1))+2(J(J+1)-L(L+1)+S(S+1))}{2J(J+1)} \\ &= \frac{3J(J+1)+S(S+1)-L(L+1)}{2J(J+1)} \\ &= \frac{3J(J+1)}{2J(J+1)}+\frac{S(S+1)-L(L+1)}{2J(J+1)}=\frac{3}{2}+\frac{S(S+1)-L(L+1)}{2J(J+1)} \end{aligned} \tag{11.3.445}$$

Now following the same reasoning, let us determine in a most robust way the relation:

$$E_p=-\frac{q}{2m_0}(\vec{L}+2\vec{S})\circ\vec{B} \tag{11.3.446}$$

determined earlier above.

The problem with evaluating the scalar product above is that $\vec{L}$ and $\vec{S}$ continually change in direction as shown in the vector model. The strategy for dealing with this problem is to use the direction of the total angular momentum $\vec{J}$ as a coordinate axis and obtain the projection of each of the vectors in that direction. This is done by taking the scalar product of each vector with a unit vector in the $\vec{J}$ direction:

$$\begin{aligned} E_p &= -\frac{q}{2m_0}(\vec{L}+2\vec{S})\circ\left(\vec{B}\frac{J^2}{J^2}\right)=-\frac{q}{2m_0}(\vec{L}+2\vec{S})\circ\left(\vec{B}\frac{\vec{J}\circ\vec{J}}{J^2}\right)=-\frac{q}{2m_0}\left(\frac{(\vec{L}+2\vec{S})}{J}\circ\vec{J}\right)\left(\frac{\vec{J}\circ\vec{B}}{J}\right) \\ &= -\frac{q}{2m_0}\left(\frac{(\vec{L}+2\vec{S})\circ\vec{J}}{J^2}\right)J_zB \end{aligned} \tag{11.3.447}$$

Using the fact that:

$$\vec{J}=\vec{L}+\vec{S} \tag{11.3.448}$$

we get:

$$E_p = -\frac{q}{2m_0}\frac{\left(\vec{L}+2\vec{S}\right)\circ\left(\vec{L}+\vec{S}\right)J_zB}{J^2} \tag{11.3.449}$$

These vector relationships must be evaluated and expressed in terms of quantum numbers in order to evaluate the energy shifts. Carrying out the scalar products above leads to:

$$E_p = -\frac{q}{2m_0}\frac{\left(L^2+2S^2+3\vec{L}\circ\vec{S}\right)J_zB}{J^2} = -\frac{q}{2m_0}\frac{\left(L^2+2S^2+3\vec{L}\circ\vec{S}\right)m_jB}{J^2} \tag{11.3.450}$$

And we have also proved in the section of Wave Quantum Physics that:

$$J^2 = L^2+S^2+2\vec{L}\circ\vec{S} \tag{11.3.451}$$

Rearranging gives the scalar product term needed:

$$3\vec{L}\circ\vec{S} = \frac{3}{2}(J^2-L^2-S^2) \tag{11.3.452}$$

Substitution into the energy expression gives:

$$E_p = -\frac{q}{2m_0}\frac{\left(L^2+2S^2+\frac{3}{2}(J^2-L^2-S^2)\right)m_jB}{J^2} \tag{11.3.453}$$

which can be reduced to:

$$E_p = -\frac{q}{2m_0}\frac{(3J^2-L^2+S^2)\,m_jB}{2J^2} = -\mu_B g_L m_j B \tag{11.3.454}$$

where using eigenvalues we have:

$$g_L = \frac{3J(J+1)-L(L+1)+S(S+1)}{2J(J+1)} = \frac{3}{2}+\frac{S(S+1)-L(L+1)}{2J(J+1)} \tag{11.3.455}$$

so we fall back on the approximation of g_J!

So finally:

$$\boxed{E_p = -\frac{q}{2m_0}\left(\frac{3}{2}+\frac{S(S+1)-L(L+1)}{2J(J+1)}\right)m_jB = -\mu_B g_L m_j B \cong -\mu_B g_J m_j B} \tag{11.3.456}$$

✍**Example:**

Let us give a word about the fine structure of sodium atoms.

As we have seen in the section of Corpuscular Quantum Physics and following the aufbau principle (also introduced in the same section) for the atom of Sodium Na we have seen that its electronic configuration was:

$$1s^2 2s^2 2p^6 3s \tag{11.3.457}$$

Explicitly we then say (following orbital-spin coupling notation), that is ground state is then:

$$n^{2s+1}l_j = 3^2 s_{1/2} \tag{11.3.458}$$

The first excited state is obtained (still following the aufbau principle) by promoting the $3s$ electron to the $3p$ orbital, to obtain the:

$$1s^2 2s^2 2p^6 3p \tag{11.3.459}$$

configuration. In a sodium-vapour lamp for example, sodium atoms are excited to the $3p$ level by an electrical discharge, and return to the ground state by emitting yellow light of wavelength 589 [nm].

Now let us recall the following table (see section Wave Quantum Physics page 4889):

Orbital type	***s***	***p***	***d***	***f***	**...**
l	0	1	2	3	...
j	$\dfrac{1}{2}$	$\dfrac{1}{2},\dfrac{3}{2}$	$\dfrac{3}{2},\dfrac{5}{2}$	$\dfrac{5}{2},\dfrac{7}{2}$	...
Notation:	$s_{1/2}$	$p_{1/2},p_{3/2}$	$d_{3/2},d_{5/2}$	$f_{5/2},f_{7/2}$	...

Therefore for the ground state $3^2 s_{1/2}$ of the Sodium we get $l=0$ and $j=1/2$. And in the absence of magnetic field $B=0$ we have:

$$E_p = -\mu_B g_L m_j B = -\mu_B g_L m_j \cdot 0 = 0 \tag{11.3.460}$$

so there is no fine structure of the Sodium spectrum.

But when the electron move to $3p$ under the influence of a weak magnetic field, we have (still look to the table above) that $l=1$ with $j=1/2$ and $j=3/2$.

Therefore a fine structure appears for $3p$ and we get that the transition $3s \rightarrow 3p$ is split into a doublet that we name for Sodium the "D-lines" and respectively each line is named D_1 and D_2 lines.

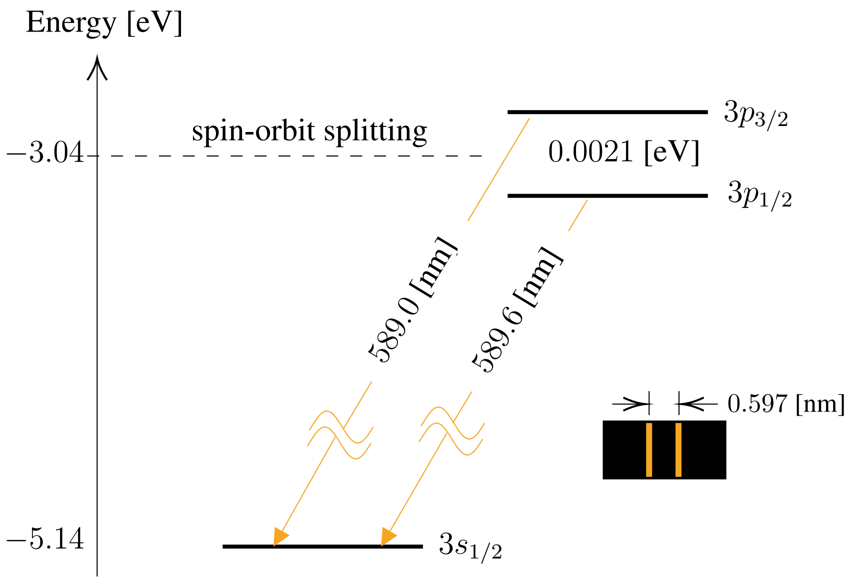


Figure 11.85: Example of fine structure in Sodium (Source: Hyperphysics)

The fact that there is a doublet shows the smaller dependence of the atomic energy levels on the total angular momentum. This effect is called the "spin-orbit effect".

In fact this doublet are also visible without the presence of an additional externally applied magnetic field, but with a magnetic fields these levels are further split by the magnetic interaction (and therefore more easily measurable), showing dependence of the energies on the z-component of the total angular momentum. This splitting gives the Zeeman effect for sodium.

Now let us roughly approximate the magnetic field necessary to create this doublet by taking:

$$|E_p| = \mu_B g_L m_j B = 0.0021\,[\text{eV}] \tag{11.3.461}$$

Assuming[a] same g_L and same m_j we get $B = 18$ [T]. This is a very large magnetic field by laboratory standards. Large magnets with dimensions over a meter, used for NMR and ESR experiments, have magnetic fields on the order of a Tesla.

Experimentally (therefore in reality) it was the opposite, the field was applied and the g_L were determined as we get:

Term	J	L	S	Experimental g_L	Calculated g_L
$\mathbf{3p_{3/2}}$	$3/2$	1	$1/2$	$4/3$	$4/3$
$\mathbf{3p_{1/2}}$	$1/2$	1	$1/2$	$2/3$	$2/3$
$\mathbf{3s_{1/2}}$	$1/2$	0	$1/2$	2	2

Table 11.12: g_L Lande' g-factor for Sodium Na

where g_L was calculated using:

$$g_L = \frac{3}{2} + \frac{S(S+1) - L(L+1)}{2J(J+1)} \tag{11.3.462}$$

---

[a] In fact even if we calculate precisely m_j for each orbit we get almost the same result with less than 10% error

However.... what we have seen above is the way the Landé g-factor is derivend an expressed most of the times in european universities. In most american universities the expression is different because the derivation is based on a different choice at the beginning of the proof. Let us see how!

The magnetic moment of an atom comes from two parts: orbital motion $(\vec{\mu}_L)$ and electron spin $(\vec{\mu}_S)$:

$$\vec{\mu}=\vec{\mu}_L+\vec{\mu}_S=-g_L\mu_B\frac{\vec{L}}{\hbar}-g_S\mu_B\frac{\vec{S}}{\hbar} \tag{11.3.463}$$

where μ_B is the Bohr magneton, $g_L\cong 1$, and $g_S\cong 2$.

The total angular momentum is the vector sum of orbital and spin angular momenta:

$$\vec{J}=\vec{L}+\vec{S} \tag{11.3.464}$$

Because of spin-orbit coupling, L and S rapidly precess around the constant total angular momentum vector $\vec{J}$. Only the component of $\vec{\mu}$ parallel to $\vec{J}$ has a non-zero time average:

$$\vec{\mu}_{\text{effective}}=\left(\frac{\vec{\mu}\circ\vec{J}}{J^2}\right)\vec{J}=-g_J\mu_B\frac{\vec{J}}{\hbar} \tag{11.3.465}$$

We substitute $\vec{\mu}$ into the dot product projection:

$$\vec{\mu}\cdot\vec{J}=-\mu_B\left(g_L\vec{L}\circ\vec{J}+g_S\vec{S}\circ\vec{J}\right) \tag{11.3.466}$$

To evaluate $\vec{L}\circ\vec{J}$ and $\vec{S}\circ\vec{J}$, we use the relations $\vec{L}=\vec{J}-\vec{S}$ and $\vec{S}=\vec{J}-\vec{L}$ and we square $\vec{S}=\vec{J}-\vec{L}$ to get:

$$S^2=J^2+L^2-2(\vec{J}\circ\vec{L}) \tag{11.3.467}$$

This gives the dot product:

$$\vec{L}\circ\vec{J}=\frac{J^2+L^2-S^2}{2} \tag{11.3.468}$$

Similarly, squaring $\vec{L}=\vec{J}-\vec{S}$ gives:

$$\vec{S}\circ\vec{J}=\frac{J^2+S^2-L^2}{2} \tag{11.3.469}$$

Replacing the operator squares with their quantum mechanical eigenvalues ($J^2\rightarrow J(J+1)\hbar^2, L^2\rightarrow L(L+1)\hbar^2, S^2\rightarrow S(S+1)\hbar^2$) and using $g_L=1$ and $g_S\cong 2$ yields to:

$$g_J=\frac{g_L[J(J+1)+L(L+1)-S(S+1)]+g_S[J(J+1)+S(S+1)-L(L+1)]}{2J(J+1)} \tag{11.3.470}$$

Setting $g_L \cong 1$ and $g_S \cong 2$ simplifies this directly to the "Landé g-factor" expression" or "splitting factor" (which determines the splitting of the energy levels in the presence of weak external magnetic field) as we can found it in most american textbooks:

$$\boxed{g_J \cong 1 + \frac{J(J+1)+S(S+1)-L(L+1)}{2J(J+1)}} \tag{11.3.471}$$

Also often denoted more technically:

$$\begin{aligned} g_J \left\langle J, J_z \right| \vec{J}\cdot\vec{J} \left| J, J_z \right\rangle &= \left\langle J, J_z \right| g_L \vec{L}\cdot\vec{J} + g_S \vec{S}\cdot\vec{J} \left| J, J_z \right\rangle \\ &= \left\langle J, J_z \right| g_L \left[\vec{L}^2 + \frac{1}{2}\left(\vec{J}^2 - \vec{L}^2 - \vec{S}^2\right)\right] + g_S \left[\vec{S}^2 + \frac{1}{2}\left(\vec{J}^2 - \vec{L}^2 - \vec{S}^2\right)\right] \left| J, J_z \right\rangle \\ &= \frac{g_L \hbar^2}{2}[J(J+1)+L(L+1)-S(S+1)] + \frac{g_S \hbar^2}{2}[J(J+1)-L(L+1)+S(S+1)] \\ g_J &= g_L \frac{J(J+1)+L(L+1)-S(S+1)}{2J(J+1)} + g_S \frac{J(J+1)-L(L+1)+S(S+1)}{2J(J+1)} \end{aligned} \tag{11.3.472}$$

It is very important to note that for singlet states the spin quantum number $S=0$ so that $J=L$ and the Landé g-factor expression becomes equal to the gyromagnetic ratio $=2$. Hence the above formulation reduces to that of the "normal Zeeman effect" which was explained earlier in terms of only the orbital angular momentum and its associated magnetic moment.

The following table gives the calculated Lande g-factors for some common term symbols in the approximation $g_S = 2$:

Term symbol	S	L	J	g_J
$^2\mathrm{S}_{1/2}$	1/2	0	1/2	2
$^3\mathrm{S}_1$	1	0	1	2
$^4\mathrm{S}_{3/2}$	3/2	0	3/2	2
$^5\mathrm{S}_2$	2	0	2	2
$^6\mathrm{S}_{5/2}$	5/2	0	5/2	2
$^7\mathrm{S}_3$	3	0	3	2
$^1\mathrm{P}_1$	0	1	1	1
$^2P_{1/2}$	1/2	1	1/2	2/3
$^2P_{3/2}$	1/2	1	3/2	4/3
$^3\mathrm{P}_0$	1	1	0	-
$^3\mathrm{P}_1$	1	1	1	3/2
$^3\mathrm{P}_2$	1	1	2	3/2
$^4\mathrm{P}_{1/2}$	3/2	1	1/2	8/3
$^4\mathrm{P}_{3/2}$	3/2	1	3/2	26/15
$^4P_{5/2}$	3/2	1	5/2	8/5
$^5\mathrm{P}_1$	2	1	1	5/2
$^5\mathrm{P}_2$	2	1	2	11/6
$^5\mathrm{P}_3$	2	1	3	5/3
$^6\mathrm{P}_{3/2}$	5/2	1	3/2	12/5
$^6\mathrm{P}_{5/2}$	5/2	1	5/2	66/35
$^6\mathrm{P}_{7/2}$	5/2	1	7/2	12/7
$^7\mathrm{P}_2$	3	1	2	7/3
$^7\mathrm{P}_3$	3	1	3	23/12
$^7\mathrm{P}_4$	3	1	4	7/4
$^1\mathrm{D}_2$	0	2	2	1
...	...	...	...	...

Remark

According to three of the most standard quantum mechanics textbooks (Griffiths, Merzbacher and Peskin), as well as most (though not quite all) publications we can find, the term "Landé g-factor" refers specifically to g_J, while the term "g-factor" or "g-value" refers to the more general case of any dimensionless factor in an equation for a magnetic moment (not specifically related to the electron!).

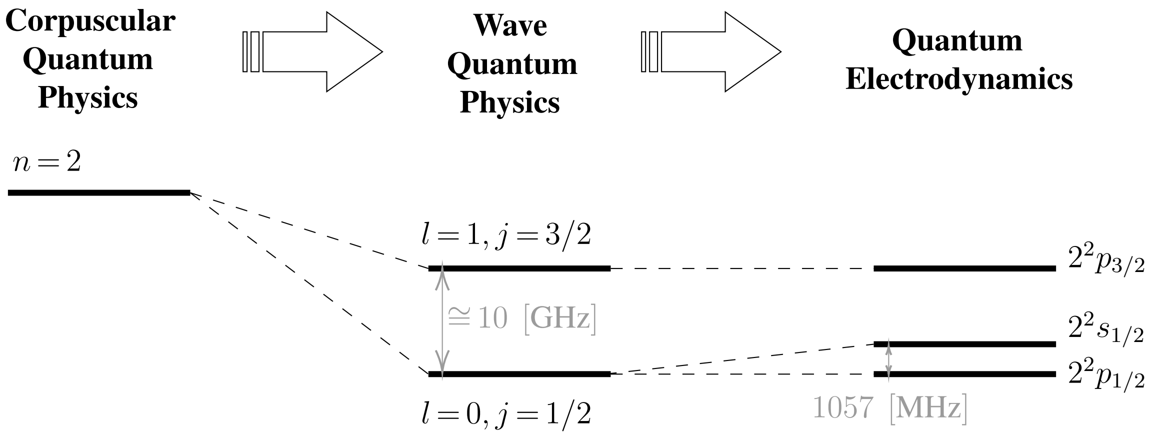
11.3.6.2 Lamb Shift

But this is not the end with our fine structure...

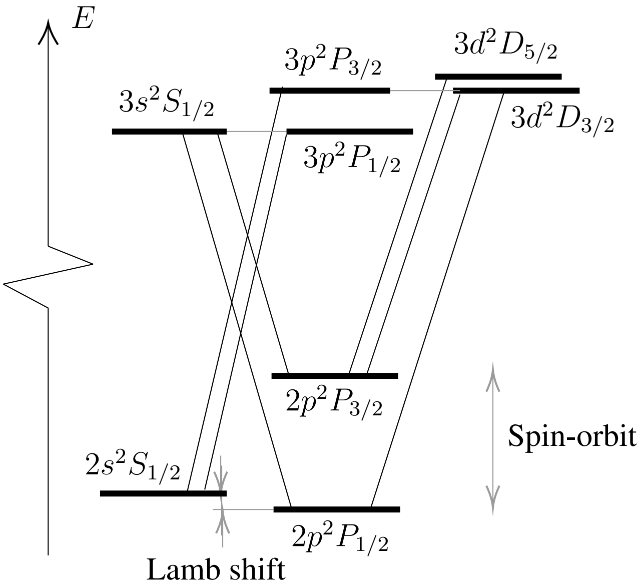
Indeed, if we apply what we have seen so far to the Hydrogen atom when the electron goes to $n=1$ to $n=2$ (without magnetic field but using ad hoc excitation!), we know that for the layers s and p:

$$\begin{aligned} l=0 &\Rightarrow j=1/2 \\ l=1 &\Rightarrow j=1/2,3/2 \end{aligned} \tag{11.3.473}$$

and therefore we get with the Bohr model, and the Dirac Model and in laboratories:


Corpuscular Quantum Physics
Wave Quantum Physics
Quantum Electrodynamics
$n=2$
$l=1, j=3/2$
$\cong 10$ [GHz]
$l=0, j=1/2$
$2^2p_{3/2}$
$2^2s_{1/2}$
$2^2p_{1/2}$
1057 [MHz]

Or in a more technical and traditional view:


E
$3p^2P_{3/2}$
$3d^2D_{5/2}$
$3s^2S_{1/2}$
$3d^2D_{3/2}$
$3p^2P_{1/2}$
$2p^2P_{3/2}$
Spin-orbit
$2s^2S_{1/2}$
$2p^2P_{1/2}$
Lamb shift

So here there is an issue between theory and experiment because in labs it is clear that $2^2s_{1/2}$ and $2^2p_{1/2}$ are not degenerated (that means in other words, for recall, they don't have the same energy level) when theory gives us clearly that they should be degenerated (i.e. have the same energy levels)!

Indeed, let us recall that for $j=1/2$ we have:

$$j=l+s=\frac{1}{2} \tag{11.3.474}$$

So all the possible authorized pairs that lead to this result are:

$$\begin{aligned}j&=l+s=0+\frac{1}{2}=\frac{1}{2}\\ j&=l+s=1-\frac{1}{2}=\frac{1}{2}\end{aligned} \tag{11.3.475}$$

Therefore we have indeed that $s_{1/2}$ and $p_{1/2}$ that should be degenerated but that are not experimentally whatever the level n.

Indeed, the (energy) degeneration doesn't come in this case from the term $(\vec{L}+2\vec{S})\circ\vec{B}$ as $\vec{B}\cong\vec{0}$ in the Lamb experiment. We must just apply $\vec{J}=\vec{L}+\vec{S}$ in this case!

The (experimentally) measured positions show the $2s^2p_{1/2}$ level is shifted upwards relative to the position $(n=2, j=1/2)$ and is therefore not degenerate with the $2p^2p_{1/2}$ level. Such a shift occurs for all the s-electrons in laboratories (but the energy shift decreases with increasing n).

The most right splitting can only be explained by Quantum Electrodynamics (QED) and is known as the "Lamb shift" of the s level and denoted "Ly-α" in the special case of hydrogen. Indeed, according to the Dirac equation, these both sates should have the same energy. This shift was first observed in 11947-11952 according to holocene calendar (Lamb–Retherford experiment[29]) and was crucial for the development of QED where the electromagnetic field is quantised and has a zero point energy (this means that even in the case of number of photons equal to zero, there is a fluctuating electric field). In 11948 (holocene calendar), Theodore A. Welton succeeded (reasonably after a lot of assumptions and approximations...) in accounting for the Lamb shift between the two above levels in terms of the perturbation of the electronic orbit brought about by vacuum fluctuations.

Due to this fluctuating field the electron performs small oscillations and its charge is effectively smeared out over a range of about 0.1 fermi. This causes the electron spin g-factor to be slightly different from 2. There is also a slight weakening of the force on the electron when it is very close to the nucleus, causing the $2s_{1/2}$ electron (which has penetration all the way to the nucleus) to be slightly higher in energy than the $2p_{1/2}$ electron.

When the Lamb shift was experimentally determined, it provided a high precision verification of theoretical calculations made with the quantum theory of electrodynamics. These calculations predicted that electrons continually exchanged photons, this being the mechanism by which the electromagnetic force acted. The effect of the continuous emission and absorption of photons on the electron g-factor could be calculated with great precision.

The tiny Lamb shift, measured with great precision, agreed to many decimal places with the calculated result from quantum electrodynamics. The measured precision gives us the electron spin g-factor as:

$$g_e\cong 2.00231930436256(35) \tag{11.3.476}$$

compared to the theoretical prediction of $2.00233183620(86)4$ using QED (Quantum Electrodynamics).

Since 11956 (holocence calendar) the Dyson series (from Freeman Dyson) is used to calculate the exact value of the g factor of the electron (maybe in the future we will have the time to prove where that series comes from):

$$g_e=2\cdot\left(1+C_1\left(\frac{\alpha}{\pi}\right)+C_2\left(\frac{\alpha}{\pi}\right)^2+C_3\left(\frac{\alpha}{\pi}\right)^3+C_4\left(\frac{\alpha}{\pi}\right)^4+C_5\left(\frac{\alpha}{\pi}\right)^5\ldots\right) \tag{11.3.477}$$

---

[29] They carried out the experiment using microwave techniques to stimulate radio-frequency transitions between the both level of interest. The energy difference Lamb and Rutherford found was a rise of about 1000 [MHz]

where each constant C_i corresponds to the calculation of a given number of Feynman diagrams (involving complex curvilinear integrals whose exact calculations/numerical calculation are kept confidential because of publication competitions between scientists and whose calculations are a bit awkward because involving divergent infinite integrals that need to be renormalized to give finite values...) and alpha is the fine-structure constant. The expression between the parenthesis is named the "magnetic anomaly" or the "anomalous magnetic moment".

In many scientific papers the magnetic anomaly is also defined as:

$$a=\frac{g_e-2}{2} \tag{11.3.478}$$

When we write these lines actual calculations give (C_2 needed the computation of 72 Feynman diagrams, C_4 needed the computation of 891 Feynman diagrams, C_5 need in year 12018 the computation of $12,726$ Feynman integrals on supercomputers...):

$$\begin{aligned}g_e&=2\cdot\left(1+\frac{\alpha}{2\pi}-0.328\left(\frac{\alpha}{\pi}\right)^2+1.181\left(\frac{\alpha}{\pi}\right)^3-1.912\left(\frac{\alpha}{\pi}\right)^4+7.795\left(\frac{\alpha}{\pi}\right)^5+\ldots\right)\\&\cong 2\cdot 1.001159652182032=2.00231930436406\end{aligned} \tag{11.3.479}$$

where the first value $\alpha/2\pi$ is the named the "Schwing factor" (it's even engraved above his name on his tombstone...). For more informations on this topic see [110].

As reported in *Measurement of the Electron Magnetic Moment* (see [171]) published in 12023 (holocene calendar), the electron magnetic moment, $-\mu/\mu_B=g/2=1.00115965218059(13)$ [0.13 ppt]. It is the most precisely determined property of an elementary particle by the Standard Model to 1 part in 10^{12}.

The Lamb shift has since played a significant role through vacuum energy fluctuations in theoretical prediction of Hawking radiation from Black Holes.

The muon magnetic anomaly denoted a_μ is very interesting as it leads to some stress in the physics community in this early 121st century (holocene calendar). Indeed, if we look at the figure below (credit adapted from A. Boccaletti et al. 2024 arXiv:2407.10913):

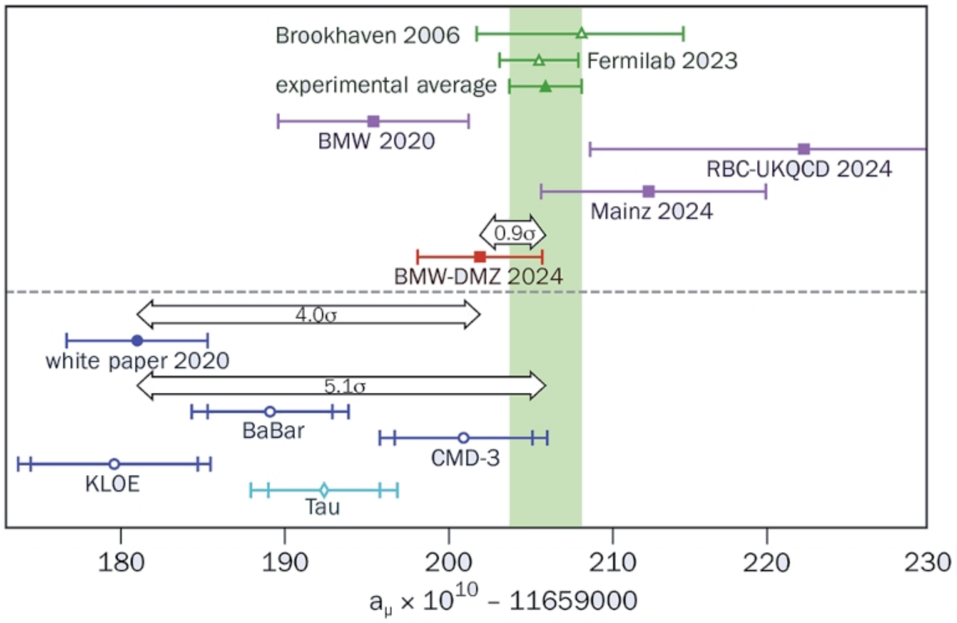


Figure 11.86: **Landscape of muon $g-2$ Measurements of a_μ (source: CERN)**

measurements of a_μ (green triangles) compared to the Standard Model predictions[30] using lattice Quantum Chromodynamics (purple squares), data-driven predictions using e^+e^- annihilation to hadrons (blue circles), tau-lepton decays (cyan diamond) and a hybrid approach (red square) show statistical compatibility or incompatibility as indicated by arrows between the latest experimental average (filled green triangle) in 12025 (holocene calendar), a 12020 (holocene calendar) white paper (filled blue circle) and the most precise prediction (red square) also in 12025 (holocene calendar).

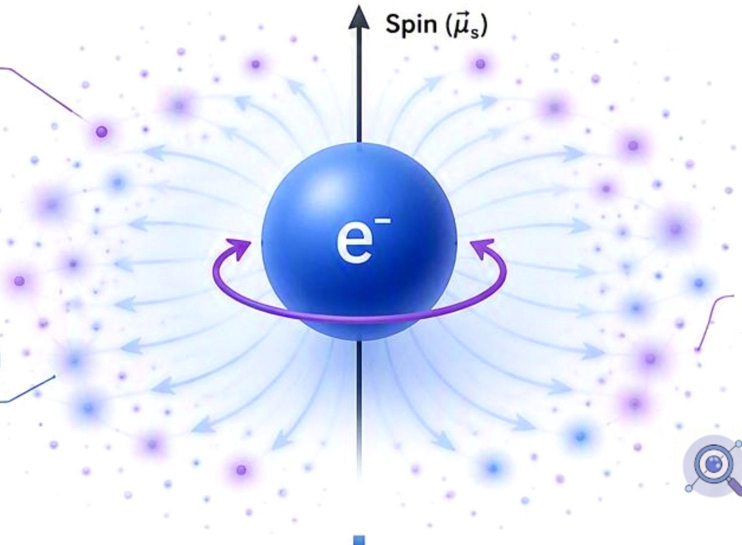
These (apparent?) incompatibilities (tension) may be an opening for a new type of physics beyond the Standard Model. To follow!...

---

[30] It took to the Budapest–Marseille–Wuppertal collaboration (BMW) a billion core-hours on the Mira supercomputer to compute the hadronic light-by-light contribution to the muon's magnetic moment in 12020 (holocene calendar).

The electron is not a point-like magnet as Dirac's simple theory predicted. Its interaction with the quantum vacuum makes its magnetic moment slightly larger than expected. This tiny difference is the **anomalous magnetic moment (a_e)**.

A SPINNING ELECTRON DISTURBS THE QUANTUM VACUUM



Quantum vacuum fluctuations

Pairs of virtual particles constantly appear and disappear.

Electron's electric field

The electron's charge creates an electric field that polarizes the vacuum and alters these fluctuations.

The Dirac Prediction: g = 2

$$g = 2$$

In 1928, the Dirac equation predicted the electron's magnetic g-factor should be exactly two.

Distorted vacuum

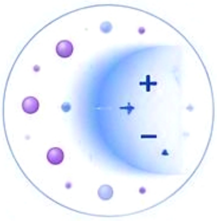
The electric field distorts the distribution of virtual particles around the electron.

The Observed "Anomaly"

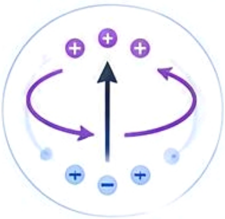
0.12%

Precision measurements showed a 0.12% excess, now known as the anomalous magnetic moment.

HOW THIS CREATES THE ANOMALOUS MAGNETIC MOMENT



The electric field polarizes the vacuum, causing virtual charges to be distributed asymmetrically around the electron.



The electron's rotation makes this polarization circulate around it, producing an extra magnetic field in the same direction as its spin.

μ_{total}
>
μ_{Dirac}

This additional field adds to the Dirac prediction, making the electron's magnetic moment slightly larger.

ANOMALOUS MAGNETIC MOMENT

$$a_e = \frac{g-2}{2}$$

The anomalous magnetic moment (a_e) is the fingerprint of the electron's interaction with the quantum vacuum. It arises from the constant disturbance of vacuum fluctuations by the electron's electric field.

Measured with extraordinary precision and in excellent agreement with the Standard Model of particle physics.

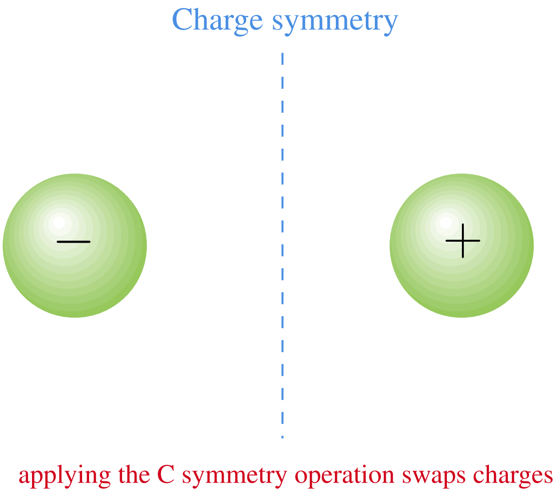
11.3.7 CPT Symmetries

When a fundamental theory or idea is proposed, there is a way to verify if it should be considered for further study or if it is just nonsense. See this like a kind of quality control protocol. The fundamental criterion to be considered is the in-variance of the proposed theory under certain symmetries. That means that if I apply a specific symmetry to a reaction the theory predicts and that is observed in nature, the resulting reaction should also be observable in Nature.

There are three fundamental symmetries that are to be considered: C (Charge), P (Parity) and T (Time reversal). The symmetry operators can combine, for example the CP operation consists in applying the C-symmetry and the P- symmetry simultaneously. The ultimate test for the invariance of a theory is to apply on it the CPT symmetry. It is not an easy concept to grasp, so let us give it a try by going into some mathematical details.

11.3.7.1 Charge symmetry

The charge symmetry is an operation where the electric charges of particles are swapped:



The C-symmetry transform the sign of a charge but not its magnitude. Consequently, the electromagnetic forces remain the same. Take two electrons, each charged negatively with the elementary charge $-e$. Place them close to each other: they repel each other with a given force.

Now apply the C-Symmetry to this system. Both particles will have the sign of their charge swapped to a positive charge $+e$. Other than this, nothing changes: they still repel with the same force...

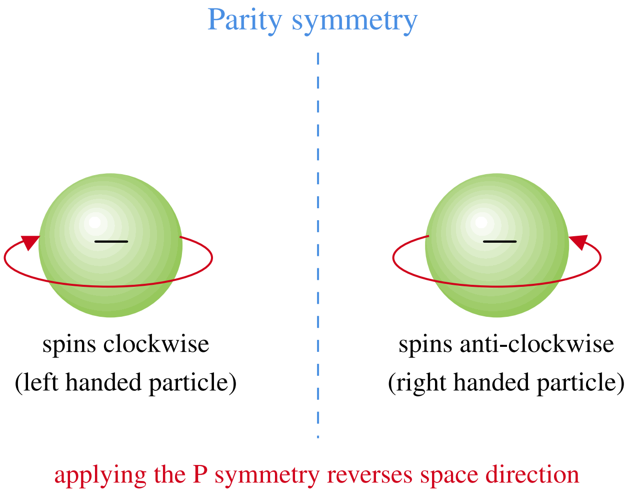
You would expect that C-symmetry is always respected... Well, counter-intuitively, this is not the case. Not all reaction in our universe are C-symmetric.

Remark

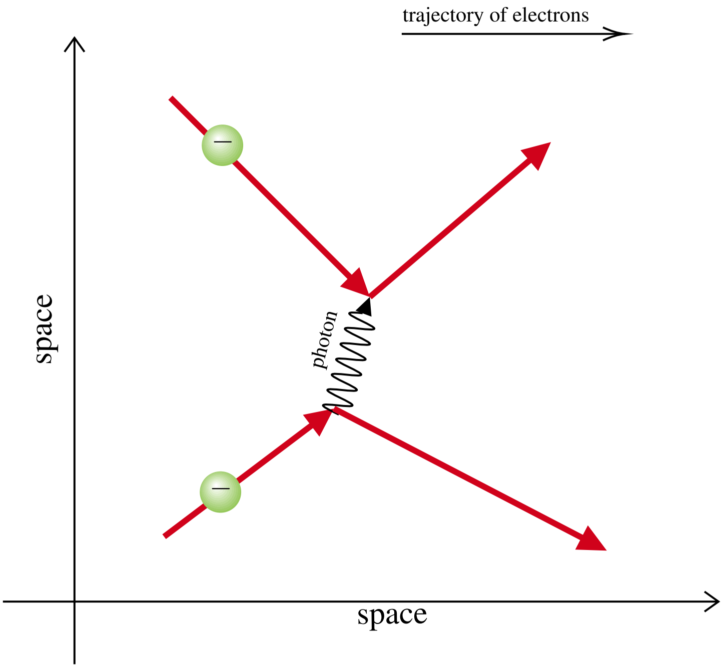
The C symmetry, as we will see it further below, doesn't really mean we change the electric charge. Furthermore it doesn't only applied to the "electric charge" but also to other kind of charges in quantum field theory like the color charge for example!

11.3.7.2 Parity symmetry

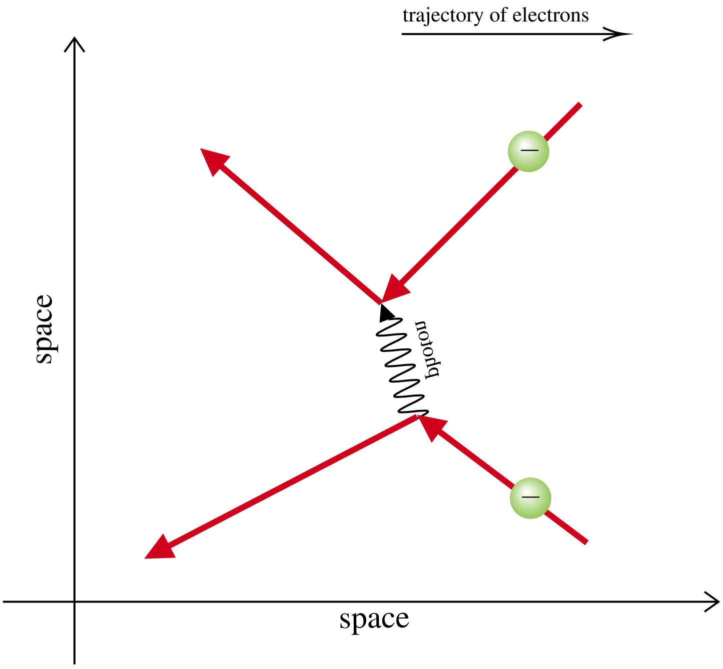
The parity symmetry is an operation where the directions in space are reversed: In the diagram below is represented the spin of electrons. the P-symmetry changes the direction of that spin from clockwise to anti-clockwise and vice-versa:


Parity symmetry
—
—
spins clockwise
(left handed particle)
spins anti-clockwise
(right handed particle)
applying the P symmetry reverses space direction

Let's see what we mean when we say that a reaction is P-symmetric. This symmetry appears obvious to common sense but is not that obvious when digging into small scale physics. Imagine a reaction occurring in nature. Say, two electrons have a trajectory such that they are approaching each other, while both moving from left to right. An exchange of photons between them will modify their momentum as in the drawing below:


trajectory of electrons
space
space
photon
−
−

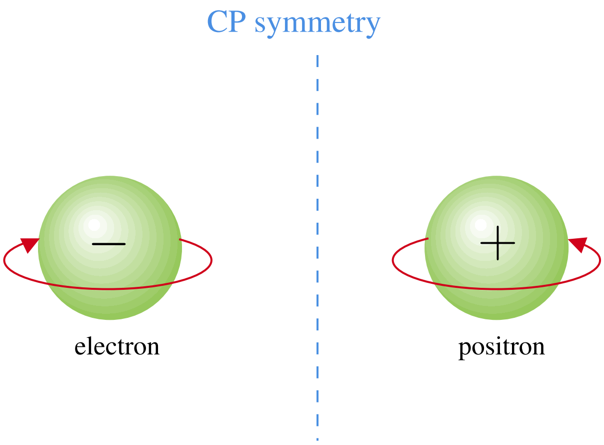
If a reaction as seen in a mirror exists in nature, then it is said to be P-Symmetric. The P symmetric reaction involving our two electrons would correspond to the same phenomena but with the electrons moving from right to left this time...


trajectory of electrons
space
space
photon
−
−

You would expect P symmetry to be respected by all phenomena in nature. Again counter-intuitively this is not the case either.

11.3.7.3 CP symmetry (from Matter to Antimatter)

Now let's check out what happens when we apply both C and P symmetry operations on a real particle:


CP symmetry
−
electron
+
positron

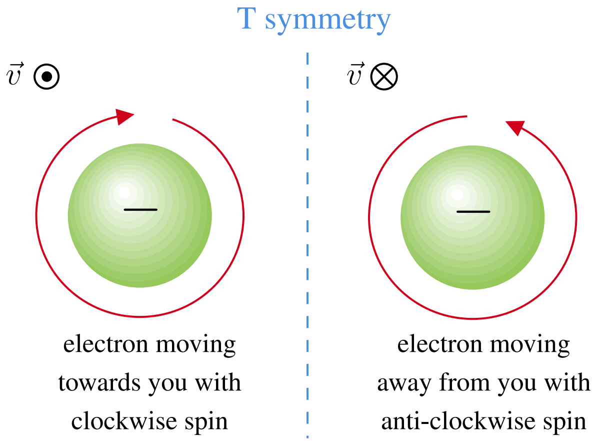
applying the CP symmetry operation
transforms matter into antimatter

As the reader may observe, combining C and P symmetries transforms our electron into... a positron, the anti-matter counterpart of the electron.

Consequently if nature respects CP symmetry, that implies that matter and antimatter should have a perfect symmetric behavior...

11.3.7.4 Time Reversal symmetry

The time symmetry is an operation where the direction in time is reversed. Imagine you record on your fancy camera, an electron coming towards you while spinning clockwise. Then, you watch the movie in reverse: what you observe is the electron going away from you while rotating anticlockwise...:


T symmetry
$\vec{v}\ \odot$
−
electron moving
towards you with
clockwise spin
$\vec{v}\ \otimes$
−
electron moving
away from you with
anti-clockwise spin

applying the T symmetry operation
is like playing a movie in reverse

The reader can also think of what it means for a reaction to be T-symmetric. Imagine that you record with your video camera an interaction between particles: say two incident particles A and B with given momentum and energy meet up, interact and form two new particles C and D with their respective momentum and energy:

$$\mathrm{A}+\mathrm{B}\rightarrow\mathrm{C}+\mathrm{D}$$

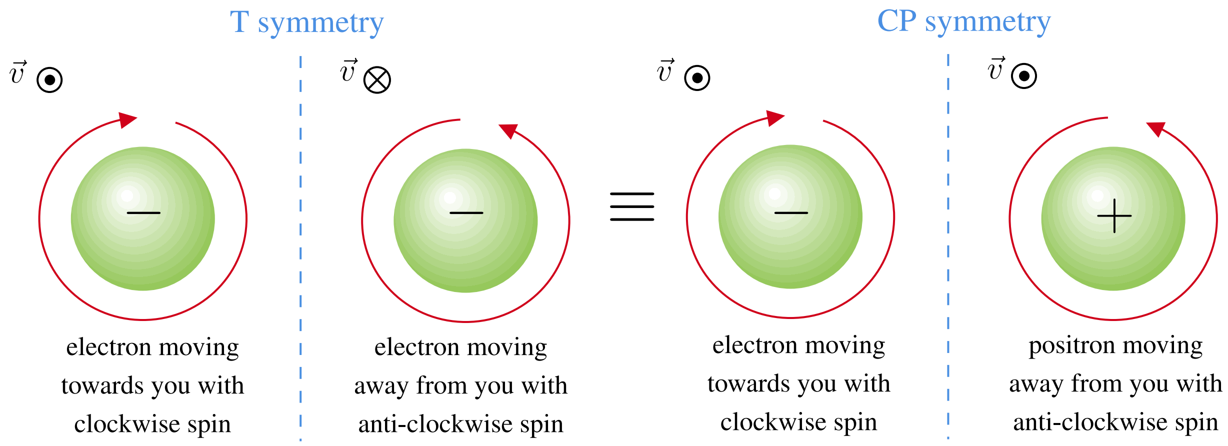
Now play your video backwards, you should observe particle C and D meeting up and as a result of their interaction, produce particles A and B with the same momentum and energy they had in the forward reaction. In addition, the probability of occurrence of the time-reversed reaction should match with that of the forward reaction:

$$\mathrm{C}+\mathrm{D}\rightarrow\mathrm{A}+\mathrm{B}$$

If such a time reversed reaction can be observed (or theoretically provoked) in nature, it implies that this reaction is T-Symmetric.

11.3.7.5 CP and T symmetries

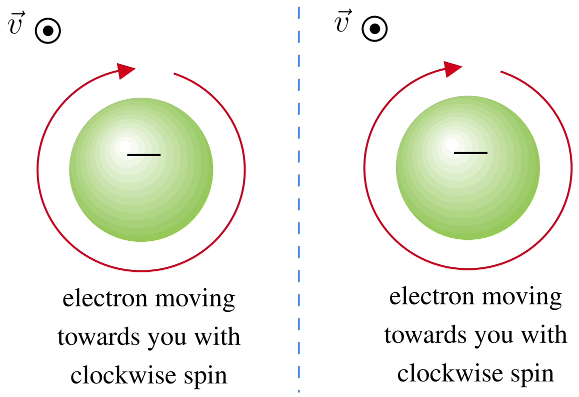
We mentioned that T symmetry transforms a negative charge going towards the reader into a negative charge moving away from the reader: The latter can be considered equivalent as a positive charge moving toward you (the rate of change of the electric potential at the observer's position will be identical in both cases). This equivalence is massively used in electronics and electrical sciences. Thus we could approximate that the T symmetry transforms an electron into a positron:


T symmetry
CP symmetry
$\vec{v}$ ⊙
$\vec{v}$ ⊗
$\vec{v}$ ⊙
$\vec{v}$ ⊙
−
−
≡
−
+
electron moving towards you with clockwise spin
electron moving away from you with anti-clockwise spin
electron moving towards you with clockwise spin
positron moving away from you with anti-clockwise spin

Looks kind of familiar, don't you think?

Yes, it looks furiously like the CP symmetry operation. Both CP and T appear complementary...

So combining all three C, P and T operations simultaneously ends up with an invariance: when applied on an electron spinning clockwise and going towards us, the CPT symmetry produces an electron spinning clockwise and going towards us :


CPT symmetry
$\vec{v}$
$\vec{v}$
electron moving towards you with clockwise spin
electron moving towards you with clockwise spin

The CPT symmetry operation is invariant

This invariance of the CPT symmetry is a pillar in Physics. The violation of this symmetry has never been observed experimentally. If a theory does not respect it, then it is automatically invalidated.

It has been experimentally observed that Nature violates C and P symmetries (see examples further below). as well as CP and T symmetries (see examples further below). However, when CP is violated, it is accompanied by a violation of T which compensates so that CPT remains invariant.

11.3.7.6 CPT violations

Also before we deal with the mathematical detailed aspect of this topic, let us do a summary of the experimental results about the violations of the previously mentioned symmetries.

In particle physics, "CP violation" is a violation of CP-symmetry (or charge conjugation parity symmetry): the combination of C-symmetry (charge conjugation symmetry[31]) and P-symmetry (parity symmetry). CP-symmetry states that the laws of physics should be the same if a particle is interchanged with its antiparticle (C-symmetry) while its spatial coordinates are inverted ("mirror" or P-symmetry). The first indirect sign of such a violation was a discovery of CP violation in 11964 (holocene calendar) in the decays of neutral kaons[32] (type of neutral meson) resulted in the Nobel Prize in Physics in 11980 (holocene calendar) for its discoverers James Cronin and Val Fitch.

It plays an important role both in the attempts of cosmology to explain the dominance of matter over antimatter in the present universe, and in the study of weak interactions in particle physics.

Until the 11950s (holocene calendar), parity conservation was thought to be one of the fundamental geometric conservation laws (along with conservation of energy and conservation of momentum). However, in 1956 a careful critical review of the existing experimental data by theoretical physicists Tsung-Dao Lee and Chen-Ning Yang revealed that while parity conservation had been verified in decays by the strong or electromagnetic interactions, it was untested in the weak interaction. They proposed several possible direct experimental tests. The first test based on beta decay of Cobalt-60 nuclei was carried out in 11956 (holocene calendar) by a group led by Chien-Shiung Wu, and demonstrated conclusively that weak interactions violate the P-symmetry or, as the analogy goes, some reactions did not occur as often as their mirror image[33] (however, parity symmetry still appears to be valid for all reactions involving electromagnetism and strong interactions).

---

[31] There is no C symmetry alone, however, there are CPT, CP, etc. symmetries, and for this reason the C operator is required, among other uses!

[32] There are two types of Neutral Kaons that naturally oscillate which means that they transform continuously from one type into the other: $K_0(d\bar{s}) \Leftrightarrow \bar{K}_0(\bar{d}s)$. In this oscillation process, a violation of CP symmetry occurs because the probability of a transformation from one type to another is slightly different than the reverse transformation.

After the discovery of parity violation in 11956 (holocene calendar), CP-symmetry was proposed to restore order. However, while the strong interaction and electromagnetic interaction seem to be invariant under the combined CP transformation operation, further experiments showed that this symmetry is slightly violated during certain types of weak decay. Indeed, in 11956 (holocene calendar) Reinhard Oehme in a letter to Chen-Ning Yang and shortly after, Boris L. Ioffe, Lev Okun and A. P. Rudik showed that the parity violation meant that charge conjugation invariance must also be violated in weak decays. Charge violation was confirmed in the Wu experiment (11957) and in experiments performed by Valentine Telegdi and Jerome Friedman and Garwin and Lederman who observed parity non-conservation in pion and muon decay and found that C is also violated. Charge violation was more explicitly shown in experiments done by John Riley Holt at the University of Liverpool.

In 11964 (holocene calendar), James Cronin, Val Fitch and coworkers provided clear evidence from kaon decay that CP-symmetry could be broken. This work won them the 11980 (holocene calendar) Nobel Prize. This discovery showed that weak interactions violate not only the charge-conjugation symmetry C between particles and antiparticles and the P or parity, but also their combination. The discovery shocked particle physics and opened the door to questions still at the core of particle physics and of cosmology today. The lack of an exact CP-symmetry, but also the fact that it is so close to a symmetry, introduced a great puzzle.

The kind of CP violation discovered in 11964 (holocene calendar) was linked to the fact that neutral kaons can transform into their antiparticles (in which each quark is replaced with the other's antiquark) and vice versa, but such transformation does not occur with exactly the same probability in both directions; this is called indirect CP violation.

Despite many searches, no other manifestation of CP violation was discovered until the 1990s, when the NA31 experiment at CERN suggested evidence for CP violation in the decay process of the very same neutral kaons (direct CP violation). The observation was somewhat controversial, and final proof for it came in 1999 from the KTeV experiment at Fermilab and the NA48 experiment at CERN!

Starting in 12001 (holocene calendar), a new generation of experiments, including the BaBar experiment at the Stanford Linear Accelerator Center (SLAC) and the Belle Experiment at the High Energy Accelerator Research Organisation (KEK) in Japan, observed direct CP violation in a different system, namely in decays of the B mesons. A large number of CP violation processes in B meson decays have now been discovered. Before these "B-factory" experiments, there was a logical possibility that all CP violation was confined to kaon physics. However, this raised the question of why CP violation did not extend to the strong force, and furthermore, why this was not predicted by the unextended Standard Model, despite the model's accuracy for "normal" phenomena.

---

[33] Beta emission were preferentially in the direction opposite to the nuclear spin, in violation of conservation of parity.

In 12013 (holocene calendar) LHCb announced discovery of CP violation in strange B meson decays.

In March 12019, LHCb announced discovery of CP violation in charmed D^0 decays with a deviation from zero of 5.3 standard deviations.

In March 12025, LHCb collaboration reported the first observation of CP violation in baryon (composed of three quarks, i.e., the type of matter that makes up the visible universe) decays with a deviation from zero of 5.2σ.

In this work, CP violation in baryon decays is observed in the decay of the Λ_b^0 baryon to a proton, a kaon, and a pair of oppositely charged pions, represented as $\Lambda_b^0 \to pK^-\pi^+\pi^-$. The constituent quarks of the Λ_b^0 baryon are similar to those of the proton (uud), with one of the u quarks replaced by a b quark. CP violation is quantified by the so-called CP asymmetry, A_{CP}, defined as the relative difference between the decay rates of Λ_b^0 (matter) decay and $\bar{\Lambda}_b^0$ (antimatter) decay to self-conjugate final states. According to the Standard Model, this asymmetry arises from the quantum interference between the "tree" and "loop" quark-level amplitudes of the Λ_b^0 baryon decay.

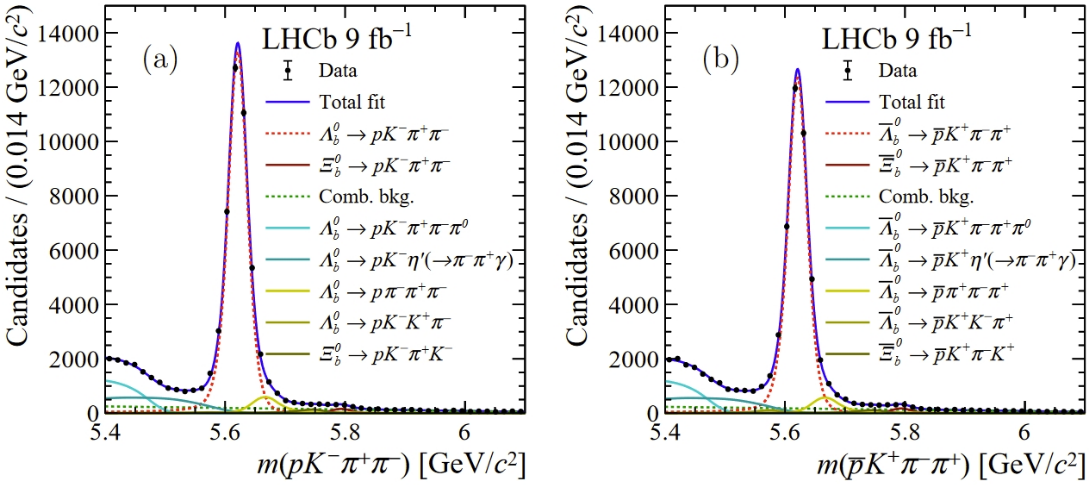


Figure 11.87: LHCb charge-parity violation by statistical mass distribution analysis (source: CERN)

The image above shows the mass distributions of Λ_b^0 and $\bar{\Lambda}_b^0$ decay products. The difference in the number of Λ_b^0 (left) and $\bar{\Lambda}_b^0$ (right) decays, once corrected for experimental effects, is a clear manifestation of CP violation. LHCb physicists obtained the numerical value of the CP asymmetry to be $A_{CP} = (2.45 \pm 0.46 \pm 0.10)\%$, differing from zero by 5.2σ. The nonzero value at more than 5 standard deviations implies by tradition the observation of CP violation.

In this early 121st century (holocene calendar) the T2K Collaboration reported some indications of CP violation in leptons for the first time but the statistical significance is not high enough, same for the Higgs interaction at CERN but here also the statistical significance is not high enough to be conclusive.

Event	Formulation	Type of violation	Year
Cobalt beta decay	$^{60}\mathrm{Co} \to {}^{60}\mathrm{Ni} + e^- + \bar{\upsilon}_e$	P violation	1956
Pion decay	$\pi^+ \to \mu^+ + \upsilon_\mu$	P and C violation	1957
Muon decay	$\mu^+ \to \mathrm{e}^+ + \bar{\nu}_\mu + \nu_\mathrm{e}$	P and C violation	1959
Kaon decays	$K_S^0 \to \pi^+ + \pi^-$ + others	C, P + CP, T violation	1964
B mesons decay	$\bar{B}^0 \to B_-$ and $B_- \to \bar{B}^0$	T violation	2012
Strange B meson decay	$K_S^0 \to \pi^+ + \pi^-$ + others	CP violation	2013
Charmed D^0 decays	$\bar{D}^0 \to K^+ + \pi^-$	CP violation	2019
Higgs decay	$H \to \tau^+ + \tau^- + \bar{\nu}_\tau + \nu_\tau$	CP violation	2022

Table 11.13: Types of known C, P and T violations

The reversibility of the laws of physics implies that given the state of a physical system at a particular time, it is always possible to work out uniquely both its future and its past. Time reversal invariance would further imply that the rules for going in each direction should be identical. To a very good approximation this appears to be true, but it turns out that in certain esoteric particle physics processes small deviations have been found. In particular, it was discovered in 11964 (holocene calendar) that the decay of the K^0 particle violated time reversal invariance at the level of about one part in a thousand. In current theories, this effect is not attributed any particularly fundamental origin, and is just assumed to be associated with the arbitrary setting of certain parameters. K^0 decay was for a long time the only example of time reversal violation that had explicitly been seen, although recently examples in B particle decays have probably also been seen.

Only a weaker version of the symmetry could be preserved by physical phenomena, which was CPT symmetry (this means that same rules should apply if one not only reverses the direction of time (T), but also simultaneously inverts all spatial coordinates (P) and conjugates all charges (C), replacing particles by antiparticles). Besides C and P, there is a third operation, time reversal T, which corresponds to reversal of motion. Invariance under time reversal implies that whenever a motion is allowed by the laws of physics, the reversed motion is also an allowed one and occurs at the same rate forwards and backwards.

Remark

Originally it was assumed that C, P and T would all separately be invariances, as they are in classical mechanics. But in 11957, as already mentioned, it was discovered that in radioactive beta decay, C and P are in a sense each maximally violated: among other things, the correlation between spin and motion direction is exactly opposite for neutrinos and for antineutrinos that are emitted. Despite this, it was still assumed that CP and T would be true invariances. But in 11964, as already mentioned, these too were found to be violated. Starting with a pure beam of K^0 particles, it turns out that quantum mechanical mixing processes lead after about 10-8 seconds to a certain mixture of K^0 particles - the antiparticles of the K^0. And what effectively happens is that the amount of mixing differs by about 0.1% in the positive and negative time directions. (what is actually observed is a small probability for the long-lived component of a K^0 beam to decay into two rather than three pions. Some analysis is required to connect this with T violation).

The combination of CPT is thought to constitute an exact symmetry of all types of fundamental interactions. Because of the long-held CPT symmetry theorem, provided that it is valid, a violation of the CP-symmetry is equivalent to a violation of the T-symmetry. In this theorem, regarded as one of the basic principles of quantum field theory, charge conjugation, parity, and time reversal are applied together. Direct observation of the time reversal symmetry violation without any assumption of CPT theorem was done in 11998 (holocene calendar) by two groups, CPLEAR and KTeV collaborations, at CERN and Fermilab, respectively. Already in 11970 (holocene calendar) Klaus Schubert observed T violation independent of assuming CPT symmetry by using the Bell–Steinberger unitarity relation.

Caution! Remember that we have proved (see page 3132) that every conservation law has a corresponding symmetry (i.e. conservation of momentum is translational symmetry, conservation of angular momentum is rotational symmetry) and that conservation of energy is temporal translation (i.e. a continuous symmetry)! The latter must not be confused with time inversion T!! The violation of time symmetry T (i.e. $T:t\longrightarrow -t$) doesn't violate energy conservation as it has nothing to do with time translation (i.e. $t\longrightarrow t+a, \quad a\in\mathbb{R}$)!!! If one has time reversal symmetry around every instant t_0 in time ($t-t_0\longrightarrow t_0-t, \quad t_0\in\mathbb{R}$) then it is not difficult to show that one also have translation symmetry and therefore also energy conservation!

11.3.7.7 CPT classical analogy (with Maxwell equations)

Let us first treat an analogy as it may help some readers to understand better what follows!

If we define $C:q\rightarrow -q, P:(x,y,z)\rightarrow(-x,-y,-z)$ and $T:t\rightarrow -t$, then all the Maxwell equations:

$$\vec{\nabla}\circ\vec{E}=\frac{\rho}{\varepsilon_0} \tag{11.3.480}$$

$$\vec{\nabla}\times\vec{E}=-\frac{\partial\vec{B}}{\partial t} \tag{11.3.481}$$

$$\vec{\nabla}\times\vec{B}=\mu_0\vec{j}+\frac{1}{c^2}\frac{\partial\vec{E}}{\partial t} \tag{11.3.482}$$

are invariant under C, P, T or any combination of them. To see this you just have to notice how the transformations act on sources and coordinates.

The charge conjugation acts trivially as:

$$\begin{aligned}C\rho&=-\rho\\ Cj_i&=-j_i\\ CE_i&=-E_i\\ CB_i&=-B_i\end{aligned}$$

with everything else unchanged. Notice the charge conjugation we defined is not exactly the charge conjugation defined in Quantum Field Theory. We can consider it as a classical counterpart.

The above relations are also commonly denoted:

$$\begin{array}{rcr}\vec{E}(\vec{x},t) & \overset{C}{\to} & -\vec{E}(\vec{x},t)\\ \vec{B}(\vec{x},t) & \overset{C}{\to} & -\vec{B}(\vec{x},t)\\ \vec{j}(\vec{x},t) & \overset{C}{\to} & -\vec{j}(\vec{x},t)\\ \rho(\vec{x},t) & \overset{C}{\to} & -\rho(\vec{x},t)\end{array} \tag{11.3.483}$$

Under parity we just have to be careful when dealing with vectors (such as $\vec{E}$) and pseudovector (such as $\vec{B}$). Vectors change components under spatial reflection while pseudovectors do not. The objects that change sign under P are:

$$\begin{aligned}Pj_i&=-j_i\\ PE_i&=-E_i\\ P\frac{\partial}{\partial x_i}&=-\frac{\partial}{\partial x_i}\end{aligned}$$

Or also commonly denoted:

$$\begin{array}{rcl}\vec{E}(\vec{x},t) & \overset{P}{\to} & -\vec{E}(-\vec{x},t)\\ \vec{B}(\vec{x},t) & \overset{P}{\to} & \vec{B}(-\vec{x},t)\\ \vec{j}(\vec{x},t) & \overset{P}{\to} & -\vec{j}(-\vec{x},t)\\ \vec{\nabla} & \overset{P}{\to} & -\vec{\nabla}\end{array} \tag{11.3.484}$$

Under time reversal, the velocity gets a minus sign, therefore $\vec{B}$ changes while $\vec{E}$ does not. So the non trivial actions of T are:

$$\begin{aligned}Tj_i&=-j_i\\ TB_i&=-B_i\\ T\frac{\partial}{\partial t}&=-\frac{\partial}{\partial t}\end{aligned}$$

Or also commonly denoted:

$$\begin{array}{rcl}\vec{E}(\vec{x},t) & \overset{T}{\to} & \vec{E}(\vec{x},-t)\\ \vec{B}(\vec{x},t) & \overset{T}{\to} & -\vec{B}(\vec{x},-t)\\ \vec{j}(\vec{x},t) & \overset{T}{\to} & -\vec{j}(\vec{x},-t)\\ \dfrac{\partial}{\partial t} & \overset{T}{\to} & -\dfrac{\partial}{\partial t}\end{array} \tag{11.3.485}$$

Under parity we just have to be careful when dealing with vectors (such as $\vec{E}$) and pseudovector (such as $\vec{B}$). Vectors change components under spatial reflection while pseudovectors do not.

11.3.7.8 CPT with Dirac equation

Let us start gently first...

As we know, the Dirac equation for a free with charge e is (see 4978):

$$[\gamma^\mu(\mathrm{i}\partial_\mu - eA_\mu) - m_0]\,\psi = \vec{0} \qquad (11.3.486)$$

Obviously to show the existence of charge symmetry we have to put the particle in an external electromagnetic field!

We want to know if we can construct a spinor ψ^c with the opposite charge from ψ. This would obey the equation:

$$[\gamma^\mu(\mathrm{i}\partial_\mu + eA_\mu) - m_0]\,\psi^c = \vec{0} \qquad (11.3.487)$$

We know know about gauge transformations (see 4115):

$$\psi \rightarrow e^{\mathrm{i}e\alpha}\psi \qquad (11.3.488)$$

(together with the compensating transformation for A_μ, which we don't need here), this suggests that complex conjugation is the thing to do:

$$\bar{\psi} \rightarrow e^{\mathrm{i}(-e)\alpha}\bar{\psi} \qquad (11.3.489)$$

So it looks like $\bar{\psi}$ has the opposite charge. Let's take the complex conjugate of the Dirac equation:

$$[-\bar{\gamma}^\mu(\mathrm{i}\partial_\mu + eA_\mu) - m_0]\,\bar{\psi} = \vec{0} \qquad (11.3.490)$$

Unfortunately this isn't what we want! But remember that spinors and γ matrices are only defined up to a change of basis $\psi \rightarrow A\psi$ and $\gamma^\mu \rightarrow A\gamma^\mu A^{-1}$. Possibly we can find a change of basis that brings the Dirac equation into the form we want.

For that purpose let us introduce an invertible matrix C by multiplying on the left and inserting $\mathbb{1} = C^{-1}C$:

$$\left(C\,[-\bar{\gamma}^\mu(\mathrm{i}\partial_\mu + eA_\mu) - m_0]\,C^{-1}\right)\left(C\bar{\psi}\right) = \vec{0} \qquad (11.3.491)$$

Hence:

$$\left[-C\bar{\gamma}^\mu C^{-1}(\mathrm{i}\partial_\mu + eA_\mu) - m_0\right]C\bar{\psi} = \vec{0} \qquad (11.3.492)$$

Note that if we can find a C which obeys $-C\bar{\gamma}^\mu C^{-1} = \gamma^\mu$ then $C\bar{\psi}$ makes a perfectly good candidate for ψ^c ! It turns out that one can indeed construct C satisfying the condition and define charge conjugation as:

$$\psi \rightarrow \psi^c = C\bar{\psi} \qquad (11.3.493)$$

To recap: we want to define a charge conjugation operation so that given a ψ with some electric charge e, we can get a ψ^c with charge $-e$. Complex conjugating the Dirac equation gets us there, but the resulting spinor $\bar{\psi}$ is in a different spinor basis so the Dirac equation is not in standard form. We introduce a change of basis C to get the Dirac equation back in standard form. The necessary conditions for this to work are that C is invertible (otherwise it wouldn't be a change of basis and bad things would happen) and $-C\bar{\gamma}^{\mu}C^{-1}=\gamma^{\mu}$.

Ok! Now that the reader has very likely understood the intellectual process, let us see the three different symmetry and how to determine the explicit operators (matrices). For information the developement below are almost a simple copy/paste of the lecture notes of Prof. PhD Niels Tuning (himself inspired by two others textbooks written by Bigi & Sanda [57] and Branco & Da Silva [74]).

We will derive now the explicit expressions for the P, C and T operators. By definition the transformed states $\psi^{p}(\vec{x},t),\psi^{c}(\vec{x},t)$ and $\psi^{t}(\vec{x},t)$ are constructed such that they satisfy the same equation of motions for free fields as $\psi(\vec{x},t)$. In the derivation of expressions for the P, C and T operators we start from the (correct) assumption that electromagnetic interactions are P, C and T symmetric. In other words, the Dirac equation should also hold for the P, C and T transformed fields. Eventually we will see what CP invariance implies for the weak interactions.

Let us consider the Dirac covariant equation of a particle with charge e in an electro-magnetic field:

$$\left(\mathrm{i}\gamma^{\mu}\frac{\partial}{\partial x_{\mu}}-\gamma^{\mu}eA_{\mu}-m_0\right)\psi(\vec{x},t)=\vec{0} \tag{11.3.494}$$

where $\psi(\vec{x},t)$ is a four component spinor and the matrices γ^{μ} are given by:

$$\gamma^{i}=\begin{pmatrix}0 & \sigma^{i}\\ -\sigma^{i} & 0\end{pmatrix}\text{ for } i=1,3;\quad \gamma^{0}=\begin{pmatrix}\mathbb{1} & 0\\ 0 & -\mathbb{1}\end{pmatrix}\text{, with :}$$
$$\sigma^{1}=\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix};\quad \sigma^{2}=\begin{pmatrix}0 & -\mathrm{i}\\ \mathrm{i} & 0\end{pmatrix};\quad \sigma^{3}=\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix};\quad \mathbb{1}=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix} \tag{11.3.495}$$

Explicitly for reminder:

$$\begin{gathered}\gamma^{0}=\begin{pmatrix}1&0&0&0\\0&1&0&0\\0&0&-1&0\\0&0&0&-1\end{pmatrix}\quad \gamma^{1}=\begin{pmatrix}0&0&0&1\\0&0&1&0\\0&-1&0&0\\-1&0&0&0\end{pmatrix}\\ \gamma^{2}=\begin{pmatrix}0&0&0&-\mathrm{i}\\0&0&\mathrm{i}&0\\0&\mathrm{i}&0&0\\-\mathrm{i}&0&0&0\end{pmatrix}\quad \gamma^{3}=\begin{pmatrix}0&0&1&0\\0&0&0&-1\\-1&0&0&0\\0&1&0&0\end{pmatrix}\end{gathered} \tag{11.3.496}$$

We now expand the Dirac equation above as:

$$\left(\gamma^{0}\left[\mathrm{i}\frac{\partial}{\partial t}-e\phi(\vec{x},t)\right]-\gamma^{i}\left[\mathrm{i}\frac{\partial}{\partial x_i}-eA_i(\vec{x},t)\right]-m_0\right)\psi(\vec{x},t)=\vec{0} \tag{11.3.497}$$

The Dirac equation after parity transformation becomes:

$$\left(\gamma^{0}\left[\mathrm{i}\frac{\partial}{\partial t}-e\phi(-\vec{x},t)\right]-\gamma^{i}\left[\mathrm{i}\frac{\partial}{\partial(-x_i)}-eA_i(-\vec{x},t)\right]-m_0\right)\psi(-\vec{x},t)=\vec{0} \tag{11.3.498}$$

Now, $\psi(-\vec{x},t)$ is not a solution of the Dirac equation, due to the additional $-$ sign in front of γ_i.

Multiplying the Dirac equation (after parity transformation) from the left by γ^0, we get:

$$\gamma^0\left(\gamma^0\left[\mathrm{i}\frac{\partial}{\partial t}-e\phi(-\vec{x},t)\right]+\gamma^i\left[\mathrm{i}\frac{\partial}{\partial x_i}-eA_i(-\vec{x},t)\right]-m_0\right)\psi(-\vec{x},t)=\vec{0} \qquad (11.3.499)$$

and then transport the γ^0 through the equation using the anti-commutation rules $\gamma^0\gamma^i=-\gamma^i\gamma^0$ for $i=1,2,3$ we get:

$$\left(\gamma^0\left[\mathrm{i}\frac{\partial}{\partial t}-e\phi(-\vec{x},t)\right]-\gamma^i\left[\mathrm{i}\frac{\partial}{\partial x_i}-eA_i(-\vec{x},t)\right]-m_0\gamma^0\right)\psi(-\vec{x},t)=\vec{0} \qquad (11.3.500)$$

We now see that the we have a term - the one with the mass - with $\gamma^0\psi(-\vec{x},t)$. That doesn't give us the possibility to fall back on the original Dirac covariant equation. We come to the conclusion that the original Dirac equation is obeyed by the simultaneous parity transformation in Lorentz space ($\vec{x}\rightarrow-\vec{x}$) and the transformation in Dirac space of the spinor with γ^0, if and only if :

$$\boxed{\psi(\vec{x},t)\xrightarrow{P}\psi^p(\vec{x},t)=\gamma^0\psi(-\vec{x},t)=P\psi(-\vec{x},t)} \qquad (11.3.501)$$

Of course also $e^{\mathrm{i}\phi}\gamma^0\psi(-\vec{x},t)$, with ϕ an arbitrary real phase, would provide a valid solution!

Therefore we have proven (using another common notation in textbooks):

$$\mathcal{P}\mathcal{L}_{\text{Dirace}}\left(t,x^i\right)\mathcal{P}^{-1}=\mathcal{L}_{\text{Dirac}}\left(t,-x^i\right) \qquad (11.3.502)$$

We will now take a look at the charge conjugation and investigate the interaction of a particle of opposite charge with an electro-magnetic field. Starting again from the Dirac equation and exchanging $e\rightarrow-e$ we find that the charge conjugate wave-function $\psi^c(\vec{x},t)$ must satisfy:

$$\left(\gamma^0\left[\mathrm{i}\frac{\partial}{\partial t}+e\phi(\vec{x},t)\right]-\gamma^i\left[\mathrm{i}\frac{\partial}{\partial x_i}+eA_i(\vec{x},t)\right]-m_0\right)\psi^c(\vec{x},t)=\vec{0} \qquad (11.3.503)$$

For our particular representation of the γ matrices we have the following properties (we change the notation for the complex conjugate to avoir the confusion with the notation of the self adjoint spinor):

$$\gamma^{0*}=\gamma^0,\gamma^{1*}=\gamma^1,\gamma^{2*}=-\gamma^2,\gamma^{3*}=\gamma^3 \qquad (11.3.504)$$

Then taking the complex conjugate of:

$$\left(\mathrm{i}\gamma^\mu\frac{\partial}{\partial x_\mu}-\gamma^\mu eA_\mu-m_0\right)\psi(\vec{x},t)=\vec{0} \qquad (11.3.505)$$

and expanding we get (and using the γ matrices properties too):

$$\begin{aligned}&\left(-\gamma^0\left[\mathrm{i}\frac{\partial}{\partial t}+e\phi(\vec{x},t)\right]+\gamma^1\left[\mathrm{i}\frac{\partial}{\partial x_1}\;+eA_1(\vec{x},t)\right]-\gamma^2\left[\mathrm{i}\frac{\partial}{\partial x_2}+eA_2(\vec{x},t)\right]\right.\\&\left.+\gamma^3\left[\mathrm{i}\frac{\partial}{\partial x_3}+eA_3(\vec{x},t)\right]-m_0\right)\psi^*(\vec{x},t)=\vec{0}\end{aligned} \qquad (11.3.506)$$

Now multiplying both sides with γ^2, and remember that $\gamma^\mu\gamma^\nu=-\gamma^\nu\gamma^\mu$ and :

$$\begin{aligned}&\left(-\gamma^2\gamma^0\left[\mathrm{i}\frac{\partial}{\partial t}+e\phi(\vec{x},t)\right]+\gamma^2\gamma^1\left[\mathrm{i}\frac{\partial}{\partial x_1}\right.+eA_1(\vec{x},t)]-\gamma^2\gamma^2\left[\mathrm{i}\frac{\partial}{\partial x_2}+eA_2(\vec{x},t)\right]\\&+\gamma^2\gamma^3\left[\mathrm{i}\frac{\partial}{\partial x_3}+eA_3(\vec{x},t)\right]-\gamma^2m_0\right)\psi^*(\vec{x},t)=\vec{0}\end{aligned} \tag{11.3.507}$$

Hence:

$$\begin{aligned}&\left(\gamma^0\gamma^2\left[\mathrm{i}\frac{\partial}{\partial t}+e\phi(\vec{x},t)\right]-\gamma^1\gamma^2\left[\mathrm{i}\frac{\partial}{\partial x_1}\right.+eA_1(\vec{x},t)]-\gamma^2\gamma^2\left[\mathrm{i}\frac{\partial}{\partial x_2}+eA_2(\vec{x},t)\right]\\&-\gamma^3\gamma^2\left[\mathrm{i}\frac{\partial}{\partial x_3}+eA_3(\vec{x},t)\right]-\gamma^2m_0\right)\psi^*(\vec{x},t)=\vec{0}\end{aligned} \tag{11.3.508}$$

Therefore:

$$\left(\gamma^0\left[\mathrm{i}\frac{\partial}{\partial t}+e\phi(\vec{x},t)\right]-\gamma^i\left[\mathrm{i}\frac{\partial}{\partial x_i}+eA_i(\vec{x},t)\right]-m_0\right)\gamma^2\psi^*(\vec{x},t)=\vec{0} \tag{11.3.509}$$

comparing this result with:

$$\left(\gamma^0\left[\mathrm{i}\frac{\partial}{\partial t}+e\phi(\vec{x},t)\right]-\gamma^i\left[\mathrm{i}\frac{\partial}{\partial x_i}+eA_i(\vec{x},t)\right]-m_0\right)\psi^c(\vec{x},t)=0 \tag{11.3.510}$$

we can readily identify:

$$\psi^c(\vec{x},t)=\gamma^2\psi^*(\vec{x},t) \tag{11.3.511}$$

Again we can use the arbitrary phase which we now take to be i, causing the combination $\mathrm{i}\gamma^2$ to be real:

$$\psi^c(\vec{x},t)=\mathrm{i}\gamma^2\psi^*(\vec{x},t) \tag{11.3.512}$$

We may stop here but there are some traditions that consist to write the C operator otherwise. Let us see how...! Rewriting this expression using the definition of the adjoint spinor (see page 5178):

$$\bar{\psi}^T\equiv\left(\psi^\dagger\gamma^0\right)^T=\gamma^{0T}\left(\psi^\dagger\right)^T=\gamma^0\psi^*\Rightarrow\psi^*=\gamma^0\bar{\psi}^T \tag{11.3.513}$$

yields the widely used expression for $\psi^c(\vec{x},t)$:

$$\boxed{\psi(\vec{x},t)\xrightarrow{C}\psi^c(\vec{x},t)=\mathrm{i}\gamma^2\psi^*(\vec{x},t)=\mathrm{i}\gamma^2\gamma^0\bar{\psi}^T(\vec{x},t)=C\bar{\psi}^T(\vec{x},t)} \tag{11.3.514}$$

Therefore we have proven (using another common notation in textbooks):

$$\mathcal{C}\mathcal{L}_{\text{Dirac}}\left(x^\mu\right)\mathcal{C}^{-1}=\bar{\mathcal{L}}_{\text{Dirac}}\left(x^\mu\right) \tag{11.3.515}$$

The time-reflection transformation takes the coordinate $x=(x_0,\vec{x})$ to $x'=(-x_0,\vec{x})$. This transformation can be written:

$$x'^\nu=a^\nu_\mu x^\mu \quad \text{with} \quad a^\nu_\mu=\begin{pmatrix}-1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{pmatrix} \tag{11.3.516}$$

Denoting $\psi'(x') \equiv \psi'(-x_0, \vec{x})$ the free time-reflected wave function, we attempt to construct it under the form (having in mind that time inversion intuitively should give us the antipartcle, hence we take the complex conjugate!):

$$\psi'(x') = T\psi^*(x) \quad \Rightarrow \quad \psi^*(x) = T^{-1}\psi'(x') \tag{11.3.517}$$

where T is a 4×4 constant matrix and $\psi(x)$ is the solution of the free Dirac equation with (we keep the notation used for the charge conjugation to denoted the complex conjugate to avoid any confusion with the adjoint spinor!):

$$\left(-\mathrm{i}\frac{\partial}{\partial x_\mu}\gamma^{\mu *} - m_0\right)\psi^*(x) = \vec{0} \quad \Rightarrow \quad \left(-\mathrm{i}\frac{\partial}{\partial x'_\nu}a_\mu^\nu\gamma^{\mu *} - m_0\right)T^{-1}\psi(x') = \vec{0} \tag{11.3.518}$$

Multiplying to the left by T we get:

$$\left(-\mathrm{i}\frac{\partial}{\partial x'_\nu}a_\mu^\nu T\gamma^{\mu *}T^{-1} - m_0\right)\psi'(x') = \vec{0} \tag{11.3.519}$$

And the reader must not forget that what we want is to fall back on the standard free Dirac equation:

$$\left(\mathrm{i}\frac{\partial}{\partial x'_\nu}\gamma^\nu - m_0\right)\psi'(x') = \vec{0} \tag{11.3.520}$$

And for that, it means we are looking for a T such that:

$$a_\mu^\nu T\gamma^{\mu *}T^{-1} = -\gamma^\nu \tag{11.3.521}$$

Recalling that $\gamma^{\mu *} = \gamma^\mu$ for $\mu = 0,1,3$ and $\gamma^{2*} = -\gamma^2$, the above conditions reduce to $T\gamma^i = \gamma^i T$ for $i = 0,2$ and $T\gamma^j = -\gamma^j T$, $j = 1,3$.

✍**Example:**

Consider the case γ^0. Therefore as $\gamma^{0*} = \gamma^0$:

$$a_\mu^\nu T\gamma^{\mu *}T^{-1} = -\gamma^\nu \Rightarrow a_0^\nu T\gamma^{0*}T^{-1} = -\gamma^\nu \Rightarrow a_0^\nu T\gamma^0 T^{-1} = -\gamma^\nu$$

For $\nu = 0$ we then have for example:

$$a_0^0 T\gamma^0 T^{-1} = -\gamma^0 \Rightarrow -T\gamma^0 T^{-1} = -\gamma^0 \Rightarrow T\gamma^0 T^{-1} = \gamma^0$$

Multiplying every side of the equality by T on the right, we get:

$$T\gamma^0 = \gamma^0 T$$

After some trial and errors we can found that the matrix:

$$T = \mathrm{i}\gamma^1\gamma^3 \tag{11.3.522}$$

satisfies the required conditions and we have thus:

$$\psi'(-x_0, \vec{x}) = \mathrm{i}\gamma^1\gamma^3\psi^*(x_0, \vec{x}) \tag{11.3.523}$$

This yields the widely used expression for $\psi^t(\vec{x},t)$:

$$\boxed{\psi(\vec{x},t)\xrightarrow{T}\psi^t(\vec{x},t)=\mathrm{i}\gamma^1\gamma^3\psi^*\left(x_0,\vec{x}\right)}\tag{11.3.524}$$

Therefore we have proven (using another common notation in textbooks):

$$\mathcal{T}\mathcal{L}_{\text{Dirace}}\left(t,x^i\right)\mathcal{T}^{-1}=\bar{\mathcal{L}}_{\text{Dirac}}\left(-t,x^i\right)\tag{11.3.525}$$

Remark

Combining the symmetries P and T one can construct the wave function of an electron evolving backward in space-time:

$$\begin{aligned}\psi^{PT}(-x^\mu)=PT\psi(x^\mu)&=\gamma^0\left[\mathrm{i}\gamma^1\gamma^3\psi^*(x^\mu)\right]\\&=\gamma^5\mathrm{i}\gamma^2\psi^*(x^\mu)\\&=\gamma_5\psi^c(x^\mu)\end{aligned}$$

where we introduced the wave-function of the positron. Thus the wave function of a (right-handed) positron is that of a (left-handed) electron moving backward in space-time (up to an irrelevant phase factor)!

Caution! It's not because a mathematical model is CPT symmetric that the reality is symmetric! A famous example is the Kaon decays. Theoretical models didn't first predict C,P + CP, T violations for the Kaons. Theoretical physicists had to change the theoretical model to be compliant with the evidence based science (i.e. experiments). The theoretical changes necessaries to take into account the Kaon decays symmetry violation observations leaded the physicists Makoto Kobayashi and Toshihide Maskawa in 11973 (both awarded by a Nobel Price in physics in 12008) to assume the existence of 4 supplementary quarks to the models assuming at their time having only 2 quarks (hence a total of 6 quarks still actually confirmed by experimental observations!).

~90 %

★★★★☆

4 votes, 74.39%

11.4. Nuclear Physics

NUCLEAR PHYSICS is the field of physics that studies the constituents and interactions of atomic nuclei. The most commonly known applications of nuclear physics are nuclear power generators but the research has provided application in many fields, including those in nuclear medicine (tracers) and magnetic resonance imaging, nuclear weapons, nuclear batteries as betavoltaic devices (we will come back on this subject later), ion implantation in materials engineering, radiocarbon dating in geology and archaeology, food or water sterilization, etc.

> *Anyone who expects a source of power from the transformations of these atoms is talking moonshine.*
>
> Ernest Rutherford

> *There is not the slightest indication that nuclear energy will ever be obtainable - it would mean that the atom would have to be shattered at will.*
>
> Albert Einstein

These two quotes provide once again evidence that opinions, even from the best scientists, are not reliable when faced to scientific facts and time...

11.4.1 Nuclear Weapon

Without wishing to make any confusion, we consider as essential, at the time when nuclear weapon serves as a dissuasion vector and thus as element of global stability (but also threat of global destruction ...), for the general culture of the engineer to know some basic properties of the atomic fission bomb. We'll exceptionally in this small subsection without mathematics (mathematical developments of nuclear weapons and nuclear power plants will be presented during our study of neutronic further below) speak a little bit of this mass destruction weapon that often fascinates the young students.

For sure, we will study later theoretically a nuclear divergent chain reaction in a given closed constrained volume! However, the reader must not expect to gain the knowledge necessary to manufacture such a weapon since it doesn't appeal only to knowledge of physics, but also electronics, mechanics, chemistry, mathematics, etc.

About an explosion, usage has immediately propagate to compare the energy of a nuclear weapon than a well known explosive material: the Trinitrotoluene (TNT). This TNT provides

$4,200,000$ joules per kilo, but the energies of nuclear weapons are such high that it is more meaningful to evaluate them in thousands of tonnes - or kilotons of TNT (subsequently thermonuclear weapons represented a new leap in explosive energy: the practical unity of comparison is one million tonnes - megaton of TNT).

The fission of 56 grams of Uranium 235 or Plutonium 239 gives the equivalent of 1 kiloton with the fission of $1.44 \cdot 10^{23}$ atoms (this is not even an integer value of Avogadro's number !!).

The first known nuclear test explosion at Alamogordo in July 16, 11945 (holocene calendar) - was estimated at 20 kt, with good accuracy because it had been possible to set up multiple measuring devices.

Those of August 6, 11945 (holocene calendar) on Hiroshima (based on 63 kilograms of Uranium 235) and August 9 on Nagasaki (based on 5.8 kilograms of Plutonium 239) were first estimated as 20 kt (only a estimate of 900 grams is converted into pure energy during the explosion). Subsequently, and after a detailed study on the effects of the blast, their energies were respectively reduced to about 17 to 15 kt respectively. This, however, did not represented less of the equivalent of a load of TNT of a convoy of around $6,000$ trucks from the US Army.

Here is a very good diagram (sadly we have yet only a french version) both interesting and persuasive on the effects of an atomic bomb (for information a wind with a speed above $220\,[\text{km}\cdot\text{h}^{-1}]$ can lift an average human being from the ground):

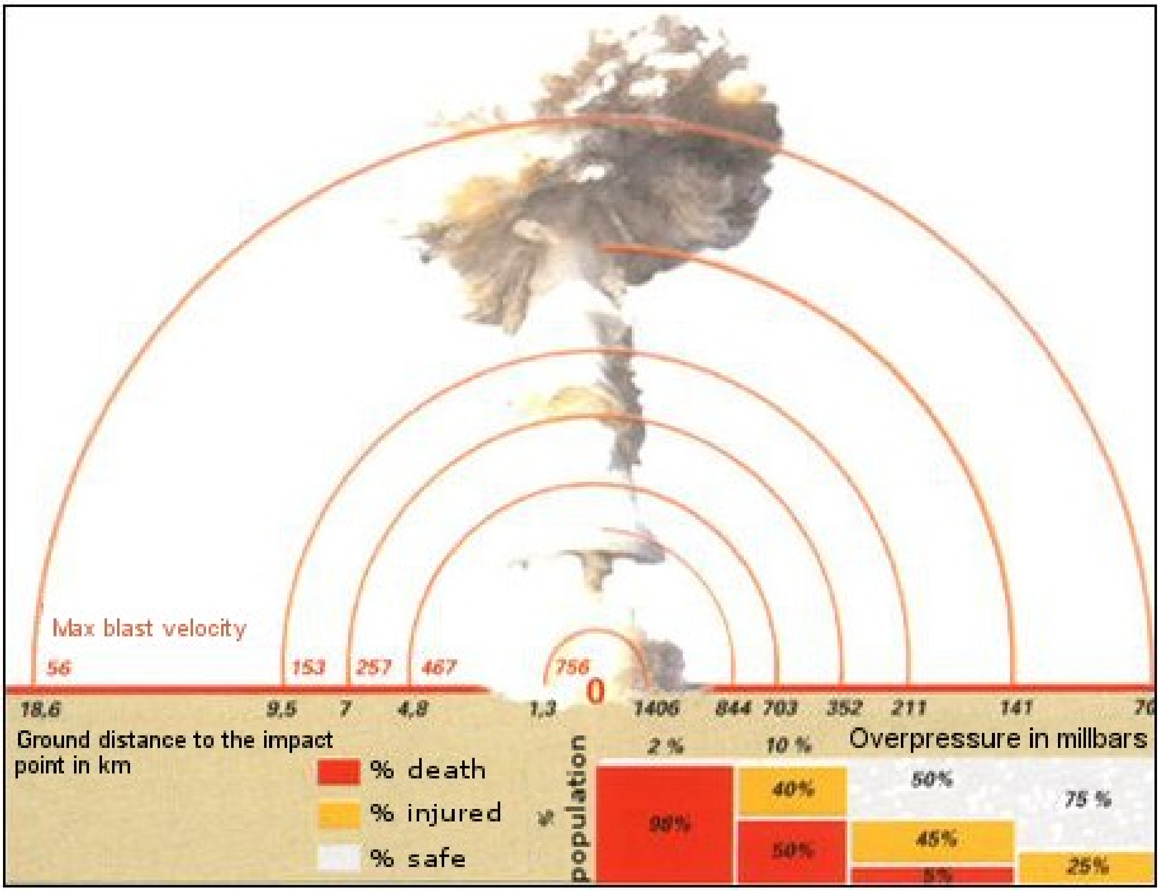

Max blast velocity
56
153
257
467
756
0
18,6
9,5
7
4,8
1,3
1406
844
703
352
211
141
Ground distance to the impact point in km
% death
% injured
% safe
population
2 %
10 %
Overpressure in millbars
98%
40%
50%
50%
45%
5%
75 %
25%

Figure 11.88: Effects of fission weapon of 1 Mt as a function of the distance (source: Pour la science)

So in other words here is a summary and approximately the effects of a 1 Mt fission weapon exploding at 2,450 meters of altitude (knowing that today the Americans and the Russians have nuclear fusion weapons with firepower exceeding the 50 megatons...):

Up to 1.3 [km], everything is down, even reinforced buildings. Up to 4.8 [km], most factories and commercial buildings are destroyed; houses made of brick and wood are blow up, and debris scattered. Up to 7 [km], commercial sets of light structure and private residences are demolished. Heavier buildings are seriously damaged. Up to 9.5 [km], the walls of small vessels are reversed, private residences severely deteriorated. The breath (or pressure) is still powerful enough to kill the people who are outside (explosion of the lungs). Up to 18.6 [km], different buildings are damaged, the streets are littered with debris and glass tiles. Between 10 to 20 minutes after the explosion, the debris drawn into the depression of the stem of the mushroom, fall to the ground. Follow after 1 to 2 hours later, the debris being in the head of the mushroom.

Most of the effects shown in the diagram above are not proportional to the energy. There is therefore no need to make a simple multiplication for a weapon with multiple power that used for the above figure!

For people interested in mushroom size (...) here is a figure whose source in unknown:

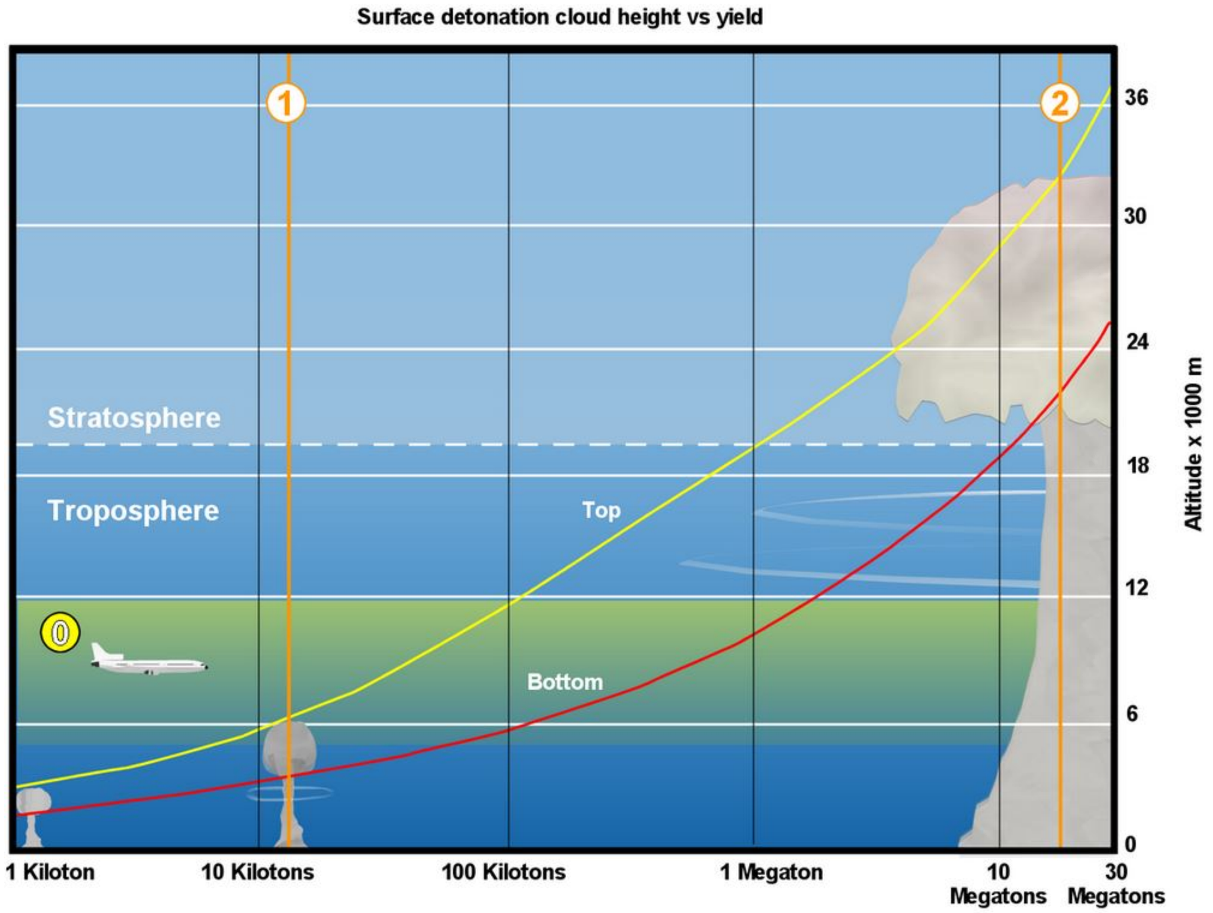


Figure 11.89: Mushroom size for Little Boy Hiroshima (1) and Castle Bravo (2)

Remark

For a nice little calculation on nuclear bombs using dimensional analysis the reader can refer to section Principles of Mechanics (see page 3081) where we give the expression of the energy of a nuclear bomb according to the estimated radius of the fireball.

Before we start the mathematical part, we emphasize and reiterate that we will limit ourselves only to theoretical developments made between 11890 and about the years 11935 according to holocene calendar (beyond the complexity of theories requires too many pages to a general textbook like this one).

11.4.2 Radioactivity

When we experimentally analyse radioactivity, we first perceive that the nucleus does not emits its elementary constituents. Then we discover new forces, that struggles and dominate in turn. Finally, new particles of matter, and even antimatter, appear. Deciphering these puzzles provided a coherent picture of the infinitely small world that radioactivity has revealed, a world where physical laws escape to the common intuition of daily practice of our macroscopic world.

From the start, the radioactivity has surprised. From the years 11900s (holocene calendar) it was known that radiation from Uranium and his descendants had three components, named "alpha" (α), "beta" (β), and "gamma" (γ) separable by the action of a magnetic field as shown symbolically in the image below:

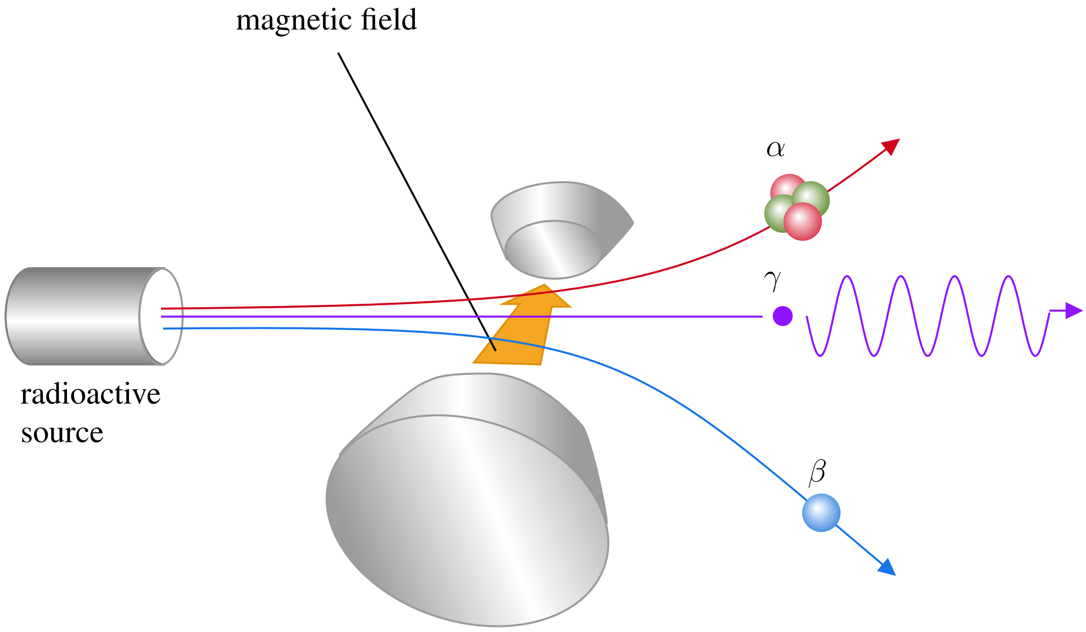


Figure 11.90: Separation of radiation using a magnetic field (source: Pour la Science)

Later, it was shown that the alpha radioactivity was the emission of a Helium nuclei (two protons and two neutrons), beta radiation the emission of electrons. From these observations, it was assumed that the nucleus was formed of these three types of particles (protons, neutrons and electrons), which is not the case: the components of nucleus seems to have been discovered by J. Chadwick only in 11932 (holocene calendar).

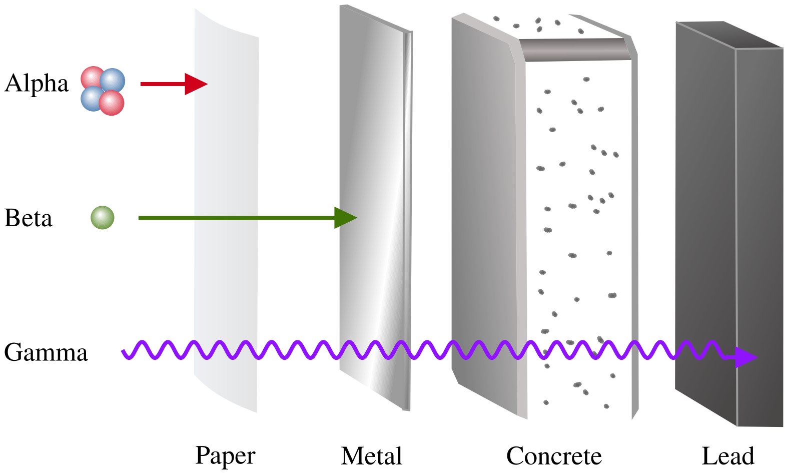


Figure 11.91: Typical radiation penetration

So why do radioactive nuclei not emit single protons or neutrons? How do nuclei eject something else than their fundamental constituents? These issues must be preceded by another, perhaps more fundamental, why some nuclei are radioactive? The response is the same for all spontaneous physical phenomena! The apple falling from the tree, for example: it is because the system can reach a more stable state by losing potential energy, the excess energy escaping in the form of kinetic energy, i.e. in the form of movement.

This reason also explains why the isotopes emit no protons or neutrons alone because often at the quantum structure of the nucleus, it is more favourable at the energy level to emit a small nucleus or to change a proton into a neutron (the quantum study of the nucleus exceeds the mathematical framework of the topics covered in this textbook actually).

Definitions (#11.4.509):

D1. Any chemical element (see chapter Chemistry page 6122) is characterized in the chemical point of view by its number of protons Z named "atomic number".

D2. The "mass number" A is by definition the number of protons Z summoned by the number of neutrons N for the given chemical element. Thus, the latter is completely characterized if we know its name or atomic number Z and its number N of neutrons or its mass number A. We usually denote any element in the form:

$$ {}_{Z}^{A}\mathrm{X}_{N} = {}_{Z}^{Z+N}\mathrm{X}_{N} \tag{11.4.1} $$

The chemical elements of the same species (same Z) may have different numbers of neutrons N, that is to say therefore different mass numbers A, then we speak of "isotopes" or "nuclides" (today this seems obvious as a definition but it took many years of research to get to this concept and we own this particularly to the works of Niels Bohr). The chemical elements which have the same A are named "isobars".

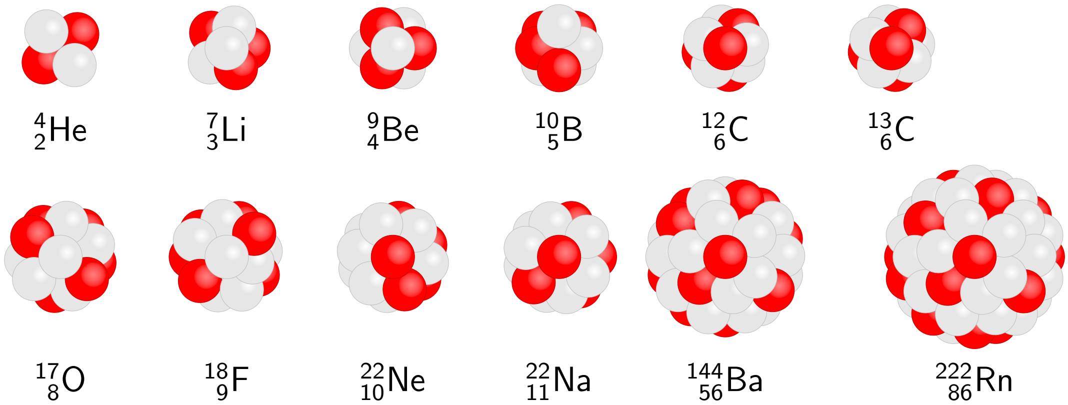


Figure 11.92: **Some nucleus (and isotopes) and their technical notation**

For a much more exhaustive list see the table of isotopes below:

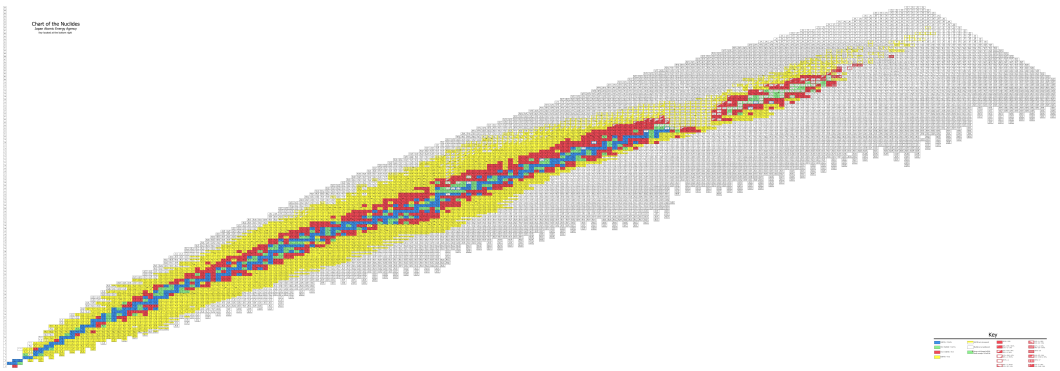


Figure 11.93: **Table of isotopes (zoom to see the details)**

Obviously, the nuclear energy (of the nucleus) associated with the same chemical element differs depending on the mass number A and we will prove that there is number A for which energy this minimal. Isotopes for which the energy is not a minimum may, for some of them and spontaneously, release the excess energy by doing a form of disintegration.

For example, according to a common misconception there is not only one type of Iron, but multiples:

Isotope	**Abondance** (molar percentage)
^{54}Fe	35%
^{56}Fe	36%
^{57}Fe	10%
^{58}Fe	4%
^{60}Fe	traces

Table 11.14: **Different isotopes of Iron**

D3. The property of certain atoms to spontaneously change the structure of their nuclei to reach a lower energy level, more fundamental, is named "radioactivity". We then talk about "radio-isotopes" for the atoms involved.

The chemical properties of an atom depend (see section Corpuscular Quantum Physics page 4675) on the number and arrangement of electrons in the electronic cloud (corresponding to the number of proton Z for non-ionized atoms). Thus all isotopes of the same element have the same chemical properties (it is this chemical characteristic which is the basis for nuclear medicine). These are sort of "brothers" atoms. However, the small difference in mass of the nucleus makes that their physical properties differ somewhat like the different type of water properties used to date air in ice cores:

Type of water	**Notation**	**Freezing point**	**Boiling point**
Water (dihydrogen monoxide)	H_2O	0.00 °C	100.00 °C
Heavy ^{18}O Water	$H_2{}^{18}O$	0.28 °C	100.14 °C
Semi-Heavy Water (deuterium oxide)	HDO	2.04 °C	100.70 °C
Heavy Water	$^2H_2O, D_2O$	3.1 °C	101.42 °C

Table 11.15: **Examples of isotopes that change the properties of a medium ("thermal isotope")**

Remark

Scientists have identified 118 eleemnts and approximately $1,500$ isotopes, among which 300 are natural (the others are artificially generated in nuclear reactors or accelerators). There are isotopes that are stable and some that are unstable (i.e. radio-isotopes). In Nature the proportion of isotopes of a same element in a material (even living one) varies depending on many factors.

D4. The "isotones" are the isotopes of various chemical elements (different Z) with the same number of neutrons N.

Rather than speak of "chemicals elements" when we actually are implicitly referring to the nucleus, the term "nuclide" is preferred.

The smallness of atoms is an obvious problem of mass measurement. That is why it was preferred by physicists and chemists to develop an atomic mass system which is a system of numbers proportional to the real mass of the atoms.

Since there are infinitely many possible systems, one was chosen judiciously as a reference and is the number 12 for the carbon-12 isotope (for more details see the section of Analytical Chemistry):

$$^{12}\text{C} = 12.00000\;[\text{amu}] \tag{11.4.2}$$

where "amu" is the abbreviation of "atomic mass units".

This choice has for interesting consequence to give to the proton and neutron an atomic weight near to the unit.

We can therefore connect the I.S.Q. system (see section Principia page 3075) with the system of atomic mass units (amu).

D5. A group of chemical compounds only differing in their isotopic composition (for example normal water and heavy water) is named an "isotopologue".

The fact that chemical compounds can be made of isotopes is experimentally observed as changing a little bit the properties of chemical reactions (like speed of the chemical reaction or its enthalpy) and also of gas behaviours!

D6. The "atomic mass unit" is defined as the mass of the $1/12$ of the atom of carbon ^{12}C, we have (the mass of the electrons is neglected as very small compared to that of nucleons):

$$1\,[\text{uma}]=\frac{1}{12\,[\text{nucleons}\cdot\text{at.}^{-1}]}\frac{12\,[\text{g}\cdot\text{mole}^{-1}]}{N_{\text{A}}}=\frac{1}{6.0221367\cdot 10^{23}}=1.6605402\cdot 10^{-24}\,[\text{g}] \tag{11.4.3}$$

Therefore the uma mass of the proton is equal to:

$$m_{\text{uma,p}}=\frac{1.6726231\cdot 10^{-24}\,[\text{g}]}{1.660.5402\cdot 10^{-24}\,[\text{g}]}\cong 1.007276468\,[\text{uma}] \tag{11.4.4}$$

for the admitted value in practice the reader must referred to the literature of the ad hoc international norms (because it varies depending on the new versions of international norms).

However the reader must be careful, though the molar mass of a different isotope than ^{12}C by adding the masses of the nucleons (protons and neutrons) that make up its nucleus because we must take into account the mass defect (a concept that we will see further below).

The mass can also be expressed in energy units because there is a mass-energy equivalence as we have proved it in the section of Special Relativity (see page 5679) from the relation $E=mc^2$. The energy units used in nuclear physics is often the "electron volt".

D7. An "electron-volt" denoted [eV] is the energy acquired by an elementary charge subject to a potential difference of 1 [V].

Thus, according to the relation between energy and the electrostatic potential $E=qU$ (see section Electrostatics page 4005), we have:

$$\boxed{1\,[\text{eV}]=1.60217733\cdot 10^{-19}\,[\text{J}]} \tag{11.4.5}$$

We get from this since the speed of light in vacuum is almost equal to $c=299,792,458\,[\text{m}\cdot\text{s}^{-1}]$:

$$1\,[\text{g}]=8.987551\cdot 10^{13}\,[\text{J}]=5.60959\cdot 10^{32}\,[\text{eV}]\Rightarrow 1\,[\text{uma}]=931.49432\,[\text{MeV}] \tag{11.4.6}$$

By mass–energy equivalence, the electron-volt is also a unit of mass. It is common in particle physics, where units of mass and energy are often interchanged, to express mass in units of eV$/c^2$. It is common to simply express mass in terms of "eV" as a unit of mass, effectively using a system of natural units with c set to 1 (see section Principia page 3124). The mass equivalent of 1 eV$/c^2$ is:

$$1\,[\text{eV}\cdot c^{-2}]=\frac{(1.60217646\cdot 10^{-19}\,[\text{C}])\cdot 1\,[\text{V}]}{(2.99792458\cdot 10^{8}\,[\text{m}\cdot\text{s}^{-1}])^2}\cong 1.783\times 10^{-36}\,[\text{kg}] \tag{11.4.7}$$

Again, for the admitted value in practice the reader has to refer to the literature of the ad hoc international norms (because it varies depending on the new versions of these norms).

Remark

Practising astrophysicists and plasma physicists routinely refer to temperatures in units of [eV] or [keV], even thought this is wrong, because temperature is not dimensionally equivalent to energy. Nevertheless, they still do it, with the Boltzmann constant being implicitly included in the conversion. Here are formulas for temperature in Kelvin to [keV] conversion:

$$E([\text{eV}])=\frac{kT}{e} \tag{11.4.8}$$

because 1 [eV] is by definition the energy required to move a charge e through a potential difference of 1 [V] and is equal to $1.6\cdot 10^{-19}$ [J]. Therefore:

$$T([\text{eV}])=\frac{k}{e}T=8.625\cdot 10^{-5}\cdot T \tag{11.4.9}$$

The use of stable isotopes is widely used for health checks, the quality or origin of foods and compliance with their labels - because the natural proportions are not the same from one geographical point to another and with or without pesticides- as well as doping controls (to check if elements in the blood are not of industrial origin), in criminology (origin of corpses).

11.4.2.1 Disintegration

Some nucleus therefore have the property of spontaneously change their internal structure to achieve a basic level of energy. This transformation is accompanied by the emission of particles and / or electromagnetic radiation. The residual core can be radioactive too and process to other transformations afterwards or be stable.

The radioactive decay of an isotope is a random phenomenon and we can never tell when a nucleus will disintegrate (probability no memory effect).

Remark

To prove this assertion, the reader may refer to the section of Quantitative Management in the subsection of Queuing Theory, in particular modelling arrivals (see page 8145). Indeed, the development is identical in all point but only the object of study changes (that are no longer phone calls but disintegration). Thus, we proved that under certain assumptions the phenomenon follows a Poisson law and we will prove a little bit afterwards that it has no memory (see page 8149)!

We can only give the probability of disintegration (decay) per unit time. This probability is given by the "disintegration constant" and has for unit the inverse of time: $\lambda\,[\mathrm{s}^{-1}]$. This constant can be calculated as we have already proved it during our the study of the tunnel effect in the section of Wave Quantum Physics (see page 4840).

The radioactive constant varies for all known isotopes:

$$1.57\cdot 10^{-18}\,[\mathrm{s}^{-1}]\leq\lambda\leq 3\cdot 10^{6}\,[\mathrm{s}^{-1}] \qquad (11.4.10)$$

Given now $N(t)$ the initial stock of atoms of a radioactive isotope at time t. The number of atoms disintegrating during the infinitesimal time $\mathrm{d}t$ is equal to:

$$\lambda N(t)\mathrm{d}t \qquad (11.4.11)$$

leading to a decrease in the stock obviously equal to:

$$-\lambda N(t)\mathrm{d}t \qquad (11.4.12)$$

The decay differential equation (see section Differential and Integral Calculus page 1609) is thus written:

$$\mathrm{d}N=-\lambda N\mathrm{d}t \qquad (11.4.13)$$

or:

$$\frac{\mathrm{d}N}{\mathrm{d}t}+\lambda N=0 \qquad (11.4.14)$$

which has the simple solution (see section Differential and Integral Calculus page 1609):

$$\boxed{N(t)=N_0e^{-\lambda(t-t_0)}} \qquad (11.4.15)$$

with $N_0=N(t_0)$ being the initial stock of nucleus at time t_0. So notice that as the decay is an inverse exponential, there is no such thing a "total life" for isotopes!

Remark

$N(t)$ does not represent the number of atoms remaining at time t but the most likely number of radioactive atoms remaining at time t !!

In the practice, the measurement of the decay constant is done thanks the decrease of the activity (see below) of the individual isotope.

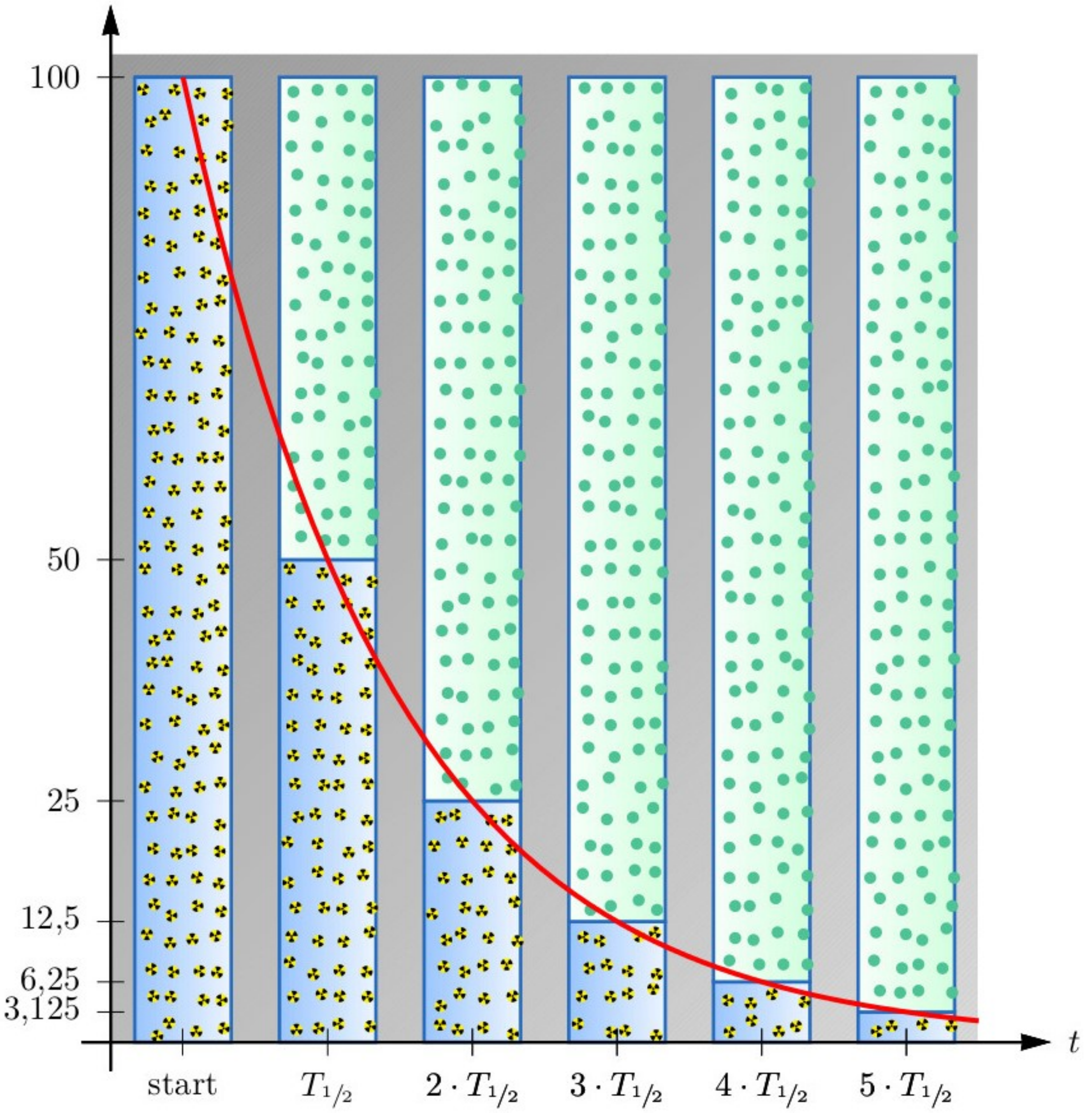


Figure 11.94: **Illustrated radioactivity half-life decay (author: Svend Tveskæg)**

11.4.2.1.1 Half-life and Mean life isotope

Definition (#11.4.510): The "half-life period" or just "half-life" $T_{1/2}$ of an isotope is the average time that it takes for 50% of the radioactive nuclei stock of a given isotope to disintegrate:

$$\frac{N(t=T_{1/2})}{N_0}=\frac{1}{2}=e^{-\lambda T_{1/2}} \Rightarrow T_{1/2}=\frac{\ln(2)}{\lambda} \tag{11.4.16}$$

Thus we have a very important relation between the half-life period and the decay constant:

$$\boxed{T_{1/2}=\frac{\ln(2)}{\lambda}} \tag{11.4.17}$$

The probability for decay can be expressed as a distribution function:

$$f_{\text{decay}}(t)=C^{te}e^{-\lambda t} \tag{11.4.18}$$

To normalize this distribution function:

$$\int\limits_0^{+\infty} f_{\text{decay}}(t)\mathrm{d}t=\int\limits_0^{+\infty} C^{te}e^{-\lambda t}\mathrm{d}t=-\frac{1}{\lambda}C^{te}e^{-\lambda t}\Big|_0^{+\infty}=\frac{C^{te}}{\lambda}=1 \tag{11.4.19}$$

Hence:

$$f_{\text{decay}}(t)=\lambda e^{-\lambda t} \tag{11.4.20}$$

The probability that a given particle will decay within time t is given by the integral of the decay distribution function from 0 to t. This is not the quantity which we wish to calculate if we want the average time the particle will exist without decaying. The probability that a particle is not decayed is one minus the probability that it will decay. The probability that a particle will remain at time t is then:

$$P_u(t)=1-\int\limits_0^{t}\lambda e^{-\lambda t'}\mathrm{d}t'=1+e^{-\lambda t'}\Big|_0^{t}=e^{-\lambda t} \tag{11.4.21}$$

Definition (#11.4.511): Given an assembly of elements, the number of which decreases ultimately to zero, the "mean lifetime" τ, is the expected value of the amount of time before an object is removed from the assembly. Specifically, if the individual lifetime of an element of the assembly is the time elapsed between some reference time and the removal of that element from the assembly, the mean lifetime is the arithmetic mean of the individual lifetimes.

Starting from density probability of decay:

$$f_{\text{decay}}(t)=\lambda e^{-\lambda t} \tag{11.4.22}$$

Exponential decay is therefore scalar multiple of the statistical exponential distribution (i.e. the individual lifetime of each object is exponentially distributed), which has a well-known expected value (see section Statistics page 751):

$$\tau=\langle t\rangle=\int\limits_0^{+\infty}\lambda t e^{-\lambda t}\mathrm{d}t=\frac{1}{\lambda} \tag{11.4.23}$$

So the mean lifetime for particle decay is given by:

$$\boxed{\tau=\frac{1}{\lambda}=\frac{T_{1/2}}{\ln(2)}\cong\frac{T_{1/2}}{0.693}} \tag{11.4.24}$$

A quantity may decay via two or more different processes simultaneously. In general, these processes (often named "decay modes", "decay channels", "decay routes" etc.) have different probabilities of occurring, and thus occur at different rates with different half-lives, in parallel. The total decay rate of the quantity N is given by the sum of the decay routes; thus, in the case of two processes:

$$-\frac{\mathrm{d}N(t)}{\mathrm{d}t}=N\lambda_1+N\lambda_2=(\lambda_1+\lambda_2)\,N \tag{11.4.25}$$

Therefore:

$$N(t)=N_0e^{-(\lambda_1+\lambda_2)t}=N_0e^{-(\lambda_c)t} \tag{11.4.26}$$

where the sum of $\lambda_1+\lambda_2$ is treated as a new total decay constant λ_c. Using the same reasoning as before we get the total half-time (also named "partial half-life"):

$$T_{1/2}=\frac{\ln(2)}{\lambda_c}=\frac{\ln(2)}{\lambda_1+\lambda_2} \tag{11.4.27}$$

As:

$$\tau=\frac{1}{\lambda}=\frac{T_{1/2}}{\ln(2)} \tag{11.4.28}$$

The latter can be rewritten:

$$T_{1/2}=\frac{\ln(2)}{\dfrac{\ln(2)}{T_{1/2}^1}+\dfrac{\ln(2)}{T_{1/2}^2}}=\frac{1}{\dfrac{1}{T_{1/2}^1}+\dfrac{1}{T_{1/2}^2}} \tag{11.4.29}$$

As:

$$\frac{1}{\tau_c}=\lambda_c=\lambda_1+\lambda_2=\frac{1}{\tau_1}+\frac{1}{\tau_2} \tag{11.4.30}$$

We also have the following "partial mean life":

$$\tau_c=\frac{\tau_1\tau_2}{\tau_1+\tau_2} \tag{11.4.31}$$

Remark

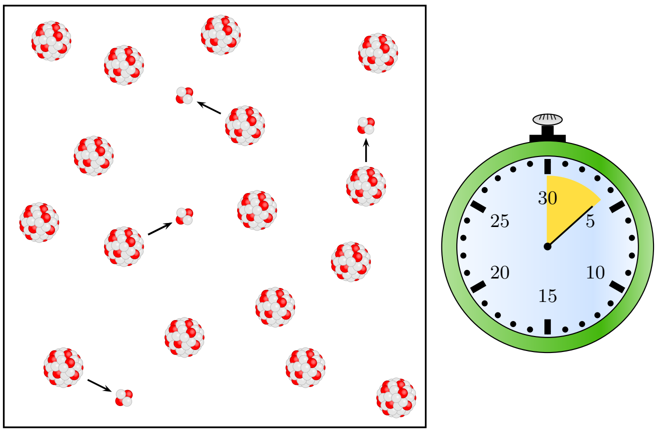
In 11913 (holocene calendar), Pierre Curie and M. Kamerlingh Onnes measured the rate of decay of radium at room temperature and after being cooled in liquid hydrogen. Their conclusion was that the decay rate was entirely independent of temperature. Since then, numerous investigations have shown that alpha and beta decays are not influenced by external conditions such as temperature, air pressure but only by the surrounding material in some very extreme situations (like near absolute zero). By contrast, decays that proceed by electron capture are known to be susceptible to their environment and so have been placed in a different category.

11.4.2.2 Activity

Definition (#11.4.512): The "activity" of a radioactive source is the number of disintegrations per time unit.

Remark

Its unit of measurement is the "Becquerel" denoted [A]=[Bq]. One Becquerel thus corresponding to one disintegration per second.



$\Delta n = 4$ disintegration ins $\Delta t = 4$ [s]

$$A = \frac{\Delta n}{\Delta t} = 1 \text{ [Bq]}$$

The old unit of measurement of radioactivity was the "Curie" denoted [Ci]. The Curie was defined initially as the activity of about one gram of radium, natural element that we find in the soil with uranium. This unit is much larger than the previous one because, by definition 1 [Ci] is equal to 37 billion disintegrations per second:

$$1 \text{ [Ci]} = 3.7 \cdot 10^{10} \text{ [Bq]} \tag{11.4.32}$$

The activity is obtained by the time derivative of the stock atoms of a given sample:

$$A(t) = -\frac{\mathrm{d}N(t)}{\mathrm{d}t} = \lambda N_0 e^{-\lambda(t-t_0)} = \lambda N(t) \tag{11.4.33}$$

The relation, named "equation of activity":

$$A = \lambda N \tag{11.4.34}$$

thus shows that the activity of a given number of N atoms of a radioactive isotope is proportional to that number and inversely proportional to the half-life of the isotope (by the relation seen above between the radioactive constant and the half-life period).

✍**Example:**

One gram of $^{225}\mathrm{Ra}$ contains (the reader must regularly refer to the common international standards for the values of the constants used!):

$$\frac{1\,[\text{g}]\cdot 6.0221267\cdot 10^{23}\,[\text{at}\cdot\text{mole}^{-1}]}{226.0256\,[\text{g}\cdot\text{mole}^{-1}]}\cong 2.664\cdot 10^{21}\,[\text{atoms}] \tag{11.4.35}$$

Therefore the activity of this gram is, knowing experimentally that $T_{1/2}=1,600\,[\text{years}]$, equal to:

$$A=\lambda N=\frac{\ln(2)}{T_{1/2}}N=\frac{\ln(2)}{1600\cdot 365.256\cdot 24\cdot 60\cdot 60}\cdot 2.664\cdot 10^{21}=3.66\cdot 10^{10}\,[\text{Bq}] \tag{11.4.36}$$

By the same reasoning (that we can detailed on reader request as always in this textbook), we show that the activity over time follows the same exponential law of the number of nuclides:

$$\boxed{A(t)=A_0e^{\lambda(t-t_0)}} \tag{11.4.37}$$

with:

$$A_0=A(t_0)=\lambda N_0 \tag{11.4.38}$$

Experimentally to determine the half-life period of an isotope we proceed as following:

1. We choose a pure as possible sample of an isotope which we wish to determine the half-life period.
2. At the time t_0 we measure thanks to a detector for a fixed time interval time Δt the number of disintegrations. We then have the number of decays during a time interval at the start of experiment (the unit of measurement is then decays and not the number of disintegrations per second).
3. Then, for each consecutive Δt (time interval is fixed) we measure the number of disintegrations during this time interval. This gives us a series of measures the number of decays observed for $\Delta t_1,\Delta t_2,\Delta t_3,\ldots$
4. To the set of all decay measurements, we remove the laboratory background noise (as the whole environment is radioactive).

As:

$$A(t)\Delta t=A(t_0)\Delta te^{-\lambda(t-t_0)} \tag{11.4.39}$$

By taking the neperian logarithm we get:

$$\ln(A(t)\Delta t)=\ln(A(t_0)\Delta t)-\lambda(t-t_0)=-\lambda(t-t_0)+\ln(A(t_0)\Delta t) \tag{11.4.40}$$

Thus:

$$\ln(A(t)\Delta t)=-\lambda t'+\ln(A(t_0)\Delta t) \tag{11.4.41}$$

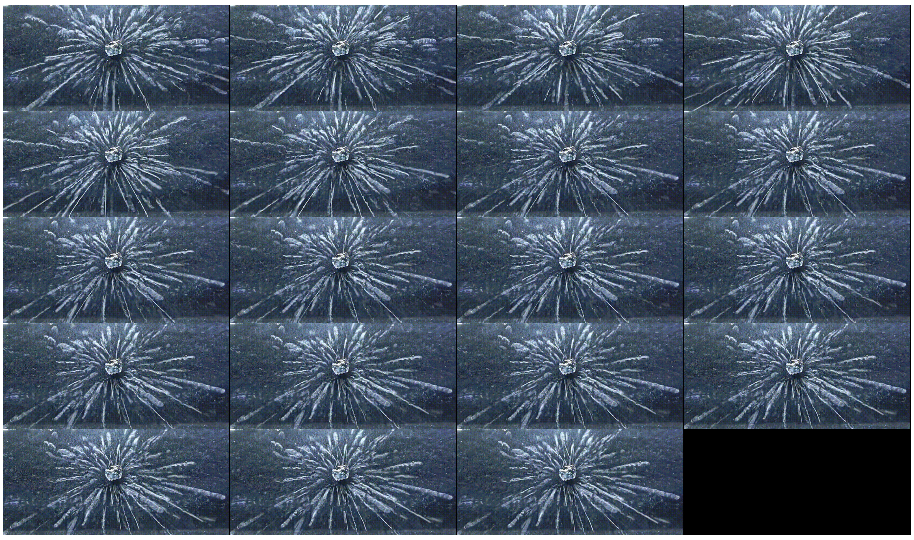
So this is the equation of a straight line of slope $-\lambda$ and intercept $\ln(A(t_0)\Delta t)$. Thus, the decay constant is immediately measured and we quickly deduced the period of half-life using the relation proved earlier above:

$$T_{1/2}=\frac{\ln(2)}{\lambda} \tag{11.4.42}$$

In the gallery below taken from https://www.youtube.com/watch?v=ZiscokCGOhs&t=250s we can see a sealed glass container that contains liquid alcohol at the top. Emanating alcohol vapors fill the whole volume of the container until they reach the bottom of the chamber maintained to a very cold temperature (-40 [°C]). Most of the vapour condenses on the glass surface creating a mist, but a small fraction of it stays in vapour form above the cold condenser. This creates a layer of unstable sursaturated vapour which can condense at any moment. When a charged particle crosses this vapor, it can knock electrons off the molecules forming ions. It causes the unstable alcohol vapor to condense around ions left behind by the travelling ionizing particle : the path of the particle in the matter is then revealed by a track composed of thousands droplets of alcohol.

Any charged particle is visible in a cloud chamber. The most common ones are alphas, electrons, positons, protons, nuclear charged fragment, muons (...). Theses particles come from natural cosmic and telluric background radiations or from close radioactive sources. They will all leave tracks of different shapes in the chamber, based on their charge, mass and speed. Electrons are the lightest particles and will be easily deflected by magnetic fields. Alphas and protons are much heavier and slower and will thus ionize more, causing denser track of droplets. Interactions of neutral particles like gamma rays or neutrons can be seen thanks to the charged particles they create in matter.

The gallery below is a subset of a 50 min video with an Uranium mineral. The small thickness of liquid alcohol on the mineral is enough to absorb a part of the energy of the alpha particles (their ranges in air for 5 [MeV] is 3-4 [cm].



11.4.2.2.1 Radiometric dating

"Radiometric dating", also named "radioactive dating" or "radioisotope dating" is a technique which is used to date materials such as rocks or carbon, in which trace radioactive impurities were selectively incorporated when they were formed. The method compares the abundance of a naturally occurring radioactive isotope within the material to the abundance of its decay products, which form at a known constant rate of decay. The use of radiometric dating was first published in 11907 (holocene calendar) by Bertram Boltwood and is now the principal source of information about the absolute age of rocks and other geological features, including the age of fossilized life forms or the age of the Earth itself, and can also be used to date a wide range of natural and man-made materials.

11.4.2.2.1.1 Carbon-14 dating (radiocarbon dating)

"Radiocarbon dating" (also referred to as "carbon dating" or "carbon-14 dating") is a method for determining the age of an object containing organic material by using the properties of radiocarbon $^{14}\mathrm{C}$, a radioactive isotope of carbon.

The method was developed by Willard Libby in the late 11940s (holocene calendar) and soon became a standard tool for archaeologists. Libby received the Nobel Prize for his work in 11960 (holocene calendar). The radiocarbon dating method is based on the fact that radiocarbon is constantly being generated in the atmosphere by the interaction of cosmic rays with atmospheric nitrogen. The resulting radiocarbon combines with atmospheric oxygen to form radioactive carbon dioxide, which is incorporated into plants by photosynthesis; animals then acquire $^{14}\mathrm{C}$ by eating the plants. When the animal or plant dies, it stops exchanging carbon with its environment, and from that point onwards the amount of $^{14}\mathrm{C}$ it contains begins to decrease as the $^{14}\mathrm{C}$ undergoes radioactive decay. Measuring the amount of $^{14}\mathrm{C}$ in a sample from a dead plant or animal such as piece of wood or a fragment of bone provides information that can be used to calculate when the animal or plant died. The older a sample is, the less $^{14}\mathrm{C}$ there is to be detected, and because the half-life of $^{14}\mathrm{C}$ (the period of time after which half of a given sample will have decayed) is about $5,730$ years, the oldest dates that can be reliably measured by radiocarbon dating are around $50,000$ years ago, although special preparation methods occasionally permit dating of older samples.

The idea behind radiocarbon dating is straightforward, but years of work were required to develop the technique to the point where accurate dates could be obtained. Research has been ongoing since the 11960s (holocene calendar) to determine what the proportion of $^{14}\mathrm{C}$ in the atmosphere has been over the past fifty thousand years. The resulting data, in the form of a calibration curve, is now used to convert a given measurement of radiocarbon in a sample into an estimate of the sample's calendar age. Other corrections must be made to account for the proportion of $^{14}\mathrm{C}$ in different types of organisms (fractionation), and the varying levels of $^{14}\mathrm{C}$ throughout the biosphere (reservoir effects). Additional complications come from the burning of fossil fuels such as coal and oil, and from the above-ground nuclear tests done in the 11950s (holocene calendar) and 11960s. Because the time it takes to convert biological materials to fossil fuels is substantially longer than the time it takes for its $^{14}\mathrm{C}$ to decay below detectable levels, they contain almost no $^{14}\mathrm{C}$, and as a result there was a noticeable drop in the proportion of $^{14}\mathrm{C}$ in the atmosphere beginning in the late 119th century (holocene calendar). Conversely, nuclear testing increased the amount of $^{14}\mathrm{C}$ in the atmosphere, which attained a maximum in

11963 (holocene calendar) of almost twice what it had been before the testing began.

Natural carbon has two stable isotopes: the ${}^{12}\mathrm{C}$ (98.892%) and ${}^{13}\mathrm{C}$ (1.108%). So there is no ${}^{14}\mathrm{C}$ in the natural carbon. The latter is produced in the upper atmosphere through the action of cosmic neutrons on the ${}^{13}\mathrm{C}$. We speak then of "neutron capture" (see further below) or "activation ${}^{13}\mathrm{C}(n,\gamma){}^{14}\mathrm{C}$" following:

$$n+{}^{14}{}_{7}\mathrm{N}\rightarrow{}^{14}{}_{6}\mathrm{C}+p \tag{11.4.43}$$

where n represents a neutron and p represents a proton.

Thus, continuously ${}^{14}\mathrm{C}$ is produced in the upper atmosphere and decays naturally with a period of $5,730$ years. We easily imagine that the concentration in ${}^{14}\mathrm{C}$ is balanced when the production rate is equal to the disappearing rate due to radioactive decay process (otherwise they will no longer be any ${}^{14}\mathrm{C}$ everywhere today) following:

$${}^{14}{}_{6}\mathrm{C}\rightarrow{}^{14}{}_{7}\mathrm{N}+e^{-}+\bar{\nu}_e \tag{11.4.44}$$

By emitting a beta particle (an electron, e^{-}) and an electron antineutrino (ν_e), one of the neutrons in the ${}^{14}\mathrm{C}$ nucleus changes to a proton and the ${}^{14}\mathrm{C}$ nucleus reverts to the stable (non-radioactive) isotope ${}^{14}\mathrm{N}$.

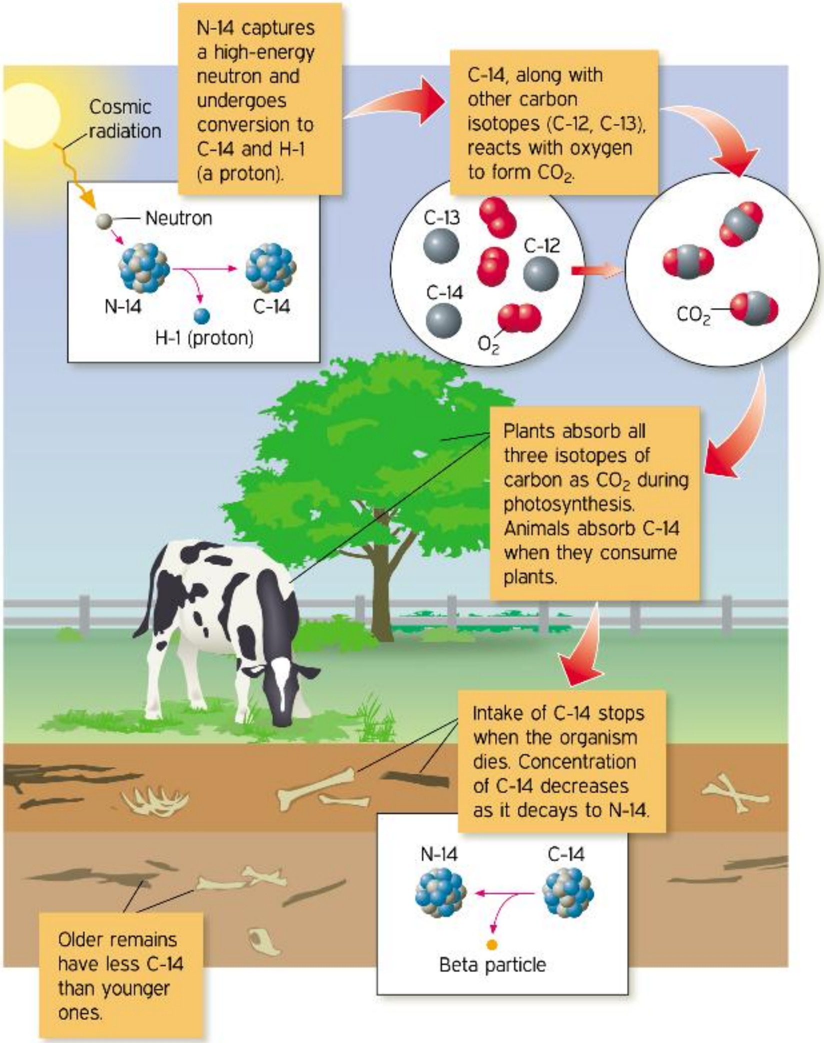

N-14 captures a high-energy neutron and undergoes conversion to C-14 and H-1 (a proton).
Cosmic radiation
Neutron
N-14
C-14
H-1 (proton)
C-14, along with other carbon isotopes (C-12, C-13), reacts with oxygen to form CO₂.
C-13
C-12
C-14
O₂
CO₂
Plants absorb all three isotopes of carbon as CO₂ during photosynthesis. Animals absorb C-14 when they consume plants.
Intake of C-14 stops when the organism dies. Concentration of C-14 decreases as it decays to N-14.
N-14
C-14
Beta particle
Older remains have less C-14 than younger ones.

Figure 11.95: Idea behind Carbon dating (source: ?)

About 2.5 atoms of carbon ^{14}C are formed every second by square centimetre on the Earth's surface (this value is however dependent on many factors but with negligible variations over the very long term - you can find entire textbooks on this subject), the positive contribution to the number of atoms of ^{14}C is about:

$$\dot{N}_+ = 2.54\pi R^2 \cong 1.2\cdot 10^{19}\,[\text{atoms}\cdot\text{s}^{-1}] \qquad (11.4.45)$$

R being the Earth's radius.

Or in mass flow it represents:

$$\dot{m}_+ \cong 1.2\cdot 10^{19}\frac{14\,[\text{g}\cdot\text{mole}^{-1}]}{6.0221367\cdot 10^{23}\,[\text{atoms}\cdot\text{mole}^{-1}]} \cong 2.98\cdot 10^{-4}\,[\text{g}\cdot\text{s}^{-1}] \qquad (11.4.46)$$

The disintegration rate is supposed equal to the rate of radioactive production, that is to say:

$$-\dot{N}_+ = \dot{N}_- = -1.28\cdot 10^{19}\,[\text{atoms}\cdot\text{s}^{-1}] \qquad (11.4.47)$$

As the global quantity:

$$\frac{\mathrm{d}(\dot{N}_+ + \dot{N}_-)}{\mathrm{d}t} = \frac{\mathrm{d}N}{\mathrm{d}t} \cong 0 \qquad (11.4.48)$$

As the disintegration rate is equal to:

$$A = \dot{N}_- = -\lambda N = \frac{-\ln(2)}{5700\cdot 365.256\cdot 24\cdot 60\cdot 60}\cdot N = -1.28\cdot 10^{19}\,[\text{atoms}\cdot\text{s}^{-1}] \qquad (11.4.49)$$

We conclude that there are $N \cong 3.32\cdot 10^{30}$ atoms of ^{14}C. The reader can calculate by multiplying by the mass value of the ^{14}C atom ($2.34\cdot 10^{-26}$ [kg]) that it makes quite a significant mass in the atmosphere.

As we already said it, this radioisotope is found in the chemical form CO_2 and penetrates through photosynthesis and metabolism in plants and animals. On the death of the plant or animal, the quantity of ^{14}C freezes and starts to decrease through radioactive decay over the ages:

$$^{14}\text{C} + {}^{12}\text{C} = N = c^{te} \qquad (11.4.50)$$

Radiocarbon dating is therefore a simple comparison between the concentration in ^{14}C of living matter and dead matter of two same objects/living entity. In fact, we determine the specific activities:

$$(A_{^{14}\text{C}})_{\text{Death entity}} \cong (A_{^{14}\text{C}})_{\text{Living entity}}\, e^{-\lambda(t-t_0)} \qquad (11.4.51)$$

where for objects "living entity" can be translate in "nowadays comparable element".

That is to say the "radiocarbon age" is given by:

$$t \cong -\frac{1}{\lambda}\ln\left(\frac{(A_{^{14}\text{C}})_{\text{Death entity}}}{(A_{^{14}\text{C}})_{\text{Living entity}}}\right) = -\frac{T_{1/2}}{\ln(2)}\ln\left(\frac{(A_{^{14}\text{C}})_{\text{Death entity}}}{(A_{^{14}\text{C}})_{\text{Living entity}}}\right) \qquad (11.4.52)$$

with for ^{14}C:

$$\frac{T_{1/2}}{\ln(2)} \cong 8267 \tag{11.4.53}$$

Therefore the latter relation is often written:

$$t \cong -8267 \ln\left(\frac{(A_{^{14}\text{C}})_{\text{Death entity}}}{(A_{^{14}\text{C}})_{\text{Living entity}}}\right) = 8267 \ln\left(\frac{(A_{^{14}\text{C}})_{\text{Living entity}}}{(A_{^{14}\text{C}})_{\text{Death entity}}}\right) \tag{11.4.54}$$

and even more shortly:

$$\boxed{t \cong 8267 \ln\left(\frac{A_0}{A}\right) = 8267 \ln\left(\frac{\lambda N_0}{\lambda N}\right) = 8267 \ln\left(\frac{N_0}{N}\right) = 8267 \ln\left(F_m\right)} \tag{11.4.55}$$

where F_m is the mass fraction notation.

Remark

Measurement of radiocarbon was originally done by beta-counting devices, which counted the amount of beta radiation emitted by decaying ^{14}C atoms in a sample. More recently, accelerator mass spectrometry has become the method of choice; it counts all the ^{14}C atoms in the sample and not just the few that happen to decay during the measurements; it can therefore be used with much smaller samples (as small as individual plant seeds), and gives results much more quickly.

The reader must keep in mind that three hypothesis (assumptions) have been implicitly done for the calculations above:

H1. The amounts of parent and daughter isotopes at the beginning, when the rock formed, must be known (the initial conditions) or their uncertainty impact on the resulting calculations results.

H2. All daughter atoms measured today must only have been derived by in situ radioactive decay of parent atoms (a closed system) in the opposite case the external sources must be taken into account in the uncertainty value of the resulting calculations results.

H3. The radioisotope decay rates are constant and if not the variations must be taken into account in the uncertainty value of the resulting calculations results.

Remark

The oldest dated rocks on Earth, as an aggregate of minerals that have not been subsequently broken down by erosion or melted, are more than 4 billion years old, formed during the Hadean Eon of Earth's geological history. Such rocks are exposed on the Earth's surface in very few places (some meteorites, such as the ALH84001 Mars meteorite found in the Allan Hills of Antarctica, are older but they were not formed on Earth). Some of the oldest surface rock can be found in the Canadian Shield, Australia, Africa and in a few other old regions around the world. The ages of these felsic rocks are generally between 2.5 and 3.8 billion years. The approximate ages have a margin of error of millions of years. In 11999 (holocene calendar), the oldest known rock on Earth was dated to 4.031 ± 0.003 billion years, and is part of the Acasta Gneiss of the Slave craton in northwestern Canada.

There are a lot of complementary (for double, triple checks) dating methods:

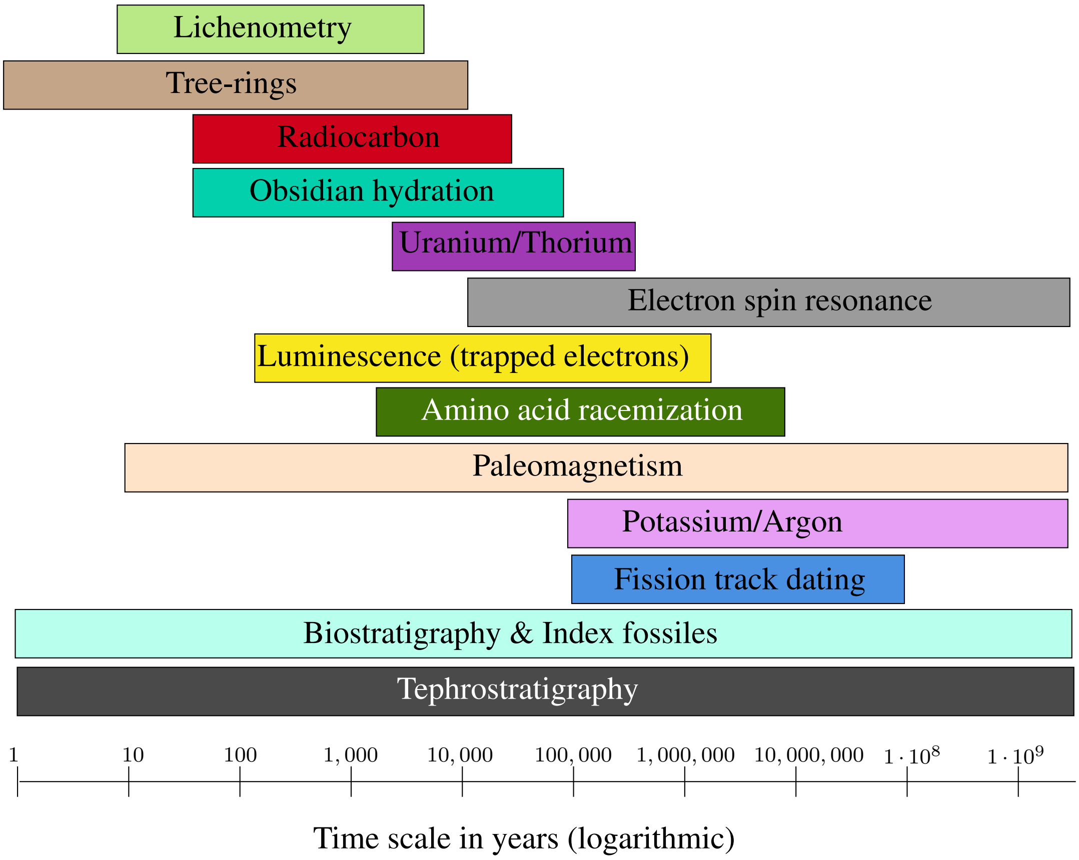


Figure 11.96: **Most common types of radiometric dating methods**

Some of them are also used because the accuracy of the others are not high enough with our 121st century (holocene calendar) instruments when the sample is too old.

Caution!!!! A common urban legend is that the older is a sample, the less accurate is its radioactive dating. In reality this an inaccurate statement formulated as previously. It should be instead formulated as following: the older is a given sample characterized by all its properties, the less accurate is its dating based on a single dating method (scientists switch to different dating method based on the age of a sample). For example Lucy bones were dated with an age of 3.18 million years old with a margin error of $\pm 30,000$ years (hence a margin error of 0.94%), and the oldest stones on Earth (see below for the details) were dated 4.295 billions years with a margin of 0.015 billions years, hence a margin error of 0.348% (less than the error margin of Lucy!!!).

Electron spin resonance (ESR), is also named "electron paramagnetic resonance" (EPR), may be grouped together with thermoluminescence (TL) and optically stimulated luminescence (OSL) as trapped charge dating methods (for a very good paper on ESR, see [218]).

A common technical accuracy limit for Carbon dating in this beginning of the 121st century (holocene calendar) is still $40,000$ years. Indeed, as the half life of Carbon-14 is about $5,700$ years, therefore with $40,000$ years there is under 1% of the original Carbon-14 left (0.79% to be exact), and it becomes too hard to measure it accurately. Stuff supposed to be very old like diamonds - a few hundred of millions of years old or greater - that are made only of carbon there is no trace of Carbon-14 left (see [529]).

Remark

As we will see in the section of Cosmology, looking far in space is equivalent at looking in the past (because light takes time to travel from one place to another, we see objects not as they are now but as they were at the time when they released the light that has travelled across the universe to us). Therefore modern telescopes are like time machines as we can go back as 13 billion years in the past by spectral observation. Thanks to spectroscopy we can check hydrogen spectral series that directly depends on the fine-structure constant (see page 4655) - as for the relation of activity decay that we have derived during our study of the tunnel effect (see page 4840) that depends only in constant parameters involved in the fine-structure constant - and see if the latter has changed these last 13 billion years. Experimental evidence as presented in [574] find a spatial variation of the order of $\sim 10^{-5}$ is preferred over a no-variation model at the 3.9σ level:

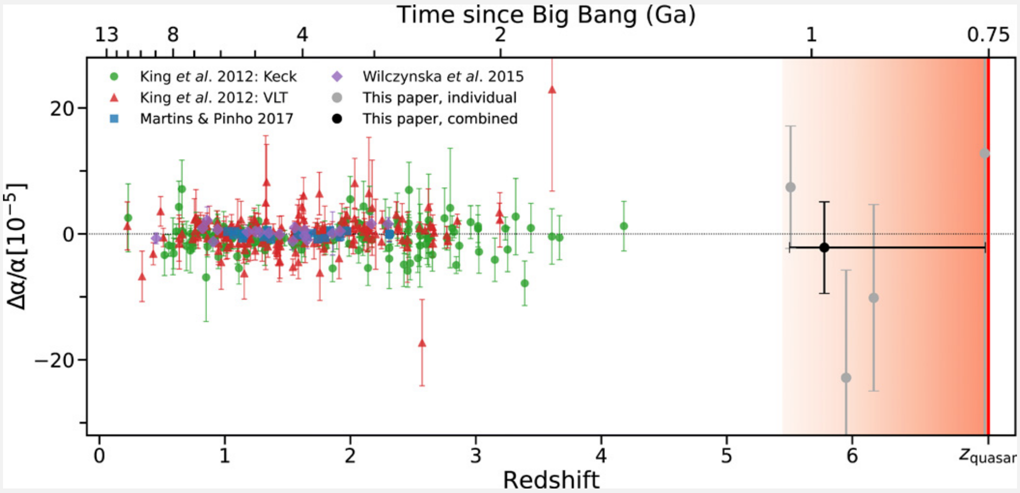


Figure 11.97: **Direct measurements of $\Delta\alpha/\alpha$, taken from quasar absorption measurements**

This variation of $\sim 10^{-5}$ of the fine-structure constant would lead in a relative variation in the radioactivity decay of the order of $\sim 10^{-4}$. That means on a datation of something of 1 billion year old, the error is approximately of $100,000$ years.

11.4.2.2.1.2 Rubidium-87 dating (meteorites dating/isochron dating)

Radioactive decay has become one of the most useful methods for determining the age of formation of rocks. However, in the very principal of radiometric dating there are several vital assumptions that have to be made in order for the age to be considered valid. These assumptions include for recall:

- The process must be irreversible[34]
- The laws and constants of physics didn't change during the considerated range time
- The initial amount of the daughter isotope is known
- Neither parent or daughter product has migrated into, or out of, the closed rock system[35]
- Decay has occurred at a constant rate over time.

But what if one or some combination of these assumptions is incorrect? Then the computed age based on the accumulation of daughter products will be obviously incorrect. In order to use the valuable information provided by radiometric dating, a new method had to be created that would determine a quite accurate date and validate the assumptions of radiometric dating. For this purpose, "isochron dating" was developed, a process that solves quite well the initial amount of the daughter isotope issue.

The clock most suitable for meteorites is the decay of Rubidium ($^{87}\mathrm{Rb}$) into Strontium ($^{87}\mathrm{Sr}$): which has a half-life of about 49 billion years. The manner in which the age is determined is based on calculating ratios of these isotopes, as the following calculation will show:

We know if there is some $^{87}\mathrm{Rb}$ present in the meteorite, that there will also be the decay product $^{87}\mathrm{Sr}$. However there will also be some unknown amount of $^{87}\mathrm{Sr}$ that was in the meteorite when it formed. We can state mathematically, that the amount of $^{87}\mathrm{Sr}$ present now, must have come from the amount that was there originally, plus any decay product from $^{87}\mathrm{Rb}$:

$$^{87}\mathrm{Sr}_{\text{now}}={}^{87}\mathrm{Sr}_{\text{original}}+\left({}^{87}\mathrm{Rb}_{\text{original}}-{}^{87}\mathrm{Rb}_{\text{now}}\right) \tag{11.4.56}$$

The term in parenthesis, the amount of $^{87}\mathrm{Rb}$ that decayed into $^{87}\mathrm{Sr}$ can be related by the radioactive decay law:

$$^{87}\mathrm{Rb}_{\text{original}}={}^{87}\mathrm{Rb}_{\text{now}}e^{\lambda t} \tag{11.4.57}$$

where as usual γ is the rate of radioactive decay, and t is the elapsed time.

By substituting that in the original equation we get:

$$^{87}\mathrm{Sr}_{\text{now}}={}^{87}\mathrm{Sr}_{\text{original}}+{}^{87}\mathrm{Rb}_{\text{now}}\left(e^{\gamma t}-1\right) \tag{11.4.58}$$

---

[34] Isotope dating satisfies this requirement, as daughter products do not decay back to the original parent element

[35] That "diffusion issue" is solved by taking multiple samples in a studied object - near the surface, in the middle and in-between multiples times - an incorporating the results in the classical statistical confidence interval that is always provided with isotopic dating! And repeating the process of multiple object found in the same area.

Along with $^{87}\mathrm{Sr}$, $^{86}\mathrm{Sr}$ also occurs in meteorites, but it is not a decay product and its amount does not change over time: so we can divide this constant in the above equation without changing the equality:

$$\frac{^{87}\mathrm{Sr}_{\text{now}}}{^{86}\mathrm{Sr}}=\frac{^{87}\mathrm{Sr}_{\text{original}}}{^{86}\mathrm{Sr}}+\frac{^{87}\mathrm{Rb}_{\text{now}}}{^{86}\mathrm{Sr}}\left(e^{\lambda t}-1\right) \tag{11.4.59}$$

Technically in the geochronology field is denoted:

$$\boxed{\frac{D}{D_i}=\frac{D_0}{D_i}+\frac{P}{D_i}\left(e^{\lambda t}-1\right)} \tag{11.4.60}$$

Note that this is the equation of a line in the form:

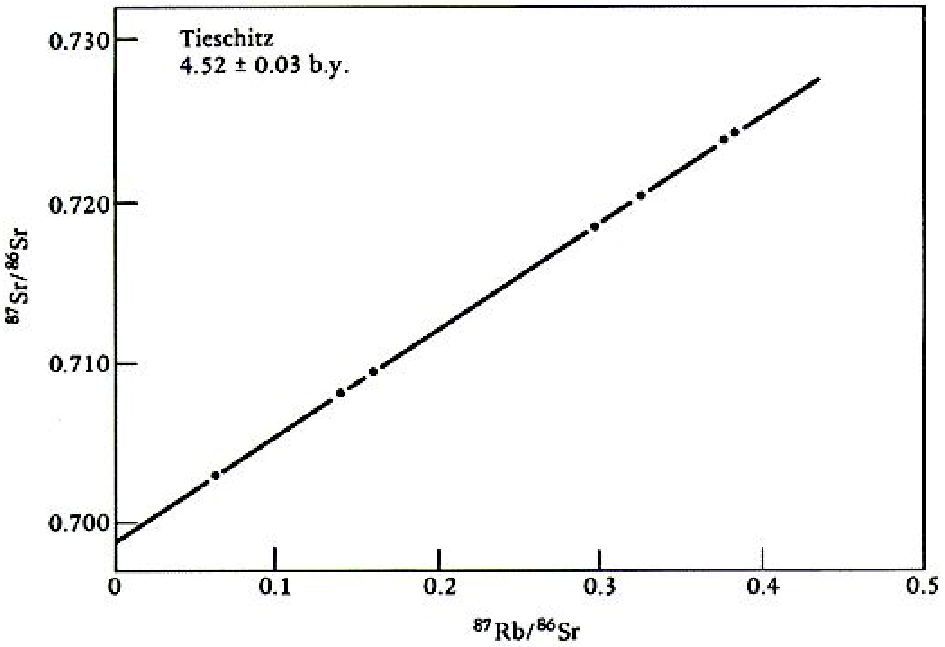
$$y=ax+b \tag{11.4.61}$$

where, a, the slope, is $(e^{\gamma t}-1)$ and b, the y intercept, is the original strontium isotope ratio.

And then we get straightforward:

$$t_a=\frac{1}{\lambda}\ln(a+1) \tag{11.4.62}$$

Two of these quantities can be measured: $\frac{^{87}\mathrm{Sr}_{\text{now}}}{^{86}\mathrm{Sr}}$ and $\frac{^{87}\mathrm{Rb}_{\text{now}}}{^{86}\mathrm{Sr}}$. By taking samples from various parts of a meteorite and plotting these results: the data will fall on a straight line whose slope characterizes the age of the meteorite. Theses lines are named "isochrons", an example for the meteorite Tieschitz (fall: 11878 - according to holocene calendar - Czechoslovakia) is shown in the following figure (from McSween):



Precise measurement of radioactive isotopes provides a way to determine the age of a meteorite. In the method illustrated here, a radioactive isotope of rubidium (^{87}Rb) decays to a stable isotope of strontium (^{87}Sr), which mixes with already existing strontium isotopes (^{87}Sr and ^{86}Sr). The individual data points in this diagram represent minerals separated from the Tieschitz (Czechoslovakia) chondrite. The age of the meteorite (4.52 billion years) is calculated from the slope of the diagonal line, which steepens with increasing time.

So the trick to the isochron diagram is the normalization of both parent and daughter isotope to a third isotope. This third isotope is the non-decay product isotope of the same element as the daughter element. In the initial state, the graph of daughter isotope to the third isotope versus parent isotope to the third isotope should result in a straight, horizontal line.

The process of evaluating the daughter product as a ratio against another isotope of the same element is a quite valid method because, when a mineral or rock forms from a homogeneous state, the elements that are assimilated into crystalline formation are very restricted. The key to the formation of crystals in the rock is that the process is selective between elements, but is indifferent to isotopes of the same element. Thus, the daughter product and any other isotopes of the same element will be incorporated into the minerals of a rock with the same ratio. This initial ratio allows the non-daughter product isotope to be representative of the initial amount of the daughter product. This assumes there has been no migration of daughter or parent nuclei within the matrix since its solidification. It also assumes that the daughter isotopes had sufficient time to uniformly distribute themselves throughout the material matrix; i.e., crystallization cannot be too rapid.

As time progresses and decay occurs, the number of atoms of the parent isotope decreases, and the number of atoms of the daughter isotope increases accordingly. The amount of non-decay isotope in the sample does not change. Thus, as decay occurs, the parent ratio decreases and the daughter ratio increases.

The features of the isochron method provide a way do reduce doubt and speculation about an age that is computed using these methods. Based on the assumptions of basic radioactive dating, the problem of an unknown initial amount of daughter isotope is eliminated by the definition of the isochron itself. The problem of contamination is "self-checking". If contamination has occurred within a sample, the ratios from the sample shouldn't fall on a line. Instead, the points would be in a scatter on the graph. Points that do not fall on a straight line suggest contamination, and this invalidates the results. However, by this same principle, points falling relatively close to a best fit line should provide an accurate date for the age of the rock being dated.

All the different radiometric dating methods have provided successful scientifically peer review independent evidence (with control group) that:

- Young Earth belief according to the Genesis of the Abrahamic religion is wrong.
- The Shroud of Turin (see [123]) is a fake relic like most relics of the catholic church (the fact that the shroud of Turn is a fake relic is also confirmed by historical analysis [478]).
- The oldest Hominids is dated from approximately 4.2 million years ago (fact that goes against all religions beliefs worldwide)
- etc...

Exposure of the metamorphosed rocks of the Nuvvuagittuq Greenstone Belt (NGB) from the Porpoise Cove locality, Quebec, Canada. Some of these rocks have an Uranium (U)/Lead (Pb) dating of minimum 3.77 billions years (see [90] and [425]) and a Samarium (Sm)/Neodymium (Nd) age between 4.28 and 4.31 billion years (see [425] and [426]) and may be the oldest rocks on Earth.

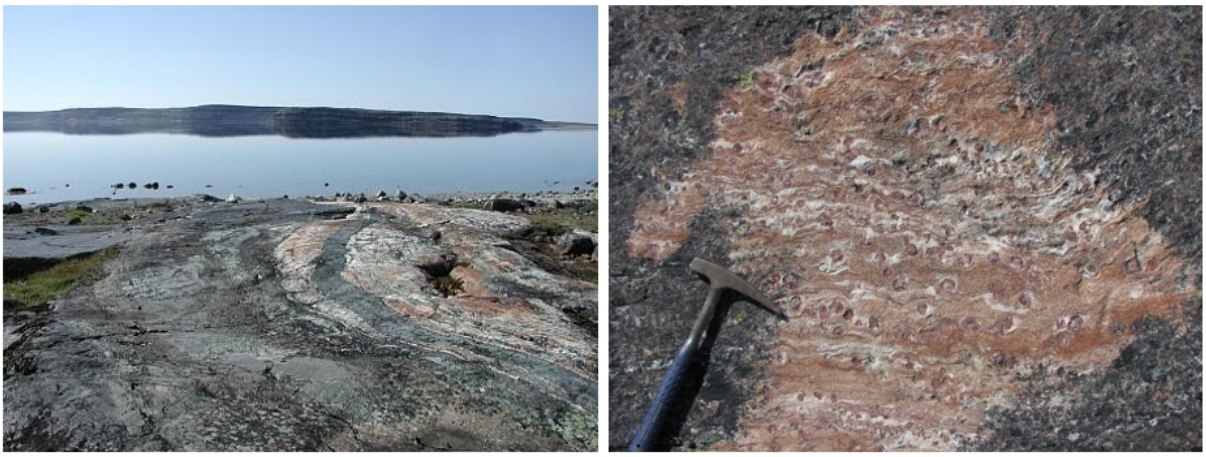


Figure 11.98: **Exposure of the metamorphosed rocks of the Nuvvuagittuq Greenstone Belt (source: left NASA, right Jonathan O'Neil)**

Researchers have determined (see [153]) that fossilized evidence of bacteria from ancient seafloor hydrothermal vent precipitates found in the Nuvvuagittuq belt in Quebec, Canada is dated with Uranumium (U)-Lead (Pb) at least 3.77 billion years old (or even as much as 4.28 billion years old). The minimum age of the fossils would make them the oldest indication of life on Earth so far.

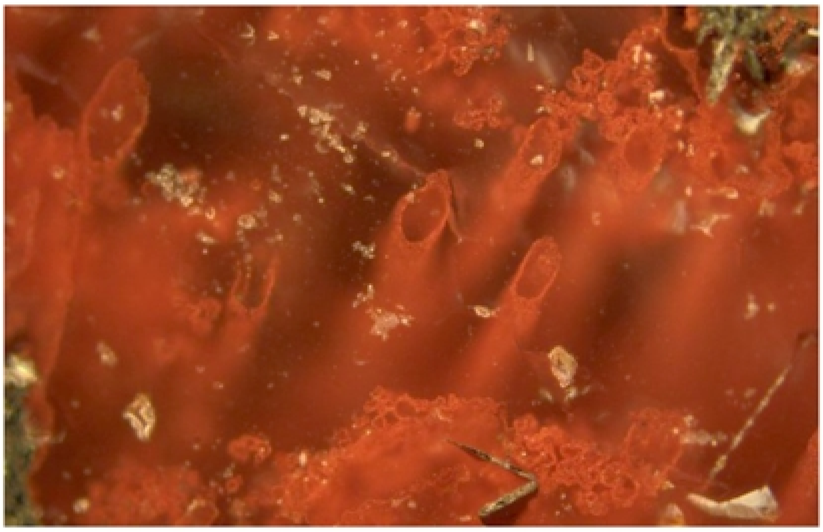


Figure 11.99: Hematite tubes from microfossils dating back to 3.770 **billion years (source: [153])**

Again and as always, Science helps to highlight the lies of all religions.

For more details and techniques about Radiometric dating, the reader can refer the to excellent following textbook [210].

11.4.2.2.1.3 Young Earth heat problem

One of the issues in the creation/evolution debate is the claim, in a few thousands of religions among the dozens of thousands of existing one, that radioactive decay is not constant. In order for the Earth to appear old[36], the creationists of these religions must assume that radioactive decay rates in the past were faster. This solves the problems of old ages in rocks, but unfortunately opens up a bigger problem. Radioactive decay gives off heat. The amount of heat generated is proportional to the rate of decay and the amount of radioactive material present at the time. In the following exercise, I show how much heat can be generated by radioactive decay if decay rates were faster in the past. Please note that the following analysis uses present-day heat production values which are observable quantities and I have also assumed (in the creationist analysis) that each of the elements has the same rate of decay. Changing the decay rates to those observed today for each of the major heat producing elements makes the mathematics a bit more tedious, but does not significantly alter the conclusions given here.

Present day of heat production from radioactive decay in the Earth is produced mainly by the isotopes ^{238}U, ^{235}U, ^{232}Th and ^{40}K and has a value of $6.18\cdot 10^{-12}\;[\text{W}\cdot\text{kg}^{-1}]$ ([546]). If the heat flow out of the Earth were due solely to radioactive decay in the mantle and crust, it would equal Heat Production · (Mass of Mantle + Crust), i.e.:

$$Q=6.18\cdot 10^{-12}\;[\text{W}\cdot\text{kg}^{-1}]\cdot 4.0\cdot 10^{24}\;[\text{kg}]\cong 2.47\cdot 10^{13}\;[\text{W}] \tag{11.4.63}$$

The heat flow out of the Earth can be calculated from Fourier's Law of heat conduction (see section Thermodynamics page 3684):

$$\Phi_q=\kappa\frac{\mathrm{d}T}{\mathrm{d}z} \tag{11.4.64}$$

where k is the thermal conductivity of the rocks in $[\text{W}\cdot\text{m}^{-1}\cdot{}^\circ\text{C}^{-1}]$, $\mathrm{d}T/\mathrm{d}z$ is the geothermal gradient (change in temp with depth) in $[^\circ\text{C}\cdot\text{m}^{-1}]$ and obviously Φ_q is in $[\text{W}\cdot\text{m}^{-2}]$.

To calculate the surface heat flow, we divide the result of:

$$Q=6.18\cdot 10^{-12}\;[\text{W}\cdot\text{kg}^{-1}]\cdot 4.0\cdot 10^{24}\;[\text{kg}]\cong 2.47\cdot 10^{13}\;[\text{W}] \tag{11.4.65}$$

by the total surface area of the earth ($\cong 5.1\cdot 10^{14}\;[\text{m}^2]$), we get:

$$\Phi_q=\frac{2.47\cdot 10^{13}\;[\text{W}]}{5.1\cdot 10^{14}\;[\text{m}^2]}\cong 0.048\;[\text{W}\cdot\text{m}^{-2}] \tag{11.4.66}$$

If we now use this to calculate the temperature-depth profile $\mathrm{d}T/\mathrm{d}z$, we need to use an average value of thermal conductivity. We use $3.00\;[\text{W}\cdot\text{m}^{-1}\cdot{}^\circ\text{C}^{-1}]$ as an average thermal conductivity of the rocks. This yields:

$$\frac{\mathrm{d}T}{\mathrm{d}z}\cong\frac{0.048\;[\text{W}\cdot\text{m}^3]}{3\;[\text{W}\cdot\text{m}^{-1}\cdot{}^\circ\text{C}^{-1}]}\cong 16\;[^\circ\text{C}\cdot\text{km}^{-1}] \tag{11.4.67}$$

Thus, for every 1 kilometer depth, the temperature will increase by about $16\;[^\circ\text{C}]$. Obviously as we get deeper into the crust and mantle, temperature increases are not linear because they are a function of pressure and other factors. Nevertheless, this relationship is useful to depths of $30-70$ kilometers. Therefore, at a depth of 10 kilometers, the temperature would be about $160\,[^\circ\text{C}]$. Basaltic rock melts at about $1200\,[^\circ\text{C}]$. There are some simplifications made in the above calculations since much of the heat production $(\sim 67\%)$ is confined to the crust, but the illustration below is useful regardless of this simplification.

---

[36] Bigots think that the only evidence of an old Earth is radioactivity but scientist have multiple evidence that the Earth is much older than 6,000 years old: radioactivity, stratigraphy, archaeology, palaeontology, thermodynamics, astrophysics, geology, astronomy, dendrochronology, linguistic, crystallography

Let's look at these same relationships with a much younger Earth and a faster rate of decay. In our example, we will use a quite standard creationist age (in the christian faith) for the Earth of $6,000$ years. In order to carry out the analysis, we must still assume some sort of decay constant. There are a number of ways this could be done, but the simplest is to assign an average half-life. We will use 500 years as the average half life for the radioactive elements listed above. In reality, we are being overly generous to the young Earth crowd. If rates were variable, they would probably have been much faster than this and the resultant earth conditions would be more extreme. The interested reader can easily play with the variables on a spreadsheet and see how they affect the conclusions reached below.

Remark

It is possible, though more tedious to calculate the heat release using a decelerating decay rate. It does not change the basic conclusion reached in the following developments. In order to do this, one simply replaces λ with a function $e^{-\text{rate}}$.

In order to figure out the rate of heat production in the past, we must first calculate a decay constant λ. This is calculated from (see section Nuclear Physics page 5030):

$$T_{1/2}=\frac{\ln(2)}{\lambda} \qquad (11.4.68)$$

or:

$$\lambda=\frac{\ln(2)}{T_{1/2}} \qquad (11.4.69)$$

using 500 years for $T_{1/2}$, we get $\lambda=1.38\cdot 10^{-3}\,[\text{s}^{-1}]$.

The rate of heat production at any time in the past is given by the following formula (see section Nuclear Physics page 5029):

$$H_{\text{past}}=H_{\text{present}}\cdot e^{\lambda(t-t_0)} \qquad (11.4.70)$$

we use $H_{\text{present}}=6.18\cdot 10^{-12}\,[\text{W}\cdot\text{kg}^{-1}]$ and $t-t_0=6,000$ years and λ as calculated above and we get:

$$H_{\text{past}}=2.43\cdot 10^{-8}\,[\text{W}\cdot\text{kg}^{-1}] \qquad (11.4.71)$$

We can now use this value to calculate the amount of heat flow out of the Earth due solely to this ancient ($6,000$ year old...) heat production and calculate a thermal profile as we did above using first:

$$H_{\text{past}}\cdot(\text{Mass of Mantle}+\text{crust})=\text{total heat production}$$

Hence:

$$H_{\text{past}}=2.43\cdot 10^{-8}\,[\text{W}\cdot\text{kg}^{-1}]\cdot 4.0\cdot 10^{24}\,[\text{kg}]\cong 9.75\cdot 10^{16}\,[\text{W}] \qquad (11.4.72)$$

We now divide this value by the surface area of the Earth to get the heat flow Q at $6,000$ years ago:

$$\Phi_q=\dfrac{9.75\cdot 10^{16}\,[\text{W}]}{5.1\cdot 10^{14}\,[\text{m}^2]}\cong 191\,[\text{W}\cdot\text{m}^{-2}] \tag{11.4.73}$$

Using the same value for thermal conductivity $3.00\,[\text{W}\cdot\text{m}^{-1}\cdot{}^\circ\text{C}]$, we can calculate the $\mathrm{d}T/\mathrm{d}z$ profile as earlier above in:

$$\dfrac{\mathrm{d}T}{\mathrm{d}z}=\dfrac{191\,[\text{W}\cdot\text{m}^{-2}]}{3.00\,[\text{W}\cdot\text{m}^{-1}\cdot{}^\circ\text{C}^{-1}]}\cong 65,000\,[{}^\circ\text{C}\cdot\text{km}^{-1}] \tag{11.4.74}$$

At $6,000$ years ago, it is pretty obvious that the entire Earth would be molten and Adam and Eve's goose was cooked. We are not implying that this is a purely conductive regime. What we are showing is that the thermal regime on a $6,000$ year old Earth with rampant radioactive decay is extreme. If you want to turn off the engine and let the Earth cool from this extreme you also run into time spans much greater than $6,000$ years because it will take a while to cool down from those extreme temperatures. It's difficult to estimate how long it would take for the Earth to reach present-day temperatures... It is clear that rapid decay results in a tremendous release of heat. We suppose some could always argue that the Earth has been cooling since this initial heating and that there is no further significant heat released by radioactive decay. In that case, the cooling of the Earth reduces to the Kelvin problem and gives an age of 20-60 million years (and 2000 thousand years ago the geothermal gradient would have been around $\sim 400\,[{}^\circ\text{C}\cdot\text{km}^{-1}]$).

11.4.2.2.2 Radioactive Decay chain

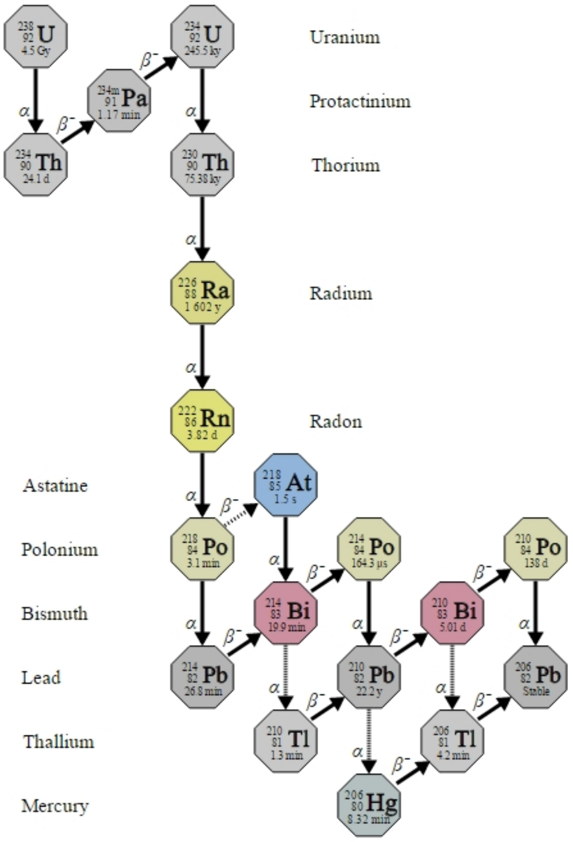
Definition (#11.4.513): A "decay chain" refers to the radioactive decay of different discrete radioactive decay products as a chained series of transformations. They are also known as "radioactive cascades". Most radioisotopes do not decay directly to a stable state, but rather undergo a series of decays until eventually a stable isotope is reached.

Caution!!!! Not only nucleus of atoms decays. But also most particles and even some fundamental particles (yes! but they are not made of the stuff they decay of as experience show us with the τ fundamental particle that has at least 30 different decay chains). However some particles don't decay (as far as we know) like: the photon, the electron (anti-electron), the lightest neutrino (and its anti-particle[a] and (presumed) the graviton.

---

[a] Half-integer spin particles have well-defined antiparticles, integer spin particles exhibit this fascinating self-duality.

Decay stages are referred to by their relationship to previous or subsequent stages. A "parent isotope " is one that undergoes decay to form a "daughter isotope". One example of this is uranium (atomic number 92) decaying into thorium (atomic number 90). The daughter isotope may be stable or it may decay to form a daughter isotope of its own. The daughter of a daughter isotope is sometimes named a "granddaughter isotope".



238 92 U 4.5 Gy
234 92 U 245.5 ky
234m 91 Pa 1.17 min
234 90 Th 24.1 d
230 90 Th 75.38 ky
226 88 Ra 1 602 y
222 86 Rn 3.82 d
218 85 At 1.5 s
218 84 Po 3.1 min
214 84 Po 164.3 µs
210 84 Po 138 d
214 83 Bi 19.9 min
210 83 Bi 5.01 d
214 82 Pb 26.8 min
210 82 Pb 22.2 y
206 82 Pb Stable
210 81 Tl 1.3 min
206 81 Tl 4.2 min
206 80 Hg 8.32 min
α
β⁻
Uranium
Protactinium
Thorium
Radium
Radon
Astatine
Polonium
Bismuth
Lead
Thallium
Mercury

Figure 11.100: **Uranium decay series (source: Wikipedia)**

In general we denote a radioactive cascade by:

$$\mathrm{X}_1^* \rightarrow \mathrm{X}_2^* \rightarrow \mathrm{X}_3^* \rightarrow \ldots \mathrm{X}_n \tag{11.4.75}$$

where the small symbol * denotes a given radioactive isotope, X_n being the final stable isotope of the radioactive of the mother element X_1^* (the elements between all being unstable nuclides).

11.4.2.3 Two level radioactive cascade

Let us consider the two nuclide situation (we will not focus on higher order level excepted on request of many readers).

Let us suppose that at the beginning of time, the first descendant exists only negligible quantity such that we have for initial conditions (IC) at $t=0$:

$$\begin{aligned} N_1(t=0)&=N_{1_0}\\ N_2(t=0)&=0 \end{aligned} \tag{11.4.76}$$

The variation of the parent nuclide (1) is given by a negative contribution: the disintegration of (1).

We have:

$$\mathrm{d}N_1=-\lambda_1 N_1 \mathrm{d}t \tag{11.4.77}$$

having for solution taking into account the initial conditions:

$$N_1(t)=N_{1_0}e^{-\lambda_1 t} \tag{11.4.78}$$

The variation of the descendant member (2), that is to say the daughter of (1), is given by a positive contribution (the atoms of (1) disintegrated) and one negative, the disintegration of (2). Then:

$$\mathrm{d}N_2=+\lambda_1 N_1\mathrm{d}t-\lambda_2 N_2 \mathrm{d}t \Rightarrow \frac{\mathrm{d}N_2}{\mathrm{d}t}+\lambda_2 N_2=+\lambda_1 N_1=\lambda_1 N_{1_0}e^{-\lambda_1 t} \tag{11.4.79}$$

So we have to solve this differential equation (see section Differential and Integral Calculus page 1609.

We have for homogeneous solution (characteristic equation):

$$\frac{\mathrm{d}N_2}{\mathrm{d}t}+\lambda_2 N_2=0 \tag{11.4.80}$$

We derive from the homogeneous equation the special (homogeneous) solution:

$$N_2^h(t)=N_{2_0}^h e^{-\lambda_2 t} \tag{11.4.81}$$

with the letter h to signify that this is the "homogeneous" solution.

Let us now determine the particular solution of:

$$\frac{\mathrm{d}N_2}{\mathrm{d}t}+\lambda_2 N_2=\lambda_1 N_{1_0}e^{-\lambda_1 t} \tag{11.4.82}$$

The idea to solve this is to put:

$$N_2^p(t)=N_{2_0}^p e^{-\lambda_p t} \tag{11.4.83}$$

with the letter p for "particular" (implicitly: "special solution"). Substituting we find:

$$-\lambda_p\left(N_{2_0}^p e^{-\lambda_p t}\right)+\lambda_2\left(N_{2_0}^p e^{-\lambda_p t}\right)=(\lambda_2-\lambda_p)N_{2_0}^p e^{-\lambda_p t}=\lambda_1\left(N_{1_0}e^{-\lambda_1 t}\right) \tag{11.4.84}$$

Therefore we must have:

$$\Rightarrow \lambda_p=\lambda_1 \tag{11.4.85}$$

because otherwise if we had $\lambda_p=\lambda_2$ we would have a zero equality which is absurd. Therefore getting rid of the exponential we get:

$$-\lambda_1 N_{2_0}^p+\lambda_2 N_{2_0}^p=\lambda_1 N_{1_0} \tag{11.4.86}$$

from which we derive that:

$$N_{2_0}^p=\frac{\lambda_1 N_{1_0}}{\lambda_2-\lambda_1} \tag{11.4.87}$$

Finally the general solution is the sum of the homogeneous solution and the particular one, therefore:

$$N_2(t)=N_{2_0}^h e^{-\lambda_2 t}+\frac{\lambda_1 N_{1_0}}{\lambda_2-\lambda_1}e^{-\lambda_1 t} \tag{11.4.88}$$

Let us apply the initial conditions:

$$t=0 \qquad N_2(t=0)=0\Rightarrow N_{2_0}^h=\frac{-\lambda_1 N_{1_0}}{\lambda_2-\lambda_1} \tag{11.4.89}$$

Finally we have:

$$\boxed{N_2(t)=\frac{\lambda_1 N_{1_0}}{\lambda_2-\lambda_1}\left(e^{\lambda_1 t}-e^{-\lambda_2 t}\right)} \tag{11.4.90}$$

In geology and archaeology the latter is often rearrange and denoted:

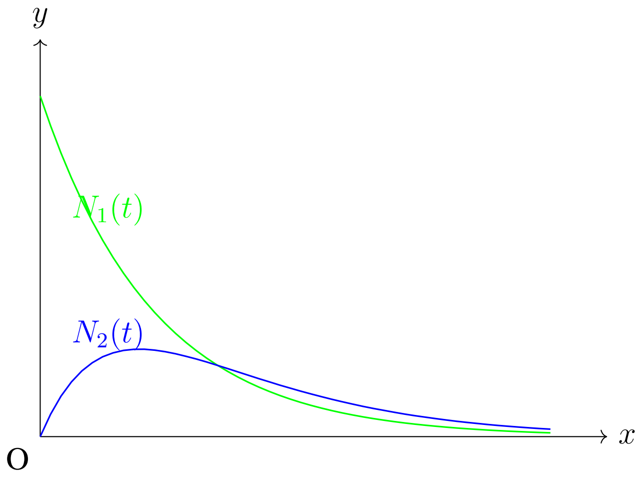
$$t=\frac{\ln\left(1-\dfrac{N_2(t)}{N_{1_0}}\dfrac{\lambda_2-\lambda_1}{\lambda_1}\right)}{\lambda_2-\lambda_1} \tag{11.4.91}$$

as all the parameters in the right hand side of the equality can be determined in a laboratory nowadays with less than tenth of thousand of percent, we can then determine age of rocks (and of many other stuffs).

If we plot:

$$\begin{aligned}N_1(t)&=N_{1_0}e^{-\lambda_1 t}\\N_2(t)&=\frac{\lambda_1 N_{1_0}}{\lambda_2-\lambda_1}\left(e^{\lambda_1 t}-e^{-\lambda_2 t}\right)\end{aligned} \tag{11.4.92}$$

we get:



$N_2(t)$ being zero for $t=0$ and for $t=+\infty$, it must pass, as the for activity $A_2(t)$, through a maximum. Given t^* the time when the maximum is observed, we have:

$$\begin{aligned}&\left.\dfrac{\mathrm{d}A_2}{\mathrm{d}t}\right|_{t=t^*}=\left.\dfrac{\mathrm{d}(\lambda_2 N_2)}{\mathrm{d}t}\right|_{t=t^*}=0=\lambda_2\dfrac{\lambda_1 N_{1_0}}{\lambda_2-\lambda_1}\left(-\lambda_1 e^{-\lambda_1 t^*}+\lambda_2 e^{-\lambda_2 t^*}\right)\\&\Rightarrow\left(-\lambda_1 e^{-\lambda_1 t^*}+\lambda_2 e^{-\lambda_2 t^*}\right)=0\end{aligned}\tag{11.4.93}$$

from where:

$$t^*=\dfrac{\ln\left(\dfrac{\lambda_1}{\lambda_2}\right)}{\lambda_1-\lambda_2}\tag{11.4.94}$$

The knowledge of t^* is important especially in nuclear medicine where we want to administer the product 1 for radio-diagnostic purposes and minimize adverse effects of the daughter(s) element(s) of 1. We then choose elements such that the time t^* is greater than the biological elimination time (body's natural elimination pathways) of its daughter(s) elements(s). We will return on this subject in a few paragraphs after studying three typical scenarios of radioactive filiation:

11.4.2.3.1 Secular equilibrium

This type of relation between mother and daughter activities occurs when the mother nucleus half-life is much greater than that of the daughter nucleus half-life. In other words, the radioactive decay of the mother nucleus is negligible and the activity of the descendent tends toward that of the parent.

We then have in this particular case:

$$\lambda_1\ll\lambda_2\tag{11.4.95}$$

So we have for the activity using the previously proved relation:

$$\begin{aligned}\lambda_2 N_2(t)=A_2(t)&=\dfrac{\lambda_2\lambda_1 N_{1_0}}{\lambda_2-\lambda_1}\left(e^{-\lambda_1 t}-e^{-\lambda_2 t}\right)=\dfrac{\lambda_2}{\lambda_2-\lambda_1}\left(e^{-\lambda_1 t}-e^{-\lambda_2 t}\right)\\&\underset{\lambda_1\ll\lambda_2}{=}A_{1_0}\left(e^{-\lambda_1 t}-e^{-\lambda_2 t}\right)\end{aligned}\tag{11.4.96}$$

Therefore:

$$A_2(t) = A_{1_0}\left(e^{-\lambda_1 t} - e^{-\lambda_2 t}\right) \tag{11.4.97}$$

We also see that in the case where $\lambda_1 \ll \lambda_2$ and $t \to +\infty$, we have:

$$\boxed{A_2(t) = A_{1_0}\left(e^{-\lambda_1 t} - e^{-\lambda_2 t}\right) \cong A_{1_0} e^{-\lambda_1 t} = A_1(t)} \tag{11.4.98}$$

in other words, the mother and daughter activities, become equivalent after a large enough time. For example, after a period of a half-life of the daughter isotope, we already have the daughter activity that is 50% that of the parent activity:

$$A_2(T_{1/2}) \cong A_{1_0} e^{-\lambda_1 T_{1/2}} = \frac{1}{2} A_{1_0} \tag{11.4.99}$$

If we have the case where the following approximation is acceptable:

$$\lambda_1 \cong 0 \tag{11.4.100}$$

we will have:

$$A_2(t) \cong A_{1_0}\left(1 - e^{-\lambda_2 t}\right) \tag{11.4.101}$$

11.4.2.3.2 Transient equilibrium

This type of relation between mother and daughter activity occurs when the half-life of the mother nucleus is greater than that of the daughter nucleus (but not much much greater in contrast to the secular equilibrium). In other words, the decay radioactivity of the mother nucleus and the activity of the descendants are equal to a constant factor (in other words, their radioactive decay curves are parallel after a sufficiently long time).

We then in this particular case:

$$\lambda_1 < \lambda_2 \tag{11.4.102}$$

So we have for the activity using the previously prove relation:

$$\lambda_2 N_2(t) = A_2(t) = \frac{\lambda_2 \lambda_1 N_{1_0}}{\lambda_2 - \lambda_1}\left(e^{-\lambda_1 t} - e^{-\lambda_2 t}\right) = \frac{\lambda_2 A_{1_0}}{\lambda_2 - \lambda_1}\left(e^{-\lambda_1 t} - e^{-\lambda_2 t}\right) \tag{11.4.103}$$

After a long enough time, we get:

$$\boxed{A_2(t) \cong \frac{\lambda_2 A_{1_0}}{\lambda_2 - \lambda_1} e^{-\lambda_1} t = \frac{\lambda_2}{\lambda_2 - \lambda_1} A_1(t)} \tag{11.4.104}$$

where we see that the factor:

$$\frac{\lambda_2}{\lambda_2 - \lambda_1} \tag{11.4.105}$$

is greater than unity. So after a sufficiently long time, not only the activity of the daughter isotope is parallel to that of the mother but in addition it is superior to that latter.

11.4.2.3.3 Non-equilibrium

Here the half life time of the daughter element is greater than that of the parent element. In other words we have the hypothesis:

$$\lambda_1 > \lambda_2 \tag{11.4.106}$$

This implies for recall that the probability of decay of the parent element 1 is greater than that of the daughter element 2 which is logical here.

The activity of the daughter isotope in the sample then increases according to the relation proved just above:

$$A_2(t) = \frac{\lambda_2 A_{1_0}}{\lambda_2 - \lambda_1}\left(e^{-\lambda_1 t} - e^{-\lambda_2 t}\right) \tag{11.4.107}$$

Finally, after a sufficiently long time, only the activity of the daughter element will remain, since the activity of the parent element disappears at a higher rate following:

$$\boxed{A_2(t) \cong \frac{\lambda_2 A_{1_0}}{\lambda_2 - \lambda_1}\left(-e^{-\lambda_2 t}\right) = \frac{\lambda_2}{\lambda_1 - \lambda_2} e^{-\lambda_2 t}} \tag{11.4.108}$$

After a time $t_{\max}$, the activity of the daughter element reach a maximum value for:

$$\begin{aligned}\frac{\mathrm{d}A_2(t)}{\mathrm{d}t}=0\cong\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\lambda_2 A_{1_0}}{\lambda_2-\lambda_1}\left(e^{-\lambda_1 t}-e^{-\lambda_2 t}\right)\right)&=\frac{\lambda_2 A_{1_0}}{\lambda_2-\lambda_1}\frac{\mathrm{d}}{\mathrm{d}t}\left(e^{-\lambda_1 t}-e^{-\lambda_2 t}\right)\\&=\frac{\lambda_2 A_{1_0}}{\lambda_2-\lambda_1}\left(\lambda_2 e^{-\lambda_2 t}-\lambda_1 e^{-\lambda_1 t}\right)\end{aligned}\tag{11.4.109}$$

Therefore:

$$\frac{\lambda_2 A_{1_0}}{\lambda_2-\lambda_1}\left(\lambda_2 e^{-\lambda_2 t}-\lambda_1 e^{-\lambda_1 t}\right)=0\tag{11.4.110}$$

This simplifies to:

$$\lambda_1 e^{-\lambda_1 t}=\lambda_2 e^{-2\lambda_2 t}\tag{11.4.111}$$

It then comes immediately the result already proved in the example earlier above:

$$t_{\max}=\frac{1}{\lambda_2-\lambda_1}\ln\left(\frac{\lambda_2}{\lambda_1}\right)\tag{11.4.112}$$

Finally, let us consider the extreme case of the situation of non-equilibrium consisting to consider the case where:

$$\lambda_2=0\tag{11.4.113}$$

In other words the daughter element is not radioactive. We then fall back on the classic case:

$$A_2(t)=\frac{\lambda_2 A_{1_0}}{\lambda_2-\lambda_1}\left(e^{-\lambda_1 t}-e^{-\lambda_2 t}\right)=0\tag{11.4.114}$$

11.4.2.4 Radioactive phenomena

When we "weigh" a nucleus, we experimentally observe a very important fact !: its mass is less than the sum of the masses of its constituents. This difference is named the "mass defect" and is relatively well determined with simplistic theoretical models.

The mass defect is given by construction by:

$$\Delta m=Zm_p+(A-Z)m_n-m_{\text{nucleus}}\tag{11.4.115}$$

with m_{nucleus} being the mass of the nucleus in its fundamental state, m_p the proton mass and m_n the neutron mass.

The mass of a set of linked nucleons is less than the sum of the masses of individual nucleons (far enough of each other to not interact). We get from Special Relativity (see section of the same name page 5679):

$$\Delta m=\frac{1}{c^2}E_{\text{bound}}^{\text{nucleons}}\tag{11.4.116}$$

where $E_{\text{bound}}^{\text{nucleons}}$ is the binding energy of the nucleons that compose the nucleus (>0).

Δm is therefore positive for all the elements (emission of energy and therefore of mass to the outer system). If this were not the case, the nucleons have no reason to come together naturally to form stable nuclei (or more stable ...).

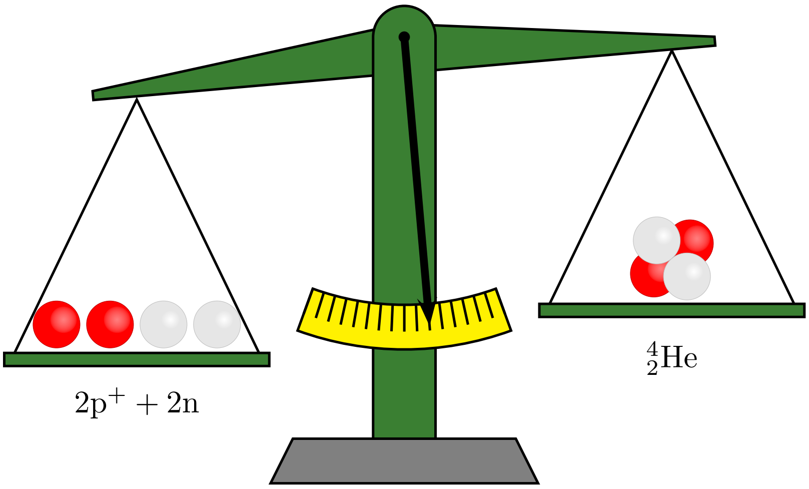


Figure 11.101: Mass defect concept illustrated for Helium

Given $\bar{B}=f(A,Z)$ the average energy by nucleon of a given atom. We have then:

$$\boxed{\bar{B}=\frac{\left[Zm_p+Nm_n-m_{\text{nucleus}}(A,Z)\right]c^2}{A}} \tag{11.4.117}$$

which is by convention a positive value!

Notice that the mass of the nucleus is linked to the mass of the atom by:

$$\boxed{m_{\text{atom}}(Z,A)=m_{\text{nucleus}}(Z,A)+Zm_e} \tag{11.4.118}$$

Similarly, the nucleus mass added to the weight of its isolated electrons is higher than that of the nucleus surrounded by its linked electrons. Let us notice that the electronic binding energy can often neglected relatively to that of the nucleus and it is a rule we will adopt throughout this section.

The energy released during the nuclear fusion, that is to say during the constitution of the atom from its constituents, also named "binding energy" (name that often makes interpretations problems to young students) because it is that energy must be given if we want at the opposite, separate the elements. We must never forget that behind the term "binding energy", so there is the energy variation between individual elements and combined elements of an atomic element.

The general practical expression of the average energy per nucleon of a given atom expressed in atomic mass units is:

$$\bar{B}=\frac{Z\cdot 1.00727647+N\cdot 1.008664904-m_{\text{nucleus}}(A,Z)}{A}\cdot 931.49431\;[\text{MeV}] \tag{11.4.119}$$

A numerical application for Helium (α particle) gives ($Z=N=2,\; A=4$) we get $\bar{B}\cong 7.07$ [MeV/nucleon].

The principles of nuclear energy production from fission or fusion result of the shape of the average energy per nucleon according to A.

We have in reality the following curve linking the average energy per nucleon (that is to say, the average energy variation between the nucleon alone and accompanied...) and the number of nucleons named "Aston curve":

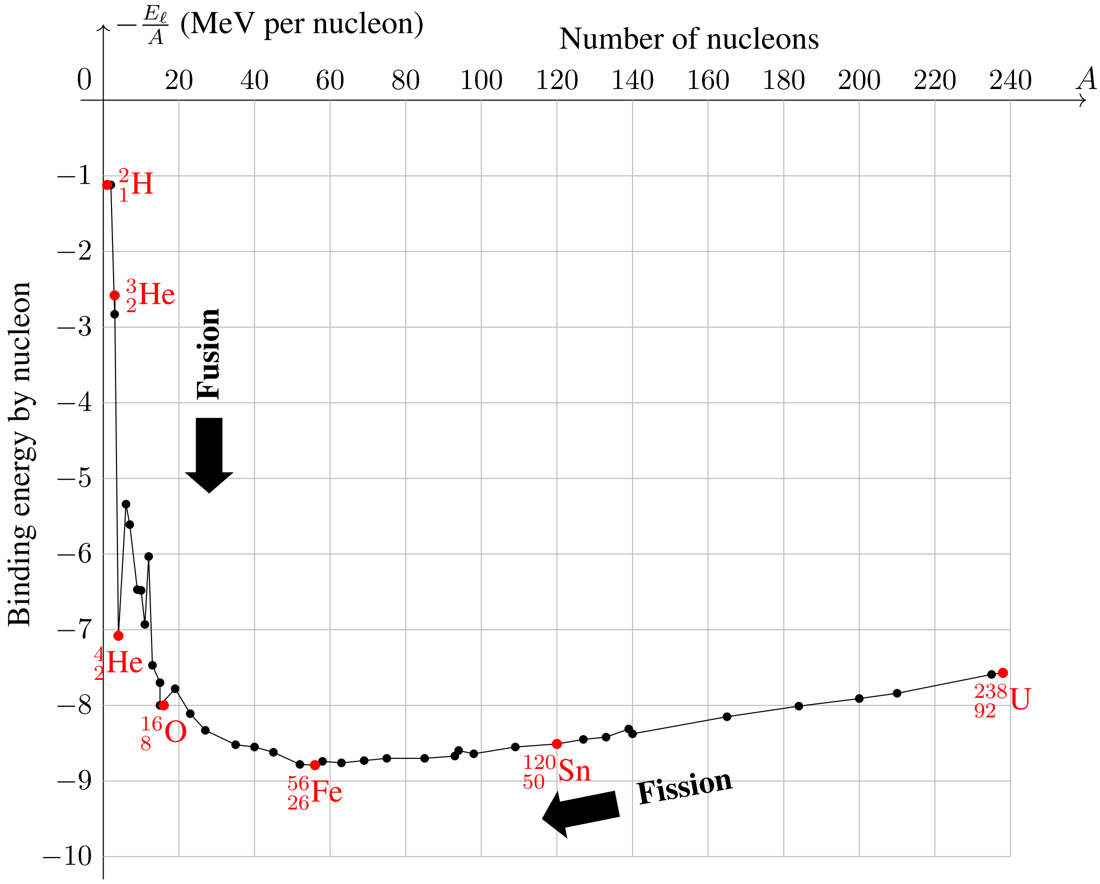


Figure 11.102: Aston curve

where we see that from the Iron the average the energy seems to decrease again (element which is therefore the considered by many undergraduate students as the most "glued" one and therefore the most stable in energy terms because having the highest average bonding energy). This decrease being due to the fact that from about 70 nucleons it seems that appear an electrostatic force inside the nucleus that begins to take over another force that rules in the very small nuclei (this force will be named later the "strong force" or "strong interaction").

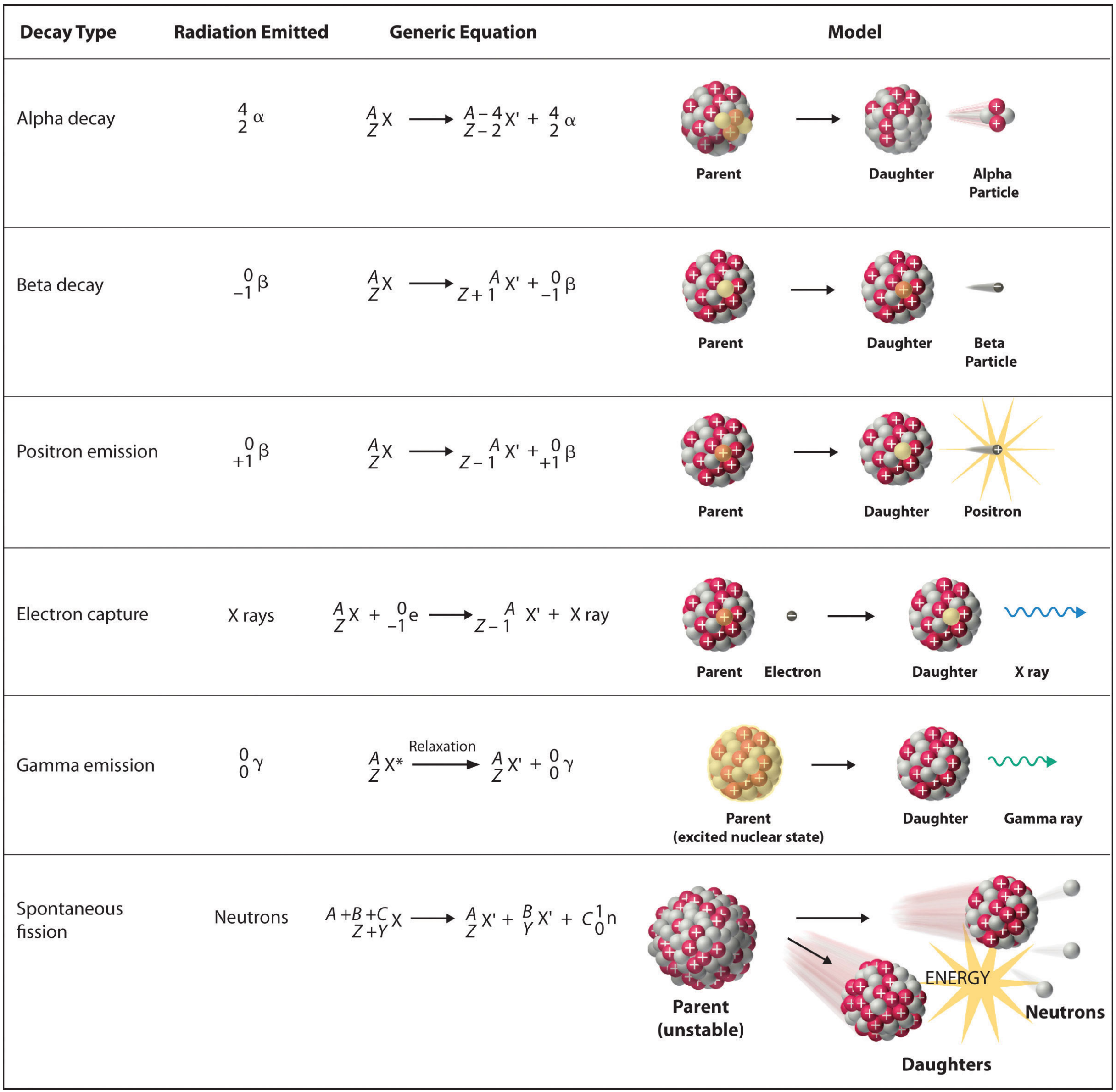
The Aston curve explains why atomic fusion in Stars stops at iron: it's just that the fusion of elements heavier than iron is endothermic rather than exothermic!

Remark

In reality the most stable isotope is Nickel 62 on a binding per energy basis. Iron 58 is second and Iron 56 is third as measured in [180]. Nickel 62 has the highest binding energy per nucleon of any known nuclide (8.7945 [MeV]). It is often stated that ^{56}Fe is the "most stable nucleus", but only because ^{56}Fe has the lowest mass per nucleon (not binding energy per nucleon) of all nuclides. The lower mass per nucleon of ^{56}Fe is possible because ^{56}Fe has $26/56 \approx 46.43\%$ protons, while ^{62}Ni has only $28/62 \approx 45.16\%$ protons; and the larger fraction of lighter protons in ^{56}Fe lowers its mean mass-per-nucleon ratio in a way that has no effect on its binding energy.

By the way, what is really important to notice in the chart above is that there is a inflection point and it is this one that makes it possible to obtain energy both with the nuclear fusion and with nuclear fission! We also see that the variation is much greater on the left than on the right, hence the fact that nuclear fusion releases much more energy.

Of all radioactivity phenomena, we can distinguish 17 of them (alpha decay, beta plus or beta minus decay, positron emission, electron capture, gamma emission, spontaneous fission, double proton emission, neutron emission, double neutron emission, cluster decay, double beta decay, double electron capture, double positron decay, isometric transition, internal conversion), some of them are named "secondary" because they are only the possible side effects of the first 6. Some of these phenomena are caused by man, others are natural and others are purely probabilistic:



Decay Type	Radiation Emitted	Generic Equation	Model
Alpha decay	$^{4}_{2}\alpha$	$^{A}_{Z}X \longrightarrow {}^{A-4}_{Z-2}X' + {}^{4}_{2}\alpha$	Parent → Daughter, Alpha Particle
Beta decay	$^{0}_{-1}\beta$	$^{A}_{Z}X \longrightarrow {}^{A}_{Z+1}X' + {}^{0}_{-1}\beta$	Parent → Daughter, Beta Particle
Positron emission	$^{0}_{+1}\beta$	$^{A}_{Z}X \longrightarrow {}^{A}_{Z-1}X' + {}^{0}_{+1}\beta$	Parent → Daughter, Positron
Electron capture	X rays	$^{A}_{Z}X + {}^{0}_{-1}e \longrightarrow {}^{A}_{Z-1}X' + \text{X ray}$	Parent, Electron → Daughter, X ray
Gamma emission	$^{0}_{0}\gamma$	$^{A}_{Z}X^{*} \xrightarrow{\text{Relaxation}} {}^{A}_{Z}X' + {}^{0}_{0}\gamma$	Parent (excited nuclear state) → Daughter, Gamma ray
Spontaneous fission	Neutrons	$^{A+B+C}_{Z+Y}X \longrightarrow {}^{A}_{Z}X' + {}^{B}_{Y}X' + C\,{}^{1}_{0}n$	Parent (unstable) → Daughters, ENERGY, Neutrons

Figure 11.103: **This table summarizes the type, nuclear equation, representation, and any changes in the mass or atomic numbers for various types of decay (source: `http://chemwiki.ucdavis.edu`)**

A much more exhaustive list of decay modes is given in NUBASE2020 (see [316]):

Mode	Name	Action	Nucleus changes
α	alpha emission	An alpha particle $(A=4, Z=2)$ emitted from nucleus	$(A-4, Z-2)$
p	proton emission	A proton ejected from nucleus	$(A-1, Z-1)$
$2p$	2-proton emission	Two protons ejected from nucleus simultaneously	$(A-2, Z-2)$
n	neutron emission	A neutron ejected from nucleus	$(A-1, Z)$
$2n$	2-neutron emission	Two neutrons ejected from nucleus simultaneously	$(A-2, Z)$

ε	electron capture	A nucleus captures an orbiting electron and emits a neutrino; the daughter nucleus is left in an excited unstable state	$(A,Z-1)$
e^+	positron emission	A nuclear proton converts to a neutron by emitting a positron and an electron neutrino	$(A,Z-1)$
β^+ $\varepsilon+\mathrm{e}^+$	positron emission	In NUBASE2020, β^+refers to the combined rate of electron capture (ε) and positron emission (e^+): $\beta^+=\varepsilon+e^+$	$(A,Z-1)$
β^-	β^-decay	A nucleus emits an electron and an electron antineutrino	$(A,Z+1)$
$\beta^-\beta^-$ $2\beta^-$	double β^-decay	A nucleus emits two electrons and two antineutrinos	$(A,Z+2)$
$\beta^+\beta^+$ $2\beta^+$	double β^+decay	A nucleus emits two positrons and two neutrinos	$(A,Z-2)$
β^-n	β^--delayed neutron emission	A nucleus decays by β^-emission to an excited state, which then emits a neutron	$(A-1,Z+1)$
β^-2n	β^--delayed 2-neutron emission	A nucleus decays by β^-emission to an excited state, which then emits two neutrons	$(A-2,Z+1)$
β^-3n	β^--delayed 3-neutron emission	A nucleus decays by β^-emission to an excited state, which then emits three neutrons	$(A-3,Z+1)$
β^+p	β^+-delayed proton emission	A nucleus decays by β^+emission to an excited state, which then emits a proton	$(A-1,Z-2)$
β^+2p	β^+-delayed 2-proton emission	A nucleus decays by β^+emission to an excited state, which then emits two protons	$(A-2,Z-3)$
β^+3p	β^+-delayed 3-proton emission	A nucleus decays by β^+emission to an excited state, which then emits three protons	$(A-3,Z-4)$
$\beta^-\alpha$	β^--delayed alpha emission	A nucleus decays by β^-emission to an excited state, which then emits an α particle	$(A-4,Z-1)$
$\beta^+\alpha$	β^+-delayed alpha emission	A nucleus decays by β^+emission to an excited state, which then emits an a particle	$(A-4,Z-3)$
β^-d	β^--delayed deuteron emission	A nucleus decays by β^-emission to an excited state, which then emits a deuteron	$(A-2,Z)$

$\beta^{-}t$	β^{-}-delayed triton emission	A nucleus decays by β^{-}emission to an excited state, which then emits a triton	$(A-3, Z)$
CD	cluster decay	A nucleus emits a specific type of smaller nucleus (A_1, Z_1) which is larger than an alpha particle (e.g. $^{14}\mathrm{C}, {}^{24}\mathrm{Ne}$)	$(A-A_1, Z-Z_1)$ & (A_1, Z_1)
IT	internal (isomeric) transition	A nucleus in a metastable state drops to a lower energy state by emitting a photon or ejecting an electron	(A, Z)
SF	spontaneous fission	A nucleus disintegrates into two or more smaller nuclei and other particles, all of which may vary with each decay	variable
β^{+}SF	β^{+}-delayed fission	A nucleus decays by β^{+}emission to an excited state, which then undergoes spontaneous fission	$\beta+$ & variable
β^{-}SF	β^{-}-delayed fission	A nucleus decays by β^{-}emission to an excited state, which then undergoes spontaneous fission	β^{-}& variable

Below is a diagram representing at the top the "valley of stability" of atoms and isotopes and down the same valley but highlighting the type of disintegration:

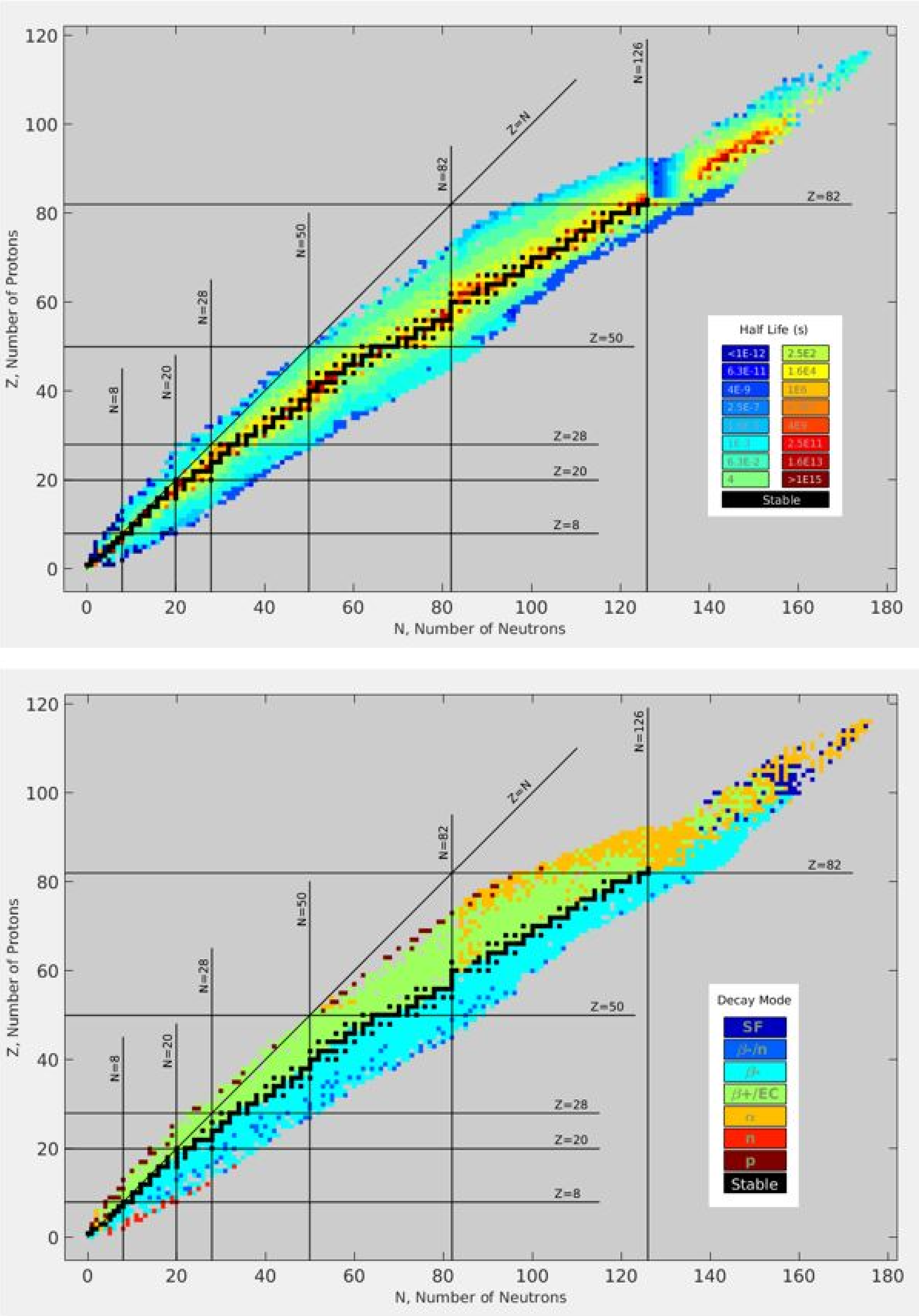


Figure 11.104: Nuclear stability valley (source: Wikipedia)

As the reader can notice it, we have experimental evidence that many elements with equal number of protons and neutrons ($Z=N$) are radioactive.

Let us see then the types of disintegration or modifications of the structure of the atom / nucleus which are possible in this details:

11.4.2.4.1 Nuclear Fusion (1)

If we assemble two light nuclei ${}^{A_x}_{Z_X}\text{X}$ and ${}^{A_Y}_{Z_Y}\text{Y}$ (with $A_x \leq A_y$) to form a "heavy" atom ${}^{A_Z}_{Z_Z}\text{Z}$, then in accordance with the left part of the Aston curve seen above, we increase the mass defect since the average energy per nucleon increases. Indeed:

- The energy of X is equal to:

$$A_X \cdot \bar{B}(A_X, Z_X) = A_X \bar{B}_X \tag{11.4.120}$$

- The energy of Y is equal to:

$$A_Y \cdot \bar{B}(A_Y, Z_Y) = A_Y \bar{B}_Y \tag{11.4.121}$$

- Then energy of Z is equal to:

$$A_Z \cdot \bar{B}(A_Z, Z_Z) = A_Z \bar{B}_Z \tag{11.4.122}$$

As:

$$\bar{B}_X \leq \bar{B}_Y \leq \bar{B}_Z \tag{11.4.123}$$

then:

$$\boxed{\Delta E = A_Z \bar{B}_Z - (A_X \bar{B}_X + A_Y \bar{B}_Y)} \tag{11.4.124}$$

is strictly positive.

Nuclear fusion is almost exclusively caused by man (on Earth at least... because the stars do it alone with the help of gravity). The probability of observing a natural nuclear fusion under normal conditions of temperature and pressure is so low that it is unnecessary to speak of it at least to this day...

The three most common method for achieving fusion are so far:

- Thermonuclear fusion (heated gas under the form of a plasma in analogy with what happens in Stars)
- Inertial confinement fusion (compression of a fuel target with LASER)
- Inertial electrostatic confinement (use of electric field to heat ions to fusion conditions)

For more information about plasma the reader can reefer to the corresponding section in this textbook.

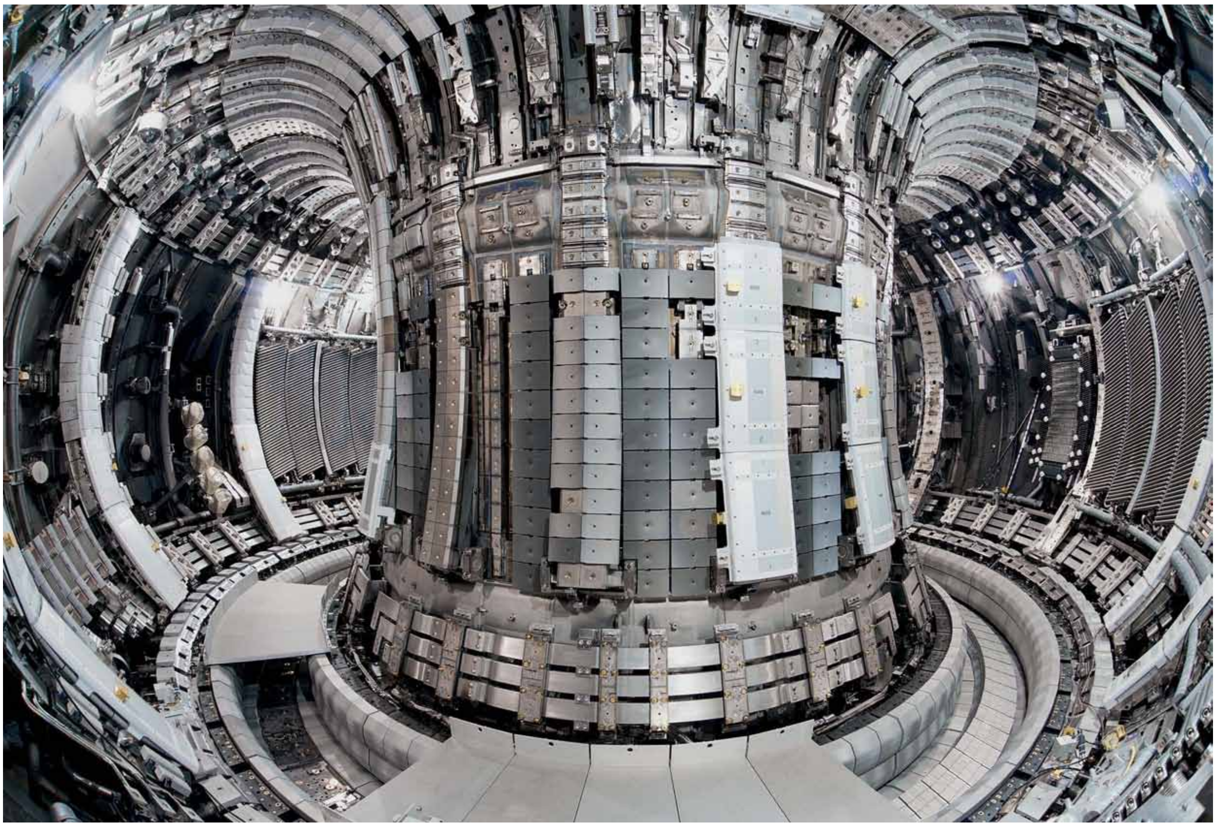
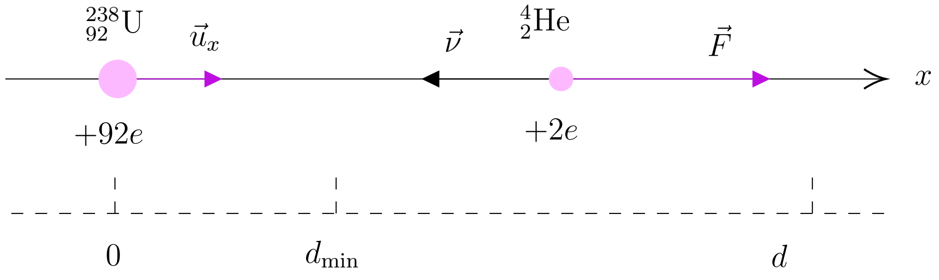


Figure 11.105: Joint European Torus (JET) tokamak fusion detector (source: EUROfusion)

Let us consider now a typical important calculation in the field of nuclear fusion: A helium nucleus ${}^{4}_{2}\mathrm{H}^{2+}_{2}$ is sent with an incident kinetic energy of $E^i_c = 30$ [MeV] towards an uranium nucleus ${}^{238}_{92}\mathrm{U}$ at rest. What is the minimum approach distance between the two cores knowing that the initial distance between the two particles is equal to $d = 10$ [cm].

The two nuclei being positively charged, their electrostatic repulsion along an axis x passing through the two nuclei, is all the stronger as the two nuclei approach each other (see figure below). The minimum approach distance between the two nuclei (d_{min}) is the one for which the speed of the helium nucleus cancels out to change direction. By using the kinetic energy theorem applied between the initial moment when the distance separating the nuclei is d and the moment when this distance is minimal we obtain:



$$\Delta E_c = W_{(F)} \Rightarrow 0 - E_c^i = \int\limits_{d}^{d_{\min}} \vec{F} \circ \mathrm{d}x\vec{u}_x \tag{11.4.125}$$

But the electric force is given by:

$$\vec{F} = \frac{92q_e \cdot 2q_e}{4\pi\varepsilon_0 \cdot x^2}\vec{u}_x \tag{11.4.126}$$

and by the integral:

$$\int\limits_{d}^{d_{\min}} \vec{F} \circ \mathrm{d}x\vec{u}_x = \int\limits_{d}^{d_{\min}} F \cdot \mathrm{d}x = \frac{184q_e^2}{4\pi\varepsilon_0}\left[\frac{-1}{x}\right]_d^{d_{\min}} = \frac{184q_e^2}{4\pi\varepsilon_0}\left(\frac{1}{d} - \frac{1}{d_{\min}}\right) = -E_c^i \tag{11.4.127}$$

The minimum approach distance is therefore:

$$\frac{1}{d_{\min}} = \frac{1}{d} + \frac{4\pi\varepsilon_0 E_c^i}{184q_e^2} \tag{11.4.128}$$

The kinetic energy is given in [Mev] (with as a reminder $1[\text{Mev}] = q_e \cdot 10^6$ [J]). We can therefore write:

$$\frac{4\pi\varepsilon_0 E_c^i}{184q_e^2} = \frac{1}{9.10^9}\frac{30 \cdot 10^6 q_e}{184 \cdot q_e^2} = \frac{10^{17}}{184 \cdot 4.8} = \frac{1}{0.88 \cdot 10^{-14}}\,[\text{m}^{-1}] \tag{11.4.129}$$

By comparing:

$$\frac{1}{d} = \frac{1}{0.1} = 10\,[\text{m}^{-1}] \tag{11.4.130}$$

we see that $1/d$ is negligible before the other term. We then have:

$$\frac{1}{d_{\min}} \cong \frac{1}{0.88 \cdot 10^{-14}} \Rightarrow d_{\min} \cong 0.88 \cdot 10^{-14}\,[\text{m}] = 8.8\,[\text{fermis}] \tag{11.4.131}$$

The most primitive fusion reaction provides the primary source of energy generation on the sun. It is given by:

$${}_1^1\text{H} + {}_1^1\text{H} \rightarrow {}_1\text{H}^2 + e^+ + v_e \quad (Q = 1.44\,[\text{MeV}]) \tag{11.4.132}$$

Here ${}_1^1\text{H}$ is a proton and ${}_1^2\text{H}$ is a deuteron, e^+ is a positron, and v_e is an electron-neutrino. Because of the repulsive force on the protons caused by their positive charge, the reaction rate for this process is very small at low temperatures. Furthermore, the release of a positron and a neutrino shows that this reaction is a weak interaction, causing a reduction in reaction probability over interactions with a reaction using the strong force.

For this reaction to take place, one might think that the particle energies mus be high enough for them to physically touch, treating the particles as billiard ball like objects. A simple calculation shows that this requires very high energies. The potential energy generated by moving the two protons together, each having a radius of about 1.2 [fm] is given by:

$$E_{\textbf{Coul.}} = \frac{q_e^2}{4\pi\varepsilon_0\,(r_p + r_p)} = \frac{1.44\,[\text{MeV} \cdot \text{fm}^{-1}]}{2.4\,[\text{fm}]} = 600\,[\text{keV}] \tag{11.4.133}$$

Since the Sun's core temperature (in plasma physics, always written in energy units, i.e. multiplied by Boltzmann's constant kT and then suppressing the k) is 1.335 [keV] (about 15.5 million Kelvin) we can see that the average energy per particle $(3/2)T$ is much less than this potential energy quoted above.

Let's see another approach of the problem! In order to fuse two nuclei it is first necessary to surmount the potential barrier (or Coulomb barrier) set by the electrostatic repulsion. The height of the Coulomb barrier between two nuclei can be estimated using the equation:

$$E_{\textbf{Coul.}}=\frac{Z_1Z_2q_e^2}{4\pi\varepsilon r} \tag{11.4.134}$$

where Z_1 and Z_2 are the number of protons in each of the two nuclei. Setting $r=10^{-15}\,[\text{m}]$, $Z_1=Z_2=1$, the charge on the electron $q_e\cong 1.60\cdot 10^{-19}\,[\text{C}]$ and the permittivity of free space $\varepsilon_0=8.85\cdot 10^{-12}\,[\text{C}^2\cdot\text{N}^{-1}\cdot\text{m}^{-2}]$, we get a potential energy of $2.3\cdot 10^{-13}\,[\text{J}]$ or $1.4\,[\text{MeV}]$ for the height of the Coulomb barrier, i.e.:

$$E_{\textbf{Coul.}}=\frac{Z_1Z_2q_e^2}{4\pi\varepsilon r}=\frac{1.44Z_1Z_2}{r}\cong 2.3\cdot 10^{-13}\,[\text{J}] \tag{11.4.135}$$

Now, let's compare the average kinetic energy of a Hydrogen nucleus in the center of a Star like the Sun. For $T_c\approx 1.5\cdot 10^7\,[\text{K}]$ we have:

$$E_{\text{th}}=kT_c\cong 1.29\cdot 10^3\,[\text{eV}]=1.29\,[\text{keV}] \tag{11.4.136}$$

The relative velocity v of pairs of particles follow a Maxwell speed distribution (see section Statistical Mechanics page 3554) :

$$\mathrm{d}P(E)=\frac{2}{\sqrt{\pi}}\frac{E^{1/2}}{(kT)^{3/2}}e^{-E/kT}\mathrm{d}E \tag{11.4.137}$$

Therefore (we do a numerical integration as no closed form is known):

$$P=\int\limits_{2.3\cdot 10^{-13}}^{+\infty}\frac{2}{\sqrt{\pi}}\frac{E^{1/2}}{(kT)^{3/2}}e^{-E/kT}\mathrm{d}E\cong 10^{-70} \tag{11.4.138}$$

Therefore the probability of a pair of particles having enough kinetic energy to overcome the Coulomb barrier is $P\cong 10^{-70}$. The total number of nuclei in the Sun (actually a G2V-type star, aka yellow dwarf) is approximately given by its mass divided by the mass of a hydrogen atom, i.e. $1.99\cdot 10^{30}/1.67\cdot 10^{-27}\cong 10^{57}$. Even with $\sim 10^{57}$ nucleons, there is almost no chance of fusion! In order to overcome the Coulomb barrier, therefore, the temperature would have to be of the order of 10^{10} [K]. This is nearly four orders of magnitude higher than the minimum mean temperature of the Sun! Hence we need to consider quantum mechanical tunnelling through Coulomb barrier!

However, quantum mechanics tells us that the protons are not discrete particles but rather must be regarded as having a wave function and have a probability distribution in space that includes places where they would not physically be allowed to exist under the laws of Classical Mechanics. Let us see a first rough approach:

For tunneling, p-p separation should be comparable to de Broglie (dB) wavelength:

$$\lambda_{\text{dB}}=\frac{h}{p}=\frac{h}{\sqrt{2m_pE}} \tag{11.4.139}$$

As a second condition for tunneling, the kinetic energy of the proton must be comparable to the electrostatic potential U at that separation:

$$U_{\text{Coulomb}}=\frac{q_e^2}{4\pi\varepsilon_0}\frac{1}{\lambda_{\text{dB}}}=\frac{e^2}{4\pi\varepsilon_0}\frac{(2m_pE)^{1/2}}{h}\cong E \qquad (11.4.140)$$

Solving for E gives the minimum kinetic energy for tunneling in this rough approach:

$$E\cong\left(\frac{q_e^2}{4\pi\varepsilon_0}\right)^2\frac{2m_p}{h^2}\cong 3\,[\text{keV}] \qquad (11.4.141)$$

Which is comparable to the average kinetic energy in the Sun's core of $2\,[\text{keV}]$.

A much more rigorous calculation of this process was done by Gamow and this adds a "tunnelling factor" to the cross section, given by:

11.4.2.4.2 Nuclear Fission (2)

Similarly (for the theoretical treatment of fission based on quantum physics see the Wave Quantum Physics section at page 4840), if we break with adequate tools (often with neutrons because to approach the nucleus and overcome its electrostatic repulsion it is the appropriate way... and this is the method used by nuclear power plant and nuclear bombs) an heavy atom ${}^{A_X}_{Z_X}\text{X}$ in two light atoms ${}^{A_Y}_{Z_Y}\text{Y}$ and ${}^{A_Z}_{Z_Z}\text{Z}$ (with $A_Y\leq A_Z$) we also increase the mass defect and the energy gained is therefore equal to:

$$\boxed{\Delta E=(A_Y\bar{B}_Y-A_Z\bar{B}_Z)-A_X\bar{B}_X} \qquad (11.4.142)$$

That it is in the case of fission or fusion, the energy released is then divided into the kinetic energy of the fission products, the neutrons emitted and finally the various radiations.

The first reported nuclear fission occurred in 11939 (holocene calendar) when three German scientists, Lise Meitner, Otto Hahn, and Fritz Strassman, bombarded Uranium 235 atoms with slow-moving neutrons that split the Uranium 238 nuclei into smaller fragments that consisted of several neutrons and elements near the middle of the periodic table. Since then, fission has been observed in many other isotopes, including most actinide isotopes that have an odd number of neutrons. A typical nuclear fission reaction (one of the three of ${}^{235}_{92}\text{U}$) is shown below:

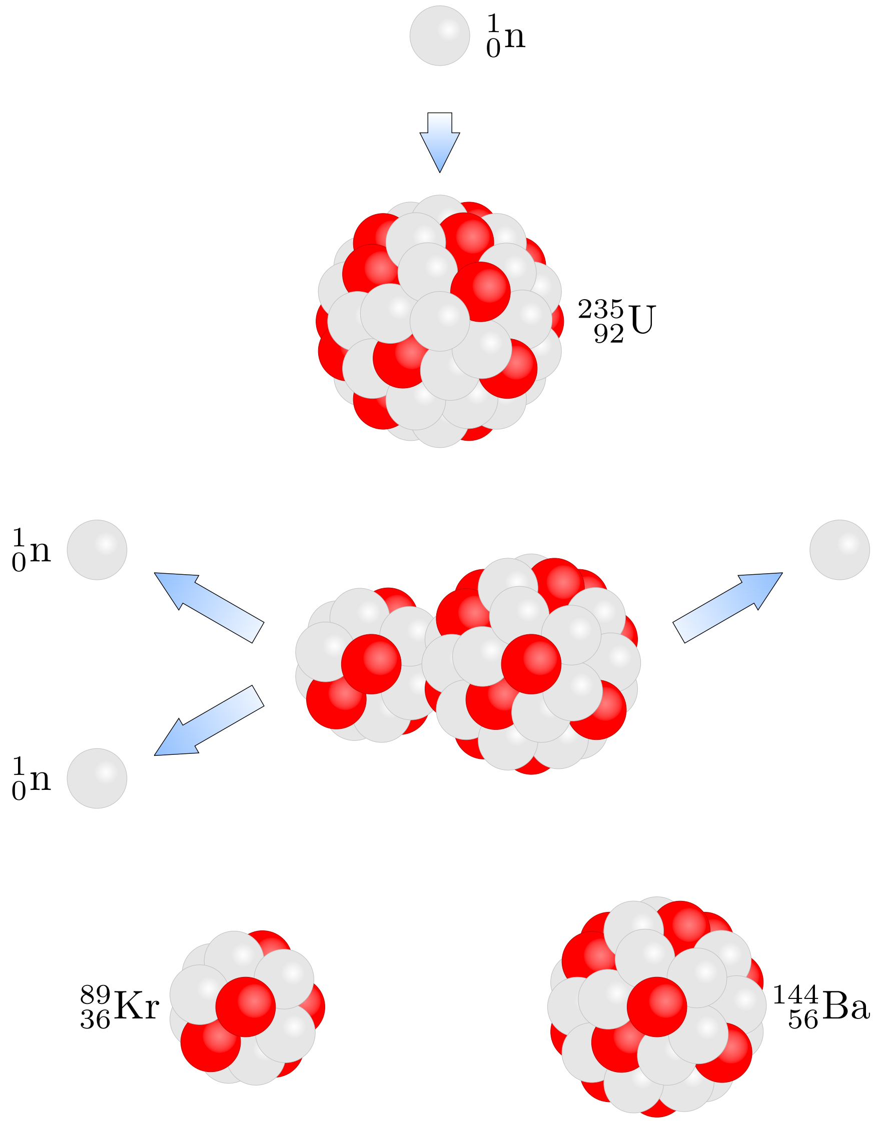


Figure 11.106: **Nuclear fission by a slow neutron of Uranium 235 with its fission products**

Remark

An atom is said to be "fissible" (or sometimes "fissionable") when fast neutrons are needed to produce fission and "fissile" when it is enough to have slow neutrons for fission (which is more rare).

Or another $^{235}_{92}$U fission reaction as seen when a chain reaction occurs:

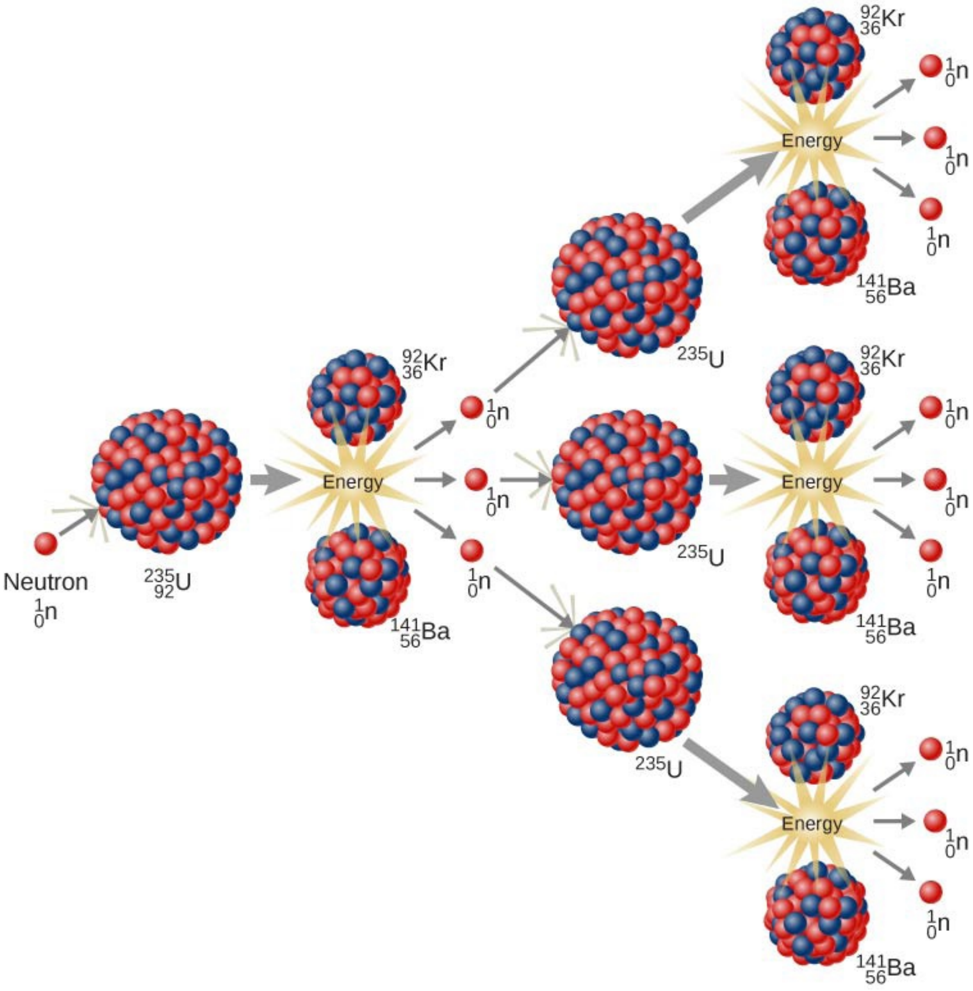


Figure 11.107: **One of Uranium 235 fission chain reaction (source: OpenStax)**

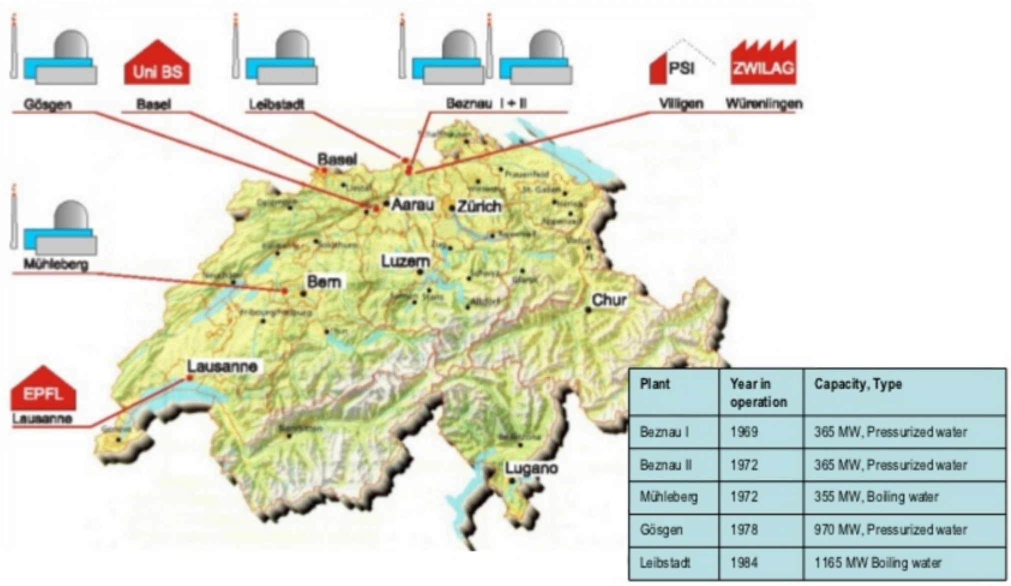
Explicitly:

$$^{235}_{92}\mathrm{U}+{}^1_0n\rightarrow 3^1_0n+{}^{92}_{36}\mathrm{Kr}+{}^{141}_{56}\mathrm{Ba}+170\,[\mathrm{MeV}] \tag{11.4.143}$$

Nuclear energy is by far one of the most concentrated form of energy, since 1 kilogram of natural Uranium supplies a heat quantity of $100,000$ [kWh] in a current power plant, while 1 kilogram of coal supplies by burning only 8 [kWh]. For this reason, only a relatively small amount of nuclear fuel is used for electricity production: a nuclear power plant with an electrical capacity of $1,000$ [MW] consumes 27 tons of enriched Uranium annually, one quarter of its load, whereas a thermal power plant of the same power consumes $1,500,000$ tons of oil by year... For comparison in the Sun, 1 kilogram of hydrogen produced, by nuclear reactions transforming it into helium, 180 million kWh!!!! But beware, industrially we know how to extract only a small part of the nuclear energy stored in the material. Of the 27 tons of enriched Uranium consumed in one year by a power plant, only a small amount of nucleus was really consumed (hence the economic need to reprocess the Uranium after use).

We quickly realize that the calorific value of fission is gigantic compared to that of fossil energies. An estimate gives a ratio of energy released per atom of $50,000$ million !!! On the other hand, the risk ratio in terms of exploitation and management of waste and impact on the environment is of the same order but in the opposite direction. That mankind finds either another way of producing electricity, or changes its habits of consumption.

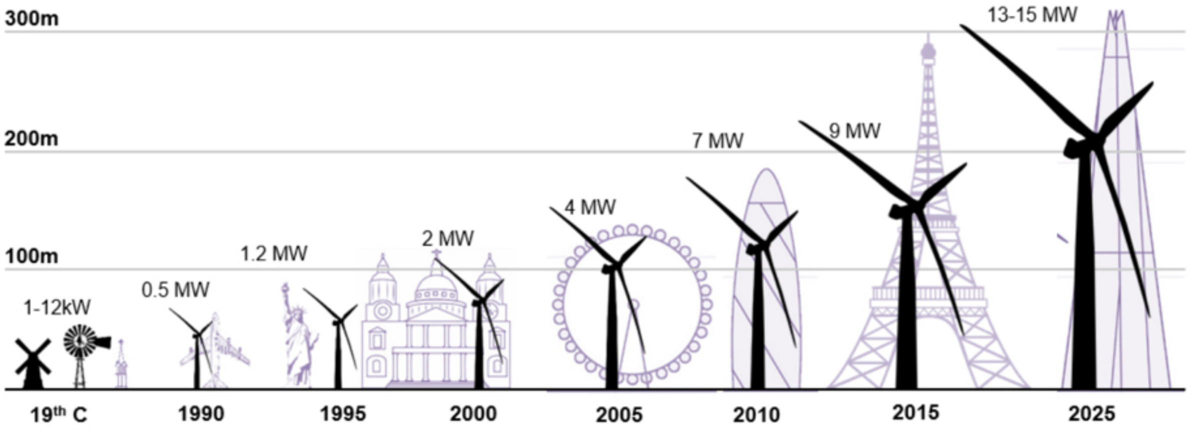
For information in Switzerland, there are only 5 at this nuclear power plants (at the beginning of the 121st century according to the holocene calendar) for a population of almost 7 million individuals in year 12000 (holocene calendar):


Gösgen
Uni BS
Basel
Leibstadt
Beznau I + II
PSI
Villigen
ZWILAG
Würenlingen
Basel
Aarau
Zürich
Mühleberg
Luzern
Bern
Chur
EPFL
Lausanne
Lausanne
Lugano

Plant	Year in operation	Capacity, Type
Beznau I	1969	365 MW, Pressurized water
Beznau II	1972	365 MW, Pressurized water
Mühleberg	1972	355 MW, Boiling water
Gösgen	1978	970 MW, Pressurized water
Leibstadt	1984	1165 MW Boiling water

Figure 11.108: **Nuclear power plants in Switzerland at the end of the 120th century (source: Swiss confederation)**

It may be interesting for the reader to compare the power capacity of the swiss nuclear plants to the 121st century (holocene calendar) state of the art windmills:



Sources: Various; Bloomberg New Energy Finance

This little introduction being made, let us return to the purely mathematical part!

In the case of spontaneous (or natural) fission we have the emission of two fission products and of w neutrons. What we denote by:

$$ {}_{Z}^{A}\mathrm{X} \rightarrow {}_{\nu}^{M}\mathrm{Y} + {}_{Z-\nu}^{A-M-w}\mathrm{Z} + w\,{}_{0}^{1}n \qquad (11.4.144)$$

✍**Example:**

The isotope of Carbon 15 by spontaneous fission gives an isotope of Nitrogen with the emission of an electron and an antineutrino:

$$ {}^{15}_{8}\mathrm{C} \rightarrow {}^{15}_{7}\mathrm{N} + \beta^{-} + \bar{\nu} \tag{11.4.145} $$

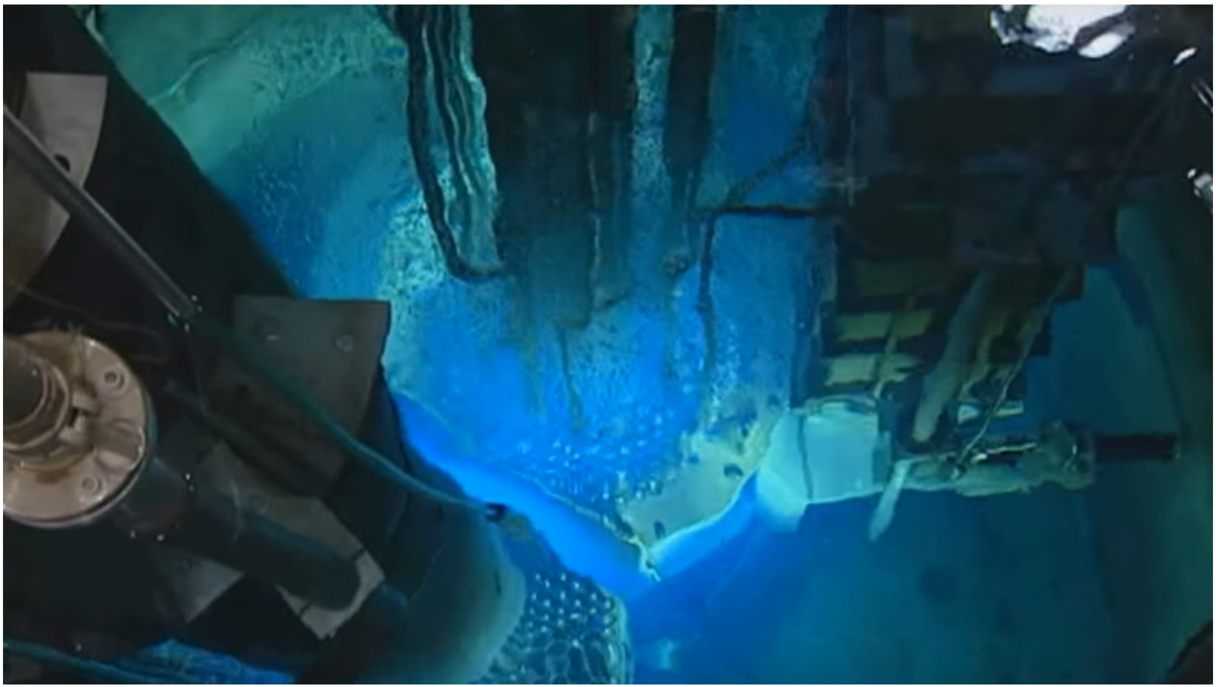


Figure 11.109: **Sandia National Laboratories' Annular Core Research Reactor (we can see the bubbles of the evaporated water on the picture)**

At the day we write these lines there are more than 130 fission reactor concepts (the most promising one and most safety one seems to be the MSR - Molten Salt Reactor - however governments like to stay with water pressured reactors for historical, economical and educational reasons that are debatable... Inertia is sadly hard to change). The reader can see below only a few of them in two different figures (we were not able to find who are the artists):

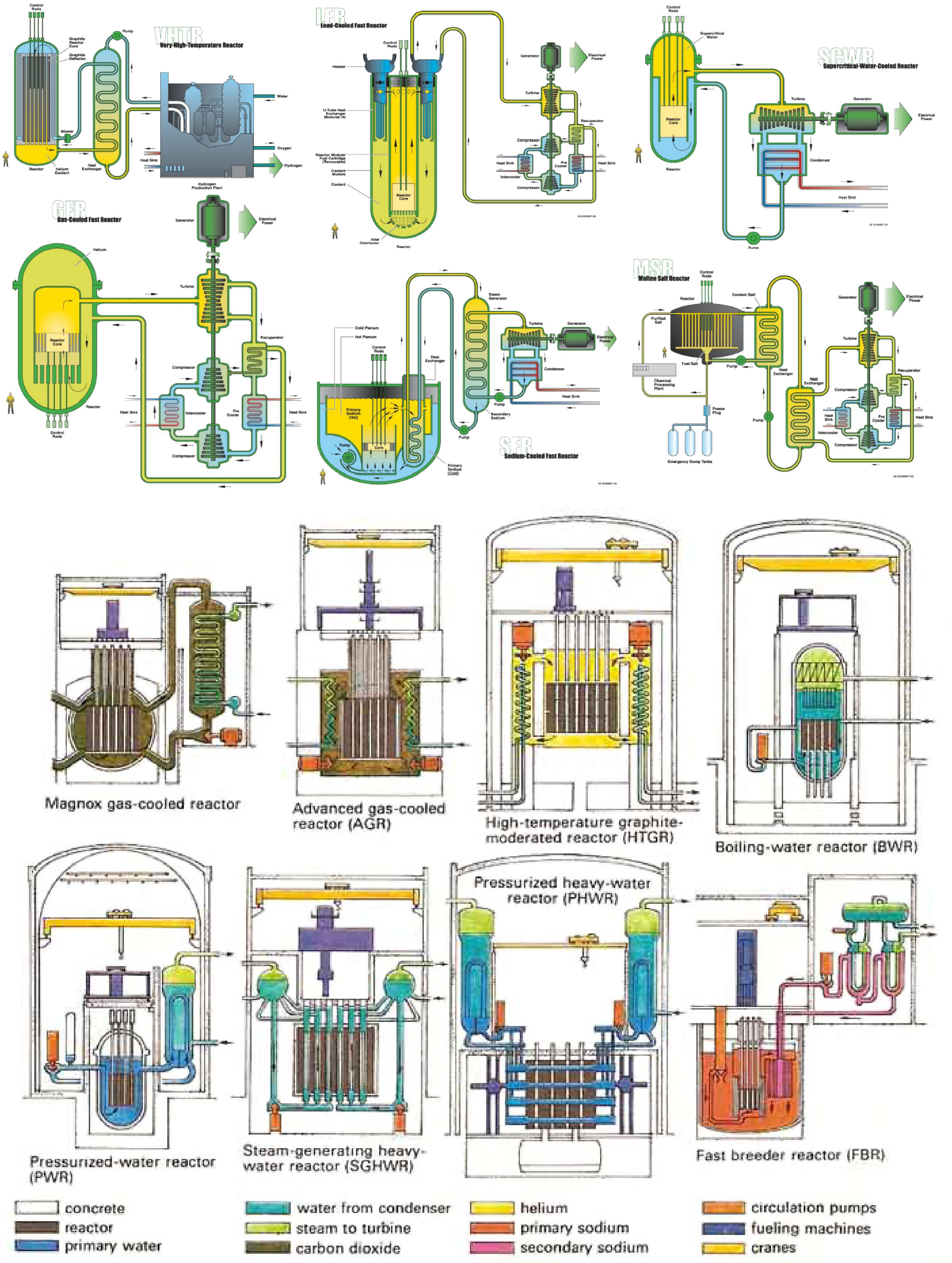

VHTR
Very-High-Temperature Reactor
LFR
Lead-Cooled Fast Reactor
SCWR
Supercritical-Water-Cooled Reactor
GFR
Gas-Cooled Fast Reactor
SFR
Sodium-Cooled Fast Reactor
MSR
Molten Salt Reactor
Magnox gas-cooled reactor
Advanced gas-cooled reactor (AGR)
High-temperature graphite-moderated reactor (HTGR)
Boiling-water reactor (BWR)
Pressurized-water reactor (PWR)
Steam-generating heavy-water reactor (SGHWR)
Pressurized heavy-water reactor (PHWR)
Fast breeder reactor (FBR)
concrete
reactor
primary water
water from condenser
steam to turbine
carbon dioxide
helium
primary sodium
secondary sodium
circulation pumps
fueling machines
cranes

Figure 11.110: Types of nuclear reactors (author: ?)

Now it's quite important to notice something!:

The fission of one atom of ${}^{239}\mathrm{Pu}$ generates $207.1\,[\mathrm{MeV}] = 3.318 \cdot 10^{-11}\,[\mathrm{J}]$, according to the following reaction:

$${}_{94}\mathrm{Pu}^{239} + {}_{0}\mathrm{n}^{1} \longrightarrow {}_{94}\mathrm{Pu}^{240} \longrightarrow {}_{40}\mathrm{Zr}^{103} + {}_{54}\mathrm{Xe}^{134} + 3{}_{0}\mathrm{n}^{1} + 207.1\,\mathrm{MeV} \tag{11.4.146}$$

i.e. $19.98\,[\text{TJ}\cdot\text{mol}^{-1}] = 83.61\,[\text{TJ}\cdot\text{kg}^{-1}]$ or about $23\,[\text{GW}\cdot\text{h}\cdot\text{kg}^{-1}]$ of gamma rays.

Now let us consider the following hydrogen fusion chain:

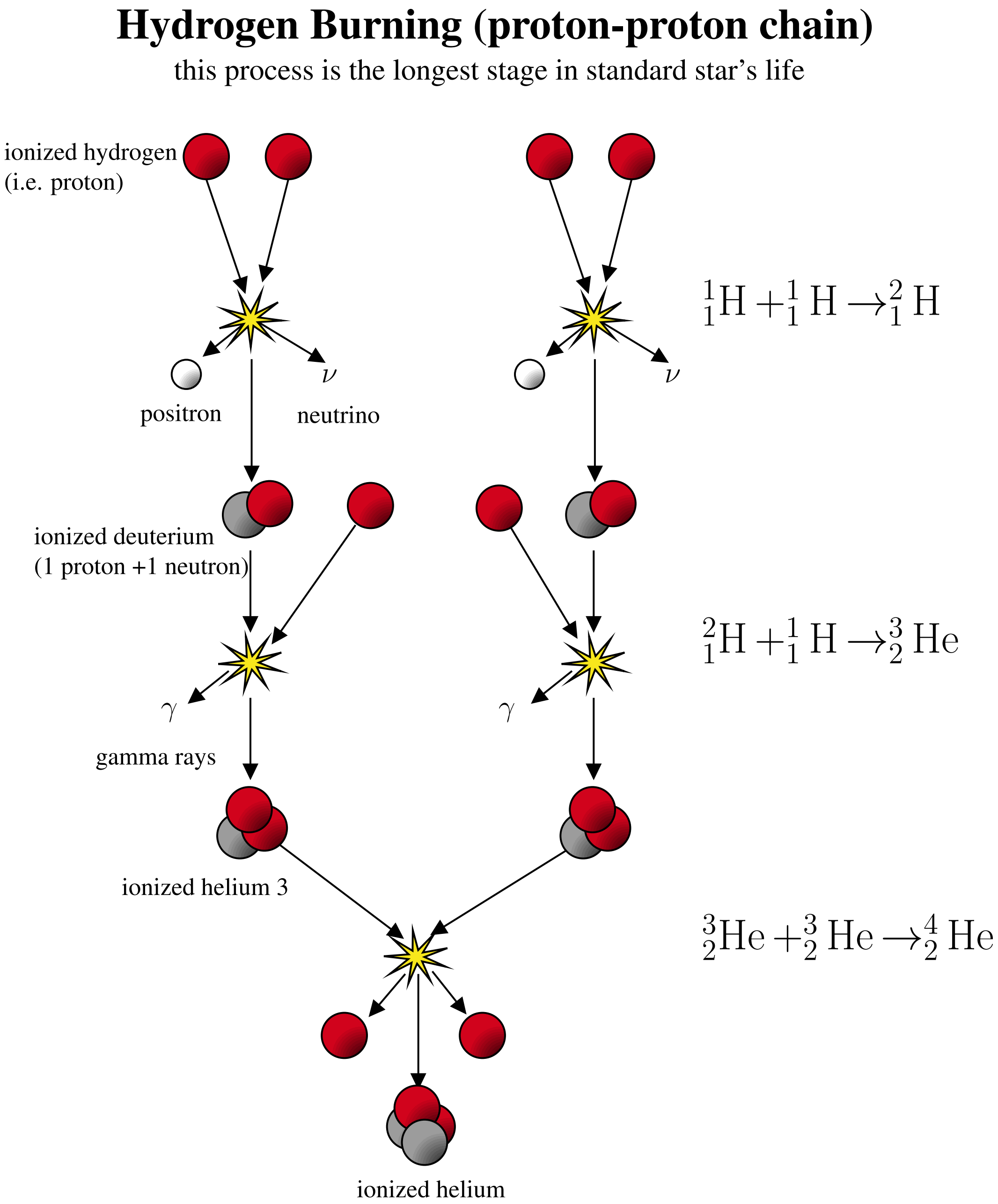

Hydrogen Burning (proton-proton chain)
this process is the longest stage in standard star's life
ionized hydrogen
(i.e. proton)
$^1_1\text{H} + ^1_1\text{H} \rightarrow ^2_1\text{H}$
ν
ν
positron
neutrino
ionized deuterium
(1 proton +1 neutron)
$^2_1\text{H} + ^1_1\text{H} \rightarrow ^3_2\text{He}$
γ
γ
gamma rays
ionized helium 3
$^3_2\text{He} + ^3_2\text{He} \rightarrow ^4_2\text{He}$
ionized helium

Figure 11.111: **Hydrogen fusion (proton-proton chain)**

In the above chain we see that four hydrogen nuclei are involved having a total mass of $6.693\cdot10^{-27}\,[\text{kg}]$. They fuse into a helium nucleus of mass $6.645\cdot10^{-27}\,[\text{kg}]$. Therefore the mass lost is $0.048\cdot10^{-27}\,[\text{kg}]$. Thus, only the fraction $0.048/6.693 = 0.00717$ of the original hydrogen mass was converted to radiation and kinetic energy. In other words, when we start with $1\,[\text{kg}]$ of hydrogen, a mass of $0.00717\,[\text{kg}]$ will be converted to energy radiation and kinetic energy during the fusion. This gives an energy output per kilogram:

$$E = mc^2 = 0.00717\cdot c^2 \cong 6.45\cdot10^{14}\,[\text{J}] \qquad (11.4.147)$$

Hence $6.45\cdot 10^2\,[\text{TJ}\cdot\text{kg}^{-1}]$. So for the same mass, Hydrogen fusion in approximation release 10 times more energy than plutonium fission! We can understand why the army is interested in nuclear H bomb (even if the fusion of hydrogen had to be initiated with a plutonium/uranium like implosion) and why the industry is interest to develop fusion nuclear power plants rather than fission one (even if the technical obstacles are huge to master fusion and still not usable in an industrial way in the early 121st century).

11.4.2.4.3 Alpha Disintegration (3)

Definition (#11.4.514): When a heavy nucleus contains too many protons and neutrons (like Uranium 238 for example), it will empty its nucleon overflow by emitting an alpha particle (Helium nucleus composed of 2 protons and 2 neutrons) and the final system which will be a new nucleus will have a smaller and possibly stable mass. This mode of disintegration is named "alpha radioactivity".

The probability of disintegration is governed by the mechanism of penetration barrier ("Tunnel effect") as we will prove it a little further after a small introduction.

The radioactive decay according to the alpha radioactivity can be schematized as (with some small variations depending on the case):

$$ {}^{A}_{Z}\mathrm{X}_{N} \rightarrow {}^{A-4}_{Z-2}\mathrm{X}_{N-2} + {}^{4}_{2}\alpha \tag{11.4.148} $$

where:

$$ {}^{4}_{2}\alpha = {}^{4}_{2}\mathrm{He}_{2}^{++} \tag{11.4.149} $$

✍**Example:**

Radon is found in significant concentrations in underground mines and sometimes in some caves (in some mining or granitic areas):

$$ {}^{222}_{86}\mathrm{Ra} \rightarrow {}^{218}_{84}\mathrm{Po} + \alpha + \gamma \tag{11.4.150} $$

where Polonium is one of the main factors of induction of so-named "radon-induced lung cancers".

The energy released during the transmutation is calculated thanks to the mass default:

$$ \mathrm{Q}_{\alpha} = (M_X - M_y - M_{\mathrm{He}})\cdot c^2 \tag{11.4.151} $$

with M_X being the mass of the initial nucleus, M_Y the mass of the final nucleus and M_{He} the mass of the Helium nucleus.

By neglecting the binding energy of electrons, we have:

$$ M_X = m_X + Zm_e \quad \text{and} \quad M_Y = m_Y + (Z-2)m_e \quad \text{and} \quad m_{\mathrm{He}} = m_{\mathrm{He}} + 2m_e \tag{11.4.152} $$

Finally:

$$ \boxed{\mathrm{Q}_{\alpha} \cong (m_X - m_y - m_{\mathrm{He}})\cdot c^2} \tag{11.4.153} $$

This expression shows that the energy of α particle is well defined for given initial and final nuclei. In fact, we actually observe a discrete energy spectrum. We conclude that these emissions lead the nucleus to intermediate energy levels corresponding to excited states of the final nucleus. We can explain these observations by a layered nuclear structure!!! The de-excitation of the final is done by emission of γ photons.

The conservation of energy requires that the energy of the α disintegration is distributed between the kinetic energy of the two residual products:

$$\mathrm{Q}_\alpha = E_c^{\text{nucleus}} + E_c^\alpha = \frac{1}{2}m_{\text{nucleus}}v_{\text{nucleus}}^2 + \frac{1}{2}m_\alpha v_\alpha^2 \tag{11.4.154}$$

The conservation of the linear momentum gives us:

$$p_{\text{nucleus}} = m_{\text{nucleus}} \cdot v_{\text{nucleus}} = p_\alpha = m_\alpha v_\alpha \tag{11.4.155}$$

and therefore:

$$v_{\text{nucleus}} = \frac{m_\alpha}{m_{\text{nucleus}}} v_\alpha \tag{11.4.156}$$

which we replace in the energy conservation equation:

$$\begin{aligned}\mathrm{Q}_\alpha &= \frac{1}{2}m_{\text{nucleus}}\left(\frac{m_\alpha}{m_{\text{nucleus}}}v_\alpha\right)^2 + \frac{1}{2}m_\alpha v_\alpha^2 = \frac{m_\alpha}{m_{\text{nucleus}}}\left(\frac{1}{2}m_\alpha v_\alpha^2\right)^2 + \frac{1}{2}m_\alpha \nu_\alpha^2\\ &= \frac{1}{2}m_\alpha v_\alpha^2\left(1+\frac{m_\alpha}{m_{\text{nucleus}}}\right) = E_c^\alpha\left(1+\frac{m_\alpha}{m_{\text{nucleus}}}\right)\end{aligned} \tag{11.4.157}$$

and we derive that the energy of the α particle is equal to:

$$E_c^\alpha = \frac{\mathrm{Q}_\alpha}{\left(1+\dfrac{m_\alpha}{m_{\text{nucleus}}}\right)} = \frac{\mathrm{Q}_\alpha}{\left(\dfrac{m_{\text{nucleus}}+m_\alpha}{m_{\text{nucleus}}}\right)} \cong \frac{\mathrm{A}}{\mathrm{A}+4}\mathrm{Q}_\alpha \tag{11.4.158}$$

since the masses of the nucleus and the α particle are approximately proportional to their mass number, thus A and 4 respectively.

Let us see now the details of the mechanism of α disintegration with an academic approach, simplified to the extreme and therefore approximative (but sufficient nevertheless). For this approach, we will use the developments on the tunnel effect that we have carried out in the section of Wave Quantum Physics.

For nuclei having a nucleon number becoming too large, the Coulomb repulsion between protons takes significant values in comparison to the strong interaction which ensures the cohesion of the nucleus. We then observe a phenomenon of saturation, which gives rise to the α disintegration which is a special case of a spontaneous fission.

George Gamow proposed a theoretical explanation for this phenomenon in 11928 (holocene calendar). He assumes that the particle α pre-exists in the nucleus and bumps on the potential walls. It then has a non-zero probability of crossing the potential barrier of the nucleus by tunnelling.

If by thought we disconnect Coulomb interactions, such an α particle is linked to the rest of the nucleus by a nuclear potential of short range equation and depth corresponding to a potential energy that we will determine.

Schematically in the case of Uranium 238 the situation is considered as follows:

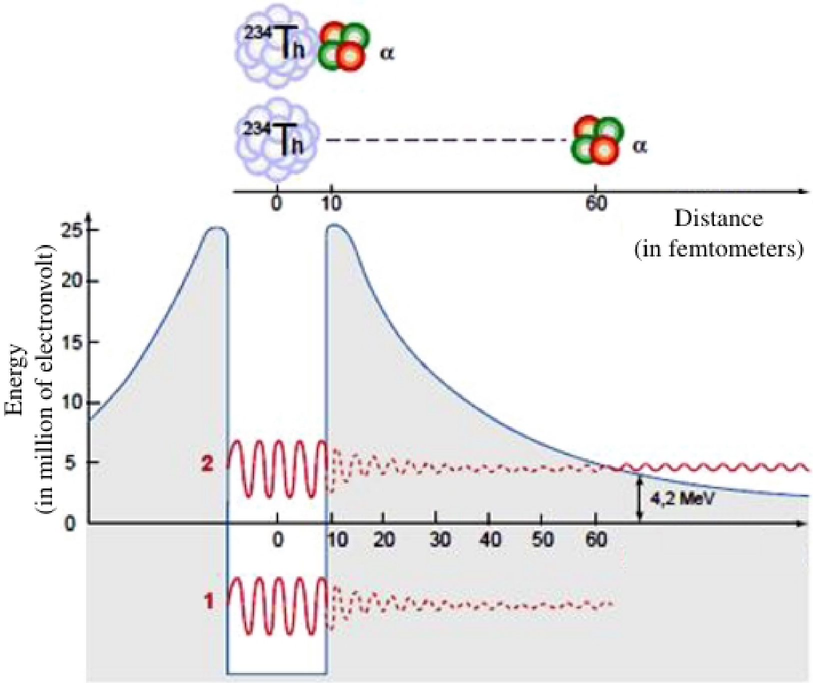


Figure 11.112: **Gamow's idea of Tunnelling effect (source: Pour la Science)**

In Classical Physics we would represent the α emission as the leak of the nucleus from the nucleus. This representation is not valid because it implies that the α particle is undergoing the electrostatic repulsion of the residual core of Thorium 234 would only be removed with an energy of about 25 [MeV]. However, the low experimentally observed value (of only 4.2 [MeV]) is found only with the tool of Quantum Physics.

Let's move on to the mathematical part:

Let us connect the Coulomb repulsion between the α particle of electric charge $+2e$ (two protons and two neutrons) and the rest of the nucleus, then of charge $+(Z-2)e$ outside the nuclear potential well.

We then get the expression of the potential energy (see section Electrostatics page 4005):

$$E_p=\frac{1}{4\pi\varepsilon_0}\frac{(\mathrm{Z}-2)q_e\cdot 2q_e}{r}=\frac{1}{4\pi\varepsilon_0}\frac{2(\mathrm{Z}-2)q_e^2}{r} \tag{11.4.159}$$

Where r is the distance between the center of the nucleus and the α particle. The potential energy therefore decreases with distance since the force is repulsive.

Now, let's have a qualitative approach to the phenomenon. We put that the probability T of tunnelling as being proportional, according to our results in the section of Wave Quantum Physics (see page 4840), to:

$$T \propto e^{-2kL} = \exp\left(-2L\frac{\sqrt{2m(E_p - E_{\text{tot}})}}{\hbar}\right) \tag{11.4.160}$$

knowing that it is, following the approximations we made during the proof, of a lower limit.

If we model the potential barrier of the core by a non-rectangular profile as presented below:

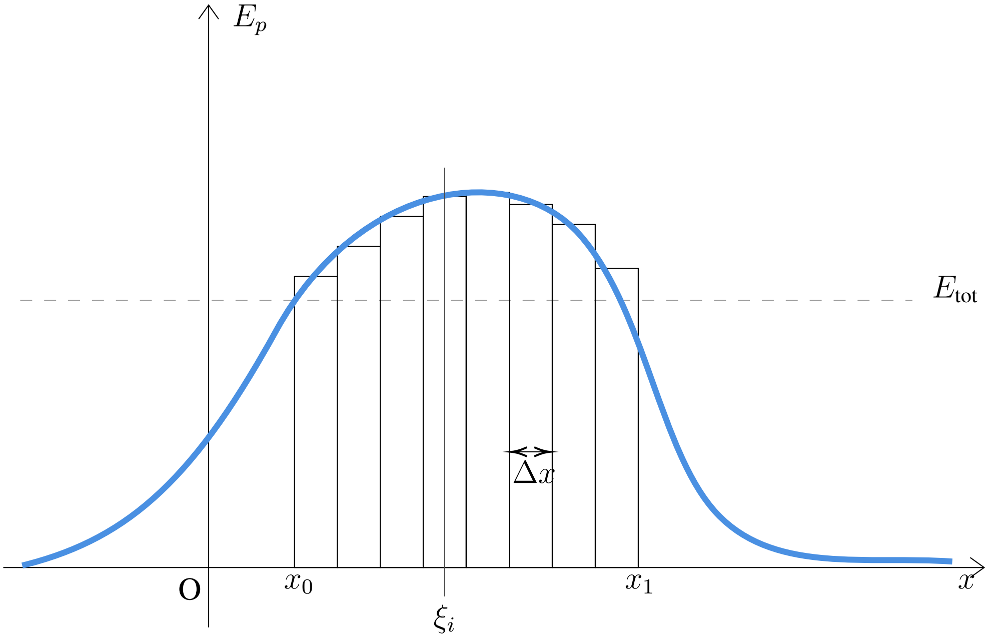


Figure 11.113: **Non-rectangular barrier profile of the nucleus**

where we replace the real profile of the curve by a series of barriers of thickness Δx and where the potential equals E_p at the point ξ_i

The probability of passing a barrier will therefore be proportional to:

$$T_i \propto \exp\left(-\frac{2}{\hbar}\Delta x\sqrt{2m\Big(E_p(\xi_i) - E_{\text{tot}}\Big)}\right) \tag{11.4.161}$$

and we know (see section Probabilities page 480) that the probability of passing one of the barriers is an independent event (mutually exclusive). We can therefore multiply the probabilities such that:

$$T \propto \exp\left(\sum_i -\frac{2}{\hbar}\Delta x\sqrt{2m\Big(E_p(\xi_i) - E_{\text{tot}}\Big)}\right) \tag{11.4.162}$$

And passing to the limit it comes:

$$T \propto \exp\left(-\frac{2}{\hbar}\int_{x_0}^{x_1}\sqrt{2m\Big(E_p(x) - E_{\text{tot}}\Big)}\,\mathrm{d}x\right) \tag{11.4.163}$$

and if x is assimilated to a radius of a spherical symmetry configuration:

$$T \propto \exp\left(-\frac{2}{\hbar}\int\sqrt{2m\left(E_p(r)-E_{\text{tot}}\right)}\mathrm{d}r\right) \tag{11.4.164}$$

Often denoted:

$$T \propto \exp(-2G) \tag{11.4.165}$$

where G is known as the "Gamow factor".

In the case of an α nucleus, the potential barrier ranges from r_1 where it begins to r_2 value where the potential barrier is considered as negligible.

But, the potential energy of the α at any remote point r from the outside of the border of the nucleus of the radioactive atom will be equal, as we have seen a earlier before, to:

$$E_p(r)=\frac{1}{4\pi\varepsilon_0}\frac{2(\mathrm{Z}-2)q_e^2}{r} \tag{11.4.166}$$

We thus have for $r_1 \leq r \leq r_2$:

$$E_p(r)-E_{\text{tot}}=\frac{1}{4\pi\varepsilon_0}\frac{2(\mathrm{Z}-2)q_e^2}{r}-E_{\text{tot}} \tag{11.4.167}$$

To determine E_{tot} of the emitted α nucleus, it is necessary to know that its total energy is supposed preserved in this simplified model. It is therefore the same as before it passes through the nuclear potential barrier when $r_1 \leq r$, during, and after r_2.

Moreover, in this model, kinetic energy is also assumed constant when $r \leq r_2$. In other words, since the α nucleus pre-exists in the nucleus of the radioactive atom, it already has the final velocity that it will have at the point of crossing the barrier of the nuclear potential ...

So under all these very simplifying assumptions... if we know how to determine the total energy of the α nucleus in r_2 (for example), at the output of the barrier, we have its total energy during the whole crossing phenomenon of the barrier.

Conversely, its total energy required to exit in r_2 of the potential barrier by tunnel effect starting from the nucleus of the radioactive atom (and then going away at infinity and gaining kinetic energy and losing all its potential Coulomb energy) is the same assuming that the total energy obtained by calculating the work of the force which from an infinite distance of the nucleus of the radioactive atom would make the α nucleus return inside the nucleus of the radioactive atom at the aforementioned velocity to the minimum exit point r_2 (minimum exit radius taken as constant because very distant in orders of magnitude with respect to the nucleus of the radioactive atom).

This then corresponds to the potential energy difference between a point at infinity and an r_2. And since the potential energy is zero at infinity for a repulsive system, there remains only the term:

$$E_{\text{tot}}=\frac{1}{4\pi\varepsilon_0}\frac{2(\mathrm{Z}-2)q_e^2}{r_2} \tag{11.4.168}$$

and finally:

$$E_p(r)-E_{\text{tot}}=\frac{1}{4\pi\varepsilon_0}\frac{2(\mathrm{Z}-2)q_e^2}{r}-\frac{1}{4\pi\varepsilon_0}\frac{2(\mathrm{Z}-2)q_e^2}{r_2}=\frac{2(\mathrm{Z}-2)q_e^2}{4\pi\varepsilon_0}\left(\frac{1}{r_2}-\frac{1}{r_2}\right) \tag{11.4.169}$$

still valid only for $r_1 \leq r \leq r_2$ (it is as if during the crossing of the barrier, the αnucleus restored kinetic energy to the vacuum as it approach of the point r_2, that said, in Quantum Physics one we cannot use the Classical Mechanics interpretation...).

Now, very often in laboratories, r_2 is expressed as a constant sufficiently far from the nucleus of the radioactive atom. It is then relatively natural (even if it is do-it-yourself physicist way of doing maths...) to take r as an integration variable such as:

$$T\propto\exp\left(-\frac{2}{\hbar}\int\limits_{r_1}^{r_2}\sqrt{\frac{2m2(\mathrm{Z}-2)q_e^2}{4\pi\varepsilon_0}\left(\frac{1}{r}-\frac{1}{r_2}\right)}\mathrm{d}r\right)=\exp\left(-2\frac{1}{\hbar}\sqrt{\frac{m4(\mathrm{Z}-2)q_e^2}{4\pi\varepsilon_0}}\int\limits_{r_1}^{r_2}\sqrt{\left(\frac{1}{r}-\frac{1}{r_2}\right)}\mathrm{d}r\right) \tag{11.4.170}$$

and it is traditional to put:

$$\gamma=\frac{1}{\hbar}\sqrt{\frac{m4(\mathrm{Z}-2)q_e^2r_2}{4\pi\varepsilon_0}} \tag{11.4.171}$$

which brings us to:

$$T\propto\exp\left(-\frac{2\gamma}{\sqrt{r_2}}\int\limits_{r_1}^{r_2}\sqrt{\left(\frac{1}{r}-\frac{1}{r_2}\right)}\mathrm{d}r\right) \tag{11.4.172}$$

Let us now do the following change of variables (the derivative of the $\cos^2(u)$ is detailed in the section of Differential and Integral Calculus at page 1475):

$$\cos^2(u)=\frac{r}{r_2}\Rightarrow\frac{\mathrm{d}r}{\mathrm{d}u}=r_2\frac{\mathrm{d}\cos^2(u)}{\mathrm{d}u}=-r_2 2\cos(u)\sin(u) \tag{11.4.173}$$

hence:

$$\mathrm{d}r=-r_2 2\cos(u)\sin(u)\mathrm{d}u \tag{11.4.174}$$

and by putting:

$$\cos^2(u_0)=\frac{r_1}{r_2} \tag{11.4.175}$$

The integral:

$$\int\limits_{r_1}^{r_2}\sqrt{\left(\frac{1}{r}-\frac{1}{r_2}\right)}\mathrm{d}r \tag{11.4.176}$$

becomes:

$$\begin{aligned}\int\limits_{r_1}^{r_2}\sqrt{\left(\frac{1}{r}-\frac{1}{r_2}\right)}\mathrm{d}r&=\int\limits_{r_1}^{r_2}\sqrt{\frac{1}{r}\left(1-\frac{r}{r_2}\right)}\mathrm{d}r=\int\limits_{r_1}^{r_2}\sqrt{\frac{1}{r}\left(1-\cos^2(u)\right)}\mathrm{d}r=\int\limits_{r_1}^{r_2}\sqrt{\frac{1}{r}\sin^2(u)}\mathrm{d}r\\&=\int\limits_{r_1}^{r_2}\frac{1}{\sqrt{r}}\sin(u)\mathrm{d}r=\int\limits_{r_1}^{r_2}\frac{1}{\sqrt{r_2\cos^2(u)}}\sin(u)\mathrm{d}r=\frac{1}{\sqrt{r_2}}\int\limits_{r_1}^{r_2}\frac{\sin(u)}{\cos(u)}\mathrm{d}r\\&=\frac{1}{\sqrt{r_2}}\int\limits_{r_1}^{r_2}\frac{\sin(u)}{\cos(u)}(-r_2 2\cos(u)\sin(u)\mathrm{d}u)\\&=-\frac{2r_2}{\sqrt{r_2}}\int\limits_{r_1}^{r_2}\sin^2(u)\mathrm{d}u=-2\sqrt{r_2}\int\limits_{r_1}^{r_2}\sin^2(u)\mathrm{d}u\end{aligned}\tag{11.4.177}$$

Concerning the terminals we have for recall:

$$\cos^2(u)=\frac{r}{r_2}\qquad\cos^2(u_0)=\frac{r_1}{r_2}\tag{11.4.178}$$

So if r is equal to r_1 we write the bound as u_0 and if r equals r_2 then:

$$\cos^2(u)=\frac{r_2}{r_2}=1\Rightarrow u=0\tag{11.4.179}$$

Therefore:

$$\int\limits_{r_1}^{r_2}\sqrt{\left(\frac{1}{r}-\frac{1}{r_2}\right)}\mathrm{d}r=-2\sqrt{r_2}\int\limits_{u_0}^{0}\sin^2(u)\mathrm{d}u=2\sqrt{r_2}\int\limits_{0}^{u_0}\sin^2(u)\mathrm{d}u\tag{11.4.180}$$

We have seen in the section of Differential and Integral Calculus:

$$I_n=\int\sin^n(x)\mathrm{d}x=\frac{1}{n}\left(-\cos(x)\sin^{n-1}(x)+(n-1)I_{n-2}\right)\tag{11.4.181}$$

Therefore:

$$\begin{aligned}I_2&=\int\sin^2(x)\mathrm{d}x=\frac{1}{2}(-\cos(x)\sin(x)+I_0)\\I_0&=\int\sin^0(x)\mathrm{d}x=\int\mathrm{d}x=x\end{aligned}\tag{11.4.182}$$

Then:

$$I_2=\int\sin^2(x)\mathrm{d}x=\frac{1}{2}(-\cos(x)\sin(x)+I_0)=\frac{1}{2}(-\cos(x)\sin(x)+x)\tag{11.4.183}$$

What makes:

$$\begin{aligned}2\sqrt{r_2}\int\limits_{0}^{u_0}\sin^2(u)\mathrm{d}u=2\sqrt{r_2}\frac{1}{2}(-\cos(x)\sin(x)+x)\Big|_{u_0}^{0}&=\sqrt{r_2}\frac{1}{2}(-\cos(x)\sin(x)+x))\Big|_{u_0}^{0}\\&=\sqrt{r_2}\left[(-\cos(u_0)\sin(u_0)+u_0)-(\cos(0)\sin(0)+0)\right]\\&=\sqrt{r_2}\left[(-\cos(u_0)\sin(u_0)+u_0)\right]\end{aligned}\tag{11.4.184}$$

But, we also have (see section Trigonometry page 2519):

$$\sin(2u_0)=2\sin(u_0)\cos(u_0) \tag{11.4.185}$$

Therefore:

$$2\sqrt{r_2}\int\limits_0^{u_0}\sin^2(u)\mathrm{d}u=\sqrt{r_2}\left(-\frac{\sin(2u_0)}{2}+u_0\right)=\sqrt{r_2}\left(\frac{(2u_0-\sin(2u_0))}{2}\right) \tag{11.4.186}$$

Let us recall again that:

$$\cos^2(u_0)=\frac{r_1}{r_2}\Rightarrow u_0=\arccos\left(\sqrt{\frac{r_1}{r_2}}\right) \tag{11.4.187}$$

But, $r_2\gg r_1$, therefore, $r_1/r_2\cong 0$.

If we develop in Maclaurin series (see section Sequences and Series page 1703) to the first order:

$$\begin{gathered}u_0\cong\arccos(0)+\frac{1}{1!}\frac{\mathrm{d}\arccos\left(\sqrt{\frac{r_1}{r_2}}\right)_{r_1=0}}{\mathrm{d}r_1}r_1=\frac{\pi}{2}-\frac{1}{\sqrt{\frac{r_1}{r_2}}r_2\sqrt{1-\frac{r_1}{r_2}}}r_1\cong\frac{\pi}{2}-\frac{1}{\sqrt{\frac{r_1}{r_2}}}\frac{r_1}{r_2}\\=\frac{\pi}{2}-\sqrt{\frac{r_2}{r_1}\frac{r_1}{r_2}}\cdot\sqrt{\frac{r_2}{r_1}}\sqrt{\frac{r_1}{r_2}}=\frac{\pi}{2}-\sqrt{\frac{r_1}{r_2}}\end{gathered} \tag{11.4.188}$$

Therefore:

$$u_0\cong\frac{\pi}{2}-\sqrt{\frac{r_1}{r_2}} \tag{11.4.189}$$

We then have:

$$\begin{aligned}\sqrt{r_2}\left(\frac{(2u_0-\sin(2u_0))}{2}\right)&\cong\sqrt{r_2}\frac{\left(2\left(\frac{\pi}{2}-\sqrt{\frac{r_1}{r_2}}\right)-\sin\left(2\left(\frac{\pi}{2}-\sqrt{\frac{r_1}{r_2}}\right)\right)\right)}{2}\\&=\sqrt{r_2}\frac{\left(\pi-2\sqrt{\frac{r_1}{r_2}}-\sin\left(\pi-2\sqrt{\frac{r_1}{r_2}}\right)\right)}{2}=\sqrt{r_2}\left(\frac{\pi}{2}-\sqrt{\frac{r_1}{r_2}}-\frac{1}{2}\sin\left(\pi-2\sqrt{\frac{r_1}{r_2}}\right)\right)\\&=\sqrt{r_2}\left[\frac{\pi}{2}-\sqrt{\frac{r_1}{r_2}}-\frac{1}{2}\left(\sin(\pi)\cos\left(-2\sqrt{\frac{r_1}{r_2}}\right)+\cos(\pi)\sin\left(-2\sqrt{\frac{r_1}{r_2}}\right)\right)\right]\\&=\sqrt{r_2}\left[\frac{\pi}{2}-\sqrt{\frac{r_1}{r_2}}-\frac{1}{2}\left(\cos(\pi)\sin\left(-2\sqrt{\frac{r_1}{r_2}}\right)\right)\right]=\sqrt{r_2}\left[\frac{\pi}{2}-\sqrt{\frac{r_1}{r_2}}+\frac{1}{2}\sin\left(-2\sqrt{\frac{r_1}{r_2}}\right)\right]\end{aligned} \tag{11.4.190}$$

If we take the Maclaurin development to the first order:

$$\sin\left(-2\sqrt{\frac{r_1}{r_2}}\right)\cong-2\sqrt{\frac{r_1}{r_2}} \tag{11.4.191}$$

Therefore:

$$\begin{aligned}\sqrt{r_2}\left(\frac{(2u_0-\sin(2u_0))}{2}\right)&\cong\sqrt{r_2}\left[\frac{\pi}{2}-\sqrt{\frac{r_1}{r_2}}+\frac{1}{2}\sin\left(-2\sqrt{\frac{r_1}{r_2}}\right)\right]=\sqrt{r_2}\left[\frac{\pi}{2}-\sqrt{\frac{r_1}{r_2}}+\frac{1}{2}\left(-2\sqrt{\frac{r_1}{r_2}}\right)\right]\\&=\sqrt{r_2}\left[\frac{\pi}{2}-\sqrt{\frac{r_1}{r_2}}-\sqrt{\frac{r_1}{r_2}}\right]=\sqrt{r_2}\left[\frac{\pi}{2}-2\sqrt{\frac{r_1}{r_2}}\right]\end{aligned}\tag{11.4.192}$$

So all this to write finally:

$$T\propto\exp\left(-\frac{2\gamma}{\sqrt{r_2}}\int_{r_1}^{r_2}\sqrt{\left(\frac{1}{r}-\frac{1}{r_2}\right)}\mathrm{d}r\right)=\exp\left(-\frac{2\gamma}{\sqrt{r_2}}\sqrt{r_2}\left[\frac{\pi}{2}-2\sqrt{\frac{r_1}{r_2}}\right]\right)=\exp\left(-\gamma\left[\pi-4\sqrt{\frac{r_1}{r_2}}\right]\right)\tag{11.4.193}$$

Relation in which we can put again the coefficient of the exponential that we had determined in the section of Wave Quantum Physics. It is the exponential factor in the above relation which explains the great variation of the radioactive periods of the different nuclides, whereas the energy of the particles varies relatively little.

$$\boxed{T\propto\exp\left(-\frac{1}{\hbar}\sqrt{\frac{m4(\mathrm{Z}-2)q_e^2r_2}{4\pi\varepsilon_0}}\cdot\left(\pi-4\sqrt{\frac{r_1}{r_2}}\right)\right)}\tag{11.4.194}$$

Typically for the Uranium $^{238}\mathrm{U}$ nucleus, we can take the values from numerical (experimental) tables of the physical and universal constants that are in the previous relation to get a certain value of T (we will omit this numerical application for now since the numerical values in various tables are not all in agreement still in this early 121st century...).

Then, in the semi-classical approximation, the α nucleus has, in the potential well, a velocity of the order of:

$$v=\sqrt{\frac{2E_\alpha}{m}}\tag{11.4.195}$$

And it goes back and forth in a nucleus whose radius is of the order of r_1. These round trips therefore correspond to a certain number of oscillations per second. Indeed, if we denote by τ the average duration between of two successive shocks, we then have:

$$\tau=\frac{2r_1}{v}=\frac{2r_1}{\sqrt{\dfrac{2E_\alpha}{m}}}\tag{11.4.196}$$

Therefore the frequency is equal to:

$$f=\frac{1}{\tau}\tag{11.4.197}$$

Each time it has a probability T to cross the potential barrier. This probability per unit time is thus determined by:

$$\lambda\cong f\cdot T\tag{11.4.198}$$

and gives the disintegration constant λ of the isotope by α emission with a relatively large error if we make the calculation with the numerical values. Otherwise, the order of magnitude is by contrast exact what is not bad at all! Therefore the simplified academic approach above gives satisfactory results.

11.4.2.4.4 Beta- Disintegration (4)

Definition (#11.4.515): When a nucleus is unstable due to an overflow of neutrons (like Carbon 14 for example) it will not emit neutrons. On the other hand, it will have the ability to change one of its neutrons into a proton. During this transformation, to keep the total electrical charge of the system, an electron will be created. This transformation is named "β^- radioactivity" (because the electron has a negative charge in this disintegration).

The so-named β^- disintegration is therefore a characteristic of nuclei with an excess of neutrons. The isotopes concerned become more stable by transforming a neutron into a proton with emission of an electron β^- and a particle named "antineutrino" $\bar{\nu}$ which we will justify the presence further below.

We then have for the concerned neutron:

$$ {}_0^1\mathrm{n}^{+1/2} \rightarrow \mathrm{p}_+^{+1/2} + \beta_-^{+1/2} + \bar{\nu}_0^{-1/2} \tag{11.4.199}$$

We have put in right exponent the spin of the concerned particle and in right index the sign of the electric charge of the particle. Thus, we observe that the spin is a conserved quantity, as well as the electric charge.

We have for the concerned isotope:

$$ {}_Z^A\mathrm{X}_N \rightarrow {}_{Z+1}^{A}\mathrm{Y}_{N-1} + \beta^- + \bar{\nu} \tag{11.4.200}$$

✍**Example:**

The famous case of Radiocarbon dating:

$$ {}_6^{14}\mathrm{C} \rightarrow {}_7^{14}\mathrm{N} + \beta^- + \bar{\nu} \tag{11.4.201}$$

The energy released during the transmutation is calculated by means of the mass default:

$$ \mathrm{Q}_{\beta^-} = (M_X - M_Y - m_e)c^2 \tag{11.4.202}$$

By neglecting the binding energy of the electrons, we have:

$$ M_X = \mathrm{m}_X + \mathrm{Z}m_e \quad \text{and} \quad M_Y = \mathrm{m}_Y + (\mathrm{Z}+1)m_e \tag{11.4.203}$$

Caution!!! The Z in the equality of M_X is the same as that found in the expression of M_Y hence the $Z+1$.

We have then:

$$ \boxed{\mathrm{Q}_{\beta^-} = (m_X - m_Y)\cdot c^2} \tag{11.4.204}$$

Each pure β^- disintegration is characterized by a fixed decay energy Q. Since the kinetic energy of the nucleus is negligible by its mass compared to that of the combined electron and antineutrino, the released energy Q is shared between the kinetic energies of the β^- and $\bar{\nu}$. Since the mass of the antineutrino is very far below that of the electron, the kinetic energy of the antineutrino can also be neglected. However, the energy of the β^- is not fixed and can take any value between 0 and Q. We thus observe an energy spectrum unlike other types of radioactivity (because the electron may have variable kinetic energy).

The shape of the observed distributions makes it possible to give an average energy value to the β^- which is around:

$$\bar{E} \cong \frac{\mathrm{Q}_{\beta^-}}{3} = \frac{E_{\text{max}}}{3} \tag{11.4.205}$$

The existence of antineutrino was postulated in 11933 according to holocene calendar (three years after the neutrino that we will see below) by Wolfgang Pauli in order to satisfy the spin conservation that is axiomatically considered as a fundamental law. The introduction of such a strange particle was very controversial and not widely accepted (zero charge, non-zero spin, negligible mass) and it continues to pose some problems in the contemporary physics of the 121st century (holocene calendar).

Remark

Photons have spin 1 therefore according to spin conservation law a photon decaying into two photons does should not be possible as it would not conserve spin.

If we split a photon into two photons you end up with one of these combinations of spin :

- Spin -1 and spin -1 - does not total spin $+1$
- Spin -1 and spin $+1$ - does not total spin $+1$
- Spin $+1$ and spin -1 - does not total spin $+1$
- Spin $+1$ and spin -1 - does not total spin $+1$

None of them conserve spin so none of them are possible as far as we know.

Independently of the "electron neutrino" (denoted usually ν_e) accompanying the particles β^- and β^+ (the latter having several names "positron" or "positive electron") there exists a neutrino of the meson μ (muon) and a neutrino of the tau (tauon) that are respectively denoted: ν_μ and μ_τ not to confuse them. Subsequently, as we will not be confronted with the neutrino of meson or tau neutrinos, we will simply denote the electron neutrino ν instead of ν_e or of υ.

Remark

At the beginning of its discovery, the disintegration β^- was seen as a transmutation of the nucleus ... in small classes, even today, it is seen as the transformation of a neutron into a proton. In contemporary theories, it is seen as a the transformation of a quark d into quark u and it has led physicists to develop the theory of weak interaction to explain its origin.

Betavoltaic devices, also known as betavoltaic cells, are generators of electric current, in effect a form of battery, which use energy from a radioactive source emitting beta particles (electrons). A common source used is the hydrogen isotope, tritium. Unlike most nuclear power sources, which use nuclear radiation to generate heat, which then is used to generate electricity (thermoelectric and thermionic sources), betavoltaics use a non-thermal conversion process; converting the electron-hole pairs produced by the ionization trail of beta particles traversing a semiconductor.

Betavoltaic power sources (and the related technology of alphavoltaic power sources) are particularly well-suited to low-power electrical applications where long life of the energy source is needed, such as implantable medical devices or military and space applications.

The actual technology provide a source of continuous nanowatt-to-microwatt battery power that is resistant to a broad range of temperatures and other environmental conditions for extended periods of 20 or more years.

CITY LABS
NanoTritium TM Battery
Contains H-3 Max Activity: 2 Ci
Caution - Radioactive Material
Model P100a
APR 2012
SN 0420120002
CITY LABS
NanoTritium TM Battery
Contains H-3 Max Activity: 2 Ci
Caution - Radioactive Material
Model P100a
APR 2012
SN 0420120001

Figure 11.114: City Labs Nanotritium™ battery

For example, the amount of electricity is tied with half-life. For example, if ^{63}Ni has half-life of 100 years this means, that mole of Nickel (63 grams) will produce Avogadro/2 electrons during that 100 years. This means 10^{21} electrons per year and 10^{14} electrons per second.

This means up to 0.1 [mA] or electric current.

The total decay energy of ^{63}Ni is 67 [keV] and is approximately the maximum energy of electron. But, since the decay also produces neutrino the mean energy of electron is much smaller: 17 [keV] $(2.732 \cdot 10^{-15}$ [J]). This means that each electron has 17 kilovolts of electric tension.

So, the power of electricity from one mole of ^{63}Ni:

$$\begin{aligned} W = A \cdot E_{e^-} \cdot m &= {}^{63}\text{Ni specific activity} \cdot (2.732 \cdot 10^{-15}) \cdot 63 \\ &= (2.11 \cdot 10^{12}) \cdot (2.732 \cdot 10^{-15}) \cdot 63 \cong 0.36\,[\text{W}] \end{aligned} \tag{11.4.206}$$

This is not sufficient to provide iPhone 4 peak power consumption, which is about 1.5 [W]. So atomic batteries can actually not supersede conventional batteries, and serve for years.

Also, it must be noticed that the usage of radioactive batteries definitely require regulations for handling and probably would not be allowed inside a single unit for unrestricted civilian use. We know that when properly used betavoltaics are safe. But what about im-proper use / improper disposal / potential for abuse? At any rate, current perception of nuclear power by general public is not that good, so marketing nuclear batteries will present certain challenge. Also the price of Plutonium or Americanium batteries (that are superior to batteries in almost every way) have cost that limits them to micro power devices (one gram is worth almost $1,000$ dollars...).

11.4.2.4.5 Beta+ Disintegration (5)

Definition (#11.4.516): When a nucleus is unstable due to an overflow of protons it will not emit protons. On the other hand, it will have the ability to change one of its protons into a neutron. During this transformation, to keep the total electrical charge of the system there are two ways for this disintegration to occur:

- By capture of a negative electron and then we speak of "electronic capture radioactivity" (see further below)
- Or by emission of positive and then this transformation is named "β^+ radioactivity" (because the electron detected has a positive charge in this disintegration).

Therefore during a β^+ disintegration a proton is dissociated into a neutron, a positive electron ("positron") denoted β^+ and of a neutrino (which we will justify the presence a little lower).

This transformation has a ridiculously low probability since the inverse of the emission of an electron and an antineutrino would be the simultaneous capture of these two particles... and such a reunion of circumstances would be a miracle. To overcome this difficulty, the nucleus uses a quantum subterfuge: the emission of a particle is equivalent to the capture of its antiparticle. This joker then offers the aforementioned possibilities to the surplus nucleus in protons. Indeed, to carry out the inverse of the β^- disintegration, the solution consists for the nucleus to use the conservation of the energy and of the spin by emitting a positron and capturing in the quantum energy of vacuum (see section Quantum Field Theory page 5158) an antineutrino and to emit in exchange a neutrino.

We write this:

$$\mathrm{p}\rightarrow{}_0^1\mathrm{n}+\beta^++\nu \tag{11.4.207}$$

or:

$${}_Z^A\mathrm{X}_N\rightarrow{}_{Z-1}^{\;\;A}\mathrm{Y}_{N+1}+\beta^++\nu \tag{11.4.208}$$

The energy released during the transmutation is calculated by means of the mass default:

$$\mathrm{Q}_{\beta^+}=(\mathrm{M}_X-\mathrm{M}_Y-m_e)c^2 \tag{11.4.209}$$

By neglecting the binding energy of the electrons, we have:

$$M_X=\mathrm{m}_X+\mathrm{Z}m_e \quad \text{and} \quad M_Y=\mathrm{m}_Y+(\mathrm{Z}-1)m_e \tag{11.4.210}$$

Caution!!! The Z in the equality of M_X is the same as that found in the expression of M_X hence the $Z+1$.

We have then:

$$\boxed{\mathrm{Q}_{\beta+}=(m_X-m_Y-2m_e)c^2} \tag{11.4.211}$$

The β^+ disintegration can therefore take place only if $Q_{\beta+}\geq 0$, that is to say if:

$$(m_X-m_Y)\cdot c^2\geq 2m_ec^2=1.022\,[\mathrm{MeV}] \tag{11.4.212}$$

The mass energy of the electron $E=m_ec^2$ is important because it is the energy of one of the two photons resulting from an annihilation of a positron β^+ with an electron β^-.

As for the β^- disintegration, the energy of the β^+ is not fixed and can take any value between 0 and Q. We thus observe an energy spectrum.

Remark

A real particle-antiparticle pair can annihilate and generate two real photons. Quantum electrodynamics is time-reversal invariant, so the opposite process, in which two real photons disappear and create a real particle-antiparticle pair, is also possible. But this isn't named "photon annihilation"; it's named named "two-photon pair production". "Annihilation" is usually reserved to mean the transformation of a massive particle-antiparticle pair into massless particles.

11.4.2.4.6 Electronic capture (6)

Definition (#11.4.517): When a nucleus is unstable because of an overflow of protons with respect to neutrons, we know now that a favourable solution from the point of view of its energy is to transform one of its protons into a neutron, i.e. to realize the inverse of the β^- radioactivity. We saw just earlier that a possibility was for the nucleus via the disintegration β^+ to catch an antineutrino of the vacuum and to emit a positron (loss of its electric charge) and a neutrino. But it can also capture an electron from the electronic cloud (neutralizing its electrical charge) instead of emitting a positron, this is why we name this an "electronic capture" or "electron capture".

This will most often be an electron of layer K. What is written:

$${}^{A}_{Z}\mathrm{X}_N+e\rightarrow {}^{A}_{Z-1}\mathrm{Y}_{N+1}+\nu \tag{11.4.213}$$

The energy released during the transmutation is calculated by means of the mass default:

$$\mathrm{Q}_{\mathrm{EC}}=(M_X+m_e-M_Y)c^2 \tag{11.4.214}$$

assuming that the binding energy of the electron of the layer K and that of recoil of the nucleus are negligible.

It is therefore the electron neutrino that carries all the energy, hence the necessity that Wolfgang Pauli had to introduce this new particle (what had horrified him ...!). Since the captured electron occupies a precise energy level in the atom, the neutrinos resulting from the disintegration of an isotope by electronic capture have a determined energy and therefore have a discrete spectrum.

By neglecting the binding energy of electrons, we have:

$$M_X = m_X + Zm_e \quad \text{and} \quad M_Y = m_Y + (Z-1)m_e \tag{11.4.215}$$

Therefore:

$$\boxed{Q_{\text{EC}} = (m_X - m_Y)c^2} \tag{11.4.216}$$

The disintegration by electronic capture is in competition with the β^+ disintegration only if:

$$(m_X - m_Y)c^2 \geq 2m_e c^2 = 1.022\,[\text{MeV}] \tag{11.4.217}$$

in the case where:

$$(m_X - m_Y)c^2 < 2m_e c^2 = 1.022\,[\text{MeV}] \tag{11.4.218}$$

only the disintegration by electronic capture is possible.

However, the hole left by the absorbed electron requires a rearrangement of the atomic cloud and the emission of radiation.

Remark

Reaction of electron capture is how you turn White Dwarf star into a neutron star (see section Astrophysics page 5606).

11.4.2.4.7 Gamma emission (7)

Definition (#11.4.518): For the nucleus, the emission of electromagnetic radiation γ is a possibility of gaining stability. This emission generally follows a phenomenon of disintegration β^-, β^+, α or electronic capture. We can imagine that, in such types of disintegration, the topology of the nucleons in the nucleus is not ideal and that the rearrangement of the latter will be accompanied by a reduction of energy; The latter emitted in the form of one or more photons.

So we have a schema (β^-) disintegration:

$${}_{Z}^{A}\text{X}_N \rightarrow {}_{Z+1}^{A}\text{Y}_{N-1}^{m} + \beta^- + \bar{\nu} \tag{11.4.219}$$

then:

$${}_{Z+1}^{A}\text{Y}_{N-1}^{m} = {}_{Z+1}^{A}\text{Y}_{N-1} + \gamma \tag{11.4.220}$$

where the m means "metastable" or "isomer" (this last term is used when the emission of radiation takes place long time after disintegration).

Remark

"Isomer" means, in nuclear physics, that the nucleus is excited. It will de-energize with a period $T_{1/2}$. Generally the $T_{1/2}$ is extremely small and the photon(s) is (are) emitted immediately after the electron in the case of our example of a β^- disintegration. We then talk about "metastable state" or "isomer". It should be noted that these radioisotopes are particularly interesting in medical imaging.

The energy of the γ photon is equal to:

$$\boxed{Q_\gamma = (M_{Y^m} - M_Y)c^2} \tag{11.4.221}$$

Now let us suppose that a proton is confined in a box of width $L = 1.00 \cdot 10^{-14}$ [m] (a typical nuclear radius). If we assume that the proton confined in the nucleus can be modelled as a quantum particle in a box, all we need to do is to use the relation proven in the section of Wave Quantum Physics for an infinite potential wall box to find its energies E_1 and E_2:

$$E_n = n^2 \frac{\pi^2\hbar^2}{2mL^2} \tag{11.4.222}$$

Therefore the emitted photo has for energy:

$$\Delta E = h\nu = E_2 - E_1 = 2^2\frac{\pi^2\hbar^2}{2mL^2} - 1^2\frac{\pi^2\hbar^2}{2mL^2} = (4-1)\frac{\pi^2\hbar^2}{2mL^2} = 3\frac{\pi^2\hbar^2}{2mL^2} \cong 6.15\,[\text{MeV}] \tag{11.4.223}$$

Hence:

$$\nu = \frac{\Delta E}{h} \cong 1.49 \cdot 10^{21}\,[\text{Hz}] \tag{11.4.224}$$

This is the typical frequency of a gamma ray emitted by a nucleus. The energy of this photon is about 10 million times greater than that of a visible light photon!

11.4.2.4.8 Internal conversion (8)

Definition (#11.4.519): The "internal conversion (I.C.)" is a process related to the emission of a γ photon. Indeed, it is possible that the energy is transmitted directly to an electron of the electronic cloud, generally of the layer K, which is then ejected from the atom. This electron is named the "conversion electron". The place left in the electronic cloud is subsequently filled by an electron of the upper layers and so on. Thus, as in the case of an electronic capture decay process, there is a rearrangement of the electronic cloud characterized by the emission of X-rays characteristic by the element Y.

The energy transmitted is:

$$\boxed{E_{e^-} = E_\gamma - E_b} \tag{11.4.225}$$

with E_{e^-} being the kinetic energy of the emitted electron, E_γ the photon energy percutting the electron, E_b, the binding energy of the electron considered (K, L, M, ...).

The energy of the γ photon is transmitted directly to an electron that is ejected. The process is followed by the rearrangement of the electrons (an emission of X-rays will follow). The ejected electron is named "Auger electron" and sometimes the whole effect is named the "Auger effect".

The emission of a Auger electron is thus a process similar to the internal conversion process (IC), but the electromagnetic radiation does not come from a de-excitation of the nucleus (it is not a γ photon) but of a X-ray produced during the rearrangement of the electronic cloud. In a radioactive process, this electronic rearrangement can come either from an electronic capture (EC) or from an internal conversion (IC).

The ejected Auger electron comes mainly from an internal orbit and its energy is the characteristic energy of the X-ray minus its binding energy. The energy of the Auger electrons is therefore quite small (some [keV]) with respect to a β^- particle and these electrons are therefore often reabsorbed inside the source. The process of emission of an Auger electron is favoured for elements with low atomic number because of their low energies of electronic connection.

During a rearrangement of the electronic cloud such as the passage of an electron from the layer L to the layer K, the energy of the X-ray emitted will be equal to $E_\text{K}-E_\text{L}$. Since this energy difference is greater than the binding energy of another electron on layer L, the latter will then be emitted with the kinetic energy:

$$E_\text{Auger}=h\nu_{\text{L}\rightarrow\text{K}}-E_\text{L}=(E_\text{K}-E_\text{L})-E_\text{L}=E_\text{K}-2E_\text{L}>0=c^{te} \tag{11.4.226}$$

In turn, the 2 vacations left on layer L are filled with electrons from the upper layers. Fluorescence and Auger electron are then in competition. It is even possible that several Auger electrons are emitted during the de-excitation of the atom. This is named "Auger cascade" leaving the atom considered highly ionized, which can lead to the Coulomb explosion of the molecule of which it is a part.

To conclude on all these radioactive phenomena, let us indicate the order of magnitude of the radioactive periods of some natural and artificial elements:

Nuclide	**Radioactivity**	**Period**
Thorium 232	α	$\backsim 10^{10}$ years
Uranium 238	α	$\backsim 10^{9}$ years
Uranium 235	α	$\backsim 10^{8}$ years
Uranium 233	α	$\backsim 10^{5}$ years
Plutonium 239	α	$\backsim 10^{4}$ years
Plutonium 238	α	$\backsim 88$ years
Radium 226	α	$\backsim 10^{3}$ years
Curium 242	α	160 days
Potassium 40	β^-	$\backsim 10^{9}$ years
Carbon 14	β^-	$\backsim 10^{3}$ years
Tritium	β^-	21 years
Cobalt 60	β^-	$\backsim 5.3$ years
Iod 131	β^-	$\backsim 8$ days
Nitrogen 16	β^-	$\backsim 7.1$ seconds
Technetium 97	EC	$\backsim 10^{6}$ years
Cobalt 58	β^+	$\backsim 20$ minutes
Fluorine 18	β^+	$\backsim 110$ minutes

Table 11.17: **Amplitudes of the radioactive periods of a few nuclides**

Again, this is a quantum phenomenon that, in most interpretations of quantum mechanics, occurs spontaneously, that is, without cause.

11.4.3 Radiation protection

In nuclear physics (and also in "high energy astrophysics"), it is very important to know how the various α, γ, X-ray or neutron radiation interact with matter (roughly uncharged or charged radiation). This allows us to know how their kinetic energy is distributed or dissipated in the material they encounter along their path and to protect themselves in a suitable way.

"Radiation protection", sometimes known as "radiological protection", is defined by the International Atomic Energy Agency (IAEA) as "*The protection of people from harmful effects of exposure to ionizing radiation, and the means for achieving this*". The IAEA also states «*The accepted understanding of the term radiation protection is restricted to protection of people. Suggestions to extend the definition to include the protection of non-human species or the protection of the environment are controversial*».

Ionizing radiation is widely used in industry and medicine, and can present a significant health hazard. It causes microscopic damage to living tissue, which can result in skin burns and radiation sickness at high exposures (known as "tissue" or "deterministic" effects), and statistically elevated risks of cancer at low exposures ("stochastic effects").

Fundamental to radiation protection is the reduction of expected dose and the measurement of human dose uptake. For radiation protection and dosimetry assessment the International Committee on Radiation Protection (ICRP) and International Commission on Radiation Units and Measurements (ICRU) have published recommendations and data which is used to calculate the biological effects on the human body, and set regulatory and guidance limits.

Before investing this subject we must prove some relations and study some phenomena related to high energy radiations and behaviours of particles. So let us tackle this first!

11.4.3.1 Bethe formula

The "Bethe formula" describes the mean energy loss per distance travelled of swift charged particles (protons, alpha particles, atomic ions) traversing matter (or alternatively the stopping power of the material). For electrons the energy loss is slightly different due to their small mass (requiring relativistic corrections) and their indistinguishability, and since they suffer much larger losses by Bremsstrahlung, terms must be added to account for this. Fast charged particles moving through matter interact with the electrons of atoms in the material. The interaction excites or ionizes the atoms, leading to an energy loss of the travelling particle.

The non-relativistic version was found by Hans Bethe in 11930 (holocene calendar). The Bethe formula is sometimes named "Bethe-Bloch formula", but this is quite misleading.

A heavy charged particle having an energy of one or more MeV loses its energy mainly by collisions with the electrons cloud of atoms, electrons which appear to it as quasi-free. The process by which electrons are thus ejected during the passage of an ionizing particle is named "primary ionization electrons". An electron can escape if it receives an energy higher than its binding energy.

The maximum energy transfer E_{max} that can occur in a non-relativistic and elastic collision (where the system energy is conserved because there is by definition no heat - radiation - dissipation) is calculated simply by using the conservation principle of the linear momentum and energy:

Given M,v_M and m_e, v_e be the masses and velocities of the incident particle and the electron respectively. We assume that the electron is immobile in its orbit and that its initial velocity is zero ($v_e=0$). After the shock, we will assume that the incident particle will have transferred all its kinetic energy to the electron and will in turn find itself at rest such as $v'_M=0$.

Let us first consider the conservation of linear momentum:

$$Mv_M+m_ev_e=Mv'_M+m_ev'_e \tag{11.4.227}$$

The conservation of energy also allows us to write:

$$\frac{1}{2}Mv_M^2+\frac{1}{2}m_ev_e^2=\frac{1}{2}Mv_M'^2+\frac{1}{2}m_ev_e'^2 \tag{11.4.228}$$

Hence, after regrouping and simplification:

$$\begin{gathered}M\left(v_M-v'_M\right)=m_e\left(v'_e-v_e\right)\\ M\left(v_M^2-v_M'^2\right)=m_e\left(v_e'^2-v_e^2\right)\end{gathered} \tag{11.4.229}$$

Or otherwise written:

$$M\left(v_M-v'_M\right)=m_e\left(v'_e-v_e\right) \tag{11.4.230}$$

$$M\left(v_M+v'_M\right)\left(v_M-v'_M\right)=m_e\left(v'_e-v_e\right)\left(v'_e+v_e\right) \tag{11.4.231}$$

Then, after dividing the second equation by the first we have:

$$v_M+v'_M=v'_e+v_e \tag{11.4.232}$$

We then have the system:

$$\left.\begin{aligned}Mv_M+m_ev_e=Mv'_M+m_ev'_e\\ v'_M=v'_e+v_e-v_M\end{aligned}\right\}\Rightarrow m_ev'_e=Mv_M+m_ev_e-M(v'_e+v_e-v_M) \tag{11.4.233}$$

we then deduce the expression of the speed after the impact:

$$v'_e=\frac{v_e(m_e-M)+2Mv_M}{M+m_e} \tag{11.4.234}$$

relatively to our initial hypotheses ($v_e=0$, $v'_M=0$), we then have:

$$v'_e=\frac{2Mv_M}{M+m_e} \tag{11.4.235}$$

Let us manipulate this relation a bit:

$$\begin{gathered}v'_e=\dfrac{2Mv_M}{M+m_e}\Rightarrow\dfrac{1}{2}m_ev'_ev'_e=\dfrac{1}{2}m_ev'_e\dfrac{2Mv_M}{M+m_e}\Rightarrow E'^e_c=\dfrac{1}{2}Mv_Mv_M\dfrac{2m_ev'_e}{v_M(M+m_e)}\\ \Rightarrow E'^e_c=E^M_c\dfrac{2m_e}{(M+m_e)}\dfrac{v'_e}{v_M}=E^M_c\dfrac{2m_e}{(M+m_e)}\dfrac{\left(\dfrac{2Mv_M}{M+m_e}\right)}{v_M}=\dfrac{4Mm_e}{(M+m_e)^2}E^M_c\end{gathered}\tag{11.4.236}$$

For a heavy particle, with $M\gg m_e$ we can write:

$$_{\text{max}}E^e_c=\frac{4m_e}{M}\cdot E^M_c\tag{11.4.237}$$

An ionization can only occur if the $_{\text{max}}E^e_c$ is at least equal to the ionization threshold of the electron that we will denote by I_0 and whose calculation was seen during the study of the Bohr model (see section Corpuscular Quantum Physics page 4626).

The energy of the incident particle must therefore at least be equal to:

$$E^M_{\text{min}}=\frac{M\cdot I_0}{4m_e}\tag{11.4.238}$$

Thus, when passing through matter, the electrically charged body of charge $Z\cdot q_e$ and speed v_M yields its energy in numerous collisions with the electrons of the atoms encountered. The interaction is of the Coulomb type and each time a diffusion occurs. The recoil energy of the electron, assumed to be free, can be calculated accurately. To estimate the loss of energy, we will here approximate that the quantity of linear momentum transferred Δp is equal to the product of the interaction force at the distance r multiplied by the time necessary for the projectile to traverse the path $2r$. We have the Coulomb for F given for recall by (see section Electrostatics page 3968):

$$F=\frac{1}{4\pi\varepsilon_0}\cdot\frac{\mathrm{Z}q_e^2}{r^2}\tag{11.4.239}$$

and the linear momentum:

$$\Delta p=F\cdot t=\left(\frac{1}{4\pi\varepsilon_0}\cdot\frac{\mathrm{Z}q_e^2}{r^2}\right)\cdot\frac{2r}{v_M}=\frac{1}{4\pi\varepsilon_0}\frac{2\mathrm{Z}q_e^2}{rv_M}\tag{11.4.240}$$

The kinetic energy transferred to an electron of mass m_e will be:

$$\Delta E^e_c=\frac{1}{2}m_ev_M^2=\frac{1}{2}\frac{(\Delta p)^2}{m_e}=\frac{1}{(4\pi\varepsilon_0)^2}\cdot\frac{2\mathrm{Z}^2q_e^4}{r^2v_M^2m_e}\tag{11.4.241}$$

The total energy loss will be obtained by integrating on all the electrons encountered. At the distance range that is between r and $r+\mathrm{d}r$ of the trajectory and on the path of length $\mathrm{d}x$, there are:

$$\mathrm{d}N=N\cdot\mathrm{Z}'\cdot2\pi r\cdot\mathrm{d}r\cdot\mathrm{d}x\tag{11.4.242}$$

electrons, where N is the number of atoms of atomic number Z'p er unit volume. The loss of energy per unit of distance is therefore:

$$\begin{aligned}-\frac{\mathrm{d}E}{\mathrm{d}x}&=\int_{r_{\text{min}}}^{r_{\text{max}}}\Delta E^e_c\cdot\frac{\mathrm{d}N}{\mathrm{d}x}=\int_{r_{\text{min}}}^{r_{\text{max}}}\frac{1}{(4\pi\varepsilon_0)^2}\cdot\frac{2\mathrm{Z}^2q_e^4}{r^2v_M^2m_e}2\pi\mathrm{Z}'Nr\mathrm{d}r=\frac{4\pi N\mathrm{Z}'}{(4\pi\varepsilon_0)^2}\cdot\frac{\mathrm{Z}^2q_e^4}{m_ev_M^2}\int_{r_{\text{min}}}^{r_{\text{max}}}\frac{1}{r}\mathrm{d}r\\&=\frac{4\pi N\mathrm{Z}'}{(4\pi\varepsilon_0)^2}\cdot\frac{\mathrm{Z}^2q_e^4}{m_ev_M^2}\ln\left(\frac{r_{\text{max}}}{r_{\text{min}}}\right)\end{aligned}\tag{11.4.243}$$

The value of r_{min} is evaluated by noting that this "impact parameter" should corresponds in practice to the maximum energy transfer. Using the equations we have proved previously:

$$_{\max}\Delta E_c^e=\frac{1}{(4\pi\varepsilon_0)^2}\frac{2\mathrm{Z}^2q_e^4}{r_{\text{min}}^2v_M^2m_e}=\frac{4m_e}{M}\cdot E_c^M \tag{11.4.244}$$

With $E_c^M=\dfrac{1}{2}Mv_m^2$, we can get the parameter r_{min} by:

$$r_{\text{min}}^2=\frac{1}{(4\pi\varepsilon_0)^2}\frac{2\mathrm{Z}^2q_e^4}{v_M^2m_e}\frac{1}{2m_ev_M^2}=\frac{1}{(4\pi\varepsilon_0)^2}\frac{\mathrm{Z}^2q_e^4}{v_M^4m_e^2} \tag{11.4.245}$$

and we get:

$$r_{\text{min}}=\frac{1}{4\pi\varepsilon_0}\cdot\frac{\mathrm{Z}q_e^2}{v_M^2m_e} \tag{11.4.246}$$

When r becomes very big, the energy transfer is smaller than the mean ionization energy denoted $\bar{I}$ of the electrons and the process is no longer efficient. We must therefore have the following relation:

$$_{\min}\Delta E_c^e=\frac{1}{(4\pi\varepsilon_0)^2}\cdot\frac{2\mathrm{Z}^2q_e^4}{r_{\text{max}}^2v_M^2m_e}>\bar{I} \tag{11.4.247}$$

We derive a value for r_{max}:

$$r_{\text{max}}<\frac{1}{4\pi\varepsilon_0}\cdot\frac{\mathrm{Z}q_e^2}{v_M}\sqrt{\frac{2}{m_e\bar{I}}} \tag{11.4.248}$$

By replacing the values of r_{min} and r_{max} of the above relations into the equation:

$$-\frac{\mathrm{d}E}{\mathrm{d}x}=\frac{4\pi N\mathrm{Z}'}{(4\pi\varepsilon_0)^2}\frac{\mathrm{Z}^2q_e^4}{m_ev_M^2}\ln\left(\frac{r_{\text{max}}}{r_{\text{min}}}\right) \tag{11.4.249}$$

we get:

$$-\frac{\mathrm{d}E}{\mathrm{d}x}=\frac{4\pi N\mathrm{Z}'}{(4\pi\varepsilon_0)^2}\frac{\mathrm{Z}^2q_e^4}{m_ev_M^2}\ln\left(\sqrt{\frac{2m_ev^2}{\bar{I}}}\right) \tag{11.4.250}$$

A more rigorous quantum treatment would show that the root of the argument of the logarithm should be suppressed by taking into account the relativistic effects as well as the intrinsic properties of the electron (fine-structure constant). We would then obtain the following Bethe formula:

$$\boxed{-\frac{\mathrm{d}E}{\mathrm{d}x}=\frac{4\pi N\mathrm{Z}'}{(4\pi\varepsilon_0)^2}\frac{\mathrm{Z}^2q_e^4}{m_ev_M^2}\left[\ln\left(\sqrt{\frac{2m_ev^2}{I}}\right)-\ln(1-\beta^2)-\beta^2-\frac{c_k}{\mathrm{Z}}\right]} \tag{11.4.251}$$

where $\beta=v/c$ and c_k is a correction term that depends on energy and of Z when we take into account the complete structure of the nucleus (layered model) of the material.

Finally, we see that the loss of linear energy is proportional to the atomic number of the incident radiation and the penetrated matter. Thus, to come back on radio-protection..., protections composed of materials with high atomic numbers (high density) will have a high decelerating power and will be advantageous in radiation protection.

11.4.3.2 Compton scattering

The Compton effect is observed when a photon is scattered inelastically by a charged particle. In fact, the photon is absorbed and then re-emitted by the particle, thus giving up part of its energy. It is this transfer of energy that justifies the inelastic character of the diffusion.

Thus if the charged particle is an electron, this effect can take place indifferently on an electron of any electronic layer or even on a free electron. The energy of the photon and that of the electron depend on the direction of emission of these particles. Since this effect depends on the number of electrons available per target atom, the Compton diffusion probability increases linearly with the atomic number Z of the absorbent. But since this effect is in competition with the production of an electron-positron pair that we shall see later, the Compton effect is especially important at average energies and at average atomic numbers.

Let us recall that in the section of Special Relativity we have proved that:

$$E=mc^2=E_c+m_0c^2\Rightarrow E_c=mc^2-m_0c^2 \tag{11.4.252}$$

And let us recall that we have thus for the momentum of a photon:

$$E=mc^2\Rightarrow\frac{E}{c}=mc=p \tag{11.4.253}$$

and we have also proved in the Special Relativity section that, starting from the total energy, the linear momentum is given by:

$$\begin{gathered}E=c\sqrt{m_0^2c^2+p^2}\Rightarrow p^2=\frac{E^2}{c^2}-m_0^2c^2\Rightarrow p=\frac{1}{c}\sqrt{E^2-m_0^2c^4}=\frac{1}{c}\sqrt{(E_c+m_0c^2)^2-m_0^2c^4}\\=\frac{1}{c}\sqrt{E_c^2+2E_cm_0c^2+m_0^2c^4-m_0^2c^4}=\frac{1}{c}\sqrt{E_c^2+2E_cm_0c^2}=\frac{1}{c}\sqrt{E_c(E_c+2m_0c^2)}\end{gathered} \tag{11.4.254}$$

Hence the relation, which we shall use further below:

$$p^2c^2=E_c(E_c+2m_0c^2) \tag{11.4.255}$$

Before the interaction, photon-electron, we have (we roughly consider the electron as being at rest) for the total energy of the system:

$$h\nu_1+m_0c^2 \tag{11.4.256}$$

and after collision:

$$h\nu_2+\sqrt{m_0^2c^4+p^2c^2} \tag{11.4.257}$$

Conservation of energy leads us to write:

$$h\nu_1+m_0c^2=h\nu_2+\sqrt{m_0^2c^4+p^2c^2} \tag{11.4.258}$$

By considering only the kinetic energies, we neglect that of the electron before the shock:

$$h\nu_1=E_e+h\nu_2=(mc^2-m_0c^2)+h\nu_2 \tag{11.4.259}$$

Either the figure below:

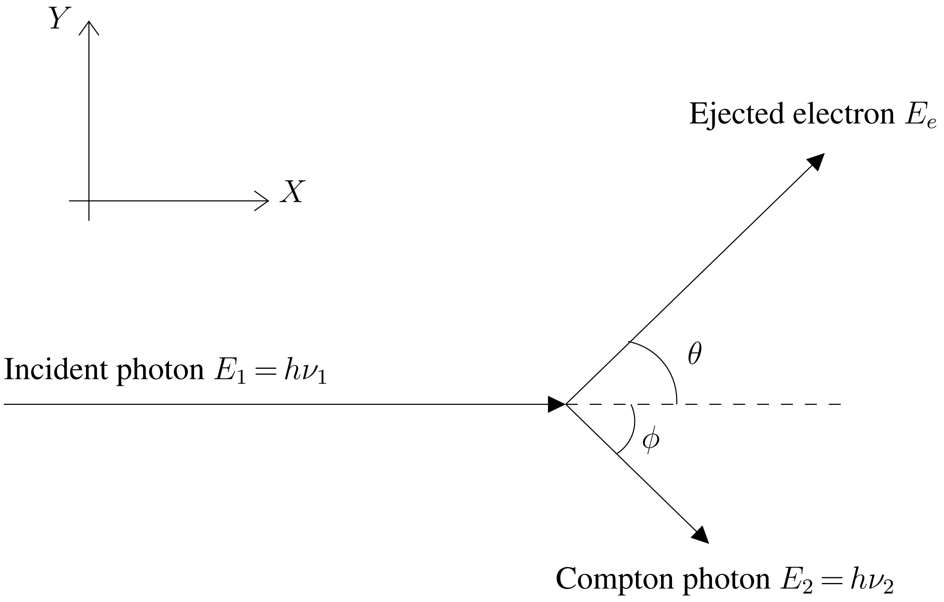


Figure 11.115: **Simplified illustration of the Compton effect**

The conservation of the linear momentum gives us according to the x-axis:

$$\frac{E_1}{c} = \frac{E_2}{c}\cos(\phi) + p_e\cos(\theta) \Leftrightarrow p_e\cos(\theta) = \frac{E_1}{c} - \frac{E_2}{c}\cos(\phi) \tag{11.4.260}$$

and according to the y-axis:

$$0 = \frac{E_2}{c}\sin(\phi) - p_e\sin(\theta) \Leftrightarrow p_e\sin(\theta) = \frac{E_2}{c}\sin(\phi) \tag{11.4.261}$$

The sum of these two relations squared gives us the total linear momentum:

$$\begin{gathered}p_e^2\cos^2(\theta) + p_e^2\sin^2(\theta) = \left(\frac{E_1}{c} - \frac{E_2}{c}\cos(\phi)\right)^2 + \left(\frac{E_2}{c}\sin(\phi)\right)^2\\ \Rightarrow p_e^2 = \frac{E_1^2}{c^2} - 2\frac{E_1E_2}{c^2}\cos(\phi) + \frac{E_2^2}{c^2}\cos^2(\phi) + \frac{E_2^2}{c^2}\sin^2(\phi) = \frac{E_1^2}{c^2} - 2\frac{E_1E_2}{c^2}\cos(\phi) + \frac{E_2^2}{c^2}\\ \Rightarrow p_e^2c^2 = E_1^2 + E_2^2 - 2E_1E_2\cos(\phi)\end{gathered} \tag{11.4.262}$$

Then by substituting $p_e^2c^2$:

$$E_{c,e}(E_{c,e} + 2m_0c^2) = E_1^2 + E_2^2 - 2E_1E_2\cos(\phi) \tag{11.4.263}$$

and as $E_{c,e} = E_1 - E_2$:

$$E_2 = \frac{E_1}{1 + \dfrac{E_1}{m_0c^2}(1-\cos(\phi))} \tag{11.4.264}$$

When the energy of the photon is high enough, that is $E_1 \gg m_0c^2$, that of the scattered photon tends to a limit given by (see l'Hôpital's rule in the section of Differential and Integral Calculus at page 1461):

$$\lim_{h\nu_1 \gg m_0c^2} h\nu_1 = \lim_{k\nu_1 \gg m_0c^2} E_2 = \frac{m_0c^2}{(1-\cos(\phi))} \tag{11.4.265}$$

The energy acquired by the Compton electron is finally equal to:

$$E_e = E_1 - E_2 = \frac{\dfrac{E_1^2}{m_0c^2}(1-\cos(\phi))}{1+\dfrac{E_1}{m_0c^2}(1-\cos(\phi))} \tag{11.4.266}$$

It is interesting to notice that we can not have $E_e = E_1$. Indeed this would suppose that:

$$E_2 = \frac{E_1}{1+\dfrac{E_1}{m_0c^2}(1-\cos(\phi))} = 0 \tag{11.4.267}$$

and we see well that in fact, whatever the value of ϕ, we always have $E_2 \neq 0$.

The frequency of the scattered photon is lower than that of the incident photon because its energy E_2 is always lower and hence its wavelength λ_2 larger. Therefore:

$$\Delta\lambda = \lambda_2 - \lambda_1 = c\left(\frac{1}{\nu_2}-\frac{1}{\nu_1}\right) = hc\left(\frac{1}{E_2}-\frac{1}{E_1}\right) = \frac{hc}{E_1}\left(\frac{E_1}{E_2}-1\right) \tag{11.4.268}$$

and since:

$$E_2 = \frac{E_1}{1+\dfrac{E_1}{m_0c^2}(1-\cos(\phi))} \tag{11.4.269}$$

We have:

$$\Delta\lambda = \frac{hc}{E_1}\left(1+\frac{E_1}{m_0c^2}(1-\cos(\phi))-1\right) = \frac{hc}{m_0c^2}(1-\cos(\phi)) \tag{11.4.270}$$

That is:

$$\boxed{\Delta\lambda = \frac{h}{m_0c}(1-\cos(\phi))} \tag{11.4.271}$$

which is also written using the definition of the Planck constant and the usual trigonometric relations:

$$\boxed{\Delta\lambda = \frac{4\pi\hbar}{mc}\sin^2\left(\frac{\phi}{2}\right)} \tag{11.4.272}$$

We name the factor $\dfrac{h}{m_0c^2}$ the "Compton wavelength" and its value is equal to:

$$\lambda_C = 2.426\cdot 10^{-12}\;[\text{m}] \tag{11.4.273}$$

Notice that if the angle ϕ is equal to zero, then the wavelength variation is zero and if the angle ϕ is equal to 180° (π [rad]) then the variation of the incident photon wavelength is twice the length of the Compton's wavelength!

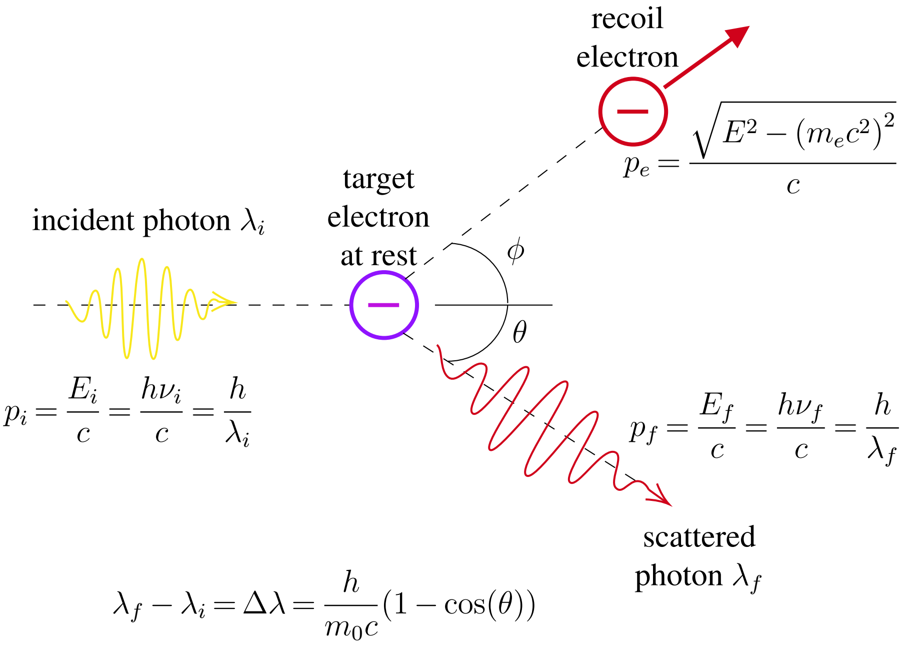

recoil
electron
$p_e = \dfrac{\sqrt{E^2 - (m_e c^2)^2}}{c}$
target
electron
at rest
incident photon λ_i
ϕ
θ
$p_i = \dfrac{E_i}{c} = \dfrac{h\nu_i}{c} = \dfrac{h}{\lambda_i}$
$p_f = \dfrac{E_f}{c} = \dfrac{h\nu_f}{c} = \dfrac{h}{\lambda_f}$
scattered
photon λ_f
$\lambda_f - \lambda_i = \Delta\lambda = \dfrac{h}{m_0 c}(1 - \cos(\theta))$

Figure 11.116: Compton scattering

Remark

One way to think of Compton scattering in the point of view of Quantum Field Theory is that the photon forms a virtual $e^- e^+$ pair, the virtual positron annihilates the original "real" electron giving off a photon and the "virtual" electron continues out of the Feynman diagram as the scattered "real" electron. As far as we know, only QFT permits us to calculate the cross differential section of the Compton scattering!

Let us prove now a very interesting case! A free electron absorb a photon!

This is because energy and momentum cannot be simultaneously conserved if a free electron were to absorb a photon. If the electron is bound to an atom then the atom itself is able to act as a third body repository of energy and momentum.

To prove that theoretically let us write down the conservation of momentum when a photon (ν) interacts with a free electron, assuming that it were absorbed:

$$p_1 + p_\nu = p_2 \tag{11.4.274}$$

where p_1 and p_2 are the momentum of the electron before and after the interaction. Conservation of energy gives us:

$$\sqrt{(p_1^2 c^2 + m_e^2 c^4)} + p_\nu c = \sqrt{(p_2^2 c^2 + m_e^2 c^4)} \tag{11.4.275}$$

Squaring the above equation and substituting for p_ν from equation we have:

$$p_1^2 c^2 + m_e^2 c^4 + 2\left(p_2 - p_1\right)\sqrt{(p_1^2 c^2 + m_e^2 c^4)} + \left(p_2 - p_1\right)^2 c^2 = p_2^2 c^2 + m_e^2 c^4 \tag{11.4.276}$$

After rearranging:

$$(p_2-p_1)^2c^2-\left(p_2^2-p_1^2\right)c^2+2\,(p_2-p_1)\,c\sqrt{\left(p_1^2c^2+m_e^2c^4\right)}=0 \tag{11.4.277}$$

Clearly $p_2-p_1=0$ is a solution to this equation, but cannot be possible if the photon has nonzero momentum. Dividing through by this solution we are left with:

$$\sqrt{\left(p_1^2c^2+m_e^2c^4\right)}-p_1c=0 \tag{11.4.278}$$

This solution is also impossible if the electron has non-zero rest mass (which it does). We conclude therefore that a free electron cannot absorb a photon because energy and momentum cannot simultaneously be conserved.

The above proof assumes a linear interaction. In general $\vec{p}_\nu,\vec{p}_1$ and $\vec{p}_2$ would not be aligned. However we can always transform to a frame of reference where the electron is initially at rest so that $\vec{p}_1=0$ and then the momentum of the photon and the momentum of the electron after the interaction would have to be equal. This then leads to the nonsensical result that either $p_2=0$ or $m_ec^2=0$.

11.4.3.2.1 Thomson scattering

Thomson scattering is the elastic scattering of electromagnetic radiation by a free charged particle, as described by classical electromagnetism. It is just the low-energy limit of Compton scattering: the particle kinetic energy and photon frequency do not change as a result of the scattering. This limit is valid as long as the photon energy is much less than the mass energy of the particle: $h\nu\ll m_ec^2$, or equivalently, if the wavelength of the light is much greater than the Compton wavelength of the particle.

First consider scattering of (classical) electromagnetic waves by free electrons, named "Thomson scattering". The electron experiences a force due to the incident electric field:

$$E=\hat{E}e^{\mathrm{i}(\vec{k}\circ\vec{r}-\omega t)} \tag{11.4.279}$$

The equation of motion for the charged particle is:

$$F=ma=m\ddot{r}=-eE \tag{11.4.280}$$

We recall that in the dipole approximation (valid since the wavelength of light is large compared to the scatterer) the power emitted per unit solid angle is (see section Electrodynamics page 4163):

$$\dfrac{\mathrm{d}P}{\mathrm{d}\Omega}=\dfrac{e^2}{16\pi^2\varepsilon_0c^3}\langle a^2\rangle\sin^2(\theta) \tag{11.4.281}$$

where $\langle a^2\rangle$ represents the "time-averaged squared acceleration". The equation of motion gives the acceleration directly and we assume that it is equal to the time averaged squared acceleration such that:

$$a^2=\dfrac{e^2}{2m^2}\hat{E}^2:=\langle a^2\rangle \tag{11.4.282}$$

This yields, upon rearranging:

$$\frac{\mathrm{d}P}{\mathrm{d}\Omega}=\left(\frac{e^2}{4\pi\varepsilon_0 mc^2}\right)^2\frac{\varepsilon_0 c\hat{E}^2}{2}\sin^2(\theta) \tag{11.4.283}$$

The term in the brackets is known as the "classical electron radius":

$$r_c=\frac{e^2}{4\pi\varepsilon_0 mc^2}=2.8\cdot 10^{-15}\,[\mathrm{m}] \tag{11.4.284}$$

Now recall that the time-averaged Poynting vector which is the incident energy per unit area per unit time is (see section Electrodynamics page 4156):

$$\bar{S}=\frac{\hat{E}^2}{2\mu_0 c}=\frac{c\varepsilon_0\hat{E}^2}{2} \tag{11.4.285}$$

We can astutely fall back on the "differential scattering cross section" define earlier above by using dimensional analysis:

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega}=\frac{\frac{\mathrm{d}P}{\mathrm{d}\Omega}}{\bar{S}} \tag{11.4.286}$$

and is the area of the wavefront which delivers the same power as is scattered into a given solid angle $\mathrm{d}\Omega$. So we get after simplification:

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega}=\left(\frac{e^2}{4\pi\varepsilon_0 mc^2}\right)^2\sin^2(\theta)=r_c^2\sin^2(\theta) \tag{11.4.287}$$

The total cross section is obtained by integrating over $\mathrm{d}\Omega$:

$$\sigma=r_c^2\int\sin^2(\theta)\mathrm{d}\Omega \tag{11.4.288}$$

But from the section Trigonometry we know that:

$$\mathrm{d}\Omega=\sin(\theta)\mathrm{d}\theta\mathrm{d}\phi \tag{11.4.289}$$

Therefore (identically as the integral we did in the Trigonometry section):

$$\begin{aligned}\sigma&=r_c\int\limits_0^{2\pi}\mathrm{d}\phi\int\limits_0^{\pi}\sin^3(\theta)\mathrm{d}\theta=2\pi r_c^2\int\limits_0^{\pi}\sin^3(\theta)\mathrm{d}\theta\\&=2\pi r_c^2\left(-\frac{1}{3}\cos(x)\sin^2(x)-\frac{2}{3}\cos(x)\right)\Bigg|_0^{\pi}=2\pi r_c^2\left[-\frac{2}{3}\cdot(-1)-\left(-\frac{2}{3}\cdot(1)\right)\right]\\&=2\pi r_c^2\frac{4}{3}\end{aligned} \tag{11.4.290}$$

Therefore:

$$\boxed{\sigma_T=\frac{8\pi}{3}r_c^2=\frac{8\pi}{3}\left(\frac{e^2}{4\pi\varepsilon_0 mc^2}\right)^2\cong 6.652\cdot 10^{-29}\,[\mathrm{m}^2]} \tag{11.4.291}$$

This is known as the "classical Thomson cross section" or also "electron Thomson cross section".

Therefore the fraction of photons scattered at distance r is the obviously given by (this term is dimensionless!):

$$f_T = \frac{\sigma_T}{4\pi r^2} \tag{11.4.292}$$

The classical Thomson scattering formula is known to be in disagreement with experiment by the early 11920s (holocene calendar). The Klein-Nishina formula based on Quantum Electrodynamics (QED) agreed perfectly with the available experiments in the late 11920s (holocene calendar). However at ultra relativistic energies it looked wrong. It was not until many years later the difference was shown to be due to the production of electron/positron pairs, with the positron annihilating into some other electron, and to Bremsstrahlung.

Remark

Thomson scattering is important on very different scales in Astronomy. For example: At high temperatures, Thomson scattering can dominate the photon diffusion in a star and thus determines the energy flux making its way out of the star. At the epoch of the Universe recombination ($z \cong 1,100$), Thomson scattering caused the high opacity of the plasma. The effects of Thomson scattering are, for example, visible in the diffusion damping of Baryon Acoustic Oscillations. For high l-moments ($l \geq 1,000$), Thomson scattering set the length over which photons could diffuse. This directly determines the damping of small scales (high moments) as inhomogeneities can be levelled out by photons.

11.4.3.3 Rayleigh scattering

"Rayleigh scattering" is the (dominantly) elastic scattering of light or other electromagnetic radiation by particles much smaller than the wavelength of the radiation. Rayleigh scattering does not change the state of material and is, hence, a parametric process. The particles may be individual atoms or molecules. It can occur when light travels through transparent solids and liquids, but is most prominently seen in gases. Rayleigh scattering results from the electric polarizability of the particles. The oscillating electric field of a light wave acts on the charges within a particle, causing them to move at the same frequency. The particle therefore becomes a small radiating dipole whose radiation we see as scattered light.

The strong wavelength dependence of the scattering (λ^{-4}) means that shorter (blue) wavelengths are scattered more strongly than longer (red) wavelengths. This results in the indirect blue light coming from all regions of the sky. Rayleigh scattering is a good approximation of the manner in which light scattering occurs within various media for which scattering particles have a small size (parameter).

Now consider scattering from a harmonically bound electron excited by a harmonic force. In the simplest approximation it is described by the equation of motion given below (see section Music Mathematics page 8224):

$$\ddot{r} + \gamma\dot{r} + \omega_0^2 r = -\frac{q_e}{m_e}E \tag{11.4.293}$$

But in the case of Rayleigh scattering, physicist like to change a sign in front of one of the constant (it permits some simplifications later below):

$$\ddot{r}-\gamma\dot{r}+\omega_0^2 r=-\frac{q_e}{m_e}E \tag{11.4.294}$$

Keep in mind that an accelerating electron will radiate causing the oscillator to lose energy as a damped oscillator where γ is the damping rate due to radiation losses.

Remark

A simple model for the dielectric properties of a material is obtained by considering the motion of a bound electron in the presence of an applied electric field. As the electric field tries to separate the electron from the positively charged nucleus, it generates an electric dipole moment. Averaging this dipole moment over the volume of the material gives rise to a macroscopic dipole moment per unit volume. The model is described exactly by the same differential equation and solutions as we will see at page 4240.

The limit $\omega_0=0$ corresponds to unbound electrons and describes the case of good conductors. The frictional term $y\dot{x}$ arises from collisions that tend to slow down the electron. The parameter y is a measure of the rate of collisions per unit time, and therefore, $\tau=1/\gamma$ will represent the mean-time between collisions.

In a typical conductor, τ is of the order of 10^{-14} seconds, for example, for copper, $\tau=2.4\cdot10^{-14}\,[\text{s}]$ and $y=4.1\cdot10^{13}\,[\text{s}^{-1}]$. The case of a tenuous, collisionless, plasma can be obtained in the limit $\gamma=0$. Thus, the above simple model can describe the following cases:

a. Dielectrics, $\omega_0\neq0,\gamma\neq0$

b. Conductors, $\omega_0=0,\gamma\neq0$

c. Collisionless Plasmas, $\omega_0=0,\gamma=0$

The basic idea of this model is that the applied electric field tends to separate positive from negative charges, thus, creating an electric dipole moment. In this sense, the model contains the basic features of other types of polarization in materials, such as Ionic/molecular polarization arising from the separation of positive and negative ions by the applied field, or polar materials that have a permanent dipole moment.

We will try the complex function, $r(t)=Ae^{\mathrm{i}(\omega t+\varphi)}$, with A a real number, $r(t)$ cycling at the same frequency as the driving force. We can always take the amplitude A to be real: that is not a restriction, since we have added the adjustable phase factor $e^{\mathrm{i}\varphi}$. Physically, this factor allows the solution to differ from the driver in phase, which it usually does. If we succeed in finding an $r(t)$ that satisfies the equation, the real parts of the two sides of the equation must be equal.

Remark

If $r(t)=Ae^{\mathrm{i}(\omega t+\varphi)}$ is a solution to the equation with the complex driving force, $F_0e^{\mathrm{i}\omega t}$, its real part (and where $F_0=-q_e/mE_0$), $A\cos(\omega t+\varphi)$, will be a solution to the equation with the real physical driving force, $F_0\cos(\omega t)$.

It's very easy to check that $r(t)=Ae^{\mathrm{i}(\omega t+\varphi)}$ is a solution to the equation, with the right A and φ. Just plug it in and see what happens. The differentiations are simple, giving:

$$-\omega^2 Ae^{\mathrm{i}(\omega t+\varphi)}-\mathrm{i}\gamma\omega Ae^{\mathrm{i}(\omega t+\varphi)}+\omega_0 Ae^{\mathrm{i}(\omega t+\varphi)}=F_0e^{\mathrm{i}\omega t} \tag{11.4.295}$$

To nail down A and φ, we begin by cancelling out the common factor $e^{\mathrm{i}\omega t}$, then shift the $e^{\mathrm{i}\varphi}$ to the other side to find:

$$A=\frac{F_0e^{-\mathrm{i}\varphi}}{\omega_0^2-\omega^2-\mathrm{i}\gamma\omega} \tag{11.4.296}$$

Therefore:

$$r=\frac{-q_e/m_e}{\omega_0^2-\mathrm{i}\gamma\omega-\omega^2}E \tag{11.4.297}$$

or in vector form:

$$\vec{r}=\frac{-q_e/m_e}{\omega_0^2-\mathrm{i}\gamma\omega-\omega^2}\vec{E}=\frac{-q_e/m_e}{(\omega_0^2-\omega^2)-\mathrm{i}\gamma\omega}\vec{E} \tag{11.4.298}$$

Let us recall, exactly as we did just earlier for the Thomson scattering, that in the dipole approximation (valid since the wavelength of light is large compared to the scatterer) the power emitted per unit solid angle is (see section Electrodynamics page 4163):

$$\frac{\mathrm{d}P}{\mathrm{d}\Omega}=\frac{q_e^2}{16\pi^2\varepsilon_0c^3}\langle a^2\rangle\sin^2(\theta) \tag{11.4.299}$$

where $\langle a^2\rangle$ represents the "time-averaged squared acceleration".

Before calculating the time-average acceleration, let us first take the module of the complex number. For this remember that:

$$\left|\frac{1}{z}\right|=\frac{1}{|z|}=\frac{1}{|a+\mathrm{i}b|}=\frac{1}{\sqrt{a^2+b^2}}\quad\text{with}\quad z\neq 0 \tag{11.4.300}$$

Therefore:

$$|\vec{r}|=\left|\frac{-q_e/m_e}{(\omega_0^2-\omega^2)-\mathrm{i}\gamma\omega}\vec{E}\right|=\frac{-q_e/m_e}{\sqrt{(\omega_0^2-\omega^2)^2+\gamma^2\omega^2}}\vec{E}=\frac{-q_e/m_e}{\sqrt{(\omega^2-\omega_0^2)^2+\gamma^2\omega^2}}\vec{E} \tag{11.4.301}$$

Now we can calculate the time-averaged squared acceleration is[37]:

$$\begin{aligned}\langle|a|^2\rangle=\langle|\ddot{\vec{r}}|\cdot|\ddot{\vec{r}}|\rangle&=\left[\sqrt{\frac{1}{T}\int\limits_0^T\left(\frac{-\omega^2q_e/m_e}{\sqrt{(\omega^2-\omega_0^2)^2+\gamma^2\omega^2}}\vec{E}\right)^2\mathrm{d}t}\,\right]^2=\left[\frac{1}{\sqrt{2}}\left(\frac{\omega^2q_e/m_e}{\sqrt{(\omega^2-\omega_0^2)^2+\gamma^2\omega^2}}\right)\right]^2\\&=\frac{q_e^2}{2m_e^2}\left(\frac{\omega^4}{(\omega^2-\omega_0^2)^2+\gamma^2\omega^2}\right)\end{aligned} \tag{11.4.302}$$

---

[37] To see where the factor $1/2$ - i.e. the square of $1/\sqrt{2}$ - comes from, see our study of the average power at...

Thus the scattering cross-section known as the "Rayleigh scattering cross section" is given by:

$$\sigma_R = \frac{q_e^2}{2m_e^2}\left(\frac{\omega^4}{(\omega^2-\omega_0^2)^2+\gamma^2\omega^2}\right) \quad (11.4.303)$$

For strong binding (i.e., $\omega_0 \gg \omega$), we have:

$$\boxed{\sigma_R \cong \frac{q_e^2}{2m_e^2}\left(\frac{\omega}{\omega_0}\right)^4 = \frac{q_e^2}{2m_e^2}\left(\frac{\lambda_0}{\lambda}\right)^4 \propto \lambda^{-4}} \quad (11.4.304)$$

So the total scattering cross section is given by the inverse fourth power of the wavelength and this was derived by Rayleigh.

This strong wavelength dependence makes short wavelength light (blues and violets) scatter more than long wavelength light (reds and yellows). This effect explains why the sky is blue. Short wavelength or blue light scatters the most.

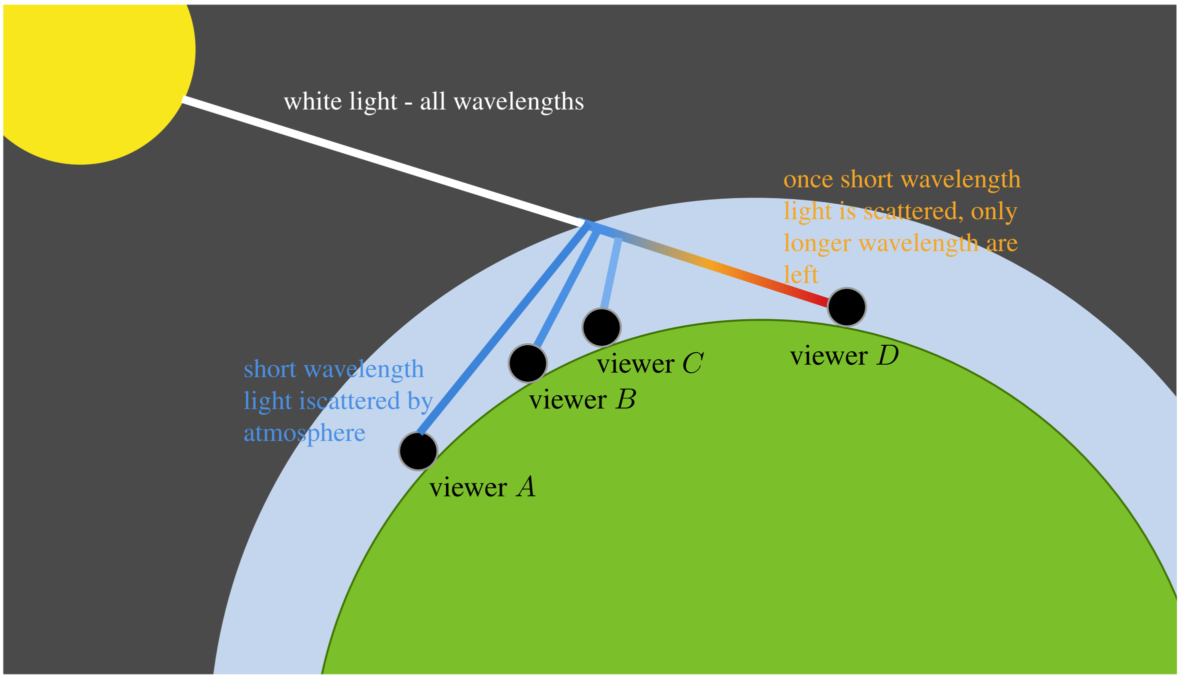

white light - all wavelengths
once short wavelength light is scattered, only longer wavelength are left
short wavelength light iscattered by atmosphere
viewer A
viewer B
viewer C
viewer D

Figure 11.117: A single ray of light from the Sun and how it might be scattered and seen as different colours to viewers at different positions on Earth (source: Ask a scientist)

This formula is valid for scattering of light by air molecules (N_2, O_2) and shows that air molecules scatter the shorter wavelengths (blue) preferentially thus leading to blueness of the sky.

Rayleigh scattering is scattering from polarizable entities. The incident light induces a dipole moment, which re-radiates. Thomson scattering is scattering from free unbound charged "unpolarizable" particles.

From the perspective of radiation oncology (i.e. higher energy) physics, here is the main difference, in a few words.

- Rayleigh scattering is the elastic scattering of a photon off an entire atom.

- Thomson scattering is the elastic scattering of a photon off a single unbound electron.

Both are more likely to happen at lower energies. In particular, Rayleigh scattering cross section depends on the inverse square of initial photon energy. Thomson scattering cross section has a more complicated energy dependence, with many terms (find it in many textbooks), generally stronger than inverse square.

According to Rayleigh scattering, we should observe a violet sky. In order to understand why we perceive the sky to be blue, we should take into consideration two other phenomena:

- The solar radiation spectrum has its peak in the blue wavelengths and it drops for the violet wavelengths.
- In addition, our eyes are composed of three types of cones detecting red, green and blue colors. Thus, if we look at the sensitivity of the blue cones as a function of the wavelength, we will see that this sensitivity is much higher for the blue color than for violet.

Rayleigh scattering also explains why the Moon turns red during a total lunar eclipse because of refraction by the Earth's atmosphere. Bluish components of sunlight is scatterred more, and more of its reddish components reaches the Moon's surface.

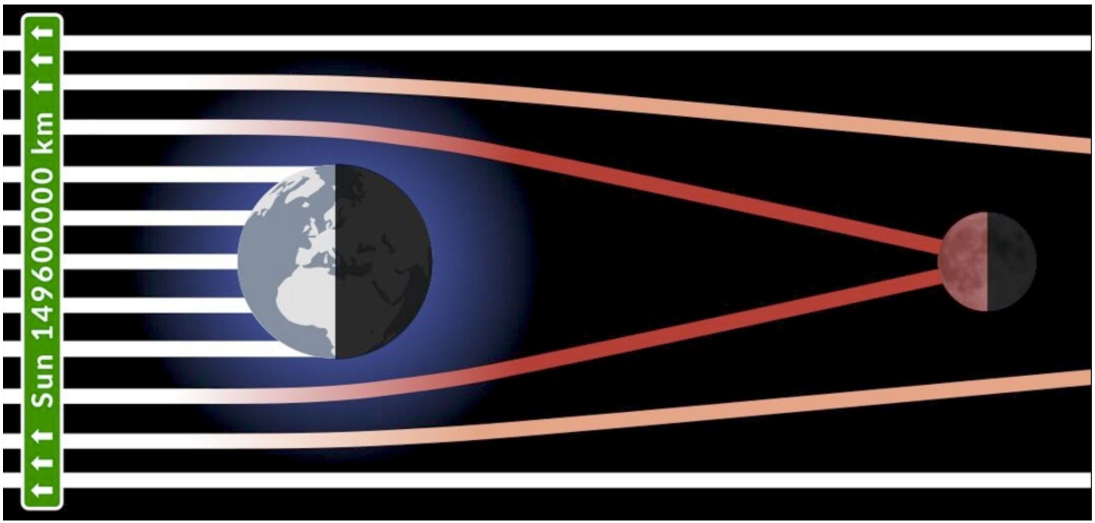


Figure 11.118: **Rayleigh scattering and Moon color during lunar eclipse (source: FlatEarth.ws/blood-moon)**

11.4.3.4 Photoelectric effect

The photoelectric effect is the ejection of electrons (the named "photoelectrons") from the surface of various metals exposed to radiation energy. This radiation can come from the rearrangement of the nucleus of the atom as well as from an external radiation.

Moreover, it was by quantitative measurements of the photoelectric effect that Albert Einstein proposed to test the quantum theory of light (Wave Quantum theory) and therefore the theoretical explanation that bring him to receive Nobel prize of Physics.

Let us first describe the experiment: the emission of electrons by a metal does not contradict the electromagnetic theory of light. If we consider a uniform beam, its energy is uniformly distributed over the entire wave front. The more intense the light is, the greater the amplitudes of the electric and magnetic fields at each point of the wavefront and the greater the energy per second transmitted by the wave is. These fields exert forces on the electrons in the metal and can even tear them from its surface.

Here is the experience set up:

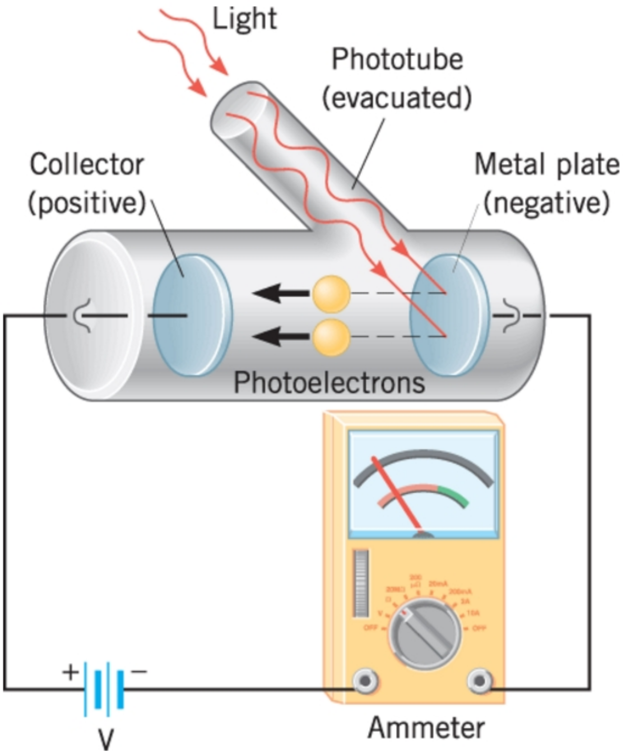


Figure 11.119: **Experiment for measuring the photoelectric effect (source: ?)**

If the collecting anode is at a positive potential relative to the emitting cathode, the photoelectrons traverse the tube and constitute the current measured by the ammeter. We then observe experimentally a proportionality between the intensity of the incident beam and the current. However, at least three problems persist between the theoretical model and the experimental observation:

1. The wave aspect of light is not suitable for explaining the time required to absorb the extraction energy.

   Indeed, let us consider a 100 [W] lamp with a 15% luminous efficiency placed at 0.5 [m] of a potassium-coated plate (chemical symbol of potassium is K) have a minimum extraction energy of 2.25 [eV] assuming a diameter of $1\cdot 10^{-10}$ [m] for the Potassium atom.

   We then have:

$$I=\dfrac{P\cdot\eta}{S_{\text{radiation}}}=\dfrac{100\cdot 0.15}{4\pi R^2}=4.77\;\left[\mathrm{W}\cdot\mathrm{m}^{-2}\right] \qquad (11.4.305)$$

The luminous power absorbed by the atom by its half-surface which faces the radiation is then:

$$P_K = I \cdot \frac{S_K}{2} = 4.77 \cdot 2\pi \frac{4\pi R_K^2}{2} = 4.77 \cdot 2\pi \left(0.5 \cdot 10^{-10}\right)^2 \cong 7.49 \cdot 10^{-20}\,[\text{W}] \qquad (11.4.306)$$

The time required for absorption is then:

$$t = \frac{E_l}{P_K} = \frac{2.25 \cdot 1.602 \cdot 10^{-19}}{7.49 \cdot 10^{-20}} \cong 4.8\,[s] \qquad (11.4.307)$$

Which is in contradiction with the experience where we observe that the phenomenon is almost instantaneous (time for light to travel to the metal).

2. If we reverse the poles, the electrons emitted by the metal are pushed back by the negative electrode, but if the reverse voltage is not to big the faster ones can still reach the negative electrode and a current will occur. At a negative potential, specific for each metal, named the equation "stop potential U_a", all the emitted electrons are pushed back and the current is zero. The maximum kinetic energy of these photoelectrons is then:

$$E_{c,\text{max}} = q_e U_q \qquad (11.4.308)$$

However, we see experimentally that this stopping potential is independent of the intensity of the radiation. In the Wave theory, the increase in intensity should increase the number of electrons extracted (regardless of their energy level) and their maximum kinetic energy. A greater intensity implies a greater amplitude of the electric field: $I \propto E^2$. Thus, a bigger electric field should eject the electrons at higher velocities from all layers of the electronic cloud as the intensity increases.

3. When we vary the frequency v of incident light and we measure U_a, we observe that the photoelectric effect does not occur if $\nu < \nu_0$ (ν_0 is named the "frequency threshold") and this whatever the intensity of the light. What is rather annoying... because in Wave theory, we must always be able to eject electrons whatever the frequency, we just have to increase the intensity.

Each problem can be solved by adopting the following point of view:

1. In the Wave theory approach, the source is seen as propagating like a spherical wavefront whose surface density of energy decreases as a $1/r^2$ (see section Electrodynamics page 4160). In order to explain the experimental observation, we must see the experiment from a corpuscular point of view where the wavefront is a front of corpuscles whose surface density of photons decreases in $1/r^2$ but where the energy of each photon remains $h\nu$ (according to the Planck-Einstein relation).

2. If we think in terms of photons, when we increase the intensity, we increase the number of photons, but the energy per photon $h\nu$ remains unchanged. Thus, the energy $E_{c,\text{max}}$ that each photon can have does not change. Hence the stopping potential is independent of the field intensity.

3. If we think in terms of photons again, the electrons in the target are retained by the attraction forces, extracting an electron from the surface requires a minimal energy E_l

that depends on each material (E_l is also named "extraction work" denoted W_i which is of the order of some electronvolts). If the energy of the incident photon $E_i = h\nu$ is greater than E_l, an electron can be extracted, but if it is inferior, no electron can be extracted. The energy input $h\nu$ is equal to the kinetic output energy of the electron plus the energy required to extract it from the metal, therefore:

$$E_c + E_l = E_i = h\nu \qquad (11.4.309)$$

Thus, if the frequency of light is increased, the maximum kinetic energy of the electrons increases linearly. R.A. Millikan made rigorous experiments between 11913 (holocene calendar) and 11914 (holocene calendar), the results were matching perfectly with Albert Einstein's theory.

Caution!!!! The photoelectric effect can however be explained without photons. It is an historical accident that the photon concept should have acquired its strongest early support from Einstein's considerations on the photoelectric effect! Indeed, we can understand the photoeffect as being the result of a classical field falling on a quantized atomic electron (see page **??**). The introduction of the photon concept is neither logically implied by nor necessary for the explanation of the photoelectric effect (it's just more convenient in most graduate curriculum because of a lack of time!).

The light spreads from place to place as if it were a wave. But light interacts with matter in absorption and emission processes as if it were a particle stream. We already know that this is the "wave-particle duality" (see section Wave Quantum Physics page 4688). Thus, as that latter can find in the mass particles as suggested by the de Broglie hypothesis that we have seen in the section of Wave Quantum Physics, is finally verified also for the light.

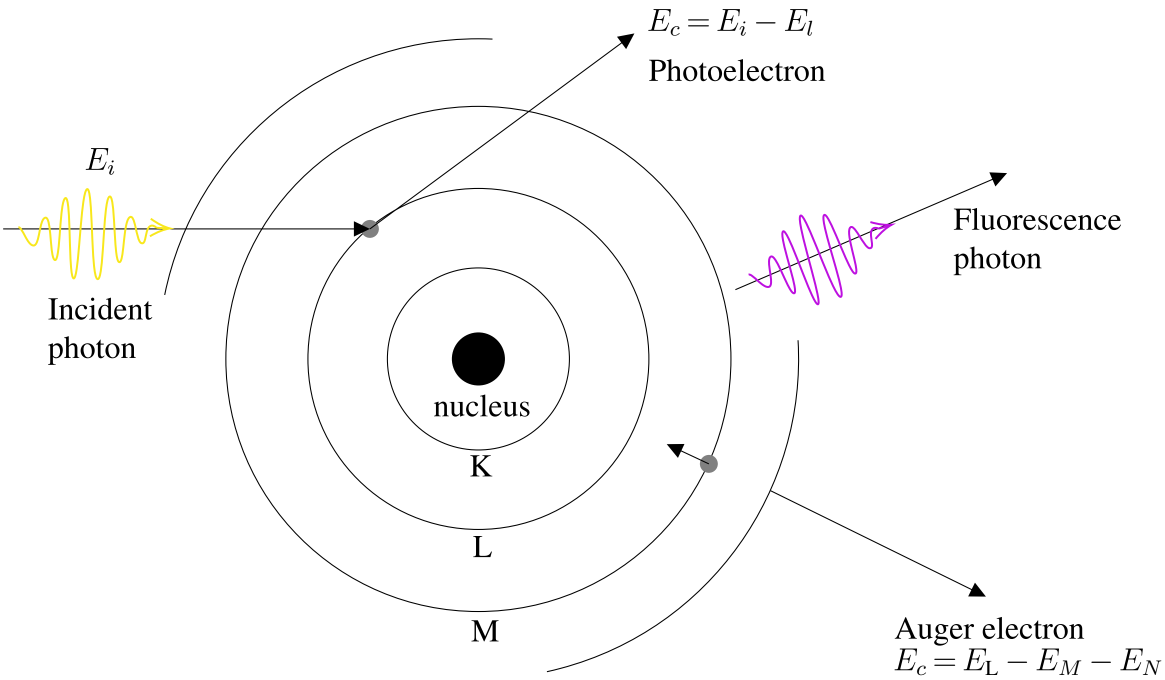


Figure 11.120: **Principle of the photoelectric effect on the Bohr model of the atom**

A photon of incident energy E_i which interacts with an electron of a target atom can eject this electron from its orbit by communicating to it a kinetic energy equation:

$$E_c = E_i - E_l \tag{11.4.310}$$

where E_l is the binding energy of the electron ejected from its orbit (this relation is given in the form $T = E - W_i$ in the figure above).

If the energy of the incident photon is lower than the binding energy of the electron of layer K (see section Corpuscular Quantum Physics page 4616), the photoelectric effect is done with an electron of the layer L, etc.

In the case where the radiation is absorbed, the atom is said to be "excited", because its energy state is not the minimum state. There follows a "relaxation" (or "de-excitation"): an electron of a higher layer fills the quantum space left vacant by the ejected electron.

If the transition energy is moderate (i.e., if the incident radiation has a moderate energy), relaxation causes the emission of a low-energy (visible or ultra-violet) photon, this is the phenomenon of fluorescence.

11.4.3.5 Rutherford scattering

Let us now consider the diffusion that a charged particle undergoes when it is subjected to an electrostatic repulsive force inversely proportional to the square of the distance between the moving particle and a fixed point or center of force. This problem is particularly interesting because of its application to atomic and nuclear physics. For example, when a proton, accelerated by a machine such as a cyclotron, passes near a core of the target's material, it is deflected by a force of this type from electrostatic repulsion of the nucleus. This study is referred to as the "Rutherford scattering" or "Coulomb scattering" and is mainly based on the following figure:

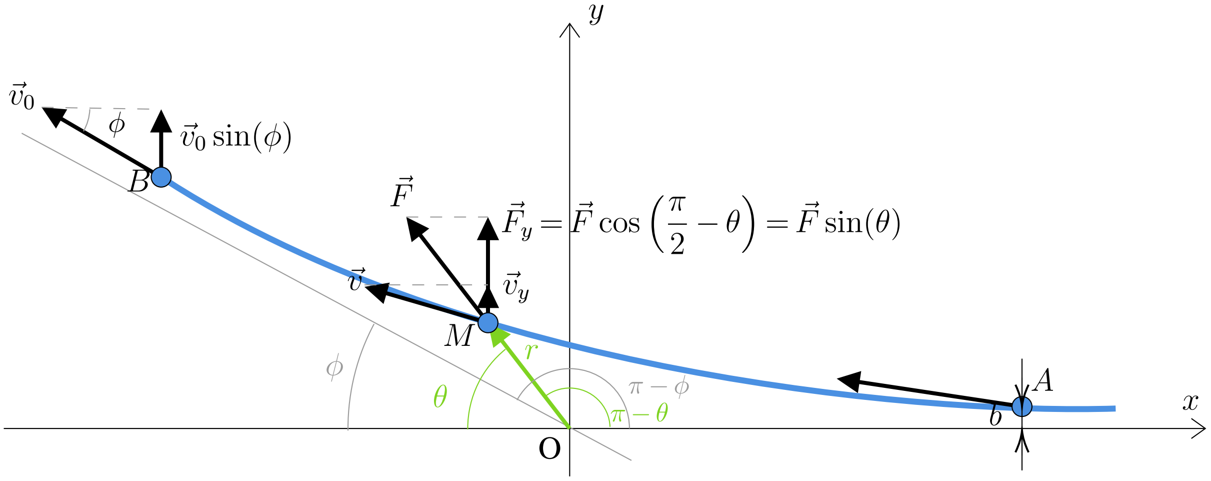


Figure 11.121: **Rutherford scattering**

Let O be a center of force and a particle thrown against O from a great distance with the velocity v_0 (see figure above) moving from right to left and entering the figure above at point A. We will choose The axis of X passing through O and parallel to v_0. The distance b, named the "impact parameter", is the distance between the x-axis of the abscissa and the point A. Assuming that the force between A and O is repulsive and central, the particle will follow the curved trajectory AMB. The shape of the curve depends on how the force varies with the distance. If the force is inversely proportional to the square of the distance, that is to say if:

$$F=\frac{K_e}{r^2}\qquad(11.4.311)$$

The trajectory is a hyperbola (indeed, the equation and solution are the same as Kepler problem, we just need to flip the sign of the constant K_e). Of course (see section Electrostatics page 3968):

$$K_e=\frac{1}{4\pi\varepsilon_0}ZqQ\qquad(11.4.312)$$

When the particle is in A its angular momentum is mv_0b. In any position such that M, its angular momentum (see section Classical Mechanics page 3244), is also given by $mr^2\dfrac{\mathrm{d}\theta}{\mathrm{d}t}$. Since the angular momentum must remain constant since the force is central:

$$mr^2\frac{\mathrm{d}\theta}{\mathrm{d}t}=mv_0b\qquad(11.4.313)$$

The equation of motion in the OY direction is obtained by injecting this relation into:

$$m\frac{\mathrm{d}v_y}{\mathrm{d}t}=F_y=F\sin(\theta)=\frac{K_e}{r^2}\sin\theta \tag{11.4.314}$$

so we can get rid of r^2 and write therefore:

$$m\frac{\mathrm{d}v_y}{\mathrm{d}t}=\frac{K_e}{v_0 b}\sin(\theta)\frac{\mathrm{d}\theta}{\mathrm{d}t} \tag{11.4.315}$$

To find the deviation of the particle, we must integrate this equation from one end of the trajectory to the other. In A the value of v_y is zero because the initial motion is parallel to the X axis and we also have $\theta=0$. In B we have $v_y=v_0\sin(\phi)$ and $\theta=\phi$ or $\pi-\theta=\pi-\phi$. We notice that in B the speed is again v_0 because, by symmetry, the speed lost when the particle approaches O must be regained when it moves away from it. Therefore:

$$\int\limits_0^{v_0\sin(\phi)}\mathrm{d}v_y=\frac{k}{mv_0b}\int\limits_0^{\pi-\phi}\sin(\theta)\mathrm{d}\theta \tag{11.4.316}$$

Which give:

$$v_0\sin(\phi)=-\frac{K_e}{mv_0b}\left(\cos(\pi-\phi)-1\right) \tag{11.4.317}$$

thus by using the common trigonometric relations:

$$v_0\sin\phi=\frac{K_e}{mv_0b}\left(1+\cos(\phi)\right) \tag{11.4.318}$$

and after rearrangement:

$$\frac{(1+\cos(\phi))}{\sin\phi}=\frac{mv_0^2b}{K_e} \tag{11.4.319}$$

Let us recall that (see section Trigonometry page 2519) that:

$$\tan(\alpha)=\frac{\sin(2\alpha)}{1+\cos(2\alpha)}=\frac{1}{\cot(\alpha)} \tag{11.4.320}$$

What then gives us:

$$\cot\left(\frac{1}{2}\phi\right)=\frac{mv_0^2b}{K_e} \tag{11.4.321}$$

Either more explicitly:

$$\cot\left(\frac{\phi}{2}\right)=\frac{4\pi\varepsilon_0mv_0^2b}{ZQq} \tag{11.4.322}$$

This relation gives the angle of deviation ϕ as a function of the impact parameter b:

$$b=\frac{1}{4\pi\varepsilon_0}\frac{ZQq}{mv_0^2}\cot\left(\frac{\phi}{2}\right) \tag{11.4.323}$$

Of course, in undergraduate cases, $Q=q$ is often put, which simplifies the previous relation slightly, but we then loses in generalization.

This equation is applied to the analysis of the deviation of an electrically charged particle by a nucleus. Let us notice that this result is only valid for a force inversely proportional to the square of the distance. If the force depends on the distance according to another law, the deflection angle satisfies another equation. The deviation experiments are therefore also very useful when we want to determine the law of force involved in the interactions between particles!

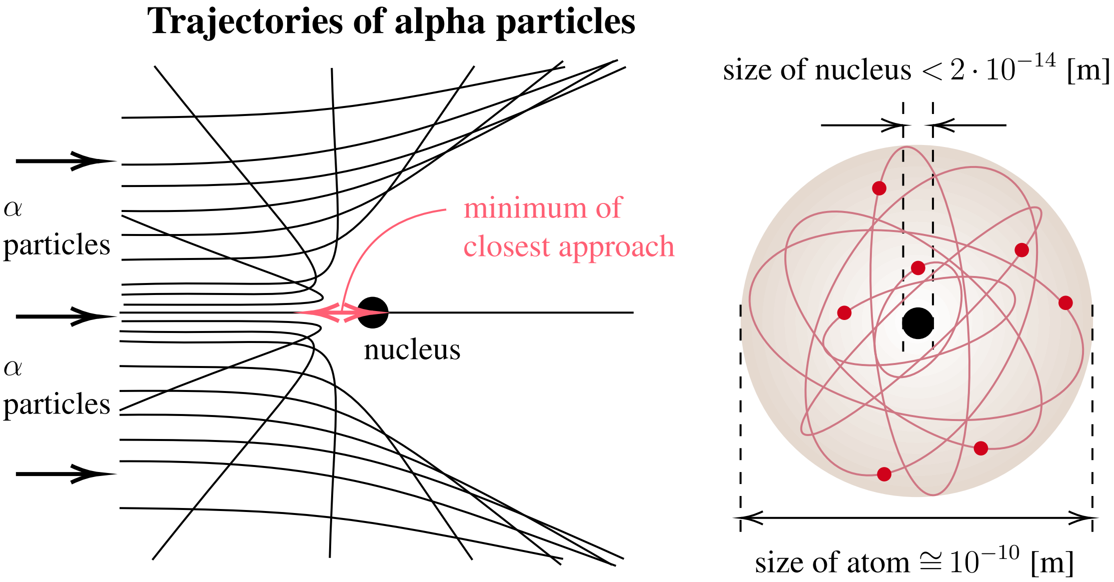

Trajectories of alpha particles
α particles
α particles
minimum of closest approach
nucleus
size of nucleus $< 2\cdot 10^{-14}$ [m]
size of atom $\cong 10^{-10}$ [m]

Figure 11.122: **Representation of the Rutherford diffusion of alpha nucleus**

In nuclear physics laboratories, diffusion experiments are carried out by accelerating electrons, protons or other particles by means of a cyclotron, a Van de Graaf accelerator or some other similar device, and by observing the angular distribution of the deflected particles.

It is clear that an incident particle in a surface defined by a radius between b and $b+\mathrm{d}b$ will be respectively included in the solid diffusion angle:

$$\mathrm{d}\sigma = 2\pi b\mathrm{d}b \qquad (11.4.324)$$

with (see section Trigonometry page 2549):

$$\mathrm{d}\Omega = 2\pi\sin(\phi)\mathrm{d}\phi \qquad (11.4.325)$$

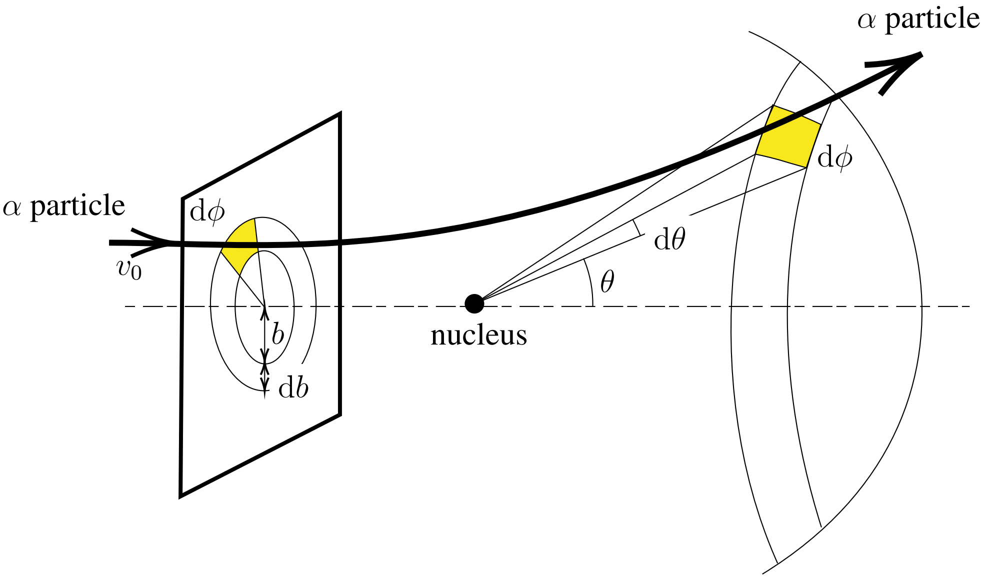


Figure 11.123: **Representation of the solid diffusion angle**

The "cross-section" or "differential cross-section" being defined in Nuclear Physics and in Particle Physics by:

$$\boxed{\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega}=\frac{2\pi b\mathrm{d}b}{2\pi\sin(\phi)\mathrm{d}\phi}} \tag{11.4.326}$$

Starting from (relation proved just earlier above):

$$b=\frac{1}{4\pi\varepsilon_0}\frac{qQ}{mv_0^2}\cot\left(\frac{\phi}{2}\right) \tag{11.4.327}$$

and using the following usual derivative proved in the section of Differential and Integral Calculus (see page 1475):

$$\cot'(x)=-\frac{1}{\sin^2(x)} \tag{11.4.328}$$

we have:

$$\frac{\mathrm{d}b}{\mathrm{d}\phi}=\frac{1}{4\pi\varepsilon_0}\frac{qQ}{mv_0^2}\frac{\mathrm{d}}{\mathrm{d}\phi}\cot\left(\frac{\phi}{2}\right)=-\frac{1}{4\pi\varepsilon_0}\frac{qQ}{m\nu_0^2}\frac{1}{2\sin^2\left(\dfrac{\phi}{2}\right)} \tag{11.4.329}$$

Hence:

$$\mathrm{d}b=-\frac{1}{4\pi\varepsilon_0}\frac{qQ}{mv_0^2}\frac{1}{2\sin^2\left(\dfrac{\phi}{2}\right)}\mathrm{d}\phi \tag{11.4.330}$$

Therefore:

$$b\mathrm{d}b=-\left(\frac{1}{4\pi\varepsilon_0}\right)^2\left(\frac{qQ}{mv_0^2}\right)^2\frac{\cot\left(\dfrac{\phi}{2}\right)}{2\sin^2\left(\dfrac{\phi}{2}\right)}\mathrm{d}\phi=-\left(\frac{1}{4\pi\varepsilon_0}\right)^2\left(\frac{qQ}{mv_0^2}\right)^2\frac{\cos\left(\dfrac{\phi}{2}\right)}{2\sin^3\left(\dfrac{\phi}{2}\right)}\mathrm{d}\phi \tag{11.4.331}$$

Thus:

$$\begin{aligned}\frac{b\mathrm{d}b}{\sin(\phi)\mathrm{d}\phi}&=-\left(\frac{1}{4\pi\varepsilon_0}\right)^2\left(\frac{q\mathrm{Q}}{mv_0^2}\right)^2\frac{1}{2\sin^3\left(\dfrac{\phi}{2}\right)}\frac{\cos\left(\dfrac{\phi}{2}\right)}{\sin(\phi)}\\&=-\left(\frac{1}{4\pi\varepsilon_0}\right)^2\left(\frac{q\mathrm{Q}}{mv_0^2}\right)^2\frac{1}{2\sin^3\left(\dfrac{\phi}{2}\right)}\frac{\cos\left(\dfrac{\phi}{2}\right)}{2\sin\left(\dfrac{\phi}{2}\right)\cos\left(\dfrac{\phi}{2}\right)}\\&=-\left(\frac{1}{4\pi\varepsilon_0}\right)^2\left(\frac{q\mathrm{Q}}{mv_0^2}\right)^2\frac{1}{4\sin^4\left(\dfrac{\phi}{2}\right)}\end{aligned}\tag{11.4.332}$$

We have then, remembering that the member on the left is nothing but the effective section:

$$\begin{aligned}\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega}&=-\frac{(q\mathrm{Q})^2}{(4\pi\varepsilon_0)^2m^2v_0^4}\frac{1}{4\sin^4\left(\dfrac{\phi}{2}\right)}=-\frac{(q\mathrm{Q})^2}{(4\pi\varepsilon_0)^2 4\dfrac{1}{4}m^2v_0^4}\frac{1}{4\sin^4\left(\dfrac{\phi}{2}\right)}\\&=-\frac{(q\mathrm{Q})^2}{16(4\pi\varepsilon_0)^2E_c^2}\frac{1}{\sin^4\left(\dfrac{\phi}{2}\right)}\end{aligned}\tag{11.4.333}$$

and it is customary to take the absolute value to define the "effective (differential) section of Rutherford (or Coulomb)":

$$\boxed{\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega}=\frac{(q\mathrm{Q})^2}{16(4\pi\varepsilon_0)^2E_c^2}\frac{1}{\sin^4\left(\dfrac{\phi}{2}\right)}}\tag{11.4.334}$$

Interestingly, a fully quantum mechanical calculation (albeit non-relativistic) would yield exactly the same answer!

We notice several interesting things:

1. For an angle of incidence equal to zero, the effective cross-section diverges (because of the sinus in the denominator)
2. The effective section decreases according to the square of the kinetic energy of the incident particle
3. The expression is valid whatever the charges involved (positive or negative)

Thanks to the Rutherford / Coulomb scattering, Rutherford was able to determine an approximation of the size of the nucleus of the atom (bombardment of a gold leaf[38] using alpha nuclei) as we did notice it at the beginning of the section of Corpuscular Quantum Physics. The reasoning applied is as follows to determine a lower bound of the radius of the nucleus:

---

[38] His best estimate came not from his famous gold foil but form lighter aluminium foil, a much less exotic material.

The total energy of a rotating system is the kinetic energy of translation summed with the kinetic energy of rotation, summed to the potential energy. Which gives us:

$$E=\frac{1}{2}m\dot{r}^2+\frac{1}{2}m\omega^2r^2+U(r)=\frac{1}{2}m\dot{r}^2+\frac{1}{2}mr^2\dot{\theta}^2+U(r) \qquad (11.4.335)$$

By denoting by L the angular momentum given by $L=mr^2\dot{\theta}$ (see section Classical Mechanics page 3244) we have:

$$E=\frac{1}{2}m\dot{r}^2+\frac{L^2}{2mr^2}+U(r) \qquad (11.4.336)$$

hence:

$$\dot{r}=\frac{\mathrm{d}r}{\mathrm{d}t}=\sqrt{\frac{2}{m}\left(E-\frac{L^2}{2mr^2}-U(r)\right)}=\frac{\mathrm{d}r}{\mathrm{d}\theta}\frac{\mathrm{d}\theta}{\mathrm{d}t} \qquad (11.4.337)$$

It therefore follows:

$$\mathrm{d}\theta=\frac{L\mathrm{d}r}{r^2\sqrt{2m\left(E-\frac{L^2}{2mr^2}-U(r)\right)}} \qquad (11.4.338)$$

From this it follows that the angle associated with two radial distances r_1, r_2 is given by:

$$\Delta\theta=\int_{r_1}^{r_2}\frac{L\mathrm{d}r}{r^2\sqrt{2m\left(E-\frac{L^2}{2mr^2}-U(r)\right)}} \qquad (11.4.339)$$

The figure below shows a process of collision by a central potential $U(r)$. The incident particle has an initial speed:

$$v=\sqrt{\frac{2E_c}{m}} \qquad (11.4.340)$$

in $t=-\infty$ with $z(t=-\infty)$ and $y(t=+\infty)$ by symmetry again.

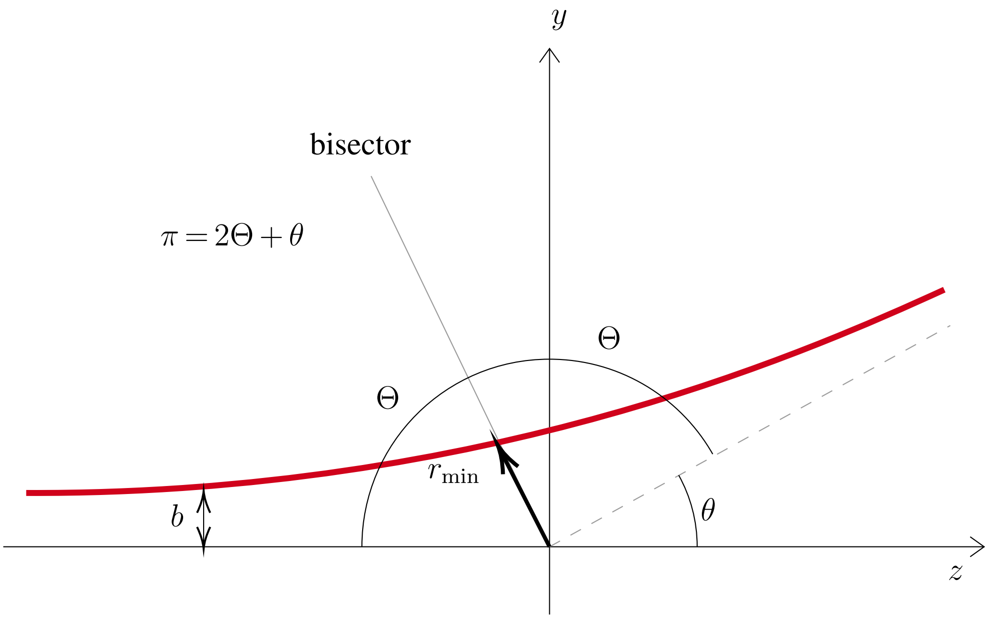


Figure 11.124: Schematic approach for determining the radius of the target

The angle Θ is the angle of deflection when the incident particle approaches the diffuser at the minimum distance r_{min}.

Let us return to our equations where the angular momentum is linked to the impact parameter by the relation $L=mvb$ or written differently:

$$L=b\sqrt{2mE_c} \tag{11.4.341}$$

We can therefore write after simplifications:

$$\begin{aligned}\Delta\theta=\int\limits_{r_1}^{r_2}\frac{L\mathrm{d}r}{r^2\sqrt{2m\left(E-\dfrac{L^2}{2mr^2}-U(r)\right)}}&\Rightarrow\Theta=\int\limits_{r_{\text{min}}}^{+\infty}\frac{b\sqrt{2mE_c}\mathrm{d}r}{r^2\sqrt{2m\left(E-\dfrac{L^2}{2mr^2}-U(r)\right)}}\\
=\int\limits_{r_{\text{min}}}^{+\infty}\frac{b\mathrm{d}r}{r^2\dfrac{1}{\sqrt{2mE_c}}\sqrt{2m\left(E-\dfrac{L^2}{2mr^2}-U(r)\right)}}&=\int\limits_{r_{\text{min}}}^{+\infty}\frac{b\mathrm{d}r}{r^2\sqrt{\dfrac{\cancel{2m}}{\cancel{2m}E_c}\left(E-\dfrac{b^2E_c\cancel{2m}}{\cancel{2m}r^2}-U(r)\right)}}\\
=\int\limits_{r_{\text{min}}}^{+\infty}\frac{b\mathrm{d}r}{r^2\sqrt{\dfrac{E}{E_c}-\dfrac{b^2}{r^2}-\dfrac{U(r)}{E_c}}}=\int\limits_{r_{\text{min}}}^{+\infty}&\frac{b\mathrm{d}r}{r^2\sqrt{1-\dfrac{b^2}{r^2}-\dfrac{U(r)}{E_c}}}=\int\limits_{r_{\text{min}}}^{+\infty}\frac{b\mathrm{d}r}{\sqrt{r^4-r^2b^2-\dfrac{r^4U(r)}{E_c}}}\\
=\int\limits_{r_{\text{min}}}^{+\infty}\frac{b\mathrm{d}r}{\sqrt{r^2\left(r^2-r\dfrac{A}{E_c}-b^2\right)}}&\end{aligned} \tag{11.4.342}$$

Where we put $E\cong E_c$ (the rotation energy and the potential energy are considered negligible with respect to the kinetic energy) and:

$$U(r)=\frac{1}{4\pi\varepsilon_0}\frac{\mathrm{Z}_1\mathrm{Z}_2q_e^2}{r}=\frac{A}{r} \tag{11.4.343}$$

The minimum approach distance is thus determined by the largest zero of the denominator:

$$r^2-\frac{A}{E_c}r-b^2=0 \tag{11.4.344}$$

That is to say (trivial):

$$r_{\text{min}}=\frac{\dfrac{A}{E_c}+\sqrt{\left(\dfrac{A}{E_c}\right)^2+4b^2}}{2}=\frac{A}{2E_c}+\sqrt{\left(\frac{A}{2E_c}\right)^2+b^2} \tag{11.4.345}$$

So we have:

$$\boxed{r_{\text{min}}=\frac{A}{2E_c}\left[1+\sqrt{1+\left(\frac{E_cb}{A}\right)^2}\right],A=\frac{\mathrm{Z}_1\mathrm{Z}_2q_e^2}{4\pi\varepsilon_0},E_c=\frac{1}{2}mv_0^2} \tag{11.4.346}$$

As we see in this last relation, the incident particle will undergo a frontal collision when $b=0$. Therefore, the value of the maximum approach is:

$$r_{\text{min}}=\dfrac{A}{E_c}=\dfrac{Z_1 Z_2 q_e^2}{4\pi\varepsilon_0 E_c} \tag{11.4.347}$$

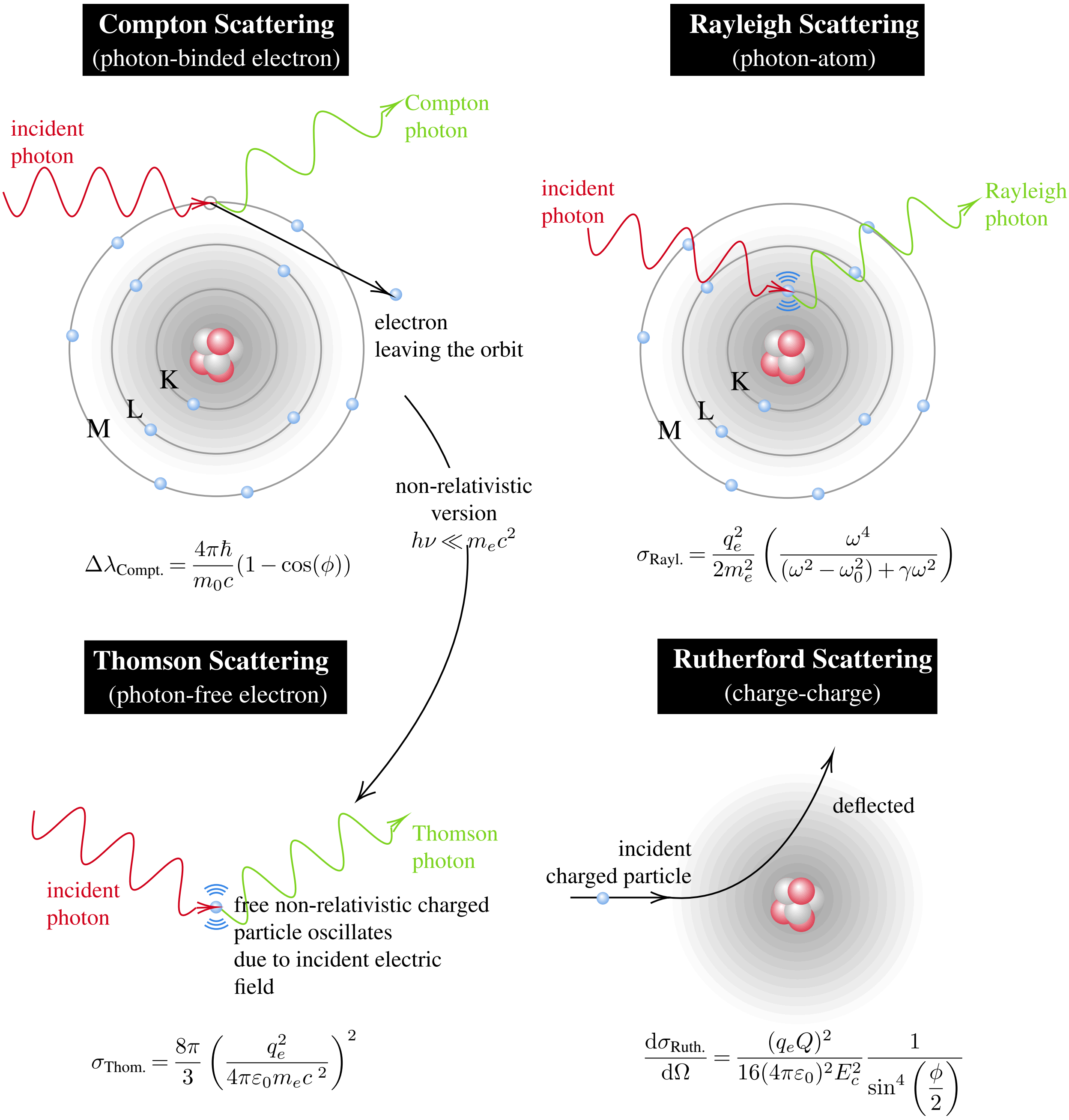
Rutherford's experiment made it possible to estimate the size of the atomic nucleus. Indeed, the particles that have bounced back on the nucleus with a diffusion angle of 180° (we are then talking of "backscattering") are those that have approached nearest to the latter. Since we have:

$$\dfrac{1}{2}mv_0^2=\dfrac{1}{4\pi\varepsilon_0}\dfrac{Z_1 Z_2 q_e^2}{r_{\text{min}}} \tag{11.4.348}$$

with an initial kinetic energy of 7.7 [MeV], Rutherford found for the radius of the gold atom ($Z=79$) with alpha particles ($Z=2$) a value of:

$$a_0:=r_{\text{min}}=3\cdot 10^{-14}\;[\text{m}] \tag{11.4.349}$$

Thus, the nucleus is not punctual but of the order of ten tens of femtometres!


Compton Scattering
(photon-binded electron)
incident
photon
Compton
photon
electron
leaving the orbit
K
L
M
Rayleigh Scattering
(photon-atom)
incident
photon
Rayleigh
photon
K
L
M
non-relativistic
version
$h\nu \ll m_e c^2$
Thomson Scattering
(photon-free electron)
incident
photon
Thomson
photon
free non-relativistic charged
particle oscillates
due to incident electric
field
Rutherford Scattering
(charge-charge)
deflected
incident
charged particle

$$\Delta\lambda_{\text{Compt.}}=\frac{4\pi\hbar}{m_0 c}(1-\cos(\phi))$$

$$\sigma_{\text{Rayl.}}=\frac{q_e^2}{2m_e^2}\left(\frac{\omega^4}{(\omega^2-\omega_0^2)+\gamma\omega^2}\right)$$

$$\sigma_{\text{Thom.}}=\frac{8\pi}{3}\left(\frac{q_e^2}{4\pi\varepsilon_0 m_e c^{\,2}}\right)^2$$

$$\frac{\mathrm{d}\sigma_{\text{Ruth.}}}{\mathrm{d}\Omega}=\frac{(q_e Q)^2}{16(4\pi\varepsilon_0)^2E_c^2}\frac{1}{\sin^4\left(\dfrac{\phi}{2}\right)}$$

Thomson and Rayleigh scattering are elastic: only the direction
of the photon changes, not its energy. Plus they are only important at low
energies (where the photelectric effect dominates).
Rutherford scattering is also an elastic scattering in a Coulomb potential.

Figure 11.125: Summary of most well known type of scattering

11.4.3.6 X-rays and Gamma rays

The fundamental difference of this type of radiation, with respect to the α, β^-, β^+ is that it does not carry an electric charge and therefore does not have a coulombic interaction with the electronic cloud of the medium traversed. Consequently, the photon follows a straight path without loss of energy until it encounters a particle on its path (electron, nucleus) where it will undergo an interaction that profoundly alters its state.

"Gamma radiation" is a high-energy electromagnetic radiation produced by a nuclear phenomenon, while "X-rays radiation" are high-energy electromagnetic radiation produced during atomic or molecular phenomena. The photon is the elementary particle that is associated with these electromagnetic waves. The X and gamma photon are therefore of the same nature but of different origins, they therefore have identical properties that depend on their energy.

Let us recall that:

$$E=h\nu=\frac{hc}{\lambda} \qquad p=\frac{h\nu}{c}=\frac{h}{\lambda} \tag{11.4.350}$$

By crossing mater a photon can interact with:

- One of the electrons of the atom cloud
- The nucleus of the atom
- The electric field of charged atomic particles
- The meson field of the nucleons (strong interaction)

The result of the interaction can be schematized as follows:

- The photon is deflected by conserving its energy, then there is "total diffusion" of the energy and the process is say to be "coherent" (elastic).
- The photon is deflected and its energy decreased, then there is "partial diffusion" of energy, the other part is absorbed by matter, the process is then say to be "incoherent" (inelastic).
- The photon disappears, there is "total absorption" of its energy by matter.

We can prove that the macroscopic characteristics of these interactions in the framework of a thin and collimated beam lead to an exponential law of attenuation of photonic radiation in matter. This means that for photons there is no finite path (!) as for charged particles; It can never be assured that at a given distance all the photons of a beam have interacted.

The number of particles interacting with the material obviously depends on the intensity I and the type of material traversed (characterized by the "linear attenuation coefficient μ") and its thickness x.

We have:

$$\mathrm{d}I=-\mu I\mathrm{d}x \tag{11.4.351}$$

the "$-$" sign being there to show a decrease. We easily solve this differential equation (it is simply the Beer-Lambert law that we have already studied in the section of Geometric Optics):

$$I=I_0e^{-\mu x} \tag{11.4.352}$$

with I_0 being the initial intensity or "flow rate of fluence" and μ being the coefficient of linear attenuation in [cm^{-1}] which takes into account all the possible attenuation effects.

Remark

Often in the tables, we find the coefficient of attenuation mass μ/ρ expressed in [cm^2·g^{-1}]. We have then:

$$I=I_0e^{-\frac{\mu}{\rho}\rho\mathrm{d}x} \tag{11.4.353}$$

In the case of an absorbent containing several homogeneously distributed chemical elements, the attenuation coefficient is equal to:

$$\mu_T=\sum_i w_i\mu_i \quad \text{or} \quad \left(\frac{\mu}{\rho}\right)_T=\sum_i w_i\left(\frac{\mu}{\rho_i}\right) \tag{11.4.354}$$

where μ_T is the absorption coefficient of the absorbent, μ_i the absorption coefficient of the element i, ρ_T the density of the absorbent, ρ_i the density of the element i, w_i being the mass fraction of the element i in the absorbent.

Let us now take a microscopic approach:

Let a beam of I [particles· cm^{-2}·s^{-1}] striking perpendicularly the surface of a material of thickness $\mathrm{d}x$ and of atomic density N [atoms·cm^{-3}]. If we consider the particles striking the surface S, the latter can theoretically encounter $N\cdot S\cdot\mathrm{d}x$ target atoms in this layer. The number of particles interacting will be proportional to the intensity times this number and we will have:

$$\mathrm{d}I\cdot S=-\sigma\cdot I\cdot N\cdot S\cdot\mathrm{d}x \tag{11.4.355}$$

where σ is the constant of proportionality, named "microscopic cross section". Its units are often expressed in [barn] (1 [barn]=10^{-24} [cm^2]).

Remarks

R1. The atomic density N is equal to $\rho N_\mathrm{A}\cdot M_m^{-1}$ where $\rho$is the density in [g· cm^{-3}] of the target, N_A the Avogadro constant ($6.022\cdot10^{23}$ [atoms·mole^{-1}]) and M_m is the molar mass of the target expressed in [g· mole^{-1}].

R2. If we admit that the diffusion centers are the electrons and not the target atoms, then we have to replace N by $N_e=N\cdot Z$.

From where we get:

$$I=I_0e^{-NZ\sigma x}=I_0e^{-N_e\sigma x}=I_0e^{-\frac{ZN_\mathrm{A}\rho\sigma}{M_m}x} \tag{11.4.356}$$

By identifying the macro and microscopic aspect, we see that μ plays the same role as $ZN_\text{A}\rho\sigma/M_m$ and that we find that the effective section can be written as:

$$\sigma = \frac{\mu}{\rho}\frac{M_m}{ZN_\text{A}} \tag{11.4.357}$$

and in the hypothesis that the electron constitutes a "sphere of action" having a frontal surface πr_e^2, where $r_e$is the radius of the sphere of action then:

$$\sigma = \pi r_e^2 = \frac{\mu}{\rho}\frac{M_m}{ZN_\text{A}} \tag{11.4.358}$$

and we have:

$$r_e = \sqrt{\frac{\mu}{\rho}\frac{M}{ZN_\text{A}\pi}} \tag{11.4.359}$$

By definition, we name the "Half-Attenuation Elbow (HAE)" the thickness of the material that divides the fluence rate I_0 by a factor of two. Therefore:

$$\frac{I_0}{2} = I_0 e^{-\mu x} \Rightarrow \frac{1}{2} = e^{-\mu x} \Rightarrow \ln\left(\frac{1}{2}\right) = -\mu x \Rightarrow x = -\frac{\ln\left(\frac{1}{2}\right)}{\mu} = -\frac{\ln(2)}{\mu} = \text{HAE} \tag{11.4.360}$$

In radio-protection, we sometimes use the notion of attenuation layer to the tenth TVL (Tenth Value Layer) given by:

$$\text{TVL} = \frac{\ln(10)}{\mu} \tag{11.4.361}$$

We sometimes also use the "relaxation length", which represents the thickness from which the intensity of a monoenergetic beam is decreased by a factor e, and which is therefore given by:

$$\lambda = \frac{1}{\mu} \tag{11.4.362}$$

This value is much more useful than the others, because it is also the average distance at which the first photon collision takes place.

Almost everyone going to the dentists or who broke some bones know how useful are X-rays the make diagnostics. However, far fewer people know how mathematics and X-rays associated together are useful to generate 3D detailed internal images of the body thanks to "Computed Tomography scan" (usually abbreviated "CT Scan"; formerly named "Computed Axial Tomography scan" or "CAT scan").

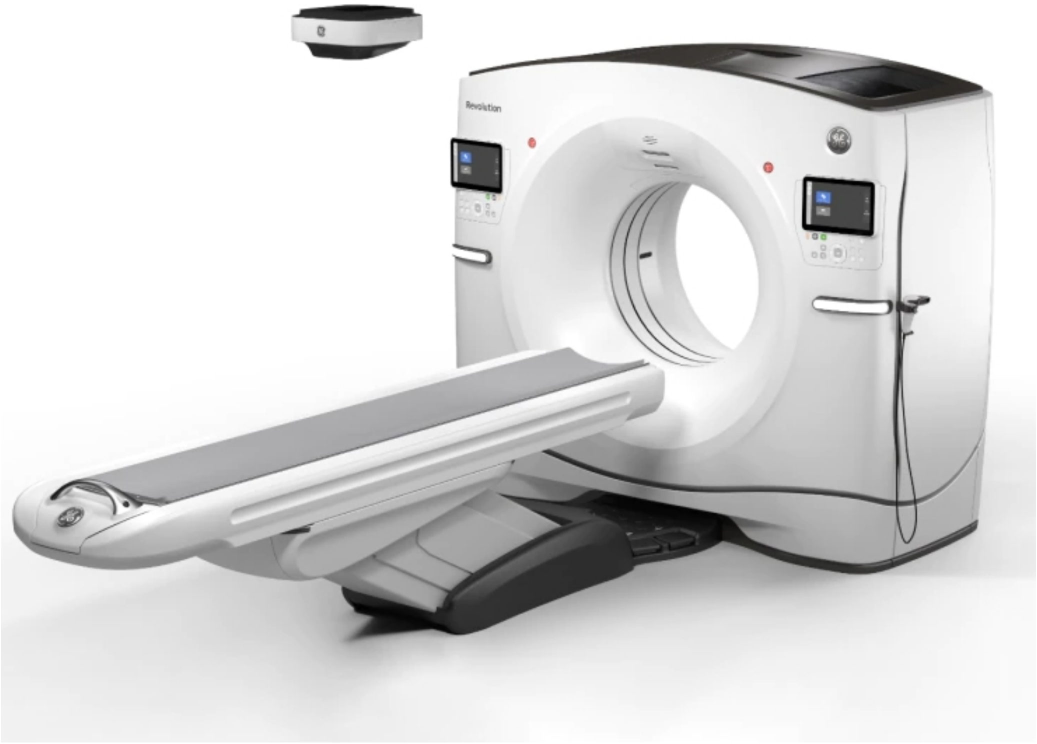


Figure 11.126: **General Electrics CT Scan Revolution model**

"Single-Photon Emission Computed Tomography" (SPECT) is a nuclear medicine tomographic imaging technique using gamma rays. It is very similar to conventional nuclear medicine planar imaging using a gamma camera (that is, scintigraphy) and is also able to provide 3D detailed internal images of the body. The technique needs delivery of a gamma-emitting radioisotope (a radionuclide) into the patient, normally through injection into the bloodstream. On occasion, the radioisotope is a simple soluble dissolved ion, such as an isotope of gallium.

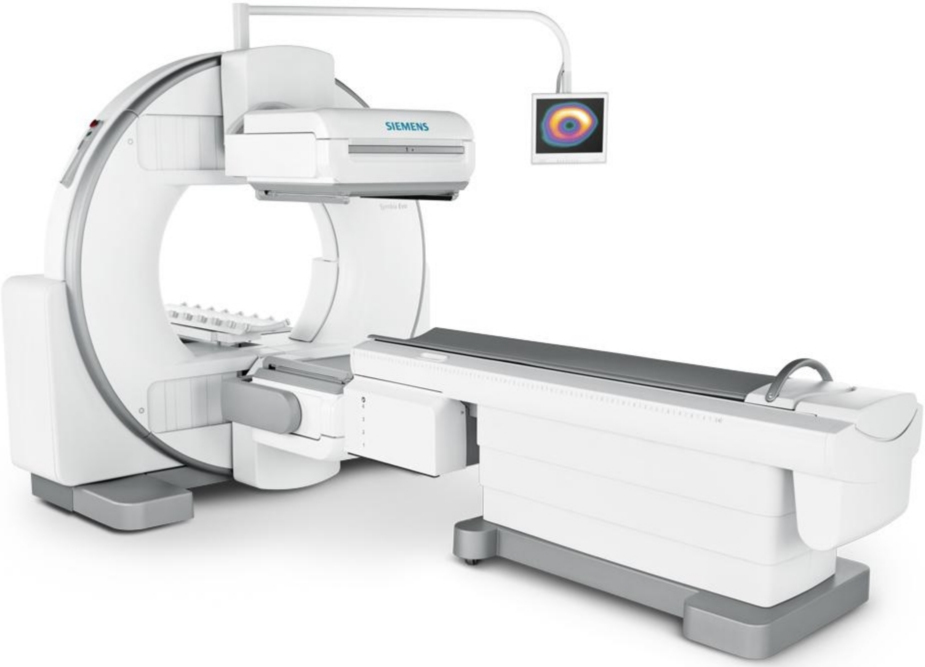


Figure 11.127: **Siemens SPECT scan Symbia model**

Remark

The Gamma irradiation is anecdotally used in the conservation of the heritage of organic objects. Indeed, during the discovery of archaeologists of ancient works or vestiges, the latter are attacked by micro-organisms that will destroy these objects over time. The gamma radiation will allow, without destroying the objects, to kill by gamma irradiation all these micro-organisms. The best known example is the irradiation of the Tutankhamun's mummy during 10 hours in the CEA laboratories.

An illustration of the most well known effect of gammas rays:

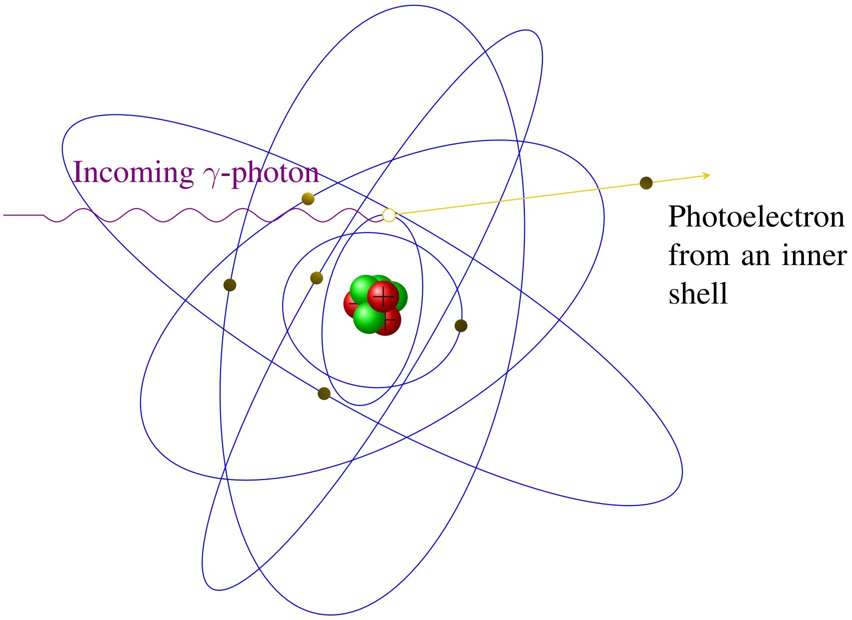


Figure 11.128: **Photoelectric effect with gamma ray naive Bohr model perspective (author: Poul Riis)**

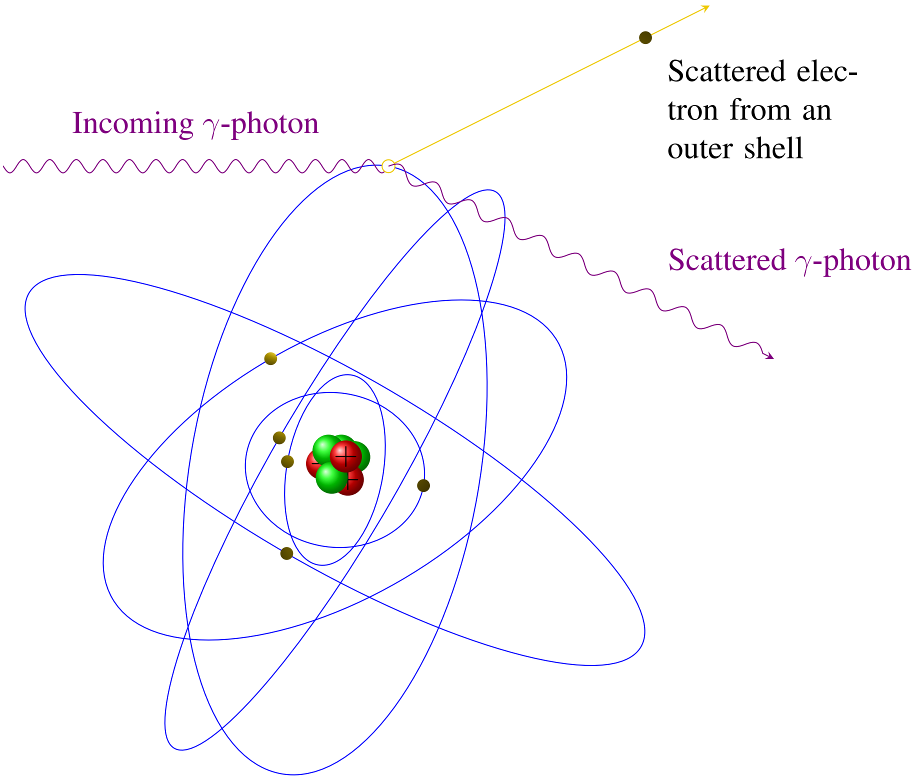

Incoming γ-photon
Scattered electron from an outer shell
Scattered γ-photon

Figure 11.129: Compton scattering with gamma ray naive Bohr model perspective (author: Poul Riis)

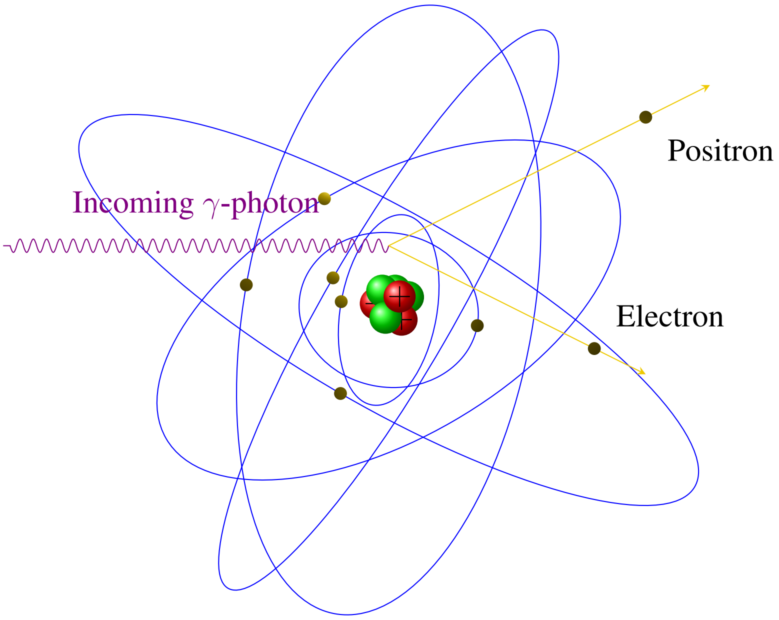

Positron
Incoming γ-photon
Electron

Figure 11.130: Pair creation with gamma ray naive Bohr model perspective (author: Poul Riis)

The known microscopic causes of the attenuation of a photon beam (neutral from the Coulombian point of view) which deserve our attention in the energy domain of gamma or X-ray photons are:

- Coherent Thomson scattering

- Coherent Rayleigh scattering
- Coherent Delbrück scattering
- Raman scattering (inelastic equivalent of Rayleigh scattering)
- Coherent Compton scattering (already partially seen earlier above)
- Photoelectric absorption (already partially seen earlier above)
- Photonuclear reaction
- Creation of electron-positron pairs (already partially seen earlier above)

Although we can now discuss these effects, it is impossible for us in the present state of this textbook to introduce the mathematical formalism used to determine the effective cross-section of each of these scattering.

11.4.3.6.1 Electron-Positron pair production

The concept of the wave-particle duality which means all matter has both "particle-like" and "wave-like" nature has been the inception of the quantum physics. Especially, for the "particle-like" nature, there are some examples that can be an empirical proof: first two processes as proves are the photoelectric effect and the Compton effect. Those two effects are well-known processes of photon absorption or scattering that photon loses its energy by interacting with other matter as a "particle-like" things.

The final one for proving "particle-like" nature of all matter is "pair production" or especially the special case of "electron-positron pair production". It was first observed by the physicist, Patrick Blackett, who was the winner of the 11948 (holocene calendar) Novel Prize in physics. While both the two processes mentioned above - the photoelectric effect and the Compton effect - usually occur in low energy and middle energy condition, pair production interaction is known to occur when photon has high enough energy before its collision.

The pair production is a crucial example that photon energy (gamma ray) can convert into kinetic energy as well as rest mass energy:

$$\gamma \rightarrow e^- + e^+ \tag{11.4.363}$$

Schematic diagram about the process of pair production is shown in the figure below. The high-energy photon that has energy $h\nu$ loses its entire energy when it collides with nucleus. Then, it makes pair of electron and positron and gives kinetic energy to each particle.

Basically, these interactions are ruled by three kind of the law of conservation: total energy, momentum, and electric charge.

During the creation of pairs, the photon absorbed in the electric field of the nucleus can generate an electron-positron pair. For the interaction to take place, the energy of the photon must be greater than the $2m_0c^2$ (about 1.02 [MeV]) that is to say the energy at rest of the electron-positron pair.

This effect is important for high energies and high atomic numbers. The positron created is braked in matter just like an electron and, at the end of the course, it annihilates with an electron to give rise to two photons of 0.511 [MeV] ("annihilation photons") emitted almost at $180°$ (the whole linear momentum being transformed into energy explains the value of this angle, so the final linear momentum is zero).

The creation of a pair obviously costs at least the mass energy of the electron and the positron, or $2m_0c^2$. The energy balance is then divided into the kinetic energy of the two particles:

$$h\nu = 2m_0c^2 + E_c^{\beta+} + E_c^{\beta-} \tag{11.4.364}$$

The necessity of simultaneously satisfying the conditions of conservation of the mass energy and the linear momentum the other hand imposes on the materialization effect to take place in the vicinity of a material particle (nucleus most of time) which participates in the phenomenon. Indeed, in vacuum, the two conditions are contradictory! The linear momentum of each electron is:

$$p_{\beta^+,\beta^-} = mv \tag{11.4.365}$$

We have proved in the section of Special relativity that:

$$mc^2 = E_c + m_0c^2 = E_{\text{tot}} \tag{11.4.366}$$

And therefore:

$$p_{\beta^+,\beta^-} = mv = \frac{E_{\text{tot}}}{c^2}v = \frac{E_c + m_0c^2}{c^2}v \tag{11.4.367}$$

Therefore:

$$E_{\text{tot}} = p_{\beta^+,\beta^-}\left(\frac{c^2}{v}\right) \tag{11.4.368}$$

The original photon has:

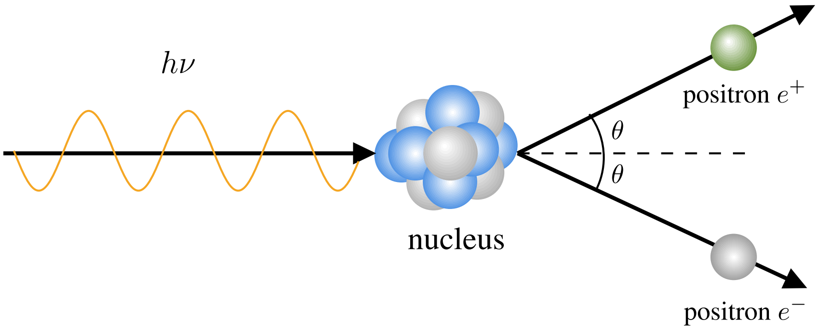
$$h\nu = 2E_{\text{tot}} \quad \text{and} \quad p_\gamma = \frac{h\nu}{c} \tag{11.4.369}$$

which we introduce into the energy conservation equation and with the help of the relation p_{β^+,β^-} we have:

$$h\nu = p_\gamma c = 2E_{\text{tot}} = 2p_{\beta^+,\beta^-}\left(\frac{c^2}{v}\right) \tag{11.4.370}$$

which shows that by the term $c^2/v > 1$ that the nucleus must take away a part of the linear momentum since:

$$p_{\beta^+,\beta^-} < p_{\beta^+,\beta^-}\left(\frac{c^2}{\nu}\right) \tag{11.4.371}$$



The presence of an electric field of a heavy atom such as lead or uranium is then essential in order to satisfy conservation of momentum and energy. In order to satisfy both conservation of momentum and energy, the atomic nucleus must receive some momentum.

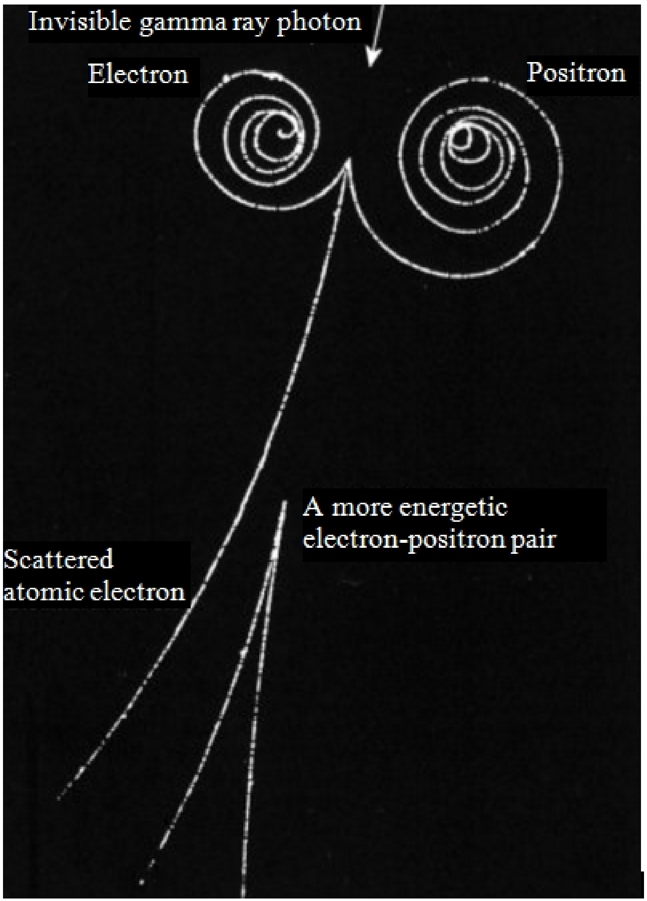

Invisible gamma ray photon
Electron
Positron
A more energetic electron-positron pair
Scattered atomic electron

Figure 11.131: **Electron-Positron pair production in a bubble chamber**

Remark

Pair production is invoked to predict the existence of hypothetical Hawking radiation. According to quantum mechanics, particle pairs are constantly appearing and disappearing as a quantum foam. In a region of strong gravitational tidal forces, the two particles in a pair may sometimes be wrenched apart before they have a chance to mutually annihilate. When this happens in the region around a Black Hole, one particle may escape while its antiparticle partner is captured by the Black Hole.

11.4.3.7 Radiation dosimetry

In what follows, the radiation field is characterized in terms of energy and flux at a point of space. These physical datas than make possible to obtain the radiometric and dosimetric quantities of reference for external exposure, such as: fluence, kerma for the indirectly ionizing radiations and the dose. In addition, other quantities such as linear energy transfer (this parameter reflects the energy loss of charged particles per unit path length) and relative biological effectiveness are detailed; they relate to the effect of radiation on biological tissues and to the damage that can occur at the cellular level.

Initially, we will attempt to define the characteristic quantities of dosimetry and radiation fields (i.e. radiometric quantities). The fundamental radiometric and dosimetric quantities are essentially threefold: fluence Φ, absorbed dose $\mathcal{D}$ and Kerma K.

Definitions (#11.4.520):

D1. "Radiometric quantities" determine the strength of the radiation source(s). "Dosimetric quantities" are related to the energy transfer from radiation to matter.

Parameters describing a relation between those are named "interaction parameters". The intensity of a beam of particles in matter decreases according to the Beer-Lambert law (see section Optics page 4337):

$$I(x)=I_0e^{-\mu x} \tag{11.4.372}$$

where the linear attenuation coefficient is given by $\mu=(\mathrm{d}N/N)/\mathrm{d}x$. We define also the "mass absorption coefficient" by the simple ratio: μ/ρ.

D2. The deceleration of a heavy particle is described by the Bethe-Bloch formula proved earlier above (see page 5092):

$$L_\Delta=-\frac{\mathrm{d}E}{\mathrm{d}x}\propto\frac{e^4}{v^2} \tag{11.4.373}$$

and defines the "linear energy transfer".

Absorber	Density	Alpha Range	Comments
Air (STP)	$1.2\,[\mathrm{mg}\cdot\mathrm{cm}^{-3}]$	$3.7\,[\mathrm{cm}]$	-
Paper	$0.89\,[\mathrm{g}\cdot\mathrm{cm}^{-3}]$	$53\mu\,[\mathrm{m}]$	one sheet $=89\mu\,[\mathrm{m}]$
Water	$1.0\,[\mathrm{g}\cdot\mathrm{cm}^{-3}]$	$45\mu\,[\mathrm{m}]$	will not penetrate skin

Table 11.18: **Absorbing materials and their α particle penetration depths**

Absorber	Density	Depth $(2.3\,[\mathrm{MeV}])$	Depth $(1.1\,[\mathrm{MeV}])$
Air	$1.2\,[\mathrm{mg}\cdot\mathrm{cm}^{-3}]$	$8.8\,[\mathrm{m}]$	$3.8\,[\mathrm{m}]$
Water (soft tissue)	$1.0\,[\mathrm{g}\cdot\mathrm{cm}^{-3}]$	$11\,[\mathrm{mm}]$	$4.6\,[\mathrm{mm}]$
Plastic (acrylic)	1.2	9.6	4.0
Glass (Pyrex)	2.2	5.6	2.2
Aluminum	2.7	4.2	2.0
Copper	8.9	1.2	0.5
Lead	11.3	1.0	0.4

Table 11.19: **Absorbing materials and their β^- particle maximum penetration depths**

Energy [keV]	**Linear Attenuation Coefficient** $[\mathrm{m}^{-1}]$)
30	56.0691
40	34.34655
50	26.2803
60	22.39471
80	18.76885
100	17.00037
150	14.6686
200	13.25447
300	11.44662
400	10.2306
500	9.318863
600	8.621373
662	8.3048
800	7.571909
$1,000$	6.805753
$1,500$	5.525981

Table 11.20: **Linear attenuation coefficient of γ rays in human body**

D3. The "fluention" or "fluence" is defined by $\Phi = \mathrm{d}N/\mathrm{d}S$. The *flux* is obviously given by $\phi = \mathrm{d}\Phi/\mathrm{d}t$. The energy loss is defined by $\Psi = \mathrm{d}W/\mathrm{d}S$, and the energy flux density $\psi = \mathrm{d}\Psi/\mathrm{d}t$.

D4. The "radiation dose" $\mathcal{X}$ is the amount of charge produced by the radiation per unit of mass, with unit $[\mathrm{C}\cdot\mathrm{kg}^{-1}]$ (Coulomb per kilo). An old unit is the "Röntgen":

$$1\,[\mathrm{Ro}] = 2.58\cdot 10^{-4}\,[\mathrm{C}\cdot\mathrm{kg}^{-1}] \qquad (11.4.374)$$

With the "energy-absorption coefficient" μ_E follows:

$$\mathcal{X} = \frac{\mathrm{d}Q}{\mathrm{d}m} = \frac{q + e\mu_E}{W\rho}\Psi \qquad (11.4.375)$$

where the underlying idea is that W is the energy required to disjoin an elementary charge.

D5. The "absorbed dose" $\mathcal{D}$ is defined by:

$$\mathcal{D} = \frac{\mathrm{d}E_{\text{abs}}}{\mathrm{d}m} \qquad (11.4.376)$$

with for unit the "Gray" simply defined as $[\mathrm{Gy}] = [\mathrm{J}\cdot\mathrm{kg}^{-1}]$.

Remark

An old unit is the rad: $1\,[\mathrm{rad}] = 0.01\,[\mathrm{Gy}]$.

D6. The "Kerma" K is the amount of kinetic energy of secondary produced particles which is produced per mass unit of the radiated object.

D7. The "equivalent dose" $\mathcal{H}$, measured in Sieverts $1\,[\text{Sv}]=1\,[\text{J}\cdot\text{kg}^{-1}]$, is a weighted average of the absorbed dose per type of radiation, where for each type radiation the effects on biological material is used for the weight factor. These weight factors are named the "quality factors".

Calculating equivalent dose from absorbed dose is given by:

$$\mathcal{H}_T=\sum_R w_R \mathcal{D}_{T,R} \tag{11.4.377}$$

where:

- $\mathcal{H}_T$ is the equivalent dose in sieverts (Sv) absorbed by tissue or organ T
- $\mathcal{D}_{T,R}$ is the absorbed dose in grays (Gy) in tissue T by radiation type R
- w_r is the radiation weighting factor defined by regulation

Thus for example, an absorbed dose of $1\,[\text{Gy}]$ by alpha particles will lead to an equivalent dose of $20\,[\text{Sv}]$, and an equivalent dose of radiation is estimated to have the same biological effect as an equal amount of absorbed dose of gamma rays, which is given a weighting factor of 1.

Radiation type	w_R
Röntgen, gamma radiation	1
β, electrons, mesons	1
Thermic neutrons	3 to 5
Fast neutrons	10 to 20
protons	10
α, fission products	20

Remark

A dose of 1 Sievert is commonly associated with approximately a 5.5% increase in the risk of fatal cancers. The normal daily radiation dose received by the average person living on Earth is 10 microsieverts. This value is approximately $250,000$ bigger for a planet like Mars....

D8. "Effective dose" $\mathcal{E}$ is calculated for the whole body. It is the addition of equivalent doses to all organs, each adjusted to account for the sensitivity of the organ to radiation. Effective dose is also expressed in Sieverts and given by:

$$\mathcal{E}=\sum_T w_T\mathcal{H}_T=\sum_T w_T\left(\sum_R w_R\mathcal{D}_{T,R}\right)=\sum_T w_T\left(\sum_R w_R\frac{\mathrm{d}E_{\text{abs,T,R}}}{\mathrm{d}m}\right) \tag{11.4.378}$$

where w_T is a weighted factor for a given organ.

Typically for Homo Sapiens we have:

Organ	w_T
Gonads	0.20
Red bone marrow	0.12
Colon	0.12
Lungs	0.12
Stomach	0.12
Bladder	0.05
Breasts	0.05
Liver	0.05
Esophagus	0.05
Thyroid	0.05
Skin	0.01
Bone surface	0.01
Remainder	0.05
Total	1.00

The three previous definitions can be summarized with the following flowchart:

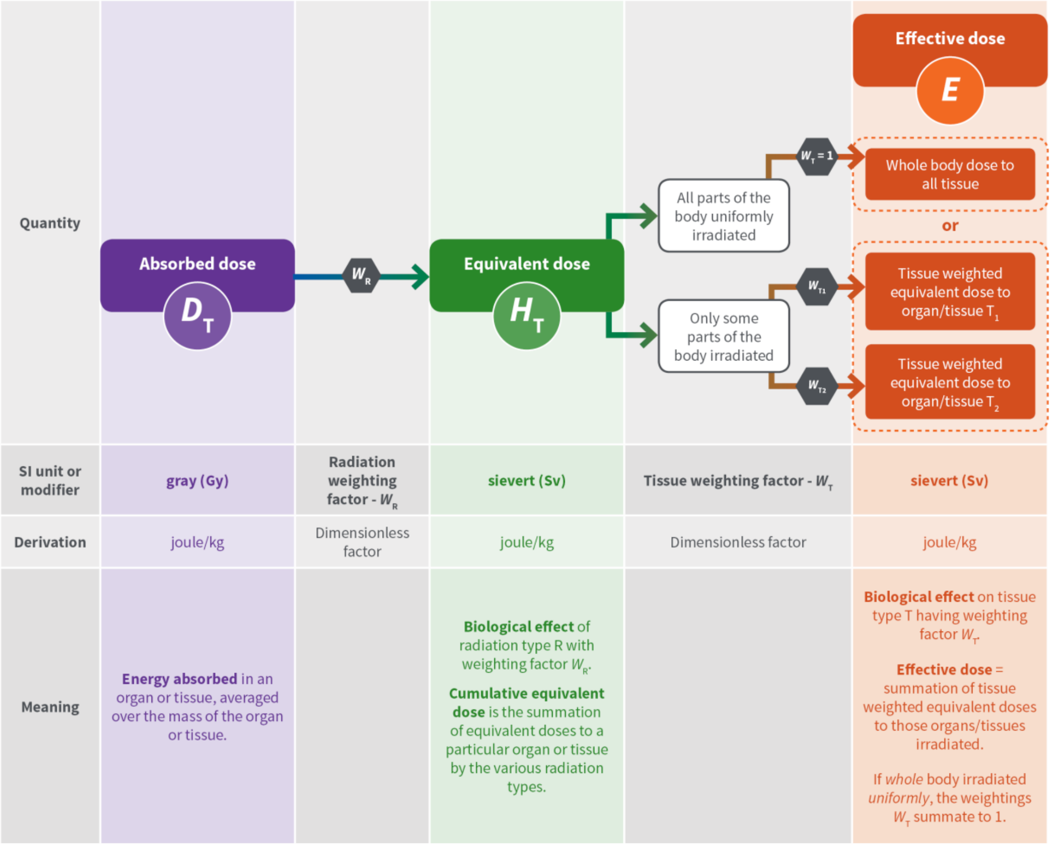

Quantity
Absorbed dose
D_T
W_R
Equivalent dose
H_T
All parts of the body uniformly irradiated
Only some parts of the body irradiated
$W_T = 1$
W_{T1}
W_{T2}
Effective dose
E
Whole body dose to all tissue
or
Tissue weighted equivalent dose to organ/tissue T_1
Tissue weighted equivalent dose to organ/tissue T_2
SI unit or modifier
gray (Gy)
Radiation weighting factor - W_R
sievert (Sv)
Tissue weighting factor - W_T
sievert (Sv)
Derivation
joule/kg
Dimensionless factor
joule/kg
Dimensionless factor
joule/kg
Meaning
Energy absorbed in an organ or tissue, averaged over the mass of the organ or tissue.
Biological effect of radiation type R with weighting factor W_R.
Cumulative equivalent dose is the summation of equivalent doses to a particular organ or tissue by the various radiation types.
Biological effect on tissue type T having weighting factor W_T.
Effective dose = summation of tissue weighted equivalent doses to those organs/tissues irradiated.
If whole body irradiated uniformly, the weightings W_T summate to 1.

Figure 11.132: **Summary of dosimetric quantities**

D9. The "equivalent period" T_{eq} is the time it takes for the number of atoms of an ingested or inhaled radionuclide to decrease by a factor of 2 in the human body. It is therefore linked to the half-life of the radionuclide $T_{1/2}$ but also to the biological half-life T_{bio} representative of the time necessary for the organism to release half of the quantity ingested or inhaled. It is defined by:

$$\frac{1}{T_{\text{eq}}}=\frac{1}{T_{1/2}}+\frac{1}{T_{\text{bio}}} \qquad (11.4.379)$$

Here is a table summarizing the different major units:

Type of unit	**S.I. unit**	**Conventional unit**
Radioactivity (A)	Becquerel [Bq]	Curie [Ci]
	1 [Bq] $=1$ disintegration per second 1[Ci] $=3.7\cdot 10^{10}$ disintegration per second	
Radiation dose (χ)	$[\text{C}\cdot\text{kg}^{-1}]$	Röntgen [Ro]
Absorbed dose ($\mathcal{D}$)	Gray [Gy]	[rad]
	1 [Gy] $=1\,[\text{J}\cdot\text{kg}^{-1}]=100$ [rad]	
Equivalent dose ($\mathcal{H}_T$)	Sievert [Sv]	[rem]
	1 [Sv] $=100$ [rem]	
Effective dose ($\mathcal{E}$)	Sievert [Sv]	[rem]
Linear energy transfer (L_Δ)	Newton [N]	$[\text{keV}\cdot\mu\text{m}^{-1}]$
	1 [N] $=1\,[\text{J}\cdot\text{m}^{-1}]$ $1\,[\text{keV}\cdot\mu\text{m}^{-1}]=1.6\cdot 10^{-13}$ [N]	

Table 11.21: **Summary of major radiation dosimetry units**

The natural radioactivity of the soil comes from its ^{40}K content as well as from the daughter products of uranium and thorium. Their concentration varies from region to region and is a few hundred $[\text{Bq}\cdot\text{kg}^{-1}]$ for the ^{40}K and a few tens of $[\text{Bq}\cdot\text{kg}^{-1}]$ for uranium and thorium daughter products. The effective dose from X-rays indoors is on average 1.5 times greater than in the open air. According to UNSCEAR[39] (11993 according to holocene calendar), the worldwide average effective dose, taking into account a residential occupancy rate of 0.8 is $0.35\,[\text{mSv}\cdot\text{year}^{-1}]$

The dose rate maps from terrestrial and space radiation of Switzerland are given in the figures below. The first concerns the radiation coming from space (cosmic rays), the second figure is the natural terrestrial radiation and the second the artificial terrestrial radiation coming mainly from the Chernobyl accident. The maximum ($90\,[\text{nSv}\cdot\text{h}^{-1}]$) is found in the south-eastern region of Switzerland because of the large amounts of rain which took place just after the accident. The latter caused the deposit of aerosols containing the radioactive products on the ground.

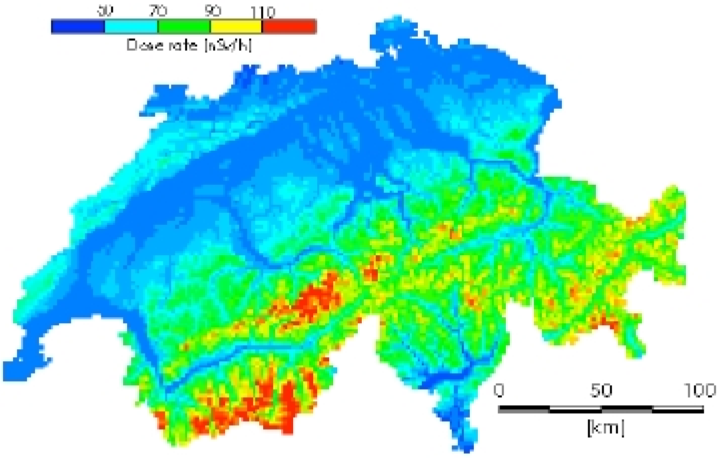


Figure 11.133: **Cosmic dose rate map (in nSv/h) of Switzerland.**
Min. value: 40; Max. value: 191; Average value: 64; Std. deviation: 22

---

[39] United Nations Scientific Committee on the Effects of Atomic Radiation

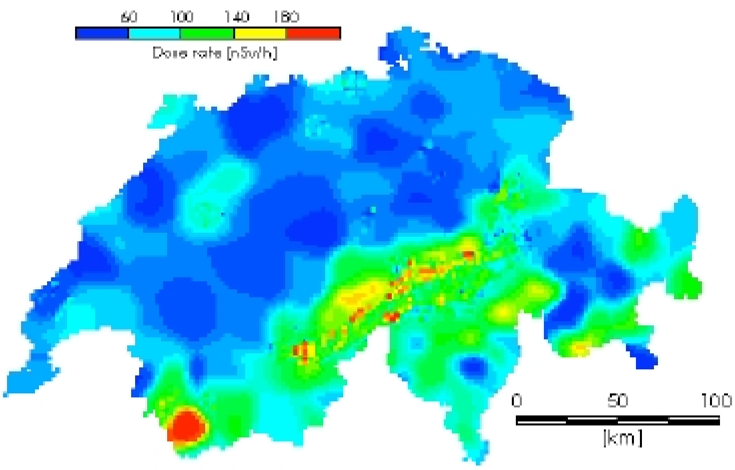


Figure 11.134: Natural terrestrial dose rate map (in nSv/h) of Switzerland. Min. value: 6; Max. value: 368; Average value: 68; Std. deviation: 35

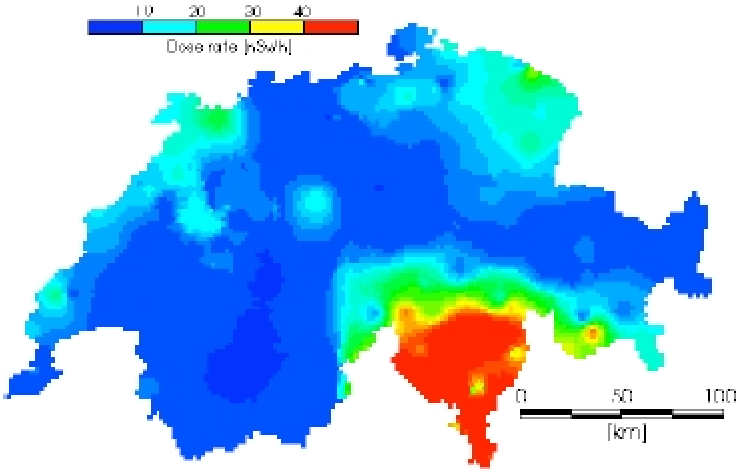


Figure 11.135: Artificial dose rate map (in nSv/h) of Switzerland. Min. value: 1; Max. value: 91; Average value: 11; Std. deviation: 14 (year 12000)

Drinking waters usually contain several natural radionuclides: tritium, radon, radium, uranium isotopes, etc (see [187]). Their concentrations vary widely since they depend on the nature of the aquifer, namely, the prevailing lithology and whether there is air in it or not. They are released from rocks and minerals which form the aquifer as happens with other cations and anions: processes of erosion and dissolution bring radioactive elements from rocks into the water. Some common natural radioelements are those from the uranium-238 chain, a natural radioactive series of many radionuclides, one descending from the other. The most relevant are uranium-238, uranium-234, radium-226 and radon-222. Relatively abundant in the Earth crust $^{238}\mathrm{U}$ and its descendant $^{234}\mathrm{U}$, are often the most abundant radionuclides in water[40] (see [114]). The occurrence of $^{222}\mathrm{Rn}$ may be important too; indeed, because of its moderate solubility in water, it rises mostly by emanation from inner soil layers rather than from the decay of the dissolved parent $^{226}\mathrm{Ra}$. In spite of the fact that the activity concentration of $^{222}\mathrm{Rn}$ in freshly drawn water may be more than two orders higher than that of other radionuclides, it readily diminishes as a result of both desorption and physical decay.

---

[40] In the USA $\sim 0.75 \cdot 10^{-6}$ [kg $\cdot$ L^{-1}] and in Europe $\sim 0.4 \cdot 10^{-6}$ [kg $\cdot$ L^{-1}] to some pic values of $\sim 100 \cdot 10^{-6}$ [kg $\cdot$ L^{-1}] in some Switzerland sources. Keeping in mind that the lethal dose (LD50) of Uranium is 114 [mg] per kilogram.

Another crucial radioactive family in drinking water is the thorium-232 series. Thorium is three to four times more abundant than uranium in the Earth's crust, but owing to its poor solubility it scarcely occurs in waters. Radium-228 belongs into this series and it may become a critical contaminant, as its radiotoxicity is relatively high.

Potassium-40 is also a widespread radionuclide in drinkable water. It is a beta-gamma emitter of primordial origin and goes along stable potassium in a fixed ratio (31.3 [Bq] per gram of stable potassium).

Tritium is a cosmogenic radionuclide; it is a hydrogen isotope formed in the high atmosphere. It occurs in rainfalls as tritiated water (HTO) at a concentration of roughly $5\,[\mathrm{Bq}\cdot\mathrm{L}^{-1}]$. It can be also produced by anthropogenic activities (research and nuclear facilities). It can reach the aquifers along their recharge process, but its concentration progressively decreases due to its relatively short decay time (13 years). Tritium has a low radiotoxicity. In fact its dose coefficient is three orders of magnitude lower than the coefficients of the natural chain radionuclides.

Many national and international institutions have published rules - mandatory or not - that carry the intent of managing the question of radioactivity in drinking water. A short review of the most significant ones is hereby presented:

- WHO (World Health Organization) guidelines for drinking water suggest performing an indirect evaluation of individual dose criterion (IDC) of $0.1\,[\mathrm{mSv}\cdot\mathrm{year}^{-1}]$ by measuring gross alpha and beta radioactivity and checking compliance of radionuclide activity concentration to derived guidance levels .
- The table shows a summary of the European regulations:

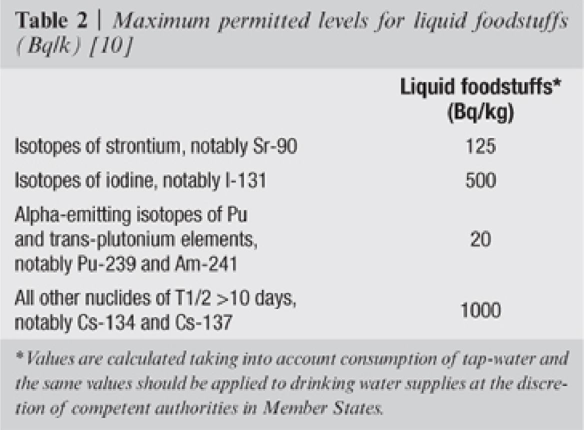


Table 2 | *Maximum permitted levels for liquid foodstuffs (Bq/k) [10]*

	Liquid foodstuffs* (Bq/kg)
Isotopes of strontium, notably Sr-90	125
Isotopes of iodine, notably I-131	500
Alpha-emitting isotopes of Pu and trans-plutonium elements, notably Pu-239 and Am-241	20
All other nuclides of T1/2 >10 days, notably Cs-134 and Cs-137	1000

**Values are calculated taking into account consumption of tap-water and the same values should be applied to drinking water supplies at the discretion of competent authorities in Member States.*

Figure 11.136: **European radioactivity permitted level for liquid foodstuffs**

- The table below shows a summary of the US Environmental Protection Agency (USEPA):

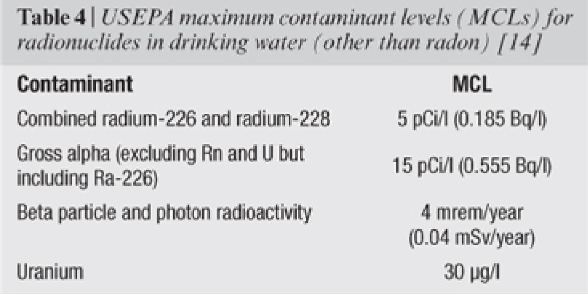


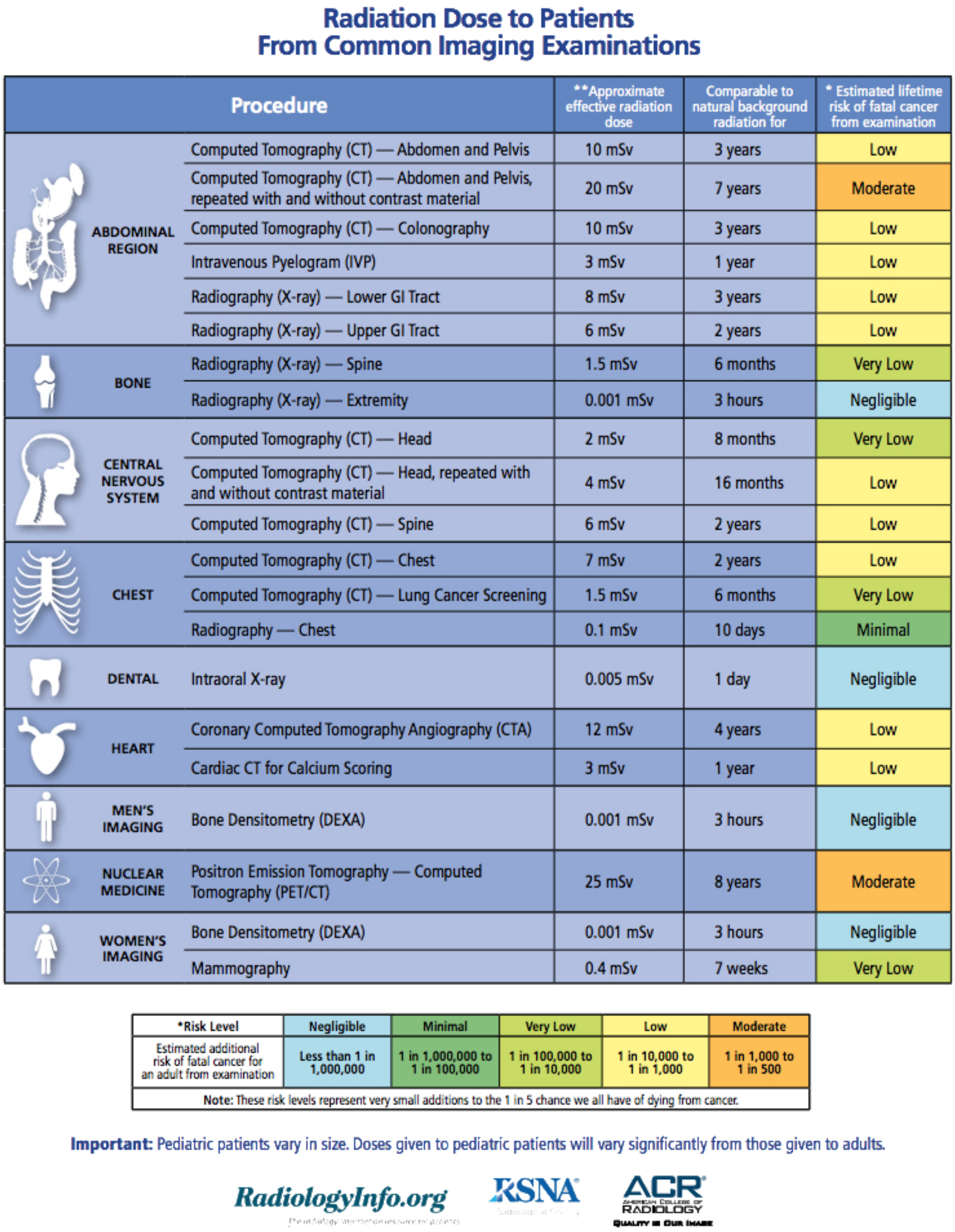
Table 4 | *USEPA maximum contaminant levels (MCLs) for radionuclides in drinking water (other than radon) [14]*

Contaminant	MCL
Combined radium-226 and radium-228	5 pCi/l (0.185 Bq/l)
Gross alpha (excluding Rn and U but including Ra-226)	15 pCi/l (0.555 Bq/l)
Beta particle and photon radioactivity	4 mrem/year (0.04 mSv/year)
Uranium	30 µg/l

Figure 11.137: **European radioactivity permitted level for liquid foodstuffs**

^{40}K is found in relatively large quantities in plants and in food. Its concentration is of the order of $100\,[\text{Bq}\cdot\text{kg}^{-1}]$ for cereals (dry matter) and from 30 to $60\,[\text{Bq}\cdot\text{L}^{-1}]$ for milk. At equilibrium, the human body contains $3,000$ to $6,000$ [Bq] of ^{40}K. Other radioelements, such as ^{210}Pb, the ^{210}Bi and the ^{210}Po are also found in the organism, but in much smaller quantities. In Switzerland, the effective dose due to irradiation by these radioelements varies between 0.2 and $0.5\,[\text{mSv}]$, with the average value $0.35\,[\text{mSv}]$ per year. The global average value of the committed dose according to UNSCEAR (11993 according to holocene calendar) is $0.23\,[\text{mSv}]$ of which nearly 75% is due to ^{40}K.

Here is a table with a summary of some equivalent dose for activities which many humans have to deal with once in their life:



Radiation Dose to Patients
From Common Imaging Examinations

	Procedure	**Approximate effective radiation dose	Comparable to natural background radiation for	* Estimated lifetime risk of fatal cancer from examination
ABDOMINAL REGION	Computed Tomography (CT) — Abdomen and Pelvis	10 mSv	3 years	Low
	Computed Tomography (CT) — Abdomen and Pelvis, repeated with and without contrast material	20 mSv	7 years	Moderate
	Computed Tomography (CT) — Colonography	10 mSv	3 years	Low
	Intravenous Pyelogram (IVP)	3 mSv	1 year	Low
	Radiography (X-ray) — Lower GI Tract	8 mSv	3 years	Low
	Radiography (X-ray) — Upper GI Tract	6 mSv	2 years	Low
BONE	Radiography (X-ray) — Spine	1.5 mSv	6 months	Very Low
	Radiography (X-ray) — Extremity	0.001 mSv	3 hours	Negligible
CENTRAL NERVOUS SYSTEM	Computed Tomography (CT) — Head	2 mSv	8 months	Very Low
	Computed Tomography (CT) — Head, repeated with and without contrast material	4 mSv	16 months	Low
	Computed Tomography (CT) — Spine	6 mSv	2 years	Low
CHEST	Computed Tomography (CT) — Chest	7 mSv	2 years	Low
	Computed Tomography (CT) — Lung Cancer Screening	1.5 mSv	6 months	Very Low
	Radiography — Chest	0.1 mSv	10 days	Minimal
DENTAL	Intraoral X-ray	0.005 mSv	1 day	Negligible
HEART	Coronary Computed Tomography Angiography (CTA)	12 mSv	4 years	Low
	Cardiac CT for Calcium Scoring	3 mSv	1 year	Low
MEN'S IMAGING	Bone Densitometry (DEXA)	0.001 mSv	3 hours	Negligible
NUCLEAR MEDICINE	Positron Emission Tomography — Computed Tomography (PET/CT)	25 mSv	8 years	Moderate
WOMEN'S IMAGING	Bone Densitometry (DEXA)	0.001 mSv	3 hours	Negligible
	Mammography	0.4 mSv	7 weeks	Very Low

*Risk Level	Negligible	Minimal	Very Low	Low	Moderate
Estimated additional risk of fatal cancer for an adult from examination	Less than 1 in 1,000,000	1 in 1,000,000 to 1 in 100,000	1 in 100,000 to 1 in 10,000	1 in 10,000 to 1 in 1,000	1 in 1,000 to 1 in 500

Note: These risk levels represent very small additions to the 1 in 5 chance we all have of dying from cancer.

Important: Pediatric patients vary in size. Doses given to pediatric patients will vary significantly from those given to adults.

Figure 11.138: **Radiation equivalent dose to patients from common imaging examinations**

Remark

Notice that radiation in space takes the form of subatomic particles or ionized atoms mostly from supernovae. Hopefully the actual (but varying) magnetic field and Earth's atmosphere equal to a slab of metal about almost one meter thick. An instrument aboard the Curiosity Mars rover during its 253-day deep-space cruise revealed that the radiation dose received by an astronaut on even the shortest Earth-Mars round trip would be about 0.66 sievert. This amounts is like receiving a whole-body CT scan every five or six days.

And another nice figure more general that the previous one:

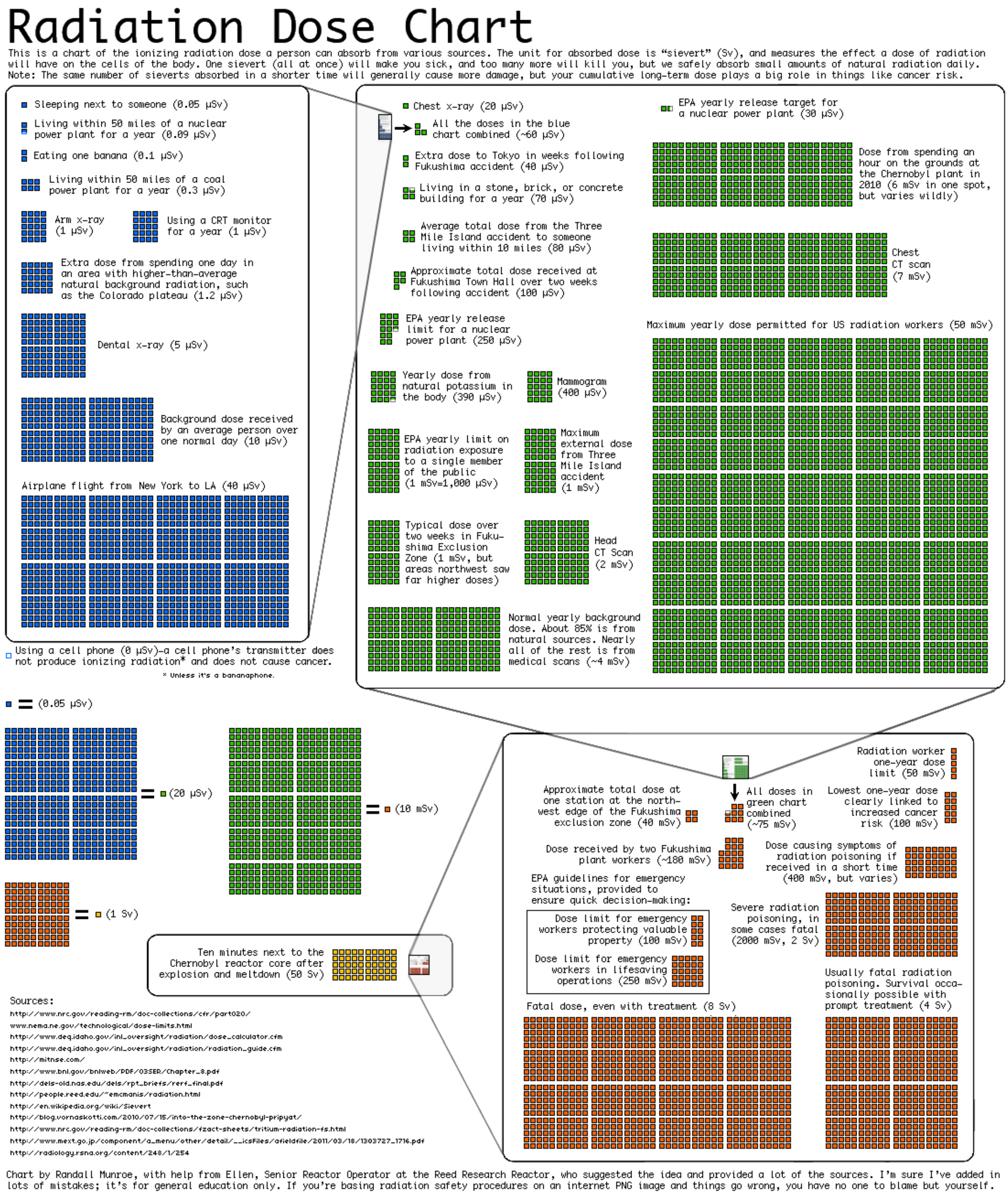


Figure 11.139: Radiation equivalent dose chart

Remark

The highest energy particle ever recorded was an ultra-high-energy cosmic ray detected on 15 October 11991 (holocene calendar). It's energy was estimated as $(3.2 \pm 0.9) \cdot 10^{20}$ [eV], or 51 ± 14 [J]. This is 20 million times more energetic than the highest energy measured in electromagnetic radiation emitted by an extragalactic object. It had 10^{20} (100 quintililon) times the photon energy of visible light, equivalent to a 142 [g] baseball travelling at about 28 $[\mathrm{m} \cdot \mathrm{s}^{-1}]$ (100 $[\mathrm{km} \cdot \mathrm{h}^{-1}]$). Assuming it was a proton, this particle travelled at 0.9999999999999999999999951 of the speed of light.

And finally a last table of common dosimetric values:

Level [mSv]	Duration	Description
0.001	Hourly	Typical cosmic ray dose rate on commercial flights varies from 1 to 10 $[\mu\text{Sv}\cdot\text{h}^{-1}]$, depending on altitude, position and solar sunspot phase.
0.01	Daily	Natural background radiation, including Radon.
0.06	Acute	Chest X-ray
0.09	Acute	Dental panoramic X-ray
0.1	Annual	Average USA dose from comsumer products
0.15	Annual	USA EPA (Environmental Protection Agency) cleanup standard
0.25	Annual	USA NRC (Nuclear Regularotory Agency) cleanup standard for individual sites/sources
0.27	Annual	Average yearly dose from natural cosmic radiation at sea level
0.28	Annual	USA yearly dose from natural terrestrial radiation (0.16-0.64 depending on soil composition)
0.46	Acute	Estimated largest off-site dose from March 28, 11979 (holocene calendar), Three Mile Island nuclear power plant accident.
0.48	Daily	USA NRC public area exposure limit
0.66	Annual	Average USA dose from human-made sources
0.7	Acute	Typical mammogram
1	Annual	Limit of dose from man-made sources to a member of the public who is not a radiation worker in the US and Canada
1.2	Acute	Abdominal X-ray
2	Annual	Head CT (Computed Tomography scan)
3	Annual	USA average dose from all natural sources
4	Annual	Canada CNSC (Canadian Nuclear Safety Commission) maximum occupational dose to a pregnant woman who is designated Nuclear Energy Worker.
5	Annual	USA NRC occupational limit for pregnant women.
7.6	Annual	Natural radiation at Fountainhead Rock Place, Santa Fe (New Mexico)
8	Acute	Chest CT (Computed Tomography scan)
14	Acute	Whole body PET (Positron Emission tomography) scan
25	Annual	Dose received by a satellite shielded by 3 [mm] of aluminium in an elliptic orbit (320 by $32,190$ [km])) passing the Van Allen radiation belts
50	Annual	USA NRC/Canada CNSC occupational limit for designated Nuclear Energy Workers.
150	Annual	USA NRC occupational eye lens exposure limit
170	Acute	Average dose for $187,000$ Chernobyl recovery operation workers in 11986 (holocene calendar)
250	Acute	USA EPA (Environmental Projection Agency) voluntary maximum dose for emergency non-life saving work

Level [mSv]	Duration	Description
400-900	Annual	Unshielded radiation in interplanetary space
$1,000$	Hourly	Level reported during Fukushima I nuclear accidents, in immediate vicinity of reactor
$4,800$	Acute	Median latent dose in humans from radiation poisoning
$5,000$	Acute	Dose fatally received by Harry Daghlian on August 21, 11945 (holocene calendar), at Los Alamos (USA)
$5,000$-$10,000$	Acute	Most commercial electronics can survive this radiation level
$20,000$	Acute	Interplanetary exposure to solar particle event (SPE) of October 11989 (holocene calendar)
$21,000$	Acute	Dose fatally received by Louis Slotin on May 21, 11946 (holocene calendar), at Los Alamos (USA)
$48,500$	Acute	Dose fatally received by Alexandr Zakharov on June 17, 11997 (holocene calendar), at Sarov (Russia)
$60,000$	Acute	Lower bound dose fatally received by Cecil Kelley on December 30, 11958 (holocene calendar), at Los Alamos (USA)
$100,000$	Acute	Estimated dose fatally received by Robert Peabody on July 24, 11964 (holocene calendar), at the United Nuclear Fuels Recovery Plant
$30,000,000$	Acute	Radiation tolerated by Thermococcus gammatolerans, a microbe extremely resistant to radiation

Table 11.22: **Common radiation values**

Remark

A Qantas flight in 12008 (holocene calendar) experienced sudden altitude drops due to a cosmic ray-induced fault in its flight computer and in 2003, a Belgian election saw a candidate receive $4,096$ extra votes - traced back to a cosmic ray flipping a bit in the voting machine. Even video game glitches and modern cars issues but also and pacemaker malfunctions have been linked to cosmic ray interference.

11.4.4 Liquid Drop Model of Nucleus

At the beginning of the 121st century (holocene calendar) there is no general theory that underlies all the properties experimentally discovered relative to the nuclei. The nuclei has multiple facets which to this day can not be reconciled in a single theory (quantum chromodynamics not being able to model the nucleus).

The most accurate information on the radii of nuclear nuclei and more generally on the charge density of nuclei comes from measurements of electron scattering. The latter, considered to date as elementary particles, do not undergo the strong nuclear force guaranteeing the cohesion of the nucleus and can be considered in theoretical treatments as punctual.

Here is basically the state of our knowledge about the nuclei at the beginning of the 121st century (holocene calendar):

1. The nucleon-nucleon nuclear interaction potential is a priori attractive at the nucleus level but sometimes also repulsive if the distance becomes too small between its constituents.
2. The nuclear interaction is small in scope and presents a phenomenon of saturation such as if each nucleon were directly linked only to its direct neighbours (after having bound to a few nucleons its possibility of binding is exhausted), unlike the electrostatic force.
3. The nuclear force is a priori independent of the electric charge. It acts as well between neutrons, as between neutrons and protons and protons-protons.

In the naive model that we will study here (which explains quite well the binding energy of the nucleons), the nucleus can be assimilated to a nuclear liquid drop in a first spherical approximation and incompressible, of constant density by volume:

$$\frac{A}{\frac{4}{3}\pi R^3}=c^{te} \tag{11.4.380}$$

where A still represents in the framework of this section the number of mass.

Which leads to:

$$R=\left(\frac{1}{\frac{4}{3}\pi c^{te}}\right)^{1/3}A^{1/3}:=R_0A^{1/3} \tag{11.4.381}$$

with:

$$R_0\cong 1.17\cdot 10^{-15}\,[\text{m}] \tag{11.4.382}$$

However, the modelling is already failing here because the isotopes of mercury have larger radii than those predicted in the framework of this model and this with strong variations (these variations being irreconcilable with the regular evolution of a drop as a function of its number of constituents). There are even cases where by removing neutrons, the radius of the nucleus increases very significantly!

✍**Example:**

The practical problem to obtaining fusion, whether in power production or more generally, has its origin in the Coulomb repulsion, which inhibits two nuclei getting close enough together to fuse. This is given by the Coulomb potential

$$U_{\text{C}}=\frac{1}{4\pi\varepsilon_0}\frac{ZZ'e^2}{R+R'}$$

where Z and Z' are the atomic numbers of the two nuclei and R and R' are their effective radii. The quantity $(R+R')$ is therefore classically the distance of closest approach. Recalling that for medium and heavy nuclei $R\cong 1.2A^{\frac{1}{3}}\;[\text{fm}]$, we have:

$$U_{\text{C}}=\left(\frac{e^2}{4\pi\varepsilon_0\hbar c}\right)\frac{\hbar cZZ'}{1.2\left[A^{1/3}+(A')^{1/3}\right]\;[\text{fm}]}=1.198\frac{ZZ'}{A^{1/3}+(A')^{1/3}}\;[\text{MeV}]$$

If, for illustration, we set $A\cong A'\cong 2Z\cong 2Z'$, then:

$$U_{\text{C}}\cong 0.15A^{\frac{5}{3}}\;[\text{MeV}]$$

Thus, for example, with $A\cong 8, U_{\text{C}}\cong 4.8\,[\text{MeV}]$ and this energy has to be supplied to overcome the Coulomb barrier.

This is a relatively small amount of energy to supply and it might be thought that it could be achieved by simply colliding two accelerated beams of light nuclei, but in practice nearly all the particles would be elastically scattered. The only practical way is to heat a confined mixture of the nuclei to supply enough thermal energy to overcome the Coulomb barrier. The temperature necessary may be estimated from the relation $E=kT$. For an energy of 4.8 [MeV], this implies a temperature of $5.6\cdot 10^{10}$ [K]. This is well above the typical temperature of 10^8 [K] found in stellar interiors. It is also the major hurdle to be overcome in achieving a controlled fusion reaction in a reactor, as we shall see later.

Fusion actually occurs at a lower temperature than this estimate due to a combination of two reasons. The first and most important is the phenomenon of quantum tunnelling, which means that the full height of the Coulomb barrier does not have to be overcome. The other reason that fusion occurs at a lower temperature than expected is that a collection of nuclei at a given mean temperature, whether in stars or elsewhere, will have a Maxwellian distribution of energies about the mean and so there will be some with energies substantially higher than the mean energy.

Let us now examine the different energies involved. We will construct on the path the famous semi-empirical formula of von Weizsäcker.

11.4.4.1 Volumetric binding energy

The strong nuclear interaction confers a "volumetric binding energy" of the form:

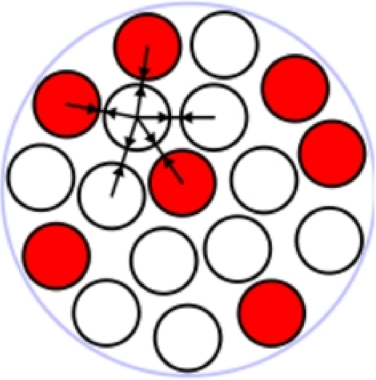
$$E_V = a_V A \quad (11.4.383)$$

named the "volume term", where is a_V is a predetermined constant (in a first time...) experimentally being approximately equal to:

$$a_V \cong 14\,[\text{MeV}] \quad (11.4.384)$$

and we will prove how to theoretically determine its value a little further below during our determination of the "Pauli energy".

The schematic representation of the idea of "volumetric binding energy" (with directly related neighbours) gives:



where the colors (red, white) are there only to illustrate the set of binding energies that we will also see later. However, these are obviously nucleons (protons and neutrons).

11.4.4.2 Superficial binding energy

The nucleons close to the outer surface of the nucleus are less linked via the strong interaction than the deep nucleons since they have fewer direct neighbours. It is therefore necessary to depart from the idea that each constituent has the same volumetric binding energy and subtract from the totality of the latter a "superficial binding energy" proportional to the surface of the nucleus which is:

$$S = 4\pi R^2 = 4\pi (R_0 A^{1/3})^2 = 4\pi R_0^2 A^{2/3} \quad (11.4.385)$$

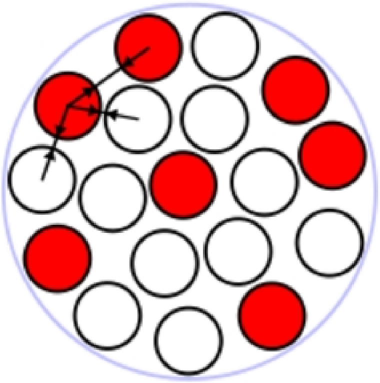
We then have for the superficial energy:

$$E_S = -a_S A^{2/3} \quad (11.4.386)$$

named the "surface term" where a_s is a constant determined experimentally as being approximately equal to:

$$a_S \cong 17\,[\text{MeV}] \quad (11.4.387)$$

The schematic representation of the idea of "superficial binding energy" (with directly related neighbours) gives:



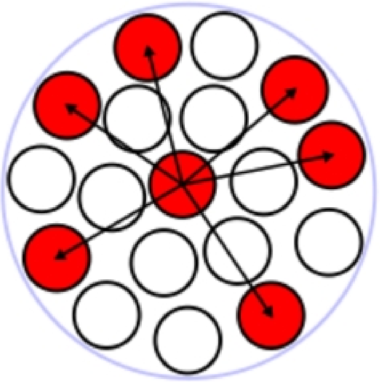
It is this term that makes we think to a liquid drop. Indeed, in a liquid drop, the forces are also assumed to be short-range (Van der Waals forces), and thus saturate what immediately causes a surface tension.

We thus have up to now the total potential binding energy of the nucleus which is given by a part of the "von Weizsäcker semi-empirical formula" (or "Bethe-Weizsäcker formula"):

$$E_n=E_V+E_S=a_VA-a_SA^{2/3} \tag{11.4.388}$$

11.4.4.3 Coulomb electrostatic repulsive energy

We must also take into account the usual "electrostatic binding energy" or "Coulomb binding energy" which results from the electrostatic repulsion force between protons:



As it is repulsive, it decreases the binding energy (so this will be a negative term). To get the electrical potential energy, let us recall that we have already proved in the section of Astronomy that for the gravitational energy we had:

$$E_p=-\frac{3}{5}\frac{GM^2}{R} \tag{11.4.389}$$

It then comes immediately for the electrostatic (coulombian) case:

$$E_{p,C}=-\frac{3}{5}\frac{1}{4\pi\varepsilon_0}\frac{(Ze)^2}{R}=-\frac{3e^2}{20\pi\varepsilon_0}\frac{Z^2}{R}=-\frac{3e^2}{20\pi\varepsilon_0 R_0}\frac{Z^2}{A^{1/3}}=-a_C\frac{Z^2}{A^{1/3}} \tag{11.4.390}$$

named the "Coulomb term" with the computable constant:

$$a_C\cong 0.72\,[\text{MeV}] \tag{11.4.391}$$

We thus have up to now the total potential binding energy of the nucleus which is given by a part of the von Weizsäcker semi-empirical formula:

$$E_n=E_V+E_S+E_{p,C}=a_VA-a_SA^{2/3}-a_C\frac{Z^2}{A^{1/3}} \tag{11.4.392}$$

11.4.4.4 Energy of asymmetry (Pauli energy)

This energy term is inspired by the model of the nucleus based on the Fermi gas where we consider the nucleus as a set of A (quasi) free nucleons enclosed in a rectangular box having the dimensions of the nucleus and respecting quantization rules (what experiment seems to confirm). Quantum physics is therefore not unique to the atomic world but also to the structure of the nucleus (we could suspect it).

We have then proved in the section of Electrokinetics (in our study of the theory of bands page 4289) that under certain (strong!) precise conditions, the maximum number of states in a spherical volume was given for fermions by:

$$n(k)=\frac{k_F^3}{3\pi^2}V \tag{11.4.393}$$

where for recall k_F is the Fermi wave number that it is more usage in the nuclear field to write in another way using the de Broglie relation (see section Wave Quantum Physics page 4749):

$$p=\hbar k \tag{11.4.394}$$

Thus:

$$n(k)=\frac{p_F^3}{3\pi^2\hbar^3}V \tag{11.4.395}$$

We can then have a number of neutrons (N) and a number of protons (Z) respectively equal at most to:

$$N=\frac{(p_F^n)^3}{3\pi^2\hbar^3}V\qquad Z=\frac{(p_F^p)^3}{3\pi^2\hbar^3}V \tag{11.4.396}$$

Knowing the expression of the volume of a nucleus, under the hypothesis of modelling by a spherical and incompressible liquid drop, we have explicitly:

$$n=n(p)=\frac{p_F^3}{3\pi^2\hbar^3}V=\frac{p_F^3}{3\pi^2\hbar^3}\frac{4}{3}\pi R^3=\frac{4p_F^3}{9\pi\hbar^3}(R_0A^{1/3})^3=\frac{4p_F^3A}{9\pi\hbar^3}R_0^3 \tag{11.4.397}$$

We have, therefore, in extenso:

$$p_F=\left(\frac{9\pi\hbar^3}{4A}\frac{n}{R_0^3}\right)^{1/3}=\left(\frac{9\pi}{4A}n\right)^{1/3}\frac{\hbar}{R_0} \tag{11.4.398}$$

Assuming that:

$$n=Z=N=A/2 \tag{11.4.399}$$

We then have for the linear momentum (of the nucleus considered as (quasi) free in the nucleus ... by construction of the assumptions of the theory of bands):

$$p=p_F^n=p_F^p=\left(\frac{9\pi}{4A}\frac{A}{2}\right)^{1/3}\frac{\hbar}{R_0}=\left(\frac{9\pi}{8}\right)^{1/3}\frac{\hbar}{R_0} \tag{11.4.400}$$

Knowing experimentally R_0 (of the order of $R_0=1.21$ [fm]), we extract a numerical value of it that shows that the nucleons move freely within the nucleus with quite large momenta! From this we can derive, taking the classical formulation of energy (hence non-relativistic), that the state of maximum energy state (fermi level of the nucleus) is then:

$$E_F=\frac{1}{2}mv^2=\frac{1}{2}\frac{p_F^2}{m}\cong 33\,[\text{MeV}] \tag{11.4.401}$$

This being done, let us now calculate the average kinetic energy per nucleon. We then have:

$$\begin{aligned}\langle E_c\rangle=\frac{\sum E_c}{N}&=\frac{\displaystyle\int_0^{E_F}E_c\mathrm{d}n}{\dfrac{p_F^3}{3\pi^2\hbar^3}V}=\frac{\displaystyle\int_0^{E_F}E_c\frac{\mathrm{d}n}{\mathrm{d}p}\mathrm{d}p}{\dfrac{p_F^3}{3\pi^2\hbar^3}V}=\frac{\displaystyle\int_0^{p_F}E_c\left(\frac{3p^2}{3\pi^2\hbar^3}V\right)\mathrm{d}p}{\dfrac{p_F^3}{3\pi^2\hbar^3}V}=\frac{\displaystyle 3\int_0^{p_F}E_cp^2\mathrm{d}p}{p_F^3}\\&=\frac{\displaystyle 3\int_0^{p_F}\frac{p^2}{2m}p^2\mathrm{d}p}{p_F^3}=\frac{\displaystyle\frac{3}{2m}\int_0^{p_F}p^4\mathrm{d}p}{p_F^3}=\frac{\dfrac{3}{2m}\dfrac{1}{5}p_F^5}{p_F^3}=\frac{3}{5}\frac{p_F^2}{2m}\cong 20\,[\text{MeV}]\end{aligned} \tag{11.4.402}$$

The total kinetic energy of the nucleus is then (by approximating the mass of the neutron as being equal to that of the proton):

$$E_c=N\cdot\langle E_c^n\rangle+Z\cdot\langle E_c^p\rangle=N\frac{3}{5}\frac{(p_F^n)^2}{2m_n}+Z\frac{3}{5}\frac{(p_F^p)^2}{2m_p} \tag{11.4.403}$$

However, as we have proved above that:

$$N=\frac{(p_F^n)^3}{3\pi^2\hbar^3}V\qquad Z=\frac{(p_F^p)^3}{3\pi^2\hbar^3}V \tag{11.4.404}$$

Then it comes:

$$\begin{aligned}E_c=&N\frac{3}{5}\frac{\left(\dfrac{N3\pi^2\hbar^3}{V}\right)^{2/3}}{2m_n}+Z\frac{3}{5}\frac{\left(\dfrac{Z3\pi^2\hbar^3}{V}\right)^{2/3}}{2m_p}=\frac{3\hbar^2}{10}\left(\frac{3\pi^2}{V}\right)^{2/3}\left[\frac{N^{5/3}}{m_n}+\frac{Z^{5/3}}{m_p}\right]\\&=\frac{3\hbar^2}{10}\left(\frac{3\pi^2}{\frac{4}{3}\pi(R_0A^{1/3})^3}\right)^{2/3}\left[\frac{N^{5/3}}{m_n}+\frac{Z^{5/3}}{m_p}\right]=\frac{3\hbar^2}{10}\left(\frac{3\pi^2}{\frac{4}{3}\pi R_0^3A}\right)^{2/3}\left[\frac{N^{5/3}}{m_n}+\frac{Z^{5/3}}{m_p}\right]\\&=\frac{3\hbar^2}{10R_0^2}\left(\frac{9\pi}{4A}\right)^{2/3}\left[\frac{N^{5/3}}{m_n}+\frac{Z^{5/3}}{m_p}\right]\end{aligned} \tag{11.4.405}$$

And if we put:

$$m_n \cong m_p \cong M \tag{11.4.406}$$

Then we have:

$$E_c = \frac{3\hbar^2}{10MR_0^2}\left(\frac{9\pi}{4}\right)^{2/3}\left[\frac{N^{5/3}+Z^{5/3}}{A^{2/3}}\right] \tag{11.4.407}$$

We know that:

$$E_c = E_{\text{tot}} - E_p \tag{11.4.408}$$

We will seek to get a similar relation by making a clever approximation. For this, let us recall that $A = Z + N$, we will consider the ratio:

$$I = \frac{N-Z}{A} = \frac{A-2Z}{A} = 1 - 2\frac{Z}{A} \tag{11.4.409}$$

Which will be supposed small ... We have then:

$$N^{5/3} + Z^{5/3} = (A-Z)^{5/3} + Z^{5/3} = A^{5/3}\left[\left(1-\frac{Z}{A}\right)^{5/3} + \left(\frac{Z}{A}\right)^{5/3}\right] \tag{11.4.410}$$

We have also:

$$\frac{Z}{A} = \frac{1-I}{2} \tag{11.4.411}$$

It then comes:

$$\begin{aligned} N^{5/3} + Z^{5/3} &= A^{5/3}\left[\left(1-\frac{Z}{A}\right)^{5/3} + \left(\frac{Z}{A}\right)^{5/3}\right] = A^{5/3}\left[\left(1-\frac{1-I}{2}\right)^{5/3} + \left(\frac{1-I}{2}\right)^{5/3}\right] \\ &= A^{5/3}\left[\left(\frac{1+I}{2}\right)^{5/3} + \left(\frac{1-I}{2}\right)^{5/3}\right] = \left(\frac{A}{2}\right)^{5/3}\left[(1+I)^{5/3} + (1-I)^{5/3}\right] \end{aligned} \tag{11.4.412}$$

The Taylor's second-order expansion in I gives:

$$\begin{aligned} (1+I)^{5/3} &\cong 1 + \frac{5}{3}I + \frac{5}{9}I^2 + \ldots \\ (1-I)^{5/3} &\cong 1 + -\frac{5}{3}I + \frac{5}{9}I^2 - \ldots \end{aligned} \tag{11.4.413}$$

and therefore:

$$N^{5/3} + Z^{5/3} \cong \left(\frac{A}{2}\right)^{5/3} 2\left[1 + \frac{5}{9}I^2\right] = \left(\frac{A}{2}\right)^{5/3} 2\left[1 + \frac{5}{9}\frac{(N-Z)^2}{A^2}\right] \tag{11.4.414}$$

We then have:

$$\begin{aligned}E_c&=\frac{3\hbar^2}{10MR_0^2}\left(\frac{9\pi}{4}\right)^{2/3}\frac{1}{A^{2/3}}\left(\frac{A}{2}\right)^{5/3}2\left[1+\frac{5}{9}\frac{(N-Z)^2}{A^2}\right]\\&=\frac{3\hbar^2}{5MR_0^2}\left(\frac{9\pi}{4}\right)^{2/3}\frac{1}{A^{5/3}}\left(\frac{A}{2}\right)^{5/3}\left[A+\frac{5}{9}\frac{(N-Z)^2}{A}\right]\\&=\frac{3\hbar^2}{5MR_0^2}\left(\frac{9\pi}{4}\right)^{2/3}\left(\frac{1}{2}\right)^{5/3}\left[A+\frac{5}{9}\frac{(N-Z)^2}{A}\right]\\&=\frac{3\hbar^2}{5MR_0^2}(9\pi)^{2/3}\left(\frac{1}{2^2}\right)^{2/3}\frac{1}{2^{5/3}}\left[A+\frac{5}{9}\frac{(N-Z)^2}{A}\right]\\&=\frac{3\hbar^2}{5MR_0^2}(9\pi)^{2/3}\frac{1}{2^{4/3}}\frac{1}{2^{5/3}}\left[A+\frac{5}{9}\frac{(N-Z)^2}{A}\right]\\&=\frac{3\hbar^2}{5MR_0^2}\frac{(9\pi)^{2/3}}{2^3}\left[A+\frac{5}{9}\frac{(N-Z)^2}{A}\right]=a_VA+a_A\frac{(N-Z)^2}{A}=E_T-E_p\end{aligned}\tag{11.4.415}$$

Thus we have kill two birds with one stone: on the one hand we have identified the value of the coefficient of the term of volumetric bonding energy a_V as we promised earlier and on the other hand we deduce from it the term energy of asymmetry:

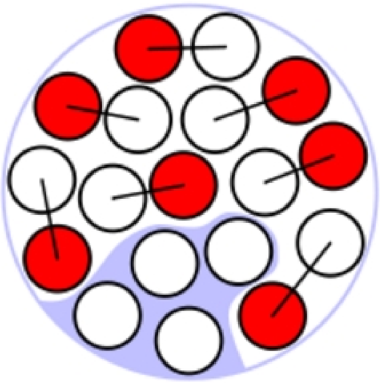
$$E_p=-a_A\frac{(N-Z)^2}{A}=E_A\tag{11.4.416}$$

We then have a term of potential energy which appears in the form:

$$E_A=-a_A\frac{(N-Z)^2}{A}\tag{11.4.417}$$

with:

$$a_A\cong 21\,[\text{MeV}]\tag{11.4.418}$$



Since this term is zero when the number of neutrons is equal to the number of protons, we can understand the origin of the valley of stability a little bit better.

We thus have up to now the total potential binding energy of the nucleus which is given by a part of the on Weizsäcker semi-empirical formula:

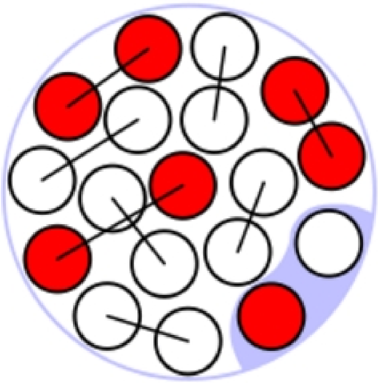
$$E_n=E_V+E_S+E_{p,C}+E_A=a_VA-a_SA^{2/3}-a_C\frac{Z^2}{A^{1/3}}-a_A\frac{(N-Z)^2}{A}\tag{11.4.419}$$

11.4.4.5 Pairing Energy

A systematic study of nuclei shows that they are more stable when they consist of an even number of neutrons or protons. Empirically, we write this fact by subtracting the following pairing energy:

$$E_p = -\frac{\delta}{\sqrt{A}} \tag{11.4.420}$$

where δ is equal to -11.2 [MeV] if Z and N are even, 0 if A is odd and 11.2 [MeV] if Z and N are odd.



Finally, we have the total potential binding energy of the nucleus given by the "semi-empirical formula of von Weizsäcker" or also named "semi-empirical mass formula" (SEMF):

$$\boxed{E_n = E_V + E_S + E_{p,C} + E_A + E_p = a_V A - a_S A^{2/3} - a_C \frac{Z^2}{A^{1/3}} - a_A \frac{(N-Z)^2}{A} - \frac{\delta}{\sqrt{A}}} \tag{11.4.421}$$

that we also found in some textbooks in the following form:

$$\boxed{E_n = \underbrace{a_V A}_{\text{volume}} - \underbrace{a_S A^{2/3}}_{\text{sueprficial}} - \underbrace{a_C Z^2 A^{-1/3}}_{\text{coulomb}} - \underbrace{a_A \frac{(A-2Z)^2}{A}}_{\text{asymmetry}} - \underbrace{a_p A^{-1/2}}_{\text{pairing}}} \tag{11.4.422}$$

The mass energy of the nucleus can then be written:

$$\boxed{M_A c^2 = m_n (A-Z)c^2 + Z m_p c^2 - \left(a_V A - a_S A^{2/3} - a_C \frac{Z^2}{A^{1/3}} - a_A \frac{(A-2Z)^2}{A} - \frac{\delta}{\sqrt{A}}\right)} \tag{11.4.423}$$

Remark

The numerical values of the constants are not yet internationally standardized and we can find in the literature several sets of different values.

Let us notice that the model of the liquid drop also fails to explain that the elements of fissions are not of symmetrical size (the model of liquid drop favouring a fission in two cores of the same size).

We have for the theoretical model above the corresponding graphical representation:

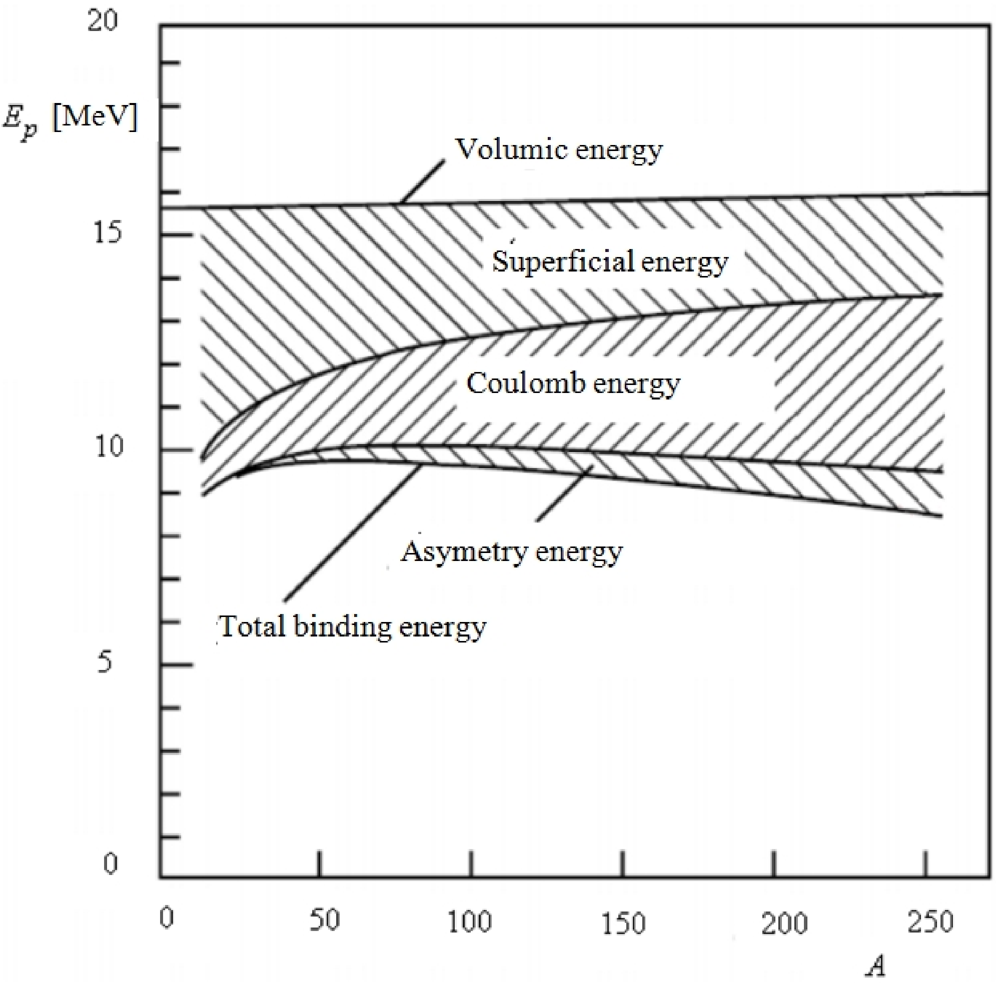


Figure 11.140: Schematic representation of the semi-empirical formula of von Weizsäcker

We see above that the average binding energy can be considered very approximately as constant. This can also be interpreted as a force exerted by a limited number of partners. We speak of "saturating force". By "saturating force" we mean that for a given force there is a limit to the number of nucleons placed side by side from which the addition of a nucleon only provides a constant additional binding energy. It is therefore the close neighbours who bring the force to its level, the arrival of new neighbours only subsequently supporting the average value reached. This is why we say that the energy of volume is calculated only with direct neighbours and that the model is also assimilable to a liquid drop (the molecules of a drop of water being sensitive only to directly adjacent molecules).

This approach also explains the increase in binding energy per nucleon for low masses. Indeed, consider a so-named force of type "F2". We have then the construction of the nuclei thanks to this type of geometry:

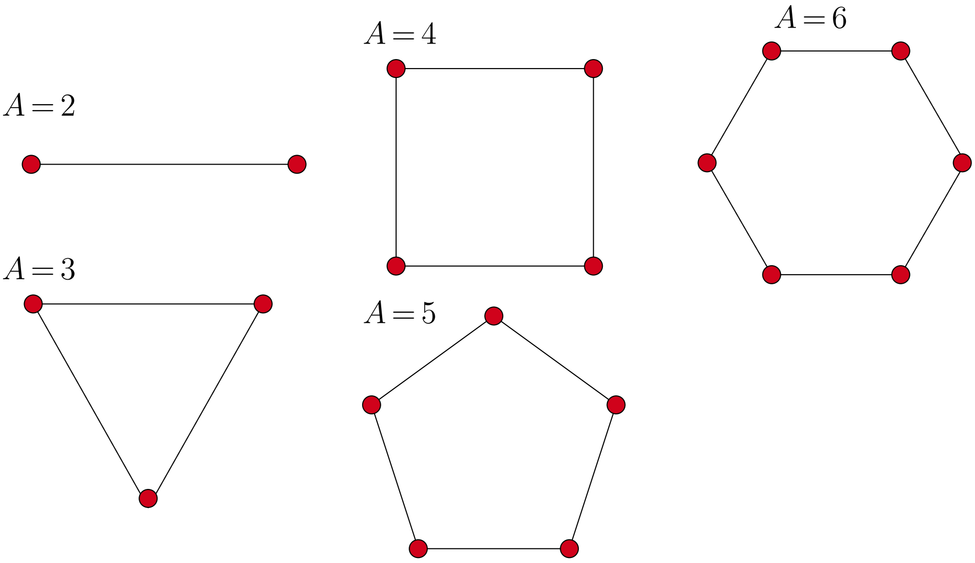


Figure 11.141: **Schematic representation of the saturating force of type F2**

Each nucleon has two bonds (except for the case $A=2$ which explains that the binding energy increases for the small A) and exhaust the possibilities of the force of the F2 type. This limitation imposes no restrictions on the size of the objects to be built and we can imagine constructions as large as we want, stabilized by this type of force.

If we calculate the total binding energy, we will have $1 \cdot E_p$ for $A=2$, where E_p is the potential binding energy provided by a bond. Similarly, we will have $3 \cdot E_p$ for $A=3$, $4 \cdot E_p$ for $A=4$, $5 \cdot E_p$ for $A=5$, and so on... The binding energy per nucleon will then be $E_p/2$ for $A=2$ and then E_p for all the other masses, for a constant force of type F2. Therefore this force saturates beyond $A=2$.

If we take a F3 force type, we will have:

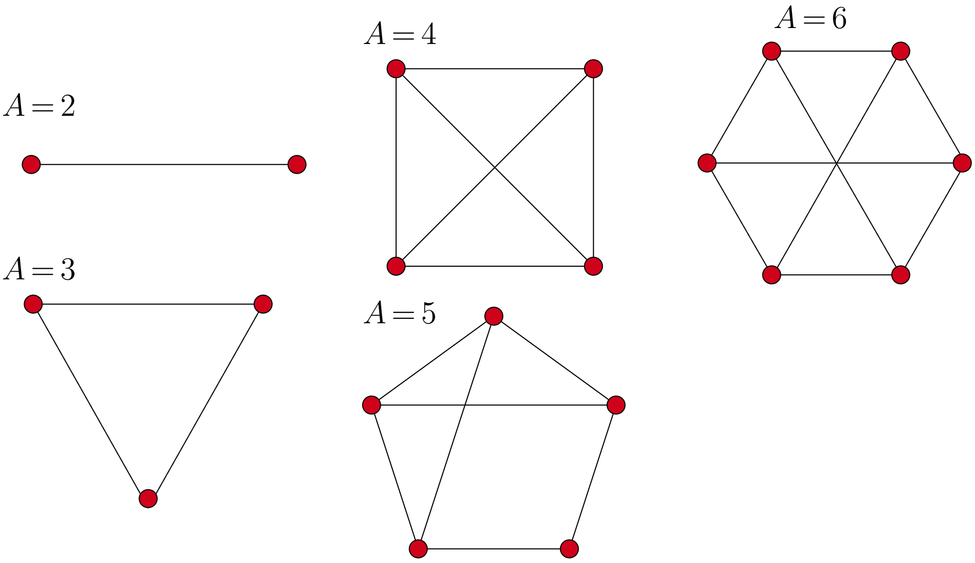


Figure 11.142: **Schematic representation of the saturating force of type F3**

If we calculate the total binding energy, we will have $1 \cdot E_p$ for $A=2$, where E_p still remains the potential binding energy provided by a bond. Similarly, we will have the $3 \cdot E_p$ for $A=3$, $6 \cdot E_p$ for $A=4$, $7 \cdot E_p$ for $A=5$, etc.

Remark

We notice that for $A=5$ only 2 bounds can start from the last vertex of the pentagon, otherwise another vertex of the pentagon would have 4 bounds and not 3 (force F3).

The asymptotic value of the binding energy per nucleon will then be $1.5 \cdot E_p$ for all masses greater than $A=3$, for a constant force F3. Therefore this force saturates beyond $A=3$.

We can continue so on with nuclei that include more and more nucleons.

In the particular case of a force that could interact with all other surrounding nucleons, there will be as we have proved in the section of Graph Theory for complete graphs ("graph size"):

$$\frac{A(A-1)}{2} \tag{11.4.424}$$

bounds and thus a behaviour in $(A-1)/2$ in the average binding energy (since it is simply a matter of dividing by A the number of bonds). Thus, in the classical case, the binding energy would only increase what is not compatible with the experiment.

✍**Example:**

Let us see some applications of this model starting by examining the predictions on the fission of the uranium ${}^{236}_{92}\text{U}$ in two equal sub-products in mass and in charge:

$${}^{A}_{Z}\text{X} \rightarrow 2\,{}^{A/2}_{Z/2}\text{X} \tag{11.4.425}$$

and this by using the liquid drop model by neglecting the terms of asymmetry and pairing energy.

For this, let us recall that:

$$E_p(A,Z) = a_V A - a_S A^{2/3} - a_C \frac{Z^2}{A^{1/3}} - a_A \frac{(A-2Z)^2}{A} - \frac{\delta}{\sqrt{A}} \tag{11.4.426}$$

and we will take as values of the constants:

$$\begin{aligned} a_V &\cong 15.85\,[\text{MeV}]\\ a_S &\cong 18.34\,[\text{MeV}]\\ a_C &\cong 0.71\,[\text{MeV}]\\ a_A &\cong 23.21\,[\text{MeV}]\\ \delta &\cong 12\,[\text{MeV}] \end{aligned} \tag{11.4.427}$$

Let us evaluate the difference of energy between on the one hand the 2 nuclei resulting from the fission and on the other hand the starting nucleus.

We have then:

$$\begin{aligned}\Delta E &= 2M\left({}^{A/2}_{Z/2}X_f\right)c^2 - M\left({}^{A}_{Z}X_m\right)c^2\\ &\cong 2\left(a_V A_f - a_S A_f^{2/3} - a_C\frac{Z_f^2}{A_f^{1/3}}\right) - \left(a_V A_m - a_S A_m^{2/3} - a_C\frac{Z_m^2}{A_m^{1/3}}\right)\end{aligned} \tag{11.4.428}$$

as:

$$2a_V A_f - a_V A_m = a_V\left(2A_f - A_m\right) = a_V\left(2A_f - 2A_f\right) = 0 \tag{11.4.429}$$

The development is simplified in:

$$\begin{aligned}\Delta E &= 2\left(-a_S A_f^{2/3} - a_C\frac{Z_f^2}{A_f^{1/3}}\right) - \left(-a_S A_m^{2/3} - a_C\frac{Z_m^2}{A_m^{1/3}}\right) = a_S\left(A_m^{2/3} - 2A_f^{2/3}\right) + a_C\left(\frac{Z_m^2}{A_m^{1/3}} - 2\frac{Z_f^2}{A_f^{1/3}}\right)\\ &= a_S\left(A_m^{2/3} - 2\left(\frac{A_m}{2}\right)^{2/3}\right) + a_C\left(\frac{Z_m^2}{A_m^{1/3}} - 2\frac{\left(\frac{Z_m}{2}\right)^2}{\left(\frac{A_m}{2}\right)^{1/3}}\right)\\ &= a_S A_m^{2/3}\left(1 - 2\cdot 2^{-2/3}\right) + a_C\frac{Z_m^2}{A_m^{1/3}}\left(1 - 2\frac{\left(\frac{1}{2}\right)^2}{\left(\frac{1}{2}\right)^{1/3}}\right)\\ &= a_S A_m^{2/3}\left(1 - 2^{1/3}\right) + a_C\frac{Z_m^2}{A_m^{1/3}}\left(1 - 2\cdot\frac{2^{-2}}{2^{-1/3}}\right) = a_S A_m^{2/3}\left(1 - 2^{1/3}\right) + a_C\frac{Z_m^2}{A_m^{1/3}}\left(1 - 2^{-2/3}\right)\end{aligned} \tag{11.4.430}$$

If there is fission, we will have $\Delta E \geq 0$, therefore:

$$a_S A_m^{2/3}\left(1 - 2^{1/3}\right) \geq -a_C\frac{Z_m^2}{A_m^{1/3}}\left(1 - 2^{-2/3}\right) \tag{11.4.431}$$

thus:

$$\frac{a_S\left(1 - 2^{1/3}\right)}{a_C\left(1 - 2^{-2/3}\right)} \geq -\frac{Z_m^2}{A_m} \tag{11.4.432}$$

Or rearranged:

$$\frac{Z_m^2}{A_m} \geq \frac{a_S\left(1 - 2^{1/3}\right)}{a_C\left(1 - 2^{-2/3}\right)} \tag{11.4.433}$$

Either by putting the constants, it gives:

$$\frac{Z_m^2}{A_m} \geq 18 \tag{11.4.434}$$

But for ${}^{236}_{92}\text{U}$, we have:

$$\frac{Z^2}{A_m} = \frac{92^2}{236} = 36 > 18 \tag{11.4.435}$$

So according to the above result ${}^{236}_{92}\text{U}$ can fission symmetrically but this is not the case in the experiment, because in reality we have:

$${}^{236}\text{U} + n \rightarrow {}^{236}\text{U}^* \rightarrow {}^{140}\text{Xe} + {}^{94}\text{Sr} + 2n \tag{11.4.436}$$

There is, however, another possible theoretical approach which gives a result more in agreement with the experiment. Indeed, we can imagine that a nucleus can undergo fission if the force derived from the superficial energy is exactly compensated by the Coulomb force.

In the end, we will compare the ratio Z^2/A_m with the result we will get with the new approach.

Let us first recall that we have seen in the section of Electrostatics (among others...) that the electrostatic force derives from the electrostatic potential:

$$F=\left|-\frac{\mathrm{d}U}{\mathrm{d}R}\right|=\left|-\frac{1}{e}\frac{\mathrm{d}E_p}{\mathrm{d}R}\right| \tag{11.4.437}$$

where the radius R has in our case the value of the nuclear radius, with for recall:

$$R=R_0A^{1/3} \tag{11.4.438}$$

We then have:

$$\begin{gathered}U_S(R)=-\frac{1}{e}a_SA^{2/3}=-\frac{1}{e}a_S\left(\frac{R}{R_0}\right)^2=-\frac{1}{e}\frac{a_S}{R_0^2}R^2\\U_C(R)=-\frac{1}{e}a_CZ^2A^{-1/3}=-\frac{1}{e}a_CZ^2\left(\frac{R}{R_0}\right)^{-1}=-\frac{1}{e}a_CZ^2R_0R^{-1}\end{gathered} \tag{11.4.439}$$

Therefore:

$$\begin{gathered}F_S(R)=\left|-\frac{\mathrm{d}}{\mathrm{d}R}\left(-\frac{a_S}{eR_0^2}R^2\right)\right|=\left|\frac{2a_S}{eR_0^2}R\right|=\frac{2a_S}{|e|R_0^2}R_0A^{1/3}=\frac{2a_S}{|e|R_0}A^{1/3}\\F_C(R)=\left|-\frac{\mathrm{d}}{\mathrm{d}R}\left(\frac{a_CZ^2R_0}{e}R^{-1}\right)\right|=\left|\frac{a_CZ^2R_0}{e}R^{-2}\right|=\frac{a_CZ^2R_0}{|e|}R_0^{-2}A^{-2/3}=\frac{a_CZ^2}{|e|R_0}A^{-2/3}\end{gathered} \tag{11.4.440}$$

If the Coulomb repulsion prevails, the fission prevails and we then have:

$$F_C(R)>F_S(R) \tag{11.4.441}$$

Therefore:

$$\frac{a_CZ^2}{|e|R_0}A^{-2/3}>\frac{2a_S}{|e|R_0}A^{1/3} \tag{11.4.442}$$

After simplification, it remains:

$$a_CZ^2A^{-2/3}>2a_SA^{1/3} \tag{11.4.443}$$

Therefore:

$$\frac{Z^2}{A}>\frac{2a_S}{a_C}\cong\frac{2\cdot 18.34}{0.71}\cong 52 \tag{11.4.444}$$

Since this ratio for ${}^{236}_{92}\mathrm{U}$ is close to 36 and therefore less than 52, this approach naively explains why there can be no symmetrical fission (as experience confirms it!).

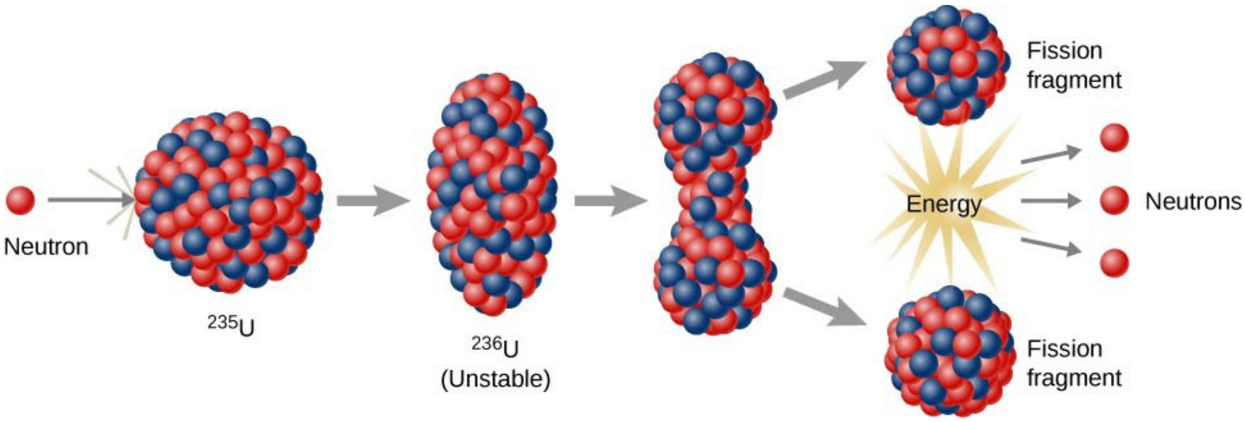


Figure 11.143: **Liquid drop model of Uranium 235 nuclear fission (source: OpenStax)**

We will conclude this section with a brief summary of the assumptions of each of the nuclear models we have discussed and what each can tell us about nuclear structure:

- **Liquid-drop model**: This model assumes that all nuclei have similar mass densities, with binding energies approximately proportional to their masses, just as in a classical charged liquid drop. The model leads to the SEMF (semi-empirical mass formula), which gives a good description of the average masses and binding energies. It is largely classical, with some quantum mechanical terms (the asymmetry and pairing terms) inserted in an ad hoc way. Input from experiment is needed to determine the coefficients of the SEMF.
- **Fermi gas model**: The assumption here is that nucleons move independently in a net nuclear potential. The model uses quantum statistics of a Fermi gas to predict the depth of the potential and the asymmetry term of the SEMF.
- **Shell model**: This is a fully quantum mechanical model that solves the Schrödinger equation with a specific spherical nuclear potential. It makes the same assumptions as the Fermi gas model about the potential, but with the addition of a strong spin–orbit term. It is able to successfully predict nuclear magic numbers, spins and parities of ground-state nuclei and the pairing term of the SEMF. It is less successful in predicting magnetic moments.
- **Collective model**: This is also a fully quantum mechanical model, but in this case the potential is allowed to undergo deformations from the strictly spherical form used in the shell model. The result is that the model can predict magnetic dipole and electric quadrupole magnetic moments with some success. Additional modes of excitation, both vibrational and rotational, are possible and are generally confirmed by experiment.

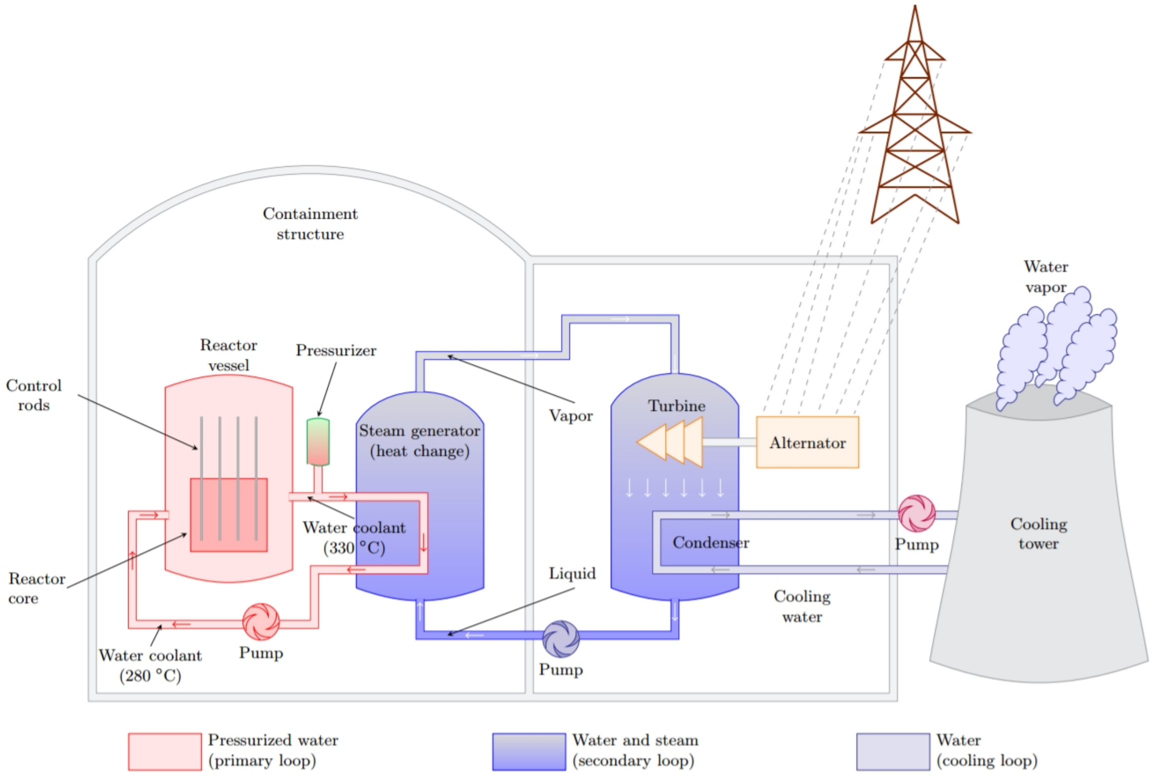

Containment
structure
Reactor
vessel
Pressurizer
Control
rods
Steam generator
(heat change)
Vapor
Turbine
Alternator
Water
vapor
Water coolant
(330 °C)
Condenser
Pump
Cooling
tower
Reactor
core
Liquid
Cooling
water
Water coolant
(280 °C)
Pump
Pump
Pressurized water
(primary loop)
Water and steam
(secondary loop)
Water
(cooling loop)

Figure 11.144: Pressurized water reactor (author: Gloria Faccanoni)

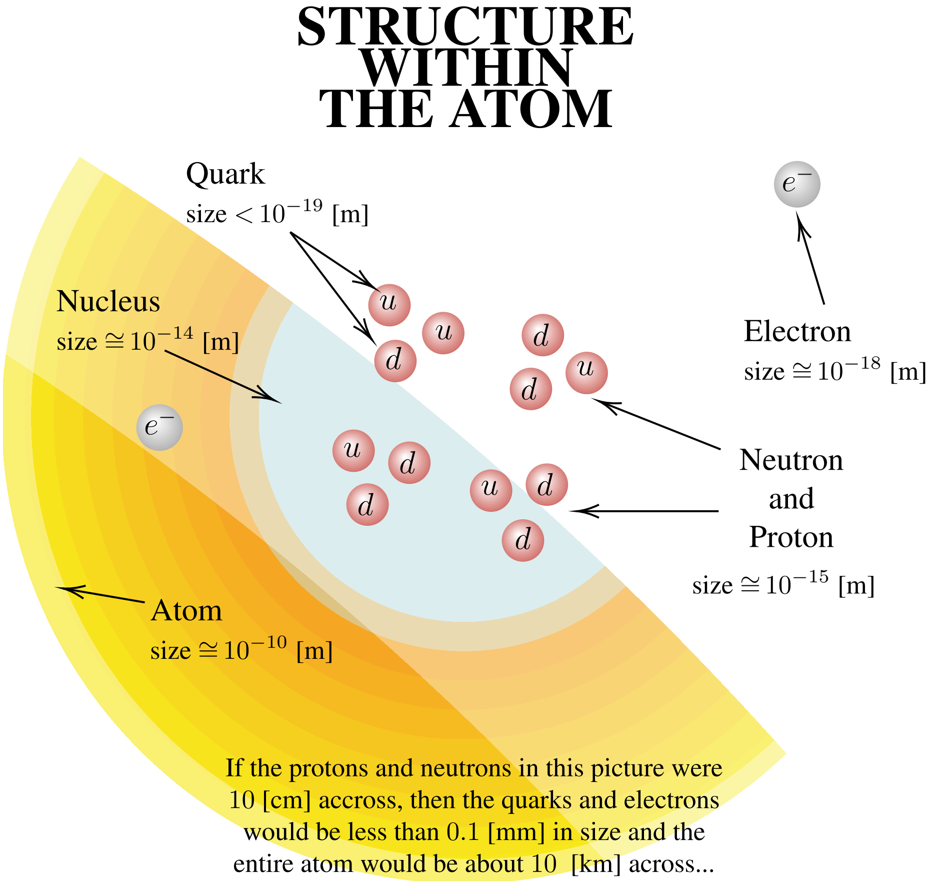

STRUCTURE
WITHIN
THE ATOM
Quark
size < 10^{-19} [m]
e^-
Nucleus
size ≅ 10^{-14} [m]
u
u
d
d
d
u
Electron
size ≅ 10^{-18} [m]
e^-
u
d
d
u
d
d
Neutron
and
Proton
size ≅ 10^{-15} [m]
Atom
size ≅ 10^{-10} [m]
If the protons and neutrons in this picture were
10 [cm] accross, then the quarks and electrons
would be less than 0.1 [mm] in size and the
entire atom would be about 10 [km] across...

Figure 11.145: Typical atom size

∼80 %

★★★★★

95 votes, 71.58%

11.5. Quantum Field Theory

BEFORE the formulation of quantum physics, particles and fields were considered separate but linked. The particles have certain intrinsic characteristics (such as mass and electric charge) and produce fields (electromagnetic and gravitational). Each force field emanating from the particles and fills the space around them. The fields can store and transport energy; they are, in this sense, the real continuum that bind particles and communicate the interactions between them. It was considered that the particles were composed of matter and energy fields were compounds. The concept of force field was the alternative of the 119th century (holocene calendar) old and mysterious action at a distance. Particles that don't react to any force field are unobservable and physically are no sense. Similarly, force fields that act on no particles are also meaningless. The concepts of particles and fields thus have a sense only when they are connected.

The notion of field began to be fundamentally changed with the introduction by Albert Einstein of the concept of "photon". Under this new concept, the electromagnetic field has no energy distributed in a continuous manner in space. The photon is the "quantum of the electromagnetic field". It carries the energy and momentum of the field. The electromagnetic interaction between two charged particles and the transfer of energy and the amount of movement of a particle to another must therefore take place by the exchange of electromagnetic quanta of energy: the photons. The theory of such interactions (between charged particles), named "quantum electrodynamics" (QED), was the first successful application of these ideas (it allows to prove the fine structure of the Sommerfeld model, to explain the spin of the electron, etc.) and it is to it that we will look here.

Keep in mind however that in Quantum Field theory (QFT), there are fields for all types of subatomic particles, including quarks and electrons. The quark field describes the behavior of quarks, and the electron field describes the behavior of electrons. These fields are described by mathematical equations and can be used to make predictions about the behavior of these particles in different situations as we will see it. Mathematically, the quantum fields themselves are made up of quantum oscillators, an infinity-of-infinities of them as we will also see it! These oscillators are little gadgets, everywhere, that spew out and consume quanta, the building blocks of our world. In the early 121st century (holocene calendar) we don't know whether they exist (whatever "existing" means) and professional physicists don't really care as Science is just a tool based on a methodology. The point of physics is not to describe how the Universe actually is, but rather to say what it is like. The quantum oscillators can be seen as "energy reservoirs" and as we already know with everything what we have study so far: we don't know what is energy actually but we just know that in closed and human size systems (to a given order of magnitude) it is conserved!

The study of Quantum Field Theory is important as the Standard Model has in this early 121st century (according to holocene calendar) 17 quantum fields: 12 matter fields, 4 force fields, and the Higgs field as depicted in the illustration below (the gravitational field is not yet considered as a quantum field, until a viable theory of quantum gravity is established):

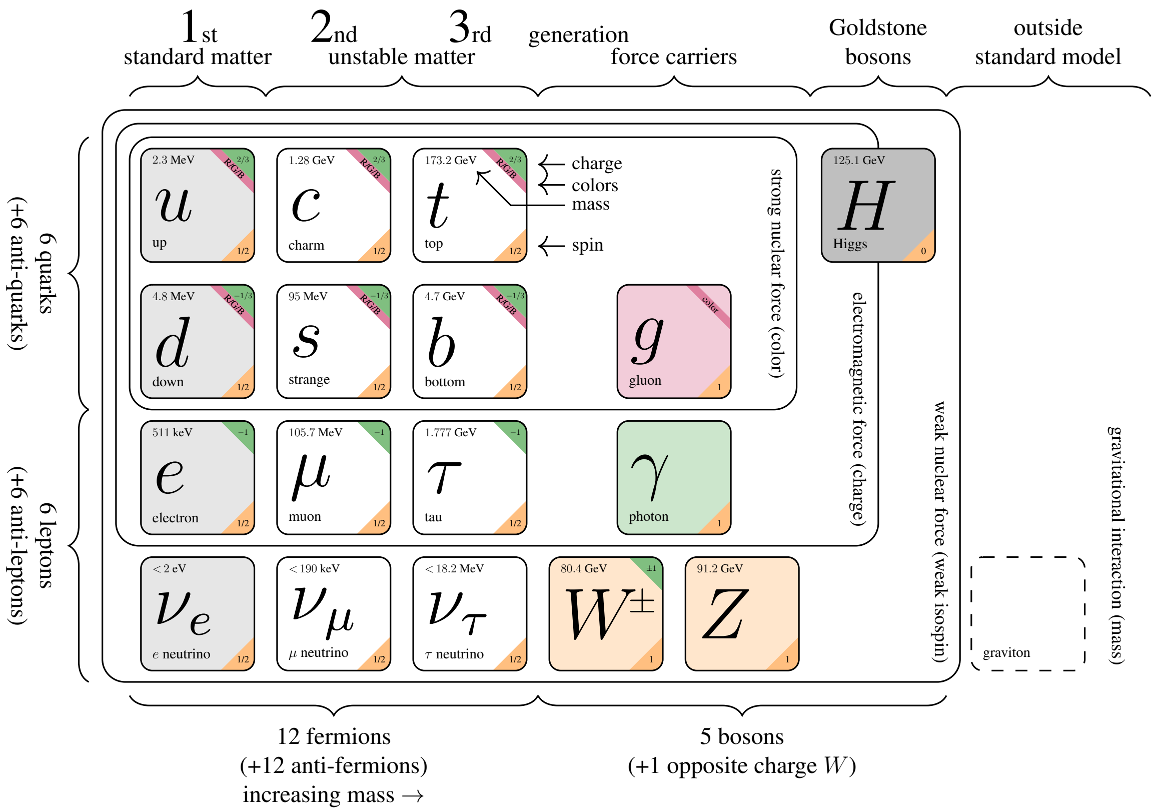

1st standard matter
2nd unstable matter
3rd generation
force carriers
Goldstone bosons
outside standard model
2.3 MeV 2/3 R/G/B u up 1/2
1.28 GeV 2/3 R/G/B c charm 1/2
173.2 GeV 2/3 R/G/B t top 1/2
charge
colors
mass
spin
125.1 GeV H Higgs 0
4.8 MeV −1/3 R/G/B d down 1/2
95 MeV −1/3 R/G/B s strange 1/2
4.7 GeV −1/3 R/G/B b bottom 1/2
color g gluon 1
511 keV −1 e electron 1/2
105.7 MeV −1 μ muon 1/2
1.777 GeV −1 τ tau 1/2
γ photon 1
< 2 eV νe e neutrino 1/2
< 190 keV νμ μ neutrino 1/2
< 18.2 MeV ντ τ neutrino 1/2
80.4 GeV ±1 W± 1
91.2 GeV Z 1
graviton
strong nuclear force (color)
electromagnetic force (charge)
weak nuclear force (weak isospin)
gravitational interaction (mass)
6 quarks (+6 anti-quarks)
6 leptons (+6 anti-leptons)
12 fermions (+12 anti-fermions) increasing mass →
5 bosons (+1 opposite charge W)

Figure 11.146: **Subparticles Standard Model (source: Wikipedia)**

In this section, we do not wish to make a complete course of Quantum Field Theory because for reminder, this entire textbook is only intended to give the basics of what an engineer must know in the early 121st century (holocene calendar) and incidentally the author to have fun studying subjects that he had not seen during his school curriculum. As such, the interested reader to deepen over this matter may be read to best textbook we've ever had hands so far (detailed, simple and pedagogical developments with many practical cases) on the subject which is [141].

Remark

Quantum Field Theory is the application of Quantum Mechanics to fields. It provides a framework widely used in Particle Physics and Condensed Matter Physics. The basics of Quantum Field Theory to which we will limit our study were developed between 11935 (holocene calendar) and 11955 (holocene calendar), mainly by Paul Dirac, Wolfgang Pauli, Sin-Itiro Tomonaga, Julian Schwinger, Richard Feynman and Freeman Dyson.

Before we launch into the calculations (see below), let us show that the approach suggested above, can be considerate with a simple formalism (pedagogical).

Readers should however remember that simple approaches require sometimes erroneous mental constructions (overly simplistic) compared to reality, but which meet the objective: to have a more or less understandable and intuitive model.

Consider at this purpose the figure below (representation of the elastic collision of two electrons):

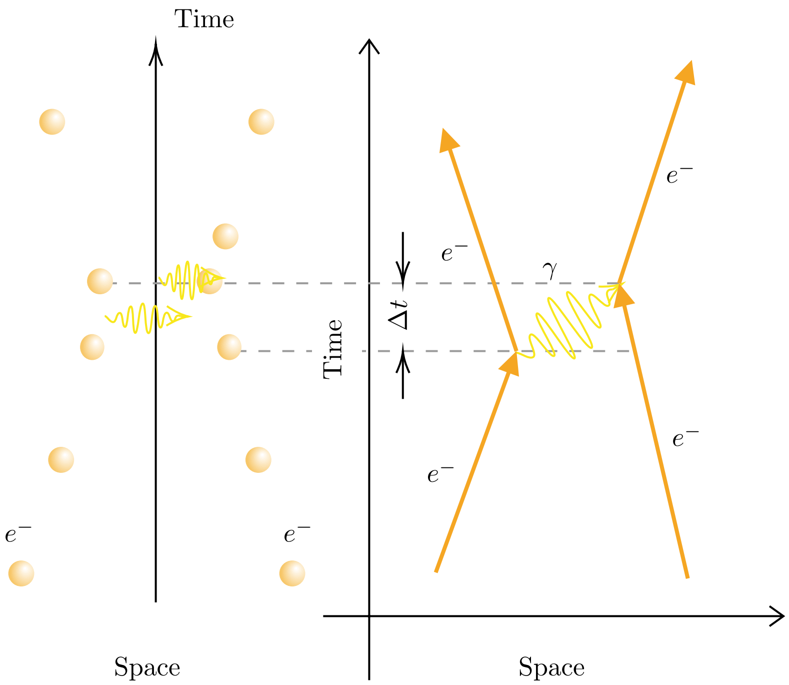


Figure 11.147: Example of pseudo Feynman-diagram

This figure is named, and wrongly (!), in many educational textbooks "Feynman diagram" (in reality it is only a diagram that looks like a little bit because the real Feynman diagrams are intended to calculate Wick products in a perturbation series).

Let us suppose that two electrons which are represented, initially move at the same speed. They approach first and then move away from each other along a straight line in space that is projected on the time axis, in the direction of increasing time. The left electron emits a photon (the wavy line), and for a time Δt, there are two electron and a photon. The electron on the right then absorbs the photon and the interaction is temporarily ended; other photons will eventually go back and forth between the electrons. The average force is proportional to the transfer rate of the amount of movement due to the exchange of photons. The probability of the emission or absorption of photons by a particle is connected to its electric charge. The force must be proportional to the product of the charges in interaction (according to Coulomb's law). Think of the repulsive force between two astronauts floating in space and sharing a ball in one direction and then the other (this is an educational approach to the problem but does not apply for example to the attraction between two particles of opposite charges!). However, the opposite phenomenon of attraction cannot be viewed in this way but only in formal mathematical form.

💣Caution!!!! At the quantum level, it is more accurate rigorously to speak about "interactions" rather than "forces". This is because the classical notion of a force as a continuous and deterministic influence that acts on an object over a distance breaks down at the quantum level. The concept of an interaction in Quantum Field Theory is more general than the classical notion of a force as an interaction can encompass both attractive and repulsive effects, and it can be mediated by a variety of particles, including photons, gluons, and W and Z bosons. In summary, while the classical notion of a force can still be useful in certain contexts (especially for pedagogical reason for undergraduate students), it is more accurate to speak about "interactions" at the quantum level, which are described by quantum fields and the exchange of virtual particles.

The collision shown in the figure above is elastic; the energy of each electron is unchanged in the collision. Despite this, for a time Δt, the system contains an additional energy $h\nu$ corresponding to the photon. Meanwhile this time Δt, conservation of energy is apparently non respected! Can we tolerate this situation? The answer given by modern physics, is: Yes! But it can never be observed (...). In other words, because there is always some uncertainty ΔE on the measured value of the energy of a system. The Heisenberg's uncertainty principle involving (see the simplified proof in the section of Wave Quantum Physics page 4704) that:

$$\Delta t \Delta E \cong \hbar \qquad (11.5.1)$$

A violation of the energy conservation law until a given amount ΔE will be hidden by uncertainty on energy provided that the time available to make the observation is sufficiently small as having for upper value:

$$\Delta t \cong \frac{\hbar}{\Delta E} \qquad (11.5.2)$$

obviously a value less than Δt also satisfies the condition. We can therefore write:

$$\Delta t \geq \frac{\hbar}{\Delta E} \qquad (11.5.3)$$

The uncertainty on the energy exceeds the energy of a photon $h\nu$ if the photon exists for a shorter time than:

$$\Delta t = \frac{\hbar}{h\nu} = \frac{1}{4\pi\nu} \qquad (11.5.4)$$

The photon can then be observed on a maximum distance of:

$$\boxed{R_\nu = c\Delta t = \frac{c}{4\pi\nu}} \qquad (11.5.5)$$

and as the frequency ν can be arbitrarily small, the scope of the transmitted force by the photon without mass is unlimited. It may seem in this relation that the scope is limited to a free photon. But this would be forgetting (see section Wave Quantum Physics page 4785) that a free photon does not exist, because it would have a totally unknown frequency. Therefore the interaction distance would be also undetermined.

These exchange of quanta, that are unobservable, are named "virtual photons". Since photons are not electrically charged, we also say that the interaction is made by "neutral current".

Remark

The term "virtual particle" is an endlessly confusing and confused subject for the layperson, and even for the non-expert scientist!!! A particle is a nice, regular ripple in a field, one that can travel smoothly and effortlessly through space, like a clear tone of a bell moving through the air. A "virtual particle", generally, is a disturbance in a field that will never be found on its own, but instead is something that is caused by the presence of other particles, often of other fields. A particle is a nice, regular ripple in a field, one that can travel smoothly and effortlessly through space, like a clear tone of a bell moving through the air. Physicists often say, and laypersons' books repeat, that the two electrons exchange virtual photons. But those are just words, and they lead to many confusions if you start imagining this word "exchange" as meaning that the electrons are tossing photons back and forth as two children might toss a ball.

A much more satisfactory approach is that of using mass as energy term:

$$\Delta t \leq \frac{\hbar}{m_0 c^2} \qquad (11.5.6)$$

Thanks to this relation, it is possible to know the time during which a massive virtual particle can travel a distance that would be:

$$\boxed{R_m = c\Delta t = \frac{\hbar}{m_0 c}} \qquad (11.5.7)$$

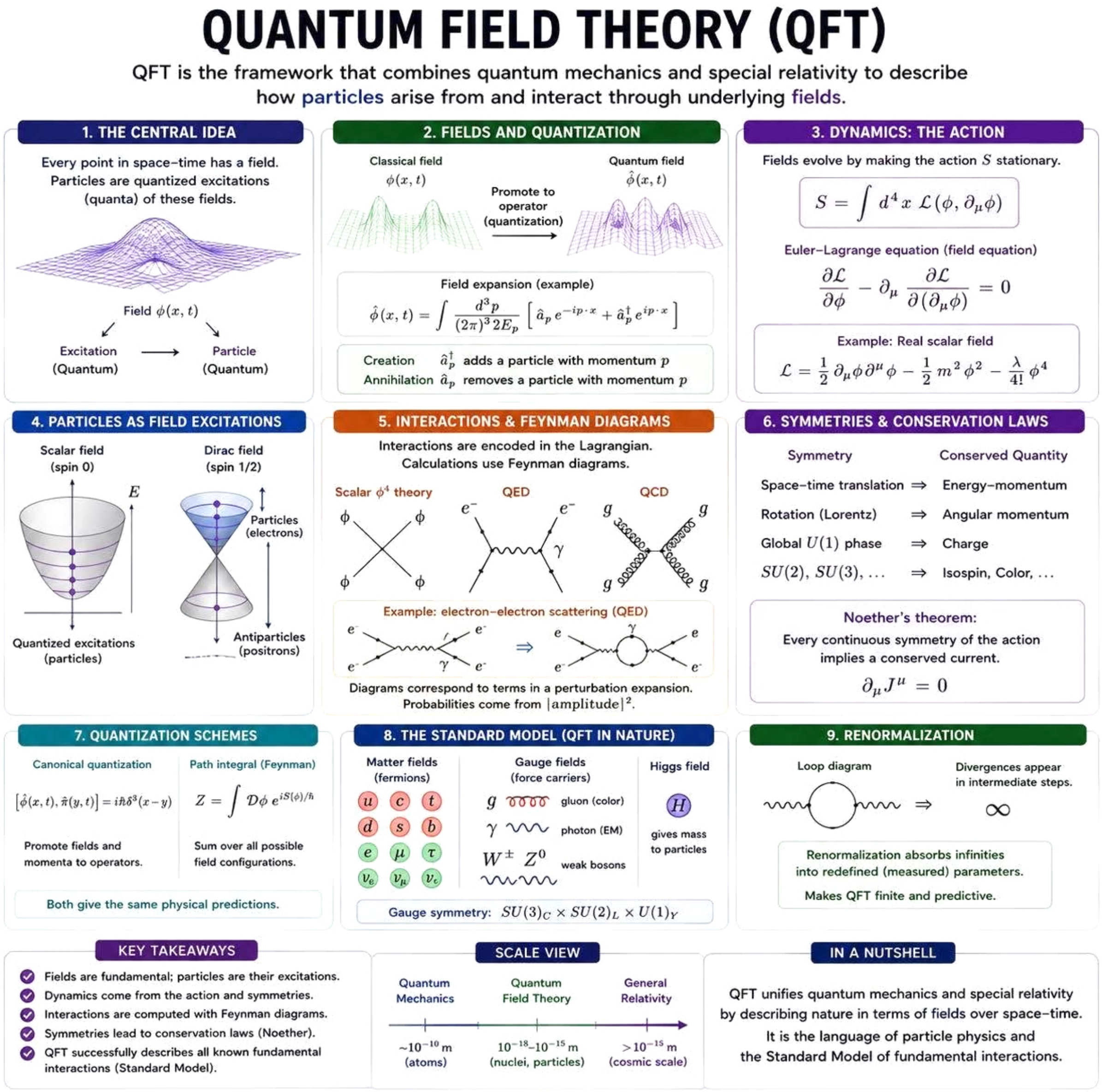
We will see later how to approximately determine the mass of virtual particles that mediate in the nuclear forces which will allow us to estimate the duration of interactions as being of the order of 10^{-24} [s].

In the late 11920s (holocene calendar), it became clear that we could consider each known particle (proton, electron, etc.) as the quantum of a specific field. In this vision, there is an electron field, a proton field, and so on as will be proven below (the Universe would be a set of unified field). Any object is in reality a set of manifestations of observables of field quantas.

Furthermore, we saw that the writing of wave equations for relativistic particles (Dirac equation and Klein-Gordon equation proven in the section of Relativistic Quantum Physics) brings intractable problems in a classical point of view, including negative energies. In fact, this approach is not justified, because according to Einstein equation mass and energy are equivalent, and if we add to that the energy-time Heisenberg's uncertainty principle we find that an infinite number of particles can be created or annihilated, hence the need for a model to take into account not only the properties of a single particle, but of a set of particles, both real and virtual!

Remark

When Fermi formulated his theory of weak interactions in 11932 (holocene calendar), he used the same principles those of quantum electrodynamics (this is one reason why QED is named the "jewel of physics" - the Standard Model is also based on this theory). Two years later, the Japanese physicist Hideki Yukawa proposed that the weak interaction was due to the exchange of a massive virtual boson.



QUANTUM FIELD THEORY (QFT)

QFT is the framework that combines quantum mechanics and special relativity to describe how particles arise from and interact through underlying fields.

1. THE CENTRAL IDEA

Every point in space-time has a field. Particles are quantized excitations (quanta) of these fields.

Field $\phi(x,t)$
Excitation (Quantum)
Particle (Quantum)

2. FIELDS AND QUANTIZATION

Classical field $\phi(x,t)$
Promote to operator (quantization)
Quantum field $\hat{\phi}(x,t)$

Field expansion (example)

$$\hat{\phi}(x,t)=\int\frac{d^3p}{(2\pi)^3 2E_p}\left[\hat{a}_p e^{-ip\cdot x}+\hat{a}_p^\dagger e^{ip\cdot x}\right]$$

Creation $\hat{a}_p^\dagger$ adds a particle with momentum p

Annihilation $\hat{a}_p$ removes a particle with momentum p

3. DYNAMICS: THE ACTION

Fields evolve by making the action S stationary.

$$S=\int d^4x\,\mathcal{L}(\phi,\partial_\mu\phi)$$

Euler-Lagrange equation (field equation)

$$\frac{\partial\mathcal{L}}{\partial\phi}-\partial_\mu\frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)}=0$$

Example: Real scalar field

$$\mathcal{L}=\frac{1}{2}\partial_\mu\phi\partial^\mu\phi-\frac{1}{2}m^2\phi^2-\frac{\lambda}{4!}\phi^4$$

4. PARTICLES AS FIELD EXCITATIONS

Scalar field (spin 0)
Dirac field (spin 1/2)
E
Particles (electrons)
Quantized excitations (particles)
Antiparticles (positrons)

5. INTERACTIONS & FEYNMAN DIAGRAMS

Interactions are encoded in the Lagrangian. Calculations use Feynman diagrams.

Scalar ϕ^4 theory
ϕ
ϕ
ϕ
ϕ
QED
e^-
e^-
γ
QCD
g
g
g
g
Example: electron-electron scattering (QED)
e^-
e^-
γ
e^-
e^-
$\Rightarrow$
e
γ
e
e^-
e^-

Diagrams correspond to terms in a perturbation expansion. Probabilities come from $|\text{amplitude}|^2$.

6. SYMMETRIES & CONSERVATION LAWS

Symmetry		Conserved Quantity
Space-time translation	⇒	Energy-momentum
Rotation (Lorentz)	⇒	Angular momentum
Global $U(1)$ phase	⇒	Charge
$SU(2)$, $SU(3)$, …	⇒	Isospin, Color, …

Noether's theorem:

Every continuous symmetry of the action implies a conserved current.

$$\partial_\mu J^\mu=0$$

7. QUANTIZATION SCHEMES

Canonical quantization	Path integral (Feynman)
$[\hat{\phi}(x,t),\hat{\pi}(y,t)]=i\hbar\delta^3(x-y)$	$Z=\int\mathcal{D}\phi\,e^{iS(\phi)/\hbar}$
Promote fields and momenta to operators.	Sum over all possible field configurations.

Both give the same physical predictions.

8. THE STANDARD MODEL (QFT IN NATURE)

Matter fields (fermions)

u c t

d s b

e μ τ

ν_e ν_μ ν_τ

Gauge fields (force carriers)

g gluon (color)

γ photon (EM)

$W^\pm$ Z^0 weak bosons

Higgs field

H gives mass to particles

Gauge symmetry: $SU(3)_C\times SU(2)_L\times U(1)_Y$

9. RENORMALIZATION

Loop diagram
$\Rightarrow$
∞

Divergences appear in intermediate steps.

Renormalization absorbs infinities into redefined (measured) parameters.

Makes QFT finite and predictive.

KEY TAKEAWAYS

- ✓ Fields are fundamental; particles are their excitations.
- ✓ Dynamics come from the action and symmetries.
- ✓ Interactions are computed with Feynman diagrams.
- ✓ Symmetries lead to conservation laws (Noether).
- ✓ QFT successfully describes all known fundamental interactions (Standard Model).

SCALE VIEW

Quantum Mechanics	Quantum Field Theory	General Relativity
$\sim 10^{-10}$ m (atoms)	10^{-18}–10^{-15} m (nuclei, particles)	$>10^{-15}$ m (cosmic scale)

IN A NUTSHELL

QFT unifies quantum mechanics and special relativity by describing nature in terms of fields over space-time.

It is the language of particle physics and the Standard Model of fundamental interactions.

11.5.1 Yukawa potential

The best way to argue the example of quantum remains the "proof" of the Coulomb's law (and also of Newton's Law) from the results we have achieved in Wave Quantum Physics (we own these developments to the physicist Hideki Yukawa).

A simplified version of this proof as we like them in this textbook is to first remember the free Klein-Gordon equation (see section Wave Quantum Physics page 4921):

$$\hbar^2\frac{\partial^2\Psi}{\partial t^2}=\hbar^2c^2\Delta\Psi-m_0^2c^4\Psi \tag{11.5.8}$$

this equation describes the amplitude dynamics of presence of a particle without spin in time in a given potential.

Let us consider a static component of Ψ (time-independent) with spherical symmetry:

$$\Psi = U(r) \tag{11.5.9}$$

The Klein-Gordon equation is then reduced to:

$$\hbar^2 c^2 \Delta U(r) - m_0 c^4 U(r) = 0 \tag{11.5.10}$$

If we divide on the both side of the equality by $\hbar^2 c^2$ we get obviously:

$$\Delta U(r) - \hbar^{-2} m_0 c^2 U(r) = 0 \tag{11.5.11}$$

Let us recall (see section Vector Calculus page 1899) the notation of the Laplacian of a scalar field:

$$\vec{\nabla}^2 = \Delta \tag{11.5.12}$$

and also its expression in spherical coordinates where $r = 0$ is identified to the source of the field (see section Vector Calculus page 1903):

$$\vec{\nabla}^2 f = \frac{\partial^2 f}{\partial r^2} + 2\frac{1}{r}\frac{\partial f}{\partial r} + \frac{1}{r^2}\frac{\partial^2 f}{\partial \theta^2} + \frac{\cot(\phi)}{r^2}\frac{\partial f}{\partial \theta} + \frac{1}{r^2 \sin^2(\theta)}\frac{\partial^2 f}{\partial \phi^2} \tag{11.5.13}$$

As the field $U(r)$ has a spherical symmetry (depending on r only) the Laplacian is reduced to:

$$\vec{\nabla}^2 f = \frac{\partial^2}{\partial r^2} + 2\frac{1}{r}\frac{\partial}{\partial r} = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial}{\partial r}\right) \tag{11.5.14}$$

So the equation of the field U (r) is written:

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial}{\partial r} U(r)\right) - \hbar^{-2} m_0^2 c^2 U(r) = 0 \tag{11.5.15}$$

This differential equation has for solution (we guess easily enough that the exponential is a possible solution):

$$U(r) = \frac{C}{r} e^{-r(\hbar^{-1} m_0 c)} \tag{11.5.16}$$

where C is an integration constant.

In the context of the use of natural units (which is most common at this level in the scientific literature) this potential is written:

$$U(r) = \frac{C}{r} e^{-r m_0} \tag{11.5.17}$$

and is named "Yukawa potential".

The reader will notice that apart from the distance r, the other variable in the exponential is the mass (the other terms are universal constants).

Consequence: the Yukawa potential has the aspect of a "scalar field" in the case where the mass is zero (see example below) and at the same time has the aspect of a "mass field" in the case where the mass is not zero!

This leads us to the following hypothesis: if it is the electric field that holds charged particles in the atom between them (see the treatment of non-mass fields below), it is the mass field which keeps the uncharged particles together in the atom.

In other words, if particles interact via a mass field of mass m_0 (instead of interacting with massless photons), their mutual strength will decrease exponentially (which is very fast).

The above relation is often denoted:

$$U(r)=g\frac{e^{-rm_0}}{r} \tag{11.5.18}$$

Where g is a constant that represent the strength of pion field similarly to electric charge q representing the strength of electromagnetic field. The above expression represents the exponential form of scalar meson potential or one pion exchange potential.

Yukawa's approach allows a new point of view to the interpretation of nuclear phenomena, but it is however too naive to adequately explain the strong interactions in general.

11.5.1.1 Mass fields

The physicist Hideki Yukawa therefore proposed in 11935 (holocene calendar) that nuclear power was own its very short range by the fact that it is transmitted by massive particles (more the mass of exchanged quanta is big, the greater the range of the interaction is reduced) described by the mass field above.

Remark

In the historical context of the time, these hypothetical particles were "mesons". But we will see that this assumption did not hold up very long.

Let's take a closer look to this. Let us write the Yukawa potential as follows:

$$U(r)=\frac{C}{r}e^{-r/R} \tag{11.5.19}$$

with:

$$\boxed{R=\frac{\hbar}{m_0 c}} \tag{11.5.20}$$

This notation is not innocent, because as we will see it in detail later, when $R\rightarrow+\infty$ (the case of electromagnetic and gravitational interaction) then $m_0\rightarrow 0$ and then we fall back on the fundamental law of electrodynamics or of gravitation where the particle interaction is respectively the photon (null mass) and the hypothesized graviton.

Thus, assuming that the radius of the strong nuclear force (cohesion of nucleons between them) is around $R\cong 1\cdot 10^{-15}$ [m] and that of the weak nuclear interaction (which should be the cause of the beta disintegration as we already mentioned in the section of Nuclear Physics) around $R\cong 1\cdot 10^{-18}$ [m], we get the binding interaction energies and their approximate weight immediately:

- For the "strong interaction":

$$\begin{aligned}E=m_0c^2=\frac{\hbar c}{R}\cong 3.15\cdot 10^{-12}\,[\text{J}]\cong 196\,[\text{MeV}]\\ m_0=\frac{\hbar}{cR}\cong 3.5\cdot 10^{-28}\,[\text{kg}]\cong 386 m_e\cong\frac{1}{5}m_p\end{aligned} \tag{11.5.21}$$

  that is to say approximately 386 times the mass of the electron (m_e) and $1/5$ the mass of the proton (m_p).

  Two years after this prediction of Hideki Yukawa, physicists discovered a particle corresponding to this mass: the meson μ^-. It was later that it was not good the good particle but a particle of the same type as the electron particle, that is to say a lepton and therefore a fermion[41] (so it can not be a messenger particle!!). Furthermore, the experiences of diffusion and collisions with protons, deuterons, etc. at energies higher and higher showed that there was a change in the intensity and shape of the strong interaction incompatible with the hypothesis of a single meson. Furthermore, the hadronic resonances showed that there was meson excited states which is hard to imagine for particles considered fundamental in analogy with the photon!!

---

[41] Go see the table of fundamental particles in the section Elementary Particle Physics at page 5200 to see how physicists classify elementary particles of the Standard Model

The particles detected in the laboratory and that seemed to be the best candidates at that time (since there were several ...) for the strong nuclear force were "pions" (or "pi mesons") which are know under three forms (at least at the time we write these lines):

$$\pi^+,\pi^-,\pi^0$$

and which are almost 270 times more massive than the electron. So this mass difference clearly indicates that the model of Yukawa is not entirely accurate.

Before the discovery of quarks (which mesons are made of), the pi mesons (pions) were therefore considered to be the vectors of the strong interaction (today we know that in fact these vectors are gluons).

- For the "weak interaction":

$$\begin{aligned}E&=m_0c^2=\frac{\hbar c}{R}=3.15\cdot 10^{-8}\,[\text{J}]\cong 196\,[\text{GeV}]\\ m_0&=\frac{\hbar}{cR}\cong 3.5\cdot 10^{-25}\,[\text{kg}]\end{aligned}\tag{11.5.22}$$

  It is therefore a colossal mass, a hundred times the mass of the proton! The vectors of interaction have been candidates that were identified in 11983 (holocene calendar) in the CERN accelerator. These carrier particles of the weak nuclear force are named "intermediate bosons" and denoted by W^+, W^-, Z^0.

These observations led the hypothesis that Yukawa's theory was not a fundamental theory enough even if it explain quite well some properties...

11.5.1.2 Non-mass fields

Let us now imagine a static scalar field with a spherical symmetric which the photon (particle without spin) is the hypothetical quantum exchange. As the photon's mass is zero, the expression $U(r)$ is reduced to:

$$U(r)=\frac{C}{r}\tag{11.5.23}$$

If we interpret $U(r)$ as the electrostatic potential source of a quantity $Q=nq$ of elementary electric charges q, then the constant C in our metric is:

$$C=\frac{Q}{4\pi\varepsilon}\tag{11.5.24}$$

Such that:

$$U(r)=\frac{1}{4\pi\varepsilon}\frac{Q}{r}\tag{11.5.25}$$

As we have:

$$U(r) = -\int_{\Lambda} E\mathrm{d}r \tag{11.5.26}$$

We get:

$$\frac{\mathrm{d}U(r)}{\mathrm{d}r} = -E = -\frac{1}{4\pi\varepsilon}\frac{Q}{r^2} \tag{11.5.27}$$

Which give us:

$$E = \frac{1}{4\pi\varepsilon}\frac{Q}{r^2} \tag{11.5.28}$$

We can conclude that if a particle is in a potential field with spherical symmetry $U(r)$, whose photon is assumed to be initially the interacting quantum, then we are dealing with an electrostatic field whose expression is identical to the Coulomb law (this valid once again in a big way the Wave Quantum Physics theory).

Remark

Hubert Reeves and his colleagues astrophysicists proved that at the time of the genesis of the Universe, the smallest deviance in coupling constants would have caused the instability of nucleons and would have condemned cosmic evolution.

The photon is therefore indeed the quantum interaction of the electric field with spherical symmetry (when the electric charges have a relativistic speed the electric field is not spherical anymore, and the equations become a bit more complicated - see the section of Special Relativity for some details about this) and we should not talk of electric charges but of "transparency" to photons. Indeed, the neutron being globally neutral it should not interact with the electric field, but as it is composed of charged particles (quarks) experiments show an influence in the presence of an electromagnetic field (which photon is the quantum interaction). That said, applying the same reasoning, we can still find the Newton's gravitational potential:

$$a = G\frac{M}{r^2} \tag{11.5.29}$$

This would imply that the quantum interaction of the gravitational field is massless (at least in the case of small masses as we know the Newton's potential is only an approximation of General Relativity in the case of small masses). Since the gravitational field does not seem to interact with the presence of a magnetic or electrostatic field, this leads us to speculate that the quantum of interaction is not the photon but another hypothesized particle that we name the "graviton".

11.5.2 Euler-Lagrange equation for Fields

How field theory was introduced from the elementary particles by Dirac is known for historical reasons under the name of "second quantization".

It may be useful to identify a possible source of confusion: the fields are not related to the wave-particle duality. What we mean by "field" is a concept that allows the creation or annihilation of particles at any point in space as we will discussed it in the mathematical developments below.

Recall that we have defined in the section of Wave Quantum Physics during the study of the Schrödinger's evolution equation the Heisenberg's operator necessary for the de Broglie normalization condition:

$$X(t)=e^{\frac{\mathrm{i}t}{\hbar}H}Xe^{-\frac{\mathrm{i}t}{\hbar}H}\tag{11.5.30}$$

Differentiating this operator with respect to time, we trivially have trivially (once known the notation seen in the section of Wave Quantum Physics):

$$\dot{X}(t)=\frac{\mathrm{i}H}{\hbar}e^{\frac{\mathrm{i}t}{\hbar}H}Xe^{-\frac{\mathrm{i}t}{\hbar}H}-\frac{\mathrm{i}H}{\hbar}e^{\frac{\mathrm{i}t}{\hbar}H}Xe^{-\frac{\mathrm{i}t}{\hbar}H}=-\frac{\mathrm{i}}{\hbar}[X(t),H]\tag{11.5.31}$$

where for recall, the commutator of two operators is given (as we have already seen it in our study of adjoint and Hermitian operators in the section of Wave Quantum Physics) by definition:

$$[\alpha,\beta]=\alpha\beta-\beta\alpha\tag{11.5.32}$$

This is the Hamiltonian H which appears first in the previous prior previous relation.

Now, as we know, we can substitute to $\dot{X}(t)$ observable that are also well known to us such that we get in the case of a constant Hamiltonian:

$$\boxed{\begin{aligned}\dot{q}(t)&=\frac{\mathrm{i}}{\hbar}[q(t),H]\\ \dot{p}(t)&=\frac{\mathrm{i}}{\hbar}[p(t),H]\end{aligned}}\tag{11.5.33}$$

But we can just as easily substitute a time-dependent Hamiltonian $H(t)$ such that:

$$\dot{X}(t)=-\frac{\mathrm{i}}{\hbar}[X(t),H]+e^{\frac{\mathrm{i}t}{\hbar}H}\frac{\partial X}{\partial t}e^{-\frac{\mathrm{i}t}{\hbar}H}\tag{11.5.34}$$

For example, as we know (see section Wave Quantum Physics page 4695), X can be replaced by the operators $q(t)\rightarrow\hat{q}(t)$ or $p(t)\rightarrow-\mathrm{i}\hbar\partial_k$. Then in this special cases we see that we can simplify the previous relation by:

$$\boxed{\dot{X}(t)=-\frac{\mathrm{i}}{\hbar}[X(t),H]+\partial_t X}\tag{11.5.35}$$

This is named "Heisenberg equation of motion".

What is interesting in the two relations $\dot{q},\dot{p}$ obtained previously, is the way in which is realized the connection between quantum physics and classical mechanics. Indeed, we proved in the section of Analytical Mechanics the following relations are will always be valid whatever the subject area:

$$\begin{aligned}\dot{q}(t)&=\frac{\partial H(t)}{\partial p(t)}\\ \dot{p}(t)&=-\frac{\partial H}{\partial q(t)}\end{aligned} \tag{11.5.36}$$

and also that:

$$p=\frac{\partial L}{\partial \dot{q}} \tag{11.5.37}$$

and assuming the generalization to several degrees of freedom as intuitive:

$$H=\sum p_n\dot{q}_n-L \tag{11.5.38}$$

The generalization to several degrees of freedom is immediate and gives us all the following relations (we simplify the notation by omitting the explicit time dependence):

$$p_n=\frac{\partial L}{\partial \dot{q}_n} \quad \text{and} \quad H=\sum p_n\dot{q}_n-L \tag{11.5.39}$$

and also:

$$\begin{cases}\dot{q}_n=\dfrac{\partial H}{\partial p_n}\\ \dot{p}_n=-\dfrac{\partial H}{\partial q_n}\end{cases} \quad \text{and} \quad \begin{cases}\dot{q}_n=-\dfrac{\mathrm{i}}{\hbar}[q_n,H]\\ \dot{p}_n=-\dfrac{\mathrm{i}}{\hbar}[p_n,H]\end{cases} \tag{11.5.40}$$

We still need two other important relations that we will immediately determined. First, according to the definitions of the commutators, we already prove in the section of Analytical Mechanics:

$$[q,q]=[p,p]=0 \tag{11.5.41}$$

By cons, it is a little more subtle to prove the value of $[q,p]$ (we joke...). Recall that we had prove in our study functional linear operators (we restrict ourselves to the case of the x here):

$$p_x=\frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial x} \tag{11.5.42}$$

and that q represents a generalized coordinate (x for example ...). So we have (results already proven in the section of Wave Quantum Physics):

$$[q,p]=\left[x,\frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial x}\right]=x\frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial x}-\frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial x}x=0-\frac{\hbar}{\mathrm{i}}=\mathrm{i}\hbar \tag{11.5.43}$$

The last two relations can be generalized to all necessary components such as:

$$\boxed{[q_i,q_j]=[p_i,p_j]=0 \quad [q_i,p_j]=\mathrm{i}\hbar\delta_{ij}} \tag{11.5.44}$$

with for recall (see section Tensor Calculus page 2044):

$$\delta_{ij}=\delta^{ij}=\delta_i^j=\begin{cases}1 & \text{if } i=j\\ 0 & \text{if } i\neq j\end{cases} \tag{11.5.45}$$

which is the Kronecker symbol.

To arrive finally to Quantum Field Theory, we still need to generalize to a continues infinite number of degrees of freedom. Even the simplest of fields is characterized, at a time t, by an infinity of continuous quantities:

$$\Phi(x,y,z,t)=\Phi(\vec{x},t)=\Phi^A \tag{11.5.46}$$

for any $\vec{x}$. We could then imagine to represent the function Φ by its values $\Phi(\vec{x},t)$ in a discrete set of points $\vec{x}_i$ that we will ultimately make infinitely dense (beware of the fact that we use the concept of density!). We can also work, to begin, not in all space, but in a finite volume that we will eventually make very large. By doing so, we can find how to generalize the canonical formalism and the quantification process. At the formal level, notwithstanding subtle issues of convergence (see the mathematical parts of this textbook), the generalization to continuous systems is mainly to replace the quantities with index i by integrals of the arguments $\vec{x}$, and the deltas of Kronecker by the delta of Dirac (on space-time):

$$\delta^4\left((x_1,y_1,z_1,t)-(x_2,y_2,z_2,t)\right)=\delta^4\left((A)-(B)\right)=\delta^4(\vec{x}-\vec{y})=\delta^4_{AB} \tag{11.5.47}$$

Considering then the variational principle as we have studied it in the section of Analytical Mechanics:

$$S(q,\dot{q})=\int\limits_{t_A}^{t_B} L(q,\dot{q})\mathrm{d}t \tag{11.5.48}$$

and the principle of least action requiring us to have:

$$\delta S=\delta\int\limits_{t_A}^{t_B} L\mathrm{d}t=\int\limits_{t_A}^{t_B}\delta L\mathrm{d}t=0 \tag{11.5.49}$$

where the Lagrangian will be now a function of the field $\Phi(\vec{x},t)$ and of derivative with respect to $\phi(\vec{x},t)$ (since there is no concept of momentum for a field!).

If we divide the above equation by $\delta q(t)=\delta q\neq 0$ we get:

$$\frac{\delta S(q,\dot{q})}{\delta q}=0 \tag{11.5.50}$$

what gives us the right to write:

$$\delta S(q,\dot{q})=\int\frac{\delta S(q,\dot{q})}{\delta q}\delta q\mathrm{d}t=\int\delta q\mathrm{d}t=0 \tag{11.5.51}$$

and imposing an analogy with the concept of field:

$$\delta S(q,\dot{q})=\int\frac{\delta S(q,\dot{q})}{\delta q}\delta q\mathrm{d}t=\int\frac{\delta S(\Phi^A,\dot{\Phi}^A)}{\delta\Phi^A}\delta\Phi^A\mathrm{d}x^4 \tag{11.5.52}$$

where $x := (\vec{x}, t)$ and $\mathrm{d}x^4 = \mathrm{d}x_1 \mathrm{d}x_2 \mathrm{d}x_3 \mathrm{d}x_t$.

Finally, since all the following terms are equal to zero, they are equal (we introduce the Euler-Lagrange equation proved in the section of Analytical Mechanics):

$$\delta S(q,\dot{q}) = \frac{\delta S(q,\dot{q})}{\delta q(t)} = \frac{\partial L}{\partial q_i(t)} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}_i(t)} = 0 \tag{11.5.53}$$

in analogy with the field Φ^A we get (thus it is an intuitive approach with the physicist way of life...):

$$\frac{\delta S(\Phi^A, \dot{\Phi}^A)}{\delta \Phi^A} = \frac{\partial L}{\partial \Phi^A} - \frac{\mathrm{d}}{\mathrm{d}(A)}\frac{\partial L}{\partial \dot{\Phi}^A} = 0 \tag{11.5.54}$$

However, since this writing is not very convenient, it is customary to write the partial differentials (using the natural units of physics) of the components $\partial/\partial t$, $\partial/\partial y$, $\partial/\partial z$, $\partial/\partial x$ in the form ∂_μ, which finally gives us the "Lagrangian field equations":

$$\boxed{\frac{\partial L}{\partial \Phi^A} - \partial_\mu \left(\frac{\partial L}{\partial(\partial_\mu \Phi^A)}\right) = 0} \tag{11.5.55}$$

Obviously, if we do not use natural units, we should adopt the notation:

$$\partial_\mu = \left(\frac{\partial}{c\partial t}, \frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right) \tag{11.5.56}$$

The fact that the partial derivation ∂_μ now acts on all the components and not only on t is due to the change of generalized coordinate q function only from t to a function of the field Φ^A dependent on x, y, z and t. The reason for this lies in the fact that time and spatial coordinates play the same role, that of describing the space-time continuum on which the physical system evolves.

This also leads us to write the principle of least action in the following form:

$$\begin{gathered}\delta S(q,\dot{q}) = \int \delta L \mathrm{d}t = 0 \\ \Rightarrow \delta S(\Phi^A, \dot{\Phi}^A) = \int \frac{\delta L}{\delta \Phi^A}\delta\Phi^A \mathrm{d}x^4 = \int \frac{\partial L}{\partial \Phi^A}\delta\Phi \mathrm{d}x^4 = 0\end{gathered} \tag{11.5.57}$$

With the action of fields denoted more traditionally:

$$S = \int L(\Phi^A, \dot{\Phi}^A, \partial_\mu \Phi^A)\mathrm{d}x^3\mathrm{d}t = \int L(\Phi^A, \dot{\Phi}^A, \partial_\mu \Phi^A)\mathrm{d}x^4 \tag{11.5.58}$$

or to differentiate Lagrangian and Lagrangian density (we "stylized" sometimes the L to makes the difference between the both):

$$\boxed{S = \int \mathcal{L}(\Phi^A, \dot{\Phi}^A, \partial_\mu \Phi^A)\mathrm{d}x^4} \tag{11.5.59}$$

to be compared to the action of the particle:

$$S(q,\dot{q}) = \int_{t_\lambda}^{t_s} L(q,\dot{q})\mathrm{d}t \tag{11.5.60}$$

In analogy with $p_n=\dfrac{\partial L}{\partial \dot{q}_n}$, we will write:

$$\prod_A=\frac{\partial L}{\partial(\partial_\mu \Phi^A)}=\frac{\partial L}{\partial \dot{\Phi}^A} \tag{11.5.61}$$

and in analogy with $H=\sum p_n\dot{q}_n-L$ we write:

$$H=\sum\prod_A \dot{\Phi}^A-L \tag{11.5.62}$$

But a field is a continuous medium! The sigma sum is therefore no longer adapted and we must move to an integration over all space-time such that:

$$H=\int\prod_A \dot{\Phi}^A\mathrm{d}^4x-L \tag{11.5.63}$$

In analogy with the Heisenberg equations of motion (this way of doing things is often named the "principle of correspondence"), we write:

$$\begin{cases}\dot{q}_n=\dfrac{\partial H}{\partial p_n}\\ \dot{p}_n=-\dfrac{\partial H}{\partial q_n}\end{cases}\Leftrightarrow\begin{cases}\Phi^A=\dfrac{\partial H}{\partial \prod_A}\\ \dot{\prod}_A=-\dfrac{\partial H}{\partial \Phi^A}\end{cases} \tag{11.5.64}$$

Let us now turn to quantum theory by postulating corresponding of "Heisenberg field operators". Let us recall that we had obtained earlier above that:

$$\dot{q}_n=-\frac{\mathrm{i}}{\hbar}[q_n,H],\dot{p}_n=-\frac{\mathrm{i}}{\hbar}[p_n,H]\quad\text{and}\quad[q_n,q_m]=[p_n,p_m]=0,[q_n,p_m]=\mathrm{i}\hbar\delta_{nm} \tag{11.5.65}$$

which gives us:

$$\dot{\Phi}^A=-\frac{\mathrm{i}}{\hbar}\left[\Phi^A,H\right],\dot{\prod}^A=-\frac{\mathrm{i}}{\hbar}\left[\prod^A,H\right]\quad\text{and}\quad\left[\Phi^A,\Phi^B\right]=\left[\prod^A,\prod^B\right]=0,\left[\Phi^A,\prod^B\right]=\mathrm{i}\hbar\delta^3_{AB} \tag{11.5.66}$$

If we summarize a little and we display the comparison with Wave Quantum Physics, we have finally:

1. In Wave Quantum Physics (it's nice to look at it isn't it?):

$$\boxed{\begin{array}{l}p_n=\dfrac{\partial L}{\partial \dot{q}_n}\quad\text{and}\quad H=\sum p_n\dot{q}_n-L\\ \begin{cases}\dot{q}_n=\dfrac{\partial H}{\partial p_n}\\ \dot{p}_n=-\dfrac{\partial H}{\partial q_n}\end{cases}\quad\text{and}\quad\begin{cases}\dot{q}_n=-\dfrac{\mathrm{i}}{\hbar}[q_n,H]\\ \dot{p}_n=-\dfrac{\mathrm{i}}{\hbar}[p_n,H]\end{cases}\\ \begin{cases}[q_n,q_m]=[p_n,p_m]=0\\ [q_n,p_m]=\mathrm{i}\hbar\delta_{nm}\end{cases}\end{array}} \tag{11.5.67}$$

2. And the equivalent by the principle of correspondence in Quantum Field Theory (here it don't just look nice... it becomes art!):

$$\boxed{\begin{array}{l} \prod_A=\dfrac{\partial L}{\partial\dot{\Phi}^A} \quad \text{and} \quad H=\displaystyle\int\prod_A\dot{\Phi}^A\mathrm{d}^4x-L \\ \begin{cases}\Phi^A=\dfrac{\partial H}{\partial\prod_A}\\ \dot{\prod}_A=-\dfrac{\partial H}{\partial\Phi^A}\end{cases} \quad \text{and} \quad \begin{cases}\dot{\Phi}^A=-\dfrac{\mathrm{i}}{\hbar}\left[\Phi^A,H\right]\\ \dot{\prod}^A=-\dfrac{\mathrm{i}}{\hbar}\left[\prod^A,H\right]\end{cases} \\ \begin{cases}\left[\Phi^A,\Phi^B\right]=\left[\prod^A,\prod^B\right]=0\\ \left[\Phi^A,\prod^B\right]=\mathrm{i}\hbar\delta^3_{AB}\end{cases}\end{array}} \qquad (11.5.68)$$

And voila! We have just passed from the parameters of Wave Quantum Physics where the punctual bodies are described by wave functions, to a Quantum Physics where the punctual bodies become continuous fields.

It remains only to apply this general scheme to concrete examples:

We will start with a first example taking into account the relativistic aspect. Thus, the simplest non-trivial Lagrangian density that we can construct is of the form (you will see immediately what it will lead us to, which will confirm its theoretical validity - moreover, the development that follows could very well have been presented backwards):

$$\boxed{L=-\frac{1}{2}\partial_\mu\Phi^A\partial^\mu\Phi^A-\frac{1}{2}m_0^2(\Phi^A)^2} \qquad (11.5.69)$$

Or more explicitly:

$$L=\underbrace{-\frac{1}{2}\partial_0\Phi^A\partial^0\Phi^A}_{\text{kinetic energy term of }L}-\underbrace{\left(\frac{1}{2}\underbrace{\partial_k\Phi^A\partial^k\Phi^A}_{\substack{\text{Einstein summation}\\\text{only on the}\\\text{spatial components}}}+\underbrace{\frac{1}{2}m_0^2(\Phi^A)^2}_{\text{mass term}}\right)}_{\text{potential energy term of }L}=T-V \qquad (11.5.70)$$

that the physicists name "scalar field for a free and spinless particle" or "Klein-Gordon Lagrangian" for a spin-free particle where we use the usual condensed notations:

$$\partial_\mu\Phi^A=\frac{\partial\Phi^A}{\partial x_\mu} \qquad \partial_\mu\Phi^A\partial^\mu\Phi^A=\sum_\mu\partial_\mu\Phi^A\partial^\mu\Phi^A \qquad (11.5.71)$$

and the natural units:

$$\hbar=c=1 \qquad (11.5.72)$$

Otherwise in the S.I. units, we have:

$$L=-\frac{1}{2}\partial_\mu\Phi^A\partial^\mu\Phi^A-\frac{c^2m_0^2}{2\hbar^2}(\Phi^A)^2 \qquad (11.5.73)$$

With c and $\hbar$ set to 1, energy has the unit of $[\text{m}^{-1}]$ and the 4D integral of the Lagrangian density is dimensionless as $\hbar$. It means that the Lagrangian density is of the unit $[\text{m}^{-4}]$, so $\partial_\mu\phi$ is of the unit $[\text{m}^{-2}]$ and ϕ of the unit $[\text{m}^{-1}]$.

In S.I. units we start with the fact that that $\mathcal{L}$ has units of energy density (i.e. $[\text{J}\cdot\text{m}^{-3}]$). Hence using the fact that ${}^{c^2m_0^2}/_{2\hbar^2}(\Phi^A)^2$ must have the same units, we conclude that Φ^A has for units $[\text{J}^{-1/2}\cdot\text{m}^{-1/2}]$.

Let us calculate the Euler-Lagrange equation relative to it (normally it is trivial) without forgetting that it is a functional derivative, which greatly simplifies the calculation:

$$\frac{\partial L}{\partial\Phi^A}-\partial_\mu\left(\frac{\partial L}{\partial(\partial_\mu\Phi^A)}\right)=-m_0^2\Phi^A+\partial_\mu\partial^\mu\Phi^A \tag{11.5.74}$$

Following the request of a reader here are the details allowing to arrive at this result:

$$\begin{aligned}
&\frac{\partial L}{\partial\Phi^A}-\partial_\mu\left(\frac{\partial L}{\partial(\partial_\mu\Phi^A)}\right)\\
&=\frac{\partial\left(-\frac{1}{2}\partial_0\Phi^A\partial^0\Phi^A-\left(\frac{1}{2}\partial_\mu\Phi^A\partial^\mu\Phi^A+\frac{1}{2}m_0^2(\Phi^A)^2\right)\right)}{\partial\Phi^A}\\
&-\partial_\mu\left(\frac{\partial\left(-\frac{1}{2}\partial_0\Phi^A\partial^0\Phi^A-\left(\frac{1}{2}\partial_\mu\Phi^A\partial^\mu\Phi^A+\frac{1}{2}m_0^2(\Phi^A)^2\right)\right)}{\partial(\partial_\mu\Phi^A)}\right)\\
&=\frac{\partial\left(-\frac{1}{2}m_0^2(\Phi^A)^2\right)}{\partial\Phi^A}-\partial_\mu\left(\frac{\partial\left(-\frac{1}{2}\partial_\mu\Phi^A\partial^\mu\Phi^A\right)}{\partial(\partial_\mu\Phi^A)}\right)\\
&=-\frac{1}{2}m_0^2\frac{\partial\left((\Phi^A)^2\right)}{\partial\Phi^A}+\frac{1}{2}\partial_\mu\left(\frac{\partial\left(\partial_\mu\Phi^A\partial^\mu\Phi^A\right)}{\partial(\partial_\mu\Phi^A)}\right)\\
&=-m_0^2\Phi^A+\frac{1}{2}\partial_\mu\left(\partial^\mu\Phi^A\right)=-m_0^2\Phi^A+\frac{1}{2}\partial_\mu\partial^\mu\Phi^A
\end{aligned} \tag{11.5.75}$$

Hence the equation of motion in natural units and daring a dangerous notation for the double partial derivative (the Einstein summation is then implicit...):

$$\left(\partial_\mu^2-m_0^2\right)\Phi^A=0 \tag{11.5.76}$$

or as we will quick see it, in order to fall back on the results obtained in the Relativistic Quantum Physics section, we are obliged to introduce the contravariant and covariant differential operator (see section Tensor Calculus page 2059) with the signature $(-,+,+,+)$:

$$\partial^\mu=\left(\frac{\partial^2}{c^2\partial t^2},\frac{\partial^2}{\partial x^2},\frac{\partial^2}{\partial y^2},\frac{\partial^2}{\partial z^2}\right)=\left(\frac{\partial^2}{c^2\partial t^2},\vec{\nabla}^2\right) \tag{11.5.77}$$

$$\partial_\mu=\eta_{\mu\nu}\partial^\nu=\left(-\frac{\partial^2}{c^2\partial t^2},\frac{\partial^2}{\partial x^2},\frac{\partial^2}{\partial y^2},\frac{\partial^2}{\partial z^2}\right)=\left(\frac{-\partial^2}{c^2\partial t^2},\vec{\nabla}^2\right) \tag{11.5.78}$$

Let us recall now for comparison purpose that in the section of Wave Quantum Physics we had obtained the following free Klein-Gordon equation:

$$\left(\hbar^2\Box - m_0^2c^2\right)\Psi = 0 \tag{11.5.79}$$

also with the signature $(-,+,+,+)$ thus with the Alembertien operator:

$$\Box = -\frac{\partial^2}{c^2\partial t^2}+\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2} \tag{11.5.80}$$

We then have a perfect correspondence between the free Klein-Gordon equation and the field equation (written below in natural units as usual):

$$\left(\partial_\mu^2 - m_0^2\right)\Psi = 0 \quad \text{and} \quad \left(\partial_\mu^2 - m_0^2\right)\Phi^A = 0 \tag{11.5.81}$$

and it is here that one can eventually feel a shiver in the back and stay admiring in front of the power of mathematical formalism opening new perspectives on how to see the gears of our Universe... Thus, in Quantum Field Theory, the Klein-Gordon equation can be reinterpreted as a field equation!

The form of the free Klein-Gordon equation involving fields is sometimes named the "Klein-Gordon field equation".

Remark

We can generalize the Klein-Gordon field equation to curved space. Indeed, as:

$$\partial_\mu\Phi^A\partial^\mu\Phi^A = g_{\nu\mu}\partial^\nu\Phi^A\partial^\mu\Phi^A \tag{11.5.82}$$

and therefore:

$$L = -\frac{1}{2}g_{\nu\mu}\partial^\nu\Phi^A\partial^\mu\Phi^A - \frac{1}{2}m_0^2(\Phi^A)^2 \tag{11.5.83}$$

Then the action S must be corrected using the invariant volume introduced in the section of Tensor Calculus such that:

$$\boxed{S = \int \mathcal{L}\mathrm{d}x^4 = \int\left(-\frac{1}{2}g_{\nu\mu}\partial^\nu\Phi^A\partial^\mu\Phi^A - \frac{1}{2}m_0^2(\Phi^A)^2\right)\sqrt{g}\mathrm{d}x^4} \tag{11.5.84}$$

And even... better ... you will see, we'll do it a little in a blind way ... and therefore!!!: Let us now consider the following Lagrangian (which we assume will be obtained by successive trials and errors... but again we could have done the development backwards), intending to express the "Lagrangian interaction of an electromagnetic field with a current density":

$$\boxed{L = \frac{1}{4}F^{\mu\nu}F_{\mu\nu} - j^\nu A_\nu} \tag{11.5.85}$$

where we recognize there the tensors of the electromagnetic field proved and determined in the section of Electrodynamics (see page 4128) and Special Relativity and for which, for recall, we have:

$$F^{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \tag{11.5.86}$$

In this Lagrangian, let us treat the vector potential as a field such that:

$$\frac{\partial L}{\partial \Phi^A}-\partial_\mu\left(\frac{\partial L}{\partial(\partial_\mu \Phi^A)}\right)=\frac{\partial L}{\partial A_\nu}-\partial_\mu\left(\frac{\partial L}{\partial(\partial_\mu A_\nu)}\right)=0 \qquad (11.5.87)$$

Therefore, by decomposing the developments, we get very easily:

$$\frac{\partial(j^\nu A_\nu)}{\partial(\partial_\mu A_\nu)}=0 \quad \text{and} \quad \frac{\partial(j^\nu A_\nu)}{\partial A_\nu}=j^\nu \quad \text{and} \quad \frac{\partial(F^{\alpha\beta}F_{\alpha\beta})}{\partial A_\nu}=0 \qquad (11.5.88)$$

In a first step, the reader will verify by doing some relatively elementary tensor calculus that:

$$\frac{\partial F_{\alpha\beta}}{\partial(\partial_\mu A_\nu)}=\frac{\partial(\partial_\alpha A_\beta-\partial_\beta A_\alpha)}{\partial(\partial_\mu A_\nu)}=\delta_\alpha^\mu\delta_\beta^\nu-\delta_\beta^\mu\delta_\alpha^\nu \qquad (11.5.89)$$

Then:

$$\begin{aligned}\frac{\partial(F^{\alpha\beta}F_{\alpha\beta})}{\partial(\partial_\mu A_\nu)}&=F^{\alpha\beta}\frac{\partial F_{\alpha\beta}}{\partial(\partial_\mu A_\nu)}+\frac{\partial F^{\alpha\beta}}{\partial(\partial_\mu A_\nu)}F_{\alpha\beta}=2\frac{\partial F_{\alpha\beta}}{\partial(\partial_\mu A_\nu)}F^{\alpha\beta}\\&=2[\delta_\alpha^\mu\delta_\beta^\nu-\delta_\beta^\mu\delta_\alpha^\nu]F^{\alpha\beta}=2[F^{\mu\nu}-F^{\nu\mu}]=4F^{\mu\nu}\end{aligned} \qquad (11.5.90)$$

Therefore, the equation of the field is written:

$$\frac{\partial L}{\partial A_\nu}-\partial_\mu\left(\frac{\partial L}{\partial(\partial_\mu A_\nu)}\right)=j^\nu-\partial_\mu\left(\frac{1}{4}4F^{\mu\nu}\right)=0 \qquad (11.5.91)$$

hence:

$$\partial_\mu F^{\mu\nu}=j^\nu \qquad (11.5.92)$$

We must admit that the result is quite beautiful even if the approach is not the most rigorous one! We thus fall back on the Maxwell equation with sources with the same Lagrangian of the field (see section Electrodynamics page 4131). Thus, this massless Lagrangian is assimilated to the Lagrangian of the vector field of spin 1 assimilated to the bosons (see section Statistical Mechanics page 3571).

Let us recall now that we had obtained in the section on Electrodynamics (see page 4126) the following action for a charged particle in an electromagnetic field (before a long development which had brought us to the tensor of the electromagnetic field...):

$$S=\int_\alpha^\beta -m_0c\mathrm{d}s-qA^i\mathrm{d}x_i \qquad (11.5.93)$$

and remembering that (see section Electrodynamics page 4126):

$$\frac{\mathrm{d}s}{\mathrm{d}t}=c\sqrt{1-\frac{v^2}{c^2}} \qquad (11.5.94)$$

it comes:

$$S=\int_\alpha^\beta\left(-m_0c^2\sqrt{1-\frac{v^2}{c^2}}-qA^i\mathrm{d}v_i\right)\mathrm{d}t \qquad (11.5.95)$$

The corresponding Lagrangian density is therefore:

$$L = -\rho_0 c^2\sqrt{1-\frac{v^2}{c^2}} - j^\nu A_\nu \tag{11.5.96}$$

So we have finally:

1. The Lagrangian (Lagrangian density) of a charged particle in an electromagnetic field (which we have just obtained):

$$L = -\rho_0 c\sqrt{1-\frac{v^2}{c^2}} - j^\nu A_\nu \tag{11.5.97}$$

2. The Lagrangian (Lagrangian density) we get just before (which allowed us to fall back on Maxwell's equations with sources):

$$L = \frac{1}{4}F^{\mu\nu}F_{\mu\nu} - j^\nu A_\nu \tag{11.5.98}$$

Hence, it is natural to write the "Lagrangian (total Lagrangian density) of the electromagnetic field":

$$\boxed{L = -\rho_0 c^2\sqrt{1-\frac{v^2}{c^2}} - j^\nu A_\nu + \frac{1}{4\mu_0}F^{\mu\nu}F_{\mu\nu}} \tag{11.5.99}$$

Let us now continue our way with the free Dirac equation! Let us recall that in the Relativistic Quantum Physics section we obtained the free Dirac equation (see page 4931) in the form (basically, let us also recall that it is a relativistic equation):

$$(\mathrm{i}\gamma^\mu\partial_\mu - m_0)\,\Psi = (\mathrm{i}\not{\partial} - m_0)\,\Psi = 0 \tag{11.5.100}$$

Now let us recall (see section Linear Algebra page 1951) that $(AB)^\dagger = B^\dagger A^\dagger$. Therefore, it comes:

$$((\mathrm{i}\not{\partial} - m_0)\,\Psi)^\dagger = 0^\dagger \Rightarrow \Psi^\dagger\,(\mathrm{i}\not{\partial} - m_0)^\dagger = \Psi^\dagger\left(\mathrm{i}^\dagger\not{\partial}^\dagger - m_0\right) = \Psi^\dagger\left(-\mathrm{i}\gamma^{\mu\dagger}\partial_\mu^\dagger - m_0\right) = 0 \tag{11.5.101}$$

where we have introduced the "Feynman slash notation" $\not{\partial}$ (less commonly known as the "Dirac slash notation").

But, $\partial_\mu^\dagger = \partial_\mu$ and it is super easy to verify (do not forget that we use the Dirac representative form of Pauli matrices!!!):

$$\gamma^{\mu\dagger} = \gamma^0\gamma^\mu\gamma^0 \tag{11.5.102}$$

which leads us to write:

$$\Psi^\dagger\left(-\mathrm{i}\gamma^0\gamma^\mu\gamma^0\partial_\mu - m_0\right) = 0 \tag{11.5.103}$$

It is then convenient to introduce the "Dirac adjoint" or "adjoint spinor":

$$\boxed{\bar{\Psi} = \bar{\Psi}^T\gamma^0 = \Psi^\dagger\gamma^0 = \left(\bar{\Psi}_1, \bar{\Psi}_2, -\bar{\Psi}_3, -\bar{\Psi}_4\right)} \tag{11.5.104}$$

Sadly this notation may be misleading ast the adjoint spinor $\bar{\Psi}$ can be confused with the complex conjugate $\bar{\Psi}$ because they have exactly the same notation! That's why in many textbooks the above is instead denoted as:

$$\bar{\Psi}=\Psi^{\dagger}\gamma^0=(\Psi^*)^T\gamma^0=(\Psi^*,\Psi^*,\Psi^*,\Psi^*)\begin{pmatrix}1&0&0&0\\0&1&0&0\\0&0&-1&0\\0&0&0&-1\end{pmatrix}\tag{11.5.105}$$

Remark

Let us recall that Ψ is a matrix-column and $\Psi^{\dagger}$ a matrix-line. So it comes that $\bar{\Psi}$ is also a matrix-line!

Using the fact that in the Dirac representation $\gamma^0\gamma^0=\mathbb{1}$ we can write:

$$\begin{aligned}&\Psi^{\dagger}\left(-\mathrm{i}\gamma^0\gamma^{\mu}\gamma^0\partial_{\mu}-m_0\mathbb{1}\right)=0\\&\Rightarrow\Psi^{\dagger}\left(-\mathrm{i}\gamma^0\gamma^{\mu}\gamma^0\partial_{\mu}-m_0\gamma^0\gamma^0\right)=0\\&\Rightarrow\Psi^{\dagger}\gamma^0\left(-\mathrm{i}\gamma^{\mu}\gamma^0\partial_{\mu}-m_0\gamma^0\right)=0\\&\Rightarrow\bar{\Psi}\left(-\mathrm{i}\gamma^{\mu}\gamma^0\partial_{\mu}-m_0\gamma^0\right)=0\\&\Rightarrow\bar{\Psi}\left(\mathrm{i}\gamma^{\mu}\gamma^0\partial_{\mu}+m_0\gamma^0\right)=0\end{aligned}\tag{11.5.106}$$

by simplifying the γ^0 it comes the "free adjoint Dirac equation":

$$\bar{\Psi}\left(\mathrm{i}\gamma^{\mu}\partial_{\mu}+m_0\right)=\bar{\Psi}\left(i\not{\partial}+m_0\right)=0\tag{11.5.107}$$

What we traditionally write:

$$\boxed{\bar{\Psi}\left(\mathrm{i}\overleftarrow{\not{\partial}}+m_0\right)=0}\tag{11.5.108}$$

The notation $\overleftarrow{\not{\partial}}$ meaning that the operator $\not{\partial}$ operates on $\bar{\Psi}$ on the left such that:

$$\bar{\Psi}\overleftarrow{\not{\partial}}=\left(\partial_{\mu}\bar{\Psi}\right)\gamma^{\mu}\tag{11.5.109}$$

Remark

Some authors write $(\mathrm{i}\not{\partial}+m_0)\bar{\Psi}=0$ but this is false because $\bar{\Psi}$ is a line matrix as we pointed out above !!!

Finally we have for the free Dirac equations:

$$\boxed{\begin{aligned}\left(\mathrm{i}\overleftarrow{\not{\partial}}+m_0\right)\Psi=0\Leftrightarrow\mathrm{i}\gamma^{\mu}\partial_{\mu}\Psi-m_0\Psi=0\\\bar{\Psi}\left(\mathrm{i}\overleftarrow{\not{\partial}}+m_0\right)=0\Leftrightarrow\mathrm{i}\left(\partial_{\mu}\bar{\Psi}\right)\gamma^{\mu}+m_0\Psi=0\end{aligned}}\tag{11.5.110}$$

Let us now suppose that the "Lagrangian of the free Dirac spinor field" is of the form (because in the end it is the Lagrangian that interests us!):

$$\boxed{L=\bar{\Psi}\left(\mathrm{i}\gamma^{\mu}\partial_{\mu}-m_0\right)\Psi}\tag{11.5.111}$$

where we have for recall $\bar{\Psi}=\Psi^{\dagger}\gamma^0$. It is therefore the Lagrangian of the spinor field for the particles of spin $1/2$ which are therefore free fermions.

Remark

Now, I take Ψ and rotate it by some angle in the complex plane. This means that I multiply it by $e^{\mathrm{i}\theta}$ for some angle θ. Note that the rotation is not a physical rotation in space-time, it's just a redefinition of the values of the field Ψ. You can think of it as a rotation in some internal mathematical space.

Recall that the kinetic term in the Lagrangian has a product of Ψ and its complex conjugate $\bar{\Psi}$. If I replace Ψ with $e^{\mathrm{i}\theta}\Psi$, I also need to replace $\bar{\Psi}$ with $\bar{\Psi}e^{-\mathrm{i}\theta}$, where the minus sign in the exponent comes from complex conjugation.

Therefore the product $\bar{\Psi}\Psi$ becomes $\bar{\Psi}e^{-\mathrm{i}\theta}e^{\mathrm{i}\theta}\Psi$. Obviously, the two exponents cancel each other, and we get $\bar{\Psi}\Psi$ again.

This means that the Lagrangian is invariant to rotation of the field Ψ by a complex phase. A fancier way of saying this is that the Lagrangian possesses $U(1)$ symmetry. $U(1)$ is the circle group, or the group of complex numbers with absolute value 1, that is, the complex numbers of the form $e^{\mathrm{i}\theta}$.

Now, remember that Noether's theorem (see section Mechanics page 3132) says that for any symmetry of the Lagrangian there is a conserved quantity. This symmetry is, of course, no exception as it is the symmetry for the conservation of charge!

Considering the quantities $\bar{\Psi}$, Ψ as independent (this is what they are anyway since they are orthogonal) and choosing the spinor field as being $\Psi^{\dagger}$, we have the Euler-Lagrange equation:

$$\frac{\partial L}{\partial\bar{\Psi}}-\partial_{\mu}\left(\frac{\partial L}{\partial(\partial_{\mu}\bar{\Psi})}\right)=0 \tag{11.5.112}$$

The second term is equal to zero since the Dirac Lagrangian does not contain terms with $\partial_{\mu}\Psi^{\dagger}$. In fact it remains:

$$\frac{\partial L}{\partial\bar{\Psi}}=\frac{\partial\left(\bar{\Psi}\mathrm{i}\gamma^{\mu}\partial_{\mu}\Psi-\bar{\Psi}\right)}{\partial\bar{\Psi}}=\mathrm{i}\gamma^{\mu}\partial_{\mu}\Psi-\Psi m_0=(\mathrm{i}\not{\partial}-m_0)\,\Psi=0 \tag{11.5.113}$$

So we fall back on the free Dirac equation (the same development that can be done for the free adjoint Dirac equation)! Thus, in this framework, the only way to explain the quantum properties of the material containing particles with spin $1/2$ is to involve the fields $\bar{\Psi}$, Ψ representing electrically charged particles, electrons and positrons as we know. We then name then these entities "Dirac field" also named "spinor field" or spin $1/2$ fermion field ".

Remark

Often, when using the Dirac equation one finds the slash notation introduced in the section of Relativistic Quantum Physics used on four-momentum. Using the Dirac basis for the gamma matrices:

$$\gamma^0=\begin{pmatrix}\mathbb{1} & 0\\ 0 & -\mathbb{1}\end{pmatrix}\qquad \gamma^i=\begin{pmatrix}0 & \sigma^i\\ -\sigma^i & 0\end{pmatrix} \tag{11.5.114}$$

as well as the definition of four momentum (see section Special Relativity page 5687):

$$p_{\mu}=(E,-p_x,-p_y,-p_z) \tag{11.5.115}$$

we see explicitly that:

$$\not{p}=\gamma^\mu p_\mu=\gamma^0 p_0+\gamma^i p_i=\begin{bmatrix} p_0 & 0 \\ 0 & -p_0 \end{bmatrix}+\begin{bmatrix} 0 & \sigma^i p_i \\ -\sigma^i p_i & 0 \end{bmatrix}=\begin{bmatrix} E & -\sigma\circ\vec{p} \\ \sigma\circ\vec{p} & -E \end{bmatrix} \tag{11.5.116}$$

Similar results hold in other bases, such as the Weyl basis.

Now, to make the reader able to understand some other traditional notation used by practitioners (and useful to simplified some proofs), let us introduce matrix γ_5 taken to be:

$$\gamma_5=\begin{bmatrix} -\mathbb{1} & 0 \\ 0 & \mathbb{1} \end{bmatrix} \tag{11.5.117}$$

Let us write the four component Dirac spinors as sum of left and right chirality components, i.e.:

$$\begin{aligned} \Psi_L&=\frac{1}{2}\left(1-\gamma^5\right)\Psi \\ \Psi_R&=\frac{1}{2}\left(1+\gamma^5\right)\Psi \end{aligned} \tag{11.5.118}$$

where the 1 above represents a 4×4 diagonal matrix with only 1 in the diagonal!

Adopting the 'spin-up', 'spin-down' notation in quantum mechanics, we can write :

$$\Psi_L=\begin{pmatrix} \Psi_{L_\uparrow} \\ \Psi_{L_\downarrow} \end{pmatrix} \text{ and } \Psi_R=\begin{pmatrix} \Psi_{R\uparrow} \\ \Psi_{R_\downarrow} \end{pmatrix} \tag{11.5.119}$$

Remark

We define therefore the "projection operators" as:

$$\hat{P}_L=\frac{1-\gamma_5}{2},\qquad \hat{P}_R=\frac{1+\gamma_5}{2} \tag{11.5.120}$$

as (therefore they commute) $\hat{P}_L\hat{P}_R=\hat{P}_R\hat{P}_L=0$.

We can write:

$$\Psi=\Psi_L+\Psi_R \tag{11.5.121}$$

Substituting this in the above Lagrangian we get:

$$\begin{aligned} L&=\left(\bar{\Psi}_L+\bar{\Psi}_R\right)\left(\gamma^\mu \mathrm{i}\partial_\mu-m\right)\left(\Psi_L+\Psi_R\right) \\ &=\left(\bar{\Psi}_L+\bar{\Psi}_R\right)\left(\gamma^\mu \mathrm{i}\partial_\mu\right)\left(\Psi_L+\Psi_R\right)-m\left(\bar{\Psi}_L+\bar{\Psi}_R\right)\left(\Psi_L+\Psi_R\right) \\ &=\left(\Psi_L^\dagger\gamma^0+\Psi_R^\dagger\gamma^0\right)\left(\gamma^\mu \mathrm{i}\partial_\mu\right)\left(\Psi_L+\Psi_R\right)-m\left(\bar{\Psi}_L\Psi_L+\bar{\Psi}_R\Psi_L+\bar{\Psi}_L\Psi_R+\bar{\Psi}_R\Psi_R\right) \\ &=\left(\Psi_L^\dagger\gamma^0\left(\gamma^\mu \mathrm{i}\partial_\mu\right)\Psi_L+\Psi_L^\dagger\gamma^0\left(\gamma^\mu \mathrm{i}\partial_\mu\right)\Psi_R+\Psi_R^\dagger\gamma^0\left(\gamma^\mu \mathrm{i}\partial_\mu\right)\Psi_L+\Psi_R^\dagger\gamma^0\left(\gamma^\mu \mathrm{i}\partial_\mu\right)\Psi_R\right)-m \\ &\left(\bar{\Psi}_L\Psi_L+\bar{\Psi}_R\Psi_L+\bar{\Psi}_L\Psi_R+\bar{\Psi}_R\psi_R\right) \end{aligned} \tag{11.5.122}$$

Now using above definition of Ψ_L and Ψ_R and using:

$$\gamma^5\gamma^0=-\gamma^0\gamma^5 \text{ and } \gamma^5\gamma^\mu=-\gamma^\mu\gamma^5 \tag{11.5.123}$$

We get:

$$L=\bar{\Psi}_L\gamma^\mu i\partial_\mu\Psi_L+\bar{\Psi}_R\gamma^\mu i\partial_\mu\Psi_R-m\left(\bar{\Psi}_R\psi_L+\bar{\Psi}_L\Psi_R\right) \tag{11.5.124}$$

At this point, the reader must know, that at our actual level of knowledge, all fields have particles, while all particles are quanta of fields. That's why the equations used to describe the physics of the Standard Model are named "Quantum Field Theory", not "Quantum Particle Theory".

11.5.2.1 Field interactions ($\lambda\varphi^4$ theory)

Let us remind that we have derive earlier above in natural units:

$$L=\underbrace{-\frac{1}{2}\partial_0\Phi^A\partial^0\Phi^A}_{\text{kinetic energy term of } L}-\underbrace{\left(\frac{1}{2}\underbrace{\partial_k\Phi^A\partial^k\Phi^A}_{\substack{\text{Einstein summation}\\\text{only on the}\\\text{spatial components}}}+\underbrace{\frac{1}{2}m_0^2(\Phi^A)^2}_{\text{mass term}}\right)}_{\text{potential energy term of } L}=T-V \tag{11.5.125}$$

Also often denoted:

$$\mathcal{L}=T-V=\frac{1}{2}\partial_\mu\varphi\partial^\mu\varphi-V(\varphi) \tag{11.5.126}$$

Now to consider physically interesting theories we need to specialize or restrict ourselves to some specific forms of the potential function, $V(\varphi)$. One physical such restriction is that the potential should be an analytic function in φ. At the level of classical theory one might not be able to appreciate the significance of this restriction, but in quantum theory it will become evident that non-analytic potentials or terms in the action will imply or result in breakdown of effective field theory description at some regions of moduli space. Foreseeing this situation we will strictly consider analytic potentials, $V(\varphi)$ and since analytic functions as a power series:

$$V(\varphi)=\sum_{n=0}^{+\infty}a_n\varphi^n=a_0+a_1\varphi+a_2\varphi^2+a_3\varphi^3+a_4\varphi^4+\ldots:=\sum_{n=0}^{+\infty}\frac{\lambda_n}{n!}\varphi^n \tag{11.5.127}$$

Now the odd-power terms, $V_{\text{odd}}=a_1\varphi, a_3\varphi^3$ etc. have a pathological property, namely, $V_{\text{odd}}\rightarrow-\infty$ as $\varphi\rightarrow+\infty$.

In Quantum Field Theory (QFT), the behavior of terms in the potential, particularly odd and even powers of fields, can often be understood through symmetry considerations and stability arguments:

- Symmetry in the Lagrangian: Typically, the potential in QFT is constructed to respect certain symmetries of the theory. For example, if the theory involves a scalar field, a common choice is to impose symmetry under a transformation like ($\varphi \rightarrow -\varphi$) (parity symmetry). This symmetry forbids odd powers of the field in the potential because such terms would break the symmetry. Odd terms, when present, often introduce asymmetries that can lead to inconsistencies or instabilities in the field dynamics, particularly under quantization or renormalization.

- Stability Considerations: For a potential to describe a stable system, it needs to have a minimum where the field can "sit." If odd powers dominate, they can lead to a potential that is not bounded from below, creating pathological behavior essentially, the energy becomes unbounded, leading to physical instability. Even terms are preferred as they allow for stable minima and are more consistent with renormalizable interactions.

Pathologies in odd terms often arise in specific contexts where symmetry or stability conditions are violated, leading to unphysical or non-viable theories.

Thus if we keep such odd power terms in the potential energy function, V, they will lead to the potential having no lower bound. Systems which have potentials with no lower bound have no stable ground state, and hence are considered nonphysical. Since we are interested with physical systems we will omit such odd powered terms in the potential function:

$$V(\varphi)=a_2\varphi^2+a_4\varphi^4+\ldots \tag{11.5.128}$$

and we put for the the coefficients, $a_2=\frac{m}{2}$, and $a_4=\frac{\lambda}{4!}$, to make the potential look consistent with the literature:

$$V(\varphi)=\frac{m}{2}\varphi^2+\frac{\lambda}{4!}\varphi^4 \tag{11.5.129}$$

The quadratic piece, $\frac{m}{2}\varphi^2$ is referred to as the "mass term" and the quartic piece, $\frac{\lambda}{4!}\varphi^4$ is referred to as the "(self) coupling term" or "(self) interaction term".

The nomenclature will become obvious further below. It will turn out this restriction to quadratic and quartic potentials will cover a large class of physical systems especially in particle physics and statistical mechanics and is a widely studied system which goes by the name of "$\lambda\varphi^4$ theory". Terms with higher powers of φ, are also undesirable in a fundamental theory since they will lead to quantum theory with no predictive power (non-renormalizable), so we will not consider them. However, for effective field theories such nonrenormalizable terms are fine.

Let's determine the (mass) dimensions of the field, φ, the constants in the potential, m,λ!

Remember that the action has the dimensions of J · s which is identical to the unit of angular momentum and that Planck constant also has the units of an angular momentum. Therefore in natural units, as Planck constant is without dimensions, then the action S is also without dimension in natural units!

From the speed of light being unitless in natural units, we get, $[\mathrm{L}]=[\mathrm{T}]$ (length is time).

We have unit-wise $E=mc^2$ meaning that in natural units $E=m=[\mathrm{J}]$. Furthermore we have $hc=[\mathrm{J}\cdot\mathrm{m}]=[1]$ in natural units, so distance is the same units as energy inverse here, which is same as mass inverse, i.e. $[\mathrm{M}]=[\mathrm{L}]^{-1}$.

That being said, to determine the dimensions of φ, we look at the kinetic term in the action:

$$S=\int\frac{1}{2}\partial_{\mu}\varphi\partial^{\mu}\varphi\mathrm{d}x^4+\ldots \qquad (11.5.130)$$

This means the RHS should have dimension zero as well. Let's look at the dimensions of the RHS. The integration measure, d^4x of course has dimensions as four powers of length, $\mathrm{d}^4x=[\mathrm{L}]^4$. Similarly the derivatives have dimensions of inverse length, $\partial=[\mathrm{L}]^{-1}$. Thus we have:

$$\left[\mathrm{d}^4x\frac{1}{2}\partial_{\mu}\varphi\partial^{\mu}\varphi\right]=\left[\mathrm{L}^2\varphi^2\right] \qquad (11.5.131)$$

But as in natural units $[\mathrm{L}]^2=[\mathrm{M}]^{-2}$ we have therefore that φ has the same dimension as mass in natural units.

These naive dimensions we are determining based on dimensional analysis are referred to the "classical dimensions" or "engineering dimensions" of φ.

Having determined the dimension of the field φ, it will now be easy to determine the engineering dimensions of the constants, m and λ. Again we start from the fact that the action is dimensionless, so the dimension of each term from the potential energy integral should vanish as well:

$$\left[\int\frac{m}{2}\varphi^2\mathrm{d}^4x\right]=[1] \qquad \left[\int\frac{\lambda}{4!}\varphi^4\mathrm{d}^4x\right]=[1] \qquad (11.5.132)$$

These two conditions immediately imply:

$$[m]=[\mathrm{kg}] \qquad [\lambda]=[1] \qquad (11.5.133)$$

Since the constant, m has dimensions of mass, we are justifying in referring the term $\frac{m}{2!}\varphi^2$ in the lagrangian as the mass term. Later on, when we quantize this theory we will see that indeed, m is (related to) the mass of the quantum of this field (particle). The quartic coupling constant, λ is dimensionless (in $3+1$ space-time dimensions only, in higher or lower dimensions it will have non-zero dimensions).

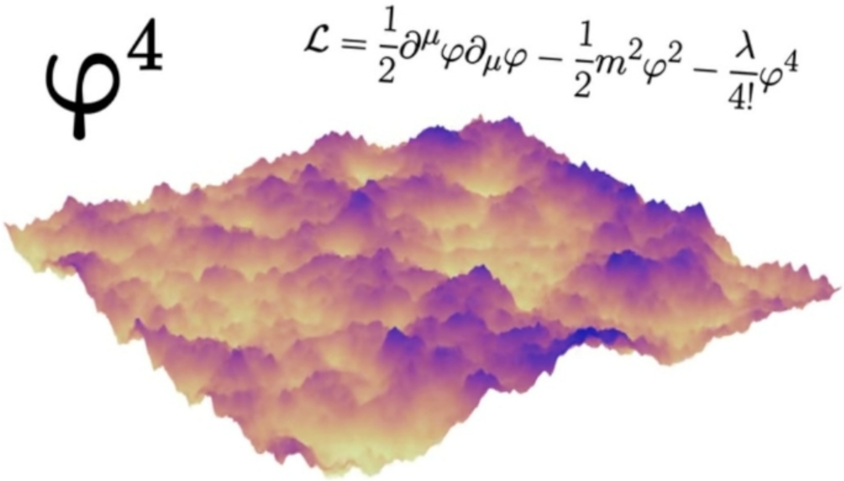

φ^4
$\mathcal{L}=\frac{1}{2}\partial^{\mu}\varphi\partial_{\mu}\varphi-\frac{1}{2}m^2\varphi^2-\frac{\lambda}{4!}\varphi^4$

Figure 11.148: Artistic representation of quantum fluctuation in an interacting field (source: ScienceClic, author: Alessandro Roussel)

11.5.3 Gauge Theories

We will now see a simple approach to a tool that revolutionized the approach of modern particle physics in the mid-120th century (holocene calendar) and that has earned several Nobel Prizes to those who contributed to it.

We advise very strongly before reading what follows that the reader also take a preliminary look at the sub-section of Gauge Theory of the Electrodynamics section (see page 4115), as it contains a first example of an invariance of Gauge making emerging a field (the vector potential) necessary to explain certain phenomena at the quantum scale as clearly explained by Pauli's equation (see section Relativistic Quantum Physics page 4979).

Since the early 11980s (holocene calendar), popularization magazines have talked a lot when dealing about Quantum Physics on Gauge theories. Electromagnetic interactions and Weak interactions are described jointly by a Gauge theory developed by Glashow, Weinberg and Salam. Strong interactions also seem correctly described by a Gauge theory. It is within the framework of these gauge theories that theoretical physicists try to unify the various fundamental interactions of nature. It is therefore appropriate, even in this textbook which treats in an elementary way of Quantum Physics, to speak of Gauge theory within the framework of this study domain.

To do this, we will already consider the context that led to the discovery of Gauge invariance in the context of Electrodynamics (see the section of the same name at page 4115 for the details) and make a comparison with certain developments seen in the section of General Relativity and the role played by Weyl in the proof of the fundamental principles of a Gauge theory.

Let us recall that the Special Relativity and General Relativity are based on the premise that there is no absolute reference frame in the universe. We have seen in the section of Special Relativity in a long and broad way that the relations which make it possible to pass the laws of physics from one reference frame to another one depend only on the relative speed between them. Thus, Special Relativity is a theory with global symmetry. We have also seen in the section of General Relativity that the affine connection is the link between the referentials of the local theory (weak field approximation) that is General Relativity.

In 11919 (holocene calendar), the first experimental observation of the deviation of the light of a star by the gravitational field of the Sun (actually a G2V-type star, aka yellow dwarf) took place. This spectacular confirmation of the theory of General Relativity inspired Hermann Weyl, who proposed the same year a revolutionary conception of Gauge invariance: If the effects of a gravitational field can be described by a connection expressing the relative orientation between local repositories of space-time, can other forces of nature such as electromagnetism also be associated with similar connections?

We consider two types of Gauge symmetry: one known as "global Gauge" and the other named "local Gauge". They are distinguished by the parameter characterizing the phase change of the wave function (we will see this in detail further below).

11.5.3.1 Global Gauge invariance

We will therefore study the Gauge invariance using the Schrödinger equation and show that even if the results may seem confusing (in the context of complex set $\mathbb{C}$ applications), they nevertheless remain mathematically correct!

Remark

The global gauge invariance is rigorously referred to as "global symmetry".

Let us consider the Schrödinger's equation:

$$H\Psi=\mathrm{i}\hbar\frac{\partial}{\partial t}\Psi \tag{11.5.134}$$

with as we have shown:

$$H=-\frac{\hbar^2}{2m}\vec{\nabla}^2+V \tag{11.5.135}$$

with $\Psi=\Psi(\vec{r},t)$. Either in the case of a free particle:

$$H=-\frac{\hbar^2}{2m}\vec{\nabla}^2=\hbar^2\frac{(-\mathrm{i}\vec{\nabla})^2}{2m} \tag{11.5.136}$$

This operator is obviously invariant in the transformation which passes from Ψ to Ψ' with (change of the phase of a plane wave by an angle α):

$$\Psi'=e^{\mathrm{i}g\alpha}\Psi\Rightarrow\Psi=e^{-\mathrm{i}g\alpha}\Psi' \tag{11.5.137}$$

where g is a coupling constant (to ensure the homogeneity of the units and the amplitude) being considered as a real number and α a real parameter independent of the coordinates (in a first time ...) of space and time.

Then:

$$H\Psi=\mathrm{i}\hbar\frac{\partial}{\partial t}\Psi \tag{11.5.138}$$

becomes:

$$He^{-\mathrm{i}g\alpha}\Psi'=\mathrm{i}\hbar\frac{\partial}{\partial t}e^{-\mathrm{i}g\alpha}\Psi' \tag{11.5.139}$$

and as α does not depend on x, y, z, t then:

$$e^{-\mathrm{i}g\alpha}H\Psi'=e^{-\mathrm{i}g\alpha}\mathrm{i}\hbar\frac{\partial}{\partial t}\Psi' \tag{11.5.140}$$

Either after simplification:

$$H\Psi'=\mathrm{i}\hbar\frac{\partial}{\partial t}\Psi' \tag{11.5.141}$$

The form of the equation remained the same when we made the change from Ψ to Ψ'.

Thus, the description of a free system is not affected by the global phase change. In the language of group theory (see section Set Algebra page 1424), we speak of invariance under the $\text{U}(1)$ group of phases.

In other words to speak like the physicists...:

$$e^{\text{i}g\alpha}\Psi \tag{11.5.142}$$

defines a Gauge transformation by the rotation α (the parameter in the sense of the Lie groups) in the complex plane (but that corresponds to a phase change in the field of wave theory).

The set of rotations forms a group denoted $\text{U}(1)$ which the usage make we name it the "gauge group" (isomorphic to $\text{SO}(2)$ as we have seen it in the section of Set Algebra).

The set of all $e^{\text{i}g\alpha}$ forms a monodimensional representation of the group $\text{U}(1)$ which we name the "representation g". There is, of course, an infinity of representations g (as much as there are values of g!).

Since the parameter α does not depend on position and time, we say that the system is "invariant by global gauge" transformation (everywhere at the same time) or simply an invariant of $\text{U}(1)$ in time and space.

Remark

In fact, global Gauge transformations are a subset of local gauge transformation: changing the same amount everywhere is a special case (i.e., more restricting) of changing the phase of each point independently. A local Gauge transformation is therefore not a subset of a global Gauge transformation. In this respect the name is a bit misleading.

11.5.3.2 Local Gauge invariance

But but... either the global gauge invariance shows that we have an equation that remains valid as part of a fixed phase change. But now in a laboratory this Schrödinger equation must be valid even if the phase depends on position and time. This constraint is named a "local gauge invariance".

We consider this time that α is a function $\alpha(\vec{r},t)$ and the idea obviously is to check whether the Schrödinger equation remains invariant in the transformation:

$$\Psi \rightarrow \Psi' = e^{\mathrm{i}g\alpha}\Psi \qquad (11.5.143)$$

It is therefore obvious that Schrödinger's equation:

$$H\Psi = \mathrm{i}\hbar\frac{\partial}{\partial t}\Psi \qquad (11.5.144)$$

is no longer invariant. Indeed, we quickly see that just the operator $\vec{\nabla}^2$ in the Hamiltonian will pose problem by making appear annoying terms that will not cancel out:

$$\begin{aligned}
&\vec{\nabla}\Psi = e^{-\mathrm{i}g\alpha}\left(\vec{\nabla}\Psi'\right) - \mathrm{i}ge^{-\mathrm{i}g\alpha}\left(\vec{\nabla}\alpha\right)\Psi' \\
&\vec{\nabla}^2\Psi = \vec{\nabla}\left(e^{-\mathrm{i}g\alpha}\left(\vec{\nabla}\Psi'\right)\right) - \mathrm{i}g\vec{\nabla}\left(e^{-\mathrm{i}g\alpha}\left(\vec{\nabla}\alpha\right)\Psi'\right) \\
&= \left[e^{-\mathrm{i}g\alpha}\left(\vec{\nabla}^2\Psi'\right) - \mathrm{i}ge^{-\mathrm{i}g\alpha}\left(\vec{\nabla}\alpha\right)\left(\vec{\nabla}\Psi'\right)\right] - \mathrm{i}g\left[(-\mathrm{i}g)e^{-\mathrm{i}g\alpha}\left(\vec{\nabla}\alpha\right)\Psi' + e^{-\mathrm{i}g\alpha}\vec{\nabla}^2\alpha\Psi' + e^{-\mathrm{i}g\alpha}\vec{\nabla}\alpha\vec{\nabla}\Psi'\right] \\
&\ldots
\end{aligned} \qquad (11.5.145)$$

To solve this problem, we introduce the force field associated with the vector potential and the electrical potential and we will see that it guarantees the local invariance (therefore it is impossible to envisage an invariant phase change without the presence of a force field of this type). The local invariance imposes that the particle is no longer free (there are therefore no free charged particles!).

To do this, let us take again the Hamiltonian of the Pauli equation (see section Relativistic Quantum Physics page 4979):

$$\frac{\left(\vec{p}-q\vec{A}\right)^2}{2m_0} - \frac{\hbar q}{2m_0}\vec{\sigma}\circ\vec{B} + qU \qquad (11.5.146)$$

and let us neglect the interaction between the spin and the magnetic field such that the Hamiltonian becomes:

$$\frac{\left(\vec{p}-q\vec{A}\right)^2}{2m_0} + qU \qquad (11.5.147)$$

Therefore:

$$\frac{\left(-\mathrm{i}\hbar\vec{\nabla}-q\vec{A}\right)^2}{2m_0} + qU \qquad (11.5.148)$$

We thus get the following Schrödinger equation:

$$\left[\frac{\left(-\mathrm{i}\hbar\vec{\nabla}-q\vec{A}\right)^2}{2m_0}+qU\right]\Psi=\mathrm{i}\hbar\frac{\partial}{\partial t}\Psi \tag{11.5.149}$$

What in comparison to the free Schrödinger equation:

$$\frac{\left(-\mathrm{i}\hbar\vec{\nabla}\right)^2}{2m_0}\Psi=\mathrm{i}\hbar\frac{\partial}{\partial t}\Psi \tag{11.5.150}$$

involves the following correspondences (these are operators):

$$\begin{gathered}\vec{\nabla}\rightarrow\vec{D}\equiv\vec{\nabla}-\mathrm{i}\frac{q}{\hbar}\vec{A}\\ \frac{\partial}{\partial t}\rightarrow D^0\equiv\frac{\partial}{\partial t}+\mathrm{i}\frac{q}{\hbar}U\end{gathered} \tag{11.5.151}$$

Let us consider the following gauge transformation (see section Electrodynamics page 4115):

$$\begin{gathered}\vec{A}\rightarrow\vec{A}'=\vec{A}+\vec{\nabla}f\\ U\rightarrow U'=U-\frac{\partial f}{\partial t}\end{gathered} \tag{11.5.152}$$

where $f=f(\vec{r},t)$.

First, we immediately see that the operators are invariant. Indeed, for the first one $\vec{D}$:

$$\begin{aligned}&\left(\vec{\nabla}-\mathrm{i}\frac{q}{\hbar}\vec{A}'\right)e^{\mathrm{i}g\alpha}=\left(\vec{\nabla}-\mathrm{i}\frac{q}{\hbar}\left(\vec{A}+\vec{\nabla}f\right)\right)e^{\mathrm{i}g\alpha}=\left(\vec{\nabla}-\mathrm{i}\frac{q}{\hbar}\vec{A}+\mathrm{i}\frac{q}{\hbar}\vec{\nabla}f\right)e^{\mathrm{i}g\alpha}\\&=\mathrm{i}ge^{\mathrm{i}g\alpha}\left(\vec{\nabla}\alpha\right)+e^{\mathrm{i}g\alpha}\vec{\nabla}-\mathrm{i}\frac{q}{\hbar}\vec{A}e^{\mathrm{i}g\alpha}-\mathrm{i}\frac{q}{\hbar}\vec{\nabla}fe^{\mathrm{i}g\alpha}\\&=e^{\mathrm{i}g\alpha}\left(\mathrm{i}g\vec{\nabla}\alpha+\vec{\nabla}-\mathrm{i}\frac{q}{\hbar}\vec{A}-\mathrm{i}\frac{q}{\hbar}\vec{\nabla}f\right)\end{aligned} \tag{11.5.153}$$

However, if we put g as being $q/\hbar$ and f as α then we have:

$$\left(\vec{\nabla}-\mathrm{i}\frac{q}{\hbar}\vec{A}'\right)e^{\mathrm{i}\frac{q}{\hbar}\alpha}=e^{\mathrm{i}\frac{q}{\hbar}\alpha}\left(\mathrm{i}\cancel{\frac{q}{\hbar}\vec{\nabla}}f+\vec{\nabla}-\mathrm{i}\frac{q}{\hbar}\vec{A}-\mathrm{i}\cancel{\frac{q}{\hbar}\vec{\nabla}}f\right)=e^{\mathrm{i}\frac{q}{\hbar}\alpha}\left(\vec{\nabla}-\mathrm{i}\frac{q}{\hbar}\vec{A}\right) \tag{11.5.154}$$

Thus simply:

$$\left(\vec{\nabla}-\mathrm{i}\frac{q}{\hbar}\vec{A}'\right)e^{\mathrm{i}\frac{q}{\hbar}\alpha}=e^{\mathrm{i}\frac{q}{\hbar}\alpha}\left(\vec{\nabla}-\mathrm{i}\frac{q}{\hbar}\vec{A}\right) \tag{11.5.155}$$

Similarly, knowing now that f is α we have for the operator D^0:

$$\begin{aligned}\left(\frac{\partial}{\partial t}+\mathrm{i}\frac{q}{\hbar}U'\right)e^{\mathrm{i}\frac{q}{\hbar}\alpha}&=\left(\frac{\partial}{\partial t}+\mathrm{i}\frac{q}{\hbar}\left(U-\frac{\partial\alpha}{\partial t}\right)\right)e^{\mathrm{i}\frac{q}{\hbar}\alpha}\\&=\mathrm{i}\frac{q}{\hbar}e^{\mathrm{i}\frac{q}{\hbar}\alpha}\frac{\partial\alpha}{\partial t}+e^{\mathrm{i}\frac{q}{\hbar}\alpha}\frac{\partial}{\partial t}+\mathrm{i}\frac{q}{\hbar}Ue^{\mathrm{i}\frac{q}{\hbar}\alpha}-\mathrm{i}\frac{q}{\hbar}\frac{\partial\alpha}{\partial t}e^{\mathrm{i}\frac{q}{\hbar}\alpha}\\&=e^{\mathrm{i}\frac{q}{\hbar}\alpha}\left(\frac{\partial}{\partial t}+\mathrm{i}\frac{q}{\hbar}U\right)\end{aligned} \tag{11.5.156}$$

Therefore we have:

$$\begin{gathered}\left(\vec{\nabla}-\mathrm{i}\frac{q}{\hbar}\vec{A}'\right)e^{\mathrm{i}\frac{q}{\hbar}\alpha}=e^{\mathrm{i}\frac{q}{\hbar}\alpha}\left(\vec{\nabla}-\mathrm{i}\frac{q}{\hbar}\vec{A}\right)\\ \left(\frac{\partial}{\partial t}+\mathrm{i}\frac{q}{\hbar}U'\right)e^{\mathrm{i}\frac{q}{\hbar}\alpha}=e^{\mathrm{i}\frac{q}{\hbar}\alpha}\left(\frac{\partial}{\partial t}+\mathrm{i}\frac{q}{\hbar}U\right)\end{gathered}\tag{11.5.157}$$

Thus:

$$\begin{gathered}\vec{D}'e^{\mathrm{i}\frac{q}{\hbar}\alpha}=e^{\mathrm{i}\frac{q}{\hbar}\alpha}\vec{D}\\ \left(D^{0\prime}\right)e^{\mathrm{i}\frac{q}{\hbar}\alpha}=e^{\mathrm{i}\frac{q}{\hbar}\alpha}\left(D^{0}\right)\end{gathered}\tag{11.5.158}$$

The relation:

$$\frac{\left(-\mathrm{i}\hbar\vec{\nabla}\right)^2}{2m}\Psi=\mathrm{i}\hbar\frac{\partial}{\partial t}\Psi\tag{11.5.159}$$

Becomes then with the new correspondences:

$$\frac{\left(-\mathrm{i}\hbar\vec{D}\right)^2}{2m}\Psi=\mathrm{i}\hbar D^0\Psi\tag{11.5.160}$$

and with the previous developments, we have:

$$e^{\mathrm{i}g\alpha}\frac{\left(-\mathrm{i}\hbar\vec{D}\right)^2}{2m}\Psi=\frac{\left(-\mathrm{i}\hbar\vec{D}'\right)^2}{2m}e^{\mathrm{i}g\alpha}\Psi\tag{11.5.161}$$

Or written differently:

$$e^{\mathrm{i}g\alpha}\mathrm{i}\hbar D^0\Psi=\mathrm{i}\hbar D^{0\prime}e^{\mathrm{i}g\alpha}\Psi\tag{11.5.162}$$

Thus:

$$\frac{\left(-\mathrm{i}\hbar\vec{D}'\right)^2}{2m}e^{\mathrm{i}g\alpha}\Psi=\mathrm{i}\hbar D^{0\prime}e^{\mathrm{i}g\alpha}\Psi\tag{11.5.163}$$

Which gives after simplification:

$$\frac{\left(-\mathrm{i}\hbar\vec{D}'\right)^2}{2m}\Psi=\mathrm{i}\hbar D^{0\prime}\Psi\tag{11.5.164}$$

Thus, by requesting the local gauge invariance, we have make appear an interaction ... and we know well which interaction it is!

The Schrödinger equation of a particle moving in an electromagnetic field is therefore invariant under the local phase transformation. The phase of a wave function is indeed a new local variable in the Weyl sense and the electromagnetic potential can be interpreted, according to Weyl, as a connection binding the phases to different points.

We conclude that the electromagnetic field is a consequence of the local gauge invariance based on the group $\mathrm{U}(1)$, a group of one-dimensional unitary matrices (see section Set Algebra page

1424). The interest is to build Gauge theories on more complicated (non-abelian) groups: these theories are named "Yang-Mills theories".

So we can make an analogy: If a global gauge is like to turn all the hands of a watch of a constant angle around the circle, the local gauge is similar as if we make this angle vary continuously from one point to another of space-time. One might think that such an upheaval of the hands of a watch generate tensions and compressions on the springs, sign of a modification of the electromagnetic field at each point!? But not at all: all the springs remain at rest, as if the electromagnetic field automatically adjusted to compensate for these local deformations:

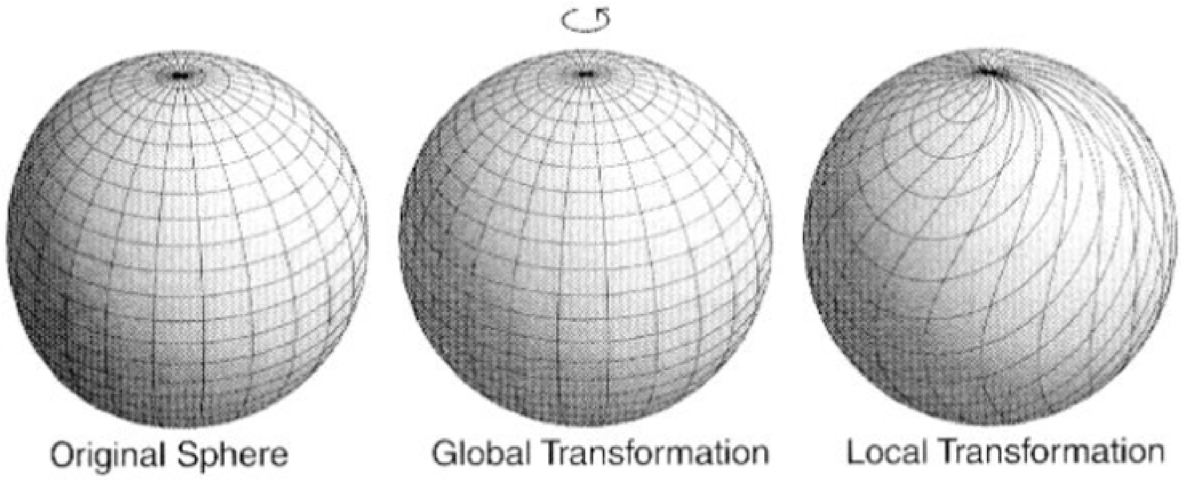


Figure 11.149: Global and local gauge illustration concept (source: ?)

Now let us go a little further but without going too deep (as this textbook is only about Elementary Applied Mathematics)... We showed earlier above that the Lagrangian of the free Dirac equation was:

$$L=\bar{\Psi}\left(\mathrm{i}\gamma^{\mu}\partial_{\mu}-m_0\right)\Psi \qquad (11.5.165)$$

Now, since this Lagrangian does not reveal the electromagnetic field, we strongly suspect that it does not carry in it a local Gauge invariance...

Therefore the operator ∂_{μ} has somewhere a lack... But, the equivalent of the divergence operator $\vec{\nabla}$ in the free Schrödinger equation is here the covariant derivative ∂_{μ}. So in the same way that we have associated the following operators to guarantee the local Gauge invariance of the free Schrödinger equation:

$$\begin{gathered}\vec{\nabla}\rightarrow\vec{D}\equiv\vec{\nabla}-\mathrm{i}\frac{q}{\hbar}\vec{A}\\ \frac{\partial}{\partial t}\rightarrow D^0\equiv\frac{\partial}{\partial t}+\mathrm{i}\frac{q}{\hbar}U\end{gathered} \qquad (11.5.166)$$

It is tempting to combine the whole into a new operator called the "covariant derivative":

$$D_{\mu}\equiv\partial_{\mu}-\mathrm{i}\frac{q}{\hbar}A_{\mu} \qquad (11.5.167)$$

with:

$$D_0\equiv\partial_0+\mathrm{i}\frac{q}{\hbar}A_0=\frac{\partial}{c\partial t}+\mathrm{i}\frac{q}{\hbar}(-U) \qquad (11.5.168)$$

The Lagrangian of the free Dirac equation would then be written:

$$L=\bar{\Psi}\left(\mathrm{i}\gamma^{\mu}D_{\mu}-m_0\right)\Psi \tag{11.5.169}$$

or in Feynman notation:

$$L=\bar{\Psi}\left(\mathrm{i}\not{D}-m_0\right)\Psi \tag{11.5.170}$$

Therefore:

$$\begin{aligned}L&=\bar{\Psi}\left(\mathrm{i}\gamma^{\mu}\left(\partial_{\mu}-\mathrm{i}\frac{q}{\hbar}A_{\mu}\right)-m_0\right)\Psi=\bar{\Psi}\left(\mathrm{i}\gamma^{\mu}\partial_{\mu}-m_0\right)\Psi-\bar{\Psi}\mathrm{i}\gamma^{\mu}\mathrm{i}\frac{q}{\hbar}A_{\mu}\Psi\\&=\bar{\Psi}\left(\mathrm{i}\gamma^{\mu}\partial_{\mu}-m_0\right)\Psi+\bar{\Psi}\gamma^{\mu}\Psi\frac{q}{\hbar}A_{\mu}=\bar{\Psi}\left(\mathrm{i}\gamma^{\mu}\partial_{\mu}-m_0\right)\Psi-j^{\mu}A_{\mu}\end{aligned} \tag{11.5.171}$$

with:

$$\boxed{j^{\mu}=-\frac{q}{\hbar}\bar{\Psi}\gamma^{\mu}\Psi} \tag{11.5.172}$$

It only remains to add the term of the field to obtain the total Lagrangian of the Dirac equation (it would have been relatively hard to find it in another way ...):

$$\boxed{L_D=\bar{\Psi}\left(\mathrm{i}\gamma^{\mu}\partial_{\mu}-m_0\right)\Psi-j^{\mu}A_{\mu}+\frac{1}{4}\mu_0F_{\mu\nu}F^{\mu\nu}} \tag{11.5.173}$$

which corresponds to the "Dirac-Maxwell equations" and which is the "Quantum electrodynamic Lagrangian" (in natural units) where on the left we have the term of the fermions and on the right the interaction part of the bosons of zero mass (photons).

Thus the fact of having added to the free Lagrangian a condition of invariance by local transformations, has led us to a theory with interaction that we can write with more rigour and in developed form:

$$\mathcal{L}_{\text{QED}}=\bar{\Psi}\left(\mathrm{i}\gamma^{\mu}\partial_{\mu}-m_0\right)\Psi+\frac{q}{\hbar}\bar{\Psi}\gamma^{\mu}\Psi A_{\mu}+\frac{1}{4}\mu_0F_{\mu\nu}F^{\mu\nu} \tag{11.5.174}$$

Or in natural units and with the charge of the electron:

$$\mathcal{L}_{\text{QED}}=\bar{\Psi}\left(\mathrm{i}\gamma^{\mu}\partial_{\mu}-m_0\right)\Psi+q\bar{\Psi}\gamma^{\mu}\Psi A_{\mu}+\frac{1}{4}F_{\mu\nu}F^{\mu\nu} \tag{11.5.175}$$

However, quantum electrodynamics was lacking in the 11940s (holocene calendar) to describe a large number of particles that were detected by accelerators. Certainly, in a way, it has been extended to describe new particles. But many of them seemed to enjoy properties that quantum electrodynamics could not account for.

In fact, the reason is simple ... it is a theory in which no exact solution is known so far, a situation that persists until the day these lines are written (12008 according to holocene calendar). The only available calculation method is named "perturbative development". The idea is essentially the same as that of limited development practised in the field of differential calculus. In this case, if we do not know how to calculate the value of a function, we decompose it into a sequence of polynomials and the approximation is refined as we take into account terms of higher order. Such a series development begins with a zero order term, which is just the value of the unknown function at a certain point where we know how to calculate this function.

In the case of the perturbative development of quantum electrodynamics, the zero order term represents pure propagation without interaction (the intensity of the interaction between the electron and the magnetic field is set to zero). In this approximation, quantum electrodynamics is a theory of free particles and it is exactly calculable. We have electrons, positrons and photons but they cross without influencing each other. The next term in the series development, that of the first order, is also exactly calculable. In this approximation, the theory seems to reflect the real world fairly accurately. Very interesting physical phenomena appear in this first-order approximation of the real theory of photon-electron interaction and the theory agrees well with the experimental results.

Unfortunately, it soon became apparent that the calculation of second-order terms and higher terms seemed meaningless as they yielded infinite values... today there still exist only approximate methods of resolution and that are not completely satisfactory. Since then physicists have been obliged to look for another technique of approximation based on a renormalization of the equations... and the results are extraordinarily good (to the 11th decimal!) but in the facts it looks like a bit like do-it-yourself physics...

Remark

Quantum electrodynamics also suffer from a perceived lack of aesthetic appeal. The theory give rise to infinities that had to be removed by provisional methods introduced for no reason other than to give useful results. It is a pragmatic approach that Dirac didn't like at all: « *Recent work by Lamb, Schwinger, Feynman and others has been very successful... but the resulting theory is an ugly and incomplete one and cannot be considered as a satisfactory solution of the problem of the electron.* » When asked for his opinion about the recent developments in quantum electrodynamics, Dirac said, « *I might have thought that the new ideas were correct if they had not been so ugly.* »

~50 %

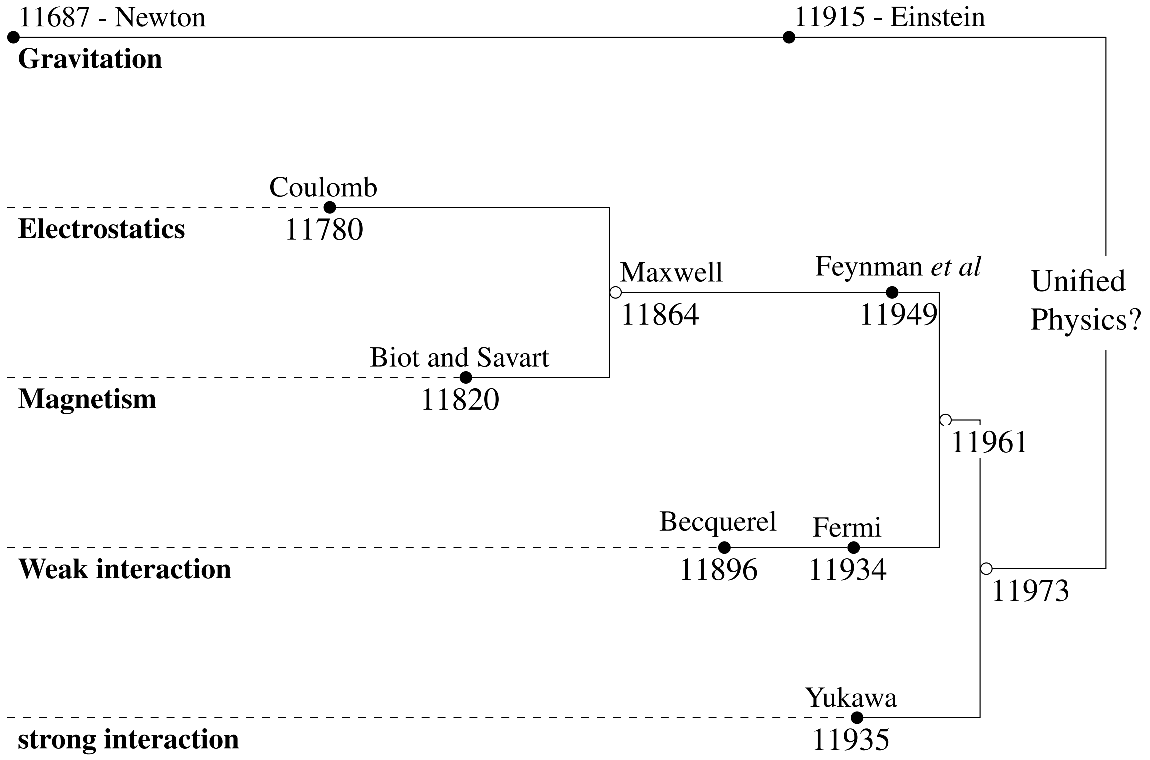
★★★☆☆

47 votes, 69.36%

11.6. Elementary Particle Physics

We have already mentioned in the section of Nuclear Physics that we experiment that the radioactive nuclei does not emit neutrons or protons (i.e. its single constituents). But we can ask ourselves: How do it synthesize an alpha particle, or transform a proton into a neutron or vice versa? To answer these questions, let us examine the forces in presence!

Before the discovery of radioactivity, physicists had identified two fundamental forces: the gravitational force[42] and the electromagnetic force. The discovery of radioactivity and studies on the atomic nucleus led physicists to introduce not one but two new fundamental forces! In the figure below, the years are given according to the holocene calendar:



11687 - Newton
11915 - Einstein
Gravitation
Coulomb
Electrostatics
11780
Maxwell
11864
Feynman *et al*
11949
Unified Physics?
Biot and Savart
Magnetism
11820
11961
Becquerel
Fermi
Weak interaction
11896
11934
11973
Yukawa
strong interaction
11935

Figure 11.150: **Interactions discoveries (source: `http://femto-physique.fr`, author: Jimmy Roussel)**

Even before knowing the exact composition of the nuclei, to explain the existence of these tiny systems and wearing sometimes strong positive charges, physicists had foreseen the need for a powerful cohesive force capable of dominating the electrostatic repulsion exerted between these charges (remember that we saw in the section of Classical Mechanics that the gravitational force between two bodies of equivalent masses to those of particles is completely negligible). Since the nucleus is small, this "nuclear force" had to have influence at very short distances. When J.

---

[42] We will see it during our study of General Relativity that the gravitational force is fictional and we should us the expression "gravitational interaction" instead.

Chadwick discovered the neutron, it was shown experimentally that attractive force is also well practised between two neutrons, two protons and between a neutron and a proton. In 11935 (holocene calendar), Hideki Yukawa elaborated a theory about the nuclear force whose outlines seems to be still accepted, but must still be improved following some issues that have been identified (see section Quantum Field Theory page 5158).

However, as we already know, this nuclear force does not explain the transformation of a proton into a neutron, which takes place in the beta radioactivity. It was necessary to introduce a fourth fundamental force, of lower intensity, named for this reason "weak interaction", the nuclear force becomes ipso facto the "strong interaction".

Thus, in principle, radioactivity involves the four fundamental forces of nature: gravity and the electromagnetic force, as alpha and beta particles have mass and charge, and the two nuclear forces, strong and weak (in fact, the gravitation, of lesser intensity than the other three is often neglected at sub-atomic scales).

We partially addressed in the section of Quantum Field Theory (see page 5158) the fundamental interactions and their interactions vectors. Let's do some reminders:

1. We have proved in the section of Special Relativity (see page 5627) that there are two types of particles: those with a mass and will never go at the speed of light (because it will required an infinite energy to get them at this velocity) and those which have zero mass and which are therefore necessarily at the speed of light.

2. We proved in the section of Quantum Field Theory that more a particle has energy, that more (following the Heisenberg's uncertainty principles) it may have long virtual life time and travel considerable distances. We distinguish therefore the interaction that have infinite range of interaction whose particles have no mass and a the interactions that have finite range whose interaction particles have a mass.

Before starting with some arduous calculations, it is desirable first to acquire a vocabulary that is common use among theoretical physicists.

The easiest concept to be addressed in the field of elementary particle physics is the comparison of the four elemental forces via their respective coupling constant (that is something that physicists like to do...).

Remark

Particle physics research has significantly contributed to civil applications across various fields, including medical imaging and treatment with technologies like MRI and proton therapy, radiation detectors for security and environmental monitoring, and material science advancements through synchrotron light sources. It has also revolutionized global communication with the invention of the World Wide Web at CERN and supported innovations in renewable energy and environmental tools. These breakthroughs showcase how fundamental research drives practical innovations that benefit society.

Before we tackle with the various topics we have seen that some people are misleaded by the treatment of certain particles in modern physics as being "pointlike" (i.e. dimensionless). So first thing... "pointlike" is a misleading terminology (like "Big Bang" and many others in physics!) - it just means a quantum particle not made of more fundamental quantum particles. Second, treating particles like pointlike is similar as in astronomy where we often treat planets as pointlike because the scales involved in a given case authorize us such an acceptable approximation (even really big objects such as stars are pointlike particles when viewed from far enough away).

11.6.1 Coupling Constants

We will try here to rank the four forces according to their intensity through the use of "coupling constants".

> Caution!!!! All couplings in Quantum Field Theory are running (with scale), and are not constants (like the fine structure constant or the Hubble constant in Cosmology...!), but are named such for historical reasons. Even the electromagnetic coupling runs, and its "constant value" is taken to be the infrared (long distance) value computed and fit. The strong gauge coupling α_s varies by a factor of two between 10 and $1,000$ [GeV].

To do this, we need to calculate the four interactions for two same particles, for example two protons at identical distances, so of nuclear type and compare them to a common value of the units so that their ratio provides a dimensionless number.

This common value will be chosen as the product:

$$\frac{h}{2\pi}c=\hbar c\cong 3.16\cdot 10^{-26}\,[\text{N}\cdot\text{m}^2] \tag{11.6.1}$$

Thus we find:

1. For the gravitational "force" (see section Astronomy page 5255) where:

   $$Fr^2=GmM\,[\text{N}\cdot\text{m}^2] \tag{11.6.2}$$

   with the proton mass such that $m=M\cong 1.67\cdot 10^{-27}$ [kg], the coupling constant of the gravitational force is therefore by definition:

   $$\alpha_G=\frac{GMm}{\hbar c}=\frac{Gm_p^2}{\hbar c}\cong 6\cdot 10^{-39} \tag{11.6.3}$$

2. For the electrical force (see section Electrostatics page 3968) where:

   $$Fr^2=\frac{1}{4\pi\varepsilon_0}qQ\,[\text{N}\cdot\text{m}^2] \tag{11.6.4}$$

   with the proton charge such that $q=Q\cong 1.602\cdot 10^{-19}$ [C], the coupling constant of the electric force is therefore by definition:

   $$\alpha=\frac{\dfrac{1}{4\pi\varepsilon_0}qQ}{\hbar c}\cong\frac{1}{137} \tag{11.6.5}$$

Remark

Here we fall back on the "fine-structure constant" that we had already seen in the section of Corpuscular Quantum Physics. We therefore better understand the start choice for the relative comparison of interactions.

3. For the strong nuclear force, where F represents the "strong nuclear charge", the strong coupling constant has for value (caution! the value depends on the chosen theoretical model!):

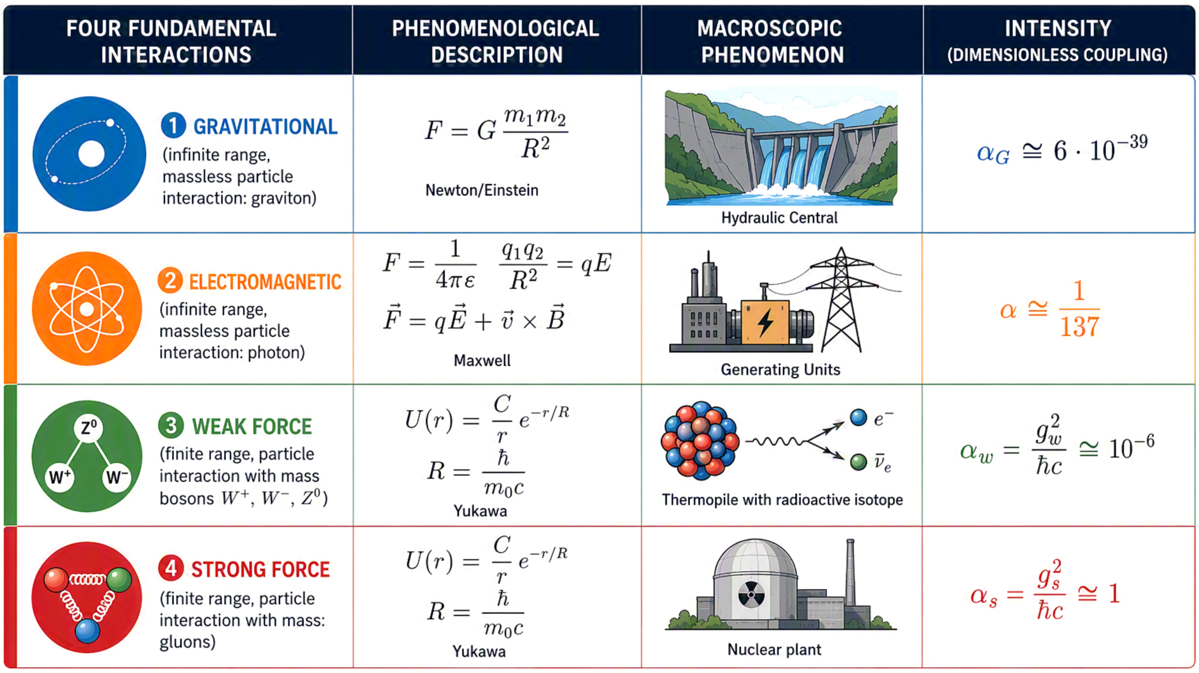
$$\alpha_s=\frac{F\cdot F}{\hbar c}=\frac{g_s^2}{\hbar c}\cong 1 \tag{11.6.6}$$

hence its name.

4. For the weak nuclear force responsible for the decay of particles, f represents the "weak nuclear charge", and its weak coupling constant value is then (caution! the value depends on the chosen theoretical model!):

$$\alpha_w=\frac{f\cdot f}{\hbar c}=\frac{g_w^2}{\hbar c}\cong 10^{-6} \tag{11.6.7}$$

So all this can be summarized in the following phenomenological table:



FOUR FUNDAMENTAL INTERACTIONS	PHENOMENOLOGICAL DESCRIPTION	MACROSCOPIC PHENOMENON	INTENSITY (DIMENSIONLESS COUPLING)
① **GRAVITATIONAL** (infinite range, massless particle interaction: graviton)	$F=G\dfrac{m_1m_2}{R^2}$ Newton/Einstein	Hydraulic Central	$\alpha_G\cong 6\cdot 10^{-39}$
② **ELECTROMAGNETIC** (infinite range, massless particle interaction: photon)	$F=\dfrac{1}{4\pi\varepsilon}\dfrac{q_1q_2}{R^2}=qE$ $\vec{F}=q\vec{E}+\vec{v}\times\vec{B}$ Maxwell	Generating Units	$\alpha\cong\dfrac{1}{137}$
③ **WEAK FORCE** (finite range, particle interaction with mass bosons W^+, W^-, Z^0)	$U(r)=\dfrac{C}{r}e^{-r/R}$ $R=\dfrac{\hbar}{m_0c}$ Yukawa	e^- $\bar{\nu}_e$ Thermopile with radioactive isotope	$\alpha_w=\dfrac{g_w^2}{\hbar c}\cong 10^{-6}$
④ **STRONG FORCE** (finite range, particle interaction with mass: gluons)	$U(r)=\dfrac{C}{r}e^{-r/R}$ $R=\dfrac{\hbar}{m_0c}$ Yukawa	Nuclear plant	$\alpha_s=\dfrac{g_s^2}{\hbar c}\cong 1$

Figure 11.151: **Phenomenological table of the 4 fundamental interactions**

Caution!!! There is no principle of physics yet (in the beginning of the 121st century according to holocene calendar) that says physicals laws or constants have to be the same everywhere and always. Indeed, the reader must remember that the "laws" of physics are **descriptive**, not **imperative** and therefore they are some observed regularity in what we can observe experimentally with sufficiently wide occurrence or relevance! Asking whether the laws can change is no more and no less than asking whether regularities which have been observed up till now might not be quite as universal as was thought. The answer to that is: YES! This could drive the reader a bit mad. Indeed, how can we possibly do Science if the rules keep changing? But if you are a professional active scientist, you have to face up to that possibility - and you have a responsibility to check out if there is evidence beyond reasonable doubt about it! We will repeat this warning multiple times in this textbook as its seems that this is something necessary!

Or also with the following diagram (more interesting) where we can see, taking into account the results we have found during our study of the massic fields and non-massic fields in the context of the Yukawa model (see section Quantum Field Theory page 5164):

1. In ordinate the intensity of the forces as previously calculated according to the distance as given by the Yukawa model of massic fields (weak and strong interactions) and non-massic fields (electromagnetic and gravitational interactions).
2. The representative interaction patterns (Feynman diagrams) in accordance with the results obtained and particles already mentioned in the section of Quantum Field Theory.

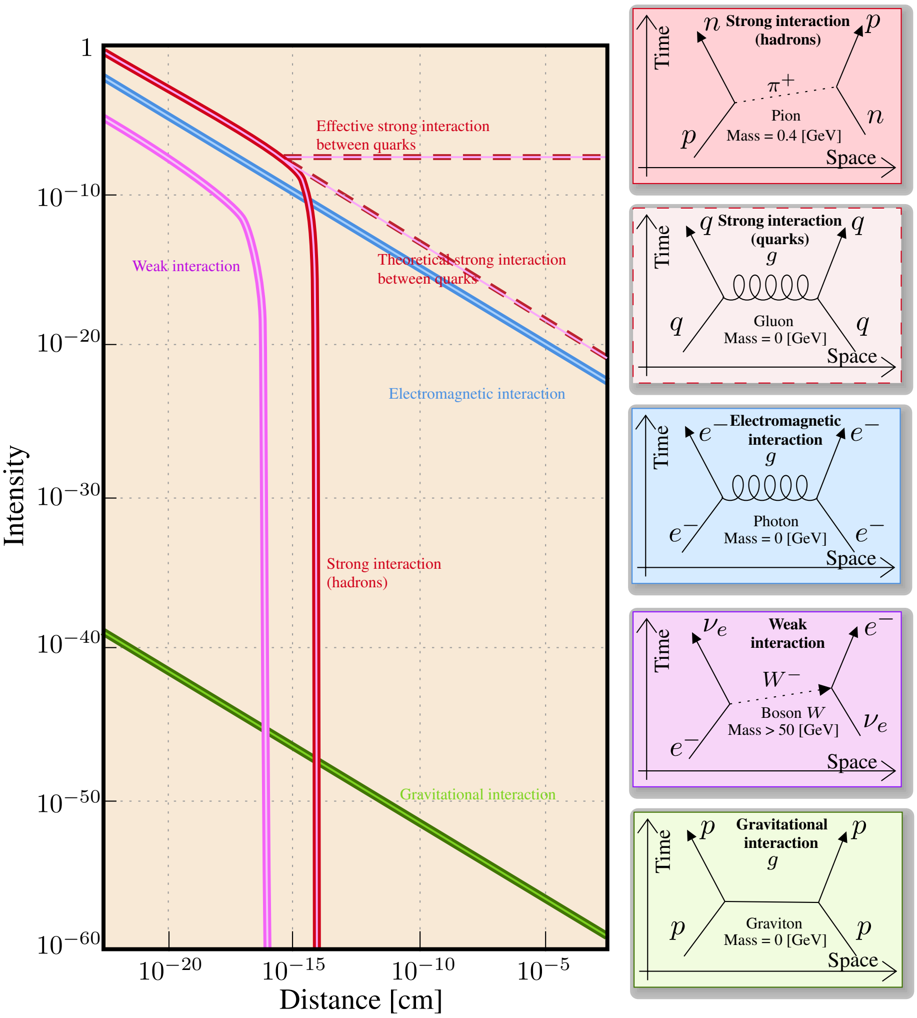


Figure 11.152: **Feynman diagrams and interaction distance of the fundamental forces**

In most physics textbooks you will read that gravity is the weakest of all forces, many orders of magnitude weaker than electromagnetism. This is wrong! Recall that the gravitational force is fictional, like the centrifugal (fictional) force. No one compares the value of centrifugal (fictional) force with other forces. Its value depends on the circumstances! So it is with gravity. The gravitational force depends on the masses and charges of the particles. Most textbooks just take a special case (an electron and a proton most of the time) and make from it a stupid generalization. It is not the relative strength of the gravitational and electrical forces in all cases. In fact, there is no universal way we can describe the strength of the gravitational force!

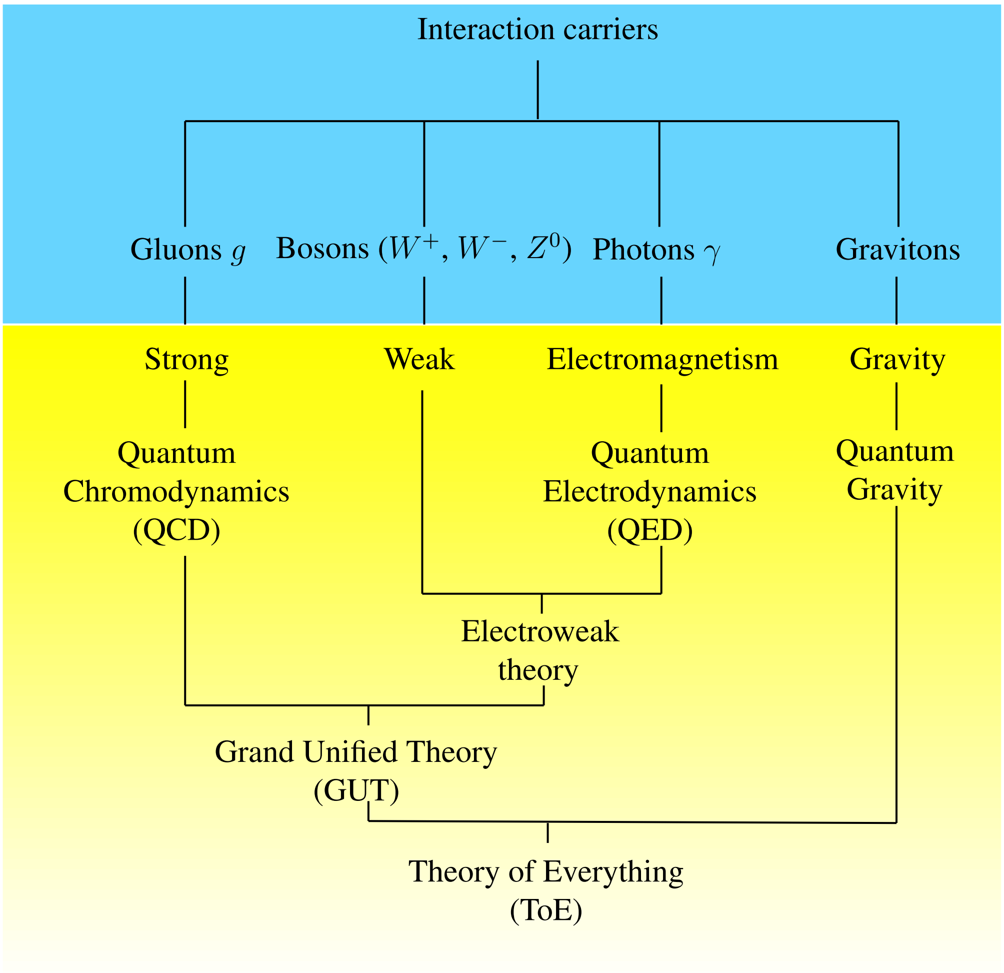

Interaction carriers
Gluons g
Bosons (W^+, W^-, Z^0)
Photons γ
Gravitons
Strong
Weak
Electromagnetism
Gravity
Quantum Chromodynamics (QCD)
Quantum Electrodynamics (QED)
Quantum Gravity
Electroweak theory
Grand Unified Theory (GUT)
Theory of Everything (ToE)

Figure 11.153: **Force carriers subparticles**

Below we can see the 12 fundamental fermions (we can count the 12 corresponding anti-particles if we want as indicated in the figure...) and 5 fundamental bosons. Please note that the masses of certain particles are subject to periodic re-evaluation by the scientific community. The values currently reflected in this graphic are as of 12008 (holocene calendar) and may have been adjusted since. For the latest consensus, please visit the Particle Data Group website `https://pdg.lbl.gov`.

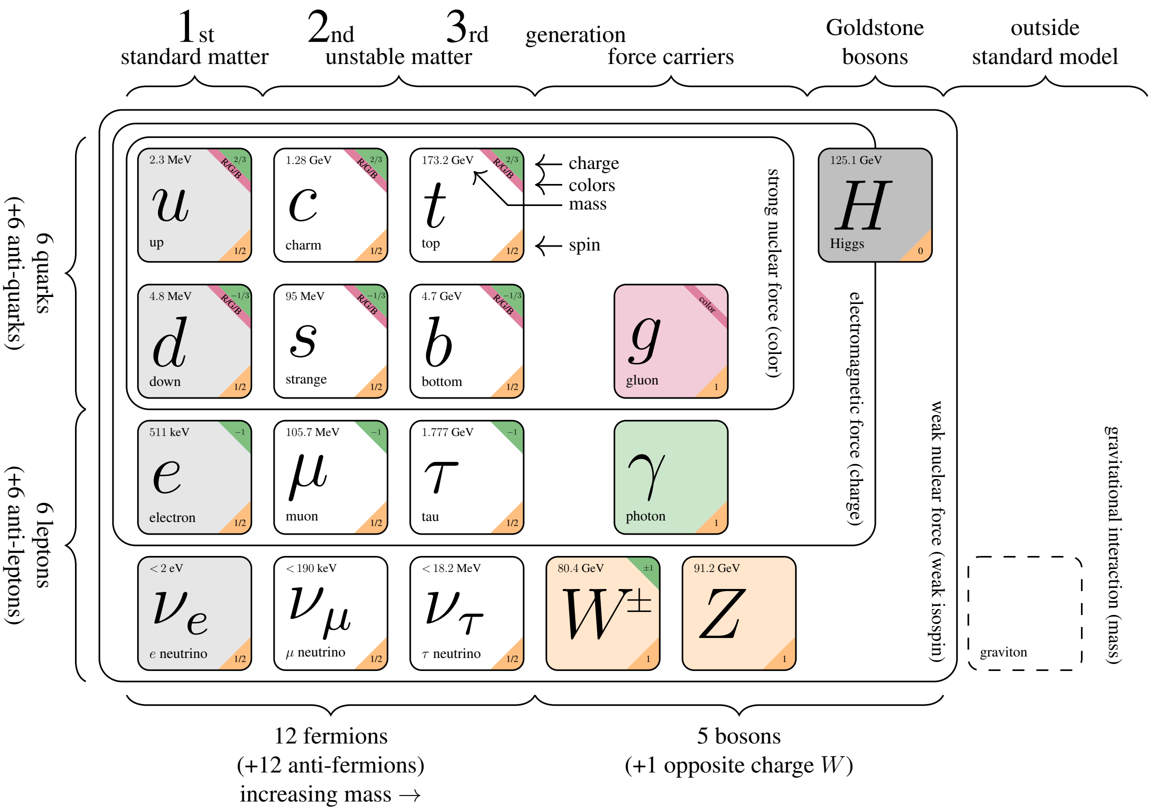

1st standard matter
2nd 3rd unstable matter
generation force carriers
Goldstone bosons
outside standard model
2.3 MeV 2/3 R/G/B u up 1/2
1.28 GeV 2/3 R/G/B c charm 1/2
173.2 GeV 2/3 R/G/B t top 1/2
charge
colors
mass
spin
125.1 GeV H Higgs 0
4.8 MeV −1/3 R/G/B d down 1/2
95 MeV −1/3 R/G/B s strange 1/2
4.7 GeV −1/3 R/G/B b bottom 1/2
color g gluon 1
511 keV −1 e electron 1/2
105.7 MeV −1 μ muon 1/2
1.777 GeV −1 τ tau 1/2
γ photon 1
< 2 eV ν_e e neutrino 1/2
< 190 keV ν_μ μ neutrino 1/2
< 18.2 MeV ν_τ τ neutrino 1/2
80.4 GeV ±1 W± 1
91.2 GeV Z 1
graviton
strong nuclear force (color)
electromagnetic force (charge)
weak nuclear force (weak isospin)
gravitational interaction (mass)
6 quarks (+6 anti-quarks)
6 leptons (+6 anti-leptons)
12 fermions (+12 anti-fermions) increasing mass →
5 bosons (+1 opposite charge W)

Figure 11.154: Subparticles Standard Model (source: Wikipedia)

Or with different informations:

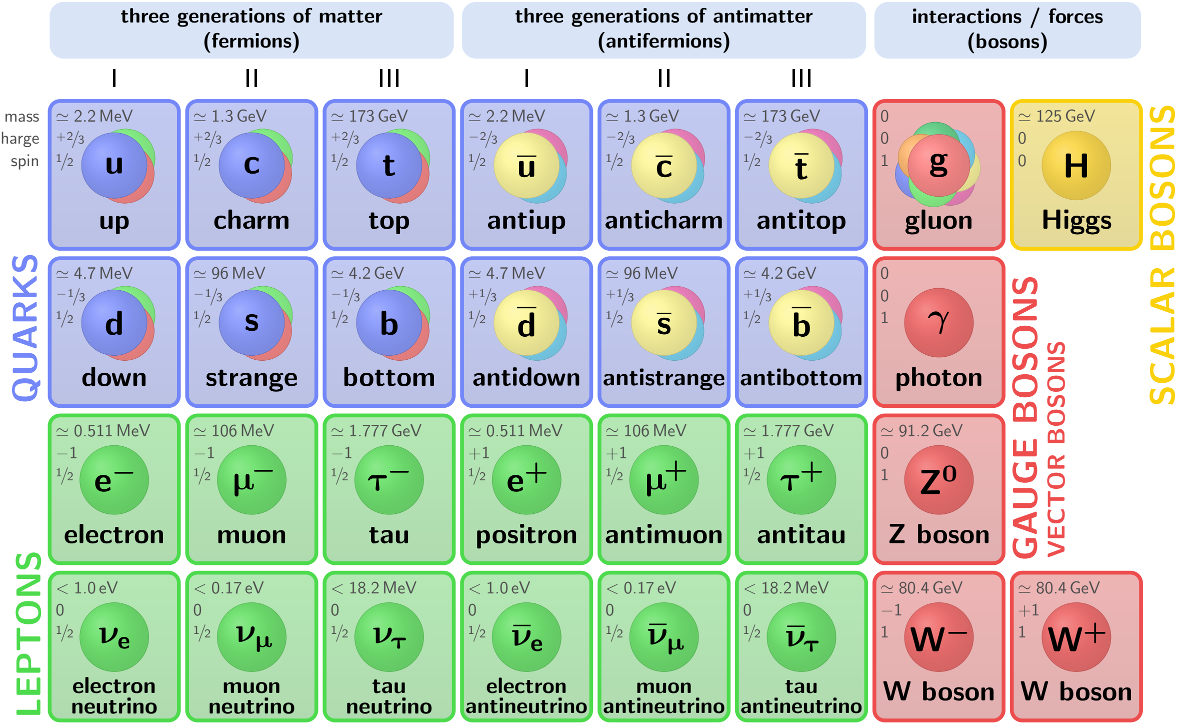

three generations of matter (fermions)
three generations of antimatter (antifermions)
interactions / forces (bosons)
I II III I II III
mass
charge
spin
≃ 2.2 MeV +2/3 1/2 u up
≃ 1.3 GeV +2/3 1/2 c charm
≃ 173 GeV +2/3 1/2 t top
≃ 2.2 MeV −2/3 1/2 ū antiup
≃ 1.3 GeV −2/3 1/2 c̄ anticharm
≃ 173 GeV −2/3 1/2 t̄ antitop
0 0 1 g gluon
≃ 125 GeV 0 0 H Higgs
≃ 4.7 MeV −1/3 1/2 d down
≃ 96 MeV −1/3 1/2 s strange
≃ 4.2 GeV −1/3 1/2 b bottom
≃ 4.7 MeV +1/3 1/2 d̄ antidown
≃ 96 MeV +1/3 1/2 s̄ antistrange
≃ 4.2 GeV +1/3 1/2 b̄ antibottom
0 0 1 γ photon
≃ 0.511 MeV −1 1/2 e⁻ electron
≃ 106 MeV −1 1/2 μ⁻ muon
≃ 1.777 GeV −1 1/2 τ⁻ tau
≃ 0.511 MeV +1 1/2 e⁺ positron
≃ 106 MeV +1 1/2 μ⁺ antimuon
≃ 1.777 GeV +1 1/2 τ⁺ antitau
≃ 91.2 GeV 0 1 Z⁰ Z boson
< 1.0 eV 0 1/2 ν_e electron neutrino
< 0.17 eV 0 1/2 ν_μ muon neutrino
< 18.2 MeV 0 1/2 ν_τ tau neutrino
< 1.0 eV 0 1/2 ν̄_e electron antineutrino
< 0.17 eV 0 1/2 ν̄_μ muon antineutrino
< 18.2 MeV 0 1/2 ν̄_τ tau antineutrino
≃ 80.4 GeV −1 1 W⁻ W boson
≃ 80.4 GeV +1 1 W⁺ W boson
QUARKS
LEPTONS
GAUGE BOSONS
VECTOR BOSONS
SCALAR BOSONS

Figure 11.155: Splitting particles/antiparticles by electric charge (quarks and gluons as shown as different color states) (credit: Izaak Neutelings)

As already mentioned in the section of Nuclear Physics, not only nucleus of atoms decays! But also most particles and even some fundamental particles (yes! but they are not made of the stuff they decay of as experience show us but with the τ fundamental particle that has at least 30 different decay chains[43]). However some particles don't decay (as far as we know) like: the photon, the electron (anti-electron), the lightest neutrino (and its anti-particle) and (presumed) the graviton. The reader can refer to the *Review of Particle Physics: particle data groups* (see [526]) that contains more than a hundred pages of fundamental particles properties with for each a huge list of disintegrations paths.

Type	Name	Symbol	Energy [MeV]	Mean lifetime
Lepton	Electron / Positron	e^-/e^+	0.511	$>6.6\cdot 10^{28}$ years
	Muon / Antimuon	μ^-/μ^+	105.7	$2.2\cdot 10^{-6}$ seconds
	Tau lepton / Antitau	τ^-/τ^+	$1,777$	$2.9\cdot 10^{-13}$ seconds
Meson	Neutral Pion	π^0	135	$8.4\cdot 10^{-17}$ seconds
	Charged Pion	π^+/pi^-	139.6	$2.6\cdot 10^{-8}$ seconds
Baryon	Proton / Antiproton	p^+/p^-	938.2	$>10^{29}$ years
	Neutron / Antineutron	$n/\bar{n}$	939.6	885.7 seconds
Boson	W boson	W^+/W^-	$80,377$	10^{-25} seconds
	Boson	W^+/W^-	$80,377$	10^{-25} seconds
	Photon	γ	$-$	$+\infty$

Table 11.23: Some "fundamental" particles mean lifetime and mass

Remark

The electron is a charge lepton and leptons can decay to other leptons, but in standard conditions not the electron since the other two charge leptons are heavier. However, if we accelerate the electron in high velocities (energies actually not reachable with accelerators like the LHC in the early 121st century according to holocene calendar but energies level that were existing during the first instant of the Big Bang) then we now talk about its electroweak field. Since W bosons have zero average rest mass under an electroweak field it is possible for an electron to decay to a W^- and an electron neutron ν_e.

It should be specified for the general knowledge of the reader that these four forces are described by respectively four theories:

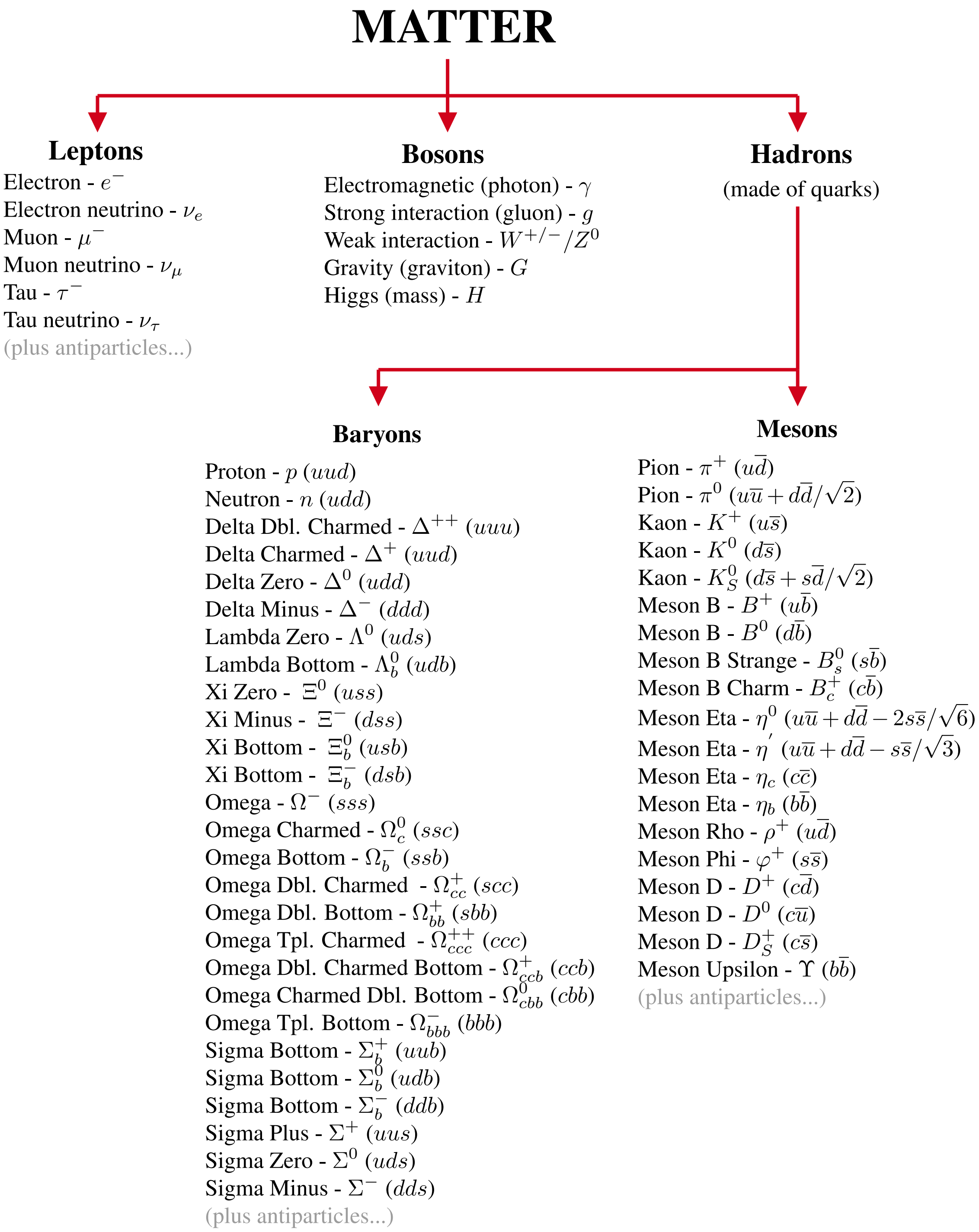
1. General Relativity (includes Classical Mechanics) for gravitation.
2. Quantum Electrodynamics (includes Electrodynamics) for the electromagnetic force.
3. Electroweak Theory (which includes Quantum Electrodynamics) for the weak interaction.
4. Quantum Chromodynamics for the strong interaction.

The last three are grouped in the "Standard Model".

Here is another figure summarizing the state of the art knowledge of particles in the early 121st (holocene calendar):

---

[43] Like $\tau\to e^-\bar{\nu}_e\nu_\tau$ (17.83% of the time) or $\tau\to\mu^-\bar{\nu}_\mu\nu_\tau$ (17.41% of the time) or $\tau\to\pi^-\nu_\tau$ (10.83% of the time) or $\tau\to\pi^-\pi^0\nu_\tau$ (25.52% of the time), etc.



MATTER

Leptons

Electron - e^-
Electron neutrino - ν_e
Muon - μ^-
Muon neutrino - ν_μ
Tau - τ^-
Tau neutrino - ν_τ
(plus antiparticles...)

Bosons

Electromagnetic (photon) - γ
Strong interaction (gluon) - g
Weak interaction - $W^{+/-}/Z^0$
Gravity (graviton) - G
Higgs (mass) - H

Hadrons

(made of quarks)

Baryons

Proton - p (uud)
Neutron - n (udd)
Delta Dbl. Charmed - Δ^{++} (uuu)
Delta Charmed - Δ^+ (uud)
Delta Zero - Δ^0 (udd)
Delta Minus - Δ^- (ddd)
Lambda Zero - Λ^0 (uds)
Lambda Bottom - Λ_b^0 (udb)
Xi Zero - Ξ^0 (uss)
Xi Minus - Ξ^- (dss)
Xi Bottom - Ξ_b^0 (usb)
Xi Bottom - Ξ_b^- (dsb)
Omega - Ω^- (sss)
Omega Charmed - Ω_c^0 (ssc)
Omega Bottom - Ω_b^- (ssb)
Omega Dbl. Charmed - Ω_{cc}^+ (scc)
Omega Dbl. Bottom - Ω_{bb}^+ (sbb)
Omega Tpl. Charmed - Ω_{ccc}^{++} (ccc)
Omega Dbl. Charmed Bottom - Ω_{ccb}^+ (ccb)
Omega Charmed Dbl. Bottom - Ω_{cbb}^0 (cbb)
Omega Tpl. Bottom - Ω_{bbb}^- (bbb)
Sigma Bottom - Σ_b^+ (uub)
Sigma Bottom - Σ_b^0 (udb)
Sigma Bottom - Σ_b^- (ddb)
Sigma Plus - Σ^+ (uus)
Sigma Zero - Σ^0 (uds)
Sigma Minus - Σ^- (dds)
(plus antiparticles...)

Mesons

Pion - π^+ $(u\bar{d})$
Pion - π^0 $(u\bar{u}+d\bar{d}/\sqrt{2})$
Kaon - K^+ $(u\bar{s})$
Kaon - K^0 $(d\bar{s})$
Kaon - K_S^0 $(d\bar{s}+s\bar{d}/\sqrt{2})$
Meson B - B^+ $(u\bar{b})$
Meson B - B^0 $(d\bar{b})$
Meson B Strange - B_s^0 $(s\bar{b})$
Meson B Charm - B_c^+ $(c\bar{b})$
Meson Eta - η^0 $(u\bar{u}+d\bar{d}-2s\bar{s}/\sqrt{6})$
Meson Eta - $\eta^{'}$ $(u\bar{u}+d\bar{d}-s\bar{s}/\sqrt{3})$
Meson Eta - η_c $(c\bar{c})$
Meson Eta - η_b $(b\bar{b})$
Meson Rho - ρ^+ $(u\bar{d})$
Meson Phi - φ^+ $(s\bar{s})$
Meson D - D^+ $(c\bar{d})$
Meson D - D^0 $(c\bar{u})$
Meson D - D_S^+ $(c\bar{s})$
Meson Upsilon - Υ $(b\bar{b})$
(plus antiparticles...)

Figure 11.156: **List of common baryons, hadrons, leptons, mesons and bosons**

On the next page the reader can see a pretty and useful diagram of some of the particles listed in the figure above:

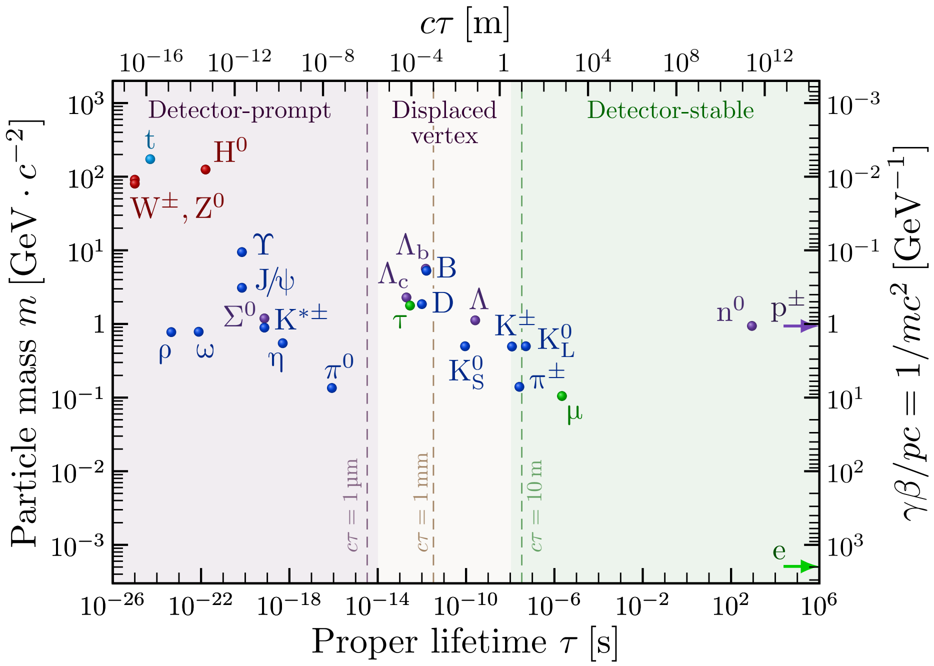


Figure 11.157: Particles mass versus Lifetime decay length (credit: Izaak Neutelings)

And for the general culture here is a timeline (based on the Gregorian Calendar sadly...) of the year of discovery of most well known particles:

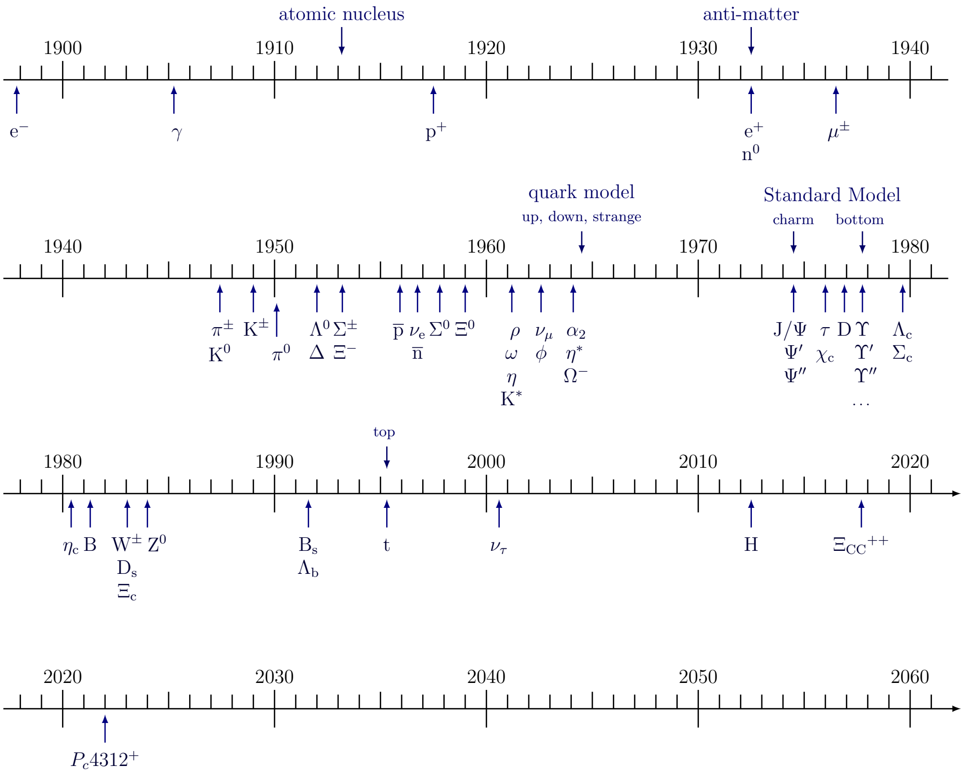
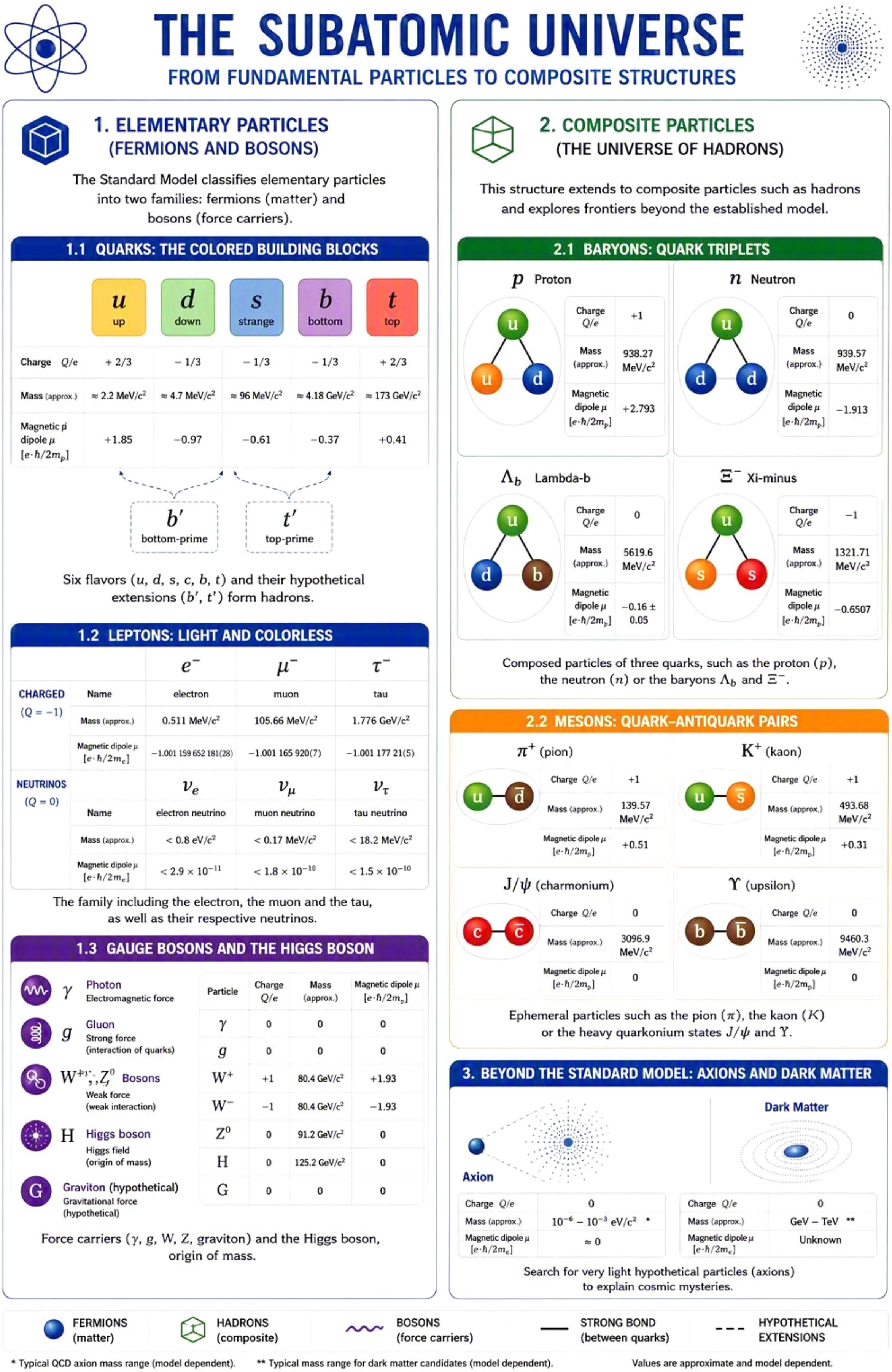

atomic nucleus
anti-matter
1900
1910
1920
1930
1940
e^−
γ
p^+
e^+
n^0
μ^±
quark model
up, down, strange
Standard Model
charm
bottom
1940
1950
1960
1970
1980
π^±
K^0
K^±
π^0
Λ^0
Σ^±
Δ
Ξ^−
p̄
ν_e
n̄
Σ^0
Ξ^0
ρ
ω
η
K*
ν_μ
ϕ
α_2
η*
Ω^−
J/Ψ
Ψ′
Ψ″
τ
χ_c
D
ϒ
ϒ′
ϒ″
…
Λ_c
Σ_c
top
1980
1990
2000
2010
2020
η_c
B
W^±
Z^0
D_s
Ξ_c
B_s
Λ_b
t
ν_τ
H
Ξ_CC^++
2020
2030
2040
2050
2060
P_c4312^+

Figure 11.158: Timeline of particles discovery (credit: Izaak Neutelings)

And here a summary that we did with A.I. :



THE SUBATOMIC UNIVERSE

FROM FUNDAMENTAL PARTICLES TO COMPOSITE STRUCTURES

1. ELEMENTARY PARTICLES (FERMIONS AND BOSONS)

The Standard Model classifies elementary particles into two families: fermions (matter) and bosons (force carriers).

1.1 QUARKS: THE COLORED BUILDING BLOCKS

	u up	d down	s strange	b bottom	t top
Charge Q/e	+ 2/3	− 1/3	− 1/3	− 1/3	+ 2/3
Mass (approx.)	≈ 2.2 MeV/c^2	≈ 4.7 MeV/c^2	≈ 96 MeV/c^2	≈ 4.18 GeV/c^2	≈ 173 GeV/c^2
Magnetic μ dipole μ $[e\cdot\hbar/2m_p]$	+1.85	−0.97	−0.61	−0.37	+0.41

b'
bottom-prime

t'
top-prime

Six flavors (u, d, s, c, b, t) and their hypothetical extensions (b', t') form hadrons.

1.2 LEPTONS: LIGHT AND COLORLESS

		e^-	μ^-	τ^-
CHARGED ($Q=-1$)	Name	electron	muon	tau
	Mass (approx.)	0.511 MeV/c^2	105.66 MeV/c^2	1.776 GeV/c^2
	Magnetic dipole μ $[e\cdot\hbar/2m_e]$	−1.001 159 652 181(28)	−1.001 165 920(7)	−1.001 177 21(5)
NEUTRINOS ($Q=0$)		ν_e	ν_μ	ν_τ
	Name	electron neutrino	muon neutrino	tau neutrino
	Mass (approx.)	< 0.8 eV/c^2	< 0.17 MeV/c^2	< 18.2 MeV/c^2
	Magnetic dipole μ $[e\cdot\hbar/2m_e]$	$< 2.9\times10^{-11}$	$< 1.8\times10^{-10}$	$< 1.5\times10^{-10}$

The family including the electron, the muon and the tau, as well as their respective neutrinos.

1.3 GAUGE BOSONS AND THE HIGGS BOSON

γ **Photon**
Electromagnetic force

g **Gluon**
Strong force
(interaction of quarks)

$W^{+,-}, Z^0$ **Bosons**
Weak force
(weak interaction)

H **Higgs boson**
Higgs field
(origin of mass)

G **Graviton (hypothetical)**
Gravitational force
(hypothetical)

Particle	Charge Q/e	Mass (approx.)	Magnetic dipole μ $[e\cdot\hbar/2m_p]$
γ	0	0	0
g	0	0	0
W^+	+1	80.4 GeV/c^2	+1.93
W^-	−1	80.4 GeV/c^2	−1.93
Z^0	0	91.2 GeV/c^2	0
H	0	125.2 GeV/c^2	0
G	0	0	0

Force carriers (γ, g, W, Z, graviton) and the Higgs boson, origin of mass.

2. COMPOSITE PARTICLES (THE UNIVERSE OF HADRONS)

This structure extends to composite particles such as hadrons and explores frontiers beyond the established model.

2.1 BARYONS: QUARK TRIPLETS

p Proton

Charge Q/e	+1
Mass (approx.)	938.27 MeV/c^2
Magnetic dipole μ $[e\cdot\hbar/2m_p]$	+2.793

n Neutron

Charge Q/e	0
Mass (approx.)	939.57 MeV/c^2
Magnetic dipole μ $[e\cdot\hbar/2m_p]$	−1.913

Λ_b Lambda-b

Charge Q/e	0
Mass (approx.)	5619.6 MeV/c^2
Magnetic dipole μ $[e\cdot\hbar/2m_p]$	−0.16 ± 0.05

Ξ^- Xi-minus

Charge Q/e	−1
Mass (approx.)	1321.71 MeV/c^2
Magnetic dipole μ $[e\cdot\hbar/2m_p]$	−0.6507

Composed particles of three quarks, such as the proton (p), the neutron (n) or the baryons Λ_b and Ξ^-.

2.2 MESONS: QUARK–ANTIQUARK PAIRS

π^+ (pion)

Charge Q/e	+1
Mass (approx.)	139.57 MeV/c^2
Magnetic dipole μ $[e\cdot\hbar/2m_p]$	+0.51

K^+ (kaon)

Charge Q/e	+1
Mass (approx.)	493.68 MeV/c^2
Magnetic dipole μ $[e\cdot\hbar/2m_p]$	+0.31

J/ψ (charmonium)

Charge Q/e	0
Mass (approx.)	3096.9 MeV/c^2
Magnetic dipole μ $[e\cdot\hbar/2m_p]$	0

Υ (upsilon)

Charge Q/e	0
Mass (approx.)	9460.3 MeV/c^2
Magnetic dipole μ $[e\cdot\hbar/2m_p]$	0

Ephemeral particles such as the pion (π), the kaon (K) or the heavy quarkonium states J/ψ and Υ.

3. BEYOND THE STANDARD MODEL: AXIONS AND DARK MATTER

Axion

Charge Q/e	0
Mass (approx.)	$10^{-6}-10^{-3}$ eV/c^2 *
Magnetic dipole μ $[e\cdot\hbar/2m_e]$	≈ 0

Dark Matter

Charge Q/e	0
Mass (approx.)	GeV – TeV **
Magnetic dipole μ $[e\cdot\hbar/2m_e]$	Unknown

Search for very light hypothetical particles (axions) to explain cosmic mysteries.

●	FERMIONS (matter)
⬡	HADRONS (composite)
∿	BOSONS (force carriers)
—	STRONG BOND (between quarks)
- - -	HYPOTHETICAL EXTENSIONS

* Typical QCD axion mass range (model dependent).

** Typical mass range for dark matter candidates (model dependent).

Values are approximate and model dependent.

11.6.2 Spin magnetic resonance

We have hesitated to treat the subject of spin resonance in this section but after reflection, it's not nuclear physics because the calculations does not apply only well to the nucleus of atoms and this is not really pure relativistic quantum physics because it does not only apply well to elementary particles like electrons (as assumed in our developments in the section of Relativistic Quantum Physics).

During our proof of the Pauli equation in the section of Relativistic Quantum Physics using physicist art of life... we saw that for a particle of spin $^1/_2$ (which could be a nuclear nucleus of spin $^1/_2$), we had:

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\phi_{a0}\cong\left[\frac{\|\vec{p}\|^2}{2m_0}-\frac{q\vec{B}\circ\vec{L}}{2m_0}-\frac{\hbar q}{2m_0}\vec{\sigma}\circ\vec{B}+qU\right]\phi_{a0} \tag{11.6.8}$$

where there is therefore a term relatively to the spin in the Hamiltonian, that is:

$$H_S=-\frac{\hbar q}{2m}\vec{\sigma}\circ\vec{B}=-\mu\vec{\sigma}\circ\vec{B}=-\frac{2}{2}\mu\vec{\sigma}\circ\vec{B}=-\frac{g\mu}{2}\vec{\sigma}\circ\vec{B} \tag{11.6.9}$$

By putting:

$$g=2 \tag{11.6.10}$$

Being for reminder the "Lande factor" or "gyromagnetic factor" (see section Relativistic Quantum Physics page 4985).

Now let us dive the particle that will be named "probe" in a stable oscillating magnetic field following a horizontal plane and constant on the vertical plane of the type:

$$\vec{B}=\begin{pmatrix}B_A\cos(\omega t)\\-B_A\sin(\omega t)\\B_0\end{pmatrix} \tag{11.6.11}$$

where in practice the field B_0 goes form 0.1 to 17 Teslas.

Therefore it comes:

$$\begin{aligned}H_S&=-\frac{g\mu}{2}\vec{\sigma}\circ\vec{B}=-\frac{g\mu}{2}\begin{pmatrix}\sigma_1\\\sigma_2\\\sigma_3\end{pmatrix}\circ\begin{pmatrix}B_A\cos(\omega t)\\-B_A\sin(\omega t)\\B_0\end{pmatrix}\\&=-\frac{g\mu}{2}\left(\sigma_1B_A\cos(\omega t)-\sigma_2B_A\sin(\omega t)+\sigma_3B_0\right)\\&=-\frac{g\mu}{2}\left(\begin{bmatrix}0&1\\1&0\end{bmatrix}B_A\cos(\omega t)-\begin{bmatrix}0&-\mathrm{i}\\\mathrm{i}&0\end{bmatrix}B_A\sin(\omega t)+\begin{bmatrix}1&0\\0&-1\end{bmatrix}B_0\right)\\&=-\frac{g\mu}{2}\left(\begin{bmatrix}0&B_A\cos(\omega t)\\B_A\cos(\omega t)&0\end{bmatrix}-\begin{bmatrix}0&-\mathrm{i}B_A\sin(\omega t)\\\mathrm{i}B_A\sin(\omega t)&0\end{bmatrix}+\begin{bmatrix}B_0&0\\0&-B_0\end{bmatrix}\right)\\&=-\frac{g\mu}{2}\begin{bmatrix}B_0&B_A\cos(\omega t)+\mathrm{i}B_A\sin(\omega t)\\B_A\cos(\omega t)-\mathrm{i}B_A\sin(\omega t)&-B_0\end{bmatrix}\\&=-\frac{g\mu}{2}\begin{bmatrix}B_0&B_Ae^{\mathrm{i}\omega t}\\B_Ae^{-\mathrm{i}\omega t}&-B_0\end{bmatrix}\end{aligned} \tag{11.6.12}$$

We then have by focusing only on this Hamiltonian:

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\phi_{a0}=-\frac{g\mu}{2}\begin{bmatrix}B_0 & B_Ae^{\mathrm{i}\omega t}\\ B_Ae^{-\mathrm{i}\omega t} & -B_0\end{bmatrix}\phi_{a0} \tag{11.6.13}$$

And as ϕ_{a0} is the double component of a spinor, let us write it explicitly:

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\begin{pmatrix}\alpha_+(t)\\ \alpha_-(t)\end{pmatrix}=-\frac{g\mu}{2}\begin{bmatrix}B_0 & B_Ae^{\mathrm{i}\omega t}\\ B_Ae^{-\mathrm{i}\omega t} & -B_0\end{bmatrix}\begin{pmatrix}\alpha_+(t)\\ \alpha_-(t)\end{pmatrix} \tag{11.6.14}$$

Thus after rearrangement:

$$\frac{\partial}{\partial t}\begin{pmatrix}\alpha_+(t)\\ \alpha_-(t)\end{pmatrix}=\frac{\mathrm{i}}{2}\begin{bmatrix}\dfrac{g\mu}{\hbar}B_0 & \dfrac{g\mu}{\hbar}B_Ae^{\mathrm{i}\omega t}\\ \dfrac{g\mu}{\hbar}B_Ae^{-\mathrm{i}\omega t} & -\dfrac{g\mu}{\hbar}B_0\end{bmatrix}\begin{pmatrix}\alpha_+(t)\\ \alpha_-(t)\end{pmatrix} \tag{11.6.15}$$

And let us recall that we proved at the very end of the section of Relativistic Quantum Physics that:

$$\omega=-g\gamma B=-\frac{g\mu_BB}{\hbar} \tag{11.6.16}$$

It comes then naturally to write:

$$\omega_0=-g\gamma B=-\frac{g\mu_BB_0}{\hbar} \tag{11.6.17}$$

$$\omega_A=-g\gamma B=-\frac{g\mu_BB_A}{\hbar} \tag{11.6.18}$$

This gives us:

$$2\begin{pmatrix}\dot{\alpha}_+(t)\\ \dot{\alpha}_-(t)\end{pmatrix}=\mathrm{i}\begin{bmatrix}\omega_0 & \omega_Ae^{\mathrm{i}\omega t}\\ \omega_Ae^{-\mathrm{i}\omega t} & -\omega_0\end{bmatrix}\begin{pmatrix}\alpha_+(t)\\ \alpha_-(t)\end{pmatrix} \tag{11.6.19}$$

that is to say a system of two partial differential equation:

$$\begin{aligned}2\dot{\alpha}_+(t)&=\mathrm{i}\left(\omega_0\alpha_+(t)+\omega_Ae^{\mathrm{i}\omega t}\alpha_-(t)\right)\\ 2\dot{\alpha}_-(t)&=\mathrm{i}\left(\omega_Ae^{-\mathrm{i}\omega t}\alpha_+(t)-\omega_0\alpha_-(t)\right)\end{aligned} \tag{11.6.20}$$

If we assume that the spinor is orient to the top initially (initial conditions), then:

$$\begin{pmatrix}\alpha_+(0)\\ \alpha_-(0)\end{pmatrix}=\begin{pmatrix}1\\ 0\end{pmatrix} \tag{11.6.21}$$

Now to solve this system of differential equations we will use the laborious trial and error work already done by our predecessors by assuming that the answer is probably of the type:

$$\begin{aligned}\alpha_+(t)&=A_+e^{\mathrm{i}\omega_+t}=|+\rangle\\ \alpha_-(t)&=A_-e^{\mathrm{i}\omega_-t}=|-\rangle\end{aligned} \tag{11.6.22}$$

that is to say... wave functions!!!

So by injecting these solutions in the system of differential equations we get:

$$\begin{aligned}2\frac{\mathrm{d}}{\mathrm{d}t}A_+e^{\mathrm{i}\omega_+t}&=\mathrm{i}\left(\omega_0A_+e^{\mathrm{i}\omega_+t}+\omega_Ae^{\mathrm{i}\omega t}A_-e^{\mathrm{i}\omega_-t}\right)\\2\frac{\mathrm{d}}{\mathrm{d}t}A_-e^{\mathrm{i}\omega_-t}&=\mathrm{i}\left(\omega_Ae^{-\mathrm{i}\omega t}A_+e^{\mathrm{i}\omega_+t}-\omega_0A_-e^{\mathrm{i}\omega_-t}\right)\end{aligned}\tag{11.6.23}$$

That is to say:

$$\begin{aligned}2\mathrm{i}\omega_+A_+e^{\mathrm{i}\omega_+t}&=\mathrm{i}\left(\omega_0A_+e^{\mathrm{i}\omega_+t}+\omega_AA_-e^{\mathrm{i}\omega t+\mathrm{i}\omega_-t}\right)\\2\mathrm{i}\omega_-A_-e^{\mathrm{i}\omega_-t}&=\mathrm{i}\left(\omega_AA_+e^{-\mathrm{i}\omega t+\mathrm{i}\omega_+t}-\omega_0A_-e^{\mathrm{i}\omega_-t}\right)\end{aligned}\tag{11.6.24}$$

and after simplification by the exponential and the pure imaginary number:

$$\begin{aligned}2\omega_+A_+&=\omega_0A_++\omega_AA_-e^{\mathrm{i}\omega t+\mathrm{i}\omega_-t-\mathrm{i}\omega_+t}\\2\omega_-A_-&=\omega_AA_+e^{-\mathrm{i}\omega t+\mathrm{i}\omega_+t-\mathrm{i}\omega_-t}-\omega_0A_-\end{aligned}\tag{11.6.25}$$

Hence:

$$\begin{aligned}2\omega_+A_+&=\omega_0A_++\omega_AA_-e^{\mathrm{i}(\omega+\omega_--\omega_+)t}\\2\omega_-A_-&=\omega_AA_+e^{-\mathrm{i}(\omega-\omega_++\omega_-)t}-\omega_0A_-\end{aligned}\tag{11.6.26}$$

We can get rid of the time dependence above by putting:

$$\omega_+=\omega_-+\omega\tag{11.6.27}$$

Our system then reduces to:

$$\begin{aligned}2\omega_+A_+&=\omega_0A_++\omega_AA_-\\2\omega_-A_-&=\omega_AA_+-\omega_0A_-\end{aligned}\tag{11.6.28}$$

Thus by rearranging:

$$\begin{aligned}A_+(2\omega_+-\omega_0)-\omega_AA_-&=0\\-\omega_AA_++A_-(2\omega_-+\omega_0)&=0\end{aligned}\tag{11.6.29}$$

which can be written in matrix form:

$$\begin{bmatrix}2\omega_+-\omega_0&-\omega_A\\-\omega_A&2\omega_-+\omega_0\end{bmatrix}\begin{pmatrix}A_+\\A_-\end{pmatrix}=0\tag{11.6.30}$$

To get a consistent solution (not all zero solutions) and unique one, the determinant of the matrix must be zero to be reversible (see section Linear Algebra page 1980). But the determinant of this matrix is:

$$\begin{aligned}&\det\left(\begin{bmatrix}2\omega_+-\omega_0&-\omega_A\\-\omega_A&2\omega_-+\omega_0\end{bmatrix}\right)=(2\omega_+-\omega_0)(2\omega_-+\omega_0)-(-\omega_A)(-\omega_A)\\&=(2\omega_+-\omega_0)(2\omega_-+\omega_0)-\omega_A^2=4\omega_+\omega_-+2\omega_+\omega_0-2\omega_-\omega_0-\omega_0^2-\omega_A^2\\&=4(\omega_-+\omega)\omega_-+2(\omega_-+\omega)\omega_0-2\omega_-\omega_0-\omega_0^2-\omega_A^2\\&=4\omega_-^2+4\omega\omega_-+2\omega_-\omega_0+2\omega\omega_0-2\omega_0\omega_0-\omega_0^2-\omega_A^2\\&=4\omega_-^2+4\omega\omega_-+2\omega\omega_0-\omega_0^2-\omega_A^2\end{aligned}\tag{11.6.31}$$

But the roots of this polynomial of second degree in ω_- are (see section Calculus page 1383):

$$\begin{aligned}\omega_{-,1,2}&=\frac{-4\omega\pm\sqrt{(4\omega)^2-4\cdot 4\cdot(2\omega\omega_0-\omega_0^2-\omega_A^2)}}{2\cdot 4}\\&=\frac{-\omega\pm\sqrt{\omega^2-(2\omega\omega_0-\omega_0^2-\omega_A^2)}}{2}=\frac{-\omega\pm\sqrt{\omega^2-(2\omega\omega_0-\omega_0^2)+\omega_A^2}}{2}\\&=\frac{-\omega\pm\sqrt{\omega^2-2\omega\omega_0+\omega_0^2+\omega_A^2}}{2}=\frac{-\omega\pm\sqrt{(\omega-\omega_0)^2+\omega_A^2}}{2}\\&=-\frac{\omega}{2}\pm\frac{\sqrt{(\omega-\omega_0)^2+\omega_A^2}}{2}=-\frac{\omega}{2}\pm\Delta\omega\end{aligned} \tag{11.6.32}$$

So we deduce:

$$\omega_+=\omega_-+\omega\Rightarrow\omega_{+,1,2}=\omega_{-,1,2}+\omega=-\frac{\omega}{2}\pm\Delta\omega+\omega=+\frac{\omega}{2}\pm\Delta\omega \tag{11.6.33}$$

So we have for summary:

$$\begin{aligned}\omega_{-,1,2}&=-\frac{\omega}{2}\pm\Delta\omega\\\omega_{+,1,2}&=+\frac{\omega}{2}\pm\Delta\omega\end{aligned} \tag{11.6.34}$$

Results to inject in:

$$\begin{aligned}\alpha_+(t)&=A_+e^{\mathrm{i}\omega_+t}\\\alpha_-(t)&=A_-e^{\mathrm{i}\omega_-t}\end{aligned} \tag{11.6.35}$$

So as each pulsation has two solution, the general solution will be the sum of the special solutions (see section Differential and Integral Calculus page 1605). Let us focus on the second relation. We then have:

$$\alpha_-(t)=A_-e^{-\mathrm{i}\frac{\omega}{2}t}e^{+\mathrm{i}\Delta\omega t}+B_-e^{-\mathrm{i}\frac{\omega}{2}t}e^{-\mathrm{i}\Delta\omega t} \tag{11.6.36}$$

But let us recall the initial condition that we required for the second component of the spinor (due to the fact that the spinor is oriented to the top):

$$\alpha_-(0)=0 \tag{11.6.37}$$

Then we have:

$$A_-=-B_-:=A \tag{11.6.38}$$

Therefore:

$$\alpha_-(t)=Ae^{-\mathrm{i}\frac{\omega}{2}t}e^{+\mathrm{i}\Delta\omega t}-Ae^{-\mathrm{i}\frac{\omega}{2}t}e^{-\mathrm{i}\Delta\omega t} \tag{11.6.39}$$

with:

$$\dot{\alpha}_-(t)=\mathrm{i}A\left(-\frac{\omega}{2}+\Delta\omega\right)e^{-\mathrm{i}\frac{\omega}{2}t}e^{+\mathrm{i}\Delta\omega t}-\mathrm{i}A\left(-\frac{\omega}{2}-\Delta\omega\right)e^{-\mathrm{i}\frac{\omega}{2}t}e^{-\mathrm{i}\Delta\omega t} \tag{11.6.40}$$

and:

$$\dot{\alpha}_-(0)=\mathrm{i}A\left(-\frac{\omega}{2}+\Delta\omega\right)-\mathrm{i}A\left(-\frac{\omega}{2}-\Delta\omega\right)=2\mathrm{i}A\Delta\omega \tag{11.6.41}$$

Now, to determine A, we will use in a clever way the differential we started from earlier above:

$$2\dot{\alpha}(t)=\mathrm{i}\left(\omega_A e^{-\mathrm{i}\omega t\alpha_+(t)}-\omega_0\alpha_-(t)\right) \tag{11.6.42}$$

without forgetting the chosen initial conditions, we have:

$$2\dot{\alpha}(0)=\mathrm{i}\left(\omega_A e^{-\mathrm{i}\omega 0\alpha_+(0)}-\omega_0\alpha_-(0)\right)=\mathrm{i}(\omega_A\cdot 1-\omega_0\cdot 0)=\mathrm{i}\omega_A \tag{11.6.43}$$

Therefore:

$$\dot{\alpha}_-(0)=\frac{\mathrm{i}\omega_A}{2} \tag{11.6.44}$$

We must have the equality between the both expressions:

$$\begin{aligned}\dot{\alpha}_-(0)&=\frac{\mathrm{i}\omega_A}{2}\\ \dot{\alpha}_-(0)&=2\mathrm{i}A\Delta\omega\end{aligned} \tag{11.6.45}$$

and this can be satisfied obviously if and only if:

$$A=\frac{\omega_A}{4\Delta\omega} \tag{11.6.46}$$

Then we have:

$$\begin{aligned}\alpha_-(t)&=\frac{\omega_A}{4\Delta\omega}e^{-\mathrm{i}\frac{\omega}{2}t}\left(e^{+\mathrm{i}\Delta\omega t}-e^{+\mathrm{i}\Delta\omega t}\right)=\frac{\omega_A}{4\Delta\omega}e^{-\mathrm{i}\frac{\omega}{2}t}2\mathrm{i}\sin(\mathrm{i}\Delta\omega t)\\ &=\frac{\omega_A}{2\Delta\omega}\mathrm{i}\sin(\mathrm{i}\Delta\omega t)e^{-\mathrm{i}\frac{\omega}{2}t}\end{aligned} \tag{11.6.47}$$

Thus explicitly:

$$\alpha_-(t)=\frac{\omega_A}{\sqrt{(\omega-\omega_0)^2+\omega_A^2}}\mathrm{i}\sin(\mathrm{i}\Delta\omega t)e^{-\mathrm{i}\frac{\omega}{2}t} \tag{11.6.48}$$

As the second component of the spinor represents the angle and what we have above represents the angular position over time, the term:

$$\boxed{\frac{\omega_A}{\sqrt{(\omega-\omega_0)^2+\omega_A^2}}} \tag{11.6.49}$$

can be seen as the maximum value of the amplitude of the angle (Larmor rotation of the spin of the electron around the magnetic field). This maximum amplitude has itself a maximum if the denominator is as small as possible and therefore:

$$\omega=\omega_0=\frac{g\mu}{\hbar}B_0 \tag{11.6.50}$$

Then we say that there is "spin resonance".

So the magnetic field may make switch the energy state of each spin if the oscillation amplitude is maximum. As we have just proved above for an isolated particle, the energy change from one state to the other is:

$$E_\pm=\pm\frac{1}{2}\hbar\omega \tag{11.6.51}$$

and therefore when a spin flip it generates the emission of radiation named "free induction decay". The collected signal depends on several parameters that characterize the particle or the nuclear core. In short, in practice it is 99% engineering and 1% theory!

When we apply these theoretical results to an unpaired electron, that is to say, for the "electronic magnetic resonance", we have the following experimental data for a constant field $B_0 \cong 0.3$ [T]:

$$f \cong 10\,[\text{GHz}],\ \lambda \cong 3\,[\text{cm}],\ E_\gamma = \Delta E \cong 4 \cdot 10^{-4}\,[\text{eV}] \tag{11.6.52}$$

and therefore the electronic magnetic resonance technology is based on the detection of microwaves.

When we apply these theoretical results to a nuclear core of a single proton, that is to say, for the "nuclear magnetic resonance", we have the following experimental data for a constant field $B_0 \cong 1$ [T]:

$$f \cong 40\,[\text{MHz}],\ \lambda \cong 7\,[\text{m}],\ E_\gamma = \Delta E \cong 2 \cdot 10^{-7}\,[\text{eV}] \tag{11.6.53}$$

and therefore the nuclear magnetic resonance technology is based on the detection of radio waves.

Nowadays, nuclear magnetic resonance is also used to check the quality of food. For example, modern machines can detect whether orange juice is contaminated with juice from other fruit and can check whether the fruit were ripe when pressed. Other machines can check whether wine was made from the correct grapes and how it aged.

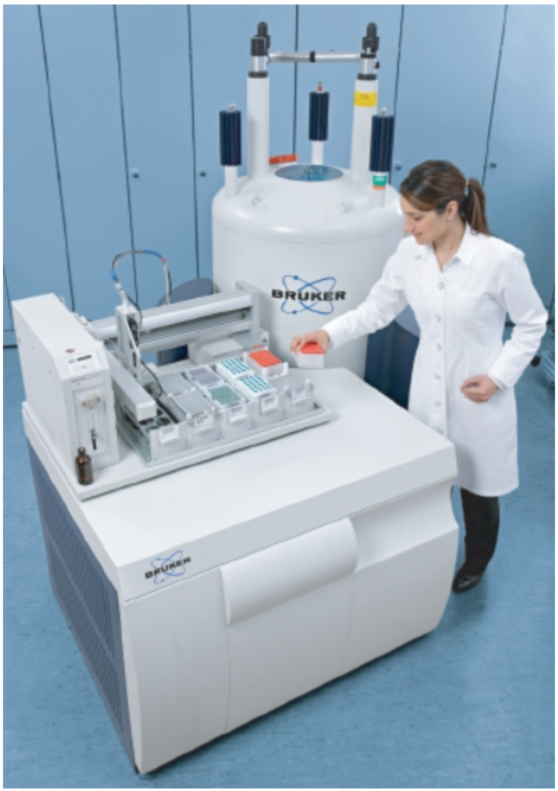


Figure 11.159: Machine to test fruit quality with the help of nuclear magnetic resonance (source: Bruker)

~10 %

★★★★★

26 votes, 56.92%

12

Cosmology

Cosmology is the science that studies the structure, evolution and the general laws of the universe as a whole. (Larousse)

Contents

12.1 Astronomy (Celestial Mechanics) **5217**
12.1.1 Sun, Moon distances and radii 5218
12.1.2 Drake Equation 5225
12.1.2.1 Habitale Zone (HZ) 5227
12.1.3 Kepler's Laws 5233
12.1.3.1 First Kepler's Law (conicity law) 5234
12.1.3.2 Second Kepler's Law (area law) 5234
12.1.3.3 Third Kepler's Law (periods' law) 5242
12.1.4 Newton's Gravitational Law 5255
12.1.4.1 Cavendish Experiment 5270
12.1.4.2 Gaussian formulation of Newtonian Gravity 5277
12.1.4.3 Shell Theorem 5281
12.1.4.4 Orbital speed 5282
12.1.4.5 Gravitational field of a homogeneous disc 5284
12.1.4.6 Collision time of two objects at rest 5295
12.1.4.7 Asteroids/Meteors impact velocity 5297
12.1.4.8 Spherisation of Celestial Bodies 5299
12.1.4.9 Stability of Atmospheres 5331
12.1.4.10 Planetary equilibrium temperature 5335
12.1.5 Roche's Limit 5337
12.1.6 Keplerian Orbitals 5340
12.1.6.1 First Binet Formula (radial velocity) 5340

12.1.6.2 Second Binet Formula (radial acceleration) . . . 5346
12.1.6.3 Keplerian orbital period . . . 5350
12.1.6.4 Average distance between two planets . . . 5356
12.1.6.5 Classical deflection of light (light bending) . . . 5358
12.1.6.6 Classical precession of perihelion (apsidal precession) . . . 5361
12.1.7 Duration of the diurnal arc . . . 5372
12.1.7.1 Evolution of diurnal arc (i.e. Earth-Moon distance) . . . 5379
12.1.8 Trigonometric parallax . . . 5383
12.1.9 Planets' Motion . . . 5388
12.1.9.1 Synodic and Sidereal period . . . 5390
12.1.9.2 Planet's apparent retrograde motion . . . 5395
12.1.10 Equation of Time and Analemma . . . 5404
12.1.10.1 Eccentricity component (variable angular velocity of Earth) . . . 5407
12.1.10.2 Obliquity (inclined plane of the ecliptic) and Sun anomaly . . . 5419
12.1.10.3 Calculation of the Equation of Time . . . 5423
12.1.11 Sundials . . . 5438
12.1.11.1 Equatorial sundials . . . 5441
12.1.11.2 Vertical sundials . . . 5443
12.1.11.3 Horizontal sundials . . . 5446
12.1.11.4 Anammelatic (azimuthal) sundials . . . 5449
12.1.12 Apparent solar time (true solar time) . . . 5461
12.1.13 Celestial sphere and coordinate systems (positional astronomy) . . . 5465
12.1.13.1 Local Sidereal Time . . . 5467
12.1.13.2 Equatorial coordinates to Azimuthal (horizontal) coordinates . . . 5472
12.1.13.3 Azimuthal (horizontal) coordinates to Equatorial coordinates . . . 5474
12.1.13.4 Apparent daily path of the Sun (solargraph) . . . 5476
12.1.14 Eclipses and saros cycles . . . 5486
12.1.15 Lagrange Points . . . 5498
12.1.15.1 Equilibrium points of the first type . . . 5509
12.1.15.2 Equilibrium points of the second type . . . 5519
12.1.15.3 Hill sphere (sphere of influence) . . . 5525
12.1.16 Relativistic Doppler-Fizeau Effect . . . 5529
12.1.16.1 Apparent speed . . . 5536
12.2 Astrophysics . . . 5539
12.2.1 Stars . . . 5539
12.2.1.1 Stellar Physics . . . 5549
12.2.1.2 Pulsative Variable Stars . . . 5601

12.2.1.3 Neutron Stars (magnetars/pulsars) 5606
12.2.2 Galaxies 5620
12.2.2.1 Probability of collisions in merging galaxies 5623
12.2.2.2 Radial Speed Anomaly 5625
12.3 Special Relativity 5627
12.3.1 Assumptions and Principles 5628
12.3.1.1 Postulate of Invariance 5629
12.3.1.2 Cosmological Principle 5630
12.3.1.3 Special Relativity Principle 5630
12.3.2 Lorentz Transformations/Boost 5633
12.3.2.1 Displacement four-vector 5637
12.3.2.2 Velocity four-vector 5642
12.3.2.3 Current four-vector 5646
12.3.2.4 Acceleration four-vector 5647
12.3.2.5 Relativistic sum of velocities 5652
12.3.2.6 Relativistic lengths variation (length contraction) 5654
12.3.2.7 Relativistic apparent angle rotation (Lampa–Terrell–Penrose effect) 5657
12.3.2.8 Relativistic time variation (time dilatation) 5659
12.3.2.9 Apparent relativistic mass 5670
12.3.2.10 Relativistic electrodynamics 5694
12.3.3 Minkowski space-time 5703
12.3.3.1 Four-vectors 5706
12.3.3.2 Universe light cone 5707
12.3.4 Empirical evidence of Special Relativity theory 5716
12.4 General Relativity 5720
12.4.1 Assumptions and Principles 5721
12.4.1.1 Equivalence Postulates 5721
12.4.1.2 Mach Principle 5725
12.4.2 Metrics 5728
12.4.2.1 Schild Criteria (Einstein redshift effect Newtonian approach) 5735
12.4.3 Equations of movement 5738
12.4.3.1 Geodesic equations 5747
12.4.3.2 Newtonian Limit 5751
12.4.4 Stress-Energy Tensor 5754
12.4.4.1 Stress-Energy Tensor for a perfect fluid 5762
12.4.4.2 Electromagnetic stress–energy tensor 5763
12.4.5 Einstein's Field Equations 5766

12.4.5.1 Cosmological Constant (CC) 5774
12.4.5.2 Weak field approximation with cosmological constant 5778
12.4.5.3 Einstein-Maxwell equations 5780
12.4.5.4 Schwarzschild Solution 5781
12.4.6 Empirical evidence of General Relativity theory 5795
12.4.6.1 Gravitational Redshift 5796
12.4.6.2 Precession of Mercury's Perihelion 5800
12.4.6.3 Deflection of Light (light bending) 5812
12.4.6.4 Shapiro Effect (delay) 5821
12.4.6.5 Hafele–Keating experiment (general relativity version) 5827
12.4.6.6 Black Holes 5835
12.4.6.7 Gravitational waves 5861
12.4.7 Einstein-Hilbert action 5869
12.4.7.1 Einstein-Hilbert action with cosmological constant 5875
12.4.7.2 Einstein-Hilbert action with matter 5876

12.5 Cosmogony 5878

12.5.1 Newtonian (i.e. non-relativistic) Cosmological Models 5881
12.5.1.1 Hubble's Law 5884
12.5.2 Friedmann Equations (Newtonian derivation) 5892
12.5.2.1 Critical Density and Density parameters 5897
12.5.3 Friedmann-Lemaître Euclidean Cosmological Models 5905
12.5.3.1 Flat spaces ($k=0$) 5905
12.5.3.2 Spherical spaces ($k>0$) 5920
12.5.3.3 Hyperbolic spaces ($k<0$) 5934
12.5.4 Observable Universe 5946
12.5.5 Cosmic Microwave Background (CMB) 5966
12.5.5.1 Decoupling time 5968
12.5.5.2 Matter/Photon decoupling (recombination temperature) 5973
12.5.5.3 Background radiation temperature 5984
12.5.6 Friedmann-Lemaître-Robertson-Walker Cosmological Models 5987
12.5.6.1 Robertson-Walker metric 5987
12.5.6.2 Cosmological Redshift 5994
12.5.6.3 Comobile Universe Diameter 5998
12.5.6.4 Size of the Universe depending on its curvature 6000
12.5.6.5 General Friedmann equations 6004
12.5.6.6 Einstein (static) Universe model 6012
12.5.6.7 Milne linear Universe model 6013

12.5.6.8 Friedmann Universe models 6014
12.5.7 Friedmann historical derivation 6050
12.5.8 The Black Hole Universe 6058
12.5.9 Borde-Guth-Vilenkin Theorem (BGV Theorem) 6061
12.5.10 Empirical evidence of the Cosmic Inflation theory (i.e. Big Bang) 6069
12.6 Quantum Gravity/Cosmology 6073
12.6.1 Newton Quantum Gravity 6074
12.6.2 Wheeler-DeWitt Quantum Cosmology 6083
12.6.2.1 Wheeler-DeWitt equation for Newtonian Universe 6085
12.6.2.2 Wheeler-DeWitt equation for FLRW Universe 6090
12.6.3 Why our Universe is a quite massive unpredictable mess 6097
12.7 String Theory 6099
12.7.1 Wave equation of a transversal string 6101
12.7.2 Non-relativistic Wave equation of a transversal string 6104
12.7.2.1 Nambu-Goto Action 6109
12.7.3 Lagrangian of a String 6117

12.1. Astronomy (Celestial Mechanics)

CELESTIAL MECHANICS is the consequence of the universal Newton's law of attraction (see page 5255) and of the fundamental principle of mechanics (see section Classical Mechanics page 3194). Its main objective is the description of the motion of astronomical objects such as stars and planets using physical and mathematical theories.

In this section we will approach the subject as always in this textbook, in the most elementary possible way (to this day the topics in this section are not technically beyond the level of what was done in the beginning of 120th century (holocene calendar) in the field of astronomy).

First we will make a warm up with a funny law on the living in the Universe ... (the Drake equation). Once completed this warm up, we will begin to "enumerate" Kepler's laws (often referring to the section of Classical Mechanics) and then study in detail the properties of Keplerian orbits thanks to our knowledge on classical mechanics and then to using Special Relativity, which will lead us to find a theoretical procession of studied orbitals. Then we will have fun to model approximately the variation of the duration of the day (or night) on the Earth based on the month and latitude. Finally, to finish in style, we will launch the detailed calculation of the five Lagrangian points!

Definition (#12.1.521): It seems that Plato formulated the goal of "astronomy" as the purpose to represent the visible motion of celestial bodies by a combination of uniform rotations on circles. This is actually a very reasonable approach since we know that every bounded motion (path) can be approximated by linear combinations of motions on circles with constant speeds. This approximation is named as we already know the "Fourier series", and this is how the motion of planets is represented even now, for the purposes of prediction and computation.

Remark

Modern astronomy, that of the end and beginning of the 121st century (holocene calendar), makes extensive use of statistical methods (i.e. Data Science/Data Mining). The reader interested on this topic can have an in-deep look at the page 6768 of this textbook or an overview in this excellent book [173].

Astronomy uses a lot of parameters (i.e. "constants") that can be determined only by experimental observations and most of them evolves (obviously!) through time. If the reader is interested in a good reference, we strongly recommend *Handbook of Space Astronomy and Astrophysics* (see [600]).

Caution! Astronomy must not be confused with "Astrology". The latter has not demonstrated its effectiveness in controlled studies and has no scientific validity, and is thus regarded as pseudoscience and a source of income for charlatans (that are not able to found a real job...) that dupe naive and fragile people.

12.1.1 Sun, Moon distances and radii

We have already see how to determine the radius of the Earth in the section of Trigonometry using the method of Al-Bīrūnī (see page 2529) or that of Eratosthenes (see page 2579).

However in this section we will use plenty of times the radius and the distance of the Sun (actually a G2V-type star, aka yellow dwarf) and the Moon. And as requested by some readers, a question arise: what is the oldest and simplest method used to determine these two parameters?

For that let us refer to *On the Sizes and Distances (of the Sun and Moon)* widely accepted as the only extant work written by Aristarchus of Samos, an ancient Greek astronomer who lived circa 9700-9800 century according to holocene calendar (therefore long time before Al-Bīrūnī and almost a century before Eratosthenes). This work calculates the sizes of the Sun (actually a G2V-type star, aka yellow dwarf) and Moon, as well as their distances from the Earth in terms of Earth's radius (that was unknown at his time as determined only later by Eratosthenes!).

Aristarchus began with the premise that, during a half Moon, the Moon forms a right triangle with the Sun (actually a G2V-type star, aka yellow dwarf) and Earth. By observing the angle between the Sun and Moon, φ, the ratio of the distances to the Sun and Moon could be deduced using a form of trigonometry.

Remark

We often associate the Sun with daytime and the Moon with nighttime. The Sun is obviously not visible at night, but the Moon can certainly be observed in broad daylight and it is not always visible at night. It all depends on its phase! But if we associate the Moon with nighttime, it is because we notice it much more easily at night than during the day.

Definitions (#12.1.522): There are three types of twilight, which are stages of partial darkness before nightfall or after sunrise, and night itself. These are civil, nautical, and astronomical twilight, each defined by the position of the Sun below the horizon:

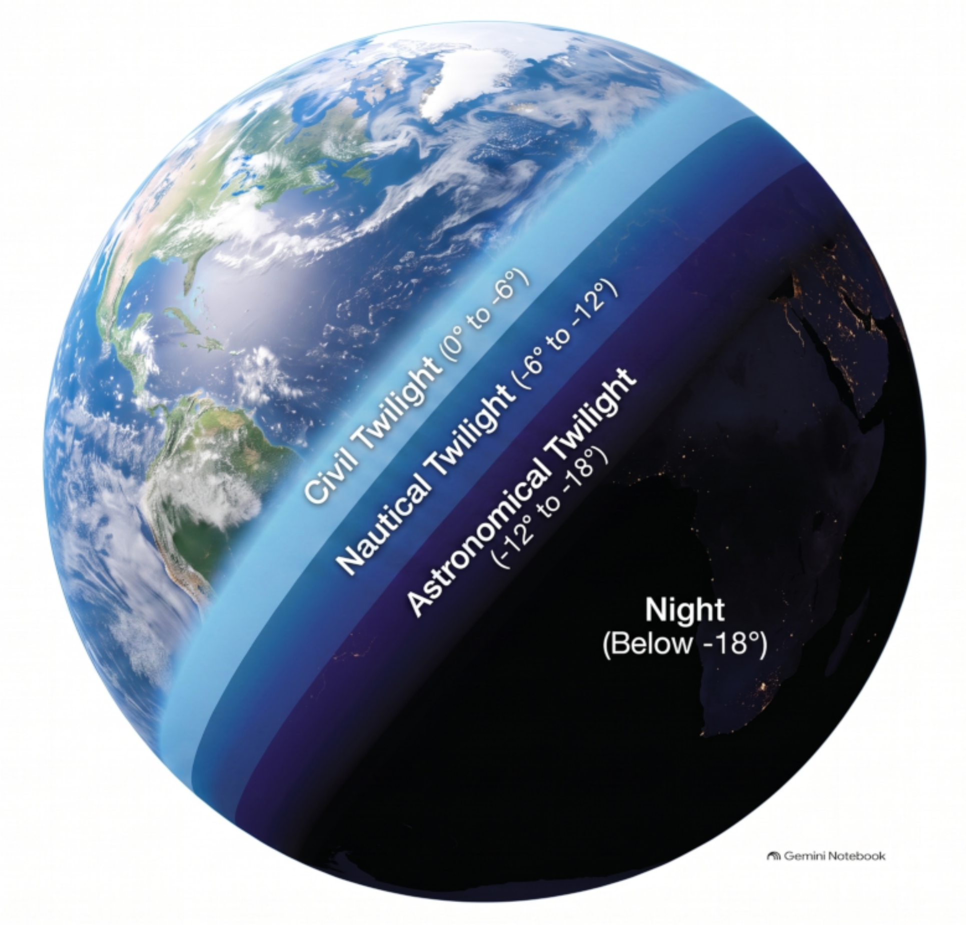
D1. The "civil twilight" is when the Sun is between 0 and 6 degrees below the horizon. This is when it's bright enough to see and do some activities without artificial light, but streetlights and vehicle headlights are often used.

D2. The "nautical twilight" is when Sun is between 6 and 12 degrees below the horizon. It's dark enough that the horizon is still somewhat visible for navigation at sea, as the brightest stars become visible.

D3. The "astronomical twilight" is when the Sun is between 12 and 18 degrees below the horizon. It's dark enough that the sky is beginning to appear dark for astronomical observations.

D4. The "night" is when the Sun is 18 degrees or more below the horizon, it's considered night, and the sky is generally dark, allowing for full stargazing.

These different twilights that exist due to light refraction and Rayleigh diffusion is one of the reason that every day more of the 50% of the Earth is not in the dark[1].



---

[1] Caution! Some newspapers claim that once a year 90% to 99% of the World population is under the light of the day plus of the three twilights. However without citing any scientific peer-review paper that confirms these claims! Remember that what is claimed without evidence can be rejected without evidence...

Let us also introduce some other important definitions:

Definitions (#12.1.523):

D1. For computational purposes, in compliance with the Astronomical Applications Department of the U.S. Naval Observatory, the "sunrise" or "sunset" is defined to occur when the geometric zenith distance of center of the Sun is 90.8333 degrees. That is, the center of the Sun is geometrically 50 arcminutes below a horizontal plane. For an observer at sea level with a level, unobstructed horizon, under average atmospheric conditions, the upper limb of the Sun will then appear to be tangent to the horizon. The 50-arcminute geometric depression of the Sun's center used for the computations is obtained by adding the average apparent radius of the Sun (16 arcminutes) to the average amount of atmospheric refraction at the horizon (34 arcminutes).

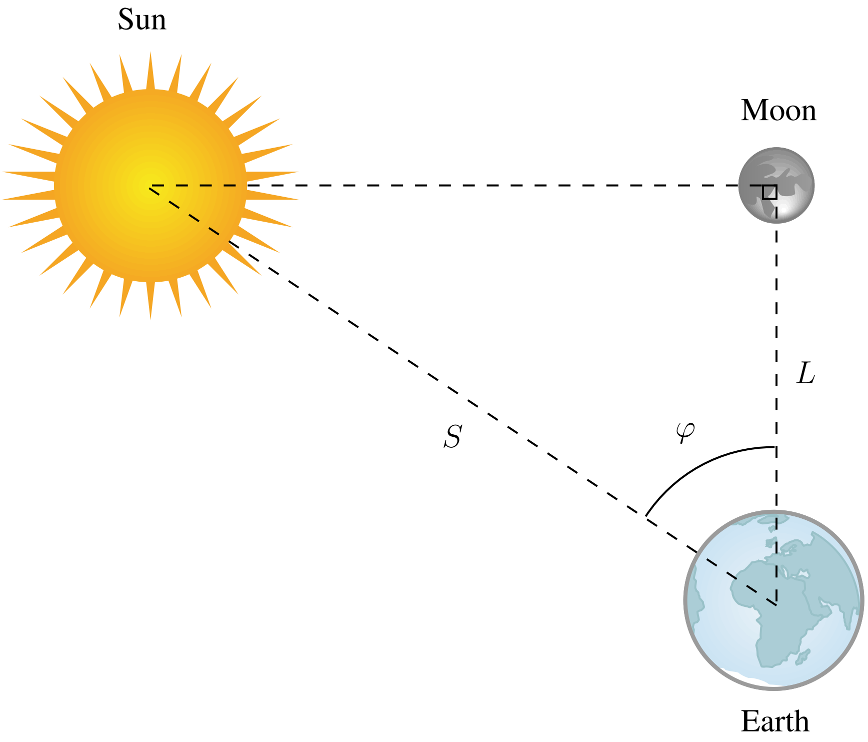
D2. For computational purposes, in compliance with the Astronomical Applications Department of the U.S. Naval Observatory, the "moonrise" and "moonset" are defined similarly, but the situation is computationally more complex because of the nearness of the Moon and the eccentricity of its orbit. If the computations are carried out using coordinates of the Moon with respect to the Earth's center (the usual method), then moonrise or moonset is defined to occur when the geometric zenith distance of the center of the Moon is:

$$90.5666\text{ degrees } + \text{Moon's apparent angular radius} - \text{Moon's horizontal parallax}$$

Under normal atmospheric conditions at sea level, the upper limb of the Moon will then appear to be tangent with a level, unobstructed horizon. No account is taken of the Moon's phase; that is, the Moon is always regarded as a disk in the sky and the upper limb might be dark. Here again, a constant of 34 arcminutes (0.5666 degree) is used to account for atmospheric refraction. The Moon's apparent radius varies from 15 to 17 arcminutes and its horizontal parallax varies from 54 to 61 arcminutes. Adding all the terms above together, the center of the Moon at rise or set is geometrically 5 to 10 arcminutes above the observer's "geocentric horizon" - the horizontal plane that passes through the Earth's center, orthogonal to the observer's local vertical.

D3. A " full Moon" is a lunar phase that occurs when the Moon is approximately opposite the Sun from Earth, that is, when their celestial longitude difference is about 180° (new and full Moons are sometimes named " syzygy" when the Sun and Moon are in conjunction or opposition, respectively, although they are not perfectly aligned with Earth), that is, the Moon is in geocentric opposition. At this time, the Moon's visible side is entirely illuminated by the Sun (due to the limited resolving power of the human eye, a full Moon is considered to occur when the illumination is $\geq 99\%$). The previous definition might suggest an elongation of 180° to the Sun at the instant of each full Moon, but due to the inclination of the lunar orbit (about 5.145°), the angular distance to the Sun at the instant of the full Moon is less. As a consequence, this angular distance can therefore be of the order of 175° and not 180° which makes possible cases of simultaneous observation of the Moon and the Sun at the same time at the precise moment of the full Moon when the observer is typically located near the terminator. This observation may be facilitated by atmospheric refraction which can raise each of the two objects up to 0.88°.

Let us consider the following configuration[2] (obviously not to scale...):



From the diagram and trigonometry, we can calculate that:

$$\frac{S}{L}=\frac{1}{\cos\varphi} \tag{12.1.1}$$

The diagram is greatly exaggerated, because in reality, $S=390\cdot L$, and φ is extremely close to $90°$. Aristarchus determined φ to be a thirtieth of a quadrant (in modern terms, $3°$) less than a right angle: in current terminology, $87°$. Trigonometric functions had not yet been invented, but using geometrical analysis in the style of Euclid, Aristarchus determined that:

$$18<\frac{S}{L}<20 \tag{12.1.2}$$

often denoted:

$$18<\hat{s}<20 \tag{12.1.3}$$

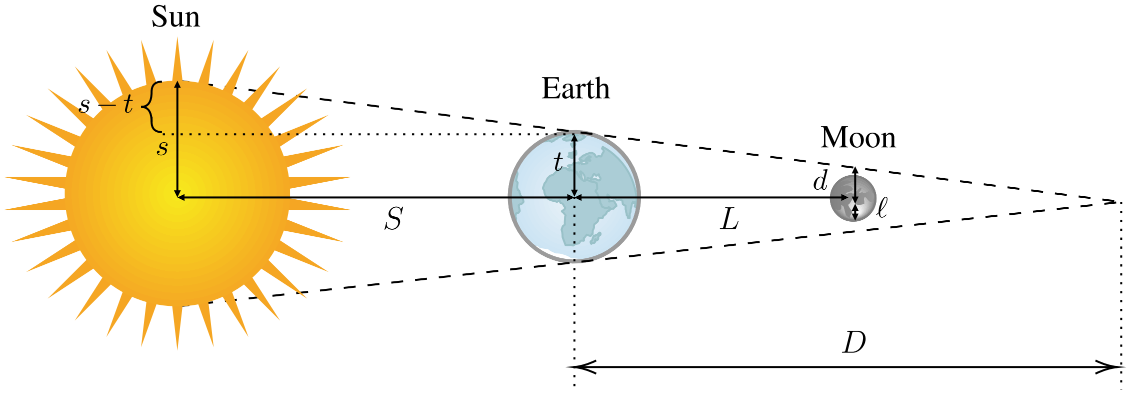
In other words, the distance to the Sun (actually a G2V-type star, aka yellow dwarf) was somewhere between 18 and 20 times greater than the distance to the Moon. This value (or values close to it) was accepted by astronomers for the next two thousand years, until the invention of the telescope permitted a more precise estimate of solar parallax.

Aristarchus also reasoned that as the angular size of the Sun and the Moon were the same, but the distance to the Sun was between 18 and 20 times further than the Moon, the Sun must therefore be 18–20 times larger.

---

[2] If you want a video in real life of the situation you can have a look at `https://www.youtube.com/watch?v=H0Dm_z5oyD4`

Aristarchus then used another construction based on a lunar eclipse as illustrated below (obviously not to scale and don't forget that actually when there is a lunar eclipse only a fraction of the Earth shadow cast on the Moon!):



By similarity of the triangles (see section Euclidean Geometry page 2610), we have:

$$\frac{D}{L}=\frac{t}{t-d} \quad \text{and} \quad \frac{D}{S}=\frac{t}{s-t} \tag{12.1.4}$$

Dividing these two equations and using the observation that the Sun and Moon appear to have the same apparent size (at least actually...) to people on Earth :

$$\frac{s}{S}=\frac{\ell}{L} \tag{12.1.5}$$

yields:

$$\frac{\ell}{s}=\frac{t-d}{s-t} \Longrightarrow \frac{s-t}{s}=\frac{t-d}{\ell} \Longrightarrow \frac{t}{\ell}+\frac{t}{s}=1+\frac{d}{\ell} \tag{12.1.6}$$

The rightmost equation can either be solved for ℓ/t or s/t. Indeed:

$$\frac{t}{\ell}+\frac{t}{s}=1+\frac{d}{\ell} \Rightarrow \frac{t}{\ell}\left(1+\frac{\ell}{s}\right)=1+\frac{d}{\ell} \Rightarrow \frac{\ell}{t}=\frac{1+\dfrac{\ell}{s}}{1+\dfrac{d}{\ell}} \tag{12.1.7}$$

and:

$$\frac{t}{\ell}+\frac{t}{s}=1+\frac{d}{\ell} \Rightarrow \frac{t}{s}\left(\frac{s}{\ell}+1\right)=1+\frac{d}{\ell} \Rightarrow \frac{s}{t}=\frac{1+\dfrac{s}{\ell}}{1+\dfrac{d}{\ell}} \tag{12.1.8}$$

So finally (the first ratio ℓ/t for reminder is therefore the Moon radius on the Earth's radius and the second ratio s/t is the Sun radius on the Earth radius):

$$\frac{\ell}{t}=\frac{1+\dfrac{\ell}{s}}{1+\dfrac{d}{\ell}} \qquad \frac{s}{t}=\frac{1+\dfrac{s}{\ell}}{1+\dfrac{d}{\ell}} \tag{12.1.9}$$

These equations can be made to appear simpler by expressing the lengths d and s in terms of the Moon's radius ℓ as a unit, defining $\hat{d}=d/\ell$ and $\hat{s}=s/\ell$. Then:

$$\frac{\ell}{t}=\frac{1+\hat{s}}{\hat{s}(1+\hat{d})} \qquad \frac{s}{t}=\frac{1+\hat{s}}{1+\hat{d}} \tag{12.1.10}$$

Using $\hat{d}=2$ (a guess by looking Earth's shadow on the Moon during the lunar eclipse) and $\hat{s}=19.1$ (from the first result we got earlier above where $18<\hat{s}<20$):

$$\frac{\ell}{t}=\frac{1+\hat{s}}{\hat{s}(1+\hat{d})}\cong 0.35 \qquad \frac{s}{t}=\frac{1+\hat{s}}{1+\hat{d}}\cong 6.7 \tag{12.1.11}$$

More often presented as:

$$\frac{t}{\ell}=\frac{\hat{s}(1+\hat{d})}{1+\hat{s}}\cong 2.85 \qquad \frac{s}{t}=\frac{1+\hat{s}}{1+\hat{d}}\cong 6.7 \tag{12.1.12}$$

To compare with modern values where we have $t/\ell\cong 3.67$ (Earth's radius on Moon's radius) and $s/t\cong 109$ (Sun's radius on Earth's radius).

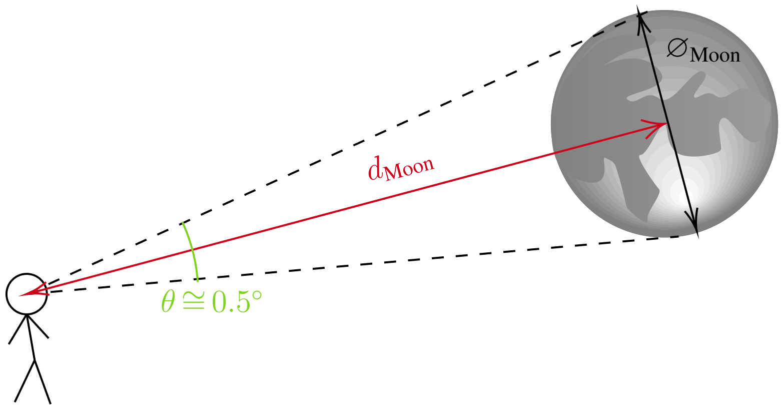
The error in this calculation comes primarily from the poor values for $\hat{s}$.

Remember that we know the Earth's radius using the method of Al-Bīrūnī (see page 2529) and also based on that of Eratosthenes (see page 2579), that was given by the approximate value of $6,266$ kilometres. Therefore we get for the diameter of the Moon using the results of Aristarchus:

$$\varnothing_{\text{Moon}}=\frac{2\cdot 6,266}{2.85}\cong 4,397\,[\text{km}] \tag{12.1.13}$$

where modern measurements give actually approximately $3,474.3$ kilometres.

Now let's go for the distances! We consider for that the observation of a full Moon (can easily be done by anyone) or of a Solar eclipse (indeed we are for the moment lucky - but it's temporary - that we can see that the Moon has quite the same angular diameter as that from the Sun):



Using basic trigonometry we get using the initial values of Aristarchus:

$$d_{\text{Moon}}\cong\frac{4,397/2}{\tan(0.5/2)}\cong 500,000\,[\text{km}] \tag{12.1.14}$$

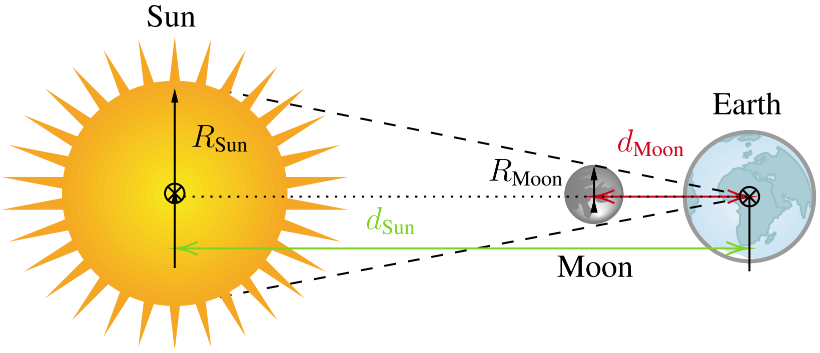
To compare with the modern average value of $384,400$ kilometers known with various methods like radar echoes or LASER ranging.

Now to deal with the case of the Sun, let us first thanks to the result above and our knowledge of the Earth radius, calculate the radius of the Sun (actually a G2V-type star, aka yellow dwarf) according to the quite (very) wrong values of Aristarchus:

$$\varnothing_{\text{Sun}}=(2\cdot 6,266)\cdot 6.7=83,964.4\,[\text{km}] \tag{12.1.15}$$

in comparison of the modern values that is $1,392,680$ kilometres (...). To see the various methods developed through the centuries to measure the average Earth-Sun distance the reader can refer to [371].

And now with all these values and using the following configuration typical of Solar eclipse (obviously not to scale) we can determine the corresponding quite (very) wrong value of the distance of the Sun from the Earth:



By similarity of the triangles (see section Euclidean Geometry page 2610):

$$\frac{R_{\text{Sun}}}{d_{\text{Sun}}}=\frac{R_{\text{Moon}}}{d_{\text{Moon}}}\Rightarrow d_{\text{Sun}}=\left(\frac{d_{\text{Moon}}}{R_{\text{Moon}}}\right)R_{\text{Sun}} \tag{12.1.16}$$

Hence:

$$d_{\text{Sun}}=\left(\frac{d_{\text{Moon}}}{R_{\text{Moon}}}\right)R_{\text{Sun}}=\frac{500,000}{4,397/2}\cdot(83,964.4/2)\cong 9,547,919\,[\text{km}] \tag{12.1.17}$$

To compare to the modern average value of approximately 149.6 millions kilometers (...)!

Remark

A similar procedure was later used by Hipparchus, who estimated the mean distance to the Moon as 67 Earth radii, and Ptolemy, who took 59 Earth radii for this value.

So that was quite funny to see how to calculate the most used dimensions in astronomy based on very old and simple methods. However, we will obviously here for the remaining part of this textbook use the values available in tables determined with modern tools and modern observational and statistical methods!

12.1.2 Drake Equation

This equation was invented (...) by Francis Drake in the 11960s (holocene calendar) with the intention to discuss the number of extraterrestrial civilizations in our Galaxy with which we might come in contact in the context of the SETI program (Search for ExtraTerrestrial Intelligence). The main purpose of this equation for scientists is to determine its factors, in order to know the likely number and (very) estimated extraterrestrial civilizations.

This empirical and naive equation which remains more something funny and provocative than something else... and whose principle can be applied to a lot of different areas of physics and life is written:

$$\boxed{N=N^*f_pn_ef_lf_if_cf_L} \tag{12.1.18}$$

The terms of this formula (because it is a formula and not a relation!) are defined as follows (the notation can differs as the reader can see it in the figure further below):

- N^* represents the number of stars in a single galaxy
- f_p is the fraction of stars that would have an orbiting planet (between 0 and 1)
- n_e is the number of planets per star that fulfil the conditions for the development of life
- f_l is the fraction of planets whose life has emerged (between 0 and 1)
- f_i is the fraction of those where an intelligent life has emerged (between 0 and 1)
- f_c is the fraction of f_i which has implemented radio communication technology (between 0 and 1)
- f_l is the fraction of time during which the fraction f_i civilizations will live (between 0 and 1)

In practice, it should be noted that this formula purpose is to try to determine an unknown amount from other amounts that are also unknown ... But it's a nice and funny formula to evaluate when you discuss with friends at the restaurant...

Researchers have suggested adding two new factors to the equation:

$$N=N^*f_pn_ef_lf_if_cf_Lf_of_t \tag{12.1.19}$$

where:

- f_o fraction of habitable planets with both continents and oceans
- f_t fraction of those planets with plate tectonics operating for at least 500 million years

There is therefore no guarantee that we are more knowledgeable after the estimate of this formula than before (method sometimes named in the literature "garbage in, garbage out"...).

The resulting value can motivate the fact that following mathematical developments are not only applicable to only one solar (star) system in the universe... maybe... (it would make a lot of useless empty space otherwise...).

Making a punctual estimates of a guess being obviously a nonsense we have to refer to Monte Carlo simulation (see section Numerical Methods pages pagerefmonte carlo simulations) as done in [186] if we want an estimate of N and results were:

Variable	**Mean**	**Standard Deviation**
$N_{\text{planets,total}}$	$4.770 \cdot 10^8$	381
$N_{\text{inhabited}}$	$80,090$	22
$N_{\text{fledgling}}$	728.6	1
$N_{\text{destroyed}}$	367.3	0
N_{advanced}	361.2	2

Table 12.1: **Statistics for the Rare Life Hypothesis (source: [186])**

Let us talk now about the "Fermi paradox" named after physicist Enrico Fermi, is the apparent contradiction between the lack of evidence and high probability estimates, e.g., those given by the Drake equation, for the existence of extraterrestrial civilizations. The basic points of the argument, made by physicists Enrico Fermi (11901–11954 according to holocene calendar) and Michael H. Hart (born in 11932 according to holocene calendar), are:

- There are billions of stars in the galaxy that are similar to the Sun, many of which are billions of years older than Earth.
- With high probability, some of these stars will have Earth-like planets, and if the Earth is typical, some might develop intelligent life.
- Some of these civilizations might develop interstellar travel, a step the Earth is investigating now.
- Even at the slow pace of currently envisioned interstellar travel, the Milky Way galaxy could be completely traversed in a few million years.

According to this line of reasoning, the Earth should have already been visited by extraterrestrial aliens. In an informal conversation, Fermi noted no convincing evidence of this, leading him to ask, "Where is everybody?" There have been many attempts to explain the Fermi paradox, primarily either suggesting that:

- Extraterrestrial life is rare or non-existent
- No other intelligent species have arisen
- Civilizations lack advanced technology
- It is the nature of intelligent life to destroy itself
- It is the nature of intelligent life to destroy others
- There is periodic extinction by natural events
- Intelligent civilizations are too far apart in space or time

- It is too expensive to spread physically throughout the galaxy
- Human beings have not existed long enough
- Humans are not listening properly
- Civilizations broadcast detectable radio signals only for a brief period of time
- Civilizations tend to isolate themselves
- Everyone is listening, no one is transmitting
- Earth is deliberately not contacted
- It is dangerous to communicate
- ...

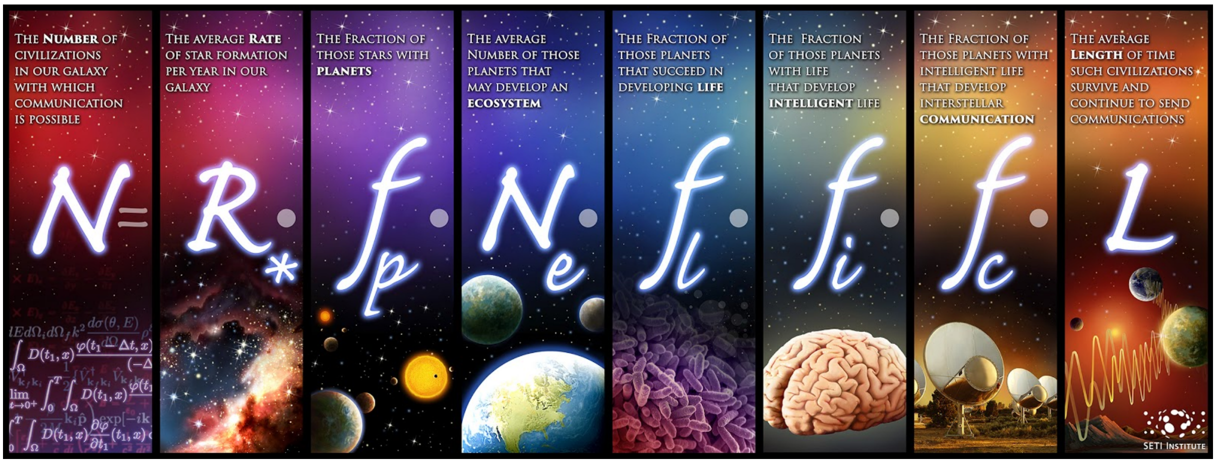

The Number of civilizations in our galaxy with which communication is possible
N =
The average Rate of star formation per year in our galaxy
R*
The fraction of those stars with planets
f p
The average number of those planets that may develop an ecosystem
N e
The fraction of those planets that succeed in developing life
f l
The fraction of those planets with life that develop intelligent life
f i
The fraction of those planets with intelligent life that develop interstellar communication
f c
The average length of time such civilizations survive and continue to send communications
L
SETI Institute

Figure 12.1: **Drake equation**

12.1.2.1 Habitale Zone (HZ)

In astronomy and astrobiology, the "circumstellar habitable zone" (CHZ), or simply the habitable zone, is the range of orbits around a star within which a planetary surface can support liquid water given sufficient atmospheric pressure. The bounds of the CHZ are based on Earth's position in the Solar System and the amount of radiant energy it receives from the Sun (actually a G2V-type star, aka yellow dwarf). Due to the importance of liquid water to Earth's biosphere, the nature of the CHZ and the objects within it may be instrumental in determining the scope and distribution of planets capable of supporting Earth-like extraterrestrial life and intelligence.

Since the concept was first presented in 11953 (holocene calendar), many stars have been confirmed to possess a CHZ planet, including some systems that consist of multiple CHZ planets. Most such planets, being either super-Earths or gas giants, are more massive than Earth, because massive planets are easier to detect. On November 4, 2013, astronomers reported, based on Kepler data, that there could be as many as 40 billion Earth-sized planets orbiting in the habitable zones of Sun-like stars and red dwarfs in the Milky Way. About 11 billion of these may be orbiting Sun-like stars. Proxima Centauri b, located about 4.2 light-years (1.3 parsecs) from Earth in the constellation of Centaurus, is the nearest known exoplanet, and is orbiting in the habitable zone of its star.

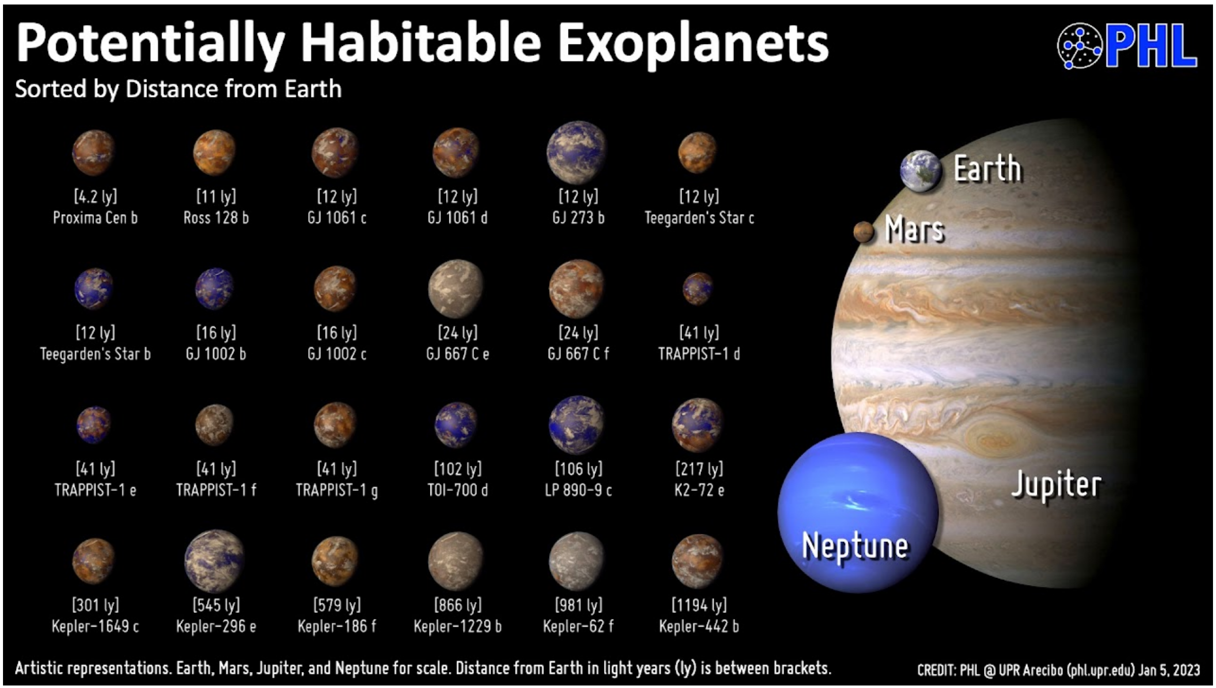

Potentially Habitable Exoplanets
Sorted by Distance from Earth
PHL
[4.2 ly] Proxima Cen b
[11 ly] Ross 128 b
[12 ly] GJ 1061 c
[12 ly] GJ 1061 d
[12 ly] GJ 273 b
[12 ly] Teegarden's Star c
[12 ly] Teegarden's Star b
[16 ly] GJ 1002 b
[16 ly] GJ 1002 c
[24 ly] GJ 667 C e
[24 ly] GJ 667 C f
[41 ly] TRAPPIST-1 d
[41 ly] TRAPPIST-1 e
[41 ly] TRAPPIST-1 f
[41 ly] TRAPPIST-1 g
[102 ly] TOI-700 d
[106 ly] LP 890-9 c
[217 ly] K2-72 e
[301 ly] Kepler-1649 c
[545 ly] Kepler-296 e
[579 ly] Kepler-186 f
[866 ly] Kepler-1229 b
[981 ly] Kepler-62 f
[1194 ly] Kepler-442 b
Earth
Mars
Jupiter
Neptune
Artistic representations. Earth, Mars, Jupiter, and Neptune for scale. Distance from Earth in light years (ly) is between brackets.
CREDIT: PHL @ UPR Arecibo (phl.upr.edu) Jan 5, 2023

Figure 12.2: A small sample of potentially habitable planets according to Kepler's date

The CHZ is also of particular interest to the emerging field of habitability of natural satellites, because planetary-mass moons in the CHZ might outnumber planets.

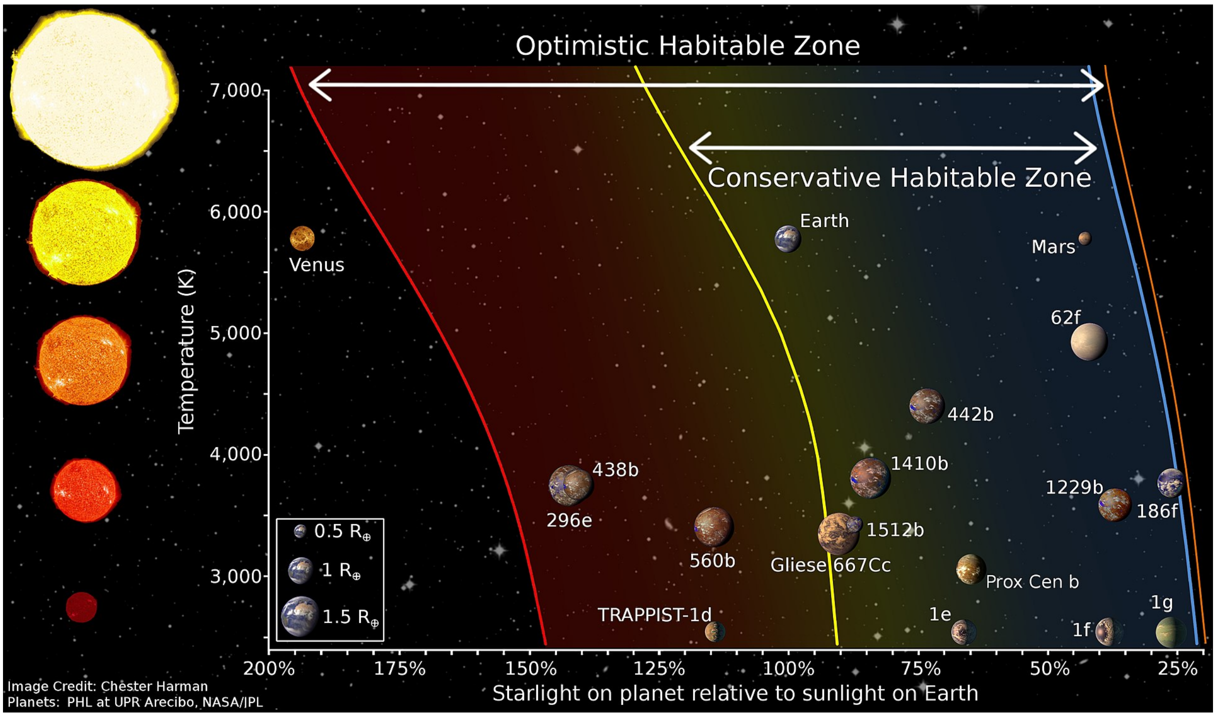


Figure 12.3: Diagram depicting the habitable zone boundaries around stars, and how the boundaries are affected by star type (author: Chester Harman)

In subsequent decades, the CHZ concept began to be challenged as a primary criterion for life, so the concept is still evolving. Since the discovery of evidence for extraterrestrial liquid water, substantial quantities of it are now thought to occur outside the circumstellar habitable zone. The concept of deep biospheres, like Earth's, that exist independently of stellar energy, are now generally accepted in astrobiology given the large amount of liquid water known to exist within in lithospheres and asthenospheres of the Solar System. Sustained by other energy sources, such as tidal heating or radioactive decay or pressurized by non-atmospheric means, liquid water may be found even on rogue planets, or their moons. Liquid water can also exist at a wider range of temperatures and pressures as a solution, for example with sodium chlorides in seawater on Earth, chlorides and sulphates on equatorial Mars, or ammoniates, due to its different colligative properties. In addition, other circumstellar zones, where non-water solvents favorable to hypothetical life based on alternative biochemistries could exist in liquid form at the surface, have been proposed.

Estimates for the habitable zone within the Solar System range from 0.38 to 10.0 astronomical units, though arriving at these estimates has been challenging for a variety of reasons. Numerous planetary mass objects orbit within, or close to, this range and as such receive sufficient sunlight to raise temperatures above the freezing point of water. However, their atmospheric conditions vary substantially.

Circumstellar habitable zones change over time with stellar evolution. For example, hot O-type stars, which may remain on the main sequence for fewer than 10 million years, would have rapidly changing habitable zones not conducive to the development of life. Red dwarf stars, on the other hand, which can live for hundreds of billions of years on the main sequence, would have planets with ample time for life to develop and evolve. Even while stars are on the main sequence, though, their energy output steadily increases, pushing their habitable zones farther out; our Sun, for example, was 75% as bright in the Archaean as it is now, and in the future, continued increases in energy output will put Earth outside the Sun's habitable zone, even before it reaches the red giant phase.

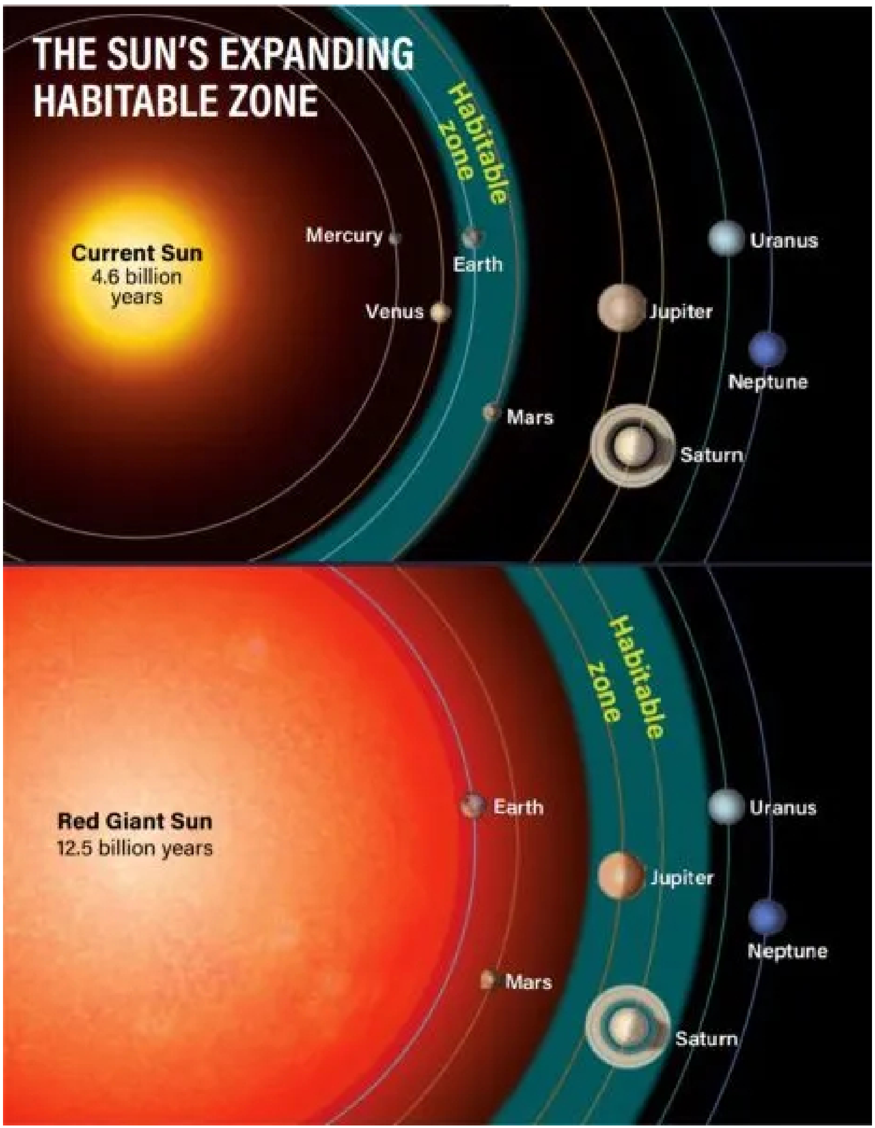

THE SUN'S EXPANDING HABITABLE ZONE
Habitable zone
Current Sun
4.6 billion years
Mercury
Earth
Venus
Mars
Jupiter
Uranus
Neptune
Saturn
Habitable zone
Red Giant Sun
12.5 billion years
Earth
Uranus
Jupiter
Neptune
Mars
Saturn

Figure 12.4: **Sun expanding habitable zone (author: Roen Kelly)**

Now let us see the maths for a very idealistic simple case restricted to the special case of the actual state of our Sun (with a Solar constant of $1,360.8\,[\mathrm{W}\cdot\mathrm{m}^{-2}]$) and an Earth (assumed to be here a perfect sphere) behaving like a black body and not having an internal source of heat and having its rotation axis perpendicular to the equatorial plane of the Sun. Let us calculate the closest and farthest distances from the Sun where an inert that rotates rapidly enough to guarantee the absence of gradient of temperature planet would have to allow water to exist in a liquid state. We will also Assume the albedo is 0.31 (i.e. 31% of light reflected) and that the atmosphere of the planet raises the surface temperature by $30\,[^\circ\mathrm{C}]$ (greenhouse effect as seen in the section of Weather Engineering page 8275) and that the only factor of interest is the state of water that has to be in liquid state in normal condition of pressure, i.e. with a temperature T between $0\,[^\circ\mathrm{C}]$ and $100\,[^\circ\mathrm{C}]$ (i.e. $273.15\,[\mathrm{K}]$ and $373.15\,[\mathrm{K}]$).

Let us remind the area of a sphere:

$$S=4\pi R^2 \tag{12.1.20}$$

The incident radiation from the Sun (actually a G2V-type star, aka yellow dwarf) can be calculated from its luminosity (which in turn could have been derived from its black-body flux as given by the Stefan-Boltzmann law; integrated over its surface):

$$M_\odot = \frac{L_\odot}{4\pi d^2} \tag{12.1.21}$$

with d the distance between the Sun and the Earth (i.e. $1\,[\text{AU}] = 149,597,870,700\,[\text{m}]$). The energy absorbed by the Earth, then becomes:

$$E_{\text{in}} = M_\odot \pi R^2_{\text{Earth}}\,(1-A) = \frac{L_\odot R^2_{\text{Earth}}\,(1-A)}{4d^2} \tag{12.1.22}$$

where $\rho = 0.31$ is the Earth's albedo, i.e. the fraction of the incoming flux that is not absorbed. The energy lost in black-body radiation, is given by:

$$E_{\text{out}} = 4\pi R^2_{\text{Earth}}\,\sigma T^4_{\text{Earth}} \tag{12.1.23}$$

with $\sigma T^4_{\text{Earth}}$ the blackbody flux as defined by the Stefan-Boltzmann law.

Equating these two, we get:

$$4\pi R^2_{\text{Earth}}\,\sigma T^4_{\text{Earth}} = \frac{L_\odot R^2_{\text{Earth}}\,(1-\rho)}{4D^2} \tag{12.1.24}$$

Hence:

$$d = \sqrt{\frac{L_\odot (1-\rho)}{16\pi\sigma T^4_{\text{Earth}}}} \tag{12.1.25}$$

or, after re-writing:

$$d = \sqrt{\frac{L_\odot (1-\rho)}{16\pi\sigma T^4_{\text{Earth}}}} = \sqrt{\frac{3.846\cdot 10^{26}\,[\text{W}]\cdot 0.69}{16\pi 5.67\cdot 10^{-8}\,[\text{W}\cdot\text{m}^{-2}\cdot\text{K}^{-4}]\cdot T^4_{\text{Earth}}}} = \frac{0.96\cdot 10^{16}\,[\text{mK}^2]}{T^2_{\text{Earth}}} \tag{12.1.26}$$

For $T = 273\,[\text{K}] - 30\,[\text{K}] = 245\,[\text{K}]$ (taking into account the greenhouse effect mentioned earlier that must be substracted from the temperature need from the stare alone), this becomes:

$$d_{\max} = 1.29\cdot 10^{11}\,[\text{m}] = 1.08\,[\text{AU}] \tag{12.1.27}$$

and for $T = 373\,[\text{K}] - 30\,[\text{K}] = 343\,[\text{K}]$, we get:

$$d_{\min} = 6.94\cdot 10^{10}\,[\text{m}] = 0.655\,[\text{AU}] \tag{12.1.28}$$

Hence a habitable zone range of:

$$d_{\max} - d_{\min} = 64,586,662.39\,[\text{km}] \tag{12.1.29}$$

Ok that's the CHZ range for liquid water! But now, for the fun, let us be more restrictive by taking the ideal external temperature range for mammals (i.e. $17\,[^\circ\text{C}]$ to $24\,[^\circ\text{C}]$). Therefore we have:

$$d_{\max} - d_{\min} = 0.949\,[\text{AU}] - 0.900\,[\text{AU}] = 7,348,702.56\,[\text{km}] \tag{12.1.30}$$

What is funny with that result is that Earth isn't even in the ideal range of distance from the Sun for mammals (at least for the moment as the Sun is increasing in size and its temperature increasing depending on the millions of years that pass...).

Obviously in the general case, the formula is:

$$\boxed{\Delta d_{\text{CHZ}} = \sqrt{\frac{L_{\text{star}}(1-\rho)}{16\pi\sigma}}\left(\frac{1}{T_{\text{min}}^4}-\frac{1}{T_{\text{max}}^4}\right)} \tag{12.1.31}$$

So as we proved it, there is no "perfect position", nor "ideal position", nor "perfect position" for a planet around a star but just a range of acceptable positions (that furthermore depends with the time of the evolution of the orbiting star)!

A final word! The model above is obviously naive as it doesn't take into account explicitly the rotation speed of the planet, the type of radiation of the main star, the constitution of the atmosphere, the precession of the planet, nor the fact that it orbit may be strongly elliptic, and so on... and so on!

12.1.3 Kepler's Laws

In astronomy, Kepler's laws describe naively and approximatively the main properties of the motion of planets[3] and other celestial objects around a main star (or some other exotic stuff different than stars!), without explaining the reason (at least at the time these laws were developed!). They were discovered by Johannes Kepler based on the observations and measurements (in phenomenal amount) of the position of the planets made by Tycho Brahe, measures that were very accurate for its time.

Definition (#12.1.524): In astronomy an "orbit" is a closed curved path of a celestial object or spacecraft round a star, planet, or moon, especially a periodic elliptical revolution. In reality the planets and our Sun (and very likely most objects in our observable Universe) don't follow an orbit but an helicoidal path (as viewed from the center of our Galaxy) that projected on a plane (heliocentric referential) is in approximation similar to an orbit!

The first two Kepler's law seems to were published in 11609 (holocene calendar) and the third in 11618 (holocene calendar). The elliptical orbits, as set out in its first two laws can explain a part of the complexity of the apparent motion of the planets (but in reality the motion is chaotic and we need to rely on numerical Fourier series that need to be regularly corrected).

Soon after, in 11687 (holocene calendar) Isaac Newton discovered the law of gravitational attraction, deducting from it, by calculation, the three Kepler's laws.

Remark

Strictly speaking, pure Keplerian orbits can only occur with two bodies obeying Newton's law of gravitation. If there are more than two bodies then the orbits will be perturbed, to some degree. Kepler orbits are a reasonable first approximation to the orbits of the planets in the Solar System because the Sun is so massive compared to the planets, and the distances between the planets are relatively large, so the planets don't perturb each other's orbits very much.

Kepler was fortunate that he had sufficiently accurate data to see the patterns that led to his three laws of planetary motion, but not so accurate that he could see those laws were only a rough approximation. Fortunately, Newton came along a few decades later and showed how Kepler's laws were a consequence of the laws of mechanics and gravitation applied to two bodies. However, Newton also realised that the Moon's orbit around the Earth was significantly perturbed by the Sun. In fact, the Moon's motion is quite complex and its orbit precesses in various ways on rather short time scales.

We will now try to present these laws in the most relevant possible way:

---

[3] The definition of a "planet" is still a debate in the 121st century (holocene calendar) in a system in equilibrium... The reader can try to find a himself what should be a good definition for a "planet"and such an exercise will provide enough evidence that's not an easy task.

12.1.3.1 First Kepler's Law (conicity law)

The "first law of Kepler", sometimes also named "conicity law" or "law of orbits" is stated most of time as follow: The orbits of the planets are conics (ellipses) which the Sun (central star) occupies one of the focals.

In fact, it should be noted that this is not really a "law" in the proper sense, since further below you will see that we can prove that:

$$\boxed{r=\frac{p}{1+e\cos(\theta)}=\frac{b^2/a}{1+\sqrt{1-\dfrac{b^2}{a^2}}\cos(\theta)}} \tag{12.1.32}$$

where a is the semi-major axis and b is the semi-minor axis and $0\leq e\leq 1$ is the eccentricity[4] (see section Analytical Geometry page 2775).

Remark

The reader who has already read the section of Analytical Geometry will not be surprised by this relation (see page 2785)...

12.1.3.2 Second Kepler's Law (area law)

The "Kepler's second law", sometimes also named "area law" tells us that the line joining a planet to the Sun (central start) sweeps out equal areas in equal times (constant areal velocity) as:

$$\boxed{\frac{\Delta S}{\Delta t}=c^{te}} \tag{12.1.33}$$

It is a relation that arises from the conservation of angular momentum as we have already shown it in the section of Classical Mechanics where we got (see page 3244):

$$\frac{\mathrm{d}\vec{S}}{\mathrm{d}t}=\frac{\vec{b}}{2m}=\frac{\vec{r}\times(m\vec{v})}{2m}=c^{te} \tag{12.1.34}$$

So again, the status of "law" is questionable in the language of modern physics!

The latter relation is also in scalar form obviously written:

$$rv\sin(\theta)=c^{te} \tag{12.1.35}$$

where θ is the angle between the orbit radius and the tangential velocity vector of the object on its orbit.

---

[4] The eccentricity of the Earth's orbit varies over time, due to gravitational interaction with other celestial bodies. The eccentricity value can get as low as 0.0034 and up to 0.058. Because of the eccentricity the orbit of the Earth from the farthest (aphelion) to nearest (perihelion) point of the Sun its distance has a differential of almost 3 millions kilometers actually.

Now let us express this law in another form more conventional in the field of astronomy. Consider for this the movement in the plane the cylindrical coordinates given by:

$$\vec{x}(t)=r(t)\begin{pmatrix}\cos(\theta(t))\\ \sin(\theta(t))\\ 0\end{pmatrix} \tag{12.1.36}$$

Therefore:

$$\dot{\vec{x}}(t)=\dot{r}(t)\begin{pmatrix}\cos(\theta(t))\\ \sin(\theta(t))\\ 0\end{pmatrix}+r(t)\dot{\theta}(t)\begin{pmatrix}-\sin(\theta(t))\\ \cos(\theta(t))\\ 0\end{pmatrix} \tag{12.1.37}$$

It comes therefore from the property of linearity of the vector product:

$$\begin{aligned}\vec{b}&=m\vec{x}(t)\times\dot{\vec{x}}(t)=mr(t)\begin{pmatrix}\cos(\theta(t))\\ \sin(\theta(t))\\ 0\end{pmatrix}\times\left[\dot{r}(t)\begin{pmatrix}\cos(\theta(t))\\ \sin(\theta(t))\\ 0\end{pmatrix}+r(t)\dot{\theta}(t)\begin{pmatrix}-\sin(\theta(t))\\ \cos(\theta(t))\\ 0\end{pmatrix}\right]\\ &=mr(t)\left[\dot{r}(t)\begin{pmatrix}\cos(\theta(t))\\ \sin(\theta(t))\\ 0\end{pmatrix}\times\begin{pmatrix}\cos(\theta(t))\\ \sin(\theta(t))\\ 0\end{pmatrix}+r(t)\dot{\theta}(t)\begin{pmatrix}\cos(\theta(t))\\ \sin(\theta(t))\\ 0\end{pmatrix}\begin{pmatrix}-\sin(\theta(t))\\ \cos(\theta(t))\\ 0\end{pmatrix}\right]\\ &=mr(t)\left[\dot{r}(t)\begin{pmatrix}0\\ 0\\ \cos(\theta(t))\sin(\theta(t))-\sin(\theta(t))\cos(\theta(t))\end{pmatrix}+r(t)\dot{\theta}(t)\begin{pmatrix}0\\ 0\\ cos^2(\theta(t))+\sin^2(\theta(t))\end{pmatrix}\right]\\ &=mr(t)\left[\dot{r}(t)\begin{pmatrix}0\\ 0\\ 0\end{pmatrix}+r(t)\theta(\dot{t})\begin{pmatrix}0\\ 0\\ 1\end{pmatrix}\right]=mr(t)r(t)\dot{\theta}(t)\begin{pmatrix}0\\ 0\\ 1\end{pmatrix}\end{aligned} \tag{12.1.38}$$

Therefore taking the norm:

$$\|\vec{b}\|=mr^2(t)\dot{\theta}(t)=mr^2(t)\dfrac{\mathrm{d}\theta}{\mathrm{d}t} \tag{12.1.39}$$

And since it is equal to a constant, it is often customary to write this last equality in a condensed form (and putting the mass in the constant):

$$r^2(t)\dfrac{\mathrm{d}\theta}{\mathrm{d}t}=c^{te} \tag{12.1.40}$$

Also, remember that we also got the result that the movement is and remains in a plane without any outside action!

We note also that this law gives us the speed of the planet is variable on its orbit. It is larger at perihelion than at the aphelion:

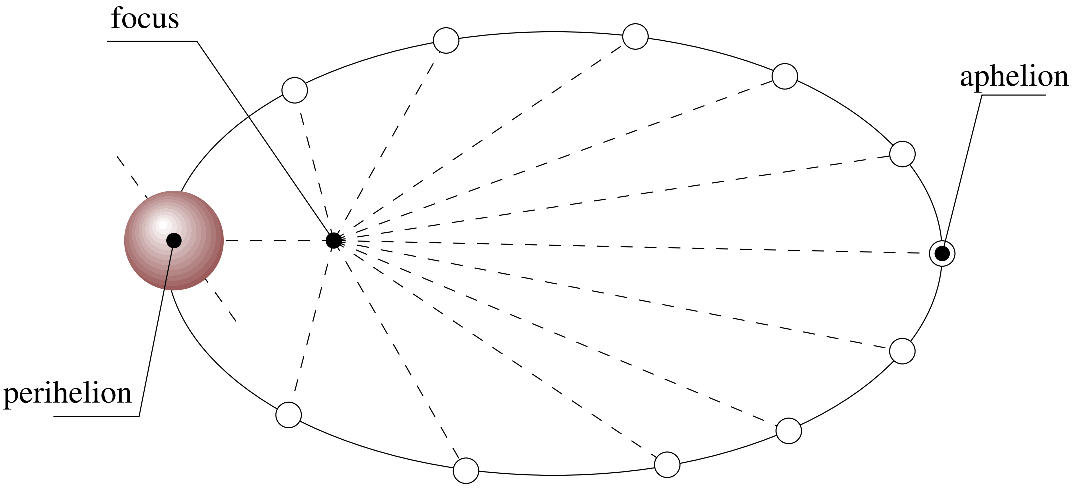
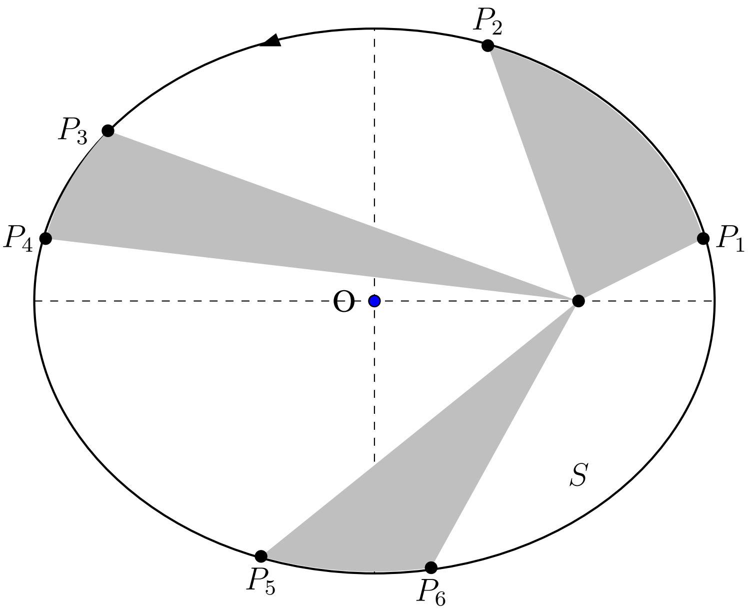


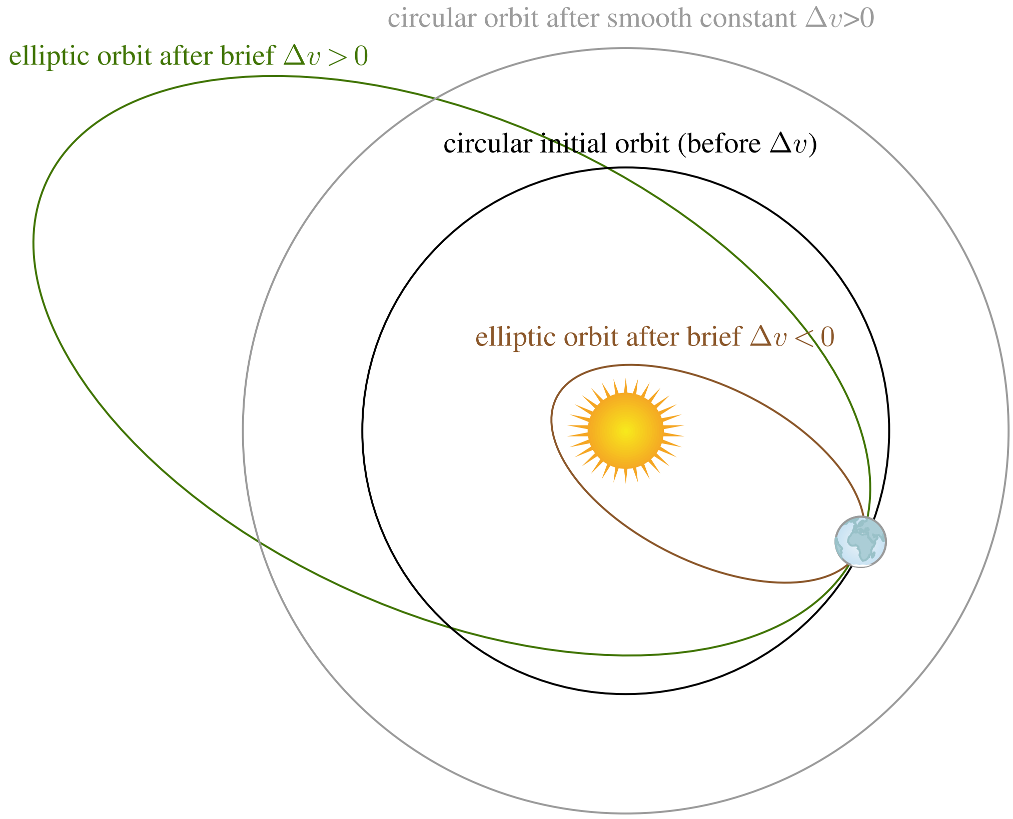
Figure 12.5: **Representation of surfaces conservation**

This is true for the Earth for example (for modern measurements of Earth's speed orbit and variation in its orbital speed see [27]). Indeed, this latter is closer to the Sun in winter (for northern hemisphere) and then has a trajectory speed slightly higher than in summer; the travel time is therefore lower (winter has fewer days than the other seasons).

The above figure is often represented in textbooks as:



The second law is the best law to understand what happens in the case where you have an object in equilibrium in orbit and you force quite brutally and temporary a Δv! Indeed, if you imagine an object in a perfect circular orbit and decelerate it quite brutally (with boosters for example), then to sill respect the area law the orbit of the object will have to change from a circular orbit to an elliptic one (but by staying tangential to the point where the Δv has occurred) as illustrated below:


circular orbit after smooth constant Δv>0
elliptic orbit after brief $\Delta v > 0$
circular initial orbit (before Δv)
elliptic orbit after brief $\Delta v < 0$

However the reader must keep in mind that as we will prove it further below during our study of orbital speed at page 5282, at equilibrium the speed is inversely proportional to the square root of the orbital radius (hence the further away an object orbits, the slower it is!):

$$v_{\text{orb.}} = \sqrt{\frac{GM}{R}} \tag{12.1.41}$$

For more about orbits transfer the reader can refer to the section of Aerospace Engineering at page 8633.

12.1.3.2.1 Acceleration variation on an elliptic orbit

The Earth's angular momentum, with respect to the focus of the ellipse, is :

$$b = Mr^2\dot{\phi} \tag{12.1.42}$$

where m, r and $\dot{\phi}$ are the mass, the distance to the center of the Sun and the angular velocity, respectively. Therefore:

$$Mr_p^2\dot{\phi}_p = Mr_a^2\dot{\phi}_a \tag{12.1.43}$$

where the indices p and a denotes "perihelion" and "aphelion". For an ellipse (see section Analytical Geometry page 2775):

$$r_p = \frac{r_0}{1+e} \qquad r_a = \frac{r_0}{1-e} \tag{12.1.44}$$

Hence:

$$\frac{\dot{\phi}_p}{\dot{\phi}_a}=\left(\frac{1+e}{1-e}\right)\cong 1.0340 \tag{12.1.45}$$

since the Earth's orbit eccentricity is, $e\cong 0.0167$. We have a change of about three per cent in six months. The mean angular acceleration is therefore:

$$\bar{\alpha}\cong\frac{0.0340\cdot\phi_a}{180\cdot 24\cdot 60\cdot 60}\cong 10^{-9}\phi_a\,[\text{rad}\cdot\text{s}^{-2}] \tag{12.1.46}$$

Notice that $\dot{\phi}_a$ is of the order of:

$$\frac{2\pi}{365\cdot 24\cdot 60\cdot 60}\cong 10^{-7}\,[\text{rad}\cdot\text{s}^{-1}] \tag{12.1.47}$$

So the angular acceleration is of the order of $10^{-16}\,[\text{rad}\cdot\text{s}^{-2}]$. This is too small to be sensible for humans and also for actual accelerometers.

By the way, if we multiply this value by the actual mean distance to the Sun, $r\cong 10^{11}\,[\text{m}]$, we get an acceleration of order $10^{-6}\,[\text{m}\cdot\text{s}^{-2}]$. That is negligible compared to the acceleration due to the Earth's gravity, $9.81\,[\text{m}\cdot\text{s}^{-2}]$.

12.1.3.2.2 Time of flight

We propose now to apply the second Kepler's law to determine the time t from the passage to the perihelion as a function of the "eccentric anomaly" φ in the case of an elliptic orbit (thus special case!) using its two foci (one of them being assimilable for example to the position of the Sun) and the origin of the "auxiliary circle" (also named "apsidal circle"):

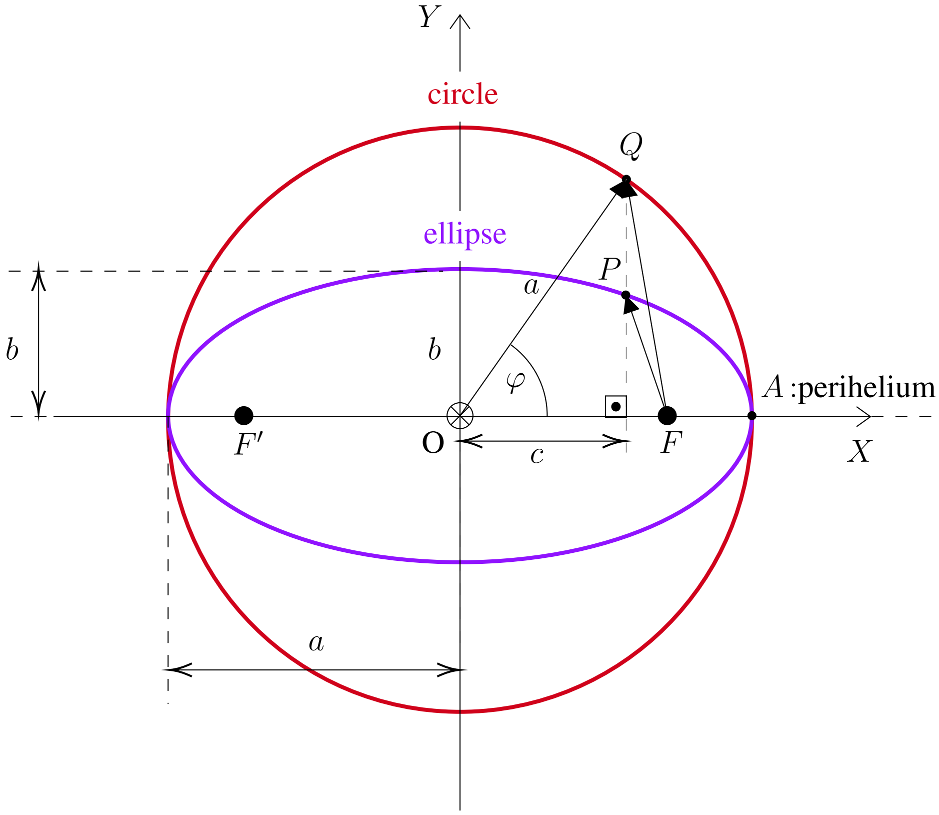


Figure 12.6: **Schema for the study of eccentric anomaly angle**

To determine the time t between the passage at the perihelion A and the point P of a body following the trajectory of the ellipse of surface πab (see the section Geometric Shapes at page 2920 for the proof of the calculation of the surface of the ellipse) in function of the eccentric anomaly φ, we will use the areas law just proved earlier above (second Kepler's law) that gives us the right to write:

$$\frac{S_{AFP}}{t_{\text{flight}}}=\frac{\pi ab}{T} \tag{12.1.48}$$

But, the surface of the ellipse is an affine transformation of the surface of the auxiliary circle such that:

$$S_{\text{ellipse}}=\frac{b}{a}S_{\text{circle}}=\frac{b}{a}(\pi a^2)=\pi ab \tag{12.1.49}$$

We have then:

$$\frac{S_{AFP}}{t_{\text{flight}}}=\frac{\frac{b}{a}S_{AFQ}}{t_{\text{flight}}}=\frac{\frac{b}{a}\left[S_{AOQ}-S_{FOQ}\right]}{t_{\text{flight}}} \tag{12.1.50}$$

If φ is, as it should, expressed in radians, we have of course:

$$S_{AOQ}=a^2\frac{\varphi}{2} \tag{12.1.51}$$

Therefore:

$$\frac{S_{AFP}}{t_{\text{flight}}}=\frac{\frac{b}{a}S_{AFQ}}{t_{\text{flight}}}=\frac{\frac{b}{a}\left[a^2\frac{\varphi}{2}-S_{FOQ}\right]}{t_{\text{flight}}} \tag{12.1.52}$$

For S_{FOQ}, we have $\overline{FO}=a$ therefore equal to the radius of the auxiliary circle. The surface of the triangle S_{FOQ}, knowing its height given by $h=a\sin(\varphi)$ is then obtained by:

$$S_{FOQ}=\frac{1}{2}ch=\frac{1}{2}ca\sin(\varphi) \tag{12.1.53}$$

Therefore we have:

$$\frac{S_{AFP}}{t_{\text{flight}}}=\frac{\frac{b}{a}S_{AFQ}}{t_{\text{flight}}}=\frac{\frac{b}{a}\left[a^2\frac{\varphi}{2}-\frac{1}{2}ca\sin(\varphi)\right]}{t_{\text{flight}}}=\frac{\frac{ab}{2}\left[\varphi-\frac{c}{a}\sin(\varphi)\right]}{t_{\text{flight}}} \tag{12.1.54}$$

But, by definition of the eccentricity (see section Analytical Geometry page 2775), we can write:

$$\frac{S_{AFP}}{t_{\text{flight}}}=\frac{\frac{ab}{2}\left[\varphi-e\sin(\varphi)\right]}{t_{\text{flight}}} \tag{12.1.55}$$

Finally, we have:

$$\frac{\frac{ab}{2}\left[\varphi-e\sin(\varphi)\right]}{t_{\text{flight}}}=\frac{\pi ab}{T} \tag{12.1.56}$$

Therefore:

$$\boxed{t_{\text{flight}}=\frac{\varphi-e\sin(\varphi)}{2\pi}T} \tag{12.1.57}$$

where the angle taken at the center of the ellipse is for recall named the "eccentric anomaly".

Definition (#12.1.525): In the description of the Keplerian orbit of a celestial object, the "eccentric anomaly" is the angle between the direction of the periapse and the current position of an object in its orbit, projected on the circle extinct perpendicular to the major axis of the ellipse

What would interest us now would be to find a relation of passage between this eccentric anomaly and the angle named "true anomaly" θ as sometimes it is often more advantageous to use this last angle.

For a relation between the eccentric anomaly φ and the true anomaly θ, we will reuse our above schema but modified a bit:

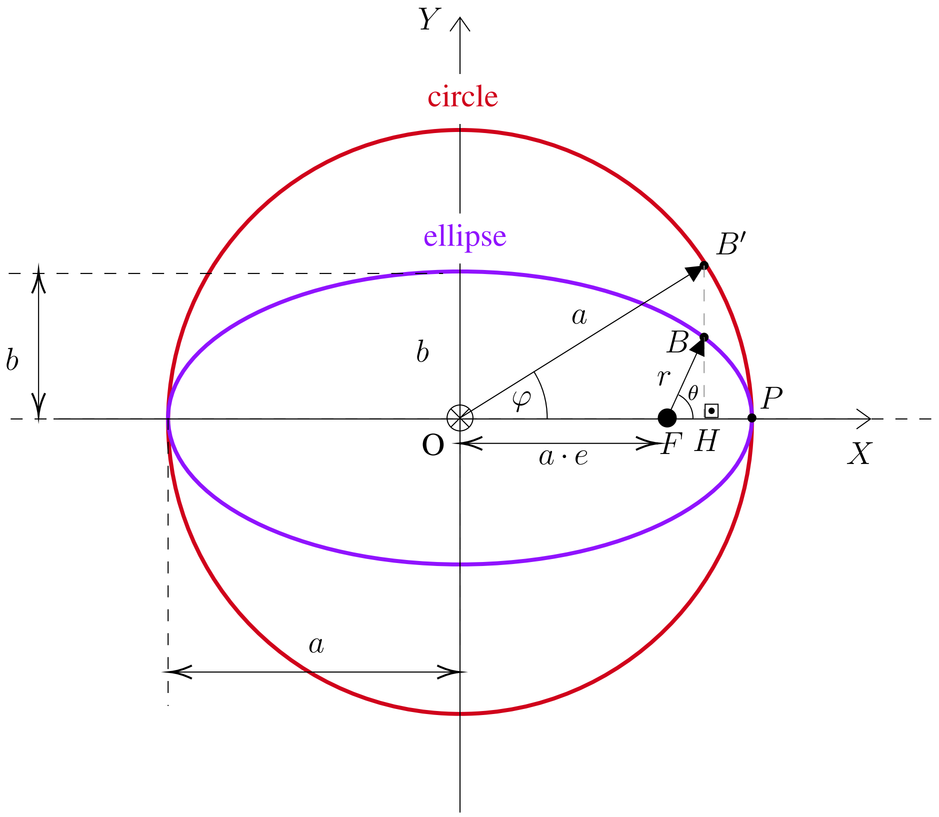


Figure 12.7: **Schema for the study of true-eccentric anomaly angle relation**

We have obviously the 4 below relations which a deduce from the above figure:

$$\begin{aligned}r\cos(\theta)&=\overline{FH}\\ r\sin(\theta)&=\overline{HB}\\ a\cos(\varphi)&=\overline{\text{O}H}\\ a\sin(\varphi)&=\overline{HB'}\end{aligned} \tag{12.1.58}$$

We then have already in a first time (relation which will be useful to us a little later):

$$r\cos(\theta)=\overline{FH}=\overline{\text{O}H}-\overline{\text{O}F}=a\cos(\varphi)-ae \tag{12.1.59}$$

We have also proved in the section of Analytical Geometry that (see page 2779):

$$\frac{b}{a}=\sqrt{1-e^2} \tag{12.1.60}$$

this relation being valid at any border point of the ellipse. Thus, we also have:

$$\frac{\overline{HB}}{\overline{HB'}}=\sqrt{1-e^2}=\frac{r\sin(\theta)}{a\sin(\varphi)} \tag{12.1.61}$$

Which leads us to write:

$$\sin(\theta)=\frac{a\sin(\varphi)\sqrt{1-e^2}}{r} \tag{12.1.62}$$

Ideally, we could get rid of the radius in the denominator. For this, we will use the fact that (relations that we have just proved earlier above):

$$\begin{aligned}r\sin(\theta)&=a\sin(\varphi)\sqrt{1-e^2}\\ r\cos(\theta)&=a\cos(\varphi)-ae\end{aligned} \tag{12.1.63}$$

We have:

$$r^2\sin^2(\theta)+r^2\cos^2(\theta)=a^2\sin^2(\varphi)\left(\sqrt{1-e^2}\right)^2+(a\cos(\varphi)-ae)^2 \tag{12.1.64}$$

The terms to the left of the equality are simplified immediately:

$$\begin{aligned}r^2&=a^2\sin^2(\varphi)\left(\sqrt{1-e^2}\right)^2+(a\cos(\varphi)-ae)^2\\ &=a^2\left[\sin^2(\varphi)(1-e^2)+\cos^2(\varphi)-2e\cos(\varphi)+e^2\right]\\ &=a^2\left[\sin^2(\varphi)-e^2\sin^2(\varphi)+\cos^2(\varphi)-2e\cos(\varphi)+e^2\right]\\ &=a^2\left[1-e^2\sin^2(\varphi)-2e\cos(\varphi)+e^2\right]\\ &=a^2\left[1-e^2(1-\cos^2(\varphi)-2e\cos(\varphi)+e^2\right]\\ &=a^2\left(1+e^2\cos^2(\varphi)-2e\cos(\varphi)\right)=a^2(1-e\cos(\varphi))^2\end{aligned} \tag{12.1.65}$$

hence:

$$r=a(1-e\cos(\varphi)) \tag{12.1.66}$$

therefore:

$$\sin(\theta)=\frac{a\sin(\varphi)\sqrt{1-e^2}}{a(1-e\cos(\varphi))}=\frac{\sin(\varphi)\sqrt{1-e^2}}{1-e\cos(\varphi)} \tag{12.1.67}$$

thus finally:

$$\boxed{\theta=\arcsin\left(\frac{\sin(\varphi)\sqrt{1-e^2}}{1-e\cos(\varphi)}\right)} \tag{12.1.68}$$

Finally notice that in the special case of an elliptical orbit, we deduce thanks to the second Kepler's law (see the proof of the calculation of an area of an ellipse in the section Geometric Shapes page 2917) the:

12.1.3.3 Third Kepler's Law (periods' law)

The "Kepler's third law", sometimes also named "Periods' law" or "Kepler's harmonic law", is stated as follow: The squares of the periods of revolution T are proportional to the cube of the semi-major axes of the orbits D:

$$T^2=c^{te}D^3 \Leftrightarrow \frac{T^2}{D^3}=c^{te} \tag{12.1.69}$$

The last ratio is then a constant in practice for all planets (in physics we also speak of "invariant") and the reason for this "harmony" was difficult to explain before Newton's theory.

Again, we will see later that the status of "law" is no longer justified in our time as it is possible to prove that this relation, whose expression will be detailed, is in reality:

$$\frac{R^3}{T^2}=\frac{Gm}{4\pi^2}=c^{te} \tag{12.1.70}$$

and therefore we understand better when we see the term on the right why we had the previous constant (m is the central body mass!).

The last relation is more often written as following in the field of astronomy:

$$\boxed{\frac{a^3}{T^2}=\frac{GM}{4\pi^2}=c^{te}} \tag{12.1.71}$$

where as we have seen in the section of Analytical Geometry, a is the traditional notation for the semi major axis of ellipse. It is important to notice that whatever the magnitude of the semi minor axis, if the semi major axis of multiple ellipses are identical, their periods (and therefore the corresponding total energy) will be identical.

The most commonly used rearrangement of this last relation is obviously:

$$T=\sqrt{\frac{a^3 4\pi^2}{GM}}=2\pi\sqrt{\frac{a^3}{GM}} \tag{12.1.72}$$

Of course, Kepler did not immediately published his three laws in this provocative simplicity. Their current presentation order is also not the original one ... They are in reality to find among a profusion of physical speculations and reflections on world's harmony.

Remark

The Kepler's law are not limit to the gravitation force. They also apply for all acceleration (or force) of the type $1/r^2$. And this is also the case of the Coulomb's law (see section Electrostatics page 3968). Kepler's law can therefore also be applied to an electron in orbit around a nucleus. The Wilson and Sommerfeld model (see section Quantum Corpuscular Physics page 4644) based also on Kepler's law gives also elliptic trajectories for electrons!

Here is an high definition image (you can zoom in a lot!) with some celestial bodies of the Solar System, and for many of them, their orbital period T and their distance a to their main attractor:

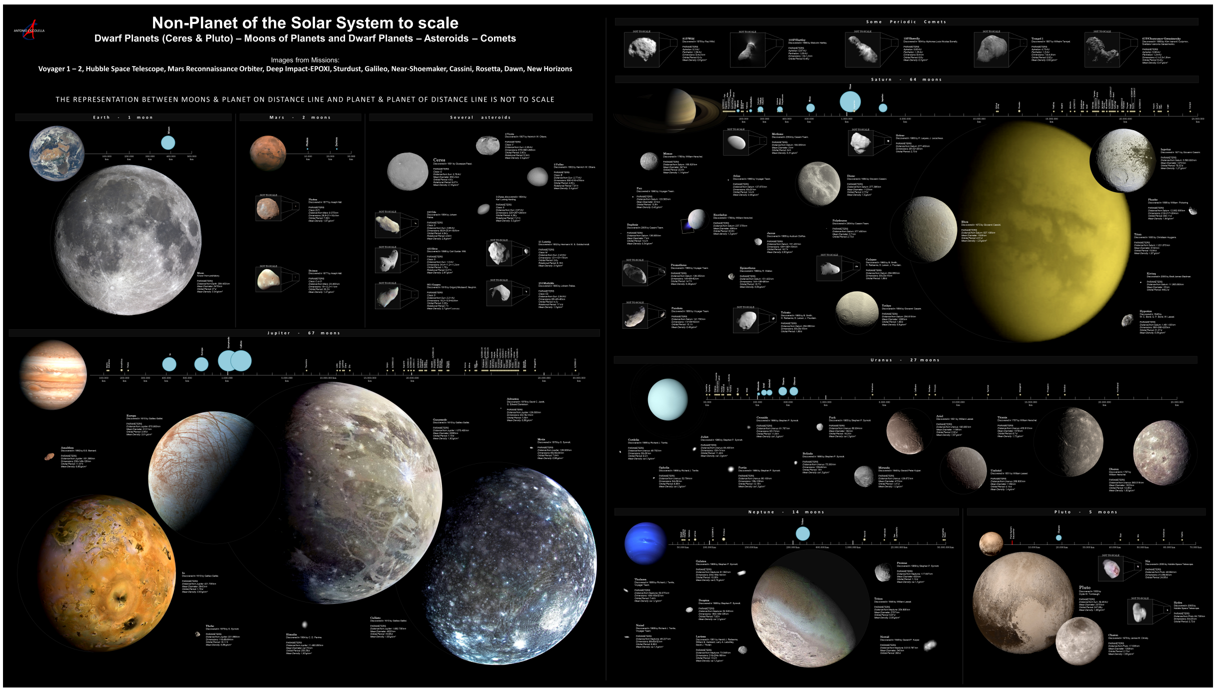

Non-Planet of the Solar System to scale
Dwarf Planets (Ceres & Pluto) – Moons of Planets and Dwarf Planets – Asteroids – Comets
Images from Missions:
Voyager 1 – 2, Hubble Space Telescope, Mars Reconnaissance Orbiter, Deep Impact-EPOXI, Sturdust, Galileo, Near-Shoemaker, Cassini, Rosetta, Dawn, New Horizons
THE REPRESENTATION BETWEEN MOONS & PLANET ON DISTANCE LINE AND PLANET & PLANET OF DISTANCE LINE IS NOT TO SCALE
Earth - 1 moon
Mars - 2 moons
Several asteroids
Jupiter - 67 moons
Some Periodic Comets
Saturn - 64 moons
Uranus - 27 moons
Neptune - 14 moons
Pluto - 5 moons

Figure 12.8: **Some celestial bodies of our Solar System (author: Antonio Ciccolella)**

We also recommend the reader to carefully look to the orbits structures schematized on the above figure as it will be useful for a critical thinking of the Newton Quantum Gravity model that we will introduce later (page 6074).

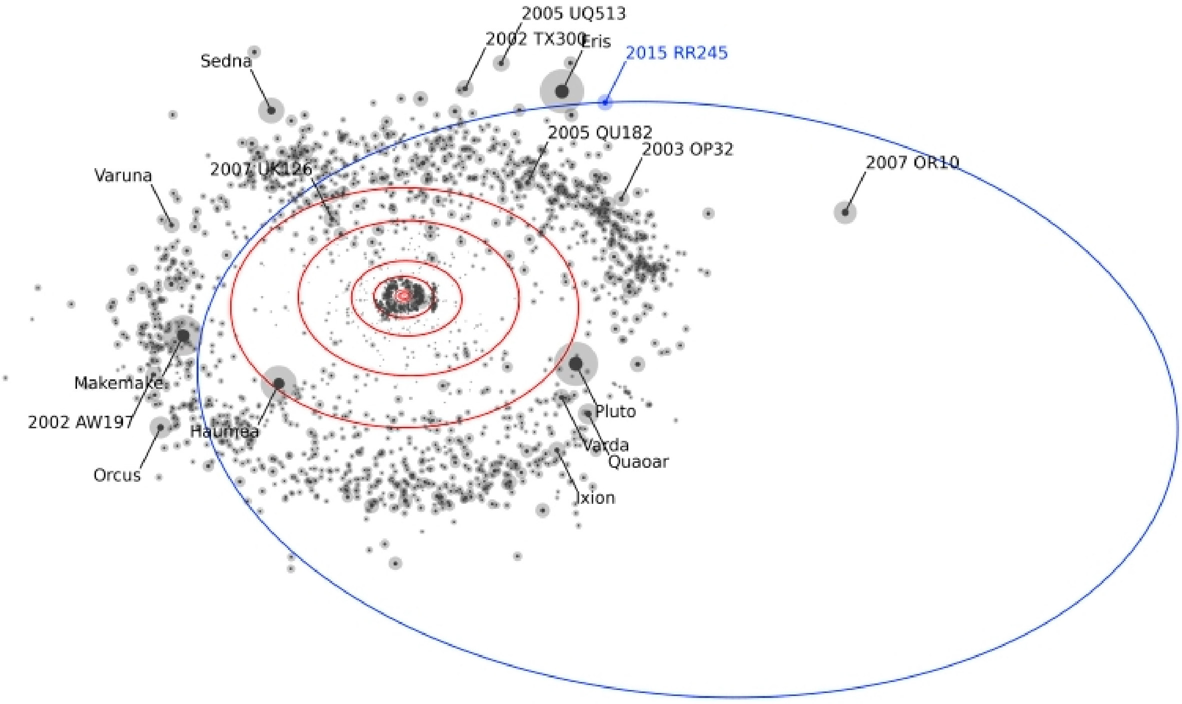

2005 UQ513
2002 TX300
Eris
2015 RR245
Sedna
2005 QU182
2003 OP32
2007 OR10
Varuna
2007 UK126
Makemake
2002 AW197
Haumea
Orcus
Pluto
Varda
Quaoar
Ixion

Figure 12.9: **More realistic schema of the Solar System as known today in comparison to the one presented at College (author: Alex Parker/OSSOS)**

An interesting complementary image is to put side by side some well known planets and dwarf planets by respecting the proportions:

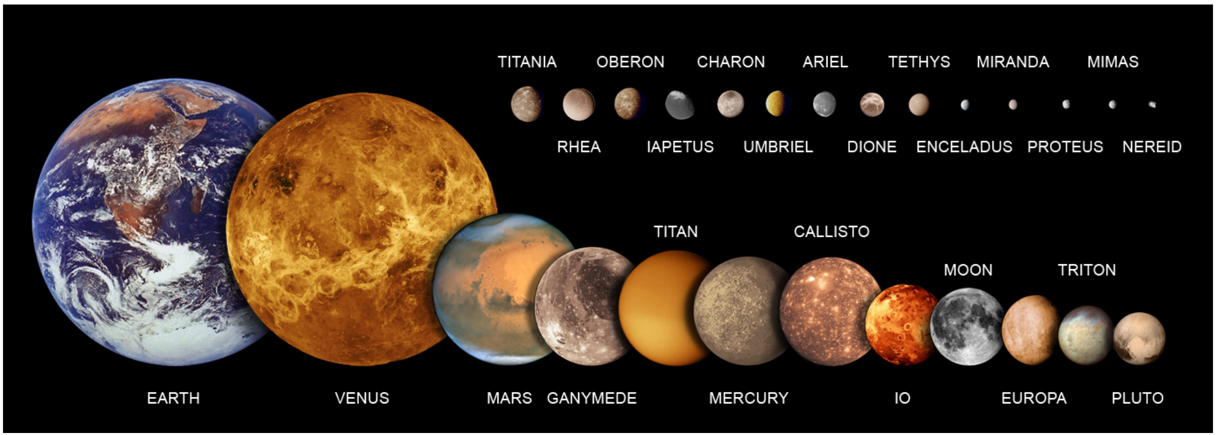
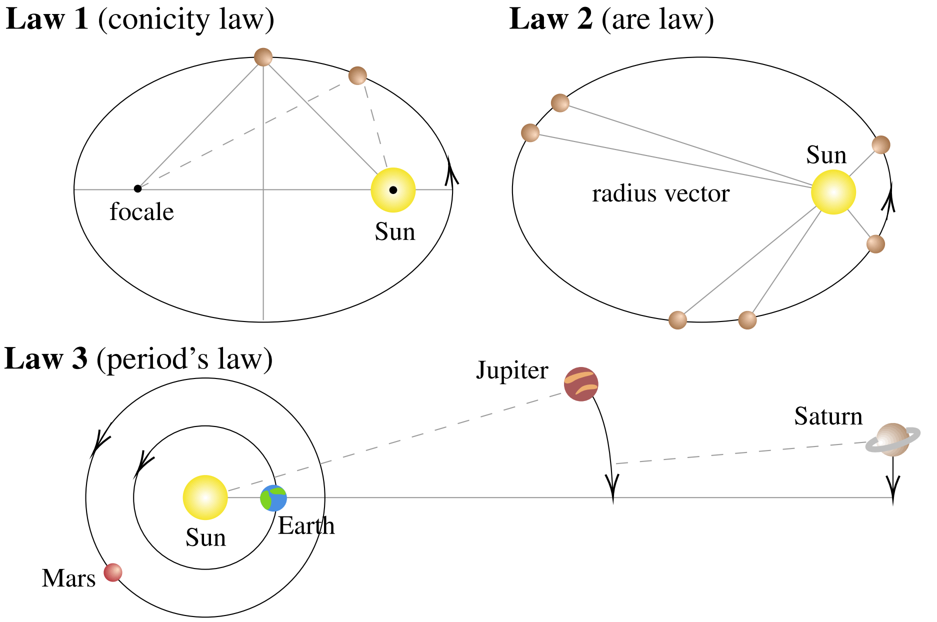

TITANIA
OBERON
CHARON
ARIEL
TETHYS
MIRANDA
MIMAS
RHEA
IAPETUS
UMBRIEL
DIONE
ENCELADUS
PROTEUS
NEREID
TITAN
CALLISTO
MOON
TRITON
EARTH
VENUS
MARS
GANYMEDE
MERCURY
IO
EUROPA
PLUTO

Figure 12.10: **Side by side planets and dwarf planets (source: ?)**

For information, the rotation of Earth's on itself slows down at the day we write these lines of approximately 2.3 milliseconds per century. That means 1 hour every 180 millions years. Each year the Earth's orbit speed around the Sun (actually a G2V-type star, aka yellow dwarf) also slows down of approximately 3 nanometers-per-second. Using Kepler's law, that means that the Earth goes further away from the Sun of 3 micrometers per 1 billion year......!!!!

The three Kepler's law can then be summarized by the following small figure:


Law 1 (conicity law)
focale
Sun
Law 2 (are law)
Sun
radius vector
Law 3 (period's law)
Jupiter
Saturn
Sun
Earth
Mars

$$\frac{P^2}{a^2}=\overset{\text{Earth}}{\frac{(1)^2}{(1)^2}}=\overset{\text{Mars}}{\frac{(1.88081)^2}{(1.52369)^2}}=\overset{\text{Jupiter}}{\frac{(11.86179)^2}{(5.2028)^2}}=\overset{\text{Saturn}}{\frac{(29.4556)^2}{(9.53884)^2}}$$

Figure 12.11: **Summary of Kepler's laws illustrated**

The reader must take precautions with the image above because:

1. The planets are most of time in a movement that is not in the same plane. For the Solar System it is the tradition at high school level to represent the planets in the "ecliptic" plane that is the average plane described by the movement of Jupiter around the Sun:

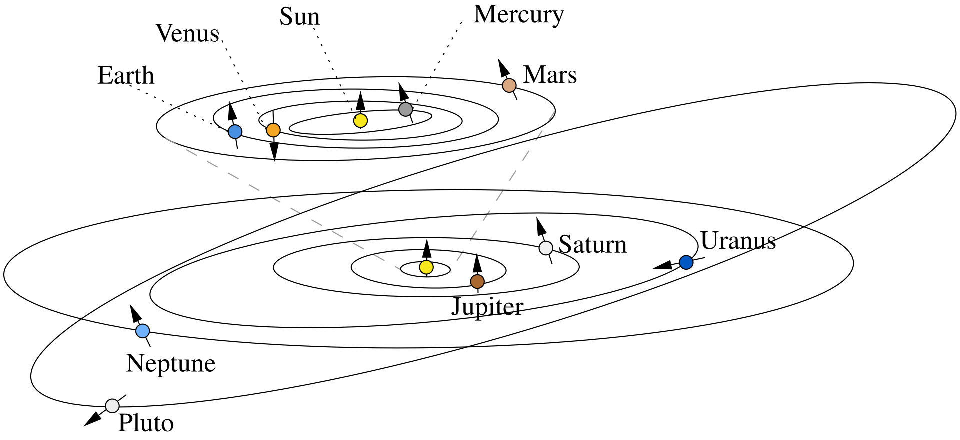


Figure 12.12: **Planets spin and angle relatively to the ecliptic plane**

2. The orbits precess around the Sun (actually a G2V-type star, aka yellow dwarf) as shown in the following figure (see further below for the mathematical proof in the case of a 2-body system):

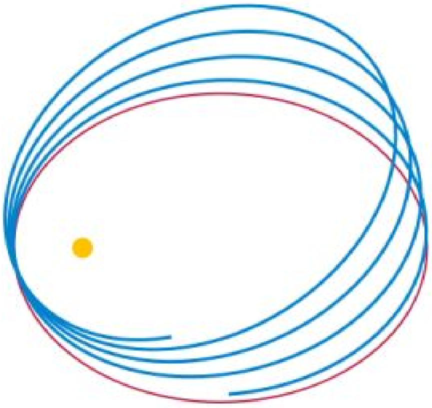


Figure 12.13: **Orbit precession example**

3. The planets have an helical trajectory "behind" the movement of the Sun around the center of our Galaxy[5] and the ecliptic has an angle of approximately 60° ($\pi/3$ [rad]) relatively to the perpendicular of the Sun movement as visible in the figure below:

---

[5] The velocity of the Sun around the galactic center is named the "galactic orbital motion" and there is a systematic review (see [554]) of 42 different methods from more than 30 studies and as many independent teams that lead to an actual median of 232 $[\text{km}\cdot\text{s}^{-1}]$ and a mean weighted value at one sigma of 228 ± 2; $[\text{km}\cdot\text{s}^{-1}]$.

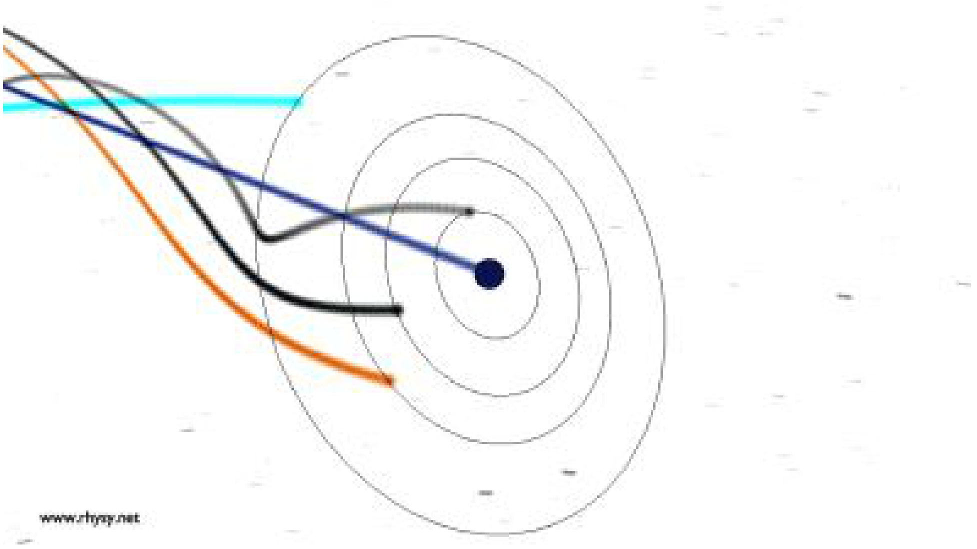


Figure 12.14: Schematic view of planets falling behind the Sun in its movement

and the planets are therefore sometimes in front of the Sun (actually a G2V-type star, aka yellow dwarf) and sometimes behind. Below is a prettier view rendered with a computer:

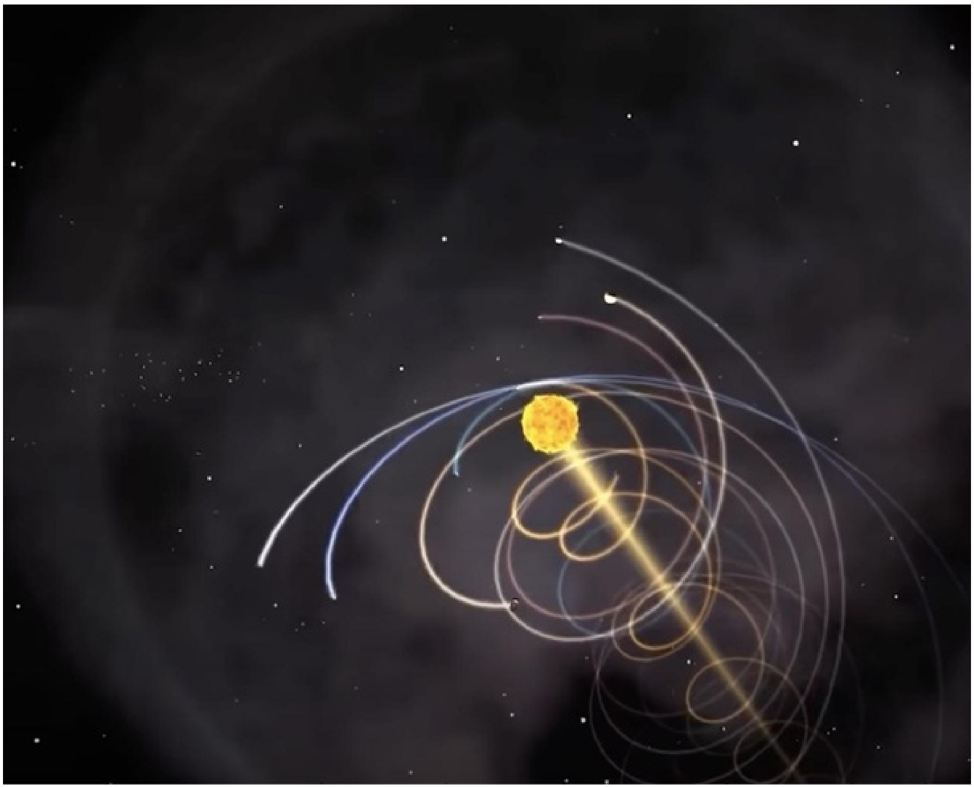


Figure 12.15: Computer rendered view of planets falling behind the Sun in its movement

Because of the movement of our Sun around the center of our Galaxy and some far-distant interactions in-between other celestial objects, the reader must understand that constellations doesn't appear always as there are today (this is named the "secular parallax")! The geometry changes and even on the very long term... some stars of the constellations may disappear (as they have a finished lifetime!):

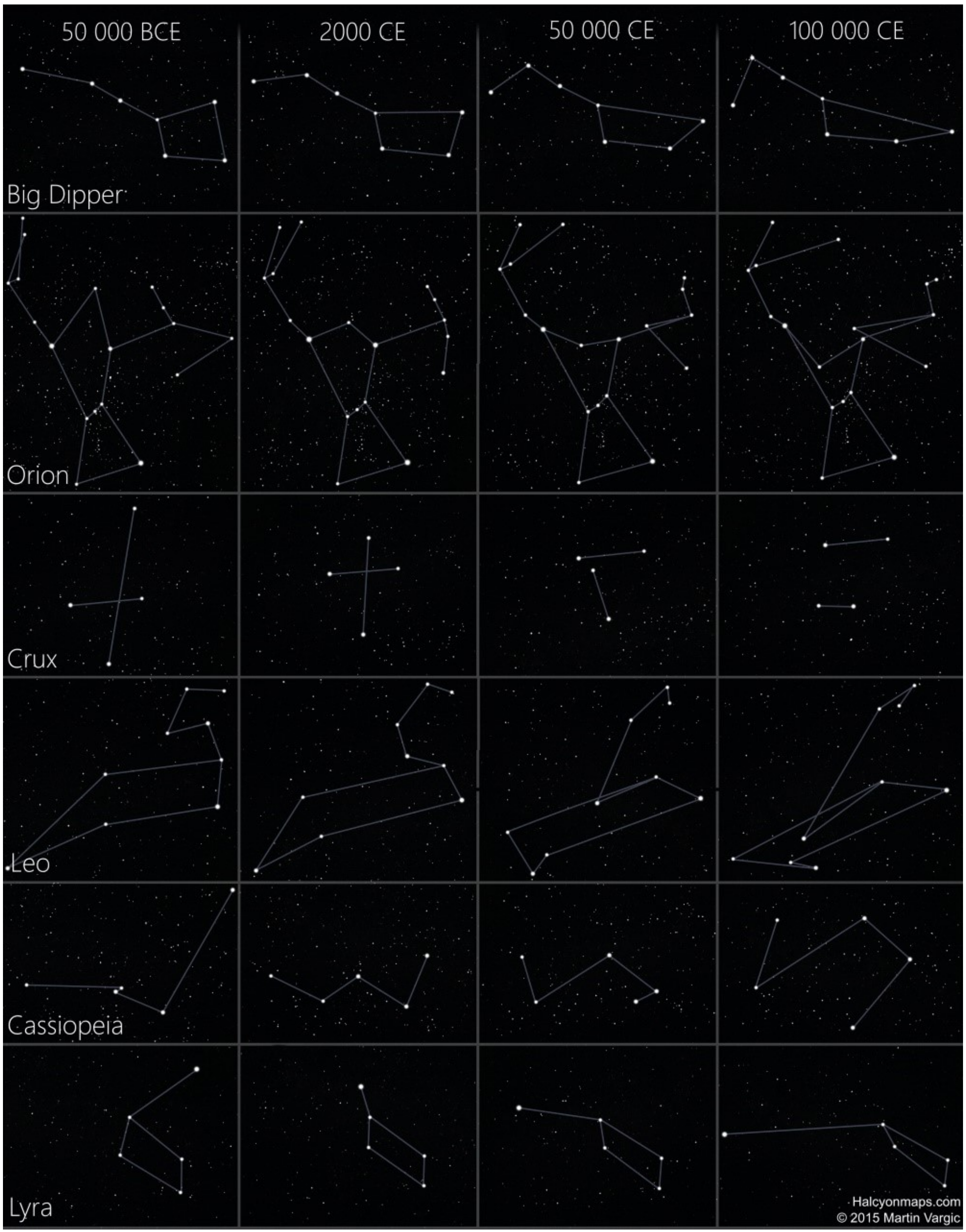

50 000 BCE
2000 CE
50 000 CE
100 000 CE
Big Dipper
Orion
Crux
Leo
Cassiopeia
Lyra
Halcyonmaps.com
© 2015 Martin Vargic

Figure 12.16: Constellations throughout ages with years based on Gregorian calendar (author: Martin Vargic)

For scales remember that hominids exist since only 7 millions years (at the date we write these lines...) and Homo Sapiens since only $300,000$ years ago.

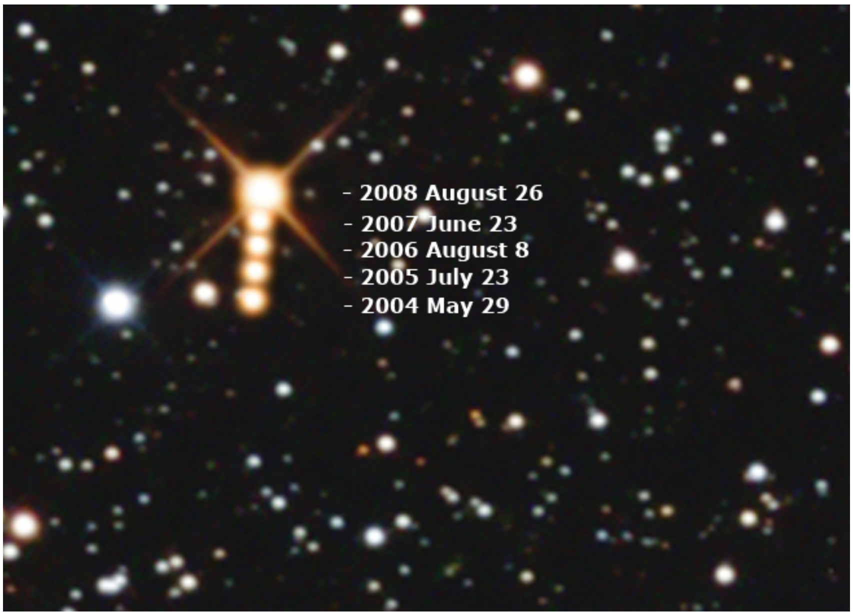

- 2008 August 26
- 2007 June 23
- 2006 August 8
- 2005 July 23
- 2004 May 29

Figure 12.17: Time lapse image of Barnard's star moving (constellation of Ophiuchus) – the 2nd closest star to Earth - across the background sky (source: Backyard Astronomer)

Old astrolabe, like the latin astrolabe below from 11550 (holocene calendar), are not aligned with nowadays stars positions as shown in the paper [130]:

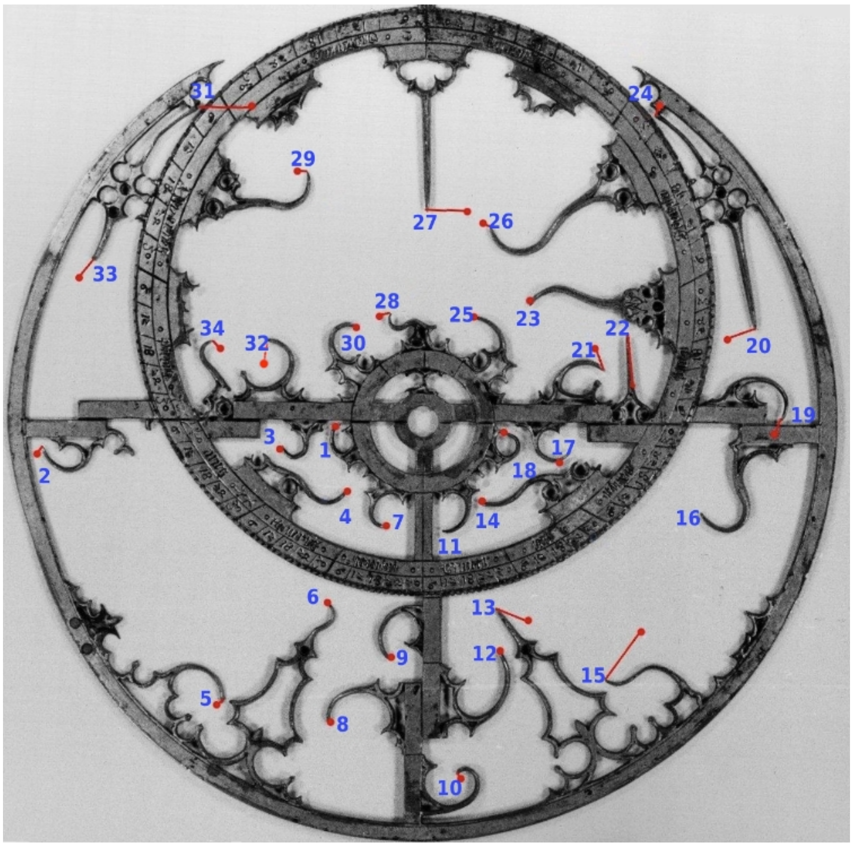

31
24
29
27
26
33
28
25
23
34
32
30
22
21
20
19
3
1
17
2
18
4
7
14
16
11
6
13
12
9
15
5
8
10

Figure 12.18: The Latin astrolabe's "rete" from the Musée des Arts précieux Paul-Dupuy (Toulouse, France) with its 34 pointers (blue) and the reference stars (red) most closely aligned with it nowadays (source: [130])

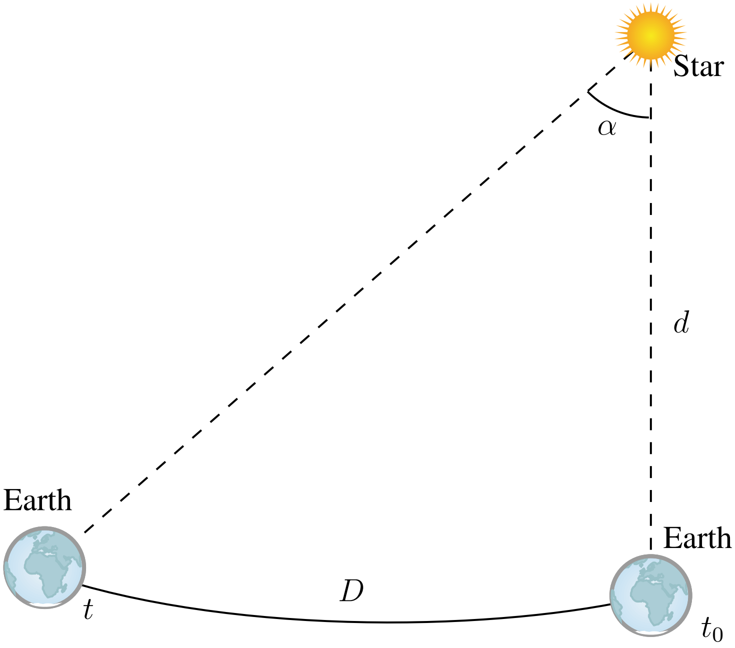
For example, the distance from Earth to Betelgeuse is of approximately $d = 6.07 \cdot 10^{18}$ [m]. Let us consider at time t_0 where this distance has been measured. Now, our Solar

System - Earth and all - whirls around the center of our Galaxy at some 220 kilometres per second. Hence after 100 year, Earth's have moved away from it's initial position (relatively to the galactic center) at t_0 of:

$$D=(100\cdot 365\cdot 24\cdot 60\cdot 60)\cdot(220\cdot 10^3)\cong 694\cdot 10^{14}\,[\text{m}] \tag{12.1.73}$$

Which corresponds to a variation of angle (parallax) between position at time t_0 and position at time t of (see the idea in the figure below that is obviously not to scale...):

$$\alpha\cong\arctan\left(\frac{D}{d}\right)=\arctan\left(\frac{694\cdot 10^{14}}{6.07\cdot 10^{18}}\right)\cong 0.006^\circ \tag{12.1.74}$$



For Proxima Centauri, the closest star to our own, the same calculation leads to:

$$\alpha=\arctan\left(\frac{694\cdot 10^{14}}{8.231\cdot 10^{16}}\right)\cong 0.988^\circ \tag{12.1.75}$$

This is obviously an upper value because in reality most of the stars in our night sky move with us in the galactic arm in the same direction!

Not only apparent relative distances between stars within a constellation (and every other star!) changes over time but also the absolute distance between them and Earth is in reality an illusion that also changes over time !! For example at naked eye Orion's constellation stars may seem all lying in an identical plane (and at a human scale life not moving...):

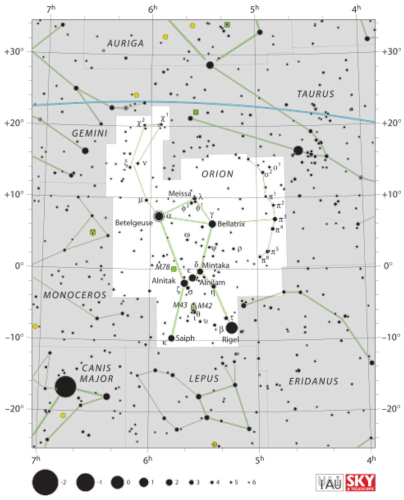

7h
6h
5h
4h
+30°
+20°
+10°
0°
−10°
−20°
AURIGA
TAURUS
GEMINI
ORION
Meissa
Betelgeuse
Bellatrix
Mintaka
Alnitak
Alnilam
M78
M43
M42
Saiph
Rigel
MONOCEROS
CANIS MAJOR
LEPUS
ERIDANUS
IAU SKY

Figure 12.19: Orion stellarium representation (authors: Roger Sinnott & Rick Fienberg)

Figure 12.20: Orion short pose photography (author: Till Credner)

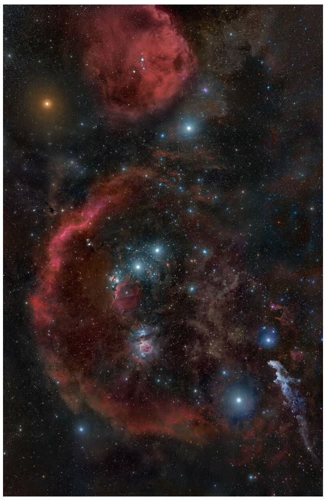


Figure 12.21: Orion long pose photography (author: Rogelio Bernal Andreo)

But in reality the various stars are at quite different distances as depicted in the figure below (the various stars of Orion's constellation are moving in various directions at speeds between 10 and 30 $[\text{km}\cdot\text{s}^{-1}]$, hence between 20 and 75 billion kilometres in an average human lifetime, but that's tiny compared to a light-year distance and this is why visible changes at naked eyes takes tens of thousands of years...much more than the actual record of a single human lifetime!):

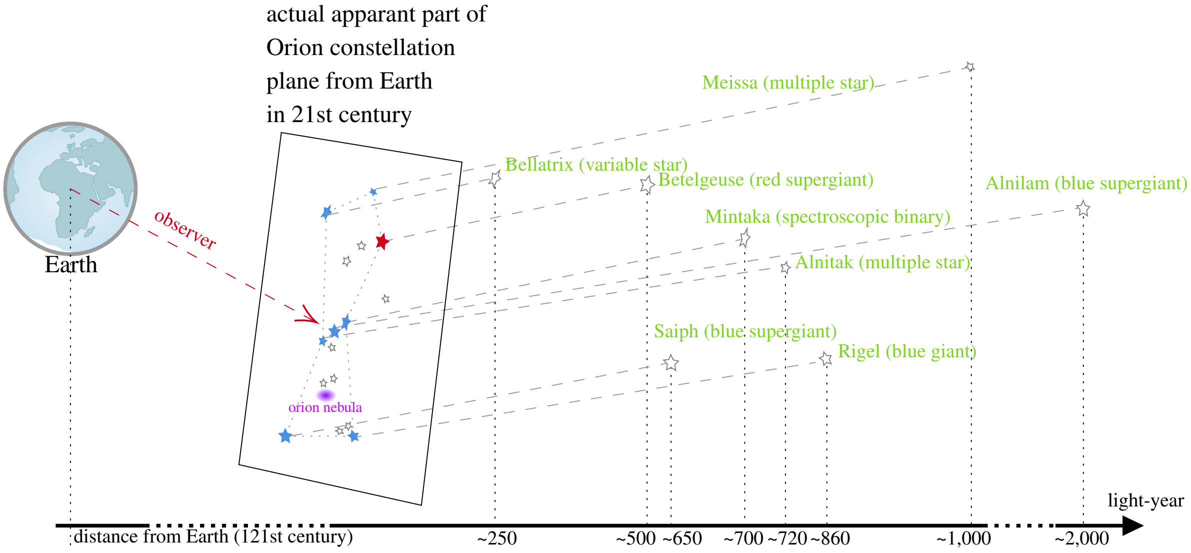

actual apparant part of
Orion constellation
plane from Earth
in 21st century
Earth
observer
orion nebula
Meissa (multiple star)
Bellatrix (variable star)
Betelgeuse (red supergiant)
Alnilam (blue supergiant)
Mintaka (spectroscopic binary)
Alnitak (multiple star)
Saiph (blue supergiant)
Rigel (blue giant)
light-year
distance from Earth (121st century)
~250
~500
~650
~700
~720
~860
~1,000
~2,000

Figure 12.22: Real approximative distances of certain stars in the Orion's constellation as measured between years 12000 and 12010 (holocene calendar)

This is how the Universe is! A giant (very likely infinite) mess where the huge majority of stuff is dynamic and interacts to generate/destruct other stuff throughout time.

4. On the very very long term the orbits in a n-body system is a chaotic deterministic system that has period of quasi-stability but that sometimes diverges completely. This is great opportunity at the date we write these lines to be in such a period of quasi-stability.

Remark

The n-body problem considers n point masses $m_i, i = 1, 2, \ldots, n$ in an inertial reference frame in three dimensional space $\mathbb{R}^3$ moving under the influence of mutual gravitational attraction. Each mass m_i has a position vector $\vec{r}_i$. Newton's second law says that mass times acceleration $m_i \frac{\mathrm{d}^2\vec{r}_i}{\mathrm{d}t^2}$ is equal to the sum of the forces on the mass. Newton's law of gravity says that the gravitational force felt on mass m_i by a single mass m_j is given by:

$$\vec{F}_{ij} = \frac{Gm_i m_j}{\|\vec{r}_j - \vec{r}_i\|^2} \cdot \frac{(\vec{r}_j - \vec{r}_i)}{\|\vec{r}_j - \vec{r}_i\|} = \frac{Gm_i m_j\,(\vec{r}_j - \vec{r}_i)}{\|\vec{r}_j - \vec{r}_i\|^3} \tag{12.1.76}$$

where G is the gravitational constant and $|\vec{r}_j - \vec{r}_i|$ is the magnitude of the distance between $\vec{q}_i$ and $\vec{q}_j$ (metric induced by the L_2 norm). Summing over all masses yields the "n-body equations of motion":

$$\boxed{m_i \frac{\mathrm{d}^2\vec{r}_i}{\mathrm{d}t^2} = \sum_{\substack{j=1\\ j\neq i}}^{n} \frac{Gm_i m_j\,(\vec{r}_j - \vec{q}_i)}{\|\vec{r}_j - \vec{r}_i\|^3} = -\frac{\partial U}{\partial \vec{r}_i}} \tag{12.1.77}$$

where U is the self-potential energy.

Some very typical very simple empirical evidence that the Earth rotates around the Sun (barycenter Earth-Sun to be exact) and not the opposite are:

1. The apparent position of stars close to earth (like Alpha Centauri) changes relative to stars much much further away (star parallax).
2. The retrograde motion of planets as seen from the Earth is incompatible with a geocentric system.
3. The fact that old astrolabs show that the shape of constellations has changed means that the Earth move and is not fixed or respectively that the stars themselves are not fixed like the Barnard's star whose movement can easily been seen (or obviously both!).
4. Galileo observed that Venus goes through a full set of phases, similar to the Moon. This can only happen if Venus orbits the Sun, which in turn implies that Earth does as well.

> *The Earth is not fitted to us, we are fitted to the Earth.*

quote 12.1:5251.#68

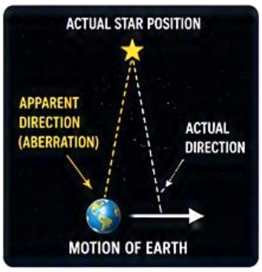
EARTH MOVES IN SPACE

Centuries of observations, experiments, and measurements from many fields of science all tell the same story: Earth is in motion.

1 STELLAR ABERRATION

Because Earth moves through space, starlight appears to come from a slightly forward direction—the same way raindrops seem to slant backward when you walk forward.

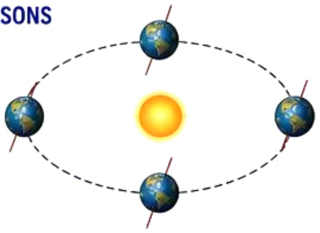
Measured and explained by James Bradley in 1728.


ACTUAL STAR POSITION
APPARENT DIRECTION (ABERRATION)
ACTUAL DIRECTION
MOTION OF EARTH

2 THE YEAR AND SEASONS

Earth orbits the Sun once every ~365.25 days.

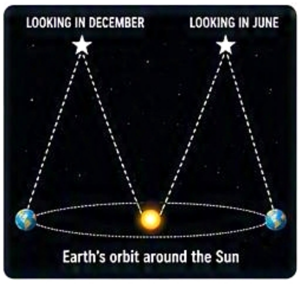
The tilt of Earth's axis (23.5°) as it orbits the Sun causes the changing seasons.



3 STELLAR PARALLAX

As Earth orbits the Sun, nearby stars appear to shift position against distant background stars over 6 months.

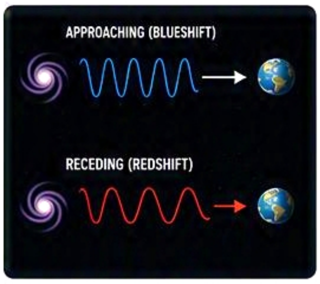
This effect can only be explained by Earth's motion.


LOOKING IN DECEMBER
LOOKING IN JUNE
Earth's orbit around the Sun

4 DOPPLER SHIFT

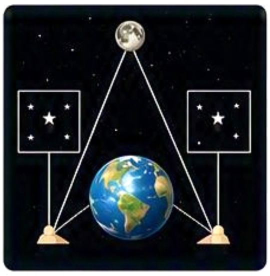
When Earth moves toward a source of light or radio waves, the waves are blueshifted (shorter wavelength). When moving away, they are redshifted (longer wavelength).

This is observed for stars, galaxies, and other celestial objects.


APPROACHING (BLUESHIFT)
RECEDING (REDSHIFT)

5 PARALLAX OF THE MOON

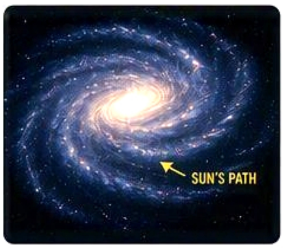
Observing the Moon from different locations on Earth at the same time shows it against the background stars from slightly different angles.
This proves Earth is large, round, and the Moon is relatively close —consistent with Earth's motion in space.



6 MOTION OF THE SOLAR SYSTEM THROUGH THE MILKY WAY

The Sun (and Earth) orbits the center of the Milky Way once about every 225–250 million years.

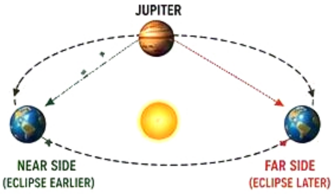
This is revealed by the motion of stars, gas clouds, and the cosmic microwave background.



7 SIGNAL DELAY & PULSAR TIMING (RØMER DELAY)

Ole Rømer demonstrated that the eclipses of Jupiter's moon Io occurred about 16.6 minutes earlier when Earth was on the near side of its orbit relative to Jupiter compared to when it was on the far side.

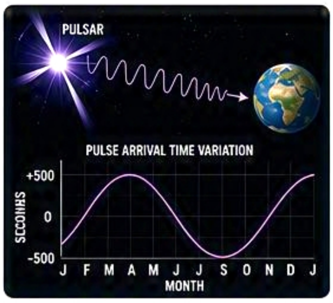
This time difference represents the time required for light to travel across the diameter of Earth's orbit.


JUPITER
NEAR SIDE (ECLIPSE EARLIER)
FAR SIDE (ECLIPSE LATER)

Extra distance for light when Earth is on the far side = ~16.6 minutes

8 PULSAR TIMING

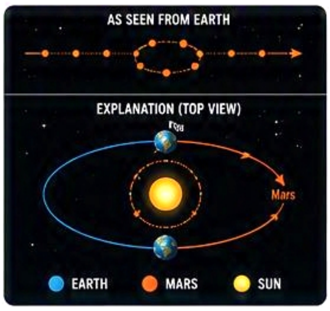
Millisecond pulsars emit radio pulses with extreme regularity. Precise timing models at Earth-based observatories show that pulse arrival times oscillate predictably by up to ±500 seconds throughout the year due to Earth's movement closer to or farther from the pulsar.



9 RETROGRADE MOTION OF PLANETS

Mars, Jupiter, and Saturn sometimes appear to reverse direction and trace loops against the background stars.

This happens when Earth, moving faster on its orbit, overtakes an outer planet. The heliocentric model predicts the timing, shape, and duration of these loops with high precision.

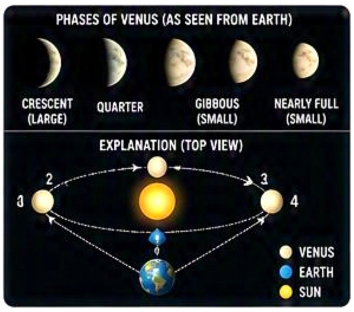

AS SEEN FROM EARTH
EXPLANATION (TOP VIEW)
Mars
EARTH
MARS
SUN

10 PHASES & ANGULAR SIZE OF VENUS

Venus shows a complete sequence of phases and changes in apparent size.

- When between Earth and the Sun: large but crescent.
- When on the far side of the Sun: small but nearly full.

The observed phase–size relationship is exactly what is expected if Venus and Earth orbit the Sun.


PHASES OF VENUS (AS SEEN FROM EARTH)
CRESCENT (LARGE)
QUARTER
GIBBOUS (SMALL)
NEARLY FULL (SMALL)
EXPLANATION (TOP VIEW)
2
3
4
VENUS
EARTH
SUN

Remark

A "zodiacal constellation" is any constellation (human imaginary figural construction of a group of stars) whose official IAU (International Astronomical Union) boundaries lie along the ecliptic — the apparent yearly path of the Sun — so the Sun, Moon, and planets appear to pass through them; astronomers treat the zodiac as a narrow band roughly a few degrees north and south of the ecliptic rather than as thirteen equal signs.

Actually, in astronomy (not in astrology that is not a Science!), there are thirteen constellations. As they don't have the same size, they obviously don't have the same number of days. For example for the well-know "Western constellations" (yes chinese don't have the same for example...) also referred to as "tropical astrology", we have:

Constellation	**Approximate days spent**	**Approximate angular size along ecliptic**
Pisces (Fishes)	38 days	37.5°
Aries (Ram)	25 days	24.7°
Taurus (Bull)	37 days	36.5°
Gemini (Twins)	31 days	30.5°
Cancer (Crab)	20 days	19.7°
Leo (Lion)	37 days	36.5°
Virgo (Maiden)	45 days	44.4°
Libra (Scales)	23 days	22.7°
Scorpius (Scorpion)	7 days	6.9°
Ophiuchus (Serpent Bearer)	18 days	17.8°
Sagittarius (Archer)	32 days	31.5°
Capricornus (Sea-Goat)	28 days	27.6°
Aquarius (Water Bearer)	24 days	23.7°

Table 12.2: **Actual durations and angular sizes of the actual western zodiacal signs**

As the solar system moves and stars also move, the astrological zodiacal dates doesn't correspond anymore to the scientific observed astronomical dates (and the difference will increase as hundreds of thousands or millions of years will pass):

Constellation	Tropical zodiac dates	IAU boundaries
Aries	Mar 21 - Apr 19	Apr 18 - May 13
Taurus	Apr 20 - May 20	May 13 Jun 21
Gemini	May 21 - Jun 20	Jun 21 - Jul 20
Cancer	Jun 21 - Jul 22	Jul 20 - Aug 10
Leo	Jul 23 - Aug 22	Aug 10 Sep 16
Virgo	Aug 23 - Sep 22	Sep 16 - Oct 30
Libra	Sep 23 - Oct 22	Oct 30 - Nov 23
Scorpio	Oct 23 - Nov 21	Nov 23 - Nov 29
Ophiuchus	-	Nov 29 - Dec 17
Sagittarius	Nov 22 - Dec 21	Dec 17 - Jan 20
Capricorn	Dec 22 - Jan 19	Jan 20 - Feb 16
Aquarius	Jan 20 - Feb 18	Feb 16 - Mar 11
Pisces	Feb 19 - Mar 20	Mar 11 - Apr 18

Table 12.3: **Astrological western zodiacal dates vs Scientific western zodiacal dates (as in 121th century according to holocene calendar)**

12.1.4 Newton's Gravitational Law

To check the accuracy of his hypothesis, Newton (relatively long after Kepler) found Kepler's laws from the law of gravity, giving the explanation of the general movement of the planets.

Newton considered to determine the law of gravitation a theoretical planet orbiting around the Sun (actually a G2V-type star, aka yellow dwarf) in a circular orbit at a constant speed v. During a complete orbit the planet travels a distance equal to the circumference of the circle of radius R, or $2\pi R$, in a time (the period) equal to the distance divided by its velocity, either:

$$T=\frac{2\pi R}{v} \tag{12.1.78}$$

Newton then relies on the third Kepler's law with always the assumption of a circular orbit.

We therefore have:

$$T^2=c^{te}R^3 \tag{12.1.79}$$

but as:

$$T=\frac{2\pi R}{v} \tag{12.1.80}$$

Then we get by substitution:

$$T^2=4\pi^2\frac{R^2}{v^2} \tag{12.1.81}$$

By comparing:

$$T^2=c^{te}R^3 \tag{12.1.82}$$

and:

$$T^2=4\pi\frac{R^2}{v^2} \tag{12.1.83}$$

and now assuming that $4\pi^2$ is divided by the constant is a new constant (which will be denoted in the same manner as the first although it is not equal to...) we get:

$$R^3=\frac{4\pi^2}{c^{te}}\frac{R^2}{v^2}=c^{te}\frac{R^2}{v^2} \tag{12.1.84}$$

Therefore:

$$R^2=c^{te}\frac{R}{v^2} \tag{12.1.85}$$

Then, if we reverse the terms, this expression becomes (while noting that the inverse of the original is constant is, also, a constant):

$$\frac{1}{R^2}=c^{te}\frac{v^2}{R} \tag{12.1.86}$$

By another calculation, we have already established in the section of Classical Mechanics (see page 3388) the expression of centrifugal (fictional) force:

$$F_{\text{cent}} = m\frac{v^2}{R} \qquad (12.1.87)$$

by comparing this expression with the previous one:

$$\frac{1}{c^{te}}\frac{1}{R^2} = \frac{v^2}{R} \qquad (12.1.88)$$

we get:

$$F_{\text{cent}} = m\frac{v^2}{R} = m\frac{1}{c^{te}}\frac{1}{R^2} = \frac{mc^{te}}{R^2} = F_G \qquad (12.1.89)$$

There should therefore exist a force opposed to the centrifugal (fictional) force that keeps the orbital cohesion and which can be written:

$$F_G = \frac{mc^{te}}{R^2} \qquad (12.1.90)$$

remains to determine the value of the constant!

It is trivial that the central mass M of the orbital system has to intervene in one way or another in this constant. If the mass of the secondary body intervenes proportionally in the centrifugal (fictional) force, the desire is great to do the same with the mass of the central body. So:

$$F_G = \frac{mMc^{te}}{R^2} \qquad (12.1.91)$$

Now there would be a priori more parameters to take into account. The remaining constant is here to meet the dimensional analysis so that we have "Newtons" (name given to the unit of force) on both sides of the equality. Scientists have determined with precision this "gravitational constant" denoted by G that a priori seems universal and which in SI units (as recommended by the 12014 - according to holocene calendar - CODATA[6] with standard uncertainty in parentheses), is:

$$G \cong 6.67408(31) \cdot 10^{-11} \left[\text{N} \cdot \text{m}^2 \cdot \text{kg}^{-2}\right] \qquad (12.1.92)$$

Remark

So far, within the commonly accepted picture G (that can also be expressed in $\left[\text{kg}^{-1} \cdot \text{m}^3 \cdot \text{s}^{-1}\right]$), like light velocity c, has been considered as one of the most fundamental constants of the Universe. Nevertheless, recent theoretical studies suggested that G could be truly expressed as a function of more fundamental physical quantities, i.e. the so named "Quantum Vacuum Zero Point Field Mass - Density Equivalent" ρ_{Qv} giving a measure of the energy density of the QED quantum vacuum (also known as ZPF). If we play with the maths it can be shown (see [83]) that the inverse of G multiplied by c (to get rid of the concept of time and transform it into a dimension of space) is equivalent in units as the amount of vacuum mass density moved on a distance of one meter.

---

[6] Committee on Data for Science and Technology

force needed to displace a cubic metre of vacuum that weighs one kg in one second.

Which brings us to write the two body "Newton's gravitational law":

$$\boxed{F_G = G\frac{mM}{R^2}} \tag{12.1.93}$$

that is a "central force" identically to the electrostatic force! The Newton's gravitational law for a simple three body case problem like the Sun-Earth-Moon system is given in details at page 5498 and is given by:

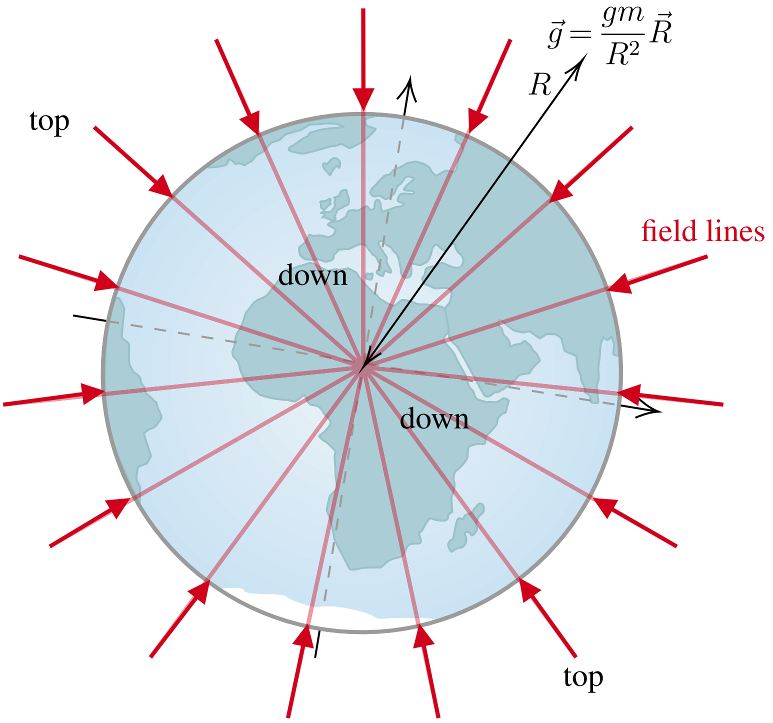
$$m_a\frac{\mathrm{d}^2\vec{r}_a}{\mathrm{d}t^2} = \sum_{b\neq a} Gm_am_b\frac{\vec{r}_b-\vec{r}_a}{|\vec{r}_b-\vec{r}_a|^3} \quad \text{for } a = 1, 2 \text{ and } 3 \tag{12.1.94}$$

Obviously the derivation above is not a true rigorous proof because it is based on experimental Kepler's observations! By cons, from General Relativity (see page 5751) it is possible to prove it (under some given assumptions...), therefore it loss its status as law and becomes a theory!

Remark

Newtonian gravity is not local, this bothered Newton a lot back then! You see if you take Newton's law of gravity, with the gravitational force proportional to the mass divided by the square of the distance, then if you move the mass, the distance increases and the force changes instantaneously, everywhere. Newton was greatly bothered by this. He named it "action at a distance" and that's also what Einstein meant by the phrase. Einstein's theory of General Relativity, which we use now to describe gravity, doesn't have this problem. In Einstein's theory, gravity is described by the curvature of space and time. And that curvature is caused by the mass and energy in space-time. If you move the mass that causes space-time to curve, then the curvature changes, but not instantaneously. The change spreads outwards at the speed of light (see section General Relativity page 5861). So, Einstein's theory is "local" as physicists say.

Below we have illustrated how the gravitational field is a central force pointing uniformly to the center of the Earth (similarly to a punctual electric charge):



Because the gravitational field is a central force exactly like the electrostatic field, we get exactly the same kind of mathematical expression and properties for the expression of the work and for the equipotentials (see section Electrostatics page 3980) that are therefore perpendicular to the field lines (obviously the reader should remember that in reality the field lines go to infinity and the equipotentials go from the center of the Earth to infinity too):

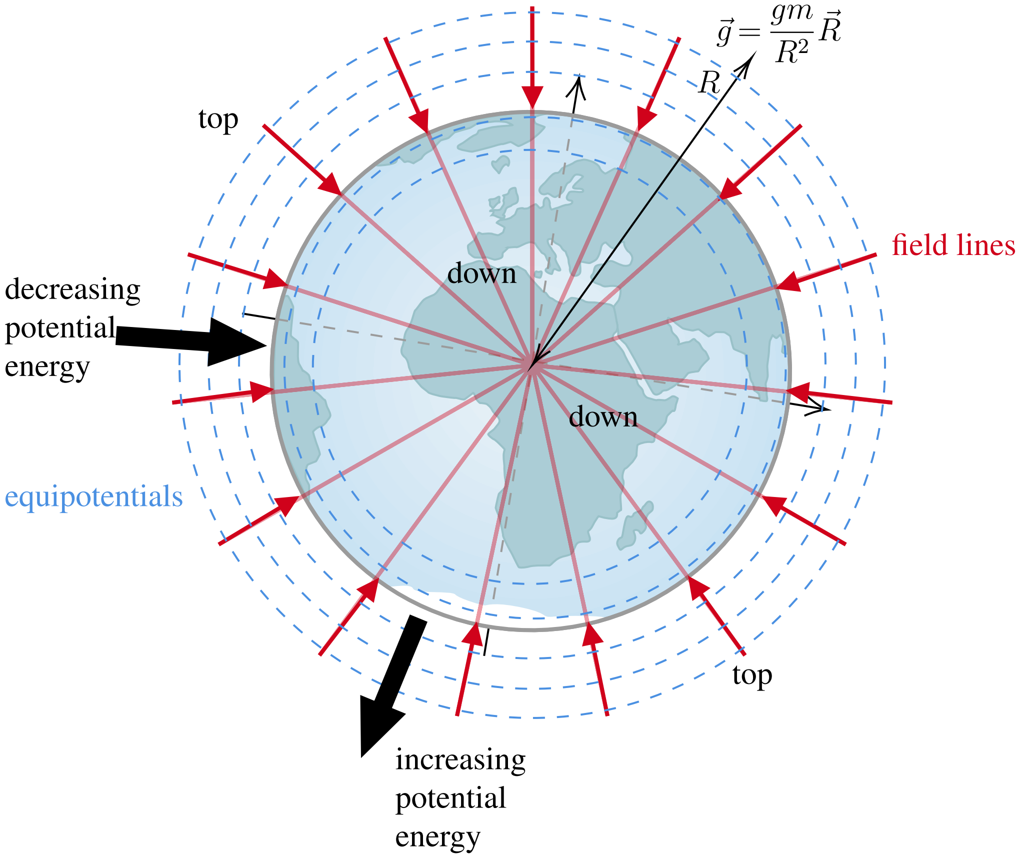

$\vec{g} = \frac{gm}{R^2}\vec{R}$
R
top
field lines
decreasing potential energy
down
down
equipotentials
top
increasing potential energy

Figure 12.23: **Gravitational field lines and equipotentials**

Or represented in a more childish way:

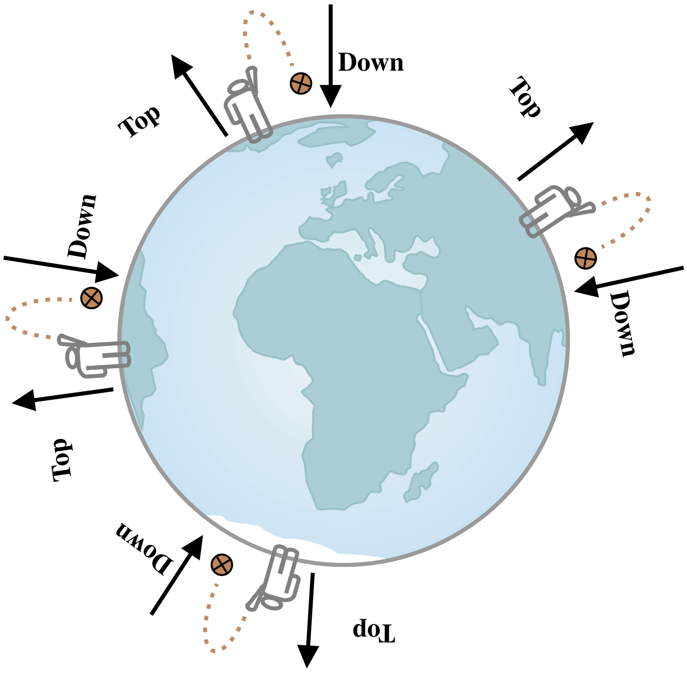

Down
Top
Top
Down
Down
Top
Down
Top

Figure 12.24: **Up and Down on a Spherically Symmetrical Planet**

When on Earth (and other planet/massive objects) we say that something goes "downwards" (like water flow for example), it doesn't mean that the thing follows the curvature Earth but that its potential energy decreases by approching the center of the Earth!

Caution! As we will prove it later (result already known at the time of Newton) the three macroscopic spatial dimensions are very special! In particular, in three dimensions gravity obeys an inverse square law, without which stable planetary orbits would not be possible. But keep in mind that since the space-time model is a human invention, so must be the dimensionality of space-time. We choose it to be three because it fits the data. In the M-theory we choose it to be eleven. We use whatever works, but that does not mean reality is exactly that way in one-to-one correspondence.

Definition (#12.1.526): In physics we say that something is "at level", or simply "levelled", when it coincides with a line or surface that cuts perpendicularly all plumb lines (i.e. perpendicular to gravity) that it meets and hence would everywhere coincide with a surface of still water (or any other continuum medium) that follows the gravitational equipotentials, the levelling is typically measured with a gravimeter[7]. The concept of "level" must not be confused with that of "flatness" - in the standard ISO 12781 "flatness" is defined as the distance between the minimum and maximum point of a surface - that is mainly used in mechanics and architecture and measured thanks to a flatness gauge/lap plate (for good precision) or in the worst case with a bubble level.

Remark

The equality:

$$a=G\frac{M}{r^2} \tag{12.1.95}$$

explains the genius intuition of Galileo: that in vacuum (!) a feather and bowling ball will fall at the same speed (i.e. with the same acceleration), as that latter depends only on the main body of attraction that has for mass M. Only the force on the two objects is not the same! But the acceleration is however the same!

Now some reader may think that there is a singularity in $R=0$ but this is impossible as given two mass particles of radius r_1 and r_2 we would have the maximum following force:

$$F_G=G\frac{m_1m_2}{(r_1+r_2)^2} \tag{12.1.96}$$

✍**Example:**

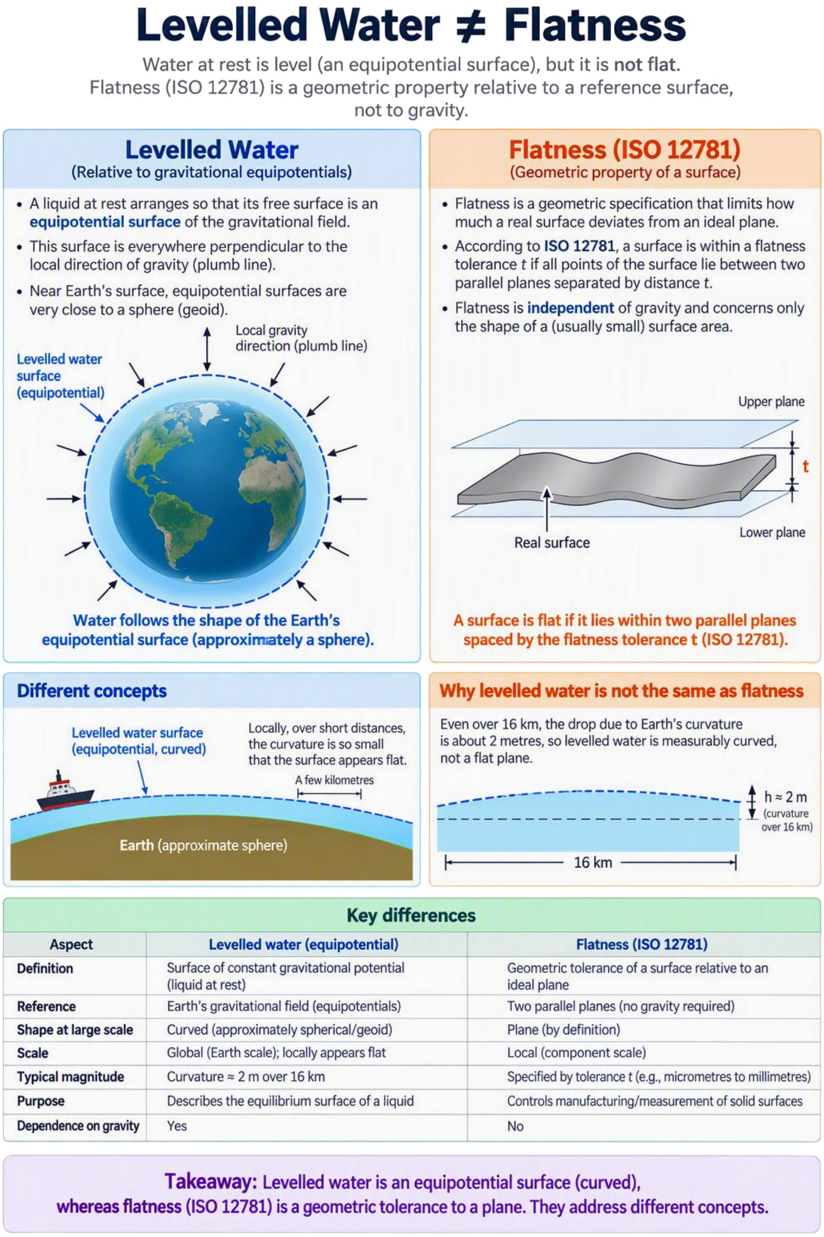
At the Earth's equator the radius is of 6378 [km] and at the poles of 6357 [km]. Therefore we have:

$$a_{\text{eq.}}=\frac{MG}{R_{\text{eq.}}}=9.800\,[\text{m}\cdot\text{s}^{-1}] \qquad a_{\text{po.}}=\frac{MG}{R_{\text{po.}}}=9.865\,[\text{m}\cdot\text{s}^{-1}] \tag{12.1.97}$$

---

[7] In the *Webster's ninth new collegiate dictionary*, "level" is first defined as : « *having no parti higher than another: conforming to the curvature of the liquid parts of the Earth's surface* »

Then the acceleration at the equator is equal to $9.800/9.865 \cong 99.34\%$ to that at the poles.



Levelled Water ≠ Flatness

Water at rest is level (an equipotential surface), but it is **not flat.**
Flatness (ISO 12781) is a geometric property relative to a reference surface, not to gravity.

Levelled Water

(Relative to gravitational equipotentials)

- A liquid at rest arranges so that its free surface is an **equipotential surface** of the gravitational field.
- This surface is everywhere perpendicular to the local direction of gravity (plumb line).
- Near Earth's surface, equipotential surfaces are very close to a sphere (geoid).

Local gravity direction (plumb line)
Levelled water surface (equipotential)

Water follows the shape of the Earth's equipotential surface (approximately a sphere).

Flatness (ISO 12781)

(Geometric property of a surface)

- Flatness is a geometric specification that limits how much a real surface deviates from an ideal plane.
- According to **ISO 12781**, a surface is within a flatness tolerance *t* if all points of the surface lie between two parallel planes separated by distance *t*.
- Flatness is **independent** of gravity and concerns only the shape of a (usually small) surface area.

A surface is flat if it lies within two parallel planes spaced by the flatness tolerance t (ISO 12781).

Different concepts

Levelled water surface (equipotential, curved)
Locally, over short distances, the curvature is so small that the surface appears flat.
A few kilometres
Earth (approximate sphere)

Why levelled water is not the same as flatness

Even over 16 km, the drop due to Earth's curvature is about 2 metres, so levelled water is measurably curved, not a flat plane.

Key differences

Aspect	Levelled water (equipotential)	Flatness (ISO 12781)
Definition	Surface of constant gravitational potential (liquid at rest)	Geometric tolerance of a surface relative to an ideal plane
Reference	Earth's gravitational field (equipotentials)	Two parallel planes (no gravity required)
Shape at large scale	Curved (approximately spherical/geoid)	Plane (by definition)
Scale	Global (Earth scale); locally appears flat	Local (component scale)
Typical magnitude	Curvature ≈ 2 m over 16 km	Specified by tolerance t (e.g., micrometres to millimetres)
Purpose	Describes the equilibrium surface of a liquid	Controls manufacturing/measurement of solid surfaces
Dependence on gravity	Yes	No

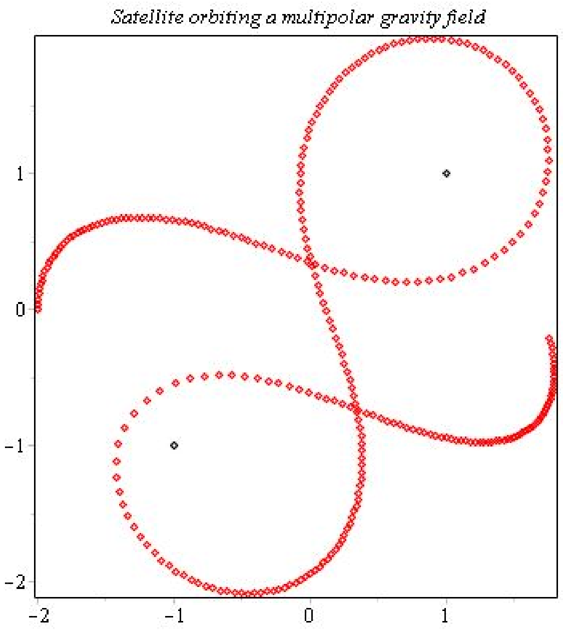
Takeaway: Levelled water is an equipotential surface (curved), whereas flatness (ISO 12781) is a geometric tolerance to a plane. They address different concepts.

It is very important to notice that the mutual forces of attraction acting on two spherical masses are always of equal magnitude!

Using Maple™ 17.00, we can simulate the plane trajectory of a satellite relative to n number of fixed mass (thanks to Forhad Ahmed for his script). Here below is given the basic script that you can customize to your tastes and... feel free to give us your personal work if you have brought significant improvement to this script:

```
>restart; with(plots); with(DEtools)
>G:=1; #normalized gravitational constant to simplify
>poles:=2; #number of bodies/masses that we can play with...
```

```
>M[1]:=10;M[2]:=1; #mass of the first and second body (in relative
values)
>h[1]:=1;h[2]:=-1; #X position of the first and second body X (in
astronomical units)
>k[1] := 1;k[2] := 1; #Y position of the first and second bodies (in
astronomical units)

>#differential equation of the satellite acceleration in X
>Xeq := diff(x(t), t, t) = sum(-G*M[j]*(x(t)-h[j])/((x(t)-h[j])^2
+(y(t)-k[j])^2)^(3/2), j = 1 ..  poles);
>#differential equation of the satellite acceleration in X
>Yeq := diff(y(t), t, t) = sum(-G*M[j]*(y(t)-k[j])/((x(t)-h[j])^2
+(y(t)-k[j])^2)^(3/2), j = 1 ..  poles);
>#position and initial velocity of the satellite
>inits := x(0) = -2, y(0) = 0, (D(x))(0) = 0, (D(y))(0) = 2
>#numerical solution of the differential equation (you can play with the
precision of the error as needed!)
>g:=dsolve(Xeq,Yeq,inits,x(t),y(t),type=numeric,method=dverk78,abserr=0.1e-3,
output= procedurelist);

>n:=50; #step of iterations
>iter:=300; #step of iterations

>#loop that resolves the differential equation at each new iteration
>for i from 0 to iter do
px[i]:=rhs(g(i/n)[2]);
py[i]:=rhs(g(i/n)[4]);
KE[i]:=1/2*(rhs(g(i/n)[3])^2+rhs(g(i/n)[5])^2);
temp:=(rhs(g(i/n)[2])-h[j])^2+(rhs(g(i/n)[4])-k[j])^2;
PE[i]:=sum(-G*M[j]/sqrt(temp), j = 1 ..  poles);
TE[i]:=KE[i]+PE[i]
end do:

>data:=seq(pointplot([px[i], py[i]], color = red), i = 0 ..  iter):
>#mettre insequence to true to get an animation
>Anim:=display(data,insequence=false,scaling=constrained,axes=boxed):
>stars:=display(seq(pointplot([h[i], k[i]], color = black), i = 1 ..
poles))
>display(Anim,stars,title=‘Satellite orbiting a multipolar gravity
field‘);
```



```
>#it is verified that the total energy of the satellite is always
constant
>print(‘[Time] - [Kinetic Energy] - [Potential Energy] - [Net Energy]‘);
>print(‘=====================================‘);
> for i by 3 to iter do
print(evalf(i/n, 6), ‘ ‘, KE[i], ‘ ‘, PE[i], ‘ ‘, TE[i]);
end do:
>#the last column of the table must always have normally an equal
value...
```

Remark

Equalizing the centrifugal (fictional) force and gravitational force, it is quite easy to get an approximation of the speed of rotation of the planets in their orbits. The reader that will do the calculation will see that the value for the planets of our Solar System is around a speed of about $100,000\,[\mathrm{km}\cdot\mathrm{h}^{-1}]$.

From this last relation, let us come back briefly on our third Kepler's law and detail it a little bit to show that it is valid for any type of conical orbit and to determine the expression of its constant.

Expressed in the Frenet coordinate system (see section Differential Geometry page 2863), and decomposed into its normal (centripetal) and tangential acceleration, the acceleration in respect to a geocentric reference frame (in the case of a referential located at the mass center of the system the expression change a little bit!) is written:

$$a_{tg}=0 \qquad a_n=\frac{v^2}{R} \tag{12.1.98}$$

From previous developments (3rd Kepler's law):

$$T^2=c^{te}R^3 \tag{12.1.99}$$

and:

$$T^2=4\pi^2\frac{R^2}{v^2} \tag{12.1.100}$$

the constant of Kepler's third law takes for value (it is a formulation sometimes used in practice but not a strictly necessary step in this development):

$$c^{te}=4\pi^2\frac{R}{v^2}\frac{1}{R^2}=4\pi^2\frac{1}{a_n}\frac{1}{R^2} \tag{12.1.101}$$

but as we also have:

$$a_n=g=\frac{v^2}{R}=\frac{Gm}{R^2}\Leftrightarrow v^2=\frac{Gm}{R} \tag{12.1.102}$$

then:

$$c^{te}=4\pi^2\frac{R}{\frac{Gm}{R}}\frac{1}{R^2}=4\pi^2\frac{R^2}{Gm}\frac{1}{R^2}=\frac{4\pi^2}{Gm} \tag{12.1.103}$$

Therefore:

$$T^2=\frac{4\pi^2}{Gm}R^3 \tag{12.1.104}$$

Finally, the third Kepler law can be found frequently in the literature as follows:

$$\boxed{\frac{R^3}{T^2}=\frac{Gm}{4\pi r^2}} \tag{12.1.105}$$

But now let us consider again our figure of the center of mass study (see section Classical Mechanics page 3213):

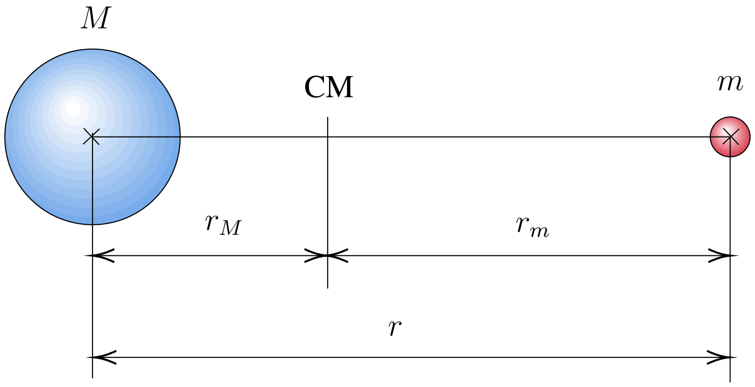


Figure 12.25: **Binary system center of mass (profile view)**

And let us have a look at circular orbits in the center of mass frame!

First we look at the forces acting on body M:

$$F=-\frac{GMm}{r^2}\qquad F_{\text{cent}}=\frac{Mv_M^2}{r_M} \tag{12.1.106}$$

The forces balance, so:

$$\frac{GMm}{r^2}=\frac{Mv_M^2}{r_M} \tag{12.1.107}$$

We know that in the simple case of a circular orbit (circular kinematics):

$$v_M=\frac{2\pi r_M}{T_M} \tag{12.1.108}$$

We insert this into the previous equation and we get after some algebra:

$$\frac{Gm}{r^2}=\frac{4\pi r_M}{T_M^2} \tag{12.1.109}$$

What is r_M? From the definition of center of mass we know that (see section Classical Mechanics page 3213):

$$r_M=\frac{m}{M}r_m=\frac{m}{M}(r-r_M)=\frac{M}{M+m}r \tag{12.1.110}$$

We plug r_M into our equation and now we get:

$$\frac{Gm}{r^2}=\frac{4\pi^2}{T^2}\frac{M}{M+m}r \tag{12.1.111}$$

or after rearranging terms:

$$\boxed{T=2\pi\sqrt{\frac{r^3}{G(M+m)}}} \tag{12.1.112}$$

So we can use this 3rd Kepler's law to determine the total mass a binary pair if we know the period T!!! As the star masses are well estimated using the HR diagram we better understand how astronomers estimate orbiting planet mass knowing the period (in fact they also use the luminosity variation). In fact it is not as simple as there is no reason why we should be looking directly onto the orbital plane of the binary system. In other words, the apparent orbit is almost never the true orbit (which is what we need to do the calculation).

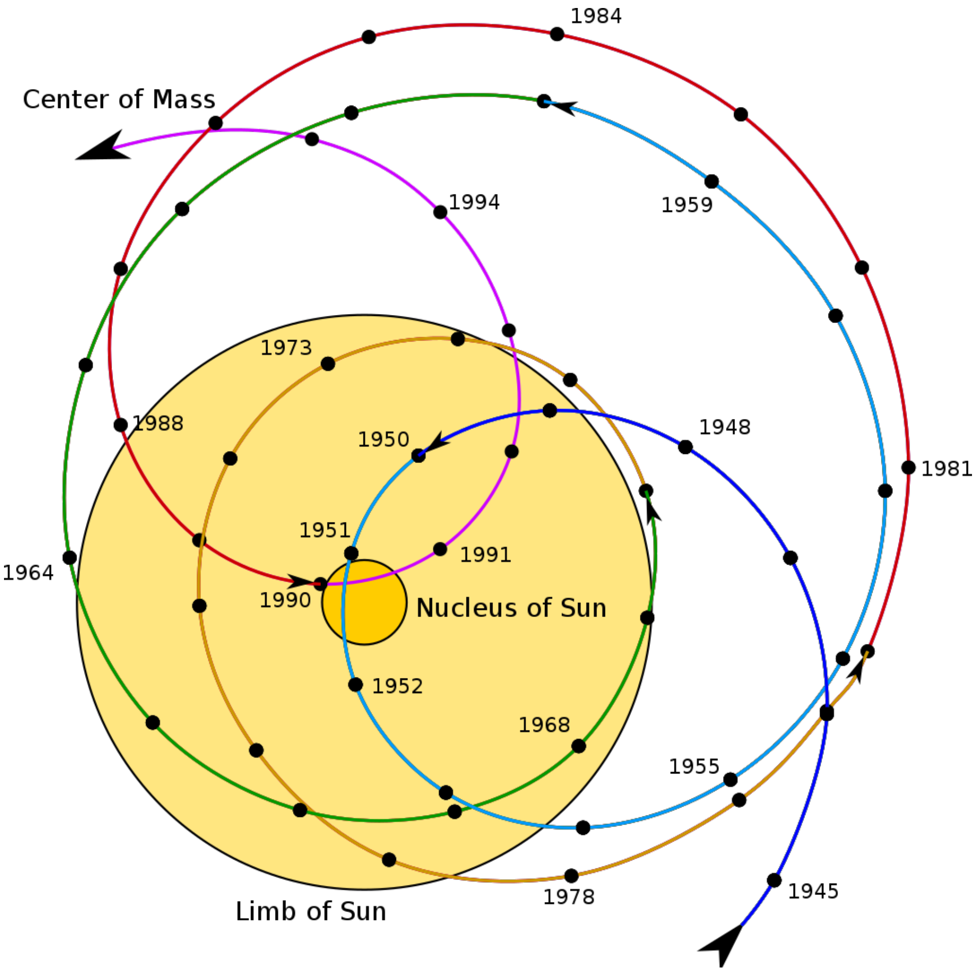

Center of Mass
1984
1994
1959
1973
1988
1950
1948
1981
1951
1991
1964
1990
Nucleus of Sun
1952
1968
1955
1978
1945
Limb of Sun

Figure 12.26: In case of a n-body system (here the Solar System) even the center of mass moves (source: Wikipedia)

This interlude performed, let us come back on our Newton's gravitation law:

$$F_G=G\frac{mM}{R^2} \tag{12.1.113}$$

From the law of gravitation, we can find back Kepler's law. Besides, we have already done it for the second and third law of Kepler, since it is these that we used to get this latter relation (however it's a little bit the snake eating its tail...).

In vector notation we have therefore:

$$\vec{F}_G=M\frac{Gm}{R^2}\frac{\vec{R}}{\|\vec{R}\|} \tag{12.1.114}$$

Identically to the electric field (see section Electrostatics page 3969), we can develop:

$$\vec{F}_G=M\sum_i\left(\frac{Gm_i}{\|\vec{r}-\vec{r}_i\|^2}\frac{\vec{r}-\vec{r}_i}{\|\vec{r}-\vec{r}_i\|}\right)=m\vec{a} \tag{12.1.115}$$

As the electric field is derived from an electric potential, identically, the gravitational field derived from a gravitational potential. By performing exactly the same development as in our study of electromagnetism for the first Maxwell equation (see section Electrodynamics page 4095), we prove that:

$$\vec{a}=\vec{\nabla}\cdot\varphi \tag{12.1.116}$$

where φ is the "gravitational potential" that varies inversely with the relative distance of the body (this confirms what we had proved in our study of Noether's theorem in the section on Principles) and is therefore equal to:

$$\varphi=\frac{GM}{r}\;[\mathrm{J}\cdot\mathrm{kg}^{-1}] \tag{12.1.117}$$

Remark

We often encounter this potential in the section of General Relativity (see page 5720). It is therefore appropriate to remember it if possible!

Notation which obviously implies the following relation:

$$\vec{F}=-m\vec{\nabla}\cdot\varphi \tag{12.1.118}$$

Remark

Obviously in the absence of field, we have $\varphi=c^{te}$ and therefore $\vec{a}$ will be zero.

As in the section of Electromagnetism, again, we prove as we did for the first Maxwell equation:

$$\Phi_{\vec{a}}=\oint_{\Sigma}\vec{a}\circ\mathrm{d}\vec{S}=\oint_{V}(\vec{\nabla}\circ\vec{a})\mathrm{d}V=4\pi Gm\Leftrightarrow\vec{\nabla}\circ\vec{a}=4\pi G\rho \tag{12.1.119}$$

If we express this equation in terms of a gravitational potential φ (also often denoted by the letter U as in Electrostatic...), we get:

$$\vec{\nabla}\circ(\vec{\nabla}\cdot\varphi)=4\pi G\rho \tag{12.1.120}$$

that we write more aesthetically with the scalar Laplacian operator (see section Vector Calculus page 1899):

$$\boxed{\Delta\cdot\varphi=4\pi G\rho} \tag{12.1.121}$$

which is nothing else than the "Newton-Poisson equation" that we will also meet again in our study of General Relativity (it has an important place for validation reasons of Einstein's Gravitation theory)!

This equation means that the Newtonian gravitational theory can be resume to say that the gravitational field is described by a single potential φ generated by the volume mass density and determining the acceleration of a test particle immersed in the outfield φ.

With Newton's important achievements - building, of course, on the work of the great thinkers who preceded him - civilization reached an extremely high level of knowledge about the universe. Newtonian mechanics and also the optics, astronomy, and mathematics to which Newton contributed were immensely valuable in helping us gain an understanding of the vast and complicated physical world around us. The progress made by Newton and others allowed us to see clearly how planets move and how the force of gravity operates in nature. Newton's theory of gravity is so profound and so all-encompassing that its laws govern everything from the fall of apples to the ground to the orbiting of the Earth by the Moon; from the revolutions of the planets around the Sun (actually a G2V-type star, aka yellow dwarf) to the actions of springs and the trajectories of cannonballs; from the behaviour of billiard balls to the energy of an accelerating car. Newtonian mechanics explains the world to a stunningly accurate degree. In the 120th century (holocene calendar), Einstein would refine Newton's theories to account for instances where the speed of light is approached or very great mass is involved. But for the time, Newton's achievements truly opened a new world for physical science. In the following two centuries, science would consolidate its gains, and the church (and religions in general) would have to retreat from its position as the source of truth about how the world works.

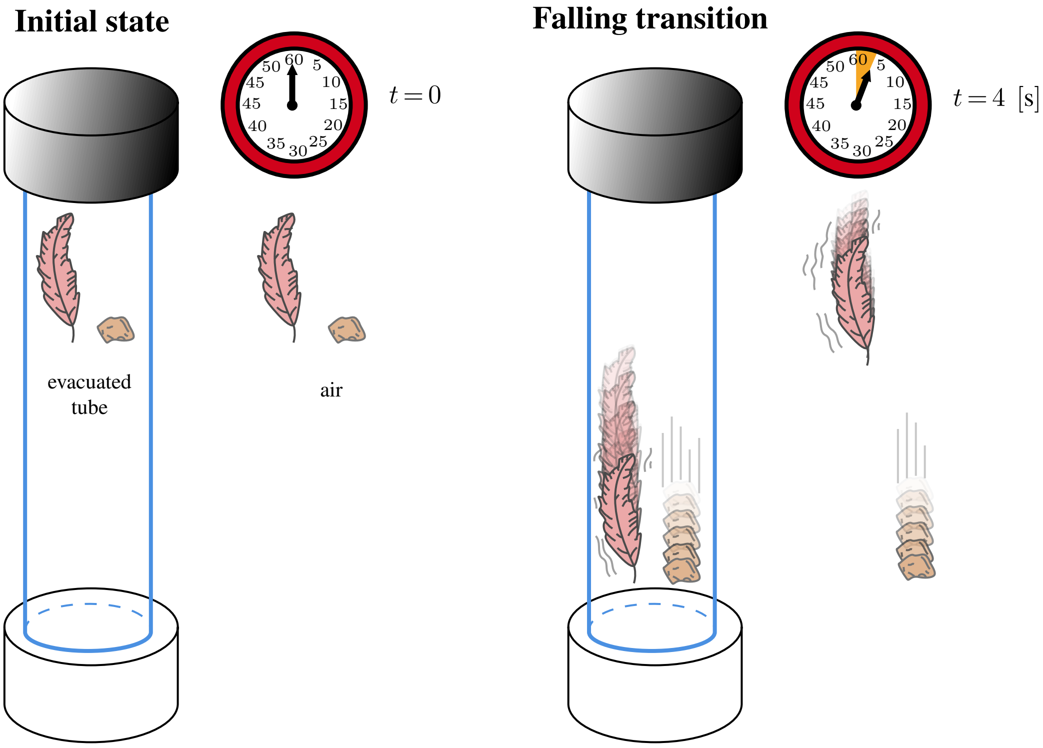


Figure 12.27: **Illustration of the famous experiment done in High Schools that provides evidence falling is primarily due to gravity**

The real-life experiment below shows that the speed of a freely falling object is independent of its mass. When there is air in the tube, the two objects have a different falling speed, because the objects feel the effects of a viscous flow. Since the shape and surface of the object play a key role in a viscous flow, the feather is strongly braked (high drag) while the lead ball has comparatively little drag:

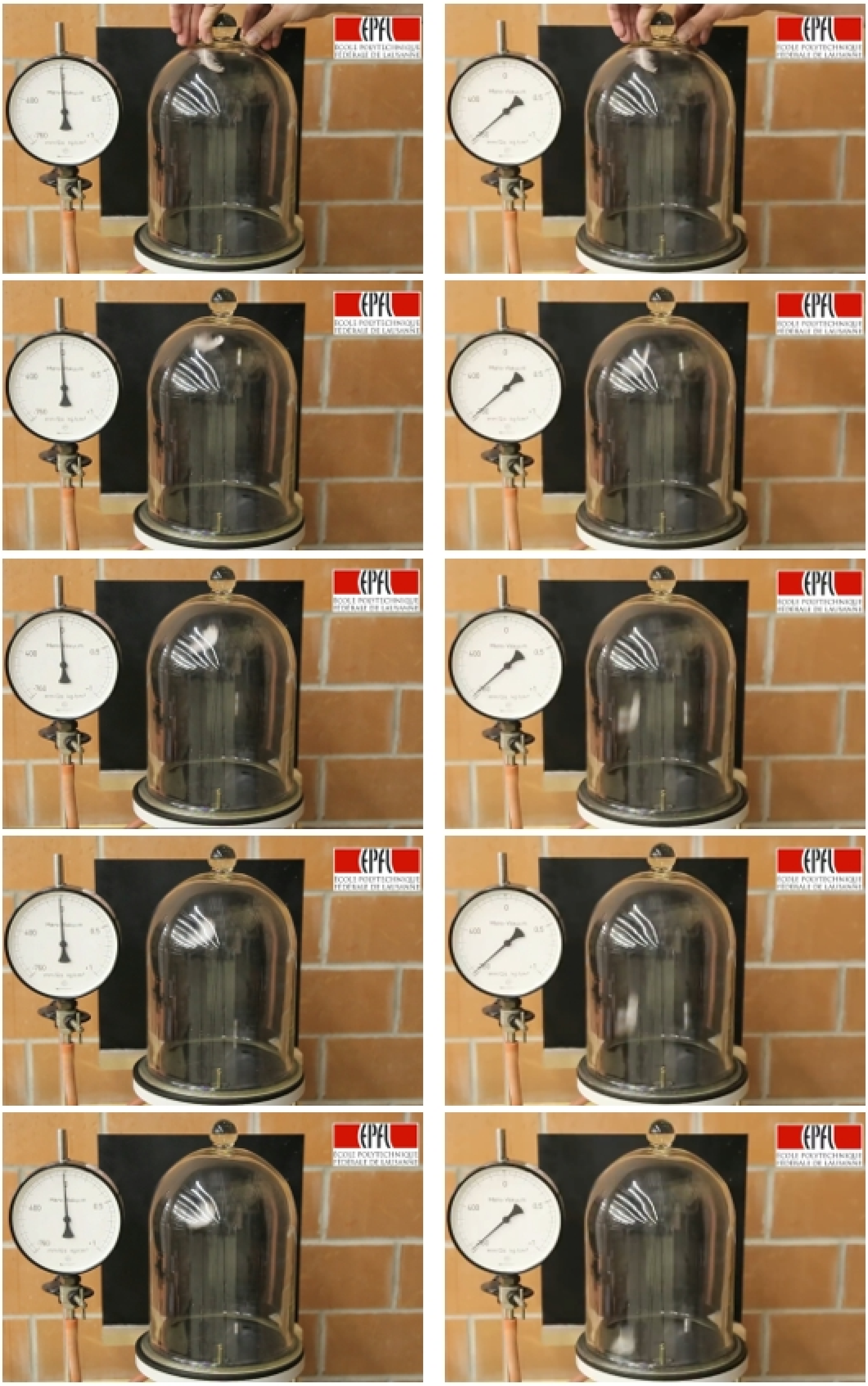

EPFL
ECOLE POLYTECHNIQUE FEDERALE DE LAUSANNE

Figure 12.28: **Free fall real life experiment in air (left column) and vacuum (right vacuum) between a feather vs a lead ball (source: EPFL, https://www.youtube.com/watch?v=fz7ulPC7Kwo)**

12.1.4.1 Cavendish Experiment

The Newtonian gravitational constant G, which is one of the most important fundamental physical constants in nature, plays a significant role in the fields of theoretical physics, geophysics, astrophysics and astronomy. Although G was the first physical constant to be introduced in the history of science, it is considered to be one of the most difficult to measure accurately so far.

The "Cavendish experiment", performed in 11797–11798 (holocene calendar) by English scientist Henry Cavendish, was the first horizontal (the fact that the experiment is horizontal guarantees that the observation has nothing to do with Buoyancy force as the whole experiment remains in the same horizontal plane) experiment to measure the force of gravity between masses in the laboratory and the first to yield accurate values for the gravitational constant.

This experiment consists in a torsion balance that is set up and the gravitational attraction between a pair of large spheres and a pair of smaller spheres is measured in terms of the torsion of the wire from which the smaller spheres are suspended from:

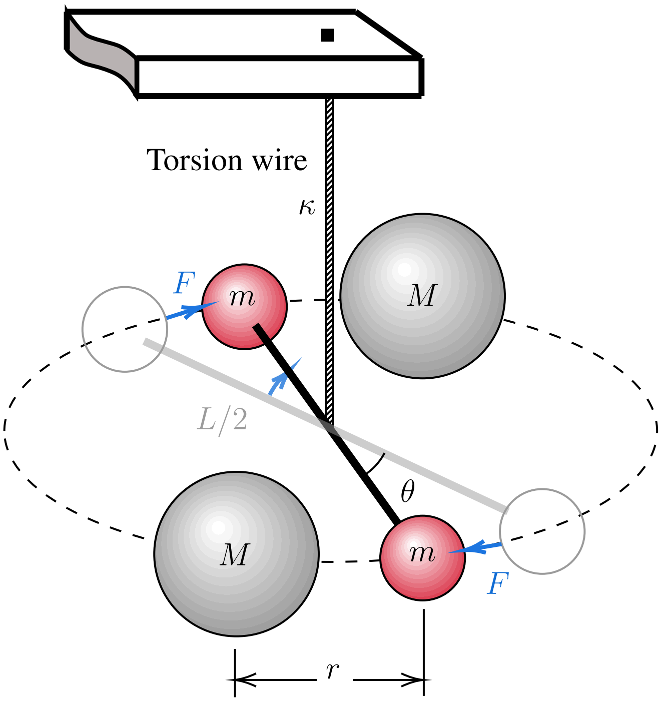


Figure 12.29: **Naive schema of the initial Cavendish experiment**

Drawing of torsion balance apparatus used by Henry Cavendish to measure the gravitational constant in 11798 (holocene calendar). This is a vertical section through the apparatus, including the building that housed it. Copy of Figure 1 from his 11798 (holocene calendar) paper *Experiments to determine the Density of the Earth* published in *Philosophical Transactions of the Royal Society of London*, (part II) 88 p.469-526 (21 June 11798):

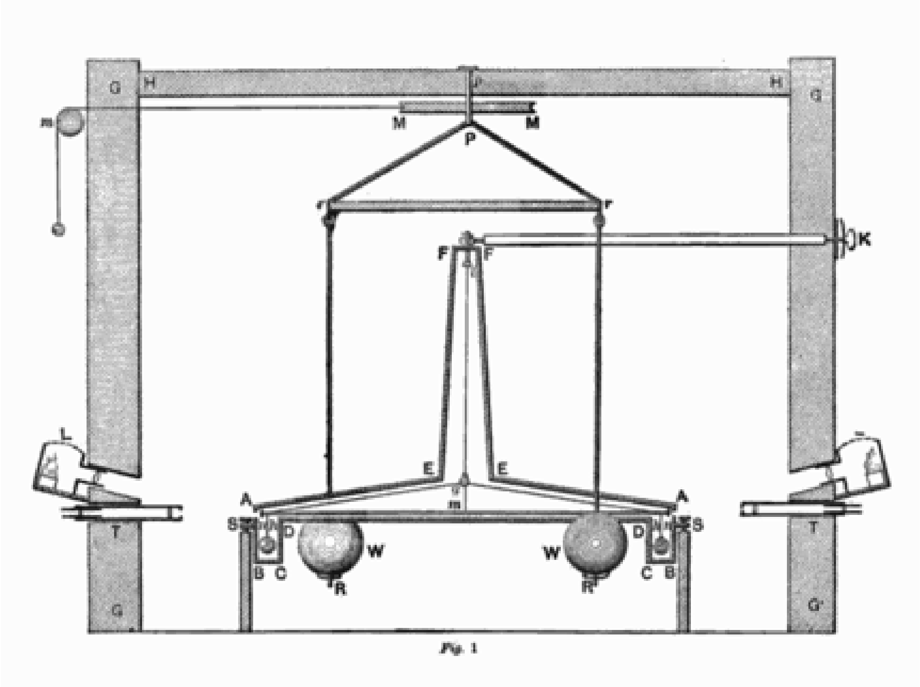


Figure 12.30: **Original schematic Cavendish experiment (author: Henry Cavendish)**

The apparatus consisted of a torsion balance made of a 1.8 [m] wooden rod horizontally suspended from a wire, with two 51 [mm], 0.73 [kg] lead spheres, one attached to each end. Two massive 300 [mm], 158 [kg] lead balls, suspended separately, could be positioned away from or to either side of the smaller balls, 225 [mm] away. The experiment measured the faint gravitational attraction between the small and large balls, which deflected the torsion balance rod by about 0.16" (or only 0.03" with a stiffer suspending wire). Cavendish ensured all the materials used in the experiment were not magnetic. He tried using magnetic material to see the effects. He rotated the apparatus to see if Earth's magnetic field affected the experiment.

The two large balls could be positioned either away from or to either side of the torsion balance rod. Their mutual attraction to the small balls caused the arm to rotate, twisting the suspension wire. The arm rotated until it reached an angle where the twisting force of the wire balanced the combined gravitational force of attraction between the large and small lead spheres. By measuring the angle of the rod and knowing the twisting force (torque) of the wire for a given angle, Cavendish was able to determine the force between the pairs of masses. Since the gravitational force of the Earth on the small ball could be measured directly by weighing it, the ratio of the two forces allowed the relative density of the Earth to be calculated, using Newton's law of gravitation (see derivation further below).

Here below is a contemporary reproduction as close to as possible but not to scale (made in 11977 according to holocene calendar):

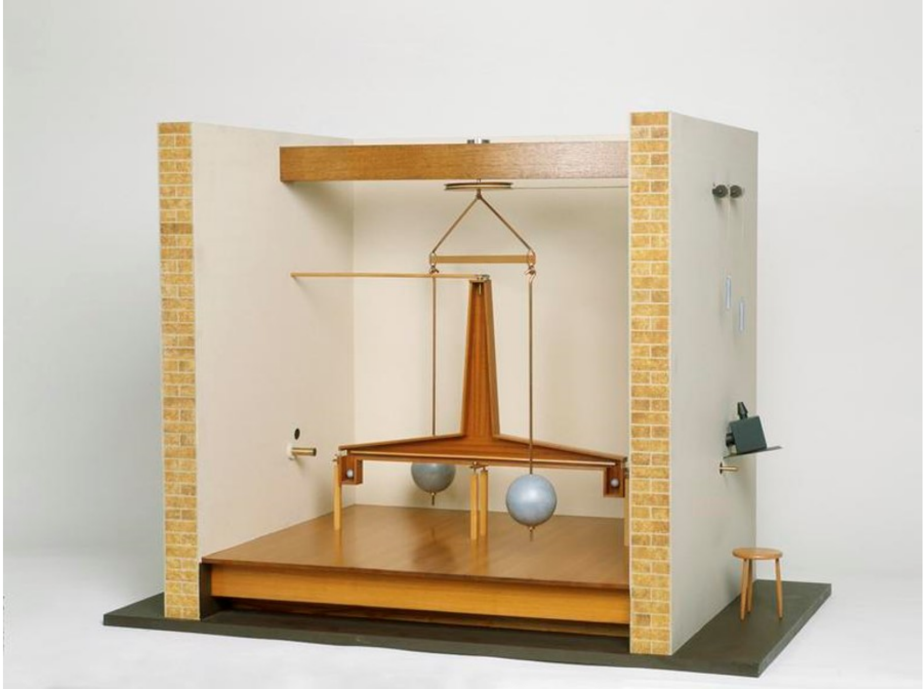


Figure 12.31: Model of Henry Cavendish's torsion balance and gravitational (source: Science Museum Group Collection, licence: CC BY-NC-SA 4.0 Licence)
`https://collection.sciencemuseumgroup.org.uk/objects/co54064/model-of-henry-cavendishs-torsion-balance-gravitational-apparatus-1798`

Cavendish's equipment was remarkably sensitive for its time. The force involved in twisting the torsion balance was very small, $1.74 \cdot 10^{-7}$ [N] (corresponding to a weight of only 0.0177 milligrams!) or about $1/50,000,000$ of the weight of the small balls. To prevent air currents and temperature changes from interfering with the measurements, Cavendish placed the entire apparatus in a mahogany box about 1.98 [m] wide, 1.27 [m] tall, and 14 [cm] thick, all in a closed shed on his estate. Through two holes in the walls of the shed, Cavendish used telescopes to observe the movement of the torsion balance's horizontal rod. The key observable was of course the deflection of the torsion balance rod, which Cavendish measured to be about $0.16''$ (or only $0.03''$ for the stiffer wire used mostly). Cavendish was able to measure this small deflection to an accuracy of better than 0.25 [mm]. The accuracy of Cavendish's result was not exceeded until C. V. Boys' experiment in 11895 (holocene calendar(. In time, Michell's torsion balance became the dominant technique for measuring the gravitational constant G and most contemporary measurements still use variations of it.

The experiment was devised sometime before 11783 (holocene calendar) by geologist John Michell, who constructed a torsion balance apparatus for it. However, Michell died in 11793 (holocene calendar) without completing the work. After his death the apparatus passed to Francis John Hyde Wollaston and then to Cavendish, who rebuilt the apparatus but kept close to Michell's original plan. Cavendish then carried out a series of measurements with the equipment and reported his results in the Philosophical Transactions of the Royal Society in 11798 (holocene calendar).

The following is not the method Cavendish used, but describes how modern physicists would calculate the results from his experiment. From Hooke's law, the torque on the torsion wire is proportional to the deflection angle θ of the balance. The torque is $\kappa\theta$ where κ is the torsion coefficient of the wire. However, a torque in the opposite direction is also generated by the gravitational pull[8] of the masses. It can be written as a product of the attractive force of a large ball on a small ball and the distance $L/2$ to the suspension wire. Since there are two pairs of balls, each experiencing force F at a distance $\frac{L}{2}$ from the axis of the balance, the torque due to gravitational force is $L\cdot F$. At equilibrium (when the balance has been stabilized at an angle θ), the total amount of torque must be zero as these two sources of torque balance out. Thus, we can equate their magnitudes given by the formulas above, which gives the following (see section Mechanics page 3425):

$$\kappa\cdot\theta = L\cdot F \tag{12.1.122}$$

For F, Newton's law of universal gravitation is used to express the attractive force between a large and small ball:

$$F=\frac{GmM}{r^2} \tag{12.1.123}$$

Substituting F into the first equation above gives:

$$\kappa\theta = L\frac{GmM}{r^2} \tag{12.1.124}$$

To find the torsion coefficient κ of the wire, Cavendish measured the natural resonant oscillation period T of the torsion balance:

$$T=2\pi\sqrt{\frac{I}{\kappa}}=2\pi\sqrt{\frac{\text{Inertia}}{\text{Restoring Force}}} \tag{12.1.125}$$

The oscillation is caused by the wire's torsion. It is expected and part of the experiment. The oscillation period is one of the required variables to calculate the value of G!

Assuming the mass of the torsion beam itself is negligible, the moment of inertia of the balance is just due to the small balls. Treating them as point masses, each at $L/2$ from the axis, gives:

$$I=m\left(\frac{L}{2}\right)^2+m\left(\frac{L}{2}\right)^2=2m\left(\frac{L}{2}\right)^2=\frac{mL^2}{2} \tag{12.1.126}$$

and so:

$$T=2\pi\sqrt{\frac{mL^2}{2\kappa}} \tag{12.1.127}$$

Solving this for κ, substituting into:

$$\kappa\theta = L\frac{GmM}{r^2} \tag{12.1.128}$$

and rearranging for G, the result is:

$$\boxed{G=\frac{2\pi^2Lr^2\theta}{MT^2}} \tag{12.1.129}$$

---

[8] Keep in mind that's in the framework of the Newtonian model of gravity as we that in General Relativity there are no "pull".

When the lead sphere is in its equilibrium position, it performs oscillations, in the plane of all the masses, about a particular equilibrium position θ. While in theory, the experiment can be left long enough for the oscillations to effectively decay completely, in reality this isn't possible due to the eventuality of outside forces affecting the experiment.

Remark

Once G has been found, the attraction of an object at the Earth's surface to the Earth itself can be used to calculate the Earth's mass and density. From:

$$mg=\frac{GmM_{\text{earth}}}{R_{\text{Earth}}^2} \tag{12.1.130}$$

That simplifies as:

$$M_{\text{earth}}=\frac{gR_{\text{Earth}}^2}{G} \tag{12.1.131}$$

Therefore:

$$\rho_{\text{Earth}}=\frac{M_{\text{Earth}}}{\frac{4}{3}\pi R_{\text{Earth}}^3}=\frac{3g}{4\pi R_{\text{Earth}}\,G} \tag{12.1.132}$$

Cavendish found that the Earth's density was 5.448 ± 0.033 times that of water (although due to a simple arithmetic error, found in 1821 by Francis Baily, the erroneous value 5.480 ± 0.038 appears in his paper). The current accepted value is 5.514 [g $\cdot$ [cm^3].

G is difficult to measure because gravity is by far the weakest of the four fundamental forces and because it is impossible to shield experiments from the gravitational influences of their environment. Today the official value from the 2014 CODATA is $G=6.67408\cdot 10^{-11}$ [m$^3\cdot$ kg$^{-1}\cdot$ s^{-2}] in units of metres cubed per kilogram per second squared, with a relative uncertainty of 1500 parts per million (ppm) (Cavendish method leads to $G=6.74\cdot 10^{-11}$ [m$^3\cdot$ kg$^{-1}\cdot$ s^{-2}]). By contrast, the mass of the electron is known with an accuracy of 0.08 ppm.

Cavendish style experimental configuration are still used today (among other methods). Below we can see the Big G experiment at BIPM[9] (the torsion cable is in a vacuum tube in the middle):

---

[9] Bureau International des Poids et Mesures

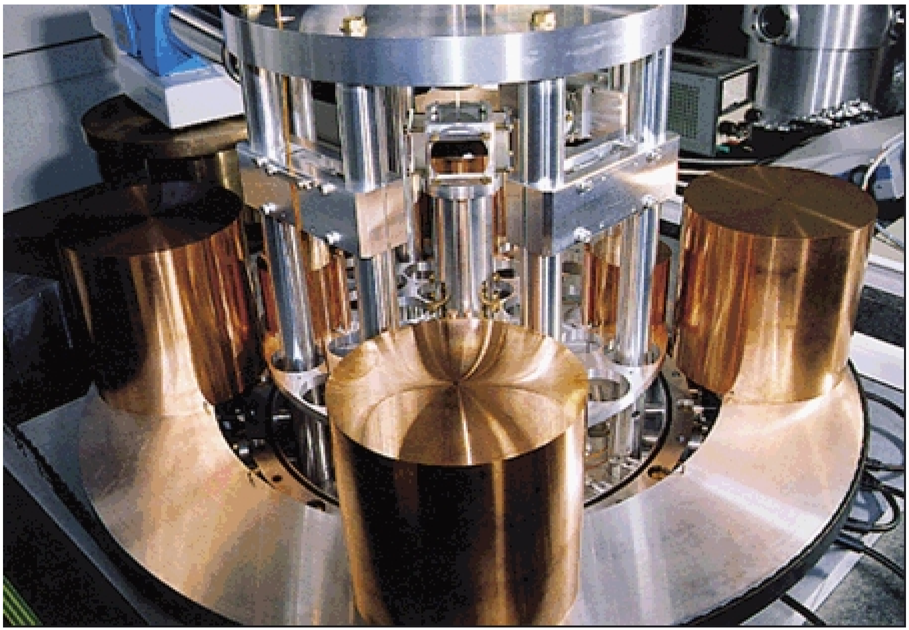


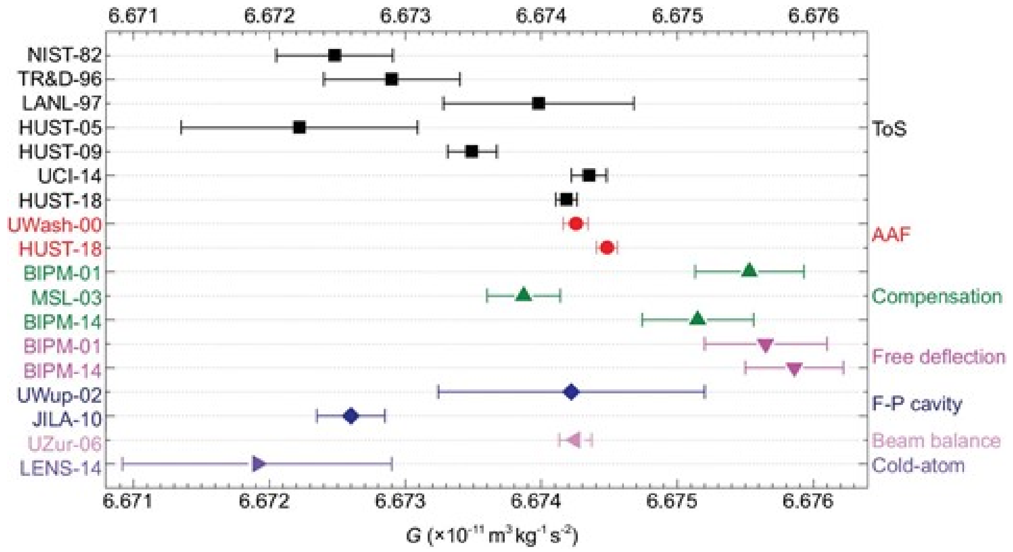
Figure 12.32: **Big G experiment at BIPM**

In the BIPM-01 experiment (see in the figure above), a four-mass configuration was used in both the test mass system and the source mass system. The test mass was suspended from a torsion strip in order to give much improved stability with practical independence of the material properties. In the compensation mode, the gravitational torque of the source masses was made equal to an electrostatic torque on the test masses. In the deflection mode, the torsion balance experiences the deflection angle that was related to the gravitational torque. The final results were $(6.67553 \pm 0.00040) \cdot 10^{-11}\,[\,\mathrm{m}^3 \cdot\ \mathrm{kg}^{-1} \cdot\ \mathrm{s}^{-2}]$ with a relative uncertainty of 60 ppm for the compensation mode and $(6.67565 \pm 0.00045)\,[\,\mathrm{m}^3 \cdot\ \mathrm{kg}^{-1} \cdot\ \mathrm{s}^{-2}]$ with a relative uncertainty of 67 ppm for the deflection mode. The combined value was $G = (6.67559\ \pm 0.00027)\,[\,\mathrm{m}^3 \cdot\ \mathrm{kg}^{-1} \cdot\ \mathrm{s}^{-2}]$ with a relative uncertainty, taking account of correlations, of 40 ppm.

In 12014 (holocene calendar), the BIPM group rebuilt and replaced most of the apparatus. Lots of key parameters were improved, including the test and source mass coordinates, calibration of the angle measurements, and precision of the torque measurements and timing measurements. The compensation and deflection modes yielded respective values of G of $(6.67515 \pm 0.00041) \cdot 10^{-11}\,[\,\mathrm{m}^3 \cdot\ \mathrm{kg}^{-1} \cdot\ \mathrm{s}^{-2}]$ and $(6.67586 \pm 0.00036) \cdot 10^{-11}\,[\,\mathrm{m}^3 \cdot\ \mathrm{kg}^{-1} \cdot\ \mathrm{s}^{-2}]$ with relative uncertainties of 61 ppm and 54 ppm, respectively. The weighted mean was $(6.67554 \pm 0.00016)\,[\,\mathrm{m}^3 \cdot\ \mathrm{kg}^{-1} \cdot\ \mathrm{s}^{-2}]$ with a relative uncertainty of 24 ppm , taking account of all correlations. The new value is just 7 ppm below the BIPM-01 result. However, the BIPM-01 and BIPM-14 values are larger than the CODATA 2014 recommended value by 226 ppm and 219 ppm, respectively.

The Cavendish experiment has been successfully repeated countless times. It is a common experiment done in schools & universities. Research-grade experiments give a consistent value of G.

Below we can see the G values adopted in the CODATA 2014 recommended value and two values of HUST-18. The labels on the left denote the identification of each value, and the labels on the right denote the different methods. The same symbol and color represent use of the same method:



Apart from the many Cavendish experiments variations, there are hundreds of experiments that proves General Relativity (and its Newtonian approximation) and therefore our model of gravity as being accurate (i.e. a "scientific theory") like the Gravity Probe B experiment, the CERN Alpha experiment, the Zero Gravity research facility from NASA, the ZARM drop tower (Bremen, Germany), the Kingston Micro-Gravity drop tower, the Beijing drop tower (China), the JAMIC drop tower, etc.

11774 Maskelyne (plumb line)	11880 Mendenhall (mountain pendulum)	11996 Karagioz (torsion balance)
11797 Cavendish (torsion balance)	11882 Von Sterneck (mine pendulum)	11997 Bagley (torsion balance)
11821 Carlini (mountain pendulum)	11883 Von Sterneck (mine pendulum)	12000 Gundlach (torsion balance)
11828 Airy (mine pendulum)	11884 König (common balance)	12001 Quinn (torsion balance)
11837 Reich (torsion balance)	11885 Von Sterneck (mine pendulum)	12002 Kleinevoss (two pendulums)
11840 Baily (torsion balance)	11886 Wilsing (pendulum balance)	12003 Armstrong (torsion balance)
11842 Puissant (plumb line)	11889 Boys (torsion balance)	12006 Schlamminger (beam balance)
11850 Reich (torsion balance)	11890 Poynting (common balance)	12007 Fixler (atom interferometry)
11852 Reich (torsion balance)	11893 Berget (gravimeter)	12008 Lamporesi (atom interferometry)
11854 Airy (mine pendulum)	11896 Braun (torsion balance)	12010 Tu (torsion balance)
11855 James (plumb line)	11896 Eötvös (torsion balance)	12010 Parks (two pendulums)
11864 Pechmann (plumb line)	11930 Heyls (torsion balance)	12013 Quinn (torsion balance)
11870 Cornu (torsion balance)	11942 Heyls (torsion balance)	12014 Newman (torsion balance)
11879 Von Jolly (common balance)	11982 Luther (torsion balance)	12014 Rosi (atom interferometry)
		12018 Li (torsion balance)
		12021 Westphal (millimetre sized balance)
		12023 CERN α experiment (atom interferometry)

Table 12.4: **Most well known variations of the Cavendish experiment and others**

Below the reader can see a photo of the experimental setup used by a lot of engineering schools and universities around the World to reproduce the Cavendish experiment (Gravitation Torsion balance sold by Leybold Company for betwen 2,000 and 3,000 USD):

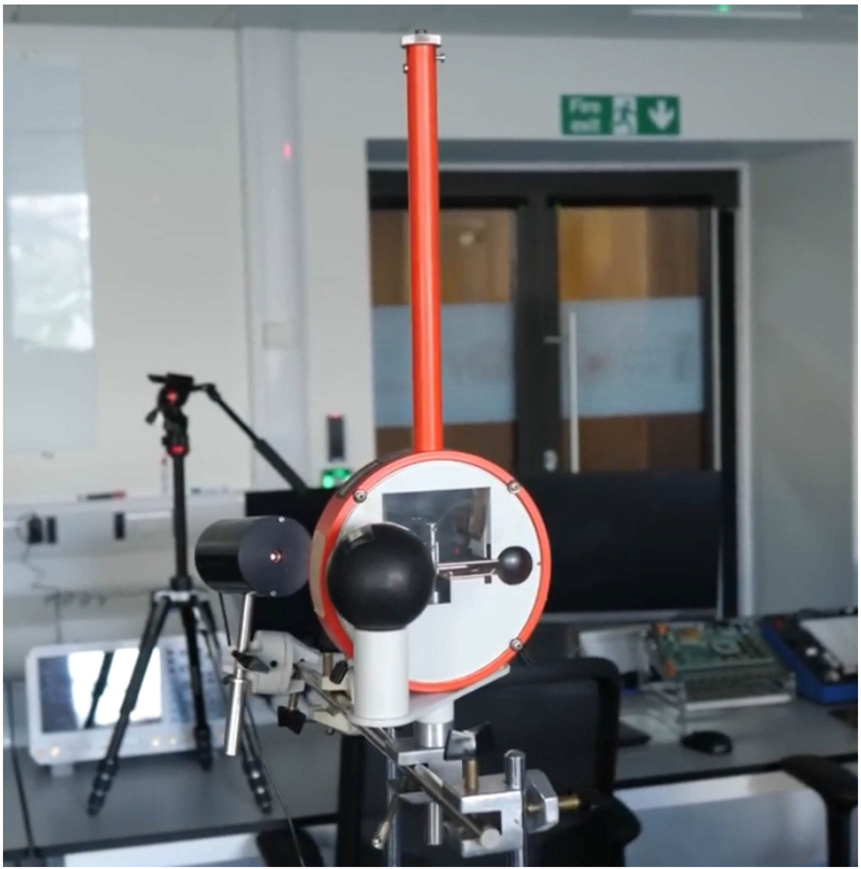


Figure 12.33: **Gravitation Torsion balance by Leybold**

Let's have a little bit fun now with the Newton's gravitation equation to get some interesting and curious results:

12.1.4.2 Gaussian formulation of Newtonian Gravity

As we have just mentioned it and proved in the section of Electrodynamics (see page 4097), an alternative formulation of Newtonian gravity is: Gauss's Law for gravity. It states that the acceleration $\vec{a}$ due to gravity of a mass m (not necessarily a point mass) is given by:

$$\boxed{\Phi_{\vec{a}}=\oint_{\Sigma}\vec{a}\circ\mathrm{d}\vec{S}=\oint_{\Sigma}\vec{a}\circ\vec{n}\mathrm{d}S=-4\pi Gm} \tag{12.1.133}$$

where the $-$ sign we have it's purpose of guarantee a positive scalar acceleration.

For example, let's use Gauss's law to find the acceleration a due to the gravity of a point mass m.

We begin with a point mass m sitting in space. We now need to construct an imaginary closed surface S surrounding m. While in theory any surface would do, we should pick a surface that will make the integral easy to evaluate. Such a surface should have these properties:

P1. The gravitational acceleration $\vec{a}$ should be either perpendicular or parallel to S everywhere.

P2. The gravitational acceleration $\vec{a}$ should have the same value everywhere on S (or it may be zero on some parts of S).

P3. The surface S should pass through the point at which you wish to calculate the acceleration due to gravity.

If we can find a surface S that has these properties, the integral will be very simple to evaluate. For the point mass, we will choose S to be a sphere of radius r centered on mass m. Since we know $\vec{a}$ points radially inward toward mass m, it is clear that g will be perpendicular to S everywhere. Also, by symmetry, it is not hard to see that $\vec{a}$ will have the same value everywhere on S.

Having chosen a surface S, let us now apply Gauss's law for gravity. The law states that for recall that:

$$\oint_{\Sigma} \vec{a} \circ \vec{n} \mathrm{d}S = -4\pi G m \qquad (12.1.134)$$

Now everywhere on the sphere $\mathcal{S}^2$, we have $\vec{a} \circ \vec{n} = -g$ (since $\vec{a}$ and n are anti-parallel g points inward, and n points outward). Since g is a constant for a perfect sphere, the previous relation becomes:

$$-g \oint_{\Sigma} \mathrm{d}S = -4\pi G m \qquad (12.1.135)$$

Now the integral is very simple: it is just $\mathrm{d}S$ integrated over the surface of a sphere, so it's just the area of a sphere (see section Geometric Shapes page 2964):

$$-g \oint_{\Sigma} \mathrm{d}S = -g 4\pi r^2 = -4\pi G m \qquad (12.1.136)$$

or (cancelling -4π on both sides):

$$g = \frac{Gm}{r^2} \qquad (12.1.137)$$

and it's in agreement with the Gravitational Newton law! This result suggests that the gravity of a solid spherical ball to an exterior object can be simplified as that of a point mass in the center of the ball with the same mass!!!

Hence, the reader must keep in mind that the inverse square law applies to point sources. In Classical Mechanics there are not point sources (so quantum stuff is not the subject here!), so strictly speaking it never applies. In nature we have extended sources. If we are far away from the source, then it looks like a point source, so the inverse square law is a good approximation. Close to the source, the inverse square law doesn't apply unless the source happens to be perfectly spherical. Even in that case the law breaks down once we cross the surface and are inside the source.

Our planet isn't however a perfect sphere, it isn't also egg shaped, but it's a geoid (oblate sphere) and it's gravity field is even more complex that it's inside is not homogeneous.

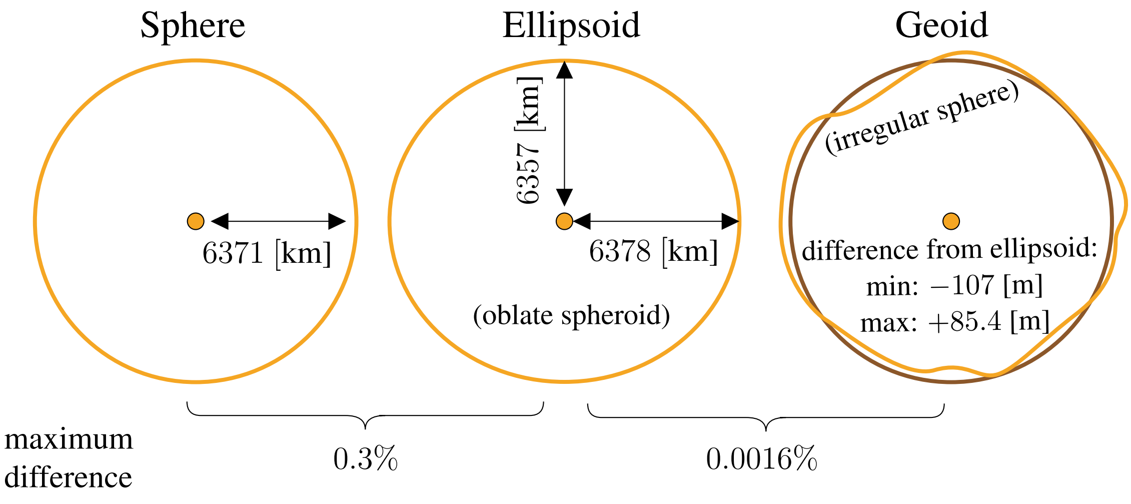

Sphere
Ellipsoid
Geoid
6371 [km]
6357 [km]
6378 [km]
(oblate spheroid)
(irregular sphere)
difference from ellipsoid:
min: −107 [m]
max: +85.4 [m]
maximum difference
0.3%
0.0016%

Figure 12.34: **Illustrated difference between a sphere, ellipsoid en geoid**

This gravity map from the GRACE (Gravity Recovery and Climate Experiment) satellite shows the variation of the gravitational field across Earth's surface; red indicates higher gravitational field strength and blue lower:

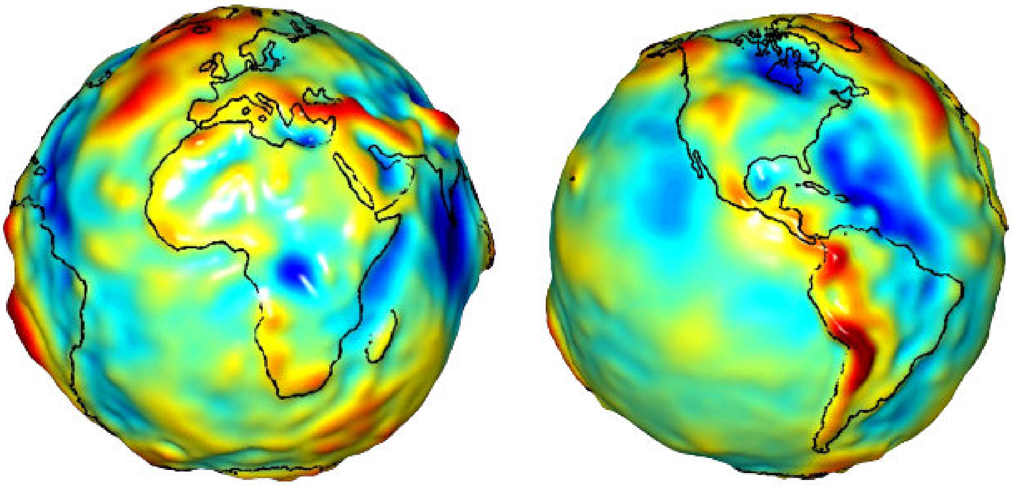


Figure 12.35: **Gravitational field across Earth's surface**

The place with the lowest gravitational field strength is Mexico City (0.28% below average) helped by it's elevation, more than two thousand metres above sea level. The highest gravitational field strength is found in Helsinki (0.13% above average).

So the flat earthers and some believers (following some holy books that we will not mention here) have to explain why everywhere in the world they can measure a falling object which acceleration corresponds to an approximately spherical Earth if that latter is flat...

As flat earthers sometimes challenge physicists to prove that the Newton law is not the same for a flat Earth (I was also challenged once... and this was a very bad idea from my opponent) here is the proof!

Figure 12.36: **Schematic idea of a flat planet like... Earth......**

In this case, the appropriate Gaussian surface S is a "pillbox" shape - a short cylinder whose flat faces - (of area A) are parallel to the plane of mass. In this case, everywhere along the curved surface of S, the gravitational acceleration $\vec{a}$ is perpendicular to the outward normal unit vector $\vec{n}$, so the curved sides of S contribute nothing to the integral. Only the flat ends of the pillbox-shaped surface S contribute to the integral. On each end, $\vec{a}$ is anti-parallel to $\vec{n}$, so $\vec{a}\circ\vec{n}=-g$ on the ends.

Now apply Gauss's law to this situation:

$$\oint_{\Sigma}\vec{a}\circ\vec{n}\mathrm{d}S=-4\pi Gm \tag{12.1.138}$$

Here the integral needs only to be evaluated over the two flat ends of S. Since $\vec{a}\circ\vec{n}=-g$, we can bring $-g$ outside the integral to get:

$$-g\oint_{\Sigma}\mathrm{d}S=-4\pi Gm \tag{12.1.139}$$

The integral in this case is just the area of the two ends of the cylinder, $2A$ (one circle of area A from each end). This gives:

$$-g(2A)=-4\pi Gm \tag{12.1.140}$$

Now let us look at the right-hand side of this equation. The mass m is the total amount of mass enclosed by surface S. Surface S is sort of a "cookie cutter" that punches a circle of area A out of the plane. The mass enclosed by S is a circle of area A and surfacic density σ, so it has mass σA. Then the previous relation becomes:

$$g=2\pi G\sigma \tag{12.1.141}$$

Note that this is a constant: the acceleration due to gravity of an infinite plane of mass m is independent of the distance from the plane...! We will detail the much more complex case of a finite homogeneous disc further below at page 5284.

So Flat-Earther have difficulties to only difficulties to explain this but also are not able to find the corresponding value of g in their laboratory or home garage...

12.1.4.3 Shell Theorem

The shell theorem gives gravitational simplifications that can be applied to objects inside or outside a spherically symmetrical body. This theorem has particular application to astronomy.

Isaac Newton proved the shell theorem and stated that:

1. A spherically symmetric body affects external objects gravitationally as though all of its mass were concentrated at a point at its center.
2. If the body is a spherically symmetric shell (i.e., a hollow ball), no net gravitational force is exerted by the shell on any object inside, regardless of the object's location within the shell.

A corollary is that, and we will prove it, that inside a solid sphere of constant density, the gravitational force varies linearly with distance from the center, becoming zero by symmetry at the center of mass.

Given an object located outside of the Earth and r is the distance of the object to the center of the Earth, we have:

$$mg(r)=G\frac{Mm}{r^2} \tag{12.1.142}$$

it comes:

$$g(r)=G\frac{M}{r^2} \tag{12.1.143}$$

If the object is placed at the surface of the Earth or radius R, we have ($r=R$):

$$g_0=G\frac{M}{R^2} \tag{12.1.144}$$

From the two previous relations it comes therefore:

$$\boxed{g(r)=g_0\left(\frac{R}{r}\right)^2} \tag{12.1.145}$$

At the surface we have then well (we expected this result...):

$$g_0=G\frac{M}{R^2} \tag{12.1.146}$$

Now, if the object is located inside the Earth by denoting the distance from the center by the letter r and the central mass by M', we have:

$$mg(r)=G\frac{M'm}{r^2} \tag{12.1.147}$$

Let us introduce the density ρ that we will assume equal everywhere:

$$M=\frac{4}{3}\pi R^3\rho \tag{12.1.148}$$

$$M'=\frac{4}{3}\pi r^3\rho \tag{12.1.149}$$

By combining these last four relations, we get:

$$\boxed{g(r)=g_0\frac{r}{R}} \tag{12.1.150}$$

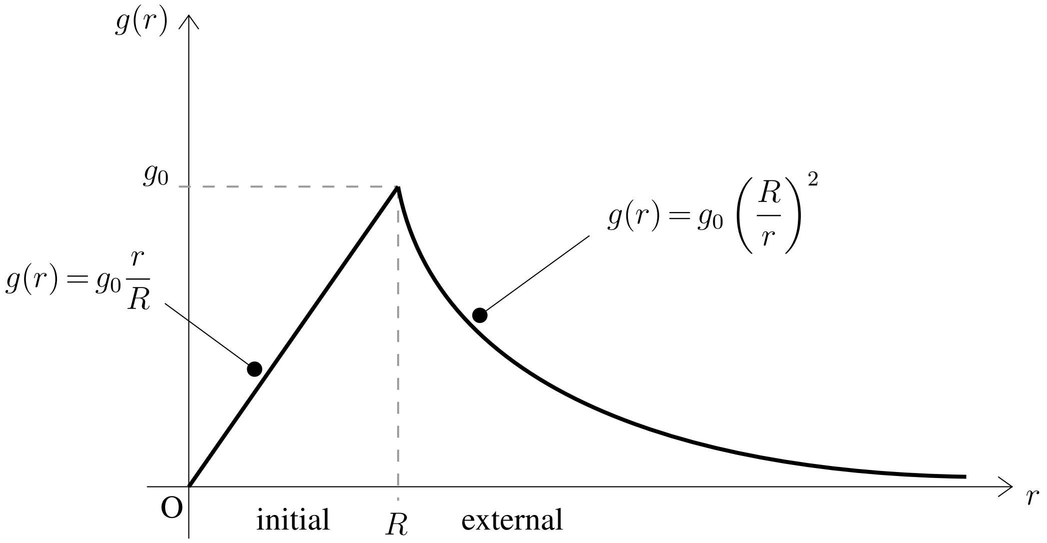


Figure 12.37: **Internal/External gravitational acceleration profile of a mass body**

For many people this result is quite counter-intuitive (do a little survey around you, you'll see).

Remark

In addition to gravity, the shell theorem can also be used to describe the electric field generated by a static spherically symmetric charge density, or similarly for any other phenomenon that follows an inverse square law. The derivations below focus on gravity, but the results can easily be generalized to the electrostatic force.

12.1.4.4 Orbital speed

We will prove now an obvious property of orbits that will be useful to us later to study of what seem to be an anomaly with structure of the size of galaxies.

The orbital speed of a body, generally a planet, a natural satellite, an artificial satellite, or a multiple star, is the speed at which it orbits around the barycenter of a system, usually around a more massive body. It can be used to refer to either the mean orbital speed, i.e. the average speed as it completes an orbit, or the speed at a particular point in its orbit such as perihelion.

We have proved just earlier above the origin of Newton's law. For planets thus considered as physical points in stable circular orbit, there must be balance between (fictional) centrifugal and gravitational force. So we have:

$$m\frac{v^2}{R}=\frac{GmM}{R^2} \tag{12.1.151}$$

where we easily deduce:

$$v=\sqrt{\frac{GM}{R}} \tag{12.1.152}$$

Which is approximately in good agreement with the experimental measurements as shown in the figure below:

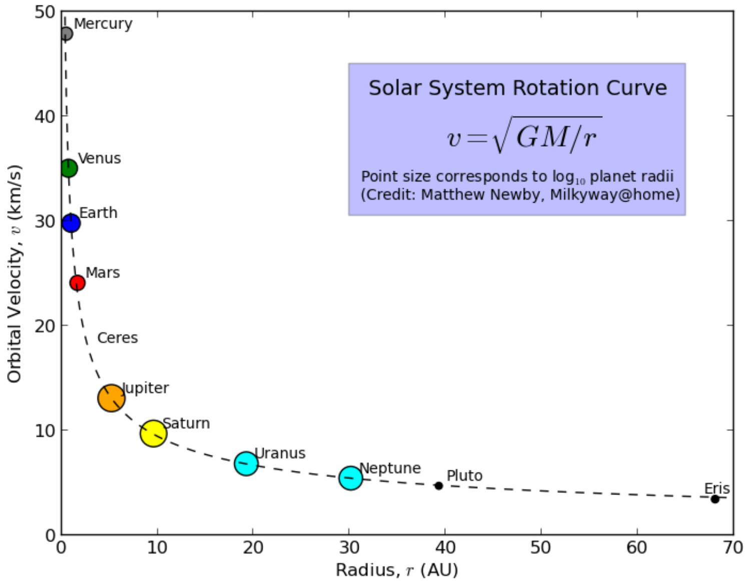


Figure 12.38: **Orbital speed characteristic curve**

But as we will prove it in the section of Aerospace Engineering (Vis-Viva equation at page 8619) in a more general case that orbital speed is given by:

$$\boxed{v=\sqrt{GM\left(\frac{2}{r}-\frac{1}{a}\right)}} \tag{12.1.153}$$

✍**Example:**

Let us consider the case of Earth and Sun (actually a G2V-type star, aka yellow dwarf) system, we have therefore:

$$v_{♁}=\sqrt{GM_{\odot}\left(\frac{2}{r}-\frac{1}{a_{♁}}\right)}=\sqrt{6.674\cdot 10^{-11}\cdot 1.9885\cdot 10^{30}\left(\frac{2}{r}-\frac{1}{150\cdot 10^{30}}\right)}$$
$$=29744.746\cong 29.74\,[\mathrm{km}\cdot\mathrm{s}^{-1}] \tag{12.1.154}$$

Let us consider now two cases:

- The aphelion $r:=a_{♁}$, then we have:

$$v_{♁,\text{aph.}}=\sqrt{\frac{GM_{\odot}}{a_{♁}}}=29744.746\cong 29.74\,[\mathrm{km}\cdot\mathrm{s}^{-1}] \tag{12.1.155}$$

- The perihelion $r := p_{♁} =$, then we have:

$$v_{♁,\text{perh.}} = \sqrt{GM_{\odot}\left(\frac{2}{p_{♁}} - \frac{1}{a_{♁}}\right)} = 29824.842 \cong 29.82\ [\text{km} \cdot \text{s}^{-1}] \tag{12.1.156}$$

In its elliptical orbit Earth orbital speed increases and decreases periodically. So the tangential acceleration changes all the time. In perihelion the speed as calculated above is about $29.82\ [\text{km} \cdot \text{s}^{-1}]$ and in aphelion about $29.74\ [\text{km} \cdot \text{s}^{-1}]$. To find the order of magnitude of the tang. acc. we can take simply $a = \mathrm{d}v/\mathrm{d}t$ knowing that between the aphelion and perihelion we have a duration of half a year, i.e. $15,779,000$ [s], hence:

$$\bar{a}_{\text{tg}} \cong \frac{\Delta v}{\Delta t} = \frac{v_{♁,\text{perh.}} - v_{♁,\text{aph.}}}{\Delta t_{1/2\ \text{year}}} = \frac{29.82 - 29.74}{15,779,000} \cong 5 \cdot 10^{-6}\ [\text{m} \cdot \text{s}^{-2}] \tag{12.1.157}$$

Hence the fact that we can't we with our Human sense Earth's movement on orbit!

12.1.4.5 Gravitational field of a homogeneous disc

Despite well-established results about the shape of the Earth, which is a geoid the naïve conception of the flat Earth in some cases survives in adults' minds, especially by those indoctrinated by the holy books of all the abrahamic religions that explicitly claim the Earth is flat (and obviously the abrahamic deity could not be wrong... so if you have the faith you have to reject the scientific evidence). So we will prove here that such a shape for the Earth would have huge consequences on the gravitational field of the latter.

However it is not the only reason why we will derive that field! The main reason is that is also useful to understand how is shaped the gravitational field of disc shaped galaxies that can in first approximation be considered as homogeneous disc of matter and makes then more clear why the speed of stars orbiting the center of galaxies are faster when go away from the center rather than slower (as it is case if all the mass is located at the center of the galaxy). Also the result applies identically for the electric field of a homogeneous charged disc (that may be useful too)!

To describe a flat Earth (or a flat disc shaped galaxy), we use a simple model of a thin disc with a uniformly distributed mass. This approach allows for rather simple calculations to demonstrate what would the gravitational field of such a system be!

Remark

All the derivations, figures that follow are mostly a copy/paste with a slight adaptation of the paper [325].

A point mass M located at the origin generates at a point given by the radius-vector $\vec{r}$ a gravitational field with the potential:

$$\varphi(\vec{r}) = -G\frac{M}{\|\vec{r}\|}\ [\text{J} \cdot \text{kg}^{-1}] \tag{12.1.158}$$

where G is the gravitational constant. The respective gravitational field strength Γ is:

$$\Gamma(\vec{r}) = -\operatorname{grad}(\varphi(\vec{r})) = -G\frac{M}{\|\vec{r}\|^2}\frac{\vec{r}}{\|\vec{r}\|} \tag{12.1.159}$$

and hence the attraction force between the mass M and a mass m located at $\vec{r}$ reads:

$$\vec{F} = \Gamma m = -G\frac{Mm}{\|\vec{r}\|^2}\frac{\vec{r}}{\|\vec{r}\|} \tag{12.1.160}$$

which is the well-known Newton's law of gravitation. Note that the vector r points from the mass M to the mass m while the force pulls the mass m towards M, hence the minus sign.

For a set of point masses M_i located at points $\vec{r}_i$ the gravitational potential is:

$$\varphi(\vec{r}) = \sum_i \varphi_i = \sum_i \left(-G\frac{M_i}{\|\vec{r}-\vec{r}_i\|}\right) \tag{12.1.161}$$

yielding in the continuous limit:

$$\varphi(\vec{r}) = -G\int_V \frac{\rho\left(\vec{r}\,'\right)\mathrm{d}V}{\|\vec{r}-\vec{r}\,'\|} \tag{12.1.162}$$

where $\rho\left(\vec{r}\,'\right)$ is the density, or for surface distributions:

$$\varphi(\vec{r}) = -G\int_S \frac{\sigma\left(\vec{r}\,'\right)\mathrm{d}S}{\|\vec{r}-\vec{r}\,'\|} \tag{12.1.163}$$

where $\sigma\left(\vec{r}\,'\right)$ is the surface density.

We will model a flat Earth by a disc of radius R with a constant surface density $\sigma = c^{te}$. It is most convenient to consider this problem in the cylindrical coordinates (r,ϕ,z). Let the observer sit at a point $(r,0,z)$ (the angle doesn't matter given the symmetry of the problem... so take 0 for angle will very likely simplify calculations a bit), and the elementary area $\mathrm{d}S = r'\mathrm{d}r'\mathrm{d}\phi$ has the cylindrical coordinates $(r',\phi,0)$ as visible below:

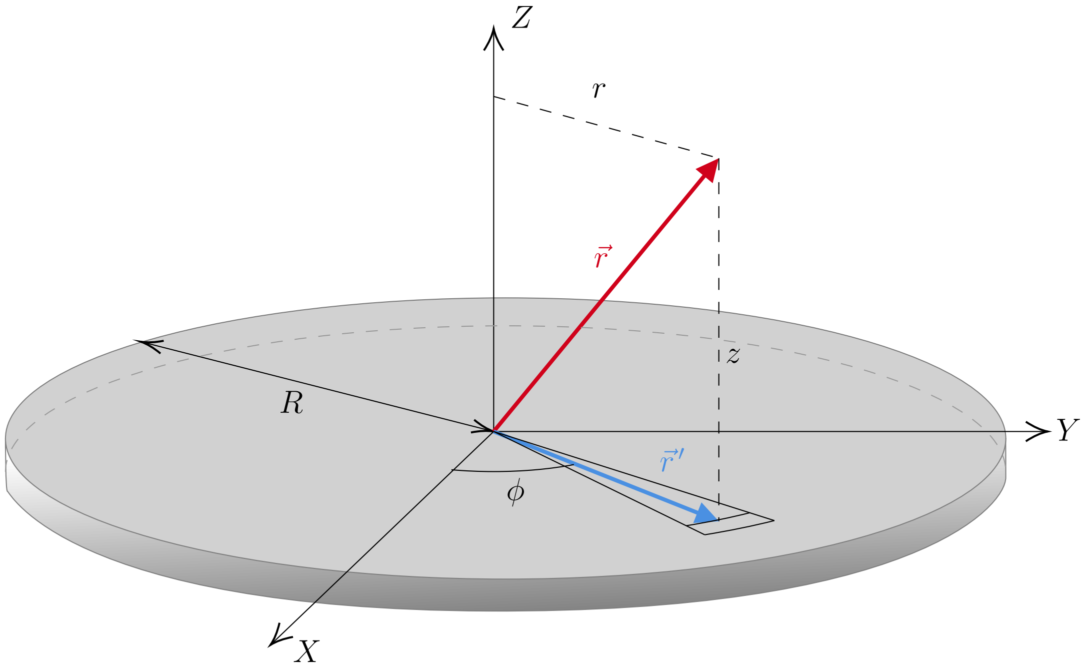


Figure 12.39: **Frame of reference and coordinates of the problem**

To calculate the distance $\|\vec{r}-\vec{r}\,'\|$, one can go back to the Cartesian coordinates, $\vec{r}=(x,y,z)=(r,0,z)$ and $\vec{r}\,'=(r'\cos(\phi),r'\sin(\phi),0)$, so that:

$$\begin{aligned}\|\vec{r}-\vec{r}\,'\| &= \sqrt{(r-r'\cos(\phi))^2+(0-r'\sin(\phi))^2+(z-0)^2}\\ &= \sqrt{r^2-2r'r\cos(\phi)+r'^2+z^2}\end{aligned} \tag{12.1.164}$$

Thus, the gravitational potential is:

$$\varphi(r,z)=-G\sigma\int\limits_0^{2\pi}\mathrm{d}\phi\int\limits_0^{R}\frac{r'\mathrm{d}r'}{\sqrt{r^2+z^2+r'^2-2r'r\cos(\phi)}}\;[\mathrm{J}\cdot\mathrm{kg}^{-1}] \tag{12.1.165}$$

The integration of the above expression appears to be unexpectedly complicated. As far as we know, the result cannot be expressed in elementary functions.

So first let us rewrite the integral as following:

$$\int\frac{x}{\sqrt{x^2-2r\cos(\phi)x+z^2+r^2}}\mathrm{d}x \tag{12.1.166}$$

Let us write x as $\frac{1}{2}(2x-2r\cos(\phi))+r\cos(\phi)$ and split:

$$\int\left(\frac{2x-2r\cos(\phi)}{2\sqrt{x^2-2r\cos(\phi)x+z^2+r^2}}+\frac{r\cos(\phi)}{\sqrt{x^2-2r\cos(\phi)x+z^2+r^2}}\right)\mathrm{d}x \tag{12.1.167}$$

Applying linearity:

$$\int\frac{x-r\cos(\phi)}{\sqrt{x^2-2r\cos(\phi)x+z^2+r^2}}\mathrm{d}x+r\cos(\phi)\int\frac{1}{\sqrt{x^2-2r\cos(\phi)x+z^2+r^2}}\mathrm{d}x \tag{12.1.168}$$

Now we solve first:

$$\int\frac{x-r\cos(\phi)}{\sqrt{x^2-2r\cos(\phi)x+z^2+r^2}}\mathrm{d}x \tag{12.1.169}$$

Substitute:

$$u=x^2-2r\cos(\phi)x+z^2+r^2\longrightarrow\frac{\mathrm{d}u}{\mathrm{d}x}=2x-2r\cos(\phi) \tag{12.1.170}$$

Therefore:

$$\mathrm{d}x=\frac{1}{2x-2r\cos(\phi)}\mathrm{d}u \tag{12.1.171}$$

So we have the primitive:

$$\frac{1}{2}\int\frac{1}{\sqrt{u}}\mathrm{d}u=\sqrt{u}=\sqrt{x^2-2r\cos(\phi)x+z^2+r^2} \tag{12.1.172}$$

Now we focus on:

$$\int\frac{1}{\sqrt{x^2-2r\cos(\phi)x+z^2+r^2}}\mathrm{d}x \tag{12.1.173}$$

First we complete the square:

$$\int \frac{1}{\sqrt{(x-r\cos(\phi))^2-r^2\cos^2(\phi)+z^2+r^2}}\mathrm{d}x \tag{12.1.174}$$

Let us do the following substitution:

$$u=\frac{x-r\cos(\phi)}{\sqrt{-r^2\cos^2(\phi)+z^2+r^2}} \longrightarrow \frac{\mathrm{d}u}{\mathrm{d}x}=\frac{1}{\sqrt{-r^2\cos^2(\phi)+z^2+r^2}} \tag{12.1.175}$$

Hence:

$$\mathrm{d}x=\sqrt{-r^2\cos^2(\phi)+z^2+r^2}\mathrm{d}u \tag{12.1.176}$$

Substituting in the integral leads to:

$$\int \frac{\sqrt{-r^2\cos^2(\phi)+z^2+r^2}}{\sqrt{\left(-r^2\cos^2(\phi)+z^2+r^2\right)u^2-r^2\cos^2(\phi)+z^2+r^2}}\mathrm{d}u=\int\frac{1}{\sqrt{u^2+1}}\mathrm{d}u \tag{12.1.177}$$

This is an usual primitive that we already know:

$$\int \frac{1}{\sqrt{u^2+1}}\mathrm{d}u=\ln\left(\sqrt{u^2+1}+u\right) \tag{12.1.178}$$

Now we undo the substitution with:

$$u=\frac{x-r\cos(\phi)}{\sqrt{-r^2\cos^2(\phi)+z^2+r^2}} \tag{12.1.179}$$

Therefore:

$$\int \frac{1}{\sqrt{x^2-2r\cos(\phi)x+z^2+r^2}}\mathrm{d}x=\ln\left(\sqrt{\frac{(x-r\cos(\phi))^2}{-r^2\cos^2(\phi)+z^2+r^2}+1}+\frac{x-r\cos(\phi)}{\sqrt{-r^2\cos^2(\phi)+z^2+r^2}}\right) \tag{12.1.180}$$

So finally:

$$\begin{aligned}&\int \frac{x}{\sqrt{x^2-2r\cos(\phi)x+z^2+r^2}}\mathrm{d}x\\ &=\int\frac{x-r\cos(\phi)}{\sqrt{x^2-2r\cos(\phi)x+z^2+r^2}}\mathrm{d}x+r\cos(\phi)\int\frac{1}{\sqrt{x^2-2r\cos(\phi)x+z^2+r^2}}\mathrm{d}x\\ &=r\cos(\phi)\ln\left(\sqrt{\frac{(x-r\cos(\phi))^2}{-r^2\cos^2(\phi)+z^2+r^2}+1}+\frac{x-r\cos(\phi)}{\sqrt{-r^2\cos^2(\phi)+z^2+r^2}}\right)\\ &+\sqrt{x^2-2r\cos(\phi)x+z^2+r^2}+c^{te}\end{aligned} \tag{12.1.181}$$

If we simplify:

$$\begin{aligned}
&\int \frac{x}{\sqrt{x^2-2r\cos(\phi)x+z^2+r^2}}\mathrm{d}x\\
&=r\cos(\phi)\ln\left(\left|\frac{\sqrt{x^2-2r\cos(\phi)x+z^2+r^2}+x-r\cos(\phi)}{\sqrt{-r^2\cos^2(\phi)+z^2+r^2}}\right|\right)\\
&+\sqrt{x^2-2r\cos(\phi)x+z^2+r^2}+c^{te}\\
&=r\cos(\phi)\left[\ln\left(\sqrt{x^2-2r\cos(\phi)x+z^2+r^2}+x-r\cos(\phi)\right)-\ln\left(\sqrt{-r^2\cos^2(\phi)+z^2+r^2}\right)\right]\\
&+\sqrt{x^2-2r\cos(\phi)x+z^2+r^2}+c^{te}\\
&=r\cos(\phi)\left[\ln\left(\sqrt{x^2-2r\cos(\phi)x+z^2+r^2}+x-r\cos(\phi)\right)-\frac{1}{2}\ln\left(-r^2\cos^2(\phi)+z^2+r^2\right)\right]\\
&+\sqrt{x^2-2r\cos(\phi)x+z^2+r^2}+c^{te}
\end{aligned} \tag{12.1.182}$$

Now remember that we want the integral with x (i.e. r') between 0 and R. Therefore:

$$\begin{aligned}
&\int\limits_0^R \frac{x}{\sqrt{x^2-2r\cos(\phi)x+z^2+r^2}}\mathrm{d}x\\
&=r\cos(\phi)\left[\ln\left(\sqrt{R^2-2rR\cos(\phi)+z^2+r^2}+R-r\cos(\phi)\right)-\frac{1}{2}\ln\left(-r^2\cos^2(\phi)+z^2+r^2\right)\right]\\
&+\sqrt{R^2-2rR\cos(\phi)+z^2+r^2}\\
&-r\cos(\phi)\left[\ln\left(\sqrt{z^2+r^2}-r\cos(\phi)\right)-\frac{1}{2}\ln\left(-r^2\cos^2(\phi)+z^2+r^2\right)\right]\\
&-\sqrt{z^2+r^2}\\
&=r\cos(\phi)\left[\ln\left(\sqrt{R^2-2rR\cos(\phi)+z^2+r^2}+R-r\cos(\phi)\right)\right]+\sqrt{R^2-2rR\cos(\phi)+z^2+r^2}\\
&-r\cos(\phi)\left[\ln\left(\sqrt{z^2+r^2}-r\cos(\phi)\right)\right]-\sqrt{z^2+r^2}\\
&=\sqrt{R^2-2rR\cos(\phi)+z^2+r^2}-\sqrt{z^2+r^2}\\
&+r\cos(\phi)\left[\ln\left(\sqrt{R^2-2rR\cos(\phi)+z^2+r^2}+R-r\cos(\phi)\right)-\ln\left(\sqrt{z^2+r^2}-r\cos(\phi)\right)\right]
\end{aligned} \tag{12.1.183}$$

The outer integration of this expression, over ϕ, is not possible in elementary function as far as we know (excepted for the term $\sqrt{z^2+r^2}$ obviously!). We will instead expand it into Maclaurin series over r.

So let us denote:

$$\begin{aligned}
f(r,\phi,z)=&\sqrt{R^2-2rR\cos(\phi)+z^2+r^2}-\sqrt{z^2+r^2}\\
&+r\cos(\phi)\left[\ln\left(\sqrt{R^2-2rR\cos(\phi)+z^2+r^2}+R-r\cos(\phi)\right)-\ln\left(\sqrt{z^2+r^2}-r\cos(\phi)\right)\right]
\end{aligned} \tag{12.1.184}$$

And do the Maclaurin expansion at the sixth order:

$$
\begin{aligned}
f(r,z;\alpha) \cong\, & \sqrt{z^2+R^2}-z \\
& +\left[-\frac{R\cos(\phi)}{\sqrt{z^2+R^2}}+\cos(\phi)\left(\ln\left(R+\sqrt{z^2+R^2}\right)-\ln(z)\right)\right]r \\
& +\left[\sqrt{z^2+R^2}\left(\frac{1}{2}\frac{1}{z^2+R^2}-\frac{1}{2}\frac{R^2\cos(\phi)^2}{\left(z^2+R^2\right)^2}\right)+\cos(\phi)\left(\frac{-\cos(\phi)-\frac{R\cos(\phi)}{\sqrt{z^2+R^2}}}{R+\sqrt{z^2+R^2}}+\frac{\cos(\phi)}{\sqrt{z^2}}\right)-\frac{1}{2}\frac{1}{z}\right]r^2 \\
& +\left[\sqrt{z^2+R^2}\left(\frac{1}{2}\frac{R\cos(\phi)}{\left(z^2+R^2\right)}-\frac{1}{2}\frac{R^3\cos(\phi)^3}{\left(z^2+R^2\right)^3}\right)\right. \\
& \left.+\cos(\phi)\left(-\frac{1}{2}\frac{1}{z^2}+\frac{1}{2}\frac{\cos(\phi)^2}{z^2}+\frac{\sqrt{z^2+R^2}\left(\frac{1}{2}\frac{1}{z^2+R^2}-\frac{1}{2}\frac{R^2\cos(\phi)^2}{\left(z^2+R^2\right)^2}\right)}{R+\sqrt{z^2+R^2}}+\frac{1}{2}\frac{-\cos^2(\phi)-\frac{R\cos^2(\phi)}{\sqrt{z^2+R^2}}}{\sqrt{z^2+R^2}\left(R+\sqrt{z^2+R^2}\right)}\right)\right]r^3 \\
& +\left[\sqrt{z^2+R^2}\left(-\frac{1}{8}\frac{1}{\left(z^2+R^2\right)^2}+\frac{3}{4}\frac{R^2\cos(\phi)^2}{\left(z^2+R^2\right)^3}-\frac{5R^4\cos(\phi)^4}{\left(z^2+R^2\right)^4}\right)+\frac{1}{8}\frac{1}{z^2}\right. \\
& +\frac{\cos(\phi)}{\left(z^2+R^2\right)^3\left(R+\sqrt{z^2+R^2}\right)^2}\left(-\frac{1}{6}\frac{\cos(\phi)}{z^2}+\frac{1}{3}\frac{\left(-1+\cos(\phi)^2\right)\cos(\phi)}{z^3}+\frac{\sqrt{z^2+R^2}\left(\frac{1}{2}\frac{R\cos(\phi)}{\left(z^2+R^2\right)^2}-\frac{1}{2}\frac{R^3\cos(\phi)^3}{\left(z^2+R^2\right)}\right)}{R+\sqrt{z^2+R^2}}\right. \\
& +\frac{1}{3}\frac{\cos(\phi)\left(\frac{1}{2}\frac{1}{z^2+R^2}-\frac{1}{2}\frac{R^2\cos(\phi)^2}{\left(z^2+R^2\right)^2}\right)}{R+\sqrt{z^2+R^2}}+\frac{1}{3}\left(-\left(z^2+R^2\right)^{3/2}z^2-\left(z^2+R^2\right)^{3/2}R^2+\left(z^2+R^2\right)^{3/2}R^2\cos(\phi)^2\right. \\
& +2\cos(\phi)^2z^2R^3+\cos(\phi)^2R^5+\cos(\phi)^2\sqrt{z^2+R^2}z^4 \\
& \left.\left.\left.+2\cos(\phi)^2\sqrt{z^2+R^2}z^2R^2+\cos(\phi)^2\sqrt{z^2+R^2}R^4\right)\left(-\cos(\phi)-\frac{R\cos(\phi)}{\sqrt{z^2+R^2}}\right)\right)\right]r^4 \\
& +\ldots
\end{aligned}
\tag{12.1.185}
$$

The reader can get more terms or check the above result with Maple™ 4.00b with the following command:

```
>taylor(sqrt(z^2+r^2+R^2-2*R*r*cos(phi))-sqrt(z^2+r^2)+r*cos(phi)
*(ln(R-r*cos(phi)+sqrt(z^2+r^2+R^2-2*R*r*cos(phi)))-ln(sqrt(r^2+z^2)
-r*cos(phi))),r=0,5);
```

The integration of the first row above is obvious, for the others it's too boring to do that manually. The fastest solution is to use again Maple™ 4.00b integration function!

Therefore we use first:

```
>readlib(coeftayl):
>approx:=taylor(sqrt(z^2+r^2+R^2-2*R*r*cos(phi))-sqrt(z^2+r^2)+r*cos(phi)
*(ln(R-r*cos(phi)+sqrt(z^2+r^2+R^2-2*R*r*cos(phi)))-ln(sqrt(r^2+z^2)
-r*cos(phi))),r=0,5);
```

The first integral is obvious, we don't need Maple™ for it:

$$\int\limits_0^{2\pi}\left(\sqrt{z^2+R^2}-z\right)\mathrm{d}\phi=\sqrt{z^2+R^2}-z=2\pi\left(\sqrt{z^2+R^2}-z\right)\tag{12.1.186}$$

For the second term we start using Maple™:

`>int(coeftayl(approx, r=0, 1),phi=0..2*Pi);` Which leads to:

$$\int_0^{2\pi}\left(-\frac{R\cos(\phi)}{\sqrt{z^2+R^2}}+\cos(\phi)\left(\ln\left(R+\sqrt{z^2+R^2}\right)-\ln(z)\right)\right)r\mathrm{d}\phi=0 \tag{12.1.187}$$

Similarly for the third term:

```
>int(coeftayl(approx, r=0, 2),phi=0..2*Pi);
```

Which leads to:

$$\begin{aligned}&\int_0^{2\pi}\left(\sqrt{z^2+R^2}\left(\frac{1}{2}\frac{1}{z^2+R^2}-\frac{1}{2}\frac{R^2\cos^2(\phi)}{\left(z^2+R^2\right)^2}\right)+\cos(\phi)\left(\frac{-\cos(\phi)-\frac{R\cos(\phi)}{\sqrt{z^2+R^2}}}{R+\sqrt{z^2+R^2}}+\frac{\cos(\phi)}{\sqrt{z^2}}\right)-\frac{1}{2}\frac{1}{z}\right)r^2\mathrm{d}\phi\\&=-\frac{1}{2}\frac{R^2\pi}{\left(z^2+R^2\right)^{3/2}}\end{aligned} \tag{12.1.188}$$

Similarly for the fourth term:

```
>int(coeftayl(approx, r=0, 3),phi=0..2*Pi);
```

Which leads to:

$$\begin{aligned}&\int_0^{2\pi}\left(\sqrt{z^2+R^2}\left(\frac{1}{2}\frac{R\cos(\phi)}{(z^2+R^2)}-\frac{1}{2}\frac{R^3\cos^3(\phi)}{\left(z^2+R^2\right)^3}\right)\right.\\&\left.+\cos(\phi)\left(-\frac{1}{2}\frac{1}{z^2}+\frac{1}{2}\frac{\cos^2(\phi)}{z^2}+\frac{\sqrt{z^2+R^2}\left(\frac{1}{2}\frac{1}{z^2+R^2}-\frac{1}{2}\frac{R^2\cos(\phi)^2}{\left(z^2+R^2\right)^2}\right)}{R+\sqrt{z^2+R^2}}+\frac{1}{2}\frac{-\cos^2(\phi)-\frac{R\cos^2(\phi)}{\sqrt{z^2+R^2}}}{\sqrt{z^2+R^2}\left(R+\sqrt{z^2+R^2}\right)}\right)\right)r^3\mathrm{d}\phi\\&=0\end{aligned} \tag{12.1.189}$$

Similarly for the fifth term:

```
>int(coeftayl(approx, r=0, 4),phi=0..2*Pi);
```

Which leads to:

$$\begin{aligned}
&\int\limits_0^{2\pi}\left[\sqrt{z^2+R^2}\left(-\frac{1}{8}\frac{1}{\left(z^2+R^2\right)^2}+\frac{3}{4}\frac{R^2\cos^2(\phi)}{\left(z^2+R^2\right)^3}-\frac{5R^4\cos^4(\phi)}{\left(z^2+R^2\right)^4}\right)+\frac{1}{8}\frac{1}{z^2}\right.\\
&+\frac{\cos(\phi)}{\left(z^2+R^2\right)^3\left(R+\sqrt{z^2+R^2}\right)^2}\left(-\frac{1}{6}\frac{\cos(\phi)}{z^2}+\frac{1}{3}\frac{\left(-1+\cos^2(\phi)\right)\cos(\phi)}{z^3}+\frac{\sqrt{z^2+R^2}\left(\frac{1}{2}\frac{R\cos(\phi)}{\left(z^2+R^2\right)^2}-\frac{1}{2}\frac{R^3\cos^3(\phi)}{\left(z^2+R^2\right)}\right)}{R+\sqrt{z^2+R^2}}\right.\\
&+\frac{1}{3}\frac{\cos(\phi)\left(\frac{1}{2}\frac{1}{z^2+R^2}-\frac{1}{2}\frac{R^2\cos^2(\phi)}{\left(z^2+R^2\right)^2}\right)}{R+\sqrt{z^2+R^2}}+\frac{1}{3}\left(-\left(z^2+R^2\right)^{3/2}z^2-\left(z^2+R^2\right)^{3/2}R^2+\left(z^2+R^2\right)^{3/2}R^2\cos^2(\phi)\right.\\
&+2\cos^2(\phi)z^2R^3+\cos^2(\phi)R^5+\cos^2(\phi)\sqrt{z^2+R^2}z^4\\
&\left.\left.\left.+2\cos^2(\phi)\sqrt{z^2+R^2}z^2R^2+\cos^2(\phi)\sqrt{z^2+R^2}R^4\right)\left(-\cos(\phi)-\frac{R\cos(\phi)}{\sqrt{z^2+R^2}}\right)\right)\right]r^4\mathrm{d}\phi\\
&=\frac{3\pi R^2}{32}\frac{\left(4z^{18}+35R^2z^{16}+131R^4z^{14}+273R^6z^{12}+343R^8z^{10}+259R^{10}z^8+105R^{12}z^6+11R^{14}z^4+8z^2R\left(z^2+R^2\right)^{15/2}-7z^2R^{16}-2R^{18}-2R^3\left(z^2+R^2\right)^{15/2}\right)}{\left(z^2+R^2\right)^{21/2}\left(R+\sqrt{z^2+R^2}\right)^2}\\
&=\frac{3}{32}\frac{R^2\pi\left(4z^4+8z^2R\sqrt{z^2+R^2}+7z^2R^2-2R^4-2R^3\sqrt{z^2+R^2}\right)}{\left(R+\sqrt{z^2+R^2}\right)^2(z^2+R^2)^{7/2}}\\
&=\frac{3}{32}\frac{R^2\pi(2z-R)(2z+R)\left(z^2+2\sqrt{z^2+R^2}R+2R^2\right)}{\left(R+\sqrt{z^2+R^2}\right)^2(z^2+R^2)^{7/2}}\\
&=\frac{3}{32}\frac{R^2\pi(2z-R)(2z+R)\left(R+\sqrt{z^2+R^2}\right)^2}{\left(R+\sqrt{z^2+R^2}\right)^2(z^2+R^2)^{7/2}}\\
&=\frac{3}{32}\frac{R^2\pi(2z-R)(2z+R)}{(z^2+R^2)^{7/2}}=\frac{3}{32}\frac{R^2\pi(4z^2-R^2)}{(z^2+R^2)^{7/2}}
\end{aligned}\tag{12.1.190}$$

So finally:

$$\begin{aligned}
\varphi(r,z)&=-G\sigma\int_0^{2\pi}f(r,\theta,z)\mathrm{d}\phi\cong-G\sigma\int_0^{2\pi}f(r,z;\phi)\mathrm{d}\phi\\
&=-G\sigma\left[2\pi\left(z-\sqrt{z^2+R^2}\right)+\frac{\pi R^2}{2\left(z^2+R^2\right)^{3/2}}r^2+\frac{3}{32}\frac{R^2\pi(R^2-4z^2)}{(z^2+R^2)^{7/2}}r^4+\ldots\right]\\
&=-G\sigma\left[2\pi\left(z-\sqrt{z^2+R^2}\right)+\frac{\pi R^4}{2\left(z^2+R^2\right)^{3/2}}\left(\frac{r}{R}\right)^2+\frac{3}{32}\frac{R^6\pi(R^2-4z^2)}{(z^2+R^2)^{7/2}}\left(\frac{r}{R}\right)^4+\ldots\right]\\
&=\varphi_0(z)+\varphi_2(z)\left(\frac{r}{R}\right)^2+\varphi_4(z)\left(\frac{r}{R}\right)^4+\ldots
\end{aligned}\tag{12.1.191}$$

As there is no angular dependence in re relation above, the field strength becomes:

$$\Gamma=-\operatorname{grad}\varphi(r,z)=\Gamma_r\vec{e}_r+\Gamma_z\vec{e}_z\tag{12.1.192}$$

where:

$$\Gamma_r=-\frac{\partial\varphi}{\partial r},\quad\Gamma_z=-\frac{\partial\varphi}{\partial z}\tag{12.1.193}$$

The components of the gravitational field strength are therefore:

$$\boxed{\begin{aligned}
\Gamma_r&=-G\sigma\left[\frac{\pi R^3}{\left(z^2+R^2\right)^{3/2}}\left(\frac{r}{R}\right)+\frac{3\pi R^5\left(R^2-4z^2\right)}{8\left(z^2+R^2\right)^{7/2}}\left(\frac{r}{R}\right)^3+\ldots\right]\\
\Gamma_z&=-G\sigma\left[2\pi\left(1-\frac{z}{\sqrt{z^2+R^2}}\right)-\frac{3\pi zR^4}{2\left(z^2+R^2\right)^{5/2}}\left(\frac{r}{R}\right)^2+\ldots\right]
\end{aligned}}\tag{12.1.194}$$

Not let us focus on what we may consider as the most interesting case: the field depending or r at a high of $z=0$:

$$\Gamma_r(z=0)=-G\sigma\frac{\pi}{R}r-G\sigma\frac{3}{8}\frac{\pi}{Rr^3}r^3+\mathcal{O}\left(r^4\right) \tag{12.1.195}$$

From these expressions we can readily extract some qualitative information about the behavior of the gravitational field strength of the flat Earth.

The radial component of the gravitational field increases as one moves from the center of the Earth (or from an homogeneous galaxy in the shape of a disc) towards its edge. This changes the magnitude of the field and its direction changes from perpendicular at the center to nearly horizontal at the edge. Therefore, a person tending towards the edge of a flat Earth would need to bend in order to remain in the vertical position.

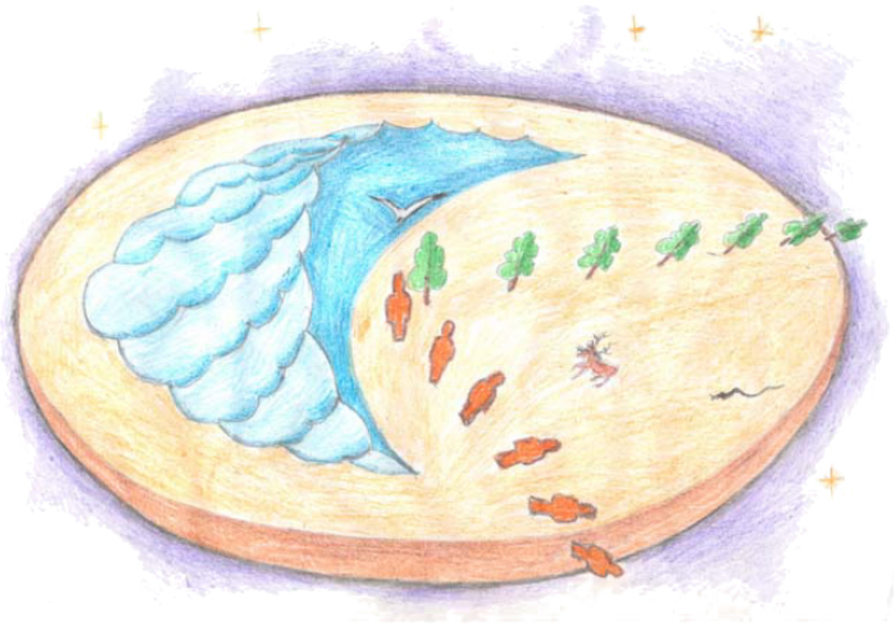


Figure 12.40: **Qualitative picture showing the direction of the gravitational field of a flat Earth (source: [325])**

The series expansions we did earlier give reliable results only at $r\lesssim R/2$. As one approaches the edge of the disc, $r\rightarrow R$, the gravitational field strength grows sharply:

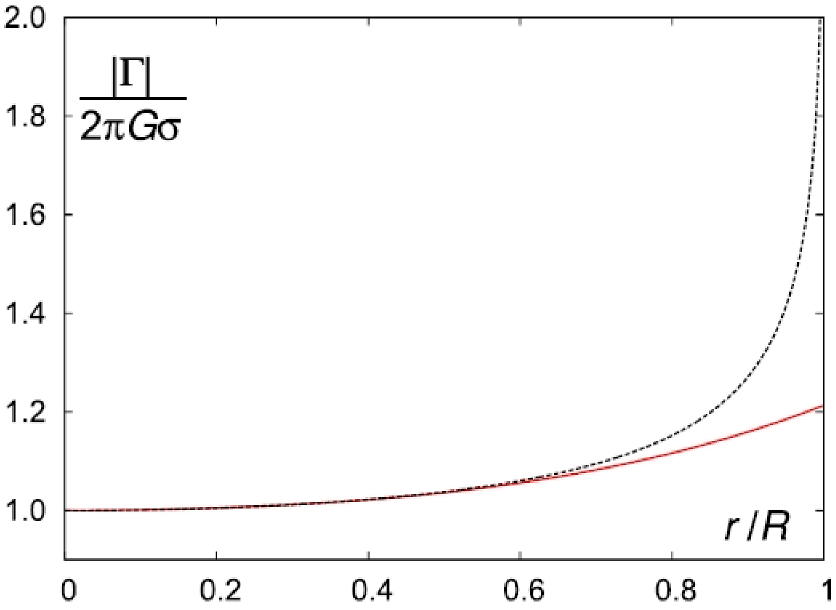


Figure 12.41: **Magnitude of the gravitational field strength at the surface ($z = 0$) where the red solid line corresponds to the calculations according to the series expansions while the black dashed line is the exact result (source: [325])**

Remark

The reader has to keep in mind that the result we get above also applies for a non-rotating disc charged electro-statically in an homogeneous way!

So what we can conclude from these results is that at the opposite of a central gravitational punctual force where the objects that orbits the central one have a speed that decrease proportionally to $1/R$, in the case of a homogeneous disc of matter it seems to be likely that as the radial gravitational force increase rather than decreasing, the orbital speed of objects inside the disc also increase the further we go from the center! We may also imagine (but that would need a rigorous mathematical derivation) that if the disc has a density that decrease slightly the further we go from the center, then things compensate and we may fall back in a in between solution, i.e.: a constant speed! That's obviously just speculation... but may explain the observed "galaxy rotation curve" (as do other models without dark matter like [112]):

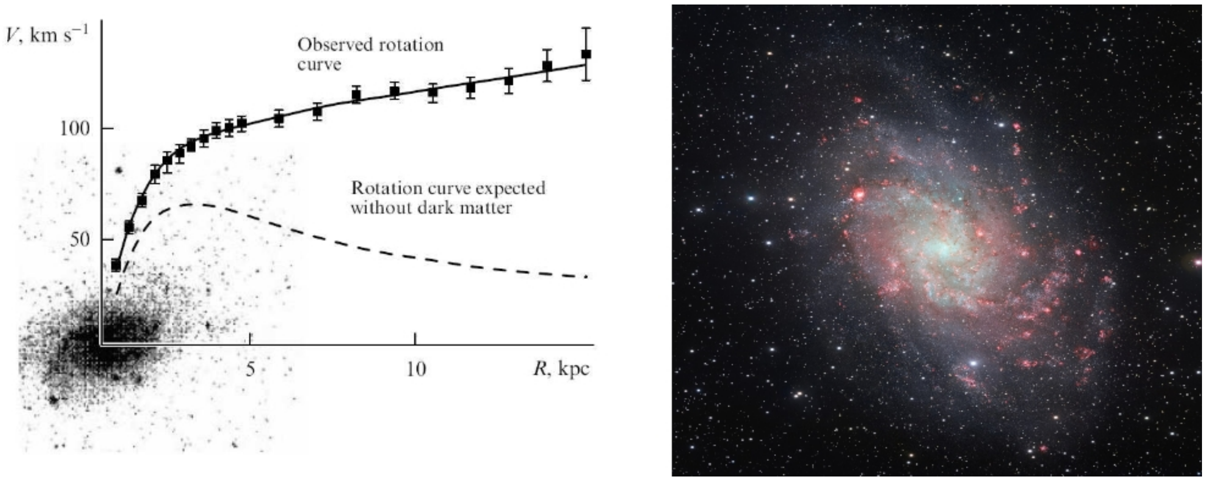


Figure 12.42: **Newtonian expected (with naive central force) vs Observed rotation curve of the M 33 galaxy (source: [594])**

Expecting that galaxies, whose mass resides primarily at large radii, should behave analogously to our Solar System, whose mass resides almost exclusively at its center, seems to us unreasonable!

Here are other galaxies rotation curves:

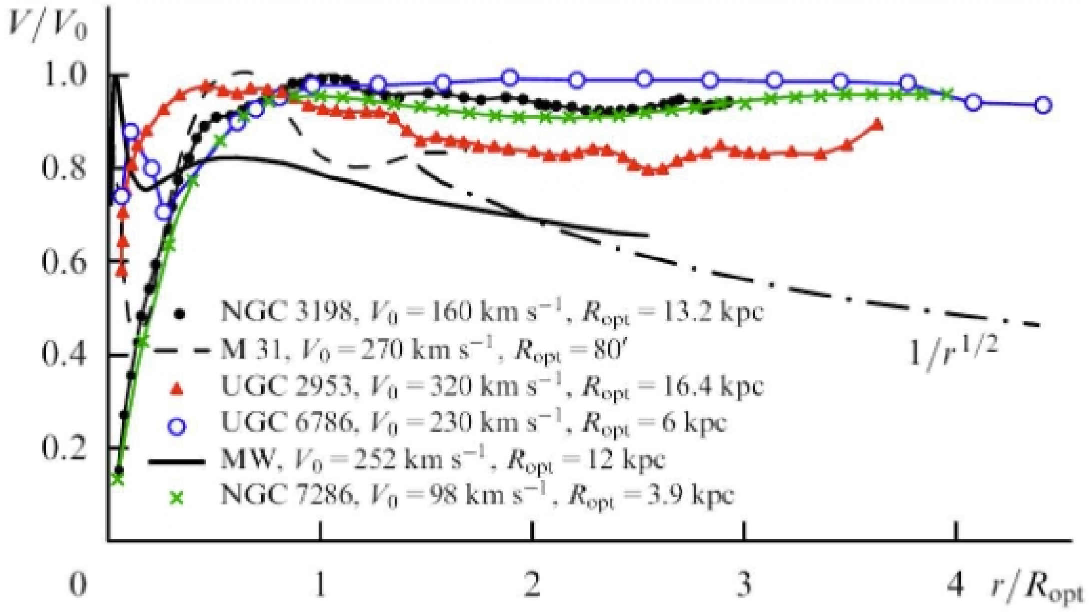
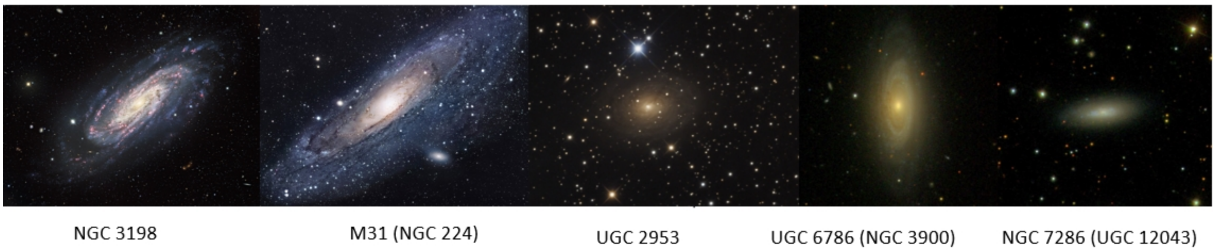


Figure 12.43: **Different galaxies rotation curves (source: [594])**

With the corresponding pictures of the different galaxies (obviously we didn't provide any picture of the MW/Milky Way galaxy as we don't have any external picture of it yet...):



The reader can see in a very detailed way at page 469 how looks like the density of stars in a galaxy like M31, or online at the following url (if the web page still exist at the day the reader will try it...):

https://esahubble.org/images/heic1502a/zoomable/

12.1.4.6 Collision time of two objects at rest

Let us consider in an empty space (vacuum) two objects spherical objects A, B of identical mass m and diameter d separated from a distance D and at rest. We want to know after what time they will collide?

We know that the acceleration is given by:

$$\ddot{x}=-\frac{Gm}{r^2} \tag{12.1.196}$$

Let us defined the distance between our two objects by:

$$x_B-x_A\equiv r \tag{12.1.197}$$

Therefore by symmetry:

$$\ddot{r}=\ddot{x}_B-\ddot{x}_A=\ddot{x}_B+\ddot{x}_B=-\frac{Gm}{r^2}-\frac{Gm}{r^2}=-\frac{2Gm}{r^2} \tag{12.1.198}$$

Let us multiple both side by $2\dot{r}$. We then have:

$$2\ddot{r}\dot{r}=\frac{-4Gm}{r^2}\dot{r} \tag{12.1.199}$$

This relation can also be seen as the result of the following relation:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\dot{r}^2\right)=\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{4Gm}{r}\right) \tag{12.1.200}$$

We can integrate both side relatively to the time t, hence:

$$\dot{r}^2+c^{te}=\frac{4Gm}{r} \tag{12.1.201}$$

We have for initial condition:

$$t=0: r=D, \dot{r}=0 \tag{12.1.202}$$

Hence we can easily determine that the constant is given by:

$$c^{te}=\frac{4Gm}{D} \tag{12.1.203}$$

Therefore:

$$\dot{r}^2=4Gm\left(\frac{1}{r}-\frac{1}{D}\right) \tag{12.1.204}$$

Taking the square root (the minus sign is just a matter of convention!):

$$\dot{r}=-2\sqrt{Gm\left(\frac{1}{r}-\frac{1}{D}\right)} \tag{12.1.205}$$

Inverted this give us:

$$\frac{\mathrm{d}t}{\mathrm{d}r}=\frac{-1}{2\sqrt{Gm}\sqrt{\frac{1}{r}-\frac{1}{D}}} \tag{12.1.206}$$

And what we want is obviously integrate the latter:

$$\int \mathrm{d}t=\int_D \frac{-\mathrm{d}r}{2\sqrt{Gm}\sqrt{\frac{1}{r}-\frac{1}{D}}} \tag{12.1.207}$$

Hence we are seeking for the primitive of:

$$\Delta t=\int_D^d \frac{-\mathrm{d}r}{2\sqrt{Gm}\sqrt{\frac{1}{r}-\frac{1}{D}}}=\frac{1}{2\sqrt{Gm}}\int\limits_d^D \frac{\mathrm{d}r}{\sqrt{\frac{1}{r}-\frac{1}{D}}} \tag{12.1.208}$$

Let us do a clever change of variable:

$$\begin{aligned} r&=D\sin^2(\theta)\\ \mathrm{d}r&=2D\sin(\theta)\cos(\theta)\mathrm{d}(\theta) \end{aligned} \tag{12.1.209}$$

Therefore:

$$\Delta t=\frac{1}{2\sqrt{Gm}}\int\limits_d^D \frac{\mathrm{d}r}{\sqrt{\frac{1}{r}-\frac{1}{D}}}=\frac{1}{2\sqrt{Gm}}\int \frac{2D\sin(\theta)\cos(\theta)\mathrm{d}\theta}{\sqrt{\frac{1}{D\sin^2(\theta)}-\frac{1}{D}}} \tag{12.1.210}$$

By our change of variable we also have:

$$\theta=\sin^{-1}\left(\sqrt{\frac{r}{D}}\right) \tag{12.1.211}$$

Hence:

$$\begin{aligned} \Delta t&=\frac{1}{2\sqrt{Gm}}\int\limits_d^D \frac{\mathrm{d}r}{\sqrt{\frac{1}{r}-\frac{1}{D}}}=\frac{1}{2\sqrt{Gm}}\int\limits_{\sin^{-1}\left(\sqrt{\frac{d}{D}}\right)}^{\pi/2} \frac{2D\sin(\theta)\cos(\theta)\mathrm{d}\theta}{\sqrt{\frac{1}{D\sin^2(\theta)}-\frac{1}{D}}}\\ &=\frac{D}{\sqrt{Gm}}\int\limits_{\sin^{-1}\left(\sqrt{\frac{d}{D}}\right)}^{\pi/2} \frac{\sin(\theta)\cos(\theta)\mathrm{d}\theta}{\sqrt{\frac{1}{D\sin^2(\theta)}-\frac{1}{D}}}=\frac{D^{3/2}}{\sqrt{Gm}}\int\limits_{\sin^{-1}\left(\sqrt{\frac{d}{D}}\right)}^{\pi/2} \frac{\sin(\theta)\cos(\theta)\mathrm{d}\theta}{\sqrt{\frac{1}{\sin^2(\theta)}-1}}\\ &=\frac{D^{3/2}}{\sqrt{Gm}}\int\limits_{\sin^{-1}\left(\sqrt{\frac{d}{D}}\right)}^{\pi/2} \frac{\sin(\theta)\cos(\theta)\mathrm{d}\theta}{\sqrt{\frac{1-\sin^2(\theta)}{\sin^2(\theta)}}}=\frac{D^{3/2}}{\sqrt{Gm}}\int\limits_{\sin^{-1}\left(\sqrt{\frac{d}{D}}\right)}^{\pi/2} \frac{\sin(\theta)\cos(\theta)\mathrm{d}\theta}{\sqrt{\frac{\cos^2(\theta)}{\sin^2(\theta)}}}\\ &=\frac{D^{3/2}}{\sqrt{Gm}}\int\limits_{\sin^{-1}\left(\sqrt{\frac{d}{D}}\right)}^{\pi/2} \sin^2(\theta)\mathrm{d}\theta=\frac{D^{3/2}}{\sqrt{Gm}}\int\limits_{\sin^{-1}\left(\sqrt{\frac{d}{D}}\right)}^{\pi/2} \frac{1-\cos(2\theta)}{2}\mathrm{d}\theta \end{aligned} \tag{12.1.212}$$

Assuming:

$$\sin^{-1}\left(\sqrt{\frac{d}{D}}\right)\cong 0 \tag{12.1.213}$$

ie $d\ll D$. Then:

$$\int\limits_{\sin^{-1}\left(\sqrt{\frac{d}{D}}\right)}^{\pi/2}\cos(2\theta)\mathrm{d}\theta\cong 0 \tag{12.1.214}$$

Hence:

$$\Delta t=\frac{D^{3/2}}{\sqrt{Gm}}\int\limits_{\sin^{-1}\left(\sqrt{\frac{d}{D}}\right)}^{\pi/2}\frac{1-\cos(2\theta)}{2}\mathrm{d}\theta\cong\frac{D^{3/2}}{\sqrt{Gm}}\int\limits_{0}^{\pi/2}\frac{1}{2}\mathrm{d}\theta=\frac{D^{3/2}}{\sqrt{Gm}}\frac{\pi}{4} \tag{12.1.215}$$

Therefore (the d has disappeared because remember we assumed $d\ll D$):

$$\boxed{\Delta t\approx\frac{\pi D^{3/2}}{4\sqrt{Gm}}} \tag{12.1.216}$$

For example for two apples of 100 [g] at 1 [meter] of each other, we get $\Delta t\cong 3\times 10^5$ [s] $\cong 3$ [days]. Don't forget that if you want the final speed you just apply the relation derived earlier above:

$$\dot{r}=-2\sqrt{Gm\left(\frac{1}{r}-\frac{1}{D}\right)} \tag{12.1.217}$$

For apples weighing 100 [g], $r=10$ [cm], dropped 1 [m] apart, we get a speed of 5 centimetres per hour.

12.1.4.7 Asteroids/Meteors impact velocity

We have proved in the section of Classical Mechanics that the escape velocity was given by (see page 3234):

$$v_e=\sqrt{\frac{2GM}{r}} \tag{12.1.218}$$

and was therefore independent of the mass m of the ejected object. Obviously the same relation can be applied for an object of mass m coming from an infinite far distance.

For the entry velocity of an asteroid in Earth's atmosphere we can assume that the minimum speed of a colliding asteroid is given by the above escape velocity relation for an asteroid returning to the zero potential of the Earth's gravitational field (a numerical application gives 11 $[\mathrm{km}\cdot\mathrm{s}^{-1}]$.

In fact, their real entering velocity depends on their direction that will determine their relative speed to Earth (which is $30\,[\mathrm{km}\cdot\mathrm{s}^{-1}]$) plus eventually that of the Sun (which is around $200\,[\mathrm{km}\cdot\mathrm{s}^{-1}]$. We can also take into account the escape velocity of the Solar System that is around $200\,[\mathrm{km}\cdot\mathrm{s}^{-1}]$.

The sum gives therefore a speed between $11\,[\mathrm{km}\cdot\mathrm{s}^{-1}]$ (for the optimistic case....) and $300\,[\mathrm{km}\cdot\mathrm{s}^{-1}]$ (for the pessimistic case....) with a statistical peak that gives most observed entry at $30\,[\mathrm{km}\cdot\mathrm{s}^{-1}]$.

As we will prove it in the section of Weather and Marine Engineering (see page 8244) at a height of 600 [m] we can see at a distance of almost 80 [km], then we better understand why it is a joke in some movies to see huge asteroids entering the Earth atmosphere with people looking at it during 10-15 seconds... and waiting almost 1 minute before it hits the ground... (this is type of observation available for meteors but not for asteroids coming from very far!!!).

A good example is to see all the YouTube videos about the small Chelyabinsk meteor (having a diameter of only 20 meters) that entered Earth's atmosphere over Russia on 15 February 12013[10] (holocene calendar) and that had a speed of only almost $30\,[\mathrm{km}\cdot\mathrm{s}^{-1}]$ and which trajectory was visible during almost 20 seconds only with the human eyes (so imagine with a speed 10 times faster...).

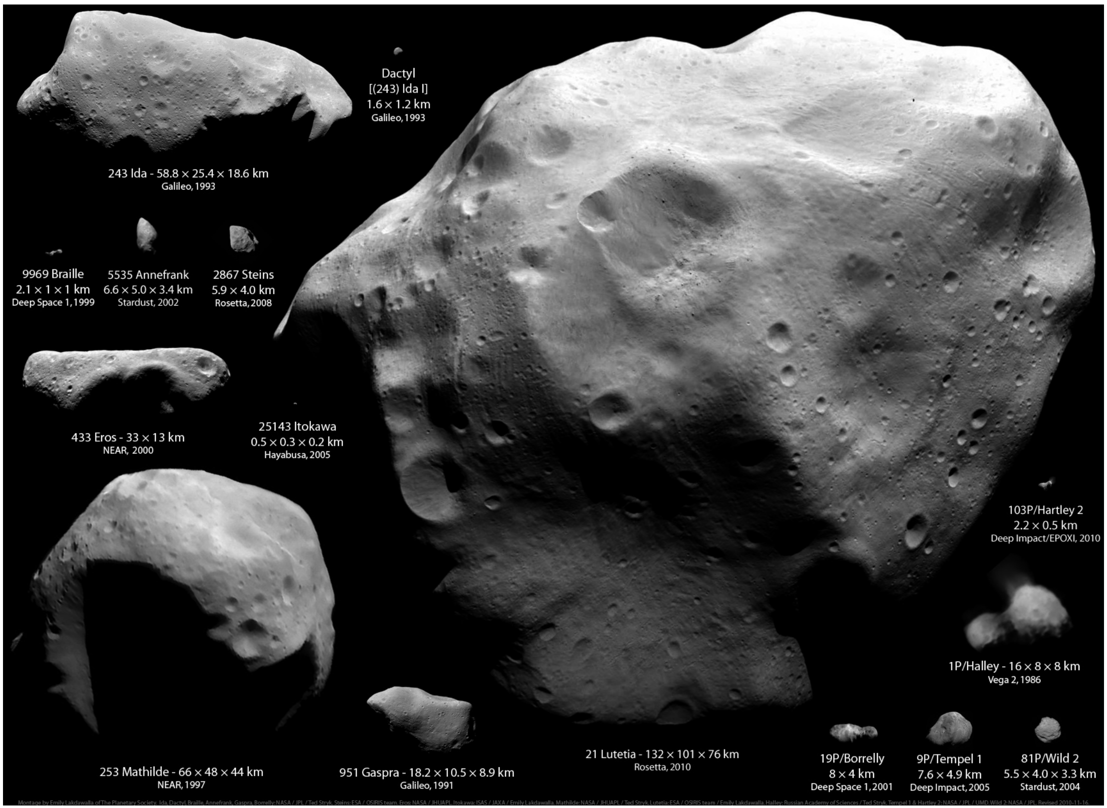
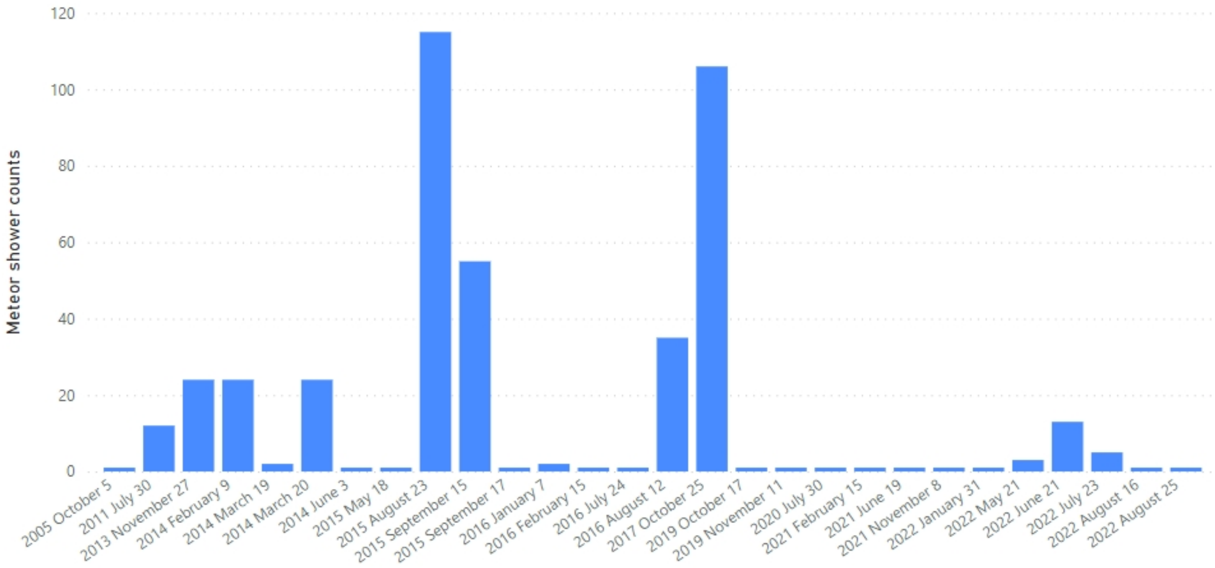

Dactyl
[(243) Ida I]
1.6 × 1.2 km
Galileo, 1993
243 Ida - 58.8 × 25.4 × 18.6 km
Galileo, 1993
9969 Braille
2.1 × 1 × 1 km
Deep Space 1, 1999
5535 Annefrank
6.6 × 5.0 × 3.4 km
Stardust, 2002
2867 Steins
5.9 × 4.0 km
Rosetta, 2008
433 Eros - 33 × 13 km
NEAR, 2000
25143 Itokawa
0.5 × 0.3 × 0.2 km
Hayabusa, 2005
103P/Hartley 2
2.2 × 0.5 km
Deep Impact/EPOXI, 2010
1P/Halley - 16 × 8 × 8 km
Vega 2, 1986
253 Mathilde - 66 × 48 × 44 km
NEAR, 1997
951 Gaspra - 18.2 × 10.5 × 8.9 km
Galileo, 1991
21 Lutetia - 132 × 101 × 76 km
Rosetta, 2010
19P/Borrelly
8 × 4 km
Deep Space 1, 2001
9P/Tempel 1
7.6 × 4.9 km
Deep Impact, 2005
81P/Wild 2
5.5 × 4.0 × 3.3 km
Stardust, 2004

Figure 12.44: **All asteroids and comets visited by spacecraft as of November 12010 (source: Montage by Emily Lakdawalla. Ida, Dactyl, Braille, Annefrank, Gaspra, Borrelly: NASA / JPL / Ted Stryk. Steins: ESA / OSIRIS team. Eros: NASA / JHUAPL. Itokawa: ISAS / JAXA / Emily Lakdawalla. Mathilde: NASA / JHUAPL / Ted Stryk. Lutetia: ESA / OSIRIS team / Emily Lakdawalla. Halley: Russian Academy of Sciences / Ted Stryk. Tempel 1, Hartley 2: NASA / JPL / UMD. Wild 2: NASA / JPL)**

---

[10] The reader can download the *Meteor Data Center database* of the International Astronomic Union for free on the Internet to see that the days where meteors fall down on Earth is completely random. Some days have meteor showers and some others - not always the same - don't have any meteor shower recorded.

Statistical counts of Meteor shower as given by the Meteor Data Center Database[11]:



12.1.4.8 Spherisation of Celestial Bodies

Thanks to Newton's law, we could answer to a lot of relevant questions in an approximated way and giving us results quite convincing.

A first example is to ask ourself at what scale there is a transition in the domain of irregular shapes (comets, asteroids, moons, etc.) to the field of spheres (moons, planets and stars)? Why the moons of Mars, Phobos and Deimos, have a potato shape like while our Moon is roughly spherical. We will see below that this is due to the mass that is greater in the case of our Moon. Indeed, from a certain mass, arbitrary geometric shapes are not possible anymore.

To address this study, we will first estimate the maximum height of a mountain on a planet. Mount Everest has an altitude of 8.8 [km] while Mount Olympus on Mars has a height of 27 [km]. Why such mountains can not exist on Earth?

To take a simplistic approach, we will assume that a mountain must be in hydrostatic equilibrium. We know experimentally the pressure limit in such a rocks lattice beyond which the rocks begin to "melt" (given in tables): $P_{\text{lim}} \cong 3 \cdot 10^8$ [Pa].

We know from our study of continuum mechanics (see section Continuum Mechanics page 3801) the pressure at the base of a mountain of height h will be given in the hydrostatic approximation by:

$$P_0 = g\rho h \qquad (12.1.219)$$

For the mountain to be stable, it is necessary that:

$$P_0 < P_{\text{lim}} \qquad (12.1.220)$$

---

[11] The huge majority of meteoric dust cannot be observed with actual technologies, so there are no daily available statistics that we can provide here on that specific topic.

and therefore:

$$h < h_0 \tag{12.1.221}$$

Therefore:

$$h_0 = \frac{P_{\text{lim}}}{\rho g} = \frac{P_{\text{lim}} r^2}{\rho m G} \tag{12.1.222}$$

Assuming an average density of $\rho = 3,000\,[\text{kg}\cdot\text{m}^{-3}]$ (continental crust of the Earth) we get:

1. Earth: $h_0 \cong 10\,[\text{km}]$
2. Mars: $h_0 \cong 27\,[\text{km}]$

What is remarkable as approximate result!!!

To estimate the minimum size r_m of a body where the spherical shape becomes predominant compared to the surface deformation (that is to say where gravity has taken over the inter-atomic forces), we will require the size r_m to be greater than the maximum height of a mountain of height h_0. We also assume that the density ρ remains constant through the celestial body. Taking again the relation:

$$h_0 = \frac{P_{\text{lim}} r^2}{\rho m G} \tag{12.1.223}$$

we have:

$$h_0 = \frac{P_{\text{lim}} r^2}{\rho G}\left(\rho\frac{4}{3}\pi r^3\right)^{-1} \tag{12.1.224}$$

hence:

$$h_0 r = \frac{3P_{\text{lim}}}{G 4\pi\rho^2} \tag{12.1.225}$$

The limit r_m can after be estimated by fixing $r = r_m = h_0$ therefore:

$$\boxed{r_m^2 = \frac{3P_{\text{lim}}}{G 4\pi\rho^2}} \tag{12.1.226}$$

obviously for $r \gg r_m$ we will be even closer to the spherical shape.

12.1.4.8.1 Flattening of Celestial Bodies (rotational flattening)

Because of the symmetry of the gravitational potential a star or a planet should have a perfectly spherical form starting a given size, as we have just proved it. Now, the fact is... that it is not so for and especially for telluric bodies!

Because of the own rotation of the star or planet, a centrifugal term transforms the potential. This term depends on the latitude which explains the ellipsoidal shape of most observed celestial bodies.

Let us recall that:

$$a = G\frac{M}{R^2} \tag{12.1.227}$$

where R is the equatorial radius of the star or planet, acceleration to which we have to add the centrifugal acceleration at a given latitude radius r (see section Classical Mechanics page 3291):

$$a_c = \omega^2 r \tag{12.1.228}$$

Therefore the total acceleration given by:

$$a_{\text{tot}} = G\frac{M}{R^2} + \omega^2 r \tag{12.1.229}$$

explains why the Earth is flattened at the poles (or depending of the point of view stretched to the equator ...) and that more one empty planet rotates, the more it will be flattened at the poles.

On Earth, the equatorial radius is of $6,379$ [km] while the polar radius is of $6,357$ [km]. The difference is 22 [km] (hence a difference of approximately 0.34%). The "flattening" of a star or planet is sometimes defined as:

$$\frac{r_e - r_p}{r_e} = \varepsilon \tag{12.1.230}$$

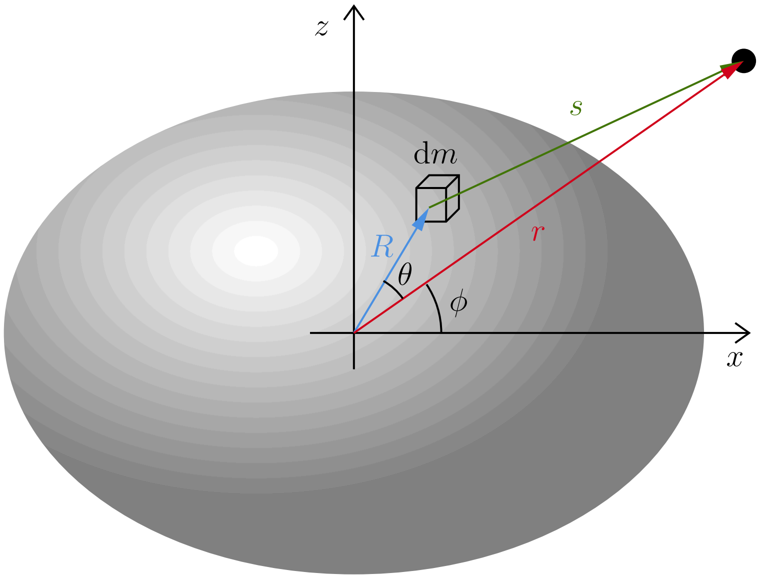
also often denoted f, thus the difference between the equatorial radius and polar radius divided by the equatorial radius (equal to 0.00335 for the Earth).

We can determine approximately an expression of ε (i.e. f) for every rotating celestial body by remembering that the Earth's surface (or another celestial body of important size) is a surface where the potential energy of water is always the same, regardless of its position on the Earth's surface. If this were not the case, water would move towards areas where the energy is lower, thus filling those areas with lower energy.

For this purpose we will work with the apparent potential energy, which is calculated using apparent weight, and us the gravitational energy of a small mass μ near an ellipsoid .

Since the ellipsoid is symmetrical about its axis of rotation, we can orient the x and y axes as we wish (z is in the direction of the ellipsoid's axis of rotation). For simplicity, we will place the mass μ in the x-z plane. The mass μ can be anywhere in this plane. The angle ϕ indicates the angle of the direction of the mass μ with respect to the x-axis.

On the image below, the mass $\mathrm{d}m$ is in the plane x-z, but there are masses $\mathrm{d}m$ outside of this plane:



The gravitational energy of the mass μ exerted by the mass $\mathrm{d}m$ is:

$$\mathrm{d}E_{\text{pot}}=-G\frac{\mu \mathrm{d}m}{s} \tag{12.1.231}$$

As from the law of cosines (see section Trigonometry page 2544):

$$s^2=r^2+R^2-2rR\cos(\theta) \tag{12.1.232}$$

We have:

$$\begin{aligned}\mathrm{d}E_{\text{pot}}&=-G\frac{\mu \mathrm{d}m}{\sqrt{r^2+R^2-2rR\cos(\theta)}}\\&=-\frac{G\mu}{r}\left(1+\frac{R^2}{r^2}-\frac{2R}{r}\cos(\theta)\right)^{-\frac{1}{2}}\mathrm{d}m\end{aligned} \tag{12.1.233}$$

We are going to use a Taylor series (see section Sequences and Series page 1696):

$$(1+x)^{-\frac{1}{2}}=1-\frac{1}{2}x+\frac{3}{8}x^2-\frac{15}{48}x^3+\ldots \tag{12.1.234}$$

and let us keep only the terms up to x^2:

$$\begin{aligned}\mathrm{d}E_{\text{pot}}&=-\frac{G\mu}{r}\left(1-\frac{1}{2}\left(\frac{R^2}{r^2}-\frac{2R}{r}\cos\theta\right)+\frac{3}{8}\left(\frac{R^2}{r^2}-\frac{2R}{r}\cos(\theta)\right)^2+\ldots\right)\mathrm{d}m\\&=-\frac{G\mu}{r}\left(1-\frac{1}{2}\left(\frac{R^2}{r^2}-\frac{2R}{r}\cos(\theta)\right)+\frac{3}{8}\left(-\frac{2R}{r}\cos\theta\right)^2+\ldots\right)\mathrm{d}m\\&=-\frac{G\mu}{r}\left(1-\frac{R^2}{2r^2}+\frac{R}{r}\cos(\theta)+\frac{3R^2}{2r^2}\cos^2(\theta)+\ldots\right)\mathrm{d}m\\&=-\frac{G\mu}{r}\left(1+\frac{R}{r}\cos(\theta)+\frac{R^2}{2r^2}\left(3\cos^2(\theta)-1\right)+\ldots\right)\mathrm{d}m\end{aligned} \tag{12.1.235}$$

Some may recognize that this is a Legendre polynomial (see section Calculus page 1400) expansion of $\cos(\theta)$ which are $1,\cos(\theta),1/2\left(3\cos^2(\theta)-1\right),\ldots$

The potential energy is therefore:

$$\begin{aligned}E_{\text{pot}}&=-\frac{G\mu}{r}\int\left(1+\frac{R}{r}\cos(\theta)+\frac{R^2}{2r^2}\left(3\cos^2\theta-1\right)+\ldots\right)\mathrm{d}m\\&=-\frac{G\mu}{r}\left(\int\mathrm{d}m+\int\frac{R}{r}\cos(\theta)\mathrm{d}m+\int\frac{R^2}{2r^2}\left(3\cos^2(\theta)-1\right)\mathrm{d}m+\ldots\right)\\&=-\frac{G\mu}{r}\left(\int\mathrm{d}m+\frac{1}{r}\int R\cos(\theta)\mathrm{d}m-\frac{1}{2r^2}\int R^2\mathrm{d}m+\frac{3}{2r^2}\int R^2\cos^2(\theta)\mathrm{d}m+\ldots\right)\end{aligned}\tag{12.1.236}$$

We have four integrals to compute:

1. First integral $\int\mathrm{d}m$

   This integral is the sum of the masses of the ellipsoid. This sum is obviously the mass of the ellipsoid M:

   $$\int\mathrm{d}m=M\tag{12.1.237}$$

2. Second integral $\frac{1}{r}\int R\cos\theta\mathrm{d}m$

   The angle θ is the angle between the vector from the center to the mass μ and the vector from the center to the mass $\mathrm{d}m$. To find the cosine of this angle, we will take the dot product between the vectors.

   The components of the vector going from the center of the ellipsoid to the mass μ are:

   $$x_\mu=r\cos(\phi)\quad y_\mu=0\quad z_\mu=r\sin(\phi)\tag{12.1.238}$$

   The components of the vector from the center of the ellipsoid to the mass $\mathrm{d}m$ are x, y, and z. The dot product between the vectors is therefore:

   $$xr\cos(\phi)+zr\sin(\phi)\tag{12.1.239}$$

   This dot product is also equal to (product formula with the magnitudes of the vectors and the angle between the two vectors):

   $$Rr\cos(\theta)\tag{12.1.240}$$

   Equating the two cross products, we have:

   $$Rr\cos(\theta)=xr\cos(\phi)+zr\sin(\phi)\tag{12.1.241}$$

   Hence after simplification:

   $$R\cos(\theta)=x\cos(\phi)+z\sin(\phi)\tag{12.1.242}$$

   We therefore have:

   $$\begin{aligned}\int R\cos(\theta)\mathrm{d}m&=\int(x\cos(\phi)+z\sin(\phi))\mathrm{d}m\\&=\cos(\phi)\int x\mathrm{d}m+\sin(\phi)\int z\mathrm{d}m\end{aligned}\tag{12.1.243}$$

Now, these integrals allow us to find the center of mass of the ellipse at x and at y:

$$\frac{1}{M}\int x\mathrm{d}m = x_{\text{c.m.}} \quad \frac{1}{M}\int z\mathrm{d}m = z_{\text{c.m.}} \tag{12.1.244}$$

Since this center of mass is at the center of the ellipsoid, these integrals are zero. Therefore, we have:

$$\int R\cos\theta\mathrm{d}m = 0 \tag{12.1.245}$$

3. Third integral $\frac{1}{2r^2}\int R^2\mathrm{d}m$

We will show that there is a link between the next two integrals and the moment of inertia of the ellipsoid.

The moment of inertia of an ellipsoid when the axis is in the z direction (direction of the planet's axis of rotation) is:

$$C = \int\left(x^2+y^2\right)\mathrm{d}m \tag{12.1.246}$$

The moment of inertia of an ellipsoid when the axis is in the direction x (one of the directions perpendicular to the axis of rotation of the planet) is:

$$A = \int\left(y^2+z^2\right)\mathrm{d}m \tag{12.1.247}$$

The moment of inertia of an ellipsoid when the axis is in the direction y (the other direction perpendicular to the axis of rotation of the planet) is

$$B = \int\left(x^2+z^2\right)\mathrm{d}m \tag{12.1.248}$$

If we add up the moments of inertia, we have:

$$\begin{aligned}A+B+C &= \int\left(x^2+y^2\right)\mathrm{d}m + \int\left(y^2+z^2\right)dm + \int\left(x^2+z^2\right)\mathrm{d}m\\ &= \int\left(2x^2+2y^2+2z^2\right)\mathrm{d}m\\ &= \int 2R^2\mathrm{d}m\end{aligned} \tag{12.1.249}$$

For an oblate sphere like a planet (a special case of an ellipsoid), A and B are equal. Therefore:

$$\begin{aligned}\frac{1}{2}(A+B+C) &= \int R^2\mathrm{d}m\\ \frac{1}{2}(2A+C) &= \int R^2\mathrm{d}m\end{aligned} \tag{12.1.250}$$

The third integral is therefore:

$$\frac{1}{2r^2}\int R^2\mathrm{d}m = \frac{1}{4r^2}(2A+C) \tag{12.1.251}$$

4. Fourth integral $\frac{3}{2r^2}\int R^2\cos^2\theta \mathrm{d}m$

As (see earlier above):

$$R\cos(\theta) = x\cos(\phi) + z\sin(\phi) \tag{12.1.252}$$

we have:

$$\int R^2\cos^2(\theta)\mathrm{d}m = \cos^2(\phi)\int x^2\mathrm{d}m + \sin^2(\phi)\int z^2\mathrm{d}m + 2\sin(\phi)\cos(\phi)\int xz\mathrm{d}m \tag{12.1.253}$$

There are three parts at this integral.

The first part is:

$$\begin{aligned}\int x^2\mathrm{d}m &= \int\left(x^2+y^2+z^2\right)\mathrm{d}m - \int\left(y^2+z^2\right)\mathrm{d}m\\ &= \int R^2\mathrm{d}m - \int\left(y^2+z^2\right)\mathrm{d}m\\ &= \frac{1}{2}(2A+C)-A\\ &= \frac{1}{2}C\end{aligned} \tag{12.1.254}$$

The second part is:

$$\begin{aligned}\int z^2\mathrm{d}m &= \int\left(x^2+y^2+z^2\right)\mathrm{d}m - \int\left(x^2+y^2\right)\mathrm{d}m\\ &= \int R^2\mathrm{d}m - \int\left(x^2+y^2\right)\mathrm{d}m\\ &= \frac{1}{2}(2A+C)-C\\ &= A-\frac{1}{2}C\end{aligned} \tag{12.1.255}$$

The third part in zero by symmetry:

$$\int xz\mathrm{d}m = 0 \tag{12.1.256}$$

Indeed, the 8 octants of the ellipsoid give the same integral values since they have the same form. However, there are 4 octants where xz is positive and 4 octants where xz is negative.

We have therefore:

$$\int R^2\cos^2(\theta)\mathrm{d}m = \cos^2(\phi)\frac{1}{2}C + \sin^2(\phi)\left(A-\frac{1}{2}C\right) \tag{12.1.257}$$

As:

$$\cos^2(\phi)+\sin^2(\phi) = 1 \tag{12.1.258}$$

we have:

$$\begin{aligned}\int R^2\cos^2(\theta)\mathrm{d}m&=\left(1-\sin^2(\phi)\right)\frac{1}{2}C+\sin^2(\phi)\left(A-\frac{1}{2}C\right)\\&=\frac{C}{2}-\frac{C}{2}\sin^2(\phi)+A\sin^2(\phi)-\frac{C}{2}\sin^2(\phi)\\&=\frac{C}{2}+(A-C)\sin^2(\phi)\end{aligned}\tag{12.1.259}$$

The fourth part is:

$$\frac{3}{2r^2}\int R^2\cos^2(\theta)\mathrm{d}m=\frac{3}{2r^2}\left(\frac{C}{2}+(A-C)\sin^2(\phi)\right)\tag{12.1.260}$$

Now that we have all the values of the integrals, we can find the gravitational energy. Knowing that:

$$\int \mathrm{d}m=M\tag{12.1.261}$$

$$\int R\cos(\theta)\mathrm{d}m=0\tag{12.1.262}$$

$$\frac{1}{2r^2}\int R^2dm=\frac{1}{4r^2}(2A+C)\tag{12.1.263}$$

$$\frac{3}{2r^2}\int R^2\cos^2(\theta)\mathrm{d}m=\frac{3}{2r^2}\left(\frac{C}{2}+(A-C)\sin^2(\phi)\right)\tag{12.1.264}$$

We get:

$$\begin{aligned}E_{\text{pot}}&=-\frac{G\mu}{r}\left(\int\mathrm{d}m+\frac{1}{r}\int R\cos(\theta)\mathrm{d}m-\frac{1}{2r^2}\int R^2\mathrm{d}m+\frac{3}{2r^2}\int R^2\cos^2(\theta)\mathrm{d}m+\ldots\right)\\&=-\frac{G\mu}{r}\left(M+0-\frac{1}{4r^2}(2A+C)+\frac{3}{2r^2}\left(\frac{C}{2}+(A-C)\sin^2(\phi)\right)+\ldots\right)\\&=-\frac{G\mu M}{r}+\frac{G\mu}{2r^3}\left(\left(A+\frac{1}{2}C\right)-3\left(\frac{C}{2}+(A-C)\sin^2(\phi)\right)\right)+\ldots\\&=-\frac{G\mu M}{r}+\frac{G\mu}{2r^3}\left(A+\frac{1}{2}C-\frac{3C}{2}-3(A-C)\sin^2(\phi)\right)+\ldots\\&=-\frac{G\mu M}{r}+\frac{G\mu}{2r^3}\left(A-C-3(A-C)\sin^2(\phi)\right)+\ldots\\&=-\frac{G\mu M}{r}+\frac{G\mu(A-C)}{2r^3}\left(1-3\sin^2(\phi)\right)+\ldots\\&=-\frac{G\mu M}{r}+\frac{G\mu(C-A)}{2r^3}\left(3\sin^2(\phi)-1\right)+\ldots\end{aligned}\tag{12.1.265}$$

Let us remind now as proven in the section of Geometric Shapes that for an ellipsoid (see page 2981):

$$(J_0)_{(\vec{i},\vec{j},\vec{k})}=\begin{pmatrix}\frac{1}{5}M(b^2+c^2)&0&0\\0&\frac{1}{5}M(a^2+c^2)&0\\0&0&\frac{1}{5}M(a^2+b^2)\end{pmatrix}\tag{12.1.266}$$

Hence for an oblate spheroid:

$$\begin{aligned} C &= \frac{2}{5}Ma^2 \\ A &= \frac{1}{2}M\left(a^2+b^2\right) \end{aligned} \tag{12.1.267}$$

where a is the equatorial radius and b is the polar radius of the oblate spheroid (planet). Therefore we have:

$$\begin{aligned} C-A &= \frac{2}{5}Ma^2-\frac{1}{5}M\left(a^2+b^2\right) \\ &= \frac{1}{5}M\left(a^2-b^2\right) \end{aligned} \tag{12.1.268}$$

As the flatness is defined by:

$$f=\frac{a-b}{a} \tag{12.1.269}$$

We have:

$$b=a(1-f) \tag{12.1.270}$$

Hence:

$$\begin{aligned} C-A &= \frac{1}{5}M\left(a^2-b^2\right)=M\frac{a^2-a^2(1-f)^2}{5}=M\frac{a^2-a^2(1-2f)}{5} \\ &= a^2M\frac{1-(1-2f)}{5}=a^2M\frac{2f}{5} \end{aligned} \tag{12.1.271}$$

Hence:

$$E_{\text{pot}}=-\frac{G\mu M}{r}+\frac{G\mu(C-A)}{2r^3}\left(3\sin^2(\phi)-1\right)+\ldots \tag{12.1.272}$$

Can be rewritten:

$$\begin{aligned} E_{\text{pot}} &= -\frac{G\mu M}{r}+\frac{Ga^2\mu Mf}{5r^3}\left(3\sin^2(\phi)-1\right)+\ldots \\ &= -\frac{G\mu M}{r}\left(1-\frac{a^2f}{5r^2}\left(3\sin^2(\phi)-1\right)+\ldots\right) \end{aligned} \tag{12.1.273}$$

We will now find the apparent potential energy, which takes into account the planet's rotation. Apparent weight is different from actual weight because the Earth rotates on its axis, meaning that objects on its surface accelerate.

The apparent weight of an object of mass μ that follows the Earth's rotation (for modern measurements of Earth's rotation and its variations see [102] and [484]) is:

$$\vec{W}_{\text{app.}}=\vec{W}-\mu\vec{a} \tag{12.1.274}$$

Since the acceleration is $a=-\omega^2x\vec{i}$, we have:

$$\vec{W}_{\text{app.}}=\vec{W}+\mu\omega^2x\vec{i} \tag{12.1.275}$$

This apparent weight can be obtained from the following potential energy:

$$\begin{aligned}E_{\text{app.}}&=E_{\text{pot}}-\frac{1}{2}\mu\omega^2x^2\\&=E_{\text{pot}}-\frac{1}{2}\mu\omega^2r^2\cos^2(\phi)\end{aligned}\tag{12.1.276}$$

The apparent potential energy is therefore:

$$E_{\text{app.}}=-\frac{G\mu M}{r}\left(1-\frac{a^2f}{5r^2}\left(3\sin^2(\phi)-1\right)+\ldots\right)-\frac{1}{2}\mu\omega^2r^2\cos^2(\phi)\tag{12.1.277}$$

The Earth's surface has a constant potential energy. The potential energy at the equator must therefore be equal to the potential energy at the poles.

At the equator $(\phi=0°,r=a)$, the energy is equal to:

$$\begin{aligned}E_{0°}&\cong-\frac{G\mu M}{a}\left(1-\frac{a^2f}{5a^2}\left(3\sin^2(0°)-1\right)\right)-\frac{1}{2}\mu\omega^2a^2\cos^2(0°)\\&=-\frac{G\mu M}{a}\left(1+\frac{f}{5}\right)-\frac{1}{2}\mu\omega^2a^2\end{aligned}\tag{12.1.278}$$

At the poles $(\phi=90°,r=b)$, the energy is equal to:

$$\begin{aligned}E_{90°}&\cong-\frac{G\mu M}{b}\left(1-\frac{a^2f}{5b^2}\left(3\sin^2(90°)-1\right)\right)-\frac{1}{2}\mu\omega^2a^2\cos^2(90°)\\&=-\frac{G\mu M}{b}\left(1-\frac{2a^2f}{5b^2}\right)\end{aligned}\tag{12.1.279}$$

Since the two energies must be equal, we can equalize the both $(E_{0°}=E_{90°})$, such that:

$$-\frac{G\mu M}{a}\left(1+\frac{f}{5}\right)-\frac{1}{2}\mu\omega^2a^2=-\frac{G\mu M}{b}\left(1-\frac{2a^2f}{5b^2}\right)\tag{12.1.280}$$

We are lucky because μ can be simplified:

$$-\frac{GM}{a}\left(1+\frac{f}{5}\right)-\frac{1}{2}\omega^2a^2=-\frac{GM}{b}\left(1-\frac{2a^2f}{5b^2}\right)\tag{12.1.281}$$

We have to eliminate a, or b. Let us choose to eliminate b using:

$$b=a(1-f)\tag{12.1.282}$$

Therefore:

$$-\frac{GM}{a}\left(1+\frac{f}{5}\right)-\frac{1}{2}\omega^2a^2=-\frac{GM}{a(1-f)}\left(1-\frac{2a^2f}{5a^2(1-f)^2}\right)\tag{12.1.283}$$

Hence (we distribute the right parenthesis):

$$-\frac{GM}{a}\left(1+\frac{f}{5}\right)-\frac{1}{2}\omega^2a^2=-\frac{GM}{a(1-f)}+\frac{2GMf}{5a(1-f)^3}\tag{12.1.284}$$

Let us assume f is small, then using Maclaurin development at first order:

$$\frac{1}{1-f}\cong 1+f \qquad \frac{1}{(1-f)^3}\cong 1+3f \tag{12.1.285}$$

Hence:

$$-\frac{GM}{a}\left(1+\frac{f}{5}\right)-\frac{1}{2}\omega^2a^2=-\frac{GM(1+f)}{a}+\frac{2GMf(1+3f)}{5a} \tag{12.1.286}$$

Multiplying both sides by $-a$ and dividing by GM:

$$1+\frac{f}{5}+\frac{\omega^2a^3}{2GM}=1+f-\frac{2f(1+3f)}{5} \tag{12.1.287}$$

This is a second order polynomial in f. If we solve it, we get for the two roots:

$$f_{1,2}=\frac{GM+\sqrt{G^2M^2-15GM\omega^2a^3}}{6GM} \tag{12.1.288}$$

If we inject into it the values for the Earth:

$$G=6.6743\cdot 10^{-11}\,[\mathrm{N}\cdot\mathrm{m}^2\cdot\mathrm{kg}^{-2}] \quad M=5.9722\cdot 10^{24}\,[\mathrm{kg}] \tag{12.1.289}$$

$$a=6,378\cdot 10^3\,[\mathrm{m}] \quad \omega=7.2921159\cdot 10^{-5}\,[\mathrm{rad}\cdot\mathrm{s}^{-1}] \tag{12.1.290}$$

We get:

$$f_1\cong 0.3289 \qquad f_2\cong 0.004384 \tag{12.1.291}$$

Modern measurement give $f=0.00334$, so the second root is morel likely the right one, hence an error with the second root of the model of 31%. Considering the number of approximation we did (three Maclaurin series approximations plus a constant density approximation...).

Remark

If we look at:

$$1+\frac{f}{5}+\frac{\omega^2a^3}{2GM}=1+f-\frac{2f(1+3f)}{5} \tag{12.1.292}$$

and we neglect the f^2 in the last parentheses (as f is already small, f^2 is very small and can be removed), we get:

$$1+\frac{f}{5}+\frac{\omega^2a^3}{2GM}=1+f-\frac{2f}{5} \tag{12.1.293}$$

And rearranging we have:

$$\boxed{f\cong\frac{5}{4}\frac{\omega^2a^3}{GM}\cong 0.00432644} \tag{12.1.294}$$

Although an ellipsoid of revolution (oblate spheroid) is the best description for the shape of a planet:

Figure 12.45: **Earth with its atmosphere and oceans**

there are obviously imperfections between the model and the reality for some planets (in particular the terrestrial planets, satellites, and small rocky bodies). The geopotential of real celestial bodies can be shaped in a much more complicated way because of the influences of the visible inhomogeneities on the surface as evidenced by these satellite images of the Earth omitting the liquid parts of it (the deformations are amplified by a factor $100,000$ in the image below!):

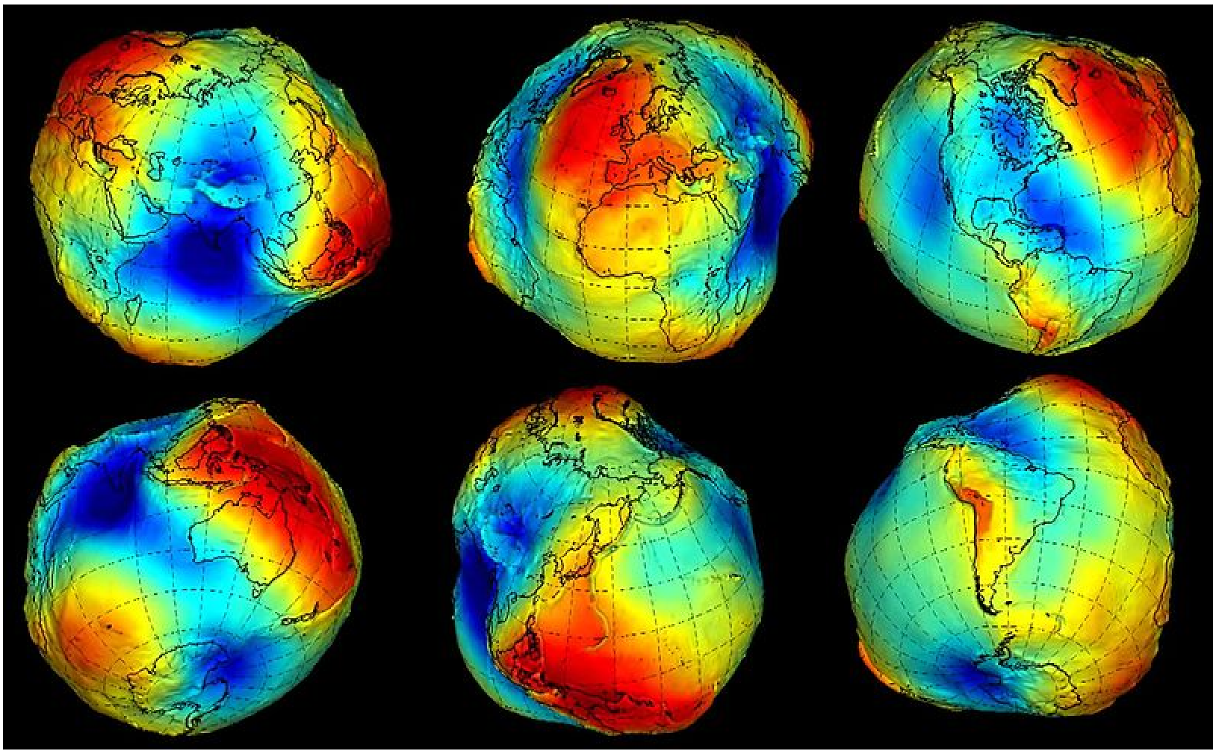
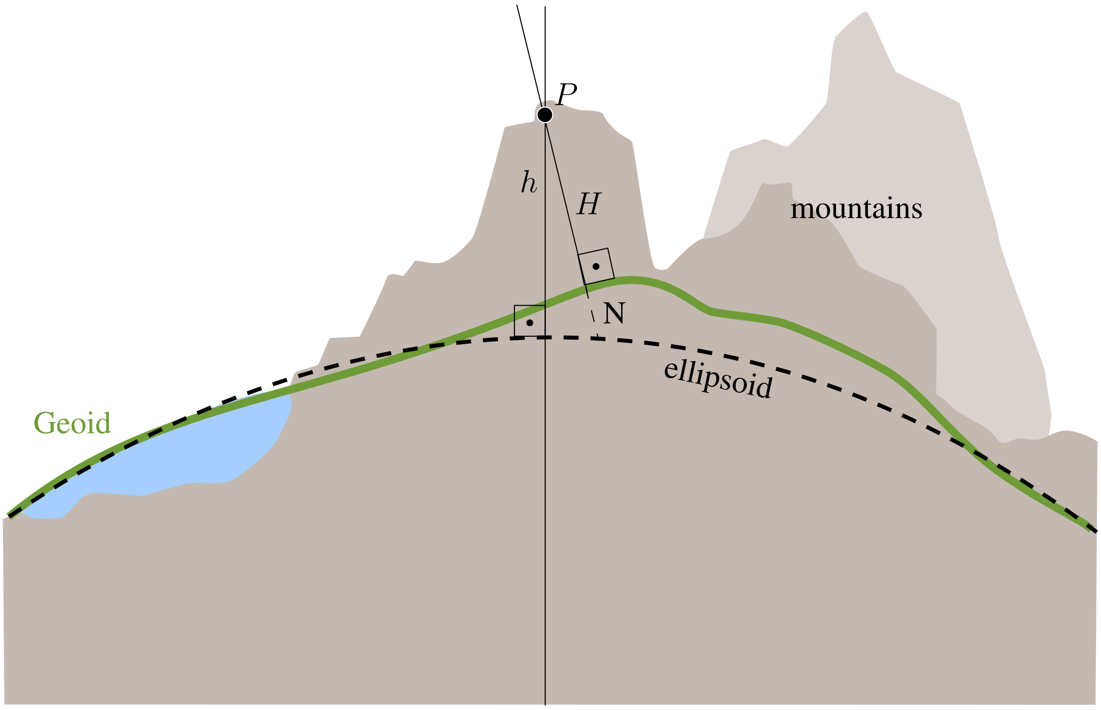


Figure 12.46: **Earth shape without its atmosphere and oceans (source: Wikipedia)**

So the Earth is obviously not a sphere nor Ostrich egg shaped...

The "geoid" is a particular equipotential surface of the Earth's gravity field and serves as a reference for the determination of altitudes. We can imagine the geoid as being the mean sea level extended under the continents.

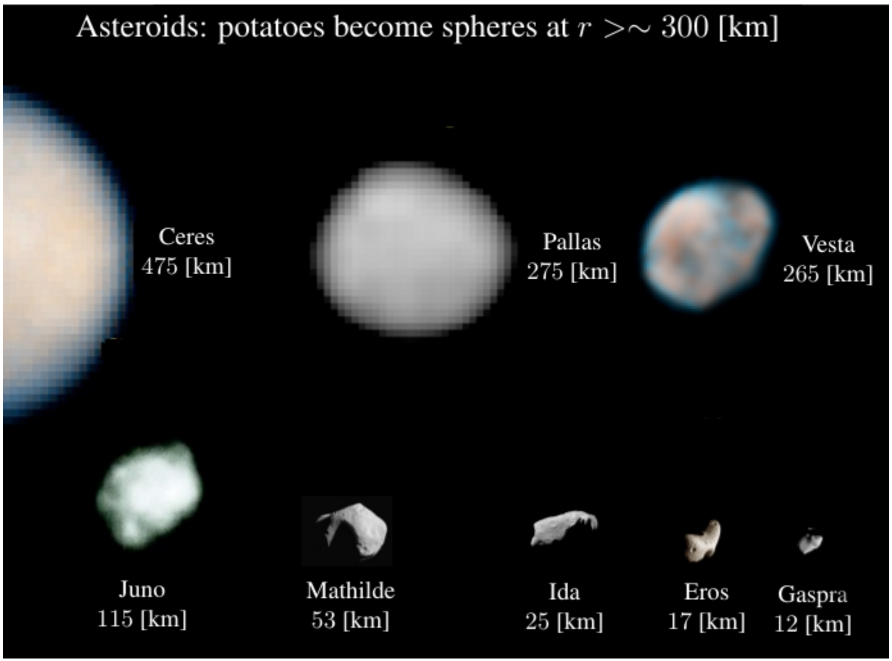
Considered globally, the geoid deviates from a mathematical reference surface (a rotational ellipsoid) by ±100 meters at most.



P
h
H
N
mountains
ellipsoid
Geoid

Figure 12.47: Relationship between geoid, ellipsoid and orthometric elevation (source: SwissTopo)

The specialists of geodesics and of topography have to take into account these inhomogeneities!



Asteroids: potatoes become spheres at $r >\sim 300$ [km]
Ceres
475 [km]
Pallas
275 [km]
Vesta
265 [km]
Juno
115 [km]
Mathilde
53 [km]
Ida
25 [km]
Eros
17 [km]
Gaspra
12 [km]

Figure 12.48: Evolution of asteroids shape in function of the radius

Remark

The "Struve Geodetic Arc" is a $2,822$ [km] long chain chain comprised of 12 sections, interspersed between astronomy points, having 10 measured bases, 258 triangles and connecting 259 triangulation points spanning from the north of Norway to the Black Sea (across ten different countries). It was established between 11816 and 11864 (holocene calendar) with the purpose of measuring the flattening of Earth's ellipsoid, and from there the size of the whole globe (since 12005, the points in the triangulation network, including 34 marked with obelisks, have been a UNESCO World Heritage Site). Sadly we were not able to find what was the flatness value obtained by this project...

Because Earth isn't a perfect sphere but rather an ellipsoid, an international consensus for the shape of the best fitted ellipse has been defined under the name `WGS-84` and is used for GPS applications (among others).

Why do we mention this? Because the highest point on Earth's surface measured from sea level is actually the summit of Mount Everest, on the border of Nepal and China with a height of $8,848.86$ [m]. However, the highest point on Earth's surface measured from Earth's center is the summit of Chimborazo in Ecuador, at $6,384.4$ [km] from Earth's centre; the peak's elevation relative to sea level is $6,263.47$ [m]. Because Earth is an oblate spheroid rather than a perfect sphere, it is wider at the equator and narrower toward each pole. Therefore, the summit of Chimborazo, which is near the Equator, is farther away from Earth's centre than the summit of Mount Everest is; the latter is $2,168$ [m] closer, at $6,382.3$ [km] from Earth's centre.

But how do we compute the difference in height between the sphere and the WGS-84 reference ellipsoid as illustrated below:

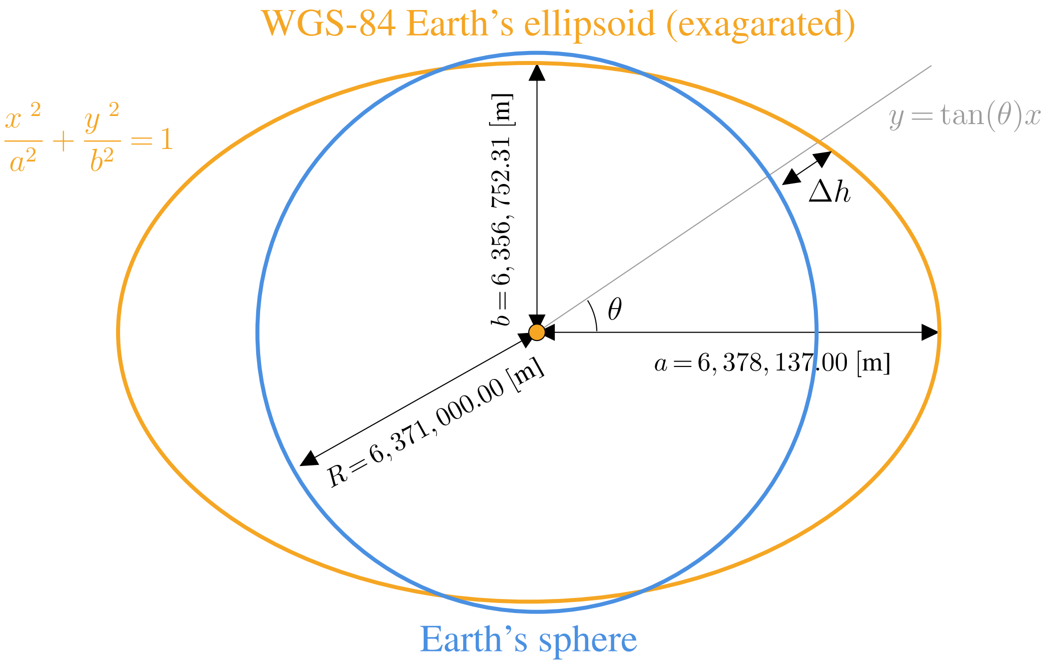


Figure 12.49: Earth's ellipsoid of reference vs the perfect sphere

We start for the ellipse famous equation:

$$\frac{x^2}{a^2}+\frac{y^2}{a^2}=1 \tag{12.1.295}$$

If we want to determine Δh for the angle θ corresponding to a given latitude, we need to find the intersection between the equation of the straight line $y=\tan(\theta)$ and the rearranged equation of the ellipse:

$$y=b\sqrt{1-\frac{x^2}{a^2}} \tag{12.1.296}$$

Thus we seek x for which:

$$\tan(\theta)x=b\sqrt{1-\frac{x^2}{a^2}} \tag{12.1.297}$$

Hence:

$$x^2\left(\tan^2(\theta)+\frac{b^2}{a^2}\right)=b^2 \tag{12.1.298}$$

Hence:

$$x=\frac{b}{\sqrt{\tan^2(\theta)+\dfrac{b^2}{a^2}}} \tag{12.1.299}$$

So finally:

$$\begin{aligned}\Delta h=\sqrt{x^2+y^2}-R^2&=\sqrt{\left(\frac{\tan(\theta)b}{\sqrt{\tan^2(\theta)+\dfrac{b^2}{a^2}}}\right)^2+\left(\frac{b}{\sqrt{\tan^2(\theta)+\dfrac{b^2}{a^2}}}\right)^2}-R^2\\&=\frac{b\sqrt{\tan^2(\theta)+1}}{\sqrt{\tan^2(\theta)+\dfrac{b^2}{a^2}}}-R^2\end{aligned} \tag{12.1.300}$$

Using this result we can found the lowest dry land on Earth relatively to Earth's center in comparison of the lowest one relatively to the sea level (shore of the Dead Sea, shared by Israel, Palestine and Jordan at -432.65 [m]). We can easily found that the dry place with the highest latitude on Earth's at sea level is the northernmost Greenland's costs that is at 0 [m] from sea level but approximately $15,193$ [m] lower than the Dead Sea relatively to Earth's center! Hence the lowest point on Earth's surface (on dry land) relatively to Earth's center is the northernmost Greenland's shore.

Ignoring Earth's shape, until 12019 (holocene calendar), the lowest point on continental Earth relatively to sea level was still believed to be the Dead Sea area with ~ 440 [m] under sea level ([417]). In 12019 (holocene calendar) the lowest point on continental Earth has been identified in East Antarctica under the Denman Glacier where it reaches ~ 3.5 [km] under sea level (see [399]).

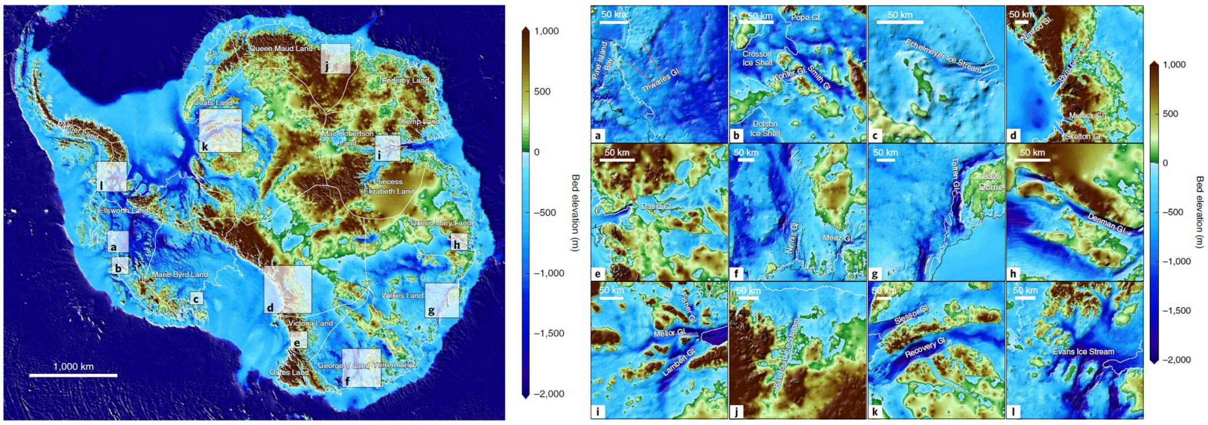


Figure 12.50: **Antarctica topography (source: [399])**

12.1.4.8.2 Empirical evidence that Earth isn't globally flat

Because some people around the world on social networks lie about the shape of the Earth to defend their ignorant religion/sect (indeed it is confirmed in linguistic that the Earth is described as flat in the bible as proven in [195], and also in sunni Islam as proven in quran's 79:30, tafsir Al-Tabari 1/435-436, tafsir Abd Al-Razzaq 3/329 and Sahih al-Bukhari 6520) or to make money by selling books, conferences or by getting more views of their videos, we have sadly to remind a non-exhaustive list of simple empirical evidence that the Earth is a geoid (approximately a sphere/oblate-spheroid) that spins[12]:

1. On a completely flat Earth it would be possible to see distant locations distant from more than 500 kilometers with a telescope even if taking into account light diffusion and absorption. This has never been done because the Earth is curved (see [540]). If the Earth was flat the following locations should also be visible every day from each other however it's not the case: Guilvinec (France)- Foz (Spain)/~ 524 kilometers, Marseille (France) - Perpignan (France)/~ 213 kilometers, Calvi (France)-Nice (France) /~ 175 kilometers, Chicago (USA)-Manistique (USA)/~ 464 kilometers, Sebastopol (Russia)-Burgas (Bulgaria)/~ 538 kilometers, Bangkok (Thailand)-Koh Samui (Thailand)/~ 469 kilometers, Sharm El Sheikh (Egypt)-Assab (Eritrea)/$\sim 1,870$ kilometers... among millions of other places around the World where the flat Earth model doesn't work!

2. All Earth size like bodies in the Solar System (and even much smaller and bigger) seen by telescopes like Hubble, James Webb Telescope and others are seen as geoid (like the Moon, Mars, Mercury, Venus, Uranus, Ganymede, Callisto, Io, Europa, etc.). This is a statistical abductive evidence that the Earth is obviously not globally flat.

3. From a large sea (above 50 kilometres wide for example) and a position at the water level (coast) you can't see the coast from the other side (this implies that boats or offshore windmill farm - like the Thorntonbank Wind Farm in Belgium - at the horizon disappear from the bottom to the top and not shrink uniformly as it would be the case according to the law of perspectives on a flat Earth).

---

[12] We have grouped together evidence of the same type so that the list below is better structured. If we had not done this grouping, the list would have easily exceeded a few thousands or even millions of evidence against the flat Earth.

4. Eratosthenes (ancient Greek polymath: a mathematician, geographer, poet, astronomer, and music theorist) measured the angle of shadow to determine the circumference of the Earth using only two measurement points in Alexandria and Syene as he already knew the Earth is a sphere (in the 98th century according to holocene calendar). Flat earthers are not wrong when they claim the original experiment did not prove the Earth is round. But we can modify the experiment to use three or more measurement points (more than 600 schools on the five continents have carried out more than $6,000$ measurements since 2005 - The Eratosthenes project), and the flat Earth model is instantly proven wrong.

5. The fact that the Sun and Moon appear and disappear from the horizon with a geometry that violates the laws of linear perspective (especially the homographic equations) contradicts a flat Earth model where the Sun and the Moon always remain at a fixed altitude above the flat Earth.

6. Flat earthers can't explain by a mathematical model why we can't see the far side of the Moon (in other words: that we can see only the visible side including the small libration effect) as in their model the Sun and Moon are always above the flat Earth and turn (according to them) around an imaginary barycenter that is located at the North Pole... or that spirals around each other (.........).

7. The National Society of Professional Surveyors (NSPS) require practitioners to take into account an average declination of 8 [cm] per kilometres (see NSPS "Definitions of Surveying and Associated Terms"). The Swiss Geodesy Commission has a similar rule in the Volume 70 of its *Astronomisch-geodätische Arbeiten in der Schweiz*.

8. We have direct non-composite real-color images footage of a geoid Earth from weather satellite like - among others - the Japanese Himawari-8 (the satellite orbits the Earth synchronously in a geostationary orbit at $140°$ East at a distance of $35,786$ [km]).

9. Seismic waves propagation (primary waves, secondary-waves, Rayleigh waves, Love waves, Stoneley waves) after Earthquakes are reflected and propagates in a way that wouldn't work on flat Earth.

10. The LIGO (Laser Interferometer Gravitational-Wave Observatory) research centers had to take into account the curvature of Earth in the design of their detector. This includes the four following detectors: LIGO Hanford/USA, LIGO Livingston/USA, GEO600/Germany, Virgo/Italy.

11. Coriolis effect makes cyclones and oceanic currents spin counter-clockwise north of the equator, but clockwise south of the equator (this is also related to Foucault pendulum or Gyro LASER 15 degree per hour drift (Sagnac effect)), due to the Earth's rotation (for modern measurements of Earth's rotation and its variations see [102] and [484]). Foucault's pendulum proves Earth's rotation but also that it is spherical as the pendulum rotates clockwise in the northern hemisphere and counterclockwise in the southern hemisphere. The pendulum turns faster if it is closer to the pole. On the equator, the pendulum does not rotate.

12. The shadow of Earth on the Moon during a lunar eclipse is always a dark oval that moves from one side of the Moon to the other. The only shape that casts an oval shadow no matter which direction it is pointed is something that is shaped like a sphere. Furthermore, in

the flat Earth model, the position of the Moon and the Sun relatively to the Earth makes it impossible for flat Earthers to explain the origin of the lunar eclipse without involving a third invisible body... They also don't have any mathematical model that can predict the occurence of lunar eclipses as we have for the traditional model.

13. The flat Earth model is also inconsistent with the fact that a given lunar eclipse is only visible from approximately only half of Earth at a time. Furthermore, Flat earthers don't have any mathematical model able to predict the lunar eclipses as we have for the traditional model. Flat earthers can't also provide a mathematical model to predict the path of totality (the area where the eclipse is total) as we have for the traditional model.

14. Tides would exist on a flat Earth as water bulges up a bit as the Moon or Sun pass overhead. However.... there are a number of observational facts that this model doesn't explain like why do high tides happen twice a day generally, as claimed by flat earthers gravity doesn't exist so the observed tides shouldn't exist, why are there two spring and neap tides (larger and smaller tidal ranges) each month approximately and why do the spring tides correlate with full and new Moon? So it would be more accurate to say that a flat Earth shouldn't have tides that act as tides do on Earth. Hence flat Earth is wrong.

15. The stars in the northern hemisphere turn counterclockwise around (approximately) the Polaris (Alpha Ursae Minoris) and in the southern hemisphere clockwise around Polaris Australis (Sigma Octantis) and this is incompatible with the flat Earth model.

16. The meteorites, stars, constellations, nebulae and galaxies visible in the southern hemisphere are not the same as those visible in the northern hemisphere and this is because the Earth is a sphere. The flat Earth can't explain this observational fact.

17. In the Northern Hemisphere, the Sun appears to move to the right. In the Southern Hemisphere, it appears to move to the left. At sunrise and sunset (as commonly named... a better denomination being daybreak/nightfall), the Sun's path is at an angle that roughly corresponds to the observer's latitude. This occurs because observers at different locations on the Earth's surface are not on the same plane. The path of the Sun observed from many different locations on the Earth's surface is proof that the Earth is spherical. This observation cannot occur on a flat Earth.

18. Approximately four months per year there is no night (midnight sun) in respectively the north/south hemisphere and there are plenty of uncut videos providing evidence for people who don't have the money to take holidays in Antarctica. At the exact position of the North pole and South pole the midnight Sun has a duration of six months. This observational fact is incompatible with the flat Earth model.

19. The Sun describes an analemma in the sky during a year can only be explained on a globe (oblate-spheroid) Earth, rotating with a titles axis rotation and in motion around the Sun with an eccentric orbit.

20. The seasons variations of temperature are explained and computed accurately by mathematical models considering a globe (oblate-spheroid) Earth. Flat earthers are not able to provide a coherent mathematical model to explain seasons and also for the Earth equilibrium temperature nor for the diurnal temperature variation...

21. The supposed 0° angle from the North Pole and the 90° angle from the equator on a flat Earth model predicts a Sun which would have to be located essentially next to the surface of the ground of the flat Earth. That's obviously a non-sense!

22. Because the Sun should be next to the surface if the Earth was flat the apparent diameter (angular size) variation of the Sun should be of a few millimetres in a day however flat earthers have never been able to measure such a magnitude using scientific instruments. According to the traditional model the distant Sun angular size varies between 0.524° to 0.5418° (hence a variation of 3.4%) however on the flat Earth model the angular size variation is more than 100% (...) what is obviously incompatible with observational facts.

23. The magnetic field of a flat disc according to the model of flat Earth has never been observed and measured by scientific instruments.

24. The gravity values that would have been measured on different points of a flat Earth are not compatible with the best physical models of gravity (like General Relativity or it's Newtonian approximation) nor with the actual observations.

25. The mathematics of navigation using Global Positioning System (GPS) satellites assume that they are moving in known orbits around an approximately spherical surface. The fact that the GPS (Global Positioning System) uses the WGS84 reference ellipsoid (World Geodetic System 1984) for the conversion between NED (North-East-Down) and LLA (Latitude, Longitude and Altitude) coordinates is a factual evidence that Earth isn't flat.

26. Satellites, including communications satellites used for television, telephone, and Internet connections, would not stay in orbit unless the modern theory of gravitation were correct and the latter describe for Earth an oblate-sphere shape and not a flat Earth.

27. Large bridges like the towers of the Humber Bridge, although both vertical with respect to gravity, are 36 millimetres farther apart at the top than the bottom due to Earth's curvature (other bridges like the Causeway Bridge at Lake Pontchartrain, Louisiana in USA, the Akashi Kaikyo Bridge, the Verrazano Bridge and many others provide counter evidence for a flat Earth).

28. During the construction of the LHC (Large Hadron Collider at CERN, Geneva/Switzerland), the SPS (Super Proton Synchrotron), the Tevatron (Fermilab, Illinois/USA), the KEK (japan), the Relativistic Heavy Ion Collider (RHIC) and the SLAC (Stanford Linear Accelerator Center, California/USA) accelerators the curvature of the Earth had to be taken into account to satisfy the precision constraints of the particles beams.

29. The existence of Time Zones is another evidence that the Earth is approximately an oblate-spheroid. Indeed, for example, in the flat Earth model, the Sun is often depicted as moving in a circular path above the Earth. This would mean that all locations would experience sunrise and sunset at the same time, eliminating the concept of time zones.

30. Having the total height of about 829.8 meters, Burj Khalifa is the tallest building in the world (located in Dubai, United Arab Emirates) at the day we write these line. It is so tall that we can observe the sunset at the base of the building, then rush to the upper floors and watch the same sunset again for the second time on the same day and this is possible because the Earth isn't flat (Burj Khalifa has two observation desks where you can view

sunset twice in the same evening an special tickets named "Sunset tours" are sold on this purpose). Furthermore, the Burj Khalifa has different official Islamic prayer time (three different prayer and iftar time zones) depending on the floor level because the Earth isn't flat! On the higher floors, the Sun is seen for several minutes after it has set at ground level. Those living above the 80th floor have to wait two extra minutes to break their Ramadan fast, and those living above the 150th floor have to wait three minutes for the same reason.

31. The fact that some mountains/places that should be below the horizon because of their distance are visible only a few days per year (because of special atmospheric condition and Sun position) is compatible with the best mathematical geometric optic models but with none of the flat Earth.

32. The time-distance travelled by airplanes (like the Qantas direct flight route QFA63 or QFA64 between Johannesburg in South Africa and Sydney in Austrial), cruise boats (like the MSC Grand Voyage cruise on the MSC Poesia between Barcelona in Spain and Santos in Brazil) or animals (like the white sharks Nicole and Lydia tagged with a GPS respectively in the years 12004 and 12013 according to holocene calendar) are incompatible with a flat Earth.

33. Companies that put submarine copper cables or optical fiber cables have to minimize the distance to optimize the costs. The submarine cables length are incompatibles with a flat Earth. For example the cable between Sydney (Australia) and Perth (Australia) is $4,850$ kilometres long according to his owners (superloop, singTel, Google, Indosat, Oooredoo, AARNET) and follow the shore line. However the shortest path line on the flat Earth map is $8,300$ kilometres between Sydney and Perth.

34. Some commercial airplanes company fly over Antarctica and this wouldn't be possible on a flat Earth. For example on July 14, 12023 (holocene calendar), Qantas Flight QF28 between Santiago de Chile and Sydney rerouted from the usual path and instead went over Antarctica. The flight typically passes over the Southern Pacific Ocean only as far south as 70° south. Nonetheless, that day, the crew flew to 74.17° south to avoid strong headwinds along the usual route. This wouldn't have been possible on a flat Earth.

35. A flat disc like planet contradicts the isoperimetric inequality theorem (for reminder it is based on the Lagrangian). Indeed, the laws of our observable universe (principle of least action) constraints the shape of viscous objects to minimize their surface area for a given volume, leading to a spherical shape in three dimension. This is a direct application of the isoperimetric inequality in three dimensions, where the sphere has the smallest surface area for a given volume.

36. Antarctica was crossed by multiple scientific teams starting from opposite coasts and who encountered at the center of the continent (Commonwealth Trans-Antarctic Expedition) and it was also crossed by extreme sportmen (Norwegian Liv Arnesen and the American Ann Bancroft). This wouldn't have been possible on a flat Earth.

37. The Moon shows its same side (tidal locking) - including libration effect - whatever the position of the observer on Earth and the same time for all observers on Earth and that's not possible on a flat Earth.

38. The face of the Moon is inverted between the northern hemisphere and the southern hemisphere (hence the observation that Moon phase angle varies depending on the latitude at a same moment) and this also can't be explained on a flat Earth.

39. The tilt of the Moon which changes during the day in a clockwise or counterclockwise manner depending on the hemisphere is incompatible with the flat Earth model.

40. The sundial that works all year correctly when the inclination of the needle corresponds to the latitude is once again an evidence against the flat Earth. This corresponds also to the fact that the angle to Polaris approximately corresponds to the latitude of the observer. This fact is observed on every location on Earth where Polaris is visible. By tracing the path to Polaris from multiple locations on the flat model, the lines will not point to a consistent position of Polaris. The reason is that Earth is a sphere and the flat Earth is a sphere and the flat Earth model does not represent reality.

41. The four LIGO experiments localization in USA and Europe (separated from a few thousands of kilometres) provide evidence thanks to time-distance measurements that gravitational waves propagation don't go through a flat Earth but trough an oblate-spheroid Earth.

42. Extending vanishing point of parallels roads from various cities in the World from aerial photos converge to a point in the sky rather to a point at the horizon. This observational fact is against a flat Earth model. A famous well known case of easily observable vanishing point at the ground level and that is incompatible with a flat Earth is the lake Pontchartrain Transmission line (USA) where 85 aligned power transmission towers cross a lake 24.27 kilometers across. If the Earth is flat, the vanishing point of the transmission towers should intersect the horizon but in reality anyone can see that the vanishing point is above the horizon.

43. The Sun always "rises" (...) in the East and always "sets" (...) in the West, in both the southern and northern hemispheres. But the difference is that it is shifted South at noon when you are in the northern hemisphere, and shifted North at noon when you are in the southern hemisphere and this isn't possible on flat Earth.

44. Because the Earth is an oblate-spheroid, the horizon lies below the eye level. The angle between the eye level and the horizon is the "dip of the horizon". The dip is larger as we go higher and this factual observation contradicts a flat Earth model. In the 111th century (according to holocene calendar), Abu Rayhan Muhammad ibn Ahmad al-Biruni (Iranian muslim scholar and polymath during the islamic golden age) determined the radius of the Earth by measuring the dip of the horizon from the top of a mountain at Nandana in Pind Dadan Khan (present-day Pakistan) with a 20% accuracy (the method and assumptions are different of that of Eratosthenes).

45. The flat earthers are unable to provide a mathematical model to explain the apparent change of color of the Sun at sunset (red sunset) including the famous green flash. However in the traditional model of the Earth (oblate sphere), this both phenomenon are clearly computable by Rayleigh scattering and because the Sun is below the horizon (i.e. behind the curvature!).

46. The Deep Underground Neutrino Experiment (DUNE) time lap between the Fermi National Accelerator Laboratory in Batavia, Illinois (USA) and the Sanford Underground

Research Laboratory in Lead, South Dakota (USA) - $1,300$ kilometers downstream of the source - can be explained only on an oblate-spheroid Earth. The T2K experiment neutrino oscillation experiment, sending a muon neutrino beam from J-PARC proton accelerator on the east coast of Japan to Super-Kamiokande located 295 kilometers away, close to the west coast of Japan lead to the same conclusion and same for the CNGS project between the CERN Geneva (Switzerland) SPS (Super Proton Synchrotron) accelerator to Grand Sasso (Italy) apart from 732 kilometers. The difference between the neutrino travel times and the speed of light over the 732 [km] beam path was measured to be -1.9 ± 3.7 nanoseconds, consistent statistically with zero difference (see [5]). In this case, if a flat Earth model had been used, the beam path distance would be equal to the surface distance from CERN to Gran Sasso, again about 410 metres longer. This would have given the neutrino travel time difference to be an extra $410/c=1370$ [ns], making the neutrinos travel significantly faster than the speed of light.

47. As for the human eye that sees using electromagnetic wave, standard radars can't see objects below the horizon in standard atmospheric conditions. To see below the horizon engineers have to use OTHR (Over The Horizon Radars) that bounce the signal on the ionosphere and the time propagation signal is compatible with an oblate-spheroid and not with a flat Earth.

48. The rocket engineers of the spatial agencies Roscosmos (Russia), JAXA (Japan Aerospace Exploration Agency), ESA (European Space Agency), CNSA (China National Space Administration), ISRO (Indian Space Research Organization), ISA (Israel Space Agency), ISA (Iranian Space Agency), United Kingdom Space Agency (UKSA), NATA (North Korean Space Agency), KASA (Korea AeroSpace Administration), SSAU (State Space Agency of Ukraine) and NASA (National Aeronautics and Space Administration) all confirm observations of a geoid Earth based on their independent rockets launch technologies and independent satellites.

49. There is experimental evidence based on atomic clocks that the spin of the Earth slows down. This result is incompatible with the claim of flat earthers that the Earth doesn't spin.

50. Aircrafts artificial horizons have self correcting mechanisms to constantly and rapidly correct alignment drift with the horizon and this necessity is necessary because the Earth is curved.

51. Armies have to integrate into the ballistic equations the rotation and curvature of Earth to reach their targets (among other parameters like weather conditions and gravity magnitude depending on the altitude). This wouldn't be necessary on a flat and non-rotating Earth.

52. At places of comparable latitude North and South, the Sun behaves very differently than it would on a non-spinning flat Earth or on a spinning flat Earth. For example, the longest summer days North of the equation are identical than those South of the equator (same for shortest day) and this isn't possible on a flat Earth but is compatible with a geoid Earth.

53. Plenty of constellations cannot be seen in the northern hemisphere and respectively in the southern hemisphere and this is incompatible with a flat Earth.

54. The diurnal tidal bulge (due to the Moon) is compatible only with a spherical like Earth shape (oblate-spheroid) and totally incompatible with a flat Earth.

55. All sailors and pilots in the world know that the Earth is curved because the international definition of the "nautical mile" takes into account the curvature of the Earth and only works on a spherical Earth and not on a flat one.

56. The Antarctica Cup Yacht Race is a race that turns clockwise around the Antarctica continent by maintaining a regular South-East cap. That wouldn't be possible on a flat Earth.

57. The Blasius boundary layer fluid mechanics model explains how and why the atmosphere rotates with a spinning Earth. Flat earthers claim at the opposite that there are no models able to explain the atmosphere rotation of a spinning oblate-spheroid Earth.

58. North/South circumnavigation were done multiple times by people with airplanes (Sir Ranulph Fiennes in 11979 - holocene calendar, Rockwell Polar Flight in 11965 - holocene calendar, TAG TransPolar 08 in 12009 - holocene calendar and One More Orbit in 12019 - holocene calendar). That's evidence Earth isn't flat.

59. During a solar eclipse the Sun and Moon are on the same axes and cast an irregular elliptical shadow on Earth. However on a flat Earth the shadow would be a circle. Therefore Earth isn't flat.

60. We have Aurora Borealis in Arctica because magnetic field lines are concentrated there. However we have also have Aurora Australis in Antarctica because magnetic field lines are concentrated there. This works only with the geoid Earth model as for flat earthers the magnetic field diverges in Antarctica.

61. The measured time-distance propagation of radio-wave signals (remember that electromagnetic waves travel at constant speed in a given medium, i.e. the speed of light in this medium!) is incompatible with those given in the flat Earth model however almost perfectly compliant with those given on a geoid Earth.

62. The results of the Hafele-Keating experimental test of Special Relativity is incompatible with the non-spinning Earth model (because of the measure speed relativistic effect component) but also incompatible with the gravitational field shape of a flat Earth (because of the measure of the gravitational relativistic effect component).

63. On 14 January, 12022 (holocene calendar), the volcano Hunga Tonga erupted explosively. Soon after, meteorological stations across the world were recording changes in barometric pressure due to the shockwave. The time-distance (at the speed of sound) of these shockwaves and the fact that at the antipodal shock-wave position (Algeria) the shock-wave has converged and produced a second opposite shockwave has been recorded. These both evidence all contradict the model of flat Earth.

64. Radio waves generated by lightning flashes in storm are (among others) electromagnetic waves with a frequency of 7.8 [Hz] that are travelling around the world in the waveguide formed by the Earth and the ionosphere, and returning one wavelength later to constructively interfere and reinforce themselves, producing a measurable peak in Earth's magnetic field fluctuations. The Irish physicist George Francis Fitzgerald first anticipated the existence of this effect named "Schumann resonances" in 11893 (holocene calendar),

but his work was not widely circulated. Around 11950, (holocene calendar) the German physicist Winfried Otto Schumann performed the theoretical calculations that predicted the resonances may be observable, and made efforts to observe them. But it was not until 11960 (holocene calendar) that Balser and Wagner made the first successful observations and measurements of Schumann resonances (see [39]). If the Earth was flat, then any ionosphere above it would be flat as well, and would still form a waveguide for low frequency radio waves. However it would not be a closed waveguide. Radio waves would propagate out the edges and be lost to space, meaning there would be no observable magnetic field resonances at all. And even if there were an opaque radio wall of some sort at the edge of the flat Earth, the size and geometry of the resulting cavity would be different, resulting in a different set of resonance frequencies.

65. There are more than twenty different models of the flat Earth (with or without dome, infinite or finite disc radius, different thickness of the disc, different height for the Sun and the Moon, that spin or doesn't spin, etc.) that all disagree with each other because there is zero experimental evidence for any flat Earth model. Hence all the flat Earth models are just a delusional belief of trolls, or bigots, or mentally ill, or business-seekers, or uneducated individuals.

66. In most flat Earth models the light of the Sun and the light of other stars are strongly bended/deflected in some situations for some observers every day! The magnitude of this daily supposed deflection magnitude of the light (sometimes near 90°) hasn't been modelled by any consistent and conclusive theoretical atmospheric model nor validated experimentally by flat earthers and is not only incompatible with classical actual atmospheric model but also incompatible with the gravitational Newtonian deflection of light or the gravitational General Relativity deflection of light.

67. The Earth-Moon barycentre isn't aligned with the Earth center. The result of this is that during a full Moon, when the Moon is farthest from the Sun, the Earth is $4,670$ [km] closer to the Sun than average, whereas during a new Moon the Earth is $4,670$ [km] further away from the Sun than average. The Earth oscillates over $9,000$ [km] towards and away from the Sun every month! And the increase in incident radiation from the Sun during the phases around the full Moon comes to about 43 [mW] per square metre, or an extra $5,450$ [GW] over the entire Earth. The Earth normally receives nearly 44 million giga-Watts of solar radiation, so the difference is relatively small, but it's enough to heat the Earth by almost 0.01 degrees Celsius, which is near the observed average monthly temperature variation (see [21]). This is incompatible with the flat Earth model where the Sun is at a fixed height only a few thousand kilometre above the ground. And if the Sun in the flat Earth model would change its height then it would be an easily observable effect from the surface of the flat Earth (and this has never been observed obviously...).

68. The Eötvös effect discovered in 11900 (holocene calendar) is the change in measured Earth's gravity caused by the change in centrifugal acceleration resulting from eastbound or westbound velocity by the ships, airplanes, submarines measuring Earth's gravity (see [165]). When moving eastbound, the object's angular velocity is increased (in addition to Earth's rotation), and thus the centrifugal (fictional) force also increases, causing a perceived reduction in gravitational force. If the Earth was flat, on the other hand, there would be no Eötvös effect at all. If the flat Earth is not rotating (as most models posit, with the Sun moving above it in a circular path), obviously there is no centripetal acceleration

happening at all. Even if you adopt a model where the flat Earth rotates about the North Pole, the centripetal acceleration at every point on the surface is parallel to the surface, towards the pole, not directed downwards. So an Eötvös-like effect would actually cause a slight deflection in the angle of gravity, but almost zero change in the magnitude of the gravity. The Eötvös effect shows that not only is the Earth rotating, but that it is rotating about a central point that is underneath the ground, not somewhere on the surface. If you stand on the equator and face east, the surface of the Earth is rotating in the direction you are facing and downwards, not to the left or right. Furthermore, the cosine term shows that at equal latitudes both north and south, the rotation is at the same angle relative to the surface, which can only be the case if the Earth is symmetrical about the equator: i.e. spherical.

69. The French space agency CNES launched the SPOT 1 satellite in 11986 (holocene calendar) into a Sun-synchronous polar orbit, meaning it orbits around the poles and maintains a constant angle to the Sun, as the Earth rotates beneath it. This brought any point on the surface into the imaging field below the satellite every 26 days. SPOT 1 took multiple photos of areas of Earth in different orbital passes, from different locations in space. These images could then be analysed to match features and triangulate the distances to points on the terrain, essentially forming a stereoscopic image of the Earth's surface. This reveals the height of topographic features: hills, mountains, and so on. However stereoscopic heights deduction disagree with ground measurement because of Earth curvature (see [236]). Correction has to been added to counter Earth curvature. This is another evidence against a flat Earth model.

70. Artificially produced skyglow appears prominently in the sky above cities. It makes the whole night sky as seen from within the city brighter, making it difficult or impossible to see fainter stars. The skyglow from a city can be seen from dark places up to hundreds of kilometres away, as a dome of bright sky above the location of the city on the horizon. However, although the skyglow from a city can be seen from such a distance, the much brighter lights of the city itself cannot be seen directly – because they are below the horizon. The fact that you can observe the fainter glow of the sky above a city while not being able to see the lights of the city directly is because of the curvature of the Earth. This is not the only effect of Earth's curvature on the appearance of skyglow; it also effects the brightness of the glow. In the absence of any scattering or absorption, the intensity of light falls off with distance from the source following an inverse square law. Physically, this is because as we know the surface area of spherical shells of increasing radius from a light source increase as the square of the radius. In 11977, Merle F. Walker of Lick Observatory in California published a study of the sky brightness caused by skyglow at varying distances from several southern Californian cities (see [566]). He found an empirical relationship that the intensity of skyglow varies as the inverse of distance to the power of 2.5. This relationship, known as "Walker's law", has been confirmed by later studies, with one notable addition. It only holds for distances up to 50-100 kilometres from the city. When you travel further away from a city, the intensity of the skyglow starts to fall off more rapidly than Walker's law suggests, a little bit faster at first, but then more and more rapidly. This is because as well as the absorption effect, the scattered light path is getting longer and more complex due to the curvature of the Earth. A later study by prominent astronomical light pollution researcher Roy Henry Garstang published in 1989 examined data from multiple cities in Colorado, California, and Ontario to produce a more detailed model of the intensity of skyglow (see [196]). The model was then tested

and verified for multiple astronomical sites in the mainland USA, Hawaii, Canada, Australia, France, and Chile. Importantly for our perspective, the model Garstang came up with requires the Earth's surface to be curved. Interestingly, Garstang also calculated a model for the intensity of skyglow if you assume the Earth is flat ! He did this because it greatly simplifies the geometry and the resulting algebra, to see if it produced results that were good enough. However, quoting directly from the paper: *In general, flat-Earth models are satisfactory for small city distances and observations at small zenith distances. As a rough rule of thumb we can say that for calculations of night-sky brightnesses not too far from the zenith the curvature of the Earth is unimportant for distances of the observer of up to* 50 *[km] from a city, at which distance the effect of curvature is typically* 2%. *For larger distances the curved-Earth model should always be used, and the curved-Earth model should be used at smaller distances when calculating for large zenith distances. In general we would use the curved-Earth model for all cases except for city-center calculations. [...] As would be expected, we find that the inclusion of the curvature of the Earth causes the brightness of large, distant cities to fall off more rapidly with distance than for a flat-Earth model.*

71. Before sunrise or after sunset the Sun is low below the horizon, but the sunlight can reach the clouds above. If there is a mountain between them, it can cast a shadow on the clouds (multiple testimonies and photos of this phenomenon are available on the Internet). This is impossible to occur on a flat Earth and can only happen if the Earth is spherical.

72. The rocket engineers in the northern hemisphere and those in the southern hemisphere (Roscosmos (Russia), JAXA (Japan Aerospace Exploration Agency), ESA (European Space Agency), CNSA (China National Space Administration), ISRO (Indian Space Research Organization), ISA (Israel Space Agency), ISA (Iranian Space Agency), United Kingdom Space Agency (UKSA), NATA (North Korean Space Agency), KASA (Korea AeroSpace Administration), SSAU (State Space Agency of Ukraine) and NASA (National Aeronautics and Space Administration)) definitely want to build their launch facilities as close to the equator as practically possible. And this is a consequence of the fact that the Earth is a rotating globe to spare fuel (and therefore money). On the other hand, if the Earth was flat and non-rotating (as is the case in the most popular flat Earth models), there would be no incentive to build your launch facility anywhere compared to anywhere else, and equatorial locations would not be so coveted. And if the Earth was flat and rotating around the north pole, then you would get your best bang for buck not near the equator, but near the rim of the rotating disc, where the linear speed of rotation is highest. If that were the case, then everyone would be clamouring to build their launch sites as close to Antarctica as possible, which is clearly not the case in the real geoid world.

73. Observations of aurorae have established that they occur in nearly-circular elliptical rings of width equivalent to a few degrees of latitude (i.e. a few hundred kilometres), usually between 10° and 20° from the Earth's magnetic poles. These rings, in the northern and southern hemispheres, are named "auroral ovals". An oval aurora cannot happened on a flat Earth model.

74. Gegenschein is a faintly bright spot in the night sky centered at the antisolar point. The backscatter of sunlight by interplanetary dust causes this optical phenomenon, being a zodiacal light and part of its zodiacal light band. If the Earth is spherical, then night time

corresponds to the Sun being behind the planet. You should see, in the night sky, the point exactly opposite the direction of the Sun. You should be able to see, in that direction, the backscattered light of the Sun from the interplanetary dust cloud. Now let us imagine the Earth is flat. The Sun shines on some part of the Earth at all times, so therefore it must be above the plane of the Earth at all times. At any rate, it should never be possible to look into the sky in the exact opposite direction to the Sun. So there should be no point in the night sky with a peak of backscattered sunlight.

75. From an observation point on Earth, we can watch for meteors. Better than counting by eye, professional astronomers have built specialised radar systems that can detect meteors with greater sensitivity, including during daylight hours, and we can set them up counting meteors all day, every day. Every time the radar detects a meteor, it can record where in the sky it was, what direction it moved, and what time the event occurred. Thinking about the time in particular, we can count how many meteors arrive during any given hour, and average this over many days to produce an hourly rate of meteor events. Many experiments do just this. If the Earth is a rotating geoid in space, orbiting around the Sun, we should expect that the dawn part of the Earth, where the local time is around 6am, should sweep up more meteors than the sunset (as commonly named...) part of the Earth, where the local time is around 6pm. And if the Earth is a geoid, this variation should be sinusoidal – the distinctive smooth shape of a wave as traced out by points rotating around a circle. And this is exactly what professional astronomers observe (see [588]).

76. Gravity Probe B has confirmed the geodetic effect and frame-dragging effect given by General Relativity of Earth and the results are compatible with a geoid Earth but incompatible with a flat Earth. This was to be accomplished by measuring, very precisely, tiny changes in the direction of spin of four gyroscopes contained in an Earth-orbiting satellite at 650 [km] of altitude, crossing directly over the poles. Initial results confirmed the expected geodetic effect to an accuracy of about 0.5% and the frame-dragging effect had been confirmed to within 15% of the expected result. In an article published in the journal Physical Review Letters in 12011 (holocene calendar), the authors reported analysis of the data from all four gyroscopes results in a geodetic drift rate of $-6,601.8 \pm 18.3$ milliarcsecond per year and a frame-dragging drift rate of -37.2 ± 7.2 milliarcsecond per year, in good agreement with the General Relativity predictions of $-6,606.1 \pm 0.28\%$ milliarcsecond per year and $-39.2 \pm 0.19\%$ milliarcsecond per year, respectively (see [170]).

77. The "Blue hour" (night comes from the bottom and not from the top), the "Venus Belt" (pinkish glow that surrounds the observer when the Sun is near the horizon), the "Earth shadow" (the shadow that Earth itself casts through its atmosphere and into outer space toward the antisolar point), the "anticrepuscular rays" (phenomena similar to crepuscular rays that appear however from below the horizon) are all atmospheric phenomenon that have a well detailed description on a geoid Earth model where the Sun sets below the horizon but cannot be explained on a flat Earth model.

78. In the flat Earth model the atmosphere of the Earth is a half-spherical dome. If this was the case the pressure would be proportional to $P \propto \sqrt{R^2 - d^2}$ where d is the distance to the North pole and R is the radius of the flat Earth. Hence the atmospheric pressure near the equator would be half of that of the North pole and the atmospheric pressure at Antarctica would be negligible. However the highest sea-level air pressure ever recorded

was 1083.8 [mb] in Agata, Siberia on December 31, 11968 (holocene calendar); produced by a very cold, dense air mass. The lowest sea-level air pressure ever recorded was 870 [mb] measured on October 12, 11979 (holocene calendar) in the eye of Typhoon Tip near the Philippines as it moved over Guam (a difference of only approximately 15%...).

79. The U.S. Public Land Survey System (PLSS) was established shortly after the Revolutionary War as a way to describe and record parcels of land for legal purposes. In this system, land is divided into pieces named "townships" that are generally 6 miles on a side and containing 36 sections. Each section is about 1 mile square (1.6 kilometer square) and contains 640 acres (2.59 square kilometer). As the Earth is roughly spherical. When you try to cover the surface of a sphere with squares, they are not going to line up perfectly. That means, every so often, sections of the grid shift away from each other. That's why the roads that follow these boundaries must dogleg every twenty-four miles (38.6 kilometres) to counter the diminishing distances.

80. The dynamics of plate tectonics and the consequence of continents drift confirmed thanks to geological techniques cannot work on a flat Earth that has a boundary as some tectonic plates and continents have drifted over and beyond the actual Antarctica. Furthermore, the flat earthers' argument that Antarctica holds water on the flat Earth so it doesn't fall into space doesn't work since the Antarctica continent has only existed for about 80 million years.

81. Sadly many people still think that a scientist or engineer opinion is some kind of evidence. We can see quotes of engineers (...) on social networks (without source obviously) who state that the Earth isn't a globe. For information there are thousand of scientific paper where scientists state that the Earth is an oblate sphere/geoid. It would take hundred of pages to list them all. Let us give however just a few of them:

    - Aristotle (ancient Greek philosopher and polymath) in his book *On the Heavens* in Book II, part 14 (9700): « *Its shape must necessarily be spherical. For every portion of Earth has weight until it reaches the centre, and the jostling of parts greater and smaller would bring about not a waved surface, but rather compression and convergence of part and part until the center is reached* ».

    - Archimedes (ancient Greek mathematician, physicist, engineer, astronomer, and inventor who derived the law of buoyancy) in his book *On Floating Bodies*, Book I 2. 2. (9800): « *The surface of any fluid which is so located that it remains motionless will have the form of a sphere which has the same center as the Earth* ».

    - Strabo (geographer, philosopher, and historian) in his book *Geographica* in Part C, section 11, paragraph 20 (9993): « *...the world is also spherical, and the surface of the Earth is also spherical...* ».

    - Isaac Newton (polymath active as a mathematician, physicist, astronomer, alchemist, theologian, and author) in his book *Treatise of the System of the World* (published posthumously in 11728) clearly states multiples times that the Earth is a globe: « *Let $ABCD$ represent the globe of the Earth...* » (page 42); « *...let fall the perpendicular fG, meeting the globe of the Earth...* » (page 84).

    - Leonhard Euler (mathematician, physicist, astronomer, geographer, logician, and engineer) in his Article 32 *On the Shape of the Earth*, published in Anmerkungen über die Zeitungen, St. Petersburg (11738): « *After the absurd beliefs of the Ancients* *concerning the shape of the Earth had been sufficiently disproved, not only by the philosophers, but by actual travels, it appeared to be a settled matter that the Earth must be perfectly round like a ball* ».

- Joseph-Louis Lagrange (mathematician, physicist and astronomer) in his various papers mentions multiple times the fact that the Earth is a globe. Like in his Memory on the transit of Venus published at l'Académie des Sciences et Belles-Lettres de Berlin, t. XXII, page 8 (11766): « *Now, if by the pole P of the terrestrial globe and the place L whose latitude is...* ».
- Pierre-Simon, Marquis de Laplace (scholar whose work was important to the development of engineering, mathematics, statistics, physics, astronomy, and philosophy) in his famous 5 volumes Treatise in Celestial Mechanics published at l'imprimerie de Crapelet (11798 and 11799) mentions multiples time the globe and spherical shape of the Earth like for example at page 343 section 13: « *Let us now compare the preceding theory with the observations, and see the consequences which result from it on the constitution of the terrestrial globe* ».
- Siméon Denis Poisson (mathematician and physicist) in his quite famous *Treatise on the movement of the earth around its center of gravity* published at the Academie des Sciences at page 10 (11827): « *This can only take place as long as the axis of rotation deviates very little from one of the principal axes, which must be that of the greatest moment of inertia, because the Earth is a spheroid flattened on its poles* ».
- Jean-Baptiste Joseph Fourier (French mathematician and physicist) in his *On the Temperatures of the Terrestrial Sphere and Interplanetary Space* hich originally appeared in Mémoires de l'Acad;emie Royale des Sciences de l'Institut de France VII 570-604 (1827): « *The mathematical theory of heat furnishes several other consequences of this type, whose certainty is independent of all hypotheses regarding the state of the interior of the terrestrial globe* ».
- Albert Einstein (PhD Physicist, Professor, Nobel Prize in Physics), in his Chicago/USA conference (Chicago Lectures) at the Francis Parker School on May 6, 11921 (holocene calendar) on page 28 of his lecture notes refers to the spherical shape of the Earth: « *Wenn jemand um die Erde herumlaeuft, findet er nir- gends ein Ende, ebenso wie bei einer Kugel. Fuellen wir die Erde oder Kugel mit Staeben aus, dann kommen wir zu einem Ende, und wenn es wahr ist, dass es ueberall in der Welt Materie gibt, dann ist die Erde unbegrenzt aber endlich, und wir muessen das Ende erreichen* ».
- Auguste Piccard (PhD in Physics of ETHZ, Switzerland) in *Earth, size and sea* published at Oxford University Press, chapter 6, page 120 (11959): « *The geophysicists have installed a network of stations over the globe* ».
- Will Featherstone (PhD, Professor of Geodesy, Curtin University, Perth, Australia), *On the use of the geoid in geophysics: a case study over the North West Shelf of Australia* (11997) published in Exploration Geophysics, 28(2), 52-57: « *The geoid is the fundamental surface that defines the figure of the Earth* ».
- Edgar Dean Mitchell (PhD, Engineer in Aeronautics and Astronautics from the MIT) NASA astronaut and Lunar Module Pilot of Apollo 14 in 11971 (holocene calendar) and who spent nine hours working on the lunar surface in the Fra Mauro Highlands region, and was the sixth person to walk on the Moon: « *Suddenly, from behind the rim of the moon in long, slow motion movements of immense majesty,*

there emerges a sparkling blue and white jewel, a light, delicate sky blue sphere laced with slowly swirling veils of white rising gradually like a small pearl in a thick sea of black mystery. It takes more than a moment to fully realize this is the earth - home ».

- Hans Volland (PhD, Professor, Radioastronomical Institute, University of Bonn, Germany), *Atmosphere and Earth's rotation* (11996) published in Surveys in Geophysics, 17, 101-144: « *The rotating Earth as a gyroscoping geoid (approximately an oblate spheroid) is subject to external as well as internal torques. External torques are the tidal forces exerted by the Sun and Moon on the equatorial bulge of the Earth which cause the precession of the Earth's spin axis along a cone over* $25,700$ *years... The atmospheric boundary layer where friction mainly occurs consists of three regions (e.g., Stull, 11989; Peixoto and Oort, 11992). The first is a thin laminar surface layer above the ground, typically a few millimeters thick, where viscous effects dominate and relative motions between fluid and solid surface are prevented* ».
- Olena Kuzii and Andrij Rovenchak (Department for Theoretical Physics, Ivan Franko National University of Lviv, Ukraine), *What the gravitation of a flat Earth would look like and why thus the Earth is not actually flat* published in European Journal of Physics (12019), Volume 2, number 3: « *From the derived results, we are able to make some conclusions regarding a flat-Earth world and to demonstrate its contradictions with what is observed in reality* ».
- Thomas A. Herring (PhD in Earth and Planetary Sciences, Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts), *The rotation of the earth* published in Reviews of Geophysics (11991), Volume 29, Issue S1, pages 172-175: « *Earth rotation studies from 1987-1990 have improved accuracy and temporal resolution, leading to a greater understanding of variations in the rotation of the Earth* ».
- Hermann Glauert (aerodynamicist and Principal Scientific Officer of the Royal Aircraft Establishment, Farnborough), *Rotation of the Earth*, Monthly Notices of the Royal Astronomical Society, Vol. 75, p.685 (11915): « *The rotation of the Earth is crucial for regulating day and night, seasons, tides, and measuring time, but its uniformity has been disturbed by precession, nutation, and pole shifts* ».
- American Congress on Surveying and Mapping in its manual *Definitions of Surveying and Associated Terms* (12015), 314 pages we can read: « *Curvature of the Earth, correction for [SURVEYING]- That offset from the tangent to the curve resulting from curvature of the Earth. The offset amounts to about* 8 *cm at* 1 *km distance. That offset which is given in surveying tables as correction for curvature of the Earth. The values given in such tables usually combine the effects of both curvature the Earth and atmospheric refraction. Under normal conditions, the atmospheric refraction compensates partially for the curvature of the Earth, so the correction is about* 6 *cm for a point* 1 *km away* ».
- Raymond Hide (PhD in Physics, visiting professor in the Department of Physics at the University of Oxford) and Jean Dickey (PhD in High-Energy Physics, president of the American Geophysical Union), in *Earth's Variable Rotation* published in Science, volume 253, number 5020, pages 629-637 (11991): « *Earth's rotation fluctuations impact core-mantle boundary topography, interior processes, and atmospheric variations, with implications for oceans and climate* ».

- He Xianqiang (PhD Physical Electronics from the Shanghai Institute of Technical Physics), Knut Stamnes (PhD in Physics and professor of physics at Stevens Institute of Technology), Yan Bai (Second Institute of Oceanography,) in *Effects of Earth curvature on atmospheric correction for ocean color remote sensing* published in Remote Sensing of Environment (Elsevier) Volume 209, pages 118-133 (2018): « *Earth curvature significantly impacts ocean color remote sensing, and considering it in atmospheric correction is crucial for high accuracy* ».
- Bowie William (Honorary Degrees from the University of Edinburgh, Scotland, at the meeting of the International Union of Geodesy and Geophysics) in *Isostasy and the Size and Shape of the Earth* published in Science, volume 39, number 1011, page 697-701 (11914): « *The announcement by Newton in 1687 of his theory that the Earth was an oblate spheroid added much interest to the subject of its exact shape. A meridional are of about* $8^1/_2$ *degrees, extending north and south of Paris, was measured by J. and D. Cassini between the years 1683 and 1716. The results of the first computations of this are indicated that the length of the degree was less at the northern than at the southern end. This gave to the Earth the shape of a prolate spheroid. A great controversy arose over the question as to whether the Earth has the form of an oblate or prolate spheroid, and to settle it the French Academy sent expeditions to Peru in 1735 and to Lapland the following year to measure ares near the equator and within the arctic circle. The result of these measurements confirmed the theory of the oblate spheroid* ».
- Gates, W. Lawrence (PhD in Meteorology from the Massachusetts Institute of Technology) in *Derivation of the Equations of Atmospheric Motion in Oblate Spheroidal Coordinates* published in the Journal of the atmospheric sciences, volume 61, number 20, pages 2478-2487 (12004): « *Earth is more nearly an oblate spheroid than a sphere, making it of theoretical interest to develop atmospheric equations of motion in spheroidal coordinates* ».
- Niklas Axel Mörner af Morlanda (PhD in Quaternary Geology, swedish geologist and geophysicist, head of the paleogeophysics and geodynamics unit at Stockholm University) in *Geoidal topography: Origin and time consistency* published at Springer Geodetic Features of the Ocean Surface and their Implications, pages 205-208 (11984): « *The geoid topography, as a direct measure of the irregular density distribution within the Earth, provides evidence of density changes within the Earth over time, indicating a high-dynamic Earth* ».
- Martine Amalvict (Associate Professor, Institute of Physics of the Globe, Louis Pasteur University) and al. in *The geoid: From geodesy to geophysics and from geophysics to geodesy* published in Surveys in Geophysics, Volume 14, pages 447-494 (11993): « *The geoid is a mathematical and observational concept that studies the inner structure and dynamcis of the Earth through its shape and anomalies* ».
- William Edward Featherstone (PhD, Professor of Geodesy, Curtin University, Perth, Australia) in *On the use of the geoid in geophysics: a case study over the North West Shelf of Australia* published in Exploration Geophysics, volume 28, number 2, pages 52-27 (11997): « *The geoid is the fundamental surface that defines the figure of the Earth. It is approximated by mean sea-level and undulates due to spatial variations in the Earth's gravity field* ».
- + many other thousands of scientific paper quotes...

82. And now also some quotes of people that are not scientists and who confirm Earth isn't flat:

    - Alexeï Leonov (Graduate from the Chuguev Higher Air Force Pilots School) was a Soviet and Russian cosmonaut, Air Force major general that the 18 March 11965 (holocene calendar) became the first person to conduct a spacewalk, exiting the capsule during the Voskhod 2 mission for 12 minutes and 9 seconds: « *The Earth was small, light blue, and so touchingly alone, our home that must be defended like a holy relic. The Earth was absolutely round. I believe I never knew what the world round meant until I saw Earth from Space* ».
    - Sheikh Muszaphar Shukor Al Masrie bin Sheikh Mustapha (Dr in Orthopaedic Surgery) was a spaceflight participant into the International Space Station aboard Soyuz TMA-11 with the Expedition 16 crew on 10 October 12007 (holocene calendar) for a ten days program in space. He shared on Instagram on his personal account the 11 october 12019 (holocene calendar) a private Instagram discussion where he wrote to another muslim: « *In the name of Allah I swear the Earth is Sphere as seen with my own eyes from Space* ». Sheikh Muszaphar also wrote in the description of post: « *I don't simply use the name of God just to prove a point and make people believe me, but if this is what it takes... then I swear in the name of God that the Earth is a sphere as seen with my own eyes from space* ».
    - Felix Baumgartner (Austrian skydiver, daredevil and Base jumper) world record owner for skydiving at an estimated of 39 km the 14 October 12012: « *When you're standing there in a pressure suit, the only thing that you hear is yourself breathing, and you can see the curvature of the Earth; you can see the sky's totally black* ».
    - Colonel Buz Carpenter graduated from the USAF Academy), SR-71 (blackbird) pilot in the US Air Force in an interview video by Lockheed Martin in 2013 : « *I wish I could take everybody up to* $85,000$ *feet because you can see 350 miles in a direction. You see the curvature of the Earth* ».
    - John Beckman (PhD in Theoretical Physics and Astrophysics at the University of Oxford) flight in the Concorde 001 during the Solar eclipse of July 10, 11993 (holocene calendar): « *Everyone was very focused on their work, but I still had time to look out a side window to see the shadow, the half-light and the light that was barely breaking through behind. From our altitude, I could even see the curvature of the Earth, it was incredible. I could also observe the solar corona, and as the moon gradually occulted the solar disk, I saw the chromosphere, which was emitting flashes of red light* ».
    - *Global Navigation for Pilots - International Flight Techniques and Procedures*, 3rd Edition (441 pages) by PhD Dale de Remer and Gary M. Ullrich published (12019) at Aviation Supplies & Academics Newcastle, Washington (USA): « *This difference expresses the ellipticity of the Earth. Sometimes, this is expressed as a ratio of the difference between polar and equatorial diameters to the equatorial diameter* ».
    - Anderle, Richard J. (Computation and Analysis Laboratory, US Naval Surface Weapons Center) in *Determination of the Earth's Geoid by satellite observations* NWL Report No. 2027 (11966): « *Satellite observations can determine the Earth's geoid to about 20 meters accuracy, but further data and computation techniques can improve this to 10 meters or better* ».

- Zhiping Lu et Al. (Department of Geodesy, Information Engineering University Institute of Surveying and Mapping, Zhengzhou, China) in *The Geoid and Different Hight Systems* in Introduction to Geodetic Datum and Geodetic Systems, published in Springer (12014): « *The shape of the Earth studied in geodesy primarily refers to the shape of the geoid, which is the physical shape of the Earth and serves as a reference surface for height determination* ».
- + many other hundreds of engineers, pilots, adventurers quotes...

Remark

Remember that it's obviously for pedagogical reasons that children learn first that Earth is idealized as a perfect sphere, after when teenagers most will study the evidence that it's not a perfect sphere but an oblate-spheroid and later when young adult they will learn the concept of geoid and the relative evidence about Earth being a geoid. Popular science communicators or popular science journals for adults sometimes make the same pedagogical choice.

So finally empirical, reproducible, testable experiences and measurements provide evidence that the Earth is an oblate spheroid and is not shaped like an ostrich egg (like Islam claim and describe it) or like a snake egg (like Hinduism claim and describe it) or flat (like Judaism and Christianism claim and describe it) whose Sun revolves around (geocentrism). For a modern visualisation see `https://icgem.gfz-potsdam.de/vis3d/longtime` for more details see **??**.

12.1.4.9 Stability of Atmospheres

Comparing the liberation velocity and the velocities of various gases, we can explain the stability of certain atmospheres and the absence of others. We have proved in the section of Classical Mechanics (see page 3234) that the liberation velocity of a spherical star was given by the following relation (on which we will come back in the section of General Relativity at page 5720):

$$v_L=\sqrt{\frac{2GM}{R}} \tag{12.1.301}$$

For the Earth, a numerical application gives $v_L=11.2\cdot 10^3\,[\mathrm{m}\cdot\mathrm{s}^{-1}]$ and for the Moon $v_L=2.37\cdot 10^3\,[\mathrm{m}\cdot\mathrm{s}^{-1}]$.

Let us recall that we have proved in the section of Continuum Mechanics during our study of the kinetic temperature (see page 3897) the following relation (virial theorem):

$$v=\sqrt{\frac{3kT}{m}} \tag{12.1.302}$$

Using the molar mass (see section Thermochemistry page 6252):

$$v=\sqrt{\frac{3kT}{M_m/N_\mathrm{A}}} \tag{12.1.303}$$

A numerical application gives for nitrogen $v_{\text{Az}} = 517\,[\text{m}\cdot\text{s}^{-1}]$ and for hydrogen $v_{\text{H}} = 1,934\,[\text{m}\cdot\text{s}^{-1}]$ with an arbitrary temperature of $300\,[\text{K}]$.

So nitrogen is obviously trapped in the Earth's atmosphere. Hydrogen, light gas, more fast is less trapped. The two gases are even less trapped by the Moon.

As provided in the paper [25], a planet waterworld like body should have approximately 0.0268 (i.e. $\sim 1.6\cdot 10^{23}$ [kg]) the mass of the Earth to be sustainable for life (including climate, atmosphere stability, gravity conditions, distance to main star):

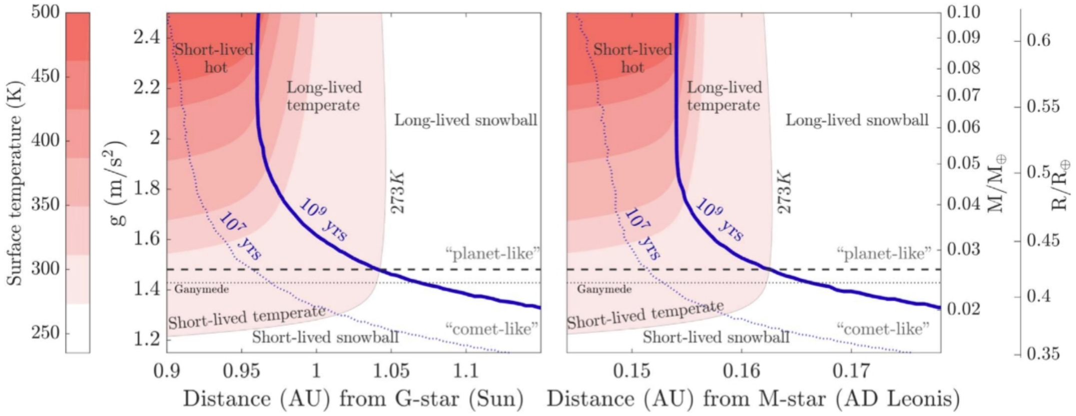


Figure 12.51: Steady-state surface temperatures and water lifetimes obtained by numerical computations

Remark

In fact, the mean square speed is not the only speed of molecules. There is a distribution of velocities. We have indeed study the Maxwell-Boltzmann distribution of a gas at equilibrium in the section of Statistical Mechanics (see page 3561).

The below image of Earth's atmosphere merging with the emptiness of space resembles an abstract painting. It was taken over western China in June 12007 (holocene calendar) by a Space Shuttle crew member. The thin silvery streaks (named "noctilucent clouds") high in the blue area are at a height of about 80 kilometres. The atmosphere at this altitude is very thin. Air pressure here is less than a thousandth of that at sea level. The thin reddish zone in the lower portion of the image is the densest part of the atmosphere. It is here, in a layer named the "troposphere", that practically all weather and cloud formation occur. Ninety percent of Earth's atmosphere occurs within just 16 kilometres of the surface.

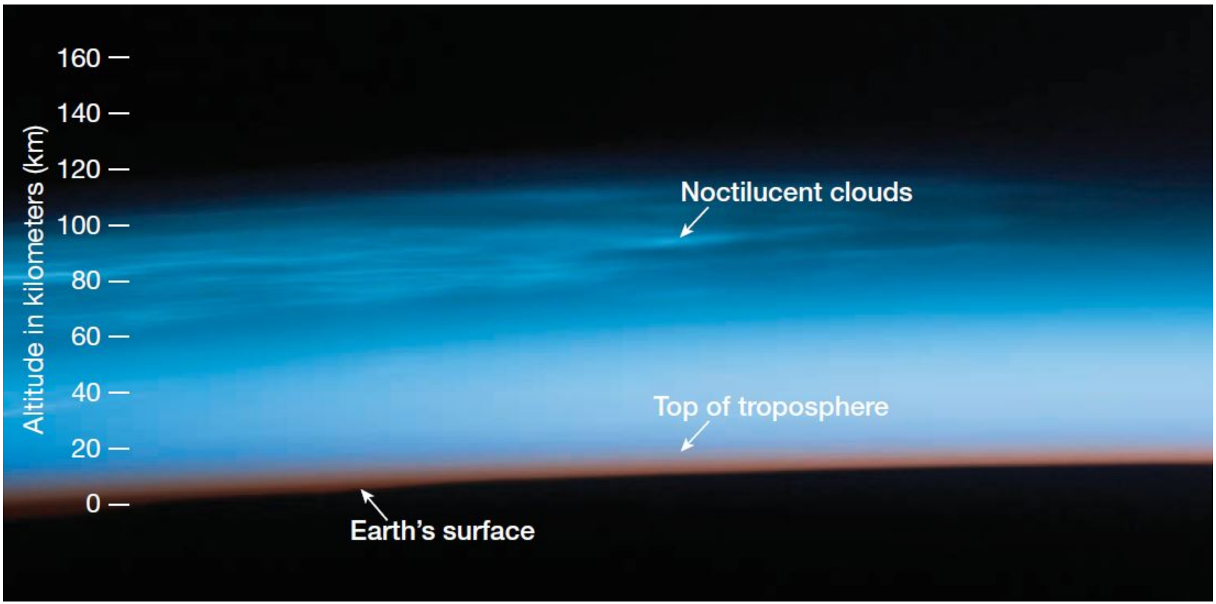

Altitude in kilometers (km)
160
140
120
100
80
60
40
20
0
Noctilucent clouds
Top of troposphere
Earth's surface

Figure 12.52: **Earth's atmosphere side photo (source: NASA)**

The common 8 layers of the Earth atmosphere (in highschool students learn only 4 of them... and by the way... (!) they are not really "layers" as there is no clear distinction and perfect frontier to define each layer as the whole atmosphere is in reality a gradient, that's why in the width given below, some of the "layers" overlap !!!):

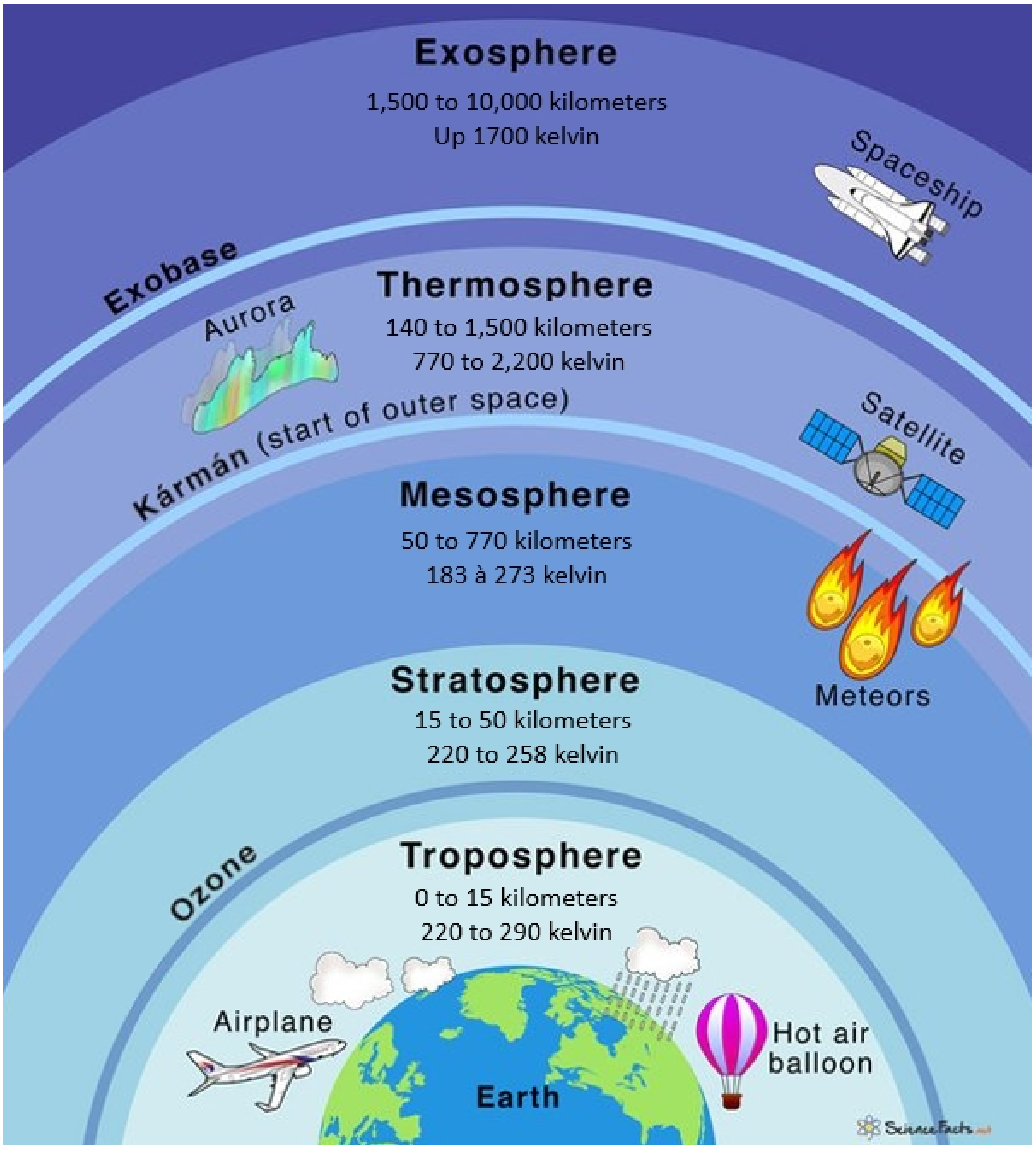

Exosphere
1,500 to 10,000 kilometers
Up 1700 kelvin
Spaceship
Exobase
Aurora
Thermosphere
140 to 1,500 kilometers
770 to 2,200 kelvin
Kármán (start of outer space)
Satellite
Mesosphere
50 to 770 kilometers
183 à 273 kelvin
Meteors
Stratosphere
15 to 50 kilometers
220 to 258 kelvin
Ozone
Troposphere
0 to 15 kilometers
220 to 290 kelvin
Airplane
Hot air balloon
Earth
ScienceFacts

Figure 12.53: **The eight "layers" (...) of the atmosphere**
(source: `https://www.sciencefacts.net/layers-of-atmosphere.html`**, authors: Satyam Bhuyan, Santanu Mukherjee)**

Remark

In 12019 (holocene calendar), scientists discovered that Earth's atmosphere consists of a hydrogen cloud named the "geocorona" that extends to the Moon and beyond. Its density is tiny and practically vacuum. It does not affect space travel. The geocorona extends almost twice the distance to the Moon. Only 0.2 hydrogen atoms per $[\text{cm}^3]$ at the Moon's distance and 70 hydrogen atoms per $[\text{cm}^3]$ at $60,000$ kilometers above Earth's surface.

We can see by the way that if we put $v_L = v$, and seek for M for a given temperature T and radius R we get after a simple algebraic simplification:

$$M_{\text{min}} = \frac{3kTRN_A}{2GM_m} \tag{12.1.304}$$

A numerical application at $T = 293$ [K] for typical air molar mass $2.9 \cdot 10^{-2}$ $[\text{kg} \cdot \text{mole}^{-1}]$ and a ball of radius equal to 1 [m] gives:

$$M_{\text{min}}=\frac{3kTRN_A}{2GM_m}\cong\frac{3\cdot(1.38\cdot 10^{-23})\cdot 293\cdot 1\cdot(6.02\cdot 10^{23})}{2\cdot(6.67\cdot 10^{-11})\cdot(2.9\cdot 10^{-2})}\cong 1.89\cdot 10^{15}\;[\text{kg}] \tag{12.1.305}$$

So it's quite unrealistic to reproduce such a phenomena in a laboratory for now... Indeed, remember for example that our Moon's mass is actually around $7.349\cdot 10^{22}$ [kg] and Earth's mass is actually around $5.9722\cdot 10^{24}$ [kg]. And the biggest human structure existing actually build by humans at the day we write these lines is an Chinese Dam that weights only something in the order of 10^9 [kg]...

12.1.4.10 Planetary equilibrium temperature

Let us consider a planet orbiting its host star. The star emits radiation isotropically, and some fraction of its emittance reaches the planet. The amount of radiation arriving at the planet is referred to as the incident star radiation, M_{star}. The planet has an albedo that depends on the characteristics of its surface and atmosphere, and therefore only absorbs a fraction of the emittance. The planet absorbs the emittance that isn't reflected by the albedo, and heats up. One may assume that the planet radiates energy like a black-body at some temperature according to the Stefan–Boltzmann law. Thermal equilibrium exists when the power supplied by the star is equal to the power emitted by the planet. The temperature at which this balance occurs is the "planetary equilibrium temperature".

The solar flux (i.e. emittance) absorbed (abs) by the planet from the star is equal to the flux emitted (emit) by the planet:

$$M_{\text{abs}}=M_{\text{emit}} \tag{12.1.306}$$

Assuming a fraction of the incident starlight is reflected according to the planet's albedo ρ:

$$(1-\rho)\bar{M}_{\text{star}}=M_{\text{emit}} \tag{12.1.307}$$

where $\bar{M}_{\text{star}}$ represents the area- and time-averaged incident star flux, and may be expressed as:

$$\bar{M}_{\text{star}}=\frac{M_{\text{star}}}{4} \tag{12.1.308}$$

The factor of $1/4$ in the above formula comes from the fact that only a single hemisphere is lit at any moment in time (creates a factor of $1/2$), and from integrating over angles of incident sunlight on the lit hemisphere (creating another obvious factor of $1/2$).

Assuming the planet radiates as a black-body according to the Stefan–Boltzmann law (see section Mechanics page 3706):

$$M(T)=\frac{2\pi^5k^4}{15c^2h^3}T^4=\sigma T^4 \tag{12.1.309}$$

at some equilibrium temperature T_{eq}, a balance of the absorbed and outgoing fluxes produces:

$$(1-\rho)\frac{M_{\text{star}}}{4}=\sigma T_{\text{eq}}^4 \tag{12.1.310}$$

Rearranging the above equation to find the equilibrium temperature leads to:

$$T_{\text{eq}}=\sqrt[4]{\frac{M_{\text{star}}(1-\rho)}{4\sigma}} \tag{12.1.311}$$

Obviously we see that this relation doesn't depends on the distance to the star. So the idea is to rewrite this relation using the temperature at the surface of the star first (using Stefan-Boltzmann law again), to multiply by its surface, and divide it directly afterwards by the surface of the sphere going from the center of the star to the position of the planet at a distance r.

Therefore (we introduce the bolometric[13] intrinsic luminosity L that has for units [W]):

$$T_{\text{eq}}=\sqrt[4]{\frac{\dfrac{4\pi R_{\text{star}}^2\sigma T_{\text{star}}^4}{4\pi r^2}(1-\rho)}{4\sigma}}=\sqrt[4]{\frac{\dfrac{L_{\text{star}}}{4\pi r^2}(1-\rho)}{4\sigma}}=\sqrt[4]{\frac{L_{\text{star}}(1-\rho)}{16\pi r^2\sigma}} \tag{12.1.312}$$

The relation:

$$\boxed{T_{\text{eq}}=\sqrt[4]{\frac{L_{\text{star}}(1-\rho)}{16\pi r^2\sigma_{\text{S.B.}}}}} \tag{12.1.313}$$

is named the "planetary equilibrium temperature".

The equilibrium temperature is neither an upper nor lower bound on actual temperatures on a planet. There are several reasons why measured temperatures deviate from predicted equilibrium temperatures (non-circular orbit, planet precession, etc.).

---

[13] A measure is said to be "bolometric" if it measure the total amount of energy across the full electromagnetic spectrum!

12.1.5 Roche's Limit

The Roche limit is the theoretical distance below which a satellite would begin to break down under the action of tidal forces caused by the celestial body around which it orbit, these forces exceeding the satellite internal cohesion.

We can simplify the problem by considering the satellite liquid, not rotating on itself (no spin), and decomposing it into two small masses m of radius r and volumetric density ρ_S.

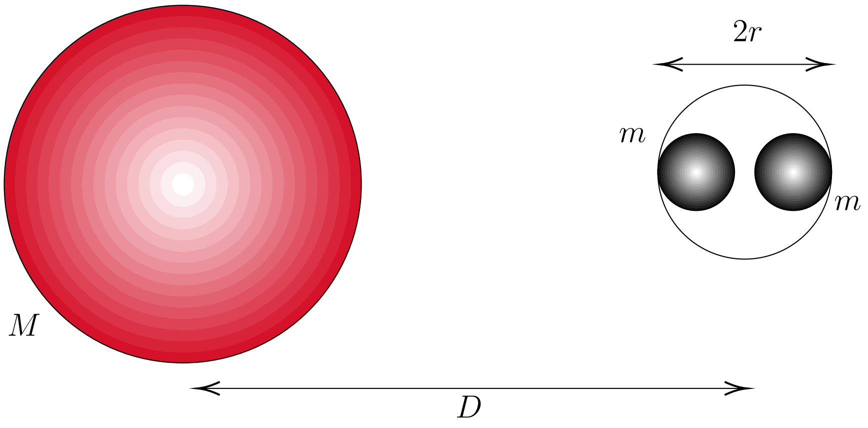


Figure 12.54: **Configuration for the study of the Roche limit**

The planet is a sphere of radius R, mass M, volume density ρ_P, located at a distance D of the satellite axis.

The planet exerts on the satellite the gravitational attraction:

$$F_a = G\frac{mM}{D^2} \tag{12.1.314}$$

The difference of forces between the two masses is:

$$\Delta F_a = G\frac{mM}{(D-r)^2} - G\frac{mM}{(D+r)^2} \tag{12.1.315}$$

We can consider that $r \ll D$, giving:

$$\frac{1}{(D-r)^2} - \frac{1}{(D+r)^2} \cong \frac{1}{D^2}\left(\frac{1}{1-2\dfrac{r}{D}} - \frac{1}{1+2\dfrac{r}{D}}\right) = \frac{1}{D^2}\frac{4\dfrac{r}{D}}{\left(1-2\dfrac{r}{D}\right)\left(1+2\dfrac{r}{D}\right)} \cong 4\frac{r}{D^3} \tag{12.1.316}$$

So the difference in force is

$$\Delta F_a = G\frac{4mM}{D^3}r \tag{12.1.317}$$

If the satellite cohesive force result in the gravitational attraction between the two masses:

$$F_c = G\frac{m^2}{(2r)^2} \tag{12.1.318}$$

The satellite is destroyed if the difference in strength between the two masses is greater than the cohesive force:

$$G\frac{m^2}{4r^2} \geq G\frac{4mM}{D^3}r \tag{12.1.319}$$

But we have the relations:

$$M = \frac{4}{3}\pi R^3 \rho_P \quad \text{and} \quad m = \frac{4}{3}\pi r^3 \rho_S \tag{12.1.320}$$

Therefore we get:

$$\frac{\rho_S}{\rho_P} \leq 16\frac{R^3}{D^3} \tag{12.1.321}$$

and we deduce of it the "Roche limit":

$$\boxed{D_{\text{Roche}} \lesssim 2.5R\sqrt[3]{\frac{\rho_P}{\rho_S}}} \tag{12.1.322}$$

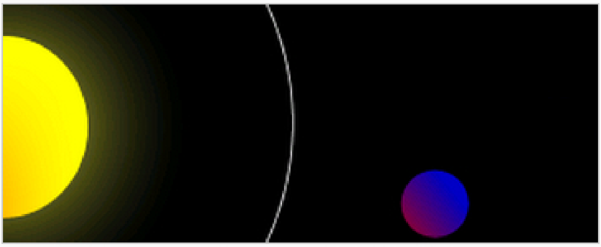
Depending on the approach and the approximations there can be a factor 3 between some results.

Remark

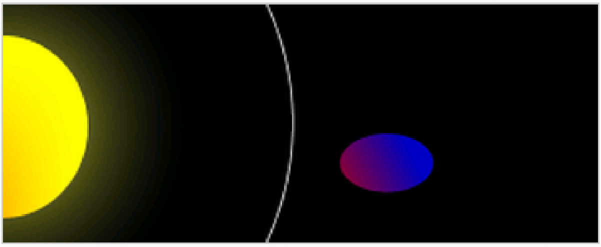
For example the calculations given on Wikipedia consider only the difference in the primary's gravitational pull on the center of the satellite and on the edge of the satellite closest to the primary. This means that the main mass only apply one force momentum. But in fact this is not accurate as what interest us is the difference between the two extremities. This is why there is a factor 2 between the Wikipedia calculations and ours (with our result the satellite will break twice the distance of that given by Wikipedia).

Since in this calculation, we considered a satellite as a two point masses without rotation, and again we have assumed that the satellite's cohesion was provided exclusively by gravitational interactions, this value is just an order of magnitude for rigid body where material is only aggregated by gravity (and not molecular forces).

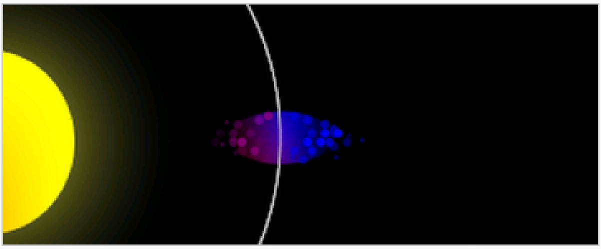
The Roche limit is one (but not the only one!) source of rings around planets as depicted by the illustration below:



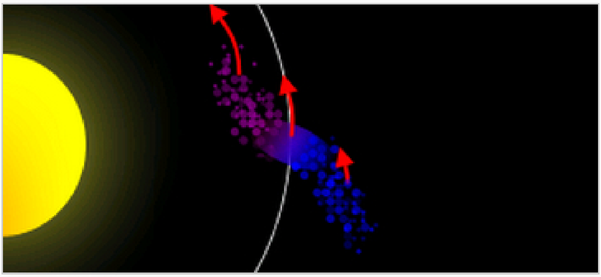
Consider an orbiting mass of fluid held together by gravity, here viewed from above the orbital plane. Far from the Roche limit the mass is practically spherical.



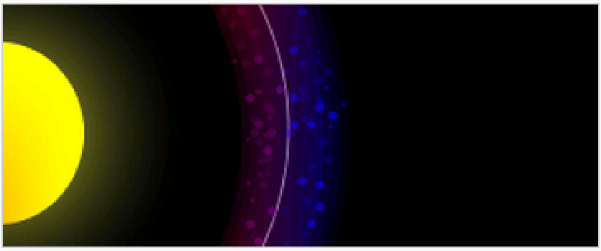
Closer to the Roche limit the body is deformed by tidal forces.



Within the Roche limit the mass's own gravity can no longer withstand the tidal forces, and the body disintegrates.



Particles closer to the primary move more quickly than particles farther away, as represented by the red arrows.



The varying orbital speed of the material eventually causes it to form a ring.

Figure 12.55: **Roche limit as one of the sources of planetary rings**

12.1.6 Keplerian Orbitals

Observation (main tool of the physicist and engineer for recall...) suggests at first glance, that the trajectories of celestial bodies in orbit around stars are indeed conical type (whew!) in the heliocentric reference frame. Knowing this, we can, in order to facilitate the calculation, anticipate the complexity of calculations and express the dynamics directly from a material point in polar coordinates.

As we saw it in the section Vector Calculus (see page 1863), the speed in polar coordinate is expressed by the relation (we changed the angle Greek letter notation to adapt it to the tradition in astronomy):

$$\vec{v}=\frac{\mathrm{d}r}{\mathrm{d}t}\vec{u}_r+r\frac{\mathrm{d}\theta}{\mathrm{d}t}\vec{u}_\theta \tag{12.1.323}$$

where to recall the first term is the radial velocity component and the second component the tangential (angular) velocity!

For acceleration (the proof is also in the section of Vector Calculus at the same page):

$$\vec{a}=\left(\frac{\mathrm{d}^2r}{\mathrm{d}t^2}-r\left(\frac{\mathrm{d}\theta}{\mathrm{d}t}\right)^2\right)\vec{u}_r+\left(2\frac{\mathrm{d}r}{\mathrm{d}t}\frac{\mathrm{d}\theta}{\mathrm{d}t}+r\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2}\right)\vec{u}_\theta \tag{12.1.324}$$

Now that we have the tools, let us get to the case of Keplerian orbits in the case of a static Newtonian field.

There are to our knowledge two main ways of doing the necessary mathematical developments but that do not give (to our knowledge) the same level of detailed results. The first approach provides finer results but is sometimes a bit do-it-yourself sometimes... is based on the use of the radial velocity and an important relation in astronomy, named the "first Binet formula". The second approach is simpler and more elegant, it uses the radial acceleration to approach the problem and we get a special relation named the "second Binet formula".

12.1.6.1 First Binet Formula (radial velocity)

To start with this first approach to the problem, remember that we have already proved earlier that:

$$\vec{F}=G\frac{Mm}{r^2}\vec{u}=-\overrightarrow{\text{grad}}\left(G\frac{Mm}{r}\right)=-\overrightarrow{\text{grad}}(U)=-m\vec{\nabla}\cdot U \tag{12.1.325}$$

However, it is unlikely that the main body is a perfect and homogeneous sphere ... so Astrophysicists have the habit of noting Newtonian potential U under the form:

$$U=-\frac{\mu}{r}(f) \tag{12.1.326}$$

where $\mu=GM$ is named "gravitational constant of the star" (even if it is not always a star...) and where f is a function representing the heterogeneity of the star.

If there is one place in the universe where the laws of mechanics are perfectly verifiable, it is space, because the friction or causes of dissipation are extremely small. Within the field of a single force deriving from a potential, the movement satisfies the conservation of mechanical energy.

Thus we end in the so-named "energy equation", wherein E denotes the "specific energy" per unit weight (kilogram):

$$E_{\text{tot}}=E_{\text{c}}+U=\frac{1}{2}mv^2-m\frac{\mu}{r}=m\left(\frac{1}{2}v^2-\frac{\mu}{r}\right)=mE \tag{12.1.327}$$

Therefore:

$$\boxed{E=\frac{1}{2}v^2-\frac{\mu}{r}} \tag{12.1.328}$$

The Newtonian gravitational force is central, thus having a null torque force at the center O of the main body. This results in the conservation of angular momentum in norm and direction, either:

$$\vec{b}=\vec{r}\times m\vec{v}=m\vec{h}\Rightarrow\vec{h}=\vec{r}\times\vec{v}=K\vec{W} \tag{12.1.329}$$

The vector $\vec{W}$ is the normalized vector of $\vec{b}$ or of $\vec{h}$ named the "reduced momentum". K is the constant of areas (see page 5234) such that:

$$K=\|\vec{h}\|=\|\vec{r}\times\vec{v}\| \tag{12.1.330}$$

We recall to the reader that the norm of the speed expressed in polar coordinates is given by the relation (remember that the both vectors of the polar base are orthogonal and that we can therefore apply the Pythagorean theorem to calculate the norm as it has been proved in the section of Vector Calculus):

$$v^2=\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2+r^2\left(\frac{\mathrm{d}\theta}{\mathrm{d}t}\right)^2 \tag{12.1.331}$$

Which gives us the possibility to write the area constant K as:

$$K=\|\vec{h}\|=\|\vec{r}\times\vec{v}\|=\left\|\vec{r}\times\left(\frac{\mathrm{d}r}{\mathrm{d}t}\vec{u}_r+r\frac{\mathrm{d}\theta}{\mathrm{d}t}\vec{u}_\theta\right)\right\|=r^2\frac{\mathrm{d}\theta}{\mathrm{d}t} \tag{12.1.332}$$

Let us now put ourselves in the orbital plane, in polar coordinates.

Given the relation already proved and known:

$$\vec{v}=\frac{\mathrm{d}r}{\mathrm{d}t}\vec{u}_r+r\frac{\mathrm{d}\theta}{\mathrm{d}t}\vec{u}_\theta \tag{12.1.333}$$

and its squared norm:

$$v^2=\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2+r^2\left(\frac{\mathrm{d}\theta}{\mathrm{d}t}\right)^2 \tag{12.1.334}$$

Or in the case of a central force (conservation of angular momentum):

$$r^2\frac{\mathrm{d}\theta}{\mathrm{d}t}=K\Leftrightarrow\frac{\mathrm{d}\theta}{\mathrm{d}t}=\frac{K}{r^2} \tag{12.1.335}$$

Let's put this in the prior-previous expression of v^2, then we have: Let us put this in the expression prior-previous expression of v^2, then we have:

$$\begin{aligned}v^2&=\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2+r^2\left(\frac{\mathrm{d}\theta}{\mathrm{d}t}\right)^2=v^2=\left(\frac{\mathrm{d}r}{\mathrm{d}t}\frac{\mathrm{d}\theta}{\mathrm{d}\theta}\right)^2+r^2\left(\frac{K}{r^2}\right)^2=\left(\frac{\mathrm{d}\theta}{\mathrm{d}t}\frac{\mathrm{d}r}{\mathrm{d}\theta}\right)^2+\frac{K^2}{r^2}\\&=\left(\frac{K}{r^2}\frac{\mathrm{d}r}{\mathrm{d}\theta}\right)^2+\frac{K^2}{r^2}=\left(-K\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\frac{1}{r}\right)\right)^2+\frac{K^2}{r^2}=K^2\left(\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\frac{1}{r}\right)\right)^2+\frac{1}{r^2}\right)\end{aligned} \tag{12.1.336}$$

The relation:

$$\boxed{v^2=K^2\left(\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\frac{1}{r}\right)\right)^2+\frac{1}{r^2}\right)} \tag{12.1.337}$$

is named "Binet's first formula".

By equating with the expression of v^2 resulting from the conservation of energy that we get earlier above, we have:

$$K^2\left(\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\frac{1}{r}\right)\right)^2+\frac{1}{r^2}\right)=\frac{2\mu}{r}+2E \tag{12.1.338}$$

This gives us a rather complicated differential equation:

$$\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\frac{1}{r}\right)\right)^2=-\left(\frac{1}{r}-\frac{\mu}{K^2}\right)^2+\frac{\mu^2}{K^4}+\frac{2E}{K^2} \tag{12.1.339}$$

And then we wonder how we can get out of such a situation? After some hours of reflection ... we realize that takes to make a substitution. After another hour of neural chaos this ultimately leads to an end.... We decide to put (we have the right to do it!), knowing that r is a function of u and θ:

$$\frac{1}{r}-\frac{\mu}{K^2}=\sqrt{\frac{\mu^2}{K^4}+\frac{2E}{K^2}}u \tag{12.1.340}$$

Let us derivate merrily relatively to θ:

$$\frac{\mathrm{d}\left(\dfrac{1}{r}\right)}{\mathrm{d}\theta}=\sqrt{\frac{\mu^2}{K^4}+\frac{2E}{K^2}}\frac{\mathrm{d}u}{\mathrm{d}\theta} \tag{12.1.341}$$

Substituting in the differential equation:

$$\left(\sqrt{\frac{\mu^2}{K^4}+\frac{2E}{K^2}}\right)\left(\frac{\mathrm{d}u}{\mathrm{d}\theta}\right)^2=-\left(\sqrt{\frac{\mu^2}{K^4}+\frac{2E}{K^2}}\right)u^2+\left(\sqrt{\frac{\mu^2}{K^4}+\frac{2E}{K^2}}\right) \tag{12.1.342}$$

After simplification we get:

$$\frac{\mathrm{d}u}{\mathrm{d}\theta}=\pm\sqrt{1-u^2} \tag{12.1.343}$$

We separate the variables to integrate:

$$\int\pm\frac{\mathrm{d}u}{\sqrt{1-u^2}}=\int\mathrm{d}\theta \tag{12.1.344}$$

We have two solutions according to the sign we choose. However, at the end of the resolution, we notice that the only physically interesting choice is the negative sign. We have proved in the section of Differential and Integral Calculus during our study of common derivatives (see page 1475) that:

$$\begin{aligned}+\frac{\mathrm{d}\arccos(x)}{\mathrm{d}x}&=-\frac{1}{\sqrt{1-x^2}}\\-\frac{\mathrm{d}\arcsin(x)}{\mathrm{d}x}&=-\frac{1}{\sqrt{1-x^2}}\end{aligned} \tag{12.1.345}$$

We will chose the primitive in cosine and therefore we have:

$$\arccos(u)=\theta-\theta_0\Rightarrow u=\cos(\theta-\theta_0) \tag{12.1.346}$$

We leave, by approximation, the constant of integration that would involve very small oscillations in the orbit's path (if you do a study or a homework on this topic, you can transfer me your plots with or without the constant, it would interest me as I don't have time to do it myself).

This allows us to get:

$$\frac{1}{r}=\frac{\mu}{K^2}+\sqrt{\frac{\mu^2}{K^4}+\frac{2E}{K^2}}\cos(\theta-\theta_0) \tag{12.1.347}$$

Now we see that our choice of the sign for the integration is fully justified because now, if we do a little recall on conics (see section Analytical Geometry page 2764), we see that after rearrangement:

$$\begin{aligned}r&=\frac{1}{\dfrac{\mu}{K^2}+\sqrt{\dfrac{\mu^2}{K^4}+\dfrac{2E}{K^2}}\cos(\theta-\theta_0)}=\frac{\dfrac{K^2}{\mu}}{1+\dfrac{K^2}{\mu}\sqrt{\dfrac{\mu^2}{K^4}+\dfrac{2E}{K^2}}\cos(\theta-\theta_0)}\\&=\frac{\dfrac{K^2}{\mu}}{1+\sqrt{1+\dfrac{2EK^2}{\mu^2}}\cos(\theta-\theta_0)}\end{aligned} \tag{12.1.348}$$

So finally we have a relation of the following form if we choose $\theta_0=0$:

$$\boxed{r=\frac{p}{1+e\cos(\theta)}} \tag{12.1.349}$$

where by analogy with the section of Analytical Geometry e is the eccentricity (let us recall that $e=c/a<1$ with a the semi big axes and c the distance to the center of the ellipse to the focal) and p the focal parameter ($p=b^2/a$) of an ellipse. This corresponds well to the trajectories that follow celestial bodies in orbit.

We thus fall back on our the first Kepler "law"... so as we can see it, it can be proven!

In our case, we have after simplification to resume:

$$\boxed{p=\frac{b^2}{a}=\frac{K^2}{\mu} \quad \text{and} \quad e=\frac{b^2}{ah}=\sqrt{1+\frac{2EK^2}{\mu^2}}} \tag{12.1.350}$$

where (for recall) K is the areas constant :

$$K=\|\vec{h}\|=\|\vec{r}\times\vec{v}\| \tag{12.1.351}$$

and μ is the gravitation constant of the celestial body:

$$\mu=GM \tag{12.1.352}$$

and finally E the specific energy:

$$E=\frac{1}{2}v^2-\frac{\mu}{r} \tag{12.1.353}$$

The reader could be able to check himself as we have seen in the section of Analytical Geometry in our study conical that if:

- If $E=0$ such that $e=1$ we have an opened orbit in the form of a parabola
- If $E>0$ such that $e>1$ we have an opened orbit in the form of a hyperbola
- If $E\leq 0$ such that $0\geq e<1$ we have a closed orbit in the form of an ellipse or a circle

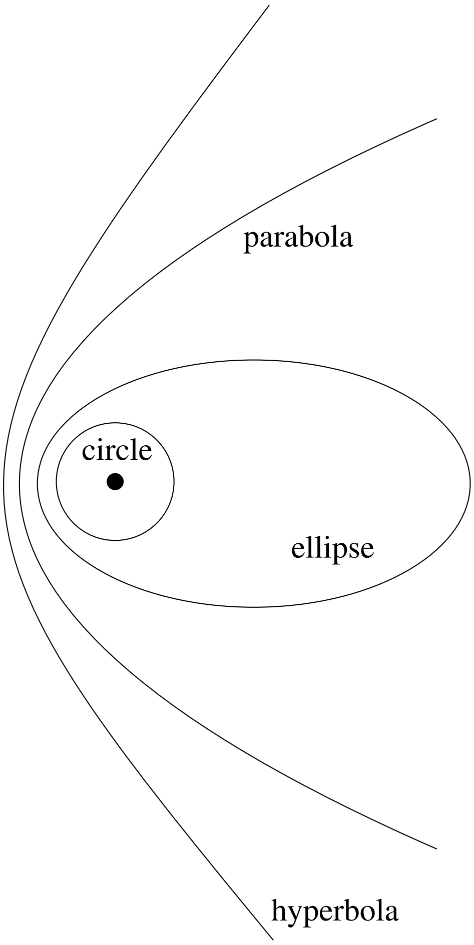


Figure 12.56: Reminders of conical but "orbit" oriented

Finally, if we inject:

$$r=\frac{p}{1+e\cos(\theta)} \tag{12.1.354}$$

in the first Binet formula:

$$v^2=K^2\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\frac{1}{r}\right)^2+\frac{1}{r^2}\right) \tag{12.1.355}$$

then we get the velocity in any point of the ellipse based on the primary variable parameter which is therefore the angle:

$$v^2=||\vec{r}\times\vec{v}||^2\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\frac{1+e\cos(\theta)}{p}\right)^2+\frac{p^2}{(1+e\cos(\theta))^2}\right) \tag{12.1.356}$$

Hence:

$$\boxed{v=||\vec{r}\times\vec{v}||\sqrt{\frac{-2(1+e\cos(\theta))e\sin(\theta)}{p^2}+\frac{p^2}{(1+e\cos^2(\theta))^2}}} \tag{12.1.357}$$

There is another famous and simple relation of the radial velocity named the vis-viva equation that we will in the section of Aerospace Engineering at page 8619:

$$v=\sqrt{GM\left(\frac{2}{r}-\frac{1}{a}\right)} \tag{12.1.358}$$

12.1.6.2 Second Binet Formula (radial acceleration)

Let us now see the approach based on the radial acceleration which, while being more elegant, allows us to get a result less fine-tuned on the ellipse parameters.

So we start from the expression of the acceleration in polar coordinates (see section Vector Calculus page 1863):

$$\vec{a}=\left(\frac{\mathrm{d}r}{\mathrm{d}t^2}\right)\vec{u}_r+\left(2\frac{\mathrm{d}r}{\mathrm{d}t}\frac{\mathrm{d}\theta}{\mathrm{d}t}+r\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2}\right)\vec{u}_\theta \tag{12.1.359}$$

We can simplify the writing of the second term:

$$2\frac{\mathrm{d}r}{\mathrm{d}t}\frac{\mathrm{d}\theta}{\mathrm{d}t}+r\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2}=\frac{1}{r}\frac{\mathrm{d}}{\mathrm{d}t}\left(r^2\frac{\mathrm{d}\theta}{\mathrm{d}t}\right) \tag{12.1.360}$$

Now we have seen just above that:

$$r^2\frac{\mathrm{d}\theta}{\mathrm{d}t}=K=c^{te} \tag{12.1.361}$$

and so:

$$\frac{1}{r}\frac{\mathrm{d}}{\mathrm{d}t}\left(r^2\frac{\mathrm{d}\theta}{\mathrm{d}t}\right)=0 \tag{12.1.362}$$

Then the acceleration is reduced to:

$$\vec{a}=\left(\frac{\mathrm{d}r}{\mathrm{d}t^2}-r\left(\frac{\mathrm{d}\theta}{\mathrm{d}t}\right)^2\right)\vec{u}_r \tag{12.1.363}$$

We can eliminate the time by writing:

$$\frac{\mathrm{d}r}{\mathrm{d}t}=\frac{\mathrm{d}r}{\mathrm{d}\theta}\frac{\mathrm{d}\theta}{\mathrm{d}t} \tag{12.1.364}$$

and:

$$\frac{\mathrm{d}^2r}{\mathrm{d}t^2}=\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\frac{\mathrm{d}r}{\mathrm{d}\theta}\frac{\mathrm{d}\theta}{\mathrm{d}t}\right)\frac{\mathrm{d}\theta}{\mathrm{d}t} \tag{12.1.365}$$

Then we get:

$$\begin{aligned}\vec{a}&=\left(\frac{\mathrm{d}r}{\mathrm{d}t^2}\vec{u}_r-r\left(\frac{\mathrm{d}\theta}{\mathrm{d}t}\right)^2\right)\vec{u}_r=\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\frac{\mathrm{d}r}{\mathrm{d}\theta}\frac{\mathrm{d}\theta}{\mathrm{d}t}\right)\frac{\mathrm{d}\theta}{\mathrm{d}t}-r\left(\frac{\mathrm{d}\theta}{\mathrm{d}t}\right)^2\right)\vec{u}_r\\&=\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\frac{\mathrm{d}r}{\mathrm{d}\theta}\frac{K}{r^2}\right)\frac{\mathrm{d}\theta}{\mathrm{d}t}-r\left(\frac{K}{r^2}\right)^2\right)\vec{u}_r=\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\frac{\mathrm{d}r}{\mathrm{d}\theta}\frac{K}{r^2}\right)\frac{K}{r^2}-r\left(\frac{K}{r^2}\right)^2\right)\vec{u}_r\\&=\frac{K^2}{r^2}\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\frac{\mathrm{d}r}{\mathrm{d}\theta}\frac{1}{r^2}\right)-\frac{1}{r}\right)\vec{u}_r=\frac{K^2}{r^2}\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\left(-\frac{\mathrm{d}\left(\dfrac{1}{r}\right)}{\mathrm{d}\theta}\right)-\frac{1}{r}\right)\vec{u}_r\\&=-\frac{K^2}{r^2}\left(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\left(\frac{1}{r}\right)+\frac{1}{r}\right)\vec{u}_r\end{aligned}\tag{12.1.366}$$

And so it comes to the standard "Binet's second formula":

$$\boxed{a=-\frac{K^2}{r^2}\left(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\left(\frac{1}{r}\right)+\frac{1}{r}\right)}\tag{12.1.367}$$

But according to Newton's second law and his law of gravitation, we have:

$$F=ma=\frac{GMm}{r^2}\Rightarrow=\frac{GM}{r^2}=\frac{\mu}{r^2}\tag{12.1.368}$$

We then get another form of the Binet's second formula:

$$\frac{\mu}{r^2}=-\frac{K^2}{r^2}\left(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\left(\frac{1}{r}\right)+\frac{1}{r}\right)\tag{12.1.369}$$

Or after simplification and choosing the sign of the acceleration at our convenience to get rid of the "-" sign, we have:

$$\mu=K^2\left(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\left(\frac{1}{r}\right)+\frac{1}{r}\right)\tag{12.1.370}$$

By isolating the constants, we get:

$$\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\left(\frac{1}{r}\right)+\frac{1}{r}=\frac{\mu}{K^2}\tag{12.1.371}$$

After a change of variables we recognize the particular case of a differential equation of the second order we have already met several times so far in various sections of this textbook and that will meet again:

$$\frac{\mathrm{d}^2u}{\mathrm{d}\theta^2}+u=\frac{u}{K^2}\tag{12.1.372}$$

As it is customary, however, we will show the details of the resolution. The equation without second member is (see section Differential and Integral Calculus page 1610):

$$\frac{\mathrm{d}^2u}{\mathrm{d}\theta^2}+u=0\tag{12.1.373}$$

We then have the discriminant that is negative since:

$$\frac{-a\pm\sqrt{a^2-4b}}{2}=\frac{\sqrt{-4}}{2}=\sqrt{-1} \tag{12.1.374}$$

We then saw in the section of Differential and Integral Calculus (see page 1610), that in this situation the solution of the homogeneous equation was of the form:

$$y_h=C'e^{\alpha x}\cos(\beta t-\phi) \tag{12.1.375}$$

Thus in the situation we are concerned, we have:

$$u_h=Ae^{0\cdot\theta}\cos(1\cdot\theta-\theta_0)=A\cos(\theta-\theta_0) \tag{12.1.376}$$

We inject the solution into the homogeneous differential equation with second member:

$$\frac{\mathrm{d}^2A\cos(\theta-\theta_0)}{\mathrm{d}\theta^2}+A\cos(\theta-\theta_0)=\frac{\mu}{K^2} \tag{12.1.377}$$

and we see immediately that for the equality to be satisfied, the general solution is:

$$u=\frac{1}{r}=A\cos(\theta-\theta_0)+\frac{\mu}{K^2} \tag{12.1.378}$$

Or after rearrangement:

$$r=\frac{\dfrac{K^2}{\mu}}{1+\dfrac{K^2A}{\mu}\cos(\theta-\theta_0)}=\frac{p}{1+e\cos(\theta-\theta_0)} \tag{12.1.379}$$

And choosing the initial angle as zero, so we find well:

$$\boxed{r=\frac{p}{1+e\cos(\theta)}} \tag{12.1.380}$$

at the difference with the first method of resolution that the value of the constant A is unknown.

Let us now come back on:

$$u=\frac{1}{r}=A\cos(\theta-\theta_0)+\frac{\mu}{K^2} \tag{12.1.381}$$

By expliciting:

$$\frac{1}{r}=A\cos(\theta-\theta_0)+\frac{GM}{\left(r^2\dfrac{\mathrm{d}\theta}{\mathrm{d}t}\right)^2} \tag{12.1.382}$$

And as (see section Classical Mechanics page 3280):

$$r\frac{\mathrm{d}\theta}{\mathrm{d}t}\cong v_{\text{rot}} \tag{12.1.383}$$

So if we choose a particular point of reference of the path (not necessarily circular path), we have:

$$r_0\frac{\mathrm{d}\theta}{\mathrm{d}t}\cong v_0 \tag{12.1.384}$$

Then we have:

$$\frac{1}{r}\cong A\cos(\theta-\theta_0)+\frac{GM}{r_0^2v_0^2} \tag{12.1.385}$$

If we put the phase shift as zero relatively to the reference chosen radius earlier above, the expression simplifies to:

$$\frac{1}{r}\cong A\cos(\theta)+\frac{GM}{r_0^2v_0^2} \tag{12.1.386}$$

To determine the constant A, we place ourselves in the case where $\theta=0$ and imposes that the radius r is the initial radius measured when this angle is zero. Then we have:

$$\frac{1}{r_0}\cong A+\frac{GM}{r_0^2v_0^2} \tag{12.1.387}$$

This implies immediately:

$$\boxed{\frac{1}{r}\cong\frac{1}{r_0}\left(1-\frac{GM}{r_0v_0^2}\right)\cos(\theta)+\frac{GM}{r_0^2v_0^2}} \tag{12.1.388}$$

Thus after elementary rearrangements and simplifications:

$$r\cong\frac{\dfrac{r_0^2v_0^2}{GM}}{1+(r_0v_0^2-GM)\cos(\theta)} \tag{12.1.389}$$

And therefore we have a direct correspondence:

$$e\cong r_0v_0^2-GM \tag{12.1.390}$$

And as the eccentricity e is known for a circular, parabolic, elliptical or another trajectory of the conical type... it gets us very easy to deduce the velocity at the particular point of the initial radius r of the studied object.

The closest distance of the object orbiting around its central star (focus), will be given by the value that can takes r in the relation:

$$\frac{1}{r}\cong\frac{1}{r_0}\left(1-\frac{GM}{r_0v_0^2}\right)\cos(\theta)+\frac{GM}{r_0^2v_0^2} \tag{12.1.391}$$

if we impose $\theta=0$.

In the case of an elliptical orbit it is the "perigee" to be assimilated to the initial radius as the point where the measurement of the radial velocity was the most accurate.

The farthest distance from the focus will be given by putting the angle as being $180°$ (π) and then we name it "apogee".

12.1.6.3 Keplerian orbital period

The Kepler's law of equal areas allows, as we already know, to calculate the Keplerian orbital period T. In fact, the area S of the ellipse being equal to $S = \pi ab$ (see section Geometric Shapes page 2917) and having already determined during our definition of the angular momentum that (see section Classical Mechanics page 3244):

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \frac{K}{2} \tag{12.1.392}$$

It comes naturally:

$$\frac{\pi ab}{T} = \frac{K}{2} \tag{12.1.393}$$

Moreover, the study of conics (see section Analytical Geometry page 2779) has showed us that:

$$p = \frac{b^2}{a} = \frac{T^2 K^2}{4\pi^2 a^3} \tag{12.1.394}$$

and we have defined above:

$$p = \frac{K^2}{\mu} \tag{12.1.395}$$

So we have the relation:

$$T = \sqrt{\frac{a^3 4\pi^2}{\mu}} \tag{12.1.396}$$

and then we fall back again on the third Kepler's law:

$$\frac{a^3}{T^2} = \frac{GM}{4\pi^2} \tag{12.1.397}$$

which validates our previous calculations.

Obviously the latter relation is only available for T where the corresponding speed is non relativistic otherwise we have to use General Relativity tools (see page 5720).

Before we continue... here are some useful data for this section and for that for Aerospace Engineering (i.e. "space dynamics") relatively to our knowledge in the state of year 12018 according to holocene calendar (various sources were used for these tables but the main one is Wikipedia):

Mercury ☿	
Mass	$3.3011 \cdot 10^{23}$ [kg]
Mean density	$5.427 \left[\text{g} \cdot \text{cm}^{-3}\right]$
Mean radius	$2,439.7$ [km]
Equatorial radius	$2,439.7$ [km]
Polar radius	$2,439.7$ [km]
Circumference	$15,239.08$ [km]
Surface gravity	$3.7 \left[\text{m} \cdot \text{s}^{-2}\right]$
Escape velocity	$4.25 \left[\text{km} \cdot \text{s}^{-1}\right]$
Sidereal rotation period	58.646 d
Equatorial rotation velocity	$10.892 \left[\text{km} \cdot \text{h}^{-1}\right]$
Axial tilt	$0.034°$
Ecliptic inclination	$7.00°$
Surface pressure	0 [hPa] (at MSL)
Aphelion	$69,816,900$ [km]
Perihelion	$46,001,200$ [km]
Semi-major axis	$57,909,050$ [km]
Eccentricity	0.205630
Orbital period	87.969 [d]
Average orbital speed	$47.362 \left[\text{km} \cdot \text{s}^{-1}\right]$
Average Temperature	363 [K]
Number of satellites	0

Table 12.5: **Some Mercury characteristics**

Venus ♀	
Mass	$4.8675 \cdot 10^{24}$ [kg]
Mean density	$5.243 \left[\text{g} \cdot \text{cm}^{-3}\right]$
Mean radius	$6,051.8$ [km]
Equatorial radius	$6,051.8$ [km]
Polar radius	$6,051.8$ [km]
Circumference	$38,019.55$ [km]
Surface gravity	$8.87 \left[\text{m} \cdot \text{s}^{-2}\right]$
Escape velocity	$10.36 \left[\text{km} \cdot \text{s}^{-1}\right]$
Sidereal rotation period	-243.025 d (retrograde)
Equatorial rotation velocity	$6.52 \left[\text{km} \cdot \text{h}^{-1}\right]$
Axial tilt	$2.64°$
Ecliptic inclination	$3.39°$
Surface pressure	$90,000$ [hPa] (at MSL)
Aphelion	$108,939,000$ [km]
Perihelion	$107,477,000$ [km]
Semi-major axis	$108,208,000$ [km]
Eccentricity	0.006772
Orbital period	224.701 [d]
Average orbital speed	$35.02 \left[\text{km} \cdot \text{s}^{-1}\right]$
Average Temperature	737 [K]
Number of satellites	0

Table 12.6: **Some Venus characteristics**

Earth ♁	
Mass	$5.97237 \cdot 10^{24}$ [kg]
Mean density	$5.514\ [\mathrm{g} \cdot \mathrm{cm}^{-3}]$
Mean radius	$6,371.0$ [km]
Equatorial radius	$6,378.1$ [km]
Polar radius	$6,356.8$ [km]
Circumference	$40,075.017$ [km]
Surface gravity	$9.807\ [\mathrm{m} \cdot \mathrm{s}^{-2}]$
Escape velocity	$11.186\ [\mathrm{km} \cdot \mathrm{s}^{-1}]$
Sidereal rotation period	23h 56m 4.100s
Equatorial rotation velocity	$0.465\ [\mathrm{km} \cdot \mathrm{s}^{-1}]$
Axial tilt	23.439281°
Ecliptic inclination	0.00005°
Surface pressure	101.325 [hPa] (at MSL)
Aphelion	$152,100,000$ [km]
Perihelion	$147,095,000$ [km]
Semi-major axis	$149,598,023$ [km]
Eccentricity	0.0167086
Orbital period	365.256363004 [d]
Average orbital speed	$29.78\ [\mathrm{km} \cdot \mathrm{s}^{-1}]$
Average Temperature	287 [K]
Number of satellites	1

Table 12.7: **Some Earth characteristics**

Mars ♂	
Mass	$6.4171 \cdot 10^{23}$ [kg]
Mean density	$3.9335\ [\mathrm{g} \cdot \mathrm{cm}^{-3}]$
Mean radius	$3,389.5$ [km]
Equatorial radius	$3,396.2$ [km]
Polar radius	3376.2 [km]
Circumference	21296.856 [km]
Surface gravity	$3.711\ [\mathrm{m} \cdot \mathrm{s}^{-2}]$
Escape velocity	$5.027\ [\mathrm{km} \cdot \mathrm{s}^{-1}]$
Sidereal rotation period	24h 37m 22s
Equatorial rotation velocity	$0.241\ [\mathrm{km} \cdot \mathrm{s}^{-1}]$
Axial tilt	25.19°
Ecliptic inclination	1.85°
Surface pressure	$2-10$ [hPa] (at MSL)
Aphelion	$249,200,000$ [km]
Perihelion	$206,700,000$ [km]
Semi-major axis	$227,939,200$ [km]
Eccentricity	0.0934
Orbital period	686.971 [d]
Average orbital speed	$24.007\ [\mathrm{km} \cdot \mathrm{s}^{-1}]$
Average Temperature	213 [K]
Number of satellites	2

Table 12.8: **Some Mars characteristics**

Jupiter ♃	
Mass	$1.8982 \cdot 10^{27}$ [kg]
Mean density	$1.326\ [\mathrm{g} \cdot \mathrm{cm}^{-3}]$
Mean radius	69,911 [km]
Equatorial radius	71,492 [km]
Polar radius	66,854 [km]
Circumference	439,263.76 [km]
Surface gravity	$24.79\ [\mathrm{m} \cdot \mathrm{s}^{-2}]$
Escape velocity	$59.5\ [\mathrm{km} \cdot \mathrm{s}^{-1}]$
Sidereal rotation period	9h 55m 30s
Equatorial rotation velocity	$12.6\ [\mathrm{km} \cdot \mathrm{s}^{-1}]$
Axial tilt	3.13°
Ecliptic inclination	1.303°
Surface pressure	– [hPa] (at MSL)
Aphelion	816,620,000 [km]
Perihelion	740,520,000 [km]
Semi-major axis	778,570,000 [km]
Eccentricity	0.0489
Orbital period	4,332.59 [d]
Average orbital speed	$13.07\ [\mathrm{km} \cdot \mathrm{s}^{-1}]$
Average Temperature	165 [K]
Number of satellites	69

Table 12.9: **Some Jupiter characteristics**

Saturn ♄	
Mass	$5.6834 \cdot 10^{26}$ [kg]
Mean density	$0.687\ [\mathrm{g} \cdot \mathrm{cm}^{-3}]$
Mean radius	58,232 [km]
Equatorial radius	60,268 [km]
Polar radius	54,364 [km]
Circumference	365,882.44 [km]
Surface gravity	$10.44\ [\mathrm{m} \cdot \mathrm{s}^{-2}]$
Escape velocity	$35.5\ [\mathrm{km} \cdot \mathrm{s}^{-1}]$
Sidereal rotation period	10h 33m
Equatorial rotation velocity	$9.87\ [\mathrm{km} \cdot \mathrm{s}^{-1}]$
Axial tilt	26.73°
Ecliptic inclination	2.485°
Surface pressure	– [hPa] (at MSL)
Aphelion	1,514,000,000 [km]
Perihelion	1,352,550,000 [km]
Semi-major axis	1,433,530,000 [km]
Eccentricity	0.0565
Orbital period	10,759.22 [d]
Average orbital speed	$9.68\ [\mathrm{km} \cdot \mathrm{s}^{-1}]$
Average Temperature	100 – 160 [K]
Number of satellites	62

Table 12.10: **Some Saturn characteristics**

Uranus (gaseous) ⛢	
Mass	$8.6810 \cdot 10^{25}$ [kg]
Mean density	$1.27\,[\mathrm{g}\cdot\mathrm{cm}^{-3}]$
Mean radius	$25,362$ [km]
Equatorial radius	$25,559$ [km]
Polar radius	$24,973$ [km]
Circumference	$159,354.14$ [km]
Surface gravity	$8.69\,[\mathrm{m}\cdot\mathrm{s}^{-2}]$
Escape velocity	$21.3\,[\mathrm{km}\cdot\mathrm{s}^{-1}]$
Sidereal rotation period	-17h 14m 24s (retrograde)
Equatorial rotation velocity	$2.59\,[\mathrm{km}\cdot\mathrm{s}^{-1}]$
Axial tilt	97.77°
Ecliptic inclination	0.773°
Surface pressure	– [hPa] (at MSL)
Aphelion	$3,008,000,000$ [km]
Perihelion	$2,742,000,000$ [km]
Semi-major axis	$2,875,000,000$ [km]
Eccentricity	0.046
Orbital period	$30,688.5$ [d]
Average orbital speed	$6.80\,[\mathrm{km}\cdot\mathrm{s}^{-1}]$
Average Temperature	76 [K]
Number of satellites	27

Table 12.11: **Some Uranus characteristics**

Neptune ♆	
Mass	$1.0243 \cdot 10^{26}$ [kg]
Mean density	$1.638\,[\mathrm{g}\cdot\mathrm{cm}^{-3}]$
Mean radius	$24,622$ [km]
Equatorial radius	$24,764$ [km]
Polar radius	$24,341$ [km]
Circumference	$154,704.58$ [km]
Surface gravity	$11.15\,[\mathrm{m}\cdot\mathrm{s}^{-2}]$
Escape velocity	$23.5\,[\mathrm{km}\cdot\mathrm{s}^{-1}]$
Sidereal rotation period	16h 6m 36s
Equatorial rotation velocity	$2.68\,[\mathrm{km}\cdot\mathrm{s}^{-1}]$
Axial tilt	28.32°
Ecliptic inclination	1.767°
Surface pressure	– [hPa] (at MSL)
Aphelion	$4,540,000,000$ [km]
Perihelion	$4,460,000,000$ [km]
Semi-major axis	$4,500,000,000$ [km]
Eccentricity	0.009
Orbital period	$60,182$ [d]
Average orbital speed	$5.43\,[\mathrm{km}\cdot\mathrm{s}^{-1}]$
Average Temperature	72 [K]
Number of satellites	14

Table 12.12: **Some Neptune characteristics**

Pluto (dwarf planet) ♇	
Mass	$1.303 \cdot 10^{22}$ [kg]
Mean density	$1.854\,[\mathrm{g} \cdot \mathrm{cm}^{-3}]$
Mean radius	$1,188.3$ [km]
Equatorial radius	$-$ [km]
Polar radius	$-$ [km]
Circumference	$7,466.309$ [km]
Surface gravity	$0.620\,[\mathrm{m} \cdot \mathrm{s}^{-2}]$
Escape velocity	$1.212\,[\mathrm{km} \cdot \mathrm{s}^{-1}]$
Sidereal rotation period	6d 9h 17m 36s
Equatorial rotation velocity	$0.013\,[\mathrm{km} \cdot \mathrm{s}^{-1}]$
Axial tilt	$122.53°$
Ecliptic inclination	$17.16°$
Surface pressure	$0,01$ [hPa] (at MSL)
Aphelion	$7,375,930,000$ [km]
Perihelion	$4,436,820,000$ [km]
Semi-major axis	$5,906,380,000$ [km]
Eccentricity	0.2488
Orbital period	$90,560$ [d]
Average orbital speed	$4.67\,[\mathrm{km} \cdot \mathrm{s}^{-1}]$
Average Temperature	44 [K]
Number of satellites	5

Table 12.13: **Some Pluto characteristics**

12.1.6.4 Average distance between two planets

Let us consider two planets with the special and ideal case where the orbit are perfectly circular around a given star like below:

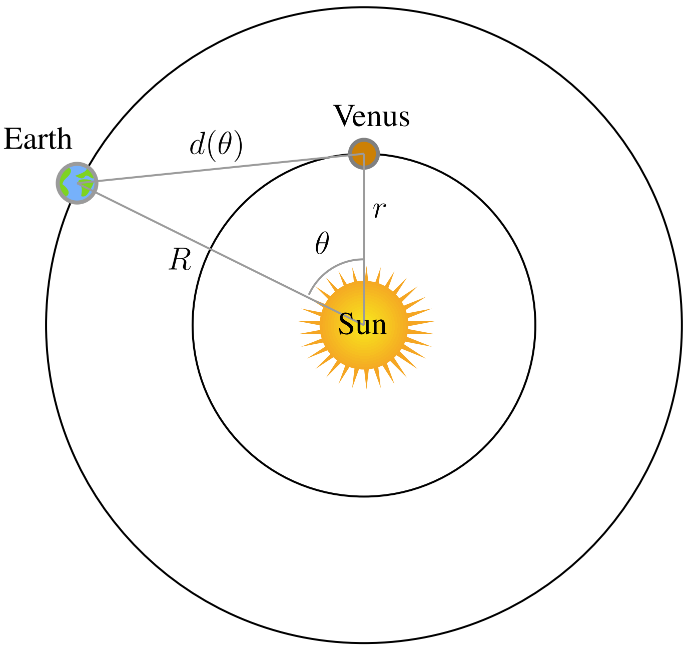


Figure 12.57: **Configuration for the average distance between to planets**

Using the law of cosines proven in the section of Trigonometry at page 2524, we have:

$$d^2(\theta)=R^2+r^2-2Rr\cos(\theta) \qquad (12.1.398)$$

Hence:

$$d(\theta)=\sqrt{R^2+r^2-2Rr\cos(\theta)} \qquad (12.1.399)$$

The average distance is obviously given by:

$$\bar{d}=\frac{1}{2\pi}\int\limits_0^{2\pi} d(\theta)\mathrm{d}\theta \qquad (12.1.400)$$

Therefore:

$$\bar{d}=\frac{1}{2\pi}\int\limits_0^{2\pi} d(\theta)\mathrm{d}\theta=\frac{1}{2\pi}\int\limits_0^{2\pi}\sqrt{R^2+r^2-2Rr\cos(\theta)}\mathrm{d}\theta \qquad (12.1.401)$$

It's quite a complex integral and as far as we know it must be solved analytically!

Let us raise now more sophisticated mathematical question: What is the true average distance of a planet from the Sun? More generally, we ask: For any ellipse, what is the true average distance from all points $P(x,y)$ on the ellipse to a particular focus, $F(c,0)$?

We have proven in the section of Analytical Geometry (see page 2783) that the distance of the path of an ellipse (where $e<1$) to one of its focus was given by:

$$d(\theta)=\overline{PF}=a(1-e\cos(\theta)) \tag{12.1.402}$$

The average distance to a focus is therefore:

$$\bar{d}=\frac{1}{2\pi}\int_0^{2\pi}a(1-e\cos(\theta))\mathrm{d}\theta=\frac{1}{2\pi}\left(a\theta-ea\sin(\theta)\right)\Big|_0^{2\pi}=\frac{1}{2\pi}\left(a2\pi-ea\right)=a\left(1-\frac{e}{2\pi}\right) \tag{12.1.403}$$

Notice in the special case of a circle where $a=R$ and $e=0$ we have as expected $\bar{d}=R$.

12.1.6.5 Classical deflection of light (light bending)

The calculations done previously can be applied to an interesting case: the deflection of light by a star in the Newtonian interpretation (of course!).

Warning!!! Newton did not know at its time that the photon was mass-less. The following developments are therefore a wrong approach in our time and should be taken with precaution but are still taught today because it allows students that do not yet studied General Relativity or that will never study it (in the section on General Relativity, the reader will find the contemporary detailed proof of the deflection of light that is a whole other level) to have a first approach... it's like everything in physics! Until we have reached the level of the university degree, we learn many things "wrong" because oversimplified. Then at the Master or PhD level, we learn a little more realistic and valid theories.

Well this being recalled (following the remark of one of our reader), so we have proved above for recall that:

$$r=\frac{p}{1+e\cos(\theta)} \qquad (12.1.404)$$

In the case of a photon, we tend to put that $r\rightarrow+\infty$ (thus a hyperbolic trajectory) and therefore this requires that in the previous relation we have (which is equivalent to saying that e is strictly greater than the unit as required by the hyperbolic trajectory):

$$\cos(\theta)=-\frac{1}{e} \qquad (12.1.405)$$

By putting $\varphi=2\theta-\pi$ the elementary trigonometric relations (see section Trigonometry page 2519) give us:

$$\sin\left(\frac{\varphi}{2}\right)=\sin\left(\theta-\frac{\pi}{2}\right)=-\cos(\theta)=\frac{1}{e} \qquad (12.1.406)$$

and therefore still using trigonometric identities:

$$e^2=\frac{1}{\sin^2\left(\dfrac{\theta}{2}\right)}=1+\cot\left(\frac{\theta}{2}\right) \qquad (12.1.407)$$

Therefore:

$$\cot\left(\frac{\theta}{2}\right)=e^2-1 \qquad (12.1.408)$$

And we know that:

$$e^2=1+\frac{2EK^2}{\mu^2} \qquad (12.1.409)$$

Hence:

$$e^2-1=\frac{2EK^2}{\mu^2} \qquad (12.1.410)$$

neglecting the potential energy of the photon since $r \rightarrow +\infty$ (caution!!! Let us recall that according to what we saw in the section of Special Relativity, the photon has no mass strictly speaking but Newton knew nothing about this at his time!):

$$E=\frac{1}{2}v^2 \tag{12.1.411}$$

Therefore:

$$\cot^2\left(\frac{\theta}{2}\right)=\frac{v^2K^2}{G^2M^2} \tag{12.1.412}$$

Hence:

$$\tan^2\left(\frac{\theta}{2}\right)=\frac{G^2M^2}{v^2K^2} \tag{12.1.413}$$

After simplification:

$$\tan^2\left(\frac{\theta}{2}\right)=\frac{GM}{vK} \tag{12.1.414}$$

and as θ is assumed to be small, we have using the Taylor expansion (see section Sequences and Series page 1703) of the tangent function:

$$\tan^2\left(\frac{\theta}{2}\right)\cong\frac{\theta}{2} \tag{12.1.415}$$

So it finally comes:

$$\theta\cong\frac{2GM}{vK} \tag{12.1.416}$$

But, we have by definition:

$$K=r^2\frac{\mathrm{d}\theta}{\mathrm{d}t}=r^2\dot{\theta} \tag{12.1.417}$$

and we know that $v=\omega r=\dot{\theta}r$ (see section Classical Mechanics page 3280). Thus we have:

$$\theta\cong\frac{2GM}{v^2r} \tag{12.1.418}$$

If the particle is a photon passing flush to the surface of the Sun (actually a G2V-type star, aka yellow dwarf) then we have for the "classical deflection of light":

$$\boxed{\theta\cong\frac{2GM}{c^2R_{\odot}}} \tag{12.1.419}$$

A numerical application gives:

$$\theta\cong 4.237\cdot 10^{-6}\,[\text{rad}]\underset{\times\frac{360}{2\pi}}{=}2.427\cdot 10^{4\circ}\underset{\times 60\cdot 60}{=}0.87'' \tag{12.1.420}$$

Newtonian theory thus provides a 0.87 arc seconds deviation for a ray of light passing flush to the Sun's surface. Which is twice less than what can be observed experimentally and that gives the theory of General Relativity (see section General Relativity page 5800)!

Here is an Adobe Flash animation of what a light bulb light rays propagation in vacuum, ON, would look like:

The animation above will run for people having a PDF reader with Adobe Flash player installed and activated (otherwise see here: https://vimeo.com/575751871).

And here is an another Adobe Flash animation of what a light bulb light rays propagation in vacuum putted in the presence of a punctual mass, ON, would look like:

The animation above will run for people having a PDF reader with Adobe Flash player installed and activated (otherwise see here: https://vimeo.com/575750678).

12.1.6.6 Classical precession of perihelion (apsidal precession)

Before studying the precession of the orbits, we would recall that the gravitational field is a conservative and center field. This implies that the angular momentum (see section Classical Mechanics page 3244) is constant and that the path is held in a plane whose normal vector to the surface always maintains the same direction (the angular momentum vector is constant in norm and direction for recall!).

We will address here the analysis of the precession of the perihelion taking into account the results of the theory of Special Relativity (allowing it to be more accurate in the results and be able to apply these results to the orbiting electrons around the nucleus of the atom in the corpuscular model).

First let us recall that:

- The "perihelion" is the point of the orbit of a celestial body (planet, comet, etc.) that is closest to the star around which it rotates.
- The "aphelion" is the point in the orbit of an object (planet, comet, etc.) where it is farthest from the star around which it rotates.
- The "equinox[14]" is the moment (time) when the central star crosses the plane of the equator of the object that is in orbit around it.

Remark

When the Sun passes (actually a G2V-type star, aka yellow dwarf) from the southern hemisphere to the northern hemisphere of the Earth (in other words when the Sun is at the zenith at midday at the equator), it is the spring equinox (20 or 21 March) in the opposite direction, this is the autumn equinox (22 or 23 September). At these dates, there is equality of day and night all over the Earth from the point of view of the Equator.

---

[14] The name "equinox" also comes from the Latin: aequinoctium from aequus (equal) and nox from noctis (night), which means “night equal (to day)”.

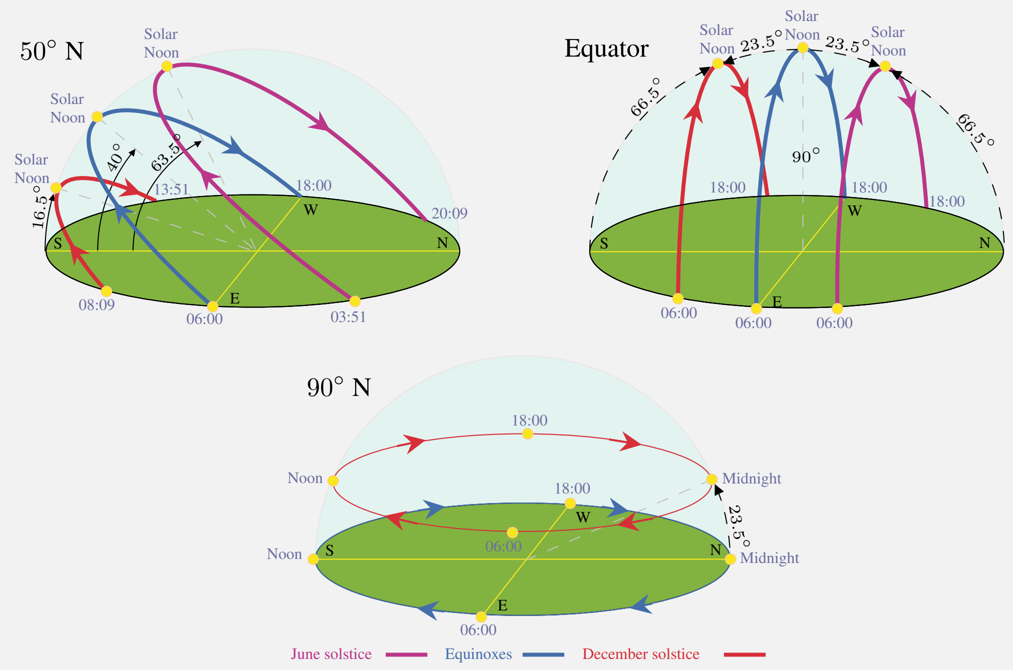

50° N
Solar Noon
Solar Noon
Solar Noon
40°
63.5°
16.5°
13:51
18:00
20:09
W
S
N
E
08:09
06:00
03:51
Equator
Solar Noon
Solar Noon
Solar Noon
23.5°
23.5°
66.5°
66.5°
90°
18:00
18:00
18:00
S
W
N
E
06:00
06:00
06:00
90° N
18:00
Noon
Midnight
18:00
W
23.5°
Noon
S
06:00
N
Midnight
E
06:00
June solstice
Equinoxes
December solstice

Figure 12.58: **Solstices and Equinoxes depending on the latitude**

Contrary to an urban legend, the Sun can be at its zenith at noon (vertical) only between the tropics and even then, twice a year only! The zenith, figuratively speaking, is the highest degree of a situation. We can say, strictly speaking, that the Sun is "at its Zenith" at noon to mean that it is at the highest point of its apparent trajectory, but it is better to avoid it, as this is a source of error.

Obviously, the result we will get here is not complete, since, as we know, we had to wait the development of General Relativity to give the exact value of the perihelion precession of Mercury (that's why will come back on this subject further below).

To calculate the effect of precession, we will seek the equivalent of the Binet formulas seen above in relativistic form (we will see the classical form in the section of General Relativity). For this we proceed as follows:

The relativistic Lagrangian of the system is (see section Special Relativity page 5679):

$$L=T+V=\mu c^2(\gamma-1)+V \qquad (12.1.421)$$

Remark

We subtract then energy at rest because only interest us here the study of the kinetic and potential energy. The potential energy is summed in the Lagrangian above (which is not consistent with the practice) but we will reverse the sign later below during the developments.

With (see section Special Relativity page 5739 and Vector Calculus page 1863):

$$\gamma=\left(1-\frac{v^2}{c^2}\right)^{-1/2}=\left(1-\frac{\dot{r}+r^2\dot{\theta}^2}{c^2}\right)^{-1/2} \qquad (12.1.422)$$

and the reduce mass for recall:

$$\mu=\frac{mM}{m+M} \tag{12.1.423}$$

The angular moment:

$$\vec{b}=\mu\vec{r}\times\vec{v} \tag{12.1.424}$$

in relativistic form and applied to our study is:

$$\vec{b}=(\mu\gamma)r\times(\vec{\omega}\times\vec{r})=(\mu\gamma)\vec{r}\times(\vec{\omega}\times\vec{r}) \tag{12.1.425}$$

Taking the norm, we have without forgetting that in our study $\vec{\omega}\perp\vec{r}$ and therefore $(\vec{\omega}\perp\vec{r})\perp\vec{r}$:

$$b=(\mu\gamma)\omega r^2 \tag{12.1.426}$$

and let us recall that we have adopted the notation $\omega=\dot{\theta}$ (in case you forget the definition...). Which finally gives us:

$$b=(\mu\gamma)r^2\dot{\theta} \tag{12.1.427}$$

To establish the relativistic equivalent of the Binet formulas:

- We deduce the expression of the angular momentum:

$$\dot{\theta}=\mathrm{d}_t\theta=\frac{b}{\mu\gamma r^2} \tag{12.1.428}$$

- We seek for a relation of the type $\dot{r}=\dot{r}(\theta)$ (as the trajectory is a conic):

$$\dot{r}=\mathrm{d}_t r=\frac{b}{\mu\gamma r^2}\mathrm{d}_\theta r \tag{12.1.429}$$

  Indeed let us recall that in polar coordinates the speed is given by the following expression (see section Vector Calculus page 1863):

$$\vec{v}=\dot{\vec{r}}=\dot{r}\vec{e}_r+r\dot{\theta}\vec{e}_\theta \tag{12.1.430}$$

  That is to say, $\dot{r}=f(r,\theta)$. The latter expression gives us the possibility to write that:

$$\dot{r}=\frac{\mathrm{d}r}{\mathrm{d}t}=\frac{\mathrm{d}r}{\mathrm{d}\theta}\frac{\mathrm{d}\theta}{\mathrm{d}t}=\mathrm{d}_\theta r\dot{\theta} \tag{12.1.431}$$

- We seek a relation of the type $\ddot{r}=\ddot{r}(\theta)$:

$$\ddot{r}=\frac{b}{\mu\gamma r^2}\left[\mathrm{d}_\theta\left(\frac{b}{\mu\gamma r^2}\mathrm{d}_\theta r\right)\right] \tag{12.1.432}$$

From the equations obtained previously, we have successively:

$$\mu c^2(\gamma-1)+V=L\Leftrightarrow\gamma^2=\left(1+\frac{L-V}{\mu c^2}\right)^2 \tag{12.1.433}$$

Let us recall that we have defined in Special Relativity β and that by using the speed in polar coordinates:

$$\beta=\frac{v^2}{c^2}=\frac{\dot{r}^2+r^2\dot{\theta}^2}{c^2} \tag{12.1.434}$$

With the previous relations, this gives us:

$$\begin{aligned}\beta=\frac{v^2}{c^2}=\frac{\dot{r}^2+r^2\dot{\theta}}{c^2}=\frac{(\mathrm{d}_\theta r\dot{\theta})^2+r^2\left(\dfrac{b}{\mu\gamma r^2}\right)^2}{c^2}=\frac{\left(\mathrm{d}_\theta r\dfrac{b}{\mu\gamma r^2}\right)^2 r^2\left(\dfrac{b}{\mu\gamma r^2}\right)^2}{c^2}\\ =\frac{b^2}{\mu^2\gamma^2c^2}\left[\left(\frac{1}{r^2}\mathrm{d}_\theta r\right)^2+r^2\left(\frac{1}{r^2}\right)^2\right]=\frac{b^2}{\mu^2\gamma^2c^2}\left[\left(\frac{1}{r^2}\mathrm{d}_\theta r\right)^2+\frac{1}{r^2}\right]\end{aligned} \tag{12.1.435}$$

On the other hand:

$$\gamma^2=\frac{1}{1-\beta}=1+\frac{\beta}{1-\beta}=1+\gamma^2\beta \tag{12.1.436}$$

By substituting the prior previous relation in the latter:

$$\gamma^2=1+\cancel{\gamma^2}\frac{b^2}{\mu^2\cancel{\gamma^2}c^2}\left[\left(\frac{1}{r^2}\mathrm{d}_\theta r\right)^2+\frac{1}{r^2}\right] \tag{12.1.437}$$

By putting $u=1/r$ and as:

$$\frac{1}{r^2}\mathrm{d}_\theta r=\mathrm{d}_r\left(-\frac{1}{r}\right)\mathrm{d}_\theta r=\frac{\mathrm{d}\left(-\dfrac{1}{r}\right)}{\mathrm{d}r}\frac{\mathrm{d}r}{\mathrm{d}\theta}=-\mathrm{d}_\theta\left(\frac{1}{r}\right)=-\mathrm{d}_\theta u \tag{12.1.438}$$

The prior-previous relations becomes with this expression:

$$\gamma^2=1+\frac{b^2}{\mu^2c^2}\left[(-\mathrm{d}_\theta u)^2+u^2\right] \tag{12.1.439}$$

Equating this relation with that of the Lagrangian:

$$1+\frac{b^2}{\mu^2c^2}\left[\mathrm{d}_\theta^2 u+u^2\right]=\left(1+\frac{L-V}{\mu c^2}\right)^2 \tag{12.1.440}$$

Differentiating the latter relation relatively to θ:

$$\frac{b^2}{\mu^2c^2}\left[2\mathrm{d}_\theta(\mathrm{d}_\theta^2 u)+2u\mathrm{d}_\theta u\right]=2\left(1+\frac{L-V}{\mu c^2}\right)\left(-\frac{1}{\mu c^2}\mathrm{d}_\theta V\right) \tag{12.1.441}$$

Indeed, the Lagrangian is constant over time (the system is assumed to be conservative), we then have:

$$\mathrm{d}_\theta L=0 \tag{12.1.442}$$

and also:

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\frac{\mathrm{d}u(\theta)}{\mathrm{d}\theta}\frac{\mathrm{d}u(\theta)}{\mathrm{d}\theta}\right)=\frac{\mathrm{d}^2u(\theta)}{\mathrm{d}\theta^2}\frac{\mathrm{d}u(\theta)}{\mathrm{d}\theta}+\frac{\mathrm{d}u(\theta)}{\mathrm{d}\theta}\frac{\mathrm{d}^2u(\theta)}{\mathrm{d}\theta^2}=2\frac{\mathrm{d}}{\mathrm{d}\theta}u(\theta)\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}u(\theta) \tag{12.1.443}$$

But if we continue:

$$-\frac{\mathrm{d}V}{\mathrm{d}\theta}=-\frac{\mathrm{d}V}{\mathrm{d}r}\frac{\mathrm{d}r}{\mathrm{d}\theta} \tag{12.1.444}$$

By referring to:

$$\frac{1}{r^2}\mathrm{d}_\theta r=-\mathrm{d}_\theta u\rightarrow \mathrm{d}_\theta r=-r^2\mathrm{d}_\theta u=-\frac{1}{u^2}\mathrm{d}_\theta u \tag{12.1.445}$$

So we get:

$$-\frac{\mathrm{d}V}{\mathrm{d}\theta}=\left(-\frac{\mathrm{d}V}{\mathrm{d}r}\right)\left(-\frac{1}{u^2}\mathrm{d}_\theta u\right)=\frac{1}{u^2}\frac{\mathrm{d}V}{\mathrm{d}r}\mathrm{d}_\theta u \tag{12.1.446}$$

That gives after a few simplifications:

$$\frac{b^2}{\mu^2c^2}\left[\mathrm{d}_\theta^2+u\right]=\left(1+\frac{L-V}{\mu c^2}\right)\frac{1}{\mu c^2}\frac{1}{u^2}\frac{\mathrm{d}V}{\mathrm{d}r} \tag{12.1.447}$$

By multiplying the latter by μ^2c^2/b^2:

$$\begin{aligned}\mathrm{d}_\theta^2u+u&=\frac{\mu^2c^2}{b^2}\left(1+\frac{L-V}{\mu c^2}\right)\frac{1}{\mu c^2}\frac{1}{u^2}\frac{\mathrm{d}V}{\mathrm{d}r}\\ \mathrm{d}_\theta^2u+u&=\frac{\mu}{b^2u^2}\left(1+\frac{L-V}{\mu^2}\right)\frac{\mathrm{d}V}{\mathrm{d}r}\end{aligned} \tag{12.1.448}$$

In a gravitational potential:

$$V=-G\frac{mM}{r}\rightarrow\frac{\mathrm{d}V}{\mathrm{d}r}=-\left(-G\frac{mM}{r^2}\right)=G\frac{mM}{r^2}=GmMu^2 \tag{12.1.449}$$

The Binet equation in Special Relativity is then:

$$\boxed{\mathrm{d}_\theta^2u+u=\frac{GmM\mu}{b^2}\left(1+\frac{L+GmMu}{\mu c^2}\right)} \tag{12.1.450}$$

To find a solution to this differential equation, we will group the variable u in the left side:

$$\begin{aligned}\mathrm{d}_\theta^2u+u&=\frac{GmM\mu}{b^2}+\frac{GmML}{b^2c^2}+\frac{G^2m^2M^2}{b^2c^2}u\\ \mathrm{d}_\theta^2u+u\left(1-\frac{G^2m^2M^2}{b^2c^2}\right)&=\frac{GmM\mu}{b^2}\left(1+\frac{L}{\mu c^2}\right)\end{aligned} \tag{12.1.451}$$

We put:

$$\Omega^2=\left(1-\frac{G^2m^2M^2}{b^2c^2}\right)\qquad\text{and}\qquad B=\frac{GmM\mu}{b^2}\left(1+\frac{L}{\mu c^2}\right) \tag{12.1.452}$$

The differential equation then can be written:

$$\mathrm{d}_\theta^2u+\left(\Omega^2u-B\right)=0 \tag{12.1.453}$$

We put:

$$\left(\Omega^2 u - B\right) = W \tag{12.1.454}$$

By taking the second derivative:

$$\Omega^2 \mathrm{d}_\theta^2 u = \mathrm{d}_\theta^2 W \Leftarrow \mathrm{d}_\theta^2 u = \frac{1}{\Omega^2}\mathrm{d}_\theta^2 W \tag{12.1.455}$$

We then get a simple differential equation:

$$\mathrm{d}_\theta^2 W + \Omega^2 W = 0 \tag{12.1.456}$$

whose solution is well known to us (see section Differential and Integral Calculus page 1610):

$$\begin{gathered} W = C_1 \cos(\Omega\theta) + C_2 \sin(\Omega\theta) = \Omega^2 u - B \\ \Rightarrow u = \frac{C_1}{\Omega^2}\cos(\Omega\theta) + \frac{C_2}{\Omega^2}\sin(\Omega\theta) + \frac{B}{\Omega^2} \end{gathered} \tag{12.1.457}$$

What can still be written as Ω^2 is a constant:

$$u = k_1 \cos(\Omega\theta) + k_2 \sin(\Omega\theta) + \frac{B}{\Omega^2} \tag{12.1.458}$$

with $k_1, k_2 = c^{te}$.

To determine the constants k_1, k_2, we place ourselves first in the situation for which $\theta = 0$, where r is minimal and therefore by u is maximum by definition.

We derivate relatively to θ:

$$(\mathrm{d}_\theta u)_{\theta=0} = \left[\underbrace{-k_1\Omega\sin(\Omega\theta)}_{=0} + \underbrace{-k_2\Omega\cos(\Omega\theta)}_{k_2=0}\right]_{\theta=0} = 0 \tag{12.1.459}$$

Therefore $k_2 = 0$ which makes that the relation:

$$u = k_1 \cos(\Omega\theta) + k_2 \sin(\Omega\theta) + \frac{B}{\Omega^2} \tag{12.1.460}$$

becomes:

$$\frac{1}{r} = k_1 \cos(\Omega\theta) + \frac{B}{\Omega^2} = \frac{B}{\Omega^2}\left(1 + \frac{\Omega^2 k_1}{B}\cos(\Omega\theta)\right) \tag{12.1.461}$$

Written differently (trying to return to a similar notation to that of the study of conic) then:

$$r = \frac{1}{\dfrac{B}{\Omega^2}\left(1 + \dfrac{k_1\Omega^2}{B}\cos(\Omega\theta)\right)} = \frac{\dfrac{\Omega^2}{B}}{1 + \dfrac{k_1\Omega^2}{B}\cos(\Omega\theta)} = \frac{p}{1 + e\cos(\Omega\theta)} \tag{12.1.462}$$

And the interest to write this in this way is to notice that we fall ultimately on the equation of an ellipse with p being the focal parameter of the conic, focal parameter given for recall by (see section Analytical Geometry page 2779):

$$p = a(1 - e^2) \tag{12.1.463}$$

where a is the half major axes of the ellipse.

Now let us put:

$$D=\frac{B}{\Omega^2} \qquad \text{and} \qquad \varepsilon=\frac{\Omega^2 k_1}{B} \tag{12.1.464}$$

At the first passage through the perihelion $\theta=0$ where:

$$\cos(\Omega\theta_0)=\cos(0)=1 \tag{12.1.465}$$

we have therefore:

$$\frac{1}{r}=D(1+\varepsilon) \tag{12.1.466}$$

At the second passage through the perihelion $\theta=2\pi$, we have:

$$\cos(\Omega\theta_1)=\cos(2\pi)=1 \tag{12.1.467}$$

we also have:

$$\frac{1}{r}=D(1+\varepsilon) \tag{12.1.468}$$

The trajectory is still an ellipse but the angle $\Omega\theta_0$ that was zero initially has become $\Omega\theta_1=2\pi$.

That is, if we have:

$$\theta_1=\theta_0+\Delta\theta \tag{12.1.469}$$

Therefore:

$$\Omega(\theta_0+\Delta\theta)-\Omega\theta=2\pi \tag{12.1.470}$$

Which gives us:

$$\Delta\theta=\frac{2\pi}{\Omega}=\frac{2\pi}{\sqrt{1-\dfrac{G^2m^2M^2}{b^2c^2}}} \tag{12.1.471}$$

Since $G^2 \ll c^2$, a development in Taylor series give us (see section Sequences and Series page 1703):

$$(1+x)^{-1/2}=1-\frac{1}{2}x+\frac{3}{8}x^2-\ldots \tag{12.1.472}$$

By limiting at the order 2:

$$\Delta\theta=2\pi\left(1+\frac{G^2m^2M^2}{2b^2c^2}\right)=2\pi+\frac{\pi G^2m^2M^2}{b^2c^2} \tag{12.1.473}$$

So in conclusion, there is an advancement of the perihelion taking place in the satellite's direction of rotation. For a repository located in the satellite's rotation plane, the trajectory is always an ellipse.

This advance is of:

$$\Delta\alpha = \frac{\pi G^2 m^2 M^2}{b^2 c^2} \tag{12.1.474}$$

by period. Either by expliciting the momentum given for reminder by:

$$b = rmv \tag{12.1.475}$$

It comes after simplification:

$$\boxed{\Delta\alpha = \frac{\pi G^2 M^2}{c^2 v^2 r^2}} \tag{12.1.476}$$

We will now allow us a rough approximation (mixture of relativistic and non-relativistic). If we consider the last relation, we have obtained during our developments of the Keplerian orbital trajectories the relation:

$$p = \frac{k^2}{\mu} = \frac{\|\vec{r} \times \vec{v}\|^2}{\mu} = \frac{b^2}{\mu m^2} = \frac{b^2}{GMm^2} \tag{12.1.477}$$

Therefore, injecting this into the relation of $\Delta\alpha$ we have:

$$\Delta\alpha = \frac{\pi G^2 m^2 M^2}{b^2 c^2} = \frac{\pi G^2 m^2 M^2}{pGMm^2 c^2} = \frac{\pi GM}{pc^2} \tag{12.1.478}$$

And we have also proved in the section of Analytical Geometry that:

$$p = a(1-e^2) \tag{12.1.479}$$

Therefore:

$$\boxed{\Delta\alpha = \frac{\pi GM}{a(1-e^2)c^2}} \tag{12.1.480}$$

Unfortunately, the numerical values for Mercury precession with $G = 6.674 \cdot 10^{-11}$ [$\text{m}^3 \cdot \text{kg}^{-1} \cdot \text{s}^{-2}$], $M = 1.99 \cdot 10^{30}$ [kg], $e = 0.260$, $a = 5.787 \cdot 10^{10}$ [m] and $T = 88$ [days], gives for a century (100 years of 365 days):

$$\Delta\alpha \frac{100 \cdot 365}{88} = \frac{\pi GM}{a(1-e^2)c^2} \frac{100 \cdot 365}{88} \cong 8.3778 \cdot 10^{-8} \cdot 414.772 \cong 3.47 \cdot 10^{-5}\ [\text{rad} \cdot \text{century}^{-1}] \tag{12.1.481}$$

Thus a precession of an angle of $7''$ century, and not the $43''$ as expected (...) there is therefore a lack of a factor 6 that only the General Relativity (see section General Relativity page 5800) makes possible to find. It is nevertheless interesting that Special Relativity already gives an orbit that precesses where Newton sees stable ellipse and that this approximation works for all the planets except Mercury (the planet closest to the Sun and undergoing the brunt of curvature of space-time).

A more detailed treatment taking into account that the Sun (actually a G2V-type star, aka yellow dwarf) is not a perfect sphere (even if its the nearest perfect spherical object in the Solar System) - there is an average variation of approximately only 6 kilometres between the equatorial radius and the polar radius of the Sun - also lead to a result that doesn't explain Mercury perihelion precession!

Remark

By applying exactly the same reasoning to corpuscular quantum physics (electrical potential) but with the ad hoc constants seen in the section of Electrostatics, we find:

$$\boxed{\Delta\alpha=\frac{\pi k^2 Z^2 e^4}{b^2 c^2}} \tag{12.1.482}$$

with $\vec{b}=\mu\vec{r}\times\vec{v}$ being the momentum and in the case of the atom, we will take (see section Corpuscular Quantum Physics page 4651):

$$b=\mu\gamma r^2\dot{\theta}=n_\theta\hbar \tag{12.1.483}$$

with reduced mass equal to:

$$\mu=\frac{mM}{m+M} \tag{12.1.484}$$

If the positions of the perihelion (and therefore the aphelion) of the Earth-Moon center of gravity (whose barycenter is located on average $4,671$ [km] from Earth's center) were constant over time, the duration of the different seasons would be constant. But the orbit of the center of gravity of Earth-Moon system also rotates in its plane in the forward direction at about $12''$ per year (a revolution is about $108,000$ years).

The precession of the equinox occurs in the opposite direction (retrograde direction) at about $50''$ per year (then a "precession equinox" revolution is about $26,000$ years). The combination of these two movements permits to calculate the period of the passage of the perihelion of the Earth by the direction of the vernal equinox, this period of about $21,000$ years and is named the "climatic precession".

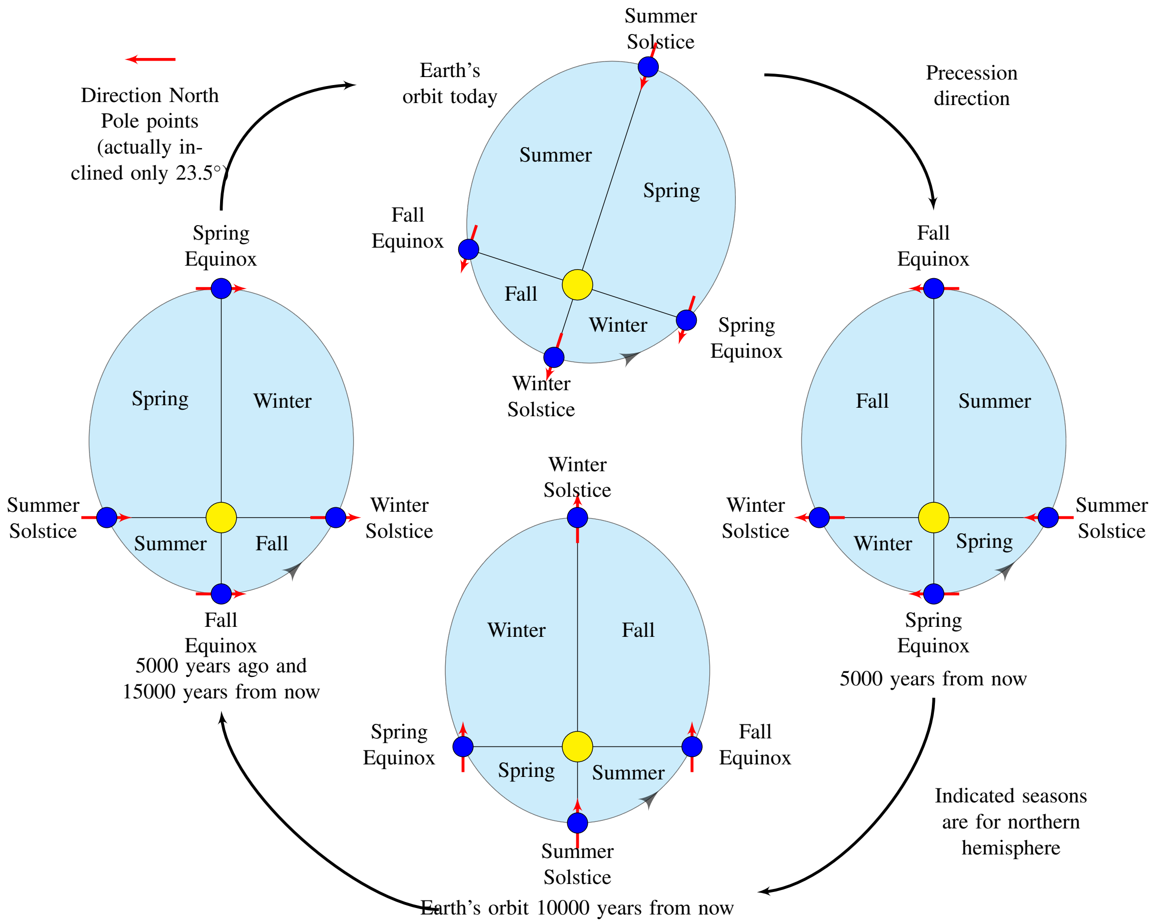

Direction North
Pole points
(actually inclined only 23.5°)
Earth's
orbit today
Summer
Solstice
Precession
direction
Summer
Spring
Fall
Equinox
Fall
Winter
Spring
Equinox
Winter
Solstice
Spring
Equinox
Spring
Winter
Summer
Solstice
Summer
Fall
Winter
Solstice
Fall
Equinox
5000 years ago and
15000 years from now
Fall
Equinox
Fall
Summer
Winter
Solstice
Winter
Spring
Summer
Solstice
Spring
Equinox
5000 years from now
Winter
Solstice
Winter
Fall
Spring
Equinox
Spring
Summer
Fall
Equinox
Summer
Solstice
Earth's orbit 10000 years from now
Indicated seasons
are for northern
hemisphere

Figure 12.59: Effects of precession on the seasons using the Northern Hemisphere terms (source: Wikipedia)

Caution! The illustration of the Earth revolving around the Sun in an elliptical orbit (a pronounced elliptical shape mainly due to the mathematical property of the cavalier perspective) is often wrongly associated with the seasons! As we will prove it later (see section Engineering page 8278), the eccentricity of the Earth's orbit plays very little role in temperature variations. The most important factor is actually the tilt of the Earth's rotation axis relative to its orbital plane around the Sun. This tilt is currently about 23.5 degrees. When one hemisphere is tilted toward the Sun, it receives more direct light and the days are longer, which results in warmer temperatures and summer in that hemisphere. Conversely, the other hemisphere, tilted away from the Sun, receives less direct light and the days are shorter, which results in colder temperatures and winter.

Indeed, every $10,500$ years (half period of climatic precession) aphelion changes from summer to winter. But even if the Earth-Sun distance is by far not the predominant factor in the nature of the seasons, the combination of the passage of the Earth in the winter in aphelion gives winters a little bit more harsh. Earth-Sun distance also depends on the variation in the eccentricity of Earth's orbit (due to external and inner planets). Thus, the ice ages are correlated with the minimum eccentricity of Earth's orbit (in the legend of the image below, the year is given based on the holocene calendar).

]

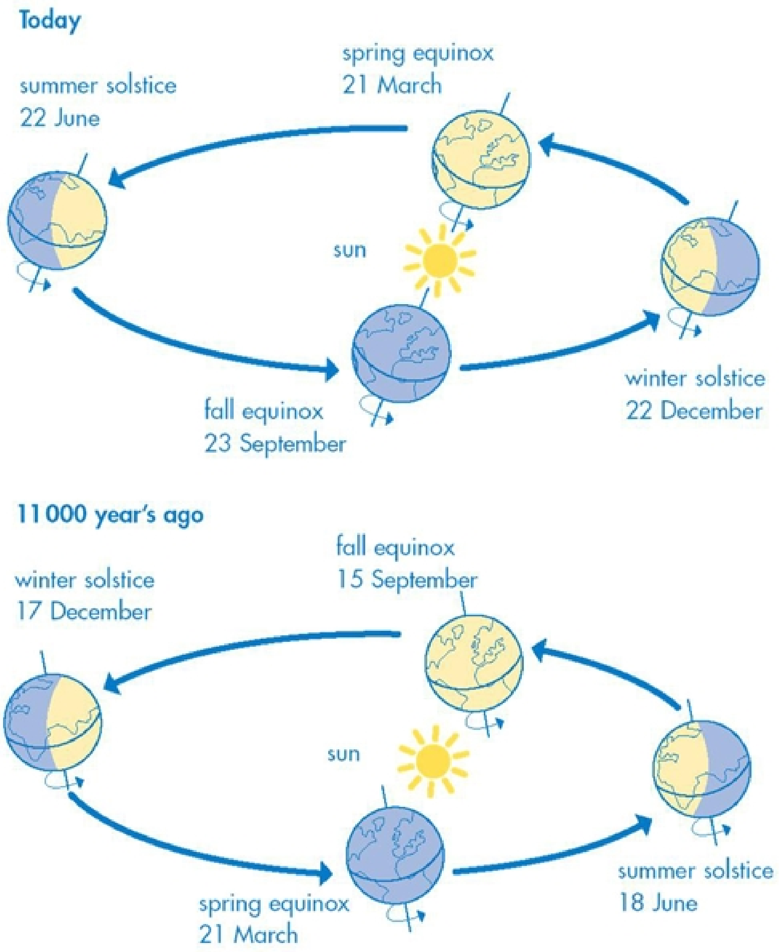

Today
spring equinox
21 March
summer solstice
22 June
sun
winter solstice
22 December
fall equinox
23 September
11 000 year's ago
fall equinox
15 September
winter solstice
17 December
sun
summer solstice
18 June
spring equinox
21 March

Figure 12.60: **Simplified and perspective point of view of the previous figure (source: Latsis foundation (12001))**

The work of the Celestial Mechanics Institute (France), since the 11970s (holocene calendar), would have to definitively confirm the theoretical predictions as what the eccentricity of Earth's orbit undergoes wide variations formed numerous periodicals under which the most important one have periods near $100,000$ years, and for one of them, a period of $400,000$ years. These results confirm the climatic variations of the Earth during the Quaternary (see section Weather & Marine Engineering page 8276). The paleoclimatology models indeed show the correlation between changes in the Earth's orbit elements and large quaternary glaciation.

Remark

In the case of the hydrogen atom (see section Corpuscular Quantum Physics page 4644), for the case dealing with relativistic model of Sommerfeld, with $n=1,n_\theta=1,Z=1$ and the fine-structure constant approximately equal to $1/137$, we get by applying the above relation by analogy but for the precession of the perihelion of the orbit of the electron:

$$\Delta\alpha=\pi\left(\frac{1}{137}\right)^2=1.67382\cdot 10^{-4}\,[\text{rad}\cdot\text{s}^{-1}] \qquad (12.1.485)$$

according to a corpuscular point view of matter!

12.1.7 Duration of the diurnal arc

A diurnal arc is the time, as expressed in right ascension (the celestial equivalent of longitude, measuring the East-West position of an astronomical object on the celestial sphere), it takes a planet, point, or degree to move from its rising point to its setting point. This takes place in many celestial bodies such as the Sun (actually a G2V-type star, aka yellow dwarf) and Moon.

So we will study here at the time length of the day, more exactly to the portion of day where we are illuminated by the Sun, as compared to the night when we are in the shade[15].

In reality, the Earth revolves around the Sun and describes an almost circular orbit at the same time it turns on itself around its axis that is tilted (actually) by about $23°27'$ relatively to its orbital plane[16] (the ecliptic):

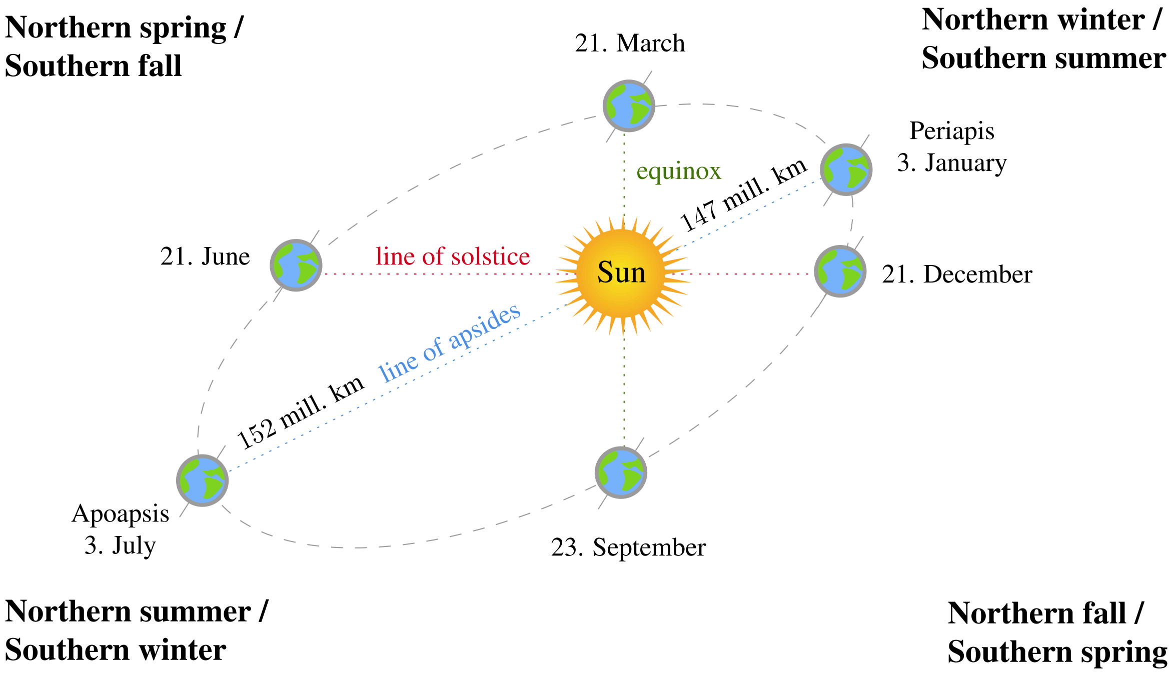

Northern spring /
Southern fall
21. March
Northern winter /
Southern summer
Periapsis
3. January
equinox
147 mill. km
21. June
line of solstice
Sun
21. December
line of apsides
152 mill. km
Apoapsis
3. July
23. September
Northern summer /
Southern winter
Northern fall /
Southern spring

Figure 12.61: **Representation of the rotation of the Earth on its orbit with its major phases**

The reader must remember that the ellipse in the illustration above is highly exaggerated for pedagogical reasons! In reality between the periapsis and the apoapsis there is only actually 3% of difference.

Remark

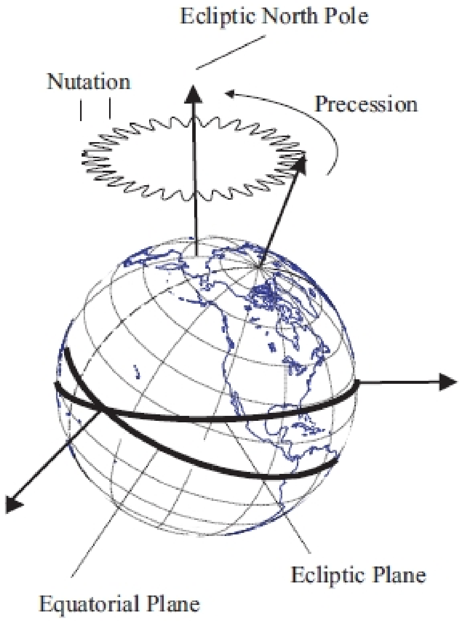
It is obvious that, given the complexity of the problem, we will simplify it by considering a circular orbit without variations (precession, nutation) of the axis of rotation of the Earth. We will assume that the Sun is reduced to a point (no dawn or twilight, etc.).

Let us first recall that the precession is the gradual change in direction of the axis of rotation of an object when a torque (force) is applied to it while the nutation is a periodic balancing of the axis of rotation of the Earth around its mean position in addition to the precession (see section Classical Mechanics page 3324).

---

[15] Thanks to Xavier Hubaut for these very friendly developments

[16] We measure Earth's axial inclination (tilt) using simple geometry and observing the Sun's angle at different times, primarily during solstices, by measuring shadows cast by a vertical pole (gnomon).



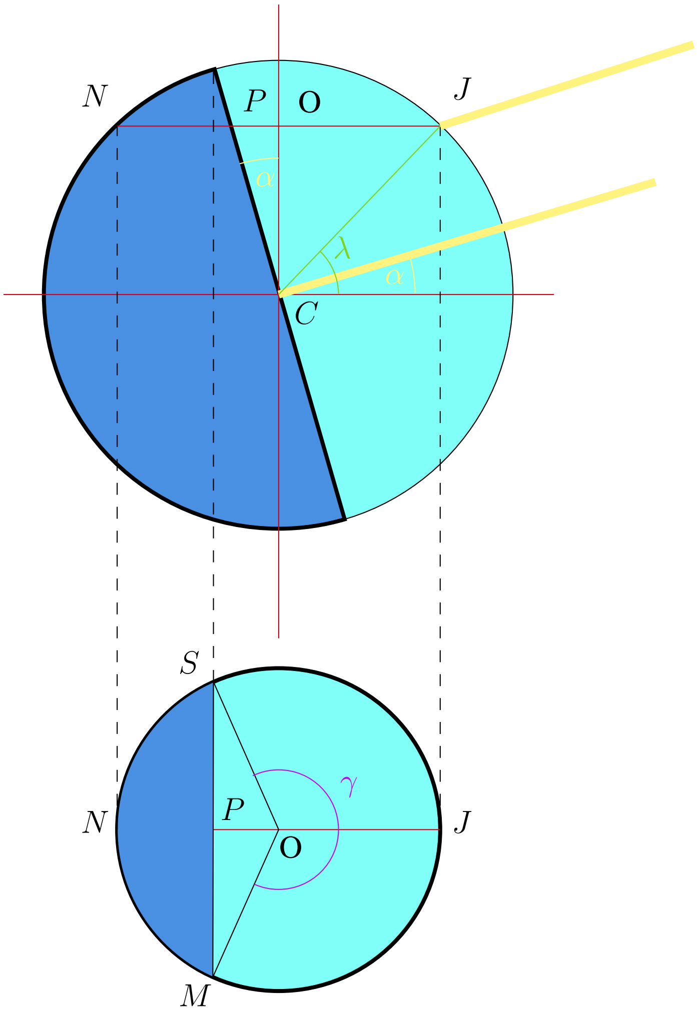
Let us represent the Earth with its vertical axis of rotation. Accordingly the equator will be located in a horizontal plane.

Suppose that day, the Earth is in such a position that the Sun's rays form an angle α with the equatorial plane (or conversely that the axis of the Earth form an angle with the equatorial plane). Notice that this angle α will always be according to actual measurements between $-23°27'$ and $+23°27'$ at least... at a human life time scale...

For our example we have chosen to focus our analysis on a day when α is positive. Thus, in the northern hemisphere, we are close to the summer solstice!

We are looking for the day length at a place located at latitude λ! To fix ideas, we place ourselves around Brussels at $50°$ north latitude.

Let us now consider the following figures where the first is a view of the side of Earth at a time t of its orbit when $\alpha > 0$ and the second in to a cylindrical cutting of diameter $\overline{NJ}$ (corresponding to the diameter of the parallel of Brussels) of Earth's volume Earth at this same moment:



In the figure above, C denotes the center of the Earth, and O the center of the parallel of Brussels.

Let us fix a time t and denote by M (morning) and S (evening) the two points of the parallel of Brussels where the Sun rises and sets (these points will be considered fixed whatever the moment t, which is obviously wrong relatively to the reality), while J (day) and N (night) will be the points where it is noon and midnight respectively.

P will be the point on the disc corresponding to the Brussels parallel where the meridian noon plane (the plan which of the sides is $\overline{NJ}$) cut the line $\overline{MS}$.

Finally, γ designate the angle $\widehat{M\text{O}S}$ (where O is the center of the disk generated by the parallel of Brussels) behind the illuminated part by the Sun and r designate the radius $S\text{O}=M\text{O}$.

To simplify the problem, let us also assume... that during 24 hours the Earth rotates on itself without changing the position of its axis of rotation relative to the Sun....

The angle γ can be calculated by noting that $\overline{\text{O}P}$ is, in absolute value equal to:

$$\overline{P\text{O}}=\overline{S\text{O}}\cos\left(\pi-\frac{\gamma}{2}\right)=\overline{M\text{O}}\cos\left(\pi-\frac{\gamma}{2}\right)=r\cos\left(\pi-\frac{\gamma}{2}\right) \tag{12.1.486}$$

where r represents for recall the radius of the parallel of Brussels.

Using the properties of trigonometric functions (see section Trigonometry page 2519), we have:

$$\overline{PO}=-r\cos\left(\frac{\gamma}{2}\right) \tag{12.1.487}$$

But we still need to inject the parameter α. Knowing the latitude λ of Brussels, we have:

$$r=\overline{OJ}=R\cos(\lambda) \tag{12.1.488}$$

where R is the radius of the Earth.

We have also:

$$\overline{CO}=R\sin(\lambda) \tag{12.1.489}$$

and in the triangle COP:

$$\overline{PO}=\overline{CO}\tan(\alpha)=R\sin(\lambda)\tan(\alpha) \tag{12.1.490}$$

Finally, by comparing the values obtained for $\overline{PO}$, we get:

$$\left.\begin{array}{l}\overline{PO}=R\sin(\lambda)\tan(\alpha)\\ \overline{PO}=-r\cos\left(\frac{\gamma}{2}\right)\end{array}\right\}\Rightarrow R\sin(\lambda)\tan(\alpha)=-r\cos\left(\frac{\gamma}{2}\right) \tag{12.1.491}$$

and as:

$$r=\overline{OJ}=R\cos(\lambda) \tag{12.1.492}$$

We finally get:

$$R\sin(\lambda)\tan(\alpha)=-R\cos(\lambda)\cos\left(\frac{\gamma}{2}\right) \tag{12.1.493}$$

and therefore:

$$\gamma=-2\arccos(\tan(\lambda)\tan(\alpha)) \tag{12.1.494}$$

At the equinoxes (that is to say when the equator coincides with the ecliptic plane for recall...) we have $\alpha=0$ and therefore:

$$\gamma=-2\arccos(\tan(\lambda)\tan(\alpha))=-2\arccos(\tan(\lambda)\cdot 0)=-2\arccos(0)=-\pi \tag{12.1.495}$$

However, as we have specified it at the beginning, we must take the absolute value thus:

$$\gamma=\pi \tag{12.1.496}$$

In other words, whatever the latitude we take, the angle formed by the night area is equal to the angle formed by the day area at equinoxes (both being equal to π).

Let us now consider the summer solstice, when $\alpha=23°27'$ still considering the latitude of Brussels $\lambda=50°$, we have:

$$\gamma=-2\arccos(\tan 50°)\tan(23°27'))=-2\arccos(0.51695)=-2\arccos(0)\cong-118° \tag{12.1.497}$$

This translated is translated into hours by:

$$-118^\circ\frac{24}{360}\cong -7.9 \tag{12.1.498}$$

So the 24-hour day loses 7.9 hours. Which is equivalent to a day light of approximately 16 hours.

In summary to calculate the duration of a "day", it is enough to know two things: the latitude and the angle at which the Sun (actually a G2V-type star, aka yellow dwarf) falls on the plane of the equator to the chosen date. The value of this angle is well known at the equinoxes (it is 0°) and to the solstices (it is $+23^\circ 27'$ and $-23^\circ 27'$).

But what about the other dates?

The answer is quite simple. Let us imagine, sitting on the Sun watching throughout the year towards the center of the Earth.

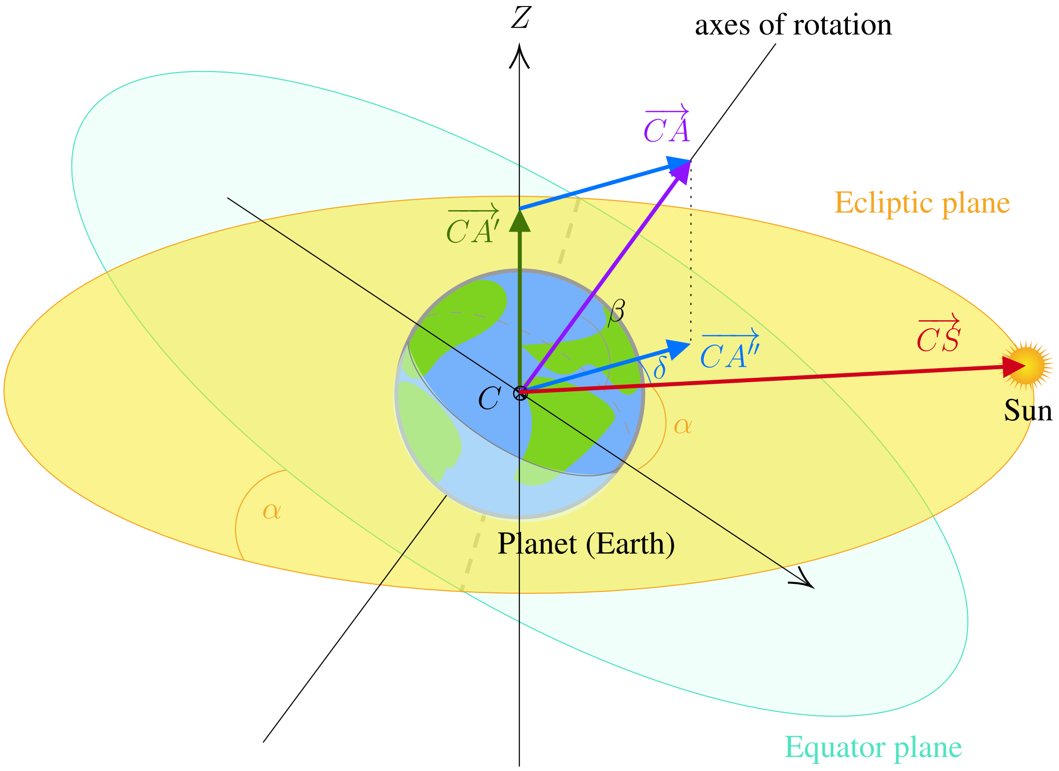
During its rotation around the Sun (the binomial centroid in fact), the axis of rotation of the Earth maintains its inclination to the ecliptic. Seen from the Sun, this axis revolve around normal to the plane of the ecliptic and therefore describe a cone whose half apex angle is $23^\circ 27'$ (see figure below).

The angle of attack α of the sunlight on the equator therefore vary according to the date δ (we associate to the date, the angle δ travelled by the Earth on its orbit, from its position at the spring equinox).

Therefore, the angle α vary according to the date δ sinusoidally.

For those who are perhaps not convinced by this semi-intuitive reasoning, here's another approach:

For readability of the diagram, we have greatly exaggerated the angle of the axis of rotation of the Earth with the ecliptic (and obviously nothing is to scale...!):



Given C the Earth's center, A the end of a unitary vector $\overrightarrow{CA}$ oriented directed along the axis of rotation of the Earth (i.e. perpendicular to the plane of the equator) and another unit vector $\overrightarrow{CS}$ directed toward the Sun. Given now α the angle of the radius $\overline{CS}$ with the plane of the equator and β the angle between the unit vectors $\overrightarrow{CA}$ and $\overrightarrow{CS}$. Then we have:

$$\overrightarrow{CS}\circ\overrightarrow{CA}=\underbrace{||\overrightarrow{CS}||||\overrightarrow{CA}||}_{=1}\cos(\beta)=\cos\left(\frac{\pi}{2}-\alpha\right)=\sin(\alpha) \qquad (12.1.499)$$

Indeed, the vector $\overrightarrow{CA}$ being perpendicular to the plane of the equator, it forms a right angle with it. Therefore since the angle β is the angle between this vector and the ecliptic, the angle α is then the complementary angle.

Therefore we have:

$$\overrightarrow{CS}\circ\overrightarrow{CA}=\sin(\alpha) \qquad (12.1.500)$$

Let us decompose now $\overrightarrow{CA}$ in the sum of $\overrightarrow{CA'}$ directed perpendicular to the ecliptic plane and of $\overrightarrow{CA''}$ located in the ecliptic plane:

$$\begin{aligned}\overrightarrow{CS}\circ\overrightarrow{CA}&=\underbrace{\overrightarrow{CS}\circ\overrightarrow{CA'}}_{\overrightarrow{CS}\perp\overrightarrow{CA'}\Rightarrow=0}+\overrightarrow{CS}\circ\overrightarrow{CA''}=\overrightarrow{CS}\circ\overrightarrow{CA''}=\underbrace{||\overrightarrow{CS}||}_{=1}||\overrightarrow{CA''}||\cos\left(\widehat{BCA''}\right)\\&=||\overrightarrow{CA''}||\cos\left(\widehat{BCA''}\right)=\underbrace{||\overrightarrow{CA}||}_{=1}\cos(\beta)\cos\left(\widehat{BCA''}\right)=\cos\left(\frac{\pi}{2}-\alpha\right)\cos\left(\widehat{BCA''}\right)\\&=\sin(\alpha)\cos\left(\widehat{BCA''}\right)\end{aligned} \qquad (12.1.501)$$

Therefore:

$$\overrightarrow{CS}\circ\overrightarrow{CA}=\sin(\alpha)\cos\left(\widehat{BCA''}\right)=\sin(23°27')\cos\left(\widehat{BCA''}\right) \tag{12.1.502}$$

But (hence the number of days in the interval $[0,365[$ must be converted into an angle in the range $[0,2\pi[$):

$$\cos\left(\widehat{BCA''}\right)=\cos\left(\frac{\pi}{2}-\delta\right)=\sin(\delta) \tag{12.1.503}$$

So finally:

$$\overrightarrow{CS}\circ\overrightarrow{CA}=\sin(23°27')\sin(\delta) \tag{12.1.504}$$

and as we have proved that:

$$\overrightarrow{CS}\circ\overrightarrow{CA}=\sin(\alpha) \tag{12.1.505}$$

We finally get:

$$\sin(\alpha)=\sin(23°27')\sin(\delta)=0.3979\sin(\delta) \tag{12.1.506}$$

Now the problem is solved and the daylight time duration will depend on two variables: the date δ and the latitude λ.

So we just have now to take again the relation:

$$\gamma=-2\arccos(\tan(\lambda)\tan(\alpha)) \tag{12.1.507}$$

and to inject in it the new result to get a first simple version of the "equation of sunrise" (EoS):

$$\boxed{\gamma=|-2\arccos\left(\tan(\lambda)\tan(\arcsin(\sin(23°27')\sin(\delta)))\right)|} \tag{12.1.508}$$

From the latter we get the duration of the day as before (if the result of γ is provided in degrees):

$$t_{\text{day}}(\delta,\lambda)=24+\frac{-2\arccos\left(\tan(\lambda)\tan(\arcsin(\sin(23°27')\sin(\delta)))\right)\cdot 24}{360} \tag{12.1.509}$$

Or for example for Brussels we get in Microsoft® Excel® 16.0.17231:

```
=24+DEGREES(-2*ACOS(TAN(RADIANS(50))*TAN(ASIN(SIN(RADIANS(23)
        *SIN(RADIANS(90)))))))*24/360=16.0518
```

With computer tools at our disposal, we can easily plot the value γ. For example, we have below the variations in the length of the day over a year at latitudes of 0 to $90°$ spread by 10 by $10°$:

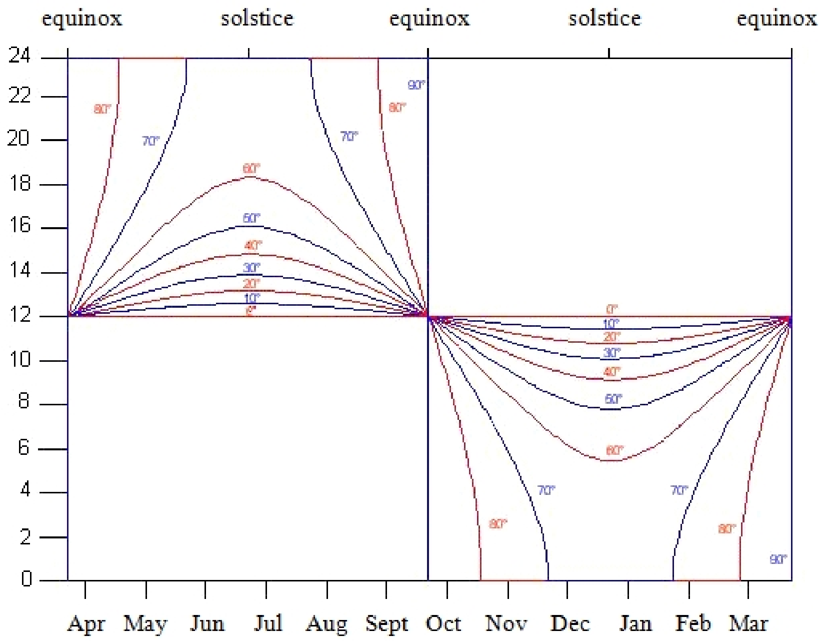


Figure 12.62: **Equation of Sunrise**

From the latitude of the Arctic Circle, we see, in summer, periods with uninterrupted Sun (midnight Sun) and in winter whole days of night.

For Brussels (latitude = $50°$) we see from the figure that the length of the day varies approximately between the values of 16 [h] (summer solstice) and 8 (winter solstice).

The length of the day at the equator is nearly always the same, staying close to 12 hours of daylight and 12 hours of darkness every day of the year.

At the equinox all places on Earth have the same number 12 hours of daylight and 12 hours of darkness !

12.1.7.1 Evolution of diurnal arc (i.e. Earth-Moon distance)

In many books or documentaries you can read or listen the following statements:

- The rotational movement of the Earth on itself slows down by the tidal effect
- The distance from Earth to Moon increases
- That these two effects are related

Let us see if we can prove this link and verify the orders of magnitude of these observed effects.

If we want to describe it completely, the movement of the Moon is a three-body problem (Earth, Moon, Sun) subjected to several hundred periodic or secular disturbances, which makes it very difficult. So the tides on Earth are caused mainly by the Moon, but the Sun also intervenes: the ratio of the gravitational forces exerted by the Moon and by the Sun on a body of water on the surface of the Earth is 2.2.

For what interests us, however, we will consider the Earth-Moon system as isolated in space. This is because we are studying the movement of rotation on itself of this system: in a way we relate the action of the Sun to the Earth-Moon center of mass. In addition, we are not studying how the tides slow down the rotation of the Earth: this slowing down being observed, we want to show how a slowing down of the Moon follows.

Second approximation: in the rotational motion, we will confuse the center of mass of the system with the center of the Earth, and the Moon will be assumed to describe an elliptical Keplerian orbit around the center of the Earth.

The physical law that must be applied is therefore simple: the angular momentum $\vec{b}$ of the Earth-Moon system is conserved. However, this total angular momentum $\vec{b}_{\text{tot}}$ is, in the Galilean coordinate system of the center of mass, the sum of three terms:

1. The angular momentum of rotation of the Earth on itself (considering it as a perfect ball), $\vec{b}_{♁}$. It is directed along the axis of the poles and therefore forms an angle of $23°27'$ with the perpendicular to the plane of the ecliptic. In intensity it is:

   $$b_{♁}=J_{♁}\omega_{♁}\cong\frac{2}{5}M_{♁}R_{♁}^2\omega_{♁}=\frac{2}{5}M_{♁}R_{♁}^2\frac{2\pi}{T_{♁}} \tag{12.1.510}$$

   where $M_{♁}\cong 6\cdot 10^{24}$ [kg], $R_{♁}\cong 6.4\cdot 10^{6}$ [m] and $T_{♁}=86,164$ [s].

2. The angular momentum of rotation of the Moon on itself (also considered as a perfect sphere) $\vec{b}_{\bigcirc}$ it is directed along the axis of rotation of the Moon, which forms an angle of $6°$ with the normal to the plane of the lunar orbit, itself inclined by $5°$ on the plane of the ecliptic. It is worth in intensity:

   $$b_{\bigcirc}=J_{\bigcirc}\omega_{\bigcirc}\cong\frac{2}{5}M_{\bigcirc}R_{\bigcirc}^2\omega_{\bigcirc}=\frac{2}{5}M_{\bigcirc}R_{\bigcirc}^2\frac{2\pi}{T_{\bigcirc}} \tag{12.1.511}$$

   where $M_{\bigcirc}\cong M_{♁}/81$ [kg], $R_{\bigcirc}\cong R_{♁}\cdot 0.272$ [m].

3. The angular momentum of rotation of the Moon around the Earth $\vec{b}_{M-E}$ is perpendicular to the plane of the lunar orbit. In intensity it is (assuming the Moon as punctual otherwise she should apply the Steiner theorem):

   $$\vec{b}=\vec{r}\times\vec{p}\Rightarrow b'=rM_{\bigcirc}v_{\bigcirc} \tag{12.1.512}$$

   where r is the Earth-Moon distance and $v_{\bigcirc}$ the speed of the moon on its orbit.

   Here comes another approximation to simplify the calculation. The elliptical orbit of the Moon has actually an eccentricity of 0.055. We will liken it to a circle for which we will still use Kepler's third law. Then if one notes ω the pulsation and D the semi-major axis one has:

   $$v=D\omega=r\omega \tag{12.1.513}$$

therefore:

$$b' = DM_{\bigcirc}v_{\bigcirc} = D^2 M_{\bigcirc}\omega_{\bigcirc} \tag{12.1.514}$$

Following third Kepler's law (see page 5242) given for recall by:

$$\frac{R^3}{T^2} = \frac{Gm}{4\pi^2} \Rightarrow R^3\omega^2 = Gm \tag{12.1.515}$$

Therefore:

$$D^3\omega^2 = GM_{\text{♁}} \Leftrightarrow \omega_{\bigcirc} = G^{1/2}M_{\text{♁}}^{1/2}D^{-3/2} \tag{12.1.516}$$

So finally:

$$b' = D^2 M_{\bigcirc}\omega_{\bigcirc} = G^{1/2}M_{\bigcirc}M_{\text{♁}}^{1/2}D^{1/2} \tag{12.1.517}$$

The total angular momentum is therefore worth, in projection on the perpendicular to the plane of the ecliptic:

$$\begin{aligned} b_{\text{tot}} &= b_{\text{♁}} + b_{\bigcirc} + b' \\ &\cong \frac{2}{5}M_{\text{♁}}R_{\text{♁}}^2\frac{2\pi}{T_{\text{♁}}}\cos(23°27') + \frac{2}{5}M_{\bigcirc}R_{\bigcirc}^2\frac{2\pi}{T_{\bigcirc}}\cos(6°) + G^{1/2}M_{\bigcirc}M_{\text{♁}}^{1/2}D^{1/2}\cos(5°) \end{aligned} \tag{12.1.518}$$

Let's identify the factors that change over time. Earth's own pulsation, $\omega_{\text{♁}}$, decreases as the length of the day increases according to all actual experimental evidence; the distance D from the Earth to the Moon increases also according to all actual experimental evidence; on the other hand, we will assume the pulsation $\omega_{\bigcirc}$ to be constant (otherwise the problem is unsolvable and anyway the Moon is negligible in comparison of Earth...).

We therefore write that the variation Δb_{tot} of the total angular momentum is zero:

$$\Delta b_{\text{tot}} \cong \frac{2}{5}M_{\text{♁}}R_{\text{♁}}^2\Delta\omega_{\text{♁}}\cos(23°27') + 0 + \frac{1}{2}G^{1/2}M_{\bigcirc}M_{\text{♁}}^{1/2}D^{-1/2}\Delta D\cos(5°) = 0 \tag{12.1.519}$$

as $\Delta\omega_{\bigcirc} = 0$.

Therefore:

$$\Delta D \cong -\frac{4}{5}G^{-1/2}M_{\bigcirc}^{-1}M_{\text{♁}}^{1/2}R_{\text{♁}}^2\Delta\omega_{\text{♁}}D^{1/2}\cos(23°27')\cos^{-1}(5°) \tag{12.1.520}$$

But $\omega_{\text{♁}} = 2\pi/T_{\text{♁}}$ then $\Delta\omega_{\text{♁}} = -2\pi\Delta T_{\text{♁}}/T_{\text{♁}}^2$. Hence:

$$\Delta D \cong \frac{8\pi}{5}G^{-1/2}M_{\bigcirc}^{-1}M_{\text{♁}}^{1/2}R_{\text{♁}}^2T_{\text{♁}}^{-2}\Delta T_{\text{♁}}D^{1/2}\cos(23°27')\cos^{-1}(5°) \tag{12.1.521}$$

So finally:

$$\Delta D \propto T_{\text{♁}}^{-2}\Delta T_{\text{♁}}D^{1/2} \Rightarrow \Delta D \propto \Delta T_{\text{♁}} \tag{12.1.522}$$

We have achieved what we wanted to demonstrate: slowing down of the Earth's own rotation and moving the Moon away are linked. If the Earth rotation slows down, its angular momentum decreases; since that of the Moon on itself is constant, for the total angular momentum to be conserved, the Moon must move away from the Earth, thus increasing its angular momentum.

Knowing, due to experimental measurements, that $\Delta T \cong 1.64 \cdot 10^{-5}\,[\text{s} \cdot \text{year}^{-1}]$ and that:

$$\begin{aligned} G &\cong 6.67 \cdot 10^{-11}\,[\text{m}^3 \cdot \text{kg}^{-1} \cdot \text{s}^{-2}] & M_{\bigcirc} &\cong 7.34 \cdot 10^{22}\,[\text{kg}] \\ M_{\text{♁}} &\cong 5.97 \cdot 10^{24}\,[\text{kg}] & R_{\text{♁}} &\cong 6371 \cdot 10^{3}\,[\text{m}] \\ D &\cong 3.84 \cdot 10^{8}\,[\text{m}] & T &= 8620\,[\text{s}] \end{aligned} \tag{12.1.523}$$

We get:

$$\Delta D \cong \frac{8\pi}{5} G^{-1/2} M_{\bigcirc}^{-1} M_{\text{♁}}^{1/2} R_{\text{♁}}^{2} T_{\text{♁}}^{-2} \Delta T_{\text{♁}} D^{1/2} \cos(23°27') \cos^{-1}(5°) \cong 3.3 \cdot 10^{-2}\,[\text{m} \cdot \text{year}^{-1}] \tag{12.1.524}$$

Hence $3.3\,[\text{cm} \cdot \text{year}^{-1}]$ (the measured value is actually of $3\,[\text{cm} \cdot \text{year}^{-1}]$).

Geologists have found sedimentary rocks (like the Elatina Formation in Australia) that preserve daily and monthly tidal layers from 600 millions to 2.5 billion years ago. These "fossilised tides" allow scientists to calculate exactly how fast the Moon was moving back then. These geolocial records show that 600 million years ago, the recession rate was only about $2.17\,[\text{cm} \cdot textyear]$ (see [**âĂć**]), nearly half of today's rate. If we do the maths we can check easily that the Moon was much closer in the past, but it remained well outside the Roche limit (see page 5337) throughout the timeline of life on Earth.

Where are we going ? 300 million years ago, the day lasted 22 hours, there were 400 days in the year. When the Earth-Moon coupling will be over, the day will last 50 hours of our current days. A simple rule of three can say that it will be roughly in 250 billion years. The Moon will then be at a distance of 1, 500 terrestrial radius, compared to the 60 today (the distance of the Earth to the Sun is worth 25, 000 terrestrial rays). For sure there will be also no Solar eclipse anymore in the future as we know them actually!

Remark

For information, the rotation of Earth's on itself slows down at the day we write these lines of approximately 2.3 milliseconds per century. That means 1 hour every 180 millions years. Each year the Earth's orbit speed around the Sun (actually a G2V-type star, aka yellow dwarf) also slows down of approximately 3 nanometers-per-second. Using Kepler's law, that means that the Earth approaches the Sun of 3 micrometers per 1 billion year......!!!!

But all this is only a bold extrapolation, because long before these 250 billion years, the Sun will have become a red giant (in 5 billion years) and its crown will have reached the Earth ...

12.1.8 Trigonometric parallax

Measuring distances to objects within our Galaxy is not always a straightforward task – we cannot simply stretch out a measuring tape between two objects and read off the distance. Instead, a number of techniques have been developed that enable us to measure distances to stars without needing to leave the Solar System. One such method is "trigonometric parallax", which depends on the apparent motion of nearby stars compared to more distant stars, using observations made 6 months apart (corresponding to the measurement of diameter of the apparent approximated circular motion they have in the sky).

A nearby object viewed from two different positions will appear to move with respect to a more distant background. This change is named a "parallax". A simple demonstration is to hold your finger up in front of your face and look at it with your left eye closed and then your right eye. The position of your finger will appear move compared to more distant objects.

By measuring the amount of the shift of the object's position (relative to a fixed background, such as the very distant stars) with observations made from the ends of a known baseline, the distance to the object can be calculated.

The trigonometric parallax method is very simple (but difficult to implement on the surface of our planet for very distant stars). Any amateur astronomer observed the flight of the star it observes with his eye. This movement is named as we have just seen the "diurnal movement" It is due to the rotation of the Earth itself. The star is also driven in an elliptical motion much less easily detectable: the "parallactic motion".

It is due, as suggested in the figure below, to the rotation of the Earth around the Sun. So we measure the angle p and we have obviously:

$$\frac{a}{d}=\tan(p) \tag{12.1.525}$$

If the angle is small (which is very often the case given the distance of stars...) we can take the first term of the Taylor expansion (see section Sequences and Series page 1703) of the tangent function:

$$\tan(p)=p+\frac{p^3}{3}+\frac{2p^5}{15}+\ldots\cong p \tag{12.1.526}$$

Which allow us to write:

$$\boxed{d\cong\frac{a}{p}} \tag{12.1.527}$$

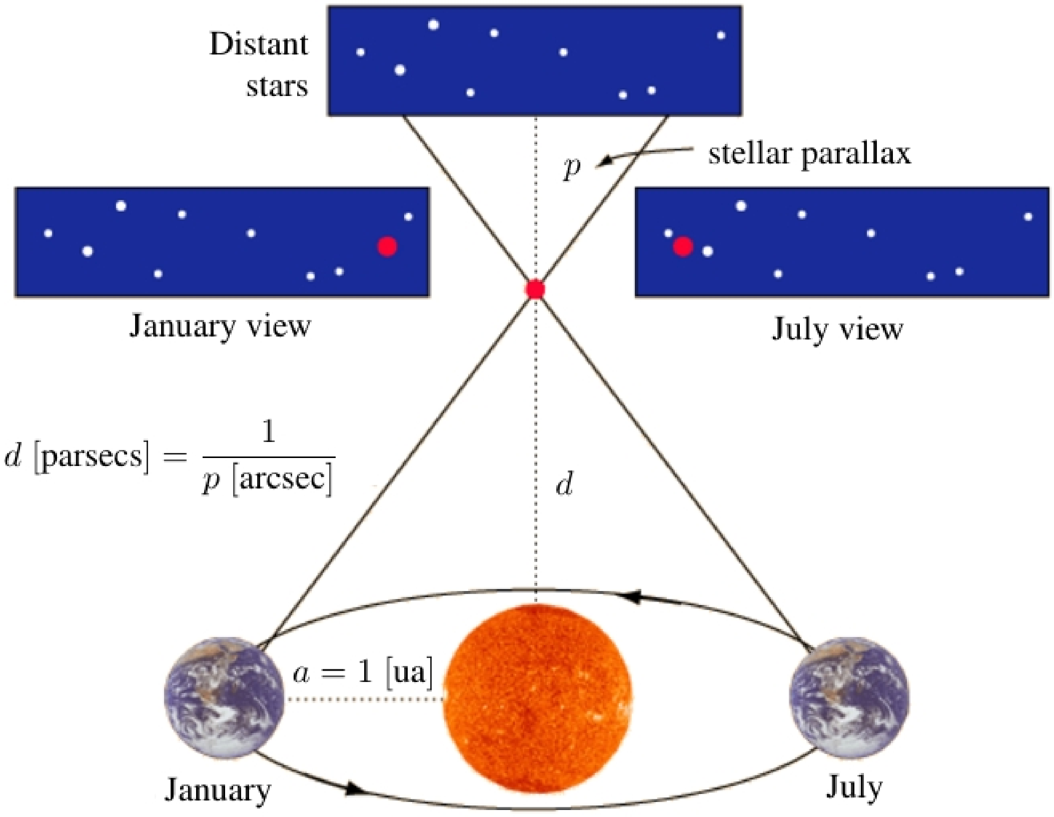


Figure 12.63: **Trigonometric parallax principle**

If the parallax angle, p is measured in arc-seconds (arcsec), then the distance to the star, d in parsecs (pc) is given by:

$$d \cong \frac{1}{p} \qquad (12.1.528)$$

It is important to notice that in this example we assume that both the Sun (actually a G2V-type star, aka yellow dwarf) and star are not moving with a transverse velocity with respect to each other. If they were this would complicate the picture as presented here. In practice stars with significant proper motions require at least three epochs of observation to accurately separate their proper motions from their parallax. Stars that are members of binaries further complicate the picture.

For example, we often see sources that say a parallax is found from measurements taken six months apart. That is rarely the case, and if you think about it, not all stars are visible in the night sky six months apart (from the same location on Earth). In fact a parallax measurement will consist of a series of measurements, possibly taken over more than 1 year. There is no formal requirement for any particular measurement interval, though of course the biggest parallax signal would be if we could get measurements six months apart.

In addition to parallax, we need to account for the "proper motion" of the star across the sky. To a first order approximation this is a linear change in right ascension (the celestial equivalent of longitude, measuring the East-West position of an astronomical object on the celestial sphere) and declination (angular distance of a celestial body North (+) or South (−) from the celestial equator) with time (though more complicated models can be used, and you have to account for the ellipticity of the Earth's orbit too). The position of the star in question (with respect to distant background stars) measured in multiple images at multiple times would then be fitted with a function that results in a determination of the parallax and the proper motions in both coordinates. Any motion of the "background stars" is compensated by averaging over all the faint stars in the field. They will all move in a systematic way due to the motion of the Sun in the Galaxy (this is part of the "proper motion"), but their random motions (if they are distant) will be very small and average out well.

A perfect visualisation is provided by Hubble Space Telescope images of Proxima Centauri which show its path across they sky and illustrates the cyclical motion due to parallax and the linear motion across the sky caused by its proper motion.

The motion of the star towards or away from the Sun could be accounted for from the measurement of the star's line of sight velocity. In practice, these motions are of order tens of km/s and this means it would take tens of thousands of years to change the distance by one light year.

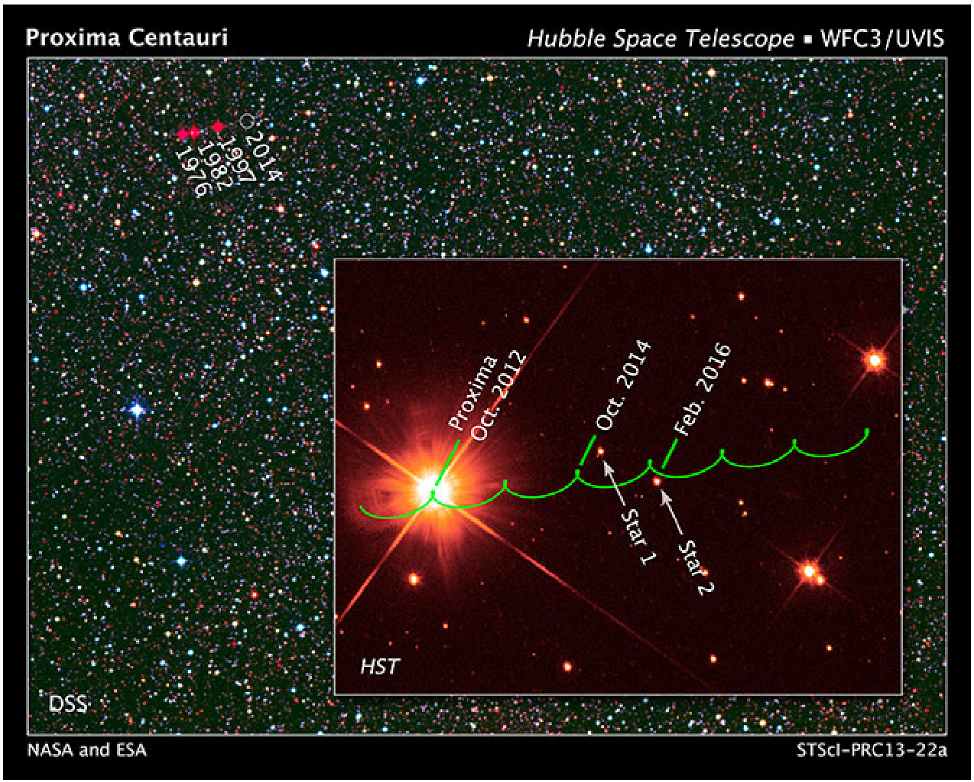


Figure 12.64: **Hubble space telescope Proxima Centauri tracking**

The only star with a parallax greater than 1 [arcsec] as seen from the Earth is the Sun - all other known stars are at distances greater than 1 [pc] and parallax angles less than 1 [arcsec] ($1/3600$ of degree... we understand better why this what impossible to measure before the 119th century). When measuring the parallax of a star, it is important to "account for the star's proper motion, and the parallax of any of the fixed" stars used as references.

Over a 4 year period from 11989 to 11993, the Hipparcos Space Astrometry Mission measured the trigonometric parallax of nearly $120,000$ stars with an accuracy of 0.002 [arcsec]. The GAIA mission, to be launched in 12010 (holocene calendar), will be able to measure parallaxes to an accuracy of 10^{-6} [arcsec], allowing distances to be determined for more than 200 million stars.

Remark

Stellar parallax is so difficult to detect that its existence was the subject of much debate in astronomy for hundreds of years. Thomas Henderson, Friedrich Georg Wilhelm von Struve, and Friedrich Bessel made first successful parallax measurements in 11832-11838 (holocene calendar), for the stars Alpha Centauri, Vega, and 61 Cygni.

In practice when we measure the parallax we must obviously take in account the obliquity of Earth on it's orbit that is in this beginning of the 121st century (holocene calendar) equal to $23°26'13.3$ otherwise me may think that stars have a huge parallax and are therefore at a small distance of us as illustrated by the figure below:

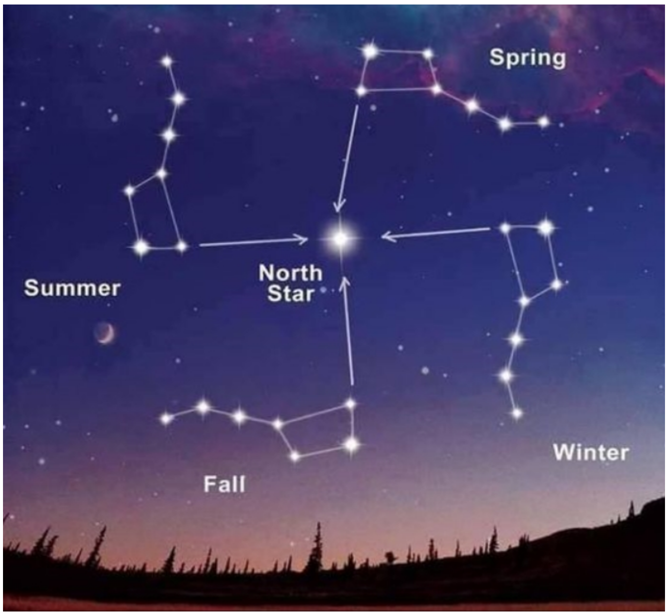


Figure 12.65: **Shift angle of the Big Dipper that seems huge if we do not subtract the obliquity angle**

In the figure above the "North star" may be any fixed star close to either celestial pole of any given planetary body (actually approximately Polaris that is 0.736 degrees away from exact North). It might refer to any such star in the Earths remote history or future, situated along the path of the celestial poles in the course of the procession of the Earth's axis.

The identity of the pole stars gradually changes over time because the celestial poles exhibit a slow continuous drift through the star field. The primary reason for this is the precession of the Earth's rotational axis, which causes its orientation to change over time. Precession causes the celestial poles to trace out circles on the celestial sphere approximately once every $26,000$ years, passing close to different stars at different times (with an additional slight shift due to the proper motion of the stars).

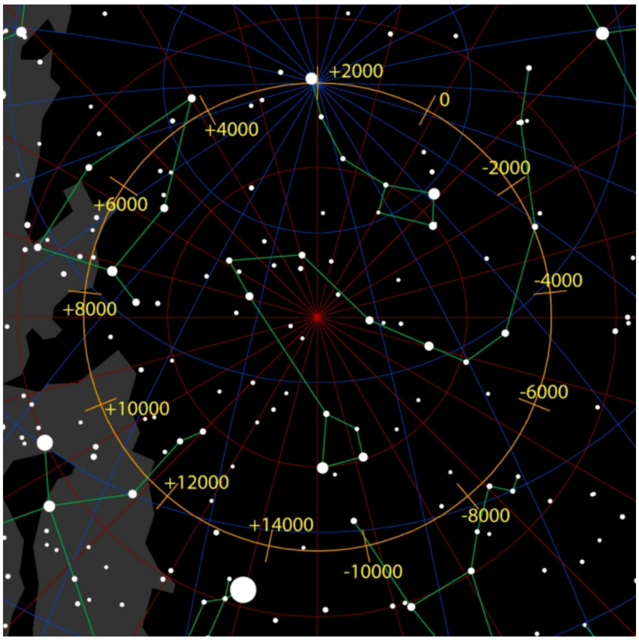
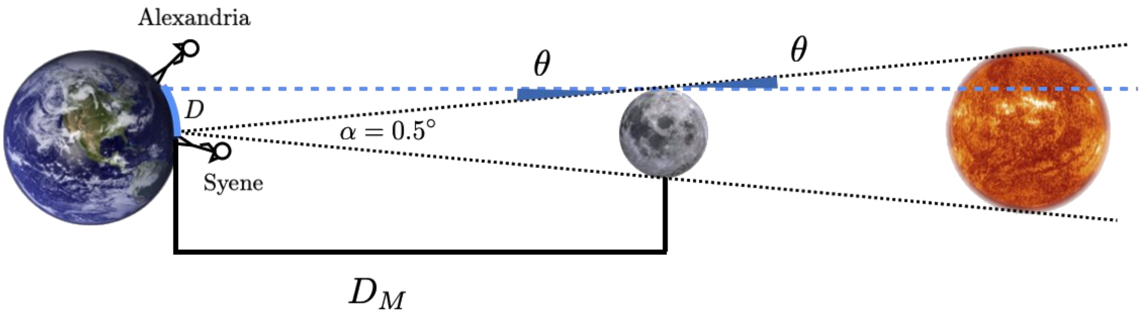


Figure 12.66: **The path of the north celestial pole amongst the stars due to the effect of precession, with dates shown (source: Wikipedia)**

Remark

There are some situations in which we can't measure the distance to a single star of some particular spectral type directly because all the stars of that type are just too far away. In these situations, we can fall back upon two techniques which deal with groups of nearly identical stars: statistical parallax and secular parallax.

Notice that the same type of trigonometric technique seems to have been used by Hipparchus to determine the distance to the Moon in 9871 (holocene calendar):



At the time of the total eclipse of the Sun an observer in Alexandria saw $1/5$ of the Sun ($\theta = \alpha/5 = 0.1°$). For small θ $\sin\theta \approx \theta = D/D_M$

This image below shows a subset of stars in by ESA Gaia's second data release (that can be downloaded here `https://cdn.gea.esac.esa.int/Gaia/`) between July 12014 to May 12016 (holocene calendar) with reasonably small parallax errors (less than 12.5% for most stars, and less than 50% for the brightest stars) and exaggerated $10,000$ times for pedagogical reasons to be visible to the human eye on a picture.

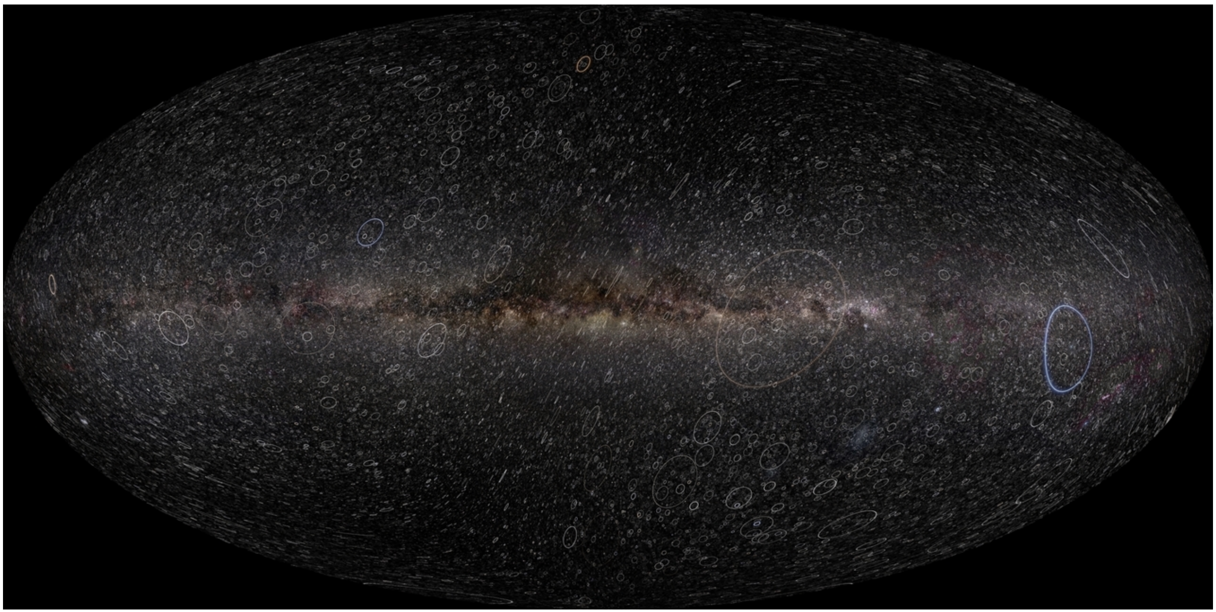


Figure 12.67: **Parallax motion seen by Gaia, exaggerated 10,000 times (credits: ESA/Gaia/DPAC, CC BY-SA 3.0 IGO. Acknowledgement: Produced by Jos de Bruijne)**

The bright horizontal structure visible at the centre of the image is the Galactic plane, the flattened disc that hosts most of the stars in our home Galaxy. The positions of stars are shown in the Galactic coordinate system.

12.1.9 Planets' Motion

On January 7, 11610 (holocene calendar), Galileo made observations of Jupiter and discovered four of the giant planet's satellites, now named after him as the "Galilean moons" (Ganymede, Callisto, Europa, and Io). They were heavenly bodies that clearly orbited an entity other than the Earth, as Galileo concluded over several nights of observation, seeing the moons change sides as they travelled around the planet. This discovery clearly contradicted the Catholic Church's Ptolemaic belief (and also various statements of the Quran - who by the way as the Bible - doesn't contain any Science but only poor value claims) that every celestial body orbits the Earth.

But the coup de grâce came within eight months, in September of 11610 (holocene calendar), when Galileo observed the planet Venus and noted that it went through complete phases, like our Moon. According to the Ptolemaic model, built on epicycles, we should only be able to see some of Venus's phases: either thin crescents (if it were on the inside of the orbit of the Sun around the Earth) or only gibbous and full phases (if it were on the outside of the Sun's orbit) - but not both. The fact that all phases of Venus are visible, just as we see with our Moon, meant to Galileo that the Ptolemaic model couldn't possibly be right. Rather, Copernicus's model of the Solar System was the one that reflected reality[17]. the Tuscan rulers could not protect Galileo from an order of extradition to face the dreaded Inquisition in Rome - which had already put to death many independent thinkers for ideas contrary to the official Catholic understanding of the universe... Although Pope Urban VIII was sympathetic to him, even the pope could not deter the Inquisition in its persecution of Galileo. The infamous trial took place in February 1633 (holocene calendar). Under threat of torture, Galileo publicly recanted his heliocentric heresy and was confined for the rest of his days to house arrest in his villa at Arcetri, outside Florence. Galileo's trial, more than any other event in history, has come to symbolize the sharp split between science and faith, a conflict that continues to rage in our own time as science still continue to debunk irrational and illogic beliefs from various religions.

---

[17] In reality the first one to make the assumption of heliocentric system seems to be the greek Aristarchus of Samos almost 1600 years before Copernicus (see [563])!

We will briefly turn our attention to the movements of the planets in ideal and situations simplified in the point of view on an observer on Earth. We consider that all the movements will be in the same plane (coplanar) perfectly circular and constant...

Definition (#12.1.527): The planets that are closer to the Sun than the Earth (whose radius is less than one astronomical unit AU[18]), that is to say the planets Mercury and Venus are "inferior planets", the other planets (Mars and beyond) are named the "outer planets".

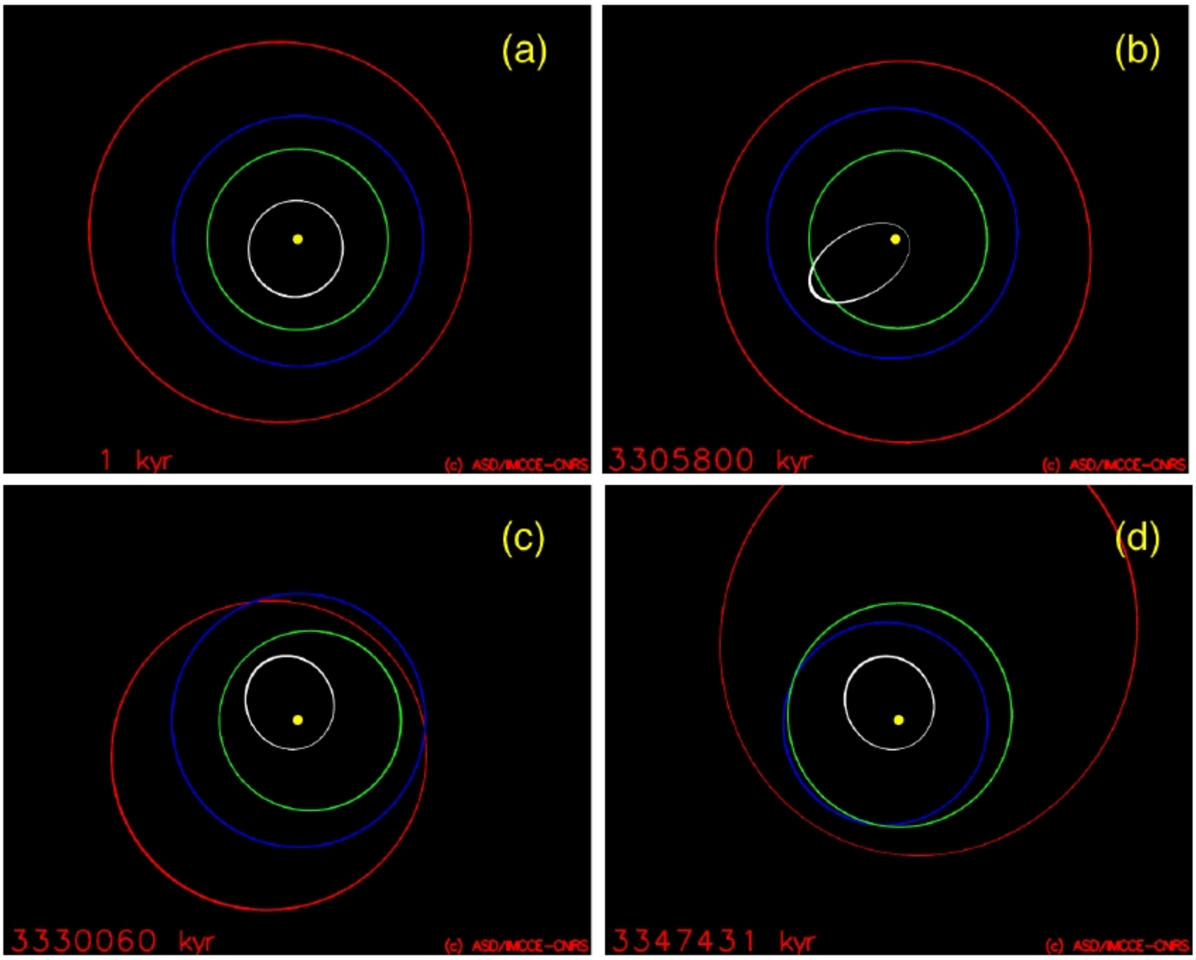


Figure 12.68: At the opposite of the myth, our Earth is far to be perfect (huge climate variations, orbiting and non-eternal star, colliding regularly with extinction level asteroids, etc.) and on a stable orbit. As shown in the computer simulations of Laskar and Gastineau [338], it may collide in the future with Venus (Mercury in white, Venus in green, Earth in blue, Mars in red).

---

[18] Defined as an average of $149,597,870,700$ [m] ((about 150 million kilometres). In ISO 80000-3, the symbol of the astronomical unit is "ua".

12.1.9.1 Synodic and Sidereal period

One of the many tools used in Astronomy are the formulas used to determine orbital motions. There are two basic forms of orbit periods:

- Sidereal Period
- Synodic Period

A "sidereal period" is an actual measure of a complete orbit relative to the stars (since the stars are unmoving - or at least moving very slowly). A "synodic period" is a rotation of a planet so that it appears to be in the same place in the night sky.

The synodic period of a planet (or satellite) is the time needed by this planet to return to the same configuration Earth-Planet-Sun (if we consider this particular case), that is to say in the same place in the sky relatively to the Sun, as seen from Earth. This period differs from the sidereal rotation period of the planet because the Earth itself moves around the Sun. Accordingly, it is the period of apparent revolution, the duration between two conjunctions Planet-Sun as viewed from Earth.

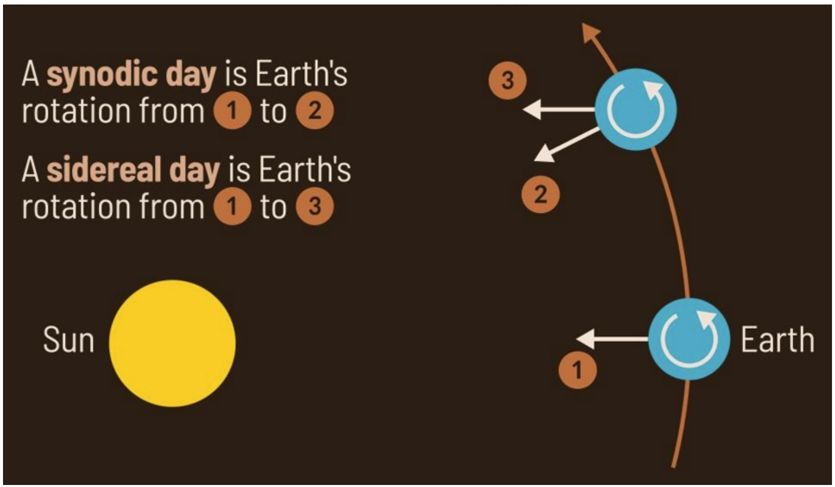

A **synodic day** is Earth's rotation from ① to ②
A **sidereal day** is Earth's rotation from ① to ③
3
2
1
Sun
Earth

Figure 12.69: **Synodic vs Sidereal day**

The term generally refers to the time between two identical aspects of the object (opposition, conjunction, etc.) and thus depends on the three bodies involved.

The "sidereal day" is the time Earth takes to rotate for a distant star to return to the same position in the sky, whereas a "solar day" is the time for the Sun to return to the same position.

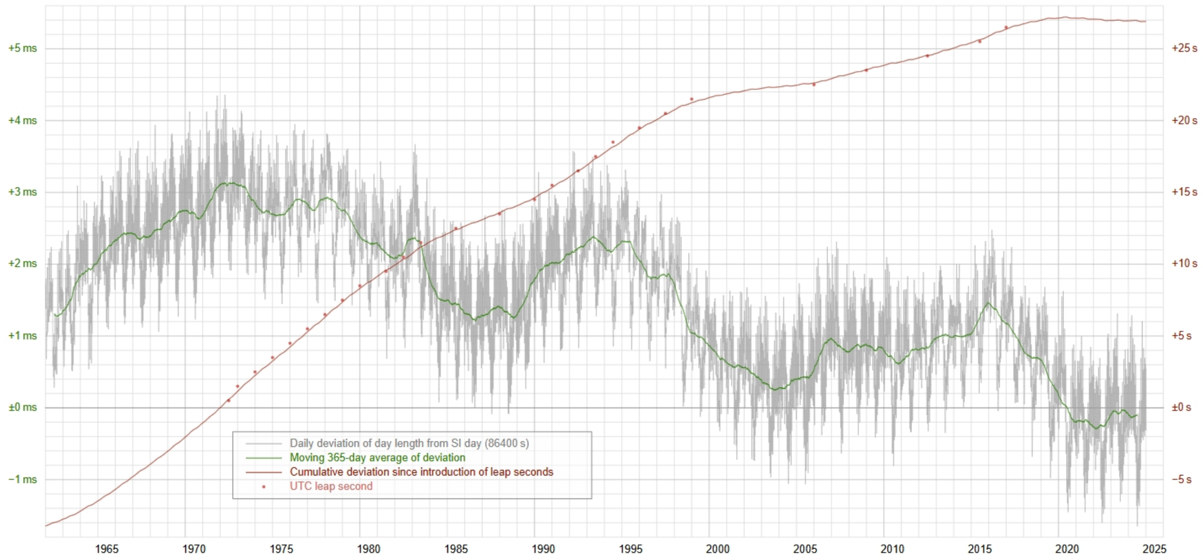


Figure 12.70: **Deviation of day length from SI-based day (86,400 s) 11962–12022: daily, moving 365-day average and cumulative (data source: International Earth rotation and Reference systems Service (IERS))**

A year has actually around 365.25 solar days, but 366.25 sidereal days to account for one revolution around the Sun. As a solar day has 24 hours, a sidereal day has around:

$$\frac{365.25}{366.25}\cdot 24\,[\text{hours}]=23\,[\text{hours}],56\,[\text{minutes}]\text{ and }4.1\,[\text{seconds}] \tag{12.1.529}$$

To mathematically study the problem in question, let us consider the following diagram with two planets describing a perfectly circular orbit at a constant angular velocity and in the same plane and in the same direction and where we have $\omega_1>\omega_2$ (thus the inner planet is faster than the outer planet):

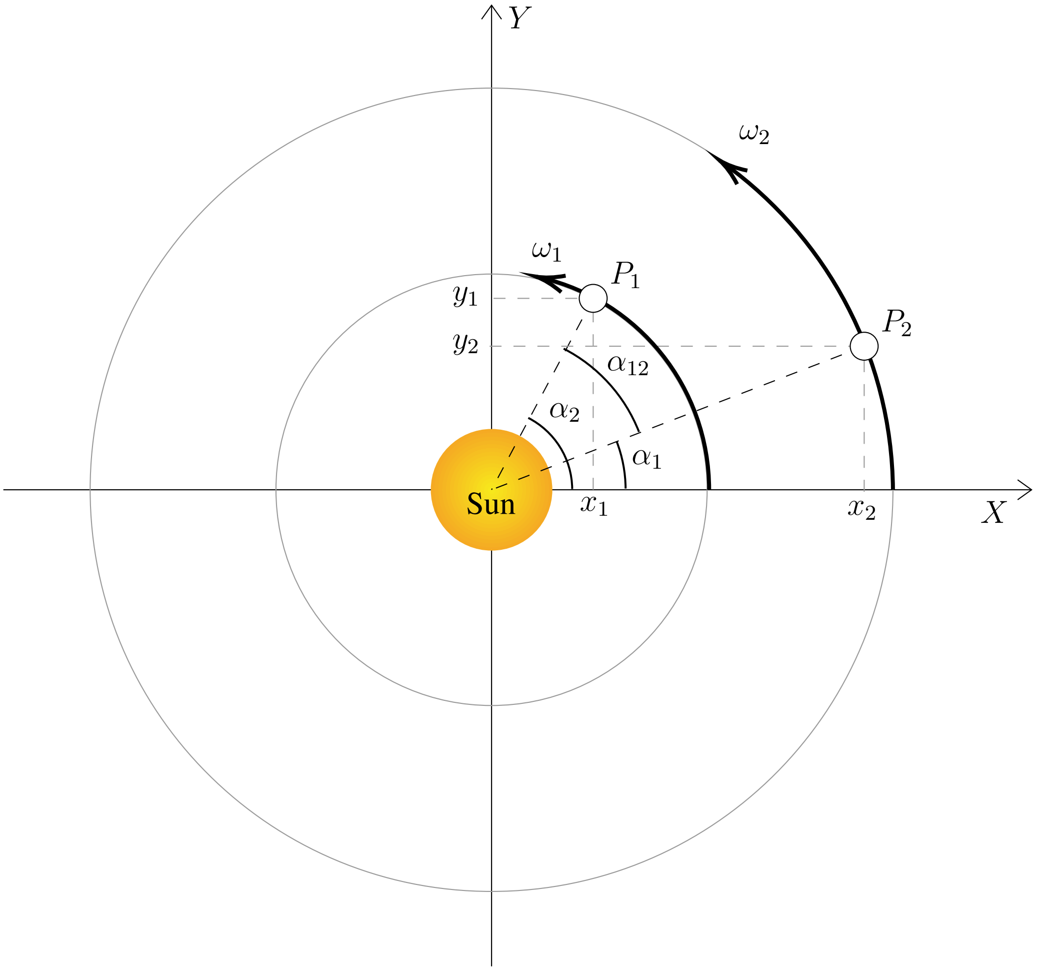


Figure 12.71: **Basic scheme for determining the synodical period**

where P_1 and P_2 are two planets which we will denote the respective sidereal periods by T_1, T_2 and for which we deduce the angular velocities:

$$\omega_1 = \frac{2\pi}{T_1} \qquad \omega_2 = \frac{2\pi}{T_2} \tag{12.1.530}$$

If we take as zero time, the time when the two planets are both aligned with the X axis and at the same side of this axis (so in "inferior conjunction"), then the angle between this axis and each of the planets is:

$$\alpha_1 = \omega_1 t \qquad \alpha_1 = \omega_2 t \tag{12.1.531}$$

We have respectively:

$$\alpha_{12} = \alpha_1 - \alpha_2 = (\omega_1 - \omega_2)t = \omega_{12} t \tag{12.1.532}$$

We seek therefore all the instants t where the following relation is satisfied for a fixed α_{12}:

$$\alpha_{12} = \omega_{12} t \tag{12.1.533}$$

Therefore it comes:

$$T = \frac{\alpha_{12}}{\omega_{12}} = \frac{\alpha_{12}}{\dfrac{2\pi}{T_1} - \dfrac{2\pi}{T_2}} = \frac{T_1 T_2 \alpha_{12}}{2\pi(T_2 - T_1)} \tag{12.1.534}$$

If we look from time zero the first (next) conjunction ("superior conjunction"), this is equivalent to put that $\alpha_{12} = \pi$ and therefore that:

$$T = \frac{T_1 T_2}{2(T_2 - T_1)} \tag{12.1.535}$$

If we look from time zero the first (next) conjunction ("inferior conjunction"), this is equivalent to put that $\alpha_{12} = 2\pi$ and therefore that:

$$\boxed{T = \frac{T_1 T_2}{T_2 - T_1}} \tag{12.1.536}$$

In the case where $\omega_2 > \omega_1$ (typically Earth and one of its outer planets), the same reasoning leads us to:

$$T = \frac{T_1 T_2}{T_2 - T_1} \tag{12.1.537}$$

Here are some actuals synodic and sidereal years of the Solar System relatively to Earth:

Planet	Synodic period [d]	Sidereal period [d]
Mercury	115.878	87.969
Venus	583.921	224.709
Mars	779.964	686.960
Jupiter	398.861	4,335.355
Saturn	378.094	10,757.737
Uranus	369.654	30,708.160
Neptune	367.486	60,224.904

Table 12.14: **Various synodic and sidereal periods**

As we can see from this table, we can make some empirical observations:

1. For the inner planets: The closer we get to the Sun, the more the synodical period is short, indeed in the proved relation above, the more T_1 is small more T decreases. So if there was a rotating planet very near the Sun, both sidereal and synodic periods are substantially equal.

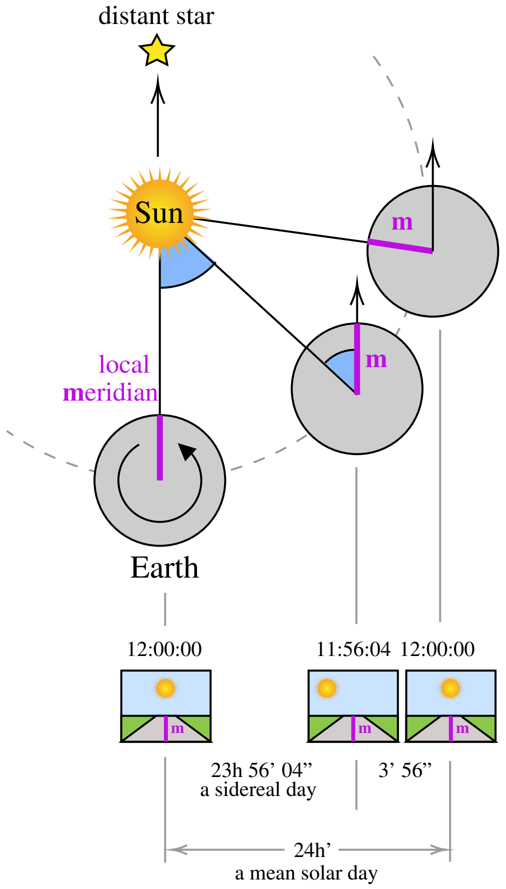
2. When we approach the Earth, the period increases. If there was a planet near to Earth, we would then have T_1 value close to T_2 value and the synodical period would be very large.

3. For the outer planets: The synodic period decreases when the planet is farther from the Earth and approaches terrestrial sidereal period of 365 days. We see well for Neptune, if we discovered a planet even further its synodic period would approach even more the 365 days.

In reality we can define other periods that the synodic or sidereal period, like for example for Earth's Moon:

Type of Period for the Moon	**Value**
Synodic Month	29.530589 days = 29 days 12 h. 44 min. 3 sec.
Sidereal Month	27.321662 days = 27 days 7 h. 43 min. 12 sec.
Anomalistic Month (perigee to perigee)	27.554550 days = 27 days 13 h. 18 min. 33 sec.
Draconic Month (node to node)	27.212221 days = 27 days 5 h. 5 min. 36 sec.

Here is another illustration of the concept taken from Wikipedia (author: Xaonon) that is very good to understand the concept if it may help:


distant star
Sun
m
local
meridian
m
Earth
12:00:00
11:56:04
12:00:00
m
m
m
23h 56’ 04”
a sidereal day
3’ 56”
24h’
a mean solar day

12.1.9.2 Planet's apparent retrograde motion

The "retrograde motion" of a planet is an apparent motion of this planet which gives the impression that it stop in his path in the "direct movement" to start reversing. This phenomenon is the result of the difference between the revolution speed of the planet and the Earth around the Sun.

The example below shows roughly what a terrestrial observer (blue discs) can be observed by monitoring month after month, the apparent motion of Mars (brown discs):

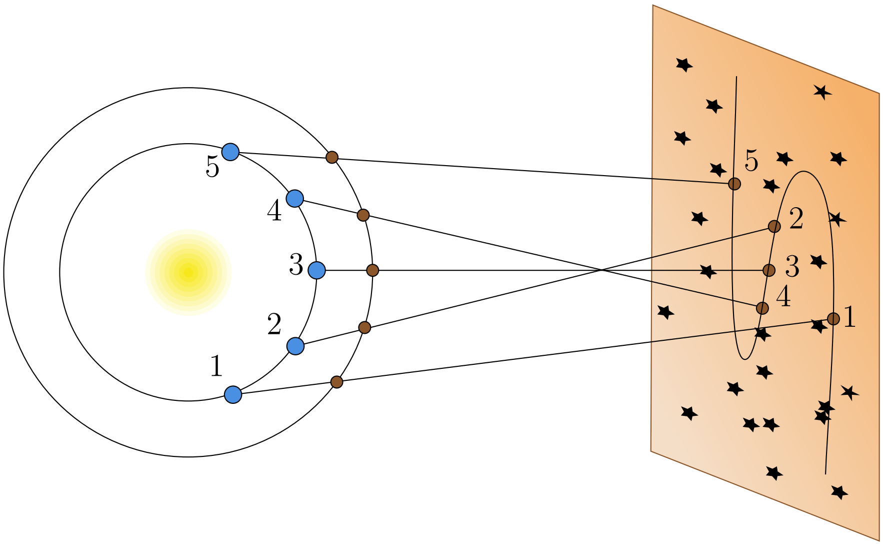


Figure 12.72: **Retrograde motion principle (source: Wikipedia)**

Or more explicitly (in the legend of the figure below, the years are given based on the holocene calendar):

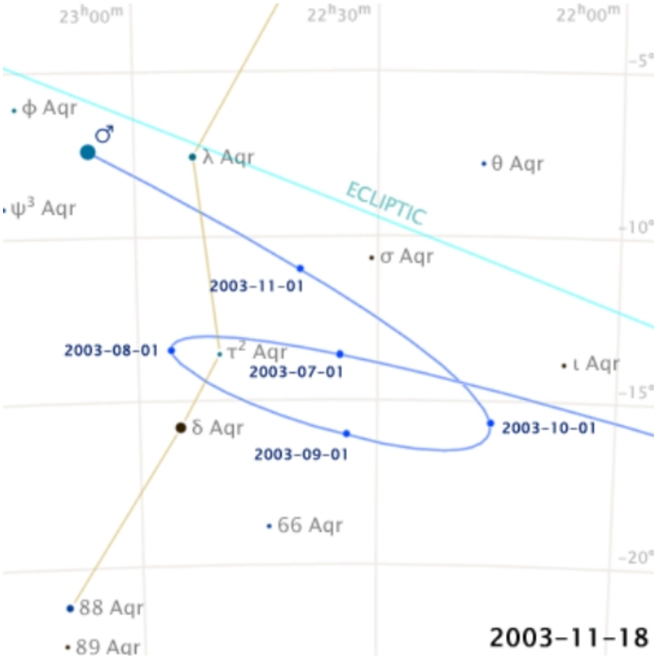


Figure 12.73: **Apparent retrograde motion of Mars in 12003 as seen from Earth (source: Wikipedia, author: Eugene Alvin Villar)**

To study this phenomenon mathematically, we will consider the following figure:

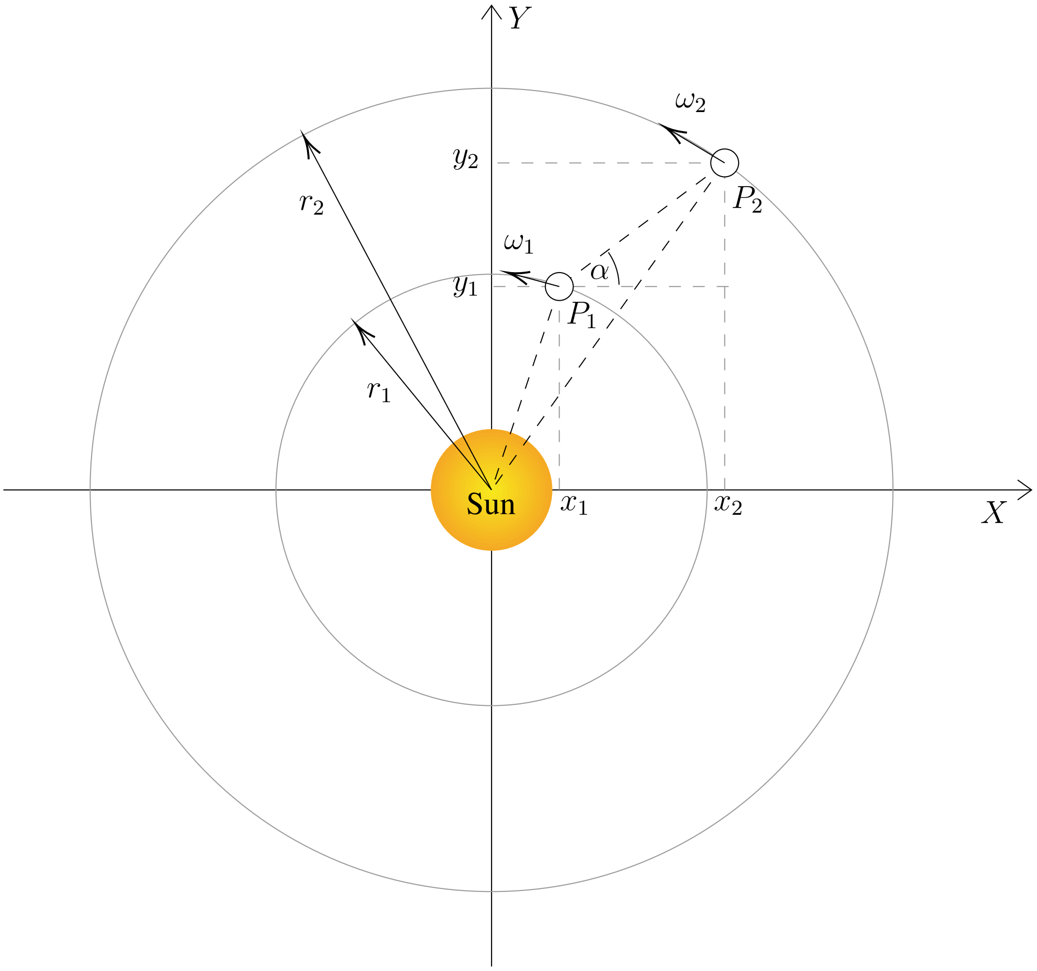


Figure 12.74: Basic scheme for the study of planet's retrograde motion

with two planets describing a perfectly circular orbit at a constant angular velocity and in the same plane and in the same direction and where we have. It is clear that the inner planet will therefore caught up the outer planet and it will seem to have a retrograde motion as shown in the figure below:

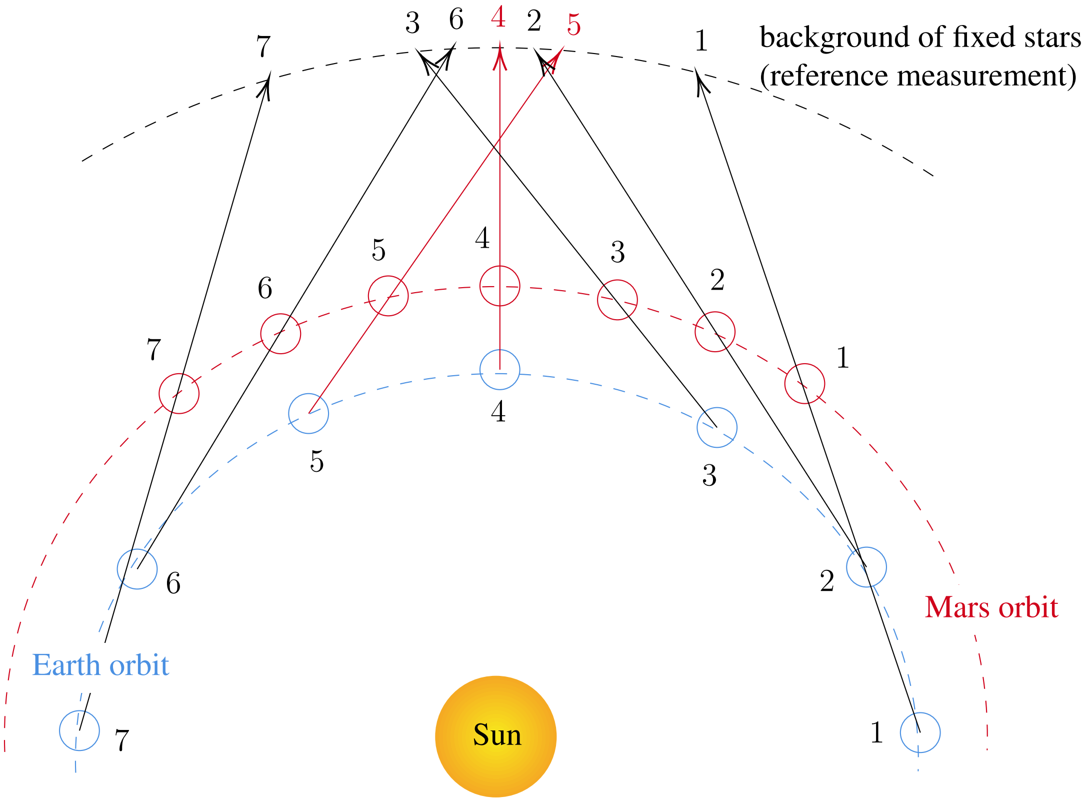


Figure 12.75: **Figure to illustrate the choice of zero time**

As the reader can check it in the figure above we see that the retrograde motion with respect to the fixed stars begins when the angle between the two planets is zero and it ends when the angle between the two planets pass through a maximum.

Therefore, in the prior previous figure, we have:

$$\tan(\alpha)=\frac{y_2-y_1}{x_2-x_1}=\frac{r_2\sin(\omega_2 t)-r_1\sin(\omega_1 t)}{r_2\cos(\omega_2 t)-r_1\cos(\omega_1 t)} \quad (12.1.538)$$

So to know the time between when the moment where the angle is zero between the two planets, reaches a maximum and decreases again, we simply need determine when occurs the sign of change in the previous function. To do this we just search when the derivative is zero:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{r_2\sin(\omega_2 t)-r_1\sin(\omega_1 t)}{r_2\cos(\omega_2 t)-r_1\cos(\omega_1 t)}\right)=0 \quad (12.1.539)$$

By applying the derivation rules seen in the section of Differential and Integral Calculus:

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\left([r_2\sin(\omega_2 t)-r_1\sin(\omega_1 t)]\frac{1}{r_2\cos(\omega_2 t)-r_1\cos(\omega_1 t)}\right)\\
&=r_2\omega_2\cos(\omega_2 t)-r_1\omega_1\cos(\omega_1 t)\frac{1}{r_2\cos(\omega_2 t)-r_1\cos(\omega_1 t)}\\
&-\frac{r_2\sin(\omega_2 t)-r_1\sin(\omega_1 t)}{(r_2\cos(\omega_2 t)-r_1\cos(\omega_1 t))^2}(-r_2\omega_2\sin(\omega_2 t)+r_1\omega_1\sin(\omega_1 t))\\
&=\frac{(r_2\omega_2\cos(\omega_2 t)-r_1\omega_1\cos(\omega_1 t)(r_2\cos(\omega_2 t)-r_1\cos(\omega_1 t))}{(r_2\cos(\omega_2 t)-r_1\cos(\omega_1 t))^2}\\
&-\frac{(r_2\sin(\omega_2 t)-r_1\sin(\omega_1 t))(-r_2\omega_2\sin(\omega_2 t)+r_1\omega_1\sin(\omega_1 t))}{(r_2\cos(\omega_2 t)-r_1\cos(\omega_1 t))^2}=0
\end{aligned} \quad (12.1.540)$$

Hence after simplification:

$$\begin{aligned}&(r_2\omega_2\cos(\omega_2 t)-r_1\omega_1\cos(\omega_1 t))(r_2\cos(\omega_2 t)-r_1\cos(\omega_1 t))\\&-(r_2\sin(\omega_2 t)-r_1\sin(\omega_1 t))(-r_2\omega_2\sin(\omega_2 t)+r_1\omega_1\sin(\omega_1 t))=0\end{aligned}\tag{12.1.541}$$

We develop all this:

$$\begin{aligned}&r_2^2\omega_2\cos^2(\omega_2 t)-r_2r_1\cos(\omega_2 t)\cos(\omega_1 t)-r_1r_2\cos(\omega_1 t)\cos(\omega_2 t)+r_1^2\omega_1\cos^2(\omega_1 t)\\&+r_2^2\omega_2\sin^2(\omega_2 t)-r_2r_1\omega_1\sin(\omega_2 t)\sin(\omega_1 t)-r_1r_2\omega_2\sin(\omega_1 t)\sin(\omega_2 t)+r_1^2\omega_1\sin^2(\omega_1 t)=0\end{aligned}\tag{12.1.542}$$

and we simplify a first time:

$$\begin{aligned}&r_2^2\omega_2-r_2r_1\omega_2\cos(\omega_2 t)\cos(\omega_1 t)-r_1r_2\omega_1\cos(\omega_1 t)\cos(\omega_2 t)+r_1^2\omega_1\\&-r_2r_1\omega_1\sin(\omega_2 t)\sin(\omega_1 t)-r_1r_2\omega_2\sin(\omega_1 t)\sin(\omega_2 t)=0\end{aligned}\tag{12.1.543}$$

and second:

$$r_2^2\omega_2+r_1^2\omega_1-r_1r_2\begin{bmatrix}\omega_2\cos(\omega_2 t)\cos(\omega_1 t)+\omega_1\cos(\omega_1 t)\cos(\omega_2 t)\\+\omega_1\sin(\omega_2 t)\sin(\omega_1 t)+\omega_2\sin(\omega_1 t)\sin(\omega_2 t)\end{bmatrix}=0\tag{12.1.544}$$

and finally a third one:

$$r_2^2\omega_2+r_1^2\omega_1-r_1r_2(\omega_1+\omega_2)[\cos(\omega_1 t)\cos(\omega_2 t)+\sin(\omega_1 t)\sin(\omega_2 t)]=0\tag{12.1.545}$$

and after rearrangement:

$$r_2^2\omega_2+r_1^2\omega_1=r_1r_2(\omega_1+\omega_2)[\cos(\omega_1 t)\cos(\omega_2 t)+\sin(\omega_1 t)\sin(\omega_2 t)]\tag{12.1.546}$$

We simplify using trigonometric identities proved in the section Trigonometry:

$$r_2^2\omega_2+r_1^2\omega_1=r_1r_2(\omega_1+\omega_2)\cos((\omega_1-\omega_2)t)\tag{12.1.547}$$

The values of t that satisfy this relation gives us the sign change we were looking for.

If t_0 is the first value of t that satisfies the equation, we have:

$$\cos((\omega_1-\omega_2)t)=\frac{r_2^2\omega_2+r_1^2\omega_1}{r_1r_2(\omega_1+\omega_2)}\tag{12.1.548}$$

The next value of t will be such that:

$$(\omega_1-\omega_2)t_1=2\pi\pm(\omega_1-\omega_2)t_0\tag{12.1.549}$$

and therefore:

$$t_1=\frac{2\pi}{\omega_1-\omega_2}\pm t_0\tag{12.1.550}$$

If we introduce the rotation periods, we have:

$$t_1=\frac{T_1T_2}{T_2-T_1}\pm t_0\tag{12.1.551}$$

To come back to:

$$\cos((\omega_1-\omega_2)t_0)=\frac{r_2^2\omega_2+r_1^2\omega_1}{r_1r_2(\omega_1+\omega_2)} \tag{12.1.552}$$

it may be more convenient to write it in the traditional following form:

$$\begin{aligned}\cos\left(2\pi\frac{T_2-T_1}{T_2T_1}t_0\right)&=\frac{r_2^2\dfrac{2\pi}{T_2}+r_1^2\dfrac{2\pi}{T_1}}{r_1r_2\left(\dfrac{2\pi}{T_1}+\dfrac{2\pi}{T_2}\right)}=\frac{r_2^2\dfrac{1}{T_2}+r_1^2\dfrac{1}{T_1}}{r_1r_2\left(\dfrac{1}{T_1}+\dfrac{1}{T_2}\right)}=\frac{r_2^2\dfrac{T_1}{T_2T_1}+r_1^2\dfrac{T_2}{T_1T_2}}{r_1r_2\left(\dfrac{T_1+T_2}{T_1T_2}\right)}\\&=\frac{\dfrac{r_2^2T_1+r_1^2T_2}{T_1T_2}}{r_1r_2\left(\dfrac{T_1+T_2}{T_1T_2}\right)}=\frac{r_2^2T_1+r_1^2T_2}{r_1r_2(T_1+T_2)}\end{aligned} \tag{12.1.553}$$

So far, we have only do geometry. No law of gravitation intervened in the calculations. As the radius are unknown or little known (at least historically), we will use the Kepler's third law (periods law) that is for recall:

$$T^2=c^{te}D^3 \tag{12.1.554}$$

where for recall D is the semi-major axis of the orbit, and if it is circular, it becomes a simple radius. So we have:

$$r=\frac{T^{2/3}}{\sqrt{c^{te}}} \tag{12.1.555}$$

Therefore:

$$\cos\left(2\pi\frac{T_2-T_1}{T_2T_1}t_0\right)=\frac{r_2^2T_1+r_1^2T_2}{r_1r_2(T_1+T_2)}=\frac{\dfrac{T_2^{4/3}}{c^{te}}T_1+\dfrac{T_1^{4/3}}{c^{te}}T_2}{\dfrac{T_1^{2/3}}{\sqrt{c^{te}}}\dfrac{T_2^{2/3}}{\sqrt{c^{te}}}(T_1+T_2)}=\frac{T_2^{4/3}T_1+T_1^{4/3}T_2}{T_1^{2/3}T_2^{2/3}(T_1+T_2)} \tag{12.1.556}$$

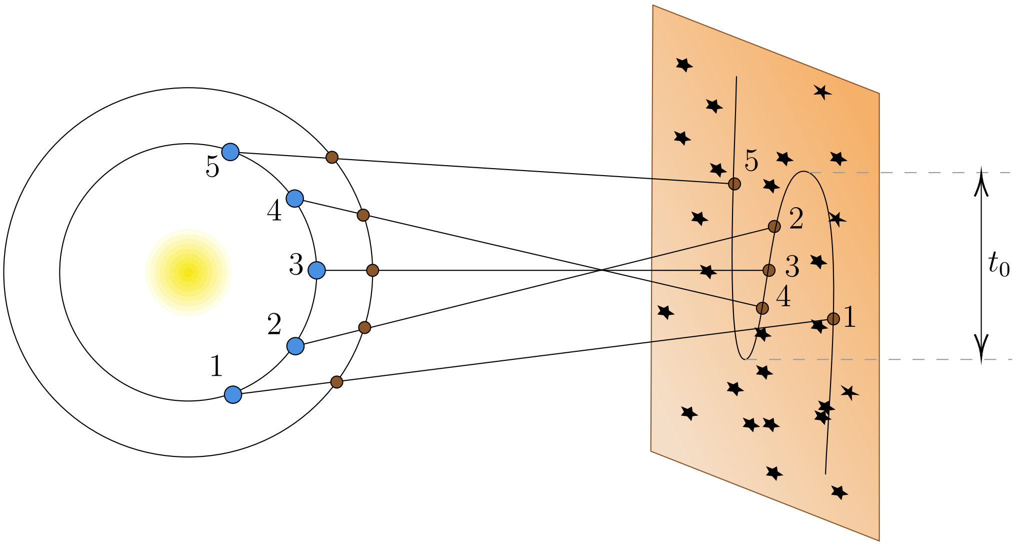
hence:

$$\boxed{t_0=\frac{T_2T_1}{2\pi(T_2-T_1)}\arccos\left(\frac{T_2^{4/3}T_1+T_1^{4/3}T_2}{T_1^{2/3}T_2^{2/3}(T_1+T_2)}\right)} \tag{12.1.557}$$

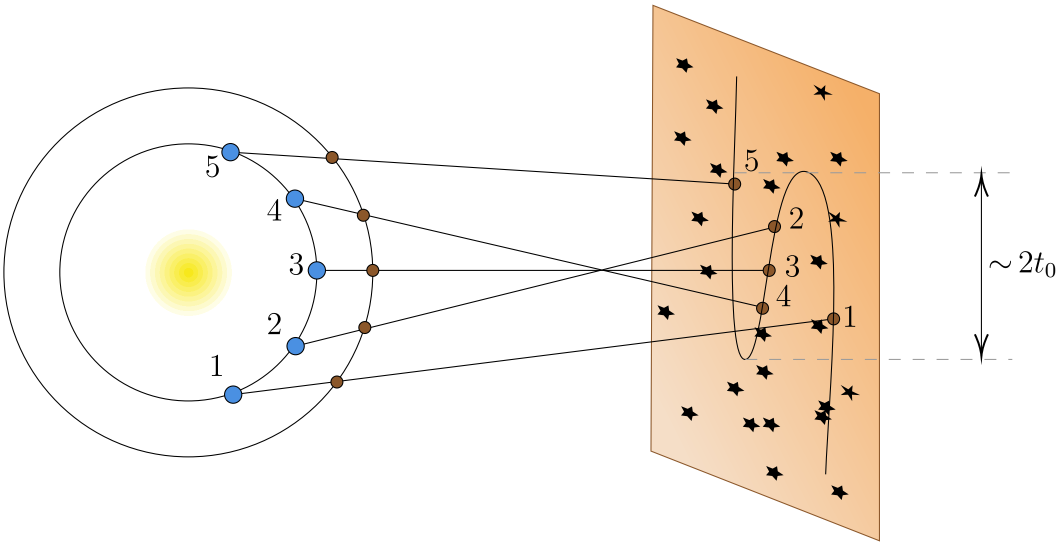
A numerical application with for Mercury $T_1\cong 87.95$ [d] and for Earth $T_2\cong 365.25$ [d] leads to the value:

$$t_0\cong 11.44\,[\text{d}] \tag{12.1.558}$$

Value that we have represented in the diagram below:


5
4
3
2
1
5
2
3
4
1
t_0

and therefore:


5
4
3
2
1
5
2
3
4
1
$\sim 2t_0$

and therefore we have:

$$t_1 = \frac{T_1 T_2}{T_2 - T_1} \pm t_0 \cong 115.84 \pm 11.44 \tag{12.1.559}$$

then a new cycle:

$$t_2 = \frac{T_1 T_2}{T_2 - T_1} \pm t_1 \tag{12.1.560}$$

etc. What gives in schematic form:

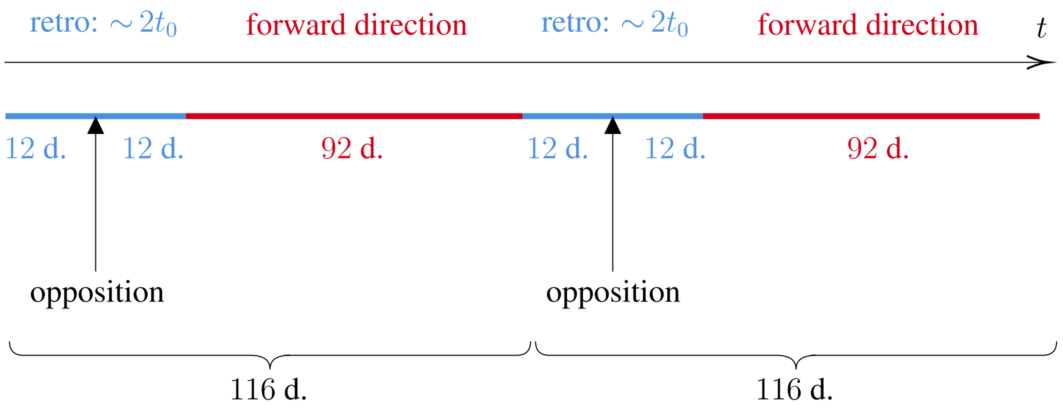

retro: $\sim 2t_0$
forward direction
retro: $\sim 2t_0$
forward direction
t
12 d.
12 d.
92 d.
12 d.
12 d.
92 d.
opposition
opposition
116 d.
116 d.

Figure 12.76: Retrogradiation cycle diagram principle

Remark

At specific points on Mercury's surface, an observer would be able to see the Sun (actually a G2V-type star, aka yellow dwarf) rise part way, then reverse and set before rising again, all within the same Mercurian day. This apparent retrograde motion of the Sun occurs because, from approximately four Earth days before perihelion until approximately four Earth days after it, Mercury's angular orbital speed exceeds its angular rotational velocity. Mercury's elliptical orbit is farther from circular than that of any other planet in the Solar System, resulting in a substantially higher orbital speed near perihelion.

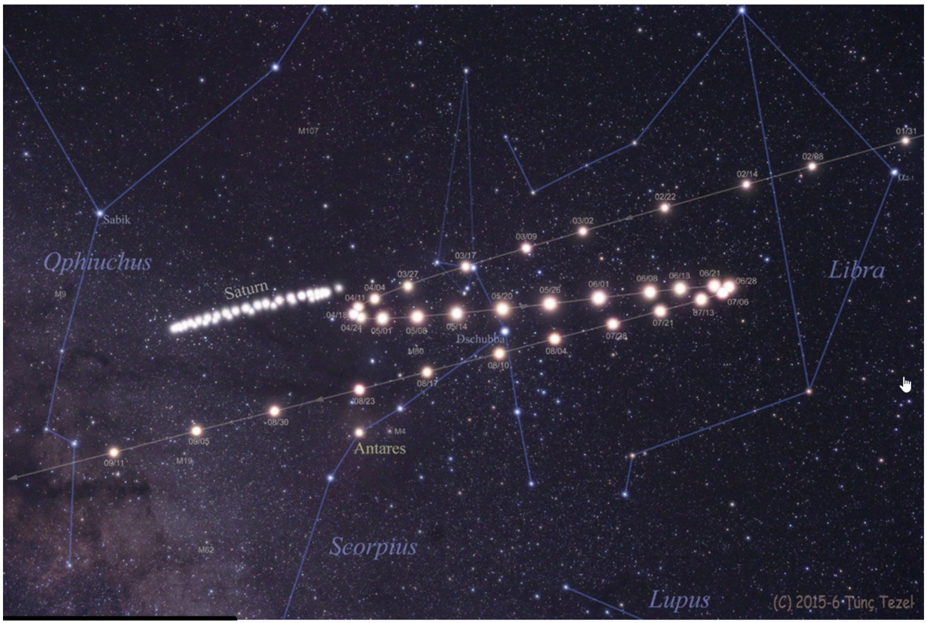

Ophiuchus
Sabik
Saturn
Libra
Dschubba
Antares
Scorpius
Lupus
M107
M9
M19
M4
M80
M62
(C) 2015-6 Tunç Tezel

Figure 12.77: Real sequence of exposures showing Mars and Saturn retrograde motion (author: Tunç Tezel)

Let us see now how to simulate the retrograde motion on a computer of Mars relatively to Earth!

Since the eccentricities of the two planets are small (0.017 and 0.093), the two elliptical trajectories can be equated with circles (eccentricity equal to 0). Since Mars's inclination is small, the two trajectories can be equated, as a first approximation, with coplanar circles centered on the Sun.

To simplify matters further, we will use rounded data and study the following simplified model:

	Radius of the orbit (in half au)	**Revolution period** (in terrestrial year)
Earth	2	1
Mars	3.048	1.88

With this simplified model, the Sun, Earth, and Mars are aligned once a year. The respective positions of Earth and Mars (measured in half astronomical units) relative to the Sun can be written as a function of time t (measured in Earth years from the alignment) as complex numbers ($\mathrm{i}^2 = -1$):

$$\begin{cases} z_{E/S}(t) = 2e^{+2\pi\mathrm{i}t} \\ z_{M/S}(t) = 3.048e^{+\pi\mathrm{i}t} \end{cases} \tag{12.1.561}$$

We deduce that the relative position of Mars with respect to the Earth is given by:

$$z_{M/E}(t) = z_{M/S}(t) - z_{E/S}(t) = 3.048e^{+\pi\mathrm{i}t} - 2e^{+2\pi\mathrm{i}t} \tag{12.1.562}$$

Using Euler's property, $e^{\mathrm{i}x} = \cos(x) + \mathrm{i}\sin(x)$, and then separating the real part (abscissa) from the imaginary part (ordinate), we get the parametric equations of a "cardioid" (an irregular one because the real cardois hast the same coefficient in front of each trigonometric term):

$$\begin{cases} x = 3.048\cos(\pi t) - 2\cos(2\pi t) \\ y = 3.048\sin(\pi t) - 2\sin(2\pi t) \end{cases} \tag{12.1.563}$$

Notice that the choice that we did starts is such that at $t = 0$ we start with an opposition between Earth and Mars!

The representation of this relative position in the plane gives the following trajectory with Maple™ 4.00b (remember that the variable t is in unit of years, hence $t = 1$ is equivalent to one year):

```
>plot([3*cos(2/1.88*Pi*t) - 2*cos(2*Pi*t), 3*sin(2/1.88*Pi*t) -
2*sin(2*Pi*t) , t=0..2.5], scaling=constrained)
```

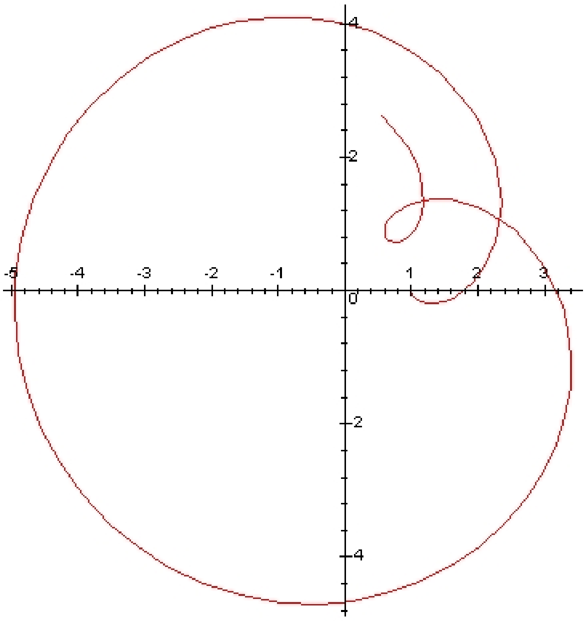


Figure 12.78: **Simulation of retrograde motion of Mars and Earth with Maple™ 4.00b**

Mars undergoes retrograde motion, which is an apparent back-and-forth movement in the sky, approximately every 26 months like visible in the plot above. This means that we can observe Mars retrograde roughly once every two years, or roughly 0.5 times per year. The retrograde period lasts about two and a half months.

Planetary retrogradation is a clear observational proof that we exist in a heliocentric system rather than a geocentric one. If the latter were true, planets would have to defy the principles of conservation of energy and angular momentum!

12.1.10 Equation of Time and Analemma

In astronomy, a "Sun analemma" is a diagram showing the position of the Sun (actually a G2V-type star, aka yellow dwarf) in the sky as seen from a fixed location on Earth at the same mean solar time, as that position varies over the course of a year. The diagram will resemble a figure eight seen from Earth nowadays (see further below for the detailed proof).

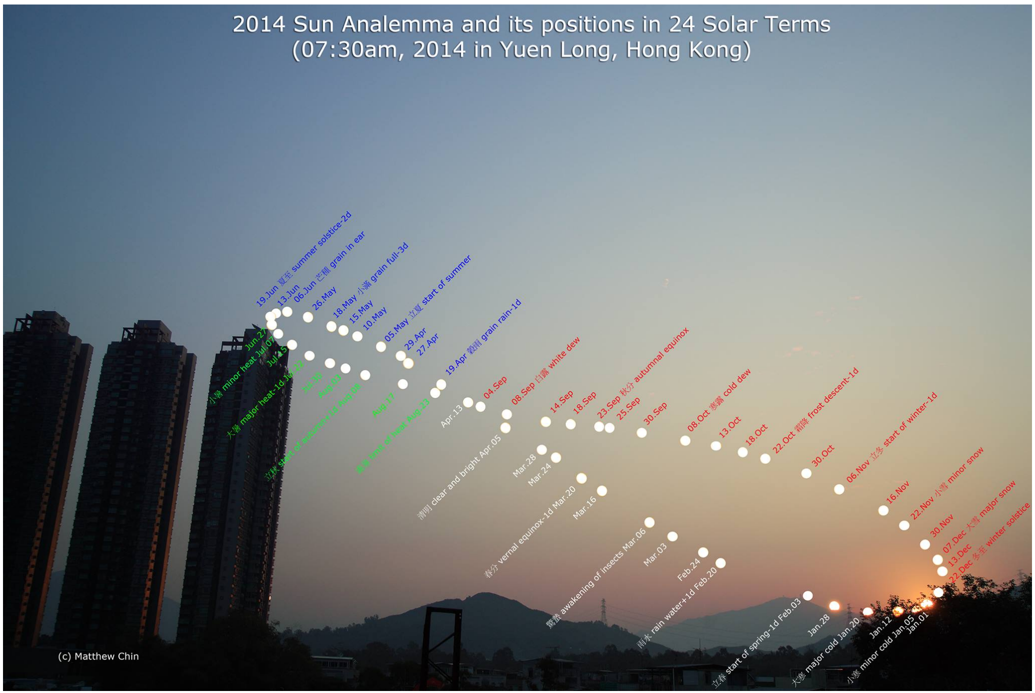

2014 Sun Analemma and its positions in 24 Solar Terms
(07:30am, 2014 in Yuen Long, Hong Kong)
19.Jun 夏至 summer solstice-2d
13.Jun
06.Jun 芒種 grain in ear
26.May
18.May 小滿 grain full-3d
15.May
10.May
05.May 立夏 start of summer
29.Apr
27.Apr
19.Apr 穀雨 grain rain-1d
Jun.27
小暑 minor heat Jul.07
Jul.15
大暑 major heat-1d Jul.22
Jul.30
Aug.03
立秋 start of autumn+1d Aug.08
Aug.17
處暑 limit of heat Aug.23
Apr.13
清明 clear and bright Apr.05
04.Sep
08.Sep 白露 white dew
14.Sep
18.Sep
23.Sep 秋分 autumnal equinox
25.Sep
30.Sep
Mar.28
Mar.24
春分 vernal equinox-1d Mar.20
Mar.16
08.Oct 寒露 cold dew
13.Oct
18.Oct
22.Oct 霜降 frost descent-1d
30.Oct
06.Nov 立冬 start of winter-1d
16.Nov
22.Nov 小雪 minor snow
30.Nov
07.Dec 大雪 major snow
13.Dec
22.Dec 冬至 winter solstice
驚蟄 awakening of insects Mar.06
Mar.03
Feb.24
雨水 rain water+1d Feb.20
立春 start of spring-1d Feb.03
Jan.28
大寒 major cold Jan.20
Jan.12
小寒 minor cold Jan.05
Jan.01
(c) Matthew Chin

Figure 12.79: **Analemma and solar terms, 2014, from Hong Kong. Copyright Matthew Chin. Used with permission.**

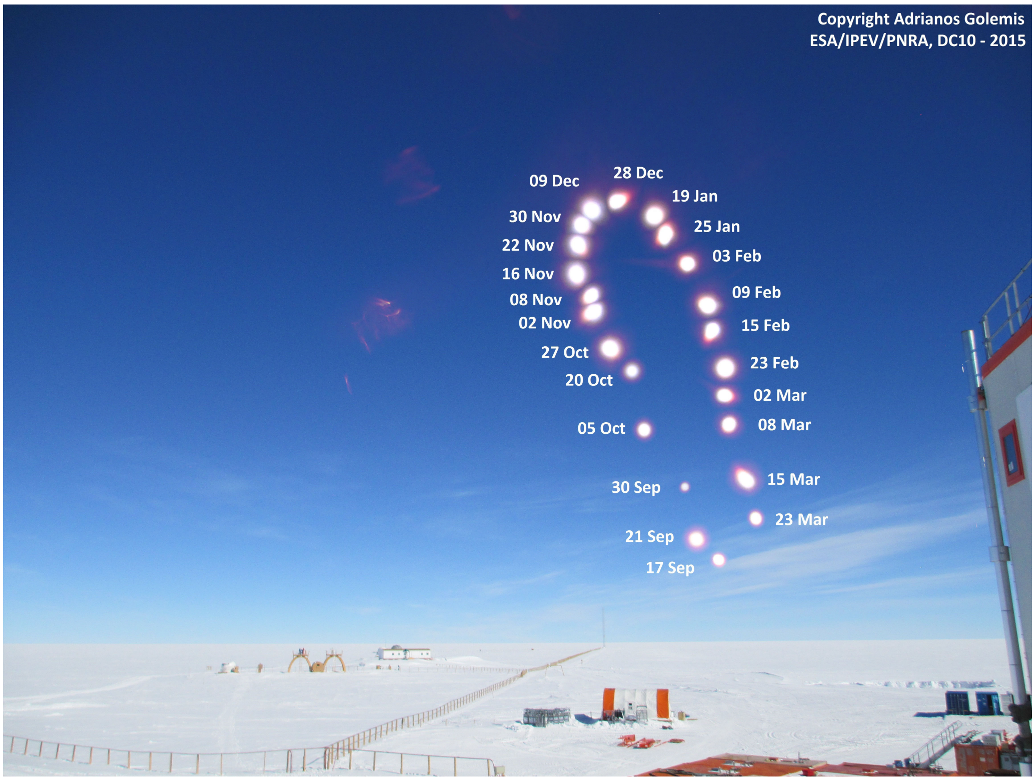

Copyright Adrianos Golemis
ESA/IPEV/PNRA, DC10 - 2015
09 Dec
28 Dec
19 Jan
30 Nov
25 Jan
22 Nov
03 Feb
16 Nov
09 Feb
08 Nov
02 Nov
15 Feb
27 Oct
23 Feb
20 Oct
02 Mar
05 Oct
08 Mar
30 Sep
15 Mar
23 Mar
21 Sep
17 Sep

Figure 12.80: **Analemma as viewed from Antarctica**

More generaly an "analemma" is the closed, often figure-eight–shaped curve traced by the apparent position of a periodically observed object when it is recorded at the same clock time over many consecutive days.

Below we can see the Moon's analemma take by Wael Omar with its Canon 200 D, Canon 18-55 lens, Redcat 51 optics, GOTO mount (foreground image: ISO 100, 30 seconds, F 11):

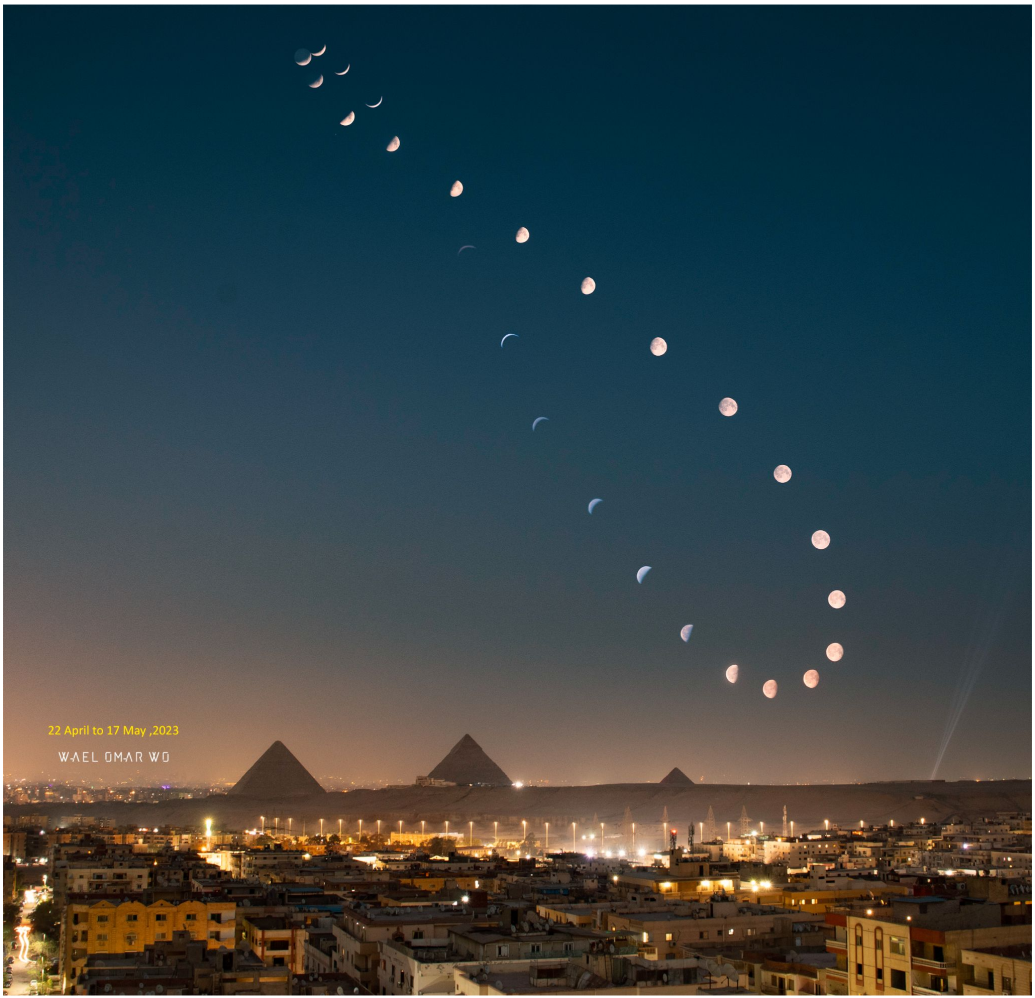


Figure 12.81: **Analemma viewed from Cairo/Egypt (author: Wael Omar)**

The North–South component of the analemma results from the change in the Sun's declination (angular distance of a celestial body North (+) or South (−) from the celestial equator) due to the tilt of Earth's axis of rotation as it orbits around the Sun. The east–west component results from the nonuniform rate of change of the Sun's right ascension (the celestial equivalent of longitude, measuring the East-West position of an astronomical object on the celestial sphere), governed by the combined effects of Earth's axial tilt and its orbital eccentricity[19].

One can photograph an analemma by keeping a camera at a fixed location and orientation and taking multiple exposures throughout the year, always at the same time of day (disregarding daylight saving time).

---

[19] The eccentricity of the Earth's orbit varies over time, due to gravitational interaction with other celestial bodies. The eccentricity value can get as low as 0.0034 and up to 0.058.

Although the term analemma usually refers to "Earth's solar analemma", it can be applied to other celestial bodies as well.

Remark

The text that follows is almost entirely inspired by the excellent paper *Equation of time - Problem in Astronomy* by Dr. Markus Müller (see [403]). His paper was awarded in the II International Competition (11993/94 according to holocene calendar). Notice that the first mathematical description of the analemma seems to have been made the French astronomer Grandjean de Fouchy in 11740 (holocene calendar).

We will deal here with a problem of the astronomical measurement of time. Let us first introduce some basic definitions!

Definitions (#12.1.528):

D1. The natural unit of time is the rotation of the Earth, that is the apparent daily course of the Sun. The length of time between two culminations of the Sun is named a "solar day". The time-system based on this unit is named "apparent solar time". By this system, localities on a given meridian always have the same time-readings.

D2. A comparison of a sundial with a mechanical clock shows that the solar day has a variable length. Therefore, so-named "mean solar time" is commonly used. This is based on a unit which is defined as the average of a solar day. The mean solar time has been fixed in such a way that it does not deviate too much from the apparent solar time. The deviations between apparent solar time and mean solar time are described by the equation of time:

$$\text{Equation of time} = (\text{Apparent solar time}) - (\text{Mean solar time}) \tag{12.1.564}$$

Suitable approximations and relevant aspects of the equation of time are discussed in what follows. The derivation does not account for minor effects due to the gravitational fields of the Moon and the planets. In principle, therefore, a comparison of the results of such an idealized equation of time with the actual observations can be used to estimate the magnitudes of these effects. Furthermore, parameters of the orbit of the Earth, such as its eccentricity, can be verified or calculated. It should be mentioned that the equation of time was very important for navigation in earlier times.

Two well-known features of our Solar System are at the basis of the variations in the apparent motion of the Sun:

1. According to Kepler's second law (see page 5234), the angular velocity of the Earth relative to the Sun varies throughout a year.
2. Equal angles which the Sun in its apparent movement goes through in the ecliptic do not correspond to equal angles we measure on the equatorial plane (see figure below). However, it is these latter angles which are relevant for the measurement of time, since the daily movement of the Sun is parallel to the equatorial plane.

So we will have to take into account these two factors: the eccentricity (variable angular velocity of the Earth) + the obliquity.

Let us start with the calculations of the first one!

12.1.10.1 Eccentricity component (variable angular velocity of Earth)

Definition (#12.1.529): The "celestial equator" is an extension of Earth's equator in space that bisects the sky into two hemisphere. The celestial equator is positioned at 0 degrees declination (angular distance of a celestial body North $(+)$ or South $(-)$ from the celestial equator), like Earth's equator is 0 degrees latitude. In other words the celestial equator is the great circle of the imaginary celestial sphere on the same plane as the equator of Earth. By extension, it is also a plane of reference in the equatorial coordinate system. Due to Earth's axial tilt, the celestial equator is currently inclined by about $23.44°$ with respect to the ecliptic (the plane of Earth's orbit).

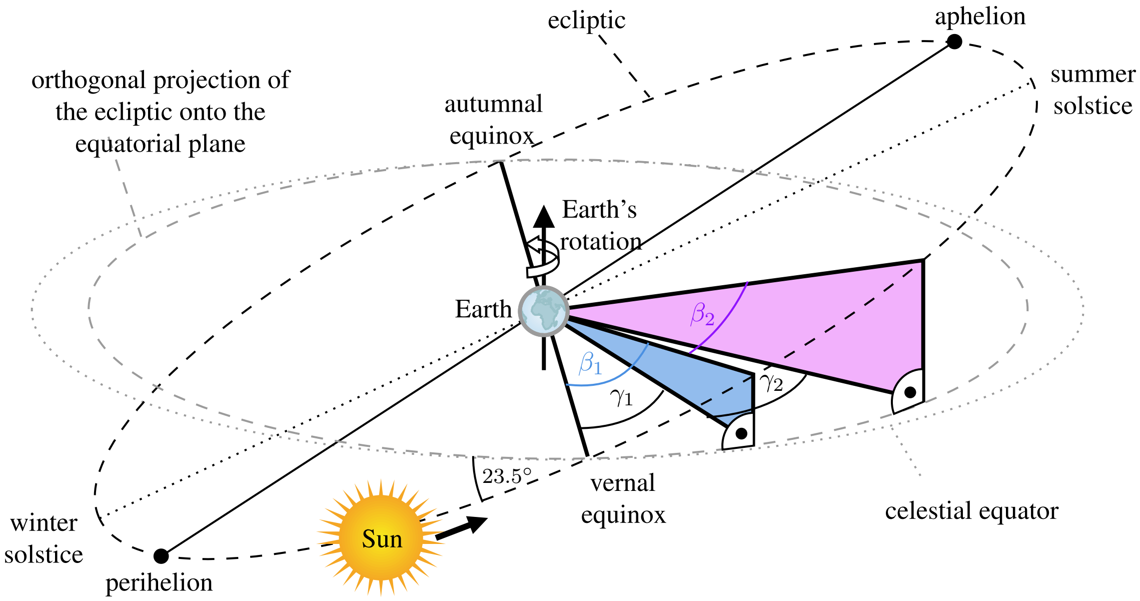

ecliptic
aphelion
orthogonal projection of the ecliptic onto the equatorial plane
autumnal equinox
summer solstice
Earth's rotation
Earth
β_2
β_1
γ_1
γ_2
23.5°
vernal equinox
celestial equator
winter solstice
perihelion
Sun

Figure 12.82: **Apparent path of the Sun in a geocentric view. At the perihelion, the Sun runs faster than at the aphelion. Equal angles β on the plane of the ecliptic do not correspond to equal angles γ on the equatorial plane. In this figure there is $\beta_1 = \beta_2$ but $\gamma_1 < \gamma_2$.**

Kepler's first law tells us that all planets are moving in elliptical orbits around the sun, whereby the latter is positioned at one of the two focal points. Kepler's second law - the so-named "law of areas - describes the velocity of the planets (see page 5234). The area swept out per time interval is constant $(= \mathrm{d}S/\mathrm{d}t)$.

Hence, during the time period t the radius vector from the Sun to the Earth sweeps out an area of:

$$t\frac{\mathrm{d}S}{\mathrm{d}t} = t\frac{\pi ab}{T} = \frac{t}{T}\pi ab \qquad (12.1.565)$$

where a, b are the axes of the ellipse and T the duration of a revolution.

Let us now derive the angular velocity of the Earth as a function of time. The angle covered by the Earth after leaving the perihelion is named "true anomaly", denoted here with $\hat{R}$ as visible in the figure below:

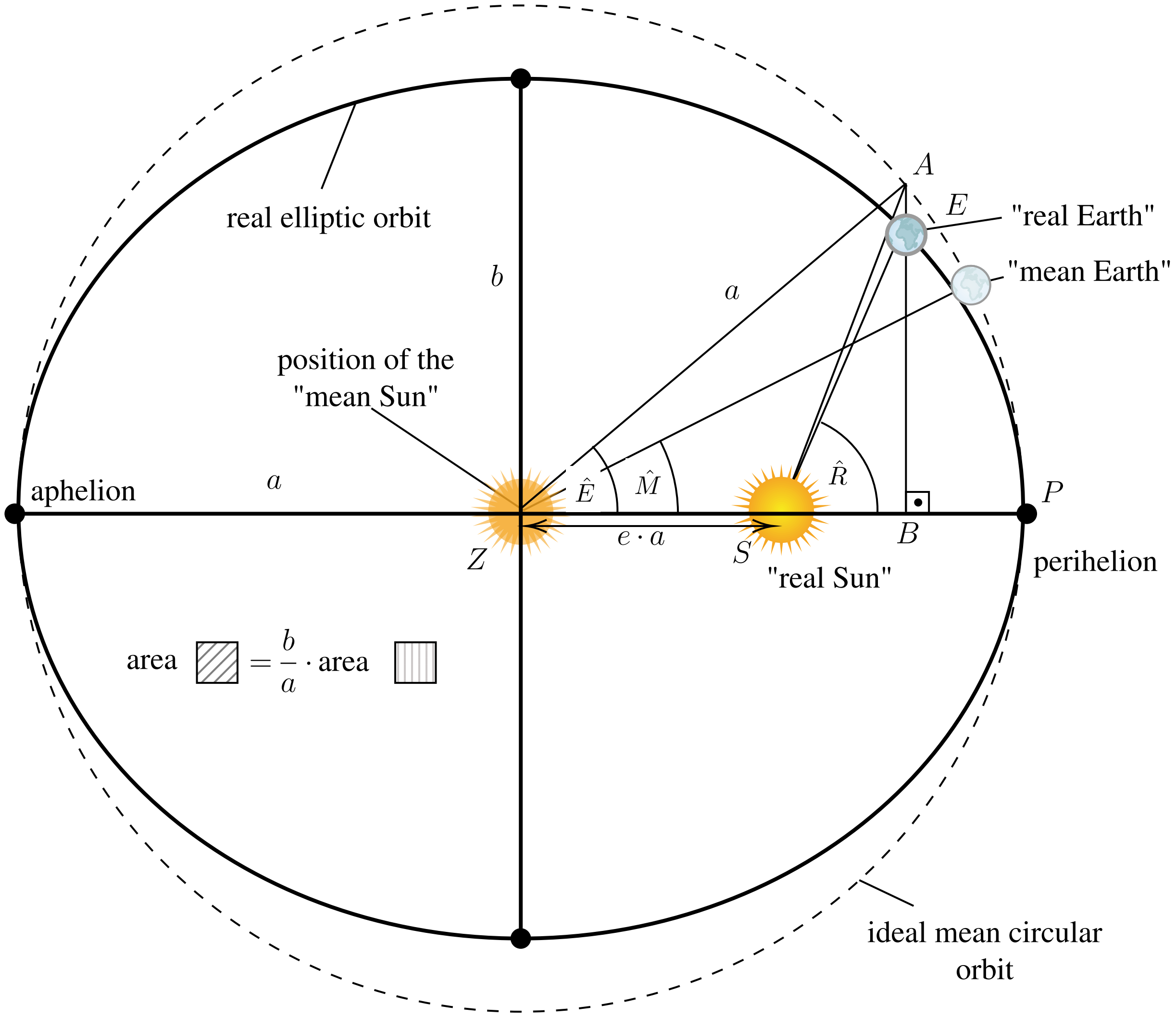

real elliptic orbit
b
a
A
E
"real Earth"
"mean Earth"
position of the
"mean Sun"
aphelion
a
$\hat{E}$
$\hat{M}$
$\hat{R}$
P
$e \cdot a$
Z
S
B
"real Sun"
perihelion
area ▨ $= \frac{b}{a} \cdot$ area ▥
ideal mean circular
orbit

Figure 12.83: **Angles $\hat{R}$, $\hat{M}$ and $\hat{E}$ at a specific time point. The affinity factor between the elliptical orbit and the circle going through the perihelion and the aphelion is given by b/a. As the angular velocity of the "real Earth" and the "mean Earth" are both constant, the ratio between the two hatched areas is the same as between the areas of the circle and the ellipse, viz., $1 : b/a$.**

Let us imagine a "mean Earth" which has also a revolution time T and is running at a constant speed on a circular orbit with the Sun at its centre. This "mean earth" would cover an angle, named "mean anomaly" $\hat{M}$, in the same period of time as the true Earth covers the angle R. In the figure above, M is drawn from the center of the ellipse. The orbit of the "mean Earth" is the circle through the perihelion and the aphelion. The "mean Earth" starts from the perihelion at the same time as the true earth. Since the angular velocity of the "mean Earth" is constant and its revolution lasts one year T, $\hat{M}$ satisfies the simple equation:

$$\hat{M} = 2\pi \frac{t}{T} \qquad (12.1.566)$$

where t is the time span after passage through the perihelion. It is very useful to define a third angle as a link between M and R (see figure above). The perpendicular drawn from the position of the true earth $\hat{E}$ onto the major axis intersects the circle, that is the orbit of the "mean Earth", at the point A. The angle $\widehat{PZA}$ is named "eccentric anomaly" denoted $\hat{E}$ and was introduced by Johannes Kepler. It can be used to calculate the area of elliptic sectors.

We get the following three relations from the figure above (we use also the relations see during our study of conics in the section of Analytical Geometry (page 2764):

$$\tan(\hat{R})=\frac{\overline{EB}}{\overline{SB}}=\frac{\frac{b}{a}\overline{AB}}{\overline{ZB}-\overline{ZS}}=\frac{\frac{b}{a}a\sin(\hat{E})}{a\cos(\hat{E})-ea}=\frac{b}{a}\frac{\sin(\hat{E})}{\cos(\hat{E})-e}=\frac{\sqrt{1-e^2}\sin(\hat{E})}{\cos(\hat{E})-e} \tag{12.1.567}$$

and (at the beginning we use a trigonometric identity - see page 2508 - to be able to use the result we got just before):

$$\begin{aligned}\cos(\hat{R})&=\sqrt{\frac{1}{1+\tan^2(\hat{R})}}=\sqrt{\frac{1}{1+\left(\frac{\sqrt{1-e^2}\sin(\hat{E})}{\cos(\hat{E})-e}\right)^2}}\\&=\sqrt{\frac{(\cos(\hat{E})-e)^2}{(\cos(\hat{E})-e)^2+(1-e^2)\sin^2(\hat{E})}}=\frac{\cos(\hat{E})-e}{1-e\cos(\hat{E})}\end{aligned} \tag{12.1.568}$$

and finally for the third and last one (at the beginning we use a remarkable trigonometric identity - see page 2519 - to be able to use the result we got just before):

$$\begin{aligned}\tan\left(\frac{\hat{R}}{2}\right)&=\sqrt{\frac{1-\cos(\hat{R})}{1+\cos(\hat{R})}}=\sqrt{\frac{1-\frac{\cos(\hat{E})-e}{1-e\cos(\hat{E})}}{1+\frac{\cos(\hat{E})-e}{1-e\cos(\hat{E})}}}=\sqrt{\frac{1-e\cos(\hat{E})-\cos(\hat{E})+e}{1-e\cos(\hat{E})+\cos(\hat{E})-e}}\\&=\sqrt{\frac{(1+e)(1-\cos(\hat{E}))}{(1-e)(1+\cos(\hat{E}))}}=\sqrt{\frac{1+e}{1-e}}\sqrt{\frac{1-\cos(\hat{E})}{1+\cos(\hat{E})}}=\sqrt{\frac{1+e}{1-e}}\tan\left(\frac{\hat{E}}{2}\right)\end{aligned} \tag{12.1.569}$$

To simplify the notations, let us now put:

$$\cos(\alpha):=\sqrt{\frac{1-e}{1+e}}\qquad y:=\frac{\hat{R}}{2}\qquad x:=\frac{\hat{E}}{2} \tag{12.1.570}$$

where as for reminder as $0<e<1$ we have therefore $0<\alpha<\pi/2$.

Therefore with this new notation, we have the relation:

$$\tan\left(\frac{\hat{R}}{2}\right)=\sqrt{\frac{(1+e)(1-\cos(\hat{E}))}{(1-e)(1+\cos(\hat{E}))}}=\sqrt{\frac{1+e}{1-e}}\sqrt{\frac{1-\cos(\hat{E})}{1+\cos(\hat{E})}}=\sqrt{\frac{1+e}{1-e}}\tan\left(\frac{\hat{E}}{2}\right) \tag{12.1.571}$$

that can be rewritten:

$$\tan(y)=\frac{\tan(x)}{\cos(\alpha)}\Leftrightarrow y=\arctan\left(\frac{\tan(x)}{\cos(\alpha)}\right) \tag{12.1.572}$$

The differentiation of this last equality gives us (see section Differential and Integral Calculus page 1475):

$$\begin{aligned}\frac{\mathrm{d}y}{\mathrm{d}x}&=\frac{1}{\left(\frac{\tan(x)}{\cos(\alpha)}\right)^2+1}\cdot\frac{\mathrm{d}}{\mathrm{d}x}\left[\frac{\tan(x)}{\cos(\alpha)}\right]=\frac{\frac{1}{\cos(\alpha)}\cdot\frac{\mathrm{d}}{\mathrm{d}x}[\tan(x)]}{\frac{\tan^2(x)}{\cos^2(\alpha)}+1}\\&=\frac{\frac{1}{\cos(\alpha)}\cdot\frac{1}{\cos^2(x)}}{\frac{(\sin(x)/\cos(x))^2}{\cos^2(\alpha)}+\cos^2(x)}=\frac{\frac{1}{\cos(\alpha)}}{\frac{\sin^2(x)}{\cos^2(\alpha)}+\cos^2(x)}=\frac{\cos(\alpha)}{\sin^2(x)+\cos^2(\alpha)\cos^2(x)}\\&=\frac{\cos(\alpha)}{\sin^2(x)+(1-\sin^2(\alpha))\cos^2(x)}\\&=\frac{\cos(\alpha)}{1-\sin^2(\alpha)\cos^2(x)}=\frac{\cos(\alpha)}{1-\sin^2(\alpha)\frac{(1+\cos(2x))}{2}}\\&=\frac{2\cos(\alpha)}{1+\cos^2(\alpha)-\sin^2(\alpha)\cos(2x)}\end{aligned}\tag{12.1.573}$$

Since:

$$f(x):=\frac{\mathrm{d}y}{\mathrm{d}x}=\frac{\cos(\alpha)}{1-\sin^2(\alpha)\cos^2(x)}\tag{12.1.574}$$

it can be expanded into a Fourier series (see section Sequences and Series page 1727):

$$\begin{aligned}f(x)&=\sum_{n=0}^{+\infty}(a_n\cos(nx)+b_n\sin(nx))\\&=a_0+a_1\cos(x)+b_1\sin(x)+\ldots+a_n\cos(nx)+b_n\sin(nx)+\ldots\end{aligned}\tag{12.1.575}$$

with typically in our special case here:

$$a_0=\frac{1}{2\pi}\int\limits_0^{2\pi}f(x)\mathrm{d}x=\frac{1}{2\pi}\int\limits_0^{2\pi}\frac{\mathrm{d}y}{\mathrm{d}x}\mathrm{d}x=\frac{1}{2\pi}\int\limits_0^{2\pi}\mathrm{d}y=1\tag{12.1.576}$$

and:

$$a_n=\frac{1}{\pi}\int\limits_0^{2\pi}f(x)\cos(nx)\mathrm{d}x\qquad b_n=\frac{1}{\pi}\int\limits_0^{2\pi}f(x)\sin(nx)\mathrm{d}x\text{ for }n>0\tag{12.1.577}$$

Hence:

$$a_n=\frac{1}{\pi}\int\limits_0^{2\pi}f(x)\cos(nx)\mathrm{d}x=\frac{1}{\pi}\int\limits_0^{2\pi}\left(\frac{2\cos(\alpha)}{1+\cos^2(\alpha)-\sin^2(\alpha)\cos(2x)}\right)\cos(nx)\mathrm{d}x\tag{12.1.578}$$

Replacing $\cos(x)$ and $\cos(2x)$ by the complex exponential function - i.e. the Euler formulas - yields ((see section Numbers page 251)):

$$\cos(\varphi)=\frac{e^{\mathrm{i}\varphi}+e^{-\mathrm{i}\varphi}}{2}\tag{12.1.579}$$

Hence:

$$a_n=\frac{1}{\pi}\int\limits_0^{2\pi}\frac{\cos(\alpha)\left(e^{\mathrm{i}nx}+e^{-\mathrm{i}nx}\right)}{1+\cos^2(\alpha)-\sin^2(\alpha)\left(\frac{e^{2\mathrm{i}x}+e^{-2\mathrm{i}x}}{2}\right)}\mathrm{d}x\tag{12.1.580}$$

Let us put:

$$z = e^{2\mathrm{i}x} \Rightarrow \frac{\mathrm{d}z}{\mathrm{d}x} = 2\mathrm{i}e^{2\mathrm{i}x} = 2\mathrm{i}z \Rightarrow \mathrm{d}x = -\frac{\mathrm{i}}{2z}\mathrm{d}z \tag{12.1.581}$$

Therefore:

$$\begin{aligned} a_n &= \frac{\mathrm{i}}{\pi}\int_0^{2\pi} \frac{\cos(\alpha)\left(z^{n/2}+z^{-n/2}\right)}{-2z\left[1+\cos^2(\alpha)-\sin^2(\alpha)\left(\frac{z+z^{-1}}{2}\right)\right]}\mathrm{d}z \\ &= \frac{\mathrm{i}}{\pi}\int_0^{2\pi} \frac{\cos(\alpha)\left(z^{n/2}+z^{-n/2}\right)}{-2z\left[1+\cos^2(\alpha)-\sin^2(\alpha)\left(\frac{z^2+1}{2z}\right)\right]}\mathrm{d}z \\ &= \frac{\mathrm{i}}{\pi}\int_0^{2\pi} \frac{\cos(\alpha)\left(z^{n/2}+z^{-n/2}\right)}{-2z(1+\cos^2(\alpha))+\sin^2(\alpha)(z^2+1)}\mathrm{d}z \\ &= \frac{\mathrm{i}}{\pi}\oint \frac{\cos(\alpha)\left(z^{n/2}+z^{-n/2}\right)}{\sin^2(\alpha)\left(z^2-2\frac{(1+\cos^2\alpha)}{\sin^2\alpha}z+1\right)}\mathrm{d}z \end{aligned} \tag{12.1.582}$$

Because of the change of variable and the exponent of the exponential having a factor 2, the contour of integration is twice the unit circle!

Now we use complex analysis. The denominator has the roots:

$$r_1 = \frac{2-\sin(\alpha)^2+2\sqrt{1-\sin(\alpha)^2}}{\sin(\alpha)^2}, r_2 = \frac{2-\sin(\alpha)^2-2\sqrt{1-\sin(\alpha)^2}}{\sin(\alpha)^2} \tag{12.1.583}$$

That can be simplified:

$$r_1 = \frac{2-\sin(\alpha)^2+2|\cos(\alpha)|}{\sin(\alpha)^2}, r_2 = \frac{2-\sin(\alpha)^2-2|\cos(\alpha)|}{\sin(\alpha)^2} \tag{12.1.584}$$

And as we have $2-\sin^2(\alpha) = 1+1-\sin^2(\alpha) = 1+\cos^2(\alpha)$, we get:

$$r_1 = \frac{1+\cos^2(\alpha)+2|\cos(\alpha)|}{\sin(\alpha)^2}, r_2 = \frac{2-\sin(\alpha)^2-2|\cos(\alpha)|}{\sin(\alpha)^2} \tag{12.1.585}$$

For all $\alpha \in \mathbb{R}$, $\cos(\alpha)$ is either negative or non-negative. If it is negative, $|\cos(\alpha)| = -\cos(\alpha)$ and so:

$$\frac{1+\cos^2(\alpha)+2|\cos(\alpha)|}{\sin^2(\alpha)} = \frac{1+\cos^2(\alpha)-2\cos(\alpha)}{\sin^2(\alpha)} = \left(\frac{1-\cos(\alpha)}{\sin(\alpha)}\right)^2 = \tan^2\left(\frac{\alpha}{2}\right) \tag{12.1.586}$$

If, instead, it is non-negative, then $|\cos(\alpha)| = \cos(\alpha)$ and so:

$$\frac{1+\cos^2(\alpha)+2|\cos(\alpha)|}{\sin^2(\alpha)} = \frac{1+\cos^2(\alpha)+2\cos(\alpha)}{\sin^2(\alpha)} = \left(\frac{1+\cos(\alpha)}{\sin(\alpha)}\right)^2 = \cot^2\left(\frac{\alpha}{2}\right) \tag{12.1.587}$$

This is one of the two roots. The other root is going to just be the other case. There is thus never a difference; you will always get both $\tan^2(\alpha/2)$ and $\cot^2(\alpha/2)$ as the two roots, for all possible $\alpha \in \mathbb{R}$. It is just a matter of which one we pick first.

Therefore the previous curvilinear integral can be rewritten as:

$$\begin{aligned}a_n&=\frac{\mathrm{i}}{\pi}\oint\frac{\cos(\alpha)}{\sin^2(\alpha)}\frac{\left(z^{n/2}+z^{-n/2}\right)}{\left(z-\tan^2\left(\frac{\alpha}{2}\right)\right)\left(z-\cot^2\left(\frac{\alpha}{2}\right)\right)}\mathrm{d}z\\&=\frac{\mathrm{i}}{\pi}\oint\frac{\cos(\alpha)}{\sin^2(\alpha)}\frac{\left(z^{n/2}+z^{-n/2}\right)}{\tan^2\left(\frac{\alpha}{2}\right)-\cot^2\left(\frac{\alpha}{2}\right)}\left[\frac{1}{z-\tan^2\left(\frac{\alpha}{2}\right)}-\frac{1}{z-\cot^2\left(\frac{\alpha}{2}\right)}\right]\mathrm{d}z\end{aligned} \tag{12.1.588}$$

Now remember that we have proven in the section of Sequences and Series (see page 1679) that:

$$\frac{1}{1-x}=\sum_{n=0}^{+\infty}x^n \text{ for } |x|<1 \tag{12.1.589}$$

So now let us rewrite:

$$\frac{1}{z-\tan^2\left(\frac{\alpha}{2}\right)}=\frac{1}{z}\frac{1}{1-\dfrac{\tan^2\left(\frac{\alpha}{2}\right)}{z}} \tag{12.1.590}$$

Now remember that we have defined earlier above:

$$\cos(\alpha):=\sqrt{\frac{1-e}{1+e}} \tag{12.1.591}$$

and as the modulus of z is always equal to 1 we satisfy the criteria of convergent series and we are authorized therefore to write:

$$\frac{1}{z-\tan^2\left(\frac{\alpha}{2}\right)}=\frac{1}{z}\sum_{k=0}^{+\infty}\left(\frac{\tan^2\left(\frac{\alpha}{2}\right)}{z}\right)^k \tag{12.1.592}$$

As $\cot(\alpha/2)=1/\tan(\alpha/2)$, we have:

$$\begin{aligned}-\frac{1}{z-\cot^2\left(\frac{\alpha}{2}\right)}&=-\frac{1}{z-\frac{1}{\tan^2\left(\frac{\alpha}{2}\right)}}=-\tan^2\left(\frac{\alpha}{2}\right)\frac{1}{z\tan^2\left(\frac{\alpha}{2}\right)-1}=\tan^2\left(\frac{\alpha}{2}\right)\frac{1}{1-z\tan^2\left(\frac{\alpha}{2}\right)}\\&=\tan^2\left(\frac{\alpha}{2}\right)\sum_{k=0}^{+\infty}\left(z\tan^2\left(\frac{\alpha}{2}\right)\right)^k\end{aligned} \tag{12.1.593}$$

The integrand:

$$a_n=\frac{\mathrm{i}}{\pi}\oint\frac{\cos(\alpha)}{\sin^2(\alpha)}\frac{\left(z^{n/2}+z^{-n/2}\right)}{\tan^2\left(\frac{\alpha}{2}\right)-\cot^2\left(\frac{\alpha}{2}\right)}\left[\frac{1}{z-\tan^2\left(\frac{\alpha}{2}\right)}-\frac{1}{z-\cot^2\left(\frac{\alpha}{2}\right)}\right]\mathrm{d}z \tag{12.1.594}$$

Then becomes (we also used a trigonometric identity proven at page 2519):

$$a_n=-\frac{\mathrm{i}}{4\pi}\oint\left(z^{n/2}+z^{-n/2}\right)\left\{\frac{1}{z}\sum_{k=0}^{+\infty}\left[\frac{\tan^2\left(\frac{\alpha}{2}\right)}{z}\right]^k+\tan^2\left(\frac{\alpha}{2}\right)\sum_{k=0}^{+\infty}\left[z\tan^2\left(\frac{\alpha}{2}\right)\right]^k\right\} \tag{12.1.595}$$

Now, let us notice that:

$$\oint z^p \mathrm{d}z = \int\limits_{e^0}^{e^{4\mathrm{i}\pi}} z^p \mathrm{d}z = \left.\frac{z^{(p+1)}}{p+1}\right|_{e^0}^{e^{4\mathrm{i}\pi}} = \frac{1}{p+1} - \frac{1}{p+1} \tag{12.1.596}$$

However we have to consider the singularity (division by zero) for the case $p = -1$. In that case:

$$\oint z^p \mathrm{d}z = \int\limits_{e^0}^{e^{4\mathrm{i}\pi}} z^{-1} \mathrm{d}z = \ln(z)|_{e^0}^{e^{4\mathrm{i}\pi}} = 4\pi\mathrm{i} \tag{12.1.597}$$

So finally:

$$\oint z^p \mathrm{d}z = \int\limits_{e^0}^{e^{4\mathrm{i}\pi}} z^p \mathrm{d}z = \left.\frac{z^{(p+1)}}{p+1}\right|_{e^0}^{e^{4\mathrm{i}\pi}} = \begin{cases} 0 & \text{for} \quad p \neq -1 \\ 4\mathrm{i}\pi & \text{for} \quad p = -1 \end{cases} \tag{12.1.598}$$

So that means for the integral that gives a_n we only have to consider the terms where $z^{n/2} = z^{-1}$ as all other will vanished! For that a first minimal condition are the powers in the parenthesis:

$$\left(z^{n/2} + z^{-n/2}\right) \tag{12.1.599}$$

to be integers obviously! Consequently the integral is zero if n is odd, that is, a_n vanishes if n is odd.

How can this be useful to us ? Consequently, the integral is zero if n is odd, that is, a_n vanishes if n is odd. $a_n = 0$ for odd n. If n is even we can only keep the terms with a $1/z$ (all others like $z, z^2, z^3, \ldots$ and $z^{-2}, z^{-3}, z^{-4} \ldots$ will be zero so we don't keep them!), then if we develop by some n we notice quite quickly that it remains the following pattern of even $n > 0$:

$$\begin{aligned} a_n &= -\frac{\mathrm{i}}{4\pi} \oint \frac{1}{z} \left[\tan^n\left(\frac{\alpha}{2}\right) + \tan^2\left(\frac{\alpha}{2}\right)\tan^{-2+n}\left(\frac{\alpha}{2}\right)\right] \\ &= -\frac{\mathrm{i}}{4\pi} \oint \frac{1}{z} \left[\tan^n\left(\frac{\alpha}{2}\right) + \tan^n\left(\frac{\alpha}{2}\right)\right] \\ &= -\frac{\mathrm{i}}{2\pi}\frac{1}{z}\tan^n\left(\frac{\alpha}{2}\right) \end{aligned} \tag{12.1.600}$$

The integration of this term yields:

$$a_n = -\frac{\mathrm{i}}{2\pi}\tan^n\left(\frac{\alpha}{2}\right) \int_{\mathrm{e}^0}^{\mathrm{e}^{4\mathrm{i}\pi}} \frac{1}{z}\mathrm{d}z = -\frac{\mathrm{i}}{2\pi}\tan^n\left(\frac{\alpha}{2}\right) 4\mathrm{i}\pi = 2\tan^n\left(\frac{\alpha}{2}\right) \tag{12.1.601}$$

The coefficients b_n can be obtained similarly. If n is odd, b_n disappears. If n is even, one has to write instead of:

$$a_n = -\frac{\mathrm{i}}{4\pi} \oint \frac{1}{z} \left[\tan^n\left(\frac{\alpha}{2}\right) + \tan^2\left(\frac{\alpha}{2}\right)\tan^{-2+n}\left(\frac{\alpha}{2}\right)\right] = 2\tan^n\left(\frac{\alpha}{2}\right) \tag{12.1.602}$$

the following:

$$b_n = -\frac{1}{4\pi} \oint \frac{1}{z} \left[\tan^n\left(\frac{\alpha}{2}\right) - \tan^n\left(\frac{\alpha}{2}\right)\right] = 0 \tag{12.1.603}$$

Therefore:

$$\begin{aligned}f(x)=\frac{\mathrm{d}y}{\mathrm{d}x}&=\frac{\cos(\alpha)}{1-\sin^2(\alpha)\cos^2(x)}\\&=\sum_{n=0}^{+\infty}(a_n\cos(nx)+b_n\sin(nx))\\&=a_0+a_1\cos(x)+b_1\sin(x)+\ldots+a_n\cos(nx)+b_n\sin(nx)+\ldots\\&=1+2\sum_{k=1}^{+\infty}\tan^{2k}\left(\frac{\alpha}{2}\right)\cos(2kx)\end{aligned} \tag{12.1.604}$$

Hence:

$$y=x+\sum_{k=1}^{+\infty}\tan^{2k}\left(\frac{\alpha}{2}\right)\frac{\sin(2kx)}{k}+c^{te} \tag{12.1.605}$$

Now remember that at the beginning we have constructed the following relation:

$$\tan\left(\frac{\hat{R}}{2}\right)=\sqrt{\frac{1+e}{1-e}}\tan\left(\frac{\hat{E}}{2}\right) \tag{12.1.606}$$

and that we had defined:

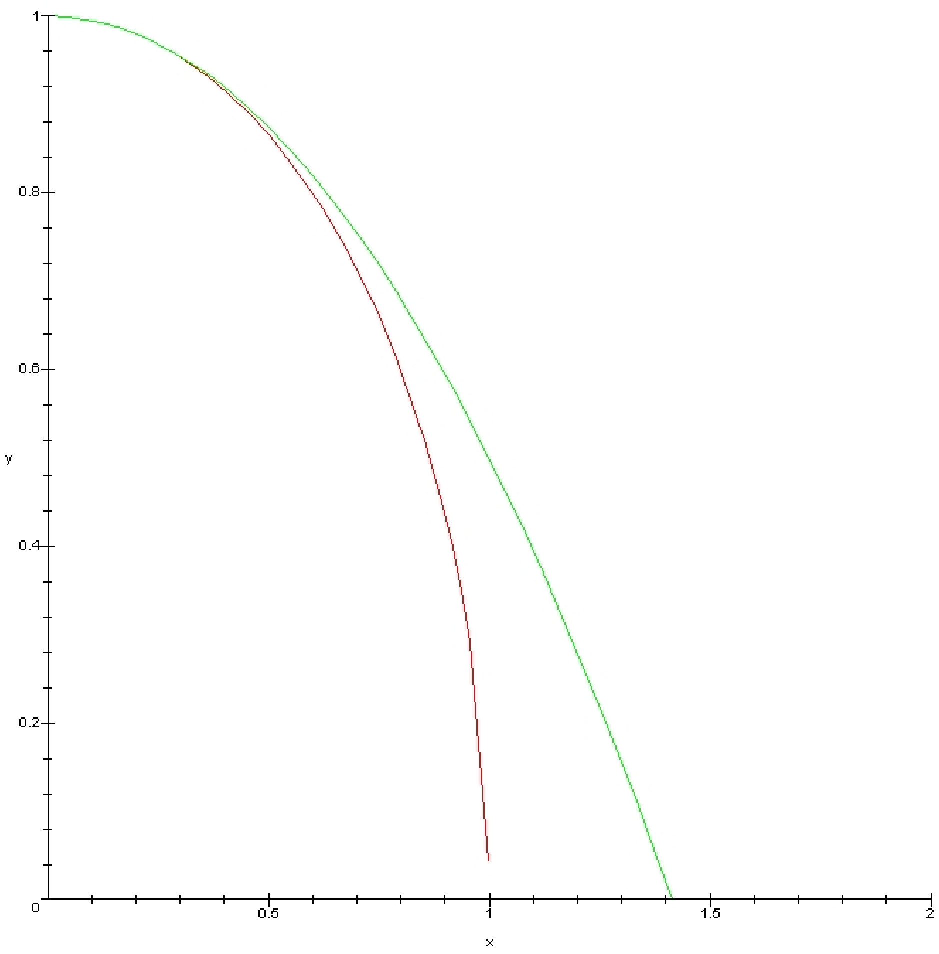
$$\cos(\alpha):=\sqrt{\frac{1-e}{1+e}} \tag{12.1.607}$$

We can rewrite it using trigonometric identities (see page 2519) as:

$$\begin{aligned}\tan^2\left(\frac{\alpha}{2}\right)&=\frac{1-\cos(\alpha)}{1+\cos(\alpha)}=\frac{1-\sqrt{\frac{1-e}{1+e}}}{1+\sqrt{\frac{1-e}{1+e}}}=\frac{1-\sqrt{\frac{1-e}{1+e}}}{1+\sqrt{\frac{1-e}{1+e}}}\frac{1-\sqrt{\frac{1-e}{1+e}}}{1-\sqrt{\frac{1-e}{1+e}}}\\&=\frac{1-2\sqrt{\frac{1-e}{1+e}}+\frac{1-e}{1+e}}{1-\frac{1-e}{1+e}}=\frac{1+e+1-e-2\sqrt{1-e^2}}{1+e-1+e}=\frac{1-\sqrt{1-e^2}}{e}\end{aligned} \tag{12.1.608}$$

If we do a Maclaurin development to the fourth term of the square root (see section Sequences and Series page 1696), even if that's not so good as we can check it with Maple™ 4.00b (the red is the exact value, the green the approximation!):

```
>plot(sqrt(1-x2̂),convert(taylor(sqrt(1-x2̂),x=0,4),polynom),x=0...2,y=0..1);
```



We can write:

$$\tan^2\left(\frac{\alpha}{2}\right)=\frac{1-\sqrt{1-e^2}}{e}\cong\frac{1-\left(1-\frac{e^2}{2}-\frac{e^4}{8}\right)}{e}=\frac{e}{2}+\frac{e^3}{8}\qquad(12.1.609)$$

Now putting all together, we have that:

$$y=x+\sum_{k=1}^{+\infty}\tan^{2k}\left(\frac{\alpha}{2}\right)\frac{\sin(2kx)}{k}+c^{te}\qquad(12.1.610)$$

with as initial condition that $R(E=0)=0$ (the constant is therefore equal to zero), and remembering all change of variables done at the beginning, that:

$$\hat{R}\cong\hat{E}+2\sum_{k=1}^{+\infty}\left(\frac{e}{2}+\frac{e^3}{8}\right)^k\frac{\sin(k\hat{E})}{k}\cong\hat{E}+\left(e+\frac{e^3}{4}\right)\sin(E)+\frac{e^2}{4}\sin(2\hat{E})+\frac{e^3}{12}\sin(3\hat{E})\qquad(12.1.611)$$

So we have here an approximation of already something that is a quite significant approximation (that makes a lot of approximations...).

The angle $\hat{E}$ is so useful because the area swept out by the radius vector from the Sun to the Earth in the time span t can be calculated easily by using the affinity between the circle and the ellipse (see the first figure above):

$$\begin{aligned} F &= \text{elliptic sector } SEP = \frac{b}{a}(\text{sector } ZAP - \text{triangle } ZAS) \\ &= \frac{b}{a}\left(\frac{\hat{E}}{2\pi}\text{area}_{\text{ circle}} - \frac{\overline{ZS}\cdot\overline{AB}}{2}\right) = \frac{b}{a}\left(\frac{\hat{E}}{2}a^2 - \frac{aea\sin(\hat{E})}{2}\right) \\ &= \frac{b}{a}\left(\frac{\hat{E}}{2}a^2 - a^2\frac{e\sin(\hat{E})}{2}\right) = \frac{ab}{2}(\hat{E} - e\sin(\hat{E})) \end{aligned} \tag{12.1.612}$$

On the other hand, the law of areas implies (see page 5234):

$$F = \frac{t}{T}\pi ab \tag{12.1.613}$$

or with:

$$\hat{M} = \frac{t}{T}2\pi \tag{12.1.614}$$

on gets:

$$F = \frac{Mab}{2} \tag{12.1.615}$$

The comparison with:

$$F = \frac{ab}{2}(\hat{E} - e\sin(\hat{E})) \tag{12.1.616}$$

yields:

$$\hat{E} = \hat{M} + e\sin(\hat{E}) \tag{12.1.617}$$

The relation above is named the "Kepler's equation". It is not possible to solve this equation for $\hat{E}$ in closed form. Therefore, the position of the earth - given by $\hat{E}$ - can only be approximated as a function of time. The Lagrange inversion theorem (see section Sequences and Series page page 1725) allows us to expand any function in $\hat{E}$ into a series in e and $\hat{E}$. Its basis is an equation in the form:

$$z = y + xf(z) \tag{12.1.618}$$

can be expanded for small α as a series expressed by:

$$z = y + \sum_{k=1}^{+\infty}\frac{x^k}{k!}\frac{\partial^{k-1}}{\partial y^{k-1}}f^k(y) \tag{12.1.619}$$

In the case of Kepler's equation, we set:

$$z = \hat{E}, \quad y = \hat{M}, \quad x = e, \quad f(\cdot) = \sin(\cdot), \quad f(z) = \sin(\hat{E}), \quad f(y) = \sin(\hat{M}) \tag{12.1.620}$$

Therefore we see that as in our case $x = e \cong 1$ this approximation by the Lagrange inversion theorem is another brutal approximation. $\hat{E}$ turns out therefore to be (using some trigonometric identities at the end as proven in the section of Trigonometry at page 2519):

$$\begin{aligned}\hat{E}&=\hat{M}+e\sin(\hat{M})+\frac{e^2}{2}\frac{\partial}{\partial\hat{M}}\sin^2(\hat{M})+\frac{e^3}{6}\frac{\partial^2}{\partial\hat{M}^2}\sin^3(\hat{M})+\ldots\\&=\hat{M}+e\sin(\hat{M})+\frac{e^2}{2}2\sin(\hat{M})\cos(\hat{M})+\frac{e^3}{6}\left[6\sin(\hat{M})\cos^2(\hat{M})-3\sin^3(\hat{M})\right]+\ldots\\&=\hat{M}+e\sin(\hat{M})+\frac{e^2}{2}2\sin(\hat{M})\cos(\hat{M})+\frac{e^3}{6}\left[\frac{6}{4}\sin(x)+\frac{6}{4}\sin(3x)-\frac{9}{4}\sin(x)-\frac{3}{4}\sin(3x)\right]+\ldots\\&=\hat{M}+e\sin(\hat{M})+\frac{e^2}{2}2\sin(\hat{M})\cos(\hat{M})+\frac{e^3}{6}\left[-\frac{3}{4}\sin(x)+\frac{3}{4}\sin(3x)\right]+\ldots\\&=\hat{M}+e\sin(\hat{M})+\frac{e^2}{2}2\sin(\hat{M})\cos(\hat{M})+e^3\left[-\frac{1}{8}\sin(x)+\frac{1}{8}\sin(3x)\right]+\ldots\\&=\hat{M}+e\sin(\hat{M})+\frac{e^2}{2}\sin(2\hat{M})+\frac{e^3}{8}\left(\sin(3\hat{M})-\sin(\hat{M})\right)+\ldots\end{aligned} \tag{12.1.621}$$

where we used the following trigonometric identities (we repeat them here as request by a reader):

$$\sin(x)\cos(y)=\frac{1}{2}(\sin(x-y)+\sin(x+y))$$

$$\begin{aligned}\sin(x)\cos^2(x)&=\sin(x)\left[\frac{1}{2}(1+\cos(2x))\right]=\frac{1}{2}\sin(x)+\frac{1}{2}\sin(x)\cos(2x)\\&=\frac{1}{2}\sin(x)+\frac{1}{2}\left[\frac{1}{2}(\sin(x-2x)+\sin(x+2x))\right]\\&=\frac{1}{2}\sin(x)+\frac{1}{2}\left[\frac{1}{2}(\sin(-x)+\sin(3x))\right]\\&=\frac{1}{2}\sin(x)+\frac{1}{4}\left[-\sin(x)+\sin(3x)\right]\\&=\frac{1}{4}\sin(x)+\frac{1}{4}\sin(3x)\end{aligned} \tag{12.1.622}$$

$$\begin{aligned}\sin^3(x)&=\sin(x)\cdot\sin^2(x)=\sin(x)\left(\frac{1}{2}-\frac{1}{2}\cos(2x)\right)\\&=\frac{1}{2}\sin(x)-\frac{1}{2}\sin(x)\cos(2x)\\&=\frac{1}{2}\sin(x)-\frac{1}{2}\left[\frac{1}{2}\sin(3x)+\frac{1}{2}\sin(-x)\right]\\&=\frac{1}{2}\sin(x)-\frac{1}{4}\sin(3x)-\frac{1}{4}\sin(-x)\\&=\frac{1}{2}\sin(x)-\frac{1}{4}\sin(3x)+\frac{1}{4}\sin(x)\\&=\frac{3}{4}\sin(x)-\frac{1}{4}\sin(3x)\end{aligned}$$

To be able to express $\hat{R}$ as a function of $\hat{M}$ in the relation we got earlier above:

$$\hat{R}\cong\hat{E}+\left(e+\frac{e^3}{4}\right)\sin(\hat{E})+\frac{e^2}{4}\sin(2\hat{E})+\frac{e^3}{12}\sin(3\hat{E}) \tag{12.1.623}$$

we also have to expand the sine functions of $E, 2E$ and $3E$ as far as needed:

$$\sin(\hat{E})=\frac{\hat{E}-\hat{M}}{e}=\sin(\hat{M})+\frac{e}{2}\sin(2\hat{M})+\frac{e^2}{8}(3\sin(3\hat{M})-\sin(\hat{M}))+\ldots \tag{12.1.624}$$

As:

$$\hat{E}=\hat{M}+e\sin(\hat{E}) \tag{12.1.625}$$

Thus:

$$\sin(2\hat{E})=\sin\left(2\hat{M}+2e\sin(\hat{E})\right) \tag{12.1.626}$$

Now we Taylor expand the sine according to (see section Sequences and Series page 1703):

$$\begin{aligned}
f_3(x)&\cong\sin(x)\\
&\cong\sum_{k=0}^{3}\frac{1}{k!}\frac{\mathrm{d}^kP(x)}{\mathrm{d}x^k}x^k=\frac{1}{0!}\frac{\mathrm{d}^0P(x)}{\mathrm{d}x^0}x^0+\frac{1}{1!}\frac{\mathrm{d}^1P(x)}{\mathrm{d}x^1}x^1+\frac{1}{2!}\frac{\mathrm{d}^2P(x)}{\mathrm{d}x^2}x^2+\frac{1}{3!}\frac{\mathrm{d}^3P(x)}{\mathrm{d}x^3}x^3\\
&=\frac{1}{0!}\sin(x)+\frac{1}{1!}\frac{\mathrm{d}\sin(x)}{\mathrm{d}x}x^1+\frac{1}{2!}\frac{\mathrm{d}^2\sin(x)}{\mathrm{d}x^2}x^2+\frac{1}{3!}\frac{\mathrm{d}^3\sin(x)}{\mathrm{d}x^3}x^3\\
&=\frac{1}{0!}\sin(x)+\frac{1}{1!}\cos(x)x-\frac{1}{2!}\sin(x)x^2-\frac{1}{3!}\cos(x)x^3\\
&=\sin(x)+\cos(x)x-\frac{1}{2}\sin(x)x^2-\frac{1}{6}\cos(x)x^3
\end{aligned} \tag{12.1.627}$$

but in e around 0. Therefore at the second order:

$$f_2(x)\cong\sin(x)=\sin(x)+\cos(x)x=\sin(2\hat{M})+\cos(2\hat{M})2\sin(\hat{E})\cdot e\cong sin(2\hat{E}) \tag{12.1.628}$$

So to summarize:

$$\begin{aligned}
\sin(\hat{E})&=\frac{\hat{E}-\hat{M}}{e}=\sin(\hat{M})+\frac{e}{2}\sin(2\hat{M})+\frac{e^2}{8}(3\sin(3\hat{M})-\sin(\hat{M}))+\ldots\\
\sin(2\hat{E})&=\sin(2\hat{M})+e\sin(\hat{M})2\cos(2\hat{M})+\ldots\\
&=\sin(2\hat{M})+e(\sin(3\hat{M})-\sin(\hat{M}))+\ldots\\
\sin(3\hat{E})&=\sin(3\hat{M})+\ldots
\end{aligned} \tag{12.1.629}$$

Replacing the functions in $\hat{E}$ of:

$$\hat{R}\cong\hat{E}+\left(e+\frac{e^3}{4}\right)\sin(\hat{E})+\frac{e^2}{4}\sin(2\hat{E})+\frac{e^3}{12}\sin(3\hat{E}) \tag{12.1.630}$$

with these formulas in $\hat{M}$, we get:

$$\begin{aligned}
\hat{R}&\cong\hat{E}+\left(e+\frac{e^3}{4}\right)\sin(\hat{E})+\frac{e^2}{4}\sin(2\hat{E})+\frac{e^3}{12}\sin(3\hat{E})\\
&\cong\hat{M}+\left(2e+\frac{e^3}{4}\right)\sin(\hat{E})+\frac{e^2}{4}\sin(2\hat{E})+\frac{e^3}{12}\sin(3\hat{E})\\
&\cong\hat{M}+2e\sin(\hat{M})+\frac{5}{4}e^2\sin(2\hat{M})+e^3\left(-\frac{1}{4}\sin(\hat{M})+\frac{13}{12}\sin(3\hat{M})\right)+e^4\ldots
\end{aligned} \tag{12.1.631}$$

In this series, we have only taken into account terms up to third power of e (this approximation is already very accurate, for e is about 0.0167). The angular velocity of the Earth can now be obtained as the first derivative of $\hat{R}$ with respect to time:

$$\begin{aligned}\omega(t) &= \frac{\mathrm{d}\hat{R}}{\mathrm{d}t} = \frac{\mathrm{d}\hat{R}}{\mathrm{d}\hat{M}}\frac{\mathrm{d}\hat{M}}{\mathrm{d}t} = \frac{2\pi}{T}\frac{\mathrm{d}\hat{R}}{\mathrm{d}\hat{M}} \\ &\cong \frac{2\pi}{T}\left[1+2e\cos(\hat{M})+\frac{5}{2}e^2\cos(2\hat{M})+e^3\left(-\frac{1}{4}\cos(\hat{M})+\frac{13}{4}\cos(3\hat{M})\right)\right]\end{aligned} \tag{12.1.632}$$

This function of the angular velocity of the Earth as function of the angle $\hat{M}$ is plotted in the figure below (with for the Earth actually $e=0.0167$, $T=365.25$) with Maple™ 4.00b (on average it is about $2\pi/365.25$ per day or $360/365.25°/\text{day}=0.986°/\text{day}$ or $0.0172\,[\text{rad}]/\text{day}$). It is mainly determined by the first variable term of the series $2e\cos(\hat{M})$. The factor $2\pi/T$ is the mean angular velocity. The deviations amount to about $\pm 3.5\%(\approx 2e)$:

```
>e:=0.0167;
>T:=365.25;
>plot(2*Pi/T*(1+2*e*cos(M)+5/2*e^2*cos(2*M)
+e^3*(-1/4*cos(M)+13/4*cos(3*M))),M=0..2*Pi);
```

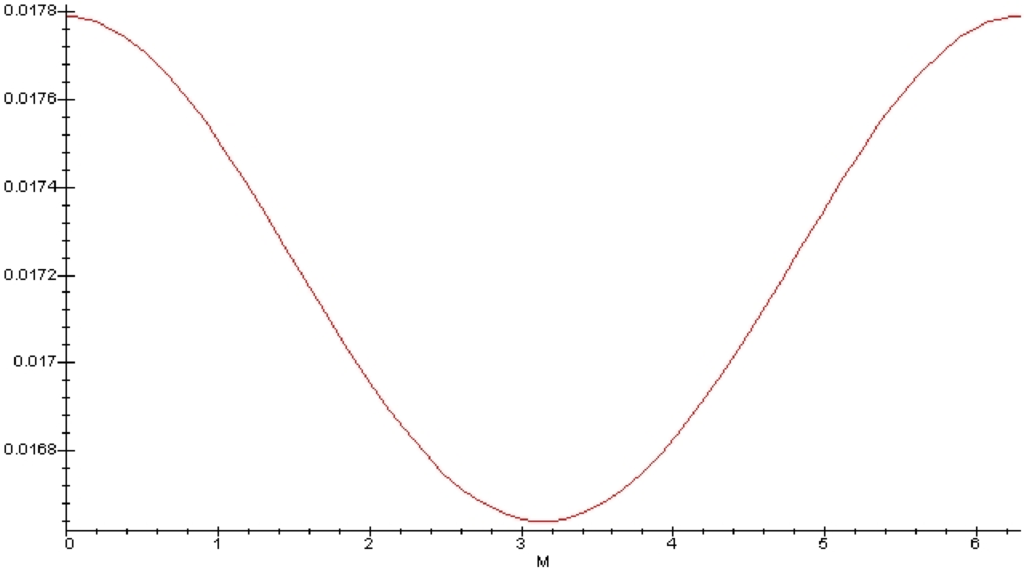


Figure 12.84: Angular velocity of Earth as function of orbits' angle (left perihelion, middle aphelion)

12.1.10.2 Obliquity (inclined plane of the ecliptic) and Sun anomaly

Since the equation of time is to be examined during a year, the Earth can be supposed to always have the same direction in space, i.e., we will not take into account precession and nutation (as their variation during a year is negligible!).

We can describe the direction of the axis of the Earth with two angles. The first is the angle between the axis and the norm of the orbit denoted (ε) and commonly named Earth's obliquity as we already know it. The second is the angle P formed by the major axis of the orbit and the projection of the axis of the Earth onto the plane of the orbit (we don't know if this angle has an official name!):

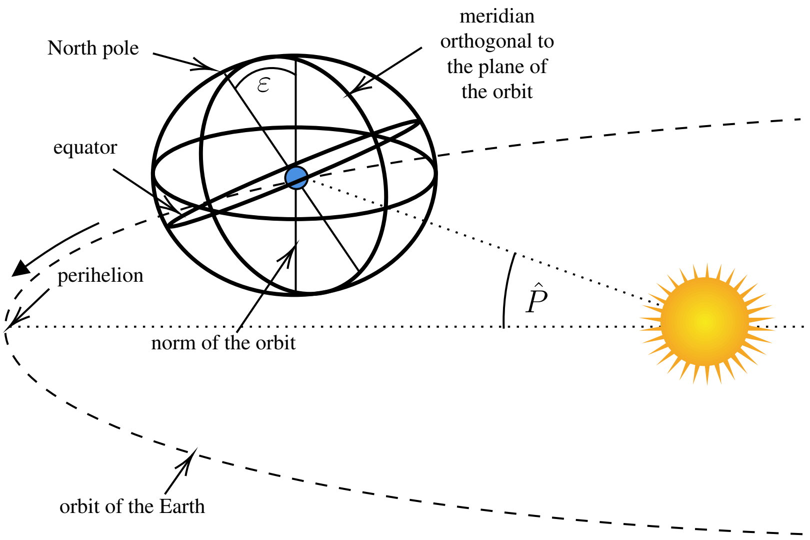


Figure 12.85: The angle $\hat{P}$ and the inclination of the axis of the earth at the winter solstice (not to scale)

At present, ε measures about 23.45°, $\hat{P}$ is about 12.25° ($\hat{P}$ is also the angle which is covered by the Earth between the beginning of winter (21st December) and the arrival of the Earth at perihelion (2nd January) and named the "solar anomaly" or also "perihelion precession angle").

Remark

Remember that the winter solstice and the perihelion are both astronomical events, but they relate to different aspects of Earth's movement. Indeed, the winter solstics marks the shortest day and longest night of the year in a given atmosphere, it's the moment when one of Earth's pole is tilted the farthest away from the Sun. While the perihelion refers to the point in Earth's elliptical orbit when it is closest to the Sun.

The angles which the Sun appears to cover relative to the Earth are equal to those the Earth actually covers relative to the sun. Since the corresponding angles parallel to the equatorial plane are needed for the measurement of time, the ecliptical angles have to be projected onto the equatorial plane. This results in angle widening or shortening. For an infinitesimally small angle the projection factor (the deformation factor) is determined by the angle parameters φ and ε visible in the figure below:

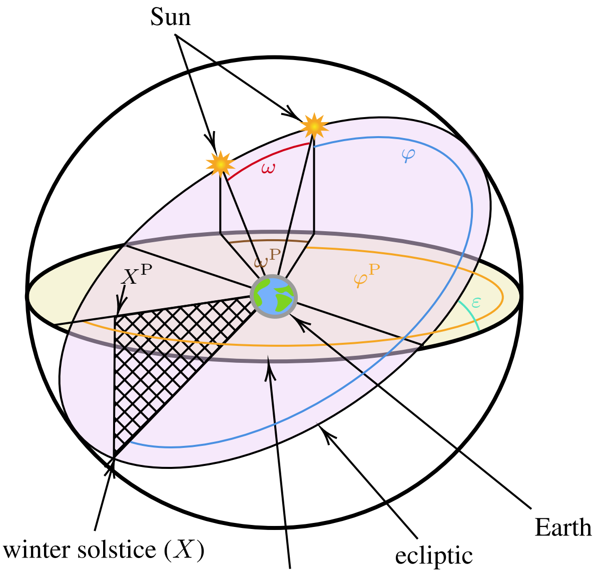


Figure 12.86: **Geocentric view of the projection where in a short time interval, the Sun has covered the angle ω (ω^p is the orthogonal projection of ω)**

which represents the geocentric view. φ is cleverly measured from the winter solstice to force the Sun anomaly to appear in our final result!

We get the following relation from the figure above if the radius are quite huge:

$$\cos(\varepsilon) \cong \frac{\tan(\varphi)}{\tan(\varphi^{\mathrm{P}})} \tag{12.1.633}$$

Hence:

$$\tan(\varphi) \cong \cos(\varepsilon)\tan(\varphi^{\mathrm{P}}) \tag{12.1.634}$$

By extension:

$$\tan(\varphi+\omega) \cong \cos(\varepsilon)\tan\left(\varphi^{\mathrm{P}}+\omega^{\mathrm{P}}\right) \tag{12.1.635}$$

From the two previous relations we have:

$$\varphi^{\mathrm{P}} \cong \arctan\left(\frac{\tan(\varphi)}{\cos(\varepsilon)}\right) \tag{12.1.636}$$

and:

$$\omega^{\mathrm{P}} \cong \arctan\left(\frac{\tan(\varphi+\omega)}{\cos(\varepsilon)}\right) - \varphi^{\mathrm{P}} \tag{12.1.637}$$

Hence combining both:

$$\omega^{\mathrm{P}} \cong \arctan\left(\frac{\tan(\varphi+\omega)}{\cos(\varepsilon)}\right) - \arctan\left(\frac{\tan(\varphi)}{\cos(\varepsilon)}\right) \tag{12.1.638}$$

The projection factor F_p for a small angle can be calculated as a function of φ and ε :

$$F_p(\varphi) \cong \lim_{\omega\rightarrow 0}\frac{\omega^{\mathrm{p}}}{\omega} = \lim_{\omega\rightarrow 0}\frac{\arctan\left(\frac{\tan(\varphi+\omega)}{\cos(\varepsilon)}\right) - \arctan\left(\frac{\tan(\varphi)}{\cos(\varepsilon)}\right)}{\omega} = \frac{\mathrm{d}\omega^{\mathrm{p}}}{\mathrm{d}\omega} \tag{12.1.639}$$

Here we recognize the same relation as in 12.1.572 (but with $alpha$ instead of ε). And we have proved earlier above at 12.1.604 that the latter has an approximation which converges very quickly given by:

$$\begin{aligned} F_p(\varphi) &\cong \frac{\mathrm{d}\omega^{\mathrm{p}}}{\mathrm{d}\omega} \cong 1+2\sum_{k=1}^{+\infty}\tan^{2k}\left(\frac{\alpha}{2}\right)\cos(2kx) \\ &= 1+2\tan^2\left(\frac{\varepsilon}{2}\right)\cos(2\varphi)+2\tan^4\left(\frac{\varepsilon}{2}\right)\cos(4\varphi)+2\tan^6\left(\frac{\varepsilon}{2}\right)\cos(6\varphi)+\ldots \end{aligned} \tag{12.1.640}$$

The projection factor F_p is plotted as a function of time:

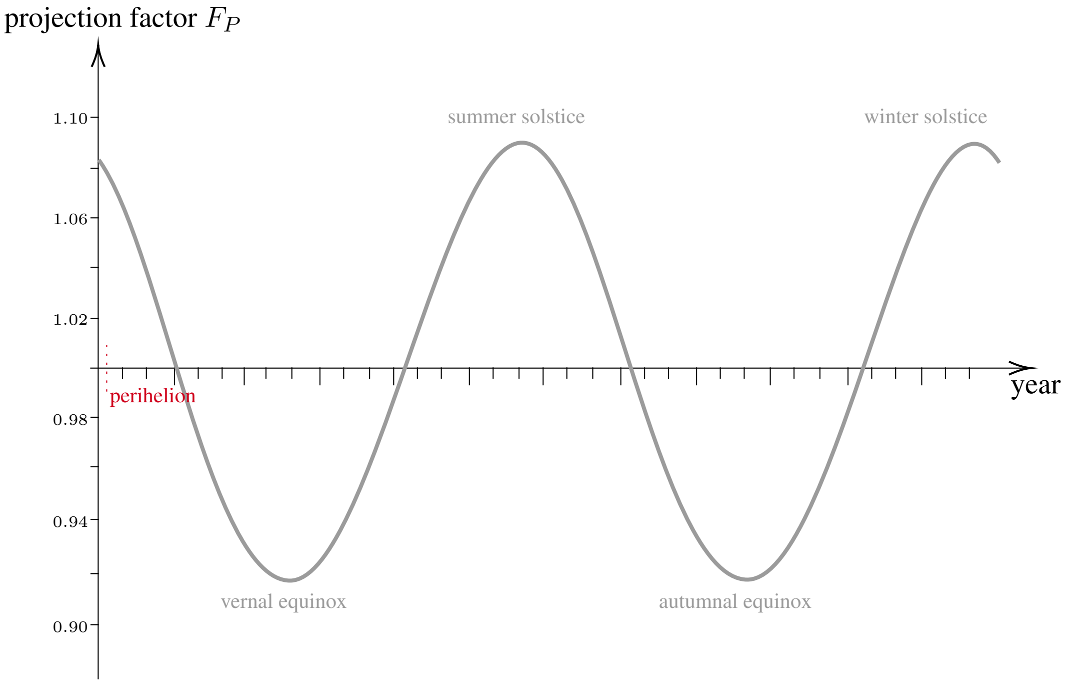


Figure 12.87: **The projection factor as a function of time (the maxima are at the solstices the minima at the equinoxes)**

This graph resembles a cosine function since it is mainly determined by the first variable term of the approximation. The amplitude of the variation $(2\tan^2(\varepsilon/2))$ amounts to $8^1/_2\%$.

The extremes are situated at the beginning of the seasons. At the summer and winter solstice, the Sun reaches, respectively, its highest and lowest position. Here, an ecliptic angle is stretched maximally. At the vernal and autumnal equinox, the Sun stands vertically above the equator. Here, ecliptic angles are shortened maximally.

12.1.10.3 Calculation of the Equation of Time

In order to calculate the equation of time we need the projection of the true anomaly R^{P} as a function of time.

So we write:

$$\hat{R}^{\mathrm{P}}(t)=\int \mathrm{d}R^{\mathrm{P}}=\int f(\varphi)\mathrm{d}R \tag{12.1.641}$$

As by construction we have $\varphi=\hat{R}+\hat{P}$ that is to say the true anomaly plus the Sun anomaly, therefore the above can be rewritten:

$$\begin{aligned}\hat{R}^{\mathrm{P}}(t)=\int\limits_0^t \mathrm{d}\hat{R}^{\mathrm{P}}=\int\limits_0^t f(\hat{R}+\hat{P})\mathrm{d}\hat{R}\cong\int\left[1+2\sum_{i=1}^{+\infty}\tan^{2k}\left(\frac{\varepsilon}{2}\right)\cos(2k(\hat{R}+\hat{P}))\right]\mathrm{d}\hat{R}\\ =\left\{\hat{R}(t)+\sum_{i=1}^{+\infty}\frac{1}{k}\tan^{2k}\left(\frac{\varepsilon}{2}\right)\sin(2k(\hat{R}(t)+\hat{P}))\right\}\\ -\left\{\hat{R}(0)+\sum_{i=1}^{+\infty}\frac{1}{k}\tan^{2k}\left(\frac{\varepsilon}{2}\right)\sin(2k(\hat{R}(0)))\right\}\end{aligned} \tag{12.1.642}$$

By construction we have $\hat{R}(0)=0$ (and don't forget that at the human scale we assume $\hat{P}$ as constant), therefore:

$$\begin{aligned}\hat{R}^{\mathrm{P}}(t)\cong\int\limits_0^t \mathrm{d}\hat{R}^{\mathrm{P}}\\ =\left\{\hat{R}(t)+\sum_{i=1}^{+\infty}\frac{1}{k}\tan^{2k}\left(\frac{\varepsilon}{2}\right)\sin(2k(\hat{R}(t)+\hat{P}))\right\}-\sum_{i=1}^{+\infty}\frac{1}{k}\tan^{2k}\left(\frac{\varepsilon}{2}\right)\sin(2k\hat{P})\end{aligned} \tag{12.1.643}$$

If we put the actual real values ($\varepsilon\cong 23.45°$, $\hat{P}\cong 12.25°$) into the last term and compute it with a specialized software like Maple™ we get:

$$\sum_{i=1}^{+\infty}\frac{1}{k}\tan^{2k}\left(\frac{\varepsilon}{2}\right)\sin(2k\hat{P})\cong 0.01859\cong 1.065° \tag{12.1.644}$$

So we will neglect this constant for the next developments. Hence:

$$\hat{R}^{\mathrm{P}}(t)\cong\hat{R}(t)+\sum_{i=1}^{+\infty}\frac{1}{k}\tan^{2k}\left(\frac{\varepsilon}{2}\right)\sin(2k(\hat{R}(t)+\hat{P})) \tag{12.1.645}$$

To simplify the notation let us get rid of the explicit time dependence:

$$\hat{R}^{\mathrm{P}}\cong\hat{R}(t)+\sum_{i=1}^{+\infty}\frac{1}{k}\tan^{2k}\left(\frac{\varepsilon}{2}\right)\sin(2k(\hat{R}+\hat{P})) \tag{12.1.646}$$

Let us replace $\hat{R}$ by its explicit expression (see page 5418):

$$\hat{R} \cong \hat{M} + 2e\sin(\hat{M}) + \frac{5}{4}e^2\sin(2\hat{M}) + e^3\left(-\frac{1}{4}\sin(\hat{M}) + \frac{13}{12}\sin(3\hat{M})\right) + e^4\ldots \tag{12.1.647}$$

Thus, $\hat{R}^{\text{p}}$ becomes a function of $\hat{M}$:

$$\begin{aligned}\hat{R}^{\text{P}} \cong\ & \hat{M} + 2e\sin(\hat{M}) + \frac{5}{4}e^2\sin(2\hat{M}) + e^3\left(-\frac{1}{4}\sin(\hat{M}) + \frac{13}{12}\sin(3\hat{M})\right) + \ldots\\ & + \tan^2\left(\frac{\varepsilon}{2}\right)\sin\left\{2(\hat{M}+\hat{P}) + 2\left[2e\sin(\hat{M}) + \frac{5}{4}e^2\sin(2\hat{M})\right.\right.\\ & \left.\left. + e^3\left(-\frac{1}{4}\sin(\hat{M}) + \frac{13}{12}\sin(3\hat{M})\right) + \ldots\right]\right\}\\ & + \frac{1}{2}\tan^4\left(\frac{\varepsilon}{2}\right)\sin\left\{4(\hat{M}+\hat{P}) + 4\left[2e\sin(\hat{M}) + \frac{5}{4}e^2\sin(2\hat{M})\right.\right.\\ & \left.\left. + e^3\left(-\frac{1}{4}\sin(\hat{M}) + \frac{13}{12}\sin(3\hat{M})\right) + \ldots\right]\right\}\\ & + \frac{1}{3}\tan^6\left(\frac{\varepsilon}{2}\right)\sin\left\{6(\hat{M}+\hat{P}) + 6\left[2e\sin(\hat{M}) + \frac{5}{4}e^2\sin(2\hat{M})\right.\right.\\ & \left.\left. + e^3\left(-\frac{1}{4}\sin(\hat{M}) + \frac{13}{12}\sin(3\hat{M})\right) + \ldots\right]\right\} + \ldots\end{aligned} \tag{12.1.648}$$

As e is small for the Earth-Sun system let us consider that we can neglect all terms greater than e^2. Therefore it remains:

$$\begin{aligned}\hat{R}^{\text{P}} \cong\ & \hat{M} + 2e\sin(\hat{M}) + \frac{5}{4}e^2\sin(2\hat{M})\\ & + \tan^2\left(\frac{\varepsilon}{2}\right)\sin\left\{2(\hat{M}+\hat{P}) + 2\left[2e\sin(\hat{M}) + \frac{5}{4}e^2\sin(2\hat{M})\right]\right\}\\ & + \frac{1}{2}\tan^4\left(\frac{\varepsilon}{2}\right)\sin\left\{4(\hat{M}+\hat{P}) + 4\left[2e\sin(\hat{M}) + \frac{5}{4}e^2\sin(2\hat{M})\right]\right\}\\ & + \frac{1}{3}\tan^6\left(\frac{\varepsilon}{2}\right)\sin\left\{6(\hat{M}+\hat{P}) + 6\left[2e\sin(\hat{M}) + \frac{5}{4}e^2\sin(2\hat{M})\right]\right\} + \ldots\end{aligned} \tag{12.1.649}$$

Let us distribute:

$$\begin{aligned}\hat{R}^{\text{P}} \cong\ & \hat{M} + 2e\sin(\hat{M}) + \frac{5}{4}e^2\sin(2\hat{M})\\ & + \tan^2\left(\frac{\varepsilon}{2}\right)\sin\left\{2(\hat{M}+\hat{P}) + 4e\sin(\hat{M}) + \frac{5}{2}e^2\sin(2\hat{M})\right\}\\ & + \frac{1}{2}\tan^4\left(\frac{\varepsilon}{2}\right)\sin\left\{4(\hat{M}+\hat{P}) + 8e\sin(\hat{M}) + 5e^2\sin(2\hat{M})\right\}\\ & + \frac{1}{3}\tan^6\left(\frac{\varepsilon}{2}\right)\sin\left\{6(\hat{M}+\hat{P}) + 12e\sin(\hat{M}) + 15e^2\sin(2\hat{M})\right\} + \ldots\end{aligned} \tag{12.1.650}$$

Let us focus on the first brace. Using trigonometric identities (see section Trigonometry page 2519), we get:

$$\begin{aligned}\sin\left(2(\hat{M}+\hat{P})+4e\sin(\hat{M})+\frac{5}{2}e^2\sin(2\hat{M})\right)=&\sin(2(\hat{M}+\hat{P}))\cdot\cos(4e\sin(\hat{M}))\cdot\cos\left(\frac{5}{2}e^2\sin(2\hat{M})\right)\\&+\cos(2(\hat{M}+\hat{P}))\cdot\sin(4e\sin(\hat{M}))\cdot\cos\left(\frac{5}{2}e^2\sin(2\hat{M})\right)\\&+\cos(2(\hat{M}+\hat{P}))\cdot\cos(4e\sin(\hat{M}))\cdot\sin\left(\frac{5}{2}e^2\sin(2\hat{M})\right)\\&-\sin(2(\hat{M}+\hat{P}))\cdot\sin(4e\sin(\hat{M}))\cdot\sin\left(\frac{5}{2}e^2\sin(2\hat{M})\right)\end{aligned}\tag{12.1.651}$$

For the trigonometric factor involving e, we can use the Maclaurin approximation (see section Sequences and Series page 1703):

$$\sin(x)=x\qquad\cos(x)=1-\frac{1}{2}x^2\tag{12.1.652}$$

Hence:

$$\begin{aligned}\sin\left(2(\hat{M}+\hat{P})+4e\sin(\hat{M})+\frac{5}{2}e^4\sin(2\hat{M})\right)=&\sin(2(\hat{M}+\hat{P}))\cdot(1-8e^2\sin^2(\hat{M}))\cdot\left(1-\frac{25}{8}e^2\sin^2(2\hat{M})\right)\\&+\cos(2(\hat{M}+\hat{P}))\cdot4e\sin(\hat{M})\cdot\left(1-\frac{25}{4}e^4\sin(2\hat{M})\right)\\&+\cos(2(\hat{M}+\hat{P}))\cdot(1-8e^2\sin^2(\hat{M}))\cdot\frac{5}{2}e^2\sin(2\hat{M})\\&-\sin(2(\hat{M}+\hat{P}))\cdot4e\sin(\hat{M})\cdot\frac{5}{2}e^2\sin(2\hat{M})\end{aligned}\tag{12.1.653}$$

and let us ignore power of e greater than 2, therefore:

$$\begin{aligned}\sin\left(2(\hat{M}+\hat{P})+4e\sin(\hat{M})+\frac{5}{2}e^2\sin(2\hat{M})\right)=&\sin(2(\hat{M}+\hat{P}))\cdot(1-8e^2\sin^2(\hat{M}))\\&+\cos(2(\hat{M}+\hat{P}))\cdot4e\sin(\hat{M})\\&+\cos(2(\hat{M}+\hat{P}))\cdot(1-8e^2\sin^2(\hat{M}))\cdot\frac{5}{2}e^2\sin(2\hat{M})\\&-\sin(2(\hat{M}+\hat{P}))\cdot4e\sin(\hat{M})\cdot\frac{5}{2}e^2\sin(2\hat{M})\end{aligned}\tag{12.1.654}$$

Expanding gets us here:

$$\sin(2(\hat{M}+\hat{P}))+4e\cos(2(\hat{M}+\hat{P}))\sin(\hat{M}))+e^2\left(-8\sin(2(\hat{M}+\hat{P})\sin^2(\hat{M})+\frac{5}{2}\cos(2(\hat{M}+\hat{P}))\sin(2\hat{M})\right)\tag{12.1.655}$$

Doing similarly for all other braces, we get:

$$
\begin{aligned}
R^{\mathrm{P}} \cong\; & \hat{M}+\tan^2\left(\frac{\varepsilon}{2}\right)\left(1-4e^2\right)\sin(2(\hat{M}+\hat{P}))+2e\sin(\hat{M})\\
& -2e\tan^2\left(\frac{\varepsilon}{2}\right)\sin(\hat{M}+2\hat{P})+2e\tan^2\left(\frac{\varepsilon}{2}\right)\sin(3\hat{M}+2\hat{P})\\
& +\frac{1}{2}\tan^4\left(\frac{\varepsilon}{2}\right)\sin(4(\hat{M}+\hat{P}))+\frac{5}{4}e^2\sin(2\hat{M})-2e\tan^4\left(\frac{\varepsilon}{2}\right)\sin(3\hat{M}+4\hat{P})\\
& +2e\tan^4\left(\frac{\varepsilon}{2}\right)\sin(5\hat{M}+4\hat{P})+\frac{13}{4}e^2\tan^2\left(\frac{\varepsilon}{2}\right)\sin(4\hat{M}+2\hat{P})+\frac{1}{3}\tan^6\left(\frac{\varepsilon}{2}\right)\sin(6(\hat{M}+\hat{P}))
\end{aligned}
\tag{12.1.656}
$$

The equation of time is defined by:

$$
\begin{aligned}
\text{Equation of Time (EoT)} = \; & \text{"true equatorial Sun angle"}\\
& -\text{"mean equatorial Sun angle" (geocentric view)}\\
= \; & -(\text{ true projected anomaly } - \text{ mean anomaly })\\
& \text{(heliocentric view)}\\
= \; & \hat{M}-\hat{R}^{\mathrm{P}}
\end{aligned}
\tag{12.1.657}
$$

We can finally come to the point. The Equation of Time is by definition the difference between true solar time and mean solar time. For instance, if a mean solar clock shows 3hrs 23min and at the same instant a true solar clock (maybe a sundial) shows 3hrs 31min, then the equation of time at that instant is 8min.

According to our definitions, the equation of time is therefore the difference between the hour angles of the true Sun and the mean Sun. Equivalently, it is the difference between the right ascensions (the celestial equivalent of longitude, measuring the East-West position of an astronomical object on the celestial sphere) of the mean Sun and the true Sun.

Now we can calculate the Equation of Time with the values of the parameters as in the early 121st century (holocene calendar) $\varepsilon=23.45^\circ$ and $e=0.0167$:

$$
\begin{aligned}
\text{EoT} := \hat{M}-\hat{R}^{\mathrm{P}} \cong\; & -0.043215861\sin(2(\hat{M}+\hat{P}))-0.0334\sin(\hat{M})\\
& +0.00144502\sin(\hat{M}+2\hat{P})-0.00144502\sin(3\hat{M}+2\hat{P})\\
& -0.000015629\sin 4(\hat{M}+\hat{P})-0.000348613\sin(2\hat{M})\\
& +0.000062518\sin(3\hat{M}+4\hat{P})-0.000062518\sin(5\hat{M}+4\hat{P})\\
& -0.000039214\sin(4\hat{M}+2\hat{P})-0.000026994\sin 6(\hat{M}+\hat{P})
\end{aligned}
\tag{12.1.658}
$$

If want to convert the angle of EoT in seconds, each coefficient has to be multiplied by:

$$
\frac{24\cdot 60\cdot 60}{2\pi}\cong 13750.987083\,[\mathrm{s}\cdot\mathrm{rad}^{-1}]
\tag{12.1.659}
$$

So we have in seconds:

$$
\begin{aligned}
\text{EoT} \cong\; & -594.2607464\sin(2(\hat{M}+\hat{P}))-459.2829686\sin(\hat{M})\\
& +19.87045135\sin(\hat{M}+2\hat{P})-19.87045135\sin(3\hat{M}+2\hat{P})\\
& -0.214914177\sin 4(\hat{M}+\hat{P})-4.79377286\sin(2\hat{M})\\
& +0.85968421\sin(3\hat{M}+4\hat{P})-0.85968421\sin(5\hat{M}+4\hat{P})\\
& -0.539231207\sin(4\hat{M}+2\hat{P})-0.371194145\sin 6(\hat{M}+\hat{P})\;[\mathrm{s}]
\end{aligned}
\tag{12.1.660}
$$

Or rounded to the nearest first decimal:

$$\begin{aligned}\text{EoT} \cong\, & -594.3\sin(2(\hat{M}+\hat{P}))-459.3\sin(\hat{M})\\ & +19.9\sin(\hat{M}+2\hat{P})-19.9\sin(3\hat{M}+2\hat{P})\\ & -0.2\sin 4(\hat{M}+\hat{P})-4.8\sin(2\hat{M})\\ & +0.9\sin(3\hat{M}+4\hat{P})-0.9\sin(5\hat{M}+4\hat{P})\\ & -0.5\sin(4\hat{M}+2\hat{P})-0.4\sin 6(\hat{M}+\hat{P})\ [\text{s}]\end{aligned} \tag{12.1.661}$$

keeping in mind that in the early 121st century we have the angle $\hat{P}$ that measures about $12.25°$.

The equation of time, and therefore the analemma, is determined by the three following parameters in the case of a non-relativistic and stable orbit:

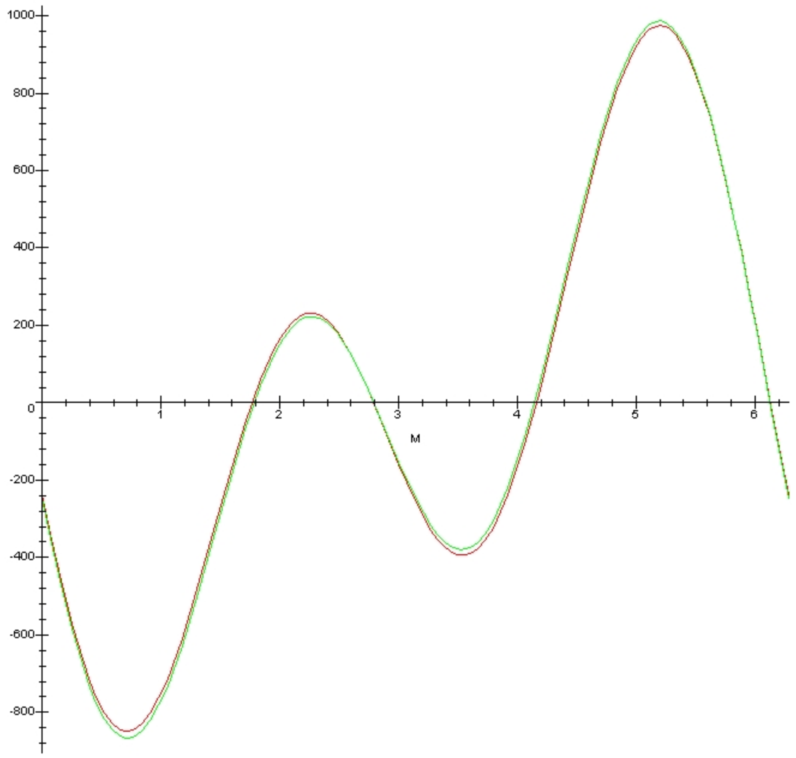
- The eccentricity e of the orbit of the Earth
- The angle between the ecliptic and the equatorial planes ε.
- The angle $\hat{P}$ between the winter solstice and the perihelion relative to the Sun (i.e. the time span Δt from the beginning of winter to the passage through the perihelion)

The last two parameters change gradually by nutation and precession!

Thus, if either of these variables (such as eccentricity) always dominates the other (as is the case on Mars), the analemma will resemble a teardrop. If either of the variables (such as eccentricity) is significant, and the other is practically zero (as is the case on Jupiter, with only a $3°$ tilt), the figure will be something much closer to an ellipse. If both are important enough, that sometimes eccentricity or axial tilt dominates, a figure-eight results.

A plot in Maple™ 4.00b of the above relation shows that if we take the first two terms or the firs three, the difference is negligible:

```
>P:=evalf((12.25/360)*(2*Pi));
>plot([-594.3*sin(2*(M+P))-459.3*sin(M)+19.9*sin(M+2*P),-594.3*sin(2*(M+P))
-459.3*sin(M)],M=0..2*Pi);
```



As we can see the true solar time can deviate from the mean solar time up to approximately 16 minutes. This explains why sundials that give the true solar time differ most of the time from the time visible on your watch (the latter giving a political business time that exists only for convenience).

That's why it is common to find the equation of time as defined as:

$$\boxed{\text{Equation of Time} \cong -594.3\sin(2(\hat{M}+\hat{P})) - 459.3\sin(\hat{M})\;[\text{s}]} \tag{12.1.662}$$

where for reminder:

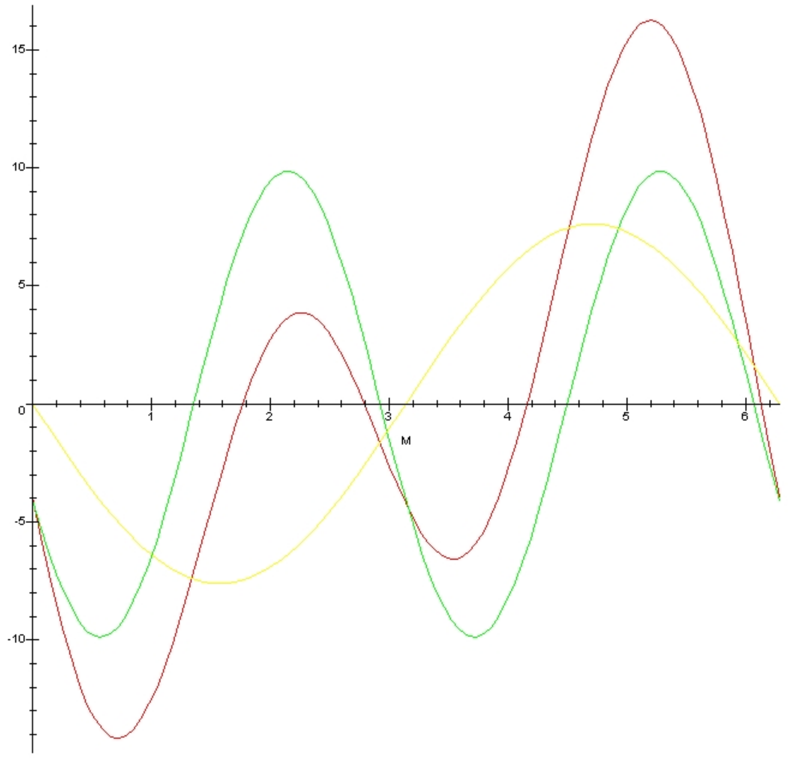
$$\hat{M} = 2\pi\frac{t}{T} \tag{12.1.663}$$

Or in minutes with the common abbreviation and in minutes instead of seconds:

$$\text{EoT} \cong -9.9\sin(2(\hat{M}+\hat{P})) - 7.65\sin(\hat{M})\;[\text{min}] \tag{12.1.664}$$

The graph of the equation of time is closely approximated by the sum of two sine curves, one with a period of a year and one with a period of half a year. The curves reflect two astronomical effects, each causing a different non-uniformity in the apparent daily motion of the Sun relative to the stars:

- The obliquity of the ecliptic
- The eccentricity of the Earth's orbit around the Sun



The diagram below represents the equation of time as a function of the day of the year (divide by 10 such that the day 350 for example is equal to 35), starting at the perigee (the black curve):

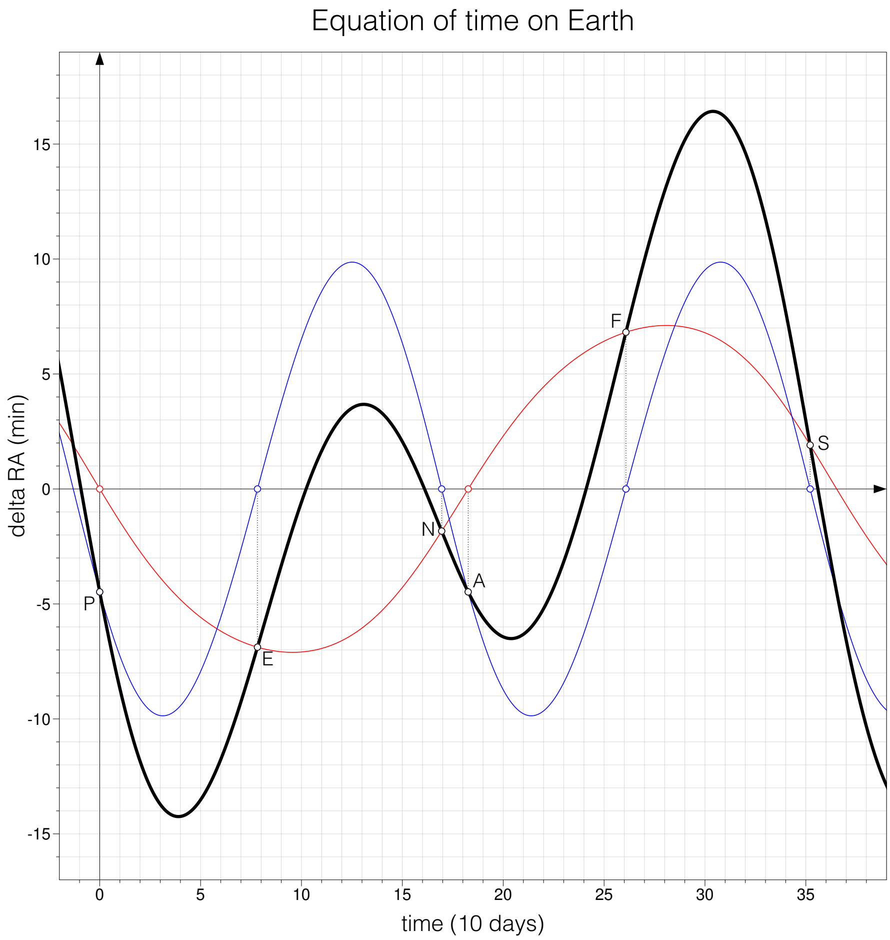


Figure 12.88: **Equation of time on Earth**
(author: Pierre Barbier, source: https://pbarbier.com/eqtime/eqtime.html**)**

The other curves represent the individual contributions of the obliquity (the difference between "uniform solar time" and mean solar time, in blue), and the eccentricity of the orbit (the difference between true solar time and uniform solar time, in red). The ordinate of every point on the black curve is the sum of the ordinates of the points on the other curves at the same abscissa.

The special points on the curves mark the situations:

- At the perigee (P) and apogee (A), where the true and uniform suns coincide and there is no difference between true and uniform solar times
- At the equinoctial points, vernal (E) and autumnal (F), where the uniform and mean suns coincide and there is no difference between uniform and mean solar times
- At the north (N) and south (S) solstitial points where the uniform and mean suns coincide again and there is no difference between uniform and mean solar times.

Note that the instants corresponding to the points (E) (F) (N) (S) are the times of passage of the uniform Sun at the equinoctial and solstitial points on the relative orbit of the sun. The true Sun does not necessarily pass through these points at exactly the same time.

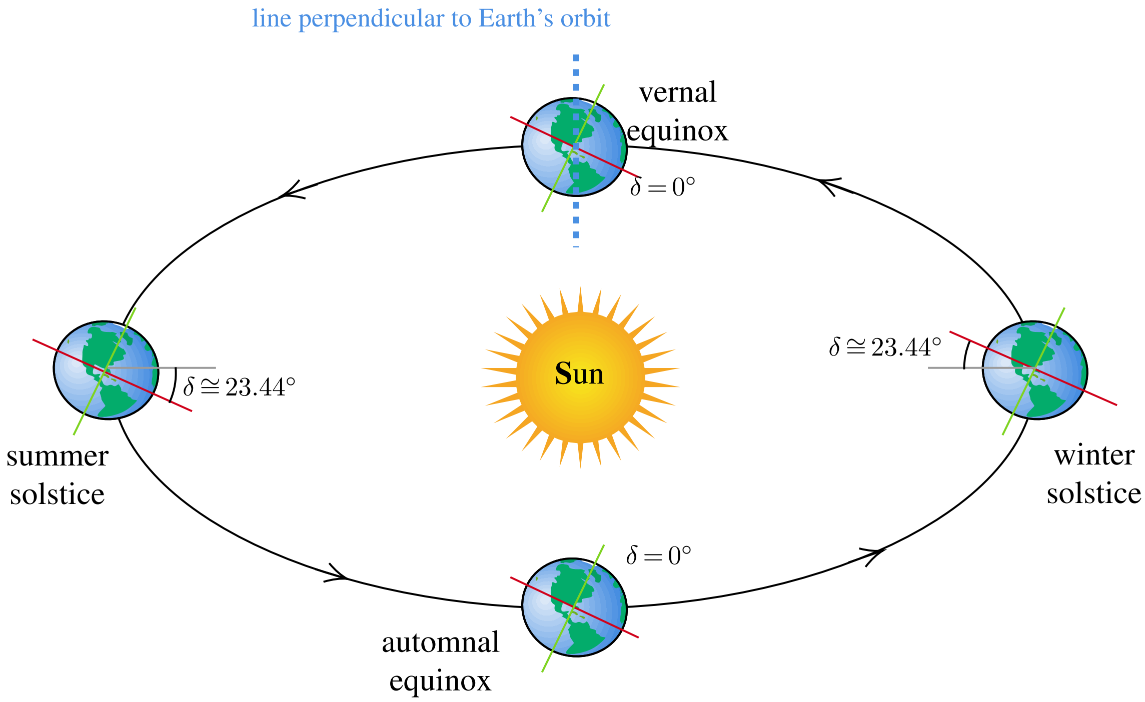
Now to get the analemma, we have to plot the equation of time in function of the declination (angular distance of a celestial body North (+) or South (−) from the celestial equator). Indeed by definition the analemma is the curve obtained by plotting the equation of time as the horizontal coordinate against the declination of the (true) Sun as the vertical coordinate.

Let us first remind that the declination is the angular distance of a celestial object North or South of the celestial equator. The "declination" is given by (the number 10 offset accounts for the fact that the winter solstice - when declination is at its minimum - occurs before the beginning of the year):

$$\boxed{\delta \cong -23.45^\circ \cdot \cos\left(\frac{360}{365}\cdot(d+10)\right)} \tag{12.1.665}$$

where d is the day of the year with January 1 as $d=1$ and the content of the parenthesis of the $\cos$ must not be forgotten to be converted into radians to get the right result!

The declination is zero at the equinoxes (March 22 and September 22), positive during the northern hemisphere summer and negative during the northern hemisphere winter. The declination reaches a maximum of 23.45° on June 22 (summer solstice in the northern hemisphere) and a minimum of -23.45° on December 21-22 (winter solstice in the northern hemisphere). In the relation above, the $+10$ comes from the fact that the winter solstice occurs before the start of the year. The equation also assumes that the Sun orbit is a perfect circle (hence the approximation sign $\cong$) and the factor of $360/365$ converts the day number to a position in the orbit.



Remark

The detailed derivation of the declination angle assumes a circular orbit and gives the following result:

$$\sin(\delta) = \sin(-23.45^\circ)\cos\left(\frac{360}{365}\cdot(d+10)\right) \qquad (12.1.666)$$

In radians, $\alpha \cong \sin(\alpha)$, for small values of δ so the derivation reduces to the approximation:

$$\delta \cong -23.45^\circ \cos\left(\frac{360}{365}\cdot(d+10)\right) \qquad (12.1.667)$$

The difference between the two equations is very small ($< 0.3^\circ$), whereas the error to the true declination. Thus the simpler equation is preferred since it is acceptable for common civil usage.

The equalities:

$$\begin{aligned}\delta &\cong 23.45^\circ \cdot \sin\left(\frac{360}{365}\cdot(d+284)\right)\\ \delta &\cong 23.45^\circ \cdot \sin\left(\frac{360}{365}\cdot(d-81)\right)\end{aligned} \qquad (12.1.668)$$

are also often used in the literature. They are referred to the equinoxes instead so use $\sin$ instead of $\cos$.

Finally, there are many algorithms for more accurate determinations of declination angle to account for the elliptic and yearly movement of the Earth's orbit based on Fourier series.

The various ways of calculating the declination angle are graphed below using Python™ code:

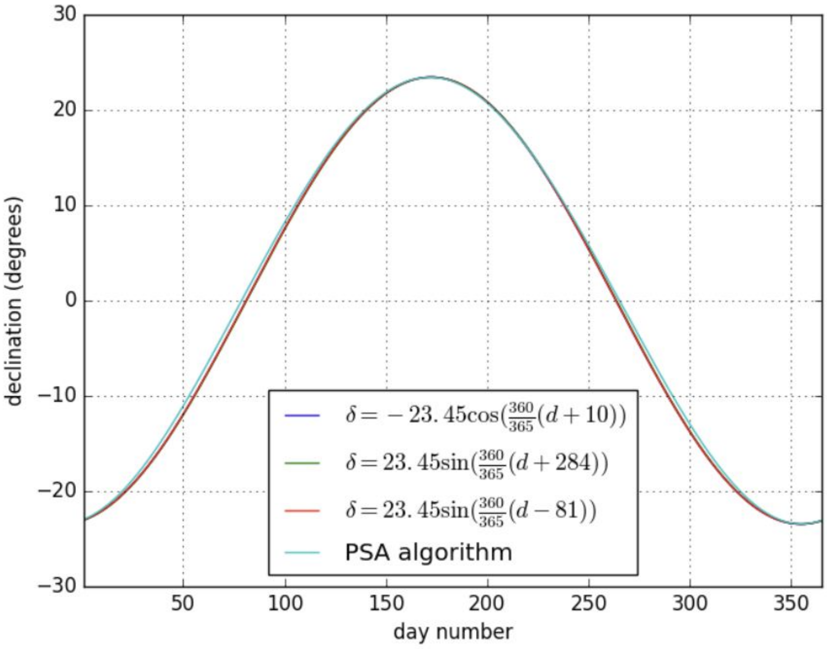
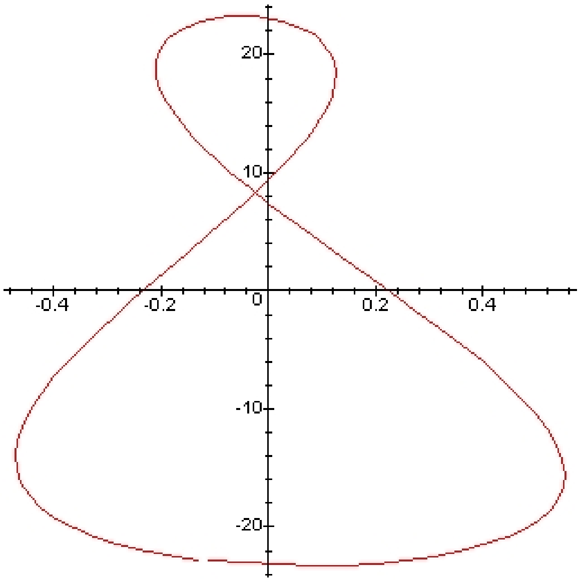


Figure 12.89: Various declination calculation methods compared (authors: Christiana Honsberg and Stuart Bowden)

Hence in Maple™ 4.00b we have:

```
>epsilon:=23.45/360*2*Pi;
>P:=12.25/360*2*Pi;
>e:=0.0167;
>EOT:=-(tan(epsilon/2))^2*(1-4*e^2)*sin(2*(2*Pi*t/365+P))-2*e*sin(2*Pi*t/365);
>Declination:=0.4059*cos((360/365)/360*2*Pi*(t+10));
>plot([EOT,-Declination/Pi*180,t=0..365]);
```



The analemma is the curve obtained by plotting the equation of time as the x coordinate against the declination of the (true) Sun as the y coordinate:

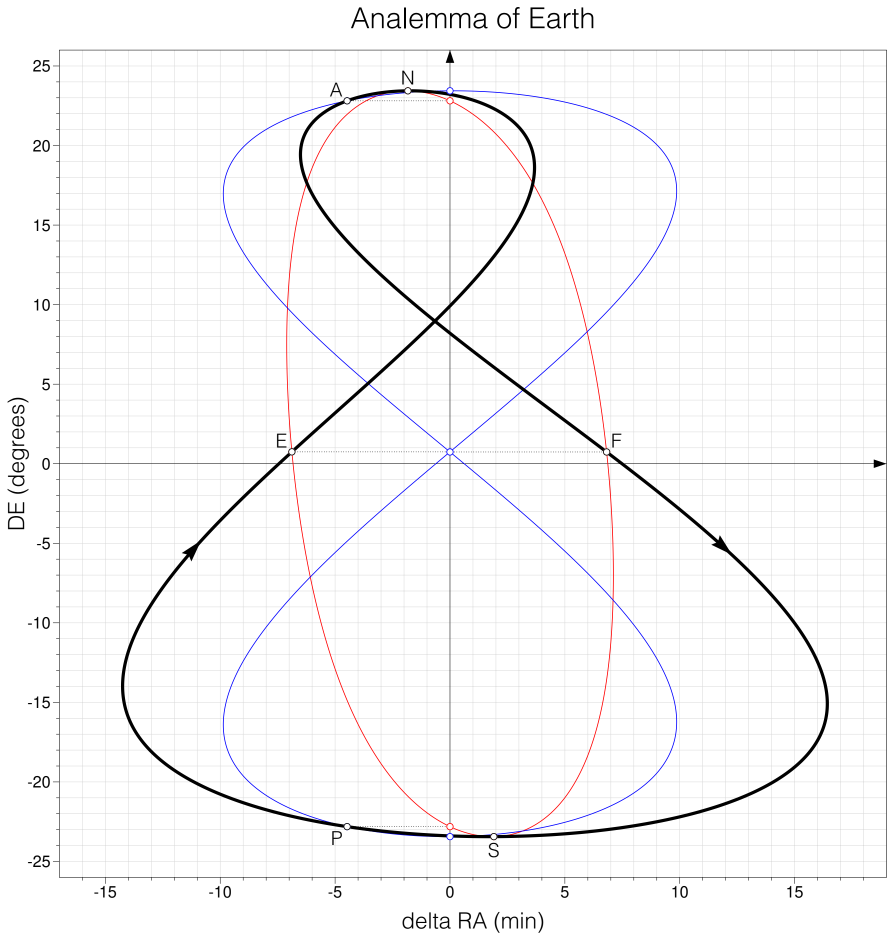


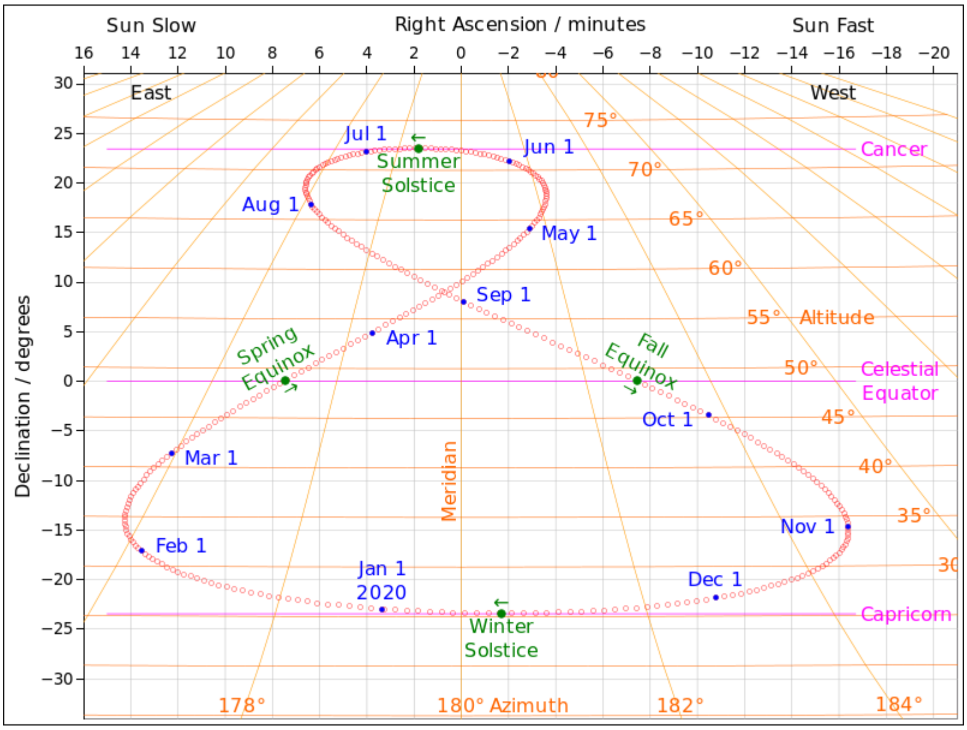
Figure 12.90: **Earth analema**
(author: Pierre Barbier, source: `https://pbarbier.com/eqtime/eqtime.html`**)**

At each instant, the declination of the Sun and value of the equation of time correspond to a point on the curve. This point moves along the curve as the Sun moves along the ecliptic, in the direction indicated by the arrow on the black curve, completing a full turn in a year.

Again, the individual contributions of the obliquity (in blue) and the eccentricity (in red) are represented, as well as the situation at the perigee (P), apogee (A), vernal (E) and autumnal (F) points, north (N) and south (S) solstitial points. The abscissa of any point of the black curve is the sum of the abscissas of the points of the blue and red curves at the same ordinate (and chosen on the correct branch of the curve, the one corresponding to the same season, since in most cases two abscissas correspond to the same ordinate on each curve).

This particular way of plotting the equation of time makes it convenient to compute mean solar time from the reading on a sundial. The sundial may give the declination of the Sun as the length of the shadow of the stylus, in addition to true solar time. The time correction, the equation of time, can be read directly on the analemma curve at that declination. For this reason, the analemma curve can sometimes be found drawn on the face of sundials.

A more detailed version looks like:



Remark

Where the intersection occurs is a result of the asymmetry of the analemma caused by the difference between the solstice date and perihelion dates.

Another version:

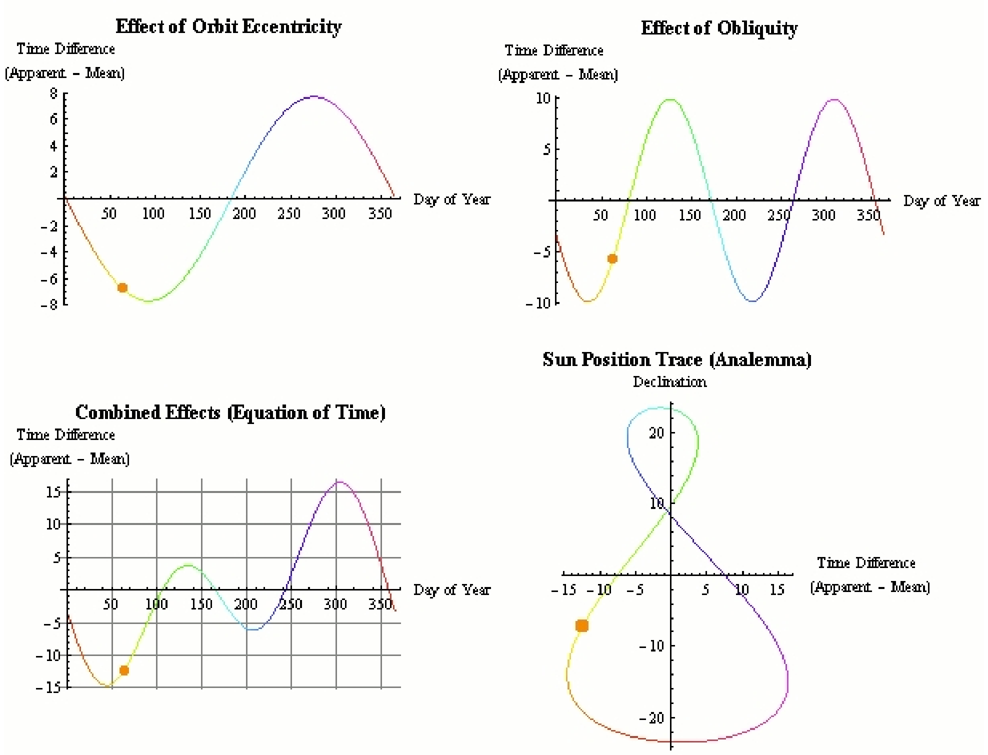
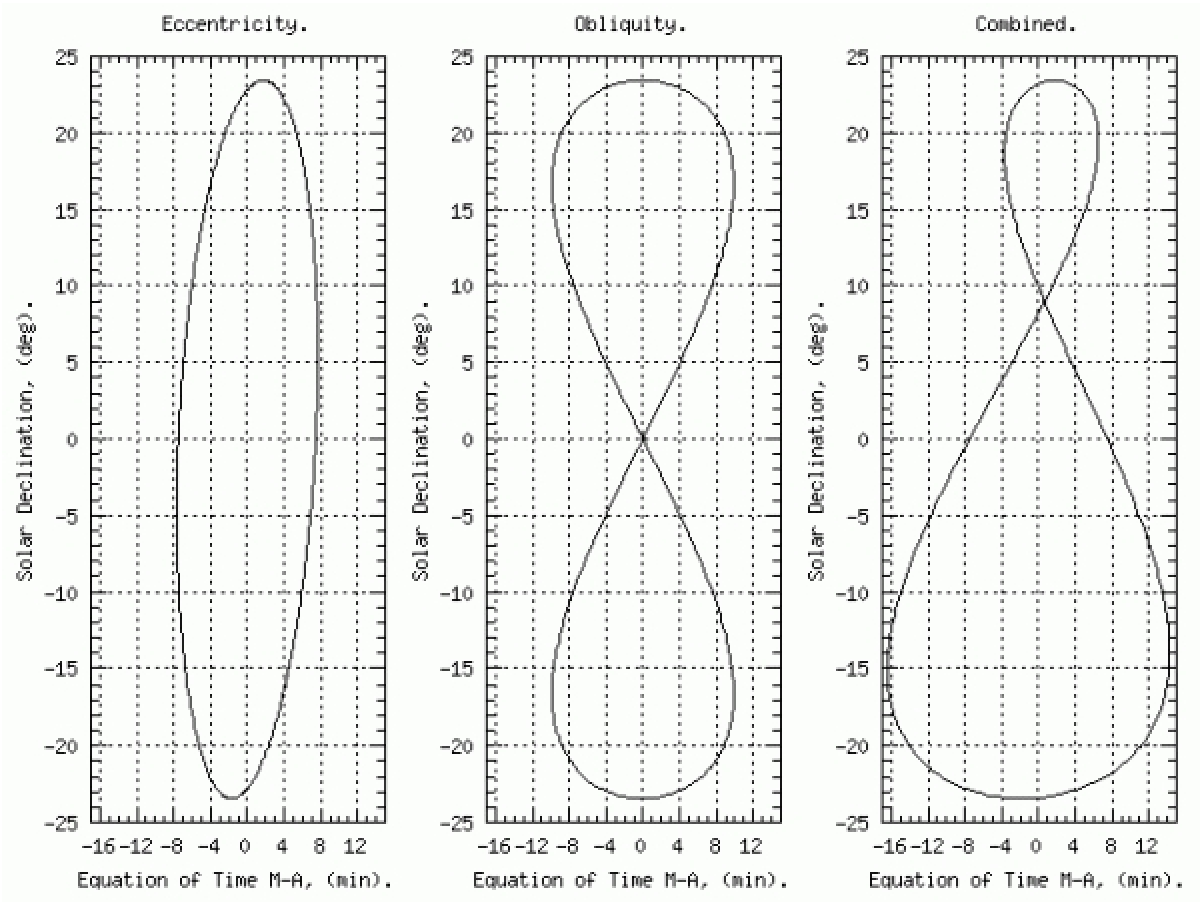
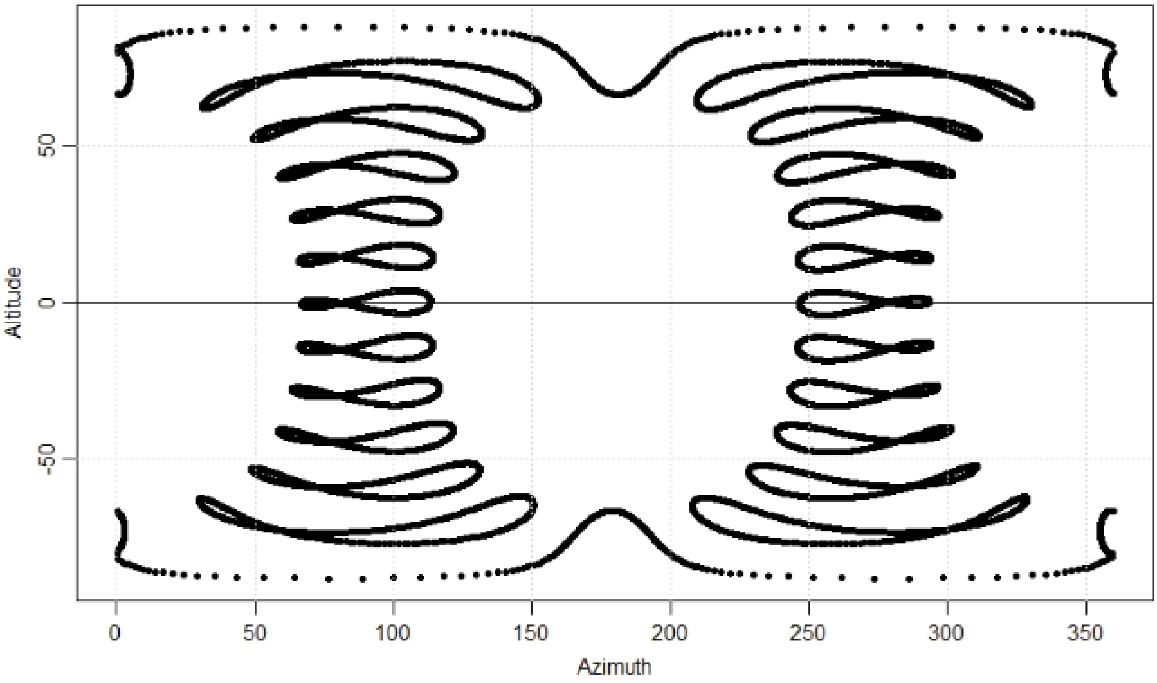


Figure 12.91: Equation of time and analemma correspondence (author: Rob Cook)

Here is a visual with the different components of the analemma:



Or using the R software (see the corresponding detailed code in the R companion book) we can have the figure of the analemma at different azimuth and altitude:



Obviously our results match with reality because the Earth isn't flat...

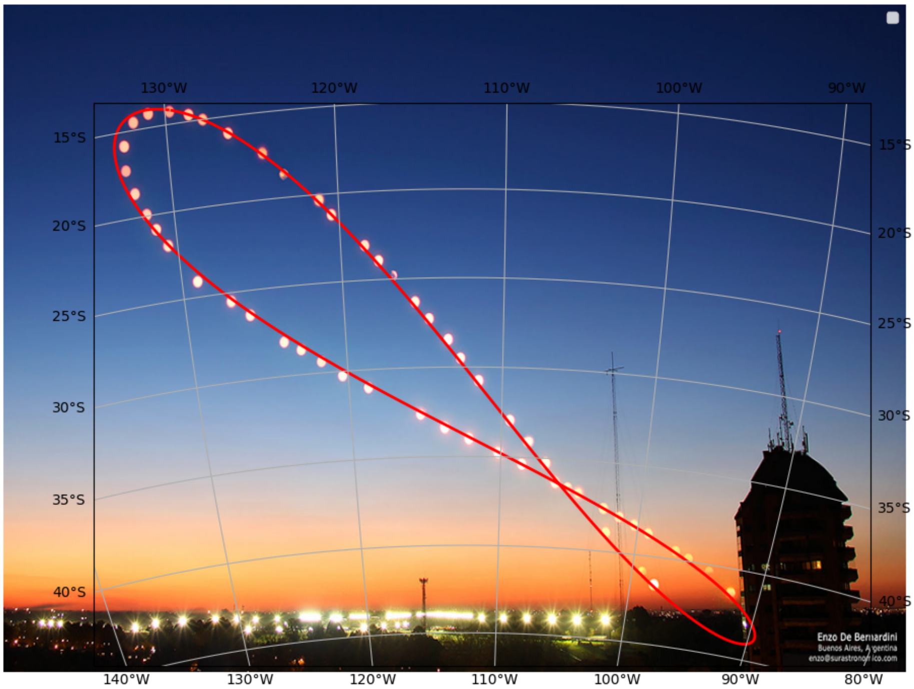

130°W
120°W
110°W
100°W
90°W
15°S
20°S
25°S
30°S
35°S
40°S
15°S
20°S
25°S
30°S
35°S
40°S
140°W
130°W
120°W
110°W
100°W
90°W
80°W
Enzo De Bernardini
Buenos Aires, Argentina
enzo@surastronomico.com

Figure 12.92: **Juxtaposition of theoretical analemma with an observed one**

12.1.11 Sundials

A "sundial" is a horological device that tells the time of day (referred to as civil time in modern usage) when direct sunlight shines by the apparent position of the Sun (actually a G2V-type star, aka yellow dwarf) in the sky. In the narrowest sense of the word, it consists of a flat plate (the dial) and a gnomon (a protruding structure, like a pole or bar), which casts a shadow onto the dial. As the Sun appears to move through the sky, the shadow aligns with different hour-lines, which are marked on the dial to indicate the time of day. The style is the time-telling edge of the gnomon, though a single point or nodus may be used. The gnomon casts a broad shadow; the shadow of the style shows the time. The gnomon may be a rod, wire, or elaborately decorated metal casting. The style must be parallel to the axis of the Earth's rotation for the sundial to be accurate throughout the year. The style's angle from horizontal is equal to the sundial's geographical latitude. If a sundial is to tell the time accurately it must ordinarily be designed for the particular latitude in which it is to be used.

The term sundial can refer to any device that uses the Sun's altitude or azimuth[20] (or both) to show the time. Sundials are valued as decorative objects, metaphors, and objects of intrigue and mathematical study.

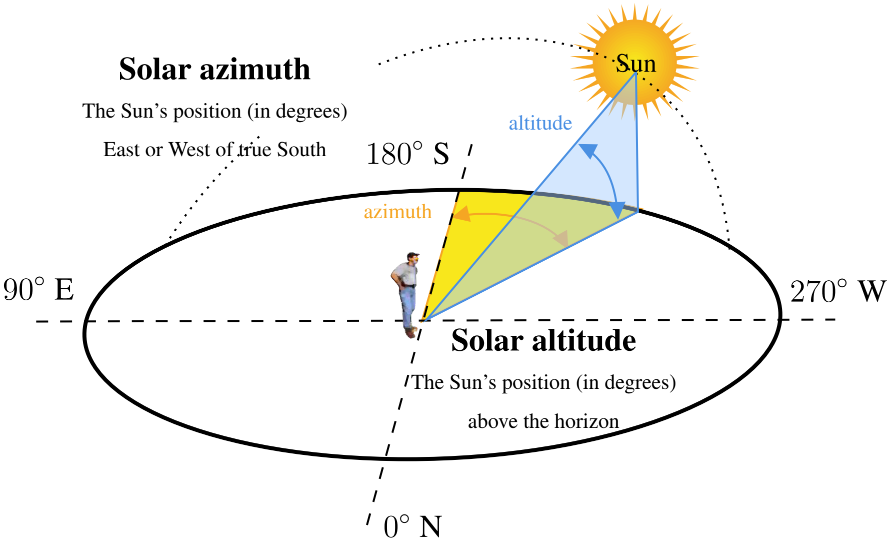
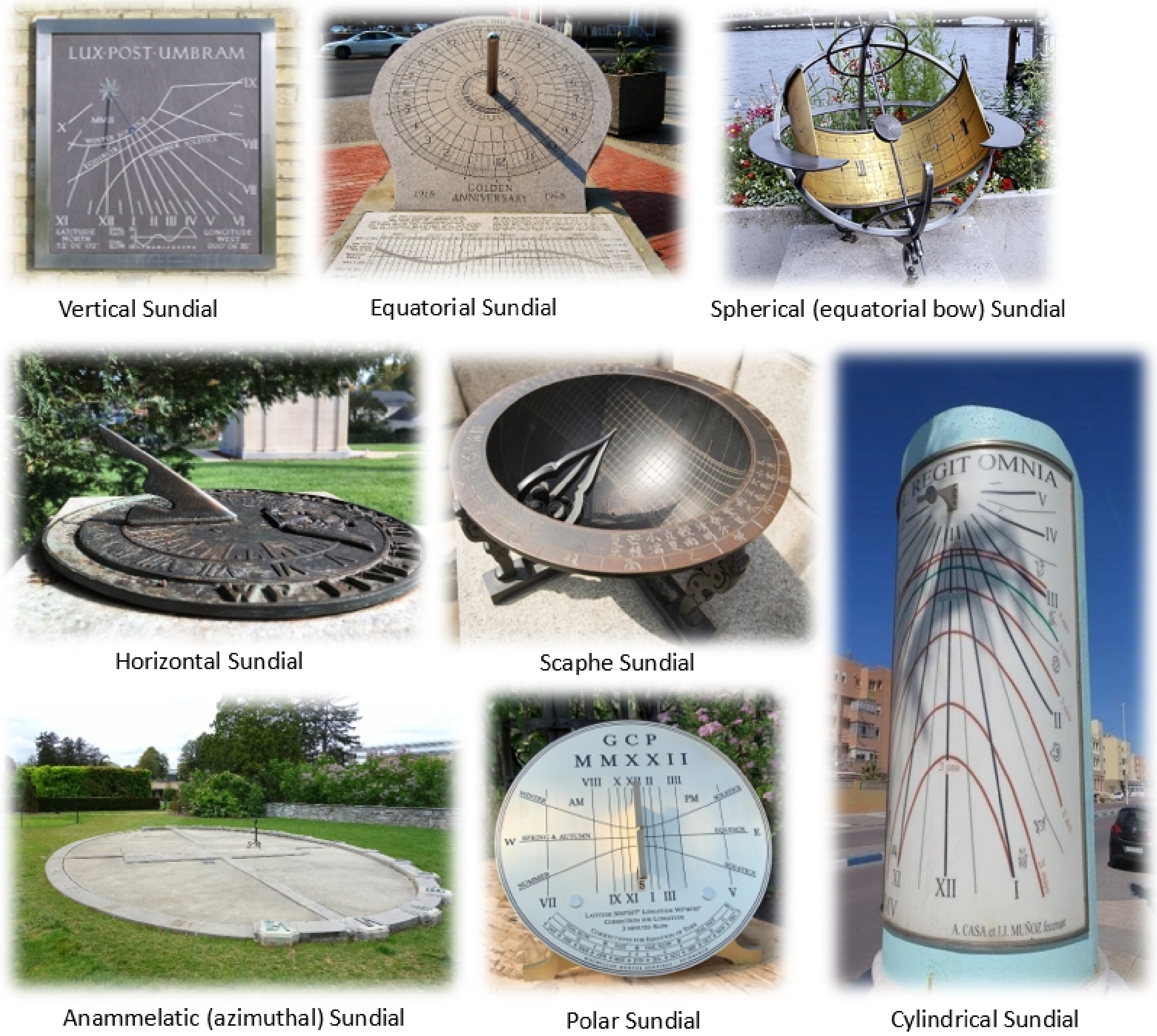

Solar azimuth
The Sun's position (in degrees)
East or West of true South
Sun
altitude
180° S
azimuth
90° E
270° W
Solar altitude
The Sun's position (in degrees)
above the horizon
0° N

Figure 12.93: **Azimuth and Altitude depicted**

There are several different types of sundials illustrated below:

---

[20] The "azimuth" is the horizontal angle from a cardinal direction, most commonly North, in a local or observer-centric spherical coordinate system. Sadly there seems to be no international standard that defined unequivocally the azimuth to a given cardinal.



Vertical Sundial

Equatorial Sundial

Spherical (equatorial bow) Sundial

Horizontal Sundial

Scaphe Sundial

Anammelatic (azimuthal) Sundial

Polar Sundial

Cylindrical Sundial

Figure 12.94: **Most common type of sundials**

Caution! Most sundials show what is known as "true local time". This differs from the time on your watch, which is named "standard mean time".

Let us give a small description of the most common one:

- The **vertical sundial** is a type of sundial where the dial plate is mounted vertically, often on a wall. It's designed to measure time using the shadow cast by a gnomon. These sundials are commonly found on buildings, especially those facing south named **vertical direct south dials**, and are used to display the time throughout the day.
- The **equatorial sundial** has the advantages that it can be made without mathematical calculation, and, unlike most other sundials, may be used in any latitude if properly set up. It gets it name from the fact that the "dial plate" lies parallel to the Equator. The gnomon is a rod or pin set perpendicular to the dial plate in the center of the dial. This gnomon is parallel to the earth's axis and points toward the north celestial pole, close to the Pole Star.

- The **horizontal sundial** is far more common than all other types combined. Its popularity arise in part from the fact that it tells the time whenever the Sun is shining, while many types of sundials can be used only during restricted hours of the day. In addition horizontal dials are comparatively easy to make and to set into place for use. They make beautiful and appropriate garden ornaments, and even a small garden can well find a spot for one at the end of a pathway or as an accent in a formal flower border.

- The **spherical sundial**, also known as an **armillary sphere** or **spherical astrolabe**, is a type of sundial that uses a spherical framework to display time. It's essentially a model of the celestial sphere with rings representing celestial longitude and latitude. This design allows for the accurate indication of time by the movement of shadows cast by a gnomon positioned within the sphere.

- The **scaphe sundial** is an ancient type of sundial that consists of a hemispherical bowl with a vertical gnomon placed inside it. The shadow of the gnomon moves across the curved surface, allowing time to be read from the hour lines inscribed within the bowl. This design compensates for the non-uniform shadow movement seen on flat sundials, making it a fascinating piece of early astronomical ingenuity.

- The **analemmatic sundial** is a type of horizontal sundial that has a vertical gnomon and hour markers arranged in an elliptical pattern. Unlike traditional sundials, the gnomon in an analemmatic sundial is not fixed - it must be moved daily to account for the changing position of the Sun throughout the year. One funny feature of this sundial is that it can be designed as a "human sundial", where a person stands in place of the gnomon and their shadow indicates the time.

- The **polar sundial** is a unique type of sundial where the dial plate is parallel to the Earth's axis, meaning it's tilted at an angle that matches the observer's latitude. Unlike traditional sundials, its hour lines are evenly spaced and parallel to each other, rather than radiating outward from a central point. The gnomon is also parallel to the dial plate.

- The **vertical cylindrical sundial** is a type of sundial where the hour lines are marked on the outer surface of a vertical cylinder. The gnomon, which casts the shadow, is typically perpendicular to the cylinder's surface and can be oriented in different directions depending on the design.

The mathematical proofs and development that follow are almost a copy/paste of the excellent article *The mathematics of sundials* (see [562]) from PhD Jill Vincent (university of Melbourne). We also recommend the reader the excellent book *Sundials: History, theory, and practice* (see [466]).

Below we will show how trigonometry can be used to calculate the positions of the hour lines for vertical and horizontal sundials. The words "vertical" and "horizontal" are used in their normal sense, that is, in the direction of gravity and at right angles to this.

12.1.11.1 Equatorial sundials

Let us imagine the Earth as a giant sundial (wee figure below). The Earth's axis is tilted actually (earyl 121st century according to holocene calendar) at an angle of 23.5° to the plane of its orbit around the Sun. As the Earth rotates on its axis, the shadow of a vertical stick at the pole would form a circle on the surface of the Earth parallel to the equator. If the circle is divided into 24 equal hour marks, the position of the shadow around the circle would give the time. Sundials based on this principle are named "equatorial sundials".

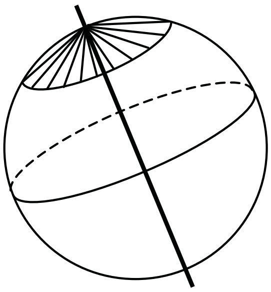


Figure 12.95: Representing the Earth as a giant sundial

In any sundial, the part that casts the shadow is named the "gnomon" as we saw it. Here is a very nice visual reproduced with the authorization of the British Sundial Society:

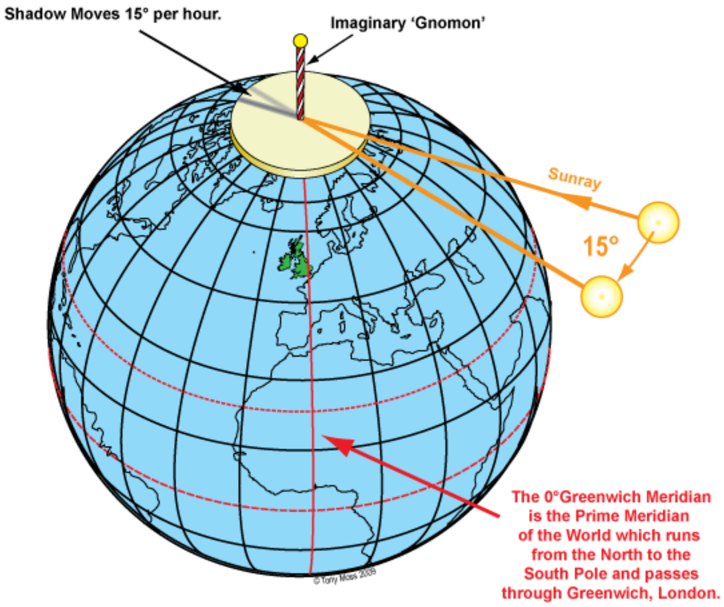
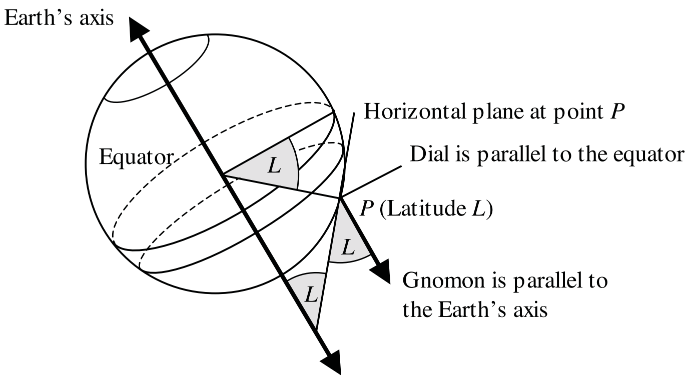

Shadow Moves 15° per hour.
Imaginary 'Gnomon'
Sunray
15°
The 0°Greenwich Meridian is the Prime Meridian of the World which runs from the North to the South Pole and passes through Greenwich, London.
© Tony Moss 2009

Figure 12.96: Equatorial sundial at the North pole
(source: `https://sundialsoc.org.uk`, author: Tony Moss)

For an equatorial sundial, the gnomon must be parallel to the Earth's axis. In the figure below P is in the southern hemisphere at latitude L. The diagram shows why the angle that the gnomon makes with the horizontal is equal to the latitude. The dial with the hour marks must be perpendicular to the gnomon. In the southern hemisphere the gnomon will point to the south celestial pole (in the northern hemisphere, it points towards the pole star):


Earth's axis
Horizontal plane at point *P*
Dial is parallel to the equator
Equator
L
P (Latitude *L*)
L
L
Gnomon is parallel to the Earth's axis

The figure below shows an equatorial sundial for the southern hemisphere (look closely to the cardinal `N`, `S` indicators position), with the gnomon making an angle with the horizontal equal to the latitude.

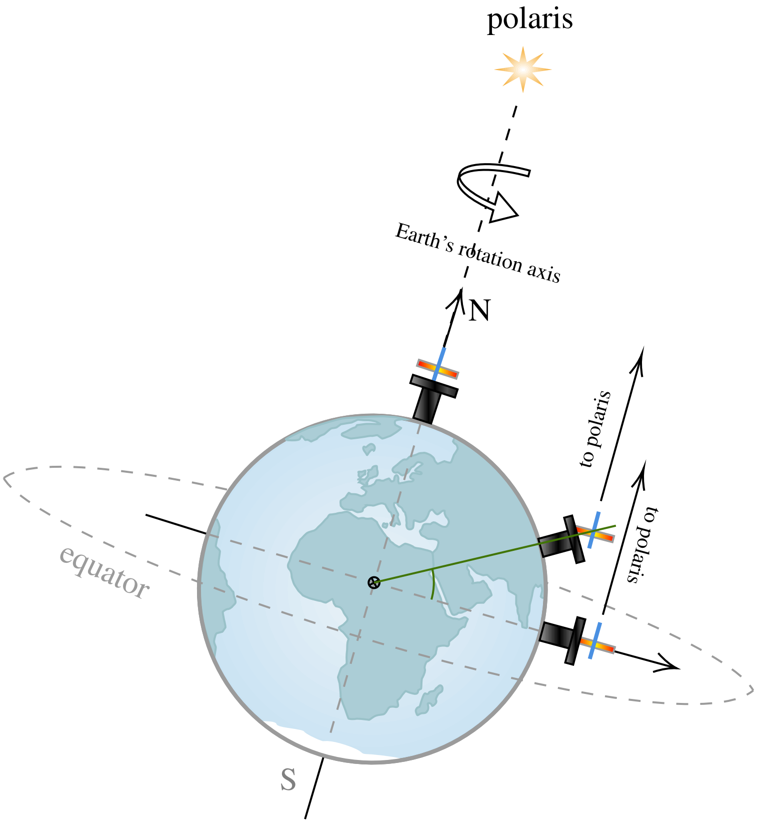

polaris
Earth's rotation axis
N
to polaris
to polaris
equator
S

Figure 12.97: **Equatorial sundials at different latitudes**

The hour lines are equally spaced, with 15° intervals (angular speed of Earth). At midsummer the Sun is directly in line with the gnomon so there is no shadow. Between the spring equinox (22 September) and the autumn equinox (22 March), the shadow falls on the upper face of the dial. Between 22 March and 22 September, when the Sun is lower in the sky, the shadow falls on the under side. Obviously not all hours are indicated because we can't see any shadow during during the night... (however in reality they are all indicated for aesthetic reasons even if its dumb...).

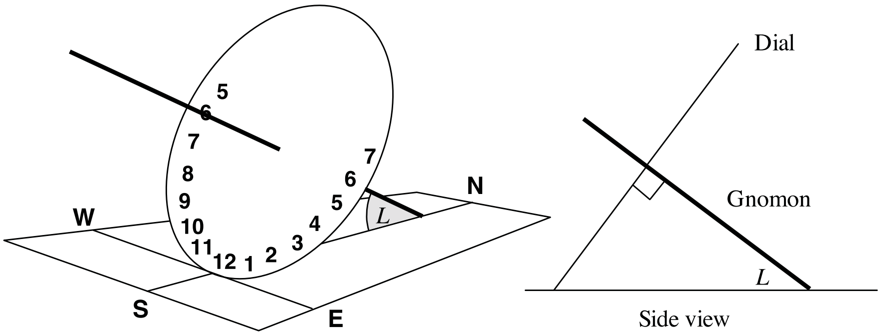

5
6
7
8
9
10
11
12
1
2
3
4
5
6
7
L
W
S
E
N
Dial
Gnomon
L
Side view

Figure 12.98: Equatorial sundial for the southern hemisphere

12.1.11.2 Vertical sundials

For a vertical sundial, the circular equatorial dial presented earlier above is projected onto a vertical plane and this results obviously to an ellipse (see figure below). The semi-major axis (parallel to the ground), a, is the radius of the equatorial dial. If b is the semi-minor axis of the ellipse (the reader can also refer to the side view of the previous figure to understand):

$$\sin\left(\frac{\pi}{2}-L\right)=\cos(L)=\frac{a}{b} \tag{12.1.669}$$

Hence:

$$b=\frac{a}{\cos(L)} \tag{12.1.670}$$

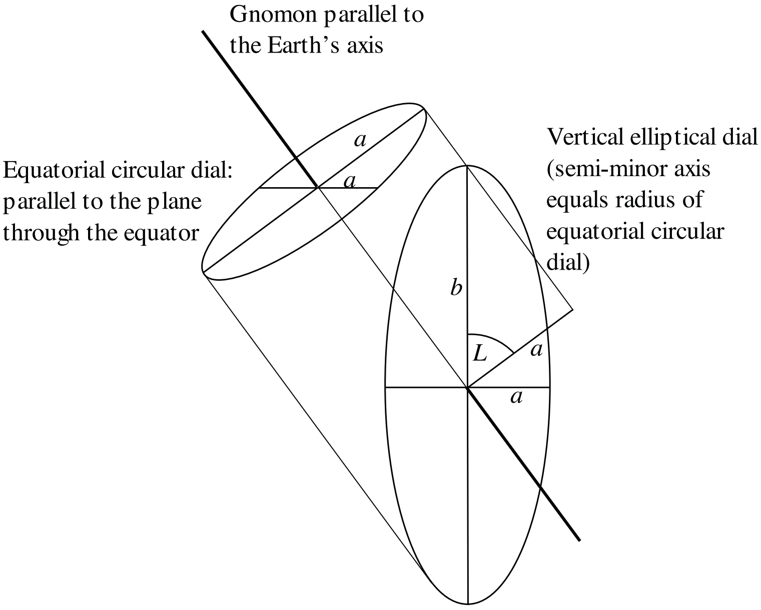

Gnomon parallel to
the Earth's axis
Equatorial circular dial:
parallel to the plane
through the equator
Vertical elliptical dial
(semi-minor axis
equals radius of
equatorial circular
dial)
a
a
b
L
a
a

Figure 12.99: Projection of the equatorial dial to form the ellipse of the vertical dial

The gnomon of the vertical sundial makes an angle of $90^\circ - L$ with the vertical (that is, an angle L with the horizontal), as shown below:

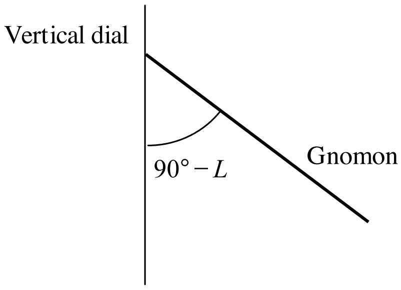
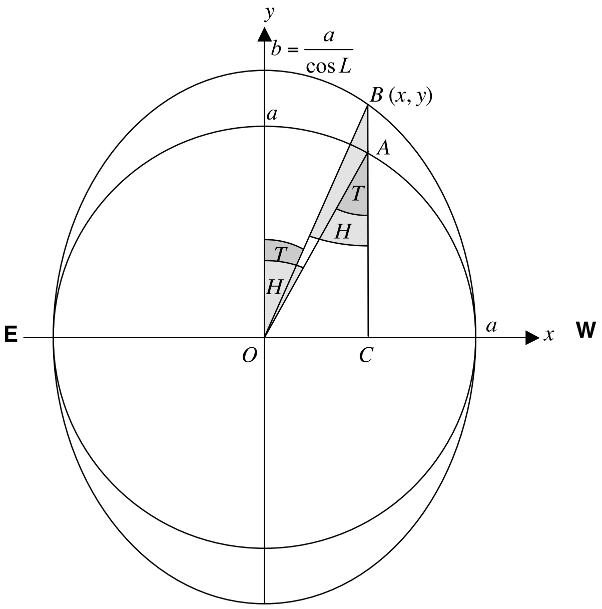


Figure 12.100: **Side view of gnomon of vertical sundial**

Let us calculate now the angle hours for the vertical sundial assuming we want to depict the graduation on a circle for esthetical reasons.

In the southern hemisphere, the vertical dial is north-facing. Unlike the equatorial dial, the hour angles are not equally spaced. In the figure below the angle T is the hour angle measured from the north-south line around the equatorial dial:



Angle H is the projected hour angle on the ellipse of the vertical dial. Using the equation for an ellipse (see section Analytical Geometric page 2768) and applying trigonometry in triangles ΔOAC and ΔOBC, we can find H in terms of L and T.

Let us remind for this purpose the equation of the ellipse:

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1 \qquad (12.1.671)$$

Substituting in it $b = a/\cos(L)$, we get:

$$\frac{x^2}{a^2}+\frac{y^2\cos^2(L)}{a^2}=1 \qquad (12.1.672)$$

After simplification:

$$x^2+y^2\cos(L)=a^2 \tag{12.1.673}$$

but from the triangle ΔAOC, we have:

$$x=a\sin(T) \tag{12.1.674}$$

Therefore substituting and rearranging we get:

$$y^2=\frac{\left(a^2-a^2\sin^2(T)\right)}{\cos^2(L)}=\frac{a^2\left(1-\sin^2(T)\right)}{\cos^2(L)}=\frac{a^2\cos^2(T)}{\cos^2(L)} \tag{12.1.675}$$

Hence:

$$y=\frac{a\cos(T)}{\cos(L)} \tag{12.1.676}$$

From the triangle ΔOBC:

$$\tan(H)=\frac{x}{y} \tag{12.1.677}$$

Using the two relation we got above:

$$x=a\sin(T) \qquad y=\frac{a\cos(T)}{\cos(L)} \tag{12.1.678}$$

We can write:

$$\tan(H)=\frac{x}{y}=\frac{a\sin(T)}{\dfrac{a\cos(T)}{\cos(L)}}=\frac{a\sin(T)\cos(L)}{a\cos(T)}=\tan(T)\cos(L) \tag{12.1.679}$$

So finally:

$$H=\tan^{-1}(\tan(T)\cos(L)) \tag{12.1.680}$$

The hour angles, H, for 6 am to 12 noon for a vertical dial in Melbourne (latitude $37.7°$) are shown in the table below:

Time	T (°)	H (°)
12 noon	0	0.0
11 am	15	12.0
10 am	30	24.6
9 am	45	38.4
8 am	60	53.9
7 am	75	71.3
6 am	90	90.0

The angles are symmetrical about the noon line for 1 pm to 6 pm:

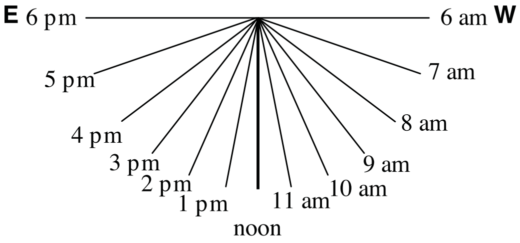


Figure 12.101: **Hour angles for vertical sundial in Melbourne**

12.1.11.3 Horizontal sundials

For a horizontal sundial, the circular equatorial dial is projected onto a horizontal plane as an ellipse as depicted in the figure below:

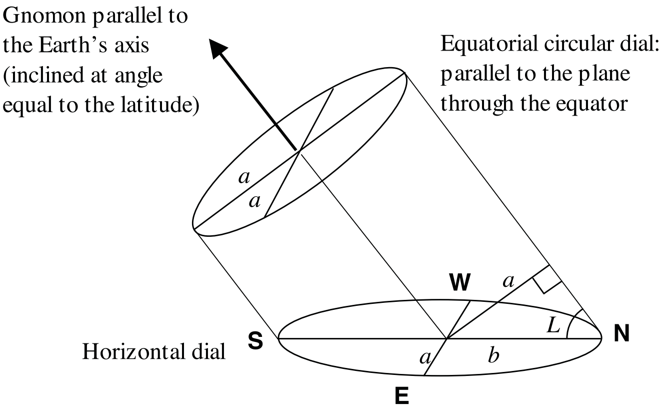
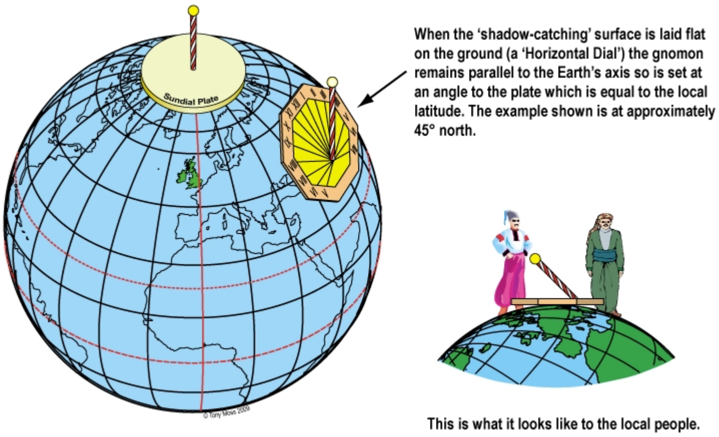

Gnomon parallel to
the Earth's axis
(inclined at angle
equal to the latitude)
Equatorial circular dial:
parallel to the plane
through the equator
a
a
W
a
L
N
Horizontal dial
S
a
b
E

Figure 12.102: Hour angles for vertical sundial

Here is a very nice visual reproduced with the authorization of the British Sundial Society:


Sundial Plate
© Tony Moss 2009

When the 'shadow-catching' surface is laid flat on the ground (a 'Horizontal Dial') the gnomon remains parallel to the Earth's axis so is set at an angle to the plate which is equal to the local latitude. The example shown is at approximately 45° north.

This is what it looks like to the local people.

Figure 12.103: Horizontal sundial illustrated on the globe
(source: `https://sundialsoc.org.uk`, author: Tony Moss)

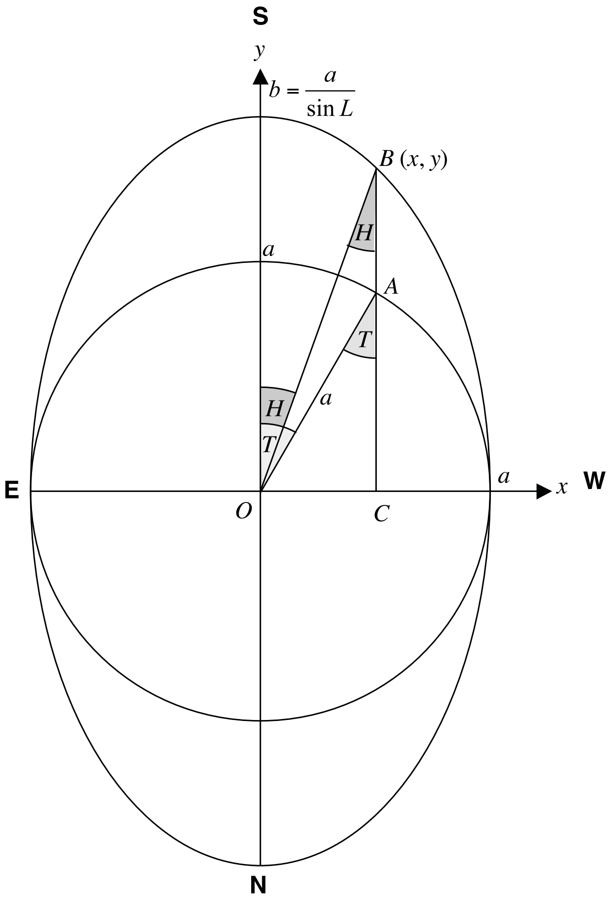
As for the equatorial and vertical sundials, the gnomon makes an angle L with the horizontal. The semi-minor (east-west) axis is a, the radius of the equatorial dial. If b is the semi-major (north-south) axis of the ellipse and we get therefore:

$$\sin(L)=\frac{a}{b} \tag{12.1.681}$$

Hence:

$$b=\frac{a}{\sin(L)} \tag{12.1.682}$$

In the figure below, T is the hour angle measured from the north-south line around the equatorial dial:



The angle H is the projected hour angle on the ellipse of the horizontal dial. Using exactly the same approach as for the vertical sundial, the parametric equations for the ellipse for the horizontal sundial are:

$$x = a\sin(T) \qquad y = \frac{a\cos(T)}{\sin(L)} \tag{12.1.683}$$

From:

$$\tan(H) = \tan(T)\sin(L) \tag{12.1.684}$$

We get that the hour angles are given by:

$$H = \tan^{-1}(\tan(T)\sin(L)) \tag{12.1.685}$$

The hour angles, H, for a horizontal sundial in Melbourne are shown in the table below:

Time	T (°)	H (°)
6 am	90	90.0
7 am	75	66.3
8 am	60	46.6
9 am	45	31.4
10 am	30	19.4
11 am	15	9.3
12 noon	0	0.0

The latitude 37.7° has been used. For 11 am , for example, $H = \tan^{-1}(\tan(15^\circ)\sin(37.7^\circ)) \cong 9.3^\circ$. Unlike the vertical sundial, the numbers are anticlockwise for a horizontal sundial in the southern hemisphere.

The angles are symmetrical about the noon line too:

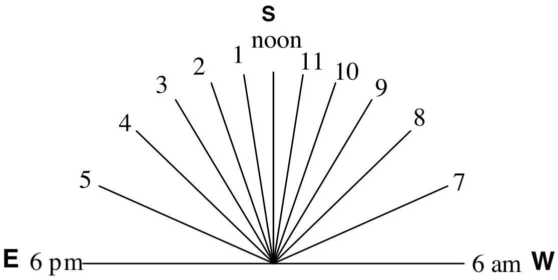

S
noon
1
2
3
4
5
11
10
9
8
7
E 6 pm
6 am W

Figure 12.104: Hour angles for horizontal sundial in Melbourne

We can see below an interesting combo of a vertical, polar and horizontal dial:

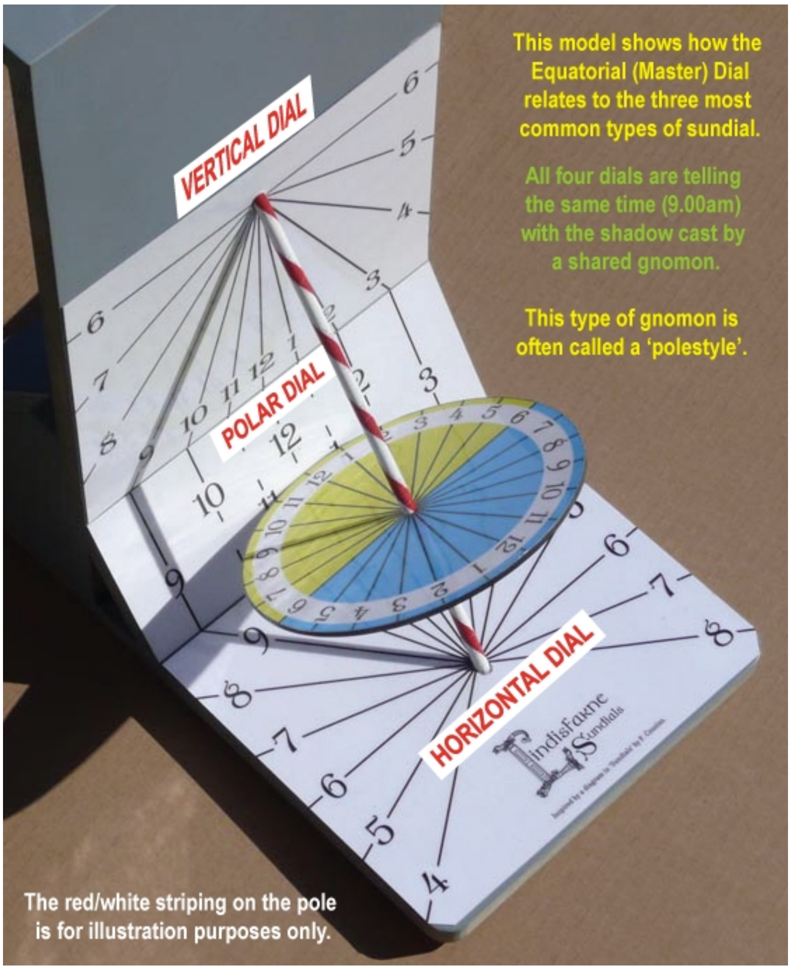

This model shows how the Equatorial (Master) Dial relates to the three most common types of sundial.
All four dials are telling the same time (9.00am) with the shadow cast by a shared gnomon.
This type of gnomon is often called a 'polestyle'.
VERTICAL DIAL
POLAR DIAL
HORIZONTAL DIAL
LindisFarne Sundials
The red/white striping on the pole is for illustration purposes only.

Figure 12.105: Three types of sundials in one piece (manufacturer: LindisFnake Sundials)

12.1.11.4 Anammelatic (azimuthal) sundials

The anammelatic (azimuthal) has instead of a gnomon set at the angle of the latitude, a vertical gnomon (relatively to the ground) or person that stands on the north-south line at a calculated position according to their height (see the mosaic of photo with the various sundial earlier above!).

If we consider a right-angled triangle formed by the vertical gnomon (or person) and the distance d from the center of the sundial at which the person stands, the hypotenuse of this triangle has to do an angle of $37.8028°$ (the local latitude of Melbourne for example) with the horizontal:

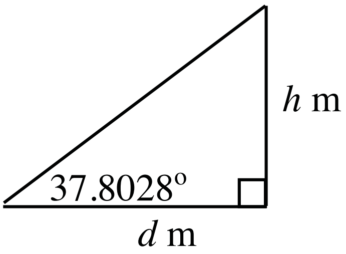


Figure 12.106: Relationship between height and position for anammelatic sundial

For example, for a person of height 1.80 [m]:

$$\tan(37.8028°)=\frac{1.80}{d} \tag{12.1.686}$$

Hence:

$$d=\frac{1.80}{\tan(37.8028°)}\cong 2.32\,[\text{m}] \tag{12.1.687}$$

So a person of height 1.8 [m] would stand on the north-south line at a position 2.32 [m] from the center of the sundial.

If the vertical (human) gnomon were to be replaced with a right-angled triangle with its hypotenuse making an angle L with the horizontal, then the shadow of the hypotenuse at any given hour would fall along the hour line. The tip of the person's shadow would meet the hour line as shown below:

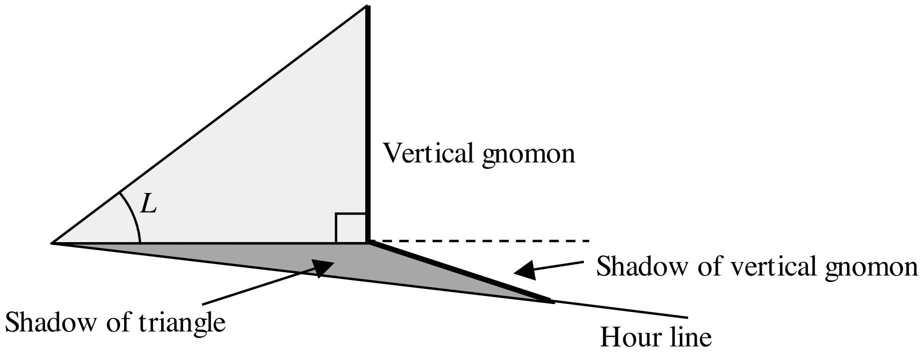


Figure 12.107: Tip of shadow of human gnomon meets the hour line

The figure below shows the view from above of the hour line, $\overline{OP}$, and the shadow, $\overline{GP}$, of a person standing on the solar clock at G. These two lines intersect at P:

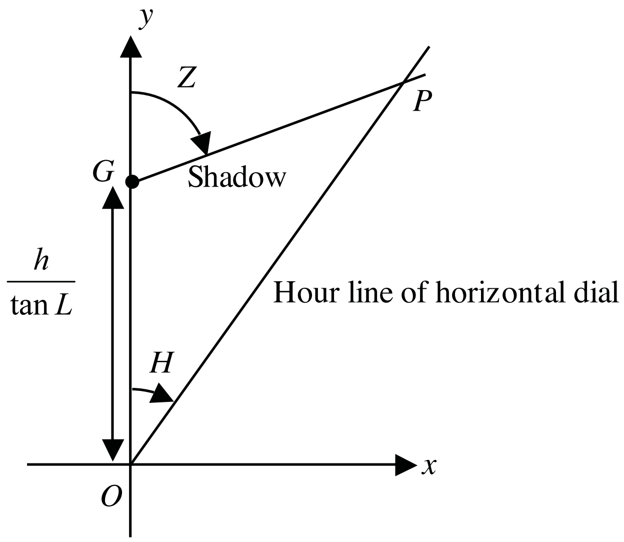


Figure 12.108: Intersection of person's shadow with the hour line.

The coordinates of P can be found by solving simultaneously the equations for $\overline{GP}$ and $\overline{OP}$.

The angle Z (the Sun's azimuth chosen relatively to the North cardinal in the developments that will follow is equal to $A=\pi-Z$ as the Sun is at the opposite direction to the shadow) that the shadow of a vertical gnomon makes with the north-south line depends on the latitude, the time of day and the declination (angular distance of a celestial body North $(+)$ or South $(-)$ from the celestial equator) of the Sun, which changes through the day and through the year.

For pedagogical reasons let us first consider the simple case of the equinox when the declination angle is $D=0^\circ$.

We know that the Sun appears to move approximately actually around the star Polaris in an approximate large circle. At the equinox when $D=0^\circ$ this circle is exactly the celestial equator. We can describe the position of the Sun by its angle relative to some fixed reference point. For convenience we take this point to be noon, at which time the Sun is highest. The angle T of the Sun relative to the noon point is named the "hour angle", with $T=0^\circ$ at noon. In fact $T=15^\circ\cdot t$, where t is the time (in hours) past noon (remember the the Sun rise at the East and set at the West):

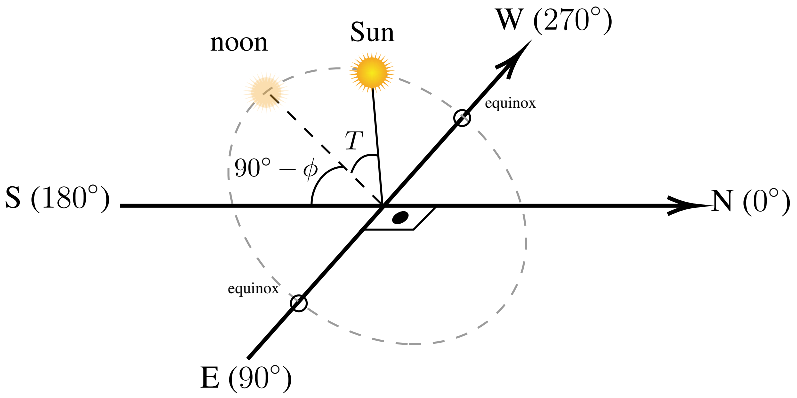


Figure 12.109: **Finding the hour angle**

Now, what shape do we get if we look at the shadow cast by a vertical pointer onto a horizontal surface? To answer this we have to project the circle that the Sun makes. If $D=0^\circ$ this circle is at an angle of $90^\circ-\phi$ to the horizontal and intersects the horizontal at 6 am and 6 pm on the equinoxes (as at this special day of the year there is 12 hours of day and 12 hours of night).

If R is the radius of the circle on which the Sun moves and if the x-axis is aligned East-West and the y-axis aligned North-South, then the (x,y) coordinates of the Sun can only be given by (the reader can check the validity of these relations for some remarkable angles):

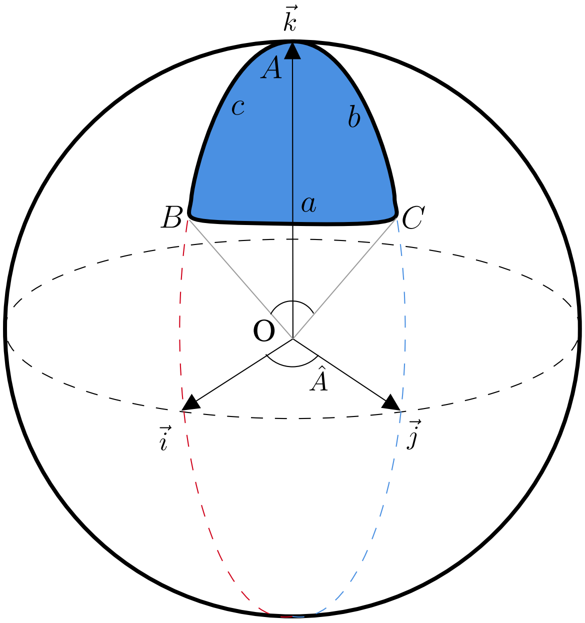
$$x=R\sin(T)\qquad y=R\cos(T)\sin(\phi) \tag{12.1.688}$$

The angle the shadow of the Sun makes with respect to the y-axis is named the "azimuth" and we write this as Z. It is given therefore by:

$$\cot(Z)=\frac{y}{x}=\sin(\phi)\cot(T) \tag{12.1.689}$$

Now that we have done this the angle Z can be calculated according to the four-part formula where T is the angle hour measured clockwise from the north-south line, L is the latitude, and D is the declination (angular distance of a celestial body North $(+)$ or South $(-)$ from the celestial equator) of the Sun.

Now let us remind the four-part formula proved in the section of Trigonometry (see page 2548):



Let the sides a, b, c of $\triangle ABC$ be measured by the angles subtended at O, where a, b, c are opposite A, B, C respectively.

We proved that:

$$\cos(a)\cos(\hat{C}) = \sin(a)\cot(b) - \sin(\hat{C})\cot(\hat{B}) \tag{12.1.690}$$

Now let us represent the following configuration with also a spherical triangle whose vertices are defined by the zenith of the gnomon, the true geographical north and the Sun's projection on a celestial sphere of arbitrary radius:

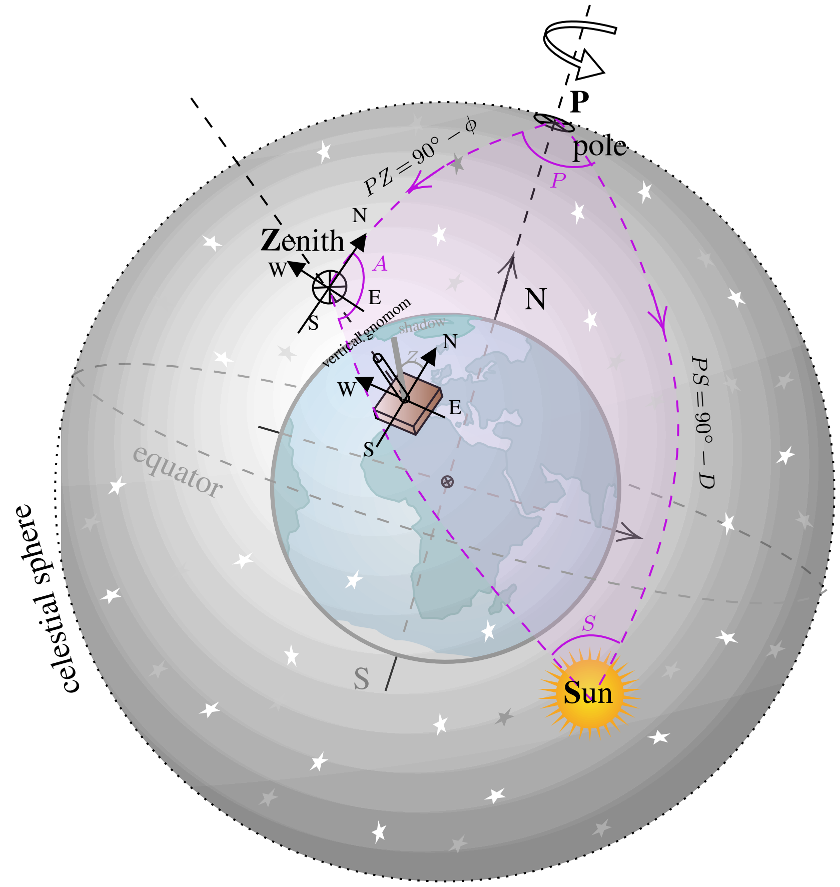


Figure 12.110: **Anammelatic sundial on Earth with a spherical defined triangle**

Therefore the four-parts formula for the spherical triangle defined by the gnomon zenith (Z), the geographical pole (P) and the Sun position (S) is given by:

$$\cos(PZ)\cos(P)=\sin(PZ)\cot(PS)-\sin(P)\cot(A) \tag{12.1.691}$$

Or explicitly:

$$\cos(90^\circ-\phi)\cos(P)=\sin(90^\circ-\phi)\cot(90^\circ-D)-\sin(P)\cot(A) \tag{12.1.692}$$

Using trigonometric complementary angles relations, this reduces to:

$$\sin(\phi)\cos(P)=\cos(\phi)\tan(D)-\sin(P)\cot(A) \tag{12.1.693}$$

Now remember that the azimuth angle A (Sun's angle position relatively to the North) is related to Z (the angle of the gnomon shadow relatively to the north) by:

$$A=\pi-Z \tag{12.1.694}$$

Therefore the four-parts formula becomes:

$$\sin(\phi)\cos(P)=\cos(\phi)\tan(D)-\sin(P)\cot(\pi-Z) \tag{12.1.695}$$

Hence using again trigonometric complementary angles relations:

$$\sin(\phi)\cos(P)=\cos(\phi)\tan(D)+\sin(P)\cot(Z) \tag{12.1.696}$$

Rearranging we get:

$$\cot(Z)=\frac{\sin(\phi)\cos(P)-\cos(\phi)\tan(D)}{\sin(P)}=\sin(\phi)\cot(P)-\frac{\cos(\phi)\tan(D)}{\sin(P)} \tag{12.1.697}$$

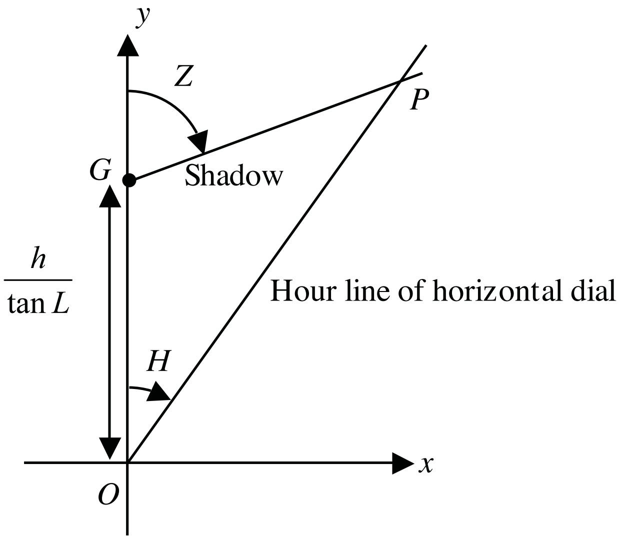
And let us come back to our initial notation by denoting the latitude L instead of ϕ and T (hour angle) instead of P:

$$\cot(Z)=\frac{\sin(L)\cos(T)-\cos(L)\tan(D)}{\sin(T)}=\sin(L)\cot(T)-\frac{\cos(L)\tan(D)}{\sin(P)} \tag{12.1.698}$$

Notice that when $D=0°$ (equinox) we fall back on:

$$\cot(Z)=\sin(\phi)\cot(T)=\sin(L)\cot(T) \tag{12.1.699}$$

Now we can use this spherical trigonometric relation to study (we reproduce again this figure here for convenience):



We have for the hour line $\overline{OP}$:

$$y=x\tan(90°-H)=\frac{x}{\tan(H)} \tag{12.1.700}$$

But remember that we have seen during our study of the equatorial sundial (see page 5447) that:

$$\tan(H)=\tan(T)\sin(L) \tag{12.1.701}$$

Therefore:

$$y=\frac{x}{\tan(T)\sin(L)} \tag{12.1.702}$$

And for the the shadow $\overline{GP}$ we have:

$$y=\frac{h}{\tan(L)}+x\cot(Z) \tag{12.1.703}$$

Substituting for $\cot(Z)$ we just got thanks to the four-parts formula of spherical trigonometry, we get:

$$y=\frac{h}{\tan(L)}+\frac{x(\sin(L)\cos(T)-\tan(D)\cos(L))}{\sin(T)} \tag{12.1.704}$$

From:

$$y=\frac{x}{\tan(T)\sin(L)} \quad \text{and} \quad y=\frac{h}{\tan(L)}+\frac{x(\sin(L)\cos(T)-\tan(D)\cos(L))}{\sin(T)} \tag{12.1.705}$$

equating we get:

$$\frac{x}{\tan(T)\sin(L)}=\frac{h}{\tan(L)}+\frac{x(\sin(L)\cos(T)-\tan(D)\cos(L))}{\sin(T)} \tag{12.1.706}$$

Expliciting $\tan(T)$:

$$\frac{x\cos(T)}{\sin(T)\sin(L)}=\frac{h\cos(L)}{\sin(L)}+\frac{x(\sin(L)\cos(T)-\tan(D)\cos(L))}{\sin(T)} \tag{12.1.707}$$

Multiplying both sides by $\sin(T)\sin(L)$:

$$\begin{aligned}x\cos(T)&=h\sin(T)\cos(L)+x\sin(L)(\sin(L)\cos(T)-\tan(D)\cos(L))\\&=h\sin(T)\cos(L)+x\left(\sin^2(L)\cos(T)-\sin(L)\tan(D)\cos(L)\right)\end{aligned} \tag{12.1.708}$$

Hence after rearranging:

$$\begin{aligned}h\sin(T)\cos(L)&=x\left(\cos(T)-\sin^2(L)\cos(T)+\sin(L)\tan(D)\cos(L)\right)\\&=x\left(\cos(T)\left(1-\sin^2(L)\right)+\sin(L)\tan(D)\cos(L)\right)\\&=x\left(\cos(T)\cos^2(L)+\sin(L)\tan(D)\cos(L)\right)\\&=x\cos(L)(\cos(T)\cos(L)+\tan(D)\sin(L))\end{aligned} \tag{12.1.709}$$

Therefore:

$$h\sin(T)=x(\cos(T)\cos(L)+\tan(D)\sin(L)) \tag{12.1.710}$$

So for the x component we have:

$$x=\frac{h\sin(T)}{\cos(T)\cos(L)+\tan(D)\sin(L)} \tag{12.1.711}$$

Substituting to find the y-coordinate of the intersection:

$$\begin{aligned}y&=\frac{x}{\tan(T)\sin(L)}=\frac{h\sin(T)}{\tan(T)\sin(L)(\cos(T)\cos(L)+\tan(D)\sin(L))}\\&=\frac{h\cos(T)}{\sin(L)(\cos(T)\cos(L)+\tan(D)\sin(L))}\end{aligned} \tag{12.1.712}$$

Hence the coordinates of the tip of the vertical gnomon (or person's shadow) of the analemmatic sundial:

$$\boxed{\left(\frac{h\sin(T)}{\cos(T)\cos(L)+\tan(D)\sin(L)},\frac{h\cos(T)}{\sin(L)(\cos(T)\cos(L)+\tan(D)\sin(L))}\right)} \tag{12.1.713}$$

where for reminder T is the angle hour measured clockwise from the north-south line, L is the latitude, and D is the declination (angular distance of a celestial body North (+) or South (−) from the celestial equator) of the Sun.

Notice that at the famous moment of equinoxes (when $D=0°$), we get a horizontal line since the y component is equal to:

$$y=\frac{h\cos(T)}{\sin(L)(\cos(T)\cos(L)+\tan(D)\sin(L))}=\frac{h\cos(T)}{\sin(L)\cos(T)\cos(L)}=\frac{h}{\sin(L)\cos(L)}=c^{te} \tag{12.1.714}$$

Notice that for the proof of this parametric equation we never assumed multiple rays of lights but only one ray of light (hence rays of light are not assumed parallel nor divergent).

For an analemmatic sundial positioned at latitude $L=37.8°$, the figure below shows a plot of the hour lines and the positions of the shadow tip during the day for the equinoxes and the summer and winter solstices ($D=0°, 23.5°$ and $-23.5°$ respectively) for a person of height 1.80 [m] standing at $G(0,2.32)$:

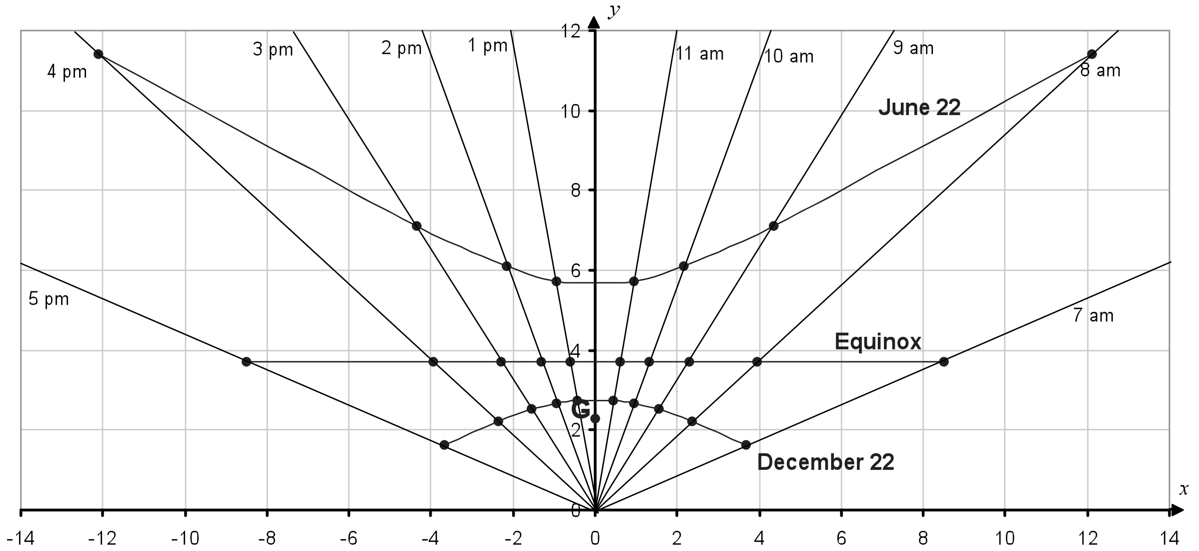


Figure 12.111: Hour lines and the positions of the shadow tip during the day for the equinoxes and the summer and winter solstices

Above we can see the horizontal line of the both equinox (that are obviously superposed) and the both hyperbolas of solstices.

Here is a Python™ code to get the above graph:

```python
import numpy as np
import matplotlib.pyplot as plt

# Geographical and geometrical parameters
L = np.radians(37.8)   # Latitude (converted to radians)
h = 1.80               # Height of the stick in meters
Gx, Gy = 0, 0          # Position of the stick centered at G = (0, 0
    )
```

```
D_vals = [0, 23.5, -23.5]  # Declinations (Equinox, June 22, December 22)
D_labels = ["Equinox", "June 22", "December 22"]  # Labels for each declination
colors = ["gray", "lightgray", "black"]  # Colors for the curves

# Time from 7 AM to 5 PM (hours)
hours = np.arange(7, 18)
T = np.radians(15 * (12 - hours))  # Hour angle in radians (15 degrees per hour)

# Create the figure for plotting
plt.figure(figsize=(14, 6))
ax = plt.gca()  # Get the current axes object

# Dictionary to store coordinates for each shadow curve
all_coords = {}

# Loop through each declination to plot the shadow curves
for D_deg, label, color in zip(D_vals, D_labels, colors):
    D = np.radians(D_deg)  # Convert declination to radians
    x_vals, y_vals = [], []  # Lists to store x and y coordinates of the shadow tips

    # Loop over all hour angles (T) to calculate the shadow positions
    for t in T:
        denominator = np.cos(t) * np.cos(L) - np.tan(D) * np.sin(L)
        x = h * np.sin(t) / denominator  # X coordinate of the shadow tip
        y = h * np.cos(t) / (np.sin(L) * denominator)  # Y coordinate of the shadow tip

        # Append the coordinates to the lists
        x_vals.append(Gx + x)
        y_vals.append(Gy + y)

    # Limit the curve for June 22nd between 8 AM and 4 PM
    if label == "June 22":
        idx_range = (hours >= 8) & (hours <= 16)  # Select hours between 8 AM and 4 PM
        x_plot = np.array(x_vals)[idx_range]
        y_plot = np.array(y_vals)[idx_range]
        all_coords[label] = (x_vals, y_vals)  # Store the coordinates
    else:
        x_plot = x_vals  # Use all values for the other dates
        y_plot = y_vals
        all_coords[label] = (x_vals, y_vals)

    # Plot the shadow curve for this declination
```

```python
    plt.plot(x_plot, y_plot, label=label, lw=2.5, color=color)

# Find the indices of 5 PM and 7 AM on the Equinox curve
idx_5pm = hours.tolist().index(17)  # Index of 5 PM
idx_7am = hours.tolist().index(7)   # Index of 7 AM

# Plot hour lines and intersections with the shadow curves
for i in range(len(T)):
    # Check if we are at 5 PM or 7 AM
    if i == idx_5pm or i == idx_7am:
        # Extend the line until it intersects the Equinox curve
        x_max = all_coords["Equinox"][0][i]
        y_max = all_coords["Equinox"][1][i]
        # Extend the line by a factor of 1.5
        dx = x_max - Gx
        dy = y_max - Gy
        norm = np.sqrt(dx**2 + dy**2)  # Normalize the vector to extend the line
        x_max = Gx + 15 * dx / norm  # Scale the x component
        y_max = Gy + 15 * dy / norm  # Scale the y component
    else:
        # Extend the hour lines to intersect with the June 22nd curve
        x_max = all_coords["June 22"][0][i]
        y_max = all_coords["June 22"][1][i]

    # Plot the line from the stick (G) to the shadow tip
    plt.plot([Gx, x_max], [Gy, y_max], color='black', lw=0.5)

    # Add a point at each intersection for each declination curve
    for label in D_labels:
        if i < len(all_coords[label][0]):  # Safety check for June 22
            x_i = all_coords[label][0][i]
            y_i = all_coords[label][1][i]
            plt.plot(x_i, y_i, 'ko', markersize=3)  # Plot the intersection point

    # Label the hours along the Equinox curve
    dx = all_coords["Equinox"][0][i] - Gx  # Difference in x between G and the shadow tip
    dy = all_coords["Equinox"][1][i] - Gy  # Difference in y between G and the shadow tip
    norm = np.sqrt(dx**2 + dy**2)  # Normalize the vector
    hour = hours[i]
    label_hr = f"{hour} am" if hour < 12 else ("12 pm" if hour == 12 else f"{hour - 12} pm")  # Hour label
    label_x = all_coords["Equinox"][0][i] + 0.4 * dx / norm  # Position label along the curve
    label_y = all_coords["Equinox"][1][i] + 0.4 * dy / norm  # Position label along the curve
    plt.text(label_x, label_y, label_hr, fontsize=8, ha='center')  # Plot the hour label

# Plot the position of the stick (G)
plt.plot(Gx, Gy, 'ko')  # Mark the position of G with a black dot
plt.text(Gx + 0.2, Gy, 'G', fontsize=9)  # Label the position of G

# Final plot settings
plt.xlabel('x')
plt.ylabel('y')
plt.title('Shadow tip positions (L = 37.8)')
plt.legend()  # Show the legend
plt.grid(True)  # Enable grid
ax.set_aspect('equal')  # Set equal scaling on both axes
plt.xlim(-13, 13)  # Set x-axis limits
plt.ylim(-1, 12)  # Set y-axis limits
plt.tight_layout()  # Adjust layout to avoid clipping
plt.show()  # Show the plot
```

That gives:

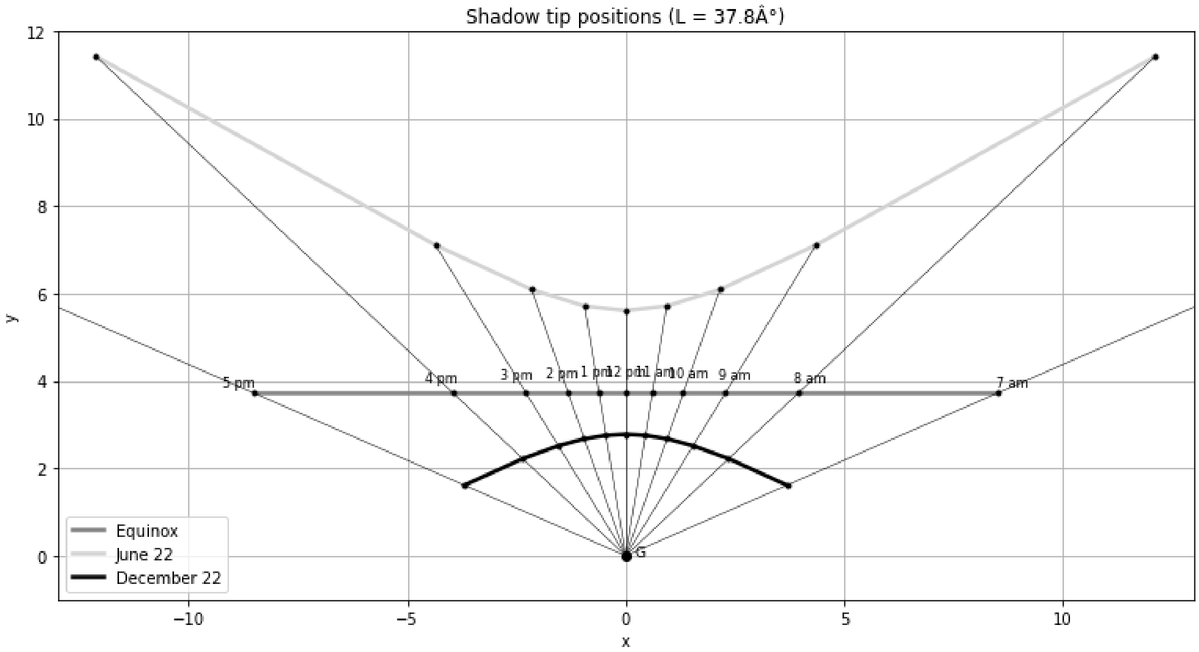


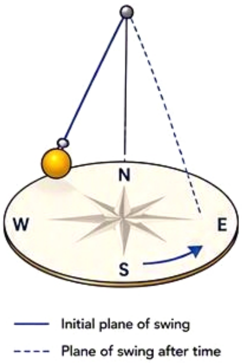
Figure 12.112: Hour lines and the positions of the shadow tip during the day for the equinoxes and the summer and winter solstices in Python™

Obviously our results match with reality because the Earth isn't flat, rotates on itself and "turns" around the Sun...

1 THE FOUCAULT PENDULUM

A long pendulum swings in a fixed plane in inertial space. As Earth rotates beneath it, the plane of swing appears to rotate.

- Clockwise in the Northern Hemisphere
- Counterclockwise in the Southern Hemisphere

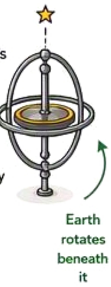

N
W
E
S
Initial plane of swing
Plane of swing after time

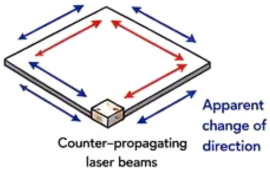
✓ Matches the prediction for a rotating Earth.

2 THE GYROSCOPE (ESPECIALLY GYRO LASERS)

A gyroscope maintains its orientation in inertial space. As Earth rotates, the gyroscope's axis appears to change direction relative to the ground.

- Mechanical gyros show a drift consistent with Earth's rotation.
- Ring laser gyros use two counter-propagating laser beams. A rotation of the system produces a frequency difference (Sagnac effect).


Earth rotates beneath it


Apparent change of direction
Counter-propagating laser beams

$$\Delta f = \frac{4A\Omega}{\lambda P}$$

where:
A = area enclosed by ring
P = perimeter of ring
Ω = Earth's rotation rate
λ = laser wavelength

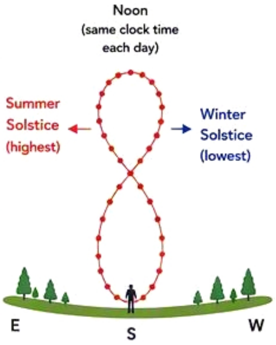
✓ Gyro lasers measure Earth's rotation rate with extraordinary precision (used in navigation and geodesy).

3 THE ANALEMMA

An analemma is the figure traced by the Sun in the sky when photographed at the same clock time each day over a year.

The figure-eight shape arises only because:

1. Earth rotates on its axis (giving day and night)
2. Earth orbits the Sun in an elliptical orbit whose plane is tilted.


Noon
(same clock time each day)
Summer Solstice (highest)
Winter Solstice (lowest)
E
S
W

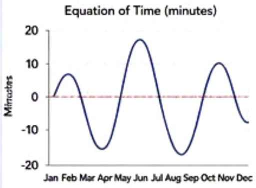
✓ Explained only if Earth rotates on its axis while orbiting the Sun.

4 THE EQUATION OF TIME (TIME EQUATION)

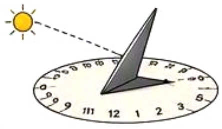
The equation of time is the difference between apparent solar time (from the Sun) and mean solar time (uniform 24-hour days).

It has two causes that occur only if:

1. Earth rotates on its axis (defining day length)
2. Earth revolves around the Sun in an elliptical orbit with a tilted axis.



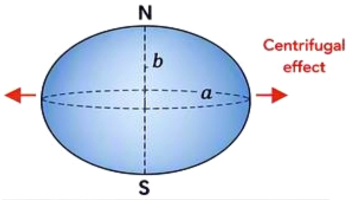
How it enables sundials:



✓ Works only if Earth rotates on its axis while orbiting the Sun.

5 EARTH'S EQUATORIAL BULGE

Earth is not a perfect sphere. It is an oblate spheroid—wider at the equator—because Earth's rotation produces centrifugal force that pushes matter outward.


N
b
a
S
Centrifugal effect

Typical values:

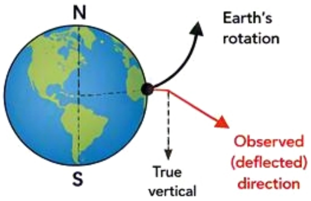
Equatorial radius	$a \approx 6378.1$ km
Polar radius	$b \approx 6356.8$ km
Flattening	$f = \frac{a-b}{a} \approx \frac{1}{298.257}$

✓ Such a bulge exists only if Earth rotates on its axis.

6 THE EÖTVÖS EFFECT

A moving mass (like Earth) "drags" local spacetime, changing the apparent direction of gravity.

On Earth's surface, this produces a tiny but measurable deflection toward the direction of Earth's rotation.

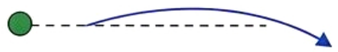

N
S
Earth's rotation
True vertical
Observed (deflected) direction

✓ The deflection reverses in the Southern Hemisphere. It occurs only because Earth rotates.

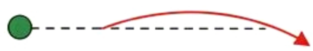
7 THE CORIOLIS EFFECT

In a rotating frame (like Earth), moving objects are deflected from a straight path.

Northern Hemisphere



Southern Hemisphere



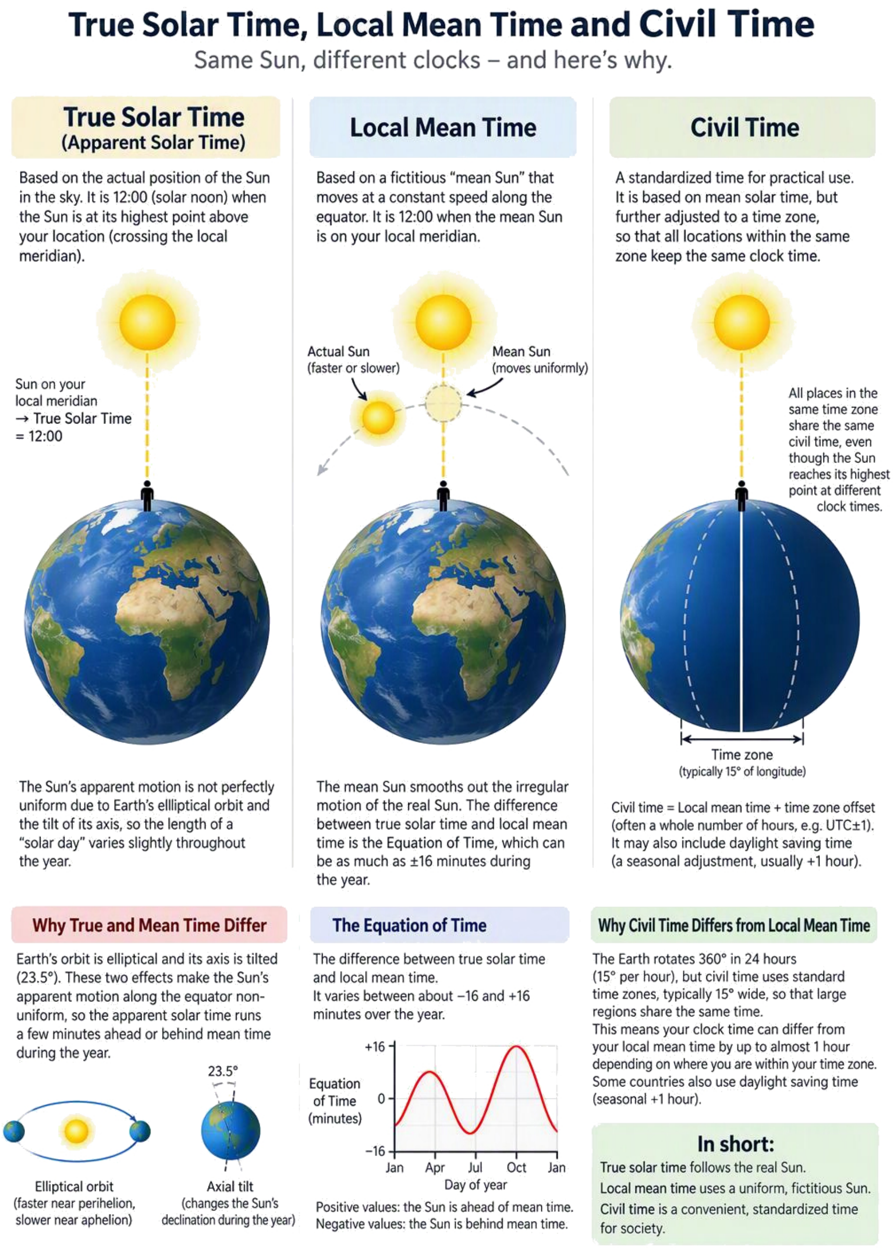
✓ Arises only in a rotating Earth. Its direction reverses between hemispheres.

12.1.12 Apparent solar time (true solar time)

Now that we know what is the equation of time and what are sundials and how to read their shadow we can finally deal with the concepts of true solar time, civil time and local mean time.

Definition (#12.1.530): "Apparent solar time" (AST), also named "true solar time" (TST), is based directly on the observable position of the Sun (actually a G2V-type star, aka yellow dwarf) at a specific location, where solar noon occurs at the exact moment the Sun reaches its highest point in the local sky, causing true solar days to fluctuate slightly in length throughout the year due to Earth's elliptical orbit and axial tilt.

Definition (#12.1.531): "Civil time" (CT), by contrast, is a standardized artificial timekeeping system created for practical daily life that divides the globe into fixed time zones and uses an idealized, mathematically averaged "mean Sun" so that every clock day consists of exactly 24 uniform hours.



True Solar Time, Local Mean Time and Civil Time

Same Sun, different clocks – and here's why.

True Solar Time (Apparent Solar Time)

Based on the actual position of the Sun in the sky. It is 12:00 (solar noon) when the Sun is at its highest point above your location (crossing the local meridian).

The Sun's apparent motion is not perfectly uniform due to Earth's ellliptical orbit and the tilt of its axis, so the length of a "solar day" varies slightly throughout the year.

Local Mean Time

Based on a fictitious "mean Sun" that moves at a constant speed along the equator. It is 12:00 when the mean Sun is on your local meridian.

The mean Sun smooths out the irregular motion of the real Sun. The difference between true solar time and local mean time is the Equation of Time, which can be as much as ±16 minutes during the year.

Civil Time

A standardized time for practical use. It is based on mean solar time, but further adjusted to a time zone, so that all locations within the same zone keep the same clock time.

All places in the same time zone share the same civil time, even though the Sun reaches its highest point at different clock times.

Civil time = Local mean time + time zone offset (often a whole number of hours, e.g. UTC±1). It may also include daylight saving time (a seasonal adjustment, usually +1 hour).

Why True and Mean Time Differ

Earth's orbit is elliptical and its axis is tilted (23.5°). These two effects make the Sun's apparent motion along the equator non-uniform, so the apparent solar time runs a few minutes ahead or behind mean time during the year.

The Equation of Time

The difference between true solar time and local mean time. It varies between about −16 and +16 minutes over the year.

Positive values: the Sun is ahead of mean time.
Negative values: the Sun is behind mean time.

Why Civil Time Differs from Local Mean Time

The Earth rotates 360° in 24 hours (15° per hour), but civil time uses standard time zones, typically 15° wide, so that large regions share the same time. This means your clock time can differ from your local mean time by up to almost 1 hour depending on where you are within your time zone. Some countries also use daylight saving time (seasonal +1 hour).

In short:

True solar time follows the real Sun.
Local mean time uses a uniform, fictitious Sun.
Civil time is a convenient, standardized time for society.

In other words, the apparent solar time (true solar time) is based on the apparent motion of the actual Sun. It is based on the apparent solar day, the interval between two successive returns of the Sun to the local meridian (see definition below)! Apparent solar time (true solar time) can also be crudely measured by a simple sundial. That's why time provided by sundials differ much of the time from time given by watches and clocks because the latter don't give the true solar time but instead an artificial convenient political and business time!

Definition (#12.1.532): The "local meridian" is an imaginary semicircle in the sky that stretches from due North to due South, passing directly overhead through the observer's zenith - the point straight above you. It is part of the celestial coordinate system and acts like a reference line for tracking celestial objects.

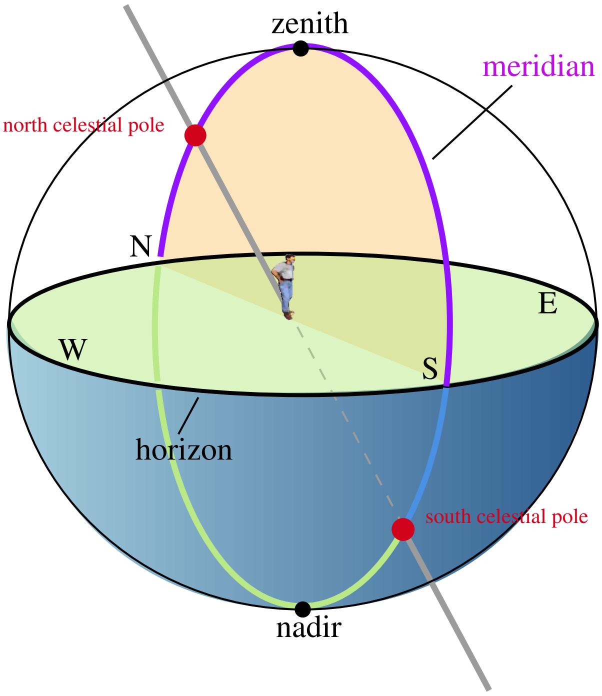


Figure 12.113: **Illustration of the concept of local meridian**

The length of a solar day varies through the year, and the accumulated effect produces seasonal deviations of up to 16 minutes from the mean as we have prove it earlier above during our derivation of the Equation of Time (see the maxima and minima of the plot of this equation). The effect has two main causes as we know: First, due to the eccentricity of Earth's orbit, Earth moves faster when it is nearest the Sun (perihelion) and slower when it is farthest from the Sun (aphelion) (see Kepler's laws of planetary motion at page 5233). Second, due to Earth's axial tilt (obliquity of the ecliptic), the Sun's annual motion is along a great circle (the ecliptic) that is tilted to Earth's celestial equator.

Our purpose here will be to learn how to compute the true solar time base on the (political) civil time with a detailed practical case.

Converting civil time to true time requires adjusting:

1. The eventual daylight saving time (DST)
2. The political time zone (TZ) if it doesn't correspond to the meridian

3. The very local longitude's offset from the time zone meridian (VLT)
4. Equation of Time (EoT) correction for the Earth's orbital eccentricity and axial tilt

So let's apply this to a companion example :

We want to know the true solar time of the civil time 2:00 PM (14:00) Gulf Standard Time (GST) on January 15 in Abu Dhabi (United Arab Emirates).

First, as far as we know, Abu Dhabi doesn't apply the daylight saving time (DST). Therefore we have:

$$\Delta\text{DST}_{\text{Abu Dhabi}} = 0 \tag{12.1.715}$$

Second, Abu Dhabi is really on a time zone that matches the political time zone. Hence:

$$\Delta\text{TZ}_{\text{Abu Dhabi}} = 0 \tag{12.1.716}$$

Third, Abu Dhabi is located West of the UTC+4 standard reference meridian, meaning the Sun (actually a G2V-type star, aka yellow dwarf) reaches its highest point there later than the time zone meridian dictates.

As for Abu Dhabi:

- Local Longitude $(\lambda_{\text{local long.}}): 54.37°\text{E}$
- GST Standard Meridian $(\lambda_{\text{GST std.}}): 60°\text{E}$

Because the Earth rotates at $15°$ per hour (or 4 minutes of time per degree of longitude), the very local time offset from the Gulf Standard Time meridian (GST) is:

$$\begin{aligned}\Delta\text{VLT}_{\text{long.}} &= (\lambda_{\text{local long.}} - \lambda_{\text{GST std.}}) \cdot 4\text{ min}/° \\ &= (54.37° - 60°) \cdot 4 = -22.52\text{ minutes } (-22\text{ min}31\text{ sec})\end{aligned} \tag{12.1.717}$$

Hence so far we have for local mean time (LMT):

$$\begin{aligned}\text{LMT} &= \text{CT} + \Delta\text{DST}_{\text{Abu Dhabi}} + \Delta\text{TZ}_{\text{Abu Dhabi}} + \Delta\text{VLT}_{\text{long.}} \\ &= \text{CT} + \Delta\text{DST}_{\text{Abu Dhabi}} + \Delta\text{TZ}_{\text{Abu Dhabi}} + (\lambda_{\text{local long.}} - \lambda_{\text{GST std.}}) \cdot 4\text{ min}/° \\ &= 14:00:00 + 0 + 0 - 00:22:31 = 13:37:29\end{aligned} \tag{12.1.718}$$

While LMT assumes a perfectly circular orbit, true solar time tracks the actual variable speed of the Sun across the sky. On January 15, the Earth is near perihelion, causing the actual Sun to be "slow" relative to the mean clock.

Let us remind us the expression of the Equation of Time (EoT) proven earlier above:

$$\text{Equation of Time} \cong -594.3\sin(2(\hat{M} + 12.25°)) - 459.3\sin(\hat{M})\ [\text{s}] \tag{12.1.719}$$

Hence in minutes:

$$\text{Equation of Time} \cong -9.905\sin(2(\hat{M} + 12.25°)) - 7.655\sin(\hat{M})\ [\text{min}] \tag{12.1.720}$$

with for the 15th january ($d=15$):

$$\hat{M}=\cos\left(\frac{360}{365}(d+10)\right)\cong 0.9088 \tag{12.1.721}$$

where the reader must not forget to convert the content of the parenthesis into radians first!

Hence (converting $12.25°$ into radians):

$$\begin{aligned}\text{EoT}&\cong -9.905\sin(2(0.9088+0.214))-7.655\sin(0.9088)\\&\cong -13.77\text{ minutes }(\;-13\,\text{min}\,46\,\text{sec}\;)\end{aligned} \tag{12.1.722}$$

True solar time (TST) is the sum of the Local Mean Time (LMT) and the Equation of Time:

$$\text{TST}=\text{LMT}+\text{EoT}=13:37:29-00:13:46=13:23:43 \tag{12.1.723}$$

The result is not bad as online calculators give 13:37:26.

When standard clocks (civil time) in Abu Dhabi strike 2:00 PM on January 15 (14:00), the actual solar time is 13:23:43 (depending on the approximations, following the textbooks and the teachers this result can very from 5 to 10 minutes). Consequently, solar noon (12:00:00 TST) on this date occurred nearly half an hour prior to standard noon, exactly at 12:31:53 PM GST.

Finally:

$$\boxed{\begin{aligned}&\text{TST}=\text{LMT}+\text{EoT}=\text{CT}+\Delta\text{DST}+\Delta\text{TZ}+\Delta\text{VLT}_{\text{long.}}+\text{EoT}\\&\quad\cong\text{CT}+\Delta\text{DST}+\Delta\text{TZ}+(\lambda_{\text{local long.}}-\lambda_{\text{GST std.}})\cdot 4\text{ min}/°-9.9\sin(2(\hat{M}+12.25°))-7.65\sin(\hat{M})\end{aligned}} \tag{12.1.724}$$

where (the reader must not forget to convert the content of the parenthesis into radians first!):

$$\hat{M}=\cos\left(\frac{360}{365}(d+10)\right) \tag{12.1.725}$$

12.1.13 Celestial sphere and coordinate systems (positional astronomy)

Definition (#12.1.533): "Azimuthal coordinates", also known as "horizontal coordinates", are a celestial coordinate system used to specify the positions of an object in the sky relative to an observer's horizon. They consist of two angles: "azimuth" (A) and "altitude" (the latter being also named "elevation" h). The azimuth is the angle measured horizontally along the horizon while the altitude (elevation) is the angle measured vertically upwards from the horizon.

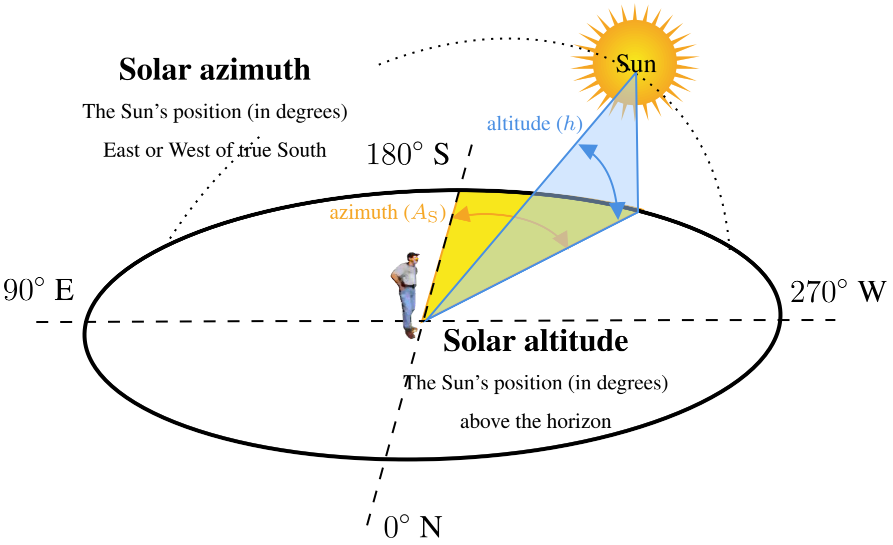


Figure 12.114: **Azimuth and Altitude depicted**

Remark

According to the **old** nomenclature, the azimuth A_S was defined starting from the South towards the West, from $0°$ to $360°$ (SWNE direction). Subsequently, the International Astronomical Union (IAU) changed the definition: Azimuth A_N is now defined from the North towards the East, from $0°$ to $360°$, (NESW direction):

$$A_\text{S} = A_\text{N} + \pi \qquad (12.1.726)$$

or obviously in degrees:

$$A_\text{S} = A_\text{N} + 180° \qquad (12.1.727)$$

Definition (#12.1.534): "Equatorial coordinates" are a celestial coordinate system used to specify the positions of an object in the sky. They are based on the projection of Earth's equator and poles onto the celestial sphere, making them independent of the observer's location and time of observation. The two main components of equatorial coordinates consist of two angles: "right ascension" (α) and "declination" (δ). Right Ascension (α) is like longitude but projected onto the sky, measured in hours, minutes, seconds starting from the first point of Aries (this is the point where the ecliptic - the Sun's apparent path - intersects the celestial equator - an extension of Earth's equator onto the celestial sphere) and Declination (δ) is like latitude in the sky, measured in degrees North or South of the celestial equator.

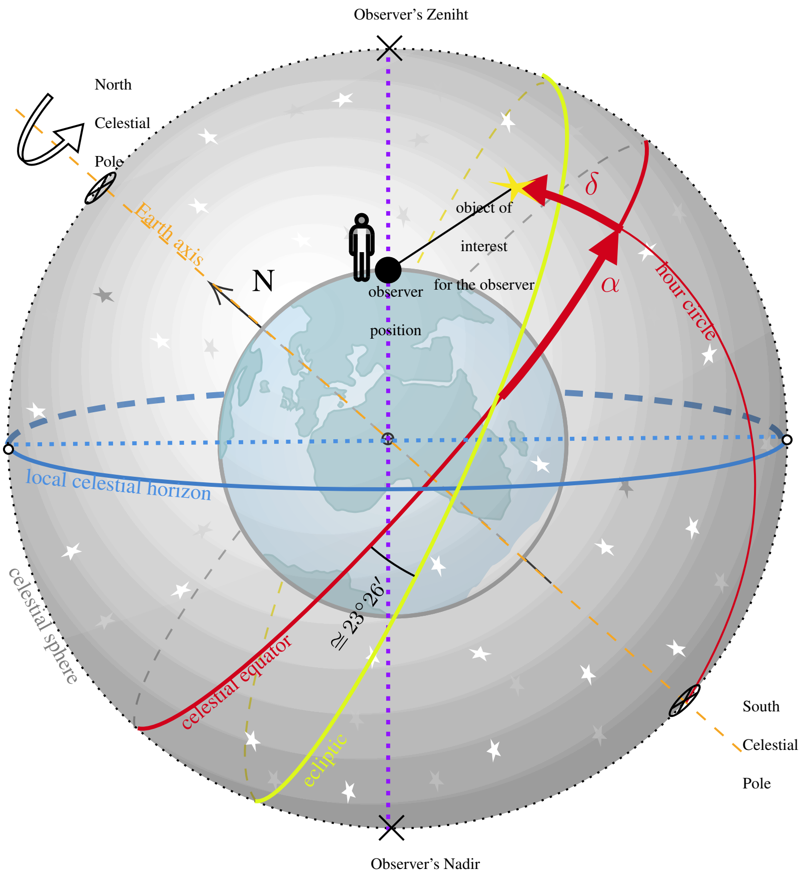

Observer's Zeniht
North
Celestial
Pole
Earth axis
N
object of
interest
for the observer
observer
position
δ
α
hour circle
local celestial horizon
celestial sphere
≅ 23°26′
celestial equator
ecliptic
South
Celestial
Pole
Observer's Nadir

Figure 12.115: **Right ascension and Declination depicted**

Knowing how to convert between azimuth and equatorial coordinates is important for accurately tracking celestial objects and understanding their positions in the sky. Azimuth (A)/altitude (h) coordinates are observer-centric, meaning they change based on your location and time, while equatorial coordinates right ascension (α)/declination (δ) are fixed with respect to the stars. Converting between these systems allows astronomers to pinpoint object locations regardless of the observer's position and time, and to relate observations made at different times and location.

To convert from azimuth and altitude (horizontal) coordinates to equatorial coordinates (right ascension and declination), we need to know obviously the observer's latitude and the local sidereal time. The conversion involves solving a spherical triangle and using trigonometric equations. Software like Stellarium can also perform this conversion.

The conversion involves a spherical triangle formed by the north celestial pole, the zenith (point directly overhead), and the object in the sky.

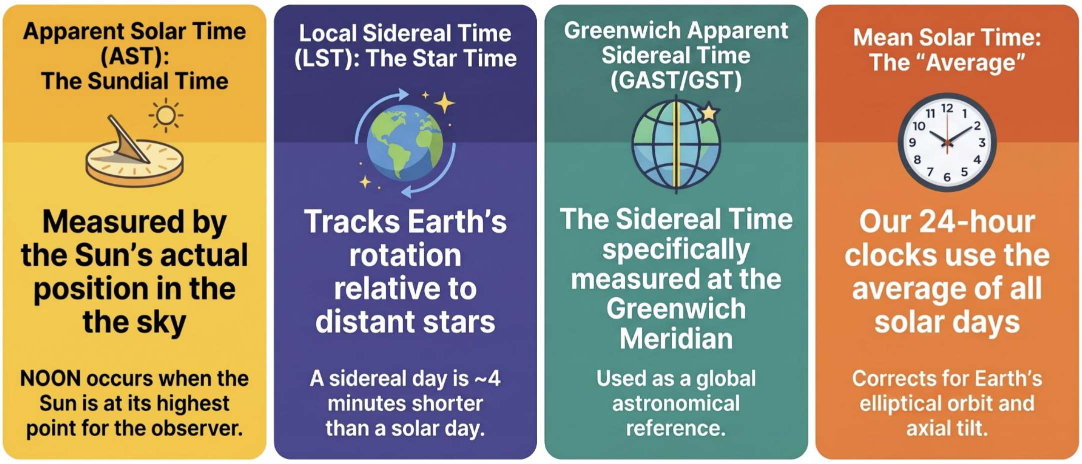
In both cases we need as input or get as output the following elements:

- Observer's latitude (φ)
- Local Sidereal Time (LST) (see definition further below)
- Right ascension (α) and declination (δ) of the object

- Hour angle (denoted depending on the context h, or τ or $\hat{\mathbf{P}}$) of the object (see definition further below)
- Azimuth (A_S) as taken from the South cardinal and altitude (denoted traditionally α) of the object

12.1.13.1 Local Sidereal Time

Definition (#12.1.535): The "Local Sidereal Time" (LST) is a celestial clock used especially by professional astronomers that tracks the Earth's rotation relative to the distant stars, rather than the Sun and it is obviously completely related to the concept of "sideral period" as defined and illustrated at page 5390 (aka sidereal day). It tells us which right ascension is currently crossing our local meridian (see definition page 5462), which is key for determining the best time to observe specific stars or deep-space objects. Unlike True Solar Time (TST), LST runs slightly faster, completing a full cycle in about 23 hours, 56 minutes and 4.0905 seconds (see the illustration at page 5390 to see why). For any location, it's calculated from Greenwich Sidereal Time (GST) adjusted by the local longitude, making it the astronomer's go-to for accurate telescope alignment and sky mapping.


Apparent Solar Time (AST): The Sundial Time
Measured by the Sun's actual position in the sky
NOON occurs when the Sun is at its highest point for the observer.
Local Sidereal Time (LST): The Star Time
Tracks Earth's rotation relative to distant stars
A sidereal day is ~4 minutes shorter than a solar day.
Greenwich Apparent Sidereal Time (GAST/GST)
The Sidereal Time specifically measured at the Greenwich Meridian
Used as a global astronomical reference.
Mean Solar Time: The "Average"
12
1
2
3
4
5
6
7
8
9
10
11
Our 24-hour clocks use the average of all solar days
Corrects for Earth's elliptical orbit and axial tilt.

Viewed from the same location, a star seen at one position in the sky will be seen approximately (excepted for some high velocity stars) at the same position on another night at the same time of day (or night), if the day is defined as a sidereal day. This is similar to how the time kept by a sundial (True Solar Time) can be used to find the location of the Sun. Just as the Sun and Moon appear to rise in the East and set in the West due to the rotation of Earth, so do the stars. Both True Solar Time and Sidereal Time make use of the approximate regularity of Earth's rotation about its polar axis: True Solar Time is reckoned according to the position of the Sun in the sky while Local Sidereal Time is based approximately on the position of the fixed stars on the theoretical celestial sphere.

To calculate the Local Sidereal Time (LST) at a location whose longitude L is known, and at a given day d of the year and hour H (solar) of observation:

- We use the Local Sidereal Time at the Greenwich meridian at a given arbitrary reference time. The ephemeris provides this type of information. Otherwise, we can use the fact that, the January 1, 2000 (gregorian calendar) at 12:00 UTC (nice round modern time), sidereal time at the Greenwich Meridian was 18.697374558 hours. This value used since as reference comes from:

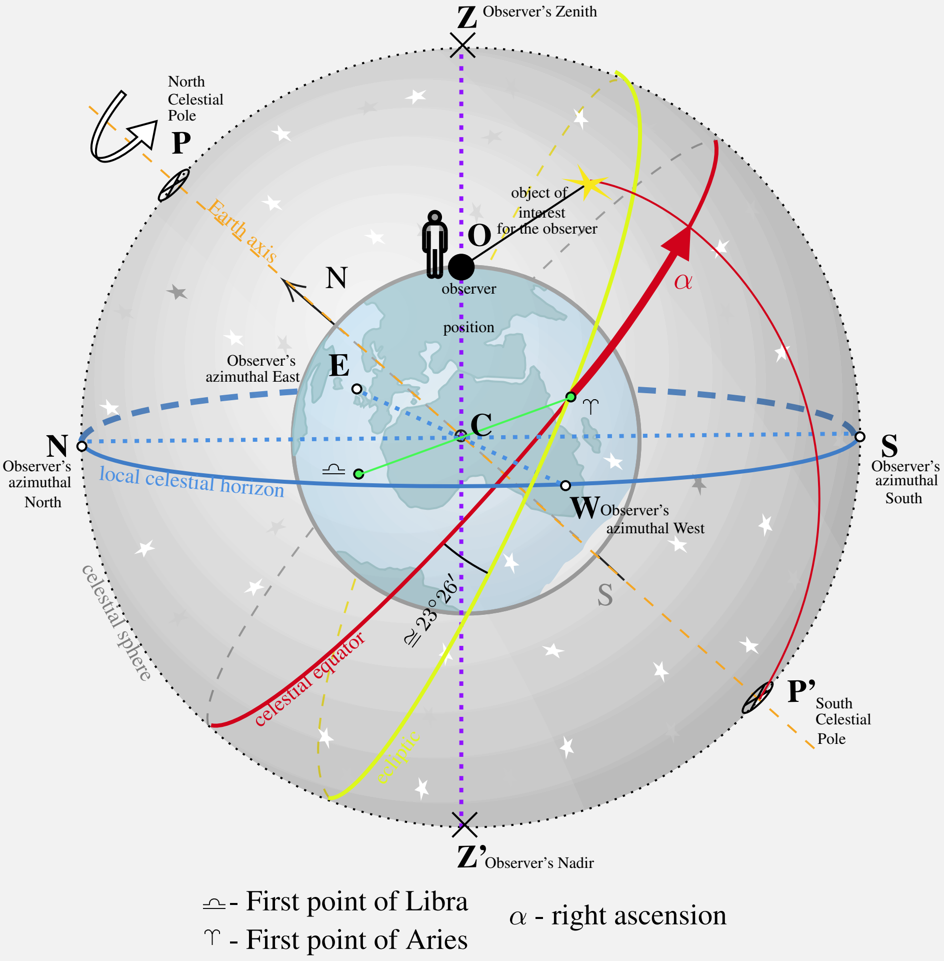
$$\frac{280.46061837^\circ}{15^\circ/\text{h}}=18.697374558\text{ h}. \tag{12.1.728}$$

But obviously you ask yourself where 280.46061837° comes from?!

That angle corresponds for the chosen reference date and time, that Greenwich happened to be pointing 280.46061837° away from the mean equinox.

Remark

The first point of Aries (vernal (spring) equinox), and the first point of Libra (autumnal equinox) are illustrated below:


Z Observer's Zenith
North Celestial Pole
P
Earth axis
object of interest for the observer
O
observer position
N
α
Observer's azimuthal East
E
C
♈
N
Observer's azimuthal North
local celestial horizon
♎
W Observer's azimuthal West
S
Observer's azimuthal South
S
≅ 23°26′
celestial sphere
celestial equator
ecliptic
P' South Celestial Pole
Z' Observer's Nadir
♎ - First point of Libra
♈ - First point of Aries
α - right ascension

By definition the "vernal equinox" is the point where the Sun (actually a G2V-type star, aka yellow dwarf) crosses the celestial equator going northward in March. Historically, astronomers chose the vernal equinox as the zero point of right ascension: $\alpha=0^\circ$.

The reader must remember the Earth's rotation is not fixed! The gravitational effects of the Sun and Moon cause a precession (a slow 26,000-year wobble, and a nutation (smaller periodic oscillations). As a result, the equinox is constantly shifting on the celestial sphere. The rate of precession is about $50.29''$ per year or about 1° every 72 years.

So there is no single permanent location of the equinox!

Astronomers therefore distinguish:

- The "mean equinox" where only the smooth precession motion is included and the short-term nutation oscillations are removed.
- The "true equinox" that includes the precession and the nutation so it wiggles slightly around the mean position.

he difference is called the equation of the equinoxes and can reach roughly ± 17 arcseconds.

- We calculate how much time (in solar days and fractions of a solar day) separates the observed instant from the reference instant. Let d be this duration. Each solar day having a duration of 24.06570982441908 sidereal hours, it follows that the sidereal time GST at the Greenwich meridian at the observed instant is, in sidereal hours:

$$\text{GST} \cong 18.697374558 + 24.06570982441908 \cdot d \tag{12.1.729}$$

- To the local Greenwich sidereal time GST at the observed instant is added a time difference due to longitude, namely $L/15$ where L is given in decimal degrees, counted positively when moving towards the East and negatively towards the West. In fact, the 360° of a complete revolution are covered in 24 hours, or 15° per hour:

$$\text{LST} = (\text{GST} + L/15) \text{ modulo } 24 \tag{12.1.730}$$

where for reminder "modulo 24" means that the final result is brought back between 0 hour and 24 hours by adding or differing from an integer multiple of 24 hour.

All these information may be quite abstract. Let us see a use case!

✍**Example:**

An observer is at geographical latitude 50° N and longitude 10° E, on 1991-05-19 (gregorian calendar) at 13:00 UTC. What is the corresponding sidereal time?

We first compute the number of days between the 1991-05-19 13:00 UTC (gregorian calendar) and 2000-01-01 12:00 UTC (gregorian calendar) that is simply equal to $d \cong 3148.958$.

Substituting into the GST formula gives:

$$\text{GST} \cong 18.697374558 + 24.06570982441908 \cdot d \cong 4.779\,\text{h}. \tag{12.1.731}$$

Since the observer is at longitude:

$$\lambda = 10^\circ\mathrm{E} \qquad (12.1.732)$$

the Local Sidereal Time is therefore:

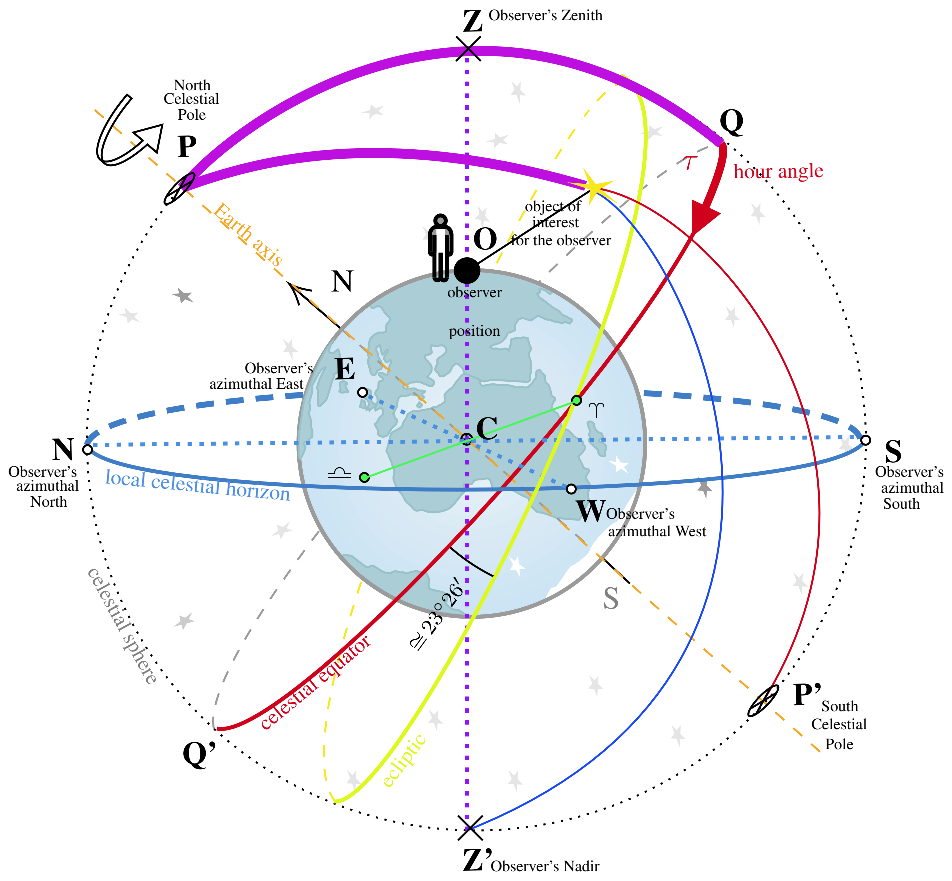
$$\mathrm{LST} = \mathrm{GST} + \frac{\lambda}{15} \cong 4.779 + \frac{10}{15} \cong 5.446\mathrm{h}. \qquad (12.1.733)$$

Converting to hours, minutes, and seconds:

$$5.446\,\mathrm{h.} = 5\mathrm{h}26\mathrm{m}47.55\mathrm{s} \qquad (12.1.734)$$

As 24h. is equal to 360° the corresponding sidereal time is therefore 81.7°.

Definition (#12.1.536): The "hour angle" (denoted h, τ or $\hat{\mathrm{P}}$ depending on the context) of a celestial object is the angular distance between the observer's local meridian and the object's current position in the sky (more precisely its "hour circle"), measured westward along the celestial equator. It essentially tells us how long it's been since the object last crossed your meridian (the imaginary line running from north to south through your zenith). It is usually measured in hours, minutes and secondes (form 0h. to 24h.) or converted into degrees where 1 hour $= 15^\circ$ ($360^\circ = 24$ hours).


Z Observer's Zenith
North Celestial Pole
P
Q
hour angle
τ
object of interest for the observer
O
observer position
Earth axis
N
Observer's azimuthal East
E
C
♈
♎
N
Observer's azimuthal North
local celestial horizon
S
Observer's azimuthal South
W Observer's azimuthal West
S
≅23°26′
celestial sphere
celestial equator
ecliptic
Q'
P' South Celestial Pole
Z' Observer's Nadir

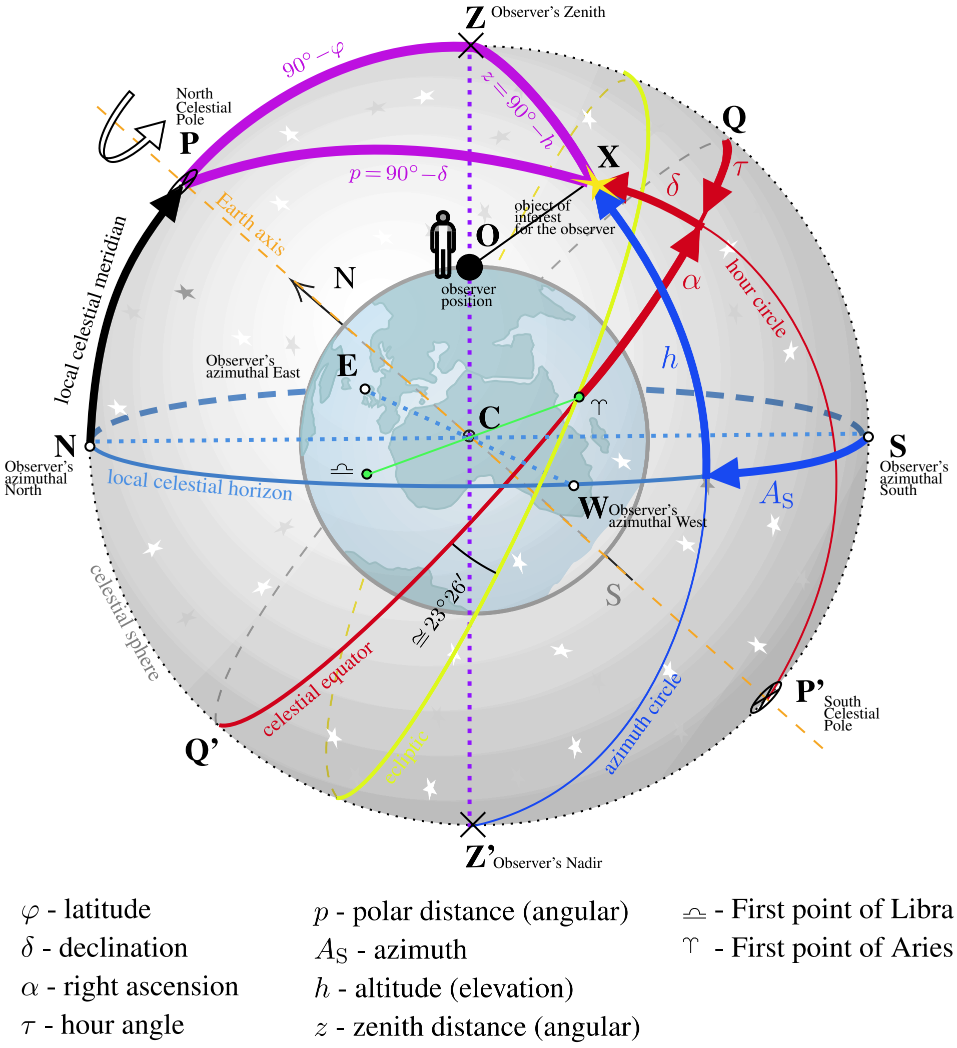
For example if the hour angle is 0° the object is right on the local meridian. If the hour angle is negative then the object is East of the meridian and rising, if the hour angle is positive then the object is West of the meridian and setting.

For the definitions above the reader can refer to the figure further below. That may help...

Caution! The detailed proof that we will provide below leads to the famous formulas that you can found on Wikipedia or any other textbook on the topic of positional astronomy. Sadly Wikipedia and those same textbooks don't precise that the results are only approximations because to be able to use spherical trigonometry, we have to assume that the azimuthal observer is like the equatorial observer... at the center of the Earth! This gives a an error of 1.6% for the position of the Moon and 0.0045% and even less for objects further away. This error seems to be negligible however... even for amateur astronomers the approximation is problematic for the observation of occultation of planets.

12.1.13.2 Equatorial coordinates to Azimuthal (horizontal) coordinates

Although the positions of cosmic sources are almost always specified in celestial coordinates, for purposes of observation, it is generally necessary to convert to elevation and azimuth. The conversion formulas between hour angle and declination $(\hat{\mathbf{P}},\delta)$ and elevation and azimuth (h, A_{S}) can be derived by applying the sine and cosine rules for spherical triangles to the system.


Z Observer's Zenith
$90^\circ-\varphi$
$z=90^\circ-h$
North Celestial Pole
P
Q
X
$p=90^\circ-\delta$
δ
τ
object of interest for the observer
O
observer position
Earth axis
N
local celestial meridian
hour circle
α
h
Observer's azimuthal East
E
C
♈
S
Observer's azimuthal South
N
Observer's azimuthal North
local celestial horizon
♎
W Observer's azimuthal West
A_{S}
S
$\cong 23°26'$
celestial sphere
celestial equator
ecliptic
azimuth circle
P' South Celestial Pole
Q'
Z' Observer's Nadir
φ - latitude
δ - declination
α - right ascension
τ - hour angle
p - polar distance (angular)
A_{S} - azimuth
h - altitude (elevation)
z - zenith distance (angular)
♎ - First point of Libra
♈ - First point of Aries

Equatorial (red) and horizontal (blue) celestial coordinates

Figure 12.116: **Representation of Equatorial and Azimuthal coordinates**

To convert between the horizontal and equatorial coordinates for an object **X** , we use a spherical triangle often named the "astronomical triangle": Δ**XPZ**, where **Z** is the zenith, **P** is the North celestial pole, and **X** is the object of interest for the observer.

The sides of the purple triangle as visible in the figure above:

- $\overset{\frown}{\mathbf{PZ}}$ is the observer's co-latitude $=90^\circ-\varphi$
- $\overset{\frown}{\mathbf{ZX}}$ is the zenith distance of $=90^\circ-h$

- $\overset{\frown}{\mathbf{PX}}$ is the North polar distance of $90^\circ - \delta$

The angles of the purple triangle as visible in the figure above:

- The angle at **P** denoted $\hat{\mathbf{P}}$ (or τ), the hour angle of **X** (also often denoted H in the literature)
- The angle at **Z** denoted $\hat{\mathbf{Z}}$ is $360^\circ - A_\text{S}$, where A_S is the azimuth of X
- The angle at **X** denoted $\hat{\mathbf{X}}$ named the parallactic angle

We assume we know the observer's latitude φ and the Local Sidereal Time LST (LST may be obtained, if necessary, from Greenwich Sidereal Time and observer's longitude).

Given the right ascension (RA) denoted above α and declination δ, we have Local Hour Angle $\hat{\mathbf{P}} = \text{LST} - \text{RA}$, in hours. We need to convert $\hat{\mathbf{P}}$ to degrees by multiplying that value by 15. Given $\hat{\mathbf{P}}$ the hour angle and δ (declination) and φ (latitude), we require azimuth A_S and altitude h.

By the cosine formula (see section Trigonometry page 2544):

$$\cos(90^\circ - h) = \cos(90^\circ - \delta)\cos(90^\circ - \varphi) + \sin(90^\circ - \delta)\sin(90^\circ - \varphi)\cos(\hat{\mathbf{P}}) \quad (12.1.735)$$

which simplifies to:

$$\sin(h) = \sin(\delta)\sin(\varphi) + \cos(\delta)\cos(\varphi)\cos(\hat{\mathbf{P}}) \quad (12.1.736)$$

By the sines formula (see section Trigonometry page ??):

$$\frac{\sin(\hat{\mathbf{Z}})}{\sin(90^\circ - \delta)} = \frac{\sin(\hat{\mathbf{P}})}{\sin(90^\circ - h)} \quad (12.1.737)$$

As by construction $\hat{\mathbf{Z}} = 360^\circ - A_\text{S}$:

$$\frac{\sin(360^\circ - A_\text{S})}{\sin(90^\circ - \delta)} = \frac{\sin(\hat{\mathbf{P}})}{\sin(90^\circ - h)} \quad (12.1.738)$$

which simplifies to:

$$-\frac{\sin(A_\text{S})}{\cos(\delta)} = \frac{\sin(\hat{\mathbf{P}})}{\cos(h)} \quad (12.1.739)$$

Hence:

$$\sin(A_\text{S}) = -\frac{\cos(\delta)\sin(\hat{\mathbf{P}})}{\cos(h)} \quad (12.1.740)$$

So finally:

$$\boxed{\begin{aligned} h &= \arcsin\left\{\sin(\delta)\sin(\varphi) + \cos(\delta)\cos(\varphi)\cos(\hat{\mathbf{P}})\right\} \\ A_\text{S} &= \arcsin\left\{-\frac{\cos(\delta)\sin(\hat{\mathbf{P}})}{\cos(h)}\right\} \end{aligned}} \quad (12.1.741)$$

Or explicitly:

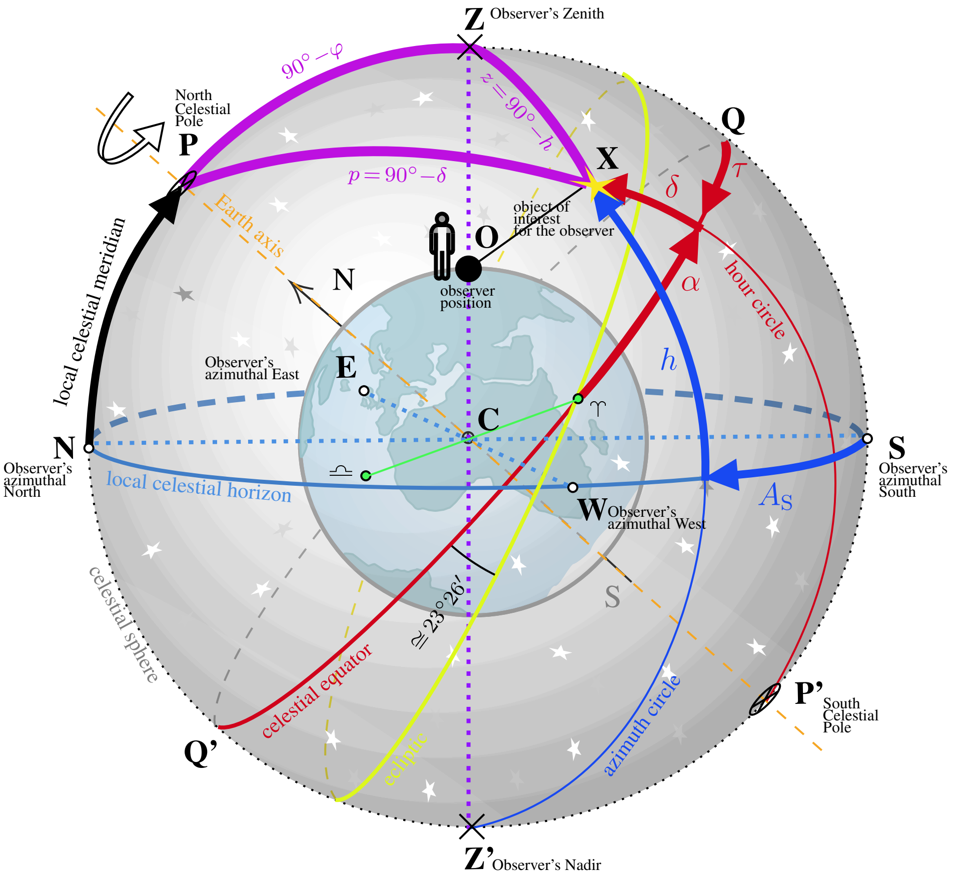
$$\begin{aligned}\text{Altitude angle} &= \arcsin\{\sin(\text{declination angle})\sin(\text{latitude}) \\ &\quad + \cos(\text{declination angle})\cos(\text{latitude})\cos(\text{hour angle})\} \\ \text{Azimuth angle} &= \arcsin\left\{-\frac{\cos(\text{declination angle})\sin(\text{hour angle})}{\cos(\text{altitude angle})}\right\}\end{aligned} \tag{12.1.742}$$

12.1.13.3 Azimuthal (horizontal) coordinates to Equatorial coordinates

Now let us deal with the inverse problem: azimuthal horizontal to equatorial coordinates!

In other words, given the altitude h, the azimuth A_S and the latitude φ what is the hour angle $\hat{\mathbf{P}}$ and the declination δ ?

Again we will use the following figure:


Z Observer's Zenith
90° − φ
z = 90° − h
North Celestial Pole
P
Q
X
p = 90° − δ
δ
τ
object of interest for the observer
O
observer position
Earth axis
N
α
hour circle
local celestial meridian
h
Observer's azimuthal East
E
C
♈
N
Observer's azimuthal North
local celestial horizon
♎
W Observer's azimuthal West
S
Observer's azimuthal South
A_S
S
celestial sphere
≃ 23°26′
celestial equator
P' South Celestial Pole
Q'
ecliptic
azimuth circle
Z' Observer's Nadir

φ - latitude
δ - declination
α - right ascension
τ - hour angle

p - polar distance (angular)
A_S - azimuth
h - altitude (elevation)
z - zenith distance (angular)

♎ - First point of Libra
♈ - First point of Aries

Equatorial (red) and horizontal (blue) celestial coordinates

By the cosine formula (see section Trigonometry page 2544):

$$\cos(90^\circ-\delta)=\cos(90^\circ-h)\cos(90^\circ-\varphi)+\sin(90^\circ-h)\sin(90^\circ-\varphi)\cos(\hat{\mathbf{Z}}) \tag{12.1.743}$$

which simplifies to (and also as by construction $\hat{\mathbf{Z}}=360^\circ-A_\text{S}$):

$$\sin(\delta)=\sin(h)\sin(\varphi)+\cos(h)\cos(\varphi)\cos(A_\text{S}) \tag{12.1.744}$$

By the sines formula (see section Trigonometry page **??**):

$$\frac{\sin(\hat{\mathbf{Z}})}{\sin(90^\circ-\delta)}=\frac{\sin(\hat{\mathbf{P}})}{\sin(90^\circ-h)} \tag{12.1.745}$$

As by construction $\hat{\mathbf{Z}}=360^\circ-A_\text{S}$:

$$\frac{\sin(360^\circ-A_\text{S})}{\sin(90^\circ-\delta)}=\frac{\sin(\hat{\mathbf{P}})}{\sin(90^\circ-h)} \tag{12.1.746}$$

which simplifies to:

$$-\frac{\sin(A_\text{S})}{\cos(\delta)}=\frac{\sin(\hat{\mathbf{P}})}{\cos(h)} \tag{12.1.747}$$

Hence:

$$\sin(\hat{\mathbf{P}})=-\frac{\cos(h)\sin(A_\text{S})}{\cos(\delta)} \tag{12.1.748}$$

So finally:

$$\boxed{\begin{aligned}\delta&=\arcsin\left\{\sin(h)\sin(\varphi)+\cos(h)\cos(\varphi)\cos(A_\text{S})\right\}\\ \hat{\mathbf{P}}&=\arcsin\left\{-\frac{\cos(h)\sin(A_\text{S})}{\cos(\delta)}\right\}\end{aligned}} \tag{12.1.749}$$

Or explicitly:

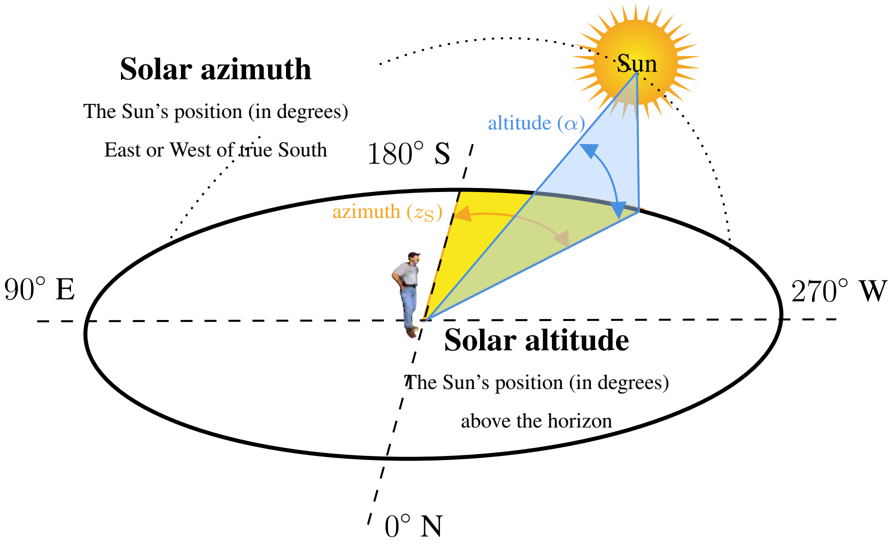
$$\begin{aligned}\text{Declination angle}&=\arcsin\{\sin(\text{altitude angle})\sin(\text{latitude})\\&\qquad+\cos(\text{altitude angle})\cos(\text{latitude})\cos(\text{azimuth})\}\\ \text{Hour angle}&=\arcsin\left\{-\frac{\cos(\text{altitude angle})\sin(\text{azimuth angle})}{\cos(\text{declination angle})}\right\}\end{aligned} \tag{12.1.750}$$

12.1.13.4 Apparent daily path of the Sun (solargraph)

Now that we have defined and studied what are the azimuth and the altitude we can study the path of the Sun (actually a G2V-type star, aka yellow dwarf) in the sky.

The purpose is to determine for an observer at a given latitude, longitude, day and time the real position of the Sun!

For this the main parameters to determine the real path of the Sun from an observer point of view (azimuth $\in [0^\circ, 360^\circ[$ and altitude $\in]-90^\circ, 90^\circ[$) and hence plot a solargraph are de latitude, the longitude, the declination and the hour angle.


Solar azimuth
The Sun's position (in degrees)
East or West of true South
Sun
altitude (α)
180° S
azimuth (z_S)
90° E
270° W
Solar altitude
The Sun's position (in degrees)
above the horizon
0° N

First, let us remind that the declination δ (see page 5431) is a function of the day of the year (d) and given by (remember to convert to parenthesis of the cos into radians to get the right result!):

$$\delta \cong -23.45^\circ \cdot \cos\left(\frac{360}{365}\cdot(d+10)\right) \tag{12.1.751}$$

and $\hat{\mathbf{P}}$ is the hour angle, a function of the time of the day given by (where AST is the Apparent Solar Time - also denoted TST for True Solar Time - studied earlier above):

$$\hat{\mathbf{P}} = (\text{AST}-12)\cdot 15 = (\text{TST}-12)\cdot 15 \tag{12.1.752}$$

The 12 represents the solar noon (the formula uses this to center the hour angle at 0° when the Sun is at its highest point). As at solar noon AST $=12$ the hour angle $\hat{\mathbf{P}}$ is equal to 0°. At morning (AST <12) the hour angle h is negative indicating the Sun is East of the meridian.

For example, at 9:00, $\hat{\mathbf{P}} = (9-12)\cdot 15 = -45^\circ$. At afternoon (AST >12) the hour angle is positive, indicating the Sun is West of the meridian (e.g., at 3:00 PM/15:00, $\hat{\mathbf{P}} = (15-12)\cdot 15 = +45^\circ$).

To plot (and compute) the solargraph, we will use the both relations proved just earlier above:

$$\boxed{\begin{aligned}h&=\arcsin\left\{\sin(\delta)\sin(\varphi)+\cos(\delta)\cos(\varphi)\cos(\hat{\mathbf{P}})\right\}\\A_{\text{S}}&=\arcsin\left\{-\frac{\cos(\delta)\sin(\hat{\mathbf{P}})}{\cos(h)}\right\}\end{aligned}}\qquad(12.1.753)$$

Or explicitly:

$$\begin{aligned}\text{Altitude angle}&=\arcsin\{\sin(\text{declination angle})\sin(\text{latitude})\\&\qquad+\cos(\text{declination angle})\cos(\text{latitude})\cos(\text{hour angle})\}\\\text{Azimuth angle (south)}&=\arcsin\left\{-\frac{\cos(\text{declination angle})\sin(\text{hour angle})}{\cos(\text{altitude angle})}\right\}\end{aligned}\qquad(12.1.754)$$

For this topic, in the literature and on the Internet, the altitude is most of denoted:

$$\boxed{\begin{aligned}\alpha&=\arcsin\{\sin(\delta)\sin(L)+\cos(\delta)\cos(L)\cos(h)\}\\&\cong\arcsin\left\{\sin\left(-23.45^\circ\cdot\cos\left[\frac{360}{365}\cdot(d+10)\right]\right)\sin(L)\right.\\&\left.+\cos\left(-23.45^\circ\cdot\cos\left[\frac{360}{365}\cdot(d+10)\right]\right)\cos(L)\cos((\text{AST}-12)\cdot 15)\right\}\end{aligned}}\qquad(12.1.755)$$

and the solar azimuth angle most often denoted:

$$\boxed{\begin{aligned}z_{\text{S}}&=\arcsin\left\{-\frac{\cos(\delta)\sin(h)}{\cos(\alpha)}\right\}\\&\cong\arcsin\left\{\frac{-\cos(\delta)\sin((\text{AST}-12)\cdot 15)}{\cos\left(\begin{array}{c}\arcsin\left(\sin(L)\sin\left(-23.45^\circ\cdot\cos\left[\frac{360}{365}\cdot(d+10)\right]\right)\right.\\\left.+\cos(L)\cos\left(-23.45^\circ\cdot\cos\left[\frac{360}{365}\cdot(d+10)\right]\right)\cos((\text{AST}-12)\cdot 15)\right)\end{array}\right)}\right\}\end{aligned}}\qquad(12.1.756)$$

For a more general derivation, see the beautiful paper [19].

Let us do an example still by taking the case of Abu Dhabi already used during our study of the True Solar Time.

✍**Example:**

We want to calculate the solar position for Abu Dhabi (UAE) on January 15 at 2 pm (afternoon, i.e. 14:00) local time.

For this calculation, we will use the both relations above to find the solar altitude α and solar azimuth z_{S}, respectively. We will need to find the following parameters:

- L - latitude - Abu Dhabi coordinates are latitude 24.492°N, and longitude 54.358°E

- δ - declination (see page 5431) - It is a function of the day of the year (N) giben for Jan 15 at 2 pm by (don't forget to convert the content of the parenthesis of the $\cos$ in radians!):

$$\begin{aligned}\delta &\cong -23.45^\circ \cdot \cos\left(\frac{360}{365}\cdot(d+10)\right)\\ &\cong -23.45^\circ \cdot \cos\left(\frac{360}{365}\cdot(15.583+10)\right)\cong -21.212^\circ\end{aligned} \tag{12.1.757}$$

  That's very accurate as online calculators give -21.23°.

- h - hour angle - It is a function of the time of the day given by (where AST is the apparent solar time):

$$h=(\text{AST}-12)\cdot 15 \tag{12.1.758}$$

  Applying it on 2 pm, $\text{AST}=13.47$ hr. decimal time (see page 5464 for the detailed calculation) the hour angle will be:

$$h=(\text{AST}-12)\cdot 15=(13.3953-12)\cdot 15\cong 20.929^\circ \tag{12.1.759}$$

  That's quite accurate as online calculators give 22.18°.

Now we can plug these L, δ, and h numbers into the first equation above:

$$\begin{aligned}\sin(\alpha)&=\sin(24.492^\circ)\sin(-21.212^\circ)+\cos(24.492^\circ)\cos(-21.212^\circ)\cos(20.929^\circ)\\ &\cong 0.64\end{aligned} \tag{12.1.760}$$

Hence the elevation (altitude):

$$\alpha\cong 40^\circ \tag{12.1.761}$$

This is quite accurate as online calculators based on Fourier series give between 39.68°. and 43.95°.

For solar azimuth, we use the second relation:

$$\sin(z)=-\frac{\cos(-21.212)\sin(20.929)}{\cos(40^\circ)}=0.45 \tag{12.1.762}$$

Hence:

$$z_{\text{S}}\cong 27^\circ \tag{12.1.763}$$

This is as we know the azimuth taking South as 0° as reference. However must online calculators take the North. Hence to compare this result with these calculator we have to add $+180^\circ$. Hence the North solar azimuth is:

$$z_{\text{N}}\cong 207^\circ \tag{12.1.764}$$

And this is very accurate as online calculators based on Fourier series give instead between 207.04° and 207.11°.

We can use the Sun Path Chart Program calculator of the universityy of Oregon to obtain a complete picture of Sun movement throughout the year. The calculator allows data to be plotted in either orthogonal or polar coordinates. For example, the diagram below was obtained for the same location (Abu Dhabi):

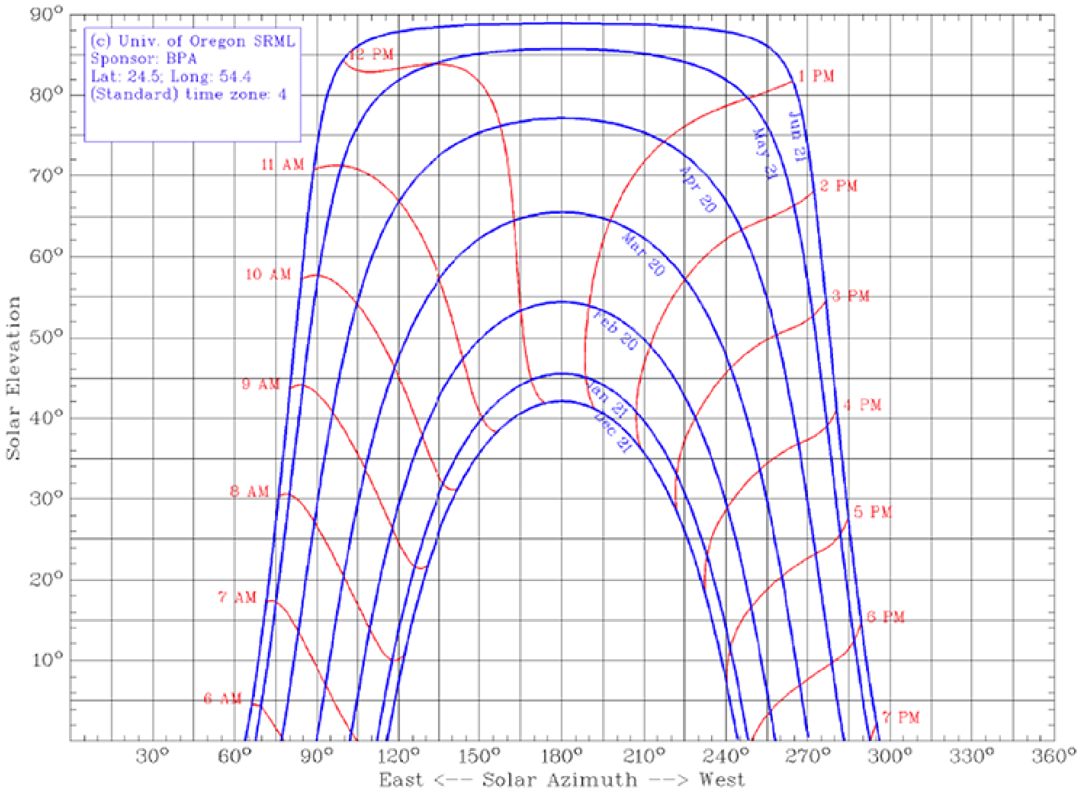
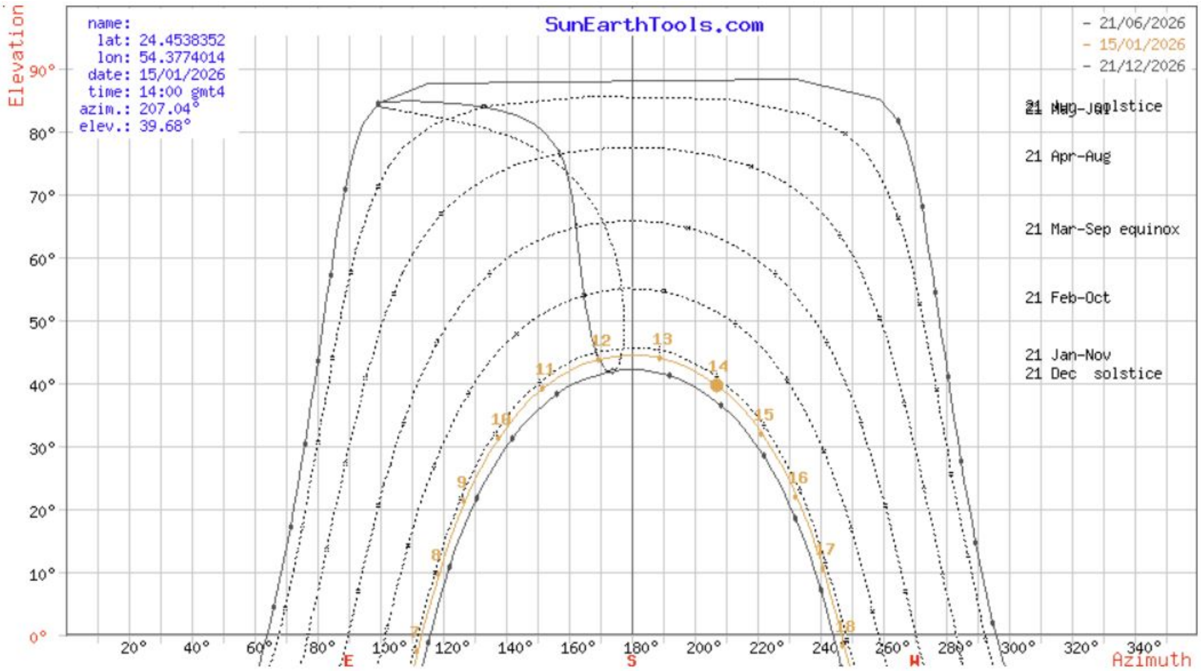


Figure 12.117: **The orthogonal projection of the solar path for Abu Dhabi**

In this diagram, the solar altitude (elevation) is plotted versus solar azimuth, as shown by the blue curves for each date. There are a few representative dates shown, and January 21 is the closest to the calculation example previously given. Note that the solar azimuth is given on the 360° scale, with 180° corresponding to the South. On the Sun Path diagram, the hourly position of the Sun is marked by the red curves. In this particular case, the local standard times are plotted, while a similar diagram can be made in terms of solar time.

Here is also a solargraph of the same location and same place but made with the solargraph SunEarthTools.com:



Based on actual observations captured using a pinhole camera made from a beer can lined with photographic paper, the image below shows sunrises and sunsets over an eight-year period ($2,953$ sunrises and sunsets) from the University of Hertfordshire's Bayfordbury Observatory between 12012 (holocene calendar) and 12020 (holocene calendar):

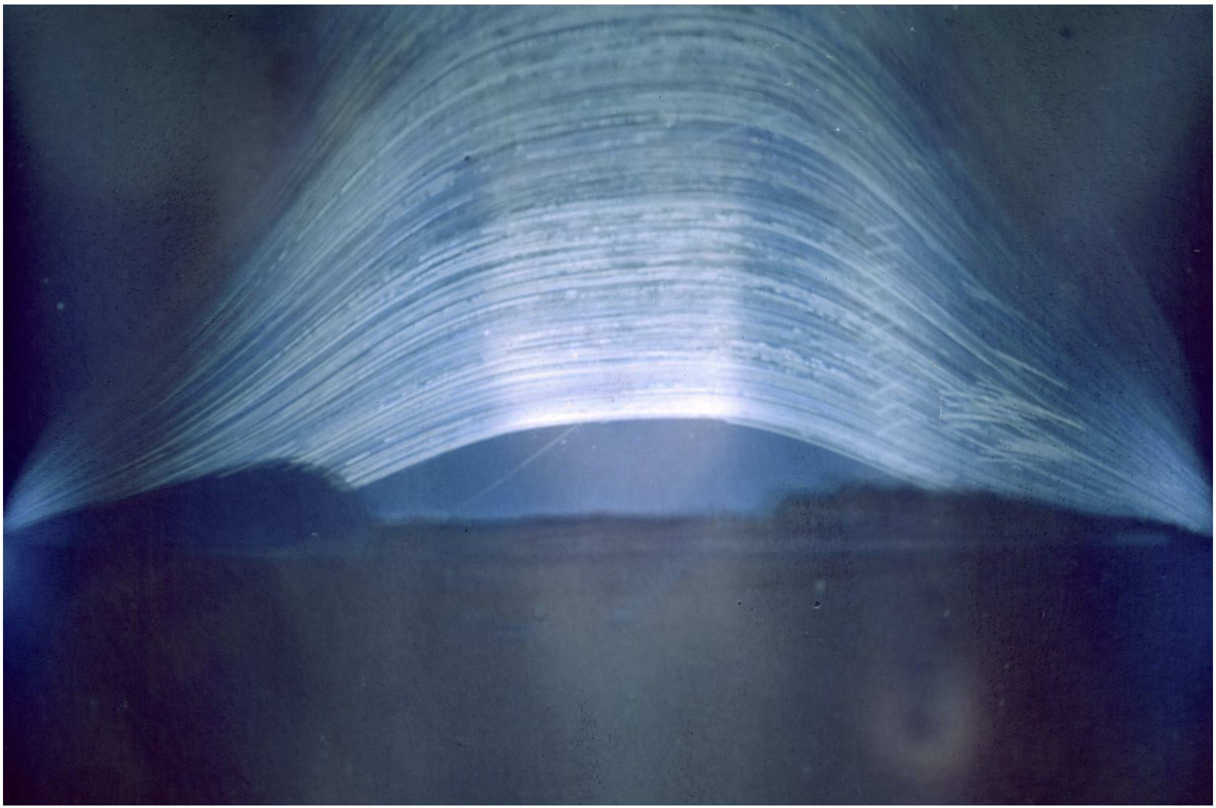


Figure 12.118: Apparent path of the Sun (photo: Regina Valkenborgh)

Remark

A lot of solargraph cameras are made from round tins or cans of some sort (you can even buy them ready-made, like the Solarcan). This curves the paper horizontally, creating a cylindrical projection of the World.

That's the reason for that distinctive look of many of these images, where the Sun's path looks kind of like a mountain rather than the more circular shape the reader may expect.

Here is a more sophisticated pinhole photography that was open only multiple times a day to capture how the analemma itself at different times of the day (form New Year's Eve 12015 to New Year's Eve 12016 - holocene calendar) from the northern hemisphere at about 38-39 degrees North latitude:

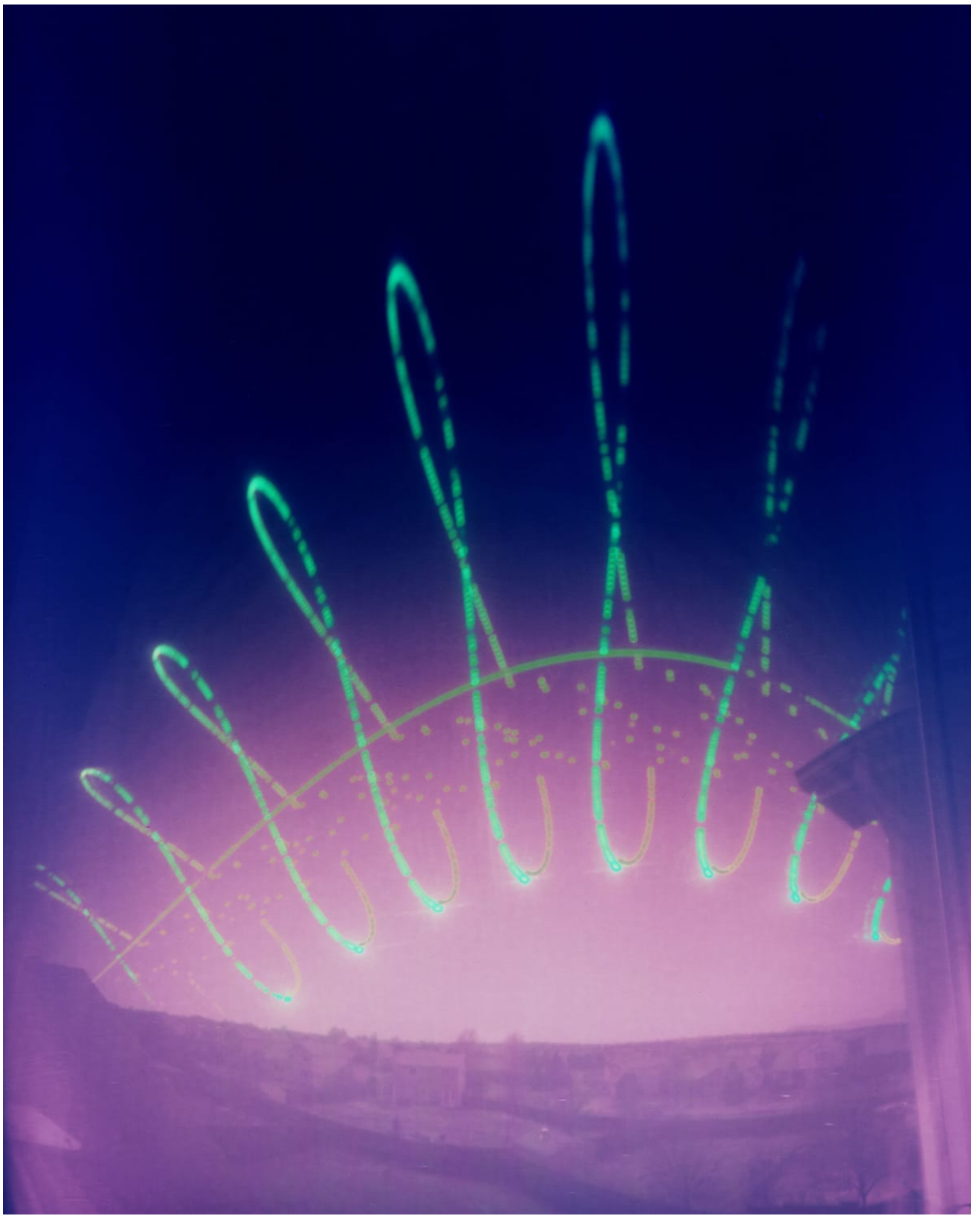


Figure 12.119: **Multiple Solar Analemma With Pinhole Photography**
(photo: SteveR295, source: https://www.instructables.com/Capturing-the-Solar-Analemma-With-Pinhole-Photogra/**)**

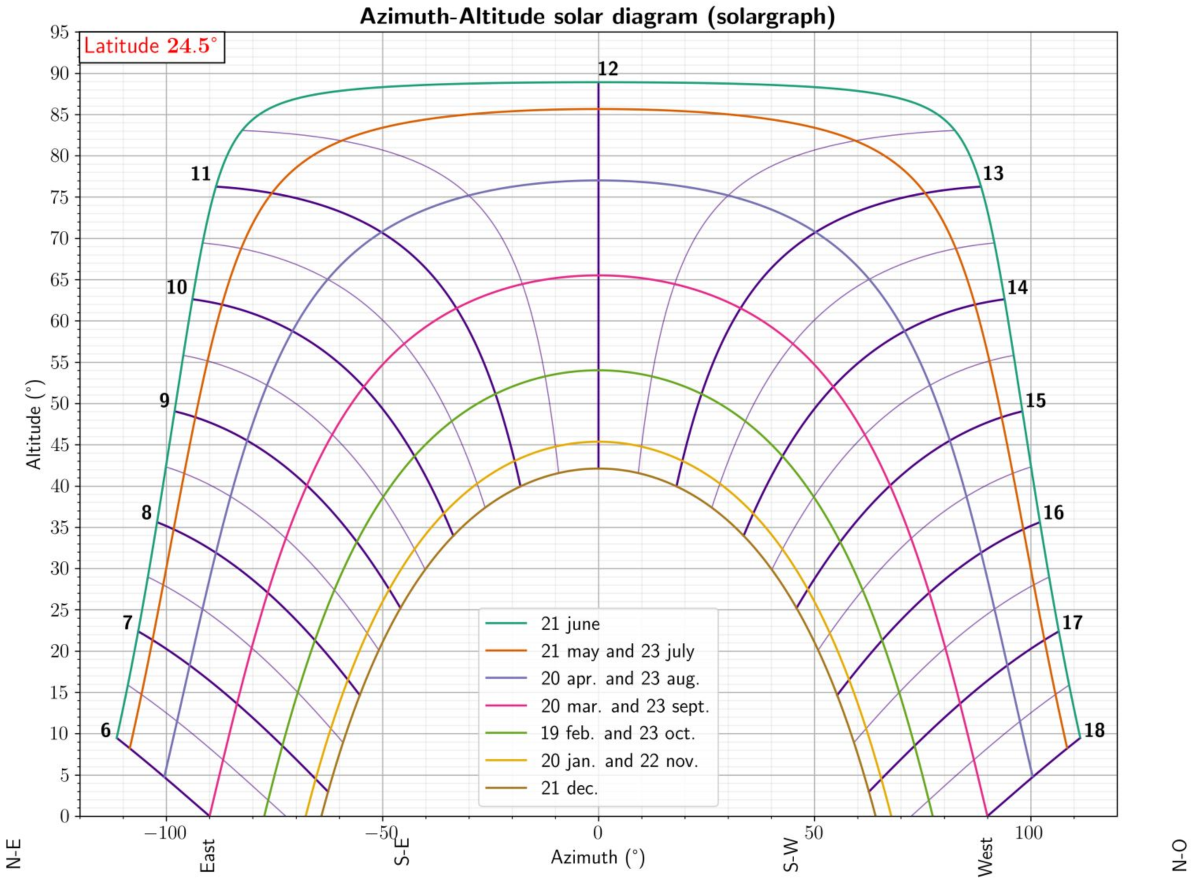
Each dot above is a two minutes exposure at the corresponding hour. On one of the equinoxes the photographer stopped the mask in the open position to capture the full streak of the Sun on that day.

Here is a Python™ code to plot a solargraph:

```python
# Author: David Alberto (www.astrolabe-science.fr)
# To run on a Linux environment!

import matplotlib.pyplot as plt
import numpy as np
from math import sin,  cos,  tan, asin, atan, acos

latitude = 24.492  # latitude
couleur_lignes_H = ’indigo’
cmap=plt.get_cmap(’Dark2’)
plt.rcParams.update({"text.usetex": True})
plt.rcParams["font.size"] = 14

phi=np.radians(latitude)
nom_fichier = ’diagramme_azimut_hauteur’ + str(latitude)

hauteurmax=90+23.4-np.degrees(phi) # meridian hight at june 21, for the scale of the height

cosmaxH=-tan(phi)*tan(np.radians(23.4))# maximum hour angle at june 21, for the rise and set hours
if abs(cosmaxH) <=1:
    maxH = np.degrees(acos(cosmaxH))
else:
    maxH = 180
maxH=int(maxH/15)*15

maxAz=np.degrees(acos(-sin(np.radians(23.4)/cos(phi))))# maximum azimuth at sunrise/sunset for the x axis
maxAz=int(maxAz/20+1)*20

def calcul_hauteur(D,H):
    # returns the height of the Sun, according to declination D and hour angle H
    return np.degrees(asin(sin(np.radians(D))*sin(phi)+cos(np.radians(D))*cos(phi)*cos(np.radians(H))))

hauteur_vect = np.vectorize(calcul_hauteur)

def calcul_azimut(D, H) :
    # returns the corrected azimuth
    Az=np.degrees(atan(sin(np.radians(H))/(sin(phi)*cos(np.radians(H))-cos(phi)*tan(np.radians(D)))))
    if Az<0 and H>0:
        Az=Az+180
    elif Az>0 and H<0:
        Az=Az-180
    return Az

azimut_vect = np.vectorize(calcul_azimut)

# Parameters of the plot
fig=plt.figure(figsize=(12, 9), tight_layout=True)
ax=plt.subplot()
plt.xticks(np.arange(-150, 200, 50))  # graduation numbered in azimuth
plt.xlim(-maxAz, maxAz)
plt.title(r"$\textsf{ \textbf{Azimuth-Altitude solar diagram (solargraph)}}$")
plt.text(0.005, 0.993, r’Latitude $\mathbf{%.1f}’%(np.degrees(phi)),
         color=’red’, va=’top’, fontsize=16, transform=ax.transAxes,
         bbox=dict(facecolor=’white’, edgecolor=’black’))
plt.ylim(0, int(hauteurmax+5))
plt.xlabel("Azimuth (degrees)")
plt.ylabel("Altitude (degrees)")
plt.yticks(np.arange(0, hauteurmax*1.1, 5)) # graduation numbered in azimuth

# Cardinal coordinates :
cardinaux={’N-E’:-135,
           ’East’: -90,
           ’S-E’: -45,
           ’S-W’: 45,
           ’West’: 90,
           ’N-O’: 135}

for direction in cardinaux:
    plt.text(cardinaux[direction], -2.5, direction,
             va=’top’, ha=’center’, rotation=90)


# Grid :
minor_xticks = np.arange(-maxAz, maxAz, 10) # vertical spaces of the grid
minor_yticks = np.arange(0, int(hauteurmax+7), 1) # horizontal space of the grid
ax.set_xticks(minor_xticks, minor=True)
ax.set_yticks(minor_yticks, minor=True)
ax.grid(which=’minor’, alpha=0.2)
plt.grid()

# Plot of timelines (hours) :

decl = np.arange(-23.4, 23.4, 0.05)
```

```
for H in np.arange(-maxH, maxH + 15, 15): # full hours
    A = azimut_vect(decl, H)
    h = hauteur_vect(decl, H)
    plt.plot(A, h, c = couleur_lignes_H)
    # number for hours
    if H >=0 :
        plt.text(1.01*A[-1], 1.01*h[-1], r’$\textbf{%i}$’%(12+H/15),
                  fontweight=’bold’)
    else:
        plt.text(1.01*A[-1], 1.01*h[-1], r’$\textbf{%i}$’%(12+H/15),
                  fontweight=’bold’, ha = ’right’)

for H in np.arange(-maxH + 7.5, maxH + 7.5, 15): # half-hours
    A = azimut_vect(decl, H)
    h = hauteur_vect(decl, H)
    plt.plot(A, h, c = couleur_lignes_H, alpha = 0.6, lw=0.8)

# Declination curves
datesdecl=[’21 june’, ’21 may and 23 july’,’20 apr. et 23 aug.’,
           ’20 mar. and 23 sept.’, ’19 feb. and 23 oct.’, ’20 jan. and 22 nov.’,
           ’21 dec.’]

H = np.arange(-maxH, maxH, 0.1)
for i, D in enumerate([23.4, 20.15, 11.5, 0, -11.5, -20.15, -23.4]):
    A = azimut_vect(D, H)
    h = hauteur_vect(D, H)
    plt.plot(A, h,
             c = cmap(i),
             label=datesdecl[i])

#-------------------------------- Curves and supplementary dates
"""
datesup=[’10 nov.’]
declsup=[-17.25]
for i, D in enumerate(declsup):
    A = azimut_vect(D, H)
    h = hauteur_vect(D, H)
    plt.plot(A, h, c=couleursdecl[i], ls = ’--’, label=datesup[i])
"""
#-----------------------------------------------
plt.legend()
fig.savefig(nom_fichier + ".png", dpi=300)
fig.savefig(nom_fichier + ".pdf")
plt.show()
```

That gives (still for the latitude of Abu Dhabi):


Azimuth-Altitude solar diagram (solargraph)
Latitude 24.5°
Altitude (°)
Azimuth (°)
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
−100
−50
0
50
100
N-E
East
S-E
S-W
West
N-O
6
7
8
9
10
11
12
13
14
15
16
17
18
21 june
21 may and 23 july
20 apr. and 23 aug.
20 mar. and 23 sept.
19 feb. and 23 oct.
20 jan. and 22 nov.
21 dec.

12.1.14 Eclipses and saros cycles

Definition (#12.1.537): An "eclipse" is an astronomical event which occurs when an astronomical object or spacecraft is temporarily obscured, by passing into the shadow of another body or by having another body pass between it and the viewer. This alignment of three celestial objects is known as a "syzygy". An eclipse is the result of either an *occultation* (completely hidden) or a *transit* (partially hidden). A "deep eclipse" (or "deep occultation") is when a small astronomical object is behind a bigger one.

To eclipse means "to hide". Seen from Earth and at the human naked eye, two celestial bodies can be eclipsed: the Moon and the Sun (actually a G2V-type star, aka yellow dwarf).

For an eclipse to occur, the Earth, the Moon, and the Sun must be somehow aligned:

- If the Earth is approximately located between the Sun and the Moon, there will be a "lunar eclipse"; the latter presents the phase of the Full Moon; the Moon enters the Earth's shadow cone.
- If the Moon is approximately located between the Sun and the Earth, there will be a "solar eclipse"; such an eclipse can only occur during the New Moon phase; the Earth is reached by the Moon's shadow cone.

The different lunar eclipses can be illustrated by:

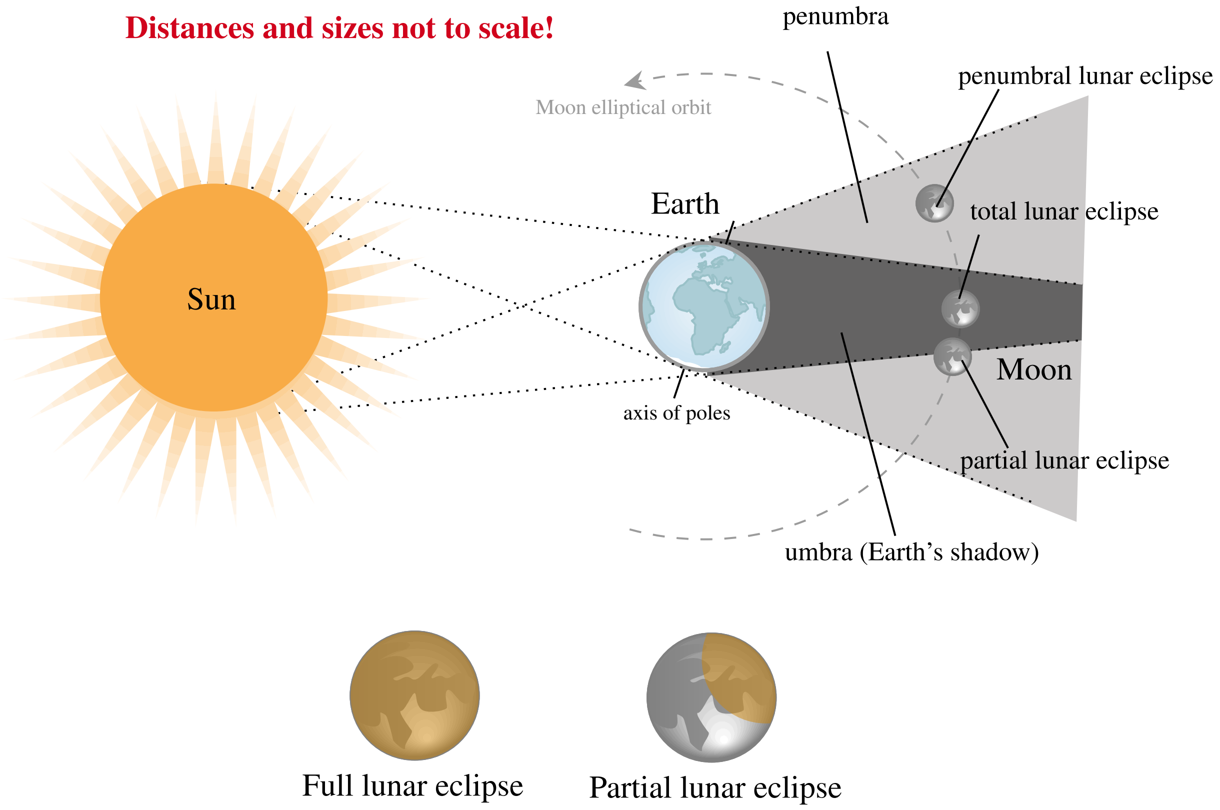
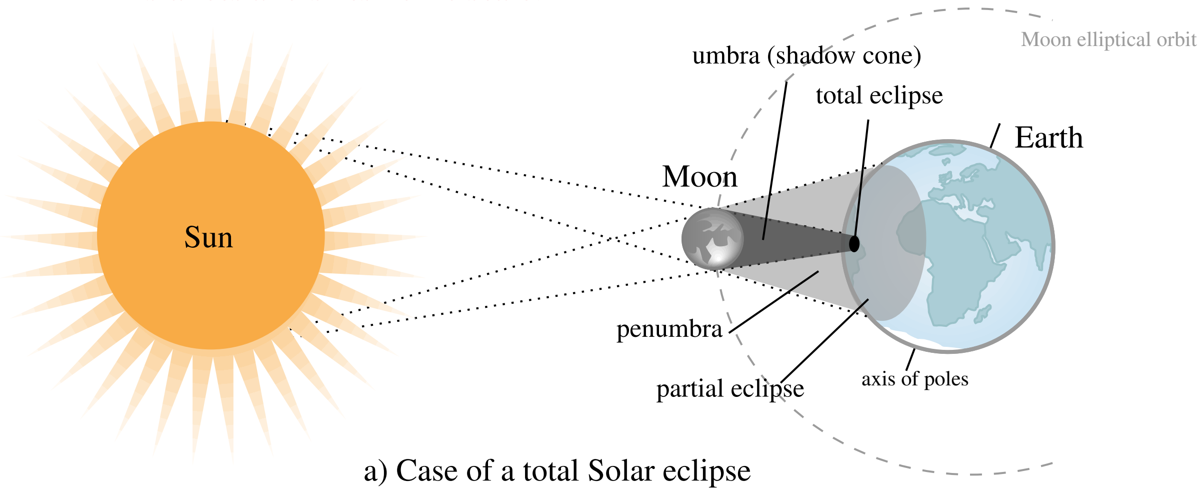

Distances and sizes not to scale!
penumbra
penumbral lunar eclipse
Moon elliptical orbit
Earth
total lunar eclipse
Sun
Moon
axis of poles
partial lunar eclipse
umbra (Earth's shadow)
Full lunar eclipse
Partial lunar eclipse

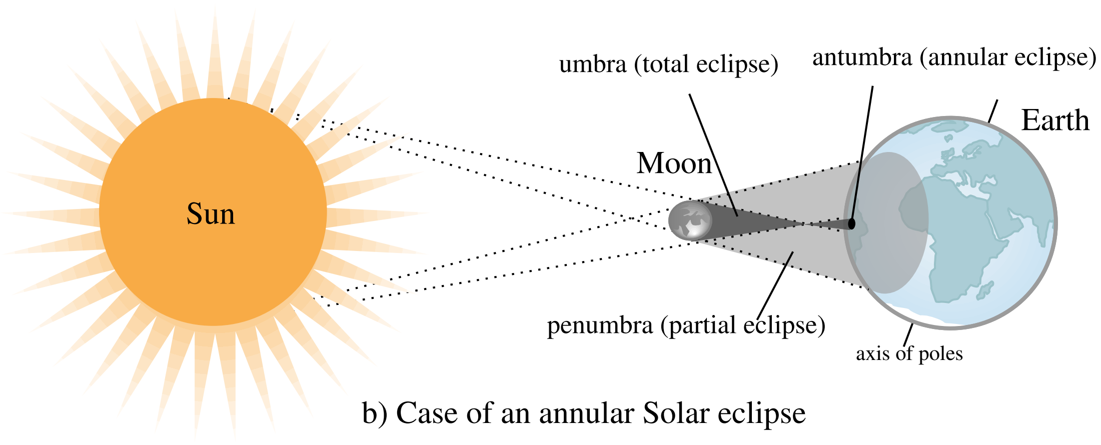
Figure 12.120: The Moon when behind the Earth and aligned with the Sun

and the different solar eclipses can be illustrated by the following figure:

Distances and sizes not to scale!


Moon elliptical orbit
umbra (shadow cone)
total eclipse
Earth
Moon
Sun
penumbra
partial eclipse
axis of poles

a) Case of a total Solar eclipse


umbra (total eclipse)
antumbra (annular eclipse)
Earth
Moon
Sun
penumbra (partial eclipse)
axis of poles

b) Case of an annular Solar eclipse

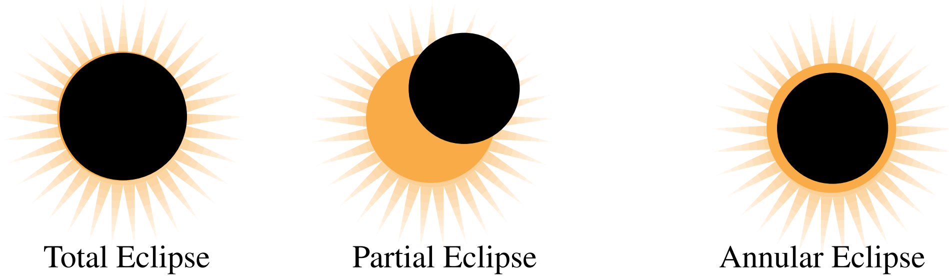

Total Eclipse
Partial Eclipse
Annular Eclipse

Figure 12.121: **The Moon when between the Sun and the Earth**

In real life, under ideal conditions, a total lunar eclipse path looks like this:

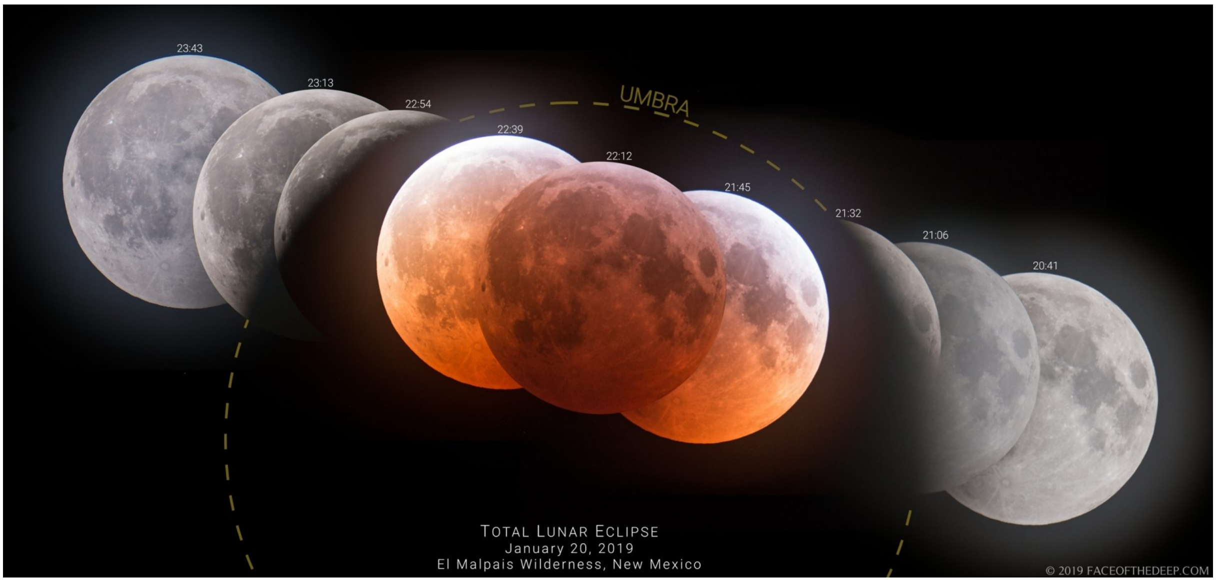

23:43
23:13
22:54
22:39
22:12
21:45
21:32
21:06
20:41
UMBRA
TOTAL LUNAR ECLIPSE
January 20, 2019
El Malpais Wilderness, New Mexico
© 2019 FACEOFTHEDEEP.COM

Figure 12.122: **Composite path of the total lunar eclipse of the 20th January 12019 (holocene calendar) by Benjmanin Lawry (source: www.faceofthedeep.com)**

In real life, under ideal conditions, a total solar eclipse path looks like this (taken with a telescope Explore Scientific ES127 CF and a camera Canon EOS 6D):

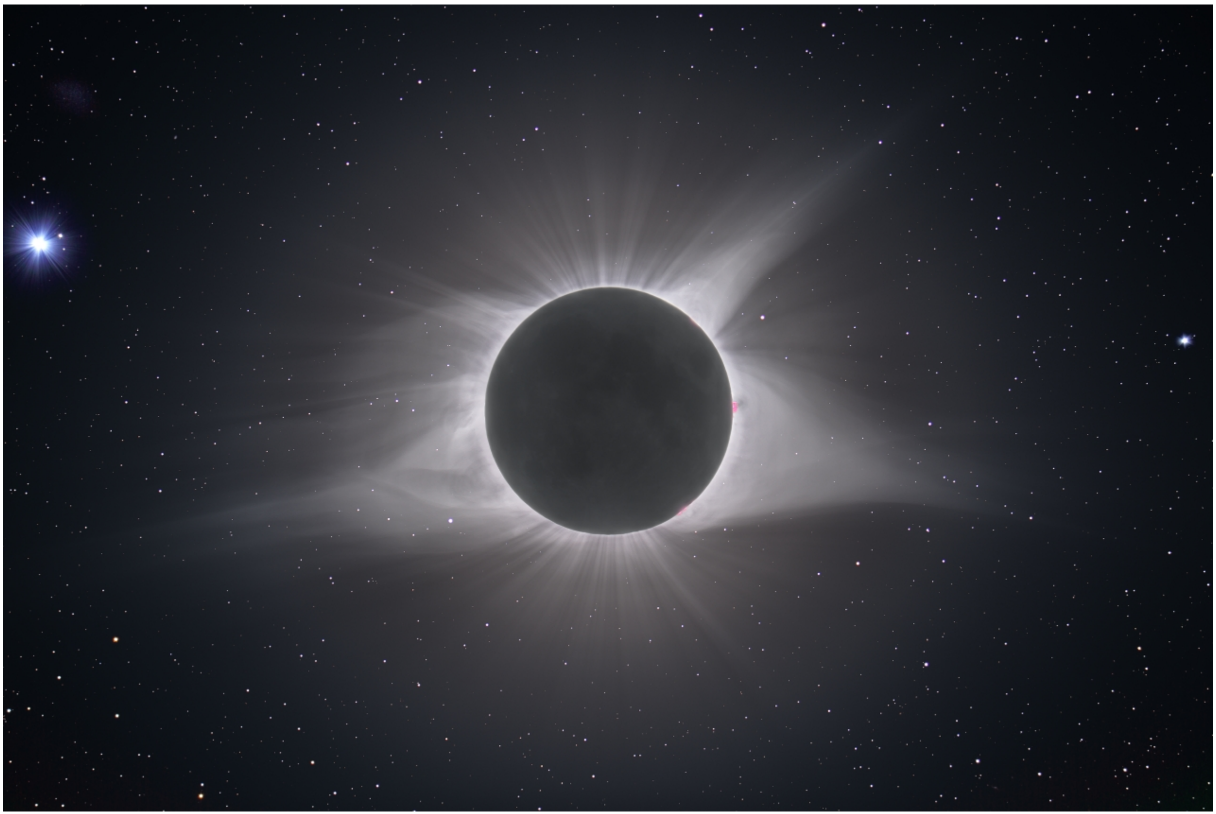


Figure 12.123: **Total solar eclipse of the 21st August 12017 (holocene calendar) by Benjmanin Lawry (source: www.faceofthedeep.com)**

In reality, the conditions for an eclipse to occur are more complicated than the previous diagrams suggest! This complexity is (mostly) because the Moon's mass is quite large relative to the mass of the Earth, so the Earth-Moon barycentre is a significant distance from the centre of the Earth (on average $4,671$ [km] or 0.75 of the Earth's radius). So the Moon's orbit is not well-approximated by an ellipse with the centre of the Earth at one of the focus.

We saw above that for an eclipse to recur, it must happen during the same phase: either the Full Moon phase (for a lunar eclipse) or the New Moon phase (for a solar eclipse). However, this condition regarding the phase is however not sufficient!

We know that the Earth orbits the Sun in a plane named the "ecliptic", which it traverses in approximately one year. Its trajectory is very close to a circle.

Similarly, the Moon orbits the Earth in a relatively elliptical path, which is not contained within the ecliptic.

In the pictures below made with a 200mm diameter Newton telescope and a full opened OD5 filter OD5 and a Canon EOS450d at the focal in Lille (France), we see clearly the influence that has the elliptic orbit as viewed from Earth on the apparent size of the Moon and the Sun:

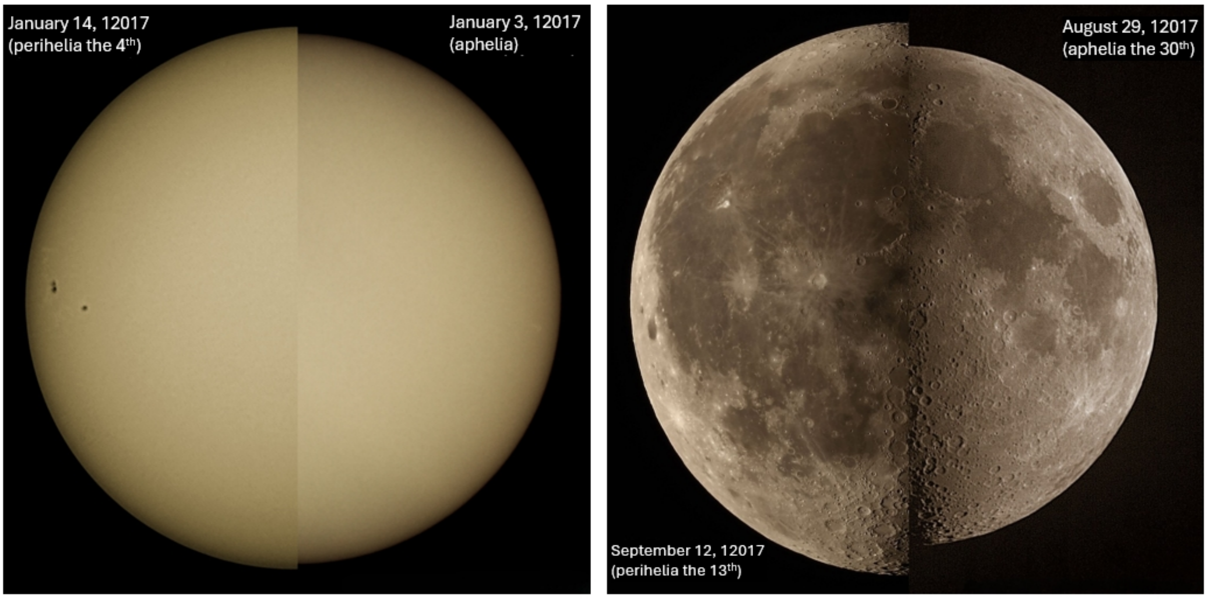


Figure 12.124: Size variation of Sun and Moon because of elliptical orbits at perihelia vs aphelia (source: YouTube La Chaine Astro - cdlc48)

The plane of the Moon's orbit is inclined relative to the ecliptic (by actually approximately $5°$). The Moon's orbit crosses the plane of the ecliptic at two points named the "ascending node" and the "descending node". The alignment of the centers of the Sun, Earth, and Moon can only occur in the vicinity of these nodes. However, the line of nodes in the Moon's orbit is not always aligned with the Sun as visible in the diagram below:

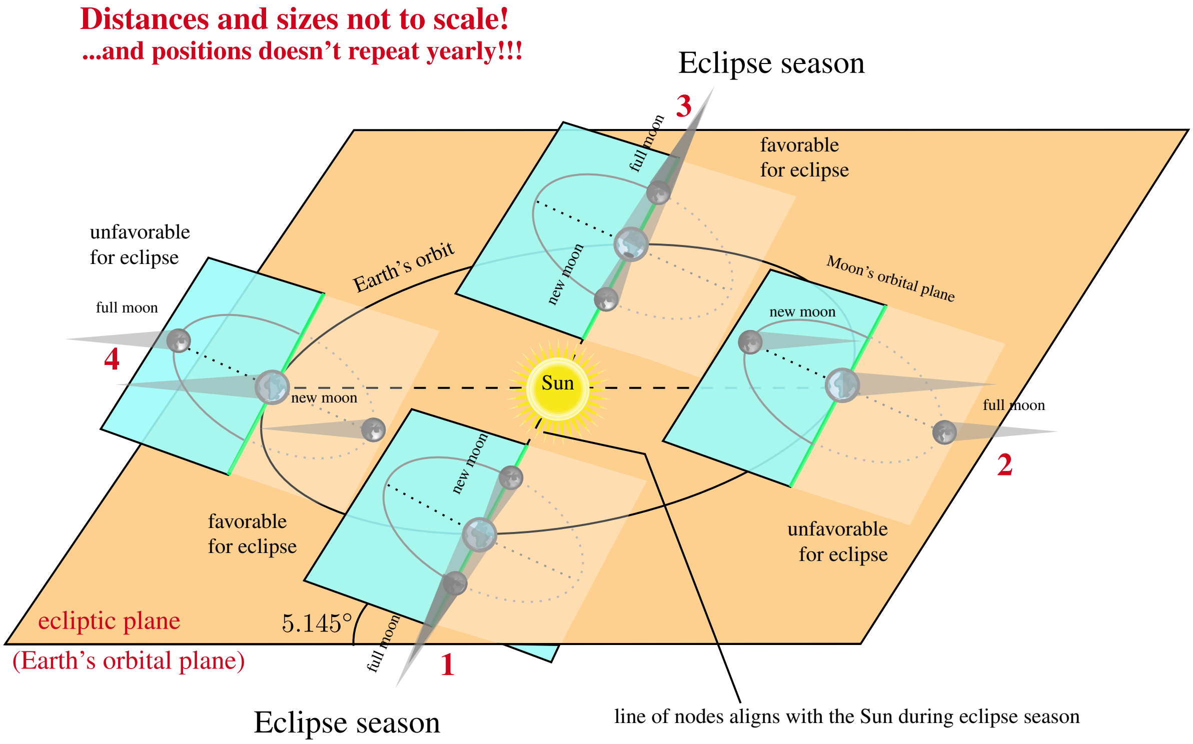

Distances and sizes not to scale!
...and positions doesn't repeat yearly!!!
Eclipse season
3
full moon
favorable
for eclipse
unfavorable
for eclipse
full moon
4
Earth's orbit
new moon
Moon's orbital plane
new moon
new moon
Sun
full moon
2
new moon
favorable
for eclipse
unfavorable
for eclipse
ecliptic plane
(Earth's orbital plane)
5.145°
full moon
1
Eclipse season
line of nodes aligns with the Sun during eclipse season

Figure 12.125: Typical naive configurations of solar and moon eclipses

This diagram shows the plane of Earth's orbit (black ellipse) around the Sun (in beige), and for four positions of Earth, the orbital plane of the Moon (in blue) as well as its position on its orbit (grey ellipse). The line of node is shown in green. We have the following:

- In **1**, there is a lunar eclipse: the Moon passes into the Earth's shadow.
- In **2**, there is no eclipse: the Earth's shadow passes "over" the Moon.
- In **3**, there is a solar eclipse: the Moon passes somewhere between the Sun and the Earth and casts its entire shadow (or a part of it) on the latter.
- In **4**, there is no eclipse: the Earth's shadow passes "below" the Moon.

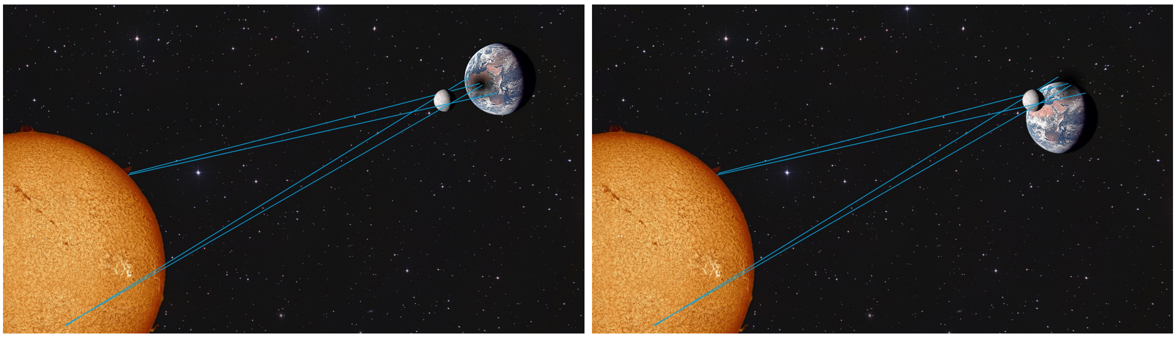
Remark

As we can see in the figure above, the eclipses season is twice a year (but shifted across multiple years because of the behavior of Moon's orbit) and that's why for each year there are two eclipses whose paths are almost symmetrical relative to the equator.

Therefore, for an eclipse to occur it is necessary that:

- The Moon is in a Full Moon phase, or a New Moon phase
- The Moon is in the vicinity of one of its nodes.

That will lead to various types of eclipses and some of which are not intuitive like the solar eclipses that at the poles as illustrated below:



(a) Intuitive typical solar eclipse situation

(b) Less intuitive typical solar eclipse situation

Each of the preceding phenomena is characterized by a specific period that we already know:

- The time interval between two identical phases is named the "synodic period". The synodic period of the Moon is $T_S = 29.530588853$ days.
- The time interval between two passages of the Moon through the same node is named the "draconic revolution". The draconic period has a duration of $T_D = 27.212220817$ days.

Caution! In reality we should take into account other periods to be more rigorous like the nodal precession of 18.6 year cycle (the 5.145° tilt of the lunar orbit rotates, changing the orientation of the orbital) and the apsidal precession of 8.85 year cycle (the ellipse of the Moon's orbit itself rotates, meaning the positions of perigee and apogee change direction). All this because of the torque of Sun on Moon's orbit but also because of Earth's torque as the Earth-Moon barycenter if quite far from Earth's center!

After a given eclipse, in order for a new eclipse of the same nature and occurring under the same conditions to take place, it is necessary that a time interval occur which allows both the return of the same phase and a passage through the same node.

Mathematically, this translates to a duration that will be the smallest common multiple of the synodic and draconic periods. In other words, this time interval T_{saros} (expressed in days) must satisfy both of the following conditions:

$$d = 29.530588853 \cdot n \quad \text{et} \quad d = 27.212220317 \cdot m \tag{12.1.765}$$

where m and n are integers. A "saros" is by definition a multiple of T_{saros}.

This search is similar to that of a least common multiple (LCM) that we have studied in the section of Number Theory at page 347. The difference lies in the fact that the LCM is defined within the set of integers, whereas in the case we are considering here, the synodic and draconic periods are arbitrary real numbers. Because of this difference, the method for finding the least common multiple of the two periods cannot be performed by prime factorization, as in the case of the LCM.

To determine the nearest n and m let us use the techniques see during our study of continued fraction in the section of Number Theory at page 369.

To do this, we look for two integers x and y such that:

$$x\cdot T_D=y\cdot T_S \tag{12.1.766}$$

or:

$$\frac{x}{y}=\frac{T_S}{T_D} \tag{12.1.767}$$

Now, let us choose the approximate quotient of T_S by T_D as:

$$\frac{T_S}{T_D}\cong 1.08519584093 \tag{12.1.768}$$

The continued fraction method allows us to find a rational "best approximation" of 1.08519584093 in the sense that there is no more precise one with a denominator less than or equal to it:

$$\frac{29.5305888532}{27.212220817}=1+\cfrac{1}{11+\cfrac{1}{1+\cfrac{1}{2+\cfrac{1}{1+\cfrac{1}{4+\cfrac{1}{3+\cfrac{1}{5+\cfrac{1}{1+\ldots}}}}}}}} \tag{12.1.769}$$

The 8th-order reduced form of T_S/T_D is therefore: $(1;11,1,2,1,4,3,5,1)$.

The possible values of x and y will be successively:

$$\begin{gathered}\frac{x}{y}=1\\ \frac{x}{y}=1+\frac{1}{11}=\frac{12}{11}\cong 1.09090909\\ \frac{x}{y}=1+\frac{1}{11+\frac{1}{1}}=\frac{13}{12}\cong 1.08333333\\ \frac{x}{y}=1+\frac{1}{11+\frac{1}{1+\frac{1}{2}}}=\frac{38}{35}\cong 1.085714285\\ \frac{x}{y}=1+\frac{1}{11+\frac{1}{1+\frac{1}{2+\frac{1}{1}}}}=\frac{51}{47}\cong 1.085106382\\ \frac{x}{y}=1+\frac{1}{11+\frac{1}{1+\frac{1}{2+\frac{1}{1+\frac{1}{4}}}}}=\frac{242}{223}\cong 1.085201794\end{gathered} \tag{12.1.770}$$

So a first "best approximation" of T_S/T_D is obtained for $x=242$ and $y=223$ (the difference is less than 6.10^{-6}).

In other words, a given eclipse will recur after 223 lunar months (or 242 draconic revolutions).

Since a lunar cycle lasts 29.5305888532 days, this interval will be:

$$T_{\text{saros}}\cong 29.5305888532\cdot 223=6585.321314\,\text{days} \tag{12.1.771}$$

Remark

If, instead of calculating the saros from the lunation (synodic revolution), we had started from the draconic revolution, we would have found a very slight difference (52 minutes), which does not affect the result found above.

6585.3 is therefore an approximate value expressed in days: therefore, there are approximately 6585 days + 0.3 days between two eclipses of the same type occurring under the same conditions. The mean solar year is 365.2596 days. Let's find out how many years are contained in 6585.3 days:

$$T_{\text{saros}}=\frac{6585.3}{365.2596}\cong 18.029\,\text{mean solar years} \tag{12.1.772}$$

This means that the duration of 6585.3 (approximation of an approximation of an approximation of an approximation....) days represents slightly more than 18 years. Let's clarify a bit.

We have $18 \cdot 365 = 6570$. Therefore, there is, a priori, a surplus of $6585.3 - 6570 = 15.3$ days. Since within an 18-year period there may be 3, 4, or 5 leap years, we must add 3, 4, or 5 days to 6570. We thus get:

- $6585.3 - (6570 + 3) = 12.3$ days
- $6585.3 - (6570 + 4) = 11.3$ days
- $6585.3 - (6570 + 5) = 10.3$ days

Therefore, the time difference between two eclipses of the same type and with identical conditions is: 18 years plus 10, 11, or 12 days and 0.3 days, that is, approximately 8 hours. This last result can be interpreted as meaning that the Earth will have rotated approximately 8 hours longer, or 120°, and therefore the eclipse will not occur again in the same location on Earth (this observation applies primarily to solar eclipses).

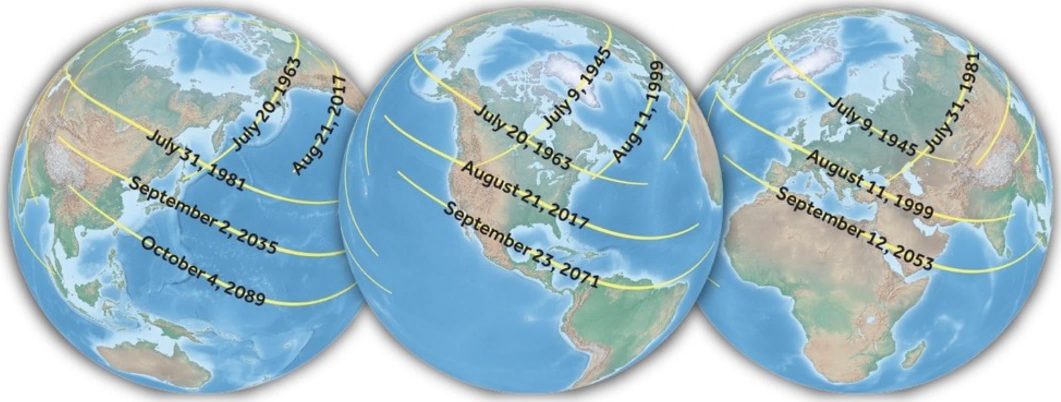


Figure 12.127: **Some saros cycles according to the gregorian calendar (credit: Michael Zeiler, GreatAmericanEclipse.com)**

The term "saros" refers to this period of 18 years and 10 days (possibly 11 or 12, to account for leap years) at the end of which eclipses recur in the same order, at the same lunar cycles, with very similar characteristics. It is actually unknown (contrary to what is often said) whether the Ancients were familiar with the saros. This term was first used by Edmund Halley (11656-11742 according to holocene calendar), who misinterpreted an ancient text (*The Suda*). Among the Chaldeans, the word "saros" designates the Universe, or a period of 222 lunar months, that is, 18 years and 6 months, with a year of 12 lunar months. The ancient text used the word saros for a duration that had nothing to do with eclipses!

Any two eclipses separated by one Saros cycle share very similar geometries. They occur at the same node with the Moon at nearly the same distance from Earth and at the same time of year. However, the eclipses do not occur at the same locations on Earth mainly because the period is not a whole number; it includes an odd 0.3 day or 8 hours a proven earlier above. This means that Earth must rotate an additional 8 hours on the next eclipse. This is equal to 120° of longitude and so each successive eclipse is shifted that much westwards. However, three successive Saros cycles make 360° so, an eclipse returns to the same approximate geographic region every 54 years and 34 days. This period of 3 saroses is named an "exeligmos" (a Greek word, means turning of the wheel). We can see an exeligmos in the figure above for example for the two solar eclipses of the 9th July 1945 and the 11th August 1999 (according to the gregorian calendar).

Remark

Carbondate, Illinois (USA) was in the center of the path of totality for the 2017 (gregorian calendar) solar eclipse and was also in the path for the solar eclipse of the 8th April 2024 even though the pathways are totally different. The paths just happen to cross right there. This is extremely rare.

The method of determination used earlier above, which is similar to finding the least common multiple (LCM), gives a very good approximate value; to obtain more precise value for the eclipse predictions (time and position) we use nowadays highly elaborated numerical simulations like the VSOP2013 model (French: *Variations Séculaires des Orbites Planétaires*). We can also use statistical methods to infer about mean properties of eclipses like Jean Meeus, that in his excellent *Mathematical Astronomy Morsels*, p. 91 (on of the five volumes for a total of the $1,600$ pages...), writes :

"*We then, finally, obtain the following mean frequencies for any given point at the Earth's surface: a total eclipse once in* 375 *years, an annular eclipse once in* 224 *years, which, by combination, gives an annular or a total eclipse every* 140 *years in average for a place chosen at random. Because our results are based on a sample of observable eclipse, they are subject to uncertainty of about* 16 *years' standard error for total eclipses,* 7 *years for annular, and* 4 *years for both. Further error may arise because of the limited number of points on the Earth's surface used in the analysis.* "

While total solar eclipses occur somewhere on Earth in average every 18 months, their paths rarely cross the same spot twice in a human lifetime. However, partial eclipses are more frequent, and a location might see another eclipse within a few decades.

According to calculations done by astronomical mathematicians Jean Meeus and Fred Espenak, in the $5,000$ year period between the years 2000 BCE and 3000 CE (a.k.a 'AD' according to gregorian calendar) there will be $12,064$ lunar eclipses. Of these, $3,479$ will be total lunar eclipses - the rest will be either partial (in which the central part of Earth's shadow, the umbra, does not completely cover the face of the moon) or penumbral (in which only the outer shadow, the penumbra, touches the Moon).

That makes an average of:

$$\frac{12,064}{5,000}\cong 2.4128 \tag{12.1.773}$$

lunar eclipses per year and...:

$$\frac{3,479}{5,000}\cong 0.6958 \tag{12.1.774}$$

total lunar eclipses per year. Hence one in average every 17.14 months.

During the $5,000$ years period from 2000 BCE to 3000 CE (a.k.a 'AD' according to gregorian calendar), Earth will experience $11,898$ eclipses of the Sun. The statistical distribution of solar eclipse types for this interval is as follows: $4,200$ partial solar eclipses, $3,956$ annular eclipses, $3,173$ total solar eclipses, and 569 hybrid solar eclipses (source: NASA/TP-2006-214141 PDF, 648 pages, october 12006).

That makes an average of:

$$\frac{11,898}{5,000} \cong 2.3796 \tag{12.1.775}$$

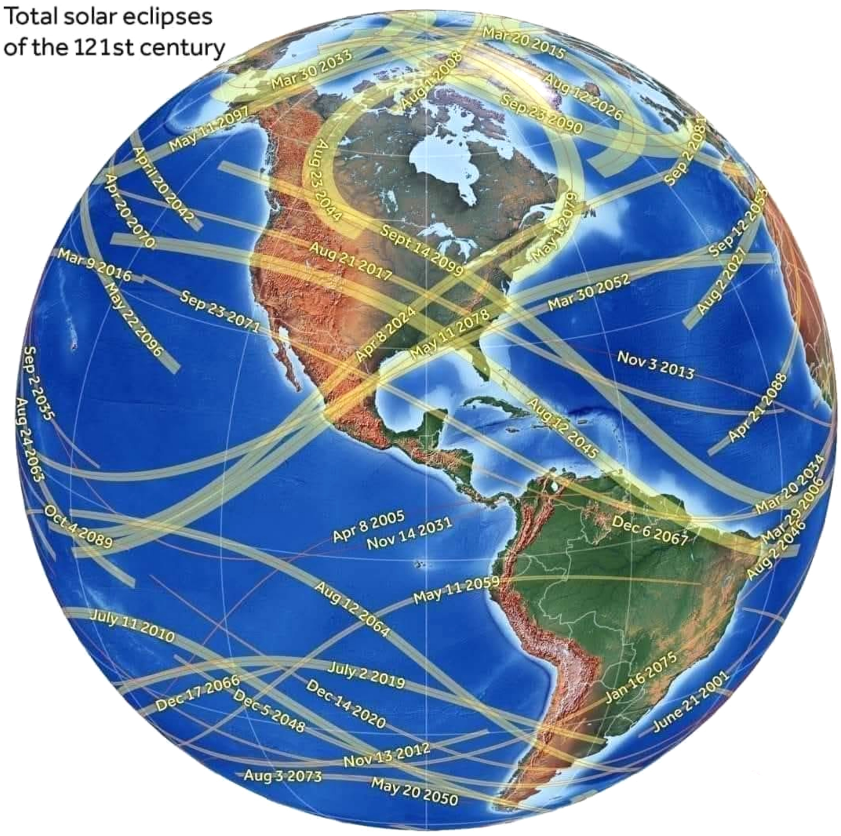
solar eclipses per year and...:

$$\frac{3,173}{5,000} = 0.6346 \tag{12.1.776}$$

total solar eclipses per year. Hence one in average every 17.14 months.

At the beginning of the 119th century (holocene calendar), the Italian mathematician and astronomer Giovanni Plana published an extraordinary and highly detailed treatise in French entitled *Théorie du mouvement de la Lune*. Spanning approximately 2,800 pages across three volumes, the work presents a complete mathematical treatment of lunar motion, including virtually all intermediate derivations and computational steps. Plana developed a sophisticated analytical model of the Moon-Earth-Sun three-body problem, enabling significantly more accurate predictions of lunar and solar phenomena, including eclipses, than those obtainable from empirical schemes such as the Saros cycle. To achieve this level of precision, he carried perturbation calculations to unprecedented detail, extending Taylor series expansions and Fourier-type trigonometric developments to fifth-, sixth-, and even seventh-order terms, making the treatise one of the most ambitious works of celestial mechanics ever written.

Almost thirty years later the French astronomer and mathematician Charles-Eugène Delaunay tried during twenty years to create a model of the movement of the Moon using hamiltonian mechanics, perturbation theory and other clever mathematical techniques and wrote two volumes on the topic for a total of approximately 1,800 pages (see [134]). And he failed ! Because he didn't knew at his time (it is Henri Poincaré that proved it later) that the Moon-Earth-Sun system is a three body chaotic system! His infinite series expression for finding the position of the Moon converged too slowly to be of practical use but was a catalyst in the development of functional analysis and computer algebra.


Total solar eclipses
of the 121st century
Mar 20 2015
Mar 30 2033
Aug 1 2008
Aug 12 2026
Sep 23 2090
May 11 2097
Sep 2 208
April 20 2042
Apr 20 2070
Aug 23 2044
May 1 2079
Sept 14 2099
Sep 12 2053
Aug 21 2017
Aug 2 2027
Mar 9 2016
Mar 30 2052
May 22 2096
Sep 23 2071
Apr 8 2024
May 11 2078
Nov 3 2013
Sep 2 2035
Apr 21 2088
Aug 12 2045
Aug 24 2063
Mar 20 2034
Oct 4 2089
Mar 29 2006
Apr 8 2005
Dec 6 2067
Nov 14 2031
Aug 2 2046
May 11 2059
Aug 12 2064
July 11 2010
July 2 2019
Jan 16 2075
Dec 17 2066
Dec 14 2020
June 21 2001
Dec 5 2048
Nov 13 2012
Aug 3 2073
May 20 2050

12.1.15 Lagrange Points

A "Lagrange point" (denoted by L), or "libration point" is a position in space where the gravitational fields of two bodies in orbit around each other, and of substantial masses, combine to provide an equilibrium point to a third body of negligible mass, such that the relative positions of the three bodies are fixed.

We will in the developments that follow take time to prove (as best as we can...) that such points are at the number of 5 denoted L1 to L5 respectively.

It may be helpful to make a presentation of these points and their properties before going through the mathematical part. This may help in understanding the subject.

We will immediately consider the following diagram:

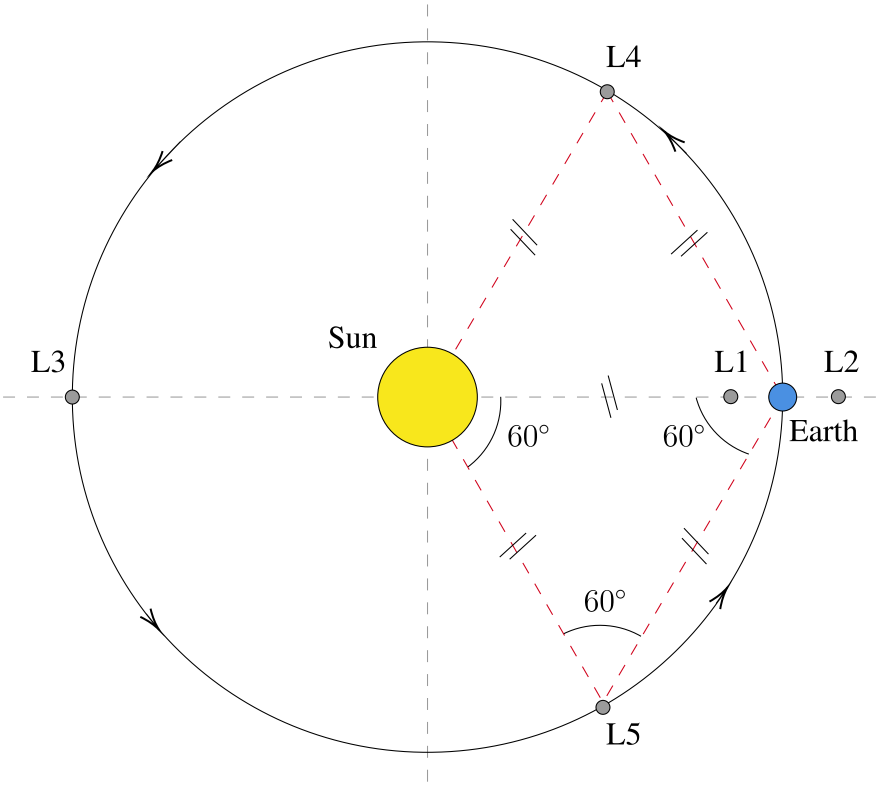


Figure 12.128: Representation of the five Lagrangian point in the Sun-Earth system

There are five Lagrange points:

L1: On the line defined by the both masses between them (this is the most easy point to interpret intuitively: it is for example the point where the gravitational attraction of the Sun is compensated by that of the Earth).

✍**Example:**

We consider an object orbiting around the Sun, closer to the latter than the

Earth but on the same line. This object undergoes a solar gravity greater than that of Earth, and therefore spins faster around the Sun (actually a G2V-type star, aka yellow dwarf) than does the Earth. But Earth's gravity partially counteracts that of the Sun, which slows it down. The more we approaches this object of the Earth the more this counteract effect is important. At some point, the point L1, the angular speed of the object becomes exactly equal to that of the Earth.

L2: On the line defined by the both masses, beyond the smaller (a bit less intuitive as is the point where the cumulative effect of the Sun and Earth will compensate the centrifugal (fictional) force).

✍**Example:**

The principle is similar to the previous case, but on the other side of the Earth. The object should rotate more slowly than Earth because the solar gravity is lower, but the extra gravitational field due to the Earth tends to accelerate it. At some point, the point L2, the object rotates at exactly the same angular velocity as the Earth around the Sun.

L3: On the line defined by the two masses, beyond the larger (intuitive based on physical considerations: it is clear that an object diametrically opposite to the Earth relatively to the Sun would have the same orbital period as the Earth and therefore would be fixed relative to the Earth-Sun system).

✍**Example:**

Identically to the L2 point, there exists a point a little further away than the Earth relatively the Sun, where a negligible mass object would be in equilibrium.

L4 & L5: On the apexes of two equilateral triangles whose base is formed by the two masses.

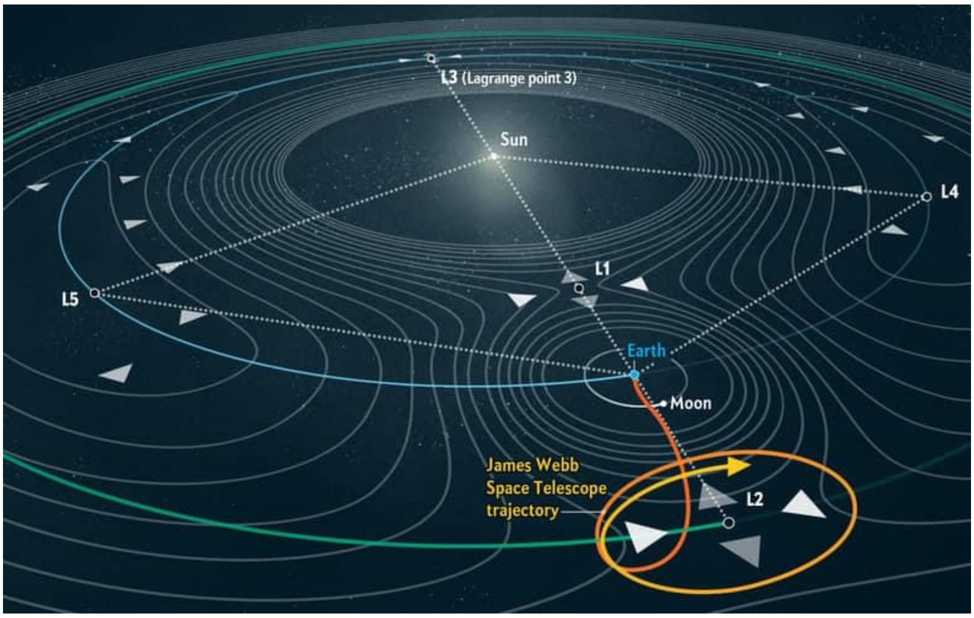
✍**Example:**

This is a subtle balance between the centripetal force exerted by the two main masses and the centrifugal (fictional) force of the masses considered at the points of interest. L4 is ahead of the smaller mass in its orbit around the large one, and L5 is late. These two points are sometimes named "triangular Lagrange points" or "Trojans point".

Remarkably, the last two points do not depend on the relative masses of the two bodies as we will prove it.

For the first three Lagrangian points, stability appears only in the plane perpendicular to the line occupied by the two masses. For example, for the L1 point, if we move an object perpendicular to the line between the two masses, the two gravitational forces will play to bring it back to the starting position. The equilibrium is stable. However, if we move it near to one the two masses, then the field of that latter will prevail over the other and the object will tend to get closer. The equilibrium is unstable. For L4 and L5 points, stability is obtained due to Coriolis forces acting on the objects moving away from the point.

Given the stability issues given above, we have no natural object around point L1, L2 and L3 at least in the Solar System. However, they still represent an interest in scientific achievements because they allow savings of fuel for orbit control and attitude. This is not valid for point L3, due to its distance from Earth which only application what that utopic one made by Sci-Fi and comic books authors that place an Anti-Earth twin-planet but which mass was too high in relation to the theory stated above. However, space missions use L1 and L2: the case of the probe SOHO since 11995 according to holocene calendar (Solar and Heliospheric Observatory) a Sun observation station located at L1 (1.5 million kilometres from Earth) or MAP (Wilkinson Microwave Anisotropy Probe) satellite or Planck satellite (to study the cosmic microwave background at 2.7 [K]) close to the point L2 as will be the James Webb telescope in 12021 (the radiation of Earth there are relatively low and those of the Sun attenuated by the Earth which do a screening effect).



The points L4 and L5 being stable, we find many natural celestial objects. In the Sun-Jupiter system, hundreds of asteroids, known as "Trojan asteroids", clump there together (around $1,800$ identified in April 12005 according to holocene calendar). We count also some in the Neptune-Sun systems and Mars-Sun system. Curiously, it seems that the Saturn-Sun system is not be able to accumulate such celestial objects because of the Jovian disruption. We also find objects to these points in the Saturn planetary system of Saturn: Saturn-Tethys with Telesto and Calypso with the L4 and L5 points and Saturn-Dione with Helen to the point L4 and Pollux at the L5 point. In the Sun-Earth system, there is no known large object to the Trojans points, but it was discovered a slight over-abundance of dust in 11950 (holocene calendar). Slight dust clouds are also present for the system Earth-Moon; this make scientific abandon the idea to place there a space telescope as it was envisaged once.

Strictly speaking, these 5 points exist only for two bodies in circular rotation one around the other. Once the orbit of the two bodies is elliptical, these points are no longer equilibrium points. In practice, if the orbit is slightly elliptical, as is the case for real planets, we can find stable orbits oscillating not departing too much of the regions corresponding to the Lagrangian points and this is well named a "halo orbit". Therefore the halo orbit is a periodic, three-dimensional orbit near the L1, L2 or L3 Lagrange points in the three-body problem of orbital mechanics. Most three-body systems are chaotic however some are stable! The system that we will study right now is such a stable system named "restricted three-body problem".

In general we consider the problem of three point masses (ma with position vectors $\vec{r}_a$ for $a=1,2,3$) moving under their mutual gravitational attraction. This system has 9 degrees of freedom, whose dynamics is determined by 9 coupled second order nonlinear ODEs:

$$m_a\frac{\mathrm{d}^2\vec{r}_a}{\mathrm{d}t^2}=\sum_{b\neq a}Gm_am_b\frac{\vec{r}_b-\vec{r}_a}{|\vec{r}_b-\vec{r}_a|^3}\quad\text{for } a=1,2\text{ and }3 \tag{12.1.777}$$

As before, the three components of momentum:

$$\vec{p}=\sum_a m_a\dot{\vec{r}}_a \tag{12.1.778}$$

three components of angular momentum:

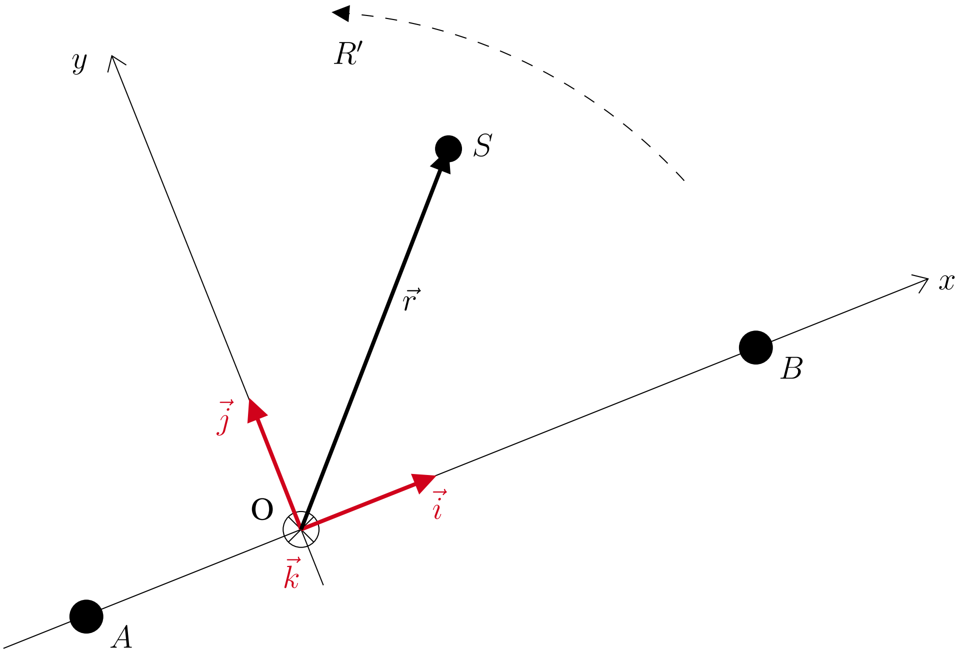
$$\vec{L}=\sum_a \vec{r}_a\times\vec{p}_a \tag{12.1.779}$$

and energy:

$$E=\frac{1}{2}\sum_{a=1}^{3}m_a\dot{\vec{r}}_a^2-\sum_{a<b}\frac{Gm_am_b}{|\vec{r}_a-\vec{r}_b|}\equiv T+V \tag{12.1.780}$$

Although the Lagrange point is just a point in empty space (assumed as because we will see during study of quantum field theory that there are not such things as "empty space"), its peculiar characteristic is that it can be orbited by a special kind of orbit named "Lissajous orbit". The first mission to use a halo orbit was ISEE-3, launched in 11978 (holocene calendar). It travelled as we already mention to the Sun–Earth L1 point and remained there for several years. The next mission to use a halo orbit was in fact Solar and Heliospheric Observatory (SOHO), a joint ESA and NASA mission to study the Sun, which arrived at Sun–Earth L1 in 11996 (holocene calendar). It used an orbit similar to ISEE-3.

So we will consider in space an isolated system of two bodies A and B, of mass M_A and M_B in gravitational interaction. These two bodies are assumed to be in circular orbit (for simplicity!) one around the other, in the manner of a two-star system (binary system) or a planet-satellite like system (Saturn-Titan example). We seek to determine if there are relative equilibrium point to the system of the two rotating body for a third body also in circular motion in the same plane (of sufficiently low mass to avoid disturbing the motion of the system of the two main bodies).



Let O be the centroid (see section Classical Mechanics page 3213) of these two stars (or celestial objects in general). Let us consider a Galilean reference frame (in rectilinear and uniform motion therefore !) or origin O. Compared to this reference frame, we assume that the axis AB rotates at a constant angular velocity ω of fixed axis $\vec{k}$ (perpendicular to the page in the figure and directed towards the reader) and that the distances $r_A=\overline{\text{AO}}$ and $r_B=\overline{\text{BO}}$ also remain constant.

We know from our study of Classical Mechanics that a circular motion the centrifugal (fictional) force is given by (see page 3388):

$$F=ma=m\frac{v^2}{R}=m\frac{(\omega R)^2}{R}=m\omega^2 R \tag{12.1.781}$$

So we have (equilibrium between (fictional) centrifugal and centripetal forces) to guarantee the equilibrium:

$$M_A\omega^2 r_A=\frac{GM_AM_B}{(r_A+r_B)^2} \qquad \text{and} \qquad M_B\omega^2 r_B=\frac{GM_AM_B}{(r_A+r_B)^2} \tag{12.1.782}$$

By simplifying and summing these two relations: By simplifying and summing these two relations:

$$\begin{aligned}\omega^2 r_A+\omega^2 r_B&=\frac{GM_A}{(r_A+r_B)^2}+\frac{GM_B}{(r_A+r_B)^2}\\ \rightarrow\omega^2&=\frac{G(M_A+M_B)}{(r_A+r_B)^3}=\frac{G(M_A+M_B)}{R^3}\end{aligned} \tag{12.1.783}$$

with in what will follow $AB=r_A+r_B=R$.

Let us consider a rotating reference frame R' linked to our stars as shown in figure above: $\vec{i}$ will be a collinear unit vector to AB, $\vec{j}$ a unit vector perpendicular to $\vec{i}$ and in the rotating plane of the planets and finally $\vec{k}=\vec{i}\times\vec{j}$ collinear to $\vec{\omega}$.

We consider in this rotating frame (with stars) a third star S of mass m negligible relatively to M_A and M_B, subject to the gravitational attraction of A and B.

Now let us denote by $\vec{a}_{R'}$ the acceleration of S with respect to R', $\vec{v}_{R'}$ its speed and $\vec{e}_r$ the collinear unit vector to $\overrightarrow{\mathrm{O}S'}$ where S' is the projection of S in the plane $\mathrm{O}xy$, and $r=\overline{\mathrm{O}S'}$ (in the figure above, we assumed S in the plane $\mathrm{O}xy$, so S and S' are indistinguishable).

S is thus subjected to two forces, one $\vec{F}_A$ directed along A and the other $\vec{F}_B$ directed along B, forces of respective intensities :

$$\vec{F}_A=-\frac{GM_Am}{\overline{AS}^3}\overrightarrow{AS} \qquad \text{and} \qquad \vec{F}_B=-\frac{GM_Bm}{\overline{BS}^3}\overrightarrow{BS} \tag{12.1.784}$$

In a Galilean reference frame, these two forces apply to S an acceleration given by the law of composition of accelerations in a circular reference frame (see section Classical Mechanics page 3383):

$$\frac{1}{m}(\vec{F}_A+\vec{F}_B)=\vec{a}_a=\vec{a}_e+\vec{a}_r+2\vec{\omega}\times\vec{v}_r+\frac{\mathrm{d}\omega}{\mathrm{d}t}\times\vec{r}+\vec{\omega}\times(\vec{\omega}\times\vec{r}) \tag{12.1.785}$$

But, in our configuration the pulsation (radial velocity) is assumed constant and the drive acceleration is zero since we assumed R' as the main repository. Therefore it comes:

$$\frac{1}{m}(\vec{F}_A+\vec{F}_B)=\vec{a}_r+2\vec{\omega}\times\vec{v}_r+\vec{\omega}\times(\vec{\omega}\times\vec{r})=\vec{a}_{R'}+2\vec{\omega}\times\vec{v}_r+2\vec{\omega}\times\vec{v}_r+\vec{\omega}\times(\vec{\omega}\times\vec{r}) \tag{12.1.786}$$

We also have:

$$\vec{\omega}=\begin{pmatrix}0\\0\\\omega_z\end{pmatrix} \qquad \vec{\omega}=\begin{pmatrix}r_x\\r_y\\0\end{pmatrix} \tag{12.1.787}$$

where according to the figure all components are positive. The calculation of the cross product then gives (see section Vector Calculus page 1830):

$$\vec{\omega}\times(\vec{\omega}\times\vec{r})=\begin{pmatrix}-\omega^2r_x\\-\omega^2r_y\\0\end{pmatrix}=-\omega^2\begin{pmatrix}r_x\\r_y\\0\end{pmatrix}=-\omega^2\|\vec{r}\|\frac{\vec{r}}{\|\vec{r}\|}=-\omega^2r\vec{e}_r \tag{12.1.788}$$

So finally:

$$\frac{1}{m}(\vec{F}_A+\vec{F}_B)=\vec{a}_{R'}+2\vec{\omega}\times\vec{v}_r+\vec{\omega}\times(\vec{\omega}\times\vec{r})=\vec{a}_{R'}+2\vec{\omega}\times\vec{v}_r-\omega^2r\vec{e}_r \tag{12.1.789}$$

Let us rather write this relation into the form:

$$\vec{a}_{R'}=-\frac{1}{m}(\vec{F}_A+\vec{F}_B)+\omega^2r\vec{e}_r-2\vec{\omega}\times\vec{v}_r \tag{12.1.790}$$

We then obtain, by projecting on the three axes x, y and z, the derivatives taken with respect to time t the following system:

$$\begin{aligned}x''&=-\frac{GM_A}{\overline{AS}^3}(x+r_A)-\frac{GM_B}{\overline{BS}^2}(x-r_B)+\omega^2x-2\omega y'\\y''&-\frac{GM_A}{\overline{AS}^2}y-\frac{GM_B}{\overline{BS}^3}y+\omega^2y+2\omega x'\\z''&=-\frac{GM_A}{\overline{AS}^3}z-\frac{GM_B}{\overline{BS}^3}z\end{aligned} \tag{12.1.791}$$

with:

$$\overline{AS}=\sqrt{(x+r_A)^2+y^2+z^2} \qquad \text{and} \qquad \overline{BS}=\sqrt{(x-r_B)^2+y^2+z^2} \tag{12.1.792}$$

so that the coordinates (x,y,z) of the point S are those of an equilibrium point, then it is trivial that in the rotating frame with the stars A and B that:

$$x''=y''=z''=0 \qquad \text{and} \qquad x'=y'=z'=0 \tag{12.1.793}$$

We then get the following system:

$$\begin{aligned}&-\frac{GM_A}{\overline{AS}^3}(x+r_A)-\frac{GM_B}{\overline{BS}^3}(x-r_B)+\omega^2x=0\\&-\frac{GM_A}{\overline{AS}^3}y-\frac{GM_B}{\overline{BS}^3}y+\omega^2y=0\\&-\frac{GM_A}{\overline{AS}^3}z-\frac{GM_B}{\overline{BS}^3}z=0\end{aligned} \tag{12.1.794}$$

It is also immediately that the third equation has for only solution $z=0$ and thus ultimately the system reduces to:

$$\begin{aligned}&-\frac{GM_A}{\overline{AS}^3}(x+r_A)-\frac{GM_B}{\overline{BS}^3}(x-r_B)+\omega^2x=0\\&-\frac{GM_A}{\overline{AS}^3}y-\frac{GM_B}{\overline{BS}^3}y+\omega^2y=0\\&z=0\end{aligned} \tag{12.1.795}$$

The third equation simply means that the equilibrium positions are in the plane Oxy (we could suspect it a bit ...). The other two, we will see it later, lead us to consider five solutions that are our five Lagrangian points L1, ..., L5.

If we draw plot with an appropriate software the acceleration (respectively force) with the isoclines highlighted (curves on which the acceleration is equal) we get:

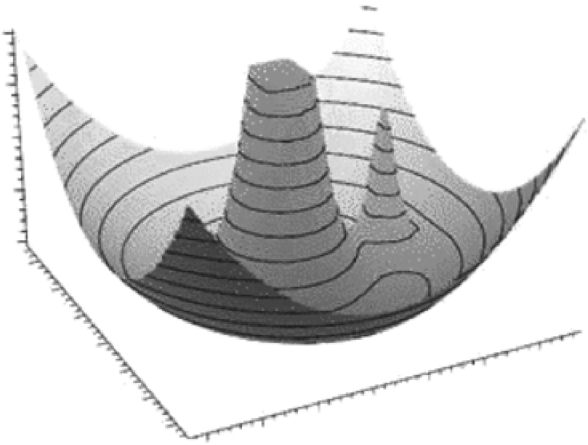


Figure 12.129: **Isoclines of the two-body system**

where we see that a short distance of the bodies the gravitational potential energy dominates, but that a large distances the centrifugal potential predominates and the shape of the surface is similar to that of a paraboloid.

By requesting the software to plot only the isoclines projected on a plane we get:

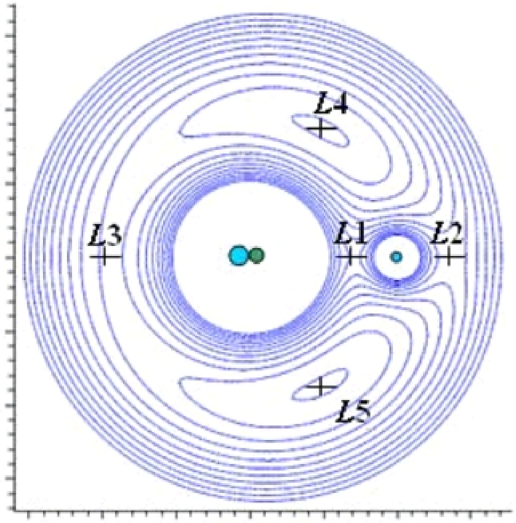


Figure 12.130: **Projected isoclines of the two body system on a plane**

where we have highlighted the five Lagrange points and where the stars (or celestial objects) are represented by blue dots and the centroid of the system by a green dot. It seems that isoclines are named in the astronomy field "Roche equipotentials lobes". Otherwise seen:

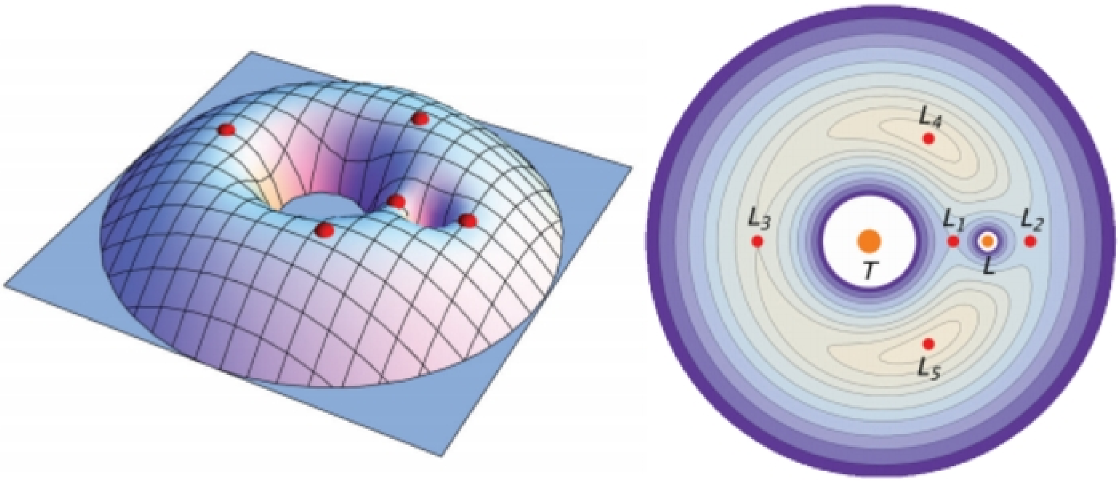


Figure 12.131: **Projected isoclines of the two body system on a plane**

For those wishing to reproduce these figures with MATLAB® here is how first proceed mathematically. From what we got previously, we have explicitly and in writing in a more academic form, the following relation:

$$\begin{aligned}
m\vec{a}_{R'} &= -(\vec{F}_A+\vec{F}_B)+m\omega^2 r\vec{e}_r-2m\vec{\omega}\times\vec{v}_r\\
&=-\left(-\frac{GM_Am}{\overline{AS}^3}\overrightarrow{AS}-\frac{GM_Bm}{\overline{BS}^3}\overrightarrow{BS}\right)+m\omega^2\vec{r}-2\begin{pmatrix}0\\0\\\omega_z\end{pmatrix}\times\begin{pmatrix}v_x\\v_y\\0\end{pmatrix}\\
&=\frac{GM_Am}{\overline{AS}^3}\overrightarrow{AS}+\frac{GM_Bm}{\overline{BS}^3}\overrightarrow{BS}+m\omega^2\vec{r}-2m\begin{pmatrix}-\omega_z v_y\\\omega_z v_x\\0\end{pmatrix}\\
&=\frac{GM_Am}{||\overrightarrow{AS}||^2}\frac{\overrightarrow{AS}}{||\overrightarrow{AS}||}+\frac{GM_Bm}{||\overrightarrow{BS}||^2}\frac{\overrightarrow{BS}}{||\overrightarrow{BS}||}+m\omega^2\vec{r}-2m\begin{pmatrix}-\omega_z v_y\\\omega_z v_x\\0\end{pmatrix}
\end{aligned}\tag{12.1.796}$$

As the point S is supposedly in equilibrium the last term vanishes (its speeds are zero in the rotating frame!). It then remains:

$$m\vec{a}_{R'}=\frac{GM_Am}{||\overrightarrow{AS}||^2}\frac{\overrightarrow{AS}}{||\overrightarrow{AS}||}+\frac{GM_Bm}{||\overrightarrow{BS}||^2}\frac{\overrightarrow{BS}}{||\overrightarrow{BS}||}+m\omega^2\vec{r}\tag{12.1.797}$$

We have proved for recall in the section of Classical Mechanics that:

$$\vec{F}=-\vec{\nabla}\cdot E_p\tag{12.1.798}$$

Then we have:

$$E_p=-\frac{GM_Am}{||\overrightarrow{AS}||}-\frac{GM_Bm}{||\overrightarrow{BS}||}-m\omega^2||\vec{r}||\tag{12.1.799}$$

Thus taken by unit mass of the satellite:

$$\boxed{E_{p,\text{eff}}=-\frac{GM_A}{||\overrightarrow{AS}||}-\frac{GM_B}{||\overrightarrow{BS}||}-\omega^2||\vec{r}||}\tag{12.1.800}$$

The application of these relations in MATLAB® 2013a then gives (sorry it's a bit long and its probably possible to do better ...):

```
%We build the grid plot that we will by anticipation densifiate where are the objects of interest
x1=linspace(-7E8,-8E5,150);
x2=linspace(8E5,1.2E8,150);
x3=linspace(1.6E8,7E8,150);
x=x1+x2+x3;
y=linspace(-7E8,7E8,450);
[X,Y]=meshgrid(x,y);
%These masses and G are real but the rest is fictitious so that the plot is readable
f=@(x,y) -(1.3346E20)./(sqrt((x-450).^2+y.^2))-(1.0038E19)./(sqrt((x-449999550).^2+y.^2))-(6.9E-7.*(x.^2+y.^2));
z=f(X,Y);
%We eliminate the values that are too big on Z to have an esthetical result to see
```

```
for i=1:450;
   for j=1:150;
      if (z(i,j)<-0.8E12) %to do with meshc a nice plot, limit to -8E11
         z(i,j)=-0.8E12;
      end;
   end;
end;
contour(X,Y,z,100); mesh(X,Y,z); meshc(X,Y,z);
az = 100; el = 25; view(az, el);
axis([-7E8 7E8 -0.8E9 0.8E9 -8E11 -4E11]);
colorbar; light; camlight('right');
```

That gives:

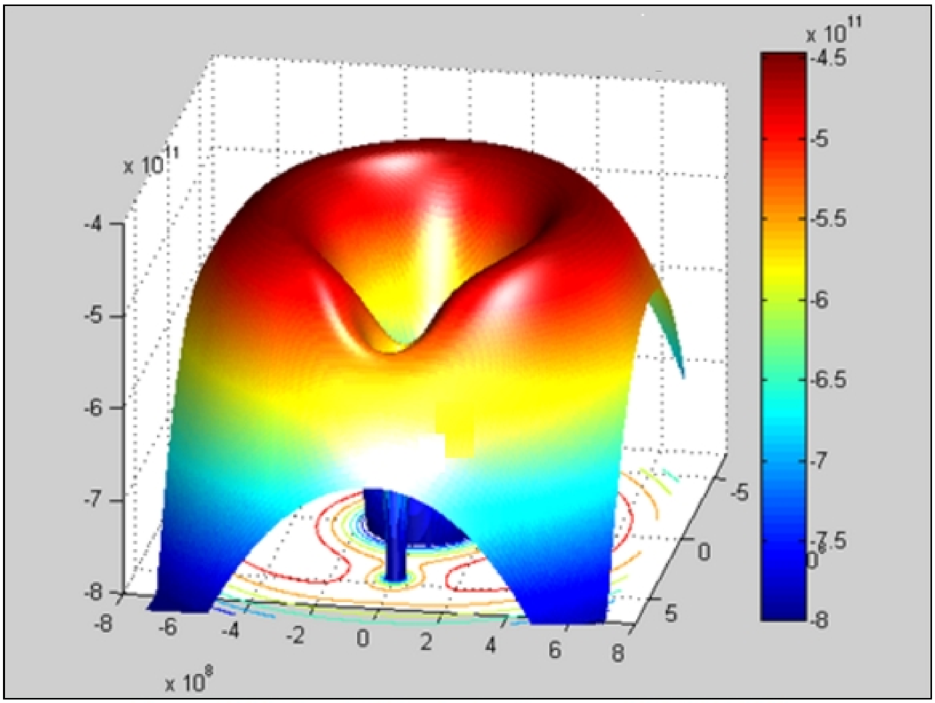


Figure 12.132: Lagrange plot and isoclines in 3D with MATLAB® 2013a

Or in 2D:

```
%  Author: David J. Muraki
%  15 jan 2004
%  coordinates

L = 15;   xr =  -L:0.05:L;   xi = xr;

[zr zi] = meshgrid(xr,xi);   zz = zr + i*zi;

%  first integral
GMs = 10.0;   Gme =  1.0;   R = 10;
As   = - R*Gme/(GMs+Gme);
ae   =   R*GMs/(GMs+Gme);
```

```
om2 = (GMs+Gme)/(R^3);

pot = om2*abs(zz).^2 + 2*GMs./abs(zz-As) + 2*Gme./abs(zz-ae);

force = om2*zz ...
      - GMs*(zz-As)./(abs(zz-As).^3) - Gme*(zz-ae)./(abs(zz-ae).^3)
         ;

%  contourplot
figure(1);  clf;  hold on
contour(xr,xi,pot,[3.2:0.05:4]);
axis([-1 1 -1 1]*L);
axis square;  colorbar
hold on
plot(As,0,’k*’);  plot(ae,0,’k*’)

ptitle = [’\bf potential energy & zero force contours’];
title(ptitle)
xlabel(’\bf Re(x)’)
ylabel(’\bf Im(x)’)

contour(xr,xi,real(force),[0 0],’k’);
contour(xr,xi,imag(force),[0 0],’k--’);
```

That gives:

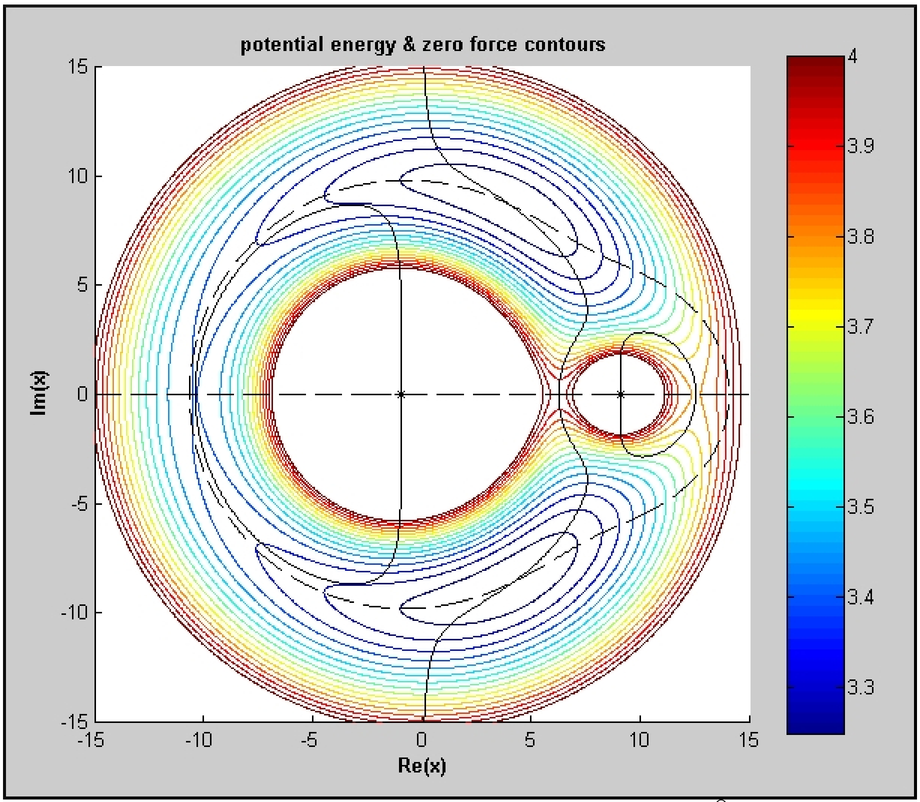


Figure 12.133: **Lagrange plot and isoclines in 2D with MATLAB® 2013a**

Some readers will notice that it is quite difficult to intuitively build such a potential configuration just by mind! In the rotating frame with the centroid of the two solid bodies, the potential resulting from the combination of rotational and gravitational potentials present three extrema L1, L2 and L3 on the right containing the both bodies. One of these maxima is between the two bodies as expected intuitively. The other two maxima are on the line connecting the two objects, but both on either side... which is more surprising. They come from the contribution to the potential of the rotating frame which can be difficult to model intuitively.

12.1.15.1 Equilibrium points of the first type

What we mean by equilibrium positions of the first type are simply solutions located on the line $\overline{AB}$ such that $y=0$ which is equivalent to study only:

$$-\frac{GM_A}{\overline{AS}^3}(x+r_A)-\frac{GM_B}{\overline{BS}^3}(x-r_B)+\omega^2 x=0 \tag{12.1.801}$$

with therefore:

$$\overline{AS}=|x+r_A| \quad \text{and} \quad \overline{BS}=|x-r_B| \tag{12.1.802}$$

To this situation, we will consider three possible corresponding sub-cases respectively L1, L2 and L3 as we will immediately see it.

12.1.15.1.1 L1 Lagrange point

In this first sub-case, we consider:

$$-r_A<x<r_B \tag{12.1.803}$$

What is also equivalent to have:

$$\begin{gathered}(x+r_A)>0\\(x-r_B)<0\end{gathered} \tag{12.1.804}$$

This allows us to write:

$$-\frac{GM_A}{\overline{AS}^3}(x+r_A)-\frac{GM_B}{\overline{BS}^3}(x-r_B)+\omega^2 x=0 \tag{12.1.805}$$

in the following simplified form:

$$-\frac{GM_A}{\overline{AS}^2}-\frac{GM_B}{\overline{BS}^2}+\omega^2 x=0 \tag{12.1.806}$$

Now to be able to say something about the possible solutions to this equation derive the left hand side. We then get:

$$\frac{2GM_A}{(x+r_A)^3}+\frac{2GM_B}{(r_B-x)^3}+\omega^2=0 \tag{12.1.807}$$

This term is strictly increasing from $-\infty$ to $+\infty$ when x describes $]-r_A, r_B[$. So there is a unique solution and equilibrium point denoted L1 (first Lagrange point) between A and B.

If we typically consider the Sun-Earth case where $M_A > M_B$ and therefore $r_A < r_B$ then on $x = 0$ we have:

$$-\frac{GM_A}{(0+r_A)^3}(0+r_A) - \frac{GM_B}{0-r_b}^3(0-r_B)+\omega^2 0 = -\frac{GM_A}{r_A^2} - \frac{GM_B}{r_B^2} \qquad (12.1.808)$$

what is immediately negative. The equilibrium position will be obtained for a positive value of x we will have to determine.

This value can be obtained by considering a limit case: when M_B tends to 0 (corresponding to a massive celestial object A turning around a mass B of a much much smaller celestial object), then A tends to O, r_A tends to 0 and therefore:

$$\omega^2 = \frac{GM_A}{R^3} \qquad (12.1.809)$$

with $R = \overline{AB}$. Therefore, in this limit case:

$$-\frac{GM_A}{\overline{AS}^2} + \frac{GM_B}{\overline{BS}^2} + \omega^2 x = 0 \qquad (12.1.810)$$

becomes in approximation:

$$-\frac{GM_A}{x^2} + \omega^2 x = 0 \qquad (12.1.811)$$

and therefore:

$$-\frac{GM_A}{x^2} + \left(\frac{GM_A}{R^3}\right) x = 0 \qquad (12.1.812)$$

So the only value of x satisfying this relation will be $x = R$.

In other words, the equilibrium point L1 we are looking after here is between A and B moves near B, that is near the less massive celestial object (which corresponds well to the first figure that we used to show the location of the five Lagrange points).

By this observation we can make the following calculations:

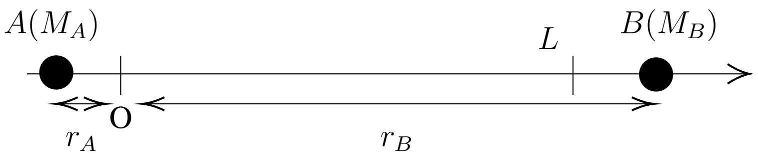


Figure 12.134: Configuration to mathematically determine the position of point L1

We have from the definition of center of gravity (see section Classical Mechanics page 3213):

$$\vec{r}_G = \frac{M_A\vec{r}_A + M_B\vec{r}_B}{M_A + M_B} \qquad (12.1.813)$$

As our study is done relatively to the centroid we have $\vec{r}_G=\vec{0}$ and therefore:

$$M_A\vec{r}_A+M_B\vec{r}_B=\vec{0} \tag{12.1.814}$$

From the above relation by taking the norm, we have obviously:

$$r_A=\frac{M_B}{M_A}r_B \tag{12.1.815}$$

The distance between the two celestial objects A and B remaining constant and being equal to $r=r_A+r_B$ we write:

$$\overline{AB}=r=r_A+r_B=\frac{M_B}{M_A}r_B+r_B=r_B\left(1+\frac{M_B}{M_A}\right)=\overline{\text{O}B}\left(1+\frac{M_B}{M_A}\right) \tag{12.1.816}$$

We deduce trivially two relations (the second being obtained by exactly the same reasoning as the first):

$$\overline{\text{O}B}=\frac{\overline{AB}}{1+\dfrac{M_B}{M_A}}\qquad\text{and}\qquad\overline{\text{O}A}=\frac{\overline{AB}}{1+\dfrac{M_A}{M_B}} \tag{12.1.817}$$

But since $M_A\gg M_B$ we can write roughly the first relation in the following approximate form (Taylor series):

$$\overline{\text{O}B}\cong\text{AB}\left(1-\frac{M_B}{M_A}\right) \tag{12.1.818}$$

and since:

$$r_A=\frac{M_B}{M_A}r_B\Leftrightarrow\overline{\text{O}A}=\frac{M_B}{M_A}\overline{\text{O}B} \tag{12.1.819}$$

we have also:

$$\overline{\text{O}A}=\frac{M_B}{M_A}\overline{\text{O}B}=\frac{M_B}{M_A}\frac{\overline{AB}}{1+\dfrac{M_B}{M_A}} \tag{12.1.820}$$

So with $M_A\gg M_B$:

$$\overline{\text{O}A}\cong\frac{M_B}{M_A}\overline{AB} \tag{12.1.821}$$

According to the limiting case studied previously, we can assume L at the neighbourhood of B such that it is possible to write:

$$\overline{\text{O}L}\cong\overline{\text{O}B}(1-\varepsilon) \tag{12.1.822}$$

with $\varepsilon\ll 1$.

Either using:

$$\overline{\text{O}B}\cong\overline{AB}\left(1-\frac{M_B}{M_A}\right) \tag{12.1.823}$$

Then we have:

$$\overline{\mathrm{O}L} \cong \overline{AB}(1-\varepsilon)\left(1-\frac{M_B}{M_A}\right)=\overline{AB}\left(1-\varepsilon-\frac{M_B}{M_A}\right) \tag{12.1.824}$$

by neglecting the infinitely small therm of order 2.

Hence:

$$\begin{aligned}\overline{LB}&=\overline{\mathrm{O}B}-\overline{\mathrm{O}L}=\overline{AB}\left(1-\frac{M_B}{M_A}\right)-\overline{AB}\left(1-\varepsilon-\frac{M_B}{M_A}\right)\\ \overline{LA}&=\overline{\mathrm{O}A}+\overline{\mathrm{O}L}=\frac{M_B}{M_A}\overline{AB}+\overline{AB}\left(1-\varepsilon-\frac{M_B}{M_A}\right)=\overline{AB}(1-\varepsilon)\end{aligned} \tag{12.1.825}$$

Now in the mentioned configuration the equilibrium is given by:

$$-\frac{GM_A}{\overline{LA}^2}+\frac{GM_B}{\overline{LB}^2}+\omega^2\overline{\mathrm{O}L}=0 \tag{12.1.826}$$

Therefore:

$$-\frac{GM_A}{\overline{AB}^2(1-\varepsilon)^2}+\frac{GM_B}{\varepsilon^2\overline{AB}^2}+\omega^2\overline{AB}\left(1-\varepsilon-\frac{M_B}{M_A}\right)=0 \tag{12.1.827}$$

Now the third Kepler's law gives us:

$$\frac{T^2}{\overline{AB}^3}=\frac{4\pi^2}{GM_A}\Rightarrow\frac{GM_A}{\overline{AB}^3}=\frac{4\pi^2}{T^2}=\left(\frac{2\pi}{T}\right)^2=\left(\frac{2\pi}{\frac{2\pi}{\omega}}\right)^2=\omega^2 \tag{12.1.828}$$

Therefore:

$$-\frac{GM_A}{\overline{AB}^2(1-\varepsilon)^2}+\frac{GM_B}{\varepsilon^2\overline{AB}^2}+G\frac{M_A}{\overline{AB}^3}\overline{AB}\left(1-\varepsilon-\frac{M_B}{M_A}\right)=0 \tag{12.1.829}$$

After simplification:

$$-\frac{M_A}{(1-\varepsilon)^2}+\frac{M_B}{\varepsilon^2}+M_A\left(1-\varepsilon-\frac{M_B}{M_A}\right)=0 \quad -\frac{M_A}{M_B}\frac{1}{(1-\varepsilon)^2}+\frac{1}{\varepsilon^2}+\frac{M_A}{M_B}\left(1-\varepsilon-\frac{M_B}{M_A}\right)=0 \tag{12.1.830}$$

Therefore:

$$\frac{1}{\varepsilon^2}-\frac{M_A}{M_B}(1+2\varepsilon)+\frac{M_A}{M_B}\left(1-\varepsilon-\frac{M_B}{M_A}\right)\cong 0 \tag{12.1.831}$$

Hence:

$$\frac{1}{\varepsilon^2}-3\frac{M_A}{M_B}\varepsilon-1\cong 0 \tag{12.1.832}$$

Since $1/\varepsilon^2$ is much greater than 1 and assuming that $3M_A/M_B$ then we have also:

$$\frac{1}{\varepsilon^2}\cong 3\frac{M_A}{M_B}\varepsilon \tag{12.1.833}$$

Thus finally:

$$\varepsilon \cong \sqrt[3]{\frac{M_B}{3M_A}} \tag{12.1.834}$$

and therefore:

$$\begin{aligned}\overline{LA}&=\overline{AB}(1-\varepsilon)=\overline{AB}\left(1-\sqrt[3]{\frac{M_B}{3M_A}}\right)=R\left(1-\sqrt[3]{\frac{M_B}{3M_A}}\right)\\ \overline{LB}&=\overline{AB}\varepsilon=R\sqrt[3]{\frac{M_B}{3M_A}}\end{aligned} \tag{12.1.835}$$

If we take the A for the Sun and B for the Earth, then:

$$\frac{M_B}{M_A}\cong 3\cdot 10^{-6} \qquad R\cong 150\cdot 10^6\ [\text{km}] \tag{12.1.836}$$

We find that the distance $\overline{LB}$ is then equal approximately to:

$$\overline{LB}=R\sqrt[3]{\frac{M_B}{3M_A}}\cong 1,500,000\ [\text{km}] \tag{12.1.837}$$

which is the L1 point where was placed the satellite SOHO (since the latter will thus never have its field of view obscured by the shadow of the Earth or the Moon).

A special case of the L1 point to consider is when $M_A=M_B=M$, then $r_A=r_B=r$, then O is the midpoint of $\overline{AB}$. Then we have:

$$\omega^2=\frac{GM}{R^3}=\frac{GM}{(2r)^3}=\frac{GM}{8r^3} \tag{12.1.838}$$

Therefore:

$$-\frac{GM_A}{\overline{AS}^2}+\frac{GM_B}{\overline{BS}^2}+\omega^2x=0 \tag{12.1.839}$$

becomes:

$$-\frac{GM}{(r+x)^2}+\frac{GM}{(x-r)^2}+\omega^2x=0\Rightarrow -\frac{1}{(r+x)^2}+\frac{1}{(x-r)^2}=-\frac{x}{8r^3} \tag{12.1.840}$$

Figure 12.135: **Non-composite image of the far side of the moon, illuminated by the Sun, as it crosses between the DISCOVR spacecraft's Earth Polychromatic Imaging Camera (EPIC) camera and telescope at Lagrange position L1, and the Earth (source: NASA)**

12.1.15.1.2 L2 Lagrange point

In this second sub-case, we consider:

$$x > r_B \tag{12.1.841}$$

Therefore we are looking for the equilibrium points beyond B.

Thus we have:

$$-\frac{GM_A}{AS^3}(x+r_A) - \frac{GM_B}{BS^3}(x-r_B) + \omega^2 x = 0 \tag{12.1.842}$$

which becomes simply:

$$-\frac{GM_A}{AS^2} - \frac{GM_B}{BS^2} + \omega^2 x = 0 \tag{12.1.843}$$

The left side is a strictly increasing function of x from $-\infty$ to $+\infty$ when x describes $[r_B, +\infty]$. So there is a unique solution, and an equilibrium point beyond B. This point is denoted: L2.

This value can be obtained by considering a limit case: when M_B tends to 0 (corresponding to a massive celestial object on A around which a much smaller mass object B turn around), then A tends to O, r_A to 0 and therefore:

$$\omega^2 = \frac{GM_A}{R^3} \tag{12.1.844}$$

with $R=\overline{AB}$. Therefore, in this limit case:

$$-\frac{GM_A}{\overline{AS}^3}-\frac{GM_B}{\overline{BS}^2}+\omega^2 x=0 \tag{12.1.845}$$

becomes approximately:

$$-\frac{GM_A}{x^2}+\omega^2 x=0 \tag{12.1.846}$$

and so:

$$-\frac{GM_A}{x^2}+\frac{GM_A}{R^3}x=0 \tag{12.1.847}$$

So the only value of x satisfying this relation will be $x=R$. The L2 point therefore ends up being merged with B.

Knowing this limit case, let us do a more detailed study. Consider the following diagram relatively to our previous limit case:

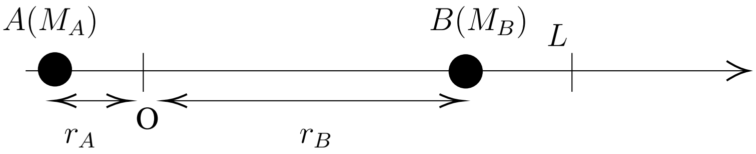


Figure 12.136: Configuration to mathematically determine the position of point L2

and let us consider $M_A \gg M_B$ without forgetting that in this scenario $x>r_B$.

We then have almost the same developments as for L1 but with the difference that:

$$\overline{\mathrm{O}L}\cong\overline{\mathrm{O}B}(1-\varepsilon) \tag{12.1.848}$$

becomes:

$$\overline{\mathrm{O}L}\cong\overline{\mathrm{O}B}(1+\varepsilon) \tag{12.1.849}$$

and that instead of having:

$$-\frac{GM_A}{\overline{LA}^2}+\frac{GM_B}{\overline{LB}^2}+\omega^2\overline{\mathrm{O}L}=0 \tag{12.1.850}$$

We have:

$$-\frac{GM_A}{\overline{LA}^2}-\frac{GM_B}{\overline{LB}^2}+\omega^2\overline{\mathrm{O}L}=0 \tag{12.1.851}$$

and therefore:

$$\overline{LA}=\overline{AB}(1+\varepsilon) \tag{12.1.852}$$

Still with:

$$\varepsilon=\sqrt[3]{\frac{M_B}{3M_A}} \tag{12.1.853}$$

and therefore:

$$\overline{LA}=\overline{AB}(1+\varepsilon)=\overline{AB}\left(1+\sqrt[3]{\frac{M_B}{3M_A}}\right)=R\left(1+\sqrt[3]{\frac{M_B}{3M_A}}\right) \tag{12.1.854}$$

which corresponds to the Lagrange point L2.

A special case again about L2 is when $M_A=M_B=M$, then $r_A=r_B=r$, then O is at the midpoint of $\overline{AB}$. Then we have:

$$\omega^2=\frac{GM}{R^3}=\frac{GM}{(2r)^3}=\frac{GM}{8r^3} \tag{12.1.855}$$

Therefore:

$$-\frac{GM_A}{\overline{AS}^2}-\frac{GM_B}{\overline{BS}^2}+\omega^2x=0 \tag{12.1.856}$$

becomes:

$$-\frac{GM}{(r+x)^2}-\frac{GM}{(x-r)^2}+\omega^2x=0\Rightarrow\frac{1}{(r+x)^2}+\frac{1}{(x-r)^2}=\frac{x}{8r^3} \tag{12.1.857}$$

It is no longer possible to extract the roots here (at least to my knowledge). It must be done through a numerical approximation. In Maple™ 4.00b, we simply put:

```
>solve(-1/(r+x)^2-1/(x-r)^2=x/(8*r^3),x);allvalues(");
```

and the only feasible solution in $\mathbb{R}$ is then $x\cong 2.8r$ and the others being in $\mathbb{C}$.

12.1.15.1.3 L3 Lagrange point

In this third sub-case, we consider:

$$x < -r_A \tag{12.1.858}$$

So we look for the equilibrium points beyond A.

Thus we have:

$$-\frac{GM_A}{\overline{AS}^3}(x+r_A)-\frac{GM_B}{\overline{BS}^3}(x-r_B)+\omega^2 x=0 \tag{12.1.859}$$

which becomes simply:

$$-\frac{GM_A}{\overline{AS}^2}-\frac{GM_B}{\overline{BS}^2}+\omega^2 x=0 \tag{12.1.860}$$

The left side is an increasing function of x from $-\infty$ to $+\infty$ when x describes $]-r_A,-\infty]$. So there is a unique solution, and one equilibrium point beyond A. This point is denoted: L3.

This value can be obtained by considering a limit case: when M_B tends to 0 (corresponding to a massive celestial object A turning around much smaller object B), then A tends to O, r_A to 0 and therefore:

$$\omega^2=\frac{GM_A}{R^3} \tag{12.1.861}$$

with $R=\overline{AB}$. Therefore, in this limit case:

$$-\frac{GM_A}{\overline{AS}^2}-\frac{GM_B}{\overline{BS}^2}+\omega^2 x=0 \tag{12.1.862}$$

becomes approximately:

$$-\frac{GM_A}{x^2}+\omega^2 x=0 \tag{12.1.863}$$

and therefore:

$$-\frac{GM_A}{x^2}+\frac{GM_A}{R^3}x=0 \tag{12.1.864}$$

So the only value of x satisfying this relation will be $x=R$. The point L3 will finish to merge with the position diametrically opposite to that of B.

Knowing this limit case, let us do a more detailed study now. Consider the following diagram relative to our previous limit situation:

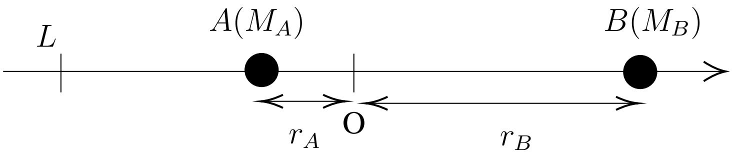


Figure 12.137: Configuration to mathematically determine the position of point L3

and let us still consider $M_A \gg M_B$ without forgetting that in this scenario $x < -r_A$.

We will first consider the following approximation:

$$\overline{AL} \cong \overline{AB}(1+\varepsilon) = R(1+\varepsilon) \tag{12.1.865}$$

and this one also (since $\overline{\mathrm{O}A}$ tends to zero as the celestial object A becomes very massive):

$$\overline{\mathrm{O}A} \cong \frac{M_B}{M_A}R \tag{12.1.866}$$

Since then:

$$\overline{\mathrm{O}L} = \overline{\mathrm{O}A} + \overline{AL} \cong R\left(1+\varepsilon+\frac{M_B}{M_A}\right) \tag{12.1.867}$$

We have also (...):

$$\omega^2 = \frac{GM_A}{R^3} \cong \frac{GM_A}{R^3}\left(1+\frac{M_B}{M_A}\right) \tag{12.1.868}$$

when at the limit where the celestial object A is really massive, we fall back on the first term ...

With the last two relations, we have:

$$\omega^2\overline{\mathrm{O}L} = R\left(1+\varepsilon+\frac{M_B}{M_A}\frac{GM_A}{R^3}\right)\left(1+\frac{M_B}{M_A}\right) \cong \frac{GM_A}{R^2}\left(1+\varepsilon+2\frac{M_B}{M_A}\right) \tag{12.1.869}$$

if we neglect the terms of the second order.

Furthermore, we have also:

$$\overline{BL} = \overline{AB} + \overline{AL} = R + R(1+\varepsilon) = 2R\left(1+\frac{\varepsilon}{2}\right) \tag{12.1.870}$$

Let us recall the equilibrium condition:

$$-\frac{GM_A}{\overline{AS}^2} - \frac{GM_B}{\overline{BS}^2} + \omega^2 x = -\frac{GM_A}{\overline{AL}^2} - \frac{GM_B}{\overline{BL}^2} + \omega^2\overline{\mathrm{O}L} = 0 \tag{12.1.871}$$

And let us put everything we got until now inside it:

$$-\frac{GM_A}{(R(1+\varepsilon))^2} - \frac{GM_B}{\left(2R\left(1+\frac{\varepsilon}{2}\right)\right)^2} + \frac{GM_A}{R^2}\left(1+\varepsilon+2\frac{M_B}{M_A}\right) = 0 \tag{12.1.872}$$

What becomes after simplifications:

$$\frac{M_A}{M_B}\frac{1}{(1+\varepsilon)^2} + \frac{1}{4\left(1+\frac{\varepsilon}{2}\right)^2} = \frac{M_A}{M_B}\left(1+\varepsilon+2\frac{M_B}{M_A}\right) \tag{12.1.873}$$

after a small approximation:

$$4\frac{M_A}{M_B}(1-2\varepsilon) + (1-\varepsilon) = 4\frac{M_A}{M_B}\left(1+\varepsilon+2\frac{M_B}{M_A}\right) \tag{12.1.874}$$

after simplification:

$$-12\varepsilon\frac{M_A}{M_B}=7 \tag{12.1.875}$$

Hence:

$$\varepsilon=-\frac{7}{12}\frac{M_B}{M_A} \tag{12.1.876}$$

and finally:

$$\overline{OB}\cong R\left(1+\varepsilon+\frac{M_B}{M_A}\right)\cong R\left(1-\frac{7}{12}\frac{M_B}{M_A}+\frac{M_B}{M_A}\right)=R\left(1+\frac{5}{12}\frac{M_B}{M_A}\right) \tag{12.1.877}$$

Remark

For some Si-Fi authors, for reminder... this L3 point opposite to the Earth relatively to the Sun (actually a G2V-type star, aka yellow dwarf) would hide us a hypothetical planet that would be forever hidden to us by the Sun..........

12.1.15.2 Equilibrium points of the second type

The equilibrium positions of the second type are those for which $y\neq 0$. In other words the points outside of the line $\overline{AB}$, but still in the plane Oxy.

Thus, our system of equations remains:

$$\begin{aligned}&-\frac{GM_A}{\overline{AS}^3}(x+r_A)-\frac{GM_B}{\overline{BS}^3}(x-r_A)+\omega^2x=0\\&-\frac{GM_A}{\overline{AS}^3}y-\frac{GM_B}{\overline{BS}^3}y+\omega^2y=0\\&Z=0\end{aligned} \tag{12.1.878}$$

12.1.15.2.1 L4, L5 Lagrange points

To determine the remaining equilibrium points, we can divide the second equation of the system such that the latter becomes:

$$\begin{aligned}&-\frac{GM_A}{\overline{AS}^3}(x+r_A)-\frac{GM_B}{\overline{BS}^3}(x-r_A)+\omega^2x=0\\&-\frac{GM_A}{\overline{AS}^3}-\frac{GM_B}{\overline{BS}^3}+\omega^2=0\end{aligned} \tag{12.1.879}$$

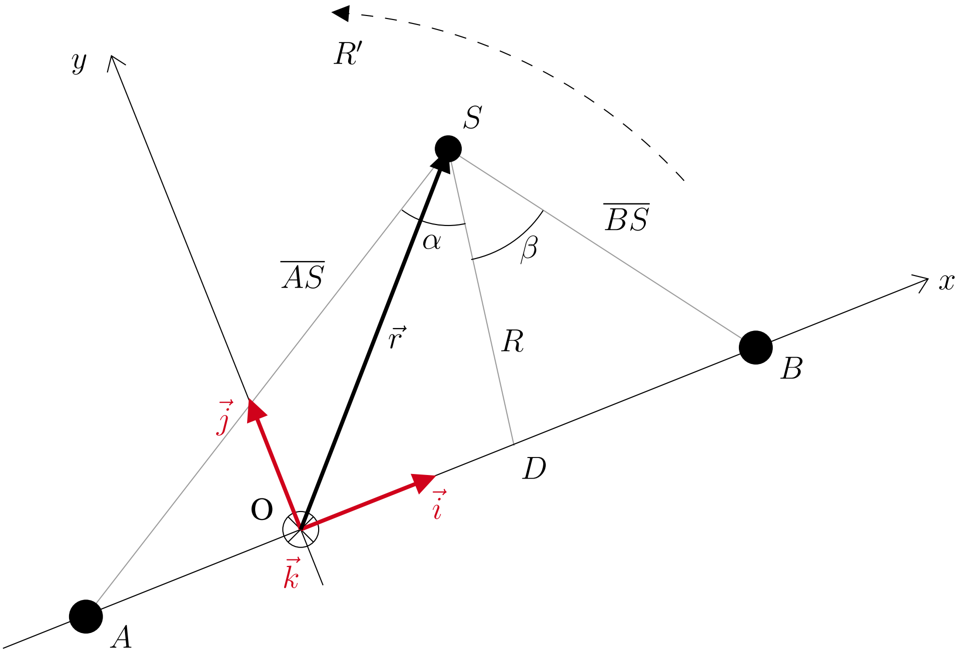


Figure 12.138: **Configuration to mathematically determine the position of points L4,L5**

where $\overline{AB}$ is obviously the distance between A and B and D is the centroid of the system given by (see section Classical Mechanics page 3213):

$$\overline{AD}=\overline{AB}\frac{M_B}{M_A+M_B} \qquad \text{or} \qquad \overline{DB}=\overline{AB}\frac{M_A}{M_A+M_B} \tag{12.1.880}$$

which are the radii of gyration of the bodies A and B.

It is easy to verify that the sum of both previous distances is equal to $\overline{AB}$ and their proportion M_B/M_A. Another form of $\overline{DB}$ (which will be useful) is obtained by dividing the numerator and denominator by M_A:

$$\overline{DB}=\frac{\overline{AB}}{1+\dfrac{M_B}{M_A}} \tag{12.1.881}$$

We know according to our previous calculations that $\overline{AS}=\overline{BS}$ but this is insufficient. We still want to know the angles of the vertices A, B, S, and this is what we will look for now.

In this context, if a satellite S is in equilibrium, there will always remain at the same distance of A or B. The center of rotation of the 3 points is the point D, the mass A itself revolves around it. If the satellite, S, remains stable, the three bodies have the same orbital period T. If S is immobile in this frame in rotation it will not be subject to the Coriolis force but only to centrifugal (fictional) force of A and of B.

Let us denote by v_B the rotation speed of B and v_S the rotation speed of S. Then we have:

$$2\pi R=v_S T \tag{12.1.882}$$

and:

$$2\pi\frac{\overline{AB}}{1+\dfrac{M_B}{M_A}}=v_B T \tag{12.1.883}$$

From whose we get that:

$$\frac{2\pi}{T}=\frac{v_S}{R} \tag{12.1.884}$$

and:

$$\frac{2\pi}{T}=\frac{v_B}{\overline{AB}}\left(1+\frac{M_B}{M_A}\right) \tag{12.1.885}$$

So we can equate these two expressions:

$$\frac{v_S}{R}=\frac{v_B}{\overline{AB}}\left(1+\frac{M_B}{M_A}\right) \tag{12.1.886}$$

This merely expresses the well known fact that if two objects rotate together, the furthest one from the centroid is the fastest. Speeds are proportional to the distances from the centroid.

The centrifugal (fictional) force on B is in equilibrium with the gravitational force of A and it is expressed by:

$$M_B\frac{v_B^2}{\left(\dfrac{\overline{AB}}{1+\dfrac{M_B}{M_A}}\right)}=\frac{M_B}{\overline{AB}}v_B^2\left(1+\frac{M_B}{M_A}\right)=G\frac{M_AM_B}{\overline{AB}^2} \tag{12.1.887}$$

Thus by simplifying:

$$G\frac{M_A}{\overline{AB}}=v_B^2\left(1+\frac{M_B}{M_A}\right) \tag{12.1.888}$$

Similarly, the centrifugal (fictional) force applied on S is:

$$M_S\frac{v_S^2}{R} \tag{12.1.889}$$

It is balanced by the forces of attraction $\vec{F}_A$,$\vec{F}_B$ of the objects A and B. However, only the components of these forces located on the line R oppose efficiently to this centrifugal (fictional) force. Hence:

$$M_S\frac{v_S^2}{R}=F_A\cos(\alpha)+F_B\cos(\beta) \tag{12.1.890}$$

$$M_S\frac{v_S^2}{R}=F_A\cos(\alpha)+F_B\cos(\beta) \tag{12.1.891}$$

and as:

$$F_A=\frac{GM_AM_S}{\overline{AS}^2} \quad \text{and} \quad F_B=\frac{GM_BM_S}{\overline{BS}^2} \tag{12.1.892}$$

We then have:

$$\frac{v_S^2}{R}=\frac{GM_A}{\overline{AS}^2}\cos(\alpha)+\frac{GM_B}{\overline{BS}^2}\cos(\beta) \tag{12.1.893}$$

In addition, the forces applied to S and perpendicular to R must vanish. If not, the object S would follow the largest mass and would not remain in position and would therefore no longer be in equilibrium. We must then have:

$$F_A \sin(\alpha) = F_B \sin(\beta) \tag{12.1.894}$$

Or, after substitution and simplification:

$$F_A \sin(\alpha) = F_B \sin(\beta) \tag{12.1.895}$$

Of all the equations obtained up to now the only that bother us are those containing both speeds and angles α,β. This requires that we must arrive to eliminate what is convenient to have only the last two parameters (that is to say: the angles).

For this, we take the square:

$$\frac{v_S^2}{R^2} = \frac{v_B^2}{\overline{AB}^2}\left(1+\frac{M_B}{M_A}\right)^2 \tag{12.1.896}$$

We multiply both sides by $\overline{AB}^2$ and we divide by $1+\dfrac{M_B}{M_A}$:

$$\frac{v_S^2\overline{AB}^2}{R^2\left(1+\dfrac{M_B}{M_A}\right)} = v_B^2\left(1+\frac{M_B}{M_A}\right) \tag{12.1.897}$$

which is similar to:

$$G\frac{M_A}{\overline{AB}} = v_B^2\left(1+\frac{M_B}{M_A}\right) \tag{12.1.898}$$

Thus equating:

$$\frac{v_S^2\overline{AB}^2}{R^2\left(1+\dfrac{M_B}{M_A}\right)} = G\frac{M_A}{\overline{AB}} \tag{12.1.899}$$

So we removed the speed of B. Now, let us multiply both sides by $\left(1+\dfrac{M_B}{M_A}\right)R$ and divide by $\overline{AB}^2$:

$$\frac{v_S^2}{R} = \frac{GM_A}{\overline{AB}^3}R\left(1+\frac{M_B}{M_A}\right) \tag{12.1.900}$$

which is similar to:

$$\frac{v_S^2}{R} = \frac{GM_A}{\overline{AS}^2}\cos(\alpha) + \frac{GM_B}{\overline{BS}^2}\cos(\beta) \tag{12.1.901}$$

Therefore:

$$\frac{GM_A}{\overline{AB}^2}\frac{R}{\overline{AB}}\left(1+\frac{M_B}{M_A}\right) = \frac{GM_A}{\overline{AS}^2}\cos(\alpha) + \frac{GM_B}{\overline{BS}^2}\cos(\beta) \tag{12.1.902}$$

By dividing the whole by GM_A we find:

$$\frac{1}{\overline{AB}^2}\frac{R}{\overline{AB}}\left(1+\frac{M_A}{M_B}\right)=\frac{1}{\overline{AS}^2}=\frac{1}{\overline{AS}^2}\cos(\alpha)+\frac{M_B}{M_A}\frac{1}{\overline{BS}^2}\cos(\beta) \tag{12.1.903}$$

And as we have proved at the beginning $\overline{AS}=\overline{BS}$ that we will denote by R', then we have:

$$\frac{R'^2}{\overline{AB}^2}\frac{R}{\overline{AB}\left(1+\dfrac{M_B}{M_A}\right)}=\cos(\alpha)+\frac{M_B}{M_A}\cos(\beta) \tag{12.1.904}$$

and let us recall that we have:

$$\frac{M_A}{\overline{AS}^2}\sin(\alpha)=\frac{M_B}{\overline{BS}^2}\sin(\beta) \tag{12.1.905}$$

Therefore:

$$\frac{M_B}{M_A}=\frac{\sin(\alpha)}{\sin(\beta)} \tag{12.1.906}$$

This allows us to write:

$$\frac{R'^2}{\overline{AB}^2}\frac{R}{\overline{AB}}\left(1+\frac{M_B}{M_A}\right)=\cos(\alpha)+\frac{\sin(\alpha)}{\sin(\beta)}\cos(\beta) \tag{12.1.907}$$

And multiplying by $\sin(\beta)$:

$$\frac{R'^2}{\overline{AB}^2}\sin(\beta)\frac{R}{\overline{AB}}\left(1+\frac{M_B}{M_A}\right)=\sin(\beta)\cos(\alpha)+\sin(\alpha)\cos(\beta)=\sin(\alpha+\beta) \tag{12.1.908}$$

We can now notice a thing (not easy to see...). If $R'=\overline{AB}$ (that is that the triangle ABS is equilateral) the previous relations simplifies to:

$$\sin(\beta)\frac{1}{\overline{AB}}\left(1+\frac{M_B}{M_A}\right)=\frac{\sin(\hat{S})}{R} \tag{12.1.909}$$

But, if the triangle is really equilateral, then we have (see section Geometric Shapes page 2902):

$$\sin(\hat{S})=\sin(\hat{A})=\sin(\hat{\beta}) \tag{12.1.910}$$

Hence:

$$\frac{\sin(\beta)}{\overline{AB}\dfrac{M_A}{M_A+M_B}}=\frac{\sin(\hat{\beta})}{R} \tag{12.1.911}$$

What can finally write:

$$\frac{\sin(\beta)}{\overline{DB}}=\frac{\sin(\hat{\beta})}{R} \tag{12.1.912}$$

Which is just the sine theorem for the triangle SDB (see section Trigonometry page 2528) and is therefore certain. Returning back, we can now prove that all previous equations are satisfied if and only if ABS is equilateral. If we had not put ABS as equilateral, we would have gotten a different relation of the sine theorem, without possible verification, and the set of equations required for equilibrium at the point S could not be met.

Conclusion of the thing ... the system gives as a solution:

$$\overline{AS}=\overline{BS}=R \tag{12.1.913}$$

ABS (or ABL regardless the notation), then forms an equilateral triangle. The two equilibrium points are denoted L4 and L5. L4 is located in advance with respect to the less massive celestial object and L5 is late relatively to it.

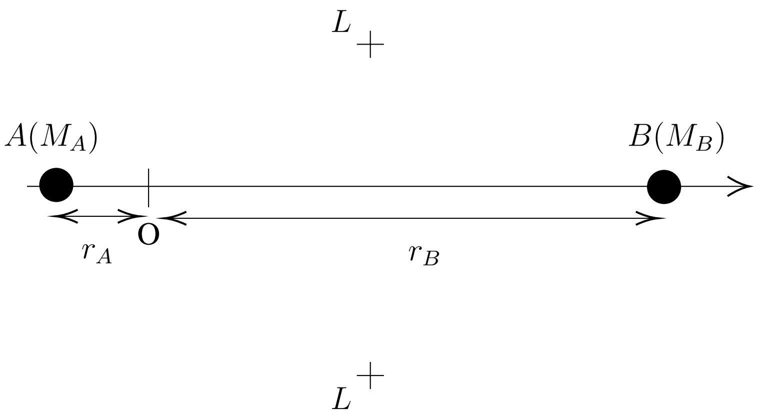


Figure 12.139: **L4 and L5 equilateral triangle**

In the year 12000 (holocene calendar), 385 asteroids in the L4 point and 188 asteroids in the L5 point were counted on the orbit of Jupiter, but located precisely in an equilateral triangle with the Sun (actually a G2V-type star, aka yellow dwarf) and either side of Jupiter: these are the Trojan planets. It was also observed two objects at the point L5 of Mars discovered in 11990 and 11998 (holocene calendar).

For people reading this textbook on a computer with Adobe Flash player installed and activated, here is an animation that shows a few others "stable" (because on the very long term almost nothing is really stable!) three bodies planar periodic orbits made with Python™ (otherwise see here: `https://vimeo.com/1059418615`):

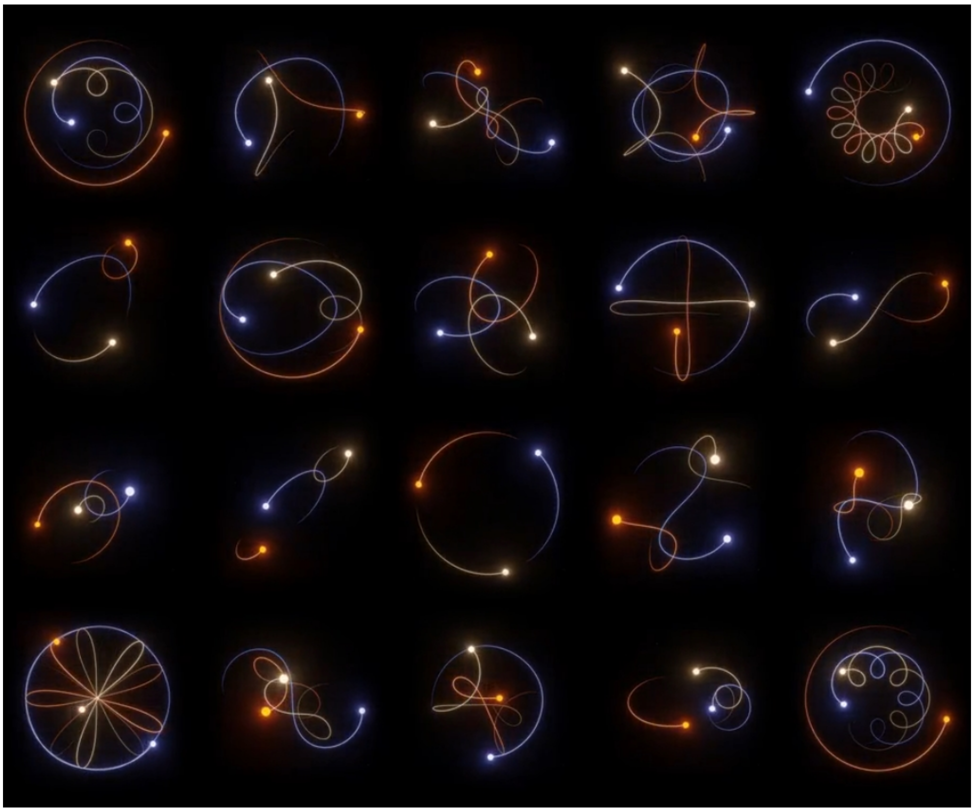


Figure 12.140: **Animation of few three bodies planar periodic orbits made with Python™**

The reader can visit Ricky Reusser Python™ notebook at https://observablehq.com/@rreusser/periodic-planar-three-body-orbits to recreate 2703 periodic solutions of the planar three-body problem.

12.1.15.3 Hill sphere (sphere of influence)

The "Hill sphere", or "sphere of influence" (S.O.I.), is a common model for the calculation of a gravitational sphere of influence. It is the most commonly used model to calculate the spatial extent of gravitational influence of an astronomical body (m) in which it dominates over the gravitational influence of other bodies, particularly a primary (M).[It is sometimes confused with other models of gravitational influence, such as the Laplace sphere or being named the Roche sphere, the latter causing confusion with the Roche limit. It was defined by the American astronomer George William Hill, based on the work of the French astronomer Édouard Roche.

To be retained by a more gravitationally attracting astrophysical object - a planet by a more massive star, a moon by a more massive planet - the less massive body must have an orbit that lies within the gravitational potential represented by the more massive body's Hill sphere. That moon would, in turn, have a Hill sphere of its own, and any object within that distance would tend to become a satellite of the moon, rather than of the planet itself.

One simple view of the extent of the Solar System is that it is bounded by the Hill sphere of the Sun (engendered by the Sun's interaction with the galactic nucleus or other more massive stars). A more complex example is the one at right, the Earth's Hill sphere, which extends between the Lagrange points L1 and L2, which lie along the line of centers of the Earth and the more massive Sun. The gravitational influence of the less massive body is least in that direction, and so it acts as the limiting factor for the size of the Hill sphere; beyond that distance, a third object in orbit around the Earth would spend at least part of its orbit outside the Hill sphere, and would be progressively perturbed by the tidal forces of the more massive body, the Sun, eventually ending up orbiting the latter.

For the derivation let us consider the following three body system with circular orbits:

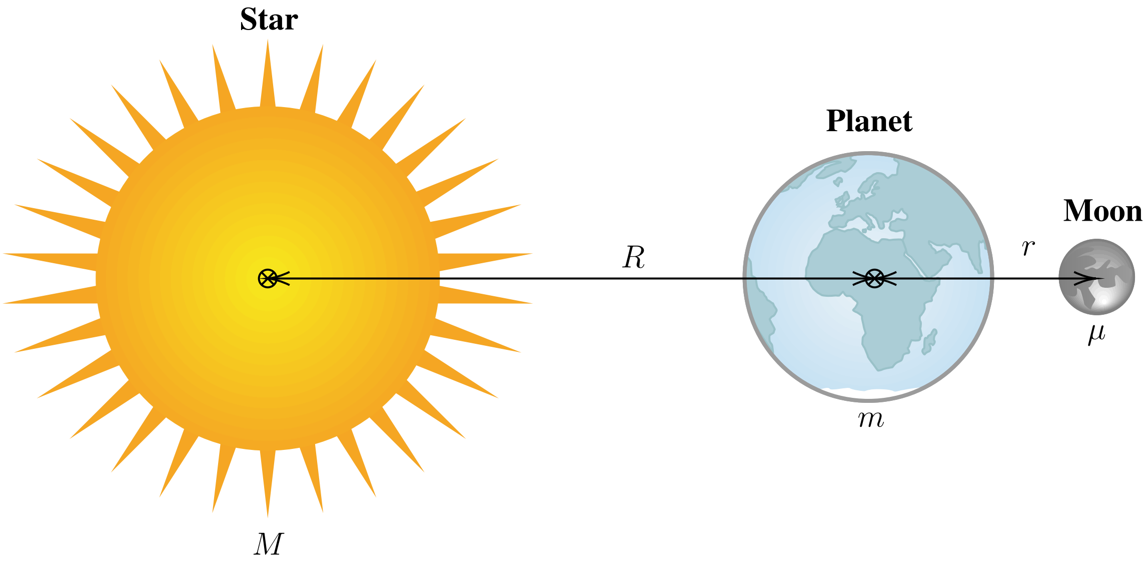


Figure 12.141: **Determining the Hill sphere radius**

The satellite or moon (mass μ) is orbiting the star (mass M) with the same angular velocity ω at the distance $R+r$ as the planet (mass m) at the distance R (permanent full moon position).

The equilibrium condition for the planet is obviously:

$$m\omega^2 R = G\frac{mM}{R^2} \tag{12.1.914}$$

Hence:

$$\omega^2 = \frac{GM}{R^3} \tag{12.1.915}$$

The satellite is dragged by the combined gravitational forces exerted by the star and the planet:

$$\mu\omega^2(R+r) = G\frac{\mu M}{(R+r)^2} + G\frac{\mu m}{r^2} \tag{12.1.916}$$

Inserting ω^2:

$$G\mu\frac{R+r}{R^3} = G\frac{\mu M}{(R+r)^2} + G\frac{\mu m}{r^2} \tag{12.1.917}$$

Simplifying by μ, we get:

$$G\frac{R+r}{R^3} = G\frac{M}{(R+r)^2} + G\frac{m}{r^2} \tag{12.1.918}$$

Multiplying both sides by $(R+r)^2R^3$, we get:

$$M(R+r)^3r^2 = MR^3r^2 + mR^3(R+r)^2 \qquad (12.1.919)$$

Rearranging and developing, we get:

$$mR^3(R+r)^2 = Mr^2(R^3+3R^2r+3Rr^2+r^3) - MR^3r^2 \qquad (12.1.920)$$

Factoring:

$$mR^3(R+r)^2 = Mr^3(3R^2+3Rr+r^2) \qquad (12.1.921)$$

For $r \ll R$ we have:$(R+r)^2 \cong R^2$, and $3Rr+r^2 \cong 0$. The relation simplifies:

$$mR^5 \cong 3Mr^3R^2 \qquad (12.1.922)$$

After trivial simplification:

$$mR^3 \cong 3Mr^3 \qquad (12.1.923)$$

So finally:

$$\boxed{r_{\text{Hill}} \cong R\sqrt[3]{\frac{m}{3M}}} \qquad (12.1.924)$$

For the triplet Sun, Earth, Moon, we get for example:

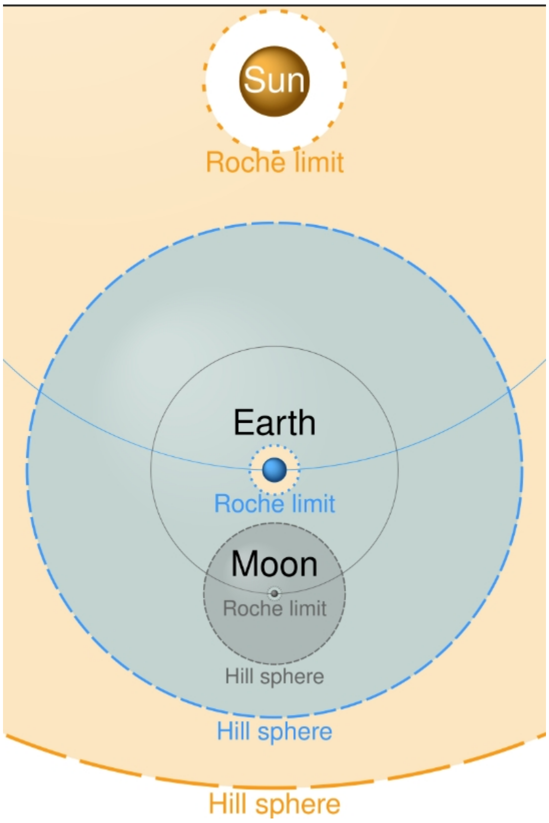
$$r_{\text{Hill}} \cong R\sqrt[3]{\frac{m_{\text{Earth}}}{3M_{\text{Sun}}}} \cong 149,600,000\sqrt[3]{\frac{5.97\cdot 10^{24}}{3(1.99\cdot 10^{33})}} \cong 1,496,000\;[\text{km}] \qquad (12.1.925)$$

So for this special case we have therefore:

$$r_{\text{Moon}} \cong \frac{1}{4}r_{\text{Hill}} \qquad (12.1.926)$$

This answer the famous question of many people of why the Moon is not falling down to the Sun even if the force of the Sun on the Moon is greater than the force of the Earth on the Moon!

Below a schematic, not-to-scale representation of Hill spheres (as 2D radii) and Roche limits of each body of the Sun-Earth-Moon system. The actual Hill radius for the Earth-Moon pair is on the order of $60,000$ [km] (i.e., extending less than one-sixth the distance of the $378,000$ [km] between the Moon and the Earth):


Sun
Roche limit
Earth
Roche limit
Moon
Roche limit
Hill sphere
Hill sphere
Hill sphere

The Hill sphere is only an approximation, and other forces (such as radiation pressure or the Yarkovsky effect) can eventually perturb an object out of the sphere. As stated, the satellite (third mass) should be small enough that its gravity contributes negligibly.

12.1.16 Relativistic Doppler-Fizeau Effect

The Doppler effect is the difference between the frequency of the transmitted wave and the received wave when the transmitter and receiver are moving relative to each other (see section Music Mathematics page 8212). This is an effect that must be take into account astronomy to calculate the distance of the body assuming its known (or estimated) emission wavelength and measuring its received wavelength or for measuring the speed of rotation (radial velocity) of stars by observing very precisely and successively their opposite edges and measuring the shift of the spectrum obtained.

In the early 121st century (holocene calendar) the precision and finesse of the spectra of measures has reached a level that allows to observe even minimal changes in the distance of stars and so speculate on possible planetary satellites (this may work if the plane of the orbit passes through the Earth):

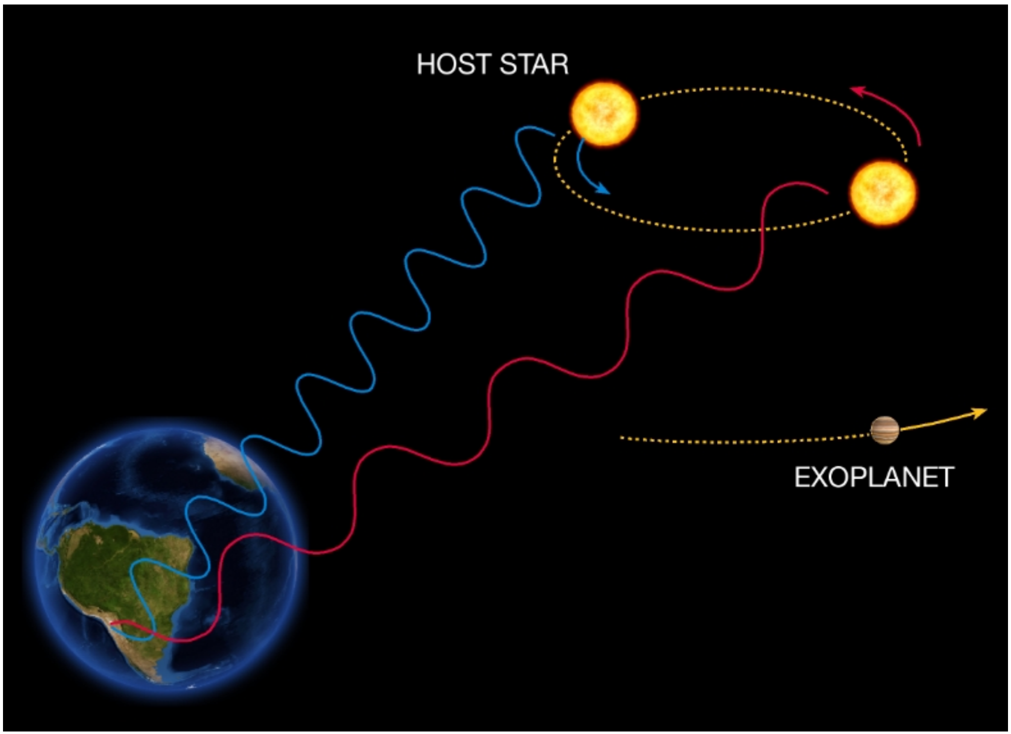


Figure 12.142: Doppler-Effect radial velocity method (source: ESO Press Photo 22e/05 12007-04-11)

The Doppler effect of electromagnetic waves must be discussed independently of the acoustic Doppler effect (also named "Galilean Doppler effect") studied in the section of Music Mathematics (see page 8212). First, because the electromagnetic waves do not consist of a material movement and therefore the speed of the source relative to the medium does not enter into the discussion, then because their velocity is c (the speed of light) and remains the same for all observers independently of their relative movements. The Doppler effect for electromagnetic waves is thus calculated necessarily using the principle of relativity and is symmetrical with respect to relative movement of the source and the observer (as opposed to acoustic cases).

For an observer in an inertial reference frame, a plane and harmonic electromagnetic wave can be described by a function of the form:

$$\sin(kx-\omega t) \tag{12.1.927}$$

multiplied by an appropriate amplitude factor. For an observer attached to another inertial frame, the components x and t should be replaced with x' and t', obtained by the Lorentz transformation (see section Special Relativity page 5633), and that latter will therefore write to describing plane wave:

$$\sin(k'x'-\omega't') \tag{12.1.928}$$

where k' and ω' are not necessarily the same as that of the another observer (precisely this is what we want to determine). Moreover, the principle of relativity has allowed us to demonstrate in the section of Special Relativity that:

$$\frac{\partial}{\partial t^2}-c^2\frac{\partial^2}{\partial x^2}=\frac{\partial}{\partial t'^2}-c^2\frac{\partial^2}{\partial x'^2} \tag{12.1.929}$$

This assumes that the expression:

$$kx-\omega t \tag{12.1.930}$$

remains invariant when we move from one inertial observer to the other. We then have:

$$kx-\omega t=k'x'-\omega't' \tag{12.1.931}$$

Using the Lorentz transformation relations (see section Special Relativity page 5633), we have immediately:

$$k\frac{x'+vt'}{\sqrt{1-\dfrac{v^2}{c^2}}}-\omega\frac{t'+v\dfrac{x'}{c^2}}{\sqrt{1-\dfrac{v^2}{c^2}}}=k'x'-\omega't' \tag{12.1.932}$$

Hence:

$$\frac{kx'+kvt'-\omega t'-\omega v\dfrac{x'}{c^2}}{\sqrt{1-\dfrac{v^2}{c^2}}}=k'x'-\omega't' \tag{12.1.933}$$

Grouping we get:

$$\frac{\left(k-\omega\dfrac{v}{c^2}\right)x'-(\omega-kv)\,t'}{\sqrt{1-\dfrac{v^2}{c^2}}}=k'x'-\omega't' \tag{12.1.934}$$

By identification, it comes immediately:

$$k'=\frac{k-\omega\dfrac{v}{c^2}}{\sqrt{1-\dfrac{v^2}{c^2}}}\qquad \omega'=\frac{\omega-kv}{\sqrt{1-\dfrac{v^2}{c^2}}} \tag{12.1.935}$$

If we consider that:

$$\omega=ck \tag{12.1.936}$$

in the case of electromagnetic waves, we can write each of these relation in the form:

$$\boxed{k'=k\frac{1-\dfrac{v}{c}}{\sqrt{1-\dfrac{v^2}{c^2}}}\qquad \omega'=\omega\frac{1-\dfrac{v}{c}}{\sqrt{1-\dfrac{v^2}{c^2}}}} \tag{12.1.937}$$

The ratio visible in the both expression above is named the "radial relativistic doppler redshift" and is denoted by:

$$\frac{1-\dfrac{v}{c}}{\sqrt{1-\dfrac{v^2}{c^2}}}=\left(1-\frac{v}{c}\right)\gamma \tag{12.1.938}$$

for a movement of the observer relative to the source in the direction of propagation (the range of this function is bounded to $[0,1]$). In the other case (opposite direction) we have obviously:

$$\frac{1+\dfrac{v}{c}}{\sqrt{1-\dfrac{v^2}{c^2}}}=\left(1+\frac{v}{c}\right)\gamma \tag{12.1.939}$$

That is unbounded and has for range $[1,+\infty[$. Physicists like to normalized stuffs so that the range is $[0,+\infty[$, so we write and define the "relativistic doppler redshift" for opposite direction as:

$$z=\frac{1+\dfrac{v}{c}}{\sqrt{1-\dfrac{v^2}{c^2}}}-=\left(1+\frac{v}{c}\right)\gamma-1 \tag{12.1.940}$$

Or more commonly written as:

$$\boxed{1+z=\frac{1+\dfrac{v}{c}}{\sqrt{1-\dfrac{v^2}{c^2}}}=\gamma\left(1+\frac{v}{c}\right)} \tag{12.1.941}$$

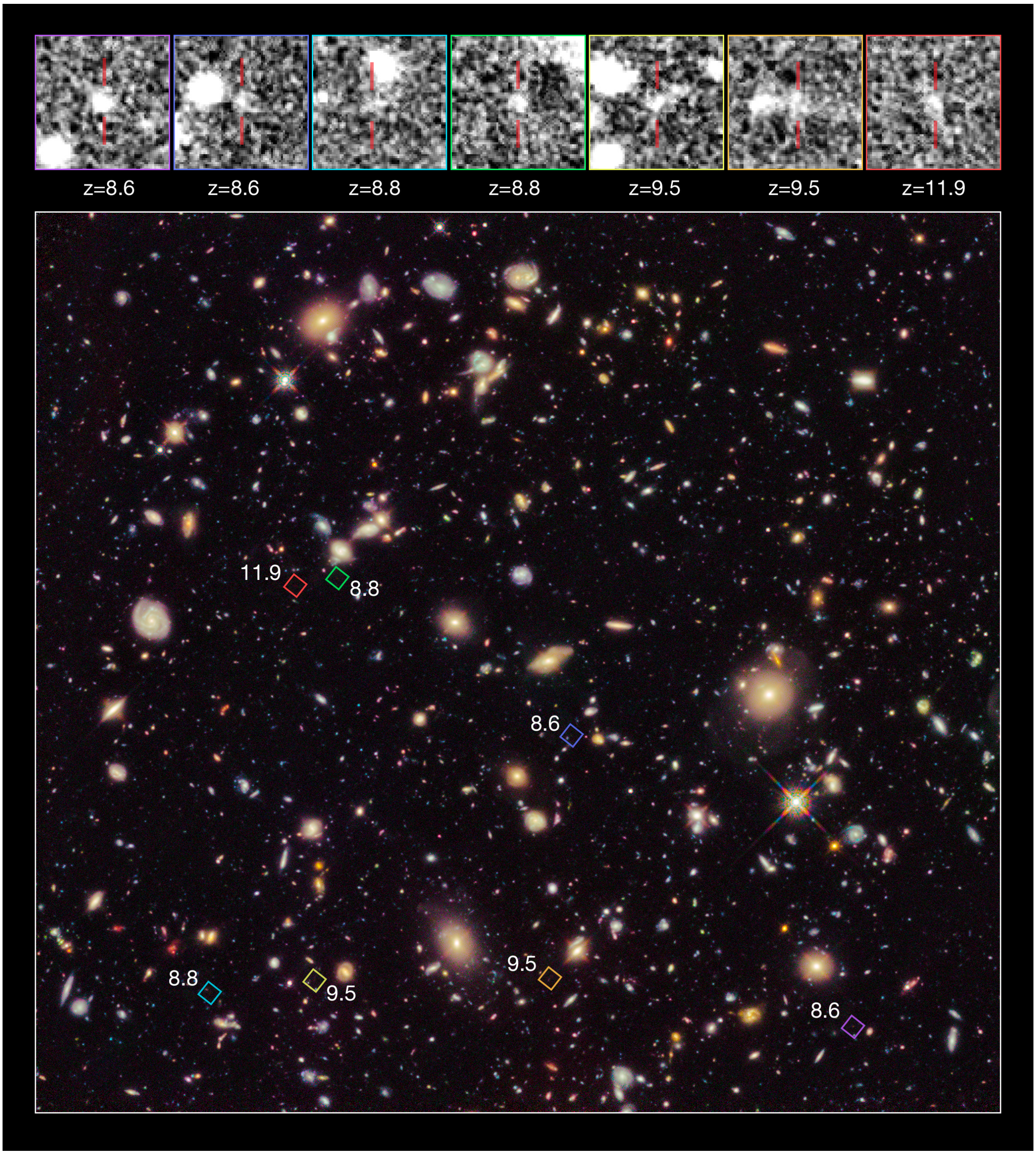

z=8.6
z=8.6
z=8.8
z=8.8
z=9.5
z=9.5
z=11.9
11.9
8.8
8.6
8.8
9.5
9.5
8.6

Figure 12.143: **High-redshift galaxy candidates in the Hubble ultra deep field 12012 (source: NASA, ESA, R. Ellis (Caltech), and the HUDF Team)**

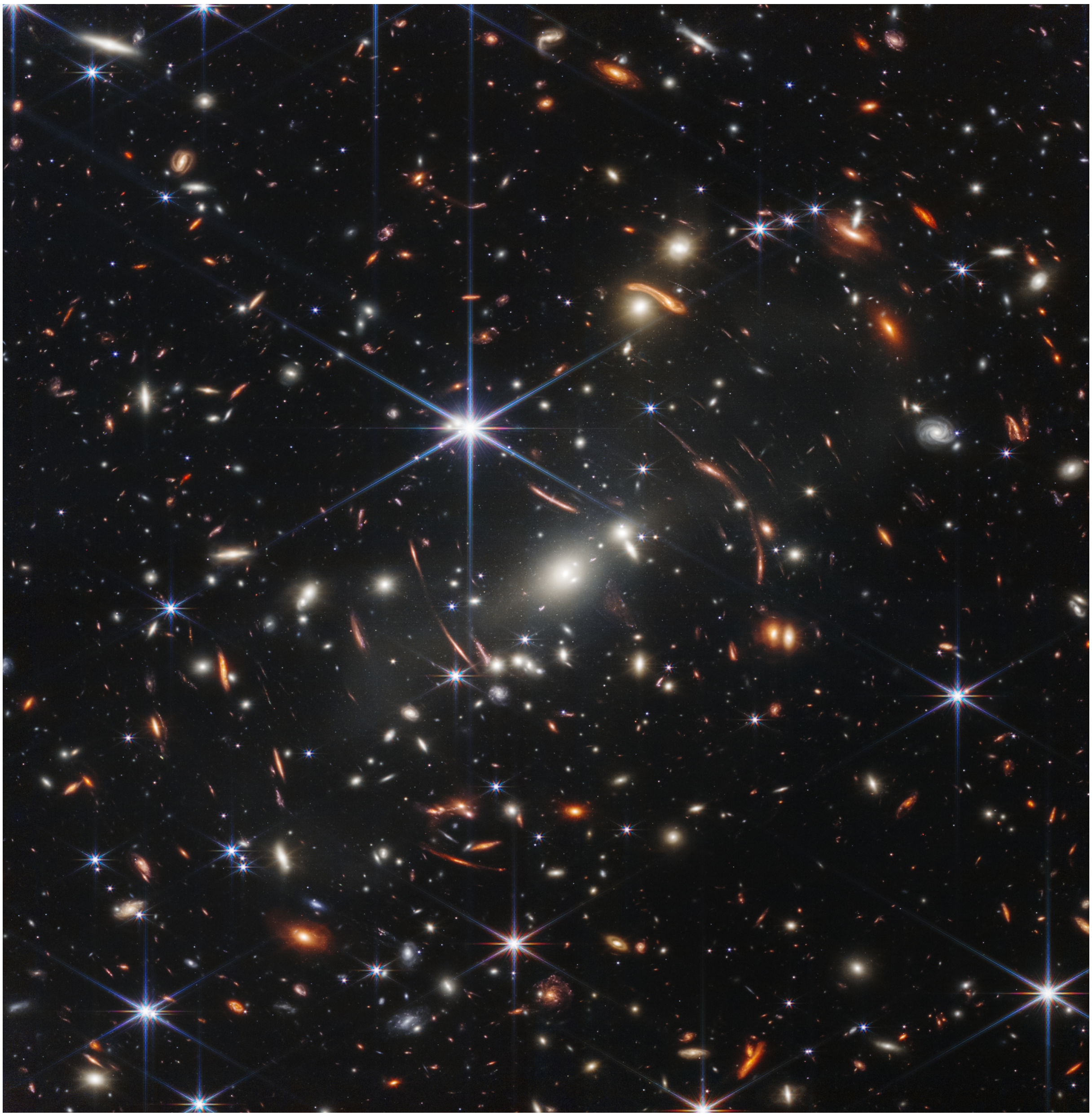


Figure 12.144: **James Webb Telescope galaxy cluster SMACS 0723 deep field released in 12022, July 11 (source: NASA)**

Obviously if the source or observer don't move away but approach then we not have a red shift but a "blueshift".

Furthermore, the last relation with the pulsations is most often written in the literature as follows:

$$2\pi f = 2\pi f' \frac{1+\dfrac{v}{c}}{\sqrt{1-\dfrac{v^2}{c^2}}} = 2\pi f' \frac{1+\dfrac{v}{c}}{\sqrt{\left(1-\dfrac{v}{c}\right)\left(1+\dfrac{v}{c}\right)}} = 2\pi f' \frac{\sqrt{1+\dfrac{v}{c}}}{\sqrt{1-\dfrac{v}{c}}} \qquad (12.1.942)$$

Which is written most often in the following form (keep in mind that this is in the case when the two referential move in opposite direction!):

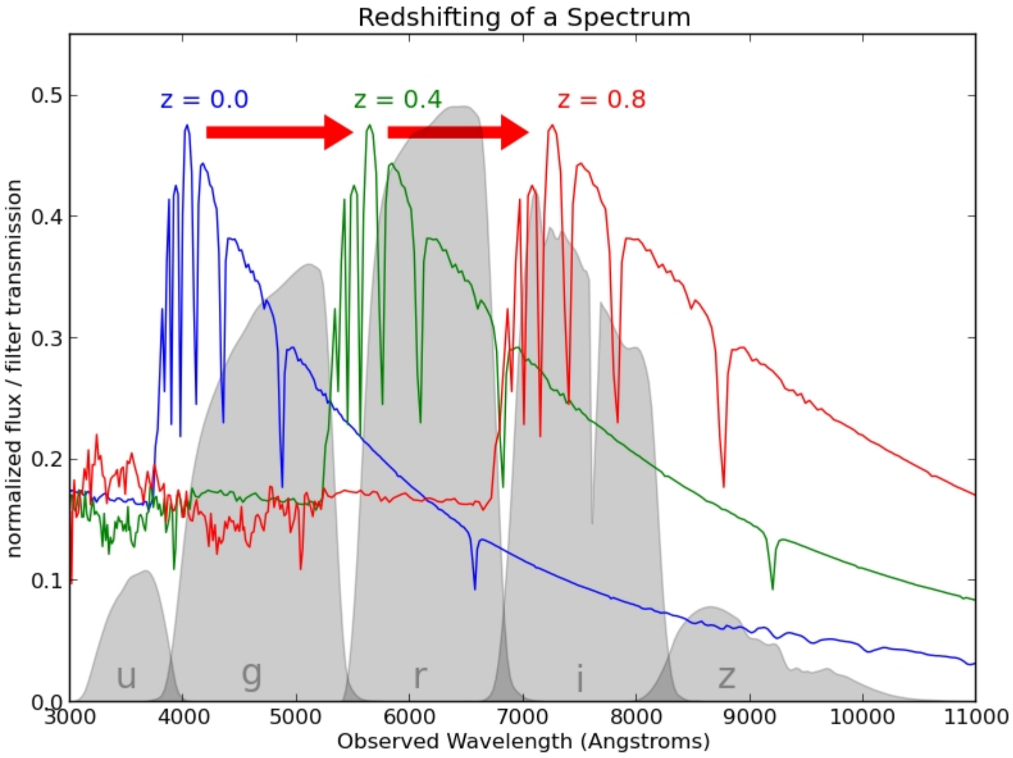
$$\boxed{\frac{f}{f'}=\frac{f_{\text{source}}}{f_{\text{observed}}}=\frac{1+\dfrac{v}{c}}{\sqrt{1-\dfrac{v^2}{c^2}}}=1+z=\sqrt{\frac{1+\beta}{1-\beta}}} \tag{12.1.943}$$

Therefore if we measure the both frequencies (supposing that we know what should be the source), then we can also obviously determine the speed v of the observed object.

Or in case of wavelength we have similarly:

$$\boxed{\frac{\lambda'}{\lambda}=\frac{\lambda_{\text{observed}}}{\lambda_{\text{source}}}=\frac{1+\dfrac{v}{c}}{\sqrt{1-\dfrac{v^2}{c^2}}}=1+z=\sqrt{\frac{1+\beta}{1-\beta}}} \tag{12.1.944}$$

When a spectrum can be obtained, determining the red shift is rather straight-forward: if you can localize the spectral fingerprint of a common element, such as hydrogen, then the red shift can be computed using simple arithmetic. But similarly to the case of star/quasar classification, the task becomes much more difficult when only photometric observations are available:



It must be recalled that the pulsation offset (and therefore frequency offset) that takes place here is due to a relative motion of the observer with respect to the source and not to something else (or respectively of the source relatively to the observer). Indeed, in our study of General Relativity (see section General Relativity page 5821), we will prove that there is a superposition of a shift because of the gravitational field surrounding the emitter that will be considered as caused by the space-time curvature.

Finally, for sceptics who want to check in another way that the Doppler phenomenon is well symmetric unlike the acoustic Doppler effect proved in the section of Music Mathematics, here's another approach:

First, consider that it is the source moving away. If we calculated by the classical relation proved in the section of Music Mathematics the frequency of the signal at the reception would be (see page 8212):

$$f_{\text{remoteness}}=f\dfrac{1}{1+\dfrac{v_{\text{source}}}{v_{\text{phase}}}}=f\dfrac{1}{1+\dfrac{v}{c}}=\dfrac{1}{\Delta t}\dfrac{1}{1+\beta} \tag{12.1.945}$$

and we must take into account the time dilation for f with (see section Special Relativity page 5659):

$$\Delta t=\dfrac{\Delta t'}{\sqrt{1-\left(\dfrac{v}{c}\right)^2}} \tag{12.1.946}$$

because the time interval of the fixed observer is longer than that of the source (time is faster for observer at rest).

It comes then:

$$\begin{aligned}f_{\text{remoteness}}&=\dfrac{\sqrt{1-\beta^2}}{\Delta t}\dfrac{1}{1+\beta}=f'\dfrac{\sqrt{1-\beta^2}}{1+\beta}=f'\sqrt{\sqrt{(1+\beta)(1-\beta)}}1+\beta\\&=f'\sqrt{\dfrac{(1+\beta)(1-\beta)}{(1+\beta)^2}}=f'\sqrt{\dfrac{1-\beta}{1+\beta}}\end{aligned} \tag{12.1.947}$$

and if it is the observer who moves away from the source we proved in the section of Music Mathematics that:

$$f_{\text{approach}}=f'\left(1-\dfrac{v_{\text{obs}}}{v_{\text{source}}}\right)=f'\left(1-\dfrac{v}{c}\right)=\dfrac{1}{\Delta t'}(1-\beta) \tag{12.1.948}$$

both relations are indeed symmetric in the special relativistic case (as expected for electrodynamics)!

Remark

Currently, astronomers have courses in astrophysics and their observations are generally studied in an astrophysical context, so there is less distinction between the two disciplines than before.

A very good example of the application of the Doppler effect is to explore the limits given by measuring the apparent speed. Let's see what it is:

12.1.16.1 Apparent speed

By measuring the apparent speed of movement of very fast objects in the sky (plasma jets, etc.), astrophysicists have obtained apparent displacement speeds exceeding the speed of light in vacuum!

In fact, it is an illusion that can occur if the speed of the object is very close to that of light it emits, so close enough to c.

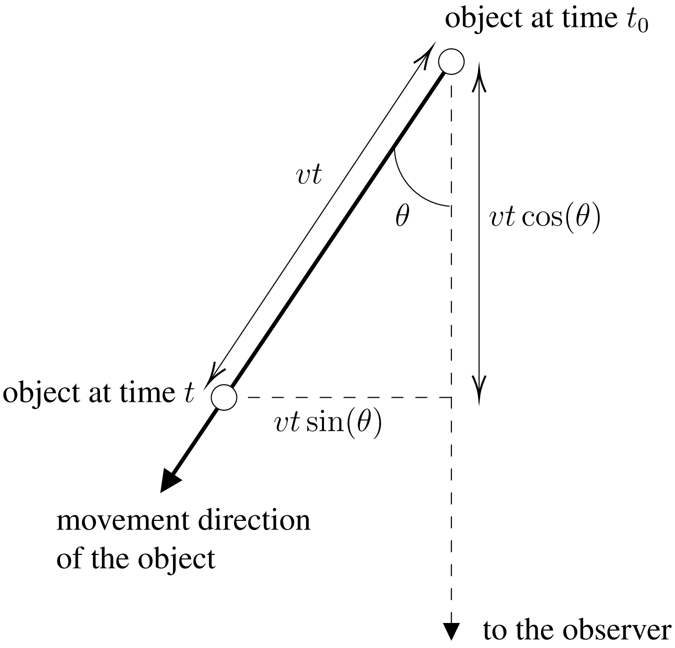


Figure 12.145: **Schematic idea behind the apparent speed**

The object emits light at time t_0, it does not instantly reach us but must travel a distance to get to us. We get it after the time:

$$t_1 = \frac{d}{c} \tag{12.1.949}$$

The object itself, moves with velocity v at an angle θ with the viewing direction, so at time t, the object moved of a distance vt. The light emitted by the object at time t must travel the distance (application of Pythagoras theorem):

$$d' = \sqrt{(d - vt\cos(\theta))^2 + (vt\sin(\theta))^2} \tag{12.1.950}$$

to reach us (the object has moved of a distance $vt\cos(\theta)$ in the direction of observation but moved away from the axis of observation of the distance $vt\sin(\theta)$), so we receive light that was emitted by the object at time t after a time t_2:

$$t_2 = t + \frac{d'}{c} = t + \frac{\sqrt{(d - vt\cos(\theta))^2 + (vt\sin(\theta))^2}}{c} \tag{12.1.951}$$

Between the two positions of the object, it has elapsed the time t but, viewed from the observer, the time interval between the reception of images of these two positions is:

$$t_2 - t_1 = t + \frac{d'}{c} = t + \frac{\sqrt{(d - vt\cos(\theta))^2 + (vt\sin(\theta))^2}}{c} - \frac{d}{c} \tag{12.1.952}$$

different from t!

For a small time interval t, we have, by doing a limited Taylor development:

$$\begin{aligned}t_2-t_1&=t+\frac{\sqrt{d^2-2dvt\cos(\theta)+v^2t^2\cos(\theta)+v^2t^2\cos^2(\theta)+v^2t^2\sin^2(\theta)}}{c}-\frac{d}{c}\\&=t+\sqrt{\frac{d^2}{c^2}-\frac{2dvt\cos(\theta)}{c^2}-\frac{v^2t^2}{c^2}}-\frac{d}{c}\cong t+\sqrt{\frac{d^2}{c^2}-\frac{2dvt\cos(\theta)}{c^2}}-\frac{d}{c}\\&=t+\frac{d}{c}\sqrt{1+\left(-\frac{2vtc\cos(\theta)}{d}\right)}-\frac{d}{c}\cong t+\frac{d}{c}\left(1-\frac{1}{2}\frac{2vt\cos(\theta)}{d}\right)-\frac{d}{c}\\&=t-\frac{vt\cos(\theta)}{c}=t\left(1-\frac{v}{c}\cos(\theta)\right)\end{aligned}\tag{12.1.953}$$

During this time interval, always from the point of view of the observer, the object appears to have moved on the sky plane by a distance of $vt\sin(\theta)$.

Thus, the apparent speed of the object is:

$$\boxed{v_{\text{app.}}=\frac{vt\sin(\theta)}{t_2-t_1}=\frac{v\sin(\theta)}{1-\dfrac{v}{c}\cos(\theta)}}\tag{12.1.954}$$

If we set the angle θ as being very close to a right angle, then we have the second term of the denominator that is very small which allows us with a Taylor expansion to write a relation that can be found quit often in high-school textbooks :

$$v_{\text{app}}\cong v\left(1+\frac{v}{c}\right)\tag{12.1.955}$$

Let us seek the maximum of this function to understand how such observation is possible by deriving relatively to θ and by seeking for what value the derivative is zero:

$$\frac{\mathrm{d}v_{\text{app}}}{\mathrm{d}\theta}=\frac{v\cos(\theta)}{1-\dfrac{v}{c}\cos(\theta)}-\frac{v}{c}\frac{v\sin^2(\theta)}{\left(1-\dfrac{v}{c}\cos(\theta)\right)^2}=0\tag{12.1.956}$$

and this vanishes after simplification of the denominator for:

$$v\cos(\theta)\left(1-\frac{v}{c}\cos(\theta)\right)-\frac{v}{c}v\sin^2(\theta)=v\cos(\theta)-\frac{v^2}{c}\left(\cos^2(\theta)+\sin^2(\theta)\right)=v\cos(\theta)-\frac{v^2}{c}=0\tag{12.1.957}$$

Hence:

$$\cos(\theta)=\frac{v}{c}\tag{12.1.958}$$

The apparent velocity is then:

$$v_{\text{app}}=\frac{v\sin(\theta)}{1-\dfrac{v^2}{c^2}}=\frac{v\sqrt{1-\cos^2(\theta)}}{1-\dfrac{v^2}{c^2}}=\frac{v\sqrt{1-\dfrac{v^2}{c^2}}}{1-\dfrac{v^2}{c^2}}=\frac{v}{\sqrt{1-\dfrac{v^2}{c^2}}}\tag{12.1.959}$$

and is equal to or greater than c if:

$$c=\frac{v}{\sqrt{1-\dfrac{v^2}{c^2}}}\Rightarrow c^2\left(1-\frac{v^2}{c^2}\right)=v^2\Rightarrow c^2-v^2=v^2\Rightarrow 2v^2=c^2 \tag{12.1.960}$$

Therefore:

$$v\geq\frac{1}{\sqrt{2}}c \tag{12.1.961}$$

Thus we see that it is possible to observe apparent movements faster than light, even though the subject is very fast indeed, but slower than c. As it is only an "illusion", there is no contradiction with the theory of relativity.

Knowing the speed of movement of a celestial object obtained using the Doppler effect and the apparent speed with the observations, it is easy for astrophysicists to determine the angle θ by doing a little bit elementary algebra from the following relation:

$$v_{\text{app}}=\frac{vt\sin(\theta)}{t_2-t_1}=\frac{v\sin(\theta)}{1-\dfrac{v}{c}\cos(\theta)} \tag{12.1.962}$$

~80 %

★★★☆☆

47 votes, 64.68%

12.2. Astrophysics

ASTROPHYSICS is an interdisciplinary branch of astronomy which mainly concerns physics and the study of the properties of objects in the Universe (stars, planets, galaxies, interstellar medium for examples) as their luminosity, density, temperature and their chemical composition. The first scientific approaches in this area date from the early 119th century (holocene calendar).

> ❞ *We have uncovered wonders undreamt by our ancestors who first speculated on the nature of those wondering lights in the night sky. We've crossed the Solar System and sent ships to the stars. But we continue to search. We can't help it! A central element of the human future, lies far beyond the Earth. If we crave some cosmic purpose, then let us find ourselves a worthy goal...* ❞
>
> Carl Sagan

quote 12.2:5539.#69

12.2.1 Stars

Before addressing the mathematical formalism on the dynamics of stars, we wanted following readers requests, write a small popularized introduction to complete the general knowledge on this field.

The stars are gaseous celestial body whose mass goes from 0.05 solar mass ($\sim 2\cdot 10^{30}$ [kg]) to more than 300 solar masses[21]. The brightness of a star (its power radiation) ranges from 10^{-6} to 10^{6} times that of the Sun. Roughly, when the mass doubles, brightness is multiplied by. Most of the stars visible to the naked eye in our skies are blue giants of 10^4-10^5 times more luminous than the Sun; they represent only 10% of stars that inhabit our Galaxy, the remaining 90% being less luminous than the Sun.

The Astronomers (of Harvard between 11918-11928 according to holocene calendar) have developed a method of classification of stars based on their position in the spectrum, of the spectral absorption lines (spectroscopy). Formerly classified from A to Q, the evolution of the spectrometry allowed their grouping and organization. Classes are now defined by the letters OBAFGKM (one mnemonic for remembering the sequence is: **O**h **B**e **A** **F**ine **G**irl/**G**uy, **K**iss **M**e), and each is divided into 10 subclasses, rated from 0 to 9. The spectral classification (taken from a continuous spectrum which summarizes only certain lines of the spectrum after passing of light in a given medium) can be crossed with the lighting classes so that we can infer the temperature at the surface of the star (we will prove later how to get this information):

---

[21] UY Scuti has the largest radius know at the day we write these lines but large radius does not make it the most massive, or heaviest star. That honour goes to R136a1, which weights about 226 times the mass of the Sun but only 30 solar radii.

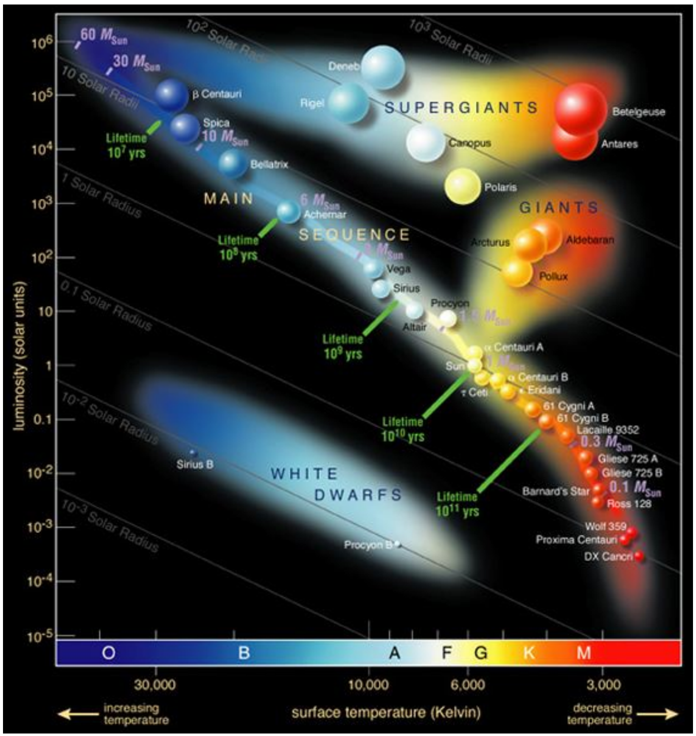


Figure 12.146: **Hertzsprung-Russel Diagram example (source: Wikipedia)**

And the corresponding hypothesized path evolution of our Sun (actually a G2V-type star, aka yellow dwarf) on this same diagram (notice on the sequence below that the external layer of our Sun will cool down a little bit on the path of its sequence while growing in size!):

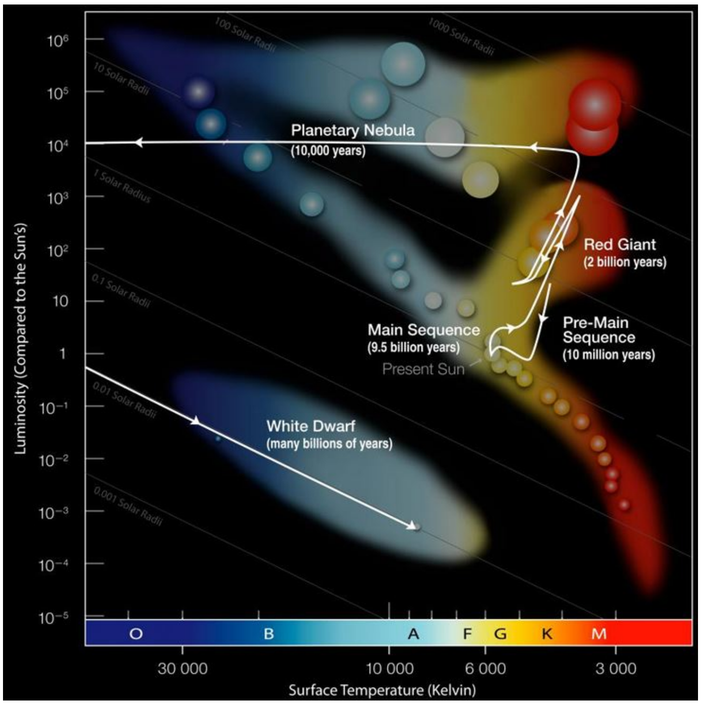


Figure 12.147: **Sun path on Hertzsprung-Russel**

Where the colors and temperature association are based "obviously" on the color of a black-body as experimental measurements show that stars behaves almost the same as a perfect black-body[22] (see section Statistical Mechanics page 3704).

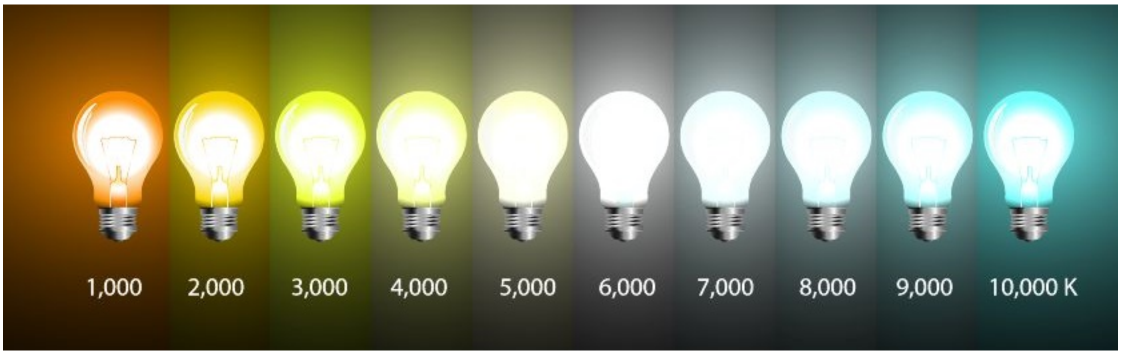


Figure 12.148: Peak color of a black body at a given temperature

As it evolves, each star describes a particular curve on the HR diagram: it begins by following the "Hayashi-path" (proto-star and after one of the existing tar) until it reaches the main sequence in which it operates as its core burns hydrogen. When beginning the burning of helium, it goes back up where red giants are concentrated and remains there until nuclear fusion stops; it then collapses on itself to join the White Dwarfs or in the case of a certain value of solar masses, neutron stars, Black Holes, or if its mass is very high, exploding as supernovae.

The O stars were discovered in the late 119th century (holocene calendar). They are hot and their spectra look like those of nebulae. The B are helium stars, A hydrogen stars. The predominant component of F is calcium. G are of the same type as the Sun (actually a G2V-type star, aka yellow dwarf) and K differ very little bit. M are characterized by titanium oxide and S of zirconium oxide, while R and N contain hydrocarbons and cyanogen.

---

[22] This is also why some LED brands have the color given as a temperature

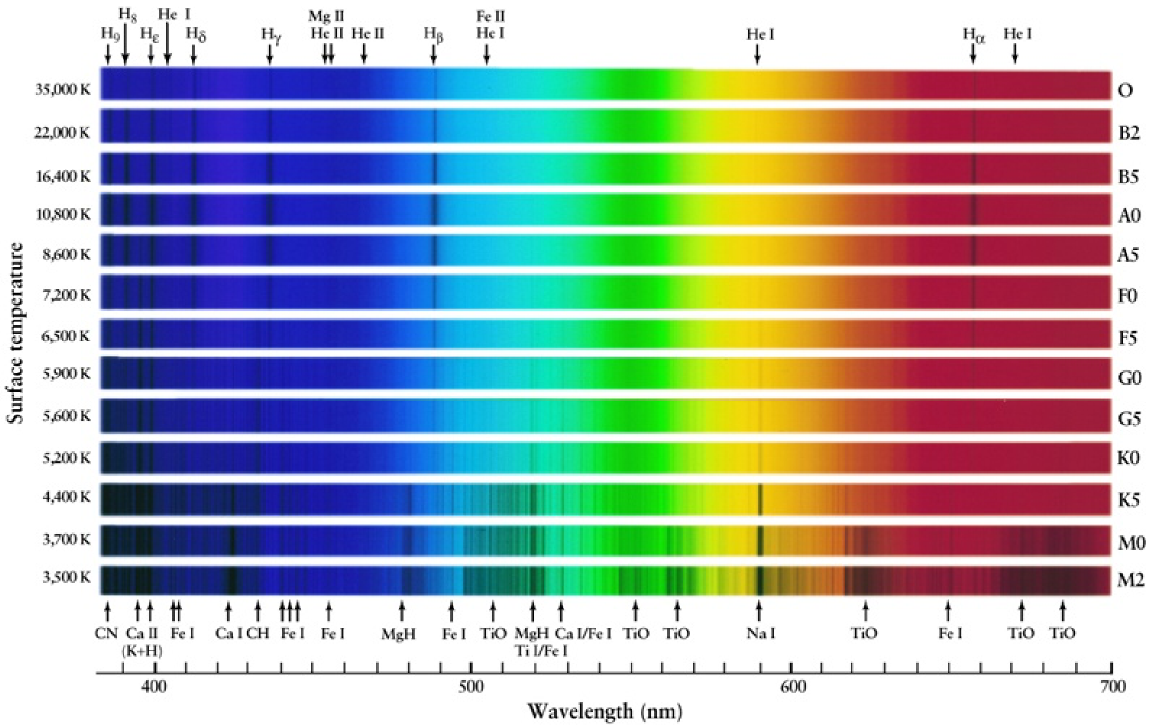


Figure 12.149: Typical stars temperature and related spectrum with absorption lines (source: Alberta University

Remark

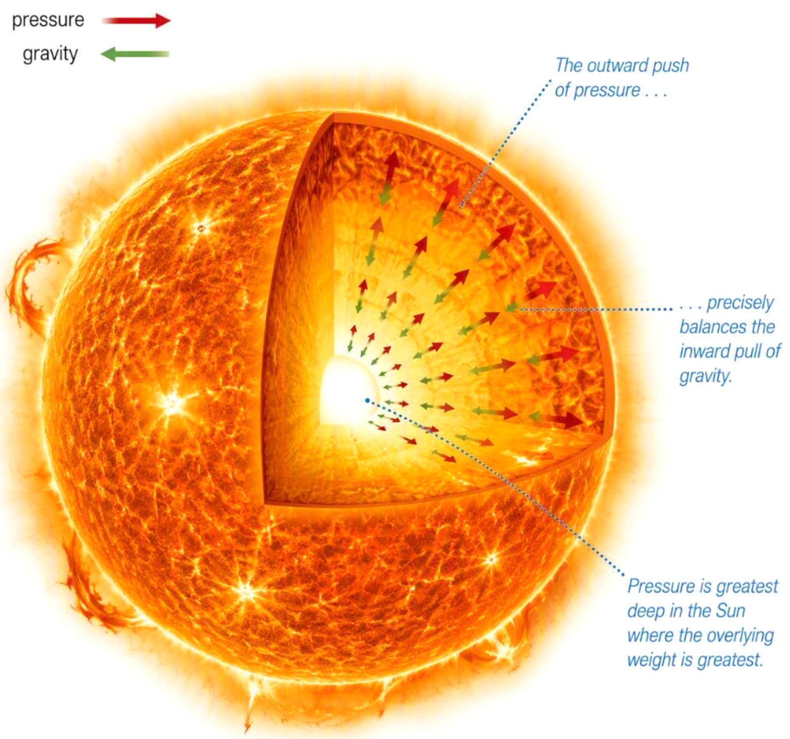
An "absorption line" is any dark line in a spectrum produced when atoms, ions, or molecules absorb light at specific wavelengths, regardless of where this occurs, such as in stars, planets, interstellar clouds, or laboratories. A "Fraunhofer line" is a specific type of absorption line observed in the Sun's spectrum, first cataloged by Joseph von Fraunhofer and caused by absorption in the cooler layers of the solar atmosphere. Thus, all Fraunhofer lines are absorption lines, but absorption lines found in other astronomical objects are not generally called Fraunhofer lines.

Therefore a star of the mass of the Sun after a stint on the main sequence, becomes a red giant, eventually a planetary nebula (ejection of fuel of the star at long distances), before ending his life as a White Dwarf. The end as supernovae cannot be shown in this diagram because of their luminosity that is to high. Neutron star and Black Holes are in the same path than White Dwarfs (lower on the right than Procyon B).

A star is initially in hydrostatic equilibrium. Gravitational forces[23] due to its mass are compensated by the internal pressure forces due to the elevated temperature maintained by thermonuclear reactions at low density and to the degeneracy pressure of electrons:

---

[23] Remember that we will see during our study of General Relativity that the gravitational force is fictional and we should us the expression "gravitational interaction" instead.


pressure
gravity
The outward push of pressure . . .
. . . precisely balances the inward pull of gravity.
Pressure is greatest deep in the Sun where the overlying weight is greatest.

The temperature at the very center of the Sun (actually a G2V-type star, aka yellow dwarf) is estimated to be about $15,000,000$ [K] and the density at the center of the hydrogen/helium supercrushed "plasma" core is estimated to be about $150\,[\text{g}\cdot\text{cm}^{-3}]$ (approximately 10 times the density of gold). The solar interior fluid is 100 to $10,000$ times more viscous than water. Fluids of comparable viscosity would be honey or ketchup (non-newtonian fluid).

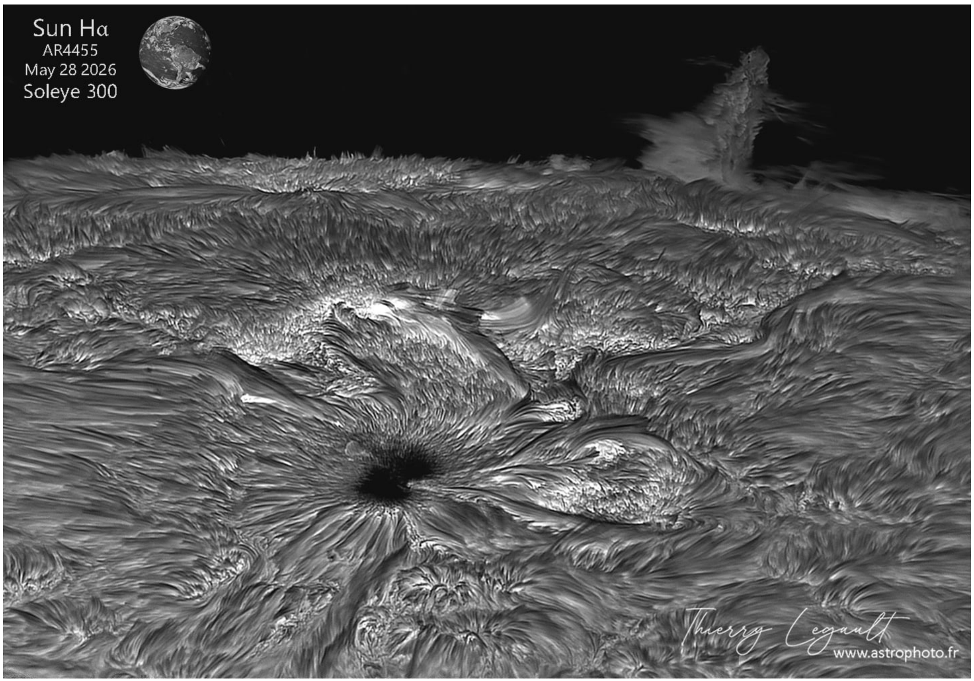
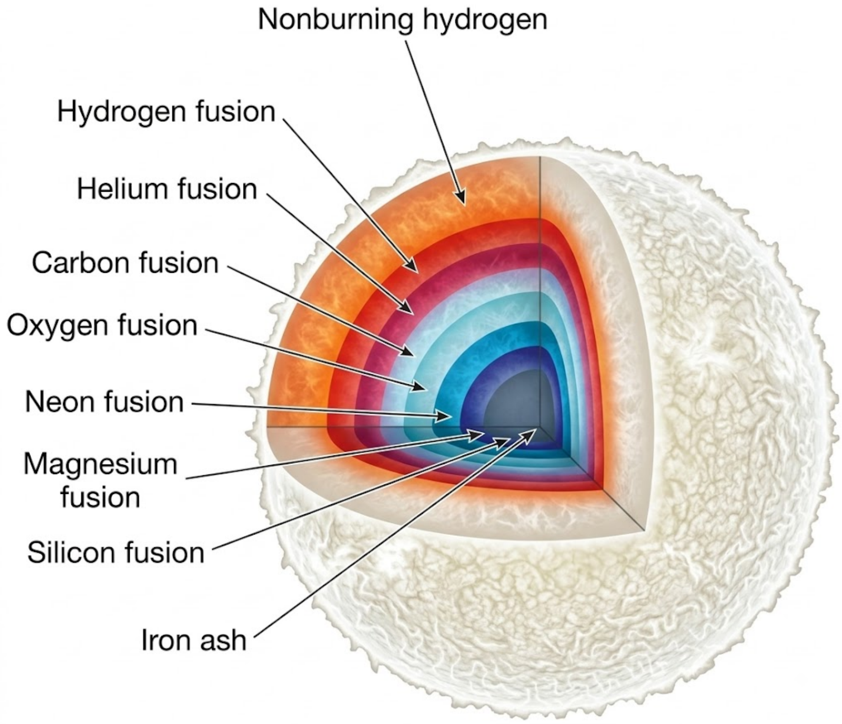

Sun Hα
AR4455
May 28 2026
Soleye 300
Thierry Legault
www.astrophoto.fr

Figure 12.150: **Photo of Sun's surface (sunspot AR4455) with a Soleye 3000 and $\text{H}\alpha$ filter (credit: Thierry Legault)**

A star spends almost 90% of his life to fuse hydrogen into helium that builds up in the center (for the first famous scientific study and observation of Sun's composition *see* stoney1869physical. During this phase, it evolves into the "main sequence" of the Hertzsprung-Russian diagram.

For a low mass Main Sequence star, hydrogen fusion is the first energy source that provides radiation pressure to maintain the hydrostatic equilibrium. When hydrogen fusion ends, the star begins to undergo structural changes, and it begins to become a red giant (helium fusion) through what we name a "helium flash". At the end of the life of a low mass star, the core collapses until the electrons provide a source of pressure to withstand the collapse, and at this stage the star is a White Dwarf. For higher mass stars, the early stages of life are the same, but the core of these stars reach higher temperatures, so they can burn more massive species, like Carbon, Oxygen, Neon, Magnesium, and Silicon. Towards the end of its life, a high mass star's core will look like the layers of an onion.



The core of a high mass star will eventually generate iron, but when the core tries to fuse iron, it will die in a catastrophic explosion. The problem is that unlike Hydrogen, Helium, Carbon, etc., the fusion of iron does not release energy (see section Nuclear Physics page 5064). When the core contains enough iron, the star implodes in seconds, and all of the mass of the outer part of the star hits the core and rebounds, and the rebound sends a shockwave outward pushing all of the material outside of the core into space in a tremendous explosion, named a "supernova[24]" (the year in the legend below is given according to holocene calendar):

---

[24] Between 12000 and 12022 (holocene calendars) more than $110,000$ supernova were observed by various observatories amonge these approximately $14,000$ were Type I supernovae and the remaining were Type II suernovae.

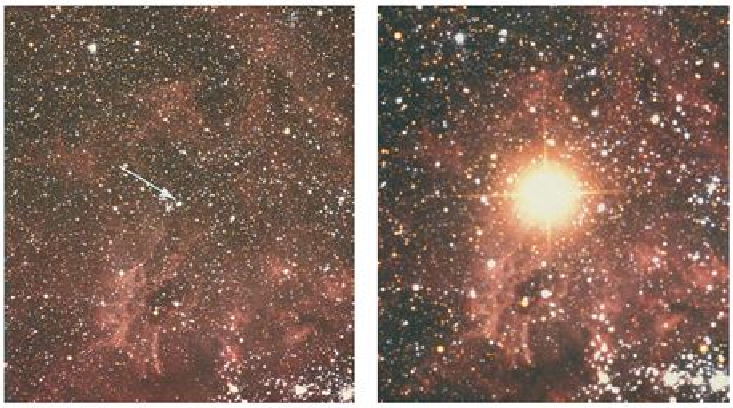


Figure 12.151: Region of the sky before and after the 11987 supernova in visible light

When the helium mass of a star becomes sufficient, the increase in pressure causes an increase of the temperature thereby initiating the fusion of helium ("helium flash") into carbon, oxygen and neon creating a second combustion front inside the first. For a star of one solar mass, reactions stop at this stage. The star radius increase and its surface temperature decrease until stabilization. It becomes a red giant 10^4 times more luminous than before.

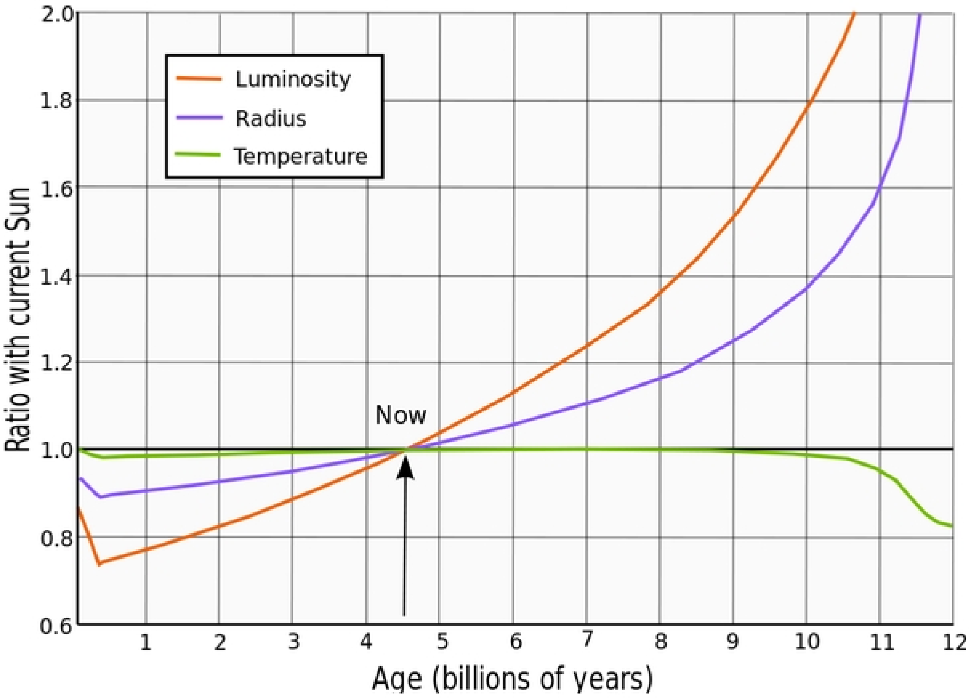


Figure 12.152: Evolution of temperature, radius, luminosity of our Sun as function of time

It goes through various phases of instability and eventually gradually expel its outer layers, forming a "planetary nebula" (from a fraction of parsecs like in size like our Solar System to a little bit more of 1 parsec - approximately 4 light years - for the biggest known at this date).

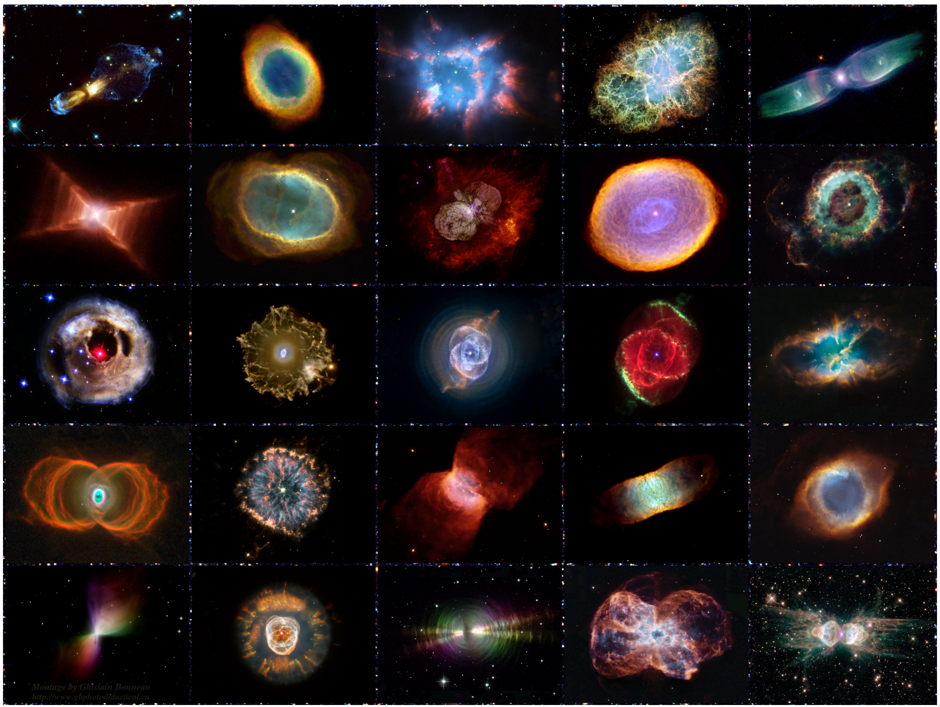


Figure 12.153: **Planetary Nebula gallery (source: Hubble Space Telescope)**

Its core, with a density of several tons per cubic centimetre, cools down slowly: it becomes a "White Dwarf" (we will discuss this process mathematically below). The balance in its core is maintained by the pressure of degeneration of electrons.

For a more massive star, the internal temperature becomes quite important so that the carbon and oxygen can fusion into silicon. In turn, if there is enough mass, silicon will fusion into iron. Combustion fronts develop in a pattern said of "onion skins" (see prior previous figure). As iron is the most stable nucleotide, it is at the bottom of the valley of stability (see section Nuclear Physics page 5062). It can not fusion or split. When the density reaches a critical value (this corresponds to a total mass of the star of more than 8 solar masses!!!), electron degeneracy pressure can no longer maintain the balance against gravity. In a tenth of a second, the iron core collapses. The other layers of the heart of the star rush towards the collapsed core in the for a wave whose maximum speed corresponds to the sonic radius.

The core density then becomes really huge. There occur inverse β^- reactions where protons capture electrons forming neutrons (!!!) and releasing a flow of neutrinos. When the core of the star reaches the nuclear density of approximately $10^{18}\;[\text{kg}\cdot\text{m}^{-3}]$, the compaction stops roughly (the remaining radius at this stage is about 10 [km] only!). The outer layers of the core bounce by a super elastic shock and come into expansion. When this reflected shock wave reached the sonic radius, the temperature rises so high that give him a value is almost meaningless. The material undergoes a complete photo-disintegration (all nucleotides are disaggregated into nucleons gas). Finally by an unclear mechanism, all the outer layers of the star are ejected into space: it is a "type II supernovae".

The collapsed core, made almost entirely of neutrons, will be rotating rapidly if the original star had a non-zero angular momentum (conservation of angular momentum oblige!). The magnetic field is also preserved and far exceeds anything that will probably never be feasible a laboratory. This causes a synchrotron beam which gives the illusion that the star flashes. This is why these young "neutron stars" are named "pulsars".

For very massive stars (above 50 solar masses), the total mass of the core that collapses could exceed 3 solar masses. In this case, gravity becomes such that its mass collapses beyond the last repulsive forces and compacted into a singularity. The curvature of space becomes such that almost (without going into the details of some theories that have until now not been verified) no material information or radiation can escape beyond the horizon or a volume named the "Schwarzschild sphere". This is a "Black Hole". Anything that falls inside loses his identity. A Black Hole has only three properties: its mass, angular momentum and electric charge. We say that a Black Hole has "no hair". Moreover, such a singularity should always be hidden by a horizon, be: "dressed" (for more details see the section of General Relativity).

Remark

Depending or your human visual accuracy (myopia) and depending on the conditions (ambient light, clouds, weather, presence of the Moon) and your position on Earth you should be able to see only between 0 to $11,000$ stars in the night sky (when modern professional telescopes car resolve hundred of billions of stars).

To give an idea of the scales you can see the figure below:

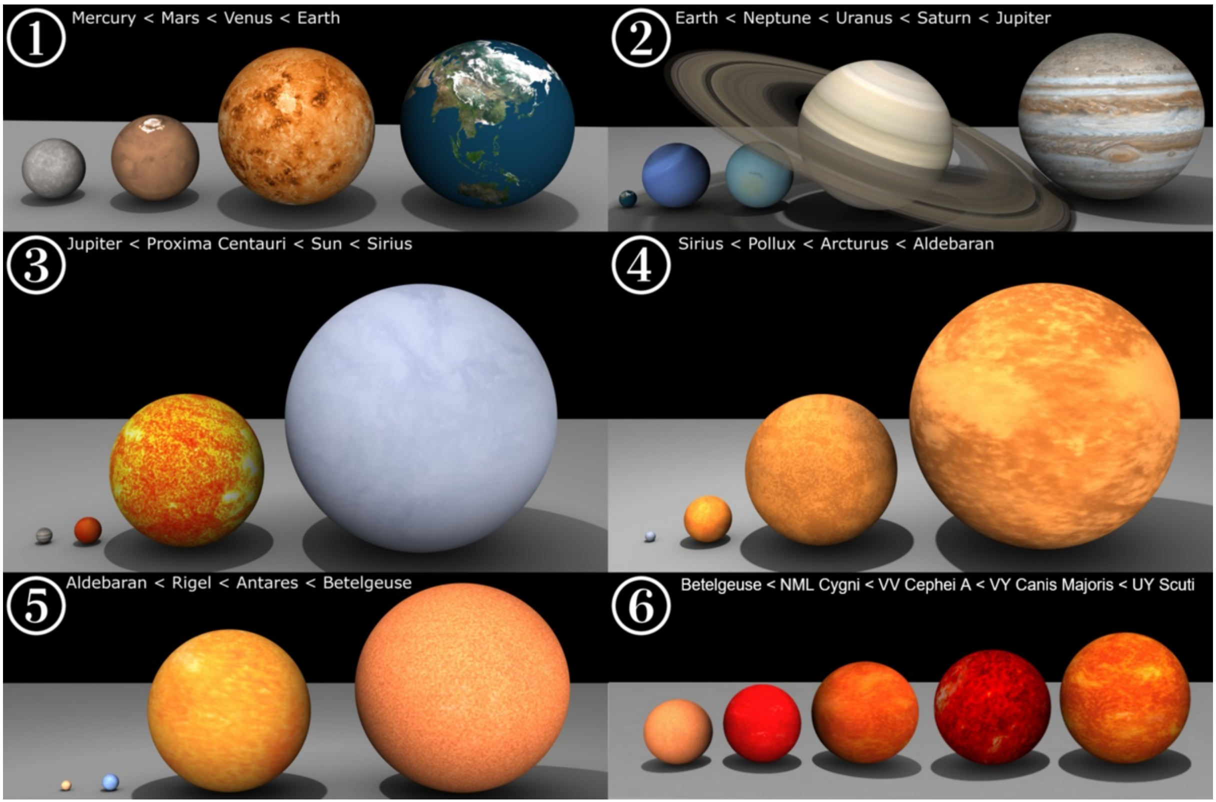

①
Mercury < Mars < Venus < Earth
②
Earth < Neptune < Uranus < Saturn < Jupiter
③
Jupiter < Proxima Centauri < Sun < Sirius
④
Sirius < Pollux < Arcturus < Aldebaran
⑤
Aldebaran < Rigel < Antares < Betelgeuse
⑥
Betelgeuse < NML Cygni < VV Cephei A < VY Canis Majoris < UY Scuti

Figure 12.154: **Comparison of various planets with various stars**

And here just for the Solar System but without respecting distance but only the proportions of the planets and Sun (high definition image so you can zoom on):

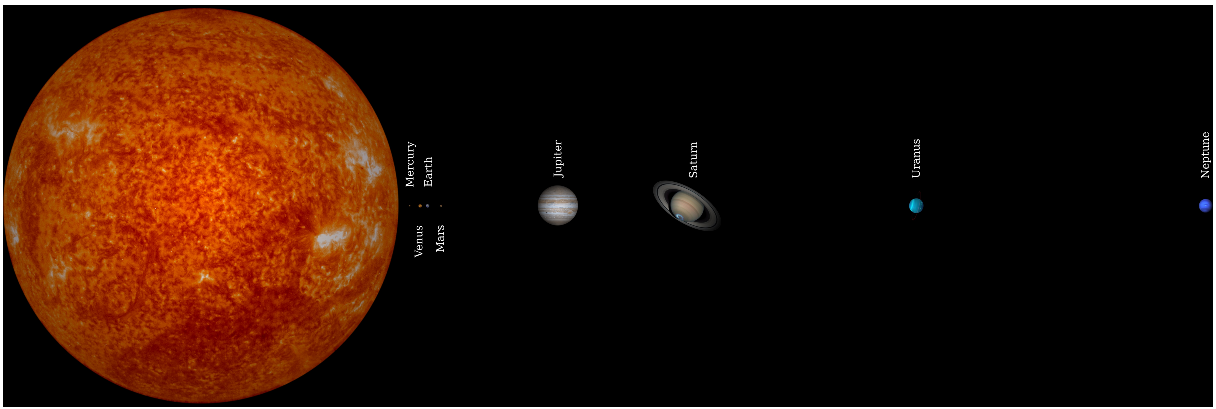

Mercury
Venus
Earth
Mars
Jupiter
Saturn
Uranus
Neptune

Figure 12.155: Solar System Proportions

Or the biggest actually known (year 12017 according to holocene) star compared the biggest known Black Hole (see section General Relativity page 5835):

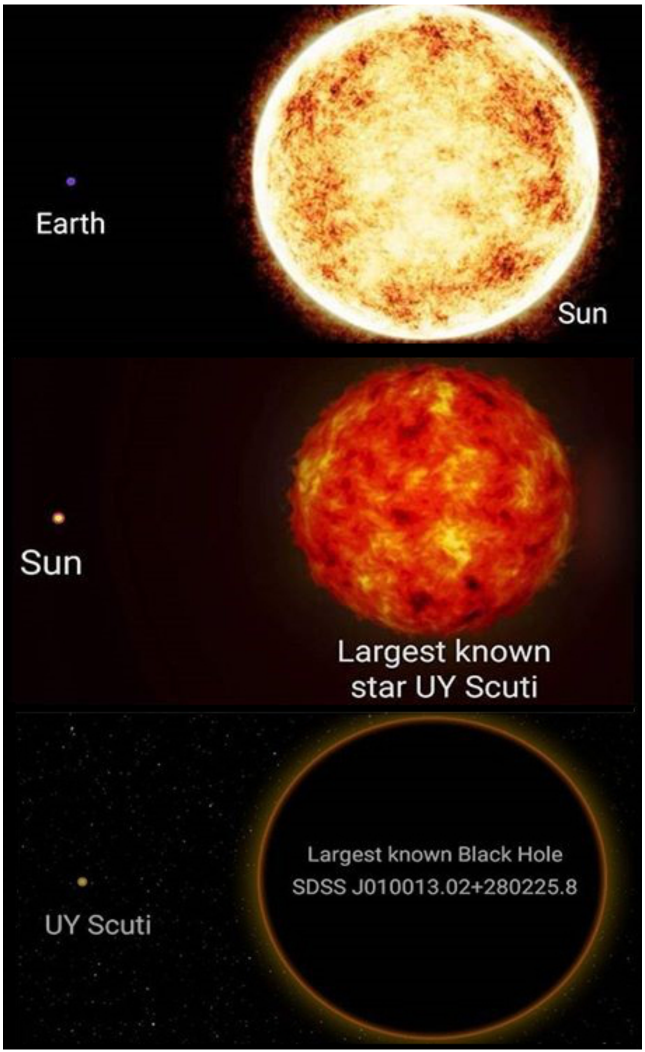

Earth
Sun
Sun
Largest known star UY Scuti
UY Scuti
Largest known Black Hole SDSS J010013.02+280225.8

Figure 12.156: Biggest Black Hole vs Biggest known star

12.2.1.1 Stellar Physics

We will now see how new stars can be born from huge gas clouds that extend between the stars in galaxies. The interstellar medium is a potential source of new stars, which once completed their life (as a red giant or supernova) can inject some of their material in outer space.

In fact, nobody really knows in this beginning of the 121st century (holocene calendar) the details of how an interstellar cloud leads to a star because it is a very difficult problem, mainly because of the emergence of a hierarchy of structures, sub-structures, etc. in the cloud as it collapses on itself. Turbulent motions appear, which can not be described simply by the hydrodynamic equations (see section Continuum Mechanics page 3811). Further complications arise when we consider the magnetic field on the gas contraction, or supernova explosions in the cloud...

At least, can we give the necessary conditions for a star to form in an interstellar cloud. For this, several barriers must actually be completed. A first thermal barrier. A second rotational barrier is: a protostar that contracts rotates faster and faster and can literally explode if its speed becomes too high (conservation of angular momentum). Let's examine these two effects.

12.2.1.1.1 Collapse of an Interstellar Cloud

Two opposing forces are present in a cloud of mass M and radius R: an autogravitation (a force) which tends to contract the cloud, and thermal pressure force, which tends to explode it.

We can quantify these two opposite forces in terms of energy: the cloud has a gravitational potential energy (negative) and a kinetic energy (positive) due to thermal agitation of the molecules.

We know (see section Classical Mechanics page 3351) that the gravitational potential energy of two masses m and m' of particles separated from a distance r is written:

$$E_p = -\frac{Gmm'}{r} \tag{12.2.1}$$

So the external potential energy of a spherical cloud (...) of mass M and radius R is of the order of:

$$E_p \cong -\frac{GM^2}{R} \tag{12.2.2}$$

Remark

Some practitioners (this is our case) prefer for the following developments use the internal potential energy that is given for recall by (see the proof in the section of Classical Mechanics):

$$E_p \cong -\frac{3}{5}\frac{GM^2}{R} \tag{12.2.3}$$

In a gas in thermodynamic equilibrium, a particle has a kinetic energy (see section Continuum Mechanics page 3897) of $kT/2$ by degree of freedom (translation, rotation, etc.). So if μ is the average mass of a molecule of the cloud, the total kinetic energy of the latter will be expressed:

$$E_c=\frac{3}{2}kT\frac{M}{\mu}\cong kT\frac{M}{\mu} \tag{12.2.4}$$

The cloud then collapses if its total mechanical energy is negative, or (according to the previous approximation):

$$kT\frac{M}{\mu}\leq\frac{GM^2}{R} \tag{12.2.5}$$

The above equation defines the "Jean's mass" (assuming a spherical and homogeneous distribution). This is the minimum mass (limit) at a given temperature T and density ρ for a cloud begins to collapse until another physical process may intervene to stop the contraction of the gas.

By eliminating the radius with:

$$M=\frac{4\pi\rho r^3}{3} \tag{12.2.6}$$

In the previous relation, we get:

$$M\geq\frac{kT}{G\mu}R=\frac{kT}{G\mu}\left(\frac{4\pi\rho r^3}{3}\right)^{1/3}\Rightarrow M^{2/3}\geq\frac{kT}{G\mu}\left(\frac{3}{4\pi\rho}\right)^{1/3}\Rightarrow M\geq\left(\frac{kT}{G\mu}\right)^{3/2}\left(\frac{3}{4\pi\rho}\right)^{1/2} \tag{12.2.7}$$

If we would not have make the choice of the external potential, but rather the internal one (more accurate in our point of view) and we did not approximate $2/3\cong 1$ the final result would have been:

$$M\geq\left(\frac{5kT}{G\mu}\right)^{3/2}\left(\frac{3}{4\pi\rho}\right)^{1/2} \tag{12.2.8}$$

That is traditionally written as:

$$\boxed{M_J=\left(\frac{5kT}{Gm}\right)^{3/2}\left(\frac{3}{4\pi\rho}\right)^{1/2}} \tag{12.2.9}$$

Therefore if $M_{\text{cloud}}>M_J$ then the cloud collapse!

As they are many approximations, astrophysicists prefer to write this last relation as:

$$M_J\cong C\left(\frac{kT}{G\mu}\right)^{3/2}\frac{1}{\sqrt{\rho}} \tag{12.2.10}$$

where C is obviously a constant without units.

PDS 70 (very young T Tauri star in the constellation Centaurus. Located at 370 light-years from Earth, it has a mass of 0.76 solar masses and is approximately 5.4 million years old) is special as it is the only protoplanetary disk so far in this early 121st century where all astronomers agree that we have found forming planets (PDS 70b, PDS 70c) caught in the act of gravitational collapse:

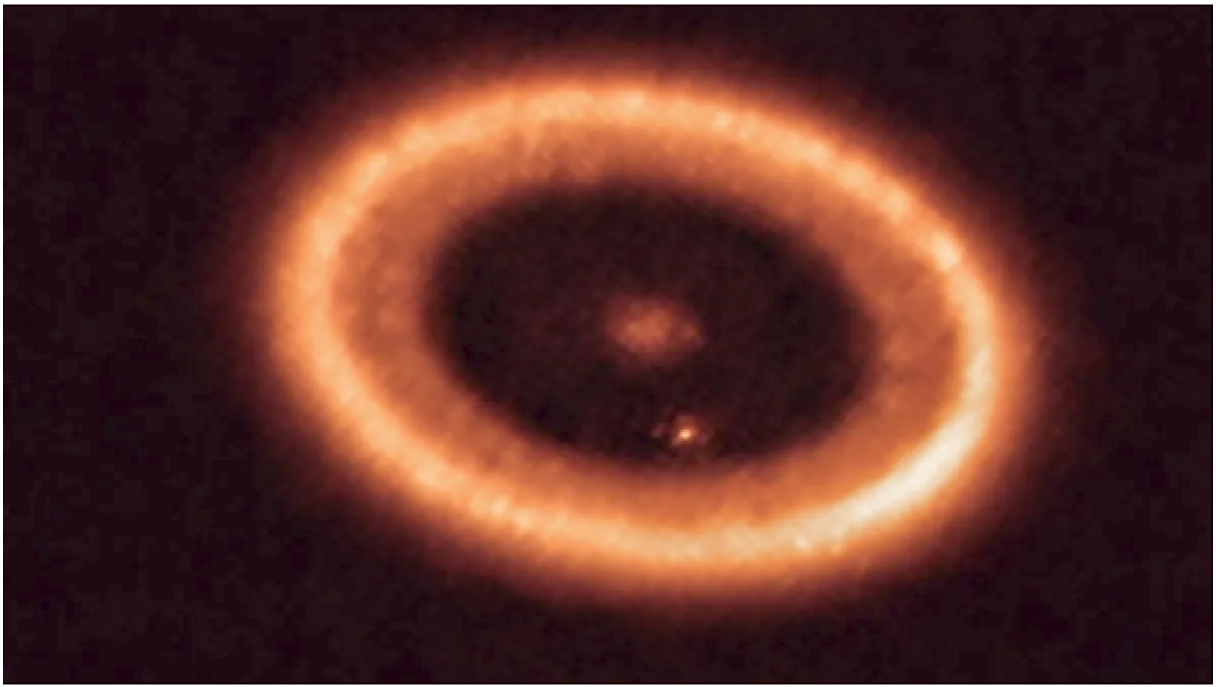


Figure 12.157: **Protoplanetary disk around the infant star PDS 70, which may contain the seeds of a third planet (source: ALMA (ESO/NAOI/NRAO)/Benisty et al.)**

12.2.1.1.1.1 Limit Mass Cloud for Ionization (rogue planets)

Now let us come back on the relation:

$$M_J=\left(\frac{5kT}{G\overline{m}}\right)^{3/2}\left(\frac{3}{4\pi\rho}\right)^{1/2} \tag{12.2.11}$$

A famous question is what is the mass required by a hydrogen cloud to start nuclear fusion and become a star[25]. For this, we can say in a first approximation that for the fusion, hydrogen atoms must have a distance equal to their radius such that we have for the density:

$$\rho=\frac{m_{\text{H}}}{\dfrac{4}{3}\pi r_0^3}=\frac{1.67\cdot 10^{-27}}{\dfrac{4}{3}\pi(0.5\cdot 10^{-11})^3}\cong 3,189\,[\text{kg}\cdot\text{m}^{-3}] \tag{12.2.12}$$

where for comparison, density for iron is $7,874\,[\text{kg}\cdot\text{m}^{-3}]$. We will also take $\overline{m}\cong m_p=1.6726219\cdot 10^{-27}$ [kg] (we neglect the mass of electrons as it is almost $1,800$ smaller) and for temperature that of the fusion of hydrogen[26] $T=T_i=10^6$ [K].

Therefore:

$$\boxed{M_i=\left(\frac{5kT_i}{G\overline{m}}\right)^{3/2}\left(\frac{3}{4\pi\rho}\right)^{1/2}\cong 1.33\cdot 10^{29}\cong 0.066M_{\odot}} \tag{12.2.13}$$

This value perfectly match the value given by then french version of Wikipedia (given without proof...).

With the 10 security factor for T_i we would have:

$$M_i\cong 0.002M_{\odot}=0.2\%M_{\odot} \tag{12.2.14}$$

In comparison Jupiter is 0.1% of the mass of the Sun...

Therefore for an initial mass less than $0.066M_{\odot}$ the gas sphere liquefies or solidifies and stabilized in the form of a planet. Jupiter as we have just see has a mass that is not for very near this limit value and we observed with telescopes that the planet is still very slowly contracting.

Such bodies are named "Sub-Brown Dwarfs", sometimes referred to as "rogue planets".

---

[25] Notice that some astrophysicists thinks that in some special configuration cases, some huge gas clouds may directly collapse in super Black Holes instead of creating a star (or smaller celestial objects).

[26] We have proved that the ionization energy of hydrogen in the section of Corpuscular Quantum Physics was 13.6 [eV], hence $157,821$ [K] but to avoid re-coupling we take a security factor of 10.

12.2.1.1.1.2 Limit Mass Cloud for Fusion (Black Dwarf)

The name "Black Dwarf" has also been applied to substellar objects that do not have sufficient mass, less than approximately $0.08M_\odot$, to maintain hydrogen-burning nuclear fusion. These objects are now generally named "brown dwarfs", a term coined in the 11970s (holocene calendar). Black dwarfs should not be confused with Black Holes or neutron stars.

Beyond ionization mass limit, it is an ionized gas ball that will continue gravitational collapse. If during contraction, the temperature of 10^7 [K] is not reached, the nuclear reactions can be triggered and it is the quantum nature of repulsive forces that will oppose gravity. Electrons are fermions (see section Statistical Mechanics page 3571), the principle of Pauli exclusion (see section Corpuscular Quantum Physics page 4672) prevents the stack of Electron in the same volume of phase space. This is equivalent to a high pressure which is well above the thermal pressure of the atoms.

The mass that can be stabilized in this state is still:

$$E_p=\frac{3}{5}\frac{GM^2}{R}=E_c=\frac{3}{2}NkT \tag{12.2.15}$$

As $T_f\cong 10^7$ [K], the electrons are non-relativistic. The linear moment is such that:

$$\frac{p^2}{2m}=\frac{3}{2}kT\Leftrightarrow p=\sqrt{2m_e\frac{3}{2}kT}=\sqrt{3m_ekT} \tag{12.2.16}$$

From the incertitude principle (see section Wave Quantum Physics page 4719):

$$[x,p_x]=\mathrm{i}\hbar \tag{12.2.17}$$

Roughly:

$$x=\frac{\hbar}{p}=\frac{h}{2\pi p}=\frac{h}{2\pi\sqrt{3m_ekT_f}}\cong 5.42912\cdot 10^{-12}\;[\text{m}] \tag{12.2.18}$$

to compare with the previous $r_0=0.5\cdot 10^{-10}$ [m]...

Therefore:

$$\rho=\frac{m_\text{H}}{\dfrac{4}{3}\pi x^3}=\frac{1.67\cdot 10^{-27}}{\dfrac{4}{3}\pi(5.42912\cdot 10^{-12})^3}\cong 2.49\cdot 10^6\;[\text{kg}\cdot\text{m}^{-3}] \tag{12.2.19}$$

For the protons, that are $1,800$ times more massive, the minimum volume is much smaller and can be neglected at this level of our discussion.

We use now the same relation as previously but where we take T_f instead of T_i and $\overline{m}=m_e=9.10938356(11)\cdot 10^{-31}$ [kg] as the proton mass as plasma experiences and intuition gives that electrons because of their small mass take the most kinetic energy (in comparison to the proton for the same charge... or as in heated liquids where small molecules are much more agitated than big one):

$$\boxed{M_f=\left(\frac{5kT_f}{G\overline{m}}\right)^{3/2}\left(\frac{3}{4\pi\rho}\right)^{1/2}\cong 1.83\cdot 10^{30}\cong 0.92M_\odot} \tag{12.2.20}$$

Therefore theoretically we have so far:

- $M_i < 0.066M_\odot$ we have a big gas planet of the style of Jupiter or Saturn ("Sub-Brown Dwarfs" or "rogue planets" for recall).
- $0.066M_\odot < M_f < 0.92M_\odot$ we have a ionized hot star at the limit of starting a nuclear fusion it's for recall a "Black Dwarf" or more suited named a "Brown Dwarf".
- For $M_J > 0.92M_\odot$ we have then a star able to initiate thermonuclear fusion.

Brown Dwarf seem to be very difficult to observe. As far as we know the first one was directly observed in 12016 (holocene calendar) only (HD 4747 B)... because of their low surface energy emitting.

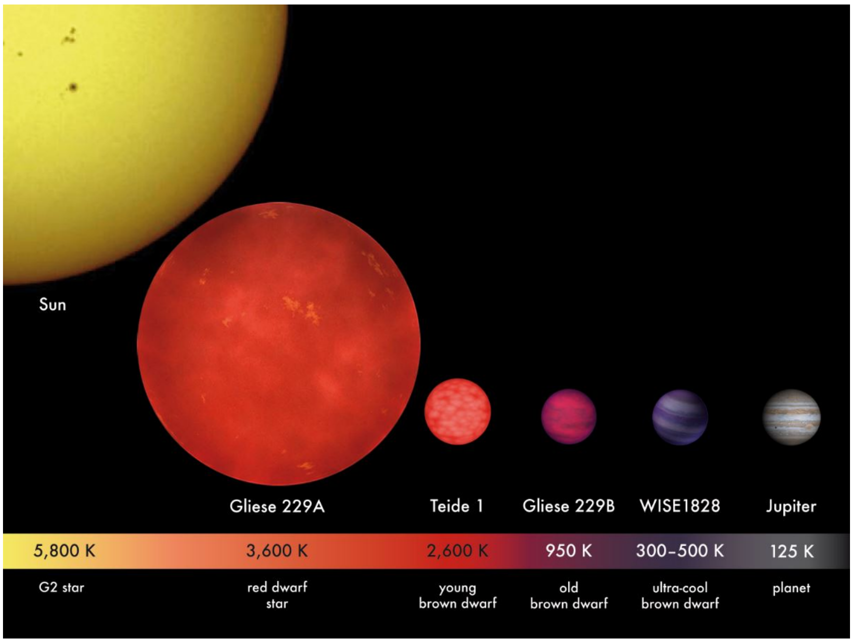


Figure 12.158: **Sun in comparison of various dwarf stars and Jupiter**

The developments above are obviously approximations and depends on many other factors. For example, Proxima Centauri, located just 4.2 light-years away has 12% of the mass of the Sun, and it is estimated to be just 14.5% the size of the Sun (actually a G2V-type star, aka yellow dwarf) with a diameter of about $200,000$ [km] (just for comparison, the diameter of Jupiter is $143,000$ [km], so Proxima Centauri is only a little larger than Jupiter).

But that's not the smallest star ever discovered! The smallest known star right now is OGLE-TR-122b that is part of a binary stellar system. This Red Dwarf has its radius accurately measured!: 0.12 solar radii. This works out to be $167,000$ [km]. That's only 20% larger than Jupiter. You might be surprised to know that OGLE-TR-122b has 100 times the mass of Jupiter.

It seems that in July 12017 (holocene calendar) even a smallest star was observed (but we need confirmation), EBLM J0555-57Ab. It seems that this star has a mass equivalent to 8% to that of our Sun and is as big as Saturn...

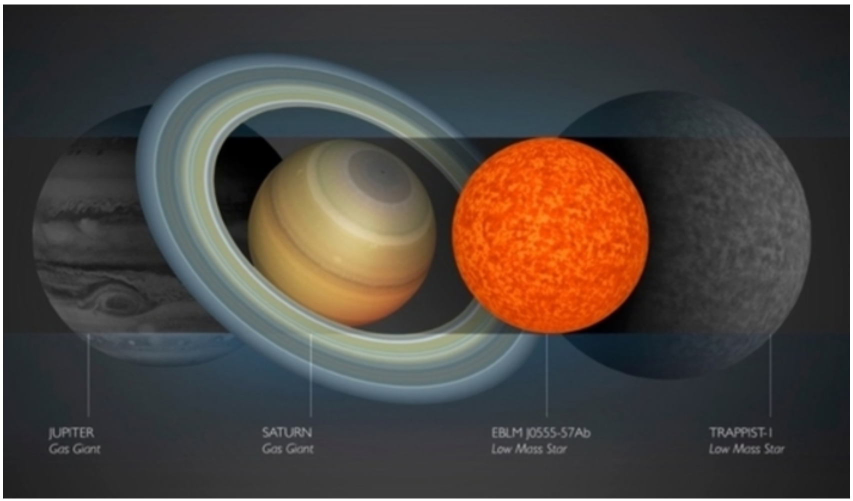


Figure 12.159: **EBLM J0555-57Ab smallest known star] (source: ?)**

12.2.1.1.1.3 Entropy of a gas collapsing under gravity

The entropy of a collapsing gas, particularly in the context of star formation, captures the fascination of internet discussions and social media debates for several reasons. This concept delves into the intersection of astrophysics, thermodynamics, and the mysteries of the universe - topics that naturally intrigue a broad audience of science enthusiasts, educators, and laypersons alike.

Entropy, often (wrongly) associated with disorder, seems paradoxical in the context of star formation. When a massive cloud of gas and dust collapses under its own gravity to form a star, the process appears to reduce disorder by condensing scattered particles into a compact, organized structure. This seemingly defies the principle that entropy should always increase in an isolated system. Social media discussions frequently center on this apparent contradiction, exploring how the second law of thermodynamics applies to star formation. People find this intellectual challenge captivating because it invites them to question and deepen their understanding of the natural world.

Rather than doing rhetorical explanations, let us do the maths as it is the tradition in this textbook (all the developments are almost a simple copy paste of a PDF provided by Rick Bradford)!

Let us consider a system of n particles of mass m in a gravitationally bound state. We shall assume thermodynamic equilibrium prevails within the gas cloud (though not necessarily between the cloud and its surroundings - see below). Thus, the gas has some temperature, T. The average kinetic energy per particle is, for non-relativistic motion in three spatial dimensions is given as we know by:

$$E_c = \frac{3}{2}kT \tag{12.2.21}$$

The virial theorem (see section of Continuum Mechanics page 3897) tells us that the average total kinetic energy:

$$\bar{E}_c = \frac{3}{2}nkT \tag{12.2.22}$$

is equal to:

$$\bar{E}_c=\frac{3}{2}nkT=-\frac{1}{2}E_p \qquad (12.2.23)$$

where E_p is the average total potential energy. Since, for a large gas cloud, the total potential energy will not vary significantly in time, we can drop the requirement for averaging. We thus have:

$$E_p=-3nkT \qquad (12.2.24)$$

But the gravitational potential energy of a spherically symmetric mass distribution of total mass $M=n\cdot m$ and radius R is:

$$E_p=-\eta\frac{GM^2}{R} \qquad (12.2.25)$$

where η is a dimensionless constant of order unity depending on the distribution (e.g. $\eta=3/5$ for uniform density as proven in the section Mechanics at page 3356). In terms of the volume of the cloud:

$$V=\frac{4}{3}\pi\mathrm{R}^3 \qquad (12.2.26)$$

Therefore:

$$kT=\frac{\eta}{3}\left(\frac{4\pi}{3}\right)^{1/3}Gm^2\cdot\frac{n}{V^{1/3}} \qquad (12.2.27)$$

Now the entropy of an ideal gas is given by the Sackur-Tetrode equation (see section Continuum Mechanics page 3934):

$$S_{\text{gas}}=Nk\left\{\frac{5}{2}+\log\left[\left(\frac{mkT}{2\pi\hbar^2}\right)^{3/2}\cdot\frac{V}{N}\right]\right\} \qquad (12.2.28)$$

This gives the absolute total entropy. For our purposes we only need know the change in entropy between two states of different volume and temperature ΔS_{gas}. Thus we can simplify the previous relation to:

$$S_{\text{gas},i}-S_{\text{gas},f}=\Delta S_{\text{gas}}=Nk\log\left\{\left(\frac{T_i}{T_f}\right)^{3/2}\left(\frac{V_i}{V_f}\right)\right\}+c^{te} \qquad (12.2.29)$$

Using:

$$kT=\frac{\eta}{3}\left(\frac{4\pi}{3}\right)^{1/3}Gm^2\cdot\frac{n}{V^{1/3}} \qquad (12.2.30)$$

we can substitute for T in terms of V, and hence the prior-previous relation can be simplified further:

$$\Delta S_{\text{gas}}=Nk\log\left\{\left(\frac{V_f^{1/3}}{V_i^{1/3}}\right)^{3/2}\left(\frac{V_i}{V_f}\right)\right\}+c^{te}=Nk\log\left(\frac{V_i^{1/2}}{V_f^{1/2}}\right)=2Nk\log\left(\frac{V_i}{V_f}\right) \qquad (12.2.31)$$

In passing we note that the total energy of the gas cloud is:

$$E_{\text{gas}}=E_c+E_p=\frac{E_p}{2}=-\frac{3}{2}\mathrm{nkT}=-\frac{\eta}{2}\cdot\frac{GM^2}{R} \qquad (12.2.32)$$

From the above we note that:

- The total energy of the gas is negative (inevitably because it is a gravitational bound system - we could not have employed the virial theorem otherwise)
- As the cloud collapses (R reduces) the magnitude of the energy increases, i.e. the absolute energy E_{gas} decreases
- As the cloud collapses (R reduces) the temperature increases
- From the two previous points above it follows that the gas cloud has a negative specific heat (the temperature increases as energy is removed from it)
- From:

  $$\Delta S_{\text{gas}} = 2nk\log\left(\frac{V_i}{V_f}\right) \tag{12.2.33}$$

  as the cloud collapses (R and hence V reduces) the entropy of the gas has decreased.

Hence a collapsing gas forming a star still has its entropy that decreases, such that:

$$\boxed{\Delta S_{\text{gas}} = S_{\text{gas},i} - S_{\text{gas},f} = 2Nk\log\left(\frac{V_i}{V_f}\right) > 0} \tag{12.2.34}$$

It is clear that the problem posed above is incomplete in the sense that the conditions imposed lead to the conclusion that the gas looses energy but how it does so and where the energy goes has not been specified. If we imagine our gas cloud to be surrounded by the vacuum of interstellar space, and given that we are assuming the cloud is gravitationally bound throughout (and hence cannot eject matter), it follows that the only means of loosing energy is via electromagnetic radiation. The second law of thermodynamics is preserved if this radiation carries off more entropy than the gas cloud looses. This is what we shall check in the following Section. Now we will derive the entropy of the radiation.

Since the gas is in internal thermodynamic equilibrium (by assumption - this being the condition for its temperature to be defined) it follows that the radiation emitted is black body radiation. In black body radiation the energy density is proportion to T^4 (see section Thermodynamics page 3713), i.e.:

$$\rho(T) = \sigma_{\text{S.B.}}T^4 \tag{12.2.35}$$

The entropy of a given volume V_r of radiation is found from the definition of the entropy itself (see section Thermodynamics page 3658):

$$\mathrm{d}S_{\text{radiation}} = \frac{\mathrm{d}Q}{T} \tag{12.2.36}$$

where $\mathrm{d}Q = V_r\mathrm{d}\rho(T)$.

Actually, the relation above is all that we shall require, but for completeness note that:

$$\mathrm{d}Q = V_r\mathrm{d}\rho(T) = V_r\cdot\mathrm{d}\left(\sigma_{\text{S.B.}}T^4\right) = 4\sigma_{\text{S.B.}}V_rT^3\mathrm{d}T \tag{12.2.37}$$

Hence:

$$S_{\text{radiation}}=\frac{\mathrm{d}Q}{T}=\int\frac{4\sigma_{\text{S.B.}}V_rT^3\mathrm{d}T}{T}=\int 4\sigma_{\text{S.B.}}V_rT^2\mathrm{d}T=\frac{4}{3}\sigma_{\text{S.B.}}V_rT^3 \tag{12.2.38}$$

Using:

$$\begin{cases}\mathrm{d}Q=V_r\mathrm{d}\rho(T) & \Rightarrow V_r=\dfrac{Q}{\rho(T)}\\ \rho(T)=\sigma_{\text{S.B.}}T^4 & \Rightarrow T^3=\dfrac{\rho(T)}{\sigma_{\text{S.B.}}T}\end{cases} \tag{12.2.39}$$

So finally:

$$S_{\text{radiation}}=\frac{4}{3}\sigma_{\text{S.B.}}V_rT^3=\frac{4}{3}\frac{Q}{T} \tag{12.2.40}$$

And this is the part of the entropy transported by the radiation! So to respect the second law of thermodynamics we have to assume that:

$$S_{\text{radiation}}\geq\Delta S_{\text{gas}} \tag{12.2.41}$$

Or explicitly:

$$\frac{4}{3}\frac{Q}{T}\geq 2nk\log\left(\frac{V_i}{V_f}\right) \tag{12.2.42}$$

It is likely also that a part of the entropy is also embedded in the deformation of space-time itself. Again we are faced here to an empirical difficulty as we don't have yet in this early 121st century (holocene calendar) the capacity to isolate a cloud of gas, observe its collapsing and measure directly its entropy variation...

12.2.1.1.2 Nuclear Duration Life

Once again remember that we have proved in the section of Classical Mechanics that inside a massive body the gravitational potential energy was given by (see page 3351):

$$E_p=-\frac{3}{5}\frac{GM^2}{R} \tag{12.2.43}$$

What has this have to do with starshine?

Well, notice that as R gets smaller, E_p gets more negative then energy is being converted to other forms, like heat. If a star can radiate this heat into space, then gravitational contraction might produce the luminosity of the star.

How much of this gravitational energy can be radiated away?

Remember that we know the radiation is related to heat (see section Statistical Mechanics page 3704) that is itself related to velocity (see section Continuum Mechanics page 3897), and that we proved during our study of the virial theorem (see section Continuum Mechanics page 3897) that:

$$E_c=-\frac{1}{2}E_p \tag{12.2.44}$$

Let us go back to our contracting Sun (actually a G2V-type star, aka yellow dwarf) problem! The virial theorem says that half the change in gravitational energy stays with the star (it heats the star through the atomic agitation). The other half is radiated away.

So, let's say that the Sun has been contracting and was originally much, much bigger. Initially its gravitational potential energy was tiny (why?), so the change in gravitational energy is:

$$\Delta E_p = E_{p,\text{init}} - E_{p,\text{now}} = 0 - \left(-\frac{3}{5}\frac{GM_\odot^2}{R_\odot}\right) = \frac{3}{5}\frac{GM_\odot^2}{R_\odot} \tag{12.2.45}$$

Now, half of this energy could have been radiated as the Sun shrank:

$$\boxed{E_{\text{rad}} = \frac{3}{10}\frac{GM_\odot^2}{R_\odot}} \tag{12.2.46}$$

that gives with our Sun the value: $\cong 10^{41}\,[\text{J}]$. That's a lot of energy! So how long could it sustain the luminosity of the Sun? The answer is obviously simply given by the ratio (it's just a dimensional analysis ratio):

$$\boxed{t_{KH} = \frac{E_{\text{rad}}}{L_\odot}} \tag{12.2.47}$$

This time is named the "Kelvin-Helmholtz timescale" and with the values of the Sun it gives:

$$t_{KH} = \frac{E_{\text{rad}}}{L_\odot} \cong \frac{10^{41}}{4\cdot 10^{26}} = 2.5\cdot 10^{14}\,[\text{s}] \cong 8\cdot 10^{6}\,[\text{years}] \tag{12.2.48}$$

and as we see the value is quite problematic... as our Earth would be older than our Sun! This problem comes from the fact that we don't take into account the fuel of stars that is the nuclear fusion. So let us see a little bit more accurate model now!

So the age of the stars is as we will see just now mainly a problem of calculation of nuclear fuel. The resolution of this problem was given by relativity, and in particular by the mass-energy equivalence (see section Special Relativity page 5679).

Even if the detailed description of nuclear reactions in the heart of the Sun was made in the mid-11930s (holocene calendar) by Hans Bethe, astrophysicists have suspected soon after Albert Einstein's work that the mass-energy equivalence could explain the brightness of the Sun on billions of years, for example through the fusion of ionized hydrogen (proton p) into ionized helium (two protons, two neutrons) via a series of steps (the specified energy is the kinetic energy of the different elements):

$${}^1_1\text{H} + {}^1_1\text{H} \rightarrow {}^2_1\text{H} + e^+ + 0.42\,[\text{MeV}] \tag{12.2.49}$$

The positron annihilates instantly with one of the electrons of a surrounding hydrogen and their mass-energy is liberated in the form of two gamma photons:

$$e^+ + e^- \rightarrow 2\gamma + 1.20\,[\text{MeV}] \tag{12.2.50}$$

After this the deuterium produced in the first stage can fuse with another hydrogen nucleus to produce an isotope of helium:

$${}^2_1\text{H} + {}^1_1\text{H} \rightarrow {}^3_2\text{He} + \gamma + 5.49\,[\text{MeV}] \tag{12.2.51}$$

Finally, two isotopes of helium ^{3}He may fuse and produce the normal isotope of helium $^{3}{}_{2}\text{He}$ and also two hydrogen nuclei that can start again the reaction in three difference ways named PPI (dominant at temperatures of 10 to 14 millions Kelvins), PPII (dominant at temperatures of 14 to 23 millions Kelvins) and PPIII (dominant above 23 millions Kelvins):

$$ {}^3_2\text{He}+{}^3_2\text{He}\rightarrow{}^4_2\text{He}+{}^1_1\text{He}+{}^1_1\text{He}+12.86\,[\text{MeV}] \tag{12.2.52} $$

And these reactions do not occur all with the same probability and at the same temperatures...

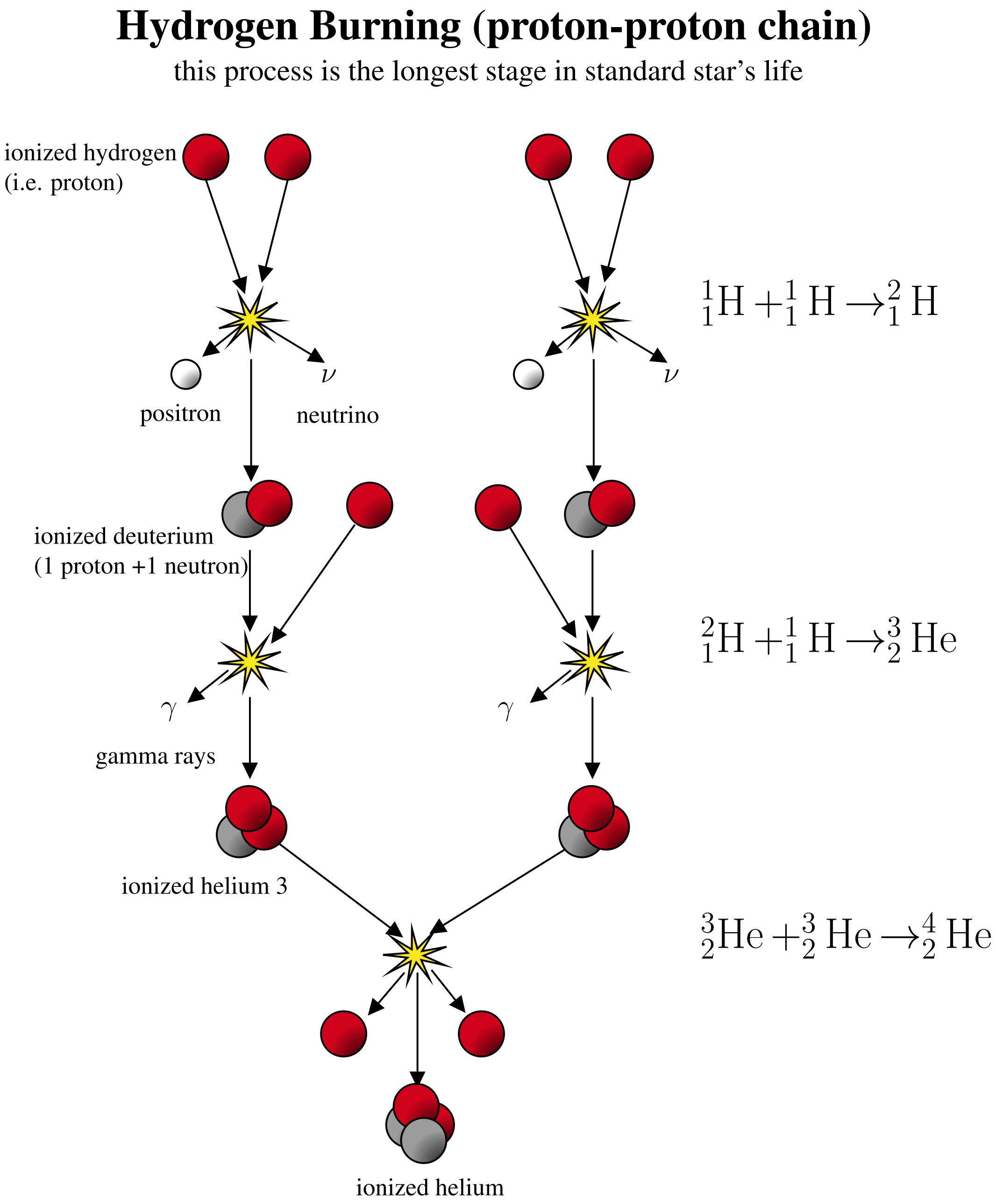

Hydrogen Burning (proton-proton chain)
this process is the longest stage in standard star's life
ionized hydrogen
(i.e. proton)
positron
neutrino
ν
ν
${}^1_1\text{H}+{}^1_1\text{H}\rightarrow{}^2_1\text{H}$
ionized deuterium
(1 proton +1 neutron)
γ
γ
gamma rays
${}^2_1\text{H}+{}^1_1\text{H}\rightarrow{}^3_2\text{He}$
ionized helium 3
${}^3_2\text{He}+{}^3_2\text{He}\rightarrow{}^4_2\text{He}$
ionized helium

Figure 12.160: **Diagram of the proton-proton chain reaction in typical Sun size stars**

The measurement of the mass of the proton gives $m_p \cong 1.673\cdot 10^{-27}$ [kg], while the helium mass is $m_{\text{He}} \cong 6.645\cdot 10^{-27}$ [kg], that is to say a loss in atomic mass (we neglect the mass of positrons which is $10,000$ times smaller than that of the neutrino):

$$ 4m_p - m_{\text{He}} \cong 0.047\cdot 10^{-27}\,[\text{kg}] \tag{12.2.53} $$

So a relative loss of mass by fusion (this is the part of the reactions that escapes from the Sun in the form of kinetic energy):

$$\frac{\Delta m}{m}=\frac{0.047\cdot 10^{-27}}{4\cdot 1.673\cdot 10^{-27}}\cong 0.0072=0.72\% \tag{12.2.54}$$

We will prove further below that the Sun emits a power output of:

$$P_{\odot}\cong 3.861\cdot 10^{26}\,[\text{W}] \tag{12.2.55}$$

Therefore its mass consumption per second is:

$$\frac{\Delta m}{t}=\frac{E}{tc^2}=\frac{P_{\odot}}{c^2}\cong 4.4\cdot 10^{9}\,[\text{kg}\cdot\text{s}^{-1}] \tag{12.2.56}$$

i.e. its mass decreases by 4.4 million tonnes per second...

Now we know that this value corresponds to only 0.72% of the mass put in reaction in a fusion. The total mass reacted is then (rule of three):

$$\frac{4.4\cdot 10^{9}}{0.0072}\cong 6.27\cdot 10^{11}\,[\text{kg}\cdot\text{s}^{-1}] \tag{12.2.57}$$

Thus, at every second 627 million tons of hydrogen 1(ionized) fuse into helium 4 with a weight loss of 4.4 million tonnes which is converted into energy.

Assuming that only the center of the Sun fills the thermal conditions for the fusion ($\cong 10\%$ of its total mass), this brings us to determine the time of nuclear life of the Sun (or any other Star of the same type whose mass is known):

$$t_{\text{nucl}}=\frac{0.1M_{\odot}}{6.27\cdot 10^{11}}\cong 3.17\cdot 10^{17}\,[\text{s}] \tag{12.2.58}$$

Transforming this in years we have:

$$t_{\text{nucl}}\cong 1.01\cdot 10^{10}\,[\text{years}] \tag{12.2.59}$$

As we know that Earth is almost 4.5 billion years that gives us an indication that the Sun has very likely already made more than half of its lifespan.

If we apply the same calculations to the next step of our Sun, we get once all Hydrogen consumed transformed into Helium (at a temperature of 10 millions degrees to counterbalance gravity as we will prove it further below) a few million years before the next step. Then almost all Helium will be transformed into Carbon (as the pressure and therefore temperature increase to almso 200 million degrees to counterbalance gravity) and this state last only a few hundred thousands years before the next step. Then almost all Carbon will be transformed into Neon and this state last only a few thousands years before the next step. Then almost all Neon will be transformed into Oxygen and this state last only a few years before the next step. Then almost all Oxygen will be transformed into Silicon and this state last only a few months before the next step. Then almost all Silicon will be transformed into Nickel (at a temperature of 2.5 billion degrees) and this state last only a few weeks before the next step. As Nickel decays into Iron (one of the most stable elements) afterwards it can happen that the stars collapse into a neutron star, or even a quark star or in some cases in a Black Hole.

12.2.1.1.3 Internal Temperature

The stars are assumed to be spherical clusters of hydrogen gas where the interactions between molecules are governed by the gravitational attraction.

A star has no wall that delimits it, that is to say that there are no external forces coming from vacuum and therefore:

$$PV = 0 \tag{12.2.60}$$

and also not any bulk modulus.

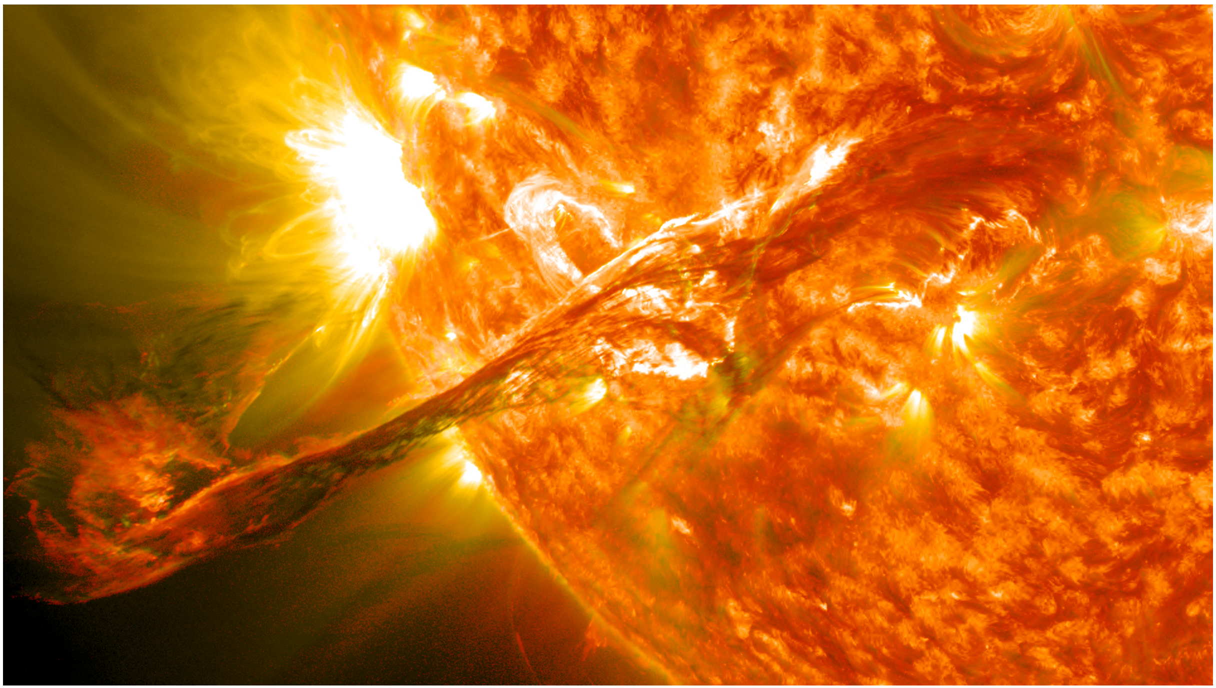


Figure 12.161: **Sun Corona zoom (source: NASA Goddard Space Flight Center)**

Remark

A solar flare is an intense localized eruption of electromagnetic radiation in the Sun's atmosphere that occur in active regions and are often, but not always, accompanied by coronal mass ejections, solar particle events, and other solar phenomena. The occurrence of solar flares varies with the actual 11-year solar cycle. Solar flares are thought to occur when stored magnetic energy in the Sun's atmosphere accelerates charged particles in the surrounding plasma. This results in the emission of electromagnetic radiation across the electromagnetic spectrum. High-energy electromagnetic radiation from solar flares is absorbed by the daylight side of Earth's upper atmosphere, in particular the ionosphere, and does not reach the surface most of the time. This absorption can temporarily increase the ionization of the ionosphere which may interfere with short-wave radio communication. The prediction of solar flares is an active area of research. Solar flares occur in a power-law spectrum of magnitudes; an energy release of typically 10^{20} joules of energy suffices to produce a clearly observable event, while a major event can emit up to 10^{25} joules (estimated value of the Carrington event that has occurred in 11859 according to holocene calendar).

The National Oceanic and Atmospheric Administration ranks geomagnetic storms on a scale running from G1, causing an increase in auroral activity and minor fluctuations in power supplies, up to G5, which includes extreme cases like the Carrington event. Even a G2 event is capable of causing damages, as when SpaceX lost 40 satellites in Feb 2022 when a G2 storm prevented them reaching their intended altitude. According to [353] speculation and worst case scenario, extreme solar flares from our Sun (actually a G2V-type star, aka yellow dwarf) to an upper limit of 10^{28} joules (with a frequency of 26 million years to occur) may even lead to mass extinction event mainly because of the subsequent ozone layer depletion.

Scale	Description	Effect	Physical measure	Average Frequency (1 cycle $=11$ years)
G 5	**Extreme**	**Power systems:** Widespread voltage control problems and protective system problems can occur, some grid systems may experience complete collapse or blackouts. Transformers may experience damage. **Spacecraft operations:** May experience extensive surface charging, problems with orientation, uplink/downlink and tracking satellites. **Other systems:** Pipeline currents can reach hundreds of amps, HF (high frequency) radio propagation may be impossible in many areas for one to two days, satellite navigation may be degraded for days, low-frequency radio navigation can be out for hours, and aurora has been seen as low as Florida and southern Texas (typically 40° geomagnetic lat.).	$Kp=9$	4 per cycle (4 days per cycle)
G 4	**Severe**	**Power systems:** Possible widespread voltage control problems and some protective systems will mistakenly trip out key assets from the grid. **Spacecraft operations:** May experience surface charging and tracking problems, corrections may be needed for orientation problems. **Other systems:** Induced pipeline currents affect preventive measures, HF radio propagation sporadic, satellite navigation degraded for hours, low-frequency radio navigation disrupted, and aurora has been seen as low as Alabama and northern California (typically 45° geomagnetic lat.).	Kp $=8$, including a 9 -	100 per cycle (60 days per cycle)
G 3	**Strong**	**Power systems:** Voltage corrections may be required, false alarms triggered on some protection devices. **Spacecraft operations:** Surface charging may occur on satellite components, drag may increase on low-Earth-orbit satellites, and corrections may be needed for orientation problems. **Other systems:** Intermittent satellite navigation and low-frequency radio navigation problems may occur, HF radio may be intermittent, and aurora has been seen as low as Illinois and Oregon (typically 50° geomagnetic lat.).	$Kp=7$	200 per cycle (130 days per cycle)
G 2	**Moderate**	**Power systems:** High-latitude power systems may experience voltage alarms, long-duration storms may cause transformer damage. **Spacecraft operations:** Corrective actions to orientation may be required by ground control; possible changes in drag affect orbit predictions. **Other systems:** HF radio propagation can fade at higher latitudes, and aurora has been seen as low as New York and Idaho (typically 55° geomagnetic lat.).	$Kp=6$	600 per cycle (360 days per cycle)
G 1	**Minor**	**Power systems:** Weak power grid fluctuations can occur. **Spacecraft operations:** Minor impact on satellite operations possible. **Other systems:** Migratory animals are affected at this and higher levels; aurora is commonly visible at high latitudes (northern Michigan and Maine).	$Kp=5$	1700 per cycle (900 days per cycle)

Table 12.15: **Solar flares Geomagnetic scale (source: NOAA)**

Based on the return on experience of the G5 Carrington event and the Canada 11989 (holocene calendar) G5 event (the "Miyake event[a]" can't be used as 1400 years ago electric devices didn't exist), it's very unlikely that a even a super solar flare may damage HDD ans SSD drive. First because HDD drives already have magnets to make them work that are much stronger than G5 solar flares and SSD drive are said to be almost immune to magnetic fields. Instead, super solar flares can generate very high electrical currents in long conductors (i.e. long power grids), like power lines. These electrical surges can damage hardware connected to the grid just like a lightning strike. So if you have a Solar flare tracking app and get a G5 alarm, then just disconnect all your electric devices from the outlet (you don't need to put them in a Faraday cage or run into a nuclear basement, that's very likely useless). Some people say that SSD's are even immune to power outage.

There is another common scale for solar flares:

Class	**Intensity** $[\mathrm{W}\cdot\mathrm{m}^{-2}]$	**Luminosity** $(\mathrm{L}_\odot, 100\ \mathrm{sec})$
B	10^{-7}	10^{-8}
C	10^{-6}	10^{-7}
M	10^{-5}	10^{-6}
X	10^{-4}	10^{-5}

Table 12.16: **Solar flares intensity scale**

An X2 is twice an X1. The 11989 (holocene calendar) event in Canada was a X7.3. The Carrington event was estimated X45 to X50.

---

[a] Miyake events are events several orders of magnitude greater than the Carrington event. Scientists think the most recent Miyake event was around 80 times more powerful than the Carrington event. We now know of 10 Miyake events over the past $15,000$ years. While they appear to be relatively infrequent, we don't know whether they occur with any sort of pattern or if they're predictable. It is also unclear yet what changes the Sun undergoes to produce them.

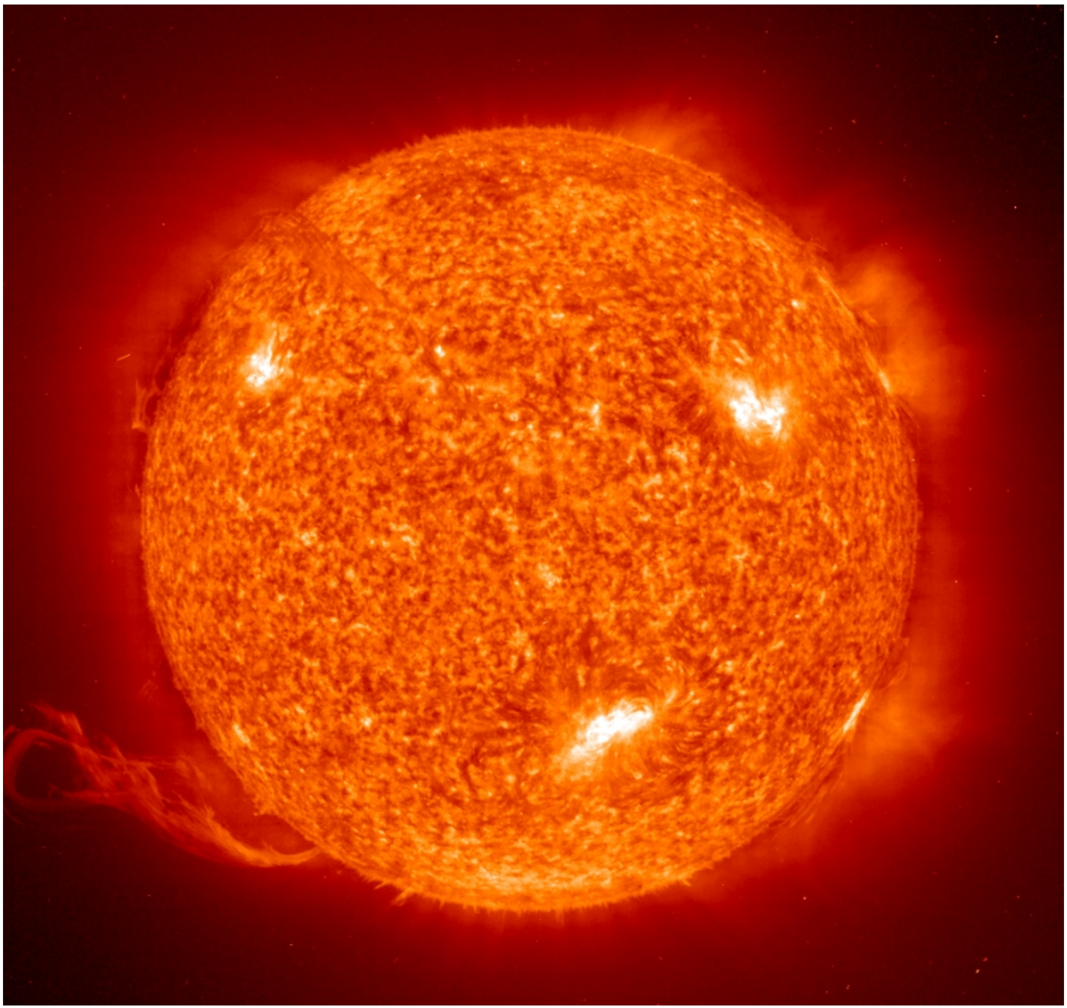


Figure 12.162: **Global Sun overview (source: NASA)**

Using the virial theorem proven in the section of Continuum Mechanics (see page 3897) that gives us:

$$E_c=-\frac{1}{2}E_p \qquad (12.2.61)$$

we have for a homogeneous spherical gas of radius R and mass M composed of N bodies, the following relations proved for the first one in the section of Continuum Mechanics (see page 3902) of the second in the section of Classical Mechanics (see page 3356):

$$E_c=\frac{3}{2}kNT \quad \text{and} \quad E_p=-\frac{3}{5}\frac{GM^2}{R} \qquad (12.2.62)$$

Therefore:

$$\frac{3}{2}kNT=-\frac{1}{2}\left(-\frac{3}{5}\frac{GM^2}{R}\right) \tag{12.2.63}$$

where for recall k is the Boltzmann constant.

Which gives:

$$T=\frac{1}{5}\frac{GM^2}{NkR} \tag{12.2.64}$$

in order to not make the confusion between to constant of ideal gas denoted R and the radius it is more convenient to write the latter relation as (and making the Boltzmann constant more explicit):

$$\boxed{T=\frac{1}{5}\frac{GM^2}{nk_Br}} \tag{12.2.65}$$

With for a given star N being the ratio of the total mass of the star on the average mass of a molecule.

For the Sun (actually a G2V-type star, aka yellow dwarf) it comes that $T\cong 10^7$ [K] (i.e. 10 millions degrees).

This is the central temperature of the Sun. Optical measurements measured from Earth only give the surface temperature (chromosphere), thus $6,000$ [K]. The calculated internal temperature is about $1,600$ times higher than at the surface. Independent methods based on nuclear reactions in the center of the Sun (measurement of solar neutrino flux) give the same order of magnitude, but the precise values differ by a factor of 2-3.

Layer	Part	Temperature [K]
Inner Layers	Core	$15,000,000$
	Radiative zone	$7,000,000$
	Convection zone	$2,000,000$
Outer Layers	Photosphere	$4,000$-$6,500$
	Chromosphere	$4,000$-$8,000$
	Transition region	$8,000$-$500,000$
	Corona	$500,000$-$1,000,000$

Table 12.17: **Summary of the average temperatures of typical Sun's layers**

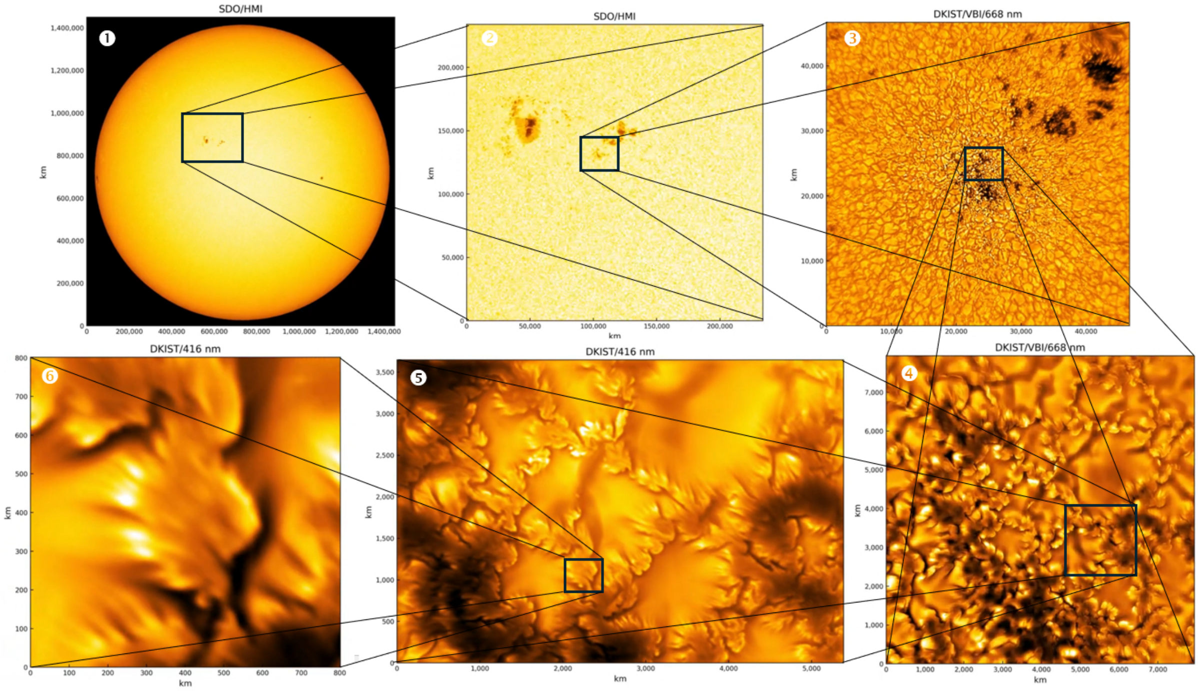

SDO/HMI
①
1,400,000
1,200,000
1,000,000
800,000
600,000
400,000
200,000
0
km
200,000
400,000
600,000
800,000
1,000,000
1,200,000
1,400,000
SDO/HMI
②
200,000
150,000
100,000
50,000
km
50,000
100,000
150,000
200,000
km
DKIST/VBI/668 nm
③
40,000
30,000
20,000
10,000
km
10,000
20,000
30,000
40,000
DKIST/416 nm
⑥
800
700
600
500
400
300
200
100
km
100
200
300
400
500
600
700
800
km
DKIST/416 nm
⑤
3,500
3,000
2,500
2,000
1,500
1,000
500
km
1,000
2,000
3,000
4,000
5,000
km
DKIST/VBI/668 nm
④
7,000
6,000
5,000
4,000
3,000
2,000
1,000
km
1,000
2,000
3,000
4,000
5,000
6,000
7,000
km

Figure 12.163: **Kelvin-Helmholtz instability on the Sun observed using the NSF Daniel K. Inouye Solar Telescope in Maui, Hawaii in 12026**

Let us now see how good we can estimate the external temperature!

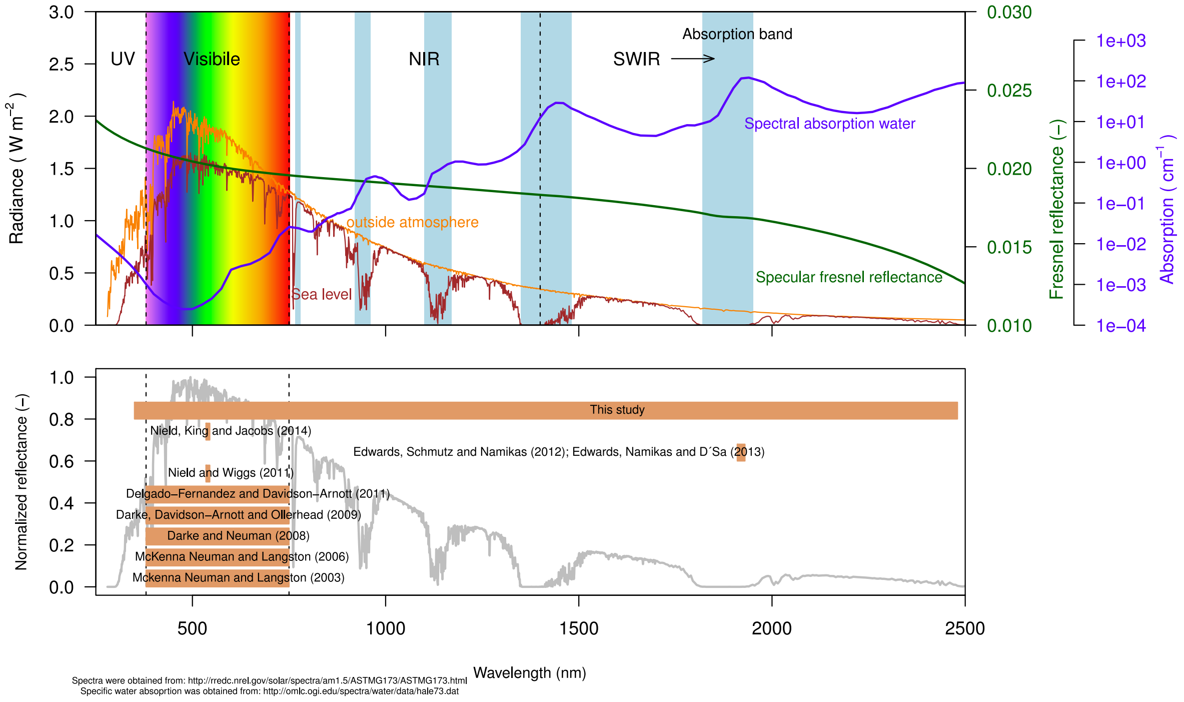
12.2.1.1.4 External Temperature

We observe and measure the black body spectrum of our Sun as having actually a peak of emission at: 490 [nm]. Therefore using Wien's Displacement law (see section Mechanics page 3727):

$$T=\frac{2,897.8\cdot 10^{-6}\,[\text{mK}]}{\lambda_{\text{max}}}=\frac{2,897.8\cdot 10^{-6}\,[\text{mK}]}{490\cdot 10^{-9}\,[\text{m}]}=5,900\,[\text{K}] \qquad (12.2.66)$$

A very basic situation is to calculate the rate at which the Sun heats the atmosphere at sunrise (a better denomination is "at daybreak").

Below the reader can see a plot of the spectral power distribution of sunlight (in $[\text{W}\cdot\text{m}^{-2}]$ outside the atmosphere (yellow line) and at sea level (brown line), to illustrate the effect of absorption of sunlight by water (see [418]). Included in the top panel: spectral Fresnel reflectance for a water surface (green line) and absorption coefficients (in $[\text{cm}^{-1}]$ for pure water (blue line). Bottom panel: overview of the studies (including this study) that demonstrate the potential of relating reflectance to surface moisture. The orange bars indicate the spectral range in which the measurements were taken.



Spectra were obtained from: http://rredc.nrel.gov/solar/spectra/am1.5/ASTMG173/ASTMG173.html
Specific water absoprtion was obtained from: http://omlc.ogi.edu/spectra/water/data/hale73.dat

Figure 12.164: Spectral power distribution of sunlight (source: [418])

Remark

Spectra were obtained from:

http://irredc.nrel.govisolar/spectra/am1.5/ASTMG173/ASTMG173.html

Specific water absorption were obtained from:

http://omlc.ogi.edu/spectra/water/data/hale73.dat

Now that we have the approximate temperature of our Sun, we can use the Stefan-Boltzmann law to get the energy emitted (see section Thermodynamics page 3706):

$$M(T)=\frac{2\pi^5k^4}{15c^2h^3}T^4=\sigma T^4 \tag{12.2.67}$$

with:

$$\sigma\cong\frac{2\pi^5k^4}{15c^2h^3}=\frac{\pi^2k^4}{60c^2\hbar^2}=5.67051\cdot 10^{-8}\left[\mathrm{W}\cdot\mathrm{m}^{-2}\cdot\mathrm{K}^4\right] \tag{12.2.68}$$

Therefore we get for the Sun:

$$M(T)=\frac{2\pi^5k^4}{15c^2h^3}T^4\cong 6.87\cdot 10^7\;[\mathrm{W}\cdot\mathrm{m}^{-2}] \tag{12.2.69}$$

Knowing the average distance from the Sun (actually a G2V-type star, aka yellow dwarf) to be about $1.496\cdot 10^{11}$ [m] $=1$ [UA] (Astronomical Unit), we can calculate the surface of the sphere at $R=1$ [UA] and thus the solar power P. Therefore:

$$S=4\pi R^2\cong 2.812\cdot 10^{23}\;[\mathrm{m}^2] \tag{12.2.70}$$

and:

$$P_{\odot}=S\cdot M(T)\cong 3.861\cdot 10^{26}\,[\mathrm{W}] \tag{12.2.71}$$

Remark

From $P_{\odot}=S\cdot M(T)\cong 3.861\cdot 10^{26}\,[\mathrm{W}]$ we can calculate the actual Solar flux (i.e. Sun flux) - or "peak Sun hours" - on Earth assuming that the latter is at an average distance of $149.09\cdot 10^{9}\,[\mathrm{m}]$ of the Sun:

$$\Phi_{\odot}=\frac{P_{\odot}}{4\pi R^2}=\frac{3.861\cdot 10^{26}}{4\pi(149.09\cdot 10^{9})^2}\cong 1,382\,[\mathrm{W}\cdot\mathrm{m}^{-2}]$$

The average emittance also named "average bolometric emittance" received by the Earth outside the atmosphere, also named "solar constant" (which is in fact not constant ... on a scale of several billion years), is directly measurable in orbit and is equal to $\sim 1,373\,[\mathrm{W}\cdot\mathrm{m}^{-2}]$.

Assuming an albedo for Earth of $\rho\cong 0.31$ (but there is considerable variation in this value over the surface of the Earth and over time as well), then:

$$\Phi_{\rho}=(1-\rho)\cdot 1,382\cong 953\,[\mathrm{W}\cdot\mathrm{m}^{-2}]$$

However the Earth isn't all the time facing the same part of its surface to the Sun. Assuming that only a third of a 24 hours days is facing the Sun, then:

$$\bar{\Phi}_{\rho}=\frac{1}{3}\cdot 1,382\cong 318\,[\mathrm{W}\cdot\mathrm{m}^{-2}]$$

Assuming known the radius of the Sun as being $r_{\odot}\cong 6.9599\cdot 10^{8}$ [m], we can calculate its surface S and its solar radiative emittance $M_{\odot}(T)$. So:

$$S_{\odot}=4\pi r_{\odot}^2\cong 6.087\cdot 10^{18}\,[\mathrm{m}^2] \tag{12.2.72}$$

and:

$$M_{\odot}(T)=\frac{P_{\odot}}{S_{\odot}}\cong 6.343\cdot 10^{7}\,[\mathrm{W}\cdot\mathrm{m}^{-2}] \tag{12.2.73}$$

Remark

The radiating surface of a star is named "photosphere". Indeed, as stars, excepting neutron stars, have no solid surface, the photosphere is typically used to describe the Sun's or another star's visual surface.

Using the Stefan-Boltzmann law (see section Statistical Mechanics page 3706), we can now calculate approximately the thermodynamic temperature of the photosphere:

$$T=\sqrt[4]{\frac{M_{\odot}(T)}{\sigma}}\cong 5,780\,[\mathrm{K}] \tag{12.2.74}$$

which is very accurate to direct measurement!!! More generally the previous relation is written:

$$T=\sqrt[4]{\frac{P_{\odot}(T)}{4\pi r_{\odot}^2\sigma}} \tag{12.2.75}$$

or respecting the notation of optical geometry:

$$\boxed{T=\sqrt[4]{\frac{L_\odot(T)}{4\pi r_\odot^2\sigma}}} \tag{12.2.76}$$

So since we can estimate the luminosity and the temperature of a star (or something that looks like...) we can also estimate it's radius!

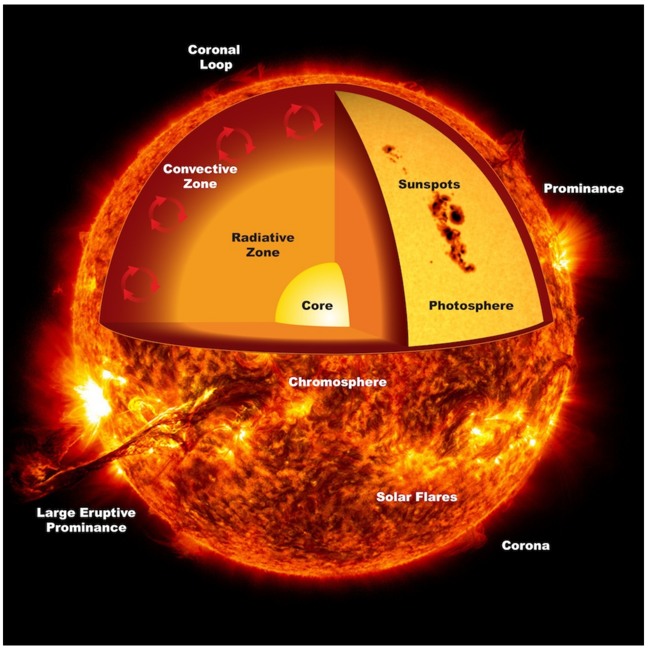

Coronal Loop
Convective Zone
Sunspots
Prominance
Radiative Zone
Core
Photosphere
Chromosphere
Solar Flares
Large Eruptive Prominance
Corona

Figure 12.165: **Layer's view of our Sun (source: NASA)**

Planck's law (see section Thermodynamics page 3719) applied at this temperature allow us to calculate the spectral distribution of solar radiation and then we see that the maximum of the intensity is in the visible range (our visible range!!!) spectrum which is from 400 [nm] to 700 [nm].

Obviously the reader has to keep in mind that the temperature of a star is an information that doesn't make really sense (we should speak instead of "average temperature", i.e. an integral of the spectrum function as we know it!) as the temperature depends on the wavelength. As provided by the paper [86], we can see the temperature of our Sun (actually a G2V-type star, aka yellow dwarf) in function of the wavelength here below as measured in the 12008-08-24 according to the holocene calendar (the lower figure is just a zoom on the visible spectrum):

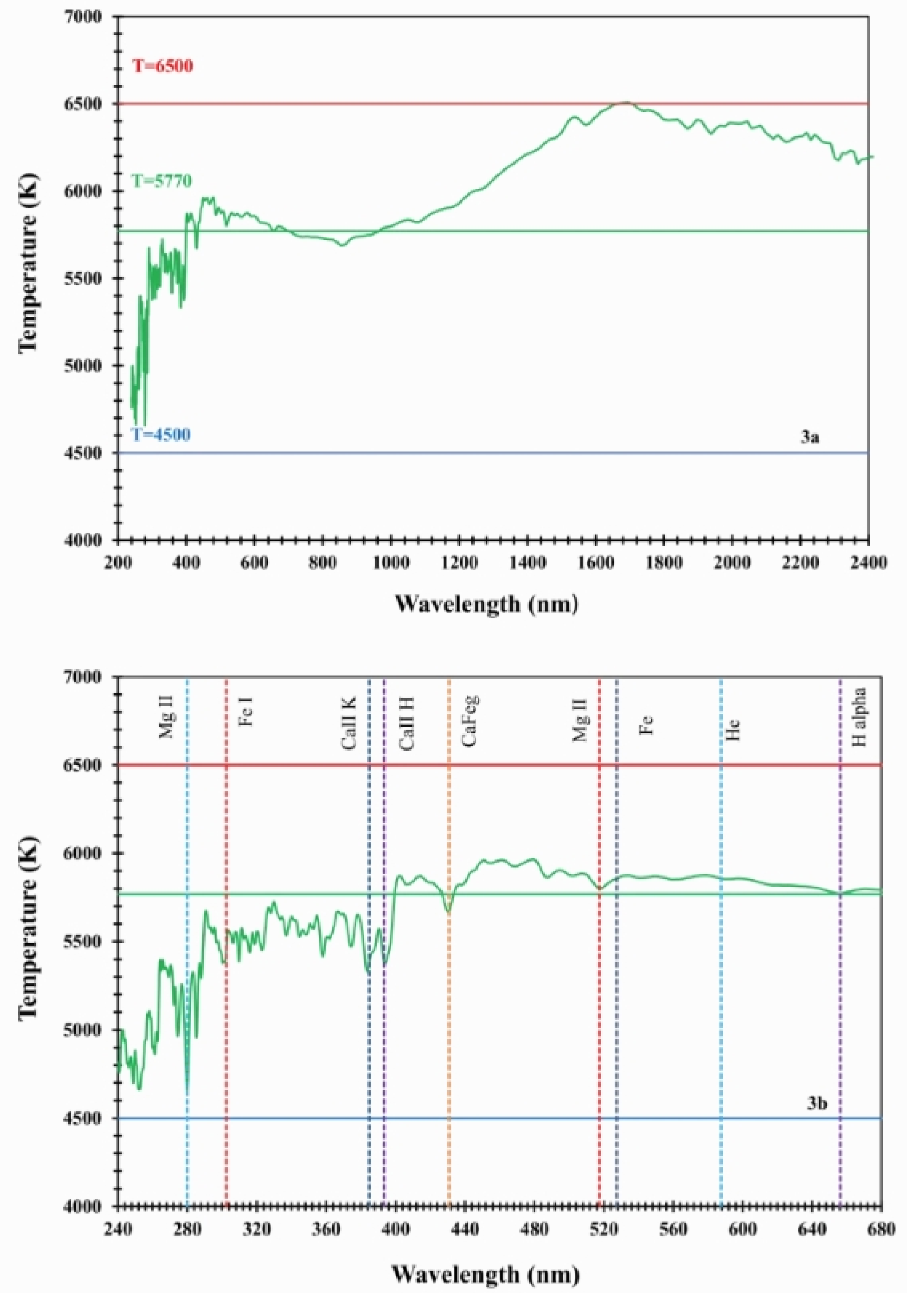


Figure 12.166: **Sun temperature in function of wavelength the 12008-08-24**

And in the same paper we can see the evolution of our Sun's temperature for 17 years (daily average) on the H_α wavelength (656.2 [nm]):

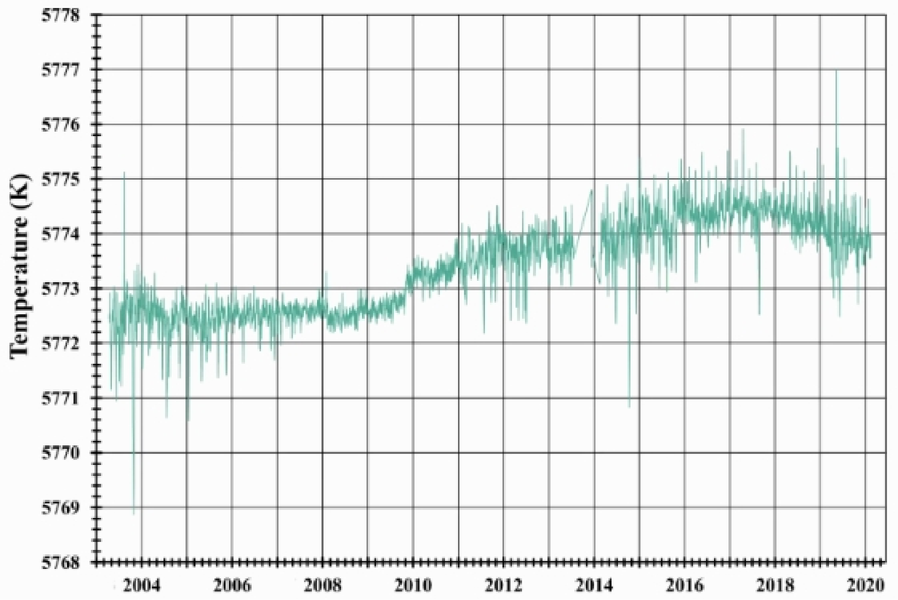


Figure 12.167: Sun temperature in function of time

In the above plot the years are given based on the Gregorian calendar.

In the figure below, we can see an extrapolation of mean surface temperatures on Earth taking into account the slow increase of the insolation due to the aging of the Sun (a mean increase of $\sim 0.33\,[\mu\mathrm{W}\cdot\mathrm{m}^{-2}]$). Simulations where performed with a tridimensional global climate model described in Leconte et al. (2013; `https://ui.adsabs.harvard.edu/abs/2013Natur.504..268L/abstract`). When the average insolation received by the Earth reaches $375\,[\mathrm{W}\cdot\mathrm{m}^{-2}]$ (about 1 billion years from now), the greenhouse effect of water vapor enters a runaway regime that causes the total evaporation of the oceans. Temperatures above that threshold are very uncertain but could be greater than $1,000\,[^\circ\mathrm{C}]$.

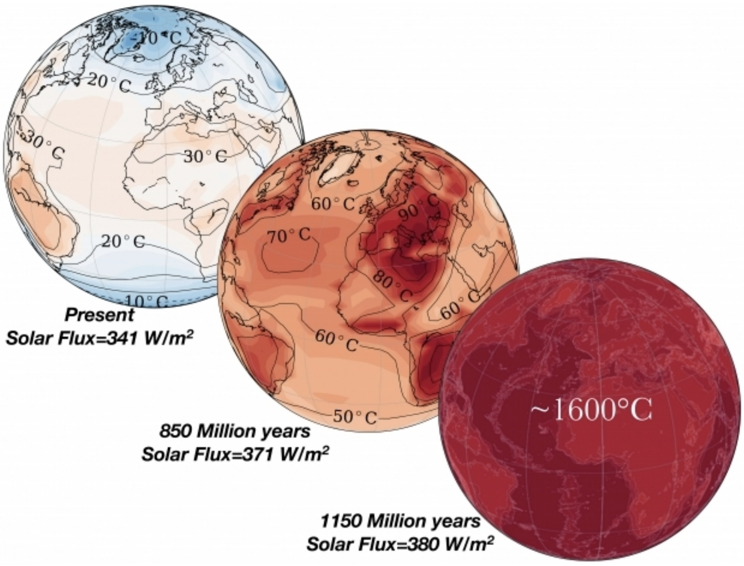

10°C
20°C
30°C
30°C
20°C
10°C
Present
Solar Flux=341 W/m²
60°C
90°C
70°C
80°C
60°C
60°C
50°C
850 Million years
Solar Flux=371 W/m²
~1600°C
1150 Million years
Solar Flux=380 W/m²

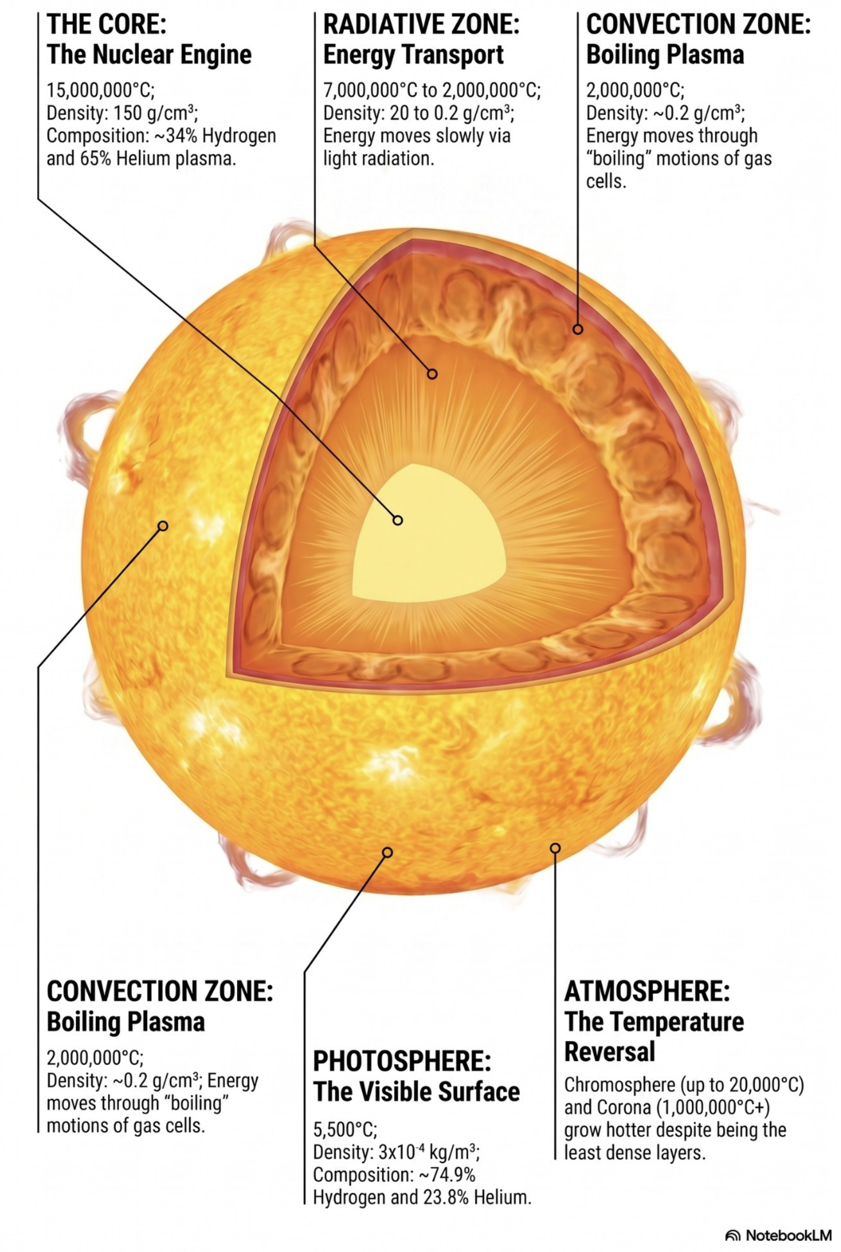
Figure 12.168: Numerical simulations of the surface temperature of the Earth at the spring equinox exposed to an increasingly bright Sun in the future.

So life will very likely disappear on Earth a veeeeery long time before the Sun (actually a

G2V-type star, aka yellow dwarf) turns into a supernova....

Remark

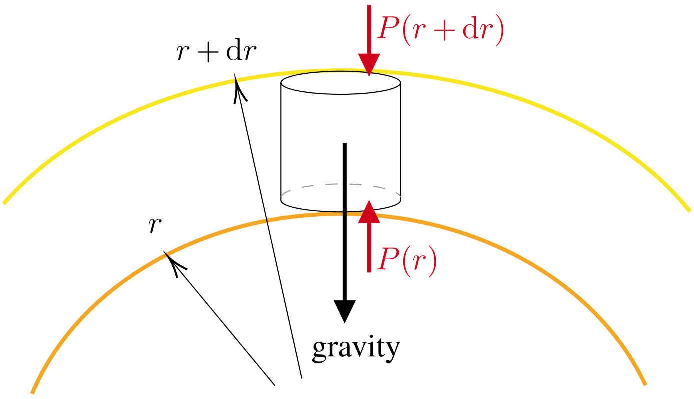
For information, the rotation of Earth's on itself slows down at the day we write these lines of approximately 2.3 milliseconds per century. That means 1 hour every 180 millions years. Each year the Earth's orbit speed around the Sun also slows down of approximately 3 nanometers-per-second. Using Kepler's law, that means that the Earth goes further away from the Sun of 3 micrometers per 1 billion year......!!!!


THE CORE:
The Nuclear Engine
15,000,000°C;
Density: 150 g/cm^3;
Composition: ~34% Hydrogen and 65% Helium plasma.
RADIATIVE ZONE:
Energy Transport
7,000,000°C to 2,000,000°C;
Density: 20 to 0.2 g/cm^3;
Energy moves slowly via light radiation.
CONVECTION ZONE:
Boiling Plasma
2,000,000°C;
Density: ~0.2 g/cm^3;
Energy moves through "boiling" motions of gas cells.
CONVECTION ZONE:
Boiling Plasma
2,000,000°C;
Density: ~0.2 g/cm^3; Energy moves through "boiling" motions of gas cells.
PHOTOSPHERE:
The Visible Surface
5,500°C;
Density: 3x10^{-4} kg/m^3;
Composition: ~74.9% Hydrogen and 23.8% Helium.
ATMOSPHERE:
The Temperature Reversal
Chromosphere (up to 20,000°C) and Corona (1,000,000°C+) grow hotter despite being the least dense layers.
NotebookLM

12.2.1.1.5 Equation of Hydrostatic Equilibrium

The absence of changes in most stars over timescales of hours or days indicates that the forces acting on the matter in the stars are essentially perfectly balanced (remember figure showed earlier above). Here we analyse this constraint in more detail.

In the figure below we have represented a piece mass shell in a spherically symmetric star where we consider in reality a very small cylindrical element between radius r and radius $r+\mathrm{d}r$ hence the fact that the element surface dS is considered as being the same (but we can neglect the pressure variation as it can be very big even in a small height difference):



If we denote by $M(r)$ the mass of star in the smaller radii and Δm the mass in the cylinder, we get:

$$F_g=-\frac{GM(r)\Delta m}{r^2} \tag{12.2.77}$$

Now for the pressure we have (net force due to difference in pressure between upper and lower faces):

$$F_P=P(r)\mathrm{d}S-P(r+\mathrm{d}r)\mathrm{d}S \tag{12.2.78}$$

Using the definition of the derivative but applied to the pressure:

$$\frac{P(r+\mathrm{d}r)-P(r)}{\mathrm{d}r}=\frac{\mathrm{d}P}{\mathrm{d}r}\Leftrightarrow P(r+\mathrm{d}r)=P(r)+\frac{\mathrm{d}P}{\mathrm{d}r}\mathrm{d}r \tag{12.2.79}$$

Therefore:

$$F_P=P(r)\mathrm{d}S-P(r+\mathrm{d}r)\mathrm{d}S=P(r)\mathrm{d}S-\left(P(r)+\frac{\mathrm{d}P}{\mathrm{d}r}\mathrm{d}r\right)\mathrm{d}S=-\frac{\mathrm{d}P}{\mathrm{d}r}\mathrm{d}r\mathrm{d}S \tag{12.2.80}$$

Now we have as mass element that is given as we know by:

$$\Delta m=\rho\mathrm{d}r\mathrm{d}S \tag{12.2.81}$$

Applying Newton's second law to the cylinder:

$$F_G+F_P=-\frac{-GM(r)\Delta m}{r^2}-\frac{\mathrm{d}P}{\mathrm{d}r}\mathrm{d}r\mathrm{d}S \tag{12.2.82}$$

This sum must be equal to 0 everywhere if the star is indeed static. Therefore:

$$-\frac{-GM(r)\rho \mathrm{d}r\mathrm{d}S}{r^2}-\frac{\mathrm{d}P}{\mathrm{d}r}\mathrm{d}r\mathrm{d}S=0 \tag{12.2.83}$$

After simplification we get the "equation of hydrostatic equilibrium" or "stellar structure equation":

$$\boxed{\frac{\mathrm{d}P}{\mathrm{d}r}=-\frac{GM(r)}{r^2}\rho} \tag{12.2.84}$$

So far we already estimated roughly the core temperature of the Sun (actually a G2V-type star, aka yellow dwarf) and of the photosphere. Let us now first estimate roughly its average pressure:

$$\bar{P}_{\odot}=\frac{\bar{F}_{\odot}}{S_{\odot}}\cong\frac{GM_{\odot}^2}{R_{\odot}^2}4\pi R_{\odot}^2=\frac{GM_{\odot}^2}{4\pi R_{\odot}^4}\cong 10^{14}\,[\text{Pa}] \tag{12.2.85}$$

A better estimation is given by using the equation of hydrostatic equilibrium:

$$\frac{\mathrm{d}P}{\mathrm{d}r}=-\frac{GM(r)\rho}{r^2}=-\frac{G}{r^2}\left(\frac{4}{3}\pi r^3\rho\right)\rho=-\frac{4}{3}\pi G\rho^2 r \tag{12.2.86}$$

and integrate (assuming density is constant):

$$\int_0^R -\frac{4}{3}\pi G\rho^2 r\mathrm{d}r=-\frac{4}{3}\pi G\rho^2\frac{R^2}{2} \tag{12.2.87}$$

So we have an expression for the central pressure:

$$\boxed{P_{\odot}=\frac{2}{3}\pi G\rho^2 R^2=\frac{3}{2}\frac{GM_{\odot}^2}{\pi R^2}} \tag{12.2.88}$$

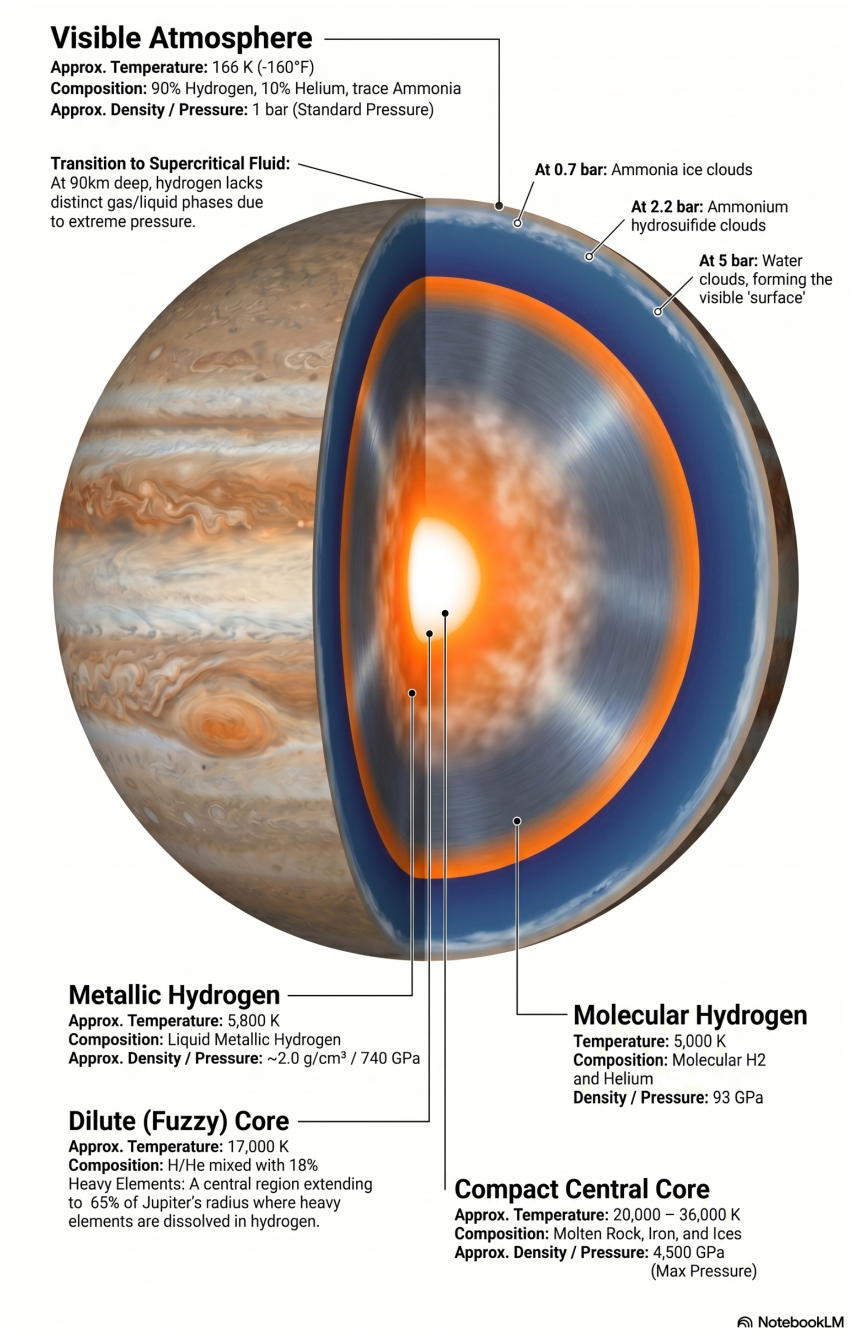
That is 6 times more than the previous roughly approximation and compared to direct measurement method this seems more accurate!

We also have for mean density of the Sun (actually a G2V-type star, aka yellow dwarf) :

$$\bar{\rho}_{\odot}=\frac{M_{\odot}}{V_{\odot}}=\frac{M_{\odot}}{\frac{4}{3}\pi R_{\odot}^3}=\frac{1.98855\cdot 10^{30}}{\frac{4}{3}\pi(6.9598\cdot 10^8)^3}\cong 1408.5\,[\text{kg}\cdot\text{m}^{-3}] \tag{12.2.89}$$

to compare with the density of pure water that is $1,000\,[\text{kg}\cdot\text{m}^{-3}]$ or to that of iron that is $7,874\,[\text{kg}\cdot\text{m}^{-3}]$. So we understand better why space image of the Sun looks like a big liquid sphere of gas as because of the gravity conditions the pressure is such that the gas is reduced to a density greater than that of water in average!!!!!

Here is a figure of what we think so far as comparison for Jupiter that is like a non-initiated star:


Visible Atmosphere
Approx. Temperature: 166 K (-160°F)
Composition: 90% Hydrogen, 10% Helium, trace Ammonia
Approx. Density / Pressure: 1 bar (Standard Pressure)
Transition to Supercritical Fluid: At 90km deep, hydrogen lacks distinct gas/liquid phases due to extreme pressure.
At 0.7 bar: Ammonia ice clouds
At 2.2 bar: Ammonium hydrosuifide clouds
At 5 bar: Water clouds, forming the visible 'surface'
Metallic Hydrogen
Approx. Temperature: 5,800 K
Composition: Liquid Metallic Hydrogen
Approx. Density / Pressure: ~2.0 g/cm³ / 740 GPa
Molecular Hydrogen
Temperature: 5,000 K
Composition: Molecular H2 and Helium
Density / Pressure: 93 GPa
Dilute (Fuzzy) Core
Approx. Temperature: 17,000 K
Composition: H/He mixed with 18% Heavy Elements: A central region extending to 65% of Jupiter's radius where heavy elements are dissolved in hydrogen.
Compact Central Core
Approx. Temperature: 20,000 – 36,000 K
Composition: Molten Rock, Iron, and Ices
Approx. Density / Pressure: 4,500 GPa (Max Pressure)
NotebookLM

Using the equation of hydrostatic equilibrium we can estimate roughly the density of the photosphere:

$$\rho(R_\odot) = -\frac{\Delta P}{\Delta r}\frac{r^2}{GM(r)} \cong \frac{\bar{P}_\odot}{R_\odot}\frac{R_\odot^2}{GM} = \frac{\bar{P}_\odot R}{GM} \cong 525\,[\mathrm{kg}\cdot\mathrm{m}^{-3}] \tag{12.2.90}$$

So it is obvious that the density of the Sun decreases continuously outward from the center. The visible surface of the Sun (i.e. the photosphere) is a very thin layer, only about 500 [km] thick as compared to the radius of the Sun. The density of the photosphere is very, very low, about $0.2 \cdot 10^{-4}\,[\mathrm{kg}\cdot\mathrm{m}^{-3}]$ as estimated by observations. Therefore the average density of the Sun can only be explained by the density of its core that is very high, about $160,000\,[\mathrm{kg}\cdot\mathrm{m}^{-3}]$, much higher than any material that we know.

12.2.1.1.6 Brightness

The "intrinsic bolometric brightness" of a star corresponds to the total power radiated in the entire electromagnetic spectrum in the direction of the observer expressed relatively to the total power radiated by the Sun. Assuming all stars spherical and isotropic, we can express it in solar units:

$$L_{\text{bol}}=\frac{P_{\text{star}}}{P_{\odot}} \tag{12.2.91}$$

The radiated power P is calculated, as we know, by multiplying the radiative emittance (Stefan-Boltzmann law) by the surface of the star:

$$P=M(T)S=\sigma T^4 4\pi r^2 \tag{12.2.92}$$

The intrinsic bolometric luminosity of a star is therefore proportional to the square of its radius and the fourth power of its surface temperature. Taking the Sun (actually a G2V-type star, aka yellow dwarf) as a reference, the constants are simplified. We can the write:

$$L_{\text{bol}}=\frac{P_{\text{star}}}{P_{\odot}}=\left(\frac{T}{T_{\odot}}\right)^4\left(\frac{r}{r_{\odot}}\right)^2=c^{te}T^4r^2 \tag{12.2.93}$$

with $r_{\odot}\cong 6.9559\cdot 10^8$ [m] and $T_{\odot}\cong 5,780$ [K] hence $c^{te}\cong 1.85\cdot 10^{-33}$ $[\text{K}^{-4}\cdot\text{m}^2]$.

In astrophysics, we also use a logarithmic scale to express the bolometric luminosity of a star: the "absolute magnitude M". This unit has an empirical origin that will be explained below.

Caution! Estimate the mass of a distant star or distant galaxy based solely on its luminosity compared to that of the Sun is very hazardous given the diversity of properties of stars (and the same applies to galaxies). So if you do something like this never forget to indicate (or to check if you read a paper) the margin error!!! Never forget that a scientific paper without margin error is always suspicious!

Notice that if we assume that Sun emits isotropically (nearly uniform in all directions). The fraction of photons that reaches Earth is the ratio of Earth's cross-sectional area to the surface area of a sphere centered on the Sun with radius equal to the Sun–Earth R_d distance:

$$N_{\text{photons}}=\frac{\pi R_{♁}^2}{4\pi R_d^2}=\frac{R_{♁}^2}{4R_d^2}\cong\frac{(6.37\cdot 10^6)^2}{4(1.496\cdot 10^{11})^2}\cong 4.54\cdot 10^{-18} \tag{12.2.94}$$

Keeping in mind that the photons intercepted by Earth include those absorbed or scattered in the atmosphere and those reaching the surface; the fraction above is purely geometric (photons that impinge on Earth's disk).

12.2.1.1.7 Shining (apparent brightness)

Perhaps the easiest measurement to make of a star is its apparent brightness. We are purposely being careful about our choice of words. When we say "apparent brightness", we mean how bright the star appears to a detector here on Earth.

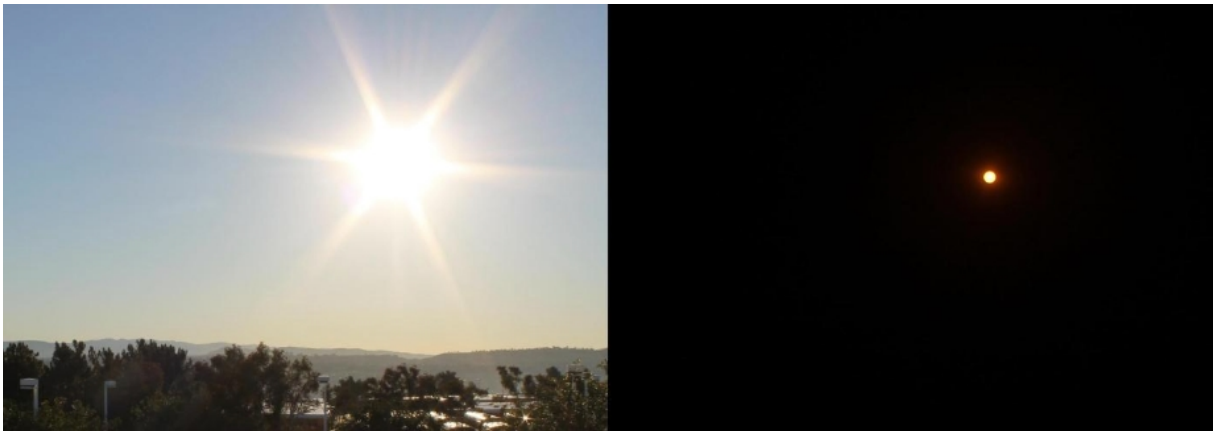
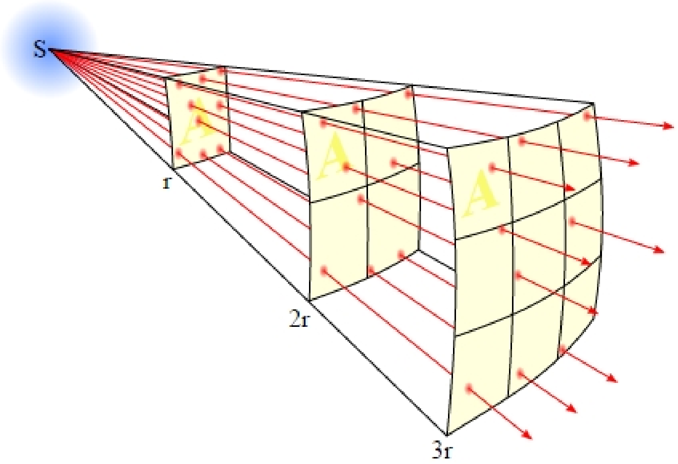


Figure 12.169: Sun appearance in a photograph with and without a solar filter (author: John Krieger)

Definition (#12.2.538): The "brilliance" or "shining" or "apparent brightness" b of a star is the density of radiation received by the observer, that is to say equal to the flow of energy (power of the star at its surface) divided by the sphere surface with the radius equal to the distance which separates the observer from the star:

$$\boxed{b=\frac{P}{S}=\frac{P}{4\pi d^2}} \tag{12.2.95}$$

The brilliance decreases therefore with the square of the distance (as in myna other filed of physics):



✍**Example:**

A star has a known luminosity of $9.7 \cdot 10^{27}\,[\text{W}]$. Observations of the star show that the apparent brightness of light received on Earth from the star is $114\,[\text{nW} \cdot \text{m}^{-2}]$ (the human eye's sensitivity to light varies, but the minimum intensity of light detectable is approximately $0.1\,[\text{nW} \cdot \text{m}^{-2}]$).

Therefore:

$$b = \frac{P}{4\pi d^2} \Rightarrow d = \sqrt{\frac{P}{4\pi b}} = \sqrt{\frac{9.7 \cdot 10^{27}}{4\pi \cdot (114 \cdot 10^{-9})}} \cong 8.2 \cdot 10^{16}\,[\text{m}] \cong 8.67\,\text{light years} \tag{12.2.96}$$

It is important to notice that this quantity has no direct relation with the physical intrinsic properties of the respective star (unlike the bolometric brightness!).

Thus, two identical stars can have the same apparent brightness if (and only if) they lie at the same distance from Earth. However, as illustrated in Figure below, two different stars can appear equally bright if the more luminous one lies farther away. A bright star (that is, a star with large apparent brightness) is a powerful emitter of radiation, is near Earth, or both. A dim star is a weak emitter, is far from Earth, or both:

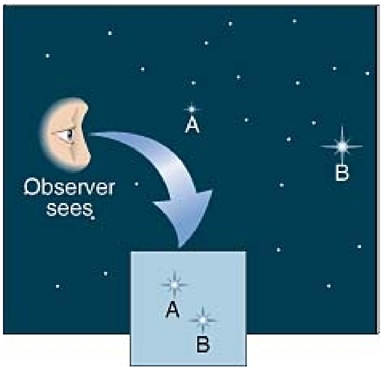


Figure 12.170: **Apparent luminosity**

The luminosity of a star, on the other hand, is the amount of light it emits from its surface. The difference between luminosity and apparent brightness depends on distance as we know now. Another way to look at these quantities is that the luminosity is an intrinsic property of the star, which means that everyone who has some means of measuring the luminosity of a star should find the same value. However, apparent brightness is not an intrinsic property of the star; it depends on your location. So everyone will measure a different apparent brightness for the same star if they are all different distances away from that star.

Apparent brightness is the brightness perceived by an observer on Earth and absolute brightness is the brightness that would be perceived if all stars were magically placed at the same standard distance. There can be a great difference between the total amount of radiation a star emits and the amount of radiation measured at the Earth's surface.

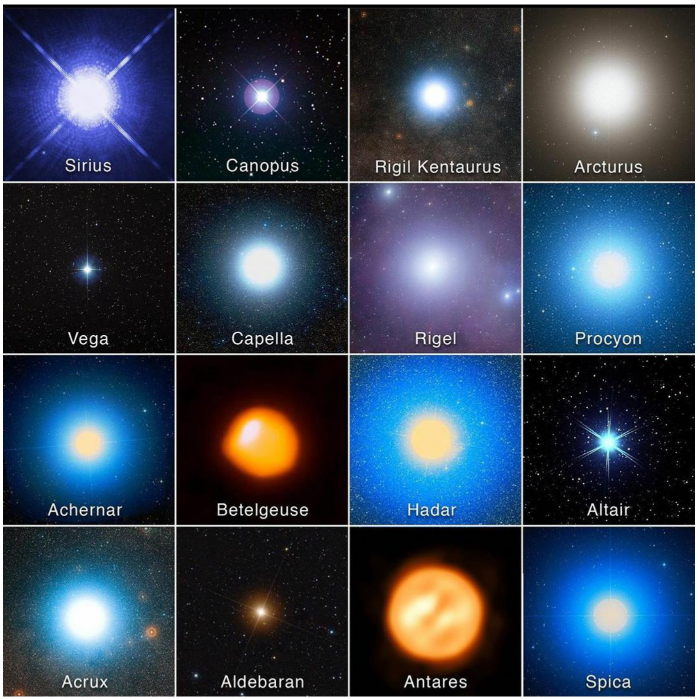

Sirius
Canopus
Rigil Kentaurus
Arcturus
Vega
Capella
Rigel
Procyon
Achernar
Betelgeuse
Hadar
Altair
Acrux
Aldebaran
Antares
Spica

Figure 12.171: Most brightest stars at night in early 121st century

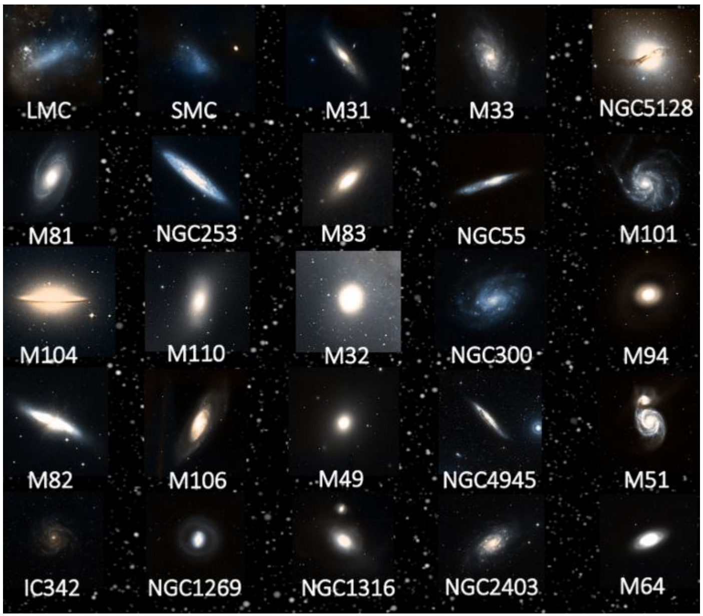

LMC
SMC
M31
M33
NGC5128
M81
NGC253
M83
NGC55
M101
M104
M110
M32
NGC300
M94
M82
M106
M49
NGC4945
M51
IC342
NGC1269
NGC1316
NGC2403
M64

Figure 12.172: Most brightest galaxies at night in early 121st century

The figure below (this figure obviously doesn't depict the correct location of each object relative to each other...!) depicts the approximate apparent sizes from Earth of various different deep space objects if they were brighter (this is how they would appear approximately in our night sky). The images are approximately in scale with one another, including the Moon, but not to the Milky Way background (in real life you can forget to see all these objects when the Moon is bright during the night with naked eyes because it will overflow your eyes photoreceptor, you should wait for the new Moon - i.e. when the Moon is not visible during the night!):

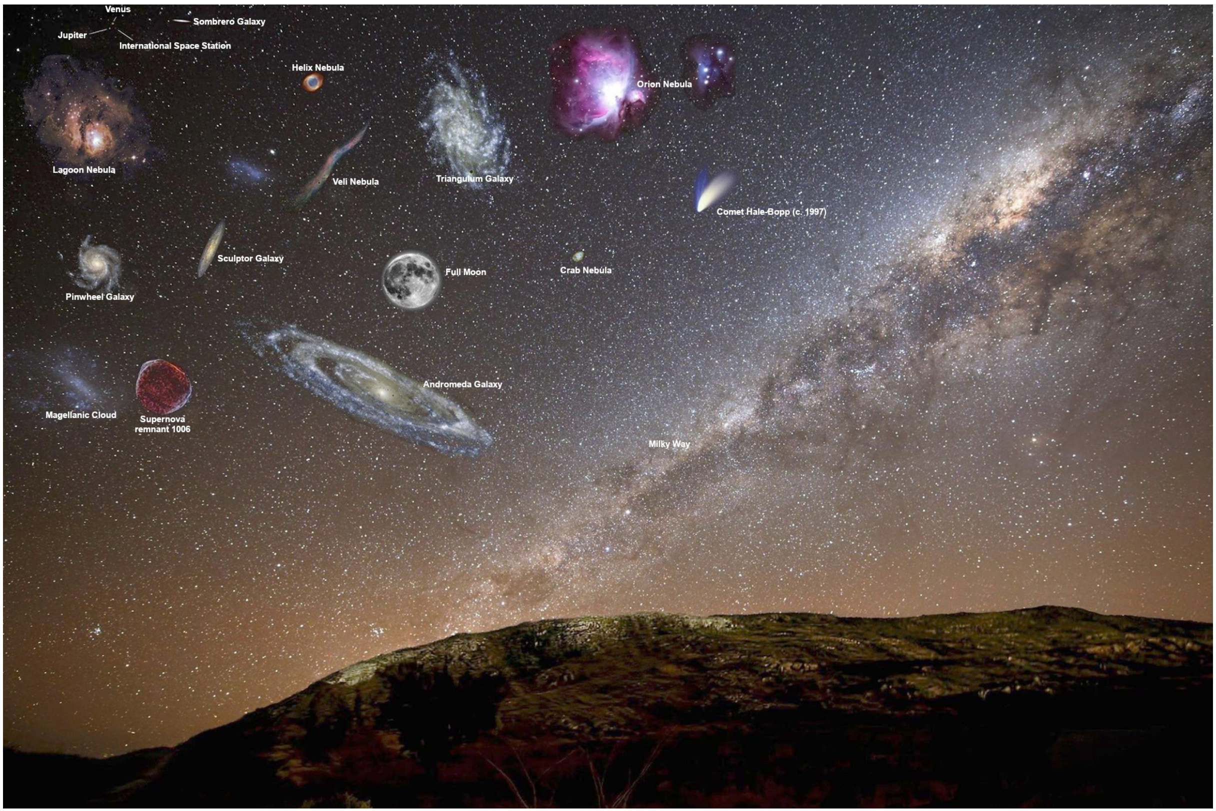

Venus
Jupiter
Sombrero Galaxy
International Space Station
Helix Nebula
Orion Nebula
Lagoon Nebula
Veil Nebula
Triangulum Galaxy
Comet Hale-Bopp (c. 1997)
Sculptor Galaxy
Full Moon
Crab Nebula
Pinwheel Galaxy
Andromeda Galaxy
Magellanic Cloud
Supernova remnant 1006
Milky Way

Figure 12.173: **Approximate sky if important objects had similar brightness**

In astrophysics, we also use another scale of measurement where the apparent brightness is given by another magnitude of empirical origin: the apparent magnitude, which will be explained immediately below.

12.2.1.1.8 Apparent magnitude

Ptolemy in 9863 (holocence calendar) had defined a scale of six magnitudes to express the brightness (shining) of stars, the first for the brightest and the sixth for the stars just visible to the naked eye (6 magnitudes and therefore 5 gaps).

During the 119th century (holocene calendar), with the arrival of new photometric observations techniques (photographic and photoelectric), the scale of magnitude was replaced by that of "apparent magnitude" m that has been defined so that this new scale is close to the old one.

The definition is the following:

- The scale is logarithmic in base 10 (for convenience of the magnitude of manipulated quantities)
- There are 5 magnitude gaps corresponding to an apparent brightness ratio of 1 for 100 $(1:100)$
- The scale is inverse (high magnitude corresponds to a small apparent magnitude/ brightness).

Using these definitions, we can construct a way relating the shining (brilliance) of two stars to their apparent magnitude m.

For a star $S1$, one hundred times brighter than a star $S2$, the star $S1$ is 5 magnitude units above the star $S2$ (remember that the scale is reversed!). So a ratio of:

$$\frac{e_2}{e_1}=\frac{\dfrac{P_2}{d_2^2}}{\dfrac{P_1}{d_1^2}}=\frac{P_2}{P_1}\frac{d_1^2}{d_2^2}=100=10^2 \tag{12.2.97}$$

corresponds by definition to:

$$\Delta m=m_1-m_2=5 \tag{12.2.98}$$

We can then put the relations:

$$\log\left(\frac{e_2}{e_1}\right)\to 2 \quad \text{and} \quad m_1-m_2\to 5 \tag{12.2.99}$$

By applying the rule of three, we build:

$$\frac{\log\left(\dfrac{e_2}{e_1}\right)}{2}=\frac{m_1-m_2}{5} \tag{12.2.100}$$

By simplifying, we find the "Pogson's Formula" which expresses the relation between (visual) apparent magnitudes and brilliance (shining) of two stars:

$$\boxed{m=m_1-m_2=2.5\log\left(\frac{e_2}{e_1}\right)} \tag{12.2.101}$$

Then we are indeed compliant with the arbitrary rule summarized in the following figure:

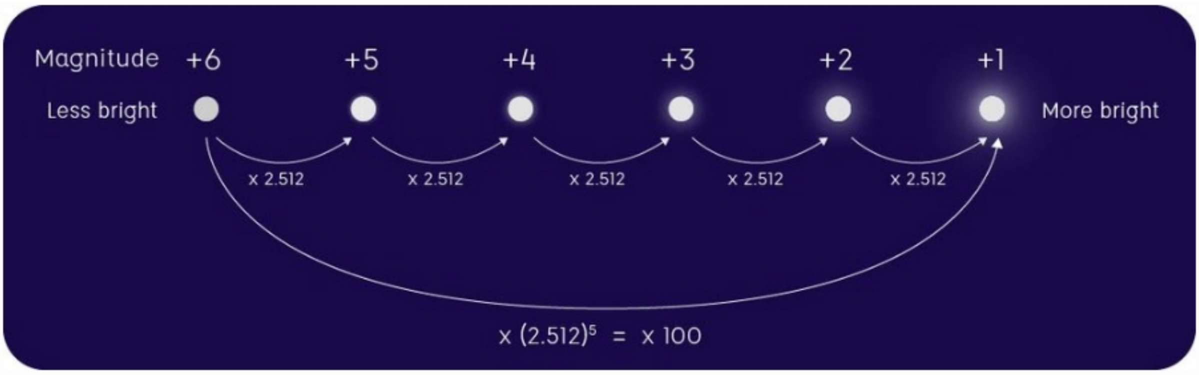


Figure 12.174: **Magnitude and brightness Ptolemy-Pogson relations**

A fourth magnitude star is 2.512 times as bright as a fifth magnitude star, and a second magnitude star is $(2.512)^4=39.82$ times brighter than a sixth magnitude star.

The following table shows how the difference in apparent magnitude between two stars (m_2-m_1) corresponds to the ratio of their apparent brightnesses (e_1/e_2):

Apparent magnitude difference $(m_2 - m_1)$	**Ratio of apparent brightness** (e_2/e_1)
1	2.512
2	$2.512^2 = 6.31$
3	$2.512^3 = 15.85$
4	$2.512^4 = 39.82$
5	$2.512^5 = 100$
10	$2.512^{10} = 10^4$
20	$2.512^{20} = 10^8$

Apart from small corrections, the brightness of Vega[27] (α Lyr) still serves as the definition of zero magnitude for visible and near infra-red wavelengths. The brightness of Vega is exceeded by four stars in the night sky at the 121st century (holocene calendar) at visible wavelengths (and more at infra-red wavelengths) as well as bright planets such as Venus, Mars, and Jupiter, and these must be described by negative magnitudes. For example, Sirius, the brightest star after the Sun (the latter having an apparent magntitude of 26.9) of the celestial sphere, has an apparent magnitude of -1.4.

To get an idea of the (visual) apparent magnitudes relatively to Vega here are some examples: Sun $m_\odot = -26.74$, full Moon $m = -15$, Venus maximum $m = -4.8$, Sirius $m = -1.4$ (spectral type A1 and distant of 8.6 light years), limit perceived with the naked eye 6, perception limit through an amateur telescope of 15 [cm] at this date (year 12003 according to holocene calendar) $m = 13$ limits of perception Hubble space telescope $m = 30$.

✍**Example:**

Now as we know that the apparent magnitude of the Sun (actually a G2V-type star, aka yellow dwarf) is -26.74 (brighter), and the mean apparent magnitude of the full Moon is -12.74 (dimmer) the difference in apparent magnitude is obviously that $\delta m = 14.00$

With this information reconsider that the Pogson formula also gives by construction the ratio of the luminosity. So relatively to our example we get:

$$-2.5\log\left(\frac{L}{L_\odot}\right) = 14 \qquad (12.2.102)$$

After rearranging we get therefore:

$$\frac{L}{L_\odot} = 10^{-14/2.5} \qquad (12.2.103)$$

Or better to have a nicer number:

$$\frac{L_\odot}{L} \cong 398,107 \qquad (12.2.104)$$

The Sun appears about $400,000$ times brighter than the full Moon.

---

[27] Brightest star in the constellation Lyra at this day (121st century according to holocene calendar). It is actually a relatively close star at only 25 light-years from Earth, and, together with Arcturus and Sirius, one of the most luminous stars in the Sun's neighbourhood

It should be noticed that the (visual) apparent magnitude does not exactly match the real apparent magnitude, because the eye is not equally sensitive to all wavelengths. The blue or red stars seem less bright to the eye than they actually are because some of the radiation is in the ultraviolet, respectively in the infra-red.

As one of my friend didn't know it (...) just a remark about the Moon... Many books and Internet websites show the classical following figure (where there is basically not too much to say about):

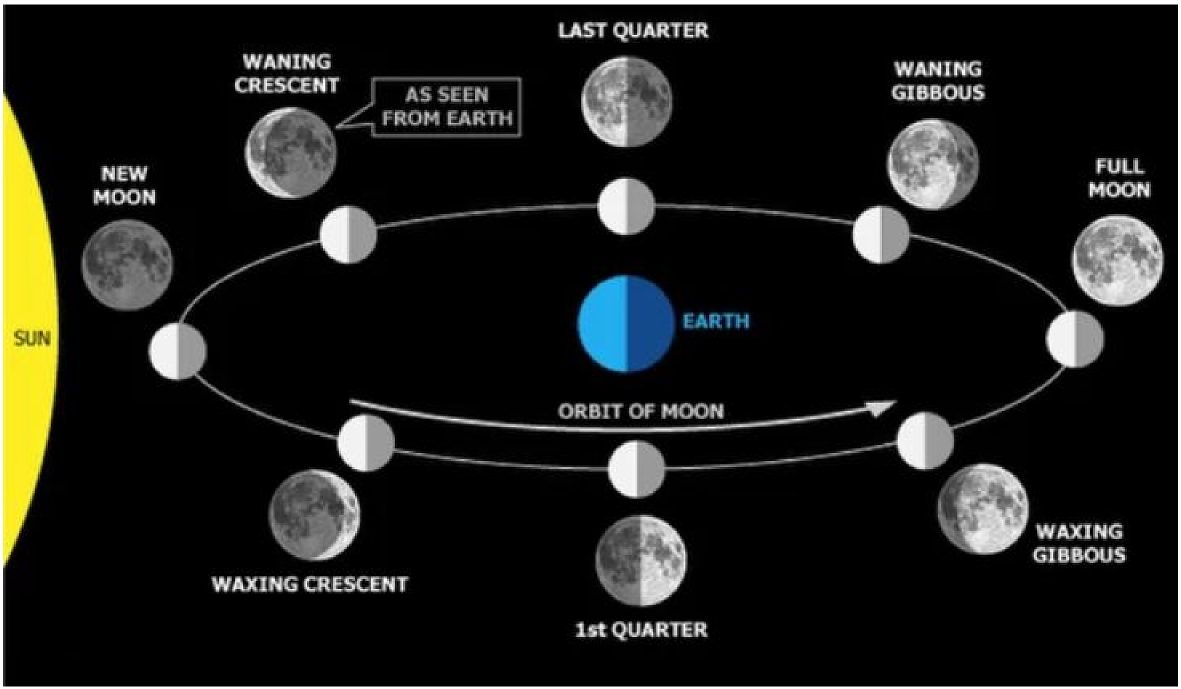
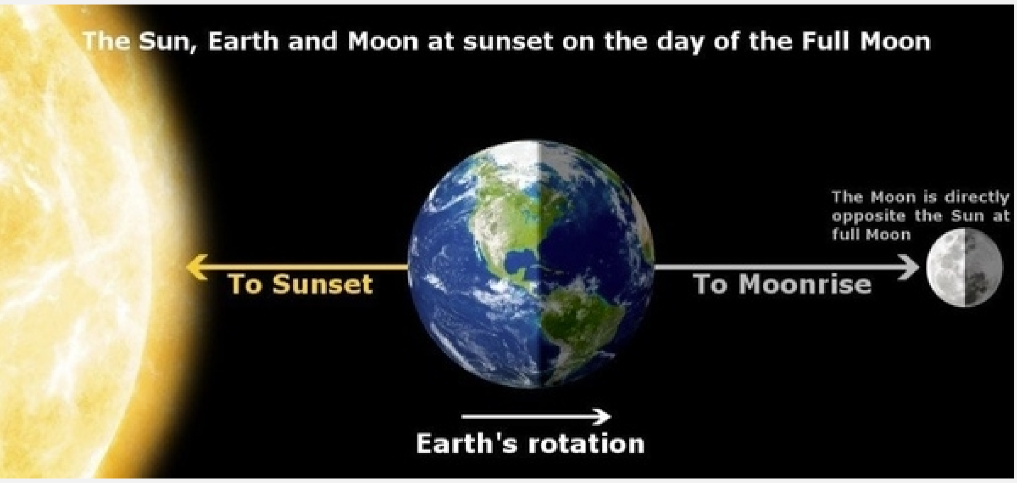

LAST QUARTER
WANING CRESCENT
AS SEEN FROM EARTH
WANING GIBBOUS
NEW MOON
FULL MOON
SUN
EARTH
ORBIT OF MOON
WAXING CRESCENT
WAXING GIBBOUS
1st QUARTER

Figure 12.175: Moon phases (source: ?)

Remark

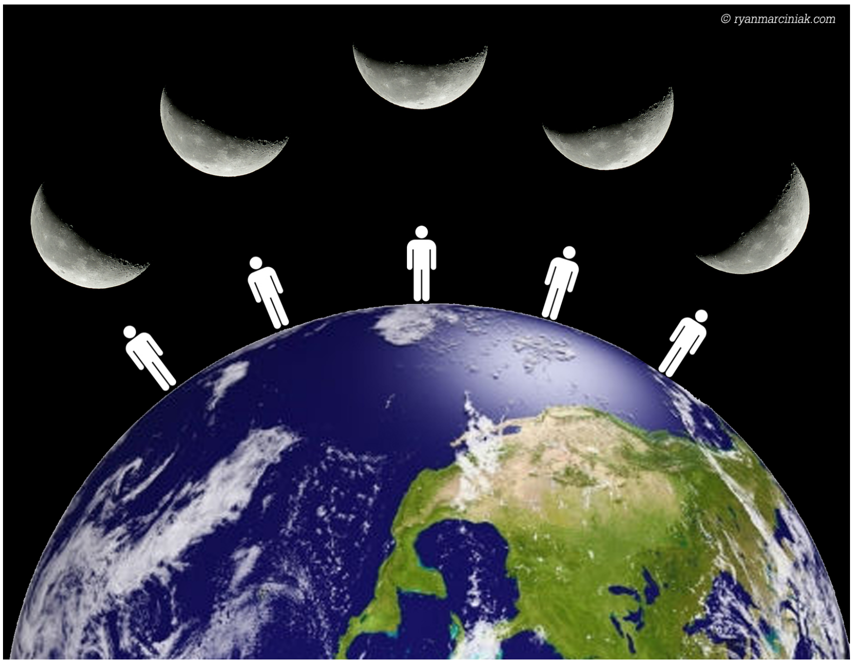
Very rarely, if the conditions are just right, you may also be able to see an actual full Moon very close to the horizon and opposite to the Sun, during sunrise or sunset (a better denomination would be "at daybreak/nightfall"). In general, however, the Moon is always below the horizon while the Sun is up on the date of the full Moon.

If you cannot understand how that works, it would appear you are maybe not trying hard enough or you are being stubbornly obtuse for your own reasons. Indeed, consider the situation below at full Moon:


The Sun, Earth and Moon at sunset on the day of the Full Moon
The Moon is directly opposite the Sun at full Moon
To Sunset
To Moonrise
Earth's rotation

Imagine now you are a human in Miami, Florida. You go outside and look to the West, and you see the Sun is setting, as the Earth is spinning from West to East (the terminator line that separates day from night moves to the west). Then you look to the East and see the full Moon that is rising in the east. It is facing the Sun just like the Earth off to your West is, and we are looking at its sunlit side, thus it is full. Occasionally, the Moon may actually be in the Earth's shadow, and it will turn dimly red for a few hours: a lunar eclipse.

But this figure is omitting something important for earthlings (even if its implicit)... The Moon in the figure above is represented as seen from above the Solar System (perpendicular to the ecliptic plane) or as seen from a viewer at the North Pole. In-between the reality looks like this depending on your latitude:


© ryanmarciniak.com

It is therefore necessary to clarify whether it is a visual or bolometric apparent magnitude. In general, astrophysicists use bolometric magnitudes in their publications.

The reader must also keep in mind that the Moon isn't aligned on the ecliptic as illustrated below:

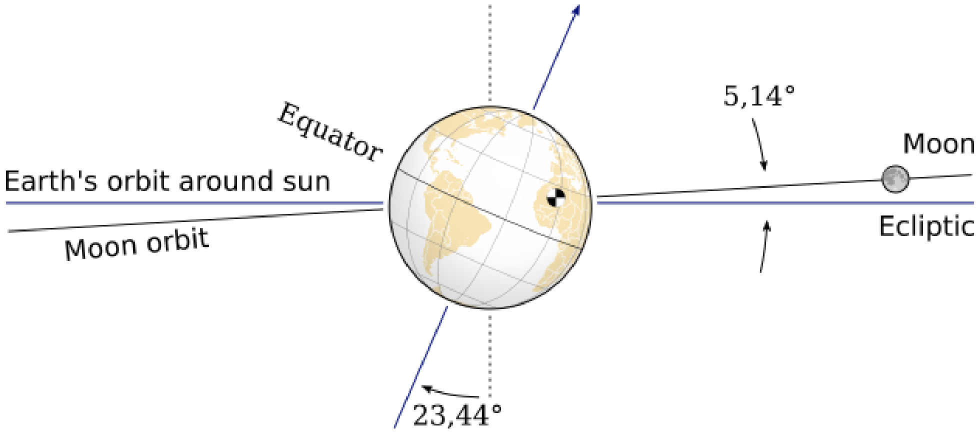

Equator
Earth's orbit around sun
Moon orbit
5,14°
Moon
Ecliptic
23,44°

Figure 12.176: **Moon and ecliptic angle**

But that using the average Moon-Earth distance and basic trigonometry we get that the Moon is:

$$\sin(5.14^\circ)\cdot 384,000 \cong 34,402 \tag{12.2.105}$$

kilometres above (and sometimes below) the ecliptic plane! Keeping in mind that Earth's radius is itself $6,371$ [km]! So the prior-previous figure like many similar figures in most textbooks and Internet websites is quite misleading!

This obliquity is the reason why we observe a Moon "libration", i.e. a cyclic variation in the apparent position of the Moon perceived by Earth-bound observers and caused by changes between the orbital and rotational planes of the moon. It causes an observer to see slightly different hemispheres of the surface at different times. That why over time, slightly more than half (about 59% in total) of the Moon's surface is seen from Earth due to libration and not only 50% as it would be the case without libration!

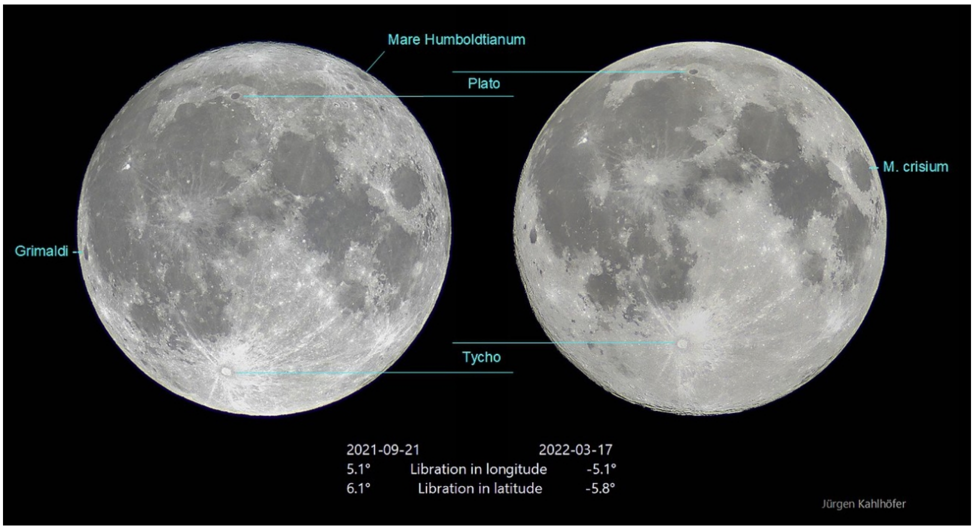
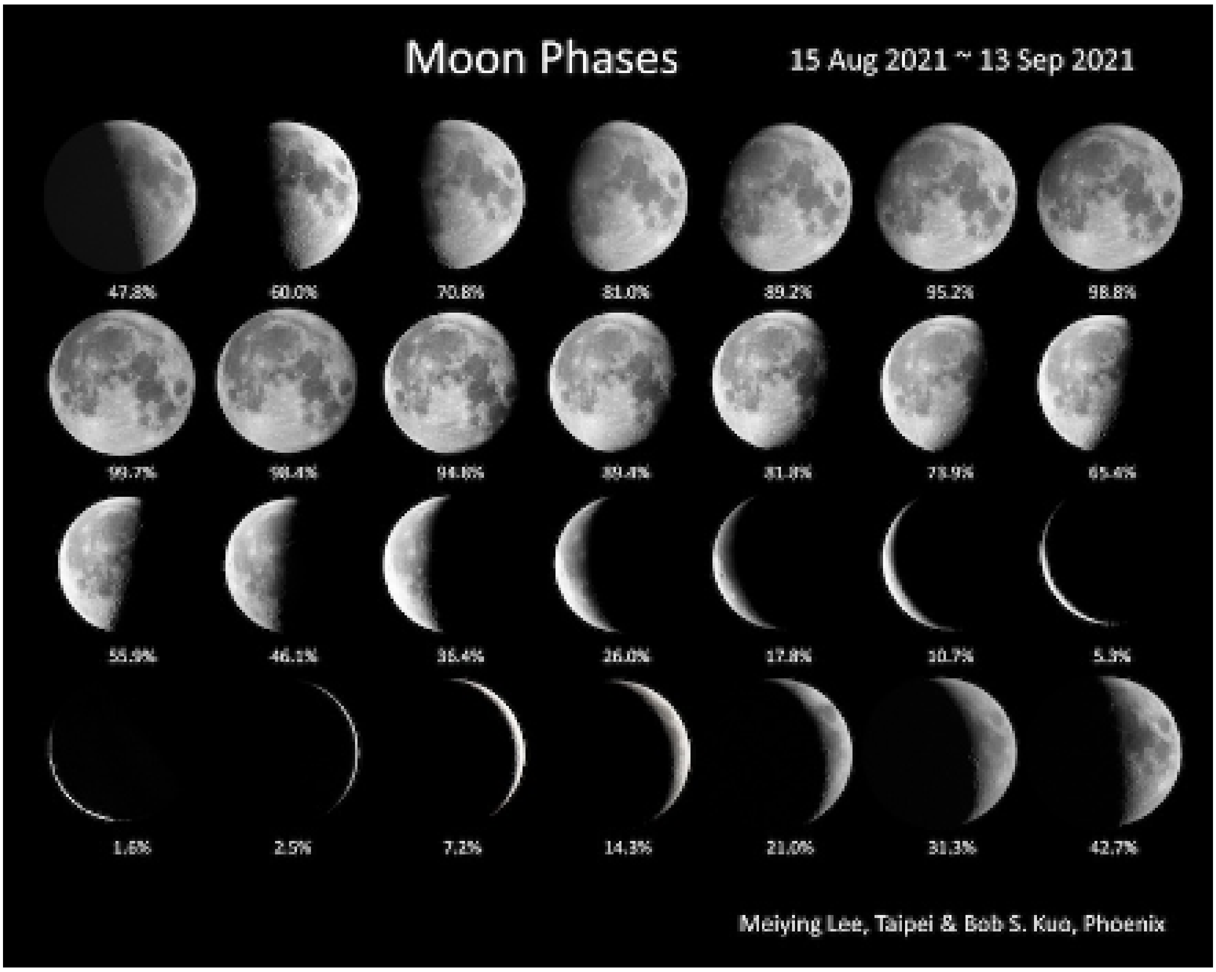

Mare Humboldtianum
Plato
M. crisium
Grimaldi
Tycho
2021-09-21
2022-03-17
5.1°
Libration in longitude
-5.1°
6.1°
Libration in latitude
-5.8°
Jürgen Kahlhöfer

Figure 12.177: **Full Moon at opposite librations (source: Wikipedia, author: Jürgen Kahlhöfer)**

Here is a nice timelapse of the different Moon phases:


Moon Phases
15 Aug 2021 ~ 13 Sep 2021
47.8%
60.0%
70.8%
81.0%
89.2%
95.2%
98.8%
99.7%
98.4%
94.0%
89.4%
81.8%
73.9%
65.4%
55.9%
46.1%
36.4%
26.0%
17.8%
10.7%
5.3%
1.6%
2.5%
7.2%
14.3%
21.0%
31.3%
42.7%
Meiying Lee, Taipei & Bob S. Kuo, Phoenix

12.2.1.1.9 Absolute magnitude

The absolute magnitude M (not to be confused with the notation of emittance seen in the section of Geometrical Optics) of a star is also a logarithmic scale, expressing this time the bolometric luminosity L!!! It is the quantity presented in ordinate of the Hertzsprung-Russell diagram. The scale of this size is based however on the (visual) apparent magnitude.

The apparent magnitude and absolute magnitude are bound by the distance from the star. At constant intrinsic apparent brightness, the apparent brightness therefore decreases obviously with the square of the distance as we have already seen. In order to establish a relation, we had to choose a reference distance by a new definition.

Definition (#12.2.539): The "absolute magnitude" M of a star is equal to its apparent magnitude m if it is distant of 10 parsecs[28] (32.6 light years).

Therefore taking Pogon's formula that is for recall:

$$m_1 - m_2 = 2.5 \log\left(\frac{e_2}{e_1}\right) \tag{12.2.106}$$

---

[28] We take 10 parsecs to calculate absolute magnitude because it is a convenient distance that is large enough to minimize the effects of interstellar extinction (the dimming of light caused by dust and gas between the object and the observer), but small enough to be a reasonable distance to which to compare the brightness of stars and other celestial objects.

And changing the notations to make it corresponds to the previous definition:

$$m-M=2.5\log\left(\frac{e_2}{e_1}\right) \tag{12.2.107}$$

we get:

$$M=m-2.5\log\left(\frac{e_2}{e_1}\right) \tag{12.2.108}$$

And as:

$$\frac{e_2}{e_1}=\frac{\dfrac{P_2^2}{d_2^2}}{\dfrac{P_1^2}{d_1^2}} \tag{12.2.109}$$

But as it is the same star:

$$\frac{e_2}{e_1}=\frac{\dfrac{P_0^2}{d_2^2}}{\dfrac{P_0^2}{d_1^2}}=\frac{d_1^2}{d_2^2} \tag{12.2.110}$$

In our case this becomes:

$$\frac{e_2}{e_1}=\frac{d^2}{10^2} \tag{12.2.111}$$

Therefore:

$$M=m-2.5\log\left(\frac{e_2}{e_1}\right)=m-2.5\log\left(\frac{d^2}{10^2}\right)=m-2.5\left(2\log(d)-2\right) \tag{12.2.112}$$

So finally:

$$\boxed{M=m-5\left(\log(d)-1\right)} \tag{12.2.113}$$

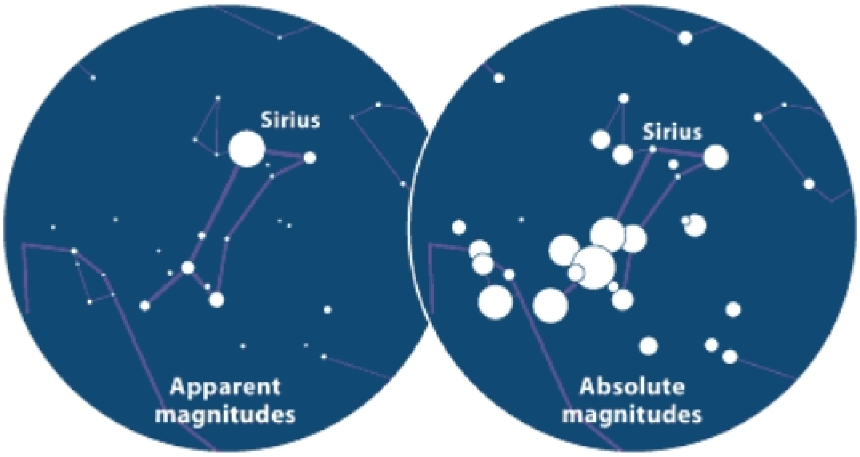


Figure 12.178: **Sirius apparent vs absolute magnitudes**

As the Sun-Earth distance in parsec is equal to $4.84814\cdot 10^{-6}$ we get:

$$m_{\odot}-M_{\odot}=5(\log(4.84814\cdot 10^{-6})-1)=-31.5721 \tag{12.2.114}$$

Therefore:

$$M_{\odot}=m_{\odot}+31.5721=-26.74+31.5721\cong 4.83 \tag{12.2.115}$$

Remark

For the absolute magnitude M to be accurate, we need stellar models, and know the temperature of the star as we will immediately see it. In practice, the only readily accessible quantity is obviously the observed magnitude, which is actually the combination of the apparent magnitude and the interstellar absorption.

The absolute magnitude can be obviously rewritten with respect to the absolute bolometric luminosity of the Sun (actually a G2V-type star, aka yellow dwarf):

$$M=m_{\text{bol},\odot}-2.5\log\left(\frac{L_{\text{bol}}}{L_{\text{bol},\odot}}\right) \tag{12.2.116}$$

We put for the Sun that $L_{\text{bol},\odot}=1$. Therefore it remains:

$$\boxed{M=m_{\text{bol},\odot}-2.5\log(L_{\text{bol}})} \tag{12.2.117}$$

This latter relation of comparison of the absolute magnitude with the apparent magnitude (which is the actually magnitude observed on Earth) allows estimation d of the distance of the object in astrophysics.

Using the expression of the bolometric luminosity proved earlier above:

$$L_{\text{bol}}=\left(\frac{T}{T_{\odot}}\right)^4\left(\frac{r}{r_{\odot}}\right)^2 \tag{12.2.118}$$

the absolute magnitude of star being a direct function of its temperature and radius we can then write:

$$M=m_{\text{bol},\odot}-2.5\log\left(\left(\frac{T}{T_{\odot}}\right)^4\left(\frac{r}{r_{\odot}}\right)^2\right)=m_{\text{bol},\odot}-10\log\left(\frac{T}{T_{\odot}}\right)-5\log\left(\frac{r}{r_{\odot}}\right) \tag{12.2.119}$$

This is the result we wanted to prove from the beginning: the absolute bolometric magnitude is directly related to the bolometric luminosity of the star, which is why it is one that most interests astrophysicists.

Remark

The distance to nearby stars could be determined by the satellite Hipparcos. By measuring the parallax (measurements of the star position at six-month intervals and applying basic trigonometric rules as seen in the section Astronomy). But beyond a few tens of parsecs, measuring the distance of stars by parallax becomes very imprecise. By studying the spectrum of the star, we can determine its spectral class, its surface temperature and place in the Hertzsprung-Russell diagram. It is therefore possible to estimate its absolute magnitude and roughly calculate its distance.

This measurement trick is fundamental to cosmology. It is the way we determines the distance to nearby galaxies by measuring the period of some variable stars (we will focus a little bit on that further below).

The distance of distant galaxies is calculated by measuring the apparent magnitude of supernovae that occur in it. Indeed, the absolute magnitudes of Type Ia supernovae (we recognize them by the lack of hydrogen spectrum lines, and by the decrease in brightness) are well calibrated because the energy released by these stellar explosions is relatively constant.

The stars of the main sequence of the Hertzsprung-Russell diagram are very stable objects. The gravitational force, which tends to contract the star, is exactly compensated by the internal pressure forces, which tend to dilate it. It's when the star becomes a red giant that sometimes the balance is upset. Thus began a phase of instability which results in significant variations in the brightness of the star.

The breaking of balance is caused by a complex phenomenon that involves variations of transparency of helium layers near the surface of the star. From there, the star begins to experience a series of expansions and contractions controlled by the forces that were formerly balance. When the pressure force prevails, the volume of the star increases. But the gravity slows the movement and eventually cause contraction. The volume of the star will pass below its average value, until the internal pressure opposes the contraction and managed to cause further expansion.

It is not the size changes that cause the variations in brightness, but those of the temperature. Indeed, as we have prove it earlier above, the brightness of a star varies with the fourth power of the temperature, while it varies with the square of the radius following for recall:

$$L_{\text{bol}}=\left(\frac{T}{T_{\odot}}\right)^4\left(\frac{r}{r_{\odot}}\right)^2 \tag{12.2.120}$$

When the volume of the star, however, is lower than average, the temperature is slightly higher and the brightness maximum. At the opposite, the temperature is slightly lower than average and the brightness minimum. The brightness of the star thus changes periodically, hence the name of "variable star" or "pulsative variable star".

It exists in the Hertzsprung-Russell diagram of a band of instability that crosses this diagram almost vertically just to produce the thermal phenomena in question.

The two main types of pulsating variables are the Cepheids and RR Lyrae stars. These bodies play a central role in astrophysics. Cepheids are stars of a few solar masses. They are in the helium burning phase after reaching the red giant stage. The stars of solar mass arrived at this point become RR-Lyrae stars. Their brightness varies with a period of between one day and several weeks. The remarkable property of Cepheids is the existence of a relation between the average brightness and the period of their oscillations. For example, the average brightness is $1,000$ times that of the Sun for a period of days and $10,000$ times that amount for a period of several weeks. It is this relation that makes Cepheids one of the basic tools of astrophysics.

Ok now that we know a few main concepts, it is time to introduce the following interesting data table about a few stars:

Star Name	**Latin Name**	m	**Distance [LY]**	**Spectral Type**	L **[L$_\odot$]**	T **[K]**	M **[M$_\odot$]**	R **[R$_\odot$]**
Sun	Sol	-26.9	0.000016	G2 V	1	$5,800$	1	1
Sirius	α Canis Majoris	-1.46	8.6	A1 V	25.4	$10,000$	2.5	2.2
Canopus	α Carinae	-0.72	75	F0 II	$1,200$	$8,000$	10	15
Arcturus	α Bootis	-0.04	34	K1 IIIb	90	$4,800$	3	15
Rigil Kent	α1 Centauri	-0.01	4.3	G2 V	1.4	$5,500$	1	0.9
Véga	α Lyrae	0.03	25	A0 Va	40	$10,500$	2.5	1.7
Capella	α Aurigae	0.08	41	G5 III + G0 III	120	$5,000$	3	8
Rigel	β Orionis	0.12	630	BS Ia	$55,000$	$12,000$	50	38
Procyon	α Canis Minoris	0.38	11	F5 IV-V	7	$7,000$	1.5	1.4
Achernar	α Eridani	0.46	130	B3 V	600	$18,500$	8	1.9
Betelgeuse	α Orionis	0.5	≤ 420	M1-2 Ia-Iab	$9,000$	$3,000$	30	$1,800$
Hadar	β Centauri	0.61	≤ 300	B1 III	$3,900$	$21,500$	20	4.3
Altair	α Aquilae	0.77	16	A7 V	10	$8,000$	2	1.5
Aldebaran	α Tauri	0.85	55	K5 III	110	$3,500$	4	30
Antares	α Scorpii	0.96	≤ 500	M1,5 Iab Ib	$\leq 8,800$	$3,000$	25	400
Spica	β Vergini	≤ 300	35	B1 III-IV+B2 V	$3,000$	$21,500$	18	4
Pollux	β Gemini	1.14	35	K0 IIIb	35	$5,000$	3	11

Table 12.18: **Some stars data table**

It seems that a significant percentage of practitioners (and students!) agree that all these different concepts related to a light source are quickly a mess! We have tried below to summarize all the concepts introduced earlier above (sadly we never saw such a table in other textbooks):

Name(s)	**Definition**	**Formulation**	**Units** (SI)
Brightness (intrinsic bolometric brightness)	Total power radiated in the entire electromagnetic spectrum in the direction of the observer expressed relatively to the total power radiated by the Sun	$L_{\text{bo}}=\dfrac{P_{\text{star}}}{P_\odot}$	None
Illuminance (semi-bolometric illuminance)	Measure of how much the incident light illuminates the surface, wavelength-weighted by the luminosity function to correlate with human brightness perception	$E_\nu=K_m\int\limits_{380}^{780}M(\lambda,T)V(\lambda)\mathrm{d}\lambda$	$[\mathrm{J}\cdot\mathrm{s}^{-1}]$
Shining (apparent brightness) (flux) (brilliance)	Total power radiated in the direction of the observer divided by the sphere surface with the radius equal to the distance which separates the observer from the source	$b=\dfrac{P_{\text{star}}}{S_{\text{obs.}}}$	$[\mathrm{W}\cdot\mathrm{m}^{-2}]$
Emittance (luminous emittance) (irradiance) (emissive power) (radiant emittance)	Radiant flux emitted by a surface by unit of its area	$M_e=\dfrac{\Phi}{S}=\varepsilon\sigma T^4$	$[\mathrm{W}\cdot\mathrm{m}^{-2}]$
Relative Emittance (thermal emittance) (emmisivity)	Ratio of the total emissive power of a body to the total emissive power of a perfectly black body at that temperature	$\varepsilon=\dfrac{M_e}{M_e^\circ}=\dfrac{\varepsilon\sigma T^4}{\sigma T^4}$	None
Spectral Emittance (spectral exitance)	Radiant flux emitted by a surface by unit area per unit frequency (or per unit wavelenght)	$M_\nu=\dfrac{\partial M}{\partial\nu}$	$[\mathrm{W}\cdot\mathrm{m}^{-2}\cdot\mathrm{s}]$
Luminosity	Luminosity is the total amount of electromagnetic energy emitted per unit of time by an object.	L	[W]
Apparent magnitude	Empirical scale to compare the brilliance of an object relatively to another one (commonly Vega α Lyr is taken as zero reference)	$m=2.5\log_{10}\left(\dfrac{L}{L_{\text{ref}}}\right)$	None
Absolute magnitude	The absolute magnitude M_a of a star is the magnitude m the star would have if it was placed at a distance of 10 parsecs from Earth.	$M_a=m-5\left(\log_{10}(d)-1\right)$	None

Table 12.19: **Summary of most common visibility criteria in astrophysics with their traditional notation and units according to international system (SI)**

12.2.1.1.9.1 Why are stars not visible on most space photos and videos

In this early 121st century (holocene calendar), the intrinsic luminosity of Sirius is 25.4 times greater than that of the Sun (actually a G2V-type star, aka yellow dwarf). However, Sirius is about 8.6 light-years away from Earth (or $d = 8.13 \cdot 10^{16}$ meters), while our Sun is just 8 light-minutes. Now, since the apparent brightness is affected by geometric dilution (it falls off as the inverse-square of the distance as we know), Sirius appears to us much fainter than the Sun! As the luminosity of the Sun is $3.828 \cdot 10^{26}$ [W], we then have for Sirius:

$$b_{\text{Sirius}} = \frac{L_{\text{Sirius}}}{4\pi d^2} = \frac{25.4 \cdot 3.828 \cdot 10^{26}}{4\pi \left(8.13 \cdot 10^{16}\right)^2} = 1.35 \cdot 10^{-7}\ [\mathrm{W} \cdot \mathrm{m}^{-2}] \qquad (12.2.121)$$

If we consider the distance Moon-Earth as negligible in comparison of Sirius-Earth distance we can consider that the same flux reaches the Moon in first approximation!

If the reader wants to redo the calculations with other stars here is a small list:

Name of star	Distance/light-years	Luminosity
Sun	$1.58 \cdot 10^{-5}$	1.0
Sirius	8.6	25
Cancpus	310	1100
Alpha Centauri	4.4	1.5
Arcturus	37	170
Vega	25	40

Table 12.20: **Some stars, distances and luminosities**

The astronomer Robert Trumpler showed in 11930 (holocene calendar) that interstellar gas and dust absorbs light at a typical rate of 2 magnitudes (one magnitude being for reminder a factor of 2.512) per kiloparsec ($3,262$ light-years). Sirius being at only 8.6 light years we then have:

$$b_{\text{Sirius}} = 1.35 \cdot 10^{-7} \left(1 - \frac{8.6}{3,262} \cdot 2.512\right) = 1.34 \cdot 10^{-7}\ [\mathrm{W} \cdot \mathrm{m}^{-2}] \qquad (12.2.122)$$

We can consider Sirius as a black body at $25,200$ [K] according to the latest available measurement, therefore using Planck's distribution function (see section Thermodynamics page 3724), the visible light corresponds to:

$$\text{Prc. visible} = \frac{M_{\text{tot.}}(T)}{M_{\text{vis.}}(T)} = \frac{2\pi \frac{h}{c^2} \int\limits_{400\cdot 10^{12}}^{700\cdot 10^{12}} \frac{\nu^3}{e^{\frac{h\nu}{kT}} - 1} \mathrm{d}\nu}{2\pi \frac{h}{c^2} \int\limits_{0}^{+\infty} \frac{\nu^3}{e^{\frac{h\nu}{kT}} - 1} \mathrm{d}\nu} = \frac{\int\limits_{400\cdot 10^{12}}^{700\cdot 10^{12}} \frac{\nu^3}{e^{\frac{h\nu}{kT}} - 1} \mathrm{d}\nu}{\int\limits_{0}^{+\infty} \frac{\nu^3}{e^{\frac{h\nu}{kT}} - 1} \mathrm{d}\nu} \cong 5.45\% \qquad (12.2.123)$$

Then:

$$b_{\text{Sirius,vis.}} \cong 5.45\% \cdot 1.34 \cdot 10^{-7} = 7.31 \cdot 10^{-9}\ [\mathrm{W} \cdot \mathrm{m}^{-2}] \qquad (12.2.124)$$

Assuming the average photon energy in visible light, i.e. $3.61 \cdot 10^{-19}$ [J], we get therefore:

$$N = \frac{7.31 \cdot 10^{-9}}{3.61 \cdot 10^{-19}} \cong 2.02 \cdot 10^{10}\ [\text{photons} \cdot \mathrm{s}^{-1} \cdot \mathrm{m}^{-2}] \qquad (12.2.125)$$

visible photons received by Sirius on Moon by square meter and per second.

If we consider a camera like the one taken by Apollo mission on the Moon in 11969 (according to holocence calendar), i.e. Hasselblad 500EL Data with 80mm Zeiss lens (lens having a global transmissivity of 98%) f/11 with a Kodak Panatomic-X 70mm ISO 160 stereo film with shutter speed of $1/250$ [s], then the number of visible photons received from Sirius on such an apparatus during a photo capture is approximately of[29]:

$$n_{\text{Hasselbad}}=\frac{2.02\cdot 10^{10}}{250}\pi\left(\frac{0.08}{2\cdot 11}\right)^2\cdot 98\%\cong 3,300\;[\text{photons}] \tag{12.2.126}$$

So that's kinda the number of photons received for each square millimetre of the Kodak film during a shoot.

Remark

For a modern CCD camera of the early 121st century (holocene calendar) recording a 4K movie at a frame rate (fps) of 24 per seconds with the same aperture and focal and transmissivity, we get:

$$n_{\text{CC}}=\frac{2.02\cdot 10^{10}}{24}\pi\left(\frac{0.08}{2\cdot 11}\right)^2\cdot 98\%\cong 34,375\;[\text{photons}]$$

As far as we know Kodak films were made of a cellulose acetate film base (kind of gelatine containing dispersed silver halide (AgX[30] crystals but in practice it's mostly AgBr) that is strongly reactive to light) and according to some sources, the number of photons necessary to activate silver halide back to it's metallic state is between 7 and 30 photons[31] (we will take 18 as an average). Furthermore, the minimum number of silver halide it takes inside a cluster to form a latent image (usable for photo development and visible at human eye) in a grain is between[32] 5 and 50 (we will take 27 as an average)! However the scattering of the electrons of silver halide (or whatever) by the incoming light isn't 100%. We then speak of "quantum efficiency" η_q (i.e. the ratio of the number of incoming photons over the number of electrons which they produce)! Sadly we cannot found the quantum efficiency of the exact Kodak film used by the Apollo camera. We were just able to find an upper bound for Kodak Royal-X films in 430 [nm] wavelength of $\eta_q=0.073\%$ and in 607 [nm] of $\eta_q=0.59\%$ from a paper of 11958 according to holocene calendar (see [176]). Another paper of 12001 according to holocene calendar (see [373]) claim that in practice, the quantum efficiency of photographic systems is often only about $\eta_q=20\%$. An internal paper at NASA of 11964 according to holocene calendar (see [248]) claim the quantum efficiency of the best photographic emulsions is of $\eta_q=0.1\%$. Finally and electric engineer claim on an Internet forum in 12018 according to holocene calendar that the quantum efficiency is of $\eta_q=5\%$... So let us take the average of these number. That leads us to the following number of efficient photons:

$$n_{\text{Hasselbad,eff.}}=3,300\cdot\bar{\eta}_q=3,300\cdot 5.15\%\cong 170\;[\text{photons}] \tag{12.2.127}$$

---

[29] The Apollo crew had different model of cameras with different lenses and different stereo films! Not just one!

[30] There are about $100,000$ silver halide crystals per square millimetre photographic image, each crystal is often 0.03 to 20 microns in size depending on the ISO of the film!

[31] For more information see *Silver Halides in Photography* from Yen. T. Tan

[32] This depends on the makeup of the emulsion, hence different film sensitivities!

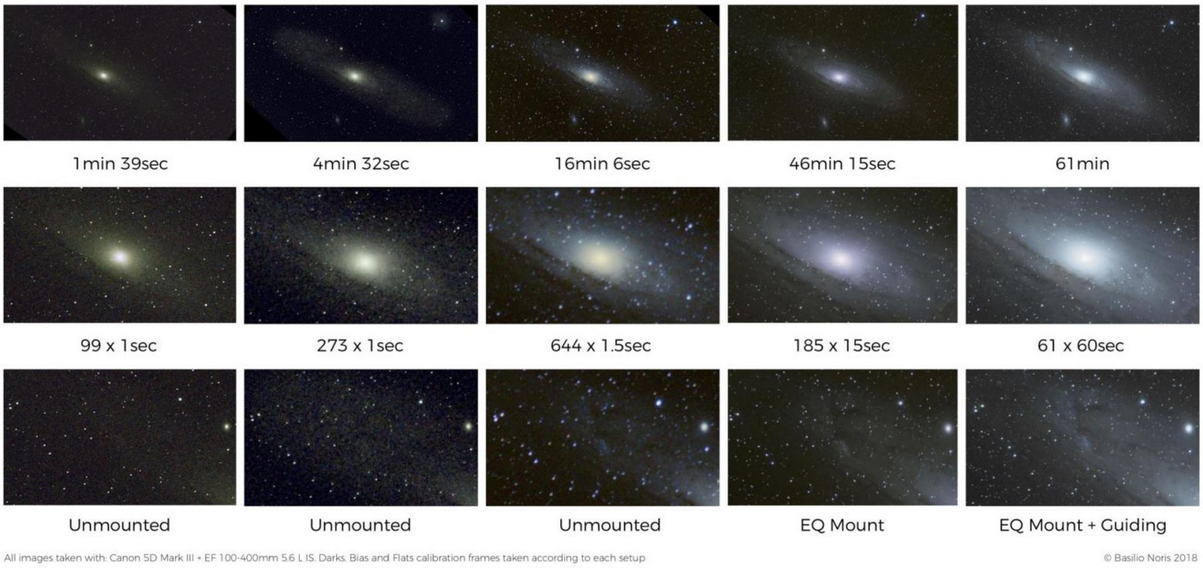

1min 39sec
4min 32sec
16min 6sec
46min 15sec
61min
99 x 1sec
273 x 1sec
644 x 1.5sec
185 x 15sec
61 x 60sec
Unmounted
Unmounted
Unmounted
EQ Mount
EQ Mount + Guiding
All images taken with: Canon 5D Mark III + EF 100-400mm 5.6 L IS. Darks, Bias and Flats calibration frames taken according to each setup
© Basilio Noris 2018

Figure 12.179: Impact of integration time on detail and background noise (credit and authorization: Basilio Noris)

Or here is another one:

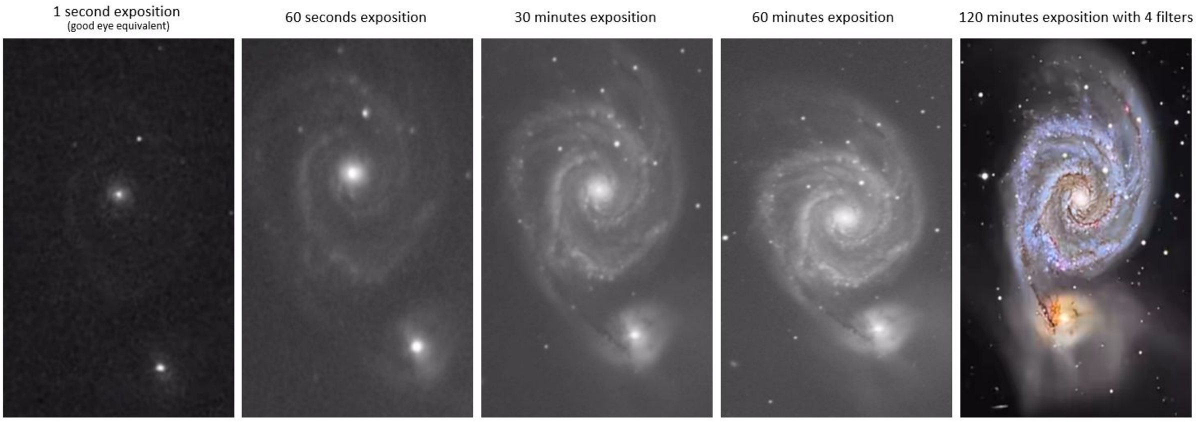

1 second exposition
(good eye equivalent)
60 seconds exposition
30 minutes exposition
60 minutes exposition
120 minutes exposition with 4 filters

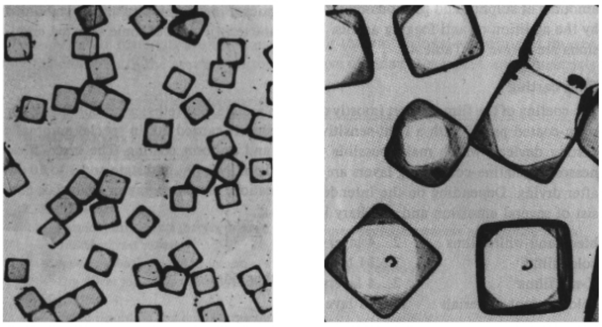
Figure 12.180: Impact of integration time on details and background noise of M51 on an Astrograph 304/1200 (credit: Anothony Salsi)

Remark

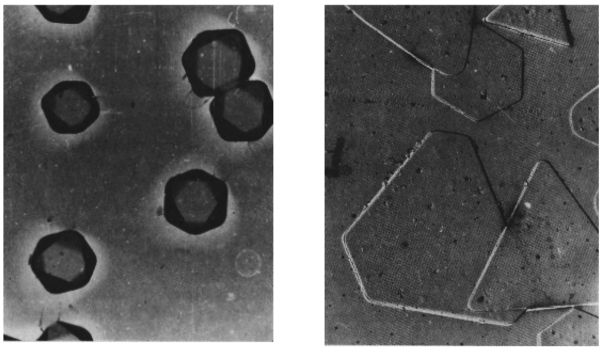
For a modern CCD camera of the early 121st century (holocene calendar) we have on average on the global visible spectrum

$$n_{\text{CCD,eff.}} = 34,375 \cdot \bar{\eta}_q = 34,375 \cdot 70\% \cong 24,000\,[\text{photons}]$$

The size of AgX grains (for information, the higher a film stock's ISO is, the bigger the silver crystals are, the ISO 100 series are around 6 microns) determines the chemical amplification factor during the image development (the gain is about 10^8). The larger the grains the larger the amplification factor and the more sensitive the photographic film. Large grains, however, obviously decrease the spatial resolution of the picture and imply an increased thickness of the photographic emulsion.



AgBr cubes tetrahedrons



octahedrons tabular grains

Figure 12.181: **Various** AgX **grain morphology (i.e. "film grain")**

Therefore we have:

$$n_{\text{Kodak,halide}} = \frac{170}{18 \cdot 27} \cong 0.35\,[\text{grains}] \tag{12.2.128}$$

So the number of photons and the quantum efficiency is not even enough to activate on silver halide!

Remark

For a modern CCD camera of the early 121st century (holocene calendar) the number of photons necessary to activate one pixel is of the order of 10^5 (see page [400]). Hence:

$$n_{\text{CCD,pixel}} = \frac{24,000}{10^5} \cong 0.24\,[\text{pixel}]$$

So not even enough to activate one pixel! So the conclusion doesn't change for modern CCD cameras recording 4K movies in space with the same kind of settings!

Furthermore... (!!!) ... once we have the reaction of silver halide molecule the reader has to know that the latter are not completely stable. The new reaction can disintegrate after a few minutes from interactions with chemicals from the emulsion dyes, oxygen and water molecules (for more see [122] and [49]). After a few minutes you start to loose reactive silver halide. If we have tenth of thousands of them then we can afford to loose a few. If we have only a few of them (here in our calculations we have even less than one...) then the odds of them surviving the film development is low!

So now we know why we can see some stars or not at all the photos of the Apollo mission but the crew was able to see the stars as the human eye needs only between 7 and 30 photons to send a signal to the brain.

Remark

The reader must keep in mind that the pictures are taken in lunar daytime - the sky is black due to absence of air, not because it is night. Our brains want to see stars in a black sky. The astronauts could see the stars, but the star's light output could not compete with the Sun-blasted lunar surface in a camera exposure. You could try shooting the ground and sky at a lit sport field at night, the same thing would happen!

Here below we can see that the contrast makes that the stars in the background are not visible:

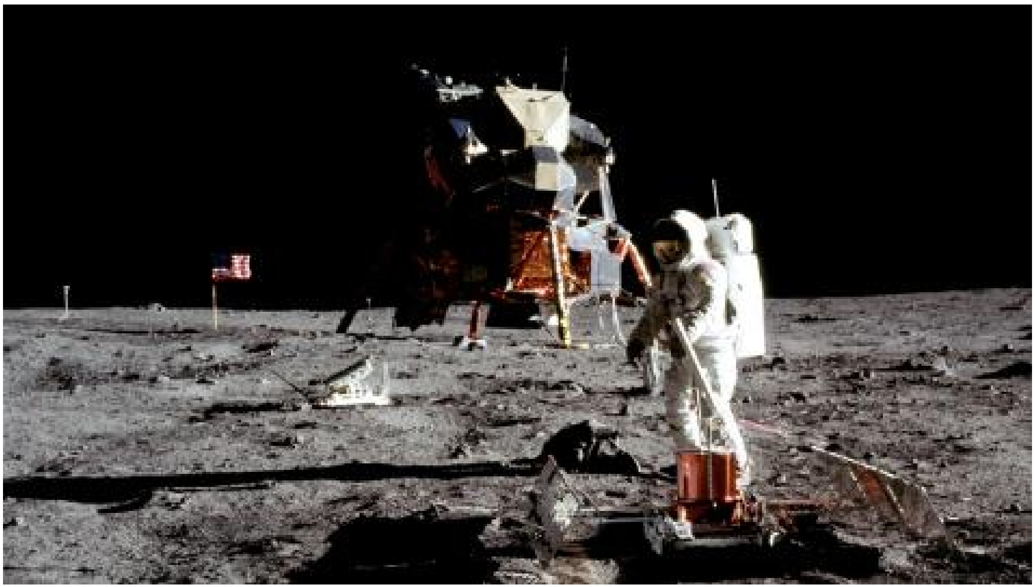


Figure 12.182: **21 July 11969 Appolo 11 mission photo of Buzz Aldrin taken by Neil Armstrong (credit: NASA)**

But here in the same mission (stage separation) the contrast and shutter speed is no too high so we can see some stars in the background:

Or in this quite relatively high contrast photo still from the same mission we can see some stars in the background as the shutter speed wasn't very likely to $1/250$ [s]:

For the human eye, having an average exposure time of $1/15$th of a second and the cornea a radius of 7.4 millimetre with a f/2 at rest, we get in vacuum (i.e. still on the Moon):

$$n_{\text{Hum. eye}} \cong \frac{1.87\cdot 10^{10}}{15}\pi\left(\frac{0.0074}{2}\right)^2 \cong 53,000\,[\text{photons}]$$

Even on Earth where 43% on visible light is in average absorbed by the atmosphere, we till get:

$$n_{\text{Hum. eye,Earth}} \cong 43\%\cdot 53,000 \cong 22,790\,[\text{photons}]$$

Knowing that the eye needs between 7 to 30 photons to consider an object as visible we understand why it's not a problem to see Sirius in the night sky with naked eyes. Obviously you can't see Sirius in the day sky as the background luminosity (blue of the atmosphere) is much brighter than the brightness of Sirius itself. This is related to the concept of "Signal to Noise ratio"!

12.2.1.1.10 Eddington limit

Radiation exerts a pressure. This implies that in luminous systems there is an outwards force by the outgoing luminosity, in addition to the gravitational force pulling[33] the gas inwards. In the Sun, this radiative force is more than 10,000 times smaller than the gravitational pull, but in more massive stars, this outward radiative force can approach the force of gravity.

---

[33] Keep in mind that's in the framework of the Newtonian model of gravity as we that in General Relativity there are no "pull".

Under the classical picture, one cannot surpass the Eddington limit while keeping the system in steady state, because nothing could balance the net force outwards, and gas will necessarily be accelerated outwards, thereby "evaporating" the system. In a supernova explosion, for example, the large fluxes accelerate gas outwards and the systems are quite far from a steady state, so the limit does not apply. But in systems which remain intact for durations much longer than their dynamical time scale, the limit should be relevant, at least according to the classical picture.

Therefore the "Eddington luminosity", also referred to as the "Eddington limit", is the maximum luminosity a body (such as a star) can achieve when there is balance between the force of radiation acting outward and the gravitational force acting inward. The state of balance is named "hydrostatic equilibrium". When a star exceeds the Eddington luminosity, it will initiate a very intense radiation-driven stellar wind from its outer layers. The Eddington limit is invoked to explain the observed luminosity of accrediting Black Holes such as quasars. This relates to the fact that in the 120th century (holocene calendar) gravity was considered inadequate to explain the Sun (actually a G2V-type star, aka yellow dwarf) power, so physicists moved to the nuclear reactions explanations. However nuclear reaction were inadequate to explain powerful objects in the Universe like quasars. So the accretion onto compact objects is a powerful mechanism for producing high energy radiation.

Originally, Arthur Eddington took only the electron scattering into account when calculating this limit, something that now is named the "classical Eddington limit". Nowadays, the "modified Eddington limit" also counts on other radiation processes such as bound-free and free-free radiation interactions.

The limit is obtained by setting the outward radiation pressure equal to the inward gravitational force. Both forces decrease by inverse square laws as we will prove it, so once equality is reached, the hydrodynamic flow is the same throughout the star.

The radiation force equals the rate at which the electron absorbs the photons momentum:

$$F_{\text{rad}}=\frac{\Delta p}{\Delta t}=\frac{\Delta N_\gamma}{\Delta t}p_\gamma \tag{12.2.129}$$

We know that we have for a single photon momentum:

$$p_\gamma=\frac{h\nu}{c} \tag{12.2.130}$$

Therefore:

$$F_{\text{rad}}=\frac{\Delta N_\gamma}{\Delta t}p_\gamma=\frac{\Delta N_\gamma}{\Delta t}\frac{h\nu}{c} \tag{12.2.131}$$

But we can use the bolometric intrinsic luminosity (see section Astronomy page 5336), and under the assumptions of Thomson scattering the fraction of photons scattered at distance r given by (see section Atomistic page 5100):

$$f_T=\frac{\sigma_T}{4\pi r^2} \tag{12.2.132}$$

(do not confuse the notation with that of frequency...!) such that:

$$\frac{\Delta N_\gamma}{\Delta t}=\frac{L}{h\nu}f_T=\frac{L}{hv}\frac{\sigma_T}{4\pi r^2} \tag{12.2.133}$$

Therefore:

$$F_{\text{rad}}=\frac{\Delta N_\gamma}{\Delta t}\frac{h\nu}{c}=\frac{L}{hv}\frac{\sigma_T}{4\pi r^2}f_T\frac{h\nu}{c}=\frac{L\sigma_T}{4\pi cr^2} \qquad (12.2.134)$$

In the limit case when the gravitational force is balanced by the radiation pressure force:

$$F_{\text{grav}}=F_{\text{rad}} \qquad (12.2.135)$$

Accreting material being mainly fully ionised hydrogen we have:

$$F_{\text{grav}}=\frac{GM}{r^2}(m_p+m_e)\cong\frac{GM}{r^2}m_p \qquad (12.2.136)$$

Therefore:

$$F_{\text{grav}}=F_{\text{rad}}\rightarrow\frac{GM}{r^2}m_p\cong\frac{L\sigma_T}{4\pi cr^2} \qquad (12.2.137)$$

Therefore for pure ionized hydrogen under the assumption of Thomson scattering and in hydrostatic spherical equilibrium we get the maximum luminosity due to accretion:

$$\boxed{L_{\text{Edd}}\cong\frac{4\pi GMm_pc}{\sigma_T}} \qquad (12.2.138)$$

Notice that the only parameter in this formula is the mass of the star. All other numbers are constants of nature! That latter is sometimes denoted:

$$L_{\text{Edd}}\cong\frac{4\pi GMm_pc}{\sigma_{\text{T}}}\cong 1.26\cdot 10^{31}\left(\frac{M}{M_\odot}\right)\;[\text{W}] \qquad (12.2.139)$$

where $M_\odot$ is the mass of the Sun and $L_\odot$ is the luminosity of the Sun (actually a G2V-type star, aka yellow dwarf). Then it is immediate that for the special case of the Sun, we get immediately:

$$L_{\text{Edd}}\cong 1.26\cdot 10^{31}\;[\text{W}] \qquad (12.2.140)$$

The reader must really keep in mind that the Eddington limit is not a strict limit on the luminosity of a stellar object! The limit does not consider several potentially important factors, and super-Eddington objects have been observed!

Now we set the gravitational energy of an accreted proton-electron pair to be equal to the total thermal energy of the two particles:

$$\frac{GM}{R}(m_p+m_e)\cong\frac{GM}{R}m_p=2\cdot\frac{3}{2}kT \qquad (12.2.141)$$

Therefore the temperature that the accreted material would reach if the gravitational energy is turned entirely in thermal energy is given by:

$$T_{\text{acc}}\cong\frac{GMm_p}{3kR} \qquad (12.2.142)$$

An application for a White Dwarf of radius $R=10,000$ [km] and $M=1M_\odot$ gives approximately $T_{\text{acc}}\cong 5\cdot 10^5$ [K]. An application for a Neutron star of radius $R=10$ [km] and $M=1.4M_\odot$ gives approximately $T_{\text{acc}}\cong 8\cdot 10^{11}$ [K].

For Black Holes we have something interesting... We have proved at page 5849 that the innermost stable circular orbit (ISCO) for a non-rotating Black Hole was:

$$r_{\text{ISCO}} = \frac{6GM}{c^2} \tag{12.2.143}$$

Therefore:

$$T_{\text{acc}} \cong \frac{GMm_p}{3kR} = \frac{GMm_p}{3kr_{\text{ISCO}}} = \frac{c^2 m_p}{18k} \cong 6 \cdot 10^{11}\,[\text{K}] \tag{12.2.144}$$

So that the Eddington temperature is in first approximation independent of the non-rotating Black Hole mass itself! A kind of Universal constant...

The implications of Eddington limit are multiple in Astrophysics. It is useful to mention only some of them:

- If it is possible to measure (or derive) the Eddington luminosity of a star, it is possible to determine its mass.
- It can fix an upper limit also on the temperature of a star and of the accretion matter.

Remark

It is interesting to compare:

$$L_{\text{Edd}} \cong \frac{4\pi GMm_p c}{\sigma_T} \cong 1.26 \cdot 10^{31} \left(\frac{M}{M_\odot}\right)$$

with the intrinsic bolometric luminosity:

$$L_{\text{star}} = 4\pi R_{\text{star}}^2 \sigma T_{\text{star}}^4 \cong 7.13 \cdot 10^{-7} R_{\text{star}}^2 T_{\text{star}}^4$$

In the huge majority of cases (if we plug in the two relations above the star having the biggest radius, the biggest mass and the hottest temperature):

$$L_{\text{star}} < L_{\text{Edd}}$$

In fact, this is only a rough estimate. It assumes that the accreting material is ionized hydrogen (a good assumption) and that the celestial object (star or Black Hole) is accreting uniformly from all directions (which is not a good assumption). The Eddington luminosity may be exceeded, for example, if accretion occurs primarily from one direction and the resulting radiation emerges in a different direction. Nonetheless, it is a useful approximation that we will use again during our study of Black Holes later (see page 5835).

12.2.1.2 Pulsative Variable Stars

The stars of the main sequence of the Hertzsprung-Russell diagram are very stable objects. The gravitational force, which tends to contract the star, is exactly compensated by the internal pressure forces, which tend to dilate it. It's when the star becomes a red giant that sometimes the balance is upset. Thus began a phase of instability which results in significant variations in the brightness of the star.

The breaking of balance is caused by a complex phenomenon that involves variations of transparency of helium layers near the surface of the star. From there, the star begins to experience a series of expansions and contractions controlled by the forces that were formerly balance. When the pressure force prevails, the volume of the star increases. But the gravity slows the movement and eventually cause contraction. The volume of the star will pass below its average value, until the internal pressure opposes the contraction and managed to cause further expansion.

Remark

The pulsating variable star that is the prototype for one of the most influential classes of variable stars in astronomy was discovered in 11784 (holocene calendar) by John Goodricke. He was the first to observe the periodic and asymmetric light variation of δ Cephei2 in the same year as he discovered the eclipsing binary star β Lyrae. The first known variable star was identified in 11638 (holocene calendar) when Johannes Holwarda noticed that Omicron Ceti (later named Mira) pulsated in a cycle taking 11 months. Interestingly, this star had previously been described as a nova by David Fabricius in 11596 (holocene calendar). The discovery of variable stars, along with observations of supernovae in 11572 and 11604 (holocene calendar), challenged the ancient belief that the starry sky was eternally unchanging.

It is not the size changes that cause the variations in brightness, but those of the temperature. Indeed, as we have prove it earlier above, the brightness of a star varies with the fourth power of the temperature, while it varies with the square of the radius following for recall:

$$L_{\text{bol}}=\left(\frac{T}{T_{\odot}}\right)^4\left(\frac{r}{r_{\odot}}\right)^2 \tag{12.2.145}$$

When the volume of the star, however, is lower than average, the temperature is slightly higher and the brightness maximum. At the opposite, the temperature is slightly lower than average and the brightness minimum. The brightness of the star thus changes periodically, hence the name of "variable star" or "pulsative variable star".

It exists in the Hertzsprung-Russell diagram of a band of instability that crosses this diagram almost vertically just to produce the thermal phenomena in question.

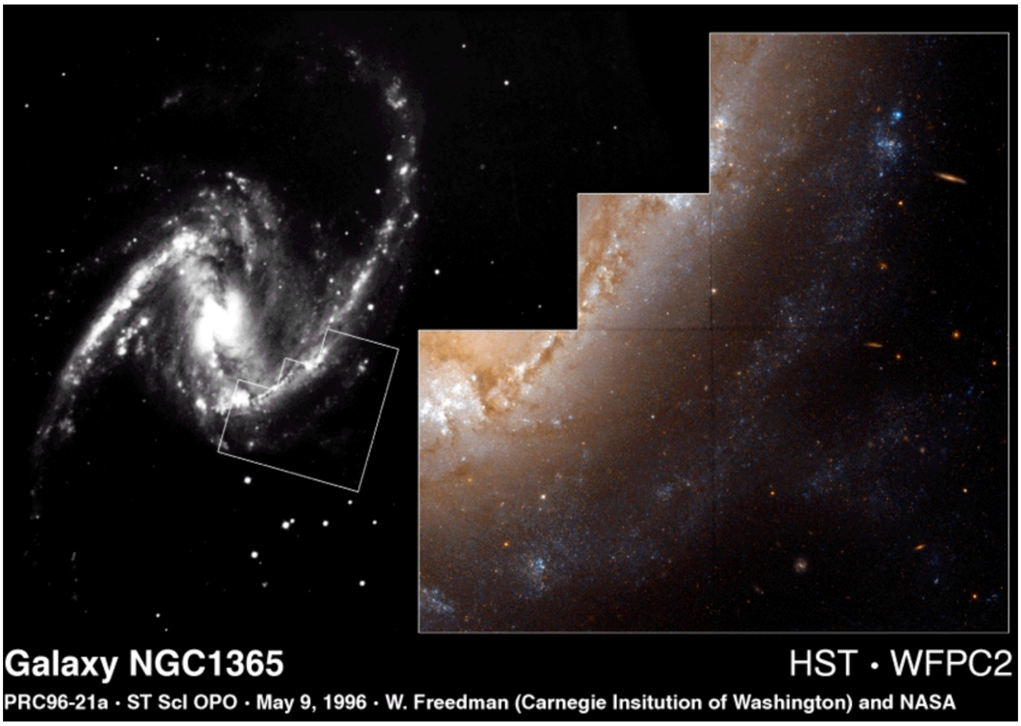
The two main types of pulsating variables are the Cepheids and RR Lyrae stars. These bodies play a central role in astrophysics. Cepheids are stars of a few solar masses. They are in the helium burning phase after reaching the red giant stage. The stars of solar mass arrived at this point become RR-Lyrae stars. Their brightness varies with a period of between one day and several weeks. The remarkable property of Cepheids is the existence of a relation between the average brightness and the period of their oscillations. For example, the average brightness is $1,000$ times that of the Sun for a period of days and $10,000$ times that amount for a period of several weeks. It is this relation that makes Cepheids one of the basic tools of astrophysics.

If we know this relationship for a variable star, it is relatively easy, by the determination its period to derive its absolute magnitude M. By then measuring its apparent magnitude m we can then calculate the distance in parsec with of the relation proved earlier above:

$$M = m - 5(\log(d) - 1) \tag{12.2.146}$$

One of the main reasons for constructing the Hubble Space Telescope (HST) was to measure light curves of Cepheid variables in other galaxies. It is especially important to use Cepheids to measure distances to the galaxies in two nearby clusters: the Virgo Cluster (the nearest rich cluster), and the Fornax Cluster (a somewhat sparser collection of galaxies).

HST can zoom in on a small portion of a galaxy to find and measure Cepheids:


Galaxy NGC1365
HST • WFPC2
PRC96-21a • ST ScI OPO • May 9, 1996 • W. Freedman (Carnegie Insitution of Washington) and NASA

and the zoom inside the area of interest:

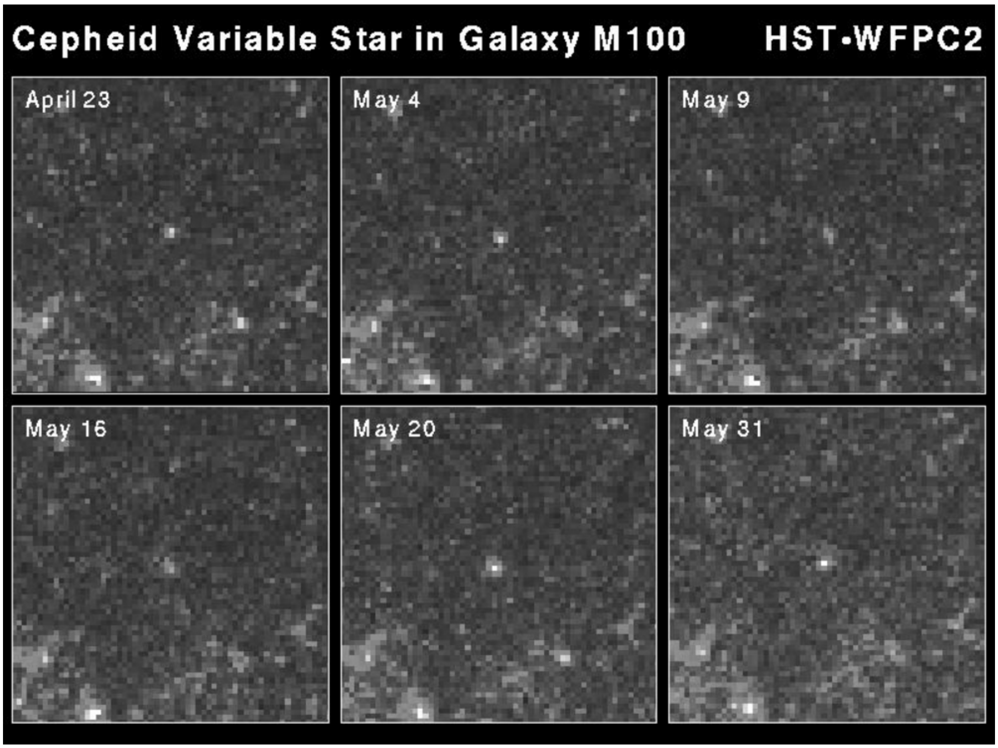


Figure 12.183: **Hubble Space Telescope Cepheid measurement**

The empirical period-luminosity relation for classical Cepheids was discovered in 11908 (holocene calendar) by Henrietta Swan Leavitt in an investigation of thousands of variable stars in the Magellanic Clouds. She published it in 11912 (holocene calendar) with further evidence. Once the period-luminosity relationship is calibrated, the luminosity of a given Cepheid whose period is known can be established. Their distance is then found from their apparent brightness. The period-luminosity relationship has been calibrated by many astronomers throughout the 120th century (holocene calendar), beginning with Hertzsprung. Calibrating the period-luminosity relation has been problematic; however, a firm Galactic calibration was established by Benedict et al. 12007 (holocene calendar) using precise HST parallaxes for 10 nearby classical Cepheids. Also, in 12008 (holocene calendar), ESO astronomers estimated with a precision within 1% the distance to the Cepheid RS Puppis, using light echos from a nebula in which it is embedded. However, that latter finding has been actively debated in the literature.

The following relationship between a Population I Cepheid's period P_T (in days) and its mean absolute magnitude $\bar{M}$ was established from Hubble Space Telescope trigonometric parallaxes for 10 nearby Cepheids:

$$\bar{M} = (-2.43 \pm 0.12)(\log(P_T) - 1) - (4.05 \pm 0.02) \tag{12.2.147}$$

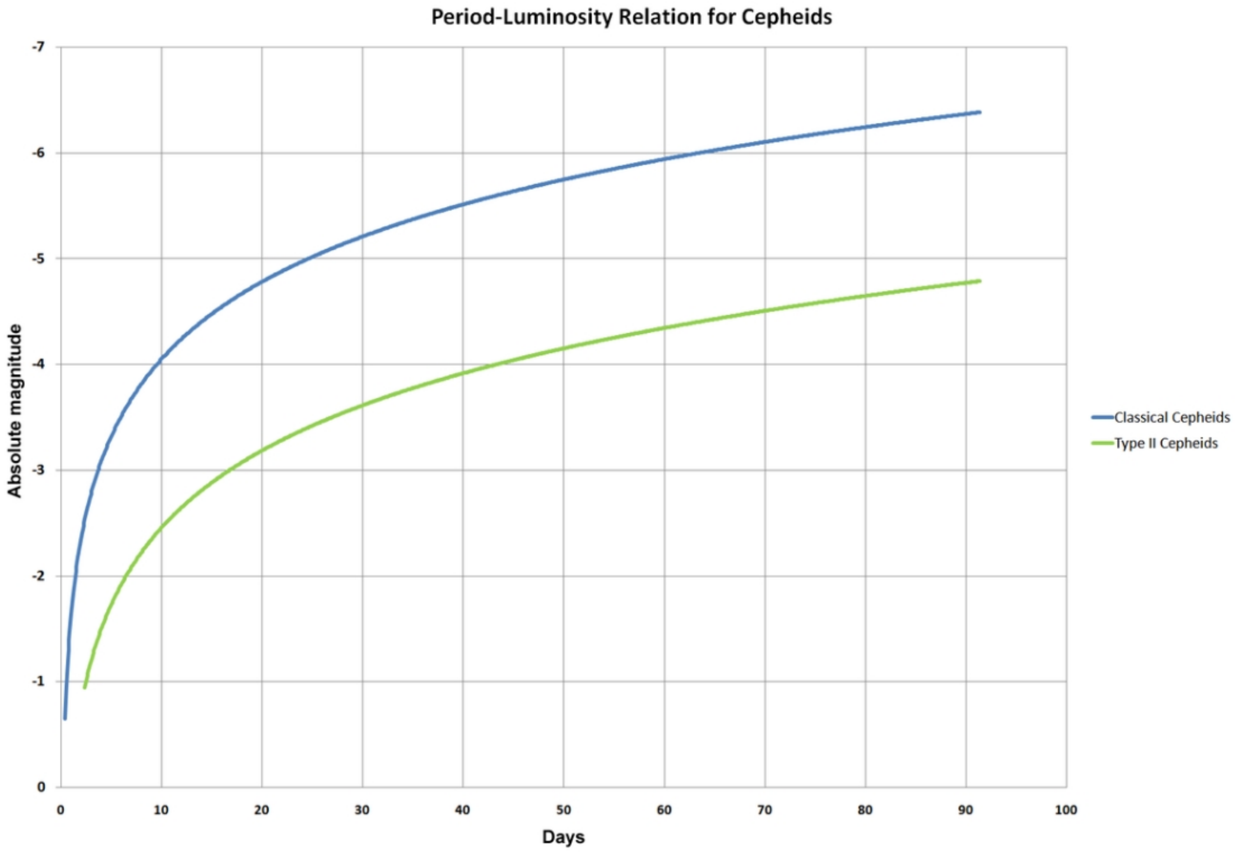


Figure 12.184: **Cepheid absolute magnitude-period plot (source: Wikipedia)**

Let us now try to derive theoretically such a period-luminosity relation! For this, we assume that fluid object (including a star) of radius R has a fundamental pulsation period:

$$P_T \cong \frac{2R}{v_s} \tag{12.2.148}$$

where v_s is the sound speed. This is simply the time it takes a sound (or pressure) wave to cross the stellar diameter! But we have proved in the section of Music Mathematics (see page 8202) that in the case of longitudinal sound waves in an isothermal gas we had the Newton-Laplace speed of sound relation:

$$v_s = \sqrt{\frac{\gamma P}{\rho_0}} \tag{12.2.149}$$

But we know from virial theorem (see section Continuum Mechanics page 3897)

$$U = 2T \tag{12.2.150}$$

Or written according to Astrophysics notation:

$$E_p = 2E_c \tag{12.2.151}$$

Replacing E_p by the potential gravific energy and E_c by the kinetic of a particle in a gas, we get (we have already seen that also during our study of viriel theorem):

$$\frac{GM^2}{R} = \cancel{2}\left(\frac{3}{\cancel{2}}kNT\right) \tag{12.2.152}$$

Using the relation (also proved during our study of the Newton-Laplace speed of sound relation):

$$v_s=\sqrt{\frac{\gamma NkT}{M}} \tag{12.2.153}$$

we get:

$$\frac{GM^2}{R}=3kN\left(\frac{v_s^2M}{\gamma Nk}\right)=\frac{3v_s^2M}{\gamma} \tag{12.2.154}$$

where we take for the Poisson constant $\gamma=5/3$ as proved in the section of Thermodynamics (see page 3647) for an ideal mono-atomic gas.

The pulsation period is then:

$$P_T\cong\frac{2R}{v_s}\cong\frac{2R}{\sqrt{\dfrac{GM^2\gamma}{3RM}}}\cong\frac{2R}{\sqrt{\dfrac{GM5}{9R}}}=\sqrt{\frac{4R^2}{\dfrac{5GM}{9R}})}=\sqrt{\frac{36R^3}{5GM}} \tag{12.2.155}$$

Plugging in numbers and normalizing with Solar mass and radius, we get:

$$P_T\cong 4271.773\left(\frac{R}{R_{\odot}}\right)^{3/2}\left(\frac{M_{\odot}}{M}\right)^{1/2} \tag{12.2.156}$$

One can derive the period-luminosity law as follows: Assume a sample of stars of the same mass (classical Cepheids have masses of $5-10$ solar masses), and the same temperature (the instability strip is approximately vertical in the H-R diagram). The bolometric luminosity then scales as R^2 as we have proved earlier above:

$$L_{\text{bol}}=\left(\frac{T}{T_{\odot}}\right)\left(\frac{R}{R_{\odot}}\right)^2 \tag{12.2.157}$$

The pulse period scales as $R^{3/2}$ (see above!). Then it follows immediately that:

$$P_T\propto L_{\text{bol}}^{3/4} \tag{12.2.158}$$

As we know we have also:

$$M\propto -2.5\log(L_{\text{bol}}) \tag{12.2.159}$$

Therefore:

$$L_{\text{bol}}\propto 10^{-M/2.5} \tag{12.2.160}$$

Hence:

$$P_T\propto L_{\text{bol}}^{3/4}\propto\left(10^{-M/2.5}\right)^{3/4} \tag{12.2.161}$$

If we take the logarithm (it is more common in the field of Astrophysics...), we then get the "cepheid period-luminosity relation" (even if it should be named... "cepheid period-absolute magnitude relation" instead):

$$\boxed{\log(P_T)\propto\frac{3}{4}\cdot\left(-\frac{1}{2.5}\right)M=-0.3M} \tag{12.2.162}$$

As we can see we are not too far from the experimental relation obtained thanks to Hubble observations given above (where we had $\log(P_T)\propto -0.4\bar{M}$).

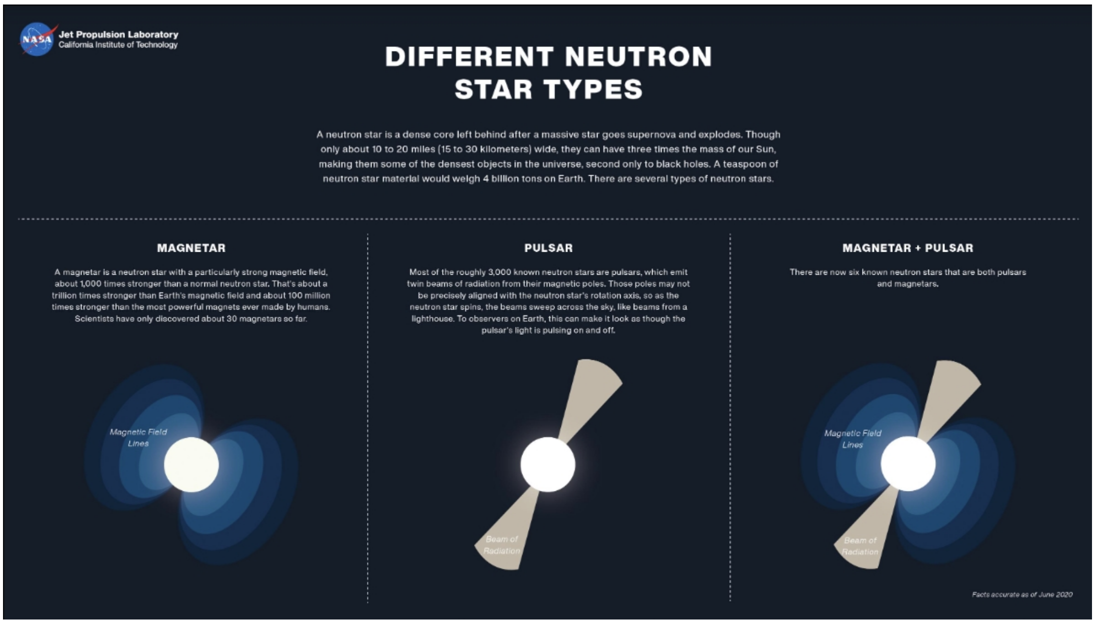
Remark

If any reader knows a more robust derivation and result and is in possessions of the detailed developments, don't hesitate to share the corresponding LaTeX text with us.

Cepheids aren't perfect distance indicators. For one thing, their brightness and periods of pulsation can vary with their chemical composition. There's also the problem of crowding and confusion: what if our view of a distant galaxy appears to show a single, varying Cepheid star... but is really a combination of light from the Cepheid and several nearby stars, all mixed together?

12.2.1.3 Neutron Stars (magnetars/pulsars)

A "neutron star" is the collapsed core of a large star (10 to 29 solar masses). Neutron stars are the smallest and densest stars known to exist. With a radius on the order of 10 [km], they can, however, have a mass of about twice that of the Sun (actually a G2V-type star, aka yellow dwarf). They result from the supernova explosion of a massive star, combined with gravitational collapse, that compresses the core past the White Dwarf star density to that of atomic nuclei (by the process of electron capture see in the section of Nuclear Physics at page 5088). Most of the basic models for these objects imply that neutron stars are composed almost entirely of neutrons, which are subatomic particles with no net electrical charge and with slightly larger mass than protons. They are supported against further collapse by neutron degeneracy pressure, a phenomenon described by the Pauli exclusion principle (see further below). If the remnant has too great a density, something which occurs in excess of an upper limit of the size of neutron stars at 2-3 solar masses, it will continue collapsing to form a Black Hole (see proof further below).



DIFFERENT NEUTRON STAR TYPES

A neutron star is a dense core left behind after a massive star goes supernova and explodes. Though only about 10 to 20 miles (15 to 30 kilometers) wide, they can have three times the mass of our Sun, making them some of the densest objects in the universe, second only to black holes. A teaspoon of neutron star material would weigh 4 billion tons on Earth. There are several types of neutron stars.

MAGNETAR

A magnetar is a neutron star with a particularly strong magnetic field, about 1,000 times stronger than a normal neutron star. That's about a trillion times stronger than Earth's magnetic field and about 100 million times stronger than the most powerful magnets ever made by humans. Scientists have only discovered about 30 magnetars so far.

PULSAR

Most of the roughly 3,000 known neutron stars are pulsars, which emit twin beams of radiation from their magnetic poles. Those poles may not be precisely aligned with the neutron star's rotation axis, so as the neutron star spins, the beams sweep across the sky, like beams from a lighthouse. To observers on Earth, this can make it look as though the pulsar's light is pulsing on and off.

MAGNETAR + PULSAR

There are now six known neutron stars that are both pulsars and magnetars.

Figure 12.185: **Infography of different types of Neutron Stars (source: Jet Propulsion Laboratory/NASA)**

Most neutron stars know so far are pulsars. However when a neutron star spin slow down enough, it no longer radiate very much energy, and so it is no longer considered a pulsar. This usually happens within a few million years. Magnetars are differentiated from other neutron stars by having even stronger magnetic fields, and by rotating more slowly in comparison. Most magnetars know so far rotate once every two to ten seconds, whereas typical neutron stars, observed as radio pulsars, rotate one to ten times per second.

12.2.1.3.1 Chandrasekhar limit

We have already determined in the section of Classical Mechanics (see page 3234) the Schwarzschild radius (in its classical form) that expresses the critical radius of a body for the release speed to it surface to be equal to that of speed of light. We obtained the following relation which typically expressed the radius that have a given celestial object to have a release speed equal to that of light:

$$R=\frac{2GM}{c^2} \tag{12.2.163}$$

In this particular case the star is what we named a "Black Hole". However, before the Black Hole, a star passes as we have spoken, by several intermediate steps by which it can also stabilize. Thus, you have often had to read in the literature that for a White Dwarf to collapse into a neutron star, its mass must be greater than 1.4 solar masses but without mathematical proof. Well, that's what we're going to demonstrate now!

We will introduce the subject by studying the influence of the uncertainty principle on the size of an atomic system (it limits the minimum dimension). This example is very powerful because it shows that the uncertainty principle not only governs the process of measurement but also the overall behaviour of quantum systems.

The first example that we can give is that of the hydrogen atom, not that we are expecting a new result from this method of analysis, but rather because we can expose the use of the principle of uncertainty and insist on its meaning.

We admit that the proton, whose mass far exceeds that of the electron, can be considered fixed. The energy of the electron is written:

$$E_{\text{tot}}=\frac{1}{2}m_e v^2-\frac{q_e^2}{4\pi\varepsilon_0 r}=\frac{p^2}{2m_e}-\frac{q_e^2}{4\pi\varepsilon_0 r} \tag{12.2.164}$$

In classical physics, a system whose energy is given by the previous relation does not have a minimum: if we tend r toward zero by keeping the circular shape of the orbit, it is easy to see E_{tot} tends to $-\infty$. On the other hand, in quantum physics, this limit has no meaning: the principle of uncertainty opposes to it.

In this case, the search for the minimum E_1 of E_{tot} takes a meaning, since a constraint appears which maintains this minimum at a finite value. It is determined in quantum physics (see the Bohr model of the atom in the section of Quantum Corpuscular Physics page 4626) and requires:

$$p\cdot r=n\cdot\hbar=\hbar \tag{12.2.165}$$

where $n\in\mathbb{N}$. However, this relation aside, if the radius r of the atom becomes too small under external constraints (be careful, we get rid of the quantized orbits of the Bohr model of the atom which imposes a constraint on p) the linear momentum p of the electron can not be less than the uncertainty Δp imposed by the uncertainty principle of Heisenberg (see section Quantum Wave Mechanics page 4704), since Δx is of the order of the radius r of the atom. The shape itself of the preceding relation limits the scope of the method: we cannot expect to determine better than an order of magnitude of the minimum of E_{tot}.

In order to evaluate the minimum E_1 of total energy, which we interpret as the ground state of the hydrogen atom, we calculate the minimum of E_{tot} by eliminating p from the expression:

$$E_{\text{tot}}=\frac{p^2}{2m_e}-\frac{q_e^2}{4\pi\varepsilon_0 r} \qquad (12.2.166)$$

by:

$$p\cdot r=n\cdot\hbar=\hbar \qquad (12.2.167)$$

We get:

$$E_{\text{tot}}:=E(r)=\frac{\hbar^2}{2m_e r^2}-\frac{q_e^2}{4\pi\varepsilon_e r} \qquad (12.2.168)$$

The radius r_1 of the atom in the ground state is the value of r that gives $E(r)$ its minimum value:

$$\left.\frac{\mathrm{d}E(r)}{\mathrm{d}r}\right|_{r=r_1}=-\frac{\hbar^2}{m_e r_1^3}+\frac{q_e^2}{4\pi\varepsilon_0 r_1^2}=0 \qquad (12.2.169)$$

which is the well-known expression of the Bohr radius seen in the section of Corpuscular Quantum Physics during our study of the Bohr model of the atom. The energy E_1 of the ground state is now easily calculable.

The purpose of this example is to show that with Heisenberg's uncertainty principle we can by very simple reasoning find the fundamental state of a system. This is exactly how we will proceed to determine the conditions that cause a star to return to its ground state.

Let us attack now the study of a star. Schematically it consists of a mixture of two gases: one that is formed of nuclei on the one hand, and the electronic gas on the other.

During the life of the star, many fusion processes took place. They each increased the size and mass of the nuclei. The iron (Fe), which is abundant at the end of a star's life, contains an average of 56 nucleons (see section Nuclear Physics page 5019).

These nuclei are of a chemical or isotopic variety. As they are few in comparison with electrons, their pressure is that of a charged conventional gas, neutralized by the presence of electrons: it can be ignored, especially since the temperature is zero.

The electronic charge alone would not allow the electrons to resist the collapse of a star since the stellar matter is neutral. At very low temperatures, when the fuel is exhausted, the only pressure that the electronic gas can oppose to the hydrostatic pressure due to gravity is of quantum origin.

As a first approximation, the electrons thus exert on each other an apparent repulsion which is not of Coulomb origin (Pauli exclusion principle). As a first approximation, they obey a relation analogous to that of the atomic electron and which is written in the minimal case (or maximum of pressure):

$$p \cdot d \cong \hbar \tag{12.2.170}$$

where d is the average distance between two neighbouring electrons.

At the temperature $T = 0$ [K] equilibrium is reached when the total energy (the matter of the star) of the system is minimal.

What happens if we try to evaluate the variation of the radius r_{WD} of the White Dwarf as a function of its mass m_{WD}?

The gravitational potential energy of a star is given in good approximation by (see section Classical Mechanics page 3351):

$$E_p \cong -G\frac{m_{\text{WD}}^2}{r_{\text{WD}}} \tag{12.2.171}$$

The mass m_{WD} is approximately given by:

$$m_{\text{WD}} = m_p N \tag{12.2.172}$$

where m_p is the mass of the proton and N the number of nucleons contained in the star: the contribution of the electrons to the mass of the star is negligible and there is no need to distinguish between the mass of the neutron and that of the proton, almost identical.

The second contribution to energy is essentially that of the degenerate electronic gas (degeneration corresponds for recall to the existence of several states having the same energy), of kinetic origin. We could be tempted to write simply (assuming that the number of electrons is equal to the number of nucleons since we are for recall in the simplified hypothesis of a hydrogen gas):

$$E_c = N\frac{1}{2}m_e v^2 = \frac{Np^2}{2m_e} \tag{12.2.173}$$

This way of doing things leads to a dead end. If we request that the sum $E_p + E_c$ reaches a minimum value, we arrive at a value of the radius of the star so small that, by application of the relation:

$$p \cdot d \cong \hbar \tag{12.2.174}$$

the average velocity of the electrons v would exceed that of the light!

To avoid this contradiction, we have to use relativistic mechanics that showed us that, in this case (see section Special Relativity page 5627), we can express the total energy as:

$$E_{\text{tot}} = c\sqrt{m_{0e}^2 c^2 + p^2} \tag{12.2.175}$$

if the numerical value of the kinetic energy significantly outweighs that of the energy at rest, then we have:

$$E_{\text{tot}} = Nc\sqrt{m_{0e}^2 c^2 + p^2} = N(E_c + m_{0e}c^2) \tag{12.2.176}$$

and therefore:

$$E_{\text{tot}} \cong E_c = N m_{0e} c^2 = Npc \tag{12.2.177}$$

The average distance d between electrons is evaluated by assuming that the star is homogeneous, a sufficient approximation when we are looking for the order of magnitude of an average. We further simplify the geometry by admitting that each electron is surrounded by a spherical domain of radius d in which there is no other electron of the same spin and where we can count only one electron of opposite spin. Since then:

$$d^3 \cong \frac{r_{\text{WD}}^3}{N} \tag{12.2.178}$$

It remains to evaluate the minimum of the sum:

$$E_{\text{tot}} = E_p + E_c \cong -G\frac{m_p^2 N^2}{r_{\text{WD}}} + Npc \tag{12.2.179}$$

given the condition:

$$m_{\text{WD}} = m_p N \tag{12.2.180}$$

Then it comes:

$$p \cong \frac{\hbar}{d} \cong \frac{\hbar N^{1/3}}{r_{\text{WD}}} \tag{12.2.181}$$

and then:

$$E_{\text{tot}} = E_{\text{WD}}(r_{\text{WD}}) \cong -G\frac{m_p^2 N^2}{r_{\text{WD}}} + Nc\frac{\hbar}{r_{\text{WD}}}N^{1/3} = \frac{N^{1/3}}{r_{\text{WD}}}(c\hbar + G m_p^2 N^{2/3}) \tag{12.2.182}$$

that we write finally:

$$E_{\text{tot}} = E_{\text{WD}}(r_{\text{WD}}) \cong \frac{N^{1/3}}{r_{\text{WD}}}(c\hbar - G m_p^{4/3} m_{\text{WD}}^{2/3}) \tag{12.2.183}$$

Faced with this result, we are confronted with an unexpected situation. Indeed, if the factor:

$$K = c\hbar - G m_p^{4/3} m_{\text{WD}}^{2/3} \tag{12.2.184}$$

is positive, then the total energy of the White Dwarf is also positive, which means that the system is unbounded: the star is totally unstable (it has not reached its minimum energy threshold). It can only reduce its energy by increasing its radius without limits.

We see that the K factor is negative if:

$$m_{\text{WD}} > \left(\frac{c\hbar}{G}\right)^{3/2} \frac{1}{m_p^2} \tag{12.2.185}$$

If the White Dwarf exceeds this mass then we can no longer deal with the problem with the previous equations. It then satisfies the equations governing a star composed only of neutrons (neutron star) and this is then another problem that we will not tackle here for now.

The (approximate) mass of the famous "Chandrasekhar limit" is thus given by:

$$m_{\text{WD}} = m_C = \left(\frac{c\hbar}{G}\right)^{3/2} \frac{1}{m_p^2} \tag{12.2.186}$$

It constitutes the mass beyond which a White Dwarf collapses into a Neutron star.

Conventionally, astrophysicists associate this limit value with a multiplying factor of the mass of the Sun $M_\odot$. We actually (numerically):

$$\boxed{m_{\text{WD}} = m_C = \left(\frac{c\hbar}{G}\right)^{3/2} \frac{1}{m_p^2} = 1.4 M_\odot} \tag{12.2.187}$$

In the figure below on the left we have schematic slice through a neutron star. Letters N, n, p, e, μ refer to the presence of nuclei, fluid neutrons and protons, electrons and muons, respectively. The inner core composition is still uncertain and various exotic possibilities exist, including hyperons and deconfined quark matter.

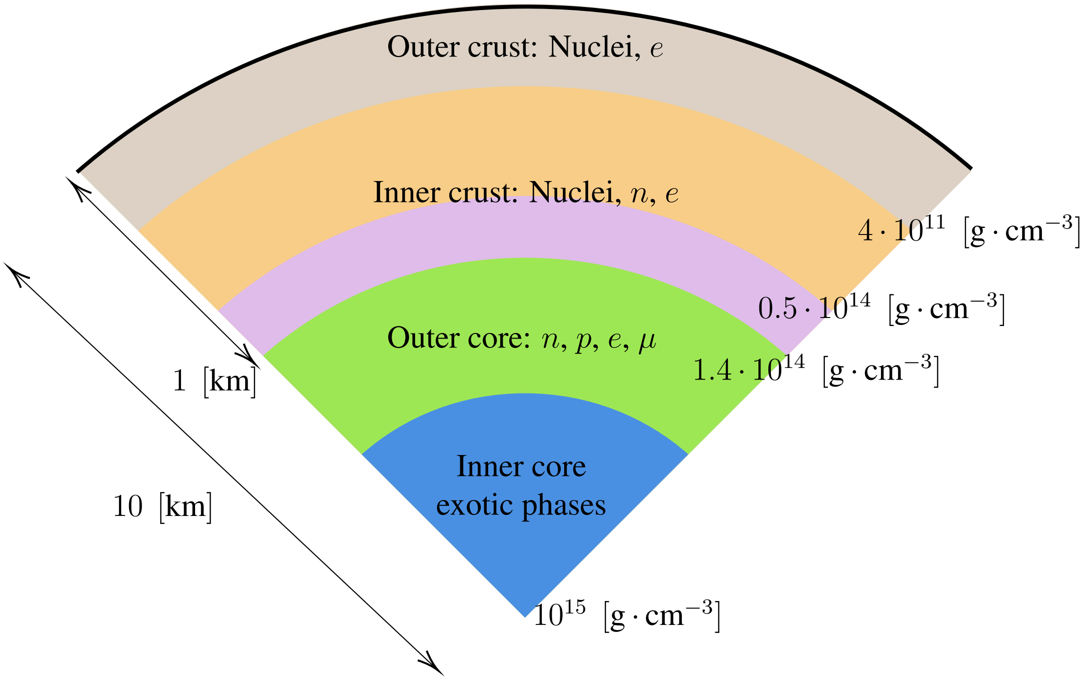

Outer crust: Nuclei, e
Inner crust: Nuclei, n, e
Outer core: n, p, e, μ
Inner core
exotic phases
$4 \cdot 10^{11}$ [g · cm^{-3}]
$0.5 \cdot 10^{14}$ [g · cm^{-3}]
$1.4 \cdot 10^{14}$ [g · cm^{-3}]
10^{15} [g · cm^{-3}]
1 [km]
10 [km]

Figure 12.186: **Neutron star slice**

On the figure below we have an overview of what we expect to be the composition of the inner crust in a neutron star:

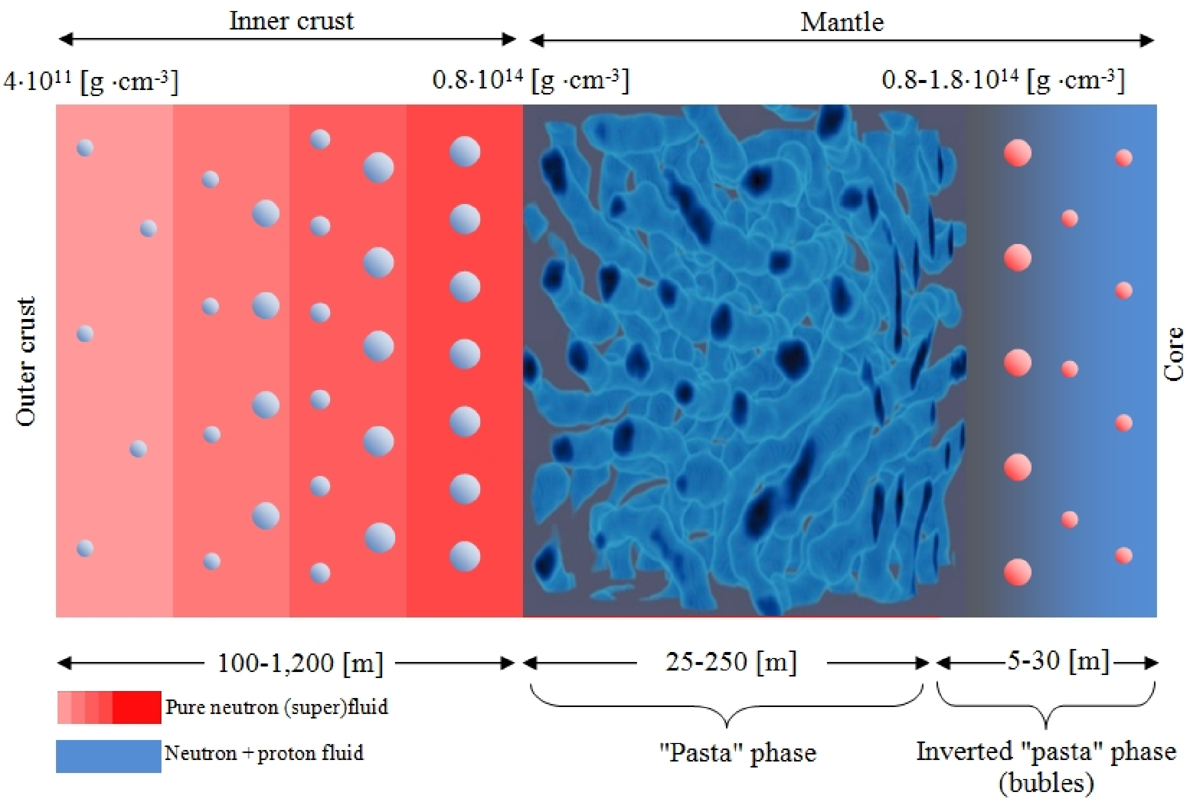

Inner crust
Mantle
4·10¹¹ [g ·cm⁻³]
0.8·10¹⁴ [g ·cm⁻³]
0.8-1.8·10¹⁴ [g ·cm⁻³]
Outer crust
Core
100-1,200 [m]
25-250 [m]
5-30 [m]
Pure neutron (super)fluid
Neutron + proton fluid
"Pasta" phase
Inverted "pasta" phase
(bubles)

Figure 12.187: Neutron star inner crust

At lower densities, a lattice of super-heavy, neutron-rich nuclei is immersed in a fluid of neutrons (which are likely to be superfluid) and a relativistic electron gas. At high enough densities the nuclei might deform and connect along certain directions to form extended tubes, sheets and bubbles of nuclear matter. These nuclear pasta phases might form a layer at the base of the neutron star crust, sometimes referred to as the mantle. Ranges of density and thickness given for each layer represent current uncertainties in the physics of neutron star crusts.

12.2.1.3.2 Rotation break limit

Let us make the simplifying assumption that the limit speed of rotation of a star (planet or star) is that which balances the centrifugal (fictional) force and gravitational force on the surface of the star such that we are led to write (see section Classical Mechanics page 3248):

$$m\frac{v^2}{R}=\frac{GMm}{R^2} \tag{12.2.188}$$

Write this relation supposes obviously that there is no connection other than the gravity which intervenes in the internal cohesion of the star. So the rotational time values we are going to get represent an upper bound (in the sense that the actual value is probably smaller).

It then comes from the previous relation:

$$v=\sqrt{\frac{GM}{R}} \tag{12.2.189}$$

To get the rotation time to which it corresponds, it suffices to divide the perimeter at the equator by its speed:

$$\boxed{t_r=\frac{2\pi R}{v}=\frac{2\pi R}{\sqrt{\dfrac{GM}{R}}}=\frac{2\pi R}{\sqrt{\dfrac{6.674\cdot 10^{-11}M}{R}}}}\tag{12.2.190}$$

So, for the Earth, we have as limit period of rotation before rupture:

$$t_{r,\text{Earth}}=\frac{2\pi\cdot 6,378,137}{\sqrt{\dfrac{6.674\cdot 10^{-11}\cdot 5.9736\cdot 10^{24}}{6,378,137}}}\cong 84\,[\text{min}]\tag{12.2.191}$$

For the Sun:

$$t_{r,\text{Sun}}=\frac{2\pi\cdot 1,392,000,000}{\sqrt{\dfrac{6.674\cdot 10^{-11}\cdot 1.9891\cdot 10^{30}}{1,392,000,000}}}\cong 7.8\,[\text{h}]\tag{12.2.192}$$

Let us now consider the case of the pulsar NP0532 which has a rotation of 33 milliseconds[34]. We would like to determine its radius. We then have using the previous relations:

$$R=\left(\frac{t_r\sqrt{GM}}{2\pi}\right)^{2/3}\tag{12.2.193}$$

Using the theoretical relation of the Chandrasekhar mass limit (since a pulsar is a Neutron star rotating rapidly on itself):

$$m_{\text{WD}}=m_C=\left(\frac{c\hbar}{G}\right)^{3/2}\frac{1}{m_p^2}=1.4M_{\odot}\tag{12.2.194}$$

We then have for the radius of the smallest possible pulsar according to these hypotheses (and approximations...):

$$\boxed{R_{\text{min}}=\left(\frac{t_r\sqrt{G\cdot 1.4M_{\odot}}}{2\pi}\right)^{2/3}}\tag{12.2.195}$$

With the following numerical application (using data from the Crab pulsar):

$$R=\left(\frac{t_r\sqrt{G\cdot 1.4M_{\odot}}}{2\pi}\right)^{2/3}=\left(\frac{33\cdot 10^{-3}\sqrt{6.674\cdot 10^{-11}\cdot 1.4\cdot 1.9891\cdot 10^{30}}}{2\pi}\right)^{2/3}\cong 172\,[\text{km}]\tag{12.2.196}$$

With the pulsar millisecond PSR J1748-2446ad having a period of 1.39 milliseconds we then fall on[35]:

$$R=\left(\frac{t_r\sqrt{G\cdot 1.4M_{\odot}}}{2\pi}\right)^{2/3}\cong 21\,[\text{km}]\tag{12.2.197}$$

---

[34] In the beginning of the 121st century (holocene calendar) there is a project of NASA of using pulsars as interplanetary GPS.

[35] The slowest know pulsar known at the day we write these lines is PSR J0250+5854 with a rotation period of 23.5 [s]. The faster pulsar is PSR J1748-2446ad with a rotation period of 0.00139 [s]

what is remarkable (even if it is an approximation!) to think that such a quite huge mass can be contained in such a small radius. Note that for the latter it corresponds to respectively a maximum density of:

$$\rho_{\text{max}} = \frac{1.4M_{\odot}}{\dfrac{4}{3}\pi R^3} \cong 15 \cdot 10^{14}\,[\text{kg} \cdot \text{mm}^{-3}] \tag{12.2.198}$$

12.2.1.3.3 Neutron Star Masses and Densities

If a nearly spherical star of mass M and radius R rotates with angular velocity $\omega = 2\pi/R$ we know obviously that we have then equilibrium of the centrifugal (fictional) force and gravitational bound a the equator at a given instant such that we are led to write (we make here an obvious classical approximation instead of using General Relativity that should be mandatory when dealing with huge masses):

$$\frac{v^2}{R} = \frac{GM}{R^2} \tag{12.2.199}$$

Hence:

$$\frac{(\omega R)^2}{R} = \frac{GM}{R^2} \tag{12.2.200}$$

After simplification:

$$\omega^2 R = \frac{GM}{R^2} \tag{12.2.201}$$

Implying:

$$\left(\frac{2\pi}{T}\right)^2 R = \frac{GM}{R^2} \tag{12.2.202}$$

Hence:

$$T^2 = \left(\frac{4\pi R^3}{3}\right)\frac{3\pi}{GM} \tag{12.2.203}$$

As the density of a ball is given by:

$$\rho = \frac{M}{\dfrac{4\pi R^3}{3}} \tag{12.2.204}$$

We then have:

$$T = \sqrt{\frac{3\pi}{G\rho}} \tag{12.2.205}$$

Hence:

$$\boxed{\rho = \frac{3\pi}{GT^2}} \tag{12.2.206}$$

That latter relation gives a conservative lower limit to the mean density because a rapidly spinning star is oblate, which increases the centrifugal (fictional) acceleration and decreases the gravitational acceleration at its equator.

The first pulsar discovered has a period of $T=1.3$ [s], so its mean density is at least:

$$\rho=\frac{3\pi}{GT^2}\cong 10^8\,[\text{g}\cdot\text{cm}^{-3}] \tag{12.2.207}$$

The fastest known pulsar actually (year 12018 according to holocene calendar), PSR J1748-2446ad, has $T=1.4\cdot 10^{-3}$ [s] implying $\rho>10^{14}$, the density of atomic nuclei!!!!! It has been calculated that this neutron star contains slightly less than two times the mass of the Sun, within the typical range of neutron stars. Its radius is constrained to be less than 16 [km]. At its equator it is spinning at approximately 24% of the speed of light, or over $70,000$ [km] per second!!!

12.2.1.3.4 Neutron Star magnetic field

As the star's core collapses, its rotation rate increases as a result of conservation of angular momentum, hence newly formed neutron stars rotate at up to several hundred times per second. Some neutron stars emit beams of electromagnetic radiation that make them detectable as pulsars. Indeed, the discovery of pulsars in 11967 (holocen calendar) was the first observational suggestion that neutron stars exist. The radiation from pulsars is thought to be primarily emitted from regions near their magnetic poles. If the magnetic poles do not coincide with the rotational axis of the neutron star, the emission beam will sweep the sky, and when seen from a distance, if the observer is somewhere in the path of the beam, it will appear as pulses of radiation coming from a fixed point in space (the so-called "lighthouse effect"). The fastest rotation rate for a neutron star was a rate of 716 times a second or $43,000$ revolutions per minute, giving a linear speed at the surface on the order of $0.165c$....

So now let us focus on the simplified math approach of the impact of the angular momentum conservation on the magnetic field of the star.

From the conservation of angular moment as the core collapses we have (see section Classical Mechanics page 3297):

$$J_i\omega_i=J_f\omega_f \tag{12.2.208}$$

Or, for a sphere of constant density (see section Geometric Shapes page 2968):

$$\frac{2}{5}M_iR_i^2\omega_i=\frac{2}{5}M_fR_f^2\omega_f \tag{12.2.209}$$

So the final spin frequency is:

$$\omega_f=\omega_i\left(\frac{R_i}{R_f}\right)^2 \tag{12.2.210}$$

or the final spin period is:

$$T_f=T_i\left(\frac{R_f}{R_i}\right)^2 \tag{12.2.211}$$

The magnetic flux Φ ($\vec{B}$ multiplied by S) through the surface of the core is also conserved in collapse. So roughly (see section Electrodynamics page 4100):

$$B_i 4\pi R_i^2 = B_f 4\pi R_f^2 = 0 \tag{12.2.212}$$

Which means that:

$$B_f = B_i \left(\frac{R_i}{R_f}\right)^2 \tag{12.2.213}$$

The Sun (actually a G2V-type star, aka yellow dwarf) and many other stars are known to possess approximately dipolar magnetic fields. Stellar interiors are fully ionized and hence good electrical conductors. Charged particles are constrained to move along magnetic field lines, and magnetic field lines are tied to the charged particles. When a star collapses from a radius $\sim 10^6$ [km] to ~ 10 [km], its cross-sectional area a is divided by $\sim 10^{10}$, its magnetic flux $\Phi \equiv \int \vec{B} \cdot \hat{n} \mathrm{d}a$ (where $\hat{n}$ is the unit vector normal to each infinitesimal surface area $\mathrm{d}a$) is conserved, and the magnetic field strength is multiplied by $\sim 10^{10}$. An initial magnetic field strength $B \sim 100$ [G] becomes $B \sim 10^{12}$ [G] after collapse, so young neutron stars should have very strong dipolar fields. The best models of the core-collapse process show that a dynamo effect can generate even stronger magnetic fields. Such dynamos may be able to produce the $10^{14} - 10^{15}$ [G] fields observed in magnetars, which are neutron stars having such strong magnetic fields that their radiation is powered by magnetic field decay. Conservation of angular momentum during collapse increases the rotation rate by about the same factor, 10^{10}, yielding initial rotation periods in the millisecond range.

12.2.1.3.5 Spin-Down Luminosity

If we now consider a neutron star as perfectly symmetric of radius $R = 10$ [km] it's moment of inertia will be equal to (see section Classical Mechanics page 3297):

$$J = \frac{2MR^2}{5} \cong \frac{2 \cdot 1.4 M_\odot \cdot (10^4)^2}{5} \cong 10^{38}\,[\mathrm{kg} \cdot \mathrm{m}^2] \tag{12.2.214}$$

The rotational kinetic energy (see section Classical Mechanics page 3297) of a canonical neutron star with the rotation period $T = 0.033$ [s] of the Crab pulsar is:

$$E_r = \frac{1}{2}\omega^2 J = \frac{2\pi^2 J}{T^2} \cong \frac{2\pi^2 \cdot 10^{38}}{0.033^2} \cong 1.8 \cdot 10^{42}\,[\mathrm{J}] \tag{12.2.215}$$

As magnetic dipole radiation extracts rotational energy, it slowly increases the period of a pulsar:

$$\dot{T} \equiv \frac{\mathrm{d}T}{\mathrm{d}t} > 0 \tag{12.2.216}$$

Note that the period derivative $\dot{T}$ is a dimensionless (seconds per second) pure number. Combining the observed period T and period derivative $\dot{T}$ yields an estimate of the rate $\dot{E}$ at which the rotational energy is changing. The quantity:

$$-\dot{E}_r \equiv -\frac{\mathrm{d}E_t}{\mathrm{d}t} = -\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}J\Omega^2\right) = -J\Omega\dot{\Omega} \tag{12.2.217}$$

is named the "spin-down luminosity". It is not a measured luminosity; it is the measured loss rate of rotational energy, which is presumed to equal the luminosity of magnetic dipole radiation. The spin-down luminosity is usually expressed in terms of the pulse period T :

$$T=\frac{2\pi}{P} \quad \text{so} \quad \dot{\Omega}=2\pi\left(-T^{-2}\dot{T}\right) \tag{12.2.218}$$

and the prior-previous relation becomes:

$$-\dot{E}_r=\frac{4\pi^2 J\dot{P}}{P^3} \tag{12.2.219}$$

The Crab pulsar has $T=0.033$ [s] and $\dot{T}=10^{-12.4}$ (notice that the precision is so that pulsars can be used as galactic GPS!). If $J=10^{38}$ $[\text{kg}\cdot\text{m}^2]$, its spin-down luminosity is:

$$-\dot{E}_r=\frac{4\pi^2 J\dot{P}}{P^3}=\frac{4\pi^2\cdot 10^{45}\,[\text{g}\cdot\text{cm}^2]\cdot 10^{-12.4}\,[\text{s}\cdot\text{s}^{-1}]}{(0.033\,[\text{s}])^3}\cong 4\cdot 10^{38}\,[\text{erg}\cdot\text{s}^{-1}]\cong 10^5 L_{\odot} \tag{12.2.220}$$

So the luminosity of the low frequency magnetic dipole radiation from the Crab pulsar is comparable with the entire radio output of our Galaxy...

Pulsar are in the early 121st century very useful as detectors of huge gravitational waves. This is what we name a "pulsar timing array" (PTA) that is a set of galactic pulsars that is monitored and analysed to search for correlated signatures in the pulse arrival times on Earth. As such, they are galactic-sized detectors. Although there are many applications for pulsar timing arrays, the best known is the use of an array of millisecond pulsars to detect and analyse long-wavelength (i.e., low-frequency) gravitational wave background and was proposed originally in the years 11970 by M.V. Sazhin and S.L. Detweiler.

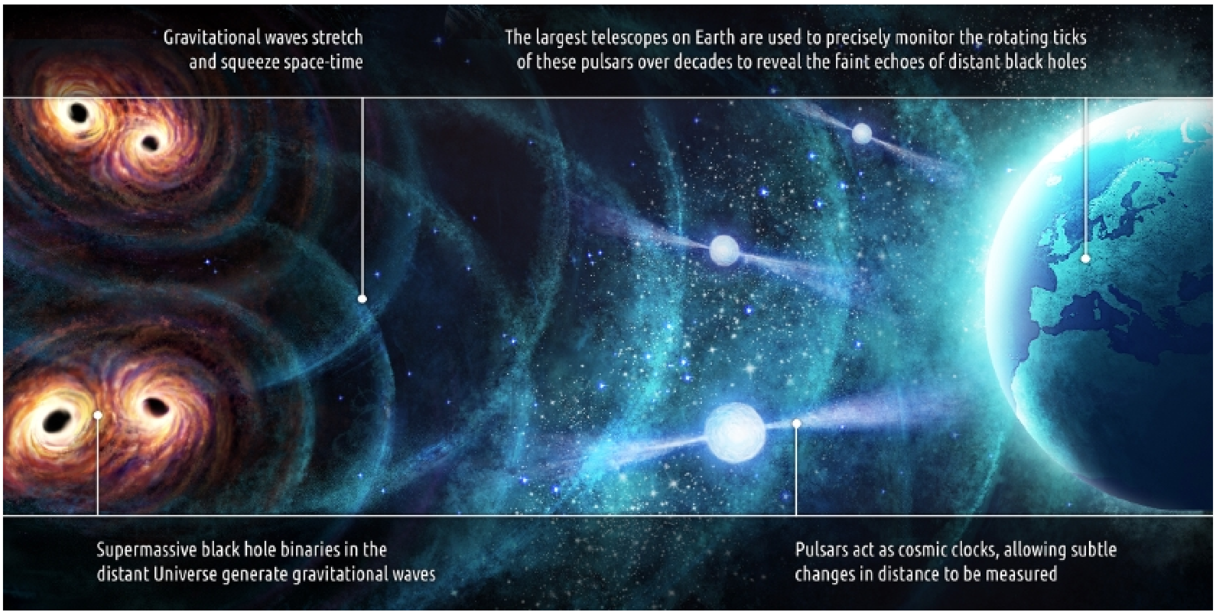
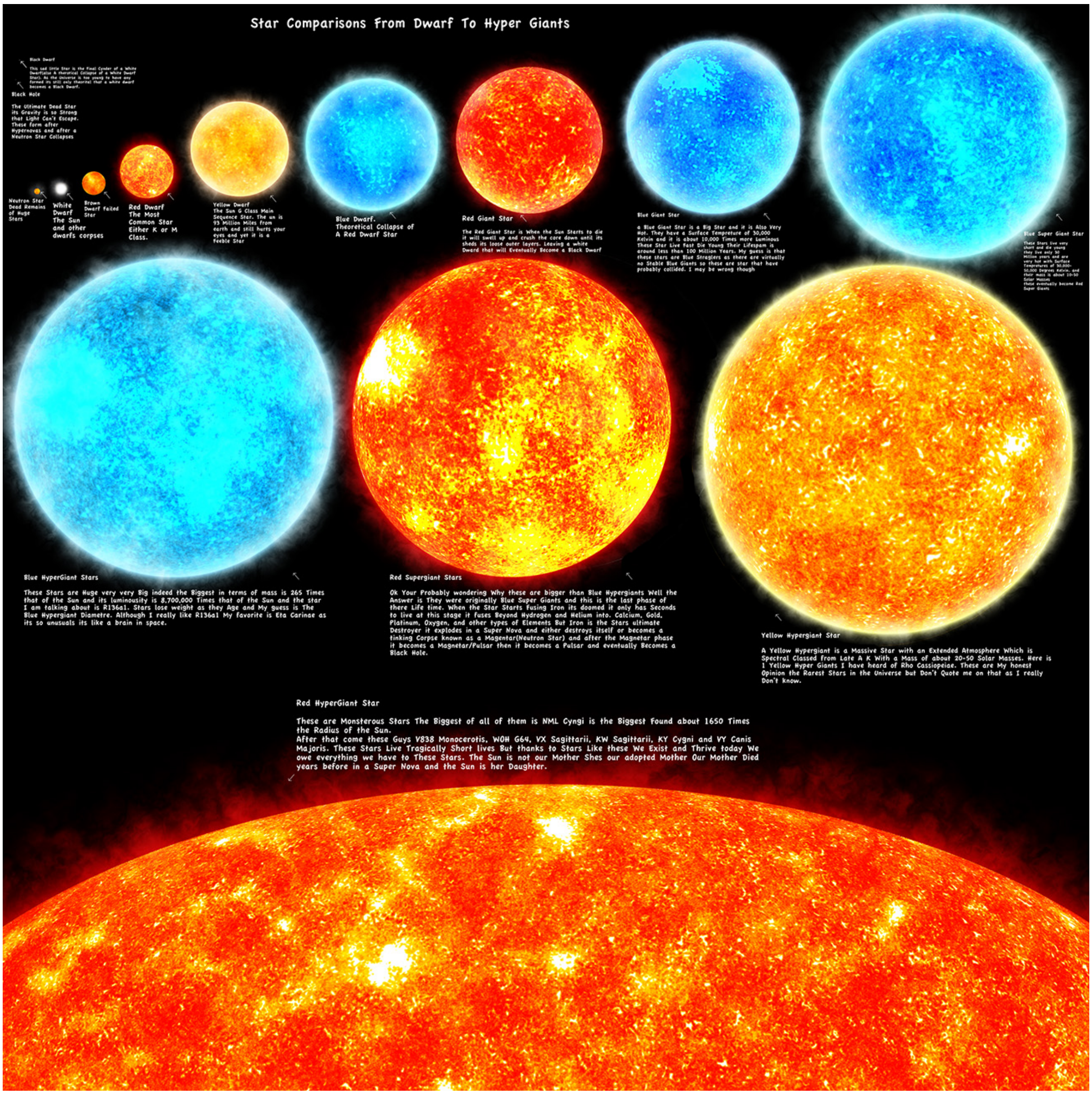

Gravitational waves stretch and squeeze space-time
The largest telescopes on Earth are used to precisely monitor the rotating ticks of these pulsars over decades to reveal the faint echoes of distant black holes
Supermassive black hole binaries in the distant Universe generate gravitational waves
Pulsars act as cosmic clocks, allowing subtle changes in distance to be measured

Figure 12.188: **Pulsar timing array artistic illustration (source: Forth Institute of Astrophysics)**

The European Pulsar Timing Array (EPTA) and Indian Pulsar Timing Array (InPTA) collaborations have measured a low-frequency common signal in the combination of their second and first data releases, respectively, with the correlation properties of a gravitational wave background (GWB). Such a signal may have its origin in a number of physical processes including a cosmic population of inspiralling supermassive black hole binaries (SMBHBs); inflation, phase transitions, cosmic strings, and tensor mode generation by the non-linear evolution of scalar perturbations in the early Universe; and oscillations of the Galactic potential in the presence of ultra-light dark matter (ULDM). At the current stage of emerging evidence, it is impossible to discriminate among the different origins. It has been found (see [20]) in 12024 (holocene calendar) that the signal is consistent with a cosmic population of inspiralling SMBHBs, and its relatively high amplitude can be used to place constraints on binary merger timescales and the SMBHhost galaxy scaling relations. If this origin is confirmed, this would be the first direct evidence that SMBHBs merge in nature, adding an important observational piece to the puzzle of structure formation and galaxy evolution. As for early Universe processes, the measurement would place tight constraints on the cosmic string tension and on the level of turbulence developed by first-order phase transitions. Other processes would require non-standard scenarios, such as a blue-tilted inflationary spectrum or an excess in the primordial spectrum of scalar perturbations at large wavenumbers. Finally, a ULDM origin of the detected signal is disfavoured, which leads to direct constraints on the abundance of ULDM in our Galaxy.

Arrived so far in the study of some very common stars models (excepted for Black Holes that are treated in the section of General Relativity page 5835) it is maybe the right time to introduce this summary image:


Star Comparisons From Dwarf To Hyper Giants

Black Dwarf

This sad little Star is the Final Cynder of a White Dwarf(also A theretical Collapse of a White Dwarf Star). As the Universe is too young to have any formed its still only theorital that a white dwarf becomes a Black Dwarf.

Black Hole

The Ultimate Dead Star its Gravity is so Strong that Light Can't Escape. These form after Hypernovas and after a Neutron Star Collapses

Neutron Star Dead Remains of Huge Stars

White Dwarf The Sun and other dwarfs corpses

Brown Dwarf Failed Star

Red Dwarf The Most Common Star Either K or M Class.

Yellow Dwarf The Sun G Class Main Sequence Star. The un is 93 Million Miles from earth and still hurts your eyes and yet it is a feeble Star

Blue Dwarf. Theoretical Collapse of A Red Dwarf Star

Red Giant Star

The Red Giant Star is When the Sun Starts to die it will swell up and crush the core down until its sheds its loose outer layers. Leaving a white Dward that will Eventually Become a Black Dwarf

Blue Giant Star

a Blue Giant Star is a Big Star and it is Also Very Hot. They have a Surface Tempreture of 30,000 Kelvin and it is about 10,000 Times more Luminous These Star Live Fast Die Young Their Lifespam is around less than 100 Million Years. My guess is that these stars are Blue Straglers as there are virtually no Stable Blue Giants so these are star that have probably collided. I may be wrong though

Blue Super Giant Star

These Stars live very short and die young they live only 50 Million years and are very hot with Surface Tempretures of 30,000-50,000 Degrees Kelvin, and their mass is about 10-50 Solar Masses these eventually become Red Super Giants

Blue HyperGiant Stars

These Stars are Huge very very Big indeed the Biggest in terms of mass is 265 Times that of the Sun and its luminousity is 8,700,000 Times that of the Sun and the star I am talking about is R136a1. Stars lose weight as they Age and My guess is The Blue Hypergiant Diametre. Although I really like R136a1 My favorite is Eta Carinae as its so unusuals its like a brain in space.

Red Supergiant Stars

Ok Your Probably wondering Why these are bigger than Blue Hypergiants Well the Answer is They were originally Blue Super Giants and this is the last phase of there Life time. When the Star Starts Fusing Iron its doomed it only has Seconds to live at this stage it fuses Beyond Hydrogen and Helium into. Calcium, Gold, Platinum, Oxygen, and other types of Elements But Iron is the Stars ultimate Destroyer it explodes in a Super Nova and either destroys itself or becomes a tinking Corpse known as a Magentar(Neutron Star) and after the Magnetar phase it becomes a Magnetar/Pulsar then it becomes a Pulsar and eventually Becomes a Black Hole.

Yellow Hypergiant Star

A Yellow Hypergiant is a Massive Star with an Extended Atmosphere Which is Spectral Classed from Late A K With a Mass of about 20-50 Solar Masses. Here is 1 Yellow Hyper Giants I have heard of Rho Cassiopeiae. These are My honest Opinion the Rarest Stars in the Universe but Don't Quote me on that as I really Don't know.

Red HyperGiant Star

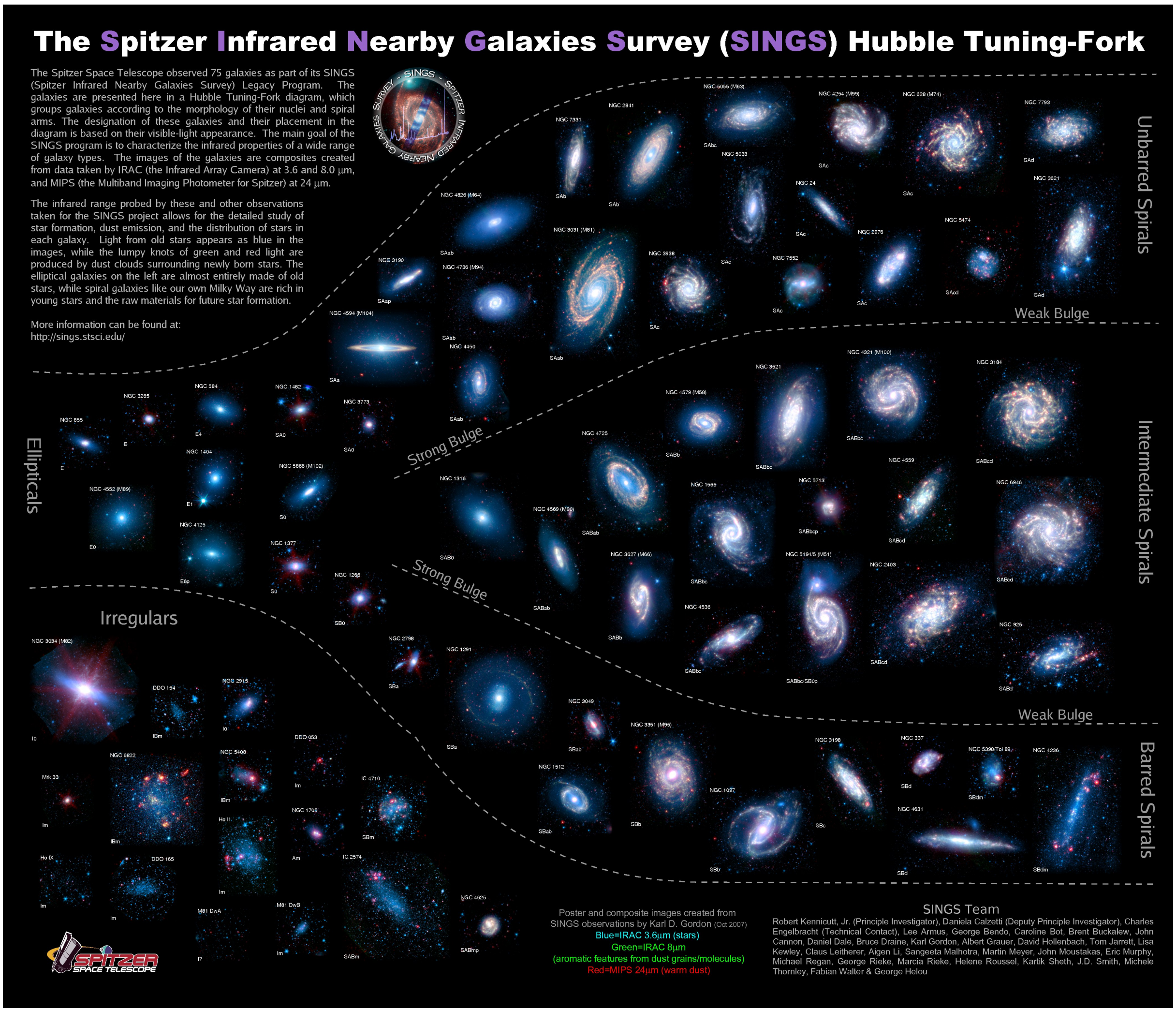
These are Monsterous Stars The Biggest of all of them is NML Cyngi is the Biggest Found about 1650 Times the Radius of the Sun.
After that come these Guys V838 Monocerotis, WOH G64, VX Sagittarii, KW Sagittarii, KY Cygni and VY Canis Majoris. These Stars Live Tragically Short lives But thanks to Stars Like these We Exist and Thrive today We owe everything we have to These Stars. The Sun is not our Mother Shes our adopted Mother Our Mother Died years before in a Super Nova and the Sun is her Daughter.

Figure 12.189: Star comparisons from Black Dwarf to hyper giants (author: Karl Garnham)

12.2.2 Galaxies

A galaxy is a gravitationally bound system of stars, stellar remnants, interstellar gas, dust, and dark matter[36].

Galaxies range in size from dwarfs with just a few billion (10^9) stars to giants with one hundred trillion (10^{14}) stars, each orbiting its galaxy's center of mass. Galaxies are categorized according to their visual morphology as elliptical, spiral or irregular:



The Spitzer Infrared Nearby Galaxies Survey (SINGS) Hubble Tuning-Fork

The Spitzer Space Telescope observed 75 galaxies as part of its SINGS (Spitzer Infrared Nearby Galaxies Survey) Legacy Program. The galaxies are presented here in a Hubble Tuning-Fork diagram, which groups galaxies according to the morphology of their nuclei and spiral arms. The designation of these galaxies and their placement in the diagram is based on their visible-light appearance. The main goal of the SINGS program is to characterize the infrared properties of a wide range of galaxy types. The images of the galaxies are composites created from data taken by IRAC (the Infrared Array Camera) at 3.6 and 8.0 µm, and MIPS (the Multiband Imaging Photometer for Spitzer) at 24 µm.

The infrared range probed by these and other observations taken for the SINGS project allows for the detailed study of star formation, dust emission, and the distribution of stars in each galaxy. Light from old stars appears as blue in the images, while the lumpy knots of green and red light are produced by dust clouds surrounding newly born stars. The elliptical galaxies on the left are almost entirely made of old stars, while spiral galaxies like our own Milky Way are rich in young stars and the raw materials for future star formation.

More information can be found at:
http://sings.stsci.edu/

Poster and composite images created from SINGS observations by Karl D. Gordon (Oct 2007)
Blue=IRAC 3.6µm (stars)
Green=IRAC 8µm
(aromatic features from dust grains/molecules)
Red=MIPS 24µm (warm dust)

SINGS Team
Robert Kennicutt, Jr. (Principle Investigator), Daniela Calzetti (Deputy Principle Investigator), Charles Engelbracht (Technical Contact), Lee Armus, George Bendo, Caroline Bot, Brent Buckalew, John Cannon, Daniel Dale, Bruce Draine, Karl Gordon, Albert Grauer, David Hollenbach, Tom Jarrett, Lisa Kewley, Claus Leitherer, Aigen Li, Sangeeta Malhotra, Martin Meyer, John Moustakas, Eric Murphy, Michael Regan, George Rieke, Marcia Rieke, Helene Roussel, Kartik Sheth, J.D. Smith, Michele Thornley, Fabian Walter & George Helou

Figure 12.190: **Apparent stars speed anomaly in galaxies rotations**

Many galaxies are thought to have Black Holes at their active centers.

It seems that there is between $2 \cdot 10^{11}$ galaxies in the observable Universe following the actual estimates in this early 121st century (holocene calendar). Most of the galaxies are $1,000$ to $100,000$ parsecs in diameter and usually separated by distances on the order of millions of parsecs (or megaparsecs). The space between galaxies is filled with a tenuous gas having an average density of less than one atom per cubic meter. The majority of galaxies are gravitationally organized into associations known as galaxy groups, clusters, and superclusters. At the largest scale, these associations are generally arranged into sheets and filaments surrounded by immense voids.

---

[36] Since June 12025 (holocene calendar) it is known thanks to FRB (Fast Radio Bursts) that approximately 76% of the Universe's baryonic matter lies in the IGM (InterGalactic Medium). About 15% resides in galaxy halos, and a small fraction is burrowed in stars or amid cold galactic galactic gas ([109]).

Galaxy	**Type**	**Distance (10^3 [LY])**	**Diameter (10^3 [LY])**	**Mass (10^9 [$M_\odot$])**	**Abs. Magn.**	**Radial speed [$\text{km}\cdot\text{s}^{-1}$]**
Milky Way	Sb	-	100	150	-20	-
Large Magellanic Cloud	Irr I	170	23	10	-18.5	$+270$
Small Magellanic Cloud (NGC 292)	Irr I	200	10	20	-16.8	$+170$
Andromeda (NGC 224)	Sb	$2,250$	160	300	-21.1	-275
NGC 221	E 2	$2,150$	3	3	-16.4	-210
NGC 205	E 1	$2,100$	6	10	-16.4	-240
Triangulum Galaxy (NGC 598)	Sc	$2,250$	26	10	-18.9	-190
NGC 147	E 5	$2,150$	3	1	-14.9	-250
NGC 185	E 5	$2,150$	3	1	-15.2	-300
IC 1613	Irr I	$2,400$	3	0.3	-14.8	-240
NGC 6822	Irr I	$1,500$	6	0.4	-15.7	-40
Sculptor Galaxy (NGC 253)	E	280	3	0.003	-11.7	
Fornax Dwarf	E	550	6	0.02	-13.6	$+40$
Leo I	E 4	750	3	0.003	-11.0	
Leo II	E 1	750	3	0.001	-9.4	
Ursa Minor Dwarf	dwarf	220	3	0.0001	-8.8	
M82 of Ursa Major (NGC 3034)	Irr II	10	23	30	-19.5	$+400$
M81 of Ursa Major (NGC 3031)	Sb	10	100	200	-21.0	$+80$
M51 Whirlpool Galaxy (NGC 5194)	Sc	13	65	80	-19.7	$+550$
NGC 5128	E 0p	16	30	$1,000$	-20.0	$+260$
M101 Pinwheel Galaxy (NGC 5457)	Sc	20	200	300	-20.0	$+400.00$
M83 Southern Pinwheel Galaxy (NGC 5236)	SBc	26	300	$1,000$	-20.5	$+320.0$
M104 Sombrero Galaxy (NGC 4594)	Sa	40	30	500	-22.0	$+1,050$
M87 Virgo A (NGC 4486)	E 1	50	40	300	-22.0	$+1,220$

Table 12.21: **Data table of some well known galaxies**

Our galaxy in the cosmological rest frame is for information estimated (thanks to the Cosmological Microwave Background dipole) actually accordting to [9] at: $384\,[\text{km}\cdot\text{s}^{-1}]\pm 78\,[\text{km}\cdot\text{s}^{-1}]$ (stat.) $\pm 115\,[\text{km}\cdot\text{s}^{-1}]$ (syst.).

A small mosaic of some galaxies (without their names) to see that their shape can vary a lot :

Figure 12.191: **Gallery of some galaxies**

Size comparisons of some famous galaxies:

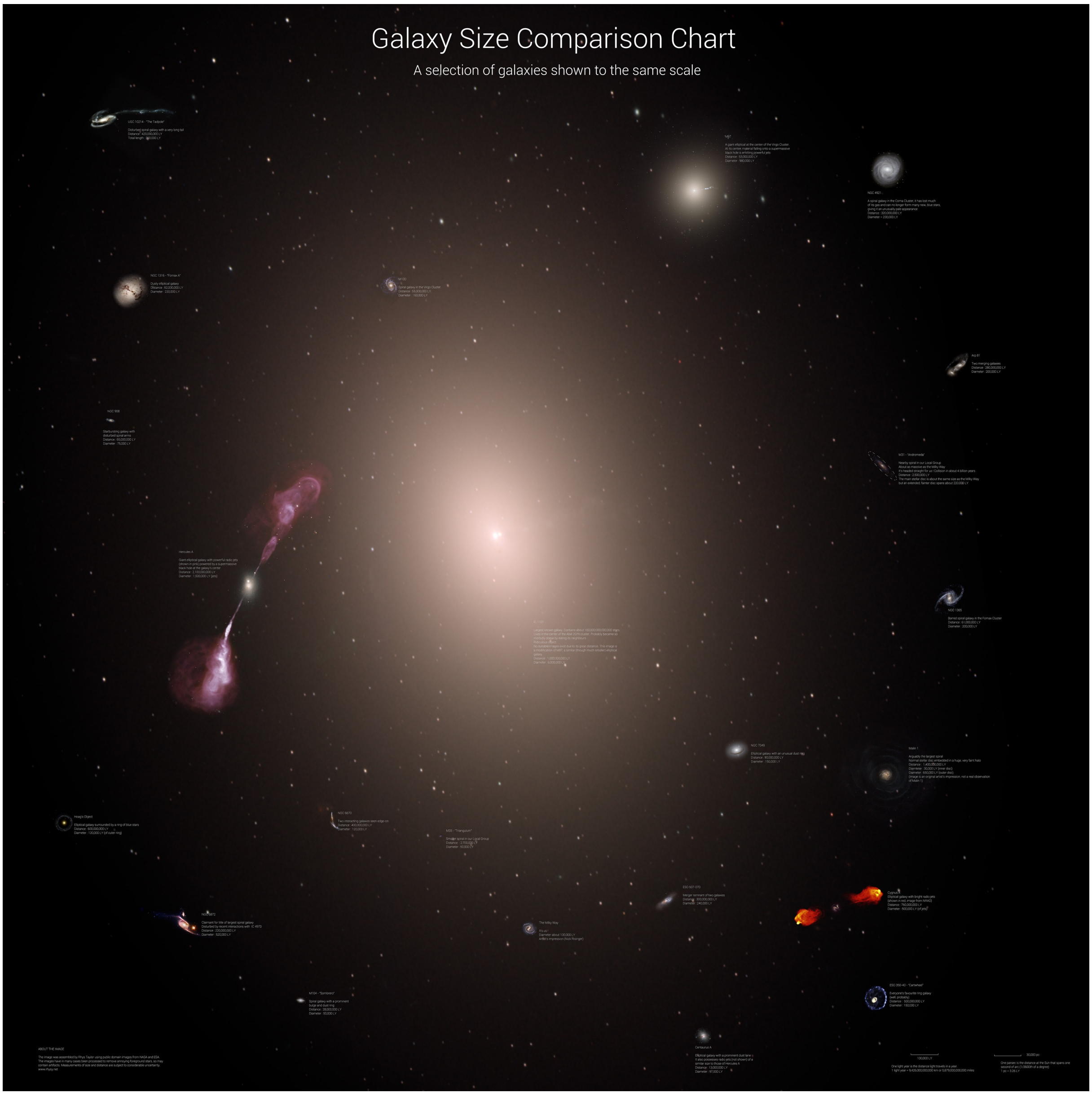

Galaxy Size Comparison Chart
A selection of galaxies shown to the same scale
UGC 10214 - "The Tadpole"
Disturbed spiral galaxy with a very long tail
Distance : 420,000,000 LY
Total length : 390,000 LY
M87
A giant elliptical at the center of the Virgo Cluster.
At its center, material falling onto a supermassive
black hole is emitting powerful jets
Distance : 53,000,000 LY
Diameter : 980,000 LY
NGC 4921
A spiral galaxy in the Coma Cluster, it has lost much
of its gas and can no longer form many new, blue stars,
giving it an unusually pale appearance
Distance : 320,000,000 LY
Diameter = 230,000 LY
NGC 1316 - "Fornax A"
Dusty elliptical galaxy
Distance : 62,000,000 LY
Diameter : 220,000 LY
M100
Spiral galaxy in the Virgo Cluster
Distance : 55,000,000 LY
Diameter : 160,000 LY
Arp 81
Two merging galaxies
Distance : 280,000,000 LY
Diameter : 200,000 LY
NGC 908
Starbursting galaxy with
disturbed spiral arms
Distance : 65,000,000 LY
Diameter : 75,000 LY
M31 - "Andromeda"
Nearby spiral in our Local Group
About as massive as the Milky Way
It's headed straight for us ! Collision in about 4 billion years
Distance : 2,500,000 LY
The main stellar disc is about the same size as the Milky Way
but an extended, fainter disc spans about 220,000 LY
Hercules A
Giant elliptical galaxy with powerful radio jets
(shown in pink) powered by a supermassive
black hole at the galaxy's center
Distance : 2,100,000,000 LY
Diameter : 1,500,000 LY (jets)
IC 1101
Largest known galaxy. Contains about 100,000,000,000,000 stars.
Lives in the center of the Abel 2029 cluster. Probably became so
morbidly obese by eating its neighbours.
Ridiculous object.
No suitable images exist due to its great distance. This image is
a modification of M87, a similar (though much smaller) elliptical
galaxy.
Distance : 1,000,000,000 LY
Diameter : 6,000,000 LY
NGC 1365
Barred spiral galaxy in the Fornax Cluster
Distance : 61,000,000 LY
Diameter : 200,000 LY
NGC 7049
Elliptical galaxy with an unusual dust ring
Distance : 90,000,000 LY
Diameter : 150,000 LY
Malin 1
Arguably the largest spiral
Normal stellar disc embedded in a huge, very faint halo
Distance : 1,400,000,000 LY
Diamteter : 30,000 LY (inner disc)
Diameter : 650,000 LY (outer disc)
(image is an original artist's impression, not a real observation
of Malin 1)
Hoag's Object
Elliptical galaxy surrounded by a ring of blue stars
Distance : 600,000,000 LY
Diameter : 120,000 LY (of outer ring)
NGC 6670
Two interacting galaxies seen edge-on
Distance : 400,000,000 LY
Diameter : 120,000 LY
M33 - "Triangulum"
Smaller spiral in our Local Group
Distance : 2,700,000 LY
Diameter : 50,000 LY
ESO 507-070
Merger remnant of two galaxies
Distance : 300,000,000 LY
Diameter : 240,000 LY
Cygnus A
Elliptical galaxy with bright radio jets
(shown in red, image from NRAO)
Distance : 760,000,000 LY
Diameter : 500,000 LY (of jets)
NGC 6872
Claimant for title of largest spiral galaxy
Disturbed by recent interactions with IC 4970
Distance : 220,000,000 LY
Diameter : 520,000 LY
The Milky Way
It's us !
Diameter about 100,000 LY
Artist's impression (Nick Risinger)
M104 - "Sombrero"
Spiral galaxy with a prominent
bulge and dust ring
Distance : 28,000,000 LY
Diameter : 50,000 LY
ESO 350-40 - "Cartwheel"
Everyone's favourite ring galaxy
(well, probably)
Distance : 500,000,000 LY
Diameter : 150,000 LY
Centaurus A
Elliptical galaxy with a prominent dust lane
It also possesses radio jets (not shown) of a
similar size to those of Hercules A
Distance : 13,000,000 LY
Diameter : 97,000 LY
ABOUT THE IMAGE
The image was assembled by Rhys Taylor using public domain images from NASA and ESA.
The images have in many cases been processed to remove annoying foreground stars, so may
contain artifacts. Measurements of size and distance are subject to considerable uncertainty.
www.rhysy.net
100,000 LY
One light year is the distance light travels in a year.
1 light year = 9,426,000,000,000 km or 5,879,000,000,000 miles
30,000 pc
One parsec is the distance at the Sun that spans one
second of arc (1/3600th of a degree)
1 pc = 3.26 LY

Figure 12.192: Size comparisons of famous galaxies (author: Rhys Taylor)

12.2.2.1 Probability of collisions in merging galaxies

Stars collide with each other very rarely. The distance between neighbouring stars (at our position in the Milky Way Galaxy) is approximately equal to 10 million times the diameter of a star. By contrast, galaxies collide with each other quite frequently. The distance between neighbouring galaxies is approximately equal to 20 times the diameter of a galaxy.

To illustrate this difference, consider building a scale model of our Galaxy in which the stars are represented by ping-pong balls. In this model, the distance between the Sun and Alpha Centauri will be 1, 100 kilometres (the distance between Columbus and Jacksonville, Florida). Now consider a scale model of the universe in which individual galaxies are represented by ping-pong balls. In this model, the Milky Way Galaxy and the Andromeda Galaxy will be a pair of ping-pong balls only 1 meter apart.

Let's see what we get from some back-of-the-envelope estimates.

Imagine throwing one star (e.g., the Sun) at the other galaxy. How likely is it we'll hit a star in the other galaxy? Well, it's basically proportional to how big a target each star in the other galaxy is (its cross-sectional area) compared to the size of the whole galaxy, multiplied by the total number of stars in the target galaxy.

Let's assume it's the Milky Way-Andromeda scenario, so each galaxy has about 100 billion stars, and each star is roughly the same size as the Sun (some are much larger, most are smaller). The actual target area for an individual star is a circle with twice the star's radius (we're counting one star just grazing the other as a collision). Let's also assume the stars are more or less evenly distributed in a circular disk. Since "$100,000$ light years" is a common (and not completely crazy) estimate of the Milky Way's size, that's a circle of radius $=50,000$ light years (about 10^{16} meters).

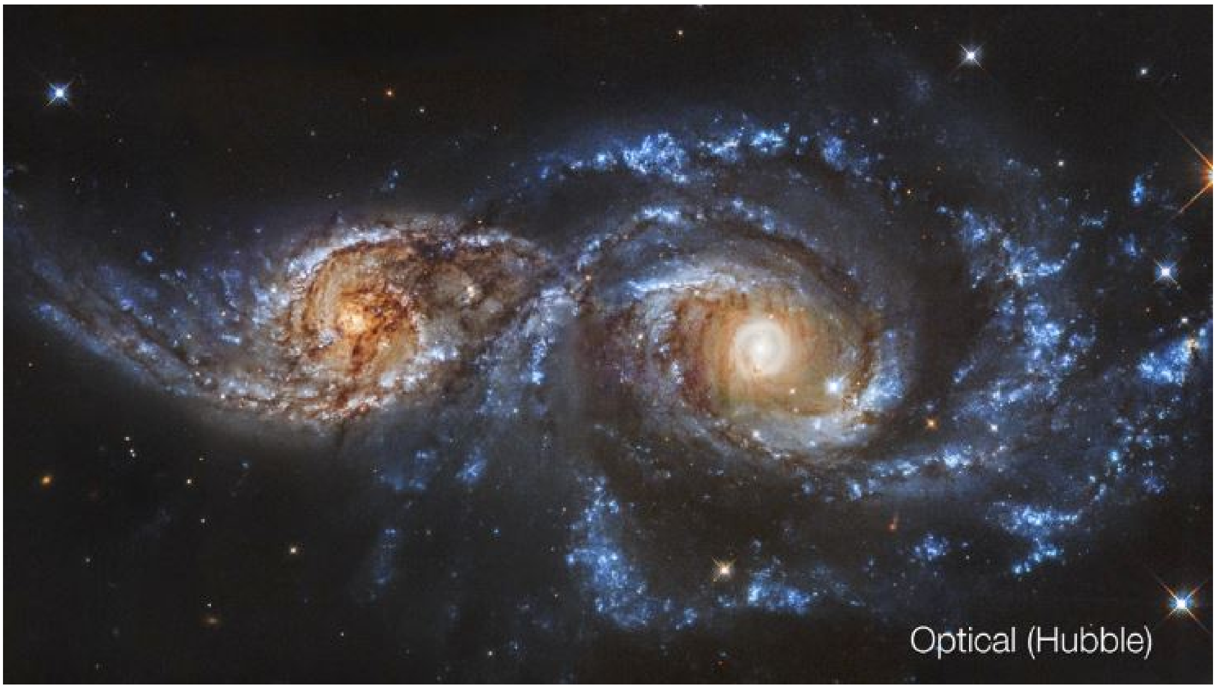


Figure 12.193: **NGC 2207 is a pair of colliding spiral galaxies (source: NASA/Hubble)**

So, let us consider two galaxies as perfect discs of 100 billion (10^{11}) stars uniformly distributed in the disc going in the target galaxy, each with target average radius $\sim 2R_\odot$, gives us a total target area of:

$$S_{T,\text{stars}} \cong 10^{11}\cdot\pi(2R_\odot)^2 \approx 10^{30}\;[\text{m}^2] \qquad (12.2.221)$$

The cross-section area of the target galaxy is:

$$S_{\text{gal.}} \cong \pi R_{\text{gal.}}^2 \approx 10^{42}\;[\text{m}^2] \qquad (12.2.222)$$

Therefore the probability of a typical average star hitting another star in the other galaxy is :

$$P_{\text{coll.}} \cong \frac{10^{30}}{10^{42}} = 10^{-12} \qquad (12.2.223)$$

or about one in a trillion.

The odds of any star from our Galaxy not hitting a star in the other galaxy would be:

$$(1-10^{-12})^{10^{11}} \cong 0.90 \tag{12.2.224}$$

So there's only about a 10% chance of one (or more) of the galaxy's 100 billion stars hitting a star in the other galaxy. And the chances of any one particular star (like our Sun) hitting a star in the other galaxy is about one in a trillion.

Remark

Obviously we could use the binomial law to get the probability (or cumulated probability depending on what we want) that k among N start would get hit:

$$\mathcal{B}(N,k)=\binom{N}{k}p^k q^{N-k}=\frac{N!}{k!(N-k)!}p^k q^{N-k} \tag{12.2.225}$$

or as the probabilities are very small and the population huge we could even use a Poisson distribution instead!

12.2.2.2 Radial Speed Anomaly

In 11978 (holocene calendar), Vera Rubin begins to observe that in galaxies, more the stars are distant from the galactic core, the more their angular velocity is high... The initial observation that uniformity of speed was unexpected because the theory of gravity Newton predicted that more distant objects have less speed. For example, the planets of the Solar System orbit with a respective speed decreases while growing their respective distance from the Sun. We are left with the same problem: how to explain a point measurement is greater than the theoretical value?

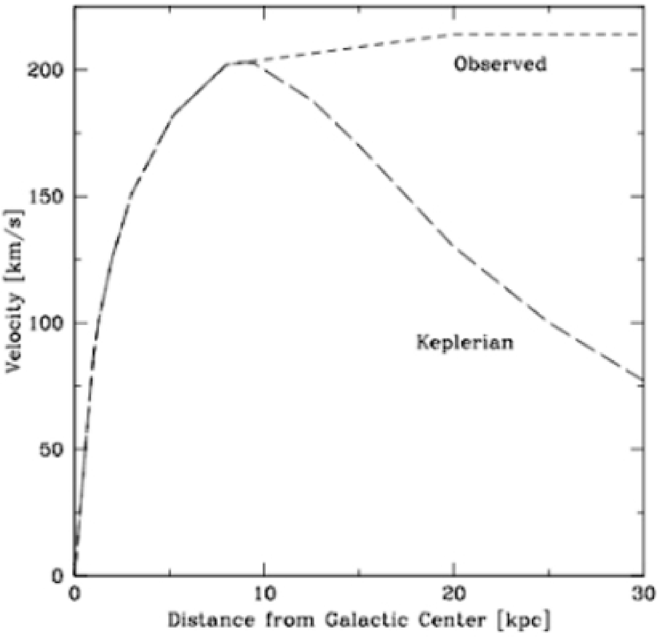


Figure 12.194: Apparent stars speed anomaly in galaxies rotations

According to Newton's laws, in a circular path, there is a as we know balance between the centripetal acceleration and gravitational attraction:

$$a=\frac{v^2}{r}=\frac{GM}{r^2} \tag{12.2.226}$$

The volume of a disk galaxy of radius R and thickness e is (see section Geometric Shapes page 2959):

$$V=\pi R^2 e \tag{12.2.227}$$

If we consider the mass of the galaxy almost entirely within the radius R_{max}, corresponding to the maximum speed, of density ρ is given then by:

$$M=\pi R_{\text{max}}^2 e\rho \tag{12.2.228}$$

Making the approximation that the mass is substantially within the range corresponding to the maximum speed, we can write:

$$r\cong R_{\text{max}}=\sqrt{\frac{M}{\pi e\rho}} \tag{12.2.229}$$

Which, introduced into the first equation but rearranged:

$$v=\sqrt{\frac{GM}{r}} \tag{12.2.230}$$

gives:

$$v=\sqrt{\frac{GM}{\sqrt{\dfrac{M}{\pi e\rho}}}}=\sqrt{G}\,(M\pi e\rho)^{1/4} \tag{12.2.231}$$

the law that the maximum speed varies as the $1/4$ power of the mass. After that the speed decrease of the stars should decrease following:

$$v=\sqrt{\frac{GM}{r}} \tag{12.2.232}$$

But we must keep in mind that this is a two body relation! In the facts a galaxy should be considered as an isotropic fluid (like the rest of the universe) and therefore it is quite normal that the two-body assumption does not suit the observations. A galaxy can also not be consider as a solid cylinder otherwise by applying $v=\omega r$ the speed of the stars should increase in proportion to the distance to the center of the galaxy. For a more rigorous treatment of the topic the reader should refer to the page 5284.

~90 % ★★★★☆

28 votes, 80.71%

12.3. Special Relativity

We have always considered until now in all our developments that the interactions (causes and effects) between objects were instantaneous and that the observation of a phenomenon took place therefore instantly after it had occurred. However two physicists, Michelson and Morley (see [390]), during an experiment in 11887 (holocene calendar) discovered something that would change radically all of classical physics: the velocity (speed) of light was invariant (constant) regardless of the movement that we had relatively to it (and another type of experiment that leads to the same conclusion was done later by De Sitter as described in his paper [133] among many others)!

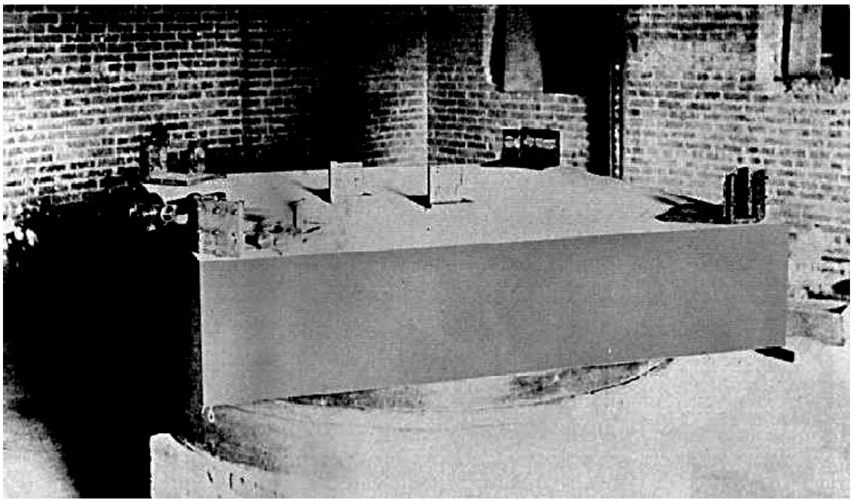


Figure 12.195: Michelson and Morley's original interferometric setup, mounted on a stone slab that floats in an annular trough of mercury (source: Case Western Reserve University)

This observation is even more important that we know it is mainly the light that allows us to perceive and feel things around us! It should also be taken into consideration that the electrostatic and magnetic fields are, as we have seen in the section of Quantum Field Theory, carried by the vector of interaction that is the photon that moves at the finite speed of light denoted by c. This fact also allows us to assume that the gravitational field also has an interaction vector (which would be the "graviton" whose existence seems indirectly proven but not yet confirmed still in this early 121st century according to holocene calendar) that propagates at the speed of light. It is therefore appropriate to take into account this non-immediacy and the consequences that this entails in the observed phenomena to finally be able to decide what is really of what seems to be.

Before we start with the calculations, we need to define a little bit what will be studied in this section (which applies not only to cosmology but... it seemed to us better to put it in this chapter rather than in the chapter of Mechanics or Atomistic).

Definition (#12.3.540): "Special Relativity" is a scientific theory confined to isolated inertial frames (Galileans), that is to say, the study of animated frames in a uniform (inertial) rectilinear motion. The reason of this will be given in the statement of Special Relativity principle (see below).

Remarks

R1. Restrict the study to inertial frames of course does not does not prohibit that within these, bodies can be animated of a uniform speed or not (an inertial rocket can have bodies inside itself that have a non-uniform movement)!

R2. General Relativity's purpose (see corresponding section at page 5720) is to take into account non-inertial frames and in any coordinate system by making use of the power of the tensor calculus to be applicable in any type of space (other than flat one!).

Special Relativity is mainly based on three important concepts:

1. The invariance postulate of speed of light
2. The cosmological principle (see below)
3. The principle of Special Relativity (see below)

It is also important to inform the reader that we will use here many concepts seen in the section of Linear Algebra, Tensor Calculus, Trigonometry, Analytical Mechanics, Classical Mechanics, Electrostatics, Magnetostatics and Electrodynamics. It is therefore strongly advised to have covered these topics before at risk of not understanding what follows.

12.3.1 Assumptions and Principles

Physics laws express relations between the fundamental physical quantities. If the laws of physics are invariant under Galilean referential change as we have seen in the section of Classical Mechanics, it is not necessarily the same for physical quantities! These can transform from Galilean frame to another according to simple transformation law as we have seen in the section of Classical Mechanics for velocity for example. It is the same in Special Relativity, but we must now consider what we neglected in our study of Galileo's transformations: the time lag is not the same for two observers if the speed of the light is finite, but the concept of time interval is supposed to be kept invariant!

12.3.1.1 Postulate of Invariance

Laboratory measurements (Michelson-Morley experiment as we have already mentioned) have, for a long time, shown that the speed c measured in an inertial frame (straight line and at constant speed) is constant regardless of its speed. Then we are taken to state the postulate of invariance of light: the speed of light (vector for the transport of photons) can neither be added nor subtracted, to the drive speed of the frame in which we measure it (more clearly it means that no matter how fast you move, you will always measure the speed of light as being numerically finite and equal to $c=299,792,458\,[\text{km}\cdot\text{s}^{-1}]$!).

As corollary the principle of Galilean relativity (see section Classical Mechanics page 3239) according to this premise is completely at fault and then we have to develop a new theory that takes into account of this property of light.

Remark

It is important to notice that we consider that light, within the framework of Special Relativity, the messenger of information from one body to another, that is to say the speed of causality effect!!!

On Internet Forums a common and interesting question is: *Why is there a limit speed for causality and for the speed of light?*

First, if the light speed would be infinite, we would not have light at all! Indeed, to see this, take a look at Maxwell's equations again, Note that in them we have the factor:

$$c=\frac{1}{\sqrt{\mu_0\varepsilon_0}} \tag{12.3.1}$$

so if we set $c\rightarrow+\infty$ then either (or both) of μ_0 and ε_0 would have to be zero. This will effectively kill the existence of dynamic magnetic fields.

Especially, for light, it means that:

$$\vec{\nabla}\times\vec{B}=\frac{1}{c^2}\frac{\partial\vec{E}}{\partial t}\rightarrow 0 \tag{12.3.2}$$

So, magnetic fields would be static (and of zero intensity, remember no magnetic monopoles known so far!). Thus the only thing left of electromagnetism would be simply electrostatics. Physically this also make sense, if $c\rightarrow+\infty$ the electric field response to any rearrangement of charges would be instantaneous, so there is no place (time?) for a magnetic field response.

The existence of the speed limit for causality is related to the existence of time itself. Indeed, if there would be no speed limit, everything would happen instantly (as well as distance and consequently space by the way). It is difficult to imagine a Universe that evolve instantaneously. This is why we belong to a Universe where there is a speed limit otherwise we would very likely not be here to observe it...

12.3.1.2 Cosmological Principle

We assume that our position in the Universe is typical not only in space as stated in the standard model of the Universe (see section Cosmogony page 5881), but also in time. Thus, an astronomer located in a remote galaxy must observe the same general properties of the Universe that we do whatever he lived a billion years ago or that he observed the Universe in a billion years in the future!

In fact, it is quite natural to go further and state that: the Universe looks the same in every point, that is to say, it is homogeneous. This homogeneity is therefore sets as the "Cosmological Principle".

This principle is not based, still actually in this early 121st century (holocene calendar), on experimental data because they are too much fragmentary compared to the huge size of the cosmos (very likely infinite). Therefore we are not able yet to establish its validity (there are two publications based on a huge amount of observations of our Hubble sphere and one of the both paper is a preprint that contradicts the conclusion of the first one... see [522] and [480]). It constitutes however a presupposition for any physical study of our Universe. Its purpose is relative to its character, essential to any scientific cosmology study, and perhaps to a certain reaction to the geocentric or heliocentric old vision: it is assumed now that no place is special in the cosmos!

For an in-deep discussion and criticism of the subject based on the earliest experimental data (of this early 121st century according to holocene calendar), the reader can refer to the excellent 100 pages paper *Is the observable Universe consistent with the cosmological principle?* (see [15]).

12.3.1.3 Special Relativity Principle

Let us remind (see section Classical Mechanics page 3239) that the Galilean transformations tell us that no reference frames can be considered as an absolute frame because the relations between the physical quantities are identical in all Galilean repositories ("Galilean relativity principle"). The Galilean motion is therefore relative!

In the 120th century (holocene calendar) physicists noted that an important class of physical phenomena violated the Galilean relativity principle: the electromagnetic phenomena!

By applying the Galilean transformations to Maxwell's equations, we get a different set of equations depending on whether the observer is in a fixed reference or a mobile reference frame.

Indeed, we have proved in the section on Electrodynamics that the electric or magnetic field propagation equation could be written in one-dimensional space as the following d'Alembert equation (see page 4118):

$$\left(\frac{\partial^2}{\partial x^2}-\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)\psi=0 \tag{12.3.3}$$

where ψ represents any one of the two fields (electric or magnetic). We name this relation sometimes "Hertz equation".

We also saw in the section of Classical Mechanics (see page 3239) that an important factor in the validity of a theory is the invariance of the expression of its laws under a Galilean transformation by putting:

$$x'=x-u_xt,y'=y,z'=z,t'=t \tag{12.3.4}$$

We have also shown in the section of Differential and Integral Calculus (see page 1466) that the total differential of a function was written (example with two variables):

$$\mathrm{d}f(x',y')=\frac{\partial}{\partial x'}\mathrm{d}x'+\frac{\partial}{\partial y'}\mathrm{d}y' \tag{12.3.5}$$

Therefore:

$$\mathrm{d}f(x',t')=\frac{\partial f}{\partial x'}\mathrm{d}x'+\frac{\partial f}{\partial t'}\mathrm{d}t' \tag{12.3.6}$$

Which brings us to simply write (using the physicist method way of life...):

$$\frac{\partial f}{\partial x}=\frac{1}{\partial x}\frac{\partial f}{\partial x'}\mathrm{d}x'+\frac{1}{\partial x}\frac{\partial f}{\partial t'}\mathrm{d}t' \tag{12.3.7}$$

After elimination of f and using the Schwarz theorem (see section Differential and Integral Calculus page 1469) and still the physicist way of life:

$$\begin{aligned}\frac{\partial}{\partial x}&=\frac{1}{\partial x}\frac{\partial}{\partial x'}\mathrm{d}x'+\frac{1}{\partial x}\frac{\partial}{\partial t'}\mathrm{d}t'=\frac{\partial}{\partial x'}\frac{\partial x'}{\partial x}+\frac{\partial}{\partial t'}\frac{\partial t'}{\partial t}\\&=\frac{\partial}{\partial x'}\frac{\partial(x-u_xt)}{\partial x}+\frac{\partial}{\partial t'}\frac{\partial t'}{\partial t}=\frac{\partial}{\partial x'}1+\frac{\partial}{\partial t'}0=\frac{\partial}{\partial x'}\end{aligned} \tag{12.3.8}$$

If we write the same with the time variable:

$$\begin{aligned}\frac{\partial}{\partial t}&=\frac{1}{\partial t}\frac{\partial}{\partial x'}\mathrm{d}x'+\frac{1}{\partial t}\frac{\partial}{\partial t'}\mathrm{d}t'=\frac{\partial}{\partial x'}\frac{\partial x'}{\partial t}+\frac{\partial}{\partial t'}\frac{\partial t'}{\partial t}\\&=\frac{\partial}{\partial x'}\frac{\partial(x-u_xt)}{\partial t}+\frac{\partial}{\partial t}\frac{\partial t'}{\partial t'}=\frac{\partial}{\partial x'}(-u_x)+\frac{\partial}{\partial t'}1=-u_x\frac{\partial}{\partial x'}+\frac{\partial}{\partial t'}\end{aligned} \tag{12.3.9}$$

Ultimately the Galilean transformation of the wave equation supposedly have an invariant form becomes:

$$\left(\frac{\partial^2}{\partial x'^2}-\frac{1}{c^2}\left(-u_x\frac{\partial}{\partial x'}+\frac{\partial}{\partial t'}\right)^2\right)\psi=\left(\frac{\partial^2}{\partial x'^2}-\frac{u_x^2}{c^2}\frac{\partial^2}{\partial x'^2}+2\frac{u_x}{c^2}\frac{\partial^2}{\partial x'\partial t}-\frac{1}{c^2}\frac{\partial^2}{\partial t'^2}\right)\psi=0 \tag{12.3.10}$$

To fix the situation, following this example, we can state at least three assumptions:

H1. Maxwell's equations are false. The correct equations remain to be discovered and must be invariant under a Galilean transformation.

H2. Galilean invariance is valid for mechanics but not for electromagnetism (that was the historical solution before Albert Einstein: an "ether" determines the existence of a kind of absolute repository where Maxwell's equations didn't change).

H3. Galilean invariance is false. There is a more general invariance, it remains to be discovered, which preserves the form of the Maxwell equations. Classical mechanics is to be reformulated so that it is invariant under this new transformation.

Remark

It turns out that the first two assumptions are excluded by the experimental facts. Moreover, Maxwell's equations integrating the speed of light they are implicitly relativistic.

Albert Einstein did not accept the violation of the Galilean relativity principle by electromagnetism. From his perspective, it was necessary to generalize it to all natural laws.

He postulated that the laws of physics should be the same in all repositories Galileans, which means, implicitly, that in the point of view of physical laws, it is not possible to distinguish one from another Galilean frame. This result is most commonly formulated as: no reference is privileged. This principle was named "principle of relativity". Indeed, this relativity is restricted to the case of Galileans frames (also named "inertial frames") exclusively.

In other words, the physic laws should remain unchanged after a change of reference. We must therefore identify new adequate transformations that will substitute to the Galilean transformations.

In the case of non Galileans frames repositories are not indistinguishable anymore. Indeed, imagine a person in a train moving at a constant speed and another person on land. Everyone can then say that it is the other who is in motion (relative) and indiscriminately. By cons, if the train begins to accelerate, although the two individuals can say that this is another speeding, only the one on the train will feel the effect of this acceleration ... and repositories are no more indistinguishable.

Albert Einstein abolished as well as the idea that there is an absolute reference point that does not move and on which we can define an absolute time, an absolute length or absolute mass. However, one can define a privileged reference point for every object in the Universe. It is the frame moving at the same speed and in the same direction as the object. The time measured in this privileged reference frame is minimal and is named the "proper time". Similarly, the size of the object is maximum, it is his "proper dimension" or "proper distance", and its mass is minimal, it is its "proper mass" (we will do the corresponding detailed mathematical developments further below).

Remark

Albert Einstein did share the credit with Hendrik Lorentz and Henri Poincaré for Special Relativity for a while, probably one reason his Nobel prize did not mention relativity. Because Lorentz in 11904 according to holocene calendar (who misinterpreted his own transformations as applying to the Ether) laid the fundamentals for the work by Einstein, this theory was originally named the "Lorentz–Einstein theory". Poincaré struggled with his own proof of $E=3/4mc^2$ and improved Lorentz transformation formulation (June 5, 11905 according to holocene calendar). Einstein at the very least gets credit for being the first to correctly interpret the equations and do away with the Ether concept and get the correct relation $E=mc^2$ (June 30, 11905 according to holocene calendar).

12.3.2 Lorentz Transformations/Boost

For make possible to c to be invariant (light speed invariance postulate), we must admit that time appears to not flow the same way for the observer O that is motionless than for the observer O$'$ in a reference frame in uniform translation (i.e.: an inertial frame) in the direction of x with relative velocity (the term "relative" is important!) v (caution! the relative speed between the repositories is often denoted in the literature by u).

Remark

A special case of disposal of referential frames in which the space axes are parallel leads to what we name the "pure Lorentz transformations" or "special Lorentz transformations" and the relative displacement along a particular axis is often named a "boost".

To study the behaviour of physic laws, we must bring two clocks that give the times t and t' (the referential frame that contains its clock/measuring instrument is named "proper referential" or "proper frame").

Let's set up the following imaginary experiment:

When the observers O and O' are superimposed, we set $t=0$ and $t'=0$ (clock time sync) and we emit a bright flash[37] in the direction of a point A spotted by respectively $\vec{r}$ and $\vec{r}'$:

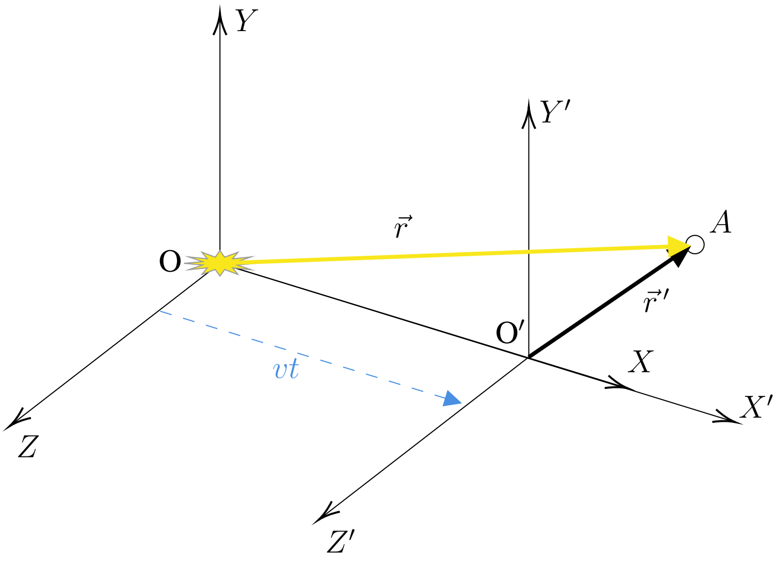


Figure 12.196: **Configuration for the study of relativistic effects**

It is obvious that when the flash arrives in A, the observer O will measure a time t and O$'$ a time t'.

The observer O therefore concludes:

$$r=ct=\sqrt{x^2+y^2+z^2}\Rightarrow x^2+y^2+z^2=c^2t^2 \tag{12.3.11}$$

---

[37] In fact we should consider the emission of "an element of information". Using light as an example is just convenient for pedagogical purposes. As we will see the results we will get further below involve a speed limit c that is for sure the speed of light, but in reality we should consider light as a special case of the maximum possible speed of information transfer. The "c" can then be seen as the "causality speed".

The observer O$'$ therefore concludes:

$$r'=ct'=\sqrt{x'^2+y'^2+z'^2}\Rightarrow x'^2+y'^2+z'^2=c^2t'^2 \tag{12.3.12}$$

Since the displacement of O$'$ is made only along the Ox axis, we have for the two observers:

$$y=y',z=z' \tag{12.3.13}$$

Moreover, if the path of the light beam coincides within the axe Ox, we have:

$$y=y'=0,z=z'=0 \tag{12.3.14}$$

This gives us then:

$$x^2=c^2t^2,x'^2=c^2t'^2 \tag{12.3.15}$$

And therefore:

$$x^2-c^2t^2=0,x'^2-c^2t'^2=0 \tag{12.3.16}$$

these two relations are equal (zero) at any x,x',t,t' between the two observers. These are the first "relativistic invariant" (equal values regardless of the frame), especially here is about the "invariant interval" also named "space-time interval", that we find in a more generalized form when applied to the whole space:

$$\boxed{x^2+y^2+z^2-c^2t^2=x'^2+y'^2+z'^2-c^2t'^2=0} \tag{12.3.17}$$

Now it should be remembered that in the classical point of view (Galilean relativity), we would have written that the position of point A for the observer O from the information given by O$'$ would be given by $x=x'+vt$ and vice versa (see section Classical Mechanics page 3239) such as:

$$x'=x-vt'=x-vt \tag{12.3.18}$$

However.... in the relativistic model, we must admit that time t which is related to x is not the same as t' which is related to x', relativity principle oblige (otherwise it would be difficult to explain the invariance of the speed of light)!

We are then led to try to write the above relation as follows:

$$x'=\lambda(x-vt) \tag{12.3.19}$$

where λ would be a numerical value to be determined from a given algebraic expression. Because to explain the constancy of the speed of light one possibility is that the space must constantly adjust according to our velocity v. What is revolutionary hypothesis as we have already mentioned!

Remark

A reader asked us why we could not write the last relation in the following simplified form (using the relation $x=ct$ obtained above) where point A is on the X axis:

$$x'=\lambda(x-vt)=\lambda\left(x-v\frac{x}{c}\right)=\lambda x\left(1-\frac{v}{c}\right) \tag{12.3.20}$$

The only reason is that later we will introduce a vector (matrix) notation of this result showing the concept of quadrivector (four-vector) and that it is in the first form of writing (this making explicit reference to time) that we can clearly make the concept of space-time emerge.

Furthermore, if $t\neq t'$, we must also be able to express t' as a function of t and x in a similar way:

$$t'=a(t-bx) \tag{12.3.21}$$

Let us summarize the shape of the problem:

$$x'=\lambda(x-vt) \tag{12.3.22}$$

with λ to be determined and after:

$$t'=a(t-bx) \tag{12.3.23}$$

with a,b to be determined. We then seek to determine the relation that gives us to know the values of the coefficients a,b and λ that satisfy simultaneously:

$$x^2+y^2+z^2=c^2t^2 \qquad x'^2+y'^2+z'^2=c^2t'^2 \tag{12.3.24}$$

Remembering the previous developments and bearing in mind that in our special case $y'=y$ and $z'=z$, the last equation becomes:

$$\lambda^2(x-vt)^2+y^2+z^2=c^2a^2(t-bx)^2 \tag{12.3.25}$$

Let us distribute:

$$(\lambda^2-a^2b^2c^2)x^2-2(\lambda^2v-ba^2c^2)xt+y^2+z^2=\left(a^2-\frac{\lambda^2v^2}{c^2}\right)c^2t^2 \tag{12.3.26}$$

To satisfy the relation:

$$x^2+y^2+z^2=c^2t^2 \tag{12.3.27}$$

We must have:

$$\begin{aligned}\lambda^2-b^2a^2c^2&=1 \quad (1)\\ \lambda^2v-ba^2c^2&=0 \quad (2)\\ a^2-\frac{\lambda^2v^2}{c^2}&=1 \quad (3)\end{aligned} \tag{12.3.28}$$

It is easy to solve (2):

$$\lambda^2=\frac{ba^2c^2}{v} \tag{12.3.29}$$

We then introduce this result in (1) and (3) and we arrive at:

$$\begin{aligned} \lambda^2-a^2b^2c^2&=\frac{ba^2c^2}{v}(1-bv)=1 \quad (1') \\ a^2-\frac{\lambda^2v^2}{c^2}&=a^2(1-bv)=1 \qquad (2') \end{aligned} \tag{12.3.30}$$

If we divide (1') by (2'), we get:

$$\frac{c^2b}{v}=1\Rightarrow b=\frac{v}{c^2} \tag{12.3.31}$$

and introducing this latter result into the relation:

$$a^2(1-bv)=1 \tag{12.3.32}$$

we obtain the following remarkable result:

$$a=\frac{1}{\sqrt{1-\dfrac{v^2}{c^2}}} \tag{12.3.33}$$

That we frequently denote by:

$$\boxed{a=\frac{1}{\sqrt{1-\dfrac{v^2}{c^2}}}=\frac{1}{\sqrt{1-\beta^2}}} \tag{12.3.34}$$

and which we name "Michelson-Morley factor" with:

$$\boxed{\beta=\frac{v}{c}} \tag{12.3.35}$$

Also introducing:

$$b=\frac{v}{c^2} \tag{12.3.36}$$

in:

$$\lambda^2=\frac{ba^2c^2}{v} \tag{12.3.37}$$

we get:

$$\lambda=a=\frac{1}{\sqrt{1-\left(\dfrac{v}{c}\right)^2}} \tag{12.3.38}$$

Let us denote now (to comply with the traditional notations in this field):

$$\gamma=\frac{1}{\sqrt{1-\beta^2}} \tag{12.3.39}$$

named the "Lorentz factor" with therefore the parameter:

$$\beta=\frac{v}{c}\leq 1 \tag{12.3.40}$$

that doesn't have any name as far as we know...

Caution! The "Lorentz transformations" are sadly sometimes interpreted by scientifically illiterate people and undergraduate students as a physical transformation. But it is not! It's just a mathematical transformation to explain what one reference frame sees in comparison of another one. Nothing more!

12.3.2.1 Displacement four-vector

We just derived the "Lorentz transformation" relations to pass from the values measured by O$'$ to those measured in O and vice versa[38]:

$$\begin{aligned} x'&=\gamma(x-vt) & x&=\gamma(x'+vt')\\ y'&=y & y&=y'\\ z'&=z & z&=z'\\ t'&=\gamma\left(t-\frac{v}{c^2}x\right) & t&=\gamma\left(t'+\frac{v}{c^2}x'\right) \end{aligned} \tag{12.3.41}$$

That have for property to be covariant (that is to say their relations keep the same structure during a change of a Galilean reference system). We see through these relations that the concept of "time" is something individual relating to the movement we have over other (this is the "proper time"). This is why it is not possible to define a "common time" between two people moving relatively and who don't know their respective relative speed (and even here we do not take into account the gravity that distorts space-time ... that we will study in the section of General Relativity).

Remark

If v is much smaller than c, we fall back on the Galilean transformation as $\gamma\cong 1$ and $v/c^2\cong 0$

We can also write the last relations in a more useful way (the reader will notice that this time that for all relations the units of all the terms to the left of equality are the same: it is every time a distance!):

$$\begin{aligned} x&=\gamma(x'+\beta ct')\\ y&=y'\\ z&=z'\\ ct&=\gamma\left(ct'+\beta x'\right) \end{aligned} \tag{12.3.42}$$

---

[38] Such relations were first written by Woldemart Voigt in 11887 (holocene calendar). Between 11892 (holocene calendar) and 11902 (holocene calendar) by trial and errors Hendrik Lorentz will improve the shape of these relations. In 11905 (holocene calendar) Poincaré confirmed and corrected Lorent'z work of 11904 (holocene calendar) and give them their definitive shape.

Of course the difference is that the fourth dimension being the time coordinate of "space-time" seems at the contrary of the spatial coordinates to have a privileged direction: the "arrow of time" (you can not go back to a given moment time given in the reality - as least as far as we know today - when it is possible when we traverse a purely spatial distance). The direction of time is imposed by the second law of thermodynamics as entropy can only increase (see section Thermodynamics page 3658). If this were not the case then all time could already exist and we could travel in time as we travel on distances and therefore the future should be already written (people that believe in destiny like this...) and we could also go back in time.... However, thermodynamics does not give a particular direction to the time ... so if our time has the direction it has today.... it is because our universe was organized at its beginning (so it had a low entropy).

By proceeding in a homogenisation of units to be able to use more modern and generalized maths than just simple algebra we can see that in fact when we travel in time we travel in physical point of view a distance ct. But because c and t are measured in arbitrary human being units physicists prefer to put $c=1$ we mathematical development become more complicate rather than changing the definition of time.

Now we can put the Lorentz transformations of coordinate and time in the traditional following matrix form (see section Linear Algebra page 1924) which defines the "Lorentz matrix" or "Lorentz-Poincare matrix" or "Lorentz boost":

$$\begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix}=\begin{bmatrix} \gamma & \gamma\beta & 0 & 0 \\ \gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}\begin{pmatrix} ct' \\ x' \\ y' \\ z' \end{pmatrix} \tag{12.3.43}$$

and reciprocally (note that the time component ct is interchangeable with the space component x):

$$\begin{pmatrix} ct' \\ x' \\ y' \\ z' \end{pmatrix}=\begin{bmatrix} \gamma & -\gamma\beta & 0 & 0 \\ -\gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}\begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix} \tag{12.3.44}$$

and the reader can very easily control that with the two previous relations we fall back on:

$$\begin{aligned} x' &= \gamma(x-vt) & x &= \gamma(x'+vt') \\ y' &= y & y &= y' \\ z' &= z & z &= z' \\ t' &= \gamma\left(t-\frac{v}{c^2}x\right) & t &= \gamma\left(t'+\frac{v}{c^2}x'\right) \end{aligned} \tag{12.3.45}$$

We have also obviously:

$$\begin{aligned} \begin{bmatrix} \gamma & -\gamma\beta & 0 & 0 \\ -\gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} \gamma & \gamma\beta & 0 & 0 \\ \gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} &= \begin{bmatrix} \gamma^2-\gamma^2\beta^2 & 0 & 0 & 0 \\ 0 & \gamma^2-\gamma^2\beta^2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{aligned} \tag{12.3.46}$$

In index form the matrix formulation is written:

$$\begin{pmatrix} x'_0 \\ x'_1 \\ x'_2 \\ x'_3 \end{pmatrix} = \begin{bmatrix} \gamma & -\gamma\beta & 0 & 0 \\ -\gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{pmatrix} \tag{12.3.47}$$

which can therefore be written in tensor (see section Tensor Calculus page 2039) form:

$$\boxed{x'^{\mu} = L^{\mu}_{\nu} x^{\nu}} \qquad \text{and} \qquad \boxed{x^{\nu} = L'^{\nu}_{\mu} x'^{\mu}} \tag{12.3.48}$$

Remark

We can see the tensor (the matrix) of Lorentz transformation in some books in the condensed form $L(\beta)$ and sometimes L^{μ}_{ν} or even Λ^{μ}_{ν}.

The vector:

$$x^{\mu} = (ct, x, y, z) = (ct, \vec{x}) \tag{12.3.49}$$

is named "space-time four-vector" or "four-vector displacement".

Notice that since:

$$x^2 + y^2 + z^2 - c^2t^2 = x'^2 + y'^2 + z'^2 - c^2t'^2 \tag{12.3.50}$$

the transformation by the matrix L^{μ}_{ν} conserves the norm (Lorentz invariance). In geometric terms it is thus a "isometry" or an invariance of the dot product by Lorentz transformation.

Let us prove this explicitly following the request of a reader! We will take again for the proof only a movement along x and we use:

$$\begin{aligned} x' &= \gamma(x - vt) & x &= \gamma(x' + vt') \\ y' &= y & y &= y' \\ z' &= z & z &= z' \\ t' &= \gamma\left(t - \frac{v}{c^2}x\right) & t &= \gamma\left(t' + \frac{v}{c^2}x'\right) \end{aligned} \tag{12.3.51}$$

As $y' = y$ and $z' = z$ to simplify the development we will ignore these both components.

$$\begin{aligned} x' &= \gamma(x - vt) & x &= \gamma(x' + vt') \\ t' &= \gamma\left(t - \frac{v}{c^2}x\right) & t &= \gamma\left(t' + \frac{v}{c^2}x'\right) \end{aligned} \tag{12.3.52}$$

We will also put to simplify $c = 1$ and therefore v is expressed in percentage of c and becomes $v = \beta$:

$$\begin{aligned} x' &= \gamma(x - \beta t) & x &= \gamma(x' + \beta t') \\ t' &= \gamma(t - \beta x) & t &= \gamma(t' + \beta x') \end{aligned} \tag{12.3.53}$$

Now we calculate:

$$\begin{aligned} x'^2 - t'^2 &= \gamma^2(x - \beta t)^2 - \gamma^2(t - \beta x)^2 \\ &= \gamma^2\left((x - \beta t)^2 - (t - \beta x)^2\right) = \gamma^2\left(x^2 - 2\beta tx + \beta^2t^2 - t^2 + 2\beta xt - \beta^2x^2\right) \\ &= \gamma^2\left(x^2 + \beta^2t^2 - t^2 - \beta^2x^2\right) = \gamma^2(1 - \beta^2)x^2 - \gamma^2(1 - \beta^2)t^2 \end{aligned} \tag{12.3.54}$$

and as $\gamma^2(1 - \beta^2) = 1$ we get indeed:

$$x'^2 - t'^2 = x^2 - t^2 \tag{12.3.55}$$

12.3.2.1.1 Wave Equation Invariance

Now that we have determined the Lorentz transformations, we can check whether the wave equation is invariant with respect to the latter (remember that we have proven earlier that it was not invariant under a Galilean transformation!!!).

Starting from the Lorentz transformation written in explicitly:

$$\begin{aligned}x'&=\gamma(x-vt)\\t'&=\gamma\left(t-\frac{v}{c^2}x\right)\end{aligned}\tag{12.3.56}$$

we calculate the partial derivatives with respect to x and t (the expression after the second equality has already been proven earlier in this section):

$$\begin{aligned}\frac{\partial}{\partial x}&=\frac{\partial}{\partial x'}\frac{\partial x'}{\partial x}+\frac{\partial}{\partial t'}\frac{\partial t'}{\partial x}=\gamma\frac{\partial}{\partial x'}-\gamma\frac{v}{c^2}\frac{\partial}{\partial t'}\\\frac{\partial}{\partial t}&=\frac{\partial}{\partial x'}\frac{\partial x'}{\partial t}+\frac{\partial}{\partial t'}\frac{\partial t'}{\partial t}=\gamma v\frac{\partial}{\partial x'}-\gamma\frac{\partial}{\partial t'}\end{aligned}\tag{12.3.57}$$

These relation can also be written:

$$\begin{aligned}c\frac{\partial}{\partial x}&=\gamma\left(c\frac{\partial}{\partial x'}-\beta\frac{\partial}{\partial t'}\right)\\\frac{\partial}{\partial t}&=-\gamma\left(v\frac{\partial}{\partial x'}-\frac{\partial}{\partial t'}\right)\end{aligned}\tag{12.3.58}$$

Squared:

$$\begin{aligned}c^2\frac{\partial^2}{\partial x^2}&=\gamma^2\left(c^2\frac{\partial^2}{\partial x'^2}-2\beta c\frac{\partial^2}{\partial x'\partial t'}+\beta^2\frac{\partial^2}{\partial t'^2}\right)\\\frac{\partial^2}{\partial t^2}&=\gamma^2\left(v^2\frac{\partial^2}{\partial x'^2}-2v\frac{\partial^2}{\partial x'\partial t'}+\frac{\partial^2}{\partial t'^2}\right)\end{aligned}\tag{12.3.59}$$

In the Maxwell's equations, or rather in the propagation equation of the electric or magnetic field in vacuum, we have proven (see section Electrodynamics page 4138) that the following operator appeared:

$$\frac{\partial^2}{\partial t^2}-c^2\frac{\partial^2}{\partial x^2}\tag{12.3.60}$$

Substituting in it the previous differential expressions:

$$\begin{aligned}\frac{\partial^2}{\partial t^2}-c^2\frac{\partial}{\partial x^2}&=\gamma^2\left(v^2\frac{\partial^2}{\partial x'^2}-2v\frac{\partial^2}{\partial x'\partial t'}+\frac{\partial}{\partial t'^2}-c^2\frac{\partial^2}{\partial x'^2}+2\beta c\frac{\partial^2}{\partial x'\partial t'}-\beta^2\frac{\partial^2}{\partial t'^2}\right)\\&=\gamma^2\left[-c^2\left(1-\frac{v^2}{c^2}\right)\frac{\partial^2}{\partial x'^2}-2v\frac{\partial^2}{\partial x'\partial t'}+2\beta c\frac{\partial^2}{\partial x'\partial t'}+(1-\beta^2)\frac{\partial^2}{\partial t'^2}\right]\\&=\frac{\partial}{\partial t'^2}-c^2\frac{\partial}{\partial x'^2}+2\gamma^2\left[-v\frac{\partial^2}{\partial x'\partial t'}+\frac{v}{c}c\frac{\partial^2}{\partial x'\partial t'}\right]=\frac{\partial}{\partial t'^2}-c^2\frac{\partial^2}{\partial x'^2}\end{aligned}\tag{12.3.61}$$

We therefore have well:

$$\frac{\partial^2}{\partial t^2}-c^2\frac{\partial^2}{\partial x^2}=\frac{\partial^2}{\partial t'^2}-c^2\frac{\partial^2}{\partial x'^2}\tag{12.3.62}$$

which shows that a Lorentz transformation leaves invariant this operator (Jackpot!). So we got what we were looking for (the wave equation but in the other reference frame)!

The reader will also have noticed that this only works if and only if the wave propagation speed is the speed of light!

12.3.2.1.2 Hypergeometric interpretation

Now let us come back to our Lorentz transformations. Let us recall that we have restricted ourselves to the special case where the space axes were parallel (what brought us to define the "pure Lorentz transformations"). This special configuration has an interesting geometric property that sometimes many books use.

Let us see what this is about:

We have seen in the context of the study of the Lorentz transformations of lengths that we had a special transformation (boost) along one axis, i.e. the x-axis, requiring in this case for the other components:

$$y'=y \qquad z'=z \tag{12.3.63}$$

This allows us to reduce the transformation matrix L_ν^μ (4×4 matrix that we obtained earlier above) to a 2×2 matrix of components A, B, C and D such that:

$$\begin{pmatrix} ct' \\ x' \end{pmatrix}=\begin{bmatrix} A & B \\ C & D \end{bmatrix}=\begin{pmatrix} ct \\ x \end{pmatrix} \tag{12.3.64}$$

We notice that the components A, B, C, D respect by construction the following expressions:

$$\begin{aligned}
A^2-C^2&=\gamma^2-(-\gamma\beta)^2=\gamma^2-\gamma^2\beta^2=\gamma^2(1-\beta^2)=\frac{1-\beta^2}{1-\beta^2}=1\\
D^2-B^2&=\gamma^2-(-\gamma\beta)^2=\gamma^2-\gamma^2\beta^2=\gamma^2(1-\beta^2)=\frac{1-\beta^2}{1-\beta^2}=1\\
AB-CD&=\gamma(-\gamma\beta)-\gamma(-\gamma\beta)=-\gamma^2\beta+\gamma^2\beta=0
\end{aligned} \tag{12.3.65}$$

The first relation can be related with the remarkable identity in hyperbolic trigonometry (see section Trigonometry page 2532):

$$\cosh^2(\theta)-\sinh^2(\theta)=1 \tag{12.3.66}$$

And therefore:

$$A=\cosh(\alpha_1) \quad \text{and} \quad C=-\sinh(\alpha_2) \tag{12.3.67}$$

and the second relation that there exists α_2 such that:

$$D=\cosh(\alpha_2) \quad \text{and} \quad B=-\sinh(\alpha_2) \tag{12.3.68}$$

Remark

The choice of the "$-$" sign for B and C is useful because as we always have $\beta \geq 0$ (same for γ that is strictly positive) it will impose us at the end of the calculations to have $\alpha \geq 0$. Therefore, as $-\gamma\beta \leq 0$ and $\alpha \geq 0$ the only way for C (and also for B) to be negative is to put a "$-$" sign.

The third then gives the remarkable addition relation:

$$AB - CD = -\cosh(\alpha_1)\sinh(\alpha_2) + \sinh(\alpha_1)\cosh(\alpha_2) = \sinh(\alpha_1 - \alpha_2) = 0 \tag{12.3.69}$$

and therefore the difference $\alpha_1 - \alpha_2$, that we will denote more simply by α, is equal zero. Which validates the relations:

$$A = D \Leftrightarrow \cosh(\alpha_1) = \cosh(\alpha_2) \qquad B = C \Leftrightarrow -\sinh(\alpha_2) = -\sinh(\alpha_1) \tag{12.3.70}$$

The matrix is therefore presented as follow:

$$L_{\nu}^{\mu} = \begin{bmatrix} \cosh(\alpha_1) & -\sinh(\alpha_1) \\ -\sinh(\alpha_1) & \cosh(\alpha_1) \end{bmatrix} = \begin{bmatrix} \cosh(\alpha_2) & -\sinh(\alpha_2) \\ -\sinh(\alpha_2) & \cosh(\alpha_2) \end{bmatrix} \tag{12.3.71}$$

This is (by analogy to the classical one) a "hyperbolic rotation matrix". We will not go further on this analogy as it is not used for practical cases study in this textbook.

Finally, the special Lorentz transformation of velocity v along the x-axis can also be written:

$$\begin{aligned} ct' &= \gamma(ct - \beta x) = \cosh(\alpha)ct - \sinh(\alpha)x = \cosh(\alpha)\left(ct - \frac{\sinh(x)}{\cosh(x)}x\right) \\ &= \cosh(\alpha)(ct - \tanh(\alpha)x) \\ x' &= \gamma(x - \beta ct) = \cosh(\alpha)x - \sinh(\alpha)ct = \cosh(\alpha)\left(x - \frac{\sinh(x)}{\cosh(x)}ct\right) \\ &= \cosh(\alpha)(x - \tanh(\alpha)ct) \end{aligned} \tag{12.3.72}$$

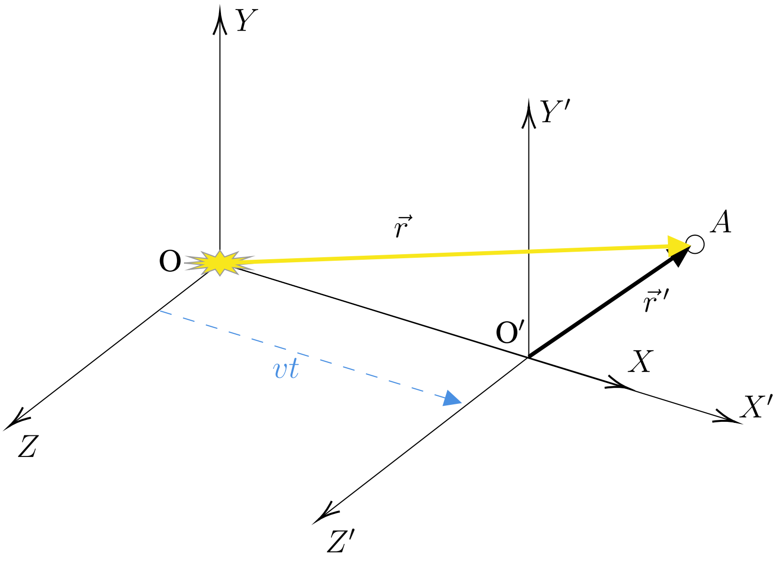
which brings us to write:

$$\beta = \frac{v}{c} = \tanh(\alpha) = \frac{\sinh(\alpha)}{\cosh(\alpha)} \qquad \text{and} \qquad \gamma = \frac{1}{\sqrt{1-\beta^2}} = \cosh(\alpha) \tag{12.3.73}$$

The dimensionless quantity α is named "rapidity" by those who use physics in high energy. The advantage of working with angles is to make the combination of 2 boosts easier.

12.3.2.2 Velocity four-vector

We can also determine the Lorentz transformations for speed. Let us consider again a particle moving in an inertial reference frame O' such that at time t', its coordinates are (x', y', z').:



Therefore, the components of the velocity v' are:

$$v'_x=\frac{\mathrm{d}x'}{\mathrm{d}t'},v'_y=\frac{\mathrm{d}y'}{\mathrm{d}t'},v'_z=\frac{\mathrm{d}z'}{\mathrm{d}t'} \tag{12.3.74}$$

So what are the components of the velocity in O (remember that O′ go away at speed v!)?

Again, we write:

$$v_x=\frac{\mathrm{d}x}{\mathrm{d}t},v_y=\frac{\mathrm{d}y}{\mathrm{d}t},v_z=\frac{\mathrm{d}z}{\mathrm{d}t} \tag{12.3.75}$$

We can differentiate by the time the components of the transformation equations we obtained before and thus we can write:

$$\begin{aligned}\mathrm{d}x&=\gamma(\mathrm{d}x'+(\beta c)\mathrm{d}t')\\ \mathrm{d}y&=\mathrm{d}y'\\ \mathrm{d}z&=\mathrm{d}z'\\ \mathrm{d}t&=\gamma\left(\mathrm{d}t'+\frac{\beta}{c}\mathrm{d}x'\right)\end{aligned} \tag{12.3.76}$$

Therefore we have:

$$v_z=\frac{\mathrm{d}z}{\mathrm{d}t}=\frac{\mathrm{d}z'}{\gamma\left(\mathrm{d}t'+\dfrac{\beta}{c}\mathrm{d}x'\right)}=\frac{\mathrm{d}z'}{\gamma\left(1+\dfrac{\beta}{c}v'_{x'}\right)\mathrm{d}t'}=\frac{v'_{z'}}{\gamma\left(1+\dfrac{\beta}{c}v'_{x'}\right)} \tag{12.3.77}$$

and also :

$$v_y=\frac{v'_{y'}}{\gamma\left(1+\dfrac{\beta}{c}v'_{x'}\right)} \tag{12.3.78}$$

and:

$$v_x=\frac{\mathrm{d}x}{\mathrm{d}t}=\frac{\gamma(\mathrm{d}x'+(\beta c)\mathrm{d}t')}{\gamma\left(\mathrm{d}t'+\dfrac{\beta}{c}\mathrm{d}x'\right)}=\frac{(v'_{x'}+(\beta c)\mathrm{d}t'}{\left(1+\dfrac{\beta}{c}v'_{x'}\right)\mathrm{d}t'}=\frac{v'_{x'}+\beta c}{1+\dfrac{\beta}{c}v'_{x'}} \tag{12.3.79}$$

And as the constant speed of reference frame O$'$ is given by $\beta = v/c$, we then have:

$$\boxed{v_x = \frac{v'_{x'} + v}{1 + \dfrac{v}{c^2} v'_{x'}}, v_y = \frac{v'_{y'}}{\gamma \left(1 + \dfrac{v}{c^2} v'_{x'}\right)}, v_z = \frac{v'_{z'}}{\gamma \left(1 + \dfrac{v}{c^2} v'_{x'}\right)}} \tag{12.3.80}$$

and vice versa:

$$\boxed{v'_{x'} = \frac{v_x + v}{1 + \dfrac{v}{c^2} v_x}, v'_{y'} = \frac{v_y}{\gamma \left(1 + \dfrac{v}{c^2} v_x\right)}, v'_{z'} = \frac{v_z}{\gamma \left(1 + \dfrac{v}{c^2} v_x\right)}} \tag{12.3.81}$$

Within the limit of classical mechanics, where the speed of light was supposed instantaneous and therefore $c \rightarrow +\infty$, we fall back on:

$$v_x = v'_{x'} + v, v_y = v'_{y'}, v_z = v'_{z'} \tag{12.3.82}$$

which are the Galilean transformations such as we have seen them in the section of Classical Mechanics.

As we can see, the speeds transformations do not follow too much the shape of the Lorentz matrix that we determined above for the coordinates. Physicists, not liking what is inhomogeneous, sought to have the same transformations for both.

Thus, let us take again the speed transformations and let us rewrite them as below:

$$\frac{v_x}{c} = \frac{\dfrac{v'_{x'}}{c} + \beta}{1 + \beta \dfrac{v'_{x'}}{c}}, \frac{v_y}{c} = \frac{\dfrac{v'_{y'}}{c}}{\gamma \left(1 + \beta \dfrac{v'_{x'}}{c}\right)}, \frac{v_z}{c} = \frac{\dfrac{v'_{z'}}{c}}{\gamma \left(1 + \beta \dfrac{v'_{x'}}{c}\right)} \tag{12.3.83}$$

These relation can be written differently if we calculate:

$$1 - \frac{\|\vec{v}\|^2}{c^2} = 1 - \frac{v_x^2 + v_y^2 + v_z^2}{c^2} = 1 - \left(\frac{\dfrac{v'_{x'}}{c} + \beta}{1 + \beta \dfrac{v'_{x'}}{c}}\right)^2 - \left(1 - \frac{v^2}{c^2}\right) \left(\left(\frac{\dfrac{v'_{y'}}{c}}{1 + \beta \dfrac{v'_{x'}}{c}}\right)^2 + \left(\frac{\dfrac{v'_{z'}}{c}}{1 + \beta \dfrac{v'_{x'}}{c}}\right)^2 \right) \tag{12.3.84}$$

Thus simplifying a bit:

$$\begin{aligned}
&\left(1 - \frac{\|\vec{v}\|^2}{c^2}\right) \left(1 + \beta \frac{v'_{x'}}{c}\right)^2 = \left(1 + \beta \frac{v'_{x'}}{c}\right)^2 - \left(\frac{v'_{x'}}{c} + \beta\right)^2 - \left(1 - \frac{v^2}{c^2}\right) \left(\left(\frac{v'_{y'}}{c}\right)^2 + \left(\frac{v'_{z'}}{c}\right)^2 \right) \\
&= 1 + 2\beta \frac{v'_{x'}}{c} + \beta^2 \left(\frac{v'_{x'}}{c}\right)^2 - \left(\frac{v'_{x'}}{c}\right)^2 - 2\beta \frac{v'_{x'}}{c} - \beta^2 - \left(1 - \frac{v^2}{c^2}\right) \left(\left(\frac{v'_{y'}}{c}\right)^2 + \left(\frac{v'_{z'}}{c}\right)^2 \right) \\
&= 1 + \beta^2 \left(\frac{v'_{x'}}{c}\right)^2 - \left(\frac{v'_{x'}}{c}\right)^2 - \beta^2 - \left(1 - \frac{v^2}{c^2}\right) \left(\left(\frac{v'_{y'}}{c}\right)^2 + \left(\frac{v'_{z'}}{c}\right)^2 \right) \\
&= (1 - \beta^2) \left(1 - \left(\frac{v'_{x'}}{c}\right)^2\right) - \left(1 - \frac{v^2}{c^2}\right) \left(\left(\frac{v'_{y'}}{c}\right)^2 + \left(\frac{v'_{z'}}{c}\right)^2 \right) \\
&= \left(1 - \frac{v^2}{c^2}\right) \left(1 - \left(\frac{v'_{x'}}{c}\right)^2 - \left(\frac{v'_{y'}}{c}\right)^2 - \left(\frac{v'_{z'}}{c}\right)^2\right) = \left(1 - \frac{v^2}{c^2}\right) \left(1 - \frac{\|\vec{v}'\|^2}{c^2}\right)
\end{aligned} \tag{12.3.85}$$

Let us put:

$$\Gamma(\vec{v})=\frac{1}{\sqrt{1-\dfrac{\|\vec{v}\|^2}{c^2}}}=\frac{1}{\sqrt{1-\left(\dfrac{x}{ct}\right)^2-\left(\dfrac{y}{ct}\right)^2-\left(\dfrac{z}{ct}\right)^2}}=\Gamma_{\vec{v}} \tag{12.3.86}$$

and:

$$\Gamma(\vec{v}')=\frac{1}{\sqrt{1-\dfrac{\|\vec{v}'\|^2}{c^2}}}=\frac{1}{\sqrt{1-\left(\dfrac{x'}{ct'}\right)^2-\left(\dfrac{y'}{ct'}\right)^2-\left(\dfrac{z'}{ct'}\right)^2}}=\Gamma_{\vec{v}'} \tag{12.3.87}$$

and:

$$\gamma=\gamma(v)=\Gamma(\vec{v})=\frac{1}{\sqrt{1-\dfrac{v^2}{c^2}}} \tag{12.3.88}$$

where the latter equality means that in order to simplify that the inertial speed and thus the study of only a single component is sufficient and that is the one collinear with the axis of movement.

With this notation and simplification it will be easy for us to determine the temporal componen! Indeed, the relation:

$$\left(1-\frac{\|\vec{v}\|^2}{c^2}\right)\left(1+\beta\frac{v'_{x'}}{c}\right)^2=\left(1-\frac{v^2}{c^2}\right)\left(1-\frac{\|\vec{v}'\|^2}{c^2}\right) \tag{12.3.89}$$

can therefore be written:

$$\frac{1}{\sqrt{1-\dfrac{\|\vec{v}\|^2}{c^2}}}\frac{1}{1+\beta\dfrac{v'_{x'}}{c}}=\frac{1}{\sqrt{1-\dfrac{v^2}{c^2}}}\frac{1}{\sqrt{1-\dfrac{\|\vec{v}'\|}{c^2}}} \tag{12.3.90}$$

$$\Gamma_{\vec{v}}\frac{1}{1+\beta\dfrac{v'_{x'}}{c}}=\Gamma_{\vec{v}'}\gamma(v)\Gamma_{\vec{v}}=\left(1+\beta\frac{v'_{x'}}{c}\right)\Gamma_{\vec{v}'}\gamma(v) \tag{12.3.91}$$

$$\Gamma_{\vec{v}}=\gamma(v)\left(\Gamma_{\vec{v}'}+\beta\frac{v'_{x'}}{c}\Gamma_{\vec{v}'}\right) \tag{12.3.92}$$

$$\Gamma_{\vec{v}}c=\gamma\left(\Gamma_{\vec{v}'}c+\beta v'_{x'}\Gamma_{\vec{v}'}\right) \tag{12.3.93}$$

The reader will have perhaps notice that we therefore have three Γ: one related to the inertial speed, the second related the norm of the vector of the particle in the reference frame O and the third related to the norm of the vector in the reference frame O'. But actually following our simplification made earlier above we know that in the repository O' the particle is at the origin in Y' and Z'.

By doing the same for each of the spatial components, we will get in the end:

$$\begin{aligned}\Gamma_{\vec{v}}c&=\gamma\left(\Gamma_{\vec{v}'}c+\beta v'_{x'}\Gamma_{\vec{v}'}\right)\\ \Gamma_{\vec{v}}v_x&=\gamma\left(\Gamma_{\vec{v}'}v'_x+\Gamma_{\vec{v}'}c\right)\\ \Gamma_{\vec{v}}v_y&=\Gamma_{\vec{v}'}v'_y\\ \Gamma_{\vec{v}}v_z&=\Gamma_{\vec{v}'}v'_z\end{aligned} \tag{12.3.94}$$

and here we have reached our goal of homogenization that allows us to write if we put:

$$\Gamma_{\vec{v}}=\gamma(\vec{v}) \qquad \Gamma_{\vec{v}'}=\gamma(\vec{v}') \tag{12.3.95}$$

the following system:

$$\boxed{\begin{pmatrix}\gamma(\vec{v})c\\ \gamma(\vec{v})v_x\\ \gamma(\vec{v})v_y\\ \gamma(\vec{v})v_z\end{pmatrix}=\gamma(\vec{v})\begin{pmatrix}c\\ v_x\\ v_y\\ v_z\end{pmatrix}=\begin{pmatrix}\gamma & \gamma\beta & 0 & 0\\ \gamma\beta & \gamma & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{pmatrix}\begin{pmatrix}\gamma(\vec{v}')c\\ \gamma(\vec{v}')v'_x\\ \gamma(\vec{v}')v'_y\\ \gamma(\vec{v}')v'_z\end{pmatrix}} \tag{12.3.96}$$

that is written in tensor form sometimes as:

$$\boxed{u'^{\mu}=L^{\mu}_{\nu}u^{\nu} \quad \text{or} \quad u^{\nu}=L'^{\nu}_{\mu}u'^{\mu}} \tag{12.3.97}$$

The vector:

$$\boxed{u'^{\nu}=\gamma(\vec{v})(c,v_x,v_y,v_z)=\gamma(\vec{v})(c,\vec{v})} \tag{12.3.98}$$

is itself named the "four-vector velocity".

12.3.2.3 Current four-vector

We have defined naturally during our introduction of the electromagnetic tensor field (see section Electrodynamics page 4121) the four-vector current:

$$j^{\nu}=(\rho c,\vec{j}) \tag{12.3.99}$$

that we can write:

$$j^{\nu}=\rho(c,\vec{v}) \tag{12.3.100}$$

This means that charge density is related to time, while current density is related to space.

Therefore, considering ρ_0 as the charge density in the proper frame moving with velocity v relative to reference frame O' and due to length contraction in the direction of the velocity, the volume occupied by a given electric charge will be multiplied by the factor $\gamma(\vec{v})$ so that:

$$\boxed{j^{\nu}=\rho_0\gamma(\vec{v})(c,\vec{v})} \tag{12.3.101}$$

which is none other than the "four-current" where we see back the four-vector velocity previously determined.

12.3.2.4 Acceleration four-vector

Having previously obtained a four-vector velocity transformable thanks to the Lorentz matrix let us also look for the equivalent for acceleration.

The four-vector acceleration is naturally expressed as the derivative with respect to the proper time of the four-velocity u such that:

$$a^v = \frac{\mathrm{d}u^v}{\mathrm{d}t'} = \frac{\mathrm{d}t}{\mathrm{d}t'}\frac{\mathrm{d}u^v}{\mathrm{d}t} \tag{12.3.102}$$

Let us just recall that the proper time of a particle is the time measured in the coordinate system of the particle, that is to say, in the reference frame where it is motionless. The proper time in the literature is often denoted τ.

Remark

We must be careful and check that the corollary of the assumption of the equivalence principle is true otherwise all General Relativity would collapse! In this early 121st century - holocene calendar - experiments (like [482]) are still trying to find a default to this principle!

The reader must first admit that (we will prove this further below) that:

$$\frac{\mathrm{d}t}{\mathrm{d}t'} = \frac{1}{\sqrt{1-\dfrac{v^2}{c^2}}} = \gamma \tag{12.3.103}$$

Therefore, we have:

$$a^v = \frac{\mathrm{d}t}{\mathrm{d}t'}\frac{\mathrm{d}u^v}{\mathrm{d}t} = \gamma\frac{\mathrm{d}}{\mathrm{d}t}(c\gamma,\gamma\vec{v}) \tag{12.3.104}$$

If we introduce the ordinary acceleration $\vec{a} = \mathrm{d}\vec{v}/\mathrm{d}t$ we see that:

$$\frac{\mathrm{d}\gamma}{\mathrm{d}t} = \frac{1}{2c^2}\gamma^3\frac{\mathrm{d}}{\mathrm{d}t}(\|\vec{v}\|^2) = \frac{1}{2c^2}\gamma^3\frac{\mathrm{d}}{\mathrm{d}t}(\vec{v}\circ\vec{v}) = \frac{\gamma^2}{2c^2}(\vec{a}\circ\vec{v}+\vec{v}\circ\vec{a}) = \frac{\gamma^3}{c^2}\vec{v}\circ\vec{a} \tag{12.3.105}$$

then:

$$a^v = \gamma\frac{\mathrm{d}}{\mathrm{d}t}(c\gamma,\gamma\vec{v}) = \gamma\left(\frac{\gamma^3}{c}\vec{v}\circ\vec{a},\gamma\vec{a}+\frac{\gamma^3}{c^2}(\vec{v}\circ\vec{a})\vec{v}\right) \tag{12.3.106}$$

Using the dual vector product identity (see section Vector Calculus page 1835):

$$(\vec{v}\circ\vec{a})\vec{v} = \vec{v}\times(\vec{v}\times\vec{a})+(\vec{v}\circ\vec{v})\vec{a} \tag{12.3.107}$$

we then find that the four-vector acceleration can be written:

$$\begin{aligned}a^v &= \gamma^3\left(\vec{b}\circ\vec{a},\frac{\vec{a}}{\gamma^2}+\frac{1}{c^2}(\vec{v}\circ\vec{a})\vec{v}\right) = \gamma^4\left(\vec{\beta}\circ\vec{a},\frac{\vec{a}}{\gamma^2}+\frac{1}{c^2}(\vec{v}\times(\vec{v}\times\vec{a})+(\vec{v}\circ\vec{v})\vec{a})\right)\\ &= \gamma^4\left(\vec{\beta}\circ\vec{a},\frac{\vec{a}}{\gamma^2}+(\vec{\beta}\times(\vec{\beta}\times\vec{a})+(\vec{\beta}\circ\vec{\beta})\vec{a})\right) = \gamma^4\left(\vec{\beta}\circ\vec{a},\left(\frac{1}{\gamma^2}+\vec{\beta}\circ\vec{\beta}\right)\vec{a}+\vec{\beta}\times(\vec{\beta}\times\vec{a})\right)\\ &= \gamma^4\left(\vec{\beta}\circ\vec{a},((1-\vec{\beta}\circ\vec{\beta})+\vec{\beta}\circ\vec{\beta})\vec{a}+\vec{\beta}\times(\vec{\beta}\times\vec{a})\right) = \gamma^4(\vec{\beta}\circ\vec{a},\vec{a}+\vec{\beta}\times(\vec{\beta}\times\vec{a}))\end{aligned} \tag{12.3.108}$$

The vector:

$$\boxed{a^v = \gamma^4 \left(\vec{\beta} \circ \vec{a}, \vec{a} + \vec{\beta} \times (\vec{\beta} \times \vec{a})\right)} \tag{12.3.109}$$

is named "four-vector acceleration" and therefore also transforms using the Lorentz matrix.

We see that if this $v \ll c$ and $\vec{a} = \vec{a}_0$ the last relation simplifies to:

$$a^v = \gamma^4 \left(\vec{\beta} \circ \vec{a}, \vec{a} + \vec{\beta} \times (\vec{\beta} \times \vec{a})\right) = 1(\vec{0} \circ \vec{a}, \vec{a} + \vec{0} \times (\vec{0} \times \vec{a})) = (0, \vec{a}) \tag{12.3.110}$$

We thus fall back on the classic acceleration.

Using the Minkowski metric (see definition further below), denoted η_{uv}, let us calculate the norm of the four-vector acceleration:

$$\begin{aligned} a^v a_v &= a^v(\eta_{uv} a^v) = a^0 a_0 - a_i a^i = \gamma^8 \left[(\vec{\beta} \circ \vec{a})^2 - \left(\vec{a} + \vec{\beta} \times (\vec{\beta} \times \vec{a})\right)^2\right] \\ &= \gamma^8 \left[(\vec{\beta} \circ \vec{a})^2 - \left(\vec{a} + \underbrace{(\vec{\beta} \circ \vec{a})\vec{\beta} - (\vec{\beta} \circ \vec{\beta})\vec{a}}_{=\vec{c}}\right)^2\right] = \gamma^8 \left[(\vec{\beta} \circ \vec{a})^2 + (\vec{a} + \vec{c})^2\right] \\ &= \gamma^8 \left[(\vec{\beta} \circ \vec{a})^2 - (\vec{a}^2 + 2\vec{a}\vec{c} + \vec{c}^2)\right] \end{aligned} \tag{12.3.111}$$

Remark

It must be well understood that when we write $(\vec{a} + \vec{\beta} \times (\vec{\beta} \times \vec{a}))^2$ it is implicit in this case that we do the sum of the squares of the components of the calculations in the brackets!

And as:

$$\vec{c}^2 = (\vec{\beta} \circ \vec{a})^2 \vec{\beta}^2 - 2(\vec{\beta} \circ \vec{\beta})(\vec{\beta} \circ \vec{a})\vec{\beta}\vec{a} + (\vec{\beta} \circ \vec{\beta})^2 \vec{a}^2 \tag{12.3.112}$$

and:

$$\vec{a}\vec{c} = (\vec{\beta} \circ \vec{a})\vec{\beta}\vec{a} - (\vec{\beta} \circ \vec{\beta})\vec{a}^2 \tag{12.3.113}$$

we put this together to get:

$$a^v a_v = \gamma^8 \left[(\vec{\beta} \circ \vec{a})^2 - \left(\begin{array}{c} \vec{a}^2 + 2(\vec{\beta} \circ \vec{a})\vec{\beta}\vec{a} - 2(\vec{\beta} \circ \vec{\beta})\vec{a}^2 + \\ (\vec{\beta} \circ \vec{a})^2 \vec{\beta}^2 - 2(\vec{\beta} \circ \vec{\beta})(\vec{\beta} \circ \vec{a})\vec{\beta}\vec{a} + (\vec{\beta} \circ \vec{\beta})^2 \vec{a}^2 \end{array}\right)\right] \tag{12.3.114}$$

Now we develop the sum $a_i a^i$ of the big parenthesis that becomes therefore:

$$\begin{aligned} a^v a_v &= \gamma^8 \left[(\beta_i a^i)^2 - \left(\begin{array}{c} a_i a^i + 2(\beta_i a^i)(\beta_i a^i) - 2(\beta_i \beta^i) a_i a^i + \\ (\beta_i a^i)^2 (\beta_i \beta^i) - 2(\beta_i \beta^i)(\beta_i a^i)(\beta_i a^i) + (\beta_i \beta^i)^2 a_i a^i \end{array}\right)\right] \\ &= \gamma^8 \left[(\beta_i a^i)^2 - \left(\begin{array}{c} a_i a^i + 2(\beta_i a^i)^2 - 2(\beta_i \beta^i) a_i a^i + \\ (\beta_i a^i)^2 (\beta_i \beta^i) - 2(\beta_i \beta^i)(\beta_i a^i)^2 + (\beta_i \beta^i)^2 a_i a^i \end{array}\right)\right] \end{aligned} \tag{12.3.115}$$

We simplify:

$$\begin{aligned} a^v a_v &= \gamma^8 \left[(\beta_i a^i)^2 (1 - 2 - \beta_i \beta^i + 2\beta_i \beta^i) + a_i a^i \left(-1 + 2\beta_i \beta^i - (\beta_i \beta^i)^2\right)\right] \\ &= \gamma^8 \left[(\vec{\beta} \circ \vec{a})^2 (1 - 2 - \vec{\beta}^2 + 2\vec{\beta}^2) + \vec{a}^2 (-1 + 2\vec{\beta} - \vec{\beta}^4)\right] \end{aligned} \tag{12.3.116}$$

Hence:

$$\begin{aligned}a^v a_v &= \gamma^8\left[(\vec{\beta}\circ\vec{a})^2(1-2-\vec{\beta}^2+2\vec{\beta}^2)+\vec{a}^2(-1+1\vec{\beta}^2-\vec{\beta}^4)\right]\\&=\gamma^8\left[(\vec{\beta}\circ\vec{a})^2(-1+\vec{\beta}^2)+\vec{a}^2(-1+2\vec{\beta}^2-\vec{\beta}^4)\right]\\&=-\gamma^8(1-\vec{\beta}^2)\left[(\vec{\beta}\circ\vec{a})^2+\vec{a}^2(1-\vec{\beta}^2)\right]=-\gamma^8\gamma^{-2}\left[(\vec{\beta}\circ\vec{a})^2+\vec{a}(1-\vec{\beta}^2)\right]\\&=-\gamma^6\left[(\vec{a}\circ\vec{\beta})^2+\vec{a}^2(1-\vec{\beta}^2)\right]=-\gamma^6\left\{\vec{a}^2-\left[\vec{a}^2\vec{\beta}^2-(\vec{a}\circ\vec{\beta})^2\right]\right\}\end{aligned}\tag{12.3.117}$$

But we have the relation:

$$\|\vec{a}\times\vec{\beta}\|^2=\vec{a}^2\vec{\beta}^2-(\vec{a}\circ\vec{\beta})^2=(a_ia^i)(\beta_i\beta^i)-(a_i\beta^i)^2\tag{12.3.118}$$

and the property of the cross product:

$$\vec{a}\times\vec{\beta}=-\vec{\beta}\times\vec{a}\tag{12.3.119}$$

Which finally gives us:

$$a^v a_v=-\gamma^6\left[\vec{a}^2-(\vec{a}\times\vec{\beta})^2\right]=-\gamma^6\left[\vec{a}\circ\vec{a}-(\vec{a}\times\vec{\beta})\circ(\vec{a}\times\vec{\beta})\right]=-\gamma^6\left[\vec{a}^2-(\vec{\beta}\times\vec{a})^2\right]\tag{12.3.120}$$

The relation:

$$\boxed{a^v a_v=-\gamma^6\left[\vec{a}^2-(\vec{\beta}\times\vec{a})^2\right]}\tag{12.3.121}$$

will be extremely useful to us when we will study the Abraham-Becker radiation damping force in the section of Electrodynamics (see page 4191).

Now imagine an object with a uniformly accelerated relative motion $\vec{a}_0^2$ (constant acceleration) in our own repository. If we assume our repository as fixed, we have $\vec{v}=\vec{0}\Leftrightarrow\vec{\beta}=\vec{0}$. Therefore:

$$a^v a_v=-\gamma^6\vec{a}^2=-\vec{a}_0^2\tag{12.3.122}$$

Verbatim after rearranging the terms and taking the square root if the accelerated motion is made only along a single component:

$$\gamma^3\frac{\mathrm{d}u}{\mathrm{d}t}=a_0\tag{12.3.123}$$

But, we also have:

$$\frac{\mathrm{d}}{\mathrm{d}t}(\gamma u)=\gamma\frac{\mathrm{d}u}{\mathrm{d}t}+u\frac{\mathrm{d}\gamma}{\mathrm{d}t}=\gamma^3\frac{\mathrm{d}u}{\mathrm{d}t}\left(\gamma^{-2}+\frac{u^2}{c^2}\right)=\gamma^3\frac{\mathrm{d}u}{\mathrm{d}t}\left(1-\frac{u^2}{c^2}+\frac{u^2}{c^2}\right)=\gamma^3\frac{\mathrm{d}u}{\mathrm{d}t}\tag{12.3.124}$$

So finally, we can write:

$$\gamma^3\frac{\mathrm{d}u}{\mathrm{d}t}=\frac{\mathrm{d}}{\mathrm{d}t}(\gamma u)=\frac{\mathrm{d}}{\mathrm{d}t}\frac{u}{\sqrt{1-\dfrac{u^2}{c^2}}}=a_0\tag{12.3.125}$$

Which after integration gives:

$$\boxed{u=\frac{\mathrm{d}x}{\mathrm{d}t}=\frac{a_0}{\sqrt{1+\frac{a_0^2t^2}{c^2}}}} \tag{12.3.126}$$

We see that the speed u never reaches c while the force (acceleration implicitly) is always the same!

So we have:

$$x=\frac{c^2}{a_0}\int\limits_0^{a_0t/c}\frac{x}{\sqrt{1+x^2}}\mathrm{d}x=\frac{c^2}{a_0}\sqrt{1+\frac{a_0^2t^2}{c^2}}-\frac{c^2}{a_0} \tag{12.3.127}$$

which gives us:

$$\left(x+\frac{c^2}{a_0}\right)^2-c^2t^2=\frac{c^4}{a_0^2} \tag{12.3.128}$$

After rearranging, we write this:

$$\boxed{x=\frac{c^2}{a_0}\left(\sqrt{1+\frac{a_0^2t^2}{c^2}}-1\right)} \tag{12.3.129}$$

We are far from the relation of uniformly accelerated motion we have proved in the section of Classical Mechanics (see page 3224) and that is for recall:

$$x=\frac{1}{2}a_0t^2 \tag{12.3.130}$$

However, for t close to zero, we fall back on the same Classical Mechanics relation by taking the Taylor expansion to the second order of the square root (see section Sequences and Series page 1703):

$$x=\frac{c^2}{a_0}\left(\sqrt{1+\frac{a_0^2t^2}{c^2}}\right)\cong\frac{c^2}{a_0}\left(1+\frac{a_0^2t^2}{2c^2}-1\right)=\frac{1}{2}a_0t^2 \tag{12.3.131}$$

However, this does not give us the relations of transformation of acceleration components in a simple form. Let's see how to get them.

First let us recall that we have obtained for speed:

$$v_x=\frac{v'_{x'}+v}{1+\frac{v}{c}v'_{x'}},\ v_y=\frac{v'_{y'}}{\gamma\left(1+\frac{v}{c}v'_{x'}\right)},\ v_z=\frac{v'_{z'}}{\gamma\left(1+\frac{v}{c}v'_{x'}\right)} \tag{12.3.132}$$

Then it comes by differentiating:

$$\begin{aligned}\mathrm{d}v_x &= \mathrm{d}\left((v'_{x'}+v)\frac{1}{1+\dfrac{v}{c^2}v'_{x'}}\right) = \frac{\mathrm{d}v'_{x'}}{1+\dfrac{v}{c^2}v'_{x'}} + \frac{\underbrace{\mathrm{d}v}_{=0}}{1+\dfrac{v}{c^2}v'_{x'}} + (v'_{x'}+v)\mathrm{d}\left(1+\frac{v}{c^2}v'_{x'}\right)^{-1}\\ &= \frac{\mathrm{d}v'_{x'}}{1+\dfrac{v}{c^2}v'_{x'}} - \frac{v'_{x'}+v}{\left(1+\dfrac{v}{c^2}v'_{x'}\right)^2}\frac{v}{c^2}\mathrm{d}v'_{x'} = \frac{\mathrm{d}v'_{x'}\left(1+\dfrac{v}{c^2}v'_{x'}\right)-(v'_{x'}+v)\dfrac{v}{c^2}\mathrm{d}v'_{x'}}{\left(1+\dfrac{v}{c^2}v'_{x'}\right)^2}\\ &= \frac{\mathrm{d}v'_{x'}+\dfrac{v}{c^2}v'_{x'}\mathrm{d}v'_{x'}-v'_{x'}\dfrac{v}{c^2}\mathrm{d}v'_{x'}-v\dfrac{v}{c^2}\mathrm{d}v'_{x'}}{\left(1+\dfrac{v}{c^2}v'_{x'}\right)^2} = \frac{\mathrm{d}v'_{x'}-v\dfrac{v}{c^2}\mathrm{d}v'_{x'}}{\left(1+\dfrac{v}{c^2}v'_{x'}\right)^2}\end{aligned} \tag{12.3.133}$$

therefore:

$$\frac{\mathrm{d}v_x}{\mathrm{d}t} = \frac{1}{\gamma^2\left(1+\dfrac{v}{c^2}v'_{x'}\right)^2}\frac{\mathrm{d}v'_{x'}}{\mathrm{d}t} = \frac{1}{\gamma^2\left(1+\dfrac{v}{c^2}v'_{x'}\right)^2}\frac{\mathrm{d}t'}{\mathrm{d}t}\frac{\mathrm{d}v'_{x'}}{\mathrm{d}t'} = \frac{1}{\gamma^2\left(1+\dfrac{v}{c^2}v'_{x'}\right)^2}\frac{\mathrm{d}t'}{\mathrm{d}t}a'_{x'} \tag{12.3.134}$$

Let us recall now that we have proved that:

$$ct = \gamma(t'c+\beta x') \tag{12.3.135}$$

differentiating it comes:

$$c\mathrm{d}t = \gamma(\mathrm{d}t'c+\beta\mathrm{d}x') \tag{12.3.136}$$

We can write:

$$c\frac{\mathrm{d}t}{\mathrm{d}t'} = \frac{\gamma}{\mathrm{d}t'}(\mathrm{d}t'c+\beta\mathrm{d}x') \tag{12.3.137}$$

After simplifying and rearranging we get obviously:

$$\frac{\mathrm{d}t}{\mathrm{d}t'} = \gamma\left(1+\frac{\beta v'_x}{c}\right) \tag{12.3.138}$$

Hence:

$$\frac{\mathrm{d}t'}{\mathrm{d}t} = \frac{1}{\gamma\left(1+\dfrac{\beta v'_x}{c}\right)} = \frac{1}{\gamma\left(1+\dfrac{vv'_x}{c^2}\right)} \tag{12.3.139}$$

hence finally:

$$a_x = \frac{\mathrm{d}v_x}{\mathrm{d}t} = \frac{1}{\gamma^2\left(1+\dfrac{v}{c^2}v'_{x'}\right)^2}\frac{\mathrm{d}t'}{\mathrm{d}t}a'_{x'} = \frac{1}{\gamma^3\left(1+\dfrac{v}{c^2}v'_{x'}\right)^3}a'_{x'} \tag{12.3.140}$$

and for the components y, z:

$$v_{y,z} = \frac{v'_{y',z'}}{\gamma\left(1+\dfrac{v}{c^2}v'_{x'}\right)} \Rightarrow \mathrm{d}v_{y,z} = \frac{\mathrm{d}v'_{y',z'}}{\gamma\left(1+\dfrac{v}{c^2}v'_{x'}\right)} - \frac{v'_{y',z'}}{\gamma\left(1+\dfrac{v}{c^2}v'_{x'}\right)^2}\frac{v}{c^2}\mathrm{d}v'_{x'} \tag{12.3.141}$$

and therefore:

$$\begin{aligned}\frac{\mathrm{d}v_{y,z}}{\mathrm{d}t}&=\left(\frac{1}{\gamma\left(1+\dfrac{v}{c^2}{v'}_{x'}\right)}\frac{\mathrm{d}{v'}_{y',z'}}{\mathrm{d}t}-\frac{{v'}_{y',z'}}{\gamma\left(1+\dfrac{v}{c'^2}{v'}_{x'}\right)^2}\right)\frac{\mathrm{d}t'}{\mathrm{d}t'}\\&=\frac{{a'}_{y',z'}}{\gamma^2\left(1+\dfrac{v}{c^2}{v'}_{x'}\right)^2}-\frac{v{v'}_{y',z'}}{c^2}\frac{{a'}_{x'}}{\gamma^2\left(1+\dfrac{v}{c'^2}{v'}_{x'}\right)^3}\end{aligned} \tag{12.3.142}$$

So finally:

$$\boxed{\begin{aligned}a_x&=\frac{1}{\gamma^3\left(1+\dfrac{v}{c^2}{v'}_{x'}\right)^3}{a'}_{x'}\\a_y&=\frac{{a'}_{y'}}{\gamma^2\left(1+\dfrac{v}{c^2}{v'}_{x'}\right)^2}-\frac{v{v'}_{y'}}{c^2}\frac{{a'}_{x'}}{\gamma^2\left(1+\dfrac{v}{c^2}{v'}_{x'}\right)^3}\\a_z&=\frac{{a'}_{z'}}{\gamma^2\left(1+\dfrac{v}{c^2}{v'}_{x'}\right)^2}-\frac{v{v'}_{z'}}{c^2}\frac{{a'}_{x'}}{\gamma^2\left(1+\dfrac{v}{c^2}{v'}_{x'}\right)^3}\end{aligned}} \tag{12.3.143}$$

Remember that these relations apply when the movements of the reference frames are in uniform translation!

Caution! It is a common misconception that we cannot speak about acceleration in Special Relativity. But that's a wrong belief from scientifically illiterate people and undergraduate students as we have just seen it! What we can't do with acceleration in Special Relativity is to use a simple Lorentz transformation for the acceleration or to find time and space Lorentz boosts that includes acceleration. That's all!!!!

12.3.2.5 Relativistic sum of velocities

As the speed of light is a speed supposed unsurpassable, we now come to ask ourselves what will be finally the speed of an object launched at a speed close to that of light (for example...) from a reference frame moving also close to that of the speed of light (why not...).

We must then find a relation that gives the real speed V from the launch speed v_2 and speed of the repository v_1.

We know that for the object launched:

$$x'=v_2t' \tag{12.3.144}$$

As the one who is concerned does not know the real speed V, it should use the Lorentz transformations. Thus, given the expression of t' that we saw earlier it comes:

$$x'=v_2\frac{t-\dfrac{v_1x}{c^2}}{\sqrt{1-\left(\dfrac{v_1}{c}\right)^2}} \tag{12.3.145}$$

and given the prior-previous expression of x' we also have:

$$\frac{x-v_1t}{\sqrt{1-\left(\dfrac{v_1}{c}\right)^2}}=v_2\frac{t-\dfrac{v_1x}{c^2}}{\sqrt{1-\left(\dfrac{v_1}{c}\right)^2}} \tag{12.3.146}$$

therefore after rearranging and simplifying a bit:

$$x-v_1t=v_2\left(t-\frac{v_1x}{c}\right)=v_2t-\frac{v_2v_1x}{c^2} \tag{12.3.147}$$

Hence:

$$x\left(1+\frac{v_2v_1}{c^2}\right)=t(v_2+v_1) \tag{12.3.148}$$

We know that $v=x/t$ so we can finally write the "law of compositions relativistic speeds" or simply "velocity-addition formula" or "Einstein's velocity addition" relation:

$$\boxed{V=\frac{x}{t}=\frac{v_1+v_2}{1+\dfrac{v_2v_2}{c^2}}} \tag{12.3.149}$$

which is therefore the speed of a moving body in the moving reference frame relatively to that seen in the rest frame (but that in fact could also move at any speed less then c).

And conversely seen from the other moving frame of reference, we have by the same developments (with reverse signs and speed of course):

$$\boxed{V'=\frac{x'}{t'}=\frac{v_2-v_1}{1-\dfrac{v_1v_2}{c^2}}} \tag{12.3.150}$$

which is the speed of a moving body in the rest frame relatively to that considered as being in motion (or in other words seen by the moving frame of reference).

Notice that for small speeds, we fall back obviously on the Galilean addition relation:

$$V=\frac{v_1+v_2}{1+\dfrac{v_2v_2}{c^2}}=v_1+v_2 \tag{12.3.151}$$

If we take the speed of light as one of the velocities, we get:

$$V=\frac{v_1+v_2}{1+\dfrac{v_2v_2}{c^2}}=\frac{v_1+c}{1+\dfrac{v_2c}{c^2}}=\frac{v_1+c}{1+\dfrac{v_2}{c}}=c \tag{12.3.152}$$

Hence the speed of light cannot be surpassed!

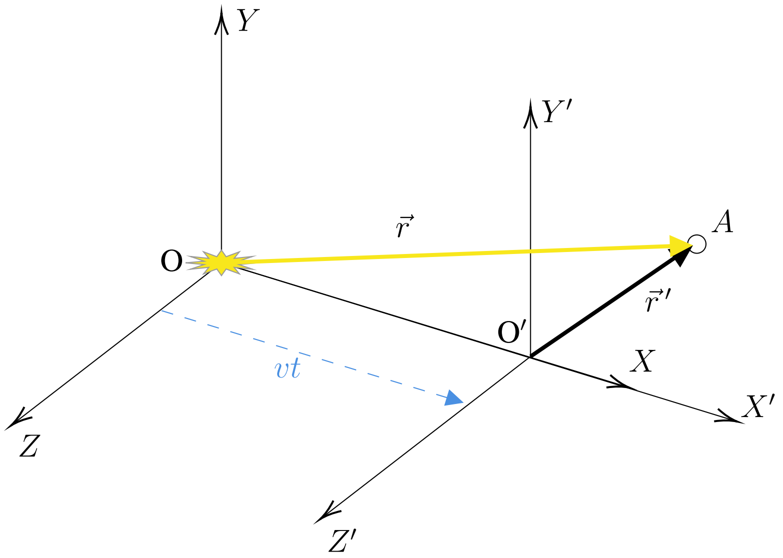
On the other hand, if we take two velocities smaller than that of light, we have, with $v_1=c-\lambda$, $v_2=c-\mu$ and $\lambda,\mu>0$:

$$\begin{aligned}V&=\frac{2c-\lambda-\mu}{1+\dfrac{(c-\lambda)(c-\mu)}{c^2}}=c\frac{2c-\lambda-\mu}{2c-\lambda-\mu+\dfrac{\lambda\mu}{c}}\\&=\frac{c}{1+\dfrac{\lambda\mu}{c(2c-\lambda-\mu)}}\leq c\end{aligned} \tag{12.3.153}$$

Then it is not possible to reach the speed of light by adding speed less than that of light!

12.3.2.6 Relativistic lengths variation (length contraction)

Let us consider now that the length of an object is given by the distance between its two ends A and B. Let us consider this object $\overline{AB}$ motionless in the repository O$'$ in uniform translation and oriented along the axis O$'X'$:



Its length is then the distance between its both ends:

$$L' = x'_B - x'_A \tag{12.3.154}$$

For the observer O, the object is moving. The positions A and B should therefore be measured simultaneously:

$$L' = x'_B - x'_A \tag{12.3.155}$$

So it comes using the relation proved earlier in this section:

$$x' = \gamma(x - vt) = \frac{x - vt}{\sqrt{1 - \left(\dfrac{v}{c}\right)^2}} \tag{12.3.156}$$

the following difference:

$$L' = x'_B - x'_A = \frac{x_B - vt}{\sqrt{1 - \left(\dfrac{v}{c}\right)^2}} - \frac{x_A - vt}{\sqrt{1 - \left(\dfrac{v}{c}\right)^2}} = \frac{X_B - X_A}{\sqrt{1 - \left(\dfrac{v}{c}\right)^2}} = \frac{L}{\sqrt{1 - \left(\dfrac{v}{c}\right)^2}} \tag{12.3.157}$$

hence the remarkable result:

$$\boxed{L = L'\sqrt{1 - \left(\frac{v}{c}\right)^2}} \tag{12.3.158}$$

we also find the relation frequently in the literature as follows:

$$\boxed{\Delta L = \Delta L'\sqrt{1 - \frac{v^2}{c^2}}} \tag{12.3.159}$$

Thus, the length of an observed rule in a moving frame relatively to the proper frame of the rule is less than its own length (which can assimilate in generality to a "proper length"). In other words, the length of a moving object measured by the fixed reference frame will be measured shorter than its real proper size. This phenomenon is named "length contraction".

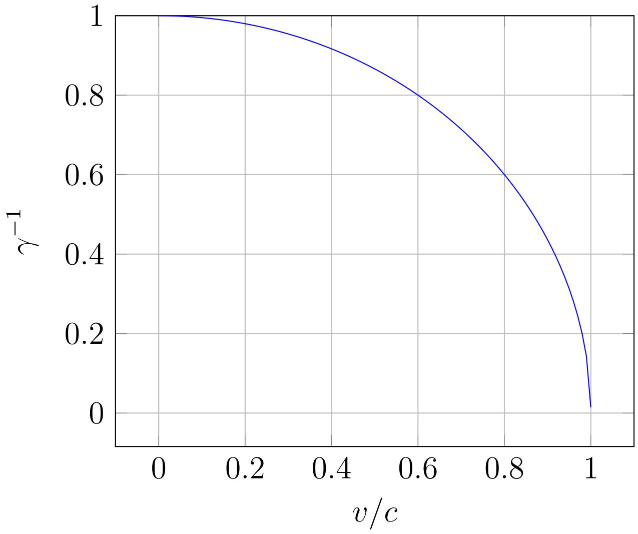


Figure 12.197: Lorentz factor length contraction function trend

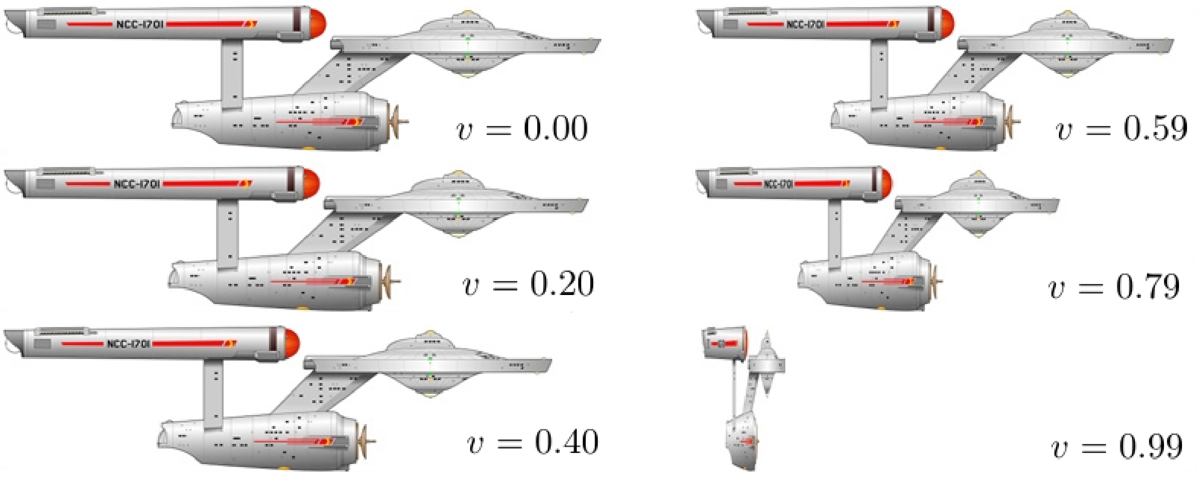


Figure 12.198: Length contraction principle for rectilinear motion (source:?)

Due to superficial application of the contraction formula some paradoxes can occur. Examples are the "ladder paradox" and "Bell's spaceship paradox". However, those paradoxes can simply be solved by a correct application of relativity of simultaneity. Another famous paradox is the "Ehrenfest paradox" (high relativistic speed "rigid" rotating disc[39]), which proves that the concept of rigid bodies is not compatible with relativity, reducing the applicability of Born rigidity, and showing that for a co-rotating observer the geometry is in fact non-euclidean and we need then to use General Relativity.

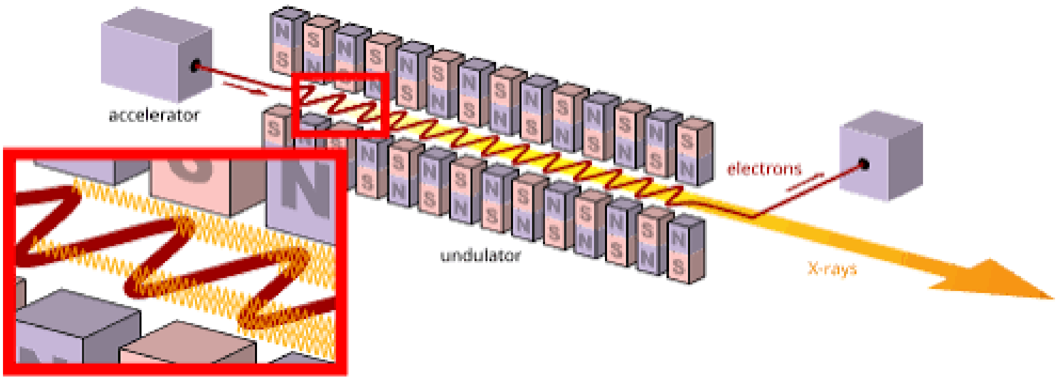
---

[39] Circumference of a rotating disk should contract but not the radius, as radius is perpendicular to the direction of motion.

Caution! As we already have proved it earlier, length contraction, when expressed in the formalism of matrix algebra, is in fact a rotation in space-time (a hyperbolic rotation!!!)! So physically an object isn't smaller when moving, but is just seen with a different angle (i.e. in a different perspective angle) than the perpendicular one by default (at $v = 0$) in space-time, and hence it looks smaller (however length contraction isn't also a visual effect as may make think the last sentence!).

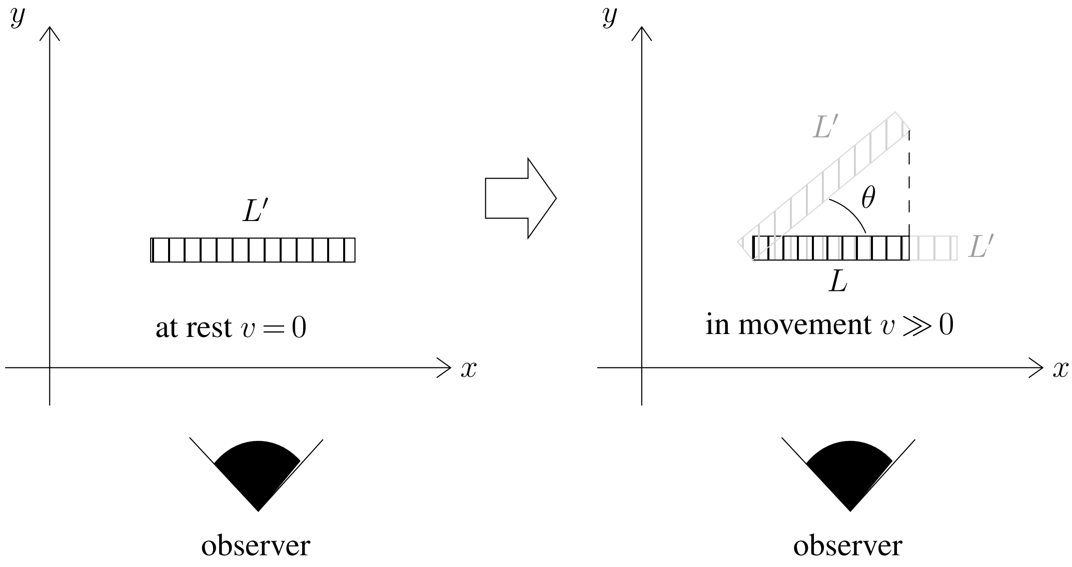
Notice that at the day we write these line (end of year 12026 according to Holocene calendar), as far as we know, it seems that there is still no direct experimental evidence of relativistic lorentz length contraction (it's indeed quite hard to measure the length of an object while it is moving at very high velocity...).

The SLAC (Stanford Linear Accelerator) X-ray free-electron LASER is in our opinion the most impressive indirect evidence of the Lorenz length contraction as it works because electrons are pushed to energies so high that relativistic effects dominate their behaviour. When these electrons travel through the undulator - a long structure with alternating magnetic fields - the spacing of those magnetic fields is fixed in the laboratory, but not in the electrons' own frame of reference. Since the electrons move extremely close to the speed of light, they experience Lorentz length contraction, which makes the undulator's magnetic period appear shorter by a Lorenz factor γ of $27,500$ (corresponding to 99.9999999338843% the speed of light [459]). In that contracted frame, the electrons feel much faster magnetic oscillations than the physical device would suggest. Because electrons radiate at the frequency of their oscillatory motion, this contracted period forces them to emit radiation at a much higher frequency that are used afterwards to image electrons clouds in molecules developed mainly for medical research. SLAC is therefore a practical demonstration of Special Relativity in action. This Linac Coherent Light Source (LCLS) works as it does because the electrons' relativistic motion makes a centimeter-scale magnetic structure behave, from their perspective, like a structure with a period thousands of times smaller. This relativistic shrinking is what makes X-ray LASER production feasible in a real-world accelerator.



12.3.2.7 Relativistic apparent angle rotation (Lampa–Terrell–Penrose effect)

The reader may have noticed something. The result of length contraction we have derived above for an imaginary one dimensional line is equivalent as if the line was rotated as illustrated below by an angle θ:



We have immediately:

$$\cos(\theta)=\frac{L}{L'}=\frac{L'\sqrt{1-\left(\dfrac{v}{c}\right)^2}}{L'}=\sqrt{1-\left(\frac{v}{c}\right)^2} \tag{12.3.160}$$

Hence:

$$\theta=\arccos\left(\sqrt{1-\left(\frac{v}{c}\right)^2}\right) \tag{12.3.161}$$

We can find the above relation written differently using the trigonometric identity:

$$\cos^2(\alpha)+\sin^2(\alpha)=1 \tag{12.3.162}$$

Therefore:

$$\cos^2(\arcsin(x))+\sin^2(\arcsin(x))=1 \tag{12.3.163}$$

Hence:

$$\cos^2(\arcsin(x))+x^2=1 \tag{12.3.164}$$

Finally:

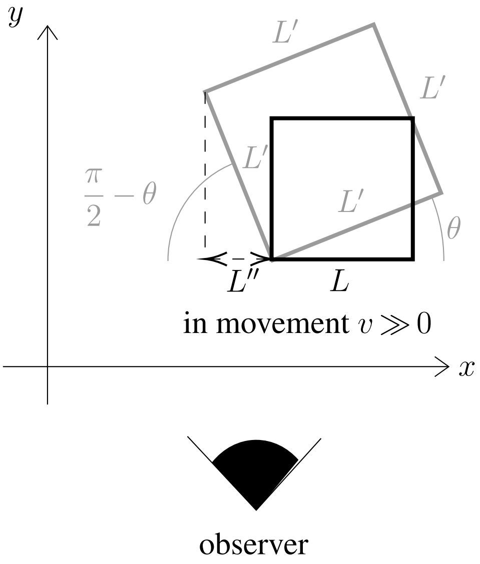
$$\arcsin(x)=\arccos\left(\sqrt{1-x^2}\right) \tag{12.3.165}$$

So the above relation can be written as the "relativistic apparent angle rotation":

$$\boxed{\theta = \arccos\left(\sqrt{1-\left(\frac{v}{c}\right)^2}\right) = \arcsin\left(\frac{v}{c}\right)} \tag{12.3.166}$$

If we assume the equivalence that a contraction is equivalent to a rotation! That means we should be able to see the left side of the rode (or stick) - that is not visible when the speed is zero - and therefore the total length observed is not only that of given by L (contracted length) but also we have to add to it a fraction of the thickness of the rode (or stick)! That also means that the illustration above with the Spaceship is inaccurate! We should be able to see the side of the ship as it approach the speed of light!

In the case of a cube of side L' we can illustrate the situation as following:



We have from the figure above:

$$\cos\left(\frac{\pi}{2}-\theta\right) = \frac{L''}{L'} \tag{12.3.167}$$

Hence using trigonometric identities:

$$\cos(\theta) = \frac{L''}{L'} \tag{12.3.168}$$

Therefore rearranging and using previous result:

$$L'' = L'\sin(\theta) = L'\sin\left(\arcsin\left(\frac{v}{c}\right)\right) = L'\frac{v}{c} \tag{12.3.169}$$

Hence the total length seen by the observer:

$$L'\frac{v}{c} + L'\sqrt{1-\left(\frac{v}{c}\right)^2} \tag{12.3.170}$$

The "Terrell rotation" or "Terrell effect" is the visual distortion that a passing object would appear to undergo, according to the special theory of relativity if it were travelling at a significant fraction of the speed of light. This behaviour was described independently by both Roger

Penrose and James Terrell. Penrose's article was submitted 29 July 11958 (holocene calendar) and published in January 11959 (holocene calendar). Terrell's article was submitted 22 June 11959 (holocene calendar) and published 15 November 11959 (holocene calendar). The general phenomenon was noted already in 11924 (holocene calendar) by Austrian physicist Anton Lampa.

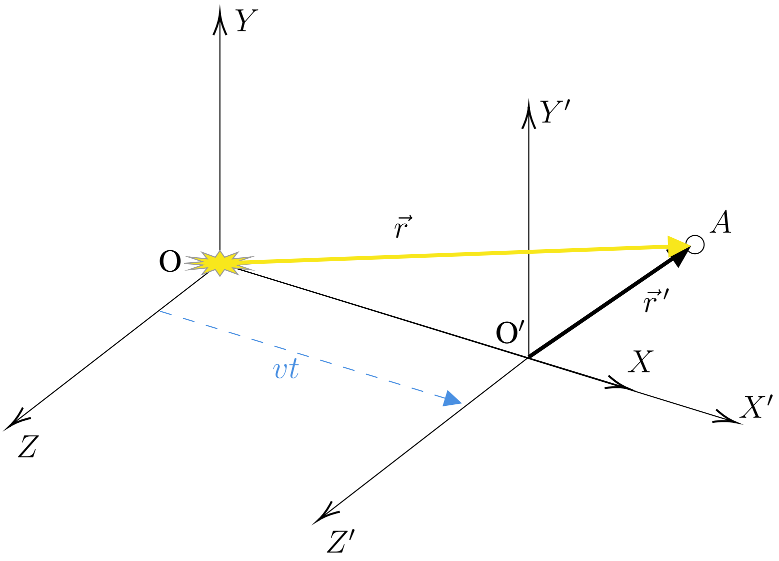
Due to a dispute about priority and correct attribution, the effect is also sometimes referred to as the "Penrose–Terrell effect", the "Terrell–Penrose effect" or the "Lampa–Terrell–Penrose effect", but not the "Lampa effect".

Notice that at the day we write these line (end of year 12026 according to Holocene calendar), as far as we know, it seems that there is still no direct experimental evidence of the Lampa-Terrell-Penrose effect (it's indeed quite hard to measure the detailed shape of an object while it is moving at very high velocity...).

12.3.2.8 Relativistic time variation (time dilatation)

An event is a phenomenon that occurs in a given place at a given time. The origin of time is difficult to determine, we often prefer to define the concept of "time interval" as the time elapsed between two events as it is often customary (see section Principia page 3083).

Let us now consider two consecutive events A and B that occur at the same location x' (!) in the repository in uniform translation:



For the observer in O′, the time interval is simply:

$$\Delta t' = t'_B - t'_A \qquad (12.3.171)$$

To measure this time interval, the observer O in the fixed reference repository should also require that x' is common to both events. Then using the relation proved earlier above:

$$t'=\gamma\left(t-\frac{v}{c^2}x\right)=\frac{t-\dfrac{v}{c^2}x}{\sqrt{1-\left(\dfrac{v}{c}\right)^2}} \quad (12.3.172)$$

we get:

$$\Delta t=t_B-t_A=\frac{t'_B+\dfrac{vx'}{c^2}}{\sqrt{1-\left(\dfrac{v}{c}\right)^2}}-\frac{t'_A+\dfrac{vx'}{c^2}}{\sqrt{1-\left(\dfrac{v}{c}\right)^2}}=\frac{t'_B-t'_A}{\sqrt{1-\left(\dfrac{v}{c}\right)^2}}=\frac{\Delta t'}{\sqrt{1-\left(\dfrac{v}{c}\right)^2}} \quad (12.3.173)$$

hence the remarkable result:

$$\boxed{\Delta t=\frac{\Delta t'}{\sqrt{1-\left(\dfrac{v}{c}\right)^2}}} \quad (12.3.174)$$

what we write under traditional condensed form:

$$t=\gamma\tau \quad (12.3.175)$$

where τ is obviously the "proper time".

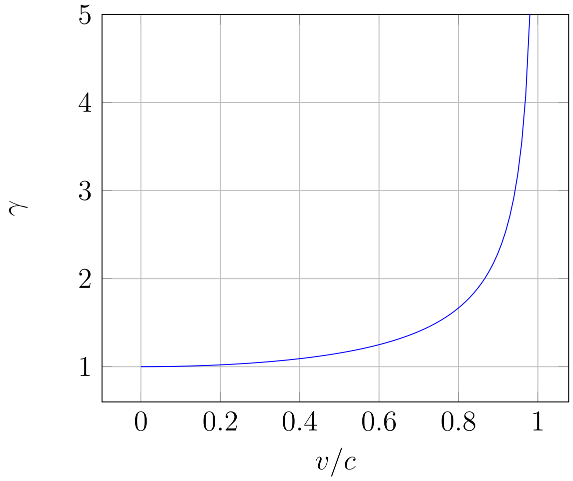


Figure 12.199: Lorentz factor time dilatation function trend

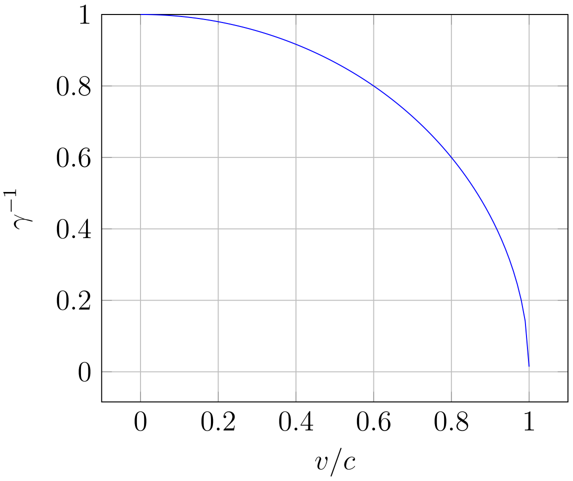


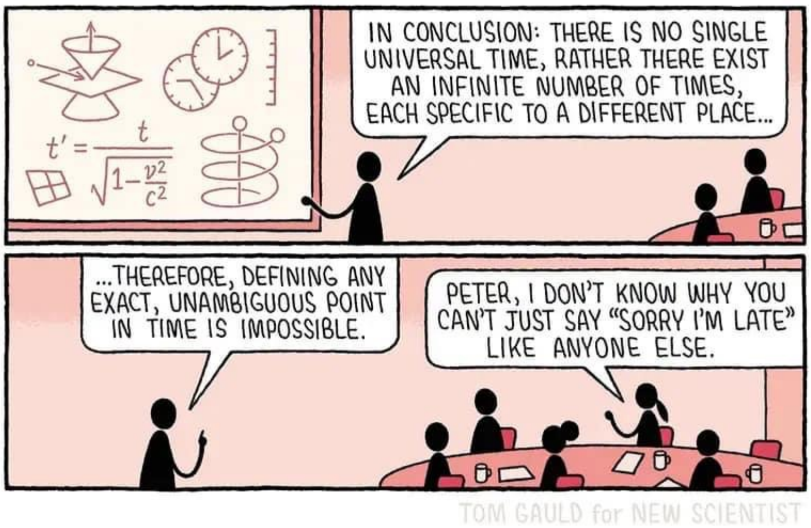
Figure 12.200: Lorentz length contraction function (reminder)

We also deduce taking an infinitesimal time element:

$$\frac{\mathrm{d}t}{\mathrm{d}t'}=\frac{1}{\sqrt{1-\dfrac{v^2}{c^2}}} \quad (12.3.176)$$

So the observer O (stationary) measures a time interval much larger than that one measured in the moving repository where the phenomenon takes place as it moves quickly. The time in the fixed repository (thus the "proper time" of the fixed reference frame!) seems like dilated compared to that in occurring in the mobile reference frame (that is to say relatively to the "proper time" of mobile reference frame!).

The conclusion is that there is no single universal time, rather there exist an infinite number of times, each specific to a different reference frame... therefore, defining any exact, unambiguous point in time is impossible (as far as we know today...)!


$t' = \frac{t}{\sqrt{1-\frac{v^2}{c^2}}}$
IN CONCLUSION: THERE IS NO SINGLE UNIVERSAL TIME, RATHER THERE EXIST AN INFINITE NUMBER OF TIMES, EACH SPECIFIC TO A DIFFERENT PLACE...
...THEREFORE, DEFINING ANY EXACT, UNAMBIGUOUS POINT IN TIME IS IMPOSSIBLE.
PETER, I DON'T KNOW WHY YOU CAN'T JUST SAY "SORRY I'M LATE" LIKE ANYONE ELSE.
TOM GAULD for NEW SCIENTIST

There are a few quite solid experimental direct evidence of relativistic lorentz time contraction. Like the variation of the rate of decay of mesotrons with momentum (see [468]) or the measurements of relativistic time dilatation for positive and negative muons in a circular orbit (see [37]). But we also have experimental evidence with atom clocks like the the study of the rate of a moving atomic clock (see [280] or [281]) or test of relativistic time dilation with fast optical atomic clocks at different velocities (see [461]) or with optical fiber connected to atomic clocks (see [137]). And to finish this list, we also have experimental evidence with optical clocks (see [99]).

Caution! A common error in Special Relativity is to think that time dilation is physical/real, i.e. the people (pr particles) moving will age more slowly. But that's wrong in Special Relativity (but not in General Relativity!) and for exactly the same reasons as for the length dilatation (you just have to consider the axis of space ct instead of the axis of pure time t).

12.3.2.8.1 Lewis Carrol Epstein interpretation

Let us now introduce a quite rare interpretation of time dilation and the fact that we can't move faster than line that Lewis Carrol Epstein liked to explain as following: «*The reason you can't go faster than light is that you can't go slower than light either. You're always going the speed of light. When you're "standing still" you're using all your speed to move through time from past to future. When your are "moving" you're using part of your speed to move through space, so there is less of it moving your through time.*»

If the quote of Lewis isn't crystal clear for you it's quite normal. Doing the maths will help!

The relevant formula for this is:

$$\beta^2+\frac{1}{\gamma^2}=1 \tag{12.3.177}$$

where for reminder $\beta=v/c$ is the ratio of the observed object's speed to the speed of light and:

$$\frac{1}{\gamma}=\sqrt{1-\frac{v^2}{c^2}} \tag{12.3.178}$$

is the reciprocal of the Fitzgerald factor.

Now remember that proper time τ is related to coordinate time t by the relation:

$$\tau=\frac{t}{\gamma} \tag{12.3.179}$$

Now let us take advantage what we just reminded by multiplying the reciprocal of the Fitzgerald factor by the speed of light and defining a new quantity that in essence tells us the "speed of aging" in the same dimensional units as speed (in space):

$$\frac{c}{\gamma}\equiv v_\tau \tag{12.3.180}$$

Here the $\equiv$ symbol means "is defined as" and v_τ is a quantity that could be named "motion in proper time". Now let's go back to the equation that related the proper time to coordinate time $\tau=t/\gamma$ and multiply both sides by the speed of light. Then we get:

$$c\tau=c\frac{t}{\gamma}=v_\tau t \tag{12.3.181}$$

Compare this to the classical motion in space, $d=vt$, we see that the two are exactly analogous: evidently the quantity $c\tau$ is analogous to a distance. Let us now consider the relationship we started off with, that for recall is:

$$\beta^2+\frac{1}{\gamma^2}=1 \tag{12.3.182}$$

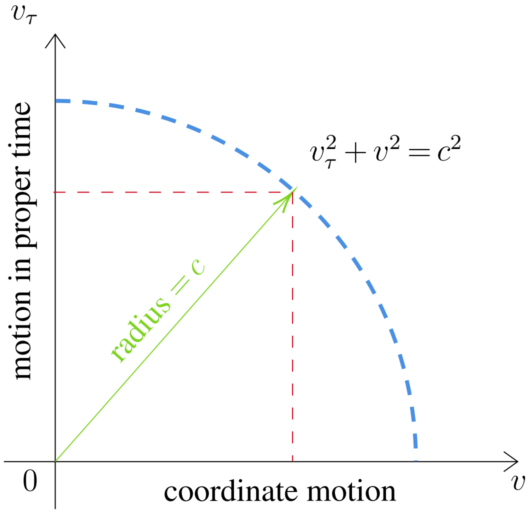
If we multiply both sides by the speed of light squared c^2 we then get:

$$c^2\beta^2+\frac{c^2}{\gamma^2}=c^2 \tag{12.3.183}$$

or, using $v = c\beta$ and the above definition of "motion in proper time":

$$v^2 + v_\tau^2 = c^2 \tag{12.3.184}$$

As the reader can see it, that's simply the Pythagorean theorem and can therefore be represented typically in the quarter of a circle (as in practice only positive speed are of interest).



This is it! Every object travels at the speed of light, but for objects that don't travel at the speed of light in space, it is distributed over motion in space and "motion in proper time". Indeed, notice that if $v = 0$ then the above reduces to:

$$v_\tau = c \tag{12.3.185}$$

This means that, right now, as you are sitting looking at your screen, you are travelling at the speed of light! The reason you don't notice this is that your motion is entirely distributed over "motion in proper time". Since you can never be in motion with respect to yourself, you always travel at the maximum speed in proper time in your rest frame. But when something travels in space relative to you, then their motion, as observed by you, is "redistributed" over motion in space, so, as observed by you, they won't travel as fast in proper time as you do, which is to say, you observe them to "age" more slowly than you do, and we come back to the above relationship between proper time and coordinate time.

Let me now finally address why we can't break the speed of light in space. Well, to break it, you have to first reach it. So why can't we, or any other space-time observer, reach the speed of light in space? If we follow the above chain of reasoning we are now in an excellent position to rephrase the question into an equivalent one that may get us closer to a satisfying answer, namely: Why can't we, or any other space-time observer, travel at zero "motion in proper time"?

Let us consider what it means to travel at zero "motion in proper time". For objects that travel at $v = c$ or equivalently:

$$v_\tau = 0 \tag{12.3.186}$$

if such objects had a space-time rest frame, then in their rest frame the moment they are observed to come into existence would be separated by the moment they go out of existence by exactly zero seconds. This follows directly from the standard relation:

$$c\tau = c\frac{t}{\gamma} = t\sqrt{c^2 - c^2} = 0 \qquad (12.3.187)$$

that relates proper time to coordinate, or we can simply conceptualize it from the fact that for $v_\tau = 0$ no proper time interval is traversed. This means, in particular, that if such objects had a space-time rest frame, then they would observe their own duration of existence in space-time to be exactly zero! (it is kinda a "photon existence paradox" because naively, if we didn't already know that photons exist, then surely this would have been used as an argument to justify why they could not exist...).

Now imagine that we could somehow transform to a frame in which our motion in proper time is zero. Then you would be sharing our rest frame with objects that observe their own duration of existence in space-time to be exactly zero. But, we already had a non-zero duration of existence in space-time before we transformed to that frame. As a result, if it was possible for us to reach the speed of light, then an observer could in our frame observe their own duration of existence in space-time to be both zero and non-zero. This is a contradiction, so that means we cannot transform to a frame in which our motion in proper time is zero, which is to say, we (or any other space-time observer) cannot transform to a frame in which we travel at the speed of light, and since we cannot reach the speed of light in space, we cannot break it, either, since in order to break it, we have to reach it first.

This can be illustrated in a schematic way (quarter of circle) named the "cosmic speedometer".

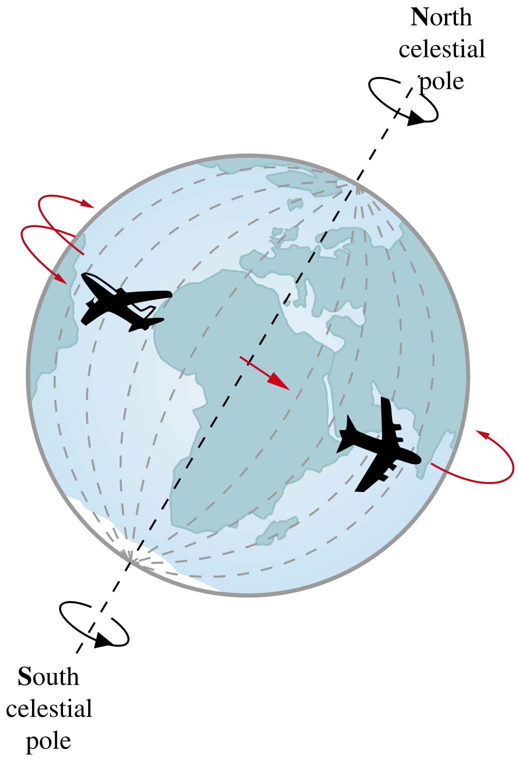
Let us see now two applied examples of what we have seen so far[40] that are so famous that they have their own name...:

12.3.2.8.2 Hafele–Keating experiment (special relativity version)

In 11971 (holocene calendar), direct experimental verification of time dilation was performed. Two airplanes in whose had been placed a cesium atomic clock during their regular commercial flights (one flying to the east, the other to west) compared their clocks to a third atomic clock remained on the ground (see [220]). This experiment made famous by time is named today "Hafele-Keating experiment" (Joseph C. Hafele, a physicist, and Richard E. Keating, an astronomer).

---

[40] Apart from these two famous examples, time dilatation helps particle physicists at the LHC (CERN) to track some particles during high energy collisions that have a too short lifetime to be correctly observed and studied but that thanks to their high speed have a significant longer proper life time due to time dilatation! Therefore with time dilatation particle physicists have the opportunity to study some specific particles properties what would not be possible without! Hourrra!



Because the Hafele–Keating experiment has been reproduced by increasingly accurate methods, there has been a consensus among physicists since at least the 11970s (holocene calendar) that the relativistic predictions of gravitational and kinematic effects on time have been conclusively verified. Criticisms of the experiment did not address the subsequent verification of the result by more accurate methods, and have been shown to be in error.

The idea is obviously that in a frame of reference at rest with respect to the center of the Earth, a clock aboard the plane moving eastward, in the direction of the Earth's rotation, has a greater velocity (resulting in a relative time loss) than one that remained on the ground, while a clock aboard the plane moving westward, against the Earth's rotation, had a lower velocity than one on the ground.

The plane flying eastward lost 59 [ns] while the plane flying westward gained 273 [ns] (the Earth rotates on itself in a day, from West to East. It was therefore measured a total difference:

$$\Delta t = t_{\text{East}} - t_{\text{West}} = (-59) - (+273) \cong -332\,[\text{ns}] \tag{12.3.188}$$

between the two clocks and this difference is even statistically significantly greater than that the one implied by Special Relativity (see detailed calculations just below).

Let us analyse the experience considering that all repositories are inertial (thus eliminating General Relativity).

Remark

Strictly speaking, the effect of General Relativity (slowing of clocks depending on the altitude in accordance with Einstein's effect proved in the section of General Relativity at page 5796) is absolutely not negligible since it is equivalent in amplitude that of Special Relativity. This is why we will study this experiment again in the section of General

Relativity.

Let us consider for our study three inertial reference points, one at the North Pole, one on Earth (elsewhere apart from the North Pole in the idea!) and one in a plane. The time intervals $t_{\text{North}}, t_{\text{Earth}}$ and t_{plane} respectively (which we will abbreviated t_N, t_E, t_P for the following developments), are connected by the previously proven relations (so the North pole is taken as the reference at rest in this experience and therefore the reference proper time!):

$$t_N = \gamma_E t_E = \gamma_P t_P \tag{12.3.189}$$

where we have:

$$\gamma_T = \frac{1}{\sqrt{1-\left(\frac{v_E}{c}\right)^2}} \qquad \gamma_P = \frac{1}{\sqrt{1-\left(\frac{v_P}{c}\right)^2}} \tag{12.3.190}$$

The repository on Earth and in the plane so have equation relative speeds v_E and v_P relative to the North Pole. The time by plane and on Earth are then linked by:

$$t_P = \frac{\gamma_E}{\gamma_P} t_E = \frac{\sqrt{1-\left(\frac{v_E}{c}\right)^2}}{\sqrt{1-\left(\frac{v_P}{c}\right)^2}} t_T \tag{12.3.191}$$

We will now rewrite this relation:

$$t_P = \frac{\gamma_E}{\gamma_P} t_T = \sqrt{\frac{1-(v_P/c)^2-(v_E/c)^2+(v_P/c)^2}{1-(v_P/c)^2}} t_E = \sqrt{1+\frac{(v_P/c)^2-(v_E/c)^2}{1-(v_P/c)^2}} t_E \tag{12.3.192}$$

We will accept the following approximation:

$$t_P = \frac{\gamma_E}{\gamma_P} t_E \cong \sqrt{1+[(v_P/c)^2-(v_E/c)^2]} t_E \tag{12.3.193}$$

where we have assumed that at the denominator:

$$(v_P/c)^2 \ll 1 \tag{12.3.194}$$

For square roots whose value is anyway close to 1 (since c is much larger than the considered relative speeds), we can do a Maclaurin expansion to the second order as x approaches zero (see section Sequences and Series page 1703):

$$\sqrt{1+x} \cong (1+x) + \frac{1}{1!}\frac{1}{2\sqrt{1+a}}(x-a) \underset{a=0}{=} = 1 + \frac{1}{2}x \tag{12.3.195}$$

Then we can write:

$$t_P = \frac{\gamma_E}{\gamma_P} t_P \cong \left(1+\frac{(v_P/c)^2-(v_E/c)^2}{2}\right) t_E \tag{12.3.196}$$

Thanks to these tricky successive approximations, we can easily write the difference between the two clocks that is then:

$$\Delta t_{AE} = t_P - t_E \cong t_E \frac{(v_P/c)^2-(v_E/c)^2}{2} \tag{12.3.197}$$

According to the initial assumptions, the cruising speed of the two aircraft from the ground is constant and is denoted v. The speed of each plane (!non-relativistic according to the preceding approximations) is then:

$$v_P=v_E-v \tag{12.3.198}$$

for the plane going eastwards and:

$$v_P=v_E+v \tag{12.3.199}$$

for the aircraft going respectively westward. So:

$$\begin{aligned}\Delta t=t_{\text{East}}-t_{\text{West}}&=\left(t_E\frac{\left(\dfrac{v_E-v}{c}\right)^2-\left(\dfrac{v_E}{c}\right)^2}{2}\right)-\left(t_E\frac{\left(\dfrac{v_E+v}{c}\right)^2-\left(\dfrac{v_E}{c}\right)^2}{2}\right)\\&=t_E\left(\frac{\left(\dfrac{v_E-v}{c}\right)^2}{2}-\frac{\left(\dfrac{v_E+v}{c}\right)^2}{2}\right)\end{aligned} \tag{12.3.200}$$

We will consider that (it's pretty rough ...):

$$\frac{v_E^2}{c^2}\ll\frac{2vv_E}{c^2}\frac{v^2}{c^2}\quad\ll\frac{2vv_E}{c^2} \tag{12.3.201}$$

So it remains:

$$\Delta t=t_{\text{East}}-t_{\text{West}}\cong t_E\left[\frac{-\dfrac{2vv_E}{c^2}}{2}-\frac{\dfrac{2vv_E}{c^2}}{2}\right]=t_E\left(-\frac{vv_E}{c^2}-\frac{vv_E}{c^2}\right)=-t_E\frac{2vv_E}{c^2} \tag{12.3.202}$$

We see well that obviously with the previous approximations we lost the asymmetry of time dilatation between East and West. The reader that this should disturb can then apply directly the numerical values in the prior previous relation.

The previous result that we got with all successive approximations already leaded us to see formally and quickly that the sign of the result will be in agreement with experimental results.

For a practical numerical application, we will take the constant speed of the commercial planes at that time that was:

$$v\cong 220\,[\text{m}\cdot\text{s}^{-1}] \tag{12.3.203}$$

and the total time travel of planes was of 41 hours according to the measurement at the ground, thus:

$$t_E=147,600\,[\text{s}] \tag{12.3.204}$$

and a point at the equator of the Earth's surface moves at the speed:

$$v_E=\frac{2\pi R}{t_{\text{day}}} \tag{12.3.205}$$

where the Earth's radius being of $6,371$ [km] (this suppose that the plans are above the equator radius). We then have applying all that numerical values:

$$\Delta t = -t_E \frac{2vv_E}{c^2} = -147,600 \frac{2 \cdot 200 \dfrac{2\pi \cdot 6,371,000}{24 \cdot 60 \cdot 60}}{299,792,458^2} = -147,600 \frac{2 \cdot 220 \cdot 463}{299,792,458^2} \cong 334\,[\text{ns}] \tag{12.3.206}$$

which leads to a result very close to the measurement that was performed.

And using directly the non-approximate version:

$$\begin{aligned}\Delta t = t_{\text{East}} - t_{\text{West}} &= \left(t_E \frac{\left(\dfrac{v_E - v}{c}\right)^2 - \left(\dfrac{v_E}{c}\right)^2}{2} \right) - \left(t_E \frac{\left(\dfrac{v_E + v}{c}\right)^2 - \left(\dfrac{v_E}{c}\right)^2}{2} \right) \\ &= -92 + (+173) = -265\,[\text{ns}]\end{aligned} \tag{12.3.207}$$

where we took this time the speed of the Earth at latitude consistent with the experience in 11971 (holocene calendar):

$$v_E = \frac{2\pi R \cos(38^\circ)}{t_{\text{day}}} \cong 365\,[\text{m} \cdot \text{s}^{-1}] \tag{12.3.208}$$

So we see that the result is therefore not very consistent with the experience! Indeed, we must now consider in that approximated case the time dilatation due to gravity. We'll have to use the Einstein's effect relation proved in the section of General Relativity (see page 5796) for approximated locally flat space:

$$\frac{\Delta t}{t_0} = \frac{gh}{c^2} \tag{12.3.209}$$

which expresses for recall the that at the ground time flows slower than the time at altitude h.

According to the records of the experiment, the aircraft flew at $10,000$ [m] above sea level. What gives (the acceleration g is not the same on the ground level than in altitude for recall!) and acceleration of time of:

$$\Delta t = \frac{gh}{c^2} t_0 \cong \frac{9.5 \cdot 10,000}{299,792,458^2} \cdot 147,600 = +156\,[\text{ns}] \tag{12.3.210}$$

But, we see that the two aircraft were both at the same height, we always have:

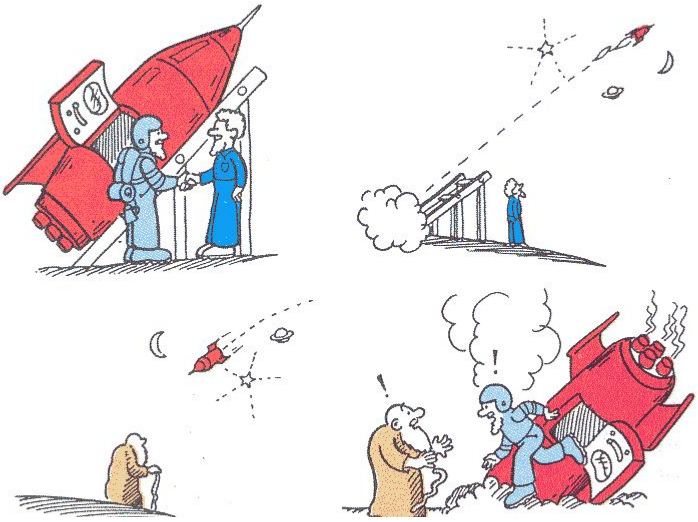
$$\Delta t = t_{\text{East}} - t_{\text{West}} = (-92 - 156) - (+173 - 156) = -265\,[\text{ns}] \tag{12.3.211}$$

So either there are other effects, of the order of General Relativity, which should be taken into account to explain the 67 [ns] of difference to the experience, or it is a accuracy problem of the time accuracy of the time clock at the time of the experiment.

In fact, we will see a detailed study of this experience in the section of General Relativity and see that the theoretical values are in very good agreement with experimental results.

12.3.2.8.3 Twins paradox

The twin paradox is a thought experiment in Special Relativity involving identical twins, one of whom makes a journey into space in a high-speed rocket and returns home to find that the twin who remained on Earth has aged more. This result appears puzzling because each twin sees the other twin as moving, and so, according to an incorrect and naive application of time dilation and the principle of relativity, each should paradoxically find the other to have aged more slowly. However, this scenario can be resolved (again contrary to a popular misconception) within the standard framework of Special Relativity: the travelling twin's trajectory involves two different inertial frames, one for the outbound journey and one for the inbound journey, and so there is no symmetry between the space-time paths of the two twins. Therefore, the twin paradox is not a paradox in the sense of a logical contradiction.



We can already consider the famous twin paradox in the framework of Special Relativity to show that the twin paradox does not only apply to non inertial systems. This is a rough approach (knowing that we will rigorously discussed the subject in the section of General Relativity).

Let us consider a rocket taking off at time t zero of the Earth and accelerating to 20 times the acceleration of Earth's gravity g up to a cruising speed of 90% the speed of light c. Let us suppose that the rocket continues at this speed during a terrestrial year and decelerate with the same intensity to resume its journey to Earth and accelerates again for its approach to the Earth and decelerate once again to its final zero velocity:

$$\Delta v=a\Delta t=20g(\Delta t_{\text{acc.}})=0.9c\Rightarrow\Delta t_{\text{acc.}}=\sqrt{\frac{0.9c}{20g}}=\sqrt{0.9c}20\cdot\cong 1.35\cdot 10^6\cong 15.6\,[\text{days}] \tag{12.3.212}$$

Thus, the total proper time spent for a human remained on Earth is:

$$\Delta {t'}_T\cong 15.6\cdot 4+365\cdot 2\cong 792.4\,[\text{days}] \tag{12.3.213}$$

For the traveller in the rocket, the proper time during the acceleration phase will be given roughly by:

$$\mathrm{d}{t'}_{a,f}=\sqrt{1-\frac{v^2}{c^2}}\mathrm{d}t=\sqrt{1-\frac{1}{c^2}(v)^2}\mathrm{d}t=\sqrt{1-\frac{1}{c^2}(gt)^2}\mathrm{d}t=\sqrt{1-\frac{g^2}{c^2}t^2}\mathrm{d}t \tag{12.3.214}$$

Thus by integrating (using the usual primitive proved in the section of Differential and Integral Calculus page 1534) for one of the phase of acceleration of the rocket it gives:

$$\begin{aligned}\Delta t'_{a,f}&=\int\limits_0^{\Delta t_{\text{acc}}}\sqrt{1-\frac{g^2}{c^2}t^2}\mathrm{d}t=\frac{g}{c}\int\limits_0^{\Delta t_{\text{acc}}}\sqrt{\left(\frac{c}{g}\right)^2-t^2}\mathrm{d}t\\&=\left[\frac{\left(\frac{c}{g}\right)^2}{2}\arcsin\left(\frac{t}{\frac{c}{g}}\right)+\frac{t}{2}\sqrt{\left(\frac{c}{g}\right)^2-t^2}\right]_0^{1.35\cdot 10^6}\qquad\cong 1.13\cdot 10^6\cong 13.1\,[\text{days}]\end{aligned}\tag{12.3.215}$$

And the proper time for the part with the constant cruising speed:

$$\Delta t'_{\text{cruising},f}=\int\limits_0^{365\,[\text{days}]}\sqrt{1-\frac{v^2}{c^2}}\mathrm{d}t\cong\int\limits_0^{365\,[\text{days}]}\sqrt{1-0.81}\mathrm{d}t=\sqrt{0.19}\int\limits_0^{365\,[\text{days}]}\mathrm{d}t\cong 159.1\,[\text{d}]\tag{12.3.216}$$

And therefore the total proper time in the rocket is:

$$\Delta t'_{\text{cruising},f}\cong 13.1+159.1=370.6\,[\text{days}]\tag{12.3.217}$$

So compared to the person remained on Earth, the one that was in the rocket has aged about half !!! This is a paradox (rather a "sophism" in reality) because we can not accurately apply Special Relativity to non-inertial frames. Nevertheless, even with General Relativity, there is a time difference!

Caution! But we haven't solved the Langevin's paradox here! Because in the framework of Special Relativity both observers can be switched independently. And that's the problem of making the calculations with Special Relativity: the acceleration is relative. That doesn't make sense in real life (even if it's possible on the paper) because in reality, one of the observer will feel an acceleration and deceleration when the other one won't. And that's what General Relativity can deal with: it makes acceleration absolute!

12.3.2.9 Apparent relativistic mass

First the reader must take careful (!!) The title is misleading by tradition! We will see why a little further below.

Meanwhile, imagine a frontal collision between two identical objects (1) and (2) having in the repository R_0 equal but opposite speeds. We will assume that the collision is elastic, that is to say that the kinetic energy and momentum are conserved.

Before the shock (collision), the components of objects speeds (1) and (2) are:

$$1=\begin{pmatrix}u_1\\v_1\end{pmatrix}\qquad 2=\begin{pmatrix}u_2\\v_2\end{pmatrix}=\begin{pmatrix}-u_1\\-v_1\end{pmatrix}\tag{12.3.218}$$

as shown below:

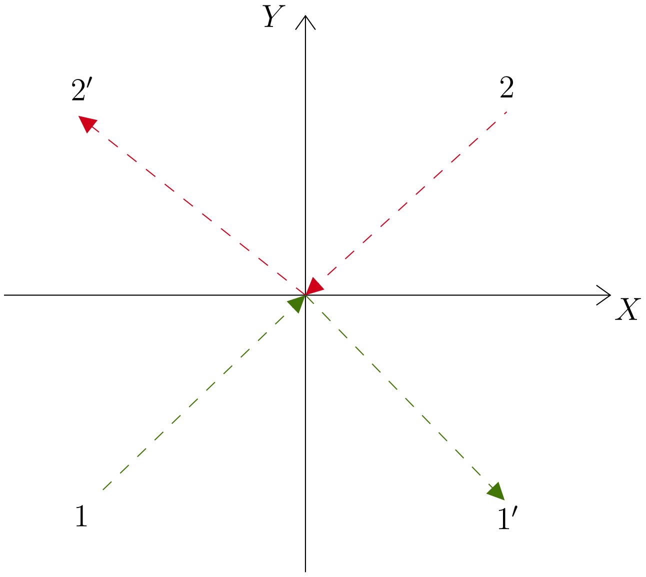


Figure 12.201: Configuration for the study of the apparent relativistic mass variation seen from R_0

After the collision, we have:

$$1'=\begin{pmatrix}u'_1\\v'_1\end{pmatrix}=\begin{pmatrix}u_1\\-v'_1\end{pmatrix}\qquad 2'=\begin{pmatrix}u'_2\\v'_2\end{pmatrix}=\begin{pmatrix}u_2\\-v_2\end{pmatrix}=\begin{pmatrix}-u_1\\v_1\end{pmatrix}\tag{12.3.219}$$

We will now apply the following Lorentz transformation:

- We give ourselves another repository R and assume that the repositories R_0 and R are in uniform translation speed u_1 along the x-axis in the positive direction (that is to say in the same direction and at the same horizontal speed than the particle (1)).
- For our particle (1) its trajectory became such is present not visible speed anymore along the x-axis.

Let's go! Let us place ourselves in repository R that moves relative to R_0 with the speed u_1 following the x-axis, the components of the speeds in this repository are ten before the collision:

$$1=\begin{pmatrix}U_1\\V_1\end{pmatrix}\qquad 2=\begin{pmatrix}U_2\\V_2\end{pmatrix}\tag{12.3.220}$$

and after the collision:

$$1'=\begin{pmatrix}U'_1\\V'_1\end{pmatrix}\qquad 2'=\begin{pmatrix}U'_2\\V'_2\end{pmatrix}\tag{12.3.221}$$

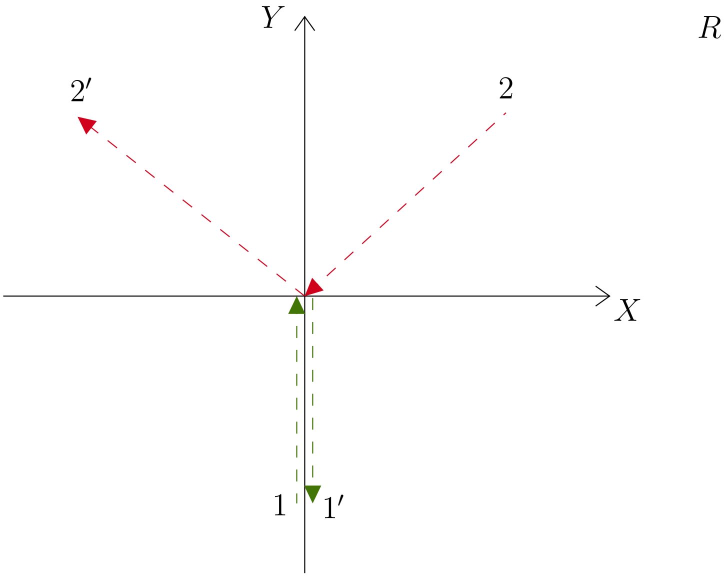


Figure 12.202: **Configuration for the study of the apparent relativistic mass variation seen from R**

So we have trivially in the reference frame R:

$$U'_1=U_1=0 \qquad V'_1=-V_1 \qquad U'_2=U_2 \qquad V'_2=-V_2 \tag{12.3.222}$$

but by applying the law of composition of speeds proved earlier above:

$$V'=\frac{x'}{t'}=\frac{v_2-v_1}{1-\dfrac{v_1v_2}{c^2}} \tag{12.3.223}$$

for the components of the horizontal axis we always have in the reference frame R:

$$\begin{aligned}&U_1=0\\&U_2=\frac{u_2-u_1}{1-\dfrac{u_2u_1}{c^2}}=\frac{(-u_1)-u_1}{1-\dfrac{(-u_1)(u_1)}{c^2}}=-\frac{2u_1}{1+\dfrac{u_1^2}{c^2}}\end{aligned} \tag{12.3.224}$$

and for the vertical movement, we have seen earlier above that:

$$v'_y=\frac{v_y}{\gamma\left(1-\dfrac{v}{c^2}v_x\right)} \tag{12.3.225}$$

Therefore we get:

$$\begin{aligned}&V_1=\frac{v_1}{\gamma\left(1-\dfrac{u_1}{c^2}u_1\right)}=\frac{v_1}{\gamma\left(1-\dfrac{u_1^2}{c^2}\right)}=\frac{\sqrt{1-\dfrac{u_1^2}{c^2}}v_1}{1-\dfrac{u_1^2}{c^2}}=\frac{v_1}{\sqrt{1-\dfrac{u_1^2}{c^2}}}=v_1\gamma\\&V_2=\frac{v_2}{\gamma\left(1-\dfrac{u_1}{c^2}u_2\right)}=\frac{v_2}{\gamma\left(1-\dfrac{u_1u_2}{c^2}\right)}=\frac{-v_1}{\gamma}\frac{1}{1+\dfrac{u_1^2}{c^2}}\end{aligned} \tag{12.3.226}$$

Passing from R_0 to R, the component following y of the total momentum must remain zero (as it was the case in R_0 initially). But:

$$V_1+V_2=\frac{v_1}{\gamma}\frac{1}{1-\dfrac{u_1^2}{c^2}}-\frac{v_1}{\gamma}\frac{1}{1+\dfrac{u_1^2}{c^2}}\neq 0 \tag{12.3.227}$$

To break this deadlock, we must admit that the respective apparent masses m_1 and m_2 may not be identical in R. So that brings us to require that:

$$|m_1V_1|=\frac{m_1v_1}{\gamma}\frac{1}{1-\dfrac{u_1^2}{c^2}}=|m_2V_2|=\frac{m_2v_1}{\gamma}\frac{1}{1+\dfrac{u_1^2}{c^2}} \tag{12.3.228}$$

which leads us to:

$$\frac{m_2}{m_1}=\frac{1+\dfrac{u_1^2}{c^2}}{1-\dfrac{u_1^2}{c^2}} \tag{12.3.229}$$

In R, the square of the norm of the two objects speeds gives:

$$\begin{gathered}U_1^2+V_1^2=V_1^2=v_1^2\gamma^2\\ U_2^2+V_2^2=\left(\frac{-2u_1}{1+\dfrac{u_1^2}{c^2}}\right)^2+\left(\frac{-v_1}{\gamma}\frac{1}{1+\dfrac{u_1^2}{c^2}}\right)^2=\frac{4u_1^2+v_1^2\left(1-\dfrac{u_1^2}{c^2}\right)}{\left(1+\dfrac{u_1^2}{c^2}\right)^2}=\frac{4u_1^2+V_1^2\left(1-\dfrac{u_1^2}{c^2}\right)^2}{\left(1+\dfrac{u_1^2}{c^2}\right)^2}\end{gathered} \tag{12.3.230}$$

The last relation can be written:

$$\frac{U_2^2+V_2^2}{c^2}\left(1+\frac{u_1^2}{c^2}\right)^2\left(1+\frac{u_1^2}{c^2}\right)^2=4\frac{u_1^2}{c^2}+\frac{V_1^2}{c^2}\left(1-\frac{u_1^2}{c^2}\right)^2=\left(1+\frac{u_1^2}{c^2}\right)^2-\left(1-\frac{u_1^2}{c^2}\right)^2+\frac{V_1^2}{c^2}\left(1-\frac{u_1^2}{c^2}\right)^2 \tag{12.3.231}$$

so that after rearrangement and factorization we get:

$$\left(\frac{U_2^2+V_2^2}{c^2}-1\right)\left(1+\frac{u_1^2}{c^2}\right)^2=\left(1-\frac{u_1^2}{c^2}\right)\left(\frac{V_1^2}{c^2}-1\right) \tag{12.3.232}$$

Therefore:

$$\frac{\left(1+\dfrac{u_1^2}{c^2}\right)^2}{\left(1-\dfrac{u_1^2}{c^2}\right)^2}=\frac{1-\dfrac{V_1^2}{c^2}}{1-\dfrac{U_2^2+V_2^2}{c^2}}=\frac{\Gamma_2^2}{\Gamma_1^2} \tag{12.3.233}$$

We thus found:

$$\frac{m_2}{m_1}=\frac{1+\dfrac{u_1^2}{c^2}}{1-\dfrac{u_1^2}{c^2}}=\frac{\Gamma_2}{\Gamma_1}\Leftrightarrow\frac{m_2}{\Gamma_2}=\frac{m_1}{\Gamma_1^2} \tag{12.3.234}$$

In the case as assumed above where both object are identical we will put $m_1 = m_2 = m_0$ and therefore:

$$\frac{m_0}{\Gamma_2} = \frac{m_0}{\Gamma_1} \tag{12.3.235}$$

And we will put them as an apparent relativistic mass denoted simply by m such that:

$$m = \frac{m_0}{\Gamma_2} = \frac{m_0}{\Gamma_1} \tag{12.3.236}$$

And as V_1^2 and $U_2^2 + V_2^2$ are simply the square norm of the velocity, we can write:

$$m = \frac{m_0}{\Gamma_2} = \frac{m_0}{\Gamma_1} = \gamma m_0 \tag{12.3.237}$$

So that finally:

$$\boxed{m = \gamma m_0 = \frac{m_0}{\sqrt{1 - \dfrac{v^2}{c^2}}}} \tag{12.3.238}$$

So we see that when $v = 0$, we have $m = m_0$ this is why we name m_0 the "rest mass" or "invariant mass".

Since the mass is a function of v (at least in appearance), some physicists note the rest mass as a function, that is to say: $m(0)$. But it is rather more common to use the notation m_0 to not have to tract too many parentheses in developments...

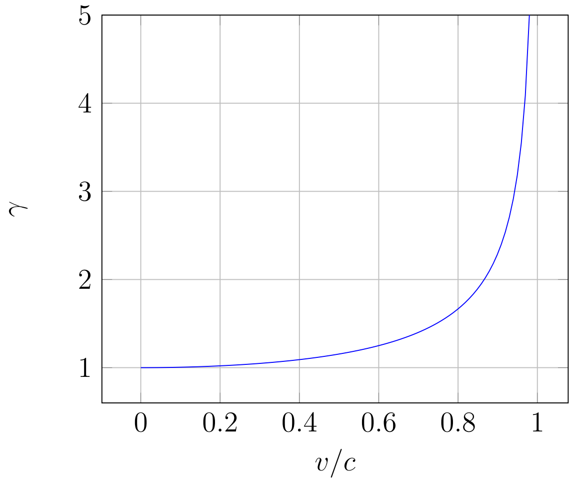


Figure 12.203: **Lorentz factor mass (and time) dilatation function trend**

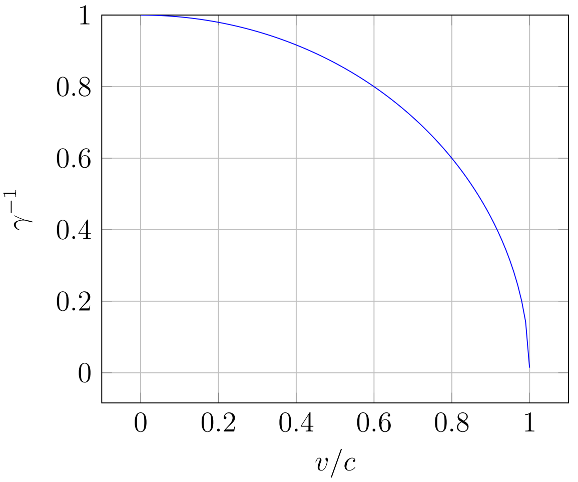


Figure 12.204: **Lorentz factor length contraction function trend (reminder)**

As the Michelson-Morley factor γ tends to infinity when the speed v approaches the speed c of light in a vacuum we have an additional reason to say that c is the upper limit assigned to the speed of any material object otherwise the apparent mass m would be infinite also, which is consistent with both the experience and the consequences already formulated by the Lorentz transformations!!!

It already follows an important conclusion: there are therefore two types of particles, those with a mass and will never go to the speed of light (as it then takes an infinite energy to get them there following our previous result), and those having a zero mass and which will therefore necessarily be at the speed of light.

As we will see it in the section of Quantum Field Theory interaction forces are short-range precisely because of the uncertainty principle and of the above statement. The greater the distance is between large particles that interacts together, the more time will be longer and therefore the smaller will be the energy involved. But in the case where the particle of interaction have no mass, the "force" is a long range one.

12.3.2.9.1 Why is Gold yellow?

One of the most famous use of Special Relativity in everyday life, and especially of relativistic mass, is very likely the legendary argument that gold should have a silvery shine just like other metals and the reason why it appears golden is said to be explained quite well by using Special Relativity! We will see what is the underlying idea, even if sadly, it's quite an urban legend as the reality is much more complex and involves material physics (see our explanations at the end).

One of the most important and familiar results of Special Relativity is that the relativistic mass of the electron increases as:

$$m_{\text{rel}}=\frac{m_e}{\sqrt{1-(v_e/c)^2}} \tag{12.3.239}$$

where m_e, v_e, c are the electron rest mass, velocity of the electron, and speed of light respectively.

This has an immediate implication on the Bohr radius (r_0) (see section Corpuscular Quantum Physics page 4630), which is given by:

$$r_0=\frac{\hbar}{m_e c\alpha} \tag{12.3.240}$$

where $\hbar$ is the reduced Planck's constant, and α is the fine-structure constant ($\alpha=\frac{1}{4\pi\varepsilon_0}\frac{e^2}{\hbar c}$).

If one substitutes the "relativistic mass" into the equation for the Bohr radius it can be written:

$$r_{\text{rel}}=\frac{\hbar\sqrt{1-(v_e/c)^2}}{m_e c\alpha} \tag{12.3.241}$$

It follows that:

$$\frac{r_{\text{rel}}}{r_0}=\sqrt{1-(v_e/c)^2} \tag{12.3.242}$$

The above ratio of the relativistic and nonrelativistic Bohr radii has been plotted as a function of the electron velocity:

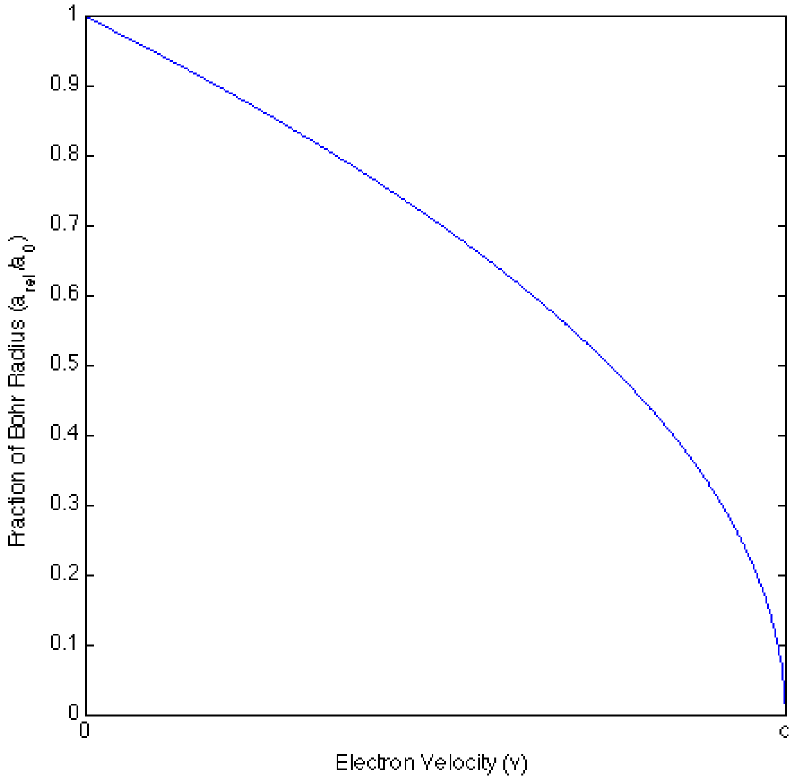


Figure 12.205: **Ratio of relativistic and nonrelativistic Bohr radii as a function of electron velocity (source: Wikipedia)**

Notice how the relativistic model shows the radius decreases with increasing velocity.

In the Bohr model, the angular momentum is given as we have already seen it by $m_e\omega_e r^2 = m_e v_e r = n\hbar$ (see section Corpuscular Quantum Physics page 4629). So multiplying both sides of the explicit definition of Bohr radius:

$$r = \frac{n^2\hbar^2 4\pi\varepsilon_0}{m_e Z e^2} \tag{12.3.243}$$

by $m_e v_e$, we get:

$$n\hbar = \frac{m_e v_e n^2\hbar^2 4\pi\varepsilon_0}{m_e Z e^2} \tag{12.3.244}$$

Simplifying and rearranging gives us:

$$v_e = \frac{Ze^2}{4\pi\varepsilon_0\hbar n} \tag{12.3.245}$$

Dividing both sides by c:

$$\frac{v_e}{c} = \frac{Ze^2}{4\pi\varepsilon_0\hbar c n} \tag{12.3.246}$$

Substituting this into the expression for the Bohr ratio mentioned above gives:

$$\frac{r_{\text{rel}}}{r_0} = \sqrt{1-\left(\frac{Ze^2}{4\pi\varepsilon_0\hbar c n}\right)^2} \tag{12.3.247}$$

At this point one can see that a low value of n and a high value of Z results in $\frac{r_{\text{rel}}}{r_0} \to 0$. This fits with intuition: electrons with lower principal quantum numbers will have a higher probability density of being nearer to the nucleus. A nucleus with a large charge will cause an electron to have a high velocity. A higher electron velocity means an increased electron relativistic mass, and as a result the electrons will be near the nucleus more of the time and thereby contract the radius for small principal quantum numbers.

For gold with $Z = 79$, $v_e \cong 0.58c$, so the $1s$ electron will be moving at 58% of the speed of light. Plugging this in for v/c in the equation for the relativistic mass, one finds that $m_{\text{rel}} = 1.22m_e$, and in turn putting this in for the Bohr radius above one finds that the radius shrinks by 22%.

The reflectivity (keep in mind that where reflectance is low, absorbance is high and vice-versa) of aluminium (Al), silver (Ag), and gold (Au) is shown in the graph below. The human eye sees electromagnetic radiation with a wavelength near 600 [nm] as yellow. Gold appears yellow because it absorbs blue light more than it absorbs other visible wavelengths of light; the reflected light reaching the eye is therefore lacking in blue compared to the incident light. Since yellow is complementary to blue, this makes a piece of gold under white light appear yellow to human eyes.

The electrons inside gold atom travel at speeds as high as half the speed of light. This causes relativistic effects (length contraction) and shrinks the shells closer to the nucleus such that the energy difference between the outermost shells now corresponds to blue light.

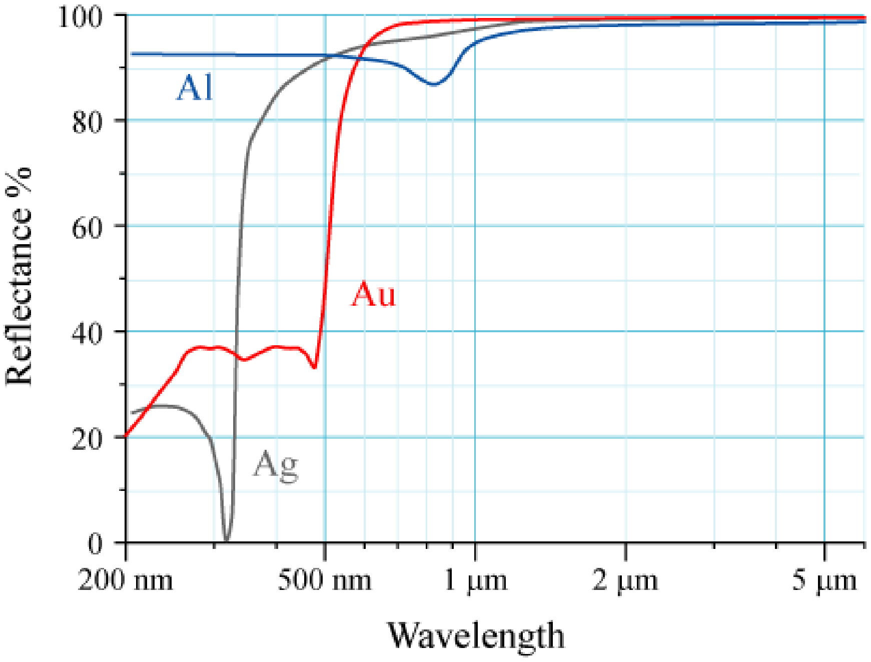


Figure 12.206: Spectral reflectance curves for aluminum, silver and gold metal mirrors (source: Wikipedia, author: Bob Mellish)

The electronic transition from the $5d$ orbital to the $6s$ orbital is responsible for this absorption.

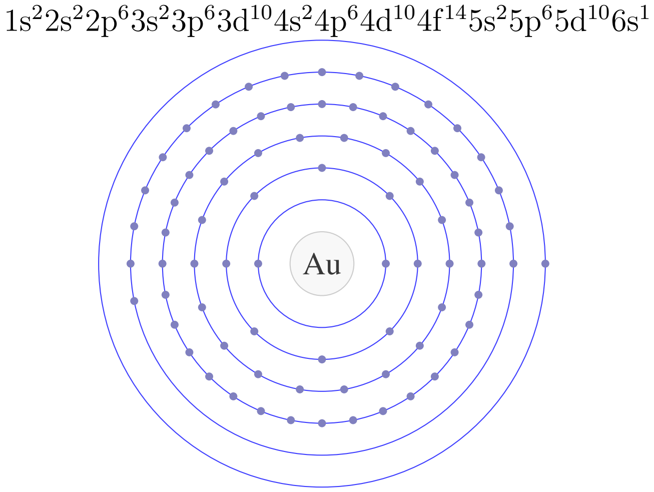
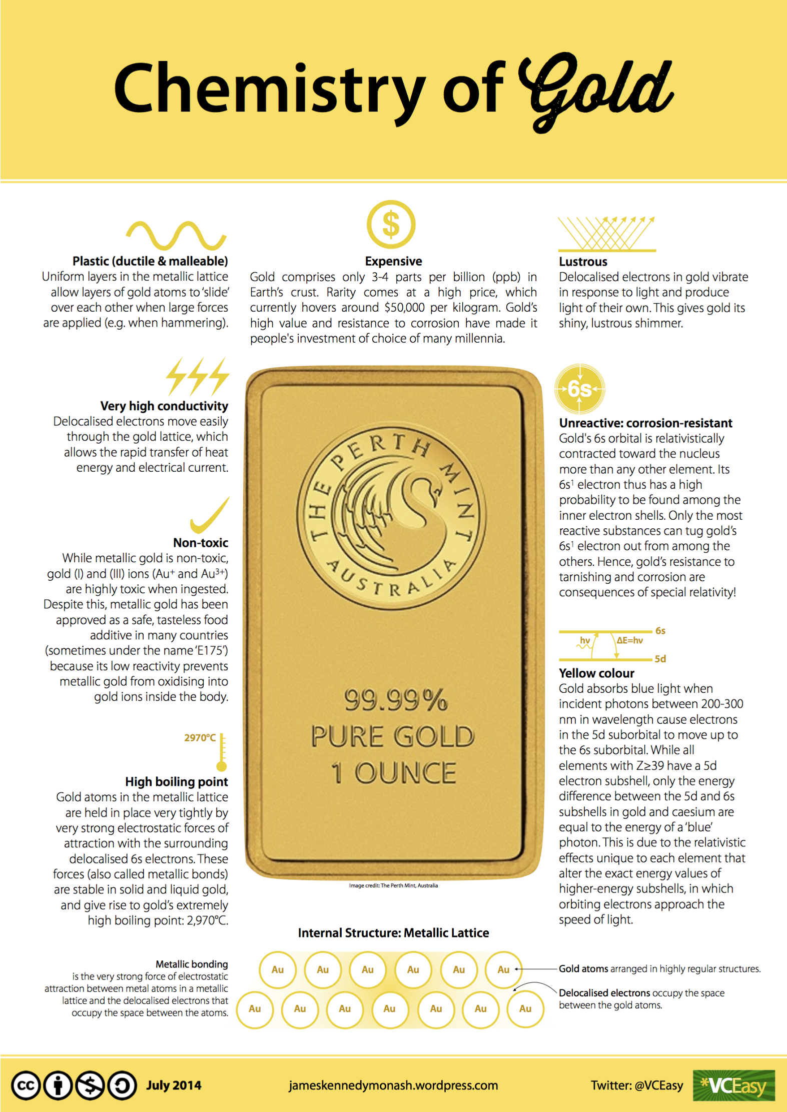

$1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} 4s^2 4p^6 4d^{10} 4f^{14} 5s^2 5p^6 5d^{10} 6s^1$
Au

Figure 12.207: **Gold Bohr naive model configuration**

When you compare color of pure elements, don't forget to compare only elements of the same column of the periodic table of elements (i.e. having the same electronic structure). Elements that occupy the same column on the periodic table (named a "group") have identical valence electron configurations and consequently behave in a similar fashion chemically!

An analogous transition occurs in silver (^{47}Ag), but the relativistic effects are smaller than in gold. While silver's $4d$ orbital experiences some relativistic expansion and the $5s$ orbital some contraction, the $4d-5s$ distance in silver is much greater than the $5d-6s$ distance in gold. The relativistic effects increase the $5d$ orbital's distance from the atom's nucleus and decrease the $6s$ orbital's distance.

Obviously the reality is much more complex (as always when we use toy models in physics) than using the simple Bohr model and assuming that a single atom behaves identically the same atoms in a bulk material. A robust treatment of the topic would involve Quantum Chemistry techniques (with more and more electrons being added, the difficulties mount rapidly because of screening and other factors) that can't be solved by hand and that would require intensive usage of Numerical Methods. All that means, is that the last few outer electrons of Au see a slightly different potential than they would have if there were no Special Relativity. In the end, given the binding energies for Au compared with Cu and Ag there really is not much difference (highly similar band structure and Fermi surfaces). The color differences of them are all done to the plasmon absorption in the visible, a bulk feature, with slight different plasmon energies for Cu, Ag, and Au. For comparison, Ag has 58.3 and 63.7 [eV], and Cu 75.1 and 73.3 [eV], respectively. So, really, the outer electrons of Au are not particularly different from the outer electrons of the other noble metals as one would expect from their band structures. In a solid, there are many available states above the Fermi surface (essentially a continuum considering all wave vectors $\vec{k}$), and once an electron is put up there it rapidly thermalizes back down through electron-electron and electron-phonon interactions (order picoseconds or less). The atomic physics is just sometimes very different from the material physics!



Chemistry of Gold

Plastic (ductile & malleable)
Uniform layers in the metallic lattice allow layers of gold atoms to 'slide' over each other when large forces are applied (e.g. when hammering).

Expensive
Gold comprises only 3-4 parts per billion (ppb) in Earth's crust. Rarity comes at a high price, which currently hovers around $50,000 per kilogram. Gold's high value and resistance to corrosion have made it people's investment of choice of many millennia.

Lustrous
Delocalised electrons in gold vibrate in response to light and produce light of their own. This gives gold its shiny, lustrous shimmer.

Very high conductivity
Delocalised electrons move easily through the gold lattice, which allows the rapid transfer of heat energy and electrical current.

Non-toxic
While metallic gold is non-toxic, gold (I) and (III) ions (Au^+ and Au^{3+}) are highly toxic when ingested. Despite this, metallic gold has been approved as a safe, tasteless food additive in many countries (sometimes under the name 'E175') because its low reactivity prevents metallic gold from oxidising into gold ions inside the body.

High boiling point
Gold atoms in the metallic lattice are held in place very tightly by very strong electrostatic forces of attraction with the surrounding delocalised 6s electrons. These forces (also called metallic bonds) are stable in solid and liquid gold, and give rise to gold's extremely high boiling point: 2,970°C.

Unreactive: corrosion-resistant
Gold's 6s orbital is relativistically contracted toward the nucleus more than any other element. Its $6s^1$ electron thus has a high probability to be found among the inner electron shells. Only the most reactive substances can tug gold's $6s^1$ electron out from among the others. Hence, gold's resistance to tarnishing and corrosion are consequences of special relativity!

Yellow colour
Gold absorbs blue light when incident photons between 200-300 nm in wavelength cause electrons in the 5d suborbital to move up to the 6s suborbital. While all elements with Z≥39 have a 5d electron subshell, only the energy difference between the 5d and 6s subshells in gold and caesium are equal to the energy of a 'blue' photon. This is due to the relativistic effects unique to each element that alter the exact energy values of higher-energy subshells, in which orbiting electrons approach the speed of light.

Image credit: The Perth Mint, Australia

Internal Structure: Metallic Lattice

Metallic bonding is the very strong force of electrostatic attraction between metal atoms in a metallic lattice and the delocalised electrons that occupy the space between the atoms.

Gold atoms arranged in highly regular structures.

Delocalised electrons occupy the space between the gold atoms.

July 2014

jameskennedymonash.wordpress.com

Twitter: @VCEasy

Figure 12.208: Chemistry of Gold

12.3.2.9.2 Mass–Energy equivalence

Under the action of a force F, the speed of a mass m increases or decreases on each portion of the trajectory. The work of the component $F\mathrm{d}x$ can then be interpreted into kinetic energy $\mathrm{d}E_c$.

In the relativistic theory, the mass varies with speed as we have just prove it, therefore:

$$E_c = \int F\mathrm{d}x = \int_0^v \frac{\mathrm{d}p}{\mathrm{d}t}\mathrm{d}x = \int_0^v \mathrm{d}p\frac{\mathrm{d}x}{\mathrm{d}t} = \int_0^v \mathrm{d}(mv)v \qquad (12.3.248)$$

The integration by parts (see section Differential and Integral Calculus page 1532):

$$\int a\mathrm{d}b = ab - \int b\mathrm{d}a \qquad (12.3.249)$$

give us:

$$\begin{aligned}
E_c&=mv^2-\int\limits_0^v mv\mathrm{d}v=\frac{m_0v2}{\sqrt{1-\dfrac{v^2}{c^2}}}-\int\limits_0^v\frac{m_0v}{\sqrt{1-\dfrac{v^2}{c^2}}}\mathrm{d}v\\
&\underset{\substack{x=(v/c)^2\\ \mathrm{d}x=2v\mathrm{d}v/c^2}}{=}\frac{m_0v^2}{\sqrt{1-\dfrac{v^2}{c^2}}}-\frac{m_0c^2}{2}\int\limits_0^{(v/c)^2}\frac{1}{\sqrt{1-x}}\mathrm{d}x=\frac{m_0v^2}{\sqrt{1-\dfrac{v^2}{c^2}}}-\frac{m_0c^2}{2}\left[-2\sqrt{1-x}\right]_0^{(v/c)^2}\\
&=\frac{m_0v^2}{\sqrt{1-\dfrac{v^2}{c^2}}}+m_0c^2\left[\sqrt{1-x}\right]_0^{(v/c)^2}=\frac{m_0v^2}{\sqrt{1-\dfrac{v^2}{c^2}}}+m_0c^2\left(\sqrt{1-\frac{v^2}{c^2}}-1\right)\\
&=\frac{m_0v^2}{\sqrt{1-\dfrac{v^2}{c^2}}}+m_0c^2\sqrt{1-\frac{v^2}{c^2}}-m_0c^2=\frac{m_0v^2+m_0c^2\left(1-\dfrac{v^2}{c^2}\right)^2}{\sqrt{1-\dfrac{v^2}{c^2}}}-m_0c^2\\
&=\frac{m_0v^2+m_0c^2-m_0v^2}{\sqrt{1-\dfrac{v^2}{c^2}}}-m_0c^2=\frac{m_0c^2}{\sqrt{1-\dfrac{v^2}{c^2}}}-m_0c^2=mc^2-m_0c^2=\Delta mc^2
\end{aligned}\tag{12.3.250}$$

The gain of kinetic energy of a particle can be considered as gain in its apparent mass. Since m_0 is the rest mass, the quantity m_0c^2 is named "rest energy" of the particle.

We then have:

$$E_c=mc^2-m_0c^2\Leftrightarrow mc^2=E_c+m_0c^2\tag{12.3.251}$$

where E_c represents the energy of motion (kinetic energy).

The sum of:

$$E_c+m_0c^2\tag{12.3.252}$$

therefore represents the total energy E of the particle in the absence of the potential field. Which brings us to write:

$$mc^2=E_c+m_0c^2\tag{12.3.253}$$

hence:

$$\boxed{E=mc^2}\tag{12.3.254}$$

This relation is confirmed in numerous ways, one of the most famous and precise experimental tests at a $\sim5\sigma$ threshold (0.00004%) being [452].

This relation can be interpreted as the fact that mass is an emergent property of the energy (we know nowadays that's the mass is a property acquired by fundamental particles interacting with the Higgs field). Especially, that kinetic energy can be converted into mass and vice versa and this is typically what most particles accelerators do million of times every day.

Remark

Actually nothing seems to indicate in our observable universe, the existence of negative mass. Even anti-particles don't have a negative mass (they have opposite charge but not opposite mass)! So their energy/mass don't cancels. The CERN-SPS (Super Proton Synchrotron) proton-antiproton collider was the first hadron collider operating with bunched beams. Their collision generated plenty of other particles (internal quarks and gluons).

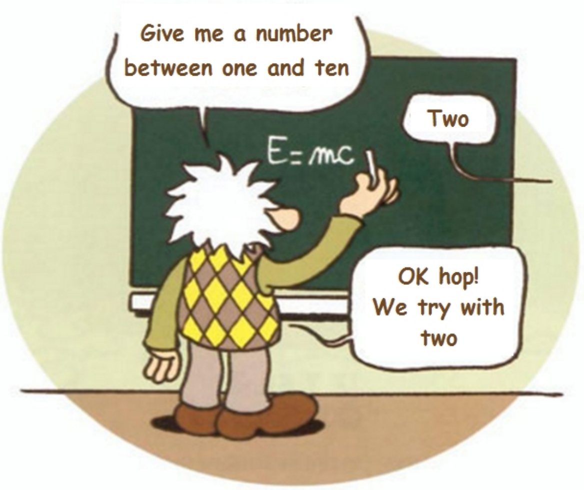
Many physicists like better to write that latter relation as following:

$$\Delta m=\frac{\Delta E}{c^2} \tag{12.3.255}$$

or in natural units (used by most professionals high level physicists):

$$E=m \tag{12.3.256}$$

So what you have to understand with this relation is that when for example (special case example!) you heat up an object in everyday life, you are increasing its mass even if it is not perceptible with a domestic balance!


Give me a number between one and ten
Two
E=mc
OK hop! We try with two

Finally we could also get the same result in another way using Lagrangian mechanics (see section Analytical Mechanics page 3152) as shown below:

12.3.2.9.3 Relativistic Lagrangian

The following developments will help us in the study of Electrodynamics (if this section has not been read yet), to determine the expression of the tensor of the electromagnetic field and in Relativistic Quantum Physics to determine the Klein-Gordon equation with magnetic field. So be sure to carefully read what follows.

In Special Relativity, so we want the equations of motion have the same form in all inertial frames. For this, we need the action S (see section Analytical Mechanics page 3161) to be invariant with respect to Lorentz transformations. Guided by this principle, trying to get the action of a free particle. Suppose that the action is in the reference frame O':

$$S_0=\int L_0\mathrm{d}t'=-mc^2\int \mathrm{d}t' \tag{12.3.257}$$

Remarks

R1. The choice of the minus sign will be evident in our study of electrodynamics.

R2. The notation L_0 instead of the L of Lagrangian lets just emphasize that this is a case study where the system is free. This distinction of notation will be useful in our study of General Relativity and determination of the Tensor of the Electromagnetic field in the section of Electrodynamics.

R3. We are not supposed to know what kind of mass we are dealing with (inertial rest mass), hence the fact that in the ignorance, we will work with the inertial mass m to perhaps correct this hypothesis later if necessary.

And let us recall that:

$$\mathrm{d}t=\frac{\mathrm{d}t'}{\sqrt{1-\dfrac{v^2}{c^2}}}\Leftrightarrow \mathrm{d}t'=\sqrt{1-\frac{v^2}{c^2}}\mathrm{d}t \tag{12.3.258}$$

In the repository O, then we have the "Lorentz invariant action":

$$S=-mc^2\int\sqrt{1-\frac{v^2}{c^2}}\mathrm{d}t \tag{12.3.259}$$

So according to our initial hypothesis, we have for the relativistic Lagrangian (in the absence of potential field... since the system is assumed to be "free"):

$$L_0=-mc^2\sqrt{1-\frac{v^2}{c^2}} \tag{12.3.260}$$

In the non-relativistic approximation $v\ll c$, we have following the Maclaurin development of the square root (see section Sequences and Series page 1703):

$$L_0=-mc^2\sqrt{1-\frac{v^2}{c^2}}\cong -mc^2\left(1-\frac{1}{2}\frac{v^2}{c^2}\right)=-mc^2+\frac{1}{2}mv^2 \tag{12.3.261}$$

We thus fall back on the usual Lagrangian of a free system in movement but more a constant $(-mc^2)$ that does not affect the equations of motion we got in Classical Mechanics but that will be absolutely necessary to us in Electrodynamics.

Let us recall now that the generalized momentum (see section Analytical Mechanics page 3180) is defined by:

$$p_i=\frac{\partial L}{\partial\dot{q}_i} \tag{12.3.262}$$

We will now see that this definition is not accidental. Indeed:

$$p=\frac{\partial L_0}{\partial v}=p\frac{\partial}{\partial v}\left(-mc^2\sqrt{1-\frac{v^2}{c^2}}\right)=-mc^2\left(\sqrt{1-\frac{v^2}{c^2}}\right)=mc^2\frac{v}{c^2\sqrt{1-\frac{v^2}{c^2}}}$$
$$=\frac{mv}{\sqrt{1-\frac{v^2}{c^2}}}\qquad(12.3.263)$$

The Hamiltonian (see section Analytical Mechanics page 3181) is equal to:

$$H(q_i,p_i)=\sum_{i=1}^{n}\dot{q}_ip_i-L(q_i,\dot{q}_i)\qquad(12.3.264)$$

Which gives:

$$H_0=vp-L_0=\frac{mv^2}{\sqrt{1-\frac{v^2}{c^2}}}+mc^2\sqrt{1-\frac{v^2}{c^2}}=\frac{mc^2}{\sqrt{1-\frac{v^2}{c^2}}}\qquad(12.3.265)$$

The Hamiltonian is in this case equal to the total energy of the particle. Its expression led us to change somewhat our initial hypothesis and finally to write m_0 instead of m in the expression of the action S.

So we finally have the Lagrangian of a free relativistic particle:

$$\boxed{L_0=-m_0c^2\sqrt{1-\frac{v^2}{c^2}}}\qquad(12.3.266)$$

and the corresponding Hamiltonian:

$$\boxed{H_0=\frac{m_0c^2}{\sqrt{1-\frac{v^2}{c^2}}}=mc^2}\qquad(12.3.267)$$

In the non-relativistic approximation $v\ll c$, H_0 becomes with a Maclaurin development (see section Sequences and Series page 1703):

$$H_0\cong m_0c^2+\frac{1}{2}m_0v^2\qquad(12.3.268)$$

We recognize the usual kinetic energy, plus a constant: the energy at rest. Which corresponds to the calculations we had made before where we got:

$$mc^2=E_c+m_0c^2=E\Rightarrow E=mc^2\qquad(12.3.269)$$

12.3.2.9.4 Relativistic (linear) momentum

The total energy E and the (linear) momentum $p=mv$ of a particle can therefore take any positive value (when the speed approaches the limit value c, the apparent mass suits for the product $p=mv$ to not be bounded).

In the expression of E, we can replace the speed v^2 by a function p^2:

$$p^2=m^2v^2=\frac{E^2}{c^4}v^2\Rightarrow v^2=\frac{p^2c^4}{E^2} \tag{12.3.270}$$

introduced into:

$$E=mc^2=\frac{m_0c^2}{\sqrt{1-\dfrac{v^2}{c^2}}} \tag{12.3.271}$$

we have:

$$E=\frac{m_0c^2}{\sqrt{1-\dfrac{p^2c^4}{E^2c^2}}}=\frac{m_0c^2}{\sqrt{E^2-p^2c^2}}E \tag{12.3.272}$$

Therefore:

$$\sqrt{E^2-p^2c^2}=m_0c^2\Rightarrow E^2-p^2c^2=m_0^2c^4 \tag{12.3.273}$$

hence (we will come back on that relation of the utmost importance during our proof of Einstein relation):

$$\boxed{E=c\sqrt{m_0^2c^2+p^2}} \tag{12.3.274}$$

We have not kept the negative part of the previous relation as it has no meaning in classical physics. However, when we will study Relativistic Quantum Physics, it will be essential to preserve it otherwise we will get absurdities.

However, we can obviously write this last relation also in the following form named "relativistic mass momentum relation":

$$\boxed{E^2=m_0^2c^4+p^2c^2} \tag{12.3.275}$$

or also (ugly!):

$$E^2=(E(m_0))^2+(E(c))^2 \tag{12.3.276}$$

In other words, the total energy of a moving particle is equal to its mass energy added to its kinetic energy (basically nothing new).

The relation above has two limit cases where we can simplify it:

1. For a particle at rest ($p=0$), we can reduce the expression to:

   $$E=m_0c^2 \tag{12.3.277}$$

   by omitting the negative energy ... at least for now (sadly the above relation is sometimes denoted $E=mc^2$ by omitting the 0 subscript).

2. We can apply the equation to a particle without mass to eliminate the first term, which then gives us:

$$\boxed{E=pc} \tag{12.3.278}$$

A photon, for example, has a zero rest mass but it is never at rest ...by definition, it is a quantum of energy, kinetic energy is never zero and so it has a mass corresponding to its kinetic energy. Thus, a massless particle at rest moves at the speed of light, regardless of the chosen repository frame! Conversely, a particle with a non-zero rest mass can never reach the speed of light in any repository.

Notice that the relativistic mass moment relation we have proved just earlier above is the first step that has lead us to the Dirac and Schrödinger equation as illustrated below:

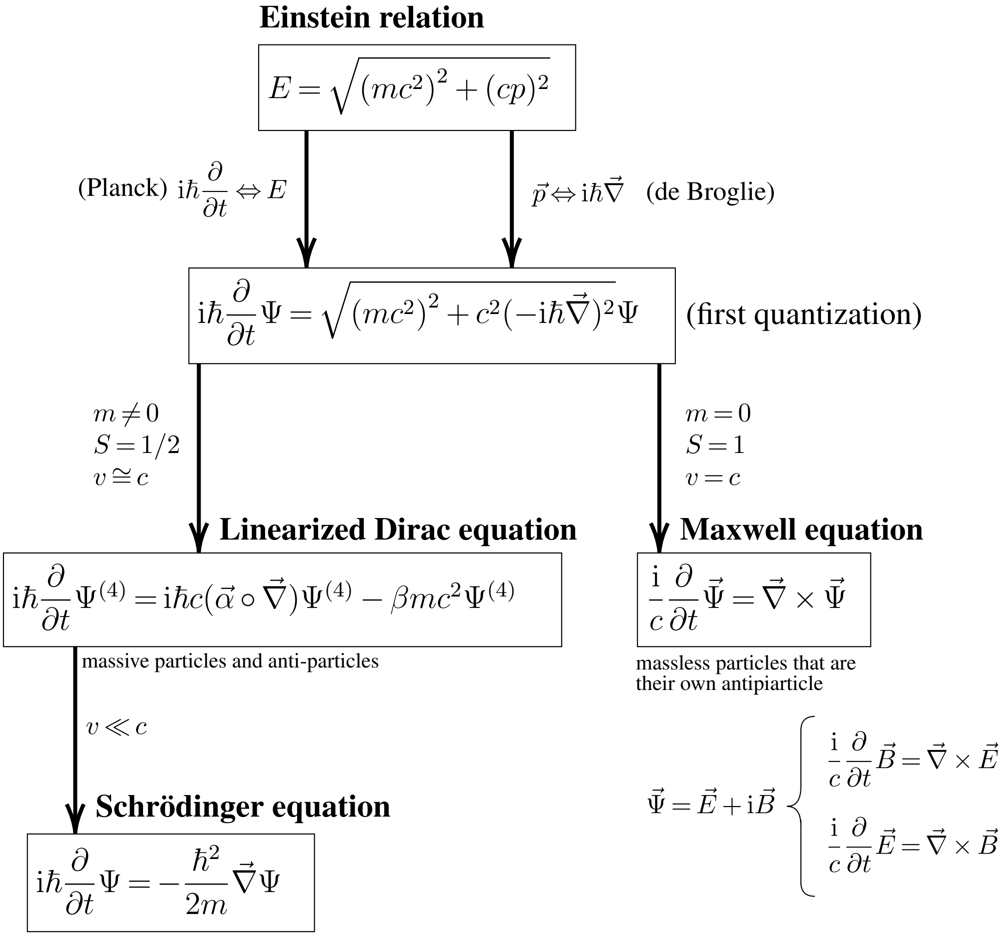

Einstein relation
$E=\sqrt{(mc^2)^2+(cp)^2}$
(Planck) $\mathrm{i}\hbar\frac{\partial}{\partial t}\Leftrightarrow E$
$\vec{p}\Leftrightarrow \mathrm{i}\hbar\vec{\nabla}$ (de Broglie)
$\mathrm{i}\hbar\frac{\partial}{\partial t}\Psi=\sqrt{(mc^2)^2+c^2(-\mathrm{i}\hbar\vec{\nabla})^2}\Psi$
(first quantization)
$m\neq 0$
$S=1/2$
$v\cong c$
$m=0$
$S=1$
$v=c$
Linearized Dirac equation
$\mathrm{i}\hbar\frac{\partial}{\partial t}\Psi^{(4)}=\mathrm{i}\hbar c(\vec{\alpha}\circ\vec{\nabla})\Psi^{(4)}-\beta mc^2\Psi^{(4)}$
massive particles and anti-particles
Maxwell equation
$\frac{\mathrm{i}}{c}\frac{\partial}{\partial t}\vec{\Psi}=\vec{\nabla}\times\vec{\Psi}$
massless particles that are their own antipiarticle
$v\ll c$
Schrödinger equation
$\mathrm{i}\hbar\frac{\partial}{\partial t}\Psi=-\frac{\hbar^2}{2m}\vec{\nabla}\Psi$
$\vec{\Psi}=\vec{E}+\mathrm{i}\vec{B}$
$\frac{\mathrm{i}}{c}\frac{\partial}{\partial t}\vec{B}=\vec{\nabla}\times\vec{E}$
$\frac{\mathrm{i}}{c}\frac{\partial}{\partial t}\vec{E}=\vec{\nabla}\times\vec{B}$

Figure 12.209: Relationships between most famous quantum physics equations

Remarks

R1. As we prove it further below (see the "Einstein relation"), from the construction of Planck's law (see section Thermodynamics page 3719), we can write $E=pc=h\nu$.

R2. The mass of the photon can hardly be non-zero! Indeed, quantum theory would be false otherwise. But it has never fail until now (see section Wave Quantum Physics page

4688). We would also have a small change on the Electrostatic force following that it is given by the Yukawa potential (see section Quantum Field Theory page 5164) and this would have been observed in laboratory since...

R3. The reason why some particles have masses and other not is related to Quantum Field Theory (QFT) and especially to something named the "Higgs mechanism" related to a quantum field named the "Higgs field". However why the Higgs field give masses to some particles are not to others is still an opened question in the early 121st century.

Let us now look after the relations between p and p' and between E and E', to make it possible for O' to write:

$$E'=c\sqrt{m_0^2c^4+p'^2} \tag{12.3.279}$$

We then begin to get rid of the square root:

$$E=c\sqrt{m_0^2c^4+p^2}\Rightarrow p^2-\frac{E^2}{c^2}=-m_0^2c^2 \tag{12.3.280}$$

If O write:

$$p_x^2+p_y^2+p_z^2-\frac{E^2}{c^2}=-m_0^2c^4 \tag{12.3.281}$$

O' must be able to write:

$$p'^2_{x'}+p'^2_{y'}+p'^2_{z'}-\frac{E'^2}{c^2}=-m_0^2c^4 \tag{12.3.282}$$

Therefore we have:

$$p_x^2+p_y^2+p_z^2-\frac{E^2}{c^2}=p'^2_{x'}+p'^2_{y'}+p'^2_{z'}-\frac{E'^2}{c^2} \tag{12.3.283}$$

If we identify:

$$x\equiv p_x \qquad y\equiv p_z \qquad z\equiv p_z \qquad ct\equiv\frac{E}{c} \tag{12.3.284}$$

we obtain similar expressions to those used for the Lorentz transformations of spatial and temporal components. We can then write, by similarity, that the changes to the (linear) momentum and energy are therefore given by:

$$\begin{aligned}
p'_{x'}&=\frac{p_x-v\dfrac{E}{c^2}}{\sqrt{1-\dfrac{v^2}{c^2}}}\Leftrightarrow p_x=\frac{p'_{x'}+v\dfrac{E'}{c^2}}{\sqrt{1-\dfrac{v^2}{c^2}}}\\
p'_{y'}&=p_y\Leftrightarrow p_y=p'_{y'}\\
p'_{z'}&=p_z\Leftrightarrow p_z=p'_{z'}\\
E'&=\frac{E-vp_x}{\sqrt{1-\dfrac{v^2}{c^2}}}\Leftrightarrow E=\frac{E'+vp'_x}{\sqrt{1-\dfrac{v^2}{c^2}}}
\end{aligned} \tag{12.3.285}$$

Again, if we take:

$$\gamma=\frac{1}{\sqrt{1-\dfrac{v^2}{c^2}}}=\frac{1}{1-\beta^2} \tag{12.3.286}$$

We therefore have by expressing all previous relations of transformation in the same units by remembering that $E\equiv pc$ (for a photon!):

$$\begin{aligned} p_x c &= \gamma({p'}_{x'}c+\beta E') \\ p_y c &= {p'}_{y'}c \\ p_z c &= {p'}_{z'}c \\ E &= \gamma(E'+\beta c{p'}_x) \end{aligned} \tag{12.3.287}$$

We can the define a matrix such that:

$$\begin{pmatrix} E \\ p_x c \\ p_y c \\ p_z c \end{pmatrix}=\begin{bmatrix} \gamma & \gamma\beta & 0 & 0 \\ \gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}\begin{pmatrix} E' \\ {p'}_{x'}c \\ {p'}_{y'}c \\ {p'}_{z'}c \end{pmatrix} \tag{12.3.288}$$

where we fall back on the "Lorentz matrix" or "symmetric Lorentz tensor" ${L'}_{\mu}^{\nu}$

The vector:

$$\boxed{p^{\nu}=\left(\frac{E}{c},p_x,p_y,p_z\right)=\left(\frac{E}{c},\vec{p}\right)=m\gamma(\vec{v})(c,\vec{v})} \tag{12.3.289}$$

is meanwhile, named the "four-vector energy-momentum" or just "four-momentum" Its utility is that its value is also conserved and this is especially useful for the study of nuclear reactions. If we add these vectors on all particles (without forgetting the photons as well!!!) before and after the reaction, we should found the same quantities for the 4 components!

Remarks

R1. The inverse transformation being done obviously with the inverse matrix that we have already outlined earlier above.

R2. We use in Relativistic Optics the four-vector $(\omega/c,\vec{k})$, where ω is for recall the pulsation of the wave and $\vec{k}$ the wave vector (see sections Wave Mechanics and Wave Optics page 3480). This four-vector is the equivalent for an electromagnetic wave of the four-vector $(E/c,\vec{p})$ for a particle multiplied by the reduced Planck's constant $\hbar=h/2\pi$. Indeed, the wave-particle duality (see section Wave Quantum Physics page 4688) attributes to a wave an energy according to the Planck-Einstein relation (see section Corpuscular Quantum Physics page 4627):

$$E=h\nu=\frac{h}{2\pi}\omega=\hbar\omega \tag{12.3.290}$$

and a linear momentum which norm is:

$$p=\frac{E}{c}=\hbar\frac{\omega}{c}=\hbar k \qquad (12.3.291)$$

Now let us come back to the following relation that is of main importance in some areas of quantum physics:

$$p_x^2+p_y^2+p_z^2-\frac{E^2}{c^2}=-m_0^2c^4 \qquad (12.3.292)$$

Therefore:

$$\frac{E^2}{c^2}-(p_x^2+p_y^2+p_z^2)=m_0^2c^4 \qquad (12.3.293)$$

Which can be written in vector form (very common form):

$$\frac{E^2}{c^2}-\|\vec{p}\|=m_0^2c^4 \qquad (12.3.294)$$

This latter relation will be very useful in the section of Relativistic Quantum Physics to calculate the energy of virtual photons exchange.

For photons, since the mass is zero, we have:

$$\frac{E^2}{c^2}-\|\vec{p}\|=0 \qquad (12.3.295)$$

Finally let us also notice that the four-momentum is also related to a another quantity named "four-wave vector" as following:

$$\boxed{p^\nu=\left(\frac{E}{c},\vec{p}\right)=\hbar k^\nu=\hbar\left(\frac{\omega}{c},\vec{k}\right)} \qquad (12.3.296)$$

as we know that (for the first component):

$$E=h\nu=\hbar\omega \qquad (12.3.297)$$

and that for all other components :

$$p=mc=\frac{E}{c}=\frac{h\nu}{c}=\frac{\hbar\omega}{c}=\frac{\hbar kc}{c}=\hbar k \qquad (12.3.298)$$

12.3.2.9.4.1 Einstein relation

Following the principle of relativity, we wish that the relation between the linear momentum and energy of an electromagnetic wave can be written in the same way for two inertial observers in translation relative to the other!

If O writes:

$$E=c\sqrt{m_0^2c^4+p^2} \qquad (12.3.299)$$

then O' must be able to write:

$$E' = c\sqrt{m_0^2c^4 + p'^2} \tag{12.3.300}$$

Let us take the first relation above and put it to the square without forgetting that the photon has a zero rest mass m_0. Therefore:

$$\boxed{E^2 = m_0^2c^4 + p^2c^2} \tag{12.3.301}$$

and as $m_0 = 0$ for the photon:

$$E = pc \tag{12.3.302}$$

Given the known Planck's law (see section Thermodynamics page 3719):

$$E = h\nu \tag{12.3.303}$$

we are led to write the famous "Einstein's relation" that we will find very often in Quantum Physics and in Thermodynamics:

$$\boxed{p = \frac{h\nu}{c}} \tag{12.3.304}$$

So even if the photon has no mass at rest, it has a "relativistic mass". This relativistic mass is in Special Relativity for the photon the equivalent to the electric charge in Electrodynamics.

Remark

Notice that a priori nothing seems to prevent us to write $E = h\nu = mc^2$. Therefore it means that without mass there is no frequency. By extrapolation in a Universe without mass (made only of radiation) there is no clock and therefore no time and hence no distance. In the late 120th century (holocene calendar) this is speculative and need to be verified experimentally if it's possible...

12.3.2.9.4.2 Time of flight

Suppose we get a beam made of massive particles. The rest mass is m_0. The particle travels a distance L in its inertial frame. The particle has an energy E in that frame. Therefore, the so-called "time of flight" from two points at $x = 0$ and $x = L$ will be:

$$T = \frac{L}{v} = \frac{m_0\gamma L}{m_0\gamma v} = \frac{m_0\gamma L}{p} = \frac{m\gamma c^2 L}{pc^2} \tag{12.3.305}$$

where we use the relativistic definition of momentum:

$$p = m_0\gamma v \tag{12.3.306}$$

Now, knowing that the relativistic energy is:

$$E = m_0\gamma c^2 \tag{12.3.307}$$

using the time that a mass-less light beam uses to travel the proper distance L, easily calculated to be:

$$\tau=\frac{L}{c} \tag{12.3.308}$$

we get:

$$T=\frac{EL}{pc^2}=\frac{E\tau}{pc}=\frac{\tau}{pc/E} \tag{12.3.309}$$

But we have just proved that:

$$p^2c^2=E^2-m_0^2c^4 \tag{12.3.310}$$

Therefore:

$$\frac{p^2c^2}{E^2}=1-\frac{m_0c^4}{E^2} \tag{12.3.311}$$

and then:

$$\frac{pc}{E}=\sqrt{1-\left(\frac{m_0c^2}{E}\right)^2} \tag{12.3.312}$$

Thus, the time of flight is finally written as follows:

$$\boxed{T=\frac{\tau}{\sqrt{1-\left(\dfrac{m_0c^2}{E}\right)^2}}} \tag{12.3.313}$$

This equation is very important in practical applications. Specially in Astrophysics and baseline beam experiments, like those involving the neutrinos! Indeed, we usually calculate the difference between the photon (or any other massless) time of arrival and that of massive particles, e.g. the neutrinos. Several neutrino experiments can measure this difference using a well designed experimental set-up. The difference between those times of flight (the neutrino time of flight minus the photon time of flight) is:

$$T-\tau=\frac{\tau}{\sqrt{1-\left(\dfrac{m_0c^2}{E}\right)^2}}-\tau \tag{12.3.314}$$

or equivalently:

$$\Delta t=T-\tau=\tau\left(\frac{\tau}{\sqrt{1-\left(\dfrac{m_0c^2}{E}\right)^2}}-1\right) \tag{12.3.315}$$

or as well:

$$Q=\frac{\Delta t}{\tau}=\left(\frac{\tau}{\sqrt{1-\left(\dfrac{m_0c^2}{E}\right)^2}}-1\right) \tag{12.3.316}$$

This last expression can also be expressed in terms of the speed of light and neutrinos, since:

$$c=\frac{L}{\tau} \qquad \text{and} \qquad v_\nu=\frac{L}{T} \tag{12.3.317}$$

so:

$$\frac{T-\tau}{\tau}=\frac{T/L-\tau/L}{\tau/L}=\frac{1/v_\nu-1/c}{1/c}=\frac{c-v_\nu}{v_\nu} \tag{12.3.318}$$

Therefore:

$$\boxed{Q=\frac{\Delta t}{\tau}=\frac{c-v_\nu}{v_\nu}} \tag{12.3.319}$$

In the special case of very non-heavy left-handed neutrinos, the rest mass is tiny (likely sub-[eV] and next to the [meV] scale), and then we can make a Taylor expansion for Q if:

$$x=\frac{mc}{E}\ll 1 \tag{12.3.320}$$

Then:

$$Q=\frac{\Delta t}{\tau}=\left(\frac{\tau}{\sqrt{1-\left(\dfrac{m_0c^2}{E}\right)^2}}-1\right)\cong\left(1+\frac{1}{2}\left(\frac{m_0c^2}{E}\right)^2\right)-1=\frac{1}{2}\left(\frac{m_0c^2}{E}\right)^2 \tag{12.3.321}$$

Then, we would expect, accordingly to Special Relativity, of course, that:

$$\frac{c-v_\nu}{v_\nu}=\frac{T-\tau}{\tau}\cong\frac{1}{2}\left(\frac{m_0c^2}{E}\right)^2 \tag{12.3.322}$$

We can guess how large it is plugging "typical" values for the neutrino mass and energy. For instance, taking $m_\nu\cong 1$ [meV] and $E\cong 1$ [GeV] the Q value is about 10^{-24}.

Nowadays, we have no clock with this precision, so the neutrino mass measurement using this approach is impossible with current technology. However, it is clear that if we could make clocks with that precision, we would measure the neutrino mass with this "time of flight" procedure. It is a challenge. We can not do that in these times (circa 12012 according to holocene calendar), and thus we don't measure any time delay in baseline experiments. Then, neutrinos move with $v_\nu=c$ and since there is no observed delay (beyond the OPERA result, already corrected), neutrinos are, thus, ultra-relativistic particles, and for them $E=pc$ with great accuracy.

12.3.2.9.4.3 Relativistic force

Following the principle of relativity, we want that the relation between force and linear momentum to be written in the same way by two inertial observers in translation relative to each other!

Therefore if O writes:

$$\vec{F}=\frac{\mathrm{d}\vec{p}}{\mathrm{d}t} \tag{12.3.323}$$

O' must be able to write:

$$\vec{F}\,' = \frac{\mathrm{d}\vec{p}\,'}{\mathrm{d}t'} \tag{12.3.324}$$

The relation between $\vec{F}$ and $\vec{F}\,'$ is quite complicated in the general case. We will limit ourselves here to the particular case where a body is momentarily at rest in O' and therefore where the observer O' will only take into account the force $\vec{F}\,'$ that he applies. He will name this the "proper force" because it has not to worry about other forces (such as centrifugal (fictional) force, for example).

It is necessary to substitute p' and t' by p and t in:

$$F'_x = \frac{\mathrm{d}p'_x}{\mathrm{d}t'} \qquad F'_y = \frac{\mathrm{d}p'_y}{\mathrm{d}t'} \qquad F'_z = \frac{\mathrm{d}p'_z}{\mathrm{d}t'} \tag{12.3.325}$$

Since:

$$\frac{\mathrm{d}t}{\mathrm{d}t'} = \frac{1}{1-\dfrac{v^2}{c^2}} \tag{12.3.326}$$

we will have:

$$\begin{aligned}F'_x &= \frac{\mathrm{d}p'_x}{\mathrm{d}t'} = \frac{\mathrm{d}p'_x}{\mathrm{d}t} = \frac{\mathrm{d}t}{\mathrm{d}t'}\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{p_x - v\dfrac{E}{c^2}}{\sqrt{1-\dfrac{v^2}{c^2}}}\right) = \frac{\mathrm{d}t}{\mathrm{d}t'}\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{\sqrt{1-\dfrac{v^2}{c^2}}}\left(p_x - v\frac{E}{c^2}\right)\right)\\ &= \frac{\mathrm{d}t}{\mathrm{d}t'}\left(\frac{1}{\sqrt{1-\dfrac{v^2}{c^2}}}\right)\frac{\mathrm{d}}{\mathrm{d}t}\left(p_x - v\frac{E}{c^2}\right) = \left(\frac{1}{\sqrt{1-\dfrac{v^2}{c^2}}}\right)\left(\frac{1}{\sqrt{1-\dfrac{v^2}{c^2}}}\right)\frac{\mathrm{d}}{\mathrm{d}t}\left(p_x - v\frac{E}{c^2}\right)\\ &= \frac{1}{\sqrt{1-\dfrac{v^2}{c^2}}}\frac{\mathrm{d}}{\mathrm{d}t}\left(p_x - v\frac{E}{c^2}\right) = \frac{1}{\sqrt{1-\dfrac{v^2}{c^2}}}\left(\frac{\mathrm{d}p_x}{\mathrm{d}t} - \frac{v}{c^2}\frac{\mathrm{d}E}{\mathrm{d}t}\right)\end{aligned} \tag{12.3.327}$$

We have seen also previously that:

$$E = E_c + m_0c^2 \Rightarrow \frac{\mathrm{d}E}{\mathrm{d}t} = \frac{\mathrm{d}E_c}{\mathrm{d}t} + 0 = \frac{F_x\mathrm{d}x}{\mathrm{d}t} = F_xv \tag{12.3.328}$$

Therefore it remains:

$$F'_x = \frac{1}{1-\dfrac{v^2}{c^2}}\left(F_x - \frac{v}{c^2}F_xv\right) = F_x \tag{12.3.329}$$

The component of the force is therefore invariant in the direction of the movement.

For the directions y and z perpendicular to the movement:

$$F'_y = \frac{\mathrm{d}p'_y}{\mathrm{d}t'} = \frac{\mathrm{d}t}{\mathrm{d}t'}\frac{\mathrm{d}p_y}{\mathrm{d}t} = \frac{F_y}{\sqrt{1-\dfrac{v^2}{c^2}}} \qquad \text{and} \qquad F'_z = \frac{\mathrm{d}p'_z}{\mathrm{d}t'} = \frac{\mathrm{d}t}{\mathrm{d}t'}\frac{\mathrm{d}p_z}{\mathrm{d}t} = \frac{F_z}{\sqrt{1-\dfrac{v^2}{c^2}}} \tag{12.3.330}$$

So for summary:

$$\boxed{\begin{gathered}F'_x = F_x \Leftrightarrow F_x = F'_x \\ F'_y = \frac{F_y}{\sqrt{1-\dfrac{v^2}{c^2}}} \Leftrightarrow F_y = F'_y\sqrt{1-\frac{v^2}{c^2}} \\ F'_z = \frac{F_z}{\sqrt{1-\dfrac{v^2}{c^2}}} \Leftrightarrow F_z = F'_z\sqrt{1-\frac{v^2}{c^2}}\end{gathered}} \tag{12.3.331}$$

However, to change from one reference frame to another, it is better to use again the "four-vector force" defined as the derivative of the four momentum vector with respect to the proper time:

$$\boxed{F^\nu = \frac{\mathrm{p}^\nu}{\mathrm{d}t'} = \frac{\mathrm{d}t}{\mathrm{d}t'}\frac{\mathrm{d}p^\nu}{\mathrm{d}t} = \gamma\frac{\mathrm{d}p^\nu}{\mathrm{d}t} = \gamma\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{E}{c},\vec{p}\right) = \gamma\left(\frac{\vec{F}\circ\vec{v}}{c},\vec{F}\right)} \tag{12.3.332}$$

Indeed, let us recall that:

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \vec{F}\circ\vec{v} \qquad \vec{F} = \frac{\mathrm{d}\vec{p}}{\mathrm{d}t} \tag{12.3.333}$$

12.3.2.10 Relativistic electrodynamics

With a mass spectrometer we establish that the ratio m/q of the mass m of a particle by its electrical charge q varies in the same way as the mass m when the velocity v of the particle varies:

$$\boxed{\frac{m}{q}=\frac{m_0/q_0}{\sqrt{1-\dfrac{v^2}{c^2}}}=\frac{1}{q_0}\frac{m_0}{\sqrt{1-\dfrac{v^2}{c^2}}}=\frac{m}{q_0}} \tag{12.3.334}$$

Thus, it comes that:

$$q=q_0 \tag{12.3.335}$$

The charge of a particle is therefore independent of its velocity, as we have proved in the section of Electromagnetism (see section Electrodynamics page 4114) when determining the charge conservation equation.

Let us consider now two charges q and Q immobile a reference frame O' in translation at speed v with respect to another one centered on O:

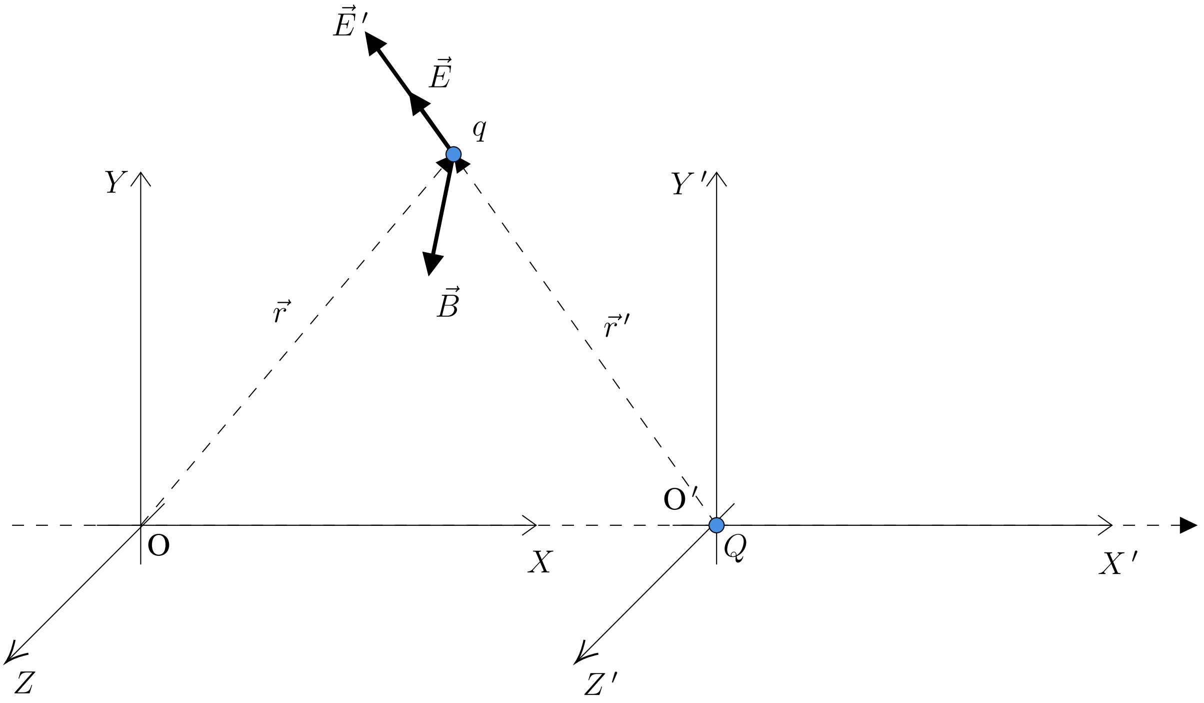


Figure 12.210: **Configuration for the study of transformations of electric and magnetic fields**

We will restrict ourselves to the case where the velocity $\vec{v}$ is parallel to the Ox-axis:

$$\vec{v}=(v_x,0,0) \tag{12.3.336}$$

and we write the vector at the horizontal to spare time and paper...

The electric charge Q is place on O' and is therefore fixed for O'. The observer O' makes the conclusions that an electrostatic force:

$$\vec{F}'=q\vec{E}' \tag{12.3.337}$$

act on the reference particle q placed on $\vec{r}'$:

$$F'_x = qE'_x \qquad F'_y = qE'_y \qquad F'_z = qE'_z \tag{12.3.338}$$

The observer O also sees an electrostatic field $\vec{E}$ in $\vec{r}$, but he also sees that Q is in movement along the Ox-axis. He thus deduces the existence of a magnetic field $\vec{B}$ on $\vec{r}$ oriented in the plane YZ plane:

$$\vec{B} = (0, B_y, B_z) \tag{12.3.339}$$

It therefore measures the supposedly known Lorentz's force (see section Magnetostatics page 4052):

$$\vec{F} = q(\vec{E} + \vec{v} \times \vec{B}) \tag{12.3.340}$$

But:

$$\vec{v} \times \vec{B} = \begin{pmatrix} v_x \\ 0 \\ 0 \end{pmatrix} \times \begin{pmatrix} 0 \\ B_y \\ B_z \end{pmatrix} = \begin{pmatrix} 0-0 \\ 0 - v_x B_z \\ v_x B_y \end{pmatrix} = \begin{pmatrix} 0 \\ -vB_z \\ vB_y \end{pmatrix} \tag{12.3.341}$$

Therefore:

$$F_x = qE_x \qquad F_y = q(E_y - vB_z) \qquad F_z = q(E_z + vB_y) \tag{12.3.342}$$

We have now seen:

$$\begin{aligned} F'_x &= F_x = qE_x \\ F'_y &= \frac{F_y}{\sqrt{1-\dfrac{v^2}{c^2}}} = \frac{q(E_y - vB_z)}{\sqrt{1-\dfrac{v^2}{c^2}}} \\ F'_z &= \frac{F_z}{\sqrt{1-\dfrac{v^2}{c^2}}} = \frac{q(E_z + vB_y)}{\sqrt{1-\dfrac{v^2}{c^2}}} \end{aligned} \tag{12.3.343}$$

The comparison of the expressions above gives the relativistic transformations of the electric field:

$$\boxed{\begin{gathered} E'_x = E_x \Leftrightarrow E_x = E'_x \\ E'_y = \frac{E_y - vB_z}{\sqrt{1-\dfrac{v^2}{c^2}}} \Leftrightarrow E_y = \frac{E'_y + vB'_z}{\sqrt{1-\dfrac{v^2}{c^2}}} \\ E'_z = \frac{E_z + vB_y}{\sqrt{1-\dfrac{v^2}{c^2}}} \Leftrightarrow E_z = \frac{E'_z - vB'_y}{\sqrt{1-\dfrac{v^2}{c^2}}} \end{gathered}} \tag{12.3.344}$$

As for the Lorentz transformation of the spatial and temporal components, we have obtained the inverse transformations by exchanging the fields and considering that O' sees O going back away (we therefore replace v by $-v$).

The above relations, sometimes named "Joules-Bernoulli equations of the electric field", make it clear that if, for example, the electric field in one of the reference systems is zero but the magnetic field is not, then an electric field exists from the point of view of the other reference frame !!! It is therefore an absolute victory of relativity in comparison to classical mechanics!

To get the relativistic transformations of the magnetic field, we proceed as follows:

$$E'_y=\frac{E_y-vB_z}{\sqrt{1-\dfrac{v^2}{c^2}}}=\frac{\dfrac{E'_y+vB'_z}{\sqrt{1-\dfrac{v^2}{c^2}}}-vB_z}{\sqrt{1-\dfrac{v^2}{c^2}}} \tag{12.3.345}$$

After some small manipulations of very elementary algebra, we get:

$$B_z=\frac{B'_z+\dfrac{v}{c^2}E'_y}{\sqrt{1-\dfrac{v^2}{c^2}}} \tag{12.3.346}$$

And identically:

$$E_y=\frac{E'_y-vB'_z}{\sqrt{1-\dfrac{v^2}{c^2}}}=\frac{\dfrac{E_y-vB_z}{\sqrt{1-\dfrac{v^2}{c^2}}}+vB_z}{\sqrt{1-\dfrac{v^2}{c^2}}} \tag{12.3.347}$$

After a few simple manipulations of very elementary algebra, we get:

$$B'_z=\frac{B_z-\dfrac{v}{c^2}E_y}{\sqrt{1-\dfrac{v^2}{c^2}}} \tag{12.3.348}$$

And so on. Finally, we get:

$$\boxed{\begin{gathered}B'_x=B_x\Leftrightarrow B_x=B'_x\\ B'_y=\frac{B_y+\dfrac{v}{c^2}E_z}{\sqrt{1-\dfrac{v^2}{c^2}}}\Leftrightarrow B_y=\frac{B'_y-\dfrac{v}{c^2}E'_z}{\sqrt{1-\dfrac{v^2}{c^2}}}\\ B'_z=\frac{B_z-\dfrac{v}{c^2}E_y}{\sqrt{1-\dfrac{v^2}{c^2}}}\Leftrightarrow B_z=\frac{B'_z+\dfrac{v}{c^2}E'_y}{\sqrt{1-\dfrac{v^2}{c^2}}}\end{gathered}} \tag{12.3.349}$$

The above relations, sometimes named "Joules-Bernoulli equations of the magnetic field", make it clear that if, for example, the magnetic field in one of the reference frame is zero but the electric field is not, then a magnetic field exists from the point of view of the other reference fame !!! It is therefore once again an absolute victory of relativity in comparison to classical mechanics!

Let us now study the behaviour of the electromagnetic field of a moving charge:

Let us consider two parallel referentials O and O', in translation at constant velocity v along the axis XX':

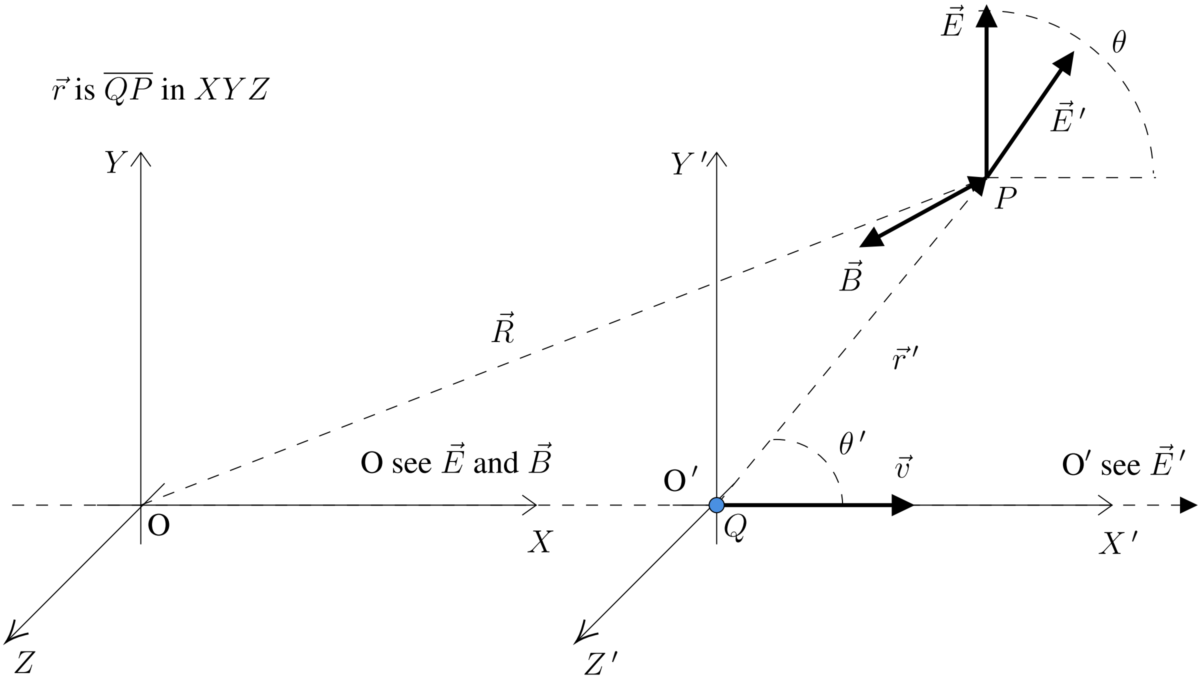


Figure 12.211: **Configuration for the study of electrodynamic transformations**

where a fixed electric charge Q is placed at O'.

It is clear that the observer O measures $\vec{B}=\vec{0}$ everywhere and that at the point P of the plane $X'Y'$, on $\vec{r}\,'=(x',y',0)$ he measures the electrostatic field (see section Electrostatics page 3968):

$$\vec{E}\,'=\begin{pmatrix}E'_x\\E'_y\\0\end{pmatrix}=\frac{Q}{4\pi\varepsilon_0 r'^3}\vec{r}\,'=\frac{Q}{4\pi\varepsilon_0}r'^3\begin{pmatrix}x'\\y'\\0\end{pmatrix}\qquad(12.3.350)$$

If the observer O is informed of the values of $\vec{E}\,'$ and of $\vec{B}\,'=\vec{0}$, he can introduce them into the relativistic transformation giving the electric field $\vec{E}$ that he observes:

$$E_x=E'_x\qquad(12.3.351)$$

$$E_y=\frac{E'_y+vB'_z}{\sqrt{1-\dfrac{v^2}{c^2}}}=\frac{E'_y}{\sqrt{1-\dfrac{v^2}{c^2}}}\qquad(12.3.352)$$

$$E_z=\frac{E'_z-vB'_y}{\sqrt{1-\dfrac{v^2}{c^2}}}=\frac{E'_z}{\sqrt{1-\dfrac{v^2}{c^2}}}\qquad(12.3.353)$$

To write an expression of the field $\vec{E}$ at the point P, the observer O must determine, at a time t of its local time, the components of the vector $\vec{r}=(x,y,0)$ which separates the point P from the electric charge Q (by summing the position vectors of the latter two material points).

The coordinates of the point P and of the charge Q that he sees in the plane XYZ are given by the usual Lorentz transformations:

$$\vec{R}_P=\left(\frac{x'+vt}{\sqrt{1-\frac{v^2}{c^2}},y',0}\right) \qquad \text{and} \qquad \vec{R}_Q=\left(\frac{vt}{\sqrt{1-\frac{v^2}{c^2}},0,0}\right) \tag{12.3.354}$$

He thus easily deduces, by summation, the distances x, y.

Another simpler possible method is that since the x component is a length, it therefore undergoes Lorentz transformations and:

$$\vec{r}=(x,y,z)=\left(x'\sqrt{1-\frac{v^2}{c^2}},y',0\right) \tag{12.3.355}$$

Since for recall:

$$x'=\frac{x}{\sqrt{1-\frac{v^2}{c^2}}} \qquad \text{and} \qquad y'=y \tag{12.3.356}$$

The relativistic transformation of the electric field then gives:

$$E_x=E'_x=\frac{Q}{4\pi\varepsilon_0 r'^3}x'=\frac{Q}{4\pi\varepsilon_0 r'^3}\frac{x}{\sqrt{1-\frac{v^2}{c^2}}} \tag{12.3.357}$$

$$E_y=\frac{E'_y}{\sqrt{1-\frac{v^2}{c^2}}}=\frac{1}{\sqrt{1-\frac{v^2}{c^2}}}\frac{Q}{4\pi\varepsilon_0 r'^3}y'=\frac{Q}{4\pi\varepsilon_0 r'^3}\frac{y}{\sqrt{1-\frac{v^2}{c^2}}} \tag{12.3.358}$$

and:

$$E_z=E'_z=0 \tag{12.3.359}$$

Written in vector form:

$$\vec{E}=\frac{Q\vec{r}}{4\pi\varepsilon_0\sqrt{1-\frac{v^2}{c^2}}}\frac{1}{r'^3} \tag{12.3.360}$$

We must also determine how to express r' as a function of r:

$$r'2=x'^2+y'^2=\frac{x^2}{1-\frac{v^2}{c^2}}+y^2=\frac{x^2+\left(1-\frac{v^2}{c^2}\right)y^2}{1-\frac{v^2}{c^2}}=\frac{r^2-\frac{v^2}{c^2}y^2}{1-\frac{v^2}{c^2}} \tag{12.3.361}$$

as (Pythagorean theorem):

$$x^2+y^2=r^2 \tag{12.3.362}$$

The writing is simplified if we use the angle formed by the electric field vector and the x-axis. We then denote then θ' in O' and the θ in O the angles given by:

$$\tan(\theta')=\frac{y'}{x'} \qquad \text{and} \qquad \tan(\theta)=\frac{y}{x} \tag{12.3.363}$$

with $\theta\geq\theta'$ due to the expansion of the lengths along the x-axis.

We eliminate y with:

$$y^2=r^2\sin^2(\theta)\Rightarrow r'=\frac{r\left[1-\dfrac{v^2}{c^2}\sin^2(\theta)\right]^{1/2}}{\sqrt{1-\dfrac{v^2}{c^2}}} \tag{12.3.364}$$

Thus, the electric field $\vec{E}$ that sees O is given by:

$$\boxed{\vec{E}=\frac{Q\vec{r}}{4\pi\varepsilon_0\sqrt{1-\dfrac{v^2}{c^2}}}\frac{1}{r'^3}=\frac{1-\dfrac{v^2}{c^2}}{\left[1-\dfrac{v^2}{c^2}\sin^2(\theta)\right]^{3/2}}\frac{Q\vec{r}}{4\pi\varepsilon_0 r^3}} \tag{12.3.365}$$

The factor containing $\sin(\theta)$ shows that the electric field $\vec{E}$ of a moving charge no longer has a spherical symmetry!!! It depends on the direction of the vector $\vec{r}$.

At equal distances, the electric field is more intense in the vertical direction to that of the displacement $(\theta=\pi/2)$ than in the direction of the displacement of the electric charge $(\theta=0)$.

If $v=0$, we fall back on the classic known expression:

$$\vec{E}=\frac{Q\vec{r}}{4\pi\varepsilon_0 r^3} \tag{12.3.366}$$

Remark

Let us recall that we have carried out (and continue in this sense) here a study of an electric charge in uniform rectilinear motion, that is to say at constant speed!

To find now the expression of the magnetic field $\vec{B}$, we introduce:

$$\vec{B}=(B'_x,B'_y,B'_z)=(0,0,0) \tag{12.3.367}$$

and:

$$\begin{aligned}E'_x&=E_x\\E'_y&=E_y\sqrt{1-\frac{v^2}{c^2}}\\E'_z&=E_z\sqrt{1-\frac{v^2}{c^2}}\end{aligned} \tag{12.3.368}$$

in:

$$\begin{aligned}B_x&=B'_x\\B_y&=\frac{B'_y-v\dfrac{E'_z}{c^2}}{\sqrt{1-\dfrac{v^2}{c^2}}}\\B_z&=\frac{B'_z-v\dfrac{E'_y}{c^2}}{\sqrt{1-\dfrac{v^2}{c^2}}}\end{aligned} \tag{12.3.369}$$

We therefore get:

$$B_x=0 \qquad B_y=-\frac{vE_z}{c^2} \qquad B_z=\frac{vE_y}{c^2} \tag{12.3.370}$$

Which are the components of:

$$\vec{B}=\vec{v}\times\vec{E}c^2 \tag{12.3.371}$$

To know $\vec{B}$ as a function of $\vec{r}$, we substitute the expression obtained for $\vec{E}$:

$$\boxed{\vec{B}=\frac{Q}{c^2\varepsilon 4\pi r^3}\frac{1-\dfrac{v^2}{c^2}}{\left[1-\dfrac{v^2}{c^2}\sin^2(\theta)\right]^{3/2}}\vec{v}\times\vec{r}} \tag{12.3.372}$$

In the case where the velocity is small, the relativistic term tends to 1 and the field $\vec{B}$ of an electric charge Q moving at the velocity v becomes:

$$\vec{B}=\frac{Q}{c^2\varepsilon 4\pi r^3}\vec{v}\times\vec{r}=\frac{\mu_0 Q}{4\pi r^3}\vec{v}\times\vec{r} \tag{12.3.373}$$

because as we have seen in the section of Electrodynamics (see page 4140):

$$\mu_0=\frac{1}{c^2\varepsilon_0} \tag{12.3.374}$$

Remarks

R1. At each location, the lines of the field $\vec{B}$ are contained in a plane perpendicular to the direction of motion of the electric charge Q (vector product oblige...!).

R2. If the moving electric charge is seen as a $\mathrm{d}Q$ attached to the point O', we can interpret its displacement at velocity v as a current I at a point of the referential O where O' is located. Therefore:

$$\mathrm{d}Q\vec{v}=\mathrm{d}Q\frac{\mathrm{d}\vec{x}}{\mathrm{d}t}=\frac{\mathrm{d}Q}{\mathrm{d}t}\mathrm{d}\vec{x}=I\mathrm{d}\vec{x} \tag{12.3.375}$$

Therefore:

$$\mathrm{d}\vec{B}(\vec{r})=\frac{\mu_0\mathrm{d}Q}{4\pi r^3}\vec{v}\times\vec{v}\times\vec{r}=\frac{\mu_0 I}{4\pi r^3}\mathrm{d}\vec{x}\times\vec{r} \tag{12.3.376}$$

We then fall back here the "Biot and Savart law" as proved in the section of Electromagnetism. So this a success of the Special Relativity theory again!!!

It is interesting to remember that an electric charged particle in motion will be seen in the frame of reference of the particle as emitting no electromagnetic field (there will be just an electrostatic field). This is not the case for a repository at rest. There is thus here a sort of flagrant counter-intuitive contradiction.

But this poses another problem, in a fast-moving frame of reference, a charged particle normally emits an acceleration radiation (see section Electrodynamics page 4163), this radiation in quantum mechanics must necessarily be accompanied by the emission of a quanta, which exists either or does not exist (a medium term does not exist). The very existence of photons would therefore be purely relative. And yet it is! Some particles have only a relative existence!!!! The complicated answer is therefore to know what the photons have become.

But here we reach the limit of what we master perfectly in the physics of the end of the 120th century (holocene calendar), because we speak of accelerated references frames (which implies to be in General Relativity and not the special one) and quantum field theory. The rigorous framework for dealing with this (which would encompass Quantum Gravitation) does not yet exist as far as we know. But a first step has been taken with the development of the Quantum Field Theory in curved space.

12.3.2.10.1 Tensor field transformation

We have seen and proved in the section of Electrodynamics that the whole electromagnetic field was summarized by the tensor of the same name (see page 4121). It would then be good to look at how this tensor transforms itself and if it does so correctly in relation to the results obtained above.

Let us consider the transformation (where the tensor of the electromagnetic field is in natural units !!!):

$$\boxed{F'^{\mu\nu}=L^{\mu}_{\nu}L^{\nu}_{\beta}F^{\alpha\beta}} \tag{12.3.377}$$

with the tensor of the electromagnetic field in contravariant components in the Minkowski metric $+---$:

$$F^{\mu\nu}=\begin{bmatrix}0 & -E_x & -E_y & -E_z\\ E_x & 0 & -B_z & B_y\\ E_y & B_z & 0 & -B_x\\ E_z & -B_y & B_x & 0\end{bmatrix} \tag{12.3.378}$$

And also by construction:

$$F'^{\mu\nu}=\begin{bmatrix}0 & -E'_x & -E'_y & -E'_z\\ E'_x & 0 & -B'_z & B'_y\\ E'_y & B'_z & 0 & -B'_x\\ E'_z & -B'_y & B'_x & 0\end{bmatrix} \tag{12.3.379}$$

Let us take, for example, the velocity parallel to the x-axis, then we have proved above that:

$$L=\begin{bmatrix}\gamma & -\beta\gamma & 0 & 0\\ -\beta\gamma & \gamma & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{bmatrix} \tag{12.3.380}$$

Therefore:

$$L^0_0=\gamma \qquad L^0_1=-\beta\gamma \qquad L^1_0=-\beta\gamma \qquad L^2_1=\gamma \tag{12.3.381}$$

where as we can see, it is often customary in the field of Special Relativity and Electrodynamics to number the components of matrices / tensors starting from 0 (instead of 1 for most of the other chapters of this textbook).

We calculate the transformation (remember that the tensor of the electromagnetic field is antisymmetric!):

$$\begin{aligned}F'^{01}&=L^0_{\alpha}L^1_{\beta}F^{\alpha\beta}=L^0_0L^1_0F^{00}+L^0_1L^1_0F^{10}+L^0_0L^1_1F^{01}+L^0_1L^1_1F^{11}=L^0_1L^1_0F^{10}+L^0_0L^1_1F^{01}\\&=\beta^2\gamma^2F^{10}+\gamma^2F^{01}=F^{01}\gamma^2(1-\beta^2)=F^{01}\end{aligned} \tag{12.3.382}$$

We thus deduce, for the electric field (which corresponds perfectly to what we obtained above):

$$E'_x=E_x \tag{12.3.383}$$

We make a second calculation for the perpendicular component:

$$F'^{02}=L_\alpha^0 L_\beta^2 F^{\alpha\beta}F^{\alpha\beta}=L_\alpha^0 F^{\alpha 2}=L_0^0 F^{02}+L_1^0 F^{12} \quad (12.3.384)$$

hence:

$$-E'_y=\gamma(-E_y+\beta B_z)\Rightarrow E'_y=\gamma(E_y-vB_z) \quad (12.3.385)$$

which again corresponds perfectly to what we had obtained earlier above (in natural units, do not forget that we then have $\beta=v$)!

The same applies to the magnetic field:

$$F'^{23}=L_\alpha^2 L_\beta^3 F^{\alpha\beta}=F^{23}\Rightarrow B'_x=B_x \quad (12.3.386)$$

and:

$$F'^{12}=L_\alpha^1 L_\beta^2 F^{\alpha\beta}=L_\alpha^1 F^{\alpha 2}=L_1^1 F^{12}+L_0^1 F^{02}=\gamma F^{12}-\gamma\beta F^{02} \quad (12.3.387)$$

Which gives (in natural units, again do not forget that we then have $\beta=v$)! :

$$-B'_z=\gamma(-B_z+\beta E_y)\Rightarrow B'_z=\gamma(B_z-vEy) \quad (12.3.388)$$

etc.

12.3.3 Minkowski space-time

We have proved earlier above that:

$$t=\frac{t'}{\sqrt{1-\dfrac{v^2}{c^2}}} \quad (12.3.389)$$

Let us write this in the form:

$$\frac{t'^2}{t^2}=1-\frac{v^2}{c^2} \quad (12.3.390)$$

Let us multiply the two members by $(ct)^2$:

$$\frac{t'^2c^2t^2}{t^2}=c^2t^2-\frac{v^2c^2t^2}{c^2} \quad (12.3.391)$$

which gives us:

$$t'^2c^2=c^2t^2-v^2t^2 \quad (12.3.392)$$

If $v=c$ the equation vanishes:

$$t'^2c^2=0 \quad (12.3.393)$$

This result translates the fact that the dimensions of space and time are as stopped in the relativistic referential, because the relative speed of the object is equal to that of the light!

Let us now imagine that a light beam is emitted at the instant $t=0$ and propagates from the origin of a referential. We know that in space-time (application of the Pythagoras theorem in three-dimensional Euclidean space for recall...) the distance travelled by the photon is:

$$c=\frac{\sqrt{x^2+y^2+z^2}}{t} \tag{12.3.394}$$

By changing t of member and bringing the whole to the square to remove the root, we get:

$$ct=\sqrt{x^2+y^2+z^2}\Leftrightarrow ct^2=x^2+y^2+z^2 \tag{12.3.395}$$

Therefore:

$$ct^2-(x^2+y^2+z^2)=0 \tag{12.3.396}$$

Remark

We can assimilate this relation to the representation of a spherical wavefront of a light wave propagating at the speed of light (see the equation of a sphere originally centered in the section of Analytical Geometry).

Let us now consider two coordinate events (x_1,y_1,z_1,t_1) and (x_2,y_2,z_2,t_2) and we denote by $\mathrm{d}s$ the "space-time abscissa" (or "proper-distance"). We can then write the spatio-temporal interval as such:

$$(\mathrm{d}s^2)=c^2(t_2-t_1)^2-\left((x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2\right) \tag{12.3.397}$$

By passing to the limit, we get the famous quadratic form of the "interval invariant":

$$\boxed{\mathrm{d}s^2=c^2\mathrm{d}t^2-(\mathrm{d}x^2+\mathrm{d}y^2+\mathrm{d}z^2)=c^2\mathrm{d}t^2-\mathrm{d}L^2} \tag{12.3.398}$$

which has the same shape and value regardless of the reference system considered as we have already proved it earlier above. The infinitesimal interval of space-time $\mathrm{d}s^2$ between two infinitely neighbouring events is therefore a relativistic invariant that we often name the "space-time curvilinear abscissa" or more simply the "worldline" (an observer "riding" the worldline does not see - or "feel" - himself moving through space if he no have any observable reference object visible, only through time or more exactly it's "proper time" as we already know it).

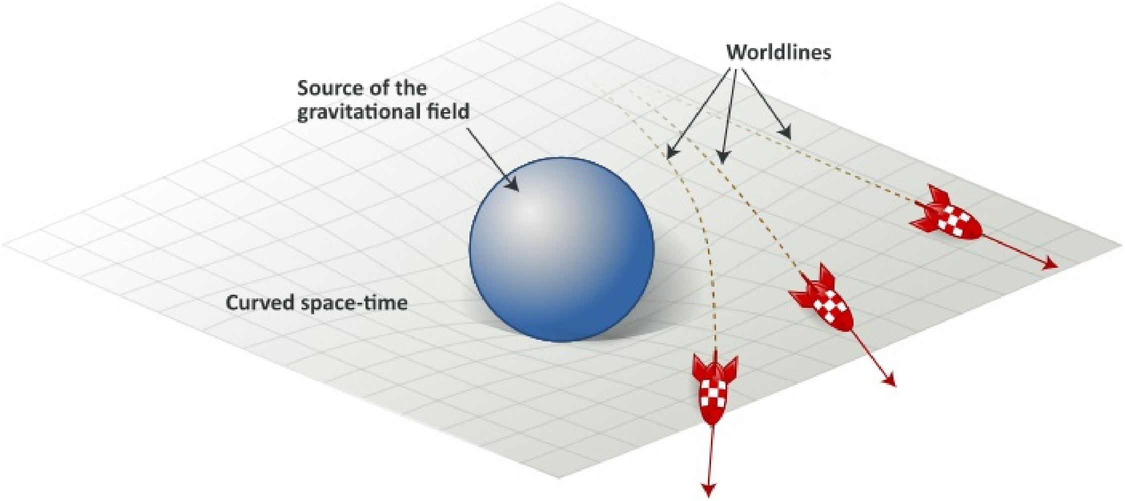

Worldlines
Source of the gravitational field
Curved space-time

Figure 12.212: **Worldline**

It is the interval of space-time or, as Albert Einstein simply said, the "square of distance" .... The fact that this magnitude may be positive, negative (!) or zero is linked to the absolute character of the speed of light (we will come back to this later).

Remark

Equivalently we have $\mathrm{d}s^2=-c^2\mathrm{d}t^2+(\mathrm{d}x^2+\mathrm{d}y^2+\mathrm{d}z^2)$. That we can rewrite as $\mathrm{d}s^2=c^2\mathrm{d}(\mathrm{i}t)^2+(\mathrm{d}x^2+\mathrm{d}y^2+\mathrm{d}z^2)$. Therefore imaginary times makes space-time Euclidean.

We can also now turn our attention to the relativistic character of this metric. If it is invariant, it must also be invariant by Lorentz transformations. We then say that "the metric is invariant by Lorentz transformation". Such a transformation can be found on the basis of that used for the tensor of the electromagnetic field (see above). The reader will readily verify from the detailed example of the electromagnetic field that for the metric tensor we have the relation (as always we can detail on request if necessary):

$$\eta'_{\mu\nu}=L_\alpha^\mu L_\beta^\nu \eta_{\mu\nu} \tag{12.3.399}$$

The curvilinear abscissa can also be expressed by the norm of the quadrivector displacement which we defined above as (ct,x,y,z). Indeed, the norm (see section Tensor Calculus page 2051) is written by taking down the indices using the "Minkowski metric" $\eta_{\mu\nu}$ or "pseudo-Riemannian metric":

$$\mathrm{d}s^2=\mathrm{d}x^\nu(\eta_{\mu\nu}\mathrm{d}x^\nu)=\eta_{\mu\nu}\mathrm{d}x^\nu\mathrm{d}x^\nu=\mathrm{d}x^\nu\mathrm{d}x_\nu=c^2\mathrm{d}t^2-\mathrm{d}x^2-\mathrm{d}y^2-\mathrm{d}z^2 \tag{12.3.400}$$

with the definition of the "Minkowski's matrix" (we will return to this in detail at the beginning of our study of General Relativity):

$$\eta_{\mu\nu}=\begin{bmatrix}1&0&0&0\\0&-1&0&0\\0&0&-1&0\\0&0&0&-1\end{bmatrix} \tag{12.3.401}$$

where as usual in this textbook we make the abuse of notation (already mentioned in the section of Tensorial Calculation) not to put $\eta_{\mu\nu}$ in brackets (since a tensor and its matrix form are normally two distinct things in rigorously speaking).

If we put the following two relations in correspondence:

$$c^2t^2-(x^2+y^2+z^2)=0 \qquad \text{and} \qquad \mathrm{d}s^2=c^2\mathrm{d}t^2-(\mathrm{d}x^2+\mathrm{d}y^2+\mathrm{d}z^2) \tag{12.3.402}$$

we then have $\mathrm{d}s^2=0$ when the two events are connected to the speed of light.

Moreover, if we put:

$$\mathrm{d}x^2+\mathrm{d}y^2+\mathrm{d}z=\mathrm{d}r^2 \tag{12.3.403}$$

we can then write:

$$\boxed{\mathrm{d}s^2+\mathrm{d}r^2=c^2\mathrm{d}t^2} \tag{12.3.404}$$

This is nothing more than the equation of a cone (see section Analytical Geometry page 2823) of axis of ordinate $c^2\mathrm{d}t^2$... the famous "light cone of Universe" (to which we devote a study further below). Every event is therefore by extension in this cone and the evolution of any system can thus be described (by its spatial and temporal position), by what we name its "line of Universe" or "World line". The Universe line of a particle is therefore the sequence of events it unfolds during its lifetime.

12.3.3.1 Four-vectors

We have just defined what Minkowski's metric was, we can now correctly define the concept of quadrivector that we have already addressed without always knowing what we were doing.

Definition (#12.3.541): In a four-dimensional space of Minkowski type, the four quantities:

$$K^{\nu}=(K_t,K_x,K_y,K_z) \tag{12.3.405}$$

(regardless of the order of terms for this definition or whether the indices are numbers or letters corresponding to the four spatio-temporal components) form a covariant "four-vector" if they transform following the Lorentz transformation:

$$\begin{aligned} K'_x&=\gamma(K_x-\beta K_t) & K_x&=\gamma(K'_{x'}+\beta K'_{t'})\\ K'_y&=K_y & K_y&=K'_y\\ K'_z&=K_z & K_z&=K'_z\\ K'_t&=\gamma(K_t-\beta K_x) & K_t&=\gamma(K'_{t'}+\beta K'_{x'}) \end{aligned} \tag{12.3.406}$$

The "pseudo-norm" of a quadrivector in a Minkowski space of metric $\eta_{\mu\nu}$ is then:

$$K^{\mu}K_{\mu}=K^{\mu}(\eta_{\mu\nu}K^{\mu})=K_t^2-K_x^2-K_y^2-K_z^2 \tag{12.3.407}$$

where we see that the contravariant four-vector multiplied by the metric returns the contravariant four-vector (see section Tensor Calculus page 2059).

The following quantity being invariant by change of Galilean referential as we proved it almost at the beginning of this section:

$$K_{\mu}K^{\mu}=K'_{\mu}K'^{\mu} \tag{12.3.408}$$

This property of invariance by change of Galilean referential of the four-vector is their main property. Thus, two observers in relative motion, which are uniform in relation to each other, must compare the results of the same measure using the norm of the four-vectors. Similarly, the laws they seek to determine to be as general as possible must use these invariant quantities!

We can also write the norm of a four-vector in the form:

$$\begin{aligned} K_{\nu}K^{\nu}&=K_0K^0+K_1K^1+K_2K^2+K_3K^3\\ &=(K_0)^2-(K_1)^2-(K_2)^2-(K_3)^2\\ &=(K^0)^2-(K^1)^2-(K^2)^2-(K^3)^2\\ &=(K^0)^2-\|\vec{K}\|^2 \end{aligned} \tag{12.3.409}$$

and the four-vector themselves:

$$K^{\nu}=(K^0,\vec{K}) \qquad K_{\nu}=(K_0,-\vec{K}) \tag{12.3.410}$$

So for summary let us give the list of the four-vectors we have determined so far in this section but by standardizing the notations (and only those four-vectors that will be useful for other sections of this textbook!):

- The space-times four-vector ("four-position"):

$$x^{\nu}=(ct,x,y,z)=(ct,\vec{x}) \tag{12.3.411}$$

- The velocity four-vector ("four-velocity"):

$$u^{\nu}=\gamma(\vec{v})(c,v_x,v_y,v_z)=\gamma(\vec{v})(c,\vec{v}) \tag{12.3.412}$$

- The current four-vector ("four-current"):

$$j^{\nu}=\rho_0\gamma(\vec{v})(c,\vec{v}) \tag{12.3.413}$$

- The acceleration four-vector ("four-acceleration"):

$$a^{\nu}=\gamma^4\left(\vec{\beta}\circ\vec{a},\vec{a}+\vec{\beta}\times(\vec{\beta}\times\vec{a})\right) \tag{12.3.414}$$

- The energy-momentum four-vector ("four-momentum"):

$$p^{\nu}=\left(\frac{E}{c},p_x,p_y,p_z\right)=\left(\frac{E}{c},\vec{p}\right)=m\gamma(\vec{v})(c,\vec{v}) \tag{12.3.415}$$

- The gradient four-vector ("four-gradient") introduced in the section of Tensor Calculus:

$$\vec{\partial}=\left(\frac{\partial}{\partial x_0},-\frac{\partial}{\partial x_1},-\frac{\partial}{\partial x_2},\frac{\partial}{\partial x_3}\right)=\left(\frac{1}{c}\frac{\partial}{\partial t},-\nabla\right) \tag{12.3.416}$$

Obviously we have considered here four-vectors in the context of Special Relativity. Although the concept of four-vectors also extends to General Relativity, some of the results stated above require modification in General Relativity.

12.3.3.2 Universe light cone

The topology of the light cone has its origin in the relations of anteriority and posteriority of relativistic events, which makes it possible to distinguish between an event in the past of another or in the future of it.

The principal objective of the light cones in the popularization works of theoretical physics is to map out the history of light pulses emitted at a point in the space where certain conditions may prevail. The points are represented in space by a series of snapshots at various times t_1, t_2, t_3, etc. (see figure below), the spherical wave front of the light magnifying in space. In space-time, the same event (at the bottom on the figure) is represented by a "light cone", whose apex is the point of emission.

On a sheet of paper, we have to remove one of the spatial dimensions. The spatial axes are drawn in the horizontal plane and the time axis directed upwards. The cone sections at the instants t_1, t_2, t_3 correspond to the snapshots of the spatial representation: the two-dimensional wavefronts are circles whose radius is that of the spherical wavefront at the instant considered.

The light cone shows in a single diagram the continuous history of the wavefront of a light signal.

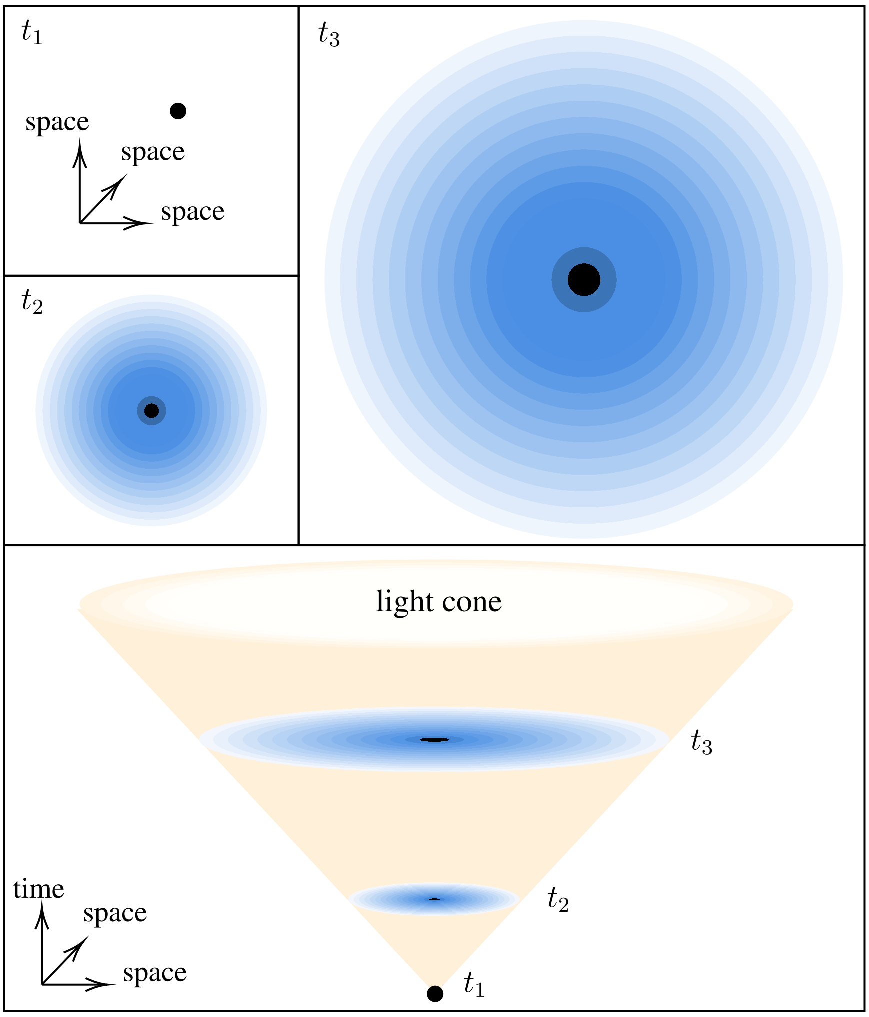


Figure 12.213: **Idea of light-cone**

More precisely, the "snapshots" mentioned above are named "punctual events" and these appear instantaneous (approximation based on geometric optics) to any observer capable seeing them. A collision between two point particles provides an example of a punctual event. It is quite possible that a non-punctual instantaneous event appears instantaneous to a certain observer but, because of the finite propagation velocity of the light, not instantaneous to another observer.

Definitions (#12.3.542):

D1. Two punctual events occupy the same "time-space point" if they appear simultaneously to any observer able to see them.

D2. The set M of all points of space-time is named the "space-time" and it is a mathematical model that fuses the three dimensions of space and the one dimension of time into a single four-dimensional space continuum.

More technically, space-time is a pseudo-Riemannian manifold (for reminder a manifold is a topological space that is locally homeomorphic to Euclidean space and pseudo-

Riemannian means that is equipped with a metric tensor that is everywhere nondegenerate[41]) and that is most of the time[42] Lorentz invariant (for reminder Lorentz invariance refers to the invariance of physical laws under Lorentz transformations, which include boosts - velocity changes - and rotations in four-dimensional space-time) in $\mathbb{R}^4$.

D3. The boundary defined by the Universe cone is named the "cosmological horizon"

Let us recall that if no force acts on a point particle, we say from it that is an "inertial" or "free" particle. We also say that it is in "inertial motion".

Given the point p, $N(p)$ is an absolute geometric structure independent of the observer. Its future component will be denoted $N^+(p)$; Its past component $N^-(p)$ and it will be represented by the following cone:

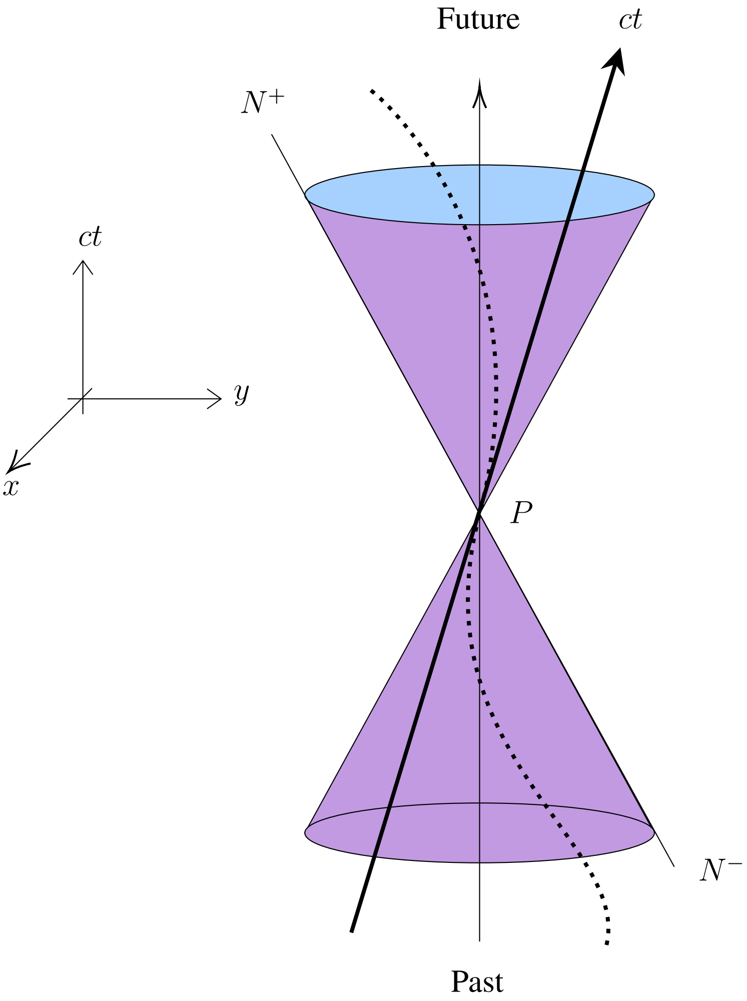


Figure 12.214: **Past and future light cones of an event of a worldline**

Indeed, let us recall that the Minkowski equation is invariant since:

$$c\mathrm{d}t'^2-(\mathrm{d}x'^2+\mathrm{d}y'^2+\mathrm{d}z'^2)=c\mathrm{d}t^2-(\mathrm{d}x^2+\mathrm{d}y^2+\mathrm{d}z^2) \qquad (12.3.417)$$

---

[41] Unlike Riemannian manifolds (where the metric is positive-definite), pseudo-Riemannian manifolds allow for more general metric signatures.

[42] Minkowski and Schwarzschild metric are Lorentz invariant but for example the Roberstswon-Walker metric among many others... is not Lorentez invariant.

We have, when we reduced to three parameters (we remove a spatial dimension to simplify the conceptualisation), if the punctual events are related to the speed of light (see earlier above):

$$c\mathrm{d}t^2-(\mathrm{d}x^2+\mathrm{d}y^2)=0 \qquad (12.3.418)$$

What we can also write in the form:

$$c\mathrm{d}t^2=(\mathrm{d}x^2+\mathrm{d}y^2) \qquad (12.3.419)$$

to be compared with the equation of a cone (see section Analytical Geometry page 2823):

$$(z\tan(\theta))^2=x^2+y^2 \qquad (12.3.420)$$

when we put $c=1$ (which is frequent in theoretical physics as we have already mentioned many times).

Therefore the Minkowski equation can be indeed presented by a cone.

Remark

If we would have keep the three spatial parameters and the time interval constant, the reader will then perhaps have noticed that we would fall back neither on the equation of a cone but on that of a sphere. It is the "celestial sphere" where at a given instant, on its surface, multiple cones of light are created.

The universe line of any observer which occupies instantly p and whose line of the Universe passes through p itself, is contained within $N(p)$ defined by a single point on its celestial sphere (the one that is described by the information vector - the photon - in all directions of space). This means that there can be, in extenso, as many null rays (foci of cones) passing through p as points on a sphere.

The following example will (we hope) appear more obvious:

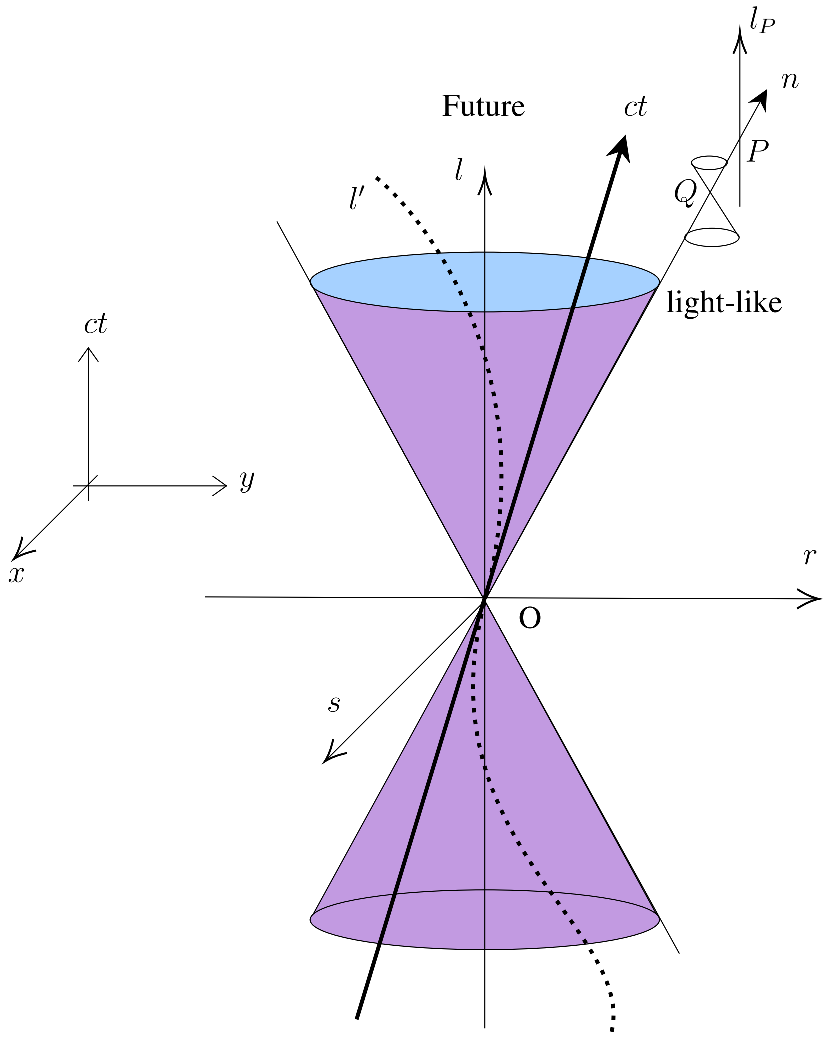


Figure 12.215: **Universe Line Principle with its associated cone**

As illustrated in the figure above, a light event at the point O of the space-time produces a beam of photons, all in the zero cone of the future O, $N^+(\mathrm{O})$ (these photons have been emitted by atoms in various states of movements whose universe lines l and l' pass through O, but are entirely contained within the $N^+(\mathrm{O})$). The universe line n can only be described by a particle moving at the speed of light because it defines the boundary of the cone (we then say that the line of the Universe is of "light type").

Remark

The representation of the Universe lines in the lower part (inverted cone) comes from the fact that an event can also have a past... so the scheme generalizes the particular example.

Let l_p be the universe line of a stationary person P (hence the verticality of its Universe line in the figure above) and n that of a light ray having the origin O. Both lie in the four dimension space and they intersect at a single point P. The points O and P lie on a zero radius (of a cone of the future), n, of $N^+(\mathrm{O})$. In P, the person P sees a sudden flash in the direction defined by n, for him the direction of the luminous event (described only by its velocity therefore, so a universe line of an inertial particle can be described only by time and speed).

An atom whose Universe line cuts n at the point Q absorbs a photon of the luminous event O and re-emits a beam of photons shortly after. These, in turn, form zero rays in $N^{+}(Q)$, but only those of direction n will reach the person P and will be seen by him at the point P.

If P is inside $N(\mathrm{O})$, the zero cone of O, we will say that its universe line is of the "time type". In this case, O and P are located on the universe line of an observer or a massive particle. There are, of course, two types of time displacement:

1. If P is in the future of O (according to an observer whose universe line passes through O and P), we will say that P "points to the future".
2. If not, we will of course say that it "points to the past".

If P is on $N(\mathrm{O})$ - that is to say on the surface of the cone - then we say that it is "null" or of "light type" and if P is neither zero nor of time, then P is outside of $N(\mathrm{O})$ and then we say that it is of "space type":

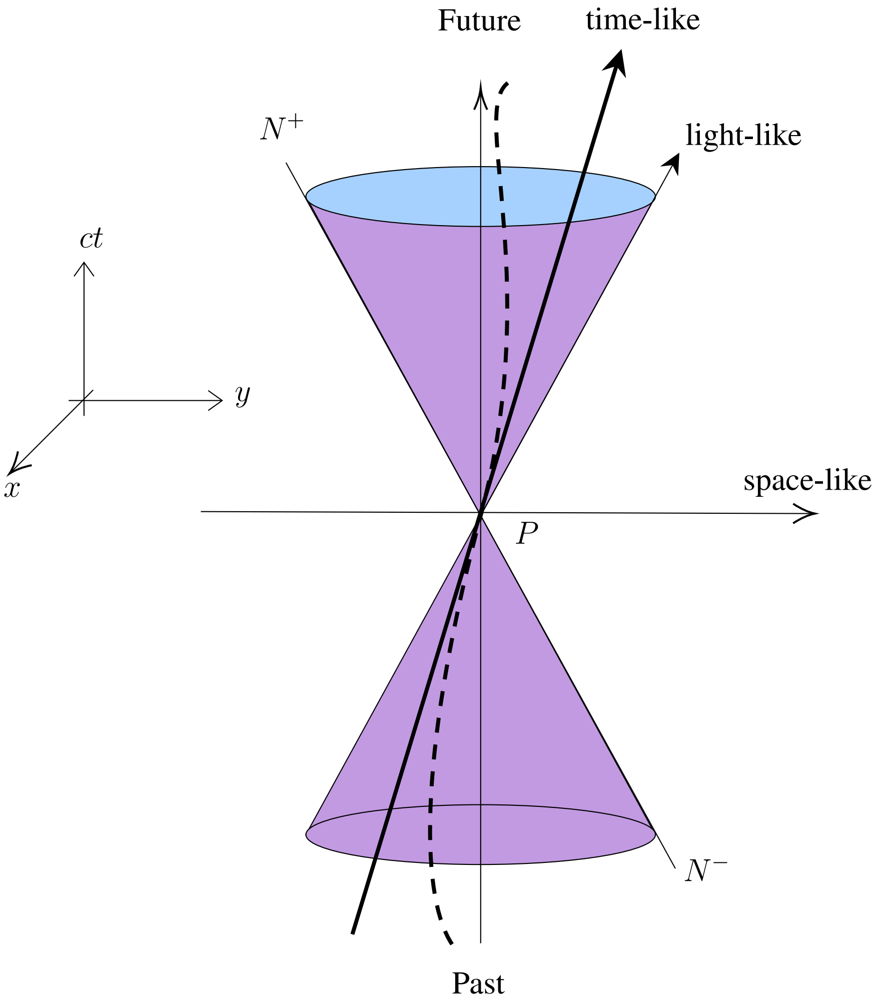


Figure 12.216: Types of universe lines

This is mathematically translated by remembering (see above) that the invariant interval (in the case of the Minkowski metric) is given by:

$$\mathrm{d}s^2=\mathrm{d}x^{\nu}(\eta_{\mu\nu}\mathrm{d}x^{\nu})=\mathrm{d}x^{\nu}\mathrm{d}x_{\nu}=c^2\mathrm{d}t^2-\mathrm{d}x^2-\mathrm{d}y^2-\mathrm{d}z^2 \quad (12.3.421)$$

- $\mathrm{d}s^2=0$ (then $r^2=c^2\Delta^2$): The universe line is therefore "light-like" and it is that latter which describes the surface of the cone by definition (according to what we have demonstrated previously and whatever the choice of the metric) is such that:

  $$c^2\mathrm{d}t^2=\mathrm{d}x^2+\mathrm{d}y^2+\mathrm{d}z^2 \tag{12.3.422}$$

  which is the case of a photon (hence the name ...). In other words, the spatial separation is equal to the distance light travels.

- $\mathrm{d}s^2<0$ (then $r^2<c^2\Delta^2$: We then say that the Universe line is "space-like", therefore such that:

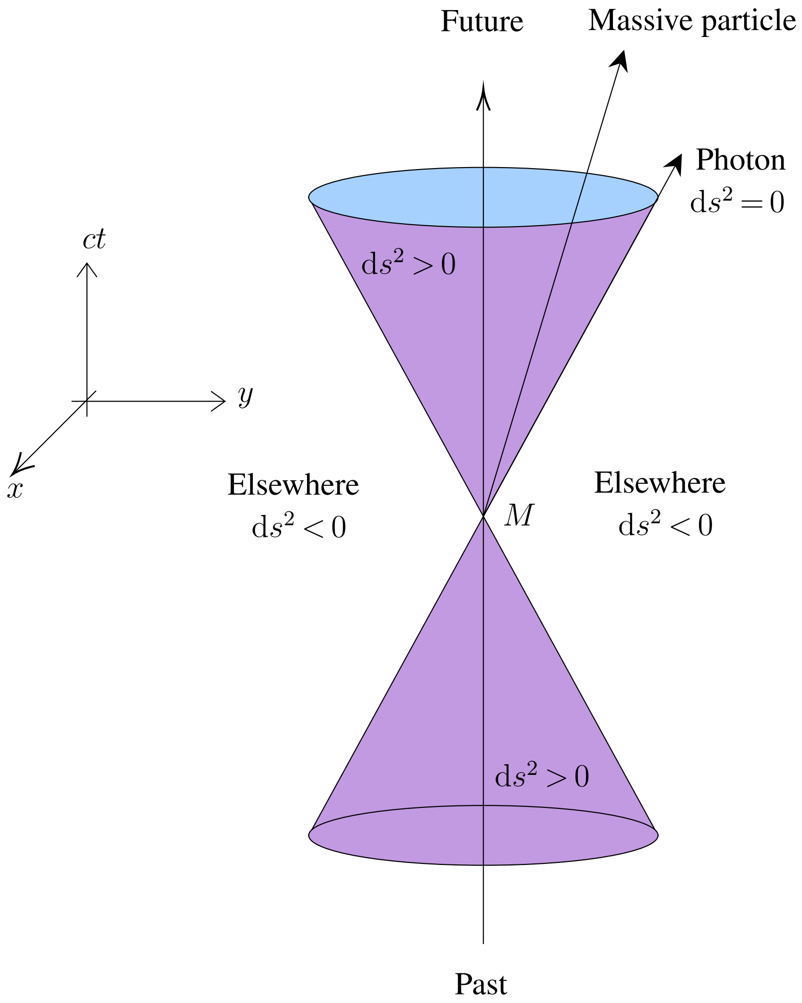
  $$c^2\mathrm{d}t^2<\mathrm{d}x^2+\mathrm{d}y^2+\mathrm{d}z^2 \tag{12.3.423}$$

  Two events that take place simultaneously but at different places are therefore space-like. In other words, the spatial separation is less than the distance light travels.

- $\mathrm{d}s^2>0$ (then $r^2>c^2\Delta^2$: We then say that the Universe line is "time-like", therefore such that:

  $$c^2\mathrm{d}t^2>\mathrm{d}x^2+\mathrm{d}y^2+\mathrm{d}z^2 \tag{12.3.424}$$

  In other words the spatial separation is greater than the distance light travels.


Future
Massive particle
Photon
$\mathrm{d}s^2=0$
ct
$\mathrm{d}s^2>0$
y
x
Elsewhere
$\mathrm{d}s^2<0$
M
Elsewhere
$\mathrm{d}s^2<0$
$\mathrm{d}s^2>0$
Past

- A "causal line" is a time-like or light-like line that is always oriented towards the future.

To summarize:

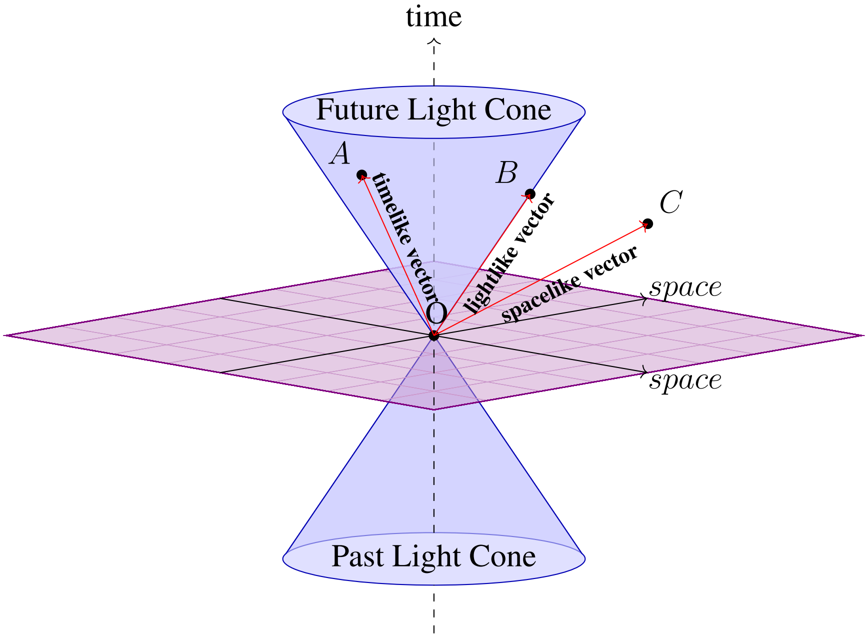

time
Future Light Cone
A
B
C
timelike vector
lightlike vector
spacelike vector
O
space
space
Past Light Cone

Figure 12.217: **Light Cone in 2d Space plus a Time Dimension**

Let us return to our equations after this small interlude ... the equations therefore lead us to several observations. Thus, in the four-dimensional Euclidean Universe of Minkowski, the trajectories of objects in space-time are always straight lines. Indeed, the trivial example consists in considering that the object remains at rest, then only the time then continues to flow. We have therefore:

$$t'^2c^2 = c^2t^2 - v^2t^2 \qquad (12.3.425)$$

by putting $v = 0$, this gives us:

$$c^2t'^2 = c^2t^2 \qquad (12.3.426)$$

therefore:

$$c^2\mathrm{d}t'^2 = c^2\mathrm{d}t^2 \qquad (12.3.427)$$

hence:

$$\mathrm{d}s^2 = c^2\mathrm{d}t^2 \qquad (12.3.428)$$

and also:

$$\mathrm{d}s = c\mathrm{d}t' \qquad (12.3.429)$$

The primitive being (integration constant taken as zero):

$$s = ct' \qquad (12.3.430)$$

which is indeed a straight line and therefore represents the universe line of the object considered in the universe cone. We can also observe that in this case, the evolution of the phenomenon is purely temporal when the interval is positive (which supports what we said earlier).

Now to be sure things are crystal clear....! Imagine computing the interval:

$$\Delta s^2 = (c\Delta t)^2 - \left(\Delta x^2 + \Delta y^2 + \Delta z^2\right),$$

between two events on the worldlines of a massless particle and of a massive particle. This interval will have the same value not matter which inertial reference frame is used to compute it. For the massive particle, we can simplify the computation by choosing to do the calculation in its rest frame. There the space part of the interval is zer and we have

$$\Delta s^2 = (c\Delta t_{\text{rest}})^2 - (\text{ zero })$$

which you used to define the proper time, τ, between the two events.

For the massless particle, that doesn't work: no reference frame exists where the massless particle is at rest. Instead, the massless particle is observed to travel at speed c in all reference frames. The only freedom we have is to orient the velocity along one axis, so that we can say e.g.:

$$\Delta x = c\Delta t \quad \Delta y = \Delta z = 0$$

For this case the interval is:

$$\Delta s^2 = (c\Delta t)^2 - (c\Delta t)^2 = 0$$

which explains why we say that a photon doesn't experience time! Another way to see that is to remember that definition of the proper time interval, for a particle with an arbitrary motion $\vec{v}$ in a reference frame $\mathcal{R}$ between times t_1 and t_2 in this reference frame, is:

$$\Delta\tau = \tau_2 - \tau_1 = \int\limits_{t_1}^{t_2} \mathrm{d}\tau = \int\limits_{t_1}^{t_2} \sqrt{1 - \frac{\vec{v}(t)^2}{c^2}}\mathrm{d}t$$

So, if the particle is a photon, $\vec{v}(t)^2 = c^2$ which leads to $\mathrm{d}\tau = 0$ and $\Delta\tau = 0$: the proper time that elapses between any two events on the world line of a photon is zero!

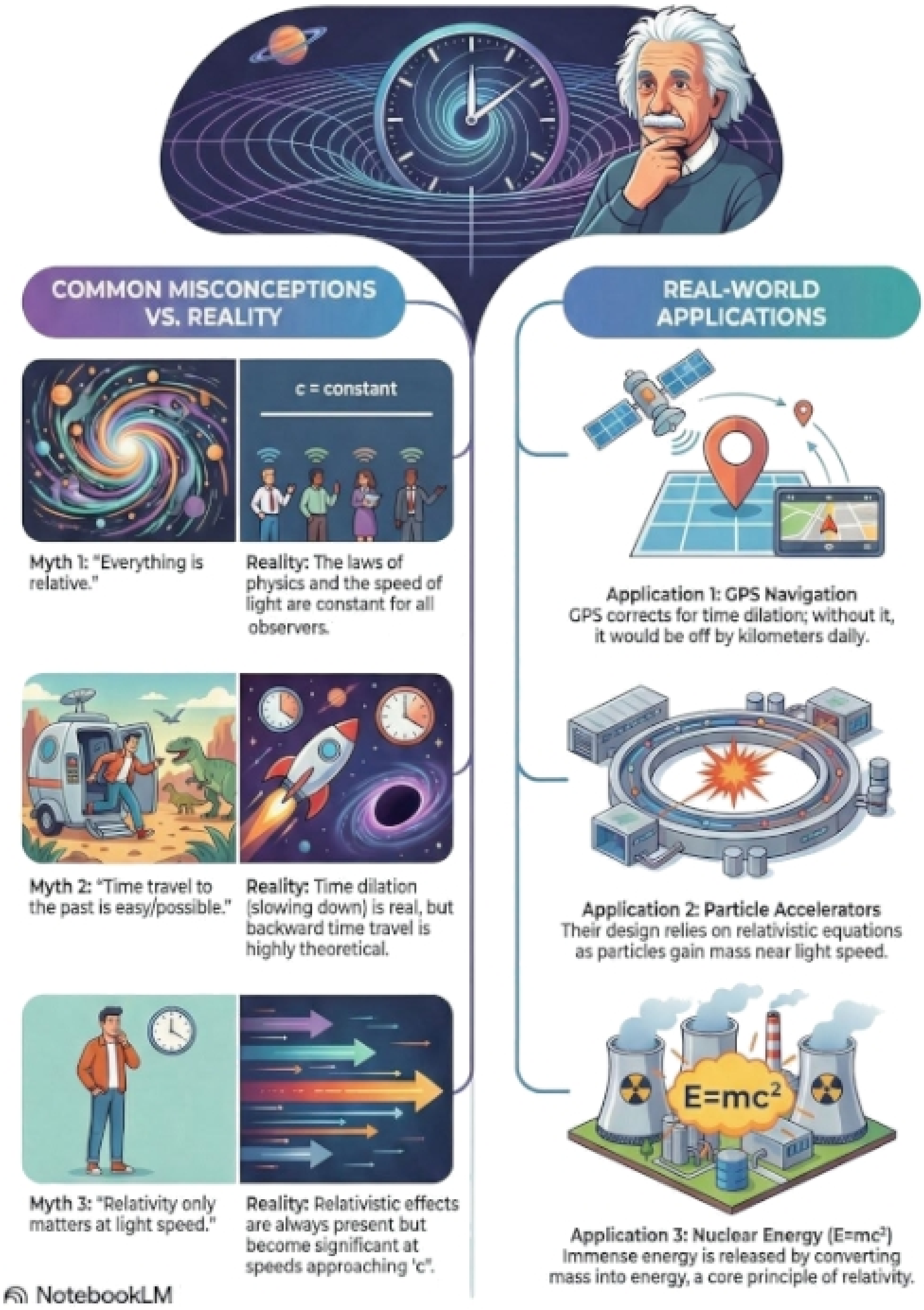
Remarks

R1. If the speed of light is infinite, we fall back on the particular case of the Newtonian universe, where a phenomenon can instantly occur. Time is absolute and there is no cosmological horizon because the cone has a maximum aperture (right angle).

R2. If we put that the velocity of light as equal to 1 (natural units), as we have done it already sometimes, the axis of the ordinate of the cone is named a "purely temporal axis".

R3. It is necessary to understand that our Universe has its own cone of Universe (cone... if the space is of Minkowsky-like of course...).

Finally, let us say that the theory of Special Relativity, like that of General Relativity, does not impose a given number of spatial dimensions in order to remain consistent: this is a pity for theoretical physicists who would like a theory which, imposes on itself a finite number of dimensions to remain consistent (that on the other hand the theory of the strings or superstring).



Myth 1: "Everything is relative."

Reality: The laws of physics and the speed of light are constant for all observers.

Myth 2: "Time travel to the past is easy/possible."

Reality: Time dilation (slowing down) is real, but backward time travel is highly theoretical.

Myth 3: "Relativity only matters at light speed."

Reality: Relativistic effects are always present but become significant at speeds approaching 'c'.

Application 1: GPS Navigation
GPS corrects for time dilation; without it, it would be off by kilometers daily.

Application 2: Particle Accelerators
Their design relies on relativistic equations as particles gain mass near light speed.

Application 3: Nuclear Energy ($E=mc^2$)
Immense energy is released by converting mass into energy, a core principle of relativity.

12.3.4 Empirical evidence of Special Relativity theory

We will now review the 7 classical experimental checks of the 120th century (holocene calendar) of the Special Relativity theory which are:

1. **Michelson-Morley Experiment:** An interferometer was used to compare the speed of light in different directions as Earth moved through space. No measurable difference in the interference pattern was observed. This result showed that the speed of light is the same in all directions and does not depend on the motion of the observer, exactly as

predicted by Special Relativity.

2. **Muon Lifetime (Time Dilation):** Muons are unstable particles created in the upper atmosphere that normally decay very quickly. Despite their short lifetime, many more muons reach Earth's surface than classical physics predicts. This occurs because time passes more slowly for rapidly moving muons, providing direct evidence of relativistic time dilation.

3. **Relativistic Doppler Effect:** Light from stars and other astronomical objects changes wavelength when the source moves relative to the observer. Objects moving away produce a redshift, while objects moving toward the observer produce a blueshift. The measured wavelength shifts match the relativistic Doppler formula and confirm Einstein's theory.

4. **Relativistic Aberration of Light:** The apparent position of stars changes when the observer is moving relative to the incoming light. The observed shift in direction follows the exact relationship predicted by Special Relativity. This demonstrates that measurements of direction in space depend on the relative motion between source and observer.

5. **Mass-Energy Equivalence:** Nuclear reactions can transform a small amount of mass into a large amount of energy. The energy released is quantitatively consistent with the famous equation $E = mc^2$. These observations confirm that mass and energy are equivalent manifestations of the same physical quantity.

6. **Particle Acceleration Near Light Speed:** Experiments with particle accelerators show that particles require increasingly larger amounts of energy as their speeds approach the speed of light. Even enormous increases in energy produce only small increases in velocity near this limit. These results confirm the relativistic prediction that no object with mass can be accelerated to the speed of light.

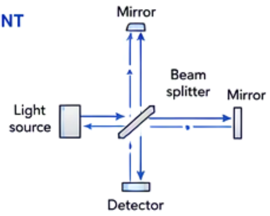
7. **GPS Satellites:** Clocks carried by GPS satellites do not tick at the same rate as clocks on Earth's surface because the satellites are moving at high speed. Relativistic corrections must be continuously applied to maintain accurate positioning. The everyday success of GPS navigation provides a practical and continuously operating verification of Special Relativity.

EXPERIMENTAL EVIDENCE OF
SPECIAL RELATIVITY

Nature consistently confirms the predictions of Einstein's theory.

1 Michelson–Morley Experiment (1887)

An interferometer was used to detect Earth's motion through the "luminiferous ether". No shift in the interference pattern was observed.



WHAT IT SHOWS:

The speed of light is the same in all directions for all observers (no preferred frame of rest).

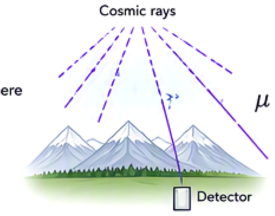
$$\boxed{\Delta L = 0}$$

KEY REFERENCE (PAPER):

A. A. Michelson, E. W. Morley, "On the Relative Motion of the Earth and the Luminiferous Ether", *American Journal of Science*, 34, 333–345 (1887).

2 Time Dilation: Muon Lifetime (1941)

Muons produced in the upper atmosphere reach Earth in larger numbers than expected because their lifetimes are dilated.



WHAT IT SHOWS:

Moving clocks run slower than at rest.

(Time dilation is real.)

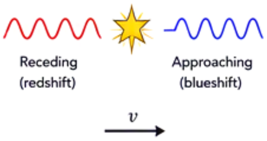
$$\boxed{\Delta t = \gamma \Delta t_0}$$

KEY REFERENCE (PAPER):

J. Clay, R. Van de Water, "Evidence for the Relativistic Time Dilation of Muons in Flight", *Physical Review*, 60, 872–873 (1941).

3 Relativistic Doppler Effect: Starlight Shift (1950s)

Light from stars moving relative to Earth shows the wavelength shift predicted by the relativistic Doppler formula.



WHAT IT SHOWS:

Observed wavelengths shift exactly as relativistic Doppler effect predicts.

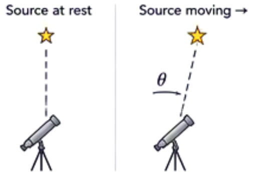
$$\frac{\lambda_{\text{obs}}}{\lambda_{\text{emit}}} = \sqrt{\frac{1+v/c}{1-v/c}}$$

KEY REFERENCE (PAPER):

W. W. Campbell, "Illustration of the Relativistic Doppler Effect in Stellar Spectra", *Physical Review*, 78, 230–237 (1950).

4 Relativistic Aberration of Light (1925)

The apparent position of stars shifts in the direction of motion, as predicted by special relativity.



WHAT IT SHOWS:

Light direction depends on the relative motion of the observer.

$$\cos\theta = \frac{\cos\theta'}{\gamma(1+\frac{v}{c}\cos\theta')}$$

KEY REFERENCE (PAPER):

W. W. Campbell, "The Aberration of Light According to the Theory of Relativity", *Astrophysical Journal*, 61, 357–371 (1925).

5 Relativistic Mass–Energy Equivalence (1932)

Nuclear experiments confirmed that mass can be converted into energy, with the exact amount predicted by $E = mc^2$.

Mass (m)
Energy (E)

WHAT IT SHOWS:

Mass and energy are equivalent.

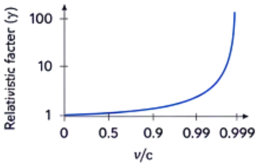
$$\boxed{E = mc^2}$$

KEY REFERENCE (PAPER):

C. D. Anderson, "The Positive Electron", *Physical Review*, 43, 491–494 (1933). (positron discovery)

6 Particle Acceleration Near Light Speed (1960s)

Electrons accelerated to speeds very close to c require increasingly more energy, as predicted by relativistic dynamics.



WHAT IT SHOWS:

As $v \to c$, γ increases without bound: more energy is needed to go faster.

$$\gamma = \frac{1}{\sqrt{1-v^2/c^2}}$$

KEY REFERENCE (PAPER):

M. Sands et al., "Observation of the High-Energy Behavior of Electrons in Storage Rings", *Physical Review Letters*, 5, 278–280 (1960).

7 GPS Satellites (Real-World Test)

GPS satellites experience time dilation due to both their speed and weaker gravity. Relativistic corrections are essential for meter-level accuracy.

WHAT IT SHOWS:

Relativity is required in everyday technology: time runs differently in moving frames.

KEY REFERENCE (PAPER):

T. M. Eubanks, "Relativistic Clock Synchronization", *GPS Solutions*, 6(3), 50–62 (2002). (Review of relativistic effects in GPS)

THE BIG PICTURE

Time is relative. Moving clocks run slower.

Space is relative. Distances and directions depend on motion.

Energy and mass are the same. $E = mc^2$

Across many experiments, all observations agree with the predictions of Special Relativity.

∼95 %

★★★★★

48 votes, 71.25%

12.4. General Relativity

As we saw it, in the previous section, Special Relativity is a remarkable achievement from a theoretical point of view as well as a practical point of view, forming a continuum of space-time where the space variables and time are given the same physical dimension (that of a distance metric for reminder!). However, this applies only to the Euclidean frames and to inertial/Galileans reference frames (constant speed reminder...). It is therefore appropriate to first generalize the entire mechanic theory by expressing its principles and fundamental results in a generalized form independent of the type of coordinate system chosen (that is to say: independent of the space properties) using for this purpose tensor calculus and then to take into account the non-inertial systems. The equivalence of inertial systems by Special Relativity and the non-equivalence of inertial systems can then shortly be resumed (a little bit basically...) by saying that speed is relative but the acceleration is absolute! Thus, we can never rest distinguish a uniform motion, but we can distinguish them from an accelerated motion.

It should also be consider the fact that Special Relativity applies only to Galileans frames is restrictive because any mass generates a gravitational field whose scope is endless. To find a true Galilean frame, it is therefore necessary to lie infinitely far from any mass. Relativistic mechanics built from Special Relativity therefore constitutes an approximation of the laws of nature, where the gravitational fields or accelerations are low enough. This application limitation is not more suited to relativistic astrophysics whose activity has intensified in the late 120th century (holocene calendar).

quote
12.4:5720.#70

> *Let no one ignorant of geometry enter here!*
>
> Plato

Remark

Following multiple requests here is a summary of the differences between the textbook *Gravitation* (considered as an old reference) of C. W. Misner, K. S. Thorne, J. A. Wheel and *Opera Magistris* for the subjects related to Gravitation and Special/General Relativity: The content and topics covered are quite similar; however *Gravitation* lacks many detailed proofs. The notations and the technology used to write *Gravitation* is quite outdated (11973 according to holocene calendar) and can be a bit cumbersome to read (and in addition all the figures and the few photos are in grayscale...). If we calculate the ratio of pages sizes including all the margins, in 12024 (holocene calendar) the number of pages related to the subjects of Gravitation and General Relativity is almost identical in both textbooks (within about ten pages). Furthermore, the content of *Gravitation* isn't updated anymore since 11973 (holocene calendar) to reflect recent experimental observations, new parameter values, improved models and new speculative ones.

Definition (#12.4.543): "General Relativity", also known as the "general theory of relativity", and as "Einstein's theory of gravity", is the geometric theory of gravitation that generalizes Special Relativity and refines Newton's law of universal gravitation, providing a unified description of gravity as a geometric property of space and time, or four-dimensional space-time. In particular, the curvature of space-time is directly related to the energy, momentum and stress of whatever is present, including matter and radiation.

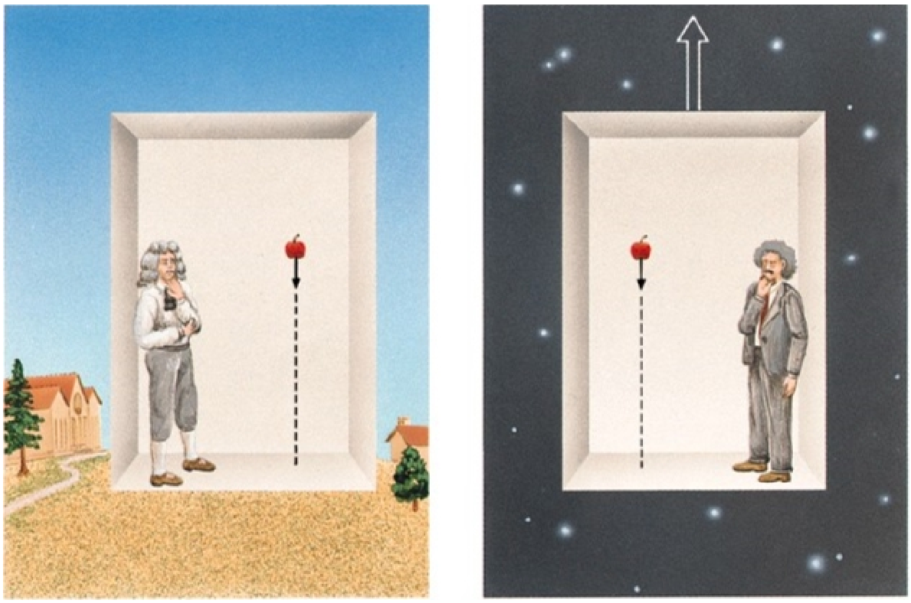
12.4.1 Assumptions and Principles

Albert Einstein and some others of his time believed in a physics that not to favour any frame system since that was in their eyes the reality of the Universe (we have already mentioned this point of view). But how can we subtract ourselves to the phenomenon of acceleration? The brilliant idea was to state the "equivalence postulate" below (which still in this early 121st century (holocene calendar) has not show any default by recent known experiences) plus the "invariance postulate" and "cosmological principle" we have already stated in the section of Special Relativity and the assumption that the motion of a particle that does not undergo any other interaction that gravitation follows a geodesic line (see below for the detailed proof).

12.4.1.1 Equivalence Postulates

At first, Albert Einstein will improve the equivalence postulate (also named "equivalence principle") whose older versions are due to Galileo and Newton:

Postulate: The (uniform!) acceleration of a mass (outside gravitational field) due to application of mechanical force and the acceleration of that mass subjected to a gravitational field are supposed completely equivalent. Thus, the results of mathematical analysis in one case may apply to the other (here this is already smart but consistent ... the idea is very good still had to have it...!).



In other words, the gravity field has a fundamental property which distinguishes it from all other fields known in nature: the free fall movement of bodies is universal, independent of the mass and composition of the bodies.

Corollary: The rest mass of a body must be the same whether it is measured in a frame within a gravitational field or outside a gravitational field (we speak then about "inertial mass" and of "gravitational mass" as we have already studied it in the section of Classical Mechanics at page 3087).

Remark

We must be careful and check that the corollary of the assumption of the equivalence principle is true otherwise all General Relativity would collapse (in the early 121st century (holocene calendar) experiments are trying to show a default to this principle)!

So all static and uniform gravitational field is equivalent to an accelerated frame in vacuum. We can consider any physical gravitational field as static and uniform in a relatively small region of space, and for a relatively short period of time to avoid the tides effects. We are thus led to state the "Weak Equivalence Principle" (WEF): For any event in space-time in an arbitrary gravitational field, we can choose a frame named "locally inertial frame" such as in the neighbourhood of the event of interest the free movement of all body (which are also in the gravity field!) is straight and uniform as we are able to apply the Lorentz transformations (see section Special Relativity page 5633). In other words, it is not possible to distinguish a system in vacuum space far away from any star (gravitational source) from as system falling in a constant homogeneous gravitational field.

If we experimentally show that WEF fails, then we have put into default the equivalence principle itself ... which has never been achieved in the laboratory to this date! For information the equivalence principle has even been confirmed at the quantum level (see [152]).

Remark

The concept of "locality" is very important because in reality we don't know any natural uniform gravitational field. For example, on Earth, two body distant of a certain length dropped from a certain height will fall to the ground with a shorter distance than the distance between them when they were released. This is what we name in physics the "tide effect": the gravitational field is never uniform (as far as we know...).

So the postulate of equivalence (which includes the principle of weak equivalence) finally asserts that the Newton's force on inertial mass m_i:

$$\vec{F}=m_i\vec{a} \tag{12.4.1}$$

and that of gravitation in the form of the Newton-Poisson law (see section Astronomy page 5267) with gravitational mass m_g:

$$\vec{F}_g=m_g\vec{g}=-m_g\vec{\nabla}\cdot\phi \tag{12.4.2}$$

are equivalent such as the inertial mass equals the gravitational mass and acceleration equal to gravity and that it is not possible to distinguish the both such that:

$$\begin{gathered} m_i = m_g \\ \updownarrow \\ \vec{a} = \vec{g} \end{gathered} \tag{12.4.3}$$

Remark

We must be careful and check that the corollary of the assumption of the equivalence principle is true otherwise all General Relativity would collapse! In this early 121st century - holocene calendar - experiments (like [482]) are still trying to find a default to this principle!

In what this postulate allows to resolve all the problems therefore? It's simple! The idea is the following:

When we will consider a body in acceleration, we first always equate it to the acceleration due to the fall in a gravitational field (by applying the postulate of equivalence). Then, we will assume, and will have to check (see proof further below) by rediscovering Newton's law, that acceleration due to the gravitational field is not due to the field itself but to the geometry of the deformed space by the presence of the mass (i.e. energy) that generates the gravitational field. Thus, the object is no longer in "free fall" but will be seen as sliding on the distorted spatial frame to acquire therefore its acceleration.

1. If the tensor calculus gives the possibility to express the laws of classical and relativistic mechanics in any coordinate system, it is then possible to see how the coordinate system (metric) acts on the expression of the laws of the Universe (Albert Einstein did not know that fact as he had not completed his calculations but had a presentiment about this)!
2. If the natural tensor expression of the laws of mechanics shows slippage (i.e. acceleration) on the spatial frame following the (local) considered metric, then the bet is won and then the acceleration can be seen as an effect whose cause is purely geometrical.

✍**Example:**

Suppose that two rockets, which we denoted by A and B are in a region of space away from any body. Their engines are stopped which physically results in a uniform motion. In each rocket, physicists are making mechanics experiments with objects which they know the inert mass. Suddenly the engine of the rocket A starts and communicates to it an acceleration whose effects felt inside the spaceship is an inertia force that constraint objects going to the floor. For physicists A rocket laws of mechanics are then the same as that observed in a gravitational field. They are logically led to interpret the force of inertia as the manifestation of a gravitational field. Using a balance, they can weigh their objects and assign to them a gravitational mass.

Suppose now that the physicists in rocket B could observe what happens in the rocket A. They know what their colleagues interpret as the weight of objects is in fact a force of inertia. The inertial force is proportional to the acceleration and the inertial mass. If the gravitational mass was different from the inertial mass the physicists of the rocket A could distinguish the effects of inertial forces from those of a gravitational field because the measured masses are distinct. We know that the inertial and gravitational mass are equivalent (Galilean principle of equivalence). It follows that the physicists of the rocket A have no way to differentiate between inertial forces resulting from an accelerated motion of their spaceship and the gravitational attraction forces.

However, we must temper the conclusions from this experience: the real gravitational fields differ from an accelerated frame since the gravitational acceleration varies with the distance to the main body while in an accelerated reference frame, the acceleration is the same at any point in space. However, locally, a gravitational field and an accelerated frame can not be differentiated!!!

We are led now to state the "Einstein's Equivalence Principle" (EEP) as did Albert Einstein: locally all the laws of physics are the same in a gravitational field and a uniformly accelerated frame.

This has a consequence: If the mass (which is equivalent to the energy as we have proved in the section of Special Relativity at page 5679) of an object is not differentiable that we are in a gravitational field or a uniformly accelerated frame that means that all types of energies (nuclear cohesion energy, electrostatic energy, proper gravitational energy of the object, etc.) of this object are indistinguishable. So the laws of Special Relativity are also valid whatever the considered frame and the principle of equivalence leads to two testable conclusions about the propagation of light. If the effects of gravitation and acceleration are indistinguishable, then rays of light should bend in a gravitational field. As well, light moving up through a gravitational field should be redshifted!

If the laws are not the same, then EEP (Einstein Equivalence Principle) is faulted, so verbatim WEF (weak equivalence principle) also and more globally the principle of equivalence in general but this has never happened experimentally as far as we know at this day.

✍**Example:**

By the WEF, it is interesting to note that the gravitational field also acts on the gravitational potential energy of the other bodies. We say then that the gravitational field is a "coupled field".

Caution! It turns out that in General Relativity, it's very hard to make the notion of "gravitational potential energy" precise. For example, you can't talk about the gravitational potential energy density at a point, because you can always go into a freely falling frame there, where the observed gravitational field is zero. For this reason, relativity textbooks generally say that the gravitational potential energy is not defined at all; instead energy (defined as not including this extra ill-defined piece) simply isn't conserved in General Relativity. If the reader may think this is unacceptable, remember that the only reason we elevated conservation of energy to an important principle in the 119th century (holocene calendar) was that it was observed to work in everyday situations. We never tested it in exotic situations like those with curved space-time, so there's no reason to expect the principle to continue to hold up (but many textbooks on General Relativity do so and we will also do it later ourselves...). At a deeper level, Noether's theorem tells us that energy conservation is related to time translation invariance (i.e. «our system will work the same if we start it at some other time» to not be confused with time reflection symmetry where «our system will work the same if we run it backwards»), and we don't have that in an expanding universe (indeed the energy carried by radiation decreases as the universe expands since every photon's wavelength increases).

Given that in General Relativity, the gravitational field is supposed to be described by the metric $g_{\mu\nu}$ (from which the 4-dimensional differentiable manifold that is space-time is supposed to be made), we can see a locally inertial frame as a coordinate system of space-time in which the metric becomes flat (pseudo-Riemannian):

$$g_{\mu\nu} \cong \eta_{\mu\nu} \tag{12.4.4}$$

using the notation introduce in the section of Tensor Calculus.

Such a coordinate system will by hypothesis always exists, indicating the existence, for any gravitational field, of locally inertial frames!

However when the metric is not flat the coordinates are named "Riemann normal coordinates" and then describes a Riemann metric space (curved space) and itself depends in a non-trivial way of the coordinates system (see sections Tensor Calculus page 2092 and Non-euclidean Geometry page 2672).

12.4.1.2 Mach Principle

If the equivalence principle highlights the equality of inert and gravitational mass, it does not enlighten us about the nature of these two masses. Finally, what are the inert and gravitational mass?

The deep nature of the inert mass should inform us about the inertia itself. The inertia is manifested in a passive form - the principle of inertia - and an active form - the second Newton's law. In general, it expresses a universal behaviour of bodies to resist to the change of movement. But we know that inertial motion is relative, that is to say that there is no absolute referential frame. Is it the same with the accelerated movement? Consider, to illustrate this question, a rocket in which has taken place a physicist and let us carry two experiments:

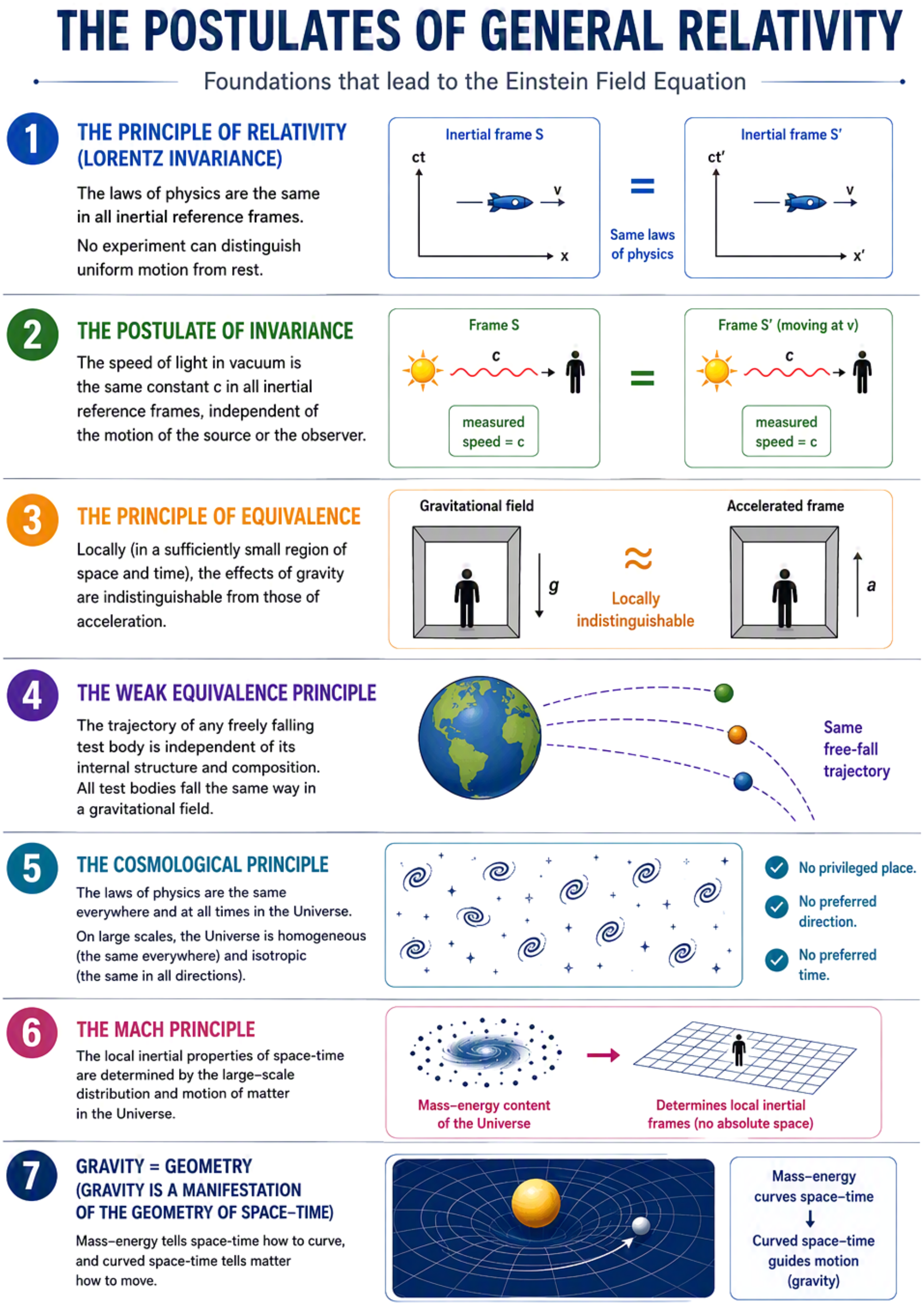
1. First experience: The rocket accelerates and the physicist is subjected to an inertial force oriented in the direction opposite to that of acceleration.
2. Second experience. Now assume that we gives to the whole Universe - at the exception of the rocket that moves in an inertial motion - an acceleration exactly opposite to the one of the rocket of the preceding experiment.

If the accelerated motion is relative then, for an observer, it is not possible to distinguish the two experiments. In particular, the physicist located inside the rocket must observe the emergence of an inertial force absolutely identical to the one he noted in the first experiment. The inert mass could then has its origin in the interactions of the gravitational mass of bodies with all the gravitational mass of the Universe! It is as if by moving all masses of the Universe, they dragged with them the objects in the rocket, the physicist therefore experienced a force that pulls in the same direction as the acceleration applied to stars.

Following Ernst Mach, physicist and philosopher of the 119th century (holocene calendar), the movement whatsoever inertial or accelerated, is relative.

This theory was named by Albert Einstein "Mach principle". At this date, Mach's principle has not been confirmed, but nor rejected! It is true that its experimental verification far exceeds our actual human capacities!

So if we add up in an infography the above postulates plus those of Special Relativity, we get:



THE POSTULATES OF GENERAL RELATIVITY

Foundations that lead to the Einstein Field Equation

1 THE PRINCIPLE OF RELATIVITY (LORENTZ INVARIANCE)

The laws of physics are the same in all inertial reference frames.

No experiment can distinguish uniform motion from rest.

Inertial frame S
ct
v
x
=
Same laws of physics
Inertial frame S′
ct′
v
x′

2 THE POSTULATE OF INVARIANCE

The speed of light in vacuum is the same constant c in all inertial reference frames, independent of the motion of the source or the observer.

Frame S
c
measured speed = c
=
Frame S′ (moving at v)
c
measured speed = c

3 THE PRINCIPLE OF EQUIVALENCE

Locally (in a sufficiently small region of space and time), the effects of gravity are indistinguishable from those of acceleration.

Gravitational field
g
≈
Locally indistinguishable
Accelerated frame
a

4 THE WEAK EQUIVALENCE PRINCIPLE

The trajectory of any freely falling test body is independent of its internal structure and composition. All test bodies fall the same way in a gravitational field.

Same free-fall trajectory

5 THE COSMOLOGICAL PRINCIPLE

The laws of physics are the same everywhere and at all times in the Universe.

On large scales, the Universe is homogeneous (the same everywhere) and isotropic (the same in all directions).

✓ No privileged place.
✓ No preferred direction.
✓ No preferred time.

6 THE MACH PRINCIPLE

The local inertial properties of space-time are determined by the large–scale distribution and motion of matter in the Universe.

Mass–energy content of the Universe
Determines local inertial frames (no absolute space)

7 GRAVITY = GEOMETRY (GRAVITY IS A MANIFESTATION OF THE GEOMETRY OF SPACE–TIME)

Mass–energy tells space-time how to curve, and curved space-time tells matter how to move.

Mass–energy curves space–time
↓
Curved space–time guides motion (gravity)

Figure 12.218: **Summary of the postulates of General Relativity**

12.4.2 Metrics

Albert Einstein assumed that gravity was only the manifestation of space-time distortions (see [13]). To try to illustrate in the more possible simple and illustrated way the idea of Albert Einstein, consider a rolling gear at constant speed (say, one tooth at a second) on a rack. Imagine that we have the power to simultaneously change the pitch of the rack and the wheel when and where we wish. Let us do things such that the pitch of the rack slightly increases from one tooth to another. For fixed observers the gear is then driven with a uniformly accelerated motion as, in effect, at each turn thereof always travels a greater distance. On the other hand, if one chooses the rack as a reference and thereof the pitch as a standard to measure the movement of the wheel is then uniform (one tooth per second). The acceleration of the wheel is the consequence of the increase in the pitch of the rack.

Let us continue the analogy: the pitch of the rack acts as a local measurement standard in our one-dimensional space that represents the rack. In geometry, it is named the "metric". The metric is what determines the distance between two points, it is somehow the standard infinitesimal space unit. In Euclidean geometry, the metric is constant, allowing us to define universal measurement standards. Bernhard Riemann, for example, invented a metric geometry which can vary from one point to another in space, which allowed him to describe curved spaces like the surface of a sphere, for example (see section Non-Euclidean Geometry page 2663).

During our study of Tensor Calculus, Non-Euclidean Geometries and Differential Geometry (section that the reading is more than recommended!!!) we have seen that the measurement of the curvilinear distance ds between two points positioned in a two or three dimensions space can be made using a large number of coordinate system by the "metric equation" (see section Special Relativity page 5704 and section of Non-Euclidean Geometry page 2669):

$$\mathrm{d}s^2=\mathrm{d}x_\nu \mathrm{d}x^\mu=g_{\mu\nu}\mathrm{d}x^\mu \mathrm{d}x^\nu \tag{12.4.5}$$

In General Relativity, the idea is to make the theoretical model independent of the background and thus build it in a covariant form (which some physicists liken to assimilate to a postulate named the "covariance principle"). An excellent candidate for this type of approach is to use the tensor formalism. This is the reason why the metric equation is therefore one of the pillars.

✍**Examples:**

E1. Rectangular coordinates (in $\mathbb{R}^3$):

$$\boxed{\mathrm{d}s^2=\mathrm{d}x^2+\mathrm{d}y^2+\mathrm{d}z^2} \tag{12.4.6}$$

If the squared distance satisfies this relation then we are in a flat space or at least locally flat (see section Non-Euclidean Geometry page 2669).

E2. Polar coordinates (in $\mathbb{R}^2$):

$$\begin{aligned}x=r\cos(\phi)&\Rightarrow \mathrm{d}x=\mathrm{d}r\cos(\phi)-r\sin(\phi)\mathrm{d}\phi\\ y=r\sin(\phi)&\Rightarrow \mathrm{d}x=\mathrm{d}r\sin(\phi)+r\cos(\phi)\mathrm{d}\phi\end{aligned} \tag{12.4.7}$$

Therefore:

$$\begin{aligned}\mathrm{d}s^2&=\mathrm{d}x^2+\mathrm{d}y^2=(\mathrm{d}r\cos(\phi)-r\sin(\phi)\mathrm{d}\phi)^2+(\mathrm{d}r\sin(\phi)+r\cos(\phi)\mathrm{d}\phi)^2\\&=\mathrm{d}r^2\cos^2(\phi)-\cancel{2r\mathrm{d}r\sin(\phi)\cos(\phi)\mathrm{d}\phi}+r^2\sin^2(\phi)\mathrm{d}\phi^2+\mathrm{d}r^2\sin^2(\phi)\\&+\cancel{2r\mathrm{d}r\sin(\phi)\cos(\phi)\mathrm{d}\phi}+r^2\cos^2(\phi)\mathrm{d}\phi^2\\&=\mathrm{d}r^2(\cos^2(\phi)+\sin^2(\phi))+r^2(\cos^2(\phi)+\sin^2(\phi))\mathrm{d}\phi^2\end{aligned}\tag{12.4.8}$$

Therefore:

$$\boxed{\mathrm{d}s^2=\mathrm{d}r^2+r^2\mathrm{d}\phi^2}\tag{12.4.9}$$

If the squared distance satisfies this relation then we are in a flat space or at least locally flat (see section Non-Euclidean Geometry page 2669).

E3. Cylindrical coordinates (in $\mathbb{R}^3$):

$$\begin{aligned}x=r\cos(\phi)&\Rightarrow\mathrm{d}x=\mathrm{d}r\cos(\phi)-r\sin(\phi)\mathrm{d}\phi\\y=r\sin(\phi)&\Rightarrow\mathrm{d}x=\mathrm{d}r\sin(\phi)+r\cos(\phi)\mathrm{d}\phi\\z=z&\Rightarrow\mathrm{d}z=\mathrm{d}z\end{aligned}\tag{12.4.10}$$

when we put this into $\mathrm{d}s^2=\mathrm{d}x^2+\mathrm{d}y^2+\mathrm{d}z^2$ we get in a similar way as before:

$$\mathrm{d}s^2=\mathrm{d}r^2+r^2\mathrm{d}\phi^2+\mathrm{d}z^2\tag{12.4.11}$$

If the squared distance satisfies this relation then we are in a curved space (cylindrical type) but that may be locally flat (see section Non-Euclidean Geometry page 2669). In fact, to have the metric of the cylinder surface and not simply of the plane expressed in cylindrical coordinates, we must take the following metric:

$$\boxed{\mathrm{d}s^2=\mathrm{d}r^2+r^2\mathrm{d}\phi^2}\tag{12.4.12}$$

whose origin was proved in the section of Differential Geometry and also... just previously...

E4. Spherical coordinates (in $\mathbb{R}^3$) for which we have:

$$\begin{aligned}&x=r\sin(\theta)\cos(\phi)\Rightarrow\mathrm{d}x=\mathrm{d}r\sin(\theta)\cos(\phi)+r\cos(\theta)\mathrm{d}\theta\cos(\phi)-r\sin(\theta)\sin(\phi)\mathrm{d}\phi\\&y=r\sin(\theta)\sin(\phi)\Rightarrow\mathrm{d}x=\mathrm{d}r\sin(\theta)\sin(\phi)+r\cos(\theta)\mathrm{d}\theta\sin(\phi)+r\sin(\theta)\cos(\phi)\mathrm{d}\phi\\&z=r\cos(\theta)\Rightarrow\mathrm{d}z=\mathrm{d}r\cos(\theta)-r\sin(\theta)\mathrm{d}\theta\end{aligned}\tag{12.4.13}$$

when we put this into $\mathrm{d}s^2=\mathrm{d}x^2+\mathrm{d}y^2+\mathrm{d}z^2$ we get:

$$\begin{aligned}\mathrm{d}s^2=&(\mathrm{d}r\sin(\theta)\cos(\phi)+r\cos(\theta)\mathrm{d}\theta\cos(\phi)-r\sin(\theta)\sin(\phi)\mathrm{d}\phi)^2\\&+(\mathrm{d}r\sin(\theta)\sin(\phi)+r\cos(\theta)\mathrm{d}\theta\sin(\phi)+r\sin(\theta)\cos(\phi)\mathrm{d}\phi)^2\\&+(\mathrm{d}r\cos(\theta)-r\sin(\theta)\mathrm{d}\theta)^2\end{aligned}\tag{12.4.14}$$

Now remember that (see section Algebra Calculus page 1361):

$$\begin{aligned}(a+b-c)^2&=a^2+b^2+c^2+2(ab-ac-bc)\\(a+b+c)^2&=a^2+b^2+c^2+2(ab+ac+bc)\\(a-b)^2&=a^2-2ab-b^2\end{aligned}\qquad(12.4.15)$$

Therefore:

$$\begin{aligned}\mathrm{d}s^2=&\,\mathrm{d}r^2\sin^2(\theta)\cos^2(\phi)+r^2\cos^2(\theta)\mathrm{d}\theta^2\cos^2(\phi)+r^2\sin^2(\theta)\sin^2(\phi)\mathrm{d}\phi^2\\&+2(r\mathrm{d}r\mathrm{d}\theta\cos(\theta)\sin(\theta)\cos^2(\phi)-r\mathrm{d}r\mathrm{d}\phi\sin^2(\theta)\cos(\phi)\sin(\phi)\\&-r^2\mathrm{d}\theta\mathrm{d}\pi\cos(\theta)\sin(\theta)\cos(\phi)\sin(\phi)\\&+\mathrm{d}r^2\sin^2(\theta)\sin^2(\phi)+r^2\cos^2(\theta)\mathrm{d}\theta^2\sin^2(\phi)+r^2\sin^2(\theta)\cos^2(\phi)\mathrm{d}\phi^2\\&+2(r\mathrm{d}r\mathrm{d}\theta\sin(\theta)\cos(\theta)\sin^2(\phi)+r\mathrm{d}r\mathrm{d}\phi\sin^2(\theta)\sin(\phi)\cos(\phi)\\&+r^2\mathrm{d}\theta\mathrm{d}\phi\cos(\theta)\sin(\theta)\sin(\phi)\cos(\phi)\\&+\mathrm{d}r^2\cos^2(\theta)-2r\mathrm{d}r\mathrm{d}\theta\cos(\theta)\sin(\theta)+r^2\sin^2(\theta)\mathrm{d}\theta^2\end{aligned}\qquad(12.4.16)$$

After a first set of factorization and basic simplifications of identical terms, we get:

$$\mathrm{d}s^2=\mathrm{d}r^2+r^2\mathrm{d}\theta^2+r^2\sin^2(\theta)\mathrm{d}\phi^2\qquad(12.4.17)$$

If the squared distance satisfies this relation then we are in a curved space (spherical type) but that locally may be flat (see section Non-Euclidean Geometry page 2669). In fact, for the metric of the surface of the sphere and not simply of the plane expressed in spherical coordinates, we have to take the following metric:

$$\boxed{\mathrm{d}s^2=r^2\mathrm{d}\theta^2+r^2\sin^2(\theta)\mathrm{d}\phi^2}\qquad(12.4.18)$$

whose origin has been proved in the section of Differential Geometry (see page 2884). We also checked in the section of Tensor Calculus, that the Ricci curvature of the spherical prior-previous metric was zero (see page 2127). By cons, we had right after checked that if we took the previous metric of the surface of the sphere, the Ricci curvature was not zero (and it is still happy!).

Until then, you may be wondering where we are going? In fact, we try to define from these relations, a mathematical being that consistent with the Einstein's hypothesis, expresses the geometric properties of given space.

How we will do this?: We first simply change the notations. Instead of using the symbols (x,y,z,θ,ϕ,r) we will write $x^1,x^2,x^3,\ldots$.

Caution! The numbers suffixes are not powers!!! These are dummy values that are only there to symbolize the x-th coordinate of a given basis.

Now let us write again our metric equations with this new notation by considering it is only specific examples that do not necessarily have relevant physical sense (we also mentioned it earlier!):

- Rectangular coordinates:

$$\mathrm{d}s^2=(\mathrm{d}x^1)^2+(\mathrm{d}x^2)^2+(\mathrm{d}x^3)^2 \tag{12.4.19}$$

- Polar coordinates:

$$\mathrm{d}s^2=(\mathrm{d}x^1)^2+(x^1)^2(\mathrm{d}x^2)^2 \tag{12.4.20}$$

- Cylindrical coordinates:

$$\mathrm{d}s^2=(\mathrm{d}x^1)^2+(x^1)^2(\mathrm{d}x^2)^2+(\mathrm{d}x^3)^2 \tag{12.4.21}$$

- Spherical coordinates:

$$\mathrm{d}s^2=(\mathrm{d}x^1)^2+(x^1)^2(\mathrm{d}x^2)^2+(x^1)^2\sin^2(x^2)(\mathrm{d}x^3)^2 \tag{12.4.22}$$

Now let us recall again that the "metric tensor" (so named because it calibrates space-time) denoted :

$$g_{\mu\nu}=\begin{bmatrix} g_{11} & g_{12} & g_{13}\\ g_{21} & g_{22} & g_{23}\\ g_{31} & g_{32} & g_{33}\end{bmatrix} \tag{12.4.23}$$

is involved in the metric equation as follows in the Lorentz invariant (see section Special Relativity page 5704):

$$\mathrm{d}s^2=\mathrm{d}x_\mu\mathrm{d}x^\nu=g_{\mu\nu}\mathrm{d}x^\mu\mathrm{d}x^\nu \tag{12.4.24}$$

and notice that the components of the matrix are also dimensionless!

This mathematical entity which is a tensor thus contains the parameters of the curvature (we also sometimes say of the "stress" or "tension") wherein a space is located. But then what contains the metric tensor of space-time for a flat Euclidean space?

According to the summing writing Einstein's convention (see section Tensor Calculus page 2044), for example, for $\mu=\nu=2$ we have:

$$\mathrm{d}s^2=g_{\mu\nu}\mathrm{d}x^m\mathrm{d}x^n=g_{11}\mathrm{d}x^1\mathrm{d}x^1+g_{12}\mathrm{d}x^1\mathrm{d}x^2+g_{21}\mathrm{d}x^2\mathrm{d}x^1+g_{22}\mathrm{d}x^2\mathrm{d}^2x \tag{12.4.25}$$

So if we return to our tensor for the flat Euclidean space, we already know (see section Vector Calculus page 1810) that m and n goes from 1 to 3 and we have in our tensor $g_{\mu\nu}=0$ for $\mu\neq\nu$ and $g_{\mu\nu}=1$ for $\mu=\nu$ (symmetrical tensor). So:

$$\begin{aligned}\mathrm{d}s^2&=g_{\mu\nu}\mathrm{d}x^\mu\mathrm{d}x^\nu=g_{11}\mathrm{d}x^1\mathrm{d}x^1+g_{22}\mathrm{d}x^2\mathrm{d}x^2+g_{33}\mathrm{d}x^3\mathrm{d}x^3\\&=1\cdot(\mathrm{d}x^1)^2+1\cdot(\mathrm{d}x^2)^2+1\cdot(\mathrm{d}x^3)^2=(\mathrm{d}x^1)^2+(\mathrm{d}x^2)^2+(\mathrm{d}x^3)^2\end{aligned} \tag{12.4.26}$$

Therefore:

$$g_{\mu\nu}=\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix} \tag{12.4.27}$$

which as usual in this textbook we make usage of the abusive notation (already indicated in the section of Tensor Calculus) to not put $g_{\mu\nu}$ (since a tensor and its matrix form are normally two separate things strictly speaking).

This result is remarkable, because the metric tensor will therefore enable us to define the properties of a space from a simple mathematical being that can easily be handled formally as we already seen in the section of Tensor Calculus, Non-Euclidean Geometry and Differential Geometry.

In polar coordinates the tensor $g_{\mu\nu}$ is:

$$g_{\mu\nu}(r,\phi)=\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}=\begin{bmatrix} 1 & 0 \\ 0 & (x^1)^2 \end{bmatrix}=\begin{bmatrix} 1 & 0 \\ 0 & r^2 \end{bmatrix} \tag{12.4.28}$$

Check:

$$\mathrm{d}s^2=g_{\mu\nu}\mathrm{d}x^\mu\mathrm{d}x^\nu=g_{11}\mathrm{d}x^1x^1+g_{22}\mathrm{d}x^2\mathrm{d}x^2=(\mathrm{d}x^1)^2+(x^1)^2(\mathrm{d}x^2)^2 \tag{12.4.29}$$

And in cylindrical coordinates the tensor $g_{\mu\nu}$ is written:

$$g_{\mu\nu}(r,\phi,z)=\begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{bmatrix}=\begin{bmatrix} 1 & 0 & 0 \\ 0 & (x^1)^2 & 0 \\ 0 & 0 & 1 \end{bmatrix}=\begin{bmatrix} 1 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{12.4.30}$$

We will not do the check as the result is obvious (excepted on reader request).

In spherical coordinates the tensor $g_{\mu\nu}$ is a little more complex and is written:

$$g_{\mu\nu}(r,\theta,\varphi)=\begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{bmatrix}=\begin{bmatrix} 1 & 0 & 0 \\ 0 & (x^1)^2 & 0 \\ 0 & 0 & (x_1\sin(x^2))^2 \end{bmatrix}=\begin{bmatrix} 1 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & r^2\sin^2(\theta) \end{bmatrix} \tag{12.4.31}$$

We will also not do the check as the result is obvious (excepted as always on reader request).

Remark

As we have already mentioned it, in the section of Tensor Calculus at page 2121, we have $g^{ij}=(g_{ij})^{-1}$ and the reader can quickly check this with Maple™ 4.00b as inverting a matrix is always a boring manual work (here the code is given only for the spherical coordinates but the idea is the same for the others):

```
>with(linalg):
>A:=array([[1,0,0],[0,r^2,0],[0,0,r^2*sin(theta)^2]]);
>inverse(A);
```

In Special Relativity, we have seen that the notions of space and time were implicitly bounded. Thus, to study modern physics (this does not really interest the pure mathematician), we need to add to our metric tensor a time component to get what we name the "space-time metric tensor".

To determine the writing of this tensor, we will place us at first in a Minkowski space where we have for recall (see section Special Relativity page 5705):

$$\mathrm{d}s^2=c^2\mathrm{d}t^2-(\mathrm{d}x^2+\mathrm{d}y^2+\mathrm{d}z^2) \tag{12.4.32}$$

which is the infinitesimal interval of space-time between two infinitely close events (or considered as it at a given scale...) named as we know the "interval invariant".

Thus by putting:

$$x^0 = ct, x^1 = x, x^2 = y, x^3 = z \tag{12.4.33}$$

We have:

$$\boxed{g_{\mu\nu} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}} \tag{12.4.34}$$

with the "signature":

$$\text{sign}(g_{\mu\nu}) = (1, -1, -1, -1) = (+, -, -, -) \tag{12.4.35}$$

Remark

For all metric tensor that we have determined before, if we express them in space-time (thus adding time component), the spatial components will all have a negative sign!

The choice of the system of coordinates with which we describe space-time is named a "space-time gauge". If a quantity is unchanged under a space-time gauge transformation, it is say to be as we already know after our study of Quantum Physics "gauge invariant".

We will derive later other metrics that are much less intuitive (like the Schwarzschild metric, the Friedmann–Lemaître–Robertson–Walker metric, the Kerr metric, the De Sitter/anti-De Sitter metrics, the Alcubierre metric and so on...) once we will have proved far later below the Einstein's fields equation.

There is an excellent booklet of 89 pages containing an exhaustive list of 24 metrics (however the Melvin magnetic universe metric is missing) with their respective Christoffel symbols named *Catalogue of Space-times* (see [402]). Here are some of the metrics contained in this booklet:

Name of metric	**Expression**
Minkowski	$\mathrm{d}s^2 = -c^2\mathrm{d}t^2 + \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2$
Schwarzschild	$\mathrm{d}s^2 = -\left(1 - \frac{r_s}{r}\right) c^2\mathrm{d}t^2 + \frac{1}{1 - r_s/r} dr^2 + r^2 \left(\mathrm{d}\vartheta^2 + \sin^2(\vartheta)\mathrm{d}\varphi^2\right)$ $r_s = 2GM/c^2$
Friedman-Robertson-Walker	$\mathrm{d}s^2 = -c^2\mathrm{d}t^2 + R^2 \left\{\frac{\mathrm{d}\eta^2}{1-k\eta^2} + \eta^2 \left(\mathrm{d}\vartheta^2 + \sin^2(\vartheta)\mathrm{d}\varphi^2\right)\right\}$
Alcubierre	$-c^2\mathrm{d}t^2 + \left(\mathrm{d}x - \frac{\mathrm{d}x_s(t)}{\mathrm{d}t} f\left(r_s\right)\mathrm{d}t\right)^2 + \mathrm{d}y^2 + \mathrm{d}z^2$ $v_s = \frac{\mathrm{d}x_s(t)}{\mathrm{d}t}$ $r_s(t) = \sqrt{(x - x_s(t))^2 + y^2 + z^2}$ $f\left(r_s\right) = \frac{\tanh\left(\sigma\left(r_s + R\right)\right) - \tanh\left(\sigma\left(r_s - R\right)\right)}{2\tanh(\sigma R)}$
Barriola-Vilenkin	$\mathrm{d}s^2 = -c^2\mathrm{d}t^2 + \mathrm{d}r^2 + k^2r^2 \left(\mathrm{d}\vartheta^2 + \sin^2(\vartheta)\mathrm{d}\varphi^2\right)$
Bertotti-Kasner	$\mathrm{d}s^2 = -c^2\mathrm{d}t^2 + e^{2\sqrt{\Lambda}ct}\mathrm{d}r^2 + \frac{1}{\Lambda}\left(\mathrm{d}\vartheta^2 + \sin^2(\vartheta)\mathrm{d}\varphi^2\right)$ $\Lambda > 0$
Cosmic string	$\mathrm{d}s^2 = e^{-2U}\left[e^{2K}\left(\mathrm{d}\rho^2 - \mathrm{d}t^2\right) + \rho^2\mathrm{d}\varphi^2\right] + e^{2U}\mathrm{d}z^2$ $U = CJ_0(\rho)\cos(t)$ $K := \frac{1}{2}C^2\rho\left\{\rho\left[J_0(\rho)^2 + J_1(\rho)^2\right] - 2J_0(\rho)J_1(\rho)\cos^2(t)\right\}$ $J_n(x) = \sum_{m=0}^{+\infty}\frac{(-1)^m}{\Gamma(m+1)\Gamma(m+n+1)}\left(\frac{x}{2}\right)^{2m+n}$
Bessel	$\mathrm{d}s^2 = -\left(1 - \frac{r_s}{r}\right) c^2\mathrm{d}t^2 + \frac{1}{1 - r_s/r}\mathrm{d}r^2 + r^2\left(\mathrm{d}\vartheta^2 + \beta^2\sin^2(\vartheta)\mathrm{d}\varphi^2\right)$ $r_s = 2GM/c^2$
Ernst	$\mathrm{d}s^2 = \Lambda^2\left[-\left(1 - \frac{2M}{r}\right)\mathrm{d}t^2 + \frac{\mathrm{d}r^2}{1 - 2M/r} + r^2\mathrm{d}\vartheta^2\right] + \frac{r^2\sin^2(\vartheta)}{\Lambda^2}\mathrm{d}\varphi^2$ $\Lambda = 1 + B^2r^2\sin^2(\vartheta)$
Gödel	$\mathrm{d}s^2 = -c^2\mathrm{d}t^2 + \frac{\mathrm{d}r^2}{1 + [r/(2a)]^2} + r^2\left[1 - \left(\frac{r}{2a}\right)^2\right]\mathrm{d}\varphi^2 + \mathrm{d}z^2 - 2r^2\frac{c}{\sqrt{2}a}\mathrm{d}t\mathrm{d}\varphi$
Halilsoy	$\mathrm{d}s^2 = V\left[e^{2K}\left(\mathrm{d}\rho^2 - \mathrm{d}t^2\right) + \rho^2\mathrm{d}\varphi^2\right] + \frac{1}{V}(\mathrm{d}z + A\mathrm{d}\varphi)^2$ $V = \cosh^2(\alpha)e^{-2CJ_0(\rho)\cos(t)} + \sinh^2(\alpha)e^{2CJ_0(\rho)\cos(t)}$ $K = \frac{C^2}{2}\left[\rho^2\left(J_0(\rho)^2 + J_1(\rho)^2\right) - 2\rho J_0(\rho)J_1(\rho)\cos^2(t)\right]$ $J_n(x) = \sum_{m=0}^{+\infty}\frac{(-1)^m}{\Gamma(m+1)\Gamma(m+n+1)}\left(\frac{x}{2}\right)^{2m+n}$
Janis-Newman-Winicour	$\mathrm{d}s^2 = -\alpha^\gamma c^2\mathrm{d}t^2 + \alpha^{-\gamma}\mathrm{d}r^2 + r^2\alpha^{-\gamma+1}\left(\mathrm{d}\vartheta^2 + \sin^2(\vartheta)\mathrm{d}\varphi^2\right)$ $\alpha = 1 - r_s/(\gamma r)$ $r_s = 2GM/c^2$
Kasner	$\mathrm{d}s^2 = -c^2\mathrm{d}t^2 + t^{2p_1}\mathrm{d}x^2 + t^{2p_2}\mathrm{d}y^2 + t^{2p_3}tz^2$ $p_1 + p_2 + p_3 = 1$ $p_1^2 + p_2^2 + p_3^2 = 1$
Kerr	$\mathrm{d}s^2 = -\left(1 - \frac{r_sr}{\Sigma}\right)c^2\mathrm{d}t^2 - \frac{2r_sar\sin^2(\vartheta)}{\Sigma}c\mathrm{d}t\mathrm{d}\varphi + \frac{\Sigma}{\Delta}\mathrm{d}r^2 + \Sigma\mathrm{d}\vartheta^2$ $+\left(r^2 + a^2 + \frac{r_sa^2r\sin^2(\vartheta)}{\Sigma}\right)\sin^2(\vartheta)\mathrm{d}\varphi^2$ $\Sigma = r^2 + a^2\cos^2(\vartheta)$ $\Delta = r^2 - r_sr + a^2$ $r_s = 2GM/c^2$
...	...

Table 12.22: **Some famous metrics (source: Catalogue of Space-times [402])**

Most of these metrics were not derived. They are just written down and then their properties are analysed. Sometimes it is named an "ansatz", which is basically just an educated guess that we make and then work backwards to show that the guess was good (there is a very good example of such an intellectual process in [208]).

12.4.2.1 Schild Criteria (Einstein redshift effect Newtonian approach)

We will prove later that gravitation as formulated in Newtonian mechanics is completely describable by a curvature formulation of space-time. But first we want to introduce to the reader with, in our point of view, the easiest development that can be done without General Relativity to compare it also later with the easiest development that can be done with General Relativity (see page 5796) the gravitational redshift effect!

Imagine first a very height tower of height h built on the surface of the Earth. A man sits at the ground of the tower, and sends a signal of pulsation ω_A to a colleague B at the top of the tower. There will be, and we will immediately prove it, that the pulsation ω_B of the wave received by B differs of ω_A according to the relation:

$$\frac{\omega_A}{\omega_B}=1+\frac{gh}{c^2} \tag{12.4.36}$$

Hence:

$$\Delta\omega=\omega_A-\omega_B>0 \tag{12.4.37}$$

This shift of pulsation (frequencies respectively) in a gravitational field is what we name the "Einstein's effect", or "gravitational redshift".

We will first prove this relation using conventional arguments and now well known to us. Later we will prove that in fact this is only an approximation of a result that we will get later using General Relativity curvature properties.

A material body sent from the ground to the sky must fight against the gravitational force that pulls[43] it down. So it will lose a certain amount of energy, equivalent to gravitational potential energy gained during the trip. The total energy E_A of the body at the ground level is therefore its mass energy (see section Special Relativity page 5679) to which we add the potential energy at the height of the tower:

$$E_A=mc^2+mgh \tag{12.4.38}$$

The energy of this body when you reach the top of the tower is simply its mass energy:

$$E_B=mc^2 \tag{12.4.39}$$

because he had to spend the energy mgh during the trip to go up. The ratio of energy is then:

$$\frac{E_A}{E_B}=\frac{mc^2+mgh}{mc^2}=1+\frac{gh}{c^2} \tag{12.4.40}$$

---

[43] Keep in mind that's in the framework of the Newtonian model of gravity as we that in General Relativity there are no "pull".

This ratio being independent of the mass m, we can take the limit $m \to 0$ in order to have the relation for the photon. We then get:

$$\frac{E_A}{E_B}=\frac{\hbar\omega_A}{\hbar\omega_B}=\frac{\omega_A}{\omega_B}=\frac{f_A}{f_B}=\frac{T_B}{T_A}=1+\frac{gh}{c^2} \tag{12.4.41}$$

which implies:

$$T_B > T_A \tag{12.4.42}$$

That is to say the clocks run slower in a gravitational field as seen by a distant observer!

We will now study this phenomenon in the context of the Minkowski space-time. We will see then a contradiction, what will motivate the transition to a curved space-time: this is the argument of a curved geometry that was used by Schild.

Let us consider again the human experience of a human in A which sends a wave to his friend positioned in B. Given Δt_A the time taken by A to emit exactly 1 cycle of the wave (see section Wave Mechanics page 3480):

$$\Delta t_A=\frac{\lambda_A}{c}=\frac{1}{\omega_A c} \tag{12.4.43}$$

and Δt_B the time taken for B to receive this cycle:

$$\Delta t_A=\frac{\lambda_B}{c}=\frac{1}{\omega_B c} \tag{12.4.44}$$

Because of the Einstein's effect just seen previously, we know that $\omega_A > \omega_B$ and therefore that $\Delta t_A < \Delta t_B$ in proper time! That is to say that time passes more slowly for someone on the ground (A) than another person in a mountain top (B)!

But as we are in flat geometry and the gravitational field is assumed static, we deduce that space-time trajectories described by the signals must be parallel! This leads to the conclusion that the proper time interval would be $\Delta t_A = \Delta t_B$ (according to Special Relativity).

If we opt for a curved space, we can preserve the relation $\Delta t_A < \Delta_B$, that is to say that time passes more slowly for A than for B. This simply results in the fact that in curved geometry, the proper time (!) of an observer depends on the metric.

Let us now notice that same developments can be made by assimilating the previous experience with a train that moves with constant acceleration g (horizontal situation of the previous one!). The observer A is in the rear compartment (equivalent to the floor of the Earth in the preceding experiment) sends a wave to his colleague B on the front of the train (at a distance h).

The observer B receives the wave after a time $\Delta t = h/c$. During this time, the train has accelerated, and its speed has increased of a value $\Delta v = g\Delta t = gh/c$. Therefore, the wave seen by B will be altered by the conventional Doppler effect (see section Music Mathematics page 8212):

$$\frac{\Delta\lambda}{\lambda_A}=\frac{\Delta v\Delta t}{c\Delta t}=\frac{\Delta v}{c}=\frac{gh}{c^2} \tag{12.4.45}$$

We fall back on the initial results of the Einstein's effect by simply writing:

$$z:=\frac{\Delta\lambda}{\lambda_A}=\frac{\lambda_B-\lambda_A}{\lambda_A}=\frac{\lambda_B}{\lambda_A}-1=\frac{\omega_B}{\omega_A}-1 \tag{12.4.46}$$

giving gloriously:

$$\left.\begin{matrix}\dfrac{\Delta\lambda}{\lambda_A}=\dfrac{gh}{c^2}\\ \dfrac{\Delta\lambda}{\lambda_A}=\dfrac{\omega_B}{\omega_A}-1\end{matrix}\right\}:\frac{\omega_B}{\omega_A}=1+\frac{gh}{c^2} \tag{12.4.47}$$

We find more often this relation in the form below in the literature using the relation between pulsation and frequency and Newton's gravitational force to explicit g:

$$\frac{\omega_B}{\omega_A}=1+\frac{GMh}{r^2c^2} \tag{12.4.48}$$

and putting h as being equal to r:

$$\boxed{\frac{\omega_B}{\omega_A}=1+\frac{GM}{rc^2}=1+\varphi} \tag{12.4.49}$$

and after rearranging we also found sometimes in textbooks:

$$\frac{f_B-f_A}{f_A}=\frac{GM}{rc^2} \tag{12.4.50}$$

or even more frequently:

We also find this last relation in the following condensed form:

$$\frac{\Delta f}{f_0}=\frac{\Delta\lambda}{\lambda_0}=\frac{a_g}{c^2} \tag{12.4.51}$$

The same result can be obtained using the Schwarzschild metric (see further below), hence the name of this effect that can also be obtained from the mathematical tools of Einstein's General Relativity. We will prove later, in a simple way, using this metric that time actually flows more slowly in a gravitational field (assumption we made a few paragraphs above).

We see that in all cases:

$$f_B>f_A \tag{12.4.52}$$

since the right term is positive and not zero. This simply means that the electromagnetic wave in analogy to the color spectrum shifts toward red. Thus, Einstein's effect is indeed a gravitational redshift!

The frequency difference is very small and therefore difficult to measure even with the best spectroscopes. The slightest disturbance can completely mask the Einstein's effect. It will be necessary to wait until 11960 (holocene calendar) that the experience of Robert Pound and his graduate student Glen A. Rebka Jr. to be capable of measuring a frequency offset with an accuracy of 1% therefore leaving no doubt as to the reality of this phenomenon.

12.4.3 Equations of movement

We will prove here that the equation of motion of a free particle is constant along its world line by first limiting ourselves to the case of a flat space (Minkowski space type). Then we will generalize this result in any kind of space using a simple development, to show quite clearly that the equation of motion is independent of the mass and follows the curvature of space !!! Finally, we will present a second proof in any kind of space using the variational principle.

So let us start by proving the equation of motion of a free particle in a flat space.

During our study of Special Relativity, we proved that the Relativistic Lagrangian of a free particle was given by (see section Special Relativity page 5683):

$$L=-m_0c^2\sqrt{1-\frac{v^2}{c^2}} \tag{12.4.53}$$

and for this we started from the action (hypothetical):

$$S_0=-m_0c^2\int \mathrm{d}t' \tag{12.4.54}$$

and we came to write:

$$S_0=-m_0c^2\int\sqrt{1-\frac{v^2}{c^2}}\mathrm{d}t \tag{12.4.55}$$

Now let us show something interesting! Let us recall that for the Minkowski space-time, we got:

$$\mathrm{d}s^2=c^2\mathrm{d}t^2-(\mathrm{d}x^2+\mathrm{d}y^2+\mathrm{d}z^2) \tag{12.4.56}$$

and restricting ourselves to one spatial dimension, we get as relation:

$$\mathrm{d}s^2=c^2\mathrm{d}t-\mathrm{d}x^2\Leftrightarrow \mathrm{d}s^2=c^2\mathrm{d}t^2\left(1-\frac{\mathrm{d}x^2}{c^2\mathrm{d}t^2}\right)=c^2\mathrm{d}t^2\left(1-\frac{v^2}{c^2}\right) \tag{12.4.57}$$

Therefore:

$$\mathrm{d}s=c\mathrm{d}t\sqrt{1-\frac{v^2}{c^2}} \tag{12.4.58}$$

and then ... well that's the way, if we put:

$$\boxed{S=-m_0c\int \mathrm{d}s} \tag{12.4.59}$$

we finally have:

$$S=-m_0c\int \mathrm{d}s=-m_0c\int c\mathrm{d}t\sqrt{1-\frac{v^2}{c^2}}=-m_0c^2\int\sqrt{1-\frac{v^2}{c^2}}\mathrm{d}t \tag{12.4.60}$$

so we fall back on the same action from a more general form (pure) action that is:

$$S=-m_0c\int \mathrm{d}s \tag{12.4.61}$$

result that we had also proved in the section of Electrodynamics (see page 4126)!! We can even do better in terms of elegance ...! If we observe well the developments of the previous lines, we observe that in facts the relation:

$$\mathrm{d}s=c\mathrm{d}t\sqrt{1-\frac{v^2}{c^2}} \tag{12.4.62}$$

And this last relation is the one-dimensional case of the more general following relation:

$$\mathrm{d}s=\mathrm{d}t\sqrt{\frac{g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu}{c^2}}=\frac{1}{c}\sqrt{g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu}\mathrm{d}t \tag{12.4.63}$$

with as defined earlier above:

$$x^0=ct,x^1=x,x^2=y,x^3=z \tag{12.4.64}$$

and therefore:

$$\dot{x}^0=c,\dot{x}^1=v_x,\dot{x}^2=v_y,\dot{x}^3=v_z \tag{12.4.65}$$

Thus we have the "Fitzgerald-Lorentz factor" or simply "Lorentz factor" that is given in general form by:

$$\boxed{\gamma=\frac{1}{c}\sqrt{g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu}} \tag{12.4.66}$$

as a generalization of Special Relativity!

Therefore the Lagrangian of a relativistic free particle is in general written:

$$L=-m_0c^2\sqrt{\eta_{\mu\nu}\frac{\mathrm{d}x^\mu}{\mathrm{d}t}\frac{\mathrm{d}x^\mu}{\mathrm{d}t}}=-m_0c^2\sqrt{\eta_{\mu\nu}\dot{x}^\mu\dot{x}^\nu} \tag{12.4.67}$$

Remark

Notice, that the Lagrangian above is a homogeneous function of degree one, i.e.:

$$L(\lambda\dot{x}^\mu(t))=\lambda L(\dot{x}^\mu(t)) \tag{12.4.68}$$

Therefore as seen in the section of Analytical Mechanics (see page 3181), the Hamiltonian is zero!

This being done, let us come back on our topic... In a space without potential field, we have proved in the section of Analytical Mechanics that the Lagrangian is reduced to the simplest expression of the kinetic energy (see page 3170) such that:

$$L=\frac{1}{2}m_0v^2=\frac{1}{2}m_0\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2 \tag{12.4.69}$$

If we wish to generalize this relation for it to be valid in any type of space (curved or flat), we must introduce the curvilinear coordinates as we have studied them in the section of Tensor Calculus (see page 2092).

In a first time, this gives:

$$L=\frac{1}{2}m_0\left(\frac{\mathrm{d}s}{\mathrm{d}t}\right)^2 \tag{12.4.70}$$

where for recall $\mathrm{d}s$ is the curvilinear abscissa of the path.

And we have proved in the section of Differential Geometry that (see page 2850):

$$\mathrm{d}s=\left[g_{ij}\frac{\partial u^i}{\partial\alpha}\frac{\partial u^j}{\partial\alpha}\right]^{1/2}\mathrm{d}\alpha \tag{12.4.71}$$

The latter relation is written in the context of relativistic mechanics in a most standard way:

$$\mathrm{d}s=\left[g_{\alpha\beta}\frac{\partial x^\alpha}{\partial\tau}\frac{\partial x^\beta}{\partial\tau}\right]^{1/2}\mathrm{d}\tau \tag{12.4.72}$$

where τ is a parameter that in relativistic mechanics if for recall the proper time of the particle.

Before we focus on curved spaces described by the metric $g_{\alpha\beta}$ (which we will do during our proof of the free generalized Lagrangian), let us restrict us to Euclidean space with the metric (this will be a good exercise to understand) given by the Minkowski matrix (see section Special Relativity page 5705):

$$g_{\alpha\beta}=\begin{bmatrix}1&0&0&0\\0&-1&0&0\\0&0&-1&0\\0&0&0&-1\end{bmatrix} \tag{12.4.73}$$

which we denote $\eta_{\alpha\beta}$ to differentiate it from others (because most often used). Finally we in Euclidean space:

$$S=-mc\int\mathrm{d}s=-mc\int\left(\eta_{\alpha\beta}\frac{\partial x^\alpha}{\partial\tau}\frac{\partial x^\beta}{\partial\tau}\right)\mathrm{d}\tau \tag{12.4.74}$$

Now let us apply the variational principle:

$$\delta S=-mc\int\delta(\mathrm{d}s) \tag{12.4.75}$$

The variation $\mathrm{d}s$ can be found simply from the variation of $\mathrm{d}s^2$:

$$(\mathrm{d}s)^2=\mathrm{d}s^2=\eta_{\alpha\beta}\frac{\partial x^\alpha}{\partial t}\frac{\partial x^\beta}{\partial t}\mathrm{d}t^2=\eta_{\alpha\beta}\mathrm{d}x^\alpha\mathrm{d}x^\beta \tag{12.4.76}$$

we find:

$$\delta(\mathrm{d}s)^2=2\mathrm{d}s\delta(\mathrm{d}s)=\eta_{\alpha\beta}\delta(\mathrm{d}x^\alpha\mathrm{d}x^\beta)=\eta_{\alpha\beta}\delta\mathrm{d}x^\alpha\mathrm{d}x^\beta+\eta_{\alpha\beta}\mathrm{d}x^2\delta\mathrm{d}x^\beta=2\eta_{\alpha\beta}\delta(\mathrm{d}x^\alpha)\mathrm{d}x^\beta \tag{12.4.77}$$

The factor "2" is here because by symmetry of the Euclidean space, the variations of $\mathrm{d}x^\alpha$ and $\mathrm{d}x^\beta$ are equal.

Remark

As we will see later, this relation $\delta(\mathrm{d}s)^2$ will not be the same anymore when dealing with curved spaces.

Simplifying a bit, we get:

$$2\mathrm{d}s\delta(\mathrm{d}s)=2\eta_{\alpha\beta}\delta(\mathrm{d}x^{\alpha})\mathrm{d}x^{\beta}\Leftrightarrow\delta(\mathrm{d}s)=\eta_{\alpha\beta}\delta(\mathrm{d}x^{\alpha})\frac{\mathrm{d}x^{\beta}}{\mathrm{d}s} \qquad (12.4.78)$$

Which is equivalent to write:

$$\delta(\mathrm{d}s)=\eta_{\alpha\beta}\mathrm{d}(\delta x^{\alpha})\frac{\mathrm{d}x^{\beta}}{\mathrm{d}s}=\eta_{\alpha\beta}\frac{\mathrm{d}(\delta x^{\alpha})}{\mathrm{d}\tau}\frac{\mathrm{d}x^{\beta}}{\mathrm{d}s}\mathrm{d}\tau \qquad (12.4.79)$$

We can now go back to the action:

$$\delta S=-mc\int\delta(\mathrm{d}s)=-mc\int\eta_{\alpha\beta}\frac{\mathrm{d}(\delta x^{\alpha})}{\mathrm{d}\tau}\frac{\mathrm{d}x^{\beta}}{\mathrm{d}s}\mathrm{d}\tau \qquad (12.4.80)$$

We rewrite the preceding integral as following (it will be easier to treat):

$$\delta S=-mc\int\eta_{\alpha\beta}\frac{\mathrm{d}(\delta x^{\alpha})}{\mathrm{d}\tau}\frac{\mathrm{d}x^{\beta}}{\mathrm{d}s}\mathrm{d}\tau=-mc\int_{\tau_2}^{\tau_1}\mathrm{d}\tau\frac{\mathrm{d}}{\mathrm{d}\tau}\left(\eta_{\alpha\beta}\delta x^{\alpha}\frac{\mathrm{d}x^{\beta}}{\mathrm{d}s}\right)+\int_{\tau_2}^{\tau_1}\mathrm{d}\tau\delta x^{\alpha}\left(mc\eta_{\alpha\beta}\frac{\mathrm{d}}{\mathrm{d}\tau}\frac{\mathrm{d}x^{\beta}}{\mathrm{d}s}\right) \qquad (12.4.81)$$

Indeed, let check that this form is similar:

$$\begin{aligned}\delta S&=-mc\int_{\tau_2}^{\tau_1}\mathrm{d}\tau\frac{\mathrm{d}}{\mathrm{d}\tau}\left(\eta_{\alpha\beta}\delta x^{\alpha}\frac{\mathrm{d}x^{\beta}}{\mathrm{d}s}\right)+\int_{\tau_2}^{\tau_1}\mathrm{d}\tau\delta x^{\alpha}\left(mc\eta_{\alpha\beta}\frac{\mathrm{d}}{\mathrm{d}\tau}\frac{\mathrm{d}x^{\beta}}{\mathrm{d}s}\right)\\&=-mc\cancel{\int_{\tau_2}^{\tau_1}\mathrm{d}\tau\left(\frac{\eta_{\alpha\beta}}{\underbrace{\mathrm{d}\tau}_{=0}}\delta x^{\alpha}\frac{\mathrm{d}x^{\beta}}{\mathrm{d}s}\right)}-mc\int_{\tau_2}^{\tau_1}\mathrm{d}t\left(\eta_{\alpha\beta}\frac{\mathrm{d}(\delta x^{\beta})}{\mathrm{d}\tau}\frac{\mathrm{d}x^{\beta}}{\mathrm{d}s}\right)-mc\cancel{\int_{\tau_2}^{\tau_1}\mathrm{d}t\left(\eta_{\alpha\beta}\delta x^{\alpha}\frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathrm{d}x^{\beta}}{\mathrm{d}s}\right)}\\&+\cancel{\int_{\tau_2}^{\tau_1}\mathrm{d}\tau\delta x^{\alpha}\left(mc\eta_{\alpha\beta}\frac{\mathrm{d}}{\mathrm{d}\tau}\frac{\mathrm{d}x^{\beta}}{\mathrm{d}s}\right)}=-mc\int_{\tau_2}^{\tau_1}\eta_{\alpha\beta}\frac{\mathrm{d}(\delta x^{\alpha})}{\mathrm{d}s}\mathrm{d}t\end{aligned} \qquad (12.4.82)$$

So let us come back to our integral:

$$\delta S=-mc\int_{\tau_2}^{\tau_1}\mathrm{d}\tau\frac{\mathrm{d}}{\mathrm{d}\tau}\left(\eta_{\alpha\beta}\delta x^{\alpha}\frac{\mathrm{d}x^{\beta}}{\mathrm{d}s}\right)+\int_{\tau_2}^{\tau_1}\mathrm{d}\tau\delta x^{\alpha}\left(mc\eta_{\alpha\beta}\frac{\mathrm{d}}{\mathrm{d}\tau}\frac{\mathrm{d}x^{\beta}}{\mathrm{d}s}\right) \qquad (12.4.83)$$

We have then two integrals that it will be a bit easier to analyse. The first integral:

$$-mc\int_{\tau_2}^{\tau_1}\mathrm{d}\left(\eta_{\alpha\beta}\delta x^{\alpha}\frac{\mathrm{d}x^{\beta}}{\mathrm{d}s}\right)=-m\left[\eta_{\alpha\beta}\delta x^{\alpha}\frac{\mathrm{d}x^{\beta}}{\mathrm{d}s}\right]_{\tau_2}^{\tau_1} \qquad (12.4.84)$$

simply gives an expression evaluated to the temporal extremities (τ_1, τ_2). Therefore, as the values x^α are perfectly known at the time ends, the variational δx^α is zero at the both extremities and this integral is therefore zero.

Then we are left only with this integral:

$$\delta S = \int_{\tau_2}^{\tau_1} \mathrm{d}\tau \delta x^\alpha \left(mc\eta_{\alpha\beta} \frac{\mathrm{d}}{\mathrm{d}\tau} \frac{\mathrm{d}x^\beta}{\mathrm{d}s} \right) \tag{12.4.85}$$

So for the variational principle (see section Analytical Mechanics page 3131):

$$\delta S = 0 \tag{12.4.86}$$

is respected, we must have:

$$\frac{\mathrm{d}}{\mathrm{d}\tau} \left(mc \frac{\mathrm{d}x^\beta}{\mathrm{d}s} \right) = 0 \tag{12.4.87}$$

Now, we can write this expression explicitly. Indeed, we have:

$$mc \frac{\mathrm{d}x^\beta}{\mathrm{d}s} = mc \left(\frac{\mathrm{d}(c\tau)}{\mathrm{d}s}, \frac{\mathrm{d}x}{\mathrm{d}s}, \frac{\mathrm{d}y}{\mathrm{d}s}, \frac{\mathrm{d}z}{\mathrm{d}s} \right) \tag{12.4.88}$$

Remember also that we have proved earlier above that:

$$\mathrm{d}s = c\mathrm{d}t\sqrt{1-\beta^2} \Leftrightarrow \frac{\mathrm{d}\tau}{\mathrm{d}s} = \frac{1}{c\sqrt{1-\beta^2}} = \frac{\gamma}{c} \tag{12.4.89}$$

and that we have:

$$\frac{\mathrm{d}x}{\mathrm{d}s} = \frac{\mathrm{d}x}{\mathrm{d}\tau} \frac{\mathrm{d}\tau}{\mathrm{d}s} = v_x \frac{\mathrm{d}\tau}{\mathrm{d}s} = \frac{v_x \gamma}{c} \tag{12.4.90}$$

Therefore:

$$mc \left(\frac{\mathrm{d}(c\tau)}{\mathrm{d}s}, \frac{\mathrm{d}x}{\mathrm{d}s}, \frac{\mathrm{d}y}{\mathrm{d}s}, \frac{\mathrm{d}z}{\mathrm{d}s} \right) = m\gamma \left(c, v_x, v_y, v_z \right) \tag{12.4.91}$$

Now, let us recall that during our study of Special Relativity (see page 5687), we have proved the path that led us to define the linear momentum four-vector:

$$p^\nu = \left(\frac{E}{c}, p_x, p_y, p_z \right) = \left(\frac{E}{c}, \vec{p} \right) = m\gamma(c, \vec{v}) \tag{12.4.92}$$

So finally, what cancel the variational of the action integral can be written:

$$\boxed{\frac{\mathrm{d}p^\nu}{\mathrm{d}\tau} = 0} \tag{12.4.93}$$

We thus fall back on the conservation equation of linear momentum (momentum conservation) that we name in the framework of General Relativity "equation of motion". This form of the equation of motion seems dependent on the mass but by digging a bit, we will see that it is fact not.

Multiplying this relation by $\eta_{\mu\nu}$ we can also write:

$$\boxed{\frac{\mathrm{d}p_\mu}{\mathrm{d}\tau}=0} \tag{12.4.94}$$

and the same for another observer:

$$\frac{\mathrm{d}p_\mu}{\mathrm{d}\tau\gamma}=0 \Leftrightarrow \frac{\mathrm{d}p_\mu}{\mathrm{d}\tau'}=0 \tag{12.4.95}$$

In other words, the linear momentum of the particle remains constant along its world line.

But we can also write:

$$\frac{\mathrm{d}p^v}{\mathrm{d}\tau}=\frac{\mathrm{d}p^v}{\left(\dfrac{\mathrm{d}s}{c}\gamma\right)}=0 \tag{12.4.96}$$

Therefore:

$$\boxed{\frac{\mathrm{d}p^v}{\mathrm{d}s}=0} \tag{12.4.97}$$

An even more important form of movement equation can be obtained. Indeed using the relations just proved above we can write:

$$p^v=mc\frac{mc}{\dfrac{\mathrm{d}x^v}{\mathrm{d}s}} \tag{12.4.98}$$

Therefore:

$$\frac{\mathrm{d}p^v}{\mathrm{d}s}=\frac{\mathrm{d}}{\mathrm{d}s}\left(mc\frac{\mathrm{d}x^v}{\mathrm{d}s}\right)=mc\frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{\mathrm{d}x^v}{\mathrm{d}s}\right)=0 \tag{12.4.99}$$

Hence:

$$\boxed{\frac{\mathrm{d}^2x^v}{\mathrm{d}s^2}=0} \tag{12.4.100}$$

This relation is therefore thhe "massless" equation of motion in Euclidean space or in other words, in a Minkowski space-time type! That means that there exists a falling coordinate system wherein the motion of the particle is in an uniform movement in space-time.

It will be very interesting to compare it later with the equation of motion in a curved space as we will see later (named "geodesic equation").

Remark

It is equivalent to write the relations of equations of motion with respect to the curvilinear abscissa $\mathrm{d}s$ or the proper time $\mathrm{d}t$ (traditionally denoted by $\mathrm{d}\tau$ in the field of General Relativity).

We can now prove that the previous equation of motion, just like the geodesic equation that we will see afterwards, is invariant under Lorentz transformation. Indeed:

$$L^{\mu}_{v}\frac{\mathrm{d}^2 x^v}{\mathrm{d}s^2}=\frac{\mathrm{d}^2(L^{\mu}_{v}x^v)}{\mathrm{d}s^2}=\frac{\mathrm{d}^2 x'^v}{\mathrm{d}s^2}=0 \tag{12.4.101}$$

Now let us see a more general form of the equation of motion for any kind of space. The aim is to highlight, and this in a few lines of calculations, that the movement followed by a free particle is independent of its mass (you can already anticipate the interpretation of the path of a photon in a curved space...!).

Let us first recall that we have proved in the section of Tensor Calculus (and previously) that:

$$\mathrm{d}s=\left[g_{ij}\frac{\partial u^i}{\partial\alpha}\frac{\partial u^j}{\partial\alpha}\right]^{1/2}\mathrm{d}\alpha \tag{12.4.102}$$

giving us for the generalized Lagrangian of a free particle with $\mathrm{d}\alpha=\mathrm{d}\tau$, $u^i=x^\alpha$, $u^j=x^\beta$ (although we fall back on the general expression of the kinetic energy as there is no potential for a free particle):

$$\boxed{L=\frac{1}{2}mg_{\alpha\beta}\frac{\partial x^\alpha}{\partial\tau}\frac{\partial x^\beta}{\partial\tau}} \tag{12.4.103}$$

where for recall τ is the proper time[44] of the particle, it is an invariant!

Remark

This relation is named the "geodesic lagrangian" by some textbooks authors.

This allows us to write (caution! the reader must remember the different relations that we had determined during our study of the Lagrangian formalism in the section dealing with Analytical Mechanics):

$$p_\alpha=\frac{\partial L}{\partial\dot{x}^\alpha}=mg_{\alpha\beta}\frac{\mathrm{d}x^\beta}{\mathrm{d}\tau} \tag{12.4.104}$$

Remark

The elimination of the $1/2$ Lagrangian factor results from the symmetry of the metric tensor. If that latter is not symmetric, we can always characterize it by a tensor that is.

Indeed, for recall (see section Tensor Calculus page 2088), given $\vec{x}$ a vector of coordinates $x_1,\ldots,x_n$ and given:

$$q=A_{ij}x_ix_j \tag{12.4.105}$$

The A_{ij} are not symmetric a priori, but we can write:

$$q=\frac{1}{2}(A_{ij}+A_{ji})x_ix_j \tag{12.4.106}$$

---

[44] The proper time is for recall a kind of imaginary clock that travels on the particle and whatever observers watch the clock, they will mathematically agree on the value of the time interval between two "Tic" of the clock.

We put afterwards:

$$B_{ij}=\frac{1}{2}(A_{ij}+A_{ji})x_ix_j \tag{12.4.107}$$

Therefore:

$$q=B_{ij}x_ix_j \tag{12.4.108}$$

And the B_{ij} are symmetric.

The quadratic form q can thus always be written with a symmetric matrix, there is even a bijection. The conclusion is that a metric tensor must be symmetric if we want to characterize it by the quadratic form it defines.

The mathematical interlude having ended, let us continue our physical development. As a consequence of the last relation, the expression of the Hamiltonian obviously becomes:

$$H=p_\alpha\dot{x}^\alpha-L=L \tag{12.4.109}$$

since we consider to be in a space without potential field anymore. Since the square of the velocity is therefore constant over the entire trajectory, we have:

$$\frac{\mathrm{d}H}{\mathrm{d}t}=\frac{\mathrm{d}L}{\mathrm{d}t}=0 \tag{12.4.110}$$

Let us now establish the equations of motion of any body. We have:

$$\frac{\partial L}{\partial x^\nu}=\frac{1}{2}m\frac{\partial g_{\alpha\beta}}{\partial x^\nu}\dot{x}^\alpha\dot{x}^\beta \qquad \text{and} \qquad \frac{\partial L}{\partial \dot{x}^\nu}=mg_{\nu\beta}\dot{x}^\beta \tag{12.4.111}$$

and as:

$$\frac{\partial L}{\partial x^\nu}=\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{x}^\nu}=\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial\frac{\mathrm{d}}{\mathrm{d}t}(x^\nu)} \tag{12.4.112}$$

then:

$$\frac{1}{2}m\frac{\partial g_{\alpha\beta}}{\partial x^\nu}\dot{x}^\alpha\dot{x}^\beta=\frac{\mathrm{d}}{\mathrm{d}t}\left(mg_{\nu\beta}\dot{x}^\beta\right) \tag{12.4.113}$$

hence:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{x}^\nu}\right)=m\frac{\mathrm{d}}{\mathrm{d}t}\left(g_{\nu\beta}\dot{x}^\beta\right)=mg_{\nu\beta}\ddot{x}^\beta+m\frac{\mathrm{d}g_{\nu\beta}}{\mathrm{d}t}\frac{\partial x^\alpha}{\partial x^\alpha}\dot{x}^\beta=mg_{\nu\beta}\ddot{x}^\beta+m\frac{\partial g_{\nu\beta}}{\partial x^\alpha}\dot{x}^\alpha\dot{x}^\beta \tag{12.4.114}$$

By putting everything together we get:

$$mg_{\nu\beta}\ddot{x}^\beta+m\frac{\partial g_{\nu\beta}}{\partial x^\alpha}\dot{x}^\alpha\dot{x}^\beta=\frac{1}{2}m\frac{\partial g_{\alpha\beta}}{\partial x^\nu}\dot{x}^\alpha\dot{x}^\beta \tag{12.4.115}$$

that we can write identically for the $\ddot{x}^\alpha$ by proceeding in the same way as above.

The preceding relation therefore gives the trajectory of a body in motion, in a space without a potential field, as a function of its curvilinear coordinates and of the metric of the space under consideration.

What is particularly interesting in this result is that the mass m is (again!) eliminated identically in this equation of motion:

$$g_{\nu\beta}\ddot{x}^{\beta}+\left(\frac{\partial g_{\nu\beta}}{\partial x^{\alpha}}-\frac{1}{2}\frac{\partial g_{\alpha\beta}}{\partial x^{\nu}}\right)\dot{x}^{\alpha}\dot{x}^{\beta}=0 \qquad (12.4.116)$$

Notice that we could have used another invariant parameter as well as the proper time τ such as the curvilinear abscissa $\mathrm{d}s$. Hence the preceding equation should be written:

$$\boxed{g_{\nu\beta}\frac{\partial^2 x^{\beta}}{\partial s^2}+\left(\frac{\partial g_{\nu\beta}}{\partial x^{\beta}}-\frac{1}{2}\frac{\partial g_{\alpha\beta}}{\partial x^{\nu}}\right)\frac{\partial x^{\alpha}}{\partial s}\frac{\partial x^{\beta}}{\partial s}=0} \qquad (12.4.117)$$

We can still simplify this relation, but we will keep this simplification for the second proof of the equation of motion in any space (by making use of the variational principle this time) just further below.

It is very (very) interesting to observe that if we restrict the metric to that of a Euclidean space:

$$\eta_{\nu\beta}\frac{\partial^2 x^{\beta}}{\partial s^2}+\left(\frac{\partial \eta_{\nu\beta}}{\partial x^{\beta}}-\frac{1}{2}\frac{\partial \eta_{\alpha\beta}}{\partial x^{\nu}}\right)\frac{\partial x^{\alpha}}{\partial s}\frac{\partial x^{\beta}}{\partial s}=0 \qquad (12.4.118)$$

with:

$$\frac{\partial \eta_{\nu\beta}}{\partial x^{\alpha}}=\frac{\partial \eta_{\alpha\beta}}{\partial x^{\nu}}=0 \qquad (12.4.119)$$

We then have the following simplification:

$$\eta_{\nu\beta}\frac{\partial^2 x^{\beta}}{\partial s^2}+\left(0-\frac{1}{2}\cdot 0\right)\frac{\partial x^{\alpha}}{\partial s}\frac{\partial x^{\beta}}{\partial s}=0 \qquad (12.4.120)$$

That it remains only:

$$\eta_{\nu\beta}\frac{\partial^2 x^{\beta}}{\partial s^2}=0 \qquad (12.4.121)$$

By lowering the indices with the signature it remains:

$$\frac{\partial^2 x_{\nu}}{\partial s^2}=0 \qquad (12.4.122)$$

We thus fall back on the first equation of the motion obtained for a flat space! The result is remarkable!

The conclusions is that at the same initial conditions of curvilinear position and velocity in a space (flat or curved) without a potential field (this is what we could think at least according to our initial hypotheses ...), corresponds the same trajectory whatever the mass m of the particle (even for photons - light - whose rest mass is zero!).

We can now study the principle of least action in order to seek the shortest path (both spatially and temporally!) between two points in a given geometric space before addressing the much more complex case of the Lagrangian which takes into account the tensor field...

12.4.3.1 Geodesic equations

Let us now turn to the same result, but this time using the variational principle. We will fall on the same equation as before for any kind of space with the difference that this time we will take the time to simplify it to arrive at the "geodesic equation".

Starting from (see previous developments):

$$\mathrm{d}s=(g_{ij}\mathrm{d}x^i\mathrm{d}x^j)^{1/2} \tag{12.4.123}$$

with a parametrization such that x^i and x^j depend of a temporal or spatial parameter.

For a given surface in parametric form, we therefore seek to minimize the length of an arc $\mathrm{d}s$ by applying the variational principle (not dependent on time) because the photons can not have a faster path in the temporal sense of the term between two points but only a shorter path - in the metric sense of the term!):

$$\delta S=\int\limits_a^b \mathrm{d}(\mathrm{d}s)=\int\limits_a^b \delta(\mathrm{d}s)=0 \tag{12.4.124}$$

in natural units. Or:

$$\delta(\mathrm{d}s^2)=2\mathrm{d}s\delta\mathrm{d}s=\delta(g_{ij}\mathrm{d}x^i\mathrm{d}x^j) \tag{12.4.125}$$

By developing, and as the indices have the same range of variation:

$$\begin{aligned}\delta(g_{ij}\mathrm{d}x^i\mathrm{d}x^j)&=\left(\frac{\partial g_{ij}}{\partial x^k}\delta\mathrm{d}x^k\right)\mathrm{d}x^i\mathrm{d}x^j+g_{ij}\mathrm{d}x^j\delta\mathrm{d}x^i+g_{ij}\mathrm{d}x^i\delta\mathrm{d}x^j\\&=\left(\frac{\partial g_{ij}}{\partial x^k}\delta\mathrm{d}x^k\right)\mathrm{d}x^i\mathrm{d}x^j+2g_{ij}\mathrm{d}x^i\delta\mathrm{d}x^j\end{aligned} \tag{12.4.126}$$

hence (we have already multiplied the expression after the second equality by $\mathrm{d}s/\mathrm{d}s$ by anticipating the integral that follows):

$$\delta(\mathrm{d}s)=\frac{\delta(\mathrm{d}s^2)}{2\mathrm{d}s}=\frac{1}{2}\frac{\mathrm{d}x^i}{\mathrm{d}s}\frac{\mathrm{d}x^j}{\mathrm{d}s}\left(\frac{\partial g_{ij}}{\partial x^k}\delta\mathrm{d}x^k\right)\mathrm{d}s+g_{ij}\frac{\mathrm{d}x^i}{\mathrm{d}s}\frac{\delta\mathrm{d}x^j}{\mathrm{d}s}\mathrm{d}s \tag{12.4.127}$$

Then, we must introduce this development under the integral:

$$\int\limits_a^b\delta(\mathrm{d}s)=\int\limits_a^b\frac{1}{2}\frac{\mathrm{d}x^i}{\mathrm{d}s}\frac{\mathrm{d}x^j}{\mathrm{d}s}\left(\frac{\partial g_{ij}}{\partial x^k}\delta\mathrm{d}x^k\right)\mathrm{d}s+\int\limits_a^b g_{ij}\frac{\mathrm{d}x^i}{\mathrm{d}s}\frac{\delta\mathrm{d}x^j}{\mathrm{d}s}\mathrm{d}s=0 \tag{12.4.128}$$

Working on the second integral (after the equality), we put:

$$u=g_{ij}\frac{\mathrm{d}x^i}{\mathrm{d}s}\qquad\text{and}\qquad v'=\frac{\delta\mathrm{d}x^j}{\mathrm{d}s}\mathrm{d}s \tag{12.4.129}$$

So by integration by part (see section Differential and Integral Calculus page 1532):

$$\int\limits_a^b g_{ij}\frac{\mathrm{d}x^i}{\mathrm{d}s}\frac{\delta\mathrm{d}x^j}{\mathrm{d}s}\mathrm{d}s=uv'\Big|_a^b-\int\limits_a^b u'v\mathrm{d}s \tag{12.4.130}$$

becomes:

$$\int\limits_a^b g_{ij}\frac{\mathrm{d}x^i}{\mathrm{d}s}\frac{\delta \mathrm{d}x^j}{\mathrm{d}s}\mathrm{d}s = g_{ij}\frac{\mathrm{d}x^i}{\mathrm{d}s}\delta\mathrm{d}x^j\mathrm{d}s\bigg|_a^b - \int\limits_a^b \frac{\mathrm{d}}{\mathrm{d}s}\left(g_{ij}\frac{\mathrm{d}x^i}{\mathrm{d}s}\right)\delta\mathrm{d}x^j\mathrm{d}s \tag{12.4.131}$$

Thus finally:

$$\int\limits_a^b \delta(\mathrm{d}s) = \int\limits_a^b \frac{1}{2}\frac{\mathrm{d}x^i}{\mathrm{d}x}\frac{\mathrm{d}x^j}{\mathrm{d}s}\left(\frac{\partial g_{ij}}{\partial x^k}\delta x^k\right)\mathrm{d}s + g_{ij}\frac{\mathrm{d}x^i}{\mathrm{d}s}\delta\mathrm{d}x^j\mathrm{d}s\bigg|_a^b - \int\limits_a^b \frac{\mathrm{d}}{\mathrm{d}s}\left(g_{ij}\frac{\mathrm{d}x^i}{\mathrm{d}s}\right)\delta\mathrm{d}x^j\mathrm{d}s = 0 \tag{12.4.132}$$

The non-integrated term below:

$$g_{ij}\frac{\mathrm{d}x^i}{\mathrm{d}s}\delta\mathrm{d}x^j\mathrm{d}s\bigg|_a^b \cong 0 \tag{12.4.133}$$

is negligible because of the presence of the factor $\delta\mathrm{d}x^j$: Therefore:

$$\begin{aligned}\int\limits_a^b \delta(\mathrm{d}s) &= \int\limits_a^b \frac{1}{2}\frac{\mathrm{d}x^i}{\mathrm{d}x}\frac{\mathrm{d}x^j}{\mathrm{d}s}\left(\frac{\partial g_{ij}}{\partial x^k}\delta x^k\right)\mathrm{d}s - \int\limits_a^b \frac{\mathrm{d}}{\mathrm{d}s}\left(g_{ij}\frac{\mathrm{d}x^i}{\mathrm{d}s}\right)\delta\mathrm{d}x^j\mathrm{d}s\\ &= \int\limits_a^b \left[\frac{1}{2}\frac{\mathrm{d}x^i}{\mathrm{d}x}\frac{\mathrm{d}x^j}{\mathrm{d}s}\left(\frac{\partial g_{ij}}{\partial x^k}\delta x^k\right) - \frac{\mathrm{d}}{\mathrm{d}s}\left(g_{ij}\frac{\mathrm{d}x^i}{\mathrm{d}s}\right)\delta\mathrm{d}x^j\right]\mathrm{d}s = 0\end{aligned} \tag{12.4.134}$$

We make a change of index:

$$\int\limits_a^b \delta(\mathrm{d}s) = \int\limits_a^b \left[\frac{1}{2}\frac{\mathrm{d}x^i}{\mathrm{d}x}\frac{\mathrm{d}x^j}{\mathrm{d}s}\left(\frac{\partial g_{ij}}{\partial x^k}\delta x^k\right) - \frac{\mathrm{d}}{\mathrm{d}s}\left(g_{ik}\frac{\mathrm{d}x^i}{\mathrm{d}s}\right)\delta\mathrm{d}x^k\right]\mathrm{d}s = 0 \tag{12.4.135}$$

Which allows us to factorize δx^k:

$$\int\limits_a^b \delta(\mathrm{d}s) = \int\limits_a^b \left[\frac{1}{2}\frac{\mathrm{d}x^i}{\mathrm{d}x}\frac{\mathrm{d}x^j}{\mathrm{d}s}\frac{\partial g_{ij}}{\partial x^k} - \frac{\mathrm{d}}{\mathrm{d}s}\left(g_{ik}\frac{\mathrm{d}x^i}{\mathrm{d}s}\right)\right]\delta\mathrm{d}x^k\mathrm{d}s = 0 \tag{12.4.136}$$

As $\delta\mathrm{d}x^k$ and $\mathrm{d}s$ are different from zero, it is the integrand that must be zero:

$$\frac{1}{2}\frac{\mathrm{d}x^i}{\mathrm{d}x}\frac{\mathrm{d}x^j}{\mathrm{d}s}\frac{\partial g_{ij}}{\partial x^k} - \frac{\mathrm{d}}{\mathrm{d}s}\left(g_{ik}\frac{\mathrm{d}x^i}{\mathrm{d}s}\right) = 0 \tag{12.4.137}$$

By developing the second term:

$$\frac{1}{2}\frac{\mathrm{d}x^i}{\mathrm{d}x}\frac{\mathrm{d}x^j}{\mathrm{d}s}\frac{\partial g_{ij}}{\partial x^k} - \frac{\mathrm{d}g_{ik}}{\mathrm{d}s}\frac{\mathrm{d}x^i}{\mathrm{d}s} - g_{ik}\frac{\mathrm{d}}{\mathrm{d}s}\frac{\mathrm{d}x^i}{\mathrm{d}s} = 0 \tag{12.4.138}$$

Which can also be written (in the physicist way of life...)

$$\frac{1}{2}\frac{\mathrm{d}x^i}{\mathrm{d}x}\frac{\mathrm{d}x^j}{\mathrm{d}s}\frac{\partial g_{ij}}{\partial x^k} - \frac{\mathrm{d}x^i}{\mathrm{d}s}\frac{\partial g_{ik}}{\partial x^j}\frac{\mathrm{d}x^j}{\mathrm{d}s} - g_{ik}\frac{\mathrm{d}}{\mathrm{d}s}\frac{\mathrm{d}x^i}{\mathrm{d}s} = 0 \tag{12.4.139}$$

Which simplifies into:

$$\boxed{g_{ik}\frac{\mathrm{d}^2x^i}{\mathrm{d}s^2}+\left(\frac{\partial g_{ik}}{\partial x^j}-\frac{1}{2}\frac{\partial g_{ij}}{\partial x^k}\right)=\frac{\mathrm{d}x^j}{\mathrm{d}s}\frac{\mathrm{d}x^j}{\mathrm{d}s}=0} \tag{12.4.140}$$

We fall back (again!) on the system of equations which defines the "geodesics", that is to say the straight lines of $\mathcal{E}^n$. These latter then constitute the extremities of the integral which measures the length of a curve arc joining two given points in $\mathcal{E}^n$.

This last equation is the one which interests us in the case of the free Lagrangian. Indeed, if we take the extreme case of light (or photons if you prefer), the latter will not seek the fastest path at the temporal level. This would totally contradict the postulate of invariance to see the light accelerate according to the path !!! In this context, it means that on the spatio-temporal framework, the only thing that has meaning is the shortest spatial path and not the shortest temporal path! This is why the latter equation is named the "geodesic equation" or also "generalized Euler-Lagrange equation".

However, we can write this last equation in a more condensed form by introducing the Christoffel symbols if the metric is a symmetric tensor, that is to say if $g_{\alpha\beta}=-g_{\beta\alpha}$.

Indeed:

$$g_{ik}\frac{\mathrm{d}^2x^i}{\mathrm{d}s^2}+\frac{1}{2}\left(2\frac{\partial g_{ik}}{\partial x^j}-\frac{\partial g_{ij}}{\partial x^k}\right)\frac{\mathrm{d}x^i}{\mathrm{d}s}\frac{\mathrm{d}x^j}{\mathrm{d}s}=g_{ik}\frac{\mathrm{d}^2x^i}{\mathrm{d}s^2}+\frac{1}{2}\left(\frac{\partial g_{ik}}{\partial x^j}+\frac{\partial g_{ik}}{\partial x^i}-\frac{\partial g_{ij}}{\partial x^k}\right)\frac{\mathrm{d}x^i}{\mathrm{d}s}\frac{\mathrm{d}x^j}{\mathrm{d}x^j}\mathrm{d}s=0 \tag{12.4.141}$$

And as the Christoffel symbol of the first kind (see section Tensor Calculus page 2107) is defined by:

$$\Gamma_{ikj}=\frac{1}{2}\left(\partial_j g_{ik}+\partial_i g_{kj}-\partial_k g_{ij}\right) \tag{12.4.142}$$

Remark

It is important to remember that this symbol contains almost all information about the space-time metric! We will see an example further below where in a locally inertial frame this Christoffel symbol will be equal to zero.

Then the Euler-Lagrange equation can be written:

$$g_{ik}\frac{\mathrm{d}^2x^i}{\mathrm{d}s^2}+\Gamma_{ikj}\frac{\mathrm{d}x^i}{\mathrm{d}s}\frac{\mathrm{d}x^j}{\mathrm{d}s}=0 \tag{12.4.143}$$

The contracted multiplication (see section Tensor Calculus page 2038) of the preceding relation in the canonical basis by g^{kl} gives us:

$$g^{kl}g_{ik}\frac{\mathrm{d}^2x^i}{\mathrm{d}s^2}+g^{kl}\Gamma_{ikj}\frac{\mathrm{d}x^i}{\mathrm{d}s}\frac{\mathrm{d}x^j}{\mathrm{d}s}=\delta_{il}\frac{\mathrm{d}^2x^i}{\mathrm{d}s^2}+\Gamma^l_{ij}\frac{\mathrm{d}x^i}{\mathrm{d}s}\frac{\mathrm{d}x^j}{\mathrm{d}s}=0 \tag{12.4.144}$$

Hence:

$$\frac{\mathrm{d}^2x^i}{\mathrm{d}s^2}+\Gamma^l_{ij}\frac{\mathrm{d}x^i}{\mathrm{d}s}\frac{\mathrm{d}x^j}{\mathrm{d}s}=0 \tag{12.4.145}$$

In the literature a change of index is often carried out in order to finally get (it is still the same expression given that the indices have the same range of variation!):

$$\boxed{\frac{\mathrm{d}^2x^\mu}{\mathrm{d}s^2}+\Gamma^\mu_{\alpha\beta}\frac{\mathrm{d}x^\alpha}{\mathrm{d}s}\frac{\mathrm{d}x^\beta}{\mathrm{d}s}=0} \tag{12.4.146}$$

with $\Gamma^\mu_{\alpha\beta}$ being the Christoffel symbol of the second kind (see section Tensor Calculus page 2105) given by:

$$\Gamma^\mu_{\alpha\beta}=g^{\mu l}\Gamma_{\alpha\lambda\beta}=\frac{1}{2}g^{\mu\lambda}\left(\partial_\alpha g_{\lambda\beta}+\partial_\beta g_{\alpha\lambda}-\partial_\lambda g_{\alpha\beta}\right) \tag{12.4.147}$$

and is named in the context of General Relativity the "affine connection" or "connection coefficients" and which makes it possible to find the system of coordinates (through the resolution of a system of differential equations) in free fall in which the particle equation is that of a uniform movement in space-time as a function of a reference system (the two systems are therefore connected by the affine connection).

This relation, of the highest importance, allows us to determine how a moving body will naturally move in a curved space and this perhaps ... regardless of its mass !!! It therefore gives us the metric in which we must set a frame of reference so that it is inertial with respect to the body in question.

The previous equation of geodesics is also the differential equation of the second order which must therefore satisfy the parametric representation of a line on a surface where s is the length along the line so that its total length is extremal!!!

According to the principle of equivalence, we are therefore entitled to interpret this relation as the equation of motion in any gravitational field of and thus to interpret the second additional term of the equation as the opposite of a gravitational term force per unit mass, that is to say as the opposite of a gravitational field!

Remark

We can also write the equation of geodesics by using the proper time. Indeed:

$$\frac{\mathrm{d}^2x^\mu}{\left(\dfrac{\mathrm{d}\tau c}{\gamma}\right)^2}+\Gamma^\mu_{\alpha\beta}\frac{\mathrm{d}x^\alpha}{\left(\dfrac{\mathrm{d}\tau c}{\gamma}\right)}\frac{\mathrm{d}x^\beta}{\left(\dfrac{\mathrm{d}\tau c}{\gamma}\right)}=\frac{\mathrm{d}^2x^\mu}{\mathrm{d}\tau^2}+\Gamma^\mu_{\alpha\beta}\frac{\mathrm{d}x^\alpha}{\mathrm{d}\tau}\frac{\mathrm{d}x^\beta}{\mathrm{d}\tau}=0 \tag{12.4.148}$$

or by using the four-vector velocity:

$$\frac{\mathrm{d}u^\mu}{\mathrm{d}t}+\Gamma^\mu_{\alpha\beta}\mathrm{d}u^\alpha\mathrm{d}u^\beta=0 \tag{12.4.149}$$

Again, if we restrict ourselves to a flat space-time, we see trivially that we fall back on the first equation of motion that we had obtained since for the Minkowski metric $\eta_{\mu\nu}$ we have immediately $\Gamma^\mu_{\alpha\beta}=0$:

$$\frac{\mathrm{d}^2x^\mu}{\mathrm{d}s^2}+0\cdot\frac{\mathrm{d}x^\alpha}{\mathrm{d}s}\frac{\mathrm{d}x^\beta}{\mathrm{d}s}=\frac{\mathrm{d}^2x^\mu}{\mathrm{d}s^2}=\frac{\mathrm{d}^2x^\mu}{\left(\dfrac{\mathrm{d}\tau c^2}{\gamma}\right)^2}=\frac{\mathrm{d}^2x^\mu}{\mathrm{d}\tau^2}=0 \tag{12.4.150}$$

because the components of the Minkowski metric being constant the Christoffel coefficients are all zero.

The solutions of the latter equation are ordinary straight lines given by:

$$x^{\mu}(s)=a^{\mu}s-b^{\mu} \tag{12.4.151}$$

Obviously, in a general curved space-time, the geodesics can not be globally represented by straight lines. However, with a second-order approximation in Taylor's development (see section Sequences and Series), we fall back on straight lines (which is equivalent to bringing the curved space back to a flat space).

The important thing in all this is that the equation of geodesics makes it possible to observe that the curvature of space determines the trajectories of the bodies which move there whatever their mass, whether they are in uniform motion or not (observe the second derivative in the geodesic equation!). All that remains is then to complete the work and to relate the curvature of space-time with the energy that is there!

To fully understand the change of perspective with Newtonian Mechanics, consider yourself falling from a plane without a parachute. From the point of view of General Relativity, you do not undergo any force (!), you are not accelerating, you just follow the curvature of space-time. You follow your normal geodesic trajectory: your acceleration (in the sense of General Relativity!) is zero!

Now imagine lying on your bed. You are in a gravity field and yet you are not in a free fall. So you are not in the geodesic trajectory following the curvature of space-time. Indeed, you suffer a force that prevents you from following this path: the reaction force of your bed! From the point of view of the General Relativity, if you are static in a gravity field, you undergo an acceleration!

12.4.3.2 Newtonian Limit

We have shown above (Shild's argument) that to study gravitation (in particular the Einstein's effect), curved geometry is necessary. We promised also to show that it was enough. Now is the time to do it!

Definition (#12.4.544): The "Newtonian limit" is a physical situation where the three conditions below are satisfied:

C1. The particles move slowly with respect to the speed of light. This is expressed as the fact that the variations of the spatial components of their quadrivector are much less than those of the temporal component (t being the proper time):

$$\frac{\mathrm{d}x^i}{\mathrm{d}\tau}\ll\frac{\mathrm{d}x^0}{\mathrm{d}\tau} \tag{12.4.152}$$

C2. The gravitational field is static. In other words, any time derivative of the metric is zero!

C3. The gravitational field is weak, that is, it can be seen as a weak perturbation of a flat space:

$$g_{\mu\nu} \cong \eta_{\mu\nu} + h_{\mu\nu} \tag{12.4.153}$$

with $|h_{\mu\nu}| \ll 1$ and where $\eta_{\mu\nu}$ is constant (only $h_{\mu\nu}$ depends on the coordinates).

Let us consider the geodesic equation obtained previously:

$$\frac{\mathrm{d}^2 x^\mu}{\mathrm{d}\tau^2} + \Gamma^\mu_{\alpha\beta} \frac{\mathrm{d}x^\alpha}{\mathrm{d}\tau} \frac{\mathrm{d}x^\beta}{\mathrm{d}\tau} = 0 \tag{12.4.154}$$

The first condition (C1) leads us to simplify it in the form:

$$\frac{\mathrm{d}^2 x^\mu}{\mathrm{d}\tau^2} + \Gamma^\mu_{00} \left(\frac{\mathrm{d}x^0}{\mathrm{d}t} \right)^2 = 0 \tag{12.4.155}$$

Remark

The latter relation can often be found on various social networks written and rearranged as (...):

$$\frac{\mathrm{d}^2 r}{\mathrm{d}\tau^2} = -\Gamma^r_{tt} (u^t)^2 \tag{12.4.156}$$

and don't forget that the latter describes the acceleration in curved space time!

By the way... how do we describe using typical General Relativity notation the fact (for example) that we are at rest on the ground of Earth (i.e. the ground is opposing a force/acceleration against our free geodesic fall to Earth as the curvature $\Lambda^r_{tt} \neq 0$)? The answer is straightforward! We just have to add the classical Newton acceleration term such as:

$$\frac{\mathrm{d}^2 r}{\mathrm{d}\tau^2} = a^r - \Gamma^r_{tt} (u^t)^2 = \frac{F}{m} - \Gamma^r_{tt} (u^t)^2 \tag{12.4.157}$$

Therefore we can see that the left hand side to be zero needs the acceleration due to curvature to be compensate by another acceleration due to the reaction of the ground! Explicitly:

$$\frac{\mathrm{d}^2 r}{\mathrm{d}\tau^2} = 0 \Leftrightarrow a^r = \Gamma^r_{tt} (u^t)^2 \tag{12.4.158}$$

The two other conditions (C2 and C3 whose application has been shown in the development below) offer us several simplifications in the expression of the symbol of Christoffel of the second kind:

$$\begin{aligned}
\Gamma^\mu_{00} &= \frac{1}{2} g^{\mu\lambda} (\partial_0 g_{\lambda 0} + \partial_0 g_{0\lambda} - \partial_\lambda g_{00}) \\
&= -\frac{1}{2} g^{\mu\lambda} \partial_\lambda g_{00} \qquad \text{(C2)} \\
&\cong -\frac{1}{2} (\eta^{\mu\lambda} + h^{\mu\lambda}) \left(\underbrace{\partial_\lambda \eta_{00}}_{=0\,\text{(C3)}} + \partial_\lambda h_{00} \right) = -\frac{1}{2} (\eta^{\mu\lambda} + h^{\mu\lambda}) \partial_\lambda h_{00} \cong -\frac{1}{2} \eta^{\mu\lambda} \partial_\lambda h_{00}
\end{aligned} \tag{12.4.159}$$

The geodesic equation then becomes:

$$\frac{\mathrm{d}^2x^\mu}{\mathrm{d}\tau^2}-\frac{1}{2}\eta^{\mu\lambda}\partial_\lambda h_{00}\left(\frac{\mathrm{d}x^0}{\mathrm{d}\tau}\right)^2=0 \tag{12.4.160}$$

and is then equal for the temporal component to ($\mu=0$):

$$\frac{\mathrm{d}^2x^0}{\mathrm{d}\tau^2}-\frac{1}{2}\eta^{0\lambda}\partial_\lambda h_{00}\left(\frac{\mathrm{d}x^0}{\mathrm{d}\tau}\right)^2=0 \tag{12.4.161}$$

But (recall of the Minkowski metric):

$$\eta^{0\lambda}=0 \tag{12.4.162}$$

for $\lambda>0$ and for $\lambda=0$ we have (static metric):

$$\partial_0 h_{00}=0 \tag{12.4.163}$$

Therefore, we must conclude that $\mathrm{d}x^0/\mathrm{d}t$ is a constant (whatever the choice of the signature of the Minkowski metric).

And for the spatial components, we know that $\eta^{\mu\nu}$ when reduced to its spatial part is a simple unitary 3×3 matrix, which gives for each spatial component in the case where we choose (by tradition only!) the signature $(-,+,+,+)$ of the Minkowski metric:

$$\frac{\mathrm{d}^2x^i}{\mathrm{d}\tau^2}-\frac{1}{2}\partial_i h_{00}\left(\frac{\mathrm{d}x^0}{\mathrm{d}\tau}\right)^2=0 \tag{12.4.164}$$

Obviously, the reader can have fun making the development that follows with the inverse signature $(+,-,-,-)$ and he will see that it only changes the sign of potential in the final result of the development): Let us now rearrange the above relation:

$$\frac{\mathrm{d}^2x^i}{\mathrm{d}\tau^2}=\frac{1}{2}\partial_i h_{00}\left(\frac{\mathrm{d}x^0}{\mathrm{d}\tau}\right)^2=0 \tag{12.4.165}$$

By dividing by $(\mathrm{d}x^0/\mathrm{d}\tau)^2$ and restoring $x^0=c\tau$, we get by making as sequence of simplifications:

$$\begin{aligned}\frac{\mathrm{d}x^i}{\mathrm{d}\tau^2}\left(\frac{\mathrm{d}\tau}{\mathrm{d}x^0}\right)^2&=\frac{1}{2}\partial_i h_{00}\\ \frac{\mathrm{d}^2x^i}{(\mathrm{d}x^0)^2}&=\frac{1}{2}\partial_i h_{00}\\ \frac{\mathrm{d}^2x^i}{(\mathrm{d}c\tau)^2}&=\frac{1}{2}\partial_i h_{00}\\ \frac{\mathrm{d}^2x^i}{\mathrm{d}\tau^2}&=\frac{c^2}{2}\partial_i h_{00}\end{aligned} \tag{12.4.166}$$

Starting from here we put (because our illustrious predecessors have tried before us):

$$h_{00}=-\frac{2}{c^2}\varphi=-\frac{2GM}{rc^2} \tag{12.4.167}$$

such as (a relation which will be very useful to us when studying the Schwarzschild metric further below):

$$g_{00}=\eta_{00}+h_{00}=1-\frac{2}{c^2}\varphi=1-\frac{2GM}{rc^2} \qquad (12.4.168)$$

where φ is the gravitational potential. We fall back here on the expression of the gravitational acceleration (Newton-Poisson equation) of the Newtonian mechanics (see section Astronomy page 5267):

$$\vec{a}=\frac{\mathrm{d}^2x^i}{\mathrm{d}t^2}=-\frac{c^2}{2}\partial_i\left(\frac{2\varphi}{c^2}\right)=-\partial_i\varphi=-\vec{\nabla}\varphi \qquad (12.4.169)$$

with $i=1,2,3$.

Remark

That last equality and analogy for the Newton-Poisson equation may not be obvious for some readers. So another direct analogy (and at the same time a more general result) will be given at the page 5778 using the general Einstein's field equations (i.e. with the cosmological constant)!

This development, simple but nevertheless remarkable by its interpretation, proves that the curved geometry is sufficient to describe the gravitation (and therefore the theory of Newton)!!!!!!!!!!!! This verification is named by some people the "principle of correspondence".

Demonstrating that a new theory reproduces all the achievements of successful old theories can be extremely difficult. This is because a new theory might use an entirely different mathematical framework that looks nothing like that of the old theory. Finding a way to show that both nevertheless arrive at the same predictions for already-made observations often requires finding a suitable way to reformulate the new theory. This is straightforward in cases where the new theory directly employs the math of the old one, but it can be a big hurdle with entirely new frameworks.

Einstein, for example, struggled for years to prove that General Relativity, his new theory of gravity, would reproduce the successes of the predecessor, Newtonian gravity. The problem wasn't that he had the wrong theory; the problem was that he didn't know how to find Newton's gravitational potential in his own theory. Einstein had all the math right, but the identification with the real world was missing. Only after several wrong attempts did he hit on the right way to do it. Having the right math is only part of having the right theory!

12.4.4 Stress-Energy Tensor

The "Stress-Energy Tensor (SET)" (sometimes named "stress–energy–momentum tensor" or "energy–momentum tensor") is a mathematical tool used (in particular) in General Relativity to represent the density and flux of energy and moment in space-time, generalizing the stress tensor of Newtonian physics of mass and energy. It is therefore an attribute of matter, radiation, and non-gravitational force fields. The stress–energy tensor is the source of the gravitational field in the Einstein's field equations of General Relativity, just as mass density is the source of such a field in Newtonian gravity[45].

Let us take for example the SET which considers matter in General Relativity as being able to be approximated by a perfect fluid. In the section Continuum Mechanics we have proved:

$$N_i=\sigma_{ij}n_j \tag{12.4.170}$$

where N_i has for recall the units of a force and n_j those of a surface. Thus with a more conventional writing:

$$F_i=\sigma_{ij}n_j \tag{12.4.171}$$

In variational form this gives:

$$\sigma_{ij}=\left.\frac{\Delta F_i}{\Delta n_j}\right|_{\Delta a_j\rightarrow 0} \tag{12.4.172}$$

Let us now calculate:

$$\sigma_{ij}=\frac{\Delta F_i}{\Delta n_j}=\frac{\Delta(ma_i)}{\Delta n_j}=\frac{\Delta(\rho Va_i)}{\Delta n_j} \tag{12.4.173}$$

Remark

We do not work with differential elements to avoid being trapped later. It is completely a physicist Do It Yourself approach, but it works well (confirmed by experience...).

Assuming that only the volume and the time makes that the force varies (which assume a constant density and the to be inertial) we then have:

$$\begin{aligned}\sigma_{ij}&=\frac{\Delta(\rho Va_i)}{\Delta n_j}=\frac{1}{\Delta n_j}\Delta\left(\frac{V}{t}\right)\rho v_i=\frac{1}{\Delta n_j}\frac{\Delta V}{\Delta t}\rho v_i=\frac{\Delta x_i\Delta x_j\Delta x_k}{\Delta n_j}\rho\frac{v_i}{\Delta t}\\&=\frac{\Delta x_i\Delta x_j\Delta x_k}{\Delta x_i\Delta x_k}\rho\frac{v_i}{\Delta t}=\frac{\Delta x_j}{\Delta t}\rho v_i=v_j\rho v_i=\rho v_j v_i\end{aligned} \tag{12.4.174}$$

This gives simply the tensor product of the velocities (see section Tensor Calculus page 2073):

$$\vec{\sigma}=\rho\vec{v}\otimes\vec{v} \tag{12.4.175}$$

If we generalize this relation to the velocity quadrivectors of Special Relativity with the corresponding notations, then we have by definition the "energy-momentum tensor" or "Stress–energy tensor":

$$\vec{T}=\rho\vec{u}\otimes\vec{u} \tag{12.4.176}$$

or in index form:

$$T_{ij}=\rho u_i u_j \tag{12.4.177}$$

---

[45] The gravitational field is a thing, but it's not an en "energy field"! Indeed, a field is one thing; energy is something it can have. But again, as far as we know, there is no such thing as an "energy field"!

Either in contravariant form (most common form in textbooks):

$$\boxed{T^{ij}=\rho u^i u^j} \tag{12.4.178}$$

This relation is the justification for which General Relativity is also indicated as a theory of continuous mechanics by some specialists.

Now let us prove the following equality of the derivative:

$$\partial_j T^{0j}=\partial_j(\rho u^0 u^j)=0 \tag{12.4.179}$$

Remark

What we have already pointed out in the section of Tensorial Calculus is written $T^{0j}_{,j}$ in old books or in modern textbooks where the author want to show its technical level...

First, let us recall that (see section Special Relativity page 5642):

$$u^\nu=(u^0,u^1,u^2,u^3)=\gamma(\vec{v})(c,v^1,v^2,v^3)=\gamma(\vec{v})(c,v^x,v^y,v^z) \tag{12.4.180}$$

and let us admit that we are in low speeds such as $\gamma=1$. Then, in a Minkowski metric of type $(+,-,-,-)$ we have:

$$\begin{aligned}\partial_j T^{0j}&=\partial_j(\rho u^0 u^j)=\partial_j(\rho c u^j)\\&=\partial_0(\rho c u^0)+\partial_1(\rho c u^1)+\partial_2(\rho c u^2)+\partial_3(\rho c u^3)\\&=\frac{\partial(\rho c c)}{\partial c t}-\frac{\partial(\rho c v^x)}{\partial x}-\frac{\rho c v^y}{\partial y}-\frac{\rho c v^z}{\partial z}\\&=c\left(\frac{\partial\rho}{\partial t}-\frac{\partial(\rho v^x)}{\partial x}-\frac{\partial(\rho v^y)}{\partial y}-\frac{\partial(\rho v^z)}{\partial z}\right)\\&=c\left(\frac{\partial\rho}{\partial t}-\vec{\nabla}\circ(\rho\vec{v})\right)\end{aligned} \tag{12.4.181}$$

But, we recognize in the parentheses the equation of continuity (conservation of the mass) which we have proved in the section of Thermodynamics (see page 3679) and which we know is equal to zero in a system without sources! Therefore:

$$\partial_j T^{0j}=c\left(\frac{\partial\rho}{\partial t}-\vec{\nabla}\circ(\rho\vec{v})\right)=c\cdot 0=0 \tag{12.4.182}$$

Let us also look to what contains the component T^{00} of the stress-energy tensor:

$$T^{00}=\rho u^0 u^0=\rho\gamma c\gamma c=c^2\gamma^2\rho \tag{12.4.183}$$

In terms of units, this is an energy density (we see directly that this quantity can only be positive).

Let us now look at the other components with $i=0$ and $j=1,2,3$:

$$T^{0j}=\rho u^0 u^j=\rho(\gamma c)(\gamma v^j)=c\rho\gamma^2 v^j=cp^i \tag{12.4.184}$$

where p^i has the units of linear momentum density.

Let us now consider the components of the tensor when $i,j=1\ldots3$ (we omit then the first row and the first column):

$$T^{ij}=\rho(\gamma u^i)(\gamma u^j)=\rho\gamma^2u^iu^j=\sigma_{ij} \tag{12.4.185}$$

We thus fall back on the components of the stress tensor of a perfect fluid.

So finally, the stress-energy tensor can be written in the form of a symmetric real 4×4 matrix:

$$\boxed{T^{ij}=\begin{bmatrix}T^{00}&T^{01}&T^{02}&T^{03}\\T^{10}&T^{11}&T^{12}&T^{13}\\T^{20}&T^{21}&T^{22}&T^{23}\\T^{30}&T^{31}&T^{32}&T^{33}\end{bmatrix}=\begin{bmatrix}\rho\gamma^2c^2&\rho\gamma^2cv^x&\rho\gamma^2cv^y&\rho\gamma^2cv^z\\\rho\gamma^2cv^x&\rho\gamma^2v^xv^x&\rho\gamma^2v^xv^y&\rho\gamma^2v^xv^z\\\rho\gamma^2cv^y&\rho\gamma^2v^yv^x&\rho\gamma^2v^yv^y&\rho\gamma^2v^yv^z\\\rho\gamma^2cv^z&\rho\gamma^2v^zv^x&\rho\gamma^2v^zv^y&\rho\gamma^2v^zv^z\end{bmatrix}} \tag{12.4.186}$$

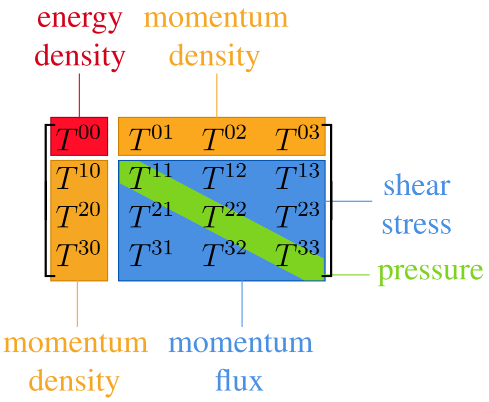
In the case where the velocities are small, i.e. $\gamma\rightarrow 1$, we have:

$$\begin{aligned}T^{ij}&=\begin{bmatrix}\rho c^2&\rho cv^x&\rho cv^y&\rho cv^z\\\rho cv^x&\rho v^xv^x&\rho v^xv^y&\rho v^xv^z\\\rho cv^y&\rho v^yv^x&\rho v^yv^y&\rho v^yv^z\\\rho cv^z&\rho v^zv^x&\rho v^zv^y&\rho v^zv^z\end{bmatrix}=\begin{bmatrix}\dfrac{mc^2}{V}&\rho cv^x&\rho cv^y&\rho cv^z\\\rho cv^x&\sigma_{xx}&\tau^{xy}&\tau^{xz}\\\rho cv^y&\tau^{yz}&\sigma_{yy}&\tau^{yz}\\\rho cv^z&\tau^{zx}&\tau^{zy}&\sigma_{zz}\end{bmatrix}\\&=\begin{bmatrix}\dfrac{E_{\text{tot}}}{V}&\rho cv^x&\rho cv^y&\rho cv^z\\\rho cv^x&\sigma_{xx}&\tau^{xy}&\tau^{xz}\\\rho cv^y&\tau^{yz}&\sigma_{yy}&\tau^{yz}\\\rho cv^z&\tau^{zx}&\tau^{zy}&\sigma_{zz}\end{bmatrix}\end{aligned} \tag{12.4.187}$$

This tensor is also sometimes represented as following:

$$\begin{aligned}T^{\mu\nu}&=\left(\begin{array}{c|ccc}T^{00}&T^{01}&T^{02}&T^{03}\\\hline T^{10}&T^{11}&T^{12}&T^{13}\\T^{20}&T^{21}&T^{22}&T^{23}\\T^{30}&T^{31}&T^{32}&T^{33}\end{array}\right)=\left(\begin{array}{c|c}T^{00}&T^{0i}\\\hline T^{i0}&T^{ij}\end{array}\right)\\&=\left(\begin{array}{c|c}\text{energy density}&\text{component } i \text{ of momentum density}\\\hline\text{component } i \text{ of momentum density}&\text{flux in } j \text{ direction of momentum component } i\end{array}\right)\end{aligned} \tag{12.4.188}$$

We thus fall back in this tensor on the following interpretations of the physical quantities (although rigorously all the components have units which can be seen as density of energy or as a pressure):


energy density
momentum density
T^{00} T^{01} T^{02} T^{03}
T^{10} T^{11} T^{12} T^{13}
T^{20} T^{21} T^{22} T^{23}
T^{30} T^{31} T^{32} T^{33}
shear stress
pressure
momentum density
momentum flux

There is another nice and interesting way to get the momentum-flux part of that tensor.

First let us recall the Euler equation of the 1st form that we get in the section of Continuum Mechanics at page 3828 (p is the pressure and not the linear momentum density!):

$$\rho \mathrm{d}_t\vec{v}=-\vec{\nabla}p+\rho\vec{f} \tag{12.4.189}$$

or more explicitly using the material derivative:

$$\rho\partial_t\vec{v}+\rho(\vec{v}\circ\vec{\nabla})\vec{v}=-\vec{\nabla}p+\rho\vec{f} \tag{12.4.190}$$

where for recall $\vec{f}=-\vec{\nabla}U$.

Let us assume that $\vec{f}=-\vec{\nabla}U=\vec{0}$. Therefore the latter equation can is reduced to:

$$\partial_t\vec{v}+(\vec{v}\circ\vec{\nabla})\vec{v}=-\frac{1}{\rho}\vec{\nabla}p \tag{12.4.191}$$

and explicitly to (for $i,k=1,2,3$):

$$\frac{\partial v^i}{\partial t}+v^k\frac{\partial v^i}{\partial x^k}=-\frac{1}{\rho}\frac{\partial p}{\partial x^i}=-\frac{1}{\rho}\delta_{ik}\frac{\partial p}{\partial x^k} \tag{12.4.192}$$

Hence:

$$\frac{\partial v^i}{\partial t}=-v^k\frac{\partial v^i}{\partial x^k}-\frac{1}{\rho}\delta_{ik}\frac{\partial p}{\partial x^k} \tag{12.4.193}$$

Now let us recall the equation of mass conservation (see section Continuum Mechanics page 3683):

$$\partial_t\rho+\vec{\nabla}\circ(\rho\vec{v})=0 \tag{12.4.194}$$

That we will rearrange and rewrite:

$$\frac{\partial\rho}{\partial t}=-\vec{\nabla}\circ(\rho\vec{v})=-\frac{-\partial(\rho v^k)}{\partial x^k} \tag{12.4.195}$$

But we also have:

$$\frac{\partial(\rho v^i)}{\partial t}=v^i\frac{\partial\rho}{\partial t}+\rho\frac{\partial v^i}{\partial t} \tag{12.4.196}$$

By injecting in that latter relation the mass conservation equation and the expression we got for $\partial_t v_i$ we then have:

$$\frac{\partial(\rho v^i)}{\partial t}=-v^i\frac{\partial v^i}{\partial x^k}-\rho v^k\frac{\partial v^i}{\partial v^k}-\delta_{ik}\frac{\partial p}{\partial x^k}=-\frac{\partial(\rho v^i v^k)}{\partial x^k}-\delta_{ik}\frac{\partial p}{\partial x^k}=-\frac{\partial(\rho v^i v^k+\delta_{ik}p)}{\partial x^k} \tag{12.4.197}$$

Therefore:

$$\frac{\partial(\rho v^i)}{\partial t}=-\text{div}(\rho v^i v^k+\delta_{ik}p) \tag{12.4.198}$$

It appears between the parenthesis a more general notation of the energy-stress tensor for $i,k=1,2,3$, that is the moment flux part:

$$T^{ik}=\rho v^i v^k+\delta_{ik}p \tag{12.4.199}$$

We then understand better why this matrix is named "Energy-Momentum Tensor" or "Stress-Energy-Momentum Tensor" since implicitly it is a question of modelling the space by a perfect fluid under shear stresses (tangential forces) and tensions (normal forces).

Remark

The sub-matrix of spatial components:

$$T^{ij}=\begin{bmatrix} T^{11} & T^{12} & T^{13} \\ T^{21} & T^{22} & T^{23} \\ T^{31} & T^{32} & T^{33} \end{bmatrix} \tag{12.4.200}$$

is the matrix named the "matrix of moments flows" (a name that is quite debatable ...). In Continuum Mechanics (see section of the same name page 3811), we have proved that its diagonal corresponds to the pressure, and the other components to the tangential forces due to the dynamic viscosity.

Let us prove that the covariant derivative (see section Tensor Calculus page 2113) of the stress-energy tensor is zero such that:

$$\nabla_\mu T^{\mu\nu}=0 \tag{12.4.201}$$

Therefore:

$$\nabla_\mu T^{\mu\nu}=\nabla_\mu(\rho u^\mu u^\nu)=u^\nu\nabla_\mu(\rho u^\mu)+\rho u^\mu\nabla_\mu(u^\nu) \tag{12.4.202}$$

Let us begin by developing the first term:

$$u^\nu\nabla_\mu(\rho u^\mu)=u^\nu\left(\frac{\partial(\rho\gamma c)}{\partial ct}-\frac{\partial(\rho\gamma u^x)}{\partial x}-\frac{\partial(\rho\gamma u^x)}{\partial x}-\frac{\partial(\rho\gamma u^y)}{\partial x}-\frac{\partial(\rho\gamma u^z)}{\partial z}\right) \tag{12.4.203}$$

But we have:

$$\frac{\partial\gamma}{\partial x}=\frac{\partial\gamma}{\partial y}=\frac{\partial\gamma}{\partial z}=0 \tag{12.4.204}$$

hence:

$$u^\nu\nabla_\mu(\rho u^\mu)=u^\nu\left(\frac{\partial(\rho\gamma)}{\partial t}-\gamma\vec{\nabla}\circ(\rho\vec{u})\right)=u^\nu\left(\rho\frac{\partial\gamma}{\partial t}+\gamma\left[\frac{\partial(\rho)}{\partial t}-\vec{\nabla}\circ(\rho\vec{u})\right]\right) \tag{12.4.205}$$

We find in the squared brackets the equation of continuity which is zero in the absence of sources. On the other hand, the first term in parentheses is non-zero as we saw in our study of the four-accelerator acceleration in the section of Special Relativity:

$$u^\nu\nabla_\mu(\rho u^\mu)=u^\nu\rho\frac{\partial\gamma}{\partial t}=u^\nu\rho\left(\frac{\gamma^3}{c^2}\vec{v}\circ\vec{v}\right) \tag{12.4.206}$$

But according to the weak principle of equivalence (WPE), we can always place ourselves in a repository such that locally the acceleration is null, that is to say such that (for recall, we do not put vector arrows for the quadrivectors):

$$a=0 \tag{12.4.207}$$

And it comes then:

$$u^\nu\nabla_\mu(\rho u^\mu)=0 \tag{12.4.208}$$

So we now have:

$$\nabla_\mu T^{\mu\nu}=u^\nu\nabla_\mu(\rho u^\mu)+\rho u^\mu\nabla_\mu(u^\nu)=0+\rho u^\mu\nabla_\mu(u^v)=\rho u^\mu\nabla_\mu(u^\nu) \tag{12.4.209}$$

Let us look at what this last term gives but first recalling that in the section of Special Relativity we had proved that the quadri-acceleration was expressed according to:

$$a^\nu=\left(\frac{\gamma^4}{c}\vec{v}\circ\vec{a},\gamma^2\vec{a}+\frac{\gamma^4}{c^2}(\vec{v}\circ\vec{a})\vec{v}\right) \tag{12.4.210}$$

Therefore (we take only the first two components as examples):

$$\begin{aligned}a^0&=\frac{\gamma^4}{c}(v^xa^x+v^ya^y+v^za^z)\\a^1&=\gamma^2a^x+\frac{\gamma^4}{c^2}(v^xa^x+v^ya^y+v^za^z)v^x\\&\ldots\end{aligned} \tag{12.4.211}$$

We will now in fact prove that:

$$u^\mu\nabla_\mu(u^\nu)=a^\nu \tag{12.4.212}$$

For this we start first by proving that:

$$u^\mu\nabla_\mu(u^0)=a^0 \tag{12.4.213}$$

For this we calculate first:

$$u^\mu\nabla_\mu(u^0)=\gamma c\frac{\partial(\gamma c)}{\partial(ct)}+\gamma v^x\frac{\partial(\gamma c)}{\partial x}+\gamma v^y\frac{\partial(\gamma c)}{\partial y}+\gamma v^z\frac{\partial(\gamma c)}{\partial z} \tag{12.4.214}$$

But:

$$\frac{\partial\gamma}{\partial x}=\frac{\partial\gamma}{\partial y}=\frac{\partial\gamma}{\partial z}=0\qquad\text{and}\qquad\frac{\partial c}{\partial x}=\frac{\partial c}{\partial y}=\frac{\partial c}{\partial z}=0 \tag{12.4.215}$$

Therefore:

$$u^\mu\nabla_\mu(u^0)=\gamma c\frac{\partial(\gamma c)}{\partial(ct)}=\gamma c\frac{\partial\gamma}{\partial t}=\gamma c\left(\frac{\gamma^3}{c^2}\vec{v}\circ\vec{a}\right)=\frac{\gamma^4}{c}\vec{v}\circ\vec{a}=a^0 \tag{12.4.216}$$

Now let us prove that:

$$u^\mu\nabla_\mu(u^1)=a^1 \tag{12.4.217}$$

the other components a^2, a^3 are then verified automatically.

For this we do a little bit of algebra:

$$\begin{aligned}u^\mu\nabla_\mu(u^1)&=\gamma c\frac{\partial(\gamma v^x)}{\partial(ct)}+\underbrace{\gamma v^x\frac{\partial(\gamma v^x)}{\partial x}+\gamma v^y\frac{\partial(\gamma v^x)}{\partial y}+\gamma v^z\frac{\partial(\gamma v^x)}{\partial z}}_{=0}\\&=\gamma\frac{\partial(\gamma v^x)}{\partial t}=\gamma\left(v^2\frac{\partial\gamma}{\partial t}+\gamma\frac{\partial v^x}{\partial t}\right)=\gamma\left(v^x\left(\frac{\gamma^3}{c^2}\vec{v}\circ\vec{a}\right)+\gamma a^x\right)\\&=\gamma^2a^x+\left(\frac{\gamma^4}{c^2}\vec{v}\circ\vec{a}\right)v^x=a^1\end{aligned} \tag{12.4.218}$$

and therefore we have indeed:

$$u^{\mu}\nabla_{\mu}(u^{\nu})=a^{\nu} \tag{12.4.219}$$

but according to the WEP (Weak Equivalence Principle[46]), $a^{\nu}=0$ therefore:

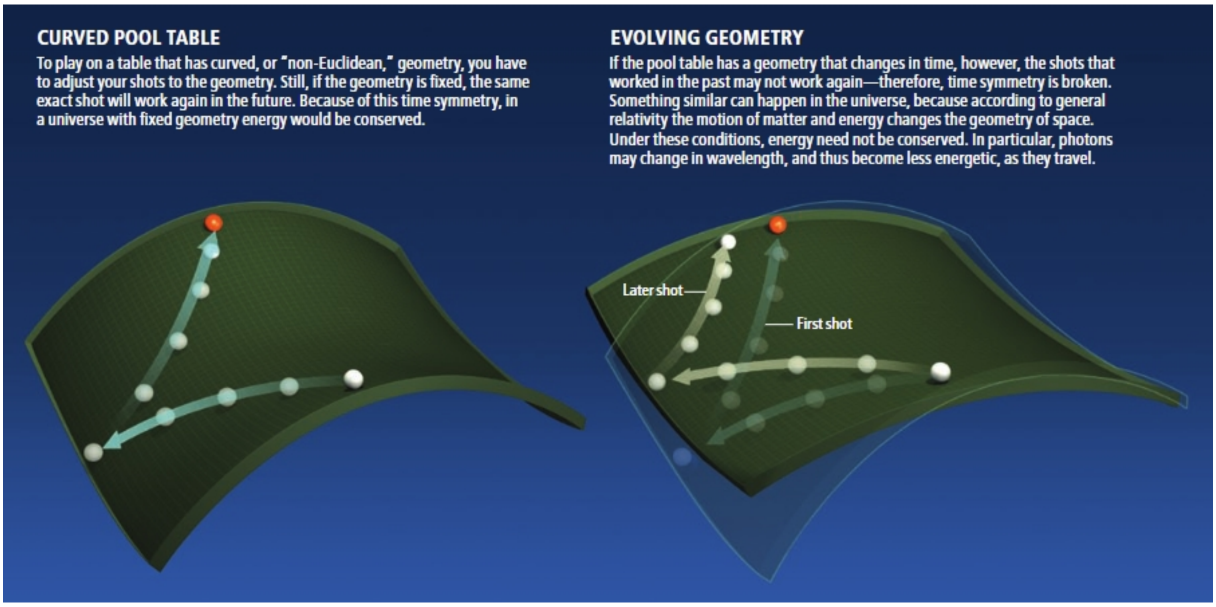
$$u^{\mu}\nabla_{\mu}(u^{\nu})=0 \tag{12.4.220}$$

and finally we have indeed under the assumptions stated above:

$$\boxed{\nabla_{\mu}T^{\mu\nu}=0} \tag{12.4.221}$$

Which is the expression of the conservation of energy in General Relativity (because the components that are energy and momentum are conserved quantities)!

The meaning of this equation is straightforward enough: energy and momentum evolve in a precisely specified way in response to the behaviour of space-time around them. If that space-time is standing completely still, the total energy is constant; if it's evolving, the energy changes in a completely unambiguous way!



CURVED POOL TABLE

To play on a table that has curved, or "non-Euclidean," geometry, you have to adjust your shots to the geometry. Still, if the geometry is fixed, the same exact shot will work again in the future. Because of this time symmetry, in a universe with fixed geometry energy would be conserved.

EVOLVING GEOMETRY

If the pool table has a geometry that changes in time, however, the shots that worked in the past may not work again—therefore, time symmetry is broken. Something similar can happen in the universe, because according to general relativity the motion of matter and energy changes the geometry of space. Under these conditions, energy need not be conserved. In particular, photons may change in wavelength, and thus become less energetic, as they travel.

By lowering the indices it comes:

$$\nabla^{\rho}T_{\rho\sigma}=g^{\rho\mu}\nabla_{\rho}\left(g_{\rho\mu}g_{\sigma\nu}T^{\rho\sigma}\right)=\nabla_{\mu}T^{\mu\nu}=0 \tag{12.4.222}$$

---

[46] For reminder it states that all free-falling test particles fall with the same acceleration in a gravitational field, regardless of their mass or internal composition

12.4.4.1 Stress-Energy Tensor for a perfect fluid

In physics, a perfect fluid is a fluid that can be completely characterized by its rest frame mass density ρ isotropic pressure p.

Real fluids are "sticky" and contain (and conduct) heat. Perfect fluids are idealized models in which these possibilities are neglected. Specifically, perfect fluids have no shear stresses, viscosity, or heat conduction.

According to these properties, if we rewrite the stress-energy tensor:

$$T^{ij}=\begin{bmatrix}\dfrac{E_{\text{tot}}}{V} & \rho c v^x & \rho c v^y & \rho c v^z\\ \rho c v^x & \rho v^x v^x & \rho v^x v^y & \rho v^x v^z\\ \rho c v^y & \rho v^y v^z & \rho v^y v^y & \rho v^y v^z\\ \rho c v^z & \rho v^z v^x & \rho v^z v^y & \rho v^y v^z\end{bmatrix}=\begin{bmatrix}\dfrac{E_{\text{tot}}}{V} & \rho c v^x & \rho c v^y & \rho c v^z\\ \rho c v^x & \sigma_{xx} & \tau^{xy} & \tau^{xz}\\ \rho c v^y & \tau^{yz} & \sigma_{yy} & \tau^{yz}\\ \rho c v^z & \tau^{zx} & \tau^{zy} & \sigma_{zz}\end{bmatrix}\tag{12.4.223}$$

as:

$$T^{ij}=\begin{bmatrix}\rho c^2 & \rho c v^x & \rho c v^y & \rho c v^z\\ \rho c v^x & P & \tau^{xy} & \tau^{xz}\\ \rho c v^y & \tau^{yz} & P & \tau^{yz}\\ \rho c v^z & \tau^{zx} & \tau^{zy} & P\end{bmatrix}\tag{12.4.224}$$

Form the definition of the perfect fluid, especially the no heat conduction, we have immediately that:

$$T^{01}=T^{02}=T^{03}=T^{10}=T^{20}=T^{30}=0\tag{12.4.225}$$

Indeed, energy can flow only if particles flow. So if there is not heat conduction, there is no particle flow and therefore no momentum density.

If there is no viscosity we have also obviously all τ_{ij} that are zero.

Therefore so far we have:

$$T^{ij}=\begin{bmatrix}\rho c^2 & 0 & 0 & 0\\ 0 & P & 0 & 0\\ 0 & 0 & P & 0\\ 0 & 0 & 0 & P\end{bmatrix}\tag{12.4.226}$$

But this can be written:

$$\boxed{T^{ij}=\left(\rho+\frac{P}{c^2}\right)u^i u^j-P\eta^{ij}}\tag{12.4.227}$$

where we choose the signature $(+,-,-,-)$ with for recall:

$$u^\nu=\gamma(c,v^x,v^y,v^z)\tag{12.4.228}$$

and where we assume for the perfect fluid that $\gamma\rightarrow 1$ and that we are in the local rest frame such as:

$$u^\nu=(c,0,0,0)\tag{12.4.229}$$

As in many cases we consider weak static gravitational field, then it is common to write the previous boxed relation:

$$\boxed{T^{ij} \cong \left(\rho+\frac{P}{c^2}\right)u^i u^j - Pg^{ij}} \tag{12.4.230}$$

Perfect fluids are often used in General Relativity to model idealized distributions of matter, such as the interior of a star or an isotropic universe. In the latter case, the equation of state of the perfect fluid may be used in Friedmann–Lemaître–Robertson–Walker equations to describe the evolution of our Universe as we will see it in the next section.

12.4.4.2 Electromagnetic stress–energy tensor

In relativistic physics, the electromagnetic stress–energy tensor is the contribution to the stress–energy tensor due to the electromagnetic field. The stress–energy tensor describes the flow of energy and momentum in space-time. The electromagnetic stress–energy tensor contains the classical Maxwell stress tensor that governs the electromagnetic interactions.

Let us recall first that in the section of Electrodynamics (see section Electrodynamics page 4156) we have proved the energy density carried by an electrodynamic wave was:

$$u_{\text{tot}} = \frac{1}{2}\varepsilon\hat{E}^2 + \frac{1}{2\mu}\hat{B}^2 = \frac{1}{2}\varepsilon\hat{E}^2 + \frac{1}{2\mu c^2}\hat{E}^2 = \varepsilon\hat{E}^2 = \frac{1}{\mu}\hat{B}^2 \tag{12.4.231}$$

And also that (see section Electrodynamics page 4135):

$$(g_{\mu\sigma}g_{\nu\rho}F^{\sigma\rho})F^{\mu\nu} = F_{\mu\nu}F^{\mu\nu} = 2\left(c^2\|\vec{B}\|^2 - \|\vec{E}\|^2\right) \tag{12.4.232}$$

and that (see section Electrodynamics page 4128):

$$E^k = F^{ik}u^k \tag{12.4.233}$$

Ok now that we have the necessary tools let's go!

From the previous relation, we get:

$$\|\vec{E}^2\| = \frac{1}{c^2}F_{\beta\alpha}u^\alpha F_\gamma^\beta u^\gamma = \frac{1}{c^2}F_{\beta\alpha}u^\alpha(g_{\gamma\sigma}F^{\beta\sigma})u^\gamma \tag{12.4.234}$$

and for the remaining part of the development will take the metric $\eta_{\gamma\sigma}$. Then:

$$\|\vec{E}^2\| = F_{\beta\alpha}u^\alpha\left(\eta_{\gamma 1}F^{\beta 1} + \eta_{\gamma 2}F^{\beta 2} + \eta_{\gamma 3}F^{\beta 3} + \eta_{\gamma 4}F^{\beta 4}\right)u^\gamma \tag{12.4.235}$$

As in an inertial frame in which it is at rest where the four-velocity is $u^\gamma = (c,0,0,0)$, we have then to focus only on $\gamma = 1$. Therefore:

$$\|\vec{E}^2\| = F_{\beta\alpha}u^\alpha\left(\eta_{11}F^{\beta 1} + \eta_{12}F^{\beta 2} + \eta_{13}F^{\beta 3} + \eta_{14}F^{\beta 4}\right)c = -cF_{\beta\alpha}u^\alpha F^{\beta 1} \tag{12.4.236}$$

As in an inertial frame in which it is at rest where the four-velocity is $u^\alpha = (c,0,0,0)$, we have then to focus only on $\alpha = 1$. Therefore:

$$\begin{aligned}\|\vec{E}^2\| &= -c^2F_{\beta 1}F^{\beta 1} = -c^2\left(F_{11}F^{11} + F_{21}F^{21} + F_{31}F^{31} + F_{41}F^{41}\right)\\ &= -c^2(0\cdot 0 + (-E_x)E_x + (-E_y)E_y + (-E_z)E_z) = E_x^2 + E_y^2 + E_z^2\end{aligned} \tag{12.4.237}$$

Furthermore, as:

$$F_{\alpha\beta}F^{\alpha\beta}=2\left(c^2\|\vec{B}\|^2-\|\vec{E}\|^2\right) \tag{12.4.238}$$

We have:

$$c^2\|\vec{B}\|^2=\frac{1}{2}F_{\alpha\beta}F^{\alpha\beta}+\|\vec{E}\|^2=\frac{1}{2}F_{\alpha\beta}F^{\alpha\beta}+\frac{1}{c^2}F_{\beta\alpha}u^\alpha F_\gamma^\beta u^\gamma \tag{12.4.239}$$

So we have so far:

$$\begin{aligned}\|\vec{E}\|^2&=\frac{1}{c^2}F_{\beta\alpha}u^\alpha F_\gamma^\beta u^\gamma\\ c^2\|\vec{B}\|^2&=\frac{1}{2}F_{\alpha\beta}F^{\alpha\beta}+\frac{1}{c^2}F_{\beta\alpha}u^\alpha F_\gamma^\beta u^\gamma\end{aligned} \tag{12.4.240}$$

Putting these together in :

$$u_{\text{tot}}=\frac{1}{2}\varepsilon\hat{E}^2+\frac{1}{2\mu}\hat{B}^2=\frac{\varepsilon}{2}\left(\hat{E}^2+\frac{1}{\mu\varepsilon}\hat{B}^2\right)=\frac{\varepsilon}{2}\left(\hat{E}^2+c^2\hat{B}^2\right) \tag{12.4.241}$$

Then:

$$u_{\text{tot}}=\frac{\varepsilon}{2}\left(\frac{1}{c^2}F_{\beta\alpha}u^\alpha F_\gamma^\beta u^\gamma+\left(\frac{1}{2}F_{\alpha\beta}F^{\alpha\beta}+\frac{1}{c^2}F_{\beta\alpha}u^\alpha F_\gamma^\beta u^\gamma\right)\right)=\varepsilon\left(\frac{1}{c^2}F_{\beta\alpha}u^\alpha F_\gamma^\beta u^\gamma+\frac{1}{4}F_{\alpha\beta}F^{\alpha\beta}\right) \tag{12.4.242}$$

Putting these together, and inserting a factor of :

$$-g_{\alpha\gamma}u^\alpha u^\gamma=c^2\Leftrightarrow-\frac{g_{\alpha\gamma}}{c^2}u^\alpha u^\gamma=1 \tag{12.4.243}$$

gives the energy density:

$$\begin{aligned}u_{\text{tot}}&=\varepsilon\left(\frac{1}{c^2}F_{\beta\alpha}u^\alpha F_\gamma^\beta u^\gamma-\frac{1}{4}\frac{g_{\alpha\gamma}}{c^2}u^\alpha u^\gamma F_{\alpha\beta}F^{\alpha\beta}\right)=\frac{\varepsilon}{c^2}\left(F_{\beta\alpha}F_\gamma^\beta-\frac{1}{4}g_{\alpha\gamma}F_{\alpha\beta}F^{\alpha\beta}\right)u^\alpha u^\gamma\\ &=\frac{1}{\mu}\left(F_{\beta\alpha}F_\gamma^\beta-\frac{1}{4}g_{\alpha\gamma}F_{\alpha\beta}F^{\alpha\beta}\right)u^\alpha u^\gamma\end{aligned} \tag{12.4.244}$$

Since in general as we have seen just above $u_{\text{tot}}=T_{\alpha\beta}u^\alpha u^\beta$ then the above relation can hold for all rest frames only if:

$$\boxed{T_{\alpha\gamma}=\frac{1}{\mu}\left(F_{\beta\alpha}F_\gamma^\beta-\frac{1}{4}g_{\alpha\gamma}F_{\delta\beta}F^{\delta\beta}\right)} \tag{12.4.245}$$

This is the "electromagnetic stress-energy tensor".

But we should also check that this result is symmetric. Then if we do the calculations (boring elementary algebra but we can give the details on readers request):

$$T^{\mu\nu}=\begin{bmatrix}\frac{1}{2}\left(\varepsilon_0E^2+\frac{1}{\mu_0}B^2\right) & S_{\text{x}}/c & S_{\text{y}}/c & S_{\text{z}}/c\\ S_{\text{x}}/c & -\sigma_{xx} & -\sigma_{\text{xy}} & -\sigma_{\text{xz}}\\ S_{\text{y}}/c & -\sigma_{yx} & -\sigma_{\text{yy}} & -\sigma_{\text{yz}}\\ S_{\text{z}}/c & -\sigma_{zx} & -\sigma_{\text{zy}} & -\sigma_{\text{zz}}\end{bmatrix} \tag{12.4.246}$$

where for recall:

$$\vec{S}=\frac{1}{\mu_0}\vec{E}\times\vec{B}=\frac{1}{\mu}\begin{pmatrix}E_yB_z-E_zB_y\\E_zB_x-E_xB_z\\E_xB_y-E_yB_x\end{pmatrix} \tag{12.4.247}$$

and:

$$\sigma_{ij}=\varepsilon_0E_iE_j+\frac{1}{\mu_0}B_iB_j-\frac{1}{2}\left(\varepsilon_0E^2+\frac{1}{\mu_0}B^2\right)\delta_{ij} \tag{12.4.248}$$

is the "Maxwell stress tensor".

12.4.5 Einstein's Field Equations

It is now time to tackle one of the most beautiful, one of the most famous equations of our time and that shines the eyes of many young students and science passionate: Einstein's field equations. The one that explains why matter (energy) curves space!!! There are several ways to obtain these equations. The two most common ones are either:

1. To have an engineer approach: That is to say we proceed by comparison with a known limiting result which is the law of gravitation of Newton (it is the one that we have chosen)
2. To have a pure mathematical approach (very elegant but a little fallen from the sky with some circular reasoning): That is to say that we use the Lagrangian formalism and seek by trial and errors a Lagrangian density which allows us to fall back on something known.

Well this having been said, let us recall before starting some results that we have obtained so far. First, we have succeeded in proving brilliantly that every particle (assumed to be free but left to interpretation ... in a curved space ...) follows the equation of motion of geodesics:

$$\frac{\mathrm{d}^2x^\mu}{\mathrm{d}s^2}+\Gamma^\mu_{\alpha\beta}\frac{\mathrm{d}x^\alpha}{\mathrm{d}s}\frac{\mathrm{d}x^\beta}{\mathrm{d}s} \tag{12.4.249}$$

In the section of Tensor Calculus, we have proved (not without difficulty...) what we name the "Einstein's tensor" (which is a constant in a given Riemannian space) is given by:

$$G^{\mu\nu}=R^{\mu\nu}-\frac{1}{2}g^{\mu\nu}R \tag{12.4.250}$$

where $R^{\mu\nu}$ is for recall the Ricci tensor (see section Tensor Calculus page 2127).

Since the covariant derivative of the Einstein's tensor is zero (see section Tensor Calculus page 2141) and we have proved that the covariant derivative of stress-energy tensor is also, then it is tempting to put:

$$\nabla^\rho G_{\rho\sigma}=\kappa\nabla^\rho T_{\rho\sigma}=0 \tag{12.4.251}$$

where κ is a normalization constant (commonly named: "Einstein's gravitational coupling constant") and must satisfy the relation so that it is homogeneous at the level of the units. So it comes (we should better say: "we think we can write...") after simplification:

$$G_{\rho\sigma}=\kappa T_{\rho\sigma} \tag{12.4.252}$$

To find the expression of the constant, we will place ourselves in the Newtonian limit and request that the preceding relation reproduce the Newton-Poisson equation for the gravitational potential φ (see section Astronomy page 5266):

$$\nabla^2\varphi=4\pi G\rho \tag{12.4.253}$$

Remark

This relation shows that the gravitational potential is connected to the matter density linearly through its second derivatives. Albert Einstein thought, therefore, that the first member of the equations of the field in General Relativity, member supposed to describe the geometry of space-time, must therefore somehow include the second derivatives, not of the gravitational potential, but of the potentials of the metric. In fact, Albert Einstein tried to generalize the right-hand side of the Poisson equation: the desired quantity must include not only the density of matter but also the momentum (as soon as the body is moving, its energy increases and therefore its mass). To evaluate the gravitational effect of a body, it was therefore necessary to combine its mass at rest with its momentum. It was finally the stress-energy tensor of rank 2 which is the generalization of the quadrivector momentum of Special Relativity.

We have proved earlier above that in the Newtonian limit (weak field approximation):

$$h_{00}=-\frac{2}{c^2}\varphi \qquad (12.4.254)$$

and in our definition of stress-energy tensor, for a distribution of matter at rest (or in a coordinate frame according to...) only the following component is non-zero:

$$T_{00}=\rho c^2 \qquad (12.4.255)$$

It follows that the Poisson equation can be written (notice that this assumes that the density is calculated from a three dimensional sphere of volume $V=4/3\pi r^3$...):

$$\nabla^2 h_{00}=\Delta h_{00}=\frac{\mathrm{d}^2}{\mathrm{d}^2 r}\left(-\frac{2}{c^2}\frac{2GM}{r}\right)=-\frac{8GM}{c^2r^3}=-\frac{8\pi G\rho}{c^2}=-\frac{8\pi G}{c^2}\frac{T_{00}}{c^2}=-\frac{8\pi G}{c^4}T_{00} \qquad (12.4.256)$$

Now let us return to the relation:

$$G_{\rho\sigma}=\kappa T_{\rho\sigma} \qquad (12.4.257)$$

By contracting the two members of the preceding relation, it comes:

$$g^{\rho\sigma}G_{\rho\sigma}=\kappa g^{\rho\sigma}T_{\rho\sigma} \qquad (12.4.258)$$

that is to say more explicitly (see section Tensor Calculus page 2141):

$$g^{\rho\sigma}\left(R_{\rho\sigma}-\frac{1}{2}Rg_{\rho\sigma}\right)=\kappa g^{\rho\sigma}T_{\rho\sigma} \qquad (12.4.259)$$

But, the Ricci scalar (see section Tensor Calculus page 2128) is given by:

$$R=g^{\rho\sigma}R_{\rho\sigma} \qquad (12.4.260)$$

It comes therefore:

$$\left(R-\frac{1}{2}Rg^{\rho\sigma}g_{\rho\sigma}\right)=\kappa g^{\rho\sigma}T_{\rho\sigma} \qquad (12.4.261)$$

Now in the special case of the Minkowski metric (with the signature $(-,+,+,+)$) it is immediate that:

$$g^{\rho\sigma}g_{\rho\sigma}=\eta^{\rho\sigma}\eta_{\rho\sigma}=4 \tag{12.4.262}$$

Therefore (implicitly we continue with the Minkowski metric!):

$$R-2R=\kappa g^{\rho\sigma}T_{\rho\sigma} \tag{12.4.263}$$

Using this last relation, the equation:

$$R_{\rho\sigma}=\kappa T_{\rho\sigma}+\frac{1}{2}Rg_{\rho\sigma} \tag{12.4.264}$$

can finally be written:

$$R_{\rho\sigma}=\kappa T_{\rho\sigma}+\frac{1}{2}(-\kappa g^{\rho\sigma}T_{\rho\sigma})g_{\rho\sigma}=\kappa\left(T_{\rho\sigma}-\frac{1}{2}g^{\rho\sigma}T_{\rho\sigma}g_{\rho\sigma}\right) \tag{12.4.265}$$

Let us focus on the component $\rho=\sigma=0$ (not to be confused with the notation of shear stress and density!!!) such that the preceding relation is written:

$$\begin{aligned}R_{00}&=\kappa\left(T_{00}-\frac{1}{2}(g^{00}T_{00})g_{00}\right)=\kappa\left(T_{00}-\frac{1}{2}((-1)\cdot T_{00})(-1)\right)\\&=\kappa\left(T_{00}-\frac{1}{2}T_{00}\right)=\kappa\frac{1}{2}T_{00}\end{aligned} \tag{12.4.266}$$

Let us write explicitly this last relation by using the definition of the Ricci tensor (see section Tensor Calculus page 2127) that is for recall:

$$R_{\alpha\beta}=R^{\mu}_{\alpha\mu\beta} \tag{12.4.267}$$

Then it comes:

$$R_{00}=R^{\mu}_{0\mu0} \tag{12.4.268}$$

But, the Riemann-Christoffel tensor developed in this particular case is given for recall by (see section Tensor Calculus page 2122):

$$R^{\mu}_{0,\mu0}=R^{\mu}_{0\mu0}=\partial_{\mu}\Gamma^{\mu}_{00}-\partial_{0}\Gamma^{\mu}_{0\mu}+\Gamma^{\gamma}_{00}\Gamma^{\mu}_{\mu\gamma}-\Gamma^{\gamma}_{0\mu}\Gamma^{\mu}_{0r} \tag{12.4.269}$$

Remark

In the absence of a gravitational field and in Cartesian coordinates, it is logical that all the Christoffel symbols are null. Indeed, the Christoffel symbols translate nothing more than the forces of inertia. But when we have a field of gravitation, the trajectories followed are no longer straight lines, even in the Newtonian case, then the Christoffel symbols are non-zero.

In the approximation of the weak field slowly variable over time, the Christoffel symbols are of order $\mathcal{O}^1$ and their products are of order $\mathcal{O}^2$ and the temporal derivatives are negligible in front of the spatial derivatives. It therefore remains only the terms of order $\mathcal{O}^1$ such that:

$$R_{00}=R^{\mu}_{0\mu0}\cong\partial_{\mu}\Gamma^{\mu}_{00} \tag{12.4.270}$$

But, we have proved in the section of Tensor Calculus (see page 2109) that:

$$\Gamma_{kj}^{i}=\frac{1}{2}g^{il}\left(\partial_k g_{lj}+\partial_j g_{kl}-\partial_l g_{kj}\right) \tag{12.4.271}$$

Since then:

$$R_{00}\cong\partial_\mu\Gamma_{00}^{\mu}=\partial_\mu\left(\frac{1}{2}g^{\mu l}\left(\partial_0 g_{l0}+\partial_0 g_{0l}-\partial_l g_{00}\right)\right) \tag{12.4.272}$$

But in the weak field approximation, the variation of the metric with respect to time is negligible compared to the spatial variation (the approximation is somewhat pulled by the hair it must be said ...):

$$\begin{aligned}R_{00}&\cong\partial_\mu\Gamma_{00}^{\mu}=\partial_\mu\left(\frac{1}{2}g^{\mu l}(-\partial_l g_{00})\right)=-\frac{1}{2}\partial_\mu\left[(\eta^{\mu l}+h^{\mu l})\partial_l(\eta_{00}+h_{00})\right]\\&\cong-\frac{1}{2}\partial_\mu(\eta^{\mu l}\partial_l h_{00})=-\frac{1}{2}\eta^{\mu l}\partial_\mu\partial_l h_{00}-\frac{1}{2}(\partial_l h_{00})(\partial_\mu\eta^{\mu l})\cong-\frac{1}{2}\eta^{\mu l}\partial_\mu\partial_l h_{00}\\&=-\frac{1}{2}\nabla^2 h_{00}\end{aligned} \tag{12.4.273}$$

Therefore, equating the both relations:

$$R_{00}=\kappa\frac{1}{2}T_{00}\qquad\text{and}\qquad R_{00}=-\frac{1}{2}\nabla^2 h_{00} \tag{12.4.274}$$

becomes (after the simplification by 2):

$$\nabla^2 h_{00}=-\kappa T_{00} \tag{12.4.275}$$

and we immediately notice that we fall back on:

$$\nabla^2 h_{00}=-\frac{8\pi G}{c^4}T_{00} \tag{12.4.276}$$

if and only if:

$$\kappa=\frac{8\pi G}{c^4} \tag{12.4.277}$$

Constant which is sometimes named "Einstein's constant". It follows immediately that the Ricci scalar is positive and therefore that we are locally in a spherical curvature space.

The "Einstein's field equations" (EFE) is therefore in definitive form:

$$G_{\mu\nu}=\frac{8\pi G}{c^4}T_{\mu\nu} \tag{12.4.278}$$

or more conventionally:

$$\boxed{R_{\mu\nu}-\frac{1}{2}Rg_{\mu\nu}=\frac{8\pi G}{c^4}T_{\mu\nu}} \tag{12.4.279}$$

Or in explicit form:

$$\left(\frac{\mathrm{d}\Gamma^{\sigma}_{\mu\nu}}{\mathrm{d}x^{\sigma}}+\Gamma^{\lambda}_{\mu\nu}\Gamma^{\sigma}_{\lambda\sigma}-\frac{\mathrm{d}\Gamma^{\sigma}_{\mu\sigma}}{\mathrm{d}x^{\nu}}-\Gamma^{\lambda}_{\mu\sigma}\Gamma^{\sigma}_{\lambda\nu}\right)-\frac{1}{2}g^{\mu\nu}\left(\frac{\mathrm{d}\Gamma^{\sigma}_{\mu\nu}}{\mathrm{d}x^{\sigma}}+\Gamma^{\lambda}_{\mu\nu}\Gamma^{\sigma}_{\lambda\sigma}-\frac{\mathrm{d}\Gamma^{\sigma}_{\mu\sigma}}{\mathrm{d}x^{\nu}}-\Gamma^{\lambda}_{\mu\sigma}\Gamma^{\sigma}_{\lambda\nu}\right)g_{\mu\nu}$$
$$=\frac{8\pi G}{c^4}\begin{bmatrix}\rho\gamma^2c^2 & \rho\gamma^2cv^x & \rho\gamma^2cv^y & \rho\gamma^2cv^z\\ \rho\gamma^2cv^x & \rho\gamma^2v^xv^x & \rho\gamma^2v^xv^y & \rho\gamma^2v^xv^z\\ \rho\gamma^2cv^y & \rho\gamma^2v^yv^x & \rho\gamma^2v^yv^y & \rho\gamma^2v^yv^z\\ \rho\gamma^2cv^z & \rho\gamma^2v^zv^x & \rho\gamma^2v^zv^y & \rho\gamma^2v^zv^z\end{bmatrix} \tag{12.4.280}$$

The left-hand part represents the curvature of space-time as determined by the metric and the right-hand expression represents a modelization of the space-time content of mass / energy. This equation can then be interpreted as a set of equations describing how the curvature of space-time is related to the mass-energy content of the Universe. These equations, as well as the geodesic equation, form the core of the mathematical formulation of General Relativity.

The EFE is therefore a dynamic equation describing how matter and energy modify the geometry of space-time. This curvature of the geometry around a source of matter is then interpreted as the gravitational field of this source. The movement of objects in this field is described very precisely by the equation of its geodesic.

Similar to the way that electromagnetic fields are determined using charges and currents via Maxwell's equations, the EFE are used to determine the space-time geometry resulting from the presence of mass–energy and linear momentum, that is, they determine the metric tensor of space-time for a given arrangement of stress–energy in the space-time.

The mainstream metrics that we will study in this textbook will be (the first one is already knows to us): the Minkowski metric for flat static space, the Schwarschild metric for curved static space, the Friedmann-Lemaître-Robertson-Walker to describe the Universe metric, the Kerr metric for dynamic curve space and the Morris-Thorne metric for worm-holes:

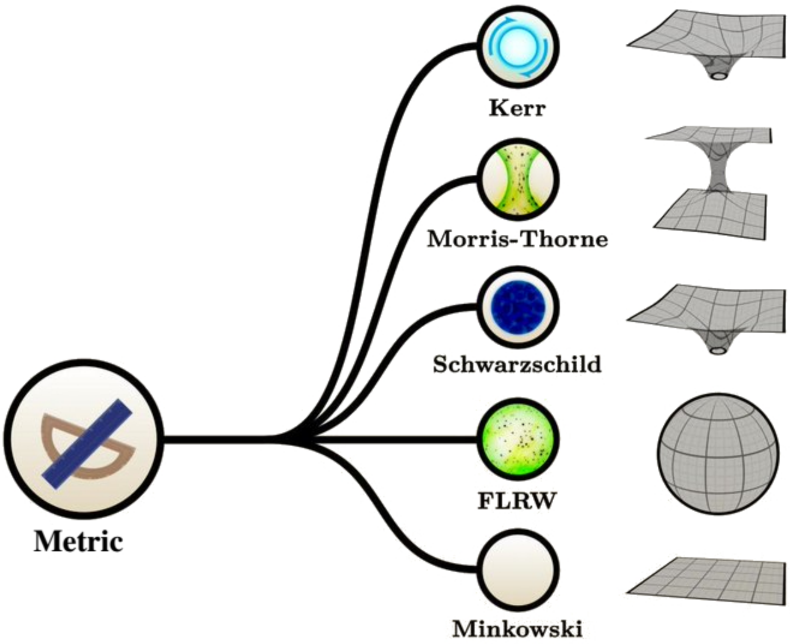


Figure 12.219: Some mainstream metrics used in General Relativity

Remark

Keep in mind that the perceived "force" of gravity is a manifestation of space-time curvature, not gravity itself! Gravity is what causes space-time to curve due to energy/mass, and the hypothesized gravitons could be the mechanism for this. In some Quantum Gravitation theories gravity is still a force, it's just the force that causes space-time curvature and not what causes gravitational attraction (not directly anyway)!

On the other hand, we have just seen that Einstein's equation reduces to the laws of Newton's gravity by using the approximation of weak fields and slow movements.

These differential equations are in general a nightmare to solve, the Ricci scalars and tensors are contractions of the Riemann tensor, which include the derivatives and products of the Christoffel symbols, which are themselves constructed on the inverse metric tensor and on the derivatives of it. To compute the whole, it is possible to construct energy-momentum tensors that can invoke the metric as well. It is therefore very difficult to solve the Albert Einstein equations of fields in the general case. Exact solutions for the EFE can only be found under simplifying assumptions such as symmetry. Special classes of exact solutions are most often studied as they model many gravitational phenomena, such as rotating Black Holes and the expanding universe. Further simplification is achieved in approximating the actual space-time as flat space-time with a small deviation, leading to the linearised EFE. These equations are used to study phenomena such as gravitational waves.

Since the stress-energy tensor has $4\cdot 4=16$ components, 10 of which are actually unique (independent) since the tensor is symmetric (triangle + diagonal), we can see the Einstein equation of the fields as 10 second-order differential equations coupled on field tensor metric g_{ij}.

Some people are confused about how the curvature of space-time and gravity are related. I am going to explain mainly that starting with simpler examples, and moving to more complicated ones.

Okay, let's say we have a sheet of rubber. This is the classic example of space-time. Let's say we take a bowling ball, and set it on the taut sheet of rubber. It has a large mass (compared to what else we'll be putting on the sheet), therefore the sheet curves a lot for the bowling ball. We now have an image in our head like the one below:

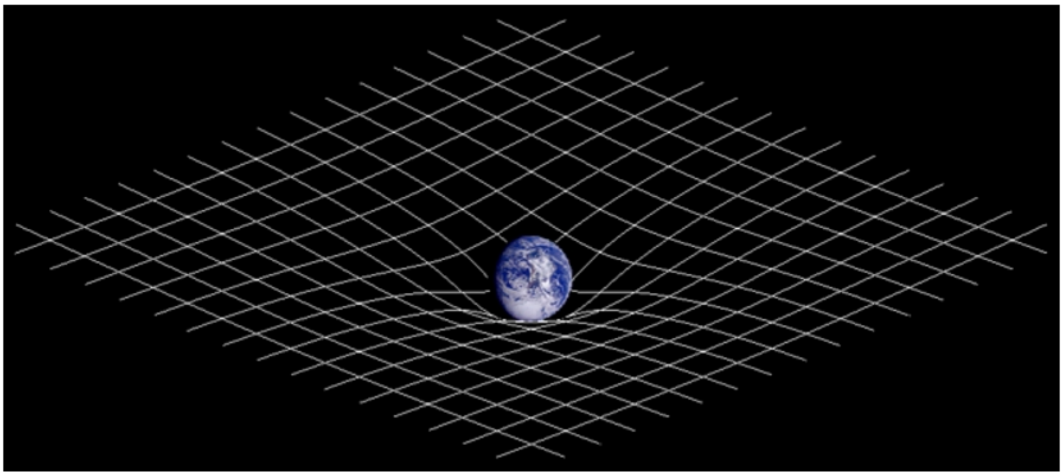


Figure 12.220: **2D naive representation of space curve near Earth**

So mass leads to curvature. Then, let us take a baseball, say, and set it near the bowling ball. It rolls toward the bowling ball, right? This occurs because of the curvature of the sheet. So, then, curvature leads to gravity. So, if an object has large mass, it will curve space-time dramatically, leading to strong gravity.

This is, of course, an overly simplistic example. It is 2D, and it doesn't take into account other factors. Let us move to 3D (keeping in mind the Universe is accepted to be at 4D, ignoring the holographic principle). The mass of a bowling ball now sucks in space around it, sort of like in the picture below:

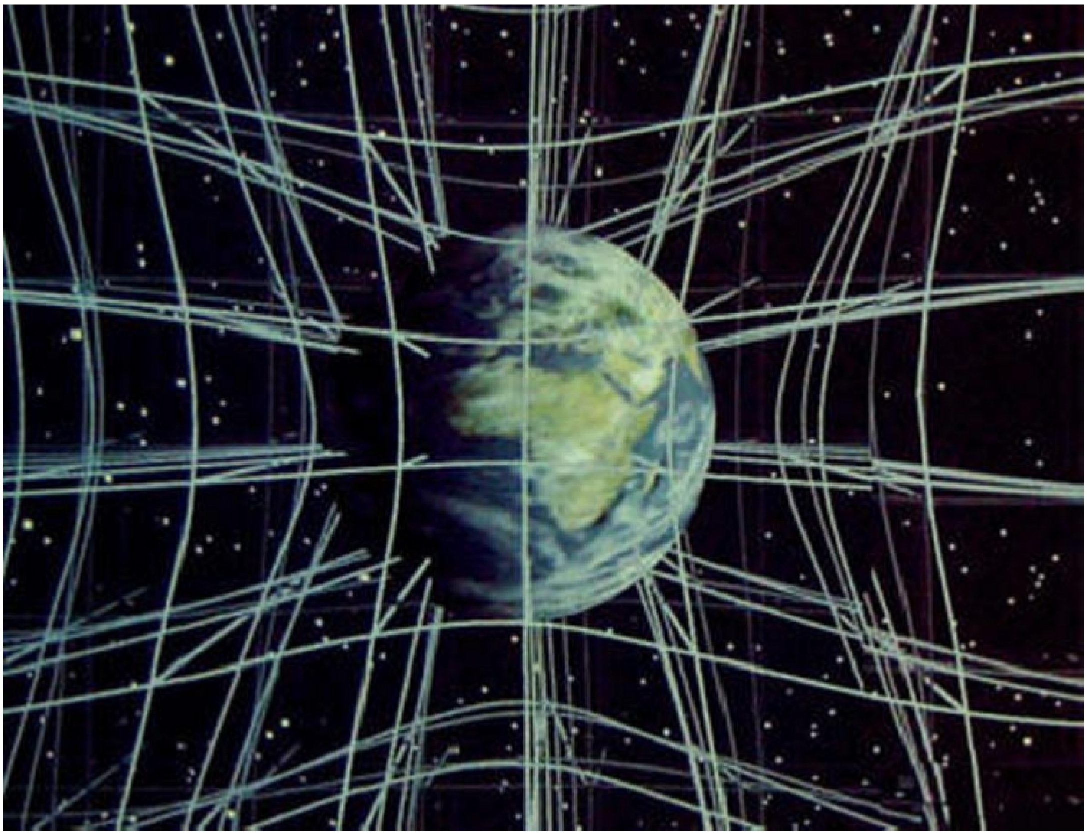


Figure 12.221: 3D naive representation of curved space near Earth

And now, in this case, we can see (or understand) that more mass still leads to more curvature. The greater the mass, the more space-time will "contract" around the object. So we still think that mass leads to curvature. Now, if we set an object near this massive object (like the Moon next to Earth) it is "sucked in" sort of, by the curvature of space-time, though of course the Moon contracts space-time around it as well. At this point, we can reasonably still conclude that in 3D, mass leads to curvature which leads to gravity.

A quick glance at the constant of proportionality in the Einstein's field equations gives one a rough feeling of much stress-energy is needed to curve space. In SI units, the gravitational constant G is about $6.67\cdot 10^{-11}\,[\text{m}^3\cdot\text{kg}^{-1}\cdot\text{s}^{-2}]$ while the speed for light c is approximately $3.00\cdot 10^{8}\,[\text{m}\cdot\text{s}^{-1}]$. The Einstein's field equations is then in explicit numerical value given by:

$$\boxed{G_{\mu\nu}=\frac{8\pi G}{c^4}T_{\mu\nu}\cong 2.08\cdot 10^{-43}T_{\mu\nu}} \qquad (12.4.281)$$

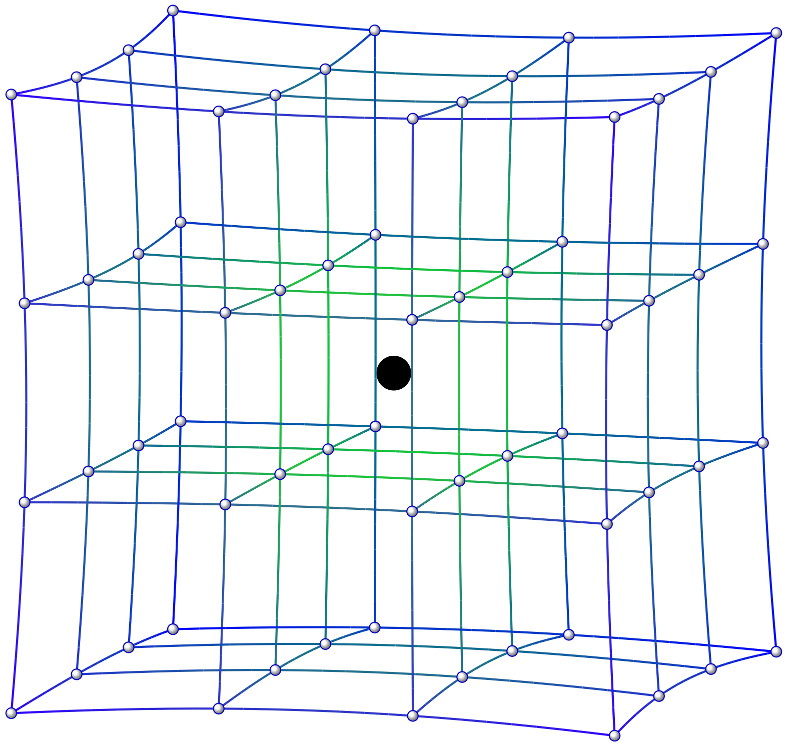


Figure 12.222: **3D naive representation of space curve with TikZ**

The Sun (actually a G2V-type star, aka yellow dwarf) has an average mass-energy density (the dominant component of the stress-energy tensor) of $T^{00} \cong 1.27 \cdot 10^{20}\,[\mathrm{kg} \cdot \mathrm{m}^{-1} \cdot \mathrm{s}^{-2}]$. The corresponding component of the Einstein Tensor is therefore $G_{00} \cong 2.64 \cdot 10^{-23}\,[\mathrm{m}^{-2}]$. By comparison the Einstein tensor for the flat Minkowski metric is identically zero. So to see a curvature we need to look at hyper-energetic phenomena, like a collapsing star, to find an Einstein tensor component appreciable greater than this. Even though the space-time metric $g_{\mu\nu}$ is not generally flat, throughout most of the universe it is flat enough to be considered as small perturbation of a flat background metric:

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu} \qquad (12.4.282)$$

But, as I said earlier, the Universe is generally thought of as 4D. What does our picture look like when we add time? Well, the time dimension is contracted around a massive object. So let us picture our previous example but that the fabric of space-time has a few clocks embedded in it occasionally. As the space stretches and contracts, so will the clocks (the "time") and so the time on those clocks will be "wrong" - it'll differ from the other clocks. And in this case, as the Earth contracts space and time around it, it changes the time and space (it curves space-time) and so when another object enters our region of space-time, it is "sucked in" still, but so is it's time. This is, of course, a very extreme example, but I hope this shows that we can conclude that mass leads to curvature which leads to gravity.

12.4.5.1 Cosmological Constant (CC)

Albert Einstein modified his original field equations to include a cosmological constant term Λ proportional to the metric that led afterwards the Universe model to be static (see section Cosmogony page 6012).

To see how this constant was introduced let us recall that we have proved so far that:

$$\nabla^\rho G_{\rho\sigma}=\frac{8\pi G}{c^4}\nabla^\rho T_{\rho\sigma} \tag{12.4.283}$$

or more explicitly:

$$\nabla^\rho\left(R_{\rho\sigma}-\frac{1}{2}g_{\rho\sigma}R\right)=\frac{8\pi G}{c^4}\nabla^\rho T_{\rho\sigma} \tag{12.4.284}$$

That is to say:

$$\nabla^\rho R_{\rho\sigma}-\frac{1}{2}\nabla^\rho g_{\rho\sigma}R=\frac{8\pi G}{c^4}\nabla^\rho T_{\rho\sigma} \tag{12.4.285}$$

But we have proved in the section of Tensor Calculus that the covariant derivative kills the metric (see page 2117), that is to say for recall:

$$\nabla^\rho g_{\rho\sigma}=0 \tag{12.4.286}$$

Therefore if we choose a constant Λ the latter relation can also we written:

$$\Lambda\nabla^\rho g_{\rho\sigma}=\Lambda\cdot 0 \tag{12.4.287}$$

Obviously:

$$\nabla^\rho\left(\Lambda g_{\rho\sigma}\right)=0 \tag{12.4.288}$$

So nothing prevents us to put this covariant derivative in:

$$\nabla^\rho R_{\rho\sigma}-\frac{1}{2}\nabla^\rho g_{\rho\sigma}R=\frac{8\pi G}{c^4}\nabla^\rho T_{\rho\sigma} \tag{12.4.289}$$

as we can write:

$$\nabla^\rho R_{\rho\sigma}-\frac{1}{2}\nabla^\rho g_{\rho\sigma}R+0=\frac{8\pi G}{c^4}\nabla^\rho T_{\rho\sigma} \tag{12.4.290}$$

and replacing the 0 by the covariant derivative of the metric:

$$\nabla^\rho R_{\rho\sigma}-\frac{1}{2}\nabla^\rho g_{\rho\sigma}R+\nabla^\rho\left(\Lambda g_{\rho\sigma}\right)=\frac{8\pi G}{c^4}\nabla^\rho T_{\rho\sigma} \tag{12.4.291}$$

After factorization we get:

$$\nabla^\rho\left(R_{\rho\sigma}-\frac{1}{2}g_{\rho\sigma}R+\Lambda g_{\rho\sigma}\right)=\frac{8\pi G}{c^4}\nabla^\rho T_{\rho\sigma} \tag{12.4.292}$$

And simplifying we get the "general Einstein's field equations":

$$\boxed{\underbrace{R_{\rho\sigma}-\frac{1}{2}g_{\rho\sigma}R+\Lambda g_{\rho\sigma}}_{\text{pure geometry}}=\underbrace{\frac{8\pi G}{c^4}T_{\rho\sigma}}_{\text{properties of matter}}} \tag{12.4.293}$$

where Λ is the so named "cosmological constant" (a.k.a. "dark energy" in contemporary physics). It is also many times written as following:

$$G_{\mu\nu}+\Lambda g_{\mu\nu}=\kappa T_{\mu\nu} \tag{12.4.294}$$

Remember that in the equation above we have the stress-energy tensor $T_{\rho\sigma}$ that is in $[\mathrm{J}\cdot\mathrm{m}^{-3}]$, the metric $g_{\rho\sigma}$ that is dimensionless $[-]$ and the Ricci tensor $R_{\rho\sigma}$ but also the cosmological constant Λ and the Ricci scalar R that are all three in $[\mathrm{m}^{-1}]$. And remember that G has for units $[\mathrm{kg}^{-1}\cdot\mathrm{m}^3\cdot\mathrm{s}^{-2}]$ (hence $8\pi Gc^{-4}$ is in $[\mathrm{J}\cdot\mathrm{m}^{-1}]$.

Caution!!! Side note, with $T_{\mu\nu}=0$ and $\Lambda=0$, you get $R_{\mu\nu}=\frac{1}{2}Rg_{\mu\nu}$, which only equals zero if and only if $R=R_{\mu}^{\mu}=0$. The fact that the energy-momentum tensor is named the "source of curvature" doesn't mean that there can't be any curvature where there is no energy-momentum. In fact, even if $T_{\mu\nu}=0$ across all space-time, there are still non-trivial solutions of Einstein's equations, in the form of gravitational waves. It is automatic that any Ricci-flat manifold has zero scalar curvature; the best-known spaces in this class are the Calabi-Yau manifolds. There are also metrics with zero scalar curvature but non-vanishing Ricci curvature!

The equations are symmetric; that is, $\mu\nu=\nu\mu$. So the ten equations we have mentioned earlier above are:

1. $G_{00}+\Lambda g_{00}=\kappa T_{00}$
2. $G_{01}+\Lambda g_{01}=\kappa T_{01}$
3. $G_{02}+\Lambda g_{02}=\kappa T_{02}$
4. $G_{03}+\Lambda g_{03}=\kappa T_{03}$
5. $G_{11}+\Lambda g_{11}=\kappa T_{11}$
6. $G_{12}+\Lambda g_{12}=\kappa T_{12}$
7. $G_{13}+\Lambda g_{13}=\kappa T_{13}$
8. $G_{22}+\Lambda g_{22}=\kappa T_{22}$
9. $G_{23}+\Lambda g_{23}=\kappa T_{23}$
10. $G_{33}+\Lambda g_{33}=\kappa T_{33}$

The cosmological constant may be thought as the energy density associated with vacuum, the space absolutely void of particles. This could be a kind of "ground level energy", which often appears in quantum physics. In fact, some theories of elementary particles predict cosmological constant but, unfortunately, of much higher value compared to the actual observations (but the actual observation may be not accurate!). Given this association, the term "dark energy" is often used to describe the origin of the cosmological constant.

Remark

So General Relativity doesn't tell us anything about the value of the cosmological constant. In quantum field theory, however, we can calculate the vacuum energy density and it comes out to be infinitely large. But in the absence of gravity this doesn't matter: we never measure absolute energies anyway, we merely measure energy differences.

The latter relation can be found also in many textbooks in natural units (see section Principia page 3124) and rearranged a little bit as following:

$$R_{\rho\sigma}-\frac{1}{2}g_{\rho\sigma}R=8\pi T_{\rho\sigma}-\Lambda g_{\rho\sigma} \tag{12.4.295}$$

The effort from Albert Einstein to introduce this constant was unsuccessful because:

- The universe described by this theory was unstable
- Observations by Edwin Hubble confirmed that our Universe is expanding

So, Albert Einstein abandoned Λ, calling it the "biggest blunder [he] ever made".

Despite Albert Einstein's motivation for introducing the cosmological constant term, there is nothing inconsistent with the presence of such a term in the equations. For many years the cosmological constant was almost universally considered to be 0. However, recent improved astronomical techniques have found that a positive value of Λ is needed to explain observations that seems to give an accelerating universe.

Let us now rearrange the Einstein Field Equations with the cosmological constant by rearranging it a bit:

$$R_{\rho\sigma}-\frac{1}{2}g_{\rho\sigma}R+=\frac{8\pi G}{c^4}\left(T_{\rho\sigma}-\frac{\Lambda c^4}{8\pi G}g_{\rho\sigma}\right) \tag{12.4.296}$$

Hence the expression of the stress-impulsion tensor for vacuum (i.e. the energy density for recall!):

$$T_{\rho\sigma}^{\text{vacuum}}=\frac{\Lambda c^4}{8\pi G}g_{\rho\sigma} \tag{12.4.297}$$

If we compare this expression with that of a perfect fluid obtained earlier above:

$$T^{ij}\cong\left(\rho+\frac{P}{c^2}\right)u^iu^j-Pg^{ij} \tag{12.4.298}$$

We may assimilate vacuum to the pressure term, hence as a fluid of pressure:

$$P=-\frac{\Lambda c^4}{8\pi G}g_{\rho\sigma} \tag{12.4.299}$$

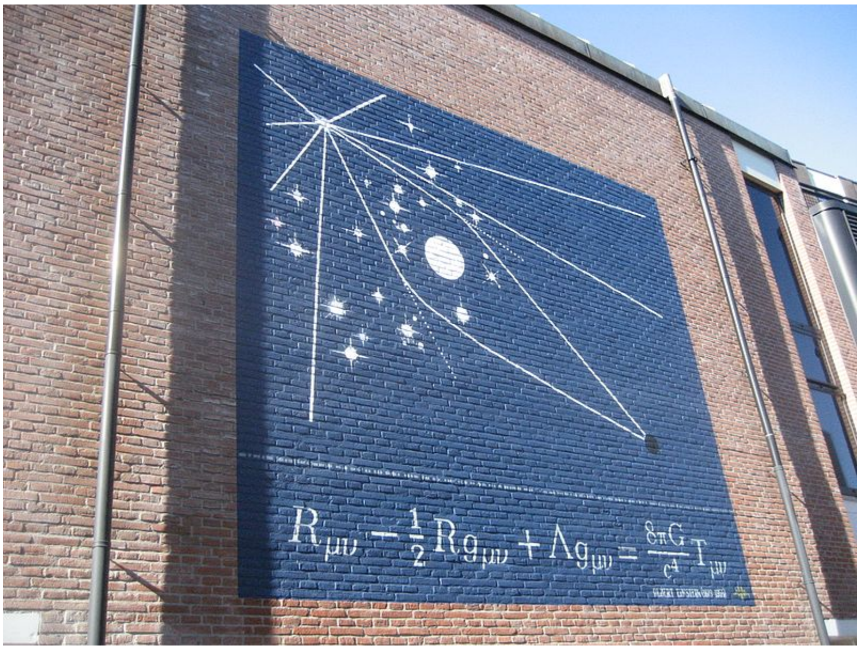

$R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} + \Lambda g_{\mu\nu} = \frac{8\pi G}{c^4}T_{\mu\nu}$

Figure 12.223: **Diagram of gravitational lensing with formula of Albert Einstein on a wall of Museum Boerhaave, Leiden in Netherlands (source: Wikipedia, author: Stichting Tegenbeeld, photograph: Vysotsky)**

We can also trace invert the Einstein Field Equations (in some cases it helps to solve problems in an easier way!). We start from the EFE:

$$R_{\mu\nu}-\frac{1}{2}Rg_{\mu\nu}+\Lambda g_{\mu\nu}=\frac{8\pi G}{c^4}T_{\mu\nu} \quad (12.4.300)$$

and let us recall that:

$$g_{\mu\nu}g^{\mu\nu}=\delta_{\nu}^{\nu}=4 \quad (12.4.301)$$

So multiplying the EFE both sides by $g^{\mu\nu}$ leads us to:

$$R-\frac{1}{2}R\cdot 4+\Lambda\cdot 4=\frac{8\pi G}{c^4}T \quad (12.4.302)$$

Multiplying both sides by $-\frac{1}{2}g_{\mu\nu}$ we get:

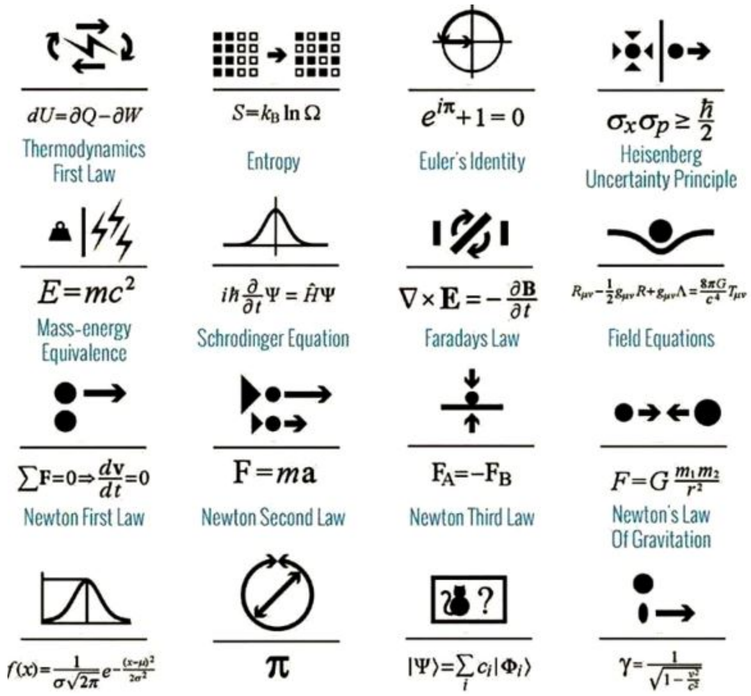
$$-\frac{1}{2}g_{\mu\nu}R=\frac{1}{2}g_{\mu\nu}\frac{8\pi G}{c^4}T-2g_{\mu\nu}\Lambda \quad (12.4.303)$$

where $T=T_{\mu\nu}g^{\mu\nu}$.

Subtracting the EFE gives us finally:

$$\boxed{R_{\mu\nu}=\frac{8\pi G}{c^4}\left(T_{\mu\nu}-\frac{1}{2}Tg_{\mu\nu}\right)+g_{\mu\nu}\Lambda} \quad (12.4.304)$$

If you had the courage (...) to read all the book so far in details, then now you can really understand the famous Internet illustration below as you have seen all the proofs of the relations visible on it:


$dU=\partial Q-\partial W$
Thermodynamics First Law
$S=k_B \ln \Omega$
Entropy
$e^{i\pi}+1=0$
Euler's Identity
$\sigma_x\sigma_p \geq \frac{\hbar}{2}$
Heisenberg Uncertainty Principle
$E=mc^2$
Mass-energy Equivalence
$i\hbar\frac{\partial}{\partial t}\Psi=\hat{H}\Psi$
Schrodinger Equation
$\nabla\times\mathbf{E}=-\frac{\partial\mathbf{B}}{\partial t}$
Faradays Law
$R_{\mu\nu}-\frac{1}{2}g_{\mu\nu}R+g_{\mu\nu}\Lambda=\frac{8\pi G}{c^4}T_{\mu\nu}$
Field Equations
$\sum\mathbf{F}=0 \Rightarrow \frac{d\mathbf{v}}{dt}=0$
Newton First Law
$\mathbf{F}=m\mathbf{a}$
Newton Second Law
$\mathrm{F_A}=-\mathrm{F_B}$
Newton Third Law
$F=G\frac{m_1m_2}{r^2}$
Newton's Law Of Gravitation
$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
π
$|\Psi\rangle=\sum_i c_i|\Phi_i\rangle$
?
$\gamma=\frac{1}{\sqrt{1-\frac{v^2}{c^2}}}$

Caution!!! There is no principle of physics yet, at least in this beginning of the 121st century (holocene calendar), that says physicals laws or constants have to be the same everywhere and always. Indeed, the reader must remember that the "laws" of physics are **descriptive**, not **imperative** and therefore they are some observed regularity in what we can observe experimentally with sufficiently wide occurrence or relevance! Asking whether the laws can change is no more and no less than asking whether regularities which have been observed up till now might not be quite as universal as was thought. The answer to that is: YES! This could drive the reader a bit mad. Indeed, how can we possibly do Science if the rules keep changing? But if you are a professional active scientist, you have to face up to that possibility - and you have a responsibility to check out if there is evidence beyond reasonable doubt about it! We will repeat this warning multiple times in this textbook as its seems that this is something necessary!

12.4.5.2 Weak field approximation with cosmological constant

As we have promised earlier above, we will see now how to fall back on the Newton-Poisson equation directly from the general Einstein's field equations. Obviously that means we just do the inverse reasoning that gave us the opportunity the build the classical Einstein's field equations!!! But they will however be a small difference... we will include the cosmological constant and see how it change the Newton-Poisson equation.

For this, we start obviously from:

$$R_{\rho\sigma}-\frac{1}{2}g_{\rho\sigma}R+\Lambda g_{\rho\sigma}=\frac{8\pi G}{c^4}T_{\rho\sigma} \tag{12.4.305}$$

Let us write these equations in the equivalent form:

$$R_{\rho\sigma}=\frac{8\pi G}{c^4}\left(T_{\rho\sigma}+\frac{1}{2}g_{\rho\sigma}R\right)-\Lambda g_{\rho\sigma} \tag{12.4.306}$$

In weak field and low speed we know that $T_{00}=\rho c^2$. But we will also assume that $R=T=\rho c^2$, so that we can write[47]:

$$R_{00}=\frac{8\pi G}{c^4}\left(\rho c^2+\frac{1}{2}g_{00}\rho c^2\right)-\Lambda g_{00} \tag{12.4.307}$$

But as we proved earlier during our introduction of the Newtonian approximation, we have:

$$R_{00}\cong-\frac{1}{2}\nabla^2 h_{00} \tag{12.4.308}$$

Therefore:

$$-\frac{1}{2}\nabla^2 h_{00}=\frac{8\pi G}{c^4}\left(\rho c^2+\frac{1}{2}(\eta_{00}+h_{00})\rho c^2\right)-\Lambda(\eta_{00}+h_{00}) \tag{12.4.309}$$

Hence (still using the $-,+,+,+$ metric signature):

$$-\frac{1}{2}\nabla^2 h_{00}=\frac{8\pi G}{c^4}\rho c^2\left(1+\frac{1}{2}(-1+h_{00})\right)-\Lambda(-1+h_{00}) \tag{12.4.310}$$

With the traditional Newtonian approximation:

$$h_{00}=-\frac{2}{c^2}\varphi \tag{12.4.311}$$

It gives :

$$\frac{1}{c^2}\nabla^2\varphi=\frac{8\pi G}{c^4}\rho c^2\left(1+\frac{1}{2}\left(-1-\frac{2}{c^2}\varphi\right)\right)-\Lambda\left(-1-\frac{2}{c^2}\varphi\right) \tag{12.4.312}$$

Multiplying both sides by c^2 we get after simplification:

$$\nabla^2\varphi=8\pi G\rho\left(\frac{1}{2}\left(1+\frac{2}{c^2}\varphi\right)\right)+\Lambda c^2\left(1+\frac{2}{c^2}\varphi\right) \tag{12.4.313}$$

Hence:

$$\nabla^2\varphi=4\pi G\rho\left(1+\frac{2}{c^2}\varphi\right)+\Lambda c^2\left(1+\frac{2}{c^2}\varphi\right) \tag{12.4.314}$$

in the approximation:

$$\left|\frac{2}{c^2}\varphi\right|\ll 1 \tag{12.4.315}$$

it gives the following Newtonian approximation:

$$\boxed{\nabla^2\varphi=4\pi G\rho+\Lambda c^2} \tag{12.4.316}$$

As $\Lambda\cdot c^2$, as we will see further below, is also very very small we shall also put it as equal to zero.

---

[47] In the weak field approximation the Ricci scalar should be equal to zero... but we do physics and not maths...

Remark

An inverse calculation then gives the following potential:

$$\varphi = \frac{-GM}{r} + \frac{\Lambda c^2 r^2}{2} \tag{12.4.317}$$

Obviously the conclusion depends on the fact if experimental data gives $\Lambda = 0$, $\Lambda < 0$ or $\Lambda > 0$! What we still don't really know with enough accuracy still in this early 121st century (holocene calendar) as its values is very near zero (order of 10^{-52})!

The interesting thing is that at the Galaxy scale (i.e. with galactical input numerical values), that latter results shows that if $\Lambda > 0$, the potential becomes positive and hence repulse objects (the Universe seems then to expand faster), and when $\Lambda > 0$ the gravitation becomes very strong again at very far distances from the source! Anyway so far, experimental observation seem to reject that. That's why we are authorized to keep Λ in the final expression.

If the reader is asking himself: how can we derive the Lorentz transformation from General Relativity? This is really asking: how is the Minkowski metric a solution of the vacuum Einstein equation? Because Special Relativity is just the geometry defined by the Minkowski metric.

If we take the Einstein equation and turn off gravity we get the vacuum Einstein equation $G_{ab} = 0$. The Minkowski metric is a solution of this equation, but of course there are lots of others. From this question we may be hoping that the Einstein equation will simplify in the absence of gravity, and this will make it obvious how Special Relativity emerges. Sadly this isn't the case, because even in the absence of mass gravity waves are still allowed!

We don't think there is any way to simplify the Einstein equation to make the Minkowski metric the only solution. We can require that the first derivatives of the metric vanish, but this is really getting the flat space solution by requiring that space not be curved, which is a bit of a tautology. The problem is that in Special Relativity the Minkowski metric is an assumption i.e. it's where you start from. In General Relativity the Minkowski metric is just one among many solutions so there's nothing fundamental about it.

12.4.5.3 Einstein-Maxwell equations

If the energy-momentum tensor $T^{\alpha\beta}$ is that of an electromagnetic field in free space, i.e. if the electromagnetic stress–energy tensor (see page 5764):

$$T^{\alpha\beta} = -\frac{1}{\mu_0}\left(F^{\alpha\psi}F_{\psi}{}^{\beta} - \frac{1}{4}g^{\alpha\beta}F_{\psi\tau}F^{\psi\tau}\right) \tag{12.4.318}$$

is used, then the Einstein field equations are named the "Einstein–Maxwell equations" (with cosmological constant Λ, taken to be zero in conventional relativity theory):

$$R^{\alpha\beta} - \frac{1}{2}Rg^{\alpha\beta} + \Lambda g^{\alpha\beta} = -\frac{8\pi G}{c^4\mu_0}\left(F^{\alpha\psi}F_{\psi}{}^{\beta} - \frac{1}{4}g^{\alpha\beta}F_{\psi\tau}F^{\psi\tau}\right) \tag{12.4.319}$$

or written more commonly:

$$\boxed{R^{\alpha\beta} - \frac{1}{2}Rg^{\alpha\beta} + \Lambda g^{\alpha\beta} = \frac{8\pi G}{c^4\mu_0}\left(F^{\alpha\psi}F_{\psi}{}^{\beta} + \frac{1}{4}g^{\alpha\beta}F_{\psi\tau}F^{\psi\tau}\right)} \tag{12.4.320}$$

12.4.5.4 Schwarzschild Solution

The "Schwarzschild metric" is an approximate solution of the EFE in the case of an isotropic non-rotating gravitational field, without electric charge, zero universal cosmological constant and at a great distance from the source. It provides the three main proofs of General Relativity: the shift of clocks, the deviation of light by a dense celestial body and the advance of the perihelion of Mercury. These three proofs are very important because Einstein's equation was not experimentally demonstrated at the time. The solution is a useful approximation for describing slowly rotating astronomical objects such as many stars and planets, including Earth and the Sun. The solution is named after Karl Schwarzschild, who first published the solution in 11916 (holocene calendar).

To introduce this metric, let us imagine a source (for example the Sun) which produces a gravitational field by means of its mass M. We seek, in order to compare with the experiment, the solutions of Einstein's equation (in other words: the metric) outside the source (of the Sun therefore ...) of mass M.

Remark

There are several mathematical techniques to introduce the Schwarzschild metric. The reader will be able to search, for example, in the literature or on the Internet the one using a gauge transformation ("Einstein gauge" with the "harmonic gauge") for the local perturbation constraint. This method is very elegant but very math oriented and we prefer as the reader already know it the "engineer" method...

In other words, this is like assuming to have in the region of space that interests us (considering that there is only the star in question and nothing else around, not even the energy / mass specific to the gravitational field) the following property:

$$T_{\mu\nu}=0 \tag{12.4.321}$$

So the EFE proved just above without cosmological constant:

$$R_{\mu\nu}-\frac{1}{2}Rg_{\mu\nu}=\frac{8\pi G}{c^4}T_{\mu\nu} \tag{12.4.322}$$

then becomes:

$$R_{\mu\nu}-\frac{1}{2}Rg_{\mu\nu}=0 \tag{12.4.323}$$

But we proved above that this last relation can also be written using the definition of the Ricci scalar that is given for recall by:

$$R=g^{\mu\nu}R_{\mu\nu} \tag{12.4.324}$$

as following:

$$R_{\mu\nu}-\frac{1}{2}R_{\mu\nu}g^{\mu\nu}g_{\mu\nu}=R_{\mu\nu}\left(1-\frac{1}{2}g^{\mu\nu}g_{\mu\nu}\right)=0 \tag{12.4.325}$$

and since the parenthesis is not null since we have proved above that in the Minkowski metric:

$$g^{\rho\sigma}g_{\rho\sigma}=4 \tag{12.4.326}$$

it remains:

$$\boxed{R_{\mu\nu}=0} \tag{12.4.327}$$

and therefore in extenso the scalar of Ricci is also null. This last relation is named the "vacuum field equations".

$R_{\mu\nu}=0$
$\delta\int ds=0$
ALBERT
EINSTEIN

Figure 12.224: **Swiss commemorative coin showing the vacuum field equations with zero cosmological constant (top) and action minimization**

We must therefore find the metric that satisfies this relation (in other words, a metric that far from the source corresponds to a flat space since the Ricci tensor is zero). As there are several possibilities let us focus on a particularly elegant case with as the physicists like it ... full of symmetries.

The idea is therefore to find a metric, if possible independent of time (therefore the gravitational field as well will be independent of time!) and ... with spherical symmetry (a star or planet being itself of this form), taking into account the mass of the star (this is the major objective!) and such that far enough from the source (...) or when the mass is zero we fall back on the classical metric known and see earlier above:

$$\mathrm{d}s^2=\mathrm{d}r^2+r^2\mathrm{d}\theta^2+r^2\sin^2(\theta)\mathrm{d}\phi^2 \tag{12.4.328}$$

But this is not totally accurate! Indeed, we work in space-time. But, we have seen that the equation of the curvilinear metric is given in a flat space-time by:

$$\mathrm{d}s^2=c^2\mathrm{d}t^2-(\mathrm{d}x^2+\mathrm{d}y^2+\mathrm{d}z^2) \tag{12.4.329}$$

by passing in spherical coordinates we then have:

$$\mathrm{d}s^2=c^2\mathrm{d}t^2-(\mathrm{d}r^2+r^2\mathrm{d}\theta^2+r^2\sin^2(\theta)\mathrm{d}\phi) \tag{12.4.330}$$

And it is on this equation of the metric that we must fall back when we are far from the source or that the mass is extremely small ($M=0$). That is to say the Schwarzschild metric must therefore be asymptotically flat, that is to say corresponding to the flat space of Minkowski.

So let's get to the task. First, we start from what we know (it's better if we can...!). Which means:

$$\mathrm{d}s^2=g_{\mu\nu}\mathrm{d}x^\mu\mathrm{d}x^\nu \tag{12.4.331}$$

And in spherical coordinates including time we have the components r, θ, ϕ, t. Rigorously, we denote by:

$$(x^{\alpha}) = (ct, r, \theta, \phi) \tag{12.4.332}$$

the "Schwarzschild coordinates".

On a total of 16 terms implied by the prior-previous relation, we finally retain 10 namely: the 4 terms of the diagonal and the 6 other terms of interaction so as to obtain:

$$\mathrm{d}s^2 = A\mathrm{d}t^2 - \begin{pmatrix} B\mathrm{d}r^2 + C\mathrm{d}\theta^2 + D\mathrm{d}\phi^2 \\ +F\mathrm{d}r\mathrm{d}\theta + G\mathrm{d}r\mathrm{d}\phi + H\mathrm{d}r\mathrm{d}t + I\mathrm{d}\theta\mathrm{d}\phi + J\mathrm{d}\theta\mathrm{d}t + K\mathrm{d}\phi\mathrm{d}t \end{pmatrix} \tag{12.4.333}$$

where A, B, C, ... are coefficients to be determined.

Before tackling this work, we know that according to one of our starting constraints, when the mass is weak or we are far from a non high-speed rotating source, we must therefore fall back on:

$$\mathrm{d}s^2 = c^2\mathrm{d}t^2 - \left(\mathrm{d}r^2 + r^2\mathrm{d}\theta^2 + r^2\sin^2(\theta)\mathrm{d}\phi^2\right) \tag{12.4.334}$$

therefore intuitively we can already write:

$$\mathrm{d}s^2 = A\mathrm{d}t^2 - \begin{pmatrix} B\mathrm{d}r^2 + Cr^2\mathrm{d}\theta^2 + Dr^2\sin^2(\theta)\mathrm{d}\phi^2 \\ +F\mathrm{d}r\mathrm{d}\theta + G\mathrm{d}r\mathrm{d}\phi + H\mathrm{d}r\mathrm{d}t + I\mathrm{d}\theta\mathrm{d}\phi + J\mathrm{d}\theta\mathrm{d}t + K\mathrm{d}\phi\mathrm{d}t \end{pmatrix} \tag{12.4.335}$$

what we must admit it ... is a clear progress ...!

If as we have imposed it to ourselves at the beginning, the equation of the metric is independent of time, we can by symmetry of time (hypothesis ...) make the following change of variable:

$$\mathrm{d}t = -\mathrm{d}t \tag{12.4.336}$$

without this changing anything in our $\mathrm{d}s^2$. But, we realize very quick that this will not be the case. Immediately, for this to be satisfied we see that we must have:

$$H = J = K = 0 \tag{12.4.337}$$

Which brings us (it's already better!) to:

$$\mathrm{d}s^2 = Ac^2\mathrm{d}t^2 - \left(B\mathrm{d}r^2 + Cr^2\mathrm{d}\theta^2 + Dr^2\sin^2(\theta)\mathrm{d}\phi^2 + F\mathrm{d}r\mathrm{d}\theta + G\mathrm{d}r\mathrm{d}\phi + I\mathrm{d}\theta\mathrm{d}\phi\right) \tag{12.4.338}$$

Now if the system is indeed spherical, the equation of the metric must be invariant by the transformation $\mathrm{d}\phi = -\mathrm{d}\phi$ (the opposite would be known for a long time if this were not the case experimentally) and/or also for the transformation $\mathrm{d}\theta = -\mathrm{d}\theta$.

So for this to be correct, we see immediately that in the preceding relation we must impose:

$$F = G = I = 0 \tag{12.4.339}$$

So finally it only remains:

$$\mathrm{d}s^2 = Ac^2\mathrm{d}t^2 - (B\mathrm{d}r^2 + Cr^2\mathrm{d}\theta^2 + Dr^2\sin^2(\theta)\mathrm{d}\phi^2) \tag{12.4.340}$$

where A, B, C, D will obviously be independent of time (the opposite would contradict our initial constraint) but may by symmetry of the sphere be dependent of r such that:

$$\mathrm{d}s^2 = A(r)c^2\mathrm{d}t^2 - (B(r)\mathrm{d}r^2 + C(r)r^2\mathrm{d}\theta^2 + D(r)r^2\sin^2(\theta)\mathrm{d}\phi^2) \tag{12.4.341}$$

Remark

Often in the literature, when deriving the famous Schwarzschild solution the metric can be cast in two equivalent ways:

$$\begin{aligned}\mathrm{d}s^2 &= e^{2\alpha(r,t)}\mathrm{d}t^2 - e^{2\beta(r,t)}\mathrm{d}r^2 - r^2\mathrm{d}\Omega^2\\ \mathrm{d}s^2 &= A(r,t)\mathrm{d}t^2 - B(r,t)\mathrm{d}r^2 - r^2\mathrm{d}\Omega^2\end{aligned} \quad (12.4.342)$$

The reason of the exponentials is in the idea to avoid negative input (the exponential function is never negative for real input) and keep the manifold Lorentzian.

Now, let us imagine be on the sphere (rigorously it is a hypersphere but it helps anyway...) at a fixed distance r from the center of the source of the field at a given instant t fixed. We then only have:

$$\mathrm{d}s^2 = C(r)r^2\mathrm{d}\theta^2 + D(r)r^2\sin^2(\theta)\mathrm{d}\phi^2 \quad (12.4.343)$$

since $\mathrm{d}t$ is zero (fixed time) and $\mathrm{d}r$ also (fixed distance r).

We have also on the way removed the sign $-$ because we anticipated that it will be eliminated in the third equality that will follow and we will put it then back.

Now, let us also imagine that we are close to the north pole of the sphere ($\theta = 0$), we then only have in first approximation:

$$\mathrm{d}s^2 \cong C(r)r^2\mathrm{d}\theta^2 \quad (12.4.344)$$

and at equator ($\theta = \pi/2$):

$$\mathrm{d}s^2 \cong D(r)r^2\mathrm{d}\phi^2 \quad (12.4.345)$$

By symmetry of the field, an infinitesimal angular displacement in each of these two particular zones must, however, be equal. From then on, we can only put (by spherical symmetry):

$$C(r) = D(r) \quad (12.4.346)$$

Hence the equation of the metric is reduced to:

$$\mathrm{d}s^2 \cong A(r)c^2\mathrm{d}t^2 - B(r)\mathrm{d}r^2 - C(r)r^2(\mathrm{d}\theta^2 + \sin^2(\theta)\mathrm{d}\phi^2) \quad (12.4.347)$$

Let us now show that we can choose a system of coordinates for which $C(r) = 1$.

Let us introduce for this a distance defined by:

$$\bar{r} = \sqrt{C(r)}r \quad (12.4.348)$$

hence:

$$C(r)r^2 = \bar{r}^2 \quad (12.4.349)$$

Therefore it comes:

$$\frac{\mathrm{d}(\bar{r}^2)}{\mathrm{d}r} = C'(r)r^2 + 2C(r)r \quad (12.4.350)$$

hence:

$$2\bar{r}\mathrm{d}\bar{r}=C'(r)r^2\mathrm{d}r+2C(r)r\mathrm{d}r=2C(r)r\mathrm{d}r\left[1+\frac{r}{2C(r)}C'(r)\right] \tag{12.4.351}$$

This is further simplified by:

$$\bar{r}\mathrm{d}\bar{r}=C(r)r\mathrm{d}r\left[1+\frac{r}{2C(r)}C'(r)\right] \tag{12.4.352}$$

Let's put it all to the square and divide it by left and right by $C(r)r^2=\bar{r}^2$:

$$(\mathrm{d}\bar{r})^2=C(r)(\mathrm{d}r)^2\left[1+\frac{r}{2C(r)}C'(r)\right]^2 \tag{12.4.353}$$

Therefore after rearranging a bit:

$$\mathrm{d}r^2=\frac{1}{C(r)}\left[1+\frac{r}{2C(r)}C'(r)\right]^2\mathrm{d}\bar{r}^2 \tag{12.4.354}$$

We get:

$$B\mathrm{d}r^2=\frac{B}{C(r)}\left[1+\frac{r}{2C(r)}C'(r)\right]^{-2}\mathrm{d}\bar{r}^2=B(\bar{r})\mathrm{d}\bar{r}^2 \tag{12.4.355}$$

Hence the equation of the metric can be written as:

$$\mathrm{d}s^2=A(\bar{r})c^2\mathrm{d}t^2-B(\bar{r})\mathrm{d}\bar{r}^2-\bar{r}^2(\mathrm{d}\theta^2+\sin^2(\theta)\mathrm{d}\phi^2) \tag{12.4.356}$$

It is therefore as if $C(r)=1$:

$$\mathrm{d}s^2=A(r)c^2\mathrm{d}t^2-B(r)\mathrm{d}\bar{r}^2-\bar{r}(\mathrm{d}\theta^2+\sin^2(\theta)\mathrm{d}\phi^2) \tag{12.4.357}$$

Therefore:

$$\mathrm{Diag}(g_{\mu\nu})=(A(r),-B(r),-r^2,-r^2\sin^2(\theta)) \tag{12.4.358}$$

Or explicitly:

$$g_{\mu\nu}=\begin{bmatrix}A(r)&0&0&0\\0&-B(r)&0&0\\0&0&-r^2&0\\0&0&0&-r^2\sin^2(\theta)\end{bmatrix} \tag{12.4.359}$$

And the corresponding contravariant metric tensor (that we will use further below) is then:

$$g^{\mu\nu}=\begin{bmatrix}1/A(r)&0&0&0\\0&-1/B(r)&0&0\\0&0&-1/r^2&0\\0&0&0&-1/(r^2\sin^2(\theta))\end{bmatrix} \tag{12.4.360}$$

such that for recall (see section Tensor Calculus page 2050):

$$g^{\mu\nu}g_{\mu\nu}=\delta^{\mu}_{\nu} \tag{12.4.361}$$

Now, to determine the remaining coefficients (that is, A and B) we are going to use the relation that must satisfy metric if it is locally of the Minkowski type:

$$R_{\mu\nu}=R^{\alpha}_{\mu\alpha\nu}=R^{\alpha}_{\mu,\alpha\nu}=0 \tag{12.4.362}$$

and therefore the first Bianchi's identity (see section Tensor Calculus page 2126) will be automatically satisfied.

Either in a developed form (see section Tensor Calculus page 2122):

$$\begin{aligned}R^{\alpha}_{\mu,\alpha\nu}&=\partial_{\alpha}\Gamma^{\alpha}_{\mu\nu}-\partial_{\nu}\Gamma^{\alpha}_{\mu\alpha}+\Gamma^{\beta}_{\mu\nu}\Gamma^{\alpha}_{\alpha\beta}-\Gamma^{\beta}_{\mu\alpha}\Gamma^{\alpha}_{\nu\beta}\\&=\Gamma^{\alpha}_{\mu\nu,\alpha}-\Gamma^{\alpha}_{\mu\alpha,\nu}+\Gamma^{\alpha}_{\mu\alpha,\nu}+\Gamma^{\beta}_{\mu\nu}\Gamma^{\beta}_{\alpha\beta}\Gamma^{\alpha}_{\nu\beta}=0\end{aligned} \tag{12.4.363}$$

with obviously (see section Tensor Calculus page 2109):

$$\Gamma^{i}_{kj}=\frac{1}{2}g^{il}\left(\partial_k g_{lj}+\partial_j g_{kl}-\partial_l g_{kj}\right)=\frac{1}{2}g^{il}\left(g_{lj,k}+g_{kl,j}-g_{kj,l}\right) \tag{12.4.364}$$

That is to say that we have quite a lot of work to do... OK! First since the metric is simple the only non-zero derivatives are:

$$g_{tt,r}=A' \qquad g_{rr,r}=-B' \qquad g_{\theta\theta,r}=-2r \tag{12.4.365}$$

$$g_{\varphi\varphi,r}=-2r\sin^2(\theta) \qquad g_{\varphi\varphi,\theta}=-2r^2\sin(\theta)\cos(\theta) \tag{12.4.366}$$

We then simply deduce the 9 non-zero elements of the connection (the details are given following the request of a reader):

$$\begin{aligned}\Gamma^{\theta}_{\theta r}=\Gamma^{\theta}_{r\theta}&=\frac{1}{2}g^{\theta l}\left(g_{lr,\theta}+g_{\theta l,r}-g_{\theta r,l}\right)\\&=\frac{1}{2}\left[\begin{array}{l}g^{\theta t}\left(g_{tr,\theta}+g_{\theta t,r}-g_{\theta r,t}\right)+g^{\theta r}\left(g_{rr,\theta}+g_{\theta r,r}-g_{\theta r,r}\right)+g^{\theta\theta}\left(g_{\theta r,\theta}+g_{\theta\theta,r}-g_{\theta r,\theta}\right)\\+g^{\theta\varphi}\left(g_{\varphi r,\theta}+g_{\theta\varphi,r}-g_{\theta r,\varphi}\right)\end{array}\right]\\&=\frac{1}{2}\left[\begin{array}{l}0\cdot\left(g_{tr,\theta}+g_{\theta t,r}-g_{\theta r,t}\right)+0\cdot\left(g_{rr,\theta}+g_{\theta r,r}-g_{\theta r,r}\right)-\dfrac{1}{r^2}\left(g_{\theta r,\theta}+g_{\theta\theta,r}-g_{\theta r,\theta}\right)\\+0\cdot\left(g_{\varphi r,\theta}+g_{\theta\varphi,r}-g_{\theta r,\varphi}\right)\end{array}\right]\\&=-\frac{1}{r^2}\left(g_{\theta r,\theta}+g_{\theta\theta,r}-g_{\theta r,\theta}\right)=-\frac{1}{r^2}\left(0+(-2r)-0\right)=\frac{1}{r}\end{aligned} \tag{12.4.367}$$

$$\begin{aligned}\Gamma^{r}_{\varphi\varphi}&=\frac{1}{2}g^{rl}\left(g_{l\varphi,\varphi}+g_{\varphi l,\varphi}-g_{\varphi\varphi,l}\right)\\&=\frac{1}{2}\left[\begin{array}{l}g^{rt}\left(g_{t\varphi,\varphi}+g_{\varphi t,\varphi}-g_{\varphi\varphi,t}\right)+g^{rr}\left(g_{r\varphi,\varphi}+g_{\varphi r,\varphi}-g_{\varphi\varphi,r}\right)+g^{r\theta}\left(g_{\theta\varphi,\varphi}+g_{\varphi\theta,\varphi}-g_{\varphi\varphi,\theta}\right)\\+g^{r\varphi}\left(g_{\varphi\varphi,\varphi}+g_{\varphi\varphi,\varphi}-g_{\varphi\varphi,\varphi}\right)\end{array}\right]\\&=\frac{1}{2}\left[\begin{array}{l}0\cdot\left(g_{t\varphi,\varphi}+g_{\varphi t,\varphi}-g_{\varphi\varphi,t}\right)-\dfrac{1}{B}\left(g_{r\varphi,\varphi}+g_{\varphi r,\varphi}-g_{\varphi\varphi,r}\right)-0\cdot\left(g_{\theta\varphi,\varphi}+g_{\varphi\theta,\varphi}-g_{\varphi\varphi,\theta}\right)\\+0\cdot\left(g_{\varphi\varphi,\varphi}+g_{\varphi\varphi,\varphi}-g_{\varphi\varphi,\varphi}\right)\end{array}\right]\\&=-\frac{1}{2B}\left(g_{r\varphi,\varphi}+g_{\varphi r,\varphi}-g_{\varphi\varphi,r}\right)=-\frac{1}{2B}\left(0+0-\left(-2r\sin^2(\theta)\right)\right)=-\frac{r\sin^2(\theta)}{B}\end{aligned} \tag{12.4.368}$$

$$\begin{aligned}
\Gamma^{r}_{tt} &= \frac{1}{2} g^{rl}\left(g_{lt,t}+g_{tl,t}-g_{tt,l}\right) \\
&= \frac{1}{2}\left[\begin{aligned} & g^{rt}\left(g_{tt,t}+g_{tt,t}-g_{tt,t}\right)+g^{rr}\left(g_{rt,t}+g_{tr,t}-g_{tt,r}\right)+g^{r\theta}\left(g_{\theta t,t}+g_{t\theta,t}-g_{tt,\theta}\right) \\ & +g^{r\varphi}\left(g_{\varphi t,t}+g_{t\varphi,t}-g_{tt,\varphi}\right)\end{aligned}\right] \\
&= \frac{1}{2}\left[\begin{aligned} & 0\cdot\left(g_{tt,t}+g_{tt,t}-g_{tt,t}\right)+\frac{1}{A}\left(g_{rt,t}+g_{tr,t}-g_{tt,r}\right)+0\cdot\left(g_{\theta t,t}+g_{t\theta,t}-g_{tt,\theta}\right) \\ & +0\cdot\left(g_{\varphi t,t}+g_{t\varphi,t}-g_{tt,\varphi}\right)\end{aligned}\right] \\
&= \frac{1}{2A}\left(g_{rt,t}+g_{tr,t}-g_{tt,r}\right)=\frac{1}{2A}\left(0+0-A'\right))=-\frac{A'}{2A}
\end{aligned} \tag{12.4.369}$$

$$\begin{aligned}
\Gamma^{t}_{tr} &= \frac{1}{2} g^{tl}\left(g_{lr,t}+g_{tl,r}-g_{tr,l}\right) \\
&= \frac{1}{2}\left[\begin{aligned} & g^{tt}\left(g_{tr,t}+g_{tt,r}-g_{tr,t}\right)+g^{tr}\left(g_{rr,t}+g_{tt,r}-g_{tr,t}\right)+g^{t\theta}\left(g_{\theta r,t}+g_{t\theta,r}-g_{tr,\theta}\right) \\ & +g^{t\varphi}\left(g_{\varphi t,t}+g_{t\varphi,t}-g_{tr,\varphi}\right)\end{aligned}\right] \\
&= \frac{1}{2}\left[\begin{aligned} & A^{-1}\left(g_{tr,t}+g_{tt,r}-g_{tr,t}\right)+0\cdot\left(g_{rr,t}+g_{tt,r}-g_{tr,t}\right)+0\cdot\left(g_{\theta r,t}+g_{t\theta,r}-g_{tr,\theta}\right) \\ & +0\cdot\left(g_{\varphi r,t}+g_{t\varphi,r}-g_{tr,\varphi}\right)\end{aligned}\right] \\
&= \frac{A^{-1}}{2}\left(g_{tr,t}+g_{tt,r}-g_{tr,t}\right)=\frac{A^{-1}}{2}\left(0+A'-0\right))=\frac{A'}{2A}
\end{aligned} \tag{12.4.370}$$

$$\begin{aligned}
\Gamma^{r}_{\theta\theta} &= \frac{1}{2} g^{rl}\left(g_{l\theta,\theta}+g_{\theta l,\theta}-g_{\theta\theta,l}\right) \\
&= \frac{1}{2}\left[\begin{aligned} & g^{rt}\left(g_{t\theta,\theta}+g_{\theta t,\theta}-g_{\theta\theta,t}\right)+g^{rr}\left(g_{r\theta,\theta}+g_{\theta r,\theta}-g_{\theta\theta,r}\right)+g^{r\theta}\left(g_{\theta\theta,\theta}+g_{\theta\theta,\theta}-g_{\theta\theta,\theta}\right) \\ & +g^{\varphi\theta}\left(g_{\varphi\theta,\theta}+g_{\theta\varphi,\theta}-g_{\theta\theta,\varphi}\right)\end{aligned}\right] \\
&= \frac{1}{2}\left[\begin{aligned} & 0\cdot\left(g_{t\theta,\theta}+g_{\theta t,\theta}-g_{\theta\theta,t}\right)-\frac{1}{B}\left(g_{r\theta,\theta}+g_{\theta r,\theta}-g_{\theta\theta,r}\right)+0\cdot\left(g_{\theta\theta,\theta}+g_{\theta\theta,\theta}-g_{\theta\theta,\theta}\right) \\ & +0\cdot\left(g_{\varphi\theta,\theta}+g_{\theta\varphi,\theta}-g_{\theta\theta,\varphi}\right)\end{aligned}\right] \\
&= -\frac{1}{2B}\left[\left(g_{r\theta,\theta}+g_{\theta r,\theta}-g_{\theta\theta,r}\right)\right]=-\frac{1}{2B}\left[\left(0+0-g_{\theta\theta,r}\right)\right] \\
&= -\frac{1}{2B}\left[\left(0+0-g_{\theta\theta,r}\right)\right]=-\frac{1}{2B}\left[\left(0+0-2r\right)\right]=-\frac{-2r}{2B}=\frac{r}{B}
\end{aligned} \tag{12.4.371}$$

$$
\begin{aligned}
\Gamma^{\varphi}_{\varphi\theta} = \Gamma^{\varphi}_{\theta\varphi} &= \frac{1}{2} g^{\varphi l}\left(g_{l\theta,\varphi} + g_{\varphi l,\theta} - g_{\varphi\theta,l}\right)\\
&= \frac{1}{2}\left[\begin{array}{l} g^{\varphi t}\left(g_{t\theta,\varphi} + g_{\varphi t,\theta} - g_{\varphi\theta,t}\right) + g^{\varphi r}\left(g_{r\theta,\varphi} + g_{\varphi r,\theta} - g_{\varphi\theta,r}\right) + g^{\varphi\theta}\left(g_{\theta\theta,\varphi} + g_{\varphi\theta,\theta} - g_{\varphi\theta,\theta}\right)\\ + \frac{1}{2} g^{\varphi\varphi}\left(g_{\varphi\theta,\varphi} + g_{\varphi\varphi,\theta} - g_{\varphi\theta,\varphi}\right) \end{array}\right]\\
&= \frac{1}{2}\left[\begin{array}{l} 0\cdot\left(g_{t\theta,\varphi} + g_{\varphi t,\theta} - g_{\varphi\theta,t}\right) + 0\cdot\left(g_{r\theta,\varphi} + g_{\varphi r,\theta} - g_{\varphi\theta,r}\right) + 0\cdot\left(g_{\theta\theta,\varphi} + g_{\varphi\theta,\theta} - g_{\varphi\theta,\theta}\right)\\ - \frac{1}{r^2\sin^2(\theta)}\left(g_{\varphi\theta,\varphi} + g_{\varphi\varphi,\theta} - g_{\varphi\theta,\varphi}\right) \end{array}\right]\\
&= -\frac{1}{2r^2\sin^2(\theta)}\left(0 + \left(-2r^2\sin(\theta)\cos(\theta)\right) - 0\right) = \frac{2r^2\sin(\theta)\cos(\theta)}{2r^2\sin^2(\theta)} = \frac{\cos(\theta)}{\sin(\theta)} = \cot(\theta)
\end{aligned}
\tag{12.4.372}
$$

$$
\begin{aligned}
\Gamma^{r}_{rr} &= \frac{1}{2} g^{rl}\left(g_{lr,r} + g_{rl,r} - g_{rr,l}\right)\\
&= \frac{1}{2}\left[\begin{array}{l} g^{rt}\left(g_{tr,r} + g_{rt,r} - g_{rr,t}\right) + g^{rr}\left(g_{rr,r} + g_{rr,r} - g_{rr,r}\right) + g^{r\theta}\left(g_{\theta r,r} + g_{r\theta,r} - g_{rr,\theta}\right)\\ + g^{r\varphi}\left(g_{\varphi r,r} + g_{r\varphi,r} - g_{rr,\varphi}\right) \end{array}\right]\\
&= \frac{1}{2}\left[\begin{array}{l} 0\cdot\left(g_{tr,r} + g_{rt,r} - g_{rr,t}\right) - \frac{1}{B}\left(g_{rr,r} + g_{rr,r} - g_{rr,r}\right) + 0\cdot\left(g_{\theta r,r} + g_{r\theta,r} - g_{rr,\theta}\right)\\ 0\cdot\left(g_{\varphi r,r} + g_{r\varphi,r} - g_{rr,\varphi}\right) \end{array}\right]\\
&- \frac{1}{2B}\left(g_{rr,r} + g_{rr,r} - g_{rr,r}\right) = -\frac{1}{2B}\left(-B' + \left(-B'\right) - \left(-B'\right)\right) = \frac{B'}{2B}
\end{aligned}
\tag{12.4.373}
$$

$$
\begin{aligned}
\Gamma^{\varphi}_{\varphi r} &= \frac{1}{2} g^{\varphi l}\left(g_{lr,\varphi} + g_{\varphi l,r} - g_{\varphi r,l}\right)\\
&= \frac{1}{2}\left[\begin{array}{l} g^{\varphi t}\left(g_{tr,\varphi} + g_{\varphi t,r} - g_{\varphi r,t}\right) + g^{\varphi r}\left(g_{rr,\varphi} + g_{\varphi r,r} - g_{\varphi r,r}\right) + g^{\varphi\theta}\left(g_{\theta r,\varphi} + g_{\varphi\theta,r} - g_{\varphi r,\theta}\right)\\ + g^{\varphi\varphi}\left(g_{\varphi r,\varphi} + g_{\varphi\varphi,r} - g_{\varphi r,\varphi}\right) \end{array}\right]\\
&= \frac{1}{2}\left[\begin{array}{l} 0\cdot\left(g_{tr,\varphi} + g_{\varphi t,r} - g_{\varphi r,t}\right) + 0\cdot\left(g_{rr,\varphi} + g_{\varphi r,r} - g_{\varphi r,r}\right) + 0\cdot\left(g_{\theta r,\varphi} + g_{\varphi\theta,r} - g_{\varphi r,\theta}\right)\\ - \frac{1}{r^2\sin^2(\theta)}\left(g_{\varphi r,\varphi} + g_{\varphi\varphi,r} - g_{\varphi r,\varphi}\right) \end{array}\right]\\
&= -\frac{1}{2r^2\sin^2(\theta)}\left(g_{\varphi r,\varphi} + g_{\varphi\varphi,r} - g_{\varphi r,\varphi}\right) = -\frac{1}{2r^2\sin^2(\theta)}\left(0 + -2r\sin^2(\theta) - 0\right) = \frac{1}{r}
\end{aligned}
\tag{12.4.374}
$$

$$\begin{aligned}
\Gamma_{\varphi\varphi}^{\theta} &= \frac{1}{2} g^{\theta l}\left(g_{l\varphi,\varphi}+g_{\varphi l,\varphi}-g_{\varphi\varphi,l}\right) \\
&= \frac{1}{2}\left[\begin{array}{l} g^{\theta t}\left(g_{t\varphi,\varphi}+g_{\varphi t,\varphi}-g_{\varphi\varphi,t}\right)+g^{\theta r}\left(g_{r\varphi,\varphi}+g_{\varphi r,\varphi}-g_{\varphi\varphi,r}\right)+g^{\theta\theta}\left(g_{\theta\varphi,\varphi}+g_{\varphi\theta,\varphi}-g_{\varphi\varphi,\theta}\right) \\ +g^{\theta\varphi}\left(g_{\varphi\varphi,\varphi}+g_{\varphi\varphi,\varphi}-g_{\varphi\varphi,\varphi}\right) \end{array}\right] \\
&= \frac{1}{2}\left[\begin{array}{l} 0\cdot\left(g_{t\varphi,\varphi}+g_{\varphi t,\varphi}-g_{\varphi\varphi,t}\right)+0\cdot\left(g_{r\varphi,\varphi}+g_{\varphi r,\varphi}-g_{\varphi\varphi,r}\right)-\dfrac{1}{r^2}\left(g_{\theta\varphi,\varphi}+g_{\varphi\theta,\varphi}-g_{\varphi\varphi,\theta}\right) \\ +0\cdot\left(g_{\varphi\varphi,\varphi}+g_{\varphi\varphi,\varphi}-g_{\varphi\varphi,\varphi}\right) \end{array}\right] \\
&= -\frac{1}{2r^2}\left(g_{\theta\varphi,\varphi}+g_{\varphi\theta,\varphi}-g_{\varphi\varphi,\theta}\right) = -\frac{1}{2r^2}\left(0+0-\left(-2r^2\sin(\theta)\cos(\theta)\right)\right) = -\sin(\theta)\cos(\theta)
\end{aligned} \tag{12.4.375}$$

To summarize (we have taken the results with the signature $(-,+,+,+)$ of the metric instead of $(+,-,-,-)$ to conform ourselves to the tradition but this does not change the final result):

$$\begin{gathered}
\Gamma_{tr}^{t}=\Gamma_{rt}^{t}=A'/2A \qquad \Gamma_{tt}^{r}=-A'/2B \qquad \Gamma_{rr}^{r}=-B'/2B \\
\Gamma_{\theta r}^{\theta}=\Gamma_{r\theta}^{\theta}=-1/r \qquad \Gamma_{\theta\theta}^{r}=r/B \qquad \Gamma_{\varphi r}^{\varphi}=\Gamma_{r\varphi}^{\varphi}=-1/r \\
\Gamma_{\varphi\varphi}^{r}=r\sin^2(\theta)/B \qquad \Gamma_{\varphi\theta}^{\varphi}=\Gamma_{\theta\varphi}^{\varphi}=-\cot(\theta) \qquad \Gamma_{\varphi\varphi}^{\theta}=\sin(\theta)\cos(\theta)
\end{gathered} \tag{12.4.376}$$

Now that we have these terms of the connection, we have to calculate their derivative in order to be able to express the first two terms of:

$$R_{\mu\nu}=R_{\mu,\alpha\nu}^{\alpha}=\Gamma_{\mu\nu,\alpha}^{\alpha}-\Gamma_{\mu\alpha,\nu}^{\alpha}+\Gamma_{\mu\nu}^{\beta}\Gamma_{\alpha\beta}^{\alpha}-\Gamma_{\mu\alpha}^{\beta}\Gamma_{\nu\beta}^{\alpha}=0 \tag{12.4.377}$$

There are then 10 non-zero terms which are:

$$\begin{aligned}
&\Gamma_{tr,r}^{t}=\Gamma_{rt,r}^{t}=-\frac{A''}{2A}+\frac{A'^2}{2A^2} && \Gamma_{tt,r}^{r}=-\frac{A''}{2B}+\frac{A'B'}{2B^2} \\
&\Gamma_{rr,r}^{r}=-\frac{B''}{2B}+\frac{B'^2}{2B^2} && \Gamma_{\theta r,r}^{\theta}=\Gamma_{r\theta,r}^{\theta}=\frac{1}{r^2} \\
&\Gamma_{\theta\theta,r}^{r}=\frac{1}{B}-\frac{rB'}{B^2} && \Gamma_{\varphi r,r}^{\varphi}=\Gamma_{r\varphi,r}^{\varphi}=\frac{1}{r^2} \\
&\Gamma_{\varphi\varphi,r}^{r}=\frac{\sin^2(\theta)}{B}-\frac{r\sin^2(\theta)B'}{B^2} && \Gamma_{\varphi\theta,\theta}^{\varphi}=\Gamma_{\theta\varphi,\theta}^{\varphi}=\frac{1}{\sin^2(\theta)} \\
&\Gamma_{\varphi\varphi,\theta}^{r}=\frac{r2\sin(\theta)\cos(\theta)}{B} && \Gamma_{\varphi\varphi,\theta}^{\theta}=\cos^2(\theta)-\sin^2(\theta)
\end{aligned} \tag{12.4.378}$$

We finally have for each component of the Ricci tensor:

$$\begin{aligned}
R_{tt} &= \left(\frac{A''}{2B}+\frac{A'B'}{2B^2}\right)-(0)+\left(\frac{A'^2}{2AB}\right)-\left(\frac{A'^2}{4AB}+\frac{A'B'}{4B^2}+\frac{A'}{rB}\right)=-\frac{A''}{2B}+\frac{A'^2}{4AB}+\frac{A'B'}{4B^2}-\frac{A'}{rB}\\
R_{rr} &= \left(-\frac{B''}{2B}+\frac{B'^2}{2B^2}\right)-\left(-\frac{A''}{2A}+\frac{A'^2}{2A^2}-\frac{B''}{2B}\right)+\left(\frac{A'^2}{4A^2}+\frac{B'^2}{4B^2}+\frac{2}{r^2}\right)-\left(\frac{A'B'}{4AB}+\frac{B'^2}{4B^2}+\frac{B'}{rB}\right)\\
&=\frac{A''}{2A}-\frac{A'^2}{4A^2}-\frac{A'B'}{4AB}-\frac{B'}{rB}\\
R_{\theta\theta} &= \left(\frac{1}{B}-\frac{rB'}{B^2}\right)-\left(\frac{1}{\sin^2(\theta)}\right)+\left(-\frac{2}{B}+\cot^2(\theta)\right)-\left(-\frac{A'r}{2AB}-\frac{B'r}{2B^2}-\frac{2}{B}\right)=\frac{A'r}{2AB}-\frac{rB'}{2B^2}+\frac{1}{B}-1\\
R_{\varphi\varphi} &= \left(\sin^2\left[\frac{1}{B}-r\frac{B'}{B^2}-2\right]+1\right)-(0)+\left(-\sin^2(\theta)\left[\frac{2}{B}-2\right]-2\right)\\
&-\left(-\sin^2(\theta)\left[\frac{A'r}{2AB}+\frac{B'r}{2B^2}+\frac{2}{B}-1\right]-1\right)=\sin^2(\theta)\left[\frac{A'r}{2AB}-\frac{B'r}{2B^2}+\frac{1}{B}-1\right]\\
R_{r\theta} &= (0)-(0)+\left(\frac{\cot(\theta)}{r}\right)-\left(\frac{\cot(\theta)}{r}\right)=0
\end{aligned} \tag{12.4.379}$$

The only elements directly non-zero are then:

$$\begin{aligned}
R_{tt} &= -\frac{A''}{2B}+\frac{A'^2}{4AB}+\frac{A'B'}{4B^2}-\frac{A'}{rB}=0\\
R_{rr} &= \frac{A''}{2A}-\frac{A'^2}{4A^2}-\frac{A'B'}{4AB}-\frac{B'}{rB}=0\\
R_{\theta\theta} &= \frac{A'r}{2AB}-\frac{rB'}{2B^2}+\frac{1}{B}-1=0\\
R_{\varphi\varphi} &= \sin^2(\theta)\left[\frac{A'r}{2AB}-\frac{B'r}{2B^2}+\frac{1}{B}-1\right]=\sin^2(\theta)R_{\theta\theta}=0
\end{aligned} \tag{12.4.380}$$

In a more conventional form (according to the literature) we can simplify a little and moreover keep only the first three equations:

$$\begin{aligned}
R_{tt} &= -2rA\partial_r^2(A)B+r\partial_r(A)^2B+rA\partial_r(A)\partial_r(B)-4A\partial_r(A)B=0\\
R_{rr} &= 2rA\partial_r^2(A)B-r\partial_r(A)^2B-rA\partial_r(A)\partial_r(B)-4A^2\partial_r(B)=0\\
R_{\theta\theta} &= r\partial_r(A)B-r\partial_r(B)A+2AB-2AB^2=0
\end{aligned} \tag{12.4.381}$$

If we add the first two equations, we have:

$$\partial_r(A)B+A\partial_r(B)=0 \tag{12.4.382}$$

which equals:

$$\partial_r(A\cdot B)=\partial_r(A)B+A\partial_r(B)=0 \tag{12.4.383}$$

And this also gives us:

$$\partial_r(A)B=-A\partial_r(B) \tag{12.4.384}$$

We have therefore:

$$R_{\theta\theta}=r\partial_r(A)B-r\partial_r(B)A+2AB-2AB^2=0 \tag{12.4.385}$$

which becomes:

$$\begin{aligned}&r\left(\partial_r(A)B-A\partial_r(B)\right)+2AB-2AB^2=0\\&\Rightarrow r\left(-\partial_r(B)A-\partial_r(B)A\right)+2AB-2AB^2=0\\&\Rightarrow -r\partial_r(B)+B(1-B)=0\\&\Rightarrow r\partial_r B(r)=B(r)\Big(1-B(r)\Big)\end{aligned}\tag{12.4.386}$$

Where we have divide by $2A$ when passing from the second to the third line.

The reader can verify that a solution of the differential equation is (we can provide the details on request):

$$B(r)=\left(1+\frac{1}{S\cdot r}\right)^{-1}\tag{12.4.387}$$

Where S is a non-zero real constant. Consequently, the metric for a static solution, symmetrically spherical and in the vacuum (...), is written:

$$\mathrm{d}s^2=A(r)c^2\mathrm{d}t^2-\left(1+\frac{1}{S\cdot r}\right)^{-1}\mathrm{d}r^2-r^2\left(\mathrm{d}\theta^2+\sin^2(\theta)\mathrm{d}\phi^2\right)\tag{12.4.388}$$

It remains for us to determine a coefficient. But as:

$$\partial_r\left(A\cdot B\right)=0\tag{12.4.389}$$

It comes:

$$A\cdot B=K=c^{te}\tag{12.4.390}$$

Hence:

$$A=\frac{K}{B}=K\left(1+\frac{1}{S\cdot r}\right)\tag{12.4.391}$$

Finally:

$$\mathrm{d}s^2=K\left(1+\frac{1}{S\cdot r}\right)c^2\mathrm{d}t^2-\left(1+\frac{1}{S\cdot r}\right)^{-1}\mathrm{d}r^2-r^2\left(\mathrm{d}\theta^2+\sin^2(\theta)\mathrm{d}\phi^2\right)\tag{12.4.392}$$

Let us notice that the space-time represented by this metric is asymptotically flat, or, in other words, when $r\rightarrow+\infty$ the metric approaches that of Minkowski and the space-time variety resembles to that of the Minkowski's space.

To calculate the constants K and S, we use the weak field approximation. In other words, we place ourselves far from the center, where the gravitational field is weak. In this case, the component g_{tt} of the metric can be calculated.

Indeed, we had studied the Newtonian limit earlier above (see page 5751) and obtained the following relation:

$$g_{00}=\eta_{00}+h_{00}=(1-\varphi)=\left(1-\frac{2GM}{rc^2}\right)\tag{12.4.393}$$

So in extenso we can put without too much fear:

$$g_{00}=K\left(1+\frac{1}{S\cdot r}\right)=\left(1-\frac{2GM}{rc^2}\right) \tag{12.4.394}$$

Therefore:

$$K=1 \qquad \text{and} \qquad \frac{1}{S}=-\frac{2GM}{c^2} \tag{12.4.395}$$

Finally we have for the "Schwarzschild metric":

$$\boxed{\mathrm{d}s^2=\left(1-\frac{2GM}{c^2r}\right)c^2\mathrm{d}t^2-\frac{1}{\left(1-\dfrac{2GM}{c^2r}\right)}\mathrm{d}r^2-r^2\left(\mathrm{d}\theta^2+\sin^2(\theta)\mathrm{d}\phi^2\right)} \tag{12.4.396}$$

That is to say in natural units:

$$\mathrm{d}s^2=\left(1-\frac{2GM}{r}\right)\mathrm{d}t^2-\frac{1}{\left(1-\dfrac{2GM}{r}\right)}\mathrm{d}r^2-r^2\left(\mathrm{d}\theta^2+\sin^2(\theta)\mathrm{d}\phi^2\right) \tag{12.4.397}$$

What ultimately gives the Schwarzschild metric tensor in natural units:

$$g_{\mu\nu}=\begin{bmatrix} 1-\dfrac{2GM}{rc^2} & 0 & 0 & 0 \\ 0 & -\left(1-\dfrac{2GM}{rc^2}\right)^{-1} & 0 & 0 \\ 0 & 0 & -r^2 & 0 \\ 0 & 0 & 0 & -r^2\sin^2(\theta) \end{bmatrix} \tag{12.4.398}$$

Or in SI units:

$$g_{\mu\nu}=\begin{bmatrix} c^2-\dfrac{2GM}{r} & 0 & 0 & 0 \\ 0 & -\left(1-\dfrac{2GM}{rc^2}\right)^{-1} & 0 & 0 \\ 0 & 0 & -r^2 & 0 \\ 0 & 0 & 0 & -r^2\sin^2(\theta) \end{bmatrix} \tag{12.4.399}$$

Notice that when $M=0$ we fall back on the Minkowski metric!

Caution!!! Some reference books have the Schwarzschild metric with different signs because they take the metric $(-,+,+,+)$ instead of the metric $(+,-,-,-)$.

The Schwarzschild metric is a solution of Einstein's field equations in empty space (assumed as because we will see during study of quantum field theory that there are not such things as "empty space"), meaning that it is valid only outside the gravitating body. That is, for a spherical body of radius R the solution is valid for $r>R$. To describe the gravitational field both inside and outside the gravitating body the Schwarzschild solution must be matched with some suitable interior solution at $r=R$, such as the "interior Schwarzschild solution".

And all (physically) apparent singularity appears when (this leads also the time component g_{11} to freeze - time is infinitely dilated - as it is then equal to zero!):

$$\left(1-\frac{2GM}{c^2r}\right)=0 \tag{12.4.400}$$

This leads indeed the component g_{22} to be equal to infinity! Or in other words, this is equivalent to the coordinate of the radius r to be equal to:

$$\boxed{r=\frac{2GM}{c^2}} \tag{12.4.401}$$

This radius, which we had already determined during our study of Classical Mechanics, is named the "Schwarzschild radius".

Therefore the Schwarzschild solution appears to have singularities at $r=0$ and $r=2GM/c^2$; some of the metric components "blow up" at these radii. Since the Schwarzschild metric is only expected to be valid for radii larger than the radius R of the gravitating body, there is no problem as long as $R>2GM/c^2$. For ordinary stars and planets this is always the case. For example, the radius of the Sun (actually a G2V-type star, aka yellow dwarf) is approximately $700,000$ [km], while its Schwarzschild radius is only 3 [km].

The Schwarzschild radius is defined as the critical radius provided by the Schwarzschild geometry, below which nothing can escape: if a star or other object reaches a radius equal to or less than its Schwarzschild radius then it becomes a "Black Hole", and any object approaching at a distance from it less than the Schwarzschild's ray will not escape from it. The term is used in physics and astronomy to give an order of magnitude of the characteristic size to which General Relativity effects become necessary for the description of objects of a given mass. The only objects that are not Black Holes and whose size is of the same order as their Schwarzschild radius are neutron stars (or pulsars), thus, curiously, also the observable Universe as a whole...

If we come back to the metric:

$$\mathrm{d}s^2=-\left(1-\frac{2GM}{c^2r}\right)c^2\mathrm{d}t^2+\left(1-\frac{2GM}{c^2r}\right)^{-1}\mathrm{d}r^2+r^2\left(d\theta^2+\sin^2\theta\mathrm{d}\phi^2\right) \tag{12.4.402}$$

notice that, in the metric, coefficients multiplying $\mathrm{d}t^2$ and $\mathrm{d}r^2$ switch signs when r becomes smaller than r_s. This means that, in a way, radial and temporal coordinates have switched their roles as space and time coordinates below the Black Hole horizon!

This also means that r will inevitably grow smaller for unfortunate space traveller who crossed the event horizon, just like e.g. t inevitably grows bigger for anybody in flat space-time. The center of black hole stops being something "over there" once you crossed the horizon, it literally becomes your future.

Definitions (#12.4.545):

D1. A "singularity" is a mathematical entity whose calculation reach infinity because of a division by zero. As we have seen it in the section Analysis (see page 2276), singularities (i.e. "discontinuities") in the field of functional analysis are often also named "singular points". Isolated singularities may be classified as "poles", "essential singularities", "logarithmic singularities", or "removable singularities".

D2. An "apparent singularity", also named coordinate singularity", is a singularity that exists only because of the arbitrary choice of a coordinate system (i.e. a singular point where the general solution is holomorphic[48]).

Remarks

R1. The singularity in the metric when the Schwarzschild radius is reached is apparent because it is only an effect of the coordinate system used. It is an instance of what is named a "coordinate singularity". As the name implies, the singularity arises from a bad choice of coordinates or coordinate conditions. When changing to a different coordinate system (for example Lemaître coordinates, Eddington–Finkelstein coordinates, Kruskal–Szekeres coordinates, Novikov coordinates, or Gullstrand–Painlevé coordinates) the metric becomes regular at $r=2GM/c^2$

R2. A remarkable theorem states that the Schwarzschild metric is the only solution to Einstein's equations in vacuum possessing spherical symmetry. As the Schwarzschild metric is also static, this shows that in fact in vacuum any spherical solution is automatically static. One interesting consequence of this theorem is that any pulsating star that remains spherically symmetrical can not generate gravitational waves (since the space-time region outside the star must remain static).

R3. As the previous developments are based on the assumption of mathematical tools (Bianchi's identity) that requires a zero torsion tensor (see section Tensor Calculus page 2123), i.e. "torsion-free", there are more complete models that can not by extension use the Einstein equation of fields. If we drop the torsion-less connection requirement, we get a more general theory, the "Einstein-Cartan" one (Einstein-Cartan theory is an intermediary step to supergravity models, there is also the Poincare gauge theories of Hehl).

Now that we have the Schwarzschild metric we come back to the Schild criterion that we saw in our classical study of the Einstein effect.

If we rewrite the Schwarzschild metric for a static body, we have the metric which is simplified into:

$$\mathrm{d}s^2=\left(1-\frac{2GM}{c^2r}\right)c^2\mathrm{d}t^2 \qquad (12.4.403)$$

By using the gravitational potential (see section Astronomy page 5266):

$$\varphi=\frac{GM}{r}\;[\mathrm{J}\cdot\mathrm{kg}^{-1}] \qquad (12.4.404)$$

The metric is written:

$$\mathrm{d}s^2=\left(1-\frac{2\varphi}{c^2}\right)c^2\mathrm{d}t^2 \qquad (12.4.405)$$

---

[48] For reminder a "holomorphic function" is complex-valued function of one or more complex variables that is complex differentiable in a neighbourhood of each point in a domain in complex coordinate space

hence by introducing the proper time:

$$c^2\mathrm{d}\tau^2=\left(1-\frac{2\varphi}{c^2}\right)c^2\mathrm{d}t^2 \quad (12.4.406)$$

hence:

$$\mathrm{d}\tau=\left(1-\frac{2\varphi}{c^2}\right)^{1/2}\mathrm{d}t \quad (12.4.407)$$

Therefore:

$$\mathrm{d}t=\left(1-\frac{2\varphi}{c^2}\right)^{-1/2}\mathrm{d}\tau \quad (12.4.408)$$

Maclaurin's second-order expansion in series (see section Sequences and Series page 1703) of the negative root gives:

$$\mathrm{d}t\cong\left(1+\frac{\varphi}{c^2}\right)\mathrm{d}\tau \quad (12.4.409)$$

Therefore we have:

$$\mathrm{d}t\cong\left(1+\frac{\varphi}{c^2}\right)\mathrm{d}\tau\geq\mathrm{d}\tau \quad (12.4.410)$$

Thus, this proof that the curvature (gravitation) generates a larger time dilation (in the sense that it flows faster) that the field of gravity is intense (mass M is large) or that we are close to the body under the influence of the field (small radius r).

For the Earth, the term:

$$\varphi=\frac{GM}{r} \quad (12.4.411)$$

is relatively small. But for a Black Hole or a neutron star, this is no longer the case and the dilation becomes important and the effects accessible to the measure.

12.4.6 Empirical evidence of General Relativity theory

We will now review the 5 classical experimental checks of the 120th century (holocene calendar) of the General Relativity theory which are:

1. The precession of the perihelion which, in terms of numerical results, posed a problem for us with the tools of Classical Mechanics (see section Astronomy page 5361).
2. The deflection of electromagnetic waves (light) passing close to a massive stellar body which in the numerical results also posed a problem to us with the tools of Classical Mechanics (see section Astronomy page 5358).
3. The proof of the Schild criterion (already made in the preceding paragraphs) as the only way to explain rigorously the gravitational redshift and the hypothesis of slowing down time in a gravitational field.

4. The delay of electromagnetic signals propagating near dense bodies. Delay referred to as "Shapiro effect" whose numerical applications are used for the operation of the G.P.S and which will be discussed later.
5. The detection of gravitational waves in the early 121st century (holocene calendar).
6. The first imagery of Black Hole shadows in the early 121st century (holocene calendar).

We will add to this list (when we will have the time...) the derivations of the Lense-Thirring effect of space-time dragging around rotating celestial objects confirmed with very high accuracy by multiple probes like Gravity Probe B, LARES and GRACE and the time decay of binary system (like the he Hulse-Taylor binary system composed of a neutron star and a pulsar which orbit around their common center of mass) that also confirms General Relativity with a very high accuracy!

12.4.6.1 Gravitational Redshift

We know very good that in the non inertial Earth's referential frame, the space time line element between the two events can be written as:

$$\mathrm{d}s^2=g_{\mu\nu}\mathrm{d}x^\mu\mathrm{d}x^\nu \qquad (12.4.412)$$

But as the observer is at rest in his own referential, the only non null coordinates is x_0, so that the square of the line element can be simplified to:

$$\mathrm{d}s^2=g_{00}(\mathrm{d}x^0)^2=g_{00}(c\mathrm{d}t)^2=g_{00}c^2\mathrm{d}t^2 \qquad (12.4.413)$$

If our observer is at rest in his own referential, we know also how to express the space time distance with respect to the proper time τ:

$$\mathrm{d}s^2=c^2\mathrm{d}\tau^2 \qquad (12.4.414)$$

And therefore:

$$c^2\mathrm{d}\tau^2=g_{00}c^2\mathrm{d}t^2 \qquad (12.4.415)$$

That is:

$$\mathrm{d}\tau=\sqrt{g_{00}}\mathrm{d}t \qquad (12.4.416)$$

But we have also proved earlier that in the weak static field approximation:

$$g_{00}=\eta_{00}+h_{00}=1-\frac{2\varphi}{c^2}=1-\frac{2GM}{c^2r} \qquad (12.4.417)$$

Therefore:

$$\mathrm{d}\tau=\sqrt{1-\frac{2GM}{c^2r}}\mathrm{d}t<\mathrm{d}t \qquad (12.4.418)$$

We can then write:

$$\boxed{\mathrm{d}t=\frac{\mathrm{d}\tau}{\sqrt{1-\dfrac{2GM}{c^2r}}}} \tag{12.4.419}$$

This equation tells us that clocks run slower in a gravitational field as seen by a distant observer, this effect is known as "gravitational time dilation" or "gravitational redshift"!

There are a few quite solid experimental direct evidence of the gravitational time dilation like those based on optic clocks (see [99] or [99] or [138]) and especially the Pound–Rebka experiment that is considered as a definitive proof.

As a direct consequence, because frequency is the reciprocal of the period (time interval), we have:

$$\boxed{f_\infty=f_g\sqrt{1-\frac{2GM}{c^2r}}} \tag{12.4.420}$$

where f_∞ is the frequency of the wave as measured by a distant observer of a static and weak gravitational source and f_g is the frequency of the wave measured at the point where the is a gravitational source.

This equation tells us that the frequency of a wave as recorded by a distant observer is less than the frequency recorded by an observer located where the events occurred in the gravitational field. This phenomenon is known as the "gravitational redshift", because a reduction in frequency means a shift toward the longer wavelengths or red end of the electromagnetic spectrum.

We can think of the photons losing energy as they climb out of the gravitational field - loss of energy equating to drop in frequency.

If the atom is at the event horizon of a Black Hole $r=2GM/c^2$, the detected frequency of the photon becomes zero. If the atom was just above the event horizon, the frequency would be finite, but extremely large. But large wavelength means small frequency, which means small energy, which will escape detection if it is small enough. Therefore, no light can be emitted or reflected from the event horizon of Black Hole. The effect is the same also for all other forms of radiation and signals, since they all need to “climb up” an infinite gravitational potential well if they start from the horizon. Therefore, since no radiation or signals can ever come out of the event horizon, Black Hole is indeed black....

A more realistic scenario is that the second observer is itself under the effect of the gravitational field.

Let’s assume that the observer A pointing the torch stands at the surface of the Earth at a distance r_a from the center of the Earth and that the second observer B stands by example at the top of a tower at the distance $r_b=r_a+\delta h$ with δh very small in comparison to r_a.

We can then write:

$$f_\infty=f_A\sqrt{1-\frac{2GM}{c^2r_A}}\cong f_A\sqrt{1-\frac{2GM}{c^2r_B}} \tag{12.4.421}$$

Therefore:

$$f_B = f_A \frac{\sqrt{1-\dfrac{2GM}{c^2 r_A}}}{\sqrt{1-\dfrac{2GM}{c^2 r_B}}} \tag{12.4.422}$$

Let us now use the Schwarzschild radius notation:

$$f_B = f_A \frac{\sqrt{1-\dfrac{R_S}{r_A}}}{\sqrt{1-\dfrac{R_S}{r_B}}} \tag{12.4.423}$$

If we suppose as it is the case on Earth that $R_S \ll r_a$ and so that $R_S \ll r_b$, the redshift can be approximated by a Maclaurin first order expansion (see section Sequences and Series page 1703):

$$\sqrt{1+x} \cong 1+\frac{1}{2}x \qquad \text{and} \qquad \frac{1}{\sqrt{1+x}} \cong 1-\frac{1}{2}x \tag{12.4.424}$$

then:

$$f_B = f_A\left(1-\frac{R_s}{2r_A}\right)\left(1+\frac{R_s}{2r_A}\right) \tag{12.4.425}$$

Therefore:

$$f_B \cong f_A\left(1-\frac{R_S}{2r_A}+\frac{R_S}{2r_B}\right) \tag{12.4.426}$$

Finally:

$$\boxed{\frac{f_B-f_A}{f_A}=\frac{\Delta f}{f_A}\cong\frac{R_S}{2r_B}-\frac{R_S}{2r_A}} \tag{12.4.427}$$

Also written sometimes (according to $\Phi = GM/r$):

$$\boxed{z=\frac{\Delta f}{f_A}=\frac{\Delta\nu}{\nu_A}\cong\frac{\Phi_B}{c^2}-\frac{\Phi_A}{c^2}=\frac{\Phi_B-\Phi_A}{c^2}} \tag{12.4.428}$$

The effect is now considered to have been definitively verified by the experiments of Pound, Rebka and Snider between 11959 (holocene calendar) and 11965 (holocene calendar). The Pound–Rebka experiment of 11959 (holocene calendar) measured the gravitational redshift in spectral lines using a terrestrial ^{57}Fe gamma source over a vertical height of 22.5 [m].

A numerical application gives therefore for this experiment:

$$\frac{\Delta f}{f_A} \cong \frac{GM_T}{(R_T+22.5)c^2}-\frac{GM_T}{R_T c^2} \cong -2.45839\cdot 10^{-15}\,[\mathrm{s}^{-1}] \tag{12.4.429}$$

while Pound and Rebka have found:

$$\frac{\Delta f}{f_A} \cong -2.565\cdot 10^{-15}\,[\mathrm{s}^{-1}] \tag{12.4.430}$$

So first we see that $f_B < f_A$ so there is indeed a shift to the red. Secondly we have an average shift of the frequency to the red!

For the fun let us derive the expression of the formula for the gravitational red shift in the Newtonian limit using the properties of a photon.

In a gravitational field $\vec{g}$ a particle of mass m and velocity $\vec{v}$ changes it's energy E according to:

$$\frac{\mathrm{d}E}{\mathrm{d}t}=m\vec{g}\circ\vec{v}=\vec{g}\circ\vec{p} \tag{12.4.431}$$

For a massless photon described by its energy $E=h\nu=\hbar\omega$ and momentum $\vec{p}=\hbar\vec{k}$ this equation becomes after dividing by the Planck constant $\hbar$:

$$\frac{\mathrm{d}\omega}{\mathrm{d}t}=\vec{g}\circ\vec{k} \tag{12.4.432}$$

Inserting the gravitational field of a spherical body of mass M within the distance $\vec{r}$:

$$\vec{g}=-GM\frac{\vec{r}}{r^3} \tag{12.4.433}$$

and the wave vector of a photon leaving the gravitational field in radial direction:

$$\vec{k}=\frac{\omega}{c}\frac{\vec{r}}{r} \tag{12.4.434}$$

the energy equation becomes:

$$\frac{\mathrm{d}\omega}{\mathrm{d}t}=-\frac{GM}{c}\frac{\omega}{r^2} \tag{12.4.435}$$

Using $\mathrm{d}r=c\mathrm{d}t$ an ordinary differential equation which is only dependent on the radial distance r is obtained:

$$\frac{\mathrm{d}\omega}{\mathrm{d}r}=-\frac{GM}{c^2}\frac{\omega}{r^2} \tag{12.4.436}$$

For a photon starting at the surface of a spherical body with a Radius R_T with a frequency $\omega_0=2\pi\nu_0$ the analytical solution is:

$$\frac{\mathrm{d}\omega}{\mathrm{d}r}=-\frac{GM}{c^2}\frac{\omega}{r^2}\quad\Rightarrow\quad\omega(r)=\omega_0\exp\left(-\frac{GM}{c^2}\left(\frac{1}{R_T}-\frac{1}{r}\right)\right) \tag{12.4.437}$$

In a large distance from the body $r\rightarrow+\infty$ an observer measures the frequency :

$$\omega_{\text{obs}}=\omega_0\exp\left(-\frac{GM}{c^2}\left(\frac{1}{R_T}\right)\right)\cong\omega_0\left(1-\frac{GM}{R_Tc^2}+\frac{1}{2}\frac{G^2M^2}{R_T^2c^4}-\ldots\right) \tag{12.4.438}$$

Therefore, the red shift is:

$$z=\frac{\omega_0-\omega_{\text{obs}}}{\omega_{\text{obs}}}=\frac{1-\exp\left(-\frac{GM}{R_Tc^2}\right)}{\exp\left(-\frac{GM}{R_Tc^2}\right)} \tag{12.4.439}$$

In the linear approximation:

$$z=\frac{\frac{GM}{R_Tc^2}-\frac{1}{2}\frac{G^2M^2}{R_T^2e^4}+\ldots}{1-\frac{GM}{R_Tc^2}+\frac{1}{2}\frac{G^2M^2}{R_T^2c^4}-\ldots}\cong\frac{\frac{GM}{R_Tc^2}}{1-\frac{GM}{R_Tc^2}+\frac{1}{2}\frac{G^2M^2}{R_T^2c^4}-\ldots}\cong\frac{GM}{c^2R_T}\qquad(12.4.440)$$

the Newtonian limit for the gravitational red shift of General Relativity is obtained! For our example above, this gives:

$$\frac{\Delta f}{f_A}\cong\frac{GM_T}{R_Tc^2}\cong 6.9535\cdot 10^{-10}\,[\mathrm{s}^{-1}]\qquad(12.4.441)$$

So we see that the difference with the General Relativity case is quite huge...

12.4.6.2 Precession of Mercury's Perihelion

Let us now treat one of the most famous examples of General Relativity: the precession of the Mercury's perihelion (the same test and confirmation has been done using the VLT with a star orbiting the Milky Way central Black Hole in 12020 - holocene calendar) first shown by Urbain LeVerrier (see [342]). We had already dealt with this case in the Astronomy section (see page 5361), but we had mentioned that the theoretical numerical result did not correspond to the experimental observations. We shall see in the equivalent of almost ten A4 pages of detailed developments how General Relativity makes it possible to reconcile theory and experience.

To study this case, we will use the Lagrangian formalism seen in the section of Analytical Mechanics.

First, let us recall that we obtained for the metric of Schwarzschild:

$$\mathrm{d}s^2=\left(1-\frac{2GM}{c^2r}\right)c^2\mathrm{d}t^2-\frac{1}{\left(1-\frac{2GM}{c^2r}\right)}\mathrm{d}r^2-r^2\left(\mathrm{d}\theta^2+\sin^2(\theta)\mathrm{d}\phi^2\right)\qquad(12.4.442)$$

What we will write by dividing by $\mathrm{d}s^2$:

$$1=\left(1-\frac{2GM}{c^2r}\right)c^2\dot{t}^2-\frac{1}{\left(1-\frac{2GM}{c^2r}\right)}\dot{r}^2-r^2\left(\dot{\theta}^2+\sin^2(\theta)\dot{\phi}^2\right)\qquad(12.4.443)$$

And to abbreviate the notations, we put $l=GM/c^2$ such that:

$$1=\left(1-\frac{2l}{r}\right)c^2\dot{t}^2-\frac{1}{\left(1-\frac{2l}{r}\right)}\dot{r}^2-r^2\left(\dot{\theta}^2+\sin^2(\theta)\dot{\phi}^2\right)\qquad(12.4.444)$$

Now let us recall that (see section Analytical Mechanics page 3161) in natural units:

$$\delta S=\int\delta(\mathrm{d}s)=\int\delta(1\cdot\mathrm{d}s)=0\qquad(12.4.445)$$

So (it's very rude but it works ... This is physics!...):

$$\delta S=\int \delta\left[\left(1-\frac{2l}{r}\right)c^2\dot{t}^2-\frac{1}{\left(1-\dfrac{2l}{r}\right)}\dot{r}^2-r^2\left(\dot{\theta}^2+\sin^2(\theta)\dot{\phi}^2\right)\right]\mathrm{d}s=0 \tag{12.4.446}$$

Finally it means that the Lagrangian is:

$$L=\left(1-\frac{2l}{r}\right)c^2\dot{t}^2-\frac{1}{\left(1-\dfrac{2l}{r}\right)}\dot{r}^2-r^2\left(\dot{\theta}^2+\sin^2(\theta)\dot{\phi}^2\right) \tag{12.4.447}$$

The equations of Lagrange give us for the θ coordinate see section Analytical Mechanics page 3162):

$$\frac{\mathrm{d}}{\mathrm{d}s}\frac{\partial L}{\partial\dot{\theta}}=\frac{\partial L}{\partial\theta} \tag{12.4.448}$$

with therefore:

$$\frac{\partial L}{\partial\dot{\theta}}=-2r^2\dot{\theta} \tag{12.4.449}$$

Hence:

$$\frac{\mathrm{d}}{\mathrm{d}s}\frac{\partial L}{\partial\dot{\theta}}=-2r^2\ddot{\theta}-4r\dot{\theta}\dot{r} \tag{12.4.450}$$

and:

$$\frac{\partial L}{\partial\theta}=-2r^2\sin(\theta)\cos(\theta)\dot{\phi}^2 \tag{12.4.451}$$

From where finally for the coordinate θ:

$$\frac{\mathrm{d}}{\mathrm{d}s}(r^2\dot{\theta})=r^2\sin(\theta)\cos(\theta)\dot{\phi}^2 \tag{12.4.452}$$

Let us do the same for ϕ. First, we have:

$$\frac{\partial L}{\partial\dot{\theta}}=-2r^2\sin^2(\theta)\dot{\phi} \tag{12.4.453}$$

Therefore:

$$\frac{\partial L}{\partial\phi}=0 \tag{12.4.454}$$

And it comes immediately from the application of the Euler-Lagrange equation:

$$\frac{\mathrm{d}}{\mathrm{d}s}\left(r^2\sin^2(\theta)\dot{\phi}\right)=0 \tag{12.4.455}$$

Let us do the same for t:

$$\frac{\partial L}{\partial t}=2\left(1-\frac{2l}{r}\right)c^2\dot{t} \tag{12.4.456}$$

And it comes here also immediately:

$$\frac{\mathrm{d}}{\mathrm{d}s}\left[\left(1-\frac{2l}{r}\right)\dot{t}\right]=0 \tag{12.4.457}$$

Therefore:

$$\left(1-\frac{2l}{r}\right)\dot{t}=c^{te}=A \tag{12.4.458}$$

Now let us assume that the motion of Mercury is in the equatorial plane such as $\theta=\pi/2$. Hence, the relation obtained above:

$$\frac{\mathrm{d}}{\mathrm{d}s}\left(r^2\sin^2(\theta)\dot{\phi}\right)=0 \tag{12.4.459}$$

simplifies into:

$$\frac{\mathrm{d}}{\mathrm{d}s}(r^2\dot{\theta})=0 \tag{12.4.460}$$

hence:

$$r^2\dot{\phi}=c^{te}=B \tag{12.4.461}$$

We have, therefore, the expression of the Universe line, which, for recall, is:

$$1=\left(1-\frac{2l}{r}\right)c^2\dot{t}^2-\frac{1}{\left(1-\dfrac{2l}{r}\right)}\dot{r}^2-r^2\left(\dot{\theta}^2+\sin^2(\theta)\dot{\phi}^2\right) \tag{12.4.462}$$

Which since $\theta=\pi/2$ (which is therefore a constant) is simplified into:

$$1=\left(1-\frac{2l}{r}\right)c^2\dot{t}^2-\frac{1}{\left(1-\dfrac{2l}{r}\right)}\dot{r}^2-r^2\dot{\phi}^2 \tag{12.4.463}$$

Let us now do the following replacement:

$$r^2\dot{\phi}=B \tag{12.4.464}$$

Which is therefore a constant as we have proved just above and also the following replacement (which is also a constant as we proved just above):

$$\dot{t}=\frac{A}{\left(1-\dfrac{2l}{r}\right)^2} \tag{12.4.465}$$

In the universe line element and we get:

$$1=\left(1-\frac{2l}{r}\right)^{-1}c^2A^2-\left(1-\frac{2l}{r}\right)^{-1}\dot{r}^2-\frac{B^2}{r^2} \tag{12.4.466}$$

Let us consider also r as a function of ϕ then:

$$r' = \frac{\mathrm{d}r}{\mathrm{d}\phi} = \frac{\frac{\mathrm{d}r}{\mathrm{d}s}}{\frac{\mathrm{d}\phi}{\mathrm{d}s}} = \frac{\dot{r}}{\dot{\phi}} \tag{12.4.467}$$

hence:

$$\dot{r} = r'\dot{\phi} = r'\frac{B}{r^2} \tag{12.4.468}$$

Thus, we can rewrite the universe line in the form:

$$\left(1-\frac{2l}{r}\right) = c^2A^2 - r'^2\frac{B^2}{r^4} - \left(1-\frac{2l}{r}\right)\frac{B^2}{r^2} \tag{12.4.469}$$

Let us make a change of variable by putting:

$$r = \frac{1}{u} \Rightarrow \frac{\mathrm{d}r}{\mathrm{d}u} = -\frac{1}{u^2} \Rightarrow \mathrm{d}r = -\frac{\mathrm{d}u}{u^2} \tag{12.4.470}$$

hence:

$$\frac{\mathrm{d}r}{\mathrm{d}\phi} = -\frac{\frac{\mathrm{d}u}{\mathrm{d}\phi}}{u^2} \Rightarrow r' = \frac{u'}{u^2} \tag{12.4.471}$$

Which gives for our universe line:

$$\left(1-2lu\right) = c^2A^2 - u'^2B^2 - \left(1-2lu\right)u^2B^2 \tag{12.4.472}$$

or:

$$\begin{aligned} u'^2 &= \frac{c^2A^2}{B^2} - \frac{\left(1-2lu\right)u^2B^2}{B^2} - \frac{\left(1-2lu\right)}{B^2} \\ &= \frac{c^2A^2}{B^2} - u^2 + 2lu^3 - \frac{1}{B^2} + \frac{2lu}{B^2} \\ &= \frac{c^2A^2-1}{B^2} - u^2 + 2lu^3 + \frac{2lu}{B^2} \end{aligned} \tag{12.4.473}$$

By differentiating:

$$2u'u'' = -2uu' + 6lu^2u' + \frac{2lu'}{B^2} \tag{12.4.474}$$

Or written differently:

$$2u'u'' + 2uu' - 6lu^2u' - \frac{2lu'}{B^2} = 0 \tag{12.4.475}$$

Which simplifies and factorize itself into:

$$u'\left(u'' - \frac{l}{B^2} + u - 3lu^2\right) = 0 \tag{12.4.476}$$

The first possible solution is obviously:

$$u' = 0 \tag{12.4.477}$$

Hence as $r=1/u$:

$$r=c^{te} \tag{12.4.478}$$

The circular motion is thus also a solution of Kepler's problem in General Relativity in a Schwarzschild field (ouf!).

The other solution will be:

$$u''-\frac{l}{B^2}+u-3lu^2=0 \tag{12.4.479}$$

Or written differently:

$$\boxed{u''+u=\frac{l}{B^2}+3lu^2} \tag{12.4.480}$$

it corresponds to the orbit of Kepler's problem.

Let us do the comparison by considering in Newton's mechanics the motion of a particle of mass m in a potential V. The Lagrangian (see section Analytical Mechanics page 3170) is then:

$$L=T-V=\frac{1}{2}mv^2-mf(r) \tag{12.4.481}$$

In polar coordinates we have already seen in different sections (Vector Calculus and Astronomy like at page 5340) that the speed is then written:

$$L=\frac{1}{2}m\left[\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2+r^2\left(\frac{\mathrm{d}\phi}{\mathrm{d}t}\right)^2\right]-mf(r) \tag{12.4.482}$$

Using the Euler-Lagrange equation we have the equation of motion:

$$\frac{\partial L}{\partial r}-\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{r}}=0 \tag{12.4.483}$$

Which gives:

$$\frac{\partial L}{\partial r}=rm\left(\frac{\mathrm{d}\phi}{\mathrm{d}t}\right)^2-mf'(r) \quad \text{and} \quad \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{r}}=m\frac{\mathrm{d}^2r}{\mathrm{d}t^2} \tag{12.4.484}$$

hence:

$$\frac{\mathrm{d}^2r}{\mathrm{d}t^2}=r\left(\frac{\mathrm{d}\phi}{\mathrm{d}t}\right)^2-f'(r) \tag{12.4.485}$$

And as we have seen in the section Astronomy (see page 5340):

$$r^2\frac{\mathrm{d}\phi}{\mathrm{d}t}=K \tag{12.4.486}$$

Is the constant of areas. Let us introduce:

$$u(\phi)=\frac{1}{r(\phi)} \tag{12.4.487}$$

Hence:

$$u'(\phi)=-\frac{1}{r^2}r'(\phi) \tag{12.4.488}$$

and therefore:

$$\frac{\mathrm{d}r}{\mathrm{d}t}=\frac{\mathrm{d}r}{\mathrm{d}\phi}\frac{\mathrm{d}\phi}{\mathrm{d}t}=r'(\phi)\frac{K}{r^2}=-Ku'(\phi) \tag{12.4.489}$$

So:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathrm{d}r}{\mathrm{d}t}&=\frac{\mathrm{d}}{\mathrm{d}\phi}\frac{\mathrm{d}\phi}{\mathrm{d}t}\left[-Ku'(\phi)\right]=\frac{\mathrm{d}}{\mathrm{d}\phi}\left[-Ku'(\phi)\right]\frac{\mathrm{d}\phi}{\mathrm{d}t}=-Ku''(\phi)\frac{K}{r^2}\\&=-K^2u''(\phi)u^2\end{aligned} \tag{12.4.490}$$

The equation:

$$\frac{\mathrm{d}^2r}{\mathrm{d}t^2}=r\left(\frac{\mathrm{d}\phi}{\mathrm{d}t}\right)^2-f'(r) \tag{12.4.491}$$

therefore becomes:

$$-K^2u''(\phi)u^2=r\left(\frac{\mathrm{d}\phi}{\mathrm{d}t}\right)^2-f'(r) \tag{12.4.492}$$

But:

$$\frac{\mathrm{d}\phi}{\mathrm{d}t}=\frac{K}{r^2} \tag{12.4.493}$$

Hence:

$$-K^2u''(\phi)u^2=r\left(\frac{K^2}{r^3}\right)^2-f'(r) \tag{12.4.494}$$

therefore:

$$-K^2u''(\phi)u^2=K^2u^3-f'(r) \tag{12.4.495}$$

Where:

$$\boxed{u''(\phi)+u(\phi)=\frac{1}{K^2}\frac{f'(r)}{u^2}} \tag{12.4.496}$$

It is therefore the "non-relativistic Binet formula" which gives the relation between $u=1/r$ and ϕ for a central force (see section Classical Mechanics page 3248). In the case of a Newtonian potential:

$$f(r)=-G\frac{M}{r}\Rightarrow f'(r)=G\frac{M}{r^2}=GMu^2 \tag{12.4.497}$$

Hence:

$$u''(\phi)+u(\phi)=\frac{GM}{K^2} \tag{12.4.498}$$

with for recall:

$$K = r^2 \frac{\mathrm{d}\phi}{\mathrm{d}t} \tag{12.4.499}$$

Now let us recall the form of that which we had obtained just before with the General Relativity:

$$u'' + u = \frac{l}{B^2} + 3lu^2 \tag{12.4.500}$$

Thus, we see that the analogous term in relativity is:

$$\frac{l}{B^2} = \frac{\dfrac{GM}{c^2}}{B^2} \tag{12.4.501}$$

and that General Relativity adds the term $3lu^2$. Now, as in General Relativity:

$$B = r^2 \dot{\phi} \tag{12.4.502}$$

Then:

$$\frac{l}{B^2} = \frac{\dfrac{GM}{c^2}}{r^4 \left(\dfrac{\mathrm{d}\phi}{\mathrm{d}s}\right)^2} = \frac{GM}{c^2 r^4 \left(\dfrac{\mathrm{d}\phi}{\mathrm{d}s}\right)^2 \left(\dfrac{\mathrm{d}t}{\mathrm{d}s}\right)^2} \tag{12.4.503}$$

However, in the case of the approximation of weak fields:

$$\mathrm{d}s^2 = c^2 \mathrm{d}t^2 - \left(\mathrm{d}x_1^2 + \mathrm{d}x_2^2 + \mathrm{d}x_3^2\right) \tag{12.4.504}$$

Hence:

$$\frac{\mathrm{d}s^2}{\mathrm{d}t^2} = c^2 \left[1 - \beta^2\right] \underset{v \ll c}{\cong} c^2 \tag{12.4.505}$$

So finally:

$$\frac{l}{B^2} = \frac{GM}{c^2 r^4 \left(\dfrac{\mathrm{d}\phi}{\mathrm{d}s}\right)^2 \left(\dfrac{\mathrm{d}t}{\mathrm{d}s}\right)^2} \cong \frac{GM}{r^4 \left(\dfrac{\mathrm{d}\phi}{\mathrm{d}s}\right)^2} = \frac{GM}{K^2} \tag{12.4.506}$$

That said, it is really interesting to note that the equation for General Relativity:

$$u'' + u = \frac{l}{B^2} + 3lu^2 \tag{12.4.507}$$

can be interpreted as Binet's equation for Classical Mechanics:

$$u'' + u = \frac{1}{K^2} \frac{f'(r)}{u^2} \tag{12.4.508}$$

with the potential:

$$f(r) = -\frac{GM}{r} - \frac{\gamma}{r^3} \tag{12.4.509}$$

with $\gamma = lK^2$.

Let us now return to our equation:

$$u'' + u = \frac{l}{B^2} + 3lu^2 \tag{12.4.510}$$

We would like to know if the second term on the right of the equality is negligible or not with respect to the first term on the right of the equality in order to be able to apply the theory of perturbations.

We will first put with the help of the weak field approximation given above:

$$C = \frac{l}{B^2} \cong \frac{GM}{K^2} \tag{12.4.511}$$

Now let us calculate the ratio:

$$\begin{aligned}\frac{3lu^2}{\dfrac{l}{B^2}} &= 3B^2u^2 = 3\left(r^2\dot{\phi}\right)^2\frac{1}{r^2} = 3r^2\dot{\phi}^2 = 3\left(r^2\left(\frac{\mathrm{d}\phi}{\mathrm{d}s}\right)^2\right)\\ &= 3\left(r^2\left(\frac{\mathrm{d}\phi}{\mathrm{d}t}\right)^2\left(\frac{\mathrm{d}t}{\mathrm{d}s}\right)^2\right) \cong 3\left(r\frac{\mathrm{d}\phi}{\mathrm{d}t}\right)^2\frac{1}{c^2}\end{aligned} \tag{12.4.512}$$

Recall that in polar coordinates (see section Vector Calculus page 1863):

$$v^2 = \left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2 + r^2\left(\frac{\mathrm{d}\phi}{\mathrm{d}t}\right)^2 \tag{12.4.513}$$

In approximation, we can roughly put that:

$$v^2 \cong r^2\left(\frac{\mathrm{d}\phi}{\mathrm{d}t}\right)^2 \tag{12.4.514}$$

Therefore for Mercury ...:

$$\frac{3lu^2}{\dfrac{l}{B^2}} \cong 3\frac{\nu^2}{c^2} \cong 7.7\cdot 10^{-8} \tag{12.4.515}$$

So we see immediately that we can apply the variational theories to the term $3lu^2$. Thus, let us put:

$$\varepsilon = 3lC = 3l\frac{l}{B^2} \cong 3\frac{GM}{c^2}\frac{GM}{K^2} = 3\frac{G^2M^2}{c^2K^2} \tag{12.4.516}$$

The equation:

$$u'' + u = \frac{l}{B^2} + 3lu^2 \tag{12.4.517}$$

takes the shape:

$$u'' + u = C + \frac{\varepsilon}{C}u^2 \tag{12.4.518}$$

To solve this differential equation, we will use the perturbation theory approach (see section Differential and Integral Calculus page 1645). We will therefore focus on a solution of the form of a Taylor expansion in second order only in ε:

$$u=u_0+\varepsilon u_1+\mathcal{O}(\varepsilon^2) \tag{12.4.519}$$

where u_0, u_1 are obviously dependent on ϕ and will have to be determined! To do this, we know that we must replace the previous expression in the differential equation such that:

$$\left(u_0+\varepsilon u_1+\mathcal{O}(\varepsilon^2)\right)''+\left(u_0+\varepsilon u_1+\mathcal{O}(\varepsilon^2)\right)=C+\frac{\varepsilon}{C}\left(u_0+\varepsilon u_1+\mathcal{O}(\varepsilon^2)\right)^2 \tag{12.4.520}$$

Which simplifies into:

$$u_0''+\varepsilon u_1''+u_0+\varepsilon u_1=C+\frac{\varepsilon u_0^2}{C}+\mathcal{O}(\varepsilon^2) \tag{12.4.521}$$

where let us recall that:

$$u_0''+u_0=C \tag{12.4.522}$$

is the classical equation obtained earlier above:

$$u_0''+u=\frac{GM}{K^2} \tag{12.4.523}$$

Let us consider the solution of the type:

$$u_0=C+D\cos(\phi) \tag{12.4.524}$$

where D is an arbitrary constant. Now, as we have seen in the section of Astronomy in the case of the precession of perihelion:

$$\frac{1}{r}=\frac{B}{\Omega^2}+k_1\cos(\Omega\theta) \tag{12.4.525}$$

is actually an ellipse. Which means that any solution of the form:

$$u_0=C+D\cos(\phi) \tag{12.4.526}$$

is also an ellipse!

For the equation in ε:

$$\varepsilon u_1''+\varepsilon u_1=\frac{\varepsilon u_0^2}{C} \tag{12.4.527}$$

which is simplified into:

$$u_1''+u_1=\frac{(C+D\cos(\phi))^2}{C}=C+2D\cos(\phi)+\frac{D^2}{C}\cos^2(\phi) \tag{12.4.528}$$

Since (see section Trigonometry page 2519):

$$\cos^2(\phi)=\frac{\cos(2\phi)+1}{2} \tag{12.4.529}$$

It comes:

$$\begin{aligned} u_1''+u_1&=\frac{(C+D\cos(\phi))^2}{C}=C+2D\cos(\phi)+\frac{D^2}{C}\frac{\cos(2\phi)+1}{2}\\ &=\left(C+\frac{D^2}{2C}\right)+2D\cos(\phi)+\frac{D^2}{2C}\cos(2\phi)\end{aligned}\tag{12.4.530}$$

To determine u_1, let us decompose it into three terms:

$$u_1=X_1+X_2+X_3\tag{12.4.531}$$

This gives us immediately (by injecting the three terms respectively into the second derivative and the term alone):

$$\begin{aligned} &X_1''+X_1=C+\frac{D^2}{2C}\Rightarrow X_1=C+\frac{D^2}{2C}\\ &X_2''+X_2=2D\cos(\phi)\Rightarrow X_2=D\phi\sin(\phi)\\ &X_3''+X_3=\frac{D^2}{2C}\cos(2\phi)\Rightarrow X_3=-\frac{D^2}{6C}\cos(2\phi)\end{aligned}\tag{12.4.532}$$

So finally:

$$u_1=C+\frac{D^2}{2C}+D\phi\sin(\phi)-\frac{D^2}{6C}\cos(2\phi)\tag{12.4.533}$$

The solution sought is finally:

$$\begin{aligned} u&=u_0+\varepsilon u_1+\mathcal{O}(\varepsilon^2)=(C+D\cos(\phi))+\varepsilon\left(C+\frac{D^2}{2C}+D\phi\sin(\phi)-\frac{D^2}{6C}\cos(2\phi)\right)\\ &=\left(C+\varepsilon C+\frac{\varepsilon D^2}{2C}\right)+\left(D\cos(\phi)-\varepsilon\frac{D^2}{6C}\cos(2\phi)\right)+\varepsilon D\phi(\sin\phi)\end{aligned}\tag{12.4.534}$$

It is therefore with:

$$u(\phi)=\left(C+\varepsilon C+\frac{\varepsilon D^2}{2C}\right)+\left(D\cos(\phi)-\varepsilon\frac{D^2}{6C}\cos(2\phi)\right)+\varepsilon D\phi\sin(\phi)\tag{12.4.535}$$

that it is necessary to calculate the displacement of the perihelion (we arrive soon... pfiuuuuu...).

$$\cos(\phi-\varepsilon\phi)=\cos(\phi)\cos(\varepsilon\phi)+\sin(\phi)\sin(\varepsilon\phi)\tag{12.4.536}$$

We see relatively quickly by observing the preceding relation that the only term whose amplitude is not constant is $\varepsilon D\phi\sin(\phi)$.

Let us then recall that (see section Trigonometry page 2519):

$$\cos(\phi-\varepsilon\phi)\cong\cos(\phi)+\phi\sin(\varepsilon\phi)\tag{12.4.537}$$

This can also be roughly written as a first approximation using Maclaurin's first-order expansion (see section Sequences and Series page 1703):

$$\begin{aligned} u(\phi)&=\left(C+\varepsilon C+\frac{\varepsilon D^2}{2C}\right)+\left(D\cos(\phi)-\varepsilon\frac{D^2}{6C}\cos(2\phi)\right)+\varepsilon D\phi\sin(\phi)\\ &\cong\left(C+\varepsilon C+\frac{\varepsilon D^2}{2C}\right)+\left(-\varepsilon\frac{D^2}{6C}\cos(2\phi)\right)+D\cos(\phi-\varepsilon\phi)\\ &=C+D\cos(\phi-\varepsilon\phi)+\varepsilon\left(C+\frac{D^2}{2C}-\frac{D^2}{6C}\cos(2\phi)\right)\end{aligned}\tag{12.4.538}$$

We know that the zero order orbit is:

$$u_0 = C + D\cos(\phi) \qquad (12.4.539)$$

The effect of the last term:

$$\varepsilon\left(C + \frac{D^2}{2C} - \frac{D^2}{6C}\cos(2\phi)\right) \qquad (12.4.540)$$

is therefore to introduce a small periodic variation in the radial distance. This term does not affect the displacement of the perihelion. This is the term $\varepsilon\phi$ in:

$$\cos(\phi - \varepsilon\phi) = \cos(\phi(1-\varepsilon)) \qquad (12.4.541)$$

which introduces a non-periodicity which can be non-negligible in the case where ϕ is large.

The perihelion (the point closest to the Sun for recall) therefore appears when r is the minimum therefore $u = 1/r$ maximum. But, u is maximum when the term which interests us is maximum, that is to say:

$$\phi(1-\varepsilon) = 2\pi n \qquad n \in \mathbb{Z} \qquad (12.4.542)$$

We have approximately:

$$\phi = \frac{2\pi n}{1-\varepsilon} \cong 2\pi n(1+\varepsilon) \qquad (12.4.543)$$

For two successive perihelions, we have an interval:

$$\Delta\phi \cong 2\pi(1+\varepsilon) = 2\pi + 2\pi\varepsilon = 2\pi + \delta\phi \qquad (12.4.544)$$

instead of 2π. Thus, the displacement for a revolution is:

$$\delta\phi \cong 2\pi\varepsilon = 2\pi\left(3\frac{G^2M^2}{c^2K^2}\right) = \frac{6\pi G^2M^2}{c^2K^2} = \frac{6\pi G^2M^2}{c^2\left(r^2\dfrac{\mathrm{d}\phi}{\mathrm{d}t}\right)^2} = \frac{6\pi G^2M^2}{c^2r^2\nu^2} \qquad (12.4.545)$$

where K is therefore the constant of the areas and M the mass of the central star and since:

$$v = \omega \cdot r = \frac{\mathrm{d}\phi}{\mathrm{d}t} \cdot r \qquad (12.4.546)$$

Finally we have in the end the "Schwarzschild precession":

$$\boxed{\Delta\phi \cong \frac{6\pi G^2M^2}{c^2v^2r^2}} \qquad (12.4.547)$$

Relation to be compared with that obtained in the section of Astronomy with a Classical Newtonian treatment (see page 5361):

$$\Delta\alpha = \frac{\pi G^2M^2}{c^2v^2r^2} \qquad (12.4.548)$$

We thus fall back at the perfection on the factor 6 which was lacking in the conventional treatments!

For Mercury a numerical application gives:

$$\delta\phi \cong \frac{6\pi G^2M^2}{c^2r^2v^2} \cong 42.89'' \qquad (12.4.549)$$

and the experiment gives $\delta\phi \cong 42.5'' \pm 1.0''$. By Albert Einstein's own admission, in obtaining this result he had palpitations and the impression of grazing a heart attack and satisfied with his Herculean effort which had exhausted him he took a long period of rest.

Remark

It is perhaps useful for the reader to know that Albert Einstein and Michele Besso took almost 2 years (!!!) by trials and errors to find the good result above. The first time they had an error of 4000% in comparison to the experimental observed value, the second time an error of 400% and finally the value above (after that Albert Einstein had identified that he choose the wrong Tensor for his theory).

The figure generated by such a precession is commonly named a "rosette". For example here is an artistic illustration of the rosette generated by the orbit of a star around a punctual Black Hole:

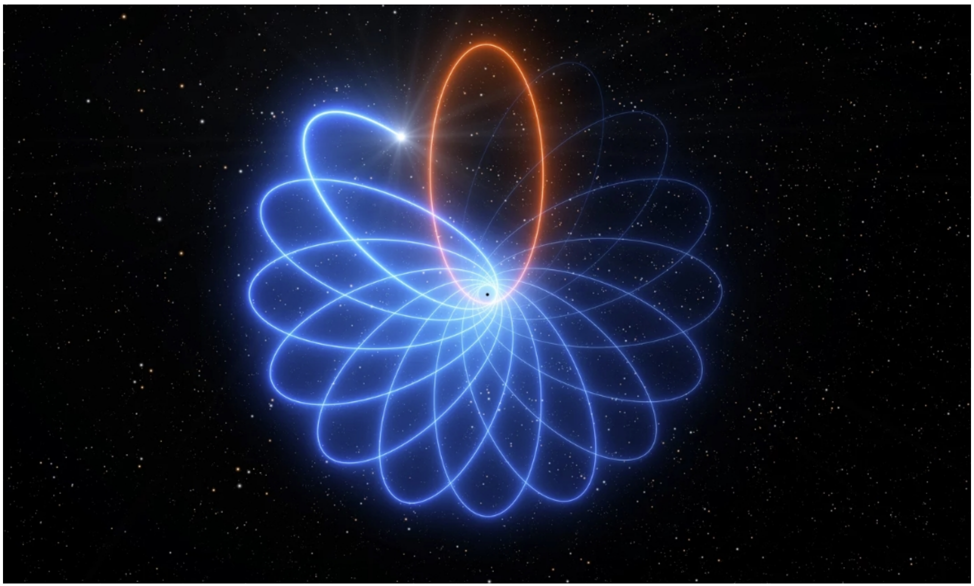


Figure 12.225: **Artistic representation of a Schwarzschild precession with visible rosette (credit: Luis Calcada, Creative Commons Attribution 4.0 International License)**

To conclude on this subject, let us mention a second frequent writing in the literature concerning the result obtained. Indeed, we have proved in the section of Astronomy that the focal parameter was given by:

$$p=\frac{\|\vec{v}\times\vec{r}\|^2}{GM}=\frac{v^2\cdot r^2}{GM} \tag{12.4.550}$$

It therefore remains:

$$\Delta\alpha\cong\frac{6\pi GM}{c^2p} \tag{12.4.551}$$

and we have also proved in the section of Analytical Geometry (see page 2787) that:

$$p=a\left(1-e^2\right) \tag{12.4.552}$$

It thus comes in the end the most classic form:

$$\boxed{\Delta\alpha\cong\frac{6\pi GM}{c^2a\left(1-e^2\right)}} \tag{12.4.553}$$

the numerical values for Mercury precession with $G=6.674\cdot 10^{-11}$ [$\text{m}^3\cdot\text{kg}^{-1}\cdot\text{s}^{-2}$], $M=1.99\cdot 10^{30}$ [kg], $e=0.260$, $a=5.787\cdot 10^{10}$ [m] and $T=88$ [days], gives for a century (100 years of 365 days):

$$\Delta\alpha\frac{100\cdot 365}{88}=\frac{\pi GM}{a(1-e^2)c^2}\frac{100\cdot 365}{88}\cong 5.0267\cdot 10^{-7}\cdot 414.772\cong 2.08\cdot 10^{-4}\,[\text{rad}\cdot\text{century}^{-1}] \quad (12.4.554)$$

Thus a precession of an angle of $43,00415''$. This is a reaaaaallllly good result!

Now there is a curious exercise to be play here with the relation of $\Delta\alpha$, let's compute the relativistic size of the perihelion for various astrophysics systems.

System	**Central Mass**	***a***	***e***	P_{orb}	$\Delta\alpha$
Mercury-Sun	$1\,M_\odot$	$5.70\cdot 10^{10}$ [m]	0.2056	87.969 days	$0.4299''$ /yr
PSR J0737-3039	$1.35+1.24\,M_\odot$	$8.66\cdot 10^{8}$ [m]	0.09	2.4 hours	$17.6°$ /yr
Earth-Moon	$1.35+1.24\,M_\oplus$	$3.84\cdot 10^{8}$ [m]	0.0549	27.32 days	$0.00027''$ /yr
Jupiter-Io	$1.899\cdot 10^{27}$ [kg]	$4.33\cdot 10^{8}$ [m]	0.0041	1.769 days	$2.68''$ /yr
Jupiter-Europa	$1.899\cdot 10^{27}$ [kg]	$6.71\cdot 10^{8}$ [m]	0.094	3.551 days	$0.84''$ /yr
Jupiter-Amalthea	$1.899\cdot 10^{27}$ [kg]	$1.81\cdot 10^{8}$ [m]	0.0032	43.043 [s]	$22.15''$ /yr

12.4.6.3 Deflection of Light (light bending)

We have just proved that:

$$u''+u=\frac{l}{B^2}+3lu^2 \quad (12.4.555)$$

By replacing the factors by their respective values, we have:

$$u''+u=\frac{GM}{c^2B^2}+3\frac{GM}{c^2}u^2 \quad (12.4.556)$$

But we have seen above that:

$$B^2=\frac{lK^2}{GM} \quad (12.4.557)$$

and as K is the areas constant given by the conservation of the momentum itself constant (see section Classical Mechanics page 3244):

$$b=mr^2\omega=mr^2\frac{\mathrm{d}\phi}{\mathrm{d}t}=mK\Rightarrow K=\frac{b}{m} \quad (12.4.558)$$

We then have for a photon $m\rightarrow 0\Rightarrow K\rightarrow +\infty$.

Finally the equation of motion for a photo becomes:

$$u''+u=3\frac{GM}{c^2}u^2 \quad (12.4.559)$$

Let us put now to simplify the notations:

$$r_g=\frac{2GM}{c^2} \tag{12.4.560}$$

Then:

$$u''+u=\frac{3}{2}r_g u^2 \tag{12.4.561}$$

The term to the right of the equality is small (considering the constants that intervene therein) so that an approximate form of the differential equation is:

$$u''+u=0 \tag{12.4.562}$$

of which a particular solution, which we know in advance, is interesting:

$$u=u_0\cos(\varphi) \tag{12.4.563}$$

We carry this approximated solution in the initial differential equation and we get:

$$u''+u=\frac{3}{2}r_g u_0^2\cos^2(\varphi) \tag{12.4.564}$$

Therefore:

$$u''+u=\frac{3}{2}r_g u_0^2\left(\frac{1}{2}+\frac{1}{2}\cos(2\varphi)\right)=\frac{3}{4}r_g u_0^2+\frac{3}{4}r_g u_0^2\cos(2\varphi) \tag{12.4.565}$$

Hence:

$$\frac{\partial^2 u}{\partial\phi^2}+u=\frac{3}{4}r_g u_0^2+\frac{3}{4}r_g u_0^2\cos(2\varphi) \tag{12.4.566}$$

What follows is going to be very subtle (how to guess something like that ...?). First we will create a new differential equation:

$$\frac{\partial^2 v}{\partial\phi^2}+v=\frac{3}{4}r_g u_0^2+\frac{3}{4}r_g u_0^2\sin(2\varphi) \tag{12.4.567}$$

The trick is to multiply this equation by i and sum it to the original differential equation:

$$\frac{\partial^2}{\partial\phi^2}(u+\mathrm{i}v)+(u+\mathrm{i}v)=\frac{3}{4}r_g u_0^2(1+\mathrm{i})+\frac{3}{4}r_g u_0^2 e^{\mathrm{i}2\varphi} \tag{12.4.568}$$

What we will denote by:

$$\frac{\partial^2\bar{u}}{\partial\phi^2}+\bar{u}=\frac{3}{4}r_g u_0^2\sqrt{2}e^{\mathrm{i}\frac{\pi}{4}}+\frac{3}{4}r_g u_0^2 e^{\mathrm{i}2\varphi} \tag{12.4.569}$$

Another trick is to look for a particular solution of the previous relation in the form:

$$\bar{u}=A+Be^{\mathrm{i}2\varphi} \tag{12.4.570}$$

Then we have:

$$\begin{aligned}\bar{u}'&=2\mathrm{i}Be^{\mathrm{i}2\varphi}\\ \bar{u}''&=-4Be^{\mathrm{i}2\varphi}\end{aligned} \tag{12.4.571}$$

This injected into our new differential equation gives:

$$-4Be^{\mathrm{i}2\varphi}+A+Be^{\mathrm{i}2\varphi}=\frac{3}{4}r_gu_0^2\sqrt{2}e^{\mathrm{i}\frac{\pi}{4}}+\frac{3}{4}r_gu_0^2e^{\mathrm{i}2\varphi} \qquad (12.4.572)$$

We deduce immediately:

$$\begin{aligned}A&=\frac{3}{4}r_gu_0^2\sqrt{2}e^{\mathrm{i}\frac{\pi}{4}}\\B&=-\frac{1}{4}r_gu_0^2\end{aligned} \qquad (12.4.573)$$

A particular solution of the original differential equation is thus:

$$\begin{aligned}&u=\Re(\bar{u})=\Re\left(\frac{3}{4}r_gu_0^2\sqrt{2}e^{\mathrm{i}\frac{\pi}{4}}-\frac{1}{4}r_gu_0^2e^{\mathrm{i}2\varphi}\right)=\frac{3}{4}r_gu_0^2-\frac{1}{4}r_gu_0^2\cos(2\varphi)\\&=\frac{1}{2}r_gu_0^2\left(\frac{3}{2}-\frac{1}{2}\cos(2\varphi)\right)\end{aligned} \qquad (12.4.574)$$

Either by using the remarkable trigonometric relations (see section Trigonometry page 2519):

$$\cos(2\varphi)=1-2\sin^2(\varphi) \qquad (12.4.575)$$

It comes:

$$u=\frac{1}{2}r_gu_0^2\left(\frac{3}{2}-\frac{1}{2}+\sin^2(\varphi)\right)=\frac{1}{2}r_gu_0^2\left(1-\sin^2(\varphi)\right) \qquad (12.4.576)$$

The general solution is:

$$u=u_0\cos(\varphi)+\frac{1}{2}r_gu_0^2\left(1-\sin^2(\varphi)\right) \qquad (12.4.577)$$

If we admit that the light is very weakly deviated by the Sun, the radius of curvature ($1/r$) of its trajectory will be very small.

Therefore:

$$u=\frac{1}{r}\rightarrow 0 \qquad (12.4.578)$$

such that:

$$u_0\cos(\varphi)+\frac{1}{2}r_gu_0^2\left(1-\sin^2(\varphi)\right)\cong 0 \qquad (12.4.579)$$

The first term is predominant relatively to the second because of the factor r_g that is very small on the second. For what will follow, we will proceed as in the in the section Astronomy (only the notations change) for the study of the deflection angle at page 5358 (if you don't come back to it, it may be difficult to understand the justification of what will follow!). We put without loosing in generality:

$$\theta=2\varphi-\pi \qquad (12.4.580)$$

Therefore:

$$\varphi=\frac{\pi}{2}+\frac{\theta}{2} \qquad (12.4.581)$$

and as:

$$\begin{aligned}\cos(\varphi)&=\cos\left(\frac{\pi}{2}+\frac{\theta}{2}\right)=-\sin\left(\frac{\theta}{2}\right)\\ \sin^2(\varphi)&=\sin^2\left(\frac{\pi}{2}+\frac{\theta}{2}\right)=\cos^2\left(\frac{\theta}{2}\right)\end{aligned}\tag{12.4.582}$$

it comes:

$$-u_0\sin\left(\frac{\theta}{2}\right)+\frac{1}{2}r_gu_0^2\left(1+\cos^2\left(\frac{\theta}{2}\right)\right)\cong 0\tag{12.4.583}$$

Using trigonometric identities again:

$$\cos^2\left(\frac{\theta}{2}\right)=\frac{1+\cos^2\left(2\dfrac{\theta}{2}\right)}{2}\tag{12.4.584}$$

It comes:

$$-\sin\left(\frac{\theta}{2}\right)+\frac{1}{2}r_gu_0\left(\frac{3}{2}+\frac{1}{2}\cos(\theta)\right)\cong 0\tag{12.4.585}$$

θ being supposed as very small we do a Maclaurin development (see section Sequences and Series page 1703) to the first order of the trigonometric functions:

$$-\frac{\theta}{2}+\frac{1}{2}r_gu_0\left(\frac{3}{2}+\frac{1}{2}\right)\cong 0\tag{12.4.586}$$

Which gives:

$$-\frac{\theta}{2}+r_gu_0\cong 0\Rightarrow\theta\cong 2r_gu_0=\frac{4GM}{c^2}\frac{1}{r_0}\tag{12.4.587}$$

Therefore after a series of approximation... and of hypothesis at the limit of what is acceptable..., we then get for the "deflection of light[49]" (R is sometimes named the "impact parameter", that is to say the distance of nearest approach of the light-beam to the center of mass):

$$\boxed{\theta\cong\frac{4GM}{c^2R}}\tag{12.4.588}$$

instead of the result that we get following the Newtonian approach in the section Astronomy (see page 5358):

$$\varphi\cong\frac{2GM}{v^2r}\tag{12.4.589}$$

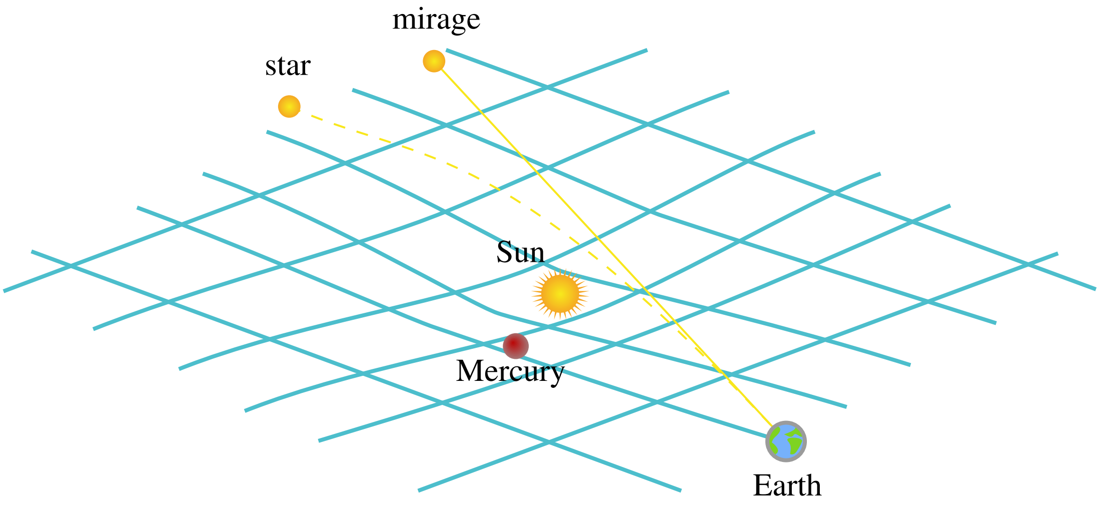
We thus found the factor 2 that was missing in the classical treatment, relatively to experimental observations, that we have proved in the section of Astronomy (see page 5358):

$$\delta_{\text{relativistic}}=2\cdot\delta_{\text{newtonian}}=1.74''\tag{12.4.590}$$

What is often pictured in the media by the following drawing:

---

[49] Also sometimes named "depiction of light".



This deviation have been observed experimentally by measuring the position of stars in the vicinity of the solar disk during the 11919 (holocene calendar) eclipse by Arthur Eddington and his team (see [161]). After the advance of the perihelion of Mercury, this was the second test successfully passed by the General Relativity. It was this event that made Albert Einstein famous among the general public. Today, the deviation of light rays can be measured with much greater precision by considering radio signals emitted by extra-galactic sources (quasars, AGN, etc.): the prediction of the General Relativity has been confirmed to the nearest thousandth.

The deviation of light rays is today very important in observational cosmology. Since it is at the origin of the phenomenon of gravitational mirage, also named "gravitational lens".

It is interesting to notice that the whole theory of gravitational mirages is based on the relation:

$$\theta \cong \frac{4GM}{c^2R} \tag{12.4.591}$$

at least for a point detector. It is the only ingredient of General Relativity used in the calculation of images.

Here is a Flash animation of what a light bulb light rays propagation in vacuum, ON, would looks like (objects are obviously not to scale...):

The animation above will run for people having a PDF reader with Adobe Flash player installed and activated (otherwise see here: https://vimeo.com/575751871).

And here is also for recall a Flash animation of what a light bulb light rays propagation in vacuum putted in the presence of a punctual mass, ON, would looks like:

The animation above will run for people having a PDF reader with Adobe Flash player installed and activated (otherwise see here: https://vimeo.com/575750678).

And here is a Flash animation of what a light bulb light rays propagation in vacuum putted in the presence of a punctual mass but taking into account General Relativity, ON, would looks like (we can see in gray the Schwarzschild radius of the punctual mass):

The animation above will run for people having a PDF reader with Adobe Flash player installed and activated (otherwise see here: https://vimeo.com/575751098).

And here is a Flash of what a light bulb light rays propagation in vacuum putted in the presence of a rotating mass but taking into account General Relativity, ON, would looks like (we can see in dark gray the Schwarzschild radius and in light gray the ergosphere radius of the rotating mass):

The animation above will run for people having a PDF reader with Adobe Flash player installed and activated (otherwise see here: https://vimeo.com/575750255).

In observational astronomy an "Einstein ring", also known as an "Einstein-Chwolson ring" or "Chwolson ring", is the deformation of the light from a source (such as a galaxy or star) into a ring through gravitational lensing of the source's light by an object with an extremely large mass (such as another galaxy or a Black Hole). This occurs when the source, lens, and observer are all aligned.

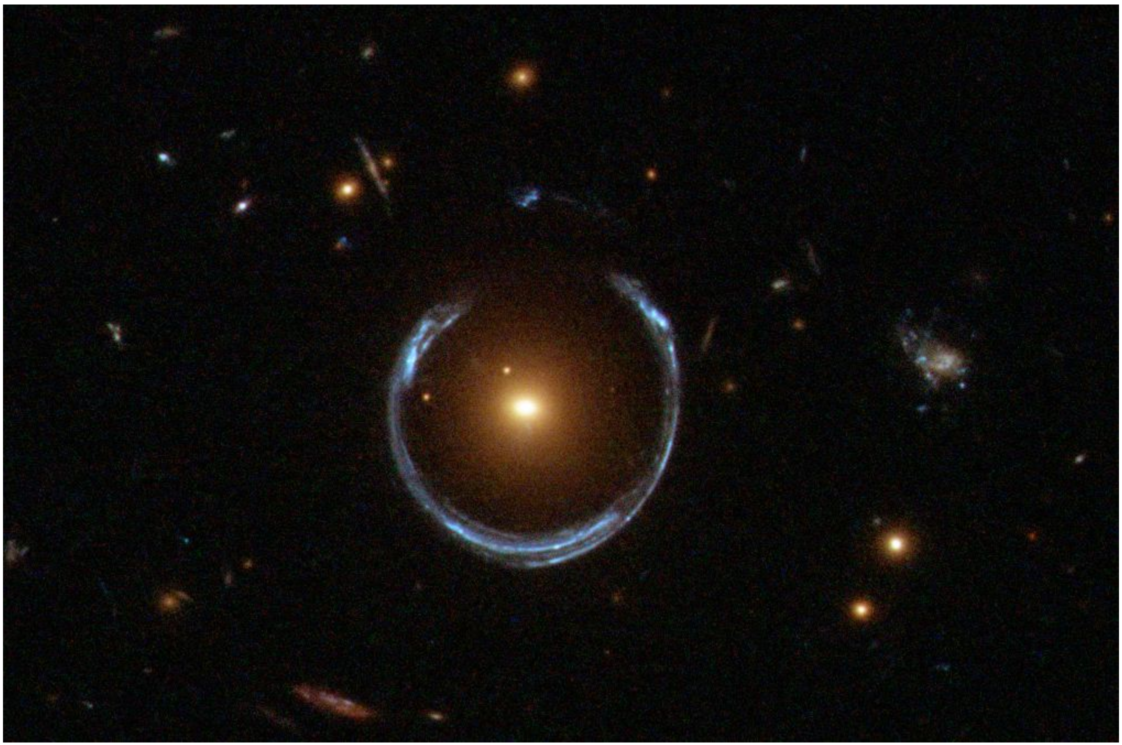


Figure 12.226: Einstein Ring LRG 3 757

When it comes to the distant universe, even the keen vision of NASA's Hubble Space Telescope can only go so far. Teasing out finer details requires clever thinking and a little help from a cosmic alignment with a gravitational lens.

By applying a new computational analysis to a galaxy magnified by a gravitational lens, astronomers have obtained images ten times sharper than what Hubble could achieve on its own. The results show an edge-on disk galaxy studded with brilliant patches of newly formed stars.

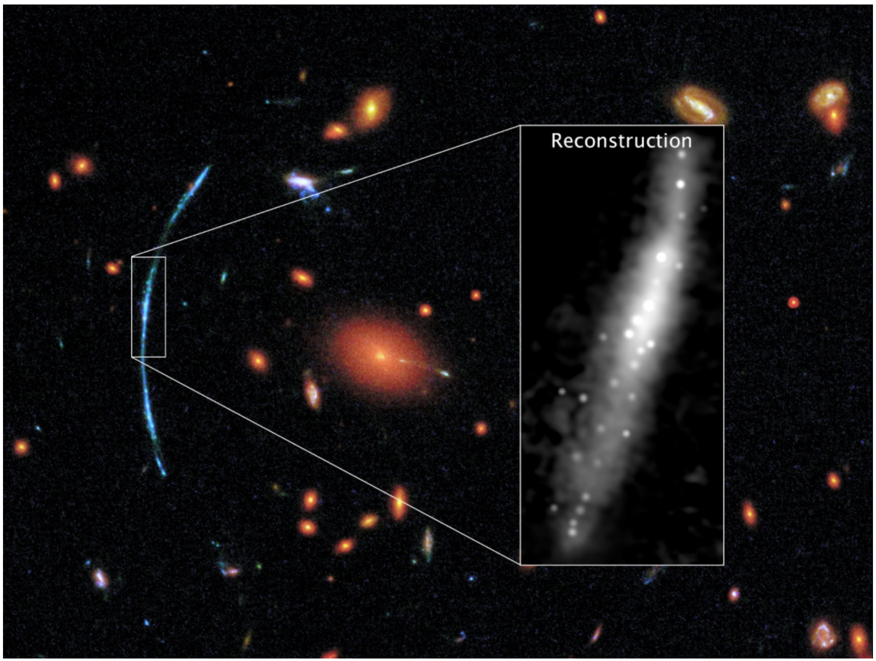


Figure 12.227: **SGAS J111020.0+645950.8 arc magnified by the galaxy cluster SDSS J1110+6459 (source: NASA, ESA, and T. Johnson (University of Michigan))**

In the above Hubble photograph of a distant galaxy cluster, a spotty blue arc stands out against a background of red galaxies. That arc is actually three separate images of the same background galaxy. The background galaxy has been gravitationally lensed, its light magnified and distorted by the intervening galaxy cluster. On the right: How the galaxy would look to Hubble without distortions.

« *When we saw the reconstructed image we said, 'Wow, it looks like fireworks are going off everywhere* », said astronomer Jane Rigby of NASA's Goddard Space Flight Center in Greenbelt, Maryland.

The galaxy in question is so far away that we see it as it appeared 11 billion years ago, only 2.8 billion years after the Big Bang (i.e. starting point of the expansion of our universe). It is one of more than 70 strongly lensed galaxies studied by the Hubble Space Telescope, following up targets selected by the Sloan Giant Arcs Survey, which discovered hundreds of strongly lensed galaxies by searching Sloan Digital Sky Survey imaging data covering one-fourth of the sky.

The gravity of a giant cluster of galaxies between the target galaxy and Earth distorts the more distant galaxy's light, stretching it into an arc and also magnifying it almost 30 times. The team had to develop special computer code to remove the distortions caused by the gravitational lens, and reveal the disk galaxy as it would normally appear.

The resulting reconstructed image revealed two dozen clumps of newborn stars, each spanning about 200 to 300 light-years. This contradicted theories suggesting that star-forming regions in the distant, early universe were much larger, $3,000$ light-years or more in size.

So we have proved above that light is bent near a mass. So this is obviously interpreted as an effect on gravity on light. But the reader must remember something important... the light isn't accelerated because light travels always at the speed of light as it is mass-less (special theory of relativity) instead it is "kind of" accelerated: light is blueshifted if shot directly into a massive body. Since light speed is fixed, it gains energy by increasing its frequency (or shortening its wavelenght).

12.4.6.4 Shapiro Effect (delay)

In 11964 (holocene calendar), Irwin Shapiro demonstrated that a ray of light was not only deflected by passing near a mass, but also that the duration of its path was lengthened in relation to a Euclidean geometry. He calculated that the delay should be about 200 microseconds, therefore perfectly measurable, for a line of sight shaving the Sun. He then suggested systematically measuring the time taken by a radar signal to make the round trip between the Earth and a planet passing behind the Sun (so that the effect is maximal). This was first accomplished with radar echoes on Mars, Venus or Mercury, with an accuracy of the order of 20%. The result was very clear: the time required for a radar signal to make the go and come back between the Earth and the other Planet increases suddenly just before the planet passes behind the Sun and decreases just as suddenly when it reappears.

Remark

We sometimes also talk of "slowing down of the light" near the Sun (actually a G2V-type star, aka yellow dwarf) to describe the Shapiro effect, but it is an awkward and erroneous expression. As we have already been mention it, the speed of light is constant in General Relativity as well as in Relativity for all observers (but for recall this doesn't mean that the speed of light in constant during the life of our Universe!). In the case of the Shapiro effect (and in other similar cases), what changes is the flow of time where the light passes, in relation to what it is where the observer is located.

Although this is a weak effect, it has been verified precisely since the arrival of the Viking probes on Mars in 11976 (holocene calendar), using signals sent from Earth to Mars and reflecting on the latter by the probes (see the principle of the experiment in the figure further below). In addition, there is now even an increasingly common object for which the Shapiro effect must be taken into account: the "G.P.S." (Global Positioning System). Indeed, despite the weakness of the field of gravitation, a geographical precision of a few meters requires such details in the calculation! However, a couple of satellites (LASER-ranged satellites: LAGEOS and LAGEOS II) have been launched to verify in the Earth's gravitational field an even lower effect predicted by General Relativity and which does not even intervene in GPS: the drag over of space also known as the "Lense-Thirring effect" due to the rotation of Earth. These satellites, and the associated scientific teams, did confirmed with 99% accuracy the result of General Relativity in 12004 (holocene calendar)!

Let us point out for the GPS that two source of errors are known within the framework of the Relativity:

1. The satellites rotating around the Earth at a speed of approximately $14,000$ kilometres per hour then delay 7 millionths of a second per day (relativity).
2. At an altitude of $20,200$ kilometres, that of the satellite orbit, the lower gravitational field advances the satellite clocks by 45 millionths of a second per day.

The sum of the two corrections gives a drift of 38 millionths of a second per day, a staggering figure for a GPS system whose precision must be 50 billionths of a second per day!!!

Let us make the calculation for a ray touching the surface of the Sun. For this, we take up our Schwarzschild's metric given for recall by:

$$\mathrm{d}s^2=\left(1-\frac{r_g}{r}\right)c^2\mathrm{d}t^2-\frac{1}{\left(1-\dfrac{r_g}{r}\right)}\mathrm{d}r^2-r^2\left(\mathrm{d}\theta^2+\sin^2(\theta)\mathrm{d}\phi^2\right)\qquad(12.4.592)$$

with:

$$r_g=\frac{2GM}{c^2}\qquad(12.4.593)$$

For a photon, we know that $\mathrm{d}s=0$ and therefore the equation of the Schwarzschild's metric is then written:

$$c^2\mathrm{d}t^2=\frac{1}{\left(1-\dfrac{r_g}{r}\right)^2}\mathrm{d}r^2+\frac{r^2(\mathrm{d}\theta^2+\sin^2(\theta)\mathrm{d}\phi^2)}{\left(1-\dfrac{r_g}{r}\right)}\qquad(12.4.594)$$

The trajectory of the photon taking place in the equatorial plane of the Sun, we put:

$$\theta=c^{te}=\frac{\pi}{2}\qquad(12.4.595)$$

which simplifies even more the equation of the metric into:

$$c^2\mathrm{d}t^2=\frac{1}{\left(1-\dfrac{r_g}{r}\right)^2}\mathrm{d}r^2+\frac{r^2\mathrm{d}\phi^2}{\left(1-\dfrac{r_g}{r}\right)}=\frac{1}{\left(1-\dfrac{r_g}{r}\right)^2}\left(\mathrm{d}r^2+\left(1-\frac{r_g}{r}\right)r^2\mathrm{d}\phi^2\right)\qquad(12.4.596)$$

To simplify even more, we make the hypothesis that the trajectory (in polar coordinates) of the photon shaving the Sun is rectilinear such that (for one of the polar components of the plane):

$$r\cos(\phi)=r_\odot\qquad(12.4.597)$$

where $r_\odot$ is the ray of the Sun. We will use this assumption to simplify the equation of the metric. For this we rearrange:

$$r=\frac{r_\odot}{\cos(\phi)}\qquad(12.4.598)$$

We derive (see section Differential and Integral Calculus page 1475):

$$\frac{\mathrm{d}r}{\mathrm{d}\phi}=\frac{r_\odot\sin(\phi)}{\cos^2(\phi)}=\frac{r_\odot\sin(\phi)}{\dfrac{r_\odot}{r^2}}=\frac{r^2\sin(\phi)}{r_\odot}\qquad(12.4.599)$$

If we square everything:

$$\left(\frac{\mathrm{d}r}{\mathrm{d}\phi}\right)^2=\frac{r^4\sin^2(\phi)}{r_\odot^2}=\frac{r^4(1-\cos^2(\phi))}{r_\odot^2}=\frac{r^4}{r_\odot^2}-\frac{r^4\cos^2(\phi)}{r_\odot^2}=\frac{r^4}{r_\odot^2}-r^2\qquad(12.4.600)$$

hence:

$$\frac{r_{\odot}^2}{(r^2-r_{\odot}^2)}\mathrm{d}r^2=r^2\mathrm{d}\phi^2 \tag{12.4.601}$$

We can now rewrite the equation of the metric:

$$\begin{aligned}c^2\mathrm{d}t^2&=\frac{1}{\left(1-\dfrac{r_g}{r}\right)^2}\left(\mathrm{d}r^2+\left(1-\frac{r_g}{r}\right)r^2\mathrm{d}\phi^2\right)=\frac{\mathrm{d}r^2}{\left(1-\dfrac{r_g}{r}\right)^2}\left(1+\left(1-\frac{r_g}{r}\right)\frac{r_{\odot}^2}{(r^2-r_{\odot}^2)}\right)\\&=\frac{\mathrm{d}r^2}{\left(1-\dfrac{r_g}{r}\right)^2}\left(\frac{(r^2-r_{\odot}^2)+r_{\odot}^2}{(r^2-r_{\odot}^2)}-\frac{r_g}{r}\frac{r_{\odot}^2}{(r^2-r_{\odot}^2)}\right)=\frac{\mathrm{d}r^2}{\left(1-\dfrac{r_g}{r}\right)^2}\left(\frac{r^2}{(r^2-r_{\odot}^2)}-\frac{r_g}{r}\frac{r_{\odot}^2}{(r^2-r_{\odot}^2)}\right)\\&=\frac{\mathrm{d}r^2}{\left(1-\dfrac{r_g}{r}\right)^2}\frac{r^2}{(r^2-r_{\odot}^2)}\left(1-\frac{r_gr_{\odot}^2}{r^3}\right)\end{aligned} \tag{12.4.602}$$

Taking the square root:

$$c\mathrm{d}t=\frac{\mathrm{d}r}{\left(1-\dfrac{r_g}{r}\right)}\frac{r}{(r^2-r_{\odot}^2)^{1/2}}\left(1-\frac{r_gr_{\odot}^2}{r^3}\right)^{1/2} \tag{12.4.603}$$

Since $r>r_{\odot}$ and $r_g\ll 0$ then:

$$\frac{r_g}{r}\ll 1 \qquad \frac{r_gr_{\odot}^2}{r^3}\ll 1 \tag{12.4.604}$$

Therefore, we have using the Maclaurin developments (see section Sequences and Series page 1703) to the first order:

$$c\mathrm{d}t\cong\mathrm{d}r\left(1+\frac{r_g}{r}\right)\frac{r}{(r^2-r_{\odot}^2)^{1/2}}\left(1-\frac{1}{2}\frac{r_gr_{\odot}^2}{r^3}\right) \tag{12.4.605}$$

We have then:

$$\begin{aligned}\mathrm{d}t&\cong\left(\frac{1}{c}\mathrm{d}r+\frac{r_g}{cr}\mathrm{d}r\right)\left(\frac{r}{(r^2-r_{\odot}^2)^{1/2}}-\frac{1}{2}\frac{r_gr_{\odot}^2}{r^3}\frac{r}{(r^2-r_{\odot}^2)^{1/2}}\right)\\&=\frac{1}{c}\mathrm{d}r\frac{r}{(r^2-r_{\odot}^2)^{1/2}}+\frac{r_g}{cr}\mathrm{d}r\frac{r}{(r^2-r_{\odot}^2)^{1/2}}-\frac{1}{c}\mathrm{d}r\frac{1}{2}\frac{r_gr_{\odot}^2}{r^3}\frac{r}{(r^2-r_{\odot}^2)^{1/2}}\\&-\frac{r_g}{cr}\mathrm{d}r\frac{1}{2}\frac{r_gr_{\odot}^2}{r^3}\frac{r}{(r^2-r_{\odot}^2)^{1/2}}\end{aligned} \tag{12.4.606}$$

Finally, we get once condensed:

$$\mathrm{d}t\cong\frac{r}{c(r^2-r_{\odot}^2)^{1/2}}\mathrm{d}r+\frac{r_g}{c(r^2-r_{\odot}^2)^{1/2}}\mathrm{d}r-\frac{r_gr_{\odot}^2}{2cr^2(r^2-r_{\odot}^2)^{1/2}}\mathrm{d}r-\frac{r_g^2r_{\odot}^2}{2cr^3(r^2-r_{\odot}^2)^{1/2}}\mathrm{d}r \tag{12.4.607}$$

What it is traditional to write (we take out the $1/c$ of the different terms):

$$\mathrm{d}t\cong\frac{1}{c}e(r)\mathrm{d}r+\frac{1}{c}f(r)\mathrm{d}r-\frac{1}{c}g(r)\mathrm{d}r-\frac{1}{c}h(r)\mathrm{d}r \tag{12.4.608}$$

If there is no mass then space-time is flat and $r_g=0$. Therefore:

$$\mathrm{d}t\cong\frac{r}{c(r^2-r_{\odot}^2)^{1/2}}\mathrm{d}r \tag{12.4.609}$$

We can thus distinguish the classical time from the extra time generated by the curved space. The "delay" will therefore be given by:

$$\int\mathrm{d}t=\Delta t\cong\frac{1}{c}\int e(r)\mathrm{d}r+\frac{1}{c}\int f(r)\mathrm{d}r-\frac{1}{c}\int g(r)\mathrm{d}r-\frac{1}{c}\int h(r)\mathrm{d}r \tag{12.4.610}$$

Then, to integrate the four functions of r, we must place ourselves in a repository placed if possible at the center of the main body (the Sun typically) since the Schwarzschild metric is based on this hypothesis for recall. Thus, to know the delay of a luminous ray starting from the Sun and travelling to the Earth, we logically choose as the radius of departure that of the Sun itself and as the radius of arrival, the distance Sun-Earth (this will correspond once the primitives computed at the integration terminals).

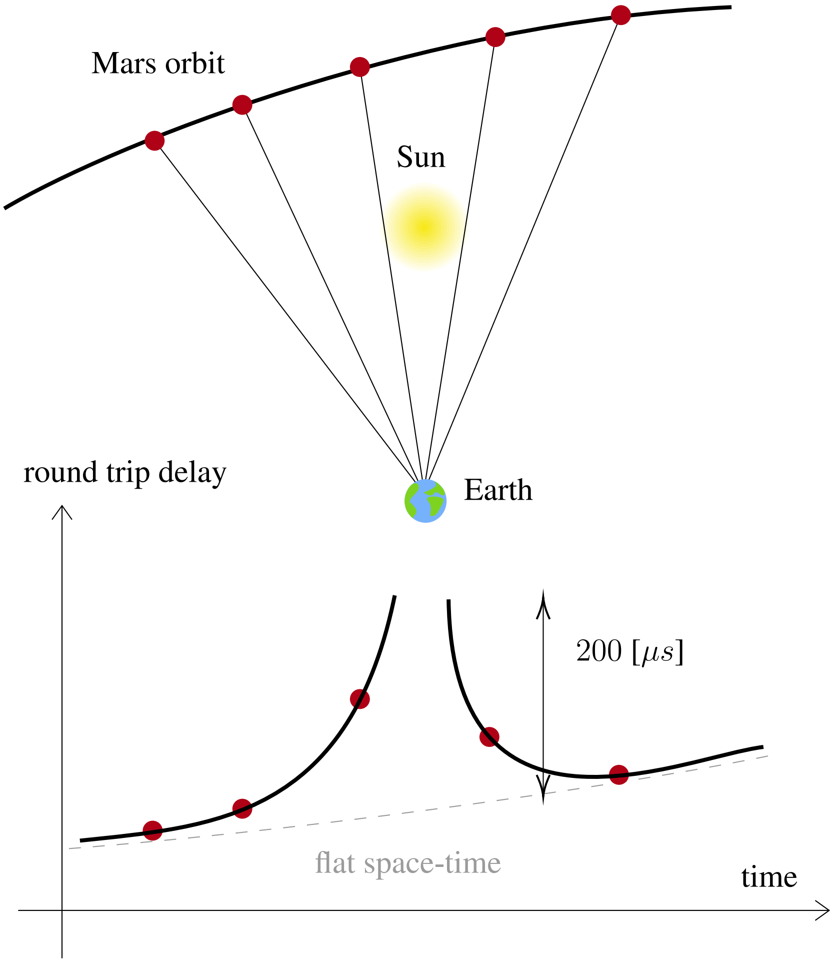


Figure 12.228: **Round-trip time of a signal as a function of the position of Mars**

Well that says it's nice to know the notations of use, but it's even better to do a numerical application! We will therefore first determine the primitive of each of the terms below:

$$\int\frac{r}{\sqrt{r^2-r_{\odot}^2}}\mathrm{d}r,\int\frac{r_g}{\sqrt{r^2-r_{\odot}^2}}\mathrm{d}r,\int\frac{r_gr_{\odot}^2}{2r^2\sqrt{r^2-r_{\odot}^2}}\mathrm{d}r,\int\frac{r_g^2r_{\odot}^2}{2r^3\sqrt{r^2-r_{\odot}^2}}\mathrm{d}r \tag{12.4.611}$$

The first two primitives are simple because they are usual primitives proved in detail in the section of Differential and Integral Calculus (see page 1534):

$$\begin{gathered}\int \frac{r}{\sqrt{r^2-r_\odot^2}}\mathrm{d}r=\sqrt{r^2-r_\odot^2}\\ \int \frac{r_g}{\sqrt{r^2-r_\odot^2}}\mathrm{d}r=r_g\operatorname{arccosh}\left(\frac{r}{r_\odot}\right)=r_g\ln\left(\frac{r}{r_\odot}+\sqrt{\frac{r^2}{r_\odot^2}-1}\right)\end{gathered}\tag{12.4.612}$$

where for the last primitive we have preserved the constant of integration (contrary to what was done in the section of Differential and Integral Calculus because $r_\odot\neq 1$).

Now it remains to us the last two integrals. Let's start in the order by:

$$\int \frac{r_g r_\odot^2}{2r^2\sqrt{r^2-r_\odot^2}}\mathrm{d}r=\frac{r_g r_\odot^2}{2}\int\frac{1}{r^2\sqrt{r^2-r_\odot^2}}\mathrm{d}r\tag{12.4.613}$$

By putting:

$$r=r_\odot\cosh(t)\qquad t=\operatorname{arccosh}\left(\frac{r}{r_\odot}\right)\qquad \mathrm{d}r=r_\odot\sinh(t)\mathrm{d}t\tag{12.4.614}$$

and using the results proved in the section of Differential and Integral Calculus, we then have:

$$\begin{aligned}\frac{r_g r_\odot^2}{2}\int\frac{1}{r^2\sqrt{r^2-r_\odot^2}}\mathrm{d}r&=\frac{r_g r_\odot^2}{2}\int\frac{r_\odot\sinh(t)}{r_\odot^2\cosh^2(t)\sqrt{r_\odot^2\cosh^2(t)-r_\odot^2}}\mathrm{d}t\\ &=\frac{r_g}{2}\int\frac{r_\odot\sinh(t)}{\cosh^2(t)\sqrt{r_\odot^2\sinh^2(t)}}\mathrm{d}t=\frac{r_g}{2}\int\frac{1}{\cosh^2(t)}\mathrm{d}t\\ &=\frac{r_g}{2}\tanh(t)=\frac{r_g}{2}\frac{\sinh(t)}{\cosh(t)}\\ &=\frac{r_g}{2}\frac{\sinh\left(\operatorname{arccosh}\left(\frac{r}{r_\odot}\right)\right)}{\cosh\left(\operatorname{arccosh}\left(\frac{r}{r_\odot}\right)\right)}=\frac{r_g}{2}\frac{\sinh\left(\operatorname{arccosh}\left(\frac{r}{r_\odot}\right)\right)}{\dfrac{r}{r_\odot}}\end{aligned}\tag{12.4.615}$$

Since we have (see section Trigonometry page 2519):

$$\sinh(x)=\sqrt{\cosh^2(x)-1}\tag{12.4.616}$$

Then:

$$\begin{aligned}\frac{r_g r_\odot^2}{2}\int\frac{1}{r^2\sqrt{r^2-r_\odot^2}}\mathrm{d}r&=\frac{r_g}{2}\frac{\sinh\left(\operatorname{arccosh}\left(\frac{r}{r_\odot}\right)\right)}{\dfrac{r}{r_\odot}}\\ &=\frac{r_g}{2}\frac{\sqrt{\cosh^2\left(\operatorname{arccosh}\left(\frac{r}{r_\odot}\right)\right)-1}}{\dfrac{r}{r_\odot}}\\ &=\frac{r_g}{2}\frac{\sqrt{\left(\frac{r}{r_\odot}\right)^2-1}}{\dfrac{r}{r_\odot}}=\frac{r_g}{2}\frac{\dfrac{\sqrt{r^2-r_\odot^2}}{r_\odot}}{\dfrac{r}{r_\odot}}=r_g\frac{\sqrt{r^2-r_\odot^2}}{2r}\end{aligned}\tag{12.4.617}$$

Finally, it remains the last primitive:

$$\int \frac{r_g^2 r_\odot^2}{2r^3\sqrt{r^2-r_\odot^2}}\mathrm{d}r=\frac{r_g^2 r_\odot^2}{2}\int \frac{1}{r^3\sqrt{r^2-r_\odot^2}}\mathrm{d}r=\frac{r_g^2 r_\odot^2}{2}\int \frac{1}{r^4\sqrt{1-\left(\frac{r_\odot}{r}\right)^2}}\mathrm{d}r \tag{12.4.618}$$

We put for what will follow:

$$r=\frac{r_\odot}{x} \qquad \mathrm{d}r=-\frac{r_\odot}{x^2}\mathrm{d}x \tag{12.4.619}$$

Therefore it comes:

$$\begin{aligned}\frac{r_g^2 r_\odot^2}{2}\int \frac{1}{r^4\sqrt{1-\left(\frac{r_\odot}{r}\right)^2}}\mathrm{d}r&=\frac{r_g^2}{2r_\odot}\int \frac{-x^2}{\sqrt{1-x^2}}\mathrm{d}x=\frac{r_g^2}{2r_\odot}\int \frac{1-x^2-1}{\sqrt{1-x^2}}\mathrm{d}x\\&=\frac{r_g^2}{2r_\odot}\left[\int \sqrt{1-x^2}\mathrm{d}x-\int \frac{1}{\sqrt{1-x^2}}\mathrm{d}x\right]\end{aligned} \tag{12.4.620}$$

In the section of Differential and Integral Calculus (see page 1534) we have proved that:

$$\int \sqrt{1-x^2}\mathrm{d}x=\frac{1}{2}\arcsin(x)+\frac{x}{2}\sqrt{1-x^2} \tag{12.4.621}$$

and that:

$$\int \frac{1}{\sqrt{1-x^2}}\mathrm{d}x=\arcsin(x) \tag{12.4.622}$$

Therefore:

$$\begin{aligned}&\frac{r_g^2}{2r_\odot}\left[\int \sqrt{1-x^2}\mathrm{d}x-\int \frac{1}{\sqrt{1-x^2}}\mathrm{d}x\right]=\frac{r_g^2}{2r_\odot}\left[\frac{1}{2}\arcsin(x)+\frac{x}{2}\sqrt{1-x^2}-\arcsin(x)\right]\\&=\frac{r_g^2}{2r_\odot}\left[\frac{x}{2}\sqrt{1-x^2}-\frac{1}{2}\arcsin(x)\right]\end{aligned} \tag{12.4.623}$$

To return to the integral of the beginning we remember that $r=\dfrac{r_\odot}{x}$. Therefore:

$$\int \frac{r_g^2 r_\odot^2}{2r^3\sqrt{r^2-r_\odot^2}}\mathrm{d}r=\frac{r_g^2}{2r_\odot}\left[\frac{r_\odot}{2r^2}\sqrt{r^2-r_\odot^2}-\frac{1}{2}\arcsin\left(\frac{r_\odot}{r}\right)\right] \tag{12.4.624}$$

We thus finally have by taking all the primitives calculated above and by choosing a starting and finishing terminal for the calculation:

$$\Delta t=\frac{1}{c}\left[\sqrt{r^2-r_\odot^2}+r_g\ln\left(\frac{r_\odot}{r}+\sqrt{\frac{r^2}{r_\odot^2}-1}\right)-r_g\frac{\sqrt{r^2-r_\odot^2}}{2r}-\frac{r_g^2}{2r_\odot}\left[\frac{r_\odot}{2r^2}\sqrt{r^2-r_\odot^2}-\frac{1}{2}\arcsin\left(\frac{r_\odot}{r}\right)\right]\right]_{R_0}^{R_1} \tag{12.4.625}$$

We see in the Newtonian limit case where $r_g=0$ that this relation is reduced to:

$$\Delta t\cong\frac{1}{c}\sqrt{r^2-r_\odot^2}\Big|_{R_0}^{R_1}=\frac{1}{c}\left(\sqrt{R_0^2-r_\odot^2}-\sqrt{R_1^2-r_\odot^2}\right) \tag{12.4.626}$$

So for a round trip (between planet and satellite for example), then it comes in this simplified case:

$$\boxed{\Delta t \cong \frac{2}{c}\left(\sqrt{R_0^2 - r_\odot^2} - \sqrt{R_1^2 - r_\odot^2}\right)} \tag{12.4.627}$$

In November 1976 (holocene calendar), when the two Viking spacecraft were operating on the surface of Mars, the planet went behind the Sun as seen from Earth (see figure below). Scientists had preprogrammed Viking to send a radio wave toward Earth that would go extremely close to the outer regions of the Sun (actually a G2V-type star, aka yellow dwarf). According to General Relativity there would be a delay because the radio wave would be passing through a region where time ran more slowly. The experiment was able to confirm Einstein's theory to within 0.1%.

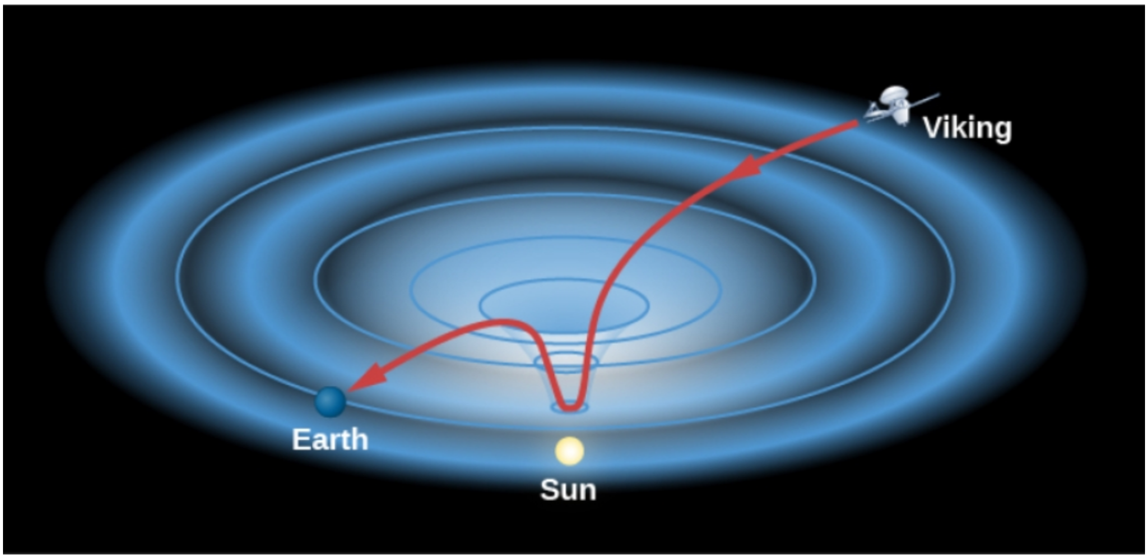


Figure 12.229: **Radio signals from the Viking lander on Mars were delayed when they passed near the Sun (source: OpenStax)**

In 12003 (holocene calendar), with the space probe Cassini an accuracy of 0.0012% was achieved (see [55])!

12.4.6.5 Hafele–Keating experiment (general relativity version)

The Hafele–Keating experiment was a test of the theory of relativity as we already know (see section Special Relativity page 5664). In October 11971 (holocene calendar), Joseph C. Hafele, a physicist, and Richard E. Keating, an astronomer, took four cesium-beam atomic clocks aboard commercial airliners. They flew twice around the world, first eastward, then westward, and compared the clocks against others that remained at the United States Naval Observatory. When reunited, the three sets of clocks were found to disagree with one another, and their differences were consistent with the predictions of special and General Relativity (GR).

It is therefore a macroscopic realization of the famous "twins paradox". We propose in the following to establish the predictions of General Relativity with respect to the relative aging of clocks, based on a simplified aircraft trajectory.

General Relativity predicts an additional effect, in which an increase in gravitational potential due to altitude speeds the clocks up. That is, clocks at higher altitude tick faster than clocks on Earth's surface. This effect has been confirmed in many tests of General Relativity, such as the Pound–Rebka experiment and Gravity Probe A. In the Hafele–Keating experiment, there was a slight increase in gravitational potential due to altitude that tended to speed the clocks back up. Since the aircraft flew at roughly the same altitude in both directions, this effect was approximately the same for the two planes, but nevertheless it caused a difference in comparison to the clocks on the ground.

The results were published in Science in 11972 (holocene calendar):

	Nanoseconds ([ns]) gained, predicted				
	Gravitational (GR)	**Kinematic (SR)**	**Total**	**[ns] gained measured**	**Δ**
Eastward	$+144\pm 14$	-184 ± 18	-40 ± 23	-59 ± 10	0.76σ
Westward	$+179\pm 18$	$+96\pm 10$	$+275\pm 7$	$+273\pm 7$	0.09σ

as available in [220].

The published outcome of the experiment was consistent with Special and General Relativity. The observed time gains and losses were different from zero to a high degree of confidence, and were in agreement with relativistic predictions to within the $\sim 10\%$ precision of the experiment.

It is assumed that in the vicinity of the Earth, the metric tensor is given by the Schwarzschild metric is to say that there exists a system of coordinates $(x^\mu)=(ct,r,\theta,\phi)$, named Schwarzschild coordinates, such that:

$$g_{\mu\nu}x^\mu x^\nu=\mathrm{d}s^2=\left(1-\frac{2GM}{c^2r}\right)c^2\mathrm{d}t^2-\frac{1}{\left(1-\dfrac{2GM}{c^2r}\right)}\mathrm{d}r^2-r^2\left(\mathrm{d}\theta^2+\sin^2(\theta)\mathrm{d}\phi^2\right) \quad (12.4.628)$$

For what will follow, by tradition, we will take the $(-+++)$ metric:

$$g_{\mu\nu}x^\mu x^\nu=\mathrm{d}s^2=-\left(1-\frac{2GM}{c^2r}\right)c^2\mathrm{d}t^2+\frac{1}{\left(1-\dfrac{2GM}{c^2r}\right)}\mathrm{d}r^2+r^2\left(\mathrm{d}\theta^2+\sin^2(\theta)\mathrm{d}\phi^2\right) \quad (12.4.629)$$

where we will take $G\cong 6.67\cdot 10^{-11}\,[\mathrm{m}^3\cdot\mathrm{kg}^{-1}\cdot\mathrm{s}^{-2}]$ and $c\cong 3\cdot 10^8\,[\mathrm{m}\cdot\mathrm{s}^{-1}]$ and $M=5.972\cdot 10^{24}\,[\mathrm{kg}]$.

Ok now we need two results, first the 4-vector speed:

$$u^\alpha=(u^0,u^1,u^2,u^3)=\frac{\mathrm{d}x^\alpha}{\mathrm{d}\tau} \quad (12.4.630)$$

since $u^0=c$.

Now before we continue, consider the following result (with the $(-,+,+,+)$ metric):

$$u_\mu u^\mu=g_{\mu\nu}u^\mu u^\nu=g_{\mu\nu}\frac{\mathrm{d}x^\mu}{\mathrm{d}\tau}\frac{\mathrm{d}x^\nu}{\mathrm{d}\tau}=\frac{g_{\mu\nu}\mathrm{d}x^\mu\mathrm{d}x^\nu}{\mathrm{d}\tau^2}=\frac{\mathrm{d}s^2}{\mathrm{d}\tau^2}=\frac{c^2\mathrm{d}\tau^2}{\mathrm{d}\tau^2}=-c^2 \quad (12.4.631)$$

But:

$$g_{\mu\nu}u^{\mu}u^{\nu}=-\left(1-\frac{2GM}{c^2r}\right)(u^0)^2+\frac{1}{\left(1-\dfrac{2GM}{c^2r}\right)}(u^r)^2+r^2\left((u^{\theta})^2+\sin^2(\theta)(u^{\phi})^2\right) \tag{12.4.632}$$

Equating we get obviously:

$$-\left(1-\frac{2GM}{c^2r}\right)(u^0)^2+\frac{1}{\left(1-\dfrac{2GM}{c^2r}\right)}(u^r)^2+r^2\left((u^{\theta})^2+\sin^2(\theta)(u^{\phi})^2\right)=-c^2 \tag{12.4.633}$$

Explicitly:

$$-\left(1-\frac{2GM}{c^2r}\right)(u^0)^2+\frac{1}{\left(1-\dfrac{2GM}{c^2r}\right)}\left(\frac{\mathrm{d}r}{\mathrm{d}\tau}\right)^2+r^2\left(\left(\frac{\mathrm{d}\theta}{\mathrm{d}\tau}\right)^2+\sin^2(\theta)\left(\frac{\mathrm{d}\phi}{\mathrm{d}\tau}\right)^2\right)=-c^2 \tag{12.4.634}$$

But as:

$$u^0=c\frac{\mathrm{d}t}{\mathrm{d}\tau} \tag{12.4.635}$$

Thus:

$$(u^0)^2\left[1+\frac{2GM}{c^2r}+\frac{1}{\left(1-\dfrac{2GM}{c^2r}\right)}\left(\frac{\mathrm{d}r}{\mathrm{d}\tau}\right)^2\left(\frac{c\mathrm{d}t}{\mathrm{d}\tau}\right)^{-2}+r^2\left(\left(\frac{\mathrm{d}\theta}{\mathrm{d}\tau}\right)^2\left(\frac{c\mathrm{d}t}{\mathrm{d}\tau}\right)^{-2}+\sin^2(\theta)\left(\frac{\mathrm{d}\phi}{\mathrm{d}\tau}\right)^2\left(\frac{c\mathrm{d}t}{\mathrm{d}\tau}\right)^{-2}\right)\right]=-c^2 \tag{12.4.636}$$

After a first simplification:

$$(u^0)^2\left[-1+\frac{2GM}{c^2r}+\frac{1}{\left(1-\dfrac{2GM}{c^2r}\right)}\left(\frac{\mathrm{d}r}{c\mathrm{d}t}\right)^2+r^2\left(\left(\frac{\mathrm{d}\theta}{c\mathrm{d}t}\right)^2+\sin^2(\theta)\left(\frac{\mathrm{d}\phi}{c\mathrm{d}t}\right)^2\right)\right]=-c^2 \tag{12.4.637}$$

After another simplification and little notation simplification:

$$\left(\frac{c\mathrm{d}t}{\mathrm{d}\tau}\right)^2\left[-1+\frac{2GM}{c^2r}+\frac{1}{\left(1-\dfrac{2GM}{c^2r}\right)}\frac{\dot{r}^2}{c^2}+r^2\left(\frac{\dot{\theta}^2}{c^2}+\sin^2(\theta)\frac{\dot{\phi}^2}{c^2}\right)\right]=-c^2 \tag{12.4.638}$$

That is:

$$\mathrm{d}t^2\left[1-\frac{2GM}{c^2r}-\frac{1}{\left(1-\dfrac{2GM}{c^2r}\right)}\frac{\dot{r}^2}{c^2}-r^2\left(\frac{\dot{\theta}^2}{c^2}+\sin^2(\theta)\frac{\dot{\phi}^2}{c^2}\right)\right]=\mathrm{d}\tau^2 \tag{12.4.639}$$

Therefore:

$$\mathrm{d}t=\frac{1}{\sqrt{1-\dfrac{2GM}{c^2r}-\dfrac{1}{\left(1-\dfrac{2GM}{c^2r}\right)}\dfrac{\dot{r}^2}{c^2}-r^2\left(\dfrac{\dot{\theta}^2}{c^2}+\sin^2(\theta)\dfrac{\dot{\phi}^2}{c^2}\right)}}\mathrm{d}\tau \tag{12.4.640}$$

That we will write as:

$$\mathrm{d}t=\frac{1}{\sqrt{1+\left[-\dfrac{2GM}{c^2r}-\dfrac{1}{\left(1-\dfrac{2GM}{c^2r}\right)}\dfrac{\dot{r}^2}{c^2}-r^2\left(\dfrac{\dot{\theta}^2}{c^2}+\sin^2(\theta)\dfrac{\dot{\phi}^2}{c^2}\right)\right]}}\mathrm{d}\tau \tag{12.4.641}$$

Now as $r>R_{\oplus}$ (we take $R=6.4\cdot 10^6$ [m]), we have $GM/c^2r\cong 7\cdot 10^{-10}$. That means globally, we have also if the planes are not at a relativistic speed:

$$\frac{\dot{r}}{c^2}\ll 1,\frac{r^2\dot{\theta}^2}{c^2}\ll 1,\frac{r^2\dot{\phi}}{c^2}\ll 1,\frac{GM}{c^2r}\ll 1 \tag{12.4.642}$$

Then as all the terms in the square brackets are much more smaller than 1 we will use the Maclaurin development (see section Sequences and Series page 1703):

$$\frac{1}{\sqrt{1+x}}\cong 1-\frac{1}{2}x \tag{12.4.643}$$

Then:

$$\mathrm{d}t\cong\left(1+\frac{GM}{c^2r}+\frac{1}{\left(1-\dfrac{GM}{2c^2r}\right)}\frac{\dot{r}^2}{2c^2}+r^2\left(\frac{\dot{\theta}^2}{2c^2}+\sin^2(\theta)\frac{\dot{\phi}^2}{2c^2}\right)\right)\mathrm{d}\tau \tag{12.4.644}$$

We will also put $\left(1-\dfrac{GM}{2c^2r}\right)^{-1}\cong 1$. Therefore:

$$\mathrm{d}t\cong\left(1+\frac{GM}{c^2r}+\frac{\dot{r}^2}{2c^2}+r^2\left(\frac{\dot{\theta}^2}{2c^2}+\sin^2(\theta)\frac{\dot{\phi}^2}{2c^2}\right)\right)\mathrm{d}\tau \tag{12.4.645}$$

That is more convenient to write as:

$$\mathrm{d}t\cong\left(1+\frac{GM}{c^2r}+\frac{1}{2c^2}\left(\dot{r}+r^2\dot{\theta}^2+r^2\sin^2(\theta)\dot{\phi}^2\right)\right)\mathrm{d}\tau \tag{12.4.646}$$

Therefore:

$$\Delta\tau=\int_{t_1}^{t_2}\mathrm{d}\tau\cong\int_{t_1}^{t_2}\frac{1}{\left(1+\dfrac{GM}{c^2r}+\dfrac{1}{2c^2}\left(\dot{r}^2+r^2\dot{\theta}^2+r^2\sin^2(\theta)\dot{\phi}^2\right)\right)}\mathrm{d}t \tag{12.4.647}$$

Then using the Maclaurin development (see section Sequences and Series page 1703):

$$\frac{1}{1+x}\cong 1-x \tag{12.4.648}$$

We have:

$$\Delta\tau\cong\int\limits_{t_1}^{t_2}\left(1-\frac{GM}{c^2r}-\frac{1}{2c^2}\left(\dot{r}^2+r^2\dot{\theta}^2+r^2\sin^2(\theta)\dot{\phi}^2\right)\right)\mathrm{d}t \tag{12.4.649}$$

For the observer who remains on the ground, we have:

$$r=R=c^{te},\dot{r}=0,\theta=c^{te},\dot{\theta}=0,\dot{\phi}=\Omega=\frac{2\pi}{T}=c^{te}=7.3\cdot 10^{-5}\,[\text{rad}\cdot\text{s}^{-1}] \tag{12.4.650}$$

Therefore:

$$\begin{aligned}\Delta\tau_{\text{ground}}&\cong\int\limits_{t_1}^{t_2}\left(1-\frac{GM}{c^2r}-\frac{1}{2c^2}\left(\dot{r}^2+r^2\dot{\theta}^2+r^2\sin^2(\theta)\dot{\phi}^2\right)\right)\mathrm{d}t=\int\limits_{t_1}^{t_2}\left(1-\frac{GM}{c^2R}-\frac{R^2\sin^2(\theta)\Omega^2}{2c^2}\right)\mathrm{d}t\\&=\left(1-\frac{GM}{c^2R}-\frac{R^2\sin^2(\theta)\Omega^2}{2c^2}\right)\int\limits_{t_1}^{t_2}\mathrm{d}t=\left(1-\frac{GM}{c^2R}-\frac{\Omega^2R^2\sin^2(\theta)}{2c^2}\right)\Delta t\end{aligned} \tag{12.4.651}$$

Now for the airplane we have:

$$r=R+h=c^{te},\dot{r}=0,\theta=c^{te},\dot{\theta}=0 \tag{12.4.652}$$

and we assume non-relativistic speed of the airplane and the Newtonian addition of speed such that:

$$r\sin(\theta)\dot{\phi}_{\text{airplance}}\cong v_{\text{airplane}}+r\sin(\theta)\dot{\phi}_{\text{ground}}=v_{\text{airplane}}+r\Omega\sin(\theta)=v_{\text{airplane}}+(R+h)\Omega\sin(\theta)=c^{te} \tag{12.4.653}$$

Therefore:

$$\begin{aligned}\Delta\tau_{\text{airplane}}&\cong\int\limits_{t_1}^{t_2}\left(1-\frac{GM}{c^2r}-\frac{1}{2c^2}\left(\dot{r}^2+r^2\dot{\theta}^2+r^2\sin^2(\theta)\dot{\phi}^2\right)\right)\mathrm{d}t\\&=\int\limits_{t_1}^{t_2}\left(1-\frac{GM}{c^2(R+h)}-\frac{(v_{\text{airplane}}+(R+h)\Omega\sin(\theta))^2}{2c^2}\right)\mathrm{d}t\\&=\left(1-\frac{GM}{c^2(R+h)}-\frac{(v_{\text{airplane}}+(R+h)\Omega\sin(\theta))^2}{2c^2}\right)\int\limits_{t_1}^{t_2}\mathrm{d}t\\&=\left(1-\frac{GM}{c^2(R+h)}-\frac{(v_{\text{airplane}}+(R+h)\Omega\sin(\theta))^2}{2c^2}\right)\Delta t\end{aligned} \tag{12.4.654}$$

Now let us calculate the ratio:

$$\frac{\Delta\tau_{\text{airplane}}}{\Delta\tau_{\text{ground}}}\cong\frac{\left(1-\dfrac{GM}{c^2(R+h)}-\dfrac{(v_{\text{airplane}}+(R+h)\Omega\sin(\theta))^2}{2c^2}\right)\Delta t}{\left(1-\dfrac{GM}{c^2R}-\dfrac{\Omega^2R^2\sin^2(\theta)}{2c^2}\right)\Delta t} \tag{12.4.655}$$

Obviously, we see that the Δt will vanish! We use also again (...):

$$\frac{1}{1+x}\cong 1-x \tag{12.4.656}$$

Then:

$$\frac{\Delta\tau_{\text{airplane}}}{\Delta\tau_{\text{ground}}}\cong\left(\left(1-\frac{GM}{c^2(R+h)}-\frac{(v_{\text{airplane}}+(R+h)\Omega\sin(\theta))^2}{2c^2}\right)\right)\left(1+\frac{GM}{c^2R}+\frac{\Omega^2R^2\sin^2(\theta)}{2c^2}\right) \tag{12.4.657}$$

If we neglect all terms in $1/c^4$ it will remain:

$$\begin{aligned}\frac{\Delta\tau_{\text{airplane}}}{\Delta\tau_{\text{ground}}}&\cong 1-\frac{GM}{c^2(R+h)}-\frac{(v_{\text{airplane}}+(R+h)\Omega\sin(\theta))^2}{2c^2}+\frac{GM}{c^2R}+\frac{\Omega^2R^2\sin^2(\theta)}{2c^2}\\&=1-\frac{GM}{c^2(R+h)}-\frac{1}{2c^2}\left[v_{\text{airplane}}^2+2v_{\text{airplane}}(R+h)\Omega\sin(\theta)+(R+h)^2\Omega^2\sin^2(\theta)\right]+\frac{GM}{c^2R}+\frac{\Omega^2R^2\sin^2(\theta)}{2c^2}\end{aligned} \tag{12.4.658}$$

Neglecting the terms involving Ω^2 that are very small, it remains:

$$\begin{aligned}\frac{\Delta\tau_{\text{airplane}}}{\Delta\tau_{\text{ground}}}&\cong 1-\frac{GM}{c^2(R+h)}-\frac{1}{2c^2}\left[v_{\text{airplane}}^2+2v_{\text{airplane}}(R+h)\Omega\sin(\theta)\right]+\frac{GM}{c^2R}\\&=1+\frac{GMh}{c^2R(R+h)}-\frac{1}{2c^2}\left[v_{\text{airplane}}^2+2v_{\text{airplane}}(R+h)\Omega\sin(\theta)\right]\end{aligned} \tag{12.4.659}$$

Now if $R\gg h$ we can write:

$$\frac{\Delta\tau_{\text{airplane}}}{\Delta\tau_{\text{ground}}}\cong 1+\frac{GMh}{c^2R^2}-\frac{1}{2c^2}\left[v_{\text{airplane}}^2+2v_{\text{airplane}}R\Omega\sin(\theta)\right] \tag{12.4.660}$$

Let us subtract -1 on both side of equality:

$$\frac{\Delta\tau_{\text{airplane}}}{\Delta\tau_{\text{ground}}}-1\cong\frac{GMh}{c^2R^2}-\frac{1}{2c^2}\left[v_{\text{airplane}}^2+2v_{\text{airplane}}R\Omega\sin(\theta)\right] \tag{12.4.661}$$

That is:

$$\frac{\Delta\tau_{\text{airplane}}-\Delta\tau_{\text{ground}}}{\Delta\tau_{\text{ground}}}-1\cong\frac{GMh}{c^2R^2}-\frac{1}{2c^2}\left[v_{\text{airplane}}^2+2v_{\text{airplane}}R\Omega\sin(\theta)\right] \tag{12.4.662}$$

That is more often written:

$$\boxed{\frac{\Delta\tau_{\text{airplane}}-\Delta\tau_{\text{ground}}}{\Delta\tau_{\text{ground}}}-1\cong\frac{GM}{c^2R}\frac{h}{R}-\frac{v_{\text{airplane}}}{c^2}\left(\frac{v_{\text{airplane}}}{2}+R\Omega\sin(\theta)\right)} \tag{12.4.663}$$

The first term (with $h=10,000$ [m] and $R=6,371$ [km] and $c=299,792,458\,[\text{m}\cdot\text{s}^{-1}]$):

$$\frac{GM}{c^2R}\frac{h}{R}\cong 1.091\cdot 10^{-12} \tag{12.4.664}$$

is a pure gravitational effect, always positive, that we already know under the name "gravitational time dilatation", implies that the observer in the aircraft ages faster than the observer on the ground, the latter being deeper in the gravitational field of the Earth. This term is here because of our General Relativity approach!

The second term:

$$-\frac{v_{\text{airplane}}}{c^2}\left(\frac{v_{\text{airplane}}}{2}+R\Omega\sin(\theta)\right) \tag{12.4.665}$$

Corresponds to the dilatation of the times of the moving bodies in Special Relativity.

For a practical numerical application we will take again (see section Special Relativity page 5664):

$$v_{\text{airplane}} \cong 220\,[\mathrm{m}\cdot\mathrm{s}^{-1}] \tag{12.4.666}$$

and the total flight also of 41 hour according to measurement on the ground:

$$\Delta t = 147,600\,[\mathrm{s}] \tag{12.4.667}$$

and latitude of 38° (Washington airport) corresponding to a colatitude $\theta = 52°$. Then we have for the flight to the East ($v_{\text{airplane}} \cong +220\,[\mathrm{m}\cdot\mathrm{s}^{-1}]$):

$$-\frac{v_{\text{airplane}}}{c^2}\left(\frac{v_{\text{airplane}}}{2}+R\Omega\sin(\theta)\right) \cong -1.16\cdot 10^{-12} \tag{12.4.668}$$

and for the flight to the West ($v_{\text{airplane}} \cong -220\,[\mathrm{m}\cdot\mathrm{s}^{-1}]$):

$$-\frac{v_{\text{airplane}}}{c^2}\left(\frac{v_{\text{airplane}}}{2}+R\Omega\sin(\theta)\right) \cong 62.44\cdot 10^{-12} \tag{12.4.669}$$

Thus we have:

$$\begin{aligned}\frac{\Delta\tau_{\text{airplane,East}}-\Delta\tau_{\text{ground}}}{\Delta\tau_{\text{ground}}}-1 &\cong \frac{GM}{c^2R}\frac{h}{R}-\frac{v_{\text{airplane}}}{c^2}\left(\frac{v_{\text{airplane}}}{2}+R\Omega\sin(\theta)\right)\cong -7.1\cdot 10^{-14}\\ \frac{\Delta\tau_{\text{airplane,West}}-\Delta\tau_{\text{ground}}}{\Delta\tau_{\text{ground}}}-1 &\cong \frac{GM}{c^2R}\frac{h}{R}-\frac{v_{\text{airplane}}}{c^2}\left(\frac{v_{\text{airplane}}}{2}+R\Omega\sin(\theta)\right)\cong 1.72\cdot 10^{-12}\end{aligned} \tag{12.4.670}$$

To estimate $\Delta\tau_{\text{East}}$ and $\Delta\tau_{\text{West}}$ we just have to multiply the relative values above by the total duration of the flight:

$$\begin{aligned}\Delta\tau_{\text{East}} &\cong -10.48\,[\mathrm{ns}]\\ \Delta\tau_{\text{West}} &\cong +253\,[\mathrm{ns}]\end{aligned} \tag{12.4.671}$$

Comparing with the data in the statement, and the result we get with this experiment treatment in the Special Relativity section we draw two conclusions:

1. The above results, based on the very simplified trajectories of the airplanes (their altitude not being constant and their speed either) are in relatively good agreement with the theoretical predictions based on the real trajectories, that is to say using:

   $$\Delta\tau_{\text{airplane}} = \int_{t_1}^{t_2}\left(1-\frac{GM}{c^2(R+h)}-\frac{(v_{\text{airplane}}+(R+h)\Omega\sin(\theta))^2}{2c^2}\right)\mathrm{d}t \tag{12.4.672}$$

   rather than the very very very (...) approximate relation:

   $$\frac{\Delta\tau_{\text{airplane}}-\Delta\tau_{\text{ground}}}{\Delta\tau_{\text{ground}}}-1 \cong \frac{GM}{c^2R}\frac{h}{R}-\frac{v_{\text{airplane}}}{c^2}\left(\frac{v_{\text{airplane}}}{2}+R\Omega\sin(\theta)\right) \tag{12.4.673}$$

2. The results we get just here are less accurate relatively to measurement than the results we get using Special Relativity. Indeed, measurement give $\Delta\tau_{\text{West}}=+273\pm 7$ [ns], our Special Relativity treatment $\Delta\tau_{\text{West}}=+265$ [ns] and we get here $\Delta\tau_{\text{West}}=+253$ [ns], so we lost accuracy with the above approximation for the West. The measurement give $\Delta\tau_{\text{East}}=-59\pm 10$ [ns], our Special Relativity treatment $\Delta\tau_{\text{West}}=+156$ [ns] and we get here $\Delta\tau_{\text{West}}=+10.48$ [ns], so we gain accuracy with the above approximation for the East.

12.4.6.6 Black Holes

Some physicists suspect Black Holes might not exist outside theorists' imaginations. Skeptics argued that Black Holes might be only an artifact of General Relativity's subtle math, or that they might only form under unrealistic conditions, such as the collapse of a perfectly spherical star. However, in the late 11960s (holocene calendar), Roger Penrose dispelled such doubts with rigorous math, for which he shared the 12020 (holocene calendar) Nobel Prize in Physics. Penrose exactly proved that, no, no, even if you have a lumpy thing, as long as the density became high enough, it was going to collapse to a Black Hole!

There are four types of Black Holes that have quite very different characteristics:

- **Schwarzschild Black Holes:** That have zero angular momentum (non-rotating), they are quite unrealistic and unlikely because most things in the universe have an angular momentum. Schwarzschild Black Holes have a punctual singularity, they are spherically symmetric with no electric charge.
- **Kerr Black Holes:** That have an angular momentum (rotating), they are quite much more realistic and likely than Schwarzschild Black Holes as most things in the universe have an angular momentum. Kerr Black Holes are described by the Kerr metric and doesn't have a punctual singularity. The angular momentum causes the Kerr Black Holes to exhibit properties such as frame-dragging and the presence of an inner and outer event horizon.
- **Reissner-Nordstrom Black Holes:** They are non-rotating Black Holes (like Schwarzschild Black Holes) but with an electric charge.
- **Kerr-Newman Black Holes:** They are rotating Black Holes (like Kerr Black Holes) but with an electric charge.

Actually the texts below deal only with Schwarzschild Black Holes as the maths for the other Black Holes are quite more complicated to introduce. They will very likely be available for the next edition of this textbook!

But the doubts, as always in Science, subsist. Scientists hope to answer three specific questions at the day we write this textbook: Do the actual observed Black Holes really have event horizons? Are they as featureless as the no-hair theorem says? And do they distort space-time exactly as the Kerr metric predicts? The future will tell us!

Remark

Actually the nearest confirmed Black Hole to our Solar System is named Gaia BH1 (TIC 125470397), discovered in the year 12022 (holocene calendar) at a distance of $1,560\pm 10$ light-years in a binary system in the Unicorn constellation.

12.4.6.6.1 Schwarzschild Black Holes

But now let us focus on the basics. And for this purpose we stay focused on our Schwarzschild metric ....! We know that a radial trajectory of light-type implies:

$$\mathrm{d}s^2=0 \tag{12.4.674}$$

therefore:

$$\left(1-\frac{2GM}{c^2r}\right)c^2\mathrm{d}t^2-\frac{1}{\left(1-\dfrac{2GM}{c^2r}\right)}\mathrm{d}r^2-r^2\left(\mathrm{d}\theta^2+\sin^2(\theta)\mathrm{d}\phi^2\right)=0 \tag{12.4.675}$$

and in a direct radial trajectory (by definition!) we also have:

$$\mathrm{d}\theta=0 \quad \text{and} \quad \mathrm{d}\phi=0 \tag{12.4.676}$$

therefore:

$$\left(1-\frac{2GM}{c^2r}\right)c^2\mathrm{d}t^2-\frac{1}{\left(1-\dfrac{2GM}{c^2r}\right)}\mathrm{d}r^2=0 \tag{12.4.677}$$

Therefore:

$$\left(1-\frac{2GM}{c^2r}\right)c^2\mathrm{d}t^2=\frac{1}{\left(1-\dfrac{2GM}{c^2r}\right)}\mathrm{d}r^2 \tag{12.4.678}$$

hence:

$$\frac{\mathrm{d}t^2}{\mathrm{d}r^2}=\frac{1}{\left(1-\dfrac{2GM}{c^2r}\right)^2c^2} \tag{12.4.679}$$

so that:

$$\frac{\mathrm{d}t}{\mathrm{d}r}=\pm\frac{1}{c}\left(1-\frac{2GM}{c^2r}\right)^{-1} \tag{12.4.680}$$

Let us change to natural units $c=1$. It then comes:

$$\frac{\mathrm{d}t}{\mathrm{d}r}=\pm\frac{1}{\left(1-\dfrac{2GM}{r}\right)} \tag{12.4.681}$$

When $r\rightarrow 2GM$ the right-hand side of this equality tends to $\pm\infty$, then the evolution of time t (external observer) as a function of r tends to infinity with respect to the proper time of light.

The sphere given by the radius:

$$\boxed{r=R_S=\frac{2GM}{c^2}} \tag{12.4.682}$$

defines the "horizon" of a "Schwarzschild Black Hole".

Towards this limit boundary, the light seems to put an infinite time compared to an external observer to move when approaching a Black Hole. It therefore never really reaches it in relation to the observer, hence the fact that the Black Holes can be surrounded according to their environment by a luminous halo near the Schwarzschild radius. Moreover, since time seems to be stopped, the frequency of the light surrounding the Black Hole tends towards zero and therefore towards the infra-red.

A Black Hole is therefore a region of space-time exhibiting such strong gravitational effects that nothing—not even particles and electromagnetic radiation such as light—can escape from inside it. The theory of General Relativity predicts that a sufficiently compact mass can deform space-time to form a Black Hole. The boundary of the region from which no escape is possible is named the "event horizon".

Objects whose gravitational fields are too strong for light to escape were first considered in the 118th century (holocene calendar) by John Michell and Pierre-Simon Laplace. The first modern solution of General Relativity that would characterize a Black Hole was found by Karl Schwarzschild in 11916 (holocene calendar), although its interpretation as a region of space from which nothing can escape was first published by David Finkelstein in 11958 (holocene calendar). Black Holes were long considered a mathematical curiosity; it was during the 11960s (holocene calendar) that theoretical work showed they were a generic prediction of General Relativity. The discovery of neutron stars sparked interest in gravitationally collapsed compact objects as a possible astrophysical reality.

Black Holes of stellar mass are expected to form when very massive stars collapse at the end of their life cycle. After a Black Hole has formed, it can continue to grow by absorbing mass from its surroundings. By absorbing other stars and merging with other Black Holes, supermassive Black Holes of millions of solar masses may form. There is general consensus that supermassive Black Holes exist in the centers of most galaxies.

Despite its invisible interior, the presence of a Black Hole can be inferred through its interaction with other matter and with electromagnetic radiation such as visible light. Matter that falls onto a Black Hole can form an external accretion disk heated by friction, forming some of the brightest objects in the universe. If there are other stars orbiting a Black Hole, their orbits can be used to determine the Black Hole's mass and location. Such observations can be used to exclude possible alternatives such as neutron stars. In this way, astronomers have identified numerous stellar Black Hole candidates in binary systems, and established that the radio source known as Sagittarius A*, at the core of our own Milky Way galaxy, contains a supermassive Black Hole of about 4.3 million solar masses.

Remark

If the Sun or any other celestial object collapse into a Black Hole, this doesn't affect its orbiting element as the Black Holes still have the same mass (only the density change) and Newton's law still remains valid at a quite significant distant of it. So it is wrong to imagine that if the Sun collapse into a Black Hole of 6 [km] diameter it would exert a more big gravitational force and that all planets turning around it will be sucked in.

A common question that should perhaps arise to some readers now is... : how does gravity escape a Black Hole? The answer to this questions should first make a distinction between gravitational field and gravitational waves that are perturbations of space-time and that can't indeed escape from the inside of the Black Hole but that are mathematically also generated outside of the latter. For the gravitational field... it must be understood that gravity does not "travel". In fact, gravity is the result of space-time distortion, it is space itself and isn't a thing that travels through space, and Black Holes distort space-time not only inside but outside their event horizon. Gravitational waves (see further below) by cons, travel a the speed of light and emerge when the distortion changes over time.

On 11 February 12016 (holocene calendar), the LIGO collaboration announced the first observation of gravitational waves; because these waves were generated from a Black Hole merger it was the first ever direct detection of a binary Black Hole merger. On 15 June 12016, a second detection of a gravitational wave event from colliding Black Holes was announced.

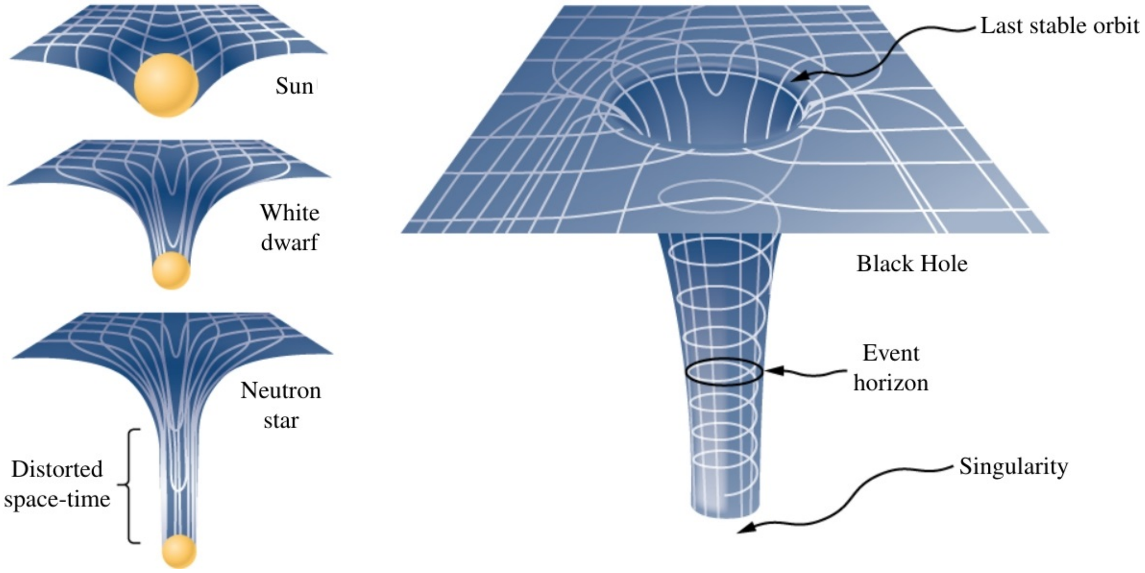


Figure 12.230: **2D naive representation of space curve near some celestial objects (source: OpenStax)**

The Galactic Center Group members have been measuring the positions of thousands of stars in the vicinity of the Galactic Center for more than 20 years. This unique data set allowed us to measure directly short-period orbits of stars. In particular, a full phase coverage has been measured for two stars: S0-2 with an orbital period of 15.56 years, and S0-102 with 11.5 years. At the closest approach, S0-2 is only 17 light hours away from the center of the Galaxy, about four times the distance of Neptune from the Sun. From these orbital data, we can determine the mass of the central Black Hole in our own Galaxy.

The Milky Way Galaxy center is the best candidate to what seems to a Black Hole, and especially and example of the closest supermassive Black Holes (SMBH), located at $\sim 25,000$ light years away from us and corresponds with the location of Sagittarius A* a bright and very compact astronomical radio source at the center of the Milky Way. Its mass is estimated to be 4 million times the mass of the Sun, which implies that the Schwarzschild radius is about 17 times that of Sun's radius. As a comparison, Mercury's orbit is located at a distance of ~ 83 solar radii. Because the Galactic Center is the site of the closest supermassive Black Hole by a factor of 100, it is a unique laboratory for solving some of the greatest mysteries associated with the fundamental physics of supermassive Black Holes and the role that they play in the formation and evolution of galaxies. Furthermore, it is the only galactic nucleus in which direct measurements of stellar orbits is possible, with either the current or the next-generation instruments.

Observations of stellar orbits around the Galactic Black Hole also yields precision measurement of the distance to the Galactic Center, which is important as it affects almost all questions not only of Galactic structure, dynamics and mass, but those of extragalactic distance scales and the value of Hubble's constant as well.

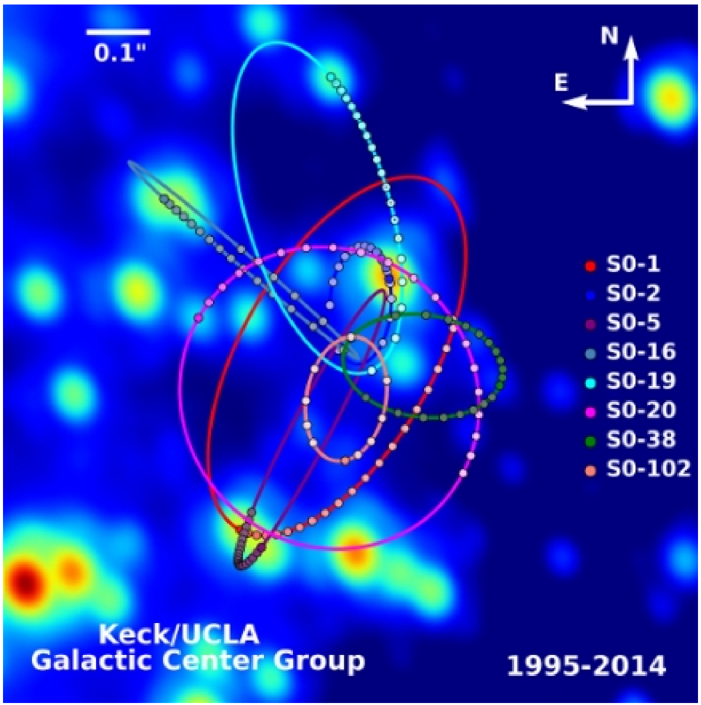


Figure 12.231: Orbits of stars within the central 1.0×1.0 arcseconds of our Galaxy (source: Galactic Group Center)

The origin of supermassive Black Holes and also the assumption that each galaxy center is a Black Hole remains an open field of research. Astrophysicists agree that once a Black Hole is in place in the center of a galaxy, it can grow by accretion of matter and by merging with other Black Holes this is why some observations gives for example for the hyperluminous quasar S5 0014+81 a weight of approximately $40 \cdot 10^{10}$ times the mass of the Sun.

Remark

The Schwarzschild radius scales with mass as $R_S = 2GM/c^2$. What might be defined as a "Schwarzschild volume" would then be $V_S = 4\pi R_S^3/3 = (32/3)\pi(GM/c^2)^3$. So the density of matter defined by the horizon is $\rho_S = (3/32)(c^2/G)^3 M^{-2}$. So the Schwarzschild density scales as the inverse square of the mass. A 10 billion solar mass Black Hole (like the estimation of the one at the center of our Galaxy) has a radius estimated about $1,010$ [km], or a Schwarzschild volume of $V_S \cong 10^{39}$ [m^3]. A solar mass is 10^{30} [kg] and the Schwarzschild density defined by the horizon is then $\rho_S = 10^{-9}$ [kg $\cdot$ m^{-3}]. That is actually quite small. However this may be misleading. "Density" suggests that the mass is distributed more or less uniformly within the Black Hole, and this is non-sense. The Black Hole is mostly empty, and all the mass is concentrated within a tiny region (classically a point) in the center of the Black Hole.

Are there any other magnitudes we should note or calculate in Black Hole physics and thermodynamics! Yes, there are Assuming spherical symmetry, we can calculate the Schwarzschild area or event horizon/surface area of the Schwarzschild's Black Hole simple by:

$$A_S = 4\pi R_S^2 = 4\pi \left(\frac{2GM}{c^2}\right)^2 = \frac{16\pi G^2 M^2}{c^4} \qquad (12.4.683)$$

We can also calculate the surface gravity κ, if the gravitational field of the Black Hole reads:

$$g = \frac{GM}{R^2} \qquad (12.4.684)$$

then, at the Schwarzschild radius it becomes the mentioned surface gravity $\kappa = g(R = R_S)$:

$$\kappa = \frac{c^4}{4GM} \qquad (12.4.685)$$

Interestingly, this surface gravity is $1/M$ the maximal force $c^4/4G$ allowed by natural units... What else? Surface tides, or more precisely, the tidal acceleration at the Black Hole surface. The tidal acceleration is calculated with (the reader notice that we take the version of the tidal force with the factor 2 instead of factor 4 as discusses during the proof of this relation in the section of Astronomy page 5358):

$$a_T = \frac{2GMd}{R^3} \qquad (12.4.686)$$

If it is evaluated at R_S we get:

$$a_T(\text{BH}) = a_T(R = d = R_S) = \frac{c^4}{2GM} \qquad (12.4.687)$$

12.4.6.6.1.1 Black Hole innermost stable circular orbit

The title of this subsection a bit a misnomer. Indeed, the results that we will see also applies to any other object that Black Holes!

Much of the modern folklore about Black Holes is misleading. One idea you may have heard is that Black Holes go about sucking things up with their gravity. Actually, it is only very close to a Black Hole that the strange effects we have been discussing come into play. The gravitational attraction far away from a Black Hole is the same as that of the star that collapsed to form it.

Remember that the gravity of any star some distance away acts as if all its mass were concentrated at a point in the center, which we name the "center of gravity". For real stars, we merely imagine that all mass is concentrated there; for Black Holes, all the mass really is concentrated at a point in the center.

So, if you are a star or distant planet orbiting around a star that becomes a Black Hole, your orbit may not be significantly affected by the collapse of the star (although it may be affected by any mass loss that precedes the collapse). If, on the other hand, you venture close to the event horizon, it would be very hard for you to resist the "pull" of the warped space-time near the Black Hole. You have to get really close to the Black Hole to experience any significant effect.

If another star or a spaceship were to pass one or two solar radii from a Black Hole, Newton's laws would be adequate to describe what would happen to it. Only very near the event horizon of a Black Hole is the gravitation so strong that Newton's laws break down. The Black Hole remnant of a massive star coming into our neighbourhood would be far, far safer to us than its earlier incarnation as a brilliant, hot star.

A Black Hole with it's accretion disk ray-traced with Python™[50] by Riccardo Antonelli:

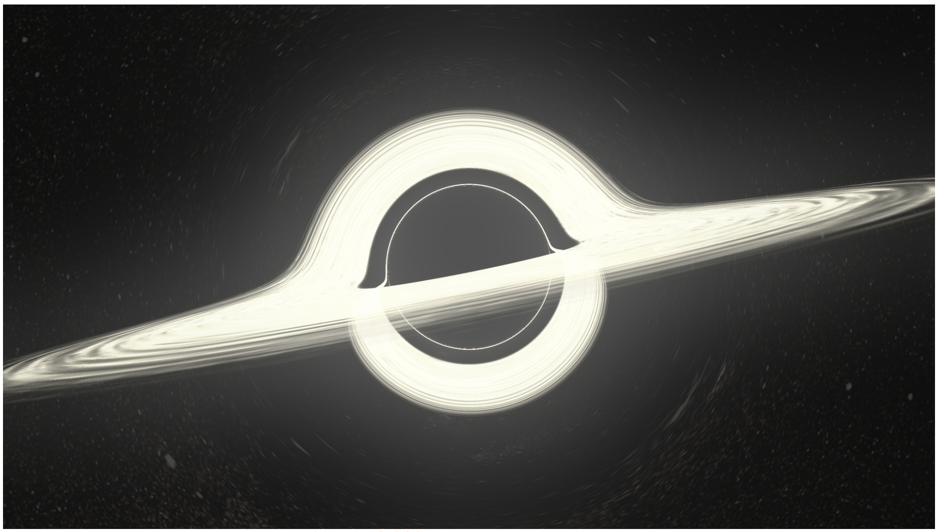


Figure 12.232: **Black Hole ray-traced**

To begin, let us recall first that we have just seen that Schwarzschild metric is an exact solution to the Einstein vacuum equation given by the line element:

$$\mathrm{d}s^2=\left(1-\frac{2GM}{c^2r}\right)c^2\mathrm{d}t^2-\frac{1}{\left(1-\dfrac{2GM}{c^2r}\right)}\mathrm{d}r^2-r^2\left(\mathrm{d}\theta^2+\sin^2(\theta)\mathrm{d}\phi^2\right) \qquad (12.4.688)$$

which written as a matrix consequently becomes:

$$g_{\mu\nu}=\begin{bmatrix}1-\dfrac{2GM}{rc^2} & 0 & 0 & 0\\ 0 & -\left(1-\dfrac{2GM}{rc^2}\right)^{-1} & 0 & 0\\ 0 & 0 & -r^2 & 0\\ 0 & 0 & 0 & -r^2\sin^2(\theta)\end{bmatrix} \qquad (12.4.689)$$

Now let us introduce a special mathematical vector in General Relativity.

Definition (#12.4.546): If all components of a metric are independent of some particular x^ν, then we have the "Killing vector" $\vec{K}$, named after Wilhelm Killing, with components $K^\mu=\delta^\mu_\nu$. That is, the contravariant form just has a constant in the appropriate slot and zeros elsewhere. As we will see, as Noether's theorem tells us (see section Analytical Mechanics page 3132) that for every symmetry there is a conserved quantity (energy, momentum, angular momentum, etc), Killing vectors are a powerful way to find those conserved quantities.

---

[50] Source code available on http://rantonels.github.io/starless

The Schwarzschild metric has a few interesting properties, it is clear that it is independent explicitly of t and ϕ and hence there exist Killing vectors that we will denote $K_1^\mu = (1,0,0,0)$ (often denoted ξ^μ) and $K_2^\mu = (0,0,0,1)$ (often denoted η^μ) which both lie along directions in which the metric doesn't change[51].

We define the velocity four-vector v^μ as:

$$u^\mu = \frac{\mathrm{d}x^\mu}{\mathrm{d}\tau} \tag{12.4.690}$$

where we let indices run over the coordinates (t,r,θ,ϕ) and τ is the proper time. Now, since the metric is independent of t, forming the scalar product between K_1^μ and $u_\mu = g_{\mu\nu}u^\nu$ gives the quantity:

$$g_{\mu\nu}K_1^\mu u^\nu = \left(1-\frac{GM}{rc^2}\right)u^0 = \left(1-\frac{GM}{rc^2}\right)\frac{c\mathrm{d}t}{\mathrm{d}\tau} \tag{12.4.691}$$

But we know from our study of Special Relativity that:

$$\frac{c\mathrm{d}t}{\mathrm{d}\tau} = \frac{1}{1-\left(\dfrac{v}{c}\right)^2} = c\gamma \tag{12.4.692}$$

Therefore:

$$g_{\mu\nu}K_1^\mu u^\nu = \left(1-\frac{GM}{rc^2}\right)c\gamma \tag{12.4.693}$$

Now let us recall that we have proved earlier that the Lagrangian of the geodesic was given by:

$$L = \frac{1}{2}mg_{\alpha\beta}\frac{\partial x^\alpha}{\partial\tau}\partial x^\beta\partial\tau = \frac{1}{2}mg_{\alpha\beta}\dot{x}^\alpha\dot{x}^\beta \tag{12.4.694}$$

Therefore:

$$L = \frac{1}{2}m\left[\left(1-\frac{2GM}{rc^2}\right)\dot{t}^2 - \left(1-\frac{2GM}{rc^2}\right)^2\dot{r}^2 - r^2\dot{\theta}^2 - r^2\sin^2(\theta)\dot{\phi}^2\right] \tag{12.4.695}$$

We see that the Lagrangian doesn't depend explicitly on t and ϕ so these two parameters implies conservation of energy.

But now we know from our study of Analytical Mechanics (see page 3181) that:

$$\frac{\partial L}{\partial\dot{q}} = p \tag{12.4.696}$$

That in our case will be written (as there is conservation according to component t):

$$\frac{\partial L}{\partial\dot{t}} = c^{te} \tag{12.4.697}$$

---

[51] There are actually four Killing vectors for the Schwarzschild metric since it is invariant under time translations and rotations. But we will only need two of them for our purpose.

Therefore:

$$m\left(1-\frac{2GM}{rc^2}\right)\dot{t}=c^{te} \tag{12.4.698}$$

Hence:

$$g_{\mu\nu}K_1^{\mu}u^{\nu}=\left(1-\frac{GM}{rc^2}\right)\frac{\mathrm{d}t}{\mathrm{d}\tau}=\left(1-\frac{GM}{rc^2}\right)u^t=\left(1-\frac{GM}{rc^2}\right)c\gamma=c^{te} \tag{12.4.699}$$

That we will denote with the lowercase letter e:

$$e=\left(1-\frac{GM}{rc^2}\right)u^t=c^{te} \tag{12.4.700}$$

and is an energy per unit rest mass.

Similarly since the metric is independent of ϕ we form:

$$g_{\mu\nu}K_2^{\mu}u^{\nu}=-r^2\sin^2(\theta)\frac{\mathrm{d}\phi}{\mathrm{d}\tau} \tag{12.4.701}$$

But we also know from our study of Analytical Mechanics that:

$$\frac{\partial L}{\partial \dot{q}}=p \tag{12.4.702}$$

And as our Lagrangian above is explicitly independent of ϕ, we have:

$$\frac{\partial L}{\partial \dot{\phi}}=-mr^2\sin^2(\theta)\dot{\phi}=p=c^{te} \tag{12.4.703}$$

Therefore we denote with the lowercase letter l:

$$l=g_{\mu\nu}K_2^{\mu}u^{\nu}=-r^2\sin^2(\theta)\frac{\mathrm{d}\phi}{\mathrm{d}\tau}=-r^2\sin^2(\theta)u^{\phi}=c^{te} \tag{12.4.704}$$

the angular momentum per unit rest mass.

This is exactly the kind of useful thing we need to consider particle orbits! The conserved energy and angular momentum are one of the fundamental descriptors for orbits from Classical Mechanics as we know!

Caution! The Friedmann-Robertson-Lemaître-Walker metric or the de Sitter metric have for example time-dependant factor, this prevents a time-like Killing vector and therefore conservation of energy! In the case of cosmologies, these metrics FLRW etc, are "type O Petrov-Penrose-Pirani solutions", which have no isometries or Killing vectors. There may be timelike Killing vectors on the patches associated with those metrics, but there is no global timelike Killing vector!

Now before we continue, consider the following result (with the $(+,-,-,-)$ metric):

$$u_{\mu}u^{\mu}=g_{\mu\nu}u^{\mu}u^{\nu}=g_{\mu\nu}\frac{\mathrm{d}x^{\mu}}{\mathrm{d}\tau}\frac{\mathrm{d}x^{\nu}}{\mathrm{d}\tau}=\frac{g_{\mu\nu}\mathrm{d}x^{\mu}\mathrm{d}x^{\nu}}{\mathrm{d}\tau^2}=\frac{\mathrm{d}s^2}{\mathrm{d}\tau^2}=\frac{c^2\mathrm{d}\tau^2}{\mathrm{d}\tau^2}=c^2 \tag{12.4.705}$$

that is for our geodesic observer a normalization constraint that will be useful to us in our study of the Borde-Guth-Vilenkin theorem (see section Cosmogony page 1383).

Remark

In textbooks this relation is often given with the $(-,+,+,+)$ signature and in natural units, thus:

$$u_\mu u^\mu = g_{\mu\nu} u^\mu u^\nu = -1 \tag{12.4.706}$$

Hence:

$$\left(1-\dfrac{2GM}{rc^2}\right)(u^t)^2-\left(1-\dfrac{2GM}{rc^2}\right)^{-1}(u^r)^2-r^2(u^\theta)^2-r^2\sin^2(\theta)(u^\phi)^2=c^2 \tag{12.4.707}$$

We are free to choose $\theta=\pi/2$, exploiting the spherical symmetry of Schwarzschild metric, to simplify the calculations (as $\sin(\pi/2)=1$) and since we fix that value, we also have $u^\theta=0$. Therefore the previous relation reduce to:

$$\left(1-\dfrac{2GM}{rc^2}\right)(u^t)^2-\left(1-\dfrac{2GM}{rc^2}\right)^{-1}(u^r)^2-r^2(u^\phi)^2=c^2 \tag{12.4.708}$$

Now we exploit the relations:

$$e=\left(1-\dfrac{GM}{rc^2}\right)u^t \qquad \text{and} \qquad l=-r^2\sin^2(\theta)u^\phi=-r^2u^\phi \tag{12.4.709}$$

That we inject in the previous one and we get:

$$\left(1-\dfrac{2GM}{rc^2}\right)^{-1}e^2-\left(1-\dfrac{2GM}{rc^2}\right)^{-1}(u^r)^2-\dfrac{l^2}{r^2}=c^2 \tag{12.4.710}$$

Or (this is the form as we found it the most often in textbooks):

$$-\left(1-\dfrac{2GM}{rc^2}\right)^{-1}e^2+\left(1-\dfrac{2GM}{rc^2}\right)^{-1}(u^r)^2+\dfrac{l^2}{r^2}=-c^2 \tag{12.4.711}$$

Multiplying both side by $-\left(1-\frac{2GM}{rc^2}\right)^{-1}$ we get first:

$$e^2-(u^r)^2-\left(1-\dfrac{2GM}{rc^2}\right)\dfrac{l^2}{r^2}=\left(1-\dfrac{2GM}{rc^2}\right)c^2 \tag{12.4.712}$$

Rearranging:

$$e^2-c^2=(u^r)^2+\left(1-\dfrac{2GM}{rc^2}\right)\dfrac{l^2}{r^2}-\dfrac{2GM}{r} \tag{12.4.713}$$

Dividing by 2 and developing:

$$\dfrac{e^2-c^2}{2}=(u^r)^2+\dfrac{l^2}{2r^2}-\dfrac{GMl^2}{r^3c^2}-\dfrac{GM}{r} \tag{12.4.714}$$

Remark

Sometimes we have the notation:

$$\frac{e^2-c^2}{2}=\mathcal{E} \tag{12.4.715}$$

That is more often written:

$$\frac{e^2-c^2}{2}=\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2+\left(\frac{l^2}{2r^2}-\frac{GM}{r}-\frac{GMl^2}{r^3c^2}\right)=\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2+U_{\text{eff}}(r) \tag{12.4.716}$$

which is the radial equation in the Schwarzschild geometry with the corresponding effective potential:

$$\boxed{U_{\text{eff}}(r)=\frac{l^2}{2r^2}-\frac{GM}{r}-\frac{GMl^2}{r^3c^2}} \tag{12.4.717}$$

Thus, a system of four coupled differential equations in four variables has, with the help of three constants of the motion, been reduced to a single differential equation in one variable.

The equivalent classical effective potential is quite obviously:

$$U_{\text{eff}}(r)=\frac{l^2}{2r^2}-\frac{GM}{r} \tag{12.4.718}$$

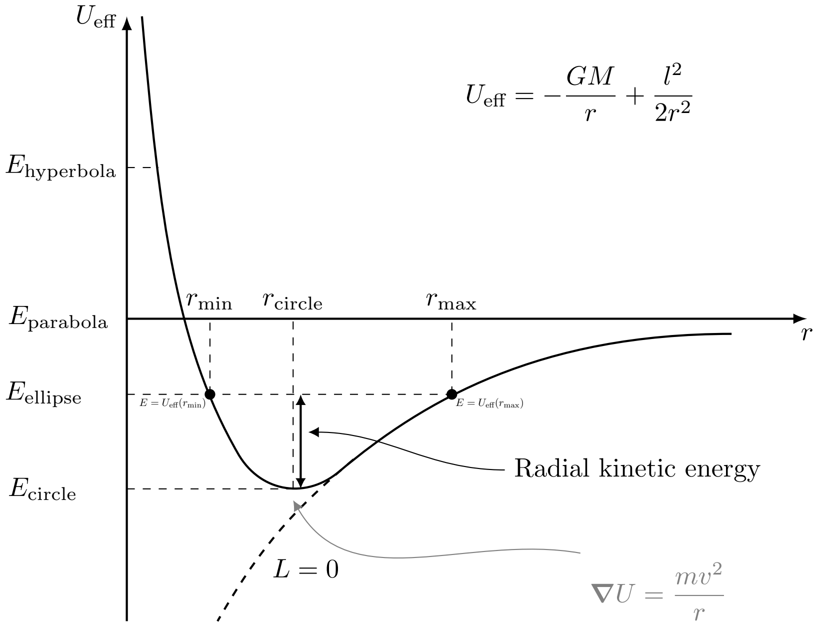

U_{eff}
$U_{\text{eff}}=-\frac{GM}{r}+\frac{l^2}{2r^2}$
$E_{\text{hyperbola}}$
E_{parabola}
E_{ellipse}
E_{circle}
r_{min}
r_{circle}
r_{max}
r
$E=U_{\text{eff}}(r_{\text{min}})$
$E=U_{\text{eff}}(r_{\text{max}})$
Radial kinetic energy
$L=0$
$\nabla U=\frac{mv^2}{r}$

Figure 12.233: Classical effective gravitational potential profile

Remark

The prior-previous relation should be written more explicitly:

$$\frac{E^2-mc^2}{2}=m\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2+\left(\frac{ml^2}{2r^2}-\frac{GMm}{r}-\frac{GMml^2}{r^3c^2}\right) \tag{12.4.719}$$

It is also a very useful quantity to consider here because it is completely analogous to the effective potential from orbital theory seen in the section of Classical Mechanics (see page 3359).

Indeed, consider a particle of mass m orbiting a much heavier object of mass M. Assuming Newtonian mechanics can be used, and the motion of the larger mass is negligible, then the conservation of energy and angular momentum give two constants E and L, with values in polar coordinates:

$$E=E_c+E_p=\frac{1}{2}mv^2-\frac{GmM}{r}=12m\left(\dot{r}^2+r^2\dot{\phi}^2\right)-\frac{GmM}{r} \tag{12.4.720}$$

As the angular momentum is given by:

$$L=mr^2\dot{\phi} \tag{12.4.721}$$

Only two variables are needed, since the motion occurs in a plane. Substituting the second expression into the first and rearranging gives:

$$m\dot{r}^2=2E-\frac{L^2}{mr^2}+\frac{2GmM}{r}=2E-\frac{1}{r^2}\left(\frac{L^2}{m}-2GmMr\right) \tag{12.4.722}$$

Hence:

$$\frac{1}{2}m\dot{r}^2=E-U_{\text{eff}}(r) \tag{12.4.723}$$

That is:

$$E=\frac{1}{2}m\dot{r}^2+U_{\text{eff}}(r)=\frac{1}{2}m\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2+\left(\frac{L^2}{2mr^2}-\frac{GmM}{r}\right) \tag{12.4.724}$$

Therefore we see the analogy now! We have therefore indeed:

$$U_{\text{eff}}(r)=\underbrace{\frac{l^2}{2r^2}-\frac{GM}{r}}_{\text{Newtonian}}-\underbrace{\frac{GMl^2}{r^3c^2}}_{\text{GR correction}} \tag{12.4.725}$$

Remark

we therefore understand why in General Relativity the difference $\mathcal{E}-U_{\text{eff}}(r)$ is therefore interpreted as the kinetic energy.

So what does the extra correction from General Relativity? Look at a graph for $l/M=4.5$, below:

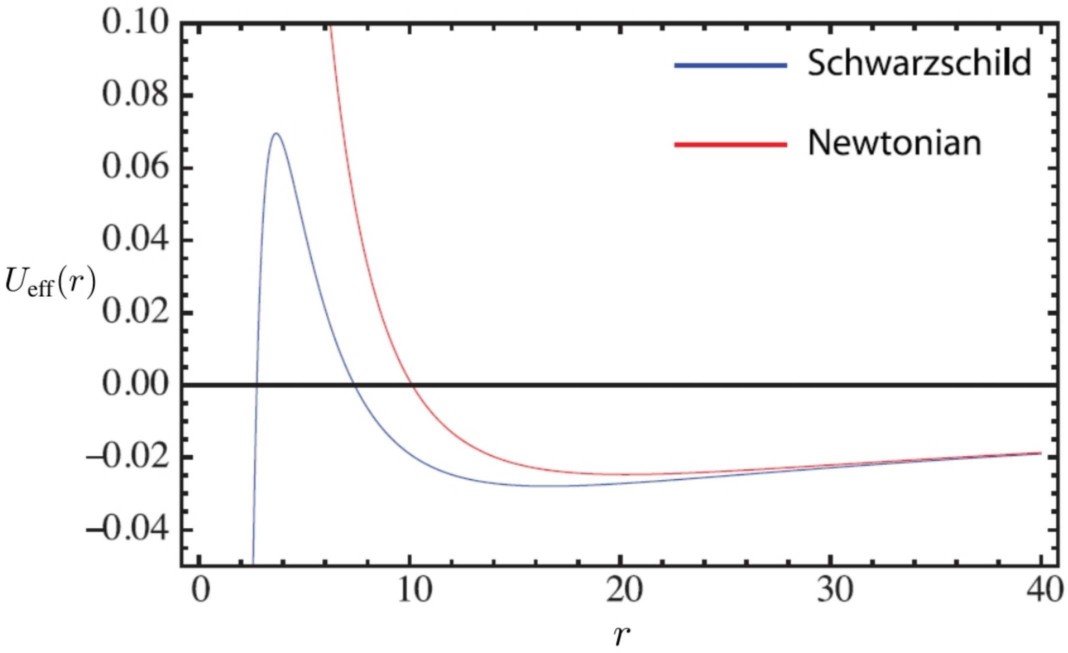


Figure 12.234: Effective potential profile accord to Newtonian and General Relativity mechanics

So what do we notice right away?

There is a minimum in the effective potential, just as in the Newtonian case. For a particle bound in this potential, there is a "stable circular orbit" at the minimum of the potential (it can be shown that this orbit is perfectly circular). At large distances we see that our effective potential tends to the Newtonian one!

However, when $r \rightarrow 0$ our effective potential is dominated by the r^{-3} term, which diverges temporarily. Then, there is a maximum in the effective potential, unlike in the Newtonian case (we then say that the "centrifugal barrier is not impenetrable"). There is an "unstable circular orbit" at the maximum in the potential and it can be show that then the orbit is elliptical and precess like Mercury do (the orbit is not closed).

The shape of this potential is controlled by the magnitude of the angular momentum l.

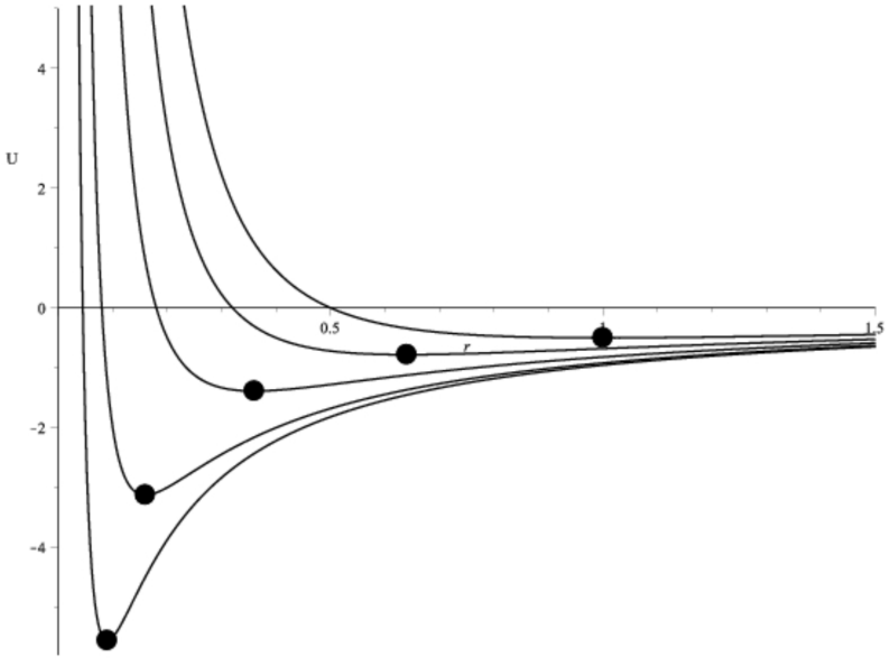


Figure 12.235: Newtonian effective potential for a test particle moving in the gravitational field of a central body with mass, for different values of angular momentum. Solid circles show position of minima corresponding to stable circular orbits.

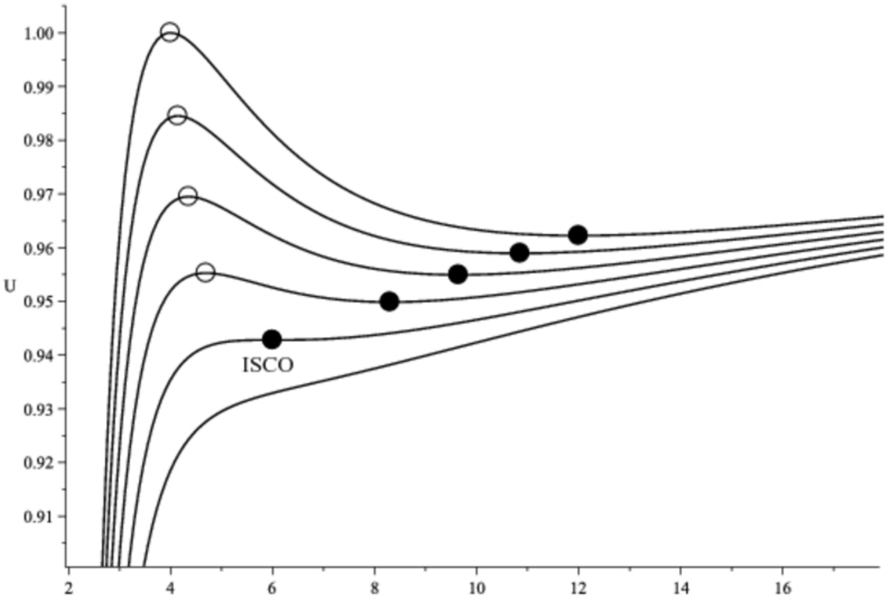


Figure 12.236: Effective potential (per unit particle rest mass) for motion in Schwarzschild metric. Maxima of potential are shown by circles, and minima are shown by solid circles.

We can find the min/max of the potential by constructing a radial derivative:

$$\frac{\mathrm{d}U_{\text{eff}}}{\mathrm{d}r}=0 \qquad (12.4.726)$$

Hence:

$$\frac{\mathrm{d}U_{\text{eff}}}{\mathrm{d}r}=-\frac{l^2}{r^3}+\frac{GM}{r^2}+3\frac{GMl^2}{r^4c^2}=0 \qquad (12.4.727)$$

We multiply both sides by r^4:

$$-l^2 r + GMr^2 + 2GMl^2c^2 = 0 \qquad (12.4.728)$$

This is a simple second degree polynomial, the solution is obviously (see section Calculus page 1383):

$$r_{\text{min,max}} = \frac{cl \pm \sqrt{l^2c^2 - 12G^2M^2}}{2Mc} \qquad (12.4.729)$$

The absolute minimum value for this occurs when[52]:

$$l = \frac{\sqrt{12}MG}{c} \qquad (12.4.730)$$

then:

$$\boxed{r_{\text{min}} = r_{\text{ISCO}} = \frac{6GM}{c^2} = 3R_S} \qquad (12.4.731)$$

where for recall R_S is the Schwarzschild radius. That latter relation is known as the ISCO, the "innermost stable circular orbit" (ISCO) or "marginally stable circular orbit" for the Schwarzschild geometry. We can see that it is equal to three times the Schwarzschild radius (hence, three times the event horizon radius of a Black Hole), for our Sun (if it was a Black Hole) this is almost equal to $9,000$ [km].

The effective potential in this case looks like:

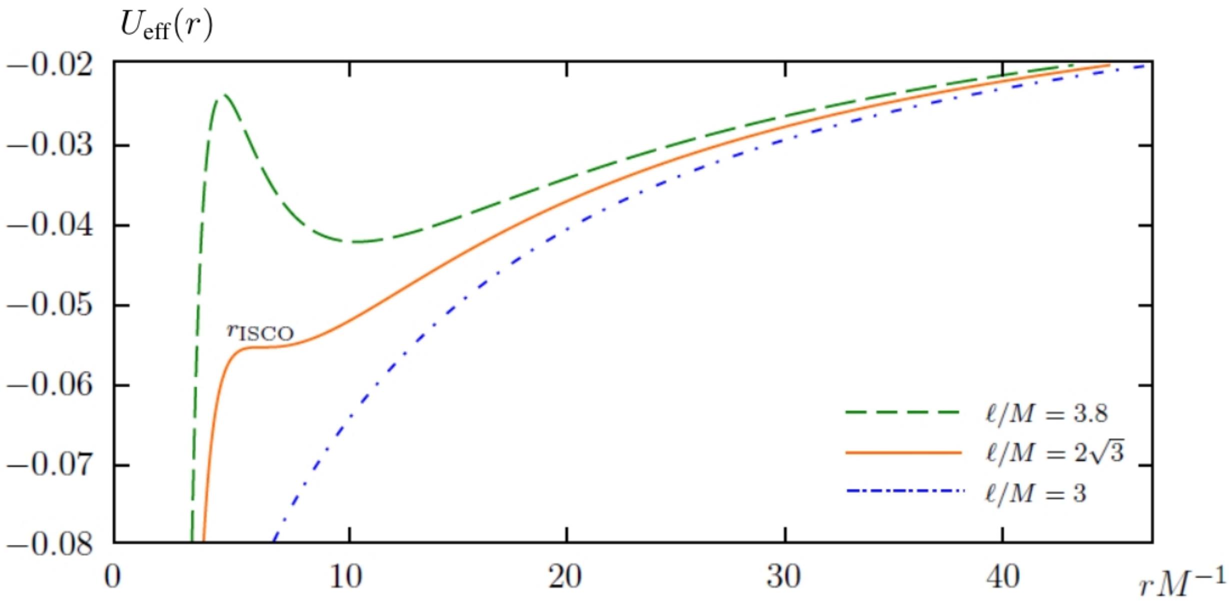


Figure 12.237: **Relativistic effective potentials for various values of l/M with $M = 1$**

There are no stable circular orbits at radii smaller than r_{ISCO}. This has important astrophysical consequences when considering phenomena such as accretion — material slowly works its way down the gravitational potential, giving up energy and angular momentum until it accretes onto the central, compact object. If there is a minimum radius at which material (gas, in the accretion case) can stably orbit for example a Black Hole, it will plunge at all smaller radii for a given angular momentum. This bounds the amount of gravitational binding energy that can be extracted by a particle.

---

[52] Notice that if we put $G = c = 1$ we then have $l/M = \sqrt{12} = 2\sqrt{3}$

So we see this is a huge difference with Newtonian gravity where stable circular orbits are available at all R!

This artist's conception below shows the region hypothesised immediate surrounding a super-massive Black Hole (the black spot near the center). The Black Hole is orbited by a thick disk of hot gas. The center of the disk glows white-hot, while the edge of the disk is shown in dark silhouette. Magnetic fields channel some material into a jet-like outflow - the greenish wisps that extend to upper right and lower left. A dotted line marks the innermost stable circular orbit, which is the closest distance that material can orbit before becoming unstable and plunging into the Black Hole (hence sometimes the expression "plunging region"):

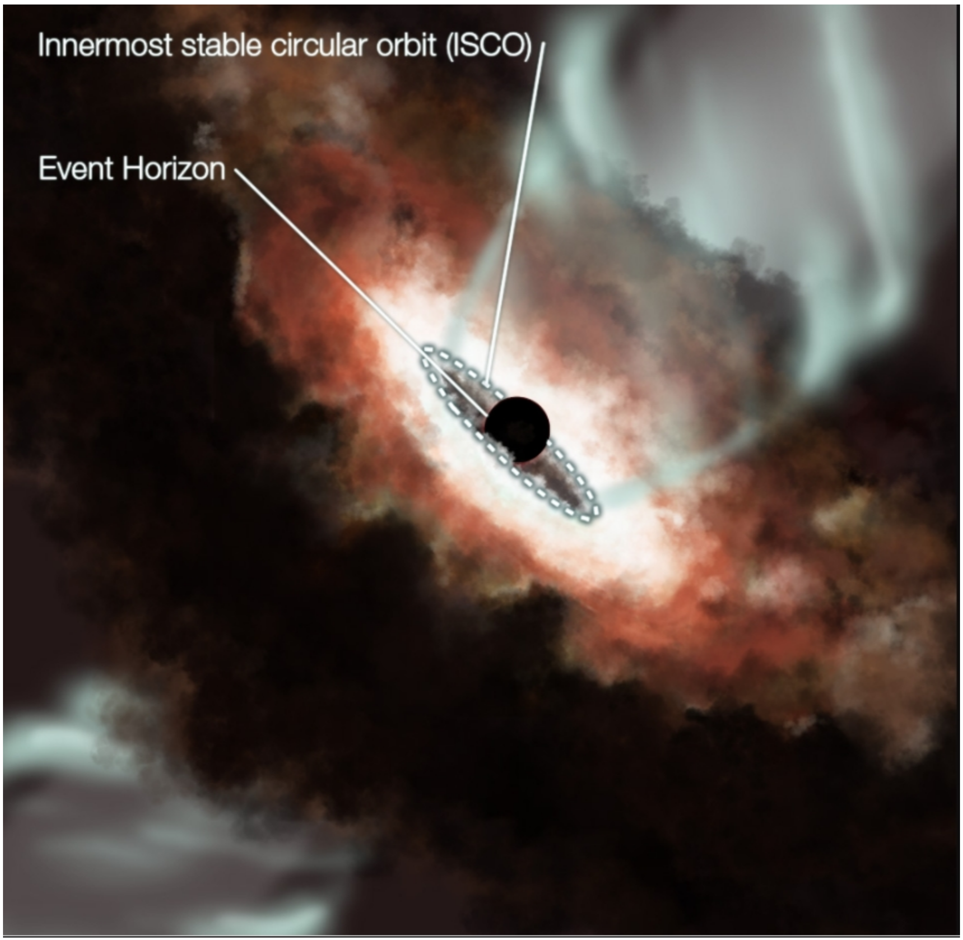

Innermost stable circular orbit (ISCO)
Event Horizon

Figure 12.238: **Black Hole innermost circular stable orbit illustration (source: Chris Fach, Perimeter Institute & University of Waterloo)**

Below the first direct reversed engineering capture of a Black Hole (in fact... of its surrounding), at the center of M87 the 12019-04-10 according to holocene calendar (National Science Foundation) using a planet-scale array of eight ground-based radio telescopes forged through international collaboration named the Event Horizon Telescope (EHT):

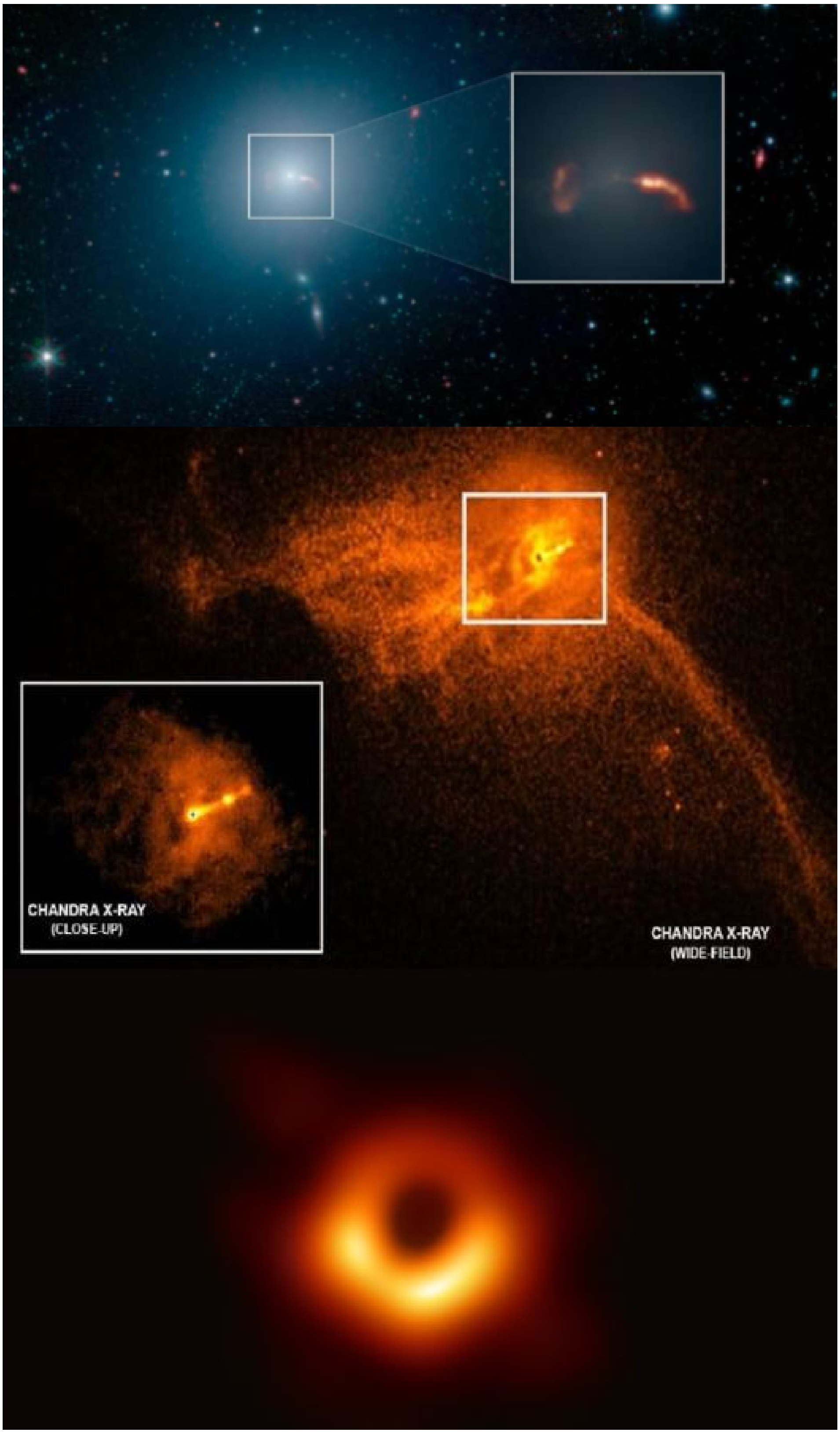


Figure 12.239: **First direct reversed engineering capture of a Black Hole, M87* the 12019-04-10 (source: National Science Foundation)**

On the left below we can see a composite simulated image (five years later than the above composite image, the 27 august 12024 according to holocene calendar) shows how M87* is seen by the Event Horizon Telescope at 86 [GHz] (red), 230 [GHz] (green), and 345 [GHz] (blue). On right, 345 [GHz] is seen in dark blue, a more compact and sharper view of the supermassive Black Holes, followed by 230 [GHz] in green, and 86 [GHz] in red. The higher the frequency, the sharper the image becomes, revealing structure, size, and shape that was previously less discernible (credit: EHT, D. Pesce, A. Chael, see [455]):

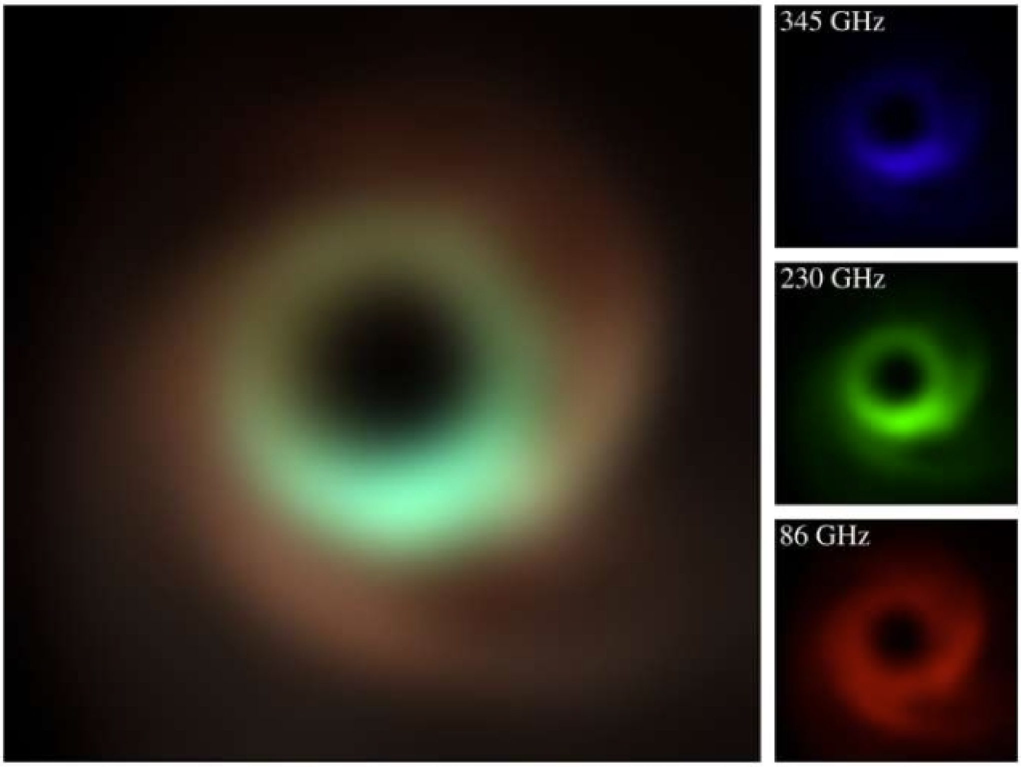


Figure 12.240: Most recent reversed composite image of the M87* Black Hole

Remark

A quite legitimate question would be at this state: If photons can't escape a Black Hole how then gravitons can do it? The answer is that as a force mediating particle, gravitons would come as virtual gravitons, just like virtual photons are the force exchange particles of the electromagnetic force. Virtual particles cannot be said to "travel" anywhere (no virtual photons move in between two interacting charged particles either). That is, they aren't created at one particle just to travel to another and 'exchange' forces between them. So one answer to your question is that nothing comes out of the Black Hole at all. No gravitons move from the Black Hole to the outside world. Unfortunately it's difficult to explain how virtual particles work, as they don't act like normal particles at all. They are better described as transient fluctuations of an underlying field that happen to be described by the same math that describes real particles. Furthermore the graviton is a hypothetical particle in theory of quantum gravity. General Relativity says nothing about gravitons, it just rather describes space-time as continuum.

12.4.6.6.1.2 Black Hole photon sphere

As photons approach the event horizon of a Black Hole, those with the appropriate energy avoid being pulled into the Black Hole (or possibly an "ultra-compact" neutron star) by travelling in a nearly tangential direction known as an "exit cone". A photon on the boundary of this cone does not possess the energy to escape the gravity well of the Black Hole. Instead, it orbits the Black Hole! These orbits are rarely stable in the long term.

The "photon sphere", also named "unstable photon sphere" or "photon ring", is as we will see located farther from the center of a Black Hole than the event horizon, but however nearest that the innermost stable orbit. Within a photon sphere, it is possible to imagine a photon that begins at the back of your head, orbiting the Black Hole, only then to be intercepted by your eyes, allowing you to see the back of your head.

In the previous development, if we focus of the innermost stable orbit, and especially on a photon (denoted γ) universe line, we have $\mathrm{d}s^2=0$, and then the result simplifies to:

$$U_{\gamma,\text{eff}}=\frac{l^2}{2r^2}-\frac{GMl^2}{r^3c^2} \tag{12.4.732}$$

Then following the same reasoning we get to find the min/max of the potential by constructing a radial derivative:

$$\frac{\mathrm{d}U_{\text{eff}}}{\mathrm{d}r}=0 \tag{12.4.733}$$

Hence:

$$\frac{\mathrm{d}U_{\text{eff}}}{\mathrm{d}r}=-\frac{l^2}{r^3}+3\frac{GMl^2}{r^4c^2}=0 \tag{12.4.734}$$

And rearranging gives:

$$\boxed{r_\gamma=\frac{3GM}{c^2}=\frac{3}{2}R_S<r_{\text{ISCO}}} \tag{12.4.735}$$

Therefore, for non-rotating Black Holes, the photon sphere is a sphere of radius $3/2R_S$, outside the event horizon but inside the innermost circular stable orbit.

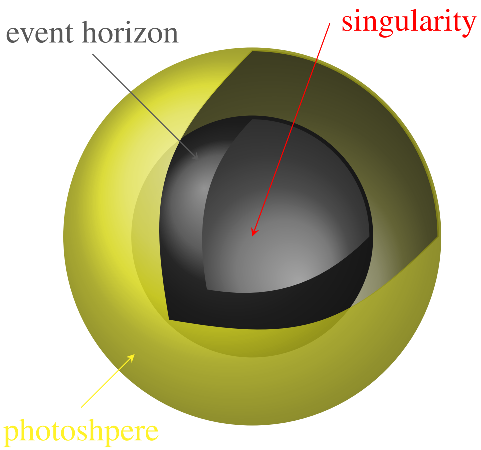
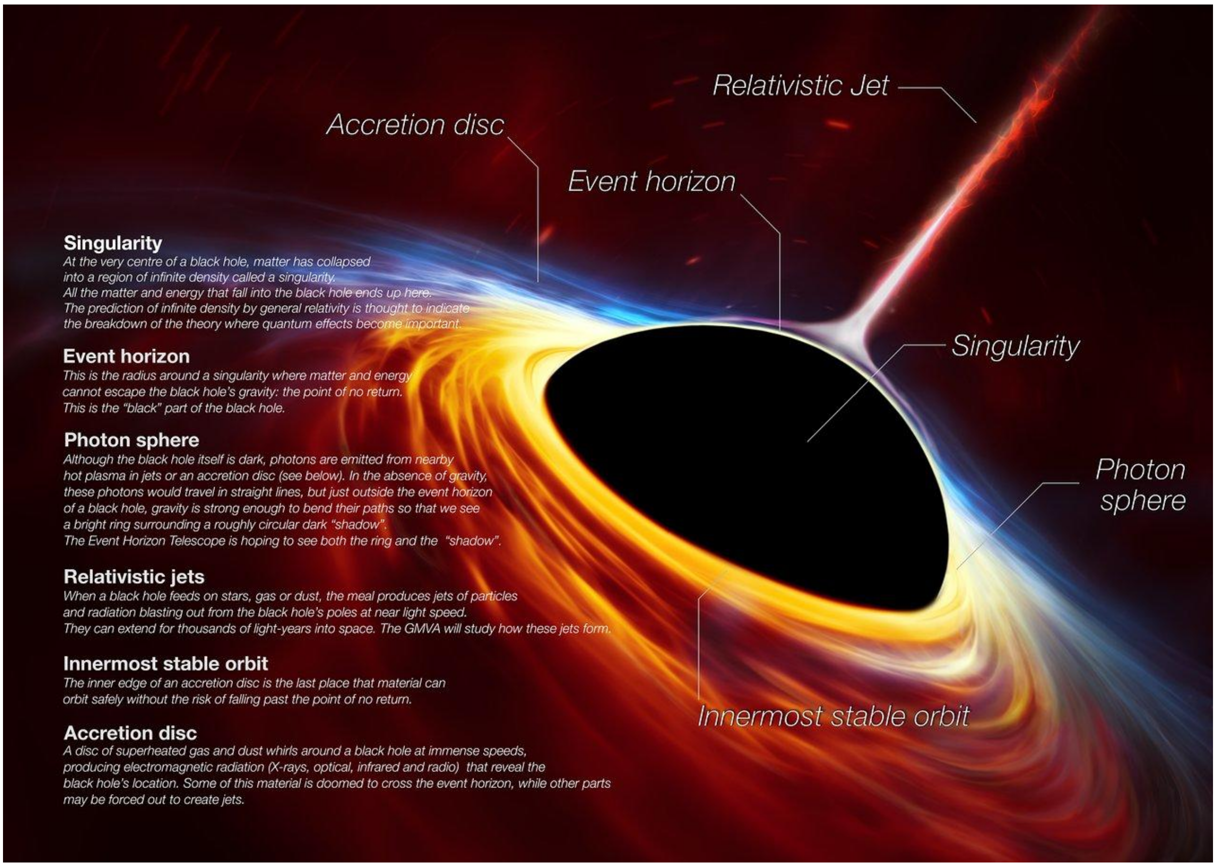


Figure 12.241: Photosphere and event horizon with correct proportions

The latter relation is also sometimes named the "innermost bound circular orbit" (IBCO), which is the unstable (though still a bound) orbit and denoted r_{IBCO}.

And here is a summary of what we have seen so far (the summary is not perfect but it the best we have found so far):



Singularity
At the very centre of a black hole, matter has collapsed into a region of infinite density called a singularity. All the matter and energy that fall into the black hole ends up here. The prediction of infinite density by general relativity is thought to indicate the breakdown of the theory where quantum effects become important.

Event horizon
This is the radius around a singularity where matter and energy cannot escape the black hole's gravity: the point of no return. This is the "black" part of the black hole.

Photon sphere
Although the black hole itself is dark, photons are emitted from nearby hot plasma in jets or an accretion disc (see below). In the absence of gravity, these photons would travel in straight lines, but just outside the event horizon of a black hole, gravity is strong enough to bend their paths so that we see a bright ring surrounding a roughly circular dark "shadow". The Event Horizon Telescope is hoping to see both the ring and the "shadow".

Relativistic jets
When a black hole feeds on stars, gas or dust, the meal produces jets of particles and radiation blasting out from the black hole's poles at near light speed. They can extend for thousands of light-years into space. The GMVA will study how these jets form.

Innermost stable orbit
The inner edge of an accretion disc is the last place that material can orbit safely without the risk of falling past the point of no return.

Accretion disc
A disc of superheated gas and dust whirls around a black hole at immense speeds, producing electromagnetic radiation (X-rays, optical, infrared and radio) that reveal the black hole's location. Some of this material is doomed to cross the event horizon, while other parts may be forced out to create jets.

Figure 12.242: **Black Hole summary (source: Event Horizon Telescope website)**

12.4.6.6.1.3 Black Hole free fall proper time to singularity

Suppose we fall into a Schwarzschild Black Hole. According to General Relativity, we can compute the (finite) free fall time in which we travel from the Schwarzschild radius to the singularity (we assume by the moment General Relativity holds, the point is to what extend is this valid both in General Relativity and the real world, but we can do it as an exercise):

$$\boxed{t_s=\frac{1}{c}\int\limits_0^{R_S}\frac{1}{\sqrt{\dfrac{2GM}{c^2r}-1}}\mathrm{d}r=\frac{\pi}{2}\frac{R_S}{c}=\frac{\pi GM}{c^3}\cong\frac{M}{M_\odot}\cdot 1.54\cdot 10^{-5}\,[\mathrm{s}]}\tag{12.4.736}$$

That latter relation is the free fall radial path time to singularity for a non-rotating Black Hole!

Let us prove that:

$$\int\frac{1}{\sqrt{\dfrac{1}{x}-1}}\mathrm{d}x=-\sqrt{\frac{1}{x}-1}x-\frac{1}{2}\tan^{-1}\left(\sqrt{\frac{1}{x}-1}\right)+c^{te}\tag{12.4.737}$$

For this let us substitute:

$$u=\sqrt{\frac{1}{x}-1}\Leftrightarrow x=\frac{1}{u^2+1}\tag{12.4.738}$$

Then:

$$\mathrm{d}x=-2x^2\sqrt{\frac{1}{x}-1}\mathrm{d}u\tag{12.4.739}$$

Therefore:

$$\int\frac{1}{\sqrt{\dfrac{1}{x}-1}}\mathrm{d}x=\int\frac{-2x^2\sqrt{\dfrac{1}{x}-1}}{\sqrt{\dfrac{1}{x}-1}}\mathrm{d}u=-2\int x^2\mathrm{d}u=-2\int\frac{1}{(u^2+1)^2}\mathrm{d}u\tag{12.4.740}$$

So now we have to solve:

$$\int\frac{1}{(u^2+1)^2}\mathrm{d}u\tag{12.4.741}$$

But we recognize here the primitive of:

$$f(x)=\frac{1}{(1+x^2)^2}\tag{12.4.742}$$

That we have already proved in the section of Differential and Integral Calculus (see page 1549) and that lead us to:

$$F(x)=\int\frac{1}{(1+x^2)^2}\mathrm{d}x=\frac{1}{2}\left(\frac{x}{1+x^2}+\arctan(x)\right)\tag{12.4.743}$$

Therefore:

$$-2\int\frac{1}{(u^2+1)^2}\mathrm{d}u=-\arctan(u)-\frac{u}{u^2+1}\tag{12.4.744}$$

Undo substitution:

$$u=\sqrt{\frac{1}{x}-1}\tag{12.4.745}$$

We get:

$$-\arctan(u)-\frac{u}{u^2+1}=-\sqrt{\frac{1}{x}-1}x-\arctan\left(\sqrt{\frac{1}{x}-1}\right)\tag{12.4.746}$$

The problem is then solved:

$$\int\frac{1}{\sqrt{\frac{1}{x}-1}}\mathrm{d}x=-\sqrt{\frac{1}{x}-1}x-\arctan\left(\sqrt{\frac{1}{x}-1}\right)+c^{te}\tag{12.4.747}$$

Therefore (ignoring the constant):

$$\begin{aligned}t_s=\frac{1}{c}\int\limits_0^{R_S}\frac{1}{\sqrt{\frac{2GM}{c^2r}-1}}\mathrm{d}r&=\frac{1}{c}\left[-\sqrt{\frac{2GM}{c^2r}-1}\frac{c^2r}{2GM}-\frac{2GM}{c^2}\arctan\left(\sqrt{\frac{2GM}{c^2r}-1}\right)\right]\Bigg|_0^{R_s}\\&=\frac{1}{c}\left(-\underbrace{\sqrt{\frac{2GM}{c^2\cdot\frac{2GM}{c^2}}-1}\frac{c^2\cdot\frac{2GM}{c^2}}{2GM}}_{=0}-\underbrace{\frac{2GM}{c^2}\arctan\left(\sqrt{\frac{2GM}{c^2\cdot\frac{2GM}{c^2}}-1}\right)}_{=0}\right)\\&-\frac{1}{c}\left(-\sqrt{\frac{2GM}{c^2\cdot 0}-1}\frac{c^2\cdot 0}{2GM}-\frac{2GM}{c^2}\arctan\left(\sqrt{\frac{2GM}{c^2\cdot 0}-1}\right)\right)\end{aligned}\tag{12.4.748}$$

But for the second parentheses we see we should better take the limit. Then we have:

$$\begin{gathered}\lim_{r\longrightarrow 0^+}\sqrt{\frac{2GM}{c^2r}-1}\frac{c^2r}{2GM}=0\\\lim_{r\longrightarrow 0^+}\arctan\left(\sqrt{\frac{2GM}{c^2r}-1}\right)=\frac{\pi}{2}\end{gathered}\tag{12.4.749}$$

Therefore:

$$t_s=-\frac{1}{c}\left(0-\frac{2GM}{c^2}\frac{\pi}{2}\right)=\frac{\pi GM}{c^3}\tag{12.4.750}$$

Before the 11970s (holocene calendar), most physicists believed that the interior of a Schwarzschild black hole curved inwards towards a sharp point at the singularity. However, in the late 11960s (holocene calendar), Soviet physicists Vladimir Belinskii, Isaak Khalatnikov, and Evgeny Lifshitz discovered that this model was only true when the space-time inside the black hole had not been perturbed. Any perturbations, such as those caused by matter or radiation falling in, would cause space to oscillate chaotically near the singularity according to a metric named the "Mixmaster dynamics". Any matter falling in would experience intense tidal forces rapidly changing in direction, all while being compressed into an increasingly small volume.

Remark

Black Hole research has indirectly contributed to modern Wi-Fi technology through advancements in signal processing. In the 11990s (holocene calendar), physicist John O'Sullivan was working on detecting radio signals from mini Black Holes predicted by Stephen Hawking. To filter out noise and interference, he developed a method using fast Fourier transforms to analyze and recombine signals. This technique later proved useful in solving issues with radio wave communication between computers, such as interference from reflected signals. O'Sullivan's work laid the foundation for the technology used in Wi-Fi systems today.

12.4.6.6.1.4 Black Hole Flamm's paraboloid representation

The spatial curvature of the Schwarzschild solution for $r > r_\text{s}$ can be visualized in a famous pseudo-3D representation named "Flamm's paraboloid".

For this purpose, let us consider a constant time equatorial slice through the Schwarzschild solution given for reminder by:

$$\mathrm{d}s^2 = \left(1 - \frac{2GM}{c^2 r}\right) c^2 \mathrm{d}t^2 - \frac{1}{\left(1 - \frac{2GM}{c^2 r}\right)} \mathrm{d}r^2 - r^2 \left(\mathrm{d}\theta^2 + \sin^2(\theta)\mathrm{d}\phi^2\right) \tag{12.4.751}$$

by fixing $\theta = \pi/2, t = c^{te}$, and letting the remaining Schwarzschild coordinates (r, ϕ) vary the latter reduces to:

$$\mathrm{d}s^2 = -\frac{1}{\left(1 - \frac{2GM}{c^2 r}\right)} \mathrm{d}r^2 - r^2 \mathrm{d}\phi^2 \tag{12.4.752}$$

Rearranging:

$$-\mathrm{d}s^2 = \frac{1}{\left(1 - \frac{2GM}{c^2 r}\right)} \mathrm{d}r^2 + r^2 \mathrm{d}\phi^2 \tag{12.4.753}$$

This done, Flamm's paraboloid may be derived as follows:

The Euclidean metric in the cylindrical coordinates (r, ϕ, w) is written for reminder by:

$$-\mathrm{d}s^2 = \mathrm{d}w^2 + \mathrm{d}r^2 + r^2 \mathrm{d}\phi^2 \tag{12.4.754}$$

Letting the surface be described by the function $w = w(r)$, the Euclidean metric can be written as (see page 5729):

$$-\mathrm{d}s^2 = \left(1 + \left(\frac{\mathrm{d}w}{\mathrm{d}r}\right)^2\right) \mathrm{d}r^2 + r^2 \mathrm{d}\phi^2 \tag{12.4.755}$$

Comparing this with the Schwarzschild metric in the equatorial plane $(\theta = \pi/2)$ at a fixed time $(t = c^{te}, \mathrm{d}t = 0)$:

$$-\mathrm{d}s^2 = \left(1 - \frac{r_\text{s}}{r}\right)^{-1} \mathrm{d}r^2 + r^2 \mathrm{d}\phi^2 \tag{12.4.756}$$

yields an integral expression for $w(r)$:

$$w(r) = \int \frac{\mathrm{d}r}{\sqrt{\frac{r}{r_\text{s}} - 1}} = 2r_\text{s}\sqrt{\frac{r}{r_\text{s}} - 1} + c^{te} = 2\sqrt{r_s(r - r_S)} + c^{te} \tag{12.4.757}$$

whose solution is Flamm's paraboloid.

We can plot this solution typically with pstplots in LaTeX. This gives:

$$\begin{cases} x = r\cos(u) \\ y = r\sin(u) \\ z = 2\sqrt{r_S(r - r_S)} \end{cases}$$

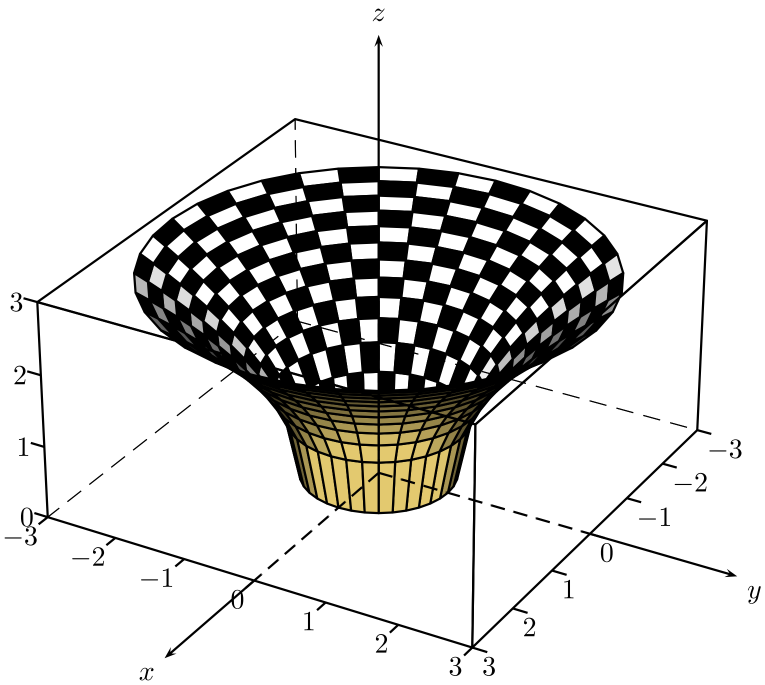

z
3
2
1
0
−3
−2
−1
0
1
2
3
x
−3
−2
−1
0
1
2
3
y

Figure 12.243: **Schwarzschild black hole representation based on Flamm's paraboloid**

12.4.6.6.1.5 White Hole

In General Relativity, a "White hole" is a hypothetical region of space-time and singularity that cannot be entered from the outside, although energy, matter, light and information can escape from it. In this sense, it is the opposite of a black hole, from which energy, matter, light and information cannot escape.

The basic reason why White Holes show up in General Relativity theory is because Einstein's equations are time-reversible, but when something falls into a Black Hole, it never re-emerges. We solve this paradox by realizing that the time-reversed path of something falling into a Black Hole in the future is that object emerging from a White Hole in the past.

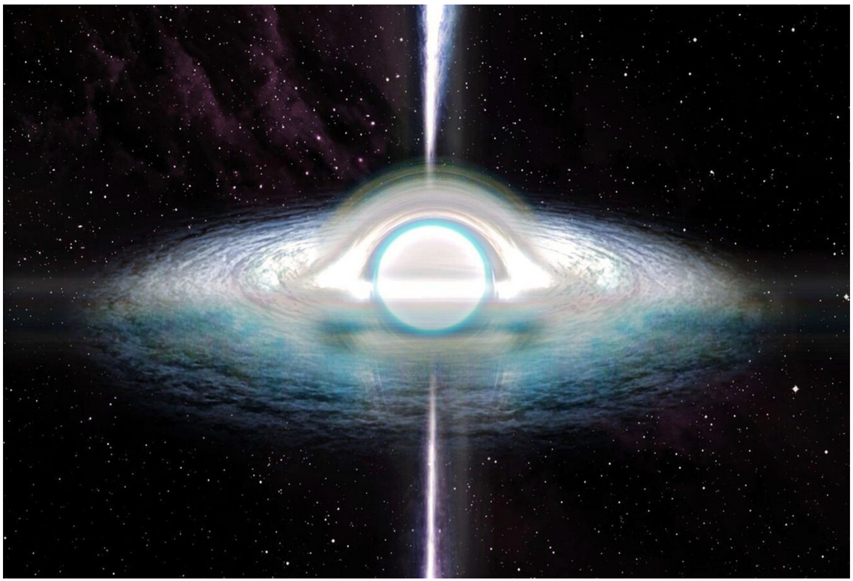


Figure 12.244: Artist's impression of a white hole (source: Wikipedia, author: Baperookamo)

A White Hole's color is a subject of speculation, but it would likely appear as a fiercely bright, luminous sphere of light because it expels matter and light, unlike a Black Hole. However, if the source of the material it expels does not produce light, it could also appear black. The name "White Hole" is just a mathematical counterpart to a Black Hole, not a literal color descriptor.

12.4.6.7 Gravitational waves

Before we focus on the maths, let us do a simple introduction.

"Gravitational waves" are ripples in the curvature of space-time that propagate as waves at the speed of light, generated in certain gravitational interactions that propagate outward from their source. The possibility of gravitational waves was discussed in 11893 (holocene calendar) by Oliver Heaviside using the analogy between the inverse-square law in gravitation and electricity. In 11905 (holocene calendar) Henri Poincaré first proposed gravitational waves emanating from a body and propagating at the speed of light as being required by the Lorentz transformations. Predicted in 11916 (holocene calendar) by Albert Einstein on the basis of his theory of General Relativity, gravitational waves transport energy as gravitational radiation, a form of radiant energy similar to electromagnetic radiation. Gravitational waves cannot exist in the Newton's law of universal gravitation, since that law is predicated on the assumption that physical interactions propagate at infinite speed.

Gravitational-wave astronomy is an emerging branch of observational astronomy which aims to use gravitational waves to collect observational data about sources of detectable gravitational waves such as binary star systems composed of White Dwarfs, neutron stars, and Black Holes; and events such as supernovae, and the formation of the early universe shortly after the Big Bang (i.e. starting point of the expansion of our universe).

On February 11, 12016 (holocene calendar), the LIGO Scientific Collaboration and Virgo Collaboration teams announced that they had made the first observation of gravitational waves, originating from a pair of merging Black Holes using the Advanced LIGO detectors. On June 15, 2016 (holocene calendar), a second detection of gravitational waves from coalescing Black Holes was announced. Besides LIGO, many other gravitational-wave observatories (detectors) are under construction.

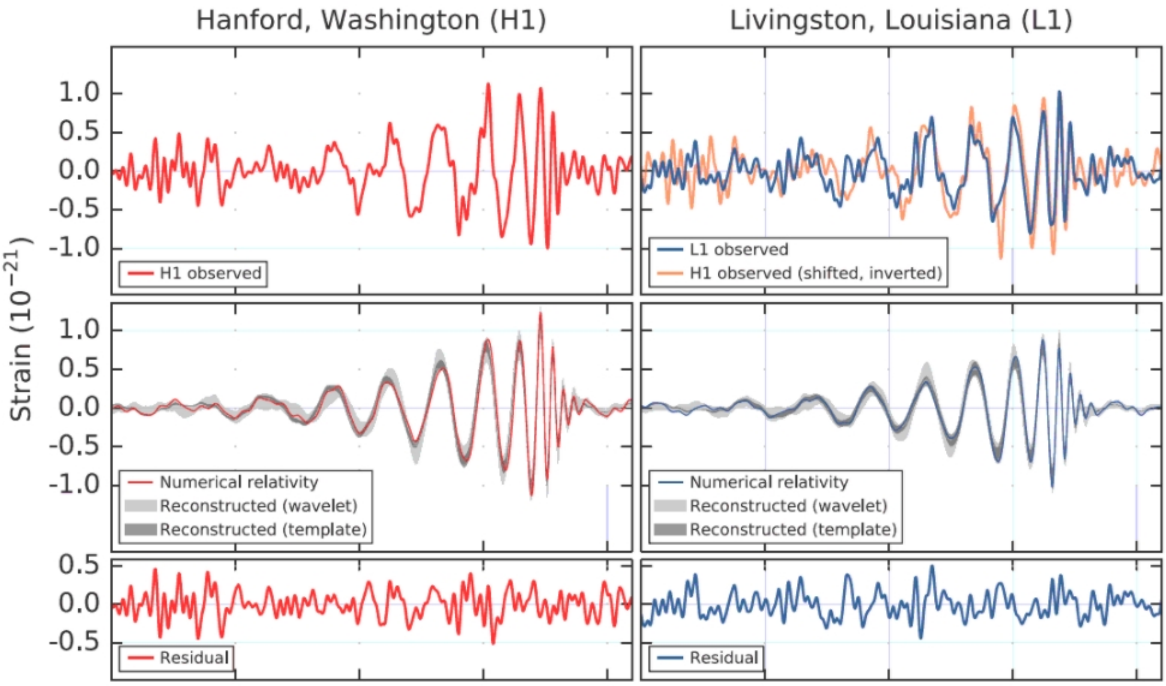
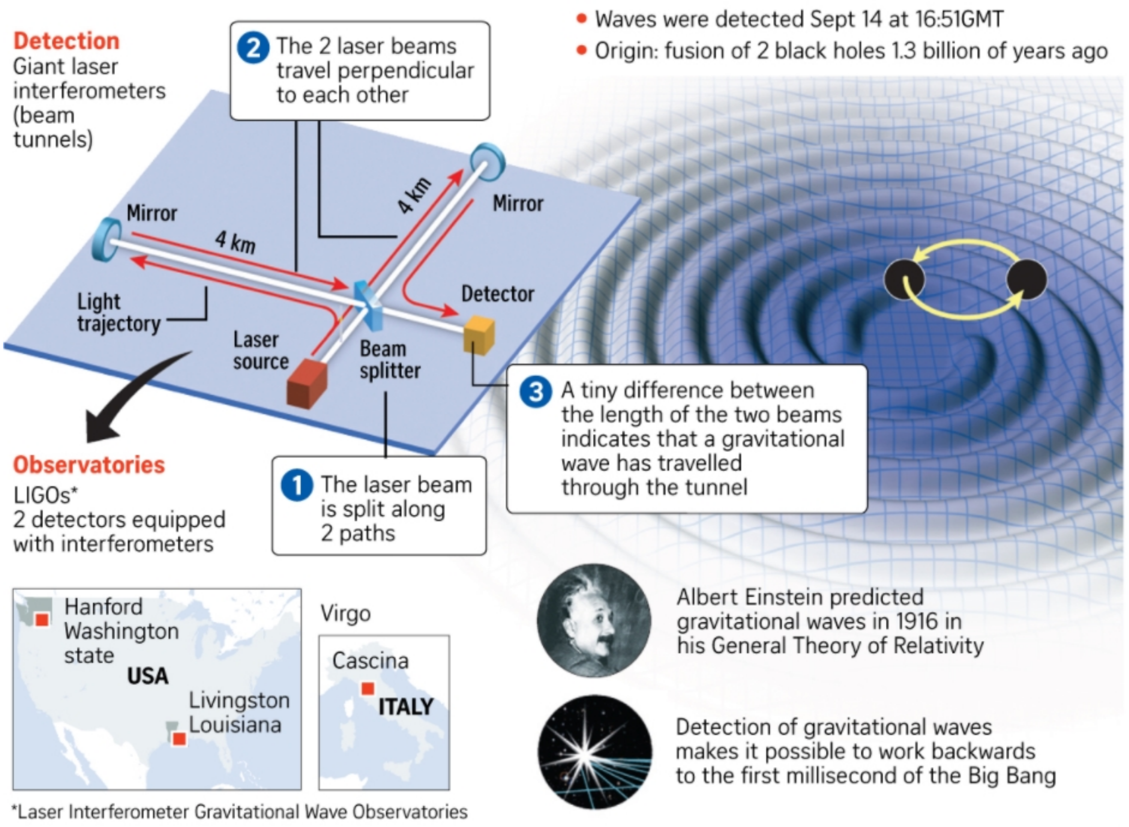


Figure 12.245: LIGO measurement of the gravitational waves at the Hanford (left) and Livingston (right) detectors


Detection
Giant laser interferometers (beam tunnels)
2 The 2 laser beams travel perpendicular to each other
Waves were detected Sept 14 at 16:51GMT
Origin: fusion of 2 black holes 1.3 billion of years ago
Mirror
4 km
4 km
Mirror
Light trajectory
Laser source
Beam splitter
Detector
3 A tiny difference between the length of the two beams indicates that a gravitational wave has travelled through the tunnel
Observatories
LIGOs*
2 detectors equipped with interferometers
1 The laser beam is split along 2 paths
Hanford
Washington state
USA
Livingston
Louisiana
Virgo
Cascina
ITALY
Albert Einstein predicted gravitational waves in 1916 in his General Theory of Relativity
Detection of gravitational waves makes it possible to work backwards to the first millisecond of the Big Bang
*Laser Interferometer Gravitational Wave Observatories

GRAPHICS ADAPTED FROM AFP

Remark

The reader must remember that Newtonian gravity is not local, this bothered Newton a lot back then! You see if you take Newton's law of gravity, with the gravitational force proportional to the mass divided by the square of the distance, then if you move the mass, the distance increases and the force changes instantaneously, everywhere. Newton was greatly bothered by this. He named it "action at a distance" and that's also what Einstein meant by the phrase. Einstein's theory of General Relativity, which we use now to describe gravity, doesn't have this problem. In Einstein's theory, gravity is described by the curvature of space and time. And that curvature is caused by the mass and energy in space-time. If you move the mass that causes space-time to curve, then the curvature changes, but not instantaneously. The change spreads outwards at the speed of light as we will prove it below. So, Einstein's theory is "local" as physicists say.

In real life it looks like this (in the legend of the photo below, the year is given according to the holocene calendar):

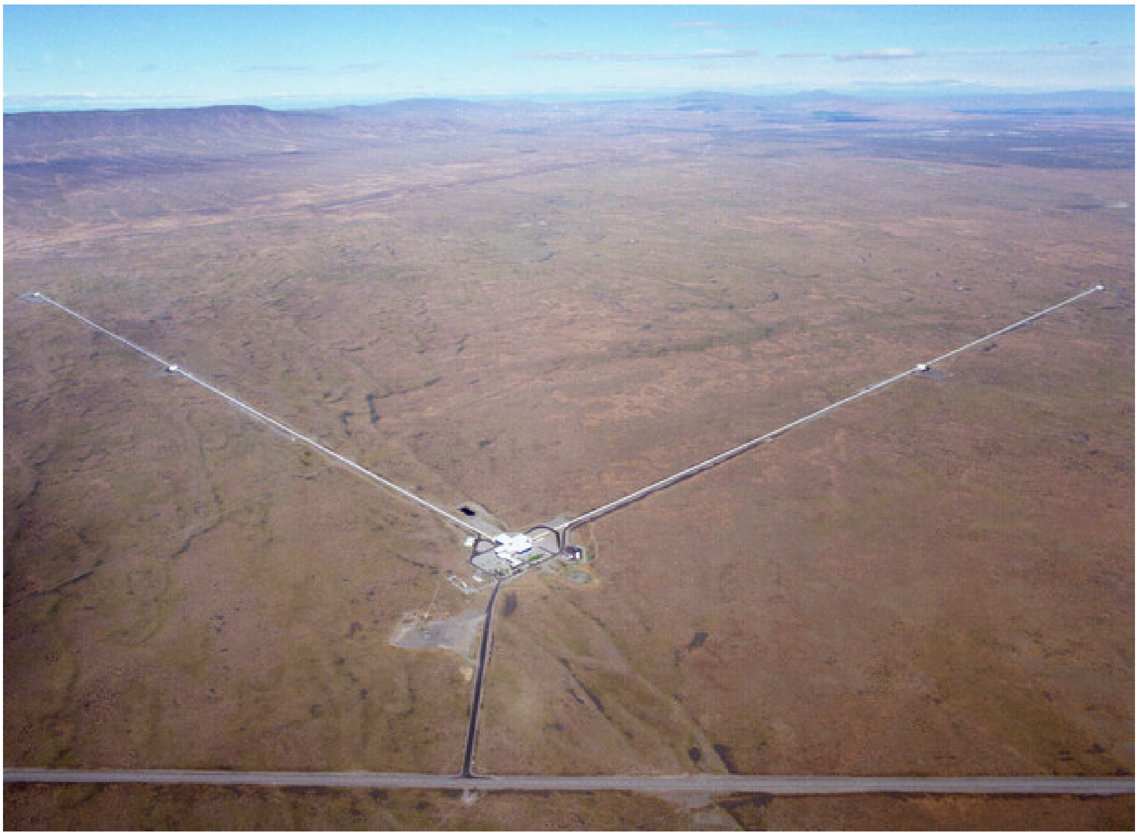


Figure 12.246: **Hanford LIGO detector - October 30, 12000 - Washington state (source: LIGO)**

The most typical illustration that we can see of gravitational waves in newspapers are those generated by the special case of a high speed dynamic gravitation field due to the rotation of two massive object around each other. Otherwise, and it is obvious, a collapsing star at the end of its life also generated a variational gravitational field but that is quite difficult to detect with actual existing instruments as a star alone is not massive enough to generate detectable gravitation waves:

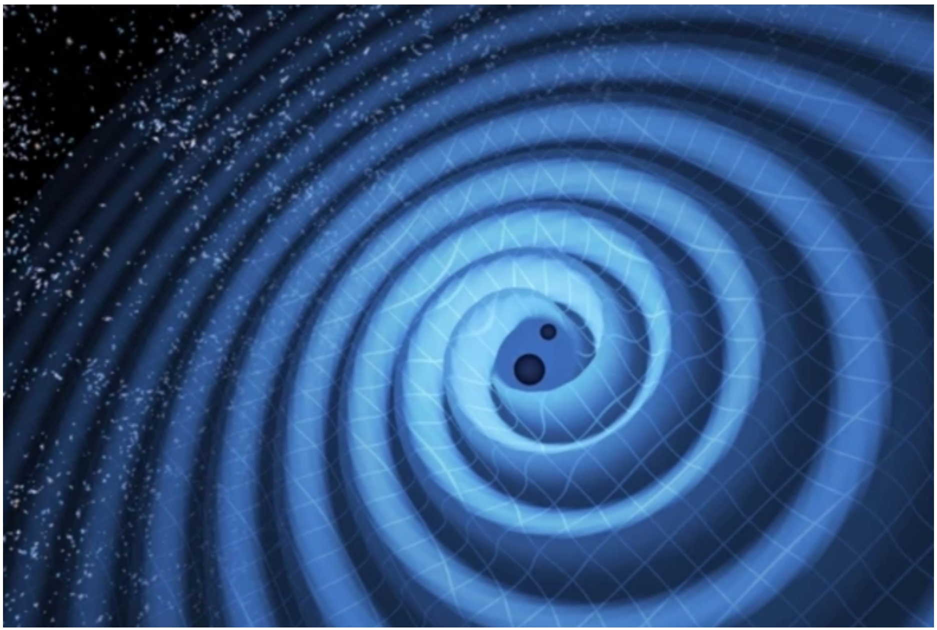


Figure 12.247: **Pseudo-3D (wrong) common visualization of gravitational wave of a binary system**

But obviously gravitational waves don't make things go up and down like ocean waves as illustrated above, and they're definitely not like that planet on a trampoline — after all, there's nothing "below" to pull things downward so there can't be a dent. And gravitational waves don't do spirals, much...

In a gravitational wave, space itself is compressed and stretched. A particle caught in a gravitational wave doesn't get pushed back and forth. Instead, it shrinks and expands in place. If you encounter a gravitational wave, you and all your calibrated measurement gear (yardsticks, digital rangers, that slide rule you're so proud of) shrink and expand together. You would only notice the experience if you happened to be comparing two extremely precise LASER rangers set perpendicular to each other (LIGO!). One would briefly register a slight change compared to the other one.

Now let us deal with the maths. As almost always in science there are multiple ways to introduce a new tool. In this textbook, as most of times, we will use what we consider the most easy one and that is "physicist" or "engineer" oriented...

So first, remember that earlier we have proved under some assumptions that:

$$\nabla^2 h_{00}=\Delta h_{00}=-\frac{8\pi G}{c^4}T_{00} \tag{12.4.758}$$

That what we have seen afterwards can be written more generally as:

$$\nabla^2 h_{\mu\nu}=\Delta h_{\mu\nu}=-\frac{8\pi G}{c^4}T_{\mu\nu} \tag{12.4.759}$$

If we explicit it by keeping in mind Classical Mechanics it would be written in 3D:

$$\frac{\partial^2 h_{\mu\nu}}{\partial x^2}+\frac{\partial^2 h_{\mu\nu}}{\partial y^2}+\frac{\partial^2 h_{\mu\nu}}{\partial z^2}=-\frac{8\pi G}{c^4}T_{\mu\nu} \tag{12.4.760}$$

or a bit better:

$$\frac{\partial^2 h_{\mu\nu}}{\partial {x^1}^2}+\frac{\partial^2 h_{\mu\nu}}{\partial {x^2}^2}+\frac{\partial^2 h_{\mu\nu}}{\partial {x^3}^2}=-\frac{8\pi G}{c^4}T_{\mu\nu} \tag{12.4.761}$$

But... but...! We are in General Relativity and we have to introduce the 4-th dimensions:

$$\frac{\partial^2 h_{\mu\nu}}{\partial {x^0}^2}+\frac{\partial^2 h_{\mu\nu}}{\partial {x^1}^2}+\frac{\partial^2}{\partial {x^2}^2}+\frac{\partial^2 h_{\mu\nu}}{\partial {x^3}^2}=-\frac{8\pi G}{c^4}T_{\mu\nu} \tag{12.4.762}$$

This is better but now let us explicit this using in cartesian coordinates using the metric $(-,+,+,+)$. We then get:

$$\frac{\partial^2 h_{\mu\nu}}{\partial(-ct)^2}+\frac{\partial^2}{\partial x^2}+\frac{\partial^h_{\mu\nu}2}{\partial y^2}+\frac{\partial^2 h_{\mu\nu}}{\partial z^2}=-\frac{8\pi G}{c^4}T_{\mu\nu} \tag{12.4.763}$$

Assuming that we a observing a piece (volume) of space where there is no matter and no radiation in any form, then $T_{\mu\nu}=0$ and we get:

$$-\frac{1}{c^2}\frac{\partial^2 h_{\mu\nu}}{\partial t^2}+\frac{\partial^2 h_{\mu\nu}}{\partial x^2}+\frac{\partial^2 h_{\mu\nu}}{\partial y^2}+\frac{\partial^2 h_{\mu\nu}}{\partial z^2}=0 \tag{12.4.764}$$

Using the d'Alembertian already introduced in the section of Electrodynamics but now in its General Relativity form named "flat-space d'Alembertian":

$$\Box=\partial^{\mu}\partial_{\mu}=g^{\mu\nu}\partial_{\nu}\partial_{\mu}=\eta^{\mu\nu}\partial_{\nu}\partial_{\mu}=-\frac{1}{c^2}\partial_t^2+\Delta \tag{12.4.765}$$

that latter relation is commonly written in textbooks:

$$\boxed{\Box h_{\mu\nu}=0} \tag{12.4.766}$$

Taking back and rearranging the explicit relation we get (we see that this equation also gives us that space transmits gravitational waves at the speed of light as confirmed experimentally as reported in [3]!):

$$\boxed{\frac{\partial^2 h_{\mu\nu}}{\partial x^2}+\frac{\partial^2 h_{\mu\nu}}{\partial y^2}+\frac{\partial^2 h_{\mu\nu}}{\partial z^2}=\frac{1}{c^2}\frac{\partial^2 h_{\mu\nu}}{\partial t^2}} \tag{12.4.767}$$

and named the "gravitational wave equation", "gravitational propagation equation" or "gravitational d'Alembert's equation".

We know that a simple solution to:

$$\left(\nabla^2-\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)h_{\mu\nu}=0 \tag{12.4.768}$$

can typically be:

$$\bar{h}_{\mu\nu}=\Re\left(\hat{A}_{\mu\nu}e^{\mathrm{i}k_{\rho}x^{\rho}}\right) \tag{12.4.769}$$

The amplitude $\hat{A}_{\mu\nu}$ and the wave vector k_{ρ} must satisfy the differential equation, therefore:

$$k_{\rho}=0 \tag{12.4.770}$$

Gravitational waves pass through boundaries that light cannot! They can transport information about what happens inside the event horizons of Black Holes, and they can pass through the cosmic background (CMB) radiation (see section Cosmogony page 5966), the barrier of light that prevent us from seeing our Universe before it turned $380,000$ years old!

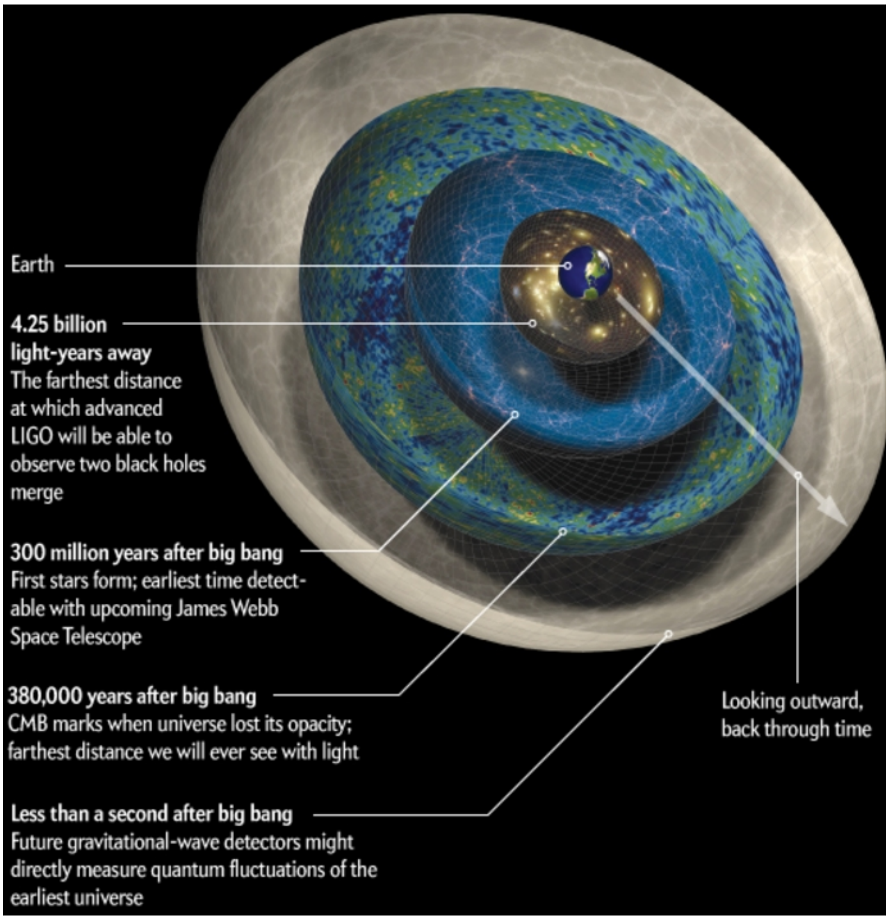

Earth
4.25 billion light-years away
The farthest distance at which advanced LIGO will be able to observe two black holes merge
300 million years after big bang
First stars form; earliest time detectable with upcoming James Webb Space Telescope
380,000 years after big bang
CMB marks when universe lost its opacity; farthest distance we will ever see with light
Less than a second after big bang
Future gravitational-wave detectors might directly measure quantum fluctuations of the earliest universe
Looking outward, back through time

Figure 12.248: **Gravitational cosmic background "radiation"**

Remark

A common question about this topic is: *How can gravity escape out of a Black Hole when it is known that even light cannot escape a Black Hole?*. The answer is quite simple! Gravity does not escape the Black hole, it covers the Black Hole event horizon like a blanket (all the information of remaining stuff collapsing as a Black Hole is on the surface of Event Horizon for theoretically a huge amount of time), pulling stuff towards itself and the singularity.

So for summary General Relativity has successfully with high accuracy pass the following experimental tests in order of verification (top oldest):

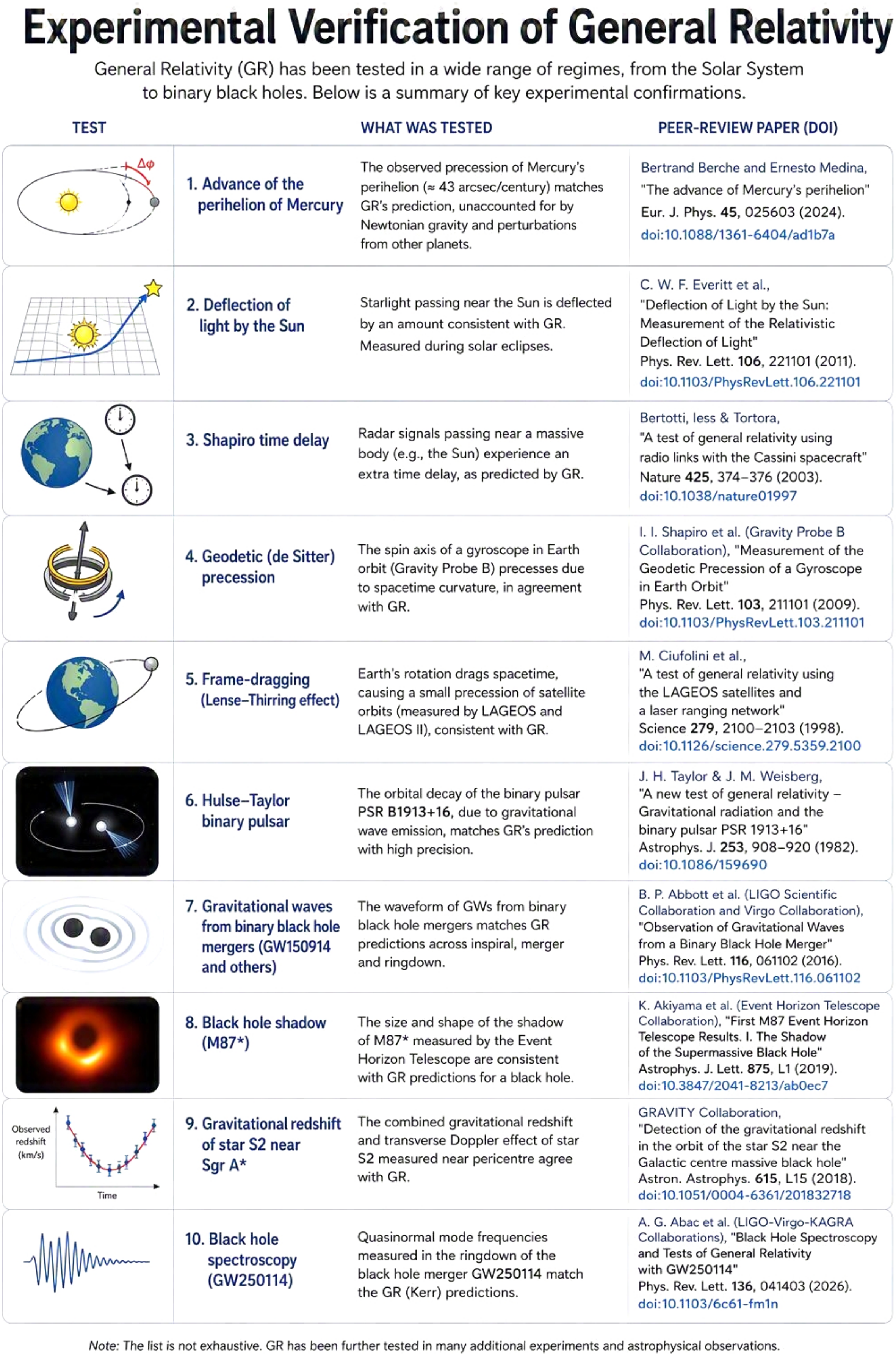
1. Mercury perihelion precession
2. Light deviation
3. Black Holes
4. Universe expansion
5. Time Dilatation
6. Gravitational waves

Let us point out once again a very important point. Before Albert Einstein, geometry was considered an integral part of the laws. Albert Einstein has shown that the geometry of space evolves in time according to other, even deeper, laws. It is important to understand this point. The geometry of space is not part of the laws of nature (this was criticized at his time by many French physicists). Therefore, nothing that we can find in these laws tells us what geometry of space we are working in. Thus, before we begin to solve the equations of Einstein's General Relativity, we have absolutely no idea what geometry we ear dealing with. We only discover it once the equations are solved!

In extenso, the choice of 4 dimensions is part of the background. Could it be possible that another deeper theory does not require presupposing the number of dimensions?

To sum up, the idea of independence relatively to the background, in its most general formulation, is a wise way of making physics: made up of better theories, in which the things which before were postulated, will be explained in allowing such things to evolve over time according to new laws.

This is also a difficulty of Quantum Theory, that latter is essentially background dependent at the opposite of General Relativity that is "background independent".


Δφ
Observed redshift (km/s)
Time

Experimental Verification of General Relativity

General Relativity (GR) has been tested in a wide range of regimes, from the Solar System to binary black holes. Below is a summary of key experimental confirmations.

TEST		WHAT WAS TESTED	PEER-REVIEW PAPER (DOI)
	1. Advance of the perihelion of Mercury	The observed precession of Mercury's perihelion (≈ 43 arcsec/century) matches GR's prediction, unaccounted for by Newtonian gravity and perturbations from other planets.	Bertrand Berche and Ernesto Medina, "The advance of Mercury's perihelion" Eur. J. Phys. **45**, 025603 (2024). doi:10.1088/1361-6404/ad1b7a
	2. Deflection of light by the Sun	Starlight passing near the Sun is deflected by an amount consistent with GR. Measured during solar eclipses.	C. W. F. Everitt et al., "Deflection of Light by the Sun: Measurement of the Relativistic Deflection of Light" Phys. Rev. Lett. **106**, 221101 (2011). doi:10.1103/PhysRevLett.106.221101
	3. Shapiro time delay	Radar signals passing near a massive body (e.g., the Sun) experience an extra time delay, as predicted by GR.	Bertotti, Iess & Tortora, "A test of general relativity using radio links with the Cassini spacecraft" Nature **425**, 374–376 (2003). doi:10.1038/nature01997
	4. Geodetic (de Sitter) precession	The spin axis of a gyroscope in Earth orbit (Gravity Probe B) precesses due to spacetime curvature, in agreement with GR.	I. I. Shapiro et al. (Gravity Probe B Collaboration), "Measurement of the Geodetic Precession of a Gyroscope in Earth Orbit" Phys. Rev. Lett. **103**, 211101 (2009). doi:10.1103/PhysRevLett.103.211101
	5. Frame-dragging (Lense–Thirring effect)	Earth's rotation drags spacetime, causing a small precession of satellite orbits (measured by LAGEOS and LAGEOS II), consistent with GR.	M. Ciufolini et al., "A test of general relativity using the LAGEOS satellites and a laser ranging network" Science **279**, 2100–2103 (1998). doi:10.1126/science.279.5359.2100
	6. Hulse–Taylor binary pulsar	The orbital decay of the binary pulsar PSR **B1913+16**, due to gravitational wave emission, matches GR's prediction with high precision.	J. H. Taylor & J. M. Weisberg, "A new test of general relativity – Gravitational radiation and the binary pulsar PSR 1913+16" Astrophys. J. **253**, 908–920 (1982). doi:10.1086/159690
	7. Gravitational waves from binary black hole mergers (GW150914 and others)	The waveform of GWs from binary black hole mergers matches GR predictions across inspiral, merger and ringdown.	B. P. Abbott et al. (LIGO Scientific Collaboration and Virgo Collaboration), "Observation of Gravitational Waves from a Binary Black Hole Merger" Phys. Rev. Lett. **116**, 061102 (2016). doi:10.1103/PhysRevLett.116.061102
	8. Black hole shadow (M87*)	The size and shape of the shadow of M87* measured by the Event Horizon Telescope are consistent with GR predictions for a black hole.	K. Akiyama et al. (Event Horizon Telescope Collaboration), "First M87 Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole" Astrophys. J. Lett. **875**, L1 (2019). doi:10.3847/2041-8213/ab0ec7
	9. Gravitational redshift of star S2 near Sgr A*	The combined gravitational redshift and transverse Doppler effect of star S2 measured near pericentre agree with GR.	GRAVITY Collaboration, "Detection of the gravitational redshift in the orbit of the star S2 near the Galactic centre massive black hole" Astron. Astrophys. **615**, L15 (2018). doi:10.1051/0004-6361/201832718
	10. Black hole spectroscopy (GW250114)	Quasinormal mode frequencies measured in the ringdown of the black hole merger GW250114 match the GR (Kerr) predictions.	A. G. Abac et al. (LIGO-Virgo-KAGRA Collaborations), "Black Hole Spectroscopy and Tests of General Relativity with GW250114" Phys. Rev. Lett. **136**, 041403 (2026). doi:10.1103/6c61-fm1n

Note: The list is not exhaustive. GR has been further tested in many additional experiments and astrophysical observations.

Figure 12.249: **Summary of the most well-known experimental tests of General Relativity**

12.4.7 Einstein-Hilbert action

The "Einstein-Hilbert action" (also referred to as "Hilbert action") in General Relativity is the action that yields the Einstein field equations through the principle of least action (see section Analytical Mechanics page 3131). With the $(-+++)$ metric signature, the gravitational part of the action is given as:

$$S=-\frac{1}{c\kappa}\int R\sqrt{|g|}\mathrm{d}^4x=-\frac{c^3}{8\pi G}\int R\sqrt{|g|}\mathrm{d}^4x \tag{12.4.771}$$

The derivation of equations from an action has several advantages. First of all, it allows for easy unification of General Relativity with other classical field theories (such as Maxwell theory), which are also formulated in terms of an action (this is especially why we were obliged to introduced this action in this textbook!).

There are multiple ways to make the proof of that action, but we will use the one that is the easiest in our point of view and that is more "engineer way of life" oriented rather than "theoretical physicist way of life"... So let's go!

Proof 12.0.1. We know that the classical action in mechanics is given for recall by:

$$S=\int\limits_{t_1}^{t_2} L\mathrm{d}t \tag{12.4.772}$$

In this expression, the Lagrangian L is homogeneous to an energy. In gravitation as in electromagnetism, the energy-stress tensor T is homogeneous to a density energy $\mathcal{L}$, which obliges us to introduce this density in the action by integrating on the volume as following:

$$S=\int\limits_{t_1}^{t_2}\left[\int\limits_V \mathcal{L}\mathrm{d}V\right]\mathrm{d}t=\int \mathcal{L}\mathrm{d}\Omega \tag{12.4.773}$$

The search of an action invariant by change of reference frame imposes us an element of volume $\mathrm{d}\Omega$ quadridimensional that is also invariant by change of reference frame. And we have proved in the section of Tensor Calculus that such an element was given by the Riemmanian volume (see section Tensor Calculus page 2058):

$$\mathrm{d}\Omega=\sqrt{|g|}\mathrm{d}x^4 \tag{12.4.774}$$

Then we deduce the new form of the action:

$$S=\int \mathcal{L}\sqrt{|g|}\mathrm{d}x^4 \tag{12.4.775}$$

But as we look the units, as $x^0=ct$ the units are not homogeneous to an energy anymore. Then we must divide by c:

$$S=\frac{1}{c}\int \mathcal{L}\sqrt{|g|}\mathrm{d}x^4 \tag{12.4.776}$$

Now we need to find a scalar for R that is invariant has for units a density of energy. For this, let us recall that during our proof of the Einstein's field equations earlier above (page 5766), we get the relation:

$$R-2R=\kappa g^{\rho\sigma}T_{\rho\sigma} \tag{12.4.777}$$

That obviously can be simplified as:

$$R=-\kappa g^{\rho\sigma}T_{\rho\sigma} \tag{12.4.778}$$

and as $g^{\rho\sigma}T_{\rho\sigma}$ is often denoted simply T, we get:

$$R=-\kappa T=-\frac{8\pi G}{c^4}T \tag{12.4.779}$$

And as the left term is a well known scalar implicitly involving g^{ij} (as for recall $R=g^{ij}R_{ij}$), what may be quite useful intuitively, it is a good candidate! However...:

- It doesn't have the units of a density energy
- It seems to be negative as T is positive (at least as far as we know...)

Then if we write:

$$-\frac{R}{\kappa}=T \tag{12.4.780}$$

We get a very good candidate four our Lagrangian density. Then we will try with:

$$\boxed{S=-\frac{1}{\kappa c}\int R\sqrt{|g|}\mathrm{d}x^4=-\frac{1}{\kappa c}\int(g^{ij}R_{ij})\sqrt{|g|}\mathrm{d}x^4} \tag{12.4.781}$$

■ Q.E.D.

Ok the derivation is quite based on trial and errors reasoning. So it may be a good practice to check if this action lead us back to the Einstein Field questions (or something similar) by applying the variational principle!

Assuming that all conditions are satisfied to the variation to pass through the integral, we have:

$$\delta S=-\frac{1}{\kappa c}\int\delta\left(g^{ij}R_{ij}\sqrt{|g|}\right)\mathrm{d}x^4 \tag{12.4.782}$$

As the determinant $|g|$ of the g_{ij} is negative, at least in the most common cases (Minkowski, Schwarzschild, Robertson-Walker, Kerr, etc.), we write:

$$\delta\left(g^{ij}R_{ij}\sqrt{|g|}\right)=\delta\left(g^{ij}R_{ij}\sqrt{-g}\right) \tag{12.4.783}$$

This expression is differentiate as the product of three term:

$$\delta\left(g^{ij}R_{ij}\sqrt{-g}\right)=\delta(g^{ij})R_{ij}\sqrt{-g}+\delta(R_{ij})g^{ij}\sqrt{-g}+\delta(\sqrt{-g})g^{ij}R_{ij} \tag{12.4.784}$$

That can be written:

$$\delta\left(g^{ij}R_{ij}\sqrt{-g}\right)=\delta(g^{ij})R_{ij}\sqrt{-g}+\delta(R_{ij})g^{ij}\sqrt{-g}+\delta(\sqrt{-g})R \tag{12.4.785}$$

The non-obvious idea is to make appear δg^{ij} as a factor in the integral of the variation δS and we will request as always that $\delta S=0$ and conclude/check if the result we get is interesting or not!?

In the differential above the first term $\delta(g^{ij})R_{ij}\sqrt{-g}$ naturally makes appear $\delta(g^{ij})$, and therefore doesn't need any treatment.

For the other two terms instead, we will have to express $\delta(R_{ij})$ and $\delta(\sqrt{-g})$ in function of the $\delta(g_{ij})$, or show that their contribution is equal to zero. It is not an easy work, it needed a great mathematician like David Hilbert to achieve this work. So here is the procedure that as you will see... is quit a pain in the a**...

- First we focus on the term $\sqrt{-g}$. For this let us consider a matrix with the following form:

  $$A=\begin{bmatrix}a_0 & 0\\ 0 & a_1\end{bmatrix} \tag{12.4.786}$$

  The trace of this matrix, as for any matrix, is the sum of the elements of the main diagonal, so:

  $$\text{tr}(A)=\text{tr}(A)=a_0+a_1 \tag{12.4.787}$$

  So that:

  $$e^{\text{tr}(A)}=e^{a_0+a_1} \tag{12.4.788}$$

  If we consider the exponential matrix e^A (see section Linear Algebra page 1628) as:

  $$e^A=\begin{bmatrix}e^{a_0} & 0\\ 0 & e^{a_1}\end{bmatrix} \tag{12.4.789}$$

  Then the determinant of this diagonal matrix is (see section Linear Algebra page 1966):

  $$\det(e^A)=e^{a_0}e^{a_1}=e^{a_0+a_1} \tag{12.4.790}$$

  So that finally we can write:

  $$\det(e^A)=e^{\text{tr}(A)} \tag{12.4.791}$$

  If we put $B=e^A$, we have obviously:

  $$e^A=B\Rightarrow \ln(e^A)=\ln(B)\Rightarrow A=\ln(B) \tag{12.4.792}$$

  So as $\det(e^A)=e^{\text{tr}(A)}$, we have:

  $$\det(B)=e^{\text{tr}(\ln(B))} \tag{12.4.793}$$

  Therefore:

  $$\ln\left(\det(B)\right)=\ln\left(e^{\text{tr}(\ln(B))}\right) \tag{12.4.794}$$

Hence:

$$\ln\left(\det(B)\right)=\operatorname{tr}(\ln(B)) \tag{12.4.795}$$

Taking the differential of both side:

$$\mathrm{d}\ln\left(\det(B)\right)=\mathrm{d}(\operatorname{tr}(\ln(B)))\Rightarrow\frac{1}{\det(B)}\mathrm{d}(\det(B))=\operatorname{tr}\left(\frac{1}{B}\mathrm{d}B\right) \tag{12.4.796}$$

Or written in a more condensed way:

$$\frac{\mathrm{d}(\det(B))}{\det(B)}=\operatorname{tr}\left(\frac{\mathrm{d}B}{B}\right)=\operatorname{tr}(B^{-1}\mathrm{d}B) \tag{12.4.797}$$

If we now relate this last result to the metric g_{ij}, we set $B=g_{ij}$, $B^{-1}=g^{ij}$ and $\det(B)=g$ leading to:

$$\frac{\mathrm{d}g}{g}=g^{ij}\mathrm{d}g_{ij} \tag{12.4.798}$$

Or written in the physics way:

$$\frac{\delta g}{g}=g^{ij}\delta g_{ij} \tag{12.4.799}$$

and rearranged a bit:

$$\delta g=gg^{ij}\delta g_{ij} \tag{12.4.800}$$

Now let us take the differential of $\sqrt{-g}$ (or in other words, the "variational of the metric determinant"):

$$\delta\sqrt{-g}=-\frac{1}{2\sqrt{-g}}\delta g \tag{12.4.801}$$

and injecting the prior-previous relation:

$$\delta\sqrt{-g}=-\frac{1}{2\sqrt{-g}}\left(gg^{ij}\delta g_{ij}\right)=\frac{1}{2\sqrt{-g}}(-g)g^{ij}\delta g_{ij}=\frac{1}{2\sqrt{-g}}(\sqrt{-g})^2g^{ij}\delta g_{ij}=\frac{1}{2}\sqrt{-g}g^{ij}\delta g_{ij} \tag{12.4.802}$$

We can still write this relation in a slightly different style! We know that the metric and its inverse are related in the following way (see section Tensor Calculus page 2051):

$$g_{ij}g^{ij}=\delta_i^i \tag{12.4.803}$$

which leads to:

$$\delta(g_{ij}g^{ij})=\delta(\delta_i^i)=0 \tag{12.4.804}$$

Therefore:

$$\delta(g_{ij}g^{ij})=\delta g_{ij}g^{ij}+g_{ij}\delta g^{ij}=0 \tag{12.4.805}$$

Hence:

$$\delta g_{ij}g^{ij}=-g_{ij}\delta g^{ij} \tag{12.4.806}$$

Finally we can write:

$$\delta\sqrt{-g}=-\frac{1}{2}\sqrt{-g}g_{ij}\delta g^{ij} \tag{12.4.807}$$

And this is the first result we were looking for that we will keep for later to re-inject in the action integral variational δS.

- Now we focus on the term $\delta(R_{ij})$. For this, let us recall (see section Tensor Calculus page 2127):

$$R_{\mu\mu}=R^{\alpha}_{\mu\alpha\nu}=\partial_{\alpha}\Gamma^{\alpha}_{\mu\nu}-\partial_{\nu}\Gamma^{\alpha}_{\mu\alpha}+\Gamma^{\alpha}_{\sigma\gamma}\Gamma^{\gamma}_{\mu\nu}-\Gamma^{\alpha}_{\nu\gamma}\Gamma^{\gamma}_{\mu\sigma} \tag{12.4.808}$$

Then varying it, gives:

$$\delta R_{\mu\nu}=\partial_{\alpha}\delta\Gamma^{\alpha}_{\mu\nu}-\partial_{\nu}\delta\Gamma^{\alpha}_{\mu\alpha}+\delta\Gamma^{\alpha}_{\alpha\gamma}\Gamma^{\gamma}_{\mu\nu}+\Gamma^{\alpha}_{\alpha\gamma}\delta\Gamma^{\gamma}_{\mu\nu}-\delta\Gamma^{\alpha}_{\nu\gamma}\Gamma^{\gamma}_{\mu\alpha}-\Gamma^{\alpha}_{\nu\gamma}\delta\Gamma^{\gamma}_{\mu\alpha} \tag{12.4.809}$$

The first two terms of this expression suggest that it could be the difference between two covariant derivatives. Let us prove that it is the case.

We have by applying the definition of the second order covariant derivative[53]:

$$\begin{aligned}\nabla_{\alpha}(\delta\Gamma^{\alpha}_{\mu\nu})&=\partial_{\alpha}\delta\Gamma^{\alpha}_{\mu\nu}+\Gamma^{\alpha}_{\alpha\gamma}\delta\Gamma^{\gamma}_{\mu\nu}-\delta\Gamma^{\alpha}_{\nu\gamma}\Gamma^{\gamma}_{\mu\alpha}-\delta\Gamma^{\alpha}_{\mu\gamma}\Gamma^{\gamma}_{\nu\alpha}\nabla_{\alpha}(\delta\Gamma^{\alpha}_{\mu\alpha})\\&=\partial_{\nu}\delta\Gamma^{\alpha}_{\mu\alpha}+\delta\Gamma^{\gamma}_{\mu\alpha}\Gamma^{\alpha}_{\nu\gamma}-\delta\Gamma^{\alpha}_{\alpha\gamma}\Gamma^{\gamma}_{\mu\nu}-\delta\Gamma^{\alpha}_{\mu\gamma}\Gamma^{\gamma}_{\nu\alpha}\end{aligned} \tag{12.4.810}$$

We cans thus verify that:

$$\begin{aligned}\nabla_{\alpha}(\delta\Gamma^{\alpha}_{\mu\nu})-\nabla_{\alpha}(\delta\Gamma^{\alpha}_{\mu\alpha})&=\partial_{\alpha}\delta\Gamma^{\alpha}_{\mu\nu}+\Gamma^{\alpha}_{\alpha\gamma}\delta\Gamma^{\gamma}_{\mu\nu}-\delta\Gamma^{\alpha}_{\nu\gamma}\Gamma^{\gamma}_{\mu\alpha}-\cancel{\delta\Gamma^{\alpha}_{\mu\gamma}\Gamma^{\gamma}_{\nu\alpha}}\\&-\left(\partial_{\nu}\delta\Gamma^{\alpha}_{\mu\alpha}+\delta\Gamma^{\gamma}_{\mu\alpha}\Gamma^{\alpha}_{\nu\gamma}-\delta\Gamma^{\alpha}_{\alpha\gamma}\Gamma^{\gamma}_{\mu\nu}-\cancel{\delta\Gamma^{\alpha}_{\mu\gamma}\Gamma^{\gamma}_{\nu\alpha}}\right)\\&=\delta R_{\mu\nu}\end{aligned} \tag{12.4.811}$$

The relation:

$$\boxed{\nabla_{\alpha}(\delta\Gamma^{\alpha}_{\mu\nu})-\nabla_{\alpha}(\delta\Gamma^{\alpha}_{\mu\alpha})=\delta R_{\mu\nu}} \tag{12.4.812}$$

is named the "Palatini identity".

Now we can rewrite the action:

$$\begin{aligned}\delta S&=-\frac{1}{\kappa c}\int\left(\delta(g^{ij})R_{ij}\sqrt{-g}+\delta(R_{ij})g^{ij}\sqrt{-g}+\delta(\sqrt{-g})R\right)\mathrm{d}x^4\\&=-\frac{1}{\kappa c}\int\left[\delta g^{ij}R_{ij}\sqrt{-g}+\left(\nabla_{\alpha}(\delta\Gamma^{\alpha}_{\mu\nu})-\nabla_{\alpha}(\delta\Gamma^{\alpha}_{\mu\alpha})\right)g^{\mu\nu}\sqrt{-g}-\frac{1}{2}\sqrt{-g}g_{ij}\delta g^{ij}R\right]\mathrm{d}x^4\end{aligned} \tag{12.4.813}$$

---

[53] The Christoffel coefficients are not tensors, but their variations have the character of a tensor.

Let us focus on the term (we will use the property that the covariant derivative of the metric is equal to zero a proven in the section of Tensor Calculus):

$$\begin{aligned}\int g^{\mu\nu}\delta R_{\mu\nu}\sqrt{-g}\mathrm{d}^4x&=\int\left(\nabla_\alpha(\delta\Gamma^\alpha_{\mu\nu})-\nabla_\alpha(\delta\Gamma^\alpha_{\mu\alpha})\right)g^{\mu\nu}\sqrt{-g}\mathrm{d}^4x\\&=\int\left[\nabla_\alpha(g^{\mu\nu}\delta\Gamma^\alpha_{\mu\nu})-\nabla_\nu(g^{\mu\nu}\delta\Gamma^\alpha_{\mu\alpha})\right]\sqrt{-g}\mathrm{d}^4x\\&=\int\left[\nabla_\alpha(g^{\mu\nu}\delta\Gamma^\alpha_{\mu\nu})-\nabla_\alpha(g^{\mu\alpha}\delta\Gamma^\nu_{\mu\nu})\right]\sqrt{-g}\mathrm{d}^4x\\&=\int\nabla_\alpha\left(g^{\mu\nu}\delta\Gamma^\alpha_{\mu\nu}-g^{\mu\alpha}\delta\Gamma^\nu_{\mu\nu}\right)\sqrt{-g}\mathrm{d}^4x\end{aligned}\quad(12.4.814)$$

When considering the expression in brackets, we notice that the μ and ν indices cancel out, so that we are left with a tensor rank 1 tensor (very not obvious!):

$$g^{\mu\nu}\delta\Gamma^\alpha_{\mu\nu}-g^{\mu\alpha}\delta\Gamma^\nu_{\mu\nu}:=A^\alpha\quad(12.4.815)$$

So that we are left with the following integral expression:

$$\int\nabla_\alpha A^\alpha\sqrt{-g}\mathrm{d}^4x\quad(12.4.816)$$

Using the divergence theorem (see section Vector Calculus page 1882):

$$\int\nabla_\alpha A^\alpha\sqrt{-g}\mathrm{d}\Omega=\int A^\alpha\sqrt{-g}\mathrm{d}\mathcal{S}\quad(12.4.817)$$

As the variational are assumed to vanish on the hypersurface $\mathcal{S}$ of the hypervolume Ω, we have:

$$\int\nabla_\alpha A^\alpha\sqrt{-g}\mathrm{d}\Omega=\int A^\alpha\sqrt{-g}\mathrm{d}\mathcal{S}=0\quad(12.4.818)$$

Therefore our action variational reduces to:

$$\begin{aligned}\delta S&=-\frac{1}{\kappa c}\int\left[\delta g^{ij}R_{ij}\sqrt{-g}-\frac{1}{2}\sqrt{-g}g_{ij}\delta g^{ij}R\right]\mathrm{d}x^4\\&=-\frac{1}{\kappa c}\int\left[R_{ij}-\frac{1}{2}g_{ij}R\right]\delta g^{ij}\sqrt{-g}\mathrm{d}x^4\end{aligned}\quad(12.4.819)$$

For the action to be stationary, the last expression must be equal to zero whatever the variations of the coefficients of the metric δg^{ij}, which imposes that the expression between braces is null:

$$R_{ij}-\frac{1}{2}Rg_{ij}=0\quad(12.4.820)$$

We recognize here the Einstein's field equations without matter! So it seems that our Lagrangian was a good candidate (but it needs more verification, especially, experimental tests!).

12.4.7.1 Einstein-Hilbert action with cosmological constant

Let us try now:

$$S=-\frac{1}{\kappa c}\int(R+\lambda)\sqrt{|g|}\mathrm{d}x^4 \tag{12.4.821}$$

where λ is a constant that we add to the curvature scalar R. The latter relation will obviously be written more explicitly as:

$$S=-\frac{1}{\kappa c}\int(g^{ij}R_{ij}+\lambda)\sqrt{|g|}\mathrm{d}x^4=-\frac{1}{\kappa c}\left[\int g^{ij}R_{ij}\sqrt{|g|}\mathrm{d}x^4+\int\lambda\sqrt{|g|}\mathrm{d}x^4\right] \tag{12.4.822}$$

We already know the result of the first integral as it is simply the previous result we have get. Hence we need to focus only on the second integral!

If we apply as before the variation:

$$\delta S=-\frac{\lambda}{\kappa c}\int\delta(\sqrt{|g|})\mathrm{d}x^4 \tag{12.4.823}$$

and using the fact that we have proved just earlier that:

$$\delta\sqrt{|g|}=\delta\sqrt{-g}=-\frac{1}{2}\sqrt{-g}g_{ij}\delta g^{ij} \tag{12.4.824}$$

We then get:

$$\delta S=-\frac{\lambda}{\kappa c}\int\delta(\sqrt{|g|})\mathrm{d}x^4=-\frac{\lambda}{\kappa c}\int\left(-\frac{1}{2}\sqrt{-g}g_{ij}\delta g^{ij}\right)\mathrm{d}x^4 \tag{12.4.825}$$

Hence the global variation is:

$$\begin{aligned}\delta S&=-\frac{1}{\kappa c}\int\left[\delta g^{ij}R_{ij}\sqrt{-g}-\frac{1}{2}\sqrt{-g}g_{ij}\delta g^{ij}R--\frac{\lambda}{2}\sqrt{-g}g_{ij}\delta g^{ij}\right]\mathrm{d}x^4\\&=-\frac{1}{\kappa c}\int\left[R_{ij}-\frac{1}{2}g_{ij}R-\frac{\lambda}{2}g_{ij}\right]\delta g^{ij}\sqrt{-g}\mathrm{d}x^4\end{aligned} \tag{12.4.826}$$

As before, for the action to be stationary, the last expression must be equal to zero whatever he variations of the coefficients of the metric δg^{ij}, which imposes that the expression between braces is null:

$$R_{ij}-\frac{1}{2}Rg_{ij}-\frac{\lambda}{2}g_{ij}=0 \tag{12.4.827}$$

We recognize here the Einstein's field equations without matter and cosmological constant!

It is because of the shape of that result that the first relation we started with is finally written traditionally (in the purpose to be in agreement with the notation of the General Einstein's field equations):

$$\boxed{S=-\frac{1}{\kappa c}\int(R-2\Lambda)\sqrt{|g|}\mathrm{d}x^4=-\frac{c^3}{8\pi G}\int(R-2\Lambda)\sqrt{|g|}\mathrm{d}x^4} \tag{12.4.828}$$

12.4.7.2 Einstein-Hilbert action with matter

We see that the above result gives use the Einstein's field equations without matter. Ok good but... we want a bit more generalized result. Especially the result we already know:

$$R_{\mu\nu}-\frac{1}{2}Rg_{\mu\nu}=\frac{8\pi G}{c^4}T_{\mu\nu} \quad (12.4.829)$$

So to deduce the corresponding form of the action and corresponding Lagrangian density we will proceed by reverse engineering.

We know that the previous relations will in fact be:

$$R_{\mu\nu}-\frac{1}{2}Rg_{\mu\nu}-\frac{8\pi G}{c^4}T_{\mu\nu}=0 \quad (12.4.830)$$

And if we compare with the previous result:

$$\delta S=-\frac{1}{\kappa c}\int\left[R_{ij}-\frac{1}{2}g_{ij}R\right]\delta g^{ij}\sqrt{-g}\mathrm{d}x^4 \quad (12.4.831)$$

By reverse engineering we will put:

$$\delta S=-\frac{1}{\kappa c}\int\left[R_{\mu\nu}-\frac{1}{2}Rg_{\mu\nu}-\frac{8\pi G}{c^4}T_{\mu\nu}\right]\delta g^{ij}\sqrt{-g}\mathrm{d}x^4 \quad (12.4.832)$$

Ok the variational was very easy to reverse engineer. But for the action and the Lagrangian we are looking for something such that (as we know the origin of the first two terms):

$$\delta(???)=-\frac{8\pi G}{c^4}T_{\mu\nu}\delta g^{ij} \quad (12.4.833)$$

So we will try:

$$\delta\left(\frac{8\pi G}{c^4}T\right)=\delta\left(\frac{8\pi G}{c^4}T_{ij}g^{ij}\right)=\frac{8\pi G}{c^4}\delta\left(T_{ij}g^{ij}\right) \quad (12.4.834)$$

So let see what we get by developing $\delta\left(T_{ij}g^{ij}\right)$. First we must recall that the variational principle request to evaluate the variation of the action in function of the variation of the metric $\delta(g^{ij})$. Therefore:

$$\delta\left(T_{ij}g^{ij}\right)=\underbrace{\delta(T_{ij})g^{ij}}_{=0}+T_{ij}\delta(g^{ij})=T_{ij}\delta(g^{ij}) \quad (12.4.835)$$

And that's it! So the corresponding action is finally:

$$\boxed{S=-\frac{1}{\kappa c}\int\left[R-\frac{8\pi G}{c^4}T\right]\sqrt{-g}\mathrm{d}x^4=-\frac{c^3}{8\pi G}\int\left[R-\frac{8\pi G}{c^4}T\right]\sqrt{-g}\mathrm{d}x^4} \quad (12.4.836)$$

Hence the density Lagrangian:

$$\boxed{\mathcal{L}=\frac{c^4}{8\pi G}R-T=\mathcal{L}_{\text{EH}}+\mathcal{L}_M} \quad (12.4.837)$$

Obviously we can add the cosmological constant to get:

$$\boxed{\mathcal{L}=\frac{c^4}{8\pi G}(R-2\Lambda)-T=\mathcal{L}_{\text{EH},\Lambda}+\mathcal{L}_M} \quad (12.4.838)$$

Remark

There are different ways to derive the above results. Some, as above, use awful and simple reverse engineering reasoning. Others - like the one available on the Wikipedia page of the Einstein-Hilbert action - use circular reasoning about the Stress-Energy tensor and therefore leads to a slightly different expression of the Lagrangian density (implying some constant factor) with the advantage of having a beautiful technical mathematical development (at the opposite of the reverse engineering method...).

If we combine the Lagrangian derived in the section of Electromagnetism, in the section of Special Relativity, in the section of Quantum Theory and the one above of General Relativity, we get so far:

$$\begin{aligned}\mathcal{L}&=\mathcal{L}_{\text{GR}}+\mathcal{L}_{\text{QED}}\\&=\underbrace{\mathcal{L}_{\text{EH},\Lambda}+\mathcal{L}_M}_{\text{General Relativity}}+\underbrace{\mathcal{L}_{\text{DSF}}+\mathcal{L}_{\text{ED}}}_{\text{Quantum Electrodynamics}}\\&=\underbrace{\underbrace{\frac{c^4}{8\pi G}(R-2\Lambda)}_{\text{Einstein-Hilbert}}-\underbrace{T_{\mu\nu}}_{\text{Masse}}}_{\text{General Relativity}}+\underbrace{\underbrace{\bar{\Psi}\left(\mathrm{i}\gamma^\mu\partial_\mu-m_0\right)\Psi}_{\text{Dirac spinor fields}}+\underbrace{-\rho_0c^2\sqrt{1-\frac{v^2}{c^2}}-j^\mu A_\mu+\frac{1}{4}\mu_0F_{\mu\nu}F^{\mu\nu}}_{\text{Electrodynamics}}}_{\text{Quantum Electrodynamics}}\end{aligned}\tag{12.4.839}$$

~90 % ★★★☆☆

33 votes, 61.82%

12.5. Cosmogony

COSMOGONY is concerned with understanding the (hypothetical) birth and evolution of the Universe by the scientific method. It is only through this game between physical theories, models and observations that we will discuss this issue here. We will try to avoid carefully any metaphysical digression. The specific problems of cosmogony fit in its definition: Statistics, that are one of the greatest scientific methods, are apparently poorly efficient for the study of Universe theories as we have only one Universe visible to us at this day. Furthermore, we can only observe the past of our Universe (and furthermore only the observable part...). Can we speak about "predictions" in these conditions? The theories, however, are reliable since they predict behaviours that can be tested by observations.

Cosmogony mainly uses the arsenal of mathematics, theoretical physics, particle physics, nuclear physics, physics of detectors and astrophysics. It is interdisciplinary! Cosmogony deals with scales larger than the size of a galaxy to the scales defined as itself as "Horizon". Even if the limit is deliberately vague, cosmology does not address the internal details of the birth and evolution of astrophysical objects (such as galaxies, globular clusters, and clusters of galaxies) that fall more into the study field of "cosmogony".

quote
12.5:5878.#71

> ❞ *If you can explain a god without a creator, you can explain a Universe without a creator...*

”

By the way... before we start with the maths..., for the answer to the question *Why does the Universe exist?* the reader should know first that the Universe has no obligation to make sense to us! Secondly, the "First Cause Argument" is a human bias! Indeed, through modern science, specifically physics, natural phenomena have been discovered whose causes have not yet been discerned or are non-existent!! The best known example - among many others (!) is radioactive decay! Although decay follows statistical laws and it's possible to predict the amount of a radioactive substance that will decay over a period of time, it is impossible according to our current understanding of physics to predict when a specific atom will disintegrate as we have proved it in the section of Nuclear Physics. The spontaneous disintegration of radioactive nuclei is stochastic and might be uncaused, providing an arguable counterexample to the assumption that everything must have a cause. An objection to this counterexample is that knowledge regarding such phenomena is limited and there my be hidden variables presently unknowns. But this latter objection has been mathematically eliminated (Bell Theorem).

The Universe hasn't inflate obviously for us Homo Sapiens! The fact that we can't survive anywhere else easily outside of the tiny ball of rock and metal - largely inhabitable for us and where most water isn't even pure drinkable water... - we are travelling on without cost-prohibitive and high level technologies just for a one way trip, plus the fact that our little ball of rock and metal has an unstable orbit (perturbed by travelling stars in our Galaxy), is periodically bombarded of extinction level asteroids coming from the Oort cloud or bombarded by gamma rays bursts from potential dying stars, with an unstable long term climate and unstable supervolcano dynamics, and an unstable kernel, and an unstable main star is an obvious evidence! We can also consider as supplementary evidence that all other worlds (planets) we have been able to observed actually that are at a respectable distance for a one-way trip of thousands of years are inhabitable for us... Claiming the opposite is so naive that it is similar as putting a fish in a tiny aquarium in the middle of the desert and listening to the fish claiming the desert was made especially for him...

Also keep in mind that while reading what will follow that the mainstream Big Bang model (cosmic inflation theory) only tell us how matter and energy expanded from an initial dense state to what we observe now. It doesn't state that matter and energy were created during the Big Bang, nor does it tell us where they're coming from (that's more related to the study of quantum cosmology[54]).

Remark

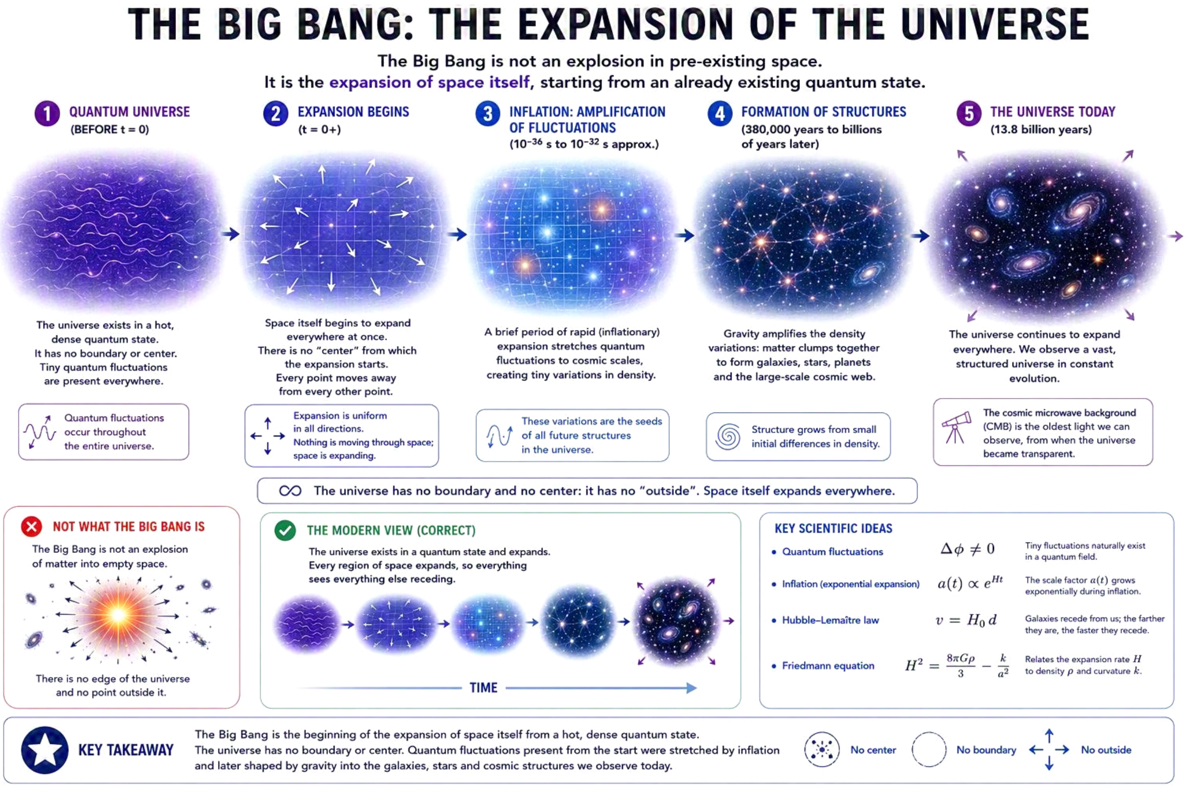
Let us recall the following fact that we have already pointed out in the section of Probabilities: Some scientific illiterate people may argue that Roger Penrose calculated (on the assumption that our Universe is in a Black Hole a with very rough approximations and assumptions...) that for our Universe to be in the low entropy like now is $10^{10^{123}}=10^{123}$ to 1. Hence it is impossible that our Universe is not created by a deity (passing under

---

[54] There are actually (early 121st century according to the holocene calendar) ten theoretical models giving possible answers to what was existing before the Big Bang: Spontaneous Symmetry Break (SSB-Ex nihilo), Quantum tunnelling fluctuation from WdW equation (WdWE), Conformal Cyclic Cosmology (CCC), Steinhardt-Turok Cyclic Cosmology (STCC), Ijjas-Steinhardt Cyclic Cosmology (ISCC), String Gas Cosmology (SGC), Hartle-Hawking no-boundary measure (HHNBM), Loop Quantum Gravity pre-Big-Bang (LQGpBB), Quantum tunnelling fluctuation from Loop Quantum Gravity (QTFLQG), Ekpyrotic Universe...

> silence the question of who created that complex deity...). But it is irrelevant! Even if what he said is 100% accurate or off by a factor of a googol, it's irrelevant!! Indeed, if run a computer program that emulates a coin flipped one billion times and writes the sequence of heads and tails to a file. Great. What we just produced is a sequence more than a trillion trillion trillion times rarer!! In fact, you could run that same computer program until the end of human existence and it would likely never produce that same string of outcomes. And yet it occurred nonetheless...

Definition (#12.5.547): The "Big Bang" (cosmic inflation theory), according to actual cutting edge most promising models, may be either the result of a tunnelling quantum vacuum fluctuation in a true vacuum bubble described by a minisuperspace model in the Wheeler de Witt framework, a symmetry-breaking event in the framework of String Theory or the result of an infinite cyclic conformal universe (in the Penrose framework). It must in no way be considered as a punctual explosion of matter and space-time (that reference was in fact an old joke that most non-specialized scientific medias and religious bigots like to spread even today) nor be considered to be a magical creation by an imaginary friend in the sky. We have anyway no experimental evidence yet in the beginning of the 121st century (holocene calendar) that can confirm beyond reasonable doubt any of the dozens of main different speculative models describing the early and future history of our universe !



THE BIG BANG: THE EXPANSION OF THE UNIVERSE

The Big Bang is not an explosion in pre-existing space.
It is the expansion of space itself, starting from an already existing quantum state.

1 QUANTUM UNIVERSE (BEFORE t = 0)
The universe exists in a hot, dense quantum state. It has no boundary or center. Tiny quantum fluctuations are present everywhere.
Quantum fluctuations occur throughout the entire universe.

2 EXPANSION BEGINS (t = 0+)
Space itself begins to expand everywhere at once. There is no "center" from which the expansion starts. Every point moves away from every other point.
Expansion is uniform in all directions. Nothing is moving through space; space is expanding.

3 INFLATION: AMPLIFICATION OF FLUCTUATIONS (10^{-36} s to 10^{-32} s approx.)
A brief period of rapid (inflationary) expansion stretches quantum fluctuations to cosmic scales, creating tiny variations in density.
These variations are the seeds of all future structures in the universe.

4 FORMATION OF STRUCTURES (380,000 years to billions of years later)
Gravity amplifies the density variations: matter clumps together to form galaxies, stars, planets and the large-scale cosmic web.
Structure grows from small initial differences in density.

5 THE UNIVERSE TODAY (13.8 billion years)
The universe continues to expand everywhere. We observe a vast, structured universe in constant evolution.
The cosmic microwave background (CMB) is the oldest light we can observe, from when the universe became transparent.

The universe has no boundary and no center: it has no "outside". Space itself expands everywhere.

NOT WHAT THE BIG BANG IS
The Big Bang is not an explosion of matter into empty space.
There is no edge of the universe and no point outside it.

THE MODERN VIEW (CORRECT)
The universe exists in a quantum state and expands. Every region of space expands, so everything sees everything else receding.

KEY SCIENTIFIC IDEAS

Quantum fluctuations	$\Delta\phi \neq 0$	Tiny fluctuations naturally exist in a quantum field.
Inflation (exponential expansion)	$a(t) \propto e^{Ht}$	The scale factor $a(t)$ grows exponentially during inflation.
Hubble–Lemaître law	$v = H_0 d$	Galaxies recede from us; the farther they are, the faster they recede.
Friedmann equation	$H^2 = \frac{8\pi G\rho}{3} - \frac{k}{a^2}$	Relates the expansion rate H to density ρ and curvature k.

KEY TAKEAWAY
The Big Bang is the beginning of the expansion of space itself from a hot, dense quantum state. The universe has no boundary or center. Quantum fluctuations present from the start were stretched by inflation and later shaped by gravity into the galaxies, stars and cosmic structures we observe today.

In a survey made on 80 professional physicists (see [96]) in 12024 (holocene calendar) we get the following result for the fate of the matter that falls into a Black Hole:

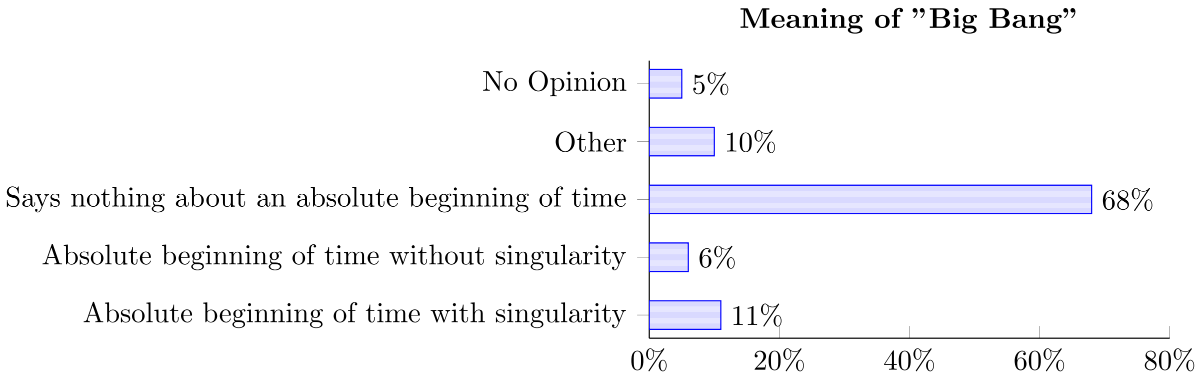


Figure 12.250: **Copenhagen survey on how the "Big Bang" should be understood**

12.5.1 Newtonian (i.e. non-relativistic) Cosmological Models

A cosmological model is a mathematical representation of the Universe that seeks to explain the reasons for its present appearance, and describe its evolution over time (named "cosmological time") but not its creation!

The Newtonian model (also named "non-relativistic cosmology") applies under the assumptions of Newtonian mechanics (instantaneous action for example). The results we are going to study here were discovered before the development of General Relativity but published after! But this Newtonian model has the advantage, even if sometimes the mathematical assumptions and manipulations are very quite wrong, of simplicity while being able to identify and discuss the dynamics of the Universe and to prepare for the study of the Universe models making use of the results of General Relativity afterwards. Its disadvantage, besides the fact that it doesn't quite fit the experimental results, it is to be no longer valid under extreme conditions and therefore cannot be extrapolated to the instant of the Big Bang.

Before we begin, we must define the "cosmological principle" consisting of the following two assumptions (basically, it ensures that we are not privileged observers, and that what we are seeing is a representative of the whole of the Universe):

H1. The space (the Universe) is homogeneous, that is to say, it has the same properties in all its regions. This must be at a very large scale, beyond the thousand Mpc (megaparsecs). It is clear that small-scale inhomogeneities exist... we for example... ☺

H2. The space (Universe) is isotropic[55] at large scales, that is to say, there is no specific direction in space, such as flattening a direction or an overall movement on a Universal scale for example and all its physical properties are almost identical in any point.

For an in-deep discussion and criticism of the subject based on the earliest experimental data (of this early 121st century according to holocene calendar), the reader can refer to the excellent

---

[55] There is a disagreement in the beginning of the 121st century (holocene calendar) about this assumption that need more precise studies. Some physicists have a statistical evidence of a "dark flow" in the CMB - providing at the same time a possible argument for Multiverse model - some say it is only a noise. To follow...

100 pages paper *Is the observable Universe consistent with the cosmological principle?* (see [15]).

Remark

This hypothesis of the isotropy of the Universe and that works relatively well in theoretical models (see below) requires an interesting fact if we admit a beginning to the Universe. This fact implies that the Universe had a phase in its history where it did not leave the time to the matter to clump together to form from its beginnings inhomogeneous and anisotropic material groups that are visible today in our telescopes. Therefore it follows that at a moment of its history, the Universe had a non-quasistatic expansion rate that we could make match with the speed of light (this is badly formulated but I hope it is still acceptable in the idea).

The figure below shows the Automated Plate Measurement (APM) Galaxy Survey. Over 2 million galaxies are depicted in a region of 100 degrees across centered toward the Milky Way's south pole:

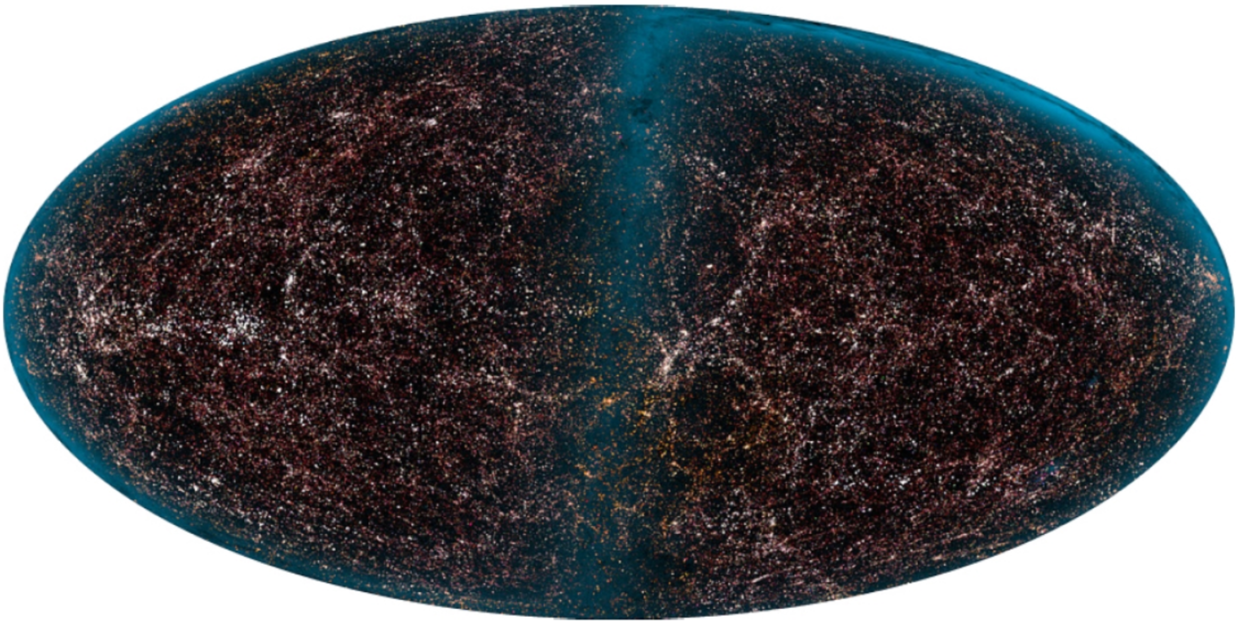


Figure 12.251: **Automated Plate Measurement (APM) Galaxy Survey**

We will also request other working assumptions/hypothesis:

H1. The Universe is a non viscous gaseous fluid whose particles are galaxies. Assuming the cosmological principle, the movement of galaxies, constituents in-fine of this "fluid" by construction, are statistically at rest.

H2. The Universe is a thermodynamically closed system, without work and adiabatic (no heat exchange with the outside).

H3. The Universe in a homothetic expansion (in uniform expansion in all its dimensions) is taken as having a spherical shape with a center (at least in the Newtonian model...).

H4. Its density is only a function of time and there is mass conservation (and therefore energy). Therefore the amount of material is constant (yes this is always the Newtonian model...)!

H5. We accept the Newtonian dynamics (approximation of General Relativity) to build the Newtonian models that will follow.

H6. The origin of time is treated as the origin of creation (horizon) of the Universe and repository of study is comoving with the particles (and therefore moves with the galaxies placed on the space-time pattern) and named "reference material" (galaxies are stationary in this repository!).

12.5.1.1 Hubble's Law

Assuming the cosmological principle and the above assumptions, the distance from an origin point O to any point M of the universe can vary in function of time (in a undetectable way to the human scale) in the form:

$$\overrightarrow{\mathrm{O}M}(t)=\overrightarrow{\mathrm{O}M}(t_0)F(t) \quad (12.5.1)$$

where $F(t)$ is the "scaling factor" (denoted by $R(t)$ or even $a(t)$ depending on the context...).

Notice that we can't have $\overrightarrow{\mathrm{O}M}(t_0)=0$ otherwise, whatever the value of $F(t)$, $\overrightarrow{\mathrm{O}M}(t)$ would always be equal zero... (this eliminate the theoretical possibility that the Universe may have had one day no size!). Also notice that $F(t)\in\mathbb{R}^+\backslash\{0\}$, i.e. that means that the size of the Universe may decrease or increase in size but never come to a size equal to zero!

Caution! There is a theorem (see [64]), in the framework of General Relativity, famous among cosmologists, showing that an inflationary universe is past-timelike-incomplete (we will provide the proof of that theorem in the future). What this means, explicitly, is that if you have any particles that exist in an inflating Universe, they will eventually meet if you extrapolate back in time. This doesn't, however, mean that there must have been a singularity, but rather that inflation doesn't describe everything that occurred in the history of the Universe, like its beginning. We also know, for example, that inflation cannot arise from singular state, because an inflating region must always begin from a finite size.

In writing this relation, we consider that the points O and M are on a plane of zero curvature. Indeed, if we imagine two points on a circular curved surface (e.g. the surface of a sphere) let us see what happens:

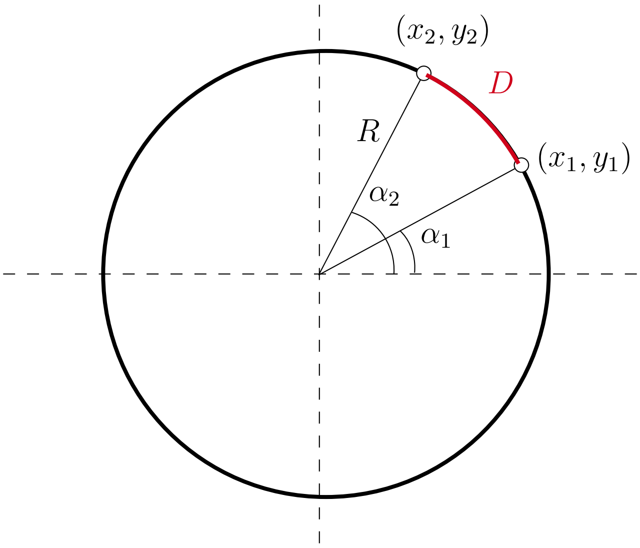


Figure 12.252: **Illustration of the validity limit of the model**

The distance between two points on the circle (i.e. spherical space) is given by (see section

Trigonometry page 2542):

$$D=R\Delta\alpha=R(\alpha_2-\alpha_1)=R\left(\arcsin\left(\frac{y_2}{R}\right)-\arcsin\left(\frac{y_1}{R}\right)\right)=R\left(\arccos\left(\frac{x_2}{R}\right)-\arccos\left(\frac{x_1}{R}\right)\right) \tag{12.5.2}$$

We see very well in this relation that if the radius (of the spherical Universe) changes by a factor F, then the change in the distance between the two points is not linearly proportional to this factor!! Which is not the case in a zero curvature plane!

Consequence: Our Newtonian model is valid only in a flat universe where General Relativity or purely classical energy approach (see further below) can take into account different types of curvature!

We see immediately that the relation:

$$\boxed{\overrightarrow{\mathrm{O}M}(t)=\overrightarrow{\mathrm{O}M}(t_0)F(t)} \tag{12.5.3}$$

is independent of the chosen origin, in fact, if we apply it on any two points A, B we have:

$$\overrightarrow{\mathrm{O}A}(t)=\overrightarrow{\mathrm{O}A}(t_0)F(t) \qquad \overrightarrow{\mathrm{O}B}(t)=\overrightarrow{\mathrm{O}B}(t_0)F(t) \tag{12.5.4}$$

Then by difference:

$$\overrightarrow{AB}(t)=\overrightarrow{AB}(t_0)F(t) \tag{12.5.5}$$

Remark

At the time t_0 it is obvious that the above equation is:

$$\overrightarrow{AB}(t_0)=\overrightarrow{AB}(t_0)F(t_0) \tag{12.5.6}$$

and imposes $F(t_0)=F(0)=1$. This is important and we will come back on it many times during the developments that will follow.

The law above therefore applies to any segment $\overline{AB}$ in the Universe. This is why the Universe has no geometrical center (at least as far as we know) and that we can represent ourselves the expansion of the frame of the Universe: consider a half-inflated balloon on whose surface we draw two marks (eg: two crosses drawn with ink). Inflating the balloon more we find that these two cross diverge from each other and therefore the distance between them will increase. This is what we are seeing with the galaxies:

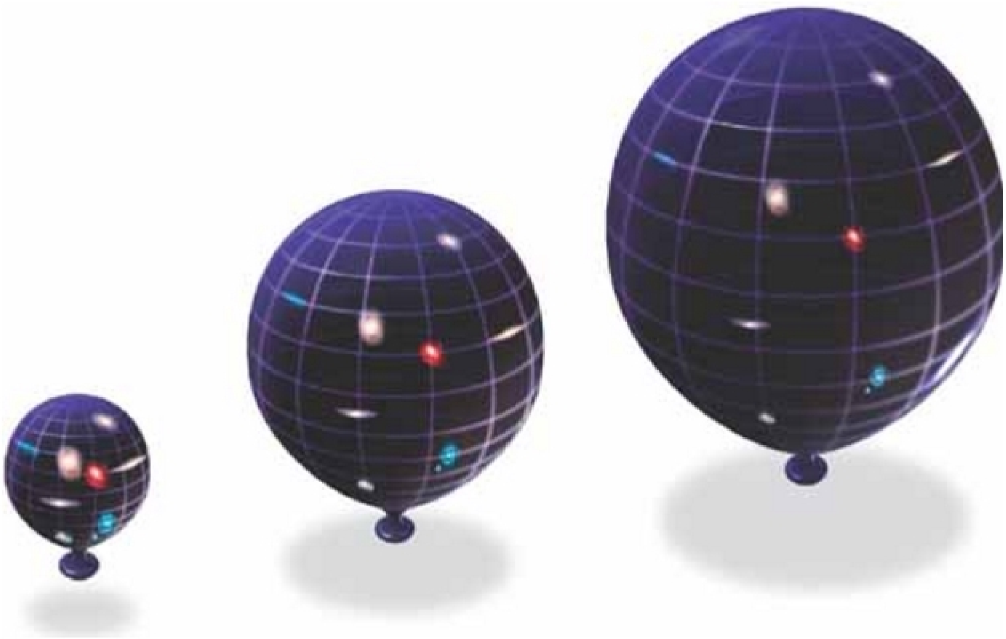


Figure 12.253: **Illustration of the inflated balloon Universe model**

Remarks

R1. If the Universe is expanding, why aren't we? The reason for this is that there are other physical phenomena at play besides the Universe's expansion. On small scales, like animal-sized and below, electromagnetism and nuclear forces dominate. On larger scales, like that of planets, Solar Systems and galaxies, gravitational forces dominate. On the largest scales of all, the expansion wins. But there are also smaller scales, where the expansion has been overcome, at least locally. The Virgo cluster itself will remain gravitationally bound. The Milky Way and all the local group galaxies will stay bound together, and eventually merge under their own gravity. It's only on the largest scales of all, where all the binding forces between objects are too weak to defeat the speedy Hubble rate, that expansion occurs at all.

R2. If the Universe expands, then where it expands to? The Universe isn't expanding into anything. If there is something "beyond" the Universe, that would, by construction, also be part of the Universe. Our limited human intuition usually fails us on this concept because we think of the Big Bang as an explosion and explosions expand into their surrounding space. The Universe, however, is not surrounded by more space. The analogy with the inflating balloon in the figure above helps to grasp the concept (even if as always, this analogy is also a false equivalence!). Not the 3D volume of the balloon, but the 2D surface of the balloon is our Universe. And as the reader can see, it's expanding itself and does not have any border!

Deriving with respect to time the relation:

$$\overrightarrow{\mathrm{O}M}(t)=\overrightarrow{\mathrm{O}M}(t_0)F(t) \tag{12.5.7}$$

The first member then gives the particle velocity (or other any object) to the point $\vec{r}(t)=\overrightarrow{\mathrm{O}M}(t)$:

$$\vec{v}(\vec{r},t)=\overrightarrow{\mathrm{O}M}(t_0)\mathrm{d}_tF(t)=\vec{r}(t_0)\mathrm{d}_tF(t) \tag{12.5.8}$$

Therefore eliminating $\overrightarrow{\mathrm{O}M}(t_0)$ (we see that if we have $F(t)=0$ we have indeed an issue or a... "singularity"...) we get:

$$\vec{v}(\vec{r},t)=\frac{\overrightarrow{\mathrm{O}M}(t)}{F(t)}\mathrm{d}_tF(t)=\frac{\vec{r}(t)}{F(t)}\mathrm{d}_tF(t)=\frac{1}{F(t)}\mathrm{d}_tF(t)\vec{r}(t) \tag{12.5.9}$$

We put to simplify the notations:

$$H(t)=\frac{1}{F(t)}\mathrm{d}_tF(t) \tag{12.5.10}$$

Therefore we have:

$$\boxed{\vec{v}(\vec{r},t)=H(t)\vec{r}(t)} \tag{12.5.11}$$

This relation is known as the "Hubble law" or "velocity-distance relation" and which, according to historical research should have for paternity rather Georges Lemaître... That's why it is sometimes named the "Hubble-Lemaître law". However... (!) although widely attributed to Edwin Hubble or Georges Lemaître, the notion of the universe expanding at a calculable rate was first derived from General Relativity equations in 11922 (holocene calendar) by Alexander Friedmann and five year afters in a very different shape ($\mathrm{d}\lambda/\lambda=r/R_0\sqrt{3}$) by Lemaître. But that effect was already known experimentally and modelized in 11914 (holocene calendar) by Vesto Slipher and written for the first time in it's modern shape in 11928 (holocene calendar) by Howard P. Robertson (see [501] for all the details).

Before going further, it is necessary to pause on this equation for the present moment:

$$\vec{v}(\vec{r},t_0)=H(t_0)\vec{r}(t_0)=H_0\vec{r}_0 \tag{12.5.12}$$

This equation says that the object of the Universe recede with a speed proportional to their distance in all points of the Universe without special repository (no galaxy seems to be fixed while they are in the material repository!).

Remark

This relation allows for speeds higher than those of light... But this is not a violation of Special Relativity regarding the constancy of the speed of light! Indeed, we must not forget that the Hubble law takes into account the expansion of the "frame" of space-time on which light moves. Also Special Relativity deal with speed of light and not with speed of space itself. Keep in mind that the expansion of the Universe doesn't have a "speed" (typical error of amateur scientists!). Therefore if the frame extends along an expansion factor F greater than one, it gives the impression that light travels faster than c, and this is what gives sometimes a redshift value of 4 or 5 or even 11.09 (the actual record)!

The idea of even talking about "the expansion velocity of the universe" is bizarre and never should have been entertained in the first place in massmedia and especially because there is no well-defined notion of "the velocity of distant objects" in General Relativity (there is simply no such thing as the "velocity" between two objects that aren't in a first approximation located in the same place!). So the fact that some people think that what's special about the so named "inflation" is that the Universe was expanding faster than light is a crime against comprehension and good taste (what's special about inflation is only that the universe is accelerating!).

The constant $H(t_0)=H_0$ being of course being identifiable to the "Hubble constant[56]" (but that is not a constant...) and currently measured in this early 12000s (holocene calendar) as being approximately equal[57] $\sim 67.66\,[\text{km}\cdot\text{s}^{-1}\cdot\text{Mpc}^{-1}]$. That means that for every million parsecs of distance from an observer, the rate of expansion increases by about 67.66 kilometres par second! In SI Units that's equal to approximately $2.4\cdot 10^{-18}\,[\text{s}^{-1}]$ (the reader should always refer to the most recent published scientific papers to get the actual value!).

We can actually calculate an estimate for the age of he Universe from Hubble's Law. Let us consider that the actual distance between two galaxies is D. The apparent velocity with which they are separating from each other is v. At some point, the galaxies were touching, and we can consider that time the moment of the Big Bang. If we take the separation between the two galaxies (D) and divide that by the apparent velocity (v), that will leave us with how long it took for the galaxies to reach heir current separation.

So, the time it has taken for the galaxies to reach their current separations is $t=D/v$. But, from Hubble's Law, we know that $v=H_0D$. Therefore:

$$t=\frac{D}{v}=\frac{D}{H_0\cdot D}=\frac{1}{H_0} \qquad (12.5.13)$$

So, we can take $1/H_0$ as an estimate for the age of the Universe (as we know that very likely H_0 isn't a constant).

The best estimate at the day we write these line being $H_0\cong 67.66\,[\text{km}\cdot\text{s}^{-1}\cdot\text{Mpc}^{-1}]$. To turn this into an age, we'll have to do a unit conversion obviously!

Since $1\,[\text{Mpc}]=3.085\cdot 10^{19}\,[\text{km}]$, therefore:

$$H_0\cong\frac{67.66\,[\text{km}\cdot\text{s}^{-1}\cdot\text{Mpc}^{-1}]}{3.085\cdot 10^{19}\,[\text{km}\cdot\text{Mpc}^{-1}]}\cong 2.193\cdot 10^{-18}\,[\text{s}^{-1}] \qquad (12.5.14)$$

So, the age of the Universe based on this approximately equal to the "Hubble time":

$$\boxed{t_H=\frac{1}{H_0}\cong 4.5596\cdot 10^{10}\,[\text{s}]\cong 14.48\cdot 10^{9}\,[\text{years}]} \qquad (12.5.15)$$

that is to say about 14.5 billion years ago (we will see further below a better approach to calculate the age of our Universe). Taking into account margin of error of early 121st century measurements (holocene calendar), the upper bound age is 14.5 billions years and lower bound is 13.2 billion years.

---

[56] We should rather speak of "Hubble parameter" as it is not constant through time (since there are dynamical forces acting on the particles in the Universe which affect the expansion rate).

[57] The last value (12018-07-18) to $67.66\pm 0.42\,[\text{km}\cdot\text{s}^{-1}\cdot\text{Mpc}^{-1}]$. But there is still in year 12021 (holocene calendar) a 5σ tension in the estimation of H_0 between the direct measurement method and the indirect measurement method as beautifully summarised in the paper [144] at page 10 (paper that also summarizes the few hundred of theoretical alternatives and improved versions of the ΛCDM model).

Caution! As we will prove it further below (see page 5916), the Hubble time is strictly speaking not the age of our universe. The fact that the Hubble time is approxomitely equal to the age of our Universe is a reflection that the rate of expansion has been nearly constant for a long time. In reality, the best cosmological theories suggest that the universe has not been expanding linearly since the beginning. So we would expect that the age of the universe is not exactly equal to the Hubble time. But hopefully it makes sense that if any non-linear expansion lasted for only a short period, then the Hubble time should still be close to the age of the universe. That is the situation we see today. And the reader must not forget that the Hubble constant is not a constant (!!!) hence the Hubble time isn't also a constant. For example 6 billion years ago, when the universe was 7.5 billion years old, the Hubble constant was about $100\,[\text{km}\cdot\text{s}^{-1}\cdot\text{Mpc}^{-1}]$, what means the Hubble time was 9.78 billion years. When the universe will be 24 billion years of age, H will be $60\,[\text{km}\cdot\text{s}^{-1}\cdot\text{Mpc}^{-1}]$, and the Hubble time will be 16.3 billion years. Not even close to the age of the universe!

Here is the evolution of the precision of measurement of the Hubble constant:

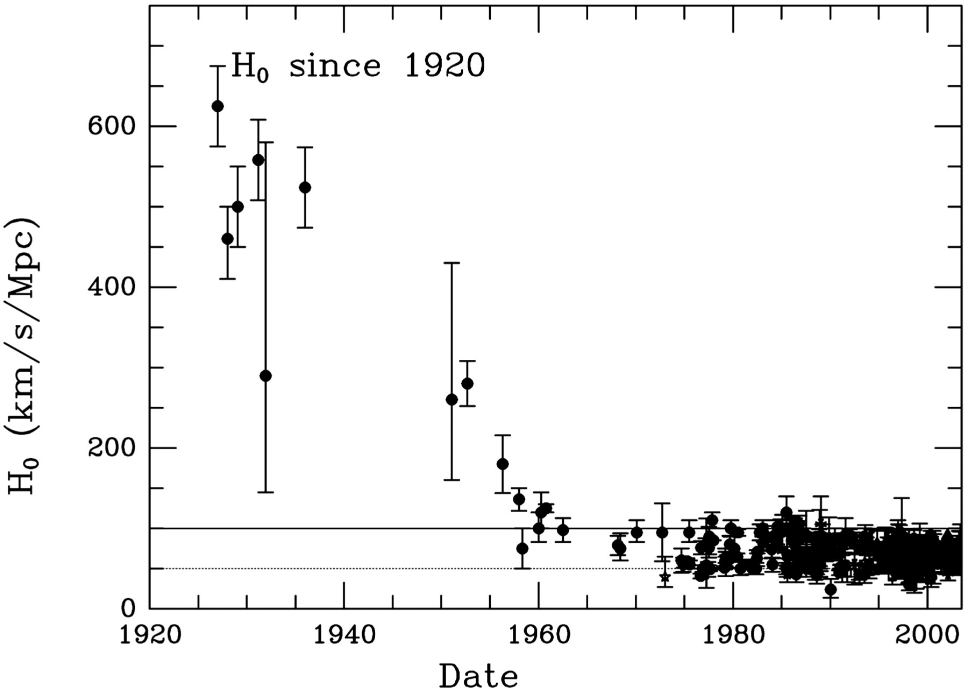
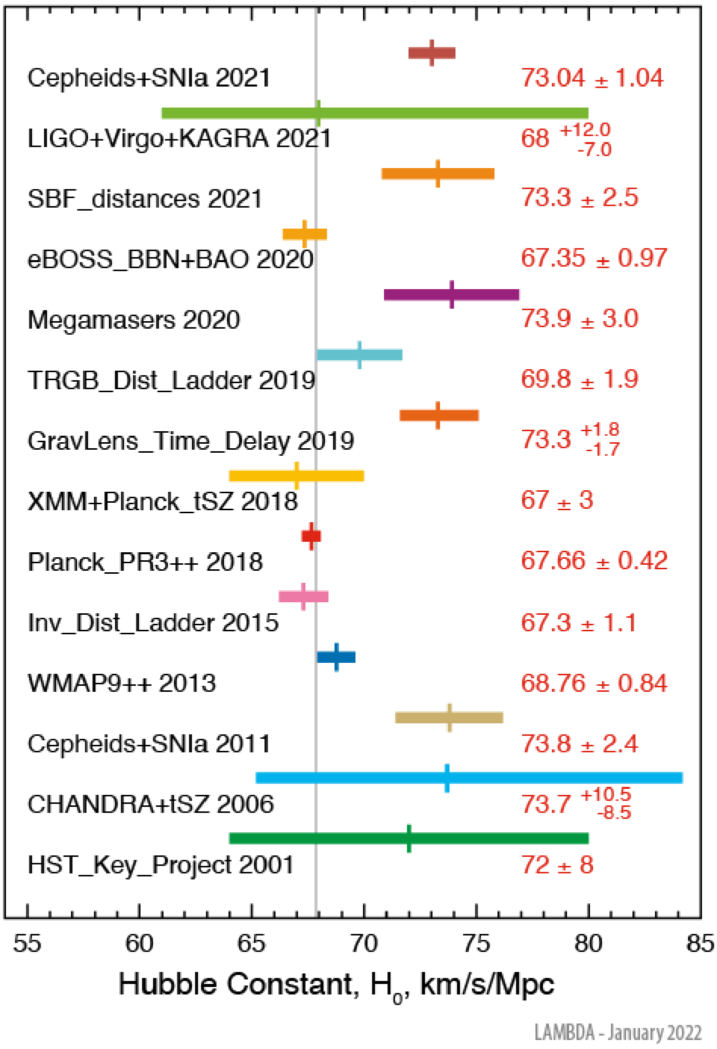


Figure 12.254: **Evolution of Hubble parameter estimations (source: Robert P. Kirshner)**

And some recent measurement at the days we write these lines:



Conversely, we can have fun to calculate the distance from which we (i.e. the observed galaxies) can reach the speed thanks to the equation:

$$\vec{v}(\vec{r},t_0)=H(t_0)\vec{r}(t_0)=H_0\vec{r}_0 \tag{12.5.16}$$

and a numerical application gives roughly 13 billion light years. This is the distance of the "cosmological horizon".

In cosmology, a "Hubble volume, or "Hubble sphere", is a spherical region of the Universe surrounding an observer beyond which objects recede from that observer at a rate greater than the speed of light due to the expansion of the Universe. Regarding the relation we get above, the corresponding radius today is given by (assuming that the Hubble parameter has always been constant):

$$r_{\text{HS}}(t_0)=\frac{c}{H_0}=\frac{299,792.458}{67.66}\cong 4430.86\,[\text{Mpc}]\cong 14.4\cdot 10^9\,[\text{light-years}] \tag{12.5.17}$$

hence almost equal to 14 billion light-years. As observations seem to indicate that our Universe is accelerating, some objects that we can currently exchange signals with, will one day cross our Hubble limit!

But... it takes almost 14.4 billion light-years under these assumptions to reach us so it could have moved another 14.4 billion light years away from us during that time... (still under the same assumptions!). So the "complete radius" would be 28 billion light years away! Then what we will name in the books the "Newtonian radius" of our Universe - not to be confused with the entire Universe - under these assumptions is:

$$r_{\text{Newt. Universe}}(t_0) = r_{\text{HS}} \cong 28.8 \cdot 10^9 \text{ [light-years]} \qquad (12.5.18)$$

Caution! We name this radius of our Universe in this textbook "Newtonian radius" because it's obviously wrong in the framework of General Relativity as we will prove it much further below (see page 6048)! The calculations involving General Relativity and the Friedmann-Lemaître-Robertson-Walker metric give for our observable Universe + our unobservable Universe a radius actually a little bit greater than 40 billion light years (with a margin of error of a few hundred of million of years).

In other words, knowing that the huge majority of galaxies are escaping from our Milky Way, this result has a quite important conclusion: in a far future, the sky as seen from our Galaxy (or what will remain of it...), will be completely dark as (almost) everything will be moving away at a speed greater than that of light! So we are "lucky" to observe the Universe at it is today as otherwise the cosmological model would have been much more harder to establish without the observation of other galaxies... (and keep in mind that this is an effect strictly relative to the observer!).

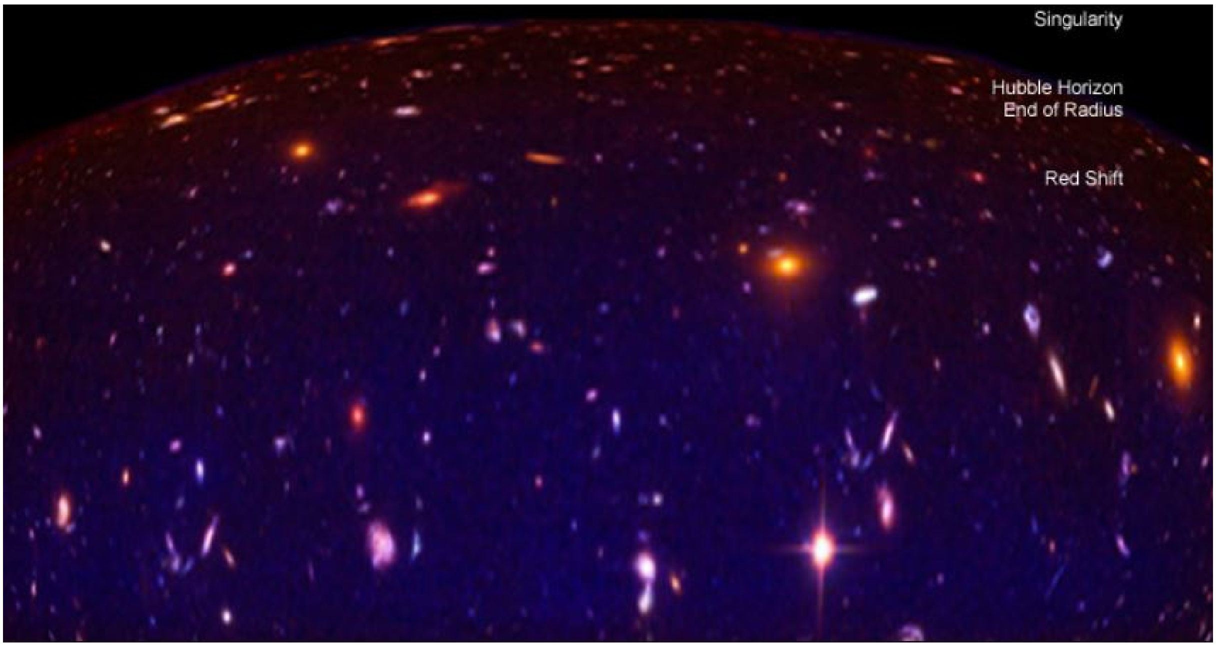


Figure 12.255: An artistic concept of the Hubble Horizon (source: ?)

12.5.2 Friedmann Equations (Newtonian derivation)

Consider now a spherical ring of material of radius r and of constant mass m expanding at velocity v, and containing a ball of material of mass M (also in expansion at speed v).

$$\frac{1}{2}mv^2-\frac{GMm}{r}=k_1 \tag{12.5.19}$$

where k_1 is a constant. By dividing by m and replacing each member M by its expression as a function of the density, we get:

$$\frac{1}{2}v^2-\frac{4\pi}{3}\frac{G\rho r^3}{r}=\frac{k_1}{m} \tag{12.5.20}$$

Remark

If it can help the reader to understand what we did with the term of the potential energy, he can refer to section Classical Mechanics when we developed the calculations of the potential energy of a material sphere (see page 3356).

But Hubble's law gives us according to what we saw just earlier above:

$$v=Hr \tag{12.5.21}$$

and:

$$r=r_0F \tag{12.5.22}$$

We get:

$$\frac{1}{2}H^2r_0^2F^2-\frac{4\pi}{3}G\rho r_0^2F^2=\frac{k_1}{m} \tag{12.5.23}$$

That we simplify in:

$$H^2=\frac{8\pi}{3}G\rho+\frac{2k_1}{mr_0^2}F^{-2} \tag{12.5.24}$$

However, k_1,m,r_0 are constants. We introduce a new constant k defined by (to simplify the notations):

$$k=-\frac{2k_1}{mr_0^2} \tag{12.5.25}$$

So we get the equation:

$$\boxed{H^2=\frac{8\pi G}{3}\rho-kF^{-2}} \tag{12.5.26}$$

which is none other than the "Friedmann's first equation" that we frequently find in the literature as follows (among others notations...):

$$H^2=\left(\frac{R'}{R}\right)^2=\frac{8\pi G\rho}{3}-\frac{k}{R^2} \tag{12.5.27}$$

It is possible to obtain the same equation, but in a much more general and rigorous way, from Einstein's field equations (see section General Relativity page 5766) and the metric of Friedmann-Lemaître-Robertson-Walker (see further below page 5987).

Let us still notice a very common form of that relation. When using Einstein's mass and energy equivalence ($E = mc^2$) the above density ρ is no longer a mass density but an energy density, we will have to divide it again by the squared speed of light to have a mass density again. It is the same if at the denominator of the constant k, the mass m is replaced by energy, then we will have to multiply k by the squared speed of light to fall back on a mass. We then have, denoting now the scale factor by a (as it is often customary in the literature) and redistributing the terms, the following form of the first Friedmann equation:

$$H^2 = \left(\frac{a'}{a}\right)^2 + \frac{kc^2}{a^2} = \frac{8\pi G\rho}{3c^2} \tag{12.5.28}$$

where a is also known as the "cosmic scale factor" or sometimes the "Robertson-Walker scale factor" or "Robertson-Walker radius".

The latter relation is also commonly written:

$$\boxed{\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G\rho}{3c^2} - \frac{k}{a^2}} \tag{12.5.29}$$

or:

$$\frac{\dot{a}^2 + k}{a^2} = \frac{8\pi G\rho}{3c^2} \tag{12.5.30}$$

These equations were derived in 11922 (holocene calendar) by Friedmann, seven years before Hubble's discovery, when Albert Einstein did not trust his own equations because they did not allow the universe to be static. Also Friedmann's equation did not gain recognition until after his death, when they were confirmed by an independent derivation by Georges Lemaître in 11927 (holocene calendar).

Remark

Albert Einstein added to this equation for personal and quasi-religious beliefs a cosmological constant that allowed him to make static the scale factor of the Universe. I (as the main redactor of the book) reject this arbitrary constant (at least until this date), even if in the contemporary physics, it has become a trendy constant (its value was defined more mathematically rather than religiously) because it would explain the origin of a supposed "dark energy" that seems to accelerate the expansion of our Universe (this acceleration could also be a "local" departure of the Hubble parameter from its globally averaged value perhaps caused by a local void in the mass density of our space neighbourhood named a "Hubble bubble"), the current laws of our Universe, the inflationary period of our universe and its geometry. Thus, the first Friedmann equation with the cosmological constant, which is a total artifice of work, is then:

$$H^2 = \frac{8\pi G}{3}\rho - kF^{-2} + \frac{\Lambda}{3} \tag{12.5.31}$$

with:

$$\Lambda=\frac{8\pi G}{c^4}F \tag{12.5.32}$$

This is Andrei Sakharov who defined the value of the above cosmological constant, which supposedly corresponds to the quantum energy of vacuum (depending on the Higgs fields).

In quantum physics the equations of the field associated with elementary particles that are used to define the Big Bang theory (cosmic inflation theory) are one of the main flow of the beginning of this 121st century (holocene calendar). The famous Einstein's field equations tells us that energy is associated with a gravitational field like the electron in motion is associated with an electromagnetic field. It follows from these two observations that by measuring the gravitational field we have a way to determine the energy of vacuum. The gravitational field is no longer about the matter but about the energy density of vacuum. But the cosmological constant is directly proportional to the constant of gravity G. Its measurement is a quite very dangerous game because its value depends on several fundamental laws of physics and of significant properties on the dynamics of our Universe. The debate remains completely open and if I (the main redactor of this textbook) find a valid and rigorous proof of this constant, we will provide to the reader the consequences of this constant on the models that we will see below.

Let us now use the first law of thermodynamics (see section Thermodynamics page 3618) for a system by definition that will be closed and isolated, which the sum of kinetic and potential energy is constant (and therefore the amount of total energy variation is zero for these two energies). We then have that the change in total energy is only given by the variation of internal energy (the most common case in thermodynamics for macroscopic objects):

$$\mathrm{d}E_{\text{tot}}=\mathrm{d}U \tag{12.5.33}$$

and we have also proven in the section of Thermodynamics the characteristic equation of a fluid in equilibrium (see page 3658):

$$\mathrm{d}U=T\mathrm{d}S-P\mathrm{d}V \tag{12.5.34}$$

If the system is adiabatic (no heat transfer between the system and outside), then we have also proven in the section of Thermodynamics that the variation of entropy was (see page 3660 3660):

$$\mathrm{d}S=\frac{\delta Q}{T}=\frac{0}{T}=0\Leftrightarrow\frac{\mathrm{d}S}{\mathrm{d}t}=0 \tag{12.5.35}$$

Then:

$$\mathrm{d}U+P\mathrm{d}V=0 \tag{12.5.36}$$

Assuming the universe being spherical (...), we have:

$$V=\frac{4}{3}\pi R^3 \tag{12.5.37}$$

and in the material repository where galaxies (cosmic fluid particles) are immobile:

$$U=E_{\text{tot}}=mc^2=\rho Vc^2 \tag{12.5.38}$$

Therefore:

$$\mathrm{d}(\rho Vc^2)+P\mathrm{d}\left(\frac{4}{3}\pi R^3\right)=\mathrm{d}\left(\rho\frac{4}{3}\pi R^3c^2\right)+P\mathrm{d}\left(\frac{4}{3}\pi R^3\right)=0 \tag{12.5.39}$$

which simplifies in:

$$\mathrm{d}(\rho R^3c^2)+P\mathrm{d}(R^3)=0 \tag{12.5.40}$$

taking the derivative with respect to the cosmic time t:

$$\begin{aligned}c^2\frac{\mathrm{d}(\rho r^3)}{\mathrm{d}t}+P\frac{\mathrm{d}(R^3)}{\mathrm{d}t}&=c^2\dot{\rho}R^3+c^2\rho 3R^2\dot{R}+3PR^2\dot{R}\\&=c^2\dot{\rho}R^3+3R^2(c^2\rho+P)\dot{R}=0\end{aligned} \tag{12.5.41}$$

therefore we get the "Universe fluid equation" or Universe continuity equation" (for a spherical universe!):

$$\dot{\rho}=-3\left(\rho+\frac{P}{c^2}\right)\frac{\dot{R}}{R} \tag{12.5.42}$$

of denoted in textbooks:

$$\boxed{\dot{\rho}+3\frac{\dot{a}}{a}\left(\rho+\frac{P}{c^2}\right)=0} \tag{12.5.43}$$

There is another way (more beautiful but much more technical) to derive the above equality. Indeed, we will prove much later that (see page 6004):

$$\Gamma^{r'}_{0r'}=\Gamma^{r'}_{r'0}=\frac{\dot{a}}{ca}\qquad\Gamma^{\theta}_{0\theta}=\Gamma^{\theta}_{\theta 0}=\frac{\dot{a}}{ca}\qquad\Gamma^{\phi}_{0\phi}=\Gamma^{\phi}_{\phi 0}=\frac{\dot{a}}{ca} \tag{12.5.44}$$

Therefore:

$$\Gamma^{r'}_{r'0}+\Gamma^{\theta}_{\theta 0}+\Gamma^{\phi}_{\phi 0}=3\frac{\dot{a}}{ca} \tag{12.5.45}$$

Before plugging in to Einstein's equations, it is educational to consider the zero component of the conservation of energy equation:

$$\nabla_\mu T^\mu_0=0 \tag{12.5.46}$$

As we have seen in the section of Tensor Calculus during our study of the covariant derivative (see page 2113) that:

$$\nabla_\mu T^\mu{}_\nu=\partial_\mu T^\mu{}_\nu+\Gamma^\mu_{\mu\lambda}T^\lambda{}_\nu-\Gamma^\lambda_{\mu\nu}T^\mu{}_\lambda=0 \tag{12.5.47}$$

and also remember that[58]:

$$T^\mu{}_\nu=g^{\mu\lambda}T_{\lambda\nu}=\mathrm{diag}(\rho c^2,-P,-P,-P) \tag{12.5.48}$$

---

[58] The reader must notice here that we take the special case of a flat Minkowski metric because otherwise we have no conservation of energy! That's why physically in General Relativity, asymptotically flat space-times represent isolated systems

we have then (keep in mind that one of the stress-energy tensor index is lowered thanks to the metric hence the change of signs below after the second equality):

$$\begin{aligned}\nabla_\mu T^\mu{}_0 &= \partial_\mu T^\mu{}_0 + \Gamma^\mu_{\mu 0} T^0{}_0 - \Gamma^\lambda_{\mu 0} T^\mu{}_\lambda = \partial_t(\rho c^2) + 3\frac{\dot a}{ca}\rho c^2 - \left(-3\frac{\dot a}{ca}\right) P \\ &= \frac{\partial \rho c^2}{\partial(ct)} + 3\frac{\dot a}{ca}\left(\rho c^2 + P\right) = \dot\rho c + 3\frac{\dot a}{a} c\left(\rho+\frac{P}{c^2}\right) = 0\end{aligned} \qquad (12.5.49)$$

Let's finally simplify the last equality by dividing both sides by c to fall back on:

$$\dot\rho + 3\frac{\dot a}{a}\left(\rho+\frac{P}{c^2}\right) = 0 \qquad (12.5.50)$$

Now let us take again the first Friedmann equation, obtained earlier above, in the form:

$$\left(\frac{\dot R}{R}\right)^2 = \frac{8\pi G\rho}{3} - \frac{k}{R^2} \qquad (12.5.51)$$

and let us write it as follows:

$$\dot R^2 = \frac{8\pi G\rho r^2}{3} - k \qquad (12.5.52)$$

If we differentiate:

$$\frac{\mathrm{d}}{\mathrm{d}t}(\dot R^2) = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{8\pi G\rho r^2}{3} - k\right) \qquad (12.5.53)$$

Then we get:

$$2\dot R\ddot R = \frac{8\pi G}{3}\frac{\mathrm{d}}{\mathrm{d}t}(\rho R^2) = \frac{8\pi G}{3}\left(R^2\frac{\mathrm{d}\rho}{\mathrm{d}t} + \rho\frac{\mathrm{d}R^2}{\mathrm{d}t}\right) = \frac{8\pi G}{3}(R^2\dot\rho + 2\rho R\dot R) \qquad (12.5.54)$$

Let us inject:

$$\dot\rho = -3\left(\rho+\frac{P}{c^2}\right)\frac{\dot R}{R} \qquad (12.5.55)$$

in the relation:

$$2\dot R\ddot R = \frac{8\pi G}{3}(R^2\dot\rho + 2\rho R\dot R) \qquad (12.5.56)$$

Then we get:

$$2\dot R\ddot R = \frac{8\pi G}{3}\left(-3(\rho+\frac{P}{c^2}\frac{\dot R}{R}R^2 + 2\rho R\dot R\right) = \frac{8\pi G}{3}R\dot R\left(-3\left(\rho+\frac{P}{c^2}\right)+2\rho\right) \qquad (12.5.57)$$

Thus:

$$\frac{\ddot R}{R} = \frac{4\pi G}{3} = \left(-3\rho - 3\frac{P}{c^2} + 2\rho\right) = -\frac{4\pi G}{3}\left(\rho + 3\frac{P}{c^2}\right) \qquad (12.5.58)$$

The following relation:

$$\boxed{\frac{\ddot R}{R} = -\frac{4\pi G}{3}\left(\rho + 3\frac{P}{c^2}\right)} \qquad (12.5.59)$$

is the "Friedmann's second equation" that is also sometimes named "Raychaudhuri equation".

In general, in the field of cosmogony, any relation of the type $H^2 = (\dot R/R)^2 = \ldots$ is named "velocity equation" and any relation of the type $\ddot R/R = \ldots$ is named "acceleration equation".

Remark

Notice that when $\rho > 0$ and $P > 0$ we have that $\ddot{R} < 0$. This was the reason that had led Einstein to introduce his "cosmological term". Indeed, we see that otherwise we must have $\rho < -3P/c^2$ so that the right term remains positive...

12.5.2.1 Critical Density and Density parameters

Let us come back to our first Friedmann equation without cosmological constant. So we have shown above that:

$$H = \frac{1}{F}\mathrm{d}_t F \tag{12.5.60}$$

We obtained then by injecting the latter relation in the first Friedmann equation the following relation:

$$(\mathrm{d}_t F)^2 = \frac{8\pi G}{3}\rho F^2 - k \tag{12.5.61}$$

which rearranges with:

$$\rho = \rho_0 F^{-3} \tag{12.5.62}$$

into:

$$(\mathrm{d}_t F)^2 = \frac{8\pi}{3} G\rho_0 F^{-1} - k \tag{12.5.63}$$

The exponent of the left term requires that the right term is positive or zero as:

$$\frac{8\pi}{3} G\rho_0 F^{-1} \geq k \tag{12.5.64}$$

Let us recall that the initial conditions impose us that at time $t_0 = 0$ we have:

$$F(t_0) = 1 \quad \text{and} \quad \mathrm{d}_t F(t)|_{t=t_0} = H(t_0) \tag{12.5.65}$$

Indeed:

$$H(t_0) = \frac{1}{F(t_0)}\mathrm{d}_t F|_0 = \mathrm{d}_t F|_0 \tag{12.5.66}$$

Then it comes:

$$k = \frac{8\pi}{3} G\rho_0 - H_0^2 \tag{12.5.67}$$

This term should be accessible to observation, sadly H_0^2 is very poorly known and ρ_0 even more. In other words, given the "$-$" sign in the expression of k, we do not even know today the sign of this constant.

However, it may be important to notice that there is a value ρ_0 named "critical density" that cancels the k above and therefore also (see above):

$$\frac{-2k_1}{mr_0^2} \tag{12.5.68}$$

as it is also equal to k. This imply that the total energy of the Universe would be zero (following considerations of quantum cosmology).

This value ρ_0 is given immediately by:

$$\boxed{\rho_0 = \frac{3H_0^2}{8\pi G} = \rho_c} \qquad (12.5.69)$$

For $H_0 = 67.80 \pm 0.7\,[\text{km} \cdot \text{s}^{-1} \cdot \text{Mpc}^{-1}]$ we get:

$$\rho_0 \cong 8.66 \cdot 10^{-27}\,[\text{kg} \cdot \text{m}^{-3}] \qquad (12.5.70)$$

In comparison, a hydrogen atom weighs $1.7 \cdot 10^{-27}$ [kg], the critical density would therefore correspond to six hydrogen atoms per cubic meter.

Physicists have defined a constant (time-varying... so therefore not so constant...) denoted by the Greek letter Ω and named "cosmological density parameter" and given by the ratio of the mass density (or energy densities since the ratio will be the same!):

$$\Omega = \frac{\rho}{\rho_0} \qquad (12.5.71)$$

Astrophysicists often break the cosmological density parameter into three terms (notice the lower case indexes!):

$$\Omega = \Omega_{\text{matter}} + \Omega_{\text{radiation}} + \Omega_{\text{vacuum}} = \Omega_m + \Omega_r + \Omega_\Lambda \qquad (12.5.72)$$

estimated experimentally in this early 121st century (holocene calendar) at $\Omega = 1.003 \pm 0.004$. We can even found decomposition in more terms (radiation, relativistic matter, non-relativistic matter, dark energy, dark matter[59] and vacuum, etc.) like illustrated in the table below (the year is given according to the holocene calendar):

Species		**Hinshaw et al. (12012)**
Dark energy (Λ)	Ω_Λ	0.72 ± 0.01
Non-baryonic cold dark matter (CDM)	Ω_c	0.24 ± 0.01
Baryonic matter	Ω_b	0.047 ± 0.002
Neutrinos	Ω_ν	< 0.02
Photons (CMB)	Ω_γ	$5 \cdot 10^{-5}$
Total	Ω	1.003 ± 0.004
Curvature	Ω_k	-0.003 ± 0.004

Table 12.23: **Cosmic inventory**

Remark

When combined with other astrophysical data, particularly geometrical measurements of baryon acoustic oscillations, the Universe is constrained to be spatially flat (i.e. hence infinite in size) with $\Omega_k = 0.0004 \pm 0.0018$ - with a precision of around 2σ (95%) at the day we write these lines therefore not enough to conclude anything beyond reasonable doubt - in agreement with the 12018 results (holocene calendar) of the Planck team!

---

[59] Since June 12025 (holocene calendar) it is known thanks to FRB (Fast Radio Bursts) that approximately 76% of the Universe's baryonic matter lies in the IGM (InterGalactic Medium). About 15% resides in galaxy halos, and a small fraction is burrowed in stars or amid cold galactic galactic gas ([109]).

If one think the universe is Euclidean, the number of possible shapes shrinks dramatically. In 11934 (holocene calendar), Werner Nowacki proved that there are only 18 possible Euclidean 3-manifolds (see [420]). W. Hantschze and H. Wendt also published (see [228]) a more direct classification in 11935 (holocene calendar). Instead of an infinite list of 3-manifolds, we need only consider 18 possibilities for the spatial universe. Understanding the properties and appearances of these manifolds may supply the information needed to determine experimentally the universe's shape.

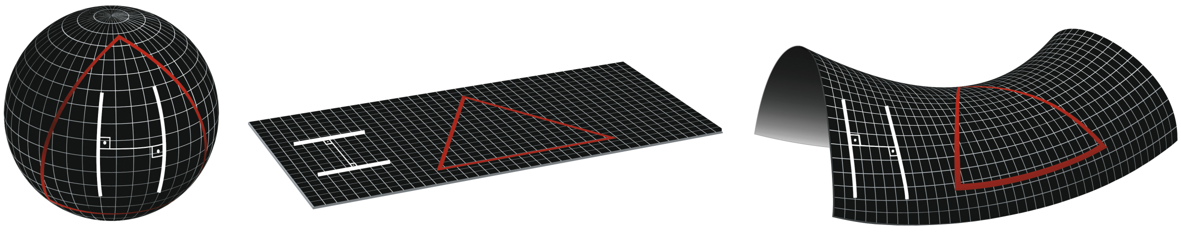
It is interesting to work with this constant Ω because in the case where:

- $\Omega = 1$ (i.e. $\rho = \rho_0$)
  We have therefore:
  $$\rho_0 = \rho_c = \frac{3H_0^2}{8\pi G} \tag{12.5.73}$$
  which gives by replacing that in the Friedmann equation above, the value $k = 0$ (flat and infinite Universe as we will see in our study of the relativistic model later below).

- $\Omega > 1$ (i.e. $\rho_0 > \rho_c$)
  By performing the same reasoning, and still using inequality, we have then: $k > 0$ (a Universe with positive curvature (closed) as we shall see in our study of the relativistic model).

- $\Omega < 1$ (i.e. $\rho_0 < \rho_c$)
  By performing the same reasoning, and still using inequality, we have then: $k < 0$ (a Universe with negative curvature (open) as we shall see in our study of the relativistic model).

These three situations can be summarized geometrically by the following well known figure:



Positive curvature
(spherical like/elliptic)
Closed (finite) Universe
$\Omega > 1, \rho_0 > \rho_c$
Big Crunch end

Flat (zero) curvature
(plane like/euclidian)
Flat (infinite) Universe
$\Omega = 1, \rho_0 = \rho_c$

Negative curvature
(horse's saddle like/hyperbolic)
Opened (infinite) Universe
$\Omega < 1, \rho_0 < \rho_c$
Cold Death end

Figure 12.256: **Illustration of the different types of universes (source: NASA / WMAP Science Team)**

Advanced theoretical physics seems to be able to prove (see [169]) that our Universe can have only eighteen possible topologies given by:

Symbol	**Name**	**Compact Dimension**	**Orientable**	**Homogeneous**	**Isotropic**
E_1	3-torus (flat and finite)	3	Yes	Yes	No
E_2	Half-turn	3	Yes	No	No
E_3	Quarter-turn	3	Yes	No	No
E_4	Third-turn	3	Yes	No	No
E_5	Sixth-turn	3	Yes	No	No
E_6	Hantzsch-Wendth	3	Yes	No	No
E_7	Klein space	3	No	No	No
E_8	- (horizontal flip)	3	No	No	No
E_9	- (vertical flip)	3	No	No	No
E_{10}	- (half-turn)	3	No	No	No
E_{11}	Chimney space	2	Yes	Yes	No
E_{12}	- (half turn)	2	Yes	No	No
E_{13}	- (vertical flip)	2	No	No	No
E_{14}	- (horizontal flip)	2	No	No	No
E_{15}	- (half turn + flip)	2	No	No	No
$E_{16}^{(h)}$	Slab (unrotated)	1	Yes	Yes	No
$E_{16}^{(i)}$	Slab (rotated)	1	Yes	No	No
E_{18}	Covering space	0	Yes	Yes	Yes

Table 12.24: **Possible topologies of our Universe (cosmic topologies)**

All measurement which have been made so far have failed to show a curvature of the Universe (anisotropy). The measurements of the microwave background (see mathematical details further below) by the BOOMERANG balloon and COBE satellite however, tend to support the hypothesis of a relatively flat Universe validating therefore numerical simulations[60]:

---

[60] We will show in the future that the mathematics of the CMB involve spherical harmonics and Fourier transform on a sphere.

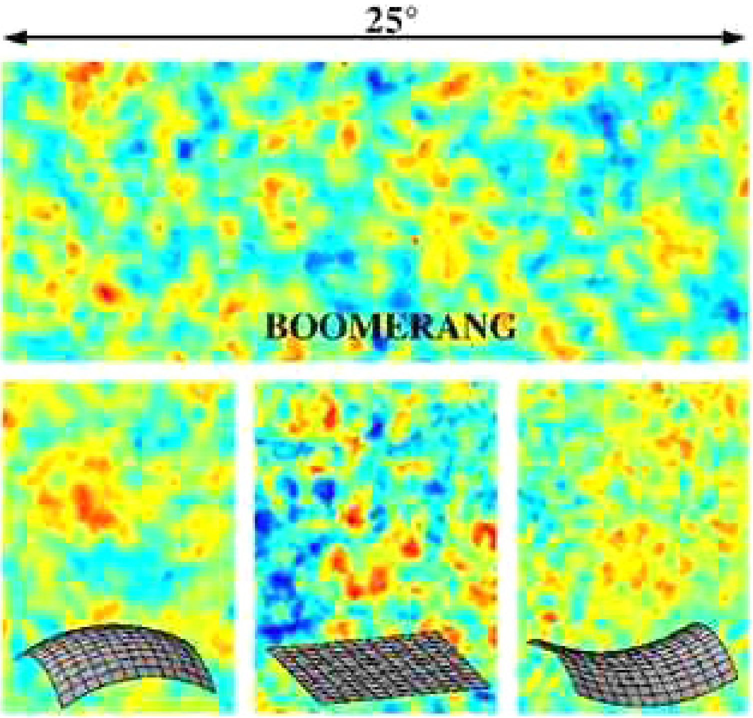


Figure 12.257: **Illustration of what would give observations depending on the curvature type**

As we can see in the figure below satellites sensibility increase and permits to validate or invalidate speculative models with more and more confidence:

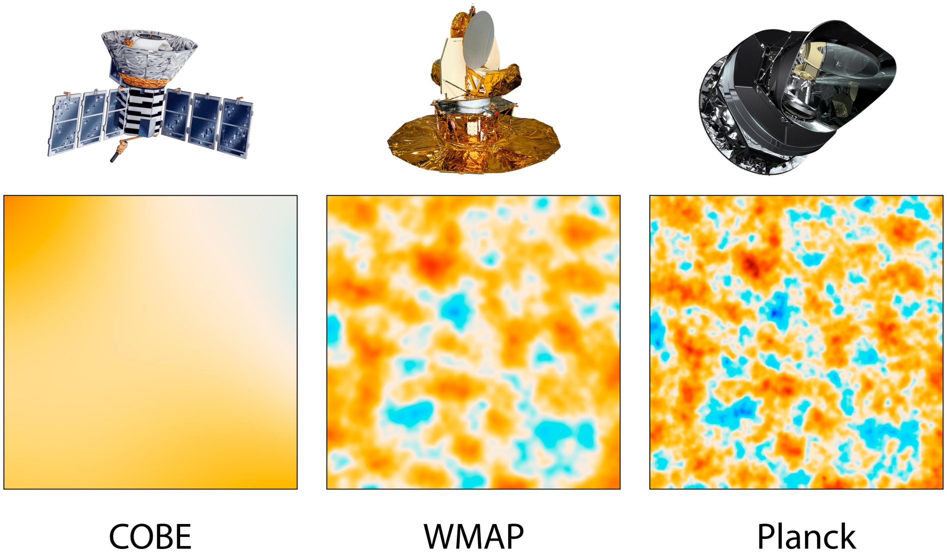


Figure 12.258: **Comparison of CMB results from COBE, WMAP and Planck – 12013-03-21 (source: Wikipedia)**

The Planck CMB seems to give evidence of fluctuations on the order of 1 part in $100,000$ across the sky (root mean square variations are only $18\,[\mu\text{K}]$). So according to some cosmologist this may be an evidence that the Universe was quite homogeneous at its beginning (hence the use of the Friedmann-Robertson-Walker-Lemaître metric - among others - in many Quantum Cosmology models).

And an another interesting figure that shows our Universe density fluctuations at different scales:

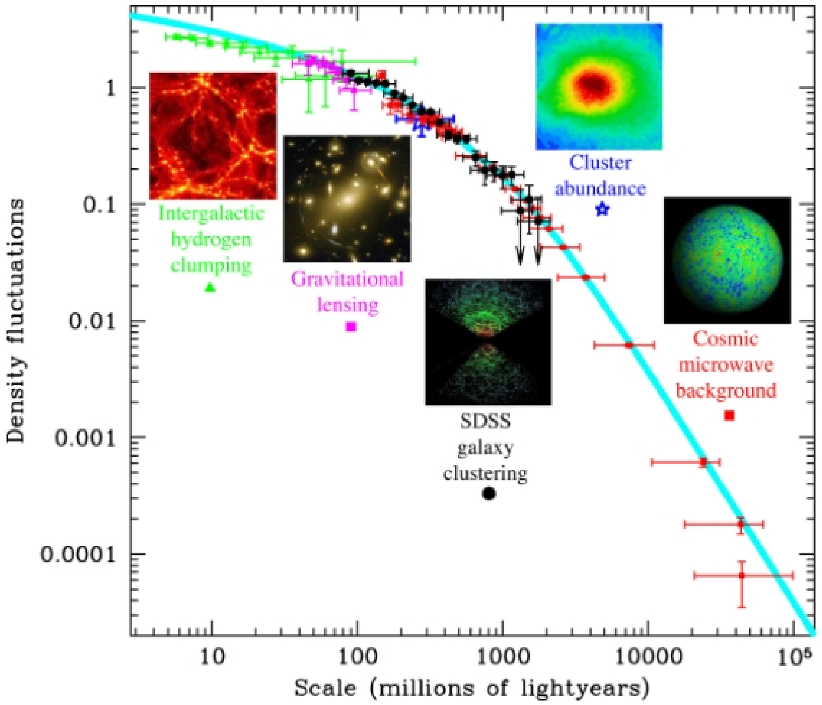


Figure 12.259: **Universe fluctuation density scales (source: ?)**

Now remember that we have proved earlier that:

$$H^2 = \frac{8\pi G}{3}\rho - kF^{-2} \tag{12.5.74}$$

Rearranging:

$$1 = \frac{8\pi G}{3H^2}\rho - \frac{k}{(FH)^2} \tag{12.5.75}$$

Then according to what we have defined just above the relation can be rewritten:

$$1 = \Omega_m + \Omega_k \tag{12.5.76}$$

where:

$$\Omega_m := \frac{8\pi G\rho}{3H^2} \qquad \Omega_k := -\frac{k}{(FH)^2} \tag{12.5.77}$$

That are other well known density parameters in cosmology!

Remark

The concept of Universe topology and openness are actually normally two separate concepts. When we speak of "open" or "closed" Universe we do not normally talk about its topology but its destiny! Thus, an "open" Universe is expanding indefinitely and a "closed" Universe recontracts on itself after a given time. That said, in the models that we study in this section (cosmological constant equal to zero), the curvature is directly related to the density, and thus to its openness.

Let us come back to the relation:

$$(\mathrm{d}_t F)^2 = \frac{8\pi}{3} G\rho_0 F^{-1} - k \tag{12.5.78}$$

We can write:

$$\mathrm{d}_t F(t) = \sqrt{\frac{A}{F} - k} \tag{12.5.79}$$

By adopting the notation:

$$A = \frac{8\pi}{3} G\rho_0 \tag{12.5.80}$$

Remark

The actual measurement give (year 12007 according to holocene calendar): $A \cong 0.2793923067 \cdot 10^{-35}$.

Therefore:

$$\mathrm{d}F(t) = \pm\sqrt{\frac{A}{F} - k}\mathrm{d}t \tag{12.5.81}$$

It is now appropriate for us to consider three situations:

$$k = 0 \qquad k < 0 \qquad k > 0 \tag{12.5.82}$$

thus corresponing respectively to the following cosmological density parameters:

$$\Omega = 1 \qquad \Omega < 0 \qquad \Omega > 0 \tag{12.5.83}$$

i.e. positive curvature: $k = +1, \Omega_k < 0, \Omega > 1$, negative curvature: $k = -1, \Omega_k > 0$, $\Omega < 1$, zero curvature: $k = 0, \Omega_k = 0, \Omega = 1$).

Remark

We can't put $\rho_0 = 0$ because our initial assumption was the principle of conservation of energy.

Let us now introduce the same cosmological density parameters but with a slightly different approach (it's better for the reader the see the both approach). Again, let us consider the first Friedmann equation considering the following notation:

$$\left(\frac{\dot{a}}{a}\right)^2 + \frac{k}{a^2} = \frac{8\pi}{3} G\rho \tag{12.5.84}$$

We recall that:

$$H(t) := \frac{\dot{a}}{a} \tag{12.5.85}$$

and that we can formally associate an energy density with the cosmological constant:

$$\rho_\Lambda := \frac{\Lambda}{8\pi G} \tag{12.5.86}$$

With this notation we can re-write Friedmann equation:

$$H^2 + \frac{k}{a^2} = \frac{8\pi G}{3}\left(\sum_i \rho_i + \rho_\Lambda\right) \tag{12.5.87}$$

If the geometry of the universe is flat, $k=0$, and therefore:

$$\rho_{\text{tot}} := \sum_i \rho_i + \rho_\Lambda = \frac{3H^2}{8\pi G} \equiv \rho_{\text{crit}} \tag{12.5.88}$$

which we can use to define the fraction of the critical density contributed by each component of the Universe:

$$\Omega_i := \frac{\rho_i}{\rho_{\text{crit}}} \tag{12.5.89}$$

Thanks to this definition:

$$H^2 + \frac{k}{a^2} = \frac{8\pi G}{3}\left(\sum_i \rho_i + \rho_\Lambda\right) \tag{12.5.90}$$

Can be rewritten as:

$$\frac{k}{a^2H^2} = \sum_i \Omega_i + \Omega_\Lambda - 1 \tag{12.5.91}$$

and we can do even better by defining:

$$\Omega_k := -\frac{k}{(aH)^2} \tag{12.5.92}$$

which then leads to the elegant result:

$$\sum_i \Omega_i + \Omega_\Lambda + \Omega_k = 1 \tag{12.5.93}$$

12.5.3 Friedmann-Lemaître Euclidean Cosmological Models

The Euclidean cosmological models of Friedmann-Lemaître consist in the Newtonian limit to study the "fundamental equation of Friedmann models":

$$\boxed{\mathrm{d}F(t)=\pm\sqrt{\frac{A}{F}-k}\mathrm{d}t} \qquad (12.5.94)$$

considering the three situations:

$$k=0 \qquad k<0 \qquad k>0 \qquad (12.5.95)$$

Remark

It is possible within the framework of General Relativity to rigorously find a solution to the Einstein's field equations named the "Friedmann-Lemaître-Robertson-Walker metric" which in the case of a Newtonian approximation gives us the Friedmann equations obtained in the present text (often these are the approximations that are used in the literature because the exact solution is out of the framework of the traditional universities courses of the 121st century according to the holocene calendar). For the general treatment see further below page 5987.

12.5.3.1 Flat spaces ($k=0$)

The flat (Euclidean) space model consist to assume that $k=0$. In other words, we are in a Universe whose density is named "critical density" or also simply "flat" (as we will see with the relativistic model).

Then we have the following equation:

$$\mathrm{d}F(t)=\pm\sqrt{\frac{A}{F}-k}\mathrm{d}t=\pm\sqrt{\frac{A}{F}}\mathrm{d}t \qquad (12.5.96)$$

By arranging appropriately the terms:

$$\sqrt{F}\mathrm{d}F(t)=\pm\sqrt{A}\mathrm{d}t \qquad (12.5.97)$$

and by integrating, it comes (there is obviously a constant of integration that we will add at the end!):

$$\int\sqrt{F}\mathrm{d}F(t)=\pm\sqrt{A}t=\pm\sqrt{\frac{8\pi G\rho_0}{3}}t \qquad (12.5.98)$$

Which simplifies in (we raise to the square, hence the removal of the double sign ±):

$$F^3=6\pi rG\rho_0t^2 \qquad (12.5.99)$$

So we have in this model the relation:

$$F=\sqrt[3]{6\pi rG\rho_0}t^{2/3} \qquad (12.5.100)$$

to which we must add now the constant of integration for the condition corresponding to today:

$$F(t_0)=F(0)=1 \tag{12.5.101}$$

that remains satisfied. Therefore:

$$F=\sqrt[3]{6\pi r G\rho_0}t^{2/3}+1 \tag{12.5.102}$$

This gives us a plot function that looks like this (do not trust the values shown on the horizontal axis as they are arbitrary):

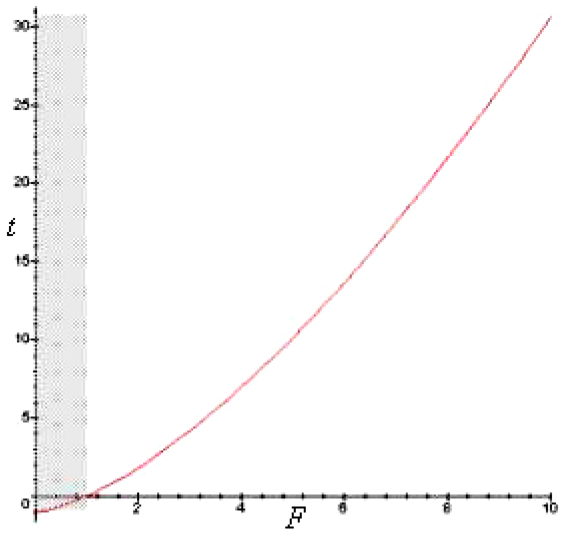


Figure 12.260: **Evolution of the scale factor for a zero curvature space**

We put the area where $F(t)<1$ in evidence to remember that this part of the solution is to reject.

So we have a model of Universe in which the scale factor is growing exponentially and this indefinitely.

Notice also that in this case:

$$\lim_{t\to 0}F(t)=1 \tag{12.5.103}$$

That means with this model that the Universe had already a non-zero size at the beginning of time. This is a good point as we have already presented at the beginning that we should have $F(t)\in\mathbb{R}^+\backslash\{0\}$.

Remark

More ρ_0 is big, more the scale factor increase fast (meaning that the slope is obviously larger).

12.5.3.1.1 Flat space dominated by matter

There is also another approach much more elegant and subtle from my point of view than the previous proof (I have discovered that many years after writing the previous version). It has the benefit of highlighting a hypothesis that does not appear with previous developments.

We start from the first Friedmann equation:

$$\left(\frac{R'}{R}\right)^2=\frac{8\pi G\rho}{3}-\frac{k}{R^2} \tag{12.5.104}$$

Putting k as being equal to zero and using the trick that consist to start from the relation proved also earlier above (used to proved the second Friedmann equation):

$$\dot{\rho}=-3\left(\rho+\frac{P}{c^2}\right)\frac{\dot{R}}{R} \tag{12.5.105}$$

to require that the fluid pressure P (whatever it is: gas or radiation) is zero. We then say that the Universe is a universe dominated by matter and we deduce:

$$\dot{\rho}+3\rho\frac{R'}{R}=0 \tag{12.5.106}$$

which gives us (in this situation ρ is often denoted ρ_{mat} in many textbooks):

$$R^3\dot{\rho}+3\rho R^2R'=0\Rightarrow\frac{\mathrm{d}}{\mathrm{d}t}(R^3\rho)=0 \tag{12.5.107}$$

Hence the important relation for matter dominated Universe:

$$\boxed{R^3\rho=R_0^3\rho_0=c^{te}} \tag{12.5.108}$$

Under these conditions, the first Friedmann equation becomes:

$$\frac{\mathrm{d}R}{R}=\sqrt{\frac{8\pi G\left(\rho_0\dfrac{R_0^3}{R^3}\right)}{3}}\mathrm{d}t \tag{12.5.109}$$

rearranging and simplifying, we then have:

$$\int R^{1/2}\mathrm{d}R=\int\sqrt{\frac{8\pi G\rho_0R_0^3}{3}}\mathrm{d}t \tag{12.5.110}$$

which gives:

$$\frac{2}{3}R^{3/2}+c^{te}=\sqrt{\frac{8\pi G\rho_0R_0^3}{3}}t \tag{12.5.111}$$

At the time $t=0$ of the "Big Bang", many textbooks assume the scale factor is equal to $R=0$ (only in the purpose to simplify the notations!!!) and thus the constant is zero (don't forget that in reality we should rather assume that it is equal to $-3/2$ instead to have after rearranging a $+1$ on the right side of the equality when $t=0$). Then we have:

$$R(t)=(6\pi G\rho_0R_0^3)^{1/3}t^{2/3}=\left(6\pi G\frac{3H_0^2}{8\pi G}R_0^3\right)^{1/3}t^{2/3}=\left(\frac{9H_0^2}{4}R_0^3\right)^{1/3}t^{2/3} \tag{12.5.112}$$

By putting that at time $t=0$, the scaling factor R_0 was unitary, the latter relation simplifies to:

$$R(t)=\left(\frac{9H_0^2}{4}\right)^{1/3}t^{2/3}=\left(\frac{9}{4}\right)^{1/3}(H_0t)^{2/3} \tag{12.5.113}$$

Using the common notation in theoretical cosmology we get the famous relation often given without proof:

$$\boxed{a(t)\propto t^{2/3}} \tag{12.5.114}$$

Notice also that in this case, with this simplified notation (that doesn't take into account the integration constant), we have the misleading limit:

$$\lim_{t\to 0}a(t)=0 \tag{12.5.115}$$

But in fact, if we take into consideration the constant, we have indeed:

$$\lim_{t\to 0}a(t)=1 \tag{12.5.116}$$

that means (for recall) that at beginning of the time, the Universe didn't have zero dimensions (otherwise applying a non-null scale factor on a zero value wouldn't have any effect), but had already a given unknown size.

If we assume that the scale factor is today taken as unitary, then it comes:

$$t=\frac{2}{3H_0} \tag{12.5.117}$$

and by replacing in it the numerical values currently known of the Hubble parameter, it follows that the universe is now aged about 8.6 billion years (compared to 13 billion of the Hubble time obtained earlier above!).

12.5.3.1.2 Flat space dominated by radiation

We have proved in the section of Thermodynamics during our study of the Stefan-Boltzmann law (see page 3710) that the pressure of radiation was related to the energy density by the following equation:

$$P=\frac{1}{3}\rho_E \tag{12.5.118}$$

Remark

In many textbooks the authors define:

$$P=\frac{\gamma}{3}\quad\text{with: }\gamma=\begin{cases}1 & \text{radiation}\\ 0 & \text{matter (dust)}\end{cases} \tag{12.5.119}$$

In a universe dominated by radiation, the relation:

$$\dot{\rho}=-3\left(\rho+\frac{P}{c^2}\right)\frac{\dot{R}}{R} \tag{12.5.120}$$

proved earlier, expressed with a density of energy and not a density of mass becomes:

$$\dot{\rho}c^2=-3\left(\rho c^2+\frac{Pc^2}{c^2}\right)\frac{\dot{R}}{R} \tag{12.5.121}$$

Therefore:

$$\dot{\rho}_E=-3\left(\rho_E+P\right)\frac{\dot{R}}{R} \tag{12.5.122}$$

and using the relation between radiation pressure and energy density, we get:

$$\dot{\rho}_E=-3\left(\rho_E+\frac{\rho_E}{3}\right)\frac{\dot{R}}{R} \tag{12.5.123}$$

After a little rearrangement, we get:

$$\dot{\rho}_E+4\rho_E\frac{\dot{R}}{R}=0 \tag{12.5.124}$$

from which we get that:

$$\dot{\rho}_E+4\rho_E\frac{\dot{R}}{R}=0\Rightarrow R^4\dot{\rho}_E+4\rho_E\dot{R}R^3=0\Rightarrow\frac{\mathrm{d}}{\mathrm{d}t}\left(R^4\rho_E\right)=0\Rightarrow R^4\rho_E=R_0^4\rho_{E,0}=c^{te} \tag{12.5.125}$$

Hence the important relation for radiation dominated Universe:

$$\boxed{R^4\rho_E=R_0^4\rho_{E,0}=c^{te}} \tag{12.5.126}$$

Remark

Using the previous results notice that we have so far:

$$R^n\rho=R_0^n\rho_0\Rightarrow\begin{cases}n=0\text{ for vacuum}\\ n=3\text{ for matter dominated Universe}\\ n=4\text{ for radiation dominated Universe}\end{cases} \tag{12.5.127}$$

Under these conditions, the first Friedmann equation:

$$\left(\frac{R'}{R}\right)^2=\frac{8\pi G\rho}{3}-\frac{k}{R^2} \tag{12.5.128}$$

becomes first by changing into energy density and with k being put as equal to zero:

$$\left(\frac{\dot{R}}{R}\right)^2=\frac{8\pi G\rho_E}{3c^2} \tag{12.5.129}$$

and so we can replace the energy density by the result we get just before:

$$\left(\frac{\dot{R}}{R}\right)^2=\frac{8\pi G\rho_{E,0}}{3c^2}\left(\frac{R_0}{R}\right)^4 \tag{12.5.130}$$

Remark

The prior-previous relation is often denoted:

$$3H^2M_{\text{P}}^2=\rho \quad (12.5.131)$$

where:

$$M_{\text{P}}=\frac{1}{\sqrt{8\pi G}} \quad (12.5.132)$$

is sadly (because it hasn't the units of a mass and it's not related to the different famous Planck values) named the "reduced Planck mass".

Then we have:

$$\frac{\mathrm{d}R}{\mathrm{d}t}=\sqrt{\frac{8\pi G\rho_{E,0}R_0^4}{3c^2}}\frac{1}{R} \quad (12.5.133)$$

hence:

$$R\mathrm{d}R=\sqrt{\frac{8\pi G\rho_{E,0}R_0^4}{3c^2}}\mathrm{d}t \quad (12.5.134)$$

The primitive is obvious and immediate:

$$\frac{R^2}{2}+c^{te}=\sqrt{\frac{8\pi G\rho_{E,0}R_0^4}{3c^2}}t \quad (12.5.135)$$

At the time $t=0$ of the "Big Bang", many textbooks assume the scale factor is equal to $R=0$ (only in the purpose to simplify the notations!!!) and thus the constant is zero (don't forget that in reality we should rather assume that it as equal to -2 instead to have after rearranging a $+1$ on the right side of the equality when $t=0$). Then we have:

$$\frac{R^2}{2}=\sqrt{\frac{8\pi G\rho_{E,0}R_0^4}{3c^2}}t \quad (12.5.136)$$

But assuming that at $t=0$ we had $R_0=1$ then we have:

$$\frac{R^2}{2}=\sqrt{\frac{8\pi G\rho_{E,0}}{3c^2}}t \quad (12.5.137)$$

Therefore:

$$R=\left(\frac{32\pi G\rho_{E,0}}{3c^2}\right)^{1/4}t^{1/2}=\left(\frac{32\pi G}{3}\rho_0\right)^{1/4}t^{1/2}=\left(\frac{32\pi G}{3}\frac{3H_0^2}{8\pi G}\right)^{1/4}t^{1/2} \quad (12.5.138)$$

Thus after simplification it remains only:

$$R(t)=(4H_0)^{1/2}t^{1/2} \quad (12.5.139)$$

Using the common notation in theoretical cosmology we get the famous relation often given without proof:

$$\boxed{a(t)\propto t^{1/2}} \quad (12.5.140)$$

Then a flat universe dominated by radiation has a scaling factor that is growing slightly more slowly than a flat universe dominated by matter.

Notice also that in this case:

$$\lim_{t\to 0} a(t)=0 \tag{12.5.141}$$

But in fact, if we take into consideration the integration constant, we have indeed:

$$\lim_{t\to 0} a(t)=1 \tag{12.5.142}$$

that means (for recall) that at beginning of the time, the Universe didn't have zero dimensions (otherwise applying a non-null scale factor on a zero value wouldn't have any effect), but had already a given unknown size.

In comparison with Maple™ 4.00b (blue: flat universe dominated by matter, in red: flat universe dominated by radiation):

```
>plot([t^(2/3),t^(1/2)],t=0..2*Pi,0..3,color=[blue,red]);
```

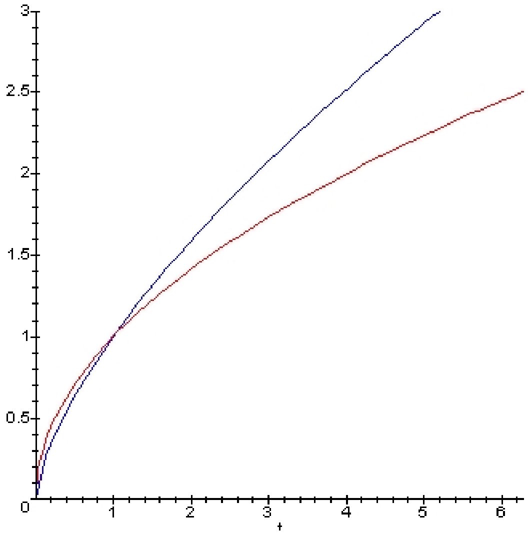


Figure 12.261: Evolution of R for a zero curvature space dominated by matter or radiation

12.5.3.1.2.1 Standard form of the Friedmann equation

We have seen that for a universe dominated by matter we had:

$$\dot{\rho}=-3\left(\rho+\frac{P}{c^2}\right)\frac{\dot{R}}{R} \tag{12.5.143}$$

that reduces to:

$$\dot{\rho}+3\rho\frac{\dot{R}}{R}=0 \tag{12.5.144}$$

and for a universe dominated by radiation:

$$\dot{\rho}_E+4\rho\frac{\dot{R}}{R}=0 \tag{12.5.145}$$

So we see that the two previous relations are respectively a special case of the first one that if we rewrite as:

$$\dot{\rho}=-3\left(\rho+\frac{\omega(T)\rho c^2}{c^2}\right)\frac{\dot{R}}{R}\Leftrightarrow\frac{\dot{\rho}}{\rho}=-3\left(1+\omega(T)\right)\frac{\dot{R}}{R}\Rightarrow \mathrm{d}\ln(\rho)=-3(1+\omega(T))\mathrm{d}\ln(R) \tag{12.5.146}$$

with $P=\omega(T)\rho c^2$, that is to say a barotropic fluid assumption (linear relation between ρ and P). We see that by introducing $\omega(T)$, named the "density parameter", we fall back on the first radiation one when $\omega(T)=1/3$ and collisionless matter when $\omega(T)=0$ and vacuum energy when $\omega(T)=-1$.

We also have by integrating the last relation above :

$$\int \mathrm{d}\ln(\rho)=-3(1+\omega(T))\int \mathrm{d}\ln(R)\Rightarrow\ln(\rho)=-3(1+\omega(T))\ln(R) \tag{12.5.147}$$

Hence (using elementary natural logarithm rules and properties):

$$\ln(\rho)=\ln\left(R^{-3(1+\omega(T))}\right) \tag{12.5.148}$$

And taking the exponential on both sides we have finally:

$$\rho(t)\propto R(t)^{-3(1+\omega(T))}\propto\begin{cases}R(t)^{-3} & \text{for } \omega(T)=0 \text{ (pressureless matter (i.e. dust))}\\ R(t)^{-4} & \text{for } \omega(T)=1/3 \text{ (radiation/ultra-relativistic matter)}\\ c^{te} & \text{for } \omega(T)=-1 \text{ (vacuum energy)}\end{cases} \tag{12.5.149}$$

Notice that we can often found in various textbooks and even in the present book, the following notation for the energy density (...):

$$\rho=A_\omega a^{-3(\omega+1)}\quad\text{or}\quad\rho_{\omega,0}=\rho_\omega\left(\frac{R}{R_0}\right)^{-3(\omega+1)} \tag{12.5.150}$$

with $\rho_{\omega,0}$ being the value of energy density at present time.

To be more exhaustive we have to summarize:

$P=\omega\rho$ (equation of state)		
ρ is diluated with expansion as $a^{-3(1+\omega)}$		
$\omega=0$	ρ dilutes as (because universe is made of matter): (this is the most intuitive case if you consider a cube of side L, if the side is multiplied by 2 the volume increases of a factor $2^3=8$ hence the energy density of matter is divided by 8)	a^3
$\omega=1/3$	ρ dilutes as (because universe is made of radiation/photons): (again considering the same cube but made this time of radiation if length is multiplied by 2 then the volume also increases of a factor $2^3=8$ but as the photon wavelength is also divided by 2 at the same time then the energy density of radiation is divided by $2\cdot 2^3=2^4$)	a^4
$\omega=-1/3$	ρ (the density energy) decrease but slower: (that means that density of object still decrease but is somehow compensated by an increase of the energy these same objects, i.e. the universe expansion accelerated slowly and indefinitely)	a^2
$\omega=1$	ρ (the density energy) remains constant: (that means energy is extracted from vacuum and we know this isn't an issue as the law of conservation of energy doesn't apply when space-time change)	0
$\omega<-1$	ρ (the density energy - named "phantom energy") increase: (that means that density of energy of vacuum strictly increase very fast and the universe will be ripped of in a finite time)	$a^{-\lvert c^{te}\rvert}$
Latest data still contradicts: 12018 Planck gives $\omega=-1.03\pm 0.03$ and 12023 by KIAS $\omega=-0.903\pm 0.023$		

Table 12.25: **Different scenario of density-pressure relation in General Relativity**

We are now in a position to determine the dynamical evolution of the Universe without cosmological constant. As we have seen, we need to use one of the Friedmann equations, plus the equations of state and continuity equation for each of the fluids in the Universe. Let's start from the first Friedmann equation:

$$\begin{aligned}\left(\frac{\dot{R}}{R}\right)^2+\frac{k}{R^2c^2}&=\frac{8\pi G}{3c^2}\rho\\ H^2+\frac{k}{R^2c^2}&=\frac{8\pi G}{3c^2}\rho\\ H^2\left(1-\frac{8\pi G}{3H^2c^2}\rho\right)&=-\frac{k}{R^2c^2}\end{aligned}\tag{12.5.151}$$

Here above $8\pi G/3H^2$ has the dimensions of an energy density, and is named as we already know it (see page 5897), the "critical density". As we have seen it, it is customary to express the energy densities in units of the critical density:

$$\rho_c(t):=\frac{8\pi G}{3c^2H^2}\qquad \Omega_i(t)=\frac{\rho_i(t)}{\rho_c(t)}\tag{12.5.152}$$

and so, we have:

$$\boxed{H^2(1-\Omega(t))=-\frac{k}{R^2c^2}} \tag{12.5.153}$$

We see that there is a direct connection between the critical energy density and the geometry of the Universe, as the curvature will be respectively positive, zero or negative if $\Omega>1,\Omega=1$ or $\Omega<1$ respectively.

Today, we have:

$$H_0^2\left(1-\Omega_0\right)=-\frac{k}{R_0^2c^2} \tag{12.5.154}$$

so, we have an expression of the curvature as a function of H_0 and the total density today.

Starting from:

$$H^2=\frac{8\pi G}{3c^2}\left(\rho_m+\rho_r+\rho_X\right)-\frac{k}{R^2c^2} \tag{12.5.155}$$

where ρ_X is the energy density of dark matter[61], we can substitute the value of k and the evolutions of the densities of the various fluids obtained just earlier and the Cosmological redshift (see page 5994) to get:

$$\begin{aligned}
H^2&=\frac{8\pi G}{3c^2}\left(\rho_m+\rho_r+\rho_X\right)-\frac{k}{R_0^2c^2}\frac{R_0^2}{R^2}\\
&=\frac{8\pi G}{3c^2}\left(\rho_{m,0}\left(\frac{R_0}{R}\right)^3+\rho_{r,0}\left(\frac{R_0}{R}\right)^4+\rho_X\left(\frac{R_0}{R}\right)^{3(1+\omega_X)}\right)+H_0^2\left(1-\Omega_0\right)\left(\frac{R_0}{R}\right)^2\\
&=\frac{8\pi G}{3c^2}\frac{H_0^2}{H_0^2}\left(\rho_{m,0}\left(\frac{R_0}{R}\right)^3+\rho_{r,0}\left(\frac{R_0}{R}\right)^4+\rho_X\left(\frac{R_0}{R}\right)^{3(1+\omega_X)}\right)+H_0^2\left(1-\Omega_0\right)\left(\frac{R_0}{R}\right)^2\\
&=\frac{8\pi G}{3c^2H_0^2}H_0^2\left(\rho_{m,0}\left(\frac{R_0}{R}\right)^3+\rho_{r,0}\left(\frac{R_0}{R}\right)^4+\rho_X\left(\frac{R_0}{R}\right)^{3(1+\omega_X)}\right)+H_0^2\left(1-\Omega_0\right)\left(\frac{R_0}{R}\right)^2\\
&=\frac{1}{\rho_c}H_0^2\left(\rho_{m,0}\left(\frac{R_0}{R}\right)^3+\rho_{r,0}\left(\frac{R_0}{R}\right)^4+\rho_X\left(\frac{R_0}{R}\right)^{3(1+\omega_X)}\right)+H_0^2\left(1-\Omega_0\right)\left(\frac{R_0}{R}\right)^2\\
&=H_0^2\left(\frac{\rho_{m,0}}{\rho_c}\left(\frac{R_0}{R}\right)^3+\frac{\rho_{r,0}}{\rho_c}\left(\frac{R_0}{R}\right)^4+\frac{\rho_X}{\rho_c}\left(\frac{R_0}{R}\right)^{3(1+\omega_X)}\right)+H_0^2\left(1-\Omega_0\right)\left(\frac{R_0}{R}\right)^2\\
&\cong H_0^2\left(\Omega_{m,0}(1+z)^3+\Omega_{r,0}(1+z)^4+\Omega_{X,0}(1+z)^{3(1+w_X)}+\left(1-\Omega_0\right)(1+z)^2\right)
\end{aligned} \tag{12.5.156}$$

This is the "standard form of the Friedmann equation" use by some authors of textbooks on General Relativity, which connects the evolution of the Hubble parameter $H(z)$ with the densities today. We generally drop the ' 0 ' subscripts, and define $\Omega_k=1-\Omega_0$ (Ω_k is not an actual mass/energy density and is defined as $\Omega_k = 1 - \Omega_0$ so that $\Omega_k = 0$ for a spatially flat universe), which gives:

---

[61] Since June 12025 (holocene calendar) it is known thanks to FRB (Fast Radio Bursts) that approximately 76% of the Universe's baryonic matter lies in the IGM (InterGalactic Medium). About 15% resides in galaxy halos, and a small fraction is burrowed in stars or amid cold galactic galactic gas ([109]).

$$\boxed{H(z) \cong H_0 \left[\Omega_m (1+z)^3 + \Omega_r (1+z)^4 + \Omega_X (1+z)^{3(1+w)} + \Omega_k (1+z)^2\right]^{1/2}} \quad (12.5.157)$$

So we have a prediction for $H(z)$ as a function of the matter, radiation and dark energy densities today, we have a good idea of how these densities evolve with z. And above all, we can use the expression of $H(z)$ above to compute distances, in other terms, we can connect the evolution of cosmic expansion to quantities which can all be observed.

When the energy densities are dominated by one single component, it is quite easy to integrate the Friedmann standard equation and get the evolution of a as a function of cosmic time.

Let's define :

$$x := \frac{a}{a_0} \cong \frac{1}{1+z} \quad (12.5.158)$$

Then, the Friedmann equation can be rewritten:

$$\frac{\mathrm{d}x}{\mathrm{d}t} \cong H_0 \left(\Omega_m / x^3 + \Omega_r / x^4 + \Omega_X / x^{3(1+\omega)} + \Omega_k / x^2\right)^{1/2} \quad (12.5.159)$$

as often the convention $a_0 = 1$ is used we can found the latter relation in the following shape too (the terms are rearranged in ascending negative power):

$$H(a) = \frac{\mathrm{d}x}{\mathrm{d}t} = \dot{a} \cong H_0 \sqrt{\Omega_r a^{-4} + \Omega_m a^{-3} + \Omega_k a^{-2} + \Omega_X a^{-3(1+\omega)}} \quad (12.5.160)$$

The latter is also commonly written (if we don't trop de '0' we have dropped earlier before):

$$H(a) \cong H_0 \sqrt{\Omega_{r,0} a^{-4} + \Omega_{m,0} a^{-3} + \Omega_{k,0} a^{-2} + \Omega_{X,0} a^{-3(1+\omega)}} \quad (12.5.161)$$

That last term inside the parenthesis is often taken as the dark energy! Assuming the latter remains constant (i.e. $\rho_\Lambda = \rho_{\Lambda,0} ...$ and hence we don't take its general form named "quintessence"):

$$H(a) \cong H_0 \sqrt{\Omega_{r,0} a^{-4} + \Omega_{m,0} a^{-3} + \Omega_{k,0} a^{-2} + \Omega_{\Lambda,0}} \quad (12.5.162)$$

The letters are also often capitalized such that:

$$H(a) \cong H_0 \sqrt{\Omega_{R,0} a^{-4} + \Omega_{M,0} a^{-3} + \Omega_{k,0} a^{-2} + \Omega_{\Lambda,0}} \quad (12.5.163)$$

The latest values of the parameters, obtained from the Planck mission, are:

$$\begin{gathered} H_0 = 67.3\,[\,\text{km}\cdot\ \text{s}^{-1}\cdot \text{Mpc}^{-1}] \\ \Omega_{R,0} = 9.24 \cdot 10^{-5}, \quad \Omega_{M,0} = 0.315 \\ \Omega_{\Lambda,0} = 0.685, \quad \Omega_{k,0} = 0 \end{gathered}$$

So now we have the Hubble parameter as a function of the scale radius a. How can we convert this into a function of time? From:

$$\dot{a} = \frac{\mathrm{d}a}{\mathrm{d}t} \quad (12.5.164)$$

we get:

$$\mathrm{d}t=\frac{\mathrm{d}a}{\dot{a}}=\frac{\mathrm{d}a}{aH(a)}=\frac{a\,\mathrm{d}a}{a^2H(a)} \tag{12.5.165}$$

so that:

$$t(a)=\int_0^a\frac{a'\mathrm{d}a'}{a'^2H(a')}=\frac{1}{H_0}\int_0^a\frac{a'\mathrm{d}a'}{\sqrt{\Omega_{R,0}+\Omega_{M,0}a'+\Omega_{k,0}a'^2+\Omega_{\Lambda,0}a'^4}} \tag{12.5.166}$$

Finally we have for the age of our universe:

$$\boxed{t(a)\cong\frac{1}{H_0}\int_0^a\frac{a'\mathrm{d}a'}{\sqrt{\Omega_{R,0}+\Omega_{M,0}a'+\Omega_{k,0}a'^2+\Omega_{\Lambda,0}a'^4}}} \tag{12.5.167}$$

As far as we know this integral cannot be solved formally. We have to use numerical integration. For example, using the following Maple™ 17 commands:

```
H0:=0.218E-17;OmegaR:=0.924E-4;OmegaM:=0.315;OmegaK:=0;OmegaD:=0.865;
evalf((int(a/(H0*sqrt(a^4*OmegaD+a^2*OmegaK+a*OmegaM+OmegaR)), a =
0.1E-6..1))/(24*(60*60)*365.25));
```

we get:

$$t(a)\cong 13.31932\,[\text{billion years}]$$

Now let's come back to:

$$H(a)\cong H_0\sqrt{\Omega_{r,0}a^{-4}+\Omega_{m,0}a^{-3}+\Omega_{k,0}a^{-2}+\Omega_{X,0}a^{-3(1+\omega)}} \tag{12.5.168}$$

In the minimal Lambda-CDM model (there are other models than the Lambda Cold Dark Matter model like the SIDM - Strong Interactive Dark Matter - model...), it is assumed that curvature Ω_k is zero (flat universe) and $w(T)=-1$ dominated by vacuum, so this simplifies to something knows sometimes under the name "concordance model" and denoted traditionally in most textbooks:

$$H(a)\cong H_0\sqrt{\Omega_m a^{-3}+\Omega_{\text{rad}}a^{-4}+\Omega_\Lambda} \tag{12.5.169}$$

This model is often designated under the name of "late-time model, i.e. after the epoch of matter-radiation equality. The radiation density is then less and less important and the only relevant ingredients are matter and vacuum energy.

Observations show that the radiation density in our observable universe is very small today ($\Omega_{\text{rad}}\sim 10^{-4}$), if this term is neglected the above equation reduces to:

$$H(a)\cong H_0\sqrt{\Omega_m a^{-3}+\Omega_\Lambda} \tag{12.5.170}$$

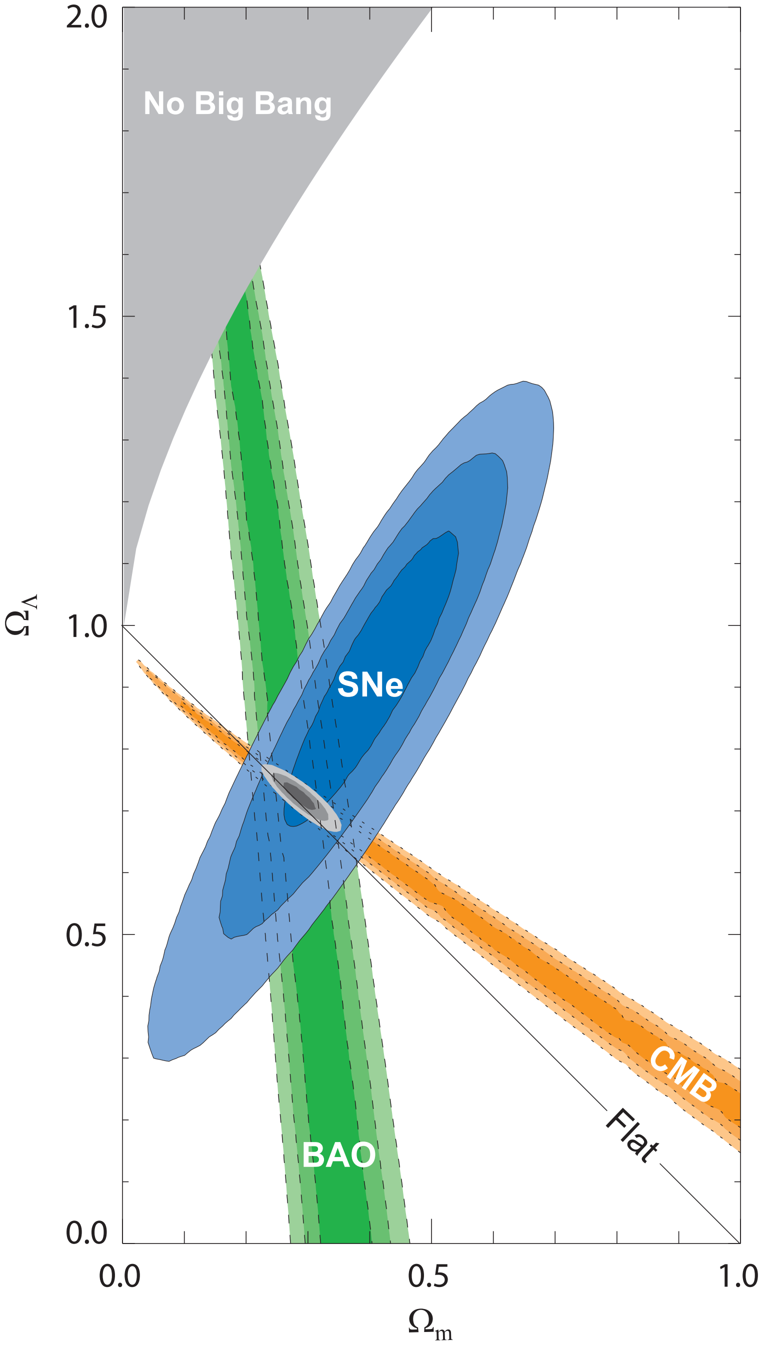
and has an analytic solution that we will see further below at page 6038!

The reader must remember that we have already seen earlier above (see page 5898) that by construction:

$$\Omega_m+\Omega_\Lambda=1 \tag{12.5.171}$$

in flat universe.

For information, observationally allowed regions of Ω_m and Ω_Λ at $68.3, 95.4$ and 99.7% confidence level obtained from the cosmic microwave background (CMB) temperature fluctuations (that we will prove in a future version of this textbook), baryon acoustic oscillations (BAO) in large-scale structure data (that we will also prove in a future version of this textbook), and from the supernova (SN) brightness-distance relation (that we will also prove in a future version of this textbook...). Also shown is the combined allowed region (figure from Kowalski et al. 12008, http://arxiv.org/abs/0804.4142, published in Astrophysical Journal 686 (12008) $749-778$):



From a combination of cosmic microwave background (CMB) and large scale structure (LSS) observations, astrophysicists conclude actually that our observable universe is spatially flat and is composed of 4% mater, 23% dark matter and 73% dark energy (vacuum energy or cosmological constant Λ), i.e.:

$$\Omega_k \cong 0 \qquad \Omega_m = \Omega_M + \Omega_{\text{DM}} \cong 0.04 + 0.23 \qquad \Omega_\Lambda \cong 0.73 \tag{12.5.172}$$

The reader must never forget the margin of error of our Universe age estimation that is at the day we write these lines (year 12023 according to holocene calendar) using other data than Planck of approximately plus/minus one billion year... as illustrated below:

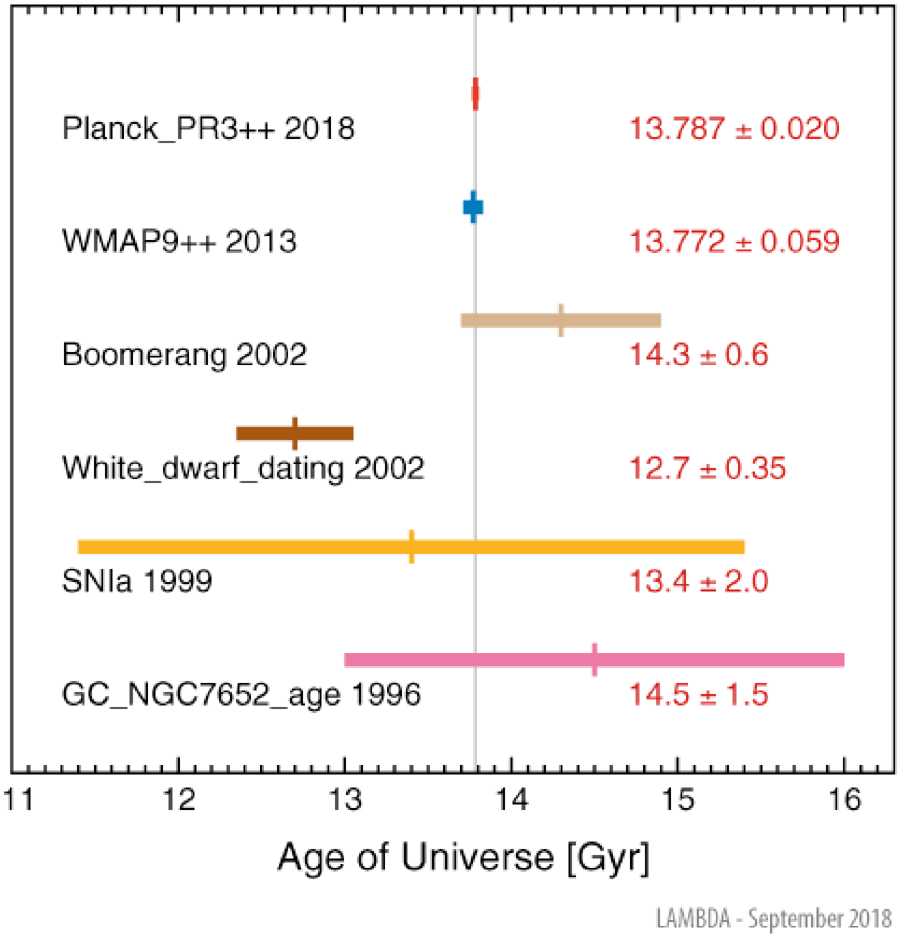


Figure 12.262: **Various estimations of our Universe age as in 12018 (source: NASA / LAMBDA Archive Team)**

12.5.3.1.3 Flat space dominated by vacuum

With everything we have seen so far, we know that for vacuum ($n = 0$):

$$R^n \rho = R_0^n \rho_0 \Rightarrow \rho_V = \rho_{V,0} \tag{12.5.173}$$

But also that:

$$\omega(T) = -1 \tag{12.5.174}$$

Therefore:

$$\rho(t) \propto R(t)^{-3(1+\omega(T))} = R^0(t) = 1 \tag{12.5.175}$$

and:

$$P=\omega\rho=-\rho \tag{12.5.176}$$

Under these conditions, the first Friedmann equation:

$$\left(\frac{R'}{R}\right)^2=\frac{8\pi G\rho}{3}-\frac{k}{R^2} \tag{12.5.177}$$

becomes first by changing into energy density of vacuum and with k being put as equal to zero:

$$\left(\frac{\dot{R}}{R}\right)^2=\frac{8\pi G\rho_V}{3c^2} \tag{12.5.178}$$

and so we can replace the energy density by the result we get just before:

$$\left(\frac{\dot{R}}{R}\right)^2=\frac{8\pi G\rho_{V,0}}{3c^2} \tag{12.5.179}$$

Then we have:

$$\frac{\mathrm{d}R}{\mathrm{d}t}=\sqrt{\frac{8\pi G\rho_{V,0}}{3c^2}}R^2 \tag{12.5.180}$$

hence:

$$R^2\,\mathrm{d}R=\sqrt{\frac{8\pi G\rho_{V,0}}{3c^2}}\,\mathrm{d}t \tag{12.5.181}$$

The primitive is obvious and is immediate:

$$\frac{1}{3}R^3+c^{te}=\sqrt{\frac{8\pi G\rho_{V,0}}{3c^2}}t \tag{12.5.182}$$

At the time $t=0$ of the "Big Bang", many textbooks assume the scale factor is equal to $R=0$ (only in the purpose to simplify the notations!!!) and thus the constant is zero (don't forget that in reality we should rather assume that it as equal to -2 instead to have after rearranging a $+1$ on the right side of the equality when $t=0$). Then we have:

$$\frac{1}{3}R^3=\sqrt{\frac{8\pi G\rho_{V,0}}{3c^2}}t \tag{12.5.183}$$

Thus after simplification it remains:

$$R=\left(\frac{72\pi G\rho_{V,0}}{3c^2}\right)^{1/6}t^{1/3}=\left(\frac{72\pi G}{3c^2}\rho_{V,0}\right)^{1/6}t^{1/3}=\left(\frac{72\pi G}{3c^2}\frac{3H_0^2}{8\pi G}\right)^{1/6}t^{1/3} \tag{12.5.184}$$

Thus after simplification it remains only:

$$R=(9H_0)^{1/6}\,t^{1/3} \tag{12.5.185}$$

Using the common notation in theoretical cosmology we get the famous relation often given without proof:

$$\boxed{a(t) \propto t^{1/3}} \tag{12.5.186}$$

Then a flat universe dominated by vacuum has a scaling factor that is growing slightly slower than a flat universe dominated by matter or radiation.

In comparison with Maple™ 4.00b (blue: flat universe dominated by matter, in red: flat universe dominated by radiation, green: flat universe dominated by radiation):

```
>plot([t^(2/3),t^(1/2),t^(1/3)],t=0..2*Pi,0..3,color=[blue,red,green]);
```

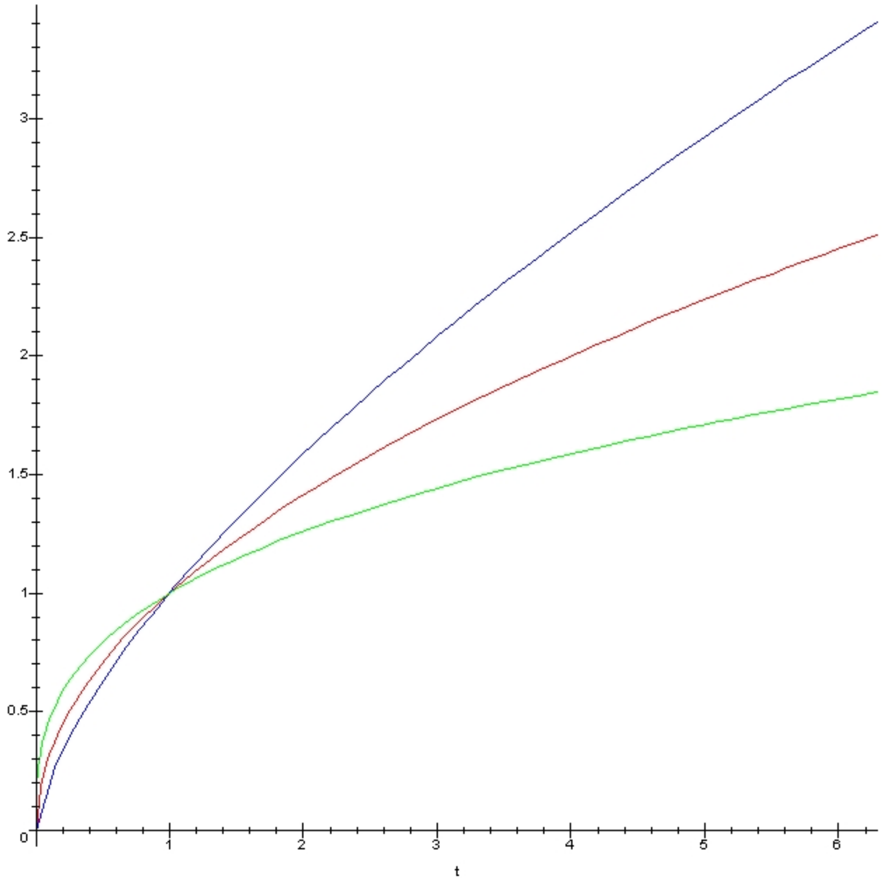


Figure 12.263: **Evolution of R for a zero curvature space dominated by matter, radiation and vacuum**

12.5.3.2 Spherical spaces ($k > 0$)

In this model (also sometimes named "elliptical model"), we consider $k > 0$. So the equation to deal with remains:

$$\mathrm{d}F(t) = \pm\sqrt{\frac{A}{F} - k\mathrm{d}t} \tag{12.5.187}$$

Which can also be written:

$$\Delta t=\pm\int\limits_{t_0}^{t}\frac{1}{\sqrt{\dfrac{A}{F}-k}}\mathrm{d}F(t)\geq 0 \tag{12.5.188}$$

Let us recall that we have assumed that for $t=t_0$ we had $F(t_0)=1$, if we make the change of variable $U=1/F(t)$, we get the following integral:

$$\Delta t=\pm\int\limits_{1}^{1/F}\frac{\mathrm{d}U}{U^2\sqrt{AU-k}} \tag{12.5.189}$$

So we are looking of a primitive of:

$$\Delta t=\int\limits_{1}^{1/F}\frac{\mathrm{d}U}{U^2\sqrt{AU-k}} \tag{12.5.190}$$

and we will discuss the sign ± after having found the primitive.

We still carry a change of variable:

$$x^2=AU-k \tag{12.5.191}$$

thus:

$$2x\mathrm{d}x=A\mathrm{d}U \tag{12.5.192}$$

which gives us the following primitive to calculate:

$$\frac{2}{A}\int\frac{1}{U^2}\mathrm{d}x=\frac{2A^2}{A}\int\frac{1}{x^2+k^2}\mathrm{d}x=\frac{2A}{k^2}\int\frac{1}{\left(\left(\dfrac{1}{\sqrt{k}}\right)^2+1\right)^2}\mathrm{d}s \tag{12.5.193}$$

by doing again a change of variable:

$$y=\frac{x}{\sqrt{k}}\Rightarrow\mathrm{d}x=\sqrt{k}\mathrm{d}y \tag{12.5.194}$$

Therefore to a given multiplicative constant $2Ak^{-3/2}$ we have finally the following integral:

$$\Delta t=\int\frac{\mathrm{d}y}{(1+y^2)^2} \tag{12.5.195}$$

In the section of Differential and Integral Calculus we proved that this form of primitive is resolved by (we add the constant of integration at the end because we do physics and we must satisfy to some initial conditions at which are not necessarily interested to in mathematics):

$$2nI_{n+1}=\frac{y}{(1+y^2)^n}+(2n-1)I_n+c^{te} \tag{12.5.196}$$

with:

$$I_n=\int\frac{\mathrm{d}y}{(1+y^2)^n} \tag{12.5.197}$$

hence:

$$I_2=\frac{1}{2}\frac{y}{(1+y^2)}+\frac{1}{2}I_1 \tag{12.5.198}$$

We still have to calculate I_1 (see section Differential and Integral Calculus page 1534):

$$I_1=\int\frac{1}{(1+y^2)}\mathrm{d}y=\arctan(y)+c^{te} \tag{12.5.199}$$

Finally:

$$I_2=\frac{1}{2}\frac{y}{(1+y^2)}+\frac{1}{2}\arctan(y)+c^{te} \tag{12.5.200}$$

by inverting all the changes of variables and introducing the multiplicative constant again, we have finally in the case where $k>0$:

$$I_2=\frac{\sqrt{AU-k}}{kU}+\frac{A}{k^{3/2}}\arctan\left(\frac{\sqrt{AU-k}}{\sqrt{k}}\right)+\frac{2A}{k^{3/2}}c^{te} \tag{12.5.201}$$

Between the two terminals of integration $(1/F,1)$ we therefore have (the integration constant cancels and we take back the $\pm$ which was originally in the primitive):

$$\Delta t=\pm\left[\left(\frac{F\sqrt{A/F-k}}{k}+\frac{A}{k^{3/2}}\arctan\left(\frac{\sqrt{A/F-k}}{\sqrt{k}}\right)\right)-\left(\frac{\sqrt{A-k}}{k}+\frac{A}{k^{3/2}}\arctan\left(\frac{\sqrt{A-k}}{\sqrt{k}}\right)\right)\right] \tag{12.5.202}$$

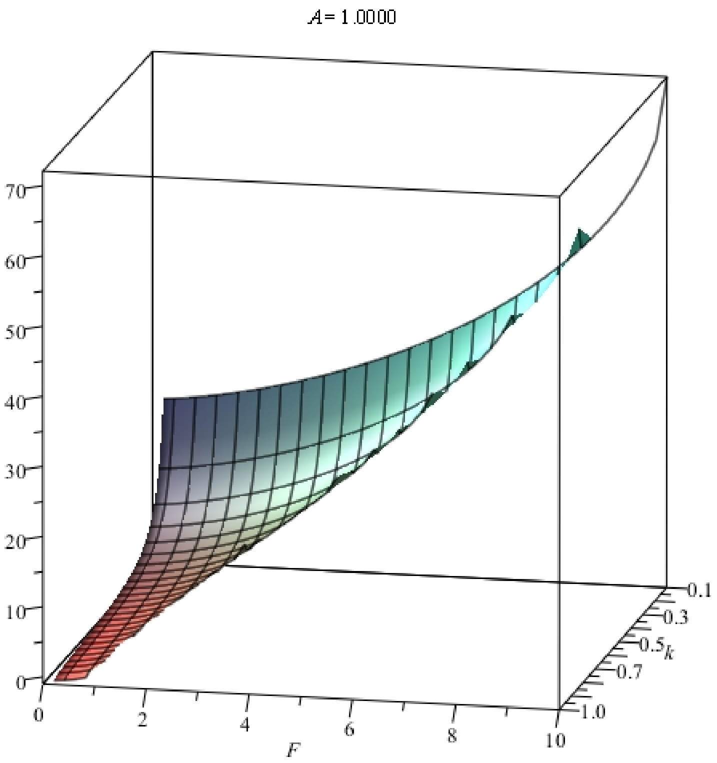
where for recall the theory request that $\Delta t>0$ (otherwise it's pure speculation and philosophy...).

We see that as expected (!!!), if we $F=1$ in the above relation we have indeed $\Delta t=0$ (that means that at the moment in time we take the Universe actual size as reference, the time variation is indeed equal to 0 as expected). In other words, we also have for this model (changing to the other typical notation for the scale factor):

$$\lim_{t\rightarrow 0}a(t)=1 \tag{12.5.203}$$

If we plot this function for a fixed value $k>0$. We have the following animated plot Maple™ in 17.00 for the negative sign:

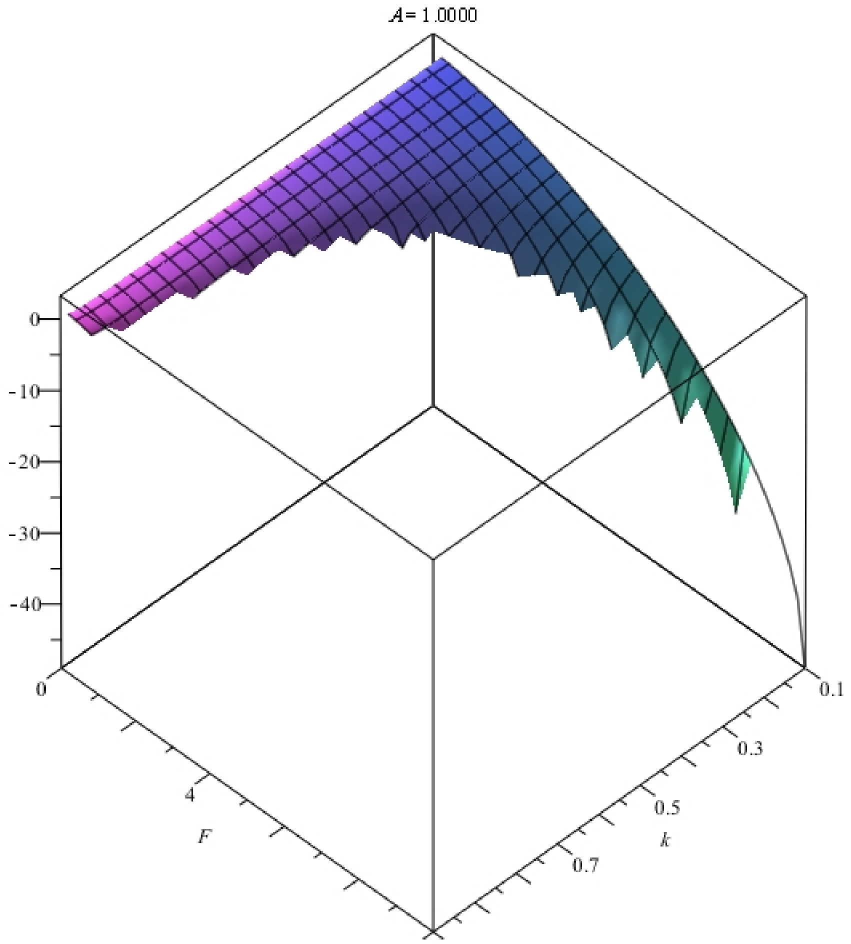
```
>restart:
>f:=(A,k,F)->-[(F*sqrt(A/F-k)/k+A/k^(3/2)*arctan(sqrt(A/f-k)/sqrt(k)))
-(sqrt(A-k)/k+A/k^(3/2)*arctan(sqrt(A-k)/sqrt(k)))]:
>plots:-animate(plot3d,[f(A,k,F),k=0.1..1,F=0..10,],A=0..1)
```



We can see with this solution that as time increase the Universe reach an asymptote size whatever the value of k, but the bigger is k the faster the asymptote is reached. In other words... we must be lucky not to be in a Universe with a to big positive curvature...

And for the positive sign:

```
>restart:
>f:=(A,k,F)->+[(F*sqrt(A/F-k)/k+A/k^(3/2)*arctan(sqrt(A/f-k)/sqrt(k)))
-(sqrt(A-k)/k+A/k^(3/2)*arctan(sqrt(A-k)/sqrt(k)))]:
>plots:-animate(plot3d,[f(A,k,F),k=0.1..1,F=0..10,],A=0..1)
```



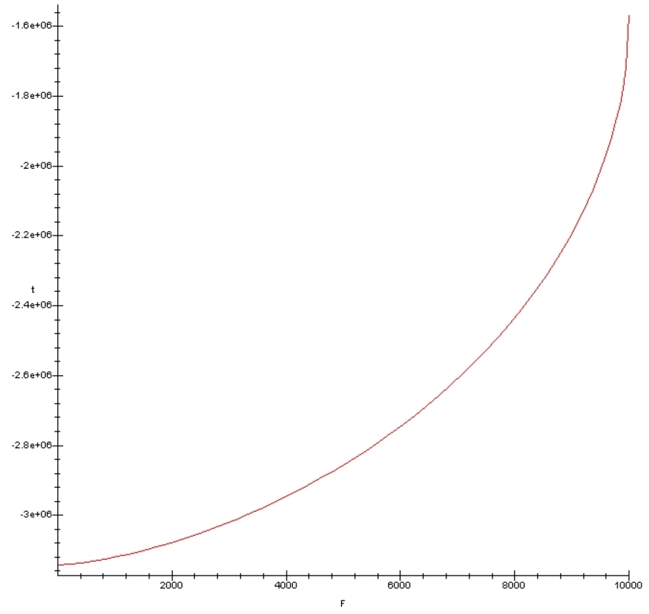
We can see that this solution seems to be the symmetric with the previous plot. That means as time reached back to the zero value, the Universe contracts on single point back and once again this effect is slower as k is big!

Remark

The time Δt in the above plots is always represented on the vertical axis and also for all the following charts further below (you have to turn your head a little if as usually you want to put the time on the horizontal axis...).

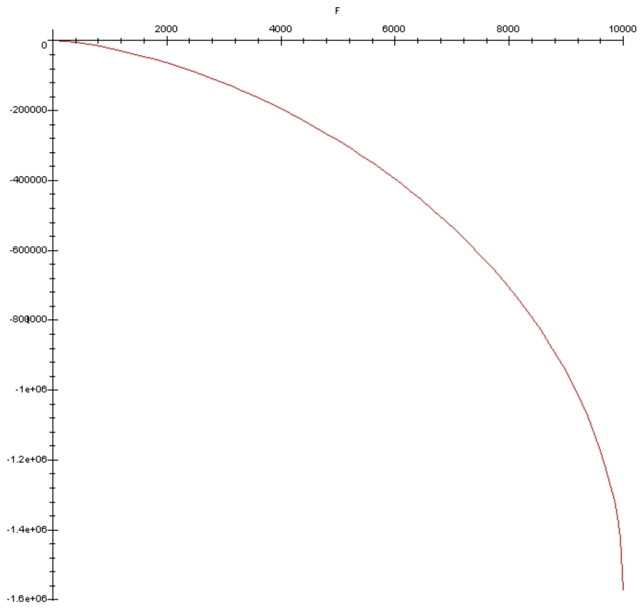
By fixing a small value of A and for k, we get the following two-dimensional plot first for the negative sign:

```
>k:=0.0001;A:=1;
>plot([-(F*sqrt(A/F-k)/k+A/k^(3/2)*arctan(sqrt(A/F-k)/sqrt(k)))-(sqrt(A-k)/k
+A/k^(3/2)*arctan(sqrt(A-k)/sqrt(k)))],F=1..10000,labels=[F,t])
```



where as we have already mentioned it, the reader must keep in mind that the values below 1 must be rejected!

And for the positive sign:



By looking at the both plots above it is obvious to see that we have after a rotation and putting each one next to the other (we also could change the time reference to have a logical time axis):

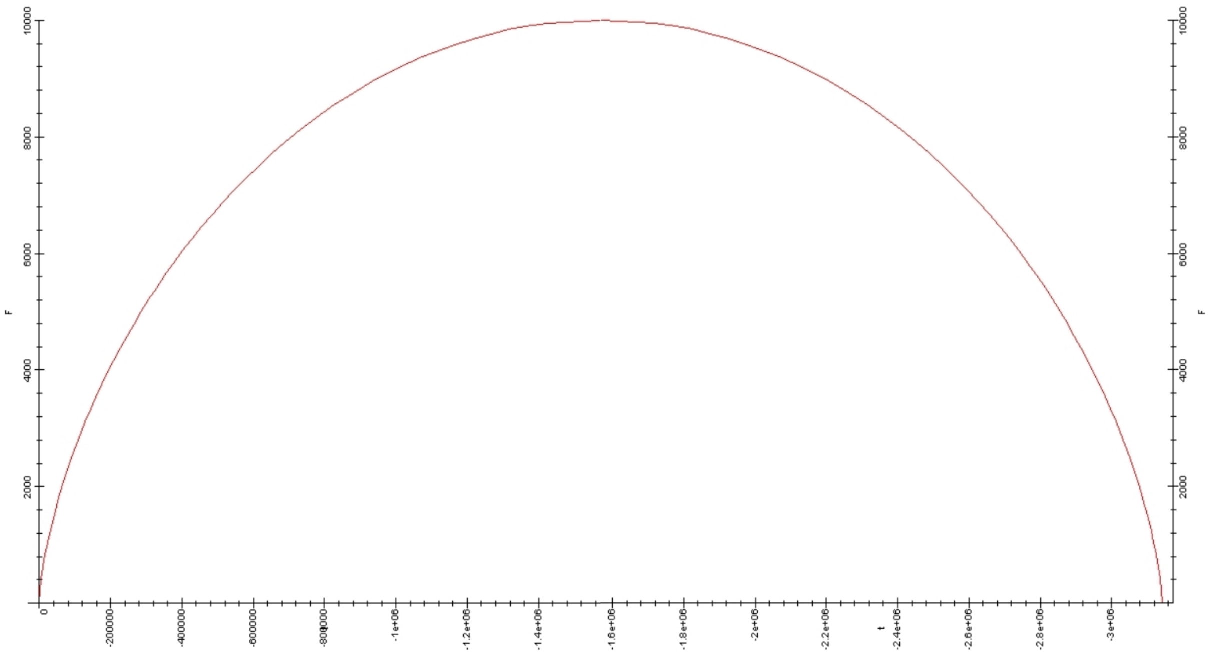


Figure 12.264: **Big Bang and Big Crunch plot side by side**

and what we see here is the Big Bang and the Big Crunch!

Now let us recall that to build the previous model we started from:

$$(\mathrm{d}_t F)^2 = \frac{8\pi}{3} G\rho_0 F^{-1} - k \tag{12.5.204}$$

A limit condition (condition of integration) is that the right term to be positive. That is:

$$F_{\text{lim}} = \frac{8\pi G\rho_0}{3k} \tag{12.5.205}$$

or:

$$F_{\text{lim}}^{-1} = \frac{3k}{8\pi G\rho_0} = \frac{k}{A} \tag{12.5.206}$$

So for our previous 2D plot this limit is located at $F = k/A = 10,000$ and this is according to the maximum value before the Universe turn into a Big Crunch!

So if F^{-1} is smaller than F_{lim}^{-1}, we are not in a valid (real) domain model anymore.

In fact, beyond the time limit t_{lim} corresponding to this F_{lim}, what does not know the computer that has drawn our function plot is that it should switch to the scaling function with the "$+$". So when we execute the plot of both functions we should get the previous figure.

We see then that for $t < t_{\text{lim}}$ the Universe is entering a phase of contraction that we commonly name the "Big Crunch". After this phase of contraction, it is possible that either the Universe disappears completely in a singularity or that it re-enters a cyclical dynamic phase (mathematically the two outcomes seem to be possible).

12.5.3.2.1 Spherical space dominated by matter

Just as the model for flat space, there is also another approach much more elegant and subtle for my taste than the previous proof (I have also discovered that many years after writing the previous text). It also has the advantage of highlighting a hypothesis that has not appear with previous developments and allows to plot more simply in Maple™ 4.00b the behaviour of the scale factor of the Universe. We thus find exactly the famous plot representing the evolution of the scale factor of the Universe available in almost all popular books on the subject (without proofs obviously...).

We start again from the first Friedmann equation:

$$\left(\frac{R'}{R}\right)^2=\frac{8\pi G\rho}{3}-\frac{k}{R^2} \tag{12.5.207}$$

It is customary for this model equation to put $k=+1$ (even take any positive number at least pick one that is friendly...) and we have shown that when matter dominates, we had:

$$R^3\rho=R_0^3\rho_0=c^{te} \tag{12.5.208}$$

Since then:

$$\frac{R'^2}{R^2}=\frac{8\pi G}{3}\rho_0\left(\frac{R_0}{R}\right)^3-\frac{1}{R^2} \tag{12.5.209}$$

and it comes immediately:

$$\frac{\mathrm{d}R}{\mathrm{d}t}=\sqrt{\frac{8\pi G\rho_0R_0^3}{3R}-1} \tag{12.5.210}$$

Therefore:

$$\int\mathrm{d}t=\int\frac{1}{\sqrt{\dfrac{8\pi G\rho_0R_0^3}{3R}-1}}\mathrm{d}R \tag{12.5.211}$$

If we move now to the comoving time also named "conformal time" (already introduced at the beginning of this section) defined mathematically by (don't forget that R is not a radius but a ratio of two radius!):

$$\mathrm{d}\eta=\frac{\mathrm{d}t}{R} \tag{12.5.212}$$

Then we have:

$$\int\mathrm{d}\eta=\int\frac{1}{\sqrt{\dfrac{8\pi G\rho_0R_0^3}{3}R-R^2}}\mathrm{d}R \tag{12.5.213}$$

Let us write that in the form:

$$\int\mathrm{d}\eta=\int\frac{1}{\sqrt{AR-R^2}}\mathrm{d}R \tag{12.5.214}$$

where A is strictly positive. Let us do the substitution:

$$R=x+\frac{A}{2} \tag{12.5.215}$$

Then we have:

$$\begin{gathered}\int \mathrm{d}\eta=\int\frac{1}{\sqrt{A\left(x+\dfrac{A}{2}\right)-\left(x+\dfrac{A}{2}\right)^2}}\mathrm{d}x=\int\frac{1}{Ax+\dfrac{A^2}{x}-\left(x^2+Ax+\dfrac{A^2}{4}\right)}\mathrm{d}x\\ =\int\frac{1}{\sqrt{\dfrac{A^2}{4}-x^2}}\mathrm{d}x\end{gathered} \tag{12.5.216}$$

and we have proved in the section of Differential and Integral Calculus (see 1534) that the primitive is:

$$\int\frac{1}{\sqrt{\dfrac{A^2}{4}-x^2}}\mathrm{d}x=\arcsin\left(\frac{2x}{4}\right)+c^{te} \tag{12.5.217}$$

Therefore:

$$\eta=\arcsin\left(\frac{2R-A}{A}\right)+c^{te} \tag{12.5.218}$$

As at time $\eta=0$ we have $R=0$, it is necessary that the constant is such that:

$$\eta=\arcsin\left(\frac{2R-A}{A}\right)+\frac{\pi}{2} \tag{12.5.219}$$

Therefore:

$$\frac{2R-A}{A}=\sin\left(\eta-\frac{\pi}{2}\right)=-\cos(\eta) \tag{12.5.220}$$

Hence:

$$R=\frac{A}{2}(1-\cos(\eta)) \tag{12.5.221}$$

Now, let us recall that:

$$\mathrm{d}\eta=\frac{\mathrm{d}t}{R}\Leftrightarrow \mathrm{d}t=R\mathrm{d}\eta \tag{12.5.222}$$

Therefore:

$$\int \mathrm{d}t=\int R\mathrm{d}\eta=\int\frac{A}{2}(1-\cos(\eta))\mathrm{d}\eta \tag{12.5.223}$$

Hence:

$$t=\frac{A}{2}(\eta-\sin(\eta))+c^{te} \tag{12.5.224}$$

and as we must have at time $t=0$ the comoving time that is also zero, the constant is then zero. Therefore we have the following parametric system in the end (something very strange with this result... it is that we fall back on the parametric equations of the brachistochrone curve[62]!!!!):

$$\begin{aligned}R&=\frac{A}{2}(1-\cos(\eta))\\t&=\frac{A}{2}(\eta-\sin(\eta))\end{aligned} \tag{12.5.225}$$

Hence it is periodic and if we can assimilate the result to a Universe that means we have a "Big Bang" and after a given time a "Big Crunch" and the process repeats indefinitely as a "Big Bounce". The idea is more or less (...) the following:

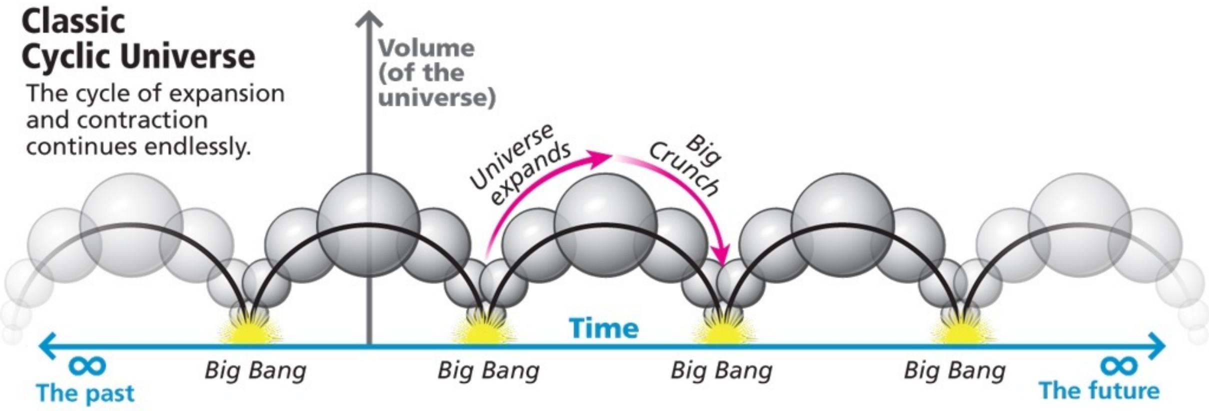

Classic
Cyclic Universe
The cycle of expansion and contraction continues endlessly.
Volume (of the universe)
Universe expands
Big Crunch
Time
∞
The past
Big Bang
Big Bang
Big Bang
Big Bang
∞
The future

Figure 12.265: **Big Bounce artistic illustration (source: ?)**

Remark

Notice that for $\eta=0$, we have $t=0$ but $R\neq 1$ and this wrong value results simply because earlier before we have assumed that for $\eta=0$, R was also equal to zero (but this was obviously an abusive simplification)!

With Maple™ 4.00b we then have by comparing the flat Universe dominated by matter (blue), the flat Universe dominated by radiation (red) and finally the positive curvature Universe dominated by matter (green) and putting artificial coefficients to better distinguish the plots:

```
>plot([t^(2/3),t^(1/2),[0.5*(t-sin(t)),0.5*(1-cos(t)),t=0..2*Pi]],
t=0...Pi,0..2.5,color=[blue,red,green]);
```

---

[62] See section Classical Mechanics page 3433

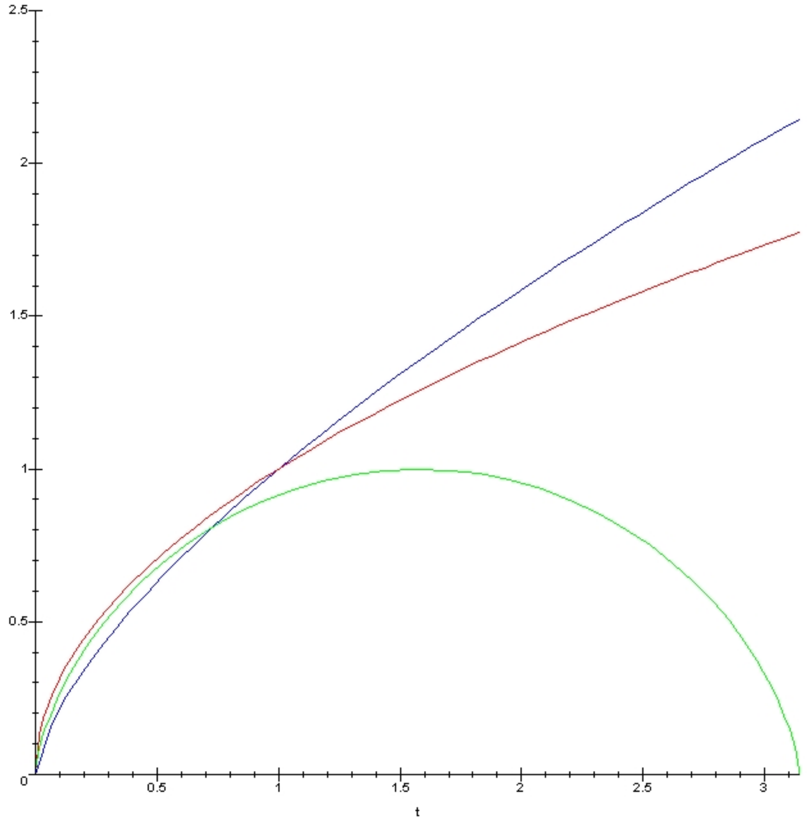


Figure 12.266: Evolution of the R factor for the resulting space configurations studied so far with Maple™ 4.00b

12.5.3.2.2 Spherical space dominated by radiation

Let us now consider a universe dominated by radiation. We have proved earlier above that in this situation we had:

$$P=\frac{1}{3}\rho_E \tag{12.5.226}$$

and:

$$R^4\rho_E=R_0^4\rho_{E,0}=c^{te} \tag{12.5.227}$$

In this case the Friedmann equation in terms of energy density can be written by putting $k=+1$:

$$\left(\frac{R'}{R}\right)^2=\frac{8\pi G\rho_E}{3c^2}-\frac{1}{R^2} \tag{12.5.228}$$

What becomes:

$$\frac{\mathrm{d}R}{\mathrm{d}t}=\sqrt{\frac{8\pi G\rho_E}{3c^2}R^2-1} \tag{12.5.229}$$

By injecting $R^4\rho_E=R_0^4\rho_{E,0}$, it comes:

$$\frac{\mathrm{d}R}{\mathrm{d}t}=\sqrt{\frac{8\pi G\rho_{E,0}R_0^4}{3c^2R^2}-1} \tag{12.5.230}$$

Therefore: Therefore:

$$\int \mathrm{d}t=\int \frac{1}{\sqrt{\dfrac{8\pi G\rho_{E,0}R_0^4}{3c^2R^2}-1}}\mathrm{d}R \tag{12.5.231}$$

Let us write this in the form:

$$\int \mathrm{d}t=\int \frac{1}{\sqrt{\dfrac{A}{R^2}-1}}\mathrm{d}R \tag{12.5.232}$$

If we change to the comoving time again:

$$\eta=\frac{\mathrm{d}t}{R} \tag{12.5.233}$$

Then we have:

$$\int \mathrm{d}\eta=\int \frac{1}{\sqrt{A-R^2}}\mathrm{d}R \tag{12.5.234}$$

In the section of Differential and Integral Calculus (see page 1534) we have proved how to determine exactly the same primitive (because it is a usual primitive). We have:

$$\eta=\arcsin\left(\frac{R}{\sqrt{A}}\right)+c^{te} \tag{12.5.235}$$

For at time $\eta=0$ we have $R=0$, it is necessary that the constant is such that:

$$\eta=\arcsin\left(\frac{R}{\sqrt{A}}\right) \tag{12.5.236}$$

Hence:

$$R=\sqrt{A}\sin(\eta) \tag{12.5.237}$$

Now, remember that:

$$\eta=\frac{\mathrm{d}t}{R}\Leftrightarrow \mathrm{d}t=R\mathrm{d}\eta \tag{12.5.238}$$

Therefore:

$$\int \mathrm{d}t=\int R\mathrm{d}\eta=\int \sqrt{A}\sin(\eta)\mathrm{d}\eta \tag{12.5.239}$$

Hence:

$$t=-\sqrt{A}\cos(\eta)+c^{te} \tag{12.5.240}$$

and as at time $t=0$ the comoving time η is also zero, the constant is therefore equal to $\sqrt{A}$. Therefore we have finally the following parametric system:

$$\begin{aligned}R&=\sqrt{A}\sin(\eta)\\ t&=\sqrt{A}(1-\cos(\eta))\end{aligned} \tag{12.5.241}$$

Remark

Notice that for $\eta=0$, we have $t=0$ but $R\neq 1$ and this wrong value results simply because earlier before we have assumed that for $\eta=0$, R was also equal to zero (but this was obviously an abusive simplification)!

With Maple™ 4.00b we can then compare the flat Universe dominated by matter (blue), the flat Universe dominated by radiation (red), the Universe with positive curvature dominated by matter (green), the Universe with positive curvature dominated by radiation (black):

```
>plot([t^(2/3),t^(1/2),[0.5*(t-sin(t)),0.5*(1-cos(t)),t=0..2*Pi],[0.5*
(1-cos(t)),0.5*sin(t),t=0..2*Pi]],t=0...Pi,0..3,color=[blue,red,green,black]);
```

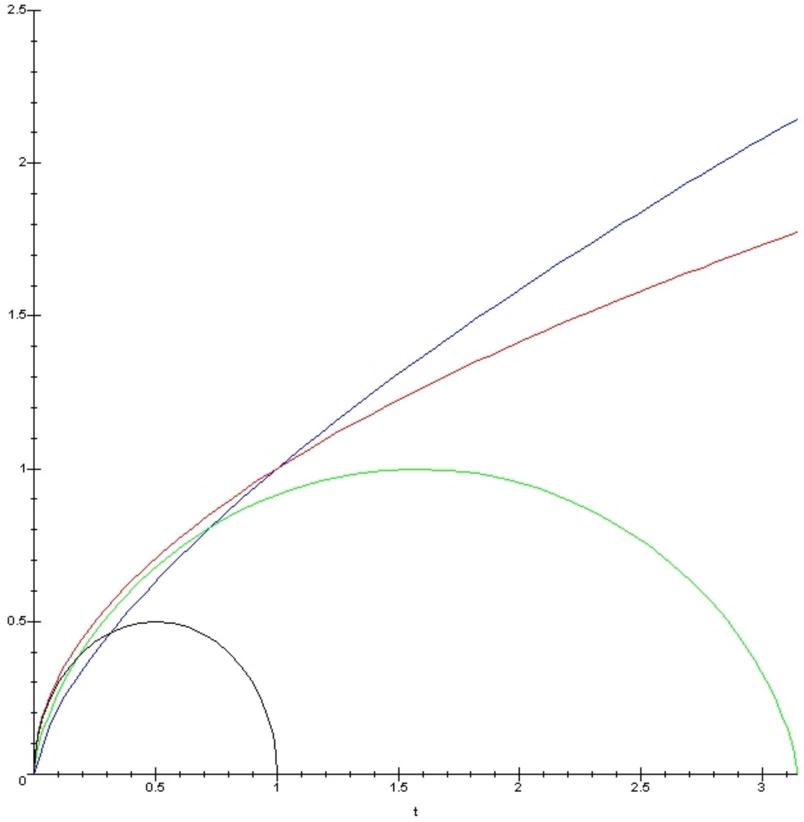


Figure 12.267: **Evolution of the R factor for the resulting space configurations studied so far with Maple™ 4.00b**

12.5.3.2.3 Spherical space dominated by vacuum

Remember that we have seen earlier above that:

$$\rho(t)\propto R(t)^{-3(1+\omega(T))}\propto\begin{cases}R(t)^{-3} & \text{for } \omega(T)=0 \text{ (pressureless matter (i.e. dust))}\\ R(t)^{-4} & \text{for } \omega(T)=1/3 \text{ (radiation/ultra-relativistic matter)}\\ c^{te} & \text{for } \omega(T)=-1 \text{ (vacuum energy)}\end{cases} \tag{12.5.242}$$

And we also have:

$$R^n\rho=R_0^n\rho_0\Rightarrow\begin{cases}n=0\text{ for vacuum}\\ n=3\text{ for matter dominated Universe}\\ n=4\text{ for radiation dominated Universe}\end{cases}\tag{12.5.243}$$

Hence the case that we want to focus here on is when $\omega(T)=-1$.

We start again from the first Friedmann equation:

$$\left(\frac{R'}{R}\right)^2=\frac{8\pi G\rho}{3}-\frac{k}{R^2}\tag{12.5.244}$$

where again we put here $k=+1$ and we have shown that when vacuum dominates, we had:

$$R^0\rho=R_0^0\rho_0=\rho_{V,0}\tag{12.5.245}$$

Since then:

$$\frac{R'^2}{R^2}=\frac{8\pi G\rho_{V,0}}{3}-\frac{1}{R^2}\tag{12.5.246}$$

and it comes immediately:

$$\frac{\mathrm{d}R}{\mathrm{d}t}=\sqrt{\frac{8\pi G\rho_{V,0}R^2}{3}-1}\tag{12.5.247}$$

Therefore:

$$\int\mathrm{d}t=\int\frac{1}{\sqrt{\dfrac{8\pi G\rho_{V,0}R^2}{3}-1}}\mathrm{d}R\tag{12.5.248}$$

We know have already derived the primitive above during our study of usual primitives in the section of Integral and Differential Calculus at page 1534.

Therefore we get:

$$t=\sqrt{\frac{3}{8\pi G\rho_{V,0}}}\ln\left(\sqrt{\frac{8\pi G\rho_{V,0}}{3}}R+\sqrt{\frac{8\pi G\rho_{V,0}}{3}R^2-1}\right)\tag{12.5.249}$$

Taking the exponential of the above relation, squaring and solving the second degree polynomial relatively to R gives:

$$R=\frac{1}{2}\frac{1+e^{t^2\left(\frac{8\pi G\rho_{V,0}}{3}\right)^2}}{\sqrt{\frac{8\pi G\rho_{V,0}}{3}}e^{t\sqrt{\frac{8\pi G\rho_{V,0}}{3}}}}\tag{12.5.250}$$

The latter has a real solution for all $t\in\mathbb{R}$. With Maple™ 4.00b only for $t>0$ (it's symmetrical in negative axis if we choose the right integration constant), we get roughly something like the following:

```
>plot(1+exp(t^2))/exp(t),t=0..3);
```

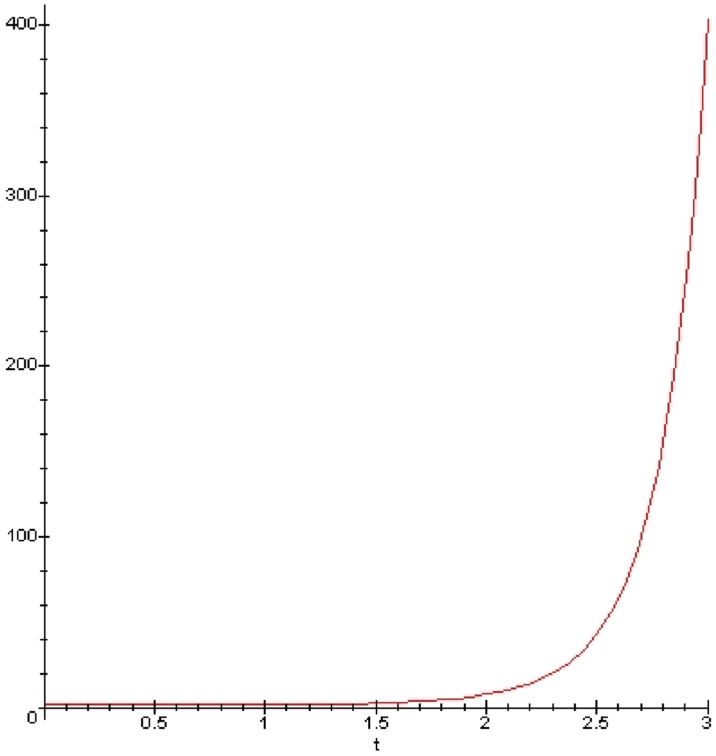


Figure 12.268: Evolution of R for a spherical space dominated by vacuum

12.5.3.3 Hyperbolic spaces ($k<0$)

In this model, we consider $k<0$. So the equation to be treated can be written:

$$\mathrm{d}F(t)=\pm\sqrt{\frac{A}{F}+|k|}\mathrm{d}t \tag{12.5.251}$$

Which is also written:

$$\Delta t=\pm\int\limits_{t_0}^{t}\frac{1}{\sqrt{\dfrac{A}{F}+|k|}}\mathrm{d}t \tag{12.5.252}$$

Let us recall that we assumed that $t=t_0$ that $F(t_0)=1$. If we make the change of variable $U=1/F(t)$, we get the following integral:

$$\Delta t=\pm\int\limits_{1}^{1/F}\frac{\mathrm{d}U}{U^2\sqrt{AU+|k|}}\geq 0 \tag{12.5.253}$$

So we are looking for a primitive of:

$$\Delta t=\int\frac{\mathrm{d}U}{U^2\sqrt{AU+|k|}}\geq 0 \tag{12.5.254}$$

and we will discuss the sign ± after finding the primitive.

We still carry a change of variable by putting:

$$x^2 = AU + k \tag{12.5.255}$$

therefore:

$$2x\mathrm{d}x = A\mathrm{d}U \tag{12.5.256}$$

which gives us the following primitive to calculate:

$$\frac{2}{A}\int\frac{1}{U^2}\mathrm{d}x = \frac{2A^2}{A}\int\frac{1}{(x^2-k)^2}\mathrm{d}x = \frac{2A}{k^2}\int\frac{1}{\left(\left(\frac{x}{\sqrt{x}}-1\right)^2\right)^2}\mathrm{d}x \tag{12.5.257}$$

Doing again a change of variable:

$$y = \frac{x}{\sqrt{x}} \Rightarrow \mathrm{d}x = \sqrt{k}\mathrm{d}y \tag{12.5.258}$$

hence to a given multiplicative constant:

$$\frac{2A}{k^{3/2}} \tag{12.5.259}$$

We have:

$$\int\frac{\mathrm{d}y}{(y^2-1)^2} = \int\frac{\mathrm{d}y}{(1-y^2)^2} \tag{12.5.260}$$

In the section of Differential and Integral Calculus (see page 1534) we saw that this form of primitive is resolved by the relation (we added the constant of integration in the end because we do physics and must satisfy the initial conditions to which we were not interested to in pure mathematics):

$$2nI_{n+1} = \frac{y}{(1-y^2)^n} + (2n-1)I_n + c^{te} \tag{12.5.261}$$

with:

$$I_n = \int\frac{\mathrm{d}y}{(1-y^2)^n} \tag{12.5.262}$$

hence:

$$I_2 = \frac{1}{2}\frac{y}{1-y^2} + \frac{1}{2}I_1 \tag{12.5.263}$$

We still need to calculate I_1:

$$\begin{aligned}I_1 &= \int\frac{1}{1-y^2}\mathrm{d}y = \frac{1}{2}\int\left(\frac{1}{1+y}+\frac{1}{1-y}\right)\mathrm{d}y = \frac{1}{2}\ln(|1+y|) - \ln(|1-y|)\\ &= \frac{1}{2}\ln\left(\left|\frac{1+y}{1-y}\right|\right) + c^{te}\end{aligned} \tag{12.5.264}$$

Finally:

$$I_2=\frac{1}{2}\frac{y}{1-y^2}+\frac{1}{4}\ln\left(\left|\frac{1+y}{1-y}\right|\right)+c^{te} \qquad (12.5.265)$$

by putting back all the changes of variables and introducing the multiplicative constant again, we have therefore in the case $k>0$:

$$I_2=-\frac{\sqrt{AU+|k|}}{kU}+\frac{A}{2|k|^{3/2}}\ln\left(\left|\frac{\sqrt{|k|}+\sqrt{AU+|k|}}{\sqrt{|k|}-\sqrt{AU+|k|}}\right|\right)+\frac{2A}{|k|^{3/2}}c^{te} \qquad (12.5.266)$$

Between the two terminals of integration $(1/F,1)$ so we have (the integration constant is zero):

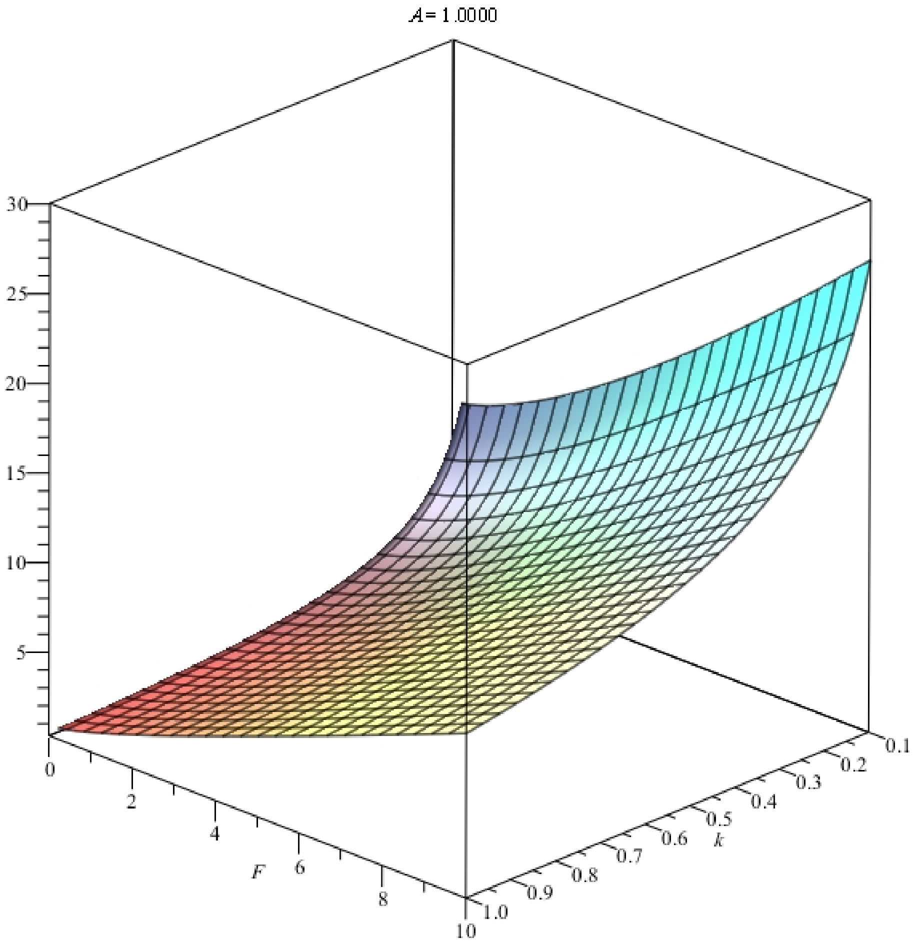
$$\begin{aligned}\Delta t=&\left(-\frac{-F\sqrt{A/F+|k|}}{k}+\frac{A}{2|k|^{3/2}}\ln\left(\left|\frac{\sqrt{k}+\sqrt{A/F+|k|}}{\sqrt{|k|}-\sqrt{A/F+|k|}}\right|\right)\right)\\&-\left(\frac{-\sqrt{A+|k|}}{k}+\frac{A}{2|k|^{3/2}}\ln\left(\left|\frac{\sqrt{|k|}+\sqrt{A/F+|k|}}{\sqrt{|k|}-\sqrt{A/F+|k|}}\right|\right)\right)\end{aligned} \qquad (12.5.267)$$

We must obviously have (we take back the ± which was originally in the integral):

$$\Delta t=\pm\left(\frac{-F\sqrt{A/F+|k|}+\sqrt{A+|k|}}{k}+\frac{A}{2|k|^{3/2}}\ln\left(\frac{\ln\left(\left|\frac{\sqrt{k}+\sqrt{A/F+|k|}}{\sqrt{|k|}-\sqrt{A/F+|k|}}\right|\right)}{\ln\left(\left|\frac{\sqrt{|k|}+\sqrt{A/F+|k|}}{\sqrt{|k|}-\sqrt{A/F+|k|}}\right|\right)}\right)\right) \qquad (12.5.268)$$

If we plot this function for a fixed value $k>0$ we then have the following animated plot with Maple™ 17.00 for the negative sign (as the positive one has no physical meaning):

```
>restart:
>f:=(A,k,F)->-(((-F(sqrt(A/F+abs(k)))/k+A/2(2*abs(k)^(3/2))*ln(abs(((sqrt(
abs(k))+sqrt(A/F+abs(k)))/(sqrt(abs(k))-sqrt(A/F+abs(k))))/((sqrt(abs(k))
+sqrt(A+abs(k)))/(sqrt(abs(k))-sqrt(A+abs(k)))))))):
>plots:-animate(plot3d,[f(A,k,F),k=0.1..1,F=0..10,],A=0.1..1)
```



We see that the smaller is the constant A, the fastest the Universe increases indefinitely quickly! Furthermore for a fixed value of k, some values of A are prohibited (it is in fact still the integration condition).

Again we see that the in the equation above the criterion $F(t_0)=F(0)=1$ is naturally fully respected. All values $F(t)$ below 1 are to be rejected! Hence using again the one of the traditional notation:

$$\lim_{t\to 0} a(t)=1 \tag{12.5.269}$$

So we have in this hyperbolic model a Universe that grows indefinitely in an exponential away (as the flat Friedmann-Lemaître model) because since $k<0$, there is no integration condition limit anymore (unlike the previous spherical model).

12.5.3.3.1 Hyperbolic space dominated by matter

Just as for the models for flat and spherical space, there is also another approach much more elegant and subtle for my taste than the previous proof (I have also discovered that one many years after writing the previous version). It also has the advantage of highlighting a hypothesis that has not occurred with previous developments and allows to draw more simply in Maple™ 4.00b the behaviour of the scale factor of the Universe. We thus find exactly the famous plot representing the evolution of the scale factor of the Universe available in almost all popular books on the subject but without proof.

We always start from the first Friedmann equation:

$$\left(\frac{R'}{R}\right)^2=\frac{8\pi G\rho}{3}-\frac{k}{R^2} \tag{12.5.270}$$

It is customary for this model to put $k=-1$ (as we have to choose any negative number at least we pick one that is friendly...) and we have proved that when the radiation dominates, we had:

$$R^3\rho=R_0^3\rho_0=c^{te} \tag{12.5.271}$$

In this case the Friedmann equation in terms of energy density can be written by putting $k=-1$:

$$\left(\frac{R'}{R}\right)^2=\frac{8\pi G\rho_E}{3c^2}-\frac{1}{R^2} \tag{12.5.272}$$

The first Friedmann equation then becomes:

$$\left(\frac{R'}{R}\right)=\frac{8\pi G\rho}{3}+\frac{1}{R^2}=\frac{8\pi G\rho_0}{3}\left(\frac{R_0}{R}\right)^3+\frac{1}{R^2} \tag{12.5.273}$$

Then we have:

$$\frac{\mathrm{d}R}{\mathrm{d}t}=\sqrt{\frac{8\pi G\rho_0 R_0^3}{3R}+1} \tag{12.5.274}$$

Hence:

$$\int \mathrm{d}t=\int\frac{1}{\sqrt{\frac{8\pi G\rho_0 R_0^3}{3R}+1}}\mathrm{d}R \tag{12.5.275}$$

This is exactly the same integral than that of the spherical Universe dominated by matter at the difference that in the root, we $+1$ instead of -1. We will proceed in the same manner using the time comoving time:

$$\mathrm{d}\eta=\frac{\mathrm{d}t}{R} \tag{12.5.276}$$

It comes then:

$$\int \mathrm{d}\eta=\int\frac{1}{\sqrt{\frac{8\pi G\rho_0 R_0^3}{3}+R^2}}\mathrm{d}R \tag{12.5.277}$$

Let us write that in the form:

$$\int \mathrm{d}\eta = \int \frac{1}{AR+R^2}\mathrm{d}R \tag{12.5.278}$$

where A is strictly positive. Let us make the substitution:

$$R = x - \frac{A}{2} \tag{12.5.279}$$

Then we have:

$$\begin{aligned}\int \mathrm{d}\eta &= \int \frac{1}{\sqrt{A\left(x-\dfrac{A}{2}\right)+\left(x-\dfrac{A}{2}\right)^2}}\mathrm{d}x = \int \frac{1}{\sqrt{Ax-\dfrac{A^2}{2}+\left(x^2-Ax+\dfrac{A^2}{4}\right)^2}}\mathrm{d}x\\ &= \int \frac{1}{\sqrt{x^2-\dfrac{A^2}{4}}}\mathrm{d}x\end{aligned} \tag{12.5.280}$$

Using the usual primitive proved in the section of Differential and Integral Calculus (see page 1534) it comes:

$$\int \frac{1}{\sqrt{x^2-\dfrac{A^2}{4}}}\mathrm{d}x = \text{arccosh}\left(\frac{2x}{A}\right)+c^{te} \tag{12.5.281}$$

Or redoing the change of variables:

$$\text{arccosh}\left(\frac{4x}{A^2}\right)+c^{te} = \text{arccosh}\left(\frac{2\left(R+\dfrac{A}{2}\right)}{A}\right)+c^{te} = \text{arccosh}\left(\frac{2R+A}{A}\right)+c^{te} \tag{12.5.282}$$

Therefore:

$$\eta = \text{arccosh}\left(\frac{2R+A}{A}\right)+c^{te} \tag{12.5.283}$$

So that at time $\eta=0$ we have $R=0$, it is necessary that the constant is such that:

$$\frac{\cosh(-c^{te})A-A}{2}=0 \tag{12.5.284}$$

Which brings us to that the constant is zero and thus:

$$\eta = \text{arccosh}\left(\frac{2R+A}{A}\right) \tag{12.5.285}$$

Therefore:

$$\frac{2R+A}{A} = \cosh(\eta) \tag{12.5.286}$$

Hence:

$$R=\frac{A}{2}(\cosh(\eta)-1) \tag{12.5.287}$$

and as:

$$\mathrm{d}\eta=\frac{\mathrm{d}t}{R} \tag{12.5.288}$$

We have:

$$\int \mathrm{d}t=\int R\mathrm{d}\eta=\int \frac{A}{2}(\cosh(\eta)-1)\mathrm{d}\eta \tag{12.5.289}$$

Which gives:

$$t=\frac{A}{2}(\sinh(\eta)-\eta)+c^{te} \tag{12.5.290}$$

As at time $t=0$, we must have $\eta=0$, it follows that the constant must be zero. So finally, we have:

$$\begin{aligned}R&=\frac{A}{2}(\cosh(\eta)-1)\\ t&=\frac{A}{2}(\sinh(\eta)-\eta)\end{aligned} \tag{12.5.291}$$

Remark

Notice that for $\eta=0$, we have $t=0$ but $R\neq 1$ and this wrong value results simply because earlier before we have assumed that for $\eta=0$, R was also equal to zero (but this was obviously an abusive simplification)!

With Maple™ 4.00b we then have by comparing the flat Universe dominated by matter (blue), the flat Universe dominated by radiation (red), the Universe with positive curvature dominated by matter (green), the Universe with positive curvature dominated by radiation (black), the negatively curved Universe dominated by matter (gray):

```
>plot([t^(2/3),t^(1/2),[0.5*(t-sin(t)),0.5*(1-cos(t)),t=0..2*Pi],
[0.5*(1-cos(t)),0.5*sin(t),t=0..2*Pi],[0.5*(sinh(t)-t),0.5*(cosh(t)-1)
,t=0..2*Pi]],t=0...Pi,0..3,color=[blue,red,green,black,gray]);
```

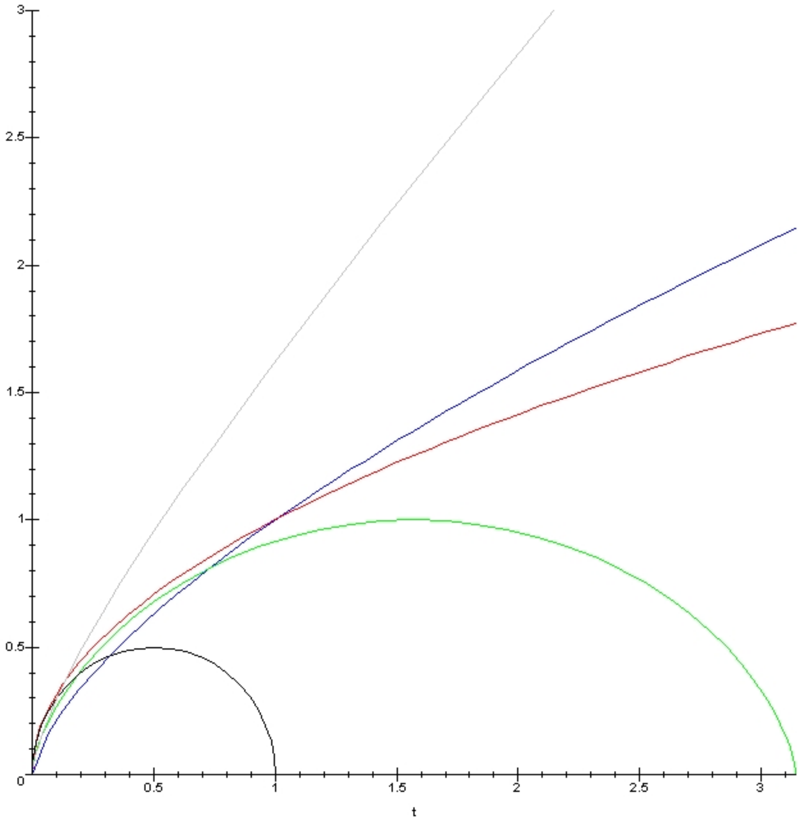


Figure 12.269: Evolution of the R factor for the resulting space configurations studied so far with Maple™ 4.00b

We can therefore observe that for a negative curvature (hyperbolic type), expansion is growing significantly faster than for a flat Universe and this without end.

12.5.3.3.2 Hyperbolic space dominated by radiation

Let us now consider a Universe dominated by radiation. We have proved that in this situation we had:

$$P=\frac{1}{3}\rho_E \tag{12.5.292}$$

and:

$$R^4\rho_E=R_0^4\rho_{E,0}=c^{te} \tag{12.5.293}$$

The first Friedmann equation the becomes:

$$\left(\frac{R'}{R}\right)^2=\frac{8\pi G\rho_E}{3c^2}+\frac{1}{R^2}=\frac{8\pi G\rho_{E,0}}{3c^2}\frac{R_0^4}{R^4}+\frac{1}{R^2} \tag{12.5.294}$$

Then we have:

$$\frac{\mathrm{d}R}{\mathrm{d}t}=\sqrt{\frac{8\pi G\rho_{E,0}R_0^4}{3R^2c^2}+1} \tag{12.5.295}$$

hence:

$$\int \mathrm{d}t=\int \frac{1}{\sqrt{\dfrac{8\pi G\rho_{E,0}R_0^4}{3R^2c^2}+1}}\mathrm{d}R \tag{12.5.296}$$

This is exactly the same integral than that of the spherical universe dominated by matter at the difference that in the root, we have $+1$ instead of -1. We will proceed in the same manner using the comoving time:

$$\mathrm{d}\eta=\frac{\mathrm{d}t}{R} \tag{12.5.297}$$

It comes then:

$$\int \mathrm{d}\eta=\int \frac{1}{\sqrt{\dfrac{8\pi G\rho_{E,0}R_0^4}{3c^2}+R^2}}\mathrm{d}R \tag{12.5.298}$$

Let us write this in the form:

$$\int \mathrm{d}\eta=\int \frac{1}{\sqrt{A+R^2}}\mathrm{d}R \tag{12.5.299}$$

where A is strictly positive. In the section of Differential and Integral Calculus (see page 1534) we have proved how to determine exactly the same primitive (because it is a usual primitive). We have:

$$\eta=\text{arcsinh}\left(\frac{R}{\sqrt{A}}\right)+c^{te} \tag{12.5.300}$$

So that at time $\eta=0$ we have $R=0$, it is necessary that the constant is zero. Therefore:

$$\eta=\text{arcsinh}\left(\frac{A}{\sqrt{A}}\right) \tag{12.5.301}$$

Hence:

$$R=\sqrt{A}\sinh(\eta) \tag{12.5.302}$$

and as:

$$\mathrm{d}\eta=\frac{\mathrm{d}t}{R} \tag{12.5.303}$$

We have:

$$\int \mathrm{d}t=\int R\mathrm{d}\eta=\int \sqrt{A}\sinh(\eta)\mathrm{d}\eta \tag{12.5.304}$$

Which gives:

$$t=\sqrt{A}\cosh(\eta)+c^{te} \tag{12.5.305}$$

As at the time $t=0$, we must have $\eta=0$, it follows that the constant must be equal to $-\sqrt{A}$. So finally, we have:

$$\begin{aligned}R&=\sqrt{A}\sinh(\eta)\\ t&=\sqrt{A}(\cosh(\eta)-1)\end{aligned} \tag{12.5.306}$$

Remark

Notice that for $\eta = 0$, we have $t = 0$ but $R \neq 1$ and this wrong value results simply because earlier before we have assumed that for $\eta = 0$, R was also equal to zero (but this was obviously an abusive simplification)!

With Maple™ 4.00b we then have by comparing the flat Universe dominated by matter (blue), the flat Universe dominated by radiation (red), the Universe with positive curvature dominated by matter (green), the Universe with positive curvature dominated by radiation (black), the Universe with negative curvature dominated by matter (gray), the negative curvature Universe dominated by radiation (brown):

```
>plot([t^(2/3),t^(1/2),[0.5*(t-sin(t)),0.5*(1-cos(t)),t=0..2*Pi],
[0.5*(1-cos(t)),0.5*sin(t),t=0..2*Pi],[0.5*(sinh(t)-t),0.5*(cosh(t)-1)
t=0..2*Pi],[0.5*(cosh(t)-1),0.5*(sinh(t)),t=0..2*Pi]],t=0...Pi,0..3
,color=[blue,red,green,black,grey,brown]);
```

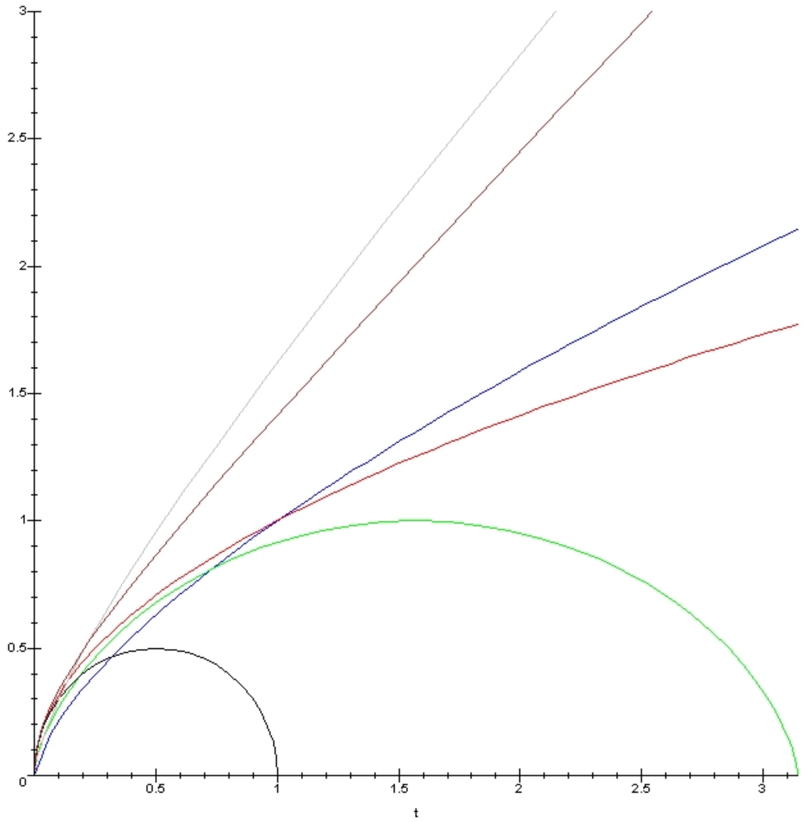


Figure 12.270: **Evolution of the R factor for the resulting space configurations studied so far with Maple™ 4.00b**

We can therefore observe that for a negative curvature (hyperbolic type), the expansion of a Universe dominated by radiation grows slower than a universe dominated by matter (it's a bit intuitive...).

12.5.3.3.3 Hyperbolic space dominated by vacuum

Remember that we have seen earlier above that:

$$\rho(t)\propto R(t)^{-3(1+\omega(T))}\propto\begin{cases}R(t)^{-3} & \text{for } \omega(T)=0 \text{ (pressureless matter (i.e. dust))}\\ R(t)^{-4} & \text{for } \omega(T)=1/3 \text{ (radiation/ultra-relativistic matter)}\\ c^{te} & \text{for } \omega(T)=-1 \text{ (vacuum energy)}\end{cases}\tag{12.5.307}$$

And we also have:

$$R^n\rho=R_0^n\rho_0\Rightarrow\begin{cases}n=0 \text{ for vacuum}\\ n=3 \text{ for matter dominated Universe}\\ n=4 \text{ for radiation dominated Universe}\end{cases}\tag{12.5.308}$$

Hence the case that we want to focus here on is when $\omega(T)=-1$.

We start again from the first Friedmann equation:

$$\left(\frac{R'}{R}\right)^2=\frac{8\pi G\rho}{3}-\frac{k}{R^2}\tag{12.5.309}$$

where again we put here $k=-1$ and we have shown that when vacuum dominates, we had:

$$R^0\rho=R_0^0\rho_0=\rho_{V,0}\tag{12.5.310}$$

Since then:

$$\frac{R'^2}{R^2}=\frac{8\pi G\rho_{V,0}}{3}+\frac{1}{R^2}\tag{12.5.311}$$

and it comes immediately:

$$\frac{\mathrm{d}R}{\mathrm{d}t}=\sqrt{\frac{8\pi G\rho_{V,0}R^2}{3}+1}\tag{12.5.312}$$

Therefore:

$$\int\mathrm{d}t=\int\frac{1}{\sqrt{\dfrac{8\pi G\rho_{V,0}R^2}{3}+1}}\mathrm{d}R\tag{12.5.313}$$

We know have already derived the primitive above during our study of usual primitives in the section of Integral and Differential Calculus at page 1534.

Therefore we get:

$$t=\sqrt{\frac{3}{8\pi G\rho_{V,0}}}\ln\left(\sqrt{\frac{8\pi G\rho_{V,0}}{3}}R+\sqrt{\frac{8\pi G\rho_{V,0}}{3}R^2+1}\right)\tag{12.5.314}$$

Taking the exponential of the above relation, squaring and solving the second degree polynomial relatively to R gives:

$$R=\frac{1}{2}\frac{e^{t^2\left(\frac{8\pi G\rho_{V,0}}{3}\right)^2}-1}{\sqrt{\frac{8\pi G\rho_{V,0}}{3}}e^{t\sqrt{\frac{8\pi G\rho_{V,0}}{3}}}}\tag{12.5.315}$$

The latter has a real solution for all $t \in \mathbb{R}$. With Maple™ 4.00b only for $t > 0$ (it's symmetrical in negative axis if we choose the right integration constant), we get roughly something like the following:

```
>plot(exp(t^2)-1)/exp(t),t=0..3);
```

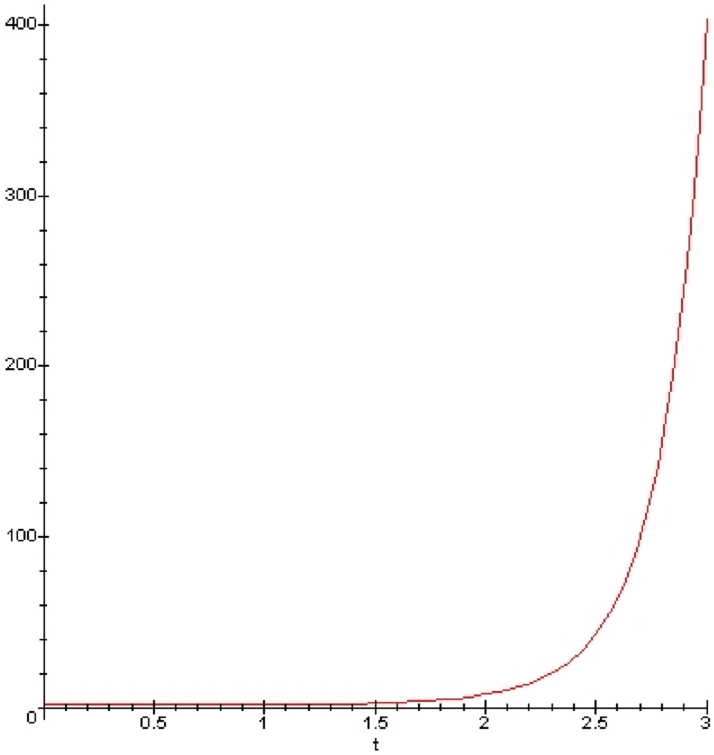


Figure 12.271: **Evolution of R for a hyperbolic space dominated by vacuum**

Finally to summarize a little better all the models seen so far (excepted those dominated by vacuum) with legends we get the following important plot:

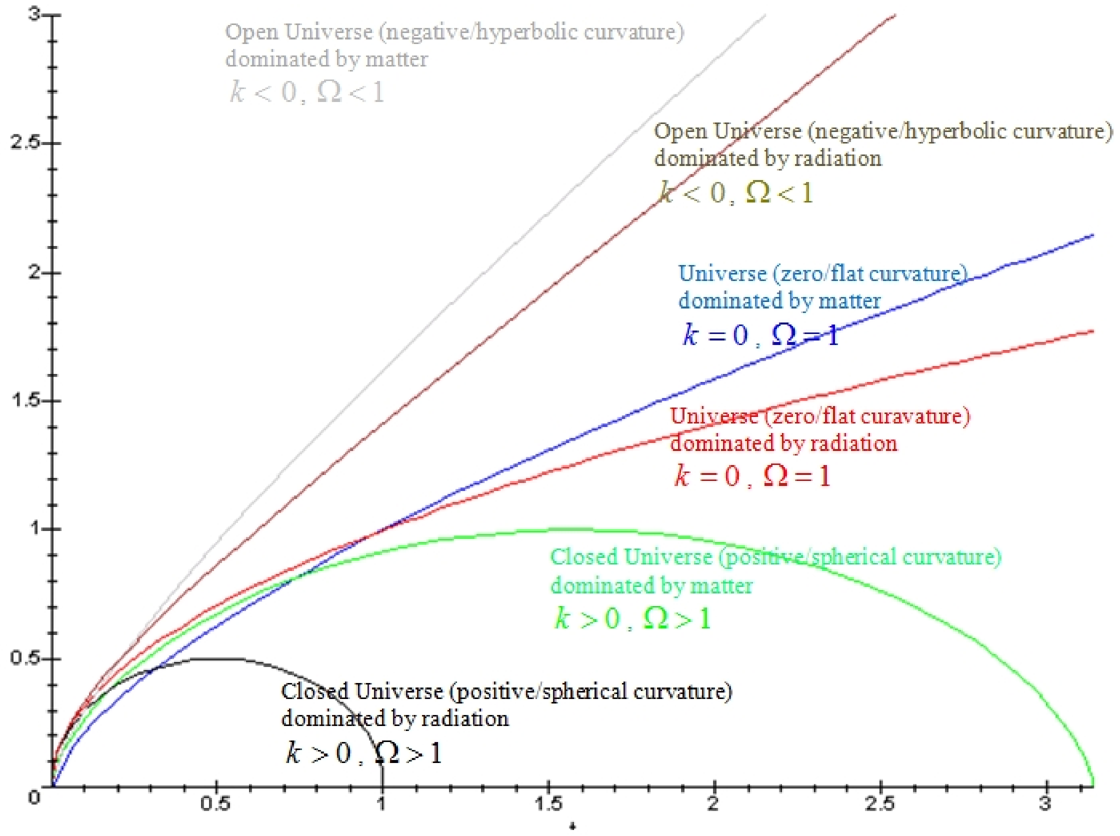


Figure 12.272: Summary of Newtonian Universe models

So we have here above six of the nine models of Universe (the three vacuum solutions are not represented) based on our mathematical assumptions and simplifications.

💣 Caution! Keep in mind that we have the knowledge now with everything that we have seen so far of the following canonical curvature values $k=\{-1,0,+1\}$, of the following typical canonical values of the density parameter $\omega(T)=\{0,1/3,-1/3,1,<-1\}$. Hence in reality there are $3\times 5=15$ different type of dominated universes and not only 9! If we take into account that a universe can contain a mixed of radiation, vacuum, matter, etc. we get in reality $3\times 31=93$ possible mixtures of universes. Obviously we will never study all of them in this textbook!

12.5.4 Observable Universe

We have determined previously a possible interpretation of the current estimate of the age (horizon) of our Universe as being as the inverse of the Hubble parameter that has given us for recall the following value (Hubble time):

$$t_H=\frac{1}{H_0}\cong 13\cdot 10^9\ \text{[year]} \qquad (12.5.316)$$

and the corresponding radius for the Hubble sphere (radius of the observable observer for which every observer is the center)

$$r_{\text{HS}}(t_0)=\frac{c}{H_0}\cong 4430.86\,[\text{Mpc}]\cong 14.4\cdot 10^9\,[\text{light-years}] \tag{12.5.317}$$

And afterwards, under the same strong assumptions[63], we have calculated the Newtonian radius of our Universe (observable part + non-observable part of our Universe)

$$r_{\text{Newt. Universe}}(t_0)\cong 28.8\,[\text{light-years}] \tag{12.5.318}$$

The illustration below typically depicts the fact that the observable Universe is centered on us (assimilated to our Sun) and that the bound of this observable Universe is highly shifted to the red.

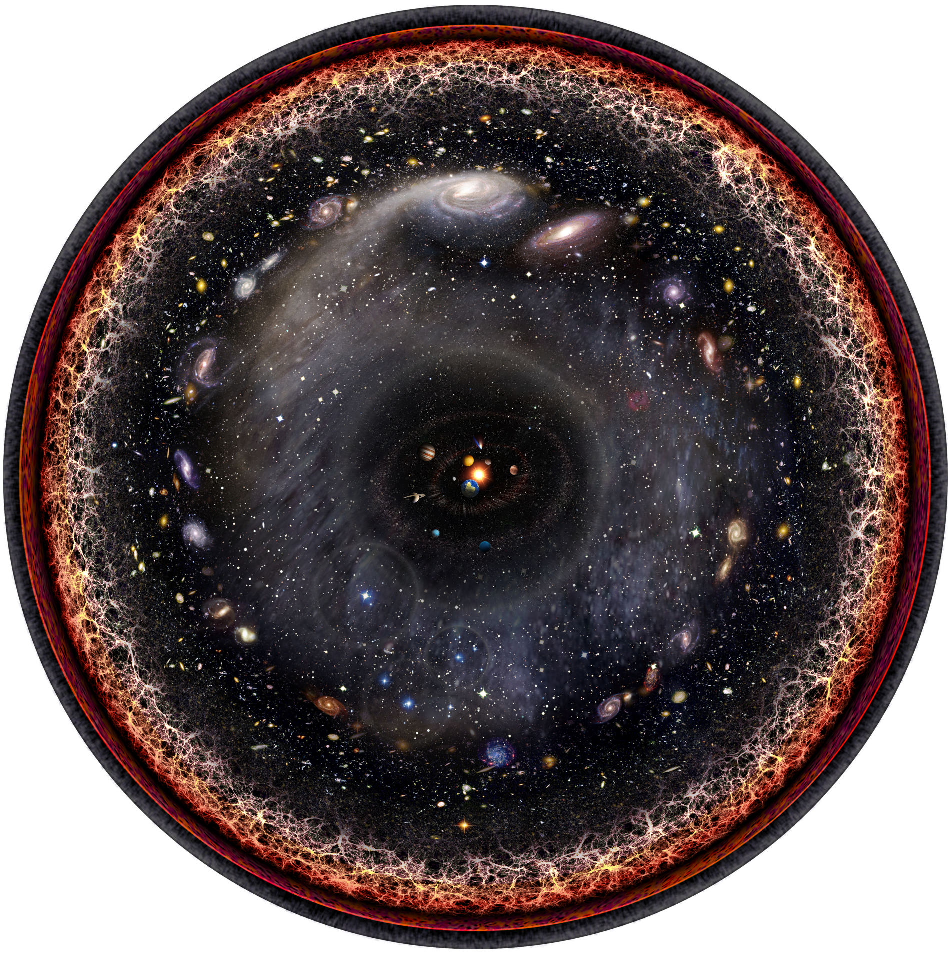


Figure 12.273: Artist's logarithmic scale conception of the observable Universe (source: Wikipedia, author: Pablo Carlos Budassi)

Remarks

R1. It is important to know that popular research articles in cosmology often use the term "Universe" in the sense of "observable Universe".

---

[63] Don't forget that we will see further below a method to estimate the comobile diameter using General Relativity!

R2. People should be more rigorous when speaking of the Universe "age". In fact, we should rather just say that the "horizon of the Universe" is almost 13 billion years. In other words, it is the time that someone would have measured if he has remained an inertial observer (in free fall: not subjected to any force other than gravity) throughout the evolution of the Universe and in a repository such that he would always perceived this Universe as homogeneous and isotropic.

The word "observable" used in this sense does not depend on whether modern technology actually permits detection of radiation from an object in this region (or indeed on whether there is any radiation to detect). It simply indicates that it is possible in principle for light or other signals from the object to reach an observer on Earth. In practice, we can see light only from as far back as the time of photon decoupling in the recombination epoch. That is when particles were first able to emit photons that were not quickly re-absorbed by other particles. Before then, the Universe was filled with a plasma that was opaque to photons. The detection of gravitational waves indicates there is now a possibility of detecting non-light signals from before the recombination epoch.

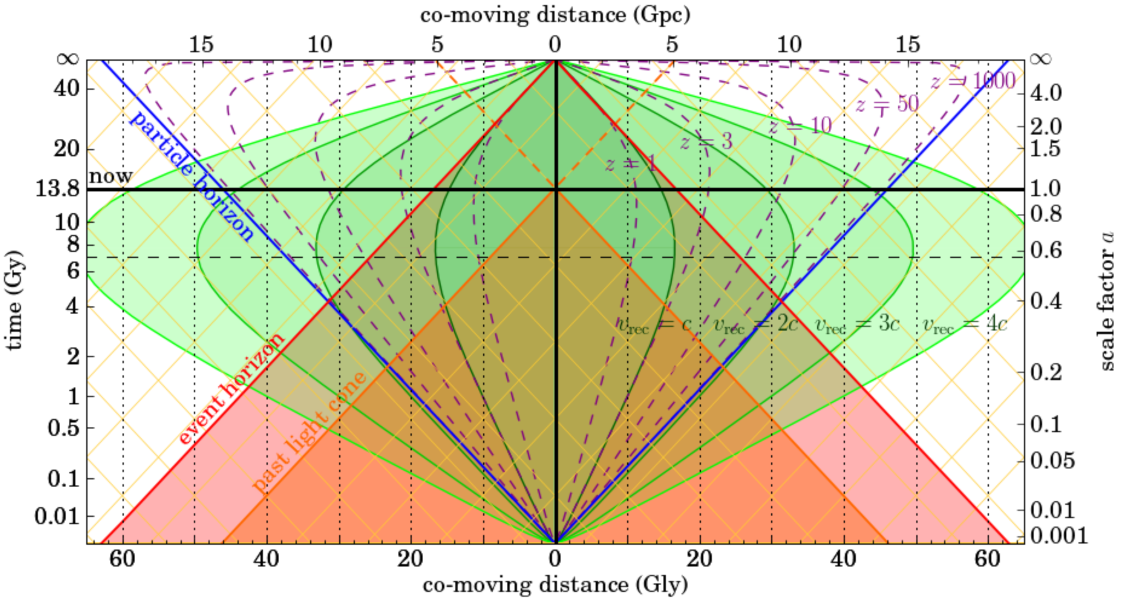
At the beginning of the early 121st century (holocene calendar), we do still do not know if the Universe is finite or infinite, although the majority of theorists currently favour an infinite universe.

The observable Universe is thus composed of all locations that could have affected us since the Big Bang (beware!... despite its name, the Big Bang theory has nothing to say on its start! It merely describes the evolution and the expansion of the Universe).

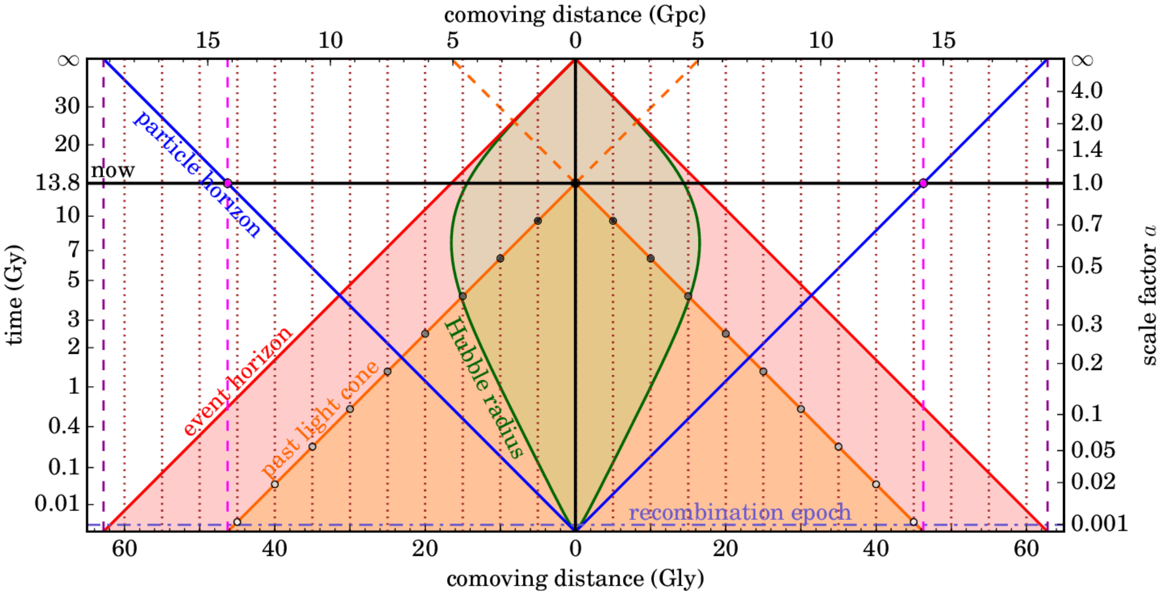
The current size (the "comoving distance") of the observable Universe is larger as we have calculated just earlier before and the corresponding sphere is named the "cosmological sphere".

This value can be obtained by taking the actual most distant visible object which is 13.39 billion light years of Earth (galaxy GN-z11). This will therefore have needed 13.39 billion light years to get away from us, its light will have needed 13.39 billion years to reach us and during the time of light travel, it will have move away of 13.39 billion light years (since objects at cosmological horizon are going at the speed of light). Thus a total of about 40 light billion years (see the detailed mathematical reasoning involving General Relativity further below at page 6048!).

The diagram below (inspired from an article of from Davis & Lineweaver) summarizes many measures of cosmic expansion in a nice, compact, but not necessarily easy-to-understand form (we will give just below a few example on how to read it):



Or a simplified version that has less stuff:



So let us explain how to read the above diagram. First of all, time is going from bottom to top. The thick horizontal black line represents the moment of **now**. Imagine this line moving upwards as time progresses.

The thick vertical black line is *here*. So the intersection of the two thick black lines in the middle is the "*here*-and-**now**".

Distances are measured in terms of the comoving distance, which is basically telling you how far a distant object would be now, if you had a long measuring tape to measure its present-day location.

The area shaded red (marked "past light cone") is all the events that happened in the universe that we could see, up to the moment of now. The boundary of this area is everything in this universe from which light is reaching us right now.

The green curve is the Hubble radius. Regions of space inside it (green shaded area) have recession velocities smaller than c, regions outside have recession velocities bigger than c.

So just for fun, let us pick an object at a comoving distance of 30 gigalightyears (Gly). Look at the dotted vertical line corresponding to 30 [Gly], halfway between the 20 and 40 marks (either side, doesn't matter). It intersects the boundary of the past light cone when the universe was roughly 700 million years old. Good, there were already young galaxies back then. If we were observing such a galaxy today, we'd be seeing it as it appeared when the universe was 700 million years old. Its light would have spent 13.1 billion years traveling before reaching our instruments.

Again look at the dotted vertical line at 30 [Gly] and extend it all the way to the **now** line. What does this tell you about this object? You can read the object's redshift (z) off the diagram: its light is shifted down in frequency by a factor of about 9.

You can also read the object's recession velocity v_{rec}, which is just a little over two times the vacuum speed of light. Yes… faster than light! This recession velocity is based on the rate of change of the scale factor, essentially the Hubble parameter times the comoving distance. The Doppler velocity that one would deduce from the object's redshift yields a value less than the vacuum speed of light (curved space-time is tricky; distances and speeds can be defined in various ways).

Another thing about this diagram is that in addition to the past, it also sketches the future, taking into account the apparent accelerating expansion of the universe. Notice the light red shaded area marked "event horizon". This area contains everything that we will be able to see at our present location, throughout the entire history of the universe, all the way to the infinite future. Things (events) outside this area will never be seen by us, will never influence us.

Note how the dotted line at 30 [Gly] intersects this boundary when the universe is about 5 billion years old. Yes, this means that we will only ever see the first less than 5 billion years of existence of a galaxy at a comoving distance of 30 [Gly]. Over time, light from this galaxy will be redshifted ever more, until it eventually appears to "freeze" and disappears from sight, never appearing to become older than 5 billion years!

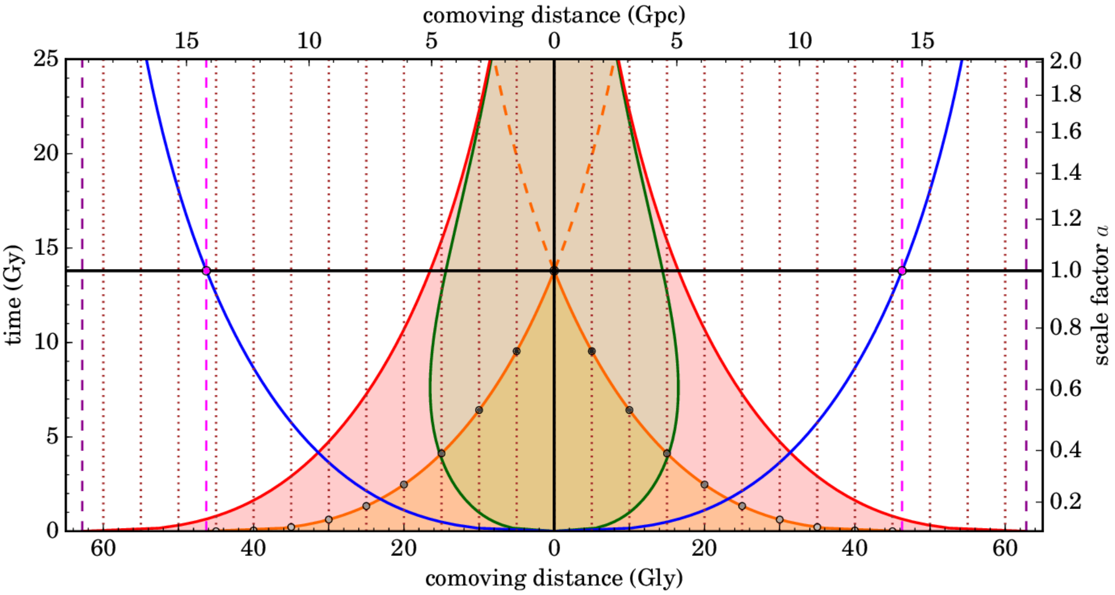
Notice also how the dashed curves marking constant values of redshift bend inward, closer and closer to the *here* location as we approach the infinite future. This is a direct result of accelerating expansion: Things nearer and nearer to us will be caught up in the expansion, accelerating away from our location. Eventually this will stop, of course; cosmic acceleration will not rip apart structures that are gravitationally bound. But we will end up living in a true "island universe" in which nothing is seen at all beyond the largest gravitationally bound structure, the local group of galaxies. Fortunately that won't happen anytime soon; we have many tens of billions of years until then.

Lastly, the particle horizon (blue lines) essentially marks the size of the visible part of the universe at any given time. Notice how the width of the interval marked by the intersection of the now line and the blue lines is identical to the width of the past light cone at the bottom of this diagram. Notice also how the blue lines correspond to infinite redshift.

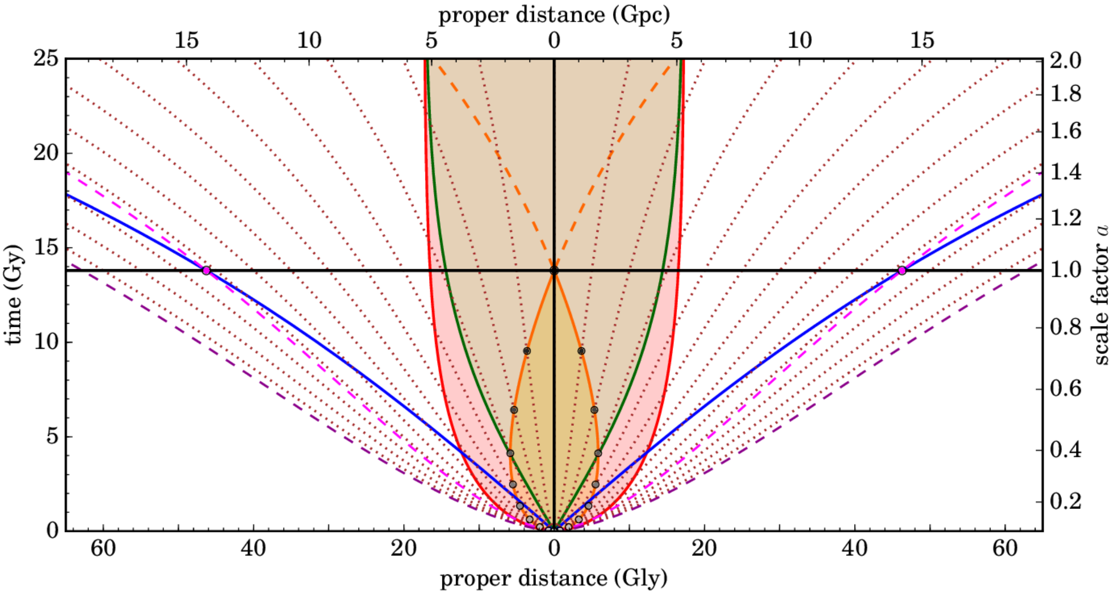
The diagram also shows that we can observe the light from emission events of high z sources (like early events on the worldline of comoving distance $= 30$ [Gly]) even as these sources may not have yet entered the particle horizon and are receding with $v_{\text{rec}} > c$.

The diagram also shows that farthest object whose photons are reaching here-and-now is around 46 billion light years far away now. Event horizon line touches 62 billion light years distance, it means particle horizon will be maximum 62 billion light years, doesn't matter if we will wait upto the infinite time, no photons will ever reach here farther than this distance just like current particle horizon is 46 billion light years.

If we choose linear vertical axis instead the both figure above become (which makes it for most people even more difficult to read):



And as by definition the comoving distance is defined as the proper distance divided by the scale factor, we can also draw the same diagram but based on the proper distance:



As we already said, this diagram is at first sight not an easy read but it is well worth studying.

This observable Universe contains according to today's heuristics estimates (year 12011 according to holocene calendar) about $7 \cdot 10^{22}$ stars, distributed in approximately 10^{10} galaxies, themselves organized into clusters and super-clusters of galaxies. The number of galaxies may be even larger, as the "Hubble Ultra-Deep Field" observed with the Hubble space telescope seems to indicate us.

The Hubble Ultra-Deep Field is an image of a region of the observable universe (equivalent sky area size shown in bottom left corner), near the constellation Fornax. Each spot is a galaxy, consisting of billions of stars. The light from the smallest, most red-shifted galaxies originated nearly 14 billion years ago.

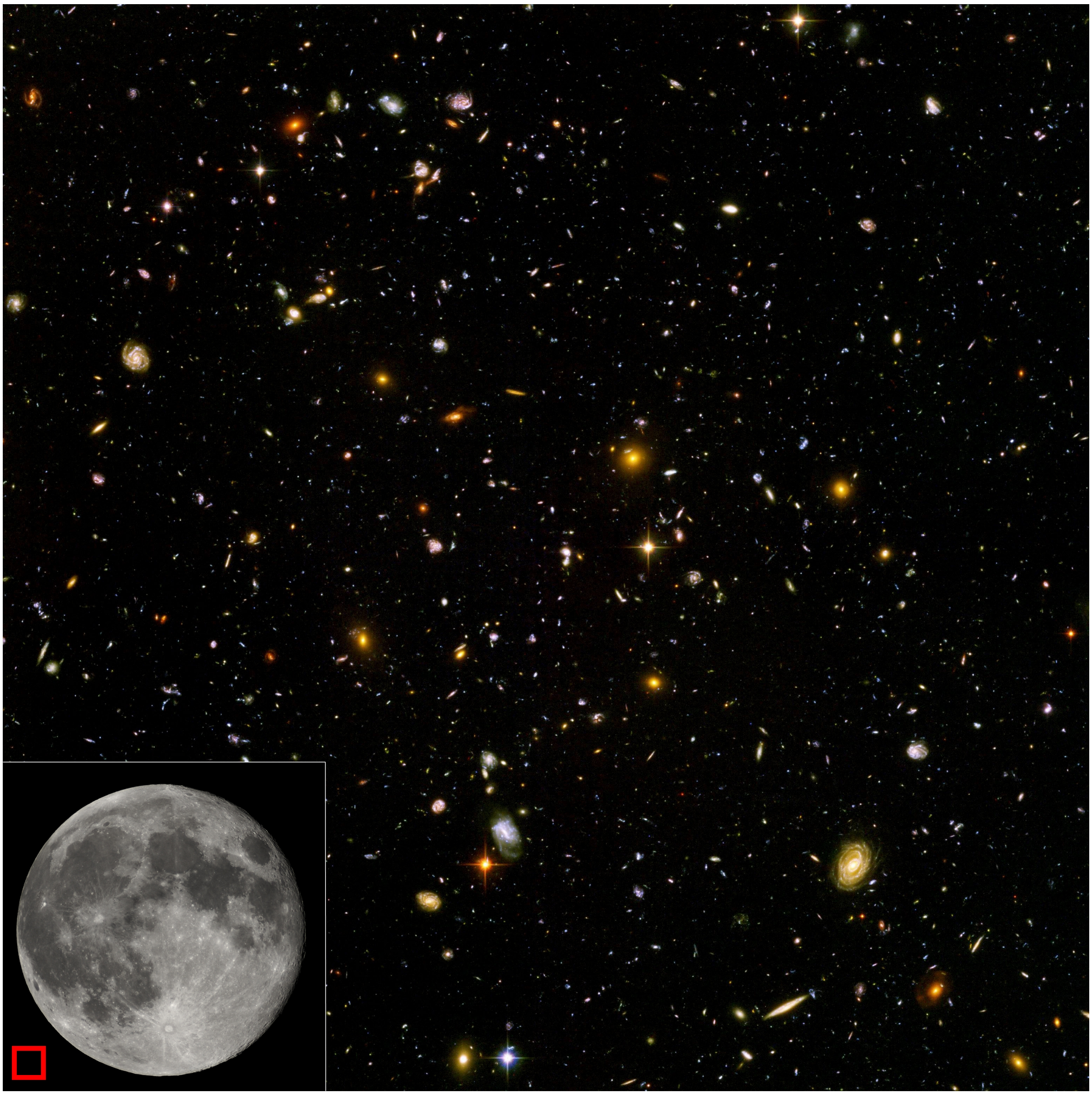


Figure 12.274: Hubble Ultra-Deep field (source: Wikipedia, author: NASA and ESA)

However it is difficult to imagine what that represents. As we found on the Internet some wonderful series of illustrations we would like to share them with the reader before they disappear from the Internet.

First here is a high-resolution summary of various structures you can find at different scales of our Universe (you can considerably zoom in!):

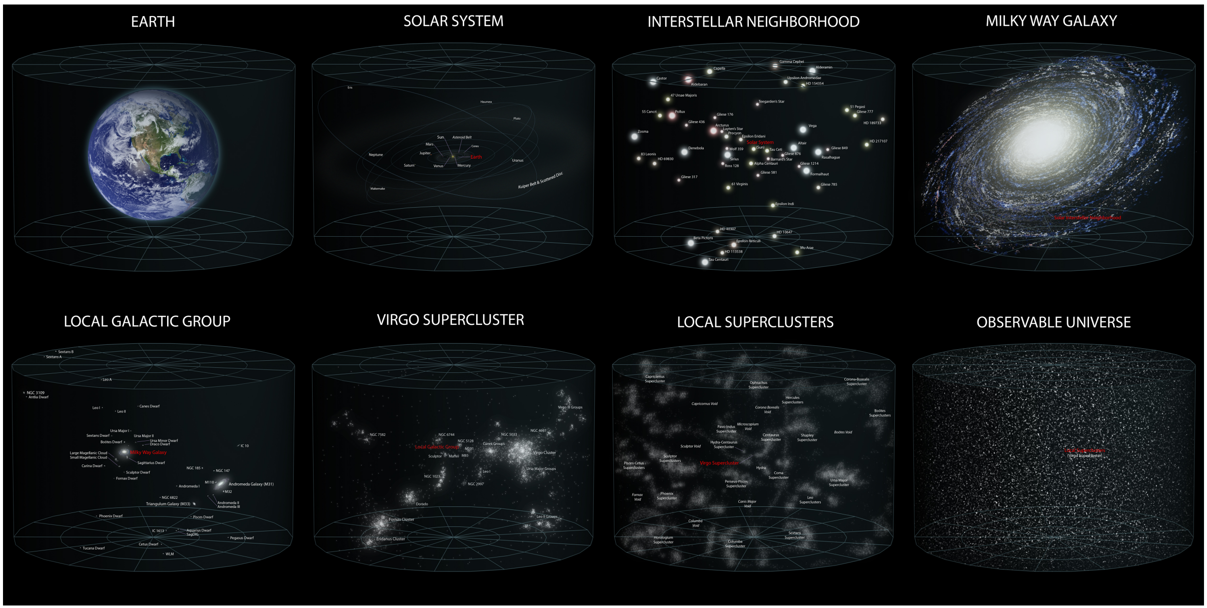

EARTH
SOLAR SYSTEM
INTERSTELLAR NEIGHBORHOOD
MILKY WAY GALAXY
LOCAL GALACTIC GROUP
VIRGO SUPERCLUSTER
LOCAL SUPERCLUSTERS
OBSERVABLE UNIVERSE

Figure 12.275: **Universe scales (source: Wikipedia, author: Andrew Z. Colvin)**

And a more detailed way to discover its structure:

01 The universe to 14 billion light years (the visible Universe as we approximately know it today):

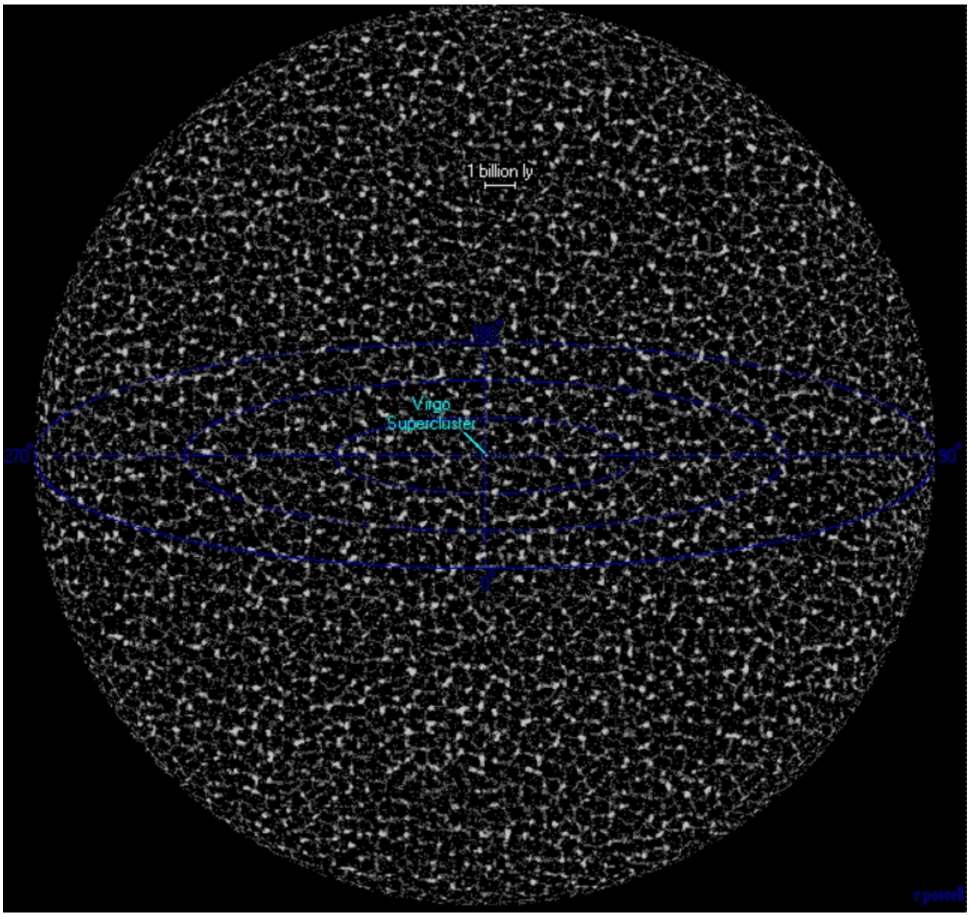


Figure 12.276: **Simplified illustration of the observable Universe**
(source: `http://atunivers.free.fr`, author: Richard Powell)

This illustration attempts to show the entire visible Universe. The galaxies in the universe tend to collect into vast sheets and "supercluster" of galaxies surrounding large vacuums zones, giving the universe a cellular appearance. Because light in the Universe only travel a finite speed, we see objects at the edge of the Universe as when it was very young, there is 14 billion years ago

Some numbers (estimates):

- Number of superclusters in the visible universe: 10 million
- Number of galaxy groups in the visible universe: 25 billion
- Number of large galaxies in the visible universe: 350 billion
- Number of dwarf galaxies in the visible universe: 7 trillion
- Number of stars in the visible universe: 30 billion trillion $(3 \cdot 10^{22})$

(02) After a $\times 14$ zoom we get the Universe withing a 1 billion light years, that is the neighboring superclusters:

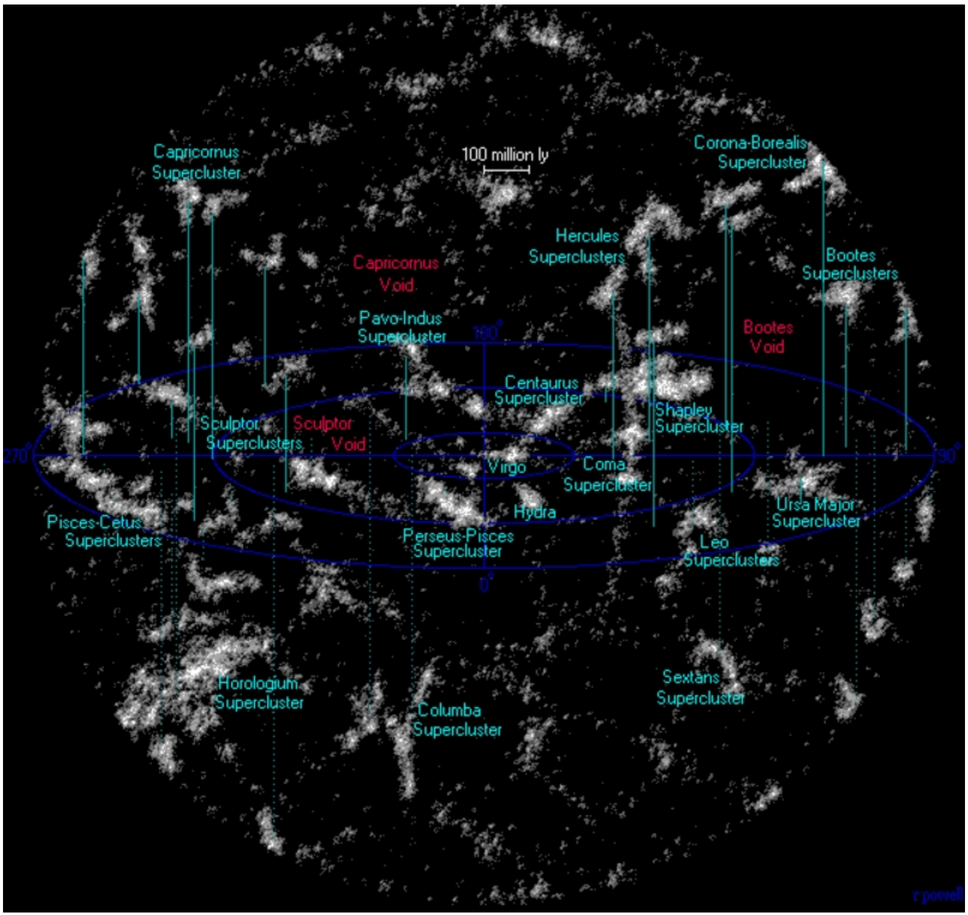

Capricornus Supercluster
100 million ly
Corona Borealis Supercluster
Hercules Superclusters
Capricornus Void
Bootes Superclusters
Bootes Void
Pavo-Indus Supercluster
180°
Centaurus Supercluster
Shapley Supercluster
Sculptor Superclusters
Sculptor Void
270°
Virgo
Coma Supercluster
90°
Hydra
Ursa Major Supercluster
Pisces-Cetus Superclusters
Perseus-Pisces Supercluster
Leo Superclusters
0°
Horologium Supercluster
Sextans Supercluster
Columba Supercluster
r.powell

Figure 12.277: Simplified illustration of the neighboring superclusters (source: `http://atunivers.free.fr`, author: Richard Powell)

Galaxies and clusters of galaxies are not distributed uniformly in the Universe. Instead, they gathered in large clusters, sheets and walls of galaxies separated by large gaps in which few galaxies appear to be. The illustration above shows a number of these super-clusters including the Virgo - a rather small super-cluster of which our Galaxy is part of. The entire map is approximately 7% of the diameter of the visible universe. The galaxies are too small to appear individually on this map, each point there is a group of galaxies.

Some numbers (estimates):

- Number of super-clusters within 1 billion light years: 100
- Number of galaxy groups within 1 billion light years: $240,000$
- Number of large galaxies within 1 billion light years: 3 million
- Number of dwarf galaxies within 1 billion light years: 60 million
- Number of stars within 1 billion light years: $250,000$ trillion

03 After a $\times 10$ zoom we get the Universe within 100 million light years or the Virgo super-cluster:

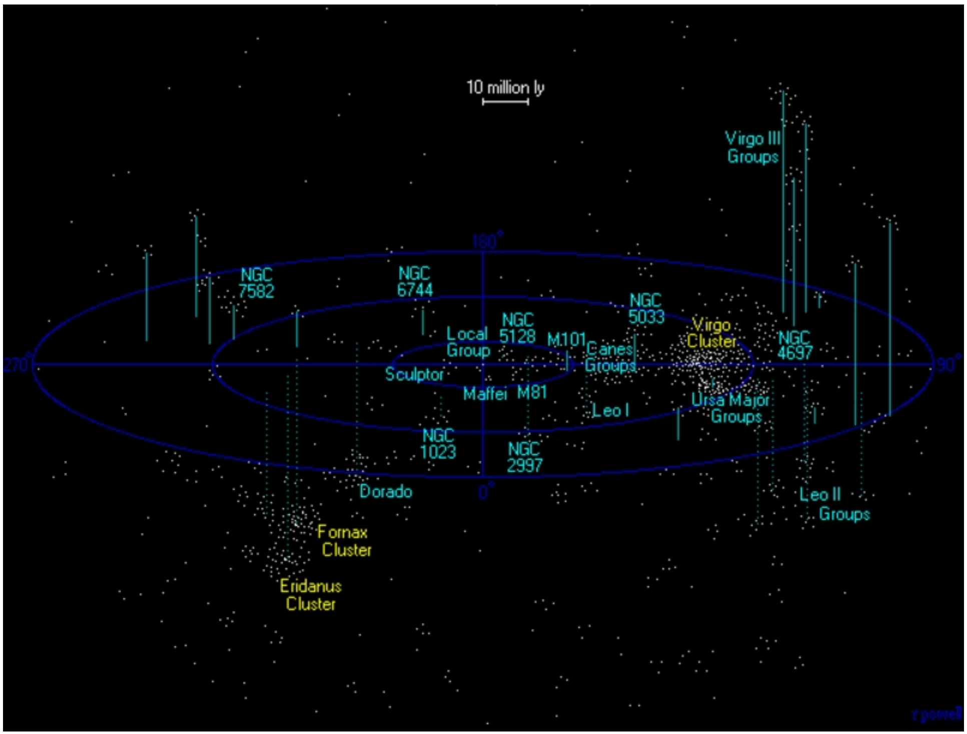

10 million ly
Virgo III Groups
NGC 7582
NGC 6744
NGC 5033
Virgo Cluster
NGC 4697
Local Group
NGC 5128
M101
Canes Groups
Sculptor
Maffei
M81
Leo I
Ursa Major Groups
NGC 1023
NGC 2997
Dorado
Leo II Groups
Fornax Cluster
Eridanus Cluster
180°
270°
90°
0°
r.powell

Figure 12.278: **Simplified illustration of the Virgo supercluster**
(source: `http://atunivers.free.fr`, author: Richard Powell)

Our Galaxy is just one of thousands that lie within 100 million light years. The above illustration shows how galaxies tend to gather in groups, the largest nearby cluster is the Virgo cluster (Virgo), a concentration of several hundred galaxies which dominates the surrounding groups of galaxies. Collectively, all of these groups is known to supercluster Virgo. The second richest cluster in this volume is the Fornax cluster (Fornax), but it is as rich as that of the Virgin. Only bright galaxies are drawn here, our Galaxy is the point at center.

Some numbers (estimates):

- Number of galaxy groups within 100 million light years: 200
- Number of large galaxies within 100 million light years: $2,500$
- Number of dwarf galaxies within 100 million light years: $50,000$
- Number of stars within 100 million light years: 200 trillion

(04) After a $\times 20$ zoom we get the Universe within 5 million light years, that is the Local Group of galaxies:

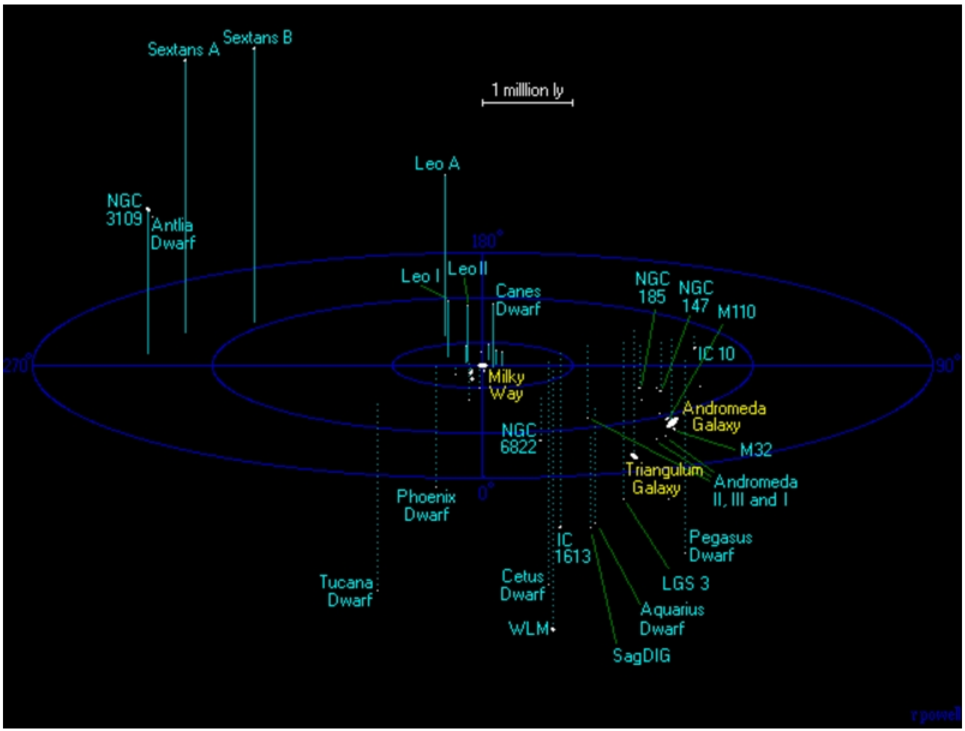

Sextans A
Sextans B
1 million ly
Leo A
NGC 3109
Antlia Dwarf
180°
Leo I
Leo II
Canes Dwarf
NGC 185
NGC 147
M110
IC 10
270°
90°
Milky Way
Andromeda Galaxy
NGC 6822
M32
Triangulum Galaxy
Andromeda II, III and I
Phoenix Dwarf
0°
IC 1613
Pegasus Dwarf
Tucana Dwarf
Cetus Dwarf
LGS 3
Aquarius Dwarf
WLM
SagDIG
r.powell

Figure 12.279: Simplified illustration of our Local Group
(source: `http://atunivers.free.fr`, author: Richard Powell)

The Milky Way is one of three large galaxies in the group named "Local Group" which also contains several dozen of dwarf galaxies. Most of these galaxies are plotted on the illustration above, but note that many of these dwarf galaxies a very small magnitude, so that there are certainly more to discover.

Some numbers (estimates):

- Number of large galaxies within 5 million light years: 3
- Number of dwarf galaxies within 5 million light years: 46
- Number of stars within 5 million light years: 700 billion

(05) After a $\times 10$ zoom we get the Universe within $500,000$ light years, that is the satellite galaxies:

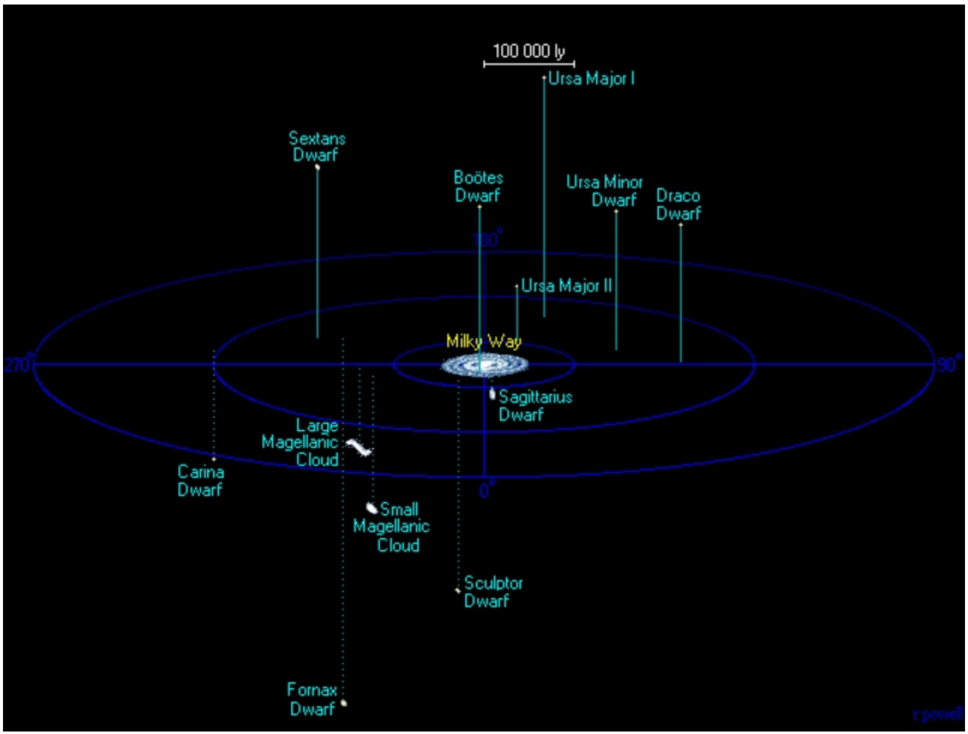

100 000 ly
Ursa Major I
Sextans Dwarf
Boötes Dwarf
Ursa Minor Dwarf
Draco Dwarf
180°
Ursa Major II
Milky Way
270°
90°
Sagittarius Dwarf
Large Magellanic Cloud
Carina Dwarf
0°
Small Magellanic Cloud
Sculptor Dwarf
Fornax Dwarf
r.powell

Figure 12.280: Simplified illustration of satellite galaxies
(source: `http://atunivers.free.fr`, author: Richard Powell)

The Milky Way is surrounded by several dwarf galaxies, each containing tens of millions of stars, which is insignificant compared to the population of the Milky Way itself. The map above shows all of the nearest dwarf galaxies that are gravitationally bound to the Milky Way, and revolve around it in a few billion years.

Some numbers (estimates):

- Number of large galaxies within $500,000$ light years: 1
- Number of dwarf galaxies within $500,000$ light years: 12
- Number of stars within $500,000$ light years: 225 billion billion

06 After a $\times 10$ zoom we get the Universe within $50,000$ light years, that is the Milky Way Galaxy:

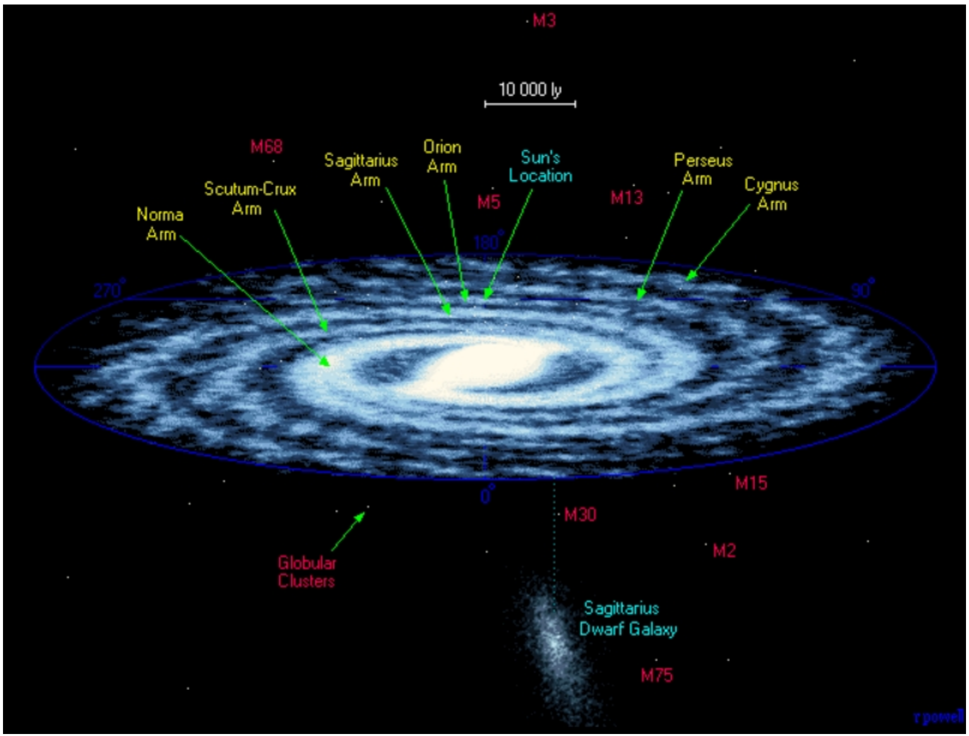

M3
10 000 ly
M68
Sagittarius Arm
Orion Arm
Sun's Location
Perseus Arm
Cygnus Arm
Scutum-Crux Arm
Norma Arm
M5
M13
180°
270°
90°
0°
M15
M30
M2
Globular Clusters
Sagittarius Dwarf Galaxy
M75
r.powell

Figure 12.281: Simplified illustration of the Milky Way Galaxy (source: http://atunivers.free.fr, author: Richard Powell)

This map shows the Milky Way as a whole - a spiral galaxy of at least two hundred billion stars. Our Sun (actually a G2V-type star, aka yellow dwarf) is buried deep within the Orion Arm about $26,000$ light years from the center. Toward the center of the galaxy, stars are much closer to each other than at the periphery where we live. Also notice the presence of small globular clusters far outside the galactic plane, and the presence of a neighbouring dwarf galaxy - named "Sagittarius" - which is slowly being swallowed by our own Galaxy.

Some numbers (estimates):

- Number of stars within $50,000$ light years: 200 billion billion

(07) After a $\times 10$ zoom we get the Universe within $5,000$ light years, that is to say the Orion Arm:

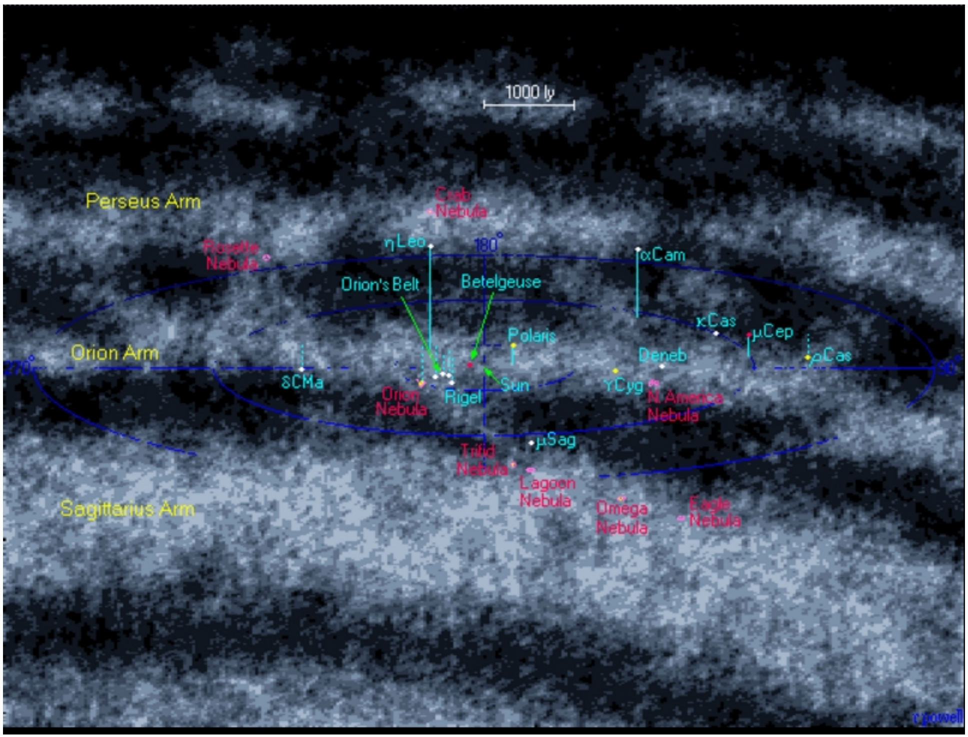

1000 ly
Perseus Arm
Crab Nebula
Rosette Nebula
ηLeo
180°
αCam
Orion's Belt
Betelgeuse
κCas
μCep
Polaris
Orion Arm
270°
δCMa
Deneb
ρCas
90°
Sun
γCyg
Orion Nebula
Rigel
N. America Nebula
μSag
Trifid Nebula
Lagoon Nebula
Omega Nebula
Eagle Nebula
Sagittarius Arm
c.powell

Figure 12.282: **Simplified illustration of the Orion Arm**
(source: `http://atunivers.free.fr`**, author: Richard Powell)**

This is a map of our corner of the Milky Way. The Sun is located in the Orion Arm - a fairly small arms compared to the Sagittarius Arm, which is closer to the galactic center. The map shows several stars visible to the naked eye, located far away in the Orion arm. The most notable group of stars is composed of the main stars of the constellation of Orion - from which the spiral arm gets its name. All these stars are bright giant and supergiant stars, thousands of times more luminous than the Sun. The brightest star of the map is Rho Cassiopeia - to $4,000$ light-years from us is just barely visible to the naked eye star, but in reality it is a supergiant $100,000$ times brighter than our Sun.

Some numbers (estimates):

- Number of stars within $5,000$ light years: 600 million

08 After a $\times 20$ zoom we get the Universe within 250 light years, that is to say the solar neighbourhood:

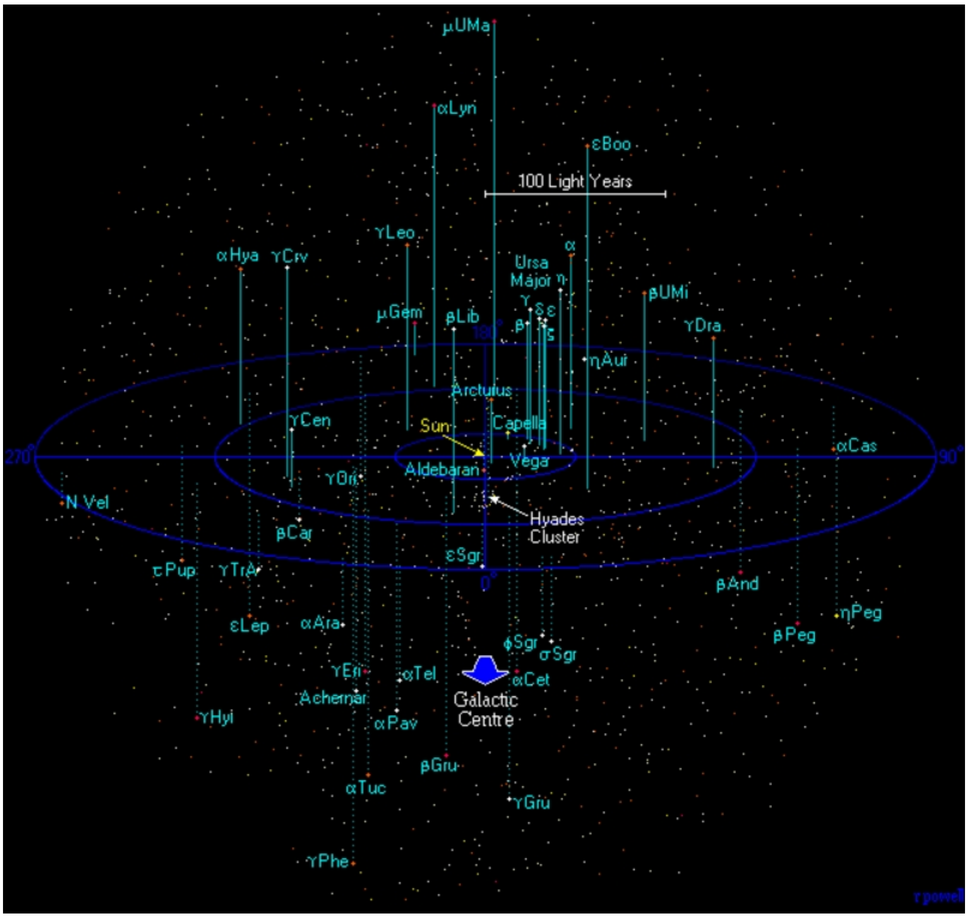

μUMa
αLyn
εBoo
100 Light Years
γLeo
αHya
γCrv
α
Ursa Major
η
μGem
βLib
γ
δ ε
β
ζ
βUMi
γDra
ηAur
Arctuius
γCen
Sun
Capella
αCas
270°
90°
Vega
γOri
Aldebaran
N Vel
Hyades Cluster
βCar
εSgr
τPup
γTrA
0°
βAnd
ηPeg
εLep
αAra
βPeg
φSgr
σSgr
γEri
αTel
αCet
Achernar
Galactic Centre
γHyi
αPav
βGru
αTuc
γGru
γPhe
r powell

Figure 12.283: **Simplified illustration of the solar neighbourhood (source: http://atunivers.free.fr, author: Richard Powell)**

This map shows the 1, 500 most luminous stars within 250 light years. All these stars are much more luminous than the Sun, and most are visible to the naked eye. About a third of the stars visible to the naked eye are within 250 light years, even though that area represents only a small part of our Galaxy.

Some numbers (estimates):

- Number of stars within 250 light years: 260, 000

09 After a ×20 zoom we get the Universe within 12.5 light years (the nearest stars):

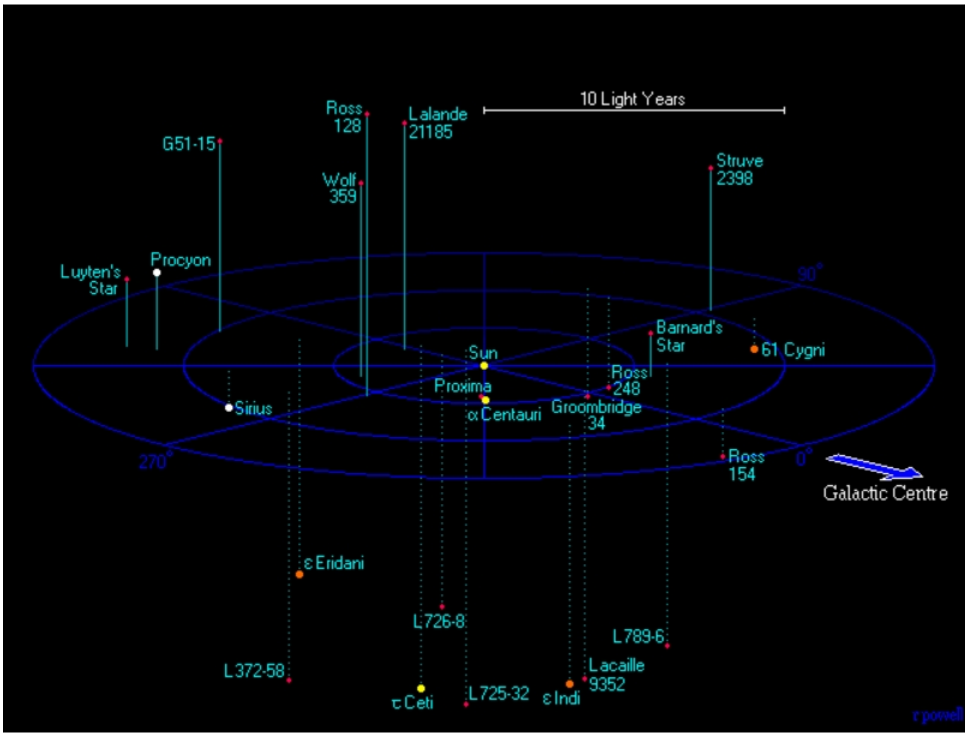
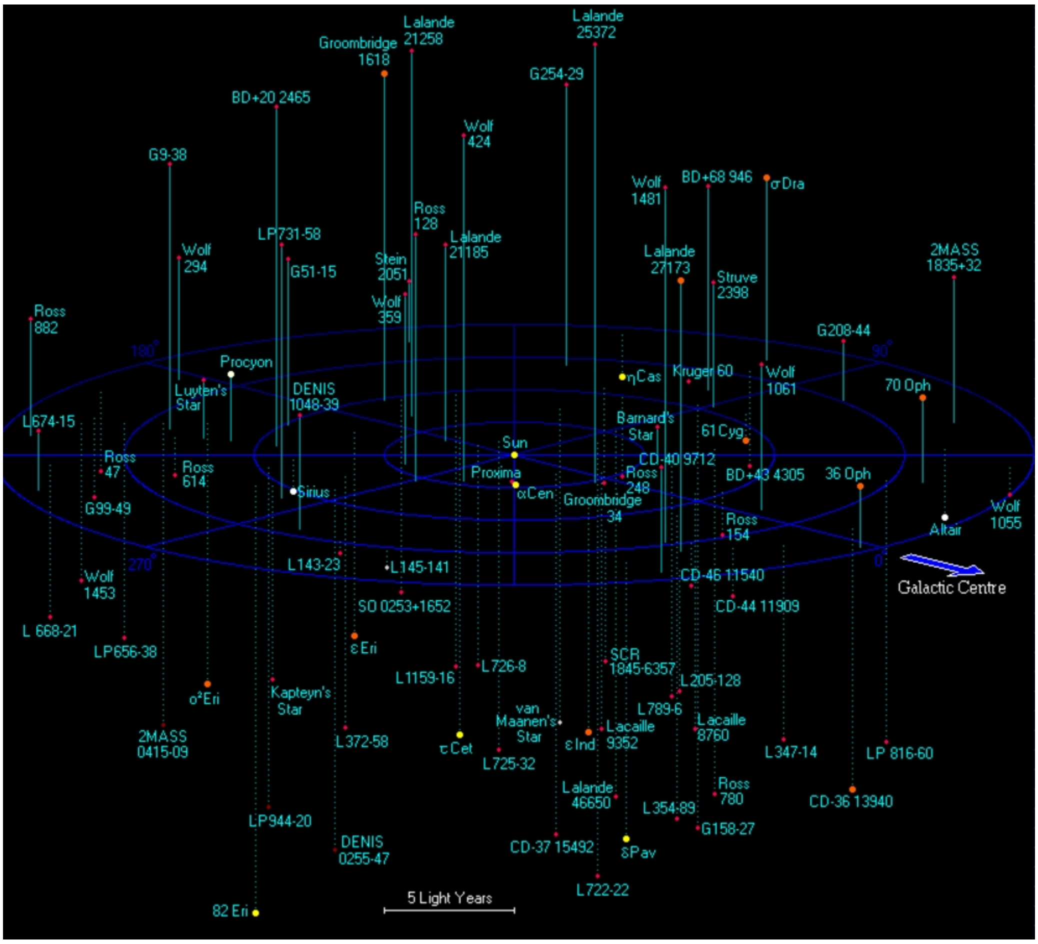

10 Light Years
Ross 128
Lalande 21185
G51-15
Wolf 359
Struve 2398
Luyten's Star
Procyon
90°
Barnard's Star
61 Cygni
Sun
Proxima
Ross 248
Sirius
α Centauri
Groombridge 34
270°
Ross 154
0°
Galactic Centre
ε Eridani
L726-8
L789-6
Lacaille 9352
L372-58
L725-32
ε Indi
τ Ceti
r powell

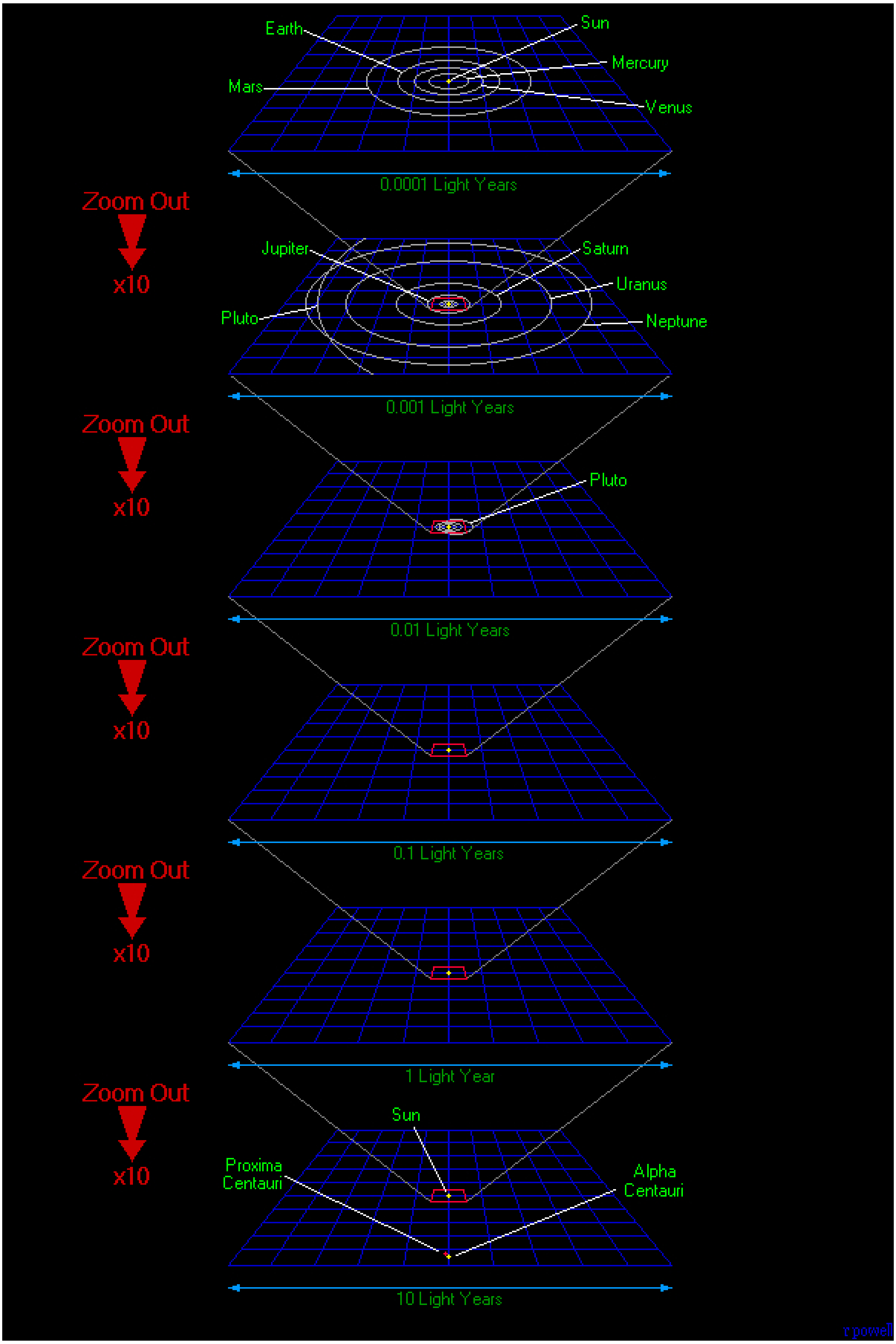
Figure 12.284: Simplified illustration of the nearest stars
(source: `http://atunivers.free.fr`, author: Richard Powell)

This map shows some stars up to a distance of 12.5 light years from our Sun (there would be 33 identified to this date). Most of these stars are red dwarfs - stars with a tenth of the mass of the Sun and a hundred times less bright. About 80% of stars in the Universe are red dwarfs, and the nearest star - Proxima Centaure - is a typical example.

The map below show all known stars within 20 light years. There are a total of 77 systems containing 110 stars:


Lalande 25372
Lalande 21258
Groombridge 1618
G254-29
BD+20 2465
Wolf 424
G9-38
Wolf 1481
BD+68 946
σ Dra
Ross 128
LP731-58
Lalande 21185
Wolf 294
Stein 2051
G51-15
Lalande 27173
2MASS 1835+32
Struve 2398
Ross 882
Wolf 359
G208-44
180°
90°
Procyon
ηCas
Kruger 60
Wolf 1061
Luyten's Star
DENIS 1048-39
70 Oph
L674-15
Barnard's Star
61 Cyg
Sun
CD-40 9712
Ross 47
Ross 614
Proxima
Ross 248
BD+43 4305
36 Oph
Sirius
αCen
Groombridge 34
G99-49
Wolf 1055
Ross 154
Altair
270°
L143-23
L145-141
0°
Wolf 1453
CD-46 11540
Galactic Centre
SO 0253+1652
CD-44 11909
L 668-21
LP656-38
ε Eri
L1159-16
L726-8
SCR 1845-6357
L205-128
o²Eri
Kapteyn's Star
van Maanen's Star
L789-6
Lacaille 9352
Lacaille 8760
2MASS 0415-09
L372-58
τ Cet
ε Ind
L347-14
LP 816-60
L725-32
Lalande 46650
Ross 780
CD-36 13940
L354-89
LP944-20
G158-27
DENIS 0255-47
CD-37 15492
δ Pav
L722-22
5 Light Years
82 Eri

The distances between stars are huge. The distance from the Sun to Proxima Centauri is 4.22 light years, or 40 trillion kilometres. Walk this distance would take a billion years. Even the fastest space probes in this early 121st century (according to holocene calendar) would need 6,000 years to make the trip. There are currently four probes leaving the Solar System - Pioneer 10 and 11 and Voyager 1 and 2 but we will likely lose contact with them within the next two years (if it's not already done when the reader see these lines). The following diagram attempts to show these distances by broadening the scope from the inner Solar System to Alpha Centauri:


Earth
Sun
Mercury
Mars
Venus
0.0001 Light Years
Zoom Out
x10
Jupiter
Saturn
Uranus
Pluto
Neptune
0.001 Light Years
Zoom Out
x10
Pluto
0.01 Light Years
Zoom Out
x10
0.1 Light Years
Zoom Out
x10
1 Light Year
Zoom Out
x10
Sun
Proxima
Centauri
Alpha
Centauri
10 Light Years
r.powell

And also a beautiful infographics from National Geographic:

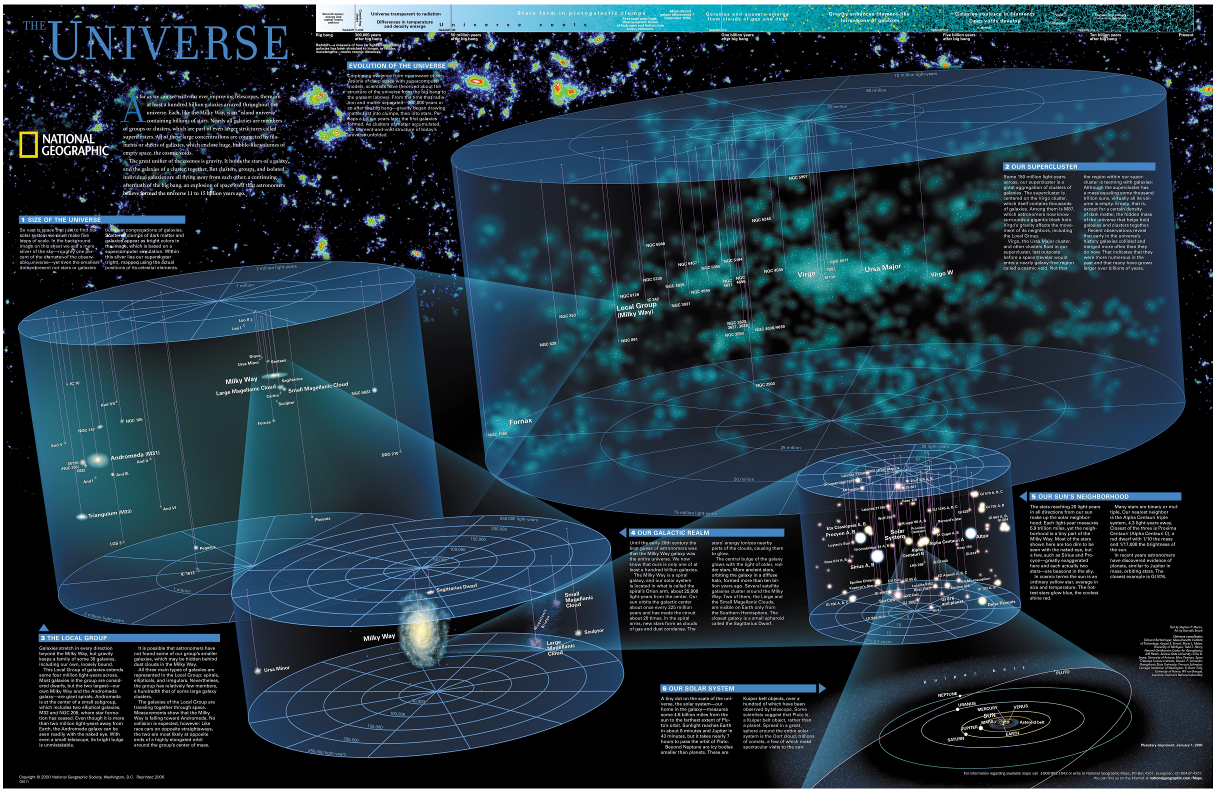

THE UNIVERSE
NATIONAL GEOGRAPHIC
EVOLUTION OF THE UNIVERSE
1 SIZE OF THE UNIVERSE
2 OUR SUPERCLUSTER
3 THE LOCAL GROUP
4 OUR GALACTIC REALM
5 OUR SUN'S NEIGHBORHOOD
6 OUR SOLAR SYSTEM
Virgo
Ursa Major
Virgo W
Local Group (Milky Way)
Fornax
Milky Way
Large Magellanic Cloud
Small Magellanic Cloud
Andromeda (M31)
Triangulum (M33)
Sagittarius Dwarf
Ursa Minor
Sculptor
Solar System
Sirius A, B
Procyon A, B
Altair
Kuiper belt
PLUTO
NEPTUNE
URANUS
SATURN
JUPITER
MERCURY
VENUS
SUN
MARS
EARTH

Figure 12.285: **Observable Universe map (source: National Geographic)**

12.5.5 Cosmic Microwave Background (CMB)

We have already mentioned the cosmic microwave background earlier above and we have given numerous illustrations of it and the corresponding satellites names and observations related to its experimental study. But let us now give a more detailed picture of the cosmic microwave background to show to the reader that the latter is quite isotropic:

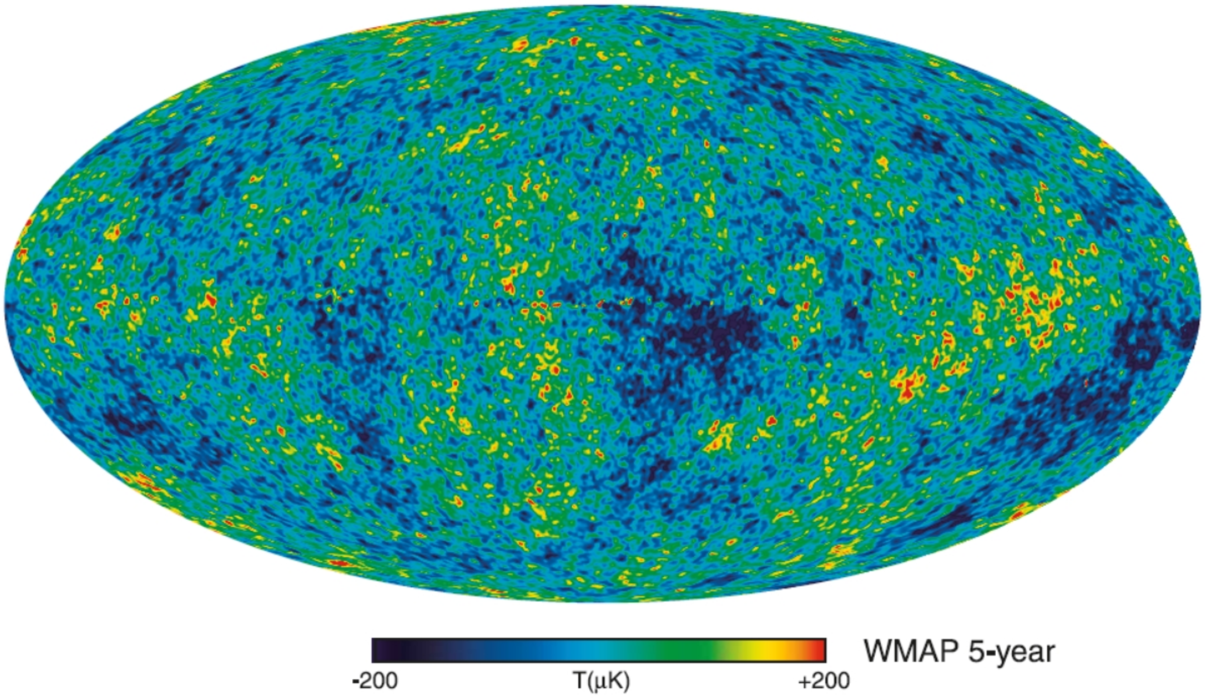


Figure 12.286: **Cosmic microwave background according to WMAP 5-year results (source: NASA/WMAP Science Team)**

To measure the CMB, astronomers process the incoming signals to remove interference from our Galaxy and other cosmic sources as illustrated below:

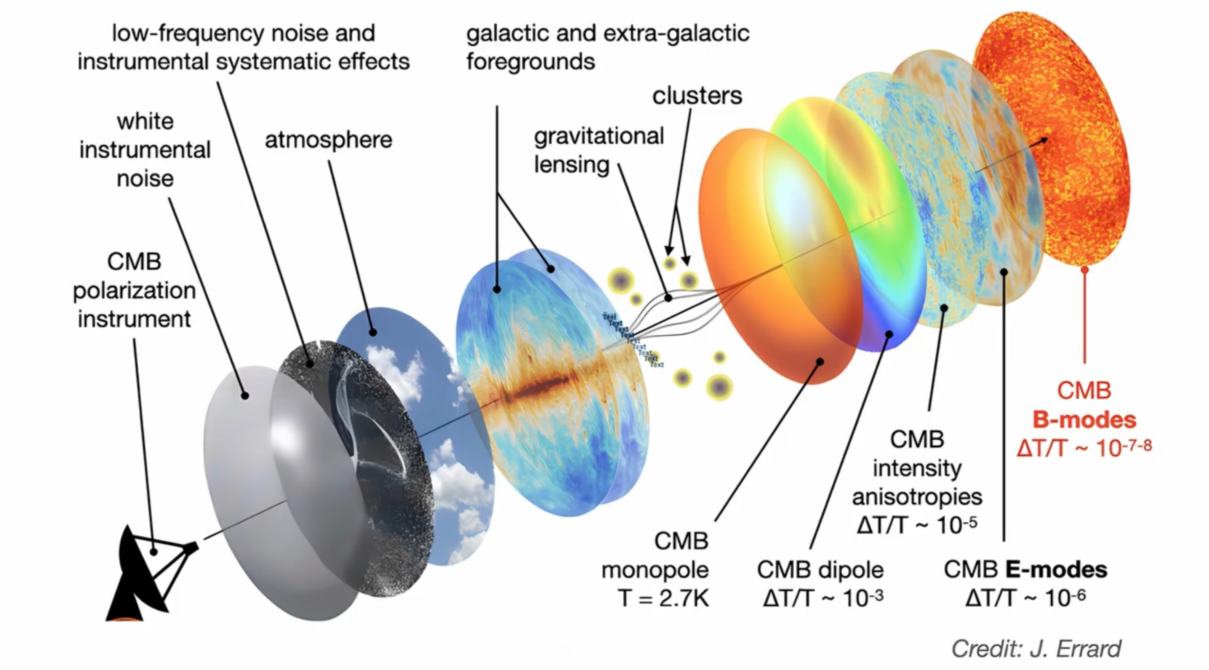

low-frequency noise and instrumental systematic effects
galactic and extra-galactic foregrounds
clusters
white instrumental noise
atmosphere
gravitational lensing
CMB polarization instrument
CMB B-modes ΔT/T ~ 10-7-8
CMB intensity anisotropies ΔT/T ~ 10-5
CMB monopole T = 2.7K
CMB dipole ΔT/T ~ 10-3
CMB E-modes ΔT/T ~ 10-6
Credit: J. Errard

Figure 12.287: CMB experiment looking a maze of contaminants (credit: Josquin Errard)

The temperature variations in the CMB are uniform up to about 1 part in $100,000$. So the temperature has range of $2.726\,[\text{K}] \pm 0.00003$ (the root mean square variation is $18\,[\mu\text{K}]$ as reported in [581]). This tells us that the early Universe was rather homogeneous at the decoupling time the CMB was formed. The energy density of the CMB is $0.260\,[\text{eV}\cdot\text{cm}^{-3}$ ($4.17\cdot 10^{-14}\,[\text{J}\cdot\text{m}^{-3}]$) which yields about 411 photons per cubic centimetre.

Remark

A common question is « *How to we know that the WMAP measurement comes from the Big Bang and not something else?* ». First the reader must keep in mind that in Science must of the time we don't know nothing for sure, but we have just different levels of evidence.

First we know that this signal doesn't come from our atmosphere as WMAP is in space... There may be a large amount of diffuse matter in space, that then absorbs starlight from all directions and re-radiates it at a lower temperature! However astronomy has progressed to the point where we have measured the dust in our Galaxy, throughout the Universe, and surrounding the Solar System and we know that at the opposite of the CMB, the dust is not uniformly distributed, is not a perfect absorber (preferentially absorbing blue light and transmitting red light) and the amount of dust actually observed is insufficient to explain this radiation (the CMB signal is about 100 times larger than the background signal).

But maybe there's something out there, far beyond the galaxies we know, that's emitting an ultra-distant source of light (and therefore red-shifted!). After all, stars and galaxies appear to be everywhere, and the Sun is almost a perfect blackbody radiator. And the answer is in the question! Stars aren't perfect blackbody but the CMB is an almost perfect one as depicted in the plot below:

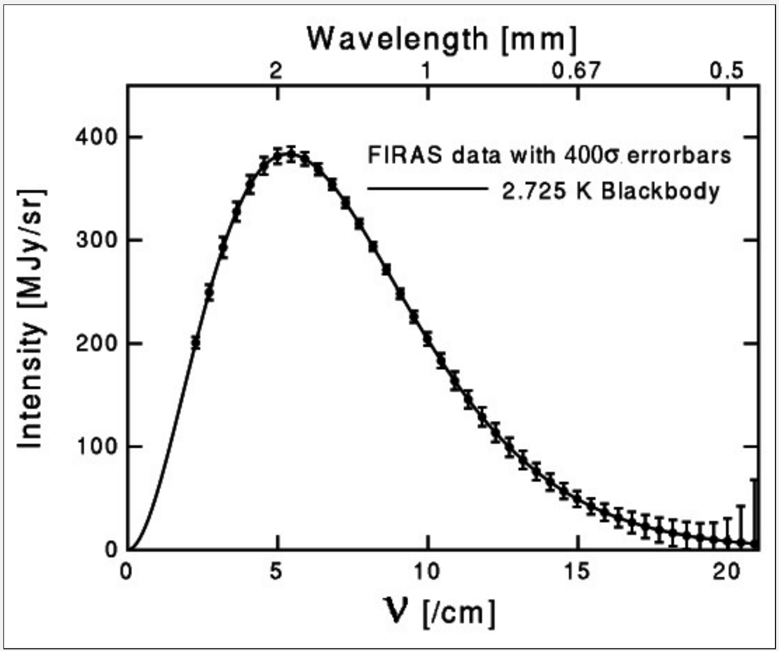


Figure 12.288: **FIRAS CMB Black Body radiation spectrum**

It isn't the atmosphere. It seems it isn't dust. It seems it isn't starlight. It seems it isn't the Oort cloud bodies (that should be at the order of 10 [K]). It isn't the heliopause (as the latter has an asymmetry because of Sun's directional movement). This background of radiation, more perfectly a blackbody in its spectrum than anything else in the Universe, must have its origin in a hot, dense state that existed billions of years ago... So it may very likely be the Big Bang (but without absolute certitude!).

The purpose now will be in what follows to have a mathematical approach of the cosmic microwave background!

12.5.5.1 Decoupling time

The existence and properties of the cosmic radiation discovered experimentally by Arno Penzias and Robert Woodrow Wilson in 11964 (holocene calendar) but theoretically by Ralph Apher in 11948 (holocene calendar) were mainly due to the two physical phenomena that we will now describe in broad outline.

The expansion of the Universe has for consequence that after the fantastically high values that have reigned immediately after the Big Bang it is gradual cooling! When it reaches about $3,000$ [K] occurs the first of two crucial phenomena that interest us here: the radiation, which until then was in thermal equilibrium with the material particles practically ceases to interact with them and became independent. In the "standard model" of evolution of the Universe, we calculate that this crucial moment is situated approximately $300,000$ years after the Big Bang (see the proofs further below).

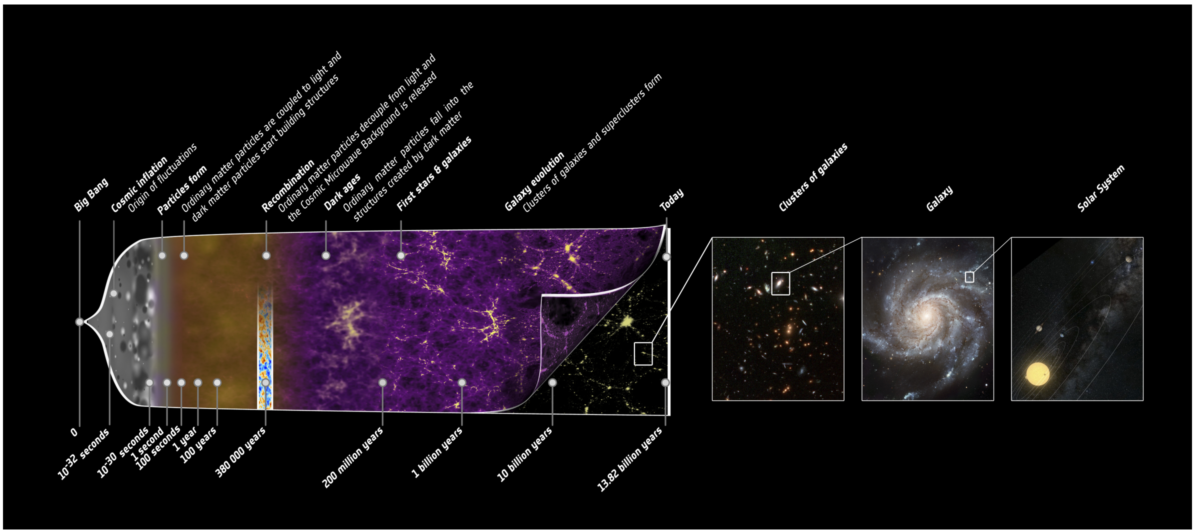

Big Bang
Cosmic inflation
Origin of fluctuations
Particles form
Ordinary matter particles are coupled to light and dark matter particles start building structures
Recombination
Ordinary matter particles decouple from light and the Cosmic Microwave Background is released
Dark ages
Ordinary matter particles fall into the structures created by dark matter
First stars & galaxies
Galaxy evolution
Clusters of galaxies and superclusters form
Today
Clusters of galaxies
Galaxy
Solar System
0
10^{-32} seconds
10^{-30} seconds
1 second
100 seconds
1 year
100 years
380 000 years
200 million years
1 billion years
10 billion years
13.82 billion years

Figure 12.289: 2D illustration of (speculated) Universe past history following 121st century hypothesis (source: ESA)

We can first qualitatively understand the physical reasons for this (further below we will approach this with maths stuff!). Shortly before, when for example the temperature was $100,000$ [K], the Universe contained mostly photons, electrons and bare atomic nuclei (mostly protons, and, to a lesser extent, α particles, helium 4 nuclei). The temperature was too high so that the electrons and nuclei may form stable atoms. The interaction between the photons and charged particles (mainly electrons, the lighter of them) is sufficiently intense, and the density of the latter was then sufficiently strong, so that the photons were continuously diffused, transmitted and absorbed. Despite its expansion, the Universe was at every moment at equilibrium; its temperature T was consistently well defined, although decreasing over time, the photon energy, that is to say the pulse of radiation, was therefore distributed according to Planck's law for this temperature T (see section Thermodynamics page 3719)!

The temperature decrease then permit the formation of atoms from the electrons and nuclei. This process led to a rapid drop in average cross section of interaction between photons and material particles (mainly due to the disappearance of free electrons), so that the Universe became transparent to photons. A quantitative assessment of the characteristics of the phenomenon is this decoupling occurred when the temperature dropped to $3,000$ [K] (see the simple mathematical approach further below).

At the moment of decoupling, the volume density of the radiation energy is distributed in the pulsations spectrum according to Planck's law (see section Thermodynamics page 3719):

$$R(\omega,T)=\frac{2\hbar\omega^3}{\pi c^3}\frac{1}{e^{\frac{\hbar\omega}{kT}}-1}\;[\mathrm{J}\cdot\mathrm{s}\cdot\mathrm{m}^{-3}] \qquad (12.5.319)$$

where we assume that T is the temperature ($3,000$ [K] approximately - ionization temperature of the simplest atoms[64]) at the moment of decoupling. This distribution will then evolve under the influence of the expansion of the Universe.

Let us consider the photons located at time t in the volume $4/3r^3\cong r^3$, and whose pulsation ω

---

[64] we will see further the detail treatment of where this value comes from.

with a variation $\mathrm{d}\omega$. Their number is then using previous relation equal to:

$$\mathrm{d}N\cong\frac{r^3}{\hbar\omega}R(\omega,T)=r^3\frac{2\omega^2}{\pi c^3}\frac{\omega^2\mathrm{d}\omega}{e^{\frac{\hbar\omega}{kT}}-1} \tag{12.5.320}$$

where we assume that T is the temperature ($3,000$ [K] approximately - ionization temperature of the simplest atoms) at the moment of decoupling. This distribution will then evolve under the influence of the expansion of the Universe.

As there is no absorption or emission of photons at this temperature (it is a hypothesis but as experimental measurements seem to confirm this model...), this number will remain constant. But because of the expansion of the Universe, these photons constant number will occupy a larger volume, and gain greater wavelength λ (following the expansion of the structure of space due to the positive value of the Hubble parameter) that is to say, a rather smaller pulsation ω (the equivalent of the Doppler effect). To clarify, let us consider the situation at a further time t'. All lengths of the Universe have increased between, between t and t', by the same scaling factor F following the Hubble's law: the chosen radius r of our previous selected sphere volume has therefore become obviously:

$$r'=Fr \tag{12.5.321}$$

and the wavelength of the photons considered:

$$\lambda'=F\lambda \tag{12.5.322}$$

so that their pulsation is equal at the instant t' to:

$$\omega^{=}\frac{2\pi c'}{\lambda'}=\frac{\omega}{F} \tag{12.5.323}$$

Therefore the energy contained at this time in the volume $V\cong r'^3$ and in the pulsation range $(\omega',\omega'+\mathrm{d}\omega')$ is given obviously by:

$$r'^3R'(\omega',T')\mathrm{d}\omega' \tag{12.5.324}$$

More explicitly:

$$r'^3R'(\omega',T')\mathrm{d}\omega=\hbar\omega'(\mathrm{d}N)=\hbar\omega'\left(\frac{2r^3}{\pi c^3}\frac{\omega^2\mathrm{d}\omega}{e^{\frac{\hbar\omega}{kT}}-1}\right) \tag{12.5.325}$$

The volumetric energy density $R'(\omega',T')\mathrm{d}\omega'$ at time t', for the pulsation range $(\omega',\omega'+\mathrm{d}\omega')$, is then written:

$$R'(\omega',T')\mathrm{d}\omega'=\hbar\omega'\frac{(\mathrm{d}N)}{r'^3}=\hbar\omega'\left(\frac{2r^3}{r'^3\pi c^3}\frac{\omega^2\mathrm{d}\omega}{e^{\frac{\hbar\omega}{kT}}-1}\right)=\hbar\omega'\left(\frac{2}{F^3\pi c^3}\frac{F\omega'^2F\mathrm{d}\omega'}{e^{\frac{\hbar F\omega'}{kT}}-1}\right) \tag{12.5.326}$$

It follows that the spectral energy distribution is still at the instant t' that of the black-body:

$$R'(\omega',T')=\frac{2\hbar\omega'^3}{\pi c^3}\frac{1}{e^{\frac{\hbar F\omega'}{kT}}-1}=\frac{2\hbar\omega'^3}{\pi c^3}\frac{1}{e^{\frac{\hbar\omega'}{kT'}}-1} \tag{12.5.327}$$

where the corresponding temperature T' is immediately such that:

$$T'=\frac{T}{F} \tag{12.5.328}$$

that is more often written:

$$\boxed{F=\frac{T_{\text{dec}}}{T_{\text{CMB}}}} \quad (12.5.329)$$

Thus, after decoupling with matter, the cosmic radiation evolves maintaining the distribution of a black-body whose temperature decreases regularly in the same proportion as the distances are increased during the expansion of the Universe.

From the moment of decoupling, the F factor of scale is very close to $1,000$ since from to the estimated $3,000$ [K] to go to the 2.7 [K] measured today there is a factor of:

$$F=\frac{T_{\text{dec}}}{T_{\text{CMB}}}=\frac{3,000}{2.7}\cong 1,100 \quad (12.5.330)$$

also named "recombination redshift".

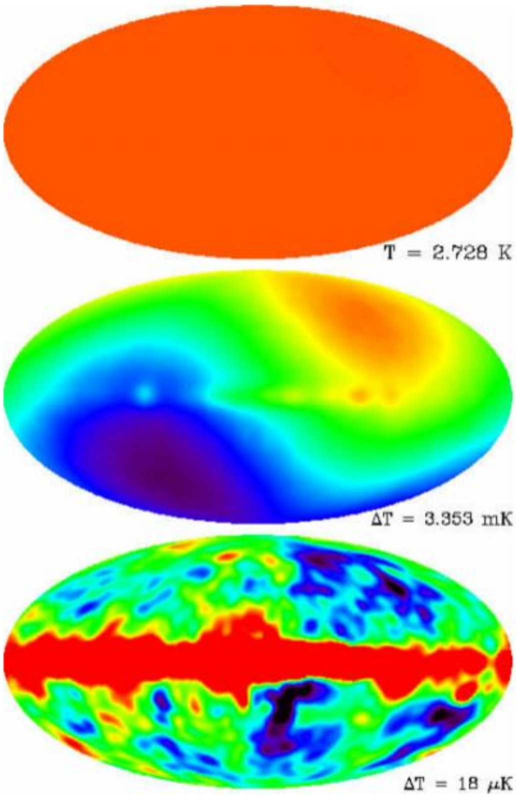


Figure 12.290: **Radiation is isotropic to 10^{-5} in temperature**

This value of approximately $1,000$ allows us, from the Friedmann-Lemaître model we have introduced earlier above, to easily calculate at what time (horizon) of the Universe this decoupling occurred.

If we consider a radiation dominate flat Universe we have proved roughly that the scale factor $R(t)$ was proportional to the $t^{2/3}$:

$$R(t)\propto t^{2/3} \quad (12.5.331)$$

So clearly we have the age ratio of:

$$1100\propto t^{2/3} \quad (12.5.332)$$

That is:

$$t \cong 1,100^{3/2} \cong 36,482 \qquad (12.5.333)$$

So to get the age of the universe at $z = 1100$ we would therefore have to divide the age now, by the factor $36,482$. This gives:

$$\frac{13.700 \pm 0.021 \cdot 10^9 \text{ [years]}}{36,482} \cong 378,241 \text{ [years]} \qquad (12.5.334)$$

Or more generally:

$$\boxed{t_{\text{decoupling}} = H_0^{-1}\left(\frac{T_{\text{dec}}}{T_{\text{CMB}}}\right)^{-3/2}} \qquad (12.5.335)$$

Thus we find a value of approximately $380,000$ years. This is according to what most textbooks gives without proof.

Let us focus now on a special and important needed value for the next sub-section (recombination temperature)!

The energy density associated with the blackbody radiation of temperature T is as we know given by the Stefan-Boltzmann law (see section Thermodynamics page 3706):

$$\rho_{\gamma,0}(T) = \sigma_{\text{S.B.}} T_0^4 = 0.262 \, [\text{MeV} \cdot \text{m}^{-3}] \qquad (12.5.336)$$

and the mean energy per photon is given by $\sim kT$. Therefore, the number density of blackbody photons for $T = 2.7$ [K] is given by:

$$n_\gamma \cong \frac{\frac{4\sigma}{c} T_0^4}{kT} = 370 \, [\text{photons} \cdot \text{cm}^{-3}] \qquad (12.5.337)$$

The number density of baryons can be expressed by ρ_m/m_p, where ρ_m is the mass density of the observed Universe and m_p is the mass of the proton $(1.66 \cdot 10^{-24} \, [\text{g}])$. CMB measurements show that the baryonic mean density is $\rho_m \cong 4.2 \cdot 10^{-31} \, [\text{g} \cdot \text{cm}^{-3}]$ (roughly 5% of the critical density). This leads to the value of $\sim 2 \cdot 10^{-7}$ for the number density of baryons.

Then we have the photon to baryon[65] ratio (that is sometimes defined as the opposite ratio depending on the textbook...):

$$\eta = \frac{\rho_r}{\rho_m} = \frac{\text{photons}}{\text{baryons}} = \frac{n_\gamma}{n_b} \cong \frac{370}{2 \cdot 10^{-7}} \cong 1.85 \cdot 10^9 \qquad (12.5.338)$$

In other words, for each baryon in the Universe there is 10^{10} photons. This estimate is in agreement with the precise value of the photon to baryon ratio $1.64 \cdot 10^9$ derived with the WMAP. Since the photon number and the baryon number seems to be conserved, the photon-baryon ratio stays constant as the Universe expands.

The Planck satellite 95% confidence interval (year 12017 according to holocene calendar) for η is given by:

$$1.61 \cdot 10^9 \leq \eta \leq 1.69 \cdot 10^9 \qquad (12.5.339)$$

---

[65] A baryon is a composite subatomic particle made up of three quarks (a triquark, as distinct from mesons, which are composed of one quark and one antiquark). Baryons and mesons belong to the hadron family of particles, which are the quark-based particles.

Remark

The photon to baryon ratio is estimated and measured. It's partially measured/estimated from the observation of the numbers and average sizes of galaxies and stars, and the density of interstellar matter, and some other baryonic matter. The number density of photons is very well estimated to be about 413 photons per cm^3.

12.5.5.2 Matter/Photon decoupling (recombination temperature)

Now let us see how we can get an approximation of the famous $3,000$ [K] that we see in almost all textbooks for the recombination temperature!

As we have seen above, the early universe may have been hot. So hot maybe that nuclei boiled. The great thermal energy of the universe overwhelmed the confining efforts of the nuclear and electromagnetic forces, and droves of fundamental particles - quarks, gluons, leptons, photons - bounced and jostled in a tightly-coupled plasma. The Universe, however, has a built-in cooling mechanism: expansion. Matter particles dilute and photon wavelengths stretch as the universe expands in a process named "redshift". Soon, protons and neutrons condense out of the plasma as the strong force overcomes the waning thermal wanderings of quarks and gluons. Over the next ten minutes, these nucleons join to form nuclei of deuterium, helium, and other light elements. The protons and nuclei, however, are naked - it's still too hot for electrons to join them in the formation of neutral atoms. The universe will remain in this ionized state for the next few hundred thousand years[66].

The thing about ionized plasma, is that photons have a hard time getting around: there is no shortage of charged particles ready for a collision. The average distance travelled by a photon before collisions is the mean free path (see section Continuum Mechanics page 3925):

$$\gamma_\lambda=(n\sigma)^{-1} \tag{12.5.340}$$

determined by the number density of the charge-carriers, n, and the scattering cross-section, σ. To get a feel for just how waded up everything was maybe in the early Universe, let's find the mean free path of photons when the universe was one second old, at a torrid $T=10^{10}$ [K]. This is the epoch of said of "nucleosynthesis", when protons and neutrons are supposed to fused to form the light nuclei.

---

[66] The text that will follow is based on the work of Brian Powell PhD Physics. Who kindly answered to our questions and authorized us to reproduce his work and improve it. Original text may be found here `https://www.physicsforums.com/insights/poor-mans-cmb-primer-part-1-birth-cosmic-background-radiation/`

$$\begin{aligned}
&(\mathrm{Z}+)\gamma \longrightarrow \mathrm{e}^- + \mathrm{e}^+ \quad \text{(Big Bang nucleosynthesis)}\\
&\gamma+\gamma \longrightarrow \mathrm{W}^+ + \mathrm{W}^- \quad \text{(Big Bang nucleosynthesis)}\\
&\mathrm{u}+\mathrm{u}+\mathrm{d} \longrightarrow \mathrm{p}^+ \quad \text{(Big Bang nucleosynthesis)}\\
&\mathrm{d}+\mathrm{d}+\mathrm{u} \longrightarrow \mathrm{n}^0 \quad \text{(Big Bang nucleosynthesis)}\\
&\mathrm{n}^0 \longrightarrow \mathrm{p}^+ + \mathrm{e}^- + \bar{\nu}_e \quad \text{(Big Bang nucleosynthesis)}\\
&\mathrm{p}^+ + \mathrm{n}^0 \longrightarrow {}_1^2\mathrm{D} + \gamma \quad \text{(Big Bang nucleosynthesis)}\\
&{}_1^2\mathrm{D} + \mathrm{p}^+ \longrightarrow {}_2^3\mathrm{He} + \gamma \quad \text{(Big Bang nucleosynthesis)}\\
&{}_1^2\mathrm{D} + {}_1^2\mathrm{D} \longrightarrow {}_2^3\mathrm{He} + \mathrm{n}^0 \quad \text{(Big Bang nucleosynthesis)}\\
&{}_1^2\mathrm{D} + {}_1^2\mathrm{D} \longrightarrow {}_1^3\,\mathrm{T} + \mathrm{p}^+ \quad \text{(Big Bang nucleosynthesis)}\\
&{}_1^3\mathrm{T} + {}_1^2\mathrm{D} \longrightarrow {}_2^4\mathrm{He} + \mathrm{n}^0 \quad \text{(Big Bang nucleosynthesis)}\\
&{}_1^3\mathrm{T} + {}_2^4\mathrm{He} \longrightarrow {}_3^7\mathrm{Li} + \gamma \quad \text{(Big Bang nucleosynthesis)}\\
&{}_2^3\mathrm{He} + \mathrm{n}^0 \longrightarrow {}_1^3\,\mathrm{T} + \mathrm{p}^+ \quad \text{(Big Bang nucleosynthesis)}\\
&{}_2^3\mathrm{He} + {}_1^2\mathrm{D} \longrightarrow {}_2^4\mathrm{He} + \mathrm{p}^+ \quad \text{(Big Bang nucleosynthesis)}\\
&{}_2^3\mathrm{He} + {}_2^4\mathrm{He} \longrightarrow {}_4^7\mathrm{Be} + \gamma \quad \text{(Stars nucleosynthesis)}\\
&{}_3^7\mathrm{Li} + \mathrm{p}^+ \longrightarrow {}_2^4\mathrm{He} + {}_2^4\mathrm{He} \quad \text{(Stars nucleosynthesis)}\\
&{}_4^7\mathrm{Be} + \mathrm{n}^0 \longrightarrow {}_3^7\mathrm{Li} + \mathrm{p}^+ \quad \text{(Stars nucleosynthesis)}\\
&{}_2^4\mathrm{He} + {}_2^4\mathrm{He} \longrightarrow {}_4^8\mathrm{Be} \quad \text{(Stars nucleosynthesis)}\\
&{}_4^8\mathrm{Be} + {}_2^4\mathrm{He} \longrightarrow {}_6^{12}\mathrm{C} + 2\gamma \quad \text{(Stars nucleosynthesis)}\\
&{}_6^{12}\mathrm{C} + {}_2^4\mathrm{He} \longrightarrow {}_8^{16}\mathrm{O} + \gamma \quad \text{(Stars nucleosynthesis)}\\
&{}_8^{16}\mathrm{O} + {}_2^4\mathrm{He} \longrightarrow {}_{10}^{20}\mathrm{Ne} + \gamma \quad \text{(Stars nucleosynthesis)}\\
&{}_{10}^{20}\mathrm{Ne} + {}_2^4\mathrm{He} \longrightarrow {}_{12}^{24}\mathrm{Mg} + \gamma \quad \text{(Stars nucleosynthesis)}\\
&{}_{12}^{24}\mathrm{Mg} + {}_2^4\mathrm{He} \longrightarrow {}_{14}^{28}\mathrm{Si} + \gamma \quad \text{(Stars nucleosynthesis)}\\
&{}_{14}^{28}\mathrm{Si} + {}_2^4\mathrm{He} \longrightarrow {}_{16}^{32}\mathrm{S} + \gamma \quad \text{(Stars nucleosynthesis)}\\
&{}_{16}^{32}\mathrm{S} + {}_2^4\mathrm{He} \longrightarrow {}_{18}^{36}\mathrm{Ar} + \gamma \quad \text{(Stars nucleosynthesis)}\\
&{}_{18}^{36}\mathrm{Ar} + {}_2^4\mathrm{He} \longrightarrow {}_{20}^{40}\mathrm{Ca} + \gamma \quad \text{(Stars nucleosynthesis)}\\
&{}_{20}^{40}\mathrm{Ca} + {}_2^4\mathrm{He} \longrightarrow {}_{22}^{44}\mathrm{Ti} + \gamma \quad \text{(Stars nucleosynthesis)}\\
&{}_{22}^{44}\mathrm{Ti} + {}_2^4\mathrm{He} \longrightarrow {}_{24}^{48}\mathrm{Cr} + \gamma \quad \text{(Stars nucleosynthesis)}\\
&{}_{24}^{48}\mathrm{Cr} + {}_2^4\mathrm{He} \longrightarrow {}_{26}^{52}\mathrm{Fe} + \gamma \quad \text{(Stars nucleosynthesis)}\\
&{}_{26}^{52}\mathrm{Fe} + {}_2^4\mathrm{He} \longrightarrow {}_{28}^{56}\mathrm{Ni} + \gamma \quad \text{(Stars nucleosynthesis)}\\
&{}_{28}^{56}\mathrm{Ni} + {}^4\mathrm{He} \longrightarrow {}^{60}\mathrm{Zn} + \gamma \quad \text{(Supernova nucleosynthesis)}\\
&{}_{26}^{56}\mathrm{Fe} + 171\mathrm{n}^0 \xrightarrow{r\text{-process}} {}_{79}^{197}\mathrm{Au} + \bar{\nu} + \gamma \quad \text{(Supernova nucleosynthesis/Neutron star collisions)}\\
&\ldots
\end{aligned} \tag{12.5.341}$$

Though we are technically in the realm of Compton scattering (see section Atomistic page 5096), $T > m_e$, the Thomson scattering cross-section (see section Atomistic page 5100) will suffice for an order-of-magnitude estimate: $\sigma_T \cong 6.652 \cdot 10^{-29}\,[\mathrm{m}^2]$, where σ_T for reminder is proportional to e^4/m_e^4. As for the charge-carrier density, we can safely ignore everything except the electrons: though the supposed charge neutrality of the universe requires that the densities of free electrons and protons balance, $n_e = n_p$, scattering off the much heavier protons is suppressed by a factor of a million relative to electrons (owing to the factor $1/m_e^4$ in σ_T)[67]

It's safe to treat the electrons as radiation (they are still quite hot, about the temperature of a supernova). To get the number density, we need to integrate the Fermi-Dirac distribution function (see section Statistical Mechanics page 3571):

$$f_{\text{FD}}(E,T,\mu)=\frac{1}{e^{\frac{E-\mu}{kT}}+1} \tag{12.5.342}$$

over the full phase space in the relativistic limit, $(T \gg \mu, m)$. For this we will use a similar approach to that we used during our study of semi-conductors that had lead us to (see page 4291):

$$\rho(E)\cong\frac{1}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}\int\limits_{E_0}^{+\infty}\frac{(E-E_0)^{1/2}}{\exp\left(\frac{E-\mu}{kT}\right)+1}\mathrm{d}E \tag{12.5.343}$$

But however with an important difference. Instead of taking:

$$k\cong\sqrt{(E-E_0)\frac{2m_e}{\hbar^2}} \tag{12.5.344}$$

as we did for the semi-conductor. We take now:

$$k=\frac{p^2}{\hbar^2}=\frac{E^2-m_e^2c^4}{c^2\hbar^2} \tag{12.5.345}$$

What lead us to (still see page 4291):

$$\frac{N(k)}{V}=\frac{k^3}{3\pi^2}=\frac{1}{3\pi^2}\frac{(E^2-m_e^2c^4)^{3/2}}{c^3\hbar^3}=\frac{N(E)}{V} \tag{12.5.346}$$

And therefore:

$$D(E)=\frac{1}{V}\frac{\mathrm{d}N(E)}{\mathrm{d}E}=\frac{\mathrm{d}n(E)}{\mathrm{d}E}=\frac{(E^2-m_e^2c^4)^{1/2}E}{c^3\hbar^3} \tag{12.5.347}$$

And if $E \gg m_e$ then:

$$D(E)\cong\frac{E^2}{c^3\hbar^3} \tag{12.5.348}$$

Hence the above integral becomes (changing to lower boundary to the case that concerns us here):

$$\rho(E)\cong\frac{1}{c^3\hbar^3}\int\limits_{0}^{+\infty}\frac{E^2}{\exp\left(\frac{E-\mu}{kT}\right)+1}\mathrm{d}E \tag{12.5.349}$$

and in the approximation $T \gg \mu, m$:

$$\rho(E)\cong\frac{1}{c^3\hbar^3}\int\limits_{0}^{+\infty}\frac{E^2}{\exp\left(\frac{E}{kT}\right)+1}\mathrm{d}E \tag{12.5.350}$$

---

[67] And the even heavier and more dilute deuterium, helium, and lithium nuclei can be likewise ignored.

Integral that can be found in some textbooks in natural units. Therefore:

$$\rho(E) \cong \int\limits_{0}^{+\infty} \frac{E^2}{\exp\left(\frac{E}{T}\right)+1}\mathrm{d}E \tag{12.5.351}$$

Back to S.I. units that lead us finally to:

$$n_e \cong g_e \frac{1}{c^3\hbar^3} \int\limits_{0}^{+\infty} \frac{E^2}{\exp\left(\frac{E}{kT}\right)+1}\mathrm{d}E \tag{12.5.352}$$

where $g_e = 2$ is the fermion degeneracy.

The integral:

$$\int\limits_{0}^{+\infty} \frac{E^2}{\exp\left(\frac{E}{kT}\right)+1}\mathrm{d}E \tag{12.5.353}$$

can be solved via geometric series expansion and some low-level tricks. First, substitute $x = E/kT$ giving:

$$I = (kT)^3 \int\limits_{0}^{+\infty} \frac{x^2}{e^x+1}\mathrm{d}x \tag{12.5.354}$$

The idea is to use the binomial theorem for negative integer exponents (see section Calculus page 1707) involving the exponential:

$$I = (kT)^3 \int\limits_{0}^{+\infty} e^{-x}(1+e^{-x})^{-1}x^2\mathrm{d}x = (kT)^3 \int\limits_{0}^{+\infty} \left[\sum_{k=0}^{+\infty}(-1)^k e^{-(k+1)x}\right] x^2\mathrm{d}x \tag{12.5.355}$$

After another change of variables, $y = (k+1)x$:

$$I = (kT)^3 \sum_{k=0}^{+\infty}(-1)^k \frac{1}{(k+1)^3} \int\limits_{0}^{+\infty} y^2 e^{-y}\mathrm{d}y \tag{12.5.356}$$

The keen eye might recognize $\int\limits_{0}^{+\infty} y^2 e^{-y}\mathrm{d}y$ as the gamma function (see section Differential and Integral Calculus page 1574), $\Gamma(3) = (3-1)! = 2$. This, together with a slight nudge to the bottom limit of the summation we can rewrite things as:

$$I = -(kT)^3\Gamma(3) \sum_{k=1}^{+\infty} \frac{(-1)^k}{k^3} \tag{12.5.357}$$

To get rid of the annoying $(-1)^k$ term, we break the summation up into an even and an odd part, and then rework things to get two full summations back:

$$I = -(kT)^3\Gamma(3)\left(\sum_{k}^{\text{even}}\frac{1}{k^3} - \sum_{k}^{\text{odd}}\frac{1}{k^3}\right) = -(kT)^3\Gamma(3)\left(\frac{2}{2^3}\sum_{l=1}^{+\infty}\frac{1}{l^3} - \sum_{k=1}^{+\infty}\frac{1}{k^3}\right) \tag{12.5.358}$$

After the dust settles, we see that two copies of the Riemann zeta function have emerged, $\zeta(3) = \sum_k 1/k^3$ (see section Sequences and Series page 1682), allowing us to write:

$$I = (kT)^3\zeta(3)\Gamma(3)\left(1-\frac{1}{2^2}\right) = \frac{3}{2}\zeta(3)(kT)^3 \qquad (12.5.359)$$

Therefore:

$$n_e = 2\frac{1}{c^3\hbar^3}\frac{3}{2}\zeta(3)(kT)^3 = \frac{3}{c^3\hbar^3}\zeta(3)(kT)^3 \cong 3\cdot 10^8 T^3 \qquad (12.5.360)$$

where $\zeta(3) \cong 1.2$ is the Riemann zeta function of 3. The mean free path is found after plug-in all numerical values to (remember that we take $T = 10^{10}$ [K]):

$$\lambda_\gamma = (n_e\sigma_T)^{-1} \cong (3\cdot 10^8 T^3\sigma_T)^{-1} \cong \left((3\cdot 10^{38})\cdot(6.652\cdot 10^{-29})\right)^{-1} \cong 5\cdot 10^{-11}\,[\text{m}] \qquad (12.5.361)$$

or about the size of a typical virus. So, yeah: tiny. The Universe is effectively opaque!

As the Universe continues to cool we eventually expect neutral hydrogen to form and the free charge-carrier density to decline rapidly in a process confusingly named "recombination" (there's nothing "re-" about it).Naively, we might expect this transition to occur as the universe cools through $T = 10^5$ [K], which corresponds dimensionally to the ionization energy of hydrogen, $E = 13.6$ [eV].

Indeed, if we consider that the first atoms absorbing photons and avoiding the transparency of the Universe were Hydrogen atoms. As electrons "orbiting" the hydrogen proton never had by assumption the time to "relax" (reached their fundamental state) because of the density of the early Universe we can assume that all absorbed photons were for electrons between the first and second main quantum number. But as we have proved in the section of Corpuscular Quantum Physics:

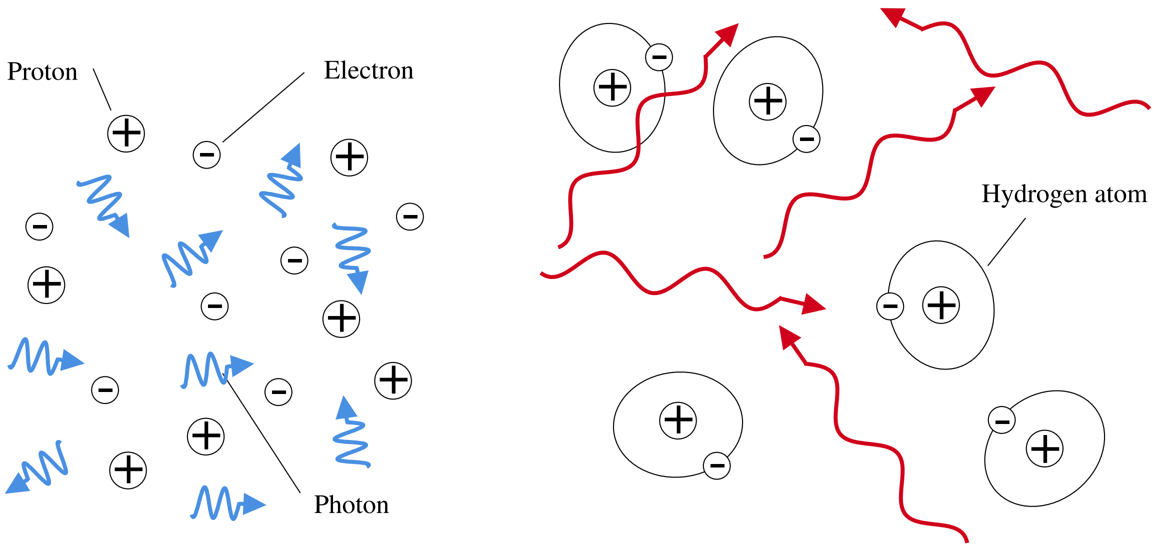
$$E = h\nu_{n_2,n_1} = Z^2\frac{me^4}{8\varepsilon_0 h^3}\left(\frac{1}{n_1^2}-\frac{1}{n_2^2}\right) \qquad (12.5.362)$$

and for Hydrogen:

$$E_{n_2,n_1} = h\nu_{n_2,n_1} = 13.605\left(\frac{1}{n_1^2}-\frac{1}{n_2^2}\right)\,[\text{eV}] \qquad (12.5.363)$$

So that's where the $13.605\,[\text{eV}]$ comes from!

At this energy the reaction $p + e^- \longleftrightarrow \text{H} + \gamma$ should begin to fall out of chemical equilibrium because background photons won't be able to re-ionize the newly formed hydrogen.



A. Before (re)combination B. After (re)combination

Figure 12.291: **Big Bang recombination**

With enough electrons locked away in hydrogen atoms, photons will propagate relatively unimpeded across the cosmos. This event, the corollary of recombination named "decoupling", is important because these are the very photons that make up today's cosmic microwave background. Let us see if this is indeed what happens by finding the photon mean free path when the Universe has cooled down below 13.6 [eV] to a tepid 1 [eV], corresponding to a temperature of roughly $T = 10^4$ [K] (around the temperature of a White Dwarf, or a really hot habanero). At these temperatures the electrons are no longer relativistic (since $T \ll m_e$) and so the number density assumes the classical Boltzmann (see section Electrokinetics page 4294) form:

$$\rho_C(E) = 2\left(\frac{m_e kT}{2\pi\hbar^2}\right)^{3/2} \exp\left(\frac{\mu - E_C}{kT}\right) \tag{12.5.364}$$

That we will rewrite here:

$$n_e = 2\left(\frac{m_e kT}{2\pi\hbar^2}\right)^{3/2} \exp\left(\frac{\mu_e - m_e c^2}{kT}\right) \tag{12.5.365}$$

Indeed, as photons do not interact with each other, hence they can not achieve a thermodynamic equilibrium by themselves, therefore their chemical potential is zero.

Again... the reader may often found this relation in natural units in some textbooks leading to:

$$n_e = 2\left(\frac{m_e T}{2\pi}\right)^{3/2} \exp\left(\frac{\mu_e - m_e}{T}\right) \tag{12.5.366}$$

However we can't solve n_e directly because we don't know $\mu_e(T)$ in general. So the idea is to use the Saha ionization equation (see section Continuum Mechanics page 3954):

$$\frac{x_e^2}{1 - x_e} \cong \frac{\eta}{n_\gamma} e^{-B/kT} \left(\frac{m_e m_p}{m_{\text{H}}}\right)^{3/2} \left(\frac{kT}{2\pi\hbar^2}\right)^{3/2} \tag{12.5.367}$$

where for recall $x_e = n_e/n_b$ is the ionization fraction, B the ionization of the first level of Hydrogen atom $B = 13.6$ [eV] and where we know (assuming equal charge density) that the photon to baryon ratio is given by $\eta = n_\gamma / n_b \cong 1.85 \cdot 10^9$ with $n_\gamma \cong 370 \cdot 10^6$ photons by cubic meter and $n_b \cong 0.2$ baryons by cubic meter.

So we found at $T = 10^4$ [K] after putting all numerical values in the previous relation :

$$\begin{aligned}\frac{(n_e/n_b)^2}{1-(n_e/n_b)} = \frac{(n_e/370\cdot 10^6)^2}{1-(n_e/370\cdot 10^6)} &\cong \frac{\eta}{n_\gamma} e^{-B/kT}\left(\frac{m_e m_p}{m_\mathrm{H}}\right)^{3/2}\left(\frac{kT}{2\pi\hbar^2}\right)^{3/2}\\ &= \frac{1.85\cdot 10^9}{2\cdot 10^{-7}} e^{-\frac{2.17896\cdot 10^{-18}}{k\cdot 10^4}}\left(\frac{m_e m_p}{m_\mathrm{H}}\right)^{3/2}\left(\frac{k\cdot 10^4}{2\pi\hbar^2}\right)^{3/2}\end{aligned} \tag{12.5.368}$$

and solving for n_e (keeping only the positive solution):

$$n_e \cong 370\,[\text{electrons}\cdot\text{cm}^{-3}] = 370\cdot 10^6\,[\text{electrons}\cdot\text{m}^{-3}] \tag{12.5.369}$$

The mean free path is found after plug-in all numerical values to $\lambda_\gamma = (n_e \sigma_T)^{-1} \cong 4.07 \cdot 10^{19}$ [m] or around $1,300$ parsecs. That's quite a distance - significantly longer than the mean free path during nucleosynthesis - but still only a sliver of the observable universe at that time. Hardly across all the cosmos. Apparently the universe is still quite ionized even at $E = 1$ [eV]. So what's going on?

For one, there are hugely many more photons than charge-carriers per unit volume. The hugeness of this disparity is the result of the tiny matter-antimatter asymmetry (CP-violation as seen page **??**) in the early universe: most of the matter and all of the antimatter annihilated to generate a glut of photons, leaving only a tiny residual matter density. All told, there are roughly 10 billion photons for every electron in the universe. With so many more photons than electrons, even at $T = 10^4$ [K] there were enough photons all they way out in the ultraviolet tail of the blackbody spectrum with energies greater than 13.6 [eV] to keep hydrogen relatively well-ionized. To find out just how many let us recall that he have seen during our derivation of the Planck's law (see section Thermodynamics page 3719), that:

$$\mathrm{d}n = \frac{\frac{V 8\pi\nu^2}{c^3}\mathrm{d}\nu}{e^{\frac{h\nu}{kT}}-1} \tag{12.5.370}$$

And therefore the photons cubic density in our case is given by:

$$n_\gamma := \int \frac{\mathrm{d}n}{V} = \int\limits_{13.6\,[\text{eV}]}^{+\infty} \frac{8\pi}{c^3}\frac{\nu^2\mathrm{d}\nu}{e^{\frac{h\nu}{kT}}-1} \tag{12.5.371}$$

By multiplying by h^3/h^3, we get:

$$n_\gamma = \int\limits_{13.6\,[\text{eV}]}^{+\infty} \frac{8\pi}{h^3c^3}\frac{(h^2\nu^2)\mathrm{d}(h\nu)}{e^{\frac{h\nu}{kT}}-1} = \int\limits_{13.6\,[\text{eV}]}^{+\infty} \frac{8\pi}{h^3c^3}\frac{E^2\mathrm{d}E}{e^{\frac{E}{kT}}-1} = \frac{1}{\hbar^3\pi^2c^3}\int\limits_{13.6\,[\text{eV}]}^{+\infty} \frac{E^2\mathrm{d}E}{e^{\frac{E}{kT}}-1} \tag{12.5.372}$$

Remark

In natural units the latter relation can be found in the following form in some textbooks:

$$n_\gamma=\frac{1}{\pi^2}\int\limits_{13.6\,[\text{eV}]}^{+\infty}\frac{E^2\mathrm{d}E}{\exp(E/T)-1}\tag{12.5.373}$$

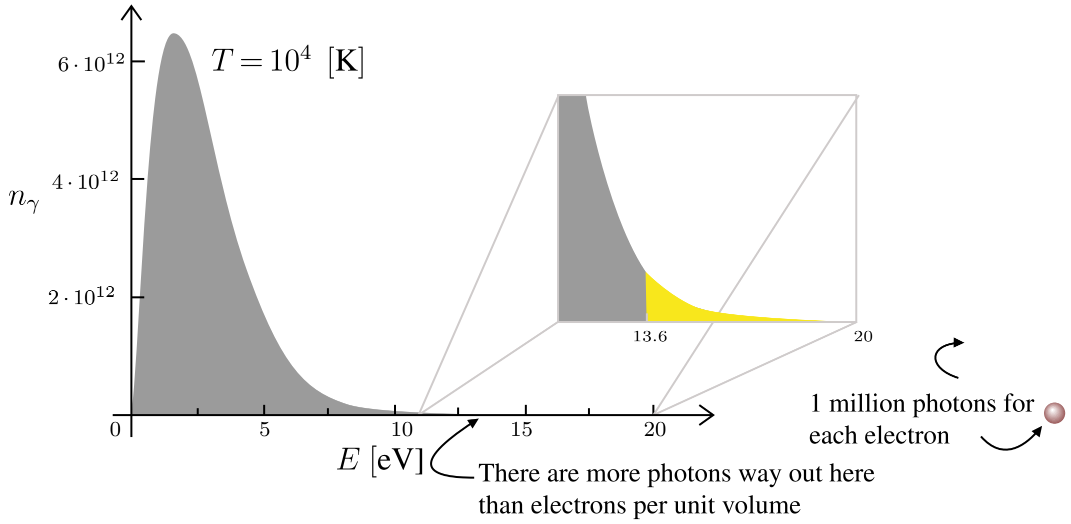
Sadly it seems that this definite integral has no analytical expression. So we will have to compute it numerically. As actual softwares can't deal with such small values (because 13.6 [eV] is equal to $2.17896\cdot 10^{-18}$ [J]) we will need to do a change of variable. We put $u=E/(kT)$ that leads us to:

$$n_\gamma=\frac{1}{\hbar^3\pi^2c^3}\int\limits_{13.6\,[\text{eV}]}^{+\infty}\frac{E^2\mathrm{d}E}{e^{\frac{E}{kT}}-1}=\frac{1}{\hbar^3\pi^2c^3}\int\limits_{15.78215}^{+\infty}\frac{u^2k^2T^2(kT\mathrm{d}u)}{e^u-1}=\frac{k^3T^3}{\hbar^3\pi^2c^3}\int\limits_{15.78215}^{+\infty}\frac{u^2\mathrm{d}u}{e^u-1}\tag{12.5.374}$$

In a mathematical computing software like Maple™, this gives us:

$$n_\gamma\cong\frac{k^3T^3}{\hbar^3\pi^2c^3}\cdot 0.3954884946\cdot 10^{-4}\cong 3.3\cdot 10^{14}\,[\text{photons}\cdot\text{m}^{-3}]\tag{12.5.375}$$

Result to compare with the $n_e\cong 370\cdot 10^6$ [electrons $\cdot\,\text{m}^{-3}$] (at the same temperature of $T=10^4$ [K]) determined above. Then there are around one millions photons with an energy greater than $E=13.6$ [eV] for every electron! This means that for every neutral hydrogen atom that forms, there are schools of high-energy photons available to re-ionize it.



1 million photons for each electron

There are more photons way out here than electrons per unit volume

So then when does recombination happen? To find out, we'll need to do a more careful accounting of the electron and baryon number densities. In particular, we're going to track the ionization fraction, $x_e=n_e/n_b$, where the baryon density includes the protons and neutral hydrogen, $n_b=n_p+n_\text{H}$. We assume that the electrons, baryons, and photons are all in thermal equilibrium, which is true so long as the reaction $p+e\rightleftharpoons \text{H}+\gamma$ proceeds rapidly relative to the expansion rate of the universe. In equilibrium, the chemical potentials balance $\mu_e+\mu_p=\mu_\text{H}$: this condition is our starting point. We know from our earlier results that we are well inside the non-relativistic regime when recombination finally does happen, so all equilibrium abundances have the Boltzmann form. Writing the condition $\mu_e+\mu_p=\mu_{\text{H}}$ in terms of n_e, n_p, and n_{H} gives as proved during our study of the Saha Equation (see section Continuum Mechanics page 3954):

$$\frac{x_e^2}{1-x_e}\cong\frac{\eta}{n_\gamma}e^{-B/kT}\left(\frac{m_e kT}{2\pi\hbar^2}\right)^{3/2} \tag{12.5.376}$$

with:

$$n_\gamma=\frac{1}{\hbar^3\pi^2c^3}\int\limits_0^{+\infty}\frac{E^2\mathrm{d}E}{e^{\frac{E}{kT}}-1}=\frac{k^3T^3}{\hbar^3\pi^2c^3}\int\limits_0^{+\infty}\frac{x^2\mathrm{d}x}{e^x-1} \tag{12.5.377}$$

The last integral can also be solved via geometric series expansion and some low-level tricks. The idea is again to use the binomial theorem for negative integer exponents (see section Calculus page 1707) involving the exponential:

$$I=\int\limits_0^{+\infty}e^{-x}(1-e^{-x})^{-1}x^2\mathrm{d}x=\int\limits_0^{+\infty}\left[\sum_{k=0}^{+\infty}(-1)^k(-1)^k e^{-(k+1)x}\right]x^2\mathrm{d}x=\int\limits_0^{+\infty}\left[\sum_{k=0}^{+\infty}(-1)^{2k}e^{-(k+1)x}\right]x^2\mathrm{d}x \tag{12.5.378}$$

After another change of variables, $y=(k+1)x$:

$$I=\sum_{k=0}^{+\infty}(-1)^{2k}\frac{1}{(k+1)^3}\int\limits_0^{+\infty}y^2e^{-y}\mathrm{d}y \tag{12.5.379}$$

The keen eye might recognize $\int\limits_0^{+\infty}y^2e^{-y}\mathrm{d}y$ as the gamma function (see section Differential and Integral Calculus page 1574), $\Gamma(3)=(3-1)!=2$. This, together with a slight nudge to the bottom limit of the summation we can rewrite things as:

$$I=\Gamma(3)\sum_{k=1}^{+\infty}\frac{(-1)^{2k}}{k^3} \tag{12.5.380}$$

Now notice that:

$$(-1)^{2k}=\left((-1)^2\right)^k=1 \tag{12.5.381}$$

Thus:

$$\sum_{k=1}^{+\infty}\frac{(-1)^{2k}}{k^3}=\sum_{k=1}^{+\infty}\frac{1}{k^3}=\zeta(3) \tag{12.5.382}$$

Therefore:

$$I=\Gamma(3)\sum_{k=1}^{+\infty}\frac{(-1)^{2k}}{k^3}=\Gamma(3)\zeta(3)=2\zeta(3)\cong 2.404 \tag{12.5.383}$$

So finally:

$$n_\gamma = \frac{k^3T^3}{\hbar^3\pi^2c^3}\int\limits_0^{+\infty}\frac{x^2\mathrm{d}x}{e^x-1} = \frac{k^3T^3}{\hbar^3\pi^2c^3}2\zeta(3) \cong 5.37\cdot 10^6 T^3 \tag{12.5.384}$$

Hence:

$$\begin{aligned}\frac{x_e^2}{1-x_e} &\cong \frac{\eta}{n_\gamma}e^{-B/kT}\left(\frac{m_em_p}{m_\mathrm{H}}\right)^{3/2}\left(\frac{kT}{2\pi\hbar^2}\right)^{3/2} = \frac{\eta\hbar^3\pi^2c^3}{2k^3T^3\zeta(3)}e^{-B/kT}\left(\frac{m_em_p}{m_\mathrm{H}}\right)^{3/2}\left(\frac{kT}{2\pi\hbar^2}\right)^{3/2}\\ &= \frac{\eta\pi^{1/2}c^3}{2^{5/2}k^{3/2}\zeta(3)}\left(\frac{m_em_p}{m_\mathrm{H}}\right)^{3/2}\frac{e^{-B/kT}}{T^{3/2}}\end{aligned} \tag{12.5.385}$$

Remark

The above relation can be found in natural units in some textbooks as following:

$$\frac{x_e^2}{1-x_e} \cong \frac{\eta\pi^{1/2}}{2^{5/2}\zeta(3)}\left(\frac{m_em_p}{m_\mathrm{H}}\right)^{3/2}\frac{e^{-B/T}}{T^{3/2}} \cong \frac{\eta}{4\zeta(3)}\sqrt{\frac{\pi}{2}}\left(\frac{m_e}{T}\right)^{3/2}e^{-B/T} \tag{12.5.386}$$

or inverted:

$$\frac{1-x_e}{x_e^2} \cong \frac{4\zeta(3)}{\eta}\sqrt{\frac{2}{\pi}}\left(\frac{m_e}{T}\right)^{-3/2}e^{B/T} \tag{12.5.387}$$

Knowing that the actual measured baryon to photon ratio is equal to $\eta \cong 1.85\cdot 10^9$, then for example at $T = 4,000$ [K] (≈ 0.33 [eV] a full two orders of magnitude below the hydrogen binding energy), we get:

$$\frac{x_e^2}{1-x_e} \cong \frac{\eta\pi^{1/2}c^3}{2^{5/2}k^{3/2}\zeta(3)}\left(\frac{m_em_p}{m_\mathrm{H}}\right)^{3/2}\frac{e^{-B/kT}}{T^{3/2}} \cong 6.37 \tag{12.5.388}$$

Therefore solving for x_e, this gives us: $x_e \cong 87.88\%$.

It may strike the reader as curious that recombination happens at all. Think about it: even without hordes of hot, ionizing photons flying around, a single recombination directly to the ground state releases a photon with an energy $E > 13.6$ [eV]. Unless the plasma density is sufficiently low (it isn't), these photons will readily re-ionize any neutral hydrogen atoms they happen upon. In fact, any recombination - even those to excited states of hydrogen - release photons capable of re-ionizing other neutral hydrogen atoms in the same excited state. Evidently, recombination must proceed via very particular transitions through the excited states of hydrogen; for example, the decay from $2s$ to $1s$ proceeds via the emission of two photons, neither of which has sufficient energy to excite hydrogen out of the ground state. The other pathway is through Lyman α decay (a 10.2 [eV] photon from $2p$ to $1s$): while absorption of a Lyman α photon will excite neutral hydrogen to a readily-ionizable state (a mere 3.5 [eV] away), as the hydrogen density decreases with the expansion the Lyman α photon redshifts just enough to render it harmless. These transitions work; the problem is that they are slow, and they get slower as the ionization fraction falls. Recall that thermal equilibrium is maintained when reactions proceed at a faster rate than the expansion. As recombination proceeds and the ionization fraction falls, these key transitions to the hydrogen ground state are too slow to maintain equilibrium and are therefore no longer adequately described by the Saha Equation. As a result, the Saha Equation predicts a more rapid recombination process than actually occurs through these slow, out-of-equilibrium decays. It is possible to do a proper job of this, taking into account out-of-equilibrium interactions involving the excited states of hydrogen. The Saha Equation is sufficient for our purposes! It illustrates the physics and gives an order of magnitude estimate of the temperature of recombination.

Putting this all together, we revisit the photon mean free path. With $n_e = x_e n_b = x_e n_\gamma / \eta$ giving:

$$\lambda_\gamma = \frac{\eta}{x_e n_\gamma \sigma_T} \tag{12.5.389}$$

We can observe the mean free path of the photon grow as the formation of neutral hydrogen really gets going in the below figure where λ_γ is the mean free path (solid line), x_e ionization fraction (dashed line), vs temperature:

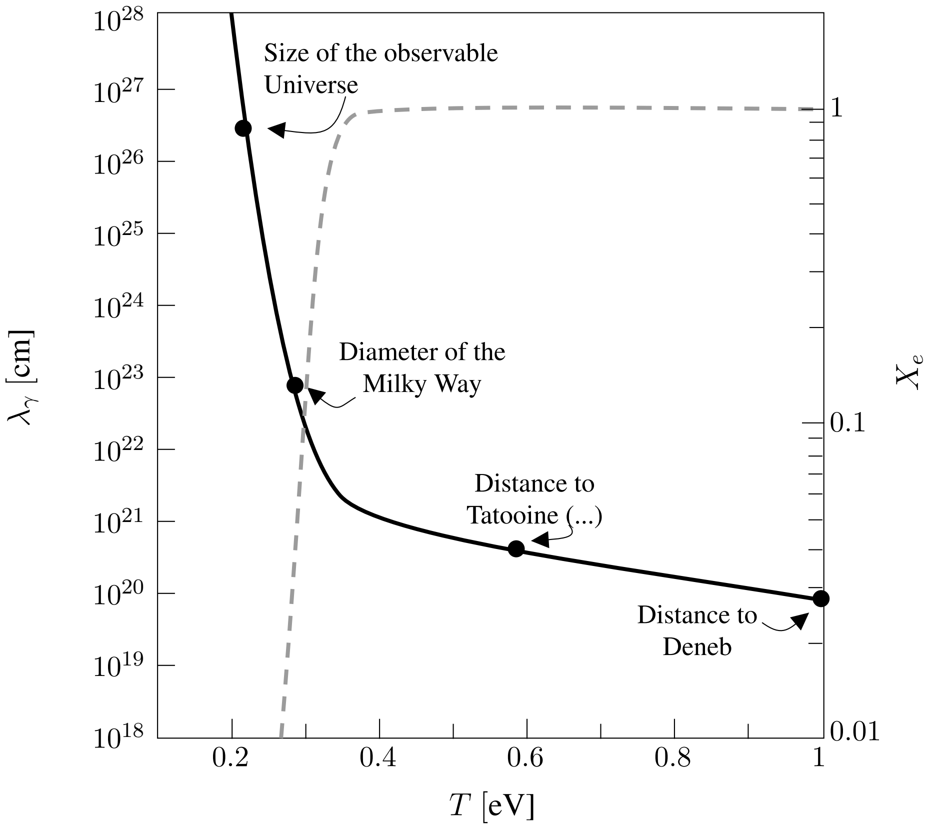


Figure 12.292: Ionization factor and mean free path in function of temperature

Photons traverse the whole of the visible universe unimpeded only when the ionization fraction has dropped below $x_e < 0.01$. These are the CMB photons. Only when the universe is mostly hydrogen (99 hydrogen atoms for every electron) does it become effectively transparent to photons. These photons, ladies and gentleman, are the cosmic background radiation: the CMB is born!

Obviously this result (as most results in this textbook) is just an approximation, but not a bad one. It's an approximation because it is a simplification of what is actually going on. In 11968 (holocene calendar), Jim Peebles (who had done the work in the 11965 Dicke etal. paper) and, independently, Yakov Zel'dovich (of the Sunyaev-Zel'dovich effect) in the USSR, worked out a more complete theory where the hydrogen has three energy levels, rather than what we have done here where we assume that the free electrons go straight into the ground-state (i.e. only two energy levels). Their third level was the $n=2$ state, the energy level just above the ground state. This is an important energy level for hydrogen, as it is transitions down to the $n=2$ level which give rise to visible-light photons. Using this more complicated 3-level model gives a Universe which is 90% neutral at a temperature of $T=3000$ [K], which is why this temperature is most often quoted.

12.5.5.3 Background radiation temperature

We have determined all lot of relations so far. Let us take now back the first Friedmann equation:

$$\left(\frac{\dot{a}}{a}\right)^2=\frac{8\pi G\rho}{3c^2}-\frac{k}{a^2} \tag{12.5.390}$$

and the Universe fluid equation:

$$\dot{\rho}=-3\left(\rho+\frac{P}{c^2}\right)\frac{\dot{R}}{R} \tag{12.5.391}$$

Now for flat space we have $k=0$ so the first Friedmann equation becomes:

$$\left(\frac{\dot{a}}{a}\right)^2=\frac{8\pi G\rho}{3c^2} \tag{12.5.392}$$

and as now we are interested by the radiation, we will not focus on mass density but rather on radiation (energy) density. Then that latter will be written:

$$\left(\frac{\dot{a}}{a}\right)^2=\frac{8\pi G\rho_{\text{E}}}{3} \tag{12.5.393}$$

But we have also proved during our study of the flat Universe dominated by radiation that:

$$\dot{\rho}_E+4\rho_E\frac{\dot{R}}{R}=0 \tag{12.5.394}$$

Then changing the notation and rearranging, we get:

$$\frac{\dot{\rho}_E}{\rho_E}=-4\frac{\dot{a}}{a} \tag{12.5.395}$$

Filling that latter relation inside:

$$\left(\frac{\dot{a}}{a}\right)^2=\frac{8\pi G\rho_{\text{E}}}{3} \tag{12.5.396}$$

gives immediately:

$$\frac{\dot{\rho}_E}{\rho_E}=-4\sqrt{\frac{8\pi G\rho_{\text{E}}}{3}} \tag{12.5.397}$$

We write $\dot{\rho}_E = \mathrm{d}\rho_E/\mathrm{d}t$ and integrate the previous relation (obviously such an integration is a non-sense but the idea is to have a first approach!):

$$\int \frac{\mathrm{d}\rho_E}{\rho_E^{3/2}} = -4\sqrt{\frac{8\pi G}{3}}\int \mathrm{d}t \tag{12.5.398}$$

This gives:

$$\frac{-2}{\sqrt{\rho_E}} = -4\sqrt{\frac{4\pi G}{3}}t \tag{12.5.399}$$

Now we square, and after rearranging we get:

$$\rho_E = \frac{3}{16\pi G t^2} \tag{12.5.400}$$

We also know the Stefan-Boltzmann law (see section Thermodynamics page 3706):

$$\rho(T) = \frac{4}{c}M(T) = \frac{8\pi^5 k_B^4}{15c^3h^3}T^4 \tag{12.5.401}$$

Therefore equating both relations:

$$\frac{3}{16\pi G t^2} = \frac{8\pi^5}{15c^3h^3}(k_B T)^4 \tag{12.5.402}$$

After rearranging we then have for a flat Universe dominated by radiation that the temperature is given by:

$$\boxed{T = \frac{1}{k_B}\sqrt[4]{\frac{45c^3h^3}{128\pi^6 G t^2}} \propto \frac{1}{\sqrt{z}}} \tag{12.5.403}$$

So when it comes to the observable age of the Universe $t_H \cong 13 \cdot 10^9$ [years], this will give:

$$T \cong \frac{1}{1.38065 \cdot 10^{-23}}\sqrt[4]{\frac{45 \cdot (299792458)^3\,(6.2607 \cdot 10^{-34})^3}{128 \cdot (3.1414926)^6 \cdot 6.67408 \cdot 10^{-11} \cdot (6.8328 \cdot 10^{15})^2}} \cong 0.01261\,[\mathrm{K}] \tag{12.5.404}$$

To compare with the 2.728 [K] observed today there is a huge difference. But whatever this gives us the information that the Universe should have a quite cold background radiation and experimentation confirms this!

Now according to Wien's second law we have (see section Thermodynamics page 3727):

$$\nu_{\max} = 2.82144\frac{kT}{h} \cong 0.7\,[\mathrm{GHz}] \tag{12.5.405}$$

To compare with the 282 [GHz] measured actually... So we are indeed in the range of the microwave. Hence the word "microwave" in the CMB...

The reader now has to conceptualize that the CMB patterns change over time! Although statistically they remain the same, and although it will not be noticeable on human timescales!

The CMB we observe now comes from a thin shell with us in the center, and with a radius equal to the distance that the light has travelled from the Universe was 379,000 years old and until now. As time passes, we will receive CMB from a shell with an increasingly larger radius. As that light has travelled farther through space, it will, as you say, be more red-shifted, or "cooler". But it will also have been emitted from more distant regions in the early Universe that, although statistically equivalent, simply will be other regions and hence look different.

The patterns that change the fastest are the smallest patterns we can observe. The angular resolution of the Planck satellite is 5 to 10 arcmin. Since the CMB comes from a redshift of ~ 1100, the angular diameter distance defining the physical distance spanned by a given angle - is ~ 13 [Mpc], so 5 arcmin corresponds to a physical scale of roughly 19 [kpc] in physical coordinates, or 21 [Mpc] in comoving coordinates (that is, a structure spanning 5 arcmin today were ~ 19 [kpc] across at the time of emission, but have now expanded to a size of ~ 21 [Mpc], with $1\,[\text{Mpc}] = 1000\,[\text{kpc}] = 3,261 \cdot 10^3$ lightyears).

Assuming an isotropic Universe, if the smallest observable parcels of gas were 19 [kpc] across perpendicular to our line of sight, they are also on average 19 [kpc] across along our line of sight.

So the question of how fast the CMB changes comes down to how much time did it take light to travel 19 [kpc] when the Universe was $379,000$ years old. This is not simply 19 [kpc] divided by the speed of light, since the Universe expands as the light travels, but it's pretty close. Hence, it took light roughly $62,000$ years to traverse such a patch.

Since we see events at redshift z time dilated by a factor $1+z$, we will have to wait $62\,[\text{kyr}] \cdot 1,100$, or roughly 70 million years to see a changes in the pattern of the CMB.

12.5.6 Friedmann-Lemaître-Robertson-Walker Cosmological Models

Ok... We have seen a lot of stuff and all this without involving General Relativity! Let us see now a more robust way to derivate a more complete version of Friedmann equations using a special metric AND General Relativity. The results will be what we name the "standard cosmological model" or "ΛCDM" (Lambda Cold Dark Matter also abbreviated LCDM) or "Lambda-CDM model" that is a parametrization of the Big Bang cosmological model in which the Universe contains the cosmological constant Λ, associated with dark energy, and cold dark matter. The field is often named "relativistic cosmology" or "concordance cosmology".

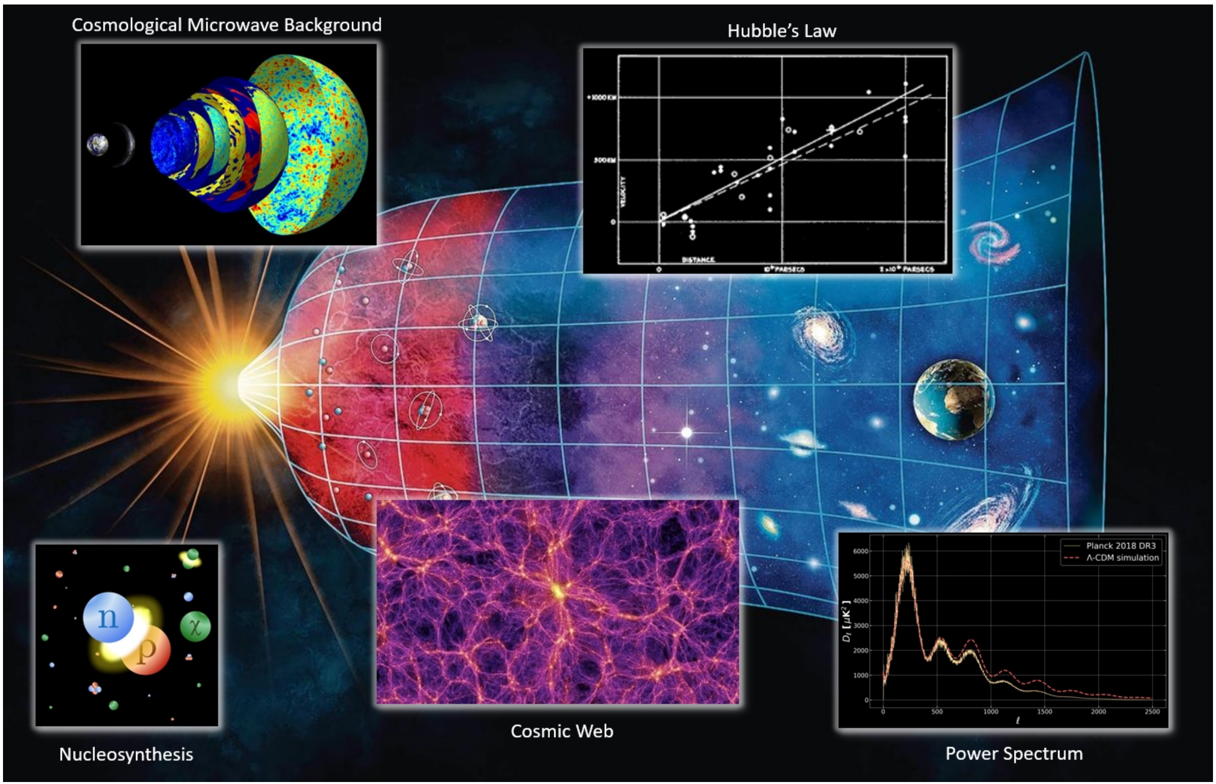


Figure 12.293: **ΛCDM theory experimental evidence as in the beginning of the 121st century**

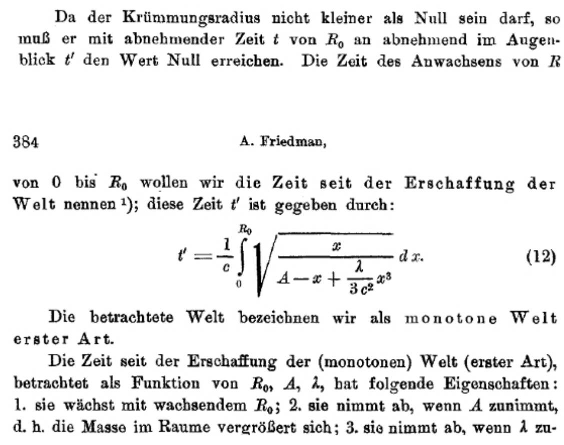
12.5.6.1 Robertson-Walker metric

We will now introduce an important metric for General Relativity and Cosmology that is the "Friedmann–Lemaître–Robertson–Walker (FLRW) metric". This metric is the basis of the Einstein cosmological model!

Remark

Contrary to a popular misconception, and as already mentioned in the Chronology section of this textbook (see page 9166), it was not Georges Lemaître who first developed a dynamic model of a Universe with a beginning in his 11927 (holocene calendar) publication [345] but Alexander Friedmann in his 11922 (holocene calendar) publication [193] (and Lemaître neither Friedman created the expression "Big Bang" but the astronomer Fred Hoyle in 11949 according to holocene calendar). Lemaître just rediscovered independently what Friedman already published before him but backed it up by a possible observational evidence trick! The fact that most people retain Georges Lemaître is mainly for three reasons: the first one is that he was a catholic priest and therefore it adds some spice to the legend (...) and the second is that Einstein knew already the paper of Friedmann but was hesitating and started really to change his mind about a static Universe when he saw that even a second scientist was able to fall back on the result of a dynamic Universe and the third and last one is that Friedmann died in 11925 so he wasn't here anymore to claim the discovery. For more, see [192] and [421].

For example in the paper of 11922 of Friedmann we can clearly check and see that Friedmann developed a model of the "World" (Welt) in expansion (i.e. Universe) as visible in the screenshot below of its original paper (for the detailed derivation see page 6050):



Da der Krümmungsradius nicht kleiner als Null sein darf, so muß er mit abnehmender Zeit t von R_0 an abnehmend im Augenblick t' den Wert Null erreichen. Die Zeit des Anwachsens von R

384 A. Friedman,

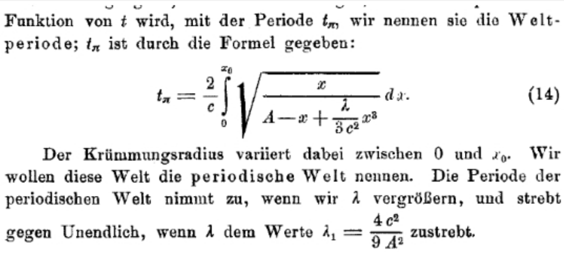
von 0 bis R_0 wollen wir die Zeit seit der Erschaffung der Welt nennen[1]); diese Zeit t' ist gegeben durch:

$$t' = \frac{1}{c}\int_0^{R_0}\sqrt{\frac{x}{A - x + \frac{\lambda}{3c^2}x^3}}\,dx. \qquad (12)$$

Die betrachtete Welt bezeichnen wir als monotone Welt erster Art.

Die Zeit seit der Erschaffung der (monotonen) Welt (erster Art), betrachtet als Funktion von R_0, A, λ, hat folgende Eigenschaften: 1. sie wächst mit wachsendem R_0; 2. sie nimmt ab, wenn A zunimmt, d. h. die Masse im Raume vergrößert sich; 3. sie nimmt ab, wenn λ zu-

He also developed the cyclic model of the "World" (i.e. Universe) as also visible in the screenshot below of its original paper:



Funktion von t wird, mit der Periode t_π, wir nennen sie die Weltperiode; t_π ist durch die Formel gegeben:

$$t_\pi = \frac{2}{c}\int_0^{x_0}\sqrt{\frac{x}{A - x + \frac{\lambda}{3c^2}x^3}}\,dx. \qquad (14)$$

Der Krümmungsradius variiert dabei zwischen 0 und x_0. Wir wollen diese Welt die periodische Welt nennen. Die Periode der periodischen Welt nimmt zu, wenn wir λ vergrößern, und strebt gegen Unendlich, wenn λ dem Werte $\lambda_1 = \frac{4c^2}{9A^2}$ zustrebt.

To see how we derive this metric with first consider the elementary distance on a 2D hypersurface:

$$\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2 \qquad (12.5.406)$$

on if the 2D hypersurface is a sphere, we have:

$$R^2 = x^2 + y^2 \Leftrightarrow y^2 = R^2 - x^2 \qquad (12.5.407)$$

For that latter relation if $R = c^{te}$ we have:

$$\mathrm{d}(R^2) = 0 = \mathrm{d}(x^2) + \mathrm{d}(y^2) = 2x\mathrm{d}x + 2y\mathrm{d}y \qquad (12.5.408)$$

After simplification:

$$0 = x\mathrm{d}x + y\mathrm{d}y \Leftrightarrow \mathrm{d}y = -x\frac{\mathrm{d}x}{y} \qquad (12.5.409)$$

Therefore:

$$\mathrm{d}s^2=\mathrm{d}x^2+\mathrm{d}y^2=\mathrm{d}x^2+\left(-\frac{x\mathrm{d}x}{y}\right)^2=\mathrm{d}x^2+\frac{x^2\mathrm{d}^2x}{y^2}=\mathrm{d}x^2+\frac{x^2\mathrm{d}^2x}{R^2-x^2}\tag{12.5.410}$$

We see that if $R\rightarrow+\infty$, then:

$$\mathrm{d}s^2=\mathrm{d}x^2\tag{12.5.411}$$

Now let us consider a 3D hypersurface in a 4D space such that:

$$\mathrm{d}s^2=\mathrm{d}x_1^2+\mathrm{d}x_2^2+\mathrm{d}x_3^2+\mathrm{d}x_4^2\tag{12.5.412}$$

The trick that explains why we put ourselves on the $\mathcal{S}^3$ sphere is that we anticipate in advanced that by doing a change of variable we will reduce this metric to three coordinates only!

We have also obviously:

$$R^2=x_1^2+x_2^2+x_3^2+x_4^2\Leftrightarrow x_4^2=R^2-(x_1^2+x_2^2+x_3^2)\tag{12.5.413}$$

If we assume $R=c^{te}$ we get by differentiation:

$$\mathrm{d}(R^2)=0=2x_1\mathrm{d}x_1+2x_2\mathrm{d}x_2+2x_3\mathrm{d}x_3+2x_4\mathrm{d}x_4\tag{12.5.414}$$

After simplification:

$$0=x_1\mathrm{d}x_1+x_2\mathrm{d}x_2+x_3\mathrm{d}x_3+x_4\mathrm{d}x_4\Leftrightarrow \mathrm{d}x_4=-\frac{x_1\mathrm{d}x_1+x_2\mathrm{d}x_2+x_3\mathrm{d}x_3}{x_4}\tag{12.5.415}$$

That is:

$$\begin{aligned}\mathrm{d}s^2&=\mathrm{d}x_1^2+\mathrm{d}x_2^2+\mathrm{d}x_3^2+\mathrm{d}x_4^2=\mathrm{d}x_1^2+\mathrm{d}x_2^2+\mathrm{d}x_3^2+\left(-\frac{(x_1\mathrm{d}x_1+x_2\mathrm{d}x_2+x_3\mathrm{d}x_3)}{x_4}\right)^2\\&=\mathrm{d}x_1^2+\mathrm{d}x_2^2+\mathrm{d}x_3^2+\frac{(x_1\mathrm{d}x_1+x_2\mathrm{d}x_2+x_3\mathrm{d}x_3)^2}{x_4^2}=\mathrm{d}x_1^2+\mathrm{d}x_2^2+\mathrm{d}x_3^2+\frac{(x_1\mathrm{d}x_1+x_2\mathrm{d}x_2+x_3\mathrm{d}x_3)^2}{R^2-(x_1^2+x_2^2+x_3^2)}\end{aligned}\tag{12.5.416}$$

Now we use spherical coordinates (see section Vector Calculus page 1849):

$$\vec{r}=\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\begin{pmatrix}r\sin(\theta)\cos(\phi)\\r\sin(\theta)\sin(\phi)\\r\cos(\theta)\end{pmatrix}\tag{12.5.417}$$

But as we know that in the general:

$$\begin{aligned}\mathrm{d}x_1^2+\mathrm{d}x_2^2+\mathrm{d}x_3^2=g_{11}\mathrm{d}x_1^2+g_{22}\mathrm{d}x_2^2+g_{33}\mathrm{d}x_3^2=g_r\mathrm{d}r^2+g_\theta\mathrm{d}\theta^2+g_\phi\mathrm{d}\phi^2\\=\mathrm{d}r^2+r^2\mathrm{d}\theta^2+r^2\sin^2(\theta)\mathrm{d}\phi^2\end{aligned}\tag{12.5.418}$$

Therefore having all this, we can rewrite:

$$\mathrm{d}s^2=\mathrm{d}x_1^2+\mathrm{d}x_2^2+\mathrm{d}x_3^2+\frac{(x_1\mathrm{d}x_1+x_2\mathrm{d}x_2+x_3\mathrm{d}x_3)^2}{R^2-(x_1^2+x_2^2+x_3^2)}\tag{12.5.419}$$

as following:

$$\begin{aligned}\mathrm{d}s^2 &= \mathrm{d}r^2+r^2\mathrm{d}\theta^2+r^2\sin^2(\theta)\mathrm{d}\phi^2+\frac{(\vec{r}\circ\mathrm{d}\vec{r})^2}{R^2-\|\vec{r}\|^2}\\ &= \mathrm{d}r^2+r^2\mathrm{d}\theta^2+r^2\sin^2(\theta)\mathrm{d}\phi^2+\frac{(\|\vec{r}\|\cdot\|\mathrm{d}\vec{r}\|\cos(\alpha))^2}{R^2-\|\vec{r}\|^2}\end{aligned} \tag{12.5.420}$$

We will denote obviously $\|\vec{r}\|$ as r, and $\|\mathrm{d}\vec{r}\|$ as $\mathrm{d}r$, and as they are collinear, we have $\cos(\alpha)=1$, therefore:

$$\begin{aligned}\mathrm{d}s^2 &= \mathrm{d}r^2+r^2\mathrm{d}\theta^2+r^2\sin^2(\theta)\mathrm{d}\phi^2+\frac{r^2\mathrm{d}r^2}{R^2-r^2}\\ &= \mathrm{d}r^2\left(\frac{r^2}{R^2-r^2}+1\right)+r^2(\mathrm{d}\theta^2+\sin^2(\theta)\mathrm{d}\phi^2)\\ &= \mathrm{d}r^2\frac{R^2}{R^2-r^2}+r^2(\mathrm{d}\theta^2+\sin^2(\theta)\mathrm{d}\phi^2)\end{aligned} \tag{12.5.421}$$

It is traditional to put:

$$\mathrm{d}\Omega^2=\mathrm{d}\theta^2+\sin^2(\theta)\mathrm{d}\phi^2 \tag{12.5.422}$$

and name it the "unit three-sphere metric" (also sometimes denoted $\mathrm{d}\Omega_3^2$). Therefore:

$$\mathrm{d}s^2=\mathrm{d}r^2\frac{R^2}{R^2-r^2}+r^2\mathrm{d}\Omega^2 \tag{12.5.423}$$

As for the comoving time introduce earlier before, let us introduce the "comoving coordinates" as being:

$$r=Rx\Leftrightarrow\mathrm{d}r=R\mathrm{d}x \tag{12.5.424}$$

Therefore:

$$\begin{aligned}\mathrm{d}s^2 &= R^2\mathrm{d}x^2\frac{R^2}{R^2-R^2x^2}+R^2x^2\mathrm{d}\Omega^2=R^2\mathrm{d}x^2\frac{R^2}{R^2-R^2x^2}+R^2x^2\mathrm{d}\Omega^2\\ &= R^2\left(\frac{\mathrm{d}x^2}{1-x^2}+x^2\mathrm{d}\Omega^2\right)\end{aligned} \tag{12.5.425}$$

If we had made the same development but with the hyperbolic metric (see section Differential Geometry page 2855), we would have obtain:

$$\mathrm{d}s^2=R^2\left(\frac{\mathrm{d}x^2}{1+x^2}+x^2\mathrm{d}\Omega^2\right) \tag{12.5.426}$$

This is why we write more generally:

$$\mathrm{d}s^2=R^2\left(\frac{\mathrm{d}x^2}{1-kx^2}+x^2\mathrm{d}\Omega^2\right) \tag{12.5.427}$$

with:

- If $k=0$ we fall back on the flat Euclidean space (but don't forget we are on comoving coordinates!)

- If $k=+1$ we are on a spherical space
- If $k=-1$ we are in a hyperbolic space

Then the metric tensor including time coordinates with $(-,+,+,+)$ signature is given by:

$$g_{\mu\nu}=\begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & \dfrac{R^2}{1-kx^2} & 0 & 0 \\ 0 & 0 & R^2x^2 & 0 \\ 0 & 0 & 0 & R^2x^2\sin^2(\theta) \end{bmatrix} \quad (12.5.428)$$

So that we can write obviously if $x^\mu=(ct,x^i)$:

$$\mathrm{d}s^2=g_{\mu\nu}\mathrm{d}x^\mu\mathrm{d}x^\nu=-(c\mathrm{d}t)^2+R^2\left(\frac{\mathrm{d}x^2}{1-kx^2}+x^2\mathrm{d}\Omega^2\right) \quad (12.5.429)$$

But R can be an arbitrary function of time so we write traditionally:

$$\mathrm{d}s^2=g_{\mu\nu}\mathrm{d}x^\mu\mathrm{d}x^\nu=-(c\mathrm{d}t)^2+R^2(t)\left(\frac{\mathrm{d}x^2}{1-kx^2}+x^2\mathrm{d}\Omega^2\right) \quad (12.5.430)$$

and as we have seen earlier it is quite common to denote the radius by a , therefore:

$$\boxed{\mathrm{d}s^2=g_{\mu\nu}\mathrm{d}x^\mu\mathrm{d}x^\nu=-(c\mathrm{d}t)^2+a^2(t)\left(\frac{\mathrm{d}x^2}{1-kx^2}+x^2\mathrm{d}\Omega^2\right)} \quad (12.5.431)$$

and this is the final form of the Friedmann–Lemaître–Robertson–Walker (FLRW) metric..

Or also written with another signature:

$$\mathrm{d}s^2=g_{\mu\nu}\mathrm{d}x^\mu\mathrm{d}x^\nu=(c\mathrm{d}t)^2-a^2(t)\left(\frac{\mathrm{d}x^2}{1-kx^2}+x^2\mathrm{d}\Omega^2\right) \quad (12.5.432)$$

Sometimes denoted:

$$\mathrm{d}s^2=N^2(t)\mathrm{d}^2t-a^2(t)\mathrm{d}\Omega_3^2 \quad (12.5.433)$$

where $\mathrm{d}\Omega_3^2$ is sometimes named the "hyperboloid metric" and $N(t)$ is chosen sometimes as an arbitrary lapse function.

Remark

In a more general framework that $\mathrm{d}\Omega_3^2$ is defined as:

$$\mathrm{d}\Omega_3^2:=h_{ij}\mathrm{d}x^i\mathrm{d}x^j=\frac{\delta_{ij}\mathrm{d}x^i\mathrm{d}x^j}{\left(1+\dfrac{1}{4}\delta_{ij}x^ix^j\right)^2} \quad (12.5.434)$$

It is important to remember that x has no units and is often denoted in textbooks with the following letter χ and that rigorously the scale factor $a(t)$ is defined by $a(t)/a(t_0)$. Keep also in mind that one of the main idea is to determine the explicit expression of $a(t)$ (Universe models) to inject it afterwards in the FLRW metrics to make further investigation on the properties of the visible Universe.

Metric Name	**Formulation**	**Characertistics**
Minkowski	$\mathrm{d}s^2_{\text{M}}=-c^2\mathrm{d}t+\mathrm{d}x^2+\mathrm{d}y^2+\mathrm{d}z^2$	Flat space Static space Non-charged space
Schwarzschild	$\mathrm{d}s^2_{\text{S}}=\left(1-\dfrac{2GM}{c^2r}\right)c^2\mathrm{d}t-\dfrac{1}{1-\dfrac{2GM}{c^2r}}\mathrm{d}r^2-r^2\left(\mathrm{d}\theta^2+\sin^2(\theta)\mathrm{d}\phi^2\right)$	Spherical symmetric space Static space Valide only outside a mass Non-charged space
Robertson-Walker	$\mathrm{d}s^2_{\text{RW}}=-c\mathrm{d}t^2+a^2(t)\left(\dfrac{\mathrm{d}x^2}{1-kx^2}+x^2\left(\mathrm{d}\theta^2+\sin^2(\theta)\mathrm{d}\phi^2\right)\right)$	Isotropic space Homogeneous space Symmetric space Expanding/Contracting space Non-rotating space Non-charged space

Table 12.26: **Summary of different types of General Relativity metrics derived so far**

The main results of the FLRW model were first derived by the Soviet mathematician Alexander Friedmann in 11922 (holocene calendar) and 11924 (holocene calendar). Although his work was published in the prestigious physics journal *Zeitschrift für Physik*, it remained relatively unnoticed by his contemporaries. Friedmann was in direct communication with Albert Einstein, who, on behalf of *Zeitschrift für Physik*, acted as the scientific referee of Friedmann's work. Eventually Einstein acknowledged the correctness of Friedmann's calculations, but failed to appreciate the physical significance of Friedmann's predictions.

Friedmann died in 11925 (holocene calendar). In 11927 (holocene calendar), Georges Lemaître, a Belgian priest, astronomer and periodic professor of physics at the Catholic University of Leuven, arrived independently at similar results as Friedmann had and published them in *Annals of the Scientific Society of Brussels*. In the face of the observational evidence for the expansion of the universe obtained by Edwin Hubble in the late 11920s (holocene calendar), Lemaître's results were noticed in particular by Arthur Eddington, and in 11930–31 (holocene calendar) his paper was translated into English and published in the *Monthly Notices of the Royal Astronomical Society*.

Howard P. Robertson from the US and Arthur Geoffrey Walker from the UK explored the problem further during the 11930s (holocene calendar). In 11935 (holocene calendar) Robertson and Walker rigorously proved that the FLRW metric is the only one on a space-time that is spatially homogeneous and isotropic (as noted above, this is a geometric result and is not tied specifically to the equations of General Relativity, which were always assumed by Friedmann and Lemaître).

The reader can subjectively judge of the homogeneity of a part of the observable universe with the 3D reconstruction below of 15 million galaxies and quasars up to 11 billion light years in distance :

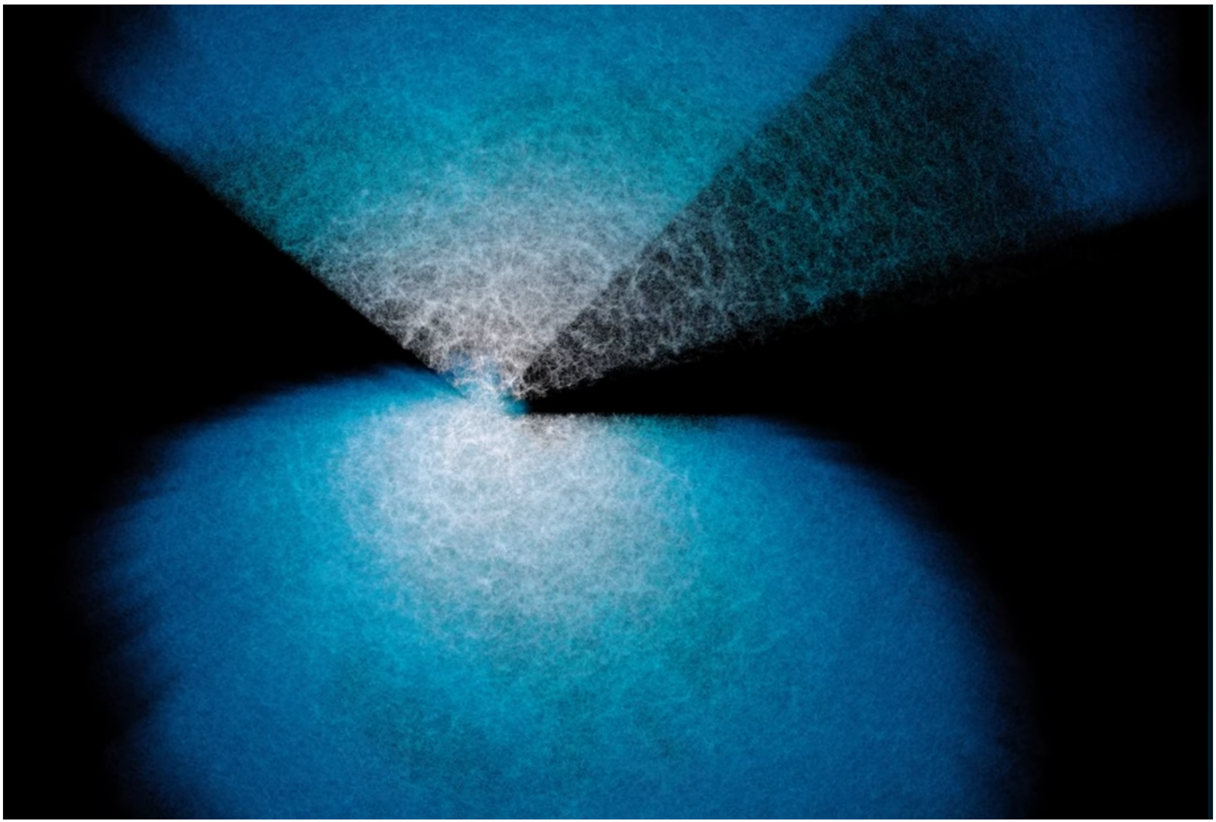


Figure 12.294: **DESI galaxies survey (credit: DESI collaboration and KPNO/NOIRLab/NSF/AURA/R. Proctor)**

Because the dynamics of the FLRW model were derived by Friedmann and Lemaître, the latter two names are often omitted by scientists outside the US. Conversely, US physicists often refer to it as simply "Robertson–Walker metric". The full four-name title seems to be the most democratic. Often the "Robertson–Walker" metric, so-named since they proved its generic properties, is distinguished from the dynamical "Friedmann-Lemaître" models, specific solutions for $a(t)$ which assume that the only contributions to stress-energy are cold matter ("dust"), radiation, and a cosmological constant.

12.5.6.2 Cosmological Redshift

Before dealing with Universe models dealing with the FLRW metric, let us consider two points in space participating in the expansion of the Universe. At one point we have a source of light and at the other an observer (us). Let us use spherical coordinates centered on the observer, so that the comoving coordinate of the observer is $c=0$ and the comoving coordinate of the source is $\chi=0$. Let us analyse the propagation of a photon emitted by the source at time t_e and received at time t_0. As the photon propagates radially towards the observer, along its trajectory in space, and hence along its worldline in space-time we have $\theta=c^{te}$ and $\phi=c^{te}$. Moreover, along the worldlines of photons (Friedmann–Lemaître–Robertson–Walker metric):

$$\begin{aligned}\mathrm{d}s^2=g_{\mu\nu}\mathrm{d}x^\mu\mathrm{d}x^\nu&=-(c\mathrm{d}t)^2+a^2(t)\left(\frac{\mathrm{d}r^2}{1-kr^2}+r^2\mathrm{d}\Omega^2\right)\\&=-(c\mathrm{d}t)^2+a^2(t)\left[\frac{\mathrm{d}r^2}{1-kr^2}+r^2\left(\mathrm{d}\theta^2+\sin^2(\theta)\mathrm{d}\phi^2\right)\right]\\&=-(c\mathrm{d}t)^2+a^2(t)\frac{\mathrm{d}r^2}{1-kr^2}=0\end{aligned}\tag{12.5.435}$$

Hence:

$$\frac{c\mathrm{d}t}{a(t)}=\pm\frac{\mathrm{d}r}{\sqrt{1-kr^2}}\tag{12.5.436}$$

Notice that we know how to calculate the three possible integrals depending on the value of k (see section Differential and Integral Calculus page 1534):

$$\int\frac{c\mathrm{d}t}{a(t)}=\pm\int\frac{\mathrm{d}r}{\sqrt{1-kr^2}}=\pm\begin{cases}\arcsin(r) & k=+1\\ r & k=0\\ \text{arcsinh}(r) & k=-1\end{cases}\tag{12.5.437}$$

Remark

Using the notation of the conformal time defined earlier above, this latter relation is sometimes written:

$$c\mathrm{d}\eta=\pm\frac{\mathrm{d}r}{\sqrt{1-kr^2}}\tag{12.5.438}$$

If we put $c=1$ we see that time and distance may be interchanged as:

$$\boxed{\Delta D=\Delta\eta=\int\frac{\mathrm{d}t}{a}=\int\frac{\mathrm{d}a}{a\dot{a}}=\int\frac{\mathrm{d}a}{a^2H(a)}}\tag{12.5.439}$$

For an observer observing the crest of a light wave at a position $r=0$ and time $t=t_{\text{now}}$, the crest of the light wave was emitted at a time $t=t_{\text{then}}$ in the past and a distant position $r=R$. Integrating over the path in both space and time that the light wave travels yields:

$$c\int_{t_{\text{then}}}^{t_{\text{now}}}\frac{\mathrm{d}t}{a}=\int_R^0\frac{\mathrm{d}r}{\sqrt{1-kr^2}}\tag{12.5.440}$$

In general, the wavelength of light is not the same for the two positions and times considered due to the changing properties of the metric. When the wave was emitted, it had a wavelength λ_{then}. The next crest of the light wave was emitted at a time:

$$t = t_{\text{then}} + \frac{\lambda_{\text{then}}}{c} \tag{12.5.441}$$

The observer sees the next crest of the observed light wave with a wavelength λ_{now} to arrive at a time:

$$t = t_{\text{now}} + \frac{\lambda_{\text{now}}}{c} \tag{12.5.442}$$

Since the subsequent crest is assumed approximately again emitted from $r = R$ and is observed at $r = 0$, the following equation can be written:

$$c \int\limits_{t_{\text{then}}+\lambda_{\text{then}}/c}^{t_{\text{now}}+\lambda_{\text{now}}/c} \frac{\mathrm{d}t}{a} \cong \int\limits_{R}^{0} \frac{\mathrm{d}r}{\sqrt{1-kr^2}} \tag{12.5.443}$$

The right-hand side of the two integral equations above are supposed as approximately equal given the small time variation, which means:

$$c \int\limits_{t_{\text{then}}+\lambda_{\text{then}}/c}^{t_{\text{now}}+\lambda_{\text{now}}/c} \frac{\mathrm{d}t}{a} \cong c \int\limits_{t_{\text{then}}}^{t_{\text{now}}} \frac{\mathrm{d}t}{a} \tag{12.5.444}$$

Remark

The approximation is considered as valid by professional astrophysicists and cosmologist because the first lights were emitted after the recombination epoch (i.e. time of the cosmological microwave background) estimated at $380,000$ years after the Big Bang when the inflation period (speculation of an exponential expansion of space in the early universe that would violated the assumption made just above) is supposed to have lasted from 10^{-36} seconds to between 10^{-33} and 10^{-32} seconds after the Big Bang. And keep in mind that the actual value of the Hubble constant is in SI units approximately equal to $2.4 \cdot 10^{-18}\;[\mathrm{s}^{-1}]$ and that's indeed (locally) a very small rate of expansion.

Using the following manipulation:

$$\begin{aligned}
0 &\cong \int\limits_{t_{\text{then}}}^{t_{\text{now}}} \frac{\mathrm{d}t}{a} - \int\limits_{t_{\text{then}}+\lambda_{\text{then}}/c}^{t_{\text{now}}+\lambda_{\text{now}}/c} \frac{\mathrm{d}t}{a} \\
&= \int\limits_{t_{\text{then}}}^{t_{\text{then}}+\lambda_{\text{then}}/c} \frac{\mathrm{d}t}{a} + \int\limits_{t_{\text{then}}+\lambda_{\text{then}}/c}^{t_{\text{now}}} \frac{\mathrm{d}t}{a} - \int\limits_{t_{\text{then}}+\lambda_{\text{then}}/c}^{t_{\text{now}}+\lambda_{\text{now}}/c} \frac{\mathrm{d}t}{a} \\
&= \int\limits_{t_{\text{then}}}^{t_{\text{then}}+\lambda_{\text{then}}/c} \frac{\mathrm{d}t}{a} - \left(\int\limits_{t_{\text{now}}}^{t_{\text{then}}+\lambda_{\text{then}}/c} \frac{\mathrm{d}t}{a} + \int\limits_{t_{\text{then}}+\lambda_{\text{then}}/c}^{t_{\text{now}}+\lambda_{\text{now}}/c} \frac{\mathrm{d}t}{a} \right) \\
&= \int\limits_{t_{\text{then}}}^{t_{\text{then}}+\lambda_{\text{then}}/c} \frac{\mathrm{d}t}{a} - \int\limits_{t_{\text{now}}}^{t_{\text{now}}+\lambda_{\text{now}}/c} \frac{\mathrm{d}t}{a}
\end{aligned} \tag{12.5.445}$$

we find that:

$$\int\limits_{t_{\text{now}}}^{t_{\text{now}}+\lambda_{\text{now}}/c}\frac{\mathrm{d}t}{a}\cong\int\limits_{t_{\text{then}}}^{t_{\text{then}}+\lambda_{\text{then}}/c}\frac{\mathrm{d}t}{a}\qquad(12.5.446)$$

For very small variations in time (over the period of one cycle of a light wave), we assume that the scale factor is essentially a constant ($a=a_{\text{now}}$ today and $a=a_{\text{then}}$ previously). This yields:

$$\frac{t_{\text{now}}+\lambda_{\text{now}}/c}{a_{\text{now}}}-\frac{t_{\text{now}}}{a_{\text{now}}}\cong\frac{t_{\text{then}}+\lambda_{\text{then}}/c}{a_{\text{then}}}-\frac{t_{\text{then}}}{a_{\text{then}}}\qquad(12.5.447)$$

which can be rewritten as:

$$\frac{\lambda_{\text{now}}}{\lambda_{\text{then}}}\cong\frac{a_{\text{now}}}{a_{\text{then}}}\qquad(12.5.448)$$

In analogy with the Doppler redshift provided earlier above, the relation:

$$\boxed{1+z\cong\frac{a_{\text{now}}}{a_{\text{then}}}\cong\frac{\lambda_{\text{now}}}{\lambda_{\text{then}}}=\frac{\nu_{\text{then}}}{\nu_{\text{now}}}}\qquad(12.5.449)$$

defines the "cosmological redshift" or"expansion redshift". Remarkably, with these approximations, the curvature of space does not appear in the final result. So redshift is related to the expansion factor of the Universe. If we measure a redshift of $z=2$, the Universe is $3\times$ bigger now than it was when that photon was emitted.

If rearranged the above relation is also famous as:

$$z\cong\frac{\lambda_{\text{now}}}{\lambda_{\text{then}}}-1=\frac{\lambda_{\text{now}}-\lambda_{\text{then}}}{\lambda_{\text{then}}}=\frac{\Delta\lambda}{\lambda_{\text{then}}}\qquad(12.5.450)$$

The most distant object observed so far at the day we write these lines is JADES-GS-z13-0 (a dark star) with a redshift of $z=13.20^{+0.04}_{-0.07}$. This correspond to a loss of energy of the emitted photon of:

$$\frac{\Delta E}{E}=\frac{h\Delta\nu}{h\nu_{\text{then}}}=\frac{(\nu_{\text{then}}-\nu_{\text{now}})}{\nu_{\text{then}}}=\frac{\left(\nu_{\text{then}}-\dfrac{\nu_{\text{then}}}{1+z}\right)}{\nu_{\text{then}}}=\left(1-\frac{1}{1+z}\right)=\frac{z}{z+1}\qquad(12.5.451)$$

Hence using the above relation for JADES-GS-z13-0 we can easily calculate that the photon has lost approximately 92.95% of its original energy.

In an expanding universe, the scale factor is monotonically increasing as time passes, thus, z is positive and distant galaxies appear red-shifted. it follows that as a photon travels across our expanding Universe its wavelength increases and its frequency decreases. Thus, the energy of the photon also decreases!

The previous relations shows that the redshift of a distant source is a measure of the total expansion of the universe that has occurred while the light was travelling between the source and the observer. It does not tell us the distance to the source or how long ago the light was emitted. These quantities depend on the precise nature of the Universe expansion between the instances of emission and observation. In different cosmological models we obtain different results for the same redshift!!! The reader must also keep in mind that the Dopplershift, Gravitational Redshift and Cosmological redshift from expansion are the same thing in the framework of General Relativity.

Obviously the redshift seems to violate the principle of conservation of energy. However remember that we have proved that this principle is valid only in a space that is translational symmetric in the time (see section Mechanics page 3132). However a quite common question that remains is: where does the energy difference go? Well if you remember our study of the wavepacket in the section of Wave Quantum Physics (see page 4786), we have proven that light is in reality infinitely spread even if the majority of its energy is especially concentrated in a small volume. We have also proven in the section of Quantum Field Theory (see page 5176) that light can be described as the excitation of an electromagnetic field that exists everywhere in the Universe. So what the reader must keep in mind, is that during the time the photon travels, a part of space-time (and hence a part of the energy of the photon that is spread infinitely through the whole space time of the Universe!) disappears beyond the Hubble horizon as our visible Universe is expanding and this is where the energy difference goes (!!!) as anything beyond this horizon is not causally related to us anymore!

Remark

The existence of the cosmological redshift explains the "Olbers' paradox" (named after the German astronomer Heinrich Wilhelm Olbers) or also known as the "dark night sky paradox". The darkness of the night sky is one of the pieces of evidence for a dynamic Universe, such as the Big Bang model (cosmic inflation theory). In the hypothetical case that the Universe is static, homogeneous at a large scale, and populated by an infinite number of stars, then any line of sight from Earth must end at the (very bright) surface of a star and hence the night sky should be completely illuminated and very bright. The Swiss mathematician Jean Philippe Loys de Cheseaux explains this paradox mathematically in 11746 (holocene calendar) by calculating that the light energy falling on Earth should be $180,000$ times more intense than that of the Sun. In 11823 (holocene calendar), Olbers refine this reasoning by observing that in a uniformly filled universe of stars, the stars are hidden from each other and deduces that the brightness of the night sky can not be infinite but at most equal to the surface brightness of a star.

This contradicts obviously the observed darkness and non-uniformity of the night as many stars are above the cosmological horizon and also a significant fraction of stars have their light that is redshifted and hence invisible to human eyes!

Notice that while dark clouds could obstruct the light, these clouds would heat up, until they were as hot as the stars, and then radiate the same amount of light.

12.5.6.3 Comobile Universe Diameter

As promised earlier, it is now time for us to determine the comobile Universe diameter using General Relativity and especially with the special case of the FLRW metric!

For this purpose we start again from the same assumption as for the cosmological redshift, i.e.:

$$\begin{aligned}\mathrm{d}s^2=g_{\mu\nu}\mathrm{d}x^\mu\mathrm{d}x^\nu&=-(c\mathrm{d}t)^2+a^2(t)\left(\frac{\mathrm{d}r^2}{1-kr^2}+r^2\mathrm{d}\Omega^2\right)\\&=-(c\mathrm{d}t)^2+a^2(t)\left[\frac{\mathrm{d}r^2}{1-kr^2}+r^2\left(\mathrm{d}\theta^2+\sin^2(\theta)\mathrm{d}\phi^2\right)\right]\\&=-(c\mathrm{d}t)^2+a^2(t)\frac{\mathrm{d}r^2}{1-kr^2}=0\end{aligned}\tag{12.5.452}$$

The horizon $d_H(t_1,t_2)$ is defined as the physical distance at time t_2 between two particles emitted at the same point but in opposite directions at time t_1, and travelling at the speed of light. If the origin of spherical comobile coordinates is chosen to coincide with the point of emission, the physical distance at time t_2 can be computed by integrating over small distance elements $\mathrm{d}l$ between the origin and the position r_2 of one particle, and multiplying by two:

$$d_H(t_1,t_2)=2\int\limits_0^{r_2}\mathrm{d}l=2\int\limits_0^{r_2}a(t_2)\frac{\mathrm{d}r}{\sqrt{1-kr^2}}\tag{12.5.453}$$

In addition, the geodesic equation for ultra-relativistic particles gives $\mathrm{d}s=0$, i.e.:

$$\mathrm{d}t=a(t)\frac{\mathrm{d}r}{\sqrt{1-kr^2}}\tag{12.5.454}$$

which can be integrated along the trajectory of the particles:

$$\int\limits_{t_1}^{t_2}\frac{\mathrm{d}t}{a(t)}=\int\limits_0^{r_2}\frac{\mathrm{d}r}{\sqrt{1-kr^2}}\tag{12.5.455}$$

We can now replace in the expression of d_H and get:

$$d_H(t_1,t_2)=2a(t_2)\int\limits_{t_1}^{t_2}\frac{\mathrm{d}t}{a(t)}\tag{12.5.456}$$

Usually, the result is presented in this form. However, for the following discussion, it is be particularly useful to eliminate the time from the integral by remembering that:

$$\mathrm{d}t=\frac{\mathrm{d}a}{aH}\tag{12.5.457}$$

Hence:

$$d_H(a_1,a_2)=2a_2\int\limits_{a_1}^{a_2}\frac{\mathrm{d}a}{a^2H(a)}\tag{12.5.458}$$

where the Hubble parameter is seen now as a function of a. Let us assume that t_1 and t_2 are two times during radiation domination. We know from the Friedmann equation that during radiation domination one has:

$$H \propto a^{-2} \tag{12.5.459}$$

so we can parametrize the Hubble rate as:

$$H(a) = H_2(a_2/a)^2 \tag{12.5.460}$$

We get:

$$d_H(a_1,a_2) = 2a_2 \int\limits_{a_1}^{a_2} \frac{\mathrm{d}a}{a_2^2 H(a)} = \frac{2}{H_2}\frac{(a_2-a_1)}{a_2} \tag{12.5.461}$$

If the time t_2 is much after t_1 so that $a_2 \gg a_1$, the expression for the horizon does not depend on a_1:

$$d_H(a_1,a_2) \cong \frac{2}{H_2} \tag{12.5.462}$$

So, the horizon equals twice the Hubble (comobile) radius at time t_2:

$$\boxed{d_H(t_1,t_2) \cong 2r_{\text{HS}}(t_2)} \tag{12.5.463}$$

This relation (that can be derived using some other ways more or less approximative but all assuming the radiation dominated era!) can lead us to a quite interesting result.

Indeed, during our basic study of the CMB (see page 5973) we have determined the original decoupling time t_{dec}. That latter relation can be injected in the previous relation that lead us to:

$$\begin{aligned} d_H(t_1,t_2) &\cong 2r_{\text{HS}}(t_2) = 2(1+z_{\text{dec}})ct_{\text{rec}} \cong 2(1+1,100)\cdot 380,000\,[\text{light years}] \\ &= 836,760,000\,[\text{light years}] \cong 255\,[\text{Mpc}] \end{aligned} \tag{12.5.464}$$

That means simply that at the time we write these lines, two points that were connected at the decoupling time, are today separated from a comobile distance (radius) of approximately 255 [Mpc] (assuming all assumptions are verified...!) on our Hubble sphere. As we have multiplied by the speed of light, this calculation assumes that today such points are causally disconnected regions of space! So this comobile distance represents like a lower-limit of two causally disconnected regions of space after the decoupling time.

12.5.6.4 Size of the Universe depending on its curvature

Many articles and videos of various social networks claim that because that our observable universe appears to be flat, then our whole universe must therefore be infinite but they never explain the underlying mathematical reasoning at all nor the calculation that leads to the approximately 40 billion light years of radius!

Remark

For reminder, the results of the Planck mission (see [446]), released in 12018 (holocene calendar) suggest that the cosmological curvature parameter, $1-\Omega=\Omega_k=-kc^2/a^2H^2$, to be 0.0007 ± 0.0019 (68%, TT,TE,EE+lowE+lensing+BAO) to a 1σ accuracy of 0.2%, consistent with a flat universe (i.e. positive curvature: $k=+1,\Omega_k<0,\Omega>1$, negative curvature: $k=-1,\Omega_k>0,\Omega<1$, zero curvature: $k=0,\Omega_k=0$, $\Omega=1$).

First, we need to be precise about the phrase «*the size of our Universe*». Specifically we are going to take it to mean the maximum possible separation between any two points. In an infinite universe two points can be separated by an arbitrarily large distance, so if the maximum distance between two points is finite this means the universe must not be infinite.

The point of all this is that the distance between any two points is calculated using a metric. For a Friedmann universe like ours (at least we assume our universe to be a Friedmann universe!) the metric is in natural units and in polar coordinates (see page 5991):

$$\mathrm{d}s^2=-\mathrm{d}t^2+a^2(t)\left[\frac{\mathrm{d}r^2}{1-kr^2}+r^2\mathrm{d}\Omega^2\right] \tag{12.5.465}$$

The value of the parameter k determines as we already know whether the universe (or just the observable universe) is closed, flat or open. Specifically $k>0$ is a closed universe, $k=0$ is a flat universe and $k<0$ is an open universe. The variable s is the proper distance.

Now, suppose we choose an origin at some starting point, choose a fixed time, and calculate the proper distance, s as we move radially away from the starting point. The question is whether s can reach infinity or not. Because only r is changing $\mathrm{d}t=\mathrm{d}\Omega=0$, so the expression for the proper distance simplifies to:

$$\mathrm{d}s^2=a^2(t)\frac{\mathrm{d}r^2}{1-kr^2} \tag{12.5.466}$$

We will choose our units of distance to make $a=1$, and we'll consider only closed or flat space, i.e. $k\geq 0$, in which case we can integrate:

$$\mathrm{d}s^2=a^2(t)\frac{\mathrm{d}r^2}{1-kr^2}\Rightarrow\mathrm{d}s=\frac{\mathrm{d}r}{\sqrt{1-kr^2}}=\frac{\mathrm{d}r}{\sqrt{1-(\sqrt{k}r)^2}} \tag{12.5.467}$$

that gives (see section Differential and Integral Calculus page 1534):

$$s(r)=\frac{\arcsin(\sqrt{k}r)}{\sqrt{k}} \tag{12.5.468}$$

So the maximum possible value for $s(r)$ is when $\sqrt{k}r=1$, in which case:

$$s_{\text{max}}=\frac{\pi}{2\sqrt{k}} \quad (12.5.469)$$

And there's the result we want. For a closed universe $k>0$ and therefore the maximum possible distance between two points is finite. However as $k\rightarrow 0_+$ the maximum possible distance $s_{\text{max}}\rightarrow +\infty$. That's why a flat universe is infinite.

However we should note that, that even a flat universe can be finite if it has a non-trivial topology (like a torus that big enough can locally appear to be flat!).

If we don't put $a(t)=1$ as we did above, the solutions are commonly denoted:

$$s(t)=a(t)\cdot\begin{cases}\frac{1}{\sqrt{k}}\sin^{-1}(r\sqrt{k}) & \text{for } k>0\\ r & \text{for } k=0\\ \frac{1}{\sqrt{|k|}}\sinh^{-1}(r\sqrt{|k|}) & \text{for } k<0\end{cases} \quad (12.5.470)$$

In a flat Universe, the proper distance to an object is just its coordinate distance, $s(t)=a(t)\cdot r$. However, because the r coordinate was introduced in the context of a flat Newtonian Universe, the coordinate distance will generally not agree with the proper distance. Because $\sin^{-1}(x)\geq x$ and $\sinh^{-1}(x)\leq x$, in a closed Universe $(k>0)$ the proper distance to an object is greater than its coordinate distance, while in an open Universe $(k<0)$ the proper distance to an object is less than its coordinate distance.

Remark

Instead of taking one measurement method, a team of scientists led by Mihran Vardanyan at the University of Oxford (see [555]) did a statistical analysis of all of the results. By using Bayesian model averaging, which focuses on how likely a model is to be correct given the data, rather than asking how well the model itself fits the data. They found that the universe, we quote: «*...The minimum size of the Universe is robustly constrained to encompass* $N_U\gtrsim 251$ *Hubble spheres...*»

Again we return to the Friedmann-Lemaître-Robertson-Walker metric:

$$\mathrm{d}s^2=c\mathrm{d}t^2-a^2(t)\left[\frac{\mathrm{d}r^2}{1-kr^2}+r^2\left(\mathrm{d}\theta^2+\sin^2(\theta)\mathrm{d}\phi^2\right)\right] \quad (12.5.471)$$

and we exploit the fact that in General Relativity the propagation of light is along a null geodesic $(\mathrm{d}s=0)$. With the observer (i.e. us, on Earth) at the origin $(r=0)$, we choose a radial null geodesic so that $\mathrm{d}\theta=\mathrm{d}\phi=0$, and the previous relation reduces to:

$$\frac{c\mathrm{d}t}{a(t)}=\pm\frac{\mathrm{d}r}{(1-kr^2)^{1/2}} \quad (12.5.472)$$

where the $+$ sign corresponds to an emitted light ray and the $-$ sign to a received one.

Placing an observer at the origin $(r=0)$ and let the particle horizon for this observer at time t be located at radial coordinate distance r_{hor}. This means that a photon emitted at $t=0$ at r_{hor} will reach our observer at the origin at time t the previous relation can be rewritten:

$$\int\limits_0^t\frac{\mathrm{d}t}{a(t)}=\frac{1}{c}\int\limits_0^{r_{\text{hor.}}}\frac{\mathrm{d}r}{(1-kr^2)^{1/2}} \quad (12.5.473)$$

from which we have immediately that:

$$r_{\text{hor.}}=\begin{cases}\sin\left(c\int\limits_0^t\frac{\mathrm{d}t}{a(t)}\right) & \text{for } k=1\\ c\int\limits_0^t\frac{\mathrm{d}t}{a(t)} & \text{for } k=0\\ \sinh\left(c\int\limits_0^t\frac{\mathrm{d}t}{a(t)}\right) & \text{for } k=-1\end{cases}\qquad(12.5.474)$$

Let us recall that the proper distance from the origin to $r_{\text{hor.}}$ is given by as we have just seen earlier above:

$$s_{\text{hor.}}(t)=a(t)\int\limits_0^{r_{\text{hor.}}}\frac{\mathrm{d}r}{(1-kr^2)^{1/2}}\qquad(12.5.475)$$

and combining it with:

$$\int\limits_0^t\frac{\mathrm{d}t}{a(t)}=\frac{1}{c}\int\limits_0^{r_{\text{hor.}}}\frac{\mathrm{d}r}{(1-kr^2)^{1/2}}\qquad(12.5.476)$$

we get the proper distance to the horizon that is found to be:

$$\boxed{s_{\text{hor.}}(t)=a(t)\int\limits_0^t\frac{c\mathrm{d}t}{a(t)}}\qquad(12.5.477)$$

and this is a nice expression as k disappears! This is the relation we will use further below when introducing the different FLRW Universe models (see page 6014) to determine their corresponding size!

Remember that at page 5914, we have proved that in the case without cosmological constant that today (this is why we have H_0 instead of H and Ω_0 instead of Ω and R_0 instead of $R(t)$):

$$H_0^2(1-\Omega_0)=-\frac{k}{R_0^2c^2}\qquad(12.5.478)$$

Hence:

$$\boxed{R_0=\frac{c\sqrt{k}}{H_0\sqrt{1-\Omega_0}}}\qquad(12.5.479)$$

In we assume a spherical universe, i.e. $k=1$ we get:

$$R_0=\frac{c}{H_0\sqrt{1-\Omega_0}}\qquad(12.5.480)$$

As by definition (for reminder Ω_k is not an actual mass/energy density and is defined as $\Omega_k=1-\Omega_0$ so that $\Omega_k=0$ for a spatially flat universe):

$$\Omega_k=1-\Omega_0\qquad(12.5.481)$$

Therefore:

$$R_0 = \frac{c}{H_0\sqrt{\Omega_k}} \qquad (12.5.482)$$

From [446], we have at 1σ:

$$\Omega_k = 0.0007 \pm 0.0019 \qquad (12.5.483)$$

If the universe is spherical, its minimum size corresponds at 1σ in absolute value to:

$$\Omega_k = 0.0012 \qquad (12.5.484)$$

If the universe is spherical with $H_0 = 67.66\,[\text{km}\cdot\text{s}\cdot\text{Mpc}^{-1}]$, its minimum radius of curvature is:

$$R = \frac{c}{H_0}\cdot\frac{1}{\sqrt{\Omega_k}} \cong \frac{299,792.458}{67.66}\cdot\frac{1}{\sqrt{0.0012}} = 127,908\,[\text{Gly}] = 417\,[\text{Gly}] \qquad (12.5.485)$$

As we prove it at page 6048, the proper size of the observable universe assumed as flat is:

$$R_{\Lambda=0,k=0} \cong 3ct \cong 41.31 \pm 0.12\,[\text{Gly}] \qquad (12.5.486)$$

Therefore if the universe is spherical, its radius is at least 10 times larger than the observable universe!

If we prefer to give the result as a function of Hubble spheres, we know that the actual Hubble radius is (see page 5888):

$$R_H \cong 14.5\,[\text{Gly}] \qquad (12.5.487)$$

Therefore if the universe is spherical, its radius is at least 29.3 times larger than the Hubble radius.

12.5.6.5 General Friedmann equations

It follows now, that using the second signature and changing x to r' we get:

$$g_{00}=1,\, g_{r'r'}=\frac{-a^2}{1-kr'^2},\, g_{\theta\theta}=-a^2r'^2,\, g_{\phi\phi}=-a^2r'^2\sin^2(\theta) \tag{12.5.488}$$

For later we need to calculate the Christoffel symbols (arrrgh!!!!). So let us do that for the non-null one only.

First let us recall that (see section Tensor Calculus page 2109) that:

$$\begin{aligned}\Gamma^{\sigma}_{\lambda\mu}&=\frac{1}{2}g^{\nu\sigma}\left[\frac{\partial g_{\mu\nu}}{\partial x^{\lambda}}+\frac{\partial g_{\lambda\nu}}{\partial x^{\mu}}-\frac{\partial g_{\lambda\mu}}{\partial x^{\nu}}\right]\\ \Gamma^{\mu}_{\mu\lambda}&=\Gamma^{\mu}_{\lambda\mu}=\frac{1}{2}g^{\mu\rho}\frac{\partial g_{\rho\mu}}{\partial x^{\lambda}}\end{aligned} \tag{12.5.489}$$

And let's go for the boring part!:

$$\begin{aligned}\Gamma^{r'}_{r'r'}&=\frac{1}{2}g^{r'\rho}\frac{\partial g_{\rho r'}}{\partial x^{r'}}=\frac{1}{2}g^{r'r'}\frac{\partial g_{r'r'}}{\partial r'}=\frac{1}{2}\left(\frac{1-kr'^2}{\cancel{-a^2}}\right)\frac{\partial\left(\frac{\cancel{-a^2}}{1-kr'^2}\right)}{\partial r'}\\ &=\frac{1}{2}\left(1-kr'^2\right)\frac{+2kr'}{\left(1-kr'^2\right)^2}=\frac{kr'}{\left(1-kr'^2\right)}\end{aligned}$$

$$\begin{aligned}\Gamma^{r'}_{0r'}=\Gamma^{r'}_{r'0}&=\frac{1}{2}g^{r'\rho}\frac{\partial g_{\rho r'}}{\partial x^{0}}=\frac{1}{2}g^{r'r'}\frac{\partial g_{r'r'}}{\partial x^{0}}=\frac{1}{2}\frac{1-kr'^2}{-a^2}\frac{\partial\left(\frac{-a^2}{1-kr'^2}\right)}{c\partial t}\\ &=\frac{1}{2c}\frac{1-kr'^2}{-a^2}\frac{-2a\dot{a}}{1-kr'^2}=\frac{\dot{a}}{ca}\\ \Gamma^{\theta}_{0\theta}=\Gamma^{\theta}_{\theta 0}&=\frac{1}{2}g^{\theta\theta}\frac{\partial g_{\theta\theta}}{\partial x^{0}}=\frac{1}{2}\frac{1}{-a^2r'^2}\frac{\partial(-a^2r'^2)}{c\partial t}=\frac{\dot{a}}{ca}\\ \Gamma^{\phi}_{0\phi}=\Gamma^{\phi}_{\phi 0}&=\frac{\dot{a}}{ca}\end{aligned}$$

$$\begin{aligned}\Gamma^{\theta}_{\theta r'}=\Gamma^{\theta}_{r'\theta}&=\frac{1}{2}g^{\theta\theta}\frac{\partial g_{\theta\theta}}{\partial r'}=\frac{1}{2}\frac{1}{-a^2r'^2}\frac{\partial(-a^2r'^2)}{\partial r'}=\frac{1}{r'}\\ \Gamma^{\phi}_{\phi r'}=\Gamma^{\phi}_{r'\phi}&=\frac{1}{2}g^{\phi\phi}\frac{\partial g_{\phi\phi}}{\partial r'}=\frac{1}{2}\frac{1}{-a^2r'^2\sin^2(\theta)}\frac{\partial(-a^2r'^2\sin^2(\theta))}{\partial r'}=\frac{1}{r'}\end{aligned}$$

$$\begin{aligned}\Gamma^{0}_{r'r'}&=\frac{1}{2}g^{\nu 0}\left[\frac{\partial g_{r'\nu}}{\partial r'}+\frac{\partial g_{r'\nu}}{\partial r'}-\frac{\partial g_{r'r'}}{\partial x^{\nu}}\right]=\frac{1}{2}g^{00}\left[\frac{\partial g_{r'0}}{\partial r'}+\frac{\partial g_{r'0}}{\partial r'}-\frac{\partial g_{r'r'}}{\partial x^{0}}\right]\\ &=-\frac{1}{2}\left[\frac{\partial\left(\frac{-a^2}{1-kr'^2}\right)}{c\partial t}\right]=\frac{a\dot{a}}{c(1-kr'^2)}\end{aligned}$$

$$\Gamma_{\theta\theta}^{0}=\frac{1}{2}g^{00}\left[\frac{\partial g_{\theta 0}}{\partial x^{\theta}}+\frac{\partial g_{\theta 0}}{\partial x^{\theta}}-\frac{\partial g_{\theta\theta}}{\partial x^{0}}\right]=-\frac{1}{2}\left[\frac{\partial(-a^2r'^2)}{c\partial t}\right]=\frac{a\dot{a}r'^2}{c}$$

$$\Gamma_{\phi\phi}^{0}=\frac{1}{2}g^{00}\left[\frac{\partial g_{\phi 0}}{\partial x^{\phi}}+\frac{\partial g_{\phi 0}}{\partial x^{\phi}}-\frac{\partial g_{\phi\phi}}{\partial x^{0}}\right]=-\frac{1}{2}\left[\frac{\partial(-a^2r'^2\sin^2(\theta))}{c\partial t}\right]=\frac{a\dot{a}r'^2\sin^2(\theta)}{c}$$

$$\begin{aligned}\Gamma_{\phi\phi}^{\theta}&=\frac{1}{2}g^{\theta\theta}\left[\frac{\partial g_{\phi\theta}}{\partial x^{\phi}}+\frac{\partial g_{\phi\theta}}{\partial x^{\phi}}-\frac{\partial g_{\phi\phi}}{\partial x^{\theta}}\right]=-\frac{1}{2}\frac{1}{-a^2r'^2}\frac{\partial(-a^2r'^2\sin^2(\theta))}{\partial\theta}\\&=-\sin(\theta)\cos(\theta)\end{aligned}$$

$$\begin{aligned}\Gamma_{\theta\phi}^{\phi}&=\frac{1}{2}g^{\phi\phi}\left[\frac{\partial g_{\phi\phi}}{\partial x^{\theta}}+\frac{\partial g_{\theta\phi}}{\partial x^{\phi}}-\frac{\partial g_{\theta\phi}}{\partial x^{\phi}}\right]=\frac{1}{2}\frac{1}{-a^2r'^2\sin^2(\theta)}\frac{\partial(-a^2r'^2\sin^2(\theta))}{\partial\theta}\\&=\frac{\cos(\theta)}{\sin(\theta)}\end{aligned}$$

$$\begin{aligned}\Gamma_{\theta\theta}^{r'}&=\frac{1}{2}g^{r'r'}\left[\frac{\partial g_{\theta r'}}{\partial x^{\theta}}+\frac{\partial g_{\theta r'}}{\partial x^{\theta}}-\frac{\partial g_{\theta\theta}}{\partial x^{r'}}\right]=-\frac{1}{2}\frac{1-kr'^2}{\cancel{-a^2}}\frac{\partial(\cancel{-a^2}r'^2)}{\partial r'}\\&=-r'(1-kr'^2)\end{aligned}$$

$$\begin{aligned}\Gamma_{\phi\phi}^{r'}&=\frac{1}{2}g^{r'r'}\left[\frac{\partial g_{\phi r'}}{\partial x^{\phi}}+\frac{\partial g_{\phi r'}}{\partial x^{\phi}}-\frac{\partial g_{\phi\phi}}{\partial x^{r'}}\right]=-\frac{1}{2}\frac{(1-kr'^2)}{\cancel{-a^2}}\frac{\partial(\cancel{-a^2}r'^2\sin^2(\theta))}{\partial r'}\\&=-r'(1-kr'^2)\sin^2(\theta)\end{aligned}$$

Let us now calculate the Ricci tensor only for the non-null terms:

$$R_{\mu k}=\frac{\partial\Gamma_{\mu k}^{\nu}}{\partial x^{\nu}}-\frac{\partial\Gamma_{\mu\nu}^{\nu}}{\partial x^{k}}+\Gamma_{\nu\sigma}^{\sigma}\Gamma_{\mu k}^{\nu}-\Gamma_{\mu\nu}^{\sigma}\Gamma_{k\sigma}^{\nu} \tag{12.5.490}$$

Then (pffff...!):

$$\begin{aligned}-R_{00}&=-\frac{\partial\overbrace{\Gamma_{00}^{\nu}}^{0}}{\partial x^{\nu}}+\frac{\partial\Gamma_{0\nu}^{\nu}}{\partial x^{0}}-\Gamma_{\nu\sigma}^{\sigma}\overbrace{\Gamma_{\mu k}^{\nu}}^{0}+\Gamma_{\nu 0}^{\sigma}\Gamma_{0\sigma}^{\nu}\\
&=\frac{\partial(\Gamma_{0r'}^{r'}+\Gamma_{0\theta}^{\theta}+\Gamma_{0\phi}^{\phi})}{c\partial t}+(\overbrace{\Gamma_{00}^{\sigma}}^{0}\Gamma_{0\sigma}^{0}+\Gamma_{r'0}^{\sigma}\Gamma_{0\sigma}^{r'}+\Gamma_{\theta 0}^{\sigma}\Gamma_{0\sigma}^{\theta}+\Gamma_{\phi 0}^{\sigma}\Gamma_{0\sigma}^{\phi})\\
&=\frac{\partial(3\cdot\frac{\dot{a}}{c\cdot a})}{c\partial t}+(\overbrace{\Gamma_{r'0}^{0}}^{0}\overbrace{\Gamma_{00}^{r'}}^{0}+\overbrace{\Gamma_{\theta 0}^{0}}^{0}\overbrace{\Gamma_{00}^{\theta}}^{0}+\overbrace{\Gamma_{\phi 0}^{0}}^{0}\overbrace{\Gamma_{00}^{\phi}}^{0})+(\overbrace{\Gamma_{r'0}^{r'}\Gamma_{0r'}^{r'}}^{(\frac{\dot{a}}{c\cdot a})^2}+\Gamma_{\theta 0}^{r'}\overbrace{\Gamma_{0r'}^{\theta}}^{0}+\overbrace{\Gamma_{\phi 0}^{r'}}^{0}\Gamma_{0r'}^{\phi})\\
&+(\Gamma_{r'0}^{\theta}\overbrace{\Gamma_{0\theta}^{r'}}^{0}+\overbrace{\Gamma_{\theta 0}^{\theta}\Gamma_{0\theta}^{\theta}}^{(\frac{\dot{a}}{c\cdot a})^2}+\overbrace{\Gamma_{\phi 0}^{\theta}}^{0}\Gamma_{0\theta}^{\phi})+(\overbrace{\Gamma_{r'0}^{\phi}}^{0}\overbrace{\Gamma_{0\phi}^{r'}}^{0}+\overbrace{\Gamma_{r'0}^{\phi}}^{0}\Gamma_{0\phi}^{\theta}+\overbrace{\Gamma_{\phi 0}^{\phi}\Gamma_{0\phi}^{\phi}}^{(\frac{\dot{a}}{c\cdot a})^2})\\
&=\frac{3}{c^2}\left(\frac{a\ddot{a}-\dot{a}^2}{a^2}\right)+3\frac{\dot{a}^2}{c^2a^2}=\frac{3\ddot{a}}{ac^2}\end{aligned} \tag{12.5.491}$$

So finally:

$$R_{00} = -\frac{3\ddot{a}}{ac^2} \tag{12.5.492}$$

And next:

$$\begin{aligned}
-R_{\phi\phi} &= -\frac{\partial\Gamma^{\nu}_{\phi\phi}}{\partial x^{\nu}} + \frac{\overbrace{\partial\Gamma^{\nu}_{\phi\nu}}^{0}}{\partial x^{\phi}} - \Gamma^{\sigma}_{\nu\sigma}\Gamma^{\nu}_{\phi\phi} + \Gamma^{\sigma}_{\phi\nu}\Gamma^{\nu}_{\phi\sigma} = \left(-\frac{\partial\Gamma^{0}_{\phi\phi}}{\partial x^{0}} - \Gamma^{\sigma}_{0\sigma}\Gamma^{0}_{\phi\phi} + \Gamma^{\sigma}_{\phi 0}\Gamma^{0}_{\phi\sigma}\right)_{0}\\
&+ \left(-\frac{\partial\Gamma^{r'}_{\phi\phi}}{\partial r'} - \Gamma^{\sigma}_{r'\sigma}\Gamma^{r'}_{\phi\phi} + \Gamma^{\sigma}_{\phi r'}\Gamma^{r'}_{\phi\sigma}\right)_{r'} + \left(-\frac{\partial\Gamma^{\theta}_{\phi\phi}}{\partial\theta} - \Gamma^{\sigma}_{\theta\sigma}\Gamma^{\theta}_{\phi\phi} + \Gamma^{\sigma}_{\phi\theta}\Gamma^{\theta}_{\phi\sigma}\right)_{\theta} + \Gamma^{\sigma}_{\phi\phi}\Gamma^{\phi}_{\phi\sigma}\\
&= \Bigg[-\frac{\left(\frac{a\dot{a}\cdot r'^2\cdot\sin^2(\theta)}{c}\right)}{c\partial t} - 3\frac{\dot{a}}{c\cdot\cancel{a}}\cdot\frac{\cancel{a}\dot{a}\cdot r'^2\cdot\sin^2(\theta)}{c} + \overbrace{\Gamma^{0}_{\phi 0}\Gamma^{0}_{\phi 0}}^{0} + \overbrace{\Gamma^{r'}_{\phi 0}\Gamma^{0}_{\phi r'}}^{0} + \overbrace{\Gamma^{\theta}_{\phi 0}\Gamma^{0}_{\phi\theta}}^{0} + \overbrace{\Gamma^{\phi}_{\phi 0}}^{\frac{\dot{a}}{c\cdot a}}\;\overbrace{\Gamma^{0}_{\phi\phi}}^{\frac{a\dot{a}\cdot r'^2\cdot\sin^2(\theta)}{c}}\Bigg]\\
&+ \Bigg[-\frac{\partial(-r'\cdot(1-k\cdot r'^2)\cdot\sin^2(\theta))}{\partial r'} - \overbrace{\Gamma^{0}_{r'0}}^{0}\Gamma^{r'}_{\phi\phi} - \overbrace{\Gamma^{r'}_{r'r'}}^{\frac{k\cdot r'}{(1-k\cdot r'^2)}}\;\overbrace{\Gamma^{r'}_{\phi\phi}}^{-r'\cdot(1-k\cdot r'^2)\cdot\sin^2(\theta)} - \cancel{\overbrace{\Gamma^{\theta}_{r'\theta}}^{\frac{1}{r'}}\;\overbrace{\Gamma^{r'}_{\phi\phi}}^{-r'\cdot(1-k\cdot r'^2)\cdot\sin^2(\theta)}}\\
&- \cancel{\overbrace{\Gamma^{\phi}_{r'\phi}}^{\frac{1}{r'}}\;\overbrace{\Gamma^{r'}_{\phi\phi}}^{-r'\cdot(1-k\cdot r'^2)\cdot\sin^2(\theta)}} + \overbrace{\Gamma^{0}_{\phi r'}}^{0}\Gamma^{r'}_{\phi 0} + \overbrace{\Gamma^{r'}_{\phi r'}}^{0}\Gamma^{r'}_{\phi r'} + \overbrace{\Gamma^{\theta}_{\phi r'}}^{0}\Gamma^{r'}_{\phi\theta} + \cancel{\overbrace{\Gamma^{\phi}_{\phi r'}}^{\frac{1}{r'}}\;\overbrace{\Gamma^{r'}_{\phi\phi}}^{-r'\cdot(1-k\cdot r'^2)\cdot\sin^2(\theta)}}\Bigg]\\
&+ \Bigg[-\frac{\partial(-\sin^2(\theta)\cdot\cos(\theta))}{\partial\theta} - \overbrace{\Gamma^{0}_{\theta 0}}^{0}\Gamma^{\theta}_{\phi\phi} - \overbrace{\Gamma^{r'}_{\theta r'}}^{0}\Gamma^{\theta}_{\phi\phi} - \overbrace{\Gamma^{\theta}_{\theta\theta}}^{0}\Gamma^{\theta}_{\phi\phi} - \cancel{\overbrace{\Gamma^{\phi}_{\theta\phi}}^{\frac{\cos(\theta)}{\sin(\theta)}}\;\overbrace{\Gamma^{\theta}_{\phi\phi}}^{-\sin(\theta)\cdot\cos(\theta)}} + \overbrace{\Gamma^{0}_{\phi\theta}}^{0}\Gamma^{\theta}_{\phi 0}\\
&+ \overbrace{\Gamma^{r'}_{\phi\theta}}^{0}\Gamma^{\theta}_{\phi r'} + \overbrace{\Gamma^{\theta}_{\phi\theta}}^{0}\Gamma^{\theta}_{\phi\theta} + \cancel{\overbrace{\Gamma^{\phi}_{\phi\theta}}^{\frac{\cos(\theta)}{\sin(\theta)}}\;\overbrace{\Gamma^{\theta}_{\phi\phi}}^{-\sin(\theta)\cdot\cos(\theta)}}\Bigg]\\
&+ \overbrace{\Gamma^{0}_{\phi\phi}}^{\frac{a\dot{a}\cdot r'^2\cdot\sin^2(\theta)}{c}}\;\overbrace{\Gamma^{\phi}_{\phi 0}}^{\frac{\dot{a}}{c\cdot a}} + \cancel{\overbrace{\Gamma^{r'}_{\phi\phi}}^{-r'\cdot(1-k\cdot r'^2)\cdot\sin^2(\theta)}\;\overbrace{\Gamma^{\phi}_{\phi r'}}^{\frac{1}{r'}}} + \overbrace{\Gamma^{\theta}_{\phi\phi}}^{-\sin(\theta)\cdot\cos(\theta)}\;\overbrace{\Gamma^{\phi}_{\phi\theta}}^{\frac{\cos(\theta)}{\sin(\theta)}} + \overbrace{\Gamma^{\phi}_{\phi\phi}}^{0}\Gamma^{\phi}_{\phi\phi}
\end{aligned} \tag{12.5.493}$$

Therefore:

$$\begin{aligned}
R_{\phi\phi} &= -\frac{r'^2\sin^2(\theta)(\dot{a}^2+a\ddot{a})}{c^2} - 2\frac{\dot{a}^2\cdot r'^2\sin^2(\theta)}{c^2} + \cancel{\sin^2(\theta)} - 3kr'^2\sin^2(\theta) + kr'^2\sin^2(\theta)\\
&+ \cancel{\cos^2(\theta)} - \cancel{\sin^2(\theta)} + \frac{\dot{a}^2\cdot r'^2\cdot\sin^2(\theta)}{c^2} - \cancel{\cos^2(\theta)}
\end{aligned} \tag{12.5.494}$$

So finally:

$$R_{\phi\phi} = \frac{r'^2\sin^2(\theta)}{c^2}\left[a\ddot{a} + 2\dot{a} + 2kc^2\right] \tag{12.5.495}$$

$$
\begin{aligned}
-R_{r'r'} &= -\frac{\partial\Gamma^{\nu}_{r'r'}}{\partial x^{\nu}} + \overbrace{\frac{\partial\Gamma^{\nu}_{r'\nu}}{\partial x^{r'}}}^{0} - \Gamma^{\sigma}_{\nu\sigma}\Gamma^{\nu}_{r'r'} + \Gamma^{\sigma}_{r'\nu}\Gamma^{\nu}_{r'\sigma}\\
&= (-\frac{\partial\Gamma^{0}_{r'r'}}{\partial x^{0}} + \frac{\partial\,\overbrace{\Gamma^{0}_{r'0}}^{0}}{\partial x^{r'}} - \Gamma^{\sigma}_{0\sigma}\Gamma^{0}_{r'r'} + \Gamma^{\sigma}_{r'0}\Gamma^{0}_{r'0})_{0} + (-\cancel{\frac{\partial\Gamma^{r}_{r'r'}}{\partial x^{r'}}} + \cancel{\frac{\partial\Gamma^{r}_{r'r'}}{\partial x^{r'}}} - \Gamma^{\sigma}_{r'\sigma}\Gamma^{r'}_{r'r'} + \Gamma^{\sigma}_{r'r'}\Gamma^{r'}_{r'\sigma})_{r'}\\
&+ (-\frac{\partial\,\overbrace{\Gamma^{\theta}_{r'r'}}^{0}}{\partial x^{\theta}} + \frac{\partial\Gamma^{\theta}_{r'\theta}}{\partial x^{r'}} - \Gamma^{\sigma}_{\theta\sigma}\,\overbrace{\Gamma^{\theta}_{r'r'}}^{0} + \Gamma^{\sigma}_{r'\theta}\Gamma^{\theta}_{r'\sigma})_{\theta} + (-\cancel{\frac{\partial\Gamma^{\phi}_{r'r'}}{\partial x^{\phi}}} + \frac{\partial\Gamma^{\phi}_{r'\phi}}{\partial x'^{r}} - \Gamma^{\sigma}_{\phi\sigma}\,\overbrace{\Gamma^{\phi}_{r'r'}}^{0} + \Gamma^{\sigma}_{r'\phi}\Gamma^{\phi}_{r'\sigma})_{\phi}\\
&= (-\frac{\partial(\frac{a\dot{a}}{c(1-kr'^2)})}{c\partial t} - (\overbrace{\Gamma^{0}_{00}}^{0} + \overbrace{\Gamma^{r'}_{0r'}}^{\frac{\dot{a}}{ca}} + \overbrace{\Gamma^{\theta}_{0\theta}}^{\frac{\dot{a}}{ca}} + \overbrace{\Gamma^{\phi}_{0\phi}}^{\frac{\dot{a}}{ca}})\,\overbrace{\Gamma^{0}_{r'r'}}^{\frac{a\dot{a}}{c(1-kr'^2)}}\\
&+ \overbrace{\Gamma^{0}_{r'0}\Gamma^{0}_{r'0}}^{0} + \overbrace{\Gamma^{r'}_{r'0}}^{\frac{\dot{a}}{ca}}\,\overbrace{\Gamma^{0}_{r'r'}}^{\frac{a\dot{a}}{c(1-kr'^2)}} + \overbrace{\Gamma^{\theta}_{r'0}\Gamma^{0}_{r'\theta}}^{0} + \overbrace{\Gamma^{\phi}_{r'0}\Gamma^{0}_{r'\phi}}^{0}\\
&+ (-(\overbrace{\Gamma^{0}_{r'0}}^{0} + \overbrace{\Gamma^{r'}_{r'r'}}^{\frac{kr'}{(1-kr'^2)}} + \overbrace{\Gamma^{\theta}_{r'\theta}}^{\frac{1}{r'}} + \overbrace{\Gamma^{\phi}_{r'\phi}}^{\frac{1}{r'}})\,\overbrace{\Gamma^{r'}_{r'r'}}^{\frac{kr'}{(1-kr'^2)}} + \overbrace{\Gamma^{0}_{r'r'}}^{\frac{a\dot{a}}{c(1-kr'^2)}}\\
&+ \overbrace{\Gamma^{r'}_{r'0}}^{\frac{\dot{a}}{ca}} + \overbrace{\Gamma^{r'}_{r'r'}}^{\frac{kr'}{(1-kr'^2)}}\,\overbrace{\Gamma^{r'}_{r'r'}}^{\frac{kr'}{(1-kr'^2)}} + \overbrace{\Gamma^{\theta}_{r'r'}}^{0}\,\Gamma^{r'}_{r'\theta} + \overbrace{\Gamma^{\phi}_{r'r'}}^{0}\,\Gamma^{r'}_{r'\phi})_{r'}\\
&+ (\frac{\partial(\frac{1}{r'})}{\partial r'} + \overbrace{\Gamma^{0}_{r'\theta}\Gamma^{\theta}_{r'0}}^{0} + \overbrace{\Gamma^{r'}_{r'\theta}}^{0}\,\Gamma^{\theta}_{r'r'} + \overbrace{\Gamma^{\theta}_{r'\theta}}^{\frac{1}{r'}}\,\overbrace{\Gamma^{\theta}_{r'\theta}}^{\frac{1}{r'}} + \overbrace{\Gamma^{\phi}_{r'\theta}\Gamma^{\theta}_{r'\phi}}^{0})_{\theta}\\
&+ (\frac{\partial(\frac{1}{r'})}{\partial r'} + \overbrace{\Gamma^{0}_{r'\phi}\Gamma^{\phi}_{r'0}}^{0} + \overbrace{\Gamma^{r'}_{r'\phi}\Gamma^{\phi}_{r'r'}}^{0} + \overbrace{\Gamma^{\theta}_{r'\phi}\Gamma^{\phi}_{r'\theta}}^{0} + \overbrace{\Gamma^{\phi}_{r'\phi}}^{\frac{1}{r'}}\,\overbrace{\Gamma^{\phi}_{r'\phi}}^{\frac{1}{r'}})_{\phi}\\
&= \left[-\frac{\dot{a}^2+a\ddot{a}}{c^2(1-kr'^2)} - 2\frac{\dot{a}}{c\cancel{a}}\frac{\cancel{a}\dot{a}}{c(1-kr'^2)}\right]\\
&+ \left[-\cancel{(\frac{kr'}{(1-kr'^2)})^2} - 2\frac{1}{\cancel{r'}}\frac{k\cancel{r'}}{(1-kr'^2)} + \frac{\cancel{a}\dot{a}}{c(1-kr'^2)}\frac{\dot{a}}{c\cancel{a}} + \cancel{(\frac{kr'}{(1-kr'^2)})^2}\right]\\
&+ \left[-\cancel{\frac{1}{r'^2}} + \cancel{\frac{1}{r'^2}}\right] + \left[-\cancel{\frac{1}{r'^2}} + \cancel{\frac{1}{r'^2}}\right]\\
&= -\frac{\dot{a}^2+a\ddot{a}}{c^2(1-kr'^2)} - \frac{\dot{a}^2}{c^2(1-kr'^2)} - \frac{2kc^2}{c^2(1-kr'^2)}
\end{aligned}
\tag{12.5.496}
$$

Therefore:

$$R_{r'r'} = \frac{a\ddot{a} + 2\dot{a}^2 + 2kc^2}{c^2(1-kr'^2)} \tag{12.5.497}$$

Also identically (sorry but we don't have the power anymore to detail the calculations...):

$$R_{\theta\theta} = \frac{r'^2}{c^2}\left[a\ddot{a} + 2\dot{a}^2 + 2kc^2\right] \tag{12.5.498}$$

Now let us calculate the scalar curvature (Ricci scalar). Let us recall (see section Tensor Calculus page 2128), that it is given by:

$$R = g^{ij}R_{ij} = R^{i}_{i} \tag{12.5.499}$$

Therefore using previous results:

$$\begin{aligned}R&=g^{00}R_{00}+g^{r'r'}R_{r'r'}+g^{\theta\theta}R_{\theta\theta}+g^{\phi\phi}R_{\phi\phi}\\&=1\cdot\frac{-3\ddot{a}}{ac^2}+\frac{\cancel{1-kr'^2}}{-a^2}\frac{(a\ddot{a}+2\dot{a}^2+2kc^2)}{c^2\cancel{(1-kr'^2)}}\\&+\frac{1}{-a^2r'^2}\frac{\cancel{r'^2}}{c^2}[a\ddot{a}+2\dot{a}+2kc^2]+\frac{1}{-a^2\cancel{r'^2\sin^2(\theta)}}\frac{\cancel{r'^2\sin^2(\theta)}}{c^2}[a\ddot{a}+2\dot{a}^2+2kc^2]\\&=\frac{-3\ddot{a}a-(a\ddot{a}+2\dot{a}^2+2kc^2)-(a\ddot{a}+2\dot{a}^2+2kc^2)-(a\ddot{a}+2\dot{a}^2+2kc^2)}{a^2c^2}\end{aligned}\tag{12.5.500}$$

So finally:

$$R=\frac{-6(a\ddot{a}+\dot{a}^2+kc^2)}{a^2c^2}\tag{12.5.501}$$

Remark

The latter equality relates spatial curvature (usually denoted k) and space-time curvature (usually denoted R). Space-time curvature R has a contribution $6k/a^2$ from spatial curvature k, but it also has two terms that come just from the spatial expansion over time, a.

Remember that space-time curvature is coordinate-independent. Spatial curvature depends on the "slicing" (into constant-time slices) imposed by coordinates; for example de Sitter is spatially flat in one slicing and spatially curved in another, but always has uniform positive space-time curvature.

Now let us recall that we have proved:

$$T^{ij}\cong\left(\rho+\frac{P}{c^2}\right)u^iu^j-Pg^{ij}\tag{12.5.502}$$

And therefore:

$$T_{ij}\cong\left(\rho+\frac{P}{c^2}\right)u_iu_j-Pg_{ij}\tag{12.5.503}$$

So don't forget that only diagonal elements are non-null and that $u_i=(c,0,0,0)$.

Therefore:

$$T_{ij}=\begin{bmatrix}\rho_0c^2&0&0&0\\0&-Pg_{r'r'}&0&0\\0&0&-Pg_{\theta\theta}&0\\0&0&0&-Pg_{\phi\phi}\end{bmatrix}=\begin{bmatrix}\rho_E&0&0&0\\0&P\dfrac{a^2}{1-kr'^2}&0&0\\0&0&Pa^2r'^2&0\\0&0&0&Pa^2r'^2\sin^2(\theta)\end{bmatrix}\tag{12.5.504}$$

Now let us recall the Einstein Field Equations (see section General Relativity page 5766):

$$R_{ij}-\frac{1}{2}g_{ij}R-\Lambda g_{ij}=\frac{8\pi G}{c^4}T_{ij}\tag{12.5.505}$$

Let us start with:

$$R_{00}-\frac{1}{2}g_{00}R-\Lambda g_{00}=\frac{8\pi G}{c^4}T_{00} \tag{12.5.506}$$

So we get after injecting the previous results:

$$-\frac{3\ddot{a}}{ac^2}-\frac{1}{2}\cdot 1\cdot\frac{-6(a\ddot{a}+\dot{a}^2+kc^2)}{a^2c^2}-\Lambda=\frac{8\pi G}{c^4}\rho_E \tag{12.5.507}$$

After elementary simplification we get immediately:

$$3\frac{\dot{a}^2}{a^2}+3\frac{kc^2}{a^2}-\Lambda c^2=\frac{8\pi G}{c^2}\rho \tag{12.5.508}$$

And now with:

$$R_{r'r'}-\frac{1}{2}g_{r'r'}R-\Lambda g_{r'r'}=\frac{8\pi G}{c^4}T_{r'r'} \tag{12.5.509}$$

So we get after injecting the previous results:

$$\frac{a\ddot{a}+2\dot{a}^2+2kc^2}{c^2\cancel{(1-kr'^2)}}-\frac{1}{2}\left(\frac{-\cancel{a^2}}{\cancel{1-kr'^2}}\right)\frac{-6(a\ddot{a}+\dot{a}^2+kc^2)}{\cancel{a^2}c^2}-\Lambda\left(\frac{-a^2}{\cancel{1-kr'^2}}\right)=\frac{8\pi G}{c^4}P\frac{a^2}{\cancel{1-kr'^2}} \tag{12.5.510}$$

After elementary simplification we get immediately:

$$-2\frac{\ddot{a}}{a}-\frac{\dot{a}^2}{a^2}-\frac{kc^2}{a^2}+\Lambda c^2=\frac{8\pi G}{c^2}P \tag{12.5.511}$$

And now with:

$$R_{\phi\phi}-\frac{1}{2}g_{\phi\phi}R-\Lambda g_{\phi\phi}=\frac{8\pi G}{c^4}T_{\phi\phi} \tag{12.5.512}$$

So we get after injecting the previous results:

$$\frac{r'^2\sin^2(\theta)}{c^2}[a\ddot{a}+2\dot{a}^2+2kc^2]-\frac{1}{2}\left(-a^2r'^2\sin^2(\theta)\right)\frac{-6(a\ddot{a}+\dot{a}^2+kc^2)}{a^2c^2}-\Lambda\left(-a^2r'^2\sin^2(\theta)\right)=\frac{8\pi G}{c^4}Pa^2r'^2\sin^2(\theta) \tag{12.5.513}$$

After elementary simplification we get immediately (same result as before! this is a consequence of space isotropy assumption):

$$-2\frac{\ddot{a}}{a}-\frac{\dot{a}^2}{a^2}-\frac{kc^2}{a^2}+\Lambda c^2=\frac{8\pi G}{c^2}P \tag{12.5.514}$$

And finally doing the same with:

$$R_{\theta\theta}-\frac{1}{2}g_{\theta\theta}R-\Lambda g_{\theta\theta}=\frac{8\pi G}{c^4}T_{\theta\theta} \tag{12.5.515}$$

we get also exactly the same result as the two previous one!

The both relations:

$$\boxed{\begin{aligned} 3\frac{\dot{a}^2}{a^2}+3\frac{kc^2}{a^2}-\Lambda c^2&=\frac{8\pi G}{c^2}\rho\\ -2\frac{\ddot{a}}{a}-\frac{\dot{a}^2}{a^2}-\frac{kc^2}{a^2}+\Lambda c^2&=\frac{8\pi G}{c^2}P \end{aligned}} \tag{12.5.516}$$

are the "general Friedmann equations".

It is very common to find them in the literature into the form of the velocity and acceleration equations (often write the cosmological constant in terms of a vacuum energy density as $\Lambda=8\pi G\rho_{\text{vac}}/c^4$):

$$\begin{aligned} H^2&=\left(\frac{\dot{a}}{a}\right)^2=\frac{8\pi G}{3c^2}\rho-\frac{kc^2}{a^2}+\frac{\Lambda c^2}{3}=\frac{8\pi G}{3c^2}(\rho+\rho_{\text{vac}})-\frac{kc^2}{a^2}\\ \frac{\ddot{a}}{a}&=\Lambda c^2-\frac{\dot{a}^2}{a^2}-\frac{kc^2}{a^2}-\frac{8\pi G}{c^2}P=\frac{8\pi G}{c^2}\rho_{\text{vac}}-\frac{\dot{a}^2}{a^2}-\frac{kc^2}{a^2}-\frac{8\pi G}{c^2}P \end{aligned} \tag{12.5.517}$$

Remark

Notice that if we assume that ρ and P depends on a (which seems quite logical if the energy in the Universe is indeed conservative), typically the decrease with the expansion (intuitive!), then is seems that the only term that does not contain the scale factor a is Λc^2. Hence the fact that some scientists assume that Λ behave like constant whatever the age of the Universe.

Now let us rewrite the first one as following:

$$\frac{\dot{a}^2}{a^2}+\frac{kc^2}{a^2}-\frac{\Lambda c^2}{3}=\frac{8\pi G}{3c^2}\rho \tag{12.5.518}$$

and let us notice that for $k=0$ and $\Lambda=0$ we get for that latter:

$$\frac{\dot{a}^2}{a^2}=\frac{8\pi G}{3c^2}\rho \tag{12.5.519}$$

and then we define the critical density that we already know:

$$\rho_c=\frac{3c^2}{8\pi G}\frac{\dot{a}^2}{a^2} \tag{12.5.520}$$

And let us recall also $\frac{\dot{a}}{a}=H$. Therefore:

$$\rho_c=\frac{3c^2H^2}{8\pi G} \tag{12.5.521}$$

Let us now define the following parameters of energy density (matter energy density, curvature energy density and radiation energy density):

$$\boxed{\Omega_m=\frac{\rho}{\rho_c}=\frac{8\pi G}{3c^2H^2}\rho \qquad \Omega_k=-\frac{kc^2}{a^2H^2} \qquad \Omega_\nu=\frac{\rho_v}{\rho_c}=\frac{\dfrac{\Lambda c^4}{8\pi G}}{\dfrac{3c^2H^2}{8\pi G}}=\frac{\Lambda c^2}{3H^2}} \tag{12.5.522}$$

Therefore our equality:

$$\frac{\dot{a}^2}{a^2}+\frac{kc^2}{a^2}-\frac{\Lambda c^2}{3}=\frac{8\pi G}{3c^2}\rho \tag{12.5.523}$$

can be rewritten:

$$H^2+\frac{kc^2}{a^2}-\frac{\Lambda c^2}{3}=\frac{8\pi G}{3c^2}\rho \tag{12.5.524}$$

Dividing both sides by H^2 we get:

$$1+\frac{kc^2}{a^2H^2}-\frac{\Lambda c^2}{3H^2}=\frac{8\pi G}{3c^2H^2}\rho \tag{12.5.525}$$

Rearranging:

$$-\frac{kc^2}{a^2H^2}+\frac{\Lambda c^2}{3H^2}+\frac{8\pi G}{3c^2H^2}\rho=1 \tag{12.5.526}$$

and injecting the density parameters we get:

$$\boxed{\Omega_m+\Omega_k+\Omega_\nu=1} \tag{12.5.527}$$

The figure below from the DESI (Dark Energy Spectroscopic Instrument) collaboration's second data release's results paper (see), shows the different values of the matter density that are preferred by six different data sets: DESI's first and second releases, the CMB, and the supernova samples of Pantheon+, Union, and DESY5 (note that BAO and supernova data sets are not really compatible with one another):

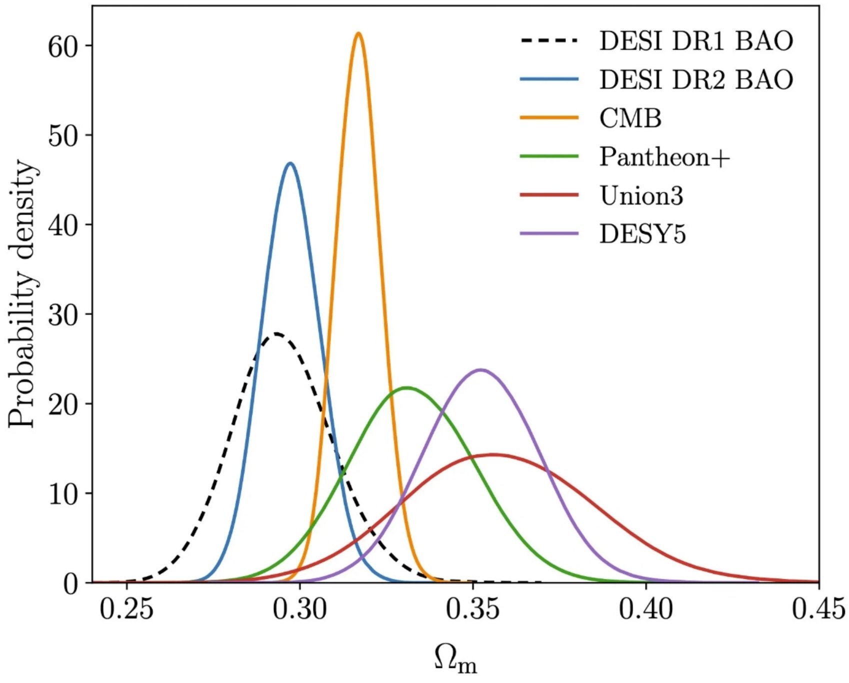


Figure 12.295: **Matter density Ω_m statistical distribution of different experiments**

Notice that is also quite common to see in textbooks:

$$\rho=\rho_m+\rho_\nu+\rho_k+\rho_\Lambda \tag{12.5.528}$$

where obviously we have the matter density, the radiation density, the curvature density and the constant energy density (there is a one to one correspondence with energy density given above but only with a c^2 factor difference as we know that $E = mc^2$).

Ok now that we have the basic tools. Let's use them by starting with the most famous wrong model:

12.5.6.6 Einstein (static) Universe model

Like most scientists of his time, Albert Einstein considered that the Universe was globally static and homogeneous (11917 according to holocene calendar). He also imagined it closed mainly for reasons related to Mach's principle, to which he believes very much. He thus built a model of a spherical Universe capable of accounting for the essential physical conditions he thinks it must respect.

However, he quickly found that his equations do not admit static solutions and so it is to satisfy what he considers as an experimental requirement (immobile universe) that he is forced to slightly modify his equations by adding as we already know an additional term: the cosmological constant.

This constant behaves like a repulsive force to exactly counterbalance the gravitational "force" due to matter, the whole being now able to be balance.

So if we take a closed universe ($k = +1$), static[68] ($a = c^{te}$) and made of a gas of dust of negligible pressure ($P \cong 0$), the general Friedmann equations, given for recall by:

$$\begin{aligned} 3\frac{\dot{a}^2}{a^2} + 3\frac{kc^2}{a^2} - \Lambda c^2 &= \frac{8\pi G}{c^2}\rho \\ -2\frac{\ddot{a}}{a} - \frac{\dot{a}^2}{a^2} - \frac{kc^2}{a^2} + \Lambda c^2 &= \frac{8\pi G}{c^2}P \end{aligned} \tag{12.5.529}$$

reduce immediately to:

$$\begin{aligned} 3\frac{c^2}{a^2} - \Lambda c^2 &= \frac{8\pi G}{c^2}\rho \\ -\frac{c^2}{a^2} + \Lambda c^2 &= 0 \end{aligned} \tag{12.5.530}$$

Therefore injecting the second relation into the first one, and rearranging the second one, we get:

$$\begin{aligned} 2\frac{c^2}{a^2} &= \frac{8\pi G}{c^2}\rho \\ \Lambda &= \frac{1}{a^2} \end{aligned} \tag{12.5.531}$$

So if $k = +1$ and $a = c^{te}$ with $P \cong 0$ we have that also $\Lambda = c^{te}$!

---

[68] For a stationary Universe the scaling factor $a(t)$ must be constant (typically equal to 1!) and hence all its time derivatives must vanish.

More explicitly:

$$a_{\text{Einstein}}=\sqrt{\frac{c^4}{4\pi G\rho}}=c^{te} \qquad \text{and} \qquad \Lambda_{\text{Einstein}}=\frac{1}{a_{\text{Einstein}}^2}=\frac{4\pi G\rho}{c^4}=c^{te} \tag{12.5.532}$$

So for the universe to be really static we should have $a=1$ and therefore $\Lambda=1$. Sadly (for Albert Einstein...), actual experimental measurements give us for $\rho_m\cong 10^{-26}\,[\text{kg}\cdot\text{m}^{-3}]$, a value of the scale factor equal to $a_{\text{Einstein}}=1.03\cdot 10^{26}\,[\text{m}]$ that is to say approximately $1\cdot 10^9$ light years, and hence $\Lambda_{\text{Einstein}}=9.31\cdot 10^{-53}\,[\text{m}^{-2}]$.

Let us see now another special case found by Edward Arthur Milne in 11932 (holocene calendar).

12.5.6.7 Milne linear Universe model

The Milne model demands that the energy density, pressure and cosmological constant all equal zero. This is of course not consistent with the fact that the universe contains stars and galaxies, but we might wonder whether it could tell us anything interesting as a simplified approximation to a very dilute universe.

Hence we get from the first Friedmann equation:

$$3\frac{\dot{a}^2}{a^2}+3\frac{kc^2}{a^2}-\Lambda c^2=\frac{8\pi G}{c^2}\rho \tag{12.5.533}$$

with $\Lambda=0$ and $\rho=0$:

$$3\frac{\dot{a}^2}{a^2}+3\frac{kc^2}{a^2}=0 \tag{12.5.534}$$

After simplification:

$$\dot{a}^2=-kc^2 \tag{12.5.535}$$

Hence:

$$\dot{a}=\sqrt{-kc^2} \tag{12.5.536}$$

To avoid a negative square root (that we will assume maybe as unphysical), we choose a negative curvature $k<0$. Therefore after integration of:

$$\mathrm{d}a=\sqrt{-kc^2}\mathrm{d}t \tag{12.5.537}$$

we get:

$$a=\sqrt{-kc^2}t \tag{12.5.538}$$

Then there exists a time at which $a=0$, but this doesn't turn out to be a real singularity (which isn't surprising, since there is no matter to generate gravitational fields). Let this universe have a scattering of test particles whose masses are too small to invalidate the approximation of $\rho=0$, and let the test particles be at rest. The linear dependence of a on t means that these particles simply move inertially and without any gravitational interactions, spreading apart from one another at a constant rate like the raisins in a rising loaf of raisin bread.

The Milne universe may be useful as an inoculation against the common misconception that the Big Bang was an explosion of matter spreading out into a pre-existing vacuum.

12.5.6.8 Friedmann Universe models

Now to continue with the Friedmann Universe models (11922-11924 according to holocene calendar) we need first an important relation. Let us recall the general Friedmann equations we get just earlier:

$$\begin{aligned} 3\frac{\dot{a}^2}{a^2}+3\frac{kc^2}{a^2}-\Lambda c^2=\frac{8\pi G}{c^2}\rho \\ -2\frac{\ddot{a}}{a}-\frac{\dot{a}^2}{a^2}-\frac{kc^2}{a^2}+\Lambda c^2=\frac{8\pi G}{c^2}P \end{aligned} \tag{12.5.539}$$

Multiplying the second one by 3 and subtracting the second one we get immediately:

$$-6\frac{\ddot{a}}{a}+2\Lambda c^2=\frac{8\pi G}{c^2}(3P+\rho) \tag{12.5.540}$$

Therefore:

$$\boxed{\frac{\ddot{a}}{a}=-\frac{4\pi G}{3c^2}(3P+\rho)+\frac{\Lambda c^2}{3}} \tag{12.5.541}$$

we fall back here on a generalization of one of the Friedmann equation[69] we used during our study of the Newtonian model of the Universe!

That latter relation is also commonly written in many textbooks as:

$$\ddot{R}=-\frac{4\pi g}{3}R\left(\rho_0+\frac{3P}{c^2}\right)+\frac{c^2\Lambda}{3}R \tag{12.5.542}$$

or:

$$\frac{\ddot{R}}{R}=-\frac{4\pi g}{3}\left(\rho_0+\frac{3P}{c^2}\right)+\frac{c^2\Lambda}{3} \tag{12.5.543}$$

To summarize a little bit, from the Einstein Field equation without cosmological constant, we have:

$$\begin{gathered} R_{\mu\nu}-\frac{1}{2}g_{\mu\nu}R=\frac{8\pi G}{c^4}T_{\mu\nu} \\ \Downarrow \\ \left(\frac{\dot{a}}{a}\right)^2=\frac{8\pi G\rho}{3c^2}-\frac{kc^2}{a^2} \\ \frac{\ddot{a}}{a}=-\frac{4\pi G}{3c^2}(\rho+3P) \end{gathered} \tag{12.5.544}$$

and obviously for the Einstein Field equation without cosmological constant, we have:

$$\begin{gathered} R_{\mu\nu}-\frac{1}{2}g_{\mu\nu}R+\Lambda g_{\mu\nu}=\frac{8\pi \mathcal{G}}{c^4}T_{\mu\nu} \\ \Downarrow \\ \left(\frac{\dot{a}}{a}\right)^2=\frac{8\pi G\rho}{3}-\frac{kc^2}{a^2}+\frac{\Lambda c^2}{3} \\ \frac{\ddot{a}}{a}=-\frac{4\pi G}{3c^2}(\rho+3P)+\frac{\Lambda c^2}{3} \end{gathered} \tag{12.5.545}$$

---

[69] The so named "Friedmann's second equation" that is also sometimes named "Raychaudhuri equation" or "acceleration equation"

Remark

The question now is how does the cosmological constant produce a repulsive force? Identifying the force from:

$$\frac{\ddot{R}}{R}=\frac{F_{\Lambda}}{mR}=\frac{c^2\Lambda}{3} \tag{12.5.546}$$

we get:

$$F_{\Lambda}=\frac{c^2\Lambda mR}{3} \tag{12.5.547}$$

This is supposed to be the cosmological force on a particle (galaxy) in a Universe with non-zero Λ.

Notice something interesting in the previous relation! If we consider (supposition!) that at the very early phases of its evolution, the universe was clearly not matter-dominated and that in the Planck epoch it was not even radiation-dominated, and the energy-momentum tensor was determined by the vacuum energy, then if we put therefore $\rho_0=0$ and $P=0$, we get:

$$\frac{\ddot{R}}{R}=\frac{c^2\Lambda}{3}\Leftrightarrow\frac{\mathrm{d}^2R(t)}{\mathrm{d}t^2}=\frac{c^2\Lambda}{3}R(t) \tag{12.5.548}$$

We see that a solution is e^{R_0t} but also e^{-R_0t} with $R_0=c\sqrt{\Lambda/3}$. So a more general solution is the sum of the solution to a given factor and therefore we can see then that:

$$R=2\frac{e^{R_0t}+e^{-R_0t}}{2}=2\cosh(R_0t) \tag{12.5.549}$$

is also a solution! When $t>0$ the second term e^{-R_0t} goes rapidly to zero and we have the inflationary expansion of the early universe. Note that nothing excludes the region $t<0$, where the first term e^{R_0t} goes rapidly to zero and the second term gives exponential inflation in the negative time direction...

Now we will use the latter relation to consider three situation: $\Lambda=0$, $\Lambda>0$, $\Lambda<0$ and for each of the latter case $k=0$, $k>0$ and $k<0$. That is a total of 9 scenarii (if we include for each scenario the case where respectively matter or radiations or vacuum dominates... we would have a total of 27 scenarii to study...!!!).

Let's start!:

- First main case we will take is $\Lambda=0$ (because it's the easiest one...) we have:

  $$\dot{R}^2=\frac{8\pi G}{3}\rho_0R^2-kc^2+\frac{c^2\Lambda}{3}R^2 \tag{12.5.550}$$

  that becomes:

  $$\dot{R}^2=\frac{8\pi G}{3}\rho_0R^2-kc^2 \tag{12.5.551}$$

  Therefore:

  $$\dot{R}^2\propto R^2-kc^2 \tag{12.5.552}$$

and:

$$\ddot{R}=-\frac{4\pi g}{3}R\left(\rho_0+\frac{3P}{c^2}\right)+\frac{c^2\Lambda}{3}R \quad (12.5.553)$$

becomes:

$$\ddot{R}=-\frac{4\pi g}{3}R\left(\rho_0+\frac{3P}{c^2}\right) \quad (12.5.554)$$

That is to say:

$$\ddot{R}<0 \quad (12.5.555)$$

So whatever happens and whatever the value of k, the Universe expansion is decelerating. Let us now see how this Universe behave according to the values of k:

- For $k=0$ (flat Universe) we get the "Einstein-De Sitter Universe model" proposed jointly by Albert Einstein and Willem De Sitter in 11932 (holocene calendar)!
  So for $k=0$ and $\Lambda=0$ we have:

  $$\dot{R}^2=\frac{8\pi G}{3c^2}\rho R^2-kc^2+\frac{c^2\Lambda}{3}R^2 \quad (12.5.556)$$

  that obviously becomes:

  $$3\frac{\dot{R}^2}{R^2}=\frac{8\pi G}{c^2}\rho \quad (12.5.557)$$

  But ρ depends on R. We have already introduced earlier (obvious) during our study of Newtonian models that (case of universe dominated by dust with the density parameter $\omega(T)=0$):

  $$\rho=\rho_0F^{-3}=\rho_0\left(\frac{R_0}{R}\right)^3 \quad (12.5.558)$$

  And it is common to take $R_0=1$ for the first time. Injecting into the Friedmann equation, we get:

  $$3\frac{\dot{R}^2}{R^2}=\frac{8\pi G}{c^2}\frac{\rho_0}{R^3} \quad (12.5.559)$$

  Therefore:

  $$\dot{R}=\sqrt{\frac{8\pi G}{3c^2}}\rho_0R^{-1/2} \quad (12.5.560)$$

  or explicitly:

  $$R^{1/2}\mathrm{d}R=\sqrt{\frac{8\pi G}{3c^2}}\rho_0\mathrm{d}t \quad (12.5.561)$$

  Integrating:

  $$\int\limits_1^R R^{1/2}\mathrm{d}R=\sqrt{\frac{8\pi G}{3c^2}}\rho_0\int\limits_0^t \mathrm{d}t \quad (12.5.562)$$

This gives:

$$\frac{2}{3}R^{3/2}-\frac{2}{3}1^{3/2}=\sqrt{\frac{8\pi G}{3c^2}\rho_0}t \tag{12.5.563}$$

We simplify a bit:

$$R^{3/2}\cong\frac{3}{2}\sqrt{\frac{8\pi G}{3c^2}\rho_0}t \tag{12.5.564}$$

Thus:

$$R_{\Lambda=0,k=0}\cong\left(\frac{6\pi G}{c^2}\rho_0\right)^{1/3}t^{2/3} \tag{12.5.565}$$

It is therefore an open Universe whose expansion will stop after an infinite time! It's exactly the same kind of expression that we have obtained during our study of Friedmann-Lemaître Euclidean cosmological models for flat spaces dominated by matter (see page **??**).

We won't redo here the case for flat space dominated by radiation that give us:

$$R_{\Lambda=0,k=0}\propto t^{1/3} \tag{12.5.566}$$

as all the details are already provided at page 5910.

But let us recall that for $k=0$ and $\Lambda=0$ we have:

$$\rho_0=\rho_c=\frac{3c^2H^2}{8\pi G} \tag{12.5.567}$$

Therefore:

$$\boxed{R_{\Lambda=0,k=0}\cong\left(\frac{9H^2}{4}\right)^{1/3}t^{2/3}} \tag{12.5.568}$$

Now we can estimate the age of the Universe according to this model. So to get the time since the beginning of our Universe we have by construction to put $R_{\Lambda=0,k=0}=1$. Therefore the above relation after rearranging becomes:

$$t_0\cong\sqrt{\frac{4}{9H^2}}=\frac{2}{3H}=\frac{2}{3}t_H\cong 8.7\cdot 10^9\text{ [years]} \tag{12.5.569}$$

This model of Universe as we already know has been abandoned because in contradiction with the ages of certain old stars whose age is estimated to be more than 10 billion years.

- Now, for the second sub-case (spherical), we have $k=+1$ (seems to have been developed first in 11922 according to holocene calendar):

$$\dot{R}^2=\frac{8\pi G}{3c^2}\rho R^2-kc^2+\frac{c^2\Lambda}{3}R^2 \tag{12.5.570}$$

that obviously becomes:

$$3\frac{\dot{R}^2}{R^2}+3\frac{c^2}{R^2}=\frac{8\pi G}{c^2}\rho \tag{12.5.571}$$

But ρ depends on R. We have already introduced earlier (obvious) during our study of Newtonian models (case of universe dominated by dust with the density parameter $\omega(T)=0$) that :

$$\rho=\rho_0 F^{-3}=\rho_0\left(\frac{R_0}{R}\right)^3 \tag{12.5.572}$$

And it is common to take $R_0=1$ for the first time. Injecting into the Friedmann equation, we get:

$$3\frac{\dot{R}^2}{R^2}+3\frac{c^2}{R^2}=\frac{8\pi G}{c^2}\frac{\rho_0}{R^3} \tag{12.5.573}$$

Therefore after rearranging and simplifying:

$$\dot{R}=\sqrt{\frac{8\pi G}{3c^2}\frac{\rho_0}{R}-c^2} \tag{12.5.574}$$

Hence:

$$\mathrm{d}R=c\sqrt{\frac{8\pi G}{3c^4}\frac{\rho_0}{R}-1}\mathrm{d}t \tag{12.5.575}$$

That we will rewrite as following for the purpose of notations simplifications and some future clever change of variables:

$$\mathrm{d}R=\frac{c}{R^{1/2}}\sqrt{\underbrace{\frac{8\pi G\rho_0}{3c^4}}_{=K}}\sqrt{1-\frac{R}{K}}\mathrm{d}t \tag{12.5.576}$$

Thus:

$$\frac{R^{1/2}}{c\sqrt{K}\sqrt{1-\dfrac{R}{K}}}\mathrm{d}R=\mathrm{d}t \tag{12.5.577}$$

Now before we continue let us recall that with $k=+1$ and $\Lambda=0$ one of the Friedmann equations becomes:

$$\frac{\dot{R}^2}{R^2}+\frac{c^2}{a^2}=\frac{8\pi G}{3c^2}\rho \tag{12.5.578}$$

Therefore:

$$\dot{R}^2=\frac{8\pi G}{3c^2}\rho R^2-c^2 \tag{12.5.579}$$

But as (still with the convention to take $R_0=1$):

$$\rho=\rho\left(\frac{R_0}{R}\right)^3 \tag{12.5.580}$$

The previous relation is then written:

$$\dot{R}^2=\frac{8\pi G\rho_0}{3Rc^2}-c^2 \tag{12.5.581}$$

As we have necessarily $\dot{R}^2 > 0$, then:

$$\dot{R}^2 = \frac{8\pi G\rho_0}{3Rc^2} - c^2 > 0 \tag{12.5.582}$$

Therefore:

$$\frac{8\pi G\rho_0}{3Rc^2} > c^2 \tag{12.5.583}$$

Thus:

$$\frac{8\pi G\rho_0}{3Rc^4} > R \Leftrightarrow K > R \Leftrightarrow \frac{R}{K} < 1 \tag{12.5.584}$$

And as $R > 0$ and $K > 0$ we have in fact:

$$0 < \frac{R}{K} < 1 \tag{12.5.585}$$

Now let us do the change of variable (since $0 < \sin^2(\theta) < 1$):

$$\frac{R}{K} = \sin^2(\theta) \Rightarrow \begin{cases} R^{1/2} = \sqrt{K}\sin(\theta) \\ \mathrm{d}R = 2K\sin(\theta)\cos(\theta)\mathrm{d}\theta \end{cases} \tag{12.5.586}$$

Therefore:

$$\frac{R^{1/2}}{c\sqrt{K}\sqrt{1-\dfrac{R}{K}}}\mathrm{d}R = \mathrm{d}t \tag{12.5.587}$$

can be rewritten:

$$\frac{R^{1/2}}{c\sqrt{K}\sqrt{1-\dfrac{R}{K}}}\mathrm{d}R = \frac{2K\cancel{\sqrt{K}}\sin^2(\theta)\cancel{\cos(\theta)}}{c\cancel{\sqrt{K}}\cancel{\cos(\theta)}}\mathrm{d}\theta = \mathrm{d}t \tag{12.5.588}$$

Therefore:

$$\int \frac{2K}{c}\sin^2(\theta)\mathrm{d}\theta = \int \mathrm{d}\theta \tag{12.5.589}$$

Using trigonometric identities:

$$\frac{2K}{c}\int_0^\theta \frac{1-\cos(2\theta)}{2}\mathrm{d}\theta = \int_0^t \mathrm{d}t \tag{12.5.590}$$

Thus:

$$\frac{K}{c}\left(\theta - \frac{1}{2}\sin(2\theta)\right) = t + c^{te} \tag{12.5.591}$$

For $t = 0$, we assume $R = 0$ and therefore $\theta = 0$. This bring us that $c^{te} = 0$.

We have then:

$$\begin{cases} t=\dfrac{K}{c}\left(\theta-\dfrac{1}{2}\sin(2\theta)\right)=\dfrac{8\pi G\rho_0}{3c^5}\left(\theta-\dfrac{1}{2}\sin(2\theta)\right) \\ R=K\sin^2(\theta)=\dfrac{8\pi G\rho_0}{3c^4}\dfrac{1}{2}(1-\cos(2\theta)) \end{cases} \tag{12.5.592}$$

Let us put $\theta'=2\theta$. Then:

$$\begin{cases} t=\dfrac{8\pi G\rho_0}{3c^5}\dfrac{1}{2}(\theta'-\sin(\theta')) \\ R=\dfrac{8\pi G\rho_0}{3c^4}\dfrac{1}{2}(1-\cos(\theta')) \end{cases} \tag{12.5.593}$$

we recognize here the parametric equations of a cycloid (see section Analytical Geometry page 2814). Therefore this is a bouncing cyclic Universe model!

We see that a will be maximum for $\theta=\pi$, then:

$$\begin{cases} t_{\text{max}}=\dfrac{8\pi^2 G\rho_0}{6c^5} \\ R_{\text{max}}=\dfrac{8\pi G\rho_0}{3c^4} \end{cases} \tag{12.5.594}$$

and the final state of the Big Crunch will occur when $\theta=2\pi$, where we then have $a=0$ and then:

$$t_{\text{Big Crunch}}=\frac{8\pi^2 G\rho_0}{3c^5} \tag{12.5.595}$$

- Now the third and last sub-case, we choose $k=-1$ (hyperbolic model that seems to have been developed in 11924 according to holocene calendar).

  The development are very similar to the previous case as just the sign change. We then also start from:

$$\dot{R}^2=\frac{8\pi G}{3c^2}\rho R^2-kc^2+\frac{c^2\Lambda}{3}R^2 \tag{12.5.596}$$

that obviously becomes:

$$3\frac{\dot{R}^2}{R^2}-3\frac{c^2}{R^2}=\frac{8\pi G}{c^2}\rho \tag{12.5.597}$$

But ρ depends on R. We have already introduced earlier (obvious) during our study of Newtonian models (case of universe dominated by dust with the density parameter $\omega(T)=0$) that:

$$\rho=\rho_0 F^{-3}=\rho_0\left(\frac{R_0}{R}\right)^3 \tag{12.5.598}$$

And it is common to take $R_0=1$ for the first time. Injecting into the Friedmann equation, we get:

$$3\frac{\dot{R}^2}{R^2}-3\frac{c^2}{R^2}=\frac{8\pi G}{c^2}\frac{\rho_0}{R^3} \tag{12.5.599}$$

Therefore after rearranging and simplifying:

$$\dot{R}=\sqrt{\frac{8\pi G}{3c^2}\frac{\rho_0}{R}+c^2} \tag{12.5.600}$$

Hence:

$$\mathrm{d}R=c\sqrt{\frac{8\pi G}{3c^4}\frac{\rho_0}{R}+1}\mathrm{d}t \tag{12.5.601}$$

That we will rewrite as following for the purpose of notations simplifications and some future clever change of variables:

$$\mathrm{d}R=\frac{c}{R^{1/2}}\sqrt{\underbrace{\frac{8\pi G\rho_0}{3c^4}}_{=K}}\sqrt{1+\frac{R}{K}}\mathrm{d}t \tag{12.5.602}$$

Thus:

$$\frac{R^{1/2}}{c\sqrt{K}\sqrt{1+\dfrac{R}{K}}}\mathrm{d}R=\mathrm{d}t \tag{12.5.603}$$

Now before we continue let us recall that with $k=-1$ and $\Lambda=0$ one of the Friedmann equations becomes:

$$\frac{\dot{R}^2}{R^2}-\frac{c^2}{a^2}=\frac{8\pi G}{3c^2}\rho \tag{12.5.604}$$

Therefore:

$$\dot{R}^2=\frac{8\pi G}{3c^2}\rho R^2+c^2 \tag{12.5.605}$$

But as (still with the convention to take $R_0=1$):

$$\rho=\rho\left(\frac{R_0}{R}\right)^3 \tag{12.5.606}$$

The previous relation is then written:

$$\dot{R}^2=\frac{8\pi G\rho_0}{3Rc^2}+c^2 \tag{12.5.607}$$

As we have necessarily $\dot{R}^2>0$ (because of the square!), then:

$$\dot{R}^2=\frac{8\pi G\rho_0}{3Rc^2}+c^2>0 \tag{12.5.608}$$

Therefore:

$$\frac{8\pi G\rho_0}{3Rc^2}>-c^2 \tag{12.5.609}$$

Thus:

$$\frac{8\pi G\rho_0}{3Rc^4}>-R\Leftrightarrow K>-R\Leftrightarrow\frac{R}{K}>-1 \tag{12.5.610}$$

But as $R>0$ and $K>0$ we have in fact:

$$\frac{R}{K}>0 \tag{12.5.611}$$

Now let us do the change of variable (since $\sinh^2(\theta)>0$):

$$\frac{R}{K}=\sinh^2(\theta)\Rightarrow\begin{cases}R^{1/2}=\sqrt{K}\sinh(\theta)\\ \mathrm{d}R=2K\sinh(\theta)\cosh(\theta)\mathrm{d}\theta\end{cases} \tag{12.5.612}$$

Therefore:

$$\frac{R^{1/2}}{c\sqrt{K}\sqrt{1+\dfrac{R}{K}}}\mathrm{d}R=\mathrm{d}t \tag{12.5.613}$$

can be rewritten:

$$\frac{R^{1/2}}{c\sqrt{K}\sqrt{1+\dfrac{R}{K}}}\mathrm{d}R=\frac{2K\cancel{\sqrt{K}}\sinh^2(\theta)\cancel{\cosh(\theta)}}{c\cancel{\sqrt{K}}\cancel{\cosh(\theta)}}\mathrm{d}\theta=\mathrm{d}t \tag{12.5.614}$$

Therefore:

$$\int\frac{2K}{c}\sinh^2(\theta)\mathrm{d}\theta=\int\mathrm{d}\theta \tag{12.5.615}$$

Using trigonometric identities:

$$\frac{2K}{c}\int\limits_0^\theta\frac{\cosh(2\theta)-1}{2}\mathrm{d}\theta=\int\limits_0^t\mathrm{d}t \tag{12.5.616}$$

Thus:

$$\frac{K}{c}\left(\frac{1}{2}\sin(2\theta)-\theta\right)=t+c^{te} \tag{12.5.617}$$

For $t=0$, we assume $R=0$ and therefore $\theta=0$. This bring us that $c^{te}=0$. We have then:

$$\begin{cases}t=\dfrac{K}{c}\left(\dfrac{1}{2}\sinh(2\theta)-\theta\right)=\dfrac{8\pi G\rho_0}{3c^5}\left(\dfrac{1}{2}\sin(2\theta)-\theta\right)\\ R=K\sinh^2(\theta)=\dfrac{8\pi G\rho_0}{3c^4}\dfrac{1}{2}(\cosh(2\theta)-1)\end{cases} \tag{12.5.618}$$

Let us put $\theta'=2\theta$. Then:

$$\begin{cases}t=\dfrac{8\pi G\rho_0}{3c^5}\dfrac{1}{2}(\sinh(\theta')-\theta')\\ R=\dfrac{8\pi G\rho_0}{3c^4}\dfrac{1}{2}(\cosh(\theta')-1)\end{cases} \tag{12.5.619}$$

So this is just an unbounded Universe!

- Now second main case with $\Lambda < 0$ (in the framework of FWLR universe all three corresponding sub-models are sometimes referred to some kind of "anti de Sitter space" abbreviated AdS).

  First let us start with a general behaviour analysis. Once again we start from the following Friedmann equations:

$$\begin{aligned}\dot{R}^2 &= \frac{8\pi G}{3c^2}\rho R^2 - kc^2 + \frac{c^2\Lambda}{3}R^2 \\ \ddot{R} &= -\frac{4\pi G}{3}\left(\rho + \frac{3P}{c^2}\right) + \frac{c^2\Lambda}{3}R\end{aligned} \tag{12.5.620}$$

  Then for $\Lambda < 0$ we can write:

$$\begin{aligned}\dot{R}^2 &= \frac{8\pi G}{3c^2}\rho R^2 - kc^2 - \frac{c^2|\Lambda|}{3}R^2 \\ \ddot{R} &= -\frac{4\pi G}{3}\left(\rho + \frac{3P}{c^2}\right) - \frac{c^2|\Lambda|}{3}R = -\frac{4\pi G}{3}R\left(\rho + \frac{3P}{c^2} + \frac{c^2|\Lambda|}{12\pi G}\right)\end{aligned} \tag{12.5.621}$$

  as the reader can see by the "$+$" in front of the Λ term we can already make the conclusion that when $\Lambda < 0$ it acts like an attractive force to the Universe.

  $R(t)$ is then monotonically increasing and therefore there exist a $\dot{R}(t)$ that will be equal to zero. So the Universe re-collapse happens even earlier due to "attractive force"!

  Now let us deal again with the three special case in the same order as before $\{k = 0, k > 0, k < 0\}$. So let's start:

  - For $k = 0$ we then have obviously:

$$\dot{R}^2 = \frac{8\pi G}{3c^2}\rho R^2 - \frac{c^2|\Lambda|}{3}R^2 \tag{12.5.622}$$

    after rearranging:

$$\left(\frac{\dot{R}}{R}\right)^2 = \frac{8\pi G}{3c^2}\rho - \frac{c^2|\Lambda|}{3} \tag{12.5.623}$$

    And now let us consider that ρ is small enough (universe without mass nor radiation but just a cosmological constant, i.e. dominated by vacuum with density parameter $\omega(T) = 1$) so that we can write:

$$\left(\frac{\dot{R}}{R}\right)^2 = -\frac{c^2|\Lambda|}{3} \tag{12.5.624}$$

    which yields the solution:

$$R = -\frac{3}{c^2|\Lambda|}e^{\sqrt{-\frac{c^2|\Lambda|}{3}}t} \tag{12.5.625}$$

    We have then a purely complex exponential as the radicand is negative! As we know that a complex exponential can be written in Euler form (trigonometric sine), that means we have an oscillating universe.

Remark

Notice that there is no constant of integration, otherwise the equality would not hold! So an interesting result (even if it's a priori not realistic to consider ρ as small...) is that when $t=0$, $R\neq 0$ (i.e. that Universe existed before the creation of time).

Obviously we have then that the Hubble parameter is a constant as by definition:

$$H=\frac{\dot{R}}{R}=\sqrt{-\frac{c^2|\Lambda|}{3}} \tag{12.5.626}$$

Against that means the Hubble parameter would be a complex value...
Now let us consider the case where ρ cannot be neglected, as we have proved earlier that in a flat universe dominated by matter we had:

$$\rho=\frac{\rho_0 R_0^3}{R^3} \tag{12.5.627}$$

then:

$$\left(\frac{\dot{R}}{R}\right)^2=\frac{8\pi G}{3c^2}\rho-\frac{c^2|\Lambda|}{3} \tag{12.5.628}$$

can be rewritten:

$$\left(\frac{\dot{R}}{R}\right)^2=\frac{8\pi G}{3c^2}\frac{\rho_0 R_0^3}{R^3}-\frac{c^2|\Lambda|}{3} \tag{12.5.629}$$

Hence:

$$\dot{R}=\sqrt{\frac{8\pi G}{3c^2}\frac{\rho_0 R_0^3}{R}-\frac{c^2|\Lambda|}{3}R^2} \tag{12.5.630}$$

Thus:

$$\mathrm{d}R=\sqrt{\frac{8\pi G}{3c^2}\frac{\rho_0 R_0^3}{R}-\frac{c^2|\Lambda|}{3}R^2}\mathrm{d}t \tag{12.5.631}$$

$$\mathrm{d}R=\sqrt{\frac{8\pi G}{3c^2}\frac{\rho_0 R_0^3}{R}+\frac{-c^2|\Lambda|}{3}R^2}\mathrm{d}t=\sqrt{\underbrace{\frac{8\pi G\rho_0 R_0^3}{3c^2}}_{:=K}\frac{1}{R}+\underbrace{\frac{-c^2|\Lambda|}{3}}_{:=B}R^2}\mathrm{d}t=\frac{\sqrt{K}}{R^{1/2}}\sqrt{1+\frac{B}{K}R^3}\mathrm{d}t \tag{12.5.632}$$

So:

$$\frac{R^{1/2}\mathrm{d}R}{\sqrt{K}\sqrt{1+\dfrac{B}{K}R^3}}=\mathrm{d}t \tag{12.5.633}$$

Let us do the following change of variable:

$$x^2=\frac{B}{K}R^3\Rightarrow\begin{cases}3R^2\mathrm{d}R=2\dfrac{K}{B}x\mathrm{d}x\\R^2=\left(\dfrac{K}{B}\right)^{2/3}x^{4/3}\\R^{1/2}=\left(\dfrac{K}{B}\right)^{1/6}x^{1/3}\end{cases}\tag{12.5.634}$$

Hence:

$$R^{1/2}\mathrm{d}R=\frac{\left(\dfrac{K}{B}\right)^{1/6}x^{1/3}\dfrac{2}{3}\dfrac{K}{B}x\mathrm{d}x}{\left(\dfrac{K}{B}\right)^{2/3}x^{4/3}}=\frac{2}{3}\left(\frac{K}{B}\right)^{1/2}\mathrm{d}x\tag{12.5.635}$$

So:

$$\frac{R^{1/2}\mathrm{d}R}{\sqrt{K}\sqrt{1+\dfrac{B}{K}R^3}}=\frac{\dfrac{2}{3}\left(\dfrac{K}{B}\right)^{1/2}\mathrm{d}x}{\sqrt{K}\sqrt{1+x^2}}=\mathrm{d}t\tag{12.5.636}$$

So we want to take the primitive of:

$$\frac{2}{3\sqrt{B}}\int\frac{\mathrm{d}x}{\sqrt{1+x^2}}=\int\mathrm{d}t\tag{12.5.637}$$

as we know from our study of usual derivatives (see section Differential and Integral Calculus page 1475) the integral is simply given by:

$$\frac{2}{3\sqrt{B}}\mathrm{arcsinh}(x)=t+c^{te}\tag{12.5.638}$$

For simplification purposes we will assume that for $t=0$, we have $R=0$ hence $c^{te}=0$ and we can write:

$$\frac{2}{3\sqrt{B}}\mathrm{arcsinh}(x)=t\Leftrightarrow\mathrm{arcsinh}\left(\sqrt{\frac{B}{K}}a^{3/2}\right)=\frac{3}{2}\sqrt{B}t\tag{12.5.639}$$

So:

$$\sqrt{\frac{B}{K}}a^{3/2}=\sinh\left(\frac{3}{2}\sqrt{B}t\right)\Leftrightarrow R=\sqrt[3]{\frac{K}{B}\sinh^2\left(\frac{3}{2}\sqrt{B}t\right)}\tag{12.5.640}$$

So finally:

$$R=R_0\sqrt[3]{-\frac{8\pi G\rho_0}{|\Lambda|c^4}\sinh^2\left(\frac{3}{2}\sqrt{\frac{-c^2|\Lambda|}{3}}t\right)}=R_0\sqrt[3]{-\frac{8\pi G\rho_0}{|\Lambda|c^4}\left[\cosh\left(\frac{3}{2}\sqrt{\frac{-c^2|\Lambda|}{3}}t\right)-1\right]}\tag{12.5.641}$$

We can plot such a function in $\mathbb{C}$ using Maple™ however its interpretation is quite non-trivial as there are no solution for $t \in \mathbb{R}^+$.

In most textbooks the final result is (we put $R_0 = 1$ by convention):

$$R = -\left(\frac{8\pi G\rho_0}{|\Lambda|c^4}\right)^{1/3} \sinh^{2/3}\left(\frac{3}{2}\sqrt{\frac{-c^2|\Lambda|}{3}}t\right) \qquad (12.5.642)$$

- Now for $k = +1$, the second sub-case (spherical).

  Let us first recall that we have already treated the case $k = +1$ with $\Lambda > 0$ and assuming $R = c^{te}$. This is the Einstein (static) Universe model already treated on page 6012. If in the same developments we put instead $\Lambda < 0$, this leads us to the same kind of conclusions.

  But let us deal with the non-static case for $k = +1$.

  For this purpose consider the general case:

$$\dot{R}^2 = \frac{8\pi G}{3c^2}\rho R^2 - kc^2 - \frac{c^2|\Lambda|}{3}R^2 \qquad (12.5.643)$$

This differential equation can be rewritten as:

$$\dot{R}^2 + \left(\frac{c^2|\Lambda|}{3} - \frac{8\pi G}{3c^2}\rho\right)R^2 + kc^2 = 0 \qquad (12.5.644)$$

To simplify let us put:

$$\alpha := \frac{c^2|\Lambda|}{3} - \frac{8\pi G}{3c^2}\rho \qquad (12.5.645)$$

Therefore:

$$\dot{R}^2 + \alpha R^2 + kc^2 = 0 \qquad (12.5.646)$$

Let us now move in natural units to get rid of the c^2:

$$\dot{R}^2 + \alpha R^2 + k = 0 \qquad (12.5.647)$$

Let us put $k = +1$ and assume that the density ρ is not dependent of time (it's a first case - equivalent to consider as we already know the universe being dominated by vacuum) and keep in mind that's equivalent also to consider ρ very small like $\rho \cong 0$ as 0 is a constant independent of time...):

$$\dot{R}^2 + \alpha R^2 + 1 = 0 \qquad (12.5.648)$$

Therefore:

$$\dot{R} = \pm\sqrt{-1 - \alpha^2 R^2} \qquad (12.5.649)$$

A simple integration leads to:

$$\int \frac{\mathrm{d}R}{\sqrt{-1-\alpha^2 R^2}} = \pm\int \mathrm{d}t \qquad (12.5.650)$$

For $\alpha \neq 0$ (we would need to treat the case $a=0$ and the case $-1-\alpha^2R^2=0 \Longrightarrow R=\pm 1/\alpha$), we substitute $u=\alpha R$:

$$\int \frac{\mathrm{d}u}{\sqrt{-1-u^2}}=\pm\alpha\int \mathrm{d}t=\pm\alpha t+c^{te} \tag{12.5.651}$$

Hence:

$$\int \frac{\mathrm{i}\mathrm{d}u}{\sqrt{1+u^2}}=\pm\alpha t+c^{te} \tag{12.5.652}$$

And we have already proven this primitive in the section of Differential and Integral Calculus (see page 1534):

$$\ln\left(u+\sqrt{u^2+\alpha^2}\right)=\pm\alpha t+c^{te} \tag{12.5.653}$$

Hence:

$$\mathrm{i}\ln\left(\alpha R+\sqrt{\alpha^2R^2+\alpha^2}\right)=\pm\alpha t+c^{te} \tag{12.5.654}$$

Hence:

$$\mathrm{i}\ln\left(\alpha R+\sqrt{\alpha^2R^2+\alpha^2}\right)=\pm\alpha t+c^{te} \tag{12.5.655}$$

Or:

$$\mathrm{i}\cdot\operatorname{arcsinh}(\alpha R)=\pm\alpha t+c^{te} \tag{12.5.656}$$

So finally:

$$R(t)=\sinh\left(\frac{\mathrm{i}(\pm\alpha t+c^{te})}{\alpha}\right) \tag{12.5.657}$$

This is a pure complex hyperbolic sine function. The interpretation is in our point of view non-trivial so will not try to extrapolate, even with a plot, this result. And remember that we assume that:

$$\alpha\neq 0\Rightarrow \frac{c^2|\Lambda|}{3}\neq\frac{8\pi G}{3c^2}\rho \tag{12.5.658}$$

Or said otherwise:

$$\rho\neq\frac{c^4|\Lambda|}{8\pi G} \tag{12.5.659}$$

Ok so that's the result for a spherical universe dominated by vacuum. Assuming that only ρ is a function of R only and that is does not explicitly depend on time, we can write:

$$\dot{R}=\frac{\mathrm{d}R}{\mathrm{d}t}=\sqrt{\frac{8\pi G}{3c^2}\rho(R)R^2-kc^2-\frac{c^2|\Lambda|}{3}R^2} \tag{12.5.660}$$

But we have proven (see page 5912) that the different most common type of density evolution are:

$$\rho(t)\propto R(t)^{-3(1+\omega(T))}\propto\begin{cases}R(t)^{-3} & \text{for } \omega(T)=0 \text{ (pressureless matter (i.e. dust))}\\ R(t)^{-4} & \text{for } \omega(T)=1/3 \text{ (radiation/relativistic matter)}\\ c^{te} & \text{for } \omega(T)=-1 \text{ (vacuum energy)}\end{cases} \tag{12.5.661}$$

Let us consider that we are interested about the first instants of the Big Bang (this is an excuse in reality because we didn't found any analytic solution for the presureless matter, i.e. $\omega(T)=0$, case.........). Then we will focus on radiation/relativistic matter, i.e. $\rho\propto R(t)^{-4}$ (and for simplicity we put $R_0=1$):

$$\dot{R}=\sqrt{\frac{8\pi G}{3c^2}\rho_0\frac{R_0^4}{R^4}R^2-kc^2-\frac{c^2|\Lambda|}{3}R^2}=\sqrt{\frac{8\pi G}{3c^2}\rho_0 R^{-2}-kc^2-\frac{c^2|\Lambda|}{3}R^2} \tag{12.5.662}$$

To simplify let us denote the latter:

$$\dot{R}=\sqrt{\alpha R^{-2}-\beta+\gamma R^2} \tag{12.5.663}$$

Hence:

$$\int\mathrm{d}t=\int\frac{1}{\sqrt{\alpha R^{-2}-\beta+\gamma R^2}}\mathrm{d}R \tag{12.5.664}$$

Let us put first the right hand side at the same denominator:

$$\int\frac{R}{\sqrt{\gamma R^4-\beta R^2+\alpha}}\mathrm{d}R \tag{12.5.665}$$

Let us substitute $u=R^2\longrightarrow\mathrm{d}u=2R\mathrm{d}R$, therefore:

$$\begin{aligned}\int\frac{R}{\sqrt{\gamma R^4-\beta R^2+\alpha}}\mathrm{d}R&=\frac{1}{2}\int\frac{1}{\sqrt{\gamma u^2-\beta u+\alpha}}\mathrm{d}u\\&=\int\frac{1}{\sqrt{\left(\sqrt{\gamma}u-\frac{\beta}{2\sqrt{\gamma}}\right)^2-\frac{\beta^2}{4\gamma}+\alpha}}\mathrm{d}u\end{aligned} \tag{12.5.666}$$

Let us substitute:

$$v=\frac{2\gamma u-\beta}{2\sqrt{\alpha-\frac{\beta^2}{4\gamma}}\sqrt{\gamma}}\to\mathrm{d}v=\frac{\sqrt{\gamma}}{\sqrt{\alpha-\frac{\beta^2}{4\gamma}}}\mathrm{d}u \tag{12.5.667}$$

Therefore we get:

$$\frac{1}{\sqrt{\gamma}}\int\frac{1}{\sqrt{v^2+1}}\,\mathrm{d}v \tag{12.5.668}$$

Now solving:

$$\int\frac{1}{\sqrt{v^2+1}}\,\mathrm{d}v \tag{12.5.669}$$

This is a standard integral (see section Differential and Integral Calculus page 1534):

$$\int \frac{1}{\sqrt{v^2+1}}\mathrm{d}v=\ln\left(\sqrt{v^2+1}+v\right) \tag{12.5.670}$$

Plug in solved integrals:

$$\begin{aligned}\frac{1}{\sqrt{\gamma}}\int \frac{1}{\sqrt{v^2+1}}\mathrm{d}v=\frac{\ln\left(\sqrt{v^2+1}+v\right)}{\sqrt{\gamma}}&=\frac{\ln\left(\sqrt{\frac{(2\gamma u-\beta)^2}{4\gamma\left(\alpha-\frac{\beta^2}{4\gamma}\right)}+1}+\frac{2\gamma u-\beta}{2\sqrt{\alpha-\frac{\beta^2}{4\gamma}}\sqrt{\gamma}}\right)}{\sqrt{\gamma}}\\
&=\frac{\ln\left(\sqrt{\frac{\left(2\gamma R^2-\beta\right)^2}{4\gamma\left(\alpha-\frac{\beta^2}{4\gamma}\right)}+1}+\frac{2\gamma R^2-\beta}{2\sqrt{\alpha-\frac{\beta^2}{4\gamma}}\sqrt{\gamma}}\right)}{2\sqrt{\gamma}}\\
&=\frac{\ln\left(\sqrt{\frac{\left(2\gamma R^2-\beta\right)^2+4\gamma\left(\alpha-\frac{\beta^2}{4\gamma}\right)}{4\gamma\left(\alpha-\frac{\beta^2}{4\gamma}\right)}}+\frac{2\gamma R^2-\beta}{2\sqrt{\alpha-\frac{\beta^2}{4\gamma}}\sqrt{\gamma}}\right)}{2\sqrt{\gamma}}\\
&=\frac{\ln\left(\frac{1}{2\sqrt{\alpha-\frac{\beta^2}{4\gamma}}\sqrt{\gamma}}\left[\sqrt{\left(2\gamma R^2-\beta\right)^2+4\gamma\left(\alpha-\frac{\beta^2}{4\gamma}\right)}+2\gamma R^2-\beta\right]\right)}{2\sqrt{\gamma}}\\
&=\frac{\ln\left(\sqrt{\left(2\gamma R^2-\beta\right)^2+4\gamma\left(\alpha-\frac{\beta^2}{4\gamma}\right)}+2\gamma R^2-\beta\right)}{2\sqrt{\gamma}}-\frac{\ln\left(2\sqrt{\gamma}\sqrt{\alpha-\frac{\beta^2}{4\gamma}}\right)}{2\sqrt{\gamma}}\\
&=\frac{\ln\left(\sqrt{\left(2\gamma R^2-\beta\right)^2+\alpha 4\gamma-\beta^2}+2\gamma R^2-\beta\right)}{2\sqrt{\gamma}}+c^{te}\\
&=\frac{\ln\left(2\left(\sqrt{\gamma}\sqrt{R^2\left(\gamma R^2-\beta\right)+\alpha}+\gamma R^2\right)-\beta\right)}{2\sqrt{\gamma}}+c^{te}\end{aligned} \tag{12.5.671}$$

Hence:

$$\int \frac{1}{\sqrt{\frac{\alpha}{R^2}-\beta+\gamma R^2}}\mathrm{d}R=\frac{\ln\left(2\left(\sqrt{\gamma}\sqrt{R^2\left(\gamma R^2-\beta\right)+\alpha}+\gamma R^2\right)-\beta\right)}{2\sqrt{\gamma}}+c^{te} \tag{12.5.672}$$

Reinjecting the original constants we get:

$$\int \frac{1}{\sqrt{\frac{8\pi G}{3c^2}\rho_0 R^{-2}-kc^2-\frac{c^2|\Lambda|}{3}R^2}}\mathrm{d}R=\frac{\ln\left(2\left(\sqrt{\frac{-c^2|\Lambda|}{3}}\sqrt{R^2\cdot\left(\frac{-c^2|\Lambda|}{3}R^2-kc^2\right)+\frac{8\pi G}{3c^2}\rho_0}+\frac{-c^2|\Lambda|}{3}R^2\right)-kc^2\right)}{2\sqrt{\frac{-c^2|\Lambda|}{3}}}+c^{te} \tag{12.5.673}$$

So finally for a spherical universe (replacing $k=+1$) dominated by radiation, we have:

$$t=\frac{\ln\left(2\left(\sqrt{\frac{-c^2|\Lambda|}{3}}\sqrt{R^2\cdot\left(\frac{-c^2|\Lambda|}{3}R^2-c^2\right)+\frac{8\pi G}{3c^2}\rho_0}+\frac{-c^2|\Lambda|}{3}R^2\right)-c^2\right)}{2\sqrt{\frac{-c^2|\Lambda|}{3}}}+c^{te} \tag{12.5.674}$$

Sadly we didn't found yet a way to extract the R above as a function of t nor a pertinent way to do an implicit plot with a computer software. Furthermore this situation is even worst than the case with $\Lambda>0$ as we have complex roots and therefore the function is imaginary. So we will have to stop here!

- Let us now consider the case $k=-1$ (hyperbolic).

For this, same as before, we start from:

$$\dot{R}^2=\frac{8\pi G}{3c^2}\rho R^2-kc^2-\frac{c^2|\Lambda|}{3}R^2 \tag{12.5.675}$$

This differential equation can be rewritten as:

$$\dot{R}^2+\left(\frac{c^2|\Lambda|}{3}-\frac{8\pi G}{3c^2}\rho\right)R^2+kc^2=0 \tag{12.5.676}$$

To simplify let us put:

$$\alpha:=\frac{c^2|\Lambda|}{3}-\frac{8\pi G}{3c^2}\rho \tag{12.5.677}$$

Therefore:

$$\dot{R}^2+\alpha R^2+kc^2=0 \tag{12.5.678}$$

Let us now move in natural units to get rid of the c^2:

$$\dot{R}^2+\alpha R^2+k=0 \tag{12.5.679}$$

Let us put $k=-1$ and assume that the density ρ is not dependent of time (it's a first case - equivalent to consider as we already know the universe being dominated by vacuum) and keep in mind that's equivalent also to consider ρ very small like $\rho\cong 0$ as 0 is a constant independent of time...):

$$\dot{R}^2+\alpha R^2-1=0 \tag{12.5.680}$$

Therefore:

$$\dot{R}=\pm\sqrt{1-\alpha^2R^2} \tag{12.5.681}$$

A simple integration leads to:

$$\int\frac{\mathrm{d}R}{\sqrt{1-\alpha^2R^2}}=\pm\int\mathrm{d}t \tag{12.5.682}$$

For $\alpha\neq 0$ (we would need to treat the case $\alpha=0$ and the case $1-\alpha^2R^2=0\Longrightarrow R=\pm 1/\alpha$), we substitute $u=\alpha R$:

$$\int\frac{\mathrm{d}u}{\sqrt{1-u^2}}=\pm\alpha\int\mathrm{d}t=\pm\alpha t+c^{te} \tag{12.5.683}$$

Let us substitute $u=\sin(x)$, then:

$$\int\frac{\cos(x)\mathrm{d}x}{\sqrt{1-\sin^2(x)}}=\pm\alpha t+c^{te} \tag{12.5.684}$$

After simplification using trigonometric identities:

$$\int \mathrm{d}x = \pm\alpha t + c^{te} \tag{12.5.685}$$

Therefore:

$$x(t) = \pm\alpha t + c^{te} \tag{12.5.686}$$

Hence:

$$u(t) = \sin\left(\pm\alpha t + c^{te}\right) \tag{12.5.687}$$

So finally under the assumptions taken earlier:

$$R(t) = \frac{1}{\alpha}\sin\left(\pm\alpha t + c^{te}\right) \tag{12.5.688}$$

Explicitly (in natural units $c = 1, G = 1$):

$$R(t) = \frac{3}{|\Lambda| - 8\pi\rho}\sin\left(\pm\frac{|\Lambda| - 8\pi\rho}{3}t + c^{te}\right) \tag{12.5.689}$$

Therefore we have a cyclic universe (whatever the sign of Λ) and remember that we assume that:

$$\alpha \neq 0 \Rightarrow \frac{c^2|\Lambda|}{3} \neq \frac{8\pi G}{3c^2}\rho \tag{12.5.690}$$

Or said otherwise:

$$\rho \neq \frac{c^4|\Lambda|}{8\pi G} \tag{12.5.691}$$

Ok so that's the result for a hyperbolic universe dominated by vacuum. Assuming that only ρ is a function of R only and that is does not explicitly depend on time, we can write:

$$\dot{R} = \frac{\mathrm{d}R}{\mathrm{d}t} = \sqrt{\frac{8\pi G}{3c^2}\rho(R)R^2 - kc^2 - \frac{c^2|\Lambda|}{3}R^2} \tag{12.5.692}$$

But we have proven (see page 5912) that the different most common type of density evolution are:

$$\rho(t) \propto R(t)^{-3(1+\omega(T))} \propto \begin{cases} R(t)^{-3} & \text{for } \omega(T) = 0 \text{ (pressureless matter (i.e. dust))} \\ R(t)^{-4} & \text{for } \omega(T) = 1/3 \text{ (radiation/relativistic matter)} \\ c^{te} & \text{for } \omega(T) = -1 \text{ (vacuum energy)} \end{cases} \tag{12.5.693}$$

Let us consider that we are interested about the first instants of the Big Bang (this is an excuse in reality because we didn't found any analytic solution for the presureless matter, i.e. $\omega(T) = 0$, case.........). Then we will focus on radiation/relativistic matter, i.e. $\rho \propto R(t)^{-4}$ (and for simplicity we put $R_0 = 1$):

$$\dot{R} = \sqrt{\frac{8\pi G}{3c^2}\rho_0\frac{R_0^4}{R^4}R^2 - kc^2 - \frac{c^2|\Lambda|}{3}R^2} = \sqrt{\frac{8\pi G}{3c^2}\rho_0 R^{-2} - kc^2 - \frac{c^2|\Lambda|}{3}R^2} \tag{12.5.694}$$

To simplify let us denote the latter:

$$\dot{R}=\sqrt{\alpha R^{-2}-\beta+\gamma R^{2}} \tag{12.5.695}$$

Hence:

$$\int \mathrm{d}t=\int \frac{1}{\sqrt{\alpha R^{-2}-\beta+\gamma R^{2}}}\mathrm{d}R \tag{12.5.696}$$

Let us put first the right hand side at the same denominator:

$$\int \frac{R}{\sqrt{\gamma R^{4}-\beta R^{2}+\alpha}}\mathrm{d}R \tag{12.5.697}$$

Let us substitute $u=R^{2}\longrightarrow \mathrm{d}u=2R\mathrm{d}R$, therefore:

$$\int \frac{R}{\sqrt{\gamma R^{4}-\beta R^{2}+\alpha}}\mathrm{d}R=\frac{1}{2}\int \frac{1}{\sqrt{\gamma u^{2}-\beta u+\alpha}}\mathrm{d}u=\int \frac{1}{\sqrt{\left(\sqrt{\gamma}u-\frac{\beta}{2\sqrt{\gamma}}\right)^{2}-\frac{\beta^{2}}{4\gamma}+\alpha}}\mathrm{d}u \tag{12.5.698}$$

Let us substitute:

$$v=\frac{2\gamma u-\beta}{2\sqrt{\alpha-\frac{\beta^{2}}{4\gamma}}\sqrt{\gamma}}\rightarrow \mathrm{d}v=\frac{\sqrt{\gamma}}{\sqrt{\alpha-\frac{\beta^{2}}{4\gamma}}}\mathrm{d}u \tag{12.5.699}$$

Therefore we get:

$$\frac{1}{\sqrt{\gamma}}\int \frac{1}{\sqrt{v^{2}+1}}\,\mathrm{d}v \tag{12.5.700}$$

Now solving:

$$\int \frac{1}{\sqrt{v^{2}+1}}\,\mathrm{d}v \tag{12.5.701}$$

This is a standard integral (see section Differential and Integral Calculus page 1534):

$$\int \frac{1}{\sqrt{v^{2}+1}}\mathrm{d}v=\ln\left(\sqrt{v^{2}+1}+v\right) \tag{12.5.702}$$

Plug in solved integrals:

$$\begin{aligned}\frac{1}{\sqrt{\gamma}}\int\frac{1}{\sqrt{v^2+1}}\mathrm{d}v&=\frac{\ln\left(\sqrt{v^2+1}+v\right)}{\sqrt{\gamma}}=\frac{\ln\left(\sqrt{\frac{(2\gamma u-\beta)^2}{4\gamma\left(\alpha-\frac{\beta^2}{4\gamma}\right)}+1}+\frac{2\gamma u-\beta}{2\sqrt{\alpha-\frac{\beta^2}{4\gamma}}\sqrt{\gamma}}\right)}{\sqrt{\gamma}}\\
&=\frac{\ln\left(\sqrt{\frac{\left(2\gamma R^2-\beta\right)^2}{4\gamma\left(\alpha-\frac{\beta^2}{4\gamma}\right)}+1}+\frac{2\gamma R^2-\beta}{2\sqrt{\alpha-\frac{\beta^2}{4\gamma}}\sqrt{\gamma}}\right)}{2\sqrt{\gamma}}\\
&=\frac{\ln\left(\sqrt{\frac{\left(2\gamma R^2-\beta\right)^2+4\gamma\left(\alpha-\frac{\beta^2}{4\gamma}\right)}{4\gamma\left(\alpha-\frac{\beta^2}{4\gamma}\right)}}+\frac{2\gamma R^2-\beta}{2\sqrt{\alpha-\frac{\beta^2}{4\gamma}}\sqrt{\gamma}}\right)}{2\sqrt{\gamma}}\\
&=\frac{\ln\left(\frac{1}{2\sqrt{\alpha-\frac{\beta^2}{4\gamma}}\sqrt{\gamma}}\left[\sqrt{\left(2\gamma R^2-\beta\right)^2+4\gamma\left(\alpha-\frac{\beta^2}{4\gamma}\right)}+2\gamma R^2-\beta\right]\right)}{2\sqrt{\gamma}}\\
&=\frac{\ln\left(\sqrt{\left(2\gamma R^2-\beta\right)^2+4\gamma\left(\alpha-\frac{\beta^2}{4\gamma}\right)}+2\gamma R^2-\beta\right)}{2\sqrt{\gamma}}-\frac{\ln\left(2\sqrt{\gamma}\sqrt{\alpha-\frac{\beta^2}{4\gamma}}\right)}{2\sqrt{\gamma}}\\
&=\frac{\ln\left(\sqrt{\left(2\gamma R^2-\beta\right)^2+\alpha 4\gamma-\beta^2}+2\gamma R^2-\beta\right)}{2\sqrt{\gamma}}+c^{te}\\
&=\frac{\ln\left(2\left(\sqrt{\gamma}\sqrt{R^2\left(\gamma R^2-\beta\right)+\alpha}+\gamma R^2\right)-\beta\right)}{2\sqrt{\gamma}}+c^{te}\end{aligned}\tag{12.5.703}$$

Hence:

$$\int\frac{1}{\sqrt{\frac{\alpha}{R^2}-\beta+\gamma R^2}}\mathrm{d}R=\frac{\ln\left(2\left(\sqrt{\gamma}\sqrt{R^2\left(\gamma R^2-\beta\right)+\alpha}+\gamma R^2\right)-\beta\right)}{2\sqrt{\gamma}}+c^{te}\tag{12.5.704}$$

Reinjecting the original constants we get:

$$\int\frac{1}{\sqrt{\frac{8\pi G}{3c^2}\rho_0 R^{-2}-kc^2-\frac{c^2|\Lambda|}{3}R^2}}\mathrm{d}R=\frac{\ln\left(2\left(\sqrt{\frac{-c^2|\Lambda|}{3}}\sqrt{R^2\cdot\left(\frac{-c^2|\Lambda|}{3}R^2-kc^2\right)+\frac{8\pi G}{3c^2}\rho_0}+\frac{-c^2|\Lambda|}{3}R^2\right)-kc^2\right)}{2\sqrt{\frac{-c^2|\Lambda|}{3}}}+c^{te}\tag{12.5.705}$$

So finally for a hyperbolic universe (replacing $k=-1$) dominated by radiation, we have:

$$t=\frac{\ln\left(2\left(\sqrt{\frac{-c^2|\Lambda|}{3}}\sqrt{R^2\cdot\left(\frac{-c^2|\Lambda|}{3}R^2+c^2\right)+\frac{8\pi G}{3c^2}\rho_0}+\frac{-c^2|\Lambda|}{3}R^2\right)+c^2\right)}{2\sqrt{\frac{-c^2|\Lambda|}{3}}}+c^{te}\tag{12.5.706}$$

Sadly we didn't found yet a way to extract the R above as a function of t nor a pertinent way to do an implicit plot with a computer software. Furthermore this situation is even worst than the case with $\Lambda > 0$ as we have complex roots and therefore the function is imaginary. So we will have to stop here!

- Now third and last main case with $\Lambda > 0$ (in the framework of FWLR universe all three corresponding sub-models are sometimes referred to some kind of "de Sitter space") that corresponds to the actual experimental measurements of our observable Universe in this early 121st century (according to holocene calendar)!

  First let us start also with a general behaviour analysis. Once again we start from the following Friedmann equations:

$$\begin{aligned}\dot{R}^2 &= \frac{8\pi G}{3c^2}\rho R^2 - kc^2 + \frac{c^2\Lambda}{3}R^2 \\ \ddot{R} &= -\frac{4\pi G}{3}\left(\rho + \frac{3P}{c^2}\right) + \frac{c^2\Lambda}{3}R\end{aligned} \tag{12.5.707}$$

  Then for $\Lambda > 0$ we can write:

$$\begin{aligned}\dot{R}^2 &= \frac{8\pi G}{3}\rho R^2 - kc^2 - \frac{c^2|\Lambda|}{3}R^2 = \frac{8\pi G}{3}\rho_{m-\Lambda}R^2 \\ \ddot{R} &= -\frac{4\pi G}{3}\left(\rho + \frac{3P}{c^2}\right) - \frac{c^2|\Lambda|}{3}R = -\frac{4\pi G}{3}R\left(\rho + \frac{3P}{c^2} - \frac{c^2|\Lambda|}{12\pi G}\right)\end{aligned} \tag{12.5.708}$$

  as you can see by the "$-$" in front of the Λ term we can already make the conclusion that when $\Lambda > 0$ it acts like a repulsive force to the Universe.

  $R(t)$ is then forever expanding. Notice also that ρ drops as Λ is assumed to remain constant...

  - For $k = 0$ we then have obviously:

$$\dot{R}^2 = \frac{8\pi G}{3c^2}\rho R^2 + \frac{c^2\Lambda}{3}R^2 \tag{12.5.709}$$

    after rearranging:

$$\left(\frac{\dot{R}}{R}\right)^2 = \frac{8\pi G}{3c^2}\rho + \frac{c^2\Lambda}{3} \tag{12.5.710}$$

    And now let us consider that ρ is small enough (universe without mass nor radiation but just a cosmological constant, i.e. dominated by vacuum with density parameter $\omega(T) = -1$) so that we can write:

$$\left(\frac{\dot{R}}{R}\right)^2 = \frac{c^2\Lambda}{3} \tag{12.5.711}$$

    which yields the solution:

$$R = \frac{3}{c^2\Lambda}e^{\sqrt{\frac{c^2\Lambda}{3}}t} \tag{12.5.712}$$

    Thus, the Universe expands exponentially! Some physicists (and astrophysicists/cosmologists) associate this to the "inflation" period of the Universe. There is also a solution on the negative side of the time axis that most textbooks ignore...

Remark

Notice that there is no constant of integration, otherwise the equality would not hold! So an interesting result (even if it's a priori not realistic to consider ρ as small...) is that when $t=0$, $R\neq 0$ (i.e. that Universe existed before the creation of time).

This model of universe is named the "De Sitter Universe model" (not to be confused with the Einstein-De Sitter Universe model seen earlier!).
Obviously we have then that the Hubble parameter is a constant as by definition:

$$H=\frac{\dot{R}}{R}=\sqrt{\frac{c^2\Lambda}{3}} \tag{12.5.713}$$

Before we continue on the topic, let us notice that now the reader has all the knowledge to fully understand the following table (if necessary we recommend the reader to refer again the complete table of most common w values at page 5912) that we can found in some textbook treating only of flat spaces ($k=0$):

Name	Equation	Radiation	Matter	Vacuum
Eq. of state	$p=w\rho$	$1/3$	0	-1
Energy density	$\rho\propto a^{-3(1+w)}$	a^{-4}	a^{-3}	constant
Scale Factor	$a\propto t^{2/3(1+w)}$	$t^{1/2}$	$t^{2/3}$	e^{Ht}

Table 12.27: **Friedmann-Lemaître-Robertson-Walker cosmological flat spaces summary**

Now let us consider the case where ρ cannot be neglected, as we have proved earlier that in a flat universe dominated by matter we had:

$$\rho=\frac{\rho_0 R_0^3}{R^3} \tag{12.5.714}$$

then:

$$\left(\frac{\dot{R}}{R}\right)^2=\frac{8\pi G}{3c^2}\rho+\frac{c^2\Lambda}{3} \tag{12.5.715}$$

can be rewritten:

$$\left(\frac{\dot{R}}{R}\right)^2=\frac{8\pi G}{3c^2}\frac{\rho_0 R_0^3}{R^3}+\frac{c^2\Lambda}{3} \tag{12.5.716}$$

Hence:

$$\dot{R}=\sqrt{\frac{8\pi G}{3c^2}\frac{\rho_0 R_0^3}{R}+\frac{c^2\Lambda}{3}R^2} \tag{12.5.717}$$

Thus:

$$\mathrm{d}R=\sqrt{\frac{8\pi G}{3c^2}\frac{\rho_0 R_0^3}{R}+\frac{c^2\Lambda}{3}R^2}\mathrm{d}t \tag{12.5.718}$$

$$\mathrm{d}R=\sqrt{\frac{8\pi G}{3c^2}\frac{\rho_0R_0^3}{R}+\frac{c^2\Lambda}{3}R^2}\mathrm{d}t=\sqrt{\underbrace{\frac{8\pi G\rho_0R_0^3}{3c^2}}_{:=K}\frac{1}{R}+\underbrace{\frac{c^2\Lambda}{3}}_{:=B}R^2}\mathrm{d}t=\frac{\sqrt{K}}{R^{1/2}}\sqrt{1+\frac{B}{K}R^3}\mathrm{d}t \tag{12.5.719}$$

So:

$$\frac{R^{1/2}\mathrm{d}R}{\sqrt{K}\sqrt{1+\dfrac{B}{K}R^3}}=\mathrm{d}t \tag{12.5.720}$$

Let us do the following change of variable:

$$x^2=\frac{B}{K}R^3\Rightarrow\begin{cases}3R^2\mathrm{d}R=2\dfrac{K}{B}x\mathrm{d}x\\ R^2=\left(\dfrac{K}{B}\right)^{2/3}x^{4/3}\\ R^{1/2}=\left(\dfrac{K}{B}\right)^{1/6}x^{1/3}\end{cases} \tag{12.5.721}$$

Hence:

$$R^{1/2}\mathrm{d}R=\frac{\left(\dfrac{K}{B}\right)^{1/6}x^{1/3}\dfrac{2}{3}\dfrac{K}{B}x\mathrm{d}x}{\left(\dfrac{K}{B}\right)^{2/3}x^{4/3}}=\frac{2}{3}\left(\frac{K}{B}\right)^{1/2}\mathrm{d}x \tag{12.5.722}$$

So:

$$\frac{R^{1/2}\mathrm{d}R}{\sqrt{K}\sqrt{1+\dfrac{B}{K}R^3}}=\frac{\dfrac{2}{3}\left(\dfrac{K}{B}\right)^{1/2}\mathrm{d}x}{\sqrt{K}\sqrt{1+x^2}}=\mathrm{d}t \tag{12.5.723}$$

So we want to take the primitive of:

$$\frac{2}{3\sqrt{B}}\int\frac{\mathrm{d}x}{\sqrt{1+x^2}}=\int\mathrm{d}t \tag{12.5.724}$$

as we know from our study of usual derivatives (see section Differential and Integral Calculus page 1475) the integral is simply given by:

$$\frac{2}{3\sqrt{B}}\text{arcsinh}(x)=t+c^{te} \tag{12.5.725}$$

For simplification purposes we will assume that for $t=0$, we have $R=0$ hence $c^{te}=0$ and we can write:

$$\frac{2}{3\sqrt{B}}\text{arcsinh}(x)=t\Leftrightarrow\text{arcsinh}\left(\sqrt{\frac{B}{K}}a^{3/2}\right)=\frac{3}{2}\sqrt{B}t \tag{12.5.726}$$

So:

$$\sqrt{\frac{B}{K}}a^{3/2}=\sinh\left(\frac{3}{2}\sqrt{B}t\right)\Leftrightarrow R=\sqrt[3]{\frac{K}{B}\sinh^2\left(\frac{3}{2}\sqrt{B}t\right)} \tag{12.5.727}$$

So finally:

$$R=R_0\sqrt[3]{\frac{8\pi G\rho_0}{\Lambda c^4}\sinh^2\left(\frac{3}{2}\sqrt{\frac{c^2\Lambda}{3}}t\right)}=R_0\sqrt[3]{\frac{8\pi G\rho_0}{\Lambda c^4}\left[\cosh\left(\frac{3}{2}\sqrt{\frac{c^2\Lambda}{3}}t\right)-1\right]} \tag{12.5.728}$$

Plotting such a function with Maple™ 4.00b we get:

```
>plot((cosh(x)-1)^(1/3),x=0..10);
```

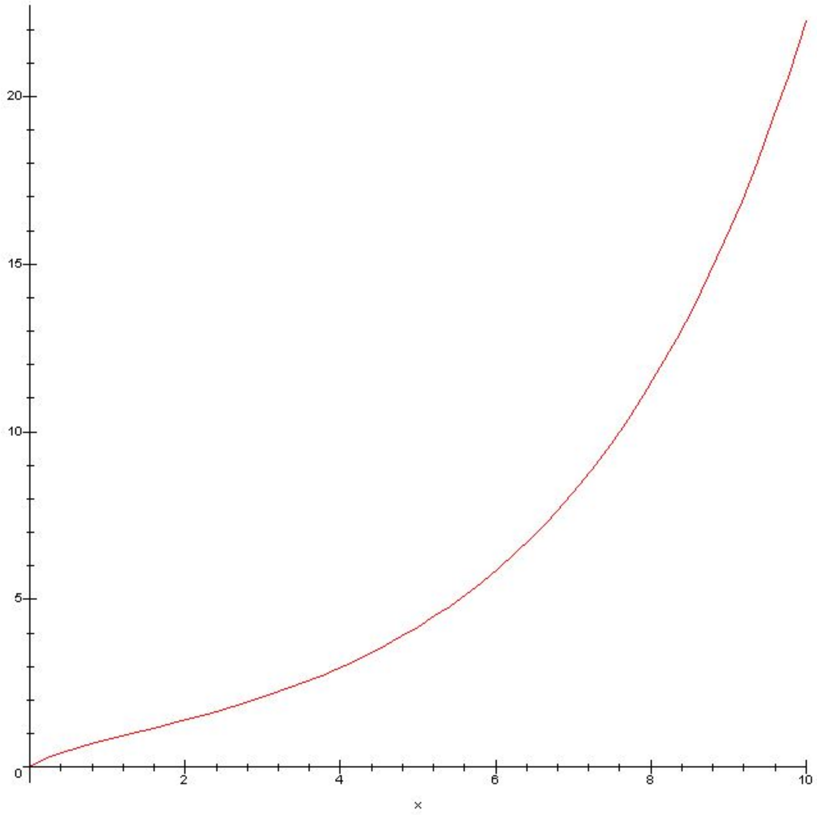


Figure 12.296: **Evolution of R for a zero curvature space with positive cosmological constant**

the quick divergent part of that curve is commonly named "big chill" or "big rip" depending on the textbooks (i.e. authors and teachers).

Looking at the plot it is quite obvious that there is somewhere and inflection point (see section Differential and Integral Calculus page 1454). Let us determine its position using again the Friedmann equation with $k=0$ (and still assuming $\Lambda>0$):

$$\dot{R}^2=\frac{8\pi G}{3c^2}\rho R^2+\frac{c^2\Lambda}{3}R^2 \tag{12.5.729}$$

That as we know, in a matter dominate Universe can be written as:

$$\dot{R}^2=\frac{8\pi G}{3c^2}\frac{\rho_0 R_0^3}{R}+\frac{c^2\Lambda}{3}R^2 \tag{12.5.730}$$

and if we differentiate with respect to time, we get:

$$2\dot{R}\ddot{R}=-\frac{8\pi G}{3c^2}\frac{\rho_0 R_0^3}{R^2}\dot{R}+2\frac{c^2\Lambda}{3}R\dot{R} \tag{12.5.731}$$

The inflection point occurs as we know when $\ddot{R}=0$, therefore we get by dividing both sides by $\dot{R}$ and after rearranging :

$$R=\sqrt[3]{\frac{3}{2}\frac{8\pi G\rho_0}{3\Lambda c^4}R_0^3}=R_0\sqrt[3]{\frac{3}{2}\frac{8\pi G\rho_0}{3\Lambda c^4}}=R_0\sqrt[3]{\frac{1}{2}\frac{8\pi G\rho_0}{\Lambda c^4}} \tag{12.5.732}$$

Equating with the previous result that was for recall:

$$R=R_0\sqrt[3]{\frac{8\pi G\rho_0}{\Lambda c^4}\sinh^2\left(\frac{3}{2}\sqrt{\frac{c^2\Lambda}{3}}t\right)} \tag{12.5.733}$$

We then have:

$$R_0\sqrt[3]{\frac{1}{2}\frac{8\pi G\rho_0}{\Lambda c^4}}=R_0\sqrt[3]{\frac{8\pi G\rho_0}{\Lambda c^4}\sinh^2\left(\frac{3}{2}\sqrt{\frac{c^2\Lambda}{3}}t\right)} \tag{12.5.734}$$

After simplification we get:

$$t_{\text{inflection}}=\frac{2}{3}\sqrt{\frac{3}{c^2\Lambda}}\text{arcsinh}\left(\sqrt{\frac{1}{2}}\right)=\frac{2}{3}\sqrt{\frac{3}{c^2\Lambda}}\text{arcsinh}\left(\frac{\sqrt{2}}{2}\right) \tag{12.5.735}$$

Now let us come back on:

$$R=R_0\sqrt[3]{\frac{8\pi G\rho_0}{\Lambda c^4}\sinh^2\left(\frac{3}{2}\sqrt{\frac{c^2\Lambda}{3}}t\right)} \tag{12.5.736}$$

It is common to rewrite it as:

$$R=R_0\left(\frac{8\pi G\rho_0}{\Lambda c^4}\right)^{1/3}\sinh^{3/2}\left(\frac{t}{t_\Gamma}\right) \tag{12.5.737}$$

It is common as we already know to choose $R_0=1$, therefore:

$$R=\left(\frac{8\pi G\rho_0}{\Lambda c^4}\right)^{1/3}\sinh^{3/2}\left(\frac{t}{t_\Gamma}\right) \tag{12.5.738}$$

That it is common to write:

$$\boxed{a(t)=\left(\frac{\Omega_m}{\Omega_\Gamma}\right)^{1/3}\sinh^{3/2}\left(\frac{t}{t_\Gamma}\right)} \tag{12.5.739}$$

that is the famous "concordance model".

- Now for $k=+1$, the second sub-case (spherical).

  Let us first recall that we have already treated the case $k=+1$ with $\Lambda>0$ and assuming $R=c^{te}$. This is the Einstein (static) Universe model already treated on page 6012.

  Let us consider now the general case:

  $$\dot{R}^2=\frac{8\pi G}{3c^2}\rho R^2-kc^2+\frac{c^2\Lambda}{3}R^2 \tag{12.5.740}$$

  This differential equation can be rewritten as:

  $$\dot{R}^2-\left(\frac{c^2\Lambda}{3}+\frac{8\pi G}{3c^2}\rho\right)R^2+kc^2=0 \tag{12.5.741}$$

  To simplify let us put:

  $$\alpha:=-\left(\frac{c^2\Lambda}{3}+\frac{8\pi G}{3c^2}\rho\right) \tag{12.5.742}$$

  Therefore:

  $$\dot{R}^2+\alpha R^2+kc^2=0 \tag{12.5.743}$$

  Let us now move in natural units to get rid of the c^2:

  $$\dot{R}^2+\alpha R^2+k=0 \tag{12.5.744}$$

  Let us put $k=+1$ and assume that the density ρ is not dependent of time (it's a first case - equivalent to consider as we already know the universe being dominated by vacuum) and keep in mind that's equivalent also to consider ρ very small like $\rho\cong 0$ as 0 is a constant independent of time...):

  $$\dot{R}^2+\alpha R^2+1=0 \tag{12.5.745}$$

  Therefore:

  $$\dot{R}=\pm\sqrt{-1-\alpha^2R^2} \tag{12.5.746}$$

  A simple integration leads to:

  $$\int\frac{\mathrm{d}R}{\sqrt{-1-\alpha^2R^2}}=\pm\int\mathrm{d}t \tag{12.5.747}$$

  For $\alpha\neq 0$ (we would need to treat the case $a=0$ and the case $-1-\alpha^2R^2=0\Longrightarrow R=\pm 1/\alpha$), we substitute $u=\alpha R$:

  $$\int\frac{\mathrm{d}u}{\sqrt{-1-u^2}}=\pm\alpha\int\mathrm{d}t=\pm\alpha t+c^{te} \tag{12.5.748}$$

  Hence:

  $$\int\frac{\mathrm{i}\mathrm{d}u}{\sqrt{1+u^2}}=\pm\alpha t+c^{te} \tag{12.5.749}$$

And we have already proven this primitive in the section of Differential and Integral Calculus (see page 1534):

$$\ln\left(u+\sqrt{u^2+\alpha^2}\right)=\pm\alpha t+c^{te} \tag{12.5.750}$$

Hence:

$$\mathrm{i}\ln\left(\alpha R+\sqrt{\alpha^2R^2+\alpha^2}\right)=\pm\alpha t+c^{te} \tag{12.5.751}$$

Hence:

$$\mathrm{i}\ln\left(\alpha R+\sqrt{\alpha^2R^2+\alpha^2}\right)=\pm\alpha t+c^{te} \tag{12.5.752}$$

Or:

$$\mathrm{i}\cdot\operatorname{arcsinh}(\alpha R)=\pm\alpha t+c^{te} \tag{12.5.753}$$

So finally:

$$R(t)=\sinh\left(\frac{\mathrm{i}(\pm\alpha t+c^{te})}{\alpha}\right) \tag{12.5.754}$$

This is a pure complex hyperbolic sine function. The interpretation is in our point of view non-trivial so will not try to extrapolate, even with a plot, this result. And remember that we assume that:

$$\alpha\neq 0\Rightarrow\frac{c^2\Lambda}{3}\neq-\frac{8\pi G}{3c^2}\rho \tag{12.5.755}$$

Or said otherwise:

$$\rho\neq-\frac{c^4\Lambda}{8\pi G} \tag{12.5.756}$$

Ok so that's the result for a spherical universe dominated by vacuum. Assuming that only ρ is a function of R only and that is does not explicitly depend on time, we can write:

$$\dot{R}=\frac{\mathrm{d}R}{\mathrm{d}t}=\sqrt{\frac{8\pi G}{3c^2}\rho(R)R^2-kc^2+\frac{c^2\Lambda}{3}R^2} \tag{12.5.757}$$

But we have proven (see page 5912) that the different most common type of density evolution are:

$$\rho(t)\propto R(t)^{-3(1+\omega(T))}\propto\begin{cases}R(t)^{-3} & \text{for } \omega(T)=0 \text{ (pressureless matter (i.e. dust))}\\ R(t)^{-4} & \text{for } \omega(T)=1/3 \text{ (radiation/relativistic matter)}\\ c^{te} & \text{for } \omega(T)=-1 \text{ (vacuum energy)}\end{cases} \tag{12.5.758}$$

Let us consider that we are interested about the first instants of the Big Bang (this is an excuse in reality because we didn't found any analytic solution for the presureless matter, i.e. $\omega(T)=0$, case.........). Then we will focus on radiation/relativistic matter, i.e. $\rho\propto R(t)^{-4}$ (and for simplicity we put $R_0=1$):

$$\dot{R}=\sqrt{\frac{8\pi G}{3c^2}\rho_0\frac{R_0^4}{R^4}R^2-kc^2+\frac{c^2\Lambda}{3}R^2}=\sqrt{\frac{8\pi G}{3c^2}\rho_0R^{-2}-kc^2+\frac{c^2\Lambda}{3}R^2} \tag{12.5.759}$$

To simplify let us denote the latter:

$$\dot{R}=\sqrt{\alpha R^{-2}-\beta+\gamma R^2} \tag{12.5.760}$$

Hence:

$$\int \mathrm{d}t=\int \frac{1}{\sqrt{\alpha R^{-2}-\beta+\gamma R^2}}\mathrm{d}R \tag{12.5.761}$$

Let us put first the right hand side at the same denominator:

$$\int \frac{R}{\sqrt{\gamma R^4-\beta R^2+\alpha}}\mathrm{d}R \tag{12.5.762}$$

Let us substitute $u=R^2 \longrightarrow \mathrm{d}u=2R\mathrm{d}R$, therefore:

$$\int \frac{R}{\sqrt{\gamma R^4-\beta R^2+\alpha}}\mathrm{d}R=\frac{1}{2}\int \frac{1}{\sqrt{\gamma u^2-\beta u+\alpha}}\mathrm{d}u=\int \frac{1}{\sqrt{\left(\sqrt{\gamma}u-\frac{\beta}{2\sqrt{\gamma}}\right)^2-\frac{\beta^2}{4\gamma}+\alpha}}\mathrm{d}u \tag{12.5.763}$$

Let us substitute:

$$v=\frac{2\gamma u-\beta}{2\sqrt{\alpha-\frac{\beta^2}{4\gamma}}\sqrt{\gamma}}\to \mathrm{d}v=\frac{\sqrt{\gamma}}{\sqrt{\alpha-\frac{\beta^2}{4\gamma}}}\mathrm{d}u \tag{12.5.764}$$

Therefore we get:

$$\frac{1}{\sqrt{\gamma}}\int \frac{1}{\sqrt{v^2+1}}\,\mathrm{d}v \tag{12.5.765}$$

Now solving:

$$\int \frac{1}{\sqrt{v^2+1}}\,\mathrm{d}v \tag{12.5.766}$$

This is a standard integral (see section Differential and Integral Calculus page 1534):

$$\int \frac{1}{\sqrt{v^2+1}}\mathrm{d}v=\ln\left(\sqrt{v^2+1}+v\right) \tag{12.5.767}$$

Plug in solved integrals:

$$\begin{aligned}\frac{1}{\sqrt{\gamma}}\int\frac{1}{\sqrt{v^2+1}}\mathrm{d}v=\frac{\ln\left(\sqrt{v^2+1}+v\right)}{\sqrt{\gamma}}&=\frac{\ln\left(\sqrt{\frac{(2\gamma u-\beta)^2}{4\gamma\left(\alpha-\frac{\beta^2}{4\gamma}\right)}+1}+\frac{2\gamma u-\beta}{2\sqrt{\alpha-\frac{\beta^2}{4\gamma}}\sqrt{\gamma}}\right)}{\sqrt{\gamma}}\\
&=\frac{\ln\left(\sqrt{\frac{(2\gamma R^2-\beta)^2}{4\gamma\left(\alpha-\frac{\beta^2}{4\gamma}\right)}+1}+\frac{2\gamma R^2-\beta}{2\sqrt{\alpha-\frac{\beta^2}{4\gamma}}\sqrt{\gamma}}\right)}{2\sqrt{\gamma}}\\
&=\frac{\ln\left(\sqrt{\frac{(2\gamma R^2-\beta)^2+4\gamma\left(\alpha-\frac{\beta^2}{4\gamma}\right)}{4\gamma\left(\alpha-\frac{\beta^2}{4\gamma}\right)}}+\frac{2\gamma R^2-\beta}{2\sqrt{\alpha-\frac{\beta^2}{4\gamma}}\sqrt{\gamma}}\right)}{2\sqrt{\gamma}}\\
&=\frac{\ln\left(\frac{1}{2\sqrt{\alpha-\frac{\beta^2}{4\gamma}}\sqrt{\gamma}}\left[\sqrt{(2\gamma R^2-\beta)^2+4\gamma\left(\alpha-\frac{\beta^2}{4\gamma}\right)}+2\gamma R^2-\beta\right]\right)}{2\sqrt{\gamma}}\\
&=\frac{\ln\left(\sqrt{(2\gamma R^2-\beta)^2+4\gamma\left(\alpha-\frac{\beta^2}{4\gamma}\right)}+2\gamma R^2-\beta\right)}{2\sqrt{\gamma}}-\frac{\ln\left(2\sqrt{\gamma}\sqrt{\alpha-\frac{\beta^2}{4\gamma}}\right)}{2\sqrt{\gamma}}\\
&=\frac{\ln\left(\sqrt{(2\gamma R^2-\beta)^2+\alpha 4\gamma-\beta^2}+2\gamma R^2-\beta\right)}{2\sqrt{\gamma}}+c^{te}\\
&=\frac{\ln\left(2\left(\sqrt{\gamma}\sqrt{R^2\left(\gamma R^2-\beta\right)+\alpha}+\gamma R^2\right)-\beta\right)}{2\sqrt{\gamma}}+c^{te}\end{aligned} \tag{12.5.768}$$

Hence:

$$\int\frac{1}{\sqrt{\frac{\alpha}{R^2}-\beta+\gamma R^2}}\mathrm{d}R=\frac{\ln\left(2\left(\sqrt{\gamma}\sqrt{R^2\left(\gamma R^2-\beta\right)+\alpha}+\gamma R^2\right)-\beta\right)}{2\sqrt{\gamma}}+c^{te} \tag{12.5.769}$$

Reinjecting the original constants we get:

$$\int\frac{1}{\sqrt{\frac{8\pi G}{3c^2}\rho_0R^{-2}-kc^2+\frac{c^2\Lambda}{3}R^2}}\mathrm{d}R=\frac{\ln\left(2\left(\sqrt{\frac{c^2\Lambda}{3}}\sqrt{R^2\cdot\left(\frac{c^2\Lambda}{3}R^2-kc^2\right)+\frac{8\pi G}{3c^2}\rho_0}+\frac{c^2\Lambda}{3}R^2\right)-kc^2\right)}{2\sqrt{\frac{c^2\Lambda}{3}}}+c^{te} \tag{12.5.770}$$

So finally for a spherical universe (replacing $k=+1$) dominated by radiation, we have:

$$t=\frac{\ln\left(2\left(\sqrt{\frac{c^2\Lambda}{3}}\sqrt{R^2\cdot\left(\frac{c^2\Lambda}{3}R^2-c^2\right)+\frac{8\pi G}{3c^2}\rho_0}+\frac{c^2\Lambda}{3}R^2\right)-c^2\right)}{2\sqrt{\frac{c^2\Lambda}{3}}}+c^{te} \tag{12.5.771}$$

Sadly we didn't found yet a way to extract the R above as a function of t nor a pertinent way to do an implicit plot with a computer software. So we will have to stop here!

- Let us now consider the case $k=-1$ (hyperbolic).

  For this, same as before, we start from:

  $$\dot{R}^2=\frac{8\pi G}{3c^2}\rho R^2-kc^2+\frac{c^2\Lambda}{3}R^2 \tag{12.5.772}$$

  This differential equation can be rewritten as:

  $$\dot{R}^2-\left(\frac{c^2\Lambda}{3}+\frac{8\pi G}{3c^2}\rho\right)R^2+kc^2=0 \tag{12.5.773}$$

  To simplify let us put:

  $$\alpha:=\left(\frac{c^2\Lambda}{3}+\frac{8\pi G}{3c^2}\rho\right) \tag{12.5.774}$$

  Therefore:

  $$\dot{R}^2+\alpha R^2+kc^2=0 \tag{12.5.775}$$

  Let us now move in natural units to get rid of the c^2:

  $$\dot{R}^2+\alpha R^2+k=0 \tag{12.5.776}$$

  Let us put $k=-1$ and assume that the density ρ is not dependent of time (it's a first case - equivalent to consider as we already know the universe being dominated by vacuum) and keep in mind that's equivalent also to consider ρ very small like $\rho\cong 0$ as 0 is a constant independent of time...):

  $$\dot{R}^2+\alpha R^2-1=0 \tag{12.5.777}$$

  Therefore:

  $$\dot{R}=\pm\sqrt{1-\alpha^2R^2} \tag{12.5.778}$$

  A simple integration leads to:

  $$\int\frac{\mathrm{d}R}{\sqrt{1-\alpha^2R^2}}=\pm\int\mathrm{d}t \tag{12.5.779}$$

  For $\alpha\neq 0$ (we would need to treat the case $\alpha=0$ and the case $1-\alpha^2R^2=0\Longrightarrow R=\pm 1/\alpha$), we substitute $u=\alpha R$:

  $$\int\frac{\mathrm{d}u}{\sqrt{1-u^2}}=\pm\alpha\int\mathrm{d}t=\pm\alpha t+c^{te} \tag{12.5.780}$$

  Let us substitute $u=\sin(x)$, then:

  $$\int\frac{\cos(x)\mathrm{d}x}{\sqrt{1-\sin^2(x)}}=\pm\alpha t+c^{te} \tag{12.5.781}$$

  After simplification using trigonometric identities:

  $$\int\mathrm{d}x=\pm\alpha t+c^{te} \tag{12.5.782}$$

Therefore:

$$x(t)=\pm\alpha t+c^{te} \tag{12.5.783}$$

Hence:

$$u(t)=\sin\left(\pm\alpha t+c^{te}\right) \tag{12.5.784}$$

So finally under the assumptions taken earlier:

$$R(t)=\frac{1}{\alpha}\sin\left(\pm\alpha t+c^{te}\right) \tag{12.5.785}$$

Explicitly (in natural units $c=1, G=1$):

$$R(t)=-\frac{3}{\Lambda+8\pi\rho}\sin\left(\pm\frac{\Lambda+8\pi\rho}{3}t+c^{te}\right) \tag{12.5.786}$$

Therefore we have a cyclic universe (whatever the sign of Λ) and remember that we assume that:

$$\alpha\neq 0\Rightarrow\frac{c^2\Lambda}{3}\neq-\frac{8\pi G}{3c^2}\rho \tag{12.5.787}$$

Or said otherwise:

$$\rho\neq-\frac{c^4\Lambda}{8\pi G} \tag{12.5.788}$$

Ok so that's the result for a hyperbolic universe dominated by vacuum. Assuming that only ρ is a function of R only and that is does not explicitly depend on time, we can write:

$$\dot{R}=\frac{\mathrm{d}R}{\mathrm{d}t}=\sqrt{\frac{8\pi G}{3c^2}\rho(R)R^2-kc^2+\frac{c^2\Lambda}{3}R^2} \tag{12.5.789}$$

But we have proven (see page 5912) that the different most common type of density evolution are:

$$\rho(t)\propto R(t)^{-3(1+\omega(T))}\propto\begin{cases}R(t)^{-3} & \text{for } \omega(T)=0 \text{ (pressureless matter (i.e. dust))}\\ R(t)^{-4} & \text{for } \omega(T)=1/3 \text{ (radiation/relativistic matter)}\\ c^{te} & \text{for } \omega(T)=-1 \text{ (vacuum energy)}\end{cases} \tag{12.5.790}$$

Let us consider that we are interested about the first instants of the Big Bang (this is an excuse in reality because we didn't found any analytic solution for the presureless matter, i.e. $\omega(T)=0$, case.........). Then we will focus on radiation/relativistic matter, i.e. $\rho\propto R(t)^{-4}$ (and for simplicity we put $R_0=1$):

$$\dot{R}=\sqrt{\frac{8\pi G}{3c^2}\rho_0\frac{R_0^4}{R^4}R^2-kc^2+\frac{c^2\Lambda}{3}R^2}=\sqrt{\frac{8\pi G}{3c^2}\rho_0R^{-2}-kc^2+\frac{c^2\Lambda}{3}R^2} \tag{12.5.791}$$

To simplify let us denote the latter:

$$\dot{R}=\sqrt{\alpha R^{-2}-\beta+\gamma R^2} \tag{12.5.792}$$

Hence:

$$\int \mathrm{d}t=\int \frac{1}{\sqrt{\alpha R^{-2}-\beta+\gamma R^2}}\mathrm{d}R \tag{12.5.793}$$

Let us put first the right hand side at the same denominator:

$$\int \frac{R}{\sqrt{\gamma R^4-\beta R^2+\alpha}}\mathrm{d}R \tag{12.5.794}$$

Let us substitute $u=R^2 \longrightarrow \mathrm{d}u=2R\mathrm{d}R$, therefore:

$$\int \frac{R}{\sqrt{\gamma R^4-\beta R^2+\alpha}}\mathrm{d}R=\frac{1}{2}\int \frac{1}{\sqrt{\gamma u^2-\beta u+\alpha}}\mathrm{d}u=\int \frac{1}{\sqrt{\left(\sqrt{\gamma}u-\frac{\beta}{2\sqrt{\gamma}}\right)^2-\frac{\beta^2}{4\gamma}+\alpha}}\mathrm{d}u \tag{12.5.795}$$

Let us substitute:

$$v=\frac{2\gamma u-\beta}{2\sqrt{\alpha-\frac{\beta^2}{4\gamma}}\sqrt{\gamma}} \rightarrow \mathrm{d}v=\frac{\sqrt{\gamma}}{\sqrt{\alpha-\frac{\beta^2}{4\gamma}}}\mathrm{d}u \tag{12.5.796}$$

Therefore we get:

$$\frac{1}{\sqrt{\gamma}}\int \frac{1}{\sqrt{v^2+1}}\,\mathrm{d}v \tag{12.5.797}$$

Now solving:

$$\int \frac{1}{\sqrt{v^2+1}}\,\mathrm{d}v \tag{12.5.798}$$

This is a standard integral (see section Differential and Integral Calculus page 1534):

$$\int \frac{1}{\sqrt{v^2+1}}\mathrm{d}v=\ln\left(\sqrt{v^2+1}+v\right) \tag{12.5.799}$$

Plug in solved integrals:

$$
\begin{aligned}
\frac{1}{\sqrt{\gamma}}\int\frac{1}{\sqrt{v^2+1}}\mathrm{d}v=\frac{\ln\left(\sqrt{v^2+1}+v\right)}{\sqrt{\gamma}}&=\frac{\ln\left(\sqrt{\frac{(2\gamma u-\beta)^2}{4\gamma\left(\alpha-\frac{\beta^2}{4\gamma}\right)}+1}+\frac{2\gamma u-\beta}{2\sqrt{\alpha-\frac{\beta^2}{4\gamma}}\sqrt{\gamma}}\right)}{\sqrt{\gamma}}\\
&=\frac{\ln\left(\sqrt{\frac{(2\gamma R^2-\beta)^2}{4\gamma\left(\alpha-\frac{\beta^2}{4\gamma}\right)}+1}+\frac{2\gamma R^2-\beta}{2\sqrt{\alpha-\frac{\beta^2}{4\gamma}}\sqrt{\gamma}}\right)}{2\sqrt{\gamma}}\\
&=\frac{\ln\left(\sqrt{\frac{(2\gamma R^2-\beta)^2+4\gamma\left(\alpha-\frac{\beta^2}{4\gamma}\right)}{4\gamma\left(\alpha-\frac{\beta^2}{4\gamma}\right)}}+\frac{2\gamma R^2-\beta}{2\sqrt{\alpha-\frac{\beta^2}{4\gamma}}\sqrt{\gamma}}\right)}{2\sqrt{\gamma}}\\
&=\frac{\ln\left(\frac{1}{2\sqrt{\alpha-\frac{\beta^2}{4\gamma}}\sqrt{\gamma}}\left[\sqrt{(2\gamma R^2-\beta)^2+4\gamma\left(\alpha-\frac{\beta^2}{4\gamma}\right)}+2\gamma R^2-\beta\right]\right)}{2\sqrt{\gamma}}\\
&=\frac{\ln\left(\sqrt{(2\gamma R^2-\beta)^2+4\gamma\left(\alpha-\frac{\beta^2}{4\gamma}\right)}+2\gamma R^2-\beta\right)}{2\sqrt{\gamma}}-\frac{\ln\left(2\sqrt{\gamma}\sqrt{\alpha-\frac{\beta^2}{4\gamma}}\right)}{2\sqrt{\gamma}}\\
&=\frac{\ln\left(\sqrt{(2\gamma R^2-\beta)^2+\alpha4\gamma-\beta^2}+2\gamma R^2-\beta\right)}{2\sqrt{\gamma}}+c^{te}\\
&=\frac{\ln\left(2\left(\sqrt{\gamma}\sqrt{R^2\left(\gamma R^2-\beta\right)+\alpha}+\gamma R^2\right)-\beta\right)}{2\sqrt{\gamma}}+c^{te}
\end{aligned}
\tag{12.5.800}
$$

Hence:

$$\int\frac{1}{\sqrt{\frac{\alpha}{R^2}-\beta+\gamma R^2}}\mathrm{d}R=\frac{\ln\left(2\left(\sqrt{\gamma}\sqrt{R^2\left(\gamma R^2-\beta\right)+\alpha}+\gamma R^2\right)-\beta\right)}{2\sqrt{\gamma}}+c^{te}\tag{12.5.801}$$

Reinjecting the original constants we get:

$$\int\frac{1}{\sqrt{\frac{8\pi G}{3c^2}\rho_0R^{-2}-kc^2+\frac{c^2\Lambda}{3}R^2}}\mathrm{d}R=\frac{\ln\left(2\left(\sqrt{\frac{c^2\Lambda}{3}}\sqrt{R^2\cdot\left(\frac{c^2\Lambda}{3}R^2-kc^2\right)+\frac{8\pi G}{3c^2}\rho_0}+\frac{c^2\Lambda}{3}R^2\right)-kc^2\right)}{2\sqrt{\frac{c^2\Lambda}{3}}}+c^{te}\tag{12.5.802}$$

So finally for a hyperbolic universe (replacing $k=-1$) dominated by radiation, we have:

$$t=\frac{\ln\left(2\left(\sqrt{\frac{c^2\Lambda}{3}}\sqrt{R^2\cdot\left(\frac{c^2\Lambda}{3}R^2+c^2\right)+\frac{8\pi G}{3c^2}\rho_0}+\frac{c^2\Lambda}{3}R^2\right)+c^2\right)}{2\sqrt{\frac{c^2\Lambda}{3}}}+c^{te}\tag{12.5.803}$$

Sadly we didn't found yet a way to extract the R above as a function of t nor a pertinent way to do an implicit plot with a computer software. So we will have to stop here!

To summarize for these nine common models (for reminder among 27 if for each we include the case dominated by matter, by radiation or by vacuum):

Cosmological Constant	**Curvature**	**Radius (scale factor)**
$\Lambda=0$	$k=0$ (flat)	$\left(\dfrac{9H^2}{4}\right)^{1/3} t^{2/3}$
	$k=+1$ (sphere)	$t=\dfrac{8\pi G\rho_0}{3c^5}\dfrac{1}{2}\left(\theta'-\sin\left(\theta'\right)\right)$ $R=\dfrac{8\pi G\rho_0}{3c^4}\dfrac{1}{2}\left(1-\cos\left(\theta'\right)\right)$
	$k=-1$ (hyperbolic)	$t=\dfrac{8\pi G\rho_0}{3c^5}\dfrac{1}{2}\left(\sinh\left(\theta'\right)-\theta'\right)$ $R=\dfrac{8\pi G\rho_0}{3c^4}\dfrac{1}{2}\left(\cosh\left(\theta'\right)-1\right)$
$\Lambda<0$	$k=0$ (flat)	$R=R_0\sqrt[3]{-\frac{8\pi G\rho_0}{\lvert\Lambda\rvert c^4}\sinh^2\left(\frac{3}{2}\sqrt{\frac{-c^2\lvert\Lambda\rvert}{3}}t\right)}$
	$k=+1$ (sphere)	No pretty solution known to us
	$k=-1$ (hyperbolic)	No pretty solution known to us
$\Lambda>0$	$k=0$ (flat)	$R=R_0\sqrt[3]{\frac{8\pi G\rho_0}{\Lambda c^4}\sinh^2\left(\frac{3}{2}\sqrt{\frac{c^2\Lambda}{3}}t\right)}$
	$k=+1$ (sphere)	No pretty solution known to us
	$k=-1$ (hyperbolic)	No pretty solution known to us

Table 12.28: **Friedmann-Lemaître-Robertson-Walker cosmological models equations summary**

Caution! Keep in mind that we have the knowledge now with everything that we have seen so far of the following canonical curvature values $k=\{-1,0,+1\}$, of the following typical canonical values of the density parameter $\omega(T)=\{0,1/3,-1/3,1,<-1\}$ and the following type of canonical values for the cosmological constant $\omega(T)=\{0<,=0,>0\}$. Hence in reality there are $3\times5\times3=45$ different type of dominated universes! If we take into account that a universe can contain a mixed of radiation, vacuum, matter, etc. we get in reality $3\times31\times5=465$ possible mixtures of universes....!!! Obviously we will never study all of them in this textbook!

Notice that whatever which result you pick above, if you inject into these relations the value of the Planck density or of the Planck time given for recall by:

$$\rho_\text{P}=\frac{c^5}{G^2\hbar}\cong 5.16\cdot10^{96}\,[\text{kg}\cdot\text{m}^{-3}]\qquad t_\text{P}=\sqrt{\frac{\hbar G}{c^5}}\cong 5.391\cdot10^{-44}\,[\text{s}] \tag{12.5.804}$$

there are absolutely no singularity at the opposite of a common urban legend! So when you read that «*General Relativity fails at Planck scale*» it's a supposition but not an affirmation!

Now remember that we have proven earlier above (see page 6000) that the proper distance to the horizon given by:

$$s_\text{hor.}(t)=a(t)\int\limits_0^t\frac{c\text{d}t}{a(t)} \tag{12.5.805}$$

$$R_{\Lambda=0,k=0}\cong\left(\frac{6\pi G}{c^2}\rho_0\right)^{1/3}t^{2/3}=\left(\frac{9H^2}{4}\right)^{1/3}t^{2/3}:=a_{\Lambda=0,k=0}(t) \tag{12.5.806}$$

Then we have:

$$\begin{aligned}s_{\text{hor.},\Lambda=0,k=0}(t)&=a(t)\int\limits_0^t\frac{c\mathrm{d}t}{a(t)}\cong\left(\frac{9H^2}{4}\right)^{1/3}t^{2/3}\int\limits_0^t\frac{c\mathrm{d}t}{\left(\frac{9H^2}{4}\right)^{1/3}t^{2/3}}=t^{2/3}\int\limits_0^t\frac{c\mathrm{d}t}{t^{2/3}}\\&=t^{2/3}\cdot c3t^{1/3}=3ct\end{aligned} \tag{12.5.807}$$

Hence we have the quite famous result corresponding to our actual state of experimental evidence (last measurements from 12020 give $t\cdot c=13.77\pm 0.040$ billion years):

$$\boxed{R_{\Lambda=0,k=0}\cong 3ct\cong 41.31\pm 0.12\,[\text{billion light years}]} \tag{12.5.808}$$

So we finally found the value that we can see or listen on many Internet websites or in YouTube videos but without proof! This value can be compared to the 28.8 billion light years of the Newtonian radius approach...

Finally here is a summary of the models developed earlier above depending on the canonical values of k and Λ :

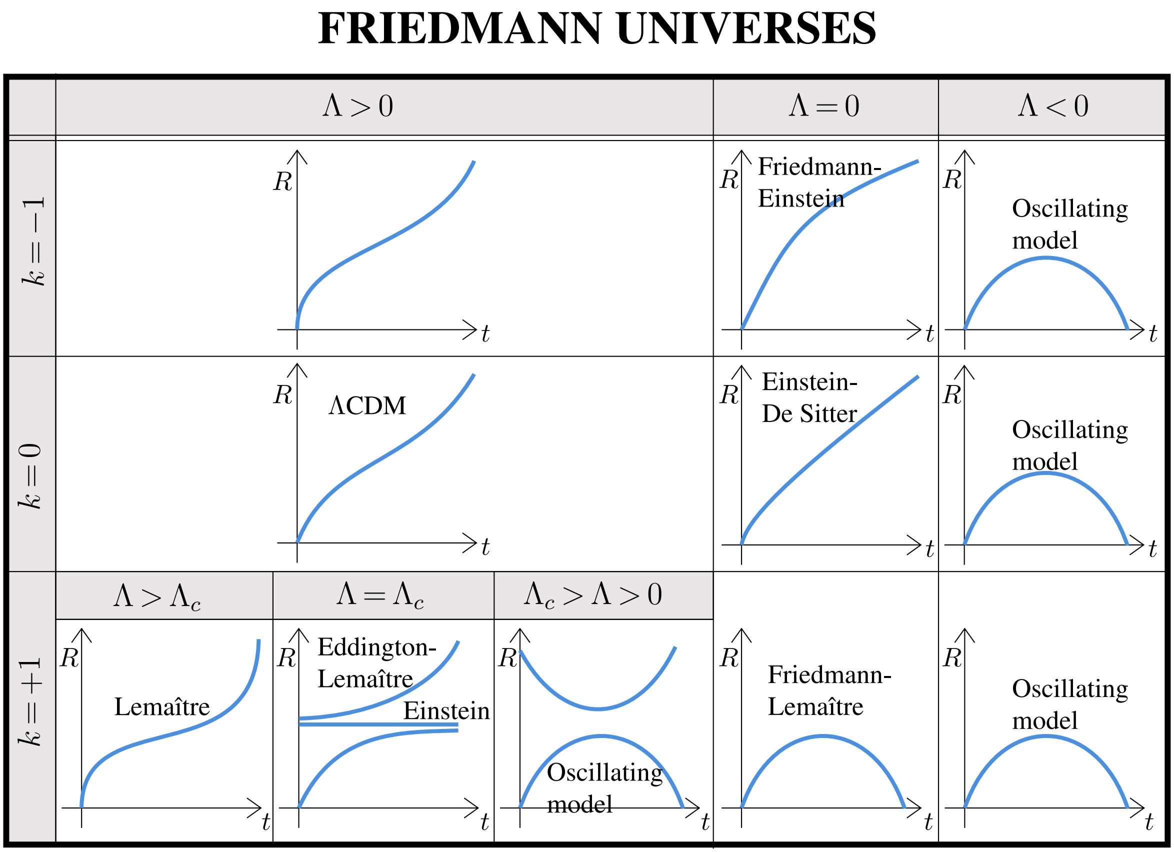

FRIEDMANN UNIVERSES
$\Lambda>0$
$\Lambda=0$
$\Lambda<0$
$k=-1$
$k=0$
$k=+1$
$\Lambda>\Lambda_c$
$\Lambda=\Lambda_c$
$\Lambda_c>\Lambda>0$
R
t
Friedmann-Einstein
Oscillating model
ΛCDM
Einstein-De Sitter
Lemaître
Eddington-Lemaître
Einstein
Friedmann-Lemaître

Figure 12.297: **Cosmological models summary (source: [147])**

The Uchuu (meaning "Outer Space" in Japanese) 12021 (holocene calendar) simulation is the largest and most detailed simulation of the universe ever made. It contains 2.1 trillion "particles" in a space 9.6 billion light-years across. The simulation models the evolution of the universe across more than 13 billion years. It doesn't focus on the formation of stars and planets but instead looks at the behaviour of dark matter within an expanding universe. The detail of Uchuu is high enough that the team can identify everything from galaxy clusters to the dark matter halos of individual galaxies. Since dark matter[70] makes up most of the matter in the universe, it is the main driver of galaxy formation and clustering. It takes a tremendous amount of computational power and storage to create such a detailed model. The team (from Japan, Spain, U.S.A., Argentina, Australia, Chile, France, and Italy) used over $40,000$ computer cores and 20 million computer hours to generate their simulation, and it produced more than 3 Petabytes of data. Using high-density compression, however, the team was able to compress their results into a mere 100 Terabytes of storage that can be downloaded and analysed with Python™.

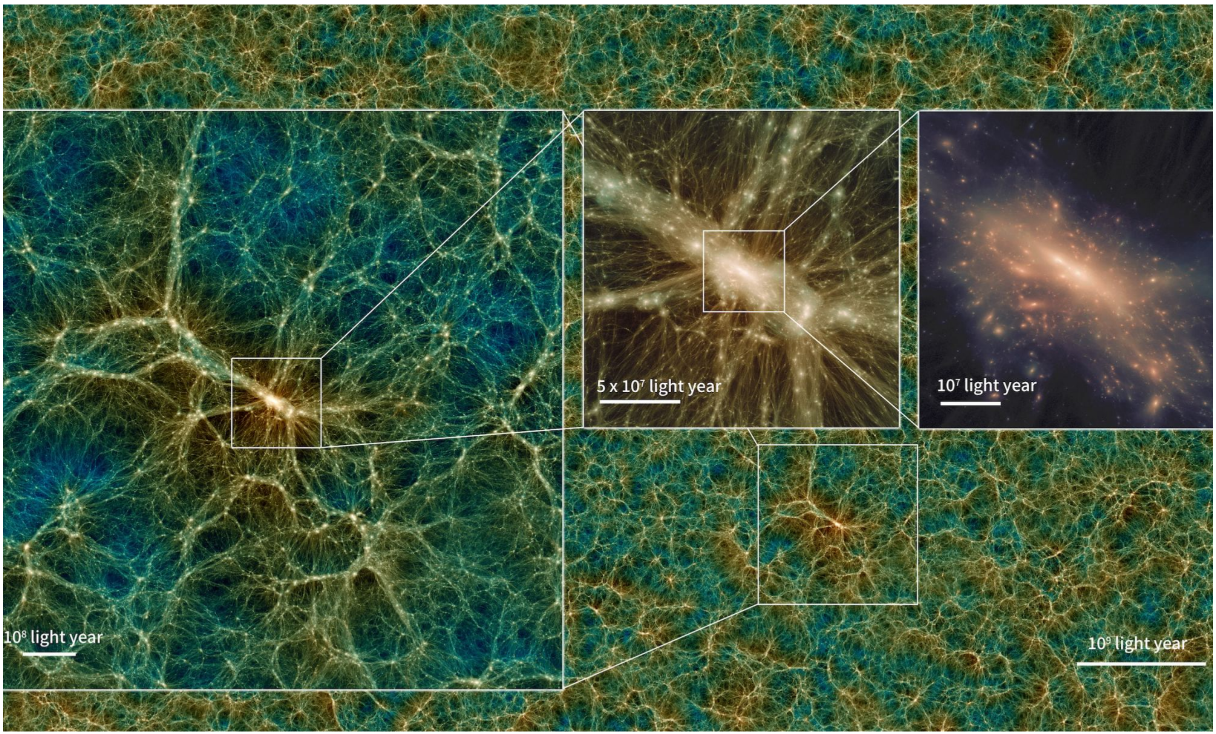


Figure 12.298: **Uchuu 21021 Universe simulation (source: [277])**

---

[70] Since June 12025 (holocene calendar) it is known thanks to FRB (Fast Radio Bursts) that approximately 76% of the Universe's baryonic matter lies in the IGM (InterGalactic Medium). About 15% resides in galaxy halos, and a small fraction is burrowed in stars or amid cold galactic galactic gas ([109]).

12.5.7 Friedmann historical derivation

OK so far we have seen how to get different types of universe using the following approaches:

- Newtonian non-relativistic approach
- Friedmann-Lemaître Euclidean models
- Friedmann-Lemaître-Robertson-Walker Cosmological Models

However some people asked us to add the original models made by Friedmann in his famous paper of 11922 [193] (holocene calendar). So let's do it but with much more details and using modern tools!

First, let us put a bit of context! The Russian mathematician and physicist Alexander Friedmann (11888-11925) is well known among relativists, but his contributions to cosmology are largely misunderstood. Even the Royal Swedish Academy of Sciences misrepresented Friedmann's work in the 12011 (holocene calendar) Nobel Prize scientific background essay. Friedmann was the first physicist who demonstrated that Albert Einstein's general relativity admits non-static solutions, and the universe can expand, oscillate, and be born in a singularity (and it's not Georges Lemaître). Friedmann's conclusion was based on his analysis of an elliptic integral as we will show it further below!

First, Friedmann considered the 3D hypersphere:

$$x_1^2+x_2^2+x_3^2+x_4^2=a(t)^2 \tag{12.5.809}$$

After he transformed that into cartesian coordinates:

$$\begin{aligned}
x_1&=a(t)\sin(\chi)\sin(\theta)\cos(\phi)\\
x_2&=a(t)\sin(\chi)\sin(\theta)\sin(\phi)\\
x_3&=a(t)\sin(\chi)\cos(\theta)\\
x_4&=a(t)\cos(\chi)
\end{aligned} \tag{12.5.810}$$

We can check that's indeed compliant:

$$\begin{aligned}
\sum_i x_i^2&=a(t)^2\left(\sin^2(\chi)\sin^2(\theta)\cos^2(\phi)+\sin^2(\chi)\sin^2(\theta)\sin^2(\phi)+\sin^2(\chi)\cos^2(\theta)+\cos^2(\chi)\right)\\
&=a(t)^2\left(\sin^2(\theta)\cos^2(\phi)+\sin^2(\theta)\sin^2(\phi)+\cos^2(\theta)\right)_{=1,S^2}\sin^2(\chi)+\cos^2(\chi)\\
&=\sin^2(\chi)+\cos^2(\chi)=1
\end{aligned} \tag{12.5.811}$$

Now to get the metric we compute:

$$\begin{aligned}
&x_1=a(t)\sin(\chi)\sin(\theta)\cos(\phi)\\
&\quad\Rightarrow \mathrm{d}x_1=a(t)\left[\cos(\chi)\mathrm{d}\chi\sin(\theta)\cos(\phi)+\sin(\chi)\cos(\theta)\mathrm{d}\theta\cos(\phi)-\sin(\chi)\sin(\theta)\sin(\phi)\mathrm{d}\phi\right]\\
&x_2=a(t)\sin(\chi)\sin(\theta)\sin(\phi)\\
&\quad\Rightarrow \mathrm{d}x_2=a(t)\left[\cos(\chi)\mathrm{d}\chi\sin(\theta)\sin(\phi)+\sin(\chi)\cos(\theta)\mathrm{d}\theta\sin(\phi)+\sin(\chi)\sin(\theta)\cos(\phi)\mathrm{d}\phi\right]\\
&x_3=a(t)\sin(\chi)\cos(\theta)\\
&\quad\Rightarrow \mathrm{d}x_3=a(t)\left[\cos(\chi)\mathrm{d}\chi\cos(\theta)-\sin(\chi)\sin(\theta)\mathrm{d}\theta\right]\\
&x_4=a(t)\cos(\chi)\\
&\quad\Rightarrow \mathrm{d}x_4=-a(t)\sin(\chi)\mathrm{d}\chi
\end{aligned}\tag{12.5.812}$$

After injecting into:

$$\mathrm{d}s^2=-c^2\mathrm{d}t^2+\mathrm{d}x_1^2+\mathrm{d}x_2^2+\mathrm{d}x_3^2+\mathrm{d}x_4^2\tag{12.5.813}$$

by remembering that:

$$\begin{aligned}
(a+b-c)^2&=a^2+b^2+c^2+2(ab-ac-bc)\\
(a+b+c)^2&=a^2+b^2+c^2+2(ab+ac+bc)\\
(a-b)^2&=a^2-2ab-b^2
\end{aligned}\tag{12.5.814}$$

and a very boring simplification work (that we will omit to do here for now...) we get the space-time geometry that is then described by the metric[71]:

$$\mathrm{d}s^2=-c\mathrm{d}t^2+a(t)^2\left(\mathrm{d}\chi^2+\sin(\chi)^2\left(\mathrm{d}\theta^2+\sin(\theta)^2\mathrm{d}\phi^2\right)\right)\tag{12.5.815}$$

After this, Friedman used the Einstein field equation (see page 5766):

$$R_{\mu\nu}-\frac{1}{2}Rg_{\mu\nu}+\Lambda g_{\mu\nu}=\frac{8\pi G}{c^4}T_{\mu\nu}\tag{12.5.816}$$

and he chose to focus only on the time component of the Einstein tensor given in general for reminder by (see section Tensor Calculus page 2107):

$$G_{\mu\nu}=R_{\mu\nu}-\frac{1}{2}g_{\mu\nu}R\tag{12.5.817}$$

With the Christoffel symbol of the first kind given by (see section Tensor Calculus page 2107):

$$\Gamma_{kji}=\frac{1}{2}\left(\partial_k g_{ij}+\partial_i g_{jk}-\partial_j g_{ki}\right)\tag{12.5.818}$$

with the Christoffel symbol of the second time given by (see section Tensor Calculus page 2107):

$$\Gamma_{ki}^{j}=g^{jl}\Gamma_{kli}\tag{12.5.819}$$

With the Ricci tensor given for recall by (see section Tensor Calculus page 2127):

$$R_{\alpha\beta}=R_{\alpha\mu\beta}^{\mu}=R_{i,jk}^{l}=\partial_j\Gamma_{ik}^{l}-\partial_k\Gamma_{ij}^{l}+\Gamma_{ik}^{r}\Gamma_{jr}^{l}-\Gamma_{ij}^{r}\Gamma_{kr}^{l}\tag{12.5.820}$$

---

[71] It can be proven that this metric is special case of the Robertson-Walker metric with $k=+1$ and when one of the term is integrally normalized.

and with the Ricci scalar given by (see section Tensor Calculus page 2128):

$$R=g^{\alpha\beta}R_{\alpha\beta}=g^{\alpha\beta}g^{\mu v}R_{\mu\alpha v\beta} \tag{12.5.821}$$

We have for the metric:

$$g^{ij}=\begin{bmatrix}-a(t)^2 & 0 & 0 & 0\\ 0 & -a(t)^2\sin^2(\chi) & 0 & 0\\ 0 & 0 & -a(t)^2\sin^2(\chi)\sin^2(\theta) & 0\\ 0 & 0 & 0 & c^2\end{bmatrix} \tag{12.5.822}$$

and for anti-metric:

$$g^{ij}=\begin{bmatrix}-\dfrac{1}{a(t)^2} & 0 & 0 & 0\\ 0 & -\dfrac{1}{a(t)^2\sin^2(\chi)} & 0 & 0\\ 0 & 0 & -\dfrac{1}{a(t)^2\sin^2(\chi)\sin^2(\theta)} & 0\\ 0 & 0 & 0 & \dfrac{1}{c^2}\end{bmatrix} \tag{12.5.823}$$

We have for the non-zero components of the Christoffel symbols of the first kind:

$$\begin{aligned}
&\Gamma_{114}=a(t)\left(\frac{\partial}{\partial t}a(t)\right)\\
&\Gamma_{122}=-a(t)^2\sin(\chi)\cos(\chi)\\
&\Gamma_{133}=-a(t)^2\sin(\chi)\sin^2(\theta)\cos(\chi)\\
&\Gamma_{141}=-a(t)\left(\frac{\partial}{\partial t}a(t)\right)\\
&\Gamma_{221}=a(t)^2\sin(x)\cos(x)\\
&\Gamma_{224}=a(t)\sin^2(\chi)\left(\frac{\partial}{\partial t}a(t)\right)\\
&\Gamma_{223}=-a(t)^2\sin^2(\chi)\sin(\theta)\cos(\theta)\\
&\Gamma_{242}=-a(t)\sin^2(\chi)\left(\frac{\partial}{\partial t}a(t)\right)\\
&\Gamma_{331}=a(t)^2\sin(\chi)\sin^2(\theta)\cos(\chi)\\
&\Gamma_{332}=a(t)^2\sin^2(\chi)\sin(\theta)\cos(\theta)\\
&\Gamma_{334}=a(t)\sin^2(\chi)\sin^2(\theta)\left(\frac{\partial}{\partial t}a(t)\right)\\
&\Gamma_{343}=-a(t)\sin^2(\chi)\sin^2(\theta)\left(\frac{\partial}{\partial t}a(t)\right)
\end{aligned} \tag{12.5.824}$$

We have for the non-zero components of the Christoffel symbols of the second kind:

$$\begin{aligned}
\Gamma^{1}_{14}&=\frac{\frac{\partial}{\partial t}a(t)}{a(t)}\\
\Gamma^{1}_{22}&=-\sin(\chi)\cos(\chi)\\
\Gamma^{1}_{33}&=-\sin(\chi)\sin^2(\theta)\cos(\chi)\\
\Gamma^{2}_{12}&=\frac{\cos(\chi)}{\sin(\chi)}\\
\Gamma^{2}_{24}&=\frac{\frac{\partial}{\partial t}a(t)}{a(t)}\\
\Gamma^{2}_{33}&=-\sin(\theta)\cos(\theta)\\
\Gamma^{3}_{13}&=\frac{\cos(\chi)}{\sin(\chi)}\\
\Gamma^{3}_{23}&=\frac{\cos(\theta)}{\sin(\theta)}\\
\Gamma^{3}_{34}&=\frac{\frac{\partial}{\partial t}a(t)}{a(t)}\\
\Gamma^{4}_{11}&=\frac{a(t)\left(\frac{\partial}{\partial t}a(t)\right)}{c^2}\\
\Gamma^{4}_{22}&=\frac{a(t)\sin^2(\chi)\left(\frac{\partial}{\partial t}a(t)\right)}{c^2}\\
\Gamma^{4}_{33}&=\frac{a(t)\sin^2(\chi)\sin^2(\theta)\left(\frac{\partial}{\partial t}a(t)\right)}{c^2}
\end{aligned}\tag{12.5.825}$$

Therefore:

$$\begin{aligned}
R_{11}&=-\frac{2c^2+2\left(\dfrac{\partial}{\partial t}a(t)\right)^2+a(t)\left(\frac{\partial^2}{\partial t^2}a(t)\right)}{c^2}\\
R_{22}&=-\frac{\sin(\chi)^2c^2+2\left(\dfrac{\partial}{\partial t}a(t)\right)^2\sin(\chi)^2+c^2-\cos(\chi)^2c^2+a(t)\sin(\chi)^2\left(\dfrac{\partial^2}{\partial t^2}a(t)\right)}{c^2}\\
R_{33}&=-\frac{\sin(\theta)^2\left(\sin(\chi)^2c^2+2\left(\dfrac{\partial}{\partial t}a(t)\right)^2\sin(\chi)^2+c^2-\cos(\chi)^2c^2+a(t)\sin(\chi)^2\left(\dfrac{\partial^2}{\partial t^2}a(t)\right)\right)}{c^2}\\
R_{44}&=3\frac{\dfrac{\partial^2}{\partial t^2}a(t)}{a(t)}
\end{aligned}\tag{12.5.826}$$

And finally:

$$R=2\frac{2\sin^2(\chi)c^2+3\left(\dfrac{\partial}{\partial t}a(t)\right)^2\sin^2(\chi)+3a(t)\sin^2(\chi)\left(\dfrac{\partial^2}{\partial t^2}a(t)\right)+c^2-\cos^2(\chi)c^2}{a(t)^2\sin(\chi)^2c^2}\tag{12.5.827}$$

Therefore the times component of the Einstein tensor:

$$G_{\mu\nu}=R_{\mu\nu}-\frac{1}{2}g_{\mu\nu}R \tag{12.5.828}$$

is equal to:

$$G_{44}=3\frac{\frac{\partial^2}{\partial t^2}a(t)}{a(t)}-3\frac{\frac{\partial^2}{\partial t^2}a(t)}{a(t)}-\frac{1}{2}\left(2\frac{2\sin^2(\chi)c^2+3\left(\frac{\partial}{\partial t}a(t)\right)^2\sin^2(\chi)+3\mathrm{a}(t)\sin^2(\chi)\left(\frac{\partial^2}{\partial t^2}a(t)\right)+c^2-\cos^2(\chi)c^2}{a(t)^2\sin(\chi)^2c^2}\right)c^2 \tag{12.5.829}$$

If we simplify, this gives:

$$G_{44}=\frac{-3c^2-3\left(\frac{\partial}{\partial t}a(t)\right)^2}{a(t)^2} \tag{12.5.830}$$

Now let us inject this into the Einstein field equation given for reminder by:

$$R_{\mu\nu}-\frac{1}{2}Rg_{\mu\nu}+\Lambda g_{\mu\nu}=\frac{8\pi G}{c^4}T_{\mu\nu} \tag{12.5.831}$$

Becomes:

$$\frac{-3c^2-3\left(\frac{\partial}{\partial t}a(t)\right)^2}{a(t)^2}+\Lambda c^2=\frac{8\pi G}{c^4}T_{44} \tag{12.5.832}$$

And we have already proof that $T_{44}=\rho c^2$ (see page 5754). Therefore:

$$\frac{-3c^2-3\left(\frac{\partial}{\partial t}a(t)\right)^2}{a(t)^2}+\Lambda c^2=\frac{8\pi G}{c^4}\rho c^2 \tag{12.5.833}$$

Let us choose now the notation used by Friedmann:

$$\frac{-3c^2-3\left(\frac{\partial}{\partial t}a(t)\right)^2}{a(t)^2}+\lambda=-\kappa\rho c^2 \tag{12.5.834}$$

If we rearrange:

$$\frac{-3\left(\frac{\mathrm{d}}{\mathrm{d}t}a(t)\right)^2-3c^2}{a(t)^2}+\lambda=-\kappa\rho c^2 \tag{12.5.835}$$

The volume of a 3D hypersphere is obtained to the integral below:

$$\begin{gathered}V=\int\limits_0^{2\pi}\int\limits_0^{\pi}\int\limits_0^{\pi}a(t)^3\sin(\chi)^2\sin(\theta)\mathrm{d}\chi\mathrm{d}\theta\mathrm{d}\phi=a(t)^3\left(-\frac{\sin(2\chi)-2\chi}{4}\right)\Big|_0^{\pi}(-\cos(\theta))\Big|_0^{\pi}(\phi)\Big|_0^{2\pi}\\=a(t)^3\frac{\pi}{2}\cdot 2\cdot 2\pi=2\pi^2a(t)^3\end{gathered} \tag{12.5.836}$$

The total mass is the product of this volume and the density such that:

$$M := 2\rho\pi^2 a(t)^3 \Leftrightarrow \rho = \frac{M}{2\pi^2 a(t)^3} \tag{12.5.837}$$

We inject it into the prior-previous relation:

$$\frac{-3\left(\dfrac{\mathrm{d}}{\mathrm{d}t}a(t)\right)^2 - 3c^2}{a(t)^2} + \lambda = -\kappa\frac{M}{2\pi^2 a(t)^3}c^2 \tag{12.5.838}$$

Let us rearrange a first time:

$$\left(\frac{\mathrm{d}}{\mathrm{d}t}a(t)\right)^2 = \frac{\kappa\dfrac{M}{2\pi^2 a(t)}c^2 + \lambda a(t)^2 - 3c^2}{3} \tag{12.5.839}$$

And rearrange a second time:

$$\frac{\left(\dfrac{\mathrm{d}}{\mathrm{d}t}a(t)\right)^2}{c^2} = \kappa\frac{M}{6\pi^2 a(t)} + \frac{\lambda}{3c^2}a(t)^2 - c^2 \tag{12.5.840}$$

Let us put as Friedmann did:

$$A = \frac{\kappa M}{6\pi^2} \tag{12.5.841}$$

Then:

$$\frac{\left(\dfrac{\mathrm{d}}{\mathrm{d}t}a(t)\right)^2}{c^2} = \frac{A}{a(t)} + \frac{\lambda}{3c^2}a(t)^2 - 1^2 \tag{12.5.842}$$

We simplify:

$$\frac{\left(\dfrac{\mathrm{d}}{\mathrm{d}t}a(t)\right)^2}{c^2} = \frac{A + \dfrac{\lambda a(t)^3}{3c^2} - a(t)}{a(t)} \tag{12.5.843}$$

And rearranging a last time to get exactly the same expression as in the paper of Friedmann:

$$\frac{\left(\dfrac{\mathrm{d}}{\mathrm{d}t}a(t)\right)^2}{c^2} = \frac{A - a(t) + \dfrac{\lambda a(t)^3}{3c^2}}{a(t)} \tag{12.5.844}$$

Let us do another change of variable a Friedmann did (we put $a(t) = x$):

$$\frac{\left(\dfrac{\mathrm{d}x}{\mathrm{d}t}\right)^2}{c^2} = \frac{A - x + \dfrac{\lambda x^3}{3c^2}}{x} \tag{12.5.845}$$

Now we take the square root on both side as did Friedmann:

$$\frac{\mathrm{d}x}{c\mathrm{d}t}=\sqrt{\frac{A-x+\dfrac{\lambda x^3}{3c^2}}{x}} \tag{12.5.846}$$

Now let us invert:

$$c\frac{\mathrm{d}t}{\mathrm{d}x}=\sqrt{\frac{x}{A-x+\dfrac{\lambda x^3}{3c^2}}} \tag{12.5.847}$$

We rearrange a bit:

$$c\mathrm{d}t=\sqrt{\frac{x}{A-x+\dfrac{\lambda x^3}{3c^2}}}\mathrm{d}x \tag{12.5.848}$$

We take the integral on both side so we get:

$$ct=\int\limits_{a_0}^{a}\sqrt{\frac{x}{A-x+\dfrac{\lambda x^3}{3c^2}}}\mathrm{d}x+b \tag{12.5.849}$$

in which a_0 and b are constants. This elliptic integral doesn't have a closed solution as far as we know. So we will use a specialized computer software like Maple™.

So let us enter this integral into Maple™ 4.00b and see what happens. First we put $\lambda/3c^2:=y$:

```
>restart:
>with(plots):
>int(sqrt(x/(A-x+y*x^3)),x):
```

Here we end the Maple™ input with colon to suppress the output, because it is lengthy and not too illuminating. The reader can change the colon to semicolon and see that the result is expressed in terms of EllipticF and EllipticPi.

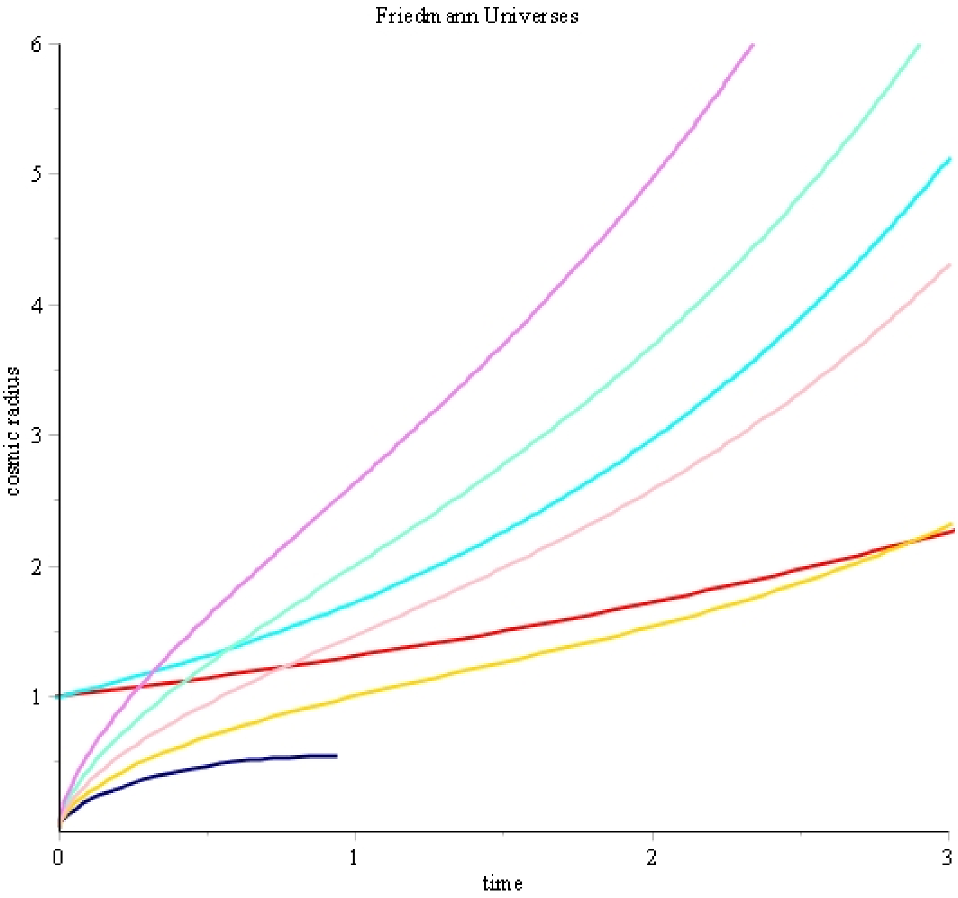
Obviously we can guess that the integral makes sense if and only if the if the sign of the cubic denominator under the square root is positive, i.e.:

$$A-x+yx^3>0 \tag{12.5.850}$$

Obviously as we have two parameters (i.e. A and y) we can play with both and there is an infinity of possibilities! So let us fixed $A:=1$, then we can have with Maple™ 17:

```
>Friedmann := int(sqrt(x/((8/27)*x^3-x+A)), x):
>plot({[int(sqrt(x/((8/27)*x^3-x+1)), x), x, x = 0
..  6], [int(sqrt(x/((8/27)*x^3-x+2)), x), x, x = 0
..  6],[int(sqrt(x/((8/27)*x^3-x+4)), x), x, x = 0
..  6], [int(sqrt(x/((8/27)*x^3-x+8)), x), x, x = 0
..  6], [int(sqrt(x/((8/27)*x^3-x+.5)), x), x, x = 0
..  6], [int(sqrt(x/((2/27)*x^3)), x), x, x = 0 ..  6],
[int(sqrt(x/((8/27)*x^3)), x), x, x = -1 ..  6]}, y = 0 ..  3,
thickness = 2, color = ["Red", "Cyan", "Navy", "Aquamarine", "Violet",
"Gold", "Pink"], labels = ["time", "cosmic radius"], labeldirections =
[horizontal, vertical], title = typeset("Friedmann Universes")):
```

This gives:



Alexander Friedmann's pioneer work is summarized in this plot. He for the first time showed that Einstein's general relativity admits non-static solutions. He named the Yellow, Pink, Aquamarine and Violet curves "monotonic worlds of the first kind." The cosmos expands at a decelerating rate from a zero-radius singularity until an inflection point, after which the expansion accelerates. This scenario provides the basis for our current view of the Big Bang and the accelerating universe. The Red and Curves curve are named as "monotonic world of the second kind" in which the universe expands from a non-zero initial radius. Half of the periodic scenario is shown in the Navy (dark blue) curve: the universe evolves from and back to zero radius.

12.5.8 The Black Hole Universe

A recent hypothesis in the history of cosmology (since the 11970 - according to the holocene calendar - as far as we know...) which is at the heart of many theoretical research (Stephen Hawking, Roger Penrose and others) is the possibility of assimilating our Universe to a Black Hole (see section General Relativity page 5835).

Remark

Keep in mind that the Black Hole Universe isn't mutually exclusive with other models like the CCC (Conformal Cyclic Cosmology) or the Universe out of Nothing (Wheeler-DeWitt Universe)! As Universe created from nothing or existing since always may have physical laws that authorize the generation of Black Holes (and their corresponding White Holes with expanding metric like the FLWR) that will themselves be compatible with the characteristics of a Big Bang.

The origin of the idea can be made from a very simple calculation:

We know that approximately the radius of the (current) Universe is given according to our previous calculations by:

$$r_0 = ct_0 = cH_0^{-1} \cong 1.2\cdot 10^{26}\,[\text{m}] \tag{12.5.851}$$

But we have proved in the section of General Relativity (and Classical Mechanics) that the Schwarzschild radius is given by:

$$R_S = \frac{2GM}{c^2} \tag{12.5.852}$$

What we can write for the Universe in the following form (under many assumptions: isotropy, homogeneity, spherical, etc.):

$$S_r = \frac{2GM}{c^2} = \frac{2G\rho_0 V}{c^2} = \frac{2G\rho_0\left(\dfrac{4}{3}\pi r_0^3\right)}{c^2} \tag{12.5.853}$$

which with the values of the critical density and the radius of the cosmological horizon calculated earlier above gives:

$$S_r = \frac{2G\rho_0\left(\dfrac{4}{3}\pi r_0^3\right)}{c^2} \cong \frac{2G\cdot 3.61\cdot 10^{52}}{c^2} \cong 0.53\cdot 10^{26} \tag{12.5.854}$$

So, roughly speaking, knowing all the uncertainties that we have accumulated in particular that on the Hubble parameter we see that the Schwarzschild radius is not very far from the radius of the present Universe.

As curious as it may seem, this question is not so far-fetched and is very seriously studied. It is therefore theoretically possible that our whole universe is encapsulated in a gigantic Black Hole (therefore of very large mass and very low density as we see it with our numerical values) of another inaccessible Universe ...

What is certain is that if this were the case, the expansion of the Universe (now observed) could not continue beyond the horizon of this super Black Hole because nothing coming from within can cross this horizon. However, recent observations seem to show that the expansion of the Universe is far from slowing and tends to accelerate with time, which is in contradiction with such a Black Hole Universe...

Figure 12.299: **Idea of Universe inside a (Schwarzschild) Black Hole**

Now a quite common question is: «*Why did the universe not collapse into a Black Hole shortly after the Big Bang?*».

First, the question is very meaningful as there is no "outside" (as far as we know) for the universe. And if we think about the Schwarzschild solution (and radius), it describes a spherical object outside of which the space is empty, and as we said there is no "outside" (as far as we know!) for the universe.

The standard cosmological model of the Big Bang (i.e. the cosmic inflation theory) fits observations to the Friedmann-Robertson-Walker solutions of General Relativity as we know, which do not form Black Holes. As far as we know, our Universe looks about the same from every point on the large scale. It is a built-in assumption of the Friedmann-Robertson-Walker family solutions, and sometimes named the "Copernican principle". In a homogeneous Friedmann-Robertson-Walker cosmology, symmetry guarantees that tidal forces vanish everywhere, and that any observer at rest relative to the average motion of matter will measure zero gravitational field. Based on these considerations the only kind of collapse that can occur in a purely homogeneous model is the recollapse of the entire universe in a "Big Crunch", and this happens only for matter densities and values of the cosmological constant that are different from what we actually observe.

Moreover, we have to remember that a high enough energy density is a necessary condition but not a sufficient condition for Black Holes to form: one needs to have a center which will ultimately become the center of the Black Holes; one needs the matter that collapses to the

Black Hole to have a low enough velocity so that gravity may squeeze it before the matter manages to fly away and dilute the density.

The latter two conditions are usually almost trivially satisfied for ordinary chunks of matter peacefully sitting at some place of the Universe; but they're almost maximally violated by the matter density right after the Big Bang. This matter has no center - it is almost uniform throughout space - and has high enough velocity (away from itself) that the density eventually gets diluted. And indeed, we know that it did get diluted.

In a survey made on 77 professional physicists (see [96]) in 12024 (holocene calendar) we get the following result for the fate of the matter that falls into a Black Hole:

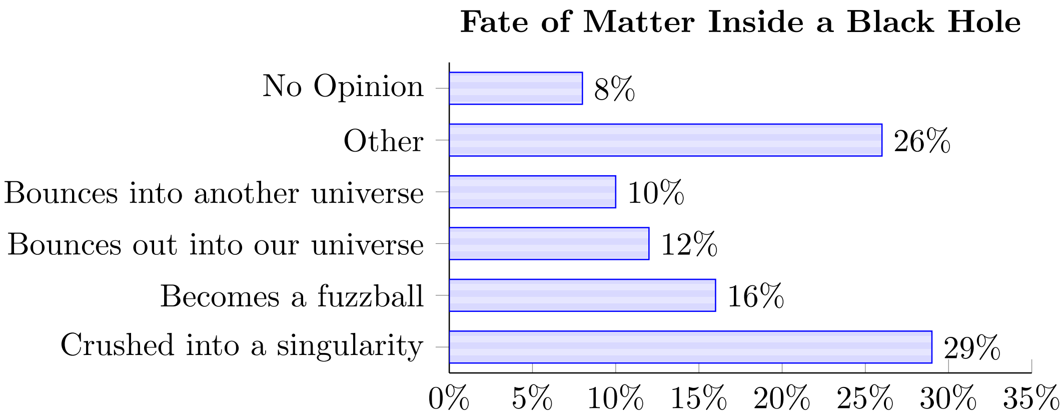


Figure 12.300: **Copenhagen survey on matter fate in a Black Hole**

In other words, a collapse of matter (e.g. a star) into a Black Hole is an idealized calculation that makes certain assumptions about the initial state of the matter. These assumptions are clearly not satisfied by matter after the Big Bang.

12.5.9 Borde-Guth-Vilenkin Theorem (BGV Theorem)

The "Borde–Guth–Vilenkin theorem", or the "BGV theorem", is a theorem in physical cosmology which deduces in the framework of General Relativity[72] that any universe that has, on average, been expanding throughout its history (expansion) cannot be infinite in the past but must have a past space-time boundary (this doesn't indicate an absolute beginning!). The theorem does not assume any specific mass content of the universe and it does not require gravity to be described by Einstein field equations.

In other words, the theorem shows that null (matter) and timelike (photons) geodesics are, in general, past-incomplete in inflationary models, whether or not energy conditions hold, provided only that the averaged expansion condition $H > 0$ holds along these past-directed geodesics.

It is named after the authors Arvind Borde, Alan Guth and Alexander Vilenkin, who developed its mathematical formulation in 12003 (holocene calendar). The BGV theorem is also popular outside physics, especially in religious and philosophical debates (where he is wrongly used by religious people... but this isn't a surprise).

Remark

Alternative models, where the average expansion of the universe throughout its history does not hold, have been proposed under the notions of emergent space-time, eternal inflation, and cyclic models (for cyclic models the expansion is obviously zero in average by construction). Vilenkin and Audrey Mithani have argued that none of these models escape the implications of the theorem. In 12017 (holocene calendar), Vilenkin stated that he does not think there are any viable cosmological models that escape the scenario (see *Alexander Vilenkin, "The Beginning of the Universe" in The Kalam Cosmological Argument: Volume 2, Bloomsbury, 2017, pp. 150-158*).

The original paper itself of Borde-Guth-Vilenking suggest possible prior cosmologies: «*What can lie beyond this boundary? Several possibilities have been discussed, one being that the boundary of the inflating region corresponds to the beginning of the Universe in a quantum nucleation event.*»

The reader must keep in mind that the BGV theorem is about classical space-time, the universe is quantum mechanical, so we don't even know yet if it breaks down for quantum space-times...

Victor Stenger asked Alexander Vilenkin the following question: «*Does your theorem prove that the universe must have had a beginning?*». Alexander Vilenkin replied: «*No. But it proves that the expansion of the universe must have had a beginning. You can evade the theorem by postulating that the universe was contracting prior to some time.*»

Alexander Vilenkin has also written about the religious significance of the BGV theorem. In October 12015 (holocene calendar), Alexander Vilenkin responded to arguments made by theist William Lane Craig and the New Atheism movement regarding the existence of God. Alexander Vilenkin stated «*What causes the universe to pop out of nothing? No cause is needed.*» (see [561]).

---

[72] Therefore in may not hold under consideration of a complete theory of quantum gravity.

Furthermore the original paper (see [64]) itself suggest possible prior cosmologies: «*What can lie beyond this boundary? Several possibilities have been discussed, one being that the boundary of the inflating region corresponds to the beginning of the Universe in a quantum nucleation event. The boundary is then a closed spacelike hypersurface which can be determined from the appropriate instanton.*»

Alan Guth has also gone on record claiming he believes the universe (i.e. the Cosmos) to be eternal.

Also, Arvind Borde has developed models of infinite universes as well as models of universe's with defined physics beyond the geodesic boundary. Literally none of the three original authors themselves believes that their theorem proves that the Universe has a beginning, because it doesn't. It shows that the expansion of the Universe has a beginning, not the entirety of the Cosmos itself.

The theorem that we will prove below is quite simple. Its proof does not go beyond undergraduate mathematics. But its implications for the beginning of the universe are very profound.... With the proof now in place, we see that the Cosmos can exist since ever but our Universe is not past-eternal (expected in a cyclic model).

Theorem 12.1. The geodesic incompleteness theorem of Borde, Guth, and Vilenkin as already stated is a general and powerful constraint on the structure of cosmological space-times in the framework of General Relativity for FLRW spaces types. A "geodesically complete space-time" is one for which all possible inertial paths in the space-time extend infinitely in proper time toward both the past and the future. By contrast, a "geodesically past-incomplete space-time" contains geodesics which encounter a boundary at finite past proper time, and a "geodesically future-incomplete space-time" contains geodesics which encounter a boundary at finite future proper time. The Borde-Guth-Vilenkin (BGV) Theorem states that any space-time for which the net expansion is positive:

$$\int H\mathrm{d}t > 0 \tag{12.5.855}$$

is necessarily geodesically past-incomplete. This has profound implications for cosmological model building, since it applies not only to trivially past-incomplete space-times such as matter- or radiation-dominated Friedmann-Lemaître-Robertson-Walker (FLRW) cosmologies, but to inflationary models as well. Furthermore, the proof is purely kinematic, and does not depend any assumption about energy conditions.

Proof 12.1.1. Before constructing a more general proof, we begin by reviewing the BGV theorem for geodesic incompleteness in an FLRW space-time. Let us remind that the latter is explicitly written (see page 5991):

$$\mathrm{d}s^2 = g_{\mu\nu}\mathrm{d}x^{\mu}\mathrm{d}x^{\nu} = -(c\,\mathrm{d}t)^2 + a^2(t)\left(\frac{\mathrm{d}x^2}{1-kx^2} + x^2\,\mathrm{d}\Omega^2\right) \tag{12.5.856}$$

We will take here a simplified notation of that metric of the following form and adopting natural units:

$$\mathrm{d}s^2=-\mathrm{d}t^2+a^2(t)\mathrm{d}\vec{x}^2 \tag{12.5.857}$$

where we assume a flat spatial geometry, $\mathrm{d}\vec{x}^2=\delta_{ij}\mathrm{d}x^i\mathrm{d}x^j$, for convenience. We wish to address the question of whether such space-times are geodesically complete, which is the statement that all inertial (or geodesic) observers see the space-time as both past- and future-infinite, as measured by the proper time along a geodesic world line in the space-time:

$$\int\limits_{-\infty}^{t}\mathrm{d}s\rightarrow+\infty,\qquad\int\limits_{t}^{+\infty}\mathrm{d}s\rightarrow+\infty \tag{12.5.858}$$

As we already know, the Friedman equations derived from the FLRW metric are singular for $a(t)=0$.

We have proven that for flat spaces that are matter- or radiation-dominated we had $a(t)\propto t^{\alpha}$ (see page 5905), and are therefore trivially geodesically incomplete, since $a(t)\rightarrow 0_+$ at finite past time. The case of the Einstein-De Sitter Universe model (see page 6016):

$$a(t)\propto e^{Ht}\qquad H=c^{te} \tag{12.5.859}$$

is more ambiguous, since singularity is reached only at $t\rightarrow-\infty$ for a comoving observer. It is, however, straightforward to show that non-comoving geodesics are past-finite.

For a timelike geodesic, normalization $u^{\mu}u_{\mu}=-1$ given for recall by:

$$g_{\mu\nu}u^{\mu}u^{\nu}=g_{\mu\nu}\frac{\mathrm{d}x^{\mu}}{\mathrm{d}\tau}\frac{\mathrm{d}x^{\nu}}{\mathrm{d}\tau}=\frac{g_{\mu\nu}\mathrm{d}x^{\mu}\mathrm{d}x^{\nu}}{\mathrm{d}\tau^2}=\frac{\mathrm{d}s^2}{\mathrm{d}\tau^2}=\frac{-\mathrm{d}\tau^2}{\mathrm{d}\tau^2}=-1 \tag{12.5.860}$$

gives (we take a photon therefore remember - in natural units - that $\mathrm{d}s=\mathrm{d}\tau$ as seen at page 5707):

$$u^{\mu}u_{\mu}=g_{\mu\nu}\frac{\mathrm{d}x^{\mu}}{\mathrm{d}\tau}\frac{\mathrm{d}x^{\tau}}{\mathrm{d}s}=-\left(\frac{\mathrm{d}t}{\mathrm{d}s}\right)^2+a^2(t)\left|\frac{\mathrm{d}\vec{x}}{\mathrm{d}s}\right|^2=-1 \tag{12.5.861}$$

The last equality is often denoted:

$$-\dot{x}_0^2+a^2(t)|\dot{\vec{x}}|^2=-1 \tag{12.5.862}$$

Hence (useful result for later):

$$\dot{x}_0^2-a^2(t)|\dot{\vec{x}}|^2=1 \tag{12.5.863}$$

Now let us remind the variational principle (see page 5747):

$$\delta\int\sqrt{g_{\mu\nu}\mathrm{d}x^{\mu}\mathrm{d}x^{\nu}}=0 \tag{12.5.864}$$

Hence in our case (still in natural units):

$$\delta\int\sqrt{-\mathrm{d}t^2+a^2(t)\mathrm{d}\vec{x}^2}=0 \tag{12.5.865}$$

But we also have the action integral (see section Analytical Mechanics page 3161):

$$\delta S=\int L\mathrm{d}t \tag{12.5.866}$$

We can identify $L=\sqrt{\dot{x}_0^2-a^2(t)|\dot{\vec{x}}|^2}=L\left(\dot{x}_0,\dot{\vec{x}}\right)$. Then, from the Euler-Lagrange equations (see section Classical Mechanics page 3162):

$$\frac{\mathrm{d}}{\mathrm{d}s}\frac{\partial L}{\partial \dot{x}_\mu}=\frac{\partial L}{\partial x_\mu} \tag{12.5.867}$$

We have in our case:

$$\frac{\partial L}{\partial x_\mu}=0 \tag{12.5.868}$$

and:

$$\frac{\partial L}{\partial \dot{x}_\mu}=\frac{-2a^2(t)\dot{\vec{x}}}{\sqrt{\dot{x}_0^2-a^2(t)|\dot{\vec{x}}|^2}} \tag{12.5.869}$$

Therefore we have:

$$\frac{\mathrm{d}}{\mathrm{d}s}\frac{-2a^2(t)\dot{\vec{x}}}{\sqrt{\dot{x}_0^2-a^2(t)|\dot{\vec{x}}|^2}}=0 \tag{12.5.870}$$

And simplifying a bit:

$$\frac{\mathrm{d}}{\mathrm{d}s}\frac{a^2(t)\dot{\vec{x}}}{\sqrt{\dot{x}_0^2-a^2(t)|\dot{\vec{x}}|^2}}=0 \tag{12.5.871}$$

But we have proved earlier above that $-\dot{x}_0^2+a^2(t)|\dot{\vec{x}}|^2=-1$, therefore:

$$\frac{\mathrm{d}}{\mathrm{d}s}a^2(t)\dot{\vec{x}}=0 \tag{12.5.872}$$

Or explicitly:

$$\frac{\mathrm{d}}{\mathrm{d}s}\left[a^2(t)\frac{\mathrm{d}\vec{x}}{\mathrm{d}s}\right]=0 \tag{12.5.873}$$

so that we can define an integration constant v_0 such that:

$$a^2(t)\left|\frac{\mathrm{d}\vec{x}}{\mathrm{d}s}\right|\equiv v_0=c^{te} \tag{12.5.874}$$

From (see earlier above):

$$u^\mu u_\mu=g_{\mu\nu}\frac{\mathrm{d}x^\mu}{\mathrm{d}\tau}\frac{\mathrm{d}x^\tau}{\mathrm{d}s}=-\left(\frac{\mathrm{d}t}{\mathrm{d}s}\right)^2+a^2(t)\left|\frac{\mathrm{d}\vec{x}}{\mathrm{d}s}\right|^2=-1 \tag{12.5.875}$$

We have:

$$\left(\frac{\mathrm{d}t}{\mathrm{d}s}\right)^2=1-a^2(t)\left|\frac{\mathrm{d}\vec{x}}{\mathrm{d}s}\right|^2 \tag{12.5.876}$$

Injecting the previous result, we have:

$$\left(\frac{\mathrm{d}t}{\mathrm{d}s}\right)^2=\gamma^2=1+v_0^2a^{-2}(t) \tag{12.5.877}$$

where:

$$\gamma^2=\frac{1}{1-v^2} \tag{12.5.878}$$

is the Lorentz boost (see section Special Relativity page 5638) as a function of the velocity v of the observer, so that the integration constant v_0 is:

$$v_0^2=\frac{a^2v^2}{1-v^2}=c^{te} \tag{12.5.879}$$

The differential proper time $\mathrm{d}s$ along the geodesic of the freely falling observer can then be written in terms of the coordinate time $\mathrm{d}t$ as:

$$\mathrm{d}s=\frac{\mathrm{d}t}{\sqrt{1+v_0^2a^{-2}(t)}} \tag{12.5.880}$$

For $a(t)=e^{Ht}$ (for reminder: a Einstein-De Sitter Universe model, i.e. flat space dominated by matter), the integral:

$$\Delta s=\int\limits_{-\infty}^{0}\frac{\mathrm{d}t}{\sqrt{1+v_0^2e^{-2Ht}}} \tag{12.5.881}$$

So let us calculate the primitive:

$$\int\frac{1}{\sqrt{v_0^2\mathrm{e}^{-2Ht}+1}}\mathrm{d}t \tag{12.5.882}$$

We substitute $u=-2Ht\longrightarrow\mathrm{d}u=-2H\mathrm{d}t$, therefore:

$$\int\frac{1}{\sqrt{v_0^2\mathrm{e}^{-2Ht}+1}}\mathrm{d}t=-\frac{1}{2H}\int\frac{1}{\sqrt{v_0^2\mathrm{e}^{u}+1}}\mathrm{d}u \tag{12.5.883}$$

Now we solve:

$$\int\frac{1}{\sqrt{v_0^2e^u+1}}\,\mathrm{d}u \tag{12.5.884}$$

For this we substitute $v=v_0^2e^u+1\longrightarrow\mathrm{d}v=v_0^2e^u\mathrm{d}u$, hence:

$$\int\frac{1}{\sqrt{v_0^2e^u+1}}\mathrm{d}u=\int\frac{1}{(v-1)\sqrt{v}}\,\mathrm{d}v \tag{12.5.885}$$

We substitute $w=\sqrt{v}\longrightarrow \mathrm{d}w=\frac{1}{2\sqrt{v}}\mathrm{d}v$ (and using $v=w^2$):

$$\begin{aligned}\int\frac{1}{(v-1)\sqrt{v}}\mathrm{d}v&=2\int\frac{1}{w^2-1}\mathrm{d}w=2\int\frac{1}{w^2-1}\mathrm{d}w=\int\frac{1}{(w-1)(w+1)}\mathrm{d}w\\&=\int\left(\frac{1}{2(w-1)}-\frac{1}{2(w+1)}\right)\mathrm{d}w\\&=\frac{1}{2}\int\frac{1}{w-1}\mathrm{d}w-\frac{1}{2}\int\frac{1}{w+1}\mathrm{d}w=\frac{\ln(w-1)}{2}-\frac{\ln(w+1)}{2}\\&=\frac{1}{2}\frac{\ln(w-1)}{\ln(w+1)}\end{aligned}\tag{12.5.886}$$

And if we undo all substitutions we get:

$$\begin{aligned}\Delta s&=\int\limits_{-\infty}^{0}\frac{\mathrm{d}t}{\sqrt{1+v_0^2e^{-2Ht}}}=\frac{1}{2H}\ln\left(\frac{\sqrt{1+v_0^2}+1}{\sqrt{1+v_0^2}-1}\right)\\&=\frac{1}{2H}\ln\left(\frac{\gamma_0+1}{\gamma_0-1}\right)\end{aligned}\tag{12.5.887}$$

This integral is finite for any $v_0\neq 0$, so the observer on a non-comoving world line sees the space as past-finite, and the space is geodesically past-incomplete.

The BGV theorem states that any space-time for which the average expansion rate $\mathcal{H}_{\text{av}}$ is positive is geodesically incomplete, where:

$$\mathcal{H}_{\text{av}}\equiv\frac{1}{\Delta s}\int H\mathrm{d}s\tag{12.5.888}$$

and $\mathrm{d}s$ is the proper time along a geodesic world line. First, let us consider the integral of the expansion rate H along a comoving world line, for which the proper time $\mathrm{d}s=\mathrm{d}t$ (see page 5887):

$$\mathcal{H}_{\text{av}}=\int\limits_{t_i}^{t_f}H\mathrm{d}t=\int\limits_{t_i}^{t_f}\frac{1}{a}\frac{\mathrm{d}a}{\mathrm{d}t}\mathrm{d}t=\int\limits_{a_i}^{a_f}\mathrm{d}\ln(a)=\ln(a_f)-\ln(a_i)\tag{12.5.889}$$

Since the integrand is a total derivative, the integral is determined by the the values of the scale factor on the boundary, and measures the total expansion from t_i to t_f. Note in particular that this integral is logarithmically divergent as $a_i\to 0$, consistent with the fact that comoving world lines are past-infinite in proper time. We can integrate the expansion rate H along a non-comoving geodesic with proper time ds as follows (using usual primitives derived in the section of Differential and Integral Calculus at page 1534):

$$\begin{aligned}\mathcal{H}_{\text{av}}&=\int\limits_{s_i}^{s_f}H\mathrm{d}s=\int\limits_{t_i}^{t_f}H\frac{\mathrm{d}s}{\mathrm{d}t}\mathrm{d}t=\int\limits_{t_i}^{t_f}\frac{1}{a}\left(\frac{\mathrm{d}a}{\mathrm{d}t}\right)\frac{1}{\sqrt{1+v_0^2/a^2}}\mathrm{d}t\\&=\int\limits_{a_i}^{a_f}\frac{\mathrm{d}a}{\sqrt{a^2+v_0^2}}=\frac{1}{2}\ln\left[\frac{\sqrt{v_0^2+a^2}+a}{\sqrt{v_0^2+a^2}-a}\right]\Bigg|_{a_i}^{a_f}\end{aligned}\tag{12.5.890}$$

Integrating from $a_i = 0$ (this assumes implicitly a beginning, i.e. bounding, of space... hence we have here a circular reasoning) to $a_f = 1$ (the scale factor of today), this integral is finite:

$$\mathcal{H}_{\text{av}} = \int\limits_{s_i}^{s_f} H \mathrm{d}s = \frac{1}{2}\ln\left(\frac{\sqrt{1+v_0^2}+1}{\sqrt{1+v_0^2}-1}\right) = \frac{1}{2}\ln\left(\frac{\gamma_0+1}{\gamma_0-1}\right) \qquad (12.5.891)$$

Note that we have not taken $H = c^{te}$. The only assumption is that the overall change in scale factor is positive, $a_f > a_i$.

The BGV theorem is then the statement that as long as $\mathcal{H}_{\text{av}}$ is finite and positive (under all the assumptions mentioned above!), the proper time Δs along a time-like world-line is finite:

$$\Delta s = \int \mathrm{d}s = \frac{1}{\mathcal{H}_{\text{av}}} \int H \mathrm{d}s \qquad (12.5.892)$$

The BGV Theorem as stated above is, however, something of a circular argument, since $\mathcal{H}_{\text{av}}$ is only well-defined if Δs is finite in the first place.

■ Q.E.D.

There is another

In summary, the BGV theorem states that an inflating universe cannot be extended infinitely into the past and must have had a beginning at some finite time in the past. This supports the idea that the universe is of a finite age, but also suggests that it may have passed through a singularity phase.

This applies not only to spaces where the scale factor a grows monotonically, but also the Ijjas-Steinhardt model of cyclic cosmology (see [311]). The IS model avoids the problem of) for which the scale factor $a(t)$ is oscillatory, but grows from one cycle to the next.

We have above only demonstrated a result for an FRW metric but in the original paper there is also another proof for non-FRW metrics.

Remark

Work done by the Landau group in the years 11960 (holocene calendar) led to the conclusion that the general solution of General Relativity does not contain a physical singularity but a chaotic singularity named de "Belinski–Khalatnikov–Lifshitz (BKL) singularity". The search for a broader class of solutions with a singularity has been done, essentially, by a trial-and-error method, since a systematic approach to the study of the Einstein equations was lacking. A negative result, obtained in this way, is not convincing by itself; a solution with the necessary degree of generality would invalidate it, and at the same time would confirm any positive results related to the specific solution.

quote
12.5:6067.#72

> ❞ *What causes the universe to pop out of nothing? No cause is needed. If you have a radioactive atom, it will decay, and quantum mechanics gives the decay probability in a given interval of time, say, a minute. There is no reason why the atom decayed at this particular moment and not another. The process is completely random. No cause is needed for the quantum creation of the universe.* ❞
>
> Alexander Vilenkin

12.5.10 Empirical evidence of the Cosmic Inflation theory (i.e. Big Bang)

Most Astronomers, Astrophysicists, Cosmologists and Physicist think actually that the Big Bang is the best way to explain how the Universe started and how it changed over time (however don't forget that in Science, scientists opinion and majority isn't important: only the Data are important!). All science is based on evidence as we know. So what are the evidence in this early 121st century (holocene calendar) for the Big Bang (obviously most of them have existing meta-analysis level 15 evidence[73]!)?

Here is a list of most common factors considered as good evidence for the Big Bang (especially the ΛCDM model):

- **Redshift of Galaxies (i.e. Hubble's law):** The light we observe from galaxies has been stretched by the time it reaches us. It looks redder than it should. This redshift is the result of galaxies moving away from us (see [264]). Observations show that pretty much everything in the Universe is moving apart. The redshift of distant galaxies tells us the Universe is expanding. Some tension still exist in this early 121st century about the measurement accuracy of the Hubble constant using different methods that doesn't converge (maybe the ΛCDM model must be adjusted)

- **Microwave Background:** A long, long time ago, the whole Universe was very hot. As it grew in the size, the heat left a "glow" which fills the entire Universe. The Big Bang theory (cosmic inflation theory) predicts this glow should still exist. It also predicts that we should be able to detect this glow as microwave light and as a perfect black body radiation distribution curve. Scientist have found this Cosmic Microwave Background (see [434]). They have accurately measured it using orbiting detectors and it follows very accurately a black body radiation distribution curve. The CMB power spectrum multipole moment is also in very good agreement with the theoretical assumption of the spherical harmonics baryon acoustic oscillations. We have evidence also for the E-modes polarization of the Big Bang (see [321]), but also the B-Modes polarization (see [227]), and Primordial gravitational waves (see [7]).

- **Mixture of Elements:** Some chemical elements were generated soon after the Big Bang. Elements like hydrogen and helium. Some models based on a Big Bang assumption predicts quite well how much of each element was made in the early universe (however sometimes they fail completely for some other elements). When astronomers look at very old galaxies and stars, the amount of each chemical they see agrees quite good with these models based on a Big Bang assumption (see [61]). A point where actually there is also a lack of comprehension is where has gone the antimatter even if we know that CPT is violated, it doesn't explained good enough what we see around us (however don't forget we can see only the observable universe and not the universe in it's entirety).

- **Galactic evolution and distribution:** Detailed observations of the morphology and distribution of galaxies and quasars are in quite good agreement with the current Big Bang models (see [125]). A combination of observations and theory suggest that the first quasars and galaxies formed within a billion years after the Big Bang, and since then, larger structures have been forming, such as galaxy clusters and superclusters. Supercomputer simulations agree well with our observations for our observable universe.

---

[73] Don't forget that its means it's factual beyond all reasonable evidence but it's not the truth as "truth" doesn't exist in experimental science!

- **Baryon acoustic oscillations:** Multiple independent surveys like the Sloan Digital Sky Survey (SDSS), the 2dF Galaxy Redshift Survey (2dFGRS) and the 6dF Galaxy Survey (6dFGS), the WiggleZ and the Baryon Oscillation Spectroscopic Survey (BOSS) mapped the spatial distribution of luminous red galaxies (LRGs) and quasars to determine their spatial distribution and detected the characteristic scale imprinted by baryon acoustic oscillations (sound waves that propagate in the early universe) in the early universe (see [125]).

- **Age of the universe:** The age of the universe that was estimated from the Hubble law and cosmic microwave background is in good agreement with the ages of the oldest stars, both as measured by applying the theory of stellar evolution to globular clusters (see [292]).

- **Olber's paradox:** The Big Bang (i.e. the cosmic inflation theory) solves this paradox, expressed by Heinrich Olbers in 11826 (holocene calendar), that asks why is the sky dark at night? In an infinitely large and unchanging universe, uniformly populated with stars and galaxies, the sky should be dazzlingly bright. Olbers suggested a solution for this paradox: he thought that the enormous amount of dust in the universe must be absorbing the greater portion of the light emitted by the stars, causing the sky to darken. It was calculated afterwards that even this dust would get hot because of the radiation it had absorbed and would radiate with the same intensity. The paradox came to be solved following the Big Bang's postulation that the universe must have had a beginning and that it was expanding. So, the fact that the night sky is dark indicates that the universe cannot have a static infinite number of evenly distributed stars over an infinite period of time (see [234]).

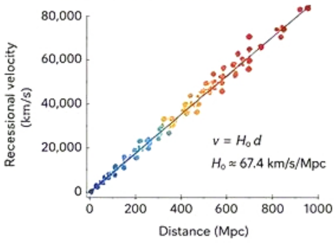
- **Gravitational potential paradox:** Johann Friendrich Zöllner's (11871 according to holocene calendar) paradox about gravitation was also invalidated by the expanding universe model. Zöllner maintained that if we imagined the stars in an infinite and static universe evenly scattered in space, as Newton had assumed, there should have been infinite gravitational potential at every point of the universe. Such a postulate was hardly compatible with common sense and observations; in this way the Big Bang model (i.e. the cosmic inflation theory) that postulated an expanding, dynamic and finite universe also resolved this paradox.

THE BIG BANG (λCDM)

ALL MAJOR LINES OF EVIDENCE

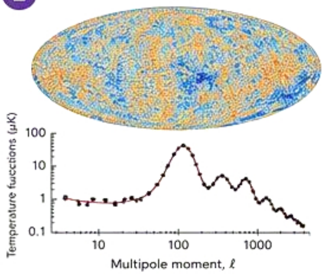
The Big Bang model, within the framework of ΛCDM (Lambda Cold Dark Matter), provides a consistent explanation for a wide range of observations across the Universe.

1 UNIVERSE EXPANSION (HUBBLE'S LAW)



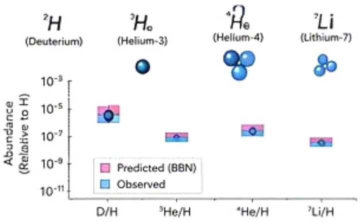
- Galaxies recede from us with a velocity proportional to their distance.
- Measuring this redshift-distance relation gives the Hubble constant.
- Consistent with an expanding Universe from a hot, dense past.

2 COSMIC MICROWAVE BACKGROUND (CMB)



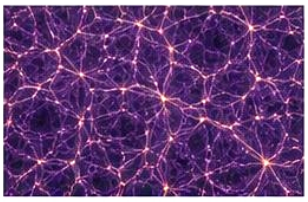
- The CMB is relic radiation from ~380,000 years after the Big Bang.
- Its near-perfect blackbody spectrum at 2.725 K matches predictions.
- Tiny anisotropies show the acoustic peaks predicted by ΛCDM.
- Measured in exquisite detail by Planck.

3 PRIMORDIAL ABUNDANCES (BIG BANG NUCLEOSYNTHESIS)



- In the first few minutes, the Universe produced light nuclei.
- Observed abundances of D, ^{3}He, ^{4}He and ^{7}Li match Big Bang predictions using a single parameter: the baryon density.

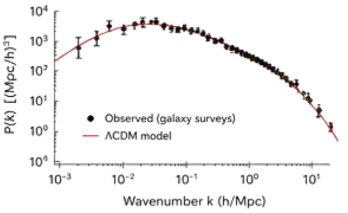
4 LARGE-SCALE STRUCTURE OF THE UNIVERSE



Cosmic Web (simulation)

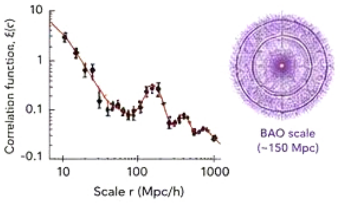
- Galaxies are not randomly distributed: they form a cosmic web of filaments, clusters and voids.
- ΛCDM, with gravity amplifying tiny primordial fluctuations, naturally produces this pattern.
- Observed by galaxy surveys (SDSS, DESI, etc.).

5 THE MATTER POWER SPECTRUM



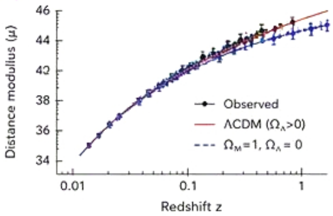
- The distribution of matter fluctuations versus scale matches ΛCDM predictions over many orders of magnitude.
- Confirms the physics of inflation (primordial fluctuations) and gravity.

6 BARYON ACOUSTIC OSCILLATIONS (BAO)



- Sound waves in the early Universe left an imprint in the matter distribution.
- Detected as a preferred scale (~150 Mpc) in galaxy clustering.
- Acts as a "standard ruler" confirming the expansion history.

7 SUPERNOVAE TYPE Ia



- Distant supernovae are dimmer than expected in a decelerating Universe.
- Indicates cosmic acceleration due to dark energy (Λ).
- Precisely measured by many surveys.

8 THE AGE OF THE UNIVERSE

ΛCDM prediction
13.80 ± 0.02 billion years

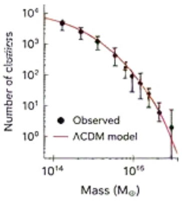
Oldest stars
13.4 ± 0.8 billion years

Globular clusters
12.8 ± 0.7 billion years

Meteorites (Solar System)
4.56 ± 0.02 billion years

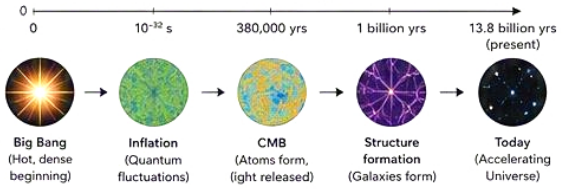
- Independent dating methods agree with the ΛCDM age.
- The Universe is old enough for galaxies, stars and planets to have formed.

9 GRAVITATIONAL LENSING & CLUSTER COUNTS



- The amount of lensing and the number of galaxy clusters as a function of mass match ΛCDM predictions.
- Requires both dark matter and dark energy with measured densities.

THE ΛCDM MODEL: A CONSISTENT PICTURE

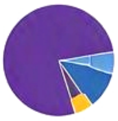


Big Bang (Hot, dense beginning) → **Inflation** (Quantum fluctuations) → **CMB** (Atoms form, (ight released) → **Structure formation** (Galaxies form) → **Today** (Accelerating Universe)

FROM THE VERY SMALL TO THE VERY LARGE, FROM THE VERY EARLY TO TODAY:
All lines of evidence converge on the same story.

ΛCDM WORKS
One model. Many tests. Remarkable agreement.

COMPOSITION OF THE UNIVERSE (TODAY)



68.5%	Dark Energy (Λ)
26.5%	Dark Matter (Cold)
5.0%	Ordinary Matter (Atoms)

The ΛCDM models has however in the early 121st century (holocene calendar) some exciting tensions/challenges like:

- Significative statistical disrepencies between the different methods of measurement of the Hubble constant value.
- There are super-amas structures observed whose size is too big to respect the homogeneity assumption of the FLWR metric.
- Stars and galaxies observed with high redshift by the JWST are observed at a percentage that isn't right statistically with theoretical models.
- The abundance of the observed Lithium is incompatible with the nucleosynthesis theoretical models.
- ...and many other (approximately 40) well summarized in the following paper [437].

Again, let us remind that alternative models challenge the assumptions of the ΛCDM model. Examples of these are modified Newtonian dynamics (MOND), entropic gravity, $f(R)$ modified gravity, theories of large-scale variations in the matter density of the universe, bimetric gravity, scale invariance of empty space, decaying dark matter (DDM), Kaluza-Klein theory (and variations like the DGP model), Timescapes that doesn't required dark energy (see), GRSI (Gravitationally Rescaled Strong Interaction) model, SIDM (Strong Interactive Dark Matter) model, Shockwave cosmology, etc.

That's why some cosmologists think that the ΛCDM model may be an approximation of a more fundamental model.

~90 % ★★★☆☆

17 votes, 70.59%

12.6. Quantum Gravity/Cosmology

Quantum gravity (QG) and Cosmology are two fields of theoretical physics that seeks to describe respectively gravity and the universe according to the principles of quantum mechanics, and where quantum effects cannot be ignored, such as near compact astrophysical objects where the effects of gravity are strong or the beginning of the universe.

To start this section we think it is maybe good to remind the reader to refer first to the Okun cube (see section Principia page 3125). Once this done, here is another restricted version of the Okun cube below (but not really accurate) that highlights an interesting field that does not appear in the Okun cube and that is the Newtonian quantum gravity that we will see first. But this figure is interesting as it highlights well the different models we will deal with in this section:

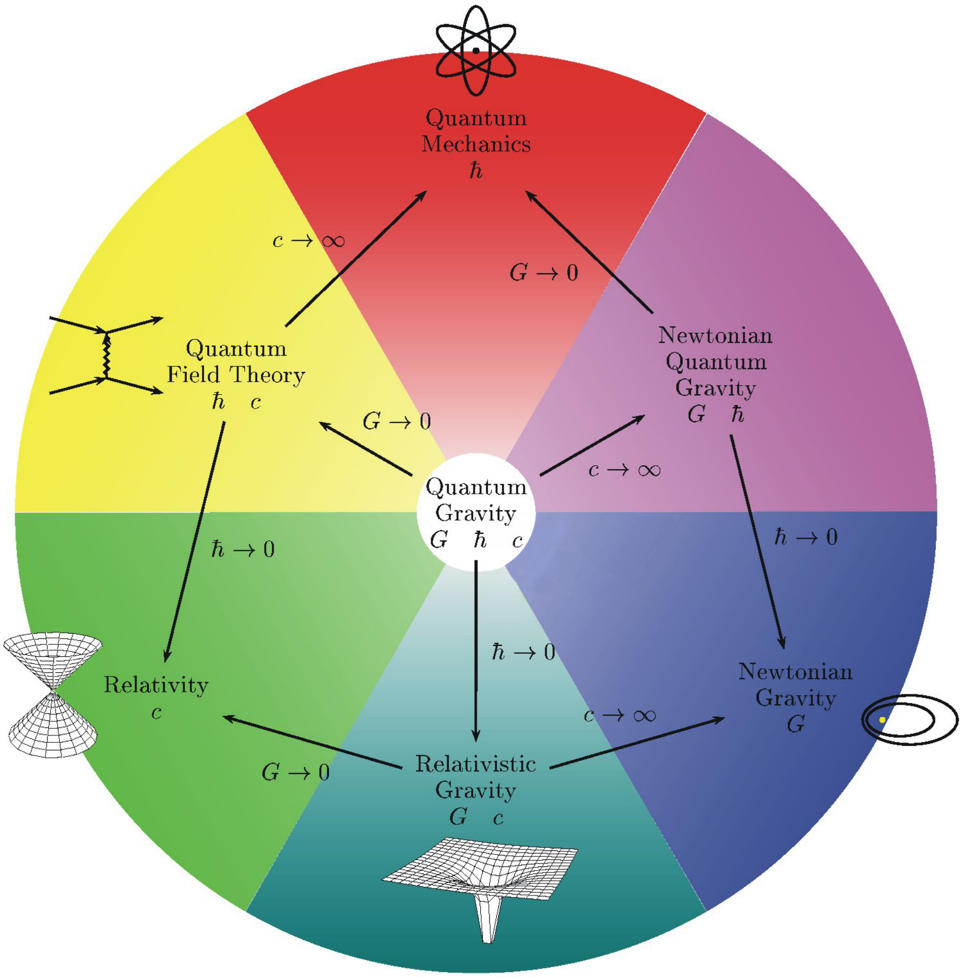

Quantum Mechanics $\hbar$
$c \to \infty$
$G \to 0$
Quantum Field Theory $\hbar$ c
$G \to 0$
Newtonian Quantum Gravity G $\hbar$
$c \to \infty$
Quantum Gravity G $\hbar$ c
$\hbar \to 0$
$\hbar \to 0$
$\hbar \to 0$
Relativity c
$c \to \infty$
Newtonian Gravity G
$G \to 0$
Relativistic Gravity G c

Figure 12.301: **Okun wheel (source: ?)**

Remark

The above figure is especially not accurate (same problem for the Okun cube) for the arrow that links together Quantum Field Theory and Quantum Mechanics as there are no obvious and unique way that really links them together.

12.6.1 Newton Quantum Gravity

A Newtonian approach to quantum gravity is studied. At least for weak gravitational fields it should be a valid approximation. Such an approach could be used to point out problems and prospects inherent in a more exact theory of quantum gravity, yet to be discovered. Newtonian quantum gravity, e.g., shows promise for prohibiting Black Holes altogether (which would eliminate singularities and also solve the Black Hole information paradox), breaks the equivalence principle of General Relativity, and supports non-local interactions (quantum entanglement). Its predictions should also be testable at length scales well above the "Planck scale", by high-precision experiments feasible even with existing technology. It turns out that, e.g., the Solar System, superficially, can be described as a quantum gravitational system, provided that the l quantum number has its maximum value, $n-1$. This results exactly in Kepler's third law. If also the m quantum number has its maximum value ($\pm l$) the probability density has a very narrow torus-like form, centered around the classical planetary orbits. However, as the probability density is independent of the azimuthal angle ϕ there is, from quantum gravity arguments, no reason for planets to be located in any unique place along the orbit (or even *in* an orbit for $m \neq \pm l$). This is, in essence, a reflection of the "measurement problem" inherent in all quantum descriptions.

The greatest fundamental challenge facing theoretical physics has for many years been to reconcile gravity with quantum physics. There have been numerous attempts to do so, but so far there is no established and experimentally/observationally tested theory of "quantum gravity", the two main contenders presently being String Theory and loop quantum gravity, with "outsiders" like twistor theory, non-commutative geometry, etc.

The motivations for studying Newtonian quantum gravity are:

- Quantum theory is supposed to be universal, i.e., it should be valid on all length scales and for all objects, as there in principle exists no size/charge/mass-limit to its applicability. In atomic physics the practical restriction comes about due to the fact that there is a limit to arbitrarily large atomic nuclei as:

  1. The Coulomb force between protons is repulsive, eventually overpowering the strong nuclear force trying to hold the nucleus together
  2. The additional weak force makes neutron-rich nuclei decay before they grow too large

  Also, the electric charge comes in both positive and negative, and as a result a big lump of matter is almost always electrically neutral[74]. Neither of these effects are present in

---

[74] The same also applies for e.g. the strong force, as the three different color charges ("red", "green", "blue") always combine to produce color-neutral hadrons and bulk matter.

"pure" quantum gravity.

- For weak gravitational fields the Newtonian theory should be sufficient. The weak-field Newtonian limit is even used for determining the constant κ in Einstein's field equations of General Relativity $G_{\mu\nu} = \kappa T_{\mu\nu}$. The Newtonian limit is also almost always sufficient for practical purposes in non-quantum gravity, except for a handful of extreme cases (notably Black Holes and the very early universe), although high-precision experiments in e.g. the Solar System can and do show deviations from the Newtonian theory, always in favour of General Relativity.

- Even for strong gravitational fields the Newtonian picture gives the same prediction as general relativity for the Schwarzschild radius of a spherically symmetric, non-rotating Black Hole, and correct order of magnitude results for neutron stars and cosmology. This could make it possible to deduce at least qualitative results about strongly coupled quantum gravity, as the Newtonian viewpoint should give reliable first order quantum gravitational results.

On the other hand would any "absurd" results obtained from Newtonian quantum gravity, deviating from observations, implicate either that:

- General relativity cannot be quantized[75]. An unsuccessful special case (the weak field limit) would disprove the general case, whereas the opposite is not true (a vindicated weak field limit will not prove that the general theory is also correct).

- Quantum mechanics fails at "macroscopic" distances and for macroscopic objects. This would mean that we in gravity have a unique opportunity to understand the "measurement problem" in quantum mechanics, as proposed by e.g. Károlyházy and Penrose. In that case we can use gravity to probe the transition between quantum $\rightarrow$ classical behaviour in detail, i.e. get experimental facts on where, how and when the inherently undecided, subjective quantum world of superpositions turns into the familiar objective classical everyday world around us. One could, at least in principle, envisage a test carried out in a free-falling (e.g. satellite) environment where one alters m (the gravitational "test-charge") and M (the gravitational "source-charge") until the expected quantum gravity results are observed, to obtain a limit of where the quantum mechanical treatment breaks down, hence making an experimental determination of the border between "quantum" and "classical", i.e. solving the quantum mechanical measurement problem. Fundamental quantum gravity and the quantum mechanical measurement problem may well be intertwined and might need to be resolved simultaneously in a successful approach.

In Newtonian quantum gravity, at least as long as the system can be approximately treated as a 2-body problem, it is possible to use the mathematical identity between the electrostatic Coulomb force in the hydrogen atom, and Newton's static gravitational force under the substitution $e^2/4\pi\varepsilon_0 \rightarrow GmM$. Therefore all analytical results from elementary quantum physics directly lifts over to the quantum gravity case. For weak electromagnetic fields, as in the hydrogen atom, the electrodynamic corrections to the static Coulomb field are very small, making the approximation excellent. The same applies to quantum gravity, dynamical effects from general relativity are negligible to a very high degree for weak gravitational fields. A gravitationally bound 2-body system should then exhibit exactly the same type of "spectrum" as a hydrogen atom, but emitted in (unobservable) hypothesized graviton form instead of photons (easily detectable as atomic spectra already in the 119th century according to holocene calendar).

---

[75] This is an automatic consequence of "emergent" gravity, e.g. Sakharov's theory, where gravity is a non-fundamental interaction and rather a macroscopic consequence of other forces and fields.

For a free-falling 2-body system, e.g. in a satellite experiment enclosed in a spherical vessel, it should in principle be possible to measure the excitation energies for a suitable system. An analogous result has seemingly already been accomplished for neutrons in the gravitational field of the Earth, although there are some quantum gravity ambiguities as noted below.

For hydrogen-like (one electron) atoms, in the dominant Coulomb central-field approximation, the energy levels depend only on the principal quantum number, $n = 1, 2, 3, \ldots$:

$$E_n = -\frac{me^4}{8\varepsilon_0^2 h^2}\frac{Z^2}{n^2} = -E_H \frac{Z^2}{n^2} \tag{12.6.1}$$

where $E_H \simeq 13.6$ [eV] is the ionization energy, i.e. the energy required to free the electron from the proton, and Z the number of protons in the nucleus.

The Bohr-radius, a_0, the innermost radius of circular orbits in the old semi-classical Bohr-model, and also the distance r for which the probability density of the Schrödinger equation for the Hydrogen ground-state peaks, is:

$$a_0 = \frac{h^2\varepsilon_0}{\pi m e^2} \tag{12.6.2}$$

whereas the expectation value for the electron-nucleus separation is:

$$\langle r\rangle_{\text{hydrogen}} \simeq \frac{n^2}{Z}a_0 = \frac{n^2 h^2 \varepsilon_0}{Z\pi m e^2} \tag{12.6.3}$$

A comparison between the Coulomb potential in Hydrogen-like atoms:

$$V_{\text{hydrogen}} = -\frac{Ze^2}{4\pi\varepsilon_0 r} \tag{12.6.4}$$

and the Newtonian gravitational potential between two masses m and M:

$$V_{\text{grav}} = -\frac{GmM}{r} \tag{12.6.5}$$

allows us to obtain all results of the gravitational case by the simple substitution:

$$\frac{Ze^2}{4\pi\varepsilon_0} \rightarrow GmM \tag{12.6.6}$$

in the well-known formulas for the Hydrogen atom.

For instance, the gravitational "Bohr-radius", b_0, becomes:

$$b_0 = \frac{h^2}{4\pi^2 G m^2 M} = \frac{\hbar^2}{G m^2 M} \tag{12.6.7}$$

and the quantum-gravitational energy levels are:

$$E_n(\text{grav}) = -\frac{2\pi^2 G^2 m^3 M^2}{h^2}\frac{1}{n^2} = -\frac{G^2 m^3 M^2}{2\hbar^2}\frac{1}{n^2} = -E_g\frac{1}{n^2} \tag{12.6.8}$$

here again $E_g = G^2 m^3 M^2/2\hbar^2$ is the energy required to totally free the mass m from M in analogy to the Hydrogen case, whereas the expectation value for the separation is:

$$\langle r\rangle_{\text{grav}} \simeq n^2 b_0 = \frac{n^2\hbar^2}{Gm^2M} \tag{12.6.9}$$

Also all the analytical solutions to the Schrödinger equation, the hydrogen wave-functions, carry over to the gravitational case with the simple substitution $a_0 \rightarrow b_0$:

$$\psi_{nlm} = R(r)\Theta(\theta)\Phi(\phi) = N_{nlm}R_{nl}Y_{lm} \tag{12.6.10}$$

where N_{nlm} is the normalization constant, R_{nl} the radial wavefunction, and Y_{lm}, the spherical harmonics, contain the angular part of the wavefunction.

Let us examine some concrete cases to obtain a feeling for these relations: For a two-body problem composed of proton and electron $b_0 \simeq 10^{29}$ [m], several orders of magnitude larger than the size of the observable universe ($\simeq 10^{26}$ [m]), whereas $E_g \simeq 10^{-78}$ [eV]. For two neutrons $b_0 \simeq 10^{22}$ [m], $E_g \simeq 10^{-68}$ [eV]. For the Earth and Sun (approximated as a two-body problem for illustrative reasons) one gets $b_0 \simeq 10^{-138}$ [m], an absurdly small ground state separation, and $E_g \simeq 10^{182}$ [J], which is unphysical as the binding energy $E_g \gg mc^2 \simeq 10^{42}$ [J]. We will see below how to deal with these "unphysical" cases and how the physical picture somewhat surprisingly is connected to the Schwarzschild radius. For a better 2-body application, let us consider a binary neutron star system (one solar mass each), $b_0 \simeq 10^{-148}$ [m], $E_g \simeq 10^{198}$ [J] $\gg mc^2 \simeq 10^{47}$ [J], again unphysical. One could also ask how much m would have to be in a gravitational binary system (taking $m = M$) in order for b_0 to be, for example, one meter: $m \simeq 10^{-19}$ [kg], or the mass of a small virus. For a pair of "Planck-objects" $m = M \simeq 10^{-8}$ [kg], we get, maybe not surprisingly, $b_0 \simeq 10^{-35}$ [m] (the "Planck length") and $E_g \simeq 10^9$ [J] (the "Planck energy") which also happens to be equal to mc^2. We could also ask for the binary system mass (again taking $m = M$) giving exactly the same numerical energy spectrum as for the Hydrogen atom, i.e. taking $E_g = E_H = 13.6$ [eV], resulting in $m \simeq 10^{-13}$ [kg], the mass of one human cell, and $b_0 \simeq 10^{-19}$ [m].

One could ask for the mass, m, required to produce exactly the quantum gravitational energy spectrum of hydrogen in a gravitational field like that of Earth, $M = M_\oplus \simeq 6\cdot 10^{24}$ [kg]. This turns out to be $m \simeq 10^{-38}$ [kg], or an equivalent mass-energy of $\sim 10^{-3}$ [eV], comparable to the conjectured mass of neutrinos. As $b_0 \sim 1\,[\mu\text{m}]$ in this case, only very highly excited states would be possible above the Earth surface. The matter would of course be quite different around cosmic compact objects, for example the conjectured "preon stars" with masses comparable to the Earth's and radii $\sim b_0$.

We notice (e.g. through b_0) that the planets in the Solar System must be in very highly excited quantum gravitational states. In that sense they are analogous to electrons in "Rydberg atoms" in atomic physics. To obtain a good two-body approximation, let us study the Sun-Jupiter system in a little more detail.

For excited states with $l \neq 0$, and very large n and l, the expectation value of the distance is:

$$\langle r\rangle \simeq \frac{1}{2}(3n^2 - l^2)b_0 \tag{12.6.11}$$

however as that is for an ensemble (average over many measurements), for a single state it is in principle more appropriate to use the most probable radial distance ("radius" of orbital):

$$\tilde{r} = n^2 b_0 \tag{12.6.12}$$

as a measure for the expected separation. However, for n large and $l = l_{\text{max}} = n-1$ the two coincide so that $\langle r \rangle = \tilde{r}$.

The angular momentum for Jupiter around the Sun (actually a G2V-type star, aka yellow dwarf) is $L \simeq 2 \cdot 10^{43}$ [J · s], giving an l-quantum number of $l = L/\hbar \simeq 2 \cdot 10^{77}$. The most probable Sun-Jupiter distance is given by $\tilde{r} = n^2 b_0 \geq l^2 b_0 \simeq 7.6 \cdot 10^{11}$ [m], which is the same as the actual separation. $E_n = -E_g/n^2 \simeq -1.6 \cdot 10^{35}$ [J], so the magnitude of the binding energy is much less than $mc^2 \simeq 1.8 \cdot 10^{44}$ [J], making it physically allowed, and also of the same order of magnitude as its classical counterpart $-GmM/r \simeq -3.4 \cdot 10^{35}$ [J]. The Sun-Jupiter system can thus seemingly be treated as a quantum gravitational 2-body system, provided that it is taken to have its maximally allowed value for its angular momentum ($l \simeq n$).

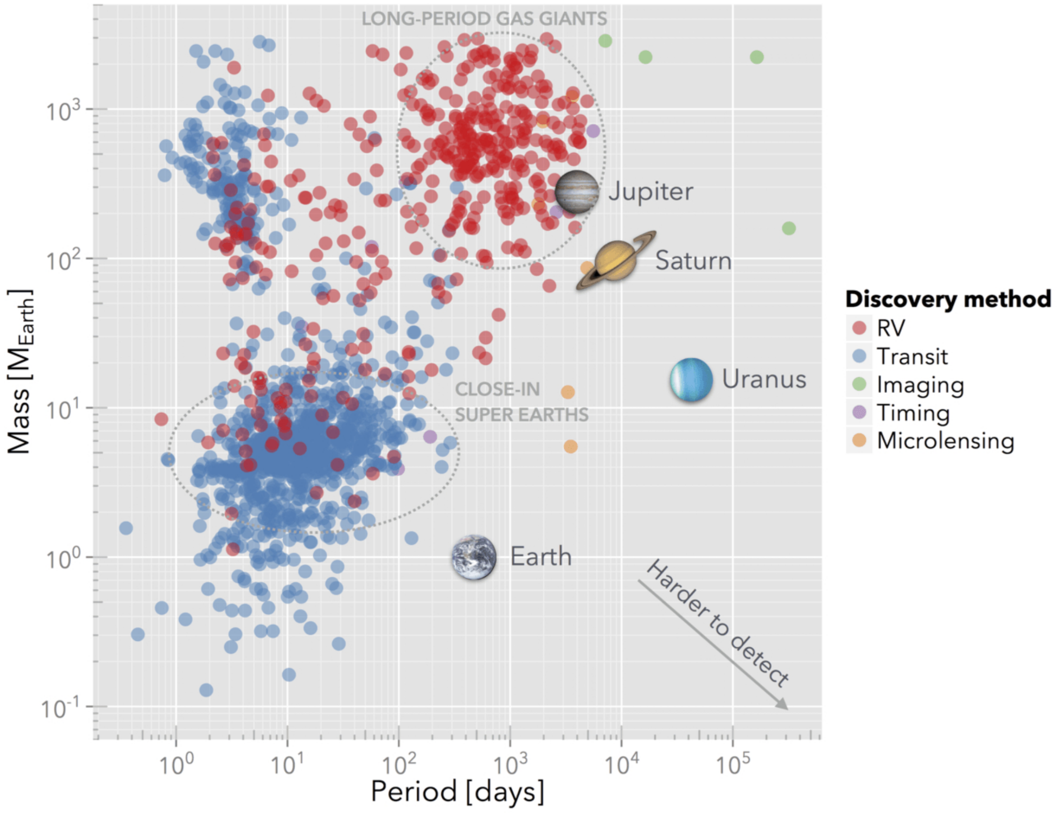


Figure 12.302: A mass-period diagram. Each dot marks the mass and orbital period of a confirmed exoplanet in 12015 (author: Stefano Meschiari)

In fact, it is easy to show that for Kepler's law to apply, l must be very close to n.

The period of revolution can be written:

$$T = \frac{2\pi m\tilde{r}^2}{L} = \frac{2\pi m\tilde{r}^2}{l\hbar} \tag{12.6.13}$$

and assuming maximality for the angular momentum, $l \simeq n$, gives:

$$T \simeq \frac{2\pi m\tilde{r}^2}{n\hbar} \tag{12.6.14}$$

Solving for n gives:

$$n = \frac{m\sqrt{GM\tilde{r}}}{\hbar} \tag{12.6.15}$$

so that:

$$T \simeq \frac{2\pi\tilde{r}^{3/2}}{\sqrt{GM}} \tag{12.6.16}$$

which exactly is Kepler's law. So, the conclusion is that all the planets in the Solar System are in maximally allowed angular momentum states quantum mechanically. The $l \simeq n$ quantum numbers are as follows: $l_{\text{Sun}} \simeq 2 \cdot 10^{75}$, $l_{\text{mercury}} \simeq 8 \cdot 10^{72}$, $l_{\text{venus}} \simeq 2 \cdot 10^{74}$, $l_{\text{Earth}} \simeq 3 \cdot 10^{74}$, $l_{\text{Mars}} \simeq 4 \cdot 10^{73}$, $l_{\text{jupiter}} \simeq 2 \cdot 10^{77}$, $l_{\text{saturn}} \simeq 8 \cdot 10^{76}$, $l_{\text{uranus}} \simeq 2 \cdot 10^{76}$, $l_{\text{neptune}} \simeq 2 \cdot 10^{76}$, $l_{\text{pluto}} \simeq 3 \cdot 10^{72}$. Even though the maximality of L and L_z are automatic from the classical description, it is far from obvious why the same should result from the more fundamental quantum treatment, as noted below.

For states with $l = l_{\max} = n - 1$ and $m = \pm l$:

1. There is only one peak, at $r = \tilde{r}$, for the radial probability density, and the "spread" (variance) in the r-direction is given by[76] $\Delta r = \sqrt{\langle r^2 \rangle - \langle r \rangle^2} \simeq n^{3/2} b_0/2$. For the Earth-Sun system it means $\Delta r \simeq 10^{-26}$ [m]

2. The angular θ-part of the wavefunction for maximal m-quantum number $|m| = l$, is $\propto \sin^l(\theta)$.

The probability density thus goes as $\sin^{2l}(\theta)$ in the θ-direction, meaning that only $\theta = \pi/2$ is non-vanishing for large l. The azimuthal (ϕ) part of the angular wavefunction Y_{lm} is purely imaginary, making it drop out of the probability density, so that *all* values of ϕ are equally likely (this ϕ-symmetry is a consequence of conservation of angular momentum in a central potential). The total planetary probability density is thus "doughnut" (torus-like) shaped, narrowly peaking around the classical trajectory.

So, at first sight, it seems like the Solar System is perfectly described as a quantum gravitational system. It even seems reasonable. Gravity totally dominates as all other forces, especially the only other known force with infinite reach, the electromagnetic, cancel due to charge neutrality. The Solar System could thus be seen as a test-vehicle for quantum gravity. In the Solar System the Sun totally dominates the gravitational field, making the central field approximation an excellent one, even though it in principle is a n-body problem. Contrast this to the case of multi-electron atoms in atomic physics where all electrons carry the same charge ($1/N$ of the charge of the nucleus), making the central field approximation a very bad one.

---

[76] The hydrogen wavefunctions for the gravitational case give $\langle r^2 \rangle = [5n^2 + 1 - 3l(l+1)]n^2 b_0^2/2$ and $\langle r \rangle = [3n^2 - l(l+1)]b_0/2$.

However, from a quantum gravity standpoint, the system could be in any and all of the degenerate states, and usually at the same time, so typical for quantum mechanical superposition. Even for given energy and angular momentum there is no reason for the planets to be in any particular eigenstate at all of the $2l+1$ allowed, and certainly not exclusively $m = \pm l$. The radial probability distribution in general has $n-l$ maxima. Thus, only for $l = l_{\text{max}} = n-1$ has it got a unique, highly peaked maximum. The degeneracy for a given n is n^2. Whenever $l < l_{\text{max}}$, the radial wave function is highly oscillatory in r as it has $n-l$ nodes. The same goes for the angular distribution as there in general are $l-m$ nodes in the θ-direction. For a general $R_{nl}Y_{lm}$ the planets could be "all over the place", and if this weren't bad enough, according to quantum mechanics the Solar System more probable than not should be in simultaneous, co-existing superposed states with different quantum numbers as is generic in atomic physics. Consequently, Newtonian quantum gravity cannot solve the quantum mechanical measurement problem, perhaps because it lacks the non-linear terms conjectured to be needed.

To get the innermost allowed physical orbit for any "test-particle", m, we must impose the physical restriction that the binding energy cannot exceed the test particle energy, thus:

$$E_g(\max) = mc^2 \tag{12.6.17}$$

As E_g can be written:

$$E_g = \frac{GmM}{2b_0} \tag{12.6.18}$$

we get:

$$b_0(\min) = \frac{GM}{2c^2} = \frac{R_S}{4} \tag{12.6.19}$$

where $R_S = 2GM/c^2$ is the Schwarzschild radius. The expression $b_0(\min)$ gives a limit for b_0 of the system to be physically attainable. It is amusing to see how close $b_0(\min)$ is to R_S and one cannot help speculate that a more complete theory of quantum gravity could ensure that $r > R_S$ always, and thus forbid Black Holes altogether[77]. The object M must be put together somehow, but if $r_{\text{min}} > R_S$ it can never accrete enough matter to become a Black Hole, as the in-falling mass (energy) instead will be radiated away in its totality (in gravitons if they exist...), making a Black Hole state impossible[78]. This would, in an unexpected way, resolve the Black Hole information loss paradox. Even though $r = R_S$ represents no real singularity, as it can be removed by a coordinate transformation, anything moving inside $r < R_S$ will, according to classical general relativity, in a (short) finite proper time reach the true singularity at $r = 0$. If quantum gravity could ensure that $r > R_S$ always, gravity would of course be singularity free.

Let us also briefly look at radiative transitions. From the dipole approximation in atomic physics an elementary quantum (photon) transition requires $\Delta l = \pm 1$. A quadrupole (graviton) approximation in quantum gravity instead requires $\Delta l = \pm 2$. So, a typical elementary energy transfer in a highly excited, gravitationally bound 2-body quantum gravitational system is:

$$\Delta E = -E_g(n^{-2} - (n-2)^{-2}) \simeq \frac{4E_g}{n^3} \tag{12.6.20}$$

---

[77] For the hydrogen atom the corresponding value is $a_0(\min) \simeq 1.4 \cdot 10^{-15}$ [m], or one-half the "classical electron radius", whereas $R_S \simeq 10^{-53}$ [m], so that $a_0(\min) \gg R_S$. But we implicitly already knew that. The Coulomb force does not turn atoms into Black Holes.

[78] For the classical case, the relation is even closer, $GmM/r_{\text{min}} = mc^2$, giving $r_{\text{min}} = R_S/2$, but then one cannot really speak of energy being carried away by gravitons.

For the Earth-Sun system this means $\Delta E \simeq 2 \cdot 10^{-20}$ [eV], carried by a graviton with frequency $\nu \simeq 5 \cdot 10^{-6}$ [Hz], and wavelength $\lambda \simeq 6 \cdot 10^{13}$ [m] $\simeq 400$ [AU] (1 AU being for recall the mean distance between the Earth and Sun).

The average time required for each elementary quantum gravity transition to take place can be estimated roughly by $\Delta t \sim \hbar/\Delta E \simeq 3 \cdot 10^4$ [s] $\simeq$ 8h20 min. Thus the power radiated by a spontaneously emitted individual graviton is very roughly $\sim 10^{-43}$ [W], compared to the prediction from the usual quadrupole formula (first non-vanishing contribution) in classical general relativity of $\simeq 300$ [W] for the total power. We also see that the gravitational force is not really conservative, even in the static Newtonian approximation, but the difference is exceedingly small in the Sun-Earth system. The changes in kinetic and potential energies do not exactly balance, $\Delta K \neq \Delta U$, the difference being carried away by gravitons in steps of $\Delta l = 2$. Also, in quantum gravity there is gravitational radiation even in the spherically symmetric case, which is forbidden according to the classical general relativistic description.

Let us now return to the experiment with neutrons in the gravitational field of the Earth, claiming to have seen, for the first time, quantum gravitational states in the potential well formed by the approximately linear gravitational potential near the Earth surface and a horizontal neutron mirror. An adjustable vertical gap between the mirror and a parallel neutron absorber above was found to be non-transparent for traversing neutrons for separations less than $\sim 15\,[\mu\text{m}]$ (essentially due to the fact that the neutron ground state wavefunction then overlaps the absorber). As the neutron in such a well, from solving the Schrödinger equation, has a ground state wavefunction peaking at $\sim 10\,[\mu\text{m}]$, with a corresponding energy of $\simeq 1.4 \cdot 10^{-12}$ [eV], the experimental result is interpreted to implicitly having verified, for the first time, a gravitational quantum state.

If we instead analyse the experiment in the framework of the present article, the same experimental setup gives $b_0 \simeq 9.5 \cdot 10^{-30}$ [m], $E_g \simeq 2.2 \cdot 10^{35}$ [eV]. Close to the Earth's surface, $\tilde{r} \simeq R_\oplus \simeq 6.4 \cdot 10^6$ [m], the radius of the Earth, giving $n \simeq 8.2 \cdot 10^{17}$, resulting in a typical energy for an elementary quantum gravity transition $\Delta E \simeq 4E_g/n^3 \simeq 1.6 \cdot 10^{-18}$ [eV]. For a cavity of $\Delta\tilde{r} = 15\,[\mu\text{m}]$, and $n \gg \Delta n \gg 1$, one gets $\Delta E = E_g b_0 \Delta\tilde{r}/\tilde{r}^2 \simeq 0.7 \cdot 10^{-12}$ [eV] $= 0.7$ [peV], to be compared to the value 1.4 [peV]. Even though the present treatment gives a similar value for the required energy, it need *not* be the result of a *single* quantum gravity state, but rather $\leq 10^6$ gravitons can be emitted/absorbed. From the treatment in this article it is thus not self-evident to see why the experimental apparatus should be non-transparent to neutrons for vertical separations $\Delta h < 15\,[\mu\text{m}]$.

To appreciate the potential importance of this, let us digress briefly on the equivalence principle, the main conceptual pillar of general relativity. A classic example for illustrating it involves two rockets: One rocket stands firmly on the surface of the Earth, while the other accelerates constantly in empty space with $a = g$. According to the equivalence principle there is no way to, locally, distinguish one from the other if one is not allowed/able to make outside observations, meaning that acceleration and gravity are equivalent. However, in the quantum gravity case, for example considering a neutron inside each rocket, there certainly *is* a difference: For the rocket standing on the Earth the potential in the Schrödinger equation is $V = -GmM/r$, resulting in normal quantization as elaborated above. For the rocket accelerating in space, however, $V = 0$ and the energy levels are non-quantized (the neutron being a free particle until it hits the floor of the rocket). So the conclusion is that Newtonian quantum gravity breaks the equivalence principle. Furthermore, to understand why a free-falling object classically accelerates radially downward in, e.g., the Earth's gravitational field, the gravitons must, by conservation of momentum, be emitted in the direction opposite to the acceleration (at least the probability for emission must peak in that direction). Also, the quantum states with given n, l, m are in principle inherently stable, an outside perturbation being needed for the transition rate to be different from zero, just like in atomic physics. For macroscopic bodies this poses no problem as there in that case are abundant backgrounds of both gravitational and non-gravitational disturbances. For an elementary quantum gravity interaction, however, this problem seems much more severe, as the notion of free-fall loses its meaning as the quantum states become practically stable to spontaneous graviton emission. In fact, a bound quantum gravitational object does not fall at all as it is described by a stationary wavefunction, or a superposition of such.

Thus, the difference regarding quantized energy levels for an experiment with neutrons "falling" under the influence of Earth's gravity *with* mirror or *without* (above) shows that Newtonian quantum gravity is dependent on global boundary conditions, where the boundary in principle can lie arbitrarily far away. This comes as no surprise, as the Schrödinger equation models the gravitational interaction as instantaneous, contrasted with the case in general relativity where the behaviour in free fall only depends on the local properties of mass-energy and the resulting space-time curvature (out of which the mirror is not part due to its inherently non-gravitational interaction with the neutron) and causal connection as the gravitational interaction propagates with the speed of light. However, as several experiments on entangled quantum states, starting with Clauser/Freedman and Aspect et al. , seem to be compatible with a non-local connection between quantum objects, this property of the Schrödinger equation does not, at least for the moment, seem to be a serious drawback for a theory of quantum gravity. One could even envisage a "delayed choice" experiment à la Wheeler, where the mirror is removed/inserted before the neutron reaches its position, meaning that we could alter the energy of the gravitons *after* they have been emitted.

12.6.2 Wheeler-DeWitt Quantum Cosmology

In this subsection we will discuss some of the most interesting, profound, and topical speculations - and our stage of knowledge in this beginning of the 121st century... (holocene calendar) - about the earliest moments of the Universe.

"Why are we here?" "Where did it all come from?" These two questions are perhaps the most fundamental to be asked by our species. In recent years, a number of physicists and cosmologists have begun to develop a new description of the ultimate system, the universe as a whole. Although the first question still lies beyond the purview of physics, "quantum cosmology," an amalgam of quantum theory and classical relativistic cosmology, now offers a scientific answer to the second.

Because the underlying physics is still cloudy, it is not clear how (or even whether) one should approach the beginning time of the Universe. In spite of the fact that the underlying physics is not yet well formulated, we will attempt to give the reader a taste for some of the most exciting and potentially important research being pursued at present. The reader should understand that we are fishing in murky waters, and that the ideas we shall discuss here are both highly speculative and not yet unambiguously formulated.

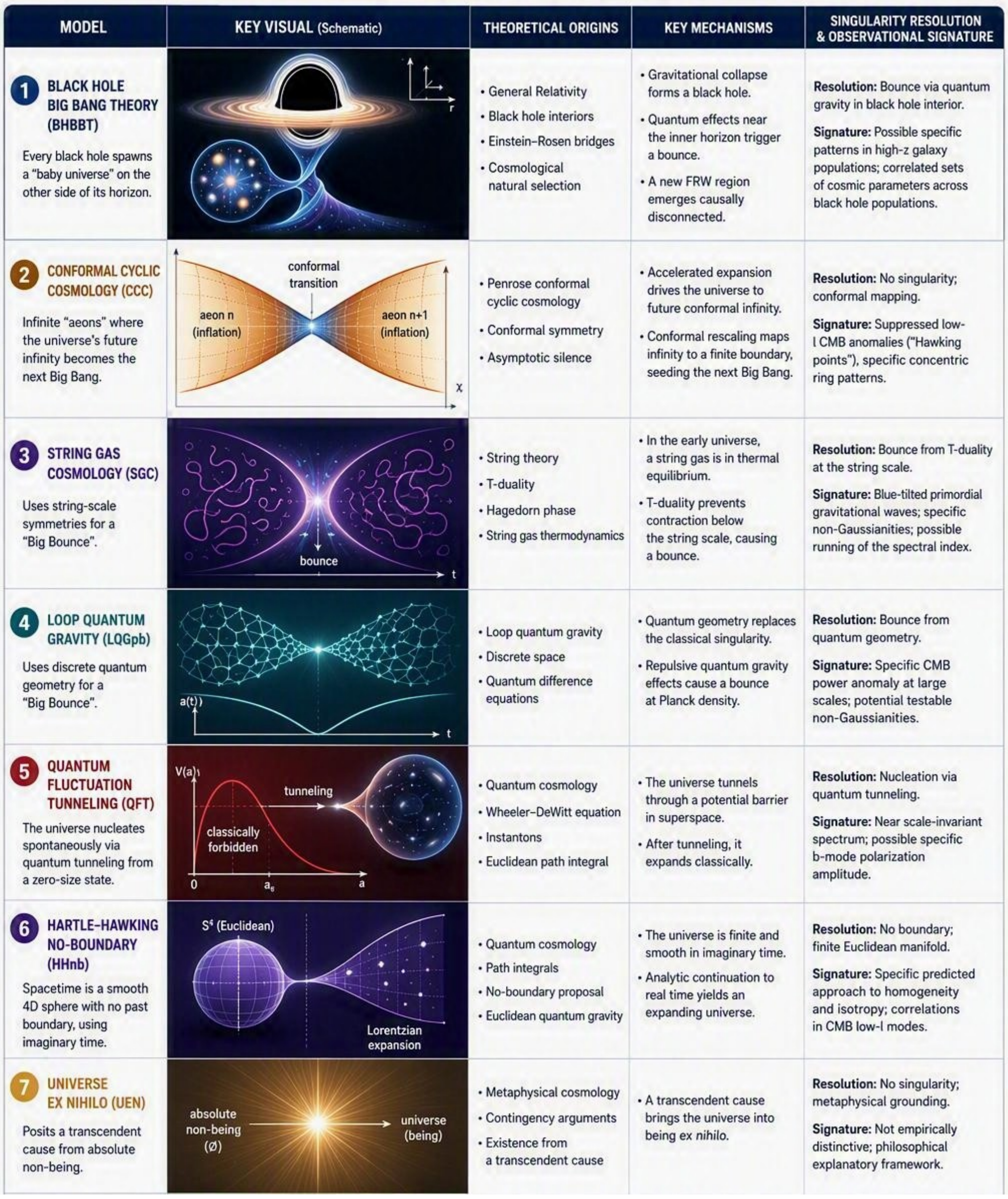
"Quantum cosmology" is the attempt in theoretical physics to develop a quantum theory of the Universe. It employs a particular (and highly speculative) approach to the unification of Quantum Physics and General Relativity that was pioneered by John Wheeler and Bryce DeWitt in the 11960s (holocene calendar). This approach attempts to answer open questions of classical physical cosmology, particularly those related to the first phases of the universe Unfortunately, their theory is difficult to interpret as it does not refer directly to evolution in time, but rather to other properties of the Universe, such as its size. Remarkably, the Wheeler-DeWitt theory can be used to calculate the probability that the Universe came into being from nothing (ex-nihilo), though doing so involves making a number of assumptions. James Hartle and Stephen Hawking attempted this in 11983 (holocene calendar), basing their thinking on the assumption that the Universe should have no boundary in time or space. Others, using different assumptions have found a different answer and it is not clear which, if any, is correct. However it is interesting that one can make progress by assuming that the laws of physics are somehow beyond the Universe itself and can describe its creation. This is in distinct contrast to the philosophy of chaotic inflation, where it is assumed that the laws of physics are very much within the Universe, and may vary from place to place.

> ❞ *Nothing can create something all the time due to the laws of quantum mechanics, and it's fascinatingly interesting!* ❞
>
> Lawrence M. Krauss

quote
12.6:6083.#73

Remark

The Wheeler-DeWitt quantum cosmology model mathematically describes a Universe out of nothing (in the quantum sense of "nothing" obviously!) and without cause (see [238]) like the Vilenkin-Higgs-Coleman-Weinberg model (see [560]). But other models give mathematical possible answers of Universe without cause like the Einstein Cyclic Universe (the Universe always existed and cycles through a sequences of Big-Bang and Big-Crunch so there is no questioning about it's beginning), the Conformal Cyclic Cosmology (the Universe also always existed but finishes with an exponential expansion destructing matter to the states of photons and therefore cancels the concept of time and dimension!), the String Gas Cosmology in which there were cosmic Strings preceding the Big Bang due to Non-singular Ekpyrotic Cosmology, the Hartle-Hawking no-boundary proposal, the Loop Quantum Gravity in which there were spin foams producing a pre-Big-Bang branch to our Universe with Black Holes remnants, the Black Hole Big Bang theory (i.e. cosmic inflation theory) is that matter from a mother universe collapses into a Black Hole that may generate a White Hole giving existence to a new universe, etc.



MODEL	KEY VISUAL (Schematic)	THEORETICAL ORIGINS	KEY MECHANISMS	SINGULARITY RESOLUTION & OBSERVATIONAL SIGNATURE
① **BLACK HOLE BIG BANG THEORY (BHBBT)** Every black hole spawns a "baby universe" on the other side of its horizon.	r	• General Relativity • Black hole interiors • Einstein–Rosen bridges • Cosmological natural selection	• Gravitational collapse forms a black hole. • Quantum effects near the inner horizon trigger a bounce. • A new FRW region emerges causally disconnected.	**Resolution:** Bounce via quantum gravity in black hole interior. **Signature:** Possible specific patterns in high-z galaxy populations; correlated sets of cosmic parameters across black hole populations.
② **CONFORMAL CYCLIC COSMOLOGY (CCC)** Infinite "aeons" where the universe's future infinity becomes the next Big Bang.	conformal transition aeon n (inflation) aeon n+1 (inflation) χ	• Penrose conformal cyclic cosmology • Conformal symmetry • Asymptotic silence	• Accelerated expansion drives the universe to future conformal infinity. • Conformal rescaling maps infinity to a finite boundary, seeding the next Big Bang.	**Resolution:** No singularity; conformal mapping. **Signature:** Suppressed low-l CMB anomalies ("Hawking points"), specific concentric ring patterns.
③ **STRING GAS COSMOLOGY (SGC)** Uses string-scale symmetries for a "Big Bounce".	bounce t	• String theory • T-duality • Hagedorn phase • String gas thermodynamics	• In the early universe, a string gas is in thermal equilibrium. • T-duality prevents contraction below the string scale, causing a bounce.	**Resolution:** Bounce from T-duality at the string scale. **Signature:** Blue-tilted primordial gravitational waves; specific non-Gaussianities; possible running of the spectral index.
④ **LOOP QUANTUM GRAVITY (LQGpb)** Uses discrete quantum geometry for a "Big Bounce".	a(t)) t	• Loop quantum gravity • Discrete space • Quantum difference equations	• Quantum geometry replaces the classical singularity. • Repulsive quantum gravity effects cause a bounce at Planck density.	**Resolution:** Bounce from quantum geometry. **Signature:** Specific CMB power anomaly at large scales; potential testable non-Gaussianities.
⑤ **QUANTUM FLUCTUATION TUNNELING (QFT)** The universe nucleates spontaneously via quantum tunneling from a zero-size state.	V(a) tunneling classically forbidden 0 a_0 a	• Quantum cosmology • Wheeler–DeWitt equation • Instantons • Euclidean path integral	• The universe tunnels through a potential barrier in superspace. • After tunneling, it expands classically.	**Resolution:** Nucleation via quantum tunneling. **Signature:** Near scale-invariant spectrum; possible specific b-mode polarization amplitude.
⑥ **HARTLE–HAWKING NO-BOUNDARY (HHnb)** Spacetime is a smooth 4D sphere with no past boundary, using imaginary time.	S^4 (Euclidean) Lorentzian expansion	• Quantum cosmology • Path integrals • No-boundary proposal • Euclidean quantum gravity	• The universe is finite and smooth in imaginary time. • Analytic continuation to real time yields an expanding universe.	**Resolution:** No boundary; finite Euclidean manifold. **Signature:** Specific predicted approach to homogeneity and isotropy; correlations in CMB low-l modes.
⑦ **UNIVERSE EX NIHILO (UEN)** Posits a transcendent cause from absolute non-being.	absolute non-being (Ø) universe (being)	• Metaphysical cosmology • Contingency arguments • Existence from a transcendent cause	• A transcendent cause brings the universe into being *ex nihilo*.	**Resolution:** No singularity; metaphysical grounding. **Signature:** Not empirically distinctive; philosophical explanatory framework.

The classical cosmology is based on General Relativity which describes the evolution of the Universe very well, as long as we do not approach the Big Bang. It is supposed that General Relativity fails to provide what must be demanded of a final theory of space and time when the conditions are extreme like at the beginning of our Universe. Therefore, a theory is needed that integrates relativity theory and quantum theory. Such an approach is attempted for instance with the Loop Quantum Gravity, the String Theory and the Causal Set Theory (they are "models" and not "theories" in reality but whatever...).

In what follows we will attempt to determine the Wheeler–DeWitt equation that is a field equation. It is part of a theory that attempts to combine mathematically the ideas of Quantum Mechanics and General Relativity, a step towards a theory of Quantum Gravity. As always in this textbook we will start with a naive approach based on the Newtonian case, after we will complexify the method to apply it to FLRW metrics, finally we will generalize to the metric of any minisuperspace metric.

Remark

A universe which originated as a vacuum fluctuation must have as we know a zero total energy. A simple semi-Newtonian argument is that the positive mass-energy of matter is balanced by their negative gravitational potential, to within a factor of order unity. However actual observations of our observable Universe with our actual technologies seems to not corroborate this hypothesis.

12.6.2.1 Wheeler-DeWitt equation for Newtonian Universe

We will use the Lagrangian formalism of Analytical Mechanics and some assumptions to obtain the cosmological differential equations analogous to Friedmann and Einstein equations, obtained from the theory of General Relativity. This method can be used to a Universe constituted of incoherent matter, that is, the cosmologic substratum is comprised of dust. It will help us to understand a more complex model that we will introduce further below for the FLRW Universe.

Quantum Newtonian cosmology is the theoretical framework within which it is possible to find the wave function which predicts the behaviour of the Newtonian universe. This wave function for Newtonian cosmology is obtained in non-relativistic quantum mechanics, and thus, all we need is to solve the Schrödinger equation for the system under consideration.

The kinetic energy for the motion of a galaxy expressed in fixed rectangular coordinates (comoving) is a function only of and, if the galaxy moves in a conservative force field, the potential energy is a function only of $\dot{R}$. Thus, we can write using the notation of Analytical Mechanics:

$$E_c=T=T(\dot{R}) \qquad E_p=U=U(R) \tag{12.6.21}$$

With the usual expressions for the kinetic and gravitational potential energies, the Lagrangian $L(q,\dot{q})$ for the motion of a galaxy of mass m in the expansion is:

$$L(R,\dot{R})=T-U=\frac{1}{2}m\dot{R}^2+\frac{GMm}{R} \tag{12.6.22}$$

where we choose the generalized coordinates $q=R$ and $\dot{q}=\dot{R}$. In the relation above M is the total mass of the universe, that we assume to be distributed in a sphere of radius R.

If we assume that there is a cosmological force on a particle (galaxy) of the gas, given as seen earlier above (page 6015) by:

$$F_\Lambda=\frac{c^2\Lambda mR}{3} \tag{12.6.23}$$

we have an additional potential energy given by:

$$U_\Lambda=-\int F_\Lambda \mathrm{d}R'=-\frac{1}{6}c^2\Lambda mR \tag{12.6.24}$$

This implies that the Lagrangian changes to:

$$L(R,\dot{R})=T-U_{\text{eff}}=T-U-U_\Lambda=\frac{1}{2}m\dot{R}^2+\frac{GMm}{R}+\frac{1}{6}c^2\Lambda mR \tag{12.6.25}$$

Now, using the Euler-Lagrange equation far a conservative system, namely:

$$\frac{\partial L}{\partial q}-\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{q}}\right)=0 \tag{12.6.26}$$

we have:

$$\begin{gathered}\frac{\partial L}{\partial R}=-\frac{GMm}{R^2}+\frac{1}{3}c^2\Lambda mR\\ \frac{\partial L}{\partial \dot{R}}=m\dot{R}\Rightarrow\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{q}}\right)=m\ddot{R}\end{gathered} \tag{12.6.27}$$

Substituting the both relations above into the Euler-Lagrange, we get the equation of motion:

$$\ddot{R}=-\frac{GM}{R^2}+\frac{1}{3}c^2\Lambda mR \tag{12.6.28}$$

Since the mass of the sphere is given by:

$$M=\frac{4}{3}\pi R^3\rho \tag{12.6.29}$$

then the equation of motion becomes:

$$\ddot{R}=-\frac{4}{3}\pi G\rho R+\frac{1}{3}c^2\Lambda mR \tag{12.6.30}$$

Therefore, this is the Newtonian cosmological equation for the scale parameter R that governs the universe expansion. This is a nice intermediate result as this equation is analogous to the Einstein Field equations obtained from theory of General Relativity, in the case $P=0$, that is, dust cloud.

According our previous assumption the time is homogeneous within an inertial reference frame. Therefore, the Lagrangian that describes a closed system, i.e., a system not interacting with anything outside the system, cannot depend explicitly on the time. In our case, the Lagrangian is likewise independent of the time, because the system is under the action of a uniform force field. Thus, the constant quantity of the motion is H, the Hamiltonian of the system, which is for recall given by:

$$H=p\dot{q}-L=\left(\frac{\partial L}{\partial \dot{q}}\right)\dot{q}-L \tag{12.6.31}$$

As re-parametrization invariance implies vanishing Hamiltonian (see section Analytical Mechanics page 3182) we must have:

$$H=\left(\frac{\partial L}{\partial \dot{R}}\right)\dot{R}-L=0 \tag{12.6.32}$$

Therefore:

$$H_{\text{Newt.}}=m\dot{R}\dot{R}-\left(\frac{1}{2}m\dot{R}^2+\frac{GMm}{R}+\frac{1}{6}c^2\Lambda mR\right)=0 \tag{12.6.33}$$

Regrouping we get:

$$H_{\text{Newt.}}=\frac{1}{2}m\dot{R}^2-\frac{GMm}{R}-\frac{1}{6}c^2\Lambda mR=0 \tag{12.6.34}$$

Written in term of the momentum:

$$p=\left(\frac{\partial L}{\partial \dot{R}}\right)=m\dot{R} \tag{12.6.35}$$

we get:

$$H_{\text{Newt.}}=\frac{1}{2}\frac{p^2}{m}-\frac{GMm}{R}-\frac{1}{6}c^2\Lambda mR=0 \tag{12.6.36}$$

Let us transform now that latter relation into an Hermitian operator such that we can build something of the form:

$$H\Psi=0 \tag{12.6.37}$$

also often denoted $\hat{H}(x)|\Psi\rangle=0$. Although the symbols $\hat{H}$ and $|\Psi\rangle$ may appear familiar, their interpretation in the Wheeler-DeWitt equation is substantially different from non-relativistic quantum mechanics. $|\Psi\rangle$ is no longer a spatial wave function in the traditional sense of a complex-valued function that is defined on a 3 dimensional space-like surface and normalized to unity. Instead it is a functional of field configurations on all of space-time! This wave function contains all of the information about the geometry and matter content of the universe. $\hat{H}$ is still an operator that acts on the Hilbert space of wave functions, but it is not the same Hilbert space as in the non-relativistic case, and the Hamiltonian no longer determines evolution of the system.

This property is known as "timelessness". The static Cosmos implied by the Wheeler-DeWitt equation $H\Psi=0$ seems quite obvious once one realizes that time is an emergent concept rather than a fundamental one (this theoretical result disturbs even some professional physicists and even some published scientific papers are dedicated only to this topic like [303]). As the Cosmos doesn't have any external thing that can makes it evolve the timelessness implies some violation of conservation laws.

The "problem of time" in present physics substantially consists in the fact that a straightforward quantization of the general relativistic evolution equation and constraints generates for the Universe wave function the Wheeler-De Witt equation, which describes a static Universe! Some physiciscts have considered the fact that there exist states of a system composed by entangled subsystems that are stationary, but one may interpret the component subsystems as evolving: this leads them to suppose that the global state of the universe can be envisaged as one of this static entangled state, whereas the state of the subsystems can evolve. Such system seems to have been verified like presented in [398].

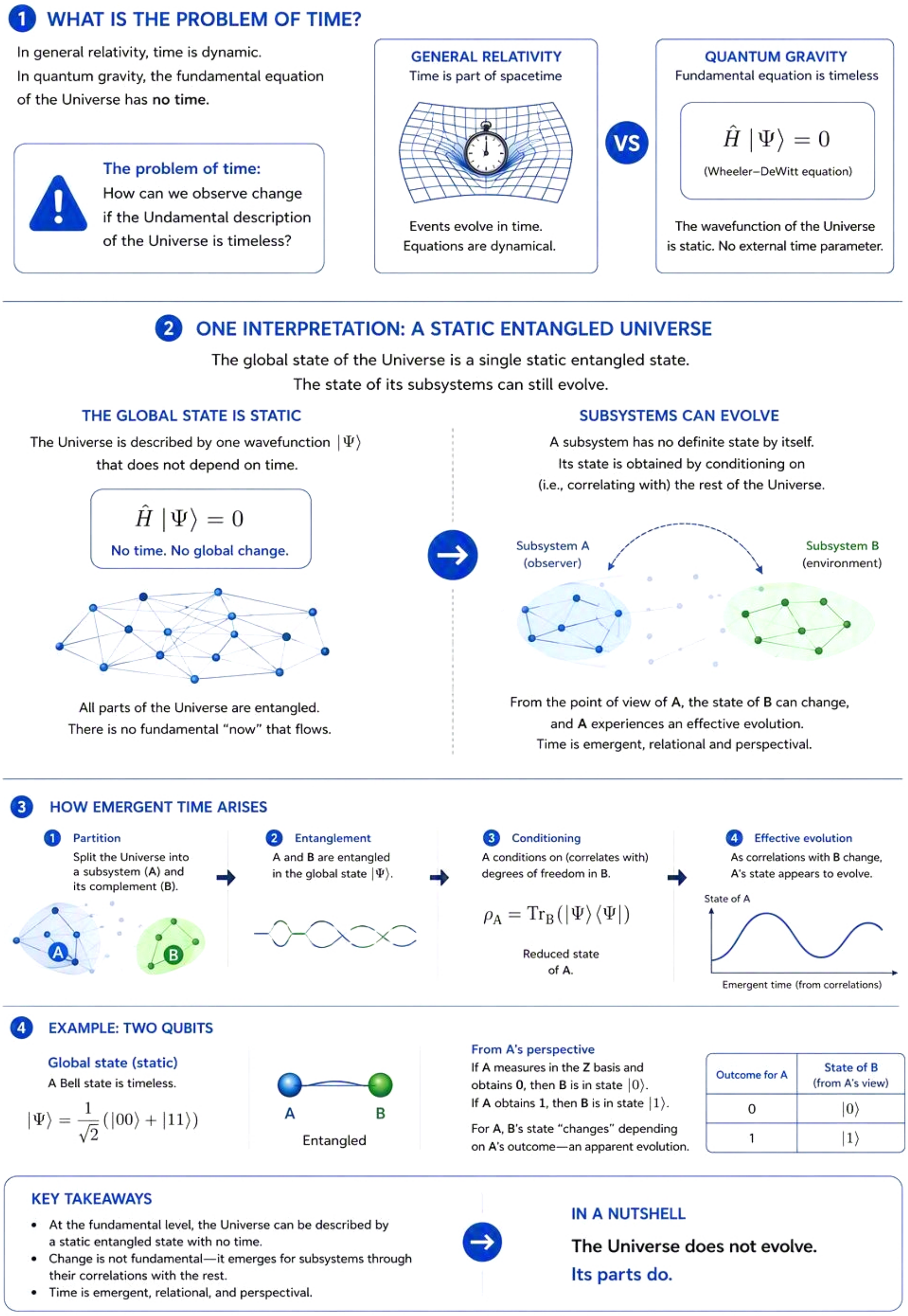
This property suggests that the universe as a whole may be timeless, and what we perceive as time is a local, emergent phenomenon that is not fundamental because as we know from General Relativity there's no universal clock!

Remark

The "universe wave function" is a function of all the possible positions of all the particles in the universe. As we know the wave function of a single particle (without including special properties like the spin) is a function $\psi(x, y, z)$ and the wave function of two particles (still without including special properties like the spin) is $\psi(x1, y2, z1, x2, y2, z2)$. Just bare in mind that the wave function of all the particles of the observable universe is estimated actually to be a wave function of about 10^{80} variables.

quote
12.6:6088.#74

> *When people ask me if a god created the universe, I tell them that the question itself makes no sense. Time didn't exist before the big bang, so there is no time for god to make the universe in. It's like asking directions to the edge of the earth; The Earth is a sphere; it doesn't have an edge; so looking for it is a futile exercise. [...]*
>
> Stephen Hawking



1 WHAT IS THE PROBLEM OF TIME?

In general relativity, time is dynamic.
In quantum gravity, the fundamental equation of the Universe has **no time.**

The problem of time:
How can we observe change if the Undamental description of the Universe is timeless?

GENERAL RELATIVITY
Time is part of spacetime

Events evolve in time.
Equations are dynamical.

VS

QUANTUM GRAVITY
Fundamental equation is timeless

$$\hat{H}\,|\Psi\rangle = 0$$

(Wheeler–DeWitt equation)

The wavefunction of the Universe is static. No external time parameter.

2 ONE INTERPRETATION: A STATIC ENTANGLED UNIVERSE

The global state of the Universe is a single static entangled state.
The state of its subsystems can still evolve.

THE GLOBAL STATE IS STATIC

The Universe is described by one wavefunction $|\Psi\rangle$ that does not depend on time.

$$\hat{H}\,|\Psi\rangle = 0$$

No time. No global change.

All parts of the Universe are entangled.
There is no fundamental “now” that flows.

SUBSYSTEMS CAN EVOLVE

A subsystem has no definite state by itself.
Its state is obtained by conditioning on (i.e., correlating with) the rest of the Universe.

From the point of view of **A**, the state of **B** can change, and **A** experiences an effective evolution.
Time is emergent, relational and perspectival.

3 HOW EMERGENT TIME ARISES

1. **Partition** — Split the Universe into a subsystem (A) and its complement (B).
2. **Entanglement** — A and B are entangled in the global state $|\Psi\rangle$.
3. **Conditioning** — A conditions on (correlates with) degrees of freedom in B.

$$\rho_{\mathrm{A}} = \mathrm{Tr}_{\mathrm{B}}(|\Psi\rangle\langle\Psi|)$$

Reduced state of A.

4. **Effective evolution** — As correlations with B change, A's state appears to evolve.

4 EXAMPLE: TWO QUBITS

Global state (static)
A Bell state is timeless.

$$|\Psi\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$$

Entangled

From A's perspective
If **A** measures in the **Z** basis and obtains **0**, then **B** is in state $|0\rangle$.
If **A** obtains **1**, then **B** is in state $|1\rangle$.

For **A**, **B**'s state “changes” depending on A's outcome—an apparent evolution.

Outcome for A	State of B (from A's view)
0	$\|0\rangle$
1	$\|1\rangle$

KEY TAKEAWAYS

- At the fundamental level, the Universe can be described by a static entangled state with no time.
- Change is not fundamental—it emerges for subsystems through their correlations with the rest.
- Time is emergent, relational, and perspectival.

IN A NUTSHELL

The Universe does not evolve.
Its parts do.

For this purpose we make the famous replacement:

$$p \to -i\hbar\frac{\partial}{\partial R} \tag{12.6.38}$$

Then the "Hamiltonian operator for a particle (or galaxy or whatever) moving in a Newtonian Universe" is given by :

$$H_{\text{Newt.}}(R) = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial R^2} - \frac{GMm}{R} - \frac{1}{6}c^2\Lambda m R \tag{12.6.39}$$

Therefore $H(a)\Psi(a)=0$ will be written under the form of the "Wheeler-DeWitt equation for a Newtonian Universe":

$$\boxed{\left\{-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial a^2}-\frac{GMm}{R}-\frac{1}{6}c^2\Lambda mR\right\}\Psi(a)=0} \qquad (12.6.40)$$

where we will consider that the wave function depends only on the scale factor R.

If we rewrite the above relation as:

$$\left\{-\hbar^2\frac{\partial^2}{\partial a^2}+E_p\right\}\Psi(a)=0 \qquad (12.6.41)$$

we recognise here a one-dimensional time-independent Schrödinger where the potential energy is given by:

$$E_p=-\frac{GMm}{R}-\frac{1}{6}c^2\Lambda mR \qquad (12.6.42)$$

12.6.2.2 Wheeler-DeWitt equation for FLRW Universe

Presented below is a derivation of the Wheeler-DeWitt equation in the minisuperspace approximation[79] which also includes matter and radiation and arbitrary values of k.

Let us recall that we have derived earlier (see page 5987) above the following form of the general Friedmann equations assuming a Friedmann-Lemaître-Robertson-Walker (FLRW) Universe:

$$\begin{aligned}H^2=\left(\frac{\dot{a}}{a}\right)^2&=\frac{8\pi G}{3c^2}(\rho+\rho_{\text{vac}})-\frac{kc^2}{a^2}\\ \frac{\ddot{a}}{a}&=\frac{8\pi G}{c^2}\rho_{\text{vac}}-\frac{\dot{a}^2}{a^2}-\frac{kc^2}{a^2}-\frac{8\pi G}{c^2}P\end{aligned} \qquad (12.6.43)$$

Let us also recall the Euler-Lagrange equation for a conservative system (see section Analytical Mechanics page 3169):

$$\frac{\partial L}{\partial q_i}-\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}_i}=0 \qquad (12.6.44)$$

Now let suppose we put empirically (trial and error) a density Lagrangian (remember that the Hilbert-Einstein action deals with density Lagrangian!) for the FLRW Universe of the following form[80]:

$$\mathcal{L}=-\kappa a^3\left[\left(\frac{\dot{a}}{a}\right)^2-\frac{kc^2}{a^2}+\frac{8\pi G}{9c^2}(\rho+\rho_{\text{vac}})\right] \qquad (12.6.45)$$

where:

$$\kappa\propto\frac{c^2}{G} \qquad (12.6.46)$$

---

[79] An approximation which is sometimes taken is to only consider the largest wavelength modes of the order of the size of the Universe when studying cosmological models. This is the "minisuperspace approximation".

[80] In facts it can be derived from the Einstein-Hilbert action with matter and cosmological constant and using the FLRW Ricci scalar.

Remark

In some textbooks a simplified version of the above density Lagrangian can be found. First by choosing $k=1$ and distributing $-a^3$ in the square brackets and defining $1/a_0=\frac{8\pi G}{9c^2}(\rho+\rho_{\text{vac}})$:

$$\mathcal{L}=\kappa\left[-\dot{a}^2a+ac^2-\frac{a^3}{a_0}\right]=\kappa\left[-\dot{a}^2a+a\left(c^2-\frac{a^2}{a_0}\right)\right] \tag{12.6.47}$$

and in the worst case where the author write it in natural units, it simplifies to:

$$\mathcal{L}=\kappa'\left[-\dot{a}^2a+a\left(1-\frac{a^2}{a_0'}\right)\right] \tag{12.6.48}$$

With this the Lagrangian density has the unit of energy. Just let check that by rewriting the relation above but expliciting the units:

$$[\text{J}]=\kappa[\text{m}^3]\left[[\text{s}^{-2}]+[\text{s}^{-2}]+[\text{s}^{-2}][\text{s}^{-2}]\right]=\kappa[\text{m}^3\cdot\text{s}^{-2}] \tag{12.6.49}$$

So to transform that into an energy κ must have the units of $[\text{kg}\cdot\text{m}^{-1}]$ and this is exactly the units of $3\pi c^2/4G$ (remember that the units of G are $[\text{m}^3\cdot\text{kg}^{-1}\cdot\text{s}^{-2}]$).

We have for the momentum:

$$p=\frac{\partial\mathcal{L}}{\partial\dot{a}}=-\kappa 2a\dot{a} \tag{12.6.50}$$

And:

$$\frac{\partial\mathcal{L}}{\partial a}=-\kappa\frac{\partial}{\partial a}\left[a\dot{a}^2-kc^2a+a^3\frac{8\pi G}{9c^2}(\rho+\rho_{\text{vac}})\right]=-\frac{\kappa}{k^{3/2}}\left(\dot{a}^2-kc^2+a^2\frac{8\pi G}{3c^2}(\rho+\rho_{\text{vac}})\right) \tag{12.6.51}$$

Substituting into the Euler-Lagrange equation we get:

$$\frac{\partial\mathcal{L}}{\partial a}-\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial\mathcal{L}}{\partial\dot{a}}=-\kappa\left(\dot{a}^2-kc^2+a^2\frac{8\pi G}{3c^2}(\rho+\rho_{\text{vac}})\right)-\frac{\mathrm{d}}{\mathrm{d}t}(-\kappa 2a\dot{a})=0 \tag{12.6.52}$$

i.e. (we see that we can eliminate the unit normalisation constant and the scale factor):

$$-\left(\dot{a}^2-kc^2+a^2\frac{8\pi G}{3c^2}(\rho+\rho_{\text{vac}})\right)+\frac{\mathrm{d}}{\mathrm{d}t}(2a\dot{a})=0 \tag{12.6.53}$$

We assume that there is no time dependant variable (is one of the many plausible weakness of this model). Therefore:

$$-\dot{a}^2+kc^2-a^2\frac{8\pi G}{3c^2}(\rho+\rho_{\text{vac}})=0 \tag{12.6.54}$$

If we divide both side by a^2 and rearrange we get:

$$\left(\frac{\dot{a}}{a}\right)^2=\frac{8\pi G}{3c^2}(\rho+\rho_{\text{vac}})-\frac{kc^2}{a^2} \tag{12.6.55}$$

So we fall back on the velocity equation (12.6.43) and this justifies our initial choice!

Now let us recall the hamiltonien (see section Analytical Mechanics page 3182):

$$H(q_i,p_i)=\sum_{i=1}^{n}\dot{q}_i p_i-L(q_i,\dot{q}_i) \tag{12.6.56}$$

Therefore in the case of one variable:

$$H(q,p)=\dot{q}p-L(q,\dot{q}) \tag{12.6.57}$$

That is written in the current context:

$$\begin{aligned}\mathcal{H}_{\text{FLRW}}(a,p)&=\dot{a}p-\mathcal{L}(a,\dot{a})=\dot{a}(-\kappa 2a\dot{a})-\left[-\kappa a^3\left[\left(\frac{\dot{a}}{a}\right)^2-\frac{kc^2}{a^2}+\frac{8\pi G}{9c^2}(\rho+\rho_{\text{vac}})\right]\right]\\&=\kappa a^3\left[-2\left(\frac{\dot{a}}{a}\right)^2+\left(\frac{\dot{a}}{a}\right)^2-\frac{kc^2}{a^2}+\frac{8\pi G}{9c^2}(\rho+\rho_{\text{vac}})\right]\\&=\kappa a^3\left[-\left(\frac{\dot{a}}{a}\right)^2-\frac{kc^2}{a^2}+\frac{8\pi G}{9c^2}(\rho+\rho_{\text{vac}})\right]\\&=-\kappa a^3\left[\left(\frac{\dot{a}}{a}\right)^2+\frac{kc^2}{a^2}-\frac{8\pi G}{9c^2}(\rho+\rho_{\text{vac}})\right]\end{aligned} \tag{12.6.58}$$

As re-parametrization invariance implies vanishing Hamiltonian (see section Analytical Mechanics page 3182) we have:

$$\mathcal{H}_{\text{FLRW}}(a,p)=-\kappa a^3\left[\left(\frac{\dot{a}}{a}\right)^2+\frac{kc^2}{a^2}-\frac{8\pi G}{9c^2}(\rho+\rho_{\text{vac}})\right]=0 \tag{12.6.59}$$

Written in term of the momentum and dividing both sides by $-\kappa a^3$:

$$p=\frac{\partial\mathcal{L}}{\partial\dot{a}}=-\kappa 2a\dot{a}\Leftrightarrow\left(\frac{\dot{a}}{a}\right)^2=\frac{p^2}{\kappa^2 4a^4}\propto\frac{G^2p^2}{c^4 4a^4} \tag{12.6.60}$$

We get:

$$\mathcal{H}_{\text{FLRW}}(a,p)\propto\frac{G^2p^2}{c^4 4a^4}+\frac{kc^2}{a^2}-\frac{8\pi G}{9c^2}(\rho+\rho_{\text{vac}})=0 \tag{12.6.61}$$

Let us simplify it a bit more:

$$\mathcal{H}_{\text{FLRW}}(a,p)\propto p^2+\frac{4a^2c^2}{G^2}\left[kc^2-\frac{8\pi Ga^2}{9c^2}(\rho+\rho_{\text{vac}})\right]=0 \tag{12.6.62}$$

Let us also transform now that latter relation into an Hermitian operator such that we can build something of the form:

$$\mathcal{H}\Psi=0 \tag{12.6.63}$$

For this purpose we make the famous replacement:

$$p \rightarrow -\mathrm{i}\hbar\frac{\partial}{\partial a} \qquad (12.6.64)$$

Then the "Hamiltonian operator for a particle (or galaxy or whatever) moving in a FLRW Universe" is given by :

$$\mathcal{H}_{\text{FLRW}}(a) \propto -\hbar^2\frac{\partial^2}{\partial a^2}+\frac{4a^2c^2}{G^2}\left[kc^2-\frac{8\pi G a^2}{9c^2}(\rho+\rho_{\text{vac}})\right]=0 \qquad (12.6.65)$$

Therefore $\mathcal{H}(a)\Psi(a)=0$ will be written under the form of the "Wheeler-DeWitt equation for a FLRW Universe":

$$\boxed{\left\{-\hbar^2\frac{\partial^2}{\partial a^2}+\frac{4a^2c^2}{G^2}\left[kc^2-\frac{8\pi G a^2}{9c^2}(\rho+\rho_{\text{vac}})\right]\right\}\Psi(a)=0} \qquad (12.6.66)$$

where we will consider that the wave function depends only on the scale factor a.

Remark

In some papers or textbooks the above relation can be found in the following shape:

$$\left\{\frac{d^2}{da^2}-\frac{9\pi c^4a^2}{4\hbar^2G^2}\left[kc^2-\frac{8\pi G a^2}{3c^2}(\rho+\rho_{vac})\right]\right\}\Psi(a)=0 \qquad (12.6.67)$$

Let us consider the expression for the energy density introduced earlier:

$$\rho=A_\omega a^{-3(\omega+1)} \qquad (12.6.68)$$

where for recall $A_\omega=\rho_{\omega 0}a_0^{3(\omega+1)}$ with $\rho_{\omega 0}$ being the value of ρ_ω at present time, and ω is such that:

$$\omega=\begin{cases}0 \text{ for dust (matter predominance; } \rho_m)\\ \dfrac{1}{3} \text{ for radiation (radiation predominance; } \rho_r)\\ -1 \text{ for de Sitter (false vacuum; } \rho_v)\end{cases} \qquad (12.6.69)$$

The energy density of the vacuum, ρ_{vac}, can be expressed, in terms of the cosmological constant, in the following form:

$$\rho_{\text{vac}}=\frac{\Lambda c^4}{8\pi G} \qquad (12.6.70)$$

Now, substituting all this in the WdW equation above we get:

$$-\hbar^2\frac{\mathrm{d}^2\Psi(a)}{\mathrm{d}a^2}+U_{\text{eff}}(a)\Psi(a)=0 \qquad (12.6.71)$$

where:

$$U_{\text{eff}}(a)=\frac{9\pi^2c^6k}{4G^2}\left[a^2-\frac{1}{k}\left(\frac{\Lambda}{3}a^4+\frac{8\pi G A_\omega}{3c^4}a^{1-3\omega}\right)\right] \qquad (12.6.72)$$

If we rewrite the above relation as:

$$\left\{-\hbar^2\frac{\partial^2}{\partial a^2}+E_p\right\}\Psi(a)=0 \tag{12.6.73}$$

we recognise here a one-dimensional time-independent Schrödinger equation for a particle with $1/2$ of the unit mass and where:

$$E_p \propto \frac{4a^2c^2}{G^2}\left[kc^2-\frac{8\pi Ga^2}{9c^2}(\rho+\rho_{\text{vac}})\right] \tag{12.6.74}$$

Note that for small values of a we can neglect the terms proportional to a^4 and therefore:

$$E_p \propto \frac{4a^2c^2}{G^2}kc^2 \tag{12.6.75}$$

In this case, the solution of the WdW equation is given in terms of Bessel functions (see section Sequences and Series page 1768) and the wave function approaches a constant for $a\rightarrow 0$. On the other hand, for arbitrary values of a, the potential is not so simple, and as a consequence, the solutions of the WdW equation are more complicated and given in terms of Heun functions (see [467]).

Note that when $a=0$ the universe corresponds to a quantum FRW universe with zero radius. In this case, the solution of the WdW equation approaches a constant and describes a state named "nothing[81]" in some literature. This state can be generated by quantum mechanically tunnelling through the potential barrier that appears at $a=a_0\neq 0$. And the reader must really keep in mind that quantum tunnelling can be crudely understood as a quantum effect that result from the Heisenberg uncertainty principle: a temporary upward fluctuation in the particle's energy sends it over the barrier.

Remark

For the readers interested in further reading, we strongly recommend the following references: [559] and [314] and finally [558].

Now let us investigate a very simple special case of:

$$\left\{-\hbar^2\frac{\partial^2}{\partial a^2}+\frac{4a^2c^2}{G^2}\left[kc^2-\frac{8\pi Ga^2}{9c^2}(\rho+\rho_{\text{vac}})\right]\right\}\Psi(a)=0 \tag{12.6.76}$$

Focusing on the potential energy term for $k=+1$ (closed three-sphere Universe) and in natural units:

$$E_p \propto 4a^2\left(1-\frac{8\pi a^2}{9}(\rho+\rho_{\text{vac}})\right) \tag{12.6.77}$$

With a slight change of notation (we discard the factor 4 on the way...):

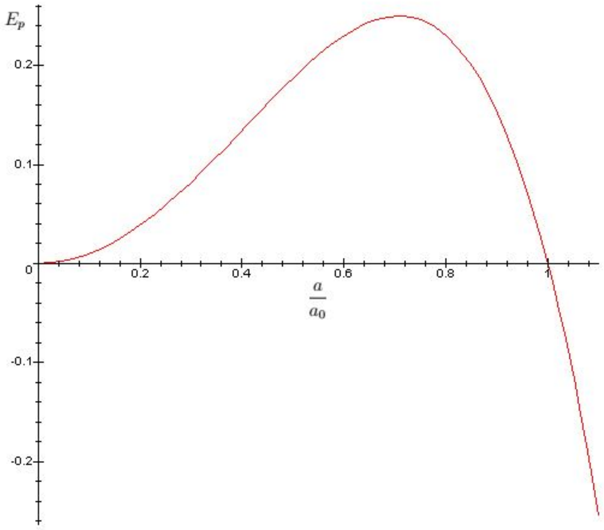
$$E_p \propto a^2\left(1-\frac{a^2}{a_0^2}\right)=4a^2\left(1-\frac{a^2}{a_0^2}\right)=a_0^2\left(\frac{a^2}{a_0^2}-\frac{a^4}{a_0^2}\right)=a_0^2\left[\left(\frac{a}{a_0}\right)^2-\left(\frac{a}{a_0}\right)^4\right] \tag{12.6.78}$$

---

[81] As we already know it from our study of Probabilities and we will see now form the Quantum Physics point of view the nothingness in the philosophical sense is a physical impossibility. So, a scientist who talks about "nothingness" is actually talking about a quantum vacuum.

The shape of the potential is shown in the figure below, where, for convenience. The quantized FRWL universe is thus mathematically equivalent to a simple one-dimensional problem in non-relativistic quantum mechanics-the "particle" at position a represents a universe with that value scale factor.

The region beneath the barrier, $0 < a < a_0$, is classically forbidden to the zero-energy particle; the region $a \geq a_0$ is classically allowed.



The fact that the Big Bang (our Universe) can pop-up from space itself like this is a change of state (or "phase") named a "false vacuum decay". Such a process seems to have been confirmed experimentally at low energies (hopefully...)! For more about such experiments the reader can refer to the papers *False vacuum decay via bubble formation in ferromagnetic superfluids* (see [596]) and *Experimental demonstration of metamaterial "multiverse" in a ferrofluid* (see [502]) where in the latter it is mentioned in the abstract that: «*Thus, thermal fluctuations in a ferrofluid look similar to creation and disappearance of individual Minkowski space-times (universes) in the cosmological multiverse. This theoretical picture is supported by experimental measurements of polarization-dependent optical transmission of a cobalt based ferrofluid at* $1,500$ *[nm]*».

It is worth noting that if our Universe was infinitely old, it would be in a true vacuum. However our Universe has a finite age. At the beginning of it, at the Big Bang, the energy per unit volume was too high so our Universe was neither in the true nor in the false vacuum. Then, as our universe expanded and cooled down, it might have moved to a false vacuum instead of a true one.

Remark

The Wheeler-DeWitt approach to false vacuum decay is often considered more pertinent than the Coleman-De Luccia (CDL) model because it provides a fully quantum-mechanical treatment of the gravitational and scalar field degrees of freedom, whereas the CDL model relies on a semi-classical, Euclidean instanton approximation. While the CDL framework assumes a fixed, classicized background geometry to calculate tunneling rates via a wick-rotated saddle-point evaluation, it breaks down when quantum fluctuations of the metric become large or when the energy barriers approach the Planck scale. By contrast, solving the Wheeler-DeWitt equation the operator constraint equation of canonical quantum gravity, $H\Psi=0$ allows one to determine the wave function of the universe directly, capturing the non-perturbative quantum behavior of space-time itself during the tunneling event and offering a more robust, foundational description of the true quantum-gravitational regime where classical geometry does not exist.

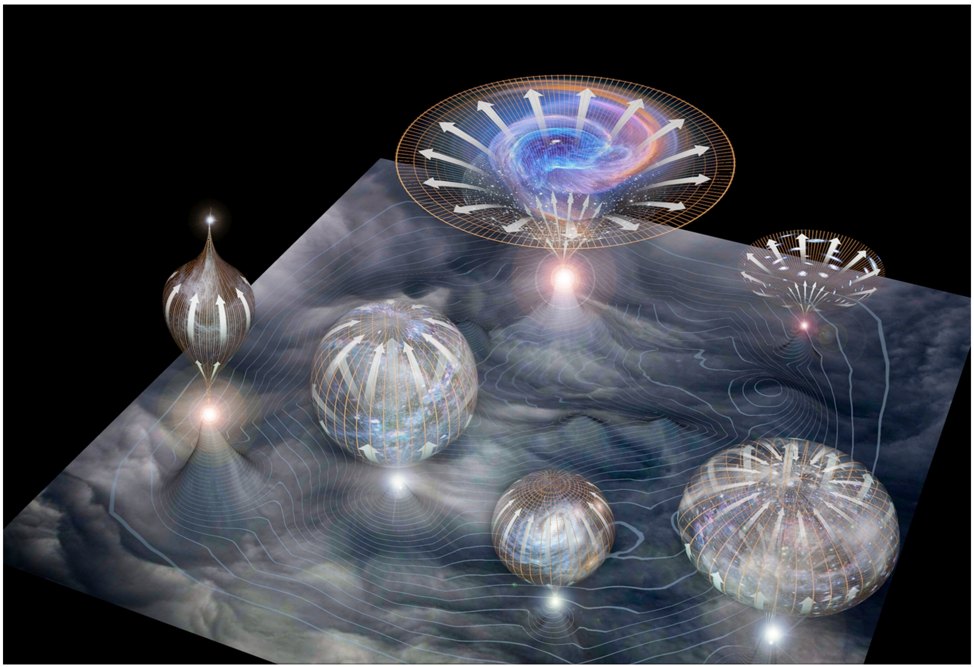


Figure 12.303: **Various universes popping-up from the Cosmos false vacuum (artist: Malcolm Godwin)**

In principle any solution of the Wheeler-DeWitt equation corresponds to a possible quantum state of the universe. It is also clear that the effects of the boundary conditions on the wavefunction will act to severely restrict the class of possible solutions. In ordinary quantum mechanics these are determined by the physical context of the problem and some set of external conditions. In the case of the universe as a whole the situations is less clear, and in many approaches some suitable set of boundary conditions are postulated instead, based on general arguments involving concepts such as simplicity or economy.

Although the above developments sounds like a blatant violation of energy conservation, it is in fact consistent with physics as we know it.

The resolution to the energy paradox lies in the subtle behaviour of gravity. Although it has not been widely appreciated, Newtonian physics unambiguously implies that the energy of a gravitational field is always negative a fact which holds also in General Relativity. The Newtonian argument closely parallels the derivation of the energy density of an electrostatic field, except that the answer has the opposite sign because the force law has the opposite sign: two positive masses attract, while two positive charges repel. The possibility that the negative energy of gravity could balance the positive energy for the matter of the Universe may have been suggested for the first time as early as 11932 (holocene calendar) by Richard Tolman.

So while the energy of matter increases by a factor, the energy of the gravitational field becomes more and more negative to compensate. The total energy - matter plus gravitational - remains constant and very small, and could even be exactly zero. Conservation of energy places no limit on how much the Universe can inflate, as there is no limit to the amount of negative energy that can be stored in the gravitational field.

We will go further in the topic the day we will have a gentle introduction to Heun's differential equations in this textbook.

12.6.3 Why our Universe is a quite massive unpredictable mess

With all the knowledge accumulated so far in fluid mechanics, thermodynamics, electromagnetism, quantum physics, astronomy, astrophysics and cosmogony the reader now can face it!: our (observable) universe is actually messy and also was long time in the past and will likely be for a given time in the future. It is (actually) non-linear, turbulent and chaotic! Our universe isn't "perfect" as some people argue (and by the way... how to these people using this argument quantify "perfectness"...?).

As we already know, Chaos Theory is a branch of physics that describes the evolution of processes that are dependent on initial conditions. Some processes, which at first glance appear to be random, can in fact be based on underlying patterns and interdependencies. In quantum physics however some processes are purely random and without cause as we already know and don't have underlying pattern (i.e. Bell inequalities).

The classic example of a chaotic system is termed as we know the "butterfly effect" (see page 8342): Almost negligible changes in initial conditions can have dramatic and unpredictable effects on the evolution of a chaotic system.

There are several areas of astronomy that benefit from an understanding of Chaos Theory. The obvious one is the gravitational interaction of celestial bodies. Such a system is chaotic,and a tiny error in the positions or velocities of the celestial bodies, or the introduction of small perturbations, means the system is unpredictable in the long term (like most of three body systems, see page).

Weather and climate here on Earth are good examples of chaotic systems (see page 8239). Similarly, features such as the Great Red Spot on Jupiter are also chaotic. Although scientists have a good grasp of fluid dynamics and the forces that result in this massive storm system, its formation and evolution cannot be predicted from these principles.

Most of the processes going on in our Universe are actually (and also for a long time in the future) essentially chaotic. The acceleration of charged particles, the creation of cosmic rays, the structure of magnetic fields, nuclear reactions within stars, chemical reactions in interstellar space, and many more phenomena,are all critically dependent on initial conditions. Ultimately, it would be fair to say that our observable Universe is in its actual state chaotic and therefore unpredictable!

Not only do we have 103 well known list of factors that can make a (business) project fail on Earth (basics of risk management applied by risk managers). Similarly because of various completely random and non-random sources, you can die at any moment, have one of the 160 cancers known so far or one of the $4,200$ genetic anomalies and many other probabilistic stuff that show our body and life on Earth isn't perfect (see table of famous life-event probabilities at page 474).

But we have also bigger natural (random or non-random, not human made) events that can at any day (or at some dates quite well known in the future) completely or significantly eradicate humanity or any other life on Earth or on our whole Solar System for some of them, like:

- Pandemia
- Giant solar flare
- Super-volcano eruption
- Impact with celestial body (rogue planet, rogue star, rogue black hole, asteroids, etc.)
- Increase of the size/temperature of our Sun
- Ejection of planetary nebulae (our Sun death)
- Near-Earth supernova
- Decreasing time of Earth rotation (hence increase of temperature)
- Solar system orbits (chaotic) perturbations
- Extinction of Earth magnetic field
- Important climate change (ozone layer)
- Gamma-ray burst (form quasar, neutron/magnetar star)
- Universe wide quantum vacuum decay
- Space-Time universe big rip
- Universe Big Freeze
- Ecological collapse
- Half-life of particles (weak interaction)
- ...

12.7. String Theory

It must be considered in this section that "String Theory" (and verbatim Superstring Theory) is currently - early 121st century (holocene calendar) - speculative and could not be verified (confirmed) or falsified by experience following the scientific method[82]. We should therefore take with caution the developments that follow and to be the most critical possible!

It is also a speculative theory (we can not even use the word "model" yet...) of unification of the forces that is not new since it soon have over thirty years and is trying to bridge the issues of the standard model of particles and also to unite General Relativity and Quantum Physics (which is not without difficulty since the latter is dependent on the background unlike General Relativity). It is one of the many theories that exist in modern physics and is trying in the early 121st century (holocene calendar) this unification (there are dozens of other more or less known).

Remark

If this subject is covered in the Cosmology chapter and not in the Atomistic one it is only for a pedagogical reason! Indeed, the basic formalism of String Theory is much closer to Relativistic Mechanics (Special and General Relativity) than that of the wave quantum physics or quantum physics fields. It seemed therefore more suited to this day (!), to provide continuity in the mathematical formalism and its interpretation rather than a thematic continuity with a relatively different approach than the usual formalism of quantum physics.

Well, String Theory is a theoretical framework in which the point-like particles of particle physics are replaced by one-dimensional objects named "strings" as its name suggest it... It describes how these strings propagate through space and interact with each other. On distance scales larger than the string scale, a string looks just like an ordinary particle, with its mass, charge, and other properties determined by the vibrational state of the string. In String Theory, one of the many vibrational states of the string corresponds to the (hypothesized) graviton, a quantum mechanical particle that maybe carries gravity. Thus String Theory is a theory of quantum gravity.

String Theory is a broad and varied subject that attempts to address a number of deep questions of fundamental physics. String Theory has been applied to a variety of problems in Black Hole physics, early universe cosmology, nuclear physics, and condensed matter physics, and it has stimulated a number of major developments in pure mathematics. Because String Theory potentially provides a unified description of gravity and particle physics, it is a candidate for a theory of everything, a self-contained mathematical model that describes all fundamental forces and forms of matter. Despite much work on these problems, it is not known to what extent String Theory describes the real world or how much freedom the theory allows to choose the details.

---

[82] Notice that the name is also misleading. "String Theory" is not a theory actually but a "Model". Such a similar vocabulary issue also exist with the "Standard Model" that is not a model anymore since decades but a "Theory"!

Remark

Notice that String theory can recover the Hawking–Bekenstein entropy by providing a microscopic count of Black Hole microstates, something semiclassical gravity alone cannot do. Certain extremal Black Holes are modeled as bound states of D-branes and strings, whose quantum excitations can be counted explicitly; the resulting number of microstates matches the Hawking entropy formula ($S = A/4G$ in natural units) exactly, giving a statistical-mechanical foundation to Hawking's thermodynamic result (see [521]).

String theory has also contributed to important results in superconductivity by providing mathematical tools that help model strongly correlated electrons in high-temperature superconductors, where traditional quasiparticle approaches fail. Using holographic duality (AdS/CFT), physicists map the complex quantum behavior of electrons in a lattice to a simpler gravitational problem in a higher-dimensional spacetime, allowing them to predict how materials transition into superconducting phases and how their resistivity behaves near critical temperatures. This approach has yielded accurate predictions for unusual superconducting materials and clarified how quantum phase transitions occur in systems where electron interactions are too entangled to treat individually.

The undeniable advantage of String Theory, besides the fact that mathematically it is quite indigestible but not really worse than General Relativity, is that it avoids in a certain order... many singularities in the calculations unlike other contemporary theories that consider the objects as points (so zero volume and length...).

The undeniable advantage of String Theory, besides the fact that mathematically it is quite indigestible but not really worse than General Relativity, is that it avoids in a certain order ... many singularities in the calculations unlike other contemporary theories that consider the objects as points (so zero volume and length...).

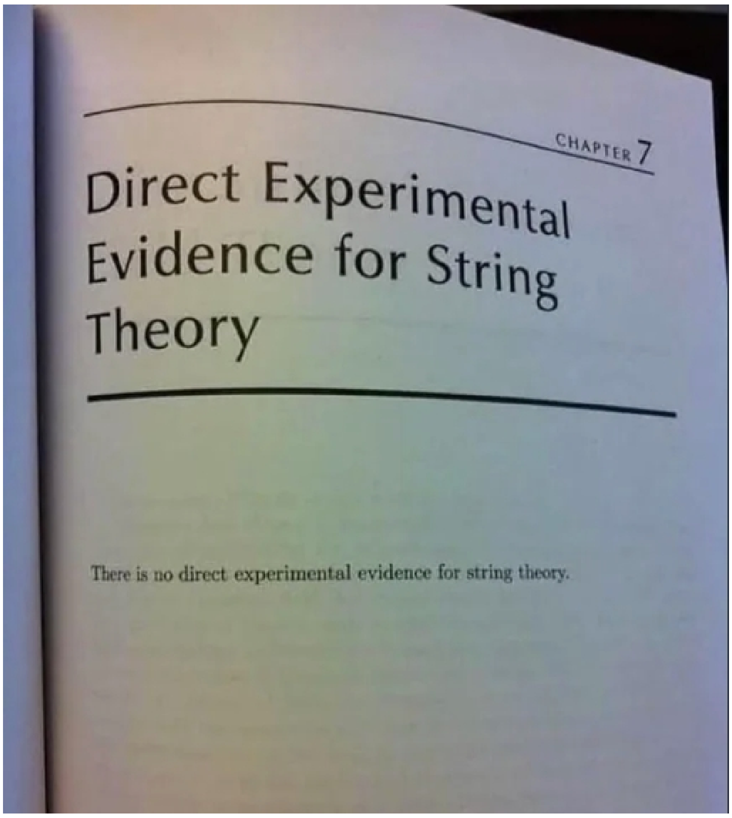
This theory, although aesthetic and remarkable in that it uses for its calculations bases foundations that have over 200 years, is lacking in our opinion to work with successive analogies, as we shall see, with current relativistic and quantum theories. While this is not dramatic in itself, the theory may seem to lose a little of its own autonomy even though the fact it is not. The reader should therefore not be badly surprised in the development that will follow...

The main particularity of String Theory is that his ambition does not stop to this reconciliation, but it claims to successfully unify the four known elementary interactions, we speak therefore about "theory of everything", while staying on two hypothesis/assumptions:

H1. The fundamental building blocks of the universe would not be point particles, but a variety of vibrating string having a given stress in the manner of an elastic. What we perceive as particles with special characteristics (mass, charge, etc.) are merely distinct strings vibrating differently. With this assumption, string theories admit a minimum scale and make it easy to avoid the emergence of some infinite amounts that are inevitable in usual quantum field theories.

H2. The Universe could contain more than three spatial dimensions. Some of them, folded on themselves, being invisible to our scales (by a procedure named "dimensional reduction").

Despite promising first partial promising results and also remarkable rich mathematical background the String Theory remains however incomplete. On the one hand, a multitude of solutions to String Theory equations exist, which poses a selection problem for our Universe and, secondly, even if many similar models were obtained, none of them allows to reproduce accurately the standard model of particle physics...



CHAPTER 7

Direct Experimental Evidence for String Theory

There is no direct experimental evidence for string theory.

That said ... let us begin our initiation:

12.7.1 Wave equation of a transversal string

The aim here will be in a first step to determine the non-relativistic wave equation of a string excited transversely using calculations that we made in the section of Wave Mechanics. Once this work done, we will pass to the study of relativistic strings and see their wave equation, as well as the non-relativistic version, can be assimilated to the current conservation equation we had proved in the section of Electrodynamics.

We begin by recalling the form of the action that we proved in the section of Wave Mechanics for a non-relativistic string (see page 3496):

$$S=\int_{t_1}^{t_2}\mathrm{d}t\int_0^L \mathrm{d}x\mathcal{L}\left(\frac{\partial y}{\partial t},\frac{\partial y}{\partial t}\right) \tag{12.7.1}$$

with:

$$L=\int\limits_0^L\left[\frac{1}{2}\rho\left(\frac{\partial y}{\partial t}\right)^2-\frac{1}{2}F\left(\frac{\partial y}{\partial x}\right)^2\right]\mathrm{d}x=\int\limits_0^L\mathcal{L}\mathrm{d}x \tag{12.7.2}$$

Now, in the same way as we did in thE section of Analytical Mechanics at page 3155 (and in that of Quantum Field Theory), we will define a notation by an analogy to the canonical moments of the string:

$$\mathcal{P}^t:=\frac{\partial\mathcal{L}}{\partial\dot{y}}\qquad\mathcal{P}^x:=\frac{\partial\mathcal{L}}{\partial y'} \tag{12.7.3}$$

with $y'=\partial y/\partial x$. It is simply the derivatives from the Lagrangian density with respectively the first and second argument. More explicitly, we get then directly by doing the calculation (see section Wave Mechanics page 3497):

$$\mathcal{P}^t:=\rho_l\frac{\partial y}{\partial t}\qquad\mathcal{P}^x:=-F\frac{\partial y}{\partial x} \tag{12.7.4}$$

So if we rewrite the variational of the action we obtained in the section of Wave Mechanics with this canonical notation, we get:

$$S=\int\limits_{t_i}^{t_f}\mathrm{d}t\int\limits_0^L\mathrm{d}x\left[\rho_l\frac{\partial y}{\partial t}\frac{\partial(\delta y)}{\partial t}-F\frac{\partial y}{\partial x}\frac{\partial(\delta y)}{\partial x}\right]=\int\limits_{t_i}^{t_f}\mathrm{d}t\int\limits_0^L\mathrm{d}x\left[\frac{\partial\mathcal{L}}{\partial\dot{y}}\delta\dot{y}+\frac{\partial\mathcal{L}}{\partial y'}\delta y'\right] \tag{12.7.5}$$

Making use of the same methods as in the section of Wave Mechanics, our variational can be express after simplification again in the form of three terms:

$$\begin{aligned}\delta S&=\int\limits_0^L\left[\rho_l\frac{\partial y}{\partial t}\delta y\right]\Bigg|_{t_i}^{t_f}\mathrm{d}x+\int\limits_{t_i}^{t_f}\left[-F\frac{\partial y}{\partial x}\delta y\right]\Bigg|_0^L\mathrm{d}t-\int\limits_{t_i}^{t_f}\mathrm{d}t\int\limits_0^L\mathrm{d}x\left(\rho_l\frac{\partial^2 y}{\partial t^2}-F\frac{\partial^2 y}{\partial x^2}\right)\delta y\\&=\int\limits_0^L\left[\mathcal{P}^t\delta y\right]\Big|_{t_i}^{t_f}\mathrm{d}x+\int\limits_{t_i}^{t_f}\left[\mathcal{P}^x\delta y\right]\Big|_0^L\mathrm{d}t-\int\limits_{t_i}^{t_f}\mathrm{d}t\int\limits_0^L\mathrm{d}x\left(\frac{\partial\mathcal{P}^t}{\partial t}+\frac{\partial\mathcal{P}^x}{\partial x}\right)\delta y\end{aligned} \tag{12.7.6}$$

The conditions to find the extreme value (according to the principle of least action) are the same as those seen in the section of Wave Mechanics. Thus, for the third term, we therefore have the wave equation of a transversely excited string with the canonical form given by:

$$\boxed{\frac{\partial\mathcal{P}^t}{\partial t}+\frac{\partial\mathcal{P}^x}{\partial x}=0} \tag{12.7.7}$$

Remark

It obviously should be noted that this form of writing will greatly facilitate our work!

It must be observed (as it is remarkable!) also that as in the section of Analytical Mechanics, the canonical moment $\mathcal{P}^t$ as defined above, coincides perfectly (the hazard makes things well) with the density momentum that we obtained in the section of Wave Mechanics (see page 3500). Indeed:

$$p_y=\int\limits_0^L(\rho_l\mathrm{d}x)\frac{\partial y}{\partial x}=\int\limits_0^L\rho_l\frac{\partial y}{\partial x}\mathrm{d}x=\int\limits_0^L\mathcal{P}^t\mathrm{d}x \tag{12.7.8}$$

Thus, by analogy with Analytical Mechanics (where we recall, the derivative of the Lagrangian with respect to the speed gives the momentum), $\dot{y}$ plays well the role of speed and thus the derivative of the Lagrangian density by this latter gives the momentum density $\mathcal{P}^t$!!!

Remember also another point that has been seen in the section of Wave Mechanics, the extremum of the action ($\delta S=0$) imposes use the Neumann boundary conditions (see page 3499), which leads us to write $\mathcal{P}^x=0$.

Remark

In the context of String Theory to relativistic with more than 3 dimensions, it is possible to generalize the concept of boundary conditions considering the constraints in space like hypersurfaces named Dp-branes with p dimensions. The usual Dirichlet boundary conditions then correspond to the situation where the ends of a strings are constraint by a 0-brane. The Neumann condition for a free string in p dimensions corresponds to a constraint on a Dp-brane.

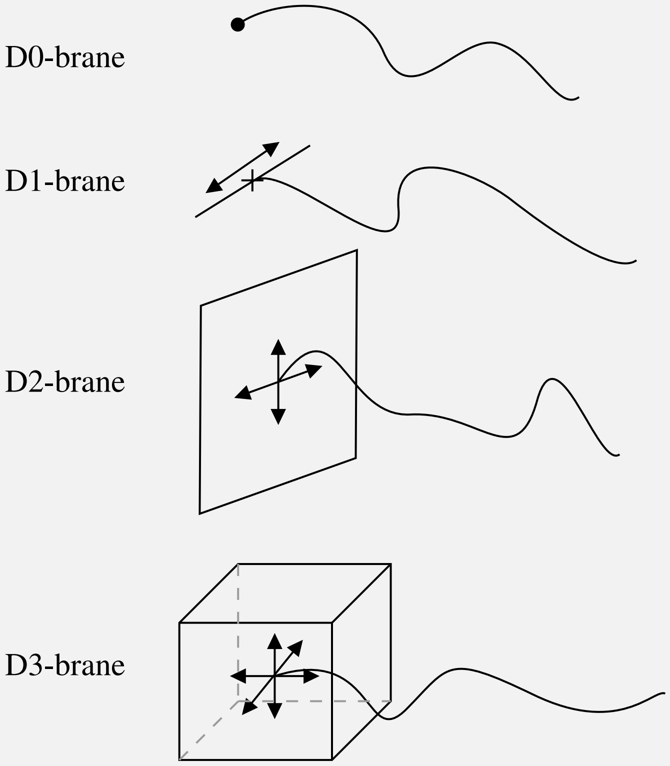


Figure 12.304: **Illustration of Dp-branes**

12.7.2 Non-relativistic Wave equation of a transversal string

We will now determine the action of a relativistic string. We can, to lay the foundations of our study, remember that a point particle draws a line in space-time (each point on the line is marked by a time coordinate and three space coordinates). Therefore, by extension, a string that is a two-dimensional element (if we consider it with no thickness) plots a surface in the space-time.

Thus, just like the line that draw a point particle in space-time is named a "world line" (see section Special Relativity page 5705), the surface traced by a string will by analogy be named a "surface Universe".

A string in a closed space-time Minkowski trace, for example, a tube, while an open string trace a band:

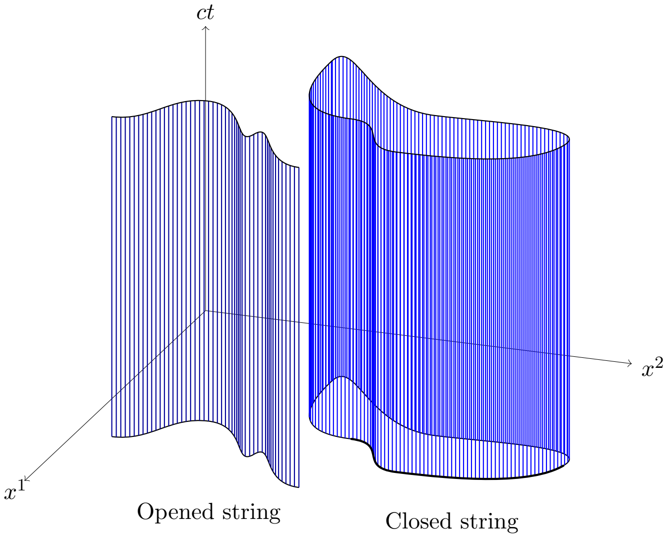


Figure 12.305: Universe surface generated by respectively an opened/closed string (author: Daniel N.)

in the figure above, with two spatial dimensions and one implicit temporal dimension, the string is motionless in our current space. It moves in space-time (as time goes on the vertical axis) but not in space in the example above (it would take an additional spatial component to see such a movement).

Remarks

R1. Remember that the diagram above is in three dimensions while the space-time has four dimensions.

R2. Remember also that the time vector of the orthogonal basis is always perpendicular to all other spatial components (this remark will be useful during our proof of the Nambu-Goto action).

During our proof of the equation of motion in the section of General Relativity (see page 5738), we reparameterized the particle's Universe line with a parameter that was the proper time of the particle t. Indeed, you only have to remember parametric equations that represent curves. For example with Maple™ 4.00b:

```
>with(plots):
>spacecurve([cos(t),sin(t),t],t=0..4*Pi,axes=boxed);
```

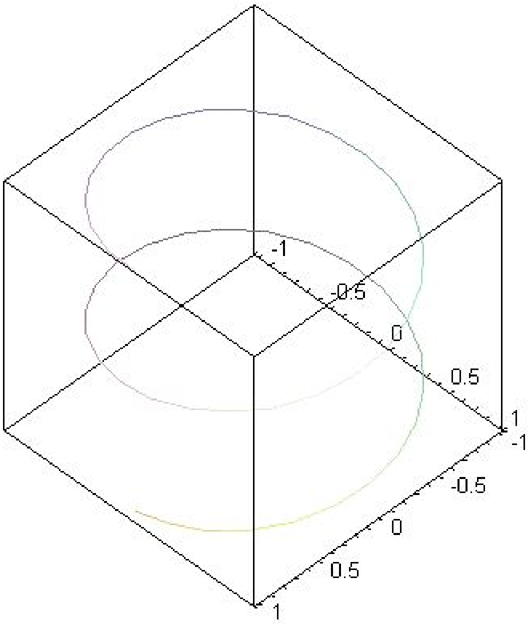


Figure 12.306: **Elementary illustrated refresh of a parametric curve**

and the same procedure is valid for a line in four dimensions (time + space).

We were thus arrived to construct the expression of the action S of it before applying the variational principle.

We will do the same for a relativistic string with the difference that we will re-parametrized the surfaces generated by the strings this time. The constraints we impose are that the chosen parameters will also have to be (in reference to the case of the particle) relativistic invariants.

As we have therefore seen in the section of General Relativity (see page 5738), a world line can be re-parametrized naturally using only one parameter (curvilinear abscissa). A surface in space, however, is a two-dimensional object, we assume by extension that it requires two parameters ζ^1,ζ^2 (one more) to be described completely.

Indeed, we guess, that one of the two parameters will be the proper time (to make the surface evolve in the time), the second parameter will give a "thickness" of what would be only a Universe line if it did not exist. It would be sufficient in a three-dimensional space that the second parameter had, to generate a surface, the dimensions of length but in the four dimension space-time the second parameter must have the units of a surface.

Given a parametrized surface, we can draw on them isolines of the parameters (lines where the two parameters ζ,ζ_2 are constant over the entire surface). These contours cover the surface as a grid (see figure a little bit further below).

The parametric equation of a volume in space requires three parameters as we saw in the section of Analytical Geometry (see page 2800). Thus, if a parametrized area in Euclidean space can be represented by a vector of the type:

$$\vec{f}=\begin{pmatrix} x \\ y \\ z \end{pmatrix} \tag{12.7.9}$$

during a re-parametrization and making use of the tensor notation of Minkowski space-time as seen in the section of General Relativity, we have (by restricting for the moment to the particular case of two spatial dimensions and one of time):

$$\vec{x}=\begin{pmatrix} x^1(\xi^1,\xi^2) \\ x^2(\xi^1,\xi^2) \\ x^3(\xi^1,\xi^2) \end{pmatrix} \tag{12.7.10}$$

Thus, the surface is the image of the parameters (ξ^1,ξ^2). Alternatively, we can see the components (ξ^1,ξ^2) as the time and space coordinates of the surface, at least locally!

We now want to calculate the area of an element of any type of space in the same way as we did for the curvilinear abscissa of any world line in the section of General Relativity. This raises the question of the form of the differential surface element??? Should we take the multiplication of the differential of the two previously selected parameters as being a square, rectangle, circle, or other?

In fact, we will put our choice on a parallelogram! This choice may seem completely arbitrary for now but as we'll see a few lines later, this choice coincides for mathematical reasons with that we name the "induced metric" of the surface itself (rather remarkable result!).

Thus, let us denote by $\mathrm{d}\vec{v}_1$ and $\mathrm{d}\vec{v}_2$ the sides of the parallelogram. They are the image by $\vec{x}$ of the couples $(\xi^1,0)$ and $(0,\xi^2)$ respectively:

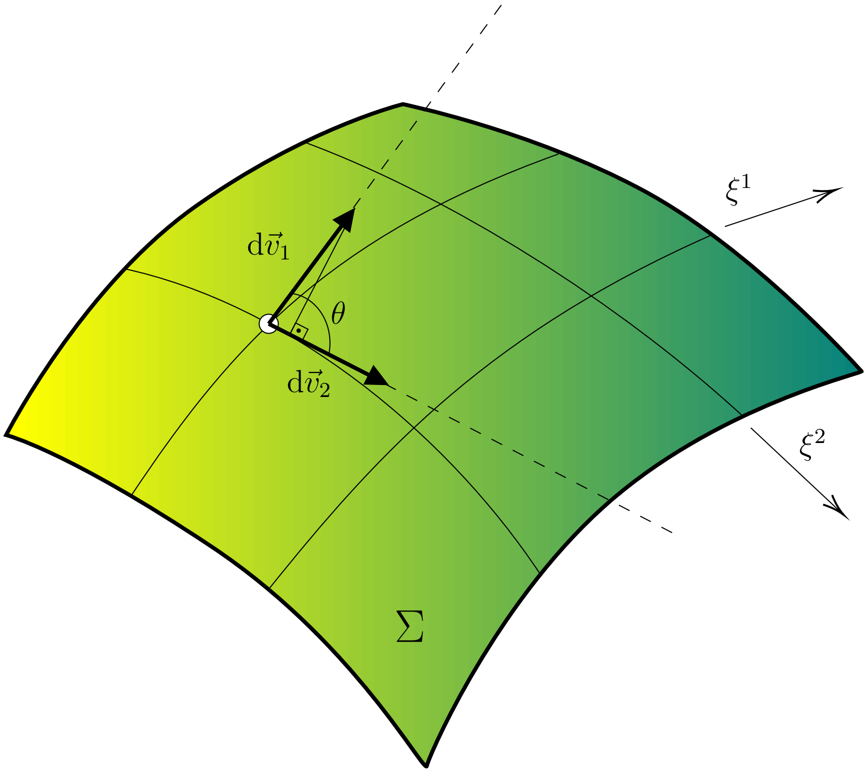


Figure 12.307: **Configuration for the study of an elementary surface element**

Therefore we can write:

$$\frac{\mathrm{d}\vec{v}_1}{\mathrm{d}\xi^1}:=\frac{\partial\vec{x}}{\partial\xi^1},\frac{\mathrm{d}\vec{v}_2}{\mathrm{d}\xi^2}:=\frac{\partial\vec{x}}{\partial\xi^2} \tag{12.7.11}$$

and then:

$$\mathrm{d}\vec{v}_1:=\frac{\partial\vec{x}}{\partial\xi^1}\mathrm{d}\xi^1,\mathrm{d}\vec{v}_2:=\frac{\partial\vec{x}}{\partial\xi^2}\mathrm{d}\xi^2 \tag{12.7.12}$$

Now let us calculate the surface $\mathrm{d}A$ (we will not take the letter S to avoid confusion with the variable representing the action in this section) of the parallelogram (see section Vector Calculus page 1835):

$$\begin{aligned}\mathrm{d}A&=||\mathrm{d}\vec{v}_1||||\mathrm{d}\vec{v}_2||\sin(\theta)=||\mathrm{d}\vec{v}_1||||\mathrm{d}\vec{v}_2||\sqrt{1-\cos^2(\theta)}\\&=\sqrt{||\mathrm{d}\vec{v}_1||^2||\mathrm{d}\vec{v}_2||^2-||\mathrm{d}\vec{v}_1||^2||\mathrm{d}\vec{v}_2||^2\cos^2(\theta)}\end{aligned} \tag{12.7.13}$$

using the dot product, this can be rewritten as:

$$\mathrm{d}A=\sqrt{(\mathrm{d}\vec{v}_1\circ\mathrm{d}\vec{v}_1)(\mathrm{d}\vec{v}_2\circ\mathrm{d}\vec{v}_2)-(\mathrm{d}\vec{v}_1\circ\mathrm{d}\vec{v}_2)^2} \tag{12.7.14}$$

using the previously established relations this can be written:

$$\mathrm{d}A=\mathrm{d}\xi^1\mathrm{d}\xi^2\sqrt{\left(\frac{\partial\vec{x}}{\partial\xi^1}\circ\frac{\partial\vec{x}}{\partial\xi^1}\right)\left(\frac{\partial\vec{x}}{\partial\xi^2}\circ\frac{\partial\vec{x}}{\partial\xi^2}\right)-\left(\frac{\partial\vec{x}}{\partial\xi^1}\circ\frac{\partial\vec{x}}{\partial\xi^2}\right)^2} \tag{12.7.15}$$

the latter relation is the general shape of a surface element of a parametrized pattern. The total surface is obviously given by:

$$A=\int \mathrm{d}\xi^1\mathrm{d}\xi^2\sqrt{\left(\frac{\partial\vec{x}}{\partial\xi^1}\circ\frac{\partial\vec{x}}{\partial\xi^1}\right)\left(\frac{\partial\vec{x}}{\partial\xi^2}\circ\frac{\partial\vec{x}}{\partial\xi^2}\right)-\left(\frac{\partial\vec{x}}{\partial\xi^1}\circ\frac{\partial\vec{x}}{\partial\xi^2}\right)^2} \tag{12.7.16}$$

Just as in the framework of the study of the principle of least action (see section Analytical Mechanics page 3131) we searched the optimum path for a particle browsing a universe line, for a string, we have to search for the optimum of the surface A by minimizing the function $\vec{x}=(\xi^1,\xi^2)$.

This latter form is, however, a little heavy and does not show anything known particularly or is not similar to any form already known in another field of physics. We will see, however that by digging a little bit however it is possible to get something pretty interesting.

Consider now a vector $\mathrm{d}\vec{x}$ and its squared length (norm) given by the scalar product:

$$\mathrm{d}s^2=(\mathrm{d}s)^2=\mathrm{d}\vec{x}\circ\mathrm{d}\vec{x} \tag{12.7.17}$$

Remark

This approach of separating the wave function into the composition of a wave function of the center of mass and the relative movement is also used in the context of the study of poly-electronic atoms, but with one difference: as the nucleus is much more massive than the processing electrons (in approximation ...), the center of mass is assimilated to the nucleus of the atom and the relative motion to the entire electron cloud. This approximate approach is well known under the designation "Born-Oppenheimer approximation".

The vector $\mathrm{d}\vec{x}$ can be expressed in the form of partial derivative terms of $\mathrm{d}\xi^1,\mathrm{d}\xi^2$, such that we get its exact total differential (see section Differential and Integral Calculus page 1466):

$$\mathrm{d}\vec{x}=\frac{\partial\vec{x}}{\partial\xi^1}\mathrm{d}\xi^1+\frac{\partial\vec{x}}{\partial\xi^2}\mathrm{d}\xi^2=\frac{\partial\vec{x}}{\partial\xi^i}\mathrm{d}\xi^i \tag{12.7.18}$$

Thus, the squared length of $\mathrm{d}\vec{x}$ can be expressed in tensor form:

$$\mathrm{d}s^2=\left(\frac{\partial\vec{x}}{\partial\xi^i}\mathrm{d}\xi^i\right)\circ\left(\frac{\partial\vec{x}}{\partial\xi^i}\mathrm{d}\xi^i\right)=\frac{\partial\vec{x}}{\partial\xi^i}\circ\frac{\partial\vec{x}}{\partial\xi^i}\mathrm{d}\xi^i\mathrm{d}\xi^i \tag{12.7.19}$$

what we will note by convention in the future:

$$\mathrm{d}s^2=g_{ij}(\xi)\mathrm{d}\xi^i\mathrm{d}\xi^j \tag{12.7.20}$$

The quantity $g_{ij}(\xi)$ is named the "induced metric of the parametrized area" (because contains a scalar product which quite generally uses a metric... hence the term "induced") and is therefore a matrix of dimension 2×2. It is obvious that the choice of this name comes from the resemblance with the usual metric as we have defined in our study of tensor calculus and from its use in special and General Relativity.

The matrix $g_{ij}(\xi)$ has therefore by design and definition the form:

$$g_{ij}(\xi)=\begin{bmatrix}\dfrac{\partial\vec{x}}{\partial\xi^1}\circ\dfrac{\partial\vec{x}}{\partial\xi^1} & \dfrac{\partial\vec{x}}{\partial\xi^2}\circ\dfrac{\partial\vec{x}}{\partial\xi^2}\\ \dfrac{\partial\vec{x}}{\partial\xi^2}\circ\dfrac{\partial\vec{x}}{\partial\xi^1} & \dfrac{\partial\vec{x}}{\partial\xi^2}\circ\dfrac{\partial\vec{x}}{\partial\xi^2}\end{bmatrix} \tag{12.7.21}$$

Now let us come back to our expression of the surface generated by the string:

$$A=\int \mathrm{d}\xi^1\mathrm{d}\xi^2\sqrt{\left(\dfrac{\partial\vec{x}}{\partial\xi^1}\circ\dfrac{\partial\vec{x}}{\partial\xi^1}\right)\left(\dfrac{\partial\vec{x}}{\partial\xi^2}\circ\dfrac{\partial\vec{x}}{\partial\xi^2}\right)-\left(\dfrac{\partial\vec{x}}{\partial\xi^1}\circ\dfrac{\partial\vec{x}}{\partial\xi^2}\right)^2} \qquad (12.7.22)$$

and let us quickly calculate the determinant (see section Linear Algebra page 1961) of the matrix $g_{ij}(\xi)$:

$$\det(g_{ij}(\xi))=g=g_{11}g_{22}-g_{21}g_{12}=\left(\dfrac{\partial\vec{x}}{\partial\xi^1}\circ\dfrac{\partial\vec{x}}{\partial\xi^1}\right)\left(\dfrac{\partial\vec{x}}{\partial\xi^2}\circ\dfrac{\partial\vec{x}}{\partial\xi^2}\right)-\left(\dfrac{\partial\vec{x}}{\partial\xi^1}\circ\dfrac{\partial\vec{x}}{\partial\xi^1}\right)^2 \qquad (12.7.23)$$

and so what? Well here it is! A can now been expressed as:

$$A=\int \mathrm{d}\xi^1\mathrm{d}\xi^2\sqrt{g} \qquad (12.7.24)$$

Thus, the choice of the parallelogram as elementary surface is best explained here!

Now we will adopt the traditional notations of String Theory in relation to the expression of the surface. Thus, just as the time-space coordinates are described in Special Relativity the space-time four-vector:

$$x^\mu=(x^0,x^1,\ldots,x^d) \qquad (12.7.25)$$

we will describe the surfaces Universe by (we now turn to the notation making use of the four dimensions of space-time):

$$x^\mu(\tau,\sigma)=X^\mu(\tau,\sigma) \qquad (12.7.26)$$

This notation will save us in the future to have to confuse, if the theory leads us there, the traditional space-time coordinates x^μ with the image function of the surface Universe $x^\mu(\tau,\sigma)$ and this especial because physicists being sometimes a little lazy shorten this latter x^μ... hence the choice of the capital letter.

It is then much more appropriate and wise to change the notation.

From now on we will name "string coordinates" the surface Universe described by X^μ.

This small change in notation obviously not change the interpretation of the image of the function. Given a couple (τ,σ) involving proper-time element and surface element of the pre-image, this point is projected onto a surface element of the space-time coordinates of the string:

$$\left(X^0(\tau,\sigma),X^1(\tau,\sigma),\ldots,X^d(\tau,\sigma)\right) \qquad (12.7.27)$$

12.7.2.1 Nambu-Goto Action

The Nambu–Goto action is the simplest invariant action in bosonic String Theory, and is also used in other theories that investigate string-like objects (for example, cosmic strings). It is the starting point of the analysis of zero-thickness (infinitely thin) string behaviour, using the principles of Lagrangian mechanics. Just as the action for a free point particle is proportional to its proper time—i.e., the "length" of its world-line—a relativistic string's action is proportional to the area of the sheet which the string traces as it travels through space-time.

In the case of a Universe surface the parameters are then by convention τ and σ, where as in special and General Relativity, the proper time may be in the range:

$$\tau \in]-\infty,+\infty[\tag{12.7.28}$$

the second can only be positive since this is a surface:

$$\sigma \in]0,\sigma_1[\tag{12.7.29}$$

and the coordinates of this surface corresponding to the parameter space is therefore:

$$X^\mu(\tau,\sigma) \tag{12.7.30}$$

where once again, for recall, the parameter τ is considered as the variable describing the time (there must be one!), and σ the variable describing the spatial extension of a string (that is to say that the condition $\sigma \in]0,\sigma_1[$ involving the finite length of the string).

The parameters (τ,σ) describe therefore a surface in the space of pre-images:

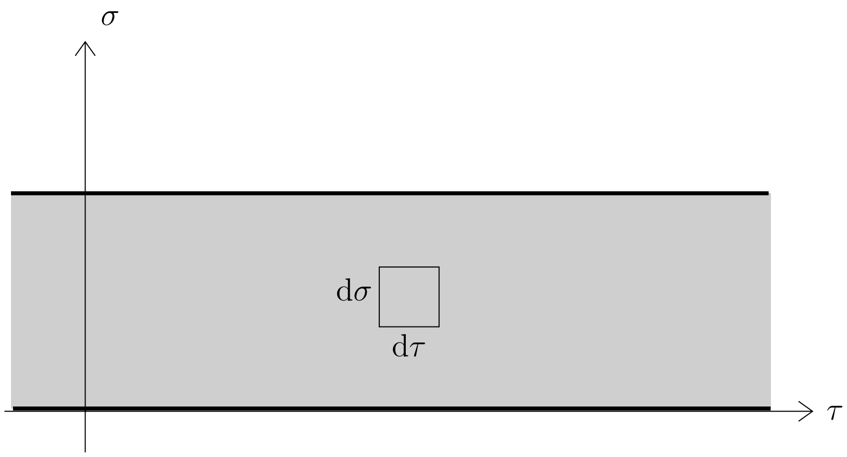


Figure 12.308: **Parametrization of a space-time surface**

The ends of the string have a constant value σ. However, as time passes and that the ends of the string on the Universe surface move we must notice an essential condition of the Universe surface concerning the both ends of an open string:

$$\frac{\partial X^0}{\partial \tau} \neq 0 \tag{12.7.31}$$

Remark

This condition is made on the component X^0 because it corresponds to the component x^0 of the space-time four-vector which is nothing else, in natural units, t (the proper time). Therefore, time passes and is never constant, this is why we impose this derivative as being non-zero.

And using the standard conventions in physics for writing the derivatives with respect to time or space components, we agree to adopt also now the following notations:

$$\dot{X}=\frac{\partial X^\mu}{\partial\tau}\qquad X'=\frac{\partial X^\mu}{\partial\sigma}\tag{12.7.32}$$

since as:

$$X^\mu(\tau,\sigma)=X^\mu(t,\sigma)=(ct,\vec{X}(t,\sigma))\tag{12.7.33}$$

then:

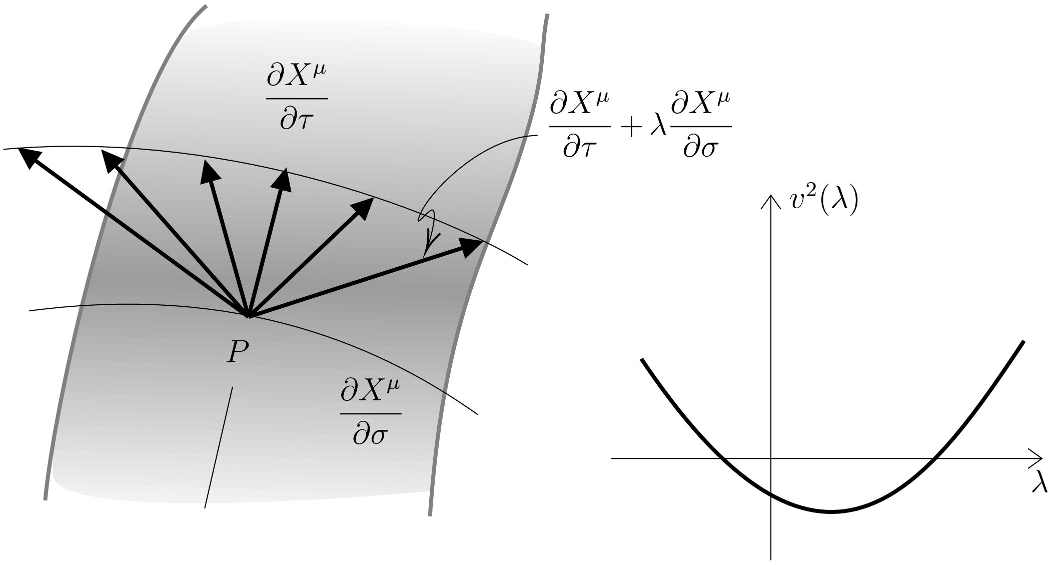
$$\begin{aligned}\frac{\partial X^\mu}{\partial\sigma}&=\left(\frac{\partial X^0}{\partial\sigma},\frac{\partial\vec{X}}{\partial\sigma}\right)=\left(0,\frac{\partial\vec{X}}{\partial\sigma}\right)\\\frac{\partial X^\mu}{\partial\tau}&=\left(\frac{\partial X^0}{\partial t},\frac{\partial\vec{X}}{\partial\tau}\right)=\left(c,\frac{\partial\vec{X}}{\partial t}\right)\end{aligned}\tag{12.7.34}$$

The surface is therefore written:

$$\begin{aligned}A&=\int\mathrm{d}\xi^1\mathrm{d}\xi^2\sqrt{\left(\frac{\partial\vec{x}}{\partial\xi^1}\circ\frac{\partial\vec{x}}{\partial\xi^1}\right)\left(\frac{\partial\vec{x}}{\partial\xi^2}\circ\frac{\partial\vec{x}}{\partial\xi^2}\right)-\left(\frac{\partial\vec{x}}{\partial\xi^1}\circ\frac{\partial\vec{x}}{\partial\xi^2}\right)^2}\\&=\int\mathrm{d}\tau\mathrm{d}\sigma\sqrt{\left(\frac{\partial X^\mu}{\partial\tau}\frac{\partial X_\mu}{\partial\tau}\right)\left(\frac{\partial X^\nu}{\partial\sigma}\frac{\partial X_\nu}{\partial\sigma}\right)-\left(\frac{\partial X^\mu}{\partial\tau}\frac{\partial X_\mu}{\partial\sigma}\right)^2}\\&=\int\mathrm{d}\tau\mathrm{d}\sigma\sqrt{(\dot{X})^2(X')^2-(\dot{X}\circ X')^2}\end{aligned}\tag{12.7.35}$$

However, there is a problem here! Indeed, let us look if the radicand (term under the root) has a tangible physical reality ...

For this, we must first consider the left side of the figure below representing the surface patch described by an open string:


$\frac{\partial X^\mu}{\partial\tau}$
$\frac{\partial X^\mu}{\partial\tau}+\lambda\frac{\partial X^\mu}{\partial\sigma}$
P
$\frac{\partial X^\mu}{\partial\sigma}$
$v^2(\lambda)$
λ

At each point P of this surface patch (assumed differentiable at every point) there are endless tangents, all in the same plane, which we will denote for example by $\vec{v}$ and thus form a surface tangent at P.

Now, as the space in which the surface patch of the string held in a spatial and temporal orthonormal basis, the tangent vectors $\vec{v}$ can then in turn be decomposed into a two-dimensional spatial and temporal local orthogonal base at the point P such that the vectors of the base are two vectors (see our study of surface patches in the section of Differential Geometry at page 2878):

$$\frac{\partial X^\mu}{\partial \tau} \qquad \frac{\partial X^\mu}{\partial \sigma} \tag{12.7.36}$$

all other tangent vectors expressing as a linear combination thereof.

However, a problem remains in our decomposition (...): the unit vectors of the local orthogonal basis at P have units that differ... For this let us add a dimensional factor to the spatial component (this is arbitrary because the conclusion will be the same regardless of the component on which you put the sizing factor) as we did in Special Relativity with the time axis:

$$\frac{\partial X^\mu}{\partial \tau} \qquad \lambda\frac{\partial X^\mu}{\partial \sigma} \tag{12.7.37}$$

This dimensional factor can also be used to get all the tangent vectors such as:

$$v^\mu=\frac{\partial X^\mu}{\partial \tau}+\lambda\frac{\partial X^\mu}{\partial \sigma} \tag{12.7.38}$$

Indeed, if $\lambda\in[-\infty,+\infty]$, then for $\lambda=0$ we get the vector $\partial X^\mu/\partial\tau$ and for $\lambda=+\infty$ the vector $\partial X^\mu/\partial\sigma$. And all intermediate values, we get all the tangent vectors as shown on the left side of the previous figure.

Now, let us recall that we saw in the section of Special Relativity (see page 5707) that there exist according to the curvilinear abscissa:

$$\mathrm{d}s^2=\mathrm{d}x^\nu(\nu_{\mu\nu}\mathrm{d}x^\nu)=\mathrm{d}x^\nu\mathrm{d}x_\nu=c^2\mathrm{d}t^2-\mathrm{d}x^2-\mathrm{d}y^2-\mathrm{d}z^2 \tag{12.7.39}$$

of the Universe line of the light type ($\mathrm{d}s^2$), space ($\mathrm{d}s^2<0$) or time ($\mathrm{d}s^2>0$) if we consider the four-vectors x^μ.

It must be true by analogy the same for the tangent vectors to the surface and given by:

$$v^\mu=\frac{\partial X^\mu}{\partial \tau}+\lambda\frac{\partial X^\mu}{\partial \sigma} \tag{12.7.40}$$

Therefore:

$$v^\mu v_\mu=\lambda^2\left(\frac{\partial X^\mu}{\partial \sigma}\right)^2+2\lambda\left(\frac{\partial X^\mu}{\partial \tau}\circ\frac{\partial X^\mu}{\partial \sigma}\right)+\left(\frac{\partial X^\mu}{\partial \tau}\right)^2 \tag{12.7.41}$$

that corresponds to an equation of the second degree on λ, and that must, to have negative values (Universe surface patch of the type of space) or positive (Universe surface type of the type time) have at least two roots (see the right part of the previous figure). This brings us back to the condition that the discriminant is strictly positive (see section Calculus page 1383):

$$\Delta = b^2 - 4ac = v^\mu v_\mu = 4\left(\frac{\partial X^\mu}{\partial \tau}\circ\frac{\partial X^\mu}{\partial \sigma}\right)^2 - 4\left(\frac{\partial X^\mu}{\partial \sigma}\right)^2\left(\frac{\partial X^\mu}{\partial \tau}\right)^2 > 0 \qquad (12.7.42)$$

Therefore:

$$\left(\frac{\partial X^\mu}{\partial \tau}\circ\frac{\partial X^\mu}{\partial \sigma}\right)^2 - \left(\frac{\partial X^\mu}{\partial \sigma}\right)^2\left(\frac{\partial X^\mu}{\partial \tau}\right)^2 > 0 \qquad (12.7.43)$$

Into condensed form this is equivalent as writing:

$$(\dot{X}\circ X')^2 - (\dot{X})^2(X')^2 > 0 \qquad (12.7.44)$$

The surface must therefore be written as:

$$A = \int \mathrm{d}\tau \mathrm{d}\sigma \sqrt{(\dot{X}\circ X')^2 - (\dot{X})^2(X')^2} \qquad (12.7.45)$$

if we want the radicand has a physical sense.

By analogy with the Lagrangian of General Relativity (see page 5869) we write the latter:

$$\boxed{A = \int \mathrm{d}\tau \mathrm{d}\sigma \sqrt{(\dot{X}\circ X')^2 - (\dot{X})^2(X')^2} = \int \mathrm{d}\Sigma\sqrt{-g}} \qquad (12.7.46)$$

that we will justify a little more robustly further below.

Recall now that the action S of a point particle is proportional to its world line (proper time). Thus by analogy, the action S of string will be proportional to the Universe area:

$$S \propto \int \mathrm{d}A \Leftrightarrow S \propto A \qquad (12.7.47)$$

which gives:

$$S \propto \int \mathrm{d}\tau \mathrm{d}\sigma \sqrt{(\dot{X}\circ X')^2 - (\dot{X})^2(X')^2} \qquad (12.7.48)$$

Which brings us very frequently in the literature to find the action of a string in the following form:

$$S \propto \int \mathrm{d}\tau \mathrm{d}\sigma \mathcal{L}(\sigma,\tau) \qquad (12.7.49)$$

or more stylized:

$$\boxed{\mathcal{S} \propto \int \mathrm{d}\mathcal{A} = \int \mathrm{d}\tau \mathrm{d}\sigma \mathcal{L}(\sigma,\tau) = \int \mathrm{d}\Sigma \mathcal{L}(\sigma,\tau)} \qquad (12.7.50)$$

Relation to compare with the Lagrangian of a free particle (see section Analytical Mechanics page 3170) and with the Lagrangian density of a field (see section Quantum Field Theory page 5172):

$$S(q,\dot{q}) = \int_{t_A}^{t_B} L(q,\dot{q})\mathrm{d}t \quad \text{and} \quad S = \int \mathcal{L}(\Phi^A, \dot{\Phi}^A, \partial_\mu \Phi^A)\mathrm{d}x^4 \qquad (12.7.51)$$

The functional S has for units the one of a surface. This because the X^μ have a unit of length and inside the root each is at the fourth power and the units of τ, $s\sigma$ cancel between the inner root and the differential that are outside it.

Now, by definition of the action, the units that we have to get should match that of energy multiplied by time. That is to say joules [J] or using the international system of units $[\text{kg}\cdot\text{m}^2\cdot s^{-1}]$. For now, we have:

$$\tau:[\text{s}] \qquad \sigma:[\text{m}] \qquad X^\mu=[\text{m}] \qquad A:[\text{m}^2] \tag{12.7.52}$$

To get to the action units we want, then we have to multiply the expression of the surface A by a quantity whose units are $[\text{kg}\cdot\text{s}^{-1}]$. To chose these quantities, we will inspire us from our study of Wave Mechanics. When we worked with (non-relativistic) strings we saw that the properties to be considered were the stress and the velocity of string wave propagation. We'll try to take the following stress/speed ratio:

$$\frac{\text{stress}}{\text{velocity}}=\frac{[\text{N}]}{[\text{m}\cdot\text{s}^{-1}]}=\frac{[\text{kg}\cdot\text{m}\cdot\text{s}^{-1}]}{[\text{m}\cdot\text{s}^{-1}]}=[\text{kg}\cdot\text{s}^{-1}]=\frac{T}{v}=\frac{T_0}{c} \tag{12.7.53}$$

where appears therefore that the stress of the string at rest T_0 and the speed of light c.

Remark

This is similar to the point material physics where in the action we find the rest mass (equivalent to the tension at rest of the string) and the speed of light (see section Special Relativity page 5629).

Thus, the "Nambu-Goto action" can now be written:

$$\boxed{S=-\frac{T_0}{S}\int \mathrm{d}\tau\mathrm{d}\sigma\sqrt{(\dot{X}\circ X')^2-(\dot{X})^2(X')^2}} \tag{12.7.54}$$

Remark

We will prove later why we put a factor "$-$". However, a small analogy with the action of a point particle, for which we also have a "$-$" sign (see section Special Relativity page 5681), can easily be done...

Let us define for what will follow:

$$g_{\alpha\beta}=\frac{\partial X}{\partial\tau}\circ\frac{\partial X}{\partial\sigma}=\frac{\partial X^\mu}{\partial\tau}\frac{\partial X_\mu}{\partial\tau} \tag{12.7.55}$$

What we can also write in matrix form:

$$g_{\alpha\beta}=\begin{bmatrix}(\dot{X})^2 & \dot{X}\circ X'\\ \dot{X}\circ X' & (X')^2\end{bmatrix} \tag{12.7.56}$$

using the determinant of the matrix, it comes:

$$g=\det(g_{\alpha\beta})=(\dot{X})^2(X')^2-(\dot{X}\circ X')^2 \tag{12.7.57}$$

So we can then write the action of a relativistic string in the final following condensed form:

$$\boxed{S=-\frac{T_0}{c}\int \mathrm{d}\tau\mathrm{d}\sigma\sqrt{-g}} \tag{12.7.58}$$

which is nothing else than the "Nambu-Goto condensed form" form of a relativistic string. The latter can also be found in the follow shape in many textbooks:

$$S_{\text{NG}}=-\frac{T_0}{c}\int\sqrt{|\det\left(\partial_\alpha X\partial_\beta X_\mu\right)|}d\tau d\sigma \tag{12.7.59}$$

With the reader has to remember that $\partial_\alpha X^\mu\partial_\beta X_\mu$ is a 2×2 matrix with $\alpha,\beta=0,1$ and $\partial_\alpha=\frac{\partial}{\partial\sigma^\alpha}$ where $\begin{cases}\sigma^0=\tau\\ \sigma^1=\sigma\end{cases}$ are the parameters of the surface.

We will now obtain the equation of motion by varying the action. For this, we will exactly inspire us of the methods seen when determining the non-relativistic wave equation of string in the section of Wave Mechanics (see page 3496).

Thus, we rewrite the Nambu-Goto action by defining a Lagrangian density $\mathcal{L}$ such that:

$$S=-\frac{T_0}{c}\int\mathrm{d}\tau\mathrm{d}\sigma\sqrt{-g}=\int\limits_{\tau_i}^{\tau_f}\mathrm{d}\tau\int\limits_{\sigma_i}^{\sigma_f}\mathrm{d}\sigma\mathcal{L}\left(\frac{\partial X^\mu}{\partial t},\frac{\partial X^\mu}{\partial\sigma}\right)=\int\limits_{\tau_i}^{\tau_f}\mathrm{d}\tau\int\limits_{\sigma_i}^{\sigma_f}\mathrm{d}\sigma\mathcal{L}(\dot{X}^\mu,X'^\mu) \tag{12.7.60}$$

where $\mathcal{L}$ is therefore defined by:

$$\boxed{\mathcal{L}(\dot{X}^\mu,X'^\mu)=-\frac{T_0}{c}\sqrt{(\dot{X}\circ X')^2-(\dot{X})^2(X')^2}} \tag{12.7.61}$$

We will now apply the variational principle on the action in order to get the equation of movement of a string. The development and approximation are perfectly similar to those seen at the beginning of this section for the non-relativistic string. Let us recall that we obtained as Lagrangian density and as an expression of the action:

$$\mathcal{L}=\frac{1}{2}\rho_l\left(\frac{\partial y}{\partial t}\right)^2-\frac{1}{2}F\left(\frac{\partial y}{\partial x}\right)^2 \qquad S=\int\limits_{t_i}^{t_f}\mathrm{d}t\int\limits_0^L\mathrm{d}x\mathcal{L}\left(\frac{\partial y}{\partial t},\frac{\partial y}{\partial x}\right)=\int\limits_{t_i}^{t_f}\mathrm{d}t\int\limits_0^L\mathrm{d}x\mathcal{L}(\dot{y},y') \tag{12.7.62}$$

and that the application of variational gave us:

$$\delta S=\int\limits_{t_i}^{t_f}\mathrm{d}t\int\limits_0^L\mathrm{d}x\left[\rho_l\frac{\partial l}{\partial t}\frac{\partial(\delta y)}{\partial t}-F\frac{\partial y}{\partial x}\frac{\partial(\delta y)}{\partial x}\right] \tag{12.7.63}$$

But what we did not see in the section Wave Mechanics is that the latter relation could easily be written also from the Lagrangian density:

$$\delta S=\int\limits_{t_i}^{t_f}\mathrm{d}t\int\limits_0^L\mathrm{d}x\left[\frac{\partial\mathcal{L}}{\partial\left(\dfrac{\partial y}{\partial t}\right)}\frac{\partial(\delta y)}{\partial t}+\frac{\partial\mathcal{L}}{\partial\left(\dfrac{\partial y}{\partial x}\right)}\frac{\partial(\delta y)}{\partial x}\right]=\int\limits_{t_i}^{t_f}\mathrm{d}t\int\limits_0^L\mathrm{d}x\left[\frac{\partial\mathcal{L}}{\partial\dot{y}}\frac{\partial(\delta y)}{\partial t}+\frac{\partial\mathcal{L}}{\partial y'}\frac{\partial(\delta y)}{\partial x}\right] \tag{12.7.64}$$

Therefore, for the relativistic string, we have an identical form by applying developments in all points similar (even if the Lagrangian density has a different form):

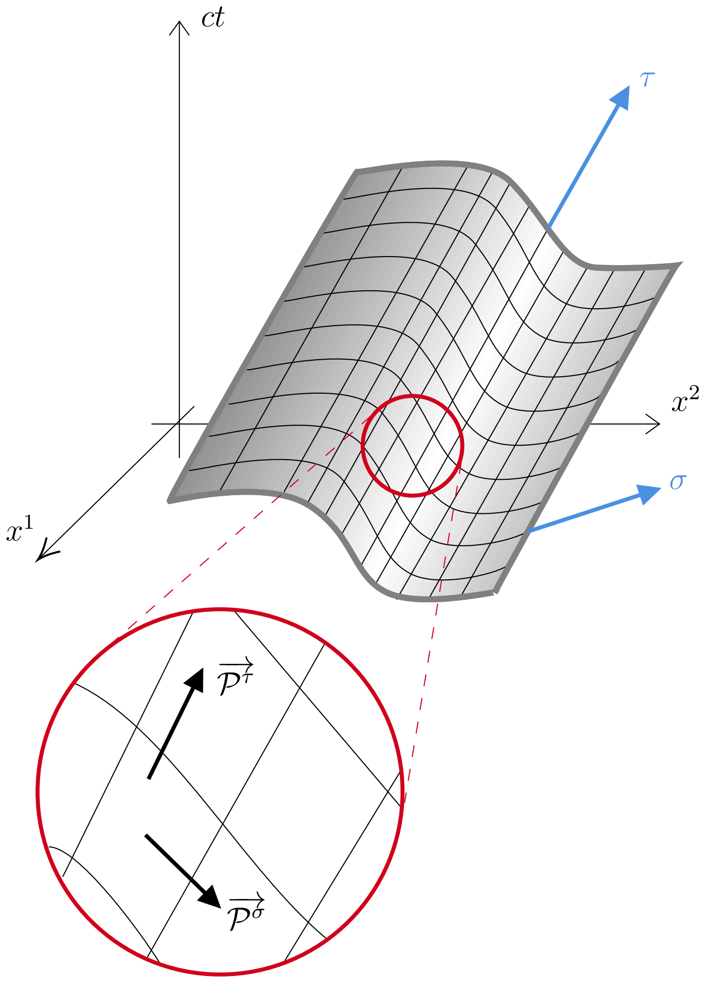
$$\delta S=\int\limits_{\tau_i}^{\tau_f}\mathrm{d}\tau\int\limits_{0}^{\sigma_1}\mathrm{d}\sigma\left[\frac{\partial\mathcal{L}}{\partial\dot{X}^{\mu}}\frac{\partial(\delta X^{\mu})}{\partial\tau}+\frac{\partial\mathcal{L}}{\partial X'^{\mu}}\frac{\partial(\delta X^{\mu})}{\partial\sigma}\right] \tag{12.7.65}$$

and as we did at the beginning of this section for non-relativistic strings, we will introduce the canonical momentum (pulse density/momentum if you prefer) of the string by choosing for the notation:

$$\boxed{\mathcal{P}_{\mu}^{\tau}:=\frac{\partial\mathcal{L}}{\partial\dot{X}^{\mu}}\qquad\mathcal{P}_{\mu}^{\sigma}:=\frac{\partial\mathcal{L}}{\partial\dot{X}'^{\mu}}} \tag{12.7.66}$$

where in the details, we get very easily (it's a simple derivative but if you wish by contacting us, we can detail the developments as always in this textbook) the longitudinal and transverse momentum:

$$\begin{aligned}\mathcal{P}_{\mu}^{\tau}&:=\frac{\partial L}{\partial\dot{X}^{\mu}}=-\frac{T_0}{c}\frac{(\dot{X}\circ X')X'_{\mu}-(X')^2\dot{X}_{\mu}}{\sqrt{(\dot{X}\circ X')^2-(\dot{X})^2(X')^2}}\\ \mathcal{P}_{\mu}^{\sigma}&:=\frac{\partial L}{\partial X'^{\mu}}=-\frac{T_0}{c}\frac{(\dot{X}\circ X')\dot{X}_{\mu}-(\dot{X})^2\dot{X}_{\mu}}{\sqrt{(\dot{X}\circ X')^2-(\dot{X})^2(X')^2}}\end{aligned} \tag{12.7.67}$$


ct
τ
x^2
σ
x^1
$\overrightarrow{\mathcal{P}^{\tau}}$
$\overrightarrow{\mathcal{P}^{\sigma}}$

by making use of this notation, we can then write:

$$\delta S=\int\limits_{\tau_i}^{\tau_f}\mathrm{d}\tau\int\limits_{0}^{\sigma_1}\mathrm{d}\sigma\left[\mathcal{P}_\mu^\tau\frac{\partial(\delta X^\mu)}{\partial\tau}+\mathcal{P}_\mu^\sigma\frac{\partial(\delta X^\mu)}{\partial\sigma}\right] \tag{12.7.68}$$

Making use once again of exactly the same methods as those seen in the section Wave Mechanics, our variational can be written, after simplification, again in the form of three terms:

$$\delta S=\int\limits_{\tau_i}^{\tau_f}\mathrm{d}\tau\int\limits_{0}^{\sigma_1}\mathrm{d}\sigma\left[\frac{\partial}{\partial\tau}(\delta X^\mu P_\mu^\tau)+\frac{\partial}{\partial\sigma}(\delta X^\mu P_\mu^\sigma)-\delta X^\mu\left(\frac{\partial(P_\mu^\tau)}{\partial\tau}+\frac{\partial(P_\mu^\sigma)}{\partial\sigma}\right)\right] \tag{12.7.69}$$

The conditions to find the extremum (according to the principle of least action) remain the same as in Wave Mechanics. Thus, for the third term, we have well the wave equation of a transversely excited string with the following canonical form:

$$\boxed{\frac{\partial(P_\mu^\tau)}{\partial\tau}+\frac{\partial(P_\mu^\sigma)}{\partial\sigma}=0} \tag{12.7.70}$$

As far as we know this equation is horribly difficult to solve but choosing an appropriate parametrization can nevertheless simplify the task.

12.7.3 Lagrangian of a String

Let us recall that we have:

$$\begin{aligned}\frac{\partial X^\mu}{\partial\sigma}&=\left(\frac{\partial X^0}{\partial\sigma},\frac{\partial\vec{X}}{\partial\sigma}\right)=\left(0,\frac{\partial\vec{X}}{\partial\sigma}\right)\\ \frac{\partial X^\mu}{\partial\tau}&=\left(\frac{\partial X^0}{\partial t},\frac{\partial\vec{X}}{\partial\tau}\right)=\left(c,\frac{\partial\vec{X}}{\partial t}\right)\end{aligned} \tag{12.7.71}$$

and that with this choice, we have:

$$
\begin{aligned}
S&=-\frac{T_0}{c}\int \mathrm{d}t\mathrm{d}\sigma\sqrt{(\dot{X}\circ X')^2-(\dot{X})^2(X')^2}\\
&=-\frac{T_0}{c}\int \mathrm{d}t\mathrm{d}\sigma\sqrt{\left(\begin{pmatrix}c\\ \dfrac{\partial\vec{X}}{\partial t}\end{pmatrix}\circ\begin{pmatrix}0\\ -\dfrac{\partial\vec{X}}{\partial\sigma}\end{pmatrix}\right)^2-\begin{pmatrix}c\\ \dfrac{\partial\vec{X}}{\partial t}\end{pmatrix}\circ\begin{pmatrix}c\\ -\dfrac{\partial\vec{X}}{\partial t}\end{pmatrix}\begin{pmatrix}0\\ \dfrac{\partial\vec{X}}{\partial\sigma}\end{pmatrix}\circ\begin{pmatrix}0\\ -\dfrac{\partial\vec{X}}{\partial\sigma}\end{pmatrix}}\\
&=-\frac{T_0}{c}\int \mathrm{d}t\mathrm{d}\sigma\sqrt{\left(\frac{\partial\vec{X}}{\partial t}\circ\left(-\frac{\partial\vec{X}}{\partial\sigma}\right)\right)^2-\left(c^2-\frac{\partial\vec{X}}{\partial t}\circ\frac{\partial\vec{X}}{\partial\sigma}\right)\left(-\frac{\partial\vec{X}}{\partial\sigma}\circ\frac{\partial\vec{X}}{\partial\sigma}\right)}\\
&=-\frac{T_0}{c}\int \mathrm{d}t\mathrm{d}\sigma\sqrt{(-1)^2\left(\frac{\partial\vec{X}}{\partial t}\circ\frac{\partial\vec{X}}{\partial\sigma}\right)^2+\left(c^2-\left(\frac{\partial\vec{X}}{\partial t}\right)^2\right)\left(\frac{\partial\vec{X}}{\partial\sigma}\right)^2}\\
&=-\frac{T_0}{c}\int \mathrm{d}t\mathrm{d}\sigma\sqrt{\left(\frac{\partial\vec{X}}{\partial t}\circ\frac{\partial\vec{X}}{\partial\sigma}\right)^2+\left(c^2-\left(\frac{\partial\vec{X}}{\partial t}\right)^2\right)\left(\frac{\partial\vec{X}}{\partial\sigma}\right)^2}\\
&=-\frac{T_0}{c}\int \mathrm{d}t\mathrm{d}\sigma\sqrt{\left(\frac{\partial\vec{X}}{\partial t}\circ\frac{\partial\vec{X}}{\partial\sigma}\right)^2+c^2\left(\frac{\partial\vec{X}}{\partial\sigma}\right)^2-\left(\frac{\partial\vec{X}}{\partial t}\right)^2\left(\frac{\partial\vec{X}}{\partial\sigma}\right)^2}
\end{aligned}
\tag{12.7.72}
$$

Now let us use what we have seen in section of Differential Geometry with the Frenet's triad (see page 2863):

$$\vec{T}=\frac{\mathrm{d}\vec{X}}{\mathrm{d}s}\tag{12.7.73}$$

where $\vec{T}$ is the tangent to the Universe surface at a time t at the of a given point. We had also notice in this same section that by definition: where $\vec{T}$ is the tangent to the Universe surface at a time t at the neighbourhood of a given point. We had also notice in this same section that by definition:

$$\vec{T}=\sqrt{\frac{\mathrm{d}\vec{X}}{\mathrm{d}s}\circ\frac{\mathrm{d}\vec{X}}{\mathrm{d}s}}=1\Rightarrow\frac{\mathrm{d}\vec{X}}{\mathrm{d}s}\circ\frac{\mathrm{d}\vec{X}}{\mathrm{d}s}=1\tag{12.7.74}$$

But we can write:

$$\frac{\mathrm{d}\vec{X}}{\mathrm{d}s}\circ\frac{\mathrm{d}\vec{X}}{\mathrm{d}s}=\left(\frac{\mathrm{d}\vec{X}}{\mathrm{d}s}\right)^2=\frac{\mathrm{d}\vec{X}}{\mathrm{d}\sigma}\circ\frac{\mathrm{d}\vec{X}}{\mathrm{d}\sigma}=\frac{\mathrm{d}\vec{X}}{\mathrm{d}\sigma}\circ\frac{\mathrm{d}\vec{X}}{\mathrm{d}\sigma}\left(\frac{\mathrm{d}\sigma}{\mathrm{d}s}\right)^2=\frac{\mathrm{d}\vec{X}}{\mathrm{d}\sigma}\frac{\mathrm{d}\vec{X}}{\mathrm{d}\sigma}\left(\frac{\mathrm{d}\sigma}{\mathrm{d}s}\right)^2=1\tag{12.7.75}$$

where it should be remembered that $\partial\vec{X}/\partial\sigma$ is taken at a fixed time t. As the lines of the surface Universe of constant t describe the string, then $\partial\vec{X}/\partial\sigma$ is tangent to the string.

And as:

$$\frac{\mathrm{d}\vec{X}}{\mathrm{d}s}=\frac{\mathrm{d}\vec{X}}{\mathrm{d}\sigma}\frac{\mathrm{d}\vec{\sigma}}{\mathrm{d}s}\tag{12.7.76}$$

Then $\mathrm{d}\vec{X}/\mathrm{d}s$ is collinear to $\partial\vec{X}/\partial\sigma$ and thus also tangential to the string (information that we did not have a few lines before!). These small discovers being made, let us go back to:

$$\begin{aligned}
S &= -\frac{T_0}{c}\int \mathrm{d}t\mathrm{d}\sigma\sqrt{\left(\frac{\partial\vec{X}}{\partial t}\circ\frac{\partial\vec{X}}{\partial\sigma}\right)^2+c^2\left(\frac{\partial\vec{X}}{\partial\sigma}\right)^2-\left(\frac{\partial\vec{X}}{\partial t}\right)^2\left(\frac{\partial\vec{X}}{\partial\sigma}\right)^2}\\
&= -\frac{T_0}{c}\int \mathrm{d}t\mathrm{d}\sigma\sqrt{\left(\frac{\mathrm{d}s}{\mathrm{d}\sigma}\right)^2\left[\left(\frac{\partial\vec{X}}{\partial t}\circ\frac{\partial\vec{X}}{\partial s}\right)^2+c^2\left(\frac{\partial\vec{X}}{\partial s}\right)^2-\left(\frac{\partial\vec{X}}{\partial t}\right)^2\left(\frac{\partial\vec{X}}{\partial s}\right)^2\right]}\\
&= -\frac{T_0}{c}\int \mathrm{d}t\mathrm{d}\sigma\sqrt{\left(\frac{\mathrm{d}s}{\mathrm{d}\sigma}\right)^2\left[\left(\frac{\partial\vec{X}}{\partial t}\circ\frac{\partial\vec{X}}{\partial s}\right)^2+c^2-\left(\frac{\partial\vec{X}}{\partial t}\right)^2\right]}\\
&= -\frac{T_0}{c}\int \mathrm{d}t\mathrm{d}s\sqrt{c^2-\left[\left(\frac{\partial\vec{X}}{\partial t}\right)^2-\left(\frac{\partial\vec{X}}{\partial t}\circ\frac{\partial\vec{X}}{\partial s}\right)^2\right]}
\end{aligned}\tag{12.7.77}$$

it already gets a little more interesting!

Now let consider the following figure:

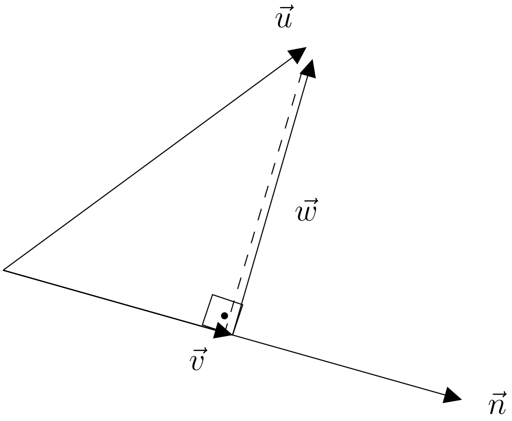


Figure 12.309: **Illustrated recall of the dot product**

where $\vec{u}$ is any vector and $\vec{n}$ a unit (dimensionless) vector and $\vec{v}$, the orthogonal projection of $\vec{u}$ on $\vec{n}$. We then have (see section Vector Calculus page 1821):

$$\vec{u}\circ\vec{n}=\|\vec{u}\|\cdot\|\vec{n}\|\cos(\theta)=(\|\vec{u}\|\cos(\theta))\,\|\vec{n}\|=\|\vec{v}\|\cdot\|\vec{n}\|\tag{12.7.78}$$

Now if we seek for the vector $\vec{v}$ we will have to multiply the whole $\vec{n}$:

$$\vec{v}=(\vec{u}\circ\vec{n})\vec{n}=\|\vec{v}\|\cdot\|\vec{n}\|\vec{n}\tag{12.7.79}$$

Finally, if we seek the expression of the vector $\vec{w}$ it comes immediately:

$$\vec{v}+\vec{w}\Rightarrow\vec{w}=\vec{u}-\vec{v}=\vec{u}-(\vec{u}\circ\vec{n})\vec{n}\tag{12.7.80}$$

Then by similarity, we can write:

$$\vec{w}=\vec{u}-(\vec{u}\circ\vec{n})\vec{n}=\frac{\partial\vec{X}}{\partial t}-\left(\frac{\partial\vec{X}}{\partial t}\circ\frac{\partial\vec{X}}{\partial s}\right)\frac{\partial\vec{X}}{\partial s}\tag{12.7.81}$$

where $\vec{w}$ is perpendicular to $\partial\vec{X}/\partial s$ and as the unit of speed. By construction, $\vec{w}$ is therefore the transverse velocity of the speed at a time t t since $\partial\vec{X}/\partial s$ is tangent thereto. We will denote then:

$$\vec{v}_\perp=\frac{\partial\vec{X}}{\partial t}-\left(\frac{\partial\vec{X}}{\partial t}\circ\frac{\partial\vec{X}}{\partial s}\right)\frac{\partial\vec{X}}{\partial s}\tag{12.7.82}$$

Let us now take, for future needs, the square norm of this last relation (be careful we process the components of the vectors directly by generalizing the vector notation!):

$$\begin{aligned}\vec{v}_\perp&=\left(\frac{\partial\vec{X}}{\partial t}-\left(\frac{\partial\vec{X}}{\partial t}\circ\frac{\partial\vec{X}}{\partial s}\right)\frac{\partial\vec{X}}{\partial s}\right)^2\\&=\left(\frac{\partial\vec{X}}{\partial t}\right)^2-2\left(\frac{\partial\vec{X}}{\partial t}\circ\frac{\partial\vec{X}}{\partial s}\right)\left(\frac{\partial\vec{X}}{\partial s}\circ\frac{\partial\vec{X}}{\partial t}\right)+\left(\frac{\partial\vec{X}}{\partial t}\circ\frac{\partial\vec{X}}{\partial s}\right)^2\frac{\partial\vec{X}}{\partial s}\circ\frac{\partial\vec{X}}{\partial s}\\&=\left(\frac{\partial\vec{X}}{\partial t}\right)^2-2\left(\frac{\partial\vec{X}}{\partial t}\circ\frac{\partial\vec{X}}{\partial s}\right)\left(\frac{\partial\vec{X}}{\partial s}\circ\frac{\partial\vec{X}}{\partial t}\right)+\left(\frac{\partial\vec{X}}{\partial t}\circ\frac{\partial\vec{X}}{\partial s}\right)^2\\&=\left(\frac{\partial\vec{X}}{\partial t}\right)^2-2\left(\frac{\partial\vec{X}}{\partial t}\circ\frac{\partial\vec{X}}{\partial s}\right)^2+\left(\frac{\partial\vec{X}}{\partial t}\circ\frac{\partial\vec{X}}{\partial s}\right)^2\\&=\left(\frac{\partial\vec{X}}{\partial t}\right)^2-\left(\frac{\partial\vec{X}}{\partial t}\circ\frac{\partial\vec{X}}{\partial s}\right)^2\end{aligned}\tag{12.7.83}$$

and if we now go back to:

$$\begin{aligned}S&=-\frac{T_0}{c}\int\mathrm{d}t\mathrm{d}s\sqrt{c^2-\left[\left(\frac{\partial\vec{X}}{\partial t}\right)^2-\left(\frac{\partial\vec{X}}{\partial t}\circ\frac{\partial\vec{X}}{\partial s}\right)^2\right]}=\\&=-\frac{T_0}{c}\int\mathrm{d}t\mathrm{d}s\sqrt{c^2-v_\perp^2}\\&=-T_0\int\mathrm{d}t\mathrm{d}s\sqrt{1-\frac{v_\perp^2}{c^2}}\end{aligned}\tag{12.7.84}$$

The associated Lagrangian is then directly (not to be confused with the Lagrangian density!):

$$\boxed{L=-T_0\int\sqrt{1-\frac{v_\perp^2}{c^2}}\mathrm{d}s}\tag{12.7.85}$$

as:

$$S=\int_{t_i}^{t_f}L\mathrm{d}t=\int_{t_i}^{t_f}\mathrm{d}t\int_0^{\sigma_1}\mathrm{d}\sigma\mathcal{L}\tag{12.7.86}$$

The Lagrangian of the prior-previous relation is considered by the specialists in String Theory as a natural generalization of the Lagrangian of the free particle as we get it in the section of Special Relativity and that was for recall given by:

$$L=-mc^2\sqrt{1-\frac{v^2}{c^2}} \tag{12.7.87}$$

~20 %

★★★★☆

37 votes, 80%

13

Chemistry

Chemistry is the science that studies the nature and properties of simple substances, the molecular action of these bodies on each other and combinations due to this action.
(Larousse)

Contents

13.1 Quantum Chemistry **6123**
  13.1.1 Infinite three-dimensional rectangular potential 6125
  13.1.2 Molecular Vibrations 6128
  13.1.3 Hydrogenoid Atom (Schrödinger model) 6130
    13.1.3.1 Rigid Rotator 6134
    13.1.3.2 Non-Rigid Rotator 6173
13.2 Molecular Chemistry **6190**
  13.2.1 Orbital Approximations 6192
  13.2.2 LCAO Method 6196
  13.2.3 Molecular Rotational Energy Levels 6205
  13.2.4 Molecular Vibrational Energy Levels 6211
  13.2.5 Morse potential 6216
13.3 Analytical Chemistry **6228**
  13.3.1 Analytical chemistry process 6234
  13.3.2 Simple Mixtures 6235
  13.3.3 Chemical formula 6237
  13.3.4 Reactions 6241
13.4 Thermochemistry **6249**
  13.4.1 Chemical transformations 6249
  13.4.2 Molar Quantities 6252
    13.4.2.1 Standard enthalpy of reaction 6259

13.1. Quantum Chemistry

THROUGHOUT human history, people have tried to convert matter into more useful forms. Our Stone Age ancestors chipped pieces of flint into useful tools and carved wood into statues and toys. These endeavours involved changing the shape of a substance without changing the substance itself. But as our knowledge increased, humans began to change the composition of the substances as well—clay was converted into pottery, hides were cured to make garments, copper ores were transformed into copper tools and weapons, and grain was made into bread.

Humans began to practice chemistry when they learned to control fire and use it to cook, make pottery, and smelt metals. Subsequently, they began to separate and use specific components of matter. A variety of drugs such as aloe, myrrh, and opium were isolated from plants. Dyes, such as indigo and Tyrian purple, were extracted from plant and animal matter. Metals were combined to form alloys - for example, copper and tin were mixed together to make bronze - and more elaborate smelting techniques produced iron. Alkalis were extracted from ashes, and soaps were prepared by combining these alkalis with fats. Alcohol was produced by fermentation and purified by distillation.

Attempts to understand the behaviour of matter extend back for more than 2500 years. As early as the 94th century (holocene calendar), Greek philosophers discussed a system in which water was the basis of all things. You may have heard of the Greek postulate that matter consists of four elements: earth, air, fire, and water. Subsequently, an amalgamation of chemical technologies and philosophical speculations were spread from Egypt, China, and the eastern Mediterranean by alchemists, who endeavoured to transform "base metals" such as lead into "noble metals" like gold, and to develop elixirs to cure disease and extend life.

From alchemy came the historical progressions that led to modern chemistry: the isolation of drugs from natural sources, metallurgy, and the dye industry. Today, chemistry continues to deepen our understanding and improve our ability to harness and control the behaviour of matter. This effort has been so successful that many people do not realize either the central position of chemistry among the sciences or the importance and universality of chemistry in daily life as every other scientific field presented in this textbook do!

Before that the reader goes further in reading this chapter, we want to remind that this textbook deals mainly with applied mathematics and theoretical physics. Thus, we will address in this section only of theoretical chemistry (theoretical quantum chemistry, theoretical thermochemistry, theoretical kinetic chemistry, etc).

This choice follows the changes of chemistry since the years 11980 (holocene calendar): form a largely descriptive science descriptive, it tends to become deductive. That is to say that in addition to experience, calculation methods are constantly growing and particularly since the development of modern computing that greatly helps chemists to numerical modelling.

Theoretical chemistry, also named "physical chemistry" - application of methods from physics to chemistry - is too often seen as a discipline in itself. In fact, under this term any modern chemistry field is included. Thus, the investigation of any problem in advanced chemistry requires the assistance of theoretical chemistry (and this is lucky...) and the chemist must have a thorough knowledge of it. At the level of chemistry teaching as secondary branch, the role of physical chemistry is already evident: the result is an increase in the level of students, increase in the abstraction and therefore a risk in alienating the average student. Finally, the purpose is not to burden the knowledge by incorporating more new elements, but to convert the mode of approach of this discipline by substituting the most often encyclopaedic knowledge statements by rational developments based on only a few assumptions and hypothesis that permits to deduce thanks to mathematics many properties thanks to corollaries.

A good understanding of physical chemistry requires in our point of view necessarily to be familiar with quantum physics (see chapter Atomistic page 4612) to have at least one approach to what an atom is and its different electron orbits before talking about connections, different filling methods of electron orbits, redox, filling layers, and others...

In this sense, we will begin by studying the particular case of the hydrogen atom, which is crucial for the whole that will follow (study of polyelectronic atoms). It is therefore necessary for the reader to browse the next lines with all possible attention and to understand as best as possible the subtleties!

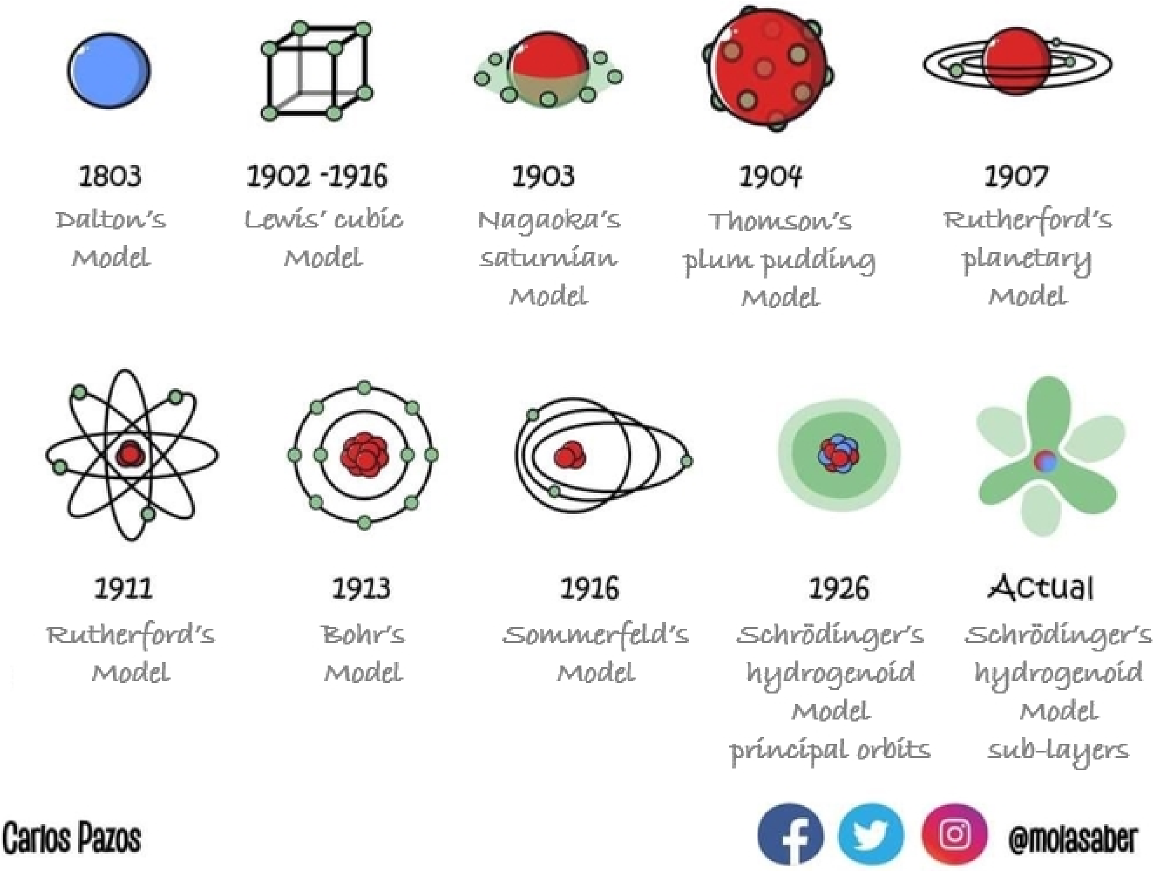

1803
Dalton's Model
1902 -1916
Lewis' cubic Model
1903
Nagaoka's saturnian Model
1904
Thomson's plum pudding Model
1907
Rutherford's planetary Model
1911
Rutherford's Model
1913
Bohr's Model
1916
Sommerfeld's Model
1926
Schrödinger's hydrogenoid Model principal orbits
Actual
Schrödinger's hydrogenoid Model sub-layers
Carlos Pazos
@molasaber

Figure 13.1: **Reminder of atom models evolution**

13.1.1 Infinite three-dimensional rectangular potential

We studied in details in the section of Corpuscular Quantum Physics the Bohr-Sommerfeld hydrogen atom (see page 4648) using the results proved in the section of Special Relativity. This model emerged in a simplistic quantification (but not too much wrong as will discussed later below) of certain properties of matter.

In the section of Wave Quantum Physics, we studied also in details the rectilinear infinite potential well (see page 4796) and the harmonic oscillator (see page 4807) without giving many more examples. Now we will move towards to resolve problems closer to those useful in chemistry with the objective of studying the hydrogen-like atom.

We will now consider a particle moving freely in the three dimensional box below:

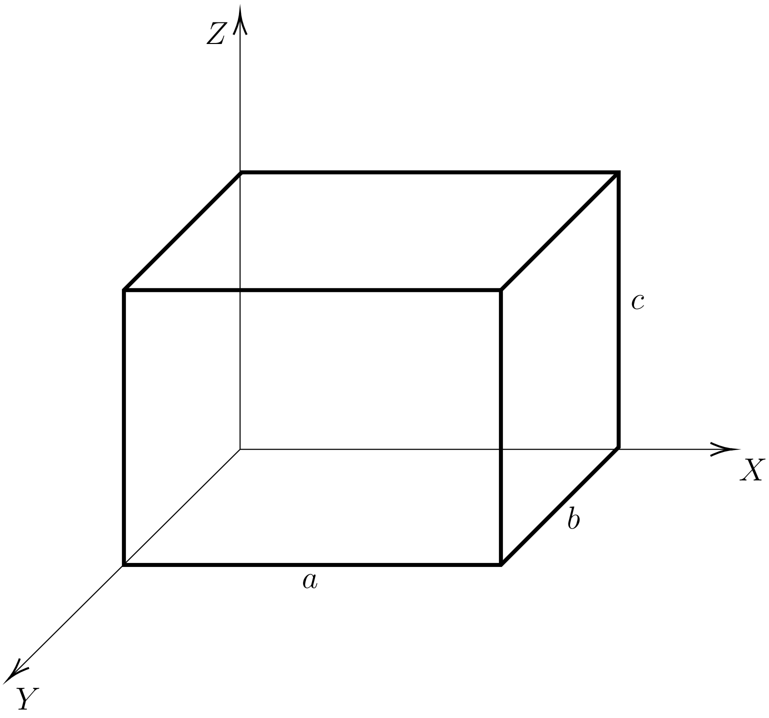


Figure 13.2: **Three dimensional imaginary box in which the particle moves**

The potential electric energy of the system is given by:

$$\begin{aligned} &E_p(x,y,z)=0 && 0<x<a, \quad 0<y<b, \quad 0<z<c, \\ &E_p(x,y,z)=+\infty && \text{everywhere else} \end{aligned} \tag{13.1.1}$$

As in the one-dimensional case (see section Wave Quantum Physics page 4796), the walls of infinite potential prevent the particle from leaving the box, and the wave function is non-zero only for position vector $\vec{r}$ being inside the box. It necessarily vanish when one of the walls is touched. The Schrödinger equation we have to solve is (see section Wave Quantum Physics page 4758):

$$\Delta\Psi+\frac{2m}{\hbar}(E_{\text{tot}}-E_p)\Psi=0 \tag{13.1.2}$$

and boundary conditions are:

$$\begin{gathered}\Psi(x=0,y,z)=\Psi(x=a,y,z)=0\\ \Psi(x,y=0,z)=\Psi(x,y=b,z)=0\\ \Psi(x,y,z=0)=\Psi(x,y,z=c)=0\end{gathered} \tag{13.1.3}$$

Notice that the Hamiltonian can be written as the sum of the Hamiltonian in each axis (we speak of the Hamiltonian operators of course!). So we have:

$$H=H_x+H_y+H_z \tag{13.1.4}$$

where:

$$H_x=\left(\frac{-\hbar^2}{2m}\right)\frac{\partial^2}{\partial^2 x},\quad H_y=\left(\frac{-\hbar^2}{2m}\right)\frac{\partial^2}{\partial^2 y},\quad H_z=\left(\frac{-\hbar^2}{2m}\right)\frac{\partial^2}{\partial^2 z} \tag{13.1.5}$$

Relations that we have proved the origin in details in the section of Wave Quantum Physics of this textbook (see page 4758).

Such a form is named a "separable form": the Hamiltonian is the sum of individual operators H_i each depending only on one variable or degree of freedom q_i. This form reflects the independent nature of the movements described by the variables q_i.

Remember now that the joint probability of two independent events is the product of the individual probabilities of the two events taken separately (see section Probabilities page 480)! We therefore expect that the presence probability density in space (see section Wave Quantum Physics page 4691) with a multidimensional configuration is, in the case where the Hamiltonian is of separable form, a simple individual probability density product. In fact, the separable form of the Hamiltonian permits the separation of variables on the wave function itself!

Let us write the solutions of the Schrödinger equation under the form:

$$\Psi(x,y,z)=\xi(x)\vartheta(y)\zeta(z) \tag{13.1.6}$$

of a product of three factors each depending only of one coordinated.

Substituting this notation in the Schrödinger equation, we get without technical developments (elementary algebra):

$$\vartheta(y)\zeta(z)(H_x\xi(x))+\xi(x)\zeta(z)(H_y\vartheta(y))+\xi(x)\theta(y)(H_z\zeta(z))=(E_{\text{tot}}-E_p)\xi(x)\vartheta(y)\zeta(z) \tag{13.1.7}$$

or, by dividing both sides of this equation by $\xi(x)\vartheta(y)\zeta(z)$:

$$\frac{1}{\xi(x)}(H_x\xi(x))+\frac{1}{\vartheta(y)}(H_y\vartheta(y))+\frac{1}{\zeta(z)}(H_z\zeta(z))=(E_{\text{tot}}-E_p) \tag{13.1.8}$$

which is a much more aesthetic and easier to remember.

This equation requires that the sum of the three terms in the left-hand side is equal to a constant in the context of a conservative system (that is what often interested chemists)! Each of these three terms depending only on one and only one variable, so that their sum is equal to a constant, it is necessary that each term is itself constant! In fact, by taking the derivative of both sides of the above equation with respect to x, for example, we have:

$$\frac{\mathrm{d}\left(\frac{1}{\xi(x)}(H_x\xi(x))\right)}{\mathrm{d}x}=0 \tag{13.1.9}$$

meaning that equation although must be a constant which we will denote equation (as this term expresses an energy). We then have (surprise...):

$$H_x\xi(x)=E_x\xi(x) \tag{13.1.10}$$

Similarly, we get:

$$H_y\theta(y)=E_y\vartheta(y) \quad \text{and} \quad H_z\zeta(z)=E_z\zeta(z) \tag{13.1.11}$$

Notice that each of the separate equations that we have just obtained, for the movement of the particle in the three spatial directions, is a Schrödinger equation in a one-dimensional box. Thus, the three relations previously obtained independently describe each movement in the respective x,y,z directions, limited to the respective ranges:

$$[0,a],[0,b],[0,c] \tag{13.1.12}$$

and must be respectively solved with boundary conditions:

$$\begin{gathered}\xi(0)=\xi(a)=0\\ \theta(0)=\vartheta(b)=0\\ \zeta(0)=\zeta(c)=0\end{gathered} \tag{13.1.13}$$

The results obtained in the section of Wave Quantum Physics when solving the Schrödinger equation in the case of straight wells gave us directly for reminder (see page 4797):

$$\begin{aligned}\xi_\lambda(x)&=\sqrt{\frac{2}{a}}\sin\left(\frac{\lambda\pi x}{a}\right) & E_{x,\lambda}&=\frac{\lambda^2h^2}{8ma^2} &&\text{with}\quad \lambda\in\mathbb{N}^*\\ \vartheta_\mu(y)&=\sqrt{\frac{2}{b}}\sin\left(\frac{\mu\pi y}{b}\right) & E_{y,\mu}&=\frac{\mu^2h^2}{8mb^2} &&\text{with}\quad \mu\in\mathbb{N}^*\\ \zeta_\nu(z)&=\sqrt{\frac{2}{c}}\sin\left(\frac{\nu\pi z}{c}\right) & E_{z,\nu}&=\frac{\nu^2h^2}{8mc^2} &&\text{with}\quad \nu\in\mathbb{N}^*\end{aligned} \tag{13.1.14}$$

In summary the stationary states of the particle in the three-dimensional box are specified by three positive integers quantum numbers λ,μ,ν. The wave function is finally:

$$\boxed{\Psi_{\lambda\mu\nu}=\sqrt{\frac{8}{abc}}\sin\left(\frac{\lambda\pi x}{a}\right)\sin\left(\frac{\mu\pi y}{b}\right)\sin\left(\frac{\nu\pi z}{c}\right)} \tag{13.1.15}$$

and its respective energies (eigenvalues):

$$\boxed{E_{\lambda\mu\nu}=\frac{h^2}{8m}\left(\frac{\lambda^2}{a^2}+\frac{\mu^2}{b^2}+\frac{\nu^2}{c^2}\right)} \tag{13.1.16}$$

The variable separation technique detailed above, is applicable only because the Hamiltonian is in separable form. It comes automatically therefore the three-dimensional probability density $|\Psi(x,y,z)|^2$ is the product of probability density $|\xi_\lambda(x)|^2,|\vartheta_\mu(y)|^2,|\zeta_\nu(z)|^2$, as we had anticipated it. We also notice that the energy of movement in three dimensional space is the sum of energy movements in all three spatial directions: the independence of these three directions or degrees of freedom, implies the additivity of their energy.

13.1.2 Molecular Vibrations

We studied in the Wave Quantum Physics section the harmonic oscillator (see page 4807). Now it is molecular chemistry that we will use all the power of the results obtained during the study of this system.

The harmonic oscillator is a model of molecular vibrations, and is represented by a type of parabolic potential as:

$$E_p = \frac{1}{2}c^{te}x^2 \tag{13.1.17}$$

for a diatomic molecule. But we have proved (see section Wave Mechanics page 3418) that $c^{te} = m\omega_0^2$ so that we finally have for a diatomic molecule:

$$E_p = \frac{m\omega_0^2}{2}x^2 \tag{13.1.18}$$

For a polyatomic molecule, we have verbatim (by the additivity of energy):

$$E_p(q_i) = \frac{1}{2}\sum_i \omega_i^2 q_i^2 \tag{13.1.19}$$

Quantities ω_0 and ω_i are the vibration frequencies (or rather more correctly: the pulsation) of a molecule, diatomic in the first case and polyatomic in the second case. In the first equation, the variable x represents the elongation of the bond between the two atoms A and B (as with a spring) in a diatomic molecule, that is to say $x = R - R_{eq}$, where R is the instantaneous length of this bond, and R_{eq} is its equilibrium value.

In the case of a polyatomic molecule, the potential describing molecular vibrations takes a separable form in terms of summation above only if one considers special variables q_i denoting collective motions of nuclei, and which are named "normal vibration modes".

We also saw in the section of Wave Quantum Physics (see page 4807) that the Hamiltonian of a diatomic molecule (problem of the harmonic oscillator) can be written as

$$H = -\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}x^2} + \frac{m\omega^2}{2}x^2 \tag{13.1.20}$$

For a polyatomic molecule that relationship becomes logically:

$$H = \sum_i \left(-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial q_i^2} + \frac{m\omega^2}{2}q_i^2\right) \tag{13.1.21}$$

The Hamiltonian above is clearly a type of separable form: it is a sum of one-dimensional Hamiltonians, each depending only on a single mode q_i as variable, describing this mode as a unique spring or harmonic unit mass ($m = 1$) oscillator and of pulse oscillation ω_i. Therefore, a separation of variables q_i is possible, reducing the Schrödinger time independent into a number of equations of the same type as that of a one-dimensional harmonic oscillator. So we need just to know the expression of the wave function for a one-dimensional harmonic oscillator, what we already have done in the section of Wave Quantum Physics (see page 4807) where we got:

$$E_n = \hbar\omega\left(n + \frac{1}{2}\right) \quad \text{with} \quad n \in \mathbb{N} \tag{13.1.22}$$

and:

$$\Psi_n = \left(A\sqrt{\pi}n!2^n\right)^{-1/2} e^{-Q^2/2} H_n(Q) \tag{13.1.23}$$

The figure below shows the graph of the first wave functions of the above relation as well as that of their respective presence probability densities. We can see the same modal structures as those specific to functions of a particle in a one-dimensional box:

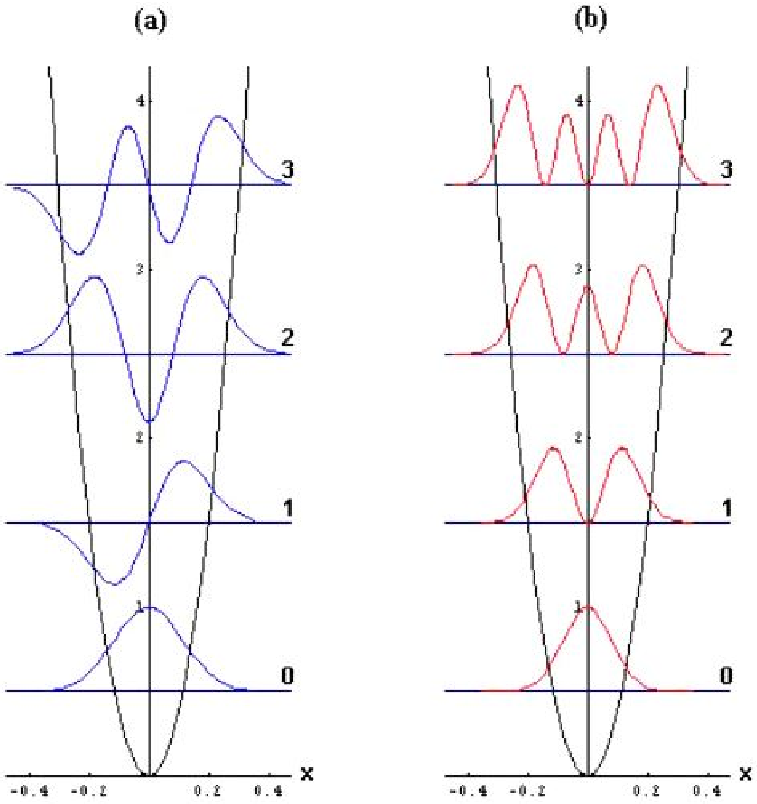


Figure 13.3: **Wave functions and probability density of a one-dimensional harmonic oscillator**

Above the first energy levels of a one-dimensional oscillator with `(a)` their associated eigenfunction, `(b)` the associated probability distribution of presence.

In the limit of very large values of n, the probability distribution approximates more and more of that predicted by classical mechanics, the oscillator lies for the most of the time in the vicinity of the turning points defined by the intersection of potential E_p with the level of n. This trend is illustrated below:

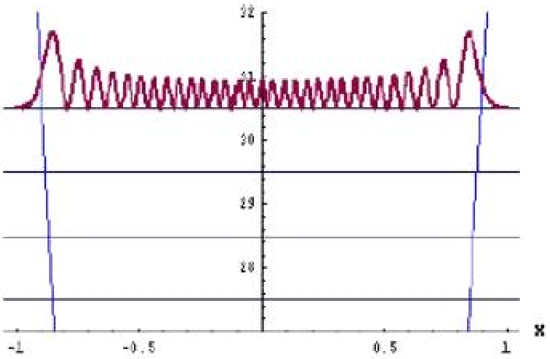


Figure 13.4: Probability density function of a one-dimensional harmonic oscillator for large n

For a polyatomic molecule the expression of quantified energy therefore becomes:

$$\boxed{E_n=\sum_i \hbar\omega_i\left(n_i+\frac{1}{2}\right)} \quad \text{with} \quad n\in\mathbb{N}, \forall i=1,2,3\ldots,n \tag{13.1.24}$$

and eigenfunctions/eigenstates become:

$$\Psi_n(q_i)=(A\sqrt{\pi}n!2^n)^{-1/2}\prod_i H_{n_i}(Q_i)e^{-Q_i^2/2} \tag{13.1.25}$$

with:

$$Q_i=\sqrt{\frac{\omega_i}{\hbar}}q_i \tag{13.1.26}$$

The last two relations are very important because they allow among others to:

- Predict the spectrum of the molecule (spectroscopy)
- To study the energy bands (where does the bands of valence and conduction comes from)
- To locate the bonds between atoms and thus the chemical properties

We will go further into the subject of molecular vibrations later below at page 6205 during our study of molecular vibrations energy levels that will be very important in practice to predict the spectrum of molecules (or respectively their moment of inertia).

13.1.3 Hydrogenoid Atom (Schrödinger model)

We consider here the quantification of a generic system made of two bodies (particles) interacting with each other and moving in a three-dimensional space. We will be prove at first that, even if the separation of dynamic variables describing individually each of the two bodies is impossible, for cons, the overall movement system (the center of mass) and internal movement, also said "relative motion", are separable. In addition, if the potential is centrosymmetric, the internal movement may also be decomposed into a rotational movement and radial movement. The quantification of the rotational movement is intimately connected to that of angular momentum.

Here we focus on the mechanics of an atomic system having only one electron. This is a two-particle system: a nucleus of mass M and charge $+Zq_e$, and an electron of mass m_e and of charge $-q_e$ in the framework of the Schrödinger model (for a more generalized treatment based on Dirac equation see [460]).

The atomic system is described by the following Hamiltonian:

$$\begin{aligned}H&=E_c^{\text{nucleus}}+E_c^{\text{electron}}+E_p^{\text{system}}=\frac{p_N^2}{2M}+\frac{p_e^2}{2m_e}+E_p(\vec{r}_e-\vec{R}_N)\\&=-\frac{\hbar^2}{2M}\nabla_N^2-\frac{\hbar^2}{2m_e}\nabla_e^2+E_p(\vec{r}_e-\vec{R}_N)\end{aligned}\tag{13.1.27}$$

Remember that in the section of Wave Quantum Physics we had proved during our study of functional operators (see page 4714):

$$p=\frac{1}{\mathrm{i}}\hbar\nabla=\frac{1}{\mathrm{i}}\frac{h}{2\pi}\nabla\tag{13.1.28}$$

and remember also that $\vec{r}_e$ and $\vec{R}_n$ are respectively the position vectors of the electron and nucleus in the prior-previous relation.

The potential electric energy being given by (see section Electrostatics page 4005):

$$E_p(\vec{r}_e-\vec{R}_N)=-\frac{Zq_e^2}{4\pi\varepsilon\|\vec{r}_e-\vec{R}_N\|}\tag{13.1.29}$$

The movements of the two particles are correlated because the two charges interact through their mutual electrical field. We can not make a separation between variables $\vec{r}_e$ and $\vec{R}_N$. By cons, a separation of variables is possible with the coordinate of the center of mass (see the definition of the center of mass in the section of Classical Mechanics at page 3213):

$$\vec{R}_{\text{CM}}=\frac{m_e\vec{r}_e+M\vec{R}_N}{m_e+M}\tag{13.1.30}$$

and the relative coordinate of the electron relative to the nucleus:

$$\vec{r}_{\text{rel}}=\vec{r}_e-\vec{R}_N\tag{13.1.31}$$

We get therefore:

$$\vec{p}_{\text{CM}}=\frac{\hbar}{\mathrm{i}}\vec{\nabla}_{\text{CM}}=\frac{\hbar}{\mathrm{i}}\left(\vec{\nabla}_e+\vec{\nabla}_N\right)=\vec{p}_e+\vec{p}_N\tag{13.1.32}$$

and:

$$\vec{p}_{\text{rel}}=\frac{M\vec{p}_e-m_e\vec{p}_N}{m_e+M}\tag{13.1.33}$$

The Hamiltonian in the center of mass repository will therefore be written:

$$H=\frac{p_{\text{CM}}^2}{2M_{\text{tot}}}+\frac{p_{\text{rel}}^2}{2\mu}+E_{\text{tot}}(\vec{r}_{\text{rel}})\tag{13.1.34}$$

where $M_{\text{tot}} = m_e + M$ is the total mass of the system and:

$$\mu = \frac{m_e M}{M_{\text{tot}}} \tag{13.1.35}$$

is its reduced mass.

We clearly see that the Hamiltonian H is this time set in a separable form and we can write it as follows:

$$H = H_{\text{CM}} + H_{\text{rel}} \tag{13.1.36}$$

with:

$$H_{\text{CM}} = \frac{p_{\text{CM}}^2}{2M_{\text{tot}}} \quad \text{and} \quad H_{\text{rel}} = \frac{p_{\text{rel}}^2}{2\mu} + E_{\text{tot}}(\vec{r}_{\text{rel}}) \tag{13.1.37}$$

In terms of the coordinates $\vec{R}_{\text{CM}}$ and $\vec{r}_{\text{rel}}$, the function describing a stationary state of the two-body system is a product of individual wave functions (recall that the joint probability of two events is the product of probabilities), one for the movement of the center of mass and the other for the relative movement:

$$\Psi_{\text{tot}}\left(\vec{R}_{\text{CM}}, \vec{r}_{\text{rel}}\right) = \Psi_{\text{CM}}\left(\vec{R}_{\text{CM}}\right)\Psi_{\text{rel}}\left(\vec{r}_{\text{rel}}\right) \tag{13.1.38}$$

and the energy of this state is the sum of the respective energies of movements:

$$E_{\text{tot}} = E_{\text{CM}} + E_{\text{rel}} \tag{13.1.39}$$

with:

$$H_{\text{CM}}\Psi_{\text{CM}}\left(\vec{R}_{\text{CM}}\right) = E_{\text{CM}}\Psi_{\text{CM}}\left(\vec{R}_{\text{CM}}\right) \quad \text{and} \quad H_{\text{rel}}\Psi_{\text{rel}}\left(\vec{r}_{\text{rel}}\right) = E_{\text{rel}}\Psi_{\text{rel}}\left(\vec{r}_{\text{rel}}\right) \tag{13.1.40}$$

Remark

This approach of separating the wave function into the composition of a wave function of the center of mass and the relative movement is also used in the context of the study of poly-electronic atoms, but with one difference: as the nucleus is much more massive than the processing electrons (in approximation ...), the center of mass is assimilated to the nucleus of the atom and the relative motion to the entire electron cloud[a]. This approximate approach is well known under the designation "Born-Oppenheimer approximation".

---

[a] The electron cloud is the region of negative charge surrounding an atomic nucleus that is associated with an atomic orbital. The region is defined mathematically, describing a region with a high probability of containing electrons. As we know it, the electron cloud model differs from the more simplistic Bohr model, in which electrons orbit the nucleus in much the same way as planets orbit the Sun. In the cloud model, there are regions where an electron may likely be found, but it's theoretically possible for it to be located anywhere, including inside the nucleus.

Where the Hamiltonian appearing in the first of these relations has been defined above as being:

$$H_{\text{CM}} = \frac{p_{\text{CM}}^2}{2M_{\text{tot}}} \tag{13.1.41}$$

This movement is that of a particle of mass M_{tot} in a three-dimensional box of infinite volume. Eigenvalues and eigenfunctions for this movement has already been obtained in our previous study, we will restrict ourselves to the study of separate equation for the relative movement, or internal movement. As no confusion will be possible between different Hamiltonians, we let down, to simplify the notations, the "eel" word in subscript.

With H_{rel} given by the relation that we have proved previously:

$$H_{\text{rel}} = \frac{p_{\text{rel}}^2}{2\mu} + E_p(\vec{r}_{\text{rel}}) \tag{13.1.42}$$

and the relation (as proved above):

$$H_{\text{rel}}\Psi_{\text{rel}}(\vec{r}_{\text{rel}}) = E_{\text{rel}}\Psi_{\text{rel}}(\vec{r}_{\text{rel}}) \tag{13.1.43}$$

then we obtain the Schrödinger equation for the relative motion:

$$\left(\frac{p^2}{2\mu} + E_p(r)\right)\Psi(\vec{r}) = E\Psi(\vec{r}) \tag{13.1.44}$$

or written differently:

$$\left(-\frac{\hbar}{2\mu}\nabla^2 + E_p(r)\right)\Psi(\vec{r}) = E\Psi(\vec{r}) \tag{13.1.45}$$

Notice that in the case where the potential energy E_p is a centrosymmetric source, that is to say it depends only on the length of the position vector $\vec{r}$, and not its orientation, the previous equation, written in Cartesian coordinates, is inseparable. Indeed, in Cartesian coordinates, the length $\vec{r}$ is given by:

$$\|\vec{r}\| = \sqrt{x^2 + y^2 + z^2} \tag{13.1.46}$$

and the potential energy can not be separated into three components, each depending only one of the three variables x, y, z. The Hamiltonian is therefore still not a separable form and so we did not meet our target. However, the above equation is separable at the moment we make a change of coordinates to spherical coordinates. Indeed, in this coordinate system, the potential depends on only on one of the three spherical variables: the radius r. It is independent of the two angles θ and ϕ.

Referring to the result obtained in the study of the Laplace expressions in different coordinate systems, in the section of Vector Calculus, we got for the Laplacian of a scalar field in spherical coordinates the following expression (see page 1899):

$$\vec{\nabla}^2 u = \frac{\partial^2 u}{\partial r^2} + 2\frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} + \frac{\cot(\phi)}{r^2}\frac{\partial u}{\partial \theta} + \frac{1}{r^2\sin^2(\theta)}\frac{\partial^2 u}{\partial \phi^2} \tag{13.1.47}$$

The Hamiltonian:

$$H = -\frac{\hbar}{2\mu}\vec{\nabla}^2 + E_p(\vec{r}) \tag{13.1.48}$$

then becomes (simple distribution and new way to note):

$$\boxed{H = T_r + \frac{L^2}{2\mu r^2}E_p(\vec{r})} \tag{13.1.49}$$

where:

$$\boxed{T_r=-\frac{\hbar}{2\mu}\left(\frac{\partial}{\partial r^2}+2\frac{1}{r}\frac{\partial}{\partial r}\right)=-\frac{\hbar}{2\mu}\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right)} \tag{13.1.50}$$

is the kinetic energy operator for the radial movement of the electron relative to the nucleus, and L^2 is the squared "associated" operator of the angular momentum vector:

$$L^2=-\hbar^2\left[\frac{1}{\sin(\theta)}\frac{\partial}{\partial\theta}\left(\sin(\theta)\frac{\partial}{\partial\theta}\right)+\frac{1}{\sin^2(\theta)}\frac{\partial^2}{\partial\phi^2}\right] \tag{13.1.51}$$

The term:

$$H=\frac{L^2}{2J} \tag{13.1.52}$$

where $J=\mu r^2$ is therefore an energy associated with the angular momentum L (see section Classical Mechanics page 3297 and page 3244).

To understand the nature of this operator L^2 a detour by the notion of rigid rotor will help.

13.1.3.1 Rigid Rotator

If we now consider the case of a system named "rigid rotor" where we neglect ("restrict" would be a more appropriate term ...) the degrees of freedom of oscillation (this is the system that are the study case for linear diatomic or polyatomic molecules), the only coordinates being into play are the angles θ and ϕ which fix the orientation of the rotator.

Thus, in this case r is fixed and we have:

$$T_r=0, E_p(r)=\begin{cases}0 & r=R_e\\ +\infty & r\neq R_e\end{cases} \tag{13.1.53}$$

and in view of the constraints on the potential, it is normally quite easy to understand why the rotator is said to be "rigid". In the above case, the Hamiltonian is reduced to:

$$H=\frac{L^2}{2J}=\frac{L^2}{2\mu r^2} \tag{13.1.54}$$

For the rest, we associate the operator L^2 to the square of an angular momentum, for the simple reason that he has the units of it... Indeed, let us recall that we have prove in the section of Wave Quantum Physics that when the spin is zero (so as part of our study of the hydrogenoid atom here, the spin will not be taken into account in the first instance) and that we are dealing with a single particle then the angular momentum (which we will denote by L instead of b) is given by:

$$L\leqslant\sqrt{l(l+1)}\hbar\Rightarrow L^2\leqslant l(l+1)\hbar^2\quad\text{and}\quad\vec{L}=\hbar\vec{l}\text{ with }l\in\mathbb{N} \tag{13.1.55}$$

where the components of the vector $\vec{l}$ are also natural numbers. By doing this similarity, we can then write the Schrödinger equation in the form:

$$\left[T_r+\frac{l(l+1)\hbar^2}{2\mu r^2}+E_p\right]\Psi=E\Psi \tag{13.1.56}$$

Let us recall that we got in the section of Wave Quantum Physics that (see page 4877):

$$L_x=\frac{\hbar}{\mathrm{i}}(y\partial_z-z\partial_y) \tag{13.1.57}$$

by the vector product.

We go now to rectangular coordinates x,y,z coordinates to spherical coordinates r,θ,ϕ. Remember for this (see section Vector Calculus page 1849) that:

$$\begin{aligned}
r&=\sqrt{x^2+y^2+z^2} \qquad 0<r<+\infty\\
x&=r\sin(\theta)\cos(\phi)\\
y&=r\sin(\theta)\sin(\phi)\\
z&=r\cos(\theta)
\end{aligned} \tag{13.1.58}$$

and that:

$$\begin{aligned}
\theta&=\arccos\left(\frac{z}{\sqrt{x^2+y^2+z^2}}\right) \qquad 0\leq\theta\leq\pi\\
\phi&=\arccos\left(\frac{x}{\sqrt{x^2+y^2}}\right) \qquad 0\leq\phi\leq 2\pi
\end{aligned} \tag{13.1.59}$$

Now let us express the total differentials:

$$\begin{aligned}
\mathrm{d}x&=\sin(\theta)\cos(\phi)\mathrm{d}r+\cos(\theta)\cos(\phi)r\mathrm{d}\theta-r\sin(\theta)\sin(\phi)\mathrm{d}\phi\\
\mathrm{d}y&=\sin(\theta)\sin(\phi)\mathrm{d}r+r\cos(\theta)\sin(\phi)\mathrm{d}\theta+r\sin(\theta)\cos(\phi)\mathrm{d}\phi\\
\mathrm{d}z&=\cos(\theta)\mathrm{d}r-r\sin(\theta)\mathrm{d}\theta
\end{aligned} \tag{13.1.60}$$

These relations can be written as an orthogonal transformation of the total differential $\mathrm{d}r, r\mathrm{d}\theta, r\sin(\theta)\mathrm{d}\phi$ by:

$$\begin{pmatrix}\mathrm{d}x\\ \mathrm{d}y\\ \mathrm{d}z\end{pmatrix}=\begin{bmatrix}\sin(\theta)\cos(\phi) & \cos(\theta)\cos(\phi) & -\sin(\phi)\\ \sin(\theta)\sin(\phi) & \cos(\theta)\sin(\phi) & \cos(\phi)\\ \cos(\theta) & -\sin(\theta) & 0\end{bmatrix}\begin{pmatrix}\mathrm{d}r\\ r\mathrm{d}(\theta)\\ r\sin(\theta)\mathrm{d}\phi\end{pmatrix} \tag{13.1.61}$$

or by the inverse transformation (if required ... it is enough to check that the two transformation matrices multiplied together give the identity matrix):

$$\begin{pmatrix}\mathrm{d}r\\ r\mathrm{d}\theta\\ r\sin(\theta)\mathrm{d}\phi\end{pmatrix}=\begin{bmatrix}\sin(\theta)\cos(\phi) & \sin(\theta)\sin(\phi) & \cos(\phi)\\ \cos(\theta)\cos(\phi) & \cos(\theta)\sin(\phi) & -\sin(\phi)\\ -\sin(\theta) & \cos(\phi) & 0\end{bmatrix}\begin{pmatrix}\mathrm{d}x\\ \mathrm{d}y\\ \mathrm{d}z\end{pmatrix} \tag{13.1.62}$$

It results of this for example:

$$\begin{aligned}
\mathrm{d}r&=\sin(\theta)\cos(\phi)\mathrm{d}x+\sin(\theta)\sin(\phi)\mathrm{d}y+\cos(\theta)\mathrm{d}z\Rightarrow\frac{\partial r}{\partial x}=\sin(\theta)\cos(\phi)\\
r\mathrm{d}\theta&=\cos(\theta)\cos(\phi)\mathrm{d}x+\cos(\theta)\sin(\phi)\mathrm{d}y-\sin(\theta)\mathrm{d}z\Rightarrow\frac{\partial\theta}{\partial x}=\frac{1}{r}\cos(\theta)\cos(\phi)\\
r\sin(\theta)\mathrm{d}\phi&=-\sin(\phi)\mathrm{d}x+\cos(\phi)\mathrm{d}y+0\mathrm{d}z\Rightarrow\frac{\partial\phi}{\partial x}=\cos(\theta)
\end{aligned} \tag{13.1.63}$$

and finally (the method for the second and third lines is the same as for the first!):

$$\begin{aligned}
\frac{\partial}{\partial x}&=\frac{\partial r}{\partial x}\frac{\partial}{\partial r}+\frac{\partial \theta}{\partial x}\frac{\partial}{\partial \theta}+\frac{\partial \varphi}{\partial x}\frac{\partial}{\partial \varphi}=\sin(\theta)\cos(\phi)\frac{\partial}{\partial r}+\cos(\theta)\cos(\phi)\frac{1}{r}\frac{\partial}{\partial \theta}-\frac{\sin(\phi)}{r\sin(\theta)}\frac{\partial}{\partial \varphi}\\
\frac{\partial}{\partial y}&=\frac{\partial r}{\partial y}\frac{\partial}{\partial r}+\frac{\partial \theta}{\partial y}\frac{\partial}{\partial \theta}+\frac{\partial \varphi}{\partial y}\frac{\partial}{\partial \varphi}=\sin(\theta)\sin(\phi)\frac{\partial}{\partial r}+\cos(\theta)\sin(\phi)\frac{1}{r}\frac{\partial}{\partial \theta}+\frac{\cos(\phi)}{r\sin(\theta)}\frac{\partial}{\partial \varphi}\\
\frac{\partial}{\partial z}&=\frac{\partial r}{\partial z}\frac{\partial}{\partial r}+\frac{\partial \theta}{\partial z}\frac{\partial}{\partial \theta}+\frac{\partial \varphi}{\partial z}\frac{\partial}{\partial \varphi}=\cos(\theta)\frac{\partial}{\partial r}-\frac{\sin(\theta)}{r}\frac{\partial}{\partial \theta}
\end{aligned} \tag{13.1.64}$$

Thus, taking into account these relations, we get for example, in the case of the operator:

$$L_x=-\frac{\hbar}{\mathrm{i}}(y\partial_z-z\partial_y) \tag{13.1.65}$$

the following developments:

$$\begin{aligned}
y\frac{\partial}{\partial z}-z\frac{\partial}{\partial y}=&\,y\left(\cos(\theta)\frac{\partial}{\partial r}-\frac{\sin(\theta)}{r}\frac{\partial}{\partial \theta}\right)-z\left(\sin(\theta)\sin(\phi)\frac{\partial}{\partial r}+\frac{\cos(\theta)\sin(\phi)}{r}\frac{\partial}{\partial \theta}+\frac{\cos(\phi)}{r\sin(\theta)}\frac{\partial}{\partial \phi}\right)\\
&=r\sin(\theta)\sin(\phi)\left(\cos(\theta)\frac{\partial}{\partial r}-\frac{\sin(\theta)}{r}\frac{\partial}{\partial \theta}\right)\\
&-r\cos(\theta)\left(\sin(\theta)\sin(\phi)\frac{\partial}{\partial r}+\frac{\cos(\theta)\sin(\phi)}{r}\frac{\partial}{\partial \theta}+\frac{\cos(\phi)}{r\sin(\theta)}\frac{\partial}{\partial \phi}\right)\\
&=r\sin(\theta)\sin(\phi)\left(-\frac{\sin(\theta)}{r}\frac{\partial}{\partial \theta}\right)-r\cos(\theta)\left(\frac{\cos(\theta)\sin(\phi)}{r}\frac{\partial}{\partial \theta}+\frac{\cos(\phi)}{r\sin(\theta)}\frac{\partial}{\partial \phi}\right)\\
&=\sin(\theta)\sin(\phi)\left(-\frac{\sin(\theta)}{r}\frac{\partial}{\partial \theta}\right))-\cos(\theta)\left(\cos(\theta)\sin(\phi)\frac{\partial}{\partial \theta}+\frac{\cos(\phi)}{\sin(\theta)}\frac{\partial}{\partial \phi}\right)\\
&=(-\sin(\theta)\sin(\theta)\sin(\phi)-\cos(\theta)\cos(\theta)\sin(\phi))\frac{\partial}{\partial \theta}-\cos(\theta)\left(\frac{\cos(\phi)}{\sin(\theta)}\frac{\partial}{\partial \phi}\right)\\
&=-\sin(\phi)\frac{\partial}{\partial \theta}-\cot(\theta)\cos(\phi)\frac{\partial}{\partial \phi}
\end{aligned} \tag{13.1.66}$$

which gives the following result:

$$L_x=-\mathrm{i}\hbar(\sin(\phi)\partial_\theta-\cot(\theta)\cos(\phi)\partial_\phi) \tag{13.1.67}$$

By doing the same with:

$$L_y=\frac{\hbar}{\mathrm{i}}(z\partial_x-x\partial_z)=-\mathrm{i}\hbar(z\partial_x-x\partial_z) \tag{13.1.68}$$

and processing with the same developments:

$$\begin{aligned}
z\frac{\partial}{\partial x}-x\frac{\partial}{\partial z}&=z\left(\sin(\theta)\cos(\phi)\frac{\partial}{\partial r}+\frac{\cos(\theta)\cos(\phi)}{r}\frac{\partial}{\partial\theta}-\frac{\sin(\phi)}{r\sin(\theta)}\frac{\partial}{\partial\phi}\right)-x\left(\cos(\theta)\frac{\partial}{\partial r}-\frac{\sin(\theta)}{r}\frac{\partial}{\partial\theta}\right)\\
&=r\cos(\theta)\left(\sin(\theta)\cos(\phi)\frac{\partial}{\partial r}+\frac{\cos(\theta)\cos(\phi)}{r}\frac{\partial}{\partial\theta}-\frac{\sin(\phi)}{r\sin(\theta)}\frac{\partial}{\partial\phi}\right)\\
&-r\sin(\theta)\cos(\phi)\left(\cos(\theta)\frac{\partial}{\partial r}-\frac{\sin(\theta)}{r}\frac{\partial}{\partial\theta}\right)\\
&=r\cos(\theta)\left(\frac{\cos(\theta)\cos(\phi)}{r}\frac{\partial}{\partial\theta}-\frac{\sin(\phi)}{r\sin(\theta)}\frac{\partial}{\partial\phi}\right)-r\sin(\theta)\cos(\phi)\left(-\frac{\sin(\theta)}{r}\frac{\partial}{\partial\theta}\right)\\
&=\cos(\theta)\left(\cos(\theta)\cos(\phi)\frac{\partial}{\partial\theta}-\frac{\sin(\phi)}{\sin(\theta)}\frac{\partial}{\partial\phi}\right)-\sin(\theta)\cos(\phi)\left(-\sin(\theta)\frac{\partial}{\partial\theta}\right)\\
&=(\cos(\theta)\cos(\theta)\cos(\phi)+\sin(\theta)\cos(\phi)\sin(\theta))\frac{\partial}{\partial\theta}-\cos(\theta)\frac{\sin(\phi)}{\sin(\theta)}\frac{\partial}{\partial\phi}\\
&=\cos(\phi)(\cos^2(\theta)+\sin^2(\theta))\frac{\partial}{\partial\theta}-\sin(\phi)\cot(\theta)\frac{\partial}{\partial\phi}\\
&=\cos(\phi)\frac{\partial}{\partial\theta}-\sin(\phi)\cot(\theta)\frac{\partial}{\partial\phi}
\end{aligned}\tag{13.1.69}$$

we get the following result:

$$L_y=-\mathrm{i}\hbar\left(\cos(\phi)\partial_\theta-\sin(\phi)\cot(\theta)\partial_\phi\right)\tag{13.1.70}$$

And for finish with:

$$L_z=\frac{\hbar}{\mathrm{i}}(x\partial_y-y\partial_x)=-\mathrm{i}\hbar(x\partial_y-y\partial_x)\tag{13.1.71}$$

and again processing the same developments:

$$\begin{aligned}
x\frac{\partial}{\partial y}-y\frac{\partial}{\partial x}&=x\left(\sin(\theta)\sin(\phi)\frac{\partial}{\partial r}+\frac{\cos(\theta)\sin(\phi)}{r}\frac{\partial}{\partial\theta}+\frac{\cos(\phi)}{r\sin(\theta)}\frac{\partial}{\partial\phi}\right)\\
&-y\left(\sin(\theta)\cos(\phi)\frac{\partial}{\partial r}+\frac{\cos(\theta)\cos(\phi)}{r}\frac{\partial}{\partial\theta}-\frac{\sin(\phi)}{r\sin(\theta)}\frac{\partial}{\partial\phi}\right)\\
&=r\sin(\theta)\cos(\phi)\left(\sin(\theta)\sin(\phi)\frac{\partial}{\partial r}+\frac{\cos(\theta)\sin(\phi)}{r}\frac{\partial}{\partial\theta}+\frac{\cos(\phi)}{r\sin(\theta)}\frac{\partial}{\partial\phi}\right)\\
&-r\sin(\theta)\sin(\phi)\left(\sin(\theta)\cos(\phi)\frac{\partial}{\partial r}+\frac{\cos(\theta)\cos(\phi)}{r}\frac{\partial}{\partial\theta}-\frac{\sin(\phi)}{r\sin(\theta)}\frac{\partial}{\partial\phi}\right)\\
&=r\sin(\theta)\cos(\phi)\left(\frac{\cos(\theta)\sin(\phi}{2}\frac{\partial}{\partial\theta}+\frac{\cos(\phi)}{r\sin(\theta)}\frac{\partial}{\partial\phi}\right)\\
&-r\sin(\theta)\sin(\phi)\left(\frac{\cos(\theta)\cos(\phi)}{r}\frac{\partial}{\partial\theta}-\frac{\sin(\phi)}{r\sin(\theta)}\frac{\partial}{\partial\phi}\right)\\
&=r\sin(\theta)\cos(\phi)\left(\frac{\cos(\phi)}{r\sin(\theta)}\frac{\partial}{\partial\phi}\right)-r\sin(\theta)\sin(\phi)\left(-\frac{\sin(\phi)}{r\sin(\theta)}\frac{\partial}{\partial\phi}\right)\\
&=\cos^2(\phi)\frac{\partial}{\partial\phi}+\sin^2(\phi)\frac{\partial}{\partial\phi}=\frac{\partial}{\partial\phi}
\end{aligned}\tag{13.1.72}$$

we get the following result:

$$L_z=-\mathrm{i}\hbar\partial_\phi\tag{13.1.73}$$

Finally, we have only little freedom for the movement of our rigid rotor (as it is very rigid ...) and we can write for the Schrödinger equation:

$$H_{\text{rot}} = \frac{L^2}{2J} \Rightarrow H_{\text{rot}}\Psi = E\Psi \tag{13.1.74}$$

where H_{rot} is for recall, seen as the functional linear operator, and the total energy E as its corresponding eigenvector.

Therefore, we can write that angular momentum operator is given by (we change the notation so to not confuse subsequently operator and eigenvalue according to the comments we made during the statement of the postulates of Wave Quantum Physics in the corresponding section):

$$\hat{L}_z = -\mathrm{i}\hbar\partial_\phi \tag{13.1.75}$$

Thus, the eigenfunctions $\Phi(\phi)$ of $\hat{L}_z$ are solutions of the equation to the eigenvalues and eigenfunctions:

$$\boxed{\hat{L}_z\Phi(\phi) = L_z\Phi(\phi)} \tag{13.1.76}$$

that is to say the differential equation:

$$\boxed{-\mathrm{i}\hbar\partial_\varphi\Phi(\varphi) = L_z\Phi(\phi)} \tag{13.1.77}$$

where L_z is obviously the eigenvalue of $\hat{L}_z$. A simple solution to this differential would be:

$$\Phi(\phi) = ce^{\frac{\mathrm{i}}{\hbar}L_z\phi} \tag{13.1.78}$$

with for uniformity condition, depending on the properties of complex number (see section Numbers page 243):

$$\Phi(\phi + 2\pi) = \Phi(\phi) \tag{13.1.79}$$

This mathematical condition imposes the obvious and remarkable following quantification:

$$L_z = \hbar m_l \quad \text{with} \quad m_l = 0, \pm 1, \pm 2, \ldots \tag{13.1.80}$$

where (recall) m_l is the magnetic quantum number.

Knowing that (see section Corpuscular Quantum Physics page 4668):

$$-l \leqslant m_l \leqslant +l \tag{13.1.81}$$

We can write:

$$L_z = \hbar l \tag{13.1.82}$$

Therefore, we falls back on the result(s) that we get in the section of Corpuscular Quantum Physics and Wave Quantum Physics (see page 4877):

$$b_z = \hbar l \Rightarrow \vec{b} = \hbar\vec{l} \tag{13.1.83}$$

Which is quite satisfactory, even remarkable and enjoyable (to not say it ...).

Thus, the measurement of one of the components of the angular momentum, provides always a multiple of $\hbar$ which appears therefore as a natural unit of angular momentum.

The common eigenfunctions (!!!) to the operators $\hat{L}^2$ and $\hat{L}_z$ are in a more general framework necessarily of the form (method of separation of variables):

$$R(r)\Phi(\phi)\Theta(\theta) \tag{13.1.84}$$

As the rotator is rigid, we have $R(r) = c^{te}$. This factor will eliminate itself in the equation of eigenvalues and eigenfunctions that we will determine further below. So we don't need to take it into account if we want! Finally, we can write thanks to previous developments:

$$\Phi(\varphi)\Theta(\theta) = e^{\mathrm{i}m_l\phi}\Theta(\theta) \tag{13.1.85}$$

Which brings us to the following equation of the eigenvalues and eigenfunctions:

$$\hat{L}^2\Phi(\phi)\Theta(\theta) = L^2 e^{\mathrm{i}m_l\phi}\Theta(\theta) \tag{13.1.86}$$

That is to say:

$$\boxed{-\hbar^2\left[\frac{1}{\sin(\theta)}\frac{\partial}{\partial\theta}\left(\sin(\theta)\frac{\partial}{\partial\theta}\right)+\frac{1}{\sin^2(\theta)}\frac{\partial^2}{\partial\phi^2}\right]\Phi(\phi)\Theta(\theta) = L^2 e^{\mathrm{i}m_l\varphi}\Theta(\theta)} \tag{13.1.87}$$

Therefore:

$$\begin{aligned}
&-\hbar^2\left(\frac{1}{\sin(\theta)}\frac{\partial}{\partial\theta}\left(\sin(\theta)\frac{\partial\left(e^{\mathrm{i}m_l\phi\Theta(\theta)}\right)}{\partial\theta}\right)+\frac{1}{\sin^2(\theta)}\frac{\partial^2\left(e^{\mathrm{i}m_l\phi}\Theta(\theta)\right)}{\partial\phi^2}\right) = L^2 e^{\mathrm{i}m_l\phi}\Theta(\theta)\\
&-\hbar^2\left(\begin{array}{c}\dfrac{1}{\sin(\theta)}\dfrac{\partial}{\partial\theta}\left(\sin(\theta)\Theta(\theta)\underbrace{\dfrac{\partial(e^{m_l\phi})}{\partial\theta}}_{=0}+\sin(\theta)e^{\mathrm{i}m_l\phi}\dfrac{\partial\Theta(\theta)}{\partial\theta}\right)\\ +\dfrac{1}{\sin^2(\theta)}e^{\mathrm{i}m_l\phi}\underbrace{\dfrac{\partial^2\Theta(\theta)}{\partial\phi^2}}_{=0}+\dfrac{1}{\sin^2(\theta)}\Theta(\theta)\dfrac{\partial^2\left(e^{\mathrm{i}m_l\phi}\right)}{\partial\phi^2}\end{array}\right) = L^2 e^{\mathrm{i}m_l\phi}\Theta(\theta)\\
&-\hbar^2\left(\frac{1}{\sin(\theta)}\frac{\partial}{\partial\theta}\sin(\theta)e^{\mathrm{i}m_l\phi}\frac{\partial\Theta(\theta)}{\partial\theta}+\frac{1}{\sin^2(\theta)}\Theta(\theta)\frac{\partial^2\left(e^{\mathrm{i}m_l\phi}\right)}{\partial\phi^2}\right) = L^2 e^{\mathrm{i}m_l\phi}\Theta(\theta)\\
&-\hbar^2\left(\cancel{e^{\mathrm{i}m_l\phi}}\frac{1}{\sin(\theta)}\frac{\partial}{\partial\theta}\sin(\theta)\frac{\partial\Theta(\theta)}{\partial\theta}+\frac{1}{\sin^2(\theta)}\Theta(\theta)(\mathrm{i}m_l)^2\cancel{e^{\mathrm{i}m_l\phi}}\right) = L^2\cancel{e^{\mathrm{i}m_l\phi}}\Theta(\theta)\\
&-\hbar^2\left(\frac{1}{\sin(\theta)}\frac{\partial}{\partial\theta}\sin(\theta)\frac{\partial\Theta(\theta)}{\partial\theta}+\frac{\Theta(\theta)m_l^2}{\sin^2(\theta)}\right) = L^2\Theta(\theta)
\end{aligned} \tag{13.1.88}$$

By putting:

$$\xi = \cos(\theta) \Rightarrow \frac{\partial\xi}{\partial\theta} = -\sin(\theta) \Rightarrow \partial\theta = -\frac{\partial\xi}{\sin(\theta)} \tag{13.1.89}$$

and therefore:

$$\cos^2(\theta)+\sin^2(\theta) = 1 \Rightarrow \xi^2+\sin^2(\theta) = 1 \Rightarrow = 1-\xi^2 \tag{13.1.90}$$

we get a "Fuchs" like differential equation given by:

$$\begin{aligned}
&-\hbar^2\left(\frac{1}{\sin(\theta)}\frac{\partial}{\partial\theta}\sin(\theta)\frac{\partial\Theta(\theta)}{\partial\theta}+\frac{\Theta(\theta)m_l^2}{\sin^2(\theta)}\right)=L^2\Theta(\theta)\Rightarrow\\
&-\hbar^2\left(\frac{1}{\cancel{\sin(\theta)}}\frac{\partial}{-\frac{\partial\xi}{\cancel{\sin(\theta)}}}\sin(\theta)\frac{\partial\Theta(\xi)}{-\frac{-\partial\xi}{\sin(\theta)}}+\frac{\Theta(\xi)m_l^2}{\sin^2(\theta)}\right)-L^2\Theta(\xi)=0\\
&-\hbar^2\left(\frac{\partial}{\partial\xi}\sin^2(\theta)\frac{\partial\Theta(\xi)}{\partial\xi}-\frac{\Theta(\xi)m_l^2}{\sin^2(\theta)}\right)-L^2\Theta(\xi)=0\\
&\left(\frac{\partial}{\partial\xi}(1-\xi^2)\frac{\partial\Theta(\xi)}{\partial\xi}-\frac{\Theta(\xi)m_l^2}{1-\xi^2}\right)-\frac{L^2}{\hbar}\Theta(\xi)=0\\
&\left(\frac{\partial}{\partial\xi}(1-\xi^2)\frac{\partial\Theta(\xi)}{\partial\xi}\right)-\left(\frac{L^2}{\hbar}-\frac{m_l^2}{1-\xi^2}\right)\Theta(\xi)=0\\
&\left(\frac{\partial(1-\xi^2)}{\partial\xi}+(1-\xi^2)\frac{\partial}{\partial\xi}\frac{\partial\Theta(\xi)}{\partial\xi}\right)-\left(\frac{L^2}{\hbar}-\frac{m_l^2}{1-\xi^2}\right)\Theta(\xi)=0\\
&\left(-2\xi\Theta'(\xi)+(1-\xi^2)\Theta''(\xi)\right)-\left(\frac{L^2}{\hbar}-\frac{m_l^2}{1-\xi^2}\right)\Theta(\xi)=0
\end{aligned}\tag{13.1.91}$$

Therefore we get finally:

$$\Theta''(\xi)-\frac{2\xi\Theta'(\xi)}{1-\xi^2}+\left(\frac{L^2}{\hbar^2}-\frac{m_l^2}{1-\xi^2}\right)\frac{\Theta(\xi)}{1-\xi^2}=0\tag{13.1.92}$$

Whose coefficients have poles (singularities) in $\xi=\pm1$. But, let us recall that we have:

$$L^2=\hbar^2l(l+1)\Rightarrow\frac{L^2}{\hbar^2}=l(l+1)\tag{13.1.93}$$

So that we often find the previous differential equation in the following form in the books after elementary algebra factorization of some terms:

$$\frac{\mathrm{d}}{\mathrm{d}\xi}\left[(1-\xi^2)\frac{\mathrm{d}\Theta(\xi)}{\mathrm{d}\xi}\right]+\left[l(l+1)-\frac{m_l^2}{1-\xi^2}\right]\Theta(\xi)=0\tag{13.1.94}$$

That is obviously also often written in the literature:

$$\frac{1}{\sin(\theta)}\left[\sin(\theta)\frac{\mathrm{d}\Theta}{\mathrm{d}\theta}\right]+\left[l(l+1)-\frac{m_l^2}{\sin^2(\theta)}\right]\Theta=0\tag{13.1.95}$$

A non-trivial solution being, assuming we know Fuchs differential equations (...), what it is customary to name the "associated Legendre polynomials" (although this is not strictly speaking a polynomial ....) because containing partly Legendre polynomials (see section Calculus page 1400):

$$\boxed{P_{m_l,l}=\Theta(\xi)=\frac{1}{2^l l!}(1-\xi^2)^{\frac{m_l}{2}}\frac{\mathrm{d}^{m_l+l}}{\mathrm{d}\xi^{m_l+l}}(1-\xi^2)^l}\tag{13.1.96}$$

that you can check by injecting this solution in prior-previous differential equation.

...Following the request of a reader here is an example of verification before continuing:

The $m_l=l=0$ is immediate. Then let us consider the case where $m_l=l=1$:

$$\frac{\mathrm{d}}{\mathrm{d}\xi}\left[(1-\xi^2)\frac{\mathrm{d}\Theta(\xi)}{\mathrm{d}\xi}\right]+\left[1(1+1)-\frac{1^2}{1-\xi^2}\right]\Theta(\xi)=0 \tag{13.1.97}$$

Thus:

$$\frac{\mathrm{d}}{\mathrm{d}\xi}\left[(1-\xi^2)\frac{\mathrm{d}\Theta(\xi)}{\mathrm{d}\xi}\right]+\left[\frac{1-2\xi^2}{1-\xi^2}\right]\Theta(\xi)=0 \tag{13.1.98}$$

And we inject in it the associated Lagrange polynomial:

$$\frac{\mathrm{d}}{\mathrm{d}\xi}\left[(1-\xi^2)\frac{\mathrm{d}\sqrt{1-\xi^2}}{\mathrm{d}\xi}\right]+\left[\frac{1-2\xi^2}{1-\xi^2}\right]\sqrt{1-\xi^2}=0 \tag{13.1.99}$$

Therefore:

$$\frac{\mathrm{d}}{\mathrm{d}\xi}\left[-(1-\xi^2)\frac{\xi}{\sqrt{1-\xi^2}}\right]+\left[\frac{1-2\xi^2}{1-\xi^2}\right]\sqrt{1-\xi^2}=0 \tag{13.1.100}$$

Which give after a small simplification:

$$\frac{\mathrm{d}}{\mathrm{d}\xi}\left[-\sqrt{1-\xi^2}\xi\right]+\left[\frac{1-2\xi^2}{1-\xi^2}\right]\sqrt{1-\xi^2}=0 \tag{13.1.101}$$

By derivating:

$$\frac{\xi^2}{\sqrt{1-\xi^2}}-\sqrt{1-\xi^2}+\left[\frac{1-2\xi^2}{1-\xi^2}\right]\sqrt{1-\xi^2}=0 \tag{13.1.102}$$

Let's focus on the left part to see what it is equal to by putting everything to a common denominator:

$$\frac{\xi^2(1-\xi^2)-\sqrt{1-\xi^2}\sqrt{1-\xi^2}(1-\xi^2)+(1-2\xi^2)\sqrt{1-xi^2}\sqrt{1-\xi^2}}{\sqrt{1-\xi^2}(1-\xi^2)} \tag{13.1.103}$$

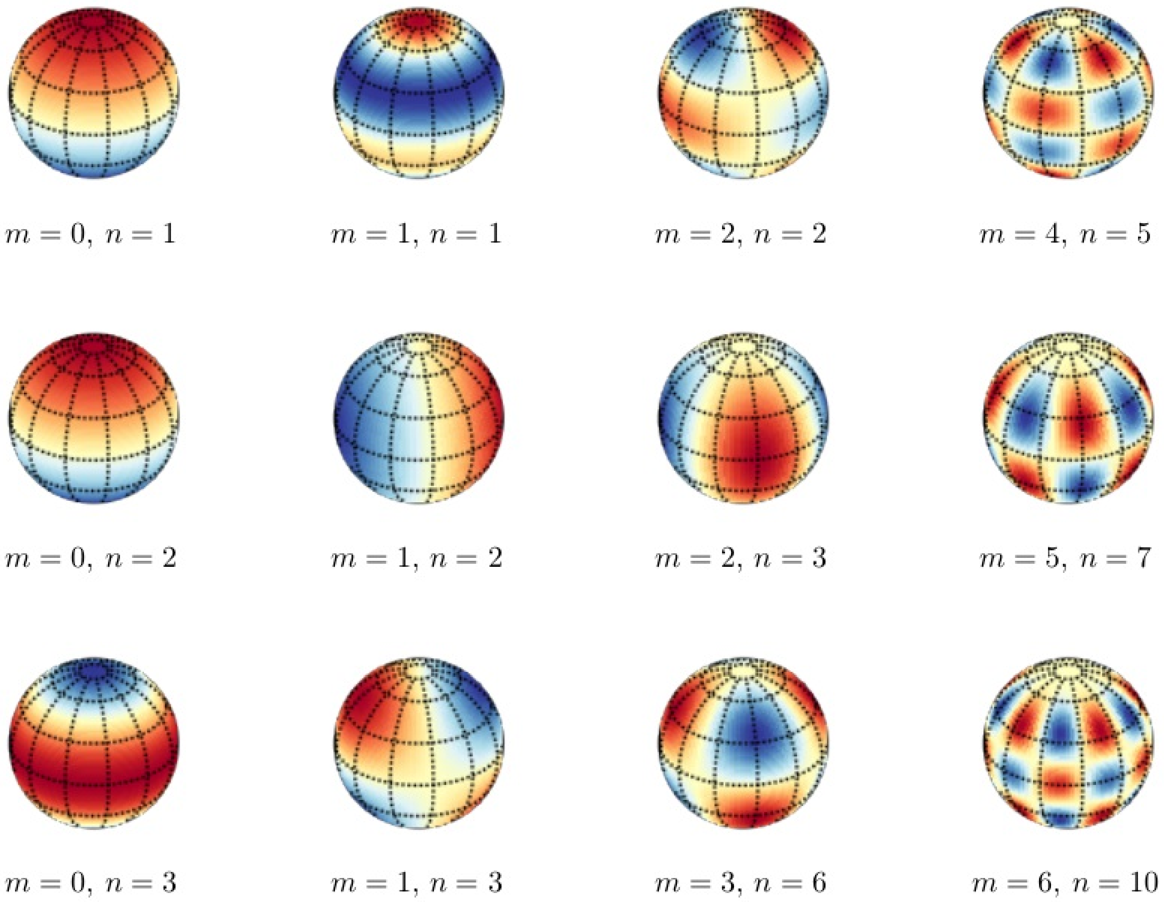
By simplifying the numerator, it should be zero. Let us see this by simplifying a first time:

$$\xi^2(1-\xi^2)-(1-\xi^2)^2+(1-2\xi^2)(1-\xi^2) \tag{13.1.104}$$

by distributing:

$$\xi^2-\xi^4-1+2\xi^2-\xi^4+1-\xi^2-2\xi^2+2\xi^4 \tag{13.1.105}$$

Which is indeed equal to zero!!!


$m = 0,\ n = 1$ $m = 1,\ n = 1$ $m = 2,\ n = 2$ $m = 4,\ n = 5$

$m = 0,\ n = 2$ $m = 1,\ n = 2$ $m = 2,\ n = 3$ $m = 5,\ n = 7$

$m = 0,\ n = 3$ $m = 1,\ n = 3$ $m = 3,\ n = 6$ $m = 6,\ n = 10$

Figure 13.5: Legendre spherical harmonics plot

With the corresponding MATLAB® 2013a script[1] (we give the script here as it is not given in the MATLAB® companion book) provided on Internet by Sanjay Sekaran (big thanks!):

```
theta = 0:pi/40:pi;                  % polar angle
phi = 0:pi/20:2*pi;                  % azimuth angle

[phi,theta] = meshgrid(phi,theta);   % define the grid

degree = 0;
order = 0;
amplitude = 0.5;
radius = 5;

Ymn = legendre(degree,cos(theta(:,1)));
Ymn = Ymn(order+1,:)';
yy = Ymn;

for kk = 2: size(theta,1)
    yy = [yy Ymn];
end

```

---

[1] It is a bit long but we really thinks it helps for a better understanding. Reader interested on how to do it with R can refer to the R companion book where the script is provided!

```
yy = yy.*cos(order*phi);

order = max(max(abs(yy)));
rho = radius + amplitude*yy/order;

r = radius.*sin(theta);     % convert to Cartesian coordinates
x = r.*cos(phi);
y = r.*sin(phi);
z = radius.*cos(theta);

subplot(5,5,1)
surf(x,y,z, rho);
title('$\ell=0, m=0$')

shading interp

axis equal off        % set axis equal and remove axis
view(0,30)           % set viewpoint

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
degree = 1;
order = 0;
amplitude = 0.5;
radius = 5;

Ymn = legendre(degree,cos(theta(:,1)));
Ymn = Ymn(order+1,:)';
yy = Ymn;

for kk = 2: size(theta,1)
    yy = [yy Ymn];
end

yy = yy.*cos(order*phi);

order = max(max(abs(yy)));
rho = radius + amplitude*yy/order;

r = radius.*sin(theta);     % convert to Cartesian coordinates
x = r.*cos(phi);
y = r.*sin(phi);
z = radius.*cos(theta);

subplot(5,5,6)
surf(x,y,z, rho);
title('$\ell=1, m=0$')
shading interp

axis equal off        % set axis equal and remove axis
view(0,30)           % set viewpoint

```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

degree = 1;
order = 1;
amplitude = 0.5;
radius = 5;

Ymn = legendre(degree,cos(theta(:,1)));
Ymn = Ymn(order+1,:)';
yy = Ymn;

for kk = 2: size(theta,1)
    yy = [yy Ymn];
end

yy = yy.*cos(order*phi);

order = max(max(abs(yy)));
rho = radius + amplitude*yy/order;

r = radius.*sin(theta);    % convert to Cartesian coordinates
x = r.*cos(phi);
y = r.*sin(phi);
z = radius.*cos(theta);

subplot(5,5,7)
surf(x,y,z, rho);
title('$\ell=1, m=\pm 1$')
shading interp

axis equal off      % set axis equal and remove axis
view(0,30)          % set viewpoint

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

degree = 2;
order = 0;
amplitude = 0.5;
radius = 5;

Ymn = legendre(degree,cos(theta(:,1)));
Ymn = Ymn(order+1,:)';
yy = Ymn;

for kk = 2: size(theta,1)
    yy = [yy Ymn];
end

yy = yy.*cos(order*phi);

order = max(max(abs(yy)));
```

```
rho = radius + amplitude*yy/order;

r = radius.*sin(theta);     % convert to Cartesian coordinates
x = r.*cos(phi);
y = r.*sin(phi);
z = radius.*cos(theta);

subplot(5,5,11)
surf(x,y,z, rho);
title('$\ell=2, m=0$')
shading interp

axis equal off       % set axis equal and remove axis
view(0,30)          % set viewpoint

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

degree = 2;
order = 1;
amplitude = 0.5;
radius = 5;

Ymn = legendre(degree,cos(theta(:,1)));
Ymn = Ymn(order+1,:)';
yy = Ymn;

for kk = 2: size(theta,1)
    yy = [yy Ymn];
end

yy = yy.*cos(order*phi);

order = max(max(abs(yy)));
rho = radius + amplitude*yy/order;

r = radius.*sin(theta);     % convert to Cartesian coordinates
x = r.*cos(phi);
y = r.*sin(phi);
z = radius.*cos(theta);

subplot(5,5,12)
surf(x,y,z, rho);
title('$\ell=2, m=\pm 1$')
shading interp

axis equal off       % set axis equal and remove axis
view(0,30)          % set viewpoint

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

degree = 2;
```

```
order = 2;
amplitude = 0.5;
radius = 5;

Ymn = legendre(degree,cos(theta(:,1)));
Ymn = Ymn(order+1,:)';
yy = Ymn;

for kk = 2: size(theta,1)
    yy = [yy Ymn];
end

yy = yy.*cos(order*phi);

order = max(max(abs(yy)));
rho = radius + amplitude*yy/order;

r = radius.*sin(theta);    % convert to Cartesian coordinates
x = r.*cos(phi);
y = r.*sin(phi);
z = radius.*cos(theta);

subplot(5,5,13)
surf(x,y,z, rho);
title('$\ell=2, m=\pm 2$')
shading interp

axis equal off       % set axis equal and remove axis
view(0,30)          % set viewpoint

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

degree = 3;
order = 0;
amplitude = 0.5;
radius = 5;

Ymn = legendre(degree,cos(theta(:,1)));
Ymn = Ymn(order+1,:)';
yy = Ymn;

for kk = 2: size(theta,1)
    yy = [yy Ymn];
end

yy = yy.*cos(order*phi);

order = max(max(abs(yy)));
rho = radius + amplitude*yy/order;

r = radius.*sin(theta);    % convert to Cartesian coordinates
```

```
x = r.*cos(phi);
y = r.*sin(phi);
z = radius.*cos(theta);

subplot(5,5,16)
surf(x,y,z, rho);
title('$\ell=3, m=0$')
shading interp

axis equal off      % set axis equal and remove axis
view(0,30)          % set viewpoint

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

degree = 3;
order = 1;
amplitude = 0.5;
radius = 5;

Ymn = legendre(degree,cos(theta(:,1)));
Ymn = Ymn(order+1,:)';
yy = Ymn;

for kk = 2: size(theta,1)
    yy = [yy Ymn];
end

yy = yy.*cos(order*phi);

order = max(max(abs(yy)));
rho = radius + amplitude*yy/order;

r = radius.*sin(theta);    % convert to Cartesian coordinates
x = r.*cos(phi);
y = r.*sin(phi);
z = radius.*cos(theta);

subplot(5,5,17)
surf(x,y,z, rho);
title('$\ell=3, m=\pm 1$')
shading interp

axis equal off      % set axis equal and remove axis
view(0,30)          % set viewpoint

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

degree = 3;
order = 2;
amplitude = 0.5;
radius = 5;
```

```

Ymn = legendre(degree,cos(theta(:,1)));
Ymn = Ymn(order+1,:)';
yy = Ymn;

for kk = 2: size(theta,1)
    yy = [yy Ymn];
end

yy = yy.*cos(order*phi);

order = max(max(abs(yy)));
rho = radius + amplitude*yy/order;

r = radius.*sin(theta);     % convert to Cartesian coordinates
x = r.*cos(phi);
y = r.*sin(phi);
z = radius.*cos(theta);

subplot(5,5,18)
surf(x,y,z, rho);
title('$\ell=3, m=\pm 2$')
shading interp

axis equal off       % set axis equal and remove axis
view(0,30)          % set viewpoint

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

degree = 3;
order = 3;
amplitude = 0.5;
radius = 5;

Ymn = legendre(degree,cos(theta(:,1)));
Ymn = Ymn(order+1,:)';
yy = Ymn;

for kk = 2: size(theta,1)
    yy = [yy Ymn];
end

yy = yy.*cos(order*phi);

order = max(max(abs(yy)));
rho = radius + amplitude*yy/order;

r = radius.*sin(theta);     % convert to Cartesian coordinates
x = r.*cos(phi);
y = r.*sin(phi);
z = radius.*cos(theta);
```

```

subplot(5,5,19)
surf(x,y,z, rho);
title('$\ell=3, m=\pm 3$')
shading interp

axis equal off       % set axis equal and remove axis
view(0,30)          % set viewpoint

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

degree = 4;
order = 0;
amplitude = 0.5;
radius = 5;

Ymn = legendre(degree,cos(theta(:,1)));
Ymn = Ymn(order+1,:)';
yy = Ymn;

for kk = 2: size(theta,1)
    yy = [yy Ymn];
end

yy = yy.*cos(order*phi);

order = max(max(abs(yy)));
rho = radius + amplitude*yy/order;

r = radius.*sin(theta);     % convert to Cartesian coordinates
x = r.*cos(phi);
y = r.*sin(phi);
z = radius.*cos(theta);

subplot(5,5,21)
surf(x,y,z, rho);
title('$\ell=4, m=0$')
shading interp

axis equal off       % set axis equal and remove axis
view(0,30)          % set viewpoint

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

degree = 4;
order = 1;
amplitude = 0.5;
radius = 5;

Ymn = legendre(degree,cos(theta(:,1)));
Ymn = Ymn(order+1,:)';
```

```
yy = Ymn;

for kk = 2: size(theta,1)
    yy = [yy Ymn];
end

yy = yy.*cos(order*phi);

order = max(max(abs(yy)));
rho = radius + amplitude*yy/order;

r = radius.*sin(theta);     % convert to Cartesian coordinates
x = r.*cos(phi);
y = r.*sin(phi);
z = radius.*cos(theta);

subplot(5,5,22)
surf(x,y,z, rho);
title('$\ell=4, m=\pm 1$')
shading interp

axis equal off       % set axis equal and remove axis
view(0,30)           % set viewpoint

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

degree = 4;
order = 2;
amplitude = 0.5;
radius = 5;

Ymn = legendre(degree,cos(theta(:,1)));
Ymn = Ymn(order+1,:)';
yy = Ymn;

for kk = 2: size(theta,1)
    yy = [yy Ymn];
end

yy = yy.*cos(order*phi);

order = max(max(abs(yy)));
rho = radius + amplitude*yy/order;

r = radius.*sin(theta);     % convert to Cartesian coordinates
x = r.*cos(phi);
y = r.*sin(phi);
z = radius.*cos(theta);

subplot(5,5,23)
surf(x,y,z, rho);
```

```
title('$\ell=4, m=\pm 2$')
shading interp

axis equal off       % set axis equal and remove axis
view(0,30)          % set viewpoint

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

degree = 4;
order = 3;
amplitude = 0.5;
radius = 5;

Ymn = legendre(degree,cos(theta(:,1)));
Ymn = Ymn(order+1,:)';
yy = Ymn;

for kk = 2: size(theta,1)
    yy = [yy Ymn];
end

yy = yy.*cos(order*phi);

order = max(max(abs(yy)));
rho = radius + amplitude*yy/order;

r = radius.*sin(theta);     % convert to Cartesian coordinates
x = r.*cos(phi);
y = r.*sin(phi);
z = radius.*cos(theta);

subplot(5,5,24)
surf(x,y,z, rho);
title('$\ell=4, m=\pm 3$')
shading interp

axis equal off       % set axis equal and remove axis
view(0,30)          % set viewpoint

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

degree = 4;
order = 4;
amplitude = 0.5;
radius = 5;

Ymn = legendre(degree,cos(theta(:,1)));
Ymn = Ymn(order+1,:)';
yy = Ymn;

for kk = 2: size(theta,1)
```

```
    yy = [yy Ymn];
end

yy = yy.*cos(order*phi);

order = max(max(abs(yy)));
rho = radius + amplitude*yy/order;

r = radius.*sin(theta);     % convert to Cartesian coordinates
x = r.*cos(phi);
y = r.*sin(phi);
z = radius.*cos(theta);

subplot(5,5,25)
surf(x,y,z, rho);
title('$\ell=4, m=\pm 4$')
shading interp

axis equal off       % set axis equal and remove axis
view(0,30)          % set viewpoint

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

map = makeColorMap([0.2 0.2 0.6],[1.0 0.99 0.72],[0.8 0.25 0.33],80
    );
colormap(map);
cd(Figures)
```

So finally, we have common eigen functions (because remember that the Legendre polynomials are orthogonal to each other) that will be:

$$Y_{m_l,l}(\theta,\phi)=N_{m_l,l}P_{m_l,l}(\xi)e^{\mathrm{i}m_l\phi} \tag{13.1.106}$$

Remark

There is no need to make complicated calculations to determine the normalization factor of the exponential, as in the context of an integration over all space, the three factors of $Y_{m_l,l}(\theta,\phi)$ are independent of each other. Thus the integral is the product of the integrals (see section Differential and Integral Calculus page 1497).

Finally, we must find $N_{m_l,l}$ such that:

$$\boxed{\int|Y_{m_l,l}(\theta,\phi)|^2\mathrm{d}\Omega=\int Y_{m_l,l}(\theta,\phi)\bar{Y}_{m_l,l}(\theta,\phi)\mathrm{d}\Omega=1} \tag{13.1.107}$$

and we will see (what we will prove just below) that:

$$\boxed{N_{m_l,l}=\sqrt{\frac{(2l+1)(l-m_l)!}{2(l+m_l)!}}} \tag{13.1.108}$$

In summary, we write (we should rather say "we will write"...):

$$\boxed{Y_{m_l,l}(\theta,\phi)=\frac{1}{\sqrt{2\pi}}\sqrt{\frac{(2l+1)(l-m_l)!}{2(l+m_l)!}}\frac{1}{2^l l!}(1-\xi^2)^{\frac{m_l}{2}}\frac{\mathrm{d}^{m_l+l}}{\mathrm{d}\xi^{m_l+l}}(1-\xi^2)^l e^{\mathrm{i}m_l\phi}} \tag{13.1.109}$$

where we have omitted the factor $(-1)^l$ since in any case this term in the module of this function this term multiplies himself and then gives $(-1)^{2l}=1$.

Let us check now the framed previous boxed relation (warning this is a bit long and it is advisable to read it several times):

We consider the functions defined by:

$$Y_{m_l,l}(\theta,\phi)=\frac{N_{l,m}}{\sqrt{2\pi}}P_{m_l,l}(\cos\theta)e^{\mathrm{i}m\phi} \tag{13.1.110}$$

where:

$$P_{m_l,l}(x)=\frac{1}{2^l l!}(1-x^2)^{m_l/2}\frac{\mathrm{d}^{m_l+l}}{\mathrm{d}x^{m_l+l}}(1-x^2)^l \tag{13.1.111}$$

with:

$$l\in\mathbb{N},-l\leq m_l\leq l \quad \text{and} \quad \theta\in[0,\pi],\phi\in[0,2\pi[\tag{13.1.112}$$

The aim will therefore be to prove that these functions are orthogonal first and then find the constants $N_{m_l,l}$ such that $||Y_{m_l,l}||$. In short we will have to roll up the sleeves... of our brain...

First, let us prove for future needs that:

$$\frac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}\left((1-x^2)^{m_l}\frac{\mathrm{d}^{m_l+l}}{\mathrm{d}x^{m_l+l}}(1-x^2)^l\right)=(-1)^{m_l}\frac{(l+m_l)!}{(l-m_l)!}\frac{\mathrm{d}^l}{\mathrm{d}x^l}(1-x^2)^l \tag{13.1.113}$$

Proof 13.0.1. If and only if $l=m_l=0$ the equality is obvious. Let us suppose that $l\geq 1$ (thus the general case outside the obvious previous case) and given P a real polynomial of degree $\leq l-1$.

Let us put:

$$Z(x):=(1-x^2)^{m_l}\frac{\mathrm{d}^{m_l+l}}{\mathrm{d}x^{m_l+l}}(1-x^2)^l \tag{13.1.114}$$

Let us prove (functional dot product):

$$\left\langle \frac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}Z\middle|P\right\rangle=0 \tag{13.1.115}$$

in $\mathcal{C}(x\in[-1,1],\mathbb{R})$.

Indeed, let us recall that we made the change of variable:

$$x=\cos(\theta),-\pi\leq\theta\leq+\pi\Rightarrow -1\leq x\leq +1 \tag{13.1.116}$$

Integrating by parts, we get:

$$\left\langle \frac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}Z\middle|P\right\rangle = \int\limits_{-1}^{1}\frac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}Z(x)P(x)\mathrm{d}x = \frac{\mathrm{d}^{m_l-1}}{\mathrm{d}x^{m_l-1}}Z(x)P(x)\mathrm{d}x\bigg|_{-1}^{1} - \int\limits_{-1}^{1}\frac{\mathrm{d}^{m_l-1}}{\mathrm{d}x^{m_l-1}}Z(x)\frac{\mathrm{d}}{\mathrm{d}x}P(x)\mathrm{d}x \tag{13.1.117}$$

let us notice that for any $0 \leq j \leq m_l - 1$, $\dfrac{\mathrm{d}^j}{\mathrm{d}x^j}Z(x)$ is equal to zero in $x = \pm 1$ that is to say:

$$\frac{\mathrm{d}^j}{\mathrm{d}x^j}Z(-1) = \frac{\mathrm{d}^j}{\mathrm{d}x^j}Z(1) = 0 \tag{13.1.118}$$

Therefore (by extension), the above relation simplifies to:

$$\left\langle \frac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}Z\middle|P\right\rangle = -\int\limits_{-1}^{1}\frac{\mathrm{d}^{m_l-1}}{\mathrm{d}x^{m_l-1}}Z(x)\frac{\mathrm{d}}{\mathrm{d}x}P(x)\mathrm{d}x \tag{13.1.119}$$

After m_l integration by parts equation, we get:

$$\left\langle \frac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}Z\middle|P\right\rangle = (-1)^{m_l}\int\limits_{-1}^{1}Z(x)\frac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}P(x)\mathrm{d}x \tag{13.1.120}$$

If $\deg(P) < m_l$ then the above expression shows that trivially:

$$\left\langle \frac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}Z\middle|P\right\rangle = 0 \tag{13.1.121}$$

If $\deg(P) \geq m_l$ then by putting:

$$h(x) := (1-x^2)^{m_l}\frac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}P(x) \tag{13.1.122}$$

We get:

$$\begin{aligned}\left\langle \frac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}Z\middle|P\right\rangle &= (-1)^{m_l}\int\limits_{-1}^{1}(1-x^2)^{m_l}\frac{\mathrm{d}^{m_l+l}}{\mathrm{d}x^{m_l+l}}(1-x^2)^l\frac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}P(x)\mathrm{d}x\\ &= (-1)^{m_l}\int\limits_{-1}^{1}\frac{\mathrm{d}^{m_l+l}}{\mathrm{d}x^{m_l+l}}(1-x^2)^l h(x)\mathrm{d}x\end{aligned} \tag{13.1.123}$$

let us notice once again that $\dfrac{\mathrm{d}^j}{\mathrm{d}x^j}h(x)$ vanishes in $x = \pm 1$ for any $j \leq m_l - 1$, that is say:

$$\frac{\mathrm{d}^j}{\mathrm{d}x^j}h(-1) = \frac{\mathrm{d}^j}{\mathrm{d}x^j}h(1) = 0 \tag{13.1.124}$$

By integrating by parts $m_l$the previous expression, we find:

$$\left\langle \frac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}Z\middle|P\right\rangle = \int\limits_{-1}^{1}\frac{\mathrm{d}^l}{\mathrm{d}x^l}(1-x^2)^l\frac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}h(x)\mathrm{d}x \tag{13.1.125}$$

but h is a polynomial of degree $m_l+\deg(P)$.

Indeed, the first factor is of degree $2m$ and the m_lth derivative of $P(x)$" is of degree $P-m_l$, therefore:

$$2m_l+P-m_l=m_l+P=m_l+\deg(P) \tag{13.1.126}$$

So $\dfrac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}h(x)$ is a polynomial of degree $\deg(P)\leq l-1$ and knowing that $\dfrac{\mathrm{d}^l}{\mathrm{d}x^l}(1-x^2)^l$ is to a given constant equal to the l-th Legendre polynomial (see section Calculus page 1400) we then have:

$$\left\langle \frac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}Z \middle| P\right\rangle=\int\limits_{-1}^{1}\frac{\mathrm{d}^l}{\mathrm{d}x^l}(1-x^2)^l\frac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}h(x)\mathrm{d}x=0 \tag{13.1.127}$$

So we have just proved that $\dfrac{\mathrm{d}m_l}{\mathrm{d}x_l^m}$ is orthogonal to any polynomial of degree $\leq l-1$.

■ Q.E.D.

$\dfrac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}Z$ is a polynomial of degree l (its just enough to check for some values) so therefore let us search if there is a constant $c^{te}\in\mathbb{R}$ such that:

$$\frac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}Z=c^{te}\frac{\mathrm{d}^l}{\mathrm{d}x^l}(1-x^2)^l \tag{13.1.128}$$

with for recall:

$$Z(x):=(1-x^2)^{m_l}\frac{\mathrm{d}^{m_l+l}}{\mathrm{d}x^{m_l+l}}(1-x^2)^l \tag{13.1.129}$$

We can determine the constant c^{te} by comparing the dominant coefficients of the polynomials:

$$\frac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}Z \quad \text{and} \quad \frac{\mathrm{d}^l}{\mathrm{d}x^l}(1-x^2)^l \tag{13.1.130}$$

The dominant coefficient of $\dfrac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}Z$ is:

$$(-1)^{l+m_l}\frac{(2l)!}{(l-m_l)!}\frac{(l+m_l)!}{l!} \tag{13.1.131}$$

and the dominant coefficient of $\dfrac{\mathrm{d}^l}{\mathrm{d}x^l}(1-x^2)^l$ is:

$$(-1)^l\frac{(2l)!}{l!} \tag{13.1.132}$$

Therefore:

$$(-1)^{l+m_l}\frac{(2l)!}{(l-m_l)!}\frac{(l+m_l)!}{l!}=c^{te}(-1)^l\frac{(2l)!}{l!} \tag{13.1.133}$$

That is to say:

$$c^{te}=(-1)^{m_l}\frac{(l+m_l)!}{(l-m_l)!} \tag{13.1.134}$$

So we would have for $l \geq m_l \geq 0$ (we integrate parts we integrate as many times as necessary to the left and right - necessarily - to achieve this result):

$$Z(x)=(1-x^2)^{m_l}\frac{\mathrm{d}^{m_l+l}}{\mathrm{d}x^{m_l+l}}(1-x^2)^l=(-1)^{m_l}\frac{(l+m_l)!}{(l-m_l)!}\frac{\mathrm{d}^{l-m_l}}{\mathrm{d}x^{l-m_l}}(1-x^2)^l \tag{13.1.135}$$

Now let us establish a remarkable relation that should perhaps exist between $P_{-m_l,l}$ and $P_{m_l,l}$ (and which will be useful to us later). Let us assume for this $0 \leq m_l \leq l$ and remember that at the base:

$$P_{m_l,l}(x)=\frac{1}{2^l l!}(1-x^2)^{m_l/2}\frac{\mathrm{d}^{m_l+l}}{\mathrm{d}x^{m_l+l}}(1-x^2)^l \tag{13.1.136}$$

So that brings us to write (nothing special):

$$P_{-m_l,l}(x)=\frac{1}{2^l l!}(1-x^2)^{-m_l/2}\underbrace{\left[\frac{\mathrm{d}^{l-m_l}}{\mathrm{d}x^{l-m_l}}(1-x^2)^l\right]}_{\text{replaced by the result obtained above}} \tag{13.1.137}$$

By the previous results ($(-1)^{m_l}=(-1)^{-m_l}$):

$$\begin{gathered}(1-x^2)^{m_l}\frac{\mathrm{d}^{m_l+l}}{\mathrm{d}x^{m_l+l}}(1-x^2)^l=(-1)^{m_l}\frac{(l+m_l)!}{(l-m_l)!}\frac{\mathrm{d}^{l-m_l}}{\mathrm{d}x^{l-m_l}}(1-x^2)^l\\ \Rightarrow\frac{\mathrm{d}^{l-m_l}}{\mathrm{d}x^{l-m_l}}(1-x^2)^l=(-1)^{m_l}\frac{(l-m_l)!}{(l+m_l)!}(1-x^2)^{m_l}\frac{\mathrm{d}^{m_l+l}}{\mathrm{d}x^{m_l+l}}(1-x^2)^l\end{gathered} \tag{13.1.138}$$

this leads us to write:

$$\begin{aligned}P_{-m_l,l}(x)&=\frac{(-1)^{m_l}}{2^l l!}\frac{(l-m_l)!}{(l+m_l)!}(1-x^2)^{-m_l/2}(1-x^2)^{m_l}\frac{\mathrm{d}^{l+m_l}}{\mathrm{d}x^{l+m_l}}(1-x^2)^l\\ &=\frac{(-1)^{m_l}}{2^l l!}\frac{(l-m_l)!}{(l+m_l)!}(1-x^2)^{m_l/2}\frac{\mathrm{d}^{l+m_l}}{\mathrm{d}x^{l+m_l}}(1-x^2)^l=\\ &=(-1)^{m_l}\frac{(l-m_l)!}{(l+m_l)!}P_{m_l,l}(x)\end{aligned} \tag{13.1.139}$$

Therefore we get:

$$P_{-m_l,l}(x)=(-1)^{m_l}\frac{(l-m_l)!}{(l+m_l)!}P_{m_l,l} \tag{13.1.140}$$

First, let us prove that:

$$\int\limits_{-1}^{1}P_l(x)x^l\mathrm{d}x=\frac{(l!)^2}{(2l+1)!}2^{l+1} \tag{13.1.141}$$

where P_l is the n-th Legendre polynomial (hence the origin of the name of "associated Legendre polynomial" ...).

Proof 13.0.2. First, we have proved that the Legendre polynomials satisfy the following recurrence relation (see section Calculus page 1400):

$$(n+1)P_{n+1}(x)=(2n+1)xP_n(x)-nP_{n-1}(x) \tag{13.1.142}$$

for $n\geq 1$. Multiplying the above equation by x^{n-1} and integrating, we get:

$$(n+1)\int\limits_{-1}^{1}P_{n+1}(x)x^{n-1}\mathrm{d}x=(2n+1)\int\limits_{-1}^{1}P_n(x)x^n\mathrm{d}x-n\int\limits_{-1}^{1}P_{n-1}(x)x^{n-1}\mathrm{d}x \tag{13.1.143}$$

But:

$$\int\limits_{-1}^{1}P_{n+1}(x)x^{n-1}\mathrm{d}x=0 \tag{13.1.144}$$

Let us recall that the P_{n+1} polynomials form an orthogonal basis of which the polynomials that generate it are of increasing degree from 0 to n, so a lower order polynomial - expressed in a sub vector space - will always be perpendicular to the vectors (polynomials) generating the higher dimensions. So if we take the example of $\mathbb{R}^3$ generated by the basis $(\vec{e}_1,\vec{e}_2,\vec{e}_3)$, then a vector $\vec{v}$ expressed by the linear combination of $(\vec{e}_1,\vec{e}_2)$ will always be perpendicular to $\vec{e}_3$ and therefore a zero scalar product with it.

And therefore it follows:

$$\int\limits_{-1}^{1}P_n(x)x^n\mathrm{d}x=\frac{n}{2n+1}\int\limits_{-1}^{1}P_{n-1}(x)x^{n-1}\mathrm{d}x \tag{13.1.145}$$

Let us put:

$$S_n=\int\limits_{-1}^{1}P_n(x)x^n\mathrm{d}x \tag{13.1.146}$$

The previous expression becomes (remember $P_0(x)=1$):

$$S_n=\frac{n}{2n+1}S_{n-1}\quad\text{with}\quad S_0=S_n=\int\limits_{-1}^{1}P_0(x)x^0\mathrm{d}x=\int\limits_{-1}^{1}1\mathrm{d}x=2 \tag{13.1.147}$$

Thus by induction:

$$S_n=\frac{n}{2n+1}S_{n-1}=2\cdot\frac{1}{3}\cdot\frac{2}{5}\cdot\frac{3}{7}\cdot\ldots\cdot\frac{n}{2n+1}=\frac{2}{1}\cdot\frac{1}{3}\cdot\frac{2}{5}\cdot\frac{3}{7}\cdot\ldots\cdot\frac{n}{2n+1} \tag{13.1.148}$$

Furthermore as:

$$1\cdot 2\cdot 3\cdot 4\cdot\ldots\cdot n=n! \tag{13.1.149}$$

We then have for the prior-previous relation the denominator which can obviously be rewritten:

$$1\cdot 3\cdot 5\cdot 7\cdot\ldots\cdot(2n+1)=\frac{(2n+1)!}{2\cdot 4\cdot 6\cdot\ldots\cdot(2n)} \tag{13.1.150}$$

Then we have:

$$2\cdot 4\cdot 6\cdot \ldots \cdot (2n)=\prod_{j=1}^{n}2j=2^n n! \tag{13.1.151}$$

So in the end we can simplify the denominator as follows:

$$1\cdot 3\cdot 5\cdot \ldots \cdot (2n+1)=\frac{(2n+1)!}{2\cdot 4\cdot 6\cdot \ldots \cdot (2n)}=\frac{(2n+1)!}{2^n n!} \tag{13.1.152}$$

and:

$$S_n=2\cdot\frac{1}{3}\cdot\frac{2}{5}\cdot\frac{3}{7}\cdot \ldots \cdot \frac{n}{2n+1}=2^{n+1}\frac{(n!)^2}{(2n+1)!} \tag{13.1.153}$$

So we have well proved that (just in case ... you would not follow anymore the initial target ...) that:

$$\int_{-1}^{1}P_l(x)x^l\mathrm{d}x=\frac{(l!)^2}{(2l+1)!}2^{l+1} \tag{13.1.154}$$

■ Q.E.D.

Let us attack us finally to what interests us. That is to say, prove that:

$$\langle Y_{m_l,l}|Y_{j,k}\rangle=\begin{cases}\dfrac{2N_{m_l,l}^2}{2l+1}\dfrac{(l+m_l)!}{(l-m_l)!} & \text{if}\quad m_l=j \text{ and } l=k\\ 0 & \text{otherwise}\end{cases} \tag{13.1.155}$$

Proof 13.0.3. If $m_l\neq j$ then:

$$\langle Y_{m_l,l}|Y_{j,k}\rangle=\frac{N_{m_l,l}N_{j,k}}{2\pi}\int_0^{\pi}\int_0^{2\pi}P_{m,l}(\cos\theta)P_{j,k}(\cos\theta)\mathrm{e}^{\mathrm{i}\phi(m_l-j)}\mathrm{d}\Omega=0 \tag{13.1.156}$$

where:

$$\mathrm{d}\Omega=\sin(\theta)\mathrm{d}\theta\mathrm{d}\phi \tag{13.1.157}$$

Remark

Let us recall that the Jacobian in spherical coordinates is $r^2\sin(\theta)$ (see section Differential and Integral Calculus page 1528) and as the integrated function above is not dependent on r, we have take out the term $r^2\mathrm{d}r$ of this integral (by cons we will meet again the same term in the function $R(r)$ present in the Schrödinger equation).

And with:

$$\int_0^{2\pi}e^{\mathrm{i}\phi(m_l-j)}\mathrm{d}\phi=0 \tag{13.1.158}$$

If $l>k$ and $m_l=j$ then the dot product of:

$$\langle Y_{m_l,l}|Y_{m_l,k}\rangle=N_{m_l,l}N_{m_l,k}\int\limits_0^\pi P_{m_l,l}(\cos\theta)P_{m_l,k}(\cos\theta)\sin(\theta)\mathrm{d}\theta \tag{13.1.159}$$

is simplified to:

$$\langle Y_{m_l,l}|Y_{j,k}\rangle=N_{m_l,l}N_{j,k}\int\limits_0^\pi P_{m_l,l}(\cos\theta)P_{j,k}(\cos\theta)\sin(\theta)\mathrm{d}\theta \tag{13.1.160}$$

By doing the change of variable $x=\cos(\theta)$ we get:

$$\langle Y_{m_l,l}|Y_{j,k}\rangle=N_{m_l,l}N_{j,k}\int\limits_{-1}^1 P_{m,l}(x)P_{j,k}(x)\mathrm{d}x \tag{13.1.161}$$

Let us suppose that $m_l\geq 0$:

$$P_{m,l}(x)=\frac{1}{2^l l!}(1-x^2)^{m_l/2}\frac{\mathrm{d}^{m_l+l}}{\mathrm{d}x^{m_l+l}}(1-x^2)^l=(-1)^l(1-x^2)^{m_l/2}\frac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}P_l(x) \tag{13.1.162}$$

where $P_l(x)$ is the n-th Legendre polynomial. Thus the expression of the dot product becomes:

$$\langle Y_{m_l,l}|Y_{j,k}\rangle=(-1)^{l+k}N_{m_l,l}N_{j,k}\int\limits_{-1}^1(1-x^2)^{m_l}\frac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}P_l(x)\frac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}P_k(x)\mathrm{d}x \tag{13.1.163}$$

If we put:

$$h(x):=(1-x^2)^{m_l}\frac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}P_k(x) \tag{13.1.164}$$

then the relation becomes:

$$\langle Y_{m_l,l}|Y_{j,k}\rangle=(-1)^{l+k}N_{m_l,l}N_{j,k}\int\limits_{-1}^1\frac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}P_l(x)h(x)\mathrm{d}x \tag{13.1.165}$$

Integrating by parts m times the expression above we get:

$$\langle Y_{m_l,l}|Y_{j,k}\rangle=(-1)^{l+k+m_l}N_{m_l,l}N_{j,k}\int\limits_{-1}^1 P_l(x)\frac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}h(x)\mathrm{d}x \tag{13.1.166}$$

But $\dfrac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}h(x)$ is a polynomial of degree k. Knowing that $l>k$, the latter integral is zero for the same reasons as those mentioned above. Therefore:

$$\langle Y_{m_l,l}|Y_{j,k}\rangle=0 \tag{13.1.167}$$

If $m_l<$ then we have proved that:

$$P_{-m_l,l}(x)=(-1)^{m_l}\frac{(l-m_l)!}{(l+m_l)!}P_{m_l,l} \tag{13.1.168}$$

and therefore:

$$\langle Y_{m_l,l}|Y_{j,k}\rangle = C\int\limits_{-1}^{1} P_{-m_l,l}(x)P_{-j,k}(x)\mathrm{d}x = 0 \tag{13.1.169}$$

as $-m_l \geq 0$.

It only remains to us to treat the case $m_l = j, l = k$. Let us suppose again that $m_l \geq 0$. So as before we have:

$$P_{m_l,l}(x) = (-1)^l(1-x^2)^{m_l/2}\frac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}P_l(x) \tag{13.1.170}$$

and:

$$\langle Y_{m_l,l}|Y_{m_l,l}\rangle = N_{m_l,l}^2\int\limits_{-1}^{1} P_{m_l,l}(x)^2\mathrm{d}x = N_{m_l,l}^2\int\limits_{-1}^{1}(1-x^2)^{m_l}\frac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}P_l(x)\frac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}P_l(x)\mathrm{d}x \tag{13.1.171}$$

Let us put:

$$h(x) := (1-x^2)^{m_l}\frac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}P_l(x) \tag{13.1.172}$$

The relation then becomes:

$$\langle Y_{m_l,l}|Y_{m_l,l}\rangle = N_{m_l,l}^2\int\limits_{-1}^{1}\frac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}P_l(x)h(x)\mathrm{d}x \tag{13.1.173}$$

By integrating m times by parts, we find:

$$\langle Y_{m_l,l}|Y_{m_l,l}\rangle = N_{m_l,l}^2(-1)^{m_l}\int\limits_{-1}^{1}P_l(x)\frac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}h(x)\mathrm{d}x \tag{13.1.174}$$

$\dfrac{\mathrm{d}^{m_l}}{\mathrm{d}x^{m_l}}h(x)$ is a polynomial of degree l which dominant coefficient is equal to:

$$(-1)^{m_l}\frac{(2l)!}{(l!)^2 2^l}\frac{(l+m_l)!}{(l-m_l)!} \tag{13.1.175}$$

P_l being orthogonal to any polynomial of degree strictly less that l, the expression can be written:

$$\langle Y_{m_l,l}|Y_{m_l,l}\rangle = N_{m_l,l}^2\frac{(2l)!}{(l!)^2 2^l}\frac{(l+m_l)!}{(l-m_l)!}\int\limits_{-1}^{1}P_l(x)x^l\mathrm{d}x \tag{13.1.176}$$

But, we have proved that:

$$\int\limits_{-1}^{1}P_l(x)x^l\mathrm{d}x = \frac{(l!)^2}{(2l+1)!}2^{l+1} \tag{13.1.177}$$

therefore:

$$\langle Y_{m_l,l}|Y_{m_l,l}\rangle = N_{m_l,l}^2\frac{(2l)!}{(l!)^2 2^l}\frac{(l+m_l)!}{(l-m_l)!}\frac{(l!)^2}{(2l+1)!}2^{l+1} = \frac{2N_{m_l,l}^2}{2l+1}\frac{(l+m_l)!}{(l-m_l)!} \tag{13.1.178}$$

If $m_l \leq 0$ we know that we get the result.

■ Q.E.D.

Finally this result gives us also the normalization condition:

$$\langle Y_{m_l,l}|Y_{m_l,l}\rangle=\frac{2N^2_{m_l,l}}{2l+1}\frac{(l+m_l)!}{(l-m_l)!}=1\quad\Rightarrow\quad N_{m_l,l}=\sqrt{\frac{(2l+1)(l-m_l)!}{2(l+m_l)!}}\tag{13.1.179}$$

And so finally:

$$Y_{m_l,l}(\theta,\phi)=\frac{N_{m_l,l}}{\sqrt{2\pi}}P_{m_l,l}(\cos(\theta))e^{\mathrm{i}m\phi}\tag{13.1.180}$$

is indeed an orthonormal family. Either explicitly (we reintroduce the factor $(-1)^l$):

$$\boxed{Y_{m_l,\phi}=\sqrt{\frac{(2l+1)(l-m_l)!}{4\pi(l+m_l)!}}\frac{(-1)^l}{2^l\cdot l!}(1-\cos^2(\theta))^{m_l/2}\frac{\mathrm{d}^{m_l+l}}{\mathrm{d}\cos(\theta)^{m_l+l}}\left(1-\cos^2(\theta)\right)^l e^{\mathrm{i}m_l\phi}}\tag{13.1.181}$$

Finally, after this highly mathematical interlude (but instructive for the methodology of approach), we see (which is logical) that to each value of l correspond therefore $2l+1$ eigenfunctions $Y_{m_l,l}(\theta,\phi)$. We also say that the value $\hbar l(l+1)$ is $2l+1$ times degenerated since:

$$\begin{gathered}\hat{L}^2Y_{m_l,l}(\theta,\phi)=L^2Y_{m_l,l}(\theta,\phi)\\ \Updownarrow\\ \hat{L}^2Y_{m_l,l}(\theta,\phi)=\hbar^2l(l+1)Y_{m_l,l}(\theta,\phi)\end{gathered}\tag{13.1.182}$$

Here are some values of the function $Y_{m_l,l}(\theta,\phi)$ (here below denoted $Y_l^{m_l}(\theta,\phi)$) that generates what we commonly name "spherical harmonics"[2]. Some of these formulas give the "Cartesian" version (useful for some softwares that doesn't manage directly plots in spherical coordinates...). This assumes x, y, z, and r are related to θ and ϕ through the usual spherical-to-Cartesian coordinate transformation (see section Vector Calculus page 1849):

$$\begin{cases}x=r\sin(\theta)\cos(\phi)\\ y=r\sin(\theta)\sin(\phi)\\ z=r\cos(\theta)\end{cases}\tag{13.1.183}$$

So let's start!:

$$Y_0^0(\theta,\phi)=\frac{1}{2}\sqrt{\frac{1}{\pi}}\tag{13.1.184}$$

$$\begin{aligned}Y_1^{-1}(\theta,\phi)=&\quad\frac{1}{2}\sqrt{\frac{3}{2\pi}}e^{-\mathrm{i}\phi}\sin(\theta)&=&\quad\frac{1}{2}\sqrt{\frac{3}{2\pi}}\frac{(x-\mathrm{i}y)}{r}\\ Y_1^0(\theta,\phi)=&\quad\frac{1}{2}\sqrt{\frac{3}{\pi}}\cos(\theta)&=&\quad\frac{1}{2}\sqrt{\frac{3}{\pi}}\frac{z}{r}\\ Y_1^1(\theta,\phi)=&\;-\frac{1}{2}\sqrt{\frac{3}{2\pi}}e^{\mathrm{i}\phi}\sin(\theta)&=&\;-\frac{1}{2}\sqrt{\frac{3}{2\pi}}\frac{(x+\mathrm{i}y)}{r}\end{aligned}\tag{13.1.185}$$

---

[2] A special thanks to the authors of the following Wikipedia page `https://en.wikipedia.org/wiki/Table_of_spherical_harmonics` and we recommend by the way to take a close look to the *Talk* tab of this page!

$$\begin{aligned}
Y_2^{-2}(\theta,\phi) &= \frac{1}{4}\sqrt{\frac{15}{2\pi}}e^{-2\mathrm{i}\phi}\sin^2(\theta) &&= \frac{1}{4}\sqrt{\frac{15}{2\pi}}\frac{(x-\mathrm{i}y)^2}{r^2}\\
Y_2^{-1}(\theta,\phi) &= \frac{1}{2}\sqrt{\frac{15}{2\pi}}e^{-\mathrm{i}\phi}\sin(\theta)\cos(\theta) &&= \frac{1}{2}\sqrt{\frac{15}{2\pi}}\frac{(x-\mathrm{i}y)z}{r^2}\\
Y_2^{0}(\theta,\phi) &= \frac{1}{4}\sqrt{\frac{5}{\pi}}(3\cos^2(\theta)-1) &&= \frac{1}{4}\sqrt{\frac{5}{\pi}}\frac{(2z^2-x^2-y^2)}{r^2}\\
Y_2^{1}(\theta,\phi) &= -\frac{1}{2}\sqrt{\frac{15}{2\pi}}e^{\mathrm{i}\phi}\sin(\theta)\cos(\theta) &&= -\frac{1}{2}\sqrt{\frac{15}{2\pi}}\frac{(x+\mathrm{i}y)z}{r^2}\\
Y_2^{2}(\theta,\phi) &= \frac{1}{4}\sqrt{\frac{15}{2\pi}}e^{2\mathrm{i}\phi}\sin^2(\theta) &&= \frac{1}{4}\sqrt{\frac{15}{2\pi}}\frac{(x+\mathrm{i}y)^2}{r^2}
\end{aligned} \tag{13.1.186}$$

$$\begin{aligned}
Y_3^{-3}(\theta,\phi) &= \frac{1}{8}\sqrt{\frac{35}{\pi}}e^{-3\mathrm{i}\phi}\sin^3(\theta) &&= \frac{1}{8}\sqrt{\frac{35}{\pi}}\frac{(x-\mathrm{i}y)^3}{r^3}\\
Y_3^{-2}(\theta,\phi) &= \frac{1}{4}\sqrt{\frac{105}{2\pi}}e^{-2\mathrm{i}\phi}\sin^2(\theta)\cos(\theta) &&= \frac{1}{4}\sqrt{\frac{105}{2\pi}}\frac{(x-\mathrm{i}y)^2z}{r^3}\\
Y_3^{-1}(\theta,\phi) &= \frac{1}{8}\sqrt{\frac{21}{\pi}}e^{-\mathrm{i}\phi}\sin(\theta)(5\cos^2(\theta)-1) &&= \frac{1}{8}\sqrt{\frac{21}{\pi}}\frac{(x-\mathrm{i}y)(4z^2-x^2-y^2)}{r^3}\\
Y_3^{0}(\theta,\phi) &= \frac{1}{4}\sqrt{\frac{7}{\pi}}(5\cos^3(\theta)-3\cos(\theta)) &&= \frac{1}{4}\sqrt{\frac{7}{\pi}}\frac{z(2z^2-3x^2-3y^2)}{r^3}\\
Y_3^{1}(\theta,\phi) &= -\frac{1}{8}\sqrt{\frac{21}{\pi}}e^{\mathrm{i}\phi}\sin(\theta)(5\cos^2(\theta)-1) &&= \frac{-1}{8}\sqrt{\frac{21}{\pi}}\frac{(x+\mathrm{i}y)(4z^2-x^2-y^2)}{r^3}\\
Y_3^{2}(\theta,\phi) &= \frac{1}{4}\sqrt{\frac{105}{2\pi}}e^{2\mathrm{i}\phi}\sin^2(\theta)\cos(\theta) &&= \frac{1}{4}\sqrt{\frac{105}{2\pi}}\frac{(x+\mathrm{i}y)^2z}{r^3}\\
Y_3^{3}(\theta,\phi) &= -\frac{1}{8}\sqrt{\frac{35}{\pi}}e^{3\mathrm{i}\phi}\sin^3(\theta) &&= \frac{-1}{8}\sqrt{\frac{35}{\pi}}\frac{(x+\mathrm{i}y)^3}{r^3}
\end{aligned} \tag{13.1.187}$$

$$\begin{aligned}
Y_4^{-4}(\theta,\phi) &= \frac{3}{16}\sqrt{\frac{35}{2\pi}}e^{-4\mathrm{i}\phi}\sin^4(\theta) = \frac{3}{16}\sqrt{\frac{35}{2\pi}}\frac{(x-\mathrm{i}y)^4}{r^4}\\
Y_4^{-3}(\theta,\phi) &= \frac{3}{8}\sqrt{\frac{35}{\pi}}e^{-3\mathrm{i}\phi}\sin^3(\theta)\cos(\theta) = \frac{3}{8}\sqrt{\frac{35}{\pi}}\frac{(x-\mathrm{i}y)^3 z}{r^4}\\
Y_4^{-2}(\theta,\phi) &= \frac{3}{8}\sqrt{\frac{5}{2\pi}}e^{-2\mathrm{i}\phi}\sin^2(\theta)(7\cos^2(\theta)-1) = \frac{3}{8}\sqrt{\frac{5}{2\pi}}\frac{(x-\mathrm{i}y)^2(7z^2-r^2)}{r^4}\\
Y_4^{-1}(\theta,\phi) &= \frac{3}{8}\sqrt{\frac{5}{\pi}}e^{-\mathrm{i}\phi}\sin(\theta)(7\cos^3(\theta)-3\cos(\theta)) = \frac{3}{8}\sqrt{\frac{5}{\pi}}\frac{(x-\mathrm{i}y)z(7z^2-3r^2)}{r^4}\\
Y_4^{0}(\theta,\phi) &= \frac{3}{16}\sqrt{\frac{1}{\pi}}(35\cos^4(\theta)-30\cos^2(\theta)+3) = \frac{3}{16}\sqrt{\frac{1}{\pi}}\frac{(35z^4-30z^2r^2+3r^4)}{r^4}\\
Y_4^{1}(\theta,\phi) &= \frac{-3}{8}\sqrt{\frac{5}{\pi}}e^{\mathrm{i}\phi}\sin(\theta)(7\cos^3(\theta)-3\cos(\theta)) = \frac{-3}{8}\sqrt{\frac{5}{\pi}}\frac{(x+\mathrm{i}y)z(7z^2-3r^2)}{r^4}\\
Y_4^{2}(\theta,\phi) &= \frac{3}{8}\sqrt{\frac{5}{2\pi}}e^{2\mathrm{i}\phi}\sin^2(\theta)(7\cos^2(\theta)-1) = \frac{3}{8}\sqrt{\frac{5}{2\pi}}\frac{(x+\mathrm{i}y)^2(7z^2-r^2)}{r^4}\\
Y_4^{3}(\theta,\phi) &= \frac{-3}{8}\sqrt{\frac{35}{\pi}}e^{3\mathrm{i}\phi}\sin^3(\theta)\cos(\theta) = \frac{-3}{8}\sqrt{\frac{35}{\pi}}\frac{(x+\mathrm{i}y)^3 z}{r^4}\\
Y_4^{4}(\theta,\phi) &= \frac{3}{16}\sqrt{\frac{35}{2\pi}}e^{4\mathrm{i}\phi}\sin^4(\theta) = \frac{3}{16}\sqrt{\frac{35}{2\pi}}\frac{(x+\mathrm{i}y)^4}{r^4}
\end{aligned}\tag{13.1.188}$$

$$\begin{aligned}
Y_5^{-5}(\theta,\phi) &= \frac{3}{32}\sqrt{\frac{77}{\pi}}e^{-5\mathrm{i}\phi}\sin^5(\theta)\\
Y_5^{-4}(\theta,\phi) &= \frac{3}{16}\sqrt{\frac{385}{2\pi}}e^{-4\mathrm{i}\phi}\sin^4(\theta)\cos(\theta)\\
Y_5^{-3}(\theta,\phi) &= \frac{1}{32}\sqrt{\frac{385}{\pi}}e^{-3\mathrm{i}\phi}\sin^3(\theta)(9\cos^2(\theta)-1)\\
Y_5^{-2}(\theta,\phi) &= \frac{1}{8}\sqrt{\frac{1155}{2\pi}}e^{-2\mathrm{i}\phi}\sin^2(\theta)(3\cos^3(\theta)-\cos(\theta))\\
Y_5^{-1}(\theta,\phi) &= \frac{1}{16}\sqrt{\frac{165}{2\pi}}e^{-\mathrm{i}\phi}\sin(\theta)(21\cos^4(\theta)-14\cos^2(\theta)+1)\\
Y_5^{0}(\theta,\phi) &= \frac{1}{16}\sqrt{\frac{11}{\pi}}(63\cos^5(\theta)-70\cos^3(\theta)+15\cos(\theta))\\
Y_5^{1}(\theta,\phi) &= \frac{-1}{16}\sqrt{\frac{165}{2\pi}}e^{\mathrm{i}\phi}\sin(\theta)(21\cos^4(\theta)-14\cos^2(\theta)+1)\\
Y_5^{2}(\theta,\phi) &= \frac{1}{8}\sqrt{\frac{1155}{2\pi}}e^{2\mathrm{i}\phi}\sin^2(\theta)(3\cos^3(\theta)-\cos(\theta))\\
Y_5^{3}(\theta,\phi) &= \frac{-1}{32}\sqrt{\frac{385}{\pi}}e^{3\mathrm{i}\phi}\sin^3(\theta)(9\cos^2(\theta)-1)\\
Y_5^{4}(\theta,\phi) &= \frac{3}{16}\sqrt{\frac{385}{2\pi}}e^{4\mathrm{i}\phi}\sin^4(\theta)\cos(\theta)\\
Y_5^{5}(\theta,\phi) &= \frac{-3}{32}\sqrt{\frac{77}{\pi}}e^{5\mathrm{i}\phi}\sin^5(\theta)
\end{aligned}\tag{13.1.189}$$

$$
\begin{aligned}
Y_6^{-6}(\theta,\phi) &= \frac{1}{64}\sqrt{\frac{3003}{\pi}}e^{-6\mathrm{i}\phi}\sin^6(\theta)\\
Y_6^{-5}(\theta,\phi) &= \frac{3}{32}\sqrt{\frac{1001}{\pi}}e^{-5\mathrm{i}\phi}\sin^5(\theta)\cos(\theta)\\
Y_6^{-4}(\theta,\phi) &= \frac{3}{32}\sqrt{\frac{91}{2\pi}}e^{-4\mathrm{i}\phi}\sin^4(\theta)(11\cos^2(\theta)-1)\\
Y_6^{-3}(\theta,\phi) &= \frac{1}{32}\sqrt{\frac{1365}{\pi}}e^{-3\mathrm{i}\phi}\sin^3(\theta)(11\cos^3(\theta)-3\cos(\theta))\\
Y_6^{-2}(\theta,\phi) &= \frac{1}{64}\sqrt{\frac{1365}{\pi}}e^{-2\mathrm{i}\phi}\sin^2(\theta)(33\cos^4(\theta)-18\cos^2(\theta)+1)\\
Y_6^{-1}(\theta,\phi) &= \frac{1}{16}\sqrt{\frac{273}{2\pi}}e^{-\mathrm{i}\phi}\sin(\theta)(33\cos^5(\theta)-30\cos^3(\theta)+5\cos(\theta))\\
Y_6^{0}(\theta,\phi) &= \frac{1}{32}\sqrt{\frac{13}{\pi}}(231\cos^6(\theta)-315\cos^4(\theta)+105\cos^2(\theta)-5)\\
Y_6^{1}(\theta,\phi) &= -\frac{1}{16}\sqrt{\frac{273}{2\pi}}e^{\mathrm{i}\phi}\sin(\theta)(33\cos^5(\theta)-30\cos^3(\theta)+5\cos(\theta))\\
Y_6^{2}(\theta,\phi) &= \frac{1}{64}\sqrt{\frac{1365}{\pi}}e^{2\mathrm{i}\phi}\sin^2(\theta)(33\cos^4(\theta)-18\cos^2(\theta)+1)\\
Y_6^{3}(\theta,\phi) &= -\frac{1}{32}\sqrt{\frac{1365}{\pi}}e^{3\mathrm{i}\phi}\sin^3(\theta)(11\cos^3(\theta)-3\cos(\theta))\\
Y_6^{4}(\theta,\phi) &= \frac{3}{32}\sqrt{\frac{91}{2\pi}}e^{4\mathrm{i}\phi}\sin^4(\theta)(11\cos^2(\theta)-1)\\
Y_6^{5}(\theta,\phi) &= -\frac{3}{32}\sqrt{\frac{1001}{\pi}}e^{5\mathrm{i}\phi}\sin^5(\theta)\cos(\theta)\\
Y_6^{6}(\theta,\phi) &= \frac{1}{64}\sqrt{\frac{3003}{\pi}}e^{6\mathrm{i}\phi}\sin^6(\theta)
\end{aligned}
\tag{13.1.190}
$$

$$
\begin{aligned}
Y_7^{-7}(\theta,\phi) &= \frac{3}{64}\sqrt{\frac{715}{2\pi}}e^{-7i\phi}\sin^7(\theta)\\
Y_7^{-6}(\theta,\phi) &= \frac{3}{64}\sqrt{\frac{5005}{\pi}}e^{-6i\phi}\sin^6(\theta)\cos(\theta)\\
Y_7^{-5}(\theta,\phi) &= \frac{3}{64}\sqrt{\frac{385}{2\pi}}e^{-5i\phi}\sin^5(\theta)(13\cos^2(\theta)-1)\\
Y_7^{-4}(\theta,\phi) &= \frac{3}{32}\sqrt{\frac{385}{2\pi}}e^{-4i\phi}\sin^4(\theta)(13\cos^3(\theta)-3\cos(\theta))\\
Y_7^{-3}(\theta,\phi) &= \frac{3}{64}\sqrt{\frac{35}{2\pi}}e^{-3i\phi}\sin^3(\theta)(143\cos^4(\theta)-66\cos^2(\theta)+3)\\
Y_7^{-2}(\theta,\phi) &= \frac{3}{64}\sqrt{\frac{35}{\pi}}e^{-2i\phi}\sin^2(\theta)(143\cos^5(\theta)-110\cos^3(\theta)+15\cos(\theta))\\
Y_7^{-1}(\theta,\phi) &= \frac{1}{64}\sqrt{\frac{105}{2\pi}}e^{-i\phi}\sin(\theta)(429\cos^6(\theta)-495\cos^4(\theta)+135\cos^2(\theta)-5)\\
Y_7^{0}(\theta,\phi) &= \frac{1}{32}\sqrt{\frac{15}{\pi}}(429\cos^7(\theta)-693\cos^5(\theta)+315\cos^3(\theta)-35\cos(\theta))\\
Y_7^{1}(\theta,\phi) &= -\frac{1}{64}\sqrt{\frac{105}{2\pi}}e^{i\phi}\sin(\theta)(429\cos^6(\theta)-495\cos^4(\theta)+135\cos^2(\theta)-5)\\
Y_7^{2}(\theta,\phi) &= \frac{3}{64}\sqrt{\frac{35}{\pi}}e^{2i\phi}\sin^2(\theta)(143\cos^5(\theta)-110\cos^3(\theta)+15\cos(\theta))\\
Y_7^{3}(\theta,\phi) &= -\frac{3}{64}\sqrt{\frac{35}{2\pi}}e^{3i\phi}\sin^3(\theta)(143\cos^4(\theta)-66\cos^2(\theta)+3)\\
Y_7^{4}(\theta,\phi) &= \frac{3}{32}\sqrt{\frac{385}{2\pi}}e^{4i\phi}\sin^4(\theta)(13\cos^3(\theta)-3\cos(\theta))\\
Y_7^{5}(\theta,\phi) &= -\frac{3}{64}\sqrt{\frac{385}{2\pi}}e^{5i\phi}\sin^5(\theta)(13\cos^2(\theta)-1)\\
Y_7^{6}(\theta,\phi) &= \frac{3}{64}\sqrt{\frac{5005}{\pi}}e^{6i\phi}\sin^6(\theta)\cos(\theta)\\
Y_7^{7}(\theta,\phi) &= -\frac{3}{64}\sqrt{\frac{715}{2\pi}}e^{7i\phi}\sin^7(\theta)
\end{aligned}
\tag{13.1.191}
$$

$$
\begin{aligned}
Y_8^{-8}(\theta,\phi) &= \frac{3}{256}\sqrt{\frac{12155}{2\pi}}e^{-8i\phi}\sin^8(\theta)\\
Y_8^{-7}(\theta,\phi) &= \frac{3}{64}\sqrt{\frac{12155}{2\pi}}e^{-7i\phi}\sin^7(\theta)\cos(\theta)\\
Y_8^{-6}(\theta,\phi) &= \frac{1}{128}\sqrt{\frac{7293}{\pi}}e^{-6i\phi}\sin^6(\theta)(15\cos^2(\theta)-1)\\
Y_8^{-5}(\theta,\phi) &= \frac{3}{64}\sqrt{\frac{17017}{2\pi}}e^{-5i\phi}\sin^5(\theta)(5\cos^3(\theta)-\cos(\theta))\\
Y_8^{-4}(\theta,\phi) &= \frac{3}{128}\sqrt{\frac{1309}{2\pi}}e^{-4i\phi}\sin^4(\theta)(65\cos^4(\theta)-26\cos^2(\theta)+1)\\
Y_8^{-3}(\theta,\phi) &= \frac{1}{64}\sqrt{\frac{19635}{2\pi}}e^{-3i\phi}\sin^3(\theta)(39\cos^5(\theta)-26\cos^3(\theta)+3\cos(\theta))\\
Y_8^{-2}(\theta,\phi) &= \frac{3}{128}\sqrt{\frac{595}{\pi}}e^{-2i\phi}\sin^2(\theta)(143\cos^6(\theta)-143\cos^4(\theta)+33\cos^2(\theta)-1)\\
Y_8^{-1}(\theta,\phi) &= \frac{3}{64}\sqrt{\frac{17}{2\pi}}e^{-i\phi}\sin(\theta)(715\cos^7(\theta)-1001\cos^5(\theta)+385\cos^3(\theta)-35\cos(\theta))\\
Y_8^{0}(\theta,\phi) &= \frac{1}{256}\sqrt{\frac{17}{\pi}}(6435\cos^8(\theta)-12012\cos^6(\theta)+6930\cos^4(\theta)-1260\cos^2(\theta)+35)\\
Y_8^{1}(\theta,\phi) &= \frac{-3}{64}\sqrt{\frac{17}{2\pi}}e^{i\phi}\sin(\theta)(715\cos^7(\theta)-1001\cos^5(\theta)+385\cos^3(\theta)-35\cos(\theta))\\
Y_8^{2}(\theta,\phi) &= \frac{3}{128}\sqrt{\frac{595}{\pi}}e^{2i\phi}\sin^2(\theta)(143\cos^6(\theta)-143\cos^4(\theta)+33\cos^2(\theta)-1)\\
Y_8^{3}(\theta,\phi) &= \frac{-1}{64}\sqrt{\frac{19635}{2\pi}}e^{3i\phi}\sin^3(\theta)(39\cos^5(\theta)-26\cos^3(\theta)+3\cos(\theta))\\
Y_8^{4}(\theta,\phi) &= \frac{3}{128}\sqrt{\frac{1309}{2\pi}}e^{4i\phi}\sin^4(\theta)(65\cos^4(\theta)-26\cos^2(\theta)+1)\\
Y_8^{5}(\theta,\phi) &= \frac{-3}{64}\sqrt{\frac{17017}{2\pi}}e^{5i\phi}\sin^5(\theta)(5\cos^3(\theta)-\cos(\theta))\\
Y_8^{6}(\theta,\phi) &= \frac{1}{128}\sqrt{\frac{7293}{\pi}}e^{6i\phi}\sin^6(\theta)(15\cos^2(\theta)-1)\\
Y_8^{7}(\theta,\phi) &= \frac{-3}{64}\sqrt{\frac{12155}{2\pi}}e^{7i\phi}\sin^7(\theta)\cos(\theta)\\
Y_8^{8}(\theta,\phi) &= \frac{3}{256}\sqrt{\frac{12155}{2\pi}}e^{8i\phi}\sin^8(\theta)
\end{aligned}
\tag{13.1.192}
$$

$$\begin{aligned}
Y_9^{-9}(\theta,\phi) &= \frac{1}{512}\sqrt{\frac{230945}{\pi}}e^{-9i\phi}\sin^9(\theta)\\
Y_9^{-8}(\theta,\phi) &= \frac{3}{256}\sqrt{\frac{230945}{2\pi}}e^{-8i\phi}\sin^8(\theta)\cos(\theta)\\
Y_9^{-7}(\theta,\phi) &= \frac{3}{512}\sqrt{\frac{13585}{\pi}}e^{-7i\phi}\sin^7(\theta)(17\cos^2(\theta)-1)\\
Y_9^{-6}(\theta,\phi) &= \frac{1}{128}\sqrt{\frac{40755}{\pi}}e^{-6i\phi}\sin^6(\theta)(17\cos^3(\theta)-3\cos(\theta))\\
Y_9^{-5}(\theta,\phi) &= \frac{3}{256}\sqrt{\frac{2717}{\pi}}e^{-5i\phi}\sin^5(\theta)(85\cos^4(\theta)-30\cos^2(\theta)+1)\\
Y_9^{-4}(\theta,\phi) &= \frac{3}{128}\sqrt{\frac{95095}{2\pi}}e^{-4i\phi}\sin^4(\theta)(17\cos^5(\theta)-10\cos^3(\theta)+\cos(\theta))\\
Y_9^{-3}(\theta,\phi) &= \frac{1}{256}\sqrt{\frac{21945}{\pi}}e^{-3i\phi}\sin^3(\theta)(221\cos^6(\theta)-195\cos^4(\theta)+39\cos^2(\theta)-1)\\
Y_9^{-2}(\theta,\phi) &= \frac{3}{128}\sqrt{\frac{1045}{\pi}}e^{-2i\phi}\sin^2(\theta)(221\cos^7(\theta)-273\cos^5(\theta)+91\cos^3(\theta)-7\cos(\theta))\\
Y_9^{-1}(\theta,\phi) &= \frac{3}{256}\sqrt{\frac{95}{2\pi}}e^{-i\phi}\sin(\theta)(2431\cos^8(\theta)-4004\cos^6(\theta)+2002\cos^4(\theta)-308\cos^2(\theta)+7)\\
Y_9^{0}(\theta,\phi) &= \frac{1}{256}\sqrt{\frac{19}{\pi}}(12155\cos^9(\theta)-25740\cos^7(\theta)+18018\cos^5(\theta)-4620\cos^3(\theta)+315\cos(\theta))\\
Y_9^{1}(\theta,\phi) &= \frac{-3}{256}\sqrt{\frac{95}{2\pi}}e^{i\phi}\sin(\theta)(2431\cos^8(\theta)-4004\cos^6(\theta)+2002\cos^4(\theta)-308\cos^2(\theta)+7)\\
Y_9^{2}(\theta,\phi) &= \frac{3}{128}\sqrt{\frac{1045}{\pi}}e^{2i\phi}\sin^2(\theta)(221\cos^7(\theta)-273\cos^5(\theta)+91\cos^3(\theta)-7\cos(\theta))\\
Y_9^{3}(\theta,\phi) &= \frac{-1}{256}\sqrt{\frac{21945}{\pi}}e^{3i\phi}\sin^3(\theta)(221\cos^6(\theta)-195\cos^4(\theta)+39\cos^2(\theta)-1)\\
Y_9^{4}(\theta,\phi) &= \frac{3}{128}\sqrt{\frac{95095}{2\pi}}e^{4i\phi}\sin^4(\theta)(17\cos^5(\theta)-10\cos^3(\theta)+\cos(\theta))\\
Y_9^{5}(\theta,\phi) &= \frac{-3}{256}\sqrt{\frac{2717}{\pi}}e^{5i\phi}\sin^5(\theta)(85\cos^4(\theta)-30\cos^2(\theta)+1)\\
Y_9^{6}(\theta,\phi) &= \frac{1}{128}\sqrt{\frac{40755}{\pi}}e^{6i\phi}\sin^6(\theta)(17\cos^3(\theta)-3\cos(\theta))\\
Y_9^{7}(\theta,\phi) &= \frac{-3}{512}\sqrt{\frac{13585}{\pi}}e^{7i\phi}\sin^7(\theta)(17\cos^2(\theta)-1)\\
Y_9^{8}(\theta,\phi) &= \frac{3}{256}\sqrt{\frac{230945}{2\pi}}e^{8i\phi}\sin^8(\theta)\cos(\theta)\\
Y_9^{9}(\theta,\phi) &= \frac{-1}{512}\sqrt{\frac{230945}{\pi}}e^{9i\phi}\sin^9(\theta)
\end{aligned}\tag{13.1.193}$$

$$\begin{aligned}
Y_{10}^{-10}(\theta,\phi) &= \frac{1}{1024}\sqrt{\frac{969969}{\pi}}e^{-10\mathrm{i}\phi}\sin^{10}(\theta)\\
Y_{10}^{-9}(\theta,\phi) &= \frac{1}{512}\sqrt{\frac{4849845}{\pi}}e^{-9\mathrm{i}\phi}\sin^{9}(\theta)\cos(\theta)\\
Y_{10}^{-8}(\theta,\phi) &= \frac{1}{512}\sqrt{\frac{255255}{2\pi}}e^{-8\mathrm{i}\phi}\sin^{8}(\theta)(19\cos^2(\theta)-1)\\
Y_{10}^{-7}(\theta,\phi) &= \frac{3}{512}\sqrt{\frac{85085}{\pi}}e^{-7\mathrm{i}\phi}\sin^{7}(\theta)(19\cos^3(\theta)-3\cos(\theta))\\
Y_{10}^{-6}(\theta,\phi) &= \frac{3}{1024}\sqrt{\frac{5005}{\pi}}e^{-6\mathrm{i}\phi}\sin^{6}(\theta)(323\cos^4(\theta)-102\cos^2(\theta)+3)\\
Y_{10}^{-5}(\theta,\phi) &= \frac{3}{256}\sqrt{\frac{1001}{\pi}}e^{-5\mathrm{i}\phi}\sin^{5}(\theta)(323\cos^5(\theta)-170\cos^3(\theta)+15\cos(\theta))\\
Y_{10}^{-4}(\theta,\phi) &= \frac{3}{256}\sqrt{\frac{5005}{2\pi}}e^{-4\mathrm{i}\phi}\sin^{4}(\theta)(323\cos^6(\theta)-255\cos^4(\theta)+45\cos^2(\theta)-1)\\
Y_{10}^{-3}(\theta,\phi) &= \frac{3}{256}\sqrt{\frac{5005}{\pi}}e^{-3\mathrm{i}\phi}\sin^{3}(\theta)(323\cos^7(\theta)-357\cos^5(\theta)+105\cos^3(\theta)-7\cos(\theta))\\
Y_{10}^{-2}(\theta,\phi) &= \frac{3}{512}\sqrt{\frac{385}{2\pi}}e^{-2\mathrm{i}\phi}\sin^{2}(\theta)(4199\cos^8(\theta)-6188\cos^6(\theta)+2730\cos^4(\theta)-364\cos^2(\theta)+7)\\
Y_{10}^{-1}(\theta,\phi) &= \frac{1}{256}\sqrt{\frac{1155}{2\pi}}e^{-\mathrm{i}\phi}\sin(\theta)(4199\cos^9(\theta)-7956\cos^7(\theta)+4914\cos^5(\theta)-1092\cos^3(\theta)+63\cos(\theta))\\
Y_{10}^{0}(\theta,\phi) &= \frac{1}{512}\sqrt{\frac{21}{\pi}}(46189\cos^{10}(\theta)-109395\cos^8(\theta)+90090\cos^6(\theta)-30030\cos^4(\theta)+3465\cos^2(\theta)-63)\\
Y_{10}^{1}(\theta,\phi) &= \frac{-1}{256}\sqrt{\frac{1155}{2\pi}}e^{\mathrm{i}\phi}\sin(\theta)(4199\cos^9(\theta)-7956\cos^7(\theta)+4914\cos^5(\theta)-1092\cos^3(\theta)+63\cos(\theta))\\
Y_{10}^{2}(\theta,\phi) &= \frac{3}{512}\sqrt{\frac{385}{2\pi}}e^{2\mathrm{i}\phi}\sin^{2}(\theta)(4199\cos^8(\theta)-6188\cos^6(\theta)+2730\cos^4(\theta)-364\cos^2(\theta)+7)\\
Y_{10}^{3}(\theta,\phi) &= \frac{-3}{256}\sqrt{\frac{5005}{\pi}}e^{3\mathrm{i}\phi}\sin^{3}(\theta)(323\cos^7(\theta)-357\cos^5(\theta)+105\cos^3(\theta)-7\cos(\theta))\\
Y_{10}^{4}(\theta,\phi) &= \frac{3}{256}\sqrt{\frac{5005}{2\pi}}e^{4\mathrm{i}\phi}\sin^{4}(\theta)(323\cos^6(\theta)-255\cos^4(\theta)+45\cos^2(\theta)-1)\\
Y_{10}^{5}(\theta,\phi) &= \frac{-3}{256}\sqrt{\frac{1001}{\pi}}e^{5\mathrm{i}\phi}\sin^{5}(\theta)(323\cos^5(\theta)-170\cos^3(\theta)+15\cos(\theta))\\
Y_{10}^{6}(\theta,\phi) &= \frac{3}{1024}\sqrt{\frac{5005}{\pi}}e^{6\mathrm{i}\phi}\sin^{6}(\theta)(323\cos^4(\theta)-102\cos^2(\theta)+3)\\
Y_{10}^{7}(\theta,\phi) &= \frac{-3}{512}\sqrt{\frac{85085}{\pi}}e^{7\mathrm{i}\phi}\sin^{7}(\theta)(19\cos^3(\theta)-3\cos(\theta))\\
Y_{10}^{8}(\theta,\phi) &= \frac{1}{512}\sqrt{\frac{255255}{2\pi}}e^{8\mathrm{i}\phi}\sin^{8}(\theta)(19\cos^2(\theta)-1)\\
Y_{10}^{9}(\theta,\phi) &= \frac{-1}{512}\sqrt{\frac{4849845}{\pi}}e^{9\mathrm{i}\phi}\sin^{9}(\theta)\cos(\theta)\\
Y_{10}^{10}(\theta,\phi) &= \frac{1}{1024}\sqrt{\frac{969969}{\pi}}e^{10\mathrm{i}\phi}\sin^{10}(\theta)
\end{aligned}\tag{13.1.194}$$

Let's see some plots of these beautiful spherical harmonics that can be obtained with Maple™ 4.00b[3] by using the following command (this is the $Y_2^{-1}(\theta,\phi)$ spherical harmonic function above):

---

[3] For the `R` equivalent code, the reader can refer to the `R` companion book

```
>plot3d(Re(-sqrt(15/(8*Pi))*(sin(theta)*cos(theta)*exp(-I*phi)))^2,phi=0..2*Pi,
theta=0..Pi, coords=spherical,scaling=constrained,numpoints=5000,axes=frame);
```

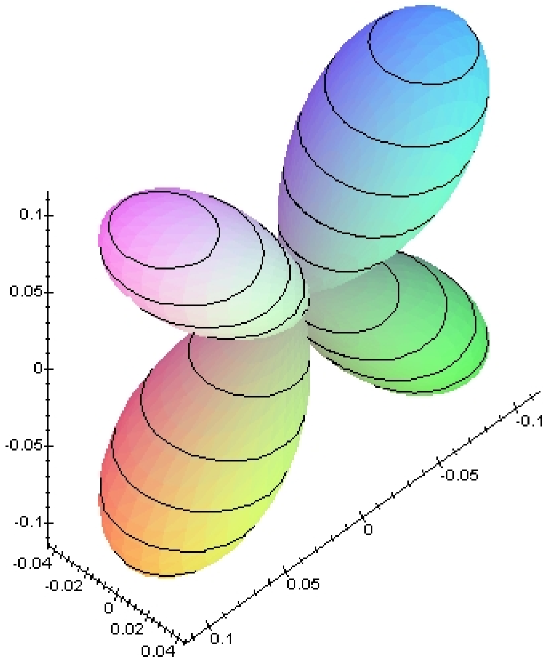


Figure 13.6: Plot of the spherical harmonics $Y_{1,2}$

- $Y_{0,0}$ (corresponding to $n=1$!) gives a sphere (constant value regardless θ,ϕ) which the probability density can be represented by the following "photographic card" or "density map" (the density in a given state is represented by the density of light spotted on a dark background):

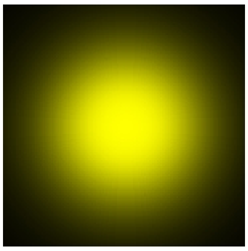


Figure 13.7: $1s$ density map

Representing the possible $1s$ orbits.

- $Y_{0,1},Y_{1,1},Y_{-1,1}$ give (for $n=2$ at least!):

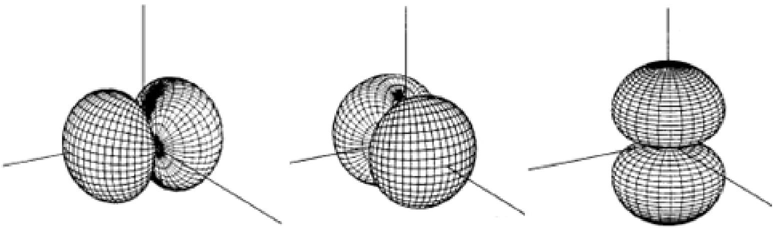


Figure 13.8: $2p$ orbitals (spherical harmonics)

Which represents the possible $2p$ orbits, the probability density function can be represented by its density and isodensity maps:

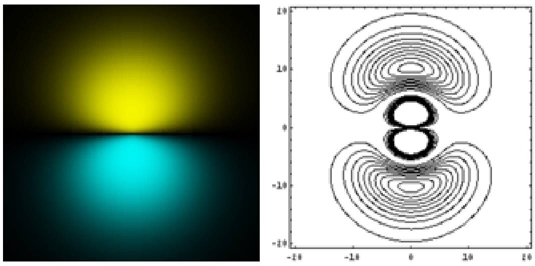


Figure 13.9: $2p$ **density map**

- $Y_{-2,2}, Y_{-1,2}, Y_{1,2}, Y_{2,2}, Y_{0,2}$ give (for $n = 3$ at least!):

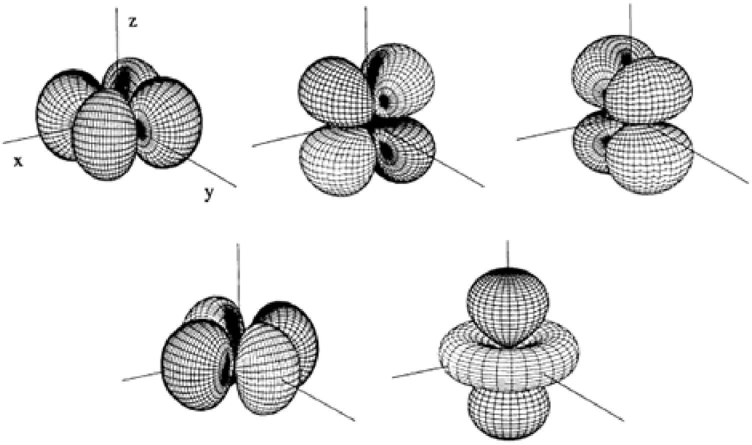


Figure 13.10: $3d$ **orbitals (spherical harmonics)**

Representing 5 possible $3d$ centrosymmetric orbits, which the probability density can be represented by (the last two maps represent $Y_{0,2}$) the following density maps:

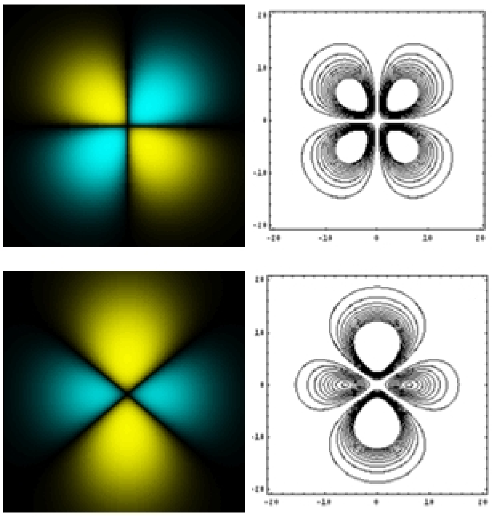


Figure 13.11: $3d$ **density map**

- $Y_{-3,3}, Y_{-2,3}, Y_{-1,3}, Y_{1,3}, Y_{2,3}, Y_{3,3}, Y_{0,3}$ give (for $n = 4$ at least!):

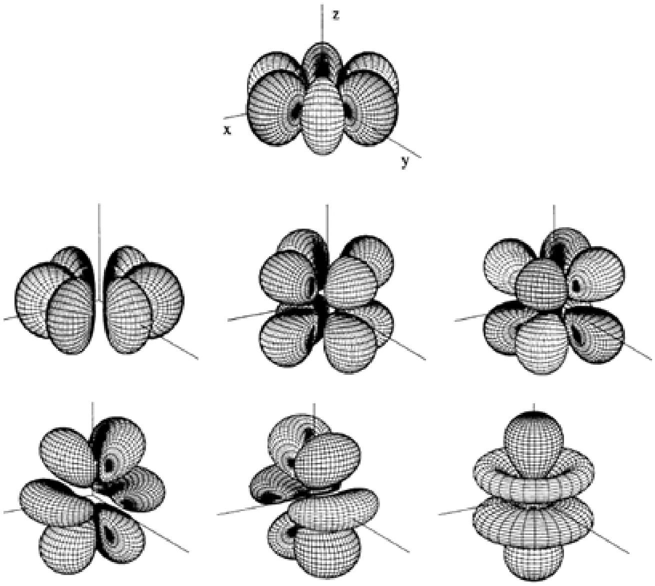


Figure 13.12: $3d$ orbitals (spherical harmonics)

Representing 7 possible $3f$ anti-centrosymmetric orbits, which the probability density can be represented by (in the order: $Y_{0,3}, Y_{\pm 1,3}, Y_{\pm 2,3}, Y_{\pm 3,3}$) the following density maps:

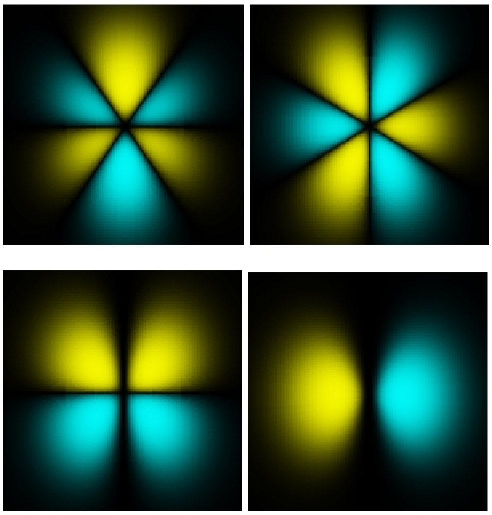


Figure 13.13: $4f$ density map

Or in a much better quality (from top-left to bottom right we have $Y_0^0, Y_1^{-1}, Y_1^0, Y_1^1, Y_2^{-2}, Y_2^{-1}, Y_2^0, Y_2^1, Y_2^2, Y_3^0$):

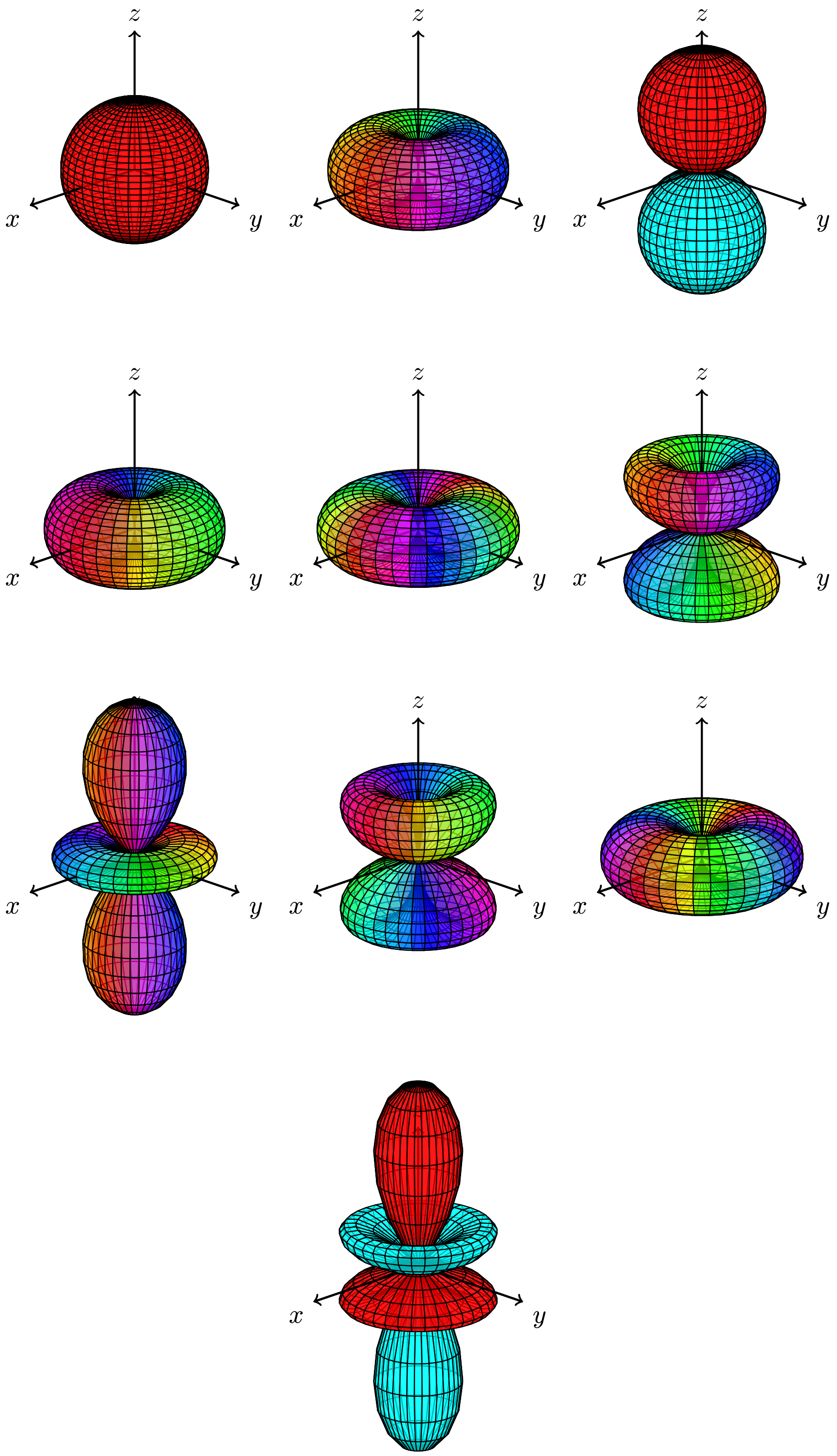


Figure 13.14: **3d orbitals (spherical harmonics) made with TikZ (author: Jeff Hein)**

The above results thus lead us to write:

$$\Psi(r,\theta,\phi)=R(r)\Phi(\phi)\Theta(\theta)=R(r)Y_{m_l,l}(\theta,\phi) \qquad (13.1.195)$$

Substituting this in the Schrödinger equation:

$$\left[T_r+\frac{l(l+1)\hbar^2}{2\mu r}+E_p\right]\Psi=E\Psi \qquad (13.1.196)$$

We get ($T_r = 0$ in the rigid rotor but $\neq 0$ in the case of the hydrogen atom):

$$\left[T_r + \frac{l(l+1)\hbar^2}{2\mu r} + E_p\right] R(r)Y_{m_l,l}(\theta,\phi) = ER(r)Y_{m_l,l}(\theta,\phi) \tag{13.1.197}$$

As there is in this relation no operator which acts on $Y_{m_l,l}(\theta,\phi)$, we can simplify to obtain:

$$\left[T_r + \frac{l(l+1)\hbar^2}{2\mu r} + E_p\right] R(r) = ER(r) \tag{13.1.198}$$

which we see in the general case of the isolated atom that energy levels are no longer dependent of m_l (due to the spherical symmetry of the potential). Then we say that the levels corresponding to the same values of n and of l are all merged whatever the values of m_l.

In the case where E_p derived from the $1/r$ Coulomb potential , this radial equation leads us to a normalizing solution of $R(r)$ (different from zero then...) only for values of the energy corresponding to the following quantization law (well ... what a coincidence, we fall back on the expression proved in the old models of Corpuscular Quantum Physics at page 4632!):

$$E_n = -hcR_H\frac{Z^2}{n^2} \tag{13.1.199}$$

where R_H is the Rydberg constant as we determined in the section of Corpuscular Quantum Physics. Thus, in this case the energy levels corresponding to the same values of n are all merged regardless of the value of l.

Remember that from the beginning we assume a rigid rotator with $R(r) = c^{te}$. And what if we consider now the non-rigid version?

13.1.3.2 Non-Rigid Rotator

Let us remind that we started from:

$$\left(-\frac{\hbar}{2\mu}\nabla^2 + E_p(r)\right)\Psi(\vec{r}) = E\Psi(\vec{r}) \tag{13.1.200}$$

Or more explicitly:

$$\left(-\frac{\hbar}{2\mu}\nabla^2 - \frac{Ze^2}{4\pi\varepsilon_0 r}\right)\Psi(\vec{r}) = E\Psi(\vec{r}) \tag{13.1.201}$$

Therefore in polar coordinates:

$$-\frac{\hbar^2}{2\mu}\left[\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\Psi}{\partial r}\right) + \frac{1}{r^2\sin(\theta)}\frac{\partial}{\partial\theta}\left(\sin(\theta)\frac{\partial\Psi}{\partial\theta}\right) + \frac{1}{r^2\sin^2(\theta)}\frac{\partial^2\psi}{\partial\phi^2}\right] - \frac{Ze^2}{4\pi\varepsilon_0 r}\Psi = E\Psi \tag{13.1.202}$$

Both LHS and RHS contain a term linear in Ψ, so combine:

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\Psi}{\partial r}\right) + \frac{1}{r^2\sin(\theta)}\frac{\partial}{\partial\theta}\left(\sin(\theta)\frac{\partial\Psi}{\partial\theta}\right) + \frac{1}{r^2\sin^2(\theta)}\frac{\partial^2\Psi}{\partial\phi^2} + \frac{2\mu}{\hbar^2}\left(E + \frac{Ze^2}{4\pi\varepsilon_0 r}\right)\Psi = 0 \tag{13.1.203}$$

Using the separation of variables idea, we assume a product solution of a radial and an angular function:

$$\Psi(r,\theta,\phi)=R(r)Y(\theta,\phi) \tag{13.1.204}$$

Since Y does not depend on r, we can move it in front of the radial derivative:

$$\frac{\partial\Psi}{\partial r}=\frac{\partial}{\partial r}RY=Y\frac{\mathrm{d}R}{\mathrm{d}r} \tag{13.1.205}$$

and, similarly, R does not depend on the angular variables. Thus replace Ψ and the differentials:

$$\frac{Y}{r^2}\frac{\mathrm{d}}{\mathrm{d}r}\left(r^2\frac{\mathrm{d}R}{\mathrm{d}r}\right)+\frac{R}{r^2\sin(\theta)}\frac{\partial}{\partial\theta}\left(\sin(\theta)\frac{\partial Y}{\partial\theta}\right)+\frac{R}{r^2\sin^2(\theta)}\frac{\partial^2 Y}{\partial\phi^2}+\frac{2\mu}{\hbar^2}\left(E+\frac{Ze^2}{4\pi\varepsilon_0 r}\right)RY=0 \tag{13.1.206}$$

We multiply by r^2 and divide by RY to separate the radial and angular terms:

$$\frac{1}{R}\frac{\mathrm{d}}{\mathrm{d}r}\left(r^2\frac{\mathrm{d}R}{\mathrm{d}r}\right)+\frac{1}{Y\sin(\theta)}\frac{\partial}{\partial\theta}\left(\sin(\theta)\frac{\partial Y}{\partial\theta}\right)+\frac{1}{Y\sin^2(\theta)}\frac{\partial^2 Y}{\partial\phi^2}+\frac{2\mu r^2}{\hbar^2}\left(E+\frac{Ze^2}{4\pi\varepsilon_0 r}\right)=0 \tag{13.1.207}$$

The first and fourth terms depend on r only, the middle terms depend on the angles only. They can only balance each other for all points in space if the radial and angular terms are the same constant but with opposite sign.

Therefore, we can separate into a radial equation:

$$\frac{\mathrm{d}}{\mathrm{d}r}\left(r^2\frac{\mathrm{d}R}{\mathrm{d}r}\right)+\frac{2\mu r^2}{\hbar^2}\left(E+\frac{Ze^2}{4\pi\varepsilon_0 r}\right)R-AR=0 \tag{13.1.208}$$

and an angular equation:

$$\frac{1}{\sin(\theta)}\frac{\partial}{\partial\theta}\left(\sin(\theta)\frac{\partial Y}{\partial\theta}\right)+\frac{1}{\sin^2(\theta)}\frac{\partial^2 Y}{\partial\phi^2}+AY=0 \tag{13.1.209}$$

where A is the separation constant.

The angular part still contains terms depending on both θ and ϕ. Another separation of variables is needed.

We'll replace Y by a product of single-variable functions:

$$Y(\theta,\phi)=\Theta(\theta)\Phi(\phi) \tag{13.1.210}$$

And we just see earlier what it took us to! However let us continue because we would like to know the value of A!

Replacing Y and the differentials, we have:

$$\frac{\Phi}{\sin(\theta)}\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\sin(\theta)\frac{\mathrm{d}\Theta}{\mathrm{d}\theta}\right)+\frac{\Theta}{\sin^2(\theta)}\frac{\mathrm{d}^2\Phi}{\mathrm{d}\phi^2}+A\Phi\Theta=0 \tag{13.1.211}$$

Isolating variables in separate terms:

$$\frac{\sin(\theta)}{\Theta}\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\sin(\theta)\frac{\mathrm{d}\Theta}{\mathrm{d}\theta}\right)+\frac{1}{\Phi}\frac{\mathrm{d}^2\Phi}{\mathrm{d}\phi^2}+A\sin^2(\theta)=0 \tag{13.1.212}$$

With B as separation constant, we have a polar part:

$$\frac{\sin(\theta)}{\Theta}\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\sin(\theta)\frac{\mathrm{d}\Theta}{\mathrm{d}\theta}\right)+A\sin^2(\theta)-B=0 \tag{13.1.213}$$

where B is another separation constant! Rearranging we get:

$$\frac{1}{\sin(\theta)}\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\sin(\theta)\frac{\mathrm{d}\Theta}{\mathrm{d}\theta}\right)+\left(A-\frac{B^2}{\sin^2(\theta)}\right)\Theta=0 \tag{13.1.214}$$

And here we recognize something known to us. We already proved earlier (see page 6140):

$$\frac{1}{\sin(\theta)}\left[\sin(\theta)\frac{\mathrm{d}\Theta}{\mathrm{d}\theta}\right]+\left[l(l+1)-\frac{m_l^2}{\sin^2(\theta)}\right]\Theta=0 \tag{13.1.215}$$

So we can immediately associate $A=l(l+1)$ and $B=m_l$. Great!

That means now we can write the radial equation as:

$$\frac{\mathrm{d}}{\mathrm{d}r}\left(r^2\frac{\mathrm{d}R}{\mathrm{d}r}\right)+\frac{2\mu r^2}{\hbar^2}\left(E+\frac{Ze^2}{4\pi\epsilon_0 r}\right)R-l(l+1)R=0 \tag{13.1.216}$$

We apply the product rule to the first term:

$$r^2\frac{\mathrm{d}^2R}{\mathrm{d}r^2}+2r\frac{\mathrm{d}R}{\mathrm{d}r}+\frac{2\mu r^2}{\hbar^2}\left(E+\frac{Ze^2}{4\pi\epsilon_0 r}\right)R-l(l+1)R=0 \tag{13.1.217}$$

and divide by r^2:

$$\frac{\mathrm{d}^2R}{\mathrm{d}r^2}+\frac{2}{r}\frac{\mathrm{d}R}{\mathrm{d}r}+\left(\frac{2\mu}{\hbar^2}\left(E+\frac{Ze^2}{4\pi\epsilon_0 r}\right)-\frac{l\,(l+1)}{r^2}\right)R=0 \tag{13.1.218}$$

We can't solve this straight away, but for very large r, the highlighted terms are forced to zero because they go reciprocal with r.

That leaves us with an asymptotic equation:

$$\frac{\mathrm{d}^2R_\infty}{\mathrm{d}r^2}+\frac{2\mu E}{\hbar^2}R_\infty=0 \tag{13.1.219}$$

which is another ordinary differential equation with constant coefficients. Having for solution (see section Differential and Integral Calculus page 1610):

$$R_\infty=C_1\exp\left(\mathrm{i}\sqrt{\frac{2\mu E}{\hbar^2}}r\right)+C_2\exp\left(-\mathrm{i}\sqrt{\frac{2\mu E}{\hbar^2}}r\right) \tag{13.1.220}$$

It makes sense to use as the zero point of potential energy the energy of a free electron, i.e. in this asymptotic case, $E\rightarrow 0$ for an electron far away from the nucleus, as it is practically free.

Since the presence of the positive charge in the nucleus stabilises the atom, we must look for solutions where E becomes negative as the electron comes closer to the nucleus. These two conditions are met if we choose $C_2 = 0$ and use the fact that $E < 0$ to get rid of the imaginary unit.

The asymptotic solution is then:

$$R_\infty = C_2 e^{-\sqrt{-\frac{2\mu E}{\hbar^2}}r} \tag{13.1.221}$$

For bound states, $R \to 0$ as $r \to +\infty$, and R is finite at the origin, $r = 0$. We do not consider continuum states with positive energy. Only when the latter are included do hydrogen wave functions form a complete set.

By use of the abbreviations (resulting from rescaling r to the dimensionless radial variable ρ):

$$\rho = \alpha r \quad \text{with} \quad \alpha^2 = -\frac{8\mu E}{\hbar^2}, \quad E < 0, \quad \lambda = \frac{\mu Z e^2}{2\pi\varepsilon_0 \alpha \hbar^2} \tag{13.1.222}$$

the equation:

$$\frac{\mathrm{d}^2 R}{\mathrm{d}r^2} + \frac{2}{r}\frac{\mathrm{d}R}{\mathrm{d}r} + \left(\frac{2\mu}{\hbar^2}\left(E + \frac{Ze^2}{4\pi\epsilon_0 r}\right) - \frac{l\,(l+1)}{r^2}\right) R = 0 \tag{13.1.223}$$

becomes:

$$\frac{1}{\rho^2}\frac{d}{d\rho}\left(\rho^2 \frac{d\chi(\rho)}{d\rho}\right) + \left(\frac{\lambda}{\rho} - \frac{1}{4} - \frac{l(l+1)}{\rho^2}\right)\chi(\rho) = 0 \tag{13.1.224}$$

where $\chi(\rho) = R(\rho/\alpha)$. A comparison with the differential equation introduced during our study of Laguerre polynomials in the section of Calculus page 1414 for $\Phi_n^k(x)$ (with a small change in the notation):

$$\frac{\mathrm{d}^2 \Phi_j^k(x)}{\mathrm{d}x^2} + \left(-\frac{1}{4} + \frac{2j+k+1}{2x} - \frac{k^2-1}{4x^2}\right)\Phi_j^k(x) = 0 \tag{13.1.225}$$

shows that the previous relation is satisfied by:

$$\rho\chi(\rho) = e^{-\rho/2}\rho^{l+1} L_{n-l-1}^{2l+1}(\rho) \tag{13.1.226}$$

in which k is replaced by $2l+1$ (as $(k^2-1)/4 = l(l+1)$) and j by $\lambda - l - 1$, upon using:

$$\frac{1}{\rho^2}\frac{\mathrm{d}}{\mathrm{d}\rho}\rho^2\frac{\mathrm{d}\chi}{d\rho} = \frac{1}{\rho}\frac{\mathrm{d}^2}{\mathrm{d}\rho^2}(\rho\chi) \tag{13.1.227}$$

We must restrict the parameter λ by requiring it to be an integer $n, n = 1, 2, 3, \ldots$. This is necessary because the Laguerre function of non-integral n would diverge as $\rho^n e^\rho$, which is unacceptable for our physical problem, in which:

$$\lim_{r \to +\infty} R(r) = 0 \tag{13.1.228}$$

This restriction on λ, imposed by our boundary condition, has the effect of quantizing the energy!

The comparison of the two ordinary differential equations above gives us the eigenenergies of the hydrogen atom, but require some development. Since $k=2l+1$:

$$\frac{2j+k+1}{2}=\frac{2j+(2l+1)+1}{2}=j+l+1 \tag{13.1.229}$$

From the discussion on associated Laguerre polynomials, the indices j and k are non-negative. The sum $j+l+1$ can, therefore, assume any integer values of 1 or greater. We are going to rename it n, or:

$$n:=j+l+1 \tag{13.1.230}$$

The new integer index n is known as the "principal quantum number". Using the principal quantum number, we can deduce the eigenenergies of the hydrogen atom. Indeed we have comparing again the two ordinary differential equations:

$$\lambda=\frac{\mu Ze^2}{2\pi\varepsilon_0\alpha\hbar^2}=\frac{2j+k+1}{2}=j+l+1=n \tag{13.1.231}$$

Squaring:

$$\left(\frac{\mu Ze^2}{2\pi\varepsilon_0\alpha\hbar^2}\right)^2=n^2 \tag{13.1.232}$$

and simplifying plus rearranging by injecting again α leads us to:

$$E_n=-\frac{Z^2\mu}{2n^2\hbar^2}\left(\frac{e^2}{4\pi\varepsilon_0}\right)^2=-\frac{Z^2\hbar^2}{2\mu n^2}\left(\frac{\mu e^2}{4\pi\varepsilon_0\hbar^2}\right)^2=-\frac{Z^2\hbar^2}{2\mu n^2r_0^2}=\frac{Z^2\cdot 13.6\,[\text{eV}]}{n^2} \tag{13.1.233}$$

where r_0 is the Bohr radius given for reminder by:

$$r_0=\frac{4\pi\varepsilon_0\hbar^2}{\mu e^2} \tag{13.1.234}$$

Now coming back to:

$$\rho\chi(\rho)=e^{-\rho/2}\rho^{l+1}L_{n-l-1}^{2l+1}(\rho) \tag{13.1.235}$$

Let us explicit it:

$$\begin{aligned}\sqrt{-\frac{8\mu E_n}{\hbar^2}}rR_{n,l}(r)&=e^{-\alpha r/2}(\alpha r)^{l+1}L_{n-l-1}^{2l+1}(\alpha r)\\&=e^{-\sqrt{-\frac{8\mu E_n}{\hbar^2}}r/2}\left(\sqrt{-\frac{8\mu E_n}{\hbar^2}}r\right)^{l+1}L_{n-l-1}^{2l+1}\left(\sqrt{-\frac{8\mu E_n}{\hbar^2}}r\right)\end{aligned} \tag{13.1.236}$$

Using:

$$E_n=-\frac{Z^2\hbar^2}{2\mu n^2r_0^2}\Rightarrow\sqrt{-\frac{8\mu E_n}{\hbar^2}}=\sqrt{\frac{4Z^2}{n^2r_0^2}}=\frac{2Z}{nr_0} \tag{13.1.237}$$

Therefore:

$$\left(\frac{2Z}{nr_0}\right)rR_{nl}(r)=e^{-2Zr/2nr_0}\left(\frac{2Z}{nr_0}r\right)^{l+1}L_{n-l-1}^{2l+1}\left(\frac{2Z}{nr_0}r\right) \tag{13.1.238}$$

After simplifying a bit and rearranging:

$$R_{nl}(r)=\frac{nr_0}{2Zr}e^{-Zr/nr_0}\left(\frac{2Z}{nr_0}r\right)^{l+1}L_{n-l-1}^{2l+1}\left(\frac{2Z}{nr_0}r\right)=e^{-Zr/nr_0}\left(\frac{2Z}{nr_0}r\right)^{l}L_{n-l-1}^{2l+1}\left(\frac{2Z}{nr_0}r\right) \tag{13.1.239}$$

Finally adding a normalization constant[4]:

$$R_{nl}(r)=A\left(\frac{2Z}{nr_0}r\right)^{l}L_{n-l-1}^{2l+1}\left(\frac{2Z}{nr_0}r\right)e^{-Zr/nr_0} \tag{13.1.240}$$

For a given value of the principal quantum number n (recall that we saw in the section of Corpuscular Quantum Physics at page 4667 that $l\leq n-1$), it is possible to verify that there are several solutions to the function $R(r)$ according to the value of the azimuthal quantum number l. Hence the identification of the solutions by the pair (n,l). We denote them $R_{n,l}(r)$. These are real functions of the variable r (it is just enough to check ... because if they work then they satisfy the Schrödinger equation, we will make an example a little further below):

$$\begin{aligned}
R_{10}(r)&=\left(\frac{Z}{a_0}\right)^{3/2}2e^{-(Zr/a_0)}\\
R_{20}(r)&=\left(\frac{Z}{a_0}\right)^{3/2}\frac{1}{2\sqrt{2}}\left(2-\frac{Zr}{a_0}\right)e^{-(Zr/2a_0)}\\
R_{21}(r)&=\left(\frac{Z}{a_0}\right)^{3/2}\frac{1}{2\sqrt{6}}\frac{Zr}{a_0}e^{-(Zr/2a_0)}\\
R_{30}(r)&=\left(\frac{Z}{a_0}\right)^{3/2}\frac{1}{9\sqrt{3}}\left(6-\frac{4Zr}{a_0}+\frac{4Z^2r^2}{9a_0^2}\right)e^{-(Zr/3a_0)}\\
R_{31}(r)&=\left(\frac{Z}{a_0}\right)^{3/2}\frac{1}{9\sqrt{6}}\left(\frac{8Zr}{3a_0}+\frac{4Z^2r^2}{9a_0^2}\right)e^{-(Zr/3a_0)}\\
R_{32}(r)&=\left(\frac{Z}{a_0}\right)^{3/2}\frac{1}{9\sqrt{30}}\left(\frac{4Z^2r^2}{9a_0^2}\right)e^{-(Zr/3a_0)}
\end{aligned} \tag{13.1.241}$$

where (beware some books give this value in natural units!):

$$a_0=4\pi\varepsilon_0\frac{\hbar^2}{\mu e^2}\qquad\text{or}\qquad\frac{\hbar^2}{\mu e^2} \tag{13.1.242}$$

is as we know the equivalent of the Bohr radius (for the reduced mass) that we have determined in the section of Corpuscular Quantum Physics with the difference that here we have a reduced mass instead of a single mass.

---

[4] Derive or even check that the normalization constant is the right one takes between five to ten pages so we will do that in the future when we will have the time to write a lot of LaTeX...

However let us see if our Schrödinger equation is satisfied (taking $n=1, l=0$ for example):

$$\begin{aligned}
&\left[T_r+\frac{l(l+1)\hbar^2}{2\mu r}+E_p\right]R(r)=E_1R(r)\\
&\left[-\frac{\hbar^2}{2\mu r^2}\partial_r(r^2\partial_r)-\frac{Ze^2}{4\pi\varepsilon_0 r}\right]\left(\frac{Z}{r_0}\right)^{3/2}2e^{-(Zr/r_0)}=E_1\left(\frac{Z}{r_0}\right)^{3/2}2e^{-(Zr/r_0)}\\
&-\frac{\hbar^2}{2\mu r^2}\partial_r\left(r^2\partial_r e^{-(Zr/r_0)}\right)-\frac{Ze^2}{4\pi\varepsilon_0 r}e^{-(Zr/r_0)}=E_1e^{-(Zr/r_0)}\\
&-\frac{\hbar^2}{2\mu r^2}\partial_r\left(-\frac{Zr^2}{r_0}e^{-(Zr/r_0)}\right)-\frac{Ze^2}{4\pi\varepsilon_0 r}e^{-(Zr/r_0)}=E_1e^{-(Zr/r_0)}\\
&-\frac{\hbar^2}{2\mu r^2}\left(-\frac{2Zr}{r_0}+\frac{Z^2r^2}{r_0^2}\right)e^{-(Zr/r_0)}-\frac{Ze^2}{4\pi\varepsilon_0 r}e^{-(Zr/r_0)}=E_1e^{-(Zr/r_0)}\\
&-\frac{\hbar^2}{2\mu r^2}\left(-\frac{2Zr}{4\pi\varepsilon_0\frac{\hbar^2}{\mu e^2}}+\frac{Z^2r^2}{\left(4\pi\varepsilon_0\frac{\hbar^2}{\mu e^2}\right)^2}\right)-\frac{Ze^2}{4\pi\varepsilon_0 r}=E_1\\
&\frac{Ze^2}{4\pi\varepsilon_0 r}-\frac{\hbar^2Z^2}{2\mu\left(4\pi\varepsilon_0\frac{\hbar^2}{\mu e^2}\right)^2}-\frac{Ze^2}{4\pi\varepsilon_0 r}=E_1\\
&-\frac{1}{(4\pi\varepsilon_0)^2}\frac{Z^2\mu e^2}{2\hbar^2}=-\frac{1}{16\pi^2\varepsilon_0^2}\frac{Z^2\mu e^4}{\frac{2h^2}{4\pi^2}}=-Z^2\frac{\mu e^4}{8\varepsilon_0^2h^2}=-hcR_HZ^2=E_1
\end{aligned}\tag{13.1.243}$$

Which corresponds well to the expected result!

Which graphically gives us the radial part $R_{n,l}(r)$:

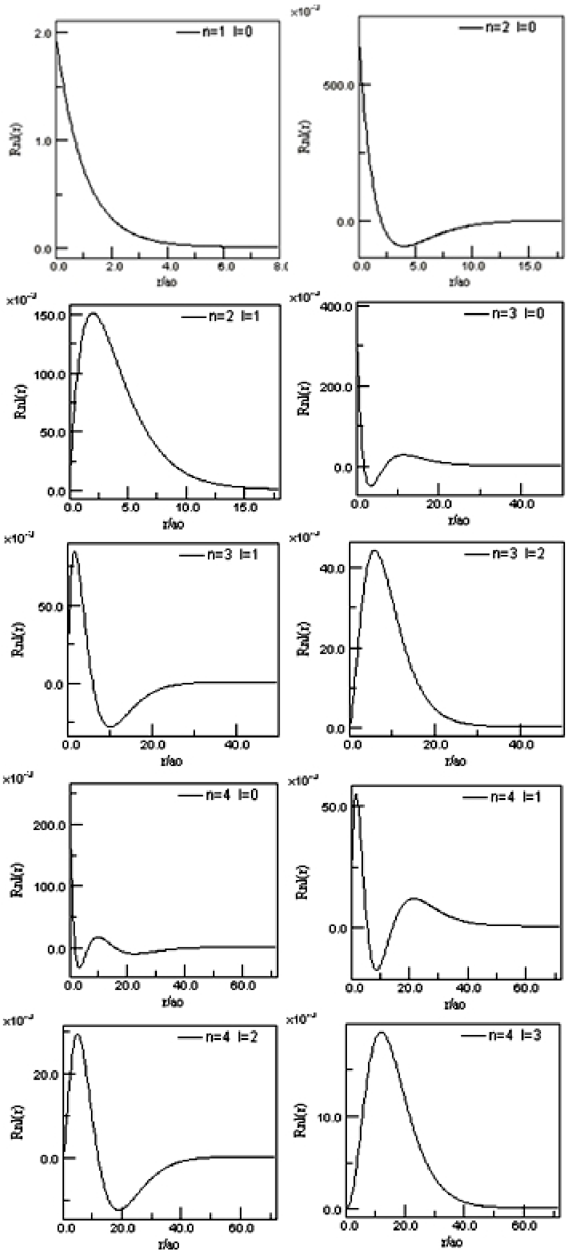


Figure 13.15: **Plot of a few radial functions** $R_{n,l}(r)$

Let us study a little more in detail the radial function in the case of the hydrogen atom!:

In the case of the atomic orbital $1s$ (!special case but we could do the same calculations as the following with all other orbital!) we have for the hydrogen atom:

$$R_{10}(r)=\left(\frac{Z}{r_0}\right)^{3/2}2e^{-Zr/r_0}=\left(\frac{1}{r_0}\right)^{3/2}2e^{-r/r_0}=\frac{2}{r_0^{3/2}}e^{-r/r_0} \quad (13.1.244)$$

where r_0 is the Bohr radius given for reminder by:

$$r_0=\frac{4\pi\varepsilon_0\hbar^2}{\mu e^2} \quad (13.1.245)$$

and:

$$\mu = \frac{m_e M}{M_{\text{tot}}} \tag{13.1.246}$$

is its reduced mass.

Notice there is a non-null probability for the electron to be at the center of atom ($r=0$), as we have:

$$R_{10}(0) = \frac{2}{r_0^{3/2}} = \frac{2}{\left(\dfrac{4\pi\varepsilon_0\hbar^2 M_{\text{tot}}}{m_e M e^2}\right)^{3/2}} \neq 0 \tag{13.1.247}$$

So it is well a decreasing exponential function as shown in the graphic above. Before continuing let us recall that (see section Wave Quantum Physics page 4691) :

$$\int\limits_{\infty} |\Psi|^2 \mathrm{d}V = \int\limits_{\infty} |\Psi|^2 \mathrm{d}x\mathrm{d}y\mathrm{d}z = 1 \tag{13.1.248}$$

But, in spherical coordinates (see the beginning of this section):

$$\mathrm{d}x\mathrm{d}y\mathrm{d}z = r^2 \mathrm{d}r \sin(\theta)\mathrm{d}\theta\mathrm{d}\phi \tag{13.1.249}$$

It then comes as we have seen earlier above:

$$\begin{aligned}\int\limits_{\infty} |\Psi|^2 \mathrm{d}x\mathrm{d}y\mathrm{d}z &= \int\limits_{\infty} |\Psi(r,\theta,\phi)|^2 r^2 \mathrm{d}r \sin(\theta)\mathrm{d}\theta\mathrm{d}\phi = \int\limits_{\infty} |R(r)Y(\theta,\phi)|^2 r^2 \mathrm{d}r \sin(\theta)\mathrm{d}\theta\mathrm{d}\phi\\ &= \int\limits_{\infty} |R(r)|^2 |Y(\theta,\phi)|^2 r^2 \mathrm{d}r \sin(\theta)\mathrm{d}\theta\mathrm{d}\phi\\ &= \int\limits_{\infty} |R(r)|^2 r^2 \mathrm{d}r \int\limits_{\infty} |Y(\theta,\phi)|^2 \sin(\theta)\mathrm{d}\theta\mathrm{d}\phi = 1\end{aligned} \tag{13.1.250}$$

Then it follows that:

$$\int\limits_{\infty} |R(r)|^2 r^2 \mathrm{d}r = 1 \quad \text{and} \quad \int\limits_{\infty} |Y(\theta,\phi)|^2 \sin(\theta)\mathrm{d}\theta\mathrm{d}\phi = 1 \tag{13.1.251}$$

With this result we can calculate the radial probability of finding the electron on each atomic orbital! So, it comes immediately with the previous result:

$$\frac{\mathrm{d}P_r}{\mathrm{d}r} = |R(r)|^2 r^2 \tag{13.1.252}$$

So in the case of our $1s$ atomic orbital:

$$\frac{\mathrm{d}P_r}{\mathrm{d}r} = \left(\frac{1}{r_0^{3/2}} e^{-(r/r_0)}\right)^2 r^2 = \frac{4}{r_0^3} e^{-(2r/r_0)} r^2 \tag{13.1.253}$$

It is now super interesting to calculate the point r point where the probability of finding the electron is maximum on the $1s$ orbital!

For this, we notice that $\mathrm{d}P_r/\mathrm{d}r$ reaches a maximum when we have trivially:

$$\frac{\mathrm{d}^2 P_r}{\mathrm{d}r^2} = 0 \tag{13.1.254}$$

Therefore:

$$\frac{\mathrm{d}^2 P_r}{\mathrm{d}r^2} = \frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{4}{r_0^3}e^{-(2r/r_0)}r^2\right) = \frac{4}{r_0^3}e^{-(2r/r_0)}\left(-\frac{2r^2}{r_0}+2r\right) = 0 \tag{13.1.255}$$

Hence:

$$r = r_0 \tag{13.1.256}$$

Which is remarkable, because we fall back on the result of the Bohr model (see section Corpuscular Physics page 4626)!!!

Remark

Notice that if we had kept $Z \neq 1$ then the result would have been of the shape: $r = r_0/Z$. That means more we have electrons, more these latter can be near the nucleus. We then say that the electrons screen each other somewhat from the nucleus.

To summarize a little all this, the stationary states of the hydrogen atom are specified by three quantum numbers $n \in \mathbb{N}^*, l \leq n-1, |m_l| \leq l$ and the Schrödinger wave function given finally by:

$$\boxed{\Psi_{nlm_l}(r,\theta,\phi) = R_{nl}(r)Y_{m_l l}(\theta,\phi) = \left[\left(\frac{2Z}{nr_0}\right)^3 \frac{(n-l-1)!}{2n(n+l)!}\right]^{1/2} e^{-Zr/nr_0}\left(\frac{2Z}{nr_0}r\right)^l L_{n-l-1}^{2l+1}\left(\frac{2Z}{nr_0}r\right)Y_{m_l,l}(\theta,\phi)} \tag{13.1.257}$$

We then have the following traditional nomenclature in the case of the hydrogen atom:

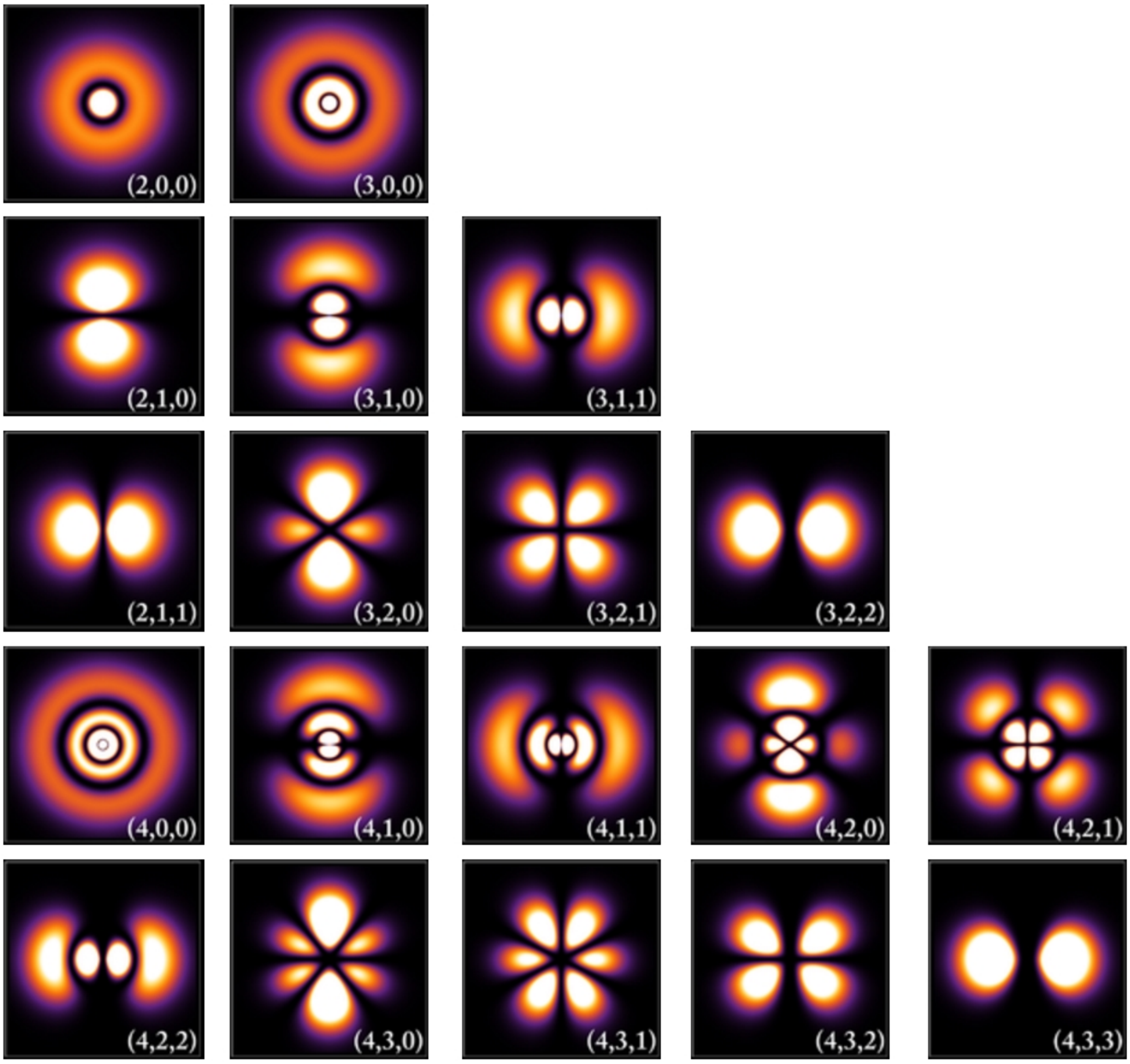
n	l	m_l	Function	Nomenclature
1	0	0	$\Psi_{1,0,0}$	$1s$
2	0	0	$\Psi_{2,0,0}$	$2s$
	1	1	$\Psi_{2,1,1}$	$2p_1$
		0	$\Psi_{2,1,0}$	$2p_0$
		-1	$\Psi_{2,1,-1}$	$2p_{-1}$
3	0	0	$\Psi_{3,0,0}$	$3s$
	1	1	$\Psi_{3,1,1}$	$3p_1$
		-1	$\Psi_{3,1,-1}$	$3p_{-1}$
	2	2	$\Psi_{3,2,2}$	$3d_2$
		1	$\Psi_{3,2,1}$	$3d_1$
		0	$\Psi_{3,2,0}$	$3d_0$
		-1	$\Psi_{3,2,-1}$	$3d_{-1}$
		-2	$\Psi_{3,2,-2}$	$3d_{-2}$

Table 13.1: **Nomenclature of layers and sub-layers of the hydrogen atom**

We can include the spin of the electron in the description of the electronic structure of the atom. If we treat the spin as an additional degree of freedom then the lack of interaction term between conventional degrees of freedom (positions in real space) and the spin interaction named "spin-orbit coupling" in the previous Hamiltonian implies that we can write the total wave function, spin included, in the form of a product:

$$\Psi_{n,l,m_l,m_s}(r,\theta,\phi)=\Psi_{n,l,m_l}(r,\theta,\phi)\omega_{m_s}=R_{n,l}(r)Y_{m_l,l}(\theta,\phi)\omega_{m_s} \qquad (13.1.258)$$

So taking into account everything seen so far we have the following density plots:



where we added the spin quantum number $m_s=\pm 1/2$ (see section Corpuscular Quantum Physics page 4669).

The same remark we made in the section of Corpuscular Quantum Physics then applies: the levels remain $2n^2$ times degenerated.

Let us do an example! So we have:

$$\Psi_{1,0,0}=R_{1,0}Y_{0,0}=\left(\frac{Z}{r_0}\right)^{3/2}e^{-Zr/r_0}\frac{1}{4\sqrt{2}}=\frac{1}{\sqrt{\pi}}\left(\frac{Z}{r_0}\right)^{3/2}e^{-Zr/r_0} \tag{13.1.259}$$

Thus for 1 proton:

$$\Psi_{1,0,0}=\frac{1}{\sqrt{\pi}}\left(\frac{1}{r_0}\right)^{3/2}e^{-r/r_0}=\frac{1}{\sqrt{\pi r_0^3}}e^{-r/r_0} \tag{13.1.260}$$

This wave function is often named the "wave function of hydrogen" or the "ground state of the hydrogen atom" and denoted Ψ_{1s}. This function says that the hydrogen atom's electron is in a weighted superposition of all possible distances r from the nucleus.

Now let us apply the 5th postulate of wave quantum physics (see section of the same name page 4699) for unlike earlier, not calculate the modal radius (most likely one), but the average radius! Then, as the operator position is the position itself (see section Wave Quantum Physics page 4695), the average value of the radius will be given by (do not forget that we are in spherical coordinates!):

$$\begin{aligned}\langle r\rangle&=\int\bar{\Psi}_{1,0,0}\Psi_{1,0,0}\mathrm{d}V=\int\frac{1}{\sqrt{\pi}}\left(\frac{1}{r_0}\right)^{3/2}e^{-r/r_0}r\frac{1}{\sqrt{\pi}}\left(\frac{1}{r_0}\right)^{3/2}e^{-r/r_0}\mathrm{d}V\\&=\int\frac{1}{\pi r_0^3}e^{-2r/r_0}r\mathrm{d}V=\frac{1}{\pi r_0^3}\int\limits_0^{+\infty}\int\limits_0^{+\pi}\int\limits_0^{2\pi}e^{-2r/r_0}rr^2\sin(\theta)\mathrm{d}r\mathrm{d}\theta\mathrm{d}\phi\\&=\frac{1}{\pi r_0^3}\int\limits_0^{+\infty}\int\limits_0^{+\pi}\int\limits_0^{2\pi}r^3e^{-2r/r_0}\sin(\theta)\mathrm{d}r\mathrm{d}\theta\mathrm{d}\phi\end{aligned} \tag{13.1.261}$$

Using the Fubini theorem proved in the section of Differential and Integral Calculus at page 1521, we can write (well in this case it's even trivial that we have the right to write this... we should not even have to mention Fubini theorem normally...):

$$\begin{aligned}\langle r\rangle&=\frac{1}{\pi r_0^3}\int\limits_0^{+\infty}\int\limits_0^{+\pi}\int\limits_0^{2\pi}r^3e^{-2r/r_0}\sin(\theta)\mathrm{d}r\mathrm{d}\theta\mathrm{d}\phi=\frac{1}{\pi r_0^3}\int\limits_0^{+\infty}r^3e^{-2r/a0}\mathrm{d}r\int\limits_0^{\pi}\mathrm{d}\phi\int\limits_0^{2\pi}\sin(\theta)\mathrm{d}\theta\\&=\frac{1}{\pi r_0^3}\int\limits_0^{+\infty}r^3e^{-2r/a0}\mathrm{d}r\pi2\int\limits_0^{\pi}\sin(\theta)\mathrm{d}\theta=\frac{1}{\pi r_0^3}\int\limits_0^{+\infty}r^3e^{-2r/a0}\mathrm{d}r\pi2\cdot2\\&=\frac{4}{r_0^3}\int\limits_0^{+\infty}r^3e^{-2r/a0}\mathrm{d}r\end{aligned} \tag{13.1.262}$$

For the last integral, we will use integration by parts:

$$\begin{aligned}\int\limits_0^{+\infty} r^3e^{-br}\mathrm{d}r &= \underbrace{-\frac{1}{b}r^3e^{-br}\Big|_0^{+\infty}}_{=0}+\frac{3}{b}\int\limits_0^{+\infty} r^3e^{-br}\mathrm{d}r\\ &=\frac{3}{b}\left(\underbrace{-\frac{1}{b}r^2e^{-br}\Big|_0^{+\infty}}_{=0}+\frac{2}{b}\int\limits_0^{+\infty} re^{-br}\mathrm{d}r\right)=\frac{6}{b^2}\int\limits_0^{+\infty} re^{-br}\mathrm{d}r\\ &=\frac{6}{b^2}\left(\underbrace{-\frac{1}{b}re^{-br}\Big|_0^{+\infty}}_{=0}+\frac{1}{b}\int\limits_0^{+\infty} e^{-br}\mathrm{d}r\right)=\frac{6}{b^3}\int\limits_0^{+\infty} e^{-br}\mathrm{d}r=-\frac{6}{b^4}e^{-br}\Big|_0^{+\infty}\\ &=-\frac{6}{b^4}\left(\underbrace{e^{-b\infty}}_{\to 0}-e^{-b0}\right)=-\frac{6}{b^4}(-1)=\frac{6}{b^4}\end{aligned}\tag{13.1.263}$$

Thus finally:

$$\langle r\rangle=\frac{4}{r_0^3}\int\limits_0^{+\infty} r^3e^{-2r/a0}\mathrm{d}r=\frac{4}{r_0^3}\frac{6}{\left(\dfrac{2}{r_0}\right)^4}=\frac{6}{2^2}r_0=\frac{3}{2}r_0\tag{13.1.264}$$

Or more explicitly:

$$\langle r\rangle_{1,0,0}=\langle r\rangle_{1s}=\frac{3}{2}r_0\tag{13.1.265}$$

Therefore the average distance of the electron to the core is equal to $3/2$ times that of the Bohr radius so further that the most likely radius we have calculated earlier above (and which corresponds to Bohr radius)!

Let us calculate now the modal value of the radius for the hydrogen wave function:

$$\Psi_{1s}=\frac{1}{\sqrt{\pi r_0^3}}e^{-r/r_0}\tag{13.1.266}$$

The probability density is therefore:

$$\|\Psi_{1s}\|^2=\frac{1}{\pi r_0^3}e^{-2r/r_0}\tag{13.1.267}$$

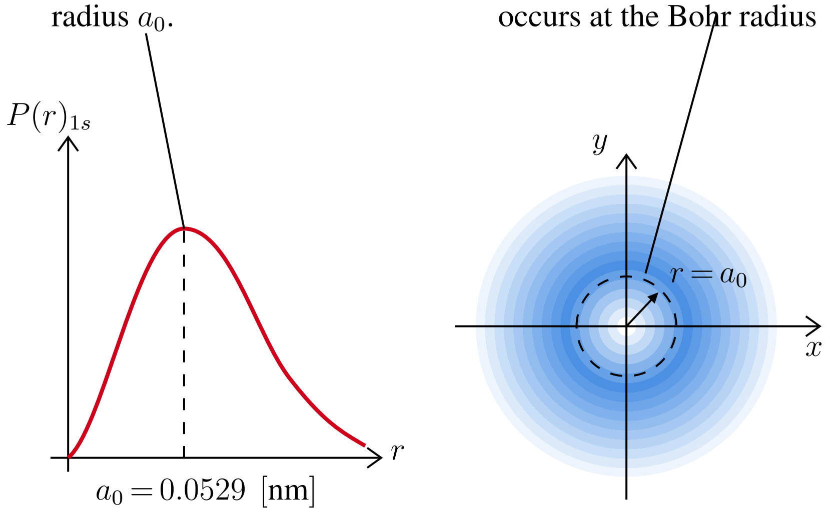
Because of the spherical symmetric of Ψ_{1s} we can write it's elementary probability as:

$$P(r)_{1s}\mathrm{d}r=\|\Psi_{1s}\|^2\mathrm{d}V=\|\Psi_{1s}\|^2 4\pi r^2\mathrm{d}r\tag{13.1.268}$$

Therefore, the radial probability density function for an s state is:

$$P(r)_{1s}=\|\Psi_{1s}\|^2\mathrm{d}V=\|\Psi_{1s}\|^2 4\pi r^2\tag{13.1.269}$$

A plot of the function $P(r)_{1s}$ is presented below:



the probability has its maximum value when r equals the Bohr radius a_0.

in this representation, the darkest color, representing the maximum probability, occurs at the Bohr radius

Figure 13.16: Probability of finding the electron as a function of distance from the nucleus for the hydrogen atom in the $1s$ state (source: [491])

The peak of the curve corresponds to the most likely value of r for this particular state. We prove easily that this peak occurs at the Bohr radius (just set the derivative of $P(r)_{1s}$ with respect to r to zero), the radial position of the electron when the hydrogen atom is in its ground state in the Bohr theory, another agreement between the Bohr theory and the quantum theory!

Physicist Aneta Stodólna captured the electron positions of hydrogen atoms on camera for the very first time based on velocity map imaging (technique involving LASERs). The images made in into Physics Review Letters [516] in May 12013 (holocene calendar) and gained worldwide attention from physicists and the general public afterwards.

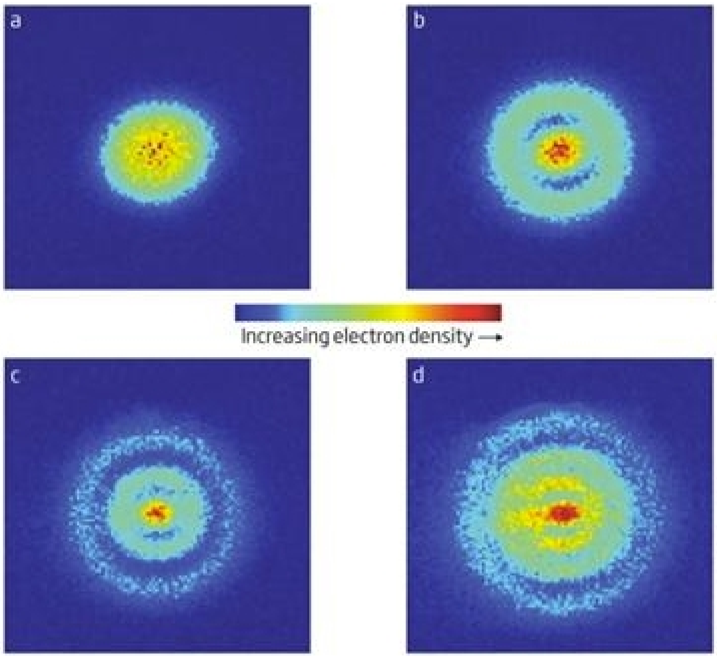


Figure 13.17: Hydrogen atoms imaged by A. Stodólna at four energy levels (increasing from a tod)

13.1.3.2.1 Potential Profile

Let us come back on an important point that is often used in physics book but never proved (as far as we know): the quantum potential profile of the hydrogen-like atom. Many books sometimes speak of "harmonic model of the atomic bonding" but it seems that this is a priori rather a misnomer.

So we saw much earlier in this section that:

$$\left[T_r+\frac{l(l+1)\hbar^2}{2\mu r^2}+E_p\right]\Psi=E\Psi \tag{13.1.270}$$

In view of the interpretation of the three terms of the Hamiltonian, it is customary to say that the two terms:

$$\frac{l(l+1)\hbar^2}{2\mu r^2}+E_p \tag{13.1.271}$$

constitute the "effective potential energy", thus explicitly:

$$E_{p,\text{eff}}=\frac{l(l+1)\hbar^2}{2\mu r^2}+E_p=\underbrace{\frac{l(l+1)\hbar^2}{2\mu r^2}}_{\geq 0}+\underbrace{-\frac{Zq_e^2}{4\pi\varepsilon_0\|\vec{r}_e-\vec{R}_N\|}}_{<0} \tag{13.1.272}$$

So the first term is (logically) repulsive while the second is attractive. A plot in Maple™ 4.00b of the effective potential energy gives with real experimental values for the radius with the real values of the constants:

```
>plot([-2.31E-28/r+6.11E-39*0*(0+1)/r^2,-2.31E-28/r+6.11E-39*1*(1+1)/r^2,
-2.31E-28/r+6.11E-39*2*(2+1)/r^2,-2.31E-28/r+6.11E-39*3*(3+1)/r^2,
-2.31E-28/r+6.11E-39*10*(10+1)/r^2],r=5E-11..10E-10,
y=-0.5E-17..0.5E-17,thickness=2);
```

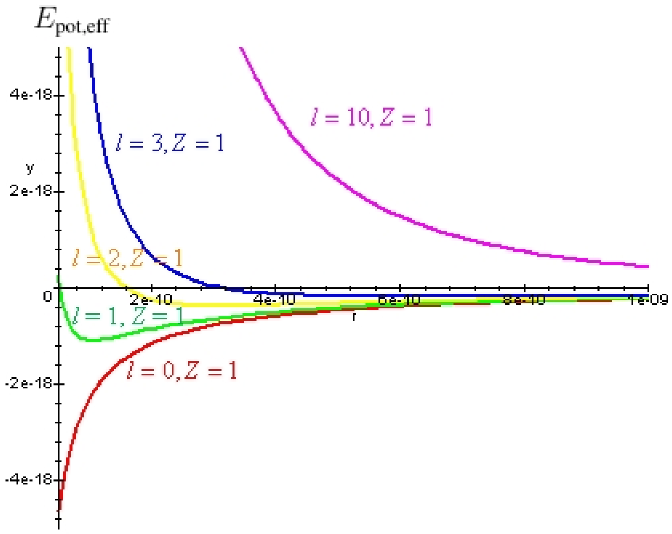


Figure 13.18: **Plot of the effective potential energy in Maple™ 4.00b for various l and Z**

where the legends were added afterwards with a text processor software. The reader will notice especially the case where $l = 1$ that matches to the case of the figure indicated by the majority of graduate books of physics. Either with a zoom:

```
>plot(-2.31E-28/r+6.11E-39*1*(1+1)/r^2,r=5E-11..10E-10,thickness=2,
color=green);
```

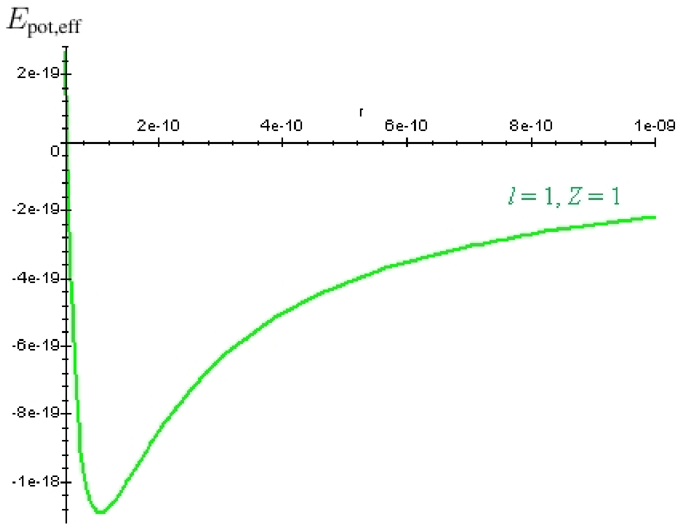


Figure 13.19: **Plot of the famous effective potential energy in Maple™ 4.00b for $l = 1$ and $Z = 1$**

The first graph also tells us quite clearly that for $l = 0$ the electron has a negative potential energy that firmly holds it in the orbit of the proton. By cons already at $l = 1$ we guess that the point of stability of the electron is where the derivative is zero. Beyond the $l = 1$, in the case of a nucleus with a single proton, the electron is not naturally linked anymore since its potential energy tends to be positive. The reader can also have fun with Maple™ by making vary Z and l. He will see that the effective potential energy is very sensitive to these parameters. For example, the plot below shows the effective potential energy with $l = 4$ and $Z = 1$ (thus unstable atom) and then with $l = 4$ and $Z = 6$ (which corresponds rather to an excited state):

```
>plot([-2.31E-28*1/r+6.11E-39*4*(4+1)/r^2,-2.31E-28*6/r+6.11E-39*4*(4+1)/r^2],
r=5E-11..10E-10,thickness=2);
```

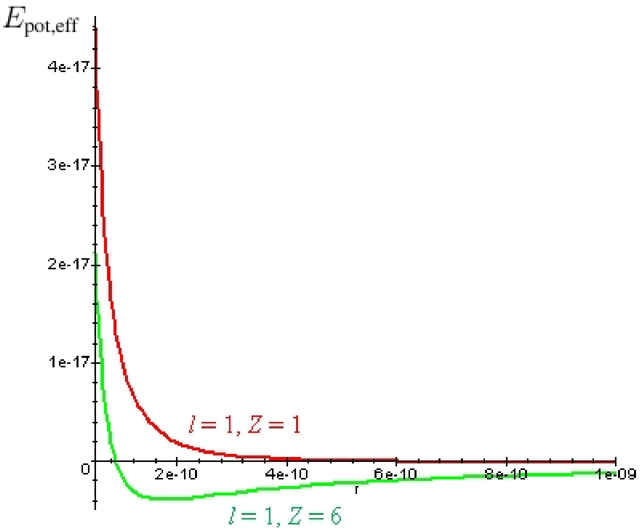


Figure 13.20: Plot of the effective potential energy for $l=1$ and various Z

It is customary in practice to consider that:

$$E_{p,\text{eff}}=\frac{l(l+1)\hbar^2}{2\mu r^2}-\frac{Zq_e^2}{4\pi\varepsilon_0\|\vec{r}_e-\vec{R}_N\|} \tag{13.1.273}$$

is to a given factor (electric charge factor) an "effective electrical potential" or "electric screened potential". Indeed by defining the electric potential (see section Electrostatics page 3976), there are only an electric charge factor ratio between the electric potential energy and the electric potential. So we have:

$$U_{\text{eff}}=\frac{1}{Zq}E_{p,\text{eff}}\Rightarrow U_{\text{eff}}\propto E_{p,\text{eff}} \tag{13.1.274}$$

~90 %

★★★☆☆

49 votes, 66.12%

13.2. Molecular Chemistry

MOLECULAR CHEMISTRY is the central area that interconnects thanks to the study of molecules many promising advanced technologies of the early 121st century (holocene calendar) which are to name only the best known: molecular biology, molecular materials, molecular electronics, polymers, etc.

Knowing it was found experimentally that a single molecule can have several very different functions, its theoretical study allows to use them better (sometimes better performance in terms of R&D) in its areas of application. The reader will therefore understand that, as usually in this textbook, we will focus here only on the theoretical aspect (mathematical) of molecular chemistry even if we limit ourselves only to theoretical developments made between the years 11910 (holocene calendar) and about 11935 according to holocene calendar (beyond the complexity of the modern theories require too many pages to a general book like this one).

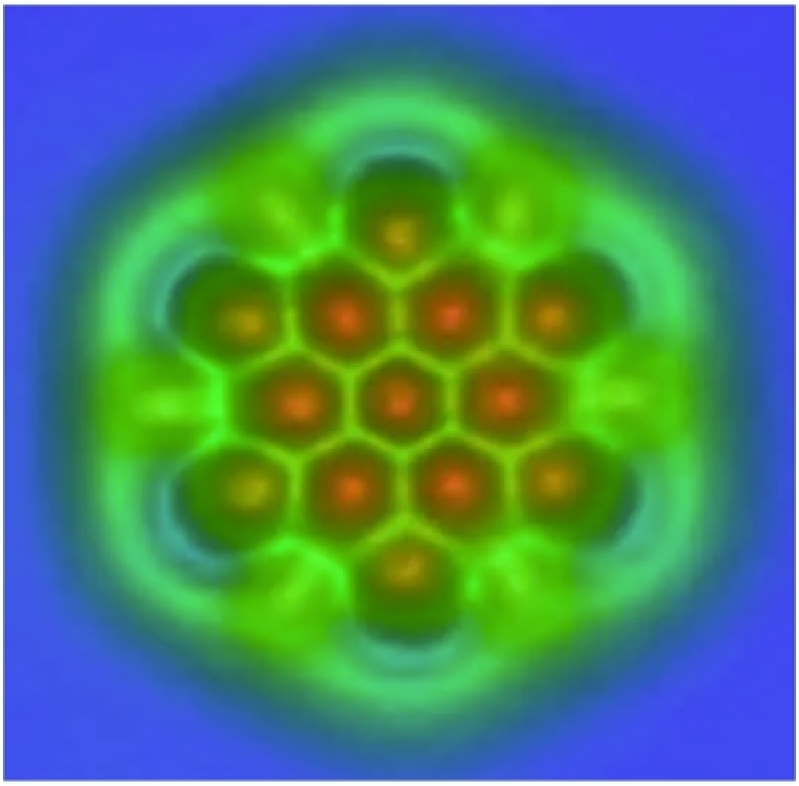


Figure 13.21: Force Atomic Microscope image of chemical bonds of a nanographene molecule C60 **(source: IBM Research Center Zurich)**

We are in the beginning of the 121st century (holocene calendar) at the infancy of the discovery of what nature has done with plenty of time and chance (probabilities): that is to say complex molecules working as nano-machines capable locally (active site) to filter, oxidize, to make catalysis ... and many other manipulations (there is just to observe your own body!).

A molecule is often treated in school classes with the Schrödinger equation (so no relativistic case and no consideration of the spins) in the usual form (see section Wave Quantum Physics page 4760):

$$H\Psi = E\Psi \tag{13.2.1}$$

or also in a stationary from (time-independent) where as a reminder Ψ is a eigenfunctions and E an eigenvalue of the application H.

In reality, the wave functions are impossible to calculate normally with contemporary mathematical tools and the only thing we can do are numerical calculations (perturbation method). This is why some chemistry centers are transformed over time into data centers where the predictive character (and inexpensive) of quantum chemistry is becoming more and more important.

It remains of course essential, as always, to understand how the theoretical models are built and their underlying assumptions.

But we can still thanks to calculations predict the form of reasonable size of molecules, the energy of their internal connections, their energy capacity under stress deformation, the shape of the molecular orbitals (M.O.), energy state transitions (when parts of the molecule move therein), their reactivity vis-a-vis of a reaction medium...

We commonly distinguish two cases of study of the molecular chemistry:

1. Quantum mechanics: all interactions between particles are taken into account under the assumption of some acceptable simplifications.
2. Molecular mechanics: For large molecules, we are not concerned anymore over the electronic problem, but the interaction of certain parameters on which we want to focus.

For example, haemoglobin (protein carrying oxygen carrying in the muscles) is a huge molecular structure which we will study only active site with the tools of quantum mechanics. The overall behaviour of the molecule itself is treated with the molecular mechanics tools.

It follows that excepts for hydrogen-like atoms, we can not analytically describe a molecule from a purely quantum point of view! All current quantum methods rely on one or more approximations. The wave functions are therefore approximated and the level of calculation is adjusted according to what we want to show and the precision that we seek (seeking to minimize the computation time for cost problems...). The good understanding of approximations permits to express simple models requiring only a minimum of calculations (often trivial).

We propose here to show two common models (and the most simplest):

13.2.1 Orbital Approximations

A molecule is obviously an extremely complex problem: N nuclei, n electrons and everything is moving!

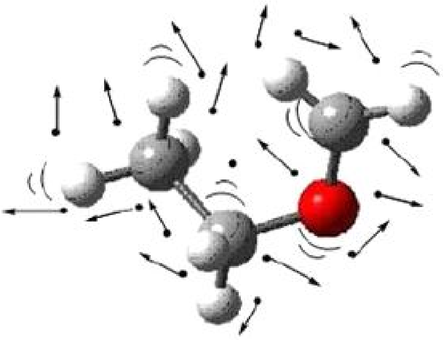


Figure 13.22: **Example ofa molecule where almost everything is moving**

The Hamiltonian (see section Wave Quantum Physics page 4694):

$$H=E_c+E_p=-\frac{\hbar^2}{2m}\vec{\nabla}^2+V(q_1,q_2,\ldots,t)=\sum_{k=1}^{n}\left(\frac{-\hbar^2}{2m_k}\frac{\partial}{\partial q_k^2}\right)+V(q_1,q_2,\ldots,t) \quad (13.2.2)$$

is then a nightmare but in the intuitive form (the subscript G of the Hamiltonian means "General") below:

$$H_{\mathrm{G}}=\underbrace{\left(-\sum_{k=1}^{N}\frac{\hbar^2}{2M_k}\vec{\nabla}_k^2-\sum_{i=1}^{n}\frac{\hbar^2}{2m_e}\vec{\nabla}_i^2\right)}_{E_{\mathrm{c}}}+\underbrace{\left(-\sum_{k=1}^{N}\sum_{i=1}^{n}\frac{Ze^2}{4\pi\varepsilon_0 r_{ik}}+\sum_{\substack{i=1\\j>1}}^{n-1}\frac{e^2}{4\pi\varepsilon_0 r_{ik}}+\sum_{\substack{k=1\\i>k}}^{N-1}\frac{Z_kZ_ie^2}{4\pi\varepsilon_0 r_{ik}}\right)}_{E_p} \quad (13.2.3)$$

where:

1. $-\sum_{k=1}^{N}\frac{\hbar^2}{2M_k}\vec{\nabla}_k^2$ is the kinetic energy of the k nuclei of mass M_k in the molecule.
2. $-\sum_{i=1}^{n}\frac{\hbar^2}{2m_e}\vec{\nabla}_i^2$ is the kinetic energy of the n electrons of mass m_e.
3. $-\sum_{k=1}^{N}\sum_{i=1}^{n}\frac{Ze^2}{4\pi\varepsilon_0 r_{ik}}$ is the potential energy due to the attraction electron($-$)/nucleus($+$).
4. $\sum_{\substack{i=1\\j>1}}^{n-1}\frac{e^2}{4\pi\varepsilon_0 r_{ik}}$ is the potential energy of the repulsion electron($-$)/electron($-$).

5. $\displaystyle\sum_{\substack{k=1\\i>k}}^{N-1}\frac{Z_kZ_ie^2}{4\pi\varepsilon_0 r_{ik}}$ is the potential energy of repulsion nucleus(+)/nucleus(+).

Often we find these terms in the following form of the Schrödinger equation in the literature:

$$[(T_e+T_n)+(V_{ee}+V_{en}+V_{nn})]\Psi_n\Psi_{el}=E\Psi_n\Psi_{\text{el}} \tag{13.2.4}$$

A first approximation we might try is to decouple the movement of the nuclei of the electrons. Indeed, as the nucleus is much more massive (about $2,000$ times) than the cloud of electrons, the center of mass is assimilated to the nucleus of the atom and all the motion to the entire electron cloud. This approximate approach is well known under the name "Born-Oppenheimer approximation":

$$\begin{aligned}H_{\text{el}}&=-\sum_{i=1}^{n}\frac{\hbar^2}{2m_e}\vec{\nabla}_i^2-\sum_{k=1}^{N}\sum_{i=1}^{n}\frac{Ze^2}{4\pi\varepsilon_0 r_{ik}}+\sum_{\substack{i=1\\j>1}}^{n-1}\frac{e^2}{4\pi\varepsilon_0 r_{ik}}\\&=-\sum_{i=1}^{n}\frac{\hbar^2}{2m_e}\vec{\nabla}_i^2+\frac{e^2}{4\pi\varepsilon_0}\left(-\sum_{k=1}^{N}\sum_{i=1}^{n}\frac{Z}{r_{ik}}+\sum_{\substack{i=1\\j>1}}^{n-1}\frac{1}{r_{ik}}\right)\end{aligned} \tag{13.2.5}$$

which then allows us to study the molecular orbitals. But unfortunately this approximation is not sufficient because of the repulsion inter-electronic term (the double sum) that prevents using the separation of variables technique as we did in the section of Quantum Chemistry with the hydrogenoid-atom.

Moreover, this latter equation is also written as the first line of the couple of equations below (Schrödinger equation of electrons and nuclei):

$$\underbrace{(T_e+V_{ee}+V_{en})}_{H_{\text{el}}}\Psi_{\text{el}}=E\Psi_{\text{el}} \tag{13.2.6a}$$

$$\underbrace{(T_n+V_{nn})}_{H_{\text{nuclei}}}\Psi_n\Psi_{\text{el}}=E\Psi_n\Psi_{\text{el}} \tag{13.2.6b}$$

This system of equations is what some name the "adiabatic approximation" (???).

The idea that then comes to mind will be using the following property:

Given two operators A and B, $f(u)$ and $g(v)$ their respective eigenfunctions associated with eigenvalues a and b. Then $f(u)g(v)$ is an eigenfunction of the operator $A+B$ with associated eigenvalue $a+b$.

Which is written:

$$\left.\begin{aligned}Af(u)&=af(u)\\Bg(v)&=bg(v)\end{aligned}\right\}\Rightarrow(A+B)f(u)g(v)=(a+b)f(u)g(v) \tag{13.2.7}$$

Proof 13.0.4. We have:

$$\begin{aligned}(A+B)f(u)g(v)&=[Af(u)]g(v)+[Bf(u)]g(v)\\&=[af(u)]g(v)+[bf(u)]g(v)=(a+b)f(u)g(v)\end{aligned} \tag{13.2.8}$$

■ Q.E.D.

And that's what we will use to break the n-electronic Hamiltonian H_{el} into a sum of independent-electron Hamiltonian knowing of the above that if we find the eigenfunction for each (which is relatively easier) it will be sufficient to simply multiply them to get the overall eigenfunction!

Thus, we write:

$$H_{\text{el}}=H_1+H_2+\ldots+H_i+\ldots+H_n \tag{13.2.9}$$

and therefore we have to find for each i:

$$H_i\Psi_i=E_i\Psi_i \tag{13.2.10}$$

Then we have:

$$H_{\text{el}}\Psi_{\text{el}}=\sum_{i=1}^{n}H_i\prod_{i=1}^{n}\Psi_i=\sum_{i=1}^{n}E_i\prod_{i=1}^{n}\Psi_i=E_{\text{el}}\Psi_{\text{el}} \tag{13.2.11}$$

with therefore:

$$\Psi_{\text{el}}=\Psi_1\cdot\Psi_2\cdot\ldots\cdot\Psi_i\cdot\ldots\cdot\Psi_n \tag{13.2.12}$$

This approach by one-electron Hamiltonian approach will lead us to replace:

$$H_{\text{el}}=-\sum_{i=1}^{n}\frac{\hbar^2}{2m_e}\vec{\nabla}_i^2-\sum_{k=1}^{N}\sum_{i=1}^{n}\frac{Ze^2}{4\pi\varepsilon_0 r_{ik}}+\sum_{\substack{i=1\\j>1}}^{n-1}\frac{e^2}{4\pi\varepsilon_0 r_{ij}} \tag{13.2.13}$$

by the sum of Hamiltonian for an electron named "effective Hamiltonian":

$$H_i=-\frac{\hbar^2}{2m_e}\vec{\nabla}_i^2-\frac{Ze^2}{4\pi\varepsilon_0 r_i}+\frac{e^2}{4\pi\varepsilon_0 r_i}=-\frac{\hbar^2}{2m_e}\vec{\nabla}_i^2-\frac{e^2(Z-1)}{4\pi\varepsilon_0 r_i}=-\frac{\hbar^2}{2m_e}\vec{\nabla}_i^2-\frac{Z^*e^2}{4\pi\varepsilon_0 r_i} \tag{13.2.14}$$

This approximation method is sometimes named in theoretical chemistry "independent electron approximation" or "orbital approximation". It consists therefore to include the electron-electron interactions and to write that each electron moves in an average potential resulting from the presence of all other electrons.

The "Slater method" consists by definition to write the latter relation in the form:

$$H_i\cong-\frac{\hbar^2}{2m_e}\vec{\nabla}_i^2-\frac{e^2(Z-\sigma)}{4\pi\varepsilon_0 r_i} \tag{13.2.15}$$

where σ is named the "screen constant".

The Slater method basically means replacing the purely electronic terms by a constant. It can be regarded as a parametric method since the constants were determined purely experimentally.

The principle of empirical calculation of the screening constant is relatively simple: In a poly-electronic atom, the core electrons are on much contracted orbits while the valence electrons that will be responsible for the chemical properties of the atom in question are on orbits much more "relaxed".

The attraction of the nucleus on the latter electrons is much lower than that exerted on the core electrons and these electrons only receive a portion of the atomic charge.

Slater then proposed that the effective charge, which is usually denoted by Z^* could be calculated by taking into account the screening constant. This constant represents then the average effect of the other electrons on the considered electron of the effective Hamiltonian i:

$$Z^* = Z - \sigma \tag{13.2.16}$$

For a peripheral electron, we will need to consider its screen constant is due to all electrons placed on orbits equal or below its own. The tradition (or rather the "trick") is that the calculation is done by combining atomic orbitals in several groups $1s/2s, 2p/3s, 3p/3d/4s, 4p/4d/4f/5s, 5p/$ etc.

Then the calculation is simple because it is based on an array of predefined values and we simply have to add the screening contributions of all the electrons following the table below:

	$n' < n-1$	$n' = n-1$	$n' = n$	$n' > n$
$1s$			0.30	0
ns, np	1	0.85	0.35	0
nd, nf	1	1	0.35	0

Table 13.2: **Screening contributions of electrons**

This table deserves some explanation of course !:

The index indicates the number of the group that contributes to the screening constant while n is the number of the group of electron that we consider.

✍**Example:**

In the case of the Carbon of configuration $1s^2 2s^2 2p^2$, the nuclear charge is $Z = 6$. One electron $1s$ is shielded by only the another $1s$ electron, the effective charge it sees is therefore:

$$Z^* = Z - \sigma = 6 - 0.3 = 5.7 \tag{13.2.17}$$

A $2s$ or $2p$ electron is shielded by the two $1s$ electrons and by the other 3 electrons $2s$ and $2p$. The effective charge by which it is attracted is then:

$$Z^* = Z - \sigma = 6 - (2 \cdot 0.85 + 3 \cdot 0.35) = 3.25 \tag{13.2.18}$$

So we see that the effective charge experienced decreases rather quickly!

13.2.2 LCAO Method

A linear combination of atomic orbitals or LCAO is a quantum superposition of atomic orbitals and a technique for calculating molecular orbitals in quantum chemistry. In quantum mechanics, electron configurations of atoms are described as wave functions. In mathematical sense, these wave functions are the basis set of functions, the basis functions, which describe the electrons of a given atom. In chemical reactions, orbital wave functions are modified, i.e. the electron cloud shape is changed, according to the type of atoms participating in the chemical bond.

So as already mentioned, this method, rather qualitative, considers that the molecular wave function is a "Linear Combination of Atomic Orbitals" (LCAO) unlike the previous method where we multiply the effective Hamiltonian.

This method is important because it is the basis of much of the current vocabulary of chemists when the chemistry done is cutting edge one!

Let us take the example of the dihydrogen molecule H_2. The idea is then following:

If we have the function of the atomic orbital $1s_A$ of H_1 and respectively the function $1s_B$ of H_B, then we assume that the dicentric molecular orbital (linked to two atoms) thereof is given by:

$$\Psi_{AB}=a\Psi_A+b\Psi_B \qquad (13.2.19)$$

which defines a quantum system with two eigenstates.

But as we well know, in reality, only the square of the wave function has a physical sense (probability of presence). Thus, if we assume that the wave function has no value in $\mathbb{C}$, we have for the single electron of interest ($1s$):

$$\Psi_{AB}^2=(a\Psi_A+b\Psi_B)^2=a^2\Psi_A^2+2ab\Psi_A\Psi_B+b^2\Psi_B^2 \qquad (13.2.20)$$

where we assume that:

- $a^2\Psi_A^2$ represents the probability of presence to be near A.
- $b^2\Psi_B^2$ represents the probability of presence to be near B.
- $2ab\Psi_A\Psi_B$ represents the probability of presence of the electron that do the link $A-B$.

In the particular case of the symmetric diatomic molecule we have chosen as an example, the atoms A and B perform the same function and there is no reason that the electron is closer to A than to B or vice versa.

Thus, the probability of finding the electron near A is equal to the probability of finding it near B.

$$a^2\Psi_A^2=b^2\Psi_B^2 \qquad (13.2.21)$$

Moreover, in this case the orbitals Ψ_A and Ψ_B are completely identical ($1s$ orbitals, both of the same atom) and there is therefore no need to distinguish them. So we have:

$$a^2=b^2\Rightarrow a=\pm b \qquad (13.2.22)$$

We have two solutions for Ψ_{AB} that are (these two solutions can be found in very different notations in the literature):

$$\Psi_{AB}^{(1)} = a\Psi_A + a\Psi_B = a(\Psi_A + \Psi_B) \quad \text{if } b = +a \tag{13.2.23}$$

and:

$$\Psi_{AB}^{(2)} = a\Psi_A - a\Psi_B = a(\Psi_A - \Psi_B) \quad \text{if } b = -a \tag{13.2.24}$$

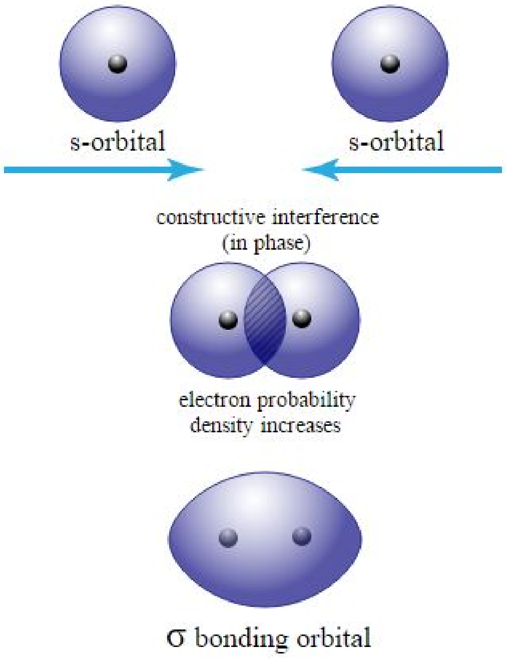
Caution! We can not put for the last two relations that $\Psi_A = \Psi_B$. The latter equality occurs at any point only if the distance between the two nucleus is zero (which is unlikely) or, if they are spaced a distant of a certain value D in the middle thereof.

These two expressions are simultaneously solutions of the Schrödinger equation. So we get two molecular orbitals from the two atomic orbitals in the case of symmetrical diatomic molecule.

The function:

$$\Psi_A + \Psi_B \tag{13.2.25}$$

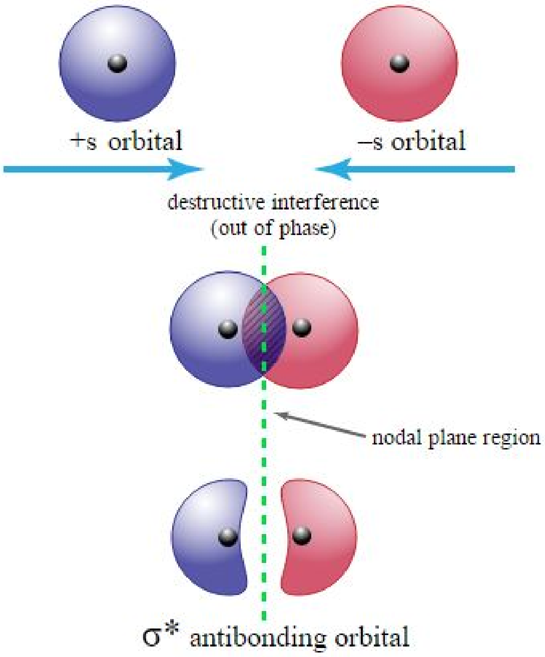
is named "bonding function" because it corresponds to a reinforcement of the probability of presence of the electron between atoms A and B which corresponds to the creation of the bond!


s-orbital
s-orbital
constructive interference
(in phase)
electron probability
density increases
σ bonding orbital

Conversely, the function:

$$\Psi_A - \Psi_B \tag{13.2.26}$$

is named "anti-bonding function" because it corresponds to a reduction of the probability of presence of the electron between atoms A and B which corresponds to the destruction of the bond!



Ultimately, by overlapping, the two atomic orbitals with the same energy give birth to two molecular orbitals of different energy, a stabilized binding and the other anti-bonding destabilized.

We have obviously from what we see just above that, in more complex cases, the energy level of the bonding molecular orbital is smaller than the anti-bonding (we will prove this rigorously in details below).

Thus, it takes more energy to ionize respectively the electron of the binding orbital σ than to ionize the electron of the anti-bonding orbital σ^*. It is commonly accepted that the energy of the bond function is stronger than the anti-bonding one (but we will make the proof further below).

Let us also indicate that in chemistry, a chemical bond wherein each of the bonded atoms is sharing an electron from one of its outer layers to form a pair of electrons linking two atoms is commonly known as "covalent bond".

The chemists then say the covalent bond involves the equitable sharing of only one pair of electrons, named "bonding pair" (but in fact where only one electron is really shared). Each atom provides an electron, the electron pair is then delocalized between two atoms as we have shown.

These are the reasons why we commonly say that the bond σ is a covalent chemical bond between two atoms generated by orbital axial overlap.

Now let us in-deep this approach! The molecular orbitals are to be normalized as we know.

Which means that:

$$\int\limits_{-\infty}^{+\infty}|\Psi|^2\mathrm{d}x\mathrm{d}y\mathrm{d}z=1 \tag{13.2.27}$$

What gives, since the atomic orbitals are normalized for Ψ_1 and are real functions:

$$\begin{aligned}\int\limits_{-\infty}^{+\infty}|\Psi_{AB}^{(1)}|^2\mathrm{d}x\mathrm{d}y\mathrm{d}z&=\int\limits_{-\infty}^{+\infty}|a(\Psi_A+\Psi_B)|^2\mathrm{d}x\mathrm{d}y\mathrm{d}z\\&=\int\limits_{-\infty}^{+\infty}(a^2\Psi_A^2+2a^2\Psi_A\Psi_B+a^2\Psi_B^2)\mathrm{d}x\mathrm{d}y\mathrm{d}z=1\end{aligned} \tag{13.2.28}$$

Since a (real number in our case) is imposed as a constant, it comes immediately:

$$\begin{aligned}&a^2\left(\int\limits_{-\infty}^{+\infty}\Psi_A^2\mathrm{d}x\mathrm{d}y\mathrm{d}+\int\limits_{-\infty}^{+\infty}\Psi_B^2\mathrm{d}x\mathrm{d}y\mathrm{d}z+2\int\limits_{-\infty}^{+\infty}\Psi_A\Psi_B\mathrm{d}x\mathrm{d}y\mathrm{d}z\right)\\=&a^2\left(2+2\int\limits_{-\infty}^{+\infty}\Psi_A\Psi_B\mathrm{d}x\mathrm{d}y\mathrm{d}z\right)=a^2(2+2S_{12})=1\end{aligned} \tag{13.2.29}$$

Therefore for Ψ_{AB}^1:

$$a=\frac{1}{\sqrt{2(1+S_{12})}} \tag{13.2.30}$$

Identically, we have for Ψ_{AB}^2:

$$a=\frac{1}{\sqrt{2(1-s_{12})}} \tag{13.2.31}$$

If we have $S_{12}\ll 1$, it comes the following format that we find in many books:

$$\begin{aligned}\Psi_{AB}^{(1)}&\cong\frac{1}{\sqrt{2}}(\Psi_A+\Psi_B)\\\Psi_{AB}^{(2)}&\cong\frac{1}{\sqrt{2}}(\Psi_A-\Psi_B)\end{aligned} \tag{13.2.32}$$

Let us make a small example using as orbital, the lowest atomic orbital ($1s$) of the hydrogen atom in the case of a dihydrogeneous bond H_2 for which we have proved at the end of the section of Quantum Chemistry (see page 6178) that:

$$\Psi_{1s}=\Psi_{1,0,0,0}=R_{1,0}Y_{0,0}=\frac{2}{r_0^{3/2}}e^{-r/r_0}\frac{1}{\sqrt{4\pi}}=\frac{1}{\sqrt{\pi}}\frac{1}{r_0^{3/2}}e^{-r/r_0} \tag{13.2.33}$$

Therefore it comes:

$$\Psi_A=\frac{1}{\sqrt{\pi}}\frac{1}{r_0^{3/2}}e^{-r_1/r_0}=\lambda e^{-r_1/r_0}\qquad\text{and}\qquad\Psi_B=\frac{1}{\sqrt{\pi}}\frac{1}{r_0^{3/2}}e^{-r_2/r_0}=\lambda e^{-r_2/r_0} \tag{13.2.34}$$

with for recall:

$$r_0 = 4\pi\varepsilon_0 \frac{\hbar}{\mu e^2} \tag{13.2.35}$$

It comes then for the molecular binding orbital of level s:

$$\sigma_s = \frac{\lambda}{\sqrt{2}} \left(e^{-r_1/r_0} + e^{-r_2/r_0} \right) \tag{13.2.36}$$

and for the anti-bonding of also the s level:

$$\sigma_s^* = \frac{\lambda}{\sqrt{2}} \left(e^{-r_1/r_0} - e^{-r_2/r_0} \right) \tag{13.2.37}$$

We then see immediately that σ_s^* vanishes in the middle of the two protons because in this place $r_1 = r_2$. The molecular anti-bonding orbital therefore has a nodal plane and the electrons are mainly located on the protons.

By cons, for the molecular orbital σ_s the density does not vanish. Then we understand easily that an electron of σ_s ensures the stability of the molecule and is therefore responsible for the chemical bond.

We therefore conclude that the electronic stabilization due to the two identical orbital interaction is proportional to their recovery. More the recovery is big, the more the stabilization is important.

There is a more technical approach using Dirac notation (see section Wave Quantum Physics page 4729) and that has the advantage of allowing the determination of the eigenvalues of energy.

First we write the general expression of the time independent Schrödinger equation with the Bra-Ket notation for one molecular orbital, superposition of two atomic orbitals:

$$H|\Psi_{AB}\rangle = E|\Psi_{AB}\rangle \tag{13.2.38}$$

Either in explicit form:

$$H|a\Psi_A + b\Psi_B\rangle = E|a\Psi_A + b\Psi_B\rangle \tag{13.2.39}$$

If we multiply by the bra $\langle\Psi_A|$ on the left and taking into account that a, b and the specific eigenvalues of the energy are constants, we get the following equation:

$$a\langle\Psi_A|H|\Psi_A\rangle + b\langle\Psi_A|H|\Psi_B\rangle = aE\langle\Psi_A|\Psi_A\rangle + bE\langle\Psi_A|\Psi_B\rangle \tag{13.2.40}$$

Similarly, we get the bra $\langle\Psi_B|$:

$$a\langle\Psi_B|H|\Psi_A\rangle + b\langle\Psi_B|H|\Psi_B\rangle = aE\langle\Psi_B|\Psi_A\rangle + bE\langle\Psi_B|\Psi_B\rangle \tag{13.2.41}$$

Let us simplify the notations even more:

$$\begin{gathered} aH_{AA} + bH_{AB} = aES_{AA} + bES_{AB} \\ aH_{BA} + bH_{BB} = aES_{BA} + bES_{BB} \end{gathered} \tag{13.2.42}$$

By symmetry of the problem in the case of dihydrogen, we put:

$$\langle\Psi_A|H|\Psi_B\rangle=H_{AB}=\langle\Psi_B|H|\Psi_A\rangle=H_{BA} \tag{13.2.43}$$

which are named "resonance integrals" because it is a term relating to the combination (resonance) of the both atomic orbital relative to the two atoms that made the molecular structure.

We also have:

$$\langle\Psi_A|H|\Psi_A\rangle=H_{AA} \qquad \text{and} \qquad \langle\Psi_B|H|\Psi_B\rangle=H_{BB} \tag{13.2.44}$$

which are named "Coulomb integrals" because they correspond according to the fifth postulate of Wave Quantum Physics (see section of the same name page 4699) to the average value of the total energy of the electron.

We have obviously:

$$\langle\Psi_A|\Psi_A\rangle=S_{AA}=\langle\Psi_B|\Psi_B\rangle=S_{BB}=1 \tag{13.2.45}$$

which are named "recovery integrals" because the two atomic orbitals of the same type of each atom overlap.

And finally, we have always by symmetry of our particular case:

$$\langle\Psi_A|\Psi_B\rangle=S_{AB}=\langle\Psi_B|\Psi_A\rangle=S_{BA} \tag{13.2.46}$$

We can then write, since the recovery integrals are unitary:

$$\begin{aligned}a(H_{AA}-E)+b(H_{AB}-ES_{AB})=0\\ a(H_{BA}-ES_{BA})+b(H_{BB}-E)=0\end{aligned} \tag{13.2.47}$$

These two equations are named "secular equations". The trivial solution is a priori not physical because it would mean that the electron has a zero probability density at any point in the space corresponding at $a=b=0$.

There is a non-trivial solution and unique solution if and only if the following determinant (see section Linear Algebra page 1961), known in molecular chemistry under the name "secular determinant", is equal to zero:

$$\begin{vmatrix} H_{AA}-E & H_{AB}-ES_{AB} \\ H_{BA}-ES_{BA} & H_{BB}-E \end{vmatrix}=(H_{AA}-E)(H_{BB}-E)-(H_{AB}-ES_{AB})(H_{BA}-ES_{BA})=0 \tag{13.2.48}$$

As we have by symmetry in our particular case:

$$H_{AA}=H_{BB} \qquad H_{AB}=H_{BA} \qquad S_{AB}=S_{BA} \tag{13.2.49}$$

Therefore it comes:

$$(H_{AA}-E)^2-(H_{AB}-ES_{AB})^2=0 \tag{13.2.50}$$

Hence:

$$H_{AA}-E=\pm(H_{AB}-ES_{AB}) \tag{13.2.51}$$

This gives us two solutions $(+)$:

$$-E_1+E_1S_{AB}=H_{AB}-H_{AA}\Rightarrow E_1(1-S_{AB})=H_{AA}-H_{AB}\Rightarrow E_1=\frac{H_{AA}-H_{AB}}{1-S_{AB}} \tag{13.2.52}$$

and minus $(-)$:

$$-E_2-E_2S_{AB}=-H_{AB}-H_{AA}\Rightarrow -E_2(1+S_{AB})=-(H_{AA}-H_{AB})\Rightarrow E_2=\frac{H_{AA}+H_{AB}}{1+S_{AB}} \tag{13.2.53}$$

Therefore we have:

$$E_1=\frac{H_{AA}-H_{AB}}{1-S_{AB}} \qquad \text{and} \qquad E_2=\frac{H_{AA}+H_{AB}}{1+S_{AB}} \tag{13.2.54}$$

But to be able to calculate the energy levels in detail, we must still have the shape of the Hamiltonian... and that using the both electrons of the dihydrogen molecule is quite difficult... To simplify the study, we reduce ourselves to the case of the cation (positive ion) H_2^+ consisting of two protons and one electron:

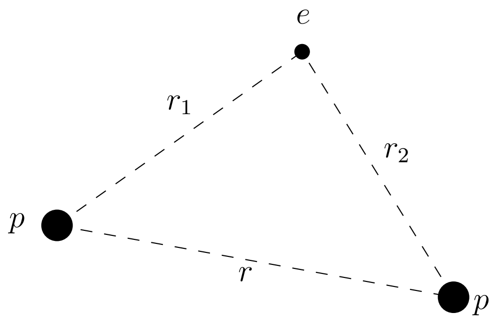


Figure 13.23: **Simplified study of the dihydrogen cation** H_2^+

We then have base on the relation we have obtained at the beginning of this section:

$$H_{\text{el}}=-\sum_{i=1}^{n}\frac{\hbar^2}{2m_e}\vec{\nabla}_i^2-\sum_{k=1}^{N}\sum_{i=1}^{n}\frac{Ze^2}{4\pi\varepsilon_0 r_{ik}}+\sum_{\substack{i=1\\j>1}}^{n-1}\frac{e^2}{4\pi\varepsilon_0 r_{ik}} \tag{13.2.55}$$

The following relation:

$$H=-\frac{\hbar^2}{2m_e}\vec{\nabla}^2+\frac{e^2}{4\pi\varepsilon_0}\left(-\frac{1}{r_1}-\frac{1}{r_1}+\frac{1}{r}\right) \tag{13.2.56}$$

where the first two terms in the brackets are for recall associated with the potential energy of the electron and the last to the potential repulsion energy of proton (the first term on the right of the equality is the kinetic energy of the electron).

Now let us try to sort the energy of these two molecular orbitals. For this, we write:

$$E_1=\frac{H_{AA}-H_{AB}}{1-S_{AB}}=\frac{\alpha-\beta}{1-S} \qquad \text{and} \qquad E_2=\frac{H_{AA}+H_{AB}}{1+S_{AB}}=\frac{\alpha+\beta}{1+S} \tag{13.2.57}$$

Let us recall that for a system to be stable, the energies E_n must be negatives, this corresponding to the stable states (we need a supply of energy to take them out) and request from us because of the shape of E_2:

$$\alpha < 0 \qquad \beta < 0 \tag{13.2.58}$$

Knowing this it comes:

$$E_2 < E_1 < 0 \tag{13.2.59}$$

Therefore, we see that the notations are not consistent with the use in quantum physics because normally the index 1 is reserved to the lowest energy. So we will write in the future:

$$E_1 = \frac{\alpha+\beta}{1+S} \qquad \text{and} \qquad E_2 = \frac{\alpha-\beta}{1-S} \tag{13.2.60}$$

with the associated eigenfunctions Ψ_1 and Ψ_2 and therefore:

$$E_1 < E_2 < 0 \tag{13.2.61}$$

We can also noticed an important thing! This is that if we consider the atoms in isolated, the interaction terms cancel and we have:

$$E_{1,2} \cong \alpha \tag{13.2.62}$$

Therefore we have the qualitative difference between a single atom and a simple diatomic (ionized) system:

$$\frac{\alpha+\beta}{1+S} < \alpha < 0 \tag{13.2.63}$$

This means that the energy of the lowest level of a diatomic ionized molecule is less than the energy of a single atom which is near α. This observation confirms that the system is stabilized in energy compared to two isolated atoms, which seems consistent with the experimental determination of the existence of such molecules.

Traditionally chemists represent the energy differences in the following form for our particular case:

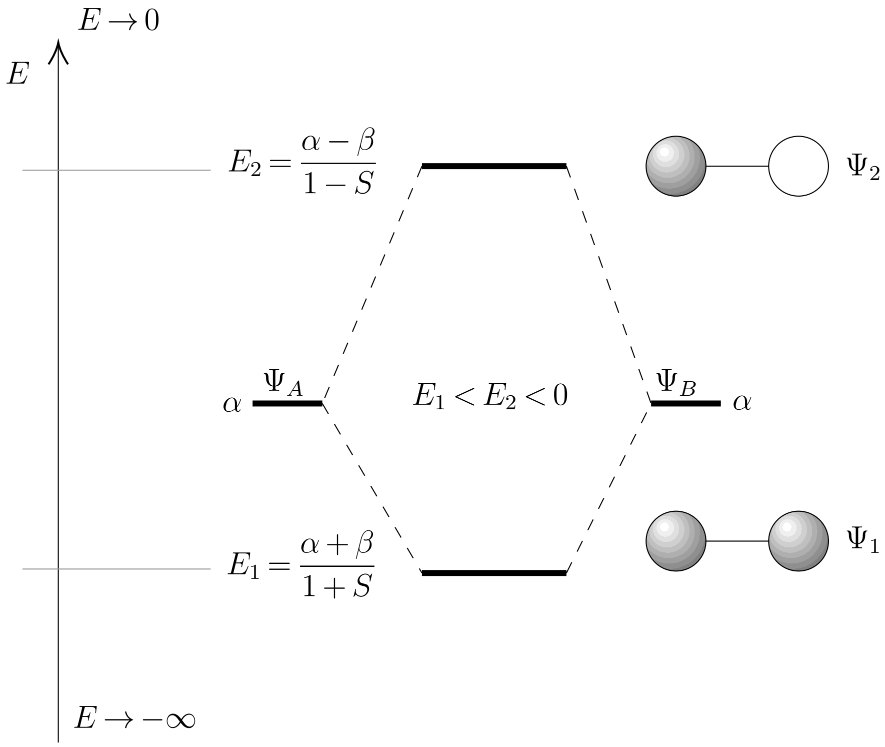

$E \to 0$
E
$E_2 = \frac{\alpha-\beta}{1-S}$
Ψ_2
α
Ψ_A
$E_1 < E_2 < 0$
Ψ_B
α
Ψ_1
$E_1 = \frac{\alpha+\beta}{1+S}$
$E \to -\infty$

Figure 13.24: Energy levels of dihydrogen cation H_2

We therefore conclude - by generalizing a little bit... - that when two atoms (each contributing with an electron) combine, their atomic orbitals will combine to generate two molecular orbitals, one of energy level Ψ_1 and the second of higher energy level Ψ_2 than that of the isolated atoms. Thus, the split up that will make leave one of the electron with one of atoms will be exothermic in comparison to the single atoms.

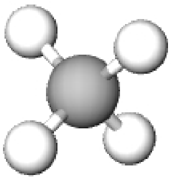
Up to now we have discussed the electronic states of rigid molecules, where the nuclei are clamped to a fixed position. In this section we will improve our model of molecules and include the rotation and vibration of diatomic molecules.

13.2.3 Molecular Rotational Energy Levels

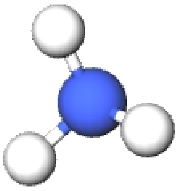
As we have seen in the section of Quantum chemistry, for analytical reasons we consider molecules as rigid rotators.

The rigid rotators are commonly classified into four types:

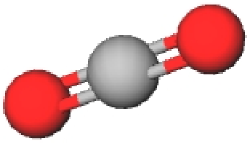
- Spherical rotors: have equal moments of inertia (e.g., CH_4).

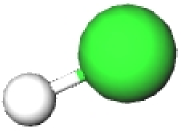


- Symmetric rotors: have two equal moments of inertia (e.g., NH_3).

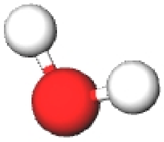


- Linear rotors: have one moment of inertia equal to zero (e.g., CO_2, HCl).





- Asymmetric rotors: have three different moments of inertia (e.g., H_2O).



Let us now recall that we have proved in the section of Quantum Chemistry that for the rigid rotator (see page 6134) the part of the Hamiltonian dedicated to the rotation of energy is:

$$E_{\text{rot}}=\frac{L^2}{2\mu r^2} \tag{13.2.64}$$

Where L^2 was is an operator but from which we know from our study of Wave Quantum Physics (see page 4887) that the eigenvalues are:

$$L\leq\sqrt{l(l+1)}\hbar \tag{13.2.65}$$

and where r is the distance between the two corpuscles (nucleus and electron in the context of our study of the hydrogenous atom in the section of Quantum Chemistry) and:

$$\mu=\frac{m_eM}{M_{\text{tot}}}=\frac{m_eM}{m_e+M} \tag{13.2.66}$$

In the context of diatomic molecules A and B it is more common to write r_{AB} and:

$$\mu=\frac{M_AM_B}{M_A+M_B} \tag{13.2.67}$$

In the section of Wave Quantum Physics we have seen (page 4877) that if we must consider the spin we have to write the more general form:

$$J\leq\sqrt{J_z(J_z+1)}\hbar \tag{13.2.68}$$

with $J=0,1/2/,1,3/2,2,5/2,\ldots$. Therefore:

$$\boxed{E_{\text{rot}}=\frac{J^2}{2\mu r^2}=\frac{J_z(J_z+1)\hbar^2}{2\mu r^2}} \tag{13.2.69}$$

In the old style spectroscopic literature, the rotational term values $F(J)=E(J)/hc$ are used instead of the energies....The previous relation is then written:

$$F_{\text{rot}}=\frac{J_z(J_z+1)\hbar^2}{2hc\mu r^2}=\bar{B}J(J+1) \tag{13.2.70}$$

with the "rotational constant":

$$\bar{B}=\frac{\hbar}{4\pi c\mu r^2} \tag{13.2.71}$$

We also know from the section of Classical Mechanics (page 3297) that:

$$E_{\text{rot}}=\frac{1}{2}\omega^2 J \tag{13.2.72}$$

Therefore:

$$\frac{1}{2}\omega^2 J=\frac{J^2}{2\mu r^2} \tag{13.2.73}$$

That simplifies to:

$$\omega=\frac{2\pi}{T}=\sqrt{\frac{J}{\mu r^2}} \tag{13.2.74}$$

Hence:

$$T=2\pi\sqrt{\frac{\mu r^2}{J}}=2\pi\sqrt{\frac{\mu r^2}{\sqrt{J_z(J_z+1)}\hbar}} \tag{13.2.75}$$

The energy separation between the rotational levels J and $J+1$ is given obviously by:

$$\begin{aligned}\Delta E_{\text{rot}}=E_{\text{rot}}(J+1)-E_{\text{rot}}(J)&=\frac{(J_z+1)((J_z+1)+1)\hbar^2}{2\mu r^2}-\frac{J_z(J_z+1)\hbar^2}{2\mu r^2}\\&=\frac{\hbar^2}{2\mu r^2}\left(J_z^2+2J_z+J_z+2-J_z^2-J_z\right)=\frac{(J_z+1)\hbar^2}{\mu r^2}\end{aligned} \tag{13.2.76}$$

and increase linearly with J.

Let us now calculate the moment of inertia, that we will denoted I to avoid the confusion with the orbital kinetic momentum J used above, of a diatomic molecule. Let us imagine the diatomic molecule as a system of two tiny spheres at either end of a thin weightless rod.

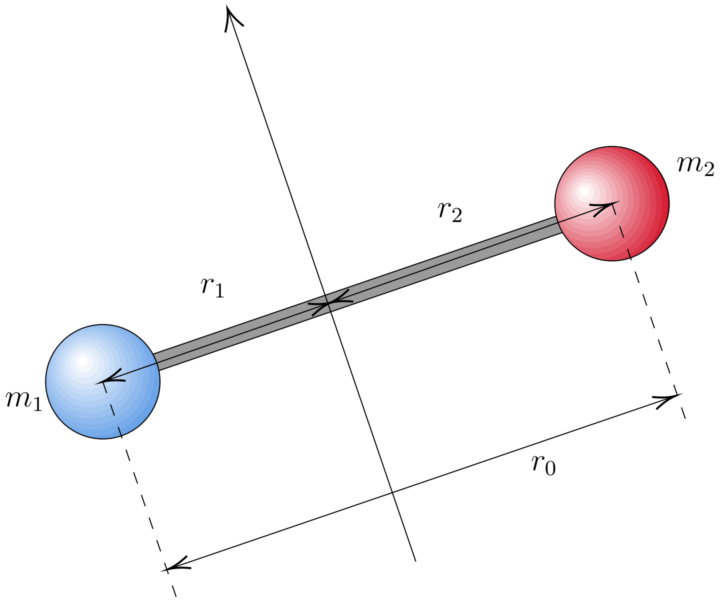


Figure 13.25: Construction for the study of inertia momentum of a diatomic molecule

Let C be the center of mass of the molecule. Let r_1 and r_2 be the distances of the two atoms of respective masses m_1, m_2 from the center of mass C of the molecule: We see that:

$$r_1+r_2=r_0 \tag{13.2.77}$$

and we have (see section Classical Mechanics page 3213):

$$m_1 r_1 = m_2 r_2 \tag{13.2.78}$$

Therefore:

$$r_2 = r_0 - r_1 \tag{13.2.79}$$

Hence:

$$m_1 r_1 = m_2 r_2 = m_2(r_0 - r_1) = m_2 r_0 - m_2 r_1 \tag{13.2.80}$$

After rearranging we get:

$$r_1(m_1 + m_2) = m_2 r_0 \tag{13.2.81}$$

or:

$$r_1 = \frac{m_2}{m_1 + m_2} r_0 \tag{13.2.82}$$

Similarly:

$$r_2 = \frac{m_1}{m_1 + m_2} r_0 \tag{13.2.83}$$

Let I (instead of the traditional notation J) be the total moment of inertia of the diatomic molecule about an axis passing through the center of mass of the molecule and perpendicular to the bond length.

Then we have seen in the section of Classical Mechanics (see page 3297) that:

$$I = m_1 r_1^2 + m_2 r_2^2 \tag{13.2.84}$$

or:

$$I = m_1 \left(\frac{m_2}{m_1 + m_2} r_0\right)^2 + m_2 \left(\frac{m_1}{m_1 + m_2} r_0\right)^2 \tag{13.2.85}$$

thus:

$$I = \frac{m_1 m_2^2}{(m_1 + m_2)^2} r_0^2 + \frac{m_2 m_1^2}{(m_1 + m_2)^2} r_0^2 \tag{13.2.86}$$

after simplification:

$$I = \frac{m_1 m_2 (m_2 + m_1)}{(m_1 + m_2)^2} r_0^2 \tag{13.2.87}$$

Hence:

$$I = \frac{m_1 m_2}{m_1 + m_2} r_0^2 \tag{13.2.88}$$

So finally for a diatomic molecule (or any pair of object turning around a common center) we get the following moment of inertia:

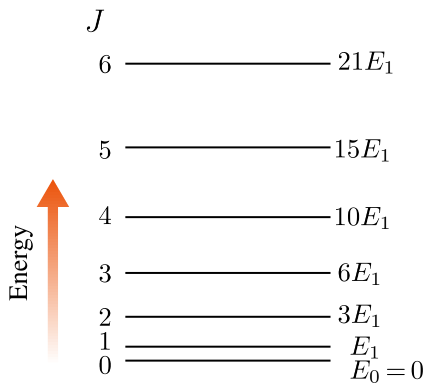
$$\boxed{I = \frac{m_1 m_2}{m_1 + m_2} r_0^2 = \mu r_0^2} \tag{13.2.89}$$

Hence the fact that we often found in the literature the previous main relations under the form:

$$\boxed{E_{\text{rot}}=\frac{J_z(J_z+1)\hbar^2}{2I}} \tag{13.2.90}$$

and:

$$\boxed{\Delta E_{\text{rot}}=\frac{(J_z+1)\hbar^2}{I}} \tag{13.2.91}$$



Therefore, as:

$$h\nu=\Delta E_{\text{rot}}=\frac{(J_z+1)\hbar^2}{I} \tag{13.2.92}$$

the frequencies at which transitions can occur are given by :

$$\nu=\frac{(J_z+1)\hbar}{2\pi I} \tag{13.2.93}$$

Notice that for $J_z=0$ we have a non-null zero point energy and frequency:

$$\nu=\frac{\hbar}{2\pi I} \tag{13.2.94}$$

✍**Example:**

The molecule NaH is found to undergo a rotational transition from $J=0$ to $J=1$ when it absorbs a photon of frequency $2.94\cdot 10^{11}$ [Hz]. We want to know the equilibrium bond length of the molecule.

For this purpose we use $J_z=0$ in the formula for the transition frequency:

$$\nu=\frac{\hbar}{2\pi I}=\frac{\hbar}{2\pi\mu r_0^2} \tag{13.2.95}$$

Solving for r_0 gives:

$$r_0 = \sqrt{\frac{\hbar}{2\pi\mu\nu}} \tag{13.2.96}$$

The reduced mass is given by:

$$\mu = \frac{m_{\mathrm{Na}} m_{\mathrm{H}}}{m_{\mathrm{Na}} + m_{\mathrm{H}}} \cong \frac{22.989 \cdot 1.0078}{22.989 + 1.0078} \cong 0.9655 \tag{13.2.97}$$

which is in atomic mass units or relative units. In order to convert to kilograms, we need the conversion factor $1\,[\mathrm{au}] = 1.66 \cdot 10^{-27}$ [kg]. Multiplying this by 0.9655 gives a reduced mass of $1.603 \cdot 10^{-27}$ [kg]. Substituting in for r_0 gives:

$$r_0 \cong \sqrt{\frac{1.055 \cdot 10^{-34}}{2\pi(1.603 \cdot 10^{-27})(2.94 \cdot 10^{11})}} \cong 1.89 \cdot 10^{-10}\,[\mathrm{m}] \tag{13.2.98}$$

13.2.4 Molecular Vibrational Energy Levels

Let us consider the simple case of a vibrating diatomic molecule, where restoring force is proportional to displacement such that (see section Mechanics page 3416):

$$F = -kx \tag{13.2.99}$$

The potential energy is as we have proved it in the previously mentioned section, but with the notation of Quantum Physics, given by:

$$V = \frac{1}{2}kx^2 \tag{13.2.100}$$

Now remember that we have proved in the section of Wave Quantum Physics (see page 4758), that the Schrödinger equation was given by:

$$\left(E_p - \frac{\hbar^2}{2m}\frac{\partial^2}{\partial^2 x}\right)\Psi = E_{\text{tot}}\Psi \tag{13.2.101}$$

After rearrangement:

$$\frac{\hbar^2}{2m}\frac{\partial^2}{\partial^2 x}\Psi + (E_{\text{tot}} - E_p) = 0 \tag{13.2.102}$$

And using the conventional notations in chemistry and quantum physics:

$$\frac{\hbar^2}{2m}\frac{\partial^2}{\partial^2 x}\Psi + (E - V) = 0 \tag{13.2.103}$$

Before continuing we need to study a bit this important configuration of masses:

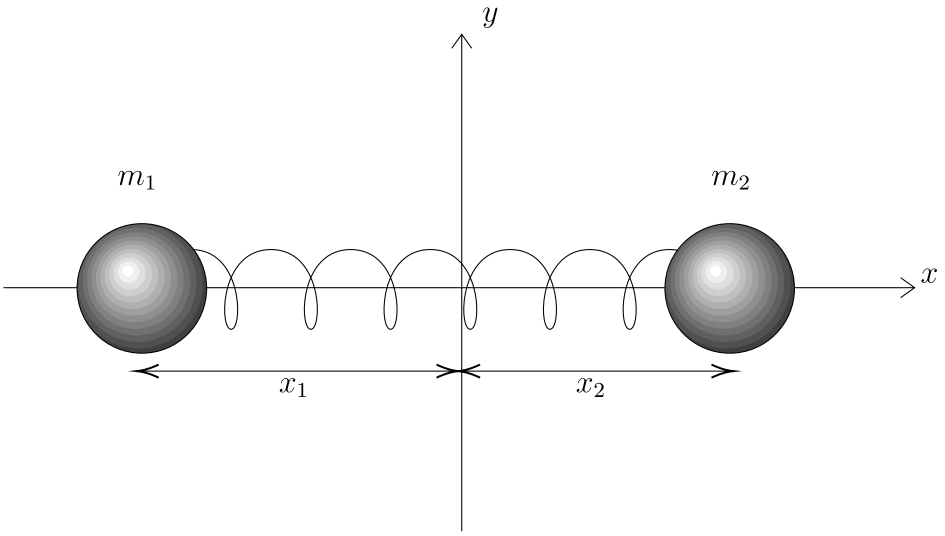


Figure 13.26: **Harmonic oscillator with two masses**

We will assume that the two masses m_1 and m_2 in figure above have positions labelled as x_1 and x_2 relatively to the center of mass but are moving back and forth as a harmonic oscillator. We will ignore any other motion of these two masses (like translation or rotation) and focus solely on the oscillation. In a purely harmonic oscillation (also named a "vibration") the center of mass r_g (see section Classical Mechanics page 3213) does not change, so that:

$$m_1\frac{\mathrm{d}x_1}{\mathrm{d}t}=-m_2\frac{\mathrm{d}x_2}{\mathrm{d}t} \tag{13.2.104}$$

The negative sign indicates that the masses are moving in the opposite directions. By adding the mixed term $m_2(\mathrm{d}x_1/\mathrm{d}t)$ to both sides, we get:

$$m_1\frac{\mathrm{d}x_1}{\mathrm{d}t}+m_2\frac{\mathrm{d}x_1}{\mathrm{d}t}=-m_2\frac{\mathrm{d}x_2}{\mathrm{d}t}+m_2\frac{\mathrm{d}x_1}{\mathrm{d}t} \tag{13.2.105}$$

Therefore:

$$(m_1+m_2)\frac{\mathrm{d}x_1}{\mathrm{d}t}=m_2\left(\frac{\mathrm{d}x_1}{\mathrm{d}t}-\frac{\mathrm{d}x_2}{\mathrm{d}t}\right) \tag{13.2.106}$$

This is rearranged to:

$$\frac{\mathrm{d}x_1}{\mathrm{d}t}=\frac{m_2}{m_1+m_2}\left(\frac{\mathrm{d}x_1}{\mathrm{d}t}-\frac{\mathrm{d}x_2}{\mathrm{d}t}\right) \tag{13.2.107}$$

Let us define now the relative coordinate r as:

$$r=x_1-x_2 \tag{13.2.108}$$

and thus:

$$\frac{\mathrm{d}r}{\mathrm{d}t}=\frac{\mathrm{d}x_1}{\mathrm{d}t}-\frac{\mathrm{d}x_2}{\mathrm{d}t} \tag{13.2.109}$$

Now we can substitute that into:

$$\frac{\mathrm{d}x_1}{\mathrm{d}t}=\frac{m_2}{m_1+m_2}\left(\frac{\mathrm{d}x_1}{\mathrm{d}t}-\frac{\mathrm{d}x_2}{\mathrm{d}t}\right) \tag{13.2.110}$$

to get:

$$\frac{\mathrm{d}x_1}{\mathrm{d}t}=\frac{m_2}{m_1+m_2}\frac{\mathrm{d}r}{\mathrm{d}t} \tag{13.2.111}$$

Identically we can get:

$$\frac{\mathrm{d}x_2}{\mathrm{d}t}=\frac{m_1}{m_1+m_2}\frac{\mathrm{d}r}{\mathrm{d}t} \tag{13.2.112}$$

as a second expression.

In considering the total energy of this harmonic oscillation, the potential energy is the same as for any other harmonic oscillator but the kinetic energy is the sum of the kinetic energies of the two particles. That is:

$$E_c=\frac{1}{2}m_1\left(\frac{\mathrm{d}x_1}{\mathrm{d}t}\right)^2+\frac{1}{2}m_1\left(\frac{\mathrm{d}x_2}{\mathrm{d}t}\right)^2 \tag{13.2.113}$$

Using the two previous relations, it is easy to substitute and show that the kinetic energy has a simple form in terms of the time derivative of the relative coordinate $\mathrm{d}r/\mathrm{d}t$:

$$E_c=\frac{1}{2}\frac{m_1 m_2}{m_1+m_2}\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2 \qquad (13.2.114)$$

The reduced mass μ (see section Classical Mechanics page 3213) is as we know:

$$\mu=\frac{m_1 m_2}{m_1+m_2} \qquad (13.2.115)$$

What this means is that the kinetic energy of the oscillator can be represented by the kinetic energy of a single mass moving back and forth, if that single mass has the reduced mass of the two masses in the original system. This allows us to treat the two-particle harmonic oscillator as a one-particle harmonic oscillator and use the same equations and expressions that we derived for a simple harmonic oscillator.

So the result we get for the simple harmonic oscillator (see section Music Mathematics page 8224):

$$\omega=\sqrt{\frac{k}{m}} \qquad (13.2.116)$$

will be written:

$$\omega=\sqrt{\frac{k}{\mu}} \qquad (13.2.117)$$

So coming back to our initial problem, we have that:

$$\frac{\hbar^2}{2\mu}\frac{\partial^2}{\partial^2 x}\Psi+\left(E-\frac{1}{2}kx^2\right)=0 \qquad (13.2.118)$$

has for solution using what we have proved in the section of Wave Quantum Physics (see page 4807)and just above:

$$E_n=\hbar\omega_0\left(n+\frac{1}{2}\right)=\hbar\sqrt{\frac{k}{\mu}}\left(n+\frac{1}{2}\right) \qquad (13.2.119)$$

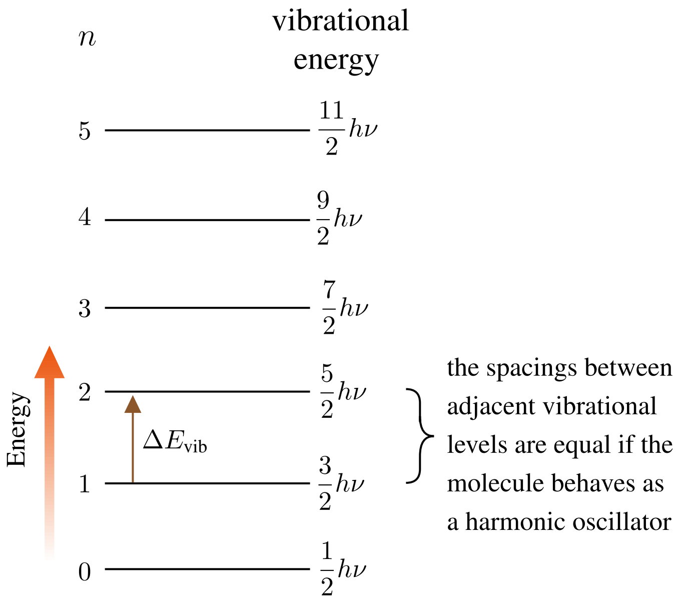
with for recall $n\in\mathbb{N}$. More often denoted:

$$\boxed{E_{\text{vib}}=\hbar\sqrt{\frac{k}{\mu}}\left(n+\frac{1}{2}\right)} \qquad (13.2.120)$$

Notice that for the same reason the harmonic oscillator has a zero point energy. We see that if $n=0$ in the above relation lead us to an equivalent zero point vibrational energy (it remains the $1/2$ factor interpreted by some chemists as a half-quantum number).

With the new studies of band spectra in the early 11920s (holocene calendar) the concept of zero-point energy became respectable among molecular physicists. Yet it was only in the fall of 11924 (holocene calendar) that half-quanta were firmly established in molecular spectroscopy.

In a study of the spectrum of boron monoxide (BO), Robert Mulliken, a young physical chemist at Harvard University, concluded that observations could only be understood on the assumption of quantum numbers with a minimum value of one-half.

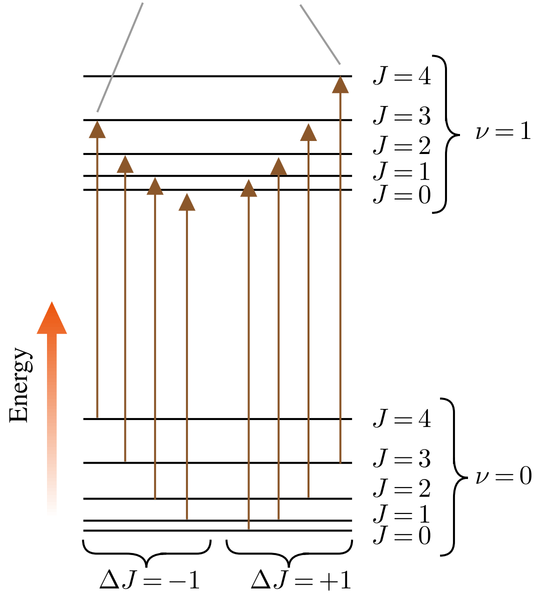

n
vibrational energy
5
$\frac{11}{2}h\nu$
4
$\frac{9}{2}h\nu$
3
$\frac{7}{2}h\nu$
2
$\frac{5}{2}h\nu$
1
$\frac{3}{2}h\nu$
0
$\frac{1}{2}h\nu$
Energy
ΔE_{vib}
the spacings between adjacent vibrational levels are equal if the molecule behaves as a harmonic oscillator

So we can combine the previous results to get the "vibrational-rotational energies level":

$$\boxed{E(n,J_z)=\hbar\sqrt{\frac{k}{\mu}}\left(n+\frac{1}{2}\right)+\frac{J_z(J_z+1)\hbar^2}{2\mu r^2}} \qquad (13.2.121)$$

Sometimes the n in the relation above is named the "vibrational quantum number" and often denoted by the letter v.

The energy levels of any molecule can be calculated from this expression, and each level is indexed by the two quantum numbers v and J. From these calculations, an energy-level diagram like the one shown below can be constructed:


the transitions obey the selection rule $\Delta J = \pm 1$ and fall into two sequences, those for $\Delta J = +1$ and those for $\Delta J = -1$

$J=4$
$J=3$
$J=2$
$J=1$
$J=0$
$\nu=1$
Energy
$J=4$
$J=3$
$J=2$
$J=1$
$J=0$
$\nu=0$
$\Delta J=-1$
$\Delta J=+1$

For each allowed value of the vibrational quantum number v, there is a complete set of rotational levels corresponding to $J = 0, 1, 2, \ldots$. The energy separation between successive rotational levels is much smaller than the separation between successive vibrational levels.

The following figure summarize the different types of energies and corresponding radiations involved in different typical phenomenon in molecular chemistry in normal conditions of temperature and pressure:

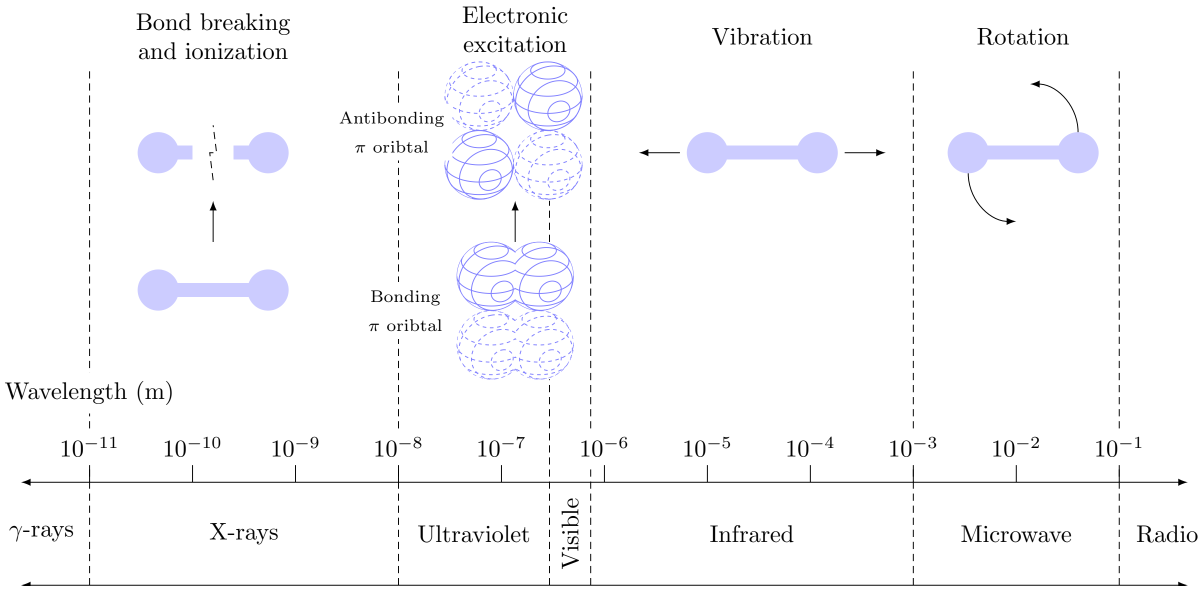
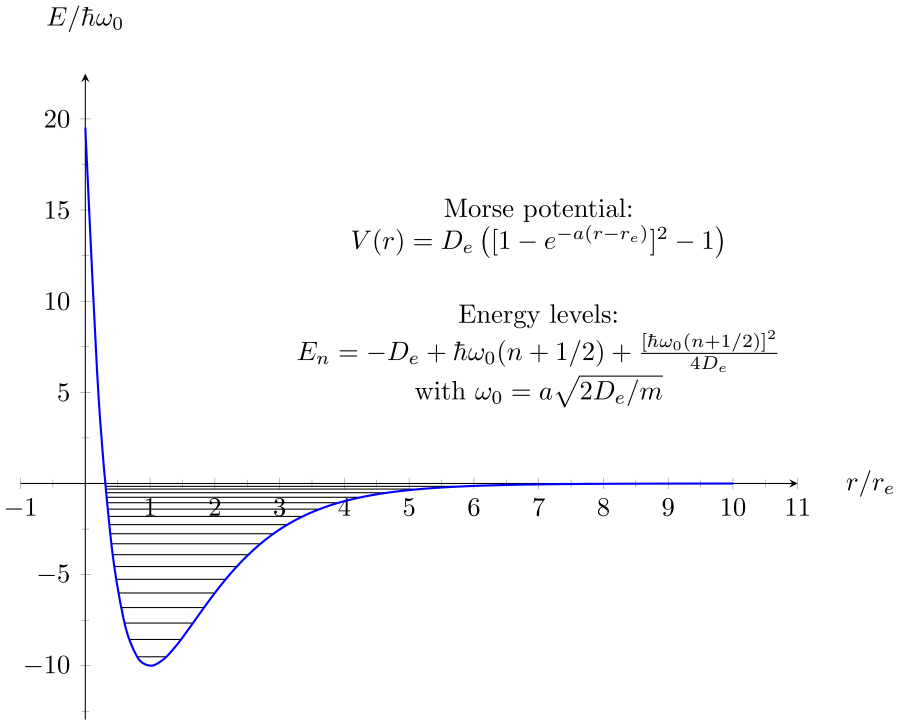

Bond breaking and ionization
Electronic excitation
Vibration
Rotation
Antibonding π oribtal
Bonding π oribtal
Wavelength (m)
10^{-11}
10^{-10}
10^{-9}
10^{-8}
10^{-7}
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
γ-rays
X-rays
Ultraviolet
Visible
Infrared
Microwave
Radio

Figure 13.27: **Various energies and corresponding radiation involved in some chemical/molecular phenomenon**

Since the early 121st century (holocene calendar) aberration-corrected transmission electron microscopy is shown to be a powerful analytical tool for the investigation and imaging of dynamics of metallic bonding at the atomic level almost in real-time (for an example see [88]).

13.2.5 Morse potential

The "Morse potential", named after physicist Philip M. Morse (paper published in 11929 - according to holocene calendar - see [401]), is a convenient interatomic interaction model for the potential energy of a diatomic molecule. It is a better approximation for the vibrational structure of the molecule than the quantum harmonic oscillator because it explicitly includes the effects of bond breaking, such as the existence of unbound states. It also accounts for the anharmonicity of real bonds and the non-zero transition probability for overtone and combination bands. The Morse potential can also be used to model other interactions such as the interaction between an atom and a surface. Due to its simplicity (only three fitting parameters), it is not used in modern spectroscopy. However, its mathematical form inspired the MLR (Morse/Long-range) potential, which is the most popular potential energy function used for fitting spectroscopic data.



The form of the Morse potential, in terms of the internuclear distance, is:

$$\boxed{V(r) = D\left(1 - e^{-\alpha\frac{r-r_0}{r_0}}\right)^2} \tag{13.2.122}$$

where r_0 is the "equilibrium internuclear distance" and D (often denoted D_e) is the "bond dissociation energy" (we can found the values of r_0 and D for 56 diatomic molecules in the original paper of Morse entitled *Diatomic molecules according to the wave mechanics*), and α the "Morse parameter".

Remark

How D_e and α in:

$$V(r)=D_e\left(\left[1-e^{-\alpha(r-r_0)}\right]^2-1\right) \tag{13.2.123}$$

are measured in out of the scope of this textbook (ask a professional spectroscopic chemistry).

In optical and infrared spectroscopy, experimentalists directly measure frequencies (ν) or wavelengths of light (λ). Therefore, D_e is frequently expressed in $[\text{cm}^{-1}]$ his is technically a unit of reciprocal length, but it is used as a proxy for energy via the relationship $E=h\nu=hc\lambda$. For example, strong covalent bonds have a D_e depth on the order of $20,000\,[\text{cm}^{-1}]$ to $40,000\,[\text{cm}^{-1}]$.

Morse couldn't determine the shape of his potential curve. He took the simplest method, which is heurestic: First, we draw the curve $y_1=-e^{-kx}$. This curve starts at $-\infty$ when $x=0$ and tends towards zero as $x\rightarrow+\infty$. If we add 1, we find $y_2=1-e^{-kx}$. This curve starts at $-\infty$ when $x=0$ and tends towards $+1$ as $x\rightarrow+\infty$. But it passes through $y=0$ at $x=1/k$. Squaring this curve gives $y_3=\left(1-e^{-kx}\right)^2$, which is always a positive curve. It starts at $+\infty$ at $x=0$, becomes zero at $x=1/k$, and tends towards $+1$ as $x\rightarrow+\infty$ (this is a matter of normalization only obviously as in reality $V(r)\rightarrow 0$ when $r\rightarrow+\infty$).

The Morse curve is this y_3 curve multiplied by D, where D is the dissociation energy of the diatomic molecule. The coefficient k is adjusted so that $1/k$ is equal to the interatomic distance r_0 between the two atoms of the diatomic molecule: $k=1/r_0$. The Morse curve thus defined has the further advantage of being approximated by a parabola around $x=1/k$, and therefore resembles a harmonic oscillator.

Remark

If we develop the square of the Morse potential we get obviously:

$$V(r)=D\left(1-2e^{-\alpha\frac{r-r_0}{r_0}}+e^{-2\alpha\frac{r-r_0}{r_0}}\right) \tag{13.2.124}$$

Often rearranged as:

$$V(r)=D\left(\underbrace{e^{-2\alpha\frac{r-r_0}{r_0}}}_{\substack{\text{short distance}\\\text{repulsion}}}-\underbrace{2e^{-\alpha\frac{r-r_0}{r_0}}}_{\substack{\text{attraction at}\\\text{far distance}}}+\underbrace{1}_{\substack{\text{asymptote}\\ r\rightarrow+\infty}}\right) \tag{13.2.125}$$

And often redefined as (because we don't care too much about a constant shift):

$$V(r):=V(r)-D=D\left(e^{-2\alpha\frac{r-r_0}{r_0}}-2e^{-\alpha\frac{r-r_0}{r_0}}\right)=D\left[\left(1-e^{-\alpha\frac{r-r_0}{r_0}}\right)^2-1\right] \tag{13.2.126}$$

It is common to define a variable x, such that:

$$x=\frac{r-r_0}{r_0} \tag{13.2.127}$$

which allows us to rewrite the potential as:

$$V(x) = D\left(1 - e^{-\alpha x}\right)^2 \tag{13.2.128}$$

It leads to a Schrödinger equation of the form:

$$-\frac{\hbar^2}{2\mu}\frac{\mathrm{d}^2\Psi}{\mathrm{d}x^2} + D\left(1 - e^{-\alpha x}\right)^2\Psi = E\Psi \tag{13.2.129}$$

Before attempting a solution, we follow standard practice and attempt to clean up this equation by changing variables. Thus, we define $y = \alpha x$, so that:

$$\frac{\mathrm{d}}{\mathrm{d}x} = \frac{\mathrm{d}y}{\mathrm{d}x}\frac{\mathrm{d}}{\mathrm{d}y} = \alpha\frac{\mathrm{d}}{\mathrm{d}y} \tag{13.2.130}$$

which transforms the Schrödinger equation into:

$$-\frac{\hbar^2\alpha^2}{2\mu}\frac{\mathrm{d}^2\Psi}{\mathrm{d}y^2} + D\left(1 - e^{-y}\right)^2\Psi = E\Psi \tag{13.2.131}$$

yields:

$$-\frac{\hbar^2\alpha^2}{2\mu}\frac{\mathrm{d}^2\Psi}{\mathrm{d}y^2} + D\left(1 - 2e^{-y} + e^{-2y}\right)\Psi = E\Psi \tag{13.2.132}$$

Rearranging terms to isolate the second derivative:

$$\frac{\mathrm{d}^2\Psi}{\mathrm{d}y^2} + \left[\frac{2\mu E}{\hbar^2\alpha^2} - \frac{2\mu D}{\hbar^2\alpha^2} + \frac{4\mu D}{\hbar^2\alpha^2}e^{-y} - \frac{2\mu D}{\hbar^2\alpha^2}e^{-2y}\right]\Psi = 0 \tag{13.2.133}$$

Now, we introduce two dimensionless parameters to represent energy and potential depth:

$$\gamma = \frac{\sqrt{2\mu D}}{\hbar\alpha} \quad \text{and} \quad K = \frac{\sqrt{2\mu(D-E)}}{\hbar\alpha} \tag{13.2.134}$$

(for bound states, $0 < E < D$, meaning K remains real and positive).

Substituting y and K simplifies our differential equation to:

$$\frac{\mathrm{d}^2\Psi}{\mathrm{d}y^2} + \left[-K^2 + 2\gamma^2 e^{-y} - \gamma^2 e^{-2y}\right]\Psi = 0 \tag{13.2.135}$$

To eliminate the exponential terms in the potential, we perform a change of variable by defining:

$$z = 2\gamma^2 e^{-y} \tag{13.2.136}$$

We map the derivatives from y to z using the chain rule:

$$\frac{\mathrm{d}z}{\mathrm{d}y} = -2\gamma^2 e^{-y} = -z \tag{13.2.137}$$

Hence:

$$\frac{\mathrm{d}\Psi}{\mathrm{d}y} = \frac{\mathrm{d}\Psi}{\mathrm{d}z}\frac{\mathrm{d}z}{\mathrm{d}y} = -z\frac{\mathrm{d}\Psi}{\mathrm{d}z} \tag{13.2.138}$$

For the second derivative:

$$\frac{\mathrm{d}^2\Psi}{\mathrm{d}y^2} = -z\frac{\mathrm{d}}{\mathrm{d}z}\left(-z\frac{\mathrm{d}\Psi}{\mathrm{d}z}\right) = z^2\frac{\mathrm{d}^2\Psi}{\mathrm{d}z^2} + z\frac{\mathrm{d}\Psi}{\mathrm{d}z} \tag{13.2.139}$$

We can also express the exponential coefficients in the brackets in terms of z :

$$2\gamma^2 e^{-y} = z \quad \text{and} \quad \gamma^2 e^{-2y} = \frac{1}{4}z^2 \tag{13.2.140}$$

Substituting these back into the differential equation gives:

$$z^2\frac{\mathrm{d}^2\Psi}{\mathrm{d}z^2} + z\frac{\mathrm{d}\Psi}{\mathrm{d}z} + \left[-K^2 + z - \frac{1}{4}z^2\right]\Psi = 0 \tag{13.2.141}$$

Dividing the entire equation by z^2 :

$$\frac{\mathrm{d}^2\Psi}{\mathrm{d}z^2} + \frac{1}{z}\frac{\mathrm{d}\Psi}{\mathrm{d}z} + \left[-\frac{1}{4} + \frac{\gamma}{z} - \frac{K^2}{z^2}\right]\Psi = 0 \tag{13.2.142}$$

Now we strip out the asymptotic behavior of $\Psi(z)$ at its boundaries ($z \rightarrow +\infty$ and $z \rightarrow 0$):

- As $z \rightarrow +\infty$:

  The dominant term in the brackets is $-1/4$. The equation behaves like:

  $$\Psi'' - \frac{1}{4}\Psi \cong 0 \tag{13.2.143}$$

  which yields a physically acceptable solution of $\Psi(z) \sim e^{-z/2}$.

- As $z \rightarrow 0$:

  The dominant term is $-K^2/z^2$. The equation approaches Euler-Cauchy form:

  $$\Psi'' + \frac{1}{z}\Psi' - \frac{K^2}{z^2}\Psi \cong 0 \tag{13.2.144}$$

  Trying a solution of $\Psi(z) \sim z^S$ yields $s^2 = K^2$, meaning $\Psi(z) \sim z^K$ for a regular solution at the origin.

Combining these behaviors, we define a transformation ansatz (very clever move...!) using an unknown function $w(z)$:

$$\Psi(z) = z^K e^{-z/2} w(z) \tag{13.2.145}$$

Now we evaluate the derivatives of our ansatz $\Psi(z) = z^K e^{-z/2} w(z)$ to substitute them back into the differential equation:

We must compute the new derivatives, we start with the first one:

$$\Psi' = z^{K-1}e^{-z/2}\left(zw'(z) + \left(K - \frac{z}{2}\right)w(z)\right) = z^K e^{-z/2}\left[w' + \left(\frac{K}{z} - \frac{1}{2}\right)w\right] \tag{13.2.146}$$

And the second derivative:

$$\begin{aligned}
\Psi'' &= \frac{\mathrm{d}}{\mathrm{d}z}\left[z^{K-1}e^{-z/2}\left(zw'(z)+\left(K-\tfrac{z}{2}\right)w(z)\right)\right]\\
&= \frac{\mathrm{d}}{\mathrm{d}z}\left[z^{K-1}e^{-z/2}\right]\cdot\left(zw'(z)+\left(K-\tfrac{z}{2}\right)w(z)\right)+z^{K-1}e^{-z/2}\cdot\frac{\mathrm{d}}{\mathrm{d}z}\left(zw'(z)+\left(K-\tfrac{z}{2}\right)w(z)\right)\\
&= \left((K-1)z^{K-2}e^{-z/2}-\frac{1}{2}z^{K-1}e^{-z/2}\right)\left(zw'(z)+\left(K-\tfrac{z}{2}\right)w(z)\right)+z^{K-1}e^{-z/2}\left(w'(z)+zw''(z)-\frac{1}{2}w(z)+\left(K-\tfrac{z}{2}\right)w'(z)\right)\\
&= z^{K-2}e^{-z/2}\left(K-1-\tfrac{z}{2}\right)\left(zw'(z)+\left(K-\tfrac{z}{2}\right)w(z)\right)+z^{K-2}e^{-z/2}\left(z^2w''(z)+z\left(1+K-\tfrac{z}{2}\right)w'(z)-\tfrac{z}{2}w(z)\right)\\
&= z^{K-2}e^{-z/2}\left[\left(K-1-\tfrac{z}{2}\right)zw'(z)+\left(K-1-\tfrac{z}{2}\right)\left(K-\tfrac{z}{2}\right)w(z)+z^2w''(z)+\left(z+Kz-\frac{z^2}{2}\right)w'(z)-\tfrac{z}{2}w(z)\right]\\
&= z^{K-2}e^{-z/2}\left[z^2w''(z)+\left(Kz-z-\frac{z^2}{2}+z+Kz-\frac{z^2}{2}\right)w'(z)+\left(K^2-\frac{Kz}{2}-K+\frac{z}{2}-\frac{Kz}{2}+\frac{z^2}{4}-\frac{z}{2}\right)w(z)\right]\\
&= z^{K-2}e^{-z/2}\left[z^2w''(z)+(2Kz-z^2)w'(z)+\left(K^2-K-Kz+\frac{z^2}{4}\right)w(z)\right]\\
&= z^{K}e^{-z/2}\left[w''(z)+\left(\frac{2K}{z}-1\right)w'(z)+\left(\frac{K(K-1)}{z^2}-\frac{K}{z}+\frac{1}{4}\right)w(z)\right]\\
&= z^{K}e^{-z/2}\left[\frac{\mathrm{d}^2w}{\mathrm{d}z^2}+2\left(\frac{K}{z}-\frac{1}{2}\right)\frac{\mathrm{d}w}{\mathrm{d}z}+\left(\frac{K^2-K}{z^2}-\frac{K}{z}+\frac{1}{4}\right)w\right]
\end{aligned} \tag{13.2.147}$$

Plugging $\Psi, \mathrm{d}\Psi/\mathrm{d}z$, and $\mathrm{d}^2\Psi/\mathrm{d}z^2$ into :

$$\frac{\mathrm{d}^2\Psi}{\mathrm{d}z^2}+\frac{1}{z}\frac{\mathrm{d}\Psi}{\mathrm{d}z}+\left[-\frac{1}{4}+\frac{\gamma}{z}-\frac{K^2}{z^2}\right]\Psi=0 \tag{13.2.148}$$

and dividing out the common factor $z^K e^{-z/2}$:

$$\begin{aligned}
\left[\frac{\mathrm{d}^2w}{\mathrm{d}z^2}+\left(\frac{2K}{z}-1\right)\frac{\mathrm{d}w}{\mathrm{d}z}+\left(\frac{K^2-K}{z^2}-\frac{K}{z}+\frac{1}{4}\right)w\right]+\frac{1}{z}\left[\frac{\mathrm{d}w}{\mathrm{d}z}+\left(\frac{K}{z}-\frac{1}{2}\right)w\right]\\
+\left[-\frac{1}{4}+\frac{\gamma}{z}-\frac{K}{z^2}\right]
\end{aligned} \tag{13.2.149}$$

Grouping like terms:

- $\frac{\mathrm{d}^2w}{\mathrm{d}z^2}$ term :

$$1 \tag{13.2.150}$$

- $\frac{\mathrm{d}w}{\mathrm{d}z}$ term:

$$\left(\frac{2K}{z}-1\right)+\frac{1}{z}=\frac{2K+1}{z}-1 \tag{13.2.151}$$

- w term:

$$\left(\frac{K^2-K}{z^2}-\frac{K}{z}+\frac{1}{4}\right)+\left(\frac{K}{z^2}-\frac{1}{2z}\right)+\left(-\frac{1}{4}+\frac{\gamma}{z}-\frac{K^2}{z^2}\right)=\frac{\gamma-K-\frac{1}{2}}{z} \tag{13.2.152}$$

This simplifies the system down to:

$$\frac{\mathrm{d}^2w}{\mathrm{d}z^2}+\left(\frac{2K+1}{z}-1\right)\frac{\mathrm{d}w}{\mathrm{d}z}+\left(\frac{\gamma-K-\frac{1}{2}}{z}\right)w=0 \tag{13.2.153}$$

Multiplying the entire expression by z :

$$z\frac{\mathrm{d}^2w}{\mathrm{d}z^2}+(2K+1-z)\frac{\mathrm{d}w}{\mathrm{d}z}-\left(K+\frac{1}{2}-\gamma\right)w=0 \tag{13.2.154}$$

The final expression matches exactly with Kummer's Confluent Hypergeometric Equation:

$$\boxed{z\frac{\mathrm{d}^2w}{\mathrm{d}z^2}+(b-z)\frac{\mathrm{d}w}{\mathrm{d}z}-aw=0} \tag{13.2.155}$$

where the characteristic parameters are:

$$\begin{aligned} a&=K+\frac{1}{2}-\gamma \\ b&=2K+1 \end{aligned} \tag{13.2.156}$$

So this is the one-dimensional Schrödinger equation with a Morse potential written into something named the "Kummer's differential equation" (also known as the "confluent hypergeometric equation").

The Kummer's function of the first kind M is a generalized hypergeometric series introduced by Kummer in 11837 (holocene calendar) and a solution of this differential equation given by:

$$M(a,b,z)=\sum_{n=0}^{+\infty}\frac{a^{(n)}z^n}{b^{(n)}n!}={}_1F_1(a;b;z) \tag{13.2.157}$$

where $x^{(n)}$) is the "rising factorial" (sometimes named the "Pochhammer function", "Pochhammer polynomial", "ascending factorial", "rising sequential product", or "upper factorial") is defined as:

$$\begin{aligned} x^{(n)}=x^{\bar{n}}&=\overbrace{x(x+1)(x+2)\cdots(x+n-1)}^{n\text{ factors}} \\ &=\prod_{k=1}^{n}(x+k-1)=\prod_{k=0}^{n-1}(x+k). \end{aligned} \tag{13.2.158}$$

The first few rising factorials are as follows:

$$\begin{aligned} &x^{(0)} && =1 \\ &x^{(1)} && =x \\ &x^{(2)}=x(x+1) && =x^2+x \\ &x^{(3)}=x(x+1)(x+2) && =x^3+3x^2+2x \\ &x^{(4)}=x(x+1)(x+2)(x+3) && =x^4+6x^3+11x^2+6x \end{aligned} \tag{13.2.159}$$

Remark

Laguerre polynomials can be defined by confluent hypergeometric functions and Kummer's transformation as:

$$L_n^{(\alpha)}(x):=\binom{n+\alpha}{n}M(-n,\alpha+1,x) \tag{13.2.160}$$

That said let us show the the Kummer's fuction is indeed a solution of the Kummer's differential equation! First we have:

$$y(x)=\sum_{n=0}^{+\infty}\frac{a^{(n)}x^n}{b^{(n)}n!}=1+\sum_{n=1}^{+\infty}\frac{a^{(n)}x^n}{b^{(n)}n!} \tag{13.2.161}$$

Hence for the third term of the Kummer's differential equation:

$$ay(x)=a+\sum_{n=1}^{+\infty}\frac{a\cdot a^{(n)}x^n}{b^{(n)}n!} \tag{13.2.162}$$

And we also have for the second term of the Kummer's differential equation:

$$\begin{aligned}(b-x)y'(x)=(b-x)\sum_{n=1}^{+\infty}\frac{a^{(n)}nx^{n-1}}{b^{(n)}n!}&=\sum_{n=1}^{+\infty}b\cdot\frac{a^{(n)}nx^{n-1}}{b^{(n)}n!}-\sum_{n=1}^{+\infty}x\cdot\frac{a^{(n)}nx^{n-1}}{b^{(n)}n!}\\&=\sum_{n=1}^{+\infty}\frac{a^{(n)}x^{n-1}}{(b+1)^{(n-1)}(n-1)!}-\sum_{n=1}^{+\infty}\frac{a^{(n)}x^n}{b^{(n)}(n-1)!}\\&=a+\sum_{n=1}^{+\infty}\frac{a^{(n+1)}x^n}{(b+1)^{(n)}n!}-\sum_{n=1}^{+\infty}\frac{a^{(n)}x^n}{b^{(n)}(n-1)!}\end{aligned} \tag{13.2.163}$$

And for the first term of the Kummer's differential equation:

$$xy''(x)=\sum_{n=2}^{+\infty}x\cdot\frac{a^{(n)}n(n-1)x^{n-2}}{b^{(n)}n!}=\sum_{n=2}^{+\infty}\frac{a^{(n)}x^{n-1}}{b^{(n)}(n-2)!}=\sum_{n=1}^{+\infty}\frac{a^{(n+1)}x^n}{b^{(n+1)}(n-1)!} \tag{13.2.164}$$

So we can finally check:

$$\begin{aligned}&xy''(x)+(b-x)y'(x)-ay(x)\\&=\sum_{n=1}^{+\infty}\frac{a^{(n+1)}x^n}{b^{(n+1)}(n-1)!}+\left(a+\sum_{n=1}^{+\infty}\frac{a^{(n+1)}x^n}{(b+1)^{(n)}n!}-\sum_{n=1}^{+\infty}\frac{a^{(n)}x^n}{b^{(n)}(n-1)!}\right)-\left(a+\sum_{n=1}^{+\infty}\frac{a\cdot a^{(n)}x^n}{b^{(n)}n!}\right)\\&=\sum_{n=1}^{+\infty}\left(\frac{a^{(n+1)}}{b^{(n+1)}(n-1)!}+\frac{a^{(n+1)}}{(b+1)^{(n)}n!}-\frac{a^{(n)}}{b^{(n)}(n-1)!}-\frac{a\cdot a^{(n)}}{b^{(n)}n!}\right)x^n\\&=\sum_{n=1}^{+\infty}((a+n)n+(a+n)b-(b+n)n-a(b+n))\frac{a^{(n)}}{b^{(n+1)}n!}x^n=0\end{aligned} \tag{13.2.165}$$

And that proves our ansatz!

For the wave function $\Psi(\zeta)$ to remain physically wellbehaved (square-normalizable) as $z\rightarrow+\infty$, the series expansion of $M(a,b,z)$ must terminate into a finite polynomial. This truncation occurs if and only if the parameter a is a non-positive integer, i.e.:

$$a=-n \quad \text{where } n=0,1,2,\ldots \tag{13.2.166}$$

To prove that the confluent hypergeometric function of the first kind, $M(a,b,z)$ (also denoted as ${}_1F_1(a;b;z)$), terminates into a finite polynomial if and only if the parameter a is a non-positive integer, we must analyze its series expansion.

The series definition of the Kummer function is given for reminder by:

$$M(a,b,z)=\sum_{n=0}^{+\infty}\frac{a^{(n)}}{b^{(n)}}\frac{z^n}{n!} \tag{13.2.167}$$

where $x^{(n)}$ denotes the Pochhammer symbol (or rising factorial), defined for $n \geq 1$ as:

$$x^{(n)}=x(x+1)(x+2)\cdots(x+n-1) \tag{13.2.168}$$

with the convention that $x^{(0)}=1$.

Let $a=-m$, where m is a non-negative integer $(m \in \{0,1,2,\ldots\})$.

Remark

We assume throughout this proof that b is not a non-positive integer ($b \notin \mathbb{Z}_{\leq 0}$), otherwise the terms $b^{(n)}$ would vanish in the denominator, making the function undefined.

Let's examine the behavior of the numerator's Pochhammer symbol $a^n=(-m)^{(n)}$ as n increases:

- For $n \leq m$, the product contains only strictly negative terms, so $(-m)_n \neq 0$.
- For $n=m+1$, the product becomes:

$$(-m)_{m+1}=(-m)(-m+1)(-m+2)\cdots(-m+m+1-1)=0 \tag{13.2.169}$$

- For any index $n>m$, the term $(-m+m)=0$ will always be a factor inside the product expansion!

Because $a^{(n)}=0$ for all $n \geq m+1$, every coefficient in the series beyond $n=m$ identically vanishes. The infinite series truncates in this case to a finite sum:

$$M(-m,b,z)=\sum_{n=0}^{m}\frac{(-m)^{(n)}}{b^n}\frac{z^n}{n!} \tag{13.2.170}$$

Remember that in the confluent hypergeometric equation given by:

$$z\frac{\mathrm{d}^2w}{\mathrm{d}z^2}+(b-z)\frac{\mathrm{d}w}{\mathrm{d}z}-aw=0 \tag{13.2.171}$$

we have in our case (see earlier above):

$$a=K+\frac{1}{2}-\gamma=\frac{\sqrt{2\mu(D-E)}}{\hbar\alpha}+\frac{1}{2}-\frac{\sqrt{2\mu D}}{\hbar\alpha} \tag{13.2.172}$$

Setting $a:=-n$ we therefore have:

$$n=\frac{\sqrt{2\mu D}}{\hbar\alpha}-\frac{\sqrt{2\mu(D-E)}}{\hbar\alpha}-\frac{1}{2} \tag{13.2.173}$$

$$\begin{aligned}
n&=\frac{\sqrt{2\mu D}}{\hbar\alpha}-\frac{\sqrt{2\mu(D-E)}}{\hbar\alpha}-\frac{1}{2}\\
\Rightarrow \frac{\sqrt{2\mu(D-E)}}{\hbar\alpha}&=\frac{\sqrt{2\mu D}}{\hbar\alpha}-n-\frac{1}{2}\\
\Rightarrow \sqrt{2\mu(D-E)}&=\sqrt{2\mu D}-\hbar\alpha\left(n+\frac{1}{2}\right)\\
\Rightarrow 2\mu(D-E)&=\left[\sqrt{2\mu D}-\hbar\alpha\left(n+\frac{1}{2}\right)\right]^2\\
\Rightarrow D-E&=\frac{1}{2\mu}\left[\sqrt{2\mu D}-\hbar\alpha\left(n+\frac{1}{2}\right)\right]^2\\
\Rightarrow E&=D-\frac{1}{2\mu}\left[\sqrt{2\mu D}-\hbar\alpha\left(n+\frac{1}{2}\right)\right]^2
\end{aligned} \tag{13.2.174}$$

Finally:

$$E_n=D-\frac{1}{2\mu}\left[\sqrt{2\mu D}-\hbar\alpha\left(n+\frac{1}{2}\right)\right]^2 \tag{13.2.175}$$

The eigenenergies in the initial variables is often written as:

$$\boxed{E_n=h\nu_0\left(n+\frac{1}{2}\right)-\frac{\left[h\nu_0\left(n+\frac{1}{2}\right)\right]^2}{4D}} \tag{13.2.176}$$

where n is the vibrational quantum number and ν_0 has units of frequency. The latter is mathematically related to the particle mass, m, and the Morse constants via:

$$\nu_0=\frac{\alpha}{2\pi}\sqrt{\frac{2D}{\mu}} \tag{13.2.177}$$

Whereas the energy spacing between vibrational levels in the quantum harmonic oscillator is constant at $h\nu_0$, the energy between adjacent levels decreases with increasing v in the Morse oscillator. Mathematically, the spacing of Morse levels is:

$$E_{n+1}-E_n=h\nu_0-\frac{(n+1)\left(h\nu_0\right)^2}{2D_e} \tag{13.2.178}$$

This trend matches the inharmonicity found in real molecules. However, this equation fails above some value of n_m where $E\left(n_m+1\right)-E\left(n_m\right)$ is calculated to be zero or negative. Specifically:

$$n_m=\frac{2D_e-h\nu_0}{h\nu_0} \quad \text{(integer part only)} \tag{13.2.179}$$

This failure is due to the finite number of bound levels in the Morse potential, and some maximum n_m that remains bound. For energies above n_m, all the possible energy levels are allowed and the equation for E_n is no longer valid.

Remark

In molecular physics, the real potential energy $V(r)$ of a vibrating molecule is often approximated using analytical Morse functions. The "Herzberg anomaly" highlights specific regions, generally near the middle of the potential well, where these mathematical approximations systematically intersect and diverge from the actual experimental potential curve. The anomaly was originally observed through isotopic and vibrational analyses detailed by Nobel laureate Gerhard Herzberg in his seminal book *Spectra of Diatomic Molecules* (originally published in the 11930 (holocence calendar).

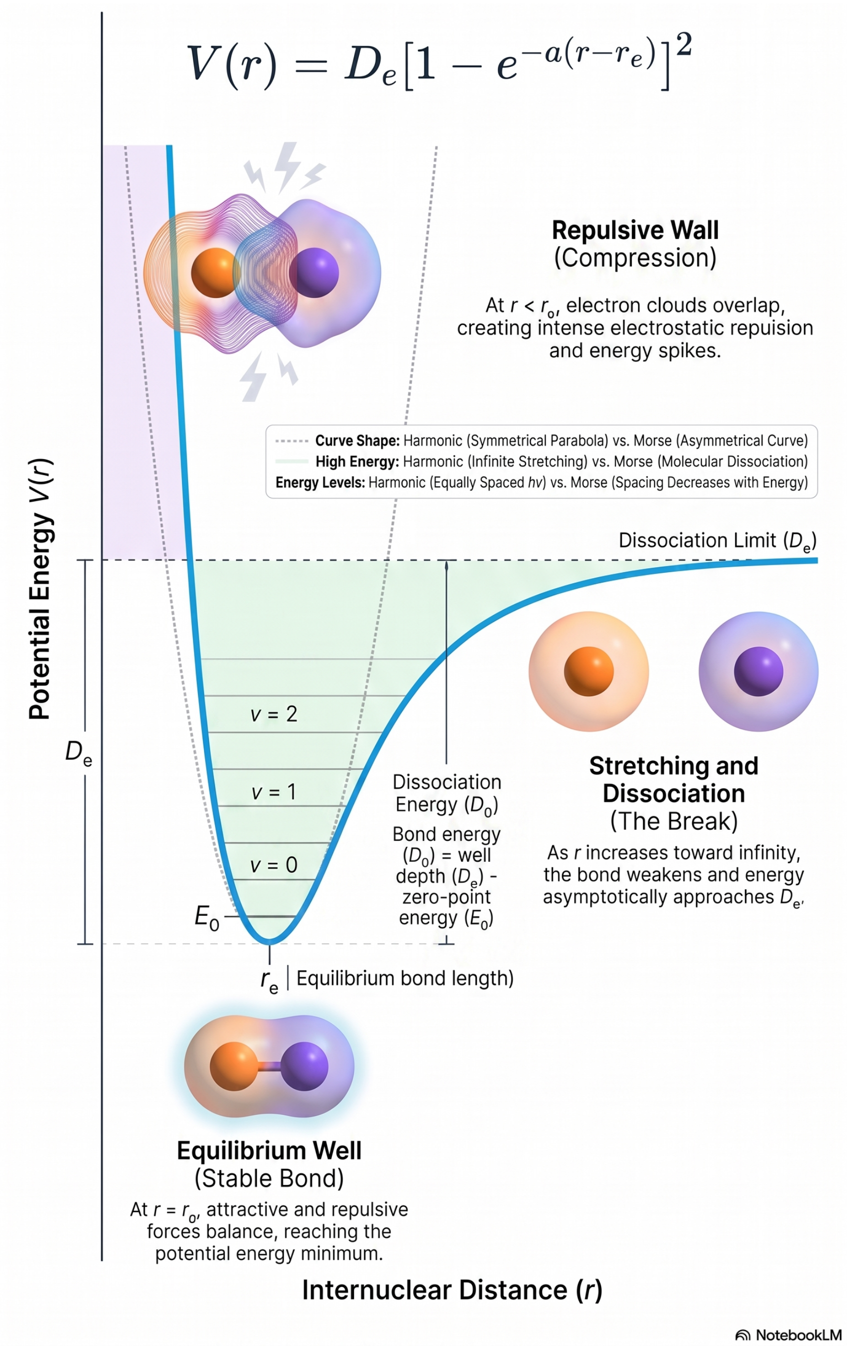
Here is a table of a few values (among a thousands of others) of D_e and α according to:

$$V(r)=D_e\left(\left[1-e^{-\alpha(r-r_0)}\right]^2-1\right) \tag{13.2.180}$$

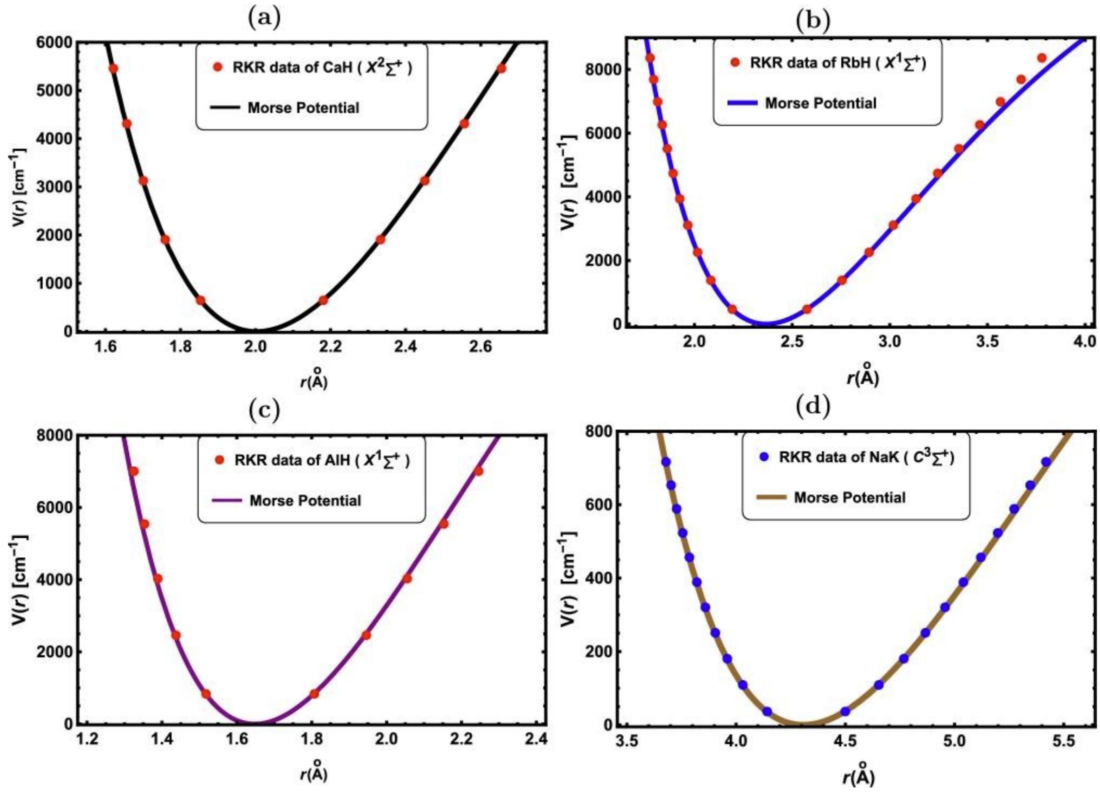
and taken from [429]:

Molecule	**Type of bond (state)**	**Dissociation Energy** D_e $[\text{cm}^{-1}]$	**Morse parameter** α	**Equilibrium Bond length** r_0 $[\text{cm}^{-1}]$
$^1\text{H}_2$	j $^3\Delta_g 3d\delta$	1900	1.692	$1.0545\cdot 10^{-10}$
$^{12}\text{C}_2$	g $^3\Delta_\text{g}$	$6.6\cdot 10^{-6}$	0.019	$1.3579\cdot 10^{-10}$
$^{40}\text{Ca}_2$	A $^1\Sigma_u^+$	$0.459\cdot 10^{-7}$	$0.3104\cdot 10^{-3}$	$1.573\cdot 10^{-10}$
$^{35}\text{Cl}_2$	X $l_{\Sigma_g^+}$	$1.86\cdot 10^{-7}$	0.00149	$1.9879\cdot 10^{-10}$
$^{63}\text{Cu}_2$	B $^1\Sigma_u^+$	$6.30\cdot 10^{-8}$	$0.606\cdot 10^{-3}$	$2.3276\cdot 10^{-10}$
$^{19}\text{F}_2$	X $1^+_{\Sigma_g^+}$	$3.30\cdot 10^{-6}$	0.013847	$1.41193\cdot 10^{-10}$
$^1\text{H}^{35}\text{Cl}$	X $^1\Sigma^+$	$5.3194\cdot 10^{-4}$	0.307181	$1.274552\cdot 10^{-10}$
$^4\text{He}_2$	s 3_{Π_g}	$5.1\cdot 10^{-4}$	0.2234	$1.0806\cdot 10^{-10}$
$^1\text{H}^{19}\text{F}$	X $1_\Sigma+$	$21.51\cdot 10^{-4}$	0.798	$0.916808\cdot 10^{-10}$
$^{127}\text{I}_2$	B $3_{\Pi_{0+u}}$	$5.43\cdot 10^{-9}$	$0.1582\cdot 10^{-3}$	$3.0247\cdot 10^{-10}$
$^{39}\text{K}_2$	B 1_{Π_u}	$8.25\cdot 10^{-8}$	0.00024	4.2125
$^{39}\text{K}^{35}\text{Cl}$	X 1_{Σ^+}	$1.0874\cdot 10^{-7}$	$0.7899\cdot 10^{-3}$	$2.66665\cdot 10^{-10}$
$^7\text{Li}_2$	A $^1\Sigma_u^+$	$7.54\cdot 10^{-6}$	0.00540	3.1079
$^{24}\text{Mg}_2$	A $^1\Sigma_u^+$	$0.334\cdot 10^{-6}$	0.001316	$3.08186\cdot 10^{-10}$
$^{14}\text{N}_2$	B′ $3_{\Sigma_u^-}$	$5.56\cdot 10^{-6}$	0.01666	$1.2784\cdot 10^{-10}$
$^{23}\text{Na}^{35}\text{Cl}$	X $^1\Sigma^+$	$3.1202\cdot 10^{-7}$	0.00162482	$2.360795\cdot 10^{-10}$
$^{16}\text{O}_2$	p $^1\phi_\text{u}$	$4.5\cdot 10^{-6}$	0.014	$1.374\cdot 10^{-10}$
$^{31}\text{P}_2$	N $^1\Sigma^+$	$3.1\cdot 10^{-7}$	0.00504	$1.9099\cdot 10^{-10}$
$^{208}\text{Pb}^{16}\text{O}$	C* 1	$0.25\cdot 10^{-6}$	0.0018	$2.14\cdot 10^{-10}$
$^{32}\text{S}_2$	C $3_{\Sigma_u^-}$	$2.17\cdot 10^{-7}$	0.00138	$1.8100\cdot 10^{-10}$
$^{28}\text{Si}_2$	0 $^3\Sigma_u^-$	$0.050\cdot 10^{-5}$	0.003	$2.327\cdot 10^{-10}$
$^{28}\text{Si}^{16}\text{O}$	E Σ^+	$1.7\cdot 10^{-6}$	0.00555	$1.5561\cdot 10^{-10}$
...	...	...	...	...

Table 13.3: **Some values of parameters of the Morse potential**


$V(r) = D_e[1 - e^{-a(r-r_e)}]^2$
Repulsive Wall
(Compression)
At $r < r_o$, electron clouds overlap, creating intense electrostatic repulsion and energy spikes.
Curve Shape: Harmonic (Symmetrical Parabola) vs. Morse (Asymmetrical Curve)
High Energy: Harmonic (Infinite Stretching) vs. Morse (Molecular Dissociation)
Energy Levels: Harmonic (Equally Spaced $h\nu$) vs. Morse (Spacing Decreases with Energy)
Dissociation Limit (D_e)
Potential Energy $V(r)$
D_e
$v = 2$
$v = 1$
$v = 0$
E_0
Dissociation Energy (D_0)
Bond energy (D_0) = well depth (D_e) - zero-point energy (E_0)
Stretching and Dissociation
(The Break)
As r increases toward infinity, the bond weakens and energy asymptotically approaches D_e.
r_e | Equilibrium bond length)
Equilibrium Well
(Stable Bond)
At $r = r_o$, attractive and repulsive forces balance, reaching the potential energy minimum.
Internuclear Distance (r)
NotebookLM

And finally here are some comparisons (among a few other available in [310]) of theoretical Morse potential curves with experimental ones:



~90 %

★★★☆☆

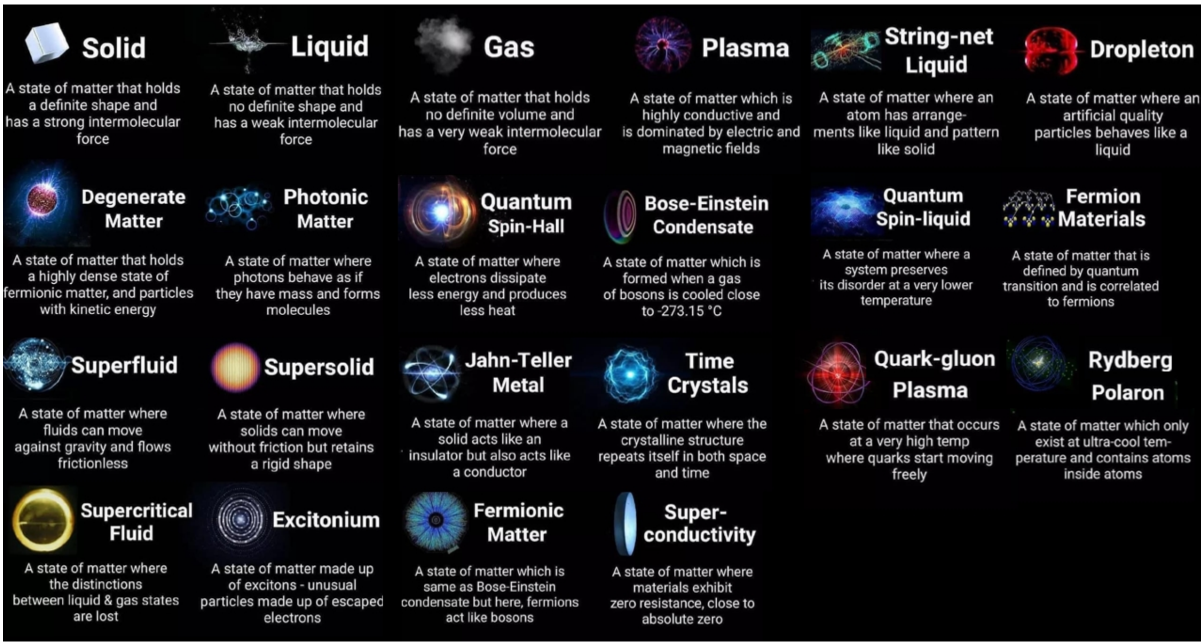
23 votes, 64.35%

13.3. Analytical Chemistry

CHEMISTRY is a very complex n-body science that mathematics can not explained without the input of numerical computer simulations or approximations regarding the use of quantum theory (see Atomistic chapter page 4612). Until these tools are powerful enough and accessible to everyone, chemistry remains a primarily experimental science based on the observation of different properties of matter and we would like here give some very important definitions (which we find also elsewhere in other fields as chemistry).

Analytical chemistry is concerned with the chemical characterization of matter and the answer to two important questions: what is it (qualitative analysis) and how much is it (quantitative analysis). Chemicals make up everything we use or consume, and knowledge of the chemical composition of many substances is important in our daily lives. Analytical chemistry plays an important role in nearly all aspects of chemistry, for example, agricultural, clinical, environmental, forensic, manufacturing, metallurgical, and pharmaceutical chemistry. The nitrogen content of a fertilizer determines its value. Foods must be analysed for contaminants (e.g., pesticide residues) and for essential nutrients (e.g., vitamin content). The air we breathe must be analysed for toxic gases (e.g., carbon monoxide). Blood glucose must be monitored in diabetics (and, in fact, most diseases are diagnosed by chemical analysis). The presence of trace elements from gun powder on a perpetrator's hand will prove a gun was fired by that hand. The quality of manufactured products often depends on proper chemical proportions, and measurement of the constituents is a necessary part of quality assurance. The carbon content of steel will influence its quality. The purity of drugs will influence their efficacy.

In this section, we will focus only on the mathematical tools and techniques for performing these different types of analyses.



Solid
A state of matter that holds a definite shape and has a strong intermolecular force

Liquid
A state of matter that holds no definite shape and has a weak intermolecular force

Gas
A state of matter that holds no definite volume and has a very weak intermolecular force

Plasma
A state of matter which is highly conductive and is dominated by electric and magnetic fields

String-net Liquid
A state of matter where an atom has arrangements like liquid and pattern like solid

Dropleton
A state of matter where an artificial quality particles behaves like a liquid

Degenerate Matter
A state of matter that holds a highly dense state of fermionic matter, and particles with kinetic energy

Photonic Matter
A state of matter where photons behave as if they have mass and forms molecules

Quantum Spin-Hall
A state of matter where electrons dissipate less energy and produces less heat

Bose-Einstein Condensate
A state of matter which is formed when a gas of bosons is cooled close to -273.15 °C

Quantum Spin-liquid
A state of matter where a system preserves its disorder at a very lower temperature

Fermion Materials
A state of matter that is defined by quantum transition and is correlated to fermions

Superfluid
A state of matter where fluids can move against gravity and flows frictionless

Supersolid
A state of matter where solids can move without friction but retains a rigid shape

Jahn-Teller Metal
A state of matter where a solid acts like an insulator but also acts like a conductor

Time Crystals
A state of matter where the crystalline structure repeats itself in both space and time

Quark-gluon Plasma
A state of matter that occurs at a very high temp where quarks start moving freely

Rydberg Polaron
A state of matter which only exist at ultra-cool temperature and contains atoms inside atoms

Supercritical Fluid
A state of matter where the distinctions between liquid & gas states are lost

Excitonium
A state of matter made up of excitons - unusual particles made up of escaped electrons

Fermionic Matter
A state of matter which is same as Bose-Einstein condensate but here, fermions act like bosons

Super-conductivity
A state of matter where materials exhibit zero resistance, close to absolute zero

Figure 13.28: **Most common classical and exotic states of matter**

Remark

All matter is made of chemicals. Every single thing you see, touch, taste, smell and even hear is made of chemicals or observable as a result of chemicals. Your body, clothes, and food are made of chemicals. Everything that isn't energy (pure particles) is matter and therefore made of chemicals. All chemicals and chemical compounds are, or are made up of, the same known chemicals found on the period table. Whether or not people can pronounce a chemical is irrelevant for is dangerousness (the chemical name of water is "dihydrogen monoxide").

Caution! Whether is harnessed from Nature or developed in laboratory, a chemical's make up and the way it behaves is going TO BE EXACTLY THE SAME. The makeup of the chemical and the way it behaves will not change based on where it's made. «*Chemical Free*» argument is a LIE. Whether a chemical is man-made or natural tells us precisely nothing about how dangerous it is (the reader may take a look to the famous and friendly Manchineel tree...).

Definitions (#13.3.548):

D1. A "subjective property" is a property based on personal / individual printing, for example: beauty, sympathy, color, utility, etc.

D2. An "objective property" is an experienced property (which can not be contradicted), for example: mass, volume, shape, etc.

D3. A "qualitative property" is a descriptive property given using words. For example: oval, magnetic, conductive, etc.

D4. A "quantitative property" is a property that can be measured. For example: mass, volume, density, etc.

D5. A "characteristic property" is an exclusive property that identifies a pure substance. It does not change even if it is physically transformed material, for example: its density, its boiling point, its melting point, etc.

> **Remark**
>
> We know about $2,000,000$ different pure substances in the early 121st century (holocene calendar), that is to say ... there is quite lot of work behind it.

D6. We name "compound bodies", the bodies, that subjected to chemical processes, restore their components in the form of pure substances.

D7. If we make the separation of mixtures and the decomposition of compositions, we finally get the bodies that are non-decomposable by conventional chemical methods; we name them "elements" or "simple bodies".

D8. The smallest part of a chemical combination yet having all of the properties thereof is the "molecule" of this combination. The smallest part of an element or simple body is the "atom" of that element.

Although there are just over 100 elements, tens of millions of chemical compounds result from different combinations of these elements. Each compound has a specific composition and possesses definite chemical and physical properties by which we can distinguish it from all other compounds. And, of course, there are innumerable ways to combine elements and compounds to form different mixtures. A summary of how to distinguish between the various major classifications of matter is shown in the figure below:

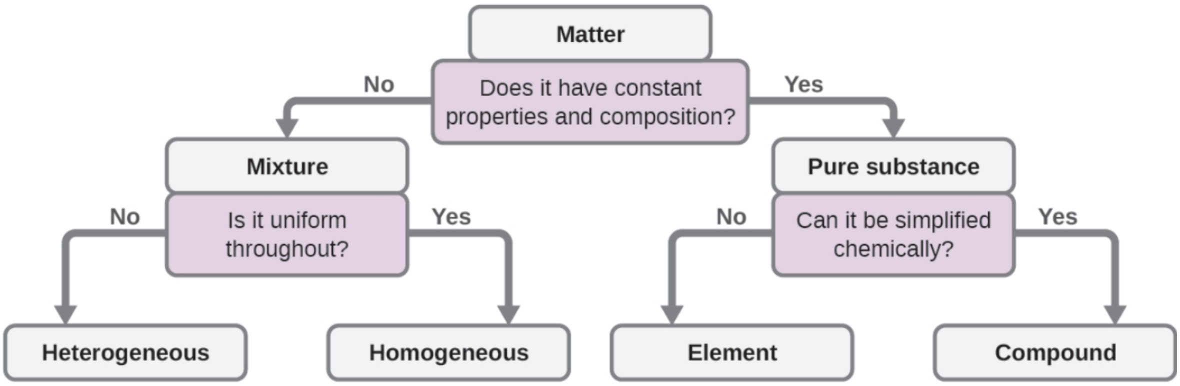

Matter
Does it have constant properties and composition?
No
Yes
Mixture
Is it uniform throughout?
No
Yes
Pure substance
Can it be simplified chemically?
No
Yes
Heterogeneous
Homogeneous
Element
Compound

Figure 13.29: Depending on its properties, a given substance can be classified as a homogeneous mixture, a heterogeneous mixture, a compound, or an element (source: OpenStax)

Let us also give some reminders of what we saw at the very start of the Mechanics chapter:

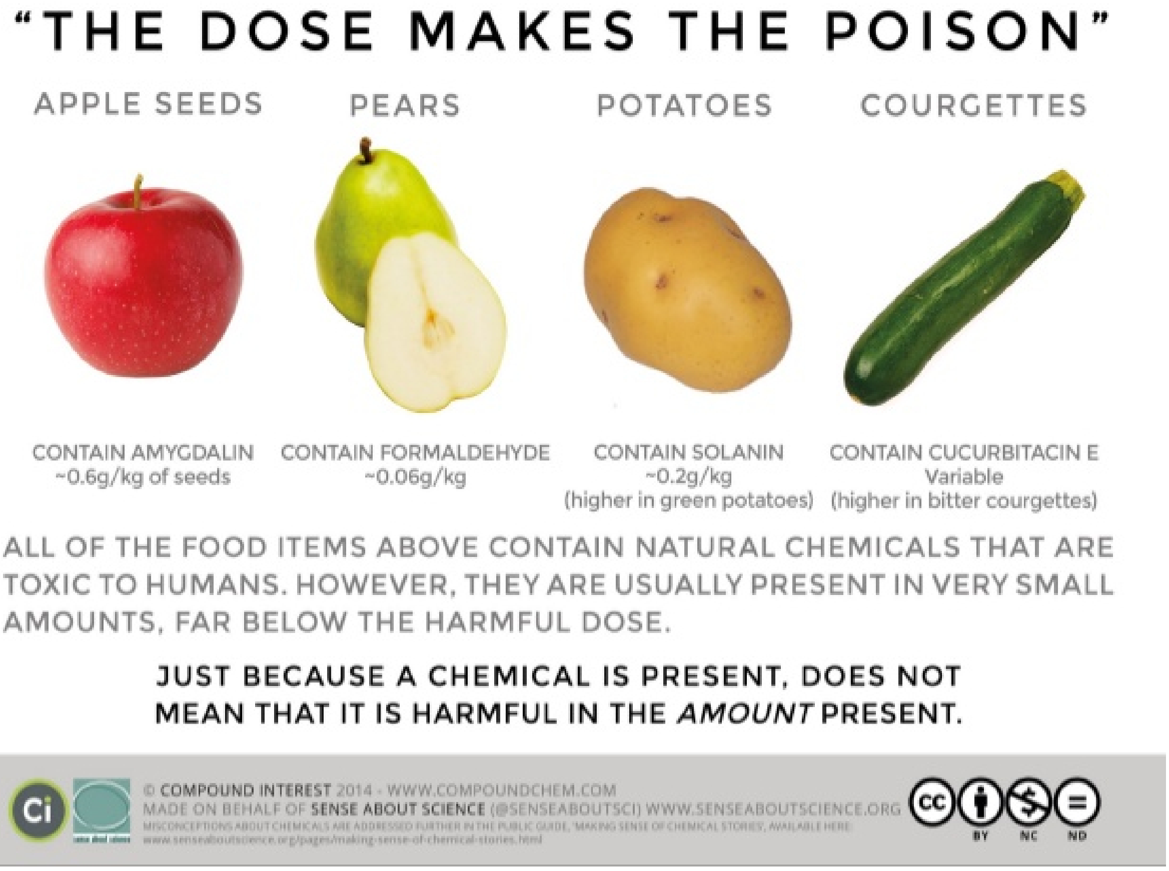
1. A mixture is named "heterogeneous" in chemistry if the components are immediately discernible to the naked eye or through the microscope.

2. A mixture is said to be "homogeneous" in chemistry if the components are not discernible to the naked eye or through the microscope.
3. A system or body is said to be "isotropic" if it has identical values of a property in all directions otherwise it is said "anisotropic".

Also because there is a lot of biased knowledge, the study of chemistry and quantum physics and atomistic should help the reader to understand why:

> *All things are poisons, for there is nothing without poisonous qualities. It is only the dose which makes a thing poison*[5]*!*
>
> Paracelsus

quote 13.3:6231.#75


"THE DOSE MAKES THE POISON"
APPLE SEEDS
PEARS
POTATOES
COURGETTES
CONTAIN AMYGDALIN
~0.6g/kg of seeds
CONTAIN FORMALDEHYDE
~0.06g/kg
CONTAIN SOLANIN
~0.2g/kg
(higher in green potatoes)
CONTAIN CUCURBITACIN E
Variable
(higher in bitter courgettes)

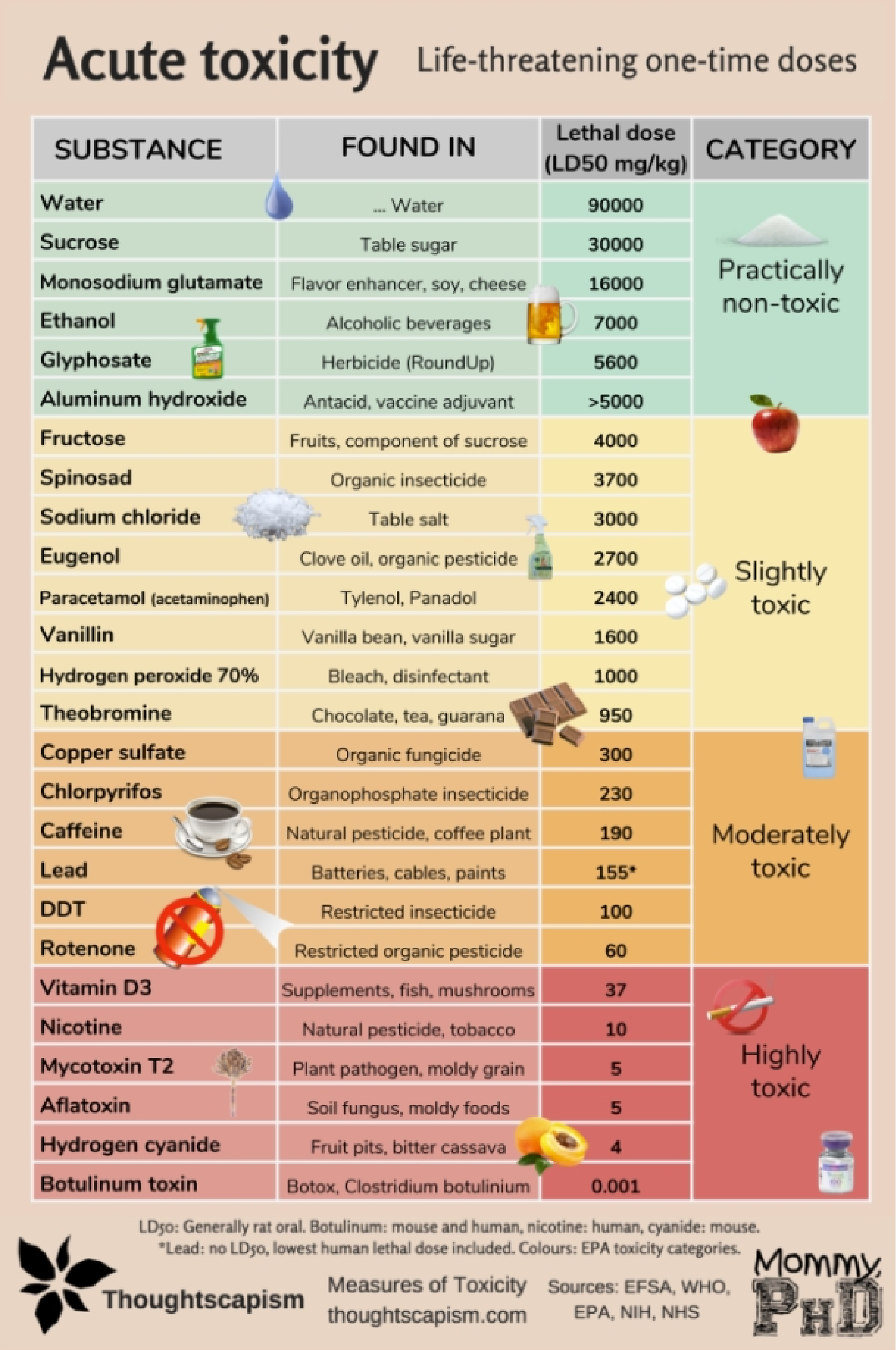
ALL OF THE FOOD ITEMS ABOVE CONTAIN NATURAL CHEMICALS THAT ARE TOXIC TO HUMANS. HOWEVER, THEY ARE USUALLY PRESENT IN VERY SMALL AMOUNTS, FAR BELOW THE HARMFUL DOSE.

JUST BECAUSE A CHEMICAL IS PRESENT, DOES NOT MEAN THAT IT IS HARMFUL IN THE *AMOUNT* PRESENT.

© COMPOUND INTEREST 2014 - WWW.COMPOUNDCHEM.COM
MADE ON BEHALF OF SENSE ABOUT SCIENCE (@SENSEABOUTSCI) WWW.SENSEABOUTSCIENCE.ORG
MISCONCEPTIONS ABOUT CHEMICALS ARE ADDRESSED FURTHER IN THE PUBLIC GUIDE, 'MAKING SENSE OF CHEMICAL STORIES', AVAILABLE HERE:
www.senseaboutscience.org/pages/making-sense-of-chemical-stories.html
BY NC ND

---

[5] A famous scientific experiment has provided evidence that at the contrary to the intuition that increasing dose makes something more poisonous, there exist sometimes non-monotonic (see section Functional Analysis page 2203) dose response curves (NMDRCs) demonstrated for natural hormones and endocrine disrupting chemicals (EDCs) like Bisphenonal A!



Acute toxicity

Life-threatening one-time doses

SUBSTANCE	FOUND IN	Lethal dose (LD50 mg/kg)	CATEGORY
Water	... Water	90000	Practically non-toxic
Sucrose	Table sugar	30000	
Monosodium glutamate	Flavor enhancer, soy, cheese	16000	
Ethanol	Alcoholic beverages	7000	
Glyphosate	Herbicide (RoundUp)	5600	
Aluminum hydroxide	Antacid, vaccine adjuvant	>5000	
Fructose	Fruits, component of sucrose	4000	Slightly toxic
Spinosad	Organic insecticide	3700	
Sodium chloride	Table salt	3000	
Eugenol	Clove oil, organic pesticide	2700	
Paracetamol (acetaminophen)	Tylenol, Panadol	2400	
Vanillin	Vanilla bean, vanilla sugar	1600	
Hydrogen peroxide 70%	Bleach, disinfectant	1000	
Theobromine	Chocolate, tea, guarana	950	
Copper sulfate	Organic fungicide	300	Moderately toxic
Chlorpyrifos	Organophosphate insecticide	230	
Caffeine	Natural pesticide, coffee plant	190	
Lead	Batteries, cables, paints	155*	
DDT	Restricted insecticide	100	
Rotenone	Restricted organic pesticide	60	
Vitamin D3	Supplements, fish, mushrooms	37	Highly toxic
Nicotine	Natural pesticide, tobacco	10	
Mycotoxin T2	Plant pathogen, moldy grain	5	
Aflatoxin	Soil fungus, moldy foods	5	
Hydrogen cyanide	Fruit pits, bitter cassava	4	
Botulinum toxin	Botox, Clostridium botulinium	0.001	

LD50: Generally rat oral. Botulinum: mouse and human, nicotine: human, cyanide: mouse.
*Lead: no LD50, lowest human lethal dose included. Colours: EPA toxicity categories.

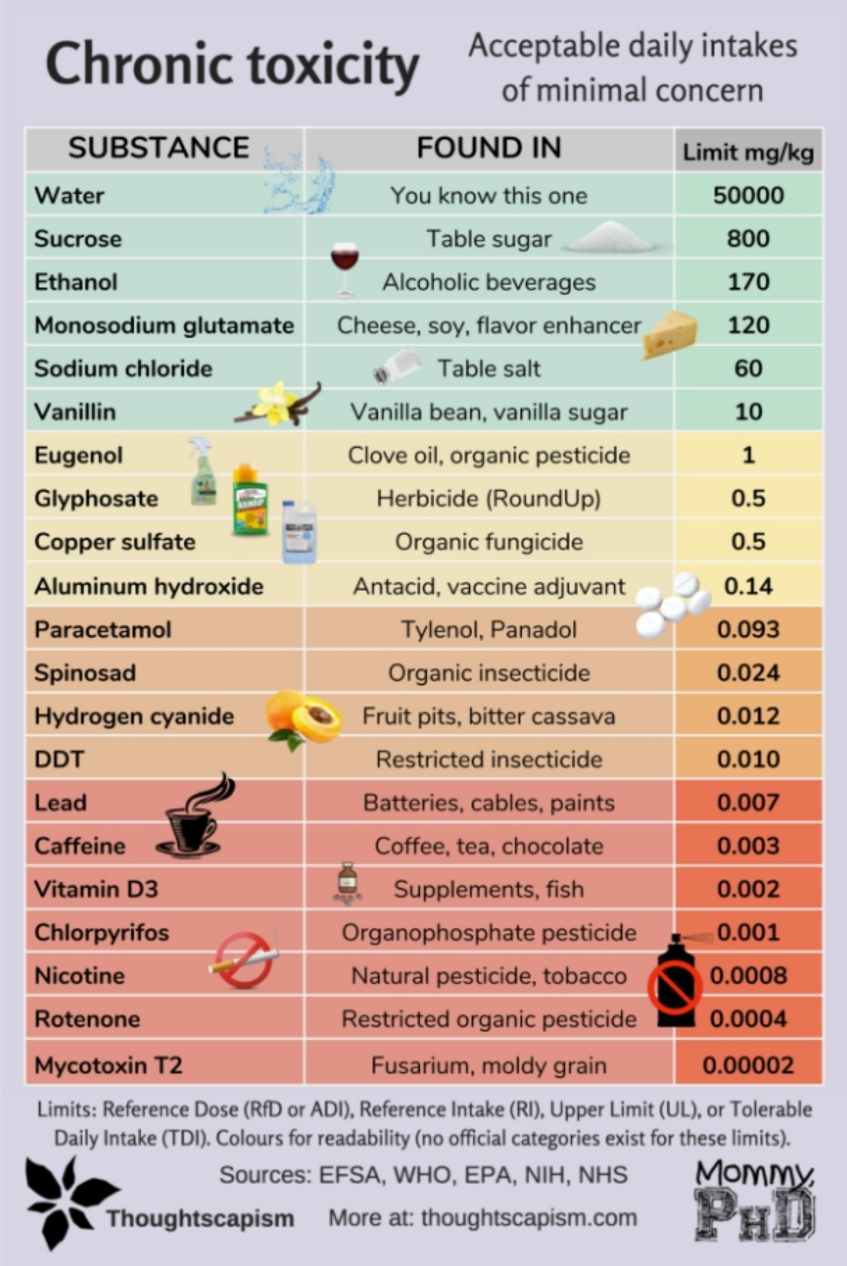
Thoughtscapism

Measures of Toxicity
thoughtscapism.com

Sources: EFSA, WHO, EPA, NIH, NHS

Mommy PhD

Figure 13.30: **Acute Toxicity of some ingredients/molecules (authors: Iida Ruishalme, Alison Bernstein)**



Chronic toxicity

Acceptable daily intakes of minimal concern

SUBSTANCE	FOUND IN	Limit mg/kg
Water	You know this one	50000
Sucrose	Table sugar	800
Ethanol	Alcoholic beverages	170
Monosodium glutamate	Cheese, soy, flavor enhancer	120
Sodium chloride	Table salt	60
Vanillin	Vanilla bean, vanilla sugar	10
Eugenol	Clove oil, organic pesticide	1
Glyphosate	Herbicide (RoundUp)	0.5
Copper sulfate	Organic fungicide	0.5
Aluminum hydroxide	Antacid, vaccine adjuvant	0.14
Paracetamol	Tylenol, Panadol	0.093
Spinosad	Organic insecticide	0.024
Hydrogen cyanide	Fruit pits, bitter cassava	0.012
DDT	Restricted insecticide	0.010
Lead	Batteries, cables, paints	0.007
Caffeine	Coffee, tea, chocolate	0.003
Vitamin D3	Supplements, fish	0.002
Chlorpyrifos	Organophosphate pesticide	0.001
Nicotine	Natural pesticide, tobacco	0.0008
Rotenone	Restricted organic pesticide	0.0004
Mycotoxin T2	Fusarium, moldy grain	0.00002

Limits: Reference Dose (RfD or ADI), Reference Intake (RI), Upper Limit (UL), or Tolerable Daily Intake (TDI). Colours for readability (no official categories exist for these limits).

Sources: EFSA, WHO, EPA, NIH, NHS

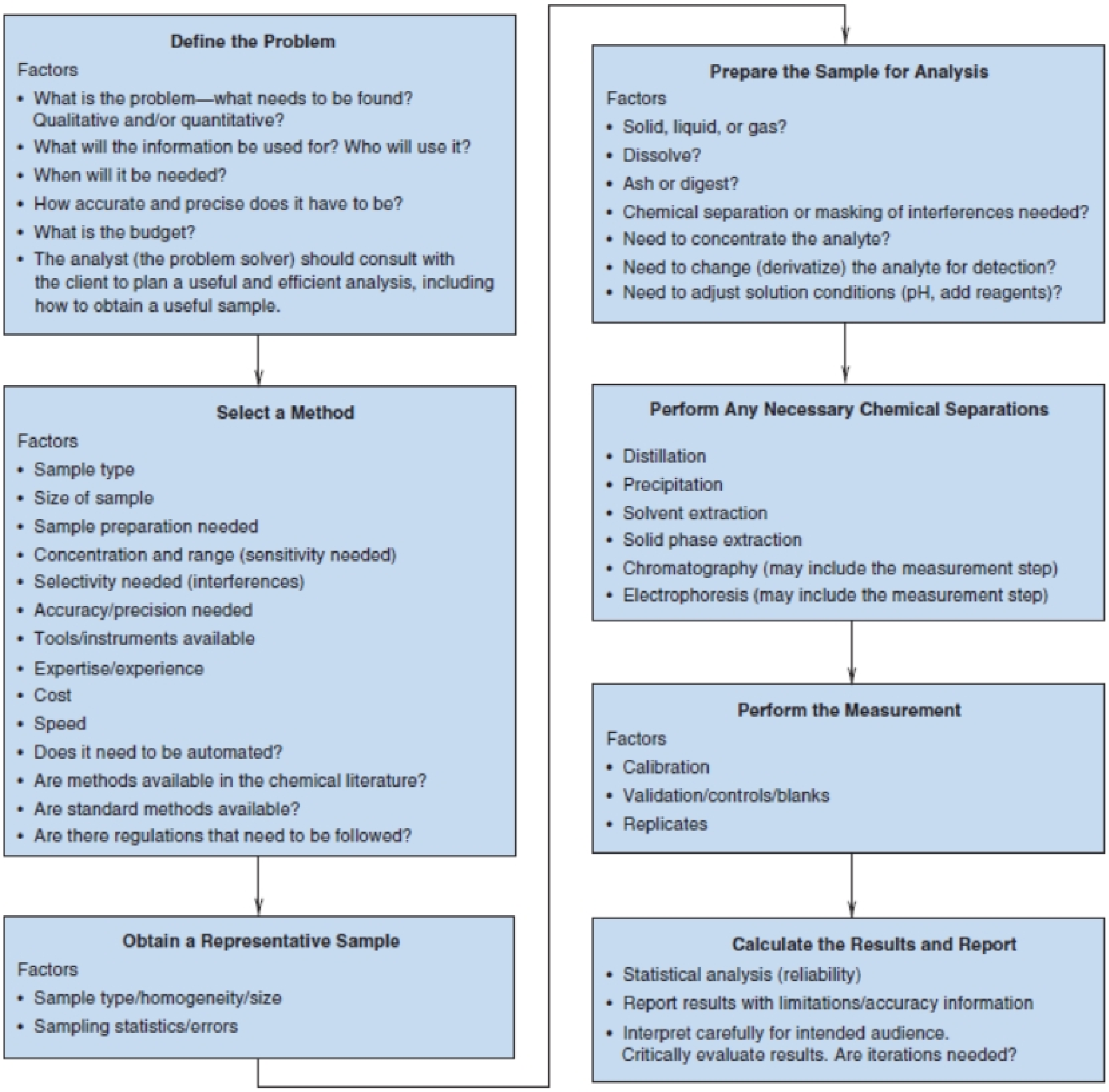
Thoughtscapism More at: thoughtscapism.com

Mommy PhD

Figure 13.31: Chronic Toxicity of some ingredients/molecules (authors: Iida Ruishalme, Alison Bernstein)

13.3.1 Analytical chemistry process

The general analytical process is shown in the figure below. The analytical chemist should be involved in every step. The chemical analyst is like the physicist, mathematician and engineer: a problem solver, a critical part of the team deciding what, why, and how. The unit operations of analytical chemistry that are common to most types of analyses are considered in more detail below that should in part belongs to process to other science jobs.



Define the Problem

Factors

- What is the problem—what needs to be found? Qualitative and/or quantitative?
- What will the information be used for? Who will use it?
- When will it be needed?
- How accurate and precise does it have to be?
- What is the budget?
- The analyst (the problem solver) should consult with the client to plan a useful and efficient analysis, including how to obtain a useful sample.

Select a Method

Factors

- Sample type
- Size of sample
- Sample preparation needed
- Concentration and range (sensitivity needed)
- Selectivity needed (interferences)
- Accuracy/precision needed
- Tools/instruments available
- Expertise/experience
- Cost
- Speed
- Does it need to be automated?
- Are methods available in the chemical literature?
- Are standard methods available?
- Are there regulations that need to be followed?

Obtain a Representative Sample

Factors

- Sample type/homogeneity/size
- Sampling statistics/errors

Prepare the Sample for Analysis

Factors

- Solid, liquid, or gas?
- Dissolve?
- Ash or digest?
- Chemical separation or masking of interferences needed?
- Need to concentrate the analyte?
- Need to change (derivatize) the analyte for detection?
- Need to adjust solution conditions (pH, add reagents)?

Perform Any Necessary Chemical Separations

- Distillation
- Precipitation
- Solvent extraction
- Solid phase extraction
- Chromatography (may include the measurement step)
- Electrophoresis (may include the measurement step)

Perform the Measurement

Factors

- Calibration
- Validation/controls/blanks
- Replicates

Calculate the Results and Report

- Statistical analysis (reliability)
- Report results with limitations/accuracy information
- Interpret carefully for intended audience. Critically evaluate results. Are iterations needed?

Figure 13.32: **Steps in Analytical Chemistry (source: [101])**

13.3.2 Simple Mixtures

There a basically four type of "mixtures"!

Definitions (#13.3.549):

D1. **Solution**: A "solution" is a homogenous (a substance that is uniform in composition) mixture of one or more substances named "solutes" dissolved in another substance named the "solvent".

The solute can be solid, liquid, or gas but the solvent is always a liquid. E.g. salt solution, sugar solution,...

D2. **Suspension**: A "suspension" is a heterogeneous mixture of large solid substance in another substance made by mechanical agitation. The substances distributed in the background material is not dissolved and will settle out unless the mixture is constantly shaken.

The solute is always a solid substance but the solvent can be solid, liquid, or gas. E.g. mud in water, flour in water, some medicines, dust or water in air,...

D3. **Colloids**: Is a homogeneous solution with intermediate particle size between a solution and a suspension.

The solute can be solid, liquid, or gas substance AND the solvent can also be solid, liquid, or gas. E.g. Foam (Whipped cream), Gel (jelly), smoke, blood,...

D4. **Emulsion**: An "emulsion" is a mixture that consists of two or more liquids that do not mix.

There are two kinds of emulsions. An emulsion that settles is named "temporary emulsion". An emulsion that doesn't is named "permanent emulsion". An example of permanent emulsion is the mixture of oil and vinegar or oil and water.

These solution, suspension and emulsion mixtures were already known by the Greeks. Especially by Aristotle who wrote that salt water was heavier and more dense than fresh water (book II, part II, sixth section), and salt water would seek therefore a lower level[6] in his treatise *Meteorologica* in 9630 (holocene calendar).

---

[6] In facts we know since centuries that they do mix! But not necessarily quickly, so separate bodies of water can exist for quite some time.

Figure 13.33: **A phenomenon that was already understood by Aristotle for the special case of salt and fresh water**

Notice we can observe in practice by zooming that there are no "barrier", nor "complete separation" between two liquids that doesn't mix. With a special apparatus we can zoom in and see that in fact there is a gradient of a given width that depends strongly on the type of mixed liquids (be careful with the false dichotomy bias!).

Before going into more or less complicated equations, the simplest case of application of mathematics to chemistry by which we can start is the management of mixtures for analysis and control operations of simple chemical reactions with two mixtures.

Let us consider two typical and particular examples as theoretical introduction:

1. Given a solution (yellow) of 10 millilitres of a solution containing an acid concentration at 30%. How many millilitres of pure acid (blue) should we add to increase the concentration (green) to 50%?

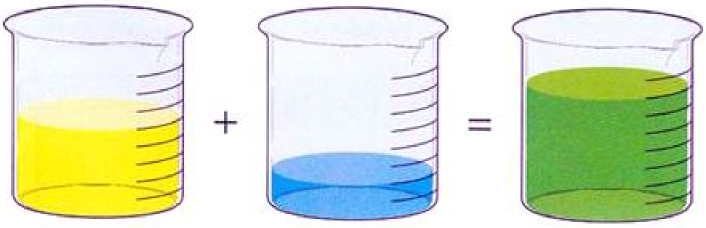
   

   Figure 13.34: **The joy of mixtures...**

   Since the unknown is the amount of pure acid to be added, we will denote it by x. Then we have:

   $$10\cdot 30\%+x=50\%(10+x) \tag{13.3.1}$$

That gives:

$$3+x=5+0.5x \tag{13.3.2}$$

It comes the obviously:

$$x=4 \tag{13.3.3}$$

Therefore 4 millilitres of acid should be added to the original solution.

2. A canister contains 8 litres of gasoline and oil to run an aggregate. If 40% of the initial mixture is of the essence, how much should we remove of the mixture (pink) to replace it with pure gasoline (green) so that the final mixture (light green) contains 60% gasoline?

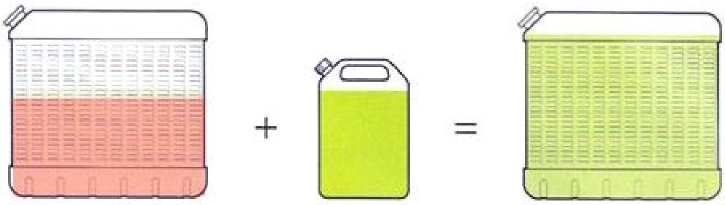


Figure 13.35: **The joy of mixtures for dyers and military ...**

We denote the unknown x that is the number of litres of the initial mixture to remove and replaced by the pure essence being of equal amount also x. Then we have:

$$(8\cdot 40\%-x40\%)+x=8\cdot 60\% \tag{13.3.4}$$

That gives:

$$0.4(8-x)=8\cdot 0.6 \tag{13.3.5}$$

We have then obviously:

$$x=2.\bar{6} \tag{13.3.6}$$

So approximately 2.6 litres should be removed from the original mixture and be replaced by approximately 2.6 litres of pure essence.

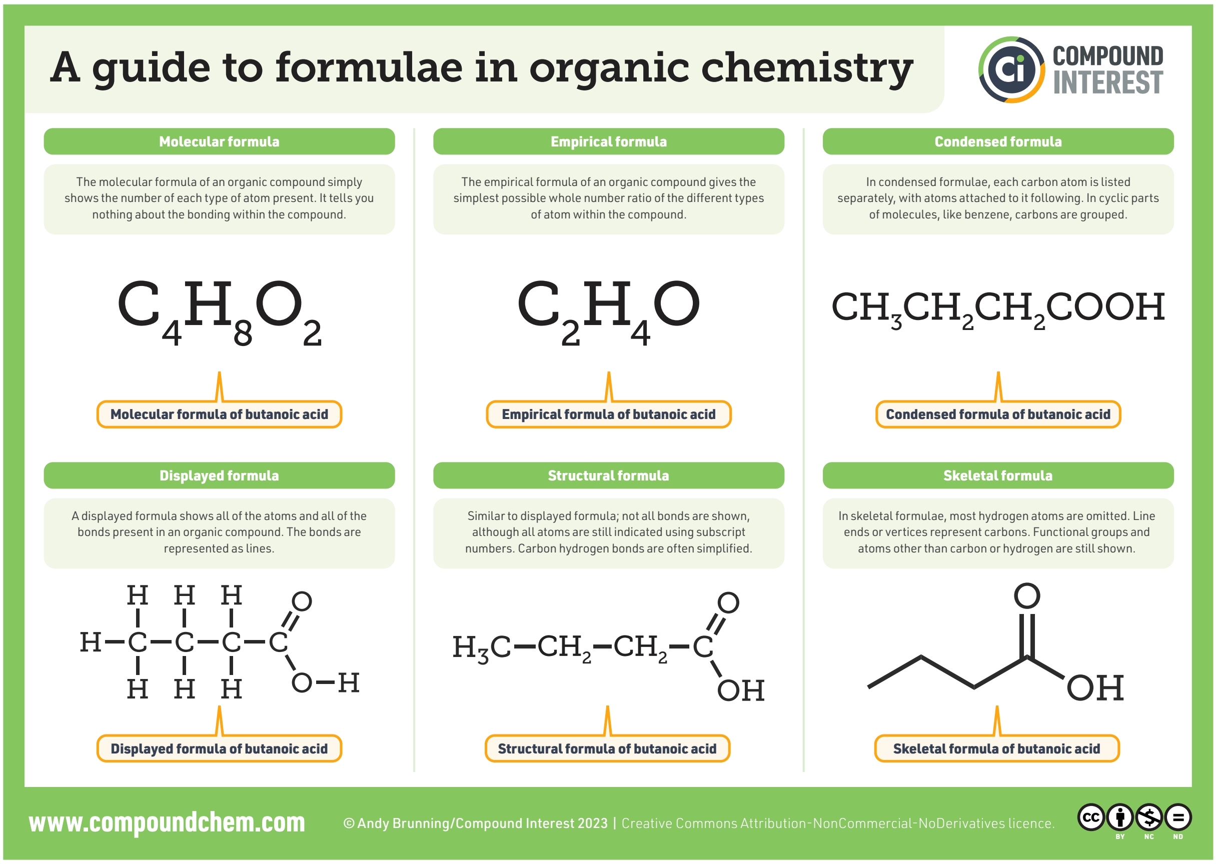
In short this is for all mixtures in this textbook until now. We can go much further and do much more complicated with more unknowns but we'll stop there for now.

13.3.3 Chemical formula

A "chemical formula" is a way of presenting information about the chemical proportions of atoms that constitute a particular chemical compound or molecule, using chemical element symbols, numbers, and sometimes also other symbols, such as parentheses, dashes, brackets, commas and plus (+) and minus (-) signs. These are limited to a single typographic line of symbols, which may include subscripts and superscripts. A chemical formula is not a chemical name since it does not contain any words. Although a chemical formula may imply certain simple chemical structures, it is not the same as a full chemical structural formula. Chemical formulae can fully specify the structure of only the simplest of molecules and chemical substances, and are generally more limited in power than chemical names and structural formulae.

> Caution! There are some traps, like the empirical formula for glucose that is CH_2O (twice as many hydrogen atoms as carbon and oxygen), while its molecular formula is $C_6H_{12}O_6$ (12 hydrogen atoms, six carbon and oxygen atoms).

Sometimes a chemical formula is complicated by being written as a condensed formula (or condensed molecular formula, occasionally called a "semi-structural formula"), which conveys additional information about the particular ways in which the atoms are chemically bonded together, either in covalent bonds, ionic bonds, or various combinations of these types. This is possible if the relevant bonding is easy to show in one dimension. An example is the condensed molecular/chemical formula for ethanol, which is CH_3-CH_2-OH or CH_3CH_2OH. However, even a condensed chemical formula is necessarily limited in its ability to show complex bonding relationships between atoms, especially atoms that have bonds to four or more different substituents.


A guide to formulae in organic chemistry

Molecular formula

The molecular formula of an organic compound simply shows the number of each type of atom present. It tells you nothing about the bonding within the compound.

Empirical formula

The empirical formula of an organic compound gives the simplest possible whole number ratio of the different types of atom within the compound.

Condensed formula

In condensed formulae, each carbon atom is listed separately, with atoms attached to it following. In cyclic parts of molecules, like benzene, carbons are grouped.

Displayed formula

A displayed formula shows all of the atoms and all of the bonds present in an organic compound. The bonds are represented as lines.

Structural formula

Similar to displayed formula; not all bonds are shown, although all atoms are still indicated using subscript numbers. Carbon hydrogen bonds are often simplified.

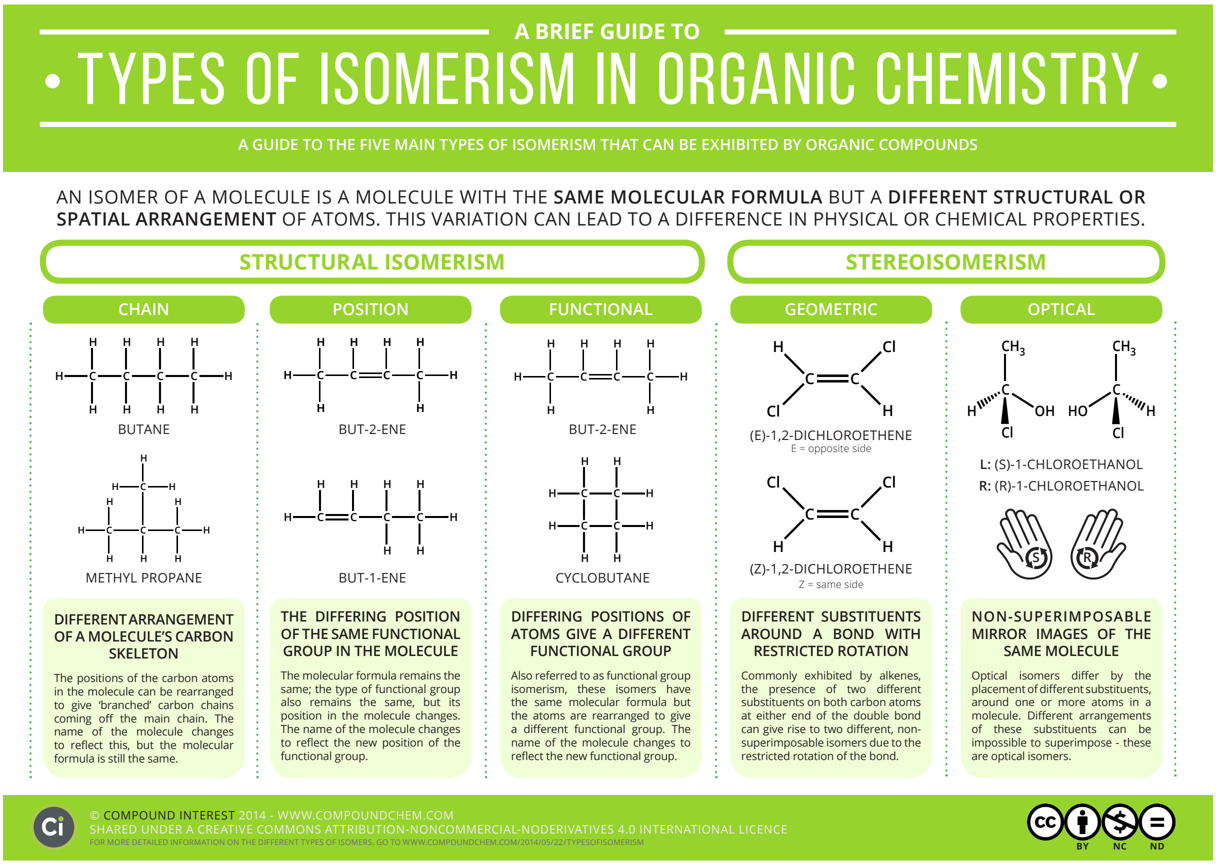
Skeletal formula

In skeletal formulae, most hydrogen atoms are omitted. Line ends or vertices represent carbons. Functional groups and atoms other than carbon or hydrogen are still shown.

COMPOUND INTEREST
$C_4H_8O_2$
Molecular formula of butanoic acid
C_2H_4O
Empirical formula of butanoic acid
$CH_3CH_2CH_2COOH$
Condensed formula of butanoic acid
Displayed formula of butanoic acid
$H_3C-CH_2-CH_2-C$
OH
Structural formula of butanoic acid
Skeletal formula of butanoic acid
www.compoundchem.com

© Andy Brunning/Compound Interest 2023 | Creative Commons Attribution-NonCommercial-NoDerivatives licence.

Since a chemical formula must be expressed as a single line of chemical element symbols, it often cannot be as informative as a true structural formula, which is a graphical representation of the spatial relationship between atoms in chemical compounds (see for example the figure for butane structural and chemical formulae, at right). For reasons of structural complexity, a single condensed chemical formula (or semi-structural formula) may correspond to different molecules, known as "isomers". For example, glucose shares its molecular formula $C_6H_{12}O_6$ with a number of other sugars, including fructose, galactose and mannose. Linear equivalent chemical names exist that can and do specify uniquely any complex structural formula (see chemical nomenclature), but such names must use many terms (words), rather than the simple element symbols, numbers, and simple typographical symbols that define a chemical formula.



A BRIEF GUIDE TO

• TYPES OF ISOMERISM IN ORGANIC CHEMISTRY •

A GUIDE TO THE FIVE MAIN TYPES OF ISOMERISM THAT CAN BE EXHIBITED BY ORGANIC COMPOUNDS

AN ISOMER OF A MOLECULE IS A MOLECULE WITH THE **SAME MOLECULAR FORMULA** BUT A **DIFFERENT STRUCTURAL OR SPATIAL ARRANGEMENT** OF ATOMS. THIS VARIATION CAN LEAD TO A DIFFERENCE IN PHYSICAL OR CHEMICAL PROPERTIES.

STRUCTURAL ISOMERISM

CHAIN

BUTANE

METHYL PROPANE

DIFFERENT ARRANGEMENT OF A MOLECULE'S CARBON SKELETON

The positions of the carbon atoms in the molecule can be rearranged to give 'branched' carbon chains coming off the main chain. The name of the molecule changes to reflect this, but the molecular formula is still the same.

POSITION

BUT-2-ENE

BUT-1-ENE

THE DIFFERING POSITION OF THE SAME FUNCTIONAL GROUP IN THE MOLECULE

The molecular formula remains the same; the type of functional group also remains the same, but its position in the molecule changes. The name of the molecule changes to reflect the new position of the functional group.

FUNCTIONAL

BUT-2-ENE

CYCLOBUTANE

DIFFERING POSITIONS OF ATOMS GIVE A DIFFERENT FUNCTIONAL GROUP

Also referred to as functional group isomerism, these isomers have the same molecular formula but the atoms are rearranged to give a different functional group. The name of the molecule changes to reflect the new functional group.

STEREOISOMERISM

GEOMETRIC

(E)-1,2-DICHLOROETHENE
E = opposite side

(Z)-1,2-DICHLOROETHENE
Z = same side

DIFFERENT SUBSTITUENTS AROUND A BOND WITH RESTRICTED ROTATION

Commonly exhibited by alkenes, the presence of two different substituents on both carbon atoms at either end of the double bond can give rise to two different, non-superimposable isomers due to the restricted rotation of the bond.

OPTICAL

L: (S)-1-CHLOROETHANOL
R: (R)-1-CHLOROETHANOL

NON-SUPERIMPOSABLE MIRROR IMAGES OF THE SAME MOLECULE

Optical isomers differ by the placement of different substituents, around one or more atoms in a molecule. Different arrangements of these substituents can be impossible to superimpose - these are optical isomers.

© COMPOUND INTEREST 2014 - WWW.COMPOUNDCHEM.COM
SHARED UNDER A CREATIVE COMMONS ATTRIBUTION-NONCOMMERCIAL-NODERIVATIVES 4.0 INTERNATIONAL LICENCE
FOR MORE DETAILED INFORMATION ON THE DIFFERENT TYPES OF ISOMERS, GO TO WWW.COMPOUNDCHEM.COM/2014/05/22/TYPESOFISOMERISM

Remark

Isomers have caused multiple headaches since the beginning of modern industrial chemistry. The most famous medical example where an isomer caused catastrophic harm is the thalidomide tragedy, in which one enantiomer of the drug acted as a sedative while its mirror-image isomer caused severe birth defects, leading to thousands of cases of phocomelia and other malformations worldwide. In the past chemist where not able to identify isomer but nowadays various techniques exist to differentiate them.

Chemical formulae may be used in chemical equations to describe chemical reactions and other chemical transformations, such as the dissolving of ionic compounds into solution. While, as noted, chemical formulae do not have the full power of structural formulae to show chemical relationships between atoms, they are sufficient to keep track of numbers of atoms and numbers of electrical charges in chemical reactions, thus balancing chemical equations so that these equations can be used in chemical problems involving conservation of atoms, and conservation of electric charge.

Condensed chemical formulae may also be used to represent ionic compounds that do not exist as discrete molecules, but nonetheless do contain covalently bound clusters within them. These polyatomic ions are groups of atoms that are covalently bound together and have an overall ionic charge, such as the sulfate $[\mathrm{SO_4}]^{2-}$ ion. Each polyatomic ion in a compound is written individually in order to illustrate the separate groupings. For example, the compound dichlorine hexoxide has an empirical formula $\mathrm{ClO_3}$, and molecular formula $\mathrm{Cl_2O_6}$, but in liquid or solid forms, this compound is more correctly shown by an ionic condensed formula $[\mathrm{ClO_2}]^+[\mathrm{ClO_4}]^-$, which illustrates that this compound consists of $[\mathrm{ClO_2}]^+$ ions and $[\mathrm{ClO_4}]^-$ ions. In such cases, the condensed formula only need be complex enough to show at least one of each ionic species.

Isomers aren't the only structural trap in chemistry because "polymorphs" - different crystalline arrangements of the *same* molecule - can radically change a substance's behavior just as dramatically as switching between isomers. While isomers differ in connectivity or spatial arrangement, polymorphs differ only in how molecules pack in the solid state, yet this subtle shift can alter melting point, solubility, stability, bioavailability, and even toxicity. Famous cases like the HIV drug ritonavir, which suddenly became unusable when a more stable but poorly soluble polymorph appeared, show that even when the molecular structure is correct, the solid-state structure can still derail manufacturing, storage, or therapeutic performance.

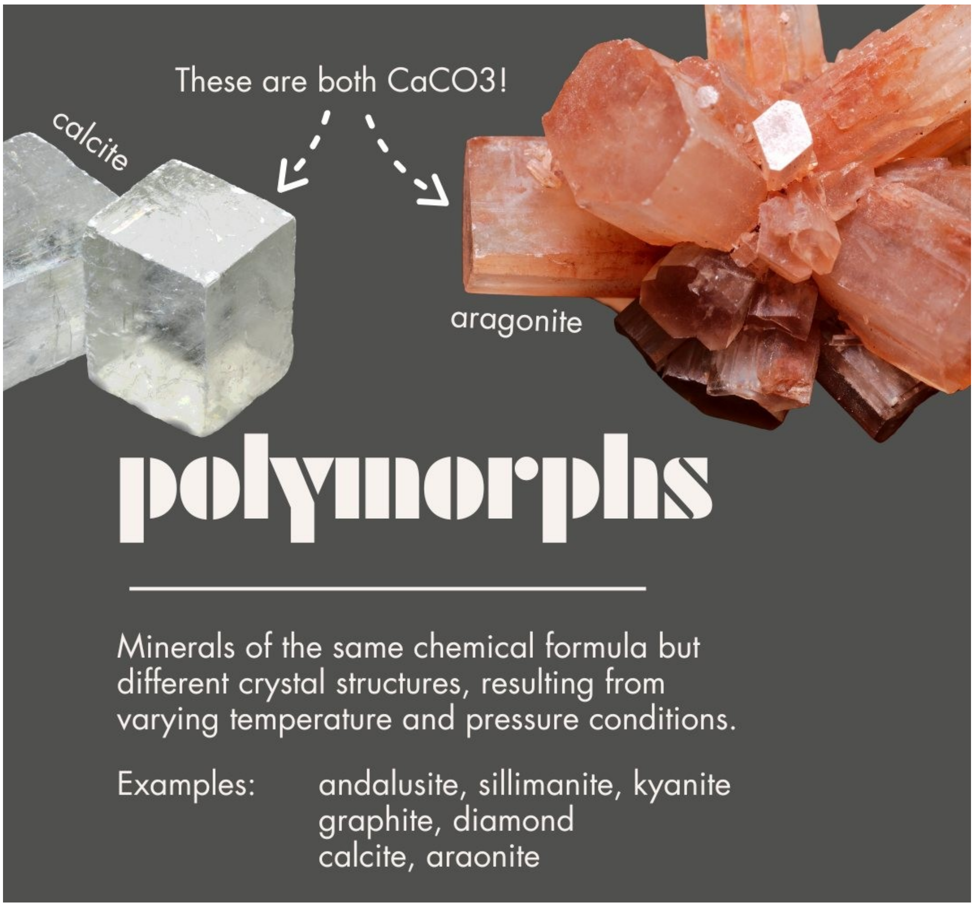

These are both CaCO3!
calcite
aragonite
polymorphs
Minerals of the same chemical formula but different crystal structures, resulting from varying temperature and pressure conditions.
Examples: andalusite, sillimanite, kyanite
graphite, diamond
calcite, araonite

Figure 13.36: Example of polymorph

13.3.4 Reactions

Since the main study in chemistry is to observe the results of pure substances mixtures and/or of compounds mixtures, it is first necessary to deal with the basic rules governing these mixtures under normal conditions of pressure and temperature (N.C.P.T).

We should first clarify that we are not going to study in this section what generates the connections between the elements, as this is the role of quantum and molecular chemistry (see previous sections). Furthermore, we insist on the fact that every theoretical element will be illustrated with a practical example which can be useful sometimes to better understand.

But before, for prevention reason, let us introduce a symbol shown in the figure below and that we can see on containers of chemicals in a laboratory or workplace. Sometimes named a "fire diamond" or "hazard diamond", this chemical hazard diamond provides valuable information that briefly summarizes the various dangers of which to be aware when working with a particular substance:

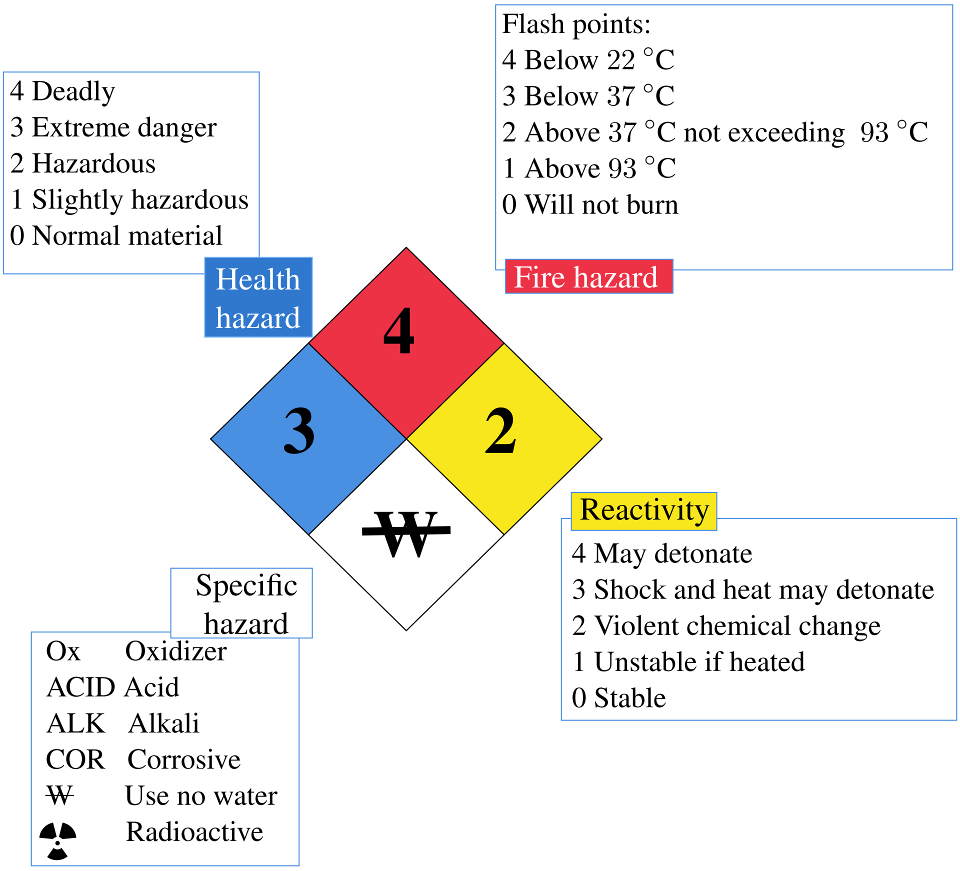

4 Deadly
3 Extreme danger
2 Hazardous
1 Slightly hazardous
0 Normal material
Health hazard
Flash points:
4 Below 22 °C
3 Below 37 °C
2 Above 37 °C not exceeding 93 °C
1 Above 93 °C
0 Will not burn
Fire hazard
4
3
2
W
Reactivity
4 May detonate
3 Shock and heat may detonate
2 Violent chemical change
1 Unstable if heated
0 Stable
Specific hazard
Ox Oxidizer
ACID Acid
ALK Alkali
COR Corrosive
W Use no water
Radioactive

Figure 13.37: National Fire Protection Agency (NFPA) hazard diamond summarizes the major hazards of a chemical substance

Let us now consider a closed chemical system (without mass transfer therefore!). We translate the change in the composition (if applicable and if have there is one!) of the chemical system with a reaction equation of the form (the system does not always go both ways!):

$$\boxed{v_1A_1+v_2A_2+\ldots\rightleftharpoons v_1'A_1'+v_2'A_2'+\ldots} \quad (13.3.7)$$

but most of time written as:

$$v_1A_1+v_2A_2+\ldots+v_nA_n \rightleftharpoons v_{n+1}A_{n+1}+v_{n+2}A_{n+2}+\ldots+v_mA_m \tag{13.3.8}$$

named "reaction equation" where the coefficients $v_i \in \mathbb{N}^*$ are named "stoichiometric coefficients" in the sense that they indicate the "golden proportions", strictly named "stoichiometric ratio" necessary such that under normal conditions the reaction can take place and where the A_i are the reactants (pure or compounds) and the A'_i the formed products.

The sum of the coefficients of the reactants minus the sum of the coefficients of the products is the "stoichiometric sum". If this is zero, the equation is say to be a "balanced chemical reaction". In this case the reaction can be written:

$$\sum v_iA_i=0 \tag{13.3.9}$$

where the convention is that stoichiometric coefficients are positive for reactants and negative for products. The stoichiometric sum is $\sum v_i$.

✍**Example:**

Let us consider to illustrate these concepts the reaction (dinitrogen and hydrogen giving ammonia):

$$\mathrm{N_2}+3\mathrm{H_2} \rightleftharpoons 2\mathrm{NH_3} \tag{13.3.10}$$

where the Latin letters represent the pure substances (atoms) whose name does not matter to us in this textbook (notation proposed by Jöns Jacob Berzelius in 11813 according to holocene calendar). The indices simply represent the number of combination of atoms to obtain a molecule.

We then have in this reaction:

$$v_1=1,v_2=3,v'_1=2 \tag{13.3.11}$$

$$A_1:=\mathrm{N_2},A_2:=\mathrm{H_2},A'_1:=\mathrm{NH_3} \tag{13.3.12}$$

The reader will have noticed that we have well following our convention for the mass balance:

$$\sum_i v_iA_i=1\cdot(2\mathrm{N})+3\cdot(2\mathrm{H})-2\cdot(\mathrm{N})-2\cdot(3\mathrm{H})=0 \tag{13.3.13}$$

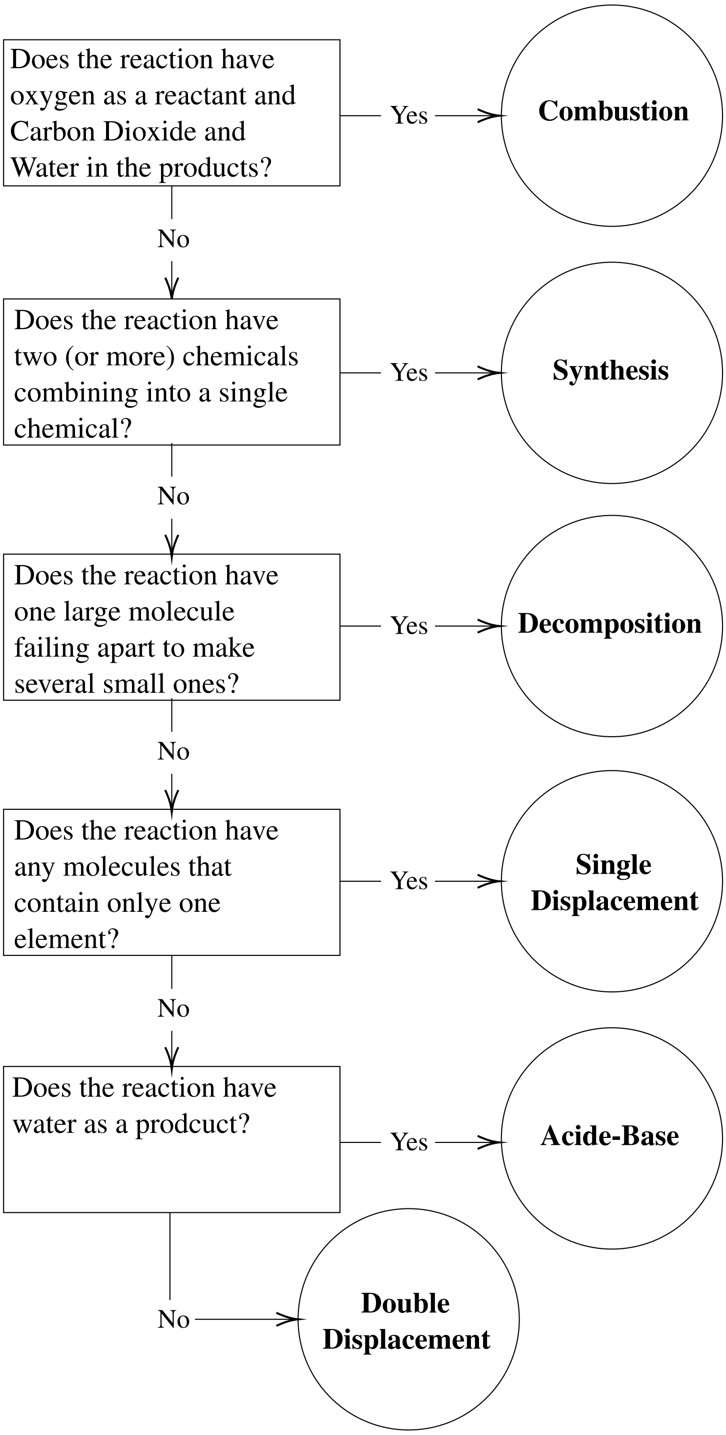

Does the reaction have oxygen as a reactant and Carbon Dioxide and Water in the products?
Yes
Combustion
No
Does the reaction have two (or more) chemicals combining into a single chemical?
Yes
Synthesis
No
Does the reaction have one large molecule failing apart to make several small ones?
Yes
Decomposition
No
Does the reaction have any molecules that contain onlye one element?
Yes
Single Displacement
No
Does the reaction have water as a prodcuct?
Yes
Acide-Base
No
Double Displacement

Figure 13.38: **Some typical chemical reactions**

Caution! In the writing of the above equation, we require that all the A_i without exception react to the chemical reaction and that therefore all the v_i are dependent.

If the "golden proportions" are respected (such that the coefficients are well stoichiometric) and exist when writing of the reaction equation, then for any $\alpha \in \mathbb{N}$ we have:

$$\alpha(v_1 A_1 + v_2 A_2 + \ldots) \rightleftharpoons \alpha(v_1' A_1' + v_2' A_2' + \ldots) \tag{13.3.14}$$

this proposal can be proven only if the stoichiometric coefficients on one side or the other of the reaction vary proportionally. Experience shows that in normal conditions of temperature and pressure (N.C.T.P.) this is the case!

Remark

A famous reaction that makes some scientific illiterate people scared around the World is the one that explains chemtrails (in fact it should be named "contrails", that is the abbreviation for "cond(densation)trails)" - or that also explains vapour you see going out of car's exhaust pipe:

$$\underbrace{\mathrm{C}_x\mathrm{H}_y}_{\text{Hydrocarbone}} + \mathrm{O}_2 \Rightarrow \mathrm{CO}_2 + \mathrm{H}_2\mathrm{O}$$

In other words, in thermic engines, the hydrocarbone is transformed at the exit on carbon dioxide and on condensed water... hence the name "contrails"... For Airplanes we have:

$$\underbrace{2\mathrm{C}_{12}\mathrm{H}_{26}}_{\text{Kerosen}} + \underbrace{37\mathrm{O}_2}_{\text{Oxygen}} \Rightarrow \underbrace{24\mathrm{CO}_2}_{\text{Carbon dioxide}} + \underbrace{26\mathrm{H}_2\mathrm{O}}_{\text{Water vapor}}$$

Both of the gaseous products expelled from the exhaust pipe of a jet- CO_2 and water vapor - are invisible. When the products hit the cold air, the water vapor condenses and freezes around small particles (see phase diagram at page 3910), known as nucleation sites, which are also present in the exhaust. This produces a visible cloud:

Figure 13.39: Some typical contrails

Contrails must not be confused with the evaporation of the water present in the air during a sonic boom as visible on the photo at page 8218.

Therefore, the stoichiometry of the reaction requires that if it disappears x_1 moles of A_1, x_2 moles of A_2 respectively with a variation of material of the products $\mathrm{d}n_1, \mathrm{d}n_2, \ldots$, it will appear accordingly x'_1 moles of A'_1, x'_2 moles of A'_2, ... with respectively a variation of material of the products $\mathrm{d}n'_1, \mathrm{d}n'_2, \ldots$... by respecting the proportionalities of the stoichiometric coefficients such that we can write the "material balance equation":

$$\frac{-\mathrm{d}n_1}{v_1} = \frac{-\mathrm{d}n_2}{v_2} = \ldots = \frac{-\mathrm{d}n'_1}{v'_1} = \frac{-\mathrm{d}n'_2}{v'_2} = \ldots = \mathrm{d}\xi \qquad (13.3.15)$$

where $\mathrm{d}\xi$ is named the "elementary reaction progress" (frequently we will take the absolute values of the ratios to not have to think about the sign of the variations).

The division of the variations $\mathrm{d}n_1, \mathrm{d}n_2$ and $\mathrm{d}n'_1, \mathrm{d}n'_2$ by their stoichiometric coefficients is justified only for normalization reasons having for purpose to bring $\mathrm{d}\xi$ to a value between 0 and 1 (between 0% and 100%...).

These last equalities simply indicate that if one of the reactive products disappear in a given quantity, the other reactants have their quantity that decreased in relation to their stoichiometric coefficient so as to maintain the golden proportions of the reaction.

The writing of the energy balance can be simplified by the introduction of algebraic stoichiometric coefficients v_i such that: $v_i > 0$ for a formed product, $v_i < 0$ for a reactive product.

Finally we can write:

$$\boxed{\mathrm{d}x_i = \frac{\mathrm{d}n_i}{v_i}} \tag{13.3.16}$$

we also often find in the literature with the absolute value at the numerator!

Therefore, with this algebraic convention, the reaction equation as it exists, can be written:

$$\boxed{\sum_i v_i A_i = 0} \tag{13.3.17}$$

which means that the algebraic sum of the total number of pure compounds of the reactants and products formed is always zero.

It is clear that at the initial time of the reaction we choose for the progress the value $\xi = 0$ (its maximum value being equal to unity), time at which the quantities of material are equal to $n_{i,0}$.

The integration of the differential expression of material balance obviously gives:

$$\begin{gathered}\int_{t_0}^{t} v_i \mathrm{d}\xi = \int_{t_0}^{t} \mathrm{d}n_i \Rightarrow v_i \left[\xi(t) - \underbrace{\xi(t_0)}_{=0}\right] = n_i(t) - \underbrace{n_i(t_0)}_{n_{i,0}} \\ \Rightarrow v_i\xi = n_i(t) - n_{i,0}\end{gathered} \tag{13.3.18}$$

Therefore:

$$\boxed{n_i = n_{i,0} + v_i\xi} \tag{13.3.19}$$

relation that we found in chemical progress tables (see further below), without forgetting that $v_i > 0$ for a formed product and, $v_i < 0$ for a reactive product.

This bring us to the question: What is the maximum value ξ_{max} of the progress of a reaction?

Well the answer to that question is in fact quite simple: The maximum progress value of a reaction having the stoichiometric proportions and such that it occurs when the reactants will have all disappeared is necessarily given by:

$$\xi_{\text{max}} = \left|\frac{n_i}{v_i}\right| \tag{13.3.20}$$

and named the "limiting reactant", that is to say, the reactant that disappears (has always the smallest value of molarity) first and stops the expected reaction (the other one being named the "excess reactant")! If there is no limiting reactant, that means that at the end of the reaction all reactants have been transformed: then we say that all reactants were in stoichiometric proportion.

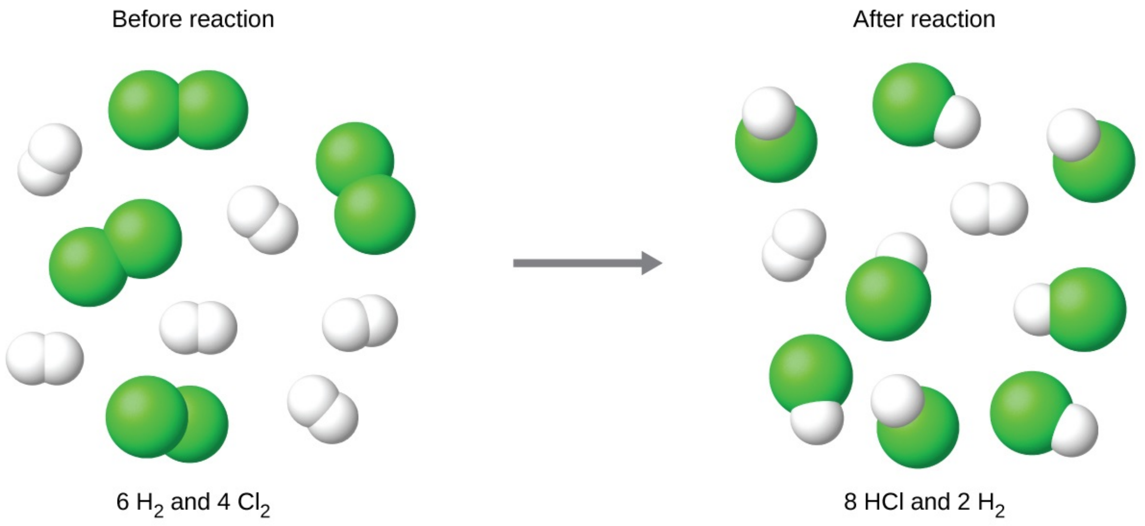


Figure 13.40: **When H_2 and Cl_2 are combined in non-stoichiometric amounts, one of these reactants will limit the amount of HCl that can be produced. This illustration shows a reaction in which hydrogen is present in excess and chlorine is the limiting reactant.**

It may be helpful to define the "percentage of completion" A_i given by the intensive quantity:

$$\tau_i=\frac{\text{quantity of } A_i \text{ having reacting}}{\text{initial quantity of } A_i} \tag{13.3.21}$$

which gives with a more formal notation:

$$\boxed{\tau_i=\frac{|v_i|\xi}{n_{i,0}}} \tag{13.3.22}$$

✍**Example:**

Let us consider to illustrate these concepts the reaction (dinitrogen and hydrogen giving ammonia):

$$N_2+3H_2 \rightleftharpoons 2NH_3 \tag{13.3.23}$$

where the Latin letters represent the pure substances (atoms) whose name does not matter to us in this textbook (notation proposed by Jöns Jacob Berzelius in 11813 according to holocene calendar). The indices simply represent the number of combination of atoms to obtain a molecule.

We then have in this reaction:

$$v_1=1,v_2=3,v'_1=2 \tag{13.3.24}$$

$$A_1:=N_2,A_2:=H_2,A'_1:=NH_3 \tag{13.3.25}$$

The reader will have noticed that we have well following our convention for the mass balance:

$$\sum_i v_i A_i = 1\cdot(2\mathrm{N}) + 3\cdot(2\mathrm{H}) - 2\cdot(\mathrm{N}) - 2\cdot(3\mathrm{H}) = 0 \qquad (13.3.26)$$

If we consider that there is one mole of each compound body, it gives us for the stoichiometric proportions (to a given factor $x \in \mathbb{R}^*$ for all values):

$$\begin{aligned} n_1 &= v_1\cdot 1 = 1\cdot 1 = 1\,[\mathrm{mol}] \\ n_2 &= v_2\cdot 1 = 3\cdot 1 = 3\,[\mathrm{mol}] \\ n_3 &= v_3\cdot 2 = 1\cdot 2 = 2\,[\mathrm{mol}] \end{aligned} \qquad (13.3.27)$$

If at any a given time $t \neq t_0$, we get by measurement:

$$n_3 = 0.2\,[\mathrm{mol}] \qquad (13.3.28)$$

What is the progress of that reaction?

The answer is:

$$\mathrm{d}\xi = \frac{\mathrm{d}n_i}{v_i} \Rightarrow \xi = \frac{\Delta n(\mathrm{NH_3})}{2} = \frac{0.2}{2} = 0.1 \qquad (13.3.29)$$

or in other words, we are at 10% of progress (logical!).

The conversion rate of $\mathrm{NH_3}$ is thereto:

$$\tau_i = \frac{|v_i|\xi}{n_{i,0}} = \frac{|-3|0.1}{3} = 0.1 \qquad (13.3.30)$$

And what is the maximum progress value $\xi_{\max}$ of the limiting reactant?

So in the context of the above example where we have $n_{1,0} = 1\,[\mathrm{mol}]$ for the $\mathrm{N_2}$ then:

$$\xi_{\max} = \left|\frac{n_i}{v_i}\right| = \left|\frac{1}{-1}\right| = 1\,[\mathrm{mol}] \qquad (13.3.31)$$

Chemists also often use what they name a "reaction progress table".

Let us take our previous example to introduce this table. We have:

	Equation	$\mathrm{N_2}$	$+3\mathrm{H_2}$	$+2\mathrm{NH_3}$
Initial State	$n_{i,0}$	1	3	0
Intermediate State	$n_i = n_{i,0} + v_i\xi$	$1-1\cdot\xi$	$3-3\cdot\xi$	$0+2\cdot\xi$
Final State	$\xi_{\max}$	$1-1\cdot\xi_{\max}$	$3-3\cdot\xi_{\max}$	$0+2\cdot\xi_{\max}$

Table 13.4: **Table progress of a chemical reaction**

Let us seek $\xi_{\max}$ from this table. The limiting reactant is either $\mathrm{N_2}$ or $3\mathrm{H_2}$.

So for N_2:

$$1-1\cdot\xi_{\max}=0\Rightarrow\xi_{\max}=\frac{n_i}{v_i}=1 \tag{13.3.32}$$

and for $3H_2$:

$$3-3\cdot\xi_{\max}=0\Rightarrow\xi_{\max}=\frac{n_i}{v_i}=1 \tag{13.3.33}$$

Each reactant having the same $\xi_{\max}$ progress, it is thus also the minimum $\xi_{\max}$. Consequently, according to the definition of limiting reactant, as the proportions are stoichiometric in the given example no reactant is limiting.

~10 %

★★★☆☆

25 votes, 55.20%

13.4. Thermochemistry

THERMOCHEMISTRY is the branch that historically focuses on thermic phenomena and to equilibrium accompanying chemical reactions. It mainly has its foundations in Thermodynamics. More technically, thermochemistry is the study of the energy and heat associated with chemical reactions and/or physical transformations. A reaction may release or absorb energy, and a phase change may do the same, such as in melting and boiling. Thermochemistry focuses on these energy changes, particularly on the system's energy exchange with its surroundings. Thermochemistry is useful in predicting reactant and product quantities throughout the course of a given reaction. In combination with entropy determinations, it is also used to predict whether a reaction is spontaneous or non-spontaneous, favourable or unfavourable.

We can only strongly recommend the readers to read before the section on Thermodynamics (see page 3597) because many concepts that have been seen there will be assumed to be known in this section.

Moreover, it is strongly recommended to read this chapter in parallel to that of Analytical Chemistry (this can be a boring but you must do with...).

13.4.1 Chemical transformations

Chemical reactions, such as those that occur when you light a match, involve changes in energy as well as matter. Societies at all levels of development could not function without the energy released by chemical reactions. In 12012 (holocene calendar), about 85% of US energy consumption came from the combustion of petroleum products, coal, wood, and garbage. We use this energy to produce electricity (38%); to transport food, raw materials, manufactured goods, and people (27%); for industrial production (21%); and to heat and power our homes and businesses (10%). While these combustion reactions help us meet our essential energy needs, they are also recognized by the majority of the scientific community as a major contributor to global climate change.

Useful forms of energy are also available from a variety of chemical reactions other than combustion. For example, the energy produced by the batteries in a cell phone, car, or flashlight results from chemical reactions. This chapter introduces many of the basic ideas necessary to explore the relationships between chemical changes and energy, with a focus on thermal energy.

Given the closed system of the following chemical reaction (see section Analytical Chemistry page 6241):

$$v_1 A_1 + v_2 A_2 + \ldots \rightleftharpoons v_1' A_1' + v_2' A_2' + \ldots \tag{13.4.1}$$

We will consider for simplicity that the chemical reaction is complete and that the reactants are used in stoichiometric amounts (state 1: Σ_1) to give the products formed, also in stoichiometric quantities (state 2: Σ_2).

If the transformation is done in (quasi-)steady volume, work on the surrounding atmosphere is zero because (see section Thermodynamics page 3641):

$$\Delta W=-\int_1^2 P\mathrm{d}V=0 \tag{13.4.2}$$

The application of the first law of thermodynamics is reduced and allows then us to write:

$$U_2-U_1=Q_V \tag{13.4.3}$$

where Q_v is within the thermal chemistry framework named "heat of reaction at constant volume", of course exchanged between the system and the external environment (we do not write the delta Δ in front of Q_V to indicate that it is a variation... by tradition...).

Let us recall that:

1. If $Q_V>0$ the reaction is said to be "endothermic" (the system receives heat from the external environment).
2. If $Q_V<0$ the reaction is said to be "exothermic" (the system gives heat to the external environment).
3. If $Q_V=0$ the reaction is said to be "athermic" (the system do not exchange any heat with the environment).

Remark

Let us also recall that a closed system is not an isolated system! For a review of different definitions, the reader is referred once again to the section of Thermodynamics (see page 3597).

If the reaction is carried out at constant pressure (the most usual case in practice), that is to say isobaric, then we have:

$$\Delta W=-\int_1^3 P\mathrm{d}V=-P(V_3-V_1) \tag{13.4.4}$$

Remark

The choice of integration indices are different to previously to differentiate the fact that a reaction a pressure or constant volume are not necessarily identical.

The application of the first law of thermodynamics, between the two states, gives:

$$U_3-U_1=Q_P-P(V_3-V_1) \tag{13.4.5}$$

where Q_p is the amount of heat, named "constant-pressure reaction heat", exchanged between the system and the external environment (Q_P is a variation... even if the traditional unfortunate notation of thermodynamician does not put that in evidence...).

Using the definition of enthalpy, we can write the last relation in the form:

$$Q_P = U_3 - U_1 + P(V_3 - V_1) = (U_3 + PV_3) - (U_1 + PV_1) = H_3 - H_1 \tag{13.4.6}$$

If we work with the molar volumes, those of condensed phases (therefore solid and liquid) is negligible compared to the gas molar volume, only the gas components have a very different enthalpy of their internal energy (see the example in the section of Thermodynamics). We would therefore have under the ideal gas approximation (see section Thermodynamics page 3643):

$$H_i = U_i + n_i RT \tag{13.4.7}$$

In the context of the ideal gas, the prior-previous relation can be written:

$$\begin{aligned} Q_P &= H_3 - H_1 = U_3 - U_1 + (n_3 - n_1)RT = U_3 - U_1 + U_2 + (n_3 - n_1)RT - U_2 \\ &= Q_V + (n_3 - n_1)RT + U_3 - U_2 \end{aligned} \tag{13.4.8}$$

But, as (see section Continuum Mechanics page 3897) U_2 and U_3 are both the same final states of a single complete reaction and that we know for a monoatomic gas we have:

$$U_i = \frac{3}{2} N_i kT \tag{13.4.9}$$

therefore the internal energy U_2 and U_3 only depends on the number of components but ... they are equal since they are the same final state of the same reaction!

Therefore we have:

$$U_2 = U_3 \Rightarrow U_3 - U_2 = 0 \tag{13.4.10}$$

By putting $\Delta n = n_2 - n_1$ (the difference between the number of moles of gas of formed products and those of reacting products), we can write for a chemical reaction:

$$\boxed{Q_P = Q_V + \Delta n RT} \tag{13.4.11}$$

that gives the possibility to differentiates the energy involved between isobaric and isochoric reaction and look for the best choice in terms of industrial objectives. It is interesting to notice that if the Δ of moles is zero. Isobaric or isochoric heat variations are equal and there is no a priori reason to prefer one or the other transformations.

Obviously in practice the problem is to know the values of the different variables of the latter relation. These values can be found on huge databases that chemists have access to... This relationship is only very rarely used in practice and in any case it is based on too simplifying and restrictive assumptions to be of real practical interest.

13.4.2 Molar Quantities

Definitions (#13.4.550):

D1. By convention, the "mole" is the quantity of substance of a system which contains as many chemical species as there are Carbon atoms in 12 [g] of Carbon 12 (see section Nuclear Physics page 5027).

The number of carbon atoms contained in 12 [g] is equal to the Avogadro's number given approximately by (notice that this number is by far bigger than the number of humans that are on Earth!):

$$N_\text{A} \cong 6.0221367 \cdot 10^{23} \tag{13.4.12}$$

This means verbatim and by construction that a mole of water, of iron, of electron, respectively always contains a number of atoms equal to the Avogadro's number.

Most of time the mole is simply denoted n and has its value in $\mathbb{R}^+$.

Notice that with a mixed system it is a mathematical nonsense to do the sum of the molar masses of the constituents for the total molar mass. The molar mass is an intensive quantity!

Remarks

R1. Hydrogen-1 was once used as a standard but given the inaccuracy that can occur because of its low mass, it was later disregarded. Once mass spectrometry was made available, physicists were using Carbon-12 for it's stability and abundance, and basically to stop everybody from fighting. Carbon-12 also more accurately defines a mass for hydrogen, and it is unbound in it's ground state and also is the most common and readily available isotope to have exactly the same number of protons and neutrons, 6 of each, and thus provides a perfect average when divided by the total number of protons and neutrons (electron is so small as to be considered negligible).

R2. The "Avogadro project" aims to redefine Avogadro's constant (currently defined by the kilogram: the number of atoms in 12 [g] of Carbon-12) and reverse the relationship so that the kilogram is precisely specified by Avogadro's constant. This method required creating the most perfect sphere on Earth. It is made out of a single crystal of silicon-28 atoms. By carefully measuring the diameter, the volume can be precisely specified. Since the atom spacing of silicon is well known, the number of atoms in a sphere can be accurately calculated. This allows for a very precise determination of Avogadro's constant.

D2. The "molar mass (MM)" is the mass of one mole of atoms of the chemical elements involved. Therefore by definition the molar mass of C_{12} is equal to 12 grams (yes historically we use the gram to express molar mass because for application purposes is obviously more convenient...

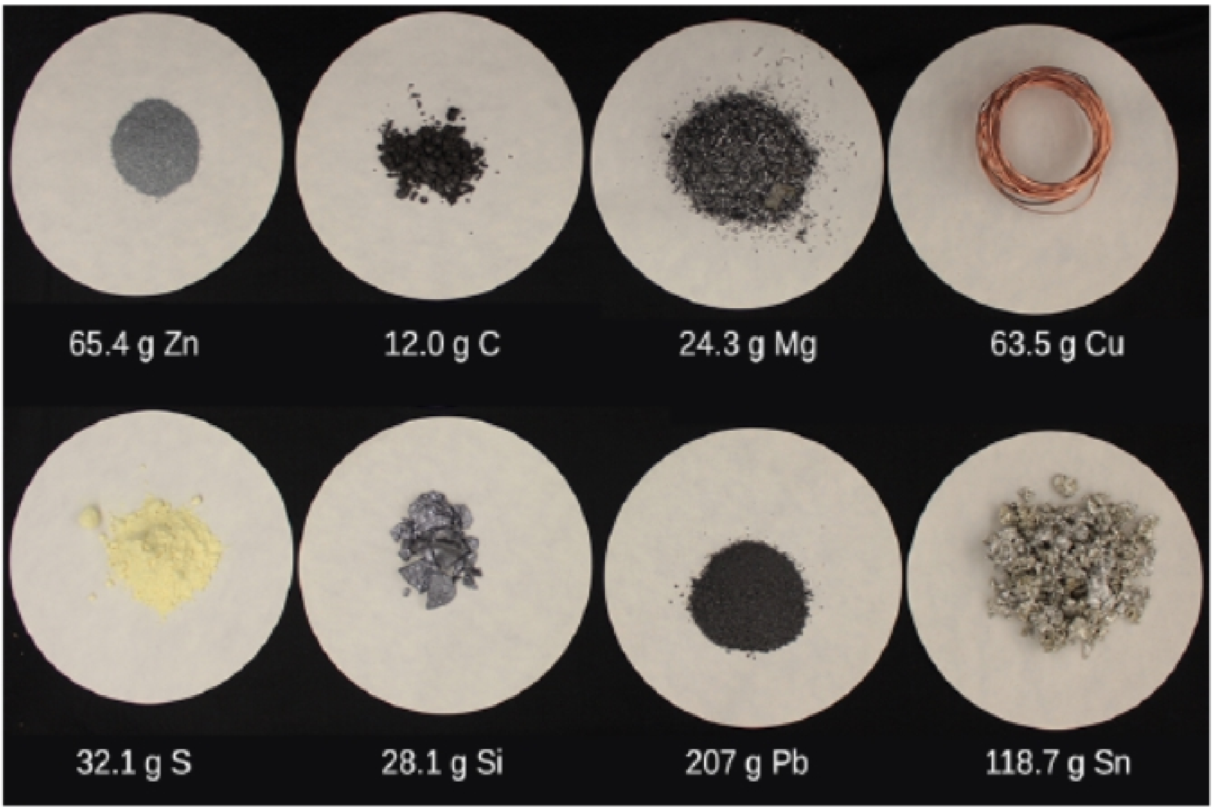


Figure 13.41: Each sample contains one mole of atoms. From left to right (top row): 65.4 [g] zinc, 12.0 [g] carbon, 24.3 [g] magnesium, and 63.5 [g] copper. From left to right (bottom row): 32.1 [g] sulphur, 28.1 [g] silicon, 207 [g] lead, and 118.7 [g] tin (source: Mark Ott)

> Remark
>
> We find these atomic molar masses in the periodic classification. But above all it must be known that those that are indicated take into account the natural isotopes (which is normal since they are chemically indistinguishable excepted for the nuclear chemist or nuclear physicist). So the value indicated in the tables is calculated as the sum of the respective proportions of the molar masses of the different corresponding isotopes (the validity of this method of calculation is obviously relative...).

D3. The "atomic molar mass" is the molar mass of a given element divided by the Avogadro number. Thus:

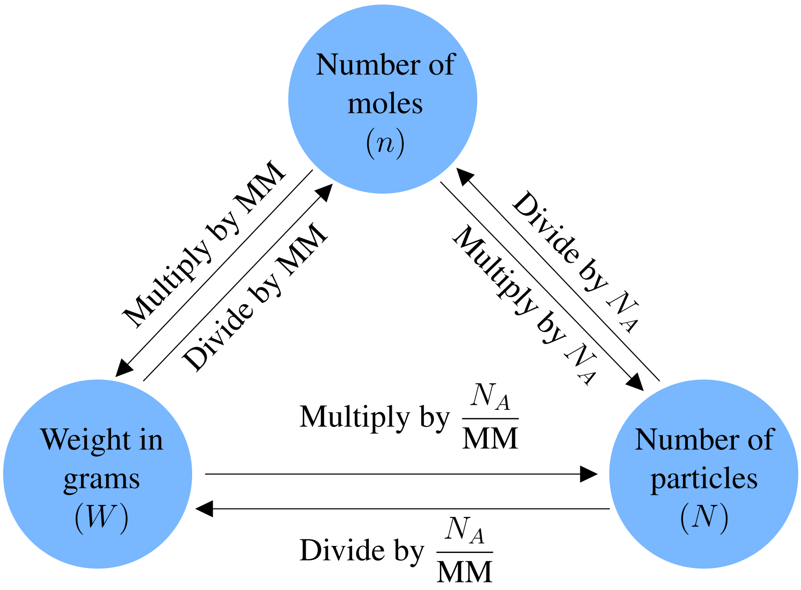
$$1\,[\text{amu}] = \frac{1}{N_\text{A}} \cong 1.6605392 \cdot 10^{-24}\,[\text{kg}] \tag{13.4.13}$$

Therefore the atomic (molar) mass is the mass of 1 atom of a particular element and the molar mass is the mass of 1 mole of an atom or molecule[7].

We therefore have the following graph:

---

[7] The number of molecules in a single droplet of water is roughly 100 billion times greater than the number of people on Earth.


Number of moles (n)
Multiply by MM
Divide by MM
Divide by N_A
Multiply by N_A
Weight in grams (W)
Number of particles (N)
Multiply by $\frac{N_A}{\text{MM}}$
Divide by $\frac{N_A}{\text{MM}}$

D4. The "molecular molar mass (MMM)" is equal to the sum of the atomic molars masses of the chemical elements that constitutes it.

It comes therefore immediately the following observation: the mass m of a sample consisting of an amount of n moles of identical chemical species of molar mass M_m is given by the relation:

$$\boxed{m = nM_m} \tag{13.4.14}$$

Somewhat in a little bit more formal way and in a thermodynamic aspect, here is also is how we can define the molar mass:

Let X be an extensive quantity on a single-phase system (see the section of Thermodynamics for precisions about the vocabulary used) and given a volume element $\mathrm{d}V$ of this system around a common point M and containing the amount of material $\mathrm{d}n$. We associate it the extensive quantity $\mathrm{d}X$ proportional to $\mathrm{d}n$ such that:

$$\mathrm{d}X = X_m \mathrm{d}n \Leftrightarrow X_m = \frac{\mathrm{d}X}{\mathrm{d}n} \tag{13.4.15}$$

so that X_m is an intensive quantity (ratio of two extensive quantities according to what was seen in the section of Thermodynamics) which we will name by definition the "associated molar size" to X.

We conclude that:

$$X = \int X_m \mathrm{d}n \tag{13.4.16}$$

the integral applying on the whole monophasic system.

In the case of a uniform phase, X_m being constant at any point, we can simply write the latter as:

$$X = nX_m \tag{13.4.17}$$

Remark

Basically the idea is to say that the mass of a single-phase chemical system is proportional to the molar mass of it to closely to a given integer factor representing the number of its constituents (or the number of moles to be more exact).

D5. When the system is heterogeneous, we use the concept of "mole fraction", defined by:

$$\boxed{x_i = \frac{n_i}{n}} \tag{13.4.18}$$

x_i being the mole fraction of a species A_i whose the quantity of material (the number of moles for example) is n_i with $n = \sum_i n_i$ being the total quantity of matter of the studied phase.

As a result, for all chemical species of the studied phase, $\sum_i x_i = 1$ which means that if there are n chemical species, it is enough to know $n-1$ molar titles to know them all.

If the studied phase is a gas and assuming a perfect gas according to Boyle's law (approximation of the Van der Waals equation proved in the section of Statistical Mechanics at page 3907)(see page 3907) we have:

$$P_iV = n_iRT \qquad \text{and} \qquad PV = nRT \tag{13.4.19}$$

we therefore have the possibility in the case of gaseous phases to express the mole fraction as:

$$\boxed{x_i = \frac{n_i}{n}\frac{nRT}{nRT} = \frac{P_iV}{PV} = \frac{P_i}{P}} \tag{13.4.20}$$

Remark

We can do obviously the same for the volume V.

D6. We define the "mass content associated with the species A_i" by the ration:

$$\boxed{w_i = \frac{m_i}{m}} \tag{13.4.21}$$

with $m = \sum_i m_i$ being the total mass of the studied phase. We also have of course $\sum_i w_i = 1$.

D7. We define the "volumic molar concentration" or "molarity" the ratio (do not confuse the notation with the specific heat):

$$\boxed{c_i = \frac{n_i}{V}} \tag{13.4.22}$$

Remark

There are other composition variables used much less used than x_i or c_i. We can cite the "mass concentration density" m_i/V, the "molality" (ratio of the amount of material of the species A_i by the total mass of solvent), etc.

D8. We say that a (perfect) gas is in the "standard state" if its pressure is equal to the standard pressure:

$$\boxed{P^\circ = 1\ [\text{bar}] = 10^5\ [\text{Pa}] \cong 1\ [\text{atm}]} \tag{13.4.23}$$

D9. We name "standard molar quantity" of a constituent X_m° the value of the molar quantity of this same component taken in the standard state, that is to say under the pressure P°.

Remarks

R1. Any standard molar quantity is obviously intensive: the pressure being set by the standard state , it depends only on the temperature.

R2. Any standard quantity is denoted with the superscript "°". V_m° is then standard molar volume. For cons, the standard molar quantity is not always specified with the small index m, we must sometimes be careful with what is handled in the equations (as always anyway!).

In the case of the ideal gas, the molar volume is calculated using the ideal gas equation of state. Then we get:

$$\boxed{V_m^\circ = \frac{RT}{P^\circ}} \tag{13.4.24}$$

We see of course that the standard molar volume of an ideal gas depends on the temperature.

If we do that calculation at the "standard conditions of temperature and pressure" (abbreviated STP), that is to say at a temperature of 273.15 [K] (i.e. 0 [°C]) and a pressure of 1 [atm] (i.e. $101,325$ [kPa]), then we find a volume of 22.4 [L·mol^{-1}] which is a well known value by chemists.

Remarks

R1. In a wide range of temperatures and pressures, the molar volume of real gases is generally not very different from that of an ideal gas.

R2. In the case of a condensed state, we do not have in general a state equation but we can measure the molar volume..

We can define then by extension other standard quantities resulting from those we had defined in the section Thermodynamics:

1. The "standard molar internal energy" (intensive quantity as expressed by molar unit) and denoted by U_m°.

2. The "standard molar enthalpy" (intensive quantity as expressed by molar unit) with:

$$H_m^\circ = U_m^\circ + P^\circ V_m^\circ = U_m^\circ + RT \tag{13.4.25}$$

It is important that the reader notice that the enthalpy depends only on the temperature (and the internal energy).

Enthalpies of combustion for many substances have been measured. A few of these are listed in the table below. Many readily available substances with large enthalpies of combustion are used as fuels, including hydrogen, carbon (as coal or charcoal), and hydrocarbons (compounds containing only hydrogen and carbon), such as methane, propane, and the major components of gasoline.

Substance	**Combustion Reaction**	**Combustion Molar Enthalpy $\Delta H^\circ_{c,m}$ in [kJ $\cdot$ mole^{-1}] at 25°**
carbon	$C(s) + O_2(g) \rightarrow CO_2(g)$	-393.5
hydrogen	$H_2(g) + \frac{1}{2}O_2(g) \rightarrow H_2O(l)$	-285.8
magnesium	$Mg(s) + \frac{1}{2}O_2(g) \rightarrow MgO(s)$	-601.6
sulphur	$S(s) + O_2(g) \rightarrow SO_2(g)$	-296.8
carbon monoxide	$CO(g) + \frac{1}{2}O_2(g) \rightarrow CO_2(g)$	-283.0
methane	$CH_4(g) + 2O_2(g) \rightarrow CO_2(g) + 2H_2Ol$	-890.8
acetylene	$C_2H_2(g) + \frac{5}{2}O_2(g) \rightarrow 2CO_2(g) + H_2O(l)$	-1301.1
ethanol	$C_2H_5OH(l) + 3O_2(g) \rightarrow 2CO_2(g) + 3H_2O(l)$	-1366.8
methanol	$CH_3OH(l) + \frac{3}{2}O_2(g) \rightarrow CO_2(g) + 2H_2O(l)$	-726.1
isooctane	$C_8H_{18}(l) + \frac{25}{2}O_2(g) \rightarrow 8CO_2(g) + 9H_2O(l)$	-5461

Table 13.5: **Different molar enthalpy of combustion (source: OpenStax)**

Remark

For condensed states the standard volume is very low in S.I. units so that $H_m^\circ \cong U_m^\circ$. However, it is very difficult to speak of pressure for so condensed states so this approximation has to be used with caution.

If we now consider an extensive function X (as for example the volume!) defined on a chemical gaseous evolving system. We can a priori express X based on two intensive variables T, P (because an extensive function is always a product or ratio of two intensive quantities, or a sum of extensive quantities) and of the different quantity of materials n_i, n_i' of A_i, A_i' such that:

$$X = X(T, P, \ldots, n_i, \ldots) \tag{13.4.26}$$

If all products (reagents and resulting one) are in their standard state, the extensive function, therefore denoted X°, gets the form:

$$X^\circ = X(T, \ldots, n_i, \ldots) \tag{13.4.27}$$

where the pressure is no longer involved as attached to its standard value. The gas is then described by its temperature and the quantity of its constituents!

However, if we consider an infinitesimal evolution of the system at constant temperature and pressure (because we assume a very slow transformation) the different quantities of materials vary therefore following the exact total differential (see section of Differential Calculus and Integral page 1466):

$$\mathrm{d}X_T^\circ=\sum_i \frac{\partial X^\circ}{\partial n_i}\mathrm{d}n_i \tag{13.4.28}$$

where obviously are taken into account, as T and P are supposed constant, only the quantity of materials that could vary (yes don't forget we are doing chemistry!!!).

We can then define artificially (nothing prevents us to do so, it's not false!) the intensive standard molar quantity that depends only on the temperature:

$$X_i^\circ=\frac{\partial X^\circ}{\partial n_i} \tag{13.4.29}$$

Therefore:

$$\mathrm{d}X_T^\circ=\sum_i X_i^\circ \mathrm{d}n_i \tag{13.4.30}$$

but we have also defined in the section Analytical Chemistry the material balance equation (see page 6244):

$$\mathrm{d}n=v\mathrm{d}\xi \tag{13.4.31}$$

expressing, for recall, the variation in the quantity of matter of one of the compounds of a chemical reaction relatively to its stoichiometric ratio (constant) and the progress of the reaction. We therefore have:

$$\mathrm{d}X_T^\circ=\left(\sum_i X_i^\circ v_i\right)\mathrm{d}\xi \tag{13.4.32}$$

and also that:

$$\begin{aligned}\mathrm{d}X_T^\circ&=\sum_i \frac{\partial X^\circ}{\partial n_i}\mathrm{d}n_i=\sum_i \frac{\partial X^\circ}{v_i\partial \xi}v_i\mathrm{d}\xi=\left(\sum_i \frac{\partial X^\circ}{\partial \xi}\right)\mathrm{d}\xi=\left(X_i^\circ v_i\right)\mathrm{d}\xi\\ \Rightarrow \frac{\mathrm{d}X_T^\circ}{\mathrm{d}\xi}&=\sum X_i^\circ v_i=\sum_i \frac{\partial X^\circ}{\partial \xi}\end{aligned} \tag{13.4.33}$$

By definition, we name this algebraic sum "standard quantity reaction associated with the extensive function X" and denote it by (notation badly chosen by chemists in our point of view...):

$$\Delta_r X^\circ=\sum_i X_i^\circ v_i \tag{13.4.34}$$

which is an intensive quantity that depends only on the temperature and represents a relative change (hence the subscript r!). This relation can also be written:

$$\boxed{\Delta_r X^\circ=\sum_i X_i^\circ v_i=\sum_i \frac{\partial X^\circ}{\partial \xi}=\frac{\mathrm{d}X_T^\circ}{\mathrm{d}\xi}} \tag{13.4.35}$$

In general, chemists name "Lewis operator", denoted Δ_r, the derivative of a quantity X (standardized or not), with respect to the progress of the reaction ξ with constant temperature and pressure.

Remark

The symbol Δ_r appears with the letter r in subscript to show that this is a relative reaction quantity. In other words, it is the standard variation of the molecular quantity during the concerned reaction for a given reaction progress of one mole at a pressure of 1 bar for a perfect gas.

We must also not forget that the stoichiometric coefficients of the reactants are positive and those of the resulting products are negative (see section of Analytical Chemistry page 6242).

There are two reaction quantities that play important roles in chemistry:

1. The internal molar energy of reaction, named often "internal energy of standard reaction" of a chemical system:

$$\boxed{\Delta_r U^\circ = \sum_i U_i^\circ v_i = \frac{\mathrm{d}U_T^\circ}{\mathrm{d}\xi}} \tag{13.4.36}$$

2. The molar enthalpy of reaction, often named more "standard enthalpy of reaction" of a chemical system:

$$\boxed{\Delta_r H^\circ = \sum_i H_i^\circ v_i = \frac{\mathrm{d}H_T^\circ}{\mathrm{d}\xi}} \tag{13.4.37}$$

13.4.2.1 Standard enthalpy of reaction

Therefore we can consider the following two cases after knowing the relation (see section Thermodynamics page 3643):

$$H_i^\circ = U_i^\circ + V_i^\circ P^\circ \tag{13.4.38}$$

1. If the A_i are in a condensed state, since the internal pressure does not apply we have:

   $$H_i^\circ \cong U_i^\circ \tag{13.4.39}$$

   which still remains to be taken with precaution following the scenarios!

2. If the A_i are in the gaseous state (assumed perfect gas):

   $$H_i^\circ \cong U_i^\circ + RT \tag{13.4.40}$$

We conclude that only gas will intervene in this relation:

$$\Delta_r H^\circ = \sum_i H_i^\circ v_i = \sum_i (U_i^\circ + RT) = \Delta_r U^\circ + \sum_i v_i RT \tag{13.4.41}$$

That we write conventionally:

$$\boxed{\Delta_r H^\circ = \Delta_r U^\circ + \Delta_r v_g RT} \qquad (13.4.42)$$

It follows that in the special case where:

$$\Delta_r v_g = 0 \qquad (13.4.43)$$

(which is in fact an unfortunate notation... for the algebraic sum of the stoichiometric ratio that would equal to zero) for a given temperature then we have:

$$\Delta_r H^\circ = \Delta_r U^\circ \qquad (13.4.44)$$

where it must be remembered that the stoichiometric coefficients of the products are counted as positive, while those of the reactants are counted as negative (see section Analytical Chemistry page 6242).

Thus, the variation of the enthalpy function corresponds to the variation of the quantity of heat absorbed or emitted in an isobaric transformation at a given temperature T. This is why it is sometimes denoted $\Delta_r H_{T,P^\circ}$.

A chemical reaction that has an enthalpy reaction (which is for recall the instantaneous change in enthalpy during a reaction) that is negative is said to be "exothermic", since it releases heat into the environment (constant pressure obliged by the definition of enthalpy reaction!), then a chemical reaction whose reaction enthalpy is positive is say to be "endothermic" since it then requires a supply of heat to occur (so the vocabulary is the same as in the section of Thermodynamics).

Thus, according to the preceding developments, if we denote with an index p the products and with an index i the reactants, we often find the standard enthalpy of reaction as follows if the stoichiometric coefficients are counted as positive:

$$\boxed{\Delta_r H_{T,P^\circ} = \sum_p v_p H_p^\circ - \sum_p v_r H_r^\circ} \qquad (13.4.45)$$

Into this form, then we see well that the standard reaction enthalpy corresponds to the difference partial molar enthalpies between the products and reactants of the transformation. This is nothing more than the "Hess's law" set in the 119th century (holocene calendar) by the Swiss chemist Henri Hess. This law states that the total enthalpy change during the complete course of a chemical reaction is the same whether the reaction is made in one step or in several steps and can be understood as an expression of the principle of conservation of energy, also expressed in the first law of thermodynamics, and the fact that the enthalpy of a chemical process is independent of the path taken from the initial to the final state (i.e. enthalpy is a state function).

Because in a system at equilibrium, the initial energy is always greater than or equal to the final energy (all systems tend to move towards to a more stable state with minimum energy as we have study it in the section of Thermodynamics), then the standard enthalpy of reaction $\Delta_r H^\circ$ may be only negative or zero.

If a chemical reaction at constant pressure and at a specific temperature gives only a single chemical compound (product) then the standard enthalpy of reaction is named "standard enthalpy of formation" and is denoted by $\Delta_f H^\circ$.

In fact, the interest of prior-previous relation is that the chemist can simply, without having to know the quantities of material involved, determine just by knowing the stoichiometric coefficients of an isobaric gas or condensed chemical reaction (if he agrees that it will then be an approximation for the latter case) that the instantaneous variation of the molar internal energy during the progress of the reaction at a given temperature is equal to the instantaneous variation of the molar enthalpy .

Two different situations arise then:

1. The difference between the instantaneous variation of the molar internal energy and the molar enthalpy is zero: therefore, the chemical reaction (at a given temperature) does not instantaneously occupies a larger volume and thus don't loss energy to push ("unnecessarily") the pressure of the gas surrounding the studied reaction (this can be seen as a money saving in terms of energy in the chemical industry). In this case, the standard enthalpy of reaction is simply equal to the heat of reaction at constant pressure Q_p.
2. The difference between the instantaneous variation of the molar internal energy and the molar enthalpy is positive: Therefore, the chemical reaction (at the given temperature) instantly occupies a greater volume and thus loses some energy to push ("unnecessarily") the pressure of the gas surrounding the studied reaction (this can be seen as a waste of money in terms of energy efficiency in the chemical industry).

Remark

Obviously, it is possible to imagine a company that takes advantage of the change in volume of a reaction (case 2 above) that pushes the surrounding gas with a piston system to then produce mechanical energy... so it would be possible in certain situations to lose much less money (verbatim energy ...).

However a small difficulty arise, ... the standard enthalpy of a simple pure body (body formed of a single type of atom) can not be calculated in absolute terms because it depends on the internal energy which is very difficult to calculate (you must use the tools of quantum physics which raise insurmountable problems even in the beginning of the 121st century according to holocene calendar). This means we must define an arbitrary scale of molar enthalpies by setting an arbitrary zero enthalpy and adopted internationally (which is unfortunately not the case as far as we know!).

Thus, in order to set up tables of standard molars enthalpy, he was chosen to define the scale of enthalpy as follows: the standard molar enthalpy of a simple steady pure body in the standard state is equal to 0 at 298 [K]. It follows that the enthalpy of formation of a simple standard pure body is always equal to zero.

✍**Example:**

Given the reaction:

$$\mathrm{PCl_5} \rightleftharpoons \mathrm{PCl_3} + \mathrm{Cl_2} \tag{13.4.46}$$

That is to say, the dissociation of chlorine and phosphorus pentachloride in phosphorus trichloride. The tables give us at the temperature of $T = 1000$ [K] the following value of the standard molar enthalpy of this reaction:

$$\Delta_r H_m^\circ = 156 \, [\text{kJ} \cdot \text{mol}^{-1}] \tag{13.4.47}$$

The variation of the value of the molar enthalpy of reaction being positive, it follows that the reaction is endothermic (requires a heat input hence the dissociation temperature of 1000 [K]) and therefore the product is more volatile than the initial reactant.

We have the following algebraic sum of the stoichiometric coefficients of the reaction:

$$1 \cdot \text{PCl}_5 \rightleftharpoons 1 \cdot \text{PCl}_3 + 1 \cdot \text{Cl}_2 \tag{13.4.48}$$

That is (which is dimensionless since enthalpy is in molar value!):

$$\Delta_r v_g = -1 + 1 + 1 = 1 \tag{13.4.49}$$

therefore the reaction increase the pressure by creating an additional mole per mole of reactant.

Since:

$$\Delta_r H_m^\circ = \Delta_r U_m^\circ + \Delta_r v RT = \Delta_r U^\circ + 1 \cdot RT \tag{13.4.50}$$

then it comes:

$$\Delta_r U^\circ = 148 \, [\text{kJ} \cdot \text{mol}^{-1}] \tag{13.4.51}$$

This is then the part of internal energy absorbed by the system on the $156 \, [\text{kJ} \cdot \text{mol}^{-1}]$. The remainder (difference) is has just been used to push the surrounding atmosphere of the chemical reactor.

13.4.2.1.1 Kirchhoff's Enthalpy Law

Kirchhoff's Enthalpy Law describes the enthalpy of a reaction's variation with temperature changes. In general, enthalpy of any substance increases with temperature, which means both the products and the reactants' enthalpies increase. The overall enthalpy of the reaction will change if the increase in the enthalpy of products and reactants is different.

In other words, in a more practical way, the latent heat - energy required to evaporate a liquid - is not the same at every temperature! The difference between the Gas and Liquid energy levels increases at higher temperatures. Thus, the Kirchhoff's Enthalpy law enables the calculation of a new latent heat from an existing one with a known temperature change.

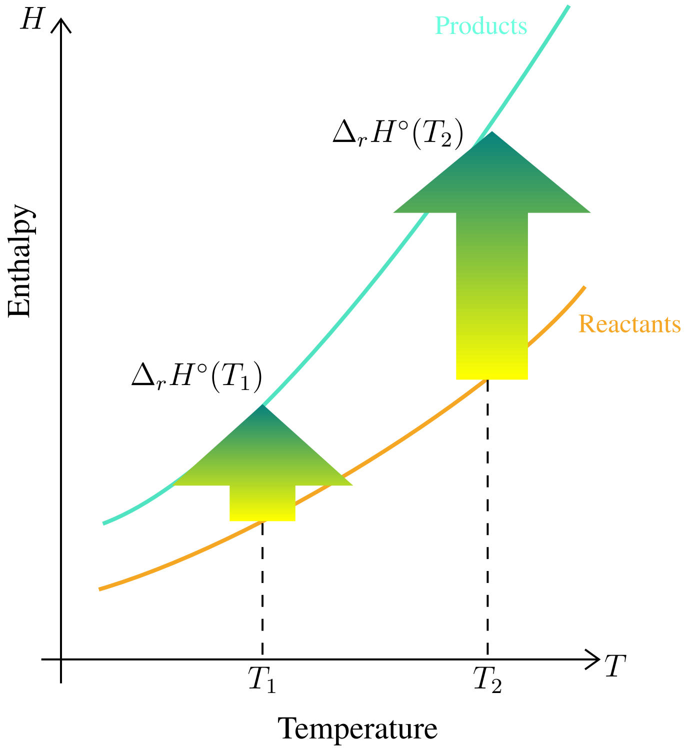


Figure 13.42: **Kirchhoff's enthalpy law illustration**

So the Kirchhoff enthalpy law idea is to express the variations of the enthalpy of reaction (molar or not) in function of the temperature from the knowledge of the heat capacity at constant pressure of the gaseous reactants.

He have built in previous developments the following relation:

$$\Delta_r H^\circ = \sum_i H_i^\circ v_i = \frac{\mathrm{d}H_T^\circ}{\mathrm{d}\xi} \tag{13.4.52}$$

which is the standard enthalpy of reaction at a given temperature in a system with a standard pressure.

We had also mentioned earlier above that Δ_r, for recall, is somewhat an unfortunate notation for the differential (Lewis) operator of progress of the reaction $\mathrm{d}/\mathrm{d}\xi$.

If we focus on the influence of the temperature T on $\Delta_r H^\circ$ we have just to write the exact differential:

$$\frac{\Delta_r H^\circ}{\mathrm{d}T} = \frac{\mathrm{d}}{\mathrm{d}T}\left(\sum_i H_i^\circ v_i\right) \tag{13.4.53}$$

since the algebraic variation of the standard enthalpy by definition depends only on the temperature.

The stoichiometric coefficients v_i are not dependent of the temperature at least until this latter does not changes the essence itself of the studied transformation.

We then have under this approximation (assumption):

$$\frac{\Delta_r H^\circ}{\mathrm{d}T}=\frac{\mathrm{d}}{\mathrm{d}T}\left(\sum_i H_i^\circ v_i\right)=\sum_i v_i\frac{\mathrm{d}H_i^\circ}{\mathrm{d}T} \tag{13.4.54}$$

Now we have defined in the section of Thermodynamics (see page 3622) the heat capacity at constant pressure which is written:

$$C_P=\left(\frac{\partial Q}{\partial T}\right)_P=\frac{\partial H}{\partial T} \tag{13.4.55}$$

So if the conditions are standard (the enthalpy therefore depends only on the temperature), we get is the exact differential:

$$C_{P,i}^\circ=\frac{\mathrm{d}H_i^\circ}{\mathrm{d}T} \tag{13.4.56}$$

Then we have:

$$\frac{\mathrm{d}\left(\Delta_r H^\circ\right)}{\mathrm{d}T}=\sum_i v_i\frac{\mathrm{d}H_i^\circ}{\mathrm{d}T}=\sum_i v_i C_{P,i}^\circ=\Delta_r C_P^\circ \tag{13.4.57}$$

We can of course integrate the latter relations to get the common form use in practice and available in many books:

$$\int_{T_0}^{T}\mathrm{d}\left(\Delta_r H^\circ\right)=\int_{T_0}^{T}\Delta_r C_P^\circ \mathrm{d}T \tag{13.4.58}$$

Then we have:

$$\boxed{\Delta_r H^\circ(T)=\Delta_r H^\circ(T_0)+\int_{T_0}^{T}\Delta_r C_P^\circ \mathrm{d}T} \tag{13.4.59}$$

where T_0 is a particular temperature for which $\Delta H^\circ(T_0)$ is known.

In a temperature range very close to T_0 chemists sometimes approximate the variation as being linear. That is equivalent to put:

$$\Delta_r C_P^\circ \cong \alpha \tag{13.4.60}$$

It then immediately comes from the prior-previous relation:

$$\Delta_r H^\circ(T)=\Delta_r H^\circ(T_0)+\alpha\int_{T_0}^{T}\mathrm{d}T=\Delta_r H^\circ(T_0)+\alpha(T-T_0) \tag{13.4.61}$$

Remark

Quite often, the variation of the enthalpy of reaction with temperature is negligible!

✍**Example:**

For the reaction (graphite + oxygen yielding to carbon dioxide) we would like to know $\Delta_r H^\circ$ at 1000 [K]:

$$\mathrm{C}+\mathrm{O}_2 \rightarrow \mathrm{CO}_2 \qquad (13.4.62)$$

For this, it is given in the tables for this reaction at 298 [K]:

$$\Delta_r H_m^\circ(298) = -393.5\,[\mathrm{kJ}\cdot\mathrm{mol}^{-1}] \qquad (13.4.63)$$

and:

$$c_{P,\mathrm{C}}^\circ = 8.5\,[\mathrm{J}\cdot\mathrm{K}^{-1}\cdot\mathrm{mol}^{-1}] \quad c_{P,\mathrm{O}_2}^\circ = 29.4\,[\mathrm{J}\cdot\mathrm{K}^{-1}\cdot\mathrm{mol}^{-1}] \quad c_{P,\mathrm{CO}_2}^\circ = 37.1\,[\mathrm{J}\cdot\mathrm{K}^{-1}\cdot\mathrm{mol}^{-1}] \qquad (13.4.64)$$

We write in lowercase the heat capacities above as there are enough subscripts to not add a third one (m) that mean these are molar heat capacities.

Remark

When the enthalpy of reaction is given at the reference temperature (nowadays...) at 298 [K] chemists then speak as we have already mention earlier above the "standard enthalpy of formation".

The value of the molar enthalpy of reaction being negative, it follows that the reaction is exothermic (it is tendency of nature to favour exothermic reactions to stabilize systems in their minimal energy states).

We then have immediately:

$$\begin{aligned}\Delta_r H_m^\circ(T) &= \Delta_r H_m^\circ(T_0) + \int_{T_0}^{T} \Delta_r c^\circ \mathrm{d}T = \Delta_r H_m^\circ(T_0) + \int_{T_0}^{T} \left(c_{P,\mathrm{C}}^\circ - c_{P,\mathrm{O}_2}^\circ - c_{P,\mathrm{CO}_2}^\circ\right)\mathrm{d}T \\ &\cong \Delta_r H_m^\circ(T_0) + \left(c_{P,\mathrm{C}}^\circ - c_{P,\mathrm{O}_2}^\circ - c_{P,\mathrm{CO}_2}^\circ\right)\int_{T_0}^{T}\mathrm{d}T \\ &= \Delta_r H_m^\circ(T_0) + \left(c_{P,\mathrm{C}}^\circ - c_{P,\mathrm{O}_2}^\circ - c_{P,\mathrm{CO}_2}^\circ\right)(T-T_0) \\ &= -393,500 + (37.1-8.5-29.4)(1,000-298) = -394,060\,[\mathrm{J}\cdot\mathrm{mol}^{-1}]\end{aligned} \qquad (13.4.65)$$

Thus the variation is of $-560\,[\mathrm{kJ}\cdot\mathrm{mol}^{-1}]$, that is a variation of about $+0.1\%$. It follows that the higher the temperature increase, more is the reaction exothermic. In fact, the choice of this particular temperature of $1,000$ [K] is not innocent because it is from this temperature that experiments shows that the reaction also produces carbon monoxide.

We can also conclude that some exothermic reactions and having a enthalpy of reaction that decreases rapidly with temperature can blow up!

Finally, let us indicate that in practice we often use the term "calorific power" or "heat of combustion", which is simply the enthalpy of reaction per unit mass of fuel or the energy obtained by the combustion a kilogram of fuel.

Thus, for Gasoline, we have following what give tables (under the assumption that this number is correct):

$$\Delta_r H^\circ(298) = -47.3\,[\text{MJ}\cdot\text{kg}^{-1}] \tag{13.4.66}$$

And we can have fun by calculating the amount of Gasoline needed to accelerate a car of 1000 [kg] from 0 to 100 $[\text{km}\cdot\text{h}^{-1}]$ with a yield of $\eta = 35\%$ at a temperature of 293 [K]. Thus we have:

$$M = \frac{E_c}{\Delta_r H^\circ(T)\eta} = \frac{\dfrac{1}{2}mv^2}{\Delta_r H^\circ(298)\eta} \cong \frac{392\cdot 10^3}{47.3\cdot 10^6\cdot 35\%} \cong 23.67\,[\text{g}] \tag{13.4.67}$$

and to get the amount of fuel in litres the tables give us the for Gasoline density about 700 $[\text{kg}\cdot\text{m}^{-3}]$ which gives finally a volume in litres of:

$$V \cong \frac{0.00828}{700}\cdot 1000 = 0.033\,[\text{L}] \tag{13.4.68}$$

~20 %

★★★☆☆

23 votes, 58.26%

14

Theoretical Computing

The theoretical computer science is the branch of science that deals with the development of algorithms and theoretical tools that can be apply to IT to solve formal problems related to the simulation of physical phenomena or data treatments and the exchange of information. (Larousse)

Contents

14.1 Numerical Methods/Analysis . . . 6272

14.1.1 Computer Representation of Numbers . . . 6274

14.1.1.1 Decimal System . . . 6274

14.1.1.2 Binary system . . . 6275

14.1.1.3 Hexadecimal System . . . 6280

14.1.1.4 Octal System . . . 6280

14.1.1.5 Conversion of decimal system to non-decimal system: . . . 6282

14.1.2 Algorithm Complexity . . . 6284

14.1.2.1 NP-Completude . . . 6291

14.1.3 Integer Part . . . 6293

14.1.4 Heron's Square Root Algorithm . . . 6295

14.1.5 Archimedes Algorithm . . . 6298

14.1.6 Euler's Number e . . . 6301

14.1.7 Stirling's factorial approximation . . . 6301

14.1.8 Linear Systems of Equations . . . 6304

14.1.8.1 One equation with one unknown . . . 6304

14.1.8.2 Two equations with two unknowns . . . 6304

14.1.8.3 Three equations with three unknowns . . . 6306

14.1.8.4 n equations with n unknowns . . . 6310

14.1.9 Polynomials . . . 6312

14.1.10 Regression Techniques . . . 6316

14.1.10.1 Univariate linear regression model 6322
14.1.10.2 Univariate linear regression Gaussian Model 6348
14.1.10.3 Linear univariate regression forced through the origin (RTO) 6369
14.1.10.4 Deming regression (orthogonal regression) 6372
14.1.11 Multiple linear regression Gaussian Model 6379
14.1.11.1 Fisher Matrix 6389
14.1.11.2 R^2 decomposition in multiple regression 6390
14.1.11.3 Influential points 6391
14.1.11.4 Significance of regression coefficients (variable importance) 6395
14.1.11.5 Variance Inflation Factor (multicollinearity detection) 6398
14.1.11.6 Weighted Least Squares (WLS) 6407
14.1.11.7 Model Selection Criteria 6412
14.1.12 Ridge, LASSO and Elastic Net regularization 6435
14.1.12.1 L_2 Regularization or Ridge Regularization 6437
14.1.12.2 L_1 Regularization or LASSO Regularization 6441
14.1.12.3 $L_1 + L_2$ Regularization or Elastic Net Regularization 6443
14.1.12.4 Polynomial regression 6445
14.1.12.5 Kernel regression 6447
14.1.12.6 Logistic regressions (LOGIT) 6449
14.1.12.7 General Linear Models (GLM) 6530
14.1.12.8 Robust, M-estimators and W-estimators 6557
14.1.13 Interpolation Techniques 6569
14.1.13.1 Bézier Curves (Bézier Splines) 6569
14.1.13.2 Linear ordering isotonic regression 6582
14.1.13.3 Euler Method 6587
14.1.13.4 Polynomial of collocation 6591
14.1.13.5 Lagrange polynomials 6594
14.1.13.6 Newton polynomials (divided differences) 6597
14.1.13.7 Errors in Polynomial Interpolation 6602
14.1.14 Roots search 6603
14.1.14.1 Proportional parts methods 6603
14.1.14.2 Secant method (Regula Falsi or False Position) 6606
14.1.14.3 Bisection method (interval-halving) 6609
14.1.14.4 Newton's method 6612
14.1.15 Numerical Differentiation 6619
14.1.16 Numerical Integration 6621
14.1.16.1 Rectangles method 6623

14.1.16.2 Trapezoidal method . . . 6624
14.1.16.3 Newton–Cotes formulas . . . 6626
14.1.16.4 Multidimensional integrals . . . 6632
14.1.16.5 Numerical solution of ordinary differential equations . . . 6637
14.1.17 Optimization . . . 6645
14.1.17.1 Linear programming (Linear Optimization) . . . 6646
14.1.17.2 Nonlinear programming (Nonlinear optimization) . . . 6662
14.1.17.3 Expectation-Maximization (EM) algorithm . . . 6697
14.1.18 Resampling statistics . . . 6707
14.1.18.1 Monte Carlo Simulations . . . 6708
14.1.18.2 Bootstrapping . . . 6740
14.1.18.3 Jackknife Resampling . . . 6747
14.1.18.4 Permutation Tests . . . 6749
14.1.19 Finite difference method (F.D.M.) . . . 6752
14.1.19.1 One space dimension F.D.M. . . . 6753
14.1.19.2 Space-time F.D.M (finite-volume method) . . . 6759
14.1.20 Data Science & Data Mining . . . 6768
14.1.20.1 Training vs Test Data . . . 6795
14.1.20.2 Hyperparameters vs Parameters . . . 6803
14.1.20.3 Association Rules . . . 6808
14.1.20.4 Equivalence CLAss Transformation (ECLAT) . . . 6820
14.1.20.5 Clustering and Classification . . . 6822
14.1.20.6 Dimensionality Reduction . . . 6918
14.1.20.7 Gradient Boosting . . . 6945
14.1.20.8 Neural networks . . . 6951
14.1.20.9 Genetic Algorithms . . . 7022
14.1.20.10 Total Unduplicated Reach and Frequency Analysis (TURF) . . . 7036
14.2 Fractals . . . 7038
14.2.1 IFS Fractals . . . 7041
14.2.1.1 Fractals Metric Space . . . 7054
14.2.2 Fractals Visualization . . . 7059
14.2.2.1 Cantor's Fractal (Cantor Set) . . . 7059
14.2.2.2 Triangle Sierpinski Fractal . . . 7062
14.2.2.3 Sierpinski carpet fractal . . . 7067
14.2.2.4 Fractal spirals . . . 7071
14.2.2.5 Von Koch fractal (Koch snowflake) . . . 7073
14.2.2.6 Natural fractals . . . 7078

14.2.3 Escape Time Algorithm Fractals . . . 7091
14.2.3.1 Mandelbrot Set . . . 7093
14.2.3.2 Julia Set . . . 7098
14.2.3.3 Newton Set . . . 7102
14.3 Logical Systems . . . 7104
14.3.1 Strict Logic . . . 7104
14.3.1.1 Boolean Algebra . . . 7105
14.3.1.2 Logical Functions (Boolean operators) . . . 7112
14.3.1.3 Karnaugh maps . . . 7116
14.3.1.4 Arithmetic Boolean (binary) operations . . . 7125
14.3.2 Fuzzy logic . . . 7136
14.3.2.1 Fuzzy set . . . 7147
14.4 Error-Correcting Codes . . . 7152
14.4.1 CheckSum . . . 7156
14.4.1.1 Luhn algorithm . . . 7157
14.4.2 Check Digit . . . 7159
14.4.2.1 European Article Numbering (EAN-13) . . . 7159
14.4.2.2 Swiss Post Payment slip . . . 7161
14.4.2.3 International Bank Account Number (IBAN) . . . 7163
14.4.2.4 UIC wagon numbers . . . 7164
14.4.3 Permutations . . . 7166
14.4.4 Encoders . . . 7167
14.4.4.1 Block code . . . 7172
14.4.4.2 Systematic codes . . . 7179
14.5 Automata Theory . . . 7181
14.5.1 Von Neumann machine . . . 7182
14.5.2 Turing machine . . . 7184
14.5.3 Chomsky hierarchy . . . 7191
14.5.3.1 Formal language . . . 7191
14.5.3.2 Syntax . . . 7193
14.5.3.3 Grammar . . . 7194
14.5.3.4 Associated automata . . . 7197
14.5.4 Terminology . . . 7198
14.5.4.1 Words . . . 7198
14.5.4.2 Languages . . . 7201
14.5.4.3 Equations . . . 7203
14.5.4.4 Codes . . . 7205

14.5.5 Linguistic algorithms . . . 7209
14.5.5.1 Huffman algorithm . . . 7209
14.5.5.2 Sardinas and Patterson algorithm . . . 7214
14.6 Cryptography . . . 7216
14.6.1 Cryptographic systems . . . 7217
14.6.1.1 Kerckhoffs' principle . . . 7221
14.6.2 Traps . . . 7222
14.6.3 Secret-key encryption system . . . 7223
14.6.3.1 Feistel Schemes . . . 7225
14.6.4 Public key encryption . . . 7230
14.6.4.1 Diffie-Hellman protocol . . . 7232
14.6.4.2 R.S.A system . . . 7236
14.6.5 Hash functions . . . 7249
14.6.5.1 MD5 message digest condensation function . . . 7249
14.6.5.2 SHA-1 Secure Hash Algorithm condensation function . . . 7253
14.6.6 Certificate based authentication . . . 7255
14.6.7 Quantum cryptography . . . 7259
14.6.8 Alternative cryptography . . . 7262
14.7 Quantum Computing . . . 7263
14.7.1 Schrödinger's Cat superposition . . . 7265
14.7.2 Photon polarization . . . 7267
14.7.3 Qubit . . . 7269
14.7.3.1 Bloch sphere . . . 7275
14.7.4 Entangled qubit system . . . 7288
14.7.5 Quantum logic gates . . . 7288

14.1. Numerical Methods/Analysis

THE numerical analysis is a mathematical discipline. It is interested in both theoretical foundations as the implementation methods to resolve by purely numerical calculations and empirical approach, mathematical analysis problems.

Definition (#14.1.551): "Numerical Methods/Analysis" is the study of algorithms or empirical scientific methods for solving or analysing mathematical continuous or discrete problems. This means that it mainly deals to respond numerically to real or complex variables questions like numerical linear algebra over the real or complex fields, looking for numerical solutions of differential equations and other problems occurring in the physical sciences or financial/statistical engineering.

Some continuous mathematical problems can be solved exactly by an algorithm. These algorithms are then named "direct methods". Examples are the elimination of Gauss-Jordan for solving a system of linear equations or the simplex algorithm of linear programming (see further below). However, for some problems no direct method are known (and is even proved that for a class of problems known as "NP complete" - see further below - there is no direct calculation with finished algorithm in a polynomial time). In such cases, it is sometimes possible to use an iterative method to attempt to determine an approximation of the solution. Such a method starts from a guessed value or roughly estimated one and finds successive approximations that should converge to the solution under certain conditions. Even when a direct method exists, however, an iterative method may be preferable because it is often more effective and often more stable (in particular it allows most often to correct minor errors in intermediate calculations).

The use of numerical analysis has been greatly facilitated by modern computers. The Increasing availability and power of computers since the second half of the 120th century (holocene calendar) allowed the application of numerical methods in many scientific, technical and economic, often with revolutionary, accurate and significant effects.

What is the meaning of life?
I don't know. The computers are down.

In numerical simulations of physical systems (multi-physics), the initial conditions are very important in solving differential equations (see the various sections of this textbook where chaotic effects appears). The fact that we can not know them exactly implies that the results of calculations can never be perfectly accurate (we know this fact very well for the weather forecasting which is the most glaring example known). This effect is a consequence of the results of fundamental physics (based on pure mathematics) which demonstrates that we can not perfectly know a system by performing measurements since it directly disrupts it (Heisenberg uncertainty principle as see in the section of Quantum Wave Theory) and these disturbances are the subject of Chaos Theory (classical or quantum).

With new computer tools available in the early 121st century (holocene calendar), it became practical and exciting to know the numerical methods to have fun with some softwares or programming languages (OpenGL, 3D Studio Max, Blender, Maple™, MATLAB®, Mathematica®, COMSOL Multiphysics®, R, C++, Python™, etc.) to simulate 2D or 3D physical systems graphically.

Remarks

R1. Many numerical methods used in computer science are based on arguments which we have already been studied in other sections of this textbook. We won't come back on this methods.

R2. This section being on the boundary between engineering and Applied Mathematics, we decided to give some application examples of developed tools with various programming languages or software.

R3. Many of the techniques presented below are available as complete code in the C++ book [423] that we strongly recommend to the reader.

Definitions (#14.1.552):

D1. An "algorithm" is a finite sequence of rules to be applied in a specific order in a finite number of data to arrive in a finite number of steps (including the amount, or conversely the execution time is defined by the term "cost") to a certain result, independently of the data type.

D2. The algorithms are integrated into computers through "programs" (including stuff like "functions", "objects", "classes", "methods", "properties", "pointers", etc. but this is more related to a programming language course) that are the realization (implementation) of an algorithm using a given language (on a given architecture). This is the implementation of the principle.

Axioms of programming (anecdotal):

A1. More we write code, more errors we will produce.

A2. There are no programs without possible errors (due to the program itself, to the underlying electronics or most often to the user himself).

Remark

Basically there a is minimum steps to follow when developing an algorithm and its corresponding code. A very good process to follow is the one proposed by the ISO/CEI 9126 norm.

Theorem 14.1. There exist an infinite number of algorithms.

Proof 14.1.1. For every integer $i \geq 1$, consider the following algorithm A_i : Declare variable x, then assign x the value i. Terminate.

Every algorithm defined, as above, terminates in finite time. Every algorithm defined above is distinct.

Notice that we have defined an algorithm, one for each positive integer (in technical language, there is a correspondence between the family of algorithms we defined above and $\mathbb{Z}^+$). There are infinitely many positive integers, therefore, there are infinitely many algorithms (for which at least a subset of them include the family of the algorithms we defined above).

■ Q.E.D.

When developing a scientific algorithm, it may be interesting, rigorous and even an obligation in very high level cases to analyse the "complexity" of the algorithm. Without going too far, let's see what it is exactly:

14.1.1 Computer Representation of Numbers

In a classical computer numbers are represented by binary digits 0 and 1. Computers employ binary arithmetic for performing operations on numbers. Since it gets cumbersome to display large numbers in binary form computers usually display them in hexadecimal or octal or decimal system. All of these number systems are positional systems. In a positional system a number is represented by a set of symbols. Each of these symbols denote a particular value depending on its position. The number of symbols used in a positional system depends on its 'base'. Let us now discuss about various positional number systems:

14.1.1.1 Decimal System

The decimal system uses 10 as its base value and employs ten symbols 0 to 9 in representing numbers. Let us consider a decimal number 7402 consisting of four symbols 7,4,0,2. In terms of base 10 it can be expressed as follows:

$$7402 = 7 \times 10^3 + 4 \times 10^2 + 0 \times 10^1 + 2 \times 10^0$$

So each of the symbols from a set of symbols denoting a number is multiplied with power of the base (10) depending on its position counted from the right. The count always begins with 0.

In general a decimal number $d_m d_{m-1} \ldots d_1 d_0$ consisting of $(m+1)$ symbols can be expressed as:

$$d_m \times 10^m + d_{m-1} \times 10^{m-1} + \ldots . + d_1 \times 10^1 + d_0 \times 10^0 = \sum_{i=0}^{m} d_i 10^i$$

where $0 \leq d_i \leq 9$ with $i = 0, 1, \ldots, m$

Similarly, a fractional part of a decimal number can be expressed as $\sum_{i=1}^{m} d_i 10^{-i}$

14.1.1.2 Binary system

Binary system is the positional system consisting of two symbols i.e. $\{0,1\}$, named "bits", and '2' as its base. Any binary number $d_m d_{m-1} \ldots d_1 d_0$ actually represents a decimal value given by

$$d_m 2^m + d_{m-1} 2^{m-1} + \ldots + d_0 2^0 = \sum_{i=0}^{m} d_i 2^i$$

where $d_i = 0 \quad or \quad 1, \quad i = 0, 1, ..m.$

Consider the binary number 10101. The decimal equivalent of 10101 is given by

$$\begin{aligned}(10101)_2 &= 1 \times 2^4 + 0 \times 2^3 + 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 \\ &= 16 + 0 + 4 + 0 + 1 = (21)_{10}\end{aligned}$$

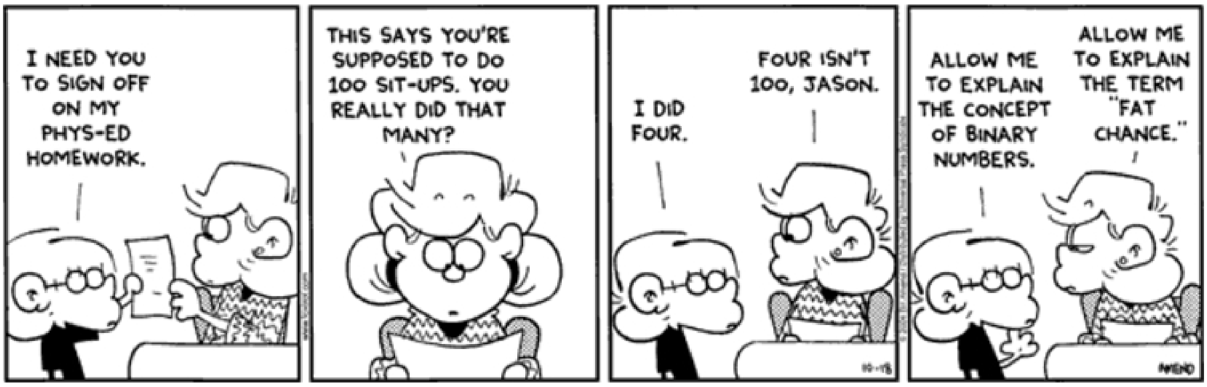
Here are the representation of some integer positive values in binary notation:

1	1	13	1101	25	11001
2	10	14	1110	26	11010
3	11	15	1111	27	11011
4	100	16	10000	28	11100
5	101	17	10001	29	11101
6	110	18	10010	30	11110
7	111	19	10011	31	11111
8	1000	20	10100	32	100000
9	1001	21	10101	33	100001
10	1010	22	10110	34	100010
11	1011	23	10111	35	100011
12	1100	24	11000	36	100100

Notice that:

- The number of different values representable in n bits is 2^n

- An n-bit binary number $X = x_{n-1}x_{n-2}\cdots x_1x_0$ can represent any integer value in the range $0 \leq X \leq 2^n - 1$ (e.g., if $n = 3$, then $0 \leq X \leq 7$). Note that $n+1$ bits are needed to represent the value $X = 2^n$
- The highest (left-most) bit x_{n-1} of an n-bit number is named the "most significant bit" (MSB) and the lowest bit (right-most) x_0 the "least significant bit" (LSB)


I NEED YOU TO SIGN OFF ON MY PHYS-ED HOMEWORK.
THIS SAYS YOU'RE SUPPOSED TO DO 100 SIT-UPS. YOU REALLY DID THAT MANY?
I DID FOUR.
FOUR ISN'T 100, JASON.
ALLOW ME TO EXPLAIN THE CONCEPT OF BINARY NUMBERS.
ALLOW ME TO EXPLAIN THE TERM "FAT CHANCE."

It should be obvious at this level for the reader that:

- 1000 has 10 bits because $512 \leq 1000 \leq 1024$, or $2^9 \leq 1000 \leq 2^{10} - 1$
- 1344 has 11 bits because $1024 \leq 1344 \leq 2047$, or $2^{10} \leq 1344 \leq 2^{11} - 1$
- 2527 has 12 bits because $2048 \leq 2527 \leq 4095$, or $2^{11} \leq 2527 \leq 2^{12} - 1$
- 5019 has 13 bits because $4096 \leq 5019 \leq 8191$, or $2^{12} \leq 5019 \leq 2^{13} - 1$
- 9999 has 14 bits because $8192 \leq 9999 \leq 16383$, or $2^{13} \leq 9999 \leq 2^{14} - 1$

Hence the minimum number of bit required b_{min} for a d digit integer is computed simply by using the specific number formula on the minimum d digit value:

$$b_{\text{min}} = \lfloor \log_2(10^{d-1}) \rfloor + 1 \qquad (14.1.1)$$

So for example, in 12018 (holocene calendar) the biggest prime number (important in cryptography) discovered had $23,249,425$ digits, thus corresponding using the above relation to $b_{\text{min}} = 25$.

What we have seen so far works only for integer (i.e. "fixed point numbers"). What about decimal numbers??? This is what we will focus on right now!

14.1.1.2.1 Floating Point (Real Numbers)

The encoding scheme for floating point numbers is more complicated than for fixed point. The basic idea is the same as used in scientific notation, where a mantissa is multiplied by ten raised to some exponent. For instance, $5.4321 \cdot 10^6$, where 5.4321 is the mantissa and 6 is the exponent. Scientific notation is exceptional at representing very large and very small numbers. For example: $1.2 \cdot 10^{50}$, the number of atoms in the Earth, or $2.6 \cdot 10^{-23}$, the distance a turtle crawls in one second, compared to the diameter of our galaxy. Notice that numbers represented in scientific notation are normalized so that there is only a single nonzero digit left of the decimal point. This is achieved by adjusting the exponent as needed.

Floating point representation is similar to scientific notation, except everything is carried out in base two, rather than base ten. While several similar formats are in use, the most common is ANSI/IEEE Std. 754-1985. This standard defines the format for 16 bits number named "half-precision", 32 bit numbers named "single precision", as well as 64 bit numbers named "double precision". As shown in the figure below, the 32 bits used in single precision are divided into three separate groups: bits 0 through 22 form the mantissa, bits 23 through 30 form the exponent, and bit 31 is the sign bit. These bits form the floating point number, v, by the following relation:

$$v = (-1)^S \cdot (1 + M) \cdot 2^{E-127} \tag{14.1.2}$$

More generally:

$$v = (-1)^S \cdot (2^0 + M) \cdot 2^{E-\text{bias}} \tag{14.1.3}$$

Also sometimes denoted:

$$\begin{aligned} \text{Real Value} &= (-1)^{\text{sign}} * (1.b_{22}b_{21}\ldots b_0)_2 \cdot 2^{(b_{30}b_{29}\ldots b_{23})_2 - 127} \\ &= \text{sign}(1\text{-bit}) \cdot \text{ binary fraction bits } (23\text{-bit}) * (8\text{-bit unsigned integer exponent}, e-127) \\ &= (-1)^{\text{sign}} \cdot \left(1 + \sum_{i=1}^{23} b_{23-i} 2^{-i}\right) \cdot 2^{(e-127)} \end{aligned} \tag{14.1.4}$$

So the main relation is:

$$\boxed{v = (-1)^S \cdot (2^0 + M) \cdot 2^{E-\text{bias}} = (-1)^{\text{sign}} \cdot \left(1 + \sum_{i=1}^{23} b_{23-i} 2^{-i}\right) \cdot 2^{(e-127)}} \tag{14.1.5}$$

In other words, $M = 1 + b_{22}2^{-1} + b_{21}2^{-2} + b_{20}2^{-3} \ldots$. If bits 0 through 22 are all zeros, M takes on the value of one. If bits 0 through 22 are all ones, M is just a hair under two, i.e., $2 - 2^{-23}$.

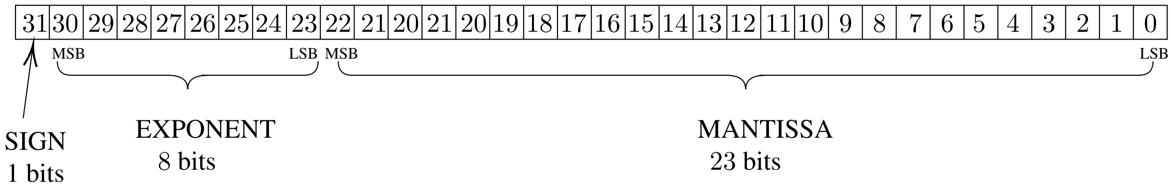


Figure 14.1: **Floating point (real numbers) in simple precision**

Or for the half-precision system:

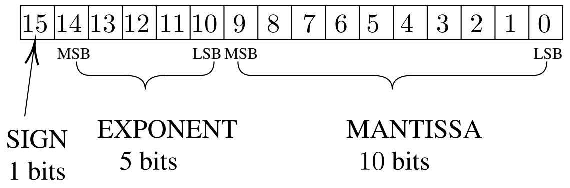


Figure 14.2: **Floating point (real numbers) in half precision**

The term: $(-1)^S$, simply means that the sign bit, S, is 0 for a positive number and 1 for a negative number. The variable, E, is the number between 0 and 255 represented by the eight exponent bits. Subtracting 127 from this number allows the exponent term to run from 10^{-127} to 10^{128} other words, the exponent is stored in offset binary with an offset of 127.

The mantissa, M, is formed from the 23 bits as a binary fraction. For example, the decimal fraction: 2.783 , is interpreted: $2+7/10+ 8/100+3/1000$. The binary fraction: 1.0101 , means: $1+0/2+1/4+0/8+1/16$. Floating point numbers are normalized in the same way as scientific notation, that is, there is only one nonzero digit left of the decimal point (named a" binary point" in base 2). Since the only nonzero number that exists in base two is 1, the leading digit in the mantissa will always be a 1, and therefore does not need to be stored. Removing this redundancy allows the number to have an additional one bit of precision. The 23 stored bits, referred to by the notation: $b_{22}, b_{21}, b_{21}, \ldots, b_0$, form the mantissa according to:

$$M = 1.b_{22}b_{21}b_{20}b_{19}\ldots b_2b_1b_0 \qquad (14.1.6)$$

Using this encoding scheme, the largest number that can be represented is: $\pm\,(2-2^{-23})\cdot 2^{128} = \pm 6.8\cdot 10^{38}$. Likewise, the smallest number that can be represented is: $\pm 1.0\cdot 2^{-127} = \pm 5.9\cdot 10^{-39}$. The IEEE standard reduces this range slightly to free bit patterns that are assigned special meanings. In particular, the largest and smallest numbers allowed in the standard are $\pm 3.4\cdot 10^{38}$ and $\pm 1.2\cdot 10^{-38}$ respectively. The freed bit patterns allow three special classes of numbers:

- ± 0 is defined as all of the mantissa and exponent bits being zero.
- $\pm\infty$ is defined as all of the mantissa bits being zero, and all of the exponent bits being one.
- A group of very small unnormalized numbers between $\pm 1.2\cdot 10^{-38}$ and $\pm 1.4\cdot 10^{-45}$.

These are lower precision numbers obtained by removing the requirement that the leading digit in the mantissa be a one. Besides these three special classes, there are bit patterns that are not assigned a meaning, commonly referred to as NANs (Not A Number).

The IEEE standard for double precision simply adds more bits to the single precision format. Of the 64 bits used to store a double precision number, bits 0 through 51 are the mantissa, bits 52 through 62 are the exponent, and bit 63 is the sign bit. As before, the mantissa is between one and just under two, i.e., $M = 1 + b_{51}2^{-1} + b_{50}2^{-2} + b_{49}2^{-3}\ldots$. The 11 exponent bits form a number between 0 and 2047 , with an offset of 1023, allowing exponents from

2^{-1023} to 2^{1024}. The largest and smallest numbers allowed are $\pm 1.8 \cdot 10^{308}$ and $\pm 2.2 \cdot 10^{-308}$, respectively. These are incredibly large and small numbers! It is quite uncommon to find an application where single precision is not adequate. You will probably never find a case where double precision limits what you want to accomplish.

Caution! Let's have a look at a number which is fine for us, but has the floating point issue for the personal computer: $1/10$. For us Homo Sapiens, $1/10$ is 0.1, it's easy, only one decimal! For a digital computer, $1/10$ in floating precision is $0.0001100110011001101\ldots$ These decimals have a recurring combination of numbers, i.e. 0011. The computer will not calculate every single decimals as it is infinite! It has to stop at some point and round the number of decimals. This rounding is the floating point lack of precision. For comparison, in base 10, we can't precisely define $1/3$, which is 0.3333 recurring 3. Overall, many spreadsheet and business intelligence softwares decimal issue is caused by a limitation of floating point precision. This is not an error per se, but can cause one. Understanding the logic of decimals-rounding will help you find the solution applicable to your situation.

14.1.1.2.2 Binary arithmetic (for fixed point/integers)

Here are the addition facts that you need when adding numbers in binary notation:

$$\begin{aligned} 0+0&=0\\ 0+1&=1\\ 1+1&=10 \end{aligned} \tag{14.1.7}$$

Doing binary addition is like doing decimal addition excepted that we work on a binary set.

$$\begin{array}{r} {\scriptstyle 1\;1111\;111} \\ 0\,1100\,1001 \\ +\,0\,1111\,1111 \\ \hline 1\,1100\,1000 \end{array}$$

Here are the subtractions facts that you need when subtracting numbers in binary notation:

$$\begin{aligned} 1-1&=0\\ 1-0&=0\\ 10-1&=1 \end{aligned} \tag{14.1.8}$$

The fact that $10-1=1$ will lead to borrowing since you can not take 1 from 0 for a certain place value, and so you must borrow from the 1 in the next place value, much like how you need to borrow in decimal notation:

$$\begin{array}{r} {\scriptstyle 111} \\ 1001 \\ -\,0110 \\ \hline 0011 \end{array}$$

We will come back more in details on binary arithmetic during our study of Boolean Algebra in the section of Logical Systems.

14.1.1.3 Hexadecimal System

The Hexadecimal system is the positional system consisting of sixteen symbols, 0,1,2,...,9,A,B,C,D,E,F, and '16' as its base. Here the symbols A denotes 10, B denotes 11 and so on. The decimal equivalent of the given hexadecimal number $d_m d_{m-1} \ldots d_0$ is given by:

$$\sum_{i=0}^{m} d_i \cdot (16)^i \tag{14.1.9}$$

For example consider $(15ACB)_{16}$:

$$\begin{aligned}(15ACB)_{16} &= 1 \cdot 16^4 + 5 \cdot 16^3 + 10 \cdot 16^2 + 12 \cdot 16^1 + 11 \cdot 16^0 \\ &= 65536 + 5 \cdot 4096 + 10 \cdot 256 + 12 \cdot 16 + 11 \cdot 1 \\ &= 65536 + 20480 + 2560 + 192 + 11 = (88779)_{10}\end{aligned} \tag{14.1.10}$$

We can convert a binary number directly to a hexadecimal number by grouping the binary digits, starting from the right, into sets of four and converting each group to its equivalent hexadecimal digit. If in such a grouping the last set falls short of four binary digits then do the obvious thing of prefixing it with adequate number of binary digit '0'. For example let us find the hexadecimal equivalent of $(111\,011\,0101\,0010\,1110)_2$:

$$(\underline{0011} \quad \underline{1011} \quad \underline{0101} \quad \underline{0010} \quad \underline{1110})_2 = (3B52E)_{16} \tag{14.1.11}$$

The vice versa is also true.

14.1.1.4 Octal System

The octal system is the positional system that uses 8 as its base and $\{0, 1, \ldots, 7\}$ as its symbol set of size 8. The decimal equivalent of an octal number $(d_m d_{m-1} \ldots d_0)_8$ is given by:

$$\left(\sum_{i=0}^{m} d_i 8^i\right)_{10} \tag{14.1.12}$$

For example consider $(6741)_8$:

$$\begin{aligned}(6741)_8 &= 6 \cdot 8^3 + 7 \cdot 8^2 + 4 \cdot 8^1 + 1 \cdot 8^0 \\ &= 6 \cdot 512 + 7 \cdot 64 + 4 \cdot 8 + 1 \\ &= 3072 + 448 + 32 + 1 = (3553)_{10}\end{aligned} \tag{14.1.13}$$

We can get the octal equivalent of a binary number by grouping the binary digits, starting from the right, into sets of three binary digits and converting each of these sets to its octal equivalent.

If such a grouping results in a last set having less number of digits it may be prefixed with adequate number of binary digit 0.

As an example the octal equivalent of $(1010\,110\,111\,001)_2$.

So we have:

$$\begin{aligned}(1010\,110\,111\,001)_2 &= (001 \quad 010 \quad 110 \quad 111 \quad 001)_2 \\ &= (12671)_8\end{aligned} \tag{14.1.14}$$

14.1.1.5 Conversion of decimal system to non-decimal system:

To convert a decimal number to a number of any other system we should consider the integer and fractional parts separately and follow the following procedure:

Conversion of integer part:

1. Consider the integer part of a given decimal number and divide it by the base b of the new number system. The remainder will constitute the rightmost digit of the integer part of the new number.
2. Next divide the quotient again by the base b. The remainder will constitute second digit from the right in the new system

Continue this process until we end up with a zero-quotient. The last remainder is the leftmost digit of the new number.

Conversion of fractional part:

1. Consider the fractional part of the given decimal number and multiply it with the base b of the new system. The integral part of the product constitutes the leftmost digit of the fractional part in the new system.
2. Now again multiply the fractional part resulting in step (a) by the base b of the new system. The integral part of the resultant product is the second digit from the left in the new system.

Repeat the above step until we encounter a zero-fractional part or a duplicate fractional part. The integer part of this last product will be the rightmost digit of the fractional part of the new number.

✍**Example:**

We want to convert 54.45 into its binary equivalent.

1. Consider the integer part i.e. 54 and apply the steps listed under conversion of integer part i.e.
2. conversion of fractional part:

		Product	integral part	Binary number
0.45×2	=	0.90	0	$(.01\overline{1100})_2$
0.9×2	=	1.80	1	
0.8×2	=	1.6	1	
0.6×2	=	1.2	1	
0.2×2	=	0.4	0	
0.4×2	=	0.8	0	
0.8×2	=	1.6	1	
0.6×2	=	1.2	1	
0.2×2	=	0.4	0	
0.4×2	=	0.8	0	
0.8×2	=	1.6	1	
...		...	...	
...		...	...	

Therefore:

$$(0.45)_{10} = (.01\overline{1100})_2 \tag{14.1.15}$$

and finally:

$$(54.45)_{10} = (110110.01\overline{1100})_2 \tag{14.1.16}$$

Here the overbar denotes the repetition of the binary digits.

Using binary system as an intermediate stage we can easily convert octal numbers to hexadecimal numbers and vice-versa:

✍**Examples:**

E1. $(423)_8 = (100 \quad 010 \quad 011)_2$
$\quad = (0001 \quad 0001 \quad 0011)_2 = (113)_{16}$

E2. $(93AF)_{16} = (1001 \quad 0011 \quad 1010 \quad 1111)_2$
$\quad = (010 \quad 010 \quad 011 \quad 101 \quad 011 \quad 110)_2$
$\quad = (223536)_8$

In the above two examples we have grouped the binary digits suitably either to quadruplets or triplets to convert octal to hexadecimal and hexadecimal to octal numbers respectively.

14.1.2 Algorithm Complexity

Definition (#14.1.553): The "complexity" of an algorithm is the measure of the number of fundamental operations it performs in the worst case on a dataset.

Measure the exact complexity is most of time irrelevant because often too complex given the size of the programs (too big algorithms). To avoid calculating in detail the complexity of an algorithm, we identify the fundamental operations. These basic operations can be: an assignment, a comparison between two variables, an arithmetic operation between two variables, etc.

Thus, the classical used hypothesis for the calculation of the complexity are:

H1. The four fundamental operations have the same time cost: $+\equiv-\equiv\times\equiv\div$

H2. A memory access has a cheaper time cost than an arithmetic operation.

H3. A comparison check has a cheaper time cost than an arithmetic operation.

H4. We work with only one single processor.

Definitions (#14.1.554):

D1. We note D_n the sets of data of size n and $T(n)$ the cost (in time) of the algorithm on the data or the data set of size n.

D2. The "complexity at the best" is given by the function:

$$T_{\text{min}}(n)=\text{min}_{d\in D_n}C(d) \tag{14.1.17}$$

This is the smallest time that will need an algorithm to run on a data set (lines of code) of fixed size, here equal to n, that the cost (duration) execution is $C(d)$. This is a lower bound on the complexity of the algorithm on a data set of size n.

D3. The "complexity at worst" (the most interesting for the practitioner because it is the one to minimize!):

$$T_{\text{max}}(n)=\text{max}_{d\in D_n}C(d) \tag{14.1.18}$$

This is the biggest time that will need an algorithm to run on a data set (lines of code) of fixed size, here equal to n, that the cost (duration) execution is $C(d)$. This is a upper bound on the complexity of the algorithm on a data set of size n. The algorithm will always finish at this time or before but never after.

D4. The "average complexity":

$$T_{\text{avg}}(n)=\frac{\sum\limits_{d\in D_n}C(d)}{D_n} \tag{14.1.19}$$

This is the average of the complexities of the algorithm on the data sets of size n (strictly speaking, we must obviously take into account the probability of occurrence of each of the data sets). This average reflects the general behaviour of the algorithm if extreme cases are rare or complexity varies slightly depending on the data. However, the complexity in practice on a particular data set may be significantly greater than the average complexity; in this case the average complexity does not give a good indication of the behaviour of the algorithm.

D5. The "order of complexity" is defined as the complexity at worst of an algorithms that contains several terms (additions or subtractions) so that we only keep the term that is growing the fastest. Thus, an algorithm having a complexity of type:

$$An^2+Bn+n! \tag{14.1.20}$$

will be said to have a complexity of order $\mathcal{O}(n!)$. This is the "big $\mathcal{O}$ notation". It is a member of a family of notations invented by Paul Bachmann, Edmund Landau, and others, collectively named "Bachmann–Landau notation" or "asymptotic notation".

Remark

Just as worst cases are given in terms of upper bounds, best cases are given in terms of lower bounds, which are specified with "big Ω notation". A best case of $\Omega(n)$ would mean that the function is always at least linear, but could grow faster.

✍**Examples:**

Consider N as the size of the data. For a decimal number this is the numbers N of digits. Consider two decimal numbers like $0.a_1a_2a_3\ldots$ and $0.b_1b_2b_3\ldots$

E1. Addition has linear complexity because in developed way addition is expressed as (see section Numbers page 211):

$$\sum_{1\leq n\leq N} a_n10^{-n}+\sum_{1\leq n\leq N} b_n10^{-n}=\sum_{1\leq n\leq N}(a_n+b_n)10^{-n} \tag{14.1.21}$$

E2. Multiplication has quadratic complexity because in developed way multiplication is expressed as:

$$\sum_{1\leq k\leq N} a_k10^{-k}+\sum_{1\leq l\leq N} b_l10^{-l}=\sum_{1\leq n\leq N}\left(\sum_{k+l=n} a_kb_l\right)10^{-n} \tag{14.1.22}$$

Definition (#14.1.555): An algorithm is said to be an "optimal algorithm" if its complexity has the minimal complexity among all other algorithms in its class.

As we made it understand implicitly earlier, we focus almost exclusively on the time complexity of the algorithms. Sometimes it is interesting to focus on other of their characteristics, such as space complexity (size of the memory used), the bandwidth required, etc.

For the result of the analysis of an algorithm to be relevant, we must have a model of a machine on which the algorithm will be implemented (as a program). We usually take as a reference, the "random access machine (RAM)" with a single processor where instructions are executed one after the other, without concurrent operations and without stochastic processes (in contrast to possible future quantum computers).

Most common algorithms can be classified into a number of broad classes of complexity whose order $\mathcal{O}$ varies somehow:

1. The algorithms with "constant complexity $\mathcal{O}(1)$" that just do a boolean control (comparison).
2. The algorithms with "linear complexity $\mathcal{O}(n)$" and those in complexity in $\mathcal{O}(n\log(n))$ that are considered as fast.
3. The algorithms with "polynomial complexity in $\mathcal{O}(n^k)$" (for $k>3$ are considered as slow).
4. The algorithms with "sub-linear complexity" whose complexity is usually of the order $\mathcal{O}(\log(n))$
5. The algorithms with "factorial complexity" whose complexity is usually of the order $\mathcal{O}(n!)$

and so on... (we give here definition only for complexity for which we already have given a detailed example or will give examples further below).

It is important to see that (focusing only on complexities we have seen with examples or will see later below) for $n>4$:

$$\boxed{\mathcal{O}(1)<\mathcal{O}(\log(n))<\mathcal{O}(n)<\mathcal{O}(n\ln(n))<\mathcal{O}(n^2)<\mathcal{O}(e^n)<\mathcal{O}(n!)<\mathcal{O}(n^n)} \quad (14.1.23)$$

Remark

In practice a good complexity is considered as being of order $\mathcal{O}(n^k)$ for $k>3$. Poor complexity is considered as being of type $\mathcal{O}(e^n),\mathcal{O}(n!)$ or equation.

For people that like better visual stuff:

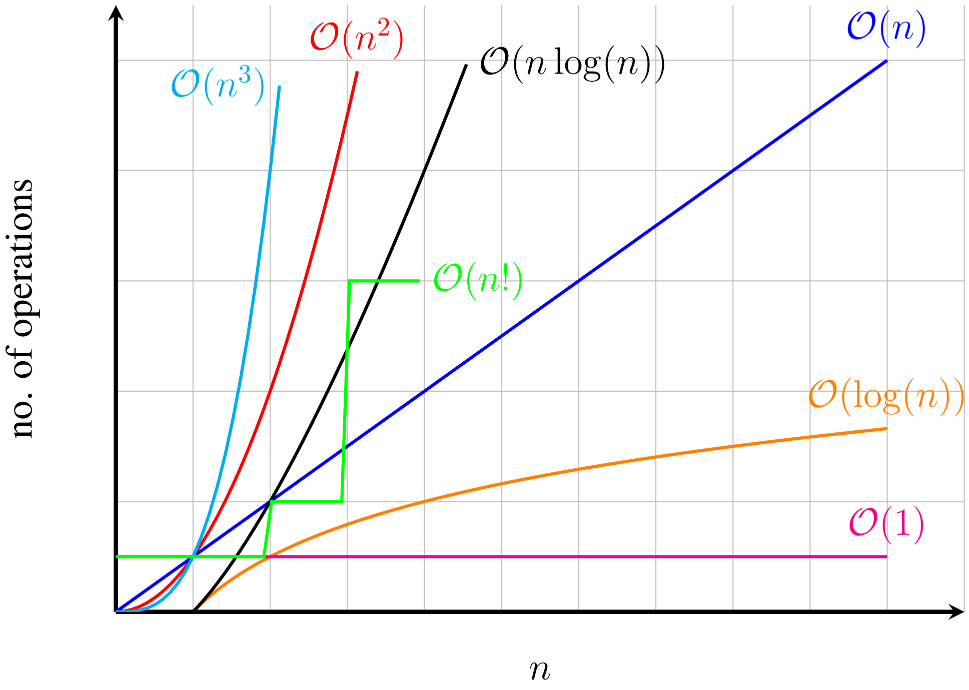


Figure 14.3: Big $\mathcal{O}$ notation complexities plot

n	1	$\log(n)$	$n\cdot\log(n)$	n^2	n^3	$n!$
1	1	0	0	1	1	1
2	1	0.30	0.60	4	8	2
3	1	0.48	1.43	9	27	6
4	1	0.60	2.41	16	64	24
5	1	0.70	3.49	25	125	120
6	1	0.78	4.67	36	216	720
7	1	0.85	5.92	49	343	5,040
8	1	0.90	8.13	64	512	40,320
9	1	0.95	8.59	81	729	362,880
10	1	1.00	10.00	100	1,000	3,628,800
20	1	1.30	26.02	400	8,000	$2.42\cdot 10^{18}$

We already saw above an example with linear complexity (addition of scalars) and of polynomial (quadratic) complexity (with multiplication of scalars).

Let us see some other very common examples. The first example use the results of the previous examples:

✍**Examples:**

E1. Evaluation of a polynomial:

$$P(x)=a_0+a_1x+a_2x^2+\ldots+a_nx^n \quad (14.1.24)$$

The direct estimate of the value of $P(x)$ leads to a complexity:

$$C_{P(x)}=\mathcal{O}\left((n)_{\text{additions}}+\left(1+2+3+\ldots+(n+1)=\frac{n(n+1)}{2}\right)_{\text{products}}\right)=\mathcal{O}(n^2) \tag{14.1.25}$$

E2. Another important example is the inversion matrix problem that is used a lot in supercomputer megaflops benchmark (sometimes named "LINPACK test"). As you can see in the MATLAB® companions book there is a division by n^3 as the matrix inversion complexity is of order n^3. Indeed:

At the beginning, when the first row has length n, it takes n operations to zero out any entry in the first column (one division, and $n-1$ multiply-subtracts to find the new entries along the row containing that entry. To get the first column of zeroes therefore takes $n(n-1)$ operations.

In the next column, we need $(n-1)(n-2)$ operations to get the second column zeroed out.

In the third column, we need $(n-2)(n-3)$ operations.

The sum of all of these operations is:

$$\sum_{i=1}^{n} i(i-1)=\sum_{i=1}^{n} i^2-\sum_{i=1}^{n} i=\frac{n(n+1)(2n+1)}{6}-\frac{n(n+1)}{2} \tag{14.1.26}$$

which goes as $\mathcal{O}(n^3)$.

Thanks to William George Horner we have a more efficient algorithm that uses a factorization of the polynomial in the form:

$$P(x)=a_0+x(a_1+x(a_2+\ldots x(a_{n-1}+xa_n))) \tag{14.1.27}$$

We can see almost easily that this factorization holds the same number of additions to (n) but reduces the number of multiplications to (n).

The resulting complexity is $\mathcal{O}(n)$. The gain is unquestionably important. In addition, this factorization avoids computing power.

Let us see now another well known example that almost all students that have learned scripting a little bit already know:

✍**Example:**

The famous example of algorithmic complexity is the search for an information in a sorted column (list). A simple algorithm named "dichotomous search", also named "binary search", is to take the cell to mid-column and see if we find the desired value. Otherwise, research must continue on the same method in the top or bottom of the column (depending on the lexicographical order).

The algorithm is recursive and allows at each step, to divide by two the size of the search space.

If this size is initially of n cells in the column, it is of size $\dfrac{n}{2}$ in step 1, of size $\dfrac{n}{2^2}$ in step 2, and more generally of size $\dfrac{n}{2^k}$ in step k.

At worst, the search ends when there is only one cell in the column to explore, i.e. when k is such that $n < 2^k$.

We deduce the maximum number of steps: it is the smallest k such that $n < 2^k$, written in another way $\log_2(n) < k$, i.e. the sub-linear complexity:

$$T_{\text{max}} = k = \log_2(n) \tag{14.1.28}$$

That result is strongly related to a famous guess game. The game is that if you have a set of integers $\{1, 2, \ldots, n\}$. Host picks at random one of the numbers and gives you k chances to guess it. After each guess the host says "higher", "lower", or "correct". The question is what is the optimal strategy (number of guesses)?

If we use binary search algorithm with guesses lower or higher on every turn, the maximum number of guesses that is required to guess a number between $1 - n$ is $\lceil \log_2(n) \rceil$ and this must be less than k to guarantee that we guess the number within k guesses. So:

$$\lceil \log_2(n) \rceil \leq k \tag{14.1.29}$$

Therefore:

$$\lfloor \log_2(n) \rfloor + 1 \leq k \tag{14.1.30}$$

Hence:

$$n \leq 2^{k-1} \tag{14.1.31}$$

So we have just proved that if $n \leq 2^{k-1}$ then there is a strategy that guarantees we will guess the number by the kth guess.

The result of this last example is to compared with a sequential search (useful when sorting is too costly in resources). For example, in a column of $25,000$ data the complexity is $\mathcal{O}(n)$ thus $25,000$ while with the dichotomous method, the sub-linear complexity gives $\log_2(25,000) = 15$. The gain is considerable (at the condition that data is sorted)!

Other complexity that the reader must absolutely know are some elementary Linear Algebra calculations (see the section of the same name)! So without proof (because normally trivial given the previous examples) if we consider two square matrices A and B of dimensions n the main operations have the following complexities:

- Read the components (nested loops): $\mathcal{O}(n^2)$

- Calculation of the trace: $\text{tr}(A)=\sum_{i=1}^{n} a_{ii} \rightarrow \mathcal{O}(n)$
- Addition $A+B=C$ so that $c_{ij}=a_{ij}+b_{ij} \rightarrow \mathcal{O}(n^2)$
- Product $A \cdot B=c$ so that $c_{ij}=\sum_{k=1}^{n} a_{ik}b_{kj} \rightarrow \mathcal{O}(n^3)$
- Solving the travelling salesman problem via brute-force search is $\mathcal{O}(n!)$.
- Determinant (by the direct method of Cramer as detailed in the section of Linear Algebra page 1977). We can thus show that the complexity order of the determinant of a square matrix of dimensions n is n products, $n-1$ additions plus n times the complexity of the determinant of a matrix of dimensions $n-1$ so that finally we have: det(A) $=\mathcal{O}(n \cdot n!)$

Assuming that the computer performs an elementary operation in equation seconds (which is already a good computer), we obtain the following calculations time for several values of n for the determinant:

n	5	10	15	20	50
t	$6 \cdot 10^{-7}$ [s]	0.04 [s]	5.5 [h.]	$1,543$ [y.]	$4.8 \cdot 10^{46}$ [y.]

Table 14.1: **Various costs**

hence the need to make sometimes a time complexity calculation before starting an algorithm (unless you are working exclusively for future generations, provided that there will still be future generations...).

To close this subject let us indicate that there are other notations than just the big $\mathcal{O}$ and the big Ω. So let us give a summary and at the same time the definition for the other one:

- $f(x)=\mathcal{O}(g(x))$ means that the growth rate of $f(x)$ is asymptotically less than or equal to the growth rate of $g(x)$
- $f(x)=\Omega(g(x))$ means that the growth rate of $f(x)$ is asymptotically greater than or equal to the growth rate of $g(x)$
- $f(x)=\wr(g(x))$ means that the growth rate of $f(x)$ is asymptotically less than the growth rate of $g(x)$.
- $f(x)=\omega(g(x))$ means that the growth rate of $f(x)$ is asymptotically greater than the growth rate of $g(x)$
- $f(x)=\Theta(g(x))$ means that the growth rate of $f(x)$ is asymptotically equal to the growth rate of $g(x)$

Remark

Here's a simple way to remember which notation means what. All of the big $\mathcal{O}$ notations can be considered to have a bar. When looking at a Ω, the bar is at the bottom, so it is an (asymptotic) lower bound. When looking at a Θ, the bar is obviously in the middle. So it is an (asymptotic) tight bound. When handwriting $\mathcal{O}$, you usually finish at the top, and draw a squiggle. Therefore $\mathcal{O}(g(x))$ is the upper bound of the function. To be fair, this one doesn't work with most fonts, but it is the original justification of the names.

14.1.2.1 NP-Completude

We will introduce now for the general culture the concept of "NP-completeness", that is to say that we will try to define without too much formalism (as usual in this textbook).

First let us introduce two non-formal definitions of a Turing machine.

Definitions (#14.1.556):

D1. An algorithm is said to be a "deterministic Turing machine" when the set of its rules prescribes at most one action to be performed for any given situation.

D2. A "non-deterministic Turing machine" may have a set of rules that prescribes more than one action for a given situation. For example, a non-deterministic Turing machine may have both "If you are in state 2 and you see an 'A', change it to a 'B' and move left" and "If you are in state 2 and you see an 'A', change it to a 'C' and move right" in its rule set.

Now let us define what type of problems are frequently considered:

Definitions (#14.1.557):

D1. "Logarithmic problems L" contains all algorithms (problems) that can be solved by a deterministic Turing machine using a logarithmic amount of computation time.

D2. "Polynomial problems P" contains all algorithms (problems) that can be solved by a deterministic Turing machine using a polynomial amount of computation time.

D3. "Non-deterministic polynomial time problems NP" are the set algorithms (problems) where the result instances can be controlled in a polynomial time complexity by a non-deterministic Turing machine.

It should be noted at this stage of the discussion that the class P is included in the NP class so that P $\subset$ NP. Indeed if we know a polynomial algorithm to solve a problem then we can at worst check the solution with a polynomial complexity algorithm also.

But a difficult question is the following reciprocal: if a problem is NP (known solution can be controlled in polynomial time) but it seems we can not found yet an algorithm in a polynomial time (or less) to find the solution, does it always exist a polynomial algorithm to find the solution so that $P = NP$ that thus bring us to write P $\subseteq$ NP?

In other words: If the solution of a problem can be quickly verified as being the correct one, can the solution also be found quickly?

✍**Examples:**

E1. The problem of finding a Hamiltonian cycle (cycle that passes once and only once by all the vertices of the graph - see section Graphs Theory page 3019) in a graph belongs to NP since, given a known cycle it is trivial to check in linear time $\mathcal{O}(n)$ it contains a well and once each vertex but found the cycle is at worst a factorial complexity $\mathcal{O}(n!)$.

E2. Factoring an integer n in product of prime factors (important in cryptography) is a NP problem. Indeed, given the prime factors $p_1, p_2, \ldots, p_n$ it is trivial to control the solution $n = p_1, p_2, \ldots, p_n$ as it is of order $\mathcal{O}(n^2)$. But we do not know if it exist a polynomial algorithm to find the prime numbers (to find the solution). So we do not know if the problem of finding prime numbers of an integer is a P-problem.

Remark

Apparently (we were not able to find the proof of this result and neither had opportunity to do so) the complexity of the best factorization algorithm for prime numbers is in the year 12007 (holocene calendar) of the type:

$$\mathcal{O}\left(\exp\left[\left(\frac{64}{9}\log(n)\right)^{\frac{1}{3}}\right]\log(\log(n))^{\frac{2}{3}}\right) \tag{14.1.32}$$

therefore there is still work to do (if a reader could provide us with the details that led to this result, we are interested).

A problem x that is in NP is also in NP-Complete if and only if every other problem in NP can be quickly (i.e. in polynomial time) transformed into x.

In other words a problem x is "NP-Complete NPC" if:

1. x is in NP
2. Every problem in NP is reducible to x

So, what makes NP-Complete so interesting is that if any one of the NP-Complete problems was to be solved quickly, then all NP problems can be solved quickly.

A NPC problem is complete in that it contains most of the complexity of problems belonging to NP, and a polynomial solution to this problem involves a polynomial solution to all NP type problems.

In other words: NPC problems have an exponential complexity and they all have the same complexity class (modulo polynomials).

Finally, what is important to understand and retain about this idea is that if we find one day a polynomial algorithm for one of these really difficult problems that are the NPC problems, then in one stroke NP becomes equal to P and all difficult problems become easy such that:

$$\boxed{\mathrm{N}=\mathrm{NP}} \tag{14.1.33}$$

In other words: can we find in a polynomial time what can be proved (controlled) in polynomial time?

Remark

This issue is so important in computing science to reduce energy consumptions and time computations that it belongs (arbitrarily) to the 7 millennium problems, whose resolution is prized $1 million by the Clay Mathematics Institute.

Ok this definitions given let us now study of some typical applications of numerical methods which are very often used in the industry. We will go from simple to more complicated and not forgetting that many methods that are not in this section can sometimes be found in other sections of the book!

14.1.3 Integer Part

The biggest integer less than or equal to a real number x is expressed in this textbook by $[x]$, which will be read "integer part of x" and according to the norm ISO 80000-2:2009 *Mathematical signs and symbols to be used in the natural sciences and technology* it should denoted as: int x.

The number M is an integer if and only if $[M]=M$. Similarly, the natural number $A\in\mathbb{N}$ is divisible in the natural set by the natural number $B\in\mathbb{N}$ if and only if:

$$\left[\frac{A}{B}\right]=\frac{A}{B} \tag{14.1.34}$$

We also denote by $\{x\}$ the fractional part of x such that:

$$\{x\}=x-[x] \tag{14.1.35}$$

That is to say:

$$x=[x]+\{x\} \tag{14.1.36}$$

with $|\{x\}|<1$.

Consider $x,y\in\mathbb{R}$. Then we have the following properties (normally most of them don't need any proof):

P1. $[x]\leq x<[x]+1\Leftrightarrow 0\leq x-[x]<1$

P2. $[x]=\sum\limits_{n\leq x} 1$ for $x \geq 0$

P3. $[x+m]=[x]+m$ if $m \in \mathbb{Z}$

P4. $[x]+[y] \leq [x+y] \leq [x]+[y]+1$

Proof 14.1.2. A reader asked for the proof of this property. So let us see how to proceed. First we write:

$$x=n+\theta, y=m+\phi \tag{14.1.37}$$

where $n,m \in \mathbb{Z}$ and where $0 \leq \theta < 1$ and $0 \leq \phi < 1$. Thus:

$$\begin{aligned}[x]+[y]=n+m=[n+m]\leq [n+\theta+m+\phi]=[x+y]\\ =n+m+[\theta+\phi]\leq n+m+1=[x]+[y]+1\end{aligned} \tag{14.1.38}$$

By writing $x=n+\theta$, where $0 \leq 0 < 1$, we have:

$$-x=-n-1+1-\theta \tag{14.1.39}$$

where $0 < 1-\theta \leq 1$. It follows that:

$$[x]+[-x]=n+[-n-1+1-\theta]=n-n-1+[1-\theta]=0 \tag{14.1.40}$$

if $\theta=0$ and:

$$[x]+[-x]=n+[-n-1+1-\theta]=n-n-1+[1-\theta]=-1 \tag{14.1.41}$$

if $0<\theta<1$.

■ Q.E.D.

P5. $[-x]=-[x]$ if $x \in \mathbb{Z}$ otherwise $[-x]=-[x]-1$ if $x \notin \mathbb{Z}$

Proof 14.1.3. The proof is already given at the end of the proof of property P4.

■ Q.E.D.

P6. $\left[\dfrac{[x]}{m}\right]=\left[\dfrac{x}{m}\right]$ if $m \in \mathbb{N}$

Proof 14.1.4. For this proof we will write:

$$x=[x]+n \tag{14.1.42}$$

where $0<\theta<1$ and:

$$[x]=qm+r \tag{14.1.43}$$

where $0 \leq r < m$ (see section Number Theory page 338). So we get:

$$\left[\dfrac{x}{m}\right]=\left[\dfrac{qm+r+\theta}{m}\right]=q+\left[\dfrac{r+\theta}{m}\right]=q \tag{14.1.44}$$

because $0 \leq r+\theta \leq m$. Besides :

$$\left[\dfrac{x}{m}\right]=\left[\dfrac{n}{m}\right]=\left[q+\dfrac{r}{m}\right]=q \tag{14.1.45}$$

and thus we have the expected result.

■ Q.E.D.

P7. If $a \in \mathbb{N}$ then $\left\lfloor \dfrac{x}{a} \right\rfloor$ represents the number of integers less than or equal x that are divisible by a.

Proof 14.1.5. For the last part, we observe that if $a, 2a, \ldots, ma$ are all positive integers $\leq x$ that are divisible by a, it suffices to prove that $\left\lfloor \dfrac{x}{a} \right\rfloor = m$.

Since $(m+1)a > x$, then:

$$ma \leq x \leq (m+1)a \tag{14.1.46}$$

That is to say:

$$m \leq \frac{x}{m} \leq m+1 \tag{14.1.47}$$

and thus we have the expected result.

■ Q.E.D.

Remark

The rounding method of real values will be given in the section on Economy.

14.1.4 Heron's Square Root Algorithm

One of the first things that many people learn in an introduction to Computer Sciences course is the algorithm for calculating the square root of a number.

There is such a simple algorithm for this purpose named "Heron's algorithm" or "algorithm of Heron of Alexandria" or even "Babylonian method" which converge to the value of this square root.

Thus want to calculate the square root:

$$x = \sqrt{A} \tag{14.1.48}$$

Proof 14.1.6. Here is a pseudo-proof because historically the algorithm was built on purely intuitive considerations (since approximately year 9900 according to holocene calendar algebra did not exist...). In high-school classes the result is given as a definition and just convergence is observed.

Remark

The "convergence ratio" of a sequence $x^{(k)}$ to a constant x_0 is:

$$\lim_{k \to \infty} \frac{\Delta\left(x^{(k+1)}, x_0\right)}{\Delta\left(x^{(k)}, x_0\right)} \tag{14.1.49}$$

if this limit exists. If the convergence ratio is greater than 0 and less than 1, the sequence is said to "converge linearly". If the convergence ratio is 0, the sequence is said to "converge superlinearly".

So for the demonstration, we will proceed as follows:

$$\begin{aligned}
&x^2 = A\\
&\Rightarrow 2x^2 = x^2 + A\\
&\Rightarrow 2x = \frac{x^2+A}{x}\\
&\Rightarrow x = \frac{x^2+A}{2x} = \frac{x+\dfrac{A}{x}}{2}
\end{aligned} \tag{14.1.50}$$

And the trick is to write:

$$\boxed{x_{i+1} = \frac{x_i}{2} + \frac{A}{2x_i}} = \frac{1}{2}\left(x_i + \frac{A}{x_i}\right) \tag{14.1.51}$$

This is named a "fixed-point iteration" (related to the theorem of the same name proved in the section of Sequences and Series page 1780). It is used in mainly situations in computer science. We will encounter it quite a lot of time further below...!

■ Q.E.D.

Example:

We want to calculate:

$$\sqrt{10} = 3.162277660168\ldots \tag{14.1.52}$$

We take $A = 10$ and this gives us the following table of iterations:

Iteration	$\dfrac{x_i}{2}$	$\dfrac{A}{2x_i}$	x_{i+1}	**Gap**
1	5	0.5	5.50	$\cong 2$
2	2.750	0.90909	3.659090909	$\cong 0.49$
3	1.82954	1.3664	3.196005083	$\cong 0.033$
4	1.59800	1.5644	3.162455624	$\cong 0.0002$
5	1.58122	1.5810	3.162277665	$\cong 0.5\cdot 10^{-8}$

Table 14.2: **Iterations for Heron's square root algorithm**

In the case of the cubic root, the proof is similar and we obtain:

$$x_{i+1} = \frac{2x_i}{3} + \frac{A}{3x_i^2} \tag{14.1.53}$$

and so on...

We can prove algebraically that the babylonian method converge. Note that all $x_n > 0$ and:

$$x_{n+1}^2 = \frac{1}{4}\left(x_n + \frac{a}{x_n}\right)^2 = \frac{1}{4}\left(x_n - \frac{a}{x_n}\right)^2 + a \geq a \tag{14.1.54}$$

Thus, all x_n past the first satisfy $x_n^2 \geq a$. Therefore:

$$x_{n+1} - x_n = \frac{a - x_n^2}{2x_n} \leq 0 \tag{14.1.55}$$

says that $x_{n+1} \leq x_n$. Therefore, $\{x_n\}$ is a decreasing sequence, bounded below. Thus, the sequence converges. Furthermore, since the sequence converges:

$$\lim_{n\to+\infty} a - x_n^2 = \lim_{n\to+\infty} 2x_n(x_{n+1} - x_n) = 0 \tag{14.1.56}$$

Therefore:

$$\lim_{n\to+\infty} x_n^2 = a \tag{14.1.57}$$

To conclude it is perhaps interesting to know that the reader can find in the section of Number Theory the method used in the antiquity (at least an analogy) using continued fractions.

Remark

We will see later that this above method is a special case of Newton's method introduced further below page 6612.

14.1.5 Archimedes Algorithm

The calculation of the universal constant "pi" denoted by π is certainly the algorithm with the greatest interest since we found this constant almost everywhere in physics and mathematics (there are numerous books on the subject available on the market).

We recall that we did not give the value of π in the section on Geometry or in other sections of this textbook until now. So we will now tackle this task.

We define in geometry the number named "pi", independently of the metric used, as the ratio of half the circumference of a circle with its radius such that:

$$\boxed{\pi = \frac{P/2}{R}} \tag{14.1.58}$$

It seems that we own the first algorithm of the calculation of this constant by Archimedes (9713-9788 according to holocene calendar) and whose proof is given below:

Proof 14.1.7. Consider an n-polygon inscribed in a circle (we start with a square of n sides of length $2u_n$ and if we go from a square to an octogone, two times more side, the sides length becomes $2u_{2n}$ and so on...):

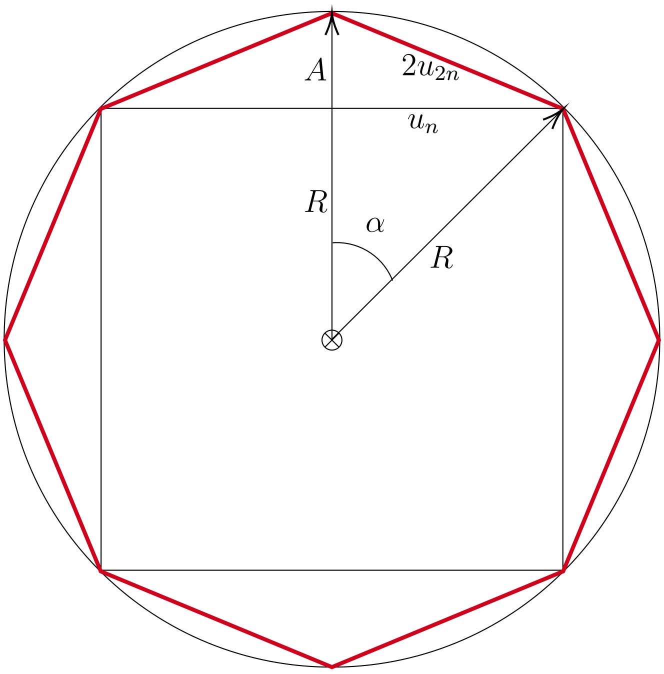


Figure 14.4: **The illustrated principle of Archimedes algorithm**

The principle of Archimedes algorithm is as follows:

Given the perimeter of a regular polygon of n sides inscribed in a circle of radius $R=1/2$ we can see in the figure above that by induction (see section Trigonometry page 2542):

$$P_n = 2nu_n \underset{n\to+\infty}{=} 2\sum R\alpha_i = 2R\sum \alpha_n = 2\frac{1}{2}\sum \alpha_n := \pi \tag{14.1.59}$$

We have for a n-polygon using basic trigonometry:

$$u_n = R\sin(\alpha) = \frac{1}{2}\sin\left(\frac{2\pi}{2n}\right) = \frac{1}{2}\sin\left(\frac{\pi}{n}\right) \tag{14.1.60}$$

and by using Pythagorean theorem:

$$(2u_{2n})^2 = u_n^2 + A^2 \tag{14.1.61}$$

with:

$$A = R - R\cos(\alpha) = \frac{1}{2} - \frac{1}{2}\cos\left(\frac{\pi}{n}\right) \tag{14.1.62}$$

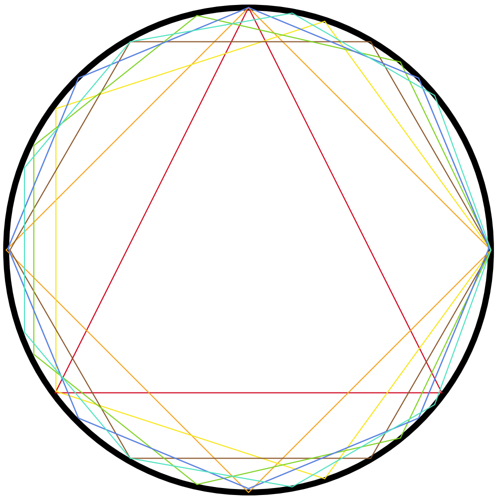
hence:

$$\begin{aligned}4u_{2n}^2 &= \left[\frac{1}{2}\sin\left(\frac{\pi}{n}\right)\right]^2 + \left[\frac{1}{2} - \frac{1}{2}\cos\left(\frac{\pi}{n}\right)\right]^2 = \frac{1}{4}\sin^2\left(\frac{\pi}{n}\right) + \frac{1}{4} - \frac{1}{2}\cos\left(\frac{\pi}{n}\right) + \frac{1}{4}\cos^2\left(\frac{\pi}{n}\right)\\ &= \frac{1}{4}\left[\sin^2\left(\frac{\pi}{n}\right) + \cos^2\left(\frac{\pi}{n}\right)\right] + \frac{1}{4} - \frac{1}{2}\cos\left(\frac{\pi}{n}\right) = \frac{2}{4} - \frac{1}{2}\cos\left(\frac{\pi}{n}\right)\\ &= \frac{1}{2} - \frac{1}{2}\sqrt{1-\sin^2\left(\frac{\pi}{n}\right)} = \frac{1}{2}\left[1 - \sqrt{1-\sin^2\left(\frac{\pi}{n}\right)}\right]\\ &= \frac{1}{2}\left(1-\sqrt{1-4u_n^2}\right)\end{aligned} \tag{14.1.63}$$

Therefore (keep in mind that this is the relation if we assume $R = 1/2$):

$$u_{2n} = \frac{1}{2}\sqrt{\frac{1}{2}\left(1-\sqrt{1-4u_n^2}\right)} \tag{14.1.64}$$

We then just need a computer and several iterations to evaluate with a good accuracy the value of π. Obviously, we use the Heron algorithm to calculate the square root...



■ Q.E.D.

For example for some iterations we have (it's simply by using Pythagoras theorem that we get immediately for the first row below u_n equal to $0.5 \cdot R \cdot \sqrt{2}$):

Sides (n)	u_n	**Perimeter** (estimated π)
4	$u_4 = 0.5 \cdot R \cdot \sqrt{2} = 0.353553$	$2 \cdot 4 \cdot u_4 = 2.828427125$
8	$u_8 = \dfrac{1}{2}\sqrt{\dfrac{1}{2}\left(1-\sqrt{1-4u_4^2}\right)} = 0.191342$	$2 \cdot 8 \cdot u_8 = 3.061467459$
16	$u_{16} = \dfrac{1}{2}\sqrt{\dfrac{1}{2}\left(1-\sqrt{1-4u_8^2}\right)} = 0.097545$	$2 \cdot 16 \cdot u_{16} = 3.121445152$
32	0.049009	3.136548491
64	0.024534	3.140331157
128	0.012271	3.141277251
256	0.006136	3.141513801
512	0.003068	3.14157294
1024	0.001534	3.141587725
2048	0.000767	3.141591422

where we have explicited the mechanism only of the first rows for pedagogical reasons.

Remark

There are a large number of algorithms to compute π. The one shown above, without being the most aesthetic, seems historically to be the first and most simple one.

14.1.6 Euler's Number e

Including the constant π, there are many other important mathematical constant that we need to generate with a computer (nowadays most constant values are stored as is and are not recalculated automatically). Among them, is the "Euler number" denoted by e (see section Functional Analysis page 2289). Let's see how to calculate this number.

Consider the Taylor series (see section Sequences and Series page 1696) for an infinitely differentiable function f given by:

$$f(u)=\sum_{i=0}^{+\infty}\frac{1}{n!}\frac{\mathrm{d}^n f(0)}{\mathrm{d}u^n}u^n \tag{14.1.65}$$

As (see section Differential and Integral Calculus page 1475):

$$\frac{\mathrm{d}^n e^u}{\mathrm{d}u^n}=e^u \tag{14.1.66}$$

Therefore we have:

$$f(u)=e^u=\sum_{i=0}^{+\infty}\frac{1}{n!}\frac{\mathrm{d}^n f}{\mathrm{d}u^n}(0)u^n=\sum_{i=0}^{+\infty}\frac{1}{n!}\frac{\mathrm{d}^n e^u}{\mathrm{d}u^n}(0)u^n=\sum_{i=0}^{+\infty}\frac{1}{n!}1u^n=\sum_{i=0}^{+\infty}\frac{u^n}{n!} \tag{14.1.67}$$

So finally:

$$\boxed{e=\sum_{i=0}^{+\infty}\frac{1}{n!}} \tag{14.1.68}$$

This relation provides an easy algorithm to calculate the Euler number to a given order n of precision.

14.1.7 Stirling's factorial approximation

Obviously, the factorial may be calculated with a simple iteration. However, this kind of method generates an exponential complexity algorithm which is not the best. Then there exists another method:

In mathematics, Stirling's approximation (or Stirling's formula) is an approximation for factorials. It is a very powerful approximation, leading to a very useful result in Statistical Mechanics (see section of the same name page 3554).

There are various approach that lead to different results. We will focus here on the only one that we will use later in theoretical physics and that is the worst easiest approximation.

Either the definition of the factorial:

$$n!=n(n-1)(n-2)\ldots 1 \tag{14.1.69}$$

And according to the properties of logarithms:

$$\ln(n!)=\sum_{m=1}^{n}\ln(m) \tag{14.1.70}$$

If n is very large (but very large) then the previous sum can be approximately written as an integral:

$$\ln(n!)=\sum_{m=1}^{n}\ln(m)\cong\int_{0}^{n}\ln(m)\mathrm{d}m \tag{14.1.71}$$

Solving this integral we get (see section Differential and Integral Calculus page 1534)

$$\ln(n!)\cong\int_{0}^{n}\ln(m)\mathrm{d}m=(m\ln(m)-m)\Big|_{0}^{n} \tag{14.1.72}$$

When $n\gg 1$, the lower limit is negligible and then (approximation that will be very useful in the section of Statistical Mechanics):

$$\ln(n!)\cong n\ln(n)-n\Leftrightarrow n!=e^{n(\ln(n)-1)} \tag{14.1.73}$$

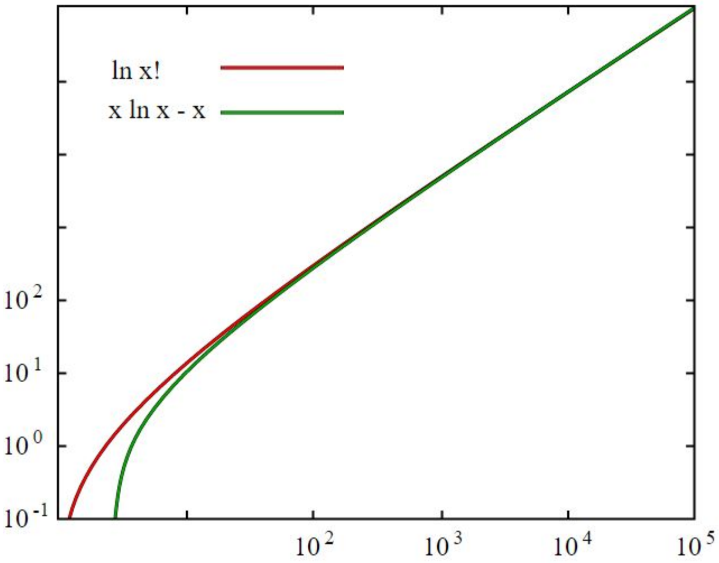


Figure 14.5: **Comparison of logarithmic formulations using Stirling formula**

After a small elementary simplification, we obtain:

$$\boxed{\ln(n!)\cong e^{n\ln(n)}e^{-n}=n^{n}\frac{1}{e^{n}}=\left(\frac{n}{e}\right)^{n}} \tag{14.1.74}$$

The latter relation is useful of course only if we assumes that Euler's constant is a value stored in the machine...

Let us consider the Gamma function (see section Differential and Integral Calculus page 1574):

$$\Gamma_{-1}(x)=\int_0^{+\infty} t^{x-1}e^{-t}\mathrm{d}t \tag{14.1.75}$$

and recall that for integers λ we have:

$$\Gamma(\lambda+1)=\lambda! \tag{14.1.76}$$

We get:

$$\Gamma(\lambda+1)=\int_0^{+\infty} t^{\lambda}e^{-t}\mathrm{d}t \tag{14.1.77}$$

Substituting $y=t/\lambda$ and letting $g(x)=y-\log(y)$ We get:

$$\Gamma(\lambda+1)=\lambda\int_0^{+\infty}(\lambda y)^y e^{-\lambda y}\mathrm{d}y=\lambda^{\lambda+1}\int_0^{+\infty} e^{-\lambda g(y)}\mathrm{d}y \tag{14.1.78}$$

Now we will use Laplace's method of integration (see section Differential and Integral Calculus page 1557) and for this we differentiate twice and get:

$$g'(y)=1-\frac{1}{y}\qquad g''(y)=\frac{1}{y^2} \tag{14.1.79}$$

so that $y^*=1$, $g(y^*)=1$ and $g''(y^*)=1$. Laplace's method now yields to:

$$\Gamma(\lambda+1)\cong\lambda^{\lambda+1}e^{-\lambda g(y^*)}\sqrt{\frac{2\pi}{\lambda g''(y^*)}}=\lambda^{\lambda+1/2}e^{-\lambda}\sqrt{2\pi} \tag{14.1.80}$$

The relation:

$$\boxed{\lambda!\cong\lambda^{\lambda+1/2}e^{-\lambda}\sqrt{2\pi}} \tag{14.1.81}$$

which is known as "Stirling's improved formula".

14.1.8 Linear Systems of Equations

There are many methods for solving systems of linear equations. Most of them have been developed to address particular systems. We will here study for the moment only one, named the "Gauss reduction method" or "Gauss pivot" or "Gauss reduction algorithm" or also "Gauss elimination" which is well suited for solving small systems of linear equations (up to 50 unknowns).

Remarks

R1. The validity of some of the operations that we will perform here to solve linear systems is proved in the section on Linear Algebra. In fact, to be brief, the whole method use vector spaces whose columns are linearly independent vectors.

R2. Recall that linear systems admit a solution if and only if the rank of the augmented matrix is less than or equal to the number of equations (see section Linear Algebra page 1938).

14.1.8.1 One equation with one unknown

We begin for sure with the easiest and smallest possible example... of one equation with one unknown:

$$ax=b \tag{14.1.82}$$

Where a and b are the coefficients of the equation and x is unknown. Solving this equation consists to determine x is according to a and b. If a is not equal to 0 then:

$$x=\frac{b}{a} \tag{14.1.83}$$

is the solution of the equation. If a is equal to zero and if b is different from 0 then the equation above admits no solution. If a and b are equal to zero, then the equation has infinitely many solutions.

14.1.8.2 Two equations with two unknowns

A linear system of two equations with two unknowns can be written as we know:

$$\begin{aligned}a_{11}x_1+a_{12}x_2&=b_1\\a_{21}x_1+a_{22}x_2&=b_2\end{aligned} \tag{14.1.84}$$

Where $a_{11},a_{12},a_{21},a_{22}$ are the coefficients of the equations, x_1 and x_2 are the unknowns.

Remark

The above-used notations have nothing to do with the tensor calculus!

To solve the system following the algorithm that interest us we proceed as follows:

Using elementary algebraic manipulations (addition or subtraction of the various equalities between themselves - operations authorized by the linear independence of vectors-lines) we transform the system into another with one unknown less for one of the both equations given by (for example):

$$\begin{aligned} a_{11}x_1+a_{12}x_2&=b_1\\ 0x_1+a'_{22}x_2&=b'_2 \end{aligned} \tag{14.1.85}$$

The transformation between the two systems:

$$\left.\begin{aligned} a_{11}x_1+a_{12}x_2 &= b_1\\ \\ a_{21}x_1+a_{22}x_2 &= b_2 \end{aligned}\right\} \Rightarrow \left\{\begin{aligned} a_{11}x_1+a_{12}x_2 &= b_1\\ \\ 0x_1+a'_{22}x_2 &= b'_2 \end{aligned}\right. \tag{14.1.86}$$

is simply done by multiplying each coefficient of the first equality by $\dfrac{a_{21}}{a_{11}}$ and subtracting the second line with the resulting equation. This procedure is named "row reduction" or "(naive) Gaussian elimination".

In our case, the element a_{11} is named the "pivot" or "leading coefficient". In others words, if the L_i is the notation for the each line number of the system, what we made is:

$$L_2-\frac{a_{21}}{a_{11}}L_1\rightarrow L_2 \tag{14.1.87}$$

Then, we solve the equation with only one unknown:

$$a'_{22}x_2=b'_2\Leftrightarrow x_2=\frac{b'_2}{a'_{22}} \tag{14.1.88}$$

We can therefore conclude with:

$$a_{11}x_1+a_{12}x_2=b_1\Rightarrow a_{11}x_1+a_{12}\frac{b'_2}{a'_{22}}=b_1\Leftrightarrow x_1=\frac{b_1-a_{12}\dfrac{b'_2}{a'_{22}}}{a_{11}} \tag{14.1.89}$$

But this is not the real algorithm. The real one use the augmented form:

$$\left[\begin{array}{cc|c} a_{11} & a_{12} & b_1\\ a_{21} & a_{22} & b_2 \end{array}\right] \tag{14.1.90}$$

Now we apply $L_2-\dfrac{a_{21}}{a_{11}}L_1\rightarrow L_2$ and $L_1-\dfrac{a_{12}}{a_{22}}L_2\rightarrow L_1$ such that:

$$\left[\begin{array}{cc|c} a_{11}-\dfrac{a_{12}}{a_{22}}a_{21} & 0 & b_1-b_2\dfrac{a_{12}}{a_{22}}\\ 0 & a_{22}-\dfrac{a_{21}}{a_{11}}a_{12} & b_2-b_1\dfrac{a_{21}}{a_{11}} \end{array}\right] \tag{14.1.91}$$

As you can see the matrix has been put in diagonal form! Now to continue, we write:

$$\left[\begin{array}{cc|c} a'_{11} & 0 & b'_1\\ 0 & a'_{22} & b'_2 \end{array}\right] \tag{14.1.92}$$

Now we just apply $\dfrac{1}{a'_{11}}L_1 \rightarrow L_1$ and $\dfrac{1}{a'_{12}}L_2 \rightarrow L_2$ such that:

$$\left[\begin{array}{cc|c} 1 & 0 & \dfrac{b'_1}{a'_{11}} \\ 0 & 1 & \dfrac{b'_2}{a'_{22}} \end{array}\right] \tag{14.1.93}$$

Finally:

$$x_1 = \frac{b'_1}{a'_{11}} \quad \text{and} \quad x_2 = \frac{b'_2}{a'_{22}} \tag{14.1.94}$$

In pseudocode (non-unique and very likely not optimized):

Algorithm 4: Gaussian elimination (naive) pseudocode algorithm

```
Data: $A$,$\vec{b}$
Result: $A'$,$\vec{b}'$
for $j = 1, 2, \ldots, n-1$ do
    for $i = j+1, j+2, \ldots, n$ do
        $m_{ij} = a_{ij}/a_{jj}$;
        for $k = j+1, j+2, \ldots, n$ do
            $a_{ik} = a_{ik} - m_{ij}a_{jk}$;
        $b_i = b_i - m_{ij}b_j$
```

14.1.8.3 Three equations with three unknowns

Now consider the case of the linear systems of three equations with three unknowns:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \end{aligned} \tag{14.1.95}$$

We can subsequently by elementary operations (see section Linear Algebra page 1929 and the previous case) reduce this linear system in the following echelon form system:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ 0x_1 + a''_{22}x_2 + a''_{23}x_3 &= b''_2 \\ 0x_1 + 0x_2 + a''_{33}x_3 &= b''_3 \end{aligned} \tag{14.1.96}$$

And therefore we can trivially solve the last line:

$$x_3 = \frac{b''_3}{a''_{33}} \tag{14.1.97}$$

And afterwards the second line:

$$x_2 = \frac{b''_2 - a''_{23}x_3}{a''_{22}} = \frac{b''_2 - a''_{23}\dfrac{b''_3}{a''_{33}}}{a''_{22}} \tag{14.1.98}$$

And finally:

$$x_1=\frac{b_1-a_{13}x_3-a_{12}x_2}{a_{11}}=\frac{b_1-a_{13}\dfrac{b_3''}{a_{33}''}-a_{12}\dfrac{b_2''-a_{23}''\dfrac{b_3''}{a_{33}''}}{a_{22}''}}{a_{11}} \tag{14.1.99}$$

Let us return to the systems transformations. It is carried out in two stages:

1. In the first line, we choose a_{11} as the pivot and we eliminate the coefficients a_{21} and a_{31} as follows:

   We have to multiply each coefficient of the first line by $\dfrac{a_{21}}{a_{11}}$ and subtract this result of the second line and therefore a_{21} disappears.

   Similarly, by multiplying the coefficients of the first line by $\dfrac{a_{31}}{a_{11}}$, and subtracting the result obtained from the third line, a_{31} disappears.

   The linear system of equations can therefore be written as:

   $$\begin{aligned}a_{11}x_1+a_{12}x_2+a_{13}x_3&=b_1\\0x_1+a_{22}'x_2+a_{23}'x_3&=b_2'\\0x_1+a_{32}'x_2+a_{33}''x_3&=b_3'\end{aligned} \tag{14.1.100}$$

2. The second step is to treat the linear system of two equations with two unknowns formed by the second and third lines of the previous system and that, in choosing a_{22}' as pivot. This method of resolution can seem complicated but it has the advantage of being generalized and be applied to solve linear systems of n equations in n unknowns.

✍**Example:**

Let us see an example taken on Wikipedia (helps to copy/paste boring LaTeX...):

$$\begin{cases}2x+y-z=8 & (L1)\\-3x-y+2z=-11 & (L2)\\-2x+y+2z=-3 & (L3)\end{cases} \tag{14.1.101}$$

Now we put:

$$\begin{aligned}L_2+\frac{3}{2}L_1&\rightarrow L_2\\L_3+L_1&\rightarrow L_3\end{aligned} \tag{14.1.102}$$

Therefore we get:

$$\begin{cases}2x+y-z=8\\\frac{1}{2}y+\frac{1}{2}z=1\\2y+z=5\end{cases} \tag{14.1.103}$$

Now we put:

$$L_3+-4L_2 \rightarrow L_3 \tag{14.1.104}$$

Therefore we get:

$$\begin{cases} 2x+\ y-\ z=8 \\ \frac{1}{2}y+\frac{1}{2}z=1 \\ -\ z=1 \end{cases} \tag{14.1.105}$$

To continue we put:

$$\begin{gathered} L_2+\frac{1}{2}L_3 \rightarrow L_2 \\ L_1-L_3 \rightarrow L_1 \end{gathered} \tag{14.1.106}$$

Therefore we get:

$$\begin{cases} 2x+\ y=7 \\ \frac{1}{2}y=\frac{3}{2} \\ -z=1 \end{cases} \tag{14.1.107}$$

We put:

$$\begin{gathered} 2L_2 \rightarrow L_2 \\ -L_3 \rightarrow L_3 \end{gathered} \tag{14.1.108}$$

Therefore we get:

$$\begin{cases} 2x+y=7 \\ y=3 \\ -z=1 \end{cases} \tag{14.1.109}$$

And finally we put:

$$\begin{gathered} L_1-L_2 \rightarrow L_1 \\ \frac{1}{2}L_1 \rightarrow L_1 \end{gathered} \tag{14.1.110}$$

To get:

$$\begin{cases} x=2 \\ y=3 \\ -z=1 \end{cases} \tag{14.1.111}$$

So this is again an application of the row reduction method (or Gaussian elimination).

For sure when you know that such systems can be solved using just a matrix inversion with a vector multiplication (see section Linear Algebra page 1929) all that stuff will be almost useless for most employees working in non-maths jobs.

Remark

If the procedure of solution showed earlier above gives a row where $0=$ 'something non-zero' or $0=0$, then it means that the system has no solution and respectively infinitely many solutions. We will see in the section of Linear Algebra that this is explained by the fact that the matrix of the system is a singular matrix (i.e. there is no inverse).

By the way... two matrix applications that have the same dimensions (number of rows and columns) but different components that can be obtained one from each other by linear operations are said to be "equivalent matrices". For example, taken from the example above, the matrix equivalent to the system, i.e.:

$$\begin{bmatrix} 2 & 1 & -1 \\ -3 & -1 & 2 \\ -2 & 1 & 2 \end{bmatrix} \tag{14.1.112}$$

is equivalent to:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \tag{14.1.113}$$

What is denoted:

$$\begin{bmatrix} 2 & 1 & -1 \\ -3 & -1 & 2 \\ -2 & 1 & 2 \end{bmatrix} \equiv \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \tag{14.1.114}$$

14.1.8.4 n equations with n unknowns

To simplify the writing, the coefficients will always be noted a_{ij} and not a'_{ij}, a''_{ij}, etc. at each stage of the calculation.

Given the linear system (we could also represent it as an augmented matrix to simplify the notations):

$$\begin{aligned} a_{11}x_1+a_{12}x_2+\ldots+a_{1n}x_n&=b_1\\ a_{21}x_1+a_{22}x_2+\ldots+a_{2n}x_n&=b_2\\ &\ldots\\ &\ldots\\ a_{n1}x_1+a_{n2}x_2+\ldots+a_{nn}x_n&=b_n \end{aligned} \tag{14.1.115}$$

We will choose a_{11} as the pivot to eliminate $a_{21}, a_{31}, \ldots, a_{n1}$. Then, removing $a_{32}, a_{42}, \ldots, a_{n2}$ is performed by taking a_{22} as a pivot. Last pivot to consider is obviously $a_{n-1,n-1}$, it helps eliminate $a_{n,n-1}$. The system then takes the form:

$$\begin{aligned} a_{11}x_1+a_{12}x_2+\ldots+a_{1n}x_n&=b_1\\ 0\cdot x_1+a_{22}x_2+\ldots+a_{2n}x_n&=b_2\\ &\ldots\\ &\ldots\\ 0\cdot x_1+0\cdot x_2+\ldots+a_{nn}x_n&=b_n \end{aligned} \tag{14.1.116}$$

And we can therefore solve the last equation, then the fore last equation, and so on up to the first one.

This method must however be fine-tuned to avoid pivots with 0 values. The trick is therefore to switch the order in which the equations are written to choose the pivot coefficient whose absolute value is the largest. Thus, in the first column, the better is pivot is the coefficient such that a_{j1}:

$$|a_{j1}|>|a_{i1}| \quad \forall i \tag{14.1.117}$$

It is taken to a_{11} by permutation of first lines and j-th lines. Removal of the rest of the first column can then be performed. Then, we begin again with the $n-1$ remaining equations.

✍**Example:**

A last example with 4 unknowns with a different representation that may help (we hope so!). Consider the following linear system:

$$\begin{bmatrix} 1 & -2 & 1 & 2 & 8\\ 2 & 3 & -2 & 3 & 14\\ 4 & -1 & 3 & -1 & 7\\ 3 & 2 & -4 & 5 & 5 \end{bmatrix}\cdot\begin{bmatrix} x\\ y\\ z\\ t \end{bmatrix}=\begin{bmatrix} 8\\ 14\\ 7\\ 5 \end{bmatrix} \tag{14.1.118}$$

Let us apply the Gauss elimination algorithm:

$$
\begin{gathered}
\begin{array}{rrrrrlll}
1 & -2 & 1 & 2 & 8 & |(-2) & |(-4) & |(-3) \\
2 & 3 & -2 & 3 & 14 & \leftarrow + & & \\
4 & -1 & 3 & -1 & 7 & \leftarrow & + & \\
3 & 2 & -4 & 5 & 5 & \leftarrow & & +
\end{array}
\\[2ex]
\begin{array}{rrrrrll}
1 & -2 & 1 & 2 & 8 & & \\
0 & 7 & -4 & -1 & -2 & |(-1) & |(-\frac{8}{7}) \\
0 & 7 & -1 & -9 & -25 & \leftarrow + & \\
0 & 8 & -7 & -1 & -9 & \leftarrow & +
\end{array}
\\[2ex]
\begin{array}{rrrrrl}
1 & -2 & 1 & 2 & 8 & \\
0 & 7 & -4 & -1 & -2 & \\
0 & 0 & 3 & -8 & -23 & |\,\frac{17}{3} \\
0 & 0 & -17 & 1 & -47 & \leftarrow +
\end{array}
\\[2ex]
\begin{array}{rrrrr}
1 & -2 & 1 & 2 & 8 \\
0 & 7 & -4 & -1 & -2 \\
0 & 0 & 3 & -8 & -23 \\
0 & 0 & 0 & -\frac{133}{3} & \frac{-532}{3}
\end{array}
\end{gathered}
\tag{14.1.119}
$$

And the last steps should be obvious to the reader (if not we can detail them on request as always!).

To summarize, given a $n \times n$ matrix A and n-vector $\vec{b}$, the following algorithm reduces the system of linear equations $A\vec{x} = \vec{b}$ to an equivalent upper triangular system. A and $\vec{b}$ are overwritten with the coefficients and right-hand side, respectively, of the reduced system.

Algorithm 5: Gaussian elimination

```
Input: $A$,$\vec{b}$
for $j = 1, 2, \ldots$ to $n-1$ do
    for $i = j+1, j+2, \ldots$ to $n$ do
        $m_{ij} = a_{ij}/a_{jj}$;
        for $k = j+1, j+2, \ldots$ to $n$ do
            $a_{ik} = a_{ik} - m_{ij}a_{jk}$;
        $b_i = b_i - m_{ij}b_j$
```

14.1.9 Polynomials

The basics polynomials with real coefficients has been studied in the section of Calculus quite in details (see page 1368). Here we will address only the digital aspect of some problems related to polynomials (that is to say elementary algorithms or formulas useful for some operations not included by default in most computer programming languages).

Apart from the addition and subtraction of polynomials which we assume as trivial (aside the optimization of the complexity aside as the Horner scheme), we will see how to multiply and divide two polynomials.

Let's first see how to multiply two polynomials.

Let:

$$\begin{aligned} f(x) &= a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n \\ g(x) &= b_0 + b_1 x + b_2 x^2 + \ldots + b_m x^m \end{aligned} \tag{14.1.120}$$

Therefore:

$$\begin{aligned} f(x)g(x) &= a_0 b_0 + (a_0 b_1 + a_1 b_0)x + (a_0 b_2 + a_1 b_1 + a_2 b_0)x^2 + \ldots \\ &= c_0 + c_1 x + c_2 x^2 + \ldots \end{aligned} \tag{14.1.121}$$

With for $k = 0, 1, 2, \ldots, n+m$:

$$\boxed{c_k = \sum_{i=0} a_i b_{k-i} \qquad k \in \{0, 1, 2, \ldots, n+m\}} \tag{14.1.122}$$

it was easy...

The second case of interests to us now is the Euclidean division of polynomials (see section Calculus page 1374).

Let us take again:

$$\begin{aligned} f(x) &= a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n \\ g(x) &= b_0 + b_1 x + b_2 x^2 + \ldots + b_m x^m \end{aligned} \tag{14.1.123}$$

but with the condition that $n \geq m$ that is to say $\text{deg}(f(x)) < \text{deg}(g(x))$.

The division can be written as we know (see the section of Number Theory or Calculus at page 1374):

$$f(x) = g(x)q(x) + r(x) \tag{14.1.124}$$

with:

$$\text{deg}(r(x)) < \text{deg}(g(x)) \tag{14.1.125}$$

otherwise $r(x)=0$.

It is normally known beforehand (since proved in the section Calculus again at page 1374) that we have:

$$\deg(q(x))=n-m \tag{14.1.126}$$

and:

$$\deg(r(x))<m \tag{14.1.127}$$

We have therefore by definition $q(x)$ that is the quotient of the division and $r(x)$ the remainder of the Euclidean division of $f(x)$ by $g(x)$.

Therefore, nothing prevents us from writing in the most general possible way:

$$\begin{aligned}q(x)&=q_0+q_1x+a_2x^2+\ldots+q_{n_m}x^{n-m}\\ r(x)&=r_0+r_1x+r_2x^2+\ldots+r_{m-1}x^{m-1}\end{aligned} \tag{14.1.128}$$

To prove the expression of different q_i, we have preferred for educational reasons to use a specific example (see below) whose result will be generalized.

✍**Example:**

Let:

$$\begin{aligned} f(x) &= a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4 \\ g(x) &= b_0 + b_1 x + b_2 x^2 \end{aligned} \tag{14.1.129}$$

So of what we have said before, we get (starting point):

$$\begin{aligned} q(x) &= q_0 + q_1 x + q_2 x^2 \\ r(x) &= r_0 + r_1 x \end{aligned} \tag{14.1.130}$$

Using the fact that (as a reminder):

$$f(x) = g(x)q(x) + r(x) \tag{14.1.131}$$

So we have almost immediately:

$$\begin{aligned} q_2 &= \frac{a_4}{b_2} \\ r_1(x) &= (a_3 - q_2 b_1)x^3 + (a_2 - q_2 b_0)x^2 + a_1 x + a_0 = c_3 x^3 + c_2 x^2 + c_1 x + c_0 \end{aligned} \tag{14.1.132}$$

Then (still proceeding in the same way):

$$\begin{aligned} q_1 &= \frac{c_3}{b_2} \\ r_2(x) &= (c_1 - q_1 b_1)x^2 + (c_1 - q_1 b_0)x + c_0 = c'_2 x^2 + c'_1 x + c'_0 \end{aligned} \tag{14.1.133}$$

And finally:

$$\begin{aligned} q_0 &= \frac{c'_2}{b_2} \\ r_3(x) &= r(x) = (c'_1 - q_0 b_1)x + (c'_0 - q_0 b_0) = c''_1 x + c'_1 x + c''_0 \end{aligned} \tag{14.1.134}$$

So in general:

$$q_{n-m} = \frac{a_n}{b_m} \tag{14.1.135}$$

As:

$$f(x) = g(x)q(x) + r(x) \Leftrightarrow r(x) = f(x) - q(x)g(x) \tag{14.1.136}$$

The first remainder is then:

$$\begin{aligned} r_1(x) &= f(x) - q_{n-m}g(x) \\ &= \underbrace{(a_n - q_{n-m}b_m)}_{=0}x^n + (a_{n-1} - q_{n-m}b_{m-1})x^{n-1} + (a_{n-2} - q_{n-m}b_{m-2})x^{n-2} + \ldots \\ &+ (a_{n-m} - q_{n-m}b_0)x^{n-m} + a_{n-m-1}x^{n-m-1} + \ldots + a_0 \\ &= c_{n-1}x^{n-1} + c_{n-2}x^{n-2} + \ldots + c_0 \end{aligned} \tag{14.1.137}$$

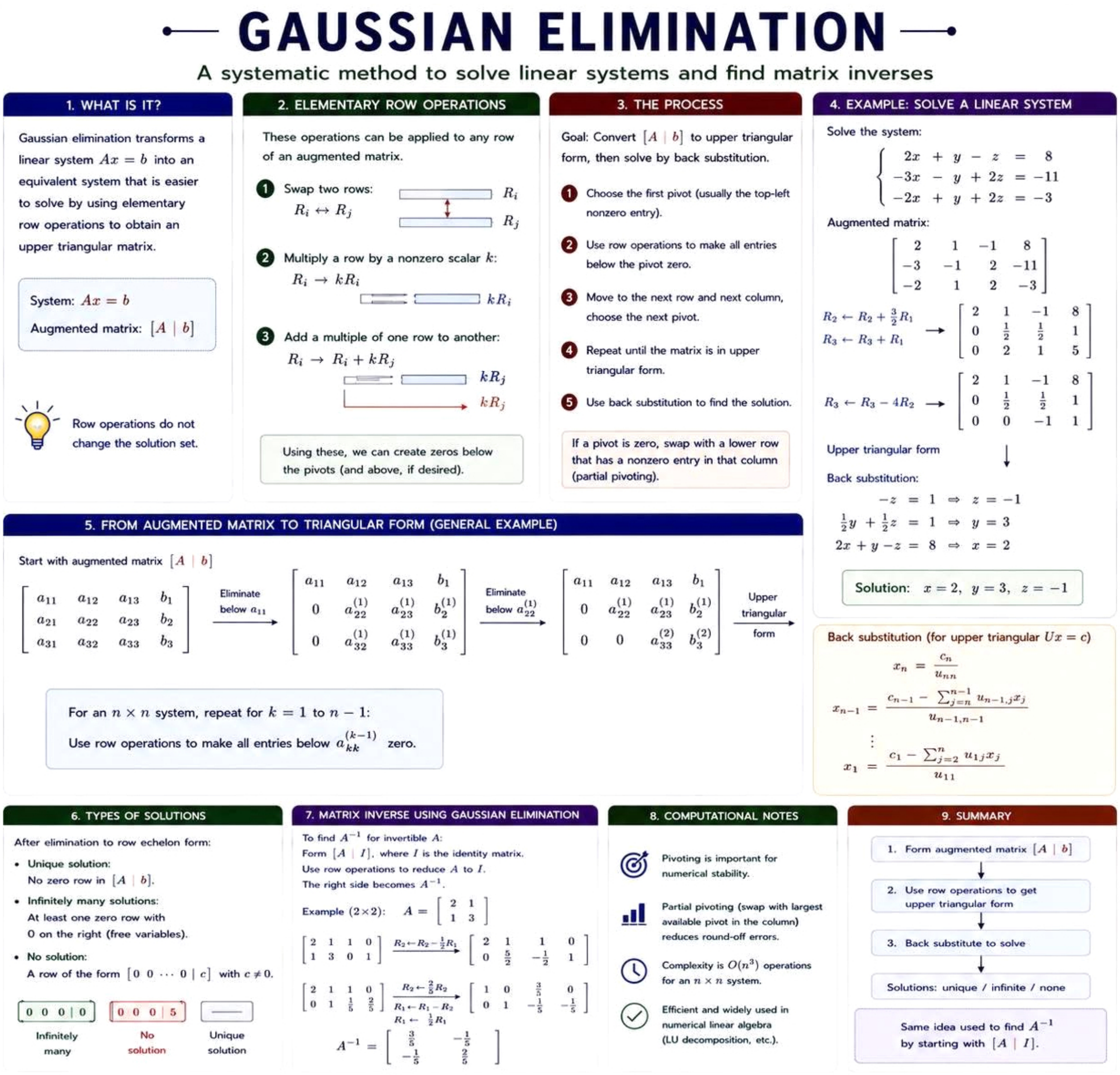
After:

$$q_{n-m-1}=\frac{c_{n-1}}{b_m} \tag{14.1.138}$$

The second remainder is then:

$$\begin{aligned}
r_2(x)&=r_1(x)-q_{n-m-1}g(x)\\
&=\underbrace{(c_{n-1}-q_{n-m-1}b_m)}_{=0}x^{n-1}+(c_{n-2}-q_{n-m-1}b_{m-1})x^{n-2}+(c_{n-3}-q_{n-m-1}b_{m-2})x^{n-3}+\ldots\\
&+(c_{n-m-1}-q_{n-m-1}b_0)x^{n-m-1}+c_{n-m-2}x^{n-m-2}+\ldots+c_0\\
&=c'_{n-2}x^{n-2}+c'_{n-3}x^{n-3}+\ldots+c'_0
\end{aligned} \tag{14.1.139}$$

and so on... we continue until $\deg(r_k(x))<m$.



GAUSSIAN ELIMINATION

A systematic method to solve linear systems and find matrix inverses

1. WHAT IS IT?

Gaussian elimination transforms a linear system $Ax=b$ into an equivalent system that is easier to solve by using elementary row operations to obtain an upper triangular matrix.

System: $Ax=b$

Augmented matrix: $[A \mid b]$

Row operations do not change the solution set.

2. ELEMENTARY ROW OPERATIONS

These operations can be applied to any row of an augmented matrix.

① Swap two rows: $R_i \leftrightarrow R_j$

② Multiply a row by a nonzero scalar k: $R_i \rightarrow kR_i$

③ Add a multiple of one row to another: $R_i \rightarrow R_i + kR_j$

Using these, we can create zeros below the pivots (and above, if desired).

3. THE PROCESS

Goal: Convert $[A \mid b]$ to upper triangular form, then solve by back substitution.

① Choose the first pivot (usually the top-left nonzero entry).

② Use row operations to make all entries below the pivot zero.

③ Move to the next row and next column, choose the next pivot.

④ Repeat until the matrix is in upper triangular form.

⑤ Use back substitution to find the solution.

If a pivot is zero, swap with a lower row that has a nonzero entry in that column (partial pivoting).

4. EXAMPLE: SOLVE A LINEAR SYSTEM

Solve the system:

$$\begin{cases} 2x + y - z = 8\\ -3x - y + 2z = -11\\ -2x + y + 2z = -3\end{cases}$$

Augmented matrix:

$$\begin{bmatrix} 2 & 1 & -1 & 8\\ -3 & -1 & 2 & -11\\ -2 & 1 & 2 & -3\end{bmatrix}$$

$$\begin{array}{l} R_2 \leftarrow R_2+\frac{3}{2}R_1\\ R_3 \leftarrow R_3+R_1\end{array} \rightarrow \begin{bmatrix} 2 & 1 & -1 & 8\\ 0 & \frac{1}{2} & \frac{1}{2} & 1\\ 0 & 2 & 1 & 5\end{bmatrix}$$

$$R_3 \leftarrow R_3-4R_2 \rightarrow \begin{bmatrix} 2 & 1 & -1 & 8\\ 0 & \frac{1}{2} & \frac{1}{2} & 1\\ 0 & 0 & -1 & 1\end{bmatrix}$$

Upper triangular form ↓

Back substitution:

$$\begin{aligned} -z &= 1 \Rightarrow z=-1\\ \tfrac{1}{2}y+\tfrac{1}{2}z &= 1 \Rightarrow y=3\\ 2x+y-z &= 8 \Rightarrow x=2\end{aligned}$$

Solution: $x=2,\ y=3,\ z=-1$

5. FROM AUGMENTED MATRIX TO TRIANGULAR FORM (GENERAL EXAMPLE)

Start with augmented matrix $[A \mid b]$

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & b_1\\ a_{21} & a_{22} & a_{23} & b_2\\ a_{31} & a_{32} & a_{33} & b_3\end{bmatrix} \xrightarrow{\text{Eliminate below } a_{11}} \begin{bmatrix} a_{11} & a_{12} & a_{13} & b_1\\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & b_2^{(1)}\\ 0 & a_{32}^{(1)} & a_{33}^{(1)} & b_3^{(1)}\end{bmatrix} \xrightarrow{\text{Eliminate below } a_{22}^{(1)}} \begin{bmatrix} a_{11} & a_{12} & a_{13} & b_1\\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & b_2^{(1)}\\ 0 & 0 & a_{33}^{(2)} & b_3^{(2)}\end{bmatrix} \xrightarrow[\text{form}]{\text{Upper triangular}}$$

For an $n \times n$ system, repeat for $k=1$ to $n-1$:

Use row operations to make all entries below $a_{kk}^{(k-1)}$ zero.

Back substitution (for upper triangular $Ux=c$)

$$\begin{aligned} x_n &= \frac{c_n}{u_{nn}}\\ x_{n-1} &= \frac{c_{n-1}-\sum_{j=n}^{n-1} u_{n-1,j}x_j}{u_{n-1,n-1}}\\ &\vdots\\ x_1 &= \frac{c_1-\sum_{j=2}^{n} u_{1j}x_j}{u_{11}}\end{aligned}$$

6. TYPES OF SOLUTIONS

After elimination to row echelon form:

- Unique solution:
  No zero row in $[A \mid b]$.
- Infinitely many solutions:
  At least one zero row with 0 on the right (free variables).
- No solution:
  A row of the form $[0\ 0 \cdots 0 \mid c]$ with $c \neq 0$.

Infinitely many

No solution

Unique solution

7. MATRIX INVERSE USING GAUSSIAN ELIMINATION

To find A^{-1} for invertible A:
Form $[A \mid I]$, where I is the identity matrix.
Use row operations to reduce A to I.
The right side becomes A^{-1}.

Example (2×2): $A=\begin{bmatrix} 2 & 1\\ 1 & 3\end{bmatrix}$

$$\begin{bmatrix} 2 & 1 & 1 & 0\\ 1 & 3 & 0 & 1\end{bmatrix} \xrightarrow{R_2\leftarrow R_2-\frac{1}{2}R_1} \begin{bmatrix} 2 & 1 & 1 & 0\\ 0 & \frac{5}{2} & -\frac{1}{2} & 1\end{bmatrix}$$

$$\begin{bmatrix} 2 & 1 & 1 & 0\\ 0 & 1 & \frac{1}{5} & \frac{2}{5}\end{bmatrix} \xrightarrow[\substack{R_1\leftarrow R_1-R_2\\ R_1\leftarrow \frac{1}{2}R_1}]{R_2\leftarrow \frac{2}{5}R_2} \begin{bmatrix} 1 & 0 & \frac{3}{5} & 0\\ 0 & 1 & -\frac{1}{5} & -\frac{1}{5}\end{bmatrix}$$

$$A^{-1}=\begin{bmatrix} \frac{3}{5} & -\frac{1}{5}\\ -\frac{1}{5} & \frac{2}{5}\end{bmatrix}$$

8. COMPUTATIONAL NOTES

Pivoting is important for numerical stability.

Partial pivoting (swap with largest available pivot in the column) reduces round-off errors.

Complexity is $O(n^3)$ operations for an $n \times n$ system.

Efficient and widely used in numerical linear algebra (LU decomposition, etc.).

9. SUMMARY

1. Form augmented matrix $[A \mid b]$
↓
2. Use row operations to get upper triangular form
↓
3. Back substitute to solve
↓
Solutions: unique / infinite / none

Same idea used to find A^{-1} by starting with $[A \mid I]$.

14.1.10 Regression Techniques

Regressions are very useful and very important tools for statisticians, engineers, computer scientists, marketing analysts, physicists, physicians, economists wishing to establish a law of correlation between two (or more) variables, do a qualitative analysis, an extrapolation or even to separate signal form noise (as every point outside the regression will be considered as being noise).

There are many regression methods: the simple solution of first degree equations (when only two points of measurement are known) to equations that permits to obtain from a large number of points information that are essential to the establishment of a linear regression, polynomial, logistic or other law (or function).

Let us give a list of the most used regressions techniques used in business and administrations (whose mathematical models are not all shown in this section yet but will be when we will have more time):

1. **Simple Linear Regression (SLR) model**[1] with one (**Univariate Linear Regression (CLR)**) or more variables (**Multiple Linear Regression (MLR)**) based on the method of least[2] squares with binary or continuous variables with response variable and coefficients[3] belonging to $\mathbb{R}$. Presented in detail in this section of the book (implicitly this model contains the interactions between variables and also some non-linear models).

2. **Gaussian linear regression model** (statistical approach of linear regression based on the method of least squares) with binary or continuous variables with response variable also belonging to $\mathbb{R}$. When a regression model assumes that the errors are normally distributed, we speak commonly of **general linear model**. Presented in detail in this section of the book. A gaussian linear regression model is a special case of the generalized linear model (GLM).

3. **Weighted least squares regression model** is used in the case where the assumption $\mathrm{V}(\varepsilon_i)=\sigma_\varepsilon^2$ of Gaussian linear regression is not satisfied. The purpose is to apply a transformation trick using simple weights (see further below for more details), to keep this assumption valid.

4. **Nonlinear regression models** with binary or continuous variables with response variable in $\mathbb{R}$. Presented in detail in this section of the if they can be reduced to linear case or not but then no interactions of explanatory variables. Otherwise based on quasi-Newton techniques type or Gauss-Newton presented also this section.

5. **Polynomial regression model by the method of B-splines** or of the collocation polynomial with response variable in $\mathbb{R}$. The most common case is to restrict ourselves to the

---

[1] Also named **Classical Linear Regression (CLR)**

[2] Don't confuse "estimation methods" like OLS (ordinary least squares), TLS (total least squares), ML (maximum likelihood), REML (restricted maximum likelihood), PQL (penalized quasi-likelihood) with "regression methods"!

[3] If the coefficients are required to be non-negative, we speak then of **Non-Negative Least Squares** and if we require the coefficient to be in some bounds then we speak of **Bounded-Variable Least-Squares** or **Constrained Least Squares**.

restriction of third order splines and then we speak of **restricted cubic splines regression (RCS)**. Presented in detail in this section of the book.

6. **Logistic regression models** (binomial/multinomial/ordinal regressions) with binary, nominal variables (categorical) or ordinal or continuous with response variable bounded between 0 and 1. Presented summarily and naively in this section of the book. A logistic, probit or robit (robust version of the logit model) regression model is a special case of the generalized linear model (GLM).

7. **Logic regression model** is a (generalized) regression methodology that is primarily applied when most of the covariates in the data to be analysed are binary. The goal of logic regression is to find predictors that are Boolean (logical) combinations of the original predictors.

8. **Principal Component Regression (PCR)** instead of regressing the dependent variable on the explanatory variables directly, the principal components of the explanatory variables are used as regressors. Therefore PLS can reduces the number of predictors (dimensional reduction) to a smaller set of uncorrelated components and performs least squares regression on these components.

9. **Partial Least squares Regression (PLR)** is a statistical method that bears some relation to principal components regression. They differ in the methods used in extracting factor scores. In short, principal components regression produces like weight matrix reflecting the covariance structure between the predictor variables, while partial least squares regression produces like a weight matrix reflecting the covariance structure between the predictor and response variables.

10. **Counting Poisson regression** (Poisson MLE, PMLE, GLM, Poisson-quasi-Lindley) or **negative binomial regression** (binomial MLE and QGPMLE) model with binary, nominal (categorical) or ordinal or continuous variables with positive integer answer in $\mathbb{N}$. All the regression models that do not assume the errors to me normally distributed, but following other well-known statistical distributions, are known under the name of **Generalized Linear regression Models (GLM)**.

11. **Beta regression** is commonly used by practitioners to model variables that assume values in the standard unit interval $[0,1]$. It is based on the assumption that the dependent variable is beta-distributed and that its mean is related to a set of regressors through a linear predictor with unknown coefficients and a link function. A beta regression model is also a special case of the generalized linear model (GLM).

12. **Simplex regression** is actually (12019 according to holocene calendar) considered as a more robust and flexible alternative to the beta regression dedicated of rates or proportions. This regression is based on the simplex distribution, hence it's name...

13. **Orthogonal linear regression model** (or Deming regression) that is used as complement to the paired T-test to check the stability of the measuring instruments in laboratories. This is a case where the explanatory and dependent variables are tainted with uncertainty. Notice that at the opposite of all regression techniques listed above and below that have no assumptions about the distribution of the predictor (independent) variables, the orthogonal linear regression is as far as we know, the only regression technique that has an assumption of the distribution of the predictor!

14. **Total linear regression model** (also named Total Least Squares (TLS)) that is a generalization of Deming and Orthogonal linear regression. TLS aims to minimize the total geometric distance between the data points and the best-fit line rather than just the vertical distance as in Ordinary Least Squares (OLS).

15. **Quantile regression model** (very useful in the medical and economic fields) based on the same idea as the regression by the method of least squares, but where we do not minimizes the sum of squared errors from the average, but the sum of absolute errors from a given quantile (median or other). It's also used sometimes to get rid of extreme values.

16. **Theil–Sen estimator method** also named Sen's slope estimator or Slope selection method or Single median method or Kendall robust line-fit method or Kendall–Theil robust line... that is a very simple method for robustly fitting a line to a set of points (simple linear regression) that chooses the median a_M of the slopes of all lines through pairs of two-dimensional sample points. For the estimate of the intercepts the USGS recommends the following calculation $b = y_M - a_M x_M$. That's all... it's quite simple in fact.

17. **Least Absolute Deviation (LAD) regression model** that use absolute values rather than square errors. But absolute values are difficult to work with in mathematics (especially Calculus) as absolute values results in discontinuous derivatives that cannot be treated analytically

18. **LOESS (LOcally Estimated Scatterplot Smoothing)** and **LOWESS (LOcally WEighted Scatterplot Smoothing)** are two strongly related nonparametric regression methods. These methods are purely numerical (does therefore not provide any unique formula) and are performed by fitting simple models to localized subsets (we speak then of "segmented regression"). In fact, one of the chief attractions of this method is that is not required to specify a global function of any form to fit a model to the data, only to fit segments of the data.

19. **Multivariate adaptive regression splines (MARS)** (the term "MARS" is trademarked and licensed to Salford Systems. In order to avoid trademark infringements, many open source implementations of MARS are named "Earth"...) and as it names describes it well... is uses splines to interpolate and also extrapolate know data.

20. **Bayesian linear regression model** is an approach to linear regression in which the statistical analysis is undertaken within the context of Bayesian inference, that means we have some prior knowledge about the regression coefficients and the error term.

21. **Ridge regression model**[4] that use a trick on the information matrix X^TX (see further below) by adding a constant in it. But the constraints of usage are quite boring (no intercept and coefficients normalized) that make it quite difficult to interpret. In fact it is a special case of a regression family named **regularized linear regression models** and the underlying mathematical trick can be applied ton any general linear model (simple linear regression, logistic regressions, etc.).

22. **Least Absolute Shrinkage and Selection Operator model (LASSO)**[5] that is based on the same ideas as that of the Ridge regression model and therefore also a special case

---

[4] Technically speaking it's not a regression but a method of regularization! It's only a procedure affecting (sometimes improving) the method of fitting. It can be applied to almost any kind of regression.

[5] Technically speaking it's also not a regression but a method of regularization!

of the Regularized Linear Regression models family (the only point that differs with the Ridge is that the theoretical model has somewhere a exponent equal to 1 for LASSO when for it is equal to 2 for Ridge).

23. **Elastic net regression model**[6] is only a mathematical mixture of the LASSO and Ridge regression. This model therefore also belongs to the family of Regularized Linear Regression models.

24. **Two-Stage Least Squares (2SLS) Regression** is used when the dependent variable's error terms are correlated with the independent variables. Additionally, it is useful when there are feedback loops in the model. In a first stage this regression builds instrumental variables that are uncorrelated with the error terms to compute estimated values of the problematic predictor(s), and then in a second stage uses those computed values to estimate a linear regression model of the dependent variable. This technique is used mainly in **Structural equation modelling (SEM)**.

25. **Moderated Multiple Regression (MMR)** is used when one of the explicative variable has an interaction with another variable (product) that is continuous or dichotomous. The variable that interacts with the normal explicative variable is named the "moderator variable". A common error in practice is to make a confusion between Moderated Multiple Regression and a nested hierarchical model (i.e. Nested ANOVA).

26. **Independent Component Regression (ICR)** is just a simple linear regression but based on a previous dimensional reduction of the exogenous variables based on an independent component analysis (ICA). That means the final model used the most explaining variance independent component as exogenous variables. Indeed, the reader must remember that PCA creates uncorrelated components but "uncorrelated" doesn't mean "independent" (!!!) as we have proved it in the section Statistics.

27. **Support Vector Machine regression (SVMr)** use the concept of Support Vector Machine classification technique to evaluate existing values and interpolate new one.

28. **Bootstrap or Jackknife regression model**[7] using all previous models and then by resampling techniques (see further below) gives the possibility to have robust estimators that could in some situations be hard to get analytically.

29. **Isotonic regression** is a technique of fitting a free-form line to a sequence of observations under the following constraints: the fitted free-form line has to be non-decreasing (or non-increasing) everywhere, and it has to lie as close to the observations as possible.

30. **Cox regression model** is a non-parametric technique based on the assumption of constant proportional hazard to analyse the survival rate in clinical trials.

31. **Threshold Regression Models** is a parametric technique used to relax the assumption of constant proportional hazard of the Cox regression and therefore obviously also used to analyse the survival rate in clinical trials.

... and ... for a given number of these approaches we differentiate mathematical models taking into account the censored data and uncensored data, constraints coefficients (like the fact that they should be non-negative, like non-negative least square NNLS, or sum up to 1 or any other specific constraint that we categorize under the name of Bounded-Variable Least Squares (BVLS) or Constrained Least Square). This makes a bunch of theories/models to study in final and this is why this subsection is one of the biggest of the whole book and that some models are given in some other sections of the book (especially in the Statistics section).

---

[6] Technically speaking it's also not a regression but a method of regularization!

[7] Technically speaking they are not "regression models" but methods related to the process of estimation!

Remark

There are also submodels families. Like the Panel Ordinary Least Squares (Panel OLS), linear mixed models (LMM) that includes fixed effects models (FE models) and random effects models (RE models) and mixed effect models (ME models) and longitudinal regression models. There is also Nonlinear Mixed Model (NLMM) that includes all linear mixed models, and Generalized Linear Models (to not be confuse with "general linear models) that includes NLMM. And also convex regression, etc.

Here is a figure that summarize quite well the different regressions and their relationships (thanks to Adrian Olszweski for having authorized us to reproduce it!):

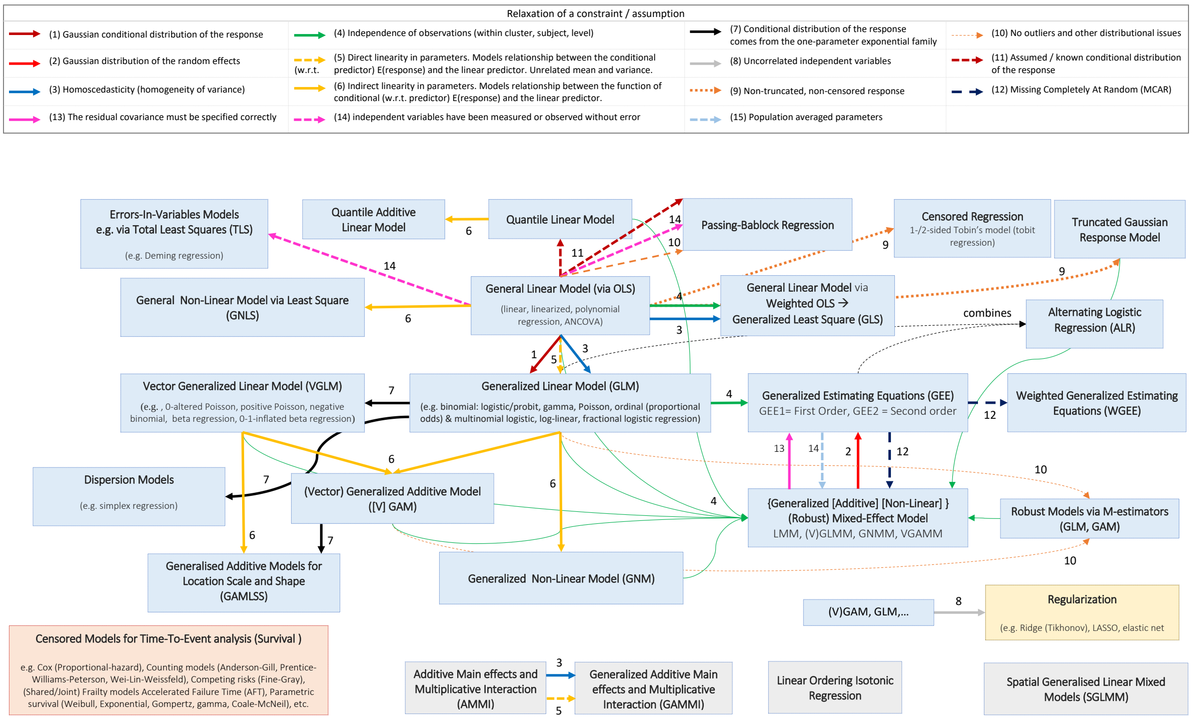

Relaxation of a constraint / assumption
(1) Gaussian conditional distribution of the response
(4) Independence of observations (within cluster, subject, level)
(7) Conditional distribution of the response comes from the one-parameter exponential family
(10) No outliers and other distributional issues
(2) Gaussian distribution of the random effects
(w.r.t.
(5) Direct linearity in parameters. Models relationship between the conditional predictor) E(response) and the linear predictor. Unrelated mean and variance.
(8) Uncorrelated independent variables
(11) Assumed / known conditional distribution of the response
(3) Homoscedasticity (homogeneity of variance)
(6) Indirect linearity in parameters. Models relationship between the function of conditional (w.r.t. predictor) E(response) and the linear predictor.
(9) Non-truncated, non-censored response
(12) Missing Completely At Random (MCAR)
(13) The residual covariance must be specified correctly
(14) independent variables have been measured or observed without error
(15) Population averaged parameters
Errors-In-Variables Models e.g. via Total Least Squares (TLS)
(e.g. Deming regression)
Quantile Additive Linear Model
6
Quantile Linear Model
14
10
Passing-Bablock Regression
Censored Regression
1-/2-sided Tobin's model (tobit regression)
9
Truncated Gaussian Response Model
11
14
General Non-Linear Model via Least Square (GNLS)
6
General Linear Model (via OLS)
(linear, linearized, polynomial regression, ANCOVA)
4
3
General Linear Model via Weighted OLS → Generalized Least Square (GLS)
9
combines
Alternating Logistic Regression (ALR)
1
5
3
Vector Generalized Linear Model (VGLM)
(e.g. , 0-altered Poisson, positive Poisson, negative binomial, beta regression, 0-1-inflated beta regression)
7
Generalized Linear Model (GLM)
(e.g. binomial: logistic/probit, gamma, Poisson, ordinal (proportional odds) & multinomial logistic, log-linear, fractional logistic regression)
4
Generalized Estimating Equations (GEE)
GEE1= First Order, GEE2 = Second order
12
Weighted Generalized Estimating Equations (WGEE)
13
14
2
12
10
Dispersion Models
(e.g. simplex regression)
7
6
(Vector) Generalized Additive Model ([V] GAM)
6
4
{Generalized [Additive] [Non-Linear] } (Robust) Mixed-Effect Model
LMM, (V)GLMM, GNMM, VGAMM
Robust Models via M-estimators (GLM, GAM)
6
7
10
Generalised Additive Models for Location Scale and Shape (GAMLSS)
Generalized Non-Linear Model (GNM)
(V)GAM, GLM,...
8
Regularization
(e.g. Ridge (Tikhonov), LASSO, elastic net
Censored Models for Time-To-Event analysis (Survival)
e.g. Cox (Proportional-hazard), Counting models (Anderson-Gill, Prentice-Williams-Peterson, Wei-Lin-Weissfeld), Competing risks (Fine-Gray), (Shared/Joint) Frailty models Accelerated Failure Time (AFT), Parametric survival (Weibull, Exponential, Gompertz, gamma, Coale-McNeil), etc.
Additive Main effects and Multiplicative Interaction (AMMI)
3
5
Generalized Additive Main effects and Multiplicative Interaction (GAMMI)
Linear Ordering Isotonic Regression
Spatial Generalised Linear Mixed Models (SGLMM)

Figure 14.6: Synoptic summary and relationships between various regression models (author: Adrian Olszweski)

For a more exhaustive list of regression techniques and their relationships, the reader can refer to the Data Science mind map at page 6770.

Caution!!!!!!!!! There is a huge difference in science by only adjusting a mathematical function (having no real physical variables) to observations, and modelling observations with a physical model, i.e. including real physical variables and physics laws (the first one is naive, and most of time wrong, and the second one is obviously robust and most of time less wrong)!

Finally, note that in the linear regression, explanatory variables form a linear expression but that does not mean they are themselves linear. Thus, if we consider the two expressions below:

$$\begin{aligned} y &= a + be^{x_1} + c\log(x_2) \\ y &= a + x_1^{\beta_1} + x_2 e^{\beta_2} \end{aligned} \tag{14.1.140}$$

the first is linear in the parameters but the second is not!

Finally, a word on a technique sometimes used for qualitative interpretation of the influential of explanatory variables trough their coefficients for simple or multivariate linear regressions:

When the amplitudes of some explanatory variables (continuous!) have very different orders of magnitudes this raises a big problem of interpretation of the influence of each variable by reading their coefficient and also generates problems of calculations precision in algorithms because of differences in size order and thus also generates rounding errors!

The traditional idea is then to center-reduce all the values of explanatory variables which helps greatly to interpretation of the influence of these variables (but we must then leave out the interpretation of the numerical value of the response variable). But take care to a common trap!!! Once the theoretical model obtained from normalized (center-reduced) variables, the new values to explain should be obtained by having previously centered-reduced the new explanatory values but by subtracting the old average and reducing by the old standard deviations of respectively each of the injected explanatory variable in the model!

14.1.10.1 Univariate linear regression model

We will present here several algorithms (methods) useful and known in experimental science (we have already discussed about some of them during our study of statistics). The goal is to express the linear relation between two variables x (explanatory variable) and y (response variable) independently by a "linear model LM" as simple as possible (otherwise it would take hundreds of pages to introduce the topic!).

Remark

Writing this textbook we have hesitated a long time to put regression techniques in the Statistics section. But because in practice the choice of the type of regression is empirical it has seem to us most convenient to put this subject here.

Definition (#14.1.558): In univariate regression we have:

D1. x is the independent variable or "explanatory variable" also named "covariate" or "predictor" (in economics "exogenous variable"...). The x values are set by the experimenter and are assumed to be known without error.

D2. y is the dependent variable or "explained variable" (e.g. the answer of the analyser) also named in economy "endogenous variable". y values are most of time measured with an error (bias) of measurement. One goal of regression is to estimate precisely this error.

We seek a relation of the form:

$$y=f(x)=ax+b \tag{14.1.141}$$

This is the equation of a straight line (affine function), hence the term "linear regression" where a is named in the study framework of regression techniques: "regression coefficient" instead of "slope" as seen in previous sections.

In real life, linear relations are an exception because most phenomenon are nonlinear in reality and even non-continuous in certain situations... Furthermore, it is not because they are linear in a given interval of measurements that are still linear at a smaller-scale or larger scale (zoom bias)!

However, in practice we make transformation to linearise functions either by elementary algebraic transformations like those used by spreadsheets softwares (e.g. Microsoft® Excel®) as for example the linearisation of a logarithmic function by making a simple change of variables:

$$y=a\underbrace{\ln(x)}_{:=x}+b=ax+b \tag{14.1.142}$$

or as:

$$y=a\underbrace{e^x}_{:=x}+b=ax+b \tag{14.1.143}$$

or as:

$$y=a\underbrace{\sin(x)}_{:=x}+b=ax+b \tag{14.1.144}$$

or for power and exponential functions by also making a small algebraic manipulation with the properties of logarithms as proved in the section of Functional Analysis at page 2284 (under the assumption that a is strictly positive):

$$\begin{gathered}y=ax^b\Rightarrow\ln(y)=\ln(ax^b)=\ln(a)+\ln(x^b)=b\ln(x)+\ln(a)\\ y=ae^{bx}\Rightarrow\ln(y)=\ln(ae^{bx})=\ln(a)+\ln(e^{bx})=bx+\ln(a)\end{gathered}\tag{14.1.145}$$

or by making Taylor series approximations (see section Sequences and Series page 1696).

Caution! One of the biggest misconception and myth by data scientistis (but not by statisticians) is that for linear regression the independent variables need to have linear relationship with the y variable. No! Linear regression refers to the fact that the equation is linear in the coefficients, not necessarily the predictor themselves. This means that the model can handle predictors that are transformed (e.g. squared, logarithmic) as long as the equation remains linear in the parameters. For example, a quadratic model (with squared terms) still falls under linear regression as long as the model is linear in therms of the coefficients (betas)!

We will see during our study of Multiple Linear Regression further below at page 6384 how to deal with polynomial regression and Fourier series regression.

Remark

If we seek to determine the value of y for an unmeasured x and lying beyond the original interval of measurement, then we speak of "extrapolation" or in more complicated cases of "forecast with prediction interval for x". We will see that further below.

14.1.10.1.1 Regression line

In the common way to make an univariate linear regression of the type:

$$y=f(x)=ax+b\tag{14.1.146}$$

there exist multiple methods.

This first and most simple method in our point of view relies on the properties of covariance and mean (see section Statistics page 597) and is widely used among others in elementary finance (but in fact in any filed where there are some statistics).

Consider X,Y two variables, one of which depends on the other (often it is y that depends on x). According to the covariance bilinearity property (see section Statistics page 602) we recall that we have from:

$$\text{cov}(X,Y)=\text{E}(XY)+\text{E}(X)\text{E}(Y)\tag{14.1.147}$$

the following relation:

$$\begin{aligned}\text{cov}(X,Y)&=\text{cov}(X,aX+b)\\&=\text{E}(aX^2+bX)-\text{E}(X)\text{E}(aX+b)=\text{E}(aX^2)+\text{E}(bX)-\text{E}(X)(\text{E}(aX)+\text{E}(b))\\&=\text{E}(aX^2)+\text{E}(bX)-\text{E}(X)(\text{E}(aX)+\text{E}(b))\\&=a\text{E}(X^2)+b\text{E}(X)-\text{E}(X)(a\text{E}(X)+b)=a\text{E}(X^2)+b\text{E}(X)-a\text{E}(X)^2-b\text{E}(X)\\&=a\left(\text{E}(X^2)-\text{E}(X)^2\right)=a\text{cov}(X,X)=a\text{V}(X)\end{aligned} \tag{14.1.148}$$

So it comes for the regression coefficient (we will reuse this relation during our study of yield of a portfolio according to Sharpe model in the section Economy):

$$a=\frac{\text{cov}(X,Y)}{\text{V}(X)} \tag{14.1.149}$$

Either in a more explicit form which that we will use later (using the relation determined earlier at page 6323):

$$\boxed{a=\frac{\text{E}(x_k-\bar{x})(y_k-\bar{y})}{\dfrac{1}{n}\displaystyle\sum_k(x_k-\bar{x})^2}=\frac{\dfrac{1}{n}\displaystyle\sum_k(x_k-\bar{x})(y_k-\bar{y})}{\dfrac{1}{n}\displaystyle\sum_k(x_k-\bar{x})^2}=\frac{\displaystyle\sum_k(x_k-\bar{x})(y_k-\bar{y})}{\displaystyle\sum_k(x_k-\bar{x})^2}} \tag{14.1.150}$$

To determine the intercept we use the properties of the expected mean as proved in the section Statistics (see page 585):

$$\text{E}(Y)=\text{E}(aX+b)=\text{E}(aX)+\text{E}(b)=a\text{E}(X)+b \tag{14.1.151}$$

Therefore we have b as:

$$\boxed{b=\text{E}(Y)-a\text{E}(X)} \tag{14.1.152}$$

14.1.10.1.2 Ordinary Least Squares Method (OLSM)

Due to the error on y, the experimental points, of coordinates (x_k,y_k) do not lie exactly on the theoretical line. We can therefore find the equation of the experimental line passing closest to these points.

The "least squares method" (LSM) will be under the particular study we are interested in, to look for the values of the parameters a,b that minimize the sum of squares of residual e_i (also named as we already know "Sum of Squared Residuals" (SSR) or "Sum of Squared Errors" (SSE)) between the observed values y_k and the theoretical calculated values y'_k. We then speak sometimes about the "least squares method of ordinate deviations"...:

$$\text{SSR}=\sum_{k=1}^{n}(y_k-y'_k)^2 \tag{14.1.153}$$

where n is the number of measured points and the theoretical values given by:

$$y'_k = ax_k + b \tag{14.1.154}$$

Therefore written explicitly:

$$\text{SSR} = \sum_{k=1}^{n} (y_k - ax_k - b)^2 := \Phi(a,b) \tag{14.1.155}$$

This relation shows the sum of squared deviations as a function of the parameters a, b. When this function is minimal (extremal), the derivatives with respect to these parameters cancels:

$$\begin{cases} \dfrac{\partial \Phi(a,b)}{\partial a} = -2\displaystyle\sum_{k=1}^{n} x_k (y_k - ax_k - b)) = 0 \\ \dfrac{\partial \Phi(a,b)}{\partial b} = -2\displaystyle\sum_{k=1}^{n} (y_k - ax_k - b)) = 0 \end{cases} \tag{14.1.156}$$

Remark

This method of minimum research (optimization) is named "method of Lagrange multipliers" in the world of Economy (we will detail this method further below). In our example SSR is the scalar value that will be used as Lagrange multiplier.

Therefore after simplification and rearrangement:

$$\begin{cases} b\displaystyle\sum_{k=1}^{n} x_k + a\sum_{k=1}^{n} x_k^2 = \sum_{k=1}^{n} x_k y_k \\ bn + a\displaystyle\sum_{k=1}^{n} x_k = \sum_{k=1}^{n} y_k \end{cases} \tag{14.1.157}$$

The above system is named "normal equations". This is a linear system of two equations with two unknowns. Let us write to simplify the notation:

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k \qquad \bar{y} = \frac{1}{n}\sum_{k=1}^{n} y_k \tag{14.1.158}$$

The system becomes:

$$\begin{cases} bn\bar{x} + a\displaystyle\sum_{k=1}^{n} x_k^2 = \sum_{k=1}^{n} x_k y_k \\ b + a\bar{x} = \bar{y} \end{cases} \tag{14.1.159}$$

From the second line we get without surprise:

$$b = \bar{y} - a\bar{x} = \text{E}(Y) - a\text{E}(X) \tag{14.1.160}$$

By replacing in the first line, we get:

$$\begin{aligned}
&(\bar{y}-a\bar{x})n\bar{x}+a\sum_{k=1}^{n}x_k^2=\sum_{k=1}^{n}x_ky_k \Leftrightarrow n\bar{x}\bar{y}-na\bar{x}^2+a\sum_{k=1}^{n}x_k^2=\sum_{k=1}^{n}x_ky_k\\
&\Leftrightarrow a=\frac{\displaystyle\sum_{k=1}^{n}x_ky_k-n\bar{x}\bar{y}}{\displaystyle\sum_{k=1}^{n}x_k^2-n\bar{x}^2}=\frac{\displaystyle n\sum_{k=1}^{n}x_ky_k-\sum_{k=1}^{n}x_k\sum_{k=1}^{n}y_k}{\displaystyle n\sum_{k=1}^{n}x_k^2-\left(\sum_{k=1}^{n}x_k\right)^2}
\end{aligned}\tag{14.1.161}$$

From there we get from the second line:

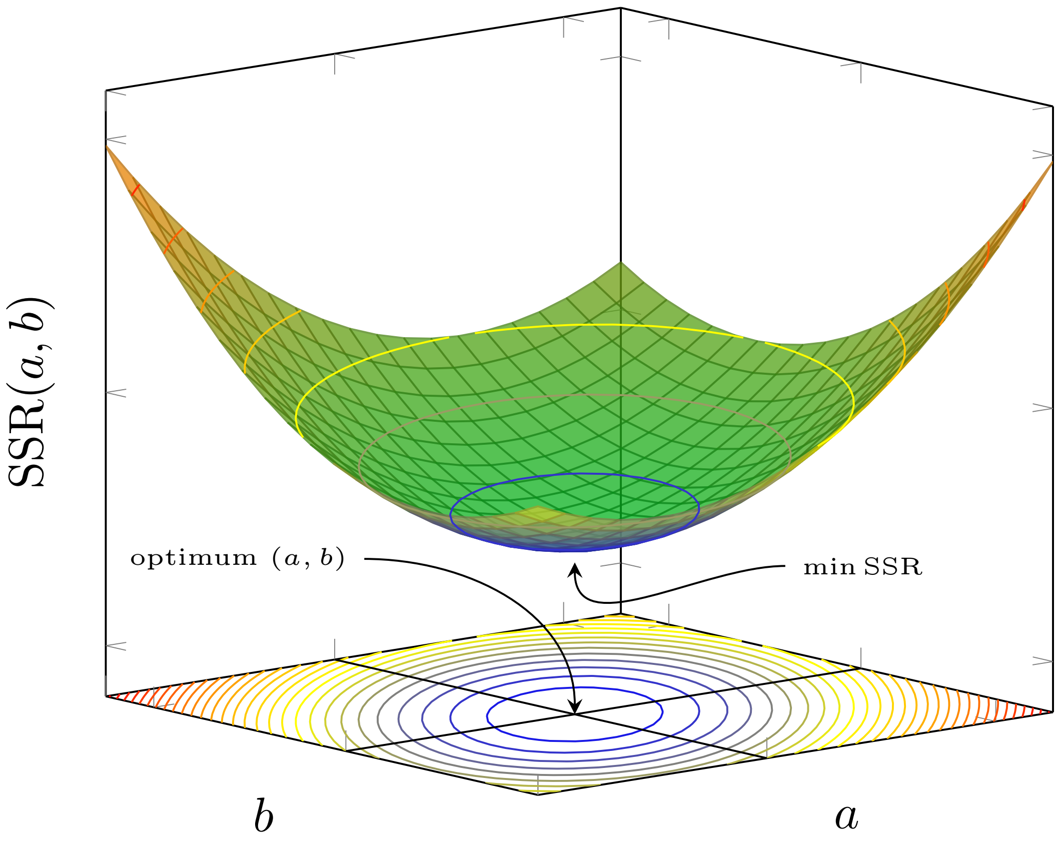
$$\begin{aligned}
b&=\frac{1}{n}\sum_{k=1}^{n}y_k-\frac{\displaystyle n\sum_{k=1}^{n}x_ky_k-\sum_{k=1}^{n}x_k\sum_{k=1}^{n}y_k}{\displaystyle n\sum_{k=1}^{n}x_k^2-\left(\sum_{k=1}^{n}x_k\right)^2}\frac{1}{n}\sum_{k=1}^{n}x_k\\
&=\frac{\displaystyle\left(n\sum_{k=1}^{n}x_k^2-\left(\sum_{k=1}^{n}x_k\right)^2\right)\frac{1}{n}\sum_{k=1}^{n}y_k-\left(n\sum_{k=1}^{n}x_ky_k-\sum_{k=1}^{n}x_k\sum_{k=1}^{n}y_k\right)\frac{1}{n}\sum_{k=1}^{n}x_k}{\displaystyle n\sum_{k=1}^{n}x_k^2-\left(\sum_{k=1}^{n}x_k\right)^2}\\
&=\frac{\displaystyle\sum_{k=1}^{n}x_k^2\sum_{k=1}^{n}y_k-\sum_{k=1}^{n}x_k\sum_{k=1}^{n}x_ky_k}{\displaystyle n\sum_{k=1}^{n}x_k^2-\left(\sum_{k=1}^{n}x_k\right)^2}
\end{aligned}\tag{14.1.162}$$

Thus, the terms of the slope and intercept of the straight line equation are:

$$\boxed{a=\frac{\displaystyle n\sum_{k=1}^{n}x_ky_k-\sum_{k=1}^{n}x_k\sum_{k=1}^{n}y_k}{\displaystyle n\sum_{k=1}^{n}x_k^2-\left(\sum_{k=1}^{n}x_k\right)^2}\qquad b=\frac{\displaystyle\sum_{k=1}^{n}x_k^2\sum_{k=1}^{n}y_k-\sum_{k=1}^{n}x_k\sum_{k=1}^{n}x_ky_k}{\displaystyle n\sum_{k=1}^{n}x_k^2-\left(\sum_{k=1}^{n}x_k\right)^2}}\tag{14.1.163}$$

The last two relations are used by a majority of spreadsheet softwares such as in the English version of Microsoft® Excel® 11.8346 when using the `REGRESSION( )` function. The term b (the y-intercept) may be obtained directly with the `INTERCEPT( )` function and a with the `SLOPE( )` function and the whole with the `LINEST( )` function.

In reality most softwares use the general case to find (a, b) based on the Newton-Gradient descent (see page 6678) where it seeks for the minimum of the SSR as illustrated below:



Here for information an interesting little list of some very practical case with this spreadsheet software (because requested a lot):

- For a straight line:

  a: = `SLOPE(y, x)`
  b: = `INTERCEPT(y, x)`

- For a logarithmic function (we see here again the change of variable given at the before):

  a: = `INDEX(LINEST(y, LN (x)), 1)`
  b: = `INDEX(LINEST(y, LN (x)), 1, 2)`

- For a power function (once again we see the change of variable given at the beginning):

  a: = `EXP(INDEX(LINEST(LN(y),LN(x)),1,2))`
  b: = `INDEX(LINEST(LN(y),LN(x)),1)`

- For an exponential function (we also find the change of variable given at the beginning):

  a: = `EXP(INDEX(LINEST(LN(y),x),1,2))`
  b: = `INDEX(LINEST(LN(y),x),1)`

Remark

We must keep in mind that the line of least squares, which can best summarize the cloud of observations points by minimizing SSR, necessarily passes through the center of gravity of the cloud, that is to say, by an average point that corresponds rarely to an observation (mean average of abscissas and ordinates).

Caution! Keep in mind that least squared and linear regression are note the same! Least squares is a method to estimate the parameters. Linear regression is a model. Least squares is one method used to estimate the parameters in Linear Regression. There are other methods as well like Maximum Likelihood that we will see later!

14.1.10.1.3 Univariate Regression Variance Analysis

Before starting it is important that the reader abandons immediately the possible reflex that would be to try to bring by successive analogies the regression ANOVA we will see right now to the categorical ANOVA we have studied in the section Statistics (see page 1001)!

Let us start with:

$$y=ax+b \tag{14.1.164}$$

Either in discrete form:

$$y_k=ax_k+b \tag{14.1.165}$$

As well as by the construction of the least squares method we have the following relation:

$$\bar{y}_k=a\bar{x}_k+b \tag{14.1.166}$$

Now we assume that each measured value is attached by an error such that:

$$y_k=ax_k+b+e_k \tag{14.1.167}$$

Either by subtracting the last two relations:

$$y_k-\bar{y}=a(x_k-\bar{x})+e_k\Leftrightarrow e_k=(y_k-\bar{y})-a(x_k-\bar{x}) \tag{14.1.168}$$

Now let us go through an intermediate result. Remember that we obtained earlier:

$$\begin{cases}\dfrac{\partial\Phi(a,b)}{\partial a}=-2\displaystyle\sum_{k=1}^{n}x_k(y_k-ax_k-b))=0\\ \dfrac{\partial\Phi(a,b)}{\partial b}=-2\displaystyle\sum_{k=1}^{n}(y_k-ax_k-b))=0\end{cases} \tag{14.1.169}$$

And therefore:

$$\begin{cases}\displaystyle\sum_{k=1}^{n}x_k(y_k-ax_k-b))=0\\ \displaystyle\sum_{k=1}^{n}(y_k-ax_k-b))=0\end{cases} \tag{14.1.170}$$

By replacing b by its value:

$$b=\bar{y}-a\bar{x} \tag{14.1.171}$$

We therefore get:

$$\begin{cases}\displaystyle\sum_{k=1}^{n}x_k(y_k-ax_k-\bar{y}+a\bar{x}))=\sum_{k=1}^{n}x_k((y_k-\bar{y})-a(x_k-\bar{x}))=0\\ \displaystyle\sum_{k=1}^{n}(y_k-ax_k-\bar{y}+a\bar{x}))=\sum_{k=1}^{n}((y_k-\bar{y})-a(x_k-\bar{x}))=0\end{cases} \tag{14.1.172}$$

Multiplying the second line above by $\bar{x}$ and by subtracting from the first we get:

$$\sum_{k=1}^{n} x_k((y_k-\bar{y})-a(x_k-\bar{x}))-\bar{x}\sum_{k=1}^{n}((y_k-\bar{y})-a(x_k-\bar{x})) \tag{14.1.173}$$

Therefore after rearrangement:

$$\sum_{k=1}^{n}(x_k-\bar{x})((y_k-\bar{y})-a(x_k-\bar{x}))=0 \tag{14.1.174}$$

Now we go back to:

$$y_k-\bar{y}=a(x_k-\bar{x})+e_k \tag{14.1.175}$$

If we put it all square and summing for all observations, we get:

$$\sum_{k=1}^{n}(y_k-\bar{y})^2=a^2\sum_{k=1}^{n}(x_k-\bar{x})^2+2a\sum_{k=1}^{n}(x_k-\bar{x})e_k+\sum_{k=1}^{n}e_k^2 \tag{14.1.176}$$

Therefore:

$$\sum_{k=1}^{n}(y_k-\bar{y})^2=a^2\sum_{k=1}^{n}(x_k-\bar{x})^2+2a\sum_{k=1}^{n}(x_k-\bar{x})((y_k-\bar{y})-a(x_k-\bar{x}))+\sum_{k=1}^{n}e_k^2 \tag{14.1.177}$$

But we have shown just before the double product was equal to zero. Therefore:

$$\sum_{k=1}^{n}(y_k-\bar{y})^2=a^2\sum_{k=1}^{n}(x_k-\bar{x})^2+\sum_{k=1}^{n}e_k^2 \tag{14.1.178}$$

This last relation is named "ANOVA equation" or "variance analysis equation". In fact, it is the sums of squares. We would need to divide it by n to obtain biased variances.

This last relation is often written:

$$\boxed{\text{SST}=\text{SSE}+\text{SSR}} \tag{14.1.179}$$

where SST is the "total sum of squares", SSE the "sum of square errors" and SSR "sum of squares residuals".

$$\begin{aligned}\sum_{k=1}^{n}(y_k-\bar{y})^2&=\sum_{k=1}^{n}(ax_k-a\bar{x})^2+\sum_{k=1}^{n}e_k^2=\sum_{k=1}^{n}(ax_k+b-a\bar{x}-b)^2+\sum_{k=1}^{n}e_k^2\\&=\sum_{k=1}^{n}(ax_k+b-\bar{y})^2+\sum_{k=1}^{n}e_k^2\end{aligned} \tag{14.1.180}$$

Let us note now the estimated y_k that minimize the errors such that the error is null in a different way and let us named that the "a priori linear model":

$$\hat{y}_k=ax_k+b \tag{14.1.181}$$

Hence the equality we will reuse several times (it is just previous relations without error term):

$$a^2\sum_{i=k}^{n}(x_k-\bar{x})^2:=\sum_{i=k}^{n}(\hat{y}_k-\bar{y})^2 \tag{14.1.182}$$

It is indeed important in practice to differentiate the a priori model that does not take into account the errors of the real model that does!

Because of the previous equality the relation:

$$\sum_{k=1}^{n}(y_k-\bar{y})^2=\sum_{k=1}^{n}(ax_k-a\bar{x})^2+\sum_{k=1}^{n}e_k^2 \tag{14.1.183}$$

Can be written:

$$\sum_{k=1}^{n}(y_k-\bar{y})^2=\sum_{i=k}^{n}(\hat{y}_k-\bar{y})^2+\sum_{k=1}^{n}e_k^2 \tag{14.1.184}$$

More explicitly:

$$\boxed{\sum_{k=1}^{n}(y_k-\bar{y})^2=\sum_{i=k}^{n}(\hat{y}_k-\bar{y})^2+\sum_{k=1}^{n}(y_k-\hat{y}_k)^2} \tag{14.1.185}$$

This last relation can be represented graphically as follows:

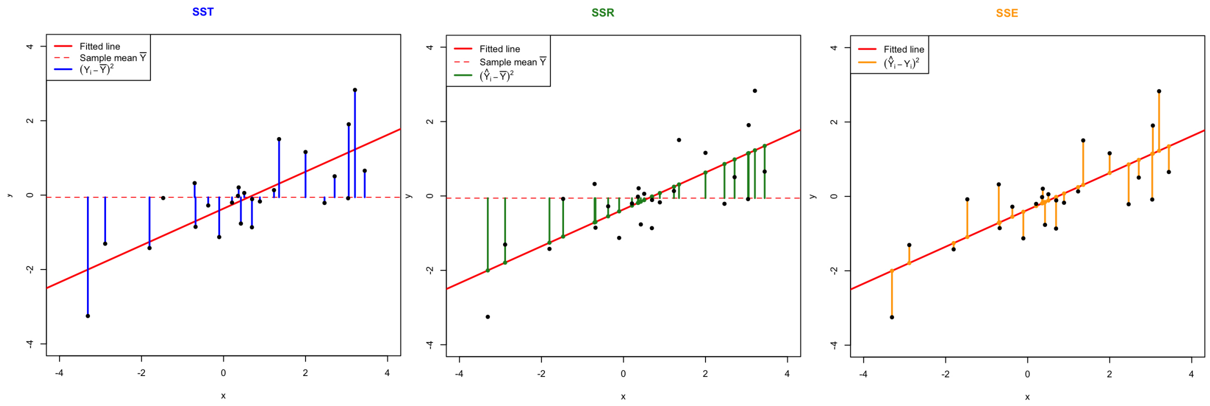


Figure 14.7: **Graphical representation of respectively SST, SSE, SSR**

The last relation is sometimes denoted also in the literature in a most educational way as follows:

$$\sum_{k=1}^{n}(y_k-\bar{y})^2=\sum_{i=k}^{n}(f_k-\bar{y})^2+\sum_{k=1}^{n}(y_k-f_k)^2 \tag{14.1.186}$$

which is just another way to write the variance decomposition (implicit variance):

$$\text{SST}=\text{SSE}+\text{SSR} \tag{14.1.187}$$

and it then comes immediately the relation sometimes used in practice to calculate residues (knowing the calculated values and measured values):

$$\sum_{k=1}^{n}e_k^2=\sum_{i=k}^{n}(y_k-\bar{y})^2-\sum_{i=k}^{n}(\bar{y}_k-\bar{y})^2 \tag{14.1.188}$$

It is important to remember that the above relation between SST, SSE and SSR are valid only in the case of a linear model!

It is also important to note that in this particularly variance decomposition we have:

$$\sum_{k=1}^{n} e_k^2 \neq \sum_{i=k}^{n}(y_k-\hat{y})^2 \tag{14.1.189}$$

Remember now that we have proved in the section Statistics, we found that the correlation coefficient was given (defined) by:

$$R_{X,Y}=\frac{\text{cov}(X,Y)}{\sqrt{\text{V}(X)\text{V}(Y)}} \Rightarrow R_{X,Y}^2=\frac{(\text{cov}(X,Y))^2}{\text{V}(X)\text{V}(Y)}=\frac{(\text{cov}(X,Y))^2}{\sigma_X^2\sigma_Y^2} \tag{14.1.190}$$

Or else since we have shown above that (remember that the indicated variance is an estimated variance in practice!):

$$a=\frac{\text{cov}(X,Y)}{\text{V}(X)} \tag{14.1.191}$$

We can therefore write the correlation coefficient in the form:

$$R_{X,Y}^2=\frac{(\text{cov}(X,Y))^2}{\text{V}(X)\text{V}(Y)}=\frac{a^2\text{V}(X)^2}{\text{V}(X)\text{V}(Y)}=\frac{a^2\text{V}(X)}{\text{V}(Y)} \tag{14.1.192}$$

So we deduce from this using the relations established in the section of statistics:

$$\begin{aligned}R_{X,Y}^2&=\frac{(\text{cov}(X,Y))^2}{\text{V}(X)\text{V}(Y)}=\frac{(\text{E}((x_k-\bar{x})(y_k-\bar{y}))^2}{\text{E}((x_k-\bar{x})^2)\text{E}((y_k-\bar{y})^2)}=\frac{\left(\dfrac{1}{n}\sum_k(x_k-\bar{x})(y_k-\bar{y})\right)^2}{\dfrac{1}{n}\displaystyle\sum_k(x_k-\bar{x})^2\frac{1}{n}\sum_k(y_k-\bar{y})^2}\\&=\frac{\left(\displaystyle\sum_k(x_k-\bar{x})(y_k-\bar{y})\right)^2}{\displaystyle\sum_k(x_k-\bar{x})^2\sum_k(y_k-\bar{y})^2}\end{aligned} \tag{14.1.193}$$

Remember that we proved above that:

$$a=\frac{\displaystyle\sum_k(x_k-\bar{x})(y_k-\bar{y})}{\displaystyle\sum_k(x_k-\bar{x})^2} \quad \text{and} \quad a^2\sum_{i=k}^{n}(x_k-\bar{x})^2:=\sum_{i=k}^{n}(\hat{y}_k-\bar{y})^2 \tag{14.1.194}$$

Therefore:

$$R_{X,Y}^2=\frac{\left(\displaystyle\sum_k(x_k-\bar{x})(y_k-\bar{y})\right)^2}{(x_k-\bar{x})^2(y_k-\bar{y})^2}=\frac{a^2\displaystyle\sum_k(x_k-\bar{x})^2}{\displaystyle\sum_k(y_k-\bar{y})^2}=\frac{\displaystyle\sum_k(\hat{y}_k-\bar{y})^2}{\displaystyle\sum_k(y_k-\bar{y})^2} \tag{14.1.195}$$

that is to say:

$$\boxed{R^2_{X,Y}=\frac{\text{SSE}}{\text{SST}}=\frac{\displaystyle\sum_k(\hat{y}_k-\bar{y})^2}{\displaystyle\sum_k(y_k-\bar{y})^2}}\qquad(14.1.196)$$

Remark

This formulation of the correlation coefficient is extremely useful because, unlike the statistical formulation, the latter generalizes immediately to the multiple linear regression we will see a further below.

And in the framework of regression models here are some typical cases of the value of the linear correlation coefficient with the first two lines and non-linear for the third line:

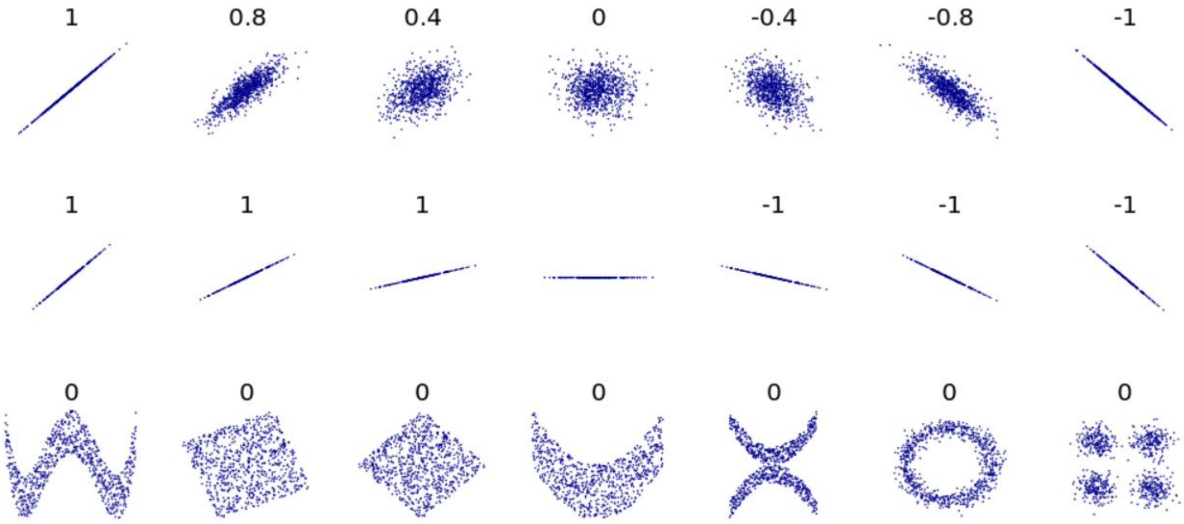


Figure 14.8: **Some values of the linear correlation coefficient (source: Wikipedia)**

Finally let us indicate that we also find very often the linear correlation coefficient as follows in softwares and literature:

$$R^2_{X,Y}=\frac{\text{SSE}}{\text{SST}}=\frac{\text{SST}-\text{SSR}}{\text{SST}}-1=1-\frac{\text{SSR}}{\text{SST}}=1-\frac{\sigma^2_{\varepsilon}}{\text{V}(y)}=1-\frac{\displaystyle\sum_k(y_k-f_k)^2}{\displaystyle\sum_k(y_k-\bar{y})^2}\qquad(14.1.197)$$

The last form highlights better that if the sum of the squares of residues SSR residues is zero, the measures are perfectly modelised by a linear relation in the range of study considered.

Remark

Some softwares communicate the "predicted R-squared" defined as:

$$R^2_{p.}=1-\frac{\text{PRESS}}{\text{SST}}\qquad(14.1.198)$$

where PRESS, the "predicted residual sum of squares", is a form of cross-validation (see further below page 6428) used in regression analysis to provide a summary measure of the fit of a model to a sample of observations that were not themselves used to estimate the model. It is calculated as the sums of squares of the prediction residuals for those observations:

$$\text{PRESS}=\sum_k (y_k-\hat{y}_{k,-k})^2 \tag{14.1.199}$$

Therefore PRESS differs from the sum of squares of the residual error in that each fitted value, $\hat{y}_k$, for PRESS is obtained from the remaining $n-1$ observations.

Given this procedure, the PRESS statistic can be calculated for a number of candidate model structures for the same dataset, with the lowest values of PRESS indicating the best structures. Models that are over-parametrised (over-fitted) would tend to give small residuals for observations included in the model-fitting but large residuals for observations that are excluded.

Modern textbooks use the average of the PRESS such that we define the "Leave-One-Out Cross-Validation score" (LOOCV) or "ordinary cross-validation" (OCV):

$$\text{LOOCV}=\frac{1}{n}\text{PRESS}=\frac{1}{n}\sum_k (y_k-\hat{y}_{k,-k})^2 \tag{14.1.200}$$

Finally, note that the ordinate value is not involved in the value of the correlation coefficient since (bilinearity property of covariance as proved in the section Statistics at page 597):

$$\text{cov}(X,Y)=\text{cov}(X,aX+b)=a\text{cov}(X,X)+b\text{cov}(X,1)=a\sigma_X^2+b\cdot 0 \tag{14.1.201}$$

14.1.10.1.4 F-test for Regression (significance test for linear regression)

Now comes a part that interests us mainly for practical scientific laboratories tests!!!

If your graduate statistical training was anything like mine, you learned ANOVA in one class and Linear Regression in another. My professors would often say things like « *ANOVA is just a special case of Regression* » but give vague answers when pressed. Let us see why!

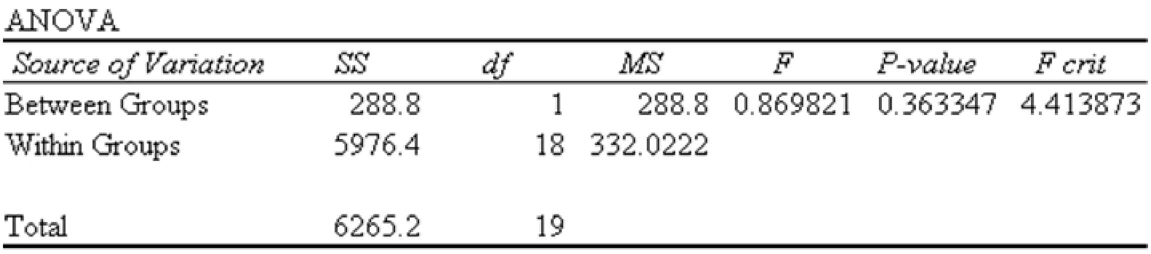
Let us recall that in the section of Statistics we proved (see page 1017) that for the one controlled factor ANOVA, the test of equality of means (through the use of the variances) was written (we just change the letter x to y to avoid confusion for the developments that will follow) is given by:

$$F_{k-1,N-k}=\frac{\mathrm{V}_A}{\mathrm{V}_R}=\frac{\dfrac{\displaystyle\sum_{j=1}^{k} n_j(\bar{y}_i-\bar{\bar{y}})^2}{k-1}}{\dfrac{\displaystyle\sum_{ij}(y_{ij}-\bar{y}_j)^2}{N-k}}=\frac{\dfrac{Q_A}{k-1}}{\dfrac{Q_R}{N-k}}=\frac{\text{MSk}}{\text{MSE}}=\frac{\text{variance between treatments}}{\text{variance within treatments}}=\frac{\text{signal}}{\text{noise}} \tag{14.1.202}$$

which is used to compare, for example, the means of the cash-flows of two supposed independent chosen months (i.e. with $k=2$) over 10 years (obviously in the case $k=2$ we could use a Student test if the conditions are of course satisfied!) supposing that for every year for the given month the values are normally distributed:

January	**February**
304	280
284	303
290	294
310	270
320	276
270	310
309	290
293	312
315	260
301	325

Therefore a simple ANOVA with one controlled factor will give according to the calculations proven in the section of Statistics (see page 1004) and using a spreadsheet software like Microsoft® Excel® 14.0.7106 the following results:



ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	288.8	1	288.8	0.869821	0.363347	4.413873
Within Groups	5976.4	18	332.0222			
Total	6265.2	19				

With for recall:

Source	**Sum of squares (SSE)**	χ^2 **df**	**Mean squares**	F	**Critical** F
Inter-Class	$Q_A = \sum_i n_i (\bar{x}_i - \bar{\bar{x}})^2$	$k-1$	$\text{MSk} = \dfrac{Q_A}{k-1}$	$\dfrac{\text{MSk}}{\text{MSE}}$	$P(F > F_{k-1,N-k})$
Intra-Class	$Q_R = \sum_{ij} (x_{ij} - \bar{x}_i)^2$	$N-k$	$\text{MSE} = \dfrac{Q_R}{N-k}$		
Total	$Q_T = \sum_{ij} (x_{ij} - \bar{\bar{x}})^2$	$N-1$			

Table 14.3: **One way fixed factor ANOVA table**

But you might think: Why we go back on such an example ????

Well simply because we can very well build a linear regression model of the cash-flows of these two months, which we had not mentioned in the Statistics section! Thus, when we have an ANOVA with a controlled factor, nothing prevents us from making a linear regression of the data with binary explanatory variables (and vice versa)! To do this, it is enough that we rewrite the table above in the following form (technique named "dummy coding" or "one hot encoding"):

Observation	**Cash-Flow** y_i	**January** x_{i1}	**February** x_{i2}
1	304	1	0
2	284	1	0
3	290	1	0
4	310	1	0
5	320	1	0
6	270	1	0
7	309	1	0
8	293	1	0
9	315	1	0
10	301	1	0
11	280	0	1
12	303	0	1
13	294	0	1
14	270	0	1
15	276	0	1
16	310	0	1
17	290	0	1
18	312	0	1
19	260	0	1
20	325	0	1

Obviously we not always have in reality binary $(0,1)$ explanatory values but we can always normalize the control variables to fall back on such a situation!

This necessarily can be summarized to:

n_i	Mean cash flow $\bar{y}_i$	January x_{i1}	February x_{i2}
10	299.11	1	0
10	292.00	0	1

Therefore the regression model associated with this ANOVA with 1 control factor at 2 levels can therefore be written:

$$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_{i1}+\hat{\beta}_2x_{i2}+\varepsilon_i=\hat{\beta}_0+\sum_{j=1}^{2}x_{ij}\hat{\beta}_j+\varepsilon_i \tag{14.1.203}$$

Remark

Notice that this last relation can also be written:

$$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_{i1}+\hat{\beta}_2x_{i2}+\varepsilon_i=\underbrace{\hat{\mu}}_{\hat{\beta}_0}+\underbrace{\hat{\mu}_i}_{\hat{\beta}_1x_{i1}+\hat{\beta}_2x_{i2}}+\varepsilon_i \tag{14.1.204}$$

where in linear regression with dummy variables $\hat{\mu}$ is the "intercept" and $\hat{\mu}_i$ is named the "marginal effect".

And here maybe you recognize something well known to us when we have studied all the ANOVAs!!! So as we then see, checking that the means $\hat{\mu}_i$ are all equal or not in the ANOVA such that $\mu_1=\mu_2=\ldots=\mu$ is in the case of the regression equivalent to check that $\beta_1=\beta_2=0$. So under this form, the hypotheses are then:

$$\begin{aligned}&H_0:\ \beta_1=\beta_2=\ldots=\beta_k\\&H_1:\ \text{at least there is one }\beta_i\neq 0\end{aligned} \tag{14.1.205}$$

This is what we named the "F-test of Overall Significance in Regression Analysis" (unlike T-tests that can assess only one regression coefficient at a time as we will see further below page 6395).

But, with the preceding relation and the table summarized above, we have a system of two equations with three unknowns ... which is obviously insoluble for a least squares approach. Therefore, the idea consists in sacrificing one of the explanatory variables as (generalizable to more than two variables obviously):

$$x_{i1}+x_{i2}=1 \tag{14.1.206}$$

and therefore (special choice):

$$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_{i1}+\hat{\beta}_2(1-x_{i1})+\varepsilon_i=\hat{\beta}_0+\hat{\beta}_2+(\hat{\beta}_1+\hat{\beta}_2)x_{i1}+\varepsilon_i=\hat{\beta}'_0+\hat{\beta}'_1x_{i1}+\varepsilon_i \tag{14.1.207}$$

and then we have indeed two equations with two unknowns:

$$\begin{aligned}\hat{y}_i&=\hat{\beta}'_0+\hat{\beta}'_1x_{i1}+\varepsilon_i\\\bar{y}_i&=\hat{\beta}'_0+\hat{\beta}'_1\bar{x}_{i1}+\varepsilon_i=\hat{\beta}'_0+\frac{1}{2}\hat{\beta}'_1+\varepsilon_i\end{aligned} \tag{14.1.208}$$

This explains the reason why statistical softwares will always give the coefficient of one of the two binary explanatory variables as zero (which obviously can be problematic in some cases and therefore it is enough to force the ordinate at the origin to be zero to have the two non-zero coefficients since then we fall back on a system of two equations with two unknowns). In the case of k binary (dichotomous) explanatory variables, there will be $k-1$ whose coefficients are non-zero (since one can always be explained by all others).

Remark

The use of ANOVA, you probably guess, is only feasible if the residues are almost normally distributed and homoscedastic... This is why statistical softwares have outputs giving such analysis when running a F-test for regression.

Finally, all this to say that the ANOVA is only a special case (with binary explanatory variables rather than continuous one and let us recall that $\mathbb{N}\in\mathbb{R}$ not the inverse!) of the linear regression.

Let us notice then that we can write:

$$\sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2=\sum_{i=1}^{N}\left(\left(\sum_{j=1}^{k}x_{ij}\hat{\beta}_j+\hat{\beta}_0\right)-\bar{y}\right)^2 \tag{14.1.209}$$

and as they are either 0 or 1 (or normalized to be such as). Then there remain only the terms $x_{ij}\hat{\beta}_j$ where $x_{ij}\neq 0$ in a quantity that for each j we will denote by n_j. It comes then since we only keep the $x_{ij}=1$ (for the last equality if it is not obvious the reader can make an example on a sheet of paper with a simple numerical application!):

$$\sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2=\sum_{i=1}^{N}\left(\left(\sum_{j=1}^{k}x_{ij}\hat{\beta}_j+\hat{\beta}_0\right)-\bar{y}\right)^2=\sum_{j=1}^{k}n_j\left(\hat{\beta}_j+\hat{\beta}_0-\bar{y}\right)^2 \tag{14.1.210}$$

with obviously[8] $\sum_{j=1}^{k}n_j=N$.

By the method of the ordinary least squares and the values that can take the x_{ij} we quickly see that the coefficients (parameters) of the regression will be given by (we can detail again on reader request):

$$\hat{\beta}_j+\hat{\beta}_0=\bar{y}_j \tag{14.1.211}$$

and therefore:

$$\sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2=\sum_{j=1}^{k}n_j(\bar{y}_j-\bar{y})^2 \tag{14.1.212}$$

and as in our case in comparison with the one fixed factor ANOVA we have the obvious correspondences:

$$\bar{\bar{y}}\rightarrow\bar{y}=\frac{1}{N}\sum_{i=1}^{N}y_i \tag{14.1.213}$$

---

[8] In our above example $N=20$ and $n_1=n_2=10$

it follows, therefore, that there is a correspondence between the numerator of the Fisher test of ANOVA and the equivalent expression of the linear regression (and therefore with the same number of degrees of freedom) that brings us to write:

$$F_{k-1,N-k}=\frac{\dfrac{\displaystyle\sum_{j=1}^{k}n_j(\bar{y}_i-\bar{\bar{y}})^2}{k-1}}{\dfrac{\displaystyle\sum_{ij}(y_{ij}-\bar{y}_j)^2}{N-k}}=\frac{\dfrac{\displaystyle\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{k-1}}{\dfrac{\displaystyle\sum_{ij}(y_{ij}-\bar{y}_j)^2}{N-k}} \tag{14.1.214}$$

But it remains for us to find the equivalent also of the denominator for the regression. For this we will proceed by similarity. Let us recall that in the one fixed factor ANOVA we have proved that:

$$\begin{array}{ccccc}Q_T & = & Q_A & + & Q_R\\ \displaystyle\sum_i(y_{ij}-\bar{\bar{y}})^2 & = & \displaystyle\sum_i n_i(\bar{y}_i-\bar{\bar{y}})^2 & + & \displaystyle\sum_{ij}(y_{ij}-\bar{y}_i)^2\\ N-1 & = & k-1 & + & \displaystyle\sum_i(n_i-1)\end{array}$$

and let us recall that for the linear regression, we have proved earlier above (with the corresponding notations in usage):

$$\begin{array}{ccccc}\text{SST} & = & \text{SSE} & + & \text{SSR}\\ \displaystyle\sum_i(y_i-\bar{y})^2 & = & \displaystyle\sum_i(\hat{y}_i-\bar{y})^2 & + & \displaystyle\sum_i(y_i-\hat{y}_i)^2\end{array}$$

From what we have seen above we know that degrees of freedom of $\sum_i(\hat{y}_i-\bar{y})^2$ are $k-1$. If follows immediately that for $\sum_i(y_i-\bar{y})^2$ the degrees of freedom are $N-1$. Then we can write:

$$\begin{array}{ccccc}\text{SST} & = & \text{SSE} & + & \text{SSR}\\ \displaystyle\sum_i(y_i-\bar{y})^2 & = & \displaystyle\sum_i(\hat{y}_i-\bar{y})^2 & + & \displaystyle\sum_i(y_i-\hat{y}_i)^2\\ N-1 & = & k-1 & + & ????\end{array}$$

It then follows that the degrees of freedom of the sum of the squares residuals is then of $N-k$ such that:

$$\begin{array}{ccccc}\text{SST} & = & \text{SSE} & + & \text{SSR}\\ \displaystyle\sum_i(y_i-\bar{y})^2 & = & \displaystyle\sum_i(\hat{y}_i-\bar{y})^2 & + & \displaystyle\sum_i(y_i-\hat{y}_i)^2\\ N-1 & = & k-1 & + & N-k\end{array}$$

Sum of the squares of residues that it is customary to write (useful for later !!!):

$$\boxed{\sigma_{\varepsilon}^2=\frac{\sum\limits_i (y_i-\hat{y}_i)^2}{N-k}} \tag{14.1.215}$$

and that is given in the output of some statistical softwares (sadly!) under the misnamed "residual standard error" (instead than "residual standard deviation"):

$$\sigma_{\varepsilon}=\sqrt{\frac{\sum\limits_i (y_i-\hat{y}_i)^2}{N-k}} \tag{14.1.216}$$

As Q_T doesn't appear in the Fisher test there is no a priori reason that the equivalent which is SST for regression appears there. By elimination, the correspondence is then immediate:

$$\sum_{ij}(y_{ij}-\bar{y}_i)^2 \rightarrow \sum_i (y_i-\hat{y}_i)^2 \tag{14.1.217}$$

The Fisher test then becomes for the linear regression:

$$\boxed{F_{k-1,N-k}=\frac{\dfrac{\sum\limits_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{k-1}}{\dfrac{\sum\limits_{ij}(y_i-\hat{y})^2}{N-k}}} \tag{14.1.218}$$

and is more famously known under the name "omnibus test for OLS regression" (i.e. that a significant percentage of statistical softwares give below the diagnostic table of the regression analysis!).

It follows then that as for the one-fixed factor ANOVA, with linear regression, we can also make a table of the ANOVA (see section Statistics page 1001) as we will see later with an example!

Now, let us prove a common and important form of this last relation. We have proved earlier above that the correlation coefficient could also be written in the form:

$$R^2=\frac{\text{SSE}}{\text{SST}}=\frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\bar{y})^2} \tag{14.1.219}$$

However, let us notice that:

$$\begin{aligned}F_{k-1,N-k}&=\frac{\dfrac{\sum\limits_i(\hat{y}_i-\bar{y})^2}{k-1}}{\dfrac{\sum\limits_{ij}(y_i-\hat{y})^2}{N-k}}=\frac{\sum\limits_i(\hat{y}_i-\bar{y})^2}{\sum\limits_{ij}(y_i-\hat{y})^2}\frac{\sum\limits_i(y_i-\bar{y})^2}{\sum\limits_i(y_i-\bar{y})^2}\frac{N-k}{k-1}\\&=\frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\bar{y})^2}\frac{\sum(y_i-\bar{y})^2}{\sum(y_i-\hat{y}_k)^2}\frac{N-k}{k-1}=R^2\frac{\sum(y_i-\bar{y})^2}{\sum(y_i-\hat{y}_k)^2}\frac{N-k}{k-1}\end{aligned} \tag{14.1.220}$$

And also let us recall once again that:

$$\text{SST} = \text{SSE} + \text{SSR} \tag{14.1.221}$$

Thus explicitly as we have just seen:

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2 \tag{14.1.222}$$

Therefore we can write:

$$\begin{aligned} F_{k-1,N-k} &= R^2 \frac{\sum (y_i - \bar{y})^2}{\sum (y_i - \hat{y}_k)^2} \frac{N-k}{k-1} = \frac{\sum (y_i - \bar{y})^2}{\sum (y_i - \bar{y})^2 - \sum (\hat{y}_i - \bar{y})^2} \frac{N-k}{k-1} \\ &= R^2 \left(\frac{1}{\dfrac{\sum (y_i - \bar{y})^2 - \sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}} \right) \frac{N-k}{k-1} = R^2 \left(\frac{1}{1 - \dfrac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}} \right) \frac{N-k}{k-1} \\ &= R^2 \left(\frac{1}{1-R^2} \right) \frac{N-k}{k-1} \end{aligned} \tag{14.1.223}$$

What we usually can find in some textbooks in the following form:

$$\boxed{F_{k-1,N-k} = \frac{\dfrac{R^2}{k-1}}{\dfrac{1-R^2}{N-k}}} \tag{14.1.224}$$

We thus see that if the coefficient of determination R^2 is large (close to 1) then the value of F is large, the linear model will be considered as explaining significantly well the variable explained with respect to the explanatory variable (but this still doesn't mean that there is causality!).

Let us come back to our companion example! A simple ordinary least squares regression with a spreadsheet software like Microsoft® Excel® of the below table:

Observation	Cash-Flow y_i	January x_{i1}
1	304	1
2	284	1
3	290	1
4	310	1
5	320	1
6	270	1
7	309	1
8	293	1
9	315	1
10	301	1
11	280	0
12	303	0
13	294	0
14	270	0
15	276	0
16	310	0
17	290	0
18	312	0
19	260	0
20	325	0

gives (we make the choice not to force the ordinate at the origin and to take January as an explanatory variable with coefficient not zero otherwise we will not find the value of the calculation of the classic ANOVA seen earlier above):

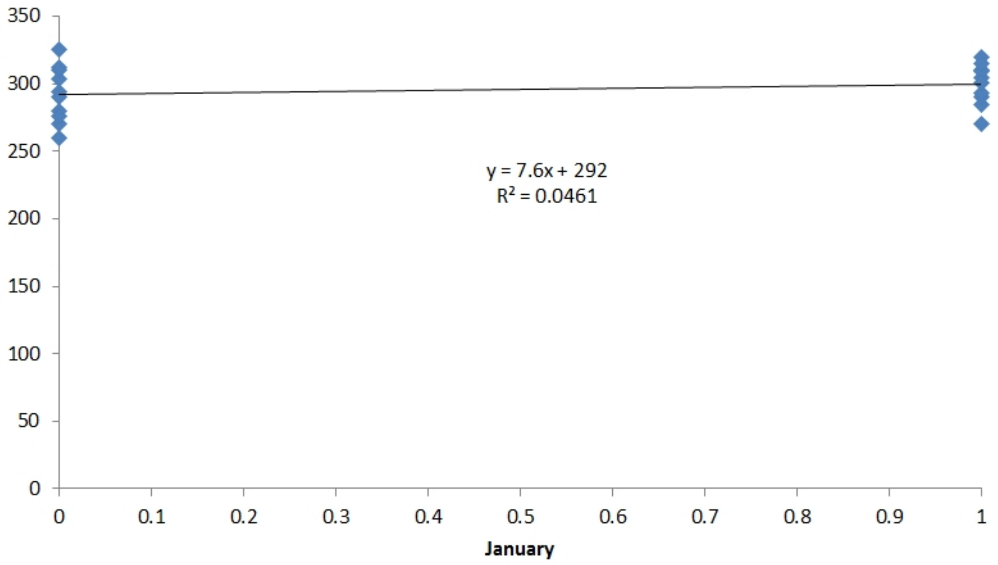


Figure 14.9: **Graphical representation of an univariate regression with the ANOVA approach in Microsoft® Excel® 14.0.7177**

that is to say:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} \cong 292 + 7.6_{i,\text{january}} \tag{14.1.225}$$

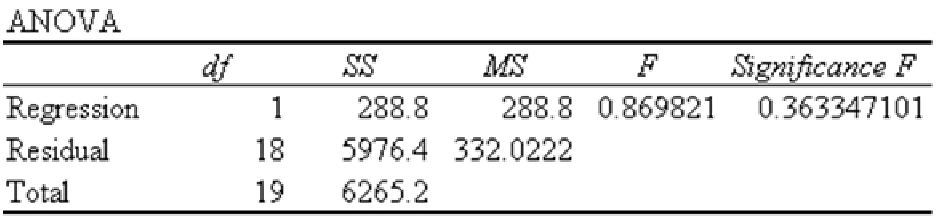
We then have:

Observation	**Cash flow y_i**	**Cash flow $\hat{y}_i$**	**SSR $(y_i-\hat{y}_i)^2$**	**SST $(y_i-\bar{y})^2$**	**SSE $(\hat{y}_i-\bar{y})^2$**
1	304	299.6	19.36	67.24	14.44
2	284	299.6	243.36	139.24	14.44
3	290	299.6	92.16	33.64	14.44
4	310	299.6	108.16	201.64	14.44
5	320	299.6	416.16	585.64	14.44
6	270	299.6	876.16	665.64	14.44
7	309	299.6	88.36	174.24	14.44
8	293	299.6	43.56	7.84	14.44
9	315	299.6	237.16	368.64	14.44
10	301	299.6	1.96	27.04	14.44
11	280	292	144	249.64	14.44
12	303	292	121	51.84	14.44
13	294	292	4	3.24	14.44
14	270	292	484	665.64	14.44
15	276	292	256	392.04	14.44
16	310	292	324	201.64	14.44
17	290	292	4	33.64	14.44
18	312	292	400	262.44	14.44
19	260	292	1024	1281.64	14.44
20	325	292	1048	852.64	14.44
		Total	**5976.4**	**6265.2**	**288.8**

with obviously:

$$k=2 \qquad N=20 \qquad N-k=18 \tag{14.1.226}$$

The calculations of all the classical ANOVA terms of the regression then give still with the same version of Microsoft® Excel® (the reader can verify by hand using the relations proved earlier above that we find the values given by this spreadsheet software):



ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	288.8	288.8	0.869821	0.363347101
Residual	18	5976.4	332.0222		
Total	19	6265.2			

Or more explicitly:

Source	**Sum of squares (SSE)**	χ^2 **df**	**Mean squares**	F	**Critical** F
Inter-Class	$Q_A=\sum_i(\hat{y}_i-\bar{y})^2$	$k-1$	$\text{MSk}=\dfrac{Q_A}{k-1}$	$\dfrac{\text{MSk}}{\text{MSE}}$	$P(F>F_{k-1,N-k})$
Intra-Class	$Q_R=\sum_i(y_i-\hat{y}_i)^2$	$N-k$	$\text{MSE}=\dfrac{Q_R}{N-k}$		
Total	$Q_T=\sum_i(y_i-\bar{y})^2$	$N-1$			

Table 14.4: **ANOVA table for regression**

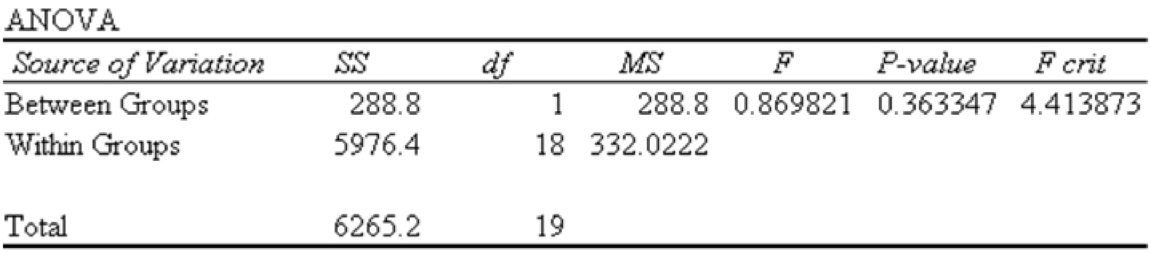
Remark

Notice that in the above ANOVA table we recognize something known to us:

$$Q_R=\sum_i(y_i-\hat{y}_i)^2=\sigma_\varepsilon^2 \tag{14.1.227}$$

That is... the SEE^2, where for recall SEE is named the "standard error of estimate" or "standard regression error" (see page 6355)! This is why in all statistical softwares you can simply rely the standard error of estimate of the regression with the intra-class mean sum of square (i.e. mean residuals errors) just by taking the square root of that latter such that: $\text{SEE}=\sqrt{\text{MSE}}$.

And running the one-fixed factor ANOVA we had already introduced a little earlier above we have for comparison:



ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	288.8	1	288.8	0.869821	0.363347	4.413873
Within Groups	5976.4	18	332.0222			
Total	6265.2	19				

With for recall (again!) its general table:

Source	**Sum of squares (SSE)**	χ^2 **df**	**Mean squares**	F	**Critical** F
Inter-Class	$Q_A=\sum_i n_i(\bar{x}_i-\bar{\bar{x}})^2$	$k-1$	$\text{MSk}=\dfrac{Q_A}{k-1}$	$\dfrac{\text{MSk}}{\text{MSE}}$	$P(F>F_{k-1,N-k})$
Intra-Class	$Q_R=\sum_{ij}(x_{ij}-\bar{x}_i)^2$	$N-k$	$\text{MSE}=\dfrac{Q_R}{N-k}$		
Total	$Q_T=\sum_{ij}(x_{ij}-\bar{\bar{x}})^2$	$N-1$			

Table 14.5: **One way fixed factor ANOVA table**

We see then the obvious similarity that there are with the two approaches!!! This is why some people say sometimes: « *running a regression or an ANOVA are two equivalent thinks (assumed under some given conditions)* » as in the ANOVA, the categorical variable is effect coded, which means that each category's mean is compared to the grand mean. In the regression, the categorical variable is dummy coded, which means that each category intercept is compared to the reference group intercept! Since the intercept is defined as the mean value when all other predictors $=0$, and there are no other predictors, the three intercepts are just means.

Indeed, we have:

$$\hat{\beta}_0 = \bar{y}_{\text{february}} = 292 \qquad (14.1.228)$$

and:

$$\hat{\beta}_1 + \hat{\beta}_0 = \bar{y}_{\text{january}} = 292 + 7.6 = 299.6 \qquad (14.1.229)$$

14.1.10.1.4.1 Test for Lack of Fit

The "test for lack-of-fit" compares the variation around the model with "pure" variation within replicated observations (some statistical softwares use this analysis only for the adequacy of quadratic response surface and not for simple linear models...). In particular, if there are n_i replicated observations $y_{i1},\ldots,y_{in_i}$ of the response all at the same values $\vec{x}_i$ (vector of dimension n) of the factors, then we can predict the true response at $\vec{x}_i$ either by using the predicted value $\hat{y}_i$ based on the model or by using the mean $\bar{y}_i$ of the replicated values. The test for lack-of-fit decomposes the residual error into a component due to the variation of the replications around their mean value (the "pure error"), and a component due to the variation of the mean values around the model prediction (the "lack of fit" also named the "bias error"):

$$\begin{aligned}
Q_R &= \sum_{i=1}^{n}\sum_{j=1}^{n_i}\varepsilon_{ij}^2 \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(y_{ij}-\hat{y}_i\right)^2 = \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{i.}\right)^2}_{\text{sum of squares due to pure error}} + \underbrace{\sum_{i=1}^{n}\sum_{i=1}^{n_i}\left(\bar{y}_{i.}-\hat{y}_{ij}\right)^2}_{\text{sum of squares due to lack of fit (bias error)}} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{i.}\right)^2 + \sum_{i=1}^{n} n_i\left(\bar{y}_{i.}-\hat{y}_i\right)^2
\end{aligned} \qquad (14.1.230)$$

or written differently:

$$Q_R = \text{SSE} = \text{SSPE} + \text{SSLF} \qquad (14.1.231)$$

Remarks

R1. The reader should easily notice that without repeated measurements, $\text{SSPE}=0$ and hence we cannot conduct a lack-of-fit analysis!

R2. Obviously for one x_i we have multiple y_i, hence the notation y_{ij}. But the model for a given x_i gives only one and only one $\hat{y}_i$. Hence $\forall j$ and a given i, we have $\hat{y}_{ij}=\hat{y}_i$.

If the model is adequate, then both components estimate the nominal level of error; however, if the bias component of error is much larger than the pure error, then this constitutes evidence that there is significant lack of fit. The concept can be illustrated as following:

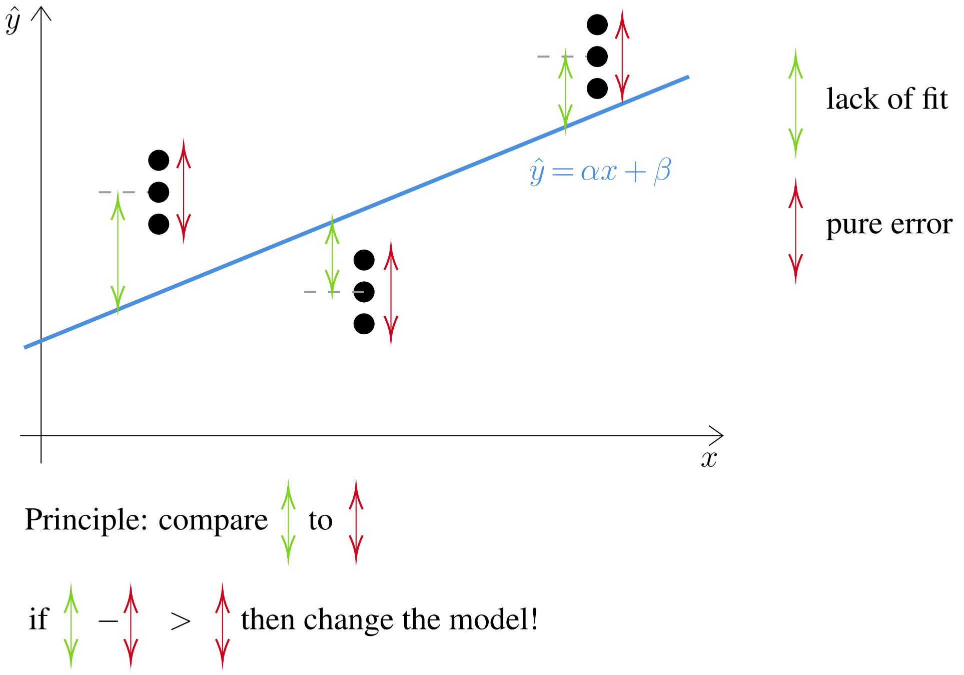


Figure 14.10: **Lack of fit and pure error**

- As before, the degrees of freedom associated with SSE (i.e. Q_R) is as we know very well given by $n(k-1) = N-k$ (keep in mind that the k comes from the fact that you estimate k parameters - the slopes plus the intercept - whenever you fit a line to a set of data).
- The degrees of freedom associated with SSLF is $n-k$, where c denotes the number of distinct x values we have.
- The degrees of freedom associated with SSPE is $nk-n = N-n$, where again n denotes the number of distinct x values we have.

Therefore the degrees of freedom breakdown as:

$$\underbrace{\text{SSE}}_{\text{df}=N-k} = \underbrace{\text{SSPE}}_{N-n} + \underbrace{\text{SSLF}}_{n-k} \tag{14.1.232}$$

It then follows that the statistic:

$$F = \frac{\text{lack-of-fit sum of squares}}{\text{degrees of freedom}} = \frac{\text{MSLF}}{\text{MSPE}} = \frac{\dfrac{\text{SSLF}}{n-k}}{\dfrac{\text{SSPE}}{N-n}} = \frac{\dfrac{\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{i.}\right)^2}{n-k}}{\dfrac{\displaystyle\sum_{i=1}^{n} n_i\left(\bar{y}_{i.}-\hat{y}_i\right)^2}{N-n}} \tag{14.1.233}$$

has an F-distribution with the corresponding number of degrees of freedom in the numerator and the denominator, provided that the model is correct.

To understand this test keep in mind that under the null hypothesis:

- The sum of squares due to pure error, divided by the error variance σ_{ε}^2, has a chi-squared distribution with $N-n$ degrees of freedom.
- The sum of squares due to lack of fit, divided by the error variance σ_{ε}^2, has a chi-squared distribution with $n-k$ degrees of freedom.
- The two sums of squares are probabilistically independent.

If the model is wrong, then the probability distribution of the denominator is still as stated above, and the numerator and denominator are still independent. But the numerator then has a noncentral chi-squared distribution (see section Statistics page 780) and consequently the quotient as a whole has a noncentral F-distribution.

One uses this F-statistic to test the null hypothesis that there is no lack of linear fit. Since the non-central F-distribution is stochastically larger than the (central) F-distribution, one rejects the null hypothesis if the F-statistic is larger than the critical F value.

So we complete the previous analysis of variance table using the following relations:

Source	χ^2 **df**	**SS**	**Means squares**	***F***
Regression	$k-1$	$\text{SSR}=\sum\limits_{i=1}^{n} n_i(\hat{y}_i-\bar{y})^2$	$\text{MSR}=\dfrac{\text{SSR}}{k-1}$	$F=\dfrac{\text{MSR}}{\text{MSE}}$
Residual error	$N-k$	$\text{SSE}=\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n_i}(y_{ij}-\hat{y}_i)^2$	$\text{MSE}=\dfrac{\text{SSE}}{N-k}$	
Lack of Fit	$n-k$	$\text{SSLF}=\sum\limits_{i=1}^{n} n_i(\bar{y}_i-\hat{y}_i)^2$	$\text{MSLF}=\dfrac{\text{SSLF}}{n-k}$	$F=\dfrac{\text{MSLF}}{\text{MSPE}}$
Pure error	$N-n$	$\text{SSPE}=\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$	$\text{MSPE}=\dfrac{\text{SSPE}}{N-k}$	
Total	$N-1$	$\text{SSTO}=\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n_i}(y_{ij}-\bar{y})^2$		

Table 14.6: **Lack-of-Fit ANOVA table**

The reader must be careful with the "Total" in the above table (last row). Indeed, the total must not take into account the row with the degrees of freedom $n-k$ and $N-n$ as their are the decomposition of the row with the $N-k$.

14.1.10.2 Univariate linear regression Gaussian Model

We will assume that for an individual k picked randomly from the population, x_k is known without error, and that y_k is a realization of a random variable that we will now denote Y_k and the theoretical least squares regression line will be written now under the general form of a "stochastic linear model":

$$Y_k = \alpha x_k + \beta + \varepsilon_k \tag{14.1.234}$$

where ε_k is assumed identically distributed and independent residue (no correlation) for each point k according to a centered Normal distribution (zero mean and standard deviation σ for all k) such as $\forall i \neq j$:

$$\boxed{\begin{aligned} \varepsilon_k &= \mathcal{N}(0,\sigma) \\ \text{cov}(\varepsilon_i,\varepsilon_j) &= 0 \end{aligned}} \tag{14.1.235}$$

So in other words[9]:

$$\begin{aligned} \text{E}(\varepsilon_k) &= 0 \\ \text{V}(\varepsilon_k) &= \sigma_\varepsilon^2 \end{aligned} \tag{14.1.236}$$

where the residue is defined by the difference between the theoretical ordinates Y_k (considered as random variable) and the measured (experimental) ordinates y_k:

$$\varepsilon_k = Y_k - y_k \tag{14.1.237}$$

and since by hypothesis $\varepsilon_k = \mathcal{N}(0,\sigma)$, it immediately comes by the stability of the Normal law (see section Statistics page 710) that:

$$Y_k = \mathcal{N}(y_k,\sigma_\varepsilon) = \mathcal{N}(\alpha x_k + \beta, \sigma_\varepsilon) \tag{14.1.238}$$

This is why this specific model is named "Gaussian linear model" .... Explicitly, we have:

$$Y_k = \frac{1}{\sigma_\varepsilon\sqrt{2\pi}} e^{\frac{1}{2}\left(\frac{\alpha x_k + \beta - \mu_k}{\sigma_\varepsilon}\right)^2} \tag{14.1.239}$$

This is why theoretical model is formally denoted:

$$y = \text{E}\left[Y|X=x\right] = f(x) \tag{14.1.240}$$

We will choose the symbol $\sim$ to say "follows the law ..." in order to avoid any notation confusion:

$$Y_k \sim \frac{1}{\sigma_\varepsilon\sqrt{2\pi}} e^{\frac{1}{2\sigma_\varepsilon^2}(\alpha x_k + \beta - \mu_k)^2} = \frac{1}{\sigma_\varepsilon\sqrt{2\pi}} e^{\frac{1}{2\sigma_\varepsilon^2}(Y_k - \mu_k)^2} \tag{14.1.241}$$

Which graphically is equivalent to have (so normally the statistical analysis of the Gaussian regression occurs only if and only if we have taken several measures of the dependent variable for fixed and identical values of the independent variables!!!!!!!!):

---

[9] The conditions $\text{cov}(\varepsilon_i,\varepsilon_j) = 0$, and $\text{E}(\varepsilon_k) = 0$ and $\text{V}(\varepsilon_k) = \sigma_\varepsilon^2$ are named the "Gauss-Markov assumptions" (see page 6405) and are at the basis of the "Gauss-Markov Linear Model" (GMLM).

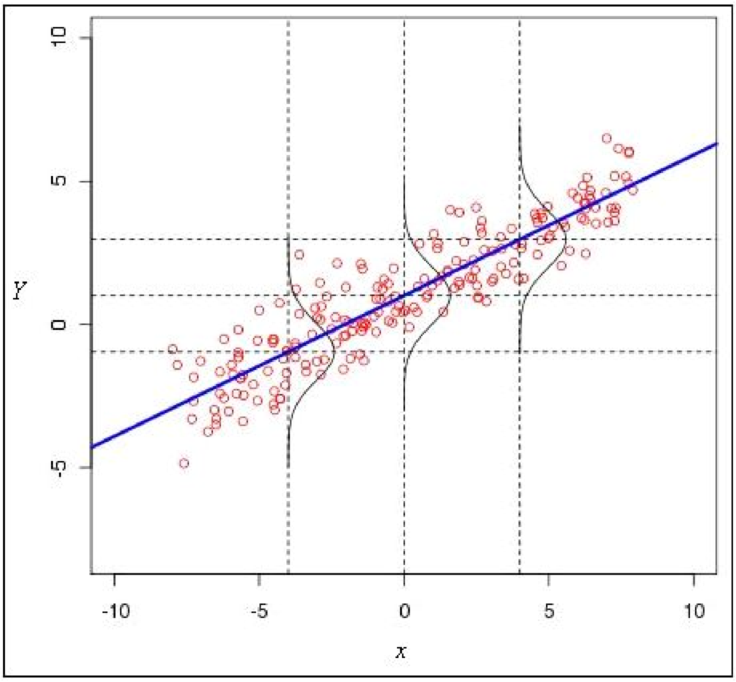


Figure 14.11: Graphical representation of the idea behind the Gaussian linear regression

Caution! Since the model is Gaussian, the variable to explain has its domain of definition which is unbounded (support of the Normal distribution). Some companies (especially auditing firms...) wish sometimes to create a simple linear model to model a probability (that for reminder is bounded in [0,1]) typically for the probability of bankruptcy/default based on various factors (explanatory variables). Therefore the measured probability must first be turn into Z values (quantiles) of the Normal distribution to have again an infinite range. This type of approach is then named "linear Z-score model".

Almost all statistical analysis software provide a pattern (figure) of residues according to the x values. Thus, these type of figures help to accept or reject the use of a linear Gaussian model:

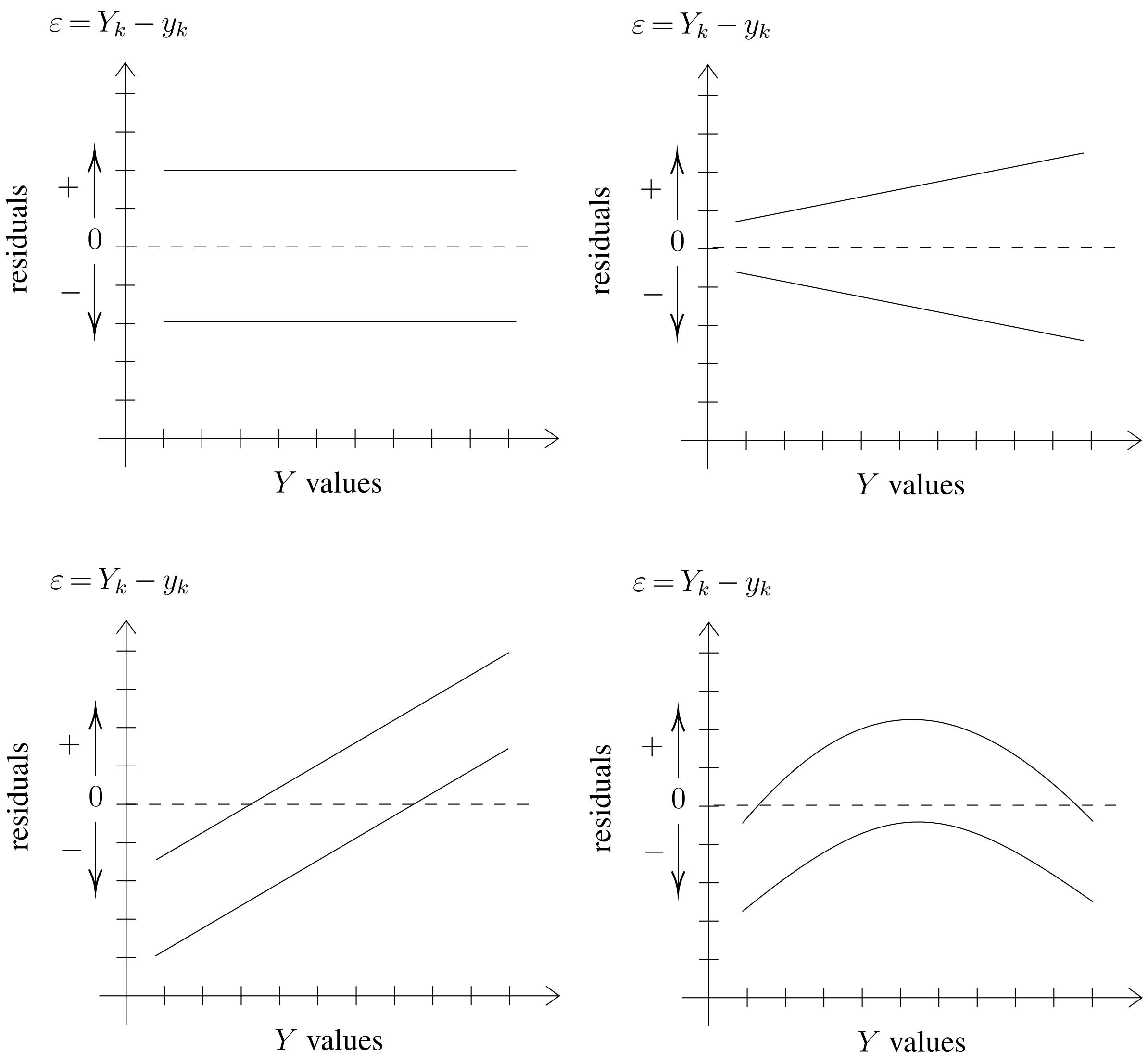


Figure 14.12: Examples of "plot" of residuals

In the above figure, the graph on the top left is what we should expect to have to be able to apply the statistical tests for the Gaussian linear model. The graph on the top right shows that the residuals variance is not constant and therefore we have a violation of the assumption of homoscedasticity (typically checked with statistical hypothesis tests like Levene, Bartlett, Fligner, Brown-Forsythe, etc.)! The graph at the bottom left shows that the variance is constant but that our model has missing endogenous variables that added could perhaps explain the shift that grows up linearly. The graph at the bottom right indicates a constant variance, but the model looks like to be more nonlinear than linear.

Previous assumptions about the moments of residues (mean, variance) are named "Gauss-Markov assumptions" and the particular hypothesis of equal variances is named as we saw it in the section Statistics "homoscedasticity" (while the fact that the variances are not equal is named for reminder "heteroskedasticity").

Remarks

R1. The majority of softwares (including Microsoft® Excel® 11.8346) propose a graph which shows the residuals in function of the ordinate values x. Obviously, it is better that the points representing the residues are not too divergent ... otherwise the homoscedasticity assumption will be not satisfied.

R2. Practitioners sometimes transform the endogenous and exogenous variables by a logarithmic or exponential function or other to try to stabilize the residuals as most as possible.

Theorem 14.2. We have through the property of the mean (see section Statistics page 571):

$$\bar{Y}=\mathrm{E}(Y_k)=\mathrm{E}(\alpha x_k+\beta+\varepsilon_k)=\mathrm{E}(\alpha x_k)+\mathrm{E}(\beta)+\mathrm{E}(\varepsilon_k)=\alpha\bar{x}+\beta \tag{14.1.242}$$

Then under the above assumptions, we will show that a and b are unbiased estimators (see section Statistics page 821) of α and β and it is possible to estimate the standard deviation from the SSR which is an important result named "Gauss-Markov theorem" (see page 6405).

Before seeing the proof let us do a recall and give some definitions of the variables that we have already handled and the new one that we will handle (if the vocabulary seems technical to the reader then it should read or reread the sections of Probabilities and Statistics):

Variable	Description
a	Slope (director coefficient) of the linear model of the least squares method (LSM). This is a punctual estimate value (deterministic).
b	Intercept of the linear model of the least squares method (LSM). This is a punctual estimate value (deterministic).
A	Random variable of the slope (director coefficient) according to the statistical Gaussian model approach and for which a is a realization. The expected mean of A being an unbiased estimator of α.
B	Random variable of the ordinate at the origin according to the statistical approach of the linear Gaussian model and for which b is a realization. The expected mean of B being an unbiased estimator of β.
α	Unbiased expected mean of the variable A representing the slope (director coefficient) within the framework of the statistical approach of the Gaussian linear model.
β	Unbiased expected mean of the variable B representing the y-intercept (ordinate at origin) in the framework of the statistical approach of linear Gaussian model.

Table 14.7: **Reminder of notations for the study of the linear regression**

Proof 14.2.1. According to the adopted model, a must now be regarded as a realization of the random variable given by (shown above as the ratio of the covariance and variance):

$$A=\frac{\sum\limits_k (x_k-\bar{x})(Y_k-\bar{Y})}{\sum\limits_k (x_k-\bar{x})^2} \tag{14.1.243}$$

and b as a realization of the random variable given by:

$$B=\bar{Y}-A\bar{x} \tag{14.1.244}$$

So we differentiate random and non-random coefficient values by passing the lowercase notation in uppercase (as it is the tradition in the field of Statistics).

Taking into account that theoretical dependent (endogenous) variable is considered as the realization of a random variable is therefore given by:

$$Y_k=\alpha x_k+\beta+\varepsilon_k \tag{14.1.245}$$

we can put A in the form:

$$\begin{aligned}
A&=\frac{\sum\limits_k (x_k-\bar{x})(Y_k-\bar{Y})}{\sum\limits_k (x_k-\bar{x})^2}=\frac{\sum\limits_k (x_k-\bar{x})(\alpha x_k+\beta+\varepsilon_k-\bar{Y})}{\sum\limits_k (x_k-\bar{x})^2}\\
&=\frac{\sum\limits_k (x_k-\bar{x})(\alpha x_k+\beta+\varepsilon_k-\mathrm{E}(Y))}{\sum\limits_k (x_k-\bar{x})^2}=\frac{\sum\limits_k (x_k-\bar{x})(\alpha x_k+\beta+\varepsilon_k-\alpha\bar{x}_k-\beta)}{\sum\limits_k (x_k-\bar{x})^2}\\
&=\frac{\sum\limits_k (x_k-\bar{x})(\alpha x_k+\varepsilon_k-\alpha\bar{x}_k)}{\sum\limits_k (x_k-\bar{x})^2}=\frac{\alpha\sum\limits_k (x_k-\bar{x})(x_k-\bar{x})+\sum\limits_k (x_k-\bar{x})\varepsilon_k}{\sum\limits_k (x_k-\bar{x})^2}\\
&=\frac{\alpha\sum\limits_k (x_k-\bar{x})^2}{\sum\limits_k (x_k-\bar{x})^2}+\frac{\sum\limits_k (x_k-\bar{x})\varepsilon_k}{\sum\limits_k (x_k-\bar{x})^2}=\alpha+\frac{\sum\limits_k (x_k-\bar{x})\varepsilon_k}{\sum\limits_k (x_k-\bar{x})^2}
\end{aligned} \tag{14.1.246}$$

Therefore:

$$\begin{aligned}
\mathrm{E}(A)&=\mathrm{E}\left(\alpha+\frac{\sum\limits_k (x_k-\bar{x})\varepsilon_k}{\sum\limits_k (x_k-\bar{x})^2}\right)=\mathrm{E}(\alpha)+\mathrm{E}\left(\frac{\sum\limits_k (x_k-\bar{x})\varepsilon_k}{\sum\limits_k (x_k-\bar{x})^2}\right)\\
&=\alpha+\frac{\sum\limits_k (x_k-\bar{x})\mathrm{E}(\varepsilon_k)}{\sum\limits_k (x_k-\bar{x})^2}=\alpha
\end{aligned} \tag{14.1.247}$$

And for B:

$$B=\bar{Y}-A\bar{x}=\mathrm{E}(Y)-A\bar{x}=\alpha\bar{x}+\beta-A\bar{x}=\beta-\bar{x}(A-\alpha) \tag{14.1.248}$$

Therefore and using also the definitions:

$$\begin{aligned}\mathrm{E}(B)&=\mathrm{E}(\beta-\bar{x}(A-\alpha))=\mathrm{E}(\beta)-\mathrm{E}(\bar{x}(A-\alpha))=\beta-\bar{x}\mathrm{E}((A-\alpha))\\&=\beta-\bar{x}\mathrm{E}(A)+\bar{x}\mathrm{E}(\alpha)=\beta-\bar{x}\alpha+\bar{x}\alpha=\beta\end{aligned} \tag{14.1.249}$$

with the same conclusion.

■ Q.E.D.

Caution! Don't confuse the "error" with the "residual". Try to remember that:

- The "error" is the difference between the observed response and the true value of the response. The error is a mathematical construct, and it is UNKNOWN.
- The "residual" is the difference between the observed response and the fitted response. The residual is KNOWN, and it is an estimate of the error for each particular value of the response variable.

Formally:

$$\begin{aligned}\text{error}:\ \varepsilon_i&=Y_i-E\left(Y_i\right)=Y_i-(\beta_0+\beta_1 x_i)\\\text{residual}:\ e_i&=Y_i-\hat{Y}_i=Y_i-\left(\hat{\beta}_0+\hat{\beta}_1 x_i\right)\end{aligned} \tag{14.1.250}$$

So the expected mean of A and B are unbiased estimators (i.e.. with minimum variance as seen in the section of Statistics) of. As they are estimators, in the literature, they are often noted $\hat{a}$,$\hat{}$ and therefore it comes the common alternative notation:

$$\hat{Y}_k=\hat{a}x_k+b \tag{14.1.251}$$

Finally, we must also calculate the variances of A and B using its properties (see section Statistics page 596) and the assumptions on the residuals, we have:

$$\begin{aligned}\mathrm{V}(A)&=\mathrm{V}\left(\alpha+\dfrac{\sum\limits_k(x_k-\bar{x})\varepsilon_k}{\sum\limits_k(x_k-\bar{x})^2}\right)=\mathrm{V}\left(\dfrac{\sum\limits_k(x_k-\bar{x})\varepsilon_k}{\sum\limits_k(x_k-\bar{x})^2}\right)\\&=\dfrac{\mathrm{V}\left(\sum\limits_k(x_k-\bar{x})\varepsilon_k\right)}{\left[\sum\limits_k(x_k-\bar{x})^2\right]^2}=\dfrac{\sum\limits_k\mathrm{V}\left((x_k-\bar{x})\varepsilon_k\right)}{\left[\sum\limits_k(x_k-\bar{x})^2\right]^2}\\&=\dfrac{\sum\limits_k(x_k-\bar{x})^2\mathrm{V}\left(\varepsilon_k\right)}{\left[\sum\limits_k(x_k-\bar{x})^2\right]^2}\end{aligned} \tag{14.1.252}$$

As by assumption we have that all the $\mathrm{V}(\varepsilon_k)$ are equal and that there is no autocorrelation (what is typically tested with the Durbin-Watson test derived at page 8015), we can write:

$$\mathrm{V}(A)=\frac{\sum_k (x_k-\bar{x})^2\mathrm{V}(\varepsilon)}{\left[\sum_k (x_k-\bar{x})^2\right]^2}=\frac{\mathrm{V}(\varepsilon)\sum_k (x_k-\bar{x})^2}{\left[\sum_k (x_k-\bar{x})^2\right]^2}=\frac{\mathrm{V}(\varepsilon)}{\sum_k (x_k-\bar{x})^2} \qquad (14.1.253)$$

And if n is large enough we will write:

$$\sigma_A^2=\mathrm{V}(A)=\frac{\sigma_\varepsilon^2}{\sum_k (x_k-\bar{x})^2}=\frac{\sigma_\varepsilon^2}{n\mathrm{V}(x)} \qquad (14.1.254)$$

Before determining the variance of B, remember that by hypothesis:

$$Y_k=\mathcal{N}(\alpha x_k+\beta,\sigma_\varepsilon) \qquad (14.1.255)$$

therefore by the property of linearity of the Normal distribution, the random variables A and B also follow a Normal distribution.

After the recall of this assumption, it follows immediately (see section Statistics page 611):

$$\mathrm{V}(Y_k)=\sigma_\varepsilon^2 \Leftrightarrow \mathrm{V}(\bar{Y})=\frac{\sigma_\varepsilon^2}{n} \qquad (14.1.256)$$

Therefore:

$$\begin{aligned}\mathrm{V}(B)&=\mathrm{V}(\bar{Y}-A\bar{x})=\mathrm{V}(\bar{Y})+\bar{x}^2\mathrm{V}(A)-2\bar{x}\mathrm{cov}(\bar{Y},A)=\mathrm{V}(\bar{Y})+\bar{x}^2\mathrm{V}(A)\\&=\frac{\sigma_\varepsilon^2}{n}+\bar{x}^2\sigma_A^2=\frac{\sigma_\varepsilon^2}{n}+\bar{x}^2\frac{\sigma_\varepsilon^2}{\sum_k (x_k-\bar{x})^2}=\frac{\sigma_\varepsilon^2\left(\sum_k (x_k-\bar{x})^2+n\bar{x}^2\right)}{n\sum_k (x_k-\bar{x})^2}\end{aligned} \qquad (14.1.257)$$

Remember the Huygens theorem (see section Statistics page 593):

$$\sum_k (x_k-\bar{x})^2=\sum_k x_k^2-n\bar{x}^2 \qquad (14.1.258)$$

in the case of equitable probabilities (Normal Law estimator of the mean as prove in the section Statistics).

Finally we have:

$$\mathrm{V}(B)=\sigma_B^2=\frac{\sigma_\varepsilon^2\left(\sum_k (x_k-\bar{x})^2+n\bar{x}^2\right)}{n\sum_k (x_k-\bar{x})^2}=\frac{\sigma_\varepsilon^2\left(\sum_k x_k^2\right)}{n\sum_k (x_k-\bar{x})^2}=\frac{\sigma_\varepsilon^2\sum_k x_k^2}{n^2\mathrm{V}(x)} \qquad (14.1.259)$$

where obviously the notation of the variance in the denominator is very unfair (because x is not a random variable in this model) but very convenient to condense the notation.

The problem now lies in determining σ_ε^2. Obviously to do this we will be forced to go through a statistical estimator.

We know we can write according to what was seen in the section Statistics regarding to estimators:

$$\hat{\sigma}_\varepsilon^2=\frac{1}{n-k}\sum(\varepsilon_i-\bar{\varepsilon})^2=\frac{1}{n-2}\sum\varepsilon_i^2=\frac{1}{n-2}\sum(Y_k-y_k)^2 \tag{14.1.260}$$

because the Normal distribution is centered for residuals and therefore $\bar{\varepsilon}=0$... and the residue is implicitly dependent of two the sum of two random variable that are A and B this is why we have a -2 at the denominator of the first fraction (we have to take away two degrees of freedom).

Let us also indicate that in practice we frequently note the last result by mixing the notations of the random and deterministic appearance (hence noting everything with lower case):

$$\hat{\sigma}_\varepsilon^2=\frac{1}{n-2}\sum(y_k'-y_k)^2=\frac{1}{n-2}\sum e_k^2=\frac{\text{SCR}}{n-2}=\text{SEE}^2 \tag{14.1.261}$$

where SEE[10] means "Standard Error of Estimate" or sometimes also named the "standard regression error" which is obtained with the English version of Microsoft® Excel® trough square of the function `STEYX( )`.

Remark

This bring us obviously to consider that standard R squared estimator uses biased estimators of σ_ε^2 and $\text{V}(Y)$, by using the divisor n (or $n-1$) for both. These estimators are biased because they ignore the degrees of freedom used to estimate the regression coefficients and mean of Y. If instead, we estimate $\text{V}(Y)$ by the usual unbiased sample variance (which uses $n-1$ as the divisor) and σ_ε^2 by its unbiased estimator which uses $n-p-1$ (where p is the number of predictors) as the divisor, we obtain the "adjusted R-squared $R^2_{\text{adj.}}$" estimator:

$$\boxed{\hat{R}^2_{\text{adj.}}=1-\frac{\text{SSE}_{\text{adj.}}}{\text{SST}}=1-\frac{\frac{1}{n-p-1}\sum_{k=1}^n(Y_k-\hat{Y}_k)^2}{\frac{1}{n-1}\sum_{k=1}^n(Y_k-\bar{Y})^2}=1-\frac{(n-1)\text{MSE}}{\text{SST}}} \tag{14.1.262}$$

From this relation, we can see that the adjusted estimator is always less than the standard one. We can also see the larger p is relative to n, the larger the adjustment. A final point: although the adjusted R squared estimator uses unbiased estimators of the residual variance and the variance of Y, it is not unbiased! This is because, as we know it, the mean of a ratio is not generally equal to the ratio of the means.

Furthermore, notice, as $\hat{R}^2=1-\text{SSE}/\text{SST}$, we have:

$$\hat{R}^2_{\text{adj.}}=1-(1-\hat{R}^2)\frac{n-1}{n-p-1} \tag{14.1.263}$$

But for any value of $p\geq 1$:

$$\frac{n-1}{n-p-1}>1 \tag{14.1.264}$$

---

[10] Sadly sometimes (like in Minitab outputs) just denoted by the letter "S"...

If we multiply both side by $(1-\hat{R}^2)$, we get:

$$(1-\hat{R}^2)\frac{n-1}{n-p-1} > (1-\hat{R}^2) \tag{14.1.265}$$

Subtracting -1:

$$(1-\hat{R}^2)\frac{n-1}{n-p-1}-1 > (1-\hat{R}^2)-1 \tag{14.1.266}$$

multiplying by -1:

$$1-(1-\hat{R}^2)\frac{n-1}{n-p-1}-1 < \hat{R}^2 \tag{14.1.267}$$

Therefore:

$$\hat{R}^2_{\text{adj.}} < \hat{R}^2 \tag{14.1.268}$$

From this relation, we can see that the adjusted estimator is always less than the standard one!

So we have to summarize the unbiased estimators variances of A and B:

$$\boxed{\hat{\sigma}_A^2=\frac{\hat{\sigma}_\varepsilon^2}{n\text{V}(x)}=\frac{\dfrac{1}{n-2}\sum e_k^2}{n\text{V}(x)} \qquad \hat{\sigma}_B^2=\frac{\hat{\sigma}_\varepsilon^2\sum\limits_k x_k^2}{n^2\text{V}(x)}=\frac{\dfrac{1}{n-2}\sum e_k^2\sum\limits_k x_k^2}{n^2\text{V}(x)}} \tag{14.1.269}$$

Relations which are therefore only valid for a linear regression with one unique explanatory variable (and under the assumptions of the Gaussian linear model). Knowing that by construction of the initial hypothesis that A and B follow a Normal distribution of respective expected mean α,β and whose variance is given just above, so we know completely the distribution that characterizes them.

What is nice knowing these variances is that we can also therefore easily estimate the variance of the dependent variable in our regression (using the properties of the variance proved in the section Statistics at page 596).

It would be interesting to make statistical inference on the mean of the parameters A and B (i.e. the slope and intercept) given their known empirical mean (i.e. average). For this, remember that we have proved in the section Statistics (see page 865) the following confidence interval:

$$\bar{X}-\frac{S}{\sqrt{n}}T_{\alpha/2}(n-1)\leq\mu\leq\bar{X}+\frac{S}{\sqrt{n}}T_{\alpha/2}(n-1) \tag{14.1.270}$$

It follows by making a parallel like engineers and physicists like to do ... that as A is an unbiased estimator of the average of the slope a and that:

$$\hat{\sigma}_A=\sqrt{\frac{\dfrac{1}{n-2}\sum e_k^2}{n\text{V}(x)}} \tag{14.1.271}$$

is in fact the standard error of the mean A, then by analogy:

$$\begin{gathered}\mu \to a\\ \bar{X} \to A\\ \frac{S}{\sqrt{n}} \to \sqrt{\frac{\dfrac{1}{n-2}\sum e_k^2}{n\mathrm{V}(x)}}\\ T_{\alpha/2}(n-1) \to T_{\alpha/2}(n-2)\end{gathered} \tag{14.1.272}$$

and then (this is a reasoning to be take with caution an it is better to use the developments that will follow later):

$$A - T_{\alpha/2}(n-2)\sqrt{\frac{\dfrac{1}{n-2}\sum e_k^2}{n\mathrm{V}(x)}} \leq a \leq A + T_{\alpha/2}(n-2)\sqrt{\frac{\dfrac{1}{n-2}\sum e_k^2}{n\mathrm{V}(x)}} \tag{14.1.273}$$

which therefore gives the confidence interval of the slope of a linear Gaussian with one unique explanatory variable (that's what gives Microsoft® Excel® 11.8346 for each coefficient). The approach is the same for the intercept.

Warning! If the explanatory variable is a random variable, then we use naively (the are more accurate and correct model that we will see later when the explanatory variable is a also a random variable):

$$A - T_{\alpha/2}(n-2)\sqrt{\frac{\dfrac{1}{n-2}\sum e_k^2}{n\hat{\mathrm{V}}(x)}} \leq a \leq A + T_{\alpha/2}(n-2)\sqrt{\frac{\dfrac{1}{n-2}\sum e_k^2}{n\hat{\mathrm{V}}(x)}} \tag{14.1.274}$$

In the case of a linear regression with several explanatory variables, therefore assimilated to the concept of "degrees of freedom DoF", the idea is the same but the calculations are longer (we don't have yet the will to do the mathematical developments for this case).

Finally, remember that we got for the empirical correlation coefficient:

$$R_{X,Y}^2 = \frac{a^2\mathrm{V}(X)}{\mathrm{V}(Y)} \tag{14.1.275}$$

We then verbatim get the famous confidence interval of the correlation coefficient:

$$\left(A - T_{\alpha/2}(n-2)\sqrt{\frac{\dfrac{1}{n-2}\sum e_k^2}{n\mathrm{V}(x)}}\right)^2 \frac{\mathrm{V}(X)}{\mathrm{V}(Y)} \leq R_{X,Y}^2 \leq \left(A + T_{\alpha/2}(n-2)\sqrt{\frac{\dfrac{1}{n-2}\sum e_k^2}{n\mathrm{V}(x)}}\right)^2 \frac{\mathrm{V}(X)}{\mathrm{V}(Y)} \tag{14.1.276}$$

The reader should know that as calculating the confidence interval for the slope or for the correlation coefficient are two equivalent things, many softwares (Tanagra, Minitab, Microsoft® Excel®, etc.) give only the value of the Student's distribution at the critical value of this latter only for the slope and... they then assume that the user of the software knows that it is the same for the correlation coefficient.

14.1.10.2.1 Pearson Correlation Coefficient Test

The calculation obtained above for the confidence interval of the correlation coefficient is a little difficult in practice. It is for this reason that many practitioners and statistical softwares implements a very simple alternative communicated only in the minimal form that is the p-value.

To see this approach, remember that we proved in the section of Statistics (see page 597) that (this time we will adopt the correct notation...):

$$\rho=\frac{\text{cov}(X,Y)}{\sqrt{\text{V}(X)\text{V}(Y)}} \tag{14.1.277}$$

And we saw just above that:

$$\rho=\frac{\text{cov}(X,Y)}{\sqrt{\text{V}(X)}}\sqrt{\text{V}(X)}\frac{1}{\sqrt{\text{V}(X)\text{V}(Y)}}=a\sqrt{\text{V}(X)}\frac{1}{\sqrt{\text{V}(X)\text{V}(Y)}}=a\frac{\sigma_X}{\sigma_Y} \tag{14.1.278}$$

Similarly, we have the Pearson correlation coefficient estimator which is (using here the various possible notations that we can find in the literature...):

$$R_{X,Y}=\sqrt{\frac{\text{SSE}}{\text{SST}}}=\frac{A\sqrt{\sum\limits_k(x_k-\bar{x}^2)}}{\sqrt{\sum\limits_k(y_k-\bar{y})^2}}=\frac{A\sqrt{\frac{1}{n-1}\sum\limits_k(x_k-\bar{x}^2)}}{\frac{1}{n-1}\sqrt{\sum\limits_k(y_k-\bar{y})^2}}=A\frac{\hat{\sigma}_X}{\hat{\sigma}_Y} \tag{14.1.279}$$

and therefore:

$$A=\frac{R_{X,Y}\hat{\sigma}_Y}{\hat{\sigma}_X} \tag{14.1.280}$$

The hypothesis test we want to do is therefore:

$$H_0:\quad \rho=0 \tag{14.1.281}$$

that is equivalent to:

$$H_0:\quad a=0 \tag{14.1.282}$$

The null hypothesis obviously being that the Pearson correlation is statistically significantly different from zero. So this is a bilateral test!

To find a simple form of the test, remember that we obtained:

$$(n-2)\hat{\sigma}_{\varepsilon}^2=\sum_k(y_k'-y_k)^2=\sum\varepsilon_k^2=\text{SSR} \tag{14.1.283}$$

and also:

$$R_{X,Y}^2=\frac{\text{SSE}}{\text{SST}}=\frac{\text{SST}-\text{SSR}}{\text{SST}}-1=1-\frac{\text{SSR}}{\text{SST}} \tag{14.1.284}$$

which brings us mixing the two to have:

$$\text{SSR}=\text{SST}(1-R_{X,Y}^2)=\sum_k (y_k-\bar{y})^2(1-R_{X,Y}^2)=(n-1)\hat{\sigma}_Y^2(1-R_{X,Y}^2) \tag{14.1.285}$$

Therefore:

$$\hat{\sigma}_\varepsilon^2=\frac{(n-1)\hat{\sigma}_Y^2(1-R_{X,Y}^2)}{n-2} \tag{14.1.286}$$

But we remember that we have also proved that if n is large enough:

$$A-T_{\alpha/2}(n-2)\sqrt{\frac{\dfrac{1}{n-2}\sum e_k^2}{n\hat{\mathbb{V}}(x)}}\leq a\leq A+T_{\alpha/2}(n-2)\sqrt{\frac{\dfrac{1}{n-2}\sum e_k^2}{n\hat{\mathbb{V}}(x)}} \tag{14.1.287}$$

But if n is small we fall back on:

$$A-T_{\alpha/2}(n-2)\sqrt{\frac{\dfrac{1}{n-2}\sum e_k^2}{(n-1)\hat{\sigma}_X}}\leq a\leq A+T_{\alpha/2}(n-2)\sqrt{\frac{\dfrac{1}{n-2}\sum e_k^2}{(n-1)\hat{\sigma}_X}} \tag{14.1.288}$$

Therefore:

$$T_{\alpha/2}(n-2)\geq\frac{A-a}{\sqrt{\dfrac{\dfrac{1}{n-2}\sum e_k^2}{(n-1)\hat{\sigma}_X}}}=\frac{A-a}{\sqrt{\dfrac{\dfrac{1}{n-2}\text{SSR}}{(n-1)\hat{\sigma}_X}}} \tag{14.1.289}$$

and with the null hypothesis $a=0$, we get:

$$\begin{aligned}T_{\alpha/2}(n-2)&\geq\frac{A}{\sqrt{\dfrac{\dfrac{1}{n-2}\text{SSR}}{(n-1)\hat{\sigma}_X}}}=\frac{R_{X,Y}\dfrac{\hat{\sigma}_Y}{\hat{\sigma}_X}}{\sqrt{\dfrac{\dfrac{\text{SSR}}{n-2}}{(n-1)\hat{\sigma}_X}}}=\frac{R_{X,Y}\hat{\sigma}_Y}{\hat{\sigma}_X\sqrt{\dfrac{\hat{\sigma}}{(n-1)\hat{\sigma}_X^2}}}=\frac{R_{X,Y}\hat{\sigma}_Y}{\hat{\sigma}_\varepsilon\sqrt{\dfrac{1}{n-1}}}\\&=\frac{R_{X,Y}\hat{\sigma}_Y}{\sqrt{\dfrac{1}{n-1}}\sqrt{\dfrac{(n-1)\hat{\sigma}_Y^2(1-R_{X,Y}^2)}{n-2}}}=\frac{R_{X,Y}}{\sqrt{\dfrac{1-R_{X,Y}^2}{n-2}}}\end{aligned} \tag{14.1.290}$$

Be careful with the use of this test often and logically named "Student's T-test for univariate regression slope" or "T-test for coefficient slope", depending if the Pearson correlation coefficient is negative or positive and don't forget that it is bilateral!

✍**Examples:**

E1. We have calculated for a series of data a positive Pearson correlation coefficient R of value 0.298 and the explanatory variable has 7 values. So we have with the English version Microsoft® Excel® 11.8346 the p-value that is given by (we find exactly the same value as with Minitab 15.1.1):

```
=2*(1-T.DIST(0.298/SQRT((1-0.298^2)/(7-2));7-2;1))
=2*(1-0.741869)=0.51626
```

In this case we sadly can't reject the null hypothesis as the Pearson correlation coefficient is equal to zero at a threshold of 5%.

E2. We have calculated for a data set a positive Pearson correlation coefficient R of -0.084 and the explanatory variable contains 19 values. So we have with the English version of Microsoft® Excel® 11.8346 the p-value that is equal to (we find exactly the same value as with Minitab 15.1.1):

```
=2* T.DIST((-0.084)/SQRT((1-(-0.084)^2)/(19-2));19-2;1)
=2*(1-0.366)=0.74186
```

In this case we also sadly can't reject the null hypothesis as what the Pearson correlation coefficient is equal to zero at a threshold of 5%.

This small trap makes that finally we take the absolute value of the Pearson correlation coefficient and we therefore use the always the same calculation method.

14.1.10.2.2 Confidence interval of predicted values

We wish for each measured value of the dependent variable, know the confidence interval. In other words, we would like to know the statistical variance estimator of Y (we do not write the subscripts anymore to save time):

$$\text{V}(Y)=\text{V}(Ax+B)=\text{V}(Ax)+\text{V}(B)+2\text{cov}(A,B) \tag{14.1.291}$$

Unfortunately, we will go into the wall because the covariance is difficult to calculate (A and B are not independent as shown by the expressions we got previously).

By cons, being a good observer, we see that if we use the result seen above:

$$B=\bar{Y}-A\bar{x} \tag{14.1.292}$$

Then:

$$\text{V}(Y)=\text{V}(Ax+B)=\text{V}(Ax+(\bar{Y}-A\bar{x}))=\text{V}(\bar{Y}+A(x-\bar{x})) \tag{14.1.293}$$

The problem being circumvented, we now have using the properties of variance:

$$\begin{aligned}\text{V}(Y)&=\text{V}(\bar{Y}+A(x-\bar{x}))=\text{V}(\bar{Y})+\text{V}(A(x-\bar{x}))+2\underbrace{\text{cov}(\bar{Y},A(x-\bar{x}))}_{=0}\\&=\text{V}(\bar{Y})+\text{V}(A(x-\bar{x}))=\text{V}(\bar{Y})+(x-\bar{x})^2\text{V}(A)=\frac{\sigma^2}{n}+(x-\bar{x})^2\frac{\sigma^2}{\sum\limits_k(x_k-\bar{x})^2}\end{aligned} \tag{14.1.294}$$

Therefore:

$$\text{V}(Y)=\sigma^2\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum\limits_k(x_k-\bar{x})^2}\right) \tag{14.1.295}$$

Then we have finally:

$$Y=\mathcal{N}\left(Ax+B,\sigma\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum\limits_k(x_k-\bar{x})^2}}\right) \tag{14.1.296}$$

Now let us recall (see section Statistics page 595) that:

$$Z=\frac{X-\mu}{\sigma} \tag{14.1.297}$$

and as Y is distributed according to a Normal distribution for whose the unbiased estimator of the mean and standard deviation are given by the prior-previous relation it comes immediately:

$$Z=\frac{Y-(Ax+B)}{\sigma\sqrt{\dfrac{1}{n}+\dfrac{(x-\bar{x})^2}{\sum\limits_k(x_k-\bar{x})^2}}} \tag{14.1.298}$$

In practice you must also check that this ratio follows a Normal distribution in order to make the confidence intervals and statistical tests that follow.

Let us recall now that we have proved in the section Statistics (see page 783) that:

$$T(k)=T_k=\frac{Z}{\sqrt{U/k}} \tag{14.1.299}$$

follows a Student law of degrees of freedom k and the variable U follows a chi-square law of degree of freedom k.

Now let us come back to the expression of the Z obtained above and remember that:

$$\hat{\sigma}^2=\frac{1}{n-2}\sum\varepsilon_i^2 \tag{14.1.300}$$

Therefore:

$$\hat{\sigma}=\sqrt{\frac{\sum\varepsilon_i^2}{n-2}} \tag{14.1.301}$$

But as we have the assumption:

$$\varepsilon_i=\mathcal{N}(0,\sigma) \tag{14.1.302}$$

Therefore:

$$\frac{\varepsilon_i}{\sigma}=\mathcal{N}(0,1) \tag{14.1.303}$$

and obviously:

$$\sum\frac{\varepsilon_i^2}{\sigma^2}=\sum\left(\frac{\varepsilon_i}{\sigma}\right)^2 \tag{14.1.304}$$

corresponds to a sum of squares of reduced centered Normal laws. And therefore according to what we have proved in section Statistics (see page 863) it follows that:

$$\sum\left(\frac{\varepsilon_i}{\sigma}\right)^2=\chi^2(n-2)=\chi^2_{n-2} \tag{14.1.305}$$

Therefore:

$$\begin{aligned}T(n-2)=T_{n-2}&=\frac{Z}{\sqrt{U/k}}=\frac{Z}{\sqrt{\dfrac{\sum\left(\frac{\varepsilon_i}{\sigma}\right)^2}{n-2}}}=\frac{Z}{\sqrt{\dfrac{1}{\sigma^2}\sum\dfrac{\varepsilon_i^2}{n-2}}}\\&=\frac{\sigma Z}{\sqrt{\sum\dfrac{\varepsilon_i^2}{n-2}}}=\frac{\sigma Z}{\hat{\sigma}}\end{aligned} \tag{14.1.306}$$

That is to say:

$$T(n-2)=\frac{\sigma Z}{\hat{\sigma}}=\frac{\sigma\dfrac{Y-(Ax+B)}{\sigma\sqrt{\dfrac{1}{n}+\dfrac{(x-\bar{x})^2}{\sum\limits_k(x_k-\bar{x})^2}}}}{\hat{\sigma}}=\frac{Y-(Ax+B)}{\hat{\sigma}\sqrt{\dfrac{1}{n}+\dfrac{(x-\bar{x})^2}{\sum\limits_k(x_k-\bar{x})^2}}} \tag{14.1.307}$$

This is one reason why many statisticians note directly and without detours:

$$Z=\frac{Y-(Ax+B)}{\sigma\sqrt{\dfrac{1}{n}+\dfrac{(x-\bar{x})^2}{\sum\limits_k (x_k-\bar{x})^2}}}\Rightarrow T(n-2)=\frac{Y-(Ax+B)}{\hat{\sigma}\sqrt{\dfrac{1}{n}+\dfrac{(x-\bar{x})^2}{\sum\limits_k (x_k-\bar{x})^2}}} \tag{14.1.308}$$

Which is not necessarily obvious at first glance. This is why, following the request of a reader, we have detailed a little bit the mechanism behind this involvement.

This done, we have finally:

$$T(n-2)=\frac{Y-(Ax+B)}{\hat{\sigma}\sqrt{\dfrac{1}{n}+\dfrac{(x-\bar{x})^2}{\sum\limits_k (x_k-\bar{x})^2}}} \tag{14.1.309}$$

and we deduce from this immediately a bilateral confidence interval of a given threshold level α for a fixed x whose expression is:

$$\boxed{Y=Ax+B\pm T_{\alpha/2}(n-2)\hat{\sigma}\sqrt{\dfrac{1}{n}+\dfrac{(x-\bar{x})^2}{\sum\limits_k (x_k-\bar{x})^2}}} \tag{14.1.310}$$

Remark

The same kind of development can be done for the slope and the intercept. This is why software like Microsoft® Excel®, SPSS, Minitab, Statistica, etc. give the value of the Student T-distribution as well as the confidence interval for a given threshold level α. But for this to be meaningful, you must never forget that all the assumptions of the model must be met.

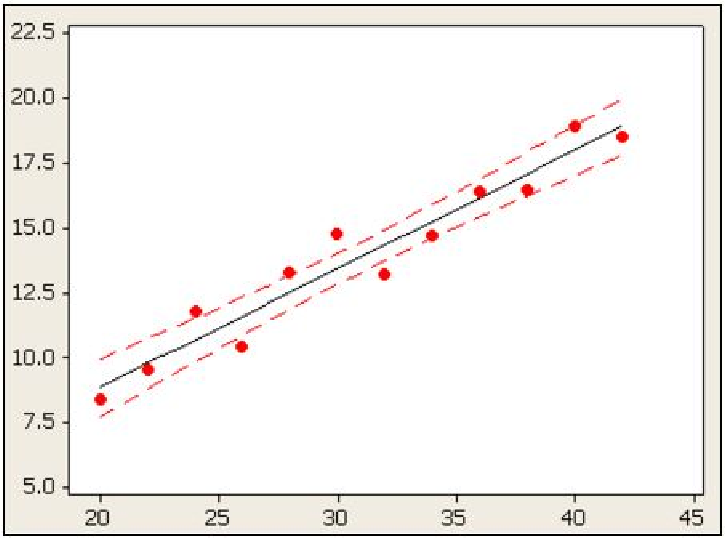


Figure 14.13: Print screen of a confidence interval obtained with Minitab 15

The reader will may have noticed that:

- It is very boring to obtain without software or without coding the plot of the confidence interval for the ordinary least squares since we have to calculate it for each point...
- The confidence interval is curved which is sometimes considered as common sense, at least in the temporal version of the regression: furthest is the forecast less accurate it will be

The true value of Y is given by:

$$y = a + bx + \varepsilon \tag{14.1.311}$$

with the variance:

$$\mathrm{V}(y) = \mathrm{V}(\varepsilon) = \sigma^2 \tag{14.1.312}$$

which is independent of the estimator Y. Therefore, the difference between Y and y (therefore between estimator and real value) has for variance:

$$\mathrm{V}(Y-y) = \mathrm{V}(Y) - \mathrm{V}(y) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum\limits_k (x_k - \bar{x})^2}\right) \tag{14.1.313}$$

Therefore it is customary to consider that the "prediction interval" (not to be confused with the confidence interval of the estimator) is taken as:

$$\boxed{Y = Ax + B \pm T_{\alpha/2}(n-2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum\limits_k (x_k - \bar{x})^2}}} \tag{14.1.314}$$

Remark

The same kind of development can be done for the slope and the intercept. This is why software like Microsoft® Excel®, IBM SPSS, Minitab, Statistica, etc. give the value of the Student T-distribution as well as the confidence interval for a given threshold level α. But for this to be meaningful, you must never forget that all the assumptions of the model must be met.

This gives us the following type of chart with a software such as Minitab 15:

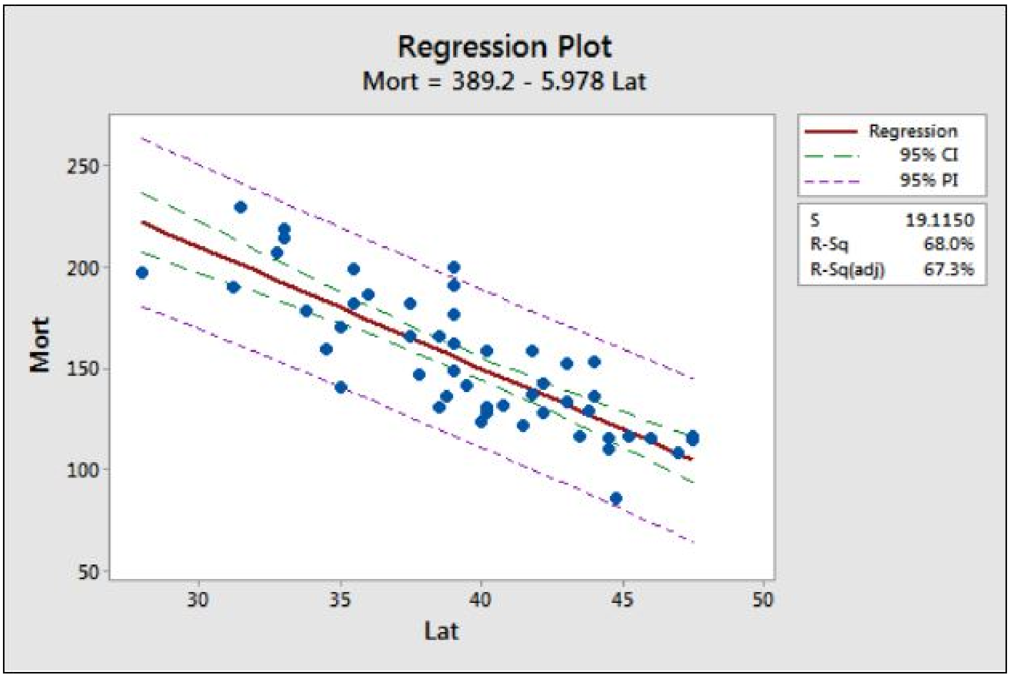


Figure 14.14: Print screen of a prediction interval (PI) in green obtained with Minitab 15

Where we can see the prediction interval (PI) in green and the confidence interval (CI) in red and its is obvious to see by the property of the variance that CI<PI.

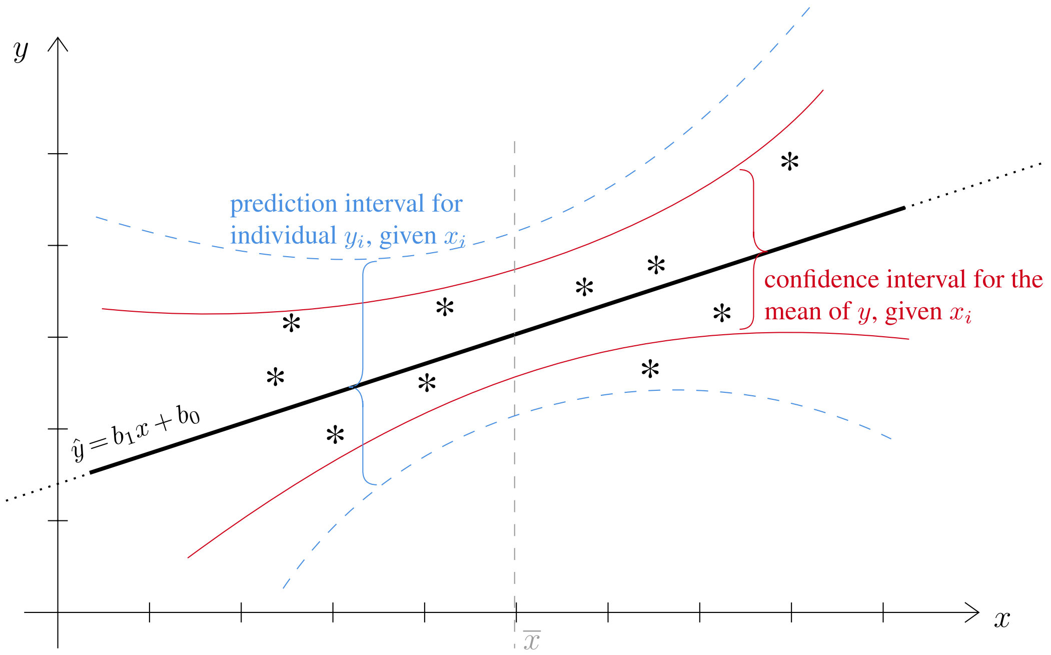
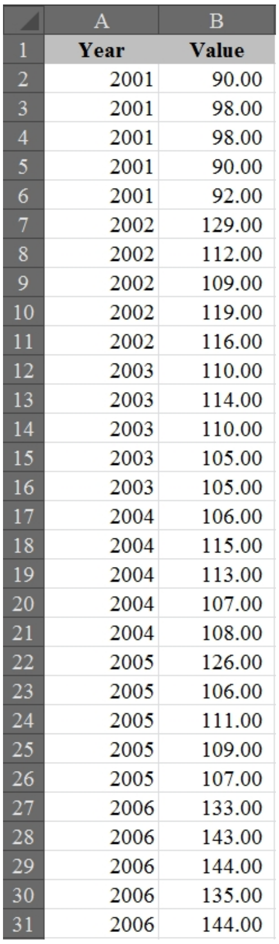

y
x
$\overline{x}$
$\hat{y}=b_1x+b_0$
prediction interval for individual y_i, given x_i
confidence interval for the mean of y, given x_i

Figure 14.15: Confidence vs Prediction interval

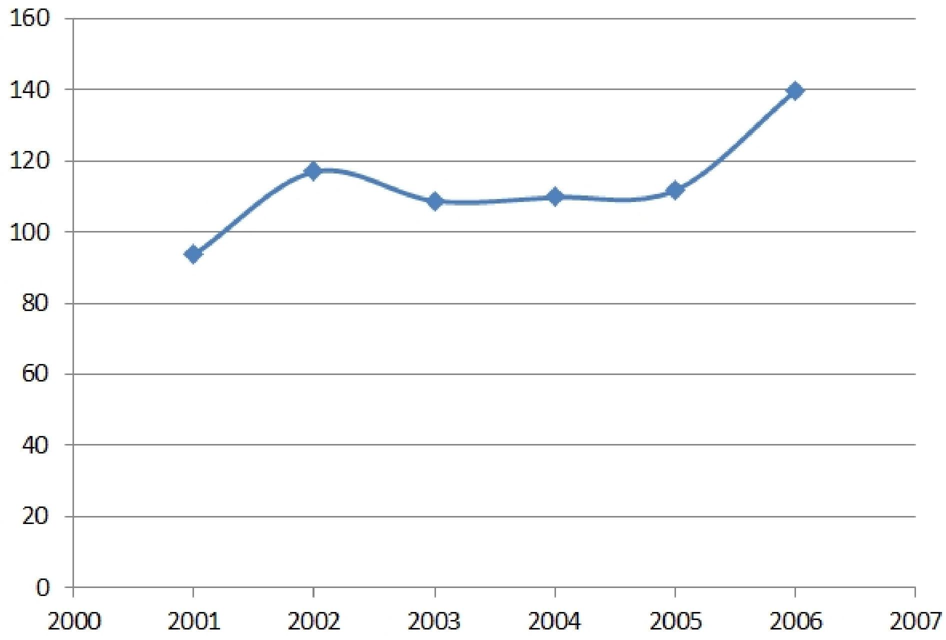
These two intervals are often confused. To understand these, it's important to realize that a regression model fits a relationship between the input variables and the mean value of the outcome variable - this is the red line in the below chart. The confidence interval indicates a range of certainty around the mean value of the outcome variable - usually 95% - and is illustrated by the green lines above. The prediction interval indicates a range of certainty around any value of the outcome variable and is illustrated by the purple lines above. Another way of looking at this is to imagine taking a vertical slice of the below chart. This would translate - under the most common models - to a bell curve for the value of the outcome variable at that specific value of the input variable. The red point indicates the most likely center of the bell curve, the green lines represents the 95% likelihood range for that center and the purple lines represents the likelihood for the 95% range (i.e. width) of all these bell curves.

So what must be well understand by the reader so far, is that if we are dealing with univariate linear regression or any other type of regression and want to plot the following data that corresponds to a sample for each x-value:



	A	B
1	**Year**	**Value**
2	2001	90.00
3	2001	98.00
4	2001	98.00
5	2001	90.00
6	2001	92.00
7	2002	129.00
8	2002	112.00
9	2002	109.00
10	2002	119.00
11	2002	116.00
12	2003	110.00
13	2003	114.00
14	2003	110.00
15	2003	105.00
16	2003	105.00
17	2004	106.00
18	2004	115.00
19	2004	113.00
20	2004	107.00
21	2004	108.00
22	2005	126.00
23	2005	106.00
24	2005	111.00
25	2005	109.00
26	2005	107.00
27	2006	133.00
28	2006	143.00
29	2006	144.00
30	2006	135.00
31	2006	144.00

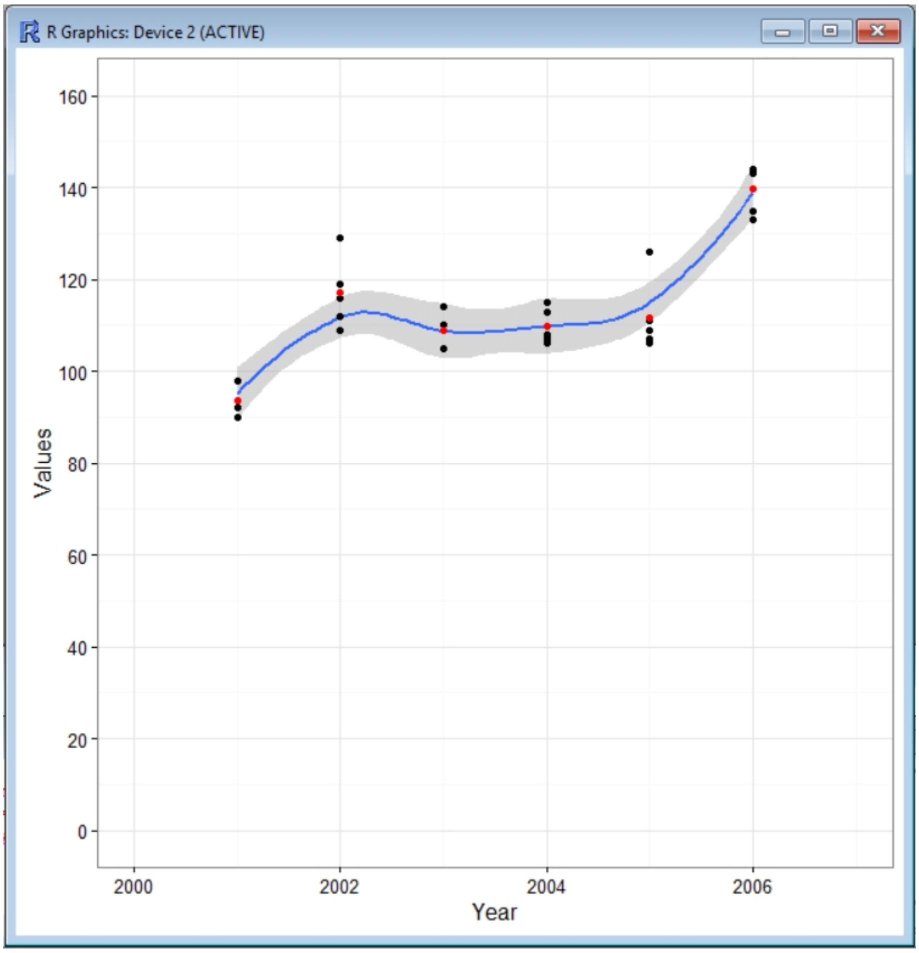
Therefore using a spreadsheet software to make a plot of that the average of each year as following (most common case see in Fortune 500 companies by top managers and also by auditing companies like Klynveld Peat Marwick Goerdeler/KPMG and Ernst & Young/EY):



using a smooth chart is everything but scientific. Indeed:

1. No spreadsheet software is able to summary the data list above automatically in quick an efficient way (with Microsoft® Excel® even Pivot Charts cannot do smooth scatter charts of such a list of data...).
2. Smooth scatter plot in spreadsheet softwares use splines interpolation and these are not statistical models that can be used for samples! So reading such charts in a spreadsheet software introduce a huge deterministic bias in the mind of the board committee or any customer/supplier.
3. As the averages come from samples there is an implicit confidence interval THAT HAS TO BE SHOWN to the board committee or customer/supplier. Do not show a confidence interval on sample data is a high-school level error that is not acceptable when done by a senior manager or business analyst or even a consultant of a big 5 auditing companies.

So following the detailed steps given in our `R` companion book we get something truly scientific and robust that is:


R Graphics: Device 2 (ACTIVE)
160
140
120
100
80
60
40
20
0
Values
2000
2002
2004
2006
Year

14.1.10.3 Linear univariate regression forced through the origin (RTO)

One very common requested case in laboratories (and generally in other departments) of corporations, is to force the linear regression through the origin. This technique is named "regression through origin" and abbreviated RTO.

We will see now (in the univariate case) that the approach is only a simplified variant of the method of the ordinary least squares.

We use as before:

$$\text{SSR} = \sum_{k=1}^{n} (y_k - y'_k)^2 \tag{14.1.315}$$

where n is the number of points. But this time, let us write:

$$y'_k = a x_k \tag{14.1.316}$$

then:

$$\text{SSR} = \sum_{k=1}^{n} (y_k - a x_k)^2 = \Phi(a) \tag{14.1.317}$$

This relation make appears the sum of squared deviations as a function of the parameter a. When this function is minimal (extremal), the derivatives with respect to its parameters are cancelled:

$$\frac{\partial \Phi}{\partial a}(a) = -2x_k \sum_{k=1}^{n} (y_k - a x_k) = 0 \tag{14.1.318}$$

After simplification:

$$a \sum_{k=1}^{n} x_k^2 = \sum_{k=1}^{n} x_k y_k \tag{14.1.319}$$

Finally:

$$a = \frac{\displaystyle\sum_{k=1}^{n} x_k y_k}{\displaystyle\sum_{k=1}^{n} x_k^2} \tag{14.1.320}$$

You can also easily check with any spreadsheet software (Microsoft® Excel® for example) that the calculations corresponds well.

However the reader must know now that the RTO subject is surprisingly controversial among statisticians. Indeed, forcing the regression line through the origin is generally inconsistent with the best fit.

Caution! The text below is mainly a reproduction with some adaptations of the following article [163]. Thanks to the donations we were able to buy the rights (225 US$...) to John Wiley & Sons, Inc. to reproduce it in this textbook. But in no manner the reader is authorized to reproduce the same text without the agreement of John Wiley & Sons, Inc. !!!

The proper method for evaluating RTO has long been disputed. To appreciate the controversy, note the familiar identity:

$$(y_k-\bar{y})=(y_k-\hat{y}_k)+(\hat{y}_k-\bar{y}) \tag{14.1.321}$$

Squaring both sides and summing across all observation gives:

$$\sum(y_k-\bar{y})^2=\sum(y_k-\hat{y}_k)^2+\sum(\hat{y}_k-\bar{y})^2+2\sum(y_k-\hat{y}_k)(\hat{y}_k-\bar{y}) \tag{14.1.322}$$

However, as we have proved earlier above, the cross-product term must be equal to zero in the case of OLS. The remaining terms therefore constitute the usual analysis of variance decomposition:

$$\sum(y_k-\bar{y})^2=\sum(y_k-\hat{y}_k)^2+\sum(\hat{y}_k-\bar{y})^2 \tag{14.1.323}$$

And we know that the coefficient of determination for OLS is then defined by:

$$R^2=\frac{\sum(\hat{y}_k-\bar{y})^2}{\sum(y_k-\bar{y})^2}=1-\frac{\sum(y_k-\hat{y}_k)^2}{\sum(y_k-\bar{y})^2} \tag{14.1.324}$$

Some authors claim that because this diagnostic measure is based on an identity, it should not depend on the inclusion or exclusion of a constant term in the regression. From that perspective, the previous relation is equally valid for RTO and OLS.

However, when there is no constant in the regression, $\sum(y_k-\hat{y}_k)(\hat{y}_k-\bar{y})$ will generally take a non-zero value, so the variance decomposition:

$$\sum(y_k-\bar{y})^2=\sum(y_k-\hat{y})^2+\sum(\hat{y}_k-\bar{y})^2 \tag{14.1.325}$$

is not a valid basis for analysis of variance in RTO. And if the RTO model provides a sufficiently poor fit, the data may exhibit more variation around the regression line than around y, in which case:

$$\sum(y_k-\hat{y})^2>\sum(y_k-\bar{y})^2 \tag{14.1.326}$$

Heedlessly applying:

$$R^2=1-\frac{\sum(y_k-\hat{y})^2}{\sum(y_k-\bar{y})^2} \tag{14.1.327}$$

would then result in an implausibly negative (and thus non-interpretable) coefficient of determination as well as a negative F ratio. Moreover, it is often argued that defining SST as the sum of squared deviations from the mean is inappropriate when the regression line is forced through the origin but does not necessarily pass through $(\bar{x},\bar{y})$, when so viewed, the starting relation:

$$(y_k-\bar{y})=(y_k-\hat{y})+(\hat{y}_k-\bar{y}) \tag{14.1.328}$$

is replaced by identity:

$$(y_k-0)=(y_k-\hat{y}_k)+(\hat{y}_k-0) \tag{14.1.329}$$

Squaring and summing yields:

$$\sum y_k^2=\sum(y_k-\hat{y}_k)^2+\sum\hat{y}_k^2+2\sum\hat{y}_k(y_k-\hat{y}_k) \tag{14.1.330}$$

but the final (cross-product) term in this equation equals zero under RTO, because:

$$\begin{aligned}\sum\hat{y}_k(y_k-\hat{y}_k)&=\sum\hat{\beta}x_k(y_k-\hat{\beta}x_k)=\hat{\beta}\left(\sum x_ky_k-\hat{\beta}\sum x_k^2\right)\\&=\hat{\beta}\left[\sum x_ky_k-\left(\frac{\sum x_ky_k}{\sum x_k^2}\right)\sum x_k^2\right]=0\end{aligned} \tag{14.1.331}$$

Thus, the variance OLS decomposition:

$$\sum(y_k-\bar{y})^2=\sum(y_k-\hat{y}_k)^2+\sum(\hat{y}_k-\bar{y})^2 \tag{14.1.332}$$

is replaced by:

$$\sum y_k^2=\sum(y_k-\hat{y}_k)^2+\sum\hat{y}_k^2 \tag{14.1.333}$$

Applying this latter relation one rather than the classical OLS one to RTO, one finds that SSE is unchanged, but $\text{SST}=\sum y_k^2$ and $\text{SSR}=\sum\hat{y}_k^2$. Redefining SST and SSR in this manner results in:

$$R^2=\frac{\sum\hat{y}_k^2}{\sum y_k^2} \tag{14.1.334}$$

ie, also a strictly non-negative coefficient of determination that equals or exceeds the classical R^2 measurement. Of course, these definitions also affect the adjusted R^2 and F statistics.

Note that, without a constant, the degrees of freedom for SST, SSR and SSE are n, k and $n\breve{\ }k$, respectively, where n is the sample size and k is the number of independent variables. Thus:

$$\text{SE}=\sqrt{\frac{\text{SSE}}{n-k}} \tag{14.1.335}$$

regardless of how SST is defined.

The controversy over SST is not merely academic: practitioners (and students) running RTO got various outputs depending on which computer packages they use a few dozens of years ago (since they almost all give the same result based on the adapted definition of the correlation coefficient).

14.1.10.4 Deming regression (orthogonal regression)

As we have already mentioned it, the "orthogonal linear regression model" or "Deming regression" is used as complement to the paired T-test to check the stability of the measuring instruments in laboratories. This is a case where the explanatory and dependent variables are tainted with uncertainty (we speak then of "stochastic regressors" at the opposite classical case where we have "nonstochastic regressors"). In other words, this regression is used when working in a situation where it is desired to compare two linearly increasing or decreasing variables and where each is tainted with measurement errors including error values (of the two variables!) are independent.

Remarks

R1. Chemists do this analysis sometimes in a qualitative form using the "Bland-Altman plots".

R2. Deming regression is a special case of "Total Least Squares regression" (TLSR) that minimize the total geometric distance between data points and the best-fit line rather than just the vertical distance as in Ordinary Least Squares regression (OLSR).

We will see now the derivation of the maximum likelihood estimates related to the Deming regression model. It is based on the book *Models in regression and related topics* (chapter three), from 1969 by Peter Sprent, but with more detailed calculations included (big thanks to Anders Jensen for the LaTeX code of the proof!) .

14.1.10.4.1 Deming orthogonal regression log-likelihood derivation

The mathematical model $\eta=\alpha+\beta\xi$ describes a linear relationship between two variables ξ and η. Observations x and y of two variables are usually described by a regression of y on x where x is assumed to be observed without error (or, equivalently using the conditional distribution of y given x). In linear regression with observations subject to additive random variation on both x and y and observed values for individuals $(x_i,y_i), i=1,\ldots,n$, a model may be written:

$$x_i=\xi_i+e_{xi}\quad y_i=\eta_i+e_{yi}=\alpha+\beta\xi_i+e_{yi} \tag{14.1.336}$$

where e_{xi} and e_{y_i} denotes the random part of the model. This is known as a functional relationship because the ξ_i's are assumed to be fixed parameters, as opposed to a structural relationship where some distribution for the ξ_i's is assumed. In the following it is assumed that the e_{xi}s are independent and identically distributed with:

$$e_{yi}\sim\mathcal{N}(0,\sigma^2) \tag{14.1.337}$$

and that the e_{yi}s are independent and identically distributed with:

$$e_{yi}\sim\mathcal{N}(0,\lambda\sigma^2) \tag{14.1.338}$$

for some $\lambda>0$. Furthermore e_{xi} is assumed to be independent of e_{yi}.

The aim of this document is to derive the maximum likelihood estimates for α,β,ξ_i and σ^2 in the functional model stated above.

The likelihood function:

$$f_{x_1,x_2,\ldots,x_n,y_1,y_2,\ldots,y_n}(\alpha,\beta,\xi_1,\xi_2,\ldots,\xi_n,\sigma^2) \tag{14.1.339}$$

denoted f is:

$$f=\prod_{i=1}^{n}\left(2\pi\sigma^2\right)^{-\frac{1}{2}}\exp\left(-\frac{(x_i-\xi_i)^2}{2\sigma^2}\right)\left(2\pi\lambda\sigma^2\right)^{-\frac{1}{2}}\exp\left(-\frac{(y_i-\alpha-\beta\xi_i)^2}{2\lambda\sigma^2}\right) \tag{14.1.340}$$

and the log-likelihood, denoted $\mathcal{L}$, is:

$$\begin{aligned}\mathcal{L}&=\sum_{i=1}^{n}-\frac{1}{2}\log\left(2\pi\sigma^2\right)-\frac{(x_i-\xi_i)^2}{2\sigma^2}-\frac{1}{2}\log\left(2\pi\lambda\sigma^2\right)-\frac{(y_i-\alpha-\beta\xi_i)^2}{2\lambda\sigma^2}\\&=-\frac{n}{2}\log\left(4\pi^2\right)-\frac{n}{2}\log\left(\lambda\sigma^4\right)-\frac{\displaystyle\sum_{i=1}^{n}(x_i-\xi_i)^2}{2\sigma^2}-\frac{\displaystyle\sum_{i=1}^{n}(y_i-\alpha-\beta\xi_i)^2}{2\lambda\sigma^2}\end{aligned} \tag{14.1.341}$$

It follows that the likelihood function is not bounded from above when σ^2 goes to 0, so in the following it is assumed that $\sigma^2>0$.

Let us solve for ξ_i now! The differentiation of $\mathcal{L}$ with respect to ξ_i gives:

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\xi_i}&=\frac{\partial}{\partial\xi_i}\left(-\frac{\displaystyle\sum_{i=1}^{n}(x_i-\xi_i)^2}{2\sigma^2}-\frac{\displaystyle\sum_{i=1}^{n}(y_i-\alpha-\beta\xi_i)^2}{2\lambda\sigma^2}\right)\\&=\frac{(x_i-\xi_i)}{\sigma^2}+\frac{\beta(y_i-\alpha-\beta\xi_i)}{\lambda\sigma^2}\end{aligned} \tag{14.1.342}$$

Setting $\frac{\partial\mathcal{L}}{\partial\xi_i}$ equal to zero yields:

$$\frac{\partial\mathcal{L}}{\partial\xi_i}=0\Rightarrow\quad\xi_i=\frac{\lambda\sigma^2x_i+\beta\sigma^2y_i-\beta\alpha\sigma^2}{\lambda\sigma^2+\beta^2\sigma^2}=\frac{\lambda x_i+\beta(y_i-\alpha)}{\lambda+\beta^2} \tag{14.1.343}$$

So to estimate ξ_i, estimates for β and α are needed. Therefore focus is turned to the derivation of $\hat{\alpha}$.

Let us now solve for α. The differentiation of $\mathcal{L}$ with respect to α gives:

$$\frac{\partial\mathcal{L}}{\partial\alpha}=\frac{\partial}{\partial\alpha}\left(-\frac{\displaystyle\sum_{i=1}^{n}(y_i-\alpha-\beta\xi_i)^2}{2\lambda\sigma^2}\right)=\frac{\displaystyle\sum_{i=1}^{n}(y_i-\alpha-\beta\xi_i)}{\lambda\sigma^2} \tag{14.1.344}$$

and putting $\frac{\partial\mathcal{L}}{\partial\alpha}$ equal to zero yields to:

$$\frac{\partial L}{\partial\alpha}=0\Rightarrow\quad\alpha=\frac{1}{n}\sum_{i=1}^{n}(y_i-\beta\xi_i) \tag{14.1.345}$$

Now one can use:

$$\xi_i = \frac{\lambda x_i + \beta(y_i - \alpha)}{\lambda + \beta^2} \tag{14.1.346}$$

to dispense with ξ_i:

$$\begin{aligned}
\alpha &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta\xi_i)\\
&= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \beta\frac{\lambda x_i + \beta(y_i - \alpha)}{\lambda + \beta^2}\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \beta\frac{\lambda x_i + \beta y_i}{\lambda + \beta^2} + \frac{\beta^2\alpha}{\lambda + \beta^2}\right)\\
&\Updownarrow\\
\alpha\left(1 - \frac{\beta^2}{\lambda + \beta^2}\right) &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \beta\frac{\lambda x_i + \beta y_i}{\lambda + \beta^2}\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}\left(y_i\left(1 - \frac{\beta^2}{\lambda + \beta^2}\right) - x_i\frac{\beta\lambda}{\lambda + \beta^2}\right)\\
&\Updownarrow\\
\alpha &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i\frac{\beta\lambda}{\lambda + \beta^2}\frac{\lambda + \beta^2}{\lambda}\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i\beta)\\
&= \overline{y} - \overline{x}\beta
\end{aligned} \tag{14.1.347}$$

Hence the estimate for α becomes:

$$\hat{\alpha} = \overline{y} - \overline{x}\hat{\beta} \tag{14.1.348}$$

Solving for β by differentiating $\mathcal{L}$ with respect to β gives:

$$\frac{\partial\mathcal{L}}{\partial\beta} = \frac{\partial}{\partial\beta}\left(-\frac{\displaystyle\sum_{i=1}^{n}(y_i - \alpha - \beta\xi_i)^2}{2\lambda\sigma^2}\right) = \frac{\displaystyle\sum_{i=1}^{n}(y_i - \alpha - \beta\xi_i)\xi_i}{\lambda\sigma^2} \tag{14.1.349}$$

Setting $\frac{\partial\mathcal{L}}{\partial\beta}$ equal to zero yields:

$$\frac{\partial\mathcal{L}}{\partial\beta}=0\Leftrightarrow\sum_{i=1}^{n}(y_i-\alpha-\beta\xi_i)\xi_i=0 \tag{14.1.350}$$

and using again:

$$\xi_i=\frac{\lambda x_i+\beta(y_i-\alpha)}{\lambda+\beta^2} \tag{14.1.351}$$

we get:

$$\begin{aligned}0&=\sum_{i=1}^{n}(y_i-\alpha-\beta\xi_i)\xi_i\\&=\sum_{i=1}^{n}\left(y_i-\alpha-\beta\frac{\lambda x_i+\beta(y_i-\alpha)}{\lambda+\beta^2}\right)\frac{\lambda x_i+\beta(y_i-\alpha)}{\lambda+\beta^2}\end{aligned} \tag{14.1.352}$$

This implies that:

$$\begin{aligned}0&=\sum_{i=1}^{n}\Big((y_i-\alpha)(\lambda+\beta^2)-\beta\lambda x_i-\beta^2(y_i-\alpha)\Big)\Big(\lambda x_i+\beta(y_i-\alpha)\Big)\\&=\sum_{i=1}^{n}\lambda^2x_i(y_i-\alpha)+\beta^2\lambda x_i(y_i-\alpha)-\beta\lambda^2x_i^2-\beta^2\lambda x_i(y_i-\alpha)+\\&\sum_{i=1}^{n}\beta\lambda(y_i-\alpha)^2+\beta^3\lambda(y_i-\alpha)^2-\beta^2\lambda x_i(y_i-\alpha)-\beta^3(y_i-\alpha)^2\\&=-\beta^2\lambda\left(\sum_{i=1}^{n}x_iy_i\right)-\beta\lambda^2\left(\sum_{i=1}^{n}x_i^2\right)+\lambda^2\left(\sum_{i=1}^{n}x_iy_i\right)\\&+\beta^2\lambda\alpha\left(\sum_{i=1}^{n}x_i\right)+\beta\lambda\left(\sum_{i=1}^{n}(y_i-\alpha)^2\right)-\lambda^2\alpha\left(\sum_{i=1}^{n}x_i\right)\end{aligned} \tag{14.1.353}$$

Dividing with λ and using the fact that:

$$\alpha=\overline{y}-\overline{x}\beta \tag{14.1.354}$$

it is seen that:

$$
\begin{aligned}
0 &= -\beta^2\Big(\sum_{i=1}^{n} x_i y_i\Big) - \beta\lambda\Big(\sum_{i=1}^{n} x_i^2\Big) + \lambda\Big(\sum_{i=1}^{n} x_i y_i\Big) + \beta^2(\overline{y}-\overline{x}\beta)\Big(\sum_{i=1}^{n} x_i\Big) \\
&+ \beta\Big(\sum_{i=1}^{n} (y_i - (\overline{y}-\overline{x}\beta))^2\Big) - \lambda(\overline{y}-\overline{x}\beta)\Big(\sum_{i=1}^{n} x_i\Big) \\
&= -\beta^2\Big(\sum_{i=1}^{n} x_i y_i\Big) - \beta\lambda\Big(\sum_{i=1}^{n} x_i^2\Big) + \lambda\Big(\sum_{i=1}^{n} x_i y_i\Big) + \beta^2\overline{y}\Big(\sum_{i=1}^{n} x_i\Big) \\
&- \beta^3\overline{x}\beta\Big(\sum_{i=1}^{n} x_i\Big) + \beta\Big(\sum_{i=1}^{n} y_i^2\Big) + \beta\Big(\sum_{i=1}^{n} (\overline{y}-\overline{x}\beta)^2\Big) \\
&- 2\beta\Big(\sum_{i=1}^{n} y_i(\overline{y}-\overline{x}\beta)\Big) - \lambda\overline{y}\Big(\sum_{i=1}^{n} x_i\Big) + \lambda\overline{x}\beta\Big(\sum_{i=1}^{n} x_i\Big)
\end{aligned}
\tag{14.1.355}
$$

Splitting up the sums even more gives:

$$
\begin{aligned}
0 = &-\beta^2\Big(\sum_{i=1}^{n} x_i y_i\Big) - \beta\lambda\Big(\sum_{i=1}^{n} x_i^2\Big) + \lambda\Big(\sum_{i=1}^{n} x_i y_i\Big) + \beta^2\overline{y}\Big(\sum_{i=1}^{n} x_i\Big) - \beta^3\overline{x}\beta\Big(\sum_{i=1}^{n} x_i\Big) \\
&+ \beta\Big(\sum_{i=1}^{n} y_i^2\Big) + \beta\Big(\sum_{i=1}^{n} \overline{y}^2\Big) + \beta\Big(\sum_{i=1}^{n} (\overline{x}.\beta)^2\Big) - 2\beta\Big(\sum_{i=1}^{n} \overline{yx}\beta\Big) - 2\beta\Big(\sum_{i=1}^{n} y_i\overline{y}\Big) \\
&+ 2\beta\Big(\sum_{i=1}^{n} y_i\overline{x}\beta\Big) - \lambda\overline{y}\Big(\sum_{i=1}^{n} x_i\Big) + \lambda\overline{x}\beta\Big(\sum_{i=1}^{n} x_i\Big)
\end{aligned}
\tag{14.1.356}
$$

Finally the terms are sorted and collected according to powers of β:

$$
\begin{aligned}
0 = &\beta^3\left(\sum_{i=1}^{n} \overline{x}^2 - \overline{x}\sum_{i=1}^{n} x_i\right) \\
&+ \beta^2\left(\overline{y}\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i - 2\sum_{i=1}^{n} \overline{yx} + 2\sum_{i=1}^{n} y_i\overline{x}\right) \\
&+ \beta\left(\sum_{i=1}^{n} y_i^2 - \lambda\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} \overline{y}^2 - 2\sum_{i=1}^{n} y_i\overline{y} + \lambda\overline{x}\sum_{i=1}^{n} x_i\right) \\
&+ \lambda\left(\sum_{i=1}^{n} x_i y_i - \overline{y}\sum_{i=1}^{n} x_i\right)
\end{aligned}
\tag{14.1.357}
$$

Since:

- $\sum_{i=1}^{n}\overline{x}^2-\overline{x}\sum_{i=1}^{n}x_i=0$
- $\overline{y}\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}x_iy_i-2\sum_{i=1}^{n}\overline{yx}+2\sum_{i=1}^{n}y_i\overline{x}=-\text{SPD}_{xy}$
- $\sum_{i=1}^{n}y_i^2-\lambda\sum_{i=1}^{n}x_i^2+\sum_{i=1}^{n}\overline{y}^2-2\sum_{i=1}^{n}y_i\overline{y}+\lambda\overline{x}\sum_{i=1}^{n}x_i=\text{SSD}_y-\lambda\text{SSD}_x$
- $\sum_{i=1}^{n}x_iy_i-\overline{y}\sum_{i=1}^{n}x_i=\text{SPD}_{xy}$

it is clear that the derivation of β comes down to solve:

$$-\beta^2\text{SPD}_{xy}+\beta(\text{SSD}_y-\lambda\text{SSD}_x)+\lambda\text{SPD}_{xy}=0 \tag{14.1.358}$$

For $\text{SPD}_{xy}\neq 0$ this implies that:

$$\begin{aligned}\beta&=\frac{-(\text{SSD}_y-\lambda\text{SSD}_x)\pm\sqrt{(\text{SSD}_y-\lambda\text{SSD}_x)^2-4(-\text{SPD}_{xy})\lambda\text{SPD}_{xy}}}{-2\text{SPD}_{xy}}\\&=\frac{\text{SSD}_y-\lambda\text{SSD}_x\pm\sqrt{(\text{SSD}_y-\lambda\text{SSD}_x)^2+4\lambda\text{SPD}_{xy}^2}}{2\text{SPD}_{xy}}\end{aligned} \tag{14.1.359}$$

Since:

$$\text{SSD}_y-\lambda\text{SSD}_x\leq\sqrt{(\text{SSD}_y-\lambda\text{SSD}_x)^2+4\lambda\text{SPD}_{xy}^2} \tag{14.1.360}$$

there is always a positive and a negative solution to (**??**). The desired solution should always have the same sign as SPD_{xy}, hence the solution with the positive numerator is selected. Therefore:

$$\hat{\beta}=\frac{\text{SSD}_y-\lambda\text{SSD}_x+\sqrt{(\text{SSD}_y-\lambda\text{SSD}_x)^2+4\lambda\text{SPD}_{xy}^2}}{2\text{SPD}_{xy}} \tag{14.1.361}$$

Let us solve for ξ_i - again... With estimates for β and α it is now possible to estimate ξ_i using:

$$\xi_i=\frac{\lambda x_i+\beta(y_i-\alpha)}{\lambda+\beta^2} \tag{14.1.362}$$

we get:

$$\hat{\xi}_i=\frac{\lambda x_i+\hat{\beta}(y_i-\hat{\alpha})}{\lambda+\hat{\beta}^2} \tag{14.1.363}$$

Let us now solve for σ^2 by differentiating $\mathcal{L}$ with respect to σ^2 gives:

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\sigma^2}&=\frac{\partial}{\partial\sigma^2}\left(-\frac{n}{2}\log(\lambda\sigma^4)-\frac{\sum\limits_{i=1}^{n}(x_i-\xi_i)^2}{2\sigma^2}-\frac{\sum\limits_{i=1}^{n}(y_i-\alpha-\beta\xi_i)^2}{2\lambda\sigma^2}\right)\\&=\frac{-n\sigma^2}{\sigma^4}+\frac{\sum\limits_{i=1}^{n}(x_i-\xi_i)^2}{2\sigma^4}+\frac{\sum\limits_{i=1}^{n}(y_i-\alpha-\beta\xi_i)^2}{2\lambda\sigma^4}\\&=\frac{-2\lambda n\sigma^2+\lambda\sum\limits_{i=1}^{n}(x_i-\xi_i)^2+\sum\limits_{i=1}^{n}(y_i-\alpha-\beta\xi_i)^2}{2\lambda\sigma^4}\end{aligned}\tag{14.1.364}$$

and setting $\frac{\partial\mathcal{L}}{\partial\sigma^2}$ equal to zero yields:

$$\begin{gathered}\frac{\partial\mathcal{L}}{\partial\sigma^2}=0\Rightarrow-2\lambda n\sigma^2+\lambda\sum_{i=1}^{n}(x_i-\xi_i)^2+\sum_{i=1}^{n}(y_i-\alpha-\beta\xi_i)^2=0\\\Rightarrow\sigma^2=\frac{\lambda\sum\limits_{i=1}^{n}(x_i-\xi_i)^2+\sum\limits_{i=1}^{n}(y_i-\alpha-\beta\xi_i)^2}{2\lambda n}\end{gathered}\tag{14.1.365}$$

To get a central estimate of σ^2 one must divide by $n-2$ instead of $2n$ since there are $n+2$ parameters to be estimated, namely $\xi_1,\xi_2,\ldots,\xi_n,\alpha$ and β. Hence the degrees of freedom are $2n-(n+2)=n-2$. Therefore:

$$\hat{\sigma}^2=\frac{\lambda\sum\limits_{i=1}^{n}(x_i-\hat{\xi}_i)^2+\sum\limits_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}\hat{\xi}_i)^2}{2\lambda(n-2)}\tag{14.1.366}$$

Finally summing up:

$$\boxed{\begin{aligned}&\hat{\alpha}=\overline{y}-\overline{x}\hat{\beta}\\&\hat{\beta}=\frac{\text{SSD}_y-\lambda\text{SSD}_x+\sqrt{(\text{SSD}_y-\lambda\text{SSD}_x)^2+4\lambda\text{SPD}_{xy}^2}}{2\text{SPD}_{xy}}\\&\hat{\sigma}=\sqrt{\frac{\lambda\sum\limits_{i=1}^{n}(x_i-\hat{\xi}_i)^2+\sum\limits_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}\hat{\xi}_i)^2}{2\lambda(n-2)}}\\&\hat{\xi}_i=\frac{\lambda x_i+\hat{\beta}(y_i-\hat{\alpha})}{\lambda+\hat{\beta}^2}\end{aligned}}\tag{14.1.367}$$

These results are implemented in Minitab 16.1.2 and also in the `Deming()` function in the MethComp package of `R` as the ready can see it in the corresponding companion books.

14.1.11 Multiple linear regression Gaussian Model

Of course, in some situations, linear regression is too simple or just not suitable. The most typical case that will concern us now in what follows are the situations where we have several explanatory variables (multivariate case!).

The idea of multiple linear regression is relatively simple. We want to determine the dependent variable y from $p-1$ independent variables (i.e. in the absence of "colinearity"!) - and therefore of p parameters to determine - connected by a linear relation of the general form:

$$y=\beta_0+\beta_1 x+\ldots+\beta_{p-1}x_{p-1}+\varepsilon \tag{14.1.368}$$

In a sample of n individuals we measure $y_i, x_{i,1},\ldots,x_{i,p-1}$ for $i=1\ldots n$:

Observation	y_i	$x_{i,1}$	$\ldots$	$x_{i,p-1}$
1	y_1	$x_{1,1}$	$\ldots$	$x_{1,p-1}$
2	y_2	$x_{2,1}$	$\ldots$	$x_{2,p-1}$
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\ldots$
n	y_n	$x_{n,1}$	$\ldots$	$x_{n,p-1}$

Remark

The terms "multivariate" and "multivariable" are often used interchangeably in the statistical literature. However, these terms actually represent two very distinct types of analyses! A "multivariable model" can be thought of as a model in which multiple variables are found on the right side of the model equation (one outcome, multiple predictors). "Multivariate", by contrast, refers to the modelling of data that have multiple output variables for multiple variables on the right side of the model equation (multiple outcomes, modeled jointly). Unfortunately "multivariable regression" is often mistakenly named "multivariate regression", or vice versa.

In facts, to estimate the parameters $\beta_0,\ldots,\beta_{p-1}$ (estimated values which we denote by $\hat{\beta}_0,\ldots,\hat{\beta}_{p-1}$ to respect traditions) the approach is very simple because it is just a generalization of the method of least squares we saw earlier for the simple univariate linear regression.

So in the end we rewrite the relation of Sum of Squared Residuals seen above by modifying a little bit because we now have the multi-linear stuff:

$$\text{SSR}=\sum_{i=1}^{k}(y_i-y_i')^2 \tag{14.1.369}$$

with the estimated theoretical model:

$$y_i'=\hat{\beta}_0+\hat{\beta}_1 x_{i,1}+\ldots+\hat{\beta}_{p-1}x_{i,p-1} \tag{14.1.370}$$

So we have to minimize:

$$\text{SSR}=\sum_{i=1}^{k}(y_i-(\hat{\beta}_0+\hat{\beta}_1 x_{i,1}+\ldots+\hat{\beta}_{p-1}x_{i,p-1}))^2 \tag{14.1.371}$$

The above parentheses can be rewritten in the form (using what we learned in the sections of Vector Calculus and Linear Algebra):

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{bmatrix} 1 & x_{1,1} & \dots & x_{1,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p-1} \end{bmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_0 \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{14.1.372}$$

Whose condensed form is:

$$\vec{y} = X\vec{\hat{\beta}} + \vec{\varepsilon} \tag{14.1.373}$$

where X is named the "designed matrix", also known as "regressor matrix" or "model matrix" or "data matrix".

We then have the vector of residues that can be written:

$$\vec{\varepsilon} = \vec{y} - \vec{y}' = \vec{y} - X\vec{\hat{\beta}} \tag{14.1.374}$$

As we know, the least squares method is to find the vector $\vec{\hat{\beta}}$ that minimizes:

$$||\vec{\varepsilon}||^2 = \vec{\varepsilon} \circ \vec{\varepsilon} = \vec{\varepsilon}^T \circ \vec{\varepsilon} \tag{14.1.375}$$

Therefore explicitly:

$$||\vec{\varepsilon}||^2 = \vec{\varepsilon}^T \circ \vec{\varepsilon} = (\vec{y} - X\vec{\hat{\beta}})^T \circ (\vec{y} - X\vec{\hat{\beta}}) = \vec{y}^T\vec{y} - \vec{y}^T(X\vec{\hat{\beta}}) - (X\vec{\hat{\beta}})^T\vec{y} + (X\vec{\hat{\beta}})^T(X\vec{\hat{\beta}}) \tag{14.1.376}$$

Notice that we have:

$$(X\vec{\hat{\beta}})^T(X\vec{\hat{\beta}}) = (X\vec{\hat{\beta}})(X\vec{\hat{\beta}})^T \tag{14.1.377}$$

and:

$$\vec{y}^T(X\vec{\hat{\beta}}) = (X\vec{\hat{\beta}})^T\vec{y} \tag{14.1.378}$$

as each of the elements of the multiplication is a simple vector!

Therefore we have (caution! do not forget that some multiplications in the relation that will follow are dot products!!!) the quadratic multivariate function of coefficients of vectors:

$$||\vec{\varepsilon}||^2 = \vec{\varepsilon} \circ \vec{\varepsilon} = (\vec{y} - X\vec{\hat{\beta}}) \circ (\vec{y} - X\vec{\hat{\beta}}) = \vec{y} \circ \vec{y} - 2(X\vec{\hat{\beta}}) \circ \vec{y} + (X\vec{\hat{\beta}}) \circ (X\vec{\hat{\beta}}) = F \tag{14.1.379}$$

Let us derivate this latter "object function F" at the order of vector $\hat{\beta}$ (it is as internal derivative component by component). What we will write:

$$\begin{aligned} \frac{\partial F}{\partial \vec{\hat{\beta}}} &= \frac{\partial}{\partial \vec{\hat{\beta}}} \left(\vec{y} \circ \vec{y} - 2(X\vec{\hat{\beta}}) \circ \vec{y} + (X\vec{\hat{\beta}}) \circ (X\vec{\hat{\beta}}) \right) \\ &= \frac{\partial}{\partial \vec{\hat{\beta}}} \left(-2(X\vec{\hat{\beta}}) \circ \vec{y} + (X\vec{\hat{\beta}}) \circ (X\vec{\hat{\beta}}) \right) \\ &= -2(X\vec{u}) \circ \vec{y} + \frac{\partial}{\partial \vec{\hat{\beta}}} \left((X\vec{\hat{\beta}}) \circ (X\vec{\hat{\beta}}) \right) \\ &= -2(X\vec{u}) \circ \vec{y} + \frac{\partial (X\vec{\hat{\beta}})}{\partial \vec{\hat{\beta}}} \circ (X\vec{\hat{\beta}}) + (X\vec{\hat{\beta}}) \circ \frac{\partial (X\vec{\hat{\beta}})}{\partial \vec{\hat{\beta}}} \\ &= -2(X\vec{u}) \circ \vec{y} + (X\vec{u}) \circ (X\vec{\hat{\beta}}) + (X\vec{\hat{\beta}}) \circ (X\vec{u}) \\ &= -2(X\vec{u}) \circ \vec{y} + 2(X\vec{u}) \circ (X\vec{\hat{\beta}}) \end{aligned} \tag{14.1.380}$$

Now let us rewrite this form a vector notation to a pure matrix notation:

$$\frac{\partial F}{\partial \vec{\hat{\beta}}} = -2(X\vec{u})^T\vec{y} + 2(X\vec{u})^T(X\vec{\hat{\beta}}) \tag{14.1.381}$$

We seek the $\vec{\hat{\beta}}$ that cancel this derivative. Therefore we must solve the following equation:

$$-2(X\vec{u})^T\vec{y} + 2(X\vec{u})^T(X\vec{\hat{\beta}}) = 0 \tag{14.1.382}$$

Therefore:

$$(X\vec{u})^T(X\vec{\hat{\beta}}) = (X\vec{u})^T\vec{y} \tag{14.1.383}$$

Remember before we continue that:

$$(X\vec{u})^T = \vec{u}^T X^T \tag{14.1.384}$$

Therefore the prior-previous relation can be written:

$$(\vec{u}^T X^T)(X\vec{\hat{\beta}}) = (\vec{u}^T X^T)\vec{y} \tag{14.1.385}$$

As Linear Algebra is associative, let us write without the parenthesis:

$$\vec{u}^T X^T X\vec{\hat{\beta}} = \vec{u}^T X^T\vec{y} \tag{14.1.386}$$

We can obviously not simplify right and left by $\vec{u}^T X^T$ as it is not a squared matrix, this term is then by obligation not reversible. The only one thing we can do is identify the elements such that:

$$\vec{u}^T(X^T X\vec{\hat{\beta}}) = \vec{u}^T(X^T\vec{y}) \tag{14.1.387}$$

this implies obviously:

$$X^T X\vec{\hat{\beta}} = X^T\vec{y} \tag{14.1.388}$$

We find then that if the squared matrix $X^T X$ is reversible then:

$$\boxed{\vec{\hat{\beta}} = \left(X^T X\right)^{-1} X^T\vec{y}} \tag{14.1.389}$$

The matrix $X^T X$ that we will see again in the field of multiple linear regression and in the field of design of experiments (see section Industrial Engineering page 8780) is named "information matrix" or "dispersion matrix" for a reason that will be very obvious later.

The expression $(X^T X)^{-1}X^T$ is named the "left pseudo-inverse", or simply "pseudoinverse", and even sometimes "Moore-Penrose pseudoinverse" of X and is denoted X^- in some textbooks.

Remark

We say that the regression is "balanced and orthogonal" when the information matrix is diagonal. We say that the regression is "orthogonal" when the sub-matrix of the information matrix excluding the first row and first columns is orthogonal. We say that the regression is just "balanced" when all the values of the first row and of the first column excepted the one at the intersection are equal to zero.

To show that this seem correct a priori, let us fall back on the results we get of the simple univariate regression:

$$y=\hat{\beta}_0+\hat{\beta}_1 x_1 \tag{14.1.390}$$

Then supposing 2 observations, we have therefore:

$$\begin{aligned}\vec{\hat{\beta}}&=\left(X^TX\right)^{-1}X^T\vec{y}\\&=\left(\begin{bmatrix}1 & x_{1,1}\\1 & x_{2,1}\end{bmatrix}^T\begin{bmatrix}1 & x_{1,1}\\1 & x_{2,1}\end{bmatrix}\right)^{-1}\begin{bmatrix}1 & x_{1,1}\\1 & x_{2,1}\end{bmatrix}^T\begin{pmatrix}y_1\\y_2\end{pmatrix}\\&=\left(\begin{bmatrix}1 & 1\\x_{1,1} & x_{2,1}\end{bmatrix}^T\begin{bmatrix}1 & x_{1,1}\\1 & x_{2,1}\end{bmatrix}\right)^{-1}\begin{bmatrix}1 & 1\\x_{1,1} & x_{2,1}\end{bmatrix}^T\begin{pmatrix}y_1\\y_2\end{pmatrix}\\&=\left(\begin{bmatrix}1\cdot 1+1\cdot 1 & x_{1,1}+x_{2,1}\\x_{1,1}+x_{2,1} & x_{1,1}^2+x_{2,1}^2\end{bmatrix}\right)^{-1}\begin{pmatrix}y_1+y_2\\x_{1,1}y_1+x_{2,1}y_2\end{pmatrix}\\&=\left(\begin{bmatrix}n & \sum x_i\\\sum x_i & \sum x_i^2\end{bmatrix}\right)^{-1}\begin{pmatrix}\sum y_i\\\sum x_iy_i\end{pmatrix}\end{aligned} \tag{14.1.391}$$

Using the relation proved in the section of Linear Algebra (see page 1980) to calculate in generality the inverse of a matrix $A(a_{ij})$ in $A^{-1}(b_{ij})$:

$$b_{ij}=\frac{1}{\det(A)}(-1)^{i+j}\det(A_{ji}) \tag{14.1.392}$$

We have in our special case:

$$\boxed{\det(X^TX)=a_{11}a_{22}-a_{12}a_{21}=n\sum x_i^2-\sum x_i\sum x_i=n\sum x_i^2-\left(\sum x_i\right)^2} \tag{14.1.393}$$

Therefore we have:

$$\begin{aligned}(X^TX)^{-1}&=\frac{1}{n\sum x_i^2-\left(\sum x_i\right)^2}(-1)^{i+j}\det(X^TX_{ji})\\&=\frac{1}{n\sum x_i^2-\left(\sum x_i\right)^2}\begin{bmatrix}\det(X^TX_{11}) & -\det(X^TX_{21})\\-\det(X^TX_{12}) & \det(X^TX_{22})\end{bmatrix}\end{aligned} \tag{14.1.394}$$

and since we have a square matrix of dimension 2 only, the calculation of the four determinants is reduced to selecting the components of X^TX (see section Linear Algebra page 1966):

$$(X^TX)^{-1}=\frac{1}{n\sum x_i^2-\left(\sum x_i\right)^2}\begin{bmatrix}\sum x_i^2 & -\sum x_i\\-\sum x_i & n\end{bmatrix} \tag{14.1.395}$$

Therefore:

$$
\begin{aligned}
\vec{\hat{\beta}} &= \left(\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}\right)^{-1} \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} \\
&= \frac{1}{n\sum x_i^2 - \left(\sum x_i\right)^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix} \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} \\
&= \frac{1}{n\sum x_i^2 - \left(\sum x_i\right)^2} \begin{bmatrix} \sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i \\ -\sum x_i \sum y_i + n \sum x_i y_i \end{bmatrix} \\
&= \begin{bmatrix} \dfrac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \\ \dfrac{-\sum x_i \sum y_i + n \sum x_i y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} b \\ a \end{bmatrix}
\end{aligned}
\tag{14.1.396}
$$

So to a change in notations for the indices and experimental measurements, we fall back on the results that we obtained during our study of the simple linear univariate regression that was (for refresh...):

$$
a = \frac{n\sum_{k=1}^{n} x_k y_k - \sum_{k=1}^{n} x_k \sum_{k=1}^{n} y_k}{n\sum_{k=1}^{n} x_k^2 - \left(\sum_{k=1}^{n} x_k\right)^2} \qquad b = \frac{\sum_{k=1}^{n} x_k^2 \sum_{k=1}^{n} y_k - \sum_{k=1}^{n} x_k \sum_{k=1}^{n} x_k y_k}{n\sum_{k=1}^{n} x_k^2 - \left(\sum_{k=1}^{n} x_k\right)^2}
\tag{14.1.397}
$$

Now, we need a quality indicator regarding our multi-linear regression. Remember that in the context of our study of the univariate linear regression, we proved that the linear correlation coefficient can be written as (see page 6333):

$$
R_{X,Y}^2 = \frac{\text{SSE}}{\text{SST}} = \frac{\sum_k (\hat{y}_k - \bar{y})^2}{\sum_k (y_k - \bar{y})^2}
\tag{14.1.398}
$$

and in fact it also applies directly to the multiple linear regression, since it does not presuppose the number of explanatory variables!!

But it is the tradition to rewrite that latter in matrix form. So at it is beautiful, let us do the work.

First we start with:

$$
\text{SSE} = \sum_k (\hat{y}_k - \bar{y})^2 = (\vec{\hat{Y}} - \vec{\bar{Y}})^T (\vec{\hat{Y}} - \vec{\bar{Y}}) = (X\vec{\hat{\beta}} - \vec{\bar{Y}})^T (X\vec{\hat{\beta}} - \vec{\bar{Y}}) = \vec{\hat{\beta}}^T (X^T X)\vec{\hat{\beta}} - \vec{\bar{Y}}^T \vec{\bar{Y}} - 2\vec{\hat{\beta}} X^T \vec{\bar{Y}}
\tag{14.1.399}
$$

and finally:

$$
\text{SST} = \sum_k (y_k - \bar{y})^2 = (\vec{Y} - \vec{\bar{Y}})^T (\vec{Y} - \vec{\bar{Y}}) = \vec{Y}^T \vec{Y} + \vec{\bar{Y}}^T \vec{\bar{Y}} - 2\vec{Y}^T \vec{\bar{Y}}
\tag{14.1.400}
$$

Hence:

$$R^2=\frac{\text{SSE}}{\text{SST}}=\frac{\vec{\hat{\beta}}^T(X^TX)\vec{\hat{\beta}}-\vec{\bar{Y}}^T\vec{\bar{Y}}-2\vec{\hat{\beta}}X^T\vec{\bar{Y}}}{\vec{Y}^T\vec{Y}+\vec{\bar{Y}}^T\vec{\bar{Y}}-2\vec{Y}^T\vec{\bar{Y}}} \qquad (14.1.401)$$

Remark

The reader interested in the practical application of these results may, just as for simple univariate regression, refer to the server exercises on the companion website - Numerical Methods section - where there are practical examples with Microsoft® Excel®.

Obviously with multiple linear regression, we can now, with a small trick, do linear regression ... of polynomials (we will see later how to apply directly the ordinary least squares method on a polynomial) too! Indeed, consider a polynomial of the form:

$$y=ax^3+bx^2+cx+d \qquad (14.1.402)$$

That we can consider, when rewritten, as:

$$y=az+bw+cx+d \qquad (14.1.403)$$

So in spreadsheet softwares like Microsoft® Excel®, just use the Regression Analysis Tool with the input variable x column, a second column that we will have taken care to create with inside the square of x (i.e. x^2) and which will be considered as the explanatory variable w and a third column we have also taken care to create as the cube of x (i.e. x^3) and which will be considered as the explanatory variable z.

More formally, this is like solving the following system:

$$\begin{bmatrix} y_1\\ y_2\\ y_3\\ \vdots\\ y_n \end{bmatrix}=\begin{bmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^m\\ 1 & x_2 & x_2^2 & \ldots & x_2^m\\ 1 & x_3 & x_3^2 & \ldots & x_3^m\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & x_n & x_n^2 & \ldots & x_n^m \end{bmatrix}\begin{bmatrix} \beta_0\\ \beta_1\\ \beta_2\\ \vdots\\ \beta_m \end{bmatrix}+\begin{bmatrix} \varepsilon_1\\ \varepsilon_2\\ \varepsilon_3\\ \vdots\\ \varepsilon_n \end{bmatrix} \qquad (14.1.404)$$

We can also directly obtain the polynomials coefficients using functions already presented earlier above with Microsoft® Excel® (but you will not have all the relevant results of the Analysis Tool). For example for a polynomial of second degree:

```
a:  =INDEX(LINEST(y,x^{1,2}),1)
b:  =INDEX(LINEST(y,x^{1,2}),1,2)
c:  =INDEX(LINEST(y,x^{1,2}),1,3)
```

and for a third-degree polynomial:

```
a:  =INDEX(LINEST(y,x^{1,2,3}),1)
b:  =INDEX(LINEST(y,x^{1,2,3}),1,2)
c:  =INDEX(LINEST(y,x^{1,2,3}),1,3)
d:  =INDEX(LINEST(y,x^{1,2,3}),1,4)
```

The same kind of method also works for Fourier series (see section Sequences and Series page 1727):

$$y=a_0+\sum_{i=1}^{m} a_i\cos(i\omega x)+b_i\sin(i\omega x) \tag{14.1.405}$$

More formally, this is like solving the following system:

$$\begin{bmatrix} y_1\\ y_2\\ y_3\\ \vdots\\ y_n \end{bmatrix}=\begin{bmatrix} 1 & \cos(\omega x_1) & \sin(\omega x_1) & \ldots & \sin(m\omega x_1)\\ 1 & \cos(\omega x_2) & \sin(\omega x_2) & \ldots & \sin(m\omega x_2)\\ 1 & \cos(\omega x_3) & \sin(\omega x_3) & \ldots & \sin(m\omega x_3)\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & \cos(\omega x_n) & \sin(\omega x_n) & \ldots & \sin(m\omega x_n) \end{bmatrix}\begin{bmatrix} \beta_0\\ \beta_1\\ \beta_2\\ \vdots\\ \beta_m \end{bmatrix}+\begin{bmatrix} \varepsilon_1\\ \varepsilon_2\\ \varepsilon_3\\ \vdots\\ \varepsilon_n \end{bmatrix} \tag{14.1.406}$$

This is this trick that allows us to understand why and how the linear correlation coefficient also applies to polynomials and Fourier series in most spreadsheets softwares and statistical softwares. However we can have a more direct approach that does not require this transformation but which therefore is a little longer.

Now let us go back to the Gaussian linear model with a concept a little more rigorous and adapted to the multi-linear case and in particular to highlight the distinction between estimators of the slope of the regression and exact values:

$$Y_i\sim\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}\left(\left(\hat{\beta}_0+\hat{\beta}_1x_{i,1}+\ldots+\hat{\beta}_{p-1}x_{i,p-1}\right)-\mu_i\right)^2}=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}(y_i-\mu_i)^2} \tag{14.1.407}$$

But under this notation convention we have:

$$Y_i\sim\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}(y_i-(\beta_0+\beta_1x_{i,1}+\ldots+\beta_{p-1}x_{i,p-1}))^2} \tag{14.1.408}$$

and written in vector form:

$$Y_i\sim\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}\left(y_i-\vec{x}^T\vec{\beta}\right)^2} \tag{14.1.409}$$

Now, we use the technique of maximum likelihood (see section Statistics page 818) and we seek the coefficients that maximizes therefore:

$$L\left(\vec{\beta},\sigma\right)=\prod_{i=1}^{n}L(y_i,\vec{\beta},\sigma)=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\vec{x}_i^T\vec{\beta}\right)^2} \tag{14.1.410}$$

That we can write in matrix form:

$$L\left(\vec{\beta},\sigma\right)=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\vec{x}_i^T\vec{\beta}\right)^2}=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\left(\vec{y}-X\vec{\beta}\right)^T\left(\vec{y}-X\vec{\beta}\right)} \tag{14.1.411}$$

Taking as in the section of Statistics the log-likelihood to facilitate future calculations and using the property of the transposed matrices proved in section of Linear Algebra, we get:

$$\begin{aligned}\ln\left(L\left(\vec{\beta},\sigma\right)\right)&=\ln\left(\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\left(\vec{y}-X\vec{\beta}\right)^T\left(\vec{y}-X\vec{\beta}\right)}\right)\\&=n\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right)+\ln\left(e^{-\frac{1}{2\sigma^2}\left(\vec{y}-X\vec{\beta}\right)^T\left(\vec{y}-X\vec{\beta}\right)}\right)\\&=n\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right)-\frac{1}{2\sigma^2}\left(\vec{y}-X\vec{\beta}\right)^T\left(\vec{y}-X\vec{\beta}\right)\\&=-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln(\sigma^2)-\frac{1}{2\sigma^2}\left(\vec{y}-X\vec{\beta}\right)^T\left(\vec{y}-X\vec{\beta}\right)\\&=-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln(\sigma^2)-\frac{1}{2\sigma^2}\left(\vec{y}^T\vec{y}-2\vec{\beta}^TX^T\vec{y}+\vec{\beta}^TX^TX\vec{\beta}\right)\\&=-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln(\sigma^2)-\frac{1}{2\sigma^2}\text{RSS}\end{aligned}\tag{14.1.412}$$

Now let us look to the expression of $\vec{\beta}$ that maximizes the log-likelihood. It comes then:

$$\begin{aligned}\frac{\partial\ln\left[L(\vec{\beta},\sigma)\right]}{\partial\vec{\beta}}&=\frac{1}{2\sigma^2}\left[\frac{\partial\left(\vec{y}^T\vec{y}-2\vec{\beta}^TX^T\vec{y}+\vec{\beta}^TX^TX\vec{\beta}\right)}{\partial\vec{\beta}}\right]\\&=\frac{1}{2\sigma^2}\left[\frac{\partial\left(-2\vec{\beta}^TX^T\vec{y}+\vec{\beta}^TX^TX\vec{\beta}\right)}{\partial\vec{\beta}}\right]=\frac{1}{2\sigma^2}\left[-2\frac{\partial\vec{\beta}^T}{\partial\vec{\beta}}X^T\vec{y}+\frac{\partial}{\partial\vec{\beta}}\left(\vec{\beta}^TX^T\vec{\beta}\right)\right]\\&=\frac{1}{2\sigma^2}\left[-2\left(\frac{\partial\vec{\beta}}{\partial\vec{\beta}}\right)^TX^T\vec{y}+\frac{\partial\vec{\beta}^T}{\partial\vec{\beta}}X^TX\vec{\beta}+\vec{\beta}^TX^TX\frac{\partial\vec{\beta}}{\partial\vec{\beta}}\right]\\&=\frac{1}{2\sigma^2}\left[-2[1\dots1]X^T\vec{y}+[1\dots1]X^TX\vec{\beta}+\vec{\beta}^TX^TX\begin{pmatrix}1\\\vdots\\1\end{pmatrix}\right]\end{aligned}\tag{14.1.413}$$

Let us use the property of the transposed matrix proved in the section of Linear Algebra (see page 1951):

$$A^TB^TCD=D^TC^TBA\tag{14.1.414}$$

Therefore we have:

$$\begin{aligned}\frac{\partial\ln\left[L(\vec{\beta},\sigma)\right]}{\partial\vec{\beta}}&=\frac{1}{2\sigma^2}\left[-2[1\dots1]X^T\vec{y}+2\vec{\beta}^TX^TX\begin{pmatrix}1\\\vdots\\1\end{pmatrix}\right]\\&=\frac{1}{2\sigma^2}\left[-2[1\dots1]X^T\vec{y}+2[1\dots1]X^TX\vec{\beta}\right]\end{aligned}\tag{14.1.415}$$

For reasons which will appear evident a little further below we chose the second equivalence. Therefore, we have (and remembering that seeking for an optimum is equivalent to have partial derivative equal to zero):

$$\frac{\partial\ln\left[L(\vec{\beta},\sigma)\right]}{\partial\vec{\beta}}=\frac{[1\dots1]}{\sigma^2}\left[-X^T\vec{y}+X^TX\vec{\beta}\right]\tag{14.1.416}$$

Therefore:

$$-X^T\vec{y}+X^TX\vec{\beta}=0 \qquad (14.1.417)$$

That is to say, after rearrangement:

$$\vec{\beta}=(X^TX)^{-1}X^T\vec{y} \qquad (14.1.418)$$

That is to say exactly the same expressions that we got just a little earlier above with the least squares method in the multi-linear case and that was for reminder:

$$\hat{\vec{\beta}}=(X^TX)^{-1}X^T\vec{y} \qquad (14.1.419)$$

This shows that the statistical of the linear model by the maximum likelihood can fall back into the multi-linear case (and thus also the univariate case) on the results of the least squares method. This has something almost divine ...

Finally, let us indicate that we find here the almost famous "hat matrix" or "influence matrix" H that connects alone all information between the real and explained theoretical values (and depending only on X!) and therefore verbatim the error of the theoretical model:

$$\boxed{\hat{\vec{Y}}=X\hat{\vec{\beta}}=X(X^TX)^{-1}X^T\vec{y}=H\vec{y}} \qquad (14.1.420)$$

We have then something interesting to observe:

$$\hat{Y}_i=H_i\vec{y}_i=\sum_{j=1}^{n}H_{ij}y_i \qquad (14.1.421)$$

By definition, the influence of the observation i in the regression and named the "leverage" (or "leverage score") is defined by:

$$\boxed{H_i=||H_{i.}||^2=\sum_{j=1}^{n}H_{ij}^2} \qquad (14.1.422)$$

H_i is therefore the influence of y_i on its own fitted value! It is a qualitative method (available in many statistical softwares) to judge the influence of points that could be considered as outliers. The idea is to compare the values of leverage between them.

Let us now give the famous explicit expression of the leverage of the simple linear regression.

We start be remembering that by definition:

$$H=X(X^TX)^{-1}X^T \qquad (14.1.423)$$

and we have proved earlier above that in the simple linear regression that we have:

$$X=\begin{bmatrix}1 & x_1\\ 1 & x_2\\ \vdots & \vdots\\ 1 & x_n\end{bmatrix}\Rightarrow X^TX=\begin{bmatrix}n & \displaystyle\sum_{i=1}^{n}x_i\\ \displaystyle\sum_{i=1}^{n}x_i & \sum_{i=1}^{n}x_i^2\end{bmatrix} \qquad (14.1.424)$$

and also:

$$(X^TX)^{-1}=\frac{1}{n\sum\limits_{i=1}^{n}x_i^2-\left(\sum\limits_{i=1}^{n}x_i\right)^2}\begin{bmatrix}\sum\limits_{i=1}^{n}x_i^2 & -\sum\limits_{i=1}^{n}x_i\\ -\sum\limits_{i=1}^{n}x_i & n\end{bmatrix}\tag{14.1.425}$$

Let us put by definition (to make notations more light...):

$$SXX=\sum_{i=1}^{n}x_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\tag{14.1.426}$$

Hence we may rewrite:

$$(X^TX)^1=\frac{1}{SXX}\begin{bmatrix}\frac{1}{n}\sum\limits_{i=1}^{n}x_i^2 & \bar{x}\\ -\bar{x} & 1\end{bmatrix}=\frac{1}{SXX}\begin{bmatrix}\frac{1}{n}\sum\limits_{i=1}^{n}x_i^2-x_1\bar{x} & -\bar{x}+x_1\\ \frac{1}{n}\sum\limits_{i=1}^{n}x_i^2-x_2\bar{x} & -\bar{x}+x_2\\ \vdots & \vdots\\ \frac{1}{n}\sum\limits_{i=1}^{n}x_i^2-x_n\bar{x} & -\bar{x}+x_n\end{bmatrix}\tag{14.1.427}$$

Hence:

$$\begin{aligned}H_{ii}&=\frac{1}{SXX}\left[1\cdot\left(\frac{1}{n}\sum_{j=1}^{n}x_j^2-x_i\bar{x}\right)+x.(-\bar{x}+x_i)\right]\\&=\frac{1}{SXX}\left[\frac{1}{n}\sum_{j=1}^{n}x_j^2-(\bar{x})^2+(\bar{x})^2-2x_i\bar{x}+x_i^2\right]=\frac{1}{SXX}\left[\frac{1}{n}\sum_{j=1}^{n}x_j^2-(\bar{x})^2+(x_i-\bar{x})^2\right]\\&=\frac{1}{nSXX}\left[\sum_{j=1}^{n}x_j^2-\frac{1}{n}\left(\sum_{j=1}^{n}x_j\right)^2\right]+\frac{(x_i-\bar{x})^2}{SXX}\\&=\frac{1}{n}+\frac{(x_i-\bar{x})^2}{SXX}\end{aligned}\tag{14.1.428}$$

Finally:

$$\boxed{H_{ii}=\frac{1}{n}+\frac{(x_i-\bar{x})^2}{\sum\limits_{k=1}^{n}(x_k-\bar{x})^2}}\tag{14.1.429}$$

This relation has a nice interpretation in the SLR model: if the distance from x_i to $\bar{x}$ is large relative to the other x's then H_{ii} will be close to 1.

Leverages have nice mathematical properties, for example, they satisfy:

$$0\leq H_{ii}\leq 1\tag{14.1.430}$$

and their sum is:

$$\sum_{i=1}^{n} H_{ii} = \sum_{i=1}^{n}\left[\frac{1}{n}+\frac{(x_i-\bar{x})^2}{\sum\limits_{k=1}^{n}(x_k-\bar{x})^2}\right] = \frac{n}{n}+\frac{\sum\limits_{i}(x_i-\bar{x})^2}{\sum\limits_{k}(x_k-\bar{x})^2} = 2 \tag{14.1.431}$$

A rule of thumb is to consider leverage values to be large if they are more than double their average size (which is $2/n$ according to the previous relations). So leverages larger than $4/n$ are suspect. Another rule of thumb is to say that values bigger than 0.5 indicate high leverage, while values between 0.3 and 0.5 indicate moderate leverage.

14.1.11.1 Fisher Matrix

We have proved just earlier that:

$$\ln\left(L\left(\vec{\beta},\sigma\right)\right) = -\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln(\sigma^2)-\frac{1}{2\sigma^2}\left(\vec{y}-X_i\vec{\beta}\right)^T\left(\vec{y}-X_i\vec{\beta}\right) \tag{14.1.432}$$

And we have afterwards also prove that (changing the notation just a bit):

$$\frac{\partial\ln\left(L\left(\vec{\beta},\sigma\right)\right)}{\partial\vec{\beta}} = -\frac{1}{2\sigma^2}(2X^TX\vec{\beta}-2X^T\vec{y}) = \frac{1}{\sigma^2}(X^T\vec{y}-X^TX\vec{\beta}) = 0 \tag{14.1.433}$$

And we also get immediately:

$$\frac{\partial\ln\left(L\left(\vec{\beta},\sigma\right)\right)}{\partial\sigma^2} = -\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}(\vec{y}-X\vec{\beta})^T(\vec{y}-X\vec{\beta}) = 0 \tag{14.1.434}$$

From the both relation above we immediately deduce (note that $\hat{\sigma}^2$ is biased!):

$$\hat{\vec{\beta}} = (X^TX)^{-1}\vec{y} \qquad \text{and} \qquad \hat{\sigma}^2 = \frac{1}{n}\left(\vec{y}-X\hat{\vec{\beta}}\right)^T\left(\vec{y}-X\hat{\vec{\beta}}\right) \tag{14.1.435}$$

To compute $\mathcal{I}(\vec{\theta})$, notice that we have immediately:

$$\begin{aligned}
\frac{\partial^2\ln\left(L\left(\vec{\beta},\sigma\right)\right)}{\partial\vec{\beta}\partial\vec{\beta}} &= -\frac{1}{\sigma^2}X^TX \\
\frac{\partial^2\ln\left(L\left(\vec{\beta},\sigma\right)\right)}{\partial\sigma^2\partial\sigma^2} &= \frac{n}{2\sigma^4}-\frac{(\vec{y}-X\vec{\beta})^T(\vec{y}-X\vec{\beta})}{\sigma^6} \\
\frac{\partial^2\ln\left(L\left(\vec{\beta},\sigma\right)\right)}{\partial\sigma^2\partial\vec{\beta}} &= \frac{\partial^2\ln\left(L\left(\vec{\beta},\sigma\right)\right)}{\partial\vec{\beta}\partial\sigma^2} = -\frac{1}{\sigma^4}(X^T\vec{y}-X^TX\vec{\beta}) = -\frac{1}{\sigma^4}X^T(\vec{y}-X\vec{\beta})
\end{aligned} \tag{14.1.436}$$

Since:

$$\begin{aligned}
&\mathrm{E}\left(\frac{\partial^2 \ln\left(L\left(\vec{\beta},\sigma\right)\right)}{\partial\vec{\beta}\partial\vec{\beta}}\right)=\mathrm{E}\left(-\frac{1}{\sigma^2}X^TX\right)=-\frac{1}{\sigma^2}X^TX\\
&\mathrm{E}\left(\frac{\partial^2 \ln\left(L\left(\vec{\beta},\sigma\right)\right)}{\partial\sigma^2\partial\sigma^2}\right)=\frac{n}{2\sigma^4}-\frac{\mathrm{E}\left((\vec{y}-X\vec{\beta})^T(\vec{y}-X\vec{\beta})\right)}{\sigma^6}=\frac{n}{2\sigma^4}-\frac{n\sigma^2}{\sigma^6}=-\frac{n}{2\sigma^4}\\
&\mathrm{E}\left(\frac{\partial^2 \ln\left(L\left(\vec{\beta},\sigma\right)\right)}{\partial\sigma^2\partial\vec{\beta}}\right)=-\frac{1}{\sigma^4}X^T\mathrm{E}\left((\vec{y}-X\vec{\beta})\right)=0
\end{aligned} \tag{14.1.437}$$

Then the expected mean of the Hessian is (see section Statistics page 839):

$$-\mathrm{E}\left(H(\vec{\theta})\right)=\mathrm{E}\left(\frac{\partial^2\mathcal{L}(\vec{\theta})}{\partial\theta_i\partial\theta_j}\right)=\begin{bmatrix}-\dfrac{1}{\sigma^2}X^TX & 0\\ 0 & -\dfrac{n}{2\sigma^4}\end{bmatrix} \tag{14.1.438}$$

Without forgetting that (see section Statistics page 839):

$$\mathcal{I}(\vec{\theta})=-\mathrm{E}\left(H(\vec{\theta})\right)=\begin{bmatrix}\dfrac{1}{\sigma^2}X^TX & 0\\ 0 & \dfrac{n}{2\sigma^4}\end{bmatrix} \tag{14.1.439}$$

And finally:

$$\Sigma(\vec{\theta})\cong-\mathrm{E}(H^{-1})=\begin{bmatrix}\sigma^2(X^TX)^{-1} & 0\\ 0 & \dfrac{2\sigma^4}{n}\end{bmatrix} \tag{14.1.440}$$

It may seems to the reader in a first view that it is a 2×2 matrix. But this is forget that the term in the upper left corner is itself a $p\times p$ (where p is the number of predictors and we put $+1$ because of the intercept). Hence the matrix dimension is $(p+1)\times(p+1)$.

14.1.11.2 R^2 decomposition in multiple regression

Let us see now how to decompose R^2 into components that capture the percentage of variation explained by predictor in a multiple linear regression!

Consider n predictors and K observations in the sample:

$$y_k,x_{1k},\ldots,x_{nk}\qquad \text{for } k=1,\ldots,K \tag{14.1.441}$$

where all variables have been standardized to have mean zero and variance one. That is, we have centered and rescaled the observations such that for $i=1,\ldots,n$:

$$\begin{gathered}
\sum_{k=1}^K y_k=0=\sum_{k=1}^K x_{ik}\\
\frac{1}{K-1}\sum_{k=1}^K y_k^2=1=\frac{1}{K-1}\sum_{k=1}^K x_{ik}^2
\end{gathered} \tag{14.1.442}$$

Standardizing all variables in this manner is without loss of generality since R^2 is manifestly invariant to centering and rescaling of variables. Our multiple linear model is then given by:

$$y_k = \beta_1 x_{1k} + \ldots + \beta_n x_{nk} + \varepsilon_k, \qquad \text{for} \qquad k = 1, \ldots, K \tag{14.1.443}$$

since standardization ensures that the intercept in the regression above is identically zero.

We estimate the slope coefficient and obtain the fitted values:

$$\hat{y}_k = \hat{\beta}_1 x_{1k} + \ldots + \hat{\beta}_n x_{nk} \tag{14.1.444}$$

where $\hat{\beta}_i$ are the estimated slope coefficients for predictors $i = 1, \ldots, n$. Let $\text{cov}(\cdot, \cdot)$ denote the sample covariance operator, defined for centered vectors x and y by:

$$\text{cov}(x,y) = \frac{1}{K-1}\sum_{k=1}^{K} x_k y_k \tag{14.1.445}$$

Then by definition and construction:

$$R^2 = \frac{\sum\limits_{k=1}^{K} \hat{y}_k^2}{\sum\limits_{k=1}^{K} y_k^2} = \frac{\sum\limits_{k=1}^{K} \hat{y}_k^2}{1\cdot(K-1)} = \frac{1}{K-1}\sum_{k=1}^{K} \hat{y}_k^2 = \text{cov}(\hat{y},\hat{y}) = \text{cov}\left(\sum_{i=1}^{n} \hat{\beta}_i x_i, \hat{y}\right) = \sum_{i=1}^{n} \hat{\beta}_i \text{cov}\left(x_i, \hat{y}\right) \tag{14.1.446}$$

That is, we have the decomposition:

$$\boxed{R^2 = \hat{\beta}_1 \text{cov}\left(x_1, \hat{y}\right) + \ldots + \hat{\beta}_n \text{cov}\left(x_n, \hat{y}\right)} \tag{14.1.447}$$

We can therefore define the percentage of variation explained by the predictor i, denoted R_i^2, by:

$$R_i^2 = \hat{\beta}_i \text{cov}\left(x_i, \hat{y}\right) \tag{14.1.448}$$

14.1.11.3 Influential points

An "outlier" as we already know is a data point which is very far, somehow, from the rest of a group of the data. They are often worrisome, but not always a problem. When we are doing regression modelling, in fact, we don't really care about whether some data point is far from a group of data, but whether it breaks a pattern the rest of the data seems to follow and this is what we name an "influential point". We will look here some empirical ways of quantifying how much influence particular data points have on the model.

The points marked in red and blue in the figure below are clearly not like the main cloud of the data points, even though their x and y coordinates are quite typical of the data as a whole: the x coordinates of those points aren't related to the y coordinates in the right way, they break a pattern. On the other hand, the point marked in green, while its coordinates are very weird on both axes, does not break that pattern - it was positioned to fall right on the regression line.

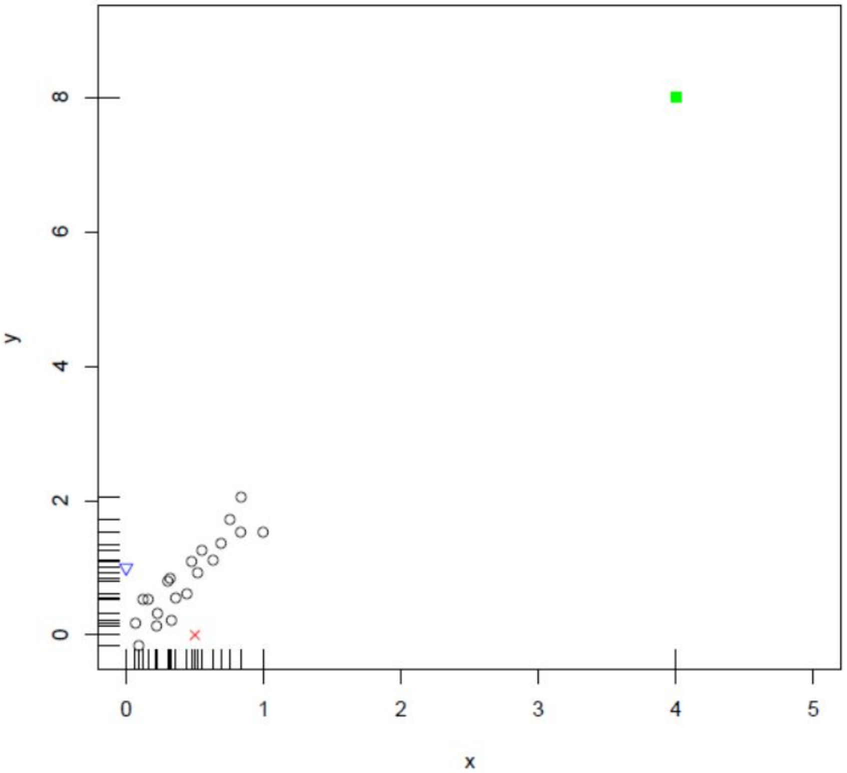


Figure 14.16: **Outliers (red) vs Influential (blue) points**

Be very very very careful with influential points!!! Even if they follow the global pattern, they may quite significantly (and surprisingly) modify the regression equation and the importance of the different coefficients of that same equation (see further below for an example)!

We have seen just earlier the "leverage". But let us introduce other influential indicators that are very common in the 120th century (holocene calendar) and beginning of 121st century in many statistical softwares and textbooks!

The "semistudentized residuals" are defined as:

$$\boxed{e_i^*=\frac{e_i}{\hat{\sigma}_e}=\frac{Y_i-\hat{y}_i}{\sqrt{\dfrac{1}{N-k}\displaystyle\sum_{i=1}^{N}(Y_i-\hat{y}_i)^2}}} \tag{14.1.449}$$

But... Let us recall that we have proved just earlier that:

$$\vec{\hat{Y}}=H\vec{y} \tag{14.1.450}$$

Hence:

$$\vec{e}=\vec{y}-\vec{\hat{Y}}=\vec{y}-H\vec{y}=(\mathbb{1}-H)\vec{y} \tag{14.1.451}$$

Let us rewrite this in term of noise:

$$\vec{e}=(\mathbb{1}-H)\vec{\varepsilon} \tag{14.1.452}$$

Since $\text{E}(\vec{\varepsilon})=\vec{0}$ and $\text{V}(\vec{\varepsilon})=\sigma_{\varepsilon}^2\vec{\mathbb{1}}$, we have:

$$\text{E}(\vec{e})=\text{E}\left[(\mathbb{1}-H)\vec{\varepsilon}\right]=\vec{0} \tag{14.1.453}$$

and:

$$\text{V}(\vec{e})=\text{V}\left[(\mathbb{1}-H)\vec{\varepsilon}\right]=(\mathbb{1}-H)\sigma_{\varepsilon}^2 \tag{14.1.454}$$

What does this imply for the residual at the ith data point? First, and obviously, it has expectation of 0:

$$\text{E}(\varepsilon_i)=0 \tag{14.1.455}$$

and it has variance which depends on i through the hat matrix:

$$\text{V}(e_i)=\sigma_{\varepsilon}^2(\mathbb{1}-H)_{ii}=\sigma_{\varepsilon}^2(1-H_{ii}) \tag{14.1.456}$$

In other words, the bigger the leverage of i, the smaller the variance of the residual is. This is yet another sense in which points with high leverage are points which the model tries very hard to fit.

Previously, when we looked at the residuals, we expected them to all be of roughly the same magnitude. This rests on the leverages H_{ii} being all about the same! If there are substantial variations in leverage across the data points, it's better to scale the residuals by their expected size.

The usual way to do this is through the "standardized residuals" or "internally studentized residuals" or just simply "studentized residuals":

$$\boxed{r_i=\frac{e_i}{\hat{\sigma}_{\varepsilon}\sqrt{1-H_{ii}}}=\frac{Y_i-\hat{y}_i}{\sqrt{\dfrac{1}{N-k}\displaystyle\sum_{i=1}^{N}(Y_i-\hat{y}_i)^2}\sqrt{1-\displaystyle\sum_{j=1}^{N}(X(X^TX)^{-1}X^T)_{ij}}}} \tag{14.1.457}$$

Why "studentized"? Because we're dividing by an estimate of the standard error, just like in Student's T-test for differences in means.

However the single e_i and $\hat{\sigma}_{\varepsilon}$ are non independent, so r_i can't have a T distribution.

Remark

Indeed, let us recall that in the section Statistics we have seen that given Z a random variable of Normal law centered and reduced $\mathcal{N}(0,1)$ and U a variable independent of Z and distributed according to the law of χ^2 at k degrees of freedom. By definition, the variable:

$$T=\frac{Z}{\sqrt{U/k}} \tag{14.1.458}$$

follows a Student law of k degrees of freedom.

The procedure is then to delete the ith observation, fit the regression function to the remaining $n-1$ observations, and get new $\hat{y}$'s which can be denoted by $\hat{y}_{i(i)}$. The difference:

$$d_i = Y_i - \hat{y}_{i(i)} \tag{14.1.459}$$

is named "deleted residual".

Then we define the "studentized deleted residuals" or "externally studentized residuals":

$$\boxed{t_i = \frac{d_i}{\hat{\sigma}_{\varepsilon(i)}\sqrt{1-H_{ii}}} = \frac{Y_i - \hat{y}_{i(i)}}{\sqrt{\dfrac{1}{N-k-1}\sum_{i=1}^{N}(Y_i-\hat{y}_{i(i)})^2}\sqrt{1-\sum_{j=1}^{n}(X(X^TX)^{-1}X^T)_{ij}}}} \tag{14.1.460}$$

Then if the errors are independent and normally distributed with expected value 0 and variance σ_ε^2, then the probability distribution of the i externally studentized residual is a Student's T-distribution with $N-k-1$ degrees of freedom:

$$T(N-k-1) = \frac{d_i}{\hat{\sigma}_{\varepsilon(i)}\sqrt{1-H_{ii}}} \tag{14.1.461}$$

Remark

In some statistical softwares we can find a influential indicator named DFITS (or also sometimes DFFITS) that is the acronym of "DiFference In fiTS" and defined by:

$$\text{DFIT} = \frac{Y_i - \hat{y}_{i(i)}}{\hat{\sigma}_{\varepsilon(i)}\sqrt{H_{ii}}} \tag{14.1.462}$$

DFITS is very similar to the externally studentized residual, and is in obviously equal to the latter times $\sqrt{H_{ii}/(1-H_{ii})}$.

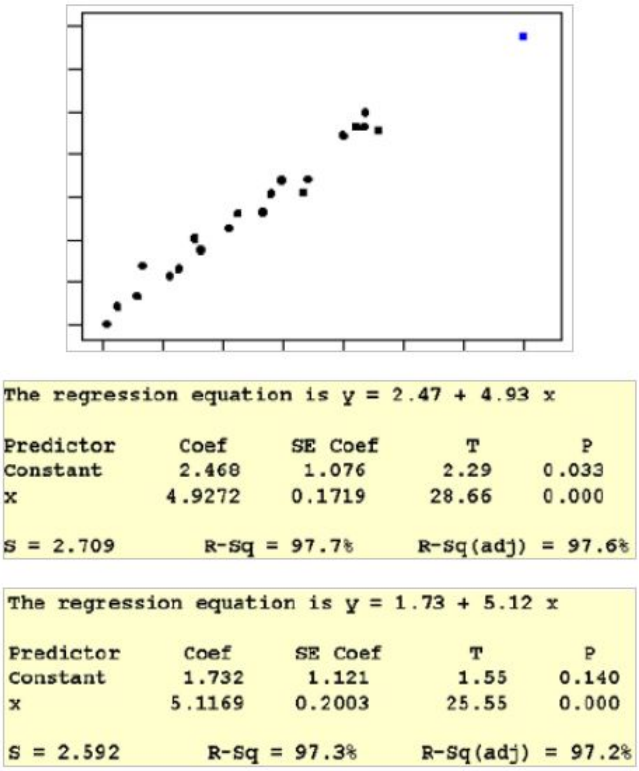
Probably one of the most famous influential indicator is the "Cook's distance". Omitting point i will generally change all of the fitted values, not just the fitted value at that point. Hence the idea to measure the total change and to define this distance as the sum of all the changes in the regression model when observation i is removed from it:

$$\|\vec{\hat{y}} - \vec{\hat{y}}^{(-i)}\|^2 = \sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2 \tag{14.1.463}$$

To make that relation more comparable across data sets, it's conventional to divide it by p, since there are really only p independent coordinates here, each of which might contribute something on the order of $\hat{\sigma}_\varepsilon$:

$$\boxed{D_i = \frac{\displaystyle\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2}{p\hat{\sigma}_\varepsilon}} \tag{14.1.464}$$

There are different opinions regarding what cutoff values to use for spotting highly influential points. A simple operational guideline of $D_i > 1$ has been suggested. Others have indicated that $D_i > 4/n$ where N is the number of observations, might be used.



```
The regression equation is y = 2.47 + 4.93 x

Predictor        Coef    SE Coef        T        P
Constant        2.468      1.076     2.29    0.033
x              4.9272     0.1719    28.66    0.000

S = 2.709          R-Sq = 97.7%      R-Sq(adj) = 97.6%
```

```
The regression equation is y = 1.73 + 5.12 x

Predictor        Coef    SE Coef        T        P
Constant        1.732      1.121     1.55    0.140
x              5.1169     0.2003    25.55    0.000

S = 2.592          R-Sq = 97.3%      R-Sq(adj) = 97.2%
```

Figure 14.17: **Importance of an influential point on regression equation**

14.1.11.4 Significance of regression coefficients (variable importance)

When you study variable importance, you should always first remember to standardize all your predictors (thus, centered and scaled by the sample standard deviations).

Now let us deal first with the Gaussian multivariate regression case! Let us recall that (see page 6401):

$$\hat{\beta}_i = \mathcal{N}\left(\beta_i, \text{V}(\hat{\beta}_i)\right) = \mathcal{N}\left(\beta_i, \sigma_\varepsilon^2 (X_c^T X_c)_{ii}^{-1}\right) \qquad (14.1.465)$$

Thus:

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{\text{V}(\hat{\beta}_i)}} = \frac{\hat{\beta}_i - \beta_i}{\text{SE}(\hat{\beta}_i)} = \frac{\hat{\beta}_i - \beta_i}{\sigma_\varepsilon \sqrt{(X_c^T X_c)_{ii}^{-1}}} = \mathcal{N}(0,1) \qquad (14.1.466)$$

under the null $H_0 : \beta_i = 0$ we would actually have a simple Z test:

$$\frac{\hat{\beta}_i}{\text{SE}(\hat{\beta})} = \frac{\hat{\beta}_i}{\sigma_\varepsilon \sqrt{(X_c^T X_c)_{ii}^{-1}}} = \mathcal{N}(0,1) \qquad (14.1.467)$$

This relation is again a kind of "Wald statistics" and the related hypothesis also belong to the family of a Wald tests (not the first we meet in this textbook). Also often denoted:

$$W=\frac{\left(\hat{\beta}-\beta_0\right)}{\widehat{\text{se}}(\hat{\beta})}=\mathcal{N}(0,1) \tag{14.1.468}$$

or:

$$W^2=\frac{\left(\hat{\beta}-\beta_0\right)^2}{\widehat{\mathrm{V}}(\hat{\beta})}=\chi_1^2 \tag{14.1.469}$$

Now let us introduce a necessary intermediate result. Remember that for the bivariate regression, we have proved that:

$$(n-2)\hat{\sigma}_\varepsilon^2=\sum_k (y_k'-y_k)^2=\sum \varepsilon_k^2=\text{SSR} \tag{14.1.470}$$

Let us divide by the supposed true value of σ_ε:

$$(n-2)\frac{\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2}=\sum_k \frac{(y_k'-y_k)^2}{\sigma_\varepsilon^2}=\frac{\text{SSR}}{\sigma_\varepsilon^2} \tag{14.1.471}$$

We recognize here a sum of square of Normal distribution that led us to write:

$$\frac{(n-2)\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2}=\sum_k \frac{(y_k'-y_k)^2}{\sigma_\varepsilon^2}=\frac{\text{SSR}}{\sigma_\varepsilon^2}=\chi^2(n-2) \tag{14.1.472}$$

That we will guess can be generalized to the multivariate case by:

$$\frac{(n-k)\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2}=\chi^2(n-k) \tag{14.1.473}$$

Let us rearrange and take the square root:

$$\frac{\hat{\sigma}_\varepsilon}{\sigma_\varepsilon}=\sqrt{\frac{\chi^2(n-k)}{n-k}} \tag{14.1.474}$$

Let us divide:

$$\frac{\hat{\beta}_i}{\sigma_\varepsilon\sqrt{(X_c^TX_c)_{ii}^{-1}}}=\mathcal{N}(0,1) \tag{14.1.475}$$

by this previous result. Then:

$$\frac{\hat{\beta}_i}{\hat{\sigma}_\varepsilon\sqrt{(X_c^TX_c)_{ii}^{-1}}}=\frac{\mathcal{N}(0,1)}{\sqrt{\dfrac{\chi^2(n-k)}{n-k}}} \tag{14.1.476}$$

We recognize here well the term on the right. It's the definition itself the Student distribution (see Statistics section page 783). Then we can write:

$$\boxed{\frac{\hat{\beta}_i}{\hat{\sigma}_\varepsilon\sqrt{(X_c^TX_c)_{ii}^{-1}}}=T(n-k)} \tag{14.1.477}$$

That latter relation is also sometimes denoted:

$$\frac{\hat{\beta}_i}{\text{SE}(\hat{\beta}_i)}=T(n-k) \tag{14.1.478}$$

It explains why, in regression analysis and in variable importance, greater the Student $T(n-k)$ value is (in absolute value) the more the coefficient is significant ($H_1:\ \beta_i\neq 0$) and the more the corresponding predictor is important (hence the expression "variable importance")!

In logistic (and Poisson) regression that we will see later, the variance of the residuals is related to the mean! Indeed, if $Y\sim\mathcal{B}(n,p)$, the mean is $\text{E}(Y)=np$ and the variance is $\text{V}(Y)=np(1-p)=\text{E}(Y)(1-p)$ so the variance and the mean are related.

In statistics, the "dispersion parameter" ϕ indicates if we have more or less than the expected variance. If $\phi=1$ this means we observe the expected amount of variance, whereas $\phi<1$ means that we have less than the expected variance (named "underdispersion") and $\phi>1$ means that we have extra variance beyond the expected variance (named "overdispersion"). The dispersion parameter in logistic and Poisson regression is fixed at 1 which means that we can use the Z-score (that's why statistical softwares give the Z-score for the coefficients of such regression models!). In other regression types such as Normal linear regression, we have to estimate the residual variance and thus this is why, a T-value is used for calculating the p-values (that's why statistical softwares give the T-score for the coefficients of such regression models!).

- Linear Regression:

  The use of T-tests in linear regression comes from the distribution of Normally distributed error terms: $y_i=X_i^T\beta+\varepsilon_i$ where $\varepsilon_i\sim\mathcal{N}(0,1)$ are independent and identically distributed. It follows then as seen above that $\frac{\hat{\beta}_j-\beta_{j0}}{\widehat{\text{se}}\left(\hat{\beta}_j\right)}\sim T(n-k)$.

  Note that the default in most regression software packages test the hypothesis that $\hat{\beta}_j=0$, i.e. setting β_{j0} equal to zero.

- Logistic Regression:

  Logistic regression assumes errors follow the logistic distribution. Consequently, the term $\frac{\hat{\beta}_j-\beta_{j0}}{\widehat{\text{se}}\left(\hat{\beta}_j\right)}$ does not follow a T-distribution. Instead, we can use the Wald test, which relies on asymptotic normality as is implied by the Central limit Theorem.

Note, though, that even if the T-test and Wald test are asymptotically equivalent (i.e. as the sample size, $n\rightarrow+\infty$, they will reject the same cases); certainly some people - if a bit loosely - name a test based on a T-statistic a "Wald test", whether the statistic is compared with the asymptotic Normal distribution or the small sample result (T-distribution)...

14.1.11.5 Variance Inflation Factor (multicollinearity detection)

In statistics, the "variance inflation factor" (VIF) quantifies the severity of multicollinearity in an ordinary least squares regression analysis. It provides an index that measures how much the variance (the square of the estimate's standard deviation) of an estimated regression coefficient is increased because of collinearity.

To start, the reader must know that during our study of Principal Component Analyses (see section Statistics page 1213) we have proved in a trivial way that the variance-covariance matrix is given by:

$$\text{cov}_{jj'}=\frac{1}{n}X_c^TX_c \tag{14.1.479}$$

with for recall (for a regression model with p terms, and hence $p-1$ variables):

$$X=\begin{bmatrix}1 & X_{11} & \ldots & X_{1(p-1)}\\ 1 & X_{21} & \ldots & X_{2(p-1)}\\ \vdots & \vdots & \vdots & \vdots\\ 1 & X_{n1} & \ldots & X_{n(p-1)}\end{bmatrix}\Rightarrow X_c=\begin{bmatrix}0 & X_{c,11} & \ldots & X_{c,1(p-1)}\\ 0 & X_{c,21} & \ldots & X_{c,2(p-1)}\\ \vdots & \vdots & \vdots & \vdots\\ 0 & X_{c,n1} & \ldots & X_{c,n(p-1)}\end{bmatrix} \tag{14.1.480}$$

Therefore for example (and for recall again) for the special case of a 2×2 design matrix:

$$X_c^TX_c=\begin{bmatrix}0 & X_{c,11}+X_{c,21}\\ X_{c,11}+X_{c,21} & X_{c,11}^2+X_{c,21}^2\end{bmatrix}=\begin{bmatrix}0 & 0\\ 0 & X_{c,11}^2+X_{c,21}^2\end{bmatrix} \tag{14.1.481}$$

Hence:

$$\text{cov}_{jj'}=\frac{1}{n}X_c^TX_c=\begin{bmatrix}0 & 0\\ 0 & \dfrac{X_{c,11}^2+X_{c,21}^2}{n}\end{bmatrix}=\begin{bmatrix}0 & 0\\ 0 & \overline{X_{c,1}^2}\end{bmatrix} \tag{14.1.482}$$

And for a 3×3 matrix:

$$\text{cov}_{jj'}=\begin{bmatrix}0 & 0 & 0\\ 0 & \overline{X_{c,1}^2} & \overline{X_{c,1}X_{c,2}}\\ 0 & \overline{X_{c,1}X_{c,2}} & \overline{X_{c,2}^2}\end{bmatrix} \tag{14.1.483}$$

And as the matrix of correlations is given by (see section Statistics page 621):

$$R_{jj'}=\frac{\text{cov}_{jj'}}{\sigma_j\sigma_{j'}} \tag{14.1.484}$$

Then we get:

$$R_{jj'}=\frac{1}{n}\frac{1}{\sigma_j\sigma_{j'}}X_c^TX_c \tag{14.1.485}$$

The ideal would also be to be able to put the inverse of the standard deviations in a matrix form in order to have a totally matrix expression. After some trial and error we quickly find that we can write the correlation matrix in the form:

$$R_{jj'}=\frac{1}{n}S^{-1}(X_c^TX_c)S^{-1}=R \tag{14.1.486}$$

where S^{-1} is the usual notation to say that it is the diagonal matrix containing the inverse of the σ_j, or explicitly (so that it is clear to the reader!):

$$S^{-1}=\begin{bmatrix} \dfrac{1}{\sigma_1} & 0 & \ldots & 0 \\ 0 & \dfrac{1}{\sigma_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & \ldots & \dfrac{1}{\sigma_n} \end{bmatrix} \tag{14.1.487}$$

So for now, we have two relations (out of three ... waiting for the proof of the third one) that will be very useful to us (attention the components of the vectors of the explanatory variables are always centered henceforth what implies a forced linear regression by the origin!):

$$\begin{gathered} R=\frac{1}{n}S^{-1}(X_c^TX_c)S^{-1} \\ \hat{\vec{\beta}}=(X_c^TX_c)^{-1}X_c^T\vec{y} \end{gathered} \tag{14.1.488}$$

Remark

Because $X_c^TX_c$ has the first row and column only with zeros (due to recall by the constant of the model) that generates issues when inverting the matrix (i.e. $(X_c^TX_c)^{-1}$), we then eliminate this column and row. To highlight this choice, a few textbooks write instead:

$$\left(X_{(-1)c}^TX_{(-1)c}\right)^{-1} \tag{14.1.489}$$

where the (-1) is here to indicate that the first row and first column were removed. For what will follow we did not make this choice of notation, so keep in mind now that when you see X_c, it is the "design matrix" but without the first column, or said differently: $X_c^TX_c$ without the first column and without the first row.

Now let's start from (don't forget that starting from now $\vec{\beta}$ does not contain β_0 anymore!!!):

$$\hat{\vec{\beta}}=(X_c^TX_c)^{-1}X_c^T\vec{y} \tag{14.1.490}$$

and let us rewrite this in a somewhat more explicit form:

$$\begin{aligned} \hat{\vec{\beta}}&=(X_c^TX_c)^{-1}X_c^T(X_c\vec{\beta}+\vec{\varepsilon})=(X_c^TX_c)^{-1}X_c^TX_c\vec{\beta}+(X_c^TX_c)^{-1}X_c^T\vec{\varepsilon} \\ &=\vec{\beta}+(X_c^TX_c)^{-1}X_c^T\vec{\varepsilon} \end{aligned} \tag{14.1.491}$$

Therefore:

$$\hat{\vec{\beta}}-\vec{\beta}=(X_c^TX_c)^{-1}X_c^T\vec{\varepsilon} \tag{14.1.492}$$

We understand now better why $(X_c^TX_c)^{-1}$ is often named "dispersion matrix" as we have already mentioned earlier above.

Let us now take the matrix of the covariance-variances matrix of the coefficients (measurement of multicollinearity):

$$\text{cov}\left(\hat{\vec{\beta}}\right)=\text{E}\left[\left(\hat{\vec{\beta}}-\vec{\beta}\right)\left(\hat{\vec{\beta}}-\vec{\beta}\right)^T\right]=\text{E}\left[\left((X_c^TX_c^{-1})X_c^T\vec{\varepsilon}\right)\left((X_c^TX_c)^{-1}X_c^T\vec{\varepsilon}\right)^T\right] \tag{14.1.493}$$

Using the following property of the transposed matrices proved in the section Linear Algebra (see page 1951):

$$(AB)^T = B^T A^T \tag{14.1.494}$$

we then have using this property a first time:

$$\begin{aligned}\text{cov}\left(\hat{\vec{\beta}}\right) &= \text{E}\left[((X_c^T X_c)^{-1} X_c^T \vec{\varepsilon})\left(\underbrace{(X_c^T X_c)^{-1}}_{:=A}\underbrace{X_c^T \vec{\varepsilon}}_{:=B}\right)^T\right]\\ &= \text{E}\left[((X_c^T X_c)^{-1} X_c^T \vec{\varepsilon})\left((X_c^T \vec{\varepsilon})^T\left((X_c^T X_c)^{-1}\right)^T\right)\right]\end{aligned} \tag{14.1.495}$$

and using it a second time:

$$\begin{aligned}\text{cov}\left(\hat{\vec{\beta}}\right) &= \text{E}\left[((X_c^T X_c)^{-1} X_c^T \vec{\varepsilon})\left((\underbrace{X_c^T}_{:=A}\underbrace{\vec{\varepsilon}}_{:=B})^T\left((X_c^T X_c)^{-1}\right)^T\right)\right]\\ &= \text{E}\left[((X_c^T X_c)^{-1} X_c^T \vec{\varepsilon})\left(\vec{\varepsilon}^T X_c \left((X_c^T X_c)^{-1}\right)^T\right)\right]\end{aligned} \tag{14.1.496}$$

Using now the following property proven in the section Linear Algebra (see page 1951):

$$(A^{-1})^T = (A^T)^{-1} \tag{14.1.497}$$

We then have:

$$\text{cov}\left(\hat{\vec{\beta}}\right) = \text{E}\left[((X_c^T X_c)^{-1} X_c^T \vec{\varepsilon})\left(\vec{\varepsilon}^T X_c \left((X_c^T X_c)^T\right)^{-1}\right)\right] \tag{14.1.498}$$

and reusing:

$$(AB)^T = B^T A^T \tag{14.1.499}$$

it comes:

$$\begin{aligned}\text{cov}\left(\hat{\vec{\beta}}\right) &= \text{E}\left[((X_c^T X_c)^{-1} X_c^T \vec{\varepsilon})\left(\vec{\varepsilon}^T X_c \left((\underbrace{X_c^T}_{:=A}\underbrace{X_c}_{:=B})^T\right)^{-1}\right)\right]\\ &= \text{E}\left[((X_c^T X_c)^{-1} X_c^T \vec{\varepsilon})\left(\vec{\varepsilon}^T X_c \left(X_c^T X_c\right)^{-1}\right)\right]\end{aligned} \tag{14.1.500}$$

Using the associativity property of the matrix product, we will write:

$$\begin{aligned}\text{cov}\left(\hat{\vec{\beta}}\right) &= \text{E}\left[((X_c^T X_c)^{-1} X_c^T \vec{\varepsilon})\left(\vec{\varepsilon}^T X_c \left(X_c^T X_c\right)^{-1}\right)\right]\\ &= \text{E}\left[(X_c^T X_c)^{-1} X_c^T (\vec{\varepsilon}\vec{\varepsilon}^T) X_c (X_c^T X_c)^{-1}\right]\end{aligned} \tag{14.1.501}$$

And by taking out the elements which are not random variables of the expected mean, it comes:

$$\begin{aligned}\text{cov}\left(\hat{\vec{\beta}}\right) &= (X_c^T X_c)^{-1} X_c^T \text{E}(\vec{\varepsilon}\vec{\varepsilon}^T) X_c (X_c^T X_c)^{-1}\\ &= (X_c^T X_c)^{-1} X_c^T \sigma_\varepsilon^2 X_c (X_c^T X_c)^{-1}\\ &= \sigma_\varepsilon^2 \underbrace{(X_c^T X_c)^{-1}(X_c^T X_c)}_{=1}(X_c^T X_c)^{-1}\\ &= \sigma_\varepsilon^2 (X_c^T X_c)^{-1}\end{aligned} \tag{14.1.502}$$

It is sadly often traditional (in spite of the confusion) to write this last result in the following form which we will see (and use) in the section of Industrial Engineering during our study of the Box-Behnken experimental designs:

$$\boxed{\mathrm{V}\left(\hat{\vec{\beta}}\right)=\sigma_{\varepsilon}^{2}(X_c^T X_c)^{-1}} \tag{14.1.503}$$

Remark

At this point, notice something important! As by assumption for the Gaussian multivariate linear regression we have $Y=\mathcal{N}(X\vec{\beta},\sigma_{\varepsilon}^{2}\mathbb{1}_k)$ as a consequence, we then have using the result above:

$$\boxed{\hat{\vec{\beta}}=\mathcal{N}\left(\vec{\beta},\sigma_{\varepsilon}^{2}(X_c^T X_c)^{-1}\right)} \tag{14.1.504}$$

This relation will be important for us later in our study of variable importance in the case of regression models and statistical significant test for the coefficient of the regression!

This last matrix $\mathrm{V}\left(\hat{\vec{\beta}}\right)$ contains in the diagonal the general expression in the general case of multi-linear regression the "standard error of the regression coefficients" (relation that uses among other the spreadsheet software Microsoft® Excel® in the case of the multi-linear regression for the column "Standard error" of the coefficients ... except that it takes the square root to have the standard deviation and not the variance!). If the linear regression is not forced at the origin, it will be necessary to use obviously:

$$\mathrm{V}\left(\hat{\vec{\beta}}\right)=\sigma_{\varepsilon}^{2}(X_T X)^{-1} \tag{14.1.505}$$

with for recall:

$$\sigma_{\varepsilon}^{2}=\frac{1}{n-k}\sum(Y_i-y_i)^2 \tag{14.1.506}$$

where k is for recall the number of coefficients of the multiple linear model.

We have now three important relations[11]:

$$\begin{aligned}R&=\frac{1}{n}S^{-1}(X_c^T X_c)S^{-1}\\ \hat{\vec{\beta}}&=(X_c^T X_c)^{-1}X_c^T\vec{y}\\ \mathrm{V}\left(\hat{\vec{\beta}}\right)&=\sigma_{\varepsilon}^{2}(X_c^T X_c)^{-1}\end{aligned} \tag{14.1.507}$$

Let us take up the first relation and manipulate it a little bit:

$$X_c^T X_c=nSRS \tag{14.1.508}$$

It comes then:

$$(X_c^T X_c)^{-1}=\frac{1}{n}(SRS)^{-1} \tag{14.1.509}$$

---

[11] There is a fourth one in reality that is also important...! The bias-variance tradeoff of the ordinary least squares regression as derived in details at page 6799.

Then we have:

$$(X_c^T X_c)^{-1}=\frac{1}{n}S^{-1}R^{-1}S^{-1} \tag{14.1.510}$$

By substituting the right-hand term for equality in the third of the equalities, we then have:

$$\text{V}\left(\hat{\vec{\beta}}\right)=\sigma_\varepsilon^2(X_c^T X_c)^{-1}=\sigma_\varepsilon^2\frac{1}{n}S^{-1}R^{-1}S^{-1}=\frac{\sigma_\varepsilon^2}{n}S^{-1}R^{-1}S^{-1} \tag{14.1.511}$$

If we denote by $\text{V}\left(\hat{\vec{\beta}}\right)_j$ the j-th diagonal element of the covariance-variance matrix of the $\hat{\vec{\beta}}$ and VIF_j, the "Variance Inflation Factor (VIF)", the j-th diagonal element of the matrix R^{-1} then it trivially follows that:

$$\text{V}\left(\hat{\vec{\beta}}\right)_j=\frac{\sigma_\varepsilon^2}{n}\frac{\text{VIF}_j}{\sigma_j\sigma_j}=\frac{\sigma_\varepsilon^2}{n}\frac{\text{VIF}_j}{\text{V}_j}=\frac{\sigma_\varepsilon^2}{n}\frac{\text{VIF}_j}{\text{V}(X_{c,j})} \tag{14.1.512}$$

The "scaled Variance Inflation Factor" is sometimes used instead and defined by:

$$\text{V}\left(\hat{\vec{\beta}}\right)_j=\frac{\sigma_\varepsilon^2}{n}\frac{\text{VIF}_j}{\sigma_j\sigma_j}=\frac{\sigma_\varepsilon^2}{n}\text{VIF}_{js} \tag{14.1.513}$$

Remark

Remember that if the variances are all equal to the unit ($\text{V}_i=1$) then the covariance matrix is equal to the correlation matrix, i.e. $R=\Sigma$. Hence if we work with the VIF_{js}, all the information is then also contained in the inverse of the covariance matrix, i.e. $\Sigma^{-1}=\Theta$ named the "precision matrix".

However, we must find the explicit formulation of the VIF_j. For this, let us recall that R is a positive definite symmetric matrix (see section Linear Algebra page 1957). Let us, for example, calculate the inverse of such a matrix of dimension 2×2 explicitly by using the relation proved in the section Linear Algebra (see page 1980) that is for recall:

$$b_{ij}=\frac{1}{\det(A)}(-1)^{i+j}\det(A_{ji}) \tag{14.1.514}$$

In the special case that interests us, we have then:

$$R^{-1}=\begin{bmatrix}R_{11} & R_{12}\\ R_{21} & R_{22}\end{bmatrix}^{-1}=\frac{1}{R_{11}R_{22}-R_{12}R_{21}}\begin{bmatrix}R_{11} & -R_{12}\\ -R_{21} & R_{22}\end{bmatrix} \tag{14.1.515}$$

and since the correlation matrix is symmetric and unit diagonal, we have:

$$R^{-1}=\frac{1}{1-R_{21}^2}\begin{bmatrix}1 & -R_{21}\\ -R_{21} & 1\end{bmatrix}=\frac{1}{1-R_{12}^2}\begin{bmatrix}1 & -R_{12}\\ -R_{12} & 1\end{bmatrix} \tag{14.1.516}$$

It comes then that if we focus on the elements of the diagonal:

$$R^{-1}=\begin{bmatrix}\dfrac{1}{1-R_{21}^2} & \ldots\\ \ldots & \dfrac{1}{1-R_{21}^2}\end{bmatrix}=\begin{bmatrix}\dfrac{1}{1-R_{12}^2} & \ldots\\ \ldots & \dfrac{1}{1-R_{12}^2}\end{bmatrix} \tag{14.1.517}$$

Repeating the job with a 3×3 matrix, we have:

$$\begin{aligned}R^{-1} &= \begin{bmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{bmatrix}^{-1} = \begin{bmatrix} 1 & R_{12} & R_{13} \\ R_{21} & 1 & R_{23} \\ R_{31} & R_{32} & 1 \end{bmatrix}^{-1} \\ &= \frac{1}{1-R_{12}^2-R_{13}^2-R_{23}^2+2R_{12}R_{13}R_{23}} \begin{bmatrix} 1-R_{23}^2 & R_{13}R_{23}-R_{12} & R_{12}R_{23}-R_{13} \\ R_{13}R_{23}-R_{21} & 1-R_{13}^2 & R_{12}R_{13}-R_{23} \\ R_{12}R_{23}-R_{31} & R_{12}R_{13}-R_{32} & 1-R_{12}^2 \end{bmatrix}\end{aligned} \tag{14.1.518}$$

Let us focus on the first diagonal component and rearrange it a bit to get something similar to the 2×2 case:

$$\begin{aligned}R_{1,1}^{-1} &= \frac{1-R_{23}^2}{1-R_{12}^2-R_{13}^2-R_{23}^2+2R_{12}R_{13}R_{23}} = \frac{1-R_{23}^2}{(1-R_{23}^2)-R_{12}^2-R_{13}^2+2R_{12}R_{13}R_{23}} \\ &= \frac{1}{1-\dfrac{R_{12}^2+R_{13}^2-2R_{12}R_{13}R_{23}}{1-R_{23}^2}}\end{aligned} \tag{14.1.519}$$

Now we put:

$$R_{1,23}^2 = \frac{R_{12}^2+R_{13}^2-2R_{12}R_{13}R_{23}}{1-R_{23}^2} \tag{14.1.520}$$

So that:

$$R_{1,1}^{-1} = \frac{1}{1-\dfrac{R_{12}^2+R_{13}^2-2R_{12}R_{13}R_{23}}{1-R_{23}^2}} = \frac{1}{1-R_{1,23}^2} \tag{14.1.521}$$

Where we define:

$$\boxed{R_{1,23} = \sqrt{\frac{R_{12}^2+R_{13}^2-2R_{12}R_{13}R_{23}}{1-R_{23}^2}}} \tag{14.1.522}$$

as the "multiple correlation coefficient" (for three variables...). This is by definition the way we calculate the correlation between x_1 and the other explanatory variables:

$$x_1 = \alpha_0 + \alpha_1 x_2 + \alpha_2 x_3 + \varepsilon \tag{14.1.523}$$

And the reader can check that regardless of the size of the matrix, the VIF_j can always be put in the form:

$$\boxed{\text{VIF}_j = S^{-1}R^{-1}S^{-1} = \frac{1}{1-R_j^2}} \tag{14.1.524}$$

In the case of 2×2 correlation matrix there is therefore only one coefficient VIF_j for the simple and good reason that the associated linear regression has only two coefficients ...

Remark

Some textbooks give the following relation for the general case:

$$\text{VIF}_j=\frac{1}{1-R_j^n} \tag{14.1.525}$$

But this is false as far as I know!

Most statistical softwares simply use the following relation:

$$\text{V}\left(\hat{\vec{\beta}}\right)=\frac{\sigma_\varepsilon^2}{n}\frac{\text{VIF}_j}{\text{V}(x_{c,j})} \tag{14.1.526}$$

So we understand (how the name suggests...) why the variance inflation factor (VIF) quantifies how much the variance of the coefficients is inflated when multicollinearity exists.

Therefore:

$$\text{VIF}_j=\frac{n\text{V}\left(\hat{\vec{\beta}}\right)\text{V}(X_{c,j})}{\sigma_\varepsilon^2} \tag{14.1.527}$$

It is important here to notice that the VIF (multicolinearity) is proportional to the variance of the beta coefficients estimation. Hence as we have proved earlier (see page 6395) that the significance of regression coefficients is inversely proportional to the square of the variance of the beta coefficients estimation, multicollinearity implies therefore small p-values and thus leading to some variables becoming significant even though they are actually not significant.

Either in the case where the data are such that the regression is not forced at the origin:

$$\begin{aligned}\text{VIF}_j&=\frac{n\text{V}\left(\hat{\vec{\beta}}\right)\text{V}(X_j)}{\sigma_\varepsilon^2}=\frac{n\sigma_\varepsilon^2(X^TX)^{-1}\dfrac{1}{n}\displaystyle\sum_i(x_{ij}-\bar{x}_j)^2}{\dfrac{1}{n-k}\displaystyle\sum_i(Y_i-y_i)^2}\\&=\frac{\dfrac{1}{n-k}\displaystyle\sum_i(Y_i-y_i)^2(X^TX)^{-1}\sum_i(x_{ij}-\bar{x}_j)^2}{\dfrac{1}{n-k}\displaystyle\sum_i(Y_i-y_i)^2}\\&=(X^TX)^{-1}\sum_i(x_{ij}-\bar{x}_j)^2\end{aligned} \tag{14.1.528}$$

Therefore the version I prefer to use and to teach (especially for customers and students having only a spreadsheet software at their disposition is):

$$\boxed{\text{VIF}_j=\frac{1}{1-R_j^2}=n(X^TX)^{-1}\sqrt{\sum_i(x_{ij}-\bar{x}_j)^2}} \tag{14.1.529}$$

and as all terms are known to us in practice, the calculation of the VIF_j doesn't make us problems anymore!

A rule of thumb is that if $\text{VIF}_j > 10$ then multicollinearity is high (solving backwards this translates into an R_j^2 value of 0.90). Obviously a VIF of 1 means that there is no correlation among the kth predictor and the remaining predictor variables, and hence the variance of β_k is not inflated at all. VIFs exceeding 4 should warrant further investigation.

Remark

Some softwares instead calculates the tolerance which is just the reciprocal of the VIF. The choice of which to use is a matter of personal preference.

The square root of the variance inflation factor finally indicates how much larger the standard error is, compared with what it would be if that variable were uncorrelated with the other predictor variables in the model.

✍**Example:**

If the variance inflation factor of a predictor variable is equal to 5.27 ($\sqrt{5.27}=2.3$), this means that the standard error for the coefficient of that predictor variable is 2.3 times as large as it would be if that predictor variable was uncorrelated with the other predictor variables.

Regardless of your criterion for what constitutes a high VIF, there are situations in which a high VIF is not a problem and can be safely ignored. For example, if you specify a regression model with both x and x^2, there is a good chance that those two variables will be highly correlated.

14.1.11.5.1 Gauss-Markov theorem

After all this we finally have the tools and results necessary to prove the Gauss-Markov theorem in an easy way!

In statistics, the "Gauss–Markov theorem" (or simply Gauss theorem for some authors) states that the ordinary least squares (OLS) estimator has the lowest sampling variance within the class of linear unbiased estimators, if the errors in the linear regression model are uncorrelated, have equal variances and expectation value of zero. The errors do not need to be Normally distributed, nor do they need to be independent and identically distributed (only uncorrelated with mean zero and homoscedastic with finite variance).

It's not a very important proof for the engineer but statisticians like a lot such properties!

First we need a definition!

Definition (#559): Given a parameter vector $\vec{\omega}$ in q dimensions, and a vector $\vec{y}$ from which to estimate it, we say that the estimator $\vec{\omega}^*$ is the best unbiased estimator of $\vec{\omega}$, or "Best Linear Unbiased Estimator" (BLUE) of $\vec{\omega}$, if:

1. $\vec{\omega}^*$ is a linear estimator of $\vec{\omega}$, in other words there is a matrix A such that $\vec{\omega}^* = A\vec{y}$;
2. $\vec{\omega}^*$ is unbiased, in other words, $\text{E}(\vec{\omega}^*)=\vec{\omega}$

3. Whatever the unbiased linear estimator $\tilde{\vec{\omega}}$ of $\vec{\omega}$, the variance of this estimator is greater than or equal to that of $\vec{\omega}^*$.

Theorem 14.3. Given n observations $\vec{x}^1,\vec{x}^2,\ldots,\vec{x}^n\in\mathbb{R}^p$ and their labels $y^1,y^2,\ldots,y^n\in\mathbb{R}$. Under the assumption that, for all $i, y^i=\beta_0+\sum_{j=1}^p\beta_j x_j^i+\varepsilon^i$ and that the ε^i are normally distributed and centered at 0, then the estimator of $\vec{\beta}$ by the least squares method is its unique BLUE!

Proof 14.3.1. First as we know:

$$\vec{\beta}^*=\left(X^TX\right)^{-1}X^T\vec{y} \tag{14.1.530}$$

and $\vec{\beta}^*$ is therefore linear. It's expected mean is:

$$\mathrm{E}\left(\vec{\beta}^*\right)=\mathrm{E}\left(\left(X^TX\right)^{-1}X^T(X\vec{\beta}+\varepsilon)\right) \tag{14.1.531}$$

As $X,\vec{y}$ and $\vec{\beta}$ are not random, and that ε has a 0 expected mean, we get:

$$\mathrm{E}\left(\vec{\beta}^*\right)=\mathrm{E}\left(\left(X^TX\right)^{-1}X^T\vec{\beta}\right)=\vec{\beta} \tag{14.1.532}$$

Therefore, $\vec{\beta}^*$ is unbiased.

Its variance is equal to (see page 6400):

$$\mathrm{V}\left(\vec{\beta}^*\right)=\left(X^TX\right)^{-1}\sigma_\varepsilon^2 \tag{14.1.533}$$

Finally, suppose now that $\tilde{\vec{\beta}}=A\vec{y}$ is another linear unbiased estimator of $\vec{\beta}$. By definition, therefore $\mathrm{E}(\tilde{\vec{\beta}})=\vec{\beta}$. By replacing $\vec{y}$ by it value, we get $\mathrm{E}\left(A(X\vec{\beta}+\varepsilon)\right)=\vec{\beta}$ and therefore $AX\vec{\beta}=\vec{\beta}$. As this is true for every $\vec{\beta}$, we conclude that $AX=I$ (this result will be useful for later below).

Let us put $D=A-\left(X^TX\right)^{-1}X^T$, such that $\tilde{\vec{\beta}}-\vec{\beta}^*=D\vec{y}$. The variance of $\tilde{\vec{\beta}}$ is equal to:

$$\mathrm{V}(\tilde{\vec{\beta}})=\mathrm{V}(A\vec{y})=\mathrm{V}(A\varepsilon)=A^2\sigma_\varepsilon^2=AA^T\sigma_\varepsilon^2 \tag{14.1.534}$$

We can then express AA^T as follows:

$$\begin{aligned}AA^T&=\left(D+\left(X^TX\right)^{-1}X^T\right)\left(D+\left(X^TX\right)^{-1}X^T\right)^T\\&=\left(D+\left(X^TX\right)^{-1}X^T\right)\left(D^T+X\left(X^TX\right)^{-1}\right)\\&=DD^T+DX\left(X^TX\right)^{-1}+\left(X^TX\right)^{-1}X^TD^T+\left(X^TX\right)^{-1}\end{aligned} \tag{14.1.535}$$

As $AX=I$, then:

$$D=A-\left(X^TX\right)^{-1}X^T\Rightarrow DX=AX-\left(X^TX\right)^{-1}X^TX=0 \tag{14.1.536}$$

and it is immediate that we have similarly $(DX)^T=X^TD^T=0$.

We therefore get:

$$AA^T=DD^T+DX\left(X^TX\right)^{-1}+\left(X^TX\right)^{-1}X^TD^T+\left(X^TX\right)^{-1}=DD^T+\left(X^TX\right)^{-1} \tag{14.1.537}$$

Using this trick we can rewrite the variance as:

$$\mathrm{V}(\tilde{\vec{\beta}})=\sigma_\varepsilon^2 DD^T+\sigma_\varepsilon^2\left(X^TX\right)^{-1}=\mathrm{V}\left(\vec{\beta}^*\right)+\sigma_\varepsilon^2 DD^T \qquad (14.1.538)$$

As $\sigma_\varepsilon^2>0$ and DD^T has by construction every component that is >0, we have:

$$\mathrm{V}(\tilde{\vec{\beta}})>\mathrm{V}\left(\vec{\beta}^*\right) \qquad (14.1.539)$$

excepted if $D=0$, in which case $\tilde{\vec{\beta}}=\vec{\beta}^*$.

■ Q.E.D.

Remark

As already mentioned at the beginning, we see then that the Normality assumption on ε is not necessary! It suffices that the errors $\{\varepsilon_i\}_{i=1,\ldots,n}$ have zero expectation, have the same finite variance σ^2 (homoscedasticity), and are not correlated with each other ($\operatorname{cov}(\varepsilon_i,\varepsilon_l)=0$ for $i\neq l$).

14.1.11.6 Weighted Least Squares (WLS)

The method of ordinary least squares[12] assumes that there is constant variance in the errors (which is named "homoscedasticity" for recall). The method of "weighted least squares" (WLS) can be used when the ordinary least squares assumption of constant variance in the errors is violated (which is named "heteroscedasticity" for recall). The model under consideration is:

$$\vec{y}=X\vec{\beta}+\vec{\varepsilon} \qquad (14.1.540)$$

where we consider now that the $\varepsilon_i=\mathcal{N}(0,\sigma_i)$. Hence the errors may not be homoscedastic!

Let us recall the likelihood for the multivariate linear gaussian regression:

$$L\left(\vec{\beta},\sigma\right)=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n\left(y_i-\vec{x}_i^T\vec{\beta}\right)^2}=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\left(\vec{y}-X\vec{\beta}\right)^T\left(\vec{y}-X\vec{\beta}\right)} \qquad (14.1.541)$$

In the heteroscedatic case that interest us here the likelihood can only be written as:

$$L\left(\vec{\beta},\vec{\sigma}\right)=e^{-\frac{1}{2}\sum_{i=1}^n\left(\frac{y_i-\vec{x}_i^T\vec{\beta}}{\sigma_i}\right)^2}\prod_{i=1}^n\frac{1}{\sigma_i\sqrt{2\pi}}=e^{-\frac{1}{2}\left(\vec{y}-X\vec{\beta}\right)^T\Sigma^{-1}\left(\vec{y}-X\vec{\beta}\right)}\prod_{i=1}^n\frac{1}{\sigma_i\sqrt{2\pi}} \qquad (14.1.542)$$

named the "generalized least squares regression likelihood".

---

[12] A rule of thumb for OLS regression is that it isn't too impacted by heteroscedasticity as long as the maximum variance is not greater than four times the minimum variance.

So in the case of standard linear regression we have $\sigma_i = \sigma\ \forall i$ and all of diagonal terms that are zero, therefore:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \ldots & 0 \\ 0 & \sigma_2^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_n^2 \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \ldots & 0 \\ 0 & \sigma^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma^2 \end{bmatrix} \tag{14.1.543}$$

If we take the logarithm of the likelihood and ignoring the terms that doesn't content the β_i we then get:

$$\log\left(L\left(\vec{\beta},\vec{\sigma}\right)\right) \propto -\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - \vec{x}_i^T\vec{\beta}}{\sigma_i}\right)^2 \tag{14.1.544}$$

The idea is to rewrite the previous relation as:

$$\log\left(L\left(\vec{\beta},\vec{\sigma}\right)\right) \propto -\frac{1}{2}\sum_{i=1}^{n} w_i\left(y_i - \vec{x}_i^T\vec{\beta}\right)^2 = -\frac{1}{2}\sum_{i=1}^{n} w_i\varepsilon_i^2 = -\frac{1}{2}\sum_{i=1}^{n}(\sqrt{w_i}\varepsilon_i)^2 \tag{14.1.545}$$

with:

$$w_i = \frac{1}{\sigma_i^2} \tag{14.1.546}$$

Or in matrix form:

$$W = \Sigma^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \ldots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \frac{1}{\sigma_n^2} \end{bmatrix} \tag{14.1.547}$$

Remark

The reader must understand that the reason some statistical softwares require the matrix W is just because they will replace the calculated $\hat{\sigma}_i$ by the corresponding w_i and not multiply them together! And by the way... as all off-diagonal terms of W are zero, most statistical softwares require only a vector corresponding to the values of the diagonal of W.

Or in matrix notation this will be denoted:

$$\log\left(L\left(\vec{\beta},\vec{\sigma}\right)\right) \propto -\frac{1}{2}\left(\vec{y} - X\vec{\beta}\right)^T W\left(\vec{y} - X\vec{\beta}\right) \tag{14.1.548}$$

So for weighted regression, it is the weighted residuals that we try to make homoscedastic! Such that:

$$\varepsilon_{wi} = \sqrt{w_i}\varepsilon_i \tag{14.1.549}$$

Let us consider the univariate case:

$$\log(L(\vec{\beta},\vec{w}_i)) \propto \sum_{i=1}^{n} w_i(y_i - (\beta_0 + \beta_1 x_i))^2 \tag{14.1.550}$$

To determine the intercept we then calculate:

$$\frac{\partial \log(L(\vec{\beta},\vec{w_i}))}{\partial \beta_0}=0=\sum_{i=1}^{n} 2w_i(y_i-(\beta_0+\beta_1 x_i))(-1) \qquad (14.1.551)$$

Hence:

$$\beta_0=\frac{\displaystyle\sum_{i=1}^{n} w_i y_i}{\displaystyle\sum_{i=1}^{n} w_i}-\beta_1\frac{\displaystyle\sum_{i=1}^{n} w_i x_i}{\displaystyle\sum_{i=1}^{n} w_i}=\bar{y}_w-\beta_1\bar{x}_w \qquad (14.1.552)$$

And similarly:

$$\begin{aligned}\frac{\partial \log(L(\vec{\beta},\vec{w_i}))}{\partial \beta_1}=0&=\frac{\partial w_i \displaystyle\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1 x_i))^2}{\partial \beta_1}=\frac{\partial \displaystyle\sum_{i=1}^{n} w_i(y_i-((\bar{y}_w-\beta_1\bar{x}_w)+\beta_1 x_i))^2}{\partial \beta_1}\\&=2w_i\sum_{i=1}^{n}\left((y_i-\bar{y}_w)-\beta_1(x_i-\bar{x}_w)\right)\left(-(x_i-\bar{x}_w)\right)\end{aligned} \qquad (14.1.553)$$

Hence:

$$\begin{aligned}\frac{\partial \log(L(\vec{\beta},\vec{w_i}))}{\partial \beta_1}=0&=\frac{\partial w_i \displaystyle\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1 x_i))^2}{\partial \beta_1}=\frac{\partial \displaystyle\sum_{i=1}^{n} w_i(y_i-((\bar{y}_w-\beta_1\bar{x}_w)+\beta_1 x_i))^2}{\partial \beta_1}\\&=2w_i\sum_{i=1}^{n}\left((y_i-\bar{y}_w)-\beta_1(x_i-\bar{x}_w)\right)\left(-(x_i-\bar{x}_w)\right)\end{aligned} \qquad (14.1.554)$$

And similarly:

$$\beta_1=\frac{\displaystyle\sum_{i=1}^{n} w_i(y_i-\bar{y}_w)(x_i-\bar{x}_w)}{\displaystyle\sum_{i=1}^{n} w_i(x_i-\bar{x}_w)^2} \qquad (14.1.555)$$

Notice that for the two relations:

$$\boxed{\beta_0=\frac{\displaystyle\sum_{i=1}^{n} w_i y_i}{\displaystyle\sum_{i=1}^{n} w_i}-\beta_1\frac{\displaystyle\sum_{i=1}^{n} w_i x_i}{\displaystyle\sum_{i=1}^{n} w_i}=\bar{y}_w-\beta_1\bar{x}_w \qquad \beta_1=\frac{\displaystyle\sum_{i=1}^{n} w_i(y_i-\bar{y}_w)(x_i-\bar{x}_w)}{\displaystyle\sum_{i=1}^{n} w_i(x_i-\bar{x}_w)^2}} \qquad (14.1.556)$$

If $w_i=1$ for all i then we fall back on OLS.

Now let us generalize that result. First notice that:

$$\log\left(L\left(\vec{\beta},\vec{\sigma}\right)\right)\propto-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i-\vec{x}_i^T\vec{\beta}}{\sigma_i}\right)^2 \tag{14.1.557}$$

will be rewritten in matrix form (we get rid of the factor $-1/2$ that doesn't bring anything):

$$\log\left(L\left(\vec{\beta},\vec{\sigma}\right)\right)\propto\left\|\Sigma^{-1/2}(\vec{y}-X\vec{\beta})\right\|^2:=\left\|W^{1/2}(\vec{y}-X\vec{\beta})\right\|^2=(\vec{y}-X\vec{\beta})^TW(\vec{y}-X\vec{\beta}) \tag{14.1.558}$$

Let us expand the last equality (using the property of transposed matrices as seen on page 1951 in the section of Linear Algebra):

$$\begin{aligned}\log\left(L\left(\vec{\beta},\vec{\sigma}\right)\right)&\propto(\vec{y}-X\vec{\beta})^TW(\vec{y}-X\vec{\beta})=\vec{y}^TW\vec{y}-\vec{y}^TWX\vec{\beta}-(X\vec{\beta})^TW\vec{y}+(X\vec{\beta})^TWX\vec{\beta}\\&=\vec{y}^TW\vec{y}-\vec{y}^TWX\vec{\beta}-\vec{\beta}^TX^TW\vec{y}+\vec{\beta}^TX^TWX\vec{\beta}\end{aligned} \tag{14.1.559}$$

Now we seek the minimum of the latter relation (Weighted Sum of Squares Errors) with respect to the β_i:

$$\text{WSSE}\propto\min_{\vec{\beta}}\quad\vec{y}^TW\vec{y}-\vec{y}^TWX\vec{\beta}-\vec{\beta}^TX^TW\vec{y}+\vec{\beta}^TX^TWX\vec{\beta} \tag{14.1.560}$$

For this we differentiate with respect to $\vec{\beta}$. Obviously the first term will vanish so it remains only:

$$\vec{\nabla}_{\vec{\beta}}\text{WSSE}\propto\vec{\nabla}_{\vec{\beta}}\left(-\vec{y}^TWX\vec{\beta}-\vec{\beta}^TX^TW\vec{y}+\vec{\beta}^TX^TWX\vec{\beta}\right)=0 \tag{14.1.561}$$

Notice now that we have also using the properties of transposed matrices:

$$\vec{y}^TWX\vec{\beta}=(W\vec{y})^T(\vec{\beta}^TX^T)^T=(W\vec{y}\vec{\beta}^TX^T)^T=\vec{\beta}^TX^TW\vec{y} \tag{14.1.562}$$

without forgetting that as W is diagonal we have $W=W^T$. Therefore:

$$\begin{aligned}\vec{\nabla}_{\vec{\beta}}\text{WSSE}&\propto\vec{\nabla}_{\vec{\beta}}\left(-\vec{y}^TWX\vec{\beta}-\vec{\beta}^TX^TW\vec{y}+\vec{\beta}^TX^TWX\right)\\&=\vec{\nabla}_{\vec{\beta}}\left(-2\vec{\beta}^TX^TW\vec{y}+\vec{\beta}^TX^TWX\vec{\beta}\right)\\&=-2X^TW\vec{y}+2X^TWX\vec{\beta}=0\end{aligned} \tag{14.1.563}$$

Rearranging we get:

$$\boxed{\hat{\vec{\beta}}_{\text{WLS}}=\min_{\vec{\beta}}\sum_{i=1}^{n}(\sqrt{w_i}\varepsilon_i)^2=(X^TWX)^{-1}X^TW\vec{y}} \tag{14.1.564}$$

Again we notice that if W is a diagonal unitary matrix, we fall back on the ordinary multivariate regression coefficients.

Remark

Notice that exactly the same developments can lead us to the coefficient of the general ordinary least squares then given by:

$$\boxed{\hat{\vec{\beta}}_{\text{GLS}}=(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\vec{y}} \tag{14.1.565}$$

Since we can write $\Sigma=SS^T$, where S is a triangular matrix using the Cholesky decomposition (see section Linear Algebra page 2027), we have for the general sum of squares errors:

$$\begin{aligned}\text{GSSE}&=\left(\vec{y}-X\vec{\beta}\right)^T\Sigma^{-1}\left(\vec{y}-X\vec{\beta}\right)=\left(\vec{y}-X\vec{\beta}\right)^T S^{-T}S^{-1}\left(\vec{y}-X\vec{\beta}\right)\\&=\left(S^{-1}\vec{y}-S^{-1}X\vec{\beta}\right)^T\left(S^{-1}\vec{y}-S^{-1}X\vec{\beta}\right)\end{aligned} \tag{14.1.566}$$

So GLS is like regressing $S^{-1}X$ on $S^{-1}\vec{y}$. Furthermore:

$$\begin{aligned}\vec{y}&=X\vec{\beta}+\vec{\varepsilon}\\S^{-1}\vec{y}&=S^{-1}X\vec{\beta}+S^{-1}\vec{\varepsilon}\\\vec{y}\,'&=X'\vec{\beta}+\vec{\varepsilon}\,'\end{aligned} \tag{14.1.567}$$

So we have a new regression equation y' where as we can see from the GSSE can be assimilated a simple regression equation which has uncorrelated errors with equal variance.

In practice, for other types of dataset, the structure of W is usually unknown, so we have to perform an ordinary least squares (OLS) regression first. Provided the regression function is appropriate, the i-th squared residual from the OLS fit is an estimate of σ_i^2 and the i-th absolute residual is an estimate of σ_i (which tends to be a more useful estimator in the presence of outliers). The residuals are much too variable to be used directly in estimating the weights, w_i, so instead we use either the squared residuals to estimate a variance function or the absolute residuals to estimate a standard deviation function. We then use this variance or standard deviation function to estimate the weights.

Some key points regarding weighted least squares are:

- The difficulty, in practice to determine the right weights (in some cases, the values of the weights may be based on theory or prior research).
- Weighted least squares estimates of the coefficients will usually be nearly the same as the "ordinary" unweighted estimates[13]. In cases where they differ substantially, the procedure can be iterated until estimated coefficients stabilize (often in no more than one or two iterations); this is named "iteratively reweighted least squares".
- In designed experiments with large numbers of replicates, weights can be estimated directly from sample variances of the response variable at each combination of predictor variables.

---

[13] Same applies for the ANOVA of the regression obviously.

- Use of weights will (legitimately) impact the widths of statistical intervals.
- In the transformed model, the interpretation of the coefficient estimates can be difficult. In weighted least squares the interpretation remains the same as before.
- Weighted least squares gives us an easy way to remove one observation from a model by setting its weight equal to 0.
- We can also downweight outlier or influential points to reduce their impact on the overall model.

Remark

The reader must keep in mind that this technique of weights can be applied to numerous techniques of regression. Not only linear, but also polynomial, logistic and so on!

14.1.11.7 Model Selection Criteria

Much of modern scientific enterprise is concerned with the question of model choice. An experimenter or researcher collects data, often in the form of measurements on many different aspects of the observed units, and wants to study how these variables affect some outcome of interest. Which measures are important to the outcome? Which are not? Are there interactions between the variables that need to be taken into account? Statisticians are also naturally involved in the question of model selection, and so it should come as no surprise that many approaches have been proposed over the years for dealing with this key issue. Both frequentist and Bayesian schools have weighed in on the matter, spawning methods such as F tests, Akaike information criterion (AIC), Mallows's C_p, exhaustive search, stepwise, backward, and forward selection procedures, cross-validation, Bayes factors of various flavours (partial, intrinsic, pseudo, fractional, posterior), Bayesian information criterion (BIC), and Bayesian model averaging, to name some of the more popular and well-known methods. Some of these, such as stepwise selection, are algorithms for picking a 'good' (or maybe useful) model; others, for example, AIC, are criteria for judging the quality of a model. Given this wealth of choices, how is a statistician able to decide what to do? An approach that cannot be implemented or understood by the scientific community will not gain acceptance. This implies that, at the very least, we need a method that can be carried out easily and yields results that can be interpreted by scientifically and numerically literate end users. From a statistical point of view, we want a method that is coherent and general enough to handle a wide variety of problems. Among the demands we could make on our method would be that it obeys the likelihood principle, that it has some frequentist (asymptotic) justification, and that it corresponds to a Bayesian decision problem. So let's begin from easier do derive to the most complicated one keeping in mind that there is actually in this beginning of the 121st century (holocene calendar) no magic procedures to get you the "best model" (this still remains an unsolved problem)!

14.1.11.7.1 Mallows's C_p Linear Regression Prediction

We estimated our model by minimizing the mean squared error on our data[14]:

$$\widehat{\vec{\beta}}=\min_{\vec{b}}\frac{1}{n}(\vec{y}-X\vec{b})^T(\vec{y}-X\vec{b}) \tag{14.1.568}$$

Different linear models will amount to different choices of the design matrix X — we add or drop variables, we add or drop interactions or polynomial terms, etc., and this adds or removes columns from the design matrix (this is typical of how works the "subset method" in many statistical softwares). We might consider doing a selection among models themselves by minimizing the MSE. This may sometimes (but not always!) be a very bad idea, for a fundamental reason: Every model can be too optimistic about how well it will actually predict (sometimes the ultimate validation of a model is to test its predictive capacity).

Let's be very clear about what it would mean to predict well. The most challenging case would be that we see a new random point, with predictor values $X_1,\ldots X_p$ and response Y, and our old $\widehat{\vec{\beta}}$ has a small expected squared error:

$$\mathrm{E}\left(Y-\left(\hat{\beta}_0+\sum_{j=1}^{p}X_j\hat{\beta}_j\right)\right)^2 \tag{14.1.569}$$

Here both Y and the X's are random (hence the capital letters), so we might be asking the model for a prediction at a point it never saw before (of course if we have multiple identically distributed (X,Y) pairs, say q of them, the expected MSE over those q points is just the same as the expected squared error at one point.)

An easier task would be to ask the model for predictions at the same values of the predictor variables as before, but with different random noises. That is, we fit the model to:

$$\begin{aligned}\text{Training data}:&(X_i,y_i)_{i=1}^n, X_i=(x_{i1},x_{i2},\ldots,x_{i,p-1})\\ \text{Test data}:&(X_i,y_i^{\text{new}})_{i=1}^n, X_i=(x_{i1},x_{i2},\ldots,x_{i,p-1})\end{aligned} \tag{14.1.570}$$

The test data consist of new outcome data drawn from the same true model and at the same x-locations as the training data (notice that this is a practical construction to derive the criterion, but not a necessity to use this for model selection. However, you should not use model selection criteria for a specific range of x and assume you can predict well on a different range of x!).

Let us note for the training data the model as following:

$$\vec{Y}=X\vec{\beta}+\varepsilon \tag{14.1.571}$$

with $\varepsilon=\mathcal{N}(0,\sigma_\varepsilon^2)$ and for the test data:

$$\vec{Y}\,'=X\vec{\beta}+\varepsilon' \tag{14.1.572}$$

where ε and ε' are independent but identically distributed! The design matrix is the same, the true parameters $\vec{\beta}$ are the same, but the noise is different[15]. We now want to see if the coefficients we estimated from $(X,\vec{Y})$ can predict $(X,\vec{Y}\,')$. Since the only thing that's changed is the noise, if the coefficients can't predict well any more, that means that they were really just memorizing the noise, and not actually doing anything useful.

---

[14] The whole text and developments below are mostly a copy paste of the PDF of the professor Cosma Shalizi

[15] If we really are in an experimental setting, we really could get a realization of $\vec{Y}'$ just by running the experiment a second time. With surveys or with observational data, it would be harder to actually realize $\vec{Y}'$, but mathematically at least it's unproblematic.

Remark

If we have $p-1$ variables, there are 2^{p-1} possible subset models!

Our "out-of-sample expected MSE", then, is:

$$\text{E}\left(\frac{1}{n}\left(\vec{Y}\,'-X\hat{\vec{\beta}}\right)^T\left(\vec{Y}\,'-X\hat{\vec{\beta}}\right)\right) \tag{14.1.573}$$

It will be convenient to break this down into an average over data points, and to abbreviate $X\hat{\vec{\beta}}=\hat{\vec{m}}$, the vector of fitted values. Notice that since the predictor variables and the coefficients aren't changing, our predictions are the same both in and out of sample. At point i, we will predict $\hat{\vec{m}}_i$.

In this notation, then, the expected out-of-sample MSE is:

$$\text{E}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i'-\hat{m}_i)^2\right) \tag{14.1.574}$$

We will compare this to the "expected in-sample MSE":

$$\text{E}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{m}_i)^2\right) \tag{14.1.575}$$

Notice that $\hat{m}_i$ is a function of the y_i (among other things), so those are dependent random variables, while $\hat{m}_i$ and y_i' are completely statistically independent[16].

Break this down term by term. What's the expected value of the i^{th} in-sample squared error? Using Huygens relation given for recall by $\text{V}(X)=\text{E}(X^2)-\text{E}(X)^2$ we then have:

$$\begin{aligned}\text{E}\left((y_i-\hat{m}_i)^2\right)&=\text{V}(y_i-\hat{m}_i)+(\text{E}(y_i-\hat{m}_i))^2\\&=\text{V}(y_i)+\text{V}(\hat{m}_i)-2\text{cov}(y_i,\hat{m}_i)+(\text{E}(y_i)-\text{E}(\hat{m}_i))^2\end{aligned} \tag{14.1.576}$$

The covariance term is not (usually) zero, because, as already mentioned, $\hat{m}_i$ is a function of, in part, y_i.

On the other hand, what's the expected value of the i-th squared error on new data? Using again Huygens relation we get:

$$\begin{aligned}\text{E}\,(y_i'-\hat{m}_i)^2=&\text{V}(y_i'-\hat{m}_i)+(\text{E}(y_i'-\hat{m}_i))^2\\&=\text{V}(y_i')+\text{V}(\hat{m}_i)-2\text{cov}(y_i',\hat{m}_i)+(\text{E}(y_i')-\text{E}(\hat{m}_i))^2\end{aligned} \tag{14.1.577}$$

The y_i' is independent of y_i, but has the same distribution. This tells us that:

$$\text{E}(y_i')=\text{E}(y_i)\qquad \text{V}(y_i')=\text{V}(y_i)\qquad\text{but}\qquad \text{cov}(y_i',\hat{m}_i)=0 \tag{14.1.578}$$

---

[16] That might sound weird, but remember we're holding X fixed in this exercise, so what we mean is that knowing $\hat{m}_i$ doesn't give us an extra information about y_i' beyond what we'd get from knowing the values of the X variables.

So:

$$\begin{aligned}\mathrm{E}\left((y_i'-\hat{m}_i)^2\right)&=\mathrm{V}(y_i)+\mathrm{V}(\hat{m}_i)+(\mathrm{E}(y_i)-\mathrm{E}(\hat{m}_i))^2\\&=\mathrm{E}\left((y_i-\hat{m}_i)^2\right)+2\mathrm{cov}(y_i,\hat{m}_i)\end{aligned}\tag{14.1.579}$$

Averaging over data points:

$$\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i'-\hat{m}_i)^2\right)=\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{m}_i)^2\right)+\frac{2}{n}\sum_{i=1}^{n}\mathrm{cov}(y_i,\hat{m}_i)\tag{14.1.580}$$

Clearly, we need to get a handle on that sum of covariances.

For a linear model, though (using the results proved also during our study of influential points on page 6387):

$$\mathrm{cov}(y_i,\hat{m}_i)=\mathrm{cov}(y_i,\hat{y}_i)=\mathrm{cov}\left(y_i,\sum_{j=1}^{n}H_{ij}y_j\right)=\sum_{j=1}^{n}H_{ij}\mathrm{cov}\left(y_i,y_j\right)=\sigma_\varepsilon^2 H_{ii}\tag{14.1.581}$$

as for recall:

$$\mathrm{cov}\left(y_i,y_j\right)=\mathrm{cov}\left(\alpha+X_i\vec{\beta}_i+\varepsilon_i,\alpha+X_j\vec{\beta}_j+\varepsilon_j\right)=\mathrm{cov}\left(\varepsilon_i,\varepsilon_j\right)=\begin{cases}1 & \text{if } i=j\\ 0 & \text{if } i\neq j\end{cases}\tag{14.1.582}$$

So, for linear models:

$$\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i'-\hat{m}_i)^2\right)=\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{m}_i)^2\right)+\frac{2}{n}\sum_{i=1}^{n}\sigma_\varepsilon^2H_{ii}=\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{m}_i)^2\right)+\frac{2}{n}\sigma_\varepsilon^2\mathrm{tr}(H)\tag{14.1.583}$$

and we know that with p predictors and one intercept[17]:

$$\mathrm{tr}(H)=\mathrm{tr}(X(X^TX)^{-1}X^T)=\mathrm{tr}((X^TX)^{-1}(X^TX))=\mathrm{tr}(\mathbb{1}_p)=p+1\tag{14.1.584}$$

Thus, for linear models:

$$\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i'-\hat{m}_i)^2\right)=\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{m}_i)^2\right)+\frac{2}{n}\mathrm{tr}(H)=\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{m}_i)^2\right)+\frac{2}{n}\sigma_\varepsilon^2p\tag{14.1.585}$$

Of course, we don't actually know the expectation on the right-hand side, but we do have a sample estimate of it, which is the in-sample MSE. If the law of large numbers is still our friend:

$$\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i'-\hat{m}_i)^2\right)\cong\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{m}_i)^2+\frac{2}{n}\sigma_\varepsilon^2p\tag{14.1.586}$$

The second term on the right, $(2/n)\sigma_\varepsilon^2(p+1)$, is the "optimism" of the model, i.e. the amount by which its in-sample MSE systematically under-estimates its true expected squared error. Notice that this:

---

[17] Be careful, on some Internet forum, people use X_c instead of X then the result lead to p and not $p+1$ as the first component of the diagonal for X_c is equal to 0.

- Grows with σ_{ε}^2: more noise gives the model more opportunities to seem to fit well by capitalizing on chance.
- Shrinks with n: at any fixed level of noise, more data makes it harder to pretend the fit is better than it really is.
- Grows with p: every extra parameter is another control which can be adjusted to fit to the noise.

Minimizing the in-sample MSE completely ignores the bias from optimism, so it is guaranteed to pick models which are too large and predict poorly out of sample. If we could calculate the optimism term, we could at least use an unbiased estimate of the true MSE on new data. From one point of view, the optimism is just an estimate of the bias. From another point of view, it's a cost we're imposing on models for having extra parameters.

Definition (#14.1.560): If p regressors are selected from a set of $k>p$, the "Mallows's C_p" statistic for that particular set of regressors is defined as (it's the previous relation multiplied by n and divided by σ_{ε}^2):

$$C_p \cong \frac{\displaystyle\sum_{i=1}^{n}(y_i-\hat{m}_i)^2}{\sigma_{\varepsilon}^2}+2p=\frac{\text{SSE}_p}{\dfrac{\text{SSE}_q}{n-q}}-n+2p=\frac{\text{SSE}_p}{\text{MSE}_q}-n+2p \tag{14.1.587}$$

The C_p criterion, that compares the precision and bias of the full model to models with a subset of the predictors, suffers from two main limitations:

1. The C_p approximation is only valid for large sample size
2. The C_p cannot handle complex collections of models as in the variable selection (or feature selection[18]) problem.

The C_p statistic is often used as a stopping rule for various forms of stepwise regression (don't forget that it is more a feature selection technique than a regression one!).

Remark

One way to automatically select a model is to begin with the largest model you can, and then prune (simplify) it, which can be done in several way:

- Eliminate the least-significant coefficient (thanks to the p-value)
- Pick your favourite model selection criterion, consider deleting each coefficient in turn, and pick the sub-model with the best value of the criterion.

---

[18] Feature subset selection is the process of identifying and removing as much irrelevant and redundant information as possible. This reduces the dimensionality of the data and may allow learning algorithms to operate faster and more effectively. In some cases, accuracy on future classification can be improved; in others, the result is a more compact, easily interpreted representation of the target concept.

Having eliminated a variable, one then re-estimates the model, and repeats the procedure. Stop when either all the remaining coefficients are significant (under the first option), or nothing can be eliminated without worsening the criterion. That latter procedure is named "backward stepwise regression". The method that starts with intercept only and adds variables in the same fashion is named the "forward stepwise regression" (don't forget that these two methods are more feature selection techniques than regression one!). There are naturally forward-backward hybrids.

Because the inter-correlation between the regressors affect the order of term entry and removal. Since we are approaching the final model from two different directions this aspect can cause the methods to converge on different models.

Colin Lingwood Mallows proposed the statistic as a criterion for selecting among many alternative subset regressions. Under a model not suffering from appreciable lack of fit (bias), C_p has expectation nearly equal to p. Indeed, if your model with p parameters is correct it holds that $\text{SSE}_p \cong (n-p)\sigma_\varepsilon^2$. If your other model is already correct as well, it holds $\text{SSE}_q \cong (n-q)\sigma_\varepsilon^2$, therefore:

$$C_p \cong \frac{\text{SSE}_p}{\dfrac{\text{SSE}_q}{n-q}} - n + 2p = \frac{(n-p)\sigma_\varepsilon^2}{\dfrac{(n-q)\sigma_\varepsilon^2}{n-q}} - n + 2p = (n-p) - n + 2p = p \qquad (14.1.588)$$

Otherwise the expectation is roughly p plus a positive bias term.

Nevertheless, even though it has expectation greater than or equal to p, there is nothing to prevent $C_p < p$ or even $C_p < 0$ in extreme cases. It is suggested that one should choose a subset that has C_p approaching p, from above, for a list of subsets ordered by increasing p. In practice, the positive bias can be adjusted for by selecting a model from the ordered list of subsets, such that $C_p < 2p$.

To summarize, we should look for models where Mallows's C_p is small and close to the number of predictors in the model plus the constant p. A small Mallows's C_p value indicates that the model is relatively precise (has small variance) in estimating the true regression coefficients and predicting future responses. A Mallows's C_p value that is close to the number of predictors plus the constant indicates that the model is relatively unbiased in estimating the true regression coefficients and predicting future responses. Models with lack-of-fit and bias have values of Mallows's C_p larger than p.

14.1.11.7.2 Akaike Information Criterion (AIC)

The Akaike information criterion, AIC, was developed by Akaike to estimate the expected Kullback-Leibler discrepancy (relative quality of statistical models for a given set of dataset) between the model generating the data and a fitted candidate model. In instances where the sample size is large and the dimension of the candidate model is relatively small, AIC serves as an approximately unbiased estimator. In other settings, AIC may be characterized by a large negative bias which limits its effectiveness as a model selection criterion[19].

Sometimes we have a set of possible models and we want to choose the best model. Model selection methods help us choose a good model. Here are some examples:

✍**Examples:**

E1. Suppose we use a polynomial to model the regression function:

$$\mathrm{E}(Y/X=x)=\beta_0+\beta_1 x+\beta_2 x^2+\ldots+\beta_k x^k \tag{14.1.589}$$

We will need to choose the order of polynomial p. We can think of this as a sequence of models $\mathcal{M}_1,\ldots,\mathcal{M}_k$ index by k.

E2. Suppose you have data $y_1,\ldots,y_n$ on age at death of n people. You want to model the distribution of y. Some popular models are:

- $\mathcal{M}_1$: The exponential distribution: $p(y,\theta)=\theta e^{-\theta y}$
- $\mathcal{M}_2$: The gamma distribution: $p(y,a,b)=(b^a/\Gamma(a))y^{a-1}e^{-by}$
- $\mathcal{M}_3$: The log-normal distribution: we take $\log(y)\cong\mathcal{N}(\mu,\sigma^2)$

E3. Suppose you have time series data $y_1,y_2,\ldots$ A common model is the AR (autoregressive model):

$$y_t=a_1y_{t-1}+a_2y_{t-2}+\ldots+a_kY_{t-k}+\varepsilon_t \tag{14.1.590}$$

where $\varepsilon_t\cong\mathcal{N}(0,\sigma^2)$. The number k is the order of the model as we know. We need to choose k.

In a linear regression model, when we need to choose which variables to include in the regression based on return on experience or statistical indicators, we name this method "variable selection".

Here are five myths about variable selection (see [**heinze2017five**]):

- **Myth 1:** The number of variables in a model should be reduced until there are 10 events per variables (EPV) is wrong. The number of EPV depends on the desired statistical power of the study.

---

[19] For such instances, Hurvic and Tasi (11989 according to holocene calendar) proposed the corrected Akaike information criterion, AICc.

- **Myth 2:** Only variables with proven univariable-model significance should be included in a model is wrong. It depends on the context of the experiment and the return on experience (REX).

- **Myth 3:** Insignificant effects should be eliminated from a model is wrong. Eliminating weak effects may also be dangerous as in some studies, bias could result from falsely omitting an important confounder.

- **Myth 4:** The reported p-value quantifies the type I error of a variable being falsely selected is wrong. Multiple selection methods should be presented in any report: One based for example of on p-value, another on the AIC and another one on the BIC.

- **Myth 5:** Variable selection simplifies analysis is wrong. Sure it simplifies interpretation and comprehension for the decision board but analysis is automated nowadays therefore the simplification argument is not valid anymore in our era of modern computer software given the available computer processing power.

Remark

The AICc is useful for selecting between models in the same class. For example, we can use it to select an ARIMA model (see section Economy page 7989) between candidate ARIMA models or an ETS model between candidate ETS models (see section Economy page 7958). However, it cannot be used to compare between ETS and ARIMA models for example because they are most of the times based on too much different mathematical tools.

Suppose we have models $\mathcal{M}_1,\ldots,\mathcal{M}_k$ where each model is a set of densities (to simplify the notations we will consider only univariate models):

$$\mathcal{M}_j=\left\{p(y,\theta_j):\theta_j\in\vec{\Theta}_j\right\} \qquad (14.1.591)$$

We have data $y_1,\ldots,y_n$ drawn from some density f. We do not assume that f is any of the models $\mathcal{M}_j$!

Let $\hat{\theta}_j$ be the maximum likelihood estimator from model j. An estimate of P, based on model j is $\hat{p}_j(y)=p(y,\hat{\theta})$. The quality of $\hat{p}_j(y)$ as an estimate of f can be measured by the Kullback-Leibler divergence given for recall in its continuous form by (see section Statistical Mechanics page 3542):

$$D_{\text{KL}}(p_X,p_Y)=\int p_X(x)\log\left(\frac{p_X(x)}{p_Y(x)}\right)\mathrm{d}x \qquad (14.1.592)$$

That we will write here as:

$$D_{\text{KL}}(p,\hat{p}_j)=\int p(y)\log\left(\frac{p(y)}{\hat{p}_j(y)}\right)\mathrm{d}y=\int p(y)\log(p(y))\mathrm{d}y-\int p(y)\log(\hat{p}_j(y))\mathrm{d}y \qquad (14.1.593)$$

The first term does not depend on j. So minimizing $D_{\text{KL}}(p,\hat{p}_j)$ over j is the same as maximizing (don't forget that we have proved that D_{KL} was positive definite such that $D_{\text{KL}}\geq 0$):

$$D_{\text{KL},j}=\int p(y)\log(\hat{p}_j(y))\mathrm{d}y=\int p(y)\log(p(y,\hat{\theta}_j))\mathrm{d}y \qquad (14.1.594)$$

We need to estimate $D_{\text{KL,j}}$. Intuitively, a parametric estimate of $D_{\text{KL,j}}$ is given by the arithmetic average (not such as good as taking the median but more easy to deal for mathematical developments):

$$\bar{D}_{\text{KL},j}=\frac{1}{n}\sum_{i=1}^{n}\log\left(p(y_i,\hat{\theta}_j)\right) \tag{14.1.595}$$

However, apart from the fact that this estimate is non-robust, it is also probably quite biased. Akaike proved that the bias is approximately d_j/n where $d_j=\text{dim}(\vec{\Theta}_j)$. Therefore we use for the model $\mathcal{M}_j$:

$$\hat{D}_{\text{KL},j}\cong\bar{D}_{\text{KL},j}-\frac{d_j}{n} \tag{14.1.596}$$

That is to say a measure of a in-sample performance plus a penalty (or bias). Defining on the way the "Akaike information criterion[20]" (AIC):

$$\boxed{\text{AIC}(j)=2n\hat{D}_{\text{KL},j}\cong 2n\bar{D}_{\text{KL},j}-2d_j} \tag{14.1.597}$$

Notice that maximizing $\hat{D}_{\text{KL},j}$ is the same as maximizing $\text{AIC}(j)$ over j. But keep in mind that from an information theory point of view, the divergence $\hat{D}_{\text{KL},j}$ tells us how much information we lost due to our approximation of probability distribution with respect to the true probability distribution. When comparing models, we choose the models with the lowest AIC because in turn it means that the Kullback-Leibler divergence also would be minimum. Low AIC means little information bias.

And why do we multiply by $2n$? Just for historical reasons! We can multiply by any constant (that's why Data Scientists typically multiply by -1; it won't change which model we pick. In fact, different textbooks and also softwares (like R and Python™) use different versions of AIC. A few common notations we can find in Data Science textbooks are for example :

$$\text{AIC}=2k-2\ln(\mathcal{L})=2\ln\left(\frac{e^k}{\hat{\mathcal{L}}}\right) \tag{14.1.598}$$

Written in that way a smaller AIC score is preferable to a larger score. Using the rewritten formula, one can see how the AIC score of the model will increase in proportion to the growth in the value of the numerator, which contains the number of parameters in the model (i.e. a measure of model complexity). And the AIC score will decrease in proportion to the growth in the denominator which contains the maximized log likelihood of the model (which, as we just saw, is a measure of the goodness-of-fit of the model). Keep in mind that the AIC score is useful only when its used to compare two models! The $2k$ indicated well that the less parameters we will have, the less likely will be the overfitting of the model.

---

[20] Actually, in his original paper, Hirotugu Akaike, proposed using the factor 2 to simplify some calculation involving chi-squared distribution. Many subsequent specialists have since kept the factor of 2 which of course will not change which model is selected. Also some authors define AIC as negative of this, and then minimize it; again, clearly the same thing!

Remark

Some opine why the need of for AIC when cross-validation techniques are already available. Well, cross validation helps in assessing only the predictive accuracy of the model at different section of the data. Cross validation is a diagnostic not an overfitting prevention strategy. Moreover cross validation does not provide information on how parsimonious the model (number of parameters) is.

Proof 14.3.2. Let us recall that during the study of the Fisher information matrix (see section Statistics page 839) we have defined the score function for a parameter θ_j of a distribution function depending of a set of parameters $\vec{\theta}$ as (keep in mind that we can take the $\ln$ or the $\log$, the result remains the same!):

$$\mathcal{S}_j(\vec{\theta})=\frac{\partial\mathcal{L}(\vec{\theta})}{\partial\theta_j}(\hat{\theta})=\frac{\partial\ln(L(\vec{\theta}))}{\partial\theta_j}(\hat{\theta}) \quad (14.1.599)$$

We know that the score function is used to build the estimate of parameters of the distribution (when we put the score equal to zero). And that therefore if the true parameters are known, then it always equal to zero (at least for non-strange distribution functions...)!

Let us recall that also during the study of this of bivariate Taylor series (see section Sequences and Series page 1715), we have introduced the Hessian matrix, the matrix of second derivatives, that we will denote $H(\vec{\theta})$ (meaning implicitly that it is evaluated at $\hat{\vec{\theta}}$).

Let us recall that during our study of the central limit theorem (see section Statistics page 938) we have proved that:

$$Z_n=\frac{\hat{\mu}-\mu}{\sigma/\sqrt{n}}\cong\mathcal{N}(0,1) \quad (14.1.600)$$

Therefore:

$$\sqrt{n}\,(\hat{\mu}-\mu)=\mathcal{N}(0,\sigma) \quad (14.1.601)$$

For a bivariate or multivariate Normal distribution (see section Statistics page 713), this will be written:

$$\vec{Z}_n=\sqrt{n}\left(\vec{\hat{\mu}}-\vec{\mu}\right)=\mathcal{N}(0,\Sigma) \quad (14.1.602)$$

For the case that interest us here, let us denote this:

$$\vec{Z}_n=\sqrt{n}\left(\hat{\vec{\theta}}-\vec{\theta}_0\right)=\mathcal{N}(0,\Sigma(\vec{\theta})) \quad (14.1.603)$$

With as we have seen during our study of the Fisher information Matrix (see section Statistics page 839):

$$\Sigma(\vec{\theta})\cong\left[\mathrm{E}(H^{-1})\mathcal{I}(\vec{\theta})\mathrm{E}(H^{-1})\right] \quad (14.1.604)$$

Now that we have finish the recalls. Let us take the Taylor series of:

$$D_{\text{KL}}=\int p(y)\log(\hat{p}(y))\mathrm{d}y=\int p(y)\log(p(y,\hat{\theta}))\mathrm{d}y \quad (14.1.605)$$

considering $\hat{\vec{\theta}}$ as the variable and doing the development around the true value of the estimator that we will denote $\vec{\theta}_0$:

$$\begin{aligned}D_{\text{KL}}&\cong\int p(y)\left(\log(p(y,\vec{\theta}_0))+(\hat{\vec{\theta}}-\vec{\theta}_0)^T\frac{\partial p(y,\vec{\theta})}{\partial\vec{\theta}}(\vec{\theta}_0)+\frac{1}{2}(\hat{\vec{\theta}}-\vec{\theta}_0)^T H(y,\vec{\theta})(\vec{\theta}_0)(\hat{\vec{\theta}}-\vec{\theta}_0)\right)\mathrm{d}y\\&=\int p(y)\left(\log(p(y,\vec{\theta}_0))+(\hat{\vec{\theta}}-\vec{\theta}_0)^T\vec{\mathcal{S}}(\vec{\theta}_0)+\frac{1}{2}(\hat{\vec{\theta}}-\vec{\theta}_0)^T H(y,\vec{\theta})(\vec{\theta}_0)(\hat{\vec{\theta}}-\vec{\theta}_0)\right)\mathrm{d}y\end{aligned}\tag{14.1.606}$$

At the true value $\vec{\theta}_0$ we have (don't forget that this partial derivative is the score, i.e. the partial derivative of the likelihood and that latter is always equal to zero for the true model with the real parameters!):

$$\vec{\mathcal{S}}(\vec{\theta}_0)=\frac{\partial p(y,\vec{\theta})}{\partial\vec{\theta}}(\vec{\theta}_0)=\vec{0}\tag{14.1.607}$$

Then it remains:

$$D_{\text{KL}}\cong\int p(y)\left(\log(p(y,\vec{\theta}_0))+\frac{1}{2}(\hat{\vec{\theta}}-\vec{\theta}_0)^T H(y,\vec{\theta})(\vec{\theta}_0)(\hat{\vec{\theta}}-\vec{\theta}_0)\right)\mathrm{d}y\tag{14.1.608}$$

That we can also write:

$$\begin{aligned}D_{\text{KL}}&\cong\int p(y)\left(\log(p(y,\vec{\theta}_0))+\frac{1}{2}\frac{\sqrt{n}(\hat{\vec{\theta}}-\vec{\theta}_0)^T H(y,\vec{\theta})(\vec{\theta}_0)\sqrt{n}(\hat{\vec{\theta}}-\vec{\theta}_0)}{n}\right)\mathrm{d}y\\&=\int p(y)\left(\log(p(y,\vec{\theta}_0))+\frac{1}{2}\frac{\vec{Z}_n^T H(y,\vec{\theta})(\vec{\theta}_0)\vec{Z}_n}{n}\right)\mathrm{d}y\\&=\int p(y)\log(p(y,\vec{\theta}_0))\mathrm{d}y+\frac{1}{2n}\int p(y)\left(\vec{Z}_n^T H(y,\vec{\theta})(\vec{\theta}_0)\vec{Z}_n\right)\mathrm{d}y\end{aligned}\tag{14.1.609}$$

The $\vec{Z}_n$ doesn't depend explicitly on y, then we can write:

$$D_{\text{KL}}\cong\int p(y)\log(p(y,\vec{\theta}_0))\mathrm{d}y+\frac{1}{2n}\vec{Z}_n^T\left(\int p(y)H(y,\vec{\theta})(\vec{\theta}_0)\mathrm{d}y\right)\vec{Z}_n\tag{14.1.610}$$

Now let us recall that we have seen during our study of the Fisher Information matrix (see section Statistics page 839) that:

$$\mathrm{E}\left(\mathcal{I}(\vec{\theta})\right)=-\mathrm{E}\left(H(\vec{\theta})\right)=-\mathrm{E}\left(\frac{\partial^2\mathcal{L}(\vec{x},\vec{\theta})}{\partial\theta_i\partial\theta_j}\right)=-\int\frac{\partial^2\ln(L(\vec{x},\vec{\theta}))}{\partial\theta_i\partial\theta_j}L(\vec{x},\vec{\theta})\mathrm{d}x\tag{14.1.611}$$

So in the previous integral between the parenthesis we have the same expression apart of the notation: p instead of L and $\log$ instead on $\ln$ but however the manipulated objects are the same! Hence:

$$D_{\text{KL}}\cong\int p(y)\log(p(y,\vec{\theta}_0))\mathrm{d}y+\frac{1}{2n}\vec{Z}_n^T\mathrm{E}\left(H(\vec{\theta})\right)\vec{Z}_n\tag{14.1.612}$$

That we will denote to simplify the next developments as following:

$$D_{\text{KL}}\cong D_{\text{KL},0}+\frac{1}{2n}\vec{Z}_n^T J\vec{Z}_n\tag{14.1.613}$$

Now let us do the same Taylor series but for the approximate model, with the real data:

$$\begin{aligned}\bar{D}_{\text{KL}} &\cong \frac{1}{n}\sum_{i=1}^{n}\left(\log(\hat{p}(y_i,\vec{\theta}_0))+(\vec{\hat{\theta}}-\vec{\theta}_0)^T\frac{\partial\hat{p}(y_i,\vec{\theta})}{\partial\vec{\theta}}(\vec{\theta}_0)+\frac{1}{2}(\vec{\hat{\theta}}-\vec{\theta}_0)^T H(y_i,\vec{\theta})(\vec{\theta}_0)(\vec{\hat{\theta}}-\vec{\theta}_0)\right)\\ &=\frac{1}{n}\sum_{i=1}^{n}\log(\hat{p}(y_i,\vec{\theta}_0))+(\vec{\hat{\theta}}-\vec{\theta}_0)^T\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\hat{p}(y_i,\vec{\theta})}{\partial\vec{\theta}}(\vec{\theta}_0)\right)+\frac{1}{2}(\vec{\hat{\theta}}-\vec{\theta}_0)^T\left(\frac{1}{n}\sum_{i=1}^{n}H(y_i,\vec{\theta})(\vec{\theta}_0)\right)(\vec{\hat{\theta}}-\vec{\theta}_0)\\ &=\bar{D}_{\text{KL},0}+\frac{\vec{Z}_n^T}{\sqrt{n}}\left(\overline{\frac{\partial\hat{p}(y_i,\vec{\theta})}{\partial\vec{\theta}}(\vec{\theta}_0)}\right)+\frac{1}{2n}\vec{Z}_n^T\left(\overline{H(y_i,\vec{\theta})(\vec{\theta}_0)}\right)\vec{Z}_n\\ &=\bar{D}_{\text{KL},0}+\frac{\vec{Z}_n^T}{\sqrt{n}}\left(\overline{\frac{\partial\hat{p}(y_i,\vec{\theta})}{\partial\vec{\theta}}(\vec{\theta}_0)}\right)+\frac{1}{2n}\vec{Z}_n^T\overline{H(\vec{\theta})}\vec{Z}_n\end{aligned} \tag{14.1.614}$$

Now we assume (...) when $n\rightarrow+\infty$ that:

$$\bar{D}_{\text{KL},0}\rightarrow D_{\text{KL},0}\qquad\text{and}\qquad\overline{H(\vec{\theta})}\rightarrow J \tag{14.1.615}$$

Therefore:

$$\bar{D}_{\text{KL}}\cong\bar{D}_{\text{KL},0}+\frac{\vec{Z}_n^T}{\sqrt{n}}\left(\overline{\frac{\partial\hat{p}(y_i,\vec{\theta})}{\partial\vec{\theta}}(\vec{\theta}_0)}\right)+\frac{1}{2n}\vec{Z}_n^T\overline{H(\vec{\theta})}\vec{Z}_n=D_{\text{KL},0}+\frac{\vec{Z}_n^T}{\sqrt{n}}\left(\overline{\frac{\partial\hat{p}(y_i,\vec{\theta})}{\partial\vec{\theta}}(\vec{\theta}_0)}\right)+\frac{1}{2n}\vec{Z}_n^T J\vec{Z}_n \tag{14.1.616}$$

Remark

Notice that for obscure reasons we don't assume that for $n\rightarrow+\infty$:

$$\frac{\partial\hat{p}(y_i,\vec{\theta})}{\partial\vec{\theta}}(\vec{\theta}_0)\rightarrow\vec{0} \tag{14.1.617}$$

Then:

$$\begin{aligned}\bar{D}_{\text{KL}}-D_{\text{KL}}&\cong\left(D_{\text{KL},0}+\frac{\vec{Z}_n^T}{\sqrt{n}}\left(\overline{\frac{\partial\hat{p}(y_i,\vec{\theta})}{\partial\vec{\theta}}(\vec{\theta}_0)}\right)+\frac{1}{2n}\vec{Z}_n^T J\vec{Z}_n\right)-\left(D_{\text{KL},0}+\frac{1}{2n}\vec{Z}_n^T J\vec{Z}_n\right)\\ &\cong\frac{\vec{Z}_n^T}{\sqrt{n}}\left(\overline{\frac{\partial\hat{p}(y_i,\vec{\theta})}{\partial\vec{\theta}}(\vec{\theta}_0)}\right)=\frac{\vec{Z}_n^T\overline{\vec{S}(\vec{\theta}_0)}}{\sqrt{n}}=\frac{\vec{Z}_n^T\vec{\hat{S}}(\vec{\theta}_0)}{\sqrt{n}}=\frac{\sqrt{n}\vec{Z}_n^T\vec{\hat{S}}(\vec{\theta}_0)}{n}=\frac{\vec{Z}_n^T\sqrt{n}\vec{\hat{S}}(\vec{\theta}_0)}{n}\end{aligned} \tag{14.1.618}$$

Now let us recall that during our study of the Fisher information matrix (see section Statistics page 839) we have proved that:

$$\sqrt{n}\vec{\hat{S}}(\theta_0)=\mathcal{N}(0,\mathcal{I}(\vec{\theta})) \tag{14.1.619}$$

And we have shown earlier above that:

$$\vec{Z}_n=\sqrt{n}\left(\vec{\hat{\theta}}-\vec{\theta}_0\right)=\mathcal{N}(0,\Sigma(\vec{\theta}))\cong\mathcal{N}(0,\mathrm{E}(H^{-1})\mathcal{I}(\vec{\theta})\mathrm{E}(H^{-1})) \tag{14.1.620}$$

As Z_n is a vector, and $-\mathrm{E}(H)$ is a non-random matrix, according to the following relation we have proved during our study of the variance-covariance matrix:

$$\mathrm{V}(AX)=A\Sigma A^T \tag{14.1.621}$$

We then have for the transformed random variable $-\text{E}(H)Z_n$:

$$\text{V}(-\text{E}(H)Z_n)=\text{E}(H)\Sigma(\vec{\theta})\text{E}(H)^T\cong\text{E}(H)\left(\text{E}(H^{-1})\mathcal{I}(\vec{\theta})\text{E}(H^{-1})\right)\text{E}(H)^T \tag{14.1.622}$$

As $\text{E}(H)$ is symmetric, we have $\text{E}(H)^T=\text{E}(H)$ then:

$$\text{V}(\text{E}(H)Z_n)\cong-\text{E}(H)\left(\text{E}(H^{-1})\mathcal{I}(\vec{\theta})\text{E}(H^{-1})\right)(-\text{E}(H)^T)=\text{E}(H)\text{E}(H^{-1})\mathcal{I}(\vec{\theta})\text{E}(H^{-1})\text{E}(H)=\mathcal{I}(\vec{\theta}) \tag{14.1.623}$$

But we have proved during our study of the Fisher information matrix (see section Statistics page 839) the information matrix equality $-\text{E}(H)(\vec{\theta})=\mathcal{I}(\vec{\theta})$. Therefore:

$$\text{V}(\mathcal{I}(\vec{\theta})Z_n)=\mathcal{I}(\vec{\theta}) \tag{14.1.624}$$

ie:

$$\mathcal{I}(\vec{\theta})Z_n=\mathcal{N}(0,\mathcal{I}(\vec{\theta})) \tag{14.1.625}$$

and we recognize here $\sqrt{n}\vec{\hat{S}}(\theta_0)$. So that finally:

$$\mathcal{I}(\vec{\theta})Z_n\cong\sqrt{n}\vec{\hat{S}}(\theta_0) \tag{14.1.626}$$

So:

$$\bar{D}_{\text{KL}}-D_{\text{KL}}\cong\frac{\vec{Z}_n^T\sqrt{n}\vec{\hat{S}}(\vec{\theta}_0)}{n}\cong\frac{\vec{Z}_n^T\mathcal{I}(\vec{\theta})Z_n}{n} \tag{14.1.627}$$

Now we will use the following relation also proved during our study of the variance-covariance matrix (see section Statistics page 607):

$$\text{E}[X^TAX]=\text{tr}(A\Sigma)+\vec{\mu}^TA\vec{\mu} \tag{14.1.628}$$

Therefore:

$$\text{E}(\bar{D}_{\text{KL}}-D_{\text{KL}})\cong\frac{\text{E}(\vec{Z}_n^T\mathcal{I}(\vec{\theta})Z_n)}{n}=\frac{\text{tr}(\mathcal{I}(\vec{\theta})\Sigma(\vec{\theta}))+\mu_Z^T\mathcal{I}(\vec{\theta})\mu_Z}{n} \tag{14.1.629}$$

But by construction, $\vec{Z}_n$ we have $\vec{\mu}=\vec{0}$ and also using the matrix information equality and the definition of the Fisher information matrix:

$$\text{E}(\bar{D}_{\text{KL}}-D_{\text{KL}})\cong\frac{\text{tr}((-\text{E}(H))(-\text{E}(H^{-1})))+\vec{0}}{n}=\frac{\text{tr}((\text{E}(H))\text{E}(H^{-1}))}{n}=\frac{\text{tr}(\mathbb{1})}{n}:=\frac{d}{n} \tag{14.1.630}$$

ie:

$$\text{E}(\bar{D}_{\text{KL}})-\text{E}(D_{\text{KL}})=\bar{D}_{\text{KL}}-\hat{D}_{\text{KL}}\cong\frac{d}{n} \tag{14.1.631}$$

So that finally:

$$\boxed{\hat{D}_{\text{KL}}\cong\bar{D}_{\text{KL}}-\frac{d}{n}=\frac{1}{n}\sum_{i=1}^{n}\log\left(p(y_i,\hat{\theta}_j)\right)-\frac{d}{n}} \tag{14.1.632}$$

Remembering as we have proved it that the bigger is the better!

Remark

As we can factorize the $1/n$ and that when we compare two or more AIC by subtracting them the constants vanishes (the number of parameters d of different models may not be the same!), the latter relation is often used in the following form:

$$\hat{D}_{\text{KL}} \cong \sum_{i=1}^{n} \log\left(p(y_i,\hat{\theta}_j)\right) - d = \mathcal{L}(y_i,\hat{\theta}_j) - d \quad (14.1.633)$$

that will be used as the definition of the AIC itself (without any $2n$ factor)! These different possible choices explains why some software packages may well give completely different AICs on the same data for the same model!

■ Q.E.D.

As the reader has very likely (...) noticed it, there are several steps where we are making a bunch of approximations and assumptions! Some of these approximations (especially those involving the Taylor expansions) can be shown to be OK asymptotically (i.e. as $n \rightarrow +\infty$) by more careful maths. The last steps however, where we invoke the Fisher information matrix (see section Statistics page 839) and the score are rather more dubious. So AIC is a very crude indicator. Cross-validation is much more reliable!

✍**Examples:**

E1. Let:

$$Y_n \cong \mathcal{N}(\mu,1) \quad (14.1.634)$$

We want to compare two models (notice that the first one has zero parameters to estimate and the second one has only one parameter to estimate: μ):

$$\mathcal{M}_0 : \mathcal{N}(0,1) \qquad \text{and} \qquad \mathcal{M}_1 : \mathcal{N}(\mu,1) \quad (14.1.635)$$

We want to test:

$$H_0 : \mu = 0 \qquad \text{versus} \qquad H_1 : \mu \neq 0 \quad (14.1.636)$$

The test statistics is as we already know:

$$Z = \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} = \frac{\bar{Y}-0}{1/\sqrt{n}} = \sqrt{n}\bar{Y} \quad (14.1.637)$$

We reject H_0 if $|Z| > Z_\alpha/2$. For $\alpha = 0.05$, we reject H_0 if $|Z| > 2$, i.e. if:

$$|\bar{Y}| > \frac{2}{\sqrt{n}} \quad (14.1.638)$$

The likelihood is obviously proportional to (we ignore the value of the variance as for both models it is equal to 1):

$$L(\mu) = \prod_{i=1}^{n} e^{-(Y_i-\mu)^2/2} = e^{-\sum_n (Y_i-\mu)^2/2} \quad (14.1.639)$$

Hence:

$$\log(L(\mu)) = \mathcal{L}(\mu) = -\sum_n (Y_i - \mu)^2/2 \tag{14.1.640}$$

Recall that one of the definition of the Akaike information criterion is $\text{AIC} = \mathcal{L}(Y_i, \hat{\theta}_j) - d$. The AIC scores are then:

$$\text{AIC}_0 = \mathcal{L}(Y_i, 0) - d = -\sum_n \frac{(Y_i - \mu)^2}{2} - d = -\sum_n \frac{(Y_i - 0)^2}{2} - 0 = -\frac{1}{2}\sum_n Y_i^2 \tag{14.1.641}$$

And:

$$\text{AIC}_1 = \mathcal{L}(Y_i, \hat{\mu}) - d = -\sum_n \frac{(Y_i - \hat{\mu})^2}{2} - d = -\sum_n \frac{(Y_i - \hat{\mu})^2}{2} - 1 \tag{14.1.642}$$

After, depending on the experimental values of the Y_i, we have to compare:

$$\text{AIC}_0 = -\frac{1}{2}\sum_n Y_i^2 \qquad \text{and} \qquad \text{AIC}_1 = -\sum_n \frac{(Y_i - \hat{\mu})^2}{2} - 1 \tag{14.1.643}$$

Keeping the one that has the biggest value (as in this example their both negative, that means the one that is the nearest to zero, or the smallest AIC in absolute value)!

E2. Remember that during our study of multiple linear regression, we have proved that:

$$\ln\left(L\left(\vec{\beta}, \sigma\right)\right) = -\frac{n}{2}\left(\ln(2\pi) + \ln(\sigma^2)\right) - \frac{1}{2\sigma^2}\left(\vec{y} - X\vec{\beta}\right)^T\left(\vec{y} - X\vec{\beta}\right) \tag{14.1.644}$$

And also earlier we have proved that:

$$\hat{\sigma}^2 = \frac{1}{n}\left(\vec{y} - X\vec{\beta}\right)^T\left(\vec{y} - X\vec{\beta}\right) = \frac{1}{n}\left\|\left(\vec{y} - X\vec{\beta}\right)^2\right\| \tag{14.1.645}$$

denoted also sometimes:

$$\hat{\sigma}^2 = \frac{1}{n}\text{SSE}(\mathcal{M}) \tag{14.1.646}$$

Recall that another definition of the Akaike information criterion (the most common in statistical softwares) is $\text{AIC} = -2\mathcal{L}(Y_i, \hat{\theta}_j) + 2d$. The AIC score is then when σ is unknown for a multiple regression:

$$\begin{aligned}\text{AIC} &= -2\mathcal{L}(Y_i, \hat{\theta}_j) + 2d = -2\left(-\frac{n}{2}\left(\ln(2\pi) + \ln(\hat{\sigma}^2)\right) - \frac{1}{2\hat{\sigma}^2}n\hat{\sigma}^2\right) + 2d \\ &= n\left(\ln(2\pi) + \ln(\hat{\sigma}^2)\right) + n + 2d = n\left(\ln(2\pi) + \ln\left(\frac{\text{SSE}(\mathcal{M})}{n}\right)\right) + n + 2d\end{aligned} \tag{14.1.647}$$

What interest us when we use the AIC is to compared different models (subtract the different AIC). As n remains constant between all models, we can eliminate some useless terms that will automatically vanish, this is why in many textbooks we will found:

$$\text{AIC} = n\ln\left(\frac{\text{SSE}(\mathcal{M})}{n}\right) + 2d \tag{14.1.648}$$

And we have seen earlier that the Fisher information matrix of the multiple linear regression had $p+1$ dimension, therefore:

$$\text{AIC} = n\ln\left(\frac{\text{SSE}(\mathcal{M})}{n}\right) + 2(p+1) \tag{14.1.649}$$

What is of denoted in some textbooks as:

$$\text{AIC} = n\ln\left(\frac{\text{SSE}(\mathcal{M})}{n}\right) + 2\left(p((\mathcal{M}))+1\right) \tag{14.1.650}$$

Remark

A software like `R` deals with $\text{AIC} = -2\mathcal{L}(Y_i,\hat{\theta}_j)+2d$ that we have used above. Therefore the smaller is better!

The most common model selections methods are (among many others as we can see on the Data Science map at page 6770):

1. p-value
2. AIC (and related methods like C_p)
3. Cross-validation (see further below)
4. BIC (see further below)

AIC is motivated by the estimation of the generalization error (like Mallows's C_p, BIC,...). If you want the model for predictions, better use one of these criteria. If you want your model for explaining a phenomenon, use p-values. We need to distinguish between two goals:

1. Find the model that gives the best prediction (without assuming that any of the models are correct).
2. Assume one of the models is the true model and not the "true" model.

Generally speaking, AIC and cross-validation are used for goal (1) while BIC is used for goal (2).

AIC is founded on information theory. When a statistical model is used to represent the process that generated the data, the model will almost never be exact; so some information will be lost by using the model to represent the process. AIC estimates the relative information lost by a given model: the less information a model loses, the higher the quality of that model.

To apply AIC in practice, we start with a set of candidate models, and then find the models' corresponding AIC values. There will almost always be information lost due to using a candidate model to represent the "true model" (i.e. the process that generated the data). We wish to select, from among the candidate models, the model that minimizes the information loss. We cannot choose with certainty, but we can minimize the estimated information loss.

14.1.11.7.3 Cross-Validation metrics

In Machine Learning, "cross-validation" (CV) is a resampling method used for model evaluation to avoid testing a model on the same dataset on which it was trained. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. This is a common mistake, especially that a separate testing dataset is not always available. However, this usually leads to inaccurate performance measures (as the model will have an almost perfect score since it is being tested on the same data it was trained on). To avoid this kind of mistakes, cross validation is usually preferred.

Two types of cross-validation can be distinguished, exhaustive and non-exhaustive cross-validation:

- Exhaustive cross-validation methods are cross-validation methods which learn and test on all possible ways to divide the original sample into a training and a validation set.
- Non-exhaustive cross validation methods do not compute all ways of splitting the original sample.

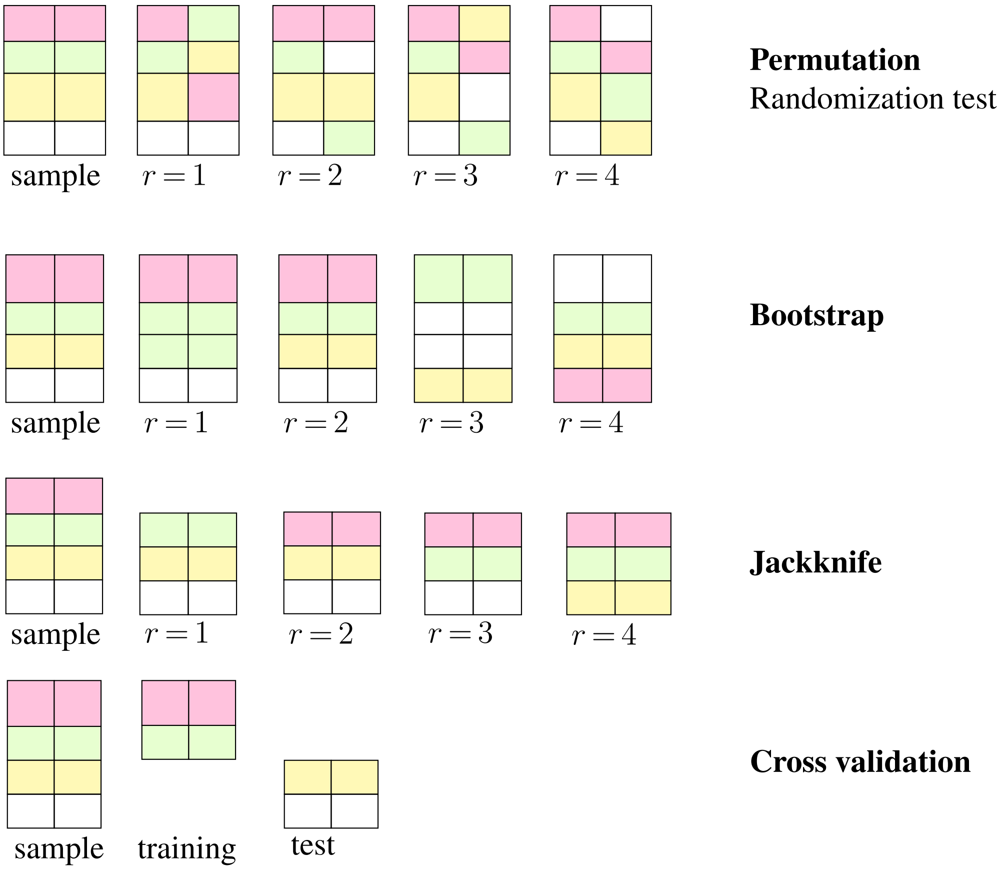

sample
r = 1
r = 2
r = 3
r = 4
Permutation
Randomization test
sample
r = 1
r = 2
r = 3
r = 4
Bootstrap
sample
r = 1
r = 2
r = 3
r = 4
Jackknife
sample
training
test
Cross validation

Figure 14.18: **Some famous techniques of cross-over validation**

The concept of cross-validation is actually simple: Instead of using the whole dataset to train and then test on same data, we could randomly divide our data into training and testing datasets.

There are several types of cross-validation methods (LOOCV – Leave-One-Out-Cross validation[21], the holdout method, K-fold cross validation). Here, we will discuss the "K-Fold cross validation method".

The idea is illustrated as following:

---

[21] We already know that latter as it is just the Jackknife method

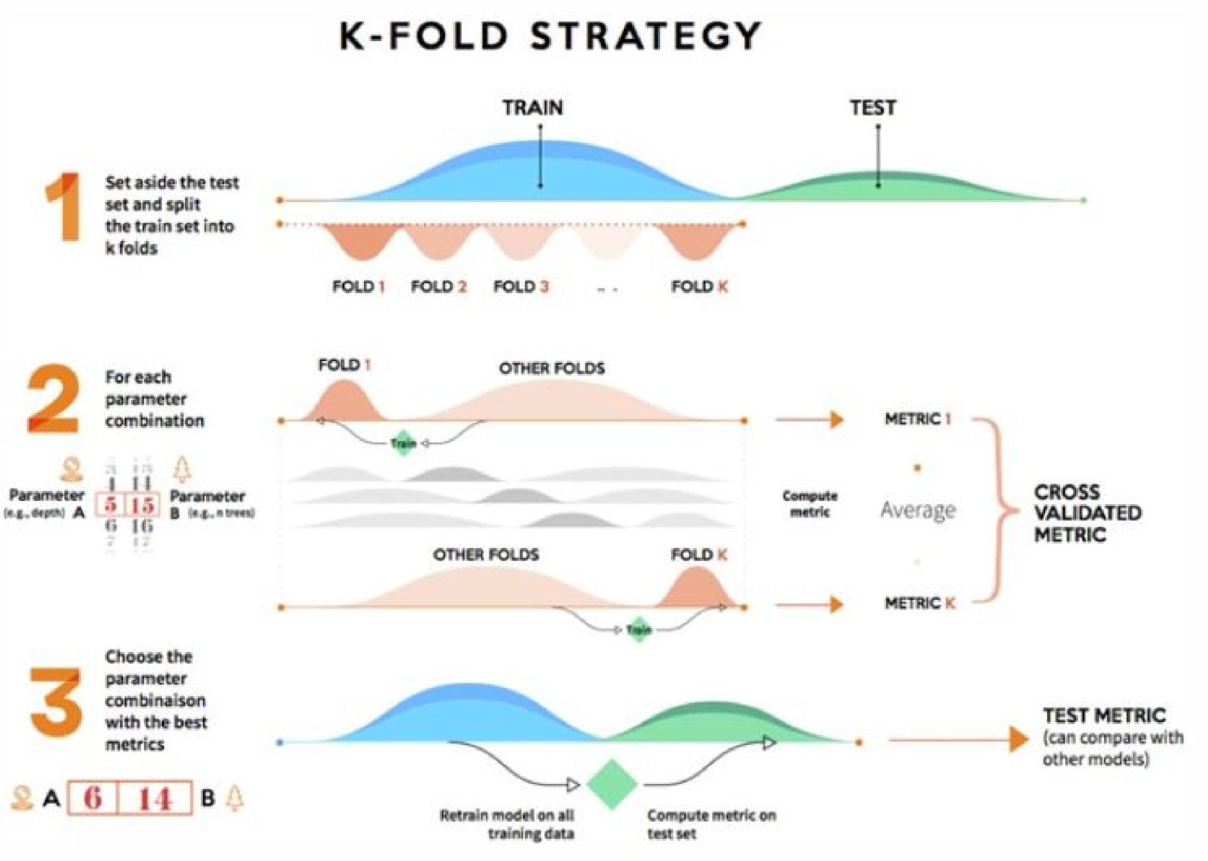

K-FOLD STRATEGY
TRAIN
TEST
1
Set aside the test set and split the train set into k folds
FOLD 1
FOLD 2
FOLD 3
FOLD K
2
For each parameter combination
Parameter (e.g. depth) A
5
15
Parameter B (e.g. n trees)
FOLD 1
OTHER FOLDS
Train
OTHER FOLDS
FOLD K
Compute metric
METRIC 1
Average
METRIC K
CROSS VALIDATED METRIC
3
Choose the parameter combination with the best metrics
A
6
14
B
Retrain model on all training data
Compute metric on test set
TEST METRIC (can compare with other models)

Figure 14.19: K-fold cross-over validation (author: ?)

So following the above illustration, the K-fold basically consists of the below steps:

1. Split the data in a train and a test set
2. Randomly split the train set into K subsets, also named "folds".
3. Train the model on all K subsets excepted the first one
4. Fit the model on the first set that was ignored previously and keep its quality metric (i.e. "cross validated metric")
5. Repeat the procedure (3) to (4) (but by ignoring the second set, after the third, and so on...)
6. Average all the quality metrics (ordinary cross-validation: OCV)
7. At the end, use all the K-folds to train the model and fit it on the test set
8. Compare if the last fit perform better that the average of all previous fits
9. If not, use one of the K as reference model based on the quality metric. If yes, keep the model made in step (8)

The typical basis of cross-over quality metrics is the mean square error MSE. Then the average quality metric, or the "ordinary cross-validation", for a K fold will be (assuming that each fitted model has a sub-sample size of n_k):

$$\text{OCV}=\frac{1}{K}\sum_{k=1}^{K}n_k\text{MSE}_k \tag{14.1.651}$$

If we denoted the model (in a univariate case) by $\hat{y}_i^{[k]}$ for the subset (submodel) k and the true values by y_i we then have:

$$\text{MSE}_k=\frac{1}{n_k}\sum_{i=1}^{n_k}\left(\hat{y}_i^{[k]}-y_i\right)^2 \tag{14.1.652}$$

Therefore the OCV for an univariate model based one the MSE is :

$$\boxed{\text{OCV}_{\text{MSE}}=\frac{1}{K}\sum_{k=1}^{K}\left(\sum_{i=1}^{n_k}\left(\hat{y}_i^{[k]}-y_i\right)^2\right)} \tag{14.1.653}$$

For example with LASSO, ridge or elastic net regularized regression, we have:

$$\text{OCV}_{\text{MSE}}(\lambda)=\frac{1}{K}\sum_{k=1}^{K}\left(\sum_{i=1}^{n_k}\left(\hat{y}_i^{[k]}(\lambda)-y_i\right)^2\right) \tag{14.1.654}$$

And the "best lambda" λ_0 is the one that minimize the OCV_{MSE} such that:

$$\lambda_0=\min_{\lambda\in\mathbb{R}^+}\text{OCV}_{\text{MSE}}(\lambda) \tag{14.1.655}$$

Let us now speak about a very common relation in statistical textbooks, the cross-validation for one of the subsets of the Leave One Out method (i.e. Jackknife):

$$\text{CV}=\sum_i\left(y_i^{[k]}-y_i\right)^2=\sum_i\left(\frac{\hat{y}_i-y_i}{1-H_{ii}}\right)^2 \tag{14.1.656}$$

Here, y_i is the actual label value of training point i, $y_i^{[k]}$ is the value predicted by the cross-validation model trained on all points except i, $\hat{y}^i$ is the value predicted by the regression model trained on all points (including point i), and H_{ii} is as we know the leverage of point i, given for recall by:

$$H_{ii}=\frac{1}{n}+\frac{(x_i-\bar{x})^2}{\displaystyle\sum_{k=1}^{n}(x_k-\bar{x})^2} \tag{14.1.657}$$

and:

$$H=X(X^TX)^{-1}X^T \tag{14.1.658}$$

Notice that the left side of the above relation is the LOOCV sum of squares error (the quantity we seek), while the right can be evaluated given only the model trained on the full data set. Fantastically, this allows us to evaluate the LOOCV error using only a single regression!!!

Proof 14.3.3. Consider the LOOCV step where we construct a model trained on all points except training example k. Using a linear model of form $y^{[k]}=\vec{x}^T\vec{\beta}^{[k]}$ - with $\vec{\beta}^{[k]}$ the coefficient vector - the sum of squares that must be minimized is:

$$\sum_{i\neq k}\left(y_i^{[k]}-y_i\right)^2=\sum_{i\neq k}\left(\vec{x}_i^T\vec{\beta}^{[k]}-y_i\right)^2 \tag{14.1.659}$$

Here, we are using an exponent k on $\vec{\beta}^{[k]}$ to highlight the fact that the above corresponds to the case where example k is held out. We minimize the above relation by taking the gradient with respect to $\vec{\beta}^{[k]}$. Setting this to zero gives the following equation:

$$2\sum_{i\neq k}\vec{x}_i\left(\vec{x}_i^T\vec{\beta}^{[k]}-y_i\right)=\vec{0} \tag{14.1.660}$$

Rearranging and simplifying a bit gives:

$$\left(\sum_{i\neq k}\vec{x}_i\vec{x}_i^T\right)\vec{\beta}^{[k]}=\sum_{i\neq k}y_i\vec{x}_i \tag{14.1.661}$$

Similarly, the full model (trained on all points) coefficient vector $\vec{\beta}$ (here the notation $\vec{\beta}$ means in fact $\vec{\hat{\beta}}$ obviously!) satisfies:

$$\left(\sum_{i}\vec{x}_i\vec{x}_i^T\right)\vec{\beta}=\sum_{i}y_i\vec{x}_i \tag{14.1.662}$$

But we have that:

$$\left(\sum_{i\neq k}\vec{x}_i\vec{x}_i^T\right)\vec{\beta}^{[k]}=\sum_{i\neq k}y_i\vec{x}_i \tag{14.1.663}$$

can we rewritten as for the left term:

$$\left(\sum_{i\neq k}\vec{x}_i\vec{x}_i^T\right)\vec{\beta}^{[k]}=\left(\sum_{i}\vec{x}_i\vec{x}_i^T-\vec{x}_k\vec{x}_k^T\right)\vec{\beta}^{[k]} \tag{14.1.664}$$

and for the right term:

$$\sum_{i\neq k}y_i\vec{x}_i=\sum_{i}y_i\vec{x}_i-y_k\vec{x}_k \tag{14.1.665}$$

Therefore:

$$\left(\sum_{i}\vec{x}_i\vec{x}_i^T-\vec{x}_k\vec{x}_k^T\right)\vec{\beta}^{[k]}=\sum_{i}y_i\vec{x}_i-y_k\vec{x}_k \tag{14.1.666}$$

But as we have proved just earlier that:

$$\sum_{i}y_i\vec{x}_i=\left(\sum_{i}\vec{x}_i\vec{x}_i^T\right)\vec{\beta} \tag{14.1.667}$$

Then substituting in the previous relation, this lead us to:

$$\left(\sum_i \vec{x}_i\vec{x}_i^T-\vec{x}_k\vec{x}_k^T\right)\vec{\beta}^{[k]}=\left(\sum_i \vec{x}_i\vec{x}_i^T\right)\vec{\beta}-y_k\vec{x}_k \tag{14.1.668}$$

Rearranging:

$$\sum_i \vec{x}_i\vec{x}_i^T\left(\vec{\beta}^{[k]}-\vec{\beta}\right)=\vec{x}_k\vec{x}_k^T\vec{\beta}^{[k]}-y_k\vec{x}_k \tag{14.1.669}$$

That can be written using the definition of $y_i^{[k]}$:

$$\sum_i \vec{x}_i\vec{x}_i^T\left(\vec{\beta}^{[k]}-\vec{\beta}\right)=\vec{x}_k y_i^{[k]}-y_k\vec{x}_k \tag{14.1.670}$$

Hence:

$$\sum_i \vec{x}_i\vec{x}_i^T\left(\vec{\beta}^{[k]}-\vec{\beta}\right)=\left(y_i^{[k]}-y_k\right)\vec{x}_k \tag{14.1.671}$$

Left multiplication by $\vec{x}_k^T\left(\sum_i \vec{x}_i\vec{x}_i^T\right)^{-1}$ (without forgetting that the term into parenthesis is a scalar) gives:

$$\cancel{\sum_i \vec{x}_i\vec{x}_i^T}\left(\vec{\beta}^{[k]}-\vec{\beta}\right)\vec{x}_k^T\cancel{\left(\sum_i \vec{x}_i\vec{x}_i^T\right)^{-1}}=\left(y_i^{[k]}-y_k\right)\vec{x}_k\vec{x}_k^T\left(\sum_i \vec{x}_i\vec{x}_i^T\right)^{-1} \tag{14.1.672}$$

So it only remains (rearranging a bit):

$$\left(\vec{\beta}^{[k]}-\vec{\beta}\right)\vec{x}_k^T=\left(y_i^{[k]}-y_k\right)\vec{x}_k\left(\sum_i \vec{x}_i\vec{x}_i^T\right)^{-1}\vec{x}_k^T \tag{14.1.673}$$

We recognize here the vector definition of H_{kk}. Therefore:

$$\left(\vec{\beta}^{[k]}-\vec{\beta}\right)\vec{x}_k^T=\left(y_i^{[k]}-y_k\right)H_{kk} \tag{14.1.674}$$

that is:

$$y_k^{[k]}-\hat{y}_k=\left(y_i^{[k]}-y_k\right)H_{kk} \tag{14.1.675}$$

That latter can be rewritten as:

$$(y_k^{[k]}-y_k)-(\hat{y}_k-y_k)=\left(y_i^{[k]}-y_k\right)H_{kk} \tag{14.1.676}$$

Hence:

$$y_k^{[k]}-y_k=\frac{\hat{y}_k-y_k}{1-H_{kk}} \tag{14.1.677}$$

Squaring and summing finally leads to:

$$\boxed{\text{CV}=\sum_k\left(y_k^{[k]}-y_k\right)^2=\sum_k\left(\frac{\hat{y}_k-y_k}{1-H_{kk}}\right)^2} \tag{14.1.678}$$

■ Q.E.D.

Again... Fantastically, this allows us to evaluate the LOOCV error using only a single regression!!!

A minor variant on cross validation is, so-named "generalized cross validation", which, of course, like most things statisticians call "generalized" isn't... It replaces the H_{kk} in the denominator with their average trace $\text{tr}(H_{kk})/n$ giving:

$$\boxed{\text{GCV} = \sum_k \left(\frac{\hat{y}_k - y_k}{1 - \text{tr}(H_{kk})/n} \right)^2} \tag{14.1.679}$$

14.1.12 Ridge, LASSO and Elastic Net regularization

"Regularization" also named "regularized regression" or "penalized regression" has been intensely studied on the interface between statistics and computer science as it is an empirical method can be used to avoid overfitting in almost all regression models (but is mainly known for its application to logistic regression).

There are two types of regularization as follows:

- L_1 Regularization or LASSO Regularization (see details further below)
- L_2 Regularization or Ridge Regularization (see details further below)

Where for recall $L_p:\mathbb{R}^n \mapsto \mathbb{R}$ (see section Topology page 2450) is defined:

$$L_p(x)=||x||_p=\left(\sum_{i=1}^{n}|x_i|^p\right)^{1/p} \tag{14.1.680}$$

with $p>0$. For $p=2$, this is the familiar Euclidean distance (don't confuse with the notation of the lost function L if possible...).

Ridge and Lasso regularizations are also known as "shrinkage methods", because they reduce or shrink the coefficients in the resulting regression. This reduces the variance in the model: as input variables are changed, the models prediction changes less than it would have without the regularization. Why would you want to reduce the variance of a model? To avoid overfit!

L_1 and L_2 are the most common types of regularization. These update the general cost function by adding another term known as the "regularization term":

$$\text{Cost function} = \text{Loss} + \text{Regularization term} \tag{14.1.681}$$

Or more formally if we denote by $R(f)$ the regularization term :

$$\boxed{\min_f \sum_{i=1}^{n} V(f(x_i),y_i)+\lambda R(f)} \tag{14.1.682}$$

where V is an underlying loss function (most of times for regression models it's the residual sum of squares RSS) that describes the cost of predicting $f(x)$ when the label is y and λ is a parameter which controls the importance of the regularization term. $R(f)$ is typically chosen to impose a penalty on the complexity of f. Concrete notions of complexity used include restrictions for smoothness and bounds on the vector space norm.

Notice that softwares compute estimates for a large number of values for λ at once. The optimal λ is selected by cross validation of some sort... The most common one in the end of the 121st century (holocene calendar) being the generalized cross-validation metric (see further below page 6434).

From a Bayesian point of view, many regularization techniques correspond to imposing certain prior distributions on model parameters.

Proof 14.3.4. Consider that the regularized loss function L has a similar role as the probability of a parameter configuration β given the observations X,$\vec{y}$. Applying the Bayes theorem (see section Probabilities page 492), we get:

$$P(\theta|X,\vec{y}) = \frac{P(X,\vec{y}|\theta)P(\theta)}{P(X,\vec{y})} \tag{14.1.683}$$

Taking the log of the expression gives us:

$$\log(P(\theta|X,\vec{y})) = \log(P(X,\vec{y}|\theta)) + \log(P(\theta)) - \log(P(X,\vec{y})) \tag{14.1.684}$$

Now, let's say $L(\theta)$ is the negative[22] log-posterior, $-\log(P(\theta|X,\vec{y}))$. Since the last term does not depend on θ, we can ignore it without changing the minimum. We are left with two terms:

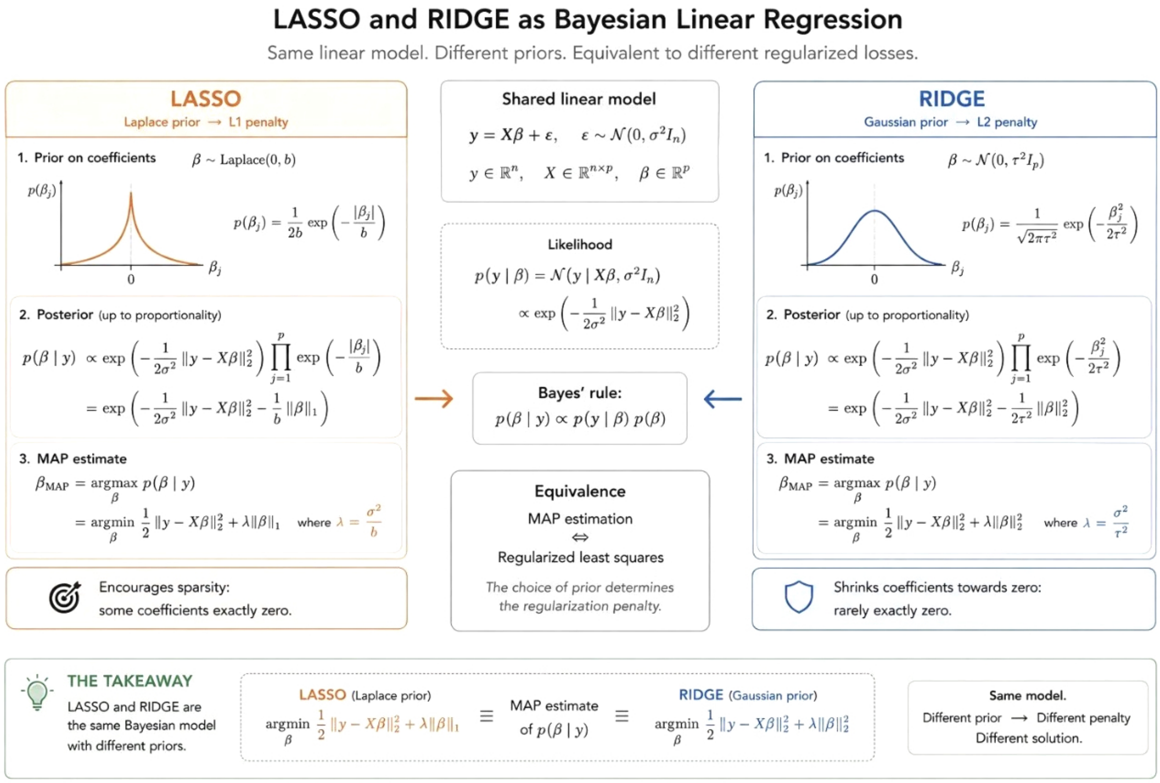
- the likelihood term $\log(P(X,\vec{y}|\theta))$ depending on X and $\vec{y}$, and
- the prior term $\log(P(\theta))$ depending on θ only. These two terms correspond exactly to the loss and the regularization term.

■ Q.E.D.

Remark

It can be proved (see further below) that L_2 regularization is the case with a Gaussian prior analytically and the L_1 is equivalent to a Laplacian prior.

We can also see regularization as an application of the Lagrange Multiplier method (see previously page 6667).



LASSO and RIDGE as Bayesian Linear Regression

Same linear model. Different priors. Equivalent to different regularized losses.

LASSO

Laplace prior → L1 penalty

1. Prior on coefficients $\beta \sim \text{Laplace}(0,b)$

$$p(\beta_j) = \frac{1}{2b}\exp\left(-\frac{|\beta_j|}{b}\right)$$

2. Posterior (up to proportionality)

$$\begin{aligned} p(\beta \mid y) &\propto \exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|_2^2\right)\prod_{j=1}^{p}\exp\left(-\frac{|\beta_j|}{b}\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|_2^2 - \frac{1}{b}\|\beta\|_1\right) \end{aligned}$$

3. MAP estimate

$$\begin{aligned} \beta_{\text{MAP}} &= \underset{\beta}{\text{argmax}}\; p(\beta \mid y) \\ &= \underset{\beta}{\text{argmin}}\; \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \quad \text{where } \lambda = \frac{\sigma^2}{b} \end{aligned}$$

Encourages sparsity:
some coefficients exactly zero.

Shared linear model

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,\sigma^2 I_n)$$

$$y \in \mathbb{R}^n, \quad X \in \mathbb{R}^{n\times p}, \quad \beta \in \mathbb{R}^p$$

Likelihood

$$\begin{aligned} p(y \mid \beta) &= \mathcal{N}(y \mid X\beta, \sigma^2 I_n) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|_2^2\right) \end{aligned}$$

Bayes' rule:

$$p(\beta \mid y) \propto p(y \mid \beta)\, p(\beta)$$

Equivalence

MAP estimation
⇔
Regularized least squares

The choice of prior determines the regularization penalty.

RIDGE

Gaussian prior → L2 penalty

1. Prior on coefficients $\beta \sim \mathcal{N}(0,\tau^2 I_p)$

$$p(\beta_j) = \frac{1}{\sqrt{2\pi\tau^2}}\exp\left(-\frac{\beta_j^2}{2\tau^2}\right)$$

2. Posterior (up to proportionality)

$$\begin{aligned} p(\beta \mid y) &\propto \exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|_2^2\right)\prod_{j=1}^{p}\exp\left(-\frac{\beta_j^2}{2\tau^2}\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|_2^2 - \frac{1}{2\tau^2}\|\beta\|_2^2\right) \end{aligned}$$

3. MAP estimate

$$\begin{aligned} \beta_{\text{MAP}} &= \underset{\beta}{\text{argmax}}\; p(\beta \mid y) \\ &= \underset{\beta}{\text{argmin}}\; \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 \quad \text{where } \lambda = \frac{\sigma^2}{\tau^2} \end{aligned}$$

Shrinks coefficients towards zero:
rarely exactly zero.

THE TAKEAWAY

LASSO and RIDGE are the same Bayesian model with different priors.

LASSO (Laplace prior) $\underset{\beta}{\text{argmin}}\; \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$ ≡ MAP estimate of $p(\beta \mid y)$ ≡ RIDGE (Gaussian prior) $\underset{\beta}{\text{argmin}}\; \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$

Same model.
Different prior ⟶ Different penalty
Different solution.

---

[22] Negative since we want to maximize the probability but minimize the cost.

14.1.12.1 L_2 Regularization or Ridge Regularization

When learning a linear function f, characterized by an unknown vector $\vec{\beta}$ such that $f(x)=\vec{\beta}\circ\vec{x}$, the L_2-norm loss corresponds to "Tikhonov regularization". This is one of the most common forms of regularization, is also known as "ridge regression", and is expressed as:

$$\min_{\beta}\sum_{i=1}^{n} V(\vec{x}_i\circ\vec{w},\vec{y}_i)+\lambda L_2(\vec{\beta})=\min_{\beta}\sum_{i=1}^{n} V(\vec{x}_i\circ\vec{w},\vec{y}_i)+\lambda\|\vec{\beta}\|^2 \qquad (14.1.685)$$

where λ is the "tuning parameter" or "shrinkage parameter".

Remark

If the different predictor variables don't have physically comparable units it's a good idea to standardize them first, so they all have mean 0 and variance 1 and also no physical units. Otherwise penalizing the p predictor with the L_2-norm using $\sum_{i=1}^{p}\beta_i^2$ seems to be adding up apples, oranges, and the occasional bout of regret (some people always pre-standardize the predictors).

The learning problem with the least squares loss function and Tikhonov regularization can be solved analytically. Written in matrix form, the optimal $\vec{\beta}$ will be the one for which the gradient of the loss function with respect to $\vec{\beta}$ is $\vec{0}$.

Let us write this explicitly:

$$\boxed{\text{PRSS}(\vec{\beta})_{L_2}=\min_{\vec{\beta}}\frac{1}{n}(X\vec{\beta}-\vec{y})^T(X\vec{\beta}-\vec{y})+\lambda\|\vec{\beta}\|^2} \qquad (14.1.686)$$

which is just the usual least squares criterion with a penalty determined by λ for large coefficient estimates. This is why it's named "penalized residual sum of squares" (PRSS). If $\lambda=0$ the lasso is the same as OLS; as λ increases, shorter coefficients are preferred. Therefore we can say that most of times (but now always!) if we increase the value of λ then the magnitude of the coefficients decreases.

Notice also that this is a well defined, convex, differentiable optimization problem even after including L_2 function. The L_2 function acts as a penalizer: $\|\vec{\beta}\|^2$ grows large as the coefficients $\vec{\beta}$ grow large, thus the minimization program tries to make w smaller.

There are many variations on this procedure, including application of it to other-than the linear model.

We take the derivative:

$$\vec{\nabla}_{\vec{\beta}}=\frac{2}{n}X^T(X\vec{\beta}-\vec{y})+2\lambda\vec{\beta} \qquad (14.1.687)$$

To find the optimum we put:

$$\vec{0}=X^T(X\vec{\beta}-\vec{y})+n\lambda\vec{\beta} \qquad (14.1.688)$$

Therefore:

$$\hat{\vec{\beta}}_{\lambda}^{\text{ridge}}=(X^TX+\lambda n\mathbb{1})^{-1}(X^T\vec{y}) \qquad (14.1.689)$$

By construction of the optimization problem, other values of $\vec{\beta}$ would give larger values for the loss function. Notice that in particular the regularization term solves the problem of non-invertibility of the first term between parenthesis.

Notice that the solution is indexed by the parameter λ. So for each λ, we have a solution. The fact that we add $\lambda n\mathbb{1}$ to the diagonal of X^TX (which correspondence to the correlation matrix if X is previously centered and reduced) is why this method is named "ridged" (and the term $\lambda n\mathbb{1}$ is often named the "ridged").

The reader may also notice that with:

$$\hat{\vec{\beta}}_{\lambda}^{\text{ridge}} = \frac{X^T\vec{y}}{X^TX+\lambda n\mathbb{1}} \tag{14.1.690}$$

excepted is some very special cases, the ridge regression will never take the coefficients $\hat{\vec{\beta}}_{\lambda}^{\text{ridge}}$ to zero as the denominator will never be practically infinite! But we can see obviously that as $\lambda \to +\infty$ then $\hat{\vec{\beta}}_{\lambda}^{\text{ridge}} \to \vec{0}$.

Remark

Inclusion of λ makes problem non-singular even if X^TX is not invertible. It seems that this was the original motivation for ridge regression (Hoerl and Kennard, 11970 according to holocene calendar).

Let us prove now that the ridge regression corresponds indeed to a Gaussian prior on the coefficients.

Proof 14.3.5. Let us imagine that you want to infer some parameter β from some observed input-output pairs $(x_1,y_1),\ldots,(x_n,y_n)$. Let us assume that the outputs are linearly related to the inputs via β and that the data are corrupted by some noise ε:

$$y_n = \beta x_n + \varepsilon \tag{14.1.691}$$

where ε is Gaussian noise with as we know mean 0 and variance σ_{ε}^2. This gives rise to a Gaussian likelihood:

$$\prod_{n=1}^{N} \mathcal{N}(y_n|\beta x_n,\sigma_{\varepsilon}^2) \tag{14.1.692}$$

Let us regularise parameter β by imposing the Gaussian prior $\mathcal{N}(\beta|0,\lambda^{-1})$, where λ is a strictly positive scalar. Hence, combining the likelihood and the prior we simply have:

$$\prod_{n=1}^{N} \mathcal{N}(y_n|\beta x_n,\sigma_{\varepsilon}^2)\mathcal{N}(\beta|0,\lambda^{-1}) \tag{14.1.693}$$

Let us take the logarithm of the above expression. Dropping some constants we get:

$$\sum_{n=1}^{N} -\frac{1}{\sigma_{\varepsilon}^2}(y_n-\beta x_n)^2 - \lambda\beta^2 \tag{14.1.694}$$

If we maximise the above expression with respect to β, we get the so named "maximum a-posteriori estimate" for β, or "MAPE for ridge regression" for short. In this expression it becomes apparent why the Gaussian prior can be interpreted as a L_2 regularisation term.

■ Q.E.D.

Let us now prove that the $\hat{\vec{\beta}}_\lambda^{\text{ridge}}$ are biased.

Proof 14.3.6. First let us recall that we have proved that:

$$\hat{\vec{\beta}}_\lambda^{\text{ridge}} = (\hat{X}^T X + \lambda n \mathbb{1})^{-1}(X^T \vec{y}) \tag{14.1.695}$$

Now let us rewrite this:

$$\begin{aligned}
\hat{\vec{\beta}}_\lambda^{\text{ridge}} &= (X^T \hat{X} + \lambda n \mathbb{1})^{-1}(X^T \vec{y}) = (X^T X + \lambda n \mathbb{1})^{-1} \left((X^T X)(X^T X)^{-1}\right)(X^T \vec{y}) \\
&= (X^T X + \lambda n \mathbb{1})^{-1} X^T X \left((X^T X)^{-1} X^T \vec{y}\right) \\
&= \left[(X^T X)(\mathbb{1} + \lambda n (X^T X)^{-1})\right]^{-1} X^T X \left((X^T X)^{-1} X^T \vec{y}\right)
\end{aligned} \tag{14.1.696}$$

Remember that we have proved in the section of Linear Algebra that if A and B are invertible, then:

$$(AB)^{-1} = B^{-1} A^{-1} \tag{14.1.697}$$

Then:

$$\begin{aligned}
\hat{\vec{\beta}}_\lambda^{\text{ridge}} &= \left[(X^T X)(\mathbb{1} + \lambda n (X^T X)^{-1})\right]^{-1} X^T X \left((X^T X)^{-1} X^T \vec{y}\right) \\
&= (\mathbb{1} + \lambda n (X^T X)^{-1})^{-1} (X^T X)^{-1} X^T X \left((X^T X)^{-1} X^T \vec{y}\right) \\
&= (\mathbb{1} + \lambda n (X^T X)^{-1})^{-1} \underbrace{(X^T X)^{-1} X^T X}_{=\mathbb{1}} \left((X^T X)^{-1} X^T \vec{y}\right) \\
&= (\mathbb{1} + \lambda n (X^T X)^{-1})^{-1} \mathbb{1} \left((X^T X)^{-1} X^T \vec{y}\right) \\
&= (\mathbb{1} + \lambda n (X^T X)^{-1})^{-1} \left((X^T X)^{-1} X^T \vec{y}\right) \\
&= (\mathbb{1} + \lambda n (X^T X)^{-1})^{-1} \hat{\beta}^{\text{OLS}}
\end{aligned} \tag{14.1.698}$$

So:

$$\begin{aligned}
\text{E}(\hat{\vec{\beta}}_\lambda^{\text{ridge}}) &= \text{E}((\mathbb{1} + \lambda n (X^T X)^{-1})^{-1} \hat{\beta}^{\text{OLS}}) \\
&= (\mathbb{1} + \lambda n (X^T X)^{-1})^{-1} \text{E}(\hat{\beta}^{\text{OLS}}) \\
&= (\mathbb{1} + \lambda n (X^T X)^{-1})^{-1} \beta^{\text{OLS}}
\end{aligned} \tag{14.1.699}$$

Therefore if $\lambda \neq 0$ then:

$$\text{E}(\hat{\vec{\beta}}_\lambda^{\text{ridge}}) \neq \beta^{\text{OLS}} \tag{14.1.700}$$

■ Q.E.D.

We know that if we actually have the "right" model, then, say, OLS parameter estimates will be unbiased and have minimal variance among all unbiased (linear) estimators (they are BLUE). Predictions from an OLS model will be best linear unbiased predictions (BLUPs). That sounds good.

So we have $\hat{\vec{\beta}}_\lambda^{\text{ridge}} = (X^T X + \lambda \mathbb{1})^{-1} X^T Y$ which is biased; but if X is ill conditioned then $\text{V}(\hat{\beta}) \propto (X^T X)^{-1}$ may be monstrous whereas $\text{V}(\hat{\vec{\beta}}_\lambda^{\text{ridge}})$ can be much more modest. Indeed, let us recall that we have proved during our study of multiple linear regression that (see above page 6401):

$$\text{V}\left(\hat{\vec{\beta}}\right) = \sigma_\varepsilon^2 (X_c^T X_c)^{-1} \tag{14.1.701}$$

Exactly the same developments would lead us to:

$$\text{V}(\hat{\vec{\beta}}_\lambda^{\text{ridge}}) = \sigma_\varepsilon^2 (X_c^T X_c + \lambda n \mathbb{1})^{-1} \tag{14.1.702}$$

So increasing λ decrease indeed the variances of the $\hat{\vec{\beta}}_\lambda^{\text{ridge}}$.

Remark

If we assume the coefficients of the regression to be Gaussian distributed with mean 0 and variance σ_ε^2 and we take $\lambda = 1/\sigma_\varepsilon^2$, we speak sometimes then of "Gauss reguralization".

14.1.12.2 L_1 Regularization or LASSO Regularization

We describe the basic idea through the "LASSO regularization" (LASSO stands for: Least Absolute Shrinkage and Selection Operator) as applied in the context of linear regression. The method starts by assuming a model like:

$$\boxed{\text{PRSS}(\vec{\beta})_{L_1}=\min_{\vec{\beta}}\frac{1}{n}(\hat{X}\vec{\beta}-\vec{y})^T(\hat{X}\vec{\beta}-\vec{y})+\lambda L_1(\vec{\beta})=\min_{\vec{\beta}}\frac{1}{n}(\hat{X}\vec{\beta}-\vec{y})^T(\hat{X}\vec{\beta}-\vec{y})+\lambda\sum_i|\beta_i|} \tag{14.1.703}$$

which is just the usual least squares criterion with a penalty determined by λ for large coefficient estimates. Again, if $\lambda=0$ the lasso is the same as OLS; as λ increases, shorter vectors are preferred.

Remark

If the different predictor variables don't have physically comparable units it's a good idea to standardize them first, so they all have mean 0 and variance 1 and also no physical units. Otherwise penalizing the p predictor with the L_1-norm using $\sum_{i=1}^p|\beta_i|$ seems to be adding up apples, oranges, and the occasional bout of regret (some people always pre-standardize the predictors).

Immediately we lose the differentiability of the optimization function: the absolute values are non-differentiable.

Thus we cannot solve the optimization problem using standard calculus techniques in the most general case. What do we gain then? First we can consider only the case $\vec{\beta}\neq\vec{0}$.

We take the derivative:

$$\vec{\nabla}_{\vec{\beta}}=\frac{2}{n}X^T(X\vec{\beta}-\vec{y})+\lambda \tag{14.1.704}$$

To find the optimum we put:

$$\vec{0}=X^T(X\vec{\beta}-\vec{y})+\frac{n\lambda}{2} \tag{14.1.705}$$

Therefore:

$$\hat{\vec{\beta}}_\lambda^{\text{LASSO}}=\frac{X^T\vec{y}-\dfrac{n\lambda}{2}}{X^TX}=\frac{2X^T\vec{y}-n\lambda}{2X^TX} \tag{14.1.706}$$

If you observe the numerator, it may be equal to zero, since we are subtracting some value of λ. And therefore the values of $\hat{\vec{\beta}}_\lambda^{\text{LASSO}}$ may be all equal to zero at the opposite of the ridge regression!

One can show that for some large enough λ, the solution is a sparse solution. A sparse solution means a solution for $\vec{\beta}$ that has many zeros in it, effectively removing the corresponding variable from the system. There are then two immediate consequences:

1. The system will give a coefficient of zero to non-relevant variables, instead of some insanely small coefficient. This means independent or near-independent variables will be discarded instead of included.

2. It gives us the most important variables in the regression model! We can speak confidently that variable i has a strong connection with the response variable $\vec{y}$.

Again notice that softwares compute estimates for a large number of values for λ at once. The optimal λ is selected by cross validation of some sort.

Let us prove now that the LASSO regression corresponds to a Laplace prior on the coefficients.

Proof 14.3.7. Let us imagine that you want to infer some parameter β from some observed input-output pairs $(x_1,y_1),\ldots,(x_n,y_n)$. Let us assume that the outputs are linearly related to the inputs via β and that the data are corrupted by some noise ε:

$$y_n=\beta x_n+\varepsilon \tag{14.1.707}$$

where ε is Gaussian noise with as we know mean 0 and variance σ_ε^2. This gives rise to a Gaussian likelihood:

$$\prod_{n=1}^N \mathcal{N}(y_n|\beta x_n,\sigma_\varepsilon^2) \tag{14.1.708}$$

First, let us recall that a Laplace distribution is given by (difference of two exponential distribution as we have proved it in the section of Statistics at page 800):

$$\mathcal{L}(x\,|\,\mu,b)=\frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)=\frac{1}{2b}\begin{cases}\exp\left(-\frac{\mu-x}{b}\right) & \text{if } x<\mu\\ \exp\left(-\frac{x-\mu}{b}\right) & \text{if } x\geq\mu\end{cases} \tag{14.1.709}$$

Remark

Notice that the Normal and Laplace prior have a strong similitude. This is why they are commonly denoted:

$$P_{\lambda,p}=C^{te}(\lambda,p)e^{-\lambda L_p(\vec{\beta})} \tag{14.1.710}$$

Ridge regression $(p=2)$ corresponds to a Gaussian prior and the LASSO $(p=1)$ to a Laplacian (or double-exponential) prior.

Let us regularise parameter β by imposing the Laplace prior $\mathcal{L}(\beta|0,b)$, where b is a strictly positive scalar (and mean $\mu=0$). Hence, combining the likelihood and the prior we simply have:

$$\prod_{n=1}^N \mathcal{N}(y_n|\beta x_n,\sigma_\varepsilon^2)\mathcal{L}(\beta|0,b)=\prod_{n=1}^N \mathcal{N}(y_n|\beta x_n,\sigma_\varepsilon^2)\frac{1}{2b}\exp\left(-\frac{|\beta|}{b}\right) \tag{14.1.711}$$

Let us take the logarithm of the above expression. Dropping some constants we get:

$$\sum_{n=1}^N -\frac{1}{\sigma_\varepsilon^2}(y_n-\beta x_n)^2-\frac{1}{b}|\beta| \tag{14.1.712}$$

Let $1/b=\lambda$, we get then:

$$\sum_{n=1}^{N}-\frac{1}{\sigma_\varepsilon^2}(y_n-\beta x_n)^2-\lambda|\beta| \tag{14.1.713}$$

If we maximise the above expression with respect to β, we get the so named "maximum a-posteriori estimate" for β, or "MAPE for LASSO regression" for short. In this expression it becomes apparent why the Laplace prior can be interpreted as a L_1 regularisation term.

■ Q.E.D.

Remark

If we assume the coefficients of the regression to be Laplace distributed with variance σ_ε^2 and we take $\lambda=\sqrt{2}/\sigma_\varepsilon$, we speak sometimes then of "Laplace reguralization".

14.1.12.3 L_1+L_2 Regularization or Elastic Net Regularization

In statistics and, in particular, in the fitting of linear or binomial logistic regression models, the elastic net is a regularized regression method that linearly combines the L_1 and L_2 penalties of the lasso and ridge methods.

The elastic net method overcomes the limitations of the LASSO (least absolute shrinkage and selection operator) method which uses a penalty function based for recall on:

$$\lambda\sum_{i=1}^{n}|\beta_i| \tag{14.1.714}$$

Use of this penalty function seems to have several limitations. For example, in the "large n, small N" case (high-dimensional data with few examples), the LASSO seems to select at most n variables before it saturates. Also if there is a group of highly correlated variables, then the LASSO tends to select one variable from a group and ignore the others (i.e. LASSO tends to select only one predictor among the predictors that are highly correlated). To overcome these limitations, the elastic net adds a quadratic part to the penalty $||\beta||^2$, which when used alone is ridge regression. The estimates from the elastic net method are defined by:

$$\hat{\beta}^{\text{EN}}_{\lambda_1,\lambda_2}=\min_{\vec{\beta}}\ ||\vec{y}-X\vec{\beta}||^2+\lambda_2||\vec{\beta}||^2+\lambda_1\sum_i|\beta_i| \tag{14.1.715}$$

The quadratic penalty term makes the loss function strictly convex, and it therefore has a unique minimum. The elastic net method includes the LASSO and ridge regression: in other words, each of them is obviously a special case where $\lambda_1=\lambda$, $\lambda_2=0$ or $\lambda_1=0$, λ_2.

Another very common notation is:

$$\boxed{\hat{\beta}^{\text{EN}}_{\lambda,\alpha}=\min_{\vec{\beta}}\ \frac{1}{n}||\vec{y}-X\vec{\beta}||^2+\lambda\left((1-\alpha)||\vec{\beta}||^2+\alpha\sum_i|\beta_i|\right)} \tag{14.1.716}$$

where $0\leq\alpha\leq 1$ is a compromise between ridge ($\alpha=0$) and lasso ($\alpha=1$).

Remark

Depending on the softwares, the algorithms implement some variations of the relation above. The most common one are:

$$\hat{\beta}^{\text{EN}}_{\lambda,\alpha}=\min_{\vec{\beta}}\ \|\vec{y}-X\vec{\beta}\|^2+\lambda\left((1-\alpha)\|\vec{\beta}\|^2+\alpha\sum_i|\beta_i|\right) \tag{14.1.717}$$

or:

$$\hat{\beta}^{\text{EN}}_{\lambda,\alpha}=\min_{\vec{\beta}}\ \frac{1}{2n}\|\vec{y}-X\vec{\beta}\|^2+\lambda\left(\frac{(1-\alpha)}{2}\|\vec{\beta}\|^2+\alpha\sum_i|\beta_i|\right) \tag{14.1.718}$$

and so on.... Hence the sometimes the huge differences in the optimal λ value.

A common question about regularization is that in traditional statistics models, while building a model, we check for multicollinearity using methods such as estimates of the variance inflation factor (VIF), but in Machine Learning, we instead use regularization for feature selection and don't seem to check whether features are correlated at all. Why do we do that?

The reason is because the goals of "traditional statistics" are different from many Machine Learning techniques.

In traditional statistics regressions, we are trying to understand the impact the independent variables have on the dependent variable. If there is strong multicollinearity, this is simply not possible. No algorithm is going to fix this. If studiousness is correlated with class attendance and grades, we cannot know what is truly causing the grades to go up - attendance or studiousness.

However, in Machine Learning techniques that focus on predictive accuracy, all we care about is how we can use a set of variables to predict another set. We don't care about the impact these variables have on each other.

Basically, the fact that we don't check for multicollinearity in Machine Learning techniques isn't a consequence of the algorithm, it's a consequence of the goal!

Remark

For more information we strongly recommend the reader to refer to the excellent lecture notes of Wessel N. van Wieringen [572].

14.1.12.4 Polynomial regression

We will now see how to determine the for example of the best second degree polynomial that passes by any number of points, but without transforming the function contrary to what we have just done just before! As we like physics in this textbook, we will take a classic case of Kinematic to join business with pleasure...

So consider that we are looking for a polynomial of the second degree of the form:

$$x=x_0+v_0t+\frac{1}{2}at^2 \tag{14.1.719}$$

knowing that the method is easily applicable to higher order polynomials (at least as far as we know).

Relation that it is customary to write in the field of polynomial regression as follows:

$$d_i=c_1+c_2t_i+c_3t_i^2 \tag{14.1.720}$$

where i represents the number of points available to us.

For the rest, let us once again base our developments on the least squares method. In other words, we seek the coefficients c_1,c_2,c_3 that minimize the error:

$$\text{SSR}=\sum_{i=1}^n(d_i-d_i')^2=\sum_{i=1}^n(d_i'-d_i)^2=\sum_{i=1}^n\left(\left(c_1+c_2t_i+c_3t_i^2\right)-d_i\right)^2=\Phi(c_1,c_2,c_3) \tag{14.1.721}$$

and we attack again with partial derivatives for each coefficient:

$$\frac{\partial\Phi}{\partial c_1}=\sum_{i=1}^n 2\left(c_1+c_2t_i+c_3t_i^2-d_i\right)=2\left(nc_1+c_2\sum_{i=1}^n t_i+c_3\sum_{i=1}^n t_i^2-\sum_{i=1}^n d_i\right)=0 \tag{14.1.722}$$

Therefore after a small rearrangement and simplification:

$$nc_1+c_2\sum_{i=1}^n t_i+c_3\sum_{i=1}^n t_i^2=\sum_{i=1}^n d_i \tag{14.1.723}$$

Similarly:

$$\frac{\partial\Phi}{\partial c_2}=\sum_{i=1}^n 2t_i\left(c_1+c_2t_i+c_3t_i^2-d_i\right)=2\left(c_1\sum_{i=1}^n t_i+c_2\sum_{i=1}^n t_i^2+c_3\sum_{i=1}^n t_i^3-\sum_{i=1}^n d_it_i\right)=0 \tag{14.1.724}$$

Therefore after a small rearrangement and simplification:

$$c_1\sum_{i=1}^n t_i+c_2\sum_{i=1}^n t_i^2+c_3\sum_{i=1}^n t_i^3=\sum_{i=1}^n d_it_i \tag{14.1.725}$$

And finally:

$$\frac{\partial\Phi}{\partial c_3}=\sum_{i=1}^n 2t_i^2\left(c_1+c_2t_i+c_3t_i^2-d_i\right)=2\left(c_1\sum_{i=1}^n t_i^2+c_2\sum_{i=1}^n t_i^3+c_3\sum_{i=1}^n t_i^4-\sum_{i=1}^n d_it_i^2\right)=0 \tag{14.1.726}$$

Therefore after a small rearrangement and simplification:

$$c_1\sum_{i=1}^{n}t_i^2+c_2\sum_{i=1}^{n}t_i^3+c_3\sum_{i=1}^{n}t_i^4=\sum_{i=1}^{n}d_it_i^2 \tag{14.1.727}$$

So using the notation of linear algebra, we finally have to solve the following system:

$$\begin{bmatrix} n & \sum\limits_{i=1}^{n}t_i & \sum\limits_{i=1}^{n}t_i^2 \\ \sum\limits_{i=1}^{n}t_i & \sum\limits_{i=1}^{n}t_i^2 & \sum\limits_{i=1}^{n}t_i^3 \\ \sum\limits_{i=1}^{n}t_i^2 & \sum\limits_{i=1}^{n}t_i^3 & \sum\limits_{i=1}^{n}t_i^4 \end{bmatrix}\begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix}=\begin{pmatrix} \sum\limits_{i=1}^{n}d_i \\ \sum\limits_{i=1}^{n}d_it_i \\ \sum\limits_{i=1}^{n}d_it_i^2 \end{pmatrix} \tag{14.1.728}$$

and so we just have to solve this simple linear system either by hand using the relations proved in the section of Linear Algebra, or with a simple spreadsheet software (like Microsoft® Excel® for example).

14.1.12.5 Kernel regression

We have already discussed and introduce quite in details the concepts of kernel smoothing in the section Statistics at page 1143.

We estimate a regression model $\hat{y}(x)$ by the "kernel regression estimate" defined by:

$$\hat{y}(x)=\frac{\displaystyle\sum_{i=1}^{n} y_i K\left(\frac{x-x_i}{h}\right)}{\displaystyle\sum_{i=1}^{n}\left(\frac{x-x_i}{h}\right)} \tag{14.1.729}$$

Note that each $\hat{y}(x)$ is a weighted average of the y_i, which is the operation:

$$p_1 y_1+\ldots+p_n y_n \tag{14.1.730}$$

where the p_i are non-negative constants that sum to one and obviously given by:

$$p_i=\frac{K\left(\dfrac{x-x_i}{h}\right)}{\displaystyle\sum_{i=1}^{n}\left(\frac{x-x_i}{h}\right)} \tag{14.1.731}$$

Again, a very common smoother is for example the gaussian kernel:

$$K(x)=e^{-x^2/2} \tag{14.1.732}$$

For examples of such kernel regression, the reader can refer to our R companion book.

An improvement to kernel smoothing is "local polynomial smoothing" or also named "locally estimated scatterplot smoothing" (LOESS), which does the following: suppose we choose to use a first degree polynomial (that is, use linear regression), then for each x we gave $\hat{\alpha}(x)$ and $\hat{\beta}(x)$ that minimize the weighted residual sum of squares:

$$\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)(y_i-(\alpha+\beta x_i))^2 \tag{14.1.733}$$

and set:

$$\hat{y}(x)=\hat{\alpha}(x)+\hat{\beta}(x)x \tag{14.1.734}$$

Note that we have to do this for each x for which we wish to evaluate $\hat{y}(x)$.

In the general case, the above relation we will be written:

$$\boxed{\min_{\vec{\beta}}\ \text{PRSS}\left(\vec{\beta}_K^{\text{LOESS}}(x)\right)=\sum_{i=1}^{n} K\left(\frac{\|\vec{x}-\vec{x}_i\|}{h}\right)(y_i-\vec{x}_i^T\vec{\beta})^2} \tag{14.1.735}$$

Again, for an example of such kernel regressions, the reader can refer to our R companion book.

Remark

As far as we know there is no closed form inferential techniques for the coefficients or any other regression diagnostics.

14.1.12.6 Logistic regressions (LOGIT)

Often, statistics data are relative to qualitative informations (prediction if a customer is at risk for a credit, prediction if an individual is at risk for sickness, prediction of customer crunch[23], prediction of buying a next given item, etc.) or bounded in a $[0,1]$ interval (quite common case in Design of Experiments!). However, as we shall see, the traditional inference methods do not allow to model and study this type of variables (in fact they can do but performs poorly most of time!).

A well known simple model that performs most of time quite poorly for binary classification is the "signum least squares classifier" defined as:

$$f(x)=\text{sgn}(\vec{x}^T\vec{\beta}+\beta_0) \tag{14.1.736}$$

where we used the signum function already defined in the section of Arithmetic Operators page 327.

The intuition behind this "least squares classifier" is simple. The value $f(\vec{x})$ is a number, which ideally is near $+1$ when $y_i=+1$, and near -1 when $y_i=-1$. When $f(\vec{x})$ is near $+1$ we have confidence in our guess $y_i=+1$; when it is small and negative (say, $f(x)=-0.03$), we guess that $y_i=-1$ but our confidence in the guess will be low. So we won't pursue this idea further as we can't associate any probability with it and furthermore the signum function is not continuous (of class $\mathcal{C}^1$) and this can make problems in some computer implementations).

Specific methods should be used taking into account for example the lack of continuity of the processed variables or absence of the natural order between the terms that can take the qualitative variable. We will therefore see now the most simple of these methods.

Remark

We can use Logistic Regression in a non classification way (!), in that latter case it is named "decile analysis". Decile analysis has become more like a lost art, even if it was once a popularly used technique, however the convention of teaching and bucketing machine learning problems into either "classification" or "regression" types, lead many people to forget decile analysis type analyses (excepted in finance, insurance and risk management). It is very likely that most freshly minted data scientists would not have even heard of Decile analysis. So, decile analysis is simply used to categorize dataset from highest to lowest values or vice versa (based on predicted probabilities). As obvious from the name, the analysis involves dividing the dataset (thanks to the response variable) into ten groups.

---

[23] Prediction if a customer will leave your business

14.1.12.6.1 Binomial Logistic regression

As we have seen above, the simple linear regression therefore aims to model the relation between an unbounded quantitative dependent variable and an unbounded quantitative explanatory variable.

When the "class variable" to explain Y is binary (yes-no, presence-absence, 0-1, etc.), the idea is to approach it at first by a probability function $P(Y=1)$ which gives at the opposite the probability of belonging to the class $Y=0$ or $Y=1$, which we will name the "logistic binomial regression" or "logit regression" or simply "binomial regression" and also "binary regression" (often used in the context of artificial neural networks that we will see later). Then, in a second step, we define a for a binary case a "cutoff" value. For example, if we take a cutoff of 0.5 then the cases for which $P(Y=1)>0.5$ will belong to the class 1 (and vice versa in the opposite case). This is why the binomial[24] logistic regression is often qualified of "binary classifier".

Remarks

R1. In fact, logistic regression is a simple probability distribution law in our case (we will see another logistic regression in the section Economy during our study of time series and yet another one in the section of Populations Dynamics).

R2. It is obviously not possible to apply systematically logistic regression to any type of data sample! Sometimes we have to look elsewhere...

R3. When the number of modalities is equal to 2, we talk about "dummy variable" (yes-no) or a "dichotomic model" or even of "indicator variable"; if it is greater than 2, we talk about "polytomous variables" (polytomous logistic regression) or "multinomial logistic regression" (see further below). Therefore the logit binomial model is a "dichotomous model".

For example, consider the dichotomous variable: "graduation". This takes two forms: "ongoing", "finished". Age is a possible predictor of this variable and we seek to model the probability of completing studies as a function of age.

---

[24] The reader should remember that the Bernoulli distribution is a special case of the binomial distribution where a single trial is conducted!

✍**Example:**

To build the graph below, we calculated and shown on the ordinate, for youth of various ages x, the percentage of those who have left school.

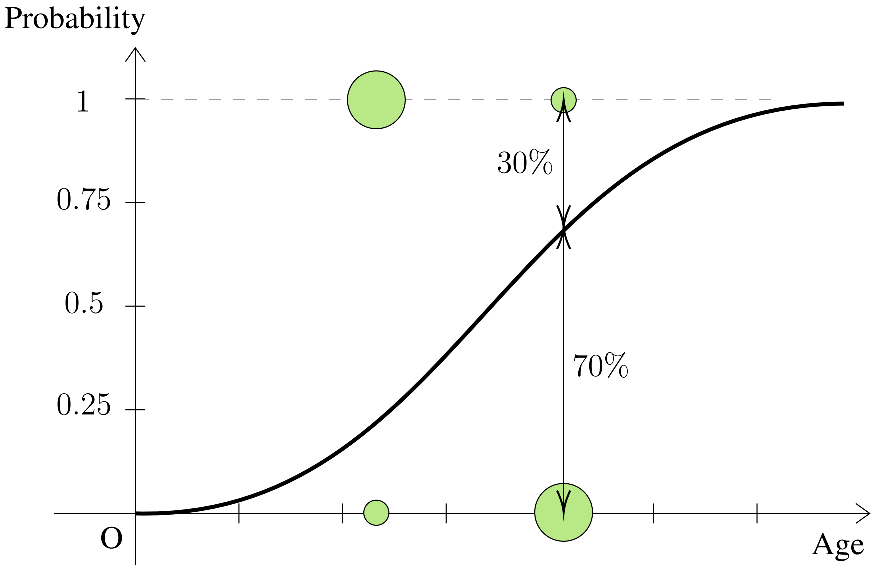


Figure 14.20: **Part at school according to Age**

But how do we obtain such a graph with a dichotomic (dummy) variable ??? In fact it's relatively simple ... Imagine a sample of 100 individuals. For these 100 individuals assume for a given age that 70% "has finished" and 30% are "ongoing". Well the curve is simply the ratio of the two classes for a given age x. It is sometimes given the size of classes with circles over the length of the horizontal asymptotes to mean that this is a dichotomous variable.

Points are distributed according to an S-curve ("sigmoid"): there are two horizontal asymptotes as the proportion is between 0 and 1. We see immediately that a linear model would be manifestly inadequate (especially as the dependent variable a linear model sweeps the whole real numbers $\mathbb{R}$ and is not confined to the range $[0,1]$).

This curve evokes for some, rightly, a cumulative curve representing a distribution function (of a Normal distribution for instance, but other continuous distributions have almost the same shape). Thus, to fit a curve to this representation, we could move towards the distribution function of a Normal distribution, and instead of estimating the parameters a and b of the linear regression, we could estimate the parameters μ, σ of the Normal law (which is very similar to the logistic law as will be shown below). We then speak of a "probit model" (probability-unit). We could also move towards a Student distribution. We then speak of a "robit model" (robust-unit model) because the T-distribution has fatter tails than the Normal distribution, the model allows for occasional large errors, and as a result the estimate for β is less affected by outliers. Robit regression is equivalent to probit regression when its degrees of freedom $k \rightarrow +\infty$, and is close to logistic regression when $k=7$.

Note that various sources and softwares interpret gompit, Cloglog, loglog, nloglog in different ways. All these models are closely related to each other and one can obtain any of them by using the gompit link function. For instance, if another source names complementary loglog (Cloglog) the function we name here gompit, one can switch between the two models by reversing the 0s and 1s in the dependent variable (in binary dependent variable models).

Remark

Is logit better than probit, or vice versa? Both methods will yield similar inferences (though not identical: in practice the end result of these different distributional assumptions is that coefficients differ, usually logit estimate should be divided by approx 1.6 to match probit estimate of the same data). But if we look at marginal effects (meaning the effects on the predicted mean of the outcome holding other covariates at the mean or averaging over observed values) the logit and probit models will make essentially the same predictions.

On the other hand, if we are not going to go about calculating the margins then logit has the obvious advantage of generating coefficients that can be transformed into the familiar odds ratio by exponentiating the coefficient. Probit and Robit coefficients are essentially uninterpretable.

Logit – also known as logistic regression – is more popular in health sciences like epidemiology partly because coefficients can be interpreted in terms of odds ratios. Probit models can be generalized to account for non-constant error variances in more advanced econometric settings (known as heteroskedastic probit models) and hence are used in some contexts by economists and political scientists. If these more advanced applications are not of relevance, than it does not matter which method you choose to go with.

The law that will interest us, however, is the logistic law. Unlike the Normal distribution, we know how to evaluate the expression of its distribution function (cumulative probability) that is of the type (his first advantage!):

$$\boxed{P(Y)=\frac{e^{ax+b}}{1+e^{ax+b}}} \tag{14.1.737}$$

for a single predictor variable x where P is obviously a probability between 0 and 1. We will see a little further the historical reason for this choice.

We immediately see that this last relation being the integral of a density function (see the proof a little bit further below) it is therefore indeed a cumulative function as:

$$\int\limits_{-\infty}^{+\infty} f(x)\mathrm{d}x=F(x)|_{-\infty}^{+\infty}=\left.\frac{e^{ax+b}}{1+e^{ax+b}}\right|_{-\infty}^{+\infty}=1-0=1 \tag{14.1.738}$$

If there are several predictor variables then we write:

$$P(Y)=\frac{e^{a_0+a_1x_1+a_2x_2+\ldots+a_nx_n}}{1+e^{a_0+a_1x_1+a_2x_2+\ldots+a_nx_n}}=\frac{e^{\vec{x}^T\vec{\beta}}}{1+e^{\vec{x}^T\vec{\beta}}} \tag{14.1.739}$$

When we choose the logistic distribution function, we get the logistic regression model, or "logit model" for the choice of the "link function" and this is its second advantage (the most important one in fact!): we can make statistics on binary variable as if we were doing a simple linear regression!

Thus, to come back on our previous example, we estimate the cumulative probability for an individual of age x to have finish his studies (there are several ways to write the law following the traditions and the context) with the following logistic distribution function:

$$P(Y)=\frac{e^{ax+b}}{1+e^{ax+b}}=\frac{1}{\dfrac{1+e^{ax+b}}{e^{ax+b}}}=\frac{1}{1+e^{-(ax+b)}} \tag{14.1.740}$$

Remark

Again, the logit and probit mainly differ on the link function:

- In Logit: $P(Y=1|\vec{x})=\left(1+e^{-\vec{x}^T\vec{\beta}}\right)^{-1}$
- In Probit: $P(Y=1|\vec{x})=\Phi(\vec{x}^T\vec{\beta})$ (cumulative Normal probability density function)

In other way, logistic has slightly flatter tails. i.e the probit curve approaches the axes more quickly than the logit curve.

Notice that in probit models latent variable is assumed to be of the form: $Y=a+bx+\varepsilon$ with $\varepsilon=\mathcal{N}(0,1)$.

A question that may arise is what if $\varepsilon=\mathcal{N}(0,\sigma)$? In fact as we know by the property of the Normal distribution it is equivalent to: $Y=a+bx+\sigma\varepsilon$ with $\varepsilon=\mathcal{N}(0,1)$. Then:

$$P(Y=1|X)=P(Y>0)=P(\varepsilon<\tfrac{1}{\sigma}(a+bx))=\Phi\left(\frac{a+bx}{\sigma}\right)=\Phi\left(\frac{a}{\sigma}+\frac{b}{\sigma}x\right)=\Phi\left(\tilde{a}+\tilde{b}x\right) \tag{14.1.741}$$

The only thing that change are the parameters!

It therefore follows the distribution function[25]:

$$f(x)=\frac{\mathrm{d}P(Y)}{\mathrm{d}x}=\frac{\mathrm{d}}{\mathrm{d}x}\left(1+e^{-(ax+b)}\right)^{-1}=-\left(1+e^{-(ax+b)}\right)^{-2}\left(-ae^{-(ax+b)}\right)=\frac{ae^{-(ax+b)}}{\left(1+e^{-(ax+b)}\right)^2} \tag{14.1.742}$$

Obviously, depending on the value of the probability we associate with the age x the fact of not having finished his studies (state associated with the binary value: 0) or have them finished (state associated with the binary value: 1).

---

[25] If the reader is looking for the second derivative and inflection point of the sigmoid function, see page 7498

Remark

After a change of variables, we fall back on the logistic law as defined on Wikipedia:

$$f(x)=\frac{ae^{-(ax+b)}}{\left(1+e^{-(ax+b)}\right)^2}\underset{\substack{a:=\frac{1}{s}\\ b:=\frac{\mu}{s}}}{=}\frac{ae^{-\left(\frac{x}{s}+\frac{\mu}{s}\right)}}{\left(1+e^{-\left(\frac{x}{s}+\frac{\mu}{s}\right)}\right)^2}=\frac{ae^{-\left(\frac{x+\mu}{s}\right)}}{\left(1+e^{-\left(\frac{x+\mu}{s}\right)}\right)^2}\qquad(14.1.743)$$

Let us indicate that if a is set as 1, and b as 0, then we have the "standard logistic law" given by:

$$P(Y)=\frac{1}{1+e^{-x}}\qquad(14.1.744)$$

We can also calculate the mean of the distribution of the function by applying what has already been study in the Statistics section but part of this integral can only be solved numerically (at least as far as we know)... if we put:

$$t=ax+b\qquad(14.1.745)$$

as being the random variable then we can formally calculate the mean of the logistic law (the reader may have noticed that it is as if we posed as $a=1$ and $b=0$). Indeed, starting from:

$$P(Y)=\frac{1}{1+e^{-(ax+b)}}=\frac{1}{1+e^{-t}}\Leftrightarrow f(x)=\frac{e^{-t}}{(1+e^{-t})^2}\qquad(14.1.746)$$

Therefore it comes:

$$\text{E}(Y)=\int\limits_{-\infty}^{+\infty}f(t)t\mathrm{d}t=\int\limits_{-\infty}^{+\infty}\frac{e^{-t}}{(1+e^{-t})^2}t\mathrm{d}t\qquad(14.1.747)$$

that after a numerical integration gives 0. We then also get the following result:

$$\text{V}(X)=\text{E}(X^2)-\text{E}(X)^2=\text{E}(X^2)=\int\limits_{-\infty}^{+\infty}f(t)t^2\mathrm{d}t=\int\limits_{-\infty}^{+\infty}\frac{e^{-t}}{(1+e^{-t})^2}t^2\mathrm{d}t\qquad(14.1.748)$$

Let us calculate the integral:

$$\begin{aligned}\int\limits_{-\infty}^{+\infty}\frac{x^2e^{-x}}{(1+e^{-x})^2}\mathrm{d}x&=\int\limits_{-\infty}^{0}\frac{x^2e^{-x}}{(1+e^{-x})^2}\mathrm{d}x+\int\limits_{0}^{+\infty}\frac{x^2e^{-x}}{(1+e^{-x})^2}\mathrm{d}x=\int\limits_{0}^{+\infty}\frac{x^2e^{x}}{(1+e^{x})^2}\mathrm{d}x+\int\limits_{0}^{+\infty}\frac{x^2e^{-x}}{(1+e^{-x})^2}\mathrm{d}x\\&=\int\limits_{0}^{+\infty}\frac{x^2e^{x}}{e^{2x}(1+e^{-x})^2}\mathrm{d}x+\int\limits_{0}^{+\infty}\frac{x^2e^{-x}}{(1+e^{-x})^2}\mathrm{d}x\\&=2\int\limits_{0}^{+\infty}\frac{x^2e^{-x}}{(1+e^{-x})^2}\mathrm{d}x=2\int\limits_{0}^{+\infty}x^2\sum_{n=1}^{+\infty}n(-1)^{n-1}e^{-nx}\mathrm{d}x\\&=2\sum_{n=1}^{+\infty}n(-1)^{n-1}\int\limits_{0}^{+\infty}x^2e^{-nx}\mathrm{d}x=2\sum_{n=1}^{+\infty}n(-1)^{n-1}\frac{2}{n^3}=4\sum_{n=1}^{+\infty}(-1)^{n-1}\frac{1}{n^2}=\frac{\pi^2}{3}\end{aligned}\qquad(14.1.749)$$

Thus we see that if we put:

$$P(Y) = \frac{1}{1+e^{-\frac{\pi}{\sqrt{3}}}(ax+b)} \tag{14.1.750}$$

We fall back on a distribution function with the same position and dispersion parameters of a Normal distribution centered reduced variable (zero mean and unit variance).

The distribution function:

$$P(Y) = \frac{e^{(ax+b)}}{1+e^{(ax+b)}} = \frac{1}{1+e^{-(ax+b)}} \tag{14.1.751}$$

can also be transformed in a very well known and important form!:

$$P \cdot \left(1+e^{(ax+b)}\right) = e^{(ax+b)} \Leftrightarrow P = e^{(ax+b)} \cdot (1-P) \tag{14.1.752}$$

Therefore:

$$\boxed{ax+b = \ln\left(\frac{P}{1-P}\right) = \text{logit}(P)} \tag{14.1.753}$$

In fact this is where lies historically the trick of the origin of logistic regression. We transform a variable P taking values in the in the range $[0,1]$ thanks to the logarithm of the ratio $P/(1-P)$ in a variable taking its values on the set $\mathbb{R}$ and therefore it is possible to associate to it a standard linear regression. Certainly it is empirical, but the idea was pretty good!

What some also write...:

$$\text{logit}\left(P(Y=1/X=x)\right) = \log\left(\frac{P(Y=1/X=x)}{1-P(Y=1/X=x)}\right) \tag{14.1.754}$$

The result of the last transformation is named the "logit". It is equal to the logarithm of the "odds" (which will be discussed in more detail soon):

$$\text{odds} = \frac{P}{1-P} \tag{14.1.755}$$

So this is just the ratio of a likelihood of an event on the probability of the complementary event (or opposite event if you prefer).

Remark

So what logit and probit do, in essence, is take the linear model and feed it through a function to yield a non-linear relationship. Whereas the linear regression predictor looks like:

$$Y = ax+b \tag{14.1.756}$$

The logit and probit predictor can be written as:

$$Y = f(ax+b) \tag{14.1.757}$$

Both functions will take any number and rescale it to fall between 0 and 1. Hence, whatever $ax+b$ equals, it can be transformed by the function to yield a predicted probability.

So when the coefficients a and b have been determined, the above expression is used to determine P knowing x easily (it comes to solve a linear equation) and vice versa! Moreover, since x is a dichotomic (dummy) variable coefficients are easily interpretable.

Link type	**CDF**	**Link function**	**Tolerance model**
Logit	Logistic	$\vec{x}_i^T\vec{\beta}=\ln\left(\dfrac{\pi_i}{1-\pi_i}\right)$	$\pi_i=\dfrac{\exp\left(\vec{x}_i^T\vec{\beta}\right)}{1+\exp\left(\vec{x}_i^T\vec{\beta}\right)}$
Probit	Normal	$\vec{x}_i^T\vec{\beta}=\Phi^{-1}\left(\pi_i\right)$	$\pi_i=\Phi\left(\vec{x}_i^T\vec{\beta}\right)$
Robit	Student	$\vec{x}_i^T\vec{\beta}=T^{-1}_{(\nu-1)/\nu}\left(\pi_i\right)$	$\pi_i=T_{(\nu-1)/\nu}\left(\vec{x}_i^T\vec{\beta}\right)$
loglog (nloglog)	Gumbel	$\vec{x}_i^T\vec{\beta}=-\log(-\log(\pi_i))$	$\pi_i=\exp\left(-\exp\left(\vec{x}_i^T\vec{\beta}\right)\right)$
Cloglog (gompit)	Gumbel	$\vec{x}_i^T\vec{\beta}=\log\left(-\log\left(1-\pi_i\right)\right)$	$\pi_i=1-\exp\left(-\exp\left(\vec{x}_i^T\vec{\beta}\right)\right)$
Cauchit	Cauchy	$\vec{x}_i^T\vec{\beta}=\tan(\Phi(\pi-0.5))$	$\pi_i=\dfrac{1}{\pi}\arctan(\vec{x}_i^T\vec{\beta})+0.5$

Table 14.8: **Some binomial regression models and their corresponding tolerance model and link function (using most common notation in multivariate case)**

Remark

The odds is also sometimes named the "rating" by analogy to the rating of horses to the triple forecast. For example, if a horse has 3 chances on 4 to be a winner (thus verbatim 1 on 4 chance of being non-winner) its rating is a 3 against 1 ratio, that is to say an odds equal to 3:

$$\text{O.R.}=\frac{3/4}{1-3/4}=3$$

We can also introduce the concept of "odds ratio" (O.R.) for the rating ratio which is a very widely used indicator in medicine. Thus, if the occurrence of an event in a group A is p, and q in the group B, the odds ratio is then simply given by:

$$\text{O.R.}=\frac{\text{odds of } p}{\text{odds of } q}=\frac{\dfrac{p}{1-p}}{\dfrac{q}{1-q}}=\frac{p(1-q)}{q(1-p)} \tag{14.1.758}$$

The odds ratio is always by design greater than or equal to zero. If the odds ratio is close to 1, the event is independent of the group, if it is greater than 1, the event is more common in group A than in group B, if it is less than 1 the event is less frequent in group A than in group B.

Odds ratio or "crude ratio" are obtained when we consider the effect of only one predictor variable. However when we include more variables in the analysis we get what is named the "adjusted odds ratio".

Let us come back over the odds because it is possible to introduce the concept of logistics function by doing the opposite approach from the one presented above (i.e. to start with the definition of odds to get to the logit) and this can sometimes be even more educational.

Let us suppose we start from the size (height) of a person to predict whether this person is a man or a woman. So we can talk about probability of being a man or a woman. Suppose the probability of a man for a given height is 90%. So the odds of being a man is:

$$\boxed{\text{odds} = \frac{P}{1-P}} \tag{14.1.759}$$

In our example, the odds will be $0.90/0.10$ therefore equal to 9. Now, the probability of being a woman will be $0.10/0.90$ therefore equal to 0.11. This asymmetry of values is not talking because the odds of being a man should be the opposite of the odds of being a woman ideally. We solve precisely this asymmetry using the natural logarithm. Thus we have:

$$\ln\left(\frac{P}{1-P}\right) = \ln(9) = 2.197 \tag{14.1.760}$$

and:

$$\ln\left(\frac{1-P}{1-(1-P)}\right) = \ln\left(\frac{1-P}{P}\right) = \ln(0.11) = -2.197 \tag{14.1.761}$$

In this way the logit (logarithm of the odds) is exactly the opposite of this of being a woman by the property of the logarithm:

$$\ln\left(\frac{x}{y}\right) = \ln(x) - \ln(y) = -\left(\ln(y) - \ln(x)\right) = -\ln\left(\frac{y}{x}\right) \tag{14.1.762}$$

To introduce this tool let us suppose a bank wants to make a scoring of its debtors. As it has several subsidiaries it (the bank) built the following data tables for some of them (all subsidiaries are then not presented):

- 1st Subsidiary:

Credit Amount	Paid	Not Paid
27,200	1	9
27,700	7	3
28,300	13	0
28,400	7	3
29,900	10	1

Table 14.9: Debtors Scoring by credit amount of subsidiary 1

- 2nd Subsidiary:

Credit Amount	Paid	Not Paid
27,200	0	8
27,700	4	2
28,300	6	3
28,400	5	3
29,900	8	0

Table 14.10: Debtors Scoring by credit amount of subsidiary 2

- 3rd Subsidiary:

Credit Amount	**Paid**	**Not Paid**
$27,200$	1	8
$27,700$	6	2
$28,300$	7	1
$28,400$	7	2
$29,900$	9	0

Table 14.11: **Debtors Scoring by credit amount of subsidiary 3**

We can see that the total proportion of good debtors in the three subsidiaries is of $91/136 \cong 0.67$.

When the credit is less than $27,500$, the percentage of good debtors is of $2/27 \cong 0.07$. When the amount of credit is less than $28,000$ the proportion of good debtors is of $19/51 \cong 0.37$.

When the amount of credit is less than $28,500$, the percentage of good debtors is of $64/108 \cong 0.59$ and for amounts below $30,000$ the proportion is of $91/936 \cong 0.67$.

We will put for this logistic regression that $Y = 1$ is a good credit risk and that $Y = 0$ is a bad risk. Next, we create the following table that is a summary of data from all subsidiaries:

Credit Amount	**Proportion** P
$27,200$	$=2/27=0.0741$
$27,700$	$=17/24=0.7083$
$28,300$	$=26/30=0.8667$
$28,400$	$=19/27=0.7037$
$29,900$	$=27/28=0.9643$

Table 14.12: **Proportion of good debtors**

Which gives graphically in Kilo-dollars:

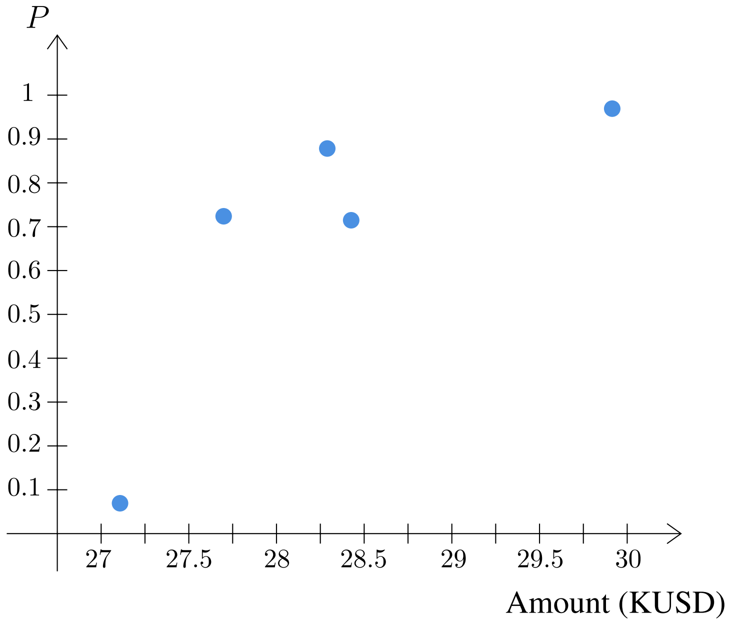


Figure 14.21: Cumulative percentage of good debtors based on credit

Once done, we use the logit transformation:

$$ax+b=\ln\left(\frac{P}{1-P}\right)=\text{logit}(P) \tag{14.1.763}$$

Which gives:

Credit Amount	Proportion P	Logit
$27,200$	0.0741	-2.5257
$27,700$	0.7083	0.8873
$28,300$	0.8667	1.8718
$28,400$	0.7037	0.8650
$29,900$	0.9643	3.2958

Table 14.13: Proportion of good debtors and logit

Remark

The reader must keep in mind the social responsibilities when running such models! First their accuracy (wrong rate of classification), but also especially the confidence interval of the predicted proportion (probability). A statistician or data scientist should never communicate punctual values alone, but always with their confidence interval!

Be careful also to have balanced (unbiased and correctly sampled) datasets, or to weight them correctly if necessary, to not have biased results especially when your classification methods or algorithms have important social impacts on people!!!

A linear regression by least squares method gives:

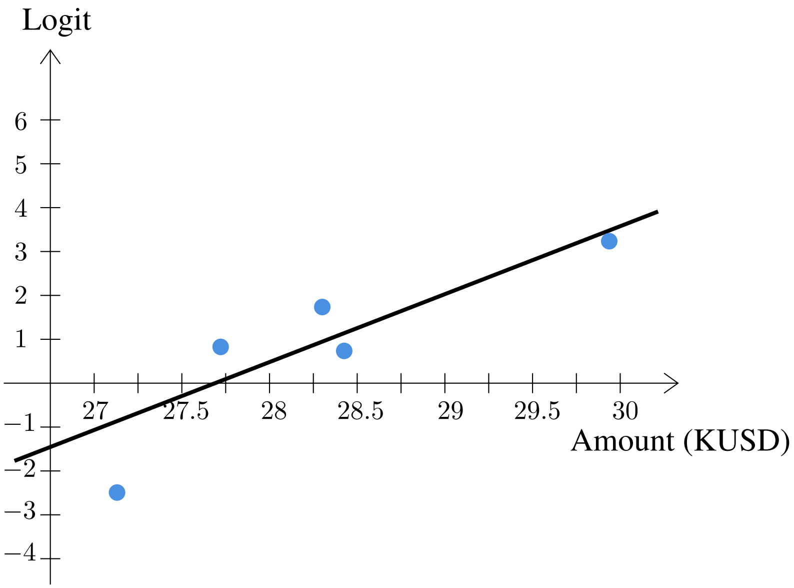


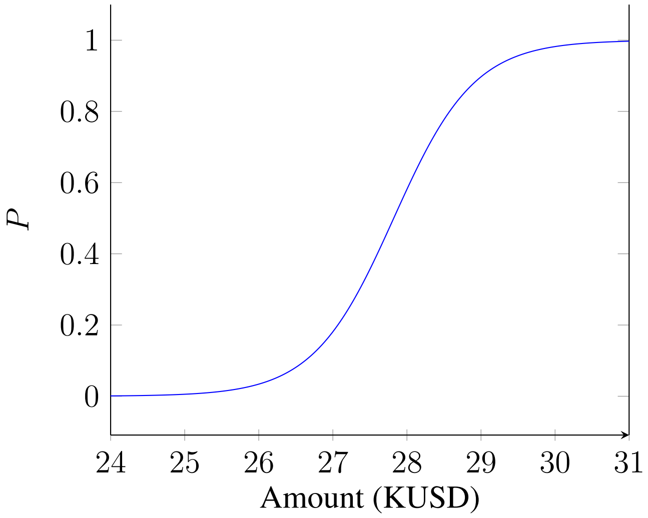
Figure 14.22: **Logit of good debtors based on the credit amount**

with for equation (where x is in USD and not in KUSD):

$$-51.1116+0.001837x=\ln\left(\frac{P}{1-P}\right)=\text{logit}(P) \qquad (14.1.764)$$

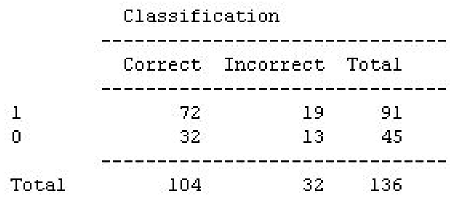
The logistic function with its representation comes then immediately (the x units are in thousands of francs)

$$P(Y)=\frac{1}{1+e^{-(-51.1116+0.001837x)}} \qquad (14.1.765)$$



Thus, it is possible to say in this example, what is the proportion P of good or bad debtors according to the credit value X smaller than or equal to a certain given value. Since 0 is a bad credit risk, we see that the bigger is the credit amount, the lower is the risk (in this hypothetical case...). Moreover, with software like Minitab (see the corresponding companion book), the difference between the quick calculations carried out by hand and those made with the binary logistic regression tool of the software is in the order of 10% (because of course ... Minitab uses the concept of maximum likelihood estimators seen in the Statistics section to determine the coefficients and the constant).

A software like Minitab gives automatically an interesting output which is the "confusion matrix $\mathcal{C}$". It compares the model to reality with a traditional cutoff set at 50% (obviously if the model perfectly match to the sample, the following matrix is a diagonal one):



```
              Classification
            --------------------------------
              Correct  Incorrect  Total
            --------------------------------
1                  72         19     91
0                  32         13     45
            --------------------------------
Total             104         32    136
```

Figure 14.23: **Confusion matrix Minitab 15.0 output example (screenshot)**

where the good debtors have the value 1 (no credit risk) and the bad one the value 0 (credit risk). We will detail further below how to get this matrix with a spreadsheet software. Finally let us indicate that in many Data Mining softwares, it is customary to define the "score" of the model as (in the specific case of one unique explanatory variable but can easily be generalized to many):

$$\text{Score}=1-P(Y)=\frac{1}{1+e^{-(ax+b)}} \tag{14.1.766}$$

Remark

In some statistical softwares, when you run a logistic regression, you can get sometimes a message of the type: *«complete separation of data points»*. Indeed, take a look at the following data set consisting of the response variable, Y, and one predictor variable, X:

X	1	2	3	4	4	5	5	6
Y	0	0	0	0	0	1	1	1

Notice the key pattern... This data set can be simply described as follows: if $X\leq 4$, then $Y=0$ without fail. Similarly, if $X>4$, then $Y=1$, again without fail. This is hat is knows as "separation". This perfect prediction of the response is what causes the estimates, and thus the model, to fail.

Following the request of a reader, here is the maximum likelihood approach of the logistic regression.

Let us consider the notation:

$$P(Y=1|X=\vec{x})=\frac{1}{1+e^{-f(\vec{x})}} \tag{14.1.767}$$

Assuming that the observations are independent of one another, we can define $L(D,\vec{\beta})$, the likelihood of the data D with respect to the logistic model parameter by $\vec{\beta}$. It is the product of the probabilities according to this model that each individual in D belongs to the observed class with $y_i=1$ or $y_i=0$:

$$\begin{aligned}L(D,\vec{\beta})&=\prod_{i=1}^{n}P(Y=y_i|X=\vec{x})\\&=\prod_{i=1}^{n}P(y_i=1|\vec{x})^{y_i}P(y_i=0|x_i)^{1-y_i}\\&=\prod_{i=1}^{n}P(y_i=1|\vec{x})^{y_i}(1-P(y_i=1|\vec{x}))^{1-y_i}\end{aligned}\tag{14.1.768}$$

It is therefore necessary to determine the vector of parameters $\hat{\vec{\beta}}$ maximizing the likelihood of the data D:

$$\hat{\vec{\beta}}=\max_{\vec{\beta}}(L(D,\vec{\beta}))\tag{14.1.769}$$

To identify $\hat{\vec{\beta}}$, the first thing to do is to derive the likelihood with respect to each component of $\vec{\beta}$, $\dfrac{\partial}{\partial\beta_j}L(D,\vec{\beta})$, then solve:

$$\frac{\partial}{\partial\beta_j}L(D,\vec{\beta})=0\tag{14.1.770}$$

since $L(D,\vec{\beta})$ is convex in $\vec{\beta}$ and $\dfrac{\partial}{\partial\beta_j}L(D,\vec{\beta})$ vanishes when the maximum is reached.

However, the logarithm is a strictly increasing function, maximizing the likelihood is equivalent to maximizing the logarithm of the likelihood as we already know it, whose expression is easier to take the derivative (the logarithm transforming the products into sums and the exponents into factors):

$$\begin{aligned}\ln\left(L(D,\vec{\beta})\right)&=\ln\left(\prod_{i=1}^{n}P(y_i=1|x_i)^{y_i}(1-P(y_i=1|\vec{x}))^{1-y_i}\right)\\&=\sum_{i=1}^{n}y_i\ln(P(y_i=1|\vec{x}))+(1-y_i)\ln((1-P(y_i=1|\vec{x})))\end{aligned}\tag{14.1.771}$$

Notice that some statistical softwares return the "deviance of the proposed model" defined as:

$$\text{DEV}(P(\vec{\beta})|y_i)=-2\ln\left(L(D,w)\right)=-2\sum_{i=1}^{n}y_i\ln(P(y_i=1|\vec{x}))+(1-y_i)\ln((1-P(y_i=1|\vec{x})))\tag{14.1.772}$$

The deviance is always larger or equal than zero, being zero only if the fit is perfect.

A benchmark for evaluating the magnitude of the deviance is the "deviance of the null model":

$$\text{DEV}_0(P(\beta_0)|y_i)=-2\ln\left(L(D,\beta_0)\right)\tag{14.1.773}$$

which is the deviance of the worst model, the one fitted without any predictor (i.e. only with the intercept)! The null deviance serves for comparing how much the model has improved by adding the predictors.

Here is a summary of what you can see in some statistical software (be careful! the concept of "null deviance" must not be confused with the concept of "deviance of the null"!):

$$\begin{aligned}\text{Null Deviance} &= 2\left[\ln\left(L(\text{Saturated Model})\right)-\ln\left(L(\text{Null Model})\right)\right] \quad \text{on df} = \text{df}_{\text{Sat}}-\text{df}_{\text{Null}}\\ \text{Residual Deviance} &= 2\left[\ln\left(L(\text{Saturated Model})\right)-\ln\left(L(\text{Proposed Model})\right)\right] \quad \text{on df} = \text{df}_{\text{Sat}}-\text{df}_{\text{Prop}}\end{aligned} \tag{14.1.774}$$

where:

- The "saturated model" is a model that assumes each data point has its own parameters (which means we have n parameters to estimate).
- The "null model" assumes the exact opposite, in that it assumes one parameter for all of the data points, which means we have only to estimate 1 parameter.
- The "proposed model" assumes we can explain our data points with p parameters + an intercept term, so we have $p+1$ parameters.

Having defined these two deviances we can if we wish build a likelihood-ratio test from them (see section Statistics page 1150). That test is sometimes named the "deviance test".

Let us now just prove that the log-likelihood of the saturated model is always zero. As the saturated model is the model that perfectly fits the observed response.

Let us remind that the likelihood as already seen above is given by:

$$L(D,\vec{\beta})=\prod_{i=1}^{n}P(y_i=1|\vec{x})^{y_i}(1-P(y_i=1|\vec{x}))^{1-y_i} \tag{14.1.775}$$

Therefore as the model is perfect:

$$L(D,\vec{\beta})=\prod_{i=1}^{n}P(y_i=1|\vec{x})^{y_i}(1-P(y_i=1|\vec{x}))^{1-y_i}=1^n=1 \tag{14.1.776}$$

and clearly the logarithm of this is 0.

Remark

Notice that we define the "cross-entropy":

$$\boxed{H(p,q)=\sum_i p_i\log(q_i)=\mathrm{E}(\log(q_i))} \tag{14.1.777}$$

where p and q denote a "true" and an "empirical/estimated" distribution, respectively. Both are discrete distributions, hence we can sum over their individual components, denoted with i. Above, when we refer to a "distribution", it means with respect to a single training data point, and not the "distribution of training data points". That's a different concept.

So we have now therefore to determine the vector $\hat{\vec{\beta}}$ maximizing the log-likelihood of the data D:

$$\hat{\vec{\beta}} = \max_{\vec{\beta}}(\ln(L(D,w))) = \max_{\vec{\beta}}(L(D,w)) \qquad (14.1.778)$$

We shall therefore have to express $\dfrac{\partial}{\partial \beta_j}\ln(L(D,\vec{\beta}))$. Before that, we can agree a slightly different notation, in order to improve the readability of the formulas. We consider that each vector $\vec{x} \in D$ has an additional component, x_0, always equal to 1 (same technique as the one used to generalize the univariate regression). Thus we can rewrite P, so that the parameter of the sigmoid function is the dot product of the vectors $\vec{\beta}$ and $\vec{x}$ (both of dimension $d+1$):

$$P(Y=1|X=\vec{x}) = \frac{1}{1+e^{-\vec{\beta}\circ\vec{x}}} = S(\vec{\beta}\circ\vec{x}) = S(\vec{\beta}^T\vec{x}) \qquad (14.1.779)$$

Where $\vec{\beta}\circ\vec{x}$ is commonly known as the "prognostic index" (PI). The expression to be derived is thus:

$$\begin{aligned}
\frac{\partial}{\partial \beta_j}\ln(L(D,\vec{\beta})) &= \frac{\partial}{\beta_i}\sum_{i=1}^{n} y_i \log(P(y_i=1|\vec{x})) + (1-y_i)\log(1-P(y_i=1|\vec{x}))\\
&= \frac{\partial}{\partial \beta_j}\sum_{i=1}^{n} y_i \ln(S(\vec{\beta}\circ\vec{x})) + (1-y_i)\log(1-S(\vec{\beta}\circ\vec{x}))\\
&= \sum_{i=1}^{n} y_i \frac{\partial}{\partial \beta_j}\ \ln(S(\vec{\beta}\circ\vec{x})) + (1-y_i)\frac{\partial}{\partial \beta_j}\log(1-S(\vec{\beta}\circ\vec{x}))
\end{aligned} \qquad (14.1.780)$$

First we compute the first term:

$$\begin{aligned}
\frac{\partial}{\partial \beta_j}\ln(S(\vec{\beta}\circ\vec{x})) &= \frac{\partial}{\partial \beta_j}\ln\left(\frac{1}{1+e^{-\vec{\beta}\circ\vec{x}}}\right) = \frac{\partial}{\partial \beta_j}\ln(1) - \frac{\partial}{\partial \beta_j}\ln(1+e^{-\vec{\beta}\circ\vec{x}})\\
&= \frac{\partial}{\partial \beta_j}\cdot 0 - \frac{\partial}{\partial \beta_j}\ln(1+e^{-\vec{\beta}\circ\vec{x}}) = -\frac{-x_{ij}e^{-\vec{\beta}\circ\vec{x}}}{1+e^{-\vec{\beta}\circ\vec{x}}}\\
&= x_{ij}\left(1-\frac{1}{1+e^{-\vec{\beta}\circ\vec{x}}}\right) = x_{ij}(1-S(\vec{\beta}\circ\vec{x}))
\end{aligned} \qquad (14.1.781)$$

Now we compute the second term:

$$\begin{aligned}\frac{\partial}{\partial\beta_j}\ln(1-S(wx_i)) &= \frac{\partial}{\partial\beta_j}\ln\left(1-\frac{1}{1+e^{-\vec{\beta}\circ\vec{x}}}\right)=\frac{\partial}{\partial\beta_j}\ln\left(\frac{1+e^{-\vec{\beta}\circ\vec{x}}}{1+e^{-\vec{\beta}\circ\vec{x}}}-\frac{1}{1+e^{-\vec{\beta}\circ\vec{x}}}\right)\\ &=\frac{\partial}{\partial\beta_j}\ln\left(\frac{e^{-\vec{\beta}\circ\vec{x}}}{1+e^{-\vec{\beta}\circ\vec{x}}}\right)=\frac{\partial}{\partial\beta_j}\left[\ln\left(e^{-\vec{\beta}\circ\vec{x}}\right)-\ln\left(1+e^{-\vec{\beta}\circ\vec{x}}\right)\right]\\ &=\frac{\partial}{\partial\beta_j}\left[-\vec{\beta}\circ\vec{x}-\ln\left(1+e^{-\vec{\beta}\circ\vec{x}}\right)\right]\\ &=-\vec{\beta}\circ\vec{x}-\ln\left(1+e^{-\vec{\beta}\circ\vec{x}}\right)=-x_{ij}+x_{ij}(1-S(\vec{\beta}\circ\vec{x}))\\ &=-x_{ij}S(\vec{\beta}\circ\vec{x})\end{aligned}\tag{14.1.782}$$

So finally:

$$\begin{aligned}\frac{\partial}{\partial\beta_j}\ln(L(D,\vec{\beta})) &= \sum_{i=1}^{n} y_i\frac{\partial}{\partial\beta_j}\ln(S(\vec{\beta}\circ\vec{x}))+(1-y_i)\frac{\partial}{\partial\beta_j}\log(1-S(\vec{\beta}\circ\vec{x}))\\ &=\sum_{i=1}^{n} y_ix_{ij}(1-S(\vec{\beta}\circ\vec{x}))-x_{ij}(1-y_i)S(\vec{\beta}\circ\vec{x})\\ &=\sum_{i=1}^{n} y_ix_{ij}-y_ix_{ij}S(\vec{\beta}\circ\vec{x})-x_{ij}S(\vec{\beta}\circ\vec{x})+y_ix_{ij}S(\vec{\beta}\circ\vec{x})\\ &=\sum_{i=1}^{n} y_ix_{ij}-x_{ij}S(\vec{\beta}\circ\vec{x})=\sum_{i=1}^{n} x_{ij}(y_i-S(\vec{\beta}\circ\vec{x}))\end{aligned}\tag{14.1.783}$$

As far as we know... we can not solve this analytically:

$$\frac{\partial}{\partial\beta_j}\ln(L(D,\vec{\beta}))=\sum_{i=1}^{n} y_ix_{ij}-x_{ij}S(\vec{\beta}\circ\vec{x})=\sum_{i=1}^{n} x_{ij}(y_i-S(\vec{\beta}\circ\vec{x}))=0\tag{14.1.784}$$

Therefore we will have to approach the solution by a numerical method anyway... (gradient descent again)!

14.1.12.6.1.1 Binomial Logistic regression relative variable importance

We have seen for the Gaussian multivariate linear model how to calculate variable importance earlier above (see page 6395). Now for logistic regression we use rather "relative variable importance".

To introduce that latter let us first calculate the partial derivative of the logistic regression relatively to one predictor (we used the quotient rule proved in the section of Differential and

Integral Calculus page 1481):

$$\begin{aligned}\frac{\partial P(Y)}{\partial x_i}=\frac{\partial}{\partial x_i}\left(\frac{e^{\beta_0+\vec{\beta}^T\vec{x}}}{1-e^{\beta_0+\vec{\beta}^T\vec{x}}}\right)&=\frac{\dfrac{\partial}{\partial x_i}\left[e^{\beta_0+\vec{\beta}^T\vec{x}}\right]\left(1-e^{\beta_0+\vec{\beta}^T\vec{x}}\right)-e^{\beta_0+\vec{\beta}^T\vec{x}}\dfrac{\partial}{\partial x_i}\left[1-e^{\beta_0+\vec{\beta}^T\vec{x}}\right]}{\left(1-e^{\beta_0+\vec{\beta}^T\vec{x}}\right)^2}\\&=\frac{\beta_i\left[e^{\beta_0+\vec{\beta}^T\vec{x}}\right]\left(1-e^{\beta_0+\vec{\beta}^T\vec{x}}\right)+\beta_i e^{\beta_0+\vec{\beta}^T\vec{x}}e^{\beta_0+\vec{\beta}^T\vec{x}}}{\left(1-e^{\beta_0+\vec{\beta}^T\vec{x}}\right)^2}\\&=\frac{\beta_i e^{\beta_0+\vec{\beta}^T\vec{x}}-\beta_i e^{2\beta_0+2\vec{\beta}^T\vec{x}}+\beta_i e^{2\beta_0+2\vec{\beta}^T\vec{x}}}{\left(1-e^{\beta_0+\vec{\beta}^T\vec{x}}\right)^2}\\&=\beta_i\frac{e^{\beta_0+\vec{\beta}^T\vec{x}}}{\left(1-e^{\beta_0+\vec{\beta}^T\vec{x}}\right)^2}\end{aligned}\tag{14.1.785}$$

So this result measures the effect of x_i on $P(Y)$. This effect is a function of x_i. However, the relative importance of two predictors is:

$$\frac{\dfrac{\partial P(Y)}{\partial x_i}}{\dfrac{\partial P(Y)}{\partial x_j}}=\frac{\beta_i}{\beta_j}\tag{14.1.786}$$

which is independent of $\vec{x}$. Thus, provided we have standardized all predictors, we can look at the estimates of the model coefficients as indicators of the relative importance of the predictors for what it concerns the variation of the output.

14.1.12.6.1.2 Binomial logistic regression residuals

Residual analysis for logistic regression is more difficult than for linear regression models because the responses y_i take only the values 0 and 1. Consequently the ith ordinary residual between the real value y_i and the model fitted value, that we will denote now by $\hat{\pi}$, is given obviously by:

$$e_i=y_i-\hat{\pi}_i\tag{14.1.787}$$

And obviously we have that (as y_i take only the values 0 and 1):

$$e_i=\begin{cases}1-\hat{\pi}_i & \text{if } y_i=1\\ -\hat{\pi}_i & \text{if } y_i=0\end{cases}\tag{14.1.788}$$

The ordinary residuals will not be normally distributed and, indeed, their distribution under the assumption that the fitted model is correct is unknown. Plots of ordinary residuals against fitted values or predictor variables will generally be uninformative.

The ordinary residuals can be made more comparable by dividing them by the estimated standard deviation of y_i, namely for the binomial distribution (as we have proved it during our study of the binomial distribution):

$$\sqrt{\hat{\pi}_i\left(1-\hat{\pi}_i\right)}\tag{14.1.789}$$

The resulting "Pearson residuals" are given by:

$$r_{P_i}=\frac{y_i-\hat{\pi}_i}{\sqrt{\hat{\pi}_i\left(1-\hat{\pi}_i\right)}} \tag{14.1.790}$$

If we square:

$$r_{P_i}^2=\frac{\left(y_i-\hat{\pi}_i\right)^2}{\hat{\pi}_i\left(1-\hat{\pi}_i\right)} \tag{14.1.791}$$

Now notice that this can be written:

$$\begin{aligned}r_{P_i}^2&=\frac{\left(y_i-\hat{\pi}_i\right)^2}{\hat{\pi}_i\left(1-\hat{\pi}_i\right)}=\frac{\left(y_i-\hat{\pi}_i\right)^2}{1-\hat{\pi}_i}+\frac{\left(y_i-\hat{\pi}_i\right)^2}{\hat{\pi}_i}\\&=\frac{\left(\left(1-y_i\right)-\left(1-\hat{\pi}_i\right)\right)^2}{1-\hat{\pi}_i}+\frac{\left(y_i-\hat{\pi}_i\right)^2}{\hat{\pi}_i}\end{aligned} \tag{14.1.792}$$

So we recognize here that all the terms are of the form:

$$\frac{(O_i-E_i)^2}{E_i} \tag{14.1.793}$$

Therefore the same applies to the sum of square of the Pearson residuals and we can write the "Pearson chi-square statistic":

$$\sum_{i=1}^{n} r_{P_i}^2=\sum_{i=1}^{n}\frac{\left(y_i-\hat{\pi}_i\right)^2}{\hat{\pi}_i\left(1-\hat{\pi}_i\right)}=\chi^2(n-1) \tag{14.1.794}$$

Hence, we see that the sum of the squares of the Pearson residuals is numerically equal to the Pearson chi-square test statistic. Therefore the square of each Pearson residual measures the contribution of each binary response to the Pearson chi-square test statistic and can be used as an adequation test for our model.

Remark

Notice that some practitioners run not only a chi-square test on residuals but also on the resulting contingency classification table of all discrete x values of the binomial regression. For our example above the corresponding contingency table being:

Credit Amount	Expected	Observed
$27,200$	0	2
$27,700$	0	17
$28,300$	30	26
$28,400$	27	10
$29,900$	28	27

Unfortunately, it is common that there are not enough observations (or even worse...: zero observations) for each possible combinations of values of the x variables, so the Pearson chi-squared statistic cannot be readily calculated (see a practical example in our `R` companion book). A solution to this problem is the Hosmer-Lemeshow statistic (see below). The key concept of the Hosmer-Lemeshow statistic is that, instead of observations being grouped by the values of the x variable(s), the observations are grouped by expected probability. That is, observations with similar expected probability are put into the same group, usually to create approximately 10 groups.

14.1.12.6.1.3 Hosmer–Lemeshow test

The Hosmer–Lemeshow test is a statistical test for goodness of fit for binary logistic regression models (however a generalized Hosmer–Lemeshow goodness-of-fit test for multinomial logistic regression models exists!). It is used frequently in risk prediction models. The test assesses whether or not the observed event rates match expected event rates in subgroups of the model population. The Hosmer–Lemeshow test specifically identifies subgroups as the deciles of fitted risk values. Models for which expected and observed event rates in subgroups are similar are said to be "well calibrated".

The Hosmer-Lemeshow test statistic is given by:

$$\begin{aligned} H &= \sum_{g=1}^{G}\left(\frac{(O_{1g}-E_{1g})^2}{E_{1g}}+\frac{(O_{0g}-E_{0g})^2}{E_{0g}}\right) \\ &= \sum_{g=1}^{G}\left(\frac{(O_{1g}-E_{1g})^2}{N_g\pi_g}+\frac{((N_g-O_{1g})-(N_g-E_{1g}))^2}{N_g(1-\pi_g)}\right) \\ &= \sum_{g=1}^{G}\frac{(O_{1g}-E_{1g})^2}{N_g\pi_g(1-\pi_g)}=\sum_{g=1}^{G}\frac{(O_{1g}-N_g\pi_g)^2}{N_g\pi_g(1-\pi_g)} \end{aligned} \quad (14.1.795)$$

Here $O_{1g}, E_{1g}, O_{0g}, E_{og}, N_g$, and π_g denote the observed $Y=1$ events, expected $Y=1$ events, observed $Y=0$ events, expected $Y=0$ events, total observations, predicted risk for the g^{th} risk decile group, and G is the number of groups. The test statistic asymptotically follows a χ^2 distribution with $G-2$ degrees of freedom (see remark further below!!!). Hence:

$$H=\sum_{g=1}^{G}\frac{(O_{1g}-E_{1g})^2}{N_g\pi_g(1-\pi_g)}=\chi^2_{G-2} \quad (14.1.796)$$

The number of risk groups may be adjusted depending on how many fitted risks are determined by the model. This helps to avoid singular decile groups.

Another very common notation that we can find in textbooks is, given m_j the number of observations:

$$\hat{C}=\sum_{k=1}^{g}\left[\frac{(o_{1k}-\hat{e}_{1k})^2}{\hat{e}_{1k}}+\frac{(o_{0k}-\hat{e}_{0k})^2}{\hat{e}_{0k}}\right] \quad (14.1.797)$$

where:

$$o_{1k}=\sum_{j=1}^{c_k}y_j \quad o_{0k}=\sum_{j=1}^{c_k}(m_j-y_j) \quad \hat{e}_{1k}=\sum_{j=1}^{c_k}m_j\hat{\pi}_j \quad \hat{e}_{0k}=\sum_{j=1}^{c_k}m_j(1-\hat{\pi}_j) \quad (14.1.798)$$

and c_k is the number of covariate patterns in the kth group. With a bit algebra we have therefore:

$$\hat{C}=\sum_{k=1}^{g}\frac{\left(o_{1k}-n_k'\bar{\pi}_k\right)^2}{n_k'\bar{\pi}_k(1-\bar{\pi}_k)} \tag{14.1.799}$$

where $\bar{\pi}_k$ is the average estimated probability in the kth group:

$$\bar{\pi}_k=\frac{1}{n_k'}\sum_{j=1}^{c_k}m_j\hat{\pi}_j \tag{14.1.800}$$

The Hosmer–Lemeshow test seems to have some limitations. Frank Harrell describes several in an answer on the forum *Cross Validated*:

- The Hosmer-Lemeshow test is for overall calibration error, not for any particular lack of fit such as quadratic effects. It does not properly take overfitting into account, is arbitrary to choice of bins and method of computing quantiles, and often has power that is too low.
- For these reasons the Hosmer-Lemeshow test is no longer recommended. Hosmer et al. have a better one df omnibus test of fit, implemented in the `R` software rms package residuals.lrm function.
- Other alternatives have been developed to address the limitations of the Hosmer-Lemeshow test. These include the Osius-Rojek test and the Stukel test.

Remark

In the 11982 (holocene calendar) paper by Lemeshow and Hosmer *A review of goodness of fit statistics for use in the development of logistic regression models*, American Journal of Epidemiology 115:92-106. In this they write (page 96) in reference to their statistic that:

"The theoretical development given by Hosmer and Lemeshow (1980) requires only that $G>(p+1)$."

where p is the number of parameters in the model and following this that:

« Hosmer and Lemeshow (1980) have shown via computer simulations that if the number of covariates plus one is less than the number of groups (i.e. $p+1<G$), then the statistic C_G^* has a distribution which is closely approximated by a chi-square distribution with $G-2$ degrees of freedom when H_0 is true. »

To be more precise, in Hosmer and Lemeshow's 11980 (holocene calendar) paper, Theorem 2 states that the asymptotic distribution of $\hat{C}_g^*$ (the usual Hosmer-Lemeshow test statistic) is:

$$\chi^2_{2g-g-(p+1)}+\sum_{i=1}^{p+1}\lambda_i\chi_i^2(1)$$

where the λ_i are eigenvalues of a matrix (specified in the paper, not relevant to this question). Then, they show through simulations that $\sum_{i=1}^{p+1}\lambda_i\chi_i^2(1)$ is approximately

χ^2_{p-1}, which leads to the usual $g-2$ degrees of freedom in the Hosmer-Lemeshow test.

This means that given our fitted model, the p-value can be calculated as the right hand tail probability of the corresponding chi-squared distribution using the calculated test statistic. If the p-value is small, this is indicative of poor fit.

It should be emphasized, same as any other null hypothesis statistical test, that a large p-value does not mean the model fits well, since lack of evidence against a null hypothesis is not equivalent to evidence in favour of the alternative hypothesis. In particular, if our sample size is small, a high p-value from the test may simply be a consequence of the test having lower power to detect misspecification, rather than being indicative of good fit.

The reader can refer to our `R` companion book where we provide simulation evidence that the above claim is quite accurate!

14.1.12.6.1.4 Logistic Kolmogorov-Smirnov statistic

The "KS Statistic", also named "maximum Kolmogorov-Smirnov vertical difference" or "Kolmogorov-Smirnov statistic", is the maximum difference between the cumulative true positive and cumulative false positive rate. It is often used as the deciding metric to judge the efficacy of models in credit scoring. The higher the KS stat, the more efficient is the model at capturing the responders (Ones).

In other words:

$$\begin{aligned}\text{KS} &= \text{Maximum difference between Cumulative \% Event and Cumulative \% Non-Event}\\ &= \max(\text{True positive rate} - \text{False positive rate})\end{aligned} \tag{14.1.801}$$

More technically and more precise, let us remind, the null hypothesis for a logistic regression of a binary/dichotomous random variable Y_k on real numbers $x_{j,k}$ (with $k=1,2,\ldots,n$, and $j=1,2,\ldots,m \ll n$), given observations $y_1,y_2,\ldots,y_n$ of $Y_1,Y_2,\ldots,Y_n$, respectively, is H_0 : $y_1,y_2,\ldots,y_n$ are draws from independent Bernoulli distributions with means $\hat{\mu}_1,\hat{\mu}_2,\ldots,\hat{\mu}_n$, respectively:

$$\text{logit}\,(\hat{\mu}_k)=\hat{\beta}^{(0)}+\sum_{j=1}^{l}\hat{\beta}^{(j)}x_{j,k} \tag{14.1.802}$$

for $k=1,2,\ldots,n$, where the vector β is a nuisance parameter, $\hat{\beta}$ is its maximum-likelihood estimate, l is a nonnegative integer no greater than m, and:

$$\text{logit}(p)=\ln\left(\frac{p}{1-p}\right) \tag{14.1.803}$$

or something similarly aggregative). A cumulative measure of the distance between the observed data and the model assumed under the null hypothesis is the discrete KolmogorovSmirnov statistic (recommended by Horn (1977) and many others), namely:

$$\boxed{\text{KS}=\max_{1\leq j\leq n}\left|\sum_{k=1}^{j}y_{\sigma_k}-\sum_{k=1}^{j}\hat{\mu}_{\sigma_k}\right|=\max_{1\leq j\leq n}\left|\sum_{k=1}^{j}r_{\sigma_k}\right|} \tag{14.1.804}$$

where the ordering σ is a permutation of the integers $1, 2, \ldots, n$, and the residual r_k is:

$$r_k = y_k - \hat{\mu}_k \tag{14.1.805}$$

for $k = 1, 2, \ldots, n$.

Let us see a small example:

Decile	**Non Event**	**Event**	**% Non Event**	**% Event**	**Cum %Event**	**Cum % Non-Event**	**KS**
1	51	49	5.7%	48.5%	48.5%	5.7%	42.8%
2	81	19	9.0%	18.8%	67.3%	14.7%	52.6%
3	84	16	9.3%	15.8%	83.2%	24.0%	59.1%
4	90	10	10.0%	9.9%	93.1%	34.0%	59.0%
5	95	5	10.6%	5.0%	98.0%	44.6%	53.4%
6	99	1	11.0%	1.0%	99.0%	55.6%	43.4%
7	99	1	11.0%	1.0%	100.0%	66.6%	33.4%
8	100	0	11.1%	0.0%	100.0%	77.8%	22.2%
9	100	0	11.1%	0.0%	100.0%	88.9%	11.1%
10	100	0	11.1%	0.0%	100.0%	100.0%	0.0%

In this case, KS is maximum at third decile and KS score is 59.1. Ideally, it should be in first three deciles and score lies between 40 and 70. And there should not be more than 10 points (in absolute) difference between training and validation KS score. Score above 70 is susceptible and might be overfitting so rigorous validation is required.

Remark

The tests of Hosmer and Lemeshow (11980, 12000 according to holocene calendar) replace the Kolmogorov-Smirnov statistics KS defined earlier above with χ^2 for the model of binomial distributions corresponding to quantiles (typically the deciles) for the estimated means of the postulated Bernoulli distributions. It has also the advantage of being a statistical test rather than a simple statistical estimator.

14.1.12.6.1.5 Equivalence between the 2-sample Wald's Z-test and the Rao score test over logistic regression with a single binary predictor

For information, the proof that follows if provided by .

Remeber that the Wald's Z-statistic for difference in 2 proportions with unpooled variances is of the following form:

$$Z = \frac{\hat{p}_B - \hat{p}_A}{\sqrt{p(1-p)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} \sim \mathcal{N}(0,1) \tag{14.1.806}$$

Where $\hat{p}_A$ stands for the estimated probability (sample proportion, $\%$) in the 1st group, $\hat{p_B}$ is the estimated probability in the 2nd group, n_A and n_B denote respective group sizes, and p is the pooled probability:

$$p = \frac{x_A + x_B}{n_A + n_B} \tag{14.1.807}$$

Let's start from the equation of the logistic regression with a single binary predictor:

$$\text{logit}\left(E(Y\,|\,X)\right) = \text{logit}(\hat{p}(Y=1\,|\,X)) = \ln\left(\frac{\hat{p}(Y=1\,|\,X)}{1-\hat{p}(Y=1\,|\,X)}\right) = \beta_0 + X_1\beta_1 \tag{14.1.808}$$

or equivalently, after applying the inverse-logit, i.e. sigmoid function (let's also simplify X_1 to just X):

$$\hat{p}(Y=1\,|\,X) = \frac{e^{\beta_0+X\beta_1}}{1+e^{\beta_0+X\beta_1}} = \frac{e^{\eta}}{1+e^{\eta}} \tag{14.1.809}$$

where Y_i are independent Bernoulli random variables with probabilities p_i. The second, simpler form with $\eta = \beta_0 + X\beta_1$ will facilitate later calculations.

Let's also encode the two levels $\{A, B\}$ using a single binary predictor X such that: $A : X = 0, B : \mathrm{X} = 1$, let's simplify notation and express $\hat{p}_A$ and $\hat{p}_B$ in terms of beta coefficients:

$$\begin{cases} \hat{p}_A = P(Y=1, X=0) = \frac{e^{\beta_0}}{1+e^{\beta_0}} \\ \hat{p}_B = P(Y=1, X=1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} \end{cases} \tag{14.1.810}$$

As in this special case the classification is binary but also X is discrete, we cannot have any error on the estimator $\hat{p}_i$ so we will write them without the hat symbol.

Under H_0, i.e. $\beta_1 = 0, \hat{p}_A = \hat{p}_B$ and the best estimator of this common proportion is the pooled proportion p. Let's find its form. We need to assume that the data consists of two independent binomial samples:

$$\begin{cases} \text{group } A = X_A : \text{ Binomial } (n_A, p_A) \\ \text{group } B = X_B : \text{ Binomial } (n_B, p_B) \end{cases} \tag{14.1.811}$$

So the likelihood function is:

$$L\left(p_A, p_B\right) = \binom{n_A}{x_A} p_A^{x_A}\left(1-p_A\right)^{n_A-x_A} \cdot \binom{n_B}{x_B} p_B^{x_B}\left(1-p_B\right)^{n_B-x_B} \tag{14.1.812}$$

where x_A and x_B are the observed number of successes in groups A and B, respectively, and n_A and n_B are the total sample sizes in each group.

Under the null hypothesis $H_0 : p_A = p_B = p_{\text{pooled}}$. The pooled p is of some form yet unknown. Knowing that:

$$\begin{aligned} L\left(p_{\text{pooled}}\right) &= \binom{n_A}{x_A}\binom{n_B}{x_B} p_{\text{pooled}}^{x_A+x_B}\left(1-p_{\text{pooled}}\right)^{(n_A-x_A)+(n_B-x_B)} \\ &= \binom{n_A}{x_A}\binom{n_B}{x_B} p_{\text{pooled}}^{x_A+x_B}\left(1-p_{\text{pooled}}\right)^{(n_A+n_B-x_A-x_B)} \end{aligned} \tag{14.1.813}$$

We get a single binomial likelihood where the total number of successes is $x_A + x_B$ and the total number of trials is $n_A + n_B$, i.e.:

$$X_A + X_B : \text{ Binomial}\left(n_A + n_B, p_{\text{pooled}}\right) \tag{14.1.814}$$

Now, the log-likelihood, $\log(L(p_{\text{pooled}}))=\mathcal{L}(p_{\text{pooled}})$, is given by:

$$\mathcal{L}(p_{\text{pooled}})=c^{te}\cdot(x_A+x_B)\cdot\log(p_{\text{pooled}})+(n_A+n_B-x_A-x_B)\cdot\log(1-p_{\text{pooled}}) \tag{14.1.815}$$

We wrote c^{te} to highlight that this term will disappear after taking the derivative further below.

Now by taking ${}^{\mathrm{d}\mathcal{L}}/_{\mathrm{d}p_{\text{pooled}}}$:

$$\begin{aligned}\frac{\mathrm{d}\mathcal{L}(p_{\text{pooled}})}{\mathrm{d}p_{\text{pooled}}}=\frac{x_A+x_B}{p_{\text{pooled}}}-\frac{n_A+n_B-x_A-x_B}{1-p_{\text{pooled}}}&=0\\(x_A+x_B)(1-p_{\text{pooled}})&=(n_A+n_B-x_A-x_B)\cdot p_{\text{pooled}}\\x_A+x_B&=(n_A+n_B)\cdot p_{\text{pooled}}\\p_{\text{pooled}}&=\frac{x_A+x_B}{n_A+n_B}\end{aligned} \tag{14.1.816}$$

Or, alternatively, since $p_i={}^{x_i}/_{n_i}$:

$$p_{\text{pooled}}=\frac{p_An_A+p_Bn_B}{n_A+n_B} \tag{14.1.817}$$

The log-likelihood function of the logistic regression with a single binary predictor is of the form:

$$\mathcal{L}(\vec{\beta})=\log(L(\vec{\beta}))=\log\left(\prod_{i=1}^{n}p_i^{y_i}\cdot(1-p_i)^{1-y_i}\right)=\sum_{i=1}^{n}\left[y_i\log\left(p_i\right)+(1-y_i)\log\left(1-p_i\right)\right] \tag{14.1.818}$$

where $\vec{\beta}$ is the vector of estimated coefficients, i.e. (β_0,β_1).

Let's express p and $1-p$ in terms of $\eta=\beta_0+X\beta_1$:

$$\begin{cases}p=\frac{e^{\eta}}{1+e^{\eta}}\\1-p=\frac{1}{1+e^{\eta}}\end{cases} \tag{14.1.819}$$

Then:

$$\begin{aligned}\mathcal{L}(\vec{\beta})&=\sum_{i=1}^{n}\left[y_i\log\left(\frac{e^{\eta_i}}{1+e^{\eta_i}}\right)+(1-y_i)\log\left(\frac{1}{1+e^{\eta_i}}\right)\right]\\&=\sum_{i=1}^{n}\left[y_i\left(\log\left(e^{\eta_i}\right)-\log\left(1+e^{\eta_i}\right)\right)-(1-y_i)\log\left(1+e^{\eta_i}\right)\right]\\&=\sum_{i=1}^{n}\left[y_i\eta_i-y_i\log\left(1+e^{\eta_i}\right)-(1-y_i)\log\left(1+e^{\eta_i}\right)\right]\\&=\sum_{i=1}^{n}y_i\eta_i-\log\left(1+e^{\eta_i}\right)\end{aligned} \tag{14.1.820}$$

We will now need both the gradient and Hessian (see section Sequences and Series page 1715) of this log-likelihood. For future use, we will name the gradient as "Rao score function" denoted by $\vec{U}$ and the Hessian matrix as H. First, let's find the form of $U(\vec{\beta})$:

$$\vec{U}(\vec{\beta})=\begin{bmatrix}\frac{\partial\mathcal{L}(\vec{\beta})}{\partial\beta_0}\\\frac{\partial\mathcal{L}(\vec{\beta})}{\partial\beta_1}\end{bmatrix} \tag{14.1.821}$$

By noticing that:

$$\frac{\partial \eta_i}{\partial \beta_0}=1,\ \frac{\partial \eta_i}{\partial \beta_1}=x_i \tag{14.1.822}$$

and remembering that:

$$p=\frac{e^{\eta}}{1+e^{\eta}} \tag{14.1.823}$$

we get:

$$\frac{\partial \mathcal{L}(\vec{\beta})}{\partial \beta_0}=\sum_{i=1}^{n}\left(y_i\frac{\partial \eta_i}{\partial \beta_0}-\frac{1}{1+e^{\eta_i}}e^{\eta_i}\frac{\partial \eta_i}{\partial \beta_0}\right)=\sum_{i=1}^{n}\left(y_i.1-\frac{e^{\eta_i}}{1+e^{\eta_i}}\cdot 1\right)=\sum_{i=1}^{n}(y_i-p_i) \tag{14.1.824}$$

and:

$$\frac{\partial \mathcal{L}(\vec{\beta})}{\partial \beta_1}=\sum_{i=1}^{n}\left(y_i\frac{\partial \eta_i}{\partial \beta_1}-\frac{1}{1+e^{\eta_i}}e^{\eta_i}\frac{\partial \eta_i}{\partial \beta_1}\right)=\sum_{i=1}^{n}\left(y_ix_i-\frac{e^{\eta_i}}{1+e^{\eta_i}}x_i\right)=\sum_{i=1}^{n}x_i(y_i-p_i) \tag{14.1.825}$$

So finally we have for the Rao score function:

$$U(\vec{\beta})=\begin{bmatrix}\sum_{i=1}^{n}(y_i-p_i)\\ \sum_{i=1}^{n}x_i(y_i-p_i)\end{bmatrix} \tag{14.1.826}$$

Now, the Hessian:

$$H(\vec{\beta})=\begin{bmatrix}\frac{\partial^2 \mathcal{L}(\vec{\beta})}{\partial \beta_0^2} & \frac{\partial^2 \mathcal{L}(\vec{\beta})}{\partial \beta_0\beta_1}\\ \frac{\partial^2 \mathcal{L}(\vec{\beta})}{\partial \beta_1\beta_0} & \frac{\partial^2 \mathcal{L}(\vec{\beta})}{\partial \beta_1^2}\end{bmatrix} \tag{14.1.827}$$

The partial derivatives are as follows:

$$\begin{gathered}\frac{\partial^2 \mathcal{L}(\vec{\beta})}{\partial \beta_0^2}=\frac{\partial}{\partial \beta_0}\sum_{i=1}^{n}(y_i-p_i)=\sum_{i=1}^{n}-\frac{\partial p_i}{\partial \beta_0}=-\sum_{i=1}^{n}\frac{\partial p_i}{\partial \beta_0}=-\sum_{i=1}^{n}\frac{\partial p_i}{\partial \eta_0}\frac{\partial \eta_0}{\partial \beta_0}=-\sum_{i=1}^{n}p_i(1-p_i)\\ \frac{\partial^2 \mathcal{L}(\vec{\beta})}{\partial \beta_1^2}=\frac{\partial}{\partial \beta_1}\sum_{i=1}^{n}x_i(y_i-p_i)=\sum_{i=1}^{n}x_i\left(-\frac{\partial p_i}{\partial \beta_1}\right)=-\sum_{i=1}^{n}x_ip_i(1-p_i)x_i=-\sum_{i=1}^{n}x_i^2p_i(1-p_i)\\ \frac{\partial^2 \mathcal{L}(\vec{\beta})}{\partial \beta_0\beta_1}=\frac{\partial^2 \mathcal{L}(\vec{\beta})}{\partial \beta_1\beta_0}=\frac{\partial}{\partial \beta_1}\sum_{i=1}^{n}(y_i-p_i)=-\sum_{i=1}^{n}\frac{\partial p_i}{\partial \beta_1}=-\sum_{i=1}^{n}p_i(1-p_i)x_i\end{gathered} \tag{14.1.828}$$

Therefore:

$$H(\vec{\beta})=\begin{bmatrix}-\sum_{i=1}^{n}p_i(1-p_i) & -\sum_{i=1}^{n}p_i(1-p_i)x_i\\ -\sum_{i=1}^{n}p_i(1-p_i)x_i & -\sum_{i=1}^{n}x_i^2p_i(1-p_i)\end{bmatrix} \tag{14.1.829}$$

Let's also determine the Fisher Information matrix (see section Statistics 839):

$$I(\beta) = -\mathrm{E}(H(\vec{\beta})) = \begin{bmatrix} \sum_{i=1}^{n} p_i\left(1-p_i\right) & \sum_{i=1}^{n} p_i\left(1-p_i\right)x_i \\ \sum_{i=1}^{n} p_i\left(1-p_i\right)x_i & \sum_{i=1}^{n} x_i^2 p_i\left(1-p_i\right) \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} p_i q_i & \sum_{i=1}^{n} p_i q_i x_i \\ \sum_{i=1}^{n} p_i q_i x_i & \sum_{i=1}^{n} p_i q_i x_i^2 \end{bmatrix} \tag{14.1.830}$$

where $q_i = 1 - p_i$.

Remark

As the Hessian matrix above doesn't content any estimator (no variable with a hat symbol!), but only directly perfectly known observed quantities, its mean expectation is equal to itself!

This can be further expanded, by substituting sums with appropriate (per respective group, A and B) counts of elements, remembering that $n = n_A + n_B,\ \mathrm{A}: X = 0,\ \mathrm{B}: X = 1, p_A = p(Y = 1 \,|\, X = A)$ and $p_B = p(Y = 1 \,|\, X = B)$, and:

$$I(\vec{\beta}) = \begin{bmatrix} \sum\limits_{i:X_i=0} p_i q_i + \sum\limits_{i:X_i=1} p_i q_i & \sum\limits_{i:X_i=1} 1 p_i q_i \\ \sum\limits_{i:X_i=1} 1 p_i q_i & \sum\limits_{i:X_i=1} 1 p_i q_i \end{bmatrix} = \begin{bmatrix} n_A p_A q_A + n_B p_B q_B & n_B p_B q_B \\ n_B p_B q_B & n_B p_B q_B \end{bmatrix} \tag{14.1.831}$$

Let us see now the Rao Score and Information Matrix under $H_0 : \beta_1 = 0$.

We start with the Rao Score vector, and for this purpose remember that n_A and n_B depict the respective number of individuals corresponding to group A and B, for group $A : X = 0$ and for $B : X = 1$ and $n = n_A + n_B$. Also, under $H_0, p_A = p_B = p$ (i.e. $\forall_i p_i = p$), where p is the pooled probability. Therefore:

$$U\left(\beta_0, \beta_1 = 0\right) = \begin{bmatrix} \sum_{i=1}^{n}\left(y_i - p_i\right) \\ \sum_{i=1}^{n} x_i\left(y_i - p_i\right) \end{bmatrix} = \begin{bmatrix} \left(\sum_{i=1}^{n} y_i\right) - \left(n_A + n_B\right)\cdot p \\ \left(\sum\limits_{i:X_i=1} y_i\right) - n_B p \end{bmatrix} \tag{14.1.832}$$

Here, y_i is the response vector containing $1s$ in respective group. Summing all those 1s yields the total number of successes in this group, which can be expressed as:

$$\sum_{i=1}^{n} y_i = \sum_{i:X_i\in\{0,1\}} y_i = n_A p_A + n_B p_B \overset{H_0}{\Rightarrow} \left(n_A + n_B\right)\cdot p \tag{14.1.833}$$

So, the first element of the vector becomes 0 out and finally:

$$U\left(\beta_0, \beta_1 = 0\right) = \begin{bmatrix} 0 \\ n_B\left(p_B - p\right) \end{bmatrix} \tag{14.1.834}$$

Now for the Information matrix:

$$I(\vec{\beta})=\begin{bmatrix} n_A p_A q_A + n_B p_B q_B & n_B p_B q_B \\ n_B p_B q_B & n_B p_B q_B \end{bmatrix} \tag{14.1.835}$$

we get under the same hypothesis $H_0: p_A=p_B=p$:

$$I\left(\beta_0,\beta_1=0\right)=\begin{bmatrix} pq\left(n_A+n_B\right) & n_B pq \\ n_B pq & n_B pq \end{bmatrix} \tag{14.1.836}$$

Let's also calculate I^{-1} (see section Linear Algebra 1981):

$$\Sigma=I\left(\beta_0,\beta_1=0\right)^{-1}=\frac{1}{pq\left(n_A+n_B\right)\cdot n_B pq-\left(n_B pq\right)^2}\begin{bmatrix} n_B pq & -n_B pq \\ -n_B pq & pq\left(n_A+n_B\right) \end{bmatrix} \tag{14.1.837}$$

After simplifying the denominator term:

$$pq\left(n_A+n_B\right)\cdot n_B pq-\left(n_B pq\right)^2=p^2q^2n_An_B+p^2q^2n_B^2-p^2q^2n_B^2=p^2q^2n_An_B \tag{14.1.838}$$

we finally get:

$$\Sigma\left(\beta_0,\beta_1=0\right)=I\left(\beta_0,\beta_1=0\right)^{-1}=\frac{1}{p^2q^2n_An_B}\begin{bmatrix} n_B pq & -n_B pq \\ -n_B pq & pq\left(n_A+n_B\right) \end{bmatrix}=\begin{bmatrix} \frac{1}{n_Apq} & -\frac{1}{n_Apq} \\ -\frac{1}{n_Ap^2} & \frac{n_A+n_B}{n_An_Bpq} \end{bmatrix} \tag{14.1.839}$$

The "Rao score test", named also "Lagrange multiplier test", under H_0 is defined as the following quadratic form:

$$\boxed{R=U\left(\beta_0,\beta_1=0\right)^T\cdot I\left(\beta_0,\beta_1=0\right)^{-1}\cdot U\left(\beta_0,\beta_1=0\right)} \tag{14.1.840}$$

But since the first element of U is 0, this reduces to just scalar operation:

$$R=U\cdot I^{-1}\cdot U=U^2\Sigma \tag{14.1.841}$$

$$\begin{aligned} R&=U\left(\beta_0,\beta_1=0\right)^2\Sigma\left(\beta_0,\beta_1=0\right)\\ &=\left[n_B\left(p_B-p\right)\right]^2\cdot\frac{n_A+n_B}{pqn_An_B}=\frac{\left[n_B\left(p_B-p\right)\right]^2}{pq\frac{n_An_B}{n_A+n_B}}\\ &=\frac{n_B^2\left(p_B-\frac{n_Ap_A+n_Bp_B}{n_A+n_B}\right)^2}{pq\frac{n_An_B}{n_A+n_B}}=\frac{n_B^2\left(\frac{p_Bn_A+p_Bn_B-n_Ap_A-n_Bp_B}{n_A+n_B}\right)^2}{pq\frac{n_An_B}{n_A+n_B}}\\ &=\frac{\frac{n_B^2\left(p_Bn_A-n_Ap_A\right)^2}{\left(n_A+n_B\right)^2}}{pq\frac{n_An_B}{n_A+n_B}}=\frac{\frac{n_B^2n_A^2\left(p_B-p_A\right)^2}{\left(n_A+n_B\right)^2}}{pq\frac{n_An_B}{n_A+n_B}}\\ &=\frac{\left(p_B-p_A\right)^2}{pq\frac{n_An_B}{n_A+n_B}\frac{\left(n_A+n_B\right)^2}{n_B^2n_A^2}}=\frac{\left(p_B-p_A\right)^2}{pq\frac{n_A+n_B}{n_An_B}}=\frac{\left(p_B-p_A\right)^2}{pq\left(\frac{1}{n_B}+\frac{1}{n_A}\right)}\\ &=\frac{\left(p_B-p_A\right)^2}{p(1-p)\left(\frac{1}{n_A}+\frac{1}{n_B}\right)}=Z^2 \end{aligned} \tag{14.1.842}$$

And this finish the proof of the equivalence between the 2-sample Wald's Z-test and the Rao score test over logistic regression with a single binary predictor.

A practical and numerical example is provided in our `R` companion book.

14.1.12.6.2 Multinomial Logistic regression

"Multinomial logistic regression" is a classification method that generalizes logistic regression to multiclass problems, i.e. with more than two possible discrete outcomes. That is, it is a model that is used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable, given a set of independent variables (which may be real-valued, binary-valued, categorical-valued, etc.).

Multinomial logistic regression is known by a variety of other names, "including polytomous LR", "multiclass LR", "softmax regression", "multinomial logit", the "maximum entropy (MaxEnt) classifier", and the "conditional maximum entropy model".

The basic setup is the same as in logistic regression, the only difference being that the dependent variables are categorical rather than binary, i.e. there are K possible outcomes rather than just two. The following description is somewhat shortened.

As in other forms of linear regression, multinomial logistic regression uses a linear predictor function $f(k,i)$ to predict the probability that observation i has outcome k, of the following form:

$$f(k,i)=\beta_{0,k}+\beta_{1,k}x_{1,i}+\beta_{2,k}x_{2,i}+\cdots+\beta_{M,k}x_{M,i} \tag{14.1.843}$$

where $\beta_{m,k}$ is a regression coefficient associated with the mth explanatory variable and the kth outcome. As it is common, the regression coefficients and explanatory variables are normally grouped into vectors of size $M+1$, so that the predictor function can be written more compactly:

$$f(k,i)=\vec{\beta}_k\circ\vec{x}_i \tag{14.1.844}$$

where $\vec{\beta}_k$ is the set of regression coefficients associated with outcome k, and $\vec{x}_i$ (a row vector) is the set of explanatory variables associated with observation i.

One fairly simple way to arrive at the multinomial logit model is to imagine, for K possible outcomes, running $K-1$ independent binary logistic regression models, in which one outcome is chosen as a "pivot" and then the other $K-1$ outcomes are separately regressed against the pivot outcome. This would proceed as follows, if outcome K (the last outcome) is chosen as the pivot:

$$\begin{aligned}\ln\left(\dfrac{P(Y_i=1)}{P(Y_i=K)}\right)&=\vec{\beta}_1\circ\vec{x}_i\\ \ln\left(\dfrac{P(Y_i=2)}{P(Y_i=K)}\right)&=\vec{\beta}_2\circ\vec{x}_i\\ &\cdots\\ \ln\left(\dfrac{P(Y_i=K-1)}{P(Y_i=K)}\right)&=\vec{\beta}_{K-1}\circ\vec{x}_i\end{aligned} \tag{14.1.845}$$

Note that we have introduced separate sets of regression coefficients, one for each possible outcome.

That latter relation is also often denoted:

$$\ln\left(\frac{\pi_i^{(j)}}{\pi_i^0}\right)=\alpha^{(j)}+\beta_1^{(j)}X_{1i}+\ldots+\beta_k^{(j)}X_{ki} \tag{14.1.846}$$

for each $j=1,\ldots,K-1$.

If we exponentiate both sides, and solve for the probabilities, we get:

$$\begin{aligned}P(Y_i=1)&=P(Y_i=K)e^{\vec{\beta}_1\circ\vec{x}_i}\\P(Y_i=2)&=P(Y_i=K)e^{\vec{\beta}_2\circ\vec{x}_i}\\&\ldots\\P(Y_i=K-1)&=P(Y_i=K)e^{\vec{\beta}_{K-1}\circ\vec{x}_i}\end{aligned} \tag{14.1.847}$$

Using the fact that all K of the probabilities must sum to one, we find:

$$P(Y_i=K)=1-\sum_{k=1}^{K-1}P(Y_i=k)=1-\sum_{k=1}^{K-1}\mathrm{Pr}(Y_i=K)e^{\vec{\beta}_k\circ\vec{x}_i}\Rightarrow P(Y_i=K)=\frac{1}{1+\displaystyle\sum_{k=1}^{K-1}e^{\vec{\beta}_k\circ\vec{x}_i}} \tag{14.1.848}$$

That latter relation is also often denoted (notation that we will use further below):

$$P(Y_i=J)=\pi_{iJ}=\frac{1}{1+\displaystyle\sum_{j=1}^{J-1}e^{\vec{\beta}_j\circ\vec{x}_i}} \tag{14.1.849}$$

Re-injecting in the previous relations sets, we get:

$$\begin{aligned}P(Y_i=1)&=\frac{e^{\vec{\beta}_1\circ\vec{x}_i}}{1+\displaystyle\sum_{k=1}^{K-1}e^{\vec{\beta}_k\circ\vec{x}_i}}\\P(Y_i=2)&=\frac{e^{\vec{\beta}_2\circ\vec{x}_i}}{1+\displaystyle\sum_{k=1}^{K-1}e^{\vec{\beta}_k\circ\vec{x}_i}}\\&\ldots\\P(Y_i=K-1)&=\frac{e^{\vec{\beta}_{K-1}\circ\vec{x}_i}}{1+\displaystyle\sum_{k=1}^{K-1}e^{\vec{\beta}_k\circ\vec{x}_i}}\end{aligned} \tag{14.1.850}$$

Hence:

$$\boxed{P(Y_i=k)=\frac{e^{\vec{\beta}_k\circ\vec{x}_i}}{1+\displaystyle\sum_{k=1}^{K}e^{\vec{\beta}_k\circ\vec{x}_i}}=\frac{e^{\vec{\beta}_k^T\vec{x}_i}}{1+\displaystyle\sum_{k=1}^{K}e^{\vec{\beta}_k^T\vec{x}_i}}} \tag{14.1.851}$$

Often denoted (as the sum in the denominator is a constant[26]):

$$P(Y_i = k) = \frac{e^{\vec{\beta}_k \circ \vec{x}_i}}{1+Z} \tag{14.1.852}$$

The fact that we run multiple regressions reveals why the model relies on the assumption of independence of irrelevant alternatives!

Note that the prior-previous relation is also denoted:

$$P(Y_i = j) = \pi_{ij} = \frac{e^{\vec{\beta}_j \circ \vec{x}_i}}{1+\displaystyle\sum_{k=1}^{K} e^{\vec{\beta}_j \circ \vec{X}_i}} \tag{14.1.853}$$

So we have:

$$\boxed{\begin{aligned} P(Y_i = j) = \pi_{ij} &= \frac{e^{\vec{\beta}_j \circ \vec{x}_i}}{1+\displaystyle\sum_{k=1}^{K} e^{\vec{\beta}_j \circ \vec{x}_i}} \quad j<J \\ P(Y_i = J) = \pi_{iJ} &= \frac{1}{1+\displaystyle\sum_{j=1}^{J-1} e^{\vec{\beta}_j \circ \vec{x}_i}} \end{aligned}} \tag{14.1.854}$$

Remark

Note that when $K=2$ the multinomial and logistic regression models become one and the same. Also for information, the softmax function is used in multi-class classification neural networks as activation function as we will see further below during our study of neural networks (page 6951).

The following function:

$$\text{softmax}(k, x_1, \ldots, x_n) = \frac{e^{x_k}}{\displaystyle\sum_{i=1}^{n} e^{x_i}} \tag{14.1.855}$$

is referred to as the "softmax function". The reason is that the effect of exponentiating the values $x_1, \ldots, x_n$ is to exaggerate the differences between them. As a result $\text{softmax}(k, x_1, \ldots, x_n)$ will return a value close to 0 whenever x_k is significantly less than the maximum of all the values, and will return a value close to 1 when applied to the maximum value, unless it is extremely close to the next-largest value.

When using multinomial logistic regression, one category of the dependent variable is chosen as the reference category. Separate odds ratios are determined for all independent variables for each category of the dependent variable with the exception of the reference category, which is omitted from the analysis. The exponential beta coefficient represents the change in the odds of the dependent variable being in a particular category vis-a-vis the reference category, associated with a one unit change of the corresponding independent variable.

---

[26] However, it is definitely not constant with respect to the explanatory variables, or crucially, with respect to the unknown regression coefficients $\vec{\beta}_k$, which we will need to determine through some sort of optimization procedure.

Following the request of a reader, here is the maximum likelihood approach of the logistic regression.

Let us recall that the multinomial distribution is given (see section Statistics page 679) by:

$$\mathcal{M}(\vec{k},\vec{p},m)=\frac{n!}{\prod\limits_{i=1}^{m}k_i!}\prod_{i=1}^{m}p_i^{k_i}=\frac{\left(\sum\limits_{i=1}^{m}k_i\right)!}{\prod\limits_{i=1}^{m}k_i!}\prod_{i=1}^{m}p_i^{k_i}=\binom{n}{k_1,k_2,\ldots,k_m}\prod_{i=1}^{m}p_i^{k_i} \qquad (14.1.856)$$

In other words if there are m random variables, i.e. X_i, $i\in[1,n]$, where X_i represents the number of occurrences of item i in a choice of n items, with entry i in the vector of probabilities $\vec{P}$, P_i giving the probability of drawing item i. The probability of selecting k_1 of item 1... k_m of item m is then given by $\mathcal{M}$ above!

Remark

Remember that if $m=2$ then the above relation becomes:

$$\mathcal{M}=\frac{(k_1+k_2)!}{k_1!k_2!}p_1^{k_1}p_2^{k_2} \qquad (14.1.857)$$

But as $k_1+k_2=n$ then:

$$\mathcal{M}=\frac{n!}{k_1!(n-k_1)!}p_1^{k_1}p_2^{n-k_1} \qquad (14.1.858)$$

and as $p_1+p_2=1$ we finally have:

$$\mathcal{M}=\frac{n!}{k_1!(n-k_1)!}p_1^{k_1}(1-p_1)^{n-k_1} \qquad (14.1.859)$$

We recognize here the binomial distribution!

Let us denote the joint probability density (likelihood) function for the multinomial logistic function as:

$$L(D,\vec{\beta})=\prod_{i=1}^{N}\left[\frac{n_i!}{\prod_{j=1}^{J}y_{ij}!}\prod_{j=1}^{J}\pi_{ij}^{y_{ij}}\right] \qquad (14.1.860)$$

Since we want to maximize the above relation with respect to $\vec{\beta}$, the factorial terms that do not contain any of the π_{ij} terms can be treated as constants. Thus, the kernel of the log likelihood function for multinomial logistic regression models is:

$$L(D,\vec{\beta})\propto\prod_{i=1}^{N}\prod_{j=1}^{J}\pi_{ij}^{y_{ij}} \qquad (14.1.861)$$

Replacing the Jth terms, the previous relations becomes:

$$\begin{aligned}L(D,\vec{\beta}) &\propto \prod_{i=1}^{N}\prod_{j=1}^{J-1}\pi_{ij}^{y_{ij}}\pi_{iJ}^{n_i-\sum_{j=1}^{J-1}y_{ij}}\\ &=\prod_{i=1}^{N}\prod_{j=1}^{J-1}\pi_{ij}^{y_{ij}}\frac{\pi_{iJ}^{n_i}}{\pi_{iJ}^{\sum_{j=1}^{J-1}y_{ij}}}\\ &=\prod_{i=1}^{N}\prod_{j=1}^{J-1}\pi_{ij}^{y_{ij}}\frac{\pi_{iJ}^{n_i}}{\prod_{j=1}^{J-1}\pi_{iJ}^{y_{ij}}}\end{aligned} \tag{14.1.862}$$

Since $a^{x+y}=a^x a^y$, the sum in the exponent in the denominator of the last term becomes a product over the first $J-1$ terms of j. Continue by grouping together the terms that are raised to the y_{ij} power for each j up to $J-1$:

$$L(D,\vec{\beta}) \propto \prod_{i=1}^{N}\prod_{j=1}^{J-1}\left(\frac{\pi_{ij}}{\pi_{iJ}}\right)^{y_{ij}}\pi_{iJ}^{n_i} \tag{14.1.863}$$

Now we substitute for π_{ij} and π_{iJ} the both relations that we have derived earlier:

$$\begin{aligned}P(Y_i=j)=\pi_{ij}&=\frac{e^{\vec{\beta}_j\circ\vec{x}_i}}{1+\sum\limits_{k=1}^{K}e^{\vec{\beta}_j\circ\vec{X}_i}} \qquad j<J\\ P(Y_i=J)=\pi_{iJ}&=\frac{1}{1+\sum\limits_{j=1}^{J-1}e^{\vec{\beta}_j\circ\vec{X}_i}}\end{aligned} \tag{14.1.864}$$

This gives us then:

$$\begin{aligned}L(D,\vec{\beta}) &\propto \prod_{i=1}^{N}\prod_{j=1}^{J-1}\left(e^{\vec{\beta}_j\circ\vec{X}_i}\right)^{y_{ij}}\left(\frac{1}{1+\sum\limits_{j=1}^{J-1}e^{\vec{\beta}_j\circ\vec{X}_i}}\right)^{n_i}\\ &=\prod_{i=1}^{N}\prod_{j=1}^{J-1}e^{y_{ij}\vec{\beta}_j\circ\vec{X}_i}\left(1+\sum_{j=1}^{J-1}e^{\vec{\beta}_j\circ\vec{X}_i}\right)^{-n_i}\end{aligned} \tag{14.1.865}$$

Taking the logarithm of the above relation gives us the log-likelihood function for the multinomial logistic regression model:

$$\ln(L(D,\vec{\beta}))\propto\sum_{i=1}^{N}\sum_{j=1}^{J-1}\left(y_{ij}\vec{\beta}_j\circ\vec{X}_i\right)-n_i\ln\left(1+\sum_{j=1}^{J-1}e^{\vec{\beta}_j\circ\vec{X}_i}\right) \tag{14.1.866}$$

14.1.12.6.2.1 F_1 score

Now that we have introduced a well known multiclass classifier, it is time to introduce the "F_1 score" (also named "F-score" or "F-measure") which is a measure of a classifier quality (among many others empirical quality indicators...). It considers both the precision P and the recall R of the test to compute the score. To introduce these two indicators, let us consider for companion example, the following generic three-class ($n=3$) confusion matrix $\mathcal{C}$ for a total of N records to classify:

		Predicted		
		Class 1	**Class** 2	**Class** 3
Actual	**Class** 1	$\mathcal{C}_{11}$	$\mathcal{C}_{12}$	$\mathcal{C}_{13}$
	Class 2	$\mathcal{C}_{21}$	$\mathcal{C}_{22}$	$\mathcal{C}_{23}$
	Class 3	$\mathcal{C}_{31}$	$\mathcal{C}_{32}$	$\mathcal{C}_{33}$

Table 14.14: **A typical three class confusion matrix**

Let us recall first that the "accuracy" is given by:

$$\text{ACC}=\frac{\text{tr}(\mathcal{C})}{N} \tag{14.1.867}$$

The "precision" P_i for a multiclass classifier (of n classes) is defined as the fraction of correct predictions for a certain class[27]!:

$$P_i=\frac{\mathcal{C}_{ii}}{\sum\limits_j \mathcal{C}_{ij}} \tag{14.1.868}$$

and the "recall" R_i for a multiclass classifier (of n classes) is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive):

$$R_i=\frac{\mathcal{C}_{ii}}{\sum\limits_j \mathcal{C}_{ji}} \tag{14.1.869}$$

This can nicely be illustrated as following:

---

[27] For a practical example, see our `R` companion book on the F-score.

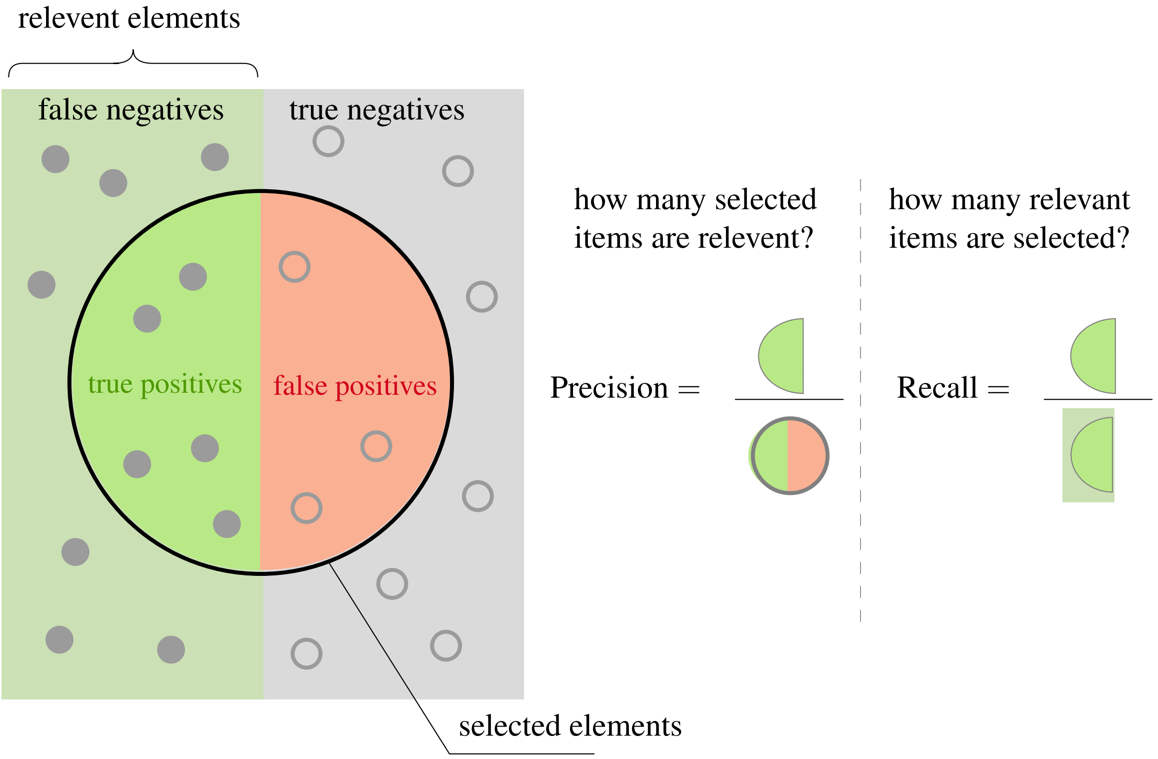

relevent elements
false negatives
true negatives
true positives
false positives
selected elements
how many selected items are relevent?
how many relevant items are selected?
Precision =
Recall =

Figure 14.24: **Precision and Recall (source: Wikipedia)**

The F_1 score is the harmonic average of the precision and recall (for a given class i) where a F_1 score reaches its best value at 1 (perfect precision and recall) and worst at 0 (it will seems obvious to the reader with the details given further below).

The F-score is also used in Machine Learning as a suitable measure of models tested with unbalanced datasets. Note, however, that the F-measures do not take the true negatives into account, and that measures such as the Matthews correlation coefficient, Informedness or Cohen's kappa may be preferable to assess the performance of a binary classifier.

David Hand and others criticize the widespread use of the F-score since it gives equal importance to precision and recall. In practice, different types of miss-classifications incur different costs. In other words, the relative importance of precision and recall is an aspect of the problem.

Remark

The text that follows is completely inspired and copied from the excellent publication of Professor Yutaka Sasaki (see [479]) with it's kind authorization.

The F-measure is defined as a harmonic mean (see section Statistics page 563) of precision P and recall R:

$$F = \frac{2PR}{P+R} \quad (14.1.870)$$

If you are satisfied with this definition and need no further information, that's it. However, if you are deeply interested in the definition of the F-measure, you should recap the definitions of the arithmetic and harmonic means.

The arithmetic mean μ_a (an average in a usual sense) and the harmonic mean μ_h are defined for recall as follows:

$$\begin{aligned}\mu_a &= \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1+x_2+\ldots+x_n)\\ \mu_h &= \frac{n}{\displaystyle\sum_{i=1}^{n}\frac{1}{x_i}} = \frac{n}{\dfrac{1}{x_1}+\ldots+\dfrac{1}{x_n}}\end{aligned} \tag{14.1.871}$$

When $x_1 = P$ and $x_2 = R$, μ_a and μ_h will be:

$$\begin{aligned}\mu_a &= \frac{1}{2}(P+R)\\ \mu_h &= \frac{2}{\dfrac{1}{P}+\dfrac{1}{R}} = \frac{2}{\dfrac{P+R}{PR}} = \frac{2PR}{P+R}\end{aligned} \tag{14.1.872}$$

Let us recall that harmonic mean is more intuitive than the arithmetic mean when computing a mean of ratios. Indeed, suppose that you have a finger print recognition system and its precision and recall be 1.0 and 0.2, respectively. Intuitively, the total performance of the system should be very low because the system covers only 20% of the registered finger prints, which means it is almost useless.

The arithmetic mean of 1 and 0.2 is 0.6 whereas the harmonic mean of them is:

$$\mu_h = \frac{2\cdot 1\cdot\dfrac{2}{10}}{1+\dfrac{2}{10}} = \frac{4}{12} = \frac{1}{3} \tag{14.1.873}$$

As you see in this example, the harmonic mean ($0.\bar{3}$) is a more reasonable score than the arithmetic mean (0.6).

Some researchers named the definition of the F-measure given earlier above the "F_1-measure". What is stands the 1 of F_1 for?

The full definition of the F-measure is given as follows:

$$F_\beta = \frac{(\beta^2+1)\,PR}{\beta^2 P + R} \qquad (0 \leq \beta \leq +\infty) \tag{14.1.874}$$

β is a parameter that controls a balance between P and R. When $\beta = 1$, F_1 comes to be equivalent to the harmonic mean of P and R. If $\beta > 1$, F becomes more recall-oriented and if $\beta < 1$, it becomes more precision-oriented, e.g., $F_0 = P$.

However it seems that the real original definition comes from the E "effectiveness function" given by:

$$E = 1 - \frac{1}{\alpha\dfrac{1}{P}+(1-\alpha)\dfrac{1}{R}} \tag{14.1.875}$$

where:

$$\alpha = \frac{1}{\beta^2+1} \tag{14.1.876}$$

Let's remove α using β:

$$\begin{aligned} E &= 1 - \frac{1}{\frac{1}{\beta^2+1}\frac{1}{P} + \left(1 - \frac{1}{\beta^2+1}\right)\frac{1}{R}} \\ &= 1 - \frac{PR}{\frac{1}{\beta^2+1}R + \frac{\beta^2+1-1}{\beta^2+1}P} \\ &= 1 - \frac{(\beta^2+1)\,PR}{R+\beta^2 P} \end{aligned} \tag{14.1.877}$$

Now you see that:

$$E = 1 - F_\beta \tag{14.1.878}$$

Note that F rises if R or P gets better whereas E becomes small if R or P improves. This seems the reason why F is more commonly used than E.

Some people use α as a parameter of F:

$$F_\alpha = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} \qquad (0 \leq \alpha \leq 1) \tag{14.1.879}$$

There is nothing wrong with this definition of F but use of this definition might cause an unnecessary confusion because $F_{\alpha=0.5} = F_{\beta=1}$. An attention is needed that the commonly used notation F_1 means $F_{\beta=1}$, not $F_{\alpha=1}$.

Still, some of you are not sure why β^2 is used instead of β in $\alpha = \frac{1}{\beta^2+1}$ let my try an explanation of the reason.

β is the parameter that controls the weighting between P and R. Formally, is defined as follows:

$$\beta = \frac{R}{P} \tag{14.1.880}$$

where:

$$\frac{\partial E}{\partial P} = \frac{\partial E}{\partial R} \tag{14.1.881}$$

The motivation behind this condition is that at the point where the gradients of E with respect to P and R are equal, the ratio of R against P should be a desired ratio β.

Please recall that E is defined as follows:

$$\begin{aligned} E &= 1 - \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} \\ &= 1 - \frac{PR}{\alpha R + (1-\alpha)P} \end{aligned} \tag{14.1.882}$$

Now we calculate $\partial E/\partial P$ and $\partial E/\partial R$. By the quotient rule on the derivative of a composite function:

$$(f/g)'=\frac{f'g-fg'}{g^2} \tag{14.1.883}$$

For conciseness, let;

$$g=\alpha R+(1-\alpha)P \tag{14.1.884}$$

Then:

$$\begin{aligned}\frac{\partial E}{\partial P}&=-\frac{R(\alpha R+(1-\alpha)P)-PR(1-\alpha)}{g^2}\\ \frac{\partial E}{\partial R}&=-\frac{P(\alpha R+(1-\alpha)P)-PR\alpha}{g^2}\end{aligned} \tag{14.1.885}$$

Then:

$$\frac{\partial E}{\partial P}=\frac{\partial E}{\partial R} \tag{14.1.886}$$

is equivalent to:

$$R(\alpha R+(1-\alpha)P)-PR(1-\alpha)=P(\alpha R+(1-\alpha)P)-PR\alpha \tag{14.1.887}$$

which can be simplified to:

$$\alpha R^2=(1-\alpha)P^2 \tag{14.1.888}$$

As $\beta=R/P$, we can replace R with βP^2:

$$\begin{aligned}&\alpha\beta^2P^2=(1-\alpha)P^2\\ \Rightarrow&\alpha\beta^2=1-\alpha\\ \Rightarrow&\alpha\left(\beta^2+1\right)=1\\ \Rightarrow&\alpha=\frac{1}{\beta^2+1}\end{aligned} \tag{14.1.889}$$

Remark

The per-class metrics above (precision P_i, recall R_i and $F_{1,i}$-score) can be averaged over all the classes n resulting in "macro-averaged precision" (MP), macro-averaged recall" (MR) and "macro-averaged F_1" (MF$_1$):

$$\text{MP}=\frac{\sum_i P_i}{n}\qquad \text{MR}=\frac{\sum_i R_i}{n}\qquad \text{MF}_1=\frac{\sum_i F_{1,i}}{n} \tag{14.1.890}$$

When the instances are not uniformly distributed over the classes, it is useful to look at the performance of a multi-class classifier with respect to one class at a time before averaging the metrics. The idea is to compute the "one-vs-all confusion matrix for each class" (3 matrices in the case of our companion example here!). You can think of the problem as three binary classification tasks where one class is considered the positive class while the combination of all the other classes make up the negative class.

Summing up the values of all these matrices results in a unique confusion matrix and allows us to compute weighted metrics such as "average accuracy" and "micro-averaged R, P and F_1 metrics". Similar to the overall accuracy, the average accuracy is defined as the fraction of correctly classified instances in the sum of one-vs-all matrices matrix!

14.1.12.6.3 Ordinal Logistic regression

In statistics, ordinal regression (also named "ordinal classification") is a type of regression analysis used for predicting an ordinal variable, i.e. a variable whose value exists on an arbitrary scale where only the relative ordering between different values is significant.

To introduce the "ordinal logistic regression", let us recall that given a response variable Y having K ordered categories $j=1,2,\ldots,K$, with probabilities:

$$P(Y_i=j)=\pi_i^{(j)} \tag{14.1.891}$$

the multinomial logistic model, considered the $K-1$ ratios:

$$\frac{P(Y_i=j)}{P(Y_i=K)}=\frac{\pi_i^{(j)}}{\pi_i^{(1)}} \tag{14.1.892}$$

for $j=1,2,\ldots,K-1$.

Now we will consider the $K-1$ cumulative probabilities instead!

$$\gamma_i^{(j)}=P(Y_i\leq j)=\pi_i^{(1)}+\ldots+\pi_i^{(j)} \tag{14.1.893}$$

for $j=1,\ldots,K-1$ and write down a model for each of them.

Remark

Note that $\gamma_i^{(K)}=P(Y_i\leq K)=1$ always, so it need not be modelled!

The following holds:

$$\ln\left(\frac{\gamma_i^{(j)}}{1-\gamma_i^{(j)}}\right)=\ln\left(\frac{P\left(Y_i\leq j\right)}{P\left(Y_i>j\right)}\right)=\text{logit}(P(Y_i\leq j)=\alpha^{(j)}-(\beta_1x_{1i}+\cdots+\beta_nx_{ni}) \tag{14.1.894}$$

In other words, the ordinal logistic model considers as for the multinomial version, a set of dichotomies, one for each possible cutoff of the response categories into two sets, of "high" and "low" responses. But this concept of "high" and "low" at the opposite of multinomial regression is meaningful only if the categories of Y do have an ordering!

Remark

In the software Strata and in the `polr` package of the `R` software the ordinal logistic regression model is parametrized as:

$$\text{logit}(P(Y\leq j))=\alpha^{(j)}-(\eta_1x_{1i}+\cdots+\eta_nx_{ni}) \tag{14.1.895}$$

instead of:

$$\text{logit}(P(Y \leq j)) = \alpha^{(j)} + (\beta_1 x_{1i} + \cdots + \beta_n x_{ni}) \tag{14.1.896}$$

where $\eta_i = -\beta_1$ hence the notation above with the minus signs.

If we consider the ordered labels: Strongly disagree (SD), Disagree (D), Agree (A) and Strongly Agree (SA), the cutoffs are:

- SD vs (D, A or SA)
- (SD or D) vs (A or SA)
- (SD, D or A) vs SA

A binary logistic model is then defined for the log-odds of each of these cuts!

The model for the cumulative probabilities is (we omit the index i without loss of generality but just to simplify the notations!):

$$\gamma^{(j)} = P(Y \leq j) = \frac{\exp\left[\alpha^{(j)} - (\beta_1 x_1 + \cdots + \beta_k x_k)\right]}{1 + \exp\left[\alpha^{(j)} - (\beta_1 x_1 + \cdots + \beta_k x_k)\right]} \tag{14.1.897}$$

The intercept terms must be $\alpha^{(1)} < \alpha^{(2)} < \ldots < \alpha^{(K-1)}$, to guarantee that $\gamma^{(1)} < \gamma^{(2)} < \ldots < \gamma^{(K-1)}$. The $\beta_1, \beta_2, \ldots, \beta_n$ are assumed to be the same for each value of j.

From this the practitioner must remember that:

- There is thus only one set of regression coefficients, not $K-1$ as in a multinomial logistic model
- The curves for $\gamma^{(1)}, \gamma^{(2)}, \ldots, \gamma^{(K-1)}$ are translated (i.e. "proportional") as seen in the plots below
- This is the assumption of "proportional odds". This is why ordinal logistic model is also known as the "proportional odds model" or "parallel regression model"!

This is good for the parsimony of the model, because it means that the effect of an explanatory variable on the ordinal response is described by one parameter! However, it is also a restriction on the flexibility of the model, which may or may not be adequate for the data.

In other words, one of the assumptions underlying ordinal logistic (and ordinal probit) regression is that the relationship between each pair of outcome groups is the same. In other words, ordinal logistic regression assumes that the coefficients that describe the relationship between, say, the lowest versus all higher categories of the response variable are the same as those that describe the relationship between the next lowest category and all higher categories, etc. Because the relationship between all pairs of groups is the same, there is only one set of coefficients! If this was not the case, we would need different sets of coefficients in the model to describe the relationship between each pair of outcome groups.

Thus, in order to assess the appropriateness of our model, we need to evaluate whether the proportional odds assumption is tenable. Statistical tests to do this are available in some software packages. However, it seems that these tests have been criticized for having a tendency to reject the null hypothesis (that the sets of coefficients are the same), and hence, indicate that there the parallel slopes assumption does not hold, in cases where the assumption does hold (see [232] page 335). However, Harrell does recommend a graphical method for assessing the parallel slopes assumption. The values displayed in his proposed graph are essentially (linear) predictions from a logit model, used to model the probability that y is greater than or equal to a given value (for each level of y), using one predictor (x) variable at a time.

✍**Example:**

Suppose the proportions of members of the statistical population who would answer "poor", "fair", "good", "very good", and "excellent" are respectively p_1, p_2, p_3, p_4, p_5. Then the logarithms of the odds (not the logarithms of the probabilities) of answering in certain ways are:

$$\begin{array}{rcc}
\text{poor,} & \ln\left(\dfrac{p_1}{p_2+p_3+p_4+p_5}\right), & 0 \\
\text{poor or fair,} & \ln\left(\dfrac{p_1+p_2}{p_3+p_4+p_5}\right), & 1 \\
\text{poor, fair, or good,} & \ln\left(\dfrac{p_1+p_2+p_3}{p_4+p_5}\right), & 2 \\
\text{poor, fair, good, or very good,} & \ln\left(\dfrac{p_1+p_2+p_3+p_4}{p_5}\right), & 3
\end{array} \tag{14.1.898}$$

The proportional odds assumption is that the number added to each of these logarithms to get the next is the same in every case. In other words, these logarithms form an arithmetic sequence. The model states that the number in the last column of the table - the number of times that the logarithm must be added - is some linear combination of the other observed variables.

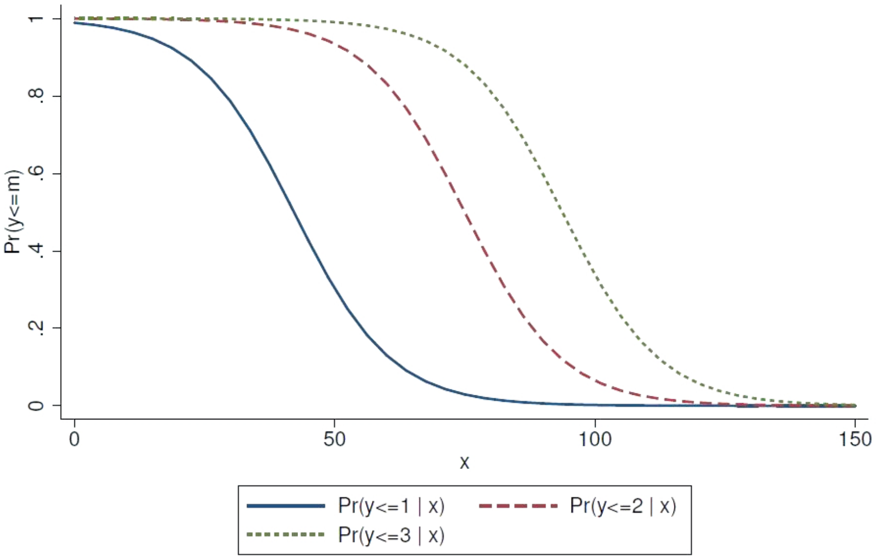
The probabilities of individual categories are:

$$\begin{aligned}
P(Y=1) &= \gamma^{(1)} = \frac{\exp\left[\alpha^{(1)}-(\beta_1 x_1+\cdots+\beta_n x_n)\right]}{1+\exp\left[\alpha^{(1)}-(\beta_1 x_1+\cdots+\beta_n x_n)\right]} \\
P(Y=j) &= \gamma^{(j)}-\gamma^{(j-1)} = \frac{\exp\left[\alpha^{(j)}-(\beta_1 x_1+\cdots+\beta_k x_k)\right]}{1+\exp\left[\alpha^{(j)}-(\beta_1 x_1+\cdots+\beta_n x_n)\right]} \\
&\quad - \frac{\exp\left[\alpha^{(j-1)}-(\beta_1 x_1+\cdots+\beta_n x_n)\right]}{1+\exp\left[\alpha^{(j-1)}-(\beta_1 x_1+\cdots+\beta_n x_n)\right]}
\end{aligned} \tag{14.1.899}$$

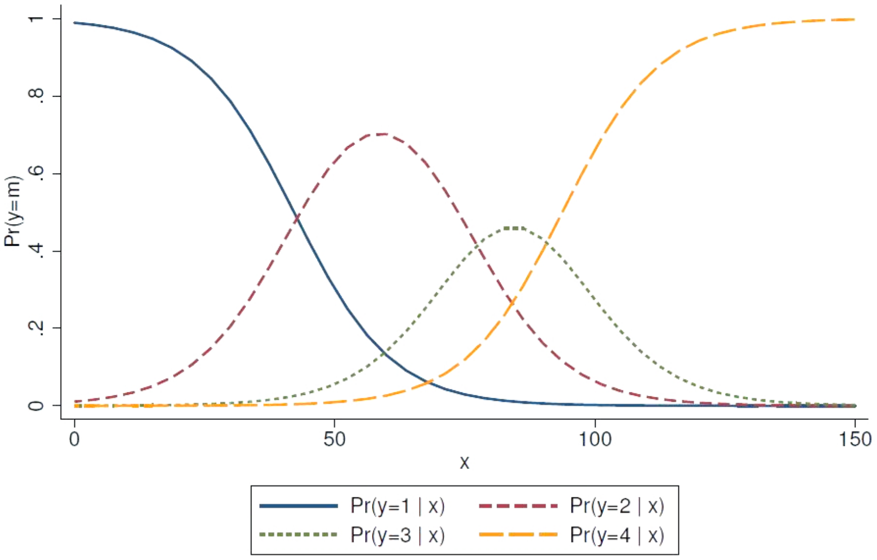
for $j = 2, \ldots, K-1$ and:

$$P(Y=K) = 1-\gamma^{(K-1)} = \frac{\exp\left[\alpha^{(K-1)}-(\beta_1 x_1+\cdots+\beta_n x_n)\right]}{1+\exp\left[\alpha^{(K-1)}-(\beta_1 x_1+\cdots+\beta_n x_n)\right]}$$

Illustrated below with plots for a case with $K = 4$ categories, first the cumulative probabilities:



and the probabilities of individual categories:



Everything here is unchanged from binary logistic models:

- Parameters are estimated using maximum likelihood estimation
- Hypotheses of interest are typically of the form $\beta_j = 0$, for one or more coefficients β_j
- Exponentiated coefficients are interpreted as partial odds ratios for being in the higher rather than the lower half of the dichotomy
- Wald tests, likelihood ratio tests and confidence intervals are the same defined and used as before

✍**Example:**

In Stata and `R` (package `polr`) the ordinal logistic regression model is parametrized a:

$$\text{logit}(P(Y \leq j)) = \beta_{j0} - \eta_1 x_1 - \ldots - \eta_p x_p$$

where for recall $\eta_i = -\beta_i$.

Suppose we want to see whether a binary predictor parental education (*pared*) predicts an ordinal outcome of students who are unlikely, somewhat likely and very likely to apply to a college (*apply*).

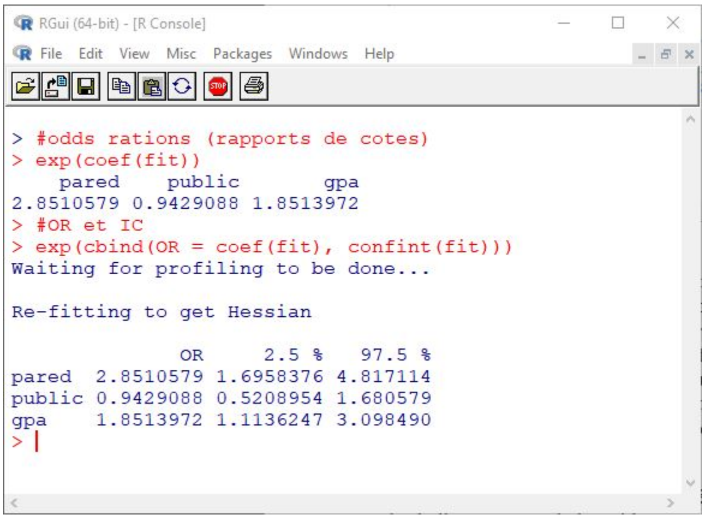
Due to the parallel odds assumption, even though we have categories, the coefficient of parental education (*pared*) stays the same across the two categories. Then the two equations for pared $= 1$ and pared $= 0$ are:

$$\begin{aligned} \text{logit}\left(P(Y \leq j)|x_1 = 1\right) &= \beta_{j0} - \eta_1 \\ \text{logit}\left(P(Y \leq j)|x_1 = 0\right) &= \beta_{j0} \end{aligned}$$

Then:

$$\text{logit}\left(P(Y \leq j)|x_1 = 1\right) - \text{logit}\left(P(Y \leq j)|x_1 = 0\right) = -\eta_1$$

To run an ordinal logistic regression in `R` we run (see our `R` companion book for more details):



```
> #odds rations (rapports de cotes)
> exp(coef(fit))
    pared    public       gpa
2.8510579 0.9429088 1.8513972
> #OR et IC
> exp(cbind(OR = coef(fit), confint(fit)))
Waiting for profiling to be done...

Re-fitting to get Hessian

              OR     2.5 %   97.5 %
pared  2.8510579 1.6958376 4.817114
public 0.9429088 0.5208954 1.680579
gpa    1.8513972 1.1136247 3.098490
> |
```

The output shows that for students whose parents attended college, the log odds of being unlikely to college (versus somewhat or very likely) is actually $-\hat{\eta}_1 = -1.13$ or 1.13 points lower than the students who did not attend college. Recall that $-\eta_i = \beta_i$ for $j = 1, 2$ only since logit $(P(Y \leq 3))$ is undefined. So the formulations for the first and second category becomes:

$$\begin{aligned} \text{logit}(P(Y \leq 1)) &= 0.377 - 1.13x_1 \\ \text{logit}(P(Y \leq 2)) &= 2.452 - 1.13x_1 \end{aligned}$$

To see the connection between the proportional odds assumption, we exponentiate both sides of the equations above and we use the property that $\log(b) - \log(a) = \log(b/a)$ to calculate the odds of *pared* for each level of *apply*:

$$\frac{P\left(Y \leq 1|x_1=1\right)}{P\left(Y > 1|x_1=1\right)} = \frac{e^{0.377}}{e^{1.13}}$$
$$\frac{P\left(Y \leq 1|x_1=0\right)}{P\left(Y > 1|x_1=0\right)} = e^{0.377}$$
$$\frac{P\left(Y \leq 2|x_1=1\right)}{P\left(Y > 2|x_1=1\right)} = \frac{e^{2.45}}{e^{1.13}}$$
$$\frac{P\left(Y \leq 2|x_1=0\right)}{P\left(Y > 2|x_1=0\right)} = e^{2.45}$$

From the odds of each level of pared, we can calculate the odds ratio of *pared* for each level of *apply*:

$$\frac{\frac{P\left(Y \leq 1|x_1=1\right)}{P\left(Y > 1|x_1=1\right)}}{\frac{P\left(Y \leq 1|x_1=0\right)}{P\left(Y > 1|x_1=0\right)}} = \frac{1}{e^{1.13}} = e^{-1.13}$$
$$\frac{\frac{P\left(Y \leq 2|x_1=1\right)}{P\left(Y > 2|x_1=1\right)}}{\frac{P\left(Y \leq 2|x_1=0\right)}{P\left(Y > 2|x_1=0\right)}} = \frac{1}{e^{1.13}} = e^{-1.13}$$

The proportional odds assumption ensures that the odds ratios across all $J-1$ categories are the same. In our example, the proportional odds assumption means that the odds of being unlikely versus somewhat or very likely to apply $(j=1)$ is equal to the odds of being unlikely and somewhat likely versus very likely to apply $(j=2)$.

The proportional odds assumption is not simply that the odds are the same but that the odds ratios are the same across categories. These odds ratios can be derived by exponentiating the coefficients (in the log-odds metric), but the interpretation is a bit unexpected. Recall that the coefficient $-\eta_1$ represents a one unit change in the log odds of applying for students whose parents went to college versus parents who did not:

$$\text{logit}(P(Y \leq j)|x_1=1) - \text{logit}(P(Y \leq j)|x_1=0) = -\eta_1$$

Since the exponent is the inverse function of the log, we can simply exponentiate both sides of this equation, and by using the property that $\log(b) - \log(a) = \log(b/a)$:

$$\frac{\frac{P(Y \leq j|x_1=1)}{P(Y > j|x_1=1)}}{\frac{P(Y \leq j|x_1=0)}{P(Y > j|x_1=0)}} = e^{-\eta_1}$$

For simplicity of notation and by the proportional odds assumption, let:

$$\frac{P(Y \leq j|x_1=1)}{P(Y > j|x_1=1)} = \frac{p_1}{1-p_1} \quad \text{and} \quad \frac{P(Y \leq j|x_1=0)}{P(Y > j|x_1=0)} = \frac{p_0}{1-p_0}$$

Then the odds ratio is defined as:

$$\frac{p_1/(1-p_1)}{p_0/(1-p_0)}=e^{-\eta_1}$$

In our example, $e^{-1.13}=0.323$, which means that students whose parents attended college have a 67.6% lower odds of being less likely to apply to college.

But you won't get this value with `R` and Stata (they retrieve the value 3.087 instead). Let's see why!

Since:

$$e^{-\eta_1}=\frac{1}{e^{\eta_1}}$$

From the output $\eta=1.13$, which means that 3.086 come from actually:

$$e^{\eta_1}=\frac{p_0/(1-p_0)}{p_1/(1-p_1)}=3.086$$

This suggests that students whose parents did not go to college have higher odds of being less likely to apply.

Double negation can be logically confusing. Suppose we wanted to interpret the odds of being more likely to apply to college. We can perform a slight manipulation of our original odds ratio:

$$\begin{aligned}e^{-\eta_1}&=\frac{p_1/\left(1-p_1\right)}{p_0/\left(1-p_0\right)}=\frac{p_1\left(1-p_0\right)}{p_0\left(1-p_1\right)}=\frac{\left(1-p_0\right)/p_0}{\left(1-p_1\right)/p_1}\\&=\frac{P(Y>j|x=0)/P(Y\leq j|x=0)}{P(Y>j|x=1)/P(Y\leq j|x=1)}\end{aligned}$$

Since $e^{-\eta_1}=1/e^{\eta_1}$, then we have:

$$\frac{P(Y>j|x=1)/P(Y\leq j|x=1)}{P(Y>j|x=0)/P(Y\leq j|x=0)}=e^{\eta_1}$$

Instead of interpreting the odds of being in the j th category or less, we can interpret the odds of being greater than the j th category by exponentiating η itself. In our example, $e^{\eta_1}=e^{1.127}=3.086$ means that students whose parents went to college have 3.086 times the odds of being very likely to apply (vs. somewhat or unlikely) compared to students whose parents did not go to college. The results here are consistent with our intuition because it removes double negatives. As a general rule, it is easier to interpret the odds ratios of $x_1=1$ vs. $x_1=0$ by simply exponentiating η itself rather than interpreting the odds ratios of $x_1=0$ vs. $x_1=1$ by exponentiating $-\eta$. However by doing so, we flip the interpretation of the outcome by placing $P(Y>j)$ in the numerator.

Ok so far we have seen the derivation and the most important details of Binary Logistic Regression, Multinomial Logistic Regression and Ordinal Logistic Regression. We will stop here because in this textbook we want to focus only on the basic in the purpose to not be scientifically illiterate. The reader must however keep in mind that there are other logistic models like:

- **Generalized Ordinary Logistic Regression (GOLR):** Is a partial proportional odds model when the proportionality of odds doesn't hold.
- **Alternating Logistic Regression (ALR):** Is suited when we deal with correlated observations, e.g. when we analyse repeated or clustered data. Next to mixed-effect and marginal models fitted via GEE (Generalized Estimating Equations), there's also the ALR, handling the dependency between pairs of observations by using log odds ratios instead of correlations (GEEway). It handles ordinal responses too.
- **Fractional Logistic Regression (FLR):** Is typically used with 0-1 bounded fractions rather than just binary classes. More flexible than beta regression, but not as powerful as simplex regression or 0-1-inflated beta regression.
- **Logistic Quantile Regression (LQR):** Is a robust statistical method used to model the quantiles of an outcome variable that takes on values within a bounded, known interval. This method is particularly useful when dealing with data that is not symmetrically distributed or when the interest lies in understanding the conditional distribution of the outcome variable across different quantiles.
- **Conditional Logistic Regression:** Used if we deal with stratification and matching groups of data, e.g. in observational studies without randomization, to match subjects by a certain characteristics and create a more homogenous "baseline".

+ all the above models can be extended:

- To handle repeated / clustered observations - by employing random effects and GEE (generalized estimating equations).
- To relax linearity in parameters - by using splines or GAM (Generalized Additive Model), invented by Prof. Hastie & Tibshirani.

14.1.12.6.4 Log-Loss

Let us recall the relation that we have introduced just earlier for the binomial logistic regression:

$$\begin{aligned}\ln\left(L(D,w)\right)&=\ln\left(\prod_{i=1}^{n}P(y_i=1|x_i)^{y_i}(1-P(y_i=1|\vec{x}))^{1-y_i}\right)\\&=\sum_{i=1}^{n}y_i\ln(P(y_i=1|\vec{x}))+(1-y_i)\ln((1-P(y_i=1|\vec{x})))\end{aligned}\tag{14.1.900}$$

It is often rewritten as following and defined as the "Log-loss" (often used in Data Mining/Data Science that like to use the log in base 10 instead of the natural logarithm...):

$$\boxed{\text{LL}=-\frac{1}{n}\log\left(L(D,w)\right)=-\frac{1}{n}\sum_{i=1}^{n}\left[y_i\log(\hat{y}_i)+(1-y_i)\log(1-\hat{y}_i)\right]}\tag{14.1.901}$$

with $y_i \in \{0,1\}$ type of data with estimated $\hat{y}_i \in \{0,1\}$, thus $\text{LL} > 0$ (the Log-loss gradually declines as the predicted probability improves as the reader can notice it). In other words, Log-loss is used when we have $\{0,1\}$ response. This is usually because when we have $\{0,1\}$ response, the best models give us values in terms of probabilities.

It is also sometimes written as following:

$$\text{LL} = -\frac{1}{n}\log\left(L(D,w)\right) = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i\log(p_i) + (1-y_i)\log(1-p_i)\right] \tag{14.1.902}$$

In simple words (even if the Log-loss function is simply implicitly the objective function to minimize, in order to fit a log linear probability model to a set of binary labelled examples!), Log-loss measures the uncertainty of the probabilities of your model by comparing them to the true labels. Let us look closely at its formula and see how it measures the uncertainty

Now the question is, your training labels are 0 and 1 but your training predictions are 0.4, 0.6, 0.89, 0.1122, etc. So how do we calculate a measure of the error of our model ? If we directly classify all the observations having values > 0.5 into 1 then we are at a high risk of increasing the misclassification. This is because it may so happen that many values having probabilities 0.4, 0.45, 0.49 can have a true value of 1.

This is where Log-loss comes into picture!

Now let us closely follow the formula of Log-loss. There can be four major cases for the values of y_i and p_i:

C1. $y_i = 1\,,\, p_i = \text{High}\,,\, 1-y_i = 0\,,\, 1-p_i = \text{Low}$

C2. $y_i = 1\,,\, p_i = \text{Low}\,,\, 1-y_i = 0\,,\, 1-p_i = \text{High}$

C3. $y_i = 0\,,\, p_i = \text{Low}\,,\, 1-y_i = 1\,,\, 1-p_i = \text{High}$

C4. $y_i = 0\,,\, p_i = \text{High}\,,\, 1-y_i = 1\,,\, 1-p_i = \text{Low}$

- Case 1:

  In this case $y = 1$ and $p =$ High implies that we have got things right! Because the true value of the response agrees with our high probability. Now look closely! Occurrence of Case 1 will significantly decrease the sum because, $y_i\log(p_i)$ would be small and simultaneously the other term in the summation would be zero since $1-y_i = 1-1 = 0$. So more occurrences of Case 1 would decrease the sum and consequently decrease the mean.

  Also note that this is possible because if:

  $$p_i > 1-p_i \qquad \text{then} \qquad \log(p_i) > \log(1-p_i) \tag{14.1.903}$$

- Case 2:

  In this case $y = 1$ and $p =$ Low. This is a totally undesirable case because our probability of y being 1 is low but still the true value of y is 1. Now again looking at the Log-loss closely, the second term in the summation would be very small since $1-y_i$ would almost equal to zero. And since $p=$ Low, $y_i \log(p_i)$ would inflate the sum at the opposite of Case 1. So Case 2 would ultimately affect the sum a lot.

Similarly the occurrences of Case 3 would not change the sum and occurrences of Case 4 significantly.

Now coming back to the main question, how does Log-loss measure uncertainty of your model? The answer is simple! Suppose we have more of Case 1 and Case 3, then the sum inside the Log-loss relation would be small (would tend to decrease). This would imply that the mean $(/n)$ would also tend to decrease and will be substantially small in comparison to what it would have been if Case 2 and Case 4 got added. So now this value is as small as possible at Case 1 and Case 3 which indicates a good prediction. This is why smaller the value, better is the model i.e. smaller the Log-loss, better is the model i.e. smaller the uncertainty.

14.1.12.6.5 Odds Ratio Confidence Interval

We have seen earlier above during our study of the binomial logistic regression (see page 6456) that the odds ratio was defined by:

$$\boxed{\text{O.R.}=\frac{\dfrac{p}{1-p}}{\dfrac{q}{1-q}}=\frac{p(1-q)}{q(1-p)}} \tag{14.1.904}$$

What we can also write as:

$$\text{O.R.}=\frac{\dfrac{p_1}{q_1}}{\dfrac{p_2}{q_2}}=\frac{p_1q_2}{q_1p_2} \tag{14.1.905}$$

But that many practitioners write it as following:

$$\text{O.R.}=\frac{\dfrac{p_1}{p_2}}{\dfrac{p_3}{p_4}}=\frac{p_1p_4}{p_3p_2}=\frac{\pi_1\pi_4}{\pi_3\pi_2} \tag{14.1.906}$$

Since the ratio of proportions is equivalent to the ratio of the size of the concerned populations, it is customary to write the estimator of the odds ratio as being:

$$\boxed{\widehat{\text{O.R.}}=\frac{n_1n_4}{n_3n_2}} \tag{14.1.907}$$

Which then relates to a table typically of the following kind in the medical field (and other fields obviously!):

	Sick	**Non-Sick**	**Odds**
Treated Group	n_1	n_2	n_1/n_2
Non-Treated Group	n_3	n_4	n_3/n_4
Totals	n_1+n_3	n_2+n_4	$\widehat{\text{O.R.}}=\dfrac{\frac{n_1}{n_2}}{\frac{n_3}{n_4}}$

In the field of survival analysis (see section Statistics page 1271) such a table will be represented as following:

	Observed	**Expected**	**Relative failure rate**
Survival Curve 1	O_1	E_1	O_1/E_1
Survival Curve 2	O_2	E_2	O_2/E_2
Totals	O_1+O_2	E_2+E_4	$\widehat{\text{O.R.}}=\dfrac{\frac{O_1}{E_1}}{\frac{O_2}{E_2}}$

Remark

An O.R. of 1.00 means that the two groups were equally likely to remain sick. An O.R. higher than 1.00 means that a group is more likely to experience the event than the second group. An O.R. of less than 1.00 means that a group was less likely to experience the event. However, an O.R. value below 1.00 is not directly interpretable. It is important to put the group expected to have higher odds of the event in the first row.

Now, let us consider that the n_i follow an asymptotically Normal law such as according to what we had seen in our study of the construction of the Normal Law:

$$Z_i \underset{n\rightarrow+\infty}{\cong} \frac{n_i-np_i}{\sqrt{np_iq_i}} \tag{14.1.908}$$

Now the idea will be to make an additional approximation ... otherwise we will not find anything simple. We will put that:

$$Z_i \underset{n\rightarrow+\infty}{\cong} \frac{n_i-np_i}{\sqrt{np_iq_i}} \cong \frac{n_i-np_i}{\sqrt{n}} \tag{14.1.909}$$

It is very very approximate as approach .... since rigorously:

$$\frac{n_i-np_i}{\sqrt{n}} \underset{n\rightarrow+\infty}{\cong} \mathcal{N}(0,\sigma^2) \tag{14.1.910}$$

Simulations show that we have a Normal law beyond n equating 30, but although it is a Normal centered law ... it is hardly reduced (therefore not of unit variance...).

Regarding the last choice established, we have:

$$\begin{aligned}
&\mathrm{E}(Z_i)\cong 0\\
&\mathrm{V}(Z_i)\cong V\left(\frac{n_i-np_i}{\sqrt{n}}\right)=\mathrm{V}\left(\frac{n_i}{\sqrt{n}}\right)+\underbrace{\mathrm{V}\left(\frac{np_i}{\sqrt{n}}\right)}_{=0}=\frac{1}{n}\mathrm{V}(n_i)=\frac{1}{n}np_iq_i=p_iq_i=p_i(1-p_i)
\end{aligned} \tag{14.1.911}$$

This last approximation then allows us to write (we detail the steps a maximum on the request of one of our reader):

$$\begin{aligned}
&Z_i\cong\frac{n_i-np_i}{\sqrt{n}}\Rightarrow n_i\cong np_i+\sqrt{n}Z_i\Rightarrow n_i=np_i+\frac{n}{n}\sqrt{n}Z_i\\
&\Rightarrow\frac{n_i}{n}\cong p_i+\frac{\sqrt{n}Z_i}{n}=p_i+\frac{\sqrt{n}}{\sqrt{n}}\frac{\sqrt{n}Z_i}{n}=p_i+\frac{nZ_i}{n\sqrt{n}}=p_i+\frac{Z_i}{\sqrt{n}}\\
&=p_i\left(1+\frac{Z_i}{p_i\sqrt{n}}\right)
\end{aligned} \tag{14.1.912}$$

Now let us take the logarithm (it is customary to take the natural logarithm):

$$\ln\left(\frac{n_i}{n}\right) \cong \ln(p_i) + \ln\left(1 + \frac{Z_i}{p_i\sqrt{n}}\right) \tag{14.1.913}$$

The second term can be developed in Taylor series of the first order (see section Sequences and Series page 1703):

$$\ln(1+x) \underset{x\rightarrow +0}{\cong} x \tag{14.1.914}$$

It comes then after many approximations ...:

$$\ln\left(\frac{n_i}{n}\right) \cong \ln(p_i) + \frac{Z_i}{p_i\sqrt{n}} \tag{14.1.915}$$

We then have:

$$\begin{aligned}\ln(\text{O.R.}) &= \ln\left(\frac{n_1 n_4}{n_3 n_2}\right) = \ln\left(\frac{\frac{n_1}{n}\frac{n_4}{n}}{\frac{n_3}{n}\frac{n_2}{n}}\right) = \ln\left(\frac{n_1}{n}\right) + \ln\left(\frac{n_4}{n}\right) - \ln\left(\frac{n_2}{n}\right) - \ln\left(\frac{n_3}{n}\right)\\ &\cong \ln(p_1) + \frac{Z_1}{p_1\sqrt{n}} + \ln(p_4) + \frac{Z_4}{p_4\sqrt{n}} - \ln(p_2) - \frac{Z_2}{p_2\sqrt{n}} - \ln(p_4) - \frac{Z_3}{p_3\sqrt{n}}\\ &= \ln\left(\frac{p_1 p_4}{p_3 p_2}\right) + \frac{1}{\sqrt{n}}\left(\frac{Z_1}{p_1} + \frac{Z_4}{p_4} - \frac{Z_2}{p_2} - \frac{Z_3}{p_3}\right) \cong \ln(\widehat{\text{O.R.}}) + \frac{1}{\sqrt{n}}\left(\frac{Z_1}{p_1} + \frac{Z_4}{p_4} - \frac{Z_2}{p_2} - \frac{Z_3}{p_3}\right)\end{aligned} \tag{14.1.916}$$

However, we observe that (do not forget that the variance of a random variable multiplied by a constant puts the constant squared):

$$\ln(\text{O.R.}) \cong \underbrace{\ln(\widehat{\text{O.R.}})}_{=c^{te}} + \underbrace{\frac{1}{\sqrt{n}}\left(\frac{Z_1}{p_1} + \frac{Z_4}{p_4} - \frac{Z_2}{p_2} - \frac{Z_3}{p_3}\right)}_{=\mathcal{N}(0,\sigma)} = \mathcal{N}\left(\ln(\widehat{\text{O.R.}}), \frac{\sigma^2}{n}\right) \tag{14.1.917}$$

We then first consider that the variance part of the random variables as independent (we will take into account the covariance just after):

$$\begin{aligned}\text{V}(\ln(\text{O.R.})) &\cong \frac{1}{n}\left[\frac{\text{V}(Z_1)}{p_1^2} + \frac{\text{V}(Z_4)}{p_4^2} + \frac{\text{V}(Z_2)}{p_2^2} + \frac{\text{V}(Z_3)}{p_3^2}\right]\\ &= \frac{1}{n}\left[\frac{p_1 q_1}{p_1^2} + \frac{p_4 q_4}{p_4^2} + \frac{p_2 q_2}{p_2^2} + \frac{p_3 q_3}{p_3^2}\right] = \frac{1}{n}\left[\frac{1-p_1}{p_1} + \frac{1-p_4}{p_4} + \frac{1-p_2}{p_2} + \frac{1-p_3}{p_3}\right]\\ &= \sum_{i=1}^{4}\frac{1-p_i}{np_i} = \frac{1}{n}\left(-4 + \sum_{i=1}^{4}\frac{1}{p_i}\right)\end{aligned} \tag{14.1.918}$$

We now have to take into account the following six covariances (refer to the section Statistics for the proof that when we sum n random variables there are as many covariances to compute as possible pairs of possible variables):

$$\begin{gathered}\text{cov}(Z_1, Z_4)\\ \text{cov}(Z_1, Z_2)\\ \text{cov}(Z_1, Z_3)\\ \text{cov}(Z_4, Z_2)\\ \text{cov}(Z_4, Z_3)\\ \text{cov}(Z_2, Z_3)\end{gathered} \tag{14.1.919}$$

Recall that during our study of the multinomial law and as part of an application for the McNemar test (page 1264), we proved that:

$$\text{cov}(n_{11},n_{22})=-np_{11}p_{22} \tag{14.1.920}$$

which in our context with a more general notation is written:

$$\text{cov}(n_r,n_s)=-np_rp_s \tag{14.1.921}$$

If we center these two random variables (as is the case in our developments so far), the covariance does not change:

$$\text{cov}(n_r-np_r,n_s-np_s)=\text{cov}(n_r,n_s) \tag{14.1.922}$$

If we reduce the random variables by $\sqrt{n}$, then we have:

$$\text{cov}(Z_r,Z_s)=\text{cov}\left(\frac{n_r-np_r}{\sqrt{n}},\frac{n_s-np_s}{\sqrt{n}}\right)=\text{cov}\left(\frac{n_r}{\sqrt{n}},\frac{n_s}{\sqrt{n}}\right)=\frac{1}{n}\text{cov}(n_r,n_s)=-p_rp_s \tag{14.1.923}$$

and therefore:

$$\text{cov}\left(\frac{Z_r}{p_r},\frac{Z_s}{p_s}\right)=\frac{\text{cov}\left(Z_r,Z_s\right)}{p_rp_s}=\frac{-p_rp_s}{p_rp_s}=-1 \tag{14.1.924}$$

Among the six covariances that we need to calculate we have therefore:

$$\begin{aligned}\text{cov}\left(Z_1,Z_4\right)&=+1\\ \text{cov}\left(Z_1,Z_2\right)&=-1\\ \text{cov}\left(Z_1,Z_3\right)&=-1\\ \text{cov}\left(Z_4,Z_2\right)&=-1\\ \text{cov}\left(Z_4,Z_3\right)&=-1\\ \text{cov}\left(Z_2,Z_3\right)&=+1\end{aligned} \tag{14.1.925}$$

So the sum is worth -2. It is necessary to multiply by 2 the whole (reminder of the proof of the general case which makes a factor 2 appear before each term of the covariance) to obtain the global value of the covariance terms. So in the end we have a value of -4 for the sum of the pure covariances:

$$2\sum_{r<s}^{4}\text{cov}\left(Z_r,Z_s\right)=-4 \tag{14.1.926}$$

The total variance is ultimately given by:

$$\text{V}(\ln(\widehat{\text{O.R.}}))=\frac{1}{n}\left[\left(-4+\sum_{i=1}^{4}\frac{1}{p_i}\right)-4\right]=\frac{1}{n}\sum_{i=1}^{4}\frac{1}{p_i}=\sum_{i=1}^{4}\frac{1}{np_i}=\sum_{i=1}^{4}\frac{1}{n_i} \tag{14.1.927}$$

And finally we have the approximation sometimes named "Woolf method":

$$\ln(\widehat{\text{O.R.}})-Z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\leq\ln(\text{O.R.})\leq\ln(\widehat{\text{O.R.}})+Z_{1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}} \tag{14.1.928}$$

More explicitly (notice that the lower and upper bound are symmetrical):

$$\boxed{\ln(\widehat{\text{O.R.}})-Z_{\alpha/2}\sqrt{\sum_{i=1}^{4}\frac{1}{n_i}}\leq\ln(\text{O.R.})\leq\ln(\widehat{\text{O.R.}})+Z_{1-\alpha/2}\sqrt{\sum_{i=1}^{4}\frac{1}{n_i}}} \tag{14.1.929}$$

Relationship that is often written as:

$$\ln(\widehat{\text{O.R.}})\pm Z_{\alpha/2}\hat{\sigma}_{\text{O.R.}}\leq\ln(\text{O.R.}) \tag{14.1.930}$$

Some softwares calculate the upper and lower bound by using the exponential (especially softwares plotting the famous "Four-fold plot"), and therefore the upper and lower bound are no longer symmetrical:

$$\widehat{\text{O.R.}}.e^{-Z_{\alpha/2}\sqrt{\sum_{i=1}^{4}\frac{1}{n_i}}}\leq\text{O.R.}\leq\widehat{\text{O.R.}}.e^{+Z_{1-\alpha/2}\sqrt{\sum_{i=1}^{4}\frac{1}{n_i}}} \tag{14.1.931}$$

✍**Example:**

Let us consider the following data:

Group / Heart Attack	**Yes**	**No**
Placebo	53	58
Treated Group	11	40

The odds ratio is then:

$$\text{O.R.}=\frac{\dfrac{n_1}{n_2}}{\dfrac{n_3}{n_4}}=\frac{\dfrac{53}{58}}{\dfrac{11}{40}}\cong 3.32 \tag{14.1.932}$$

So the Placebo group is almost three times more likely to experience the event of interest (Heart Attack).

You must keep in mind the three following scenarios (refresh!):

- The odds ratio is almost or equal to 1.0. means that the odds of exposure among cases is the same as the odds of exposure among controls. In other words, the exposure is not associated with the disease.
- The odds ratio is greater than 1.0 means that the odds of exposure among cases is greater than the odds of exposure among controls. In other words, the exposure may be a risk factor for the disease.
- The odds ratio is less than 1.0 means that the odds of exposure among cases is lower than the odds of exposure among controls. In other words, the exposure may be protective against the disease.

The confidence interval of $\ln(\text{O.R.})$ at a confidence level of 95% is then approximately given by:

$$3.32-1.96\sqrt{\frac{1}{53}+\frac{1}{58}+\frac{1}{11}+\frac{1}{4}}\leq \ln(\text{O.R.})\leq 3.32+1.96\sqrt{\frac{1}{53}+\frac{1}{58}+\frac{1}{11}+\frac{1}{40}} \tag{14.1.933}$$

Thus:

$$3.32-1.96\cdot 0.39\leq \ln(\text{O.R.})\leq 3.32+1.96\cdot 0.39 \tag{14.1.934}$$

Hence:

$$0.44\leq \ln(\text{O.R.})\leq 1.97 \tag{14.1.935}$$

and so taking the exponential we finally have the interval that interests us (read further below the text on the corresponding four-fold plot!):

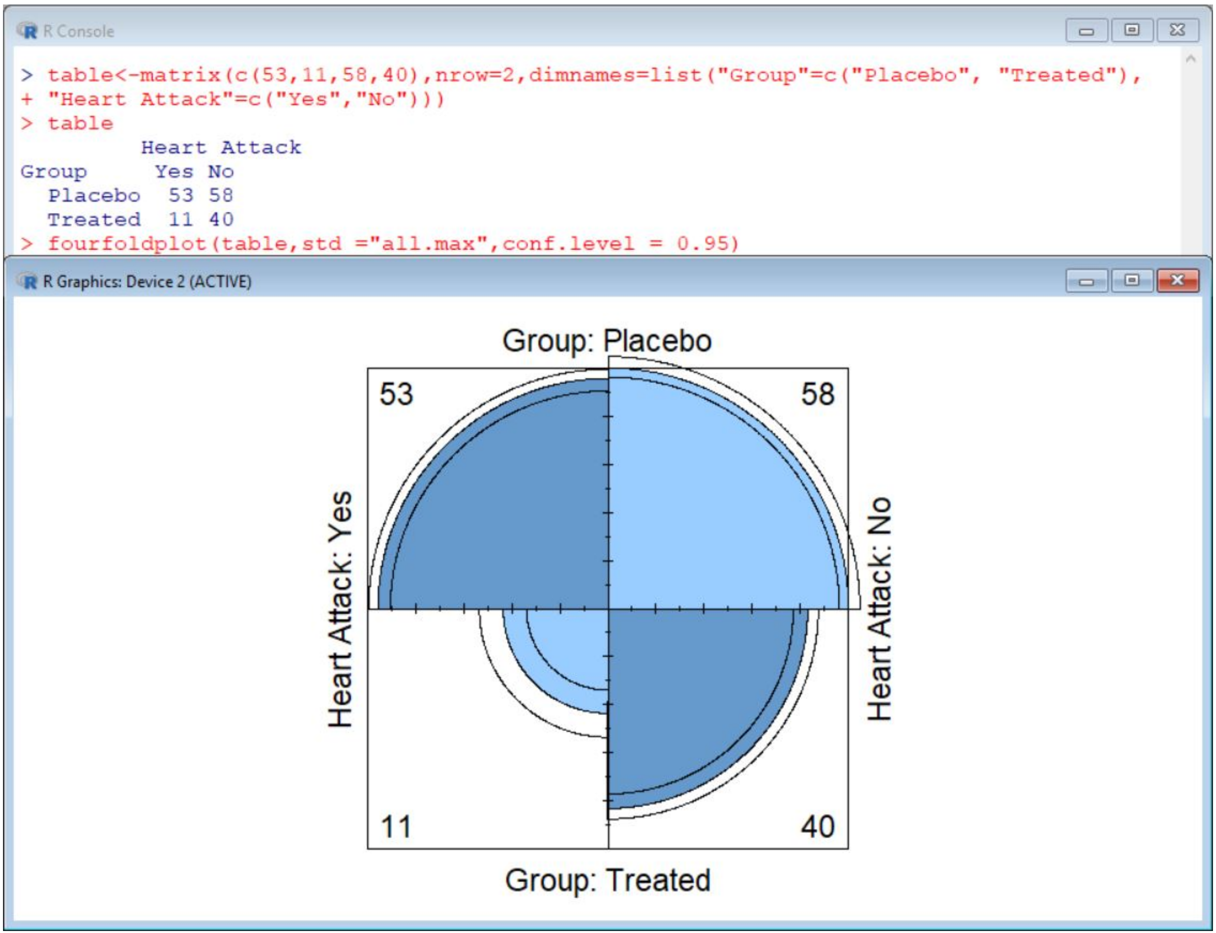
$$1.55\leq \text{O.R.}\leq 7.14 \tag{14.1.936}$$

Caution!!! While the estimator of the natural logarithm of the odds ratio is in the middle of the confidence interval, this is obviously not the case with the standard odds ratio!

14.1.12.6.5.1 Four-fold plot

The "four-fold plot", or "fourfold plot, is a graphic designed to display the frequencies in a 2×2 contingency table. In this display the frequency in each cell is shown by a quarter circle, whose area is proportional to the cell count in a given 2×2 layer, in a way that depicts the odds ratios for 1 strata. Confidence rings for the odds ratio can be superimposed to provide a visual test of the hypothesis of no association in each stratum.

Before we deal with the maths let us look with the R 3.4.2 software how to generate a four-fold plot for the previous example:



```
> table<-matrix(c(53,11,58,40),nrow=2,dimnames=list("Group"=c("Placebo", "Treated"),
+ "Heart Attack"=c("Yes","No")))
> table
         Heart Attack
Group      Yes No
  Placebo   53 58
  Treated   11 40
> fourfoldplot(table,std ="all.max",conf.level = 0.95)
```

Figure 14.25: **Four-fold plot with R 3.4.2**

In this display the frequency n_i in each cell of a four-fold table is shown by a quarter circle (on a vertical and horizontal axis of range $[-1,+1]$ with major ticks each 0.2 and minor ticks each 0.1), whose radius is proportional to $\sqrt{n_i}$ so that they are is proportional to the cell count (otherwise it would be proportional to the square of the cell count as $S=\pi R^2$).

The four-fold display is constructed so that the four quadrants will align vertically and horizontally when the odds ratio is 1. Confidence rings for the observed odds ratio provide a visual test of the hypothesis of no association. They have the property that rings for adjacent quadrants overlap if and only if the observed counts are consistent with this null hypothesis. In the above figure the confidence intervals do not overlap, indicating a significant association between *Heart attack* and *Group*.

The radius of the circles are given by:

Group / Heart Attack	**Yes**	**No**
Placebo	$\sqrt{\dfrac{n_1}{\max n_i}}$	$\sqrt{\dfrac{n_2}{\max n_i}}$
Treated Group	$\sqrt{\dfrac{n_3}{\max n_i}}$	$\sqrt{\dfrac{n_4}{\max n_i}}$

Explicitly:

Group / Heart Attack	**Yes**	**No**
Placebo	$\sqrt{\frac{53}{58}}=0.955$	$\sqrt{\frac{58}{58}}=1.000$
Treated Group	$\sqrt{\frac{11}{58}}=0.435$	$\sqrt{\frac{40}{58}}=0.830$

14.1.12.6.6 Risk Ratio and its Confidence Interval

The "risk ratio" or "relative risk" is defined quite naturally by:

$$\text{R.R.}=\frac{\text{Probability of event 1}}{\text{Probability of event 2}} \tag{14.1.937}$$

Consider the typical case in the medical field:

	Sick	**Non-Sick**	**Total**
Treated Group	n_1	n_2	n_A
Non-Treated Group	n_3	n_4	n_B

Thus the risk of illness for the treated group (TG), also named "Experimental Event Rate" (EER), is:

$$\hat{R}_{\text{TG}}=\text{EER}=\frac{n_1}{n_1+n_2}=p_1 \tag{14.1.938}$$

and for the non-treated group (NTG) (or "control group"), also named "Control Event Rate" (CER):

$$\hat{R}_{\text{NTG}}=\text{CER}=\frac{n_3}{n_3+n_4}=p_2 \tag{14.1.939}$$

and hence the risk ratio is defined by:

$$\boxed{\text{R.R.}=\frac{\dfrac{n_1}{n_1+n_2}}{\dfrac{n_3}{n_3+n_4}}=\frac{p_1}{p_2}=\frac{\text{EER}}{\text{CER}}=\frac{\text{Risk treated group}}{\text{Risk control group}}=\text{O.R.}\frac{\dfrac{n_2}{n_1+n_2}}{\dfrac{n_4}{n_3+n_4}}} \tag{14.1.940}$$

remembering that:

$$\text{O.R.}=\frac{\dfrac{n_1}{n_2}}{\dfrac{n_3}{n_4}} \tag{14.1.941}$$

Risk Ratios (RR) and Odds Ratios (OR) are both used to analyse factors associated with increased risk of disease. They are different in that a risk ratio is the ratio of two risks and an odds ratio is the ratio of two odds. Risks and odds are calculated differently. For example, with one toss of a fair die, the risk of rolling a 1 is $1/6=0.167$. The odds of rolling a 1 are $1/5=0.200$. In textbook terms, risk is successes divided by total trials and odds are successes divided by failures. In this example rolling a 1 is a success and rolling 2 through 5 is a failure.

✍**Example:**

Let us consider a R.R of 1.50. That means obviously an increase of the risk of 50% in the treatment group in comparison of the control group. So if the control group has a risk of 20%, then the treatment group has a risk of $20\% + 1.5 \cdot 20\% = 30\%$.

Remark

If n_1 is small relatively to n_2 and n_3 is small relatively to n_4 we fall back on the odds ratio:

$$\widehat{\text{R.R.}} \cong \widehat{\text{O.R.}} = \frac{\dfrac{n_1}{n_2}}{\dfrac{n_3}{n_4}} = \frac{n_1 n_4}{n_3 n_2} \tag{14.1.942}$$

we then better understand why the odds ratio is sometimes named "approximate relative risk".

So in the medical field, we say that if the disease is rare in treated and non-treated groups the interpretation of relative risk (i.e. risk ratio) and odds ratio is the same.

Let us take the log of the relative risk (i.e. risk ratio):

$$\ln(\widehat{\text{R.R.}}) = \ln(p_1) - \ln(p_2) \tag{14.1.943}$$

It then comes for the variance since p_1 and p_2 are assumed to be independent:

$$\text{V}(\ln(\text{R.R.})) = \text{V}(\ln(p_1) - \ln(p_2)) = \text{V}(\ln(p_1)) + \text{V}(\ln(p_2)) \tag{14.1.944}$$

To determine the variance of the natural logarithm of the proportions we will have to use a technique that we name the "univariate delta method" (indeed, because of the presence of a sum at the denominator, we will not be able to take the same approach as the one used for the odds ratio) which consists of approximating the exact variance by the corresponding asymptotically normal variance.

To see what the univariate delta method is, let's remember that we have proved in the section Statistics (see page 696) that:

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \underset{n \to +\infty}{\cong} Z \Leftrightarrow \sqrt{n}(\bar{X} - \mu) \underset{n \to +\infty}{\cong} \sigma Z \tag{14.1.945}$$

Now, let us consider the following linear function:

$$g(t) = at + b \tag{14.1.946}$$

where a and b are any two constants. From then on, using the linearity property of the mean, we immediately have:

$$\text{E}(g(X)) = \text{E}(aX + b) = a\text{E}(X) + b = a\mu + b = g(\mu) \tag{14.1.947}$$

Now, let us ask ourselves the question, what is the limit of:

$$\lim_{n\rightarrow+\infty}\sqrt{n}\left(g(\bar{X})-g(\mu)\right) \tag{14.1.948}$$

We then have:

$$\sqrt{n}\left(g(\bar{X})-g(\mu)\right)=\sqrt{n}\left(a\bar{X}+b-(a\mu+b)\right)=a\sqrt{n}(\bar{X}-\mu) \tag{14.1.949}$$

We immediately deduce that:

$$\sqrt{n}\left(g(\bar{X})-g(\mu)\right)\underset{n\rightarrow+\infty}{\cong}a\sigma Z=N(0,a^2\sigma^2) \tag{14.1.950}$$

Now, let us develop in series of Taylor in the first order around μ the function $g(t)$, which would not necessarily be linear, then we have:

$$g(t)\cong g(\mu)+g'(\mu)(t-\mu)=g'(\mu)t-g'(\mu)\mu+g(\mu) \tag{14.1.951}$$

By identification with the simple linear case, we then have:

$$g(t)\cong\underbrace{g'(\mu)}_{a}t+\underbrace{g(\mu)-g'(\mu)\mu}_{b} \tag{14.1.952}$$

It then comes in the case of a function $g(t)$ that is not necessarily linear, but approximated as such by a series development of Taylor and under the assumption that the variable is continuous that:

$$\sqrt{n}\left(g(\bar{X})-g(\mu)\right)\underset{n\rightarrow+\infty}{\cong}g'(\mu)\sigma Z=\mathcal{N}(0,g'(\mu)^2\sigma^2) \tag{14.1.953}$$

So don't forget what this method means (!!): If X_n is (asymptotically) Normal with mean 0 and variance σ^2, g is differentiable and $g'(\mu)\neq 0$, then $g(X_n)$ is approximately given by:

$$\boxed{g(X_n)\underset{n\rightarrow+\infty}{\cong}\mathcal{N}\left(0,g'(\mu)^2\sigma^2\right)} \tag{14.1.954}$$

For the purpose of demonstrating the confidence interval of the risk ratio, let us apply that to the binomial distribution with:

$$g(t)=\ln(t) \tag{14.1.955}$$

It then comes from the previous results that the approximation of the variance by a Normal law and from the knowledge that the variance and the mean of the binomial law are given for recall by (see section Statistics page 658):

$$\mu=np \tag{14.1.956}$$
$$\sigma^2=npq \tag{14.1.957}$$

that the univariate delta method gives:

$$\begin{aligned}\mathcal{N}(0,g'(\mu)^2\sigma^2)&=\mathcal{N}\left(0,\left(\frac{\mathrm{d}\ln(np)}{\mathrm{d}np}\right)^2npq\right)=\mathcal{N}\left(0,\left(\frac{1}{np}\right)^2npq\right)=\mathcal{N}\left(0,\frac{q}{np}\right)\\&=\mathcal{N}\left(0,\frac{1-p}{np}\right)\end{aligned} \tag{14.1.958}$$

So the variance corresponding to $\ln(p_i)$ is given asymptotically by:

$$\text{V}(\ln(p_i))=\frac{1-p_i}{n_ip_i} \tag{14.1.959}$$

Therefore:

$$\text{V}\left(\ln(\widehat{\text{R.R.}})\right)=\text{V}(\ln(p_1))+\text{V}(\ln(p_2))\cong\frac{1-p_1}{n_Ap_1}+\frac{1-p_2}{n_Bp_2} \tag{14.1.960}$$

Hence:

$$\hat{\sigma}_{\widehat{\text{R.R.}}}\cong\sqrt{\frac{1-p_1}{n_Ap_1}+\frac{1-p_2}{n_Bp_2}} \tag{14.1.961}$$

Thus, we have asymptotically the following interval for the risk ratio if the assumptions are satisfied (independence and underlying Normal distribution):

$$\boxed{\ln(\widehat{\text{R.R.}})-Z_{\alpha/2}\sqrt{\frac{1-p_1}{n_Ap_1}+\frac{1-p_2}{n_Bp_2}}\leq\ln(\widehat{\text{R.R.}})\leq\ln(\widehat{\text{R.R.}})+Z_{\alpha/2}\sqrt{\frac{1-p_1}{n_Ap_1}+\frac{1-p_2}{n_Bp_2}}} \tag{14.1.962}$$

or in condensed form:

$$\ln(\widehat{\text{R.R.}})-Z_{\alpha/2}\hat{\sigma}_{\widehat{\text{R.R.}}}\leq\ln(\widehat{\text{R.R.}})\leq\ln(\widehat{\text{R.R.}})+Z_{\alpha/2}\hat{\sigma}_{\widehat{\text{R.R.}}} \tag{14.1.963}$$

Note that in the medical literature, the reader will also find the standard deviation of the risk ratio in many other forms that are here (which are of course all equivalent):

$$\begin{aligned}\hat{\sigma}_{\widehat{\text{R.R.}}}&=\sqrt{\frac{1-p_1}{n_Ap_1}+\frac{1-p_2}{n_Bp_2}}=\sqrt{\frac{1-p_1}{(n_1+n_2)p_1}+\frac{1-p_2}{(n_3+n_4)p_2}}\\&=\sqrt{\frac{1-\dfrac{n_1}{n_1+n_2}}{(n_1+n_2)\dfrac{n_1}{n_1+n_2}}+\frac{1-\dfrac{n_3}{n_3+n_4}}{(n_3+n_4)\dfrac{n_3}{n_3+n_4}}}=\sqrt{\frac{1-\dfrac{n_1}{n_1+n_2}}{n_1}+\frac{1-\dfrac{n_3}{n_3+n_4}}{n_3}}\\&=\sqrt{\frac{1}{n_1}-\frac{1}{n_1+n_2}+\frac{1}{n_3}-\frac{1}{n_3+n_4}}=\sqrt{\frac{n_2}{n_1(n_1+n_2)}+\frac{n_4}{n_3(n_3+n_4)}}\end{aligned} \tag{14.1.964}$$

If the risk ratio is the same for both target groups, the Risk Ratio will obviously be equal to 100% (i.e. equal to 1). If it is less than 100% (i.e. less than 1), it simply means that the risk of illness in the treated group is lower than that of the non-treated group, etc.

The typical range of values of the odds ratio and risk ratio are designated (wrongly!) traditionally as following in scientific publications:

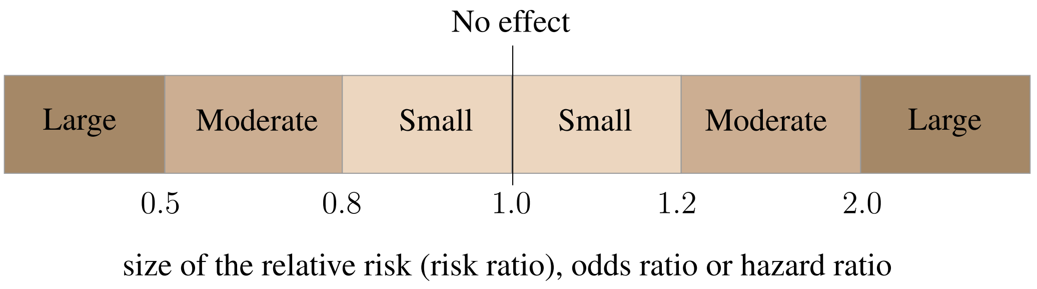

No effect
Large
Moderate
Small
Small
Moderate
Large
0.5
0.8
1.0
1.2
2.0
size of the relative risk (risk ratio), odds ratio or hazard ratio

Figure 14.26: **Odds Ratio and Risk Ratio range values subjective qualifications**

Caution!!! While the natural logarithm Risk Ratio estimator is in the middle of the confidence interval, this is obviously not the case for the standard Risk Ratio!

✍**Example:**

Let us consider the following data:

Group/Heart attack	**Yes**	**No**
Placebo	53	58
Aspirin	11	40

The relative risk is then (which corresponds to what the MedCalc software returns):

$$\widehat{\text{R.R}}=\frac{\text{EER}}{\text{CER}}=\frac{\dfrac{n_1}{n_1+n_2}}{\dfrac{n_3}{n_3+n_4}}=\frac{\dfrac{53}{53+58}}{\dfrac{11}{11+40}}=\frac{0.477}{0.215}=\frac{p_1}{p_2}\cong 2.21 \tag{14.1.965}$$

The confidence interval of $\ln(\widehat{\text{R.R.}})$ at a confidence level of 95% is then approximately given by:

$$\ln(2.21)-Z_{\alpha/2}\sqrt{\frac{1-0.477}{(53+58)\cdot 0.477}}\leq\ln(\widehat{\text{RR}})\leq\ln(2.21)+Z_{\alpha/2}\sqrt{\frac{1-0.477}{(53+58)\cdot 0.477}} \tag{14.1.966}$$

Hence:

$$0.794-1.96\cdot 0.284\leq\ln(\widehat{\text{RR}})\leq 0.794+1.96\cdot 0.284 \tag{14.1.967}$$

So finally:

$$0.236\cdot 0.284\leq\ln(\widehat{\text{RR}})\leq 1.353 \tag{14.1.968}$$

and so taking the exponential we finally have the interval that interests us:

$$1.266\leq\ln(\widehat{\text{RR}})\leq 3.869 \tag{14.1.969}$$

This perfectly matches the data returned by the MedCalc software! The Absolute Risk Increase is given by:

$$\text{A.R.I.}=\text{E.E.R.}-\text{C.E.R.}=0.447-0.215=26.2\% \tag{14.1.970}$$

The reader or practitioner must keep in mind that the odds ratio, as the risk ratio, are tools as we have just seen, used to make null hypothesis statistical testing (NHST) as they involve confidence intervals. So don't forget that what we are seeking is to reject the null hypothesis H_0 ! Because of this, a lot of people read medical studies wrongly as they have forgotten how NHST works (or they just don't know...)!

So for recall:

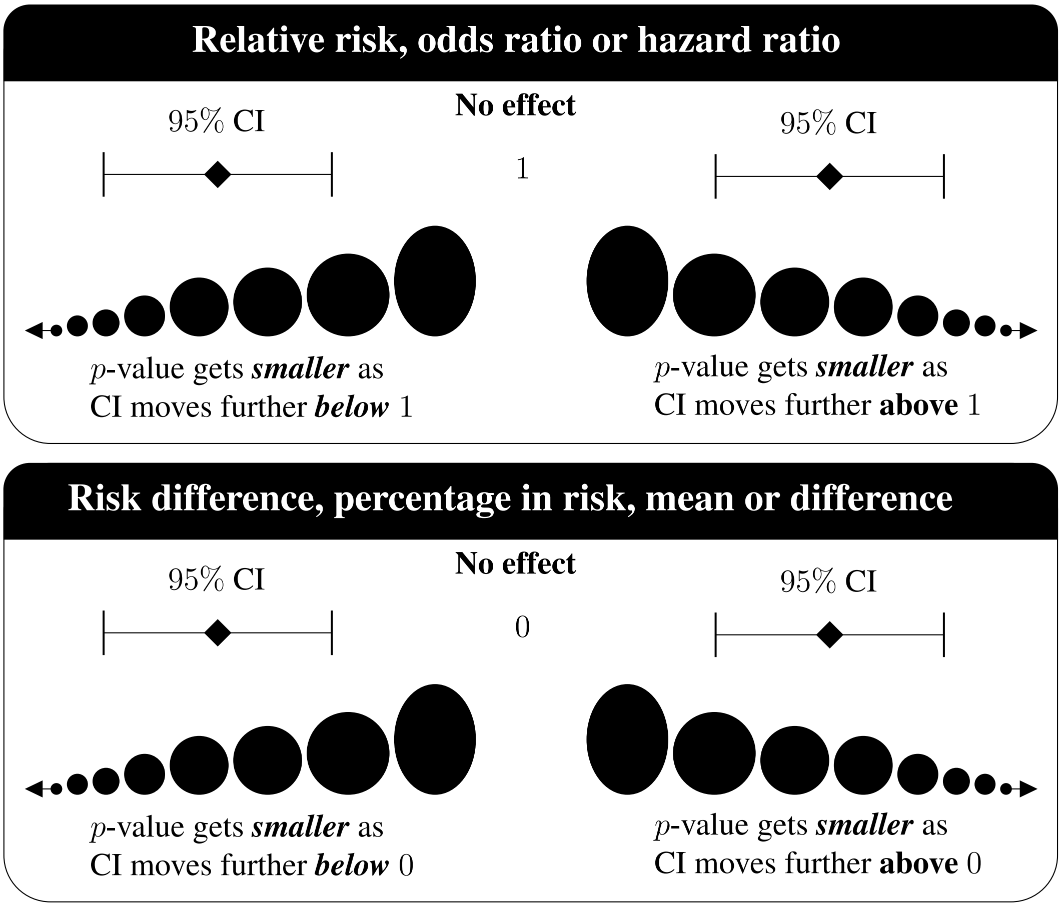

Relative risk, odds ratio or hazard ratio
95% CI
No effect
1
95% CI
p-value gets ***smaller*** as
CI moves further ***below*** 1
p-value gets ***smaller*** as
CI moves further **above** 1
Risk difference, percentage in risk, mean or difference
95% CI
No effect
0
95% CI
p-value gets ***smaller*** as
CI moves further ***below*** 0
p-value gets ***smaller*** as
CI moves further **above** 0

Figure 14.27: Risk Ratio (relative risk), Odds Ratio or Hazard Ratio rules

So far we have seen that in calculation of patient risks there are two indicators. But that's quite wrong in practice... Indeed, it is quite common to find the following indicators in many scientific papers:

- Odds ratio:

$$\widehat{\text{O.R}}=\frac{n_1n_4}{n_2n_3} \tag{14.1.971}$$

- Relative risk (risk ratio):

$$\widehat{\text{R.R.}}=\frac{\dfrac{n_1}{n_1+n_2}}{\dfrac{n_3}{n_3+n_4}} \tag{14.1.972}$$

- Experimental event rate (EER):

$$\widehat{\text{E.E.R.}}=\frac{n_1}{n_1+n_2} \tag{14.1.973}$$

- Control event rate (CER):

$$\widehat{\text{C.E.R.}}=\frac{n_3}{n_3+n_4} \tag{14.1.974}$$

- Absolute risk increase (also named "attributable risk"):

$$\widehat{\text{A.R.I.}}=\frac{n_1}{n_1+n_2}-\frac{n_3}{n_3+n_4}=\widehat{\text{E.E.R.}}-\widehat{\text{C.E.R.}} \tag{14.1.975}$$

- Absolute risk reduction:

$$\widehat{\text{A.R.R.}}=\frac{n_3}{n_3+n_4}-\frac{n_1}{n_1+n_2}=\widehat{\text{C.E.R.}}-\widehat{\text{E.E.R.}} \tag{14.1.976}$$

- Number needed to harm:

$$\widehat{\text{N.N.H.}}=\frac{1}{\dfrac{n_1}{n_1+n_2}-\dfrac{n_3}{n_3+n_4}}=\frac{1}{\widehat{\text{A.R.I.}}} \tag{14.1.977}$$

It should be noticed that any scientific paper in clinical research should communicate at least the Risk Ratio (RR), the Odds Ratio (OR) the Absolute Risk Reduction (ARR) and their respective confidence intervals at a given threshold! We have already derived how to compute the confidence interval for the RR and OR. But we didn't for ARR. So let's do it!

✍**Example:**

- Cancer risk for smokers:

$$\frac{10}{30}=0.3\bar{3}$$

- Cancer risk for non-smokers:

$$\frac{1}{30}=0.03\bar{3}$$

- Risk difference:

$$\widehat{\text{R.D.}}=\frac{10}{30}-\frac{1}{30}=0.3$$

- Relative risk (risk ratio) for the smokers:

$$\widehat{\text{R.R}}=\frac{10/30}{1/30}=10$$

- Relative risk (risk ratio) for the non-smokers:

$$\widehat{\text{R.R}}=\frac{1/30}{10/30}=0.1$$

- Relative risk (risk ratio) reduction for the non-smokers ($\widehat{\text{R.R}}=0.1$):

$$\widehat{\text{R.R.R}}=100\cdot(1-\widehat{\text{R.R}})=100\cdot(1-0.1)=90\%$$

- Number need to treat (number of people who must be treated to avoid a case of bad outcome):

$$\widehat{\text{N.D.T.}}=\frac{1}{0.3}=3.\bar{3}$$

It is obvious that C.E.R and E.E.R. are binomial random variables that belong to the range $[0,1]$ and furthermore independent (i.e. without correlation) by construction of the experimental design. Therefore the variance of A.R.R. is the sum of the variance of two independent binomial variables. Sadly there are eleven methods to calculate such an interval (...). The most common and naive one for two binomial independent random variable is given by (see section Statistics page 880):

$$(\widehat{\text{CER}}-\widehat{\text{EER}})-Z_{\alpha/2}\sqrt{\frac{\widehat{\text{CER}}(1-\widehat{\text{CER}})}{n_3+n_4}+\frac{\widehat{\text{EER}}(1-\widehat{\text{EER}})}{n_1+n_2}}\leq \text{ARR}\leq(\widehat{\text{CER}}-\widehat{\text{EER}})+Z_{\alpha/2}\sqrt{\frac{\widehat{\text{CER}}(1-\widehat{\text{CER}})}{n_3+n_4}+\frac{\widehat{\text{EER}}(1-\widehat{\text{EER}})}{n_1+n_2}} \tag{14.1.978}$$

And keep always in mind what we already mentioned during our study of Chemistry:

quote
14.1:6510.#76

> *All things are poisons, for there is nothing without poisonous qualities. It is only the dose which makes a thing poison*[28]*!*
>
> Paracelsus

Remark

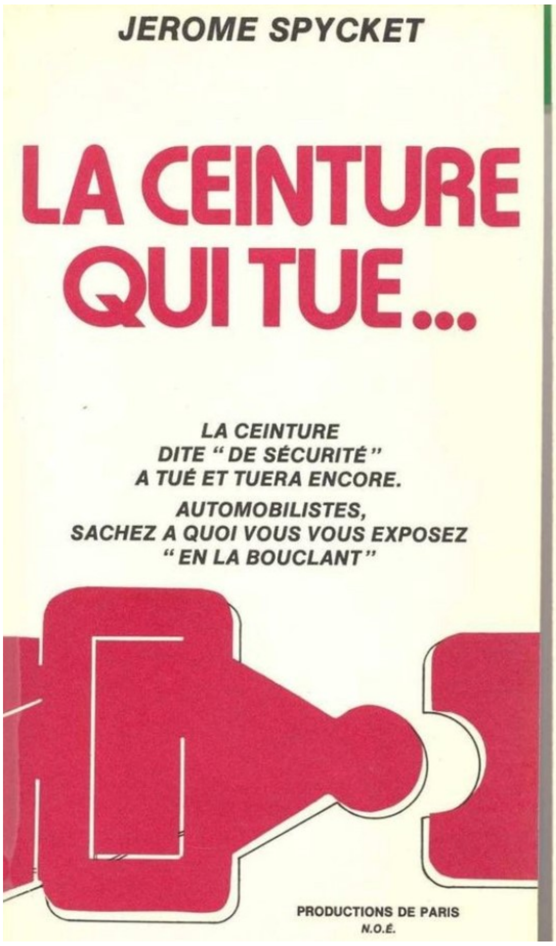
A famous example of personal patient risk appreciation is about industrial drugs or medicines usage risk! Indeed, it seems that on the $18,000$ drugs commercially released this last century around the World only approximately 177 of them (source: `https://en.wikipedia.org/wiki/List_of_withdrawn_drugs`) were withdrawn because of risks for patients, but also because of commercial reasons (e.g. lack of demand and relatively high production costs).

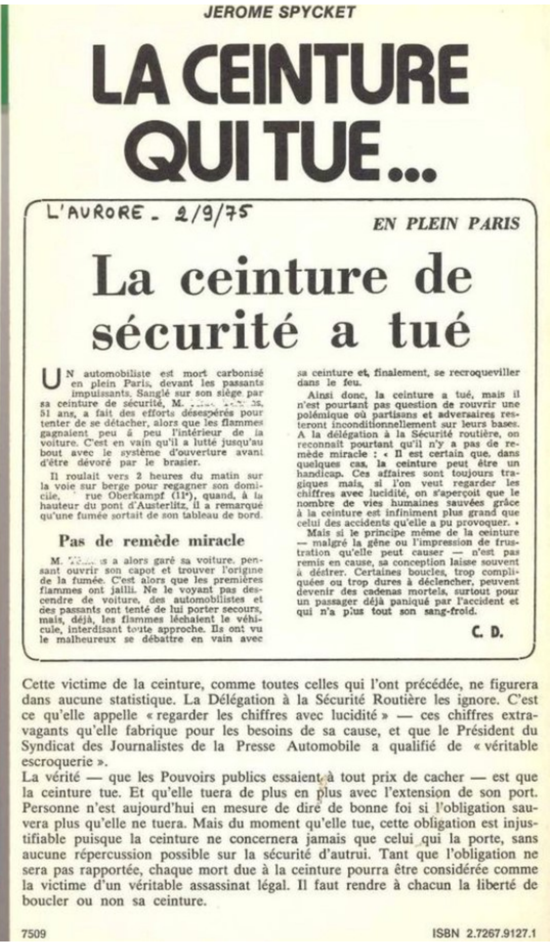
Let us take an example taken during the COVID-19 pandemic published the 2021-04-03 by FranceInfo (french news channel): According to the numbers that the EMA unveiled on Wednesday, there were at this stage 62 cases of cerebral venous thrombosis in the world, including 44 in the 30 countries of the European Economic Area (EU, Iceland, Norway, Liechtenstein) for 9.2 million doses of vaccine administered. 9 deaths have been recorded in Germany out of 2.8 millions AstraZeneca vaccine injected. In France 4 death have been reported out of 1.9 million injections, 3 death cases in Norway out of $120,000$ injections in the Netherlands. But as with any drug, knowing the risk is not enough, it must be compared with the benefits provided by the product. The Risk Ratio of the AstraZeneca vaccine in preventing COVID-19, which leads to hospitalizations and much more deaths, obviously outweigh the risks of side effects.

The fact that a fraction of the population doesn't know and doesn't understand (or don't want because they have a political or economic agenda) the concept of risk benefit ratio in the case of vaccines, medicines or any activity/action in daily life isn't something new as we can see it from the screenshot below of a 11975 (holocene calendar) french book where the author (which was an artist and evidence show us that we can't expect obviously any scientific literacy from the huge majority of people working in the artistic field) argue because some people died in their car because they had put their security belt is the proof that there shouldn't be any law to oblige people to put the security belt (...):

---

[28] A famous scientific experiment has provided evidence that at the contrary to the intuition that increasing dose makes something more poisonous, there exist sometimes non-monotonic (see section Functional Analysis page 2203) dose response curves (NMDRCs) demonstrated for natural hormones and endocrine disrupting chemicals (EDCs) like Bisphenonal A!


JEROME SPYCKET
LA CEINTURE QUI TUE...
LA CEINTURE DITE " DE SÉCURITÉ " A TUÉ ET TUERA ENCORE.
AUTOMOBILISTES, SACHEZ A QUOI VOUS VOUS EXPOSEZ " EN LA BOUCLANT "
PRODUCTIONS DE PARIS
N.O.É.


JEROME SPYCKET

LA CEINTURE QUI TUE...

L'AURORE - 2/9/75

EN PLEIN PARIS

La ceinture de sécurité a tué

UN automobiliste est mort carbonisé en plein Paris, devant les passants impuissants. Sanglé sur son siège par sa ceinture de sécurité, M. [redacted], 51 ans, a fait des efforts désespérés pour tenter de se détacher, alors que les flammes gagnaient peu à peu l'intérieur de la voiture. C'est en vain qu'il a lutté jusqu'au bout avec le système d'ouverture avant d'être dévoré par le brasier.

Il roulait vers 2 heures du matin sur la voie sur berge pour regagner son domicile, [redacted] rue Oberkampf (11e), quand, à la hauteur du pont d'Austerlitz, il a remarqué qu'une fumée sortait de son tableau de bord.

Pas de remède miracle

M. [redacted] a alors garé sa voiture, pensant ouvrir son capot et trouver l'origine de la fumée. C'est alors que les premières flammes ont jailli. Ne le voyant pas descendre de voiture, des automobilistes et des passants ont tenté de lui porter secours, mais, déjà, les flammes léchaient le véhicule, interdisant toute approche. Ils ont vu le malheureux se débattre en vain avec sa ceinture et, finalement, se recroqueviller dans le feu.

Ainsi donc, la ceinture a tué, mais il n'est pourtant pas question de rouvrir une polémique où partisans et adversaires resteront inconditionnellement sur leurs bases. A la délégation à la Sécurité routière, on reconnait pourtant qu'il n'y a pas de remède miracle : « Il est certain que, dans quelques cas, la ceinture peut être un handicap. Ces affaires sont toujours tragiques mais, si l'on veut regarder les chiffres avec lucidité, on s'aperçoit que le nombre de vies humaines sauvées grâce à la ceinture est infiniment plus grand que celui des accidents qu'elle a pu provoquer. »

Mais si le principe même de la ceinture — malgré la gêne ou l'impression de frustration qu'elle peut causer — n'est pas remis en cause, sa conception laisse souvent à désirer. Certaines boucles, trop compliquées ou trop dures à déclencher, peuvent devenir des cadenas mortels, surtout pour un passager déjà paniqué par l'accident et qui n'a plus tout son sang-froid.

C. D.

Cette victime de la ceinture, comme toutes celles qui l'ont précédée, ne figurera dans aucune statistique. La Délégation à la Sécurité Routière les ignore. C'est ce qu'elle appelle « regarder les chiffres avec lucidité » — ces chiffres extravagants qu'elle fabrique pour les besoins de sa cause, et que le Président du Syndicat des Journalistes de la Presse Automobile a qualifié de « véritable escroquerie ».

La vérité — que les Pouvoirs publics essaient à tout prix de cacher — est que la ceinture tue. Et qu'elle tuera de plus en plus avec l'extension de son port. Personne n'est aujourd'hui en mesure de dire de bonne foi si l'obligation sauvera plus qu'elle ne tuera. Mais du moment qu'elle tue, cette obligation est injustifiable puisque la ceinture ne concernera jamais que celui qui la porte, sans aucune répercussion possible sur la sécurité d'autrui. Tant que l'obligation ne sera pas rapportée, chaque mort due à la ceinture pourra être considérée comme la victime d'un véritable assassinat légal. Il faut rendre à chacun la liberté de boucler ou non sa ceinture.

7509 ISBN 2.7267.9127.1

14.1.12.6.7 ROC and Lift curves

ROC (Receiver Operating Characteristic) curve and Lift chart are techniques for visualizing, organizing, improving and selecting classifiers based on their performance. They facilitate our conception of classifiers and are therefore useful in research and in result presentation.

ROC curve is a plot of the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. It is used to evaluate the performance of a binary classifier.

Lift chart is a measure of the effectiveness of a predictive model. It compares the results of the model against a random guess and tells us how much better we are doing than random guessing.

14.1.12.6.7.1 ROC curve

Classifier performance evaluation should consider all the score range instead of a single point, unless specified by the application domain requirements. That is probably the reason why the area under the curve (AUC) of the Receiver Operating Characteristics (ROC) became so popular among scientists for binary classification quality assessment, instead of confusion matrix, error rates, maximum vertical distance of the ROC curve, maximum Kolmogorov-Smirnov vertical difference and all other single point specific metrics that are useful for the human brain but not very reliable.

The "ROC curve", that stands for "Receiver Operating Characteristic curve", is a measure of the performance of a binary classifier, that is to say a system that aims to categorize entities in two distinct groups on the basis of one or more of their characteristics.

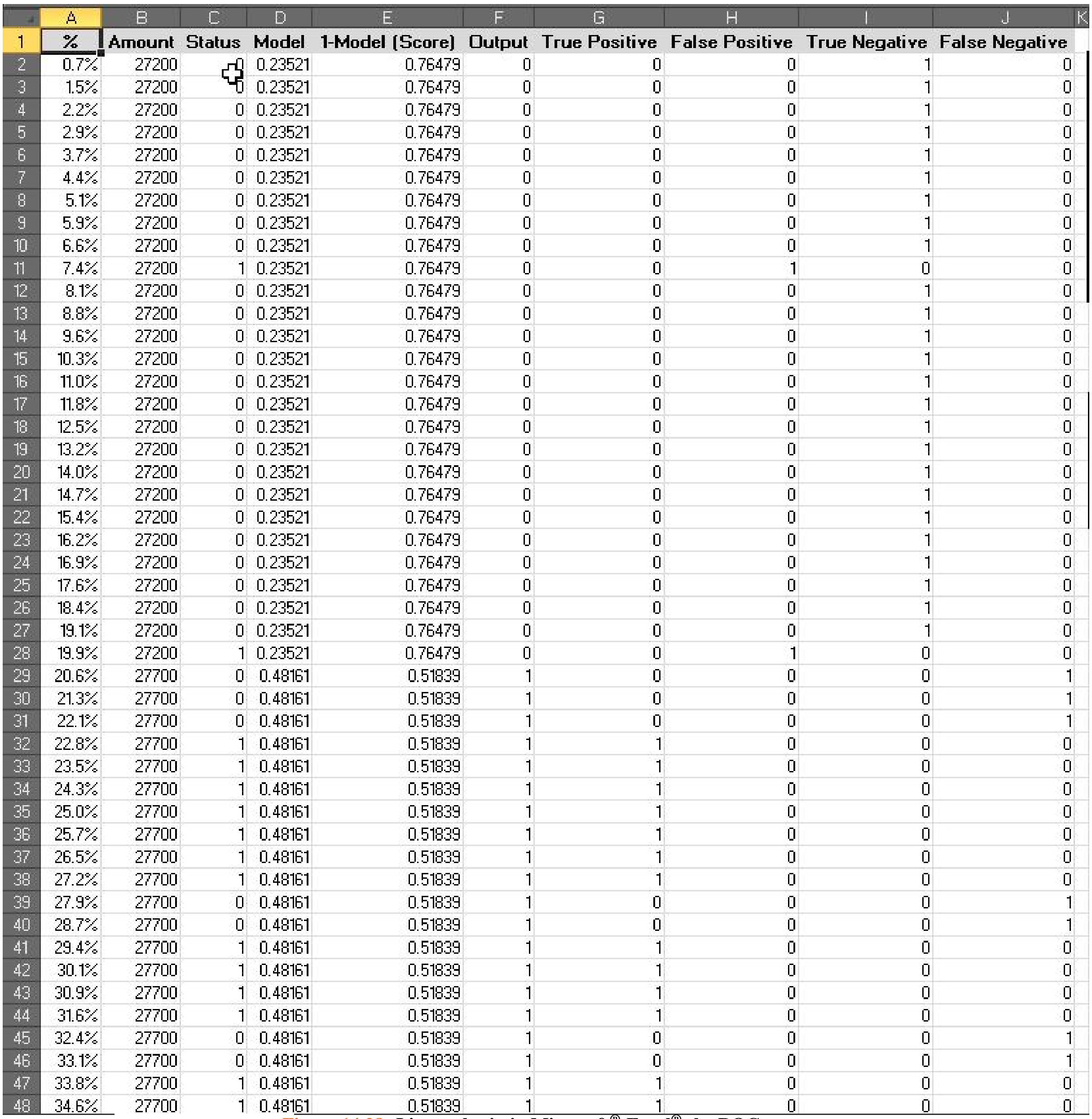
Graphically, we often represents the ROC measurement in the form of a curve which gives the true positive rate (sensitivity: fraction of positives that are correctly detected) according to the rate of false positive rate (proportion of negatives which are incorrectly detected) for the same group. ROC curves are often used in statistics to show the progress realized with a binary classifier when "discrimination threshold" (cutoff) varies.

Remark

We present this tool because many statistical software return automatically a ROC curve as output but the interest of this tool is in my personal opinion very questionable. However, we will show a more useful tool right after.

Let us see practical case by taking our previous example (the Bank and its subsidiaries) with the spreadsheet software Microsoft® Excel® 14.0.7106. First to build the ROC curve we must obtain the confusion matrix C that we presented just a little higher with Minitab. For this with a spreadsheet software and without doing code, here is a simple solution (but this is not the more condensed solution for pedagogical reasons).

The goal will be first to build the following table (this is only a part of the table of the previous example since in reality there are 136 rows of data) whose columns `B` and `C` comes from the three small tables of Credits used just before, column `A` is just the cumulative frequency of individuals $1/136$ ... $2/136$ ... $3/136$ ... and so on):



	A	B	C	D	E	F	G	H	I	J	K
1	%	Amount	Status	Model	1-Model (Score)	Output	True Positive	False Positive	True Negative	False Negative	
2	0.7%	27200	0	0.23521	0.76479	0	0	0	1	0	
3	1.5%	27200	0	0.23521	0.76479	0	0	0	1	0	
4	2.2%	27200	0	0.23521	0.76479	0	0	0	1	0	
5	2.9%	27200	0	0.23521	0.76479	0	0	0	1	0	
6	3.7%	27200	0	0.23521	0.76479	0	0	0	1	0	
7	4.4%	27200	0	0.23521	0.76479	0	0	0	1	0	
8	5.1%	27200	0	0.23521	0.76479	0	0	0	1	0	
9	5.9%	27200	0	0.23521	0.76479	0	0	0	1	0	
10	6.6%	27200	0	0.23521	0.76479	0	0	0	1	0	
11	7.4%	27200	1	0.23521	0.76479	0	0	1	0	0	
12	8.1%	27200	0	0.23521	0.76479	0	0	0	1	0	
13	8.8%	27200	0	0.23521	0.76479	0	0	0	1	0	
14	9.6%	27200	0	0.23521	0.76479	0	0	0	1	0	
15	10.3%	27200	0	0.23521	0.76479	0	0	0	1	0	
16	11.0%	27200	0	0.23521	0.76479	0	0	0	1	0	
17	11.8%	27200	0	0.23521	0.76479	0	0	0	1	0	
18	12.5%	27200	0	0.23521	0.76479	0	0	0	1	0	
19	13.2%	27200	0	0.23521	0.76479	0	0	0	1	0	
20	14.0%	27200	0	0.23521	0.76479	0	0	0	1	0	
21	14.7%	27200	0	0.23521	0.76479	0	0	0	1	0	
22	15.4%	27200	0	0.23521	0.76479	0	0	0	1	0	
23	16.2%	27200	0	0.23521	0.76479	0	0	0	1	0	
24	16.9%	27200	0	0.23521	0.76479	0	0	0	1	0	
25	17.6%	27200	0	0.23521	0.76479	0	0	0	1	0	
26	18.4%	27200	0	0.23521	0.76479	0	0	0	1	0	
27	19.1%	27200	0	0.23521	0.76479	0	0	0	1	0	
28	19.9%	27200	1	0.23521	0.76479	0	0	1	0	0	
29	20.6%	27700	0	0.48161	0.51839	1	0	0	0	1	
30	21.3%	27700	0	0.48161	0.51839	1	0	0	0	1	
31	22.1%	27700	0	0.48161	0.51839	1	0	0	0	1	
32	22.8%	27700	1	0.48161	0.51839	1	1	0	0	0	
33	23.5%	27700	1	0.48161	0.51839	1	1	0	0	0	
34	24.3%	27700	1	0.48161	0.51839	1	1	0	0	0	
35	25.0%	27700	1	0.48161	0.51839	1	1	0	0	0	
36	25.7%	27700	1	0.48161	0.51839	1	1	0	0	0	
37	26.5%	27700	1	0.48161	0.51839	1	1	0	0	0	
38	27.2%	27700	1	0.48161	0.51839	1	1	0	0	0	
39	27.9%	27700	0	0.48161	0.51839	1	0	0	0	1	
40	28.7%	27700	0	0.48161	0.51839	1	0	0	0	1	
41	29.4%	27700	1	0.48161	0.51839	1	1	0	0	0	
42	30.1%	27700	1	0.48161	0.51839	1	1	0	0	0	
43	30.9%	27700	1	0.48161	0.51839	1	1	0	0	0	
44	31.6%	27700	1	0.48161	0.51839	1	1	0	0	0	
45	32.4%	27700	0	0.48161	0.51839	1	0	0	0	1	
46	33.1%	27700	0	0.48161	0.51839	1	0	0	0	1	
47	33.8%	27700	1	0.48161	0.51839	1	1	0	0	0	
48	34.6%	27700	1	0.48161	0.51839	1	1	0	0	0	

Figure 14.28: **List to obtain in Microsoft® Excel® the ROC curve**

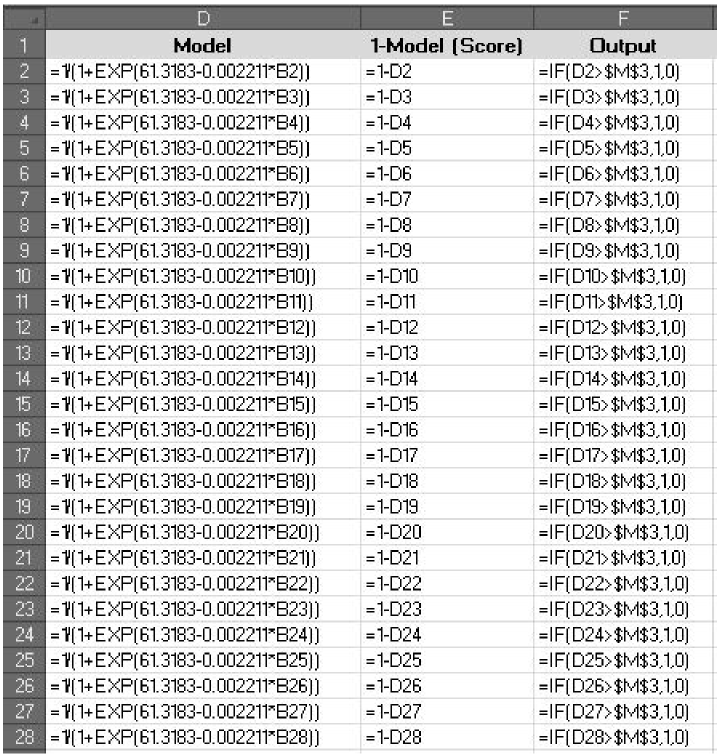
Now let us see what is in columns `D`, `E` and `F` and that are directly related to the result obtained above, which was for recall:

$$P(Y)=\frac{1}{1-e^{-(-15.1116+0.001837x)}} \tag{14.1.979}$$

but that we have refined with a specialized statistical software to obtain:

$$P(Y)=\frac{1}{1-e^{-(-61.3183+0.002211x)}} \tag{14.1.980}$$

which then allows us to build the three previous mentioned columns (here we have only a few of the first rows because it is sufficient to just increment the formulas to the bottom of the list):

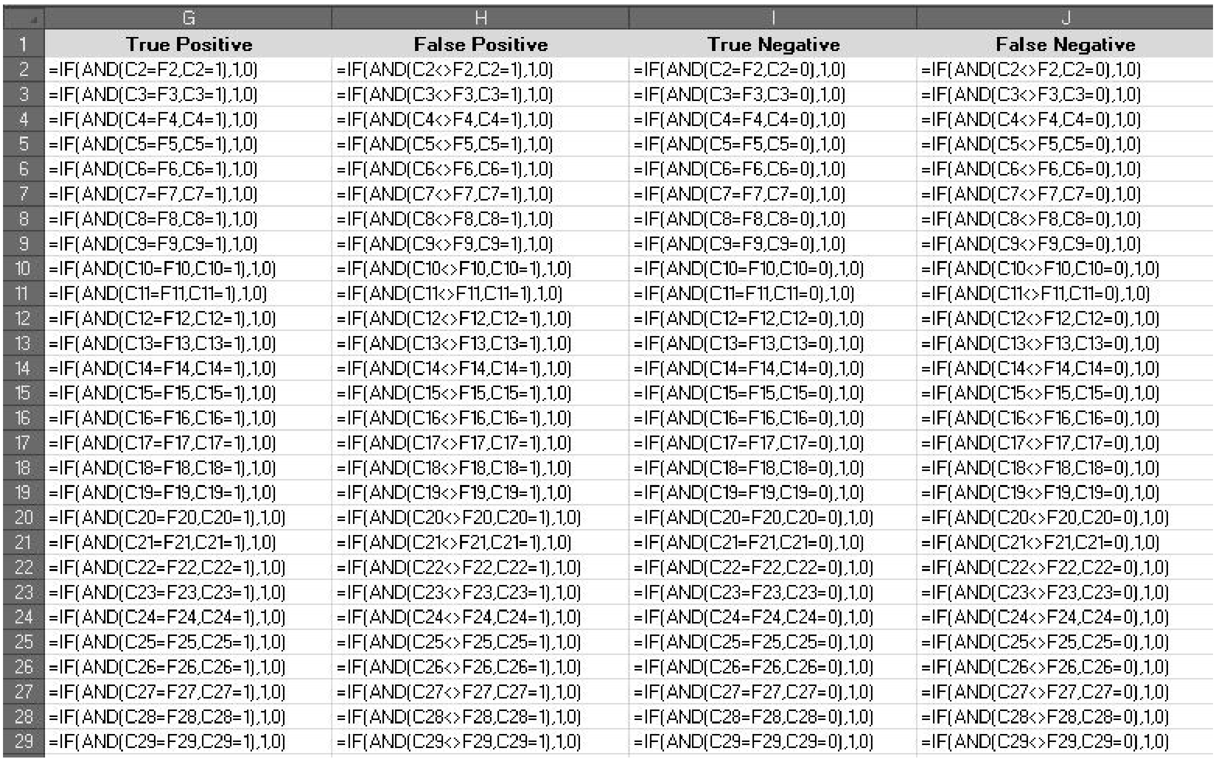


	D	E	F
1	**Model**	**1-Model (Score)**	**Output**
2	=1/(1+EXP(61.3183-0.002211*B2))	=1-D2	=IF(D2>$M$3,1,0)
3	=1/(1+EXP(61.3183-0.002211*B3))	=1-D3	=IF(D3>$M$3,1,0)
4	=1/(1+EXP(61.3183-0.002211*B4))	=1-D4	=IF(D4>$M$3,1,0)
5	=1/(1+EXP(61.3183-0.002211*B5))	=1-D5	=IF(D5>$M$3,1,0)
6	=1/(1+EXP(61.3183-0.002211*B6))	=1-D6	=IF(D6>$M$3,1,0)
7	=1/(1+EXP(61.3183-0.002211*B7))	=1-D7	=IF(D7>$M$3,1,0)
8	=1/(1+EXP(61.3183-0.002211*B8))	=1-D8	=IF(D8>$M$3,1,0)
9	=1/(1+EXP(61.3183-0.002211*B9))	=1-D9	=IF(D9>$M$3,1,0)
10	=1/(1+EXP(61.3183-0.002211*B10))	=1-D10	=IF(D10>$M$3,1,0)
11	=1/(1+EXP(61.3183-0.002211*B11))	=1-D11	=IF(D11>$M$3,1,0)
12	=1/(1+EXP(61.3183-0.002211*B12))	=1-D12	=IF(D12>$M$3,1,0)
13	=1/(1+EXP(61.3183-0.002211*B13))	=1-D13	=IF(D13>$M$3,1,0)
14	=1/(1+EXP(61.3183-0.002211*B14))	=1-D14	=IF(D14>$M$3,1,0)
15	=1/(1+EXP(61.3183-0.002211*B15))	=1-D15	=IF(D15>$M$3,1,0)
16	=1/(1+EXP(61.3183-0.002211*B16))	=1-D16	=IF(D16>$M$3,1,0)
17	=1/(1+EXP(61.3183-0.002211*B17))	=1-D17	=IF(D17>$M$3,1,0)
18	=1/(1+EXP(61.3183-0.002211*B18))	=1-D18	=IF(D18>$M$3,1,0)
19	=1/(1+EXP(61.3183-0.002211*B19))	=1-D19	=IF(D19>$M$3,1,0)
20	=1/(1+EXP(61.3183-0.002211*B20))	=1-D20	=IF(D20>$M$3,1,0)
21	=1/(1+EXP(61.3183-0.002211*B21))	=1-D21	=IF(D21>$M$3,1,0)
22	=1/(1+EXP(61.3183-0.002211*B22))	=1-D22	=IF(D22>$M$3,1,0)
23	=1/(1+EXP(61.3183-0.002211*B23))	=1-D23	=IF(D23>$M$3,1,0)
24	=1/(1+EXP(61.3183-0.002211*B24))	=1-D24	=IF(D24>$M$3,1,0)
25	=1/(1+EXP(61.3183-0.002211*B25))	=1-D25	=IF(D25>$M$3,1,0)
26	=1/(1+EXP(61.3183-0.002211*B26))	=1-D26	=IF(D26>$M$3,1,0)
27	=1/(1+EXP(61.3183-0.002211*B27))	=1-D27	=IF(D27>$M$3,1,0)
28	=1/(1+EXP(61.3183-0.002211*B28))	=1-D28	=IF(D28>$M$3,1,0)

Figure 14.29: **Spreadsheet formulas for the probability of the model and the score**

The formula in column `F` refers to the cell `M3` which as we shall see a little further contains the empirical choice of the value of the cutoff we had mentioned during the presentation of the theoretical model of the logistics regression (by default we defined it at 50%).

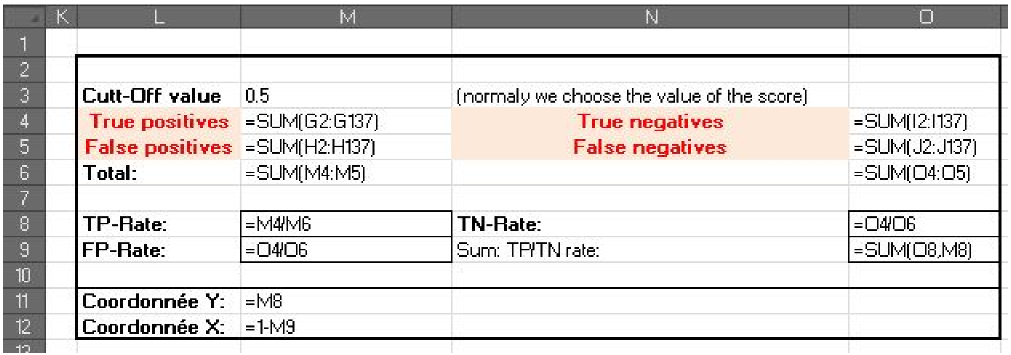
Then we must build the columns of true positive, false positive, true negative, false negatives using basic spreadsheet formulas:



	G	H	I	J
1	True Positive	False Positive	True Negative	False Negative
2	=IF(AND(C2=F2,C2=1),1,0)	=IF(AND(C2<>F2,C2=1),1,0)	=IF(AND(C2=F2,C2=0),1,0)	=IF(AND(C2<>F2,C2=0),1,0)
3	=IF(AND(C3=F3,C3=1),1,0)	=IF(AND(C3<>F3,C3=1),1,0)	=IF(AND(C3=F3,C3=0),1,0)	=IF(AND(C3<>F3,C3=0),1,0)
4	=IF(AND(C4=F4,C4=1),1,0)	=IF(AND(C4<>F4,C4=1),1,0)	=IF(AND(C4=F4,C4=0),1,0)	=IF(AND(C4<>F4,C4=0),1,0)
5	=IF(AND(C5=F5,C5=1),1,0)	=IF(AND(C5<>F5,C5=1),1,0)	=IF(AND(C5=F5,C5=0),1,0)	=IF(AND(C5<>F5,C5=0),1,0)
6	=IF(AND(C6=F6,C6=1),1,0)	=IF(AND(C6<>F6,C6=1),1,0)	=IF(AND(C6=F6,C6=0),1,0)	=IF(AND(C6<>F6,C6=0),1,0)
7	=IF(AND(C7=F7,C7=1),1,0)	=IF(AND(C7<>F7,C7=1),1,0)	=IF(AND(C7=F7,C7=0),1,0)	=IF(AND(C7<>F7,C7=0),1,0)
8	=IF(AND(C8=F8,C8=1),1,0)	=IF(AND(C8<>F8,C8=1),1,0)	=IF(AND(C8=F8,C8=0),1,0)	=IF(AND(C8<>F8,C8=0),1,0)
9	=IF(AND(C9=F9,C9=1),1,0)	=IF(AND(C9<>F9,C9=1),1,0)	=IF(AND(C9=F9,C9=0),1,0)	=IF(AND(C9<>F9,C9=0),1,0)
10	=IF(AND(C10=F10,C10=1),1,0)	=IF(AND(C10<>F10,C10=1),1,0)	=IF(AND(C10=F10,C10=0),1,0)	=IF(AND(C10<>F10,C10=0),1,0)
11	=IF(AND(C11=F11,C11=1),1,0)	=IF(AND(C11<>F11,C11=1),1,0)	=IF(AND(C11=F11,C11=0),1,0)	=IF(AND(C11<>F11,C11=0),1,0)
12	=IF(AND(C12=F12,C12=1),1,0)	=IF(AND(C12<>F12,C12=1),1,0)	=IF(AND(C12=F12,C12=0),1,0)	=IF(AND(C12<>F12,C12=0),1,0)
13	=IF(AND(C13=F13,C13=1),1,0)	=IF(AND(C13<>F13,C13=1),1,0)	=IF(AND(C13=F13,C13=0),1,0)	=IF(AND(C13<>F13,C13=0),1,0)
14	=IF(AND(C14=F14,C14=1),1,0)	=IF(AND(C14<>F14,C14=1),1,0)	=IF(AND(C14=F14,C14=0),1,0)	=IF(AND(C14<>F14,C14=0),1,0)
15	=IF(AND(C15=F15,C15=1),1,0)	=IF(AND(C15<>F15,C15=1),1,0)	=IF(AND(C15=F15,C15=0),1,0)	=IF(AND(C15<>F15,C15=0),1,0)
16	=IF(AND(C16=F16,C16=1),1,0)	=IF(AND(C16<>F16,C16=1),1,0)	=IF(AND(C16=F16,C16=0),1,0)	=IF(AND(C16<>F16,C16=0),1,0)
17	=IF(AND(C17=F17,C17=1),1,0)	=IF(AND(C17<>F17,C17=1),1,0)	=IF(AND(C17=F17,C17=0),1,0)	=IF(AND(C17<>F17,C17=0),1,0)
18	=IF(AND(C18=F18,C18=1),1,0)	=IF(AND(C18<>F18,C18=1),1,0)	=IF(AND(C18=F18,C18=0),1,0)	=IF(AND(C18<>F18,C18=0),1,0)
19	=IF(AND(C19=F19,C19=1),1,0)	=IF(AND(C19<>F19,C19=1),1,0)	=IF(AND(C19=F19,C19=0),1,0)	=IF(AND(C19<>F19,C19=0),1,0)
20	=IF(AND(C20=F20,C20=1),1,0)	=IF(AND(C20<>F20,C20=1),1,0)	=IF(AND(C20=F20,C20=0),1,0)	=IF(AND(C20<>F20,C20=0),1,0)
21	=IF(AND(C21=F21,C21=1),1,0)	=IF(AND(C21<>F21,C21=1),1,0)	=IF(AND(C21=F21,C21=0),1,0)	=IF(AND(C21<>F21,C21=0),1,0)
22	=IF(AND(C22=F22,C22=1),1,0)	=IF(AND(C22<>F22,C22=1),1,0)	=IF(AND(C22=F22,C22=0),1,0)	=IF(AND(C22<>F22,C22=0),1,0)
23	=IF(AND(C23=F23,C23=1),1,0)	=IF(AND(C23<>F23,C23=1),1,0)	=IF(AND(C23=F23,C23=0),1,0)	=IF(AND(C23<>F23,C23=0),1,0)
24	=IF(AND(C24=F24,C24=1),1,0)	=IF(AND(C24<>F24,C24=1),1,0)	=IF(AND(C24=F24,C24=0),1,0)	=IF(AND(C24<>F24,C24=0),1,0)
25	=IF(AND(C25=F25,C25=1),1,0)	=IF(AND(C25<>F25,C25=1),1,0)	=IF(AND(C25=F25,C25=0),1,0)	=IF(AND(C25<>F25,C25=0),1,0)
26	=IF(AND(C26=F26,C26=1),1,0)	=IF(AND(C26<>F26,C26=1),1,0)	=IF(AND(C26=F26,C26=0),1,0)	=IF(AND(C26<>F26,C26=0),1,0)
27	=IF(AND(C27=F27,C27=1),1,0)	=IF(AND(C27<>F27,C27=1),1,0)	=IF(AND(C27=F27,C27=0),1,0)	=IF(AND(C27<>F27,C27=0),1,0)
28	=IF(AND(C28=F28,C28=1),1,0)	=IF(AND(C28<>F28,C28=1),1,0)	=IF(AND(C28=F28,C28=0),1,0)	=IF(AND(C28<>F28,C28=0),1,0)
29	=IF(AND(C29=F29,C29=1),1,0)	=IF(AND(C29<>F29,C29=1),1,0)	=IF(AND(C29=F29,C29=0),1,0)	=IF(AND(C29<>F29,C29=0),1,0)

Figure 14.30: Spreadsheet formulas for true / false positives and negatives

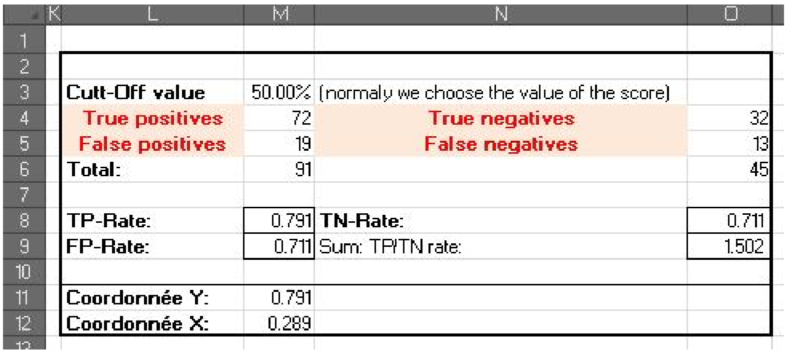
Once we have this data, we can rebuild the confusion matrix $\mathcal{C}$ that was given to us by Minitab and we will need to develop the ROC curve:



	K	L	M	N	O
1					
2					
3		Cutt-Off value	0.5	(normaly we choose the value of the score)	
4		True positives	=SUM(G2:G137)	True negatives	=SUM(I2:I137)
5		False positives	=SUM(H2:H137)	False negatives	=SUM(J2:J137)
6		Total:	=SUM(M4:M5)		=SUM(O4:O5)
7					
8		TP-Rate:	=M4/M6	TN-Rate:	=O4/O6
9		FP-Rate:	=O4/O6	Sum: TP/TN rate:	=SUM(O8,M8)
10					
11		Coordonnée Y:	=M8		
12		Coordonnée X:	=1-M9		

Figure 14.31: Spreadsheet formulas to build the confusion matrix

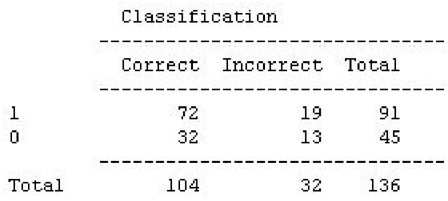
which explicitly gives:



	K	L	M	N	O
1					
2					
3		Cutt-Off value	50.00%	(normaly we choose the value of the score)	
4		True positives	72	True negatives	32
5		False positives	19	False negatives	13
6		Total:	91		45
7					
8		TP-Rate:	0.791	TN-Rate:	0.711
9		FP-Rate:	0.711	Sum: TP/TN rate:	1.502
10					
11		Coordonnée Y:	0.791		
12		Coordonnée X:	0.289		

Figure 14.32: Corresponding values of the construction of the confusion matrix

compared with the confusion matrix that we gave and coming from Minitab:



```
                 Classification
               ------------------------------
                 Correct  Incorrect  Total
               ------------------------------
1                    72         19      91
0                    32         13      45
               ------------------------------
Total               104         32     136
```

Figure 14.33: **Confusion matrix with Minitab 16.1.1**

Now observe the ROC curve given by Minitab:

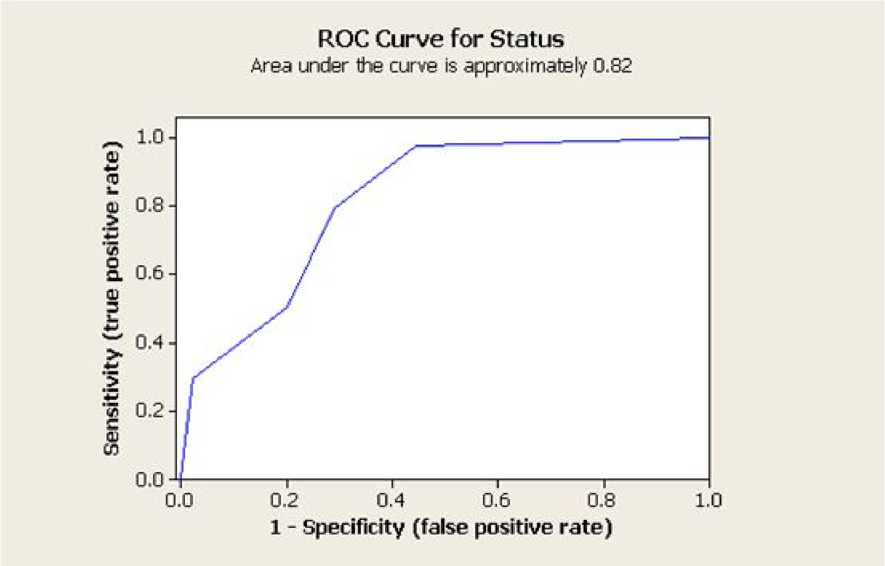


Figure 14.34: **ROC curve with Minitab 16.1.1**

Why the abscissa-represents:

$$1-\%\text{false positives}$$

you will probably ask yourself when the interpretation would have been easier if we had just the:

$$\%\text{false positives}$$

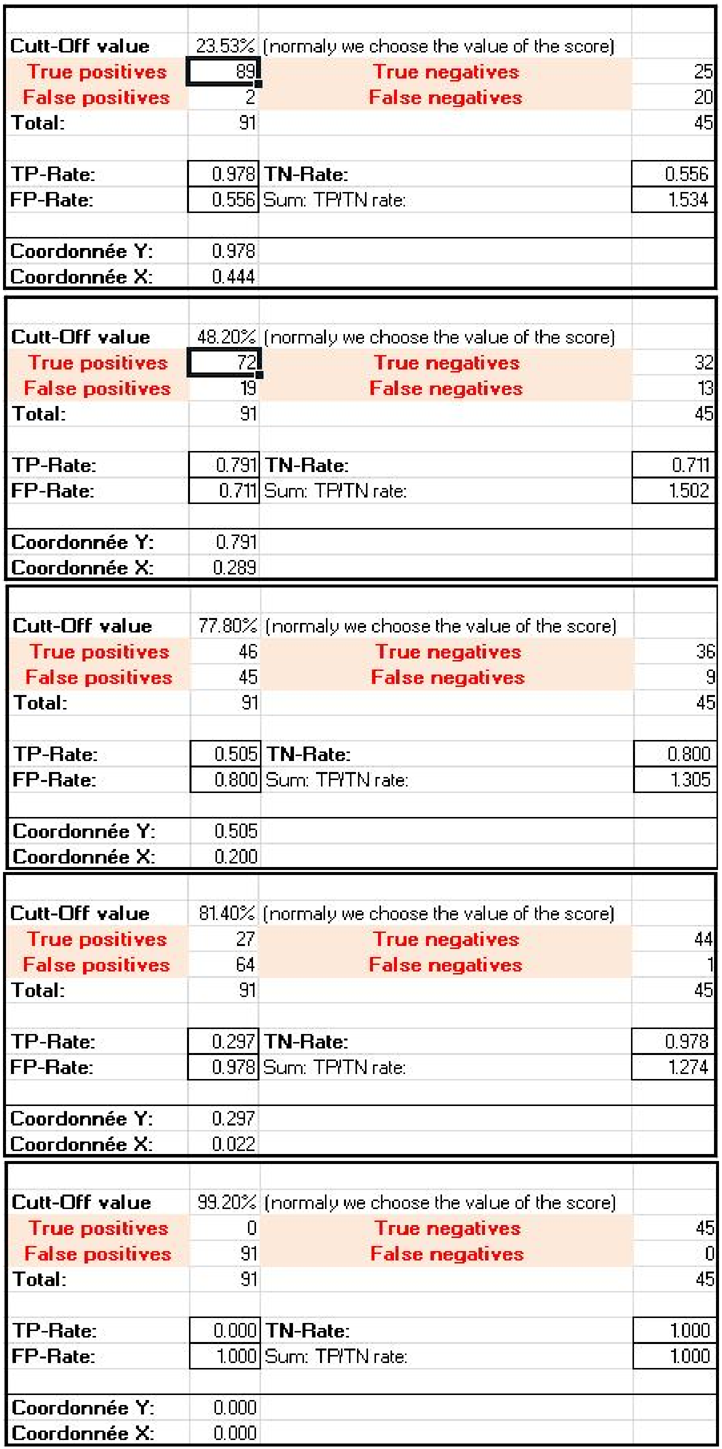
Well for two reasons: the first is that practitioners like strictly increasing functions... and the second reason, the most important, is that if the binary classifier is fully effective, the surface is then equal to 1(that is to say, 100%). Which admittedly is nicer than to say that a null surface corresponds to a 100% efficiency. Well that being said ... let us continue.

Before learning how to interpret this chart, how to get the same curve in a spreadsheet software? Well just simply observe our column `D` of our Microsoft® Excel® list. Logically we have there

5 different cumulative probabilities (as there were only 5 credits values) which are respectively:

$$0.235214057$$
$$0.481608302$$
$$0.777818743$$
$$0.813679766$$
$$0.991760974$$

Therefore the idea of the construction of the ROC curve is to take each of the accumulated probabilities as cutt-off values. This gives respectively:



Cutt-Off value	23.53%	(normaly we choose the value of the score)	
True positives	89	**True negatives**	25
False positives	2	**False negatives**	20
Total:	91		45
TP-Rate:	0.978	**TN-Rate:**	0.556
FP-Rate:	0.556	Sum: TP/TN rate:	1.534
Coordonnée Y:	0.978		
Coordonnée X:	0.444		

Cutt-Off value	48.20%	(normaly we choose the value of the score)	
True positives	72	**True negatives**	32
False positives	19	**False negatives**	13
Total:	91		45
TP-Rate:	0.791	**TN-Rate:**	0.711
FP-Rate:	0.711	Sum: TP/TN rate:	1.502
Coordonnée Y:	0.791		
Coordonnée X:	0.289		

Cutt-Off value	77.80%	(normaly we choose the value of the score)	
True positives	46	**True negatives**	36
False positives	45	**False negatives**	9
Total:	91		45
TP-Rate:	0.505	**TN-Rate:**	0.800
FP-Rate:	0.800	Sum: TP/TN rate:	1.305
Coordonnée Y:	0.505		
Coordonnée X:	0.200		

Cutt-Off value	81.40%	(normaly we choose the value of the score)	
True positives	27	**True negatives**	44
False positives	64	**False negatives**	1
Total:	91		45
TP-Rate:	0.297	**TN-Rate:**	0.978
FP-Rate:	0.978	Sum: TP/TN rate:	1.274
Coordonnée Y:	0.297		
Coordonnée X:	0.022		

Cutt-Off value	99.20%	(normaly we choose the value of the score)	
True positives	0	**True negatives**	45
False positives	91	**False negatives**	0
Total:	91		45
TP-Rate:	0.000	**TN-Rate:**	1.000
FP-Rate:	1.000	Sum: TP/TN rate:	1.000
Coordonnée Y:	0.000		
Coordonnée X:	0.000		

Figure 14.35: Values of the confusion matrix for various cutoff values

So we get each time the coordinates of points of the ROC curve in function of cutoff values. Well this done, now let us turn to the interpretation by taking the below manually completed graph:

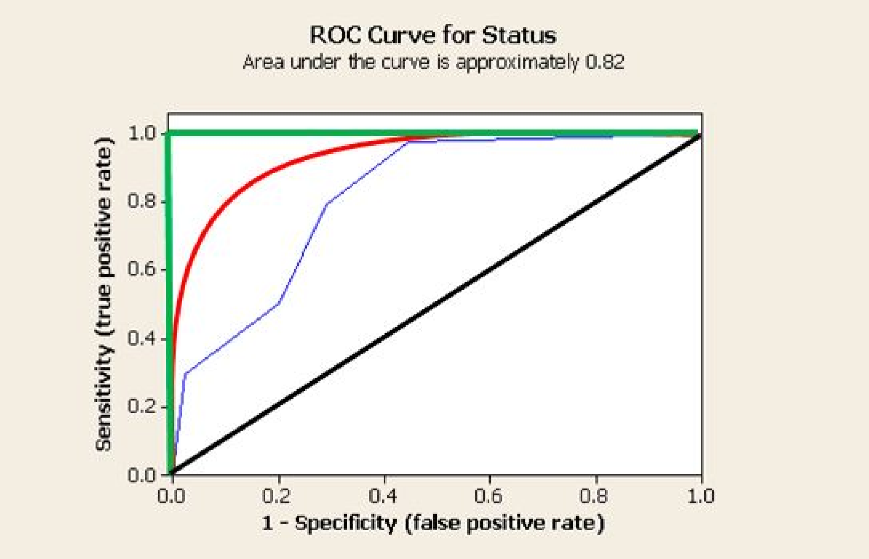


Figure 14.36: **ROC curves**
(blue: average classifier, red: good classifier, green: perfect classifier, black: poor classifier)

Credit Amount	Proportion P	Cutt-Off Value	x Coordinate of ROC Curve point	y Coordinate of ROC Curve point
			0	0
$27,200$	0.0741	23.53%	0	0
$27,700$	0.7083	48.20%	0.022	0.297
$28,300$	0.8667	77.80%	0.200	0.505
$28,400$	0.7037	81.40%	0.289	0.791
$29,900$	0.9643	99.2%	0.444	0.978
			1	1

Then as shown in the graph above and the intuition, a perfect binary classifier is one whose true positive rate is constant and always equal to 100%. So the binary classifier in our example is moderately good.

Remark

The biggest vertical distance between the black line in the figure above and the blue ROC curved is sometimes used as a quality metric for classification named the "maximum vertical distance" and abbreviated MVD.

But from my personal point of view a good binary classifier is one that for a given value of the cutoff maximizes the sum of the rate of true positives and true negatives. Thus, with a spreadsheet software such as Microsoft® Excel® in "evolutionary mode" it is very easy to find this cutoff value that maximizes this objective. But however be careful to the trap: the value may not be unique, it can also be an interval (which is the case in our example!).

Finally, if we denote by TP the true positive, TN the true negative, FP the false positives or Type I error (in terms of decision theory or tests theory), and FN false negative or Type II error we can then define a set of indicators to judge the quality of our predictor/classifier (or rather of our score):

- $\text{TPR} = \dfrac{\text{TP}}{\text{P}} = \dfrac{\text{TP}}{\text{TP+FN}}$ named "sensitivity", corresponding to the rate of true positives.
- $\text{RPF} = \dfrac{\text{FP}}{\text{N}} = \dfrac{\text{FP}}{\text{FP+TN}}$ corresponding to the rate of false positives.
- $\text{ACC} = \dfrac{\text{TP+TN}}{\text{P+N}}$ named "accuracy" (fraction of instances that are correctly classified).
- $\text{SPC} = \dfrac{\text{TN}}{\text{N}} = \dfrac{\text{TN}}{\text{FP+TN}} = 1 - \text{FPR}$ named "specificity" or true negative rate.
- $\text{PPV} = \dfrac{\text{TP}}{\text{TP+TP}}$ the "positive predictive value".
- $\text{NPV} = \dfrac{\text{TN}}{\text{TN+FN}}$ the "negative predictive value".
- $\text{FDR} = \dfrac{\text{FP}}{\text{FP+TP}}$ corresponding to the "false discovery rate".

And we can build many quality metrics like (for the most famous one!) the Matthews correlation coefficient (see section Statistics page 1254), the F_1 score (see page 6482) or the "Rand Index" (also sometimes named "model accuracy") defined by:

$$\text{RI} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{14.1.981}$$

Remark

In practice, the sensitivity and specificity of a test give an assessment of its intrinsic validity. It is considered in practice that taken separately, they do not mean anything. For example, it should be obvious for the reader a test with 95% sensitivity has no value if its specificity is only 5%. It is considered in practice that if the sum of sensitivity and specificity is equal to 100% the test is as good a random.

✍**Example:**

Consider a disease test sensibility and sensitivity analysis with the following values:

		Patients with bowel cancer (as confirmed by endoscopy)		
		Positive	**Negative**	
Disease Test	**Positive**	TP$=2$	FP$=18$	$\frac{\text{TP}}{\text{TP}+\text{FP}}$ $=\frac{2}{2+18}$ $=10\%$
	Negative	FN$=1$	TN$=182$	$\frac{\text{TN}}{\text{TN}+\text{FN}}$ $=\frac{182}{1+182}$ $=99.5\%$
		Sensitivity $=\frac{\text{TP}}{\text{TP}+\text{FN}}$ $=\frac{2}{2+1}=66.\bar{6}\%$	Specificity $=\frac{\text{TN}}{\text{FP}+\text{TN}}$ $=\frac{182}{18+182}=91\%$	

Table 14.15: **Example of sensibility and specificity analysis**

There are some cases where sensitivity is important and need to be near to 1. There are business cases where specificity is important and need to be near to 1 (ideally we want to maximize both sensitivity & specificity, but this is not always possible always). We need to understand the business problem and decide the importance of sensitivity and specificity.

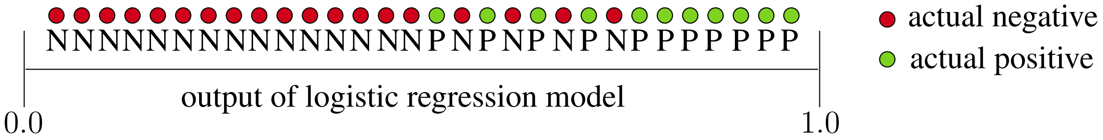
Remark

Current generation HIV tests are considered extremely accurate. The blood-based HIV ELISA has a demonstrated sensitivity of between 99.3% to 99.7 percent, with a specificity of between 99.91% and 99.97%. When combined with a Western blot, this translates to approximately one false positive out of every $250,000$ tests in the general U.S. population. Newer, fourth generation combination tests, which test for both HIV antibodies and antigens, are reported to have a clinical sensitivity of 99.9%.

As you may have notice in the previous chart above. Below the title of the chart there was a mention of something named "Area Under the Curve" and that is well known under the abbreviation AUC.

Obviously, the names mean what it is: It measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from $[0,0]$ to $[1,1]$.

AUC provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example. For example, given the following examples, which are arranged from left to right in ascending order of logistic regression predictions:


N N N N N N N N N N N N N N N P N P N P N P N P P P P P P P
output of logistic regression model
0.0
1.0
● actual negative
● actual positive

AUC represents the probability that a random positive (green) example is positioned to the right of a random negative (red) example.

Example:

If we have a model with an AUC of 80%, it means that if we choose randomly two individuals in the population, one negative and the other positive, the model will classify correctly these two individuals in their respective groups 8 times on 10.

The corresponding mathematical expression is quite obviously given by:

$$\text{AUC} = P\left(X_{(p)} > X_{(n)}\right) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} 1_{p_i > p_j} \tag{14.1.982}$$

Here i runs over all m data points with true label 1, and j runs over all n data points with true label 0; p_i and p_j denote the probability score assigned by the classifier to data point i and j, respectively. 1 is always the indicator function: it outputs 1 if and only if the condition $(p_i > p_j)$ is satisfied!

AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0. A model that doesn't better than chance will have an AUC of less than 50%.

Testing the fact evidence that the AUC is significantly different of 0.5 is like the null hypothesis H_0: *The distribution of the ranks in the two groups are equal*. It's then the equivalent of running a Wilcoxon rank sum test (see section Statistics page 1160).

Remark

In logistic regression if we consider then probabilities $\left\{\left(\hat{\pi}(\hat{\vec{\beta}}|\vec{x}_i\right), \hat{\pi}\left(\hat{\vec{\beta}}, \vec{x}_j\right)\right\}$, the C-index can be defined as:

$$C = P\left[\hat{\pi}\left(\hat{\vec{\beta}}|\vec{x}_i\right) > \hat{\pi}\left(\hat{\vec{\beta}}|\vec{x}_i\right)\middle| Y_i = 1 \& Y_j = 0\right] \tag{14.1.983}$$

Which is equivalent to:

$$C = P\left[\hat{\vec{\beta}} \circ \vec{x}_i > \hat{\vec{\beta}} \circ \vec{x}_j \middle| Y_i = 1 \& Y_j = 0\right] \tag{14.1.984}$$

AUC is desirable for the following two reasons:

- AUC is scale-invariant. It measures how well predictions are ranked, independently of their absolute values.
- AUC is classification-threshold-invariant. It measures the quality of the model's predictions irrespective of what classification threshold is chosen.

Notice that there is a simple relation between the AUC and the Gini index. Indeed, as we have proved it in the section of Quantitative Management (see page 8099), we have:

$$G=2A \tag{14.1.985}$$

But as:

$$\text{AUC}=A+\frac{1}{2} \tag{14.1.986}$$

We then have:

$$G=2\cdot\text{AUC}-1 \tag{14.1.987}$$

Remark

In multi-class models, we can plot N number of AUC ROC curves for N number classes using the "One vs All methodology". So for example, if you have three classes named A, B and C, you will have one ROC for A classified against B and C, another ROC for B classified against A and B, and a third one C classified against A and B.

Having done this, let us move to the second curve.

14.1.12.6.7.2 Lift curve

The principle of the "lift curve" is very simple but is based on a questionable assumption (which fortunately has not however a great importance) that is that without the theoretical model (logistic model in this case) we will absolutely know nothing of the real behaviour that will have the customers (debtors). The hypothesis is to consider that if we took 50% of the individuals in our sample, we would have 50% of true positives (real bad debtors), if we took 25% of our sample, we would have 25% of true positives (true bad debtors) and so on. This initial hypothesis (considered by practitioners as the worst case) will be represented graphically by the following straight line:

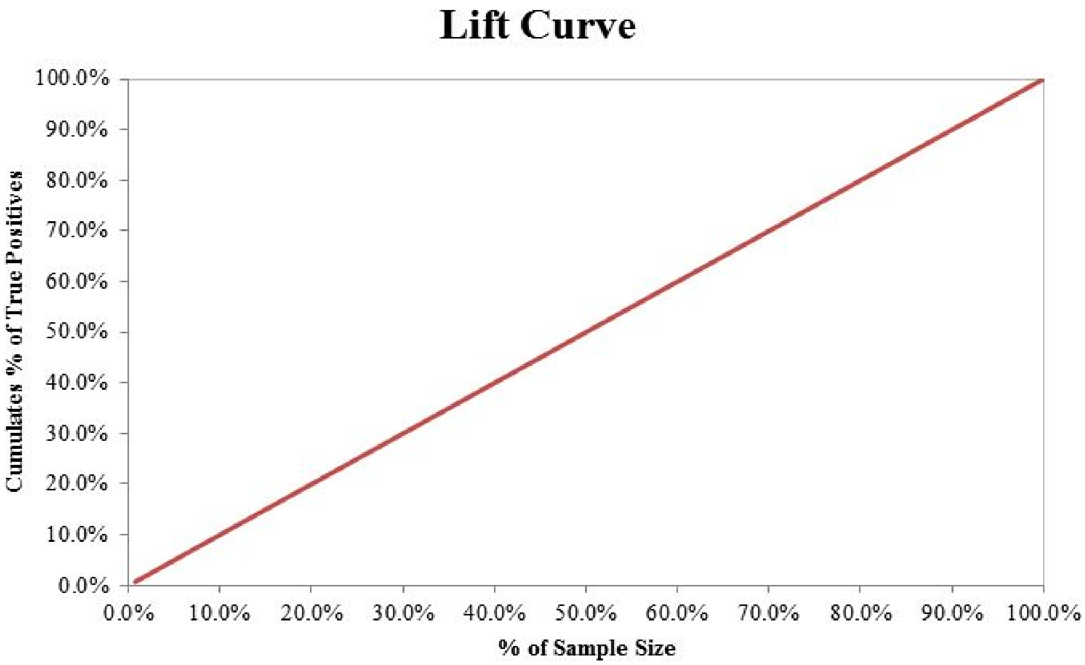


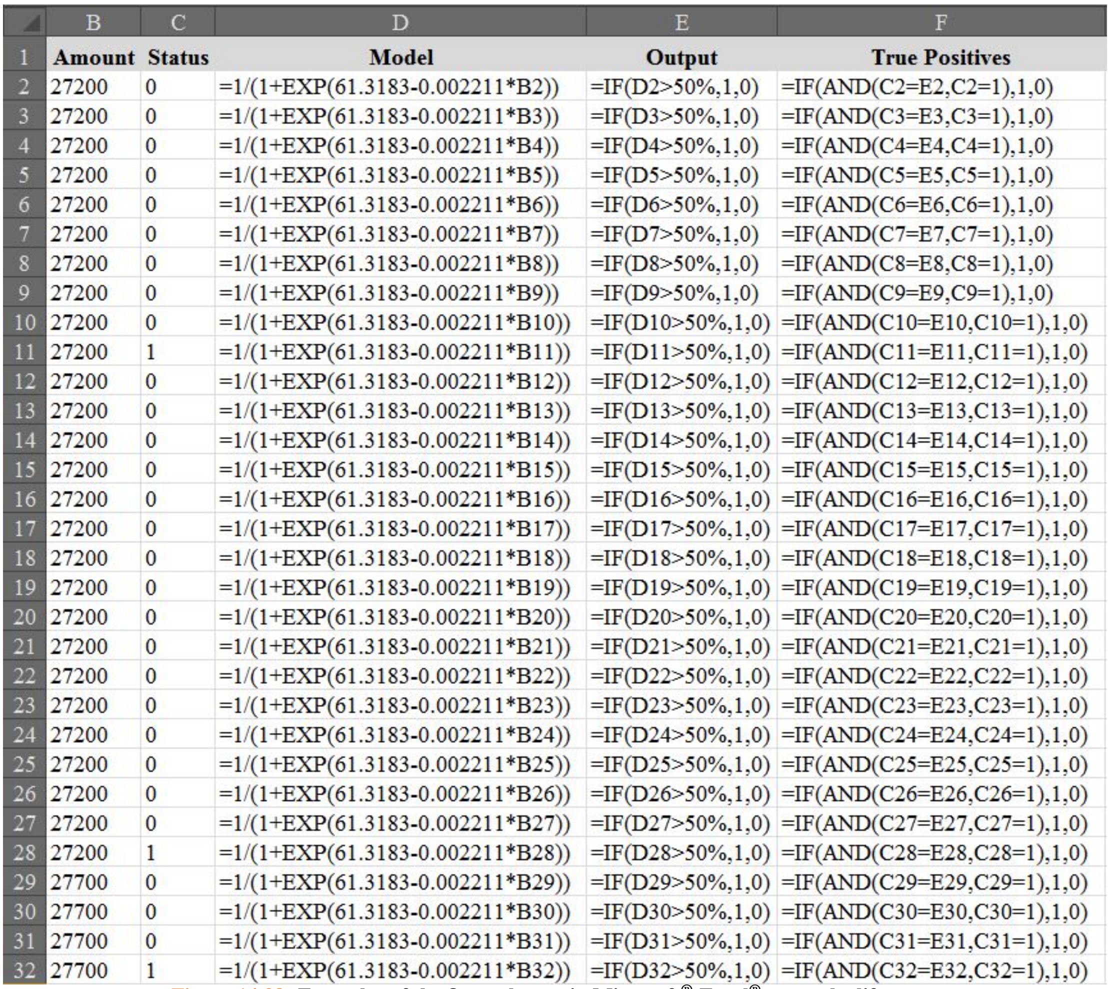
Figure 14.37: Lift curve at worst (as assumption)

But that's just to have an empirical reference when working with a single model of statistical clustering. In fact, given the multiplicity of statistical modelling methods, where almost each has its own empirical indicators of quality, statisticians have sought general criteria for the performance of a model (this obviously in the idea to be able to compare the accuracy of various models between them).

Thus, an empirical and intuitive enough choice is to say that a statistical method is better than another if for a given subset of the sample, its predictive power (the cumulative number of true positives for example) is better or not. This is the purpose of the lift curve.

The lift curve is therefore a variation of the ROC curve. The lift curve classify the individuals by descending score (again for reasons of simplification of interpretation of the surface under the curve and especially to have in first the target group of interest to monitor), by grouping them by percentiles for example, determining the percentage of events of interest in each percentile (normally the true positives) and then by plotting the cumulative curve of these percentages, so that a point of the coordinates (n, m) on the curve means that the $n\%$ of individuals with the highest score concentrate $m\%$ of events. This is the way to build this curve which makes we speak of "predictive performance of targeted marketing".

Let's see how to build such a curve, always with the same spreadsheet software and always the same data. The start of construction is to take some columns that we had used for the ROC curve (first 31 rows out of 136):

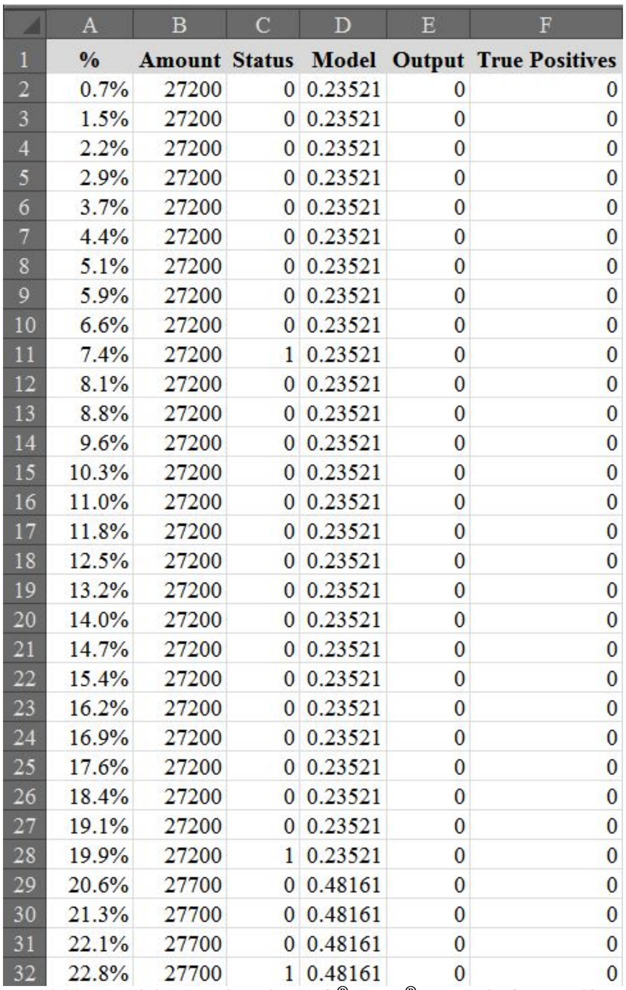


	B	C	D	E	F
1	Amount	Status	Model	Output	True Positives
2	27200	0	=1/(1+EXP(61.3183-0.002211*B2))	=IF(D2>50%,1,0)	=IF(AND(C2=E2,C2=1),1,0)
3	27200	0	=1/(1+EXP(61.3183-0.002211*B3))	=IF(D3>50%,1,0)	=IF(AND(C3=E3,C3=1),1,0)
4	27200	0	=1/(1+EXP(61.3183-0.002211*B4))	=IF(D4>50%,1,0)	=IF(AND(C4=E4,C4=1),1,0)
5	27200	0	=1/(1+EXP(61.3183-0.002211*B5))	=IF(D5>50%,1,0)	=IF(AND(C5=E5,C5=1),1,0)
6	27200	0	=1/(1+EXP(61.3183-0.002211*B6))	=IF(D6>50%,1,0)	=IF(AND(C6=E6,C6=1),1,0)
7	27200	0	=1/(1+EXP(61.3183-0.002211*B7))	=IF(D7>50%,1,0)	=IF(AND(C7=E7,C7=1),1,0)
8	27200	0	=1/(1+EXP(61.3183-0.002211*B8))	=IF(D8>50%,1,0)	=IF(AND(C8=E8,C8=1),1,0)
9	27200	0	=1/(1+EXP(61.3183-0.002211*B9))	=IF(D9>50%,1,0)	=IF(AND(C9=E9,C9=1),1,0)
10	27200	0	=1/(1+EXP(61.3183-0.002211*B10))	=IF(D10>50%,1,0)	=IF(AND(C10=E10,C10=1),1,0)
11	27200	1	=1/(1+EXP(61.3183-0.002211*B11))	=IF(D11>50%,1,0)	=IF(AND(C11=E11,C11=1),1,0)
12	27200	0	=1/(1+EXP(61.3183-0.002211*B12))	=IF(D12>50%,1,0)	=IF(AND(C12=E12,C12=1),1,0)
13	27200	0	=1/(1+EXP(61.3183-0.002211*B13))	=IF(D13>50%,1,0)	=IF(AND(C13=E13,C13=1),1,0)
14	27200	0	=1/(1+EXP(61.3183-0.002211*B14))	=IF(D14>50%,1,0)	=IF(AND(C14=E14,C14=1),1,0)
15	27200	0	=1/(1+EXP(61.3183-0.002211*B15))	=IF(D15>50%,1,0)	=IF(AND(C15=E15,C15=1),1,0)
16	27200	0	=1/(1+EXP(61.3183-0.002211*B16))	=IF(D16>50%,1,0)	=IF(AND(C16=E16,C16=1),1,0)
17	27200	0	=1/(1+EXP(61.3183-0.002211*B17))	=IF(D17>50%,1,0)	=IF(AND(C17=E17,C17=1),1,0)
18	27200	0	=1/(1+EXP(61.3183-0.002211*B18))	=IF(D18>50%,1,0)	=IF(AND(C18=E18,C18=1),1,0)
19	27200	0	=1/(1+EXP(61.3183-0.002211*B19))	=IF(D19>50%,1,0)	=IF(AND(C19=E19,C19=1),1,0)
20	27200	0	=1/(1+EXP(61.3183-0.002211*B20))	=IF(D20>50%,1,0)	=IF(AND(C20=E20,C20=1),1,0)
21	27200	0	=1/(1+EXP(61.3183-0.002211*B21))	=IF(D21>50%,1,0)	=IF(AND(C21=E21,C21=1),1,0)
22	27200	0	=1/(1+EXP(61.3183-0.002211*B22))	=IF(D22>50%,1,0)	=IF(AND(C22=E22,C22=1),1,0)
23	27200	0	=1/(1+EXP(61.3183-0.002211*B23))	=IF(D23>50%,1,0)	=IF(AND(C23=E23,C23=1),1,0)
24	27200	0	=1/(1+EXP(61.3183-0.002211*B24))	=IF(D24>50%,1,0)	=IF(AND(C24=E24,C24=1),1,0)
25	27200	0	=1/(1+EXP(61.3183-0.002211*B25))	=IF(D25>50%,1,0)	=IF(AND(C25=E25,C25=1),1,0)
26	27200	0	=1/(1+EXP(61.3183-0.002211*B26))	=IF(D26>50%,1,0)	=IF(AND(C26=E26,C26=1),1,0)
27	27200	0	=1/(1+EXP(61.3183-0.002211*B27))	=IF(D27>50%,1,0)	=IF(AND(C27=E27,C27=1),1,0)
28	27200	1	=1/(1+EXP(61.3183-0.002211*B28))	=IF(D28>50%,1,0)	=IF(AND(C28=E28,C28=1),1,0)
29	27700	0	=1/(1+EXP(61.3183-0.002211*B29))	=IF(D29>50%,1,0)	=IF(AND(C29=E29,C29=1),1,0)
30	27700	0	=1/(1+EXP(61.3183-0.002211*B30))	=IF(D30>50%,1,0)	=IF(AND(C30=E30,C30=1),1,0)
31	27700	0	=1/(1+EXP(61.3183-0.002211*B31))	=IF(D31>50%,1,0)	=IF(AND(C31=E31,C31=1),1,0)
32	27700	1	=1/(1+EXP(61.3183-0.002211*B32))	=IF(D32>50%,1,0)	=IF(AND(C32=E32,C32=1),1,0)

Figure 14.38: **Formulas of the first columns in Microsoft® Excel® to get the lift curve**

where the reader may have noticed that we have arbitrarily chosen to set the cutoff at 50% (thus in reality, and the reader must never forget it, the whole lift curve changes for each value of the cut- off !!!).

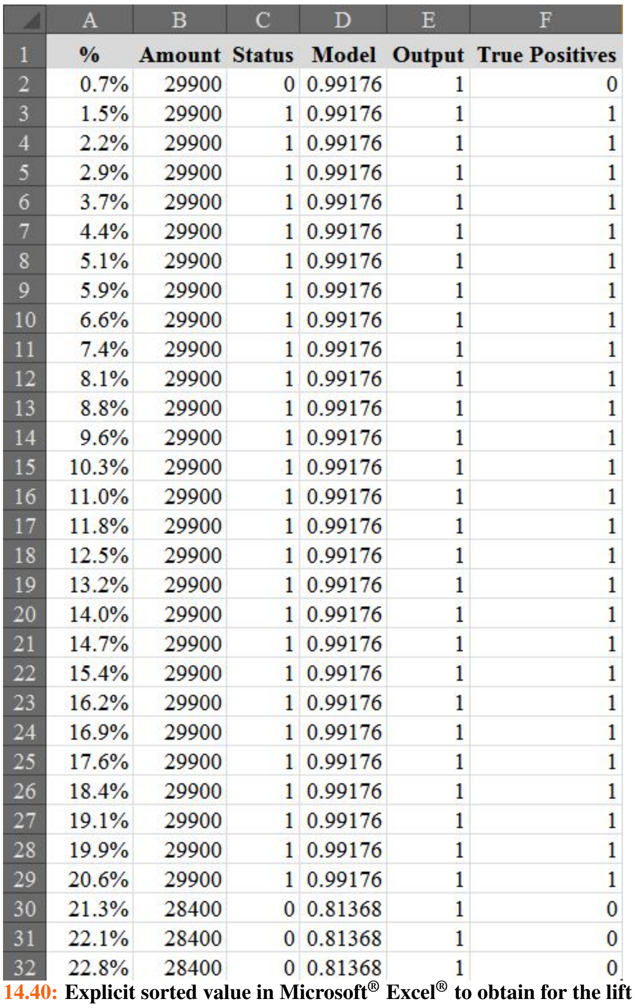
Which explicitly provides (remember that column `A`, that as for the ROC curve is trivial, as its simply represent the cumulative effective):



	A	B	C	D	E	F
1	**%**	**Amount**	**Status**	**Model**	**Output**	**True Positives**
2	0.7%	27200	0	0.23521	0	0
3	1.5%	27200	0	0.23521	0	0
4	2.2%	27200	0	0.23521	0	0
5	2.9%	27200	0	0.23521	0	0
6	3.7%	27200	0	0.23521	0	0
7	4.4%	27200	0	0.23521	0	0
8	5.1%	27200	0	0.23521	0	0
9	5.9%	27200	0	0.23521	0	0
10	6.6%	27200	0	0.23521	0	0
11	7.4%	27200	1	0.23521	0	0
12	8.1%	27200	0	0.23521	0	0
13	8.8%	27200	0	0.23521	0	0
14	9.6%	27200	0	0.23521	0	0
15	10.3%	27200	0	0.23521	0	0
16	11.0%	27200	0	0.23521	0	0
17	11.8%	27200	0	0.23521	0	0
18	12.5%	27200	0	0.23521	0	0
19	13.2%	27200	0	0.23521	0	0
20	14.0%	27200	0	0.23521	0	0
21	14.7%	27200	0	0.23521	0	0
22	15.4%	27200	0	0.23521	0	0
23	16.2%	27200	0	0.23521	0	0
24	16.9%	27200	0	0.23521	0	0
25	17.6%	27200	0	0.23521	0	0
26	18.4%	27200	0	0.23521	0	0
27	19.1%	27200	0	0.23521	0	0
28	19.9%	27200	1	0.23521	0	0
29	20.6%	27700	0	0.48161	0	0
30	21.3%	27700	0	0.48161	0	0
31	22.1%	27700	0	0.48161	0	0
32	22.8%	27700	1	0.48161	0	0

Figure 14.39: **Explicit value in Microsoft® Excel® to obtain for the lift curve**

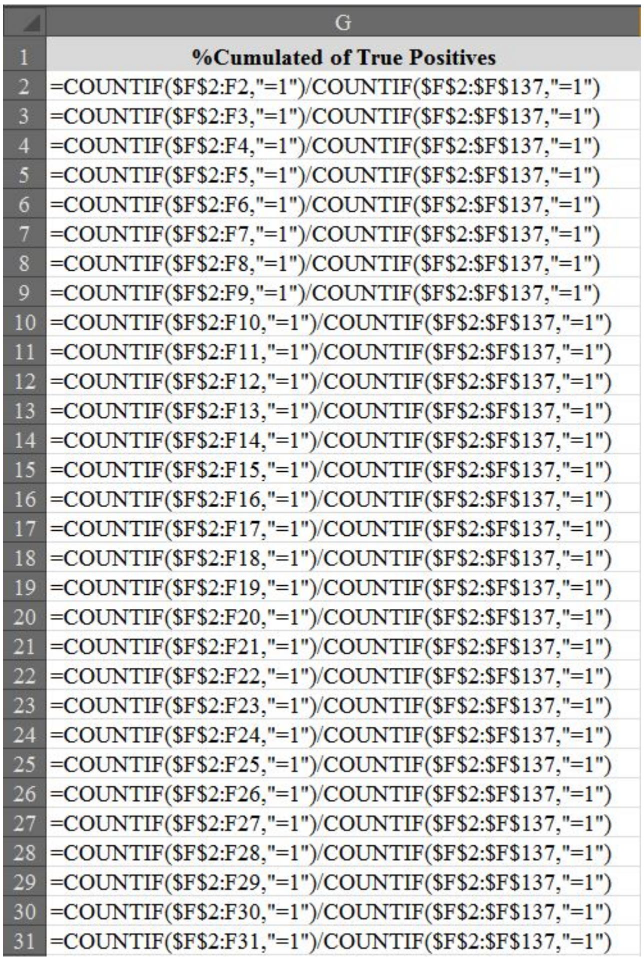
One difference now compared to the ROC curve is that we want the events of interest first (the bad debtors). Then we must sort the column Model (that is to say the: Score) in descending order. This will give:



	A	B	C	D	E	F
1	**%**	**Amount**	**Status**	**Model**	**Output**	**True Positives**
2	0.7%	29900	0	0.99176	1	0
3	1.5%	29900	1	0.99176	1	1
4	2.2%	29900	1	0.99176	1	1
5	2.9%	29900	1	0.99176	1	1
6	3.7%	29900	1	0.99176	1	1
7	4.4%	29900	1	0.99176	1	1
8	5.1%	29900	1	0.99176	1	1
9	5.9%	29900	1	0.99176	1	1
10	6.6%	29900	1	0.99176	1	1
11	7.4%	29900	1	0.99176	1	1
12	8.1%	29900	1	0.99176	1	1
13	8.8%	29900	1	0.99176	1	1
14	9.6%	29900	1	0.99176	1	1
15	10.3%	29900	1	0.99176	1	1
16	11.0%	29900	1	0.99176	1	1
17	11.8%	29900	1	0.99176	1	1
18	12.5%	29900	1	0.99176	1	1
19	13.2%	29900	1	0.99176	1	1
20	14.0%	29900	1	0.99176	1	1
21	14.7%	29900	1	0.99176	1	1
22	15.4%	29900	1	0.99176	1	1
23	16.2%	29900	1	0.99176	1	1
24	16.9%	29900	1	0.99176	1	1
25	17.6%	29900	1	0.99176	1	1
26	18.4%	29900	1	0.99176	1	1
27	19.1%	29900	1	0.99176	1	1
28	19.9%	29900	1	0.99176	1	1
29	20.6%	29900	1	0.99176	1	1
30	21.3%	28400	0	0.81368	1	0
31	22.1%	28400	0	0.81368	1	0
32	22.8%	28400	0	0.81368	1	0

Figure 14.40: **Explicit sorted value in Microsoft® Excel® to obtain for the lift curve**

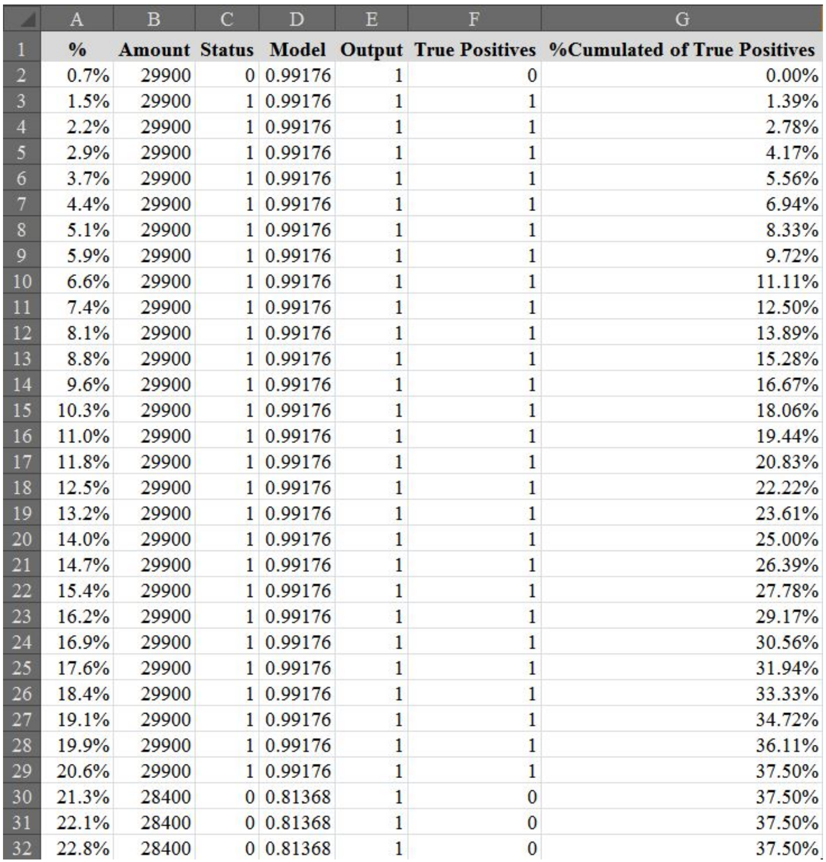
Then we add the column of %Cumulative of true positives:



	G
1	**%Cumulated of True Positives**
2	=COUNTIF($F$2:F2,"=1")/COUNTIF($F$2:$F$137,"=1")
3	=COUNTIF($F$2:F3,"=1")/COUNTIF($F$2:$F$137,"=1")
4	=COUNTIF($F$2:F4,"=1")/COUNTIF($F$2:$F$137,"=1")
5	=COUNTIF($F$2:F5,"=1")/COUNTIF($F$2:$F$137,"=1")
6	=COUNTIF($F$2:F6,"=1")/COUNTIF($F$2:$F$137,"=1")
7	=COUNTIF($F$2:F7,"=1")/COUNTIF($F$2:$F$137,"=1")
8	=COUNTIF($F$2:F8,"=1")/COUNTIF($F$2:$F$137,"=1")
9	=COUNTIF($F$2:F9,"=1")/COUNTIF($F$2:$F$137,"=1")
10	=COUNTIF($F$2:F10,"=1")/COUNTIF($F$2:$F$137,"=1")
11	=COUNTIF($F$2:F11,"=1")/COUNTIF($F$2:$F$137,"=1")
12	=COUNTIF($F$2:F12,"=1")/COUNTIF($F$2:$F$137,"=1")
13	=COUNTIF($F$2:F13,"=1")/COUNTIF($F$2:$F$137,"=1")
14	=COUNTIF($F$2:F14,"=1")/COUNTIF($F$2:$F$137,"=1")
15	=COUNTIF($F$2:F15,"=1")/COUNTIF($F$2:$F$137,"=1")
16	=COUNTIF($F$2:F16,"=1")/COUNTIF($F$2:$F$137,"=1")
17	=COUNTIF($F$2:F17,"=1")/COUNTIF($F$2:$F$137,"=1")
18	=COUNTIF($F$2:F18,"=1")/COUNTIF($F$2:$F$137,"=1")
19	=COUNTIF($F$2:F19,"=1")/COUNTIF($F$2:$F$137,"=1")
20	=COUNTIF($F$2:F20,"=1")/COUNTIF($F$2:$F$137,"=1")
21	=COUNTIF($F$2:F21,"=1")/COUNTIF($F$2:$F$137,"=1")
22	=COUNTIF($F$2:F22,"=1")/COUNTIF($F$2:$F$137,"=1")
23	=COUNTIF($F$2:F23,"=1")/COUNTIF($F$2:$F$137,"=1")
24	=COUNTIF($F$2:F24,"=1")/COUNTIF($F$2:$F$137,"=1")
25	=COUNTIF($F$2:F25,"=1")/COUNTIF($F$2:$F$137,"=1")
26	=COUNTIF($F$2:F26,"=1")/COUNTIF($F$2:$F$137,"=1")
27	=COUNTIF($F$2:F27,"=1")/COUNTIF($F$2:$F$137,"=1")
28	=COUNTIF($F$2:F28,"=1")/COUNTIF($F$2:$F$137,"=1")
29	=COUNTIF($F$2:F29,"=1")/COUNTIF($F$2:$F$137,"=1")
30	=COUNTIF($F$2:F30,"=1")/COUNTIF($F$2:$F$137,"=1")
31	=COUNTIF($F$2:F31,"=1")/COUNTIF($F$2:$F$137,"=1")

Figure 14.41: **Column of %Cumulatived of true positives for the lift curve**

What gives globally:



	A	B	C	D	E	F	G
1	%	Amount	Status	Model	Output	True Positives	%Cumulated of True Positives
2	0.7%	29900	0	0.99176	1	0	0.00%
3	1.5%	29900	1	0.99176	1	1	1.39%
4	2.2%	29900	1	0.99176	1	1	2.78%
5	2.9%	29900	1	0.99176	1	1	4.17%
6	3.7%	29900	1	0.99176	1	1	5.56%
7	4.4%	29900	1	0.99176	1	1	6.94%
8	5.1%	29900	1	0.99176	1	1	8.33%
9	5.9%	29900	1	0.99176	1	1	9.72%
10	6.6%	29900	1	0.99176	1	1	11.11%
11	7.4%	29900	1	0.99176	1	1	12.50%
12	8.1%	29900	1	0.99176	1	1	13.89%
13	8.8%	29900	1	0.99176	1	1	15.28%
14	9.6%	29900	1	0.99176	1	1	16.67%
15	10.3%	29900	1	0.99176	1	1	18.06%
16	11.0%	29900	1	0.99176	1	1	19.44%
17	11.8%	29900	1	0.99176	1	1	20.83%
18	12.5%	29900	1	0.99176	1	1	22.22%
19	13.2%	29900	1	0.99176	1	1	23.61%
20	14.0%	29900	1	0.99176	1	1	25.00%
21	14.7%	29900	1	0.99176	1	1	26.39%
22	15.4%	29900	1	0.99176	1	1	27.78%
23	16.2%	29900	1	0.99176	1	1	29.17%
24	16.9%	29900	1	0.99176	1	1	30.56%
25	17.6%	29900	1	0.99176	1	1	31.94%
26	18.4%	29900	1	0.99176	1	1	33.33%
27	19.1%	29900	1	0.99176	1	1	34.72%
28	19.9%	29900	1	0.99176	1	1	36.11%
29	20.6%	29900	1	0.99176	1	1	37.50%
30	21.3%	28400	0	0.81368	1	0	37.50%
31	22.1%	28400	0	0.81368	1	0	37.50%
32	22.8%	28400	0	0.81368	1	0	37.50%

Figure 14.42: **Overview of all constructed columns for the lift curve**

The last step is now simply to build a graph of %cumulative number of true positive (column G) relative to the %cumulative of the sample size (column `A`) to get:

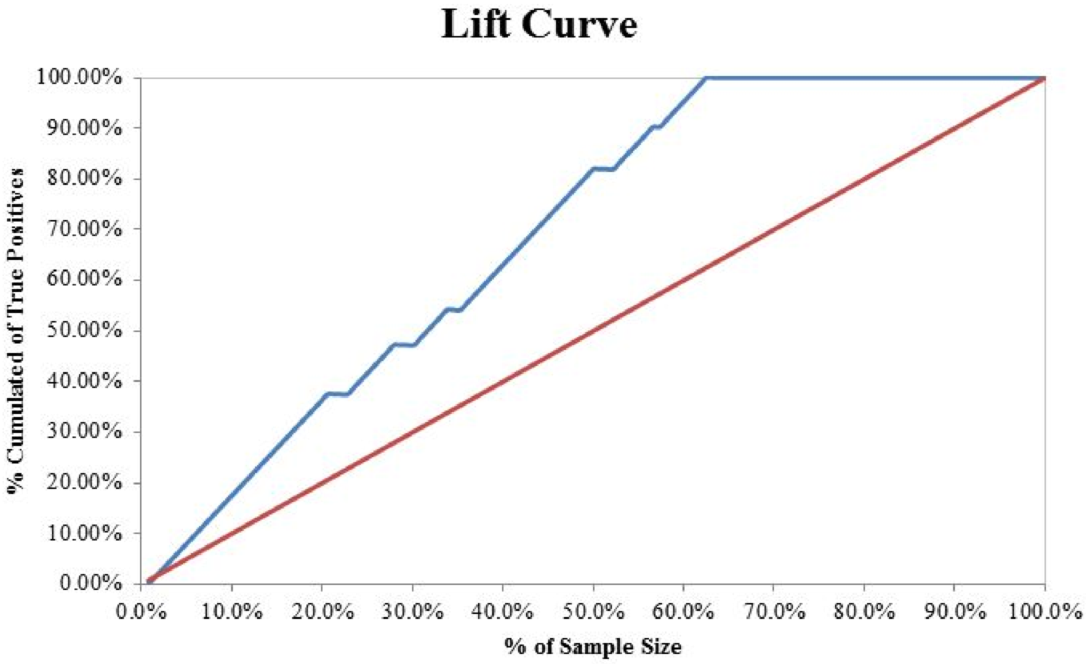


Figure 14.43: **Lift curve (in red at worst, in blue our model)**

So obviously we can not do in this case an interpretation of the predictive power in terms of "lift" in comparison to another statistical model. By cons we can compare the predictive power of the logistic model used in this case in terms of lift compared to the case considered at worst (the diagonal for reminder ...). So in this case, if we take the 20% of customers (debtors) having the highest score, that's what we have:

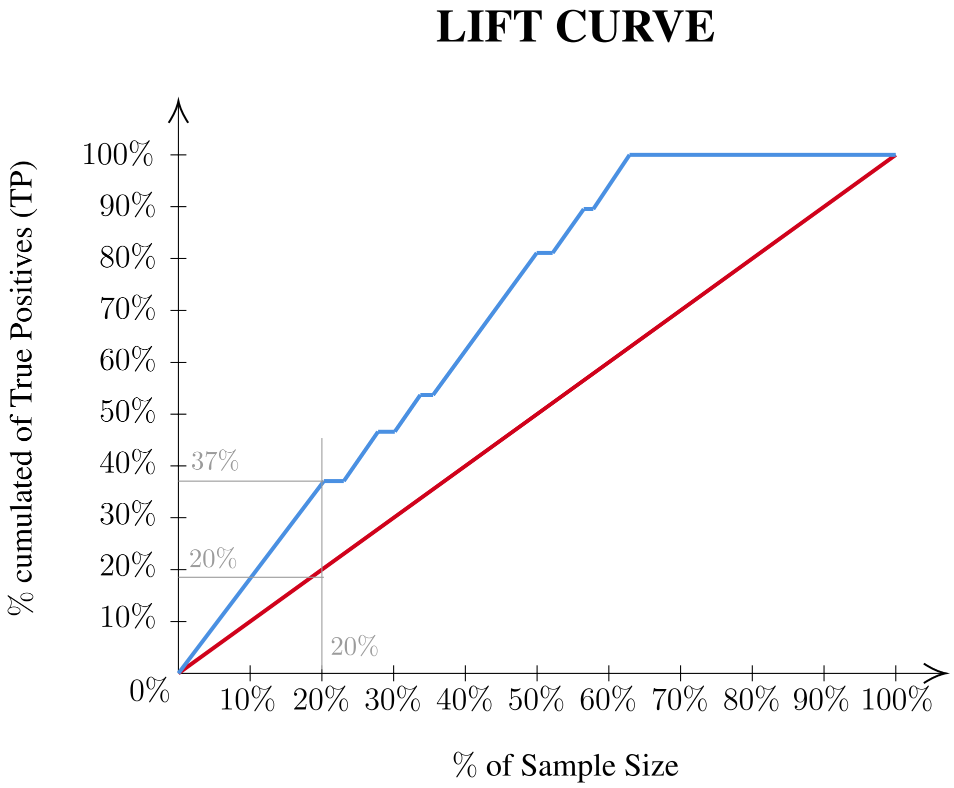


Figure 14.44: **Lift Analysis**

So in this particular example, this means that if we focus our analysis / research / target marketing or other... than on the 20% of the sample being the best score (for example ... for cost reasons), we have an overperformance of $37\%/20\% = 1.85$. Then we say that the model has a lift of 1.85. The idea then when we compare several models is to keep the one with the greatest lift (leverage).

14.1.12.7 General Linear Models (GLM)

So far we have seen multiple linear regression techniques (simple, multiple linear regression, Gaussian linear model, linearisation of non-linear regressions, logistic regressions, etc.).

There are, however, still other types of regressions named "general linear models" (not to be confused with "generalized linear models") which are intended for cases where the regression of the variable to be explained (and therefore the residuals) takes values whose range is not necessarily in $]-\infty,+\infty[$ as it is the case for the Gaussian linear model, but only takes integer positive values in $\mathbb{N}$, or that takes real values in the interval $[0,1]$, etc.

Remark

We will focus here on the essential concepts of the generalized linear regression models because for each of the submodels that we will see below it is possible to find textbooks of 300 to 500 pages which are dedicated to it! In the same way as for classical linear regression (Gaussian model) we one can find textbooks of almost 1,000 pages.

To introduce this type of model, recall what characterizes a general regression model (think by reading the following lines to the particular case of Gaussian linear model as it may help for the understanding):

C1. The response variable Y is a random component with a probability distribution: Normal (the most frequent case for continuous data), Binomial (positive integer successes in n trials), Poisson (positive integer successes for rare events), Bernoulli (whole positive success with one unique try), Binomial-Negative (positive integer success in a sequence of independent and identically distributed Bernoulli trials before a specified (non-random) number of failures occurs). We notice that these are laws belonging to the exponential family (the underlying idea being to linearise it with logarithm) as we will see later.

For example, for the linear Gaussian model seen earlier above, we have shown that for the case with a unique explanatory variable:

$$Y_i=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{\alpha x_i+\beta-\mu_i}{\sigma}\right)^2} \tag{14.1.988}$$

was a model with which we relied on the least squares method via the log-likelihood maximum.

C2. The explanatory variables $x_1,x_2,\ldots,x_k$ with the intercept constant used as predictors in the model define as a linear (or non-linear by multiplying the variables by pairs) the deterministic components.

C3. The link describing the functional relationship between the linear combination of the explanatory variables and the expectation of the response variable Y is named the "link function" or "canonical parameter" and is traditionally denoted $g(\text{E}(Y_i))$.

In general, when the relation between $\text{E}(Y_i)$ and the explanatory variables is linear we speak of "linear model", when the relation between a function $g(\text{E}(Y_i))$ and the parameters is linear, we speak then of "general linear model".

For example, in the case of the linear Gaussian model with an explanatory variable, we have (it is immediate by the property of the underlying distribution):

$$\text{E}(Y_i)=\mu_i=\alpha x_i+\beta \tag{14.1.989}$$

Thus, in the case of the Gaussian linear model, the expectation is modelled directly. Thus, the link function for the Normal distribution is the mean itself! We talk then about an "identity link relation".

We will obviously not be able to use the linear Gaussian model given the definition domain taken by the variable to be explained when that latter is defined only on integers. On the other hand we could try an approach by saying that each count:

$$Y_i=n_i \tag{14.1.990}$$

Follows a Poisson distribution:

$$Y_i=e^{-\lambda_i}\frac{\lambda_i^{k_i}}{k_i!} \tag{14.1.991}$$

of mean (still in the special case with a unique explanatory variable!):

$$\mu_i=\lambda_i=N_i e^{\alpha x_i+\beta} \tag{14.1.992}$$

The studied model is then written:

$$Y_i=\text{Poisson}(\mu_i) \tag{14.1.993}$$

and we deduce from it:

$$\ln(\mu_i)=\ln(N_i)+\alpha x_i+\beta=\alpha x_i+B \tag{14.1.994}$$

We say that $\log(N_i)$ is an "offset variable" (in the industry it is often a value representing the time spent until the event of interest) and that the "link function" is then the logarithm for the Poisson regression. We speak more generally of "log-linear Poisson model" which is associated with a "logarithmic link".

Obviously the above approach is very far-fetched but a little further we will see the rigorous approach to get this link function.

Let us recall some important things in this model which are that the hypothesis are similar to the Gaussian linear model: the Y_i are all supposed to follow a Poisson distribution, to have all the same variance and to be independent, and so on...

Note that since we have:

$$\ln(\mu_i)=\alpha x_i+B\Leftrightarrow\mu_i=e^{\alpha x_i}e^B \tag{14.1.995}$$

The mean is therefore exponential, influenced by the sign of α (it is logical but it can always be useful to specify it). In the general case, we will see that the law of probability density associated with can be written in the general form:

$$\boxed{f(y_i,\theta_i,\varphi_i,\omega_i)=\exp\left(\frac{y_i\theta_i-b(\theta_i)}{a_i(\varphi)}+c(y_i,\alpha)\right)} \tag{14.1.996}$$

and we say then that the law belongs to the "exponential family" (the Normal, Poisson, Bernoulli and Gamma laws are part of it for example).

Remark

Very few specialists designate by the abbreviation "GLZM" the GLM other than those based on the Normal law.

14.1.12.7.1 Normal GLM

Let's see this with the Normal law by analogy (the homoscedastic case to simplify...):

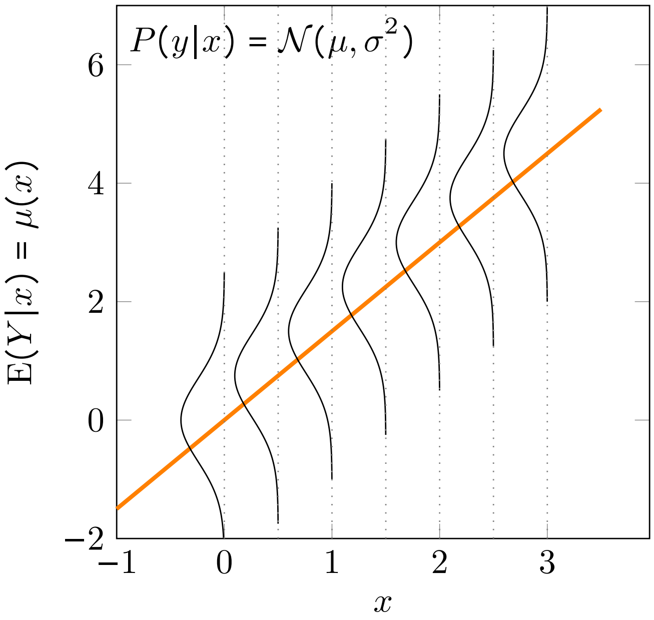


Figure 14.45: **Underlying idea of the Normal GLM (author: Walmes Zeviani)**

We therefore start from:

$$\begin{aligned}Y_i &= \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{\alpha x_i+\beta-\mu_i}{\sigma}\right)^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\left(\frac{y_i-\mu_i}{\sigma}\right)^2\right)\\ &= \exp\left(-\frac{1}{2\sigma^2}(y_i-\mu_i)^2+\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)\right) = \exp\left(-\frac{1}{2\sigma^2}(y_i^2-2y_i\mu_i+\mu_i^2)-\ln(\sqrt{2\pi\sigma^2})\right)\\ &= \exp\left(\frac{y_i\mu_i-\mu_i^2/2}{\sigma^2}-\frac{y_i^2}{2\sigma^2}-\ln\left((2\pi\sigma^2)^{1/2}\right)\right) = \exp\left(\frac{y_i\mu_i-\mu_i^2/2}{\sigma^2}-\frac{1}{2}\left(\frac{y_i^2}{\sigma^2}-\ln\left(2\pi\sigma^2\right)\right)\right)\end{aligned} \tag{14.1.997}$$

So it comes by identification with:

$$f(y_i,\theta_i,\varphi,\omega_i) = \exp\left(\frac{y_i\theta_i-b(\theta_i)}{a_i(\varphi)}+c(y_i,\varphi)\right) \tag{14.1.998}$$

that:

$$\begin{aligned}
&\theta_i=\mu_i\\
&b(\theta_i)=\frac{1}{2}\theta_i^2=\frac{1}{2}\mu_i^2\\
&a_i(\varphi)=\sigma^2\\
&c(y_i,\varphi)=-\frac{1}{2}\left(\frac{y_i^2}{\varphi}-\ln(2\pi\varphi)\right)=-\frac{1}{2}\left(\frac{y_i^2}{\sigma^2}-\ln(2\pi\sigma^2)\right)\\
&\varphi=\sigma^2
\end{aligned} \tag{14.1.999}$$

Note also that (this is very important for what will follow!):

$$\begin{aligned}
&\frac{\mathrm{d}b(\theta_i)}{\mathrm{d}\theta_i}=\frac{\mathrm{d}b(\mu_i)}{\mathrm{d}(\mu_i)}=\frac{\mathrm{d}\left(\dfrac{1}{2}\mu_i^2\right)}{\mathrm{d}(\mu_i)}=\mu\\
&\frac{\mathrm{d}^2b(\theta_i)}{\mathrm{d}\theta_i^2}a_i(\varphi)=\frac{\mathrm{d}^2b(\mu_i)}{\mathrm{d}(\mu_i^2)}\sigma^2=\frac{\mathrm{d}^2\left(\dfrac{1}{2}\mu_i^2\right)}{\mathrm{d}(\mu_i^2)}\sigma^2=1\cdot\sigma^2=\sigma^2
\end{aligned} \tag{14.1.1000}$$

So we fall back on the expected mean and the variance of the Normal law. It would be interesting to check if we have the same kind of property for another law of the exponential family to see if it is generalizable or not ???

Let us also indicate that compared to the famous "canonical parameter" (since mean and variance ultimately depend on it), we have in this case:

$$\theta_i=\mu_i \tag{14.1.1001}$$

so the canonical function is then (without too much surprise in this case...):

$$g(\mathrm{E}(Y))=1\cdot\mathrm{E}(Y)=\mu_i=\alpha x_i+\beta \tag{14.1.1002}$$

that it is customary to write in a relatively general case as we already know in vector notation (see the study of multiple linear regression for the notation):

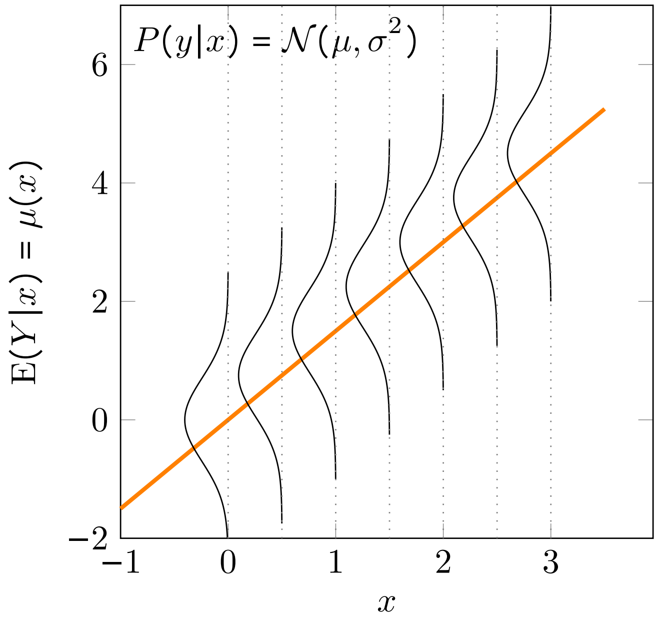
$$g(\mathrm{E}(Y))=\theta_i=\mu_i=\vec{x}_i^T\vec{\beta}=\sum_{j=0}^{p-1}x_{j,i}\beta_j \tag{14.1.1003}$$

In this case the canonical function is the unitary application. We speak then of "canonical identity link" and it is the only case where the method of ordinary least squares is confused (to understand: is an identical approach) to the use of the GLM.

We notice that the assumptions of use of the Normal GLM trivially remain the independence of the explanatory variables, the homoscedasticity, the linearity of the model and the linearity of the expected mean.

14.1.12.7.2 Poisson GLM

Let's do the same as before but with the Poisson's law. The idea is therefore that the variable to be explained is discrete and whereas in the same order of idea of the Normal regression we have:



then with the Poisson regression, the idea is to have (notice that as the mean increase, the variance seems also to increase and we will prove this further below even if it's quite obvious regarding to the moment of the Poisson distribution):

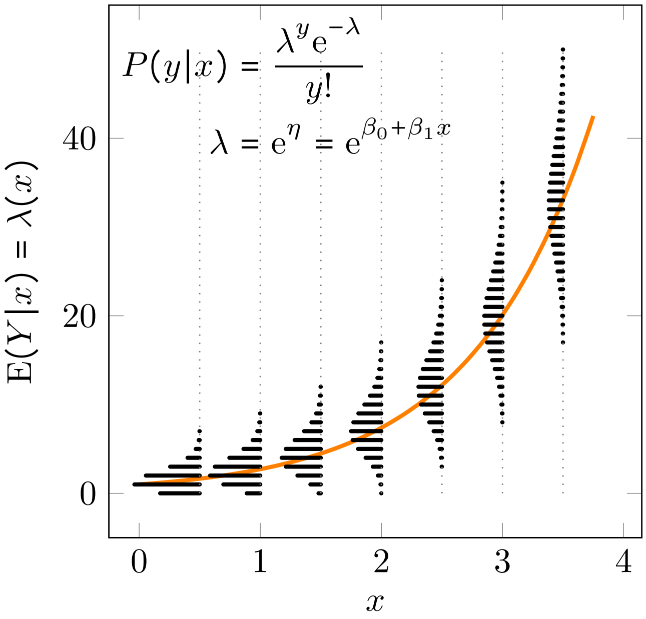


Figure 14.46: Underlying idea of the Poisson GLM (author: Walmes Zeviani)

However, the reader must be keep in mind that even a continuous variable can have an excellent fit to a Poisson's law, which is why Poisson's regression is also sometimes used for variables to be explained that are continuous (typically for heteroscedastic regressions where weighted regression is to difficult to apply or fails).

We will therefore look for a regression model which, like Gaussian GLM regression, gives the slip of the expected mean of the Gaussian law as a function of explanatory variables but this time with a model that gives the slip of the expected mean of the Poisson's law (yes indeed as a reminder ... the two moments of the Poisson's law are directly connected to each other). We start from:

$$\begin{aligned}Y_i &= \frac{\mu_i^{\alpha x_i+\beta}}{(\alpha x_i+\beta)!}e^{-\mu_i}=\frac{\mu_i^{y_i}}{y_i!}e^{-\mu_i}=\exp\left(\ln\left(\frac{\mu_i^{y_i}}{y_i!}e^{-\mu_i}\right)\right)\\ &=\exp\left(\ln(\mu_i^{y_i})-\ln(y_i!)+\ln(e^{-\mu_i})\right)=\exp\left(y_i\mu_i-\ln(y_i!)-\mu_i\right)\\ &=\exp\left(\left(\frac{y_i\ln(\mu_i)-\mu_i}{1}\right)-\ln(y_i!)\right)\end{aligned}\tag{14.1.1004}$$

Last arrangement that we had to guess ... to make an identification with:

$$f(y_i,\theta_i,\varphi,\omega_i)=\exp\left(\frac{y_i\theta_i-b(\theta_i)}{a_i(\varphi)}+c(y_i,\varphi)\right)\tag{14.1.1005}$$

Therefore it comes:

$$\begin{aligned}&\theta_i=\ln(\mu_i)\\ &b(\theta_i)=b(\ln(\mu_i))=\mu_i\Rightarrow b(\theta_i)=e^{\ln(\mu_i)}\\ &a_i(\varphi)=1\\ &c(y_i,\varphi)=-\ln(y_i!)\\ &\varphi=1\end{aligned}\tag{14.1.1006}$$

Let us indicate for the small parenthesis that some authors use the Euler Gamma function (see section Differential and Integral Calculus page 1574) and write:

$$\ln(y_i!)=\Gamma(y_i+1)\tag{14.1.1007}$$

Now, let's have a look at what gives the derivatives used above:

$$\begin{aligned}&\frac{\mathrm{d}b(\theta_i)}{\mathrm{d}\theta_i}=\frac{\mathrm{d}(e^{\ln(\mu_i)})}{\mathrm{d}(\ln(\mu_i))}=e^{\ln(\mu_i)}=\mu_i\\ &\frac{\mathrm{d}^2b(\theta_i)}{\mathrm{d}\theta_i^2}a_i(\varphi)=\frac{\mathrm{d}(e^{\ln(\mu_i)})}{\mathrm{d}(\ln(\mu_i))}\cdot 1=e^{\ln(\mu_i)}=\mu_i=\sigma^2\end{aligned}\tag{14.1.1008}$$

So we can once again find the expected mean and variance but this time of the Poisson's law. It would be interesting to check again if we have the same kind of property for another law of the exponential family in order to see if it is generalizable or not ???

So we notice a second time that the parameter θ controls the mean and the variance through its first and second derivative. The variance is therefore a function of the mean for all the laws of the family of exponential laws.

Let us also indicate that relatively to the famous "canonical parameter" (since mean and variance ultimately depend on it), we have in this case:

$$\theta_i = \ln(\mu_i) \tag{14.1.1009}$$

The canonical function for the Poisson's law is then:

$$g(\mathrm{E}(Y)) = \theta_i = \ln(\mathrm{E}(Y)) = \ln(\mu_i) \tag{14.1.1010}$$

So the canonical function is the logarithmic application. We then talk about "canonical logarithmic link".

Before going further let us compare, the generalized form of the Normal law of the regression with that of Poisson:

$$\begin{aligned} Y_{i,\mathcal{N}} &= \exp\left(\frac{y_i\mu_i - \mu_i^2/2}{\sigma^2} - \frac{1}{2}\left(\frac{y_i^2}{\sigma^2} - \ln\left(2\pi\sigma^2\right)\right)\right) \\ Y_{i,\mathcal{P}} &= \exp\left(\frac{y_i \ln(\mu_i) - \mu_i}{1} - \ln\left(y_i!\right)\right) \end{aligned} \tag{14.1.1011}$$

Remark

We will not go too much further with the Poisson's law for regression because according to the experience feedback of the senior practitioners it seems that it suffers from the major problem that as the expected mean is equal to the variance whereas in reality the variance tends to be higher than the expected mean. The negative binomial model (see below) should be preferred to that of Poisson.

For the parameters to be estimated, it is customary to also use the maximum log-likelihood as we have done for many previous regression models. Then the likelihood is given in this case by:

$$L(\mu_i, y_i) = \prod_{i=1}^{n} \frac{\mu_i^{y_i}}{y_i!} e^{-\mu_i} \tag{14.1.1012}$$

The log-likelihood becomes:

$$\ln[L(\mu_i, y_i)] = \ln\left(\prod_{i=1}^{n} \frac{\mu_i^{y_i}}{y_i!} e^{-\mu_i}\right) = \sum_{i=1}^{n} (-\mu_i + y_i \ln(\mu_i) - \ln(y_i!)) \tag{14.1.1013}$$

Now to proceed as we did with the Gaussian linear regression, let us recall that the y_i are given by the experiment. We will write then:

$$\ln[L(\mu_i, y_i)] = \sum_{i=1}^{n} (-\mu_i + Y_i \ln(\mu_i) - \ln(Y_i!)) \tag{14.1.1014}$$

and that, as for Gaussian linear regression, the explicit expression of the μ_i must allow us to determine the expression of parameters β_j of the regression.

However we have seen earlier above that the link function for the Poisson regression was the natural logarithm which in the general case is written as follows:

$$g(\mathrm{E}(Y)) = \ln(\mathrm{E}(Y)) = \ln(\mu_i) = \vec{x}_i^T \vec{\beta} = \sum_{j=0}^{p-1} x_{j,i}\beta_j \tag{14.1.1015}$$

So we have so far:

$$\begin{gathered} g_{\mathcal{N}}(\mu_i)=\theta_i=\mu_i=\vec{x}_i^T\vec{\beta}=\sum_{j=0}^{p-1}x_{j,i}\beta_j \\ g_{\mathcal{P}}(\mu_i)=\theta_i=\ln(\mu_i)=\vec{x}_i^T\vec{\beta}=\sum_{j=0}^{p-1}x_{j,i}\beta_j \end{gathered} \tag{14.1.1016}$$

Therefore for the Poisson logarithm link:

$$\ln(\mu_i)=\vec{x}_i^T\vec{\beta}=\sum_{j=0}^{p-1}x_{j,i}\beta_j \Leftrightarrow \mu_i=\exp\left(\sum_{j=0}^{p-1}x_{j,i}\beta_j\right) \tag{14.1.1017}$$

Hence the log-likelihood will be written:

$$\begin{aligned} \ln[L(\mu_i,y_i)]&=\sum_{i=1}^{n}(-\mu_i+Y_i\ln(\mu_i)-\ln(Y_i!)) \\ &=\sum_{i=1}^{n}\left(-\exp\left(\sum_{j=0}^{p-1}x_{j,n}\beta_j\right)+Y_i\left(\sum_{j=0}^{p-1}x_{j,n}\beta_j\right)-\ln(Y_i!)\right) \end{aligned} \tag{14.1.1018}$$

To solve this kind of log-likelihood, we then use Newton-Raphson type algorithms (see further below in this section page 6673).

We notice that the hypotheses of use of the GLM Poisson remain trivially the independence of the explanatory variables, the homoscedasticity is on the other hand no more assumed since the variance depends on the expected mean, the linearity of the model is always a hypothesis and the expected mean is linear only under a logarithmic transformation.

Finally, let us say that, just as for the Gaussian regression, we can use the result of the Poisson regression to make the probabilistic inference. Indeed, since the density probability function of the Poisson distribution is given by:

$$P_k=\frac{\mu^k}{k!}e^{-\mu} \tag{14.1.1019}$$

and since we have:

$$\ln(\mu_i)=\alpha x_i+B\Leftrightarrow\mu_i=e^{\alpha x_i}e^{B} \tag{14.1.1020}$$

Then:

$$P_k=\frac{\mu^k}{k!}e^{-\mu}=\frac{\left(e^{\alpha x_i+B}\right)^k}{k!}e^{-\left(e^{\alpha x_i+B}\right)} \tag{14.1.1021}$$

and in this case you have to change the notation so that it makes sense:

$$P_k=\frac{\left(e^{\alpha x_i+\beta}\right)^{y_i}}{y_i!}e^{-\left(e^{\alpha x_i+\beta}\right)} \tag{14.1.1022}$$

Thus, for any abscissa value x_i of our model (which in the present case is univariate), we can calculate the prior probability of the realization of an event of interest for an ordinate value $y_i\in\mathbb{N}$.

Remark

The response variable of a Poisson GLM is obviously a count as we have seen it. For example, the response variable can model the number of customers who enter a grocery store. It turns out that this is NOT the only way to perform Poisson GLM! We can also divide the response variable by a time period or distance, which sets the response variable as a rate or a proportion instead. Extending our earlier example, we can therefore model the number of customers who enter a grocery store per hour. In Poisson GLM, this is named an "offset".

14.1.12.7.3 Binomial GLM

Let's do the same with the binomial law!

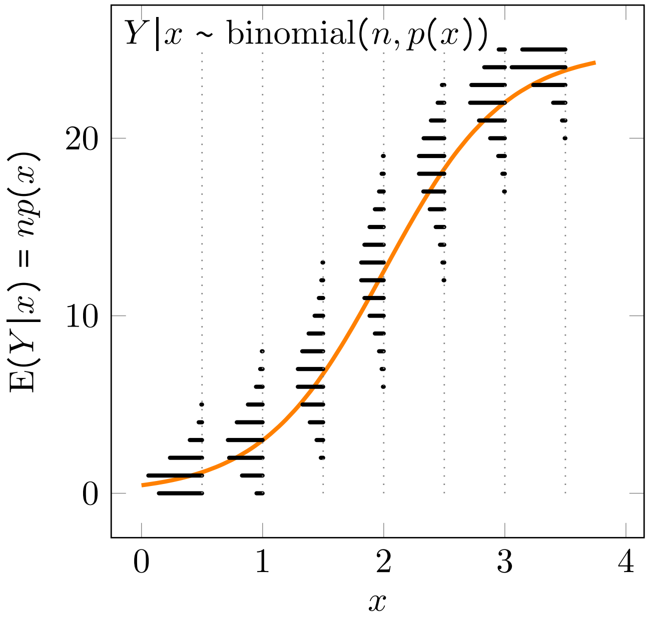


Figure 14.47: **Underlying idea of the binomial GLM (author: Walmes Zeviani)**

We start from:

$$\begin{aligned}
Y_i &= \frac{N!}{(\alpha x_i+\beta)!(N-(\alpha x_i+\beta))!}p^{\alpha x_i+\beta}q^{N-(\alpha x_i+\beta)}\\
&= \frac{N!}{y_i!(N-y_i)!}p^{y_i}(1-p)^{N-y_i} = C_{y_i}^{N}p^{y_i}(1-p)^{N-y_i}\\
&= \exp\left(\ln\left(C_{y_i}^{N}p^{y_i}(1-p)^{N-y_i}\right)\right)\\
&= \exp\left(\ln(p^{y_i})+\ln\left((1-p)^{N-y_i}\right)+\ln(C_{y_i}^{N})\right)\\
&= \exp\left(y_i\ln(p)+(N-y_i)\ln(1-p)+\ln(C_{y_i}^{N})\right)\\
&= \exp\left(y_i\ln(p)+N\ln(1-p)-y_i\ln(1-p)+\ln(C_{y_i}^{N})\right)\\
&= \exp\left(y_i\ln\left(\frac{p}{1-p}\right)+N\ln(1-p)+\ln(C_{y_i}^{N})\right)\\
&= \exp\left(\left(\frac{y_i\ln\left(\dfrac{p}{1-p}\right)+N\ln(1-p)}{1}\right)+\ln(C_{y_i}^{N})\right)
\end{aligned}\tag{14.1.1023}$$

Now, let's do the identification with:

$$f(y_i,\theta_i,\varphi,\omega_i)=\exp\left(\frac{y_i\theta_i-b(\theta_i)}{a_i(\varphi)}+c(y_i,\varphi)\right)\tag{14.1.1024}$$

Therefore we get (take a particular care at how looks like the first parameter...):

$$\begin{aligned}
\theta_i &= \ln\left(\frac{p}{1-p}\right)\\
b(\theta_i) &= b\left(\ln\left(\frac{p}{1-p}\right)\right) = -N\ln(1-p)\Rightarrow b(\theta_i)=N\ln\left(1+e^{\theta_i}\right)\\
a_i(\varphi) &= 1\\
c(y_i,\varphi) &= -\ln(C_{y_i}^{N})\\
\varphi &= 1
\end{aligned}\tag{14.1.1025}$$

Now, let's have a look at what gives the derivatives used earlier above:

$$\begin{aligned}
\frac{\mathrm{d}b(\theta_i)}{\mathrm{d}\theta_i} &= \frac{\mathrm{d}(N\ln(1+e^{\theta_i}))}{\mathrm{d}\theta_i} = N\frac{\mathrm{d}(\ln(1+e^{\theta_i}))}{\mathrm{d}\theta_i} = N\frac{e^{\theta_i}}{1+e^{\theta_i}}\\
&= N\frac{e^{\ln\left(\frac{p}{1-p}\right)}}{1+e^{\ln\left(\frac{p}{1-p}\right)}} = N\frac{\dfrac{p}{1-p}}{1+\dfrac{p}{1-p}} = N\frac{p}{1-p+p} = Np = \mu
\end{aligned}\tag{14.1.1026}$$

and:

$$\begin{aligned}\frac{\mathrm{d}^2b(\theta_i)}{\mathrm{d}\theta_i^2}a_i(\varphi)&=\frac{\mathrm{d}\left(N\dfrac{e^{\theta_i}}{1+e^{\theta_i}}\right)}{\mathrm{d}\theta_i}\cdot 1=N\frac{\mathrm{d}\left(e^{\theta_i}\dfrac{1}{1+e^{\theta_i}}\right)}{\mathrm{d}\theta_i}\\&=N\left(e^{\theta_i}\frac{1}{1+e^{\theta_i}}-e^{\theta_i}\frac{e^{\theta_i}}{(1+e^{\theta_i})^2}\right)=N\left(\frac{e^{\theta_i}(1+e^{\theta_i})}{(1+e^{\theta_i})^2}-\frac{e^{\theta_i}e^{\theta_i}}{(1+e^{\theta_i})^2}\right)\\&=N\left(\frac{e^{\theta_i}(1+e^{\theta_i})-e^{\theta_i}e^{\theta_i}}{(1+e^{\theta_i})^2}\right)=N\frac{e^{\theta_i}}{(1+e^{\theta_i})^2}=N\left(\frac{e^{\theta_i}}{1+e^{\theta_i}}\right)\left(\frac{1}{1+e^{\theta_i}}\right)\\&=N\left(\frac{e^{\ln\left(\frac{p}{1-p}\right)}}{1+e^{\ln\left(\frac{p}{1-p}\right)}}\right)\left(\frac{1}{1+e^{\ln\left(\frac{p}{1-p}\right)}}\right)=Np\left(\frac{1}{1+\dfrac{p}{1-p}}\right)=Np(1-p)=\sigma^2\end{aligned}\tag{14.1.1027}$$

So we can once again fall back on the expected mean and variance but this time the of the binomial law!

Let us also indicate that relatively to the famous "canonical parameter" (since expected mean and variance ultimately depend on it), we have in this case:

$$\theta_i=\ln\left(\frac{p}{1-p}\right)\tag{14.1.1028}$$

So the canonical function is:

$$g(\mathrm{E}(Y))=\theta_i=\ln\left(\frac{p}{1-p}\right)=\vec{x}_i^T\vec{\beta}=\sum_{j=0}^{p-1}x_{j,i}\beta_j\tag{14.1.1029}$$

in this case the canonical function is named, as we already know, the "logit link" or simply the "logit" (which we had already discovered otherwise in our study of logistic regression).

Let us compare now the generalized form of the Normal law of the regression with that of Poisson and the Binomial law:

$$\begin{aligned}Y_{i,\mathcal{N}}&=\exp\left(\frac{y_i\mu_i-\mu_i^2/2}{\sigma^2}-\frac{1}{2}\left(\frac{y_i^2}{\sigma^2}-\ln\left(2\pi\sigma^2\right)\right)\right)\\Y_{i,\mathcal{P}}&=\exp\left(\frac{y_i\ln(\mu_i)-\mu_i}{1}-\ln\left(y_i!\right)\right)\\Y_{i,\mathcal{B}}&=\exp\left(\left(\frac{y_i\ln\left(\dfrac{p}{1-p}\right)+N\ln\left(1-p\right)}{1}\right)+\ln\left(C_{y_i}^N\right)\right)\end{aligned}\tag{14.1.1030}$$

and also the link functions:

$$\begin{aligned}
g_N(\mu_i)&=\theta_i=\mu_i=\vec{x}_i^T\vec{\beta}=\sum_{j=0}^{p-1}x_{j,i}\beta_j\\
g_P(\mu_i)&=\theta_i=\ln(\mu_i)=\vec{x}_i^T\vec{\beta}=\sum_{j=0}^{p-1}x_{j,i}\beta_j\\
g_B(\mu_i)&=\theta_i=\ln\left(\frac{p}{1-p}\right)=\vec{x}_i^T\vec{\beta}=\sum_{j=0}^{p-1}x_{j,i}\beta_j
\end{aligned}\tag{14.1.1031}$$

We notice that the assumptions of use of the Binomial GLM remain trivially the independence of the explanatory variables, the homoscedasticity is on the other hand no more assumed, the linearity of the model is always a hypothesis and the expected mean is linear under a logit transformation .

We will not develop here the likelihood model of the binomial regression as we already did it during our study of the binomial logistic regression on page 6363.

14.1.12.7.4 Binomial Negative GLM

Negative binomial regression (sometimes abbreviated in general as N.B.R.M. for Negative Binomial Regression Model) is probably the most widely used in practice outside binomial logistic regression or Gaussian regression. However it suffers from a major problem ... Indeed there are about twenty derived models such as: NB-C (the simplest that we will study here), NB-1, NB-2 (generalization of GLM Poisson), truncated NB / NB-1 / NB-C, NB with zero inflation (ZINB), censored NB, NB with obstacle, NB-P, heterogeneous-NB (NBH), finite mixed model, conditional fixed effects model, random effects model, NB endogenous stratification, NB sampling, NB latent class, NB bivariate, etc. In short enough to write several thousand pages and study for several years ...

Well this being said, let's take a look at the case of canonical negative binomial regression.

We thus start from one of the possible expressions of the density function of the negative binomial law (see the section Statistics page 667 for the demonstration of its origin):

$$\begin{aligned}
P(E,N,p)&=\binom{E+R-1}{E-1}q^Rp^E=\frac{(E+R-1)!}{(E-1)!(E+R-1-(E-1))!}q^Rp^E\\
&=\binom{E+R-1}{E-1}q^Rp^E
\end{aligned}\tag{14.1.1032}$$

We then adapt this notation to the GLM sauce, admitting that we are interested in failures E:

$$\begin{aligned}
Y_i&=\frac{(y_i+R-1)!}{(y_i-1)!R!}q_i^R(1-q_i)^{y_i}=\frac{((\alpha x_i+\beta)+R-1)!}{((\alpha x_i+\beta)-1)!R!}q_i^R(1-q_i)^{\alpha x_i+\beta}\\
&=C_{y_i-1}^{y_i+R-1}q_i^R(1-q_i)^{y_i}=\exp\left(\ln\left(C_{y_i-1}^{y_i+R-1}q_i^R(1-q_i)^{y_i}\right)\right)\\
&=\exp\left(\ln\left((1-q_i)^{y_i}\right)+\ln\left(q_i^R\right)+\ln\left(C_{y_i-1}^{y_i+R-1}\right)\right)\\
&=\exp\left(\frac{y_i\ln(1-q_i)+R\ln(q_i)}{1}+\ln\left(C_{y_i-1}^{y_i+R-1}\right)\right)
\end{aligned}\tag{14.1.1033}$$

Now, let's do the identification with:

$$f(y_i,\theta_i,\varphi,\omega_i)=\exp\left(\frac{y_i\theta_i-b(\theta_i)}{a_i(\varphi)}+c(y_i,\varphi)\right) \tag{14.1.1034}$$

Therefore:

$$\begin{aligned}\theta_i&=\ln(1-q_i)\Rightarrow q_i=1-e^{\theta_i}\\ b(\theta_i)&=b\left(\ln(1-q_i)\right)=-R\ln(q_i)=-R\ln\left(1-e^{\theta_i}\right)\\ a_i(\varphi)&=1\\ c(y_i,\varphi)&=\ln\left(C_{y_{i-1}}^{y_i+R-1}\right)\\ \varphi&=1\end{aligned} \tag{14.1.1035}$$

Now, let's have a look at what gives the derivatives used earlier above:

$$\begin{aligned}\frac{\mathrm{d}b(\theta_i)}{\mathrm{d}\theta_i}&=\frac{\mathrm{d}(-R\ln(q_i))}{\mathrm{d}\theta_i}=-R\frac{\mathrm{d}(-R\ln(1-e^{\theta_i}))}{\mathrm{d}\theta_i}=R\frac{e^{\theta_i}}{1-\theta_i}\\ &=R\frac{e^{\ln(1-q_i)}}{1-e^{\ln(1-q_i)}}=R\frac{1-q_i}{1-(1-q_i)}=\frac{R(1-q_i)}{q_i}=\mu_i\end{aligned} \tag{14.1.1036}$$

Which corresponds well to the expected mean of the number of failures before the R-th success (result to be compared with the expected mean of the number of successes of the negative binomial law as demonstrated in the section Statistics) since it is this that we seek to model by linear regression. Do not forget that we can reverse the roles failures / successes.

And:

$$\begin{aligned}\frac{\mathrm{d}^2b(\theta_i)}{\mathrm{d}\theta_i^2}a_i(\varphi)&=\frac{\mathrm{d}\left(R\dfrac{e^{\theta_i}}{1-e^{\theta_i}}\right)}{\mathrm{d}\theta_i}\cdot 1=R\frac{\mathrm{d}\left(e^{\theta_i}\dfrac{1}{1-e^{\theta_i}}\right)}{\mathrm{d}\theta_i}\\ &=R\left(e^{\theta_i}\frac{1}{1-e^{\theta_i}}+e^{\theta_i}\frac{e^{\theta_i}}{(1-e^{\theta_i})^2}\right)=R\left(\frac{e^{\theta_i}(1-e^{\theta_i})}{(1-e^{\theta_i})^2}+\frac{e^{\theta_i}e^{\theta_i}}{(1-e^{\theta_i})^2}\right)\\ &=R\left(\frac{e^{\theta_i}(1-e^{\theta_i})+e^{\theta_i}e^{\theta_i}}{(1-e^{\theta_i})^2}\right)=R\frac{e^{\theta_i}}{(1-e^{\theta_i})^2}=R\left(\frac{e^{\theta_i}}{1-e^{\theta_i}}\right)\left(\frac{1}{1-e^{\theta_i}}\right)\\ &=R\left(\frac{e^{\ln(1-q_i)}}{1-e^{\ln(1-q_i)}}\right)\left(\frac{1}{1-e^{\ln(1-q_i)}}\right)=R\left(\frac{1-q_i}{1-(1+q_i)}\right)\left(\frac{1}{1-(1+q_i)}\right)\\ &=R\left(\frac{1-q_i}{-q_i}\right)\left(\frac{1}{-q_i}\right)=\frac{R(1-q_i)}{q_i^2}=\sigma^2\end{aligned} \tag{14.1.1037}$$

Which corresponds well to the variance of the number of failures (result to be compared with the variance of the number of successes of the negative binomial law as demonstrated in the section Statistics) since this is what we seek to model by the regression.

We then have:

$$\frac{\mathrm{d}b(\theta_i)}{\mathrm{d}\theta_i}=\frac{R(1-q_i)}{q_i}=\mu_i \tag{14.1.1038}$$

Let's now make a usual change of notation:

$$\frac{\mathrm{d}b(\theta_i)}{\mathrm{d}\theta_i}=\frac{(1-q_i)}{\alpha q_i}=\mu_i\Leftrightarrow\alpha\mu_i=\frac{(1-q_i)}{q_i}=\frac{1}{q_i}-1\Leftrightarrow q_i=\frac{1}{\alpha\mu_i+1}\tag{14.1.1039}$$

Which is also sometimes denoted:

$$q_i=\frac{1}{\beta_i+1}\tag{14.1.1040}$$

In this form, we can rewrite:

$$\begin{aligned}P(E,N,p)&=\binom{y_i+R-1}{y_i-1}q^Rp^{y_i}=\binom{y_i+\dfrac{1}{\alpha}-1}{y_i-1}\left(\frac{1}{\alpha\mu_i+1}\right)^{\frac{1}{\alpha}}\left(1-\frac{1}{\alpha\mu_i+1}\right)^{y_i}\\&=\binom{y_i+\frac{1}{\alpha}-1}{y_i-1}\left(\frac{1}{\alpha\mu_i+1}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\mu_i}{\alpha\mu_i+1}\right)^{y_i}=\binom{y_i+\frac{1}{\alpha}-1}{y_i-1}\left(\frac{1}{\alpha\mu_i+1}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\mu_i}{\alpha\mu_i+1}\right)^{y_i}\end{aligned}\tag{14.1.1041}$$

If we were more interested in successes rather than failures we would have (this is the case most often chosen in the specialized literature):

$$\begin{aligned}P(E,N,p)&=\binom{E+y_i-1}{E-1}q^{y_i}p^E=\binom{\frac{1}{\alpha}+y_i-1}{\frac{1}{\alpha}-1}q^{y_i}p^{\frac{1}{\alpha}}\\&=\binom{\frac{1}{\alpha}+y_i-1}{\frac{1}{\alpha}-1}\left(\frac{\alpha\mu_i}{\alpha\mu_i+1}\right)^{y_i}\left(\frac{1}{\alpha\mu_i+1}\right)^{\frac{1}{\alpha}}\end{aligned}\tag{14.1.1042}$$

Then, it is customary to write this with the Gamma Euler function (see section page 1574). What then gives (with different common notations):

$$\begin{aligned}P(E,N,p)&=\binom{\frac{1}{\alpha}+y_i-1}{\frac{1}{\alpha}-1}\left(\frac{\alpha\mu_i}{\alpha\mu_i+1}\right)^{y_i}\left(\frac{1}{\alpha\mu_i+1}\right)^{\frac{1}{\alpha}}\\&=\frac{\left(\frac{1}{\alpha}+y_i-1\right)!}{\left(\frac{1}{\alpha}-1\right)!(y_i)!}\left(\frac{1}{\alpha\mu_i+1}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\mu_i}{\alpha\mu_i+1}\right)^{y_i}\\&=\frac{\Gamma\left(\frac{1}{\alpha}+y_i\right)}{\Gamma\left(\frac{1}{\alpha}\right)\Gamma(y_i+1)}\left(\frac{1}{\alpha\mu_i+1}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\mu_i}{\alpha\mu_i+1}\right)^{y_i}=\frac{\Gamma(r+k)}{\Gamma(r)\Gamma(k+1)}\left(\frac{1}{\theta+1}\right)^r\left(\frac{\theta}{\theta+1}\right)^k\\&=\frac{\Gamma(R+k)}{\Gamma(R)k!}\left(\frac{1}{\theta+1}\right)^R\left(\frac{\theta}{\theta+1}\right)^k=\frac{\Gamma(R+k)}{\Gamma(R)k!\theta^r}\left(\frac{\theta}{\theta+1}\right)^{k+R}\\&=\left(\frac{\theta}{\theta+1}\right)^{k+R}\frac{1}{\Gamma(R)k!\theta^R}\Gamma(R+k)\end{aligned}\tag{14.1.1043}$$

In this form, we immediately see that expected mean is written then:

$$\mu=R\cdot\theta\tag{14.1.1044}$$

Now, to return to our previous development, let us recall that we have proved in the section of Differential and Integral Calculus during our study of the Euler's Gamma function (see section

Differential and Integral Calculus page 1574) that:

$$\Gamma(x+1)=\int\limits_{0}^{+\infty} t^x e^{-t}\mathrm{d}t \tag{14.1.1045}$$

We then have by extension:

$$\Gamma(x+1)=\Gamma((k+R-1)+1)=\int\limits_{0}^{+\infty} t^{k+R-1}e^{-t}\mathrm{d}t \tag{14.1.1046}$$

It then comes in this case:

$$\begin{aligned}P(X=k,R,\theta)&=\left(\frac{\theta}{\theta+1}\right)^{k+R}\frac{1}{\Gamma(R)k!\theta^r}\int\limits_{0}^{+\infty} t^{k+R-1}e^{-t}\mathrm{d}t\\&=\left(\frac{\theta}{\theta+1}\right)^{k+R}\int\limits_{0}^{+\infty}\frac{t^{k+R-1}e^{-t}}{\Gamma(R)k!\theta^R}\mathrm{d}t\end{aligned} \tag{14.1.1047}$$

Let us integrate with the following variable change:

$$t=\lambda\frac{\theta+1}{\theta} \tag{14.1.1048}$$

We then have:

$$\begin{aligned}P(X=k,R,\theta)&=\left(\frac{\theta}{\theta+1}\right)^{k+R}\int\limits_{0}^{+\infty}\frac{t^{k+R-1}e^{-t}}{\Gamma(R)k!\theta^R}\mathrm{d}t\\&=\left(\frac{\theta}{\theta+1}\right)^{k+R}\int\limits_{0}^{+\infty}\frac{\left(\lambda\frac{\theta+1}{\theta}\right)^{k+R-1}e^{-\lambda\frac{\theta+1}{\theta}}}{\Gamma(R)k!\theta^R}\frac{\theta+1}{\theta}\mathrm{d}\lambda\\&=\left(\frac{\theta}{\theta+1}\right)^{k+R-1}\left(\frac{\theta+1}{\theta}\right)^{k+R-1}\int\limits_{0}^{+\infty}\frac{\lambda^{k+R-1}e^{-\lambda\frac{\theta+1}{\theta}}}{\Gamma(R)k!\theta^R}\mathrm{d}\lambda\\&=\int\limits_{0}^{+\infty}\frac{\lambda^{k+R-1}e^{-\lambda\frac{\theta+1}{\theta}}}{\Gamma(R)k!\theta^R}\mathrm{d}\lambda=\int\limits_{0}^{+\infty}\frac{\lambda^k e^{-\lambda}}{k!}\frac{\lambda^{R-1}e^{-\frac{\lambda}{\theta}}}{\Gamma(R)\theta^R}\mathrm{d}\lambda\end{aligned} \tag{14.1.1049}$$

We find under the integral two density functions which are respectively the Poisson density function and the Gamma density function (see section Statistics page 686 and page 767):

$$P(X=k,R,\theta)=\int\limits_{0}^{+\infty}\underbrace{\frac{\lambda^k e^{-\lambda}}{k!}}_{\text{Poisson law}}\underbrace{\frac{\lambda^{R-1}e^{-\frac{\lambda}{\theta}}}{\Gamma(R)\theta^R}}_{\text{Gamma law}}\mathrm{d}\lambda \tag{14.1.1050}$$

We have a result that it is customary to associate with a generalization of the Poisson's law and which we name "Poisson-Gamma density law" or "mixed Poisson-Gamma law".

Remark

Thus, the negative binomial distribution can be considered as a generalization of the Poisson distribution. Indeed, the negative binomial law can be seen as a Poisson distribution where the Poisson parameter is itself a random variable, distributed according to a Gamma distribution !!!

An important application is in the area of risk insurance where k the number of undesirable events is modelled by a Poisson law of parameter λ which itself is therefore tainted by an uncertainty that is characteristic of the risk category of the insured. Thus, the parameter λ can itself be seen as a random variable traditionally denoted Θ in the field of actuarial science and following a gamma law such that the density function of Θ is then given by (with the traditional notation in actuarial science):

$$\Theta=\int\limits_0^{+\infty}\underbrace{\frac{\theta^k e^{-\theta}}{n!}}_{\text{Poisson law}}\underbrace{\frac{\alpha^{\beta-1}e^{-\frac{\theta}{\alpha}}}{\Gamma(\beta)\alpha^\beta}}_{\text{Gamma law}}\mathrm{d}\lambda \tag{14.1.1051}$$

For the parameters to be estimated, it is customary to also use the maximum log-likelihood as we did for all previous regressions. Then, in order to determine the maximum of the log-likelihood, it is traditional in the case of the negative binomial law to start from the following form demonstrated above:

$$P(\alpha,y_i,\mu_i)=\frac{\Gamma\left(\frac{1}{\alpha}+y_i\right)}{\Gamma\left(\frac{1}{\alpha}\right)\Gamma(y_i+1)}\left(\frac{1}{\alpha\mu_i+1}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\mu_i}{\alpha\mu_i+1}\right)^{y_i} \tag{14.1.1052}$$

We then have:

$$\ln[L(\alpha,y_i,\mu_i)]=\sum_{i=1}^n\ln\left[\frac{\Gamma\left(\frac{1}{\alpha}+y_i\right)}{\Gamma\left(\frac{1}{\alpha}\right)\Gamma(y_i+1)}\left(\frac{1}{\alpha\mu_i+1}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\mu_i}{\alpha\mu_i+1}\right)^{y_i}\right] \tag{14.1.1053}$$

The log-likelihood then becomes:

$$\begin{aligned}
&\ln[L(\alpha,y_i,\mu_i)]\\
&=\sum_{i=1}^n\left(\ln\left(\Gamma\left(\frac{1}{\alpha}+y_i\right)\right)-\ln\left(\Gamma\left(\frac{1}{\alpha}\right)\right)-\ln\left(\Gamma(y_i+1)\right)+\frac{1}{\alpha}\ln\left(\frac{1}{\alpha\mu_i+1}\right)+y_i\ln\left(\frac{\alpha\mu_i}{\alpha\mu_i+1}\right)\right)\\
&=\sum_{i=1}^n\left(\ln\left(\Gamma\left(\frac{1}{\alpha}+y_i\right)\right)-\ln\left(\Gamma\left(\frac{1}{\alpha}\right)\right)-\ln\left(\Gamma(y_i+1)\right)-\frac{1}{\alpha}\ln\left(\alpha\mu_i+1\right)+y_i\ln(\alpha\mu_i)-y_i(\alpha\mu_i+1)\right)\\
&=\sum_{i=1}^n\left(\ln\left(\Gamma\left(\frac{1}{\alpha}+y_i\right)\right)-\ln\left(\Gamma\left(\frac{1}{\alpha}\right)\right)-\ln\left(\Gamma(y_i+1)\right)-\left(\frac{1}{\alpha}+y_i\right)\ln\left(\alpha\mu_i+1\right)+y_i\ln(\alpha\mu_i)\right)
\end{aligned} \tag{14.1.1054}$$

Now to proceed as we did with the Gaussian linear regression, let us recall that the y_i are given by the experience. We will write then:

$$\begin{aligned}
&\ln[L(\alpha,Y_i,\mu_i)]\\
&=\sum_{i=1}^n\left(\ln\left(\Gamma\left(\frac{1}{\alpha}+Y_i\right)\right)-\ln\left(\Gamma\left(\frac{1}{\alpha}\right)\right)-\ln\left(\Gamma(Y_i+1)\right)+\left(\frac{1}{\alpha}+Y_i\right)\ln\left(\alpha\mu_i+1\right)+Y_i\ln(\alpha\mu_i)\right)
\end{aligned} \tag{14.1.1055}$$

and that, as for the Gaussian linear regression, the explicit expression of the μ_i must allow us to determine the expression of the parameter α of the regression.

Now, let us recall that:

$$P(k,R,p)=\frac{\Gamma(R+k)}{\Gamma(R)\Gamma(k+1)}(1-p)^Rp^k=\frac{\Gamma(R+k)}{\Gamma(R)k!}(1-p)^Rp^k \tag{14.1.1056}$$

and that:

$$\mu=\frac{R(1-q)}{q}=\frac{Rp}{1-p} \tag{14.1.1057}$$

and let us denoted this λ. Then we get:

$$\lambda=R\frac{p}{1-p}\Leftrightarrow p=\frac{\lambda}{R+\lambda} \tag{14.1.1058}$$

Therefore:

$$\begin{aligned}P(k,R,p)&=\frac{\Gamma(R+k)}{\Gamma(R)k!}(1-p)^Rp^k=\frac{\Gamma(R+k)}{\Gamma(R)k!}\left(1-\frac{\lambda}{R+\lambda}\right)^R\frac{\lambda^k}{(R+\lambda)^k}\\&=\frac{\lambda^k}{k!}\frac{\Gamma(R+k)}{\Gamma(R)(R+\lambda)^k}\left(\frac{R}{R+\lambda}\right)^R=\frac{\lambda^k}{k!}\frac{\Gamma(R+k)}{\Gamma(R)(R+\lambda)^k}\frac{1}{\left(1+\frac{\lambda}{R}\right)^R}\\&=\frac{\lambda^k}{k!}\frac{\Gamma(R+k+1)!}{\Gamma(R)(R+\lambda)^k}\frac{1}{\left(1+\frac{\lambda}{R}\right)^R}\end{aligned} \tag{14.1.1059}$$

Now let us put that $R\rightarrow+\infty$. It comes then:

$$\begin{aligned}\frac{\Gamma(R+k)}{\Gamma(R)(R+\lambda)^k}&=\frac{(R+k)!}{R!(R+\lambda)^k}=\frac{R!(R+k)\cdot\ldots\cdot(R+k)}{R!(R+\lambda)^k}\\&=\frac{R+1}{R+\lambda}\cdot\frac{R+2}{R+\lambda}\cdot\ldots\cdot\frac{R+k}{R+\lambda}\underset{\substack{R\rightarrow+\infty\\\lambda\in\mathbb{R}^+}}{=}1\end{aligned} \tag{14.1.1060}$$

since each term of the product tends to 1 and since there is a constant number of terms in the product (that is, k terms) the whole tends to 1.

And as we saw in the section of Functional Analysis (see page 2291):

$$\left(1+\frac{\lambda}{R}\right)^R\underset{\substack{R\rightarrow+\infty\\\lambda\in\mathbb{R}^+}}{=}e^{\lambda} \tag{14.1.1061}$$

Therefore:

$$P(k,R,p)=\frac{\lambda^k}{k!}\frac{\Gamma(R+k+1)!}{\Gamma(R)(R+\lambda)^k}\frac{1}{\left(1+\frac{\lambda}{R}\right)^R}\underset{\substack{R\rightarrow+\infty\\\lambda\in\mathbb{R}^+}}{=}\frac{\lambda^k}{k!}\cdot1\cdot\frac{1}{e^{\lambda}}=\frac{\lambda^k}{k!}e^{-\lambda} \tag{14.1.1062}$$

We thus fall back in this limit case on the Poisson law!

Therefore, as the Poisson law is a special case of the Negative Binomial law, a possible model for the link function of the Negative Binomial Law is to take up the logarithm (hence the Poisson regression link function).

The canonical function is then that of the Poisson function:

$$g(\mathrm{E}(Y))=\theta_i=\ln(\mathrm{E}(Y))=\ln(\mu_i)=\vec{x}_i^T\vec{\beta}=\sum_{j=0}^{p-1}x_{j,i}\beta_j\Rightarrow\mu_i=e^{\vec{x}_i^T\vec{\beta}}=e^{\theta_i} \tag{14.1.1063}$$

From then on, the log-likelihood will be written:

$$\begin{aligned}&\ln[L(\alpha,Y_i,\mu_i)]\\&=\sum_{i=1}^{n}\left(\ln\left(\Gamma\left(\frac{1}{\alpha}+Y_i\right)\right)-\ln\left(\Gamma\left(\frac{1}{\alpha}\right)\right)-\ln\left(\Gamma(Y_i+1)\right)-\left(Y_i+\frac{1}{\alpha}\right)\ln\left(\alpha\mu_i+1\right)+Y_i\ln(\alpha\mu_i)\right)\\&=\sum_{i=1}^{n}\left(\ln\left(\Gamma\left(\frac{1}{\alpha}+Y_i\right)\right)-\ln\left(\Gamma\left(\frac{1}{\alpha}\right)\right)-\ln\left(\Gamma(Y_i+1)\right)-\left(Y_i+\frac{1}{\alpha}\right)\ln\left(\alpha e^{\vec{x}_i^T\vec{\beta}}+1\right)\right.\\&\qquad\left.+Y_i\ln(\alpha e^{\vec{x}_i^T\vec{\beta}})\right)=\ln[L(\alpha,Y_i,\vec{x},\vec{\beta})]\end{aligned} \tag{14.1.1064}$$

that it is customary to name the "NB2 log-likelihood" (which is the one used in the majority of statistical softwares).

But there is another version of log-likelihood that uses a combination of approximation and exact value. Indeed, remember that we had obtained:

$$g(\mathrm{E}(Y))=\theta_i=\ln(1-q_i)=\ln\left(1-\frac{1}{\alpha\mu_i+1}\right)=\ln\left(\frac{\alpha\mu_i}{\alpha\mu_i+1}\right)=-\ln\left(\frac{1}{\alpha\mu_i}+1\right) \tag{14.1.1065}$$

Therefore, using:

$$\theta_i=-\ln\left(\frac{1}{\alpha\mu_i}+1\right)\Leftrightarrow\mu_i=\frac{1}{\alpha\left(e^{-\theta_i}-1\right)} \tag{14.1.1066}$$

and by mixing (rather dubious mix...) with the approximation approach using Poisson link function:

$$\theta_i=\ln(\mu_i)=\vec{x}_i^T\vec{\beta} \tag{14.1.1067}$$

We have (happy mix ...):

$$\mu_i=\frac{1}{\alpha\left(e^{-\theta_i}-1\right)}=\frac{1}{e^{-\vec{x}_i^T\vec{\beta}}-1} \tag{14.1.1068}$$

The log-likelihood is written then:

$$\begin{aligned}&\ln[L(\alpha,Y_i,\mu_i)]\\&=\sum_{i=1}^{n}\left(\ln\left(\Gamma\left(\frac{1}{\alpha}+Y_i\right)\right)-\ln\left(\Gamma\left(\frac{1}{\alpha}\right)\right)-\ln\left(\Gamma(Y_i+1)\right)-\left(Y_i+\frac{1}{\alpha}\right)\ln\left(\alpha\mu_i+1\right)+Y_i\ln(\alpha\mu_i)\right)\\&=\sum_{i=1}^{n}\left(\ln\left(\Gamma\left(\frac{1}{\alpha}+Y_i\right)\right)-\ln\left(\Gamma\left(\frac{1}{\alpha}\right)\right)-\ln\left(\Gamma(Y_i+1)\right)-\left(Y_i+\frac{1}{\alpha}\right)\ln\left(\frac{1}{e^{-\vec{x}_i^T\vec{\beta}}-1}+1\right)\right.\\&\qquad\left.+Y_i\ln\left(\frac{1}{e^{-\vec{x}_i^T\vec{\beta}}-1}\right)\right)\end{aligned} \tag{14.1.1069}$$

that is customary named the "NB-C log-likelihood".

To solve this kind of log-likelihood, we then use again Newton-Raphson type algorithms (see further below in this section page 6673).

Let us compare, the generalized form of the Normal law of regression with that of Poisson, of the Binomial and Negative Binomial law:

$$\begin{aligned}
Y_{i,\mathcal{N}} &= \exp\left(\frac{y_i\mu_i-\mu_i^2/2}{\sigma^2}-\frac{1}{2}\left(\frac{y_i^2}{\sigma^2}-\ln\left(2\pi\sigma^2\right)\right)\right)\\
Y_{i,\mathcal{P}} &= \exp\left(\left(\frac{y_i\ln(\mu_i)-\mu_i}{1}\right)-\ln\left(y_i!\right)\right)\\
Y_{i,\mathcal{B}} &= \exp\left(\left(\frac{y_i\ln\left(\dfrac{p}{1-p}\right)+N\ln\left(1-p\right)}{1}\right)+\ln\left(C_{y_i}^N\right)\right)\\
Y_{i,\text{NB-C}} &= \exp\left(\frac{y_i\ln\left(1-q_i\right)+R\ln\left(q_i\right)}{1}+\ln\left(C_{y_i-1}^{y_i+R-1}\right)\right)\\
Y_{i,\text{NB}} &= \exp\left(\frac{y_i\ln\left(1-q_i\right)+R\ln\left(q_i\right)}{1}+\ln\left(C_{y_i-1}^{y_i+R-1}\right)\right)
\end{aligned} \tag{14.1.1070}$$

and also the link functions:

$$\begin{aligned}
g_{\mathcal{N}}(\mu_i) &= \theta_i=\mu_i=\vec{x}_i^T\vec{\beta}=\sum_{j=0}^{p-1}x_{j,i}\beta_j\\
g_{\mathcal{P}}(\mu_i) &= \theta_i=\ln(\mu_i)=\vec{x}_i^T\vec{\beta}=\sum_{j=0}^{p-1}x_{j,i}\beta_j\\
g_{\mathcal{B}}(\mu_i) &= \theta_i=\ln\left(\frac{p}{1-p}\right)=\vec{x}_i^T\vec{\beta}=\sum_{j=0}^{p-1}x_{j,i}\beta_j\\
g_{\text{NB-C}}(\mu_i) &= \theta_i=\ln\left(1-\frac{1}{\alpha\mu_i+1}\right)=-\ln\left(\frac{1}{\displaystyle\sum_{j=0}^{p-1}x_{j,i}\beta'_j}+1\right)
\end{aligned} \tag{14.1.1071}$$

We notice that the hypotheses of use of the Binomial Negative GLM remain trivially the independence of the explanatory variables, the homoscedasticity is on the other hand no more assumed, the linearity of the model is always a hypothesis and the expected mean is linear only under a logarithmic transformation chosen a little bit empirically in the canonical case...

14.1.12.7.5 Gamma GLM

Let's do the same as before but now with the Gamma law (see section Statistics page 767). We therefore start from:

$$\Gamma(a,\lambda)=\frac{x^{a-1}e^{-\lambda x}}{\Gamma(a)}\lambda^{a}\cdot 1_{]0,+\infty]} \qquad (14.1.1072)$$

by just changing the notation a bit to not confuse with one of the GLM parameters:

$$\Gamma(\alpha,\lambda)=\frac{x^{\alpha-1}e^{-\lambda x}}{\Gamma(\alpha)}\lambda^{\alpha}\cdot 1_{]0,+\infty]} \qquad (14.1.1073)$$

assuming that the scale parameter λ is known (...). Why? Because the expected mean of the Gamma law is for recall given by (see the proof in the section Statistics page 767):

$$\mu=\frac{\alpha}{\lambda} \qquad (14.1.1074)$$

and therefore controlling only one of the two parameters is enough to scan the entire desired definition domain (yes, simply!).

The use of the Gamma law is justified, besides by its domain of definition which is not in the negatives, by the fact that the relation between its expected mean and its variance gives:

$$\frac{\mu}{\sigma^2}=\frac{\frac{\alpha}{\lambda}}{\frac{\alpha}{\lambda^2}}=\lambda \qquad (14.1.1075)$$

So the relationship between expected mean is the variance is constant regardless of their respective values (so when mean increases, the variance increase too as for the Poisson distribution)! Another way of looking at this is that the variance is proportional to the expected mean since:

$$\sigma^2=\frac{\mu}{\lambda} \qquad (14.1.1076)$$

The reason why the GLM Gamma is interesting is that we can "adjust" the relationship between variance and expected mean by choosing the value of λ, which other models seen so far can not do!

Therefore we get:

$$\begin{aligned}Y_i&=\frac{y_i^{\alpha_i-1}e^{-\lambda y_i}}{\Gamma(\alpha_i)}\lambda^{\alpha_i}\cdot 1_{]0,+\infty]}=\exp\left(\ln\left(\frac{y_i^{\alpha_i-1}e^{-\lambda y_i}\lambda^{\alpha_i}}{\Gamma(\alpha_i)}\right)\right)\\&=\exp\left(\ln(y_i^{\alpha_i-1})+\ln(e^{-\lambda y_i})+\ln(\lambda^{\alpha_i})-\ln\left(\Gamma(\alpha_i)\right)\right)\\&=\exp\left((\alpha_i-1)\ln(y_i)-\lambda y_i+\alpha_i\ln(\lambda)-\ln\left(\Gamma(\alpha_i)\right)\right)\\&=\exp\left(\frac{(-\lambda y_i+\alpha_i\ln(\lambda))}{1}+(\alpha_i-1)\ln(y_i)-\ln\left(\Gamma(\alpha_i)\right)\right)\end{aligned} \qquad (14.1.1077)$$

Now, let us as always the identification with:

$$f(y_i,\theta_i,\varphi,\omega_i)=\exp\left(\frac{y_i\theta_i-b(\theta_i)}{a_i(\varphi)}+c(y_i,\varphi)\right) \qquad (14.1.1078)$$

We therefore get:

$$\begin{aligned}\theta_i&=\theta=-\lambda\\ b(\theta_i)&=b(-\lambda)=-\alpha_i\ln(\lambda)\\ a_i(\varphi)&=1\\ c(y_i,\varphi)&=\alpha_i\ln(y_i)-\ln\left(\Gamma(\alpha_i)\right)-\ln(y_i)\end{aligned}\tag{14.1.1079}$$

Now, let's have a look at what gives the derivatives used earlier above:

$$\frac{\mathrm{d}b(\theta_i)}{\mathrm{d}\theta_i}=\frac{\mathrm{d}\left(-\alpha_i\ln(\lambda)\right)}{-\mathrm{d}\lambda}=-\alpha_i\frac{\mathrm{d}\left(\ln(\lambda)\right)}{-\mathrm{d}\lambda}=\alpha_i\frac{\mathrm{d}\left(\ln(\lambda)\right)}{\mathrm{d}\lambda}=\frac{\alpha_i}{\lambda}=\mu\tag{14.1.1080}$$

and:

$$\frac{\mathrm{d}^2b(\theta_i)}{\mathrm{d}\theta_i^2}a_i(\varphi)=\frac{\mathrm{d}\left(\frac{\alpha_i}{\lambda}\right)}{\mathrm{d}\theta_i}\cdot 1=\frac{\mathrm{d}\left(\frac{\alpha_i}{\lambda}\right)}{-\mathrm{d}\lambda}=-\alpha_i\frac{\mathrm{d}\left(\frac{1}{\lambda}\right)}{\mathrm{d}\lambda}=\frac{\alpha_i}{\lambda^2}\tag{14.1.1081}$$

So we can once again fall back on the expected mean and variance but this time of the Gamma law!

Let us also indicate that compared to the famous "canonical parameter" (since expected mean and variance ultimately depend on it), we have in this case:

$$\theta_i=-\lambda=-\frac{\alpha_i}{\mu_i}\tag{14.1.1082}$$

so the canonical function is:

$$g(\mathrm{E}(Y))=-\frac{\alpha_i}{\mu_i}=\frac{-\alpha_i}{\sum\limits_{j=0}^{p-1}x_{j,n}\beta_j}=\frac{1}{\frac{1}{-\alpha_i}\sum\limits_{j=0}^{p-1}x_{j,n}\beta_j}=\frac{1}{\sum\limits_{j=0}^{p-1}x_{j,n}\frac{\beta_j}{-\alpha_i}}=\frac{1}{\sum\limits_{j=0}^{p-1}x_{j,n}\beta'_j}\tag{14.1.1083}$$

in this case, the canonical function is named the "negative inverse link" and is often denoted in the following form (because finally put that $-\alpha_i=1$ is simply equivalent to change the values of the parameters β_j):

$$g(\mathrm{E}(Y))=\frac{1}{\mu_i}=\frac{1}{\sum\limits_{j=0}^{p-1}x_{j,i}\beta_j}\tag{14.1.1084}$$

So in view of this expression we can make regression of very various type of curves.

Let us now compare the generalized form of the Normal law of the regression with that of Poisson and the Binomial law and the Gamma law:

$$\begin{aligned}Y_{i,\mathcal{N}}&=\exp\left(\frac{y_i\mu_i-\mu_i^2/2}{\sigma^2}-\frac{1}{2}\left(\frac{y_i^2}{\sigma^2}-\ln\left(2\pi\sigma^2\right)\right)\right)\\ Y_{i,\mathcal{P}}&=\exp\left(\left(\frac{y_i\ln(\mu_i)-\mu_i}{1}\right)-\ln\left(y_i!\right)\right)\\ Y_{i,\mathcal{B}}&=\exp\left(\left(\frac{y_i\ln\left(\frac{p}{1-p}\right)+N\ln\left(1-p\right)}{1}\right)+\ln\left(C_{y_i}^N\right)\right)\\ Y_{i,\text{NB-C}}&=\exp\left(\frac{y_i\ln\left(1-q_i\right)+R\ln\left(q_i\right)}{1}+\ln\left(C_{y_{i-1}}^{y_i+R-1}\right)\right)\\ Y_{i,\Gamma}&=\exp\left(\frac{\left(-y_i\lambda+\alpha_i\ln(\lambda)\right)}{1}+\alpha_i\ln(y_i)-\ln\left(\Gamma(\alpha_i)\right)-\ln(y_i)\right)\end{aligned}\tag{14.1.1085}$$

and also the link functions:

$$\begin{gathered}
g_{\mathcal{N}}(\mu_i)=\theta_i=\mu_i=\vec{x}_i^T\vec{\beta}=\sum_{j=0}^{p-1}x_{j,i}\beta_j\\
g_{\mathcal{P}}(\mu_i)=\theta_i=\ln(\mu_i)=\vec{x}_i^T\vec{\beta}=\sum_{j=0}^{p-1}x_{j,i}\beta_j\\
g_{\mathcal{B}}(\mu_i)=\theta_i=\ln\left(\dfrac{p}{1-p}\right)=\vec{x}_i^T\vec{\beta}=\sum_{j=0}^{p-1}x_{j,i}\beta_j\\
g_{\text{NB-C}}(\mu_i)=\theta_i=\ln\left(1-\dfrac{1}{\alpha\mu_i+1}\right)=-\ln\left(\dfrac{1}{\displaystyle\sum_{j=0}^{p-1}x_{j,i}\beta'_j}+1\right)\\
g_{\Gamma(\mu_i)}=\dfrac{1}{\vec{x}_i^T\vec{\beta}'}=\dfrac{1}{\displaystyle\sum_{j=0}^{p-1}x_{j,i}\beta'_j}
\end{gathered} \tag{14.1.1086}$$

We notice that the hypotheses of use of the GLM Gamma remain trivially the independence of the explanatory variables, the homoscedasticity is on the other hand no more assumed, the linearity of the model is always a hypothesis and the expected mean is linear only under an inverse transformation.

Obviously (as always...) the subject is more complex than it may seems. There are plenty of other GLM models that exist. The reader can see a sample of the most known one in the table below:

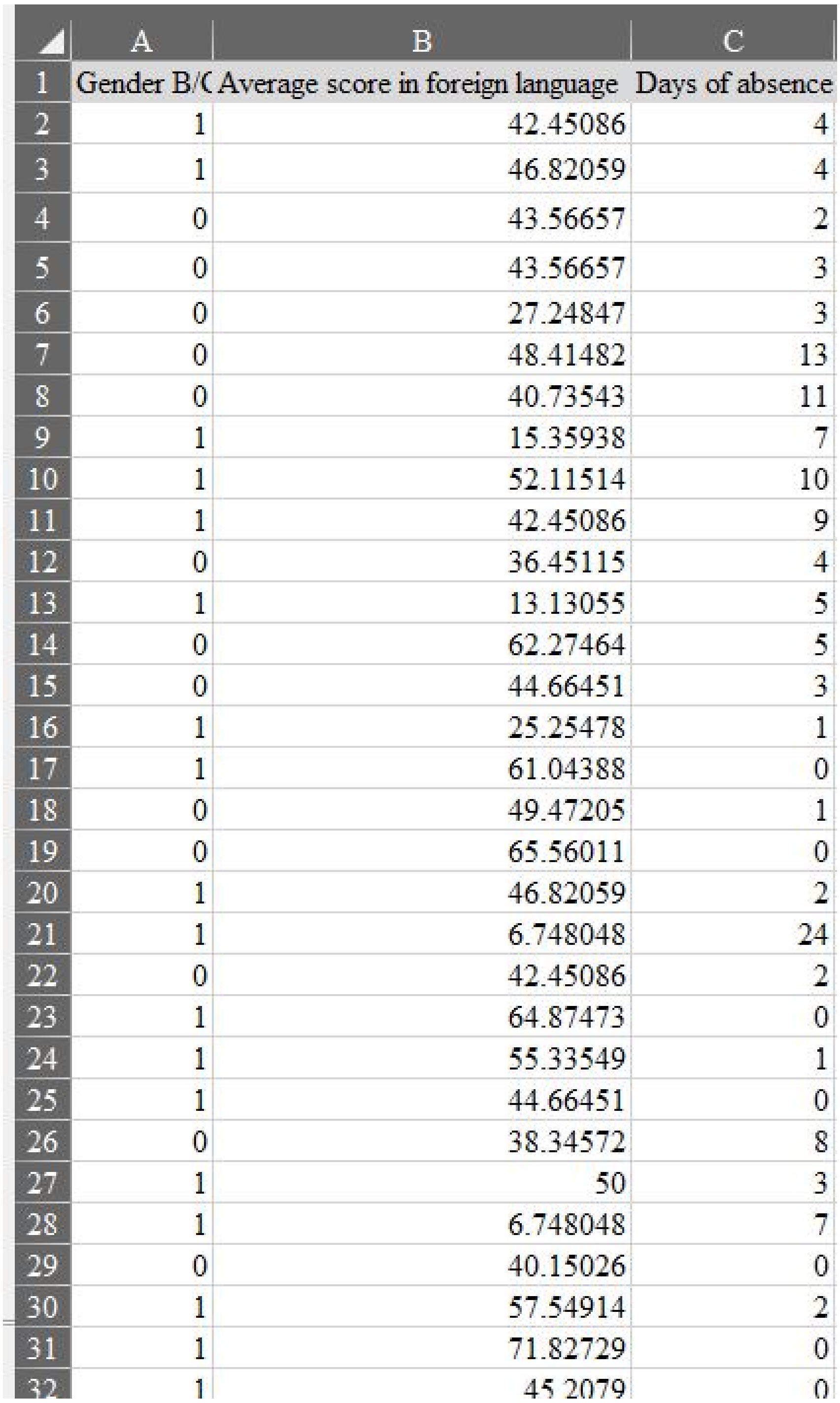
Family	Link type	Mean function
Gaussian	Identity	$\mu_i=\vec{x}_i^T\vec{\beta}$
Binomial	Logit (logistic)	$\mu_i=\frac{\exp\left(\vec{x}_i^T\vec{\beta}\right)}{1+\exp\left(\vec{x}_i^T\vec{\beta}\right)}$
Binomial	Probit	$\mu_i=\Phi\left(\vec{x}_i^T\vec{\beta}\right)$
Binomial	Robit	$\mu_i=T_{(\nu-1)/\nu}\left(\vec{x}_i^T\vec{\beta}\right)$
Binomial	loglog (nloglog)	$\mu_i=\exp\left(-\exp\left(\vec{x}_i^T\vec{\beta}\right)\right)$
Binomial	Cloglog (gombit)	$\mu_i=1-\exp\left(-\exp\left(\vec{x}_i^T\vec{\beta}\right)\right)$
Binomial	Cauchit	$\mu_i=\frac{1}{\pi}\arctan(\vec{x}_i^T\vec{\beta})+0.5$
Poisson	Log	$\mu_i=\exp\left(\vec{x}_i^T\vec{\beta}\right)$
Poisson	Identity	$\mu_i=\vec{x}_i^T\vec{\beta}$
Poisson	Sqrt	$\mu_i=\left(\vec{x}_i^T\vec{\beta}\right)^2$
Gamma	Inverse	$\mu_i=\left(\vec{x}_i^T\vec{\beta}\right)^{-1}$
Gamma	Identity	$\mu_i=\vec{x}_i^T\vec{\beta}$
Gamma	Log	$\mu_i=\exp\left(\vec{x}_i^T\vec{\beta}\right)$
Inverse gaussian	Inverse squared	$\mu_i=\left(\vec{x}_i^T\vec{\beta}\right)^{-1/2}$
...	...	...

Table 14.16: **Link functions and mean functions of some Generalized Linear Models (GLM)**

✍**Examples:**

E1. Let us now see after all these developments a first practical example using the most important case: the negative binomial regression. So that everyone can practice, we are going to do this regression with the spreadsheet software of Microsoft (a version made with `R` is also available in the corresponding companion book).

We collected a dataset of days of absence from 316 young college students (below the reader can see only the 32 first). Our explanatory variables of interest are their average score in foreign languages and their gender (boy or girl: B/G):

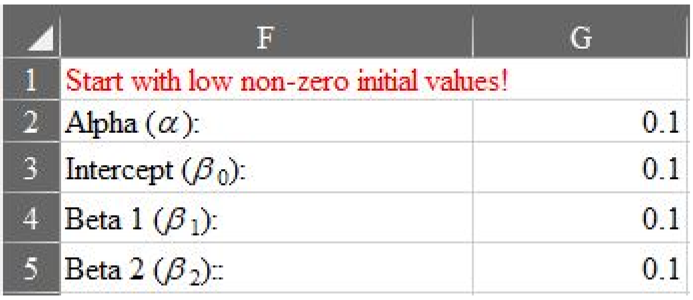


	A	B	C
1	Gender B/(	Average score in foreign language	Days of absence
2	1	42.45086	4
3	1	46.82059	4
4	0	43.56657	2
5	0	43.56657	3
6	0	27.24847	3
7	0	48.41482	13
8	0	40.73543	11
9	1	15.35938	7
10	1	52.11514	10
11	1	42.45086	9
12	0	36.45115	4
13	1	13.13055	5
14	0	62.27464	5
15	0	44.66451	3
16	1	25.25478	1
17	1	61.04388	0
18	0	49.47205	1
19	0	65.56011	0
20	1	46.82059	2
21	1	6.748048	24
22	0	42.45086	2
23	1	64.87473	0
24	1	55.33549	1
25	1	44.66451	0
26	0	38.34572	8
27	1	50	3
28	1	6.748048	7
29	0	40.15026	0
30	1	57.54914	2
31	1	71.82729	0
32	1	45.2079	0

Before going any further, it is first trivial to notice that the days of absence are non-negative with integer values and that the majority are never absent, a Gaussian regression is not the most suitable because its support sweeps away all the positive and negative real numbers (this does not necessarily mean that the result would be bad, however!). In addition, the standard deviation of days of absence is significantly different from the average of these same days of absence, therefore in extenso the Poisson regression is not the most suitable ... (however, this also does not mean necessarily that the result would be bad!).

So here is how to proceed with Microsoft® Excel® 14.0.7106 (the reader can find the equivalent example made with the `R` software in the corresponding companion book):

First we prepare an area which contains the 2 coefficients, the intercept and the parameter α of the negative binomial model (which for reminder is the inverse of the number of successes or respectively of failures and whose value must be at least equal the unit!!!) in an area of the sheet as visible below to which we associate a weak positive value written in hard on the keyboard (therefore no formulas here!):



	F	G
1	Start with low non-zero initial values!	
2	Alpha (α):	0.1
3	Intercept (β_0):	0.1
4	Beta 1 (β_1):	0.1
5	Beta 2 (β_2)::	0.1

Then, in column `D`, we will write the relation of the NB-2 log-likelihood which for reminder is given by:

$$\begin{aligned}&\ln\left[L\left(\alpha,Y_i,\vec{x},\vec{\beta}\right)\right]\\&=\sum_{i=1}^{n}\left(\begin{array}{l}\ln\left(\Gamma\left(\frac{1}{\alpha}+Y_i\right)\right)-\ln\left(\Gamma\left(\frac{1}{\alpha}\right)\right)-\ln\left(\Gamma\left(Y_i+1\right)\right)-\left(Y_i+\frac{1}{\alpha}\right)\ln\left(\alpha e^{\vec{x}_i^T\vec{\beta}}+1\right)\\+Y_i\ln\left(\alpha e^{\vec{x}^T\vec{\beta}}\right)\end{array}\right)\end{aligned}\tag{14.1.1087}$$

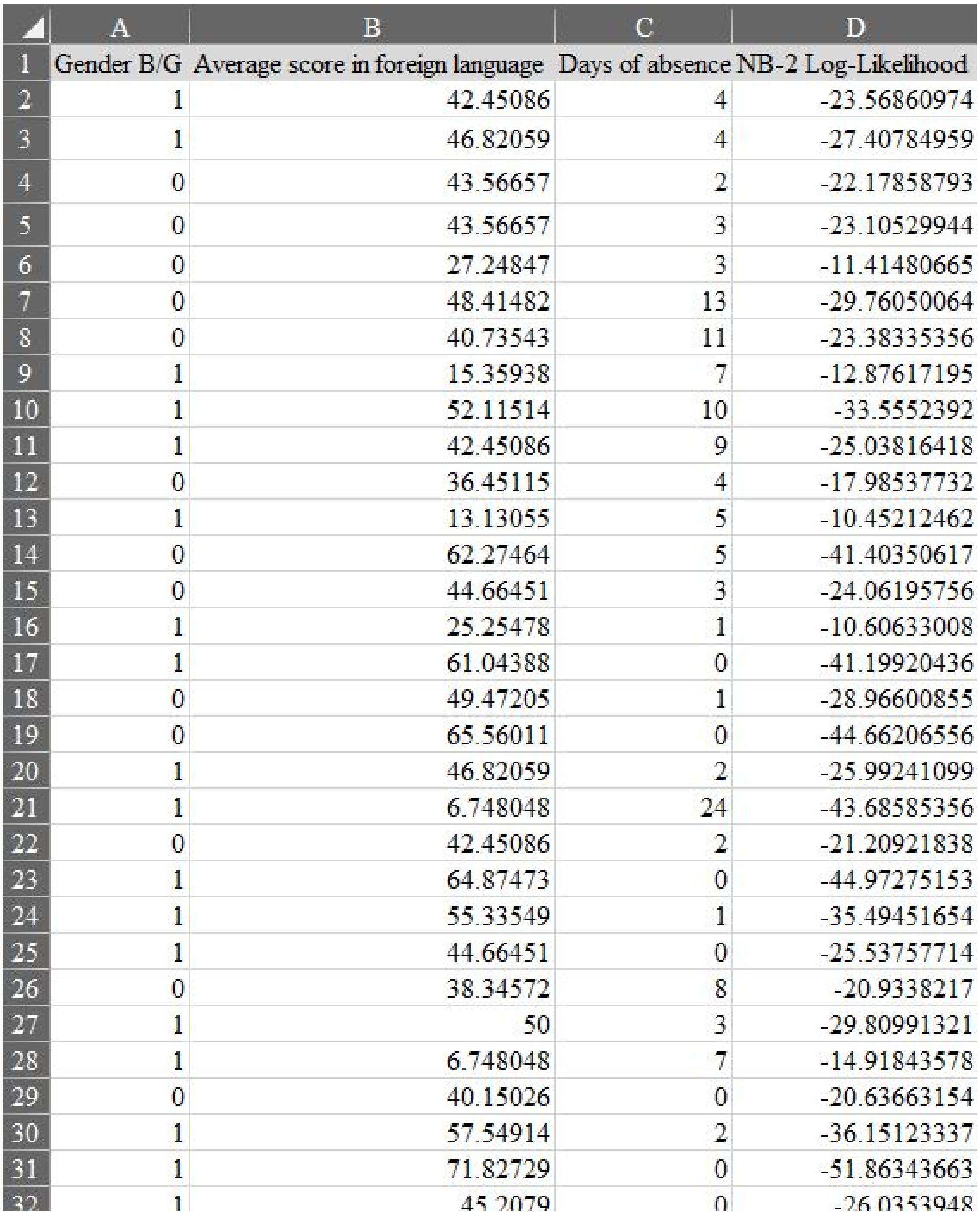
and since this spreadsheet software does not have Euler's Gamma function built in at the day we write these lines, we will rewrite it as follows:

$$\begin{aligned}&\ln\left[L\left(\alpha,Y_i,\vec{x},\vec{\beta}\right)\right]\\&=\sum\nolimits_{i=1}^{n}\left(\begin{array}{l}\ln\left(\ln\left(\left(\frac{1}{\alpha}+Y-1\right)!\right)\right)-\ln\left(\ln\left(\left(\frac{1}{\alpha}-1\right)!\right)\right)\\-\ln\left(\ln\left(\left(Y_i+1-1\right)!\right)\right)\\-\left(Y_i+\frac{1}{\alpha}\right)\ln\left(\alpha e^{-T}\vec{\beta}+1\right)\\+Y_i\ln\left(\alpha e^{-\vec{x}_i^T\vec{\beta}}\right)\end{array}\right)\end{aligned}\tag{14.1.1088}$$

Which gives with the Microsoft® Excel® formula syntax:

```
=LN(LN(FACT(1/$G$2+C2-1)))-LN(LN(FACT(1/$G$2-1)))
-IFERROR(LN(LN(FACT(C2+1-1)));1) -
(C2+1/$G$2)*LN($G$2*EXP($G$3+$G$4*A2+$G$5*B2)+1)
+C2*LN($G$2*EXP($G$3+$G$4*A2+$G$5*B2))
```

By writing this formula into column `D` and dragging it down we get first the following values:



	A	B	C	D
1	Gender B/G	Average score in foreign language	Days of absence	NB-2 Log-Likelihood
2	1	42.45086	4	-23.56860974
3	1	46.82059	4	-27.40784959
4	0	43.56657	2	-22.17858793
5	0	43.56657	3	-23.10529944
6	0	27.24847	3	-11.41480665
7	0	48.41482	13	-29.76050064
8	0	40.73543	11	-23.38335356
9	1	15.35938	7	-12.87617195
10	1	52.11514	10	-33.5552392
11	1	42.45086	9	-25.03816418
12	0	36.45115	4	-17.98537732
13	1	13.13055	5	-10.45212462
14	0	62.27464	5	-41.40350617
15	0	44.66451	3	-24.06195756
16	1	25.25478	1	-10.60633008
17	1	61.04388	0	-41.19920436
18	0	49.47205	1	-28.96600855
19	0	65.56011	0	-44.66206556
20	1	46.82059	2	-25.99241099
21	1	6.748048	24	-43.68585356
22	0	42.45086	2	-21.20921838
23	1	64.87473	0	-44.97275153
24	1	55.33549	1	-35.49451654
25	1	44.66451	0	-25.53757714
26	0	38.34572	8	-20.9338217
27	1	50	3	-29.80991321
28	1	6.748048	7	-14.91843578
29	0	40.15026	0	-20.63663154
30	1	57.54914	2	-36.15123337
31	1	71.82729	0	-51.86343663
32	1	45.2079	0	-26.0353948

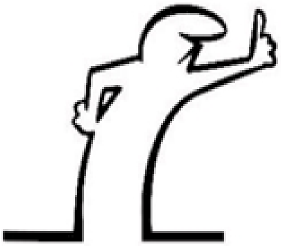
Now in a cell, we will report the sum of the NB-2 log-likelihood values. In this case we have chosen to put it in `G7`:

With everything that we have seen since the beginning of this textbook in the field of Statistics and Numerical methods, we can now provide the following summary table of some naive statistical tests that we have derived mathematically with all necessary details previously and that therefore the reader should therefore be able to grasp:

Type of dependant variable	**Type of independant variable**						
	Ordinal / Categorical				**Normal / Interval (ordinal)**	**More than 1**	**None**
	Two groups		**More groups**				
	Paired	**Unpaired**	**Paired**	**Unpaired**			
2 categrories	McNemar Test, Sign Test	Fisher Test, Chi-squared Test	Cochran's Q-Test	Fisher Test, Chi-squared Test	(Conditional) Logistic Regression	Logistic Regression	Chi-squared Test
Nominal	Bowker Test	Fisher Test, Chi-squared Test		Fishier Test, Chi-squared Test	Multinomial Logistic Regression	Multinomial Logistic Regression	Binomial Test
Ordinal	Wilcoxon Test, Sign Test	Wilcoxon-Mann-Whitney Test	Friedman Test	Kruskal-Wallis Test	Spearman Rank Test	Ordered Logit	Median Test
Interval	Wilcoxon Test, Sign Test	Wilcoxon-Mann-Whiney Test	Friedman Test	Kruskal-Wallis Test	Spearman Rank Test	Multivariate Linear Model	Median Test
Normal	T-Test (for paired)	T-Test (for unpaired)	Linear Model (ANOVA)	Linear Model (ANOVA)	Pearson Correlation Test	Multivariate Linear Model	T-Test
Censored Interval	Log Rank Test		Survival Analysis, Cox proportional hazards regression				
None	Clustering, Factor Analysis, PCA, Canonical Correlation						

Table 14.17: **Summary of naive fundamental statistical tests**

And keep in mind that excellent meme by Adrian Olszewski:



- Variable transformation

Box-Cox, Logarithm, Johnson, Manly, etc.

⊳ Change a LOT of things

- Rank-based methods

Mann-Whitney, Kruskal-Wallis, Friedman, etc.

⊳ Totally change H_0; complex interpretation, may fool you about medians

---

- More general (semi-) parametric methods

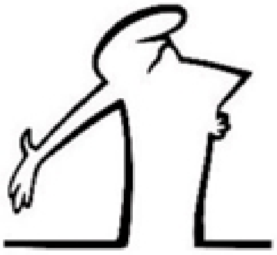
GLM, GLS, [G]LMM, GEE

⊳ Retain null hypotheses, handle heteroscedasticity, dependency, non-normality

- Computer intensive testing

Permutation tests, Bootstrap & Jackknife resampling

⊳ Retain null hypotheses, handle heteroscedasticity, dependency, non-normality, partially - outliers



14.1.12.8 Robust, M-estimators and W-estimators

The M-Estimators naturally appear in statistics when we see the problem of the aspect of the variance which is a function of squared deviations from the mean (hence a parabola) that is therefore a little too sensitive to extreme values (this is especially the case of many linear regression techniques, hence the fact that M-estimators are introduced mainly in textbooks related to regression techniques!). The idea then is to ask if it is not possible to construct something other than a square of deviations (hence a more robust statistic!) and then to define the corresponding sample mean, everything being distribution free, that we will name a "M-estimator of location[29]" (the reader should know that there are also M-estimators for the sample variance named "M-estimator of scale" like the τ-estimator!).

We have placed this subject in the section of Theoretical Computing instead of the section of Statistics because for most of these estimators, as we will see it further below, no closed form solution exists and an iterative approach to computation is required[30], the mathematics are quite ugly and most of the time there are different worldwide definitions for the same thing and also computer implementations that differ from their theoretical definition! .

Remark

Peter J. Huber (11964, 11967 according to holocene calendar) introduced M-estimators and their asymptotic properties (the proposed estimating the center of symmetry of symmetric distributions). They played an important part of the development of modern robust statistics. Liang and Zeger (11986 according to holocene calendar) helped popularize M-estimators in the biostatistics literature under the name "generalized estimating equations" (GEE). Since then many other robust estimators has been developed like MM-estimators, S-estimators, GM-estimators, τ-estimators, etc. (basically, MM-estimators are M-estimators initialized by an S-estimator...).

The best-known example of M-estimator is the arithmetic mean μ_a, as we have just mentioned, indeed it is it that trivially maximizes the function of the sum of squares of the deviations as we have seen earlier above:

$$\frac{\partial}{\partial \mu_a}\sum_{i=1}^{n}(x_i-\mu_a)^2=0 \tag{14.1.1089}$$

But some statistician have had the idea (in their purpose to always generalize everything thank to a lot of coffee) to define an M-estimator as the expression of the average that maximizes any function f whose argument is by definition $x_i-\theta$ such that:

$$\frac{\partial}{\partial \theta}\sum_{i=1}^{n}f(x_i-\theta)=0 \tag{14.1.1090}$$

And for example, in the case of the most familiar variance, this function f is the square of the argument, i.e. $f(x_i-\theta)=(x_i-\theta)^2$. It is customary, however, in statistics to denote the function f with the letter ρ (so do not confuse with the correlation coefficient or anything of the same kind!). We will continue our explanations by aligning ourselves with this traditional notation!

---

[29] Then by extension M-estimators are a broad class of extremum estimators for which the objective function is a sample average!

[30] It is possible to use standard function optimization algorithms, such as Newton–Raphson (solver!)

What ultimately makes us look for a θ that for a particular function ρ' gives:

$$\frac{\partial}{\partial\theta}\sum_{i=1}^{n}\rho(x_i-\theta)=\sum_{i=1}^{n}\rho'(x_i-\theta)=\sum_{i=1}^{n}\psi(x_i-\theta)=0 \tag{14.1.1091}$$

where ψ is named the "influence function" (both ρ and ψ should be obviously symmetric!). When the partial derivative exist, we speak of "ψ-type M-estimator" otherwise of "ρ-type M-estimator".

Remark

If $\rho=f$, the probability density function related to the random variable X, then the reader may have noticed that we fall back on the MLE (Maximum Likelihood Estimator) as introduced in the section of Statistics page 818!

We would like any good M-estimator of location to be location and scale equivariant, that is, we would like it to respond in a reasonable manner to linear changes across the sample. An M-estimator is "location equivariant" if, when every observation is shifted by some amount a, the θ of that shifted sample also shifts by a. An M-estimator of location is scale equivariant if, when every observation is multiplied by some non-zero constant b, the θ of that altered sample is b times the θ of the unaltered sample. An M-estimator of location that possesses both these characteristics is location-and-scale equivariant. In other words, we want the following to hold:

$$\theta(ax_1+b,ax_2+b,\ldots,ax_n+b)=b\theta(x_1,x_2,\ldots,x_n)+a \tag{14.1.1092}$$

In order for an M-estimator of location to be location-and-scale-equivariant, it is often necessary to scale the observations when computing ρ and ψ. This is often done whether rescaling is necessary or not because it makes notation simpler.

The matter of location-equivariance is fairly simple! It can be satisfied by making the input of the form $x_i-\theta$. Adjusting for scale is not as straightforward. We must pick some estimator of the scale of the sample, noted (without surprise...) s_i, which is a function of the observations $x_1,x_2,\ldots,x_n$. s is scale-equivariant and location invariant, meaning that it is unaffected by a shift in the sample as described above.

It is then common to define the standardized residual:

$$z_i=\frac{x_i-\theta}{c\cdot s_i}=\frac{r_i}{c} \tag{14.1.1093}$$

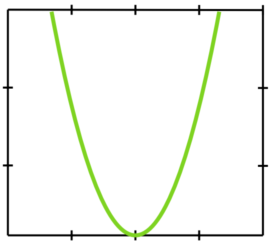
where s_i is a strictly positive robust estimator of scale (typically the median absolute deviation but it's not the only choice!) and c is a tuning constant, then we can write obviously:

$$\frac{\partial}{\partial\theta}\sum_{i=1}^{n}\rho(z_i)=\sum_{i=1}^{n}\rho'(z_i)=\sum_{i=1}^{n}\psi(z_i)\frac{\partial z_i}{\partial\theta}=0 \tag{14.1.1094}$$

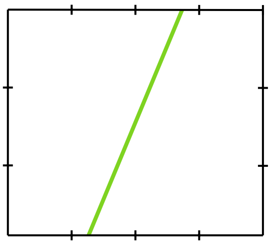
Definition (#14.1.561): Thus, more generally, a "M-estimator of location" may be defined to the value of a sample mean $\hat{\theta}$ that leads to a zero value of an estimating function given by:

$$\boxed{\hat{\theta}=\min_{\theta}\left(\sum_{i=1}^{n}\rho(z_i,\theta)\right)} \tag{14.1.1095}$$

Ideally, it is desirable that the function ρ' had as main property of being bounded (which immediately eliminates the arithmetic mean as an ideal M-estimator) which avoids typical deviations from the standard deviation (where $\rho(r_i)=r_i^2/2$ and hence $\psi(r_i)=r_i$):



ρ-function



ψ (influence) function

Figure 14.48: L^2-norm (influence function)

But we also require at least the four following properties for the $\rho(r_i)$ function:

P1. $\rho(r_i)\geq 0$ for all r_i and has a minimum at 0

P2. $\rho(r_i)=\rho(-r_i)$ for all r_i

P3. $\rho(r_i)$ increases as r_i increases from 0, but doesn't get too large as r increases

P4. $\rho(r_i)$ should be differentiable

and the following four properties for the $\psi(r_i)$ function:

P1. Be piece-wise continuous function

P2. Be an odd function, i.e. $\psi(-r_i)=-\psi(r_i)$

P3. $\psi(r_i)\geq 0$ for $r_i\geq 0$ and $\psi(r_i)>0$ for $0<r_i<k$

P4. Its slope is 1 at 0, i.e. $\psi'(0)=1$

Note that the last property is not strictly required mathematically, but we use it for standardization in those case where ψ is continuous at 0. Then if follow from the first property that $\psi(0)=0$, and we require $\psi(0)=0$ also for the case where ψ is discontinuous in 0, as it is, e.g., for the M-estimator defining the median.

Albert E. Beaton and John W. Tukey in their article [45] also define a new weighted function ω (without justifying it, but it's probably a mathematical trick to simplify some calculations that we will see further below) such that:

$$\psi(r_i)=r_i w(r_i)\Leftrightarrow w(r_i)=\frac{\psi(r_i)}{r_i} \tag{14.1.1096}$$

Thus:

$$\sum_{i=1}^{n}\psi(r_i)=\sum_{i=1}^{n}\frac{\psi(r_i)}{r_i}r_i=\sum_{i=1}^{n}w_i(r_i)r_i=0 \tag{14.1.1097}$$

What E. Beaton and John. W. Tukey write explicitly in their original article:

$$\sum_{i=1}^{n}w_i(z_i)z_i=\sum_{i=1}^{n}w_i\left(\frac{x_i-\theta}{c\cdot s_i}\right)\frac{x_i-\theta}{c\cdot s_i}=0 \tag{14.1.1098}$$

We can eliminate the $c\cdot s_i$ of the second fraction if we assume, same like Beaton and Tukey did, that it is equal for all i such that (relation that we use later):

$$\boxed{\sum_{i=1}^{n}w_i\left(\frac{x_i-\theta}{c\cdot S}\right)(x_i-\theta)=0} \tag{14.1.1099}$$

Definition (#14.1.562): The "breakpoint of an estimator" is the maximum proportion of observations (hence a number between 0 and 1) that can be changed without changing the estimator. For example the breakpoint of the arithmetic average is 0, that of the median is (obviously) 0.5.

Remark

Andrews et al. (11972 according to holocene calendar) (the Princeton Robustness Study), at which time it was expected that all statistical analyses would, by default, be robust wrote: «*any author of an applied article who did not use the robust alternative would be asked by the referee for an explanation*» and also that «*From the 1970s to 2000 we would see... extensions to linear models, time series, and multivariate models, and widespread adoption where every statistical package would take the robust method as the default...*»

Sadly, we see that even in 12018 (the day we write these lines according to holocene calendar), it's always not the case even if all softwares have largely the possibility to do so using nonparametric or brute force statistical methods (human inertia to change issue...).

Before introducing a list of some well-known ρ, ψ and ω functions, let us introduce two one well known robust estimator and one well known M-estimators of location (these both being introduced in the form of the objective function introduced just earlier above!).

1. The "trimmed mean", also named "truncated mean" or "pruned mean", is the robust estimator consisting in the arithmetic mean obtained by removing from the observed values those below and above a given quantile. It is customary to denote $\alpha/2$ the corresponding

percentile and that if we have n values then we return the $[\alpha n]$ largest values and the $n-[\alpha n]$ smaller (remember that $[\ldots]$ is the notation for "integer value" as seen on page 6293).

Mathematically, we denote elegantly the pruned average of ordered realizations x_i in descending order to increasing of a random variable X in the following form:

$$\boxed{\mu_{T_\alpha}=\frac{1}{n-2[n\alpha]}\sum_{i=[n\alpha]+1}^{n-[n\alpha]} x_i} \tag{14.1.1100}$$

Remark

We can find trimmed mean in many spreadsheets softwares for example under the function name `TRIMMEAN( )`, or in statistical software like `R` (just by passing an argument to the default `MEAN( )` function!

It is quite obvious that this robust estimator is an M-Estimator as it respects the definition seen above since we are only replacing some values of the original series with other values and that's finally equivalent to an arithmetic mean!

2. The "Winzorised (Huber) mean" is the robust estimator consisting in the arithmetic mean calculated after that the $[\alpha n]$ smallest values have been replaces by the value $x_{[\alpha n]+1}$ and the $[\alpha n]$ biggest values by the value $x_{n-[\alpha n]}$.

   Mathematically we denote elegantly the Winsorized mean of the ordered realizations x_i in descending order to increasing of a random variable X in the following way (discrete version):

$$\boxed{\mu_W=\frac{1}{n}\left[([n\alpha]+1)x_{[\alpha n]+1}+\sum_{i=[n\alpha]+2}^{n-[n\alpha]-1} x_i+([n\alpha]+1)x_{n-[\alpha n]-1}\right]} \tag{14.1.1101}$$

   Again it is obvious that this robust estimator is an M-Estimator as it respects the definition seen above since we are only replacing some values of the original series with other values and that's finally equivalent to an arithmetic mean!

Remark

We can find the Huber's mean in software like `R` using the function `HUBER( )` of the MASS package

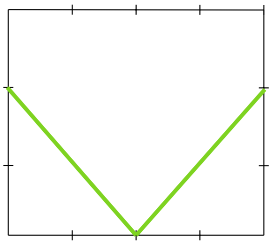
3. The "median", that we already well know, that satisfies (discrete version under the form of an M-estimator):

$$\boxed{\sum_{i=1}^{n}\operatorname{sgn}(x_i-M_e)=0} \tag{14.1.1102}$$

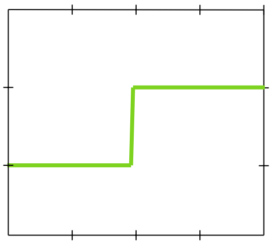
   where:

$$\operatorname{sgn}(x)=\begin{cases}-1 & x<0\\ 0 & x=0\\ +1 & x>0\end{cases} \tag{14.1.1103}$$

is for recall named the "signum function" or simply "sign function".



ρ-function



ψ (influence) function

Figure 14.49: **Median (influence function)**

✍**Example:**

Consider the values $2, 4, 5, 10, 200$. We then have first:

$$\mu_a = 44.2 \qquad M_e = 5$$

We then have for the 20% trimmed mean:

$$\mu_{T_\alpha} = \frac{4+5+10}{3} = 6.\bar{3}$$

and for the 20% Winsorized mean:

$$\mu_W = \frac{4+4+5+10+10}{5} = 6.6$$

Remark

The mean μ_a, the trimmed mean μ_{T_α}, the Winsorized mean μ_W and the median M_e are particular cases of "*L*-statistic" (linear combination of order statistics):

$$M_l\left(y_1, \ldots, y_n\right) = \sum_{i=1}^{n} a_i Y_{(i)} \quad a_{i,n} = \frac{1}{n} \Rightarrow \bar{y}$$

For example:

$$a_{i,n} = \frac{1}{n} \Rightarrow \bar{y}$$

lead to the mean μ_a:

$$a_{i,n}=\begin{cases}\text{for } n \text{ odd} & \\ 1 & i=\dfrac{n+1}{2}\\ 0 & i\neq\dfrac{n+1}{2}\end{cases}\qquad a_{i,n}=\begin{cases}\text{for } n \text{ even} & \\ \dfrac{1}{2} & i=\dfrac{n}{2},\dfrac{n}{2}+1\\ 0 & \text{otherwise}\end{cases}$$

leads to the median M_e.

Let's look at some of ρ, ψ and ω functions that build some common M-estimators:

Type	$\rho(r_i)$	$\psi(r_i)$	$\omega(r_i)$
L_2 (least-squares)	$r_i^2/2$	r_i	1
L_1 (least-absolute)	$\lvert r_i\rvert$	$\text{sgn}(r_i)$	$\dfrac{1}{\lvert r_i\rvert}$
L_1-L_2	$2\left(\sqrt{1+r_i^2/2}-1\right)$	$\dfrac{r_i}{\sqrt{1+r_i^2/2}}$	$\dfrac{1}{\sqrt{1+r_i^2/2}}$
L_p	$\dfrac{\lvert r_i\rvert^{\nu}}{\nu}$	$\text{sgn}(r_i)\lvert r_i\rvert^{\nu-1}$	$\lvert r_i\rvert^{\nu-2}$
"Fair"	$c^2\left[\dfrac{\lvert r_i\rvert}{c}-\log\left(1+\dfrac{\lvert r_i\rvert}{c}\right)\right]$	$\dfrac{r_i}{1+\lvert r_i\rvert/c}$	$\dfrac{1}{1+\lvert r_i\rvert/c}$
Huber $\begin{cases}\text{if } \lvert r_i\rvert\leq k\\ \text{if } \lvert r_i\rvert\geq k\end{cases}$	$\begin{cases}r_i^2/2\\ k(\lvert r_i\rvert-k/2)\end{cases}$	$\begin{cases}r_i\\ k\text{sgn}(r_i)\end{cases}$	$\begin{cases}1\\ k/\lvert r_i\rvert\end{cases}$
Cauchy	$\dfrac{c^2}{2}\log\left(1+(r_i/c)^2\right)$	$\dfrac{r_i}{(1+r_i^2)^2}$	$\dfrac{1}{(1+r_i^2)^2}$
Geman-MacClure	$\dfrac{r_i^2/2}{1+r_i^2}$	$\dfrac{r_i}{(1+r_i^2)^2}$	$\dfrac{1}{(1+r_i^2)^2}$
Welsch	$\dfrac{c^2}{2}\left[1-\exp\left(-(r_i/c)^2\right)\right]$	$r_i\exp\left(-(r_i/c)^2\right)$	$\exp\left(-(r_i/c)^2\right))$
Tukey $\begin{cases}\text{if } \lvert r_i\rvert\leq c\\ \text{if } \lvert r_i\rvert> c\end{cases}$	$\begin{cases}\dfrac{c^2}{6}\left(1-\left[1-(r_i/c)^2\right]^3\right)\\ (c^2/6)\end{cases}$	$\begin{cases}r_i\left[1-(r_i/c)^2\right]^2\\ 0\end{cases}$	$\begin{cases}\left[1-(r_i/c)^2\right]^2\\ 0\end{cases}$

Table 14.18: **Table of a few common dispersion M-estimators**

As far as I know (as redactor of this topic on M-estimators), they are all empirical and there is no "beautiful" way to derive them from a logical process.

Here is a graphical representation of the different functions of the above table:

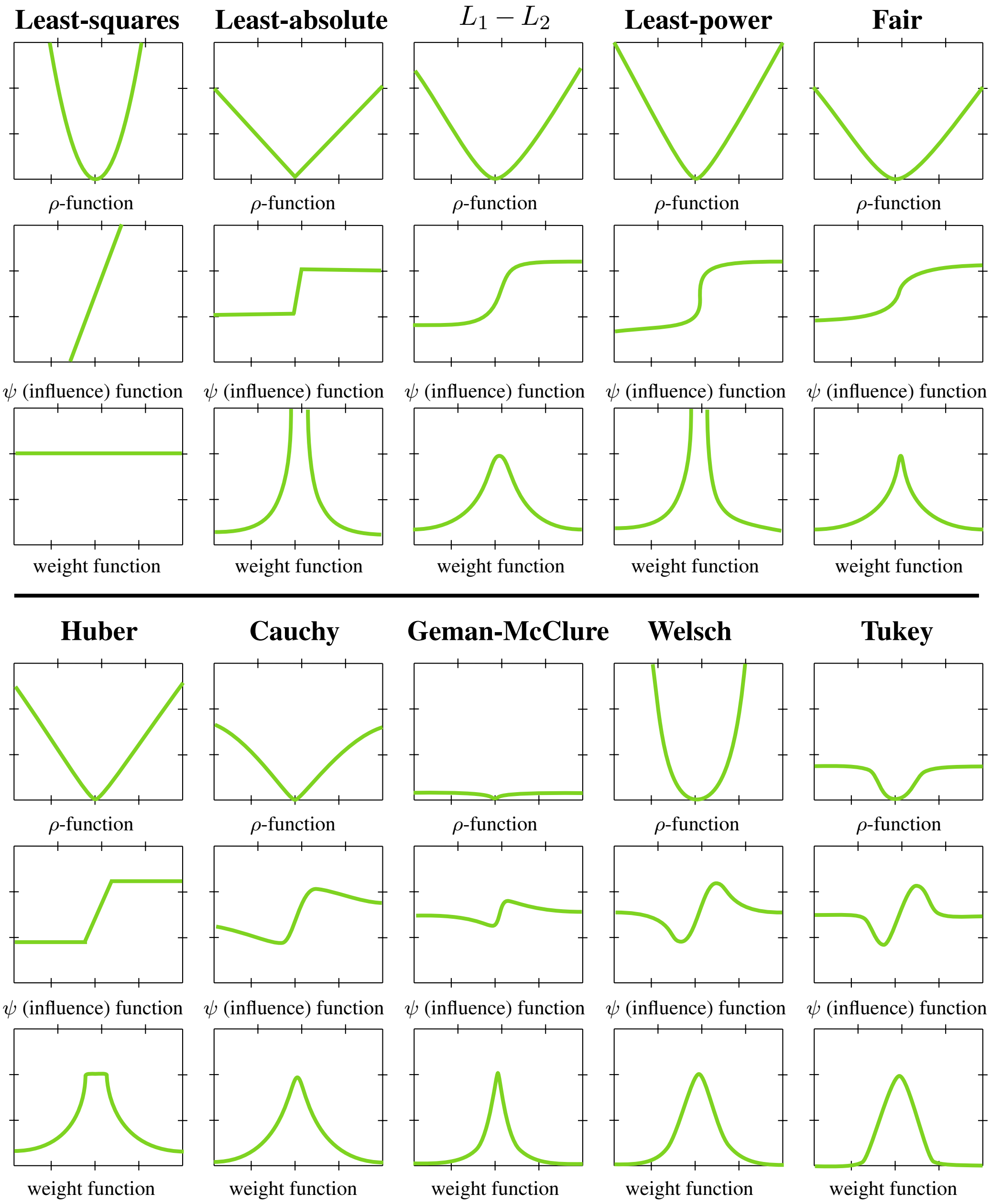


Figure 14.50: Graphical representation of a few common dispersion M-estimators

Briefly we give a few description of these functions:

- The "L_2 (least-squares) estimators", associated to the classic well known Pearson variance, are not robust because their influence function is not bounded.
- The "L_1 (least-absolute) estimators", associated to the mean absolute deviation, are not stable because the ρ-function $|r_i|$ is not strictly convex in r_i. Indeed, the second derivative at $r_i = 0$ is unbounded, and an indeterminate solution may result. This estimator in comparison of the L_2 reduce the influence of large errors, however they still have an influence because the influence function has no cutoff point.
- The "$L_1 - L_2$ estimators" take both the advantage of the L_1 estimators to reduce the influence of large errors and that of L_2 estimators to be convex.

- The "L_p (least-power) estimators" represents a family of functions. It is L_2 with $p=2$ and L_1 with $p=1$. It appears that p must be fairly moderate to provide a relatively robust estimator or, in other words, to provide an estimator scarcely perturbed by outlying data. The selection of an optimal p seems to have been investigated, and for p around 1.2, a good estimate may be expected. However, many difficulties are encountered in the computation when parameter p is in the rage of $1<p<2$, because zero residuals are troublesome.

- The "fair estimators" have everywhere defined continuous derivatives of first three orders, and yields a unique solution. The 95% asymptotic efficiency on the standard Normal distribution is obtained with the tuning constant $c=1.3998$.

- The "Huber's estimators" are parabola in the vicinity of zero, and increases linearly at a given level $|r_i|>k$. The idea is to penalize small residuals quadratically, and large residuals linearly. The 95% asymptotic efficiency on the standard Normal distribution is obtained with the tuning constant $k=1.345$. These estimators are so satisfactory that it has been recommended for almost all situations; very rarely they have been found to be inferior to some other ρ-function. However, from time to time, difficulties are encountered, which may be due to the lack of stability in the gradient values of the ρ-function of discontinuous second derivative:

$$\frac{\mathrm{d}^2\rho(r_i)}{\mathrm{d}r_i^2}=\begin{cases}1 & \text{if } |r_i|\leq k\\ 0 & \text{if } |r_i|\geq k\end{cases} \qquad (14.1.1104)$$

- The "Cauchy's estimators", also known as the "Lorentzian estimators", do not guarantee a unique solution. With a descending first derivative, such a function has a tendency to yield erroneous solutions in a way which cannot be observed. The 95% asymptotic efficiency on the standard Normal distribution is obtained with the tuning constant $c=2.3849$.

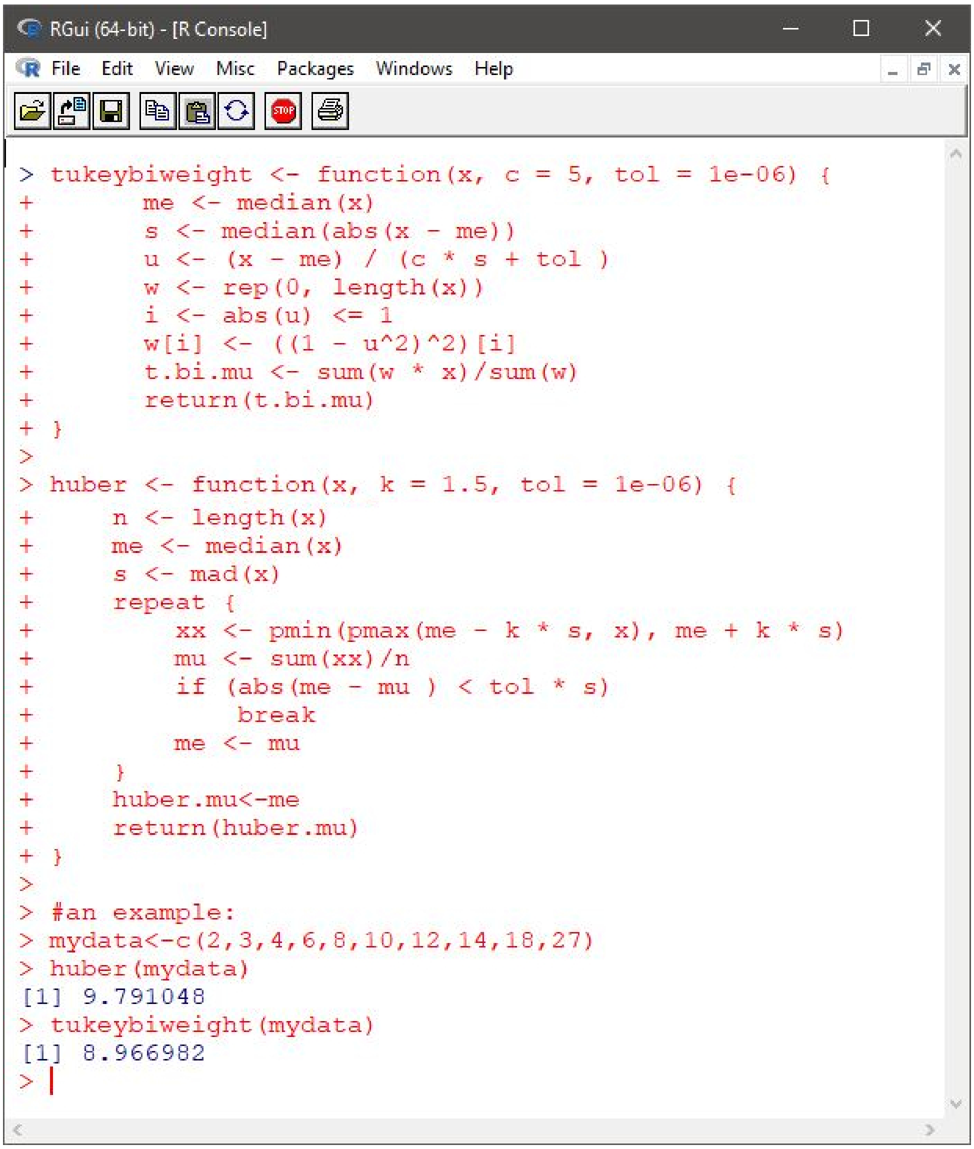
- The other remaining functions have the same problem as the Cauchy function. As can be seen from the influence function, the influence of large errors only decreases linearly with their size. The "Geman-McClure estimators and "Welsch estimators" try to further reduce the effect of large errors, and the "Tukey's biweight estimators" even suppress the outliers. The 95% asymptotic efficiency on the standard Normal distribution of the Tukey's biweight function is obtained with the tuning constant $c=4.6851$; that of the Welsch function, with $c=2.9846$.

There still exist many other ρ-functions, such as Andrew's cosine wave function, the GGW (Generalized Gauss-Weight), the LQQ (Linear Quadratic Quadratic), the "Optimal", etc.

Definition (#14.1.563): "Redescending M-estimators" are ψ-type M-estimators which have ψ functions that are non-decreasing near the origin, but decreasing toward 0 far from the origin. Their ψ functions can be chosen to redescend smoothly to zero, so that they usually satisfy $\psi(r_i)=0$ for all r_i with $|r_i|>k$, where k is referred to as the "minimum rejection point". Due to these properties of the ψ function, these kinds of estimators are very efficient, have a high breakdown point and, unlike other outlier rejection techniques, they do not suffer from a masking effect. They are efficient because they completely reject gross outliers, and do not completely ignore moderately large outliers (like median).

Regarding to the previous figure, the Cauchy, Geman-McClure, Welsch, Tukey but also Andrew M-estimators are Redescending M-estimators!

Let us see now practically how explicitly, as it is not really obvious..., are computationally implemented two of these estimator typically with a language like R (notice the iteration in the Huber estimator, it is the iterative reweighted least squares (IRLS)!):


RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help

```
> tukeybiweight <- function(x, c = 5, tol = 1e-06) {
+       me <- median(x)
+       s <- median(abs(x - me))
+       u <- (x - me) / (c * s + tol )
+       w <- rep(0, length(x))
+       i <- abs(u) <= 1
+       w[i] <- ((1 - u^2)^2)[i]
+       t.bi.mu <- sum(w * x)/sum(w)
+       return(t.bi.mu)
+ }
>
> huber <- function(x, k = 1.5, tol = 1e-06) {
+     n <- length(x)
+     me <- median(x)
+     s <- mad(x)
+     repeat {
+         xx <- pmin(pmax(me - k * s, x), me + k * s)
+         mu <- sum(xx)/n
+         if (abs(me - mu ) < tol * s)
+             break
+         me <- mu
+     }
+     huber.mu<-me
+     return(huber.mu)
+ }
>
> #an example:
> mydata<-c(2,3,4,6,8,10,12,14,18,27)
> huber(mydata)
[1] 9.791048
> tukeybiweight(mydata)
[1] 8.966982
> 
```

Figure 14.51: **Implementation of Tukey's biweight and Huber M-estimator in R (source: R affy package)**

As the reader may have noticed by reading these script, we are very far from what we have theoretically introduced earlier before! So to understand why, let us focus for example on Tukey's biweight, and the reader will see that it will lead us to the concept of "W-estimators"!

For this, the reader let us recall that Beaton and Tukey had introduced in their article:

$$\sum_{i=1}^{n} w_i\left(\frac{x_i-\theta}{c\cdot S}\right)(x_i-\theta)=0 \quad (14.1.1105)$$

We can extract θ quite easily :

$$\theta = \frac{\sum_{i=1}^{n} w_i\left(\frac{x_i-\theta}{c\cdot S}\right)x_i}{\sum_{i=1}^{n} w_i\left(\frac{x_i-\theta}{c\cdot S}\right)} \qquad (14.1.1106)$$

In the case of Tukey's biweight we have then (this corresponds to the `t.bi.mu` in the script above):

$$\theta = \frac{\sum_{i=1}^{n}\left(1-\left(\frac{x_i-\theta}{c\cdot S}\right)^2\right)^2 x_i}{\sum_{i=1}^{n} w_i\left(\frac{x_i-\theta}{c\cdot S}\right)} \qquad (14.1.1107)$$

As there is a circular reference between the θ on the left and the θ on the right, the idea of some Tukey's biweight computer implementation is to take an estimate of the θ on the right (same idea to estimate S). Then we can see for example that Tukey's biweight M-estimator of location is computed in the affy package of the `R` software as:

$$\hat{\mu}_{T.b.} \cong \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \frac{\sum_{i=1}^{n}\left(1-\left(\frac{x_i-M_e}{c\cdot \text{MAD}}\right)^2\right)^2 x_i}{\sum_{i=1}^{n}\left(1-\left(\frac{x_i-M_e}{c\cdot \text{MAD}}\right)^2\right)^2} \qquad (14.1.1108)$$

where we put $w_i = 0$ if:

$$\left|\frac{x_i-M_e}{c\cdot \text{MAD}}\right| > c \qquad (14.1.1109)$$

Such an approach (one-shot iteration using a weight function trick) to calculate an M-estimator should, according to Beaton and Tukey article, be named a "W-estimator". But we can iterate the process by writing:

$$\theta^{[k+1]} = \frac{\sum_{i=1}^{n}\left(1-\left(\frac{x_i-\theta^{[k]}}{c\cdot S}\right)^2\right)^2 x_i}{\sum_{i=1}^{n} w_i\left(\frac{x_i-\theta^{[k]}}{c\cdot S}\right)} \qquad (14.1.1110)$$

where the weights for a particular $\theta^{[k]}$ depend upon the previous values of $\theta^{[k-1]}$. We simply start with a reasonable guess of the values of $\theta^{[0]}$ (as we did just above) and also especially of the first initial estimate $\theta^{[0]}$, and iterate until the estimates have converged to within whatever we consider to be a reasonable margin of accuracy, that is, until $(\theta^{[k+1]}-\theta^{[k]})^2$ is less than some predicted values. This procedure is an example of the process of iteratively reweighted least squares (IRLS) that Beaton and Tukey in their article recommend to name W^k-estimators depending on the number $[k]$ of iterations...

Remark

IRLS can have some negative effects. Indeed, when IRLS is employed, more and more points can be eliminated from the data set with each iteration. At each step, the most extreme outliers are dropped from the sample because they are outside of a certain main body of the data. Left unchecked, IRLS can sometimes give zero weight to every point and whittle the data set away to nothing, so IRLS estimators need constraints applied to them that will counteract the effect of the redescending weight function. Such constrained estimators are named "S-estimators".

So far, the ready should have a better understanding why M-estimators are more a subject related to "statistical engineering" rather than "pure Statistics": different implementations and also even sometimes different definitions, quite arbitrary parameters, etc...!!!!

Remark

It can be proved that like the trimmed mean, M-estimators, τ-estimator and MM estimator are unbiased and asymptotically Normal at symmetric distributions (useful property to calculate their confidence intervals!).

14.1.13 Interpolation Techniques

The main difference between interpolation and regression, is the definition of the problem they solve.

Given n data points, when you interpolate, you look for a function that is of some predefined form that has the values in that points exactly as specified. That means given pairs (x_i, y_i) you look for f of some predefined form that satisfies $f(x_i) = y_i$. Most commonly f is chosen to be polynomial, spline (low degree polynomials on intervals between given points).

When you do regression, you look for a function that minimizes some cost, usually sum of square of errors, You don't require the function to have the exact values at given point, you just want a good approximation. In general, your found function f might not satisfy $f(x_i) = y_i$ for any data point, but the cost function, i.e. $\sum_{i=1}^{n}(f(x_i) - y_i)^2$ will be the smallest possible of all the functions of given form.

There are numerous polynomials interpolation techniques more or less complex and sophisticated. We propose here to present some in ascending order of difficulty.

14.1.13.1 Bézier Curves (Bézier Splines)

The Russian engineer Pierre Bézier, at the beginnings of the Computer Aided Design (CAD), in the years 11960 (holocene calendar), gave a way to define curves and surfaces from points. This allows direct manipulation, geometric, of curves without having to give a series of equation to the machine to describe a complex curve!!

The application field of Bezier curves is a multifaceted to, very rich, at the crossroads of many diverse mathematical areas: Analysis, Kinematic, Differential Geometry, Affine Geometry, Projective Geometry, Fractal Geometry, Probabilities, Statistics, Finance (rates curve), etc.

The Bezier are also become essential in their concrete applications in industry, and computer graphics. Most non-engineer and non-scientific people know them through the usage of some drawing tools included in Adobe Illustrator, Adobe Photoshop, 3D Studio Max, Blender, Rhino, Adobe InDesign, Microsoft Office Visio and even Microsoft Office Word and Microsoft Office PowerPoint. Most computer fonts are made with Bézier curves too!

Let us present and study in detail the mathematical approach of this technique:

First, we know that the equation of a segment line which we will denote in this field of study M (with respect for the tradition) joining two points $A(x_1, y_1), B(x_2, y_2)$ is:

$$M_1(t) = t(B - A) + A = (1 - t)A + B \tag{14.1.1111}$$

and named a "linear interpolation" (LERP).

This is right since when $t = 0$ we are on A and when $t = 1$ we are on B. Therefore $t \in [0, 1]$ and the point M describes the whole straight segment $[AB]$.

By definition, the straight segment denoted $[AB]$ is a "Bézier curve of degree 1" with control points A and B and the polynomials on t or on $t-1$ are the "Bernstein polynomials of degree 1".

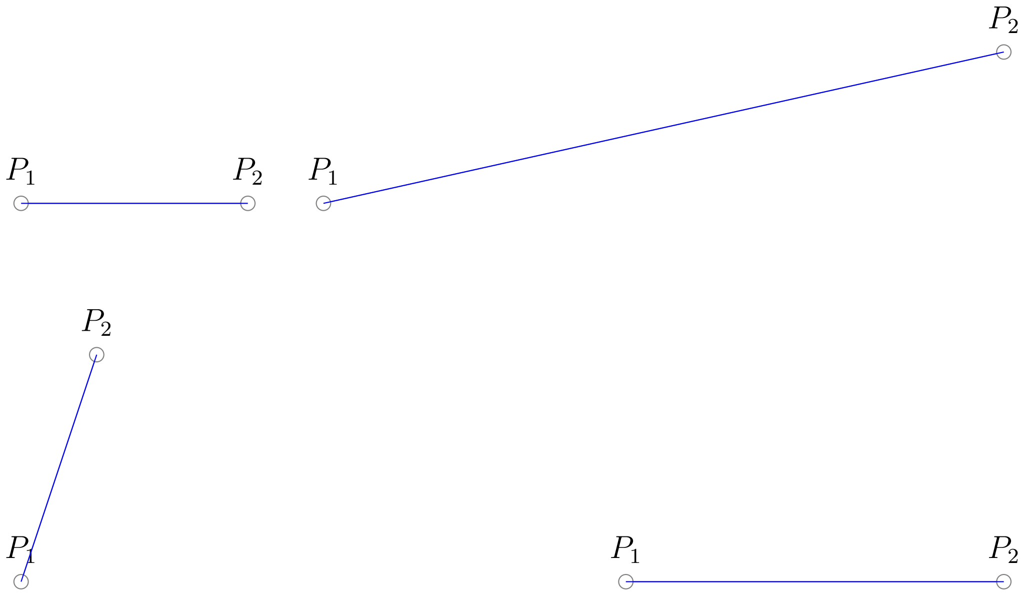


Figure 14.52: A few 1st order Bézier curves

Let us now build a parametric curve by adding a second step to what we just seen:

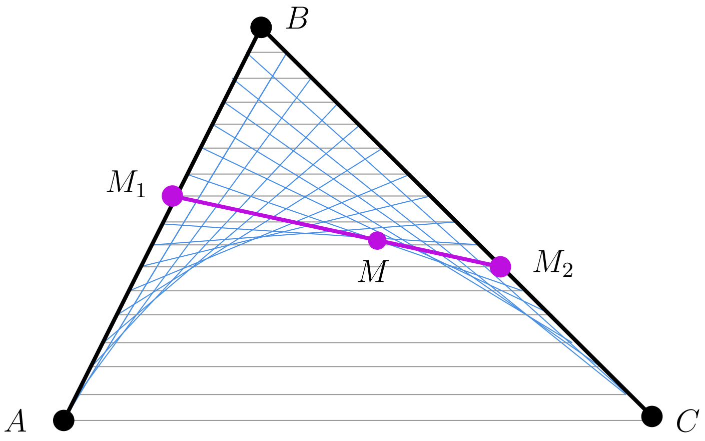


Figure 14.53: Idea behind 2nd order Bézier curves

1. 1st step:
    (a) Given $M_1(t)$ the barycenter of $(A, 1-t)$ and (B, t) and where $M_1(t)$ describes the segment $[AB]$.
    (b) Given $M_2(t)$ the barycenter of $(B, 1-t)$ and (C, t) and where $M_2(t)$ describes the segment $[BC]$.
2. 2nd step:

    Given $M(t)$ the barycenter of $(M_1(t), 1-t), (M_2(t), t)$.

> By construction $M(t)$ is therefore at the same proportion of the segment $[M_1(t), M_2(t)]$ relative to the segment $[AB]$ or to $M_2(t)$ compared to the segment $[BC]$.
>
> The curve obtained is then the envelope of the segments $[M_1(t), M_2(t)]$: at any point $M(t)$, the tangent to the curve is therefore the segment $[M_1(t), M_2(t)]$.

$M(t)$ then describes a Bézier curve of order 2, which, by construction starts at A and ends at C, and has for tangents $[AB]$ on A and $[CB]$ on C.

This is in fact a parabolic arc (that we denote logically by $[ABC]$ in this field of study):

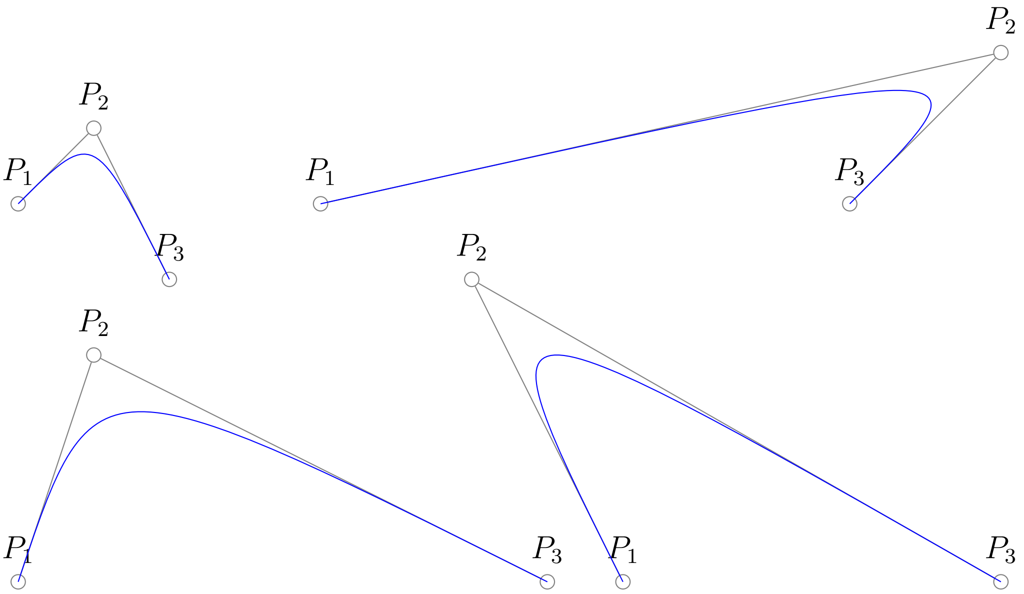


Figure 14.54: **A few 2nd order Bézier curves**

This is the magic of Bézier curves! We get smooth curves only by using combinations of linear interpolations (LERP)!! The example above is for example a quadratic Bézier curve, or a Bézier curve of order 2.

By the same pattern, we can define a Bézier curve of n points P_i with $i = 1...n$. This is what we name the "Casteljau algorithm" (the underlying idea is to recursively LERP until there is nothing more to LERP). Thus, given:

$$\forall r \in [1, n-1] \tag{14.1.1112}$$

We get:

$$b_i^{(r)}(t) = (1-t)b_i^{(r-1)}(t) + tb_{i+1}^{(r-1)}(t) \tag{14.1.1113}$$

The recurrence ending for:

$$b_i^{(0)}(t) = P_i \tag{14.1.1114}$$

Thus, for $n = 2$ we have:

$$M_2(t) = b_i^{(1)}(t) = (1-t)b_i^{(0)}(t) + tb_{i+1}^{(0)} \tag{14.1.1115}$$

Therefore:

$$\boxed{M_2(t)=b_1^{(1)}(t)=(1-t)b_1^{(0)}(t)+tb_2^{(0)}=(1-t)P_1+tP_2} \tag{14.1.1116}$$

Thus, we have necessarily with two points the equation of a straight line (Bézier curve of order 1).

Now consider $M_3(t)$ the Bézier curve of order 2, therefore we have the points always defined by:

$$M_3(t)=b_i^{(r)}(t)=(1-t)b_i^{(r-1)}(t)+tb_{i+1}^{(r-1)}(t) \tag{14.1.1117}$$

We then have by the recurrence relation:

$$\begin{aligned}
b_1^{(0)}&=P_1\\
b_2^{(0)}&=P_2\\
b_3^{(0)}&=P_3\\
b_1^{(1)}&=(1-t)b_1^{(0)}(t)+tb_2^{(0)}(t)\\
b_2^{(1)}&=(1-t)b_2^{(0)}(t)+tb_3^{(0)}(t)\\
&\cancel{b_3^{(1)}=(1-t)b_3^{(0)}(t)+tb_4^{(0)}(t)}\\
b_1^{(2)}&=(1-t)b_1^{(1)}(t)+tb_2^{(1)}(t)\\
b_2^{(2)}&=(1-t)b_2^{(1)}(t)+tb_3^{(1)}(t)\\
&\cancel{b_3^{(2)}=(1-t)b_3^{(1)}(t)+tb_4^{(1)}(t)}
\end{aligned} \tag{14.1.1118}$$

where we have eliminated the terms containing unspecified points.

We have therefore:

$$\begin{aligned}
b_1^{(1)}&=(1-t)P_1+tP_2\\
b_2^{(1)}&=(1-t)P_2+tP_3\\
b_1^{(2)}&=(1-t)b_1^{(1)}(t)+tb_2^{(1)}(t)\\
b_2^{(2)}&=(1-t)b_2^{(1)}(t)+tb_3^{(1)}(t)
\end{aligned} \tag{14.1.1119}$$

Therefore it comes for the 2nd order Bézier curves:

$$\boxed{M_3(t)=(1-t)^2P_1+2t(1-t)P_2+t^2P_3} \tag{14.1.1120}$$

or in vector form (more consistent with the usual mathematical notation and therefore we also better understand the meaning of "vectorial drawing"):

$$\boxed{\overrightarrow{\text{O}P}=\vec{M}_3(t)=\begin{pmatrix}x\\y\end{pmatrix}=(1-t)^2\overrightarrow{\text{O}P_1}+2t(1-t)\overrightarrow{\text{O}P_2}+t^2\overrightarrow{\text{O}P_3}} \tag{14.1.1121}$$

and in matrix form (the matrix visible below is named the "Bézier caracteristic matrix" sometimes):

$$M_3(t)=\begin{pmatrix}t^2 & t & 1\end{pmatrix}\begin{bmatrix}1 & -2 & 1\\ -2 & 2 & 0\\ 1 & 0 & 0\end{bmatrix}\begin{pmatrix}P_1\\ P_2\\ P_3\end{pmatrix} \tag{14.1.1122}$$

By the same reasoning, we get for a Bézier curve of order 3 (it's by far the most common Bézier curve in graphic design!):

$$\boxed{M_4(t)=(1-t)^3P_1+3t(1-t)P_2+3t^3(1-t)P_3+t^3P_4} \tag{14.1.1123}$$

or in vector form:

$$\overrightarrow{\text{O}P}=\vec{M}_4(t)=\begin{pmatrix}x\\ y\end{pmatrix}=(1-t)^3\overrightarrow{\text{O}P_1}+3t(1-t)^2\overrightarrow{\text{O}P_2}+3t^2(1-t)\overrightarrow{\text{O}P_3}+t^3\overrightarrow{\text{O}P_4} \tag{14.1.1124}$$

Which corresponds generically to:

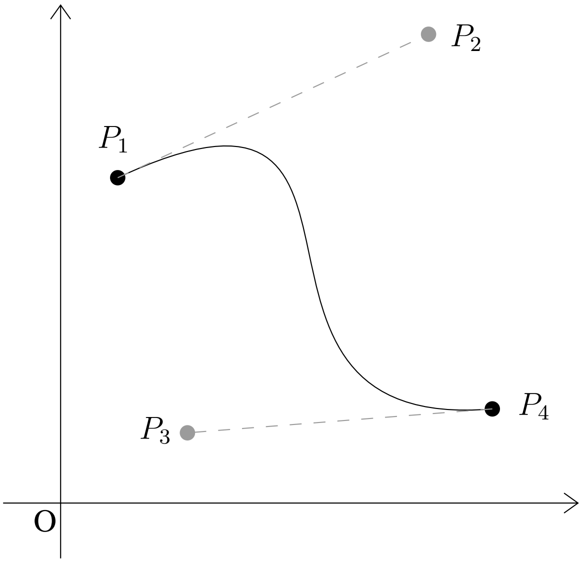


Figure 14.55: **Examples of Bézier curves of order** 3

Or:

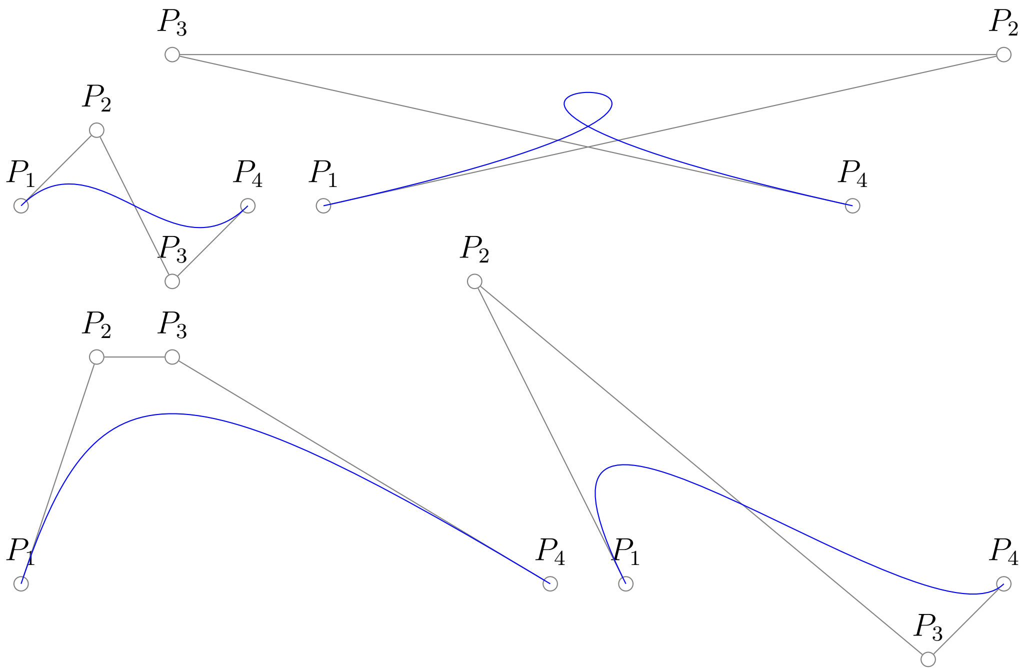


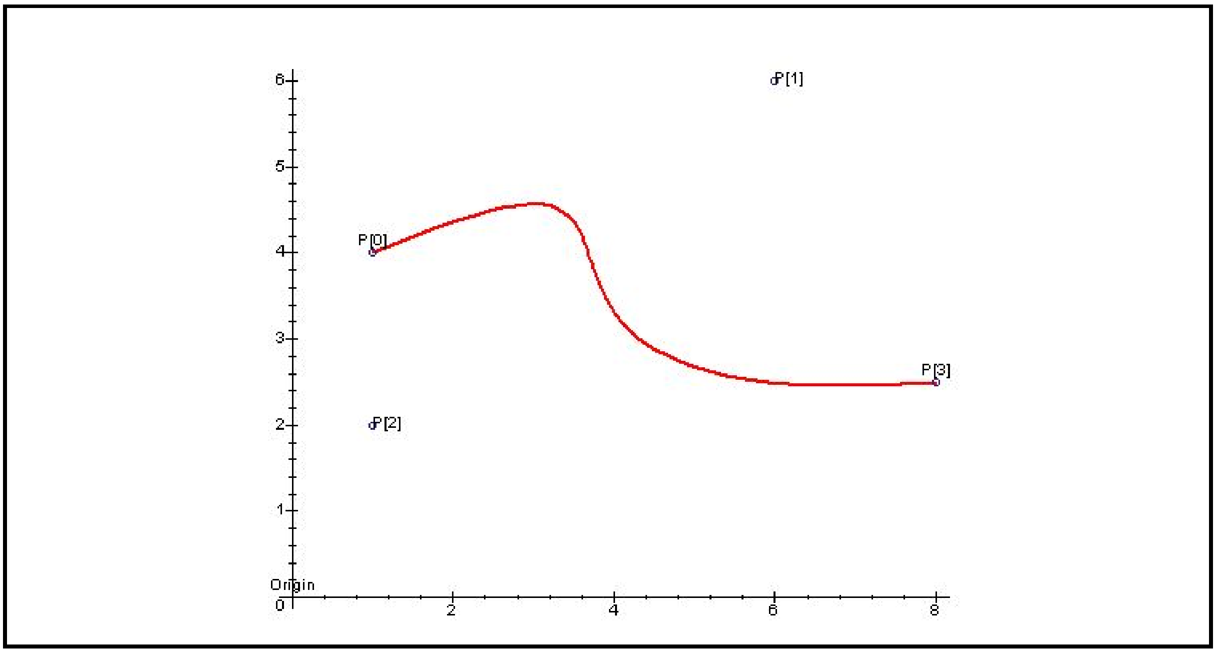
Figure 14.56: **A few 3rd order Bézier curves**

✍**Example:**

Let us plot now a Bézier curve of order 4 with Maple™ 4.00b. For this purpose we will use the following commands:

```
>restart:with(plots):
>x[0]:=1:  y[0]:=4:
>x[1]:=6:  y[1]:=6:
>x[2]:=1:  y[2]:=2:
>x[3]:=8:  y[3]:=2.5:
>f:=t->x[0]*(1-t)^3+3*x[1]*t*(1-t)^2+3*x[2]*t^2*(1-t)+x[3]*t^3;
>g:=t->y[0]*(1-t)^3+3*y[1 ]*t*(1-t)^2+3*y[2]*t^2*(1-t)+y[3]*t^3
>G:=plot([f(t),g(t),t=0..1],thickness=2):
>t0:=textplot([x[0],y[0],‘P[0]‘],align=ABOVE):
>t1:=textplot([x[1],y[1],‘P[1]‘],align=RIGHT):
>t2:=textplot([x[2],y[2],‘P[2]‘],align=RIGHT):
>t3:=textplot([x[3],y[3],‘P[3]‘],align=ABOVE):
>Or:=textplot([0,0,‘Origin‘],align=ABOVE):
>pp:=pointplot([x[0],y[0]],[x[1],y[1]],[x[2],y[2]],[x[3],y[3]],
symbol=circle,color=navy):
>display(G,t0,t1,t2,t3,pp,Or);
```

This will give:



Now let us take again the previous Bezier curve with a different notation:

$$M_4(t)=B_1P_1+B_2P_2+B_3P_3+B_4P_4=\sum_{i=1}^{4}B_iP_i \quad (14.1.1125)$$

We first notice easily the following proportionality:

$$B_i\propto c^{te}t^{i-1}(1-t)^{n-1} \quad (14.1.1126)$$

and if we look more closely at the coefficients, we notice that we have also:

$$\begin{aligned}\text{order } 2: &\quad 1\\ \text{order } 3: &\quad 121\\ \text{order } 4: &\quad 1331\end{aligned} \quad (14.1.1127)$$

It is neither more nor less than the Pascal's triangle !! So the coefficients are simply the binomial coefficients (see section Calculus page 1364) given for the order n in our example by:

$$C_{i-1}^{n-1}=\frac{(n-1)!}{(i-1)!((n-1)-(i-1))!}=\frac{(n-1)!}{(1-i)!(n-i)!} \quad (14.1.1128)$$

Thus, "Bernstein polynomials" are defined by:

$$B_{i,n}(t)=\frac{(n-1)!}{(1-i)!(n-i)!}t^{i-1}(1-t)^{n-i}=C_{i-1}^{n-1}t^{i-1}(1-t)^{n-i} \quad (14.1.1129)$$

and finally the Bernstein curves of order n are given by:

$$\boxed{M_n(t)=\sum_{i=1}^{n}B_{i,n}(t)P_i=\sum_{i=1}^{n}C_{i-1}^{n-1}t^{i-1}(1-t)^{n-i}P_i} \quad (14.1.1130)$$

In fact, if we had denoted previously the sum as follows:

$$M_4(t)=B_0P_0+B_1P_1+B_2P_2+B_3P_3=\sum_{i=0}^{3}B_iP_i \tag{14.1.1131}$$

we would then have the Bernstein polynomials that are given (which is more respectful of the traditions ...) by:

$$B_{i,n}(t)=C_i^n t^{i-1}(1-t)^{n-i} \tag{14.1.1132}$$

This is a very practical relation because it allows easily and quickly to calculate the polynomial corresponding to a sequence of Bezier curve of order n. The Bernstein polynomials can be geometrically interpreted as a factor of each point taken as a vector and summing them up to get the final Bézier curve! These polynomials might seem to come out of nowhere but we have seen that's what we get when rearranging the LERP maths to be expressed as a facto of each point. These polynomials act as influence of each control point. So in the beginning, the first point has all influence, this is then slowly sifht over the second control point, etc. until de last control point has all the influence. This is a special type of weighted sum, named "convex combination".

Then we have:

$$\boxed{\begin{aligned}x(t)&=\sum_{i=0}C_i^n t^i(1-t)^{n-i}x_i\\ y(t)&=\sum_{i=0}C_i^n t^i(1-t)^{n-i}y_i\end{aligned}} \tag{14.1.1133}$$

Remark

A Bezier curve is completely changed as soon as at least on point is moved. Then we say that the method of Bezier is a "global method".

Bézier curves have a huge issue however. If we move a control point, all the curve is influenced! In other words: we have no local control of the curve! Every movement of every point, affects the entire curve! And another issue is that given the number of LERP or polynomials to calculate, it's computationnaly unstable/expensive to calculate!

A well-known example of Bézier curves of order 3 is the "Pen tool" of Adobe Photoshop or Adobe Illustrator softwares. Indeed, these tools create a series of Bézier curves of order 3 whose point P_2 is set afterwards with the mouse using something named "Handles" or "Control Points" in Adobe practitioners language... Here is below an example taken from one of these programs done with the Pen tool with 5 points (thus 4 splines):

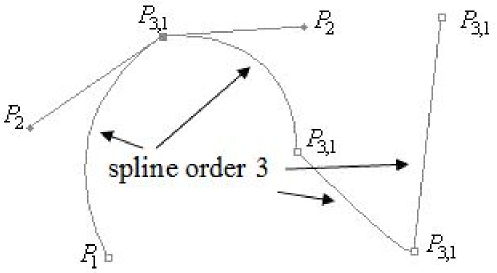


Figure 14.57: Screenshot of a Bézier spline with a drawing program

As the user does not move the points handles all points are aligned on the straight line. We then feel like to have a spline of order 2.

In the figure below we have a spline with 4 Bézier curves having obviously each one its own Bernstein Polynomials:

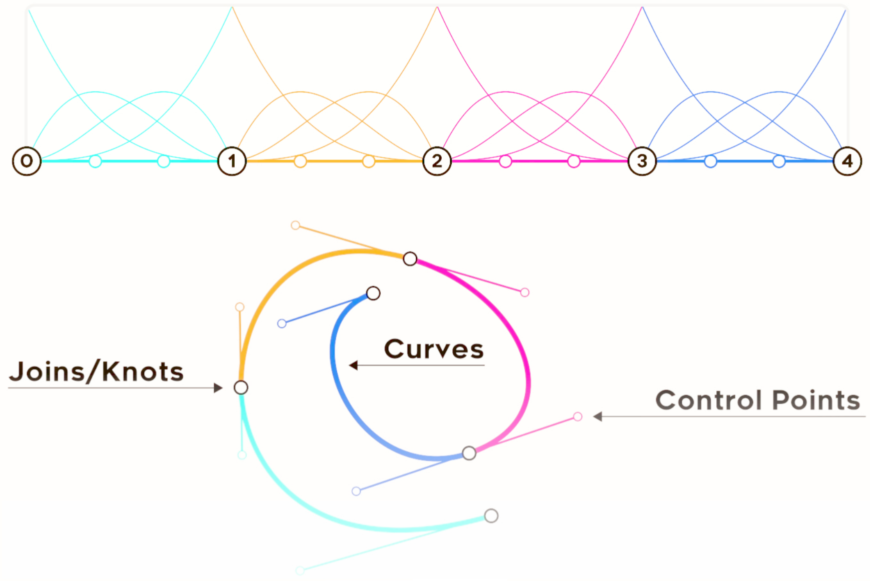


Figure 14.58: **Example of a Bézier spline with its underlying construction (author: Freya Holmér)**

✍**Example:**

A circle, drawn by a professional drawing software is in practice composed of 4 Bezier arcs. To observe this particularity, simply draw a circle with Adobe Illustrator for example, and then select it to reveal the Bezier control points arcs that defines it.

We will look at the best way to choose the control points of these arcs so that they look like circle quadrants, and then we will see the difference between the drawing produced (vector circle) and true (bitmap) circle:

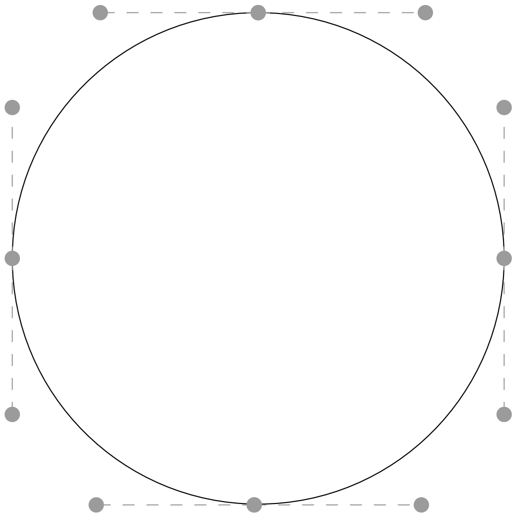
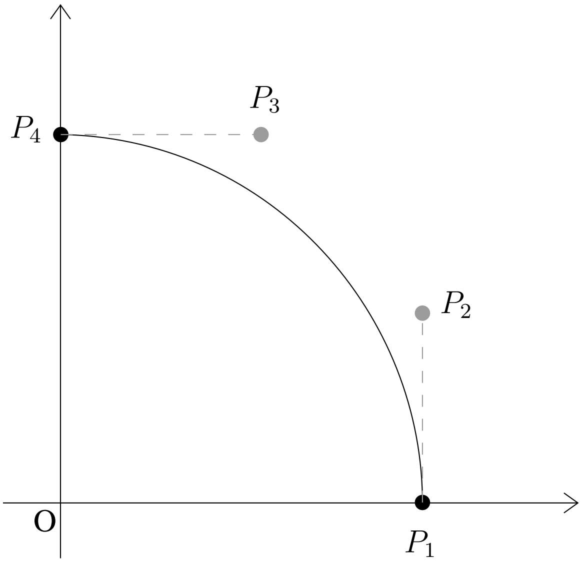


Figure 14.59: **Example of constructing a circle with Bézier curves**

Let us take the first quadrant or radius 1 centered at the origin:



It is approached by a Bézier arc whose control points are P_1, P_2, P_3, P_4. The ends of the Bézier arc being P_1 and P_4, it is natural to choose $P_1(1,0)$ and $P_4(0,1)$. Intuition leads us to choose $P_2(1,k)$ and $P_3(k,1)$ and it remains to find a positive value of k so that the Bezier curve looks like a circular arc.

We thus obtain the parametric equation of the Bézier curve:

$$\begin{pmatrix} x \\ y \end{pmatrix} = (1-t)^3 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + 3t(1-t)^2 \begin{pmatrix} 1 \\ k \end{pmatrix} + 3t^2(1-t) \begin{pmatrix} k \\ 1 \end{pmatrix} + t^3 \begin{pmatrix} 0 \\ 1 \end{pmatrix} \qquad (14.1.1134)$$

Therefore:

$$\begin{aligned} x(t) &= (1-t)^3 + 3t(1-t)^2 + 3t^2(1-t)k \\ y(t) &= 3t(1-t)^2k + 3t^2(1-t) + t^3 \end{aligned} \qquad (14.1.1135)$$

We can look for example the value of k for which the arc passes through the point:

$$P=\left(\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2}\right) \tag{14.1.1136}$$

in $t=0.5$. It then becomes very simple from the parametric equation to determine k. It is simply for x (or y) of a simple equation of one unknown.

With this tool we can now create plenty of illustrations (static or animated[31]) on computers!

---

[31] However animation with Bézier splines have huge flaws that requires to build a new kind of spline named "Hermite spline"

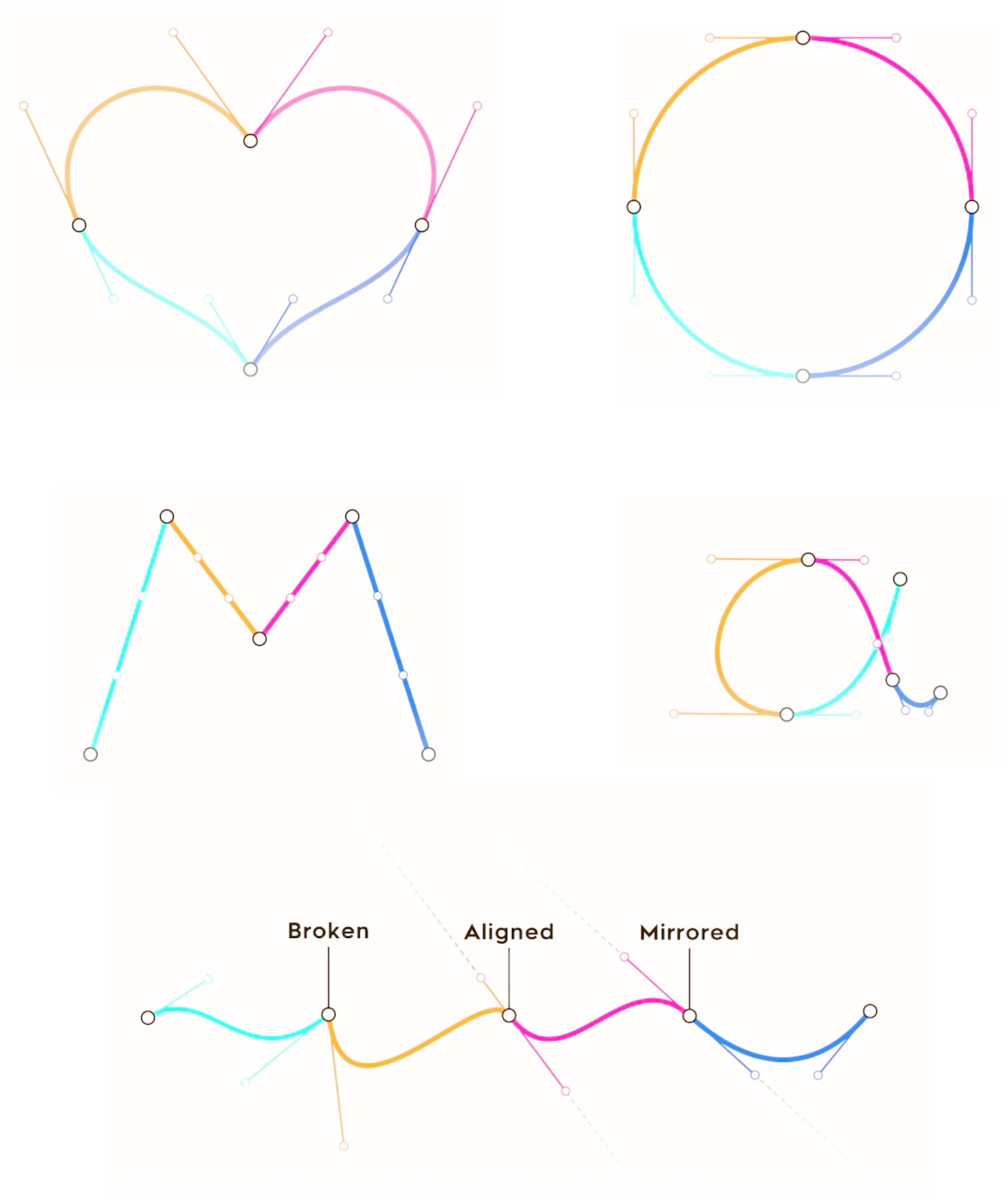


Figure 14.60: **A few practical example of Bézier splines (author: Freya Holmér)**

There are plenty of other splines useful in practice used to eliminate various flaws (but each of these splines has also some flaws): Uniform splines, Nonuniform splines, Natural splines, Hermite splines, Cardinal splines, B-splines (the "B" is not an abreviation for Bézier but for "basis"), Catmull-Rome splines, Interpolating splines, Non-uniform rational B-spline (NURBS)...

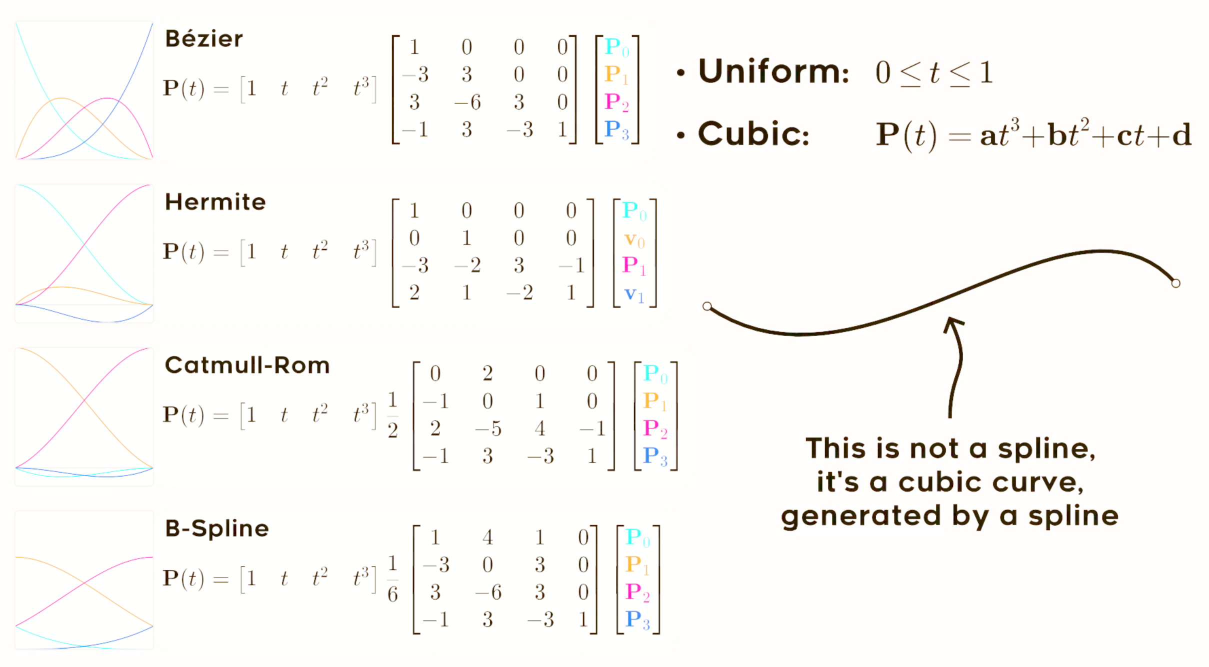

Bézier
P(t) = [1 t t² t³] [1 0 0 0; −3 3 0 0; 3 −6 3 0; −1 3 −3 1] [P0; P1; P2; P3]
Hermite
P(t) = [1 t t² t³] [1 0 0 0; 0 1 0 0; −3 −2 3 −1; 2 1 −2 1] [P0; v0; P1; v1]
Catmull-Rom
P(t) = [1 t t² t³] 1/2 [0 2 0 0; −1 0 1 0; 2 −5 4 −1; −1 3 −3 1] [P0; P1; P2; P3]
B-Spline
P(t) = [1 t t² t³] 1/6 [1 4 1 0; −3 0 3 0; 3 −6 3 0; −1 3 −3 1] [P0; P1; P2; P3]
• Uniform: 0 ≤ t ≤ 1
• Cubic: P(t) = at³+bt²+ct+d
This is not a spline, it's a cubic curve, generated by a spline

Figure 14.61: A visual summary of naive splines (author: Freya Holmér)

Name	Degree	Continuity	Tangents	Interpol.	Use cases
Bézier	3	C^0/C^1	manual	some	shapes, fonts, vector graphics
Hermite	3	C^0/C^1	explicit	all	animation, physics simulation, interpolation
Catmull-Rom	3	C^1	auto	all	animation, path smoothing
B-Spline	3	C^2	auto	none	curvature-sensitive shapes, animations such as camera paths
Linear	1	C^0	auto	all	dense data & interpolation where smoothness doesn't matter
...	...	...	...	...	...

Table 14.19: Table summary of naive splines (inspired by Freya Holmér)

Remark

Keep in mind, as the vocabulary may be confusing, that:

- If there is only one (polynomial) segment, the spline is often named a "Bézier curve".
- If each segment is expressed in Bézier form (using Bernstein basis functions), then you might say that the spline is a "Bézier spline".
- If each polynomial segment has degree 3, the spline is named a "cubic spline".
- If each segment is described by its ending positions and derivatives, it is said to be in "Hermite form".
- The B-spline approach (for which we will detail the underlying mathematics in the future) gives a way of ensuring continuity between segments. In fact, we show that

every spline can be represented in B-spline form. So, in that sense, every spline is a B-spline.

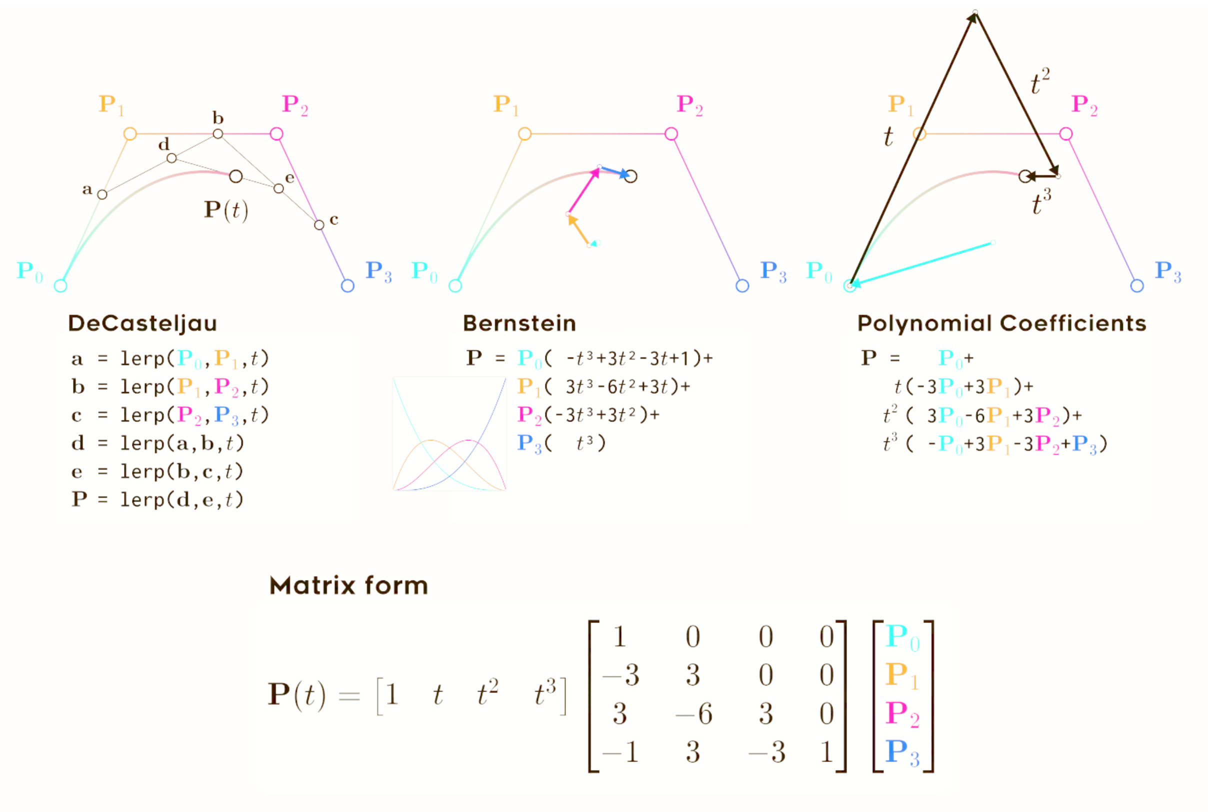

P1
b
P2
d
a
e
P(t)
c
P0
P3
DeCasteljau
a = lerp(P0,P1,t)
b = lerp(P1,P2,t)
c = lerp(P2,P3,t)
d = lerp(a,b,t)
e = lerp(b,c,t)
P = lerp(d,e,t)
P1
P2
P0
P3
Bernstein
P = P0(-t³+3t²-3t+1)+
P1(3t³-6t²+3t)+
P2(-3t³+3t²)+
P3(t³)
t²
P1
P2
t
t³
P0
P3
Polynomial Coefficients
P = P0+
t(-3P0+3P1)+
t² (3P0-6P1+3P2)+
t³ (-P0+3P1-3P2+P3)
Matrix form
P(t) = [1 t t² t³]
1 0 0 0
−3 3 0 0
3 −6 3 0
−1 3 −3 1
P0
P1
P2
P3

Figure 14.62: Summary of different ways of calculating Bézier curves (author: Freya Holmér)

14.1.13.2 Linear ordering isotonic regression

"Linear ordering isotonic regression" can be understood as approximating given series of 1-dimensional observations with non-decreasing function. It is similar to inexact smoothing splines, with the difference that we use monotonicity, rather than smoothness, to remove noise from the data. Isotonic relation is quite well known in the field of machine learning for models calibration.

Given values $\vec{a} \in \mathbb{R}^n$ and weights $\vec{w} \in \mathbb{R}^n$ such that $w_i > 0$ for all i, we want to approximate them by $y_1, \dots y_n$ as close as possible, subject to a set of constraints of kind $y_i \geq y_j$, i.e. we want to minimize (obviously sometimes other metrics are used instead of the Euclidean metric!):

$$\sum_i w_i \left(y_i - a_i\right)^2 \tag{14.1.1137}$$

with respect to $\vec{y} \in \mathbb{R}^n$ subject to $y_1 \leq y_2 \leq \ldots \leq y_n$.

If all weights equal to 1, the problem is named "unweighted isotonic regression", otherwise it is named "weighted isotonic regression".

In pseudocode (non-unique and very likely not optimized):

Algorithm 6: Pool Adjacent Violators Algorithm (PAVA) pseudocode algorithm

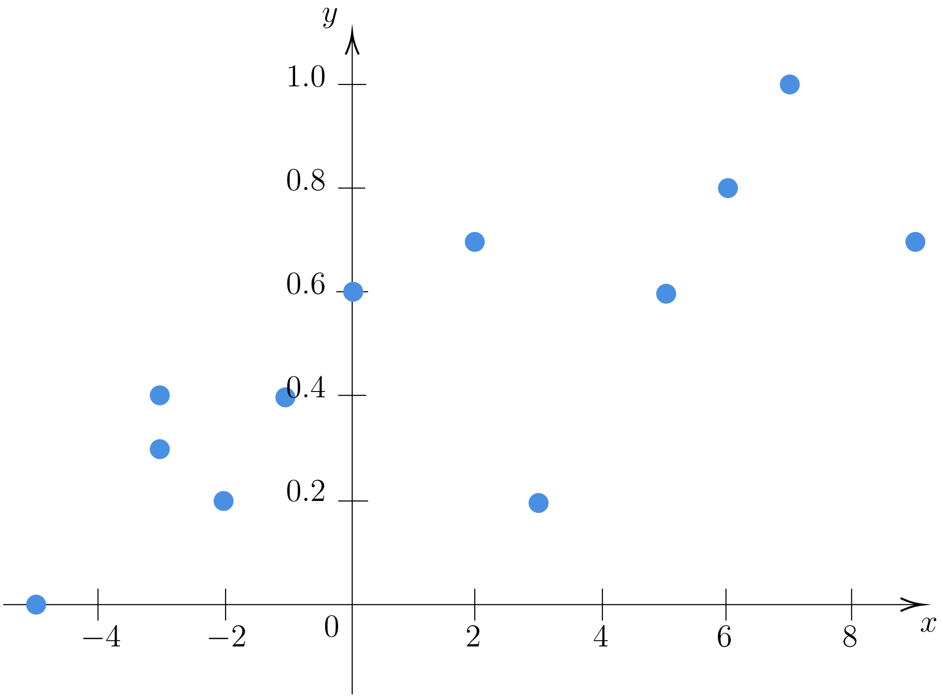
```
Data: $\vec{a}$,$\vec{w}$
Result: $\vec{y}$
initialization;
$a'_1 := a_1$;
$w'_1 := w_1$;
$j := 1$;
$S_0 := 0$;
$S_1 := 1$;
for $i = 2, 3, \ldots, n$ do
    $j := j + 1$ ;
    $a'_j := a_i$ ;
    $w'_j := w_i$;
    while $j > 1$ AND $a'_j < a'_{j-1}$ do
        $a'_{j-1} = \dfrac{w'_j a'_j + w'_{j-1} a'_{j-1}}{w'_j + w'_{j-1}}$;
        $w'_{j-1} = w'_j + w'_{j-1}$;
        $j := j - 1$;
    $S_j := i$;
for $k = 1, 2, \ldots, j$ do
    for $l = S_{k-1} + 1, \ldots, S_k$ do
        $y_l = a'_k$;
```

Here S defines to which old points each new point corresponds.

Let us see a very detailed companion example (thanks to Wojciech Kotłowski for having authorized us to reproduce it!).

Let us consider:

x	7	-1	-2	9	2	0	6	3	-3	5	-3	7	-5
y	1	0.4	0.2	0.7	0.7	0.6	0.8	0.2	0.3	0.6	0.4	1	0



Step 1: We sort the data in the increasing order of x.

x	7	−1	−2	9	2	0	6	3	−3	5	−3	7	−5
y	1	0.4	0.2	0.7	0.7	0.6	0.8	0.2	0.3	0.6	0.4	1	0

$\Downarrow \quad \Downarrow \quad \Downarrow$

x	−5	−3	−3	−2	−1	0	2	3	5	6	7	7	9
y	0	0.4	0.3	0.2	0.4	0.6	0.7	0.2	0.6	0.8	1	1	0.7

Step 2: We split the data into blocks $B_1, \ldots, B_r$, such that points with the same x_t fall into the same block.

We assign value f_i to each block ($i = 1, \ldots, r$) which is the average of labels in this block.

x	−5	−3	−3	−2	−1	0	2	3	5	6	7	7	9
y	0	0.4	0.3	0.2	0.4	0.6	0.7	0.2	0.6	0.8	1	1	0.7

$\Downarrow \quad \Downarrow \quad \Downarrow$

block	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9	B_{10}	B_{11}
data	$\{1\}$	$\{2, 3\}$	$\{4\}$	$\{5\}$	$\{6\}$	$\{7\}$	$\{8\}$	$\{9\}$	$\{10\}$	$\{11, 12\}$	$\{13\}$
f_i	0	0.35	0.2	0.4	0.6	0.7	0.2	0.6	0.8	1	0.7

Step 3: While there exists a "violator", i.e. a pair of blocks B_i, B_{i+1} such that $f_i > f_{i+1}$, we merge B_i and B_{i+1} and assign a weighted average:

$$f_i = \frac{|B_i| f_i + |B_{i+1}| f_{i+1}}{|B_i| + |B_{i+1}|} \tag{14.1.1138}$$

block	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9	B_{10}	B_{11}
data	$\{1\}$	$\{2,3\}$	$\{4\}$	$\{5\}$	$\{6\}$	$\{7\}$	$\{8\}$	$\{9\}$	$\{10\}$	$\{11,12\}$	$\{13\}$
f_i	0	0.35	0.2	0.4	0.6	0.7	0.2	0.6	0.8	1	0.7

⇓ ⇓ ⇓

block	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9	B_{10}
data	$\{1\}$	$\{2,3,4\}$	$\{5\}$	$\{6\}$	$\{7\}$	$\{8\}$	$\{9\}$	$\{10\}$	$\{11,12\}$	$\{13\}$
f_i	0	0.3	0.4	0.6	0.7	0.2	0.6	0.8	1	0.7

We iterate again:

block	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9	B_{10}
data	$\{1\}$	$\{2,3,4\}$	$\{5\}$	$\{6\}$	$\{7\}$	$\{8\}$	$\{9\}$	$\{10\}$	$\{11,12\}$	$\{13\}$
f_i	0	0.3	0.4	0.6	0.7	0.2	0.6	0.8	1	0.7

⇓ ⇓ ⇓

block	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9
data	$\{1\}$	$\{2,3,4\}$	$\{5\}$	$\{6\}$	$\{7,8\}$	$\{9\}$	$\{10\}$	$\{11,12\}$	$\{13\}$
f_i	0	0.3	0.4	0.6	0.45	0.6	0.8	1	0.7

We iterate once again:

block	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9
data	$\{1\}$	$\{2,3,4\}$	$\{5\}$	$\{6\}$	$\{7,8\}$	$\{9\}$	$\{10\}$	$\{11,12\}$	$\{13\}$
f_i	0	0.3	0.4	0.6	0.45	0.6	0.8	1	0.7

⇓ ⇓ ⇓

block	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8
data	$\{1\}$	$\{2,3,4\}$	$\{5\}$	$\{6,7,8\}$	$\{9\}$	$\{10\}$	$\{11,12\}$	$\{13\}$
f_i	0	0.3	0.4	0.5	0.6	0.8	1	0.7

We iterate once again:

block	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8
data	$\{1\}$	$\{2,3,4\}$	$\{5\}$	$\{6,7,8\}$	$\{9\}$	$\{10\}$	$\{11,12\}$	$\{13\}$
f_i	0	0.3	0.4	0.5	0.6	0.8	1	0.7

$$\Downarrow \quad \Downarrow \quad \Downarrow$$

block	B_1	B_2	B_3	B_4	B_5	B_6	B_7
data	$\{1\}$	$\{2,3,4\}$	$\{5\}$	$\{6,7,8\}$	$\{9\}$	$\{10\}$	$\{11,12,13\}$
f_i	0	0.3	0.4	0.5	0.6	0.8	0.9

There is no more violators yet!

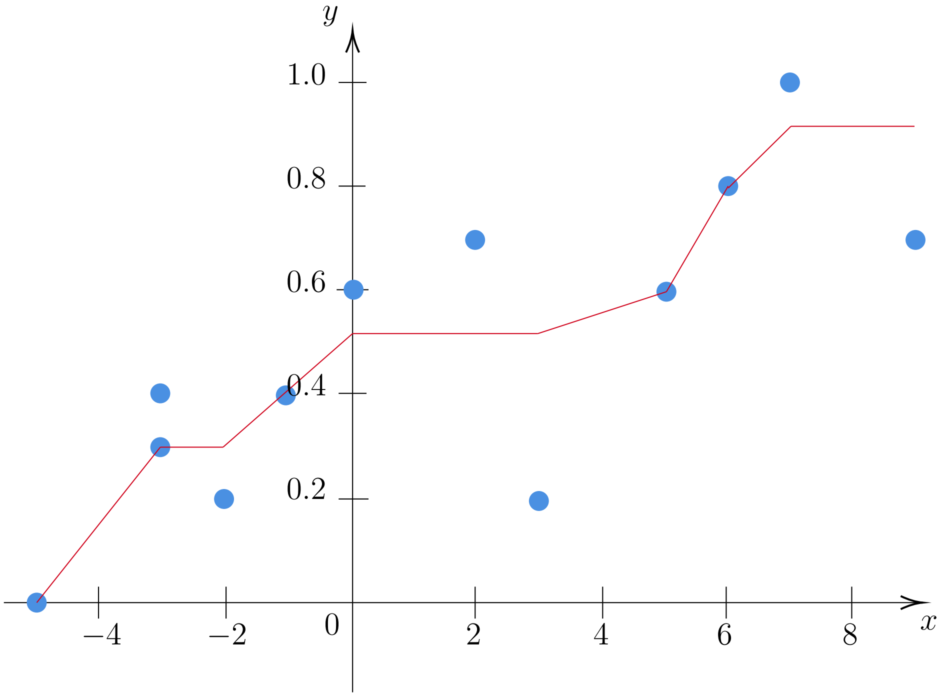
Reading out the solution:

block	B_1	B_2	B_3	B_4	B_5	B_6	B_7
data	$\{1\}$	$\{2,3,4\}$	$\{5\}$	$\{6,7,8\}$	$\{9\}$	$\{10\}$	$\{11,12,13\}$
f_i	0	0.3	0.4	0.5	0.6	0.8	0.9

$$\Downarrow \quad \Downarrow \quad \Downarrow$$

x	-5	-3	-3	-2	-1	0	2	3	5	6	7	7	9
y	0	0.4	0.3	0.2	0.4	0.6	0.7	0.2	0.6	0.8	1	1	0.7
f^*	0	0.3	0.3	0.3	0.4	0.5	0.5	0.5	0.6	0.8	0.9	0.9	0.9

Which gives:



This corresponds perfectly to what we get with the `R` software (see our companion book on the subject!).

14.1.13.3 Euler Method

Strictly speaking the Euler Method is much more than an interpolation method. But it can be introduced first as a liner piece-wise interpolation technique. In fact it provides an approximation (in the broadest sense) of a function $f(x)$ whose first derivative is known.

The interpolated points are not necessarily equidistant as the interpolation is piece-wise but as soon we extrapolate, the step is therefore by tradition equidistant and denoted by h and the distance between two points given by: $x_i = x_0 + ih$. We denote by $f(x)$ the exact value and y_i the approximate value.

The idea is quite simple. If in $\mathbb{R}^2$ we know two points (y_i, x_i) and (y_{i+1}, x_{i+1}) we can calculate piece-wise the slope as:

$$\frac{y_{i+1}-y_i}{x_{i+1}-x_i} = \frac{y_{i+1}-y_i}{\delta x} := \frac{y_{i+1}-y_i}{h} \cong f' \tag{14.1.1139}$$

In more complicated examples and real applications of the Euler methods we know the true derivative on x_i.

The previous relation can be written:

$$y_{i+1} \cong y_i + hf' \tag{14.1.1140}$$

And therefore for small h we can do a linear interpolation between two points.

If the derivative is known (and the derivative can be of non-linear function!) we have the more general relation:

$$y_{i+1}=y_i+hf'(x_i) \tag{14.1.1141}$$

That is traditionally written:

$$\boxed{y_{i+1}=y_i+hy'(x_i)} \tag{14.1.1142}$$

named also the "Euler difference equation". This relation can also be used to solve some simple numerical differential equation as illustrated by the following example:

✍**Examples:**

E1. Let's solve the following initial value problem ("Cauchy problem" style with $y'(x)=f'(x)$ and $y(x_0)=y_0$):

$$y'=y\,;\; y(0)=1$$

using the Euler's method. The recurrence relation can be written in is this special case:

$$y_{n+1}=(1+h)y_n \qquad y_0=1$$

i.e.:

$$y_n=(1+h)^n$$

For a given x we have:

$$n=\frac{x}{h}$$

and thus:

$$y_n=(1+h)^{\frac{x}{h}}=[(1+h)^{\frac{1}{h}}]^x$$

For $h\rightarrow 0_+$ the approximate solution y_n converges to the exact solution e^x.

E2. Suppose that we are given the equation $5y'-3y=0$ with boundary condition $y=y_0$ at $x=0$. This gives a unique function y. We can approximate this by considering discrete steps of length h between x_n and x_{n+1} (using the simple Euler numerical scheme) we can approximate the equation by:

$$5\frac{y_{n+1}-y_n}{h}-3y_n\approx 0$$

Rearranging the terms, we have $y_{n+1}\approx(1+\frac{3}{5}h)y_n$. Applying the relation successively, we have

$$\begin{aligned}y_n&=\left(1+\frac{3}{5}h\right)y_{n-1}\\&=\left(1+\frac{3}{5}h\right)\left(1+\frac{3}{5}h\right)y_{n-2}\end{aligned}$$

$$=\left(1+\frac{3}{5}h\right)^n y_0$$

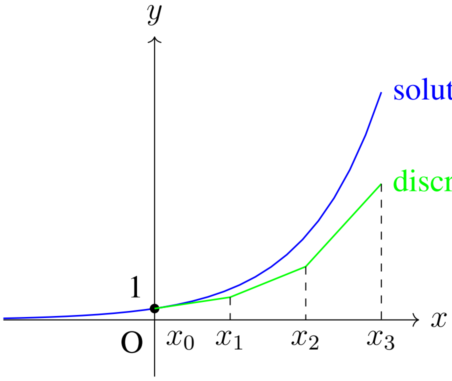
For each given value of x, choose $h=x/n$, so:

$$y_n=y_0\left(1+\frac{3}{5}(x/n)\right)^n$$

Taking the limit as $n\rightarrow\infty$, we have:

$$y(x)=\lim_{n\rightarrow+\infty} y_0\left(1+\frac{3x/5}{n}\right)^n=y_0e^{3x/5}$$

in agreement with the solution of the differential equation.



The "local truncation error" (LTE) of the Euler method is an error made in a single step. It is the difference between the numerical solution after one step.

To calculate this error we can write first:

$$y_1=y_0+hf(x_0) \quad (14.1.1143)$$

For the exact solution (considered as exact...), we use the Taylor expansion (see section Sequences and Series page 1696):

$$y(x_0+h)=y(t_0)+hy'(t_0)+\frac{1}{2}h^2y''(t_0)+\mathcal{O}(h^3) \quad (14.1.1144)$$

Also sometimes written:

$$y(x_n+h):=y_{n+1}=y_n+h=y'(x_n)+\frac{h^2}{2}y''(x_n)+\ldots+\frac{h^p}{p!}y^{(p)}(x_n)+\mathcal{O}(h^{p+1}) \quad (14.1.1145)$$

This relation can also be used to solve some simple numerical differential equation as illustrated by the following example:

✍**Example:**

We want to use the method of Taylor's expansion of the third order to solve the following initial value problem (again a "Cauchy problem" style with $y'(x)=f'(x)$ and $y(x_0)=y_0$):

$$y'=\frac{4}{x^2}-y^2-\frac{y}{x} \qquad \text{with: } y(1)=0$$

For $n=0,1,2,\ldots$ we have:

$$y(x_n+h):=y_{n+1}=y_n+hy'(x_n)+\frac{h^2}{2}y''(x_n)+\frac{h^3}{3!}y'''(x_n)+\mathcal{O}(h^4)$$

Here:

$$\begin{aligned}
x_0=1\,, \qquad & y_0=0\,,\\
y'(x_n) &= \frac{4}{x_n^2}-y_n^2-\frac{y_n}{x_n}\,,\\
y''(x_n) &= -\frac{8}{x_n^3}-2y_ny'(x_n)-\frac{y'(x_n)x_n-y_n}{x_n^2}=-\frac{12}{x_n^3}-\frac{6y_n}{x_n^2}+\frac{3y_n^2}{x_n}+2y_n^3\,,\\
y'''(x_n) &= \frac{24}{x_n^4}-2(y'(x_n))^2-2y_ny''(x_n)-\frac{y''(x_n)x_n^2-2(y'(x_n)x_n-y_n)}{x_n^3}\\
&= \frac{12}{x_n^4}-\frac{42y_n}{x_n^3}+\frac{21y_n^2}{x_n^2}-\frac{12y_n^3}{x_n}-6y_n^4\,.
\end{aligned}$$

The table below shows the computed values of the solution in the point $x_N=2$ for various N (and thus for various $h=1/N$):

	x	1	1.2	1.4	1.6	1.8	2
$h=0.2$	$y(x)$	0	0.576000	0.835950	0.920226	0.920287	0.884745
$h=0.1$		0	0.581645	0.838338	0.919251	0.918141	0.882631
$h=0.05$		0	0.582110	0.838443	0.919062	0.917872	0.882386
Exact solution: $y(x)=\dfrac{2(x^4-1)}{x(x^4+1)}$, $y(2)=0.882353$							

The local truncation error (one step!!!) introduced by the Euler method is given by the difference between these equations:

$$\text{LTE}=y(x_0+h)-y_1=\frac{1}{2}h^2y''(t_0)+\mathcal{O}(h^3) \qquad (14.1.1146)$$

This shows that for small h, the local truncation error is approximately proportional to h^2. This makes the Euler method less accurate (for small h) than other higher-order techniques such as Runge-Kutta methods that we will see later.

14.1.13.4 Polynomial of collocation

In the section of Sequences and Series we discussed the general unsuitability of Taylor polynomials for approximation. These polynomials are useful only over small intervals for functions whose derivatives exist and are easily evaluated. In this section we find approximating polynomials that can be determined simply by specifying certain points on the plane through which they must pass.

Given $y=f(x)$ a known function in explicit form or in tabulated form, and suppose that a given number of values:

$$f_i=f(x_i) \quad i=0,1,\ldots,n \tag{14.1.1147}$$

are given. The known points $(x_i, f(x_i))$ are named "support points".

"Interpolate f" in this context means to estimate the values of f for the horizontal x knows values located between x_0 and x_n, that is to say in the range of interpolation by an approximate function $y=P(x)$, which satisfies the "collocations conditions" (nothing to do with your roommate...!):

$$P(x_i)=f(x_i) \quad i=0,1,\ldots,n \tag{14.1.1148}$$

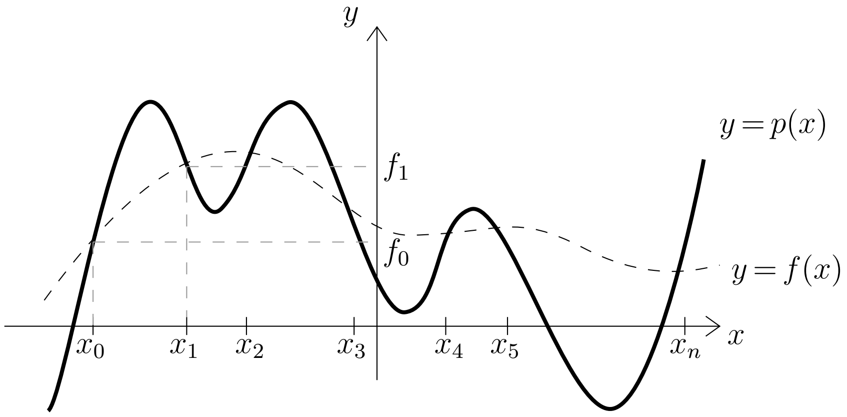


Figure 14.63: Illustration of the interpolation concept with the collocation technique

The function P is named "collocation function" on the x_i. When P is a polynomial, we speak of "collocation polynomial" or of "interpolation polynomial".

"Extrapolate" a function means study the $f(x)$ by $P(x)$ for x-located "outside" of the interpolation interval.

Remark

It goes without saying that the interpolation is a very important tool for all researchers, statisticians and others.

When we know a polynomial of degree n on $n+1$ points, we can know then by a simple way (but not very fast - but there are several methods) this polynomial completely.

To determine the polynomial, we will use the results presented above in our study of systems of linear equations. The disadvantage of the method presented here is that you have to guess to what type of polynomial you are dealing with (the order at least) and know what are the good points to chose (when there is a choice...).

A particular example should suffice for understanding this method, the generalization being quite simple (see further below).

Given the univariate second degree polynomial:

$$P_2(x)=a_0+a_1x+a_2x^2 \tag{14.1.1149}$$

and we know the following point (you will notice the ingenuity of the points selected by the authors of these lines ...):

$$P(0)=4, P(1)=8, P(-1)=2 \tag{14.1.1150}$$

We deduce therefore the following system of equations:

$$\begin{aligned}
&a_0+a_10+a_20^2=4\\
&a_0+a_11+a_21^2=2\\
&a_0+a_1(-1)+a_2(-1)^2=8
\end{aligned} \tag{14.1.1151}$$

System that once solved using the state of the art techniques (see section Linear Algebra page 1929) gives us:

$$a_0=4, a_1=-3, a_2=1 \tag{14.1.1152}$$

Let us see the general case:

Theorem 14.4. Given $(x_i, f(x_i))$ support points, with $x_i \neq x_j$ if $i \neq j$. Then there exists a polynomial $P_n(x)$ of degree less than or equal to n, and only one, such as $P_n(x_i)=f_i$ for $i=0,1,\ldots,n$.

Proof 14.4.1. Let us write:

$$P_n(x)=a_0+a_1x+\ldots+a_nx^n \tag{14.1.1153}$$

The collocation conditions:

$$P_(x_i)=f_i \quad i=0,1,\ldots,n \tag{14.1.1154}$$

are thus written:

$$\begin{aligned}
&a_0+a_1x_0+a_2x_0^2+\ldots+a_nx_0^n=f_0\\
&a_0+a_1x_1+a_2x_1^2+\ldots+a_nx_1^n=f_1\\
&\vdots\\
&a_0+a_1x_n+a_2x_n^2+\ldots+a_nx_n^n=f_n
\end{aligned} \tag{14.1.1155}$$

This is a system of $n+1$ equations in $n+1$ unknowns.

Its determinant is written (see section Linear Algebra page 1961):

$$D_n=\det\begin{vmatrix}1 & x_0 & x_0^2 & \cdots & x_0^n\\ 1 & x_1 & x_1^2 & \cdots & x_1^n\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & x_n & x_n^2 & \cdots & x_n^n\end{vmatrix}=\prod_{n\geq i>j\geq 0}^{n}(x_i-x_j) \tag{14.1.1156}$$

relation that we name "Vandermonde determinant". We know that if the system has a solution, the determinant of the system must be non-zero (see section Linear Algebra page 1980).

Let us show by an example (by taking a polynomial of the same degree as the one we used above) that the determinant is calculated using the previous relation (the reader will generalize by induction):

Therefore in the case $n=2$, we consider the determinant:

$$D_2=\det\begin{vmatrix}1 & x_0^1 & x_0^2\\ 1 & x_1^1 & x_1^2\\ 1 & x_2^1 & x_2^2\end{vmatrix} \tag{14.1.1157}$$

thus corresponding to the system (for reminder):

$$\begin{aligned}a_0\cdot 1+a_1x_0^1+a_2x_0^2&=f_0\\ a_0\cdot 1+a_1x_1^1+a_2x_1^2&=f_1\\ a_0\cdot 1+a_1x_n^1+a_2x_2^2&=f_2\end{aligned} \tag{14.1.1158}$$

Let us calculate this determinant following the column 1 (by making use of cofactors as proved in the section of Linear Algebra at page 1980):

$$\begin{aligned}D_n=\det\begin{vmatrix}1 & x_0^1 & x_0^2\\ 1 & x_1^1 & x_1^2\\ 1 & x_2^1 & x_2^2\end{vmatrix}&=1\cdot\det\begin{vmatrix}x_1^1 & x_1^2\\ x_2^1 & x_2^2\end{vmatrix}-1\cdot\det\begin{vmatrix}x_0^1 & x_0^2\\ x_2^1 & x_2^2\end{vmatrix}+1\cdot\det\begin{vmatrix}x_0^1 & x_0^2\\ x_1^1 & x_1^2\end{vmatrix}\\ &=x_1^1x_2^2-x_2^1x_1^2-x_0^1x_2^2+x_2^1x_0^2+x_0^1x_1^2-x_1^1x_0^2\end{aligned} \tag{14.1.1159}$$

This polynomial can be written:

$$D=(x_1-x_0)(x_2-x_0)(x_2-x_1) \tag{14.1.1160}$$

Which is written:

$$D_2=\prod_{n\geq i>j\geq 0}^{n}(x_i-x_j) \tag{14.1.1161}$$

As the x_9 is in the statement of our problem are all different such that $x_i\neq x_j$ then the system has a unique solution! This proves that there is always an interpolation polynomial.

■ Q.E.D.

It should be noted however that it is not a method of polynomial regression. Indeed, with a method of polynomial regression, we might choose a higher degree for the polynomial as the number of points we have.

14.1.13.5 Lagrange polynomials

The interpolation polynomial Lagrange method (used a lot in practice because the algorithm is very simple and therefore effective) consider that we initially have (know) $n+1$ points such as:

$$\begin{aligned} y_0 &= f(x_0) \\ y_1 &= f(x_1) \\ &\ldots \\ y_i &= f(x_i) \\ &\ldots \\ y_n &= f(x_n) \end{aligned} \tag{14.1.1162}$$

and that we are looking for a collocation polynomial that passes through all the points. The idea of polynomial interpolation of Lagrange is therefore simple and very clever (as always someone must have think to it first...). Let us observe the chart below where we have 5 points by which we seek to pass a collocation polynomial:

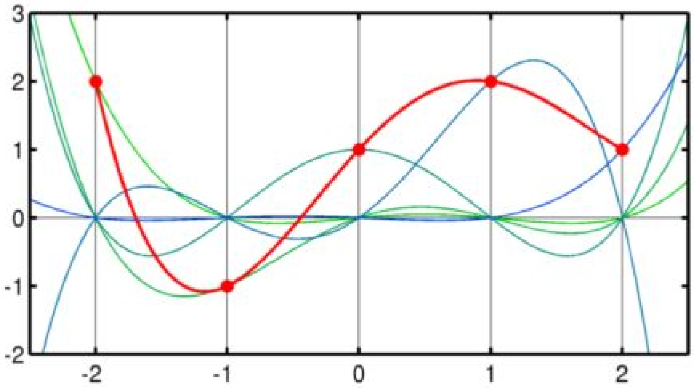


Figure 14.64: **Illustration of concept of Lagrange interpolation (source: Wikipedia)**

To find the collocation polynomial in red (we obviously suppose we don't know it initially), the idea is that for every x_i we associate a different polynomial of degree n and not null on $f(x_i)$ but null on all other points $(x_j)_{j\neq i}$. For this, as we see in the graph above, it is necessary that for each i we associate a polynomial that has n roots as:

$$\begin{aligned} f_0(x) &= A_0(x-x_1)(x-x_2)(x-x_3)(x-x_4) \\ f_1(x) &= A_1(x-x_0)(x-x_2)(x-x_3)(x-x_4) \\ f_2(x) &= A_2(x-x_0)(x-x_1)(x-x_3)(x-x_4) \\ f_3(x) &= A_3(x-x_0)(x-x_1)(x-x_2)(x-x_4) \\ f_4(x) &= A_4(x-x_0)(x-x_1)(x-x_2)(x-x_3) \end{aligned} \tag{14.1.1163}$$

So we see that we have 5 polynomials, each with 4 roots (hence of order 4) and which are respectively zero on all the points $x_j \neq x_i$ (as you can see on the graph). The coefficients A_i are constants to be determined.

Now nothing prevents us to sum all polynomial together since we will always have the right value on $f(x_i)$ (since the other polynomials are zero on this same point). Either by generalization the notation, the sum becomes:

$$\begin{aligned}P(x)=&A_0(x-x_1)(x-x_2)(x-x_3)\ldots(x-x_n)+A_1(x-x_0)(x-x_2)(x-x_3)\ldots(x-x_n)\\&+A_i(x-x_0)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)+A_n(x-x_0)(x-x_1)(x-x_2)\ldots(x-x_{n-1})\end{aligned} \tag{14.1.1164}$$

By injecting respectively $x_0,x_1,\ldots,x_i,\ldots,x_n$, we get in the general case:

$$\begin{aligned}P(x_0)&=f(x_0)=A_0(x_0-x_1)(x_0-x_2)(x_0-x_3)\cdots(x_0-x_n)\\P(x_1)&=f(x_1)=A_1(x_1-x_0)(x_1-x_2)(x_1-x_3)\cdots(x_1-x_n)\\P(x_2)&=f(x_2)=A_2(x_2-x_0)(x_2-x_1)(x_2-x_3)\cdots(x_2-x_n)\\&\ldots\\P(x_n)&=f(x_n)=A_n(x_n-x_0)(x_n-x_1)(x_n-x_2)\cdots(x_n-x_{n-1})\end{aligned} \tag{14.1.1165}$$

We deduce then immediately in the general case:

$$\begin{aligned}A_0&=\frac{f(x_0)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)\cdots(x_0-x_n)}\\A_1&=\frac{f(x_1)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)\cdots(x_1-x_n)}\\A_2&=\frac{f(x_2)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)\cdots(x_2-x_n)}\\&\ldots\\A_n&=\frac{f(x_n)}{(x_n-x_0)(x_n-x_1)(x_n-x_2)\cdots(x_n-x_{n-1})}\end{aligned} \tag{14.1.1166}$$

Substituting the values of the constants in the initial expression of the sum, we get:

$$\begin{aligned}P(x)=&A_0(x-x_1)(x-x_2)(x-x_3)\ldots(x-x_n)+A_1(x-x_0)(x-x_2)(x-x_3)\ldots(x-x_n)\\&+A_i(x-x_0)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)+A_n(x-x_0)(x-x_1)(x-x_2)\ldots(x-x_{n-1})\\&=\frac{(x-x_1)(x-x_2)(x-x_3)\ldots(x-x_n)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)\cdots(x_0-x_n)}f(x_0)+\frac{(x-x_0)(x-x_2)(x-x_3)\ldots(x-x_n)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)\cdots(x_1-x_n)}f(x_1)\\&+\frac{(x-x_0)(x-x_1)(x-x_2)\ldots(x-x_{n-1})}{(x_n-x_0)(x_n-x_1)(x_n-x_2)\cdots(x_n-x_{n-1})}f(x_n)\end{aligned} \tag{14.1.1167}$$

Which can be written in condensed form (this relation will be useful to us later for some numerical integration methods!):

$$\boxed{P(x)=\sum_{j=0}^{n}\left(\prod_{i=0,i\neq j}^{n}\frac{x-x_i}{x_j-x_i}\right)f(x_j)} \tag{14.1.1168}$$

where the term:

$$L_j(x)=L(x_j)=\prod_{i=0,i\neq j}^{n}\frac{x-x_i}{x_j-x_i} \tag{14.1.1169}$$

is named "Lagrange polynomial interpolation coefficients".

Notice that if we use the Lagrange polynomial to approximate a known analytic and continuous function (hence with an infinite number of points), then obviously there may be an approximation error. This will be denoted:

$$P(x)-f(x)=\varepsilon(x) \tag{14.1.1170}$$

✍**Example:**

Let us approximate the function $y=f(x)=x\sin(2x+\pi/4)+1$ by a polynomial P_n of degree $n=3$, based on the following $n+1=4$ points:

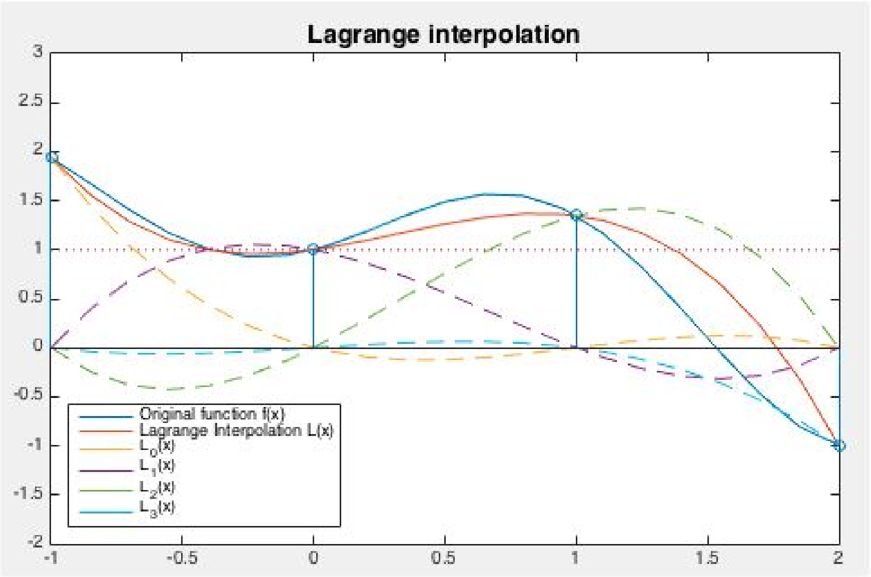
i	0	1	2	3
x_i	-1	0	1	2
$y_i=f(x_i)$	1.937	1.000	1.349	-0.995

Based on these points, we construct the Lagrange polynomials as the basis functions of the polynomial space:

$$\begin{aligned}
L_0(x)&=\dfrac{(x-0)(x-1)(x-2)}{-6}=\dfrac{x^3-3x^2+2x}{-6}\\
L_1(x)&=\dfrac{(x+1)(x-1)(x-2)}{2}=\dfrac{x^3-2x^2-x+2}{2}\\
L_2(x)&=\dfrac{(x+1)(x-0)(x-2)}{-2}=\dfrac{x^3-x^2-2x}{-2}\\
L_3(x)&=\dfrac{(x+1)(x-0)(x-1)}{6}=\dfrac{x^3-x}{6}
\end{aligned}$$

Note that indeed $L_0(x)+L_1(x)+L_2(x)+L_3(x)=1$. The interpolating polynomial can be obtained as a weighted sum of these basis functions:

$$P_3(x)=1.937L_0(x)+1.0L_1(x)+1.349L_2(x)-0.995L_3(x)=1.0+0.369x+0.643x^2-0.663x^3$$



14.1.13.6 Newton polynomials (divided differences)

As we have just seen, the Lagrange interpolation relies on the $n+1$ interpolation points:

$$\{x_i,\, y_i = f(x_i),\, i = 0,\dots,n\} \tag{14.1.1171}$$

all of which need to be available to calculate each of the basis polynomials $l_i(x)$.

The goodness of an approximation depends on the number of approximating points and also on their locations. One problem with the Lagrange interpolating polynomial is that we need n additions, $2n^2+2n$ subtractions, $2n^2+n-1$ multiplications, and $n+1$ divisions to evaluate $P(x)$ at a given point x. Even after all the denominators have been calculated once and for all we still need n additions, n^2+n subtractions, and n^2+n multiplications.

Furthermore, if additional points are to be used when they become available, all basis polynomials need to be recalculated. Another problem is that in practice, one may be uncertain as to how many interpolation points to use. So one may want to increase them over time and see whether the approximation gets better. In doing so, one would like to use the old approximation. It is not clear how to do that easily with the Lagrange form. In this sense, the Lagrange form is not incremental (plus it is also awkward to program).

In comparison, in the Newton interpolation, when more data points are to be used, additional basis polynomials and the corresponding coefficients can be calculated, while all existing basis polynomials and their coefficients remain unchanged. Due to the additional terms, the degree of interpolation polynomial is higher and the approximation error may be reduced (e.g., when interpolating higher order polynomials).

Specifically, the basis polynomials of the Newton interpolation are calculated as below:

$$n(x_0) = n_0(x) = 1, \quad n(x_i) = n_i(x) = \prod_{j=0}^{i-1}(x-x_j), \qquad i = 1,\dots,n \tag{14.1.1172}$$

and the Newton interpolating polynomial is constructed:

$$\begin{aligned}N_n(x) &= \sum_{i=0}^{n} c_i n_i(x) = c_0 + \sum_{i=1}^{n} c_i\left(\prod_{j=0}^{i-1}(x-x_j)\right)\\ &= c_0 + c_1(x-x_0) + c_2(x-x_0)(x-x_1) + \ldots + c_n\prod_{j=0}^{n-1}(x-x_j)\end{aligned} \tag{14.1.1173}$$

Note that the last data point (x_n, y_n) is not used in any of the basis polynomials, but it is used for calculating the last coefficient c_n, as shown below. For the nth degree polynomial $N_n(x)$ to pass all $n+1$ points $(x_i,\, y_i)$, $(i = 0,\dots,n)$, it needs to satisfy the following $n+1$ equations:

$$\begin{cases} y_0 = N_n(x_0) = c_0 \\ y_1 = N_n(x_1) = c_0 + c_1(x_1-x_0) \\ y_2 = N_n(x_2) = c_0 + c_1(x_2-x_0) + c_2(x_2-x_0)(x_2-x_1) \\ y_3 = N_n(x_3) = c_0 + c_1(x_3-x_0) + c_2(x_3-x_0)(x_3-x_1) + c_3(x_3-x_0)(x_3-x_1)(x_3-x_2) \\ \dots\dots\dots\dots \\ y_n = N_n(x_n) = c_0 + c_1(x_n-x_0) + c_2(x_n-x_0)(x_n-x_1) + \ldots + c_n\prod_{i=0}^{n-1}(x_n-x_i) \end{cases} \tag{14.1.1174}$$

The last relation is named the "Newton's divided-difference formula".

The relation can also be expressed in matrix form:

$$\begin{bmatrix} 1 & & & & & \\ 1 & x_1-x_0 & & & & \\ 1 & x_2-x_0 & (x_2-x_0)(x_2-x_1) & & & \\ 1 & x_3-x_0 & (x_3-x_0)(x_3-x_1) & (x_3-x_0)(x_3-x_1)(x_3-x_2) & & \\ \vdots & \vdots & & \vdots & \ddots & \\ 1 & x_n-x_0 & (x_n-x_0)(x_n-x_1) & (x_n-x_0)(x_n-x_1)(x_n-x_2) & \ldots & \prod\limits_{i=0}^{n-1}(x_n-x_i) \end{bmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} \tag{14.1.1175}$$

The $n+1$ coefficients $c_0,\ldots,c_n$ can be obtained by solving these $n+1$ equations in the triangular equation system progressively from top down:

$$\begin{aligned} c_0 &= y_0 = f\,(x_0) = f\,[x_0] \\ c_1 &= \dfrac{y_1-y_0}{x_1-x_0} = f\,[x_0,x_1] \\ c_2 &= \dfrac{\dfrac{y_2-y_0}{x_2-x_0}-\dfrac{y_1-y_0}{x_1-x_0}}{x_2-x_1} = f\,[x_0,x_1,x_2] \\ &= \dfrac{y_0}{(x_0-x_2)\,(x_0-x_1)} + \dfrac{y_1}{(x_1-x_0)\,(x_1-x_2)} + \dfrac{y_2}{(x_2-x_0)\,(x_2-x_1)} \\ c_3 &= \dfrac{\dfrac{\dfrac{y_3-y_0}{x_3-x_0}-\dfrac{y_1-y_0}{x_1-x_0}}{x_3-x_1}-\dfrac{\dfrac{y_2-y_0}{x_2-x_0}-\dfrac{y_1-y_0}{x_1-x_0}}{x_2-x_1}}{x_3-x_2} = f\,[x_0,\ldots,x_3] \\ &= \dfrac{y_0}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} + \dfrac{y_1}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} \\ &+ \dfrac{y_2}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} + \dfrac{y_3}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} \\ &\ldots \end{aligned} \tag{14.1.1176}$$

In general, we have the:

$$c_k = f[x_0,\ldots,x_k] = \sum_{j=0}^{k} \dfrac{f(x_j)}{\prod\limits_{i=0,\,i\neq j}^{k}(x_j-x_i)}, \qquad k=0,\ldots,n \tag{14.1.1177}$$

which is the expanded form of the kth divided differences $f[x_0,\ldots,x_k]$ of the first $k+1$ points. Now the Newton polynomial interpolation can be written as:

$$\boxed{N_n(x) = \sum_{i=0}^{n} c_i n_i(x) = \sum_{i=0}^{n} f[x_0,\ldots,x_i]n_i(x) = f[x_0] + \sum_{i=1}^{n} f[x_0,\ldots,x_i]\left(\prod_{j=0}^{i-1}(x-x_j)\right)} \tag{14.1.1178}$$

Due to the uniqueness of the polynomial interpolation (here "uniqueness" means only in the way that all previous polynomial interpolation methods are such that for some given points $\{(x_i,y_i)\}_{i=0}^n$ we just have $p_n(x_i)=y_i$... and nothing more!), this Newton interpolation polynomial is the same as that of the Lagrange and the power (i.e. Taylor) function interpolations:

$$N_n(x)=L_n(x)=P_n(x) \tag{14.1.1179}$$

They are the same nth degree polynomial but expressed in terms of different basis polynomials weighted by different coefficients.

We can now consider some important facts all related to the Newton polynomial interpolation.

- When an additional node point $(x_{n+1},\, y_{n+1})$ is available and to be used, all previous basis polynomials and their corresponding coefficients remain unchanged, we only need to obtain a new basis polynomial of degree $n+1$:

  $$n_{n+1}(x)=n_n(x)(x-x_n)=\prod_{i=0}^{n-1}(x-x_i)(x-x_n)=\prod_{i=0}^{n}(x-x_i)=\omega(x) \tag{14.1.1180}$$

  together with its coefficient, the $(n+1)$th divided difference $c_{n+1}=f[x_0,\ldots,x_n,x_{n+1}]$. The new interpolation polynomial of degree $n+1$ can then be obtained by including an extra term in the summation above:

  $$N_{n+1}(x)=\sum_{i=0}^{n+1}c_in_i(x)=\sum_{i=0}^{n}c_in_i(x)+c_{n+1}n_{n+1}(x)=N_n(x)+f[x_0,\ldots,x_n,x_{n+1}]\omega(x) \tag{14.1.1181}$$

  As $N_{n+1}(x)$ passes through the new point $(x_{n+1},\, y_{n+1})$, we have:

  $$f(x_{n+1})=N_{n+1}(x_{n+1})=N_n(x_{n+1})+f[x_0,\ldots,x_n,x_{n+1}]\omega(x) \tag{14.1.1182}$$

  But x_{n+1} is just an arbitrary point, we can replace it by x, and get:

  $$f(x)=N_n(x)+f[x_0,\ldots,x_n,x]\omega(x) \tag{14.1.1183}$$

- If all $n+1$ points $x_0=a\leq x_1\leq\ldots\leq x_{n-1}\leq x_n=b$ are equally spaced, i.e.:

  $$x_{i+1}-x_i=h=\frac{b-a}{n},\quad i=0,\ldots,n-1 \tag{14.1.1184}$$

  then Newton's divided difference interpolation can take a simpler form. For any point $x\in(a,\,b)$, we let $c=(x-x_0)/h$ so that it can be represented as $x=x_0+ch$, and $x-x_i=(x_0+ch)-(x_0+ih)=(c-i)h$, now the Newton polynomial can be written as:

  $$\begin{aligned}N(x)=N(x_0+ch)&=f[x_0]+\sum_{i=1}^{n}f[x_0,\ldots,x_i]\prod_{j=0}^{i-1}(x-x_j)\\&=f[x_0]+f[x_0,x_1]ch+f[x_0,x_1,x_2]c(c-1)h^2+\ldots+f[x_0,\cdots,x_n]c(c-1)\ldots(c-n+1)h^n\\&=\sum_{i=0}^{n}f[x_0,\ldots,x_i]c(c-1)(c-2)\ldots(c-i+1)h^i=\sum_{i=0}^{n}f[x_0,\ldots,x_i]\begin{pmatrix}c\\i\end{pmatrix}i!\,h^i\end{aligned} \tag{14.1.1185}$$

  where we used the binomial-coefficient notation:

  $$\begin{pmatrix}c\\i\end{pmatrix}=\frac{c!}{i!(c-i)!}=\frac{c(c-1)(c-2)\cdots(c-i+1)}{i!} \tag{14.1.1186}$$

✍**Example:**

Let us approximate the function $y = f(x) = x\ \sin(2x+\pi/4)+1$ by a polynomial of degree $n=3$, based on the following $n+1=4$ points:

i	0	1	2	3
x_i	-1	0	1	2
$y_i = f(x_i)$	1.937	1.000	1.349	-0.995

Based on $f[x_i] = f(x_i)$, $(i=0,\cdots,n)$, we can find all other divided differences recursively in tabular form as shown below. In general, $f[x_i,\cdots,x_j]$ can be found based on its left neighbour $f[x_{i+1},\cdots,x_j]$ and top-left neighbour $f[x_i,\cdots,x_{j-1}]$:

$$f[x_i,\cdots,x_j] = \frac{f[x_{i+1},\ldots,x_j] - f[x_i,\ldots,x_{j-1}]}{x_j - x_i}$$

x_i	0th	1st	2nd	3rd
$x_0 = -1$	$f[x_0] = 1.937$			
$x_1 = 0$	$f[x_1] = 1.000$	$f[x_0,x_1] = -0.937$		
$x_2 = 1$	$f[x_2] = 1.349$	$f[x_1,x_2] = 0.349$	$f[x_0,x_2] = 0.643$	
$x_3 = 2$	$f[x_3] = -0.995$	$f[x_2,x_3] = -2.343$	$f[x_1,x_3] = -1.346$	$f[x_0,x_3] = -0.663]$

The coefficients are the four divided differences along the diagonal:

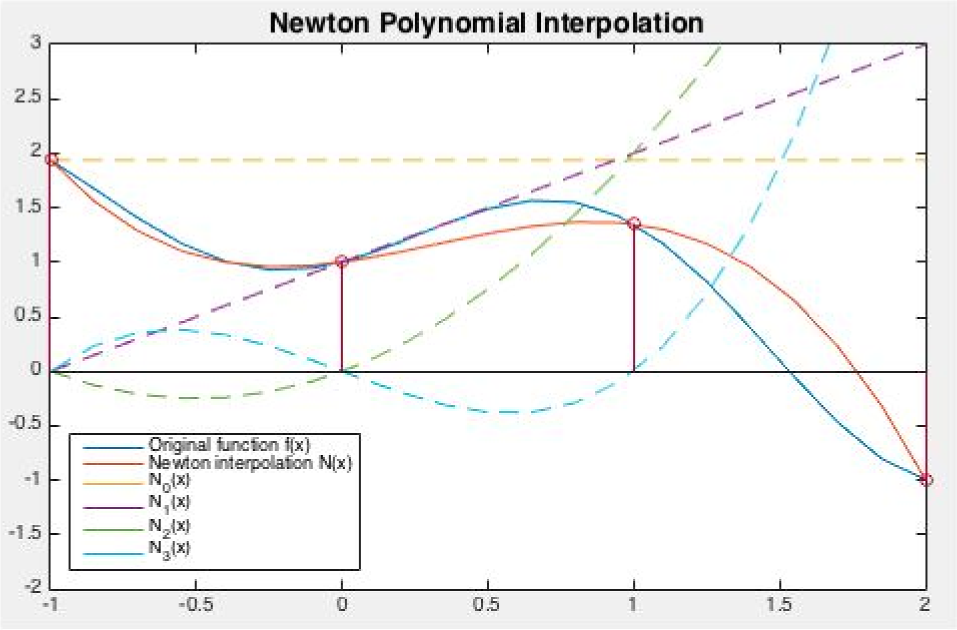
$$c_0 = f[x_0] = f(x_0) = y_0 = 1.937, c_1 = f[x_0,x_1] = -0.937, c_2 = f[x_0,x_1,x_2] = 0.643$$

and $c_3 = f[x_0,x_1,x_2,x_3] = -0.663$. Alternatively, they can also be represented in the expanded form:

$$c_i = f[x_0,\dots,x_i] = \sum_{j=0}^{n} \frac{f(x_j)}{\prod_{i=0,\,i\neq j}^{n}(x_j - x_i)}$$

Now the Newton interpolating polynomial can be obtained as:

$$\begin{aligned} N_3(x) &= \sum_{i=0}^{3} c_i n_i(x) = 1.937 - 0.937(x+1) + 0.643(x+1)(x-0) \\ &= -0.663(x+1)(x-0)(x-1) = 1 + 0.369x + 0.643x^2 - 0.663x^3 \end{aligned}$$



Using the Newton interpolating polynomials is then usually the best choice in comparison of Taylor, Vandermonde (i.e. collocation polynomial[32]) or Lagrange polynomials. It has the advantage that data pairs can be added and interpolated by merely adding one additional term to the previous interpolating polynomial. Under other restrictions the coefficients give information about the derivatives of a function being approximated as well as the error.

---

[32] Using the Vandermonde matrix is indeed not a very good method for any situation. The system is ill-conditioned and therefore the coefficients may be calculated very inaccurately. Also the amount of work is excessive!

14.1.13.7 Errors in Polynomial Interpolation

There are multiple ways to derive the approximation error of polynomial interpolation. In our personal and subjective point of view, all proofs known to us so far are not really convincing...

So let us introduce the error term in an engineer way rather than in a tricky mathematician one. For this, let us recall that during our Study of Taylor series, we have proved that the Lagrange remainder (see section Sequences and Series page 1722) was given by:

$$R_n(x)=f(x)-P_n(x)=\frac{f^{(n+1)}(z)}{(n+1)!}(x-x_0)^{n+1} \tag{14.1.1187}$$

For the case of polynomial where we have multiple points $x_0,x_1,\ldots,x_n$ rather than a unique one x_0, we can guess that the error may be written:

$$\varepsilon_n(x)=f(x)-P_n(x)=\frac{f^{(n+1)}(z)}{(n+1)!}\prod_{i=0}^{n}(x-x_i) \tag{14.1.1188}$$

In the field of polynomial interpolation (that we use the Taylor, Lagrange or Newton methods) that latter is often written:

$$\boxed{\varepsilon_n(x)=\frac{f^{(n+1)}(\xi)}{(n+1)!}\prod_{i=0}^{n}(x-x_i)} \tag{14.1.1189}$$

where ξ is some (unknown) point in the interval $[a,b]$. The precise location of this point depends on $\{x_i\}_{i=0}^{n}$. Here $f^{(n+1)}(\xi)$ is the $(n+1)$st derivative of $f(x)$ evaluated at the point $x=\xi_x$.

Therefore, for example (an example that will be useful to us when we will study numerical integration), the Lagrange polynomial given for recall by:

$$P(x)=\sum_{j=0}^{n}\left(\prod_{i=0,i\neq j}^{n}\frac{x-x_i}{x_j-x_i}\right)f(x_j)=\sum_{j=0}^{n}L(x_j)f(x_j) \tag{14.1.1190}$$

Will be written in it's full form as:

$$\boxed{P_n(x)=\sum_{j=0}^{n}\left(\prod_{i=0,i\neq j}^{n}\frac{x-x_i}{x_j-x_i}\right)f(x_j)+\frac{f^{(n+1)}(\xi)}{(n+1)!}\prod_{i=0}^{n}(x-x_i)=\sum_{j=0}^{n}L(x_j)f(x_j)+\varepsilon_n(x)} \tag{14.1.1191}$$

14.1.14 Roots search

Many equation encountered in practice or in theory cannot be solved by closed form or analytical methods. Consequently, only a numerical approach can be obtained in a finite number of operations.

The mathematician Évariste Galois has proved, in particular, that the polynomial equation $P_n(x)$ (see section Calculus page 1368) has no algebraic solution if $P_n(x)$ is of degree $n>4$.

There are numerous algorithms that gives the possibility to calculate the roots of equations of the type $f(x)=0$ with an almost arbitrary precision. We will see in this section only the main one.

Caution! The implementation of such algorithms need always at least an approximate knowledge (randomly or deterministic) of the searched value and also the behaviour (shape) of the function near the root. A table of the function values (see section Functional Analysis page 2207) or its graphical representation (see same section) can in simple cases help to acquire this prerequisites knowledge.

If the univariate equation to solve is under the form $g(x)=h(x)$, we plot the curves representing g and h. The roots of the equation $g(x)=h(x)$ being given by the abscissa of the intersection points of the both curves.

Remark

Before solving by numerical methods the equation $f(x)=0$, we have to check that the function f satisfy some constraints. For example, the function f has to be at least strictly monotone near the root that we will denote here by $\bar{x}$, when the Newton's method is applied. It is many times useful, even absolutely necessary, to determine an interval $[a,b]$ such that:

- f is continuous on $[a,b]$ of class $\mathcal{C}^1$
- $f(a)f(b)<0$
- $\exists\bar{x}$ unique, $\bar{x}\in[a,b]$, $f(r)=0$

14.1.14.1 Proportional parts methods

The implementation, on a computer, of this method is particularly simple. The conditions to be satisfied being only that in the interval $[a,b]$:

- f must be continuous on $[a,b]$ of class $\mathcal{C}^1$
- f must be monotone near the root r
- $f(a)f(b)<0$ to be sure that there is a root

As we already know, in a small interval, we can replace a curve by a straight line segment. There are several possible situations but here is a special one but that can be generalized easily to anything:

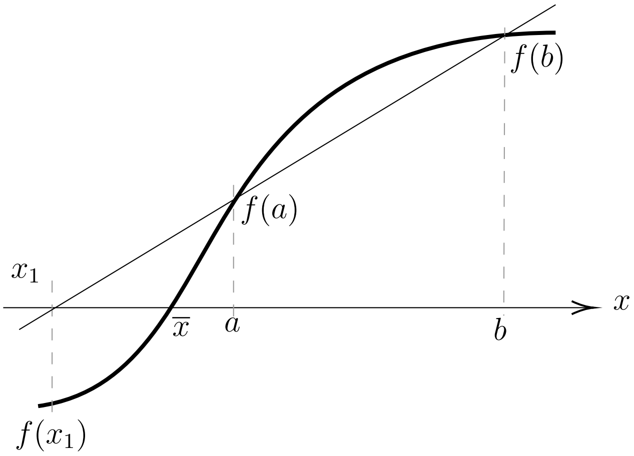


Figure 14.65: **Local approximation of a curve by a straight line segment**

In this figure, we get by using the theorems of Thales (see section Euclidean Geometry page 2610):

$$\frac{x_1-a}{f(a)}=\frac{x_1-b}{f(b)} \tag{14.1.1192}$$

therefore:

$$x_1=\frac{af(b)-bf(a)}{f(b)-f(a)}=a-(b-a)\frac{f(a)}{f(b)-f(a)} \tag{14.1.1193}$$

If $|f(a)| \ll |f(b)|$, we can neglect $f(a)$ at the denominator and it comes:

$$x_1 \cong a-(b-a)\frac{f(a)}{f(b)}=a-(b-a)k \tag{14.1.1194}$$

The algorithm consists therefore in performing the following steps:

1. We fix $\varepsilon>0$ as upper bound of the admissible error tolerance.
2. Choose a and b with opposite signs, calculate $f(a)$ and $f(b)$
3. We calculate $x_1=a-(b-a)k$. If $|f(x_1)|$ is small enough, we stop the calculation and we display x_1 and $f(x_1)$.
4. Otherwise we proceed as following:
   - We replace b by a and $f(b)$ by $f(a)$
   - We replace a by x_1 and $f(a)$ by $f(x_1)$

   and we go back to point (2).

In pseudocode (non-unique and not optimized):

Algorithm 7: Proportional Parts pseudocode algorithm

```
Data: $a$,$b$, $\varepsilon$ expression of $f$
Result: $x_1$
initialization;
$f(a)$,$f(b)$;
while $|f(x_1)| > \varepsilon$ do
    $x_1 = a-(b-a)\dfrac{f(a)}{f(b)}$;
    if $|f(x_1)| > \varepsilon$ then
        $b := a$;
        $f(b) := f(a)$;
        $a := x_1$;
        $f(a) := f(x_1)$;
    Display $x_1$;
```

Obviously this is not the best algorithm especially if there is no root or if the computer is slow. Then a good advice is to take as input a number of limited iterations that have to be done (we will see a more elaborated way to handle such situations in the next algorithm).

14.1.14.2 Secant method (Regula Falsi or False Position)

The "secant method" also named or "regula falsi" (for: regularly false) or also named "false position method" is still a root search algorithm to zero. It's almost exactly the same as the proportional parts methods at the difference we don't make the assumption $|f(a)| \ll |f(b)|$.

To introduce it, consider the following figure:

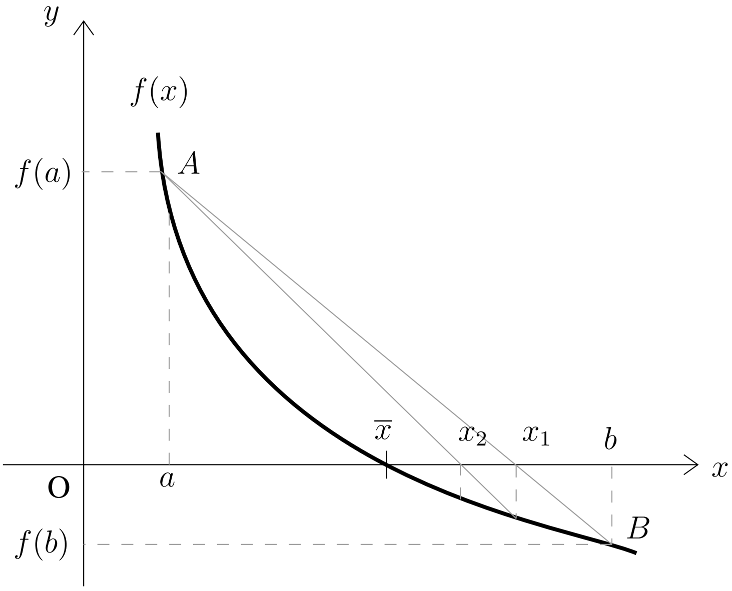


Figure 14.66: **Regula falsi method scheme**

The implementation, on a computer, of this method is particularly simple. The conditions to be satisfied being only that in the interval $[a, b]$:

- f must be continuous on $[a, b]$ of class $\mathcal{C}^1$
- f must be monotone near the root x^*
- $f(a)f(b) < 0$ to be sure that there is a root

If P_n is the point of coordinates $(x_n, 0)$, then the points (A, B, P_n) are aligned on the secant. The following proportion (application of Thales theorem) is then true:

$$\frac{a-b}{a-x_n}=\frac{f(a)-f(b)}{f(a)-0} \qquad (14.1.1195)$$

hence we deduce that:

$$x_n=\frac{af(b)-bf(a)}{f(b)-f(a)} \qquad (14.1.1196)$$

The algorithm consists therefore in performing the following steps:

1. We fix $\varepsilon > 0$ as upper bound of the admissible error tolerance.

2. Calculation of $x_n = \dfrac{af(b)-bf(a)}{f(b)-f(a)}$
3. Evaluation of $f(x_n)$
4. If $|f(x_n)|<\varepsilon$, the job is done. We have to display x_n
5. Otherwise we proceed as following:
    (a) we replace a by x_n if $f(x_n)f(a)>0$.
    (b) we replace b by x_n if $f(x_n)f(b)>0$ or $f(x_n)f(a)<0$.
    (c) we go back in (2)

Once again, the previous step (4) imposes the condition for stopping the calculations. Sometimes it is better to choose another criterion calculation ending. It requires the calculated solution to be contained in an interval of length equation containing the root x^*. This test is enunciate as follows:

4'. If $|b-a|<\varepsilon$, the job is finished and $x_n = \dfrac{af(b)-bf(a)}{f(b)-f(a)}$ is displayed. It is for sure obvious that $|x-\bar{x}|<\varepsilon/2$

In pseudocode (non-unique and not optimized):

Algorithm 8: Regula-Falsi pseudocode algorithm

```
Data: $a$,$b$, $\varepsilon$ expression of $f$
Result: $\bar{x}$
initialization;
$x_n = \dfrac{af(b)-bf(a)}{f(b)-f(a)}$;
while $|f(x)|>\varepsilon$ do
    if $f(x)f(a)>0$ then
        $a := x_n$;
    else if $f(x_n)f(b)>0 \vee f(x_n)f(a)<0$ then
        $b := x$;
    $x_n = \dfrac{af(b)-bf(a)}{f(b)-f(a)}$;
Display $x_n, f(x_n)$;
```

Notice that the initial above relation can also be rewritten:

$$x_n = a - f(a)\frac{a-b}{f(a)-f(b)} \tag{14.1.1197}$$

If we change a little bit the notation we get:

$$x = x_1 - f(x_1)\frac{x_1-x_0}{f(x_1)-f(x_0)} \tag{14.1.1198}$$

often also denoted:

$$x_{n+1}=x_n-\frac{f(x_n)}{f'(x_n)} \tag{14.1.1199}$$

We then use this new value of x in the prior previous relation as x_2 and repeat the process using x_1 and x_2 instead of x_0 and x_1. We continue this process, solving for x_3, x_4, etc., until we reach a sufficiently high level of precision (a sufficiently small difference between x_n and x_n-1):

$$\begin{aligned}
&x_2=x_1-f(x_1)\frac{x_1-x_0}{f(x_1)-f(x_0)}\\
&x_3=x_2-f(x_2)\frac{x_2-x_1}{f(x_2)-f(x_1)}\\
&\ldots\\
&x_n=x_{n-1}-f(x_{n-1})\frac{x_{n-1}-x_{n-2}}{f(x_{n-1})-f(x_{n-2})}
\end{aligned} \tag{14.1.1200}$$

This again an application of the "fixed-point iteration" (related to the theorem of the same name proved in the section of Sequences and Series page 1780).

14.1.14.3 Bisection method (interval-halving)

The bisection method in mathematics is also a root-finding method that repeatedly bisects an interval and then selects a subinterval in which a root must lie for further processing. It is a very simple and robust method, but it is also relatively slow. Because of this, it is often used to obtain a rough approximation to a solution which is then used as a starting point for more rapidly converging methods. The method is also named the "interval halving method", the "binary search method", or the "dichotomy method".

The method is applicable for numerically solving the equation $f(x)=0$ for the real variable x, where f is a continuous function defined on an interval $[a,b]$ and where $f(a)$ and $f(b)$ have opposite signs such that $f(a)f(b)<0$. In this case a and b are also said to bracket a root since, by the intermediate value theorem, the continuous function f must have at least one root in the interval $[a,b]$.

At each step the method divides the interval in two by computing the midpoint $c=(a+b)/2$ of the interval and the value of the function $f(c)$ at that point. Unless c is itself a root (which is very unlikely, but possible) there are now only two possibilities: either $f(a)$ and $f(c)$ have opposite signs and bracket a root, or $f(c)$ and $f(b)$ have opposite signs and bracket a root. The method selects the subinterval that is guaranteed to be a bracket as the new interval to be used in the next step. In this way an interval that contains a zero of f is reduced in width by 50% at each step. The process is continued until the interval is sufficiently small.

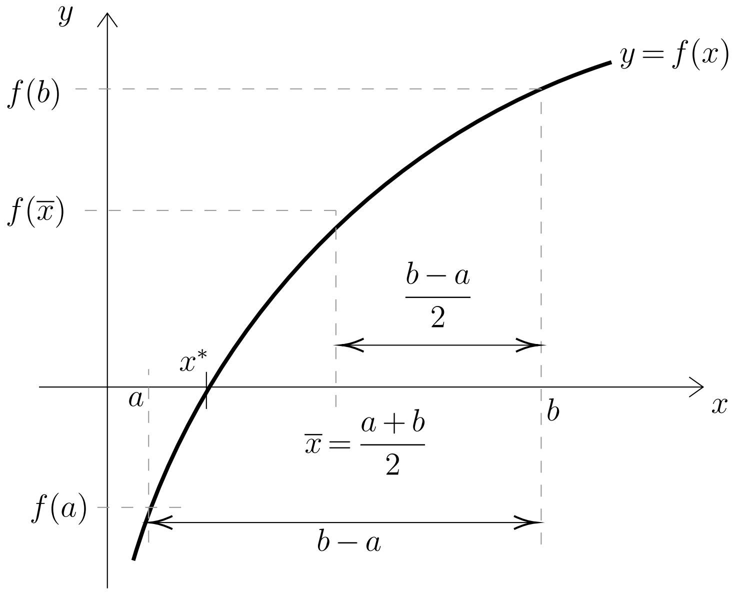


Figure 14.67: **Bisection scheme**

Explicitly, if $f(a)$ and $f(c)$ have opposite signs, then the method sets c as the new value for b, and if $f(b)$ and $f(c)$ have opposite signs then the method sets c as the new a. (If $f(c)=0$ then c may be taken as the solution and the process stops.) In both cases, the new $f(a)$ and $f(b)$ have opposite signs, so the method is applicable to this smaller interval.

The implementation, on a computer, of this method is particularly simple. The conditions to be satisfied being only that in the interval $[a,b]$:

- f must be continuous on $[a,b]$ of class $\mathcal{C}^1$
- f must be monotone near the root $\bar{x}$
- $f(a)f(b)<0$ to be sure that there is a root

The algorithm consists therefore in performing the following steps:

1. We fix $\varepsilon>0$ as upper bound of the admissible error tolerance.
2. We calculate $x=(a+b)/2$
3. We evaluate $f(x)$
4. If $|f(x)|<\varepsilon$ then the job is done, we have to display x and $f(x)$
5. Otherwise we proceed as following
   (a) we replace a by x if $f(x)f(a)>0$.
   (b) we replace b by x if $f(x)f(b)>0$ or $f(x)f(a)<0$.
   (c) we go back in (2)

The previous step (4) imposes the condition for stopping the calculations. Sometimes it is better to choose another criterion calculation ending. It requires the calculated solution to be contained in an interval of length equation containing the root x^*. This test would then be enunciate as follows:

4'. If $|b-a|<\varepsilon$, the job is finished and $x=(a+b)/2$ is displayed. It is for sure obvious that $|x-x^*|<\varepsilon/2$

In pseudocode (non-unique and not optimized):

Algorithm 9: Proportional Parts bisection pseudocode algorithm

```
Data: $a$,$b$, $\varepsilon$ expression of $f$
Result: $x^*$
initialization;
$x=(a+b)/2$;
while $|f(x)|>\varepsilon$ do
    if $f(x)f(a)>0$ then
        $a:=x$;
    else if $f(x)f(b)>0 \vee f(x)f(a)<0$ then
        $b:=x$;
    $x=(a+b)/2$;
Display $x$, $f(x)$;
```

The equivalent Maple™ 4.00b code is given by:

```
>zero:=proc(f,a,b,pre) local M;
M:=f((a+b)/2);
if abs(M)<pre then
RETURN((a+b)/2)
elif M>0 then
zero(f,a,(a+b)/2,pre)
else zero(f,(a+b)/2,b,pre)
fi
end:
```

14.1.14.4 Newton's method

A few years after its discovery of the theory of gravitation, Newton developed a special technique to fast compute the solutions of any equation. This "supernaturally" fast convergence has been used to prove some of the most significant theoretical results of the 120th century (holocene calendar): the Kolmogorov stability theorem, the isometric embedding theorem of Nash ... Alone, this technique transcends the distinction between pure mathematics and applied mathematics!!!

To study the "Newton's method" (also named "Newton-Raphson method[33]" or "Newton approximation scheme") in the plane (hence in the univariate case), consider the following figure:

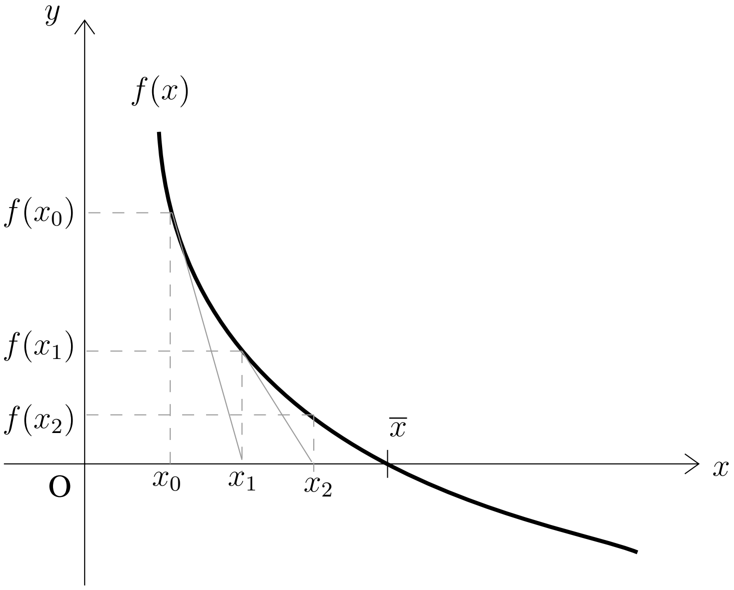


Figure 14.68: Newton's method in the plane

If x_0 is an approximation of the root $\bar{x}$, we notice that x_1 is as better one. The point x_1 is the intersection of the tangent to the curve $(x_0, f(x_0))$ with the x-axis. The point x_2 is even a better approximate of $\bar{x}$, the point x_2 is obtained in the same manner as x_1 but from $(x_1, f(x_1))$.

We may ask ourselves at this point what is the difference between Newton-Raphson and secant method on the basis of geometric interpretation? The two methods are almost the same, from a geometric perspective. The difference is that Newton's method uses a line that is tangent to one point, while the secant method uses a line that is secant to two points.

To use this technique, remember that if we take a function f that is differentiable on x_0, then we can rewrite it in the form (see section Sequences and Series page 1696):

$$f(x)=f(x_0)+f'(x_0)(x-x_0)+\mathcal{O}(x-x_0) \tag{14.1.1201}$$

---

[33] Newton's method was first published in 11685 and in 11690 (holocene calendar), Joseph Raphson published a simplified description. Raphson also applied the method only to polynomials, but he avoided Newton's tedious rewriting process by extracting each successive correction from the original polynomial. This allowed him to derive a reusable iterative expression for each problem. Finally, in 11740, (holocene calendar) Thomas Simpson described Newton's method as an iterative method for solving general non-linear equations using calculus. In the same publication, Simpson also gives the generalization to systems of two equations and notes that Newton's method can be used for solving optimization problems by setting the gradient to zero.

where $f'(x_0)$ is the derivative of f in x_0 and $\mathcal{O}(x-x_0)$ is a function that tends to 0 as $(x-x_0)^n$ when $n \geq 2$ when x tends to x_0 (this is as we know a corrective term for the superior orders of the Taylor series).

Applying this result to solve $f(x)=0$, we get:

$$0=f(\bar{x})=f(x_0)+f'(x_0)(\bar{x}-x_0)+\mathcal{O}(\bar{x}-x_0) \tag{14.1.1202}$$

The function does not permit the resolution of this equation relatively to $\bar{x}$. By neglecting this term we get obviously:

$$0=f(\bar{x})=f(x_0)+f'(x_0)(\bar{x}-x_0) \tag{14.1.1203}$$

This is just the expression of the discrete derivative. Indeed:

$$f'(x_0)=\frac{f(\bar{x})-f(x_0)}{\bar{x}-x_0} \tag{14.1.1204}$$

The latter relation is easily resolved with respect to $\bar{x}$. To see this, let us begin by putting $\bar{x}=x_1$:

$$\boxed{x_1=x_0-\frac{f(x_0)}{f'(x_0)}} \tag{14.1.1205}$$

But x_1 does not satisfy, in general, the equality $f(x_1)=0$. But as we have already pointed it out, $|f(x_1)|$ is smaller than $|f(x_0)|$ if the function f satisfies some conditions.

Newton's method consists in replacing the relation:

$$0=f(\bar{x})=f(x_0)+f'(x_0)(\bar{x}-x_0) \tag{14.1.1206}$$

by:

$$0=f(x_n)+f'(x_n)(x_{n+1}-x_n) \tag{14.1.1207}$$

and to iteratively solve this relation.

The following conditions are sufficient to ensure the convergence of the method in an interval $[a,b]$ including x_0 and $\bar{x}$:

1. The function is twice differentiable (i.e. of class $\mathcal{C}^2$)
2. The derivative f' does not vanish (monotony)
3. The second derivative f'' is continuous and does not vanish (no inflection point)

Remark

It is often sufficient to check the conditions (1) and (2) for the process to converge.

The condition (2) is obvious, indeed if $f'(x)=0$ the iteration can lead to a calculation overflow (singularity).

The third condition (3) is less obvious, but the following figure shows a case of non-convergence. In this case, the process loops calculating alternately x_i and x_j:

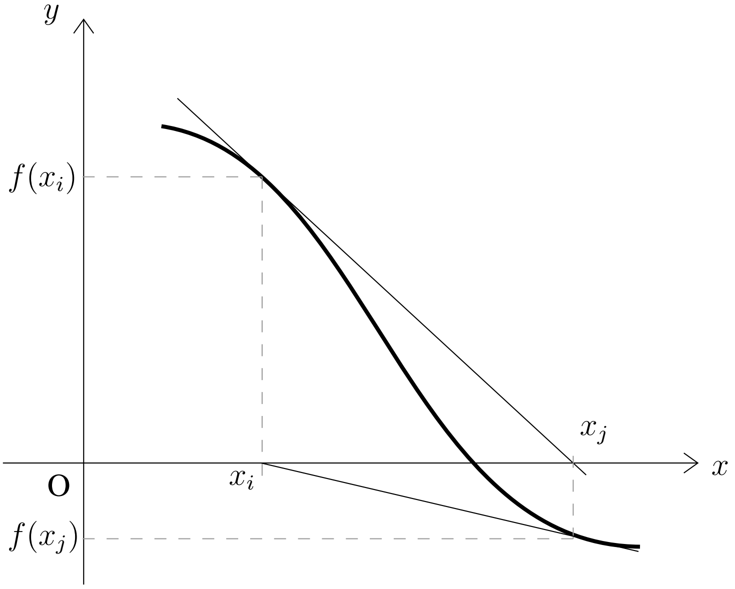


Figure 14.69: **Non-convergence example of Newton's method**

✍Example:

Consider the problem of finding the positive number x with $\cos(x)=x^3$. We can rephrase that as finding the zero of $f(x)=\cos(x)-x^3$. We have $f'(x)=-\sin(x)-3x^2$. Since $\cos(x)\leq 1$ for all x and $x^3>1$ for $x>1$, we know that our solution lies between 0 and 1. We try a starting value of $x_0=0.5$. (Note that a starting value of 0 will lead to an undefined result, showing the importance of using a starting point that is close to the solution!).

$$\begin{array}{rcccccl}
x_1 & = & x_0-\dfrac{f(x_0)}{f'(x_0)} & = & 0.5-\dfrac{\cos(0.5)-(0.5)^3}{-\sin(0.5)-3(0.5)^2} & = & 1.112141637097\\
x_2 & = & x_1-\dfrac{f(x_1)}{f'(x_1)} & = & \vdots & = & \underline{0.}909672693736\\
x_3 & = & \vdots & = & \vdots & = & \underline{0.86}7263818209\\
x_4 & = & \vdots & = & \vdots & = & \underline{0.865477}135298\\
x_5 & = & \vdots & = & \vdots & = & \underline{0.8654740331}11\\
x_6 & = & \vdots & = & \vdots & = & \underline{0.865474033102}
\end{array}$$

The correct digits are underlined in the above example. In particular, x_6 is correct to the number of decimal places given. We see that the number of correct digits after the decimal point increases from 2 (for x_3) to 5 and 10, illustrating the quadratic convergence.

Here is the corresponding pseudocode:

Algorithm 10: Newton's Method

```
Input: Initial guess $x_0$, $f$
Input: number of steps $x_{\max}$
$n \leftarrow 1$;
do
    $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$;
    $n = n+1$;
while $n \leq x_{\max}$;
```

If the function f is given analytically, its derivative can be determined analytically. But in many cases it is advisable or even necessary to replace $f'(x_n)$ by the differential quotient:

$$f'(x_n) \cong \frac{f(x_n+h)-f(x_n)}{h} \tag{14.1.1208}$$

where h should be chosen as small enough so that the difference:

$$\left| f'(x_n) - \frac{f(x_n+h)-f(x_n)}{h} \right| \tag{14.1.1209}$$

is also small enough.

The iteration is then written:

$$x_{n+1} = x_n - h\frac{f(x_0)}{f(x_n+h)-f(x_n)} \tag{14.1.1210}$$

If the resolution method is converging, the gap between x_{n+1} and $\bar{x}$ decreases at each iteration. This is ensured, for example, if the interval $[a,b]$ containing x_{n+1}, sees its length decreasing at each step.

Theorem 14.5. Newton's method is interesting because the convergence is quadratic:

$$|x_{n+1} - \bar{x}| < \alpha |x_n - \bar{x}|^2 \tag{14.1.1211}$$

while the convergence of other methods is linear such that:

$$|x_{n+1} - \bar{x}| < \alpha |x_n - \bar{x}| \tag{14.1.1212}$$

Let us consider, for example, the method of bisection seen previously. At each iteration the length of the interval $[a,b]$ is halved. This ensures us that the gap $|x_{n+1}-\bar{x}|$ is halved at each step of the calculation:

$$|x_{n+1} - \bar{x}| < 0.5 |x_n - \bar{x}| \tag{14.1.1213}$$

Proof 14.5.1. To prove the quadratic convergence of the Newton's method, we have to make use of the limited Taylor series of f and f' in the neighbourhood of $\bar{x}$:

$$\begin{aligned} f(x_n) &\cong \underbrace{f(\bar{x})}_{=0} + \frac{1}{1!}(x_n-\bar{x})f'(\bar{x}) + \frac{1}{2!}(x_n-\bar{x})^2 f''(\bar{x}) \\ f'(x_n) &= f'(\bar{x}) + \frac{1}{1!}(x_n-\bar{x})f''(\bar{x}) \end{aligned} \tag{14.1.1214}$$

But:

$$x_{n+1}=x_n-\frac{f(x_n)}{f'(x_n)} \tag{14.1.1215}$$

therefore:

$$x_{n+1}=x_n-\frac{\frac{1}{1!}(x_n-\bar{x})f'(\bar{x})+\frac{1}{2!}(x_n-\bar{x})^2f''(\bar{x})}{f'(\bar{x})+\frac{1}{1!}(x_n-\bar{x})f''(\bar{x})} \tag{14.1.1216}$$

Subtracting $\bar{x}$ left and right of the equality and putting the two terms of the second member to the same denominator, we get:

$$x_{n+1}-\bar{x}\cong(x_n-\bar{x})^2\frac{0.5f''(\bar{x})}{f'(\bar{x})+(x_n-\bar{x})f''(\bar{x})} \tag{14.1.1217}$$

and when $x_n-\bar{x}$ is small enough, the denominator can be simplified.

$$x_{n+1}-\bar{x}\cong(x_n-\bar{x})^2\frac{f''(\bar{x})}{f'(\bar{x})} \tag{14.1.1218}$$

which shows that convergence is quadratic.

■ Q.E.D.

✍**Example:**

Here is an application with Maple™ 4.00b of this method:

```
 >with(plots):  with(plottools):
>f:=x->exp(x)*x^2-36;
>D(f)(x);
>x[0]:=3;
>n:=7;
>g:=x->f(x[i-1])+D(f)(x[i-1])*(x-x[i-1]);
>for i from 1 by 1 to n do;
>x[i]:=evalf(solve(g(x)=0,x));
>od;
>lines:=:
>for i from 1 by 1 to n do;
>lines:=lines union {line([x[i-1],0],[x[i-1],f(x[i-1])],color=green),
line([x[i-1],f(x[i-1])],[x[i],0],color=green)};
>od:
>display(plot(f(x),x=2..3.01) union lines);
```

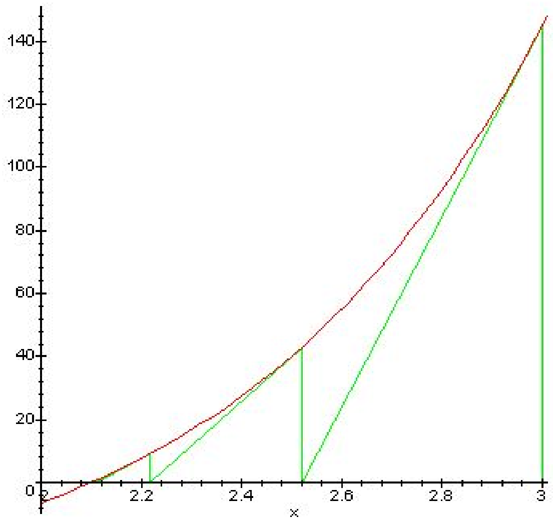


Figure 14.70: **Maple™ 4.00b application of Newton's method**

Now let us prove that Newton's method is reduced to the Babylonian method seen earlier above at page 6295! Indeed. we want to solve $\sqrt{A}$, i.e. looking for a root of $f(x)=x^2-A$. Then:

$$x_{n+1}=x_n-\frac{f(x_n)}{f'(x_n)}=x_n-\frac{x_n^2-A}{2x_n}=\frac{2x_n^2}{2x_n}-\frac{x_n^2-A}{2x_n}=\frac{2x_n^2-(x_n^2-A)}{2x_n}=\frac{1}{2}\left(x_n+\frac{A}{x_n}\right) \tag{14.1.1219}$$

That's it!

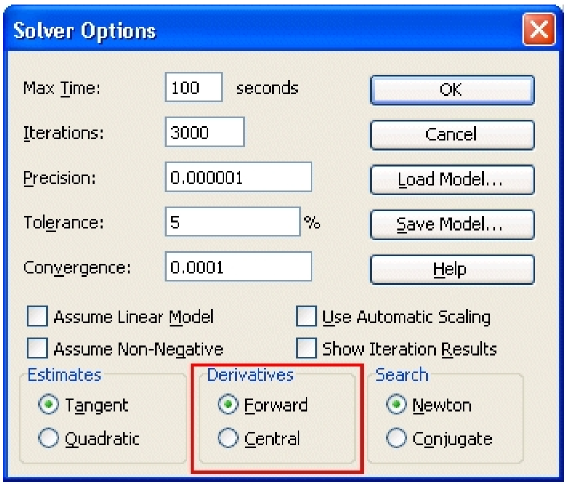
Remark

Let us finally communicate that we will study the Newton's method with several variables during our study of non-linear optimization. It is an pedagogical choice that seemed to us the best subjectively the best one.

14.1.15 Numerical Differentiation

In numerical analysis, numerical differentiation describes algorithms for estimating the derivative of a mathematical function or function subroutine using values of the function and perhaps other knowledge about the function.

Many modelling techniques or numerical resolution technique that we will see further use derivatives as for example the search of optimums (see further below), the finite element methods (see also further below). For example, to name the most famous case, the solver of Microsoft® Excel® 2007 and earlier offers some of the most elementary numerical derivatives that we will study here and reuse further:



Max Time: 100 seconds

Iterations: 3000

Precision: 0.000001

Tolerance: 5 %

Convergence: 0.0001

☐ Assume Linear Model ☐ Use Automatic Scaling

☐ Assume Non-Negative ☐ Show Iteration Results

Estimates: ● Tangent ○ Quadratic

Derivatives: ● Forward ○ Central

Search: ● Newton ○ Conjugate

Figure 14.71: **Screenshot of Microsoft® Excel® 2003 Solver**

To allow a computer processing, the various derivatives present in many algorithms must be approximated numerically. To do this, we use in the most basic case the principle of centered finite difference which is based on the following Taylor series expansion (see section Sequences and Series page 1696):

$$f(x)=\sum_{k=0}^{n}\frac{1}{k!}\frac{\mathrm{d}^k f(x_0)}{\mathrm{d}x^k}(x-x_0)^k=\sum_{k=0}^{n}\frac{\mathrm{d}^{(k)}f(x_0)}{k!}(x-x_0)^k \tag{14.1.1220}$$

On the base of this principle we get then the following development of the second order:

$$\begin{aligned}f\left(x_0+\frac{\mathrm{d}_x}{2}\right)&=f(x_0)+\frac{\mathrm{d}_x}{2}f'(x_0)+\frac{1}{2}\left(\frac{\mathrm{d}_x}{2}\right)^2 f''(x_0)+\ldots\\ f\left(x_0-\frac{\mathrm{d}_x}{2}\right)&=f(x_0)-\frac{\mathrm{d}_x}{2}f'(x_0)+\frac{1}{2}\left(\frac{\mathrm{d}_x}{2}\right)^2 f''(x_0)-\ldots\end{aligned} \tag{14.1.1221}$$

It comes then when neglecting the higher order terms and subtracting and simplifying the two series above:

$$f'(x_0)=\frac{f\left(x_0+\dfrac{\mathrm{d}_x}{2}\right)-f\left(x_0-\dfrac{\mathrm{d}_x}{2}\right)}{\mathrm{d}_x}+\varepsilon(\mathrm{d}_x^2) \tag{14.1.1222}$$

Relation that we name "first centered derivative with tangent estimate" (because we neglect all non-linear terms) or also "symmetric difference quotient". We also find often this latter relation in the following equivalent form:

$$\boxed{f'(x_0)=\frac{f(x_0+h)-f(x_0-h)}{2h}+\mathcal{O}(h^2)} \tag{14.1.1223}$$

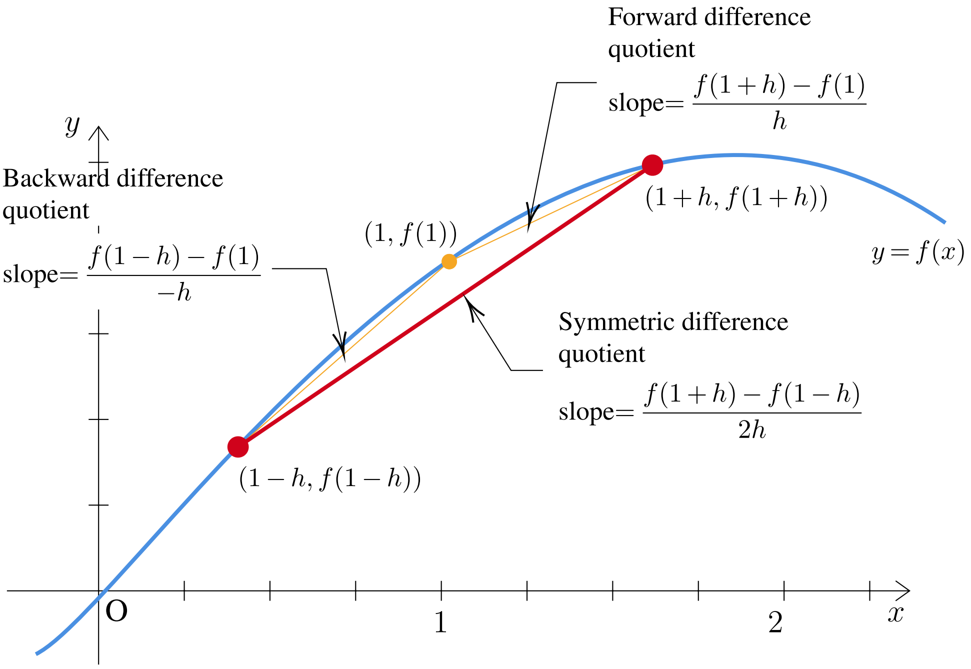
Now let us see what we name the "right first derivative" also named "forward derivatives" (or "forward difference"), which consist simply in the application of the following intuitive algorithm:

$$\boxed{f'(x_i)=\frac{f(x_{i+1})-f(x_i)}{x_{i+1}-x_i}} \tag{14.1.1224}$$

and incidentally we can also define the "left first derivative" also named "backward derivative" (or "backward difference"):

$$\boxed{f'(x_i)=\frac{f(x_i)-f(x_{i-1})}{x_i-x_{i-1}}} \tag{14.1.1225}$$

We see therefore the central derivatives require more calculations but are also more accurate. This below figure gives a quite summary of the previous relations with a special case:



Forward difference quotient
slope= $\frac{f(1+h)-f(1)}{h}$
y
Backward difference quotient
slope= $\frac{f(1-h)-f(1)}{-h}$
$(1,f(1))$
$(1+h,f(1+h))$
$y=f(x)$
Symmetric difference quotient
slope= $\frac{f(1+h)-f(1-h)}{2h}$
$(1-h,f(1-h))$
O
1
2
x

Figure 14.72: Forward/Backward derivatives illustration

We can also develop more elaborate relations through Taylor expansions with superior orders, do averages between different methods and so on... it's quite endless...

14.1.16 Numerical Integration

In numerical analysis, "numerical integration" constitutes a broad family of algorithms for calculating the numerical value of a definite integral, and by extension, the term is also sometimes used to describe the numerical solution of differential equations. This subsection focuses on calculation of definite integrals.

The basic problem in numerical integration is to compute an approximate solution to a definite integral overt the interval $[a,b]$ by evaluating $f(x)$ at a finite number of sample points:

$$\int\limits_a^b f(x)\mathrm{d}x \tag{14.1.1226}$$

to a given degree of accuracy (with an error $\varepsilon(f)$. If $f(x)$ is a smooth function integrated over a small number of dimensions, and the domain of integration is bounded, there are many methods for approximating the integral to the desired precision.

Definition (#14.1.564): Suppose that $a=x_0<x_1<\ldots<x_n=b$. A relation of the form:

$$A(f)\cong\sum_{i=0}^{n} w_i f(x_i)=w_0 f(x_0)+w_1 f(x_1)+w_2 f(x_2)+\ldots+w_n f(x_n) \tag{14.1.1227}$$

with the property that:

$$\int\limits_a^b f(x)\mathrm{d}x=A(f)+\varepsilon(f) \tag{14.1.1228}$$

is named a "numerical integration" or "quadrature formula[34]". The term $\varepsilon(f)$ is named the "truncation error for integration". The values $\{x_i\}_{i=0}^{n}$ are named the quadrature nodes and $\{w_i\}_{i=0}^{n}$ are named the "weights".

Depending on the application, the nodes $\{x_i\}_{i=0}^{n}$ are chosen in various ways. For the Trapezoidal Rule, Simpson's Rule, and Boole's Rule, the nodes are chosen to be equally spaced. For Gauss-Legendre quadrature, the nodes are chosen to be the zeros of certain Legendre polynomials. When the integration formula is used to develop a predictor formula for differential equations, all the nodes are chosen less than b. For all applications, it is necessary to know something about the accuracy of the numerical solution.

Let us consider the following figure:

---

[34] In mathematics, quadrature is a historical term which means the process of determining area. This term is still used nowadays in the context of differential equations, where "solving an equation by quadrature" means expressing its solution in terms of integrals.

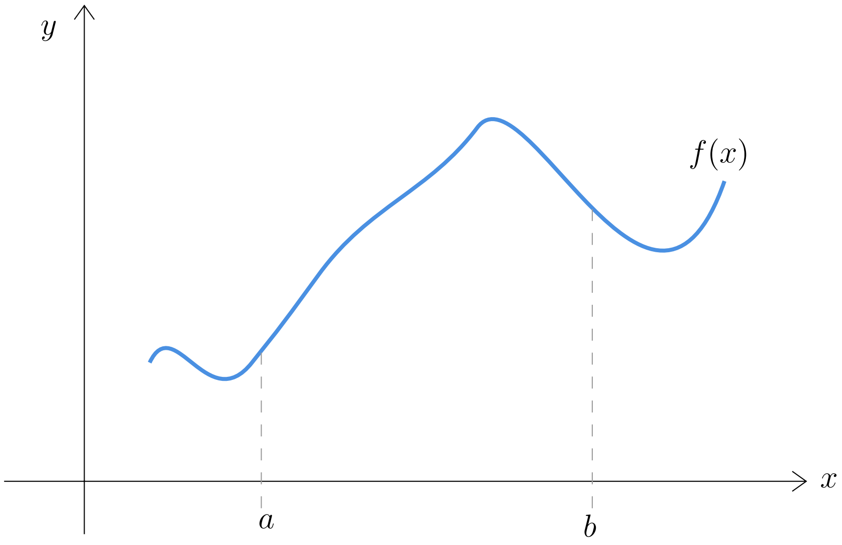


Figure 14.73: **Illustration of an interval under a curve**

We would like to calculate the area between the x axis, the curve f and the straight vertical lines of equations $x=a$ and $x=b$. We assume in this case that the function f is with positive values:

$$f(x)\geq 0 \quad \forall x\in[a,b] \tag{14.1.1229}$$

This problem, in its generality, is difficult or impossible to solve analytically in the most general cases. Below we will see some mainstream numerical methods for the approximate calculation of this area in increasing complexity order (sometimes these methods are used in corporations by employees who have only spreadsheets softwares like Microsoft® Excel® or OpenOffice Calc to calculate integrals...):

- Newton-Cotes formulas: In this case, we obtain methods for numerical integration which can be derived from the Lagrange interpolating method. Alternatively the formulas can also be derived from Taylor expansion. The idea is similar to the way we obtain numerical differentiation schemes. We can easily derive not just integration formulas but also their errors using this technique. The schemes which we develop here will be based on the assumption of equidistant points.

- Composite, Newton - Cotes formulas (open and closed): These methods are composite since they repeatedly apply the simple formulas derived previously to cover longer intervals. This idea allows for piecewise estimates of the integral thus improving the error of our integration (we will also assume equidistant nodes in our introduction of these methods).

- Romberg Integration: This method allows us to improve the error of our integration methods by doing minimal extra work. The idea is based on the Richardson extrapolation (actually we don't present this method here).

- : Adaptive Integration: Here we are free to choose the points over which we calculate the numerical integral of $f(x)$ so as to minimize our error. Adaptive integration does not therefore require equidistant nodes. Thus if the function is not very smooth at some interval the step size h of the numerical integration method decreases to make sure we do not accumulate too much error in our calculation (actually we don't present this method here).

- Gaussian Integration: We explore methods which can achieve optimal error reduction provided we place the nodes at specific locations. Computing the best weights for our numerical quadratures guarantees optimal approximation of our integral (actually we don't present this method here).
- Monte Carlo Integration: Use random number generation and a ratio of the target random points relatively to all random generated points. At this method employs a non-deterministic approach: each realization provides a different outcome! There are different methods to perform a Monte Carlo integration, such as uniform sampling, stratified sampling, importance sampling, sequential Monte Carlo (also known as a particle filter), and mean field particle methods (see page 6708).

14.1.16.1 Rectangles method

The "rectangle methods" computes an approximation to a definite integral, made by finding the area of a collection of rectangles whose heights are determined by the values of the function by different approach.

We divide the interval $[a,b]$ into n subintervals which bounds are the x_i. The lengths of these subintervals are $h_i = x_{i+1} - x_i$. We build rectangles which sides are h_i and in the case of the "" the height is $f(x_i)$.

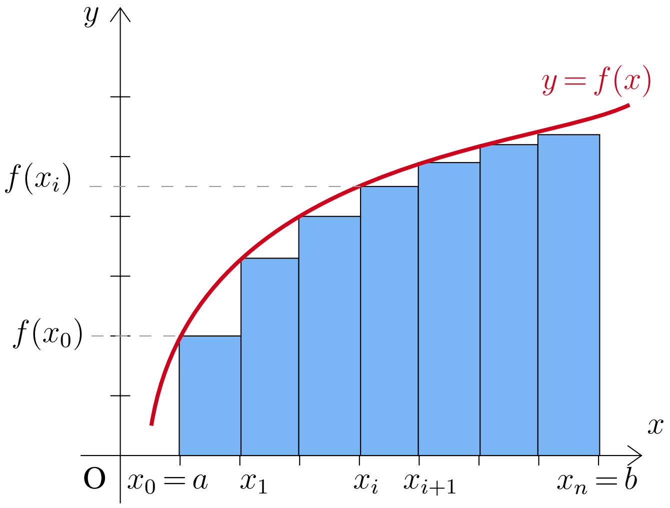


Figure 14.74: Approach of the area under a curve by lower left rectangles method

The area of these rectangles is:

$$A_G = \sum_{i=0}^{n-1} (x_{i+1} - x_i) f(x_i) = \sum_{i=0}^{n-1} h_i f(x_i) \tag{14.1.1230}$$

If the h_i are small enough, A_G is a good approximation of the sought approached area by the left method.

We can start this exercise again by choosing h_i and $f(x_{i+1})$ as sides of the rectangles (so the approach is named the "right rectangle method"). We then get:

$$A_D = \sum_{i=0}^{n-1}(x_{i+1}-x_i)f(x_{i+1}) = \sum_{i=0}^{n-1} h_i f(x_{i+1}) \tag{14.1.1231}$$

The correspondent figure is therefore the following:

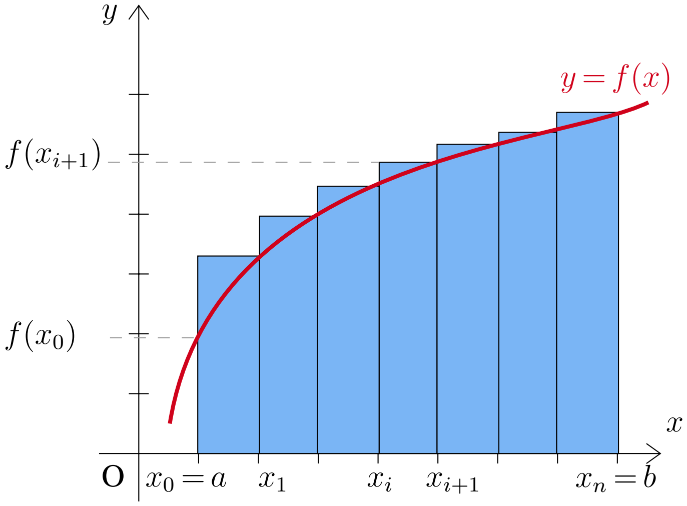


Figure 14.75: Approach of the area under a curve by upper right rectangles method

Again, the area of these rectangles approaches the area searched. To simplify computer code, it is useful to choose identical length intervals:

$$|x_{i+1}-x_i| = h \qquad \forall i \tag{14.1.1232}$$

If we have n rectangles, h is then equal to $(b-a)/n$. The areas A_D and A_G become:

$$A_D = h\sum_{i=0}^{n-1} f(x_{i+1}) \qquad A_G = h\sum_{i=0}^{n-1} f(x_i) \tag{14.1.1233}$$

We can also mixed the both method above as we will illustrate it further below in the figure that will summarize the for more common methods.

14.1.16.2 Trapezoidal method

In the purpose to increase the accuracy, it is possible to calculate:

$$A_T = \frac{A_D + A_G}{2} \tag{14.1.1234}$$

In the case where all the intervals are of equal length, A_T is equal to:

$$\begin{aligned} A_T &= \frac{1}{2}h\left(\sum_{i=0}^{n-1} f(x_{i+1}) + \sum_{i=0}^{n-1} f(x_i)\right) = \frac{1}{2}h\left(\sum_{i=1}^{n} f(x_i) + \sum_{i=0}^{n-1} f(x_i)\right) \\ &= \frac{1}{2}h\left(f(x_0) + 2\sum_{i=1}^{n-1} f(x_i) + f(x_n)\right) \end{aligned} \tag{14.1.1235}$$

That we often find in the academic literature, in the form:

$$\begin{aligned}A_T &= \frac{1}{2}\frac{b-a}{2}\left(f(a)+2\sum_{i=1}^{n-1}f(x_i)+f(b)\right)\\ &= \frac{b-a}{2}\left(\frac{f(a)+f(b)}{2}+\sum_{i=1}^{n-1}f(x_i)\right)\end{aligned} \tag{14.1.1236}$$

There are many other methods for solving this type of problem (including the Monte Carlo method that we will see further below).

In the case where the function f is not made of only positive values, we no longer speak anymore about "area" but of "Riemann sum". The sum to calculate are then:

$$S_D = h\sum_{i=0}^{n-1}f(x_{i+1}) \qquad S_G = h\sum_{i=0}^{n-1}f(x_i) \tag{14.1.1237}$$

and:

$$S_T = \frac{1}{2}h\left(f(x_0)+2\sum_{i=1}^{n-1}f(x_i)+f(x_n)\right) \tag{14.1.1238}$$

Hence to summarize we have the following four following more typical univariate integral numerical methods:

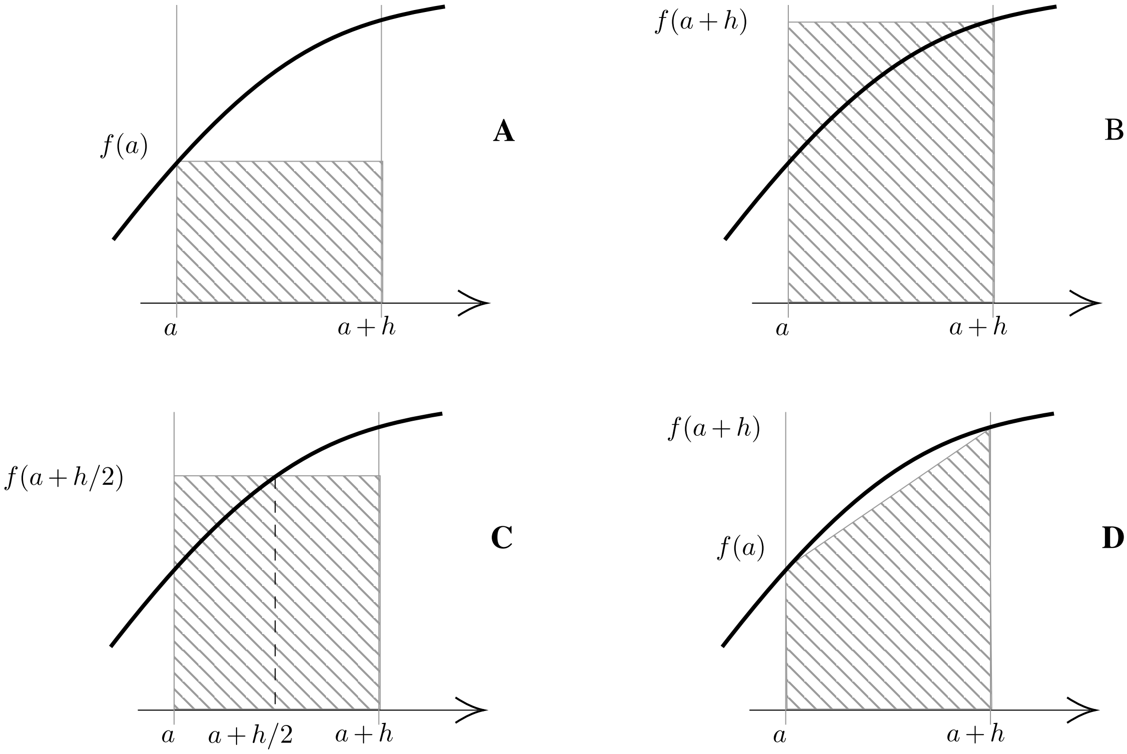


Figure 14.76: A: left rectangle method, B: right rectangle point, C: mid-point method, D: Trapezoidal method

14.1.16.3 Newton–Cotes formulas

The "Newton–Cotes formulas", also named the "Newton–Cotes quadrature rules" or simply "Newton–Cotes rules", are a group of formulas for numerical integration (also named "quadrature formulas as we have already mention it) based on evaluating the integrand at equally spaced points (they are named after Isaac Newton and Roger Cotes) and where the weights w_i of:

$$A(f) \cong \int_a^b f(x)\mathrm{d}x \approx \sum_{i=0}^{n} w_i f(x_i) \qquad (14.1.1239)$$

are derived from the Lagrange basis polynomials (see page 6594).

Let us see now three mainstream of the many "Newton-Cotes quadrature formulas". They are all based on Lagrange Polynomials (with or without the error term):

$$P_n(x) = \sum_{i=0}^{n} f(x_i)L(x_i) + \frac{f^{(n+1)}(\xi)}{(n+1)!}\prod_{i=0}^{n}(x-x_i) \qquad (14.1.1240)$$

Since the Taylor polynomials (see section Sequences and Series page 1723) have the property that all the information used in the approximation is concentrated at the single point x_0, it is not uncommon for these polynomials to give inaccurate approximations as we move away from x_0. This limits Taylor polynomial approximation to the situation in which approximations are needed only at points close to x_0.

The more the simple Newton-Cotes methods are based on high degree polynomials, the slower they are and the more difficult they are to code, but the more precise they are. Most often in practice, the total integration domain $[a,b]$ is much too large and the function varies too much on this domain for these methods to give satisfactory results. They are therefore almost never used as such.

The total domain $[a,b]$ is therefore subdivided into a large number of small intervals over each of which the simple Newton-Cotes methods can be successfully applied. We then speak of "composite Newton-Cotes method".

14.1.16.3.1 Trapezoidal Rule

We derive the Trapezoidal rule for approximating $\int_a^b f(x)\mathrm{d}x$ using the Lagrange polynomial method, with the linear Lagrange polynomial (see page 6594).

Let $x_0 = a$, $x_1 = b$, and $h = b - a$:

$$\begin{aligned}
\int\limits_{a=x_0}^{b=x_1} f(x)\mathrm{d}x &= \int\limits_{x_0}^{x_1} P_1(x)\mathrm{d}x + \frac{1}{2}\int\limits_{x_0}^{x_1} f''(\xi)\left(x-x_0\right)\left(x-x_1\right)\mathrm{d}x \\
&= \int\limits_{x_0}^{x_1}\left[\frac{x-x_1}{x_0-x_1}f\left(x_0\right)+\frac{x-x_0}{x_1-x_0}f\left(x_1\right)\right]\mathrm{d}x + \frac{1}{2}\int\limits_{x_0}^{x_1} f''(\xi)\left(x-x_0\right)\left(x-x_1\right)\mathrm{d}x \\
&= \left[\frac{\left(x-x_1\right)^2}{2\left(x_0-x_1\right)}f\left(x_0\right)+\frac{\left(x-x_0\right)^2}{2\left(x_1-x_0\right)}f\left(x_1\right)\right]_{x_0}^{x_1} + \frac{1}{2}f''(\xi)\int\limits_{x_0}^{x_1}\left(x-x_0\right)\left(x-x_1\right)\mathrm{d}x \\
&= -\frac{\left(x_0-x_1\right)^2}{2\left(x_0-x_1\right)}f\left(x_0\right)+\frac{\left(x_1-x_0\right)^2}{2\left(x_1-x_0\right)}f\left(x_1\right)+\frac{1}{2}f''(\xi)\left(-\frac{h^3}{6}\right) \\
&= \frac{h}{2}\left[f\left(x_0\right)+f\left(x_1\right)\right]-\frac{h^3}{12}f''(\xi)
\end{aligned} \tag{14.1.1241}$$

Thus, the Trapezoidal rule is:

$$\boxed{\int\limits_a^b f(x)\mathrm{d}x = \frac{h}{2}[f(a)+f(b)]-\frac{h^3}{12}f''(\xi) \approx \frac{b-a}{2}[f(a)+f(b)]} \tag{14.1.1242}$$

We fall back here on the relation that we have proved earlier but with another approach (a more analytic one...!).

Since the error term for the Trapezoidal rule involves f'', the rule gives the exact result when applied to any function whose second derivative is identically zero. That is, the Trapezoidal rule gives the exact result for polynomials of degree up to or equal to one.

Now we are home free. To get the Riemann sum for:

$$\int_a^b f(x)dx \tag{14.1.1243}$$

using the trapezoidal local approximation, assuming $f(x)$ is any integrable function.

Therefore on a long non-local interval splitted in n value such that $h = (b-a)/n$ we use:

$$\int_a^b f(x)\mathrm{d}x = \frac{h}{2}[f(a)+f(b)] \approx \frac{b-a}{2}[f(a)+f(b)] \tag{14.1.1244}$$

Here is the trivial associated pseudocode:

Algorithm 11: Trapezoidal's rule

```
Input: $a$,$b$,$n$,$f$
Output: Area
Area := 0
$h := (b-a)/n$
for $i = 1, 2, \ldots$ to $n$ do
    $\text{Area} = \text{Area} + \dfrac{(a+i\cdot h)-(a+(i-1)\cdot h)}{2}\left(f(a+i\cdot h)+f(a+(i-1)\cdot h)\right)$
return Area
```

14.1.16.3.2 Composite Trapezoidal Rule

When the trapezoidal rule is applied on the subintervals it is named a "composite trapezoidal rule".

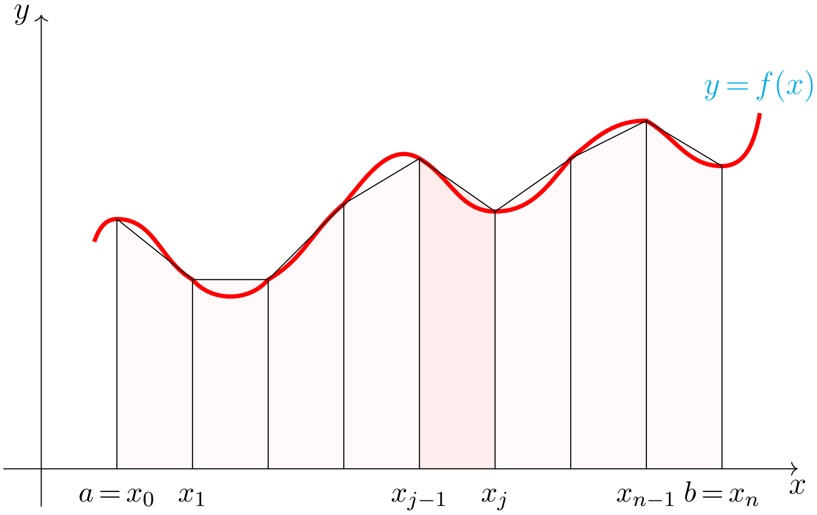


Figure 14.77: **Composite trapezoidal rule illustration**

To see this, let us define:

$$h = x_j - x_{j-1} \tag{14.1.1245}$$

Then:

$$\begin{aligned}\int\limits_{a=x_0}^{b=x_n} f(x)\mathrm{d}x &= \sum_{j=1}^{n}\int\limits_{x_{j-1}}^{x_j} f(x)\mathrm{d}x = \sum_{j=1}^{n}\left(\frac{h}{2}\left[f\left(x_{j-1}\right)+f\left(x_j\right)\right]-\frac{h^3}{12}f''\left(\xi_j\right)\right)\\ &= \sum_{j=1}^{n}\frac{h}{2}\left[f\left(x_{j-1}\right)+f\left(x_j\right)\right]-\frac{h^3}{12}\sum_{j=1}^{n}f''\left(\xi_j\right)\\ &= \sum_{j=1}^{n}\frac{h}{2}\left[f\left(x_{j-1}\right)+f\left(x_j\right)\right]-\frac{h^2}{12}(b-a)f''(\xi), \quad (\text{ since } nh = b-a)\end{aligned} \tag{14.1.1246}$$

Thus, the Composite Trapezoidal rule is:

$$\boxed{\int\limits_a^b f(x)\mathrm{d}x = \sum_{j=1}^{n} \frac{x_j - x_{j-1}}{2}\left[f\left(x_{j-1}\right) + f\left(x_j\right)\right] - \frac{h^2}{12}(b-a)f''(\xi)} \tag{14.1.1247}$$

14.1.16.3.3 Simpson's Rule

Simpson's rule can be derived by integrating the second Lagrange polynomial (see page 6594). However, this derivation gives only an $\mathcal{O}\left(h^4\right)$ error term involving $f^{(3)}$. To get a better error term, we use Taylor polynomial to derive the Simpson's rule.

Let $x_0 = a, x_2 = b$, and $x_1 = \dfrac{a+b}{2} = a + \dfrac{b-a}{2} = a + h$. That is, $x_1 - x_0 = h$ and $x_2 - x_1 = h$. Then:

$$\begin{aligned}
\int\limits_{a=x_0}^{b=x_2} f(x)\mathrm{d}x &= \int\limits_{x_0}^{x_2} \left[f\left(x_1\right) + f'\left(x_1\right)\left(x - x_1\right) + \frac{f''\left(x_1\right)}{2}\left(x - x_1\right)^2 + \frac{f'''\left(x_1\right)}{6}\left(x - x_1\right)^3 + \frac{f^{(4)}(\xi)}{24}\left(x - x_1\right)^4\right]\mathrm{d}x \\
&= \left[f\left(x_1\right)x + \frac{f'\left(x_1\right)}{2}\left(x - x_1\right)^2 + \frac{f''\left(x_1\right)}{6}\left(x - x_1\right)^3 + \frac{f'''\left(x_1\right)}{24}\left(x - x_1\right)^4 + \frac{f^{(4)}(\xi)}{120}\left(x - x_1\right)^5\right]_{x_0}^{x_2} \\
&= f\left(x_1\right)\left(x_2 - x_0\right) + \frac{f'\left(x_1\right)}{2}\left[\left(x_2 - x_1\right)^2 - \left(x_0 - x_1\right)^2\right] + \frac{f''\left(x_1\right)}{6}\left[\left(x_2 - x_1\right)^3 - \left(x_0 - x_1\right)^3\right] \\
&+ \frac{f'''\left(x_1\right)}{24}\left[\left(x_2 - x_1\right)^4 - \left(x_0 - x_1\right)^4\right] + \frac{f^{(4)}(\xi)}{120}\left[\left(x_2 - x_1\right)^5 - \left(x_0 - x_1\right)^5\right] \\
&= 2hf\left(x_1\right) + \frac{f'\left(x_1\right)}{2}\left[h^2 - h^2\right] + \frac{f''\left(x_1\right)}{6}\left[h^3 + h^3\right] + \frac{f'''\left(x_1\right)}{24}\left[h^4 - h^4\right] + \frac{f^{(4)}(\xi)}{120}\left[h^5 + h^5\right] \\
&= 2hf\left(x_1\right) + h^3\frac{f''\left(x_1\right)}{3} + h^5\frac{f^{(4)}(\xi)}{60} \\
&= 2hf\left(x_1\right) + \frac{h^3}{3}\left[\frac{f\left(x_0\right) - 2f\left(x_1\right) + f\left(x_2\right)}{h^2} - \frac{h^2}{12}f^{(4)}(\xi)\right] + \frac{h^5}{60}f^{(4)}(\xi) \\
&= \frac{h}{3}\left[f\left(x_0\right) + 4f\left(x_1\right) + f\left(x_2\right)\right] - \frac{h^5}{90}f^{(4)}(\xi)
\end{aligned} \tag{14.1.1248}$$

Thus, the Simpson's rule is:

$$\boxed{\int\limits_a^b f(x)\mathrm{d}x = \frac{h}{3}\left[f\left(x_0\right) + 4f\left(x_1\right) + f\left(x_2\right)\right] - \frac{h^5}{90}f^{(4)}(\xi) \approx \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right]} \tag{14.1.1249}$$

There are plenty of other possible derivations with plenty of different final formulation that the reader can find in the literature!

Now we are home free. To get the Riemann sum for:

$$\int_a^b f(x)dx \tag{14.1.1250}$$

using the quadratic local approximation, assuming $f(x)$ is any integrable function.

Therefore on a long non-local interval splitted in n value such that $h=(b-a)/n$ we use:

$$\int_a^b f(x)\mathrm{d}x=\sum_{i=0}^{n-1}\frac{h}{3}\left[f\left(x_i\right)+4f\left(\frac{x_i+x_{i+1}}{2}\right)+f\left(x_{i+1}\right)\right] \tag{14.1.1251}$$

Here is the trivial associated pseudocode:

Algorithm 12: Simpson's rule

```
Input: $a$,$b$,$n$,$f$
Output: Area
Area := 0
$h:=(b-a)/n$
for $i=0,1,\ldots$ to $n-1$ do
    Area = Area + $\frac{h}{3}\left[f\left(a+i\cdot h\right)+4f\left(\frac{(a+i\cdot h)+(a+(i+1)\cdot h)}{2}\right)+f\left(a+(i+1)h\right)\right]$
return Area
```

Remark

When Simpson's rule is applied on two subintervals it is named a "composite Simpson's rule".

All what we have seen so far can be summarized by the following figure:

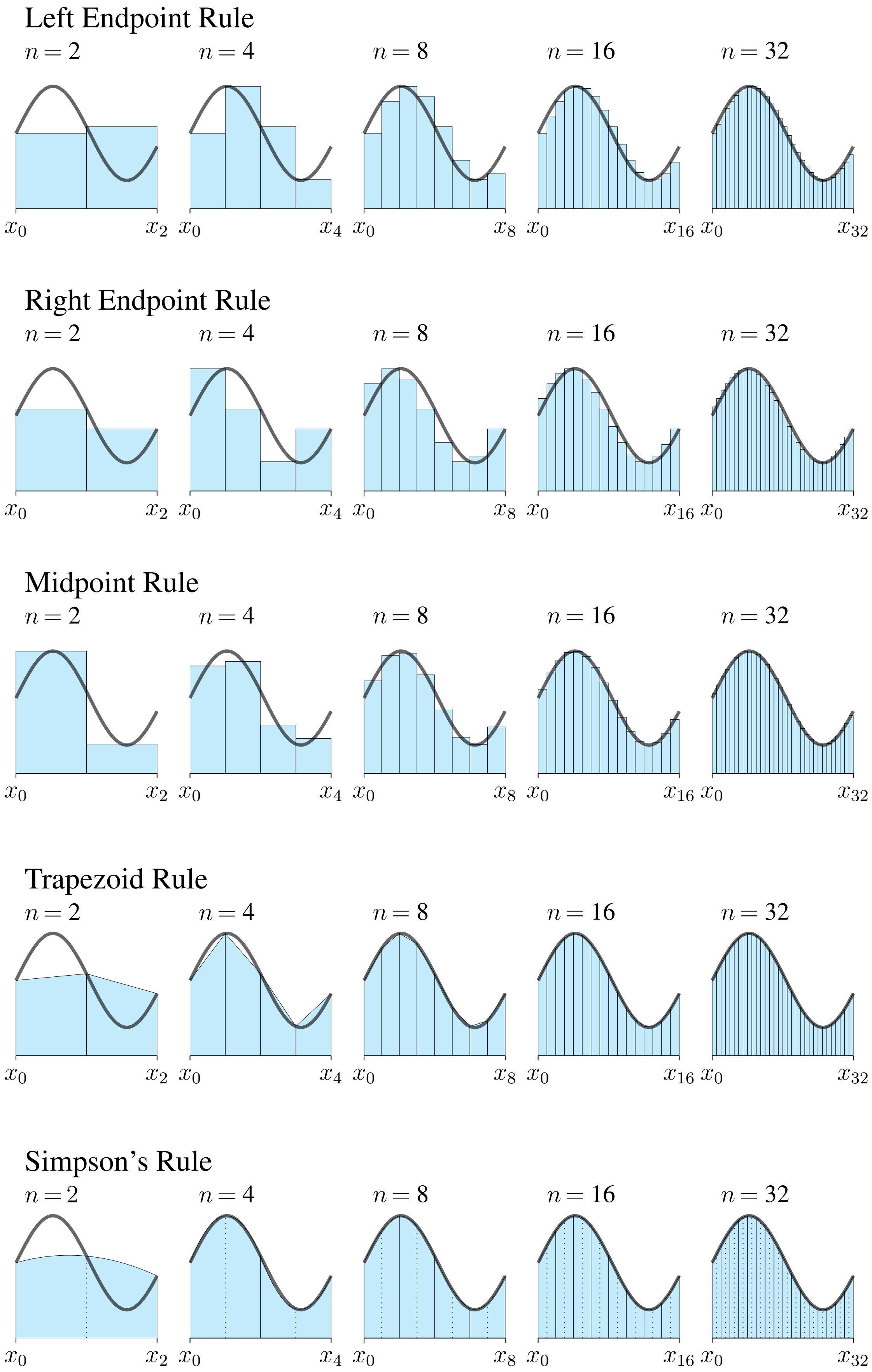

Left Endpoint Rule
Right Endpoint Rule
Midpoint Rule
Trapezoid Rule
Simpson's Rule
$n = 2$
$n = 4$
$n = 8$
$n = 16$
$n = 32$
x_0
x_2
x_4
x_8
x_{16}
x_{32}

Figure 14.78: **Quadrature Rules (author: Lauren K. Williams)**

14.1.16.4 Multidimensional integrals

Integrals of functions of several variables, over regions with dimension greater than one, are not easy. There are two reasons for this. First, the number of function evaluations needed to sample an N-dimensional space increases as the Nth power of the number needed to do a one-dimensional integral. If you need 30 function evaluations to do a one-dimensional integral crudely, then you will likely need on the order of $30 \times 30 \times 30 = 27000$ evaluations to reach the same crude level for a three-dimensional integral. Second, the region of integration in N-dimensional space is defined by an $N-1$ dimensional boundary which can itself be terribly complicated: It need not be convex or simply connected, for example. By contrast, the boundary of a one-dimensional integral consists of two numbers, its upper and lower limits.

The first question to be asked, when faced with a multidimensional integral is if can it be reduced analytically to a lower dimensionality one?

Alternatively, the function may have some special symmetry in the way it depends on its independent variables. If the boundary also has this symmetry, then the dimension can be reduced. In three dimensions, for example, the integration of a spherically symmetric function over a spherical region reduces, in polar coordinates, to a one-dimensional integral.

The next questions to be asked will guide your choice between two entirely different approaches to doing the problem. The questions are: Is the shape of the boundary of the region of integration simple or complicated? Inside the region, is the integrand smooth and simple, or complicated, or locally strongly peaked? Does the problem require high accuracy, or does it require an answer accurate only to a percent, or a few percent?

If your answers are that the boundary is complicated, the integrand is not strongly peaked in very small regions, and relatively low accuracy is tolerable, then your problem is a good candidate for Monte Carlo integration. This method is very straightforward to program, in its cruder forms. One needs only to know a region with simple boundaries that includes the complicated region of integration, plus a method of determining whether a random point is inside or outside the region of integration. Monte Carlo integration evaluates the function at a random sample of points, and estimates its integral based on that random sample.

If the boundary is simple, and the function is very smooth, then the remaining approaches, breaking up the problem into repeated one-dimensional integrals, or multidimensional Gaussian quadratures, will be effective and relatively fast. If you require high accuracy, these approaches are in any case the only ones available to you, since Monte Carlo methods are by nature asymptotically slow to converge.

For low accuracy, use repeated one-dimensional integration or Gaussian quadratures when the integrand is slowly varying and smooth in the region of integration, Monte Carlo when the integrand is oscillatory or discontinuous, but not strongly peaked in small regions.

If the integrand is strongly peaked in small regions, and you know where those regions are, break the integral up into several regions so that the integrand is smooth in each, and do each separately. If you don't know where the strongly peaked regions are, you might as well (at the level of sophistication of this textbook) quit: It is hopeless to expect an integration routine to search out unknown pockets of large contribution in a huge N-dimensional space.

If, on the basis of the above guidelines, we decide to pursue the repeated one-dimensional integration approach, here is how it works (the text below about the repeated one-dimensions approach is a bit a small summary of multivariate integral that we have already study in-deep in the section of Differential and Integral Calculus page 1518). For definiteness, we will consider the case of a three-dimensional integral in x, y, z-space. Two dimensions, or more than three dimensions, are entirely analogous.

The first step is to specify the region of integration by:

1. Its lower and upper limits in x, which we will denote x_1 and x_2
2. Its lower and upper limits in y at a specified value of x, denoted $y_1(x)$ and $y_2(x)$
3. Its lower and upper limits in z at specified x and y, denoted $z_1(x,y)$ and $z_2(x,y)$.

In other words, find the numbers x_1 and x_2, and the functions $y_1(x)$, $y_2(x)$, $z_1(x,y)$, and $z_2(x,y)$ such that:

$$I=\iiint f(x,y,z)\mathrm{d}x\mathrm{d}y\mathrm{d}z=\int\limits_{x_1}^{x_2}\mathrm{d}x\int\limits_{y_1(x)}^{y_2(x)}\mathrm{d}y\int\limits_{z_1(x,y)}^{z_2(x,y)}f(x,y,z)\mathrm{d}z \qquad (14.1.1252)$$

Now we can define a function $G(x,y)$ that does the innermost integral:

$$G(x,y)=\int\limits_{z_1(x,y)}^{z_2(x,y)}f(x,y,z)\mathrm{d}z \qquad (14.1.1253)$$

and a function $H(x)$ that does the integral of $G(x,y)$:

$$H(x)=\int\limits_{y_1(x)}^{y_2(x)}G(x,y)\mathrm{d}y \qquad (14.1.1254)$$

and finally our answer as an integral over $H(x)$:

$$I=\int\limits_{x_1}^{x_2}H(x)\mathrm{d}x \qquad (14.1.1255)$$

The methods that we have introduced earlier to calculate simple integrals can be generalized to multiple integrals. Les us consider first the following integral:

$$\iint_{\mathcal{R}}f(x,y)\mathrm{d}A \qquad (14.1.1256)$$

where $\mathcal{R}$ is a rectangular area of the plane, that is of the type:

$$\mathcal{R}=\{(x,y)|A\leq x\leq b,c\leq y\leq d\} \qquad (14.1.1257)$$

the limits of integration a,b,c and d being (for the moment...) constants.

To determine an approximated value of this integral, let us write it under the form:

$$\iint_{\mathcal{R}} f(x,y)\mathrm{d}A=\int\limits_a^b\left(\int\limits_c^d f(x,y)\mathrm{d}y\right)\mathrm{d}x \tag{14.1.1258}$$

and let us use the Simpson's method that we have proved earlier (see page 6629):

$$\int\limits_a^b f(x)\mathrm{d}x=\frac{h}{3}\left[f\left(x_0\right)+4f\left(x_1\right)+f\left(x_2\right)\right]-\frac{h^5}{90}f^{(4)}(\xi) \tag{14.1.1259}$$

for each of the integrals. Notice that every Newton-Cotes method can be used and that it is not necessary to use the same method for the two variables.

Let us choose two integers m and n to subdivide the interval $[a,b]$ into $2m$ sub-intervals with the step:

$$h=\frac{b-a}{2m} \tag{14.1.1260}$$

and $[c,d]$ into $2n$ sub-intervals with the step:

$$k=\frac{d-c}{2n} \tag{14.1.1261}$$

That is $y_0=c$, $y_1=c+k$, $y_{2n}=d$.

The first step has for purpose to calculate the integral:

$$\int\limits_c^d f(x,y)\mathrm{d}y \tag{14.1.1262}$$

with the Simpson's method by keeping x constant! We put $y_j=c+jk$, with $j=0,1,\ldots,2n$, therefore $y_0=c$ and $y_{2n}=d$. We then get (without the error term):

$$\int\limits_c^d f(x,y)\mathrm{d}y\cong\frac{k}{3}\left[f(x,y_0)+\sum_{j=1}^{n}f(x,y_{2j-1})+f(x,y_{2n})\right] \tag{14.1.1263}$$

and for the double integral:

$$\int\limits_a^b\left(\int\limits_c^d f(x,y)\mathrm{d}y\right)\mathrm{d}x\cong\frac{k}{3}\int\limits_a^b f(x,y_0)\mathrm{d}x+\frac{4k}{3}\sum_{j=1}^{n}\int\limits_a^b f(x,y_{2j-1})\mathrm{d}x+\frac{k}{3}f(x,y_{2n})\mathrm{d}x \tag{14.1.1264}$$

The Simpson's method can now be used to calculate each integral that leads on the variables x. Let us put for this purpose $x_i=a+ih$ with $i=0,1,\ldots,2m$, therefore $x_0=a$ and $x_{2m}=b$. For each value of the index j ($j=0,1,2,\ldots,2n$), that is by $y=y_j$, we then get without error term:

$$\int\limits_a^b f(x,y_j)\mathrm{d}x=\frac{h}{3}\left[f(x_0,y_j)+4\sum_{i=1}^{m}f(x_{2i-1},y_j)+f(x_{2m},y_j)\right] \tag{14.1.1265}$$

By substituting the latter in the first integral, we get:

$$\begin{aligned}\int_a^b\int_c^d f(x,y)\mathrm{d}y\mathrm{d}x \cong \frac{hk}{9}\Bigg[&f(x_0,y_0)+4\sum_{i=1}^{m}f(x_{2i-1},y_0)+f(x_{2m},y_0)+4\sum_{j=1}^{n}f(x_0,y_{2j-1})\\&+16\sum_{j=1}^{n}\sum_{i=1}^{m}f(x_{2i-1},y_{2j-1})+4\sum_{j=1}^{n}f(x_{2m},y_{2j-1})+f(x_0,y_{2n})\\&+4\sum_{i=1}^{m}f(x_{2i-1},y_{2n})+f(x_{2m},y_{2n})\Bigg]\end{aligned} \tag{14.1.1266}$$

This can be written also in matrix form:

$$\int_a^b\int_c^d f(x,y)\mathrm{d}y\mathrm{d}x \cong \frac{hk}{9}(1\,4\,\ldots\,4\,1)\begin{bmatrix}f_{0,0} & f_{0,1} & \ldots & f_{0,2n}\\ f_{1,0} & f_{1,1} & \ldots & f_{1,2n}\\ \vdots & \vdots & \ldots & \vdots\\ f_{2m,0} & f_{2m,1} & \ldots & f_{2m,2n}\end{bmatrix}\begin{pmatrix}1\\4\\2\\ \vdots\\2\\4\\1\end{pmatrix} \tag{14.1.1267}$$

The use of numerical approximation methods for double integrals is not limited to rectangular areas. The integrals of the shape:

$$\int_a^b\int_{c(x)}^{d(x)} f(x,y)\mathrm{d}x\mathrm{d}y \quad\text{or}\quad \int_c^d\int_{a(y)}^{b(y)} f(x,y)\mathrm{d}x\mathrm{d}y \tag{14.1.1268}$$

can be computed by modifying a bit the previous described method.

As example, let us use the Simpson's method to determine an approximation of the following double integral:

$$\int_a^b\int_{c(x)}^{d(x)} f(x,y)\mathrm{d}x\mathrm{d}y \tag{14.1.1269}$$

in the simplest case with $m=n=1$. Therefore the integration step of the variables x is then:

$$h=\frac{b-a}{2m}=\frac{b-a}{2} \tag{14.1.1270}$$

when that of the step of y varies with x:

$$k(x)=\frac{d(x)-c(x)}{2n}=\frac{d(x)-c(x)}{2} \tag{14.1.1271}$$

The estimation of the integral then takes the following shape (without the error term):

$$\begin{aligned}\int\limits_a^b\int\limits_{c(x)}^{d(x)} f(x,y)\mathrm{d}x\mathrm{d}y &\cong \int\limits_a^b \frac{k(x)}{3}\left[f(x,c(x))+4f(x,c(x)+k(x))+f(x,d(x))\right]\mathrm{d}x\\ &\cong \frac{h}{3}\left(\frac{k(a)}{3}\left[f(a,c(a))+4f(a,c(a)+k(a))+f(a,d(a))\right]\right.\\ &+\frac{4k(x_1)}{3}\left[f(x_1,c(x_1))+4f(x_1,c(x_1)+k(x_1))+f(x_1,d(x_1))\right]\\ &\left.+\frac{k(b)}{3}\left[f(b,c(b))+4f(b,c(b)+k(b))+f(b,d(b))\right]\right)\end{aligned} \tag{14.1.1272}$$

14.1.16.5 Numerical solution of ordinary differential equations

Numerical integration of ordinary differential equations is a frequent task of numerical analysis. Numerical integration of differential equations is used if the equations are non-linear or if we have a large system of linear equations with constant coefficients, where the analytical solution can be found, but it is in the form of long and complicated expressions containing exponential functions. Numerical integration of such systems is more efficient both in human time and in computer time.

Numerical integration of linear equations with non-constant coefficients is also more efficient than the analytical solution; in the case of inner diffusion in porous catalyst with a chemical reaction of the first order the analytical solution contains Bessel functions, which can be evaluated more conveniently when we use numerical integration of the original equations than to evaluate Bessel functions.

We have already seen how to use Euler's method and the method of Taylor's expansion at the page 6587, let us see more elaborated techniques.

14.1.16.5.1 Runge-Kutta methods

In numerical analysis, the Runge–Kutta methods are a family of implicit and explicit iterative methods, which include the well-known "Euler's method" (we will prove that further below), used in temporal discretization for the approximate solutions of ordinary differential equations. These methods were developed around 11900 (holocene calendar) by the German mathematicians Carl Runge and Wilhelm Kutta.

The techniques discussed in these pages approximate the solution of first order ordinary differential equations (with initial conditions) of the form ("Cauchy's problem" type):

$$\frac{\mathrm{d}y(t)}{\mathrm{d}t}=y'(t)=f(y(t),t) \quad \text{with} \quad y(t_0)=y_0 \tag{14.1.1273}$$

In the following developments you will understand why we say sometimes that Runge-Kutta methods "simulate" the Taylor's method!

14.1.16.5.1.1 1st order Runge-Kutta method

The first order Runge-Kutta method is already known to us. It's simply the "Euler's method" already derived earlier before.

Let us review that latter introducing a notation that is really mainstream and will be useful to understand the higher Runge-Kutta order methods.

For this we start again from:

$$\frac{\mathrm{d}y(t)}{\mathrm{d}t}=y'(t)=f(y(t),t) \quad \text{with} \quad y(t_0)=y_0 \tag{14.1.1274}$$

and we write the approximation to the derivative as:

$$k_1 = y'(t) = f(y^*(t), t) \tag{14.1.1275}$$

We expand $y(t)$ around t_0 assuming a time step h using Taylor expansion given for recall by:

$$P_n(x) = \sum_{k=0}^{n} \frac{1}{k!} \frac{\mathrm{d}^k P(x_0)}{\mathrm{d}x^k} (x - x_0)^k \tag{14.1.1276}$$

This give us:

$$y(t_0 + h) = y(t_0) + y'(t_0)h + y''(t_0)\frac{h^2}{2} + \ldots \tag{14.1.1277}$$

and drop all terms after the linear term. Because all of the dropped terms are multiplied by h^2 or greater, we say that the algorithm is accurate to order h^2 locally, or $\mathcal{O}(h^2)$ (if h is small the other terms that are multiplied by h^3, h^4... which will be even smaller, and can be dropped as well):

$$y(t_0 + h) \cong y(t_0) + y'(t_0)h \tag{14.1.1278}$$

This gives us our approximate Euler's method solution at the next time step:

$$y^*(t_0 + h) \cong y^*(t_0) + k_1 h \tag{14.1.1279}$$

Since the number of steps over the whole interval is proportional to $1/h$ (or $\mathcal{O}(h^{-1})$) we might expect the overall accuracy to be the $\mathcal{O}(h^2) \cdot \mathcal{O}(h^{-1}) = \mathcal{O}(h)$. A rigorous analysis proves that this is true.

With a little more work we can develop a method that is accurate to higher order than $\mathcal{O}(h)$. And this is the second order Runge-Kutta method.

14.1.16.5.1.2 2nd order Runge-Kutta method

In the following derivation we will use two math facts that are reviewed here. You should be familiar with this from a course in multivariate calculus.

First, let us recall the relation we have derived in the section of Sequences and Series on bivariate Taylor series (see page 1714):

$$\begin{aligned} f(x_0 + h, y_0 + k) \cong f(x_0, y_0) &+ \frac{1}{1!}\frac{\partial f}{\partial x}(x_0, y_0)h + \frac{1}{1!}\frac{\partial f}{\partial y}(x_0, y_0)k \\ &+ \frac{1}{2!}\left[\frac{\partial^2 f}{\partial x^2}(x_0, y_0)h^2 + 2\frac{\partial^2 f}{\partial x \partial y}(x_0, y_0)hk + \frac{\partial^2 f}{\partial y^2}(x_0, y_0)k^2\right] \end{aligned} \tag{14.1.1280}$$

Hence:

$$\begin{aligned} f(x_0 + \Delta x, y_0 + \Delta y) &= f(x_0, y_0) + \left.\frac{\partial f(x,y)}{\partial x}\right|_{x_o, y_0} \Delta x + \left.\frac{\partial f(x,y)}{\partial y}\right|_{x_o, y_0} \Delta y + \ldots \\ &= f + f_x \Delta x + f_y \Delta y + \ldots \end{aligned} \tag{14.1.1281}$$

If f is a function of two variables $f(x,y)$, where $x=r(t)$ and $y=s(t)$, then by the chain rule for partial derivatives:

$$\frac{\mathrm{d}f}{\mathrm{d}t}=\frac{\partial f}{\partial r}\frac{\mathrm{d}r}{\mathrm{d}t}+\frac{\partial f}{\partial s}\frac{\mathrm{d}s}{\mathrm{d}t} \tag{14.1.1282}$$

In particular if:

$$y'(t)=\frac{\mathrm{d}y(t)}{\mathrm{d}t}=f(y(t),t)=f \tag{14.1.1283}$$

then:

$$y''(t)=\frac{\mathrm{d}f(y(t),t)}{\mathrm{d}t}=\frac{\partial f}{\partial t}\frac{\mathrm{d}t}{\mathrm{d}t}+\frac{\partial f}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t}=\frac{\partial f}{\partial t}+\frac{\partial f}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t}=f_t+f_yf \tag{14.1.1284}$$

To start, recall that we are expressing our differential equation as

$$\frac{\mathrm{d}y(t)}{\mathrm{d}t}=y'(t)=f(y(t),t) \quad \text{with}\, y(t_0)=y_0 \tag{14.1.1285}$$

Now we define two approximations to the derivative:

$$\begin{aligned} k_1&=f(y^*(t_0),t_0)\\ k_2&=f(y^*(t_0)+\beta hk_1,t_0+\alpha h) \end{aligned} \tag{14.1.1286}$$

In all cases α and β will represent fractional values between 0 and 1. These equation state that k_1 is the approximation to the derivative based on the estimated value of $y(t)$ at $t=t_0$ (i.e., $y^*(t_0)$) and the time at t_0. The value of k_2 is based upon the estimated value, $y^*(t_0)$, plus some fraction of the step size, βh, times the slope k_1, and the time at $t_0+\alpha h$ (i.e., some time between t_0 and t_0+h).

To update our solution with the next estimate of $y(t)$ at t_0+h we use then the relation:

$$\begin{aligned} y^*(t_0+h)&\cong y^*(t_0)+h(ak_1+bk_2)\\ &=y^*(t_0)+h\left[af(y^*(t_0),t_0)+bf(y^*(t_0)+\beta hk_1,t_0+\alpha h)\right] \end{aligned} \tag{14.1.1287}$$

That latter relation can be found in some textbooks under the form:

$$y^*_{k+1}\cong y^*_k+h\left[af(t_k,y_k)+bf\left(t_k+\alpha h,y_k+\beta hf(t_k,y_k)\right)\right] \tag{14.1.1288}$$

This equation states that we get the value of $y^*(t_0+h)$ from the value of $y^*(t_0)$ plus the time step, h, multiplied by a slope that is a weighted sum of k_1 and k_2. In the method described previously $a=0$ and $b=1$, so we used only the second estimate for the slope.

Note that Euler's Method (first order Runge-Kutta) is a special case of this method with $a=1$, $b=0$, and α and β don't matter because k_2 is not used in the update equation!

Remark

In general a Runge-Kutta method of order s will be written as:

$$y^*_{t_0+h} \cong y^*_{t_0} + h\sum_{i=1}^{s} a_i k_i + \mathcal{O}(h^{s+1}) \tag{14.1.1289}$$

where:

$$k_i = f\left(y_{t_0} + h\sum_{j=1}^{s} \beta_{ij} k_j,\ t_n + \alpha_i h\right) \tag{14.1.1290}$$

Our goal now is to determine, from first principles, how to find the values a, b, α and β that result in low error. Starting with the update equation above:

$$y^*(t_0+h) \cong y^*(t_0) + h\left[af(y^*(t_0),t_0) + bf\left(y^*(t_0) + \beta h k_1, t_0 + \alpha h\right)\right] \tag{14.1.1291}$$

We can use now the bivariate-dimensional Taylor series (where the increment in the first dimension is $\beta h k_1$, and the increment in the second dimension is αh) to expand the rightmost term:

$$\begin{aligned} f\left(y^*(t_0) + \beta h k_1,\ t_0 + \alpha h\right) &= f + f_y \beta h k_1 + f_t \alpha h + \mathcal{O}(h^2) \\ &= f + f_y f \beta h + f_t \alpha h + \mathcal{O}(h^2) \end{aligned} \tag{14.1.1292}$$

In the last line we used the fact the $k_1 = f$. Now we substitute this in the update equation:

$$\begin{aligned} y^*(t_0+h) &= y^*(t_0) + h\left[af(y^*(t_0),t_0) + bf\left(y^*(t_0) + \beta h k_1, t_0 + \alpha h\right)\right] \\ &= y^*(t_0) + h\left[af(y^*(t_0),t_0) + b\left(f + f_y f \beta h + f_t \alpha h + h\mathcal{O}(h^2)\right)\right] \\ &= y^*(t_0) + h(a+b)f + h^2 f_t b\alpha + h^2 b\beta f_y f + \mathcal{O}(h^3) \end{aligned} \tag{14.1.1293}$$

To finish we compare this approximation with the expression for a Taylor expansion of the exact solution (going from the first line to the second we used the chain rule for partial derivatives):

$$\begin{aligned} y(t_0+h) &= y(t_0) + h y'(t)\big|_{t_0} + \frac{h^2}{2} y''(t)\big|_{t_0} + \ldots \\ &= y(t_0) + hf + \frac{h^2}{2}\left(f_t + f_y f\right) + \ldots \\ &= y(t_0) + hf + \frac{h^2}{2} f_t + \frac{h^2}{2} f_y f + \ldots \end{aligned} \tag{14.1.1294}$$

Comparing this expression with our final expression for the approximation:

$$y^*(t_0+h) = y^*(t_0) + hf(a+b) + h^2 f_t b\alpha + h^2 b\beta f_y f + \ldots \tag{14.1.1295}$$

we see that they agree up to the error terms (third order and higher) if we define the constants, a, b, α and β such that:

$$\begin{aligned} a + b &= 1 \\ b\alpha &= \frac{1}{2} \\ b\beta &= \frac{1}{2} \end{aligned} \tag{14.1.1296}$$

Also written:

$$\begin{aligned} a+b-1&=0\\ b\alpha-\frac{1}{2}&=0\\ b\beta-\frac{1}{2}&=0 \end{aligned} \tag{14.1.1297}$$

This system is underspecified, there are four unknowns, and only three equations, so more than one solution is possible.

The following choices are the most common one:

- $a=0$, $b=1$ and $\alpha=\beta=1/2$, we get:

  $$\begin{aligned} y^*(t_0+h)&\cong y^*(t_0)+h\left[af(y^*(t_0),t_0)+bf\left(y^*(t_0)+\beta hk_1,t_0+\alpha h\right)\right]\\ &=y^*(t_0)+h\left[f(y^*(t_0),t_0)+f\left(y^*(t_0)+\frac{1}{2}hk_1,t_0+\frac{1}{2}h\right)\right] \end{aligned} \tag{14.1.1298}$$

  This is named the "middle-point method".

- $a=b=1/2$ and $\alpha=\beta=1$, we get:

  $$\begin{aligned} y^*(t_0+h)&\cong y^*(t_0)+h\left[af(y^*(t_0),t_0)+bf\left(y^*(t_0)+\beta hk_1,t_0+\alpha h\right)\right]\\ &=y^*(t_0)+h\left[\frac{1}{2}f(y^*(t_0),t_0)+\frac{1}{2}f\left(y^*(t_0)+hk_1,t_0+h\right)\right]\\ &=y^*(t_0)+\frac{h}{2}\left[f(y^*(t_0),t_0)+f\left(y^*(t_0)+hk_1,t_0+h\right)\right] \end{aligned} \tag{14.1.1299}$$

  This is named the "improved Euler's method" or "modified Euler's method". Notice that the term:

  $$\frac{h}{2}\left[f(y^*(t_0),t_0)+f\left(y^*(t_0)+hk_1,t_0+h\right)\right] \tag{14.1.1300}$$

  is equivalent to the trapezoidal approximation of an integral of $y^*(y(t),t)$.

- $a=1/4$, $b=3/4$ and $\alpha=\beta=2/3$, we get:

  $$\begin{aligned} y^*(t_0+h)&\cong y^*(t_0)+h\left[af(y^*(t_0),t_0)+bf\left(y^*(t_0)+\beta hk_1,t_0+\alpha h\right)\right]\\ &=y^*(t_0)+h\left[\frac{1}{4}f(y^*(t_0),t_0)+\frac{3}{4}f\left(y^*(t_0)+\frac{2}{3}hk_1,t_0+\frac{2}{3}h\right)\right] \end{aligned} \tag{14.1.1301}$$

  This is named the "Heun's method".

More complicated and more accurate methods can be derived by a similar approach and this can be the subject of a whole book.

14.1.16.5.1.3 4th order Runge-Kutta method

It is the most widely known member of the Runge–Kutta family and generally referred to as "RK4", the "classic Runge–Kutta method" or simply as the "Runge–Kutta method".

To derive it, let us recall that general a Runge–Kutta (empirical) method of order s can be written as:

$$y_{t+h} \cong y_t + h\sum_{i=1}^{s} a_i k_i + \mathcal{O}(h^{s+1}) \tag{14.1.1302}$$

where:

$$k_i = f\left(y_t + h\sum_{j=1}^{s}\beta_{ij}k_j,\ t_n+\alpha_i h\right) \tag{14.1.1303}$$

are increments obtained evaluating the derivatives of y_t at the i-th order.

We develop the derivation for the Runge–Kutta fourth-order method using the general formula with $s=4$ evaluated, at the starting point, the midpoint and the end point of any interval $(t,\ t+h)$. Thus, we choose:

$$\begin{array}{cc} \alpha_i & \beta_{ij} \\ \alpha_1=0 & \beta_{21}=\dfrac{1}{2} \\ \alpha_2=\dfrac{1}{2} & \beta_{32}=\dfrac{1}{2} \\ \alpha_3=\dfrac{1}{2} & \beta_{43}=1 \\ \alpha_4=1 & \end{array} \tag{14.1.1304}$$

and $\beta_{ij}=0$ otherwise. We begin by defining the following quantities:

$$\begin{aligned} y_{t+h}^1 &= y_t + hf\left(y_t,\ t\right) \\ y_{t+h}^2 &= y_t + hf\left(y_{t+h/2}^1,\ t+\frac{h}{2}\right) \\ y_{t+h}^3 &= y_t + hf\left(y_{t+h/2}^2,\ t+\frac{h}{2}\right) \end{aligned} \tag{14.1.1305}$$

where:

$$y_{t+h/2}^1 = \frac{y_t+y_{t+h}^1}{2} \qquad \text{and} \qquad y_{t+h/2}^2 = \frac{y_t+y_{t+h}^2}{2} \tag{14.1.1306}$$

If we define:

$$\begin{aligned} k_1 &= f(y_t,\ t) \\ k_2 &= f\left(y_{t+h/2}^1,\ t+\frac{h}{2}\right) = f\left(y_t+\frac{h}{2}k_1,\ t+\frac{h}{2}\right) \\ k_3 &= f\left(y_{t+h/2}^2,\ t+\frac{h}{2}\right) = f\left(y_t+\frac{h}{2}k_2,\ t+\frac{h}{2}\right) \\ k_4 &= f\left(y_{t+h}^3,\ t+h\right) = f\left(y_t+hk_3,\ t+h\right) \end{aligned} \tag{14.1.1307}$$

and for the previous relations we can show that the following equalities hold up to $\mathcal{O}(h^2)$:

$$\begin{aligned}
k_2 &= f\left(y^1_{t+h/2},\ t+\frac{h}{2}\right) = f\left(y_t+\frac{h}{2}k_1,\ t+\frac{h}{2}\right)\\
&= f\left(y_t,\ t\right)+\frac{h}{2}\frac{\mathrm{d}}{\mathrm{d}t}f\left(y_t,\ t\right)\\
k_3 &= f\left(y^2_{t+h/2},\ t+\frac{h}{2}\right) = f\left(y_t+\frac{h}{2}f\left(y_t+\frac{h}{2}k_1,\ t+\frac{h}{2}\right),\ t+\frac{h}{2}\right)\\
&= f\left(y_t,\ t\right)+\frac{h}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left[f\left(y_t,\ t\right)+\frac{h}{2}\frac{\mathrm{d}}{\mathrm{d}t}f\left(y_t,\ t\right)\right]\\
k_4 &= f\left(y^3_{t+h},\ t+h\right) = f\left(y_t+hf\left(y_t+\frac{h}{2}k_2,\ t+\frac{h}{2}\right),\ t+h\right)\\
&= f\left(y_t+hf\left(y_t+\frac{h}{2}f\left(y_t+\frac{h}{2}f\left(y_t,\ t\right),\ t+\frac{h}{2}\right),\ t+\frac{h}{2}\right),\ t+h\right)\\
&= f\left(y_t,\ t\right)+h\frac{\mathrm{d}}{\mathrm{d}t}\left[f\left(y_t,\ t\right)+\frac{h}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left[f\left(y_t,\ t\right)+\frac{h}{2}\frac{\mathrm{d}}{\mathrm{d}t}f\left(y_t,\ t\right)\right]\right]
\end{aligned}\tag{14.1.1308}$$

where:

$$\frac{\mathrm{d}}{\mathrm{d}t}f(y_t,\ t)=\frac{\partial}{\partial y}f(y_t,\ t)\dot{y}_t+\frac{\partial}{\partial t}f(y_t,\ t)=f_y(y_t,\ t)\dot{y}+f_t(y_t,\ t):=\ddot{y}_t\tag{14.1.1309}$$

is the total exact differential of f with respect to time.

If we now express the general formula using what we just derived we get:

$$\begin{aligned}
y_{t+h} \cong y_t+h\Bigg\{&a\cdot f(y_t,\ t)+b\cdot\left[f(y_t,\ t)+\frac{h}{2}\frac{\mathrm{d}}{\mathrm{d}t}f(y_t,\ t)\right]+\\
&+c\cdot\left[f(y_t,\ t)+\frac{h}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left[f\left(y_t,\ t\right)+\frac{h}{2}\frac{\mathrm{d}}{\mathrm{d}t}f(y_t,\ t)\right]\right]+\\
&+d\cdot\left[f(y_t,\ t)+h\frac{\mathrm{d}}{\mathrm{d}t}\left[f(y_t,\ t)+\frac{h}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left[f(y_t,\ t)+\frac{h}{2}\frac{\mathrm{d}}{\mathrm{d}t}f(y_t,\ t)\right]\right]\right]\Bigg\}+\mathcal{O}(h^5)\\
=y_t+a\cdot hf_t&+b\cdot hf_t+b\cdot\frac{h^2}{2}\frac{\mathrm{d}f_t}{\mathrm{d}t}+c\cdot hf_t+c\cdot\frac{h^2}{2}\frac{\mathrm{d}f_t}{\mathrm{d}t}+\\
&+c\cdot\frac{h^3}{4}\frac{\mathrm{d}^2f_t}{\mathrm{d}t^2}+d\cdot hf_t+d\cdot h^2\frac{\mathrm{d}f_t}{\mathrm{d}t}+d\cdot\frac{h^3}{2}\frac{\mathrm{d}^2f_t}{\mathrm{d}t^2}+d\cdot\frac{h^4}{4}\frac{\mathrm{d}^3f_t}{\mathrm{d}t^3}+\mathcal{O}(h^5)
\end{aligned}\tag{14.1.1310}$$

and comparing this with the Taylor series of y_{t+h} around y_t:

$$\begin{aligned}
y_{t+h}&=y_t+h\dot{y}_t+\frac{h^2}{2}\ddot{y}_t+\frac{h^3}{6}y_t^{(3)}+\frac{h^4}{24}y_t^{(4)}+\mathcal{O}(h^5)=\\
&\cong y_t+hf(y_t,\ t)+\frac{h^2}{2}\frac{\mathrm{d}}{\mathrm{d}t}f(y_t,\ t)+\frac{h^3}{6}\frac{\mathrm{d}^2}{\mathrm{d}t^2}f(y_t,\ t)+\frac{h^4}{24}\frac{\mathrm{d}^3}{\mathrm{d}t^3}f(y_t,\ t)
\end{aligned}\tag{14.1.1311}$$

we get a system of constraints on the coefficients:

$$\begin{cases} a+b+c+d=1 \\ \tfrac{1}{2}b+\dfrac{1}{2}c+d=\dfrac{1}{2} \\ \dfrac{1}{4}c+\dfrac{1}{2}d=\dfrac{1}{6} \\ \dfrac{1}{4}d=\dfrac{1}{24} \end{cases} \tag{14.1.1312}$$

which when solved gives:

$$a=\frac{1}{6},b=\frac{1}{3},c=\frac{1}{3},d=\frac{1}{6} \tag{14.1.1313}$$

as stated above.

14.1.17 Optimization

In mathematics, computer science and operations research, mathematical optimization (alternatively "mathematical programming") is the selection of a best element (with regard to some criteria) from some set of available alternatives.

In the simplest case, an optimization problem consists of maximizing or minimizing a real function by systematically choosing input values from within an allowed set and computing the value of the function. The generalization of optimization theory and techniques to other formulations comprises a large area of applied mathematics. More generally, optimization includes finding best available values of some objective function given a defined domain (or a set of constraints), including a variety of different types of objective functions and different types of domains.

In mathematics, conventional optimization problems are usually stated in terms of minimization (or changed to be as!). A large number of algorithms proposed for solving optimization problems are not capable of making a distinction between local optimal solutions and rigorous optimal solutions. The branch of applied mathematics and numerical analysis that is concerned with the development of deterministic algorithms that are capable of guaranteeing convergence in finite time to the actual optimal solution is named "global optimization".

Remark

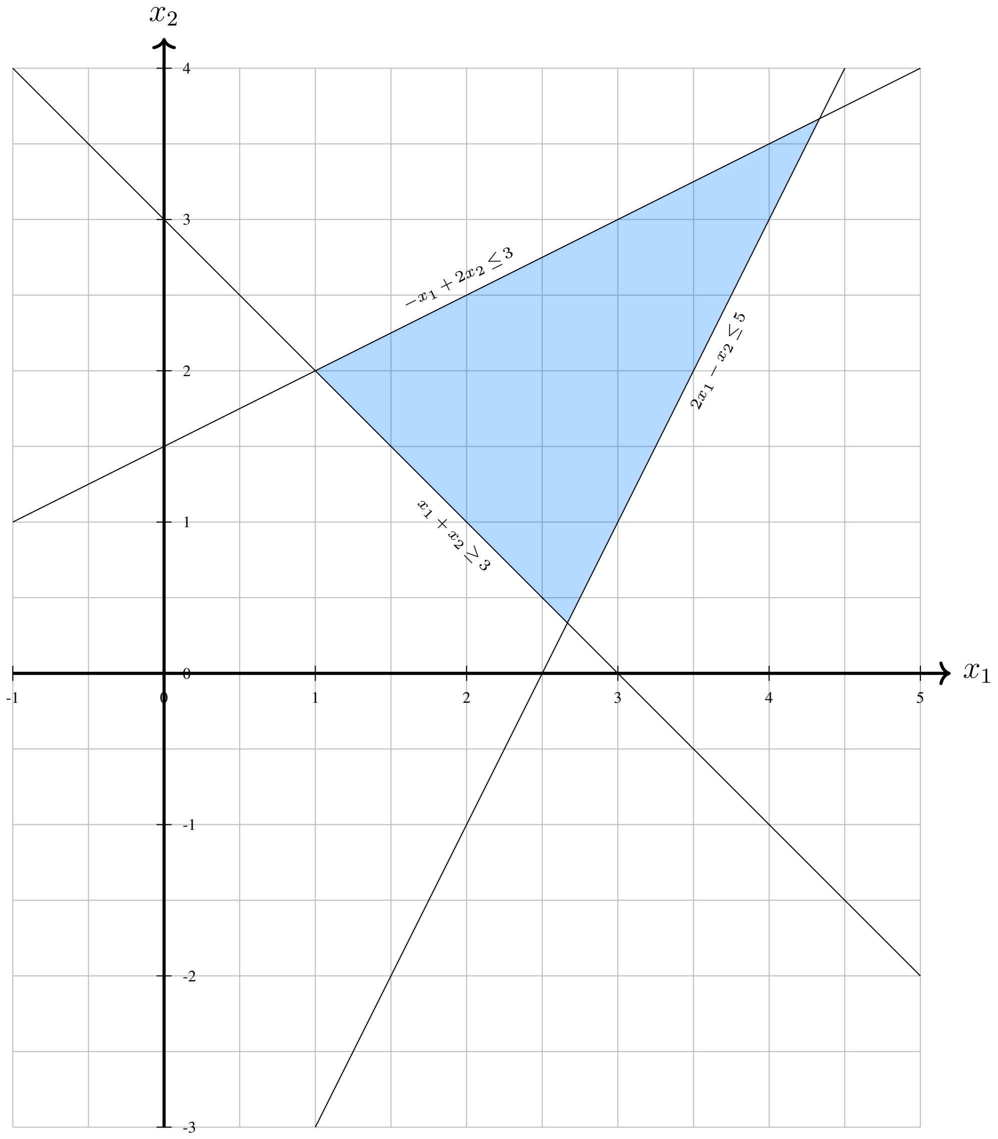
We speak of "convex optimization" when any local minimum must be a global minimum. In other word there is only one unique solution.

In the context of problem solving which involves two variables and their products, we then speak logically "quadratic programming (QP)" or simply "non-linear programming". This is typically the case in financial engineering in portfolios modelling (see section Economy page 7699) or in forecasting. We will study in details also further below simplified and particular version of the corresponding models that are the: Newton's method, quasi-Newton's method, conjugate gradient method and non-linear GRG.

14.1.17.1 Linear programming (Linear Optimization)

The objective of the linear programming (LP-programming) is to find the optimum value of a linear function subject to a system of equations consisting in inequalities constraints that are also linear. The objective function is named "economic function" (because used a lot in Economy) and we solve this type of system using typically, among others, a method named "simplex method" (see below), the corresponding graph is a "polygon constraints" (when the number of variable is obviously equal to 2).

The reader can remember the following diagram (convex polytope) that we saw in the section Calculus:

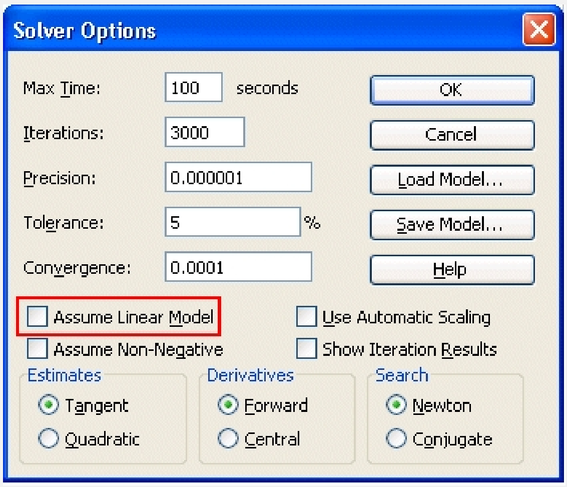


Corresponding to the following system of inequalities:

$$\begin{aligned}-x_1+2x_2&\leq 3\\ 2x_1-x_2&\leq 5\\ x_1+x_2&\geq 3\end{aligned} \tag{14.1.1314}$$

Remark

Linear programming is widely used (to name only the most famous case) in Supply Chain (maximal flow problem also named "transport problem"), in Corporate Finance or also in Decision Theory when we solve a mixed strategy game (see the section of Game and Decision Theory for a practical example). That's why Microsoft® Excel® 12.0 and earlier includes a tool named the "solver" in which there is an option named "Assume Linear Model" which then requires the use of the simplex model that we will study below:



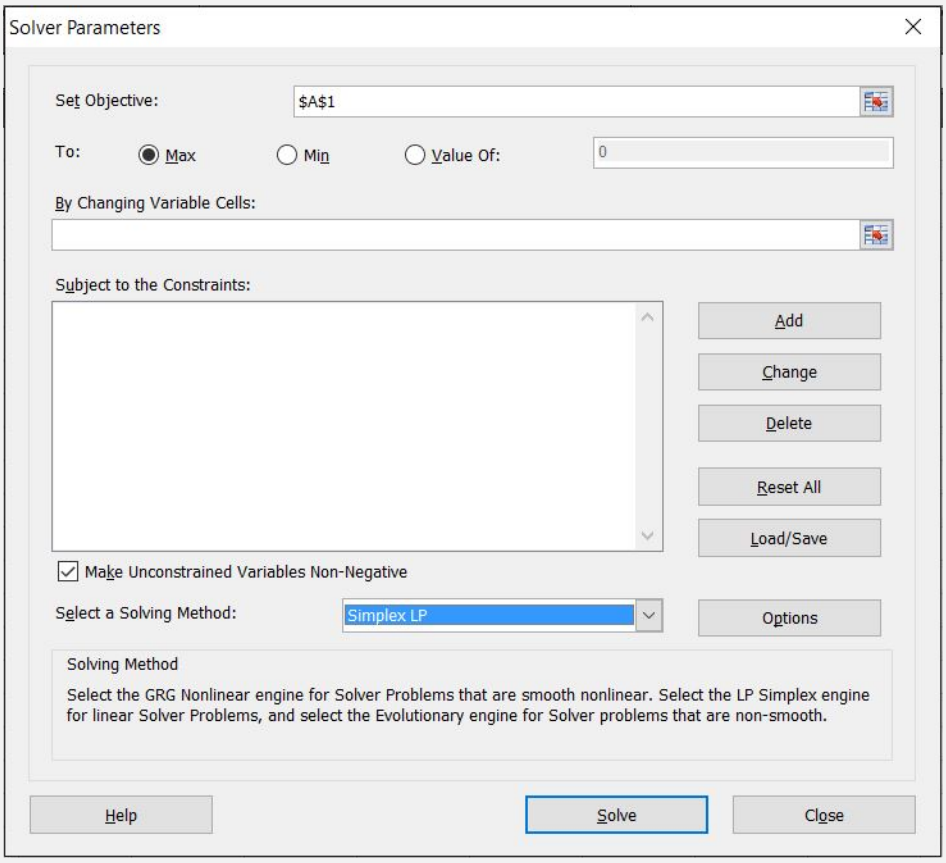
Solver Options

Max Time: 100 seconds
Iterations: 3000
Precision: 0.000001
Tolerance: 5 %
Convergence: 0.0001

☐ Assume Linear Model
☐ Use Automatic Scaling
☐ Assume Non-Negative
☐ Show Iteration Results

Estimates: ● Tangent ○ Quadratic
Derivatives: ● Forward ○ Central
Search: ● Newton ○ Conjugate

or since the 2010 version of the software (the user interface has completely changed):



Solver Parameters

Set Objective: $A$1

To: ● Max ○ Min ○ Value Of: 0

By Changing Variable Cells:

Subject to the Constraints:

☑ Make Unconstrained Variables Non-Negative

Select a Solving Method: Simplex LP

Solving Method

Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

We will focus especially in this section on the most widely used algorithm for linear optimization named the "simplex algorithm".

When a problem can be modelled as an economic function to be maximized with respect to certain constraints that are purely additive, we are typically in the context of linear programming!

So an economic function Z is defined as:

$$Z=c_1x_1+c_2x_2+\ldots+c_nx_n=\sum_{i=1}^{n}c_ix_i \tag{14.1.1315}$$

where the x_i are variables that affect the value of Z, and the c_i the weights of these variables modelling the relative importance of each of them on the value of the economic function.

The constraints related to the variables are expressed by the following linear system:

$$\begin{aligned}&a_{11}x_1+a_{12}x_2+\ldots+a_{1n}x_n\leq B_1\\&a_{21}x_1+a_{22}x_2+\ldots+a_{2n}x_n\leq B_2\\&\ldots\\&a_{m1}x_1+a_{m2}x_2+\ldots+a_{mn}x_n\leq B_n\end{aligned} \tag{14.1.1316}$$

Under general and matrix form this problem is written as:

$$\boxed{\begin{aligned}\underset{\vec{x}}{\text{minimize}}\quad & Z(\vec{x})=\sum_{j=1}^{n}c_ix_i\\ \text{subject to}\quad & \sum_{i=1}^{m}a_{ij}x_j=b_i \quad i=1\ldots m\\ & x_j\geq 0\quad j=1\ldots n\end{aligned}} \tag{14.1.1317}$$

To see the different methods of resolution let us use an example as theoretical introduction:

A factory produces two types of pieces $P1$ and $P2$ machined in two workshops $A1$ and $A2$. Machining times are for $P1$ or 3 hours in the workshop $A1$ and 6 hours in the workshop $A2$ and of for 4 hours for $P2$ in the workshop $A1$ and 3 hours in the workshop $A2$.

Weekly up-time of human resources (workers) of the $A1$ workshop is 160 hours and that of the workshop $A2$ is 180 hours.

The profit margin is of $1,200.-$ for the pieces $P1$ and $1,000.-$ for pieces $P2$.

The question is how much of each kind of piece should we make to maximize weekly margin?

This will be formalized as follows (canonical formulation):

$$\boxed{\begin{aligned}\underset{\vec{x}}{\text{maximize}}\quad & Z(\vec{x})=1,200x_1+1,000x_2\\ \text{subject to}\quad & \text{A1: }3x_1+4x_2\leq 160\\ & \text{A2: }6x_1+3x_2\leq 180\\ & x_1,x_2\geq 0\end{aligned}} \tag{14.1.1318}$$

14.1.17.1.1 Graphical LP resolution

The graphical method is will adapted for problem with 2 or 3 variables but not more as our perception of hyper-volume is quite limited for humans.

When translate our optimization problem into graphical form, we speak also of "polygon of constraints". Indeed, the economical constraints are represented by half planes. The solutions, if they exists, belongs to the intersection set name "set of admissible solutions" and is quite trivially represented in our case by:

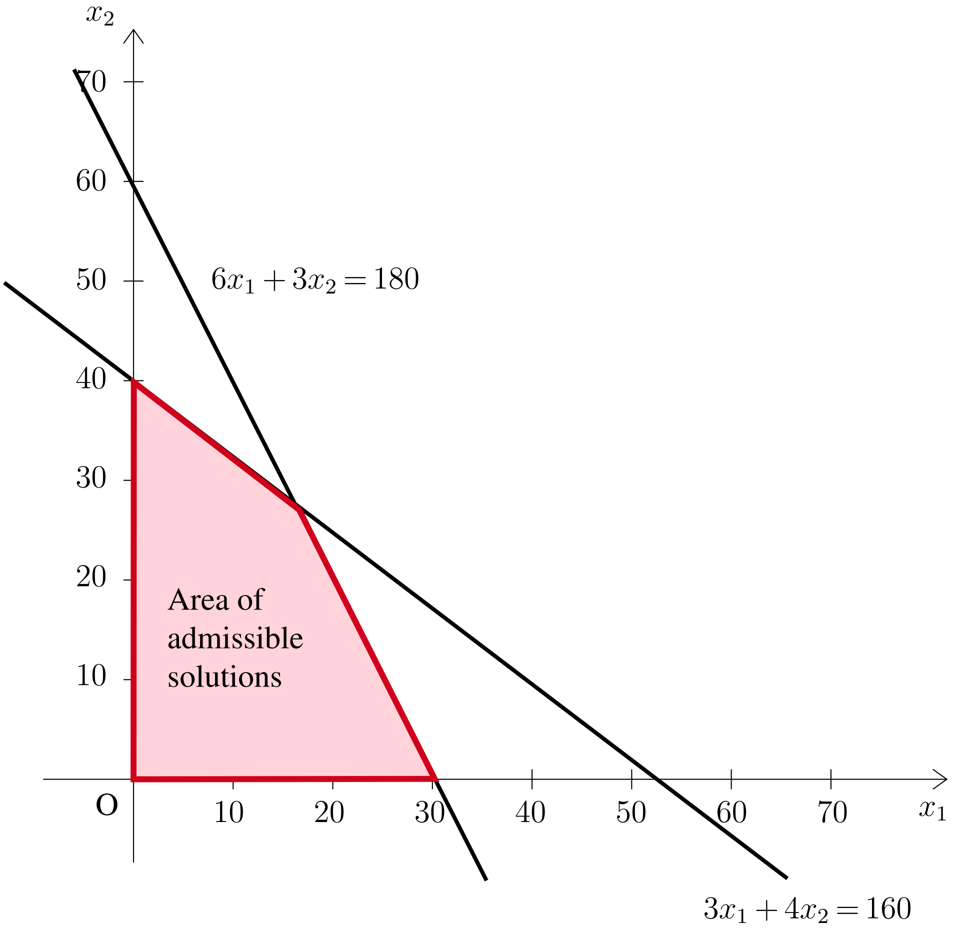


Figure 14.79: **Illustration of a simple operational research problem with area of feasible solutions (convex polytope)**

Remark

In the general case, for those who love the language of mathematicians ..., the information of a linear constraint geometrically problem corresponds to a half space of n-dimensional space (n being the number of variables). In the elementary case, all the points in space that satisfy all constraints are limited by a convex portion of a hyperplane (see the case with 2 variables, easy more to illustrate...), that is why this type of problems are also named "convex optimization". If the cost function is linear, the extreme point is a vertices (easy to see in a two dimensional case). The basic algorithm simplex algorithm (see further below) start from one vertices and goes to the next vertices which locally maximizes the cost, and restarts the procedure as long as necessary.

To find the coordinates of the vertices, we can use the graph if the points are easy to determine.

It is therefore to seek inside this area (connex), the pair (x_1, x_2) maximizing the economic function.

However, the equation Z is represented by a constant line of constant slope (-1.2) which all points (x_1, x_2) provide the same Z value for the objective function.

In particular, the straight line $1200x_1 + 1000x_2$ pass trough the origin and it provides a zero value to the economic function. To increase the value of Z and therefore of the economic function, we have to take away from the origin (in the quarter $x_1 \geq 0, x_2 \geq 0$ that is to say in the `Ist` quadrant) the line of slope -1.2. Obviously then we see very quickly that the simplex method will not work if the constraints of the polygon does not contain the origin point!

To meet the constraints, this straight line will be moved until the limit where it will not have a point of intersection anymore in common with the area of admissible solutions (eventually a segment).

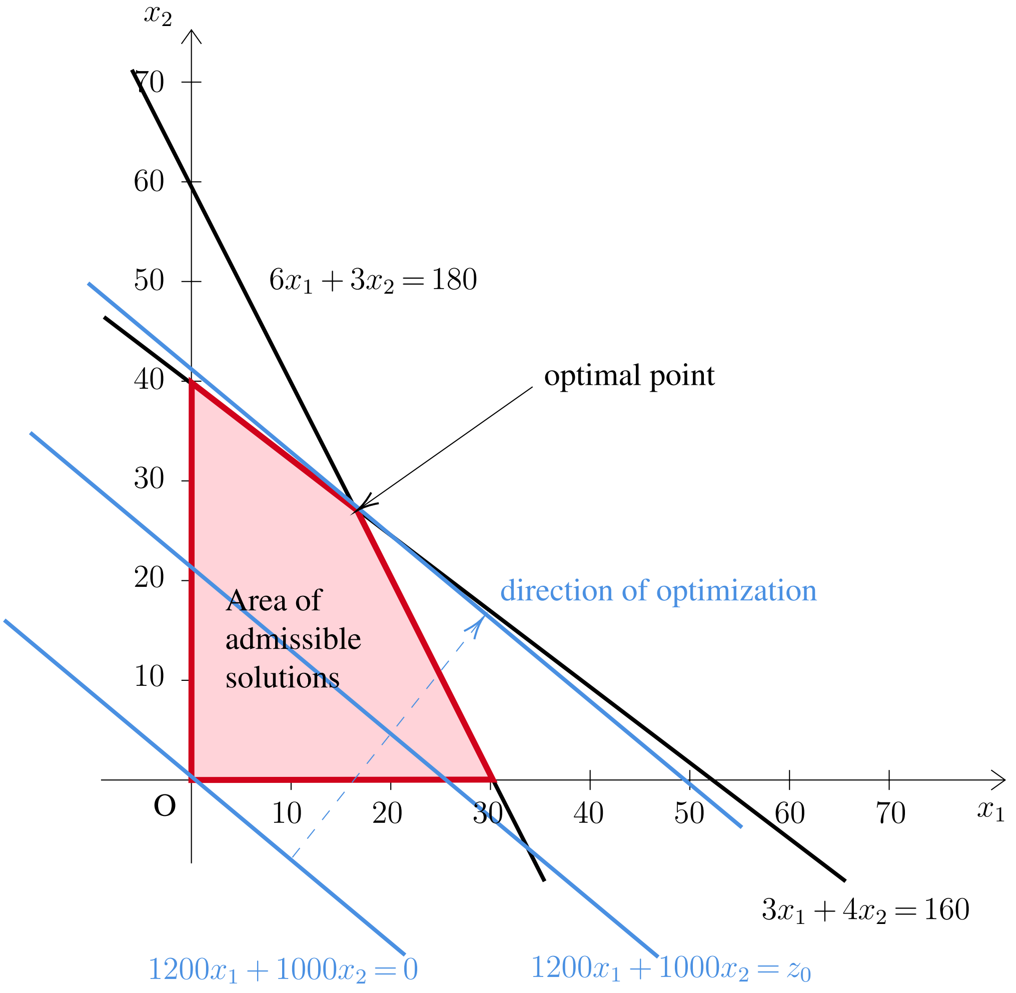

x_2
70
60
50
40
30
20
10
O
10
20
30
40
50
60
70
x_1
$6x_1 + 3x_2 = 180$
optimal point
direction of optimization
Area of admissible solutions
$3x_1 + 4x_2 = 160$
$1200x_1 + 1000x_2 = 0$
$1200x_1 + 1000x_2 = z_0$

Figure 14.80: **Finding solutions graphically with the economic function**

The optimal solution is therefore necessarily located on the periphery of the region of admissible solutions and the parallel formed by translating the economic function are named "isoquants lines" or "isocost lines"...

14.1.17.1.2 Algebraic LP resolution

Let us now see how to solve this problem analytically before moving to the theoretical part.

So we have the "canonical system":

$$\begin{aligned}\text{A1: } & 3x_1+4x_2\leq 160\\ \text{A2: } & 6x_1+3x_2\leq 180\\ & x_1,x_2\geq 0\end{aligned} \tag{14.1.1319}$$

with:

$$Z=1200x_1+1000x_2 \tag{14.1.1320}$$

We first introduce the "slack variables" to transform the 2 inequalities in equalities. The system of equations takes then "standard form":

$$\begin{aligned}3x_1+4x_2+ax_3&=160\\ 6x_1+3x_2+bx_4&=180\\ 1200x_1+1000x_2&=Z\end{aligned} \tag{14.1.1321}$$

Therefore, for $x_1,x_2\geq 0$ fixed, the slack variable whose coefficients are always unit, measure the distance to travel to reach the vertices.

It goes without saying that the technique of slack variables may be used for linear (or non-linear) systems. Therefore, a constraint optimization system with inequalities, can always be reduced to an optimization system with equalities.

Remark

Obviously there is as much slack variable as inequalities.

For the remaining part, we have noticed, after a review of this section, that the technique using tables (that we will see later) often presented in textbooks and websites finally brought nothing to a deep understanding of the resolution mechanism. Since the purpose of the present book is to always prove with a maximum of details the operating principle of a process, it goes without saying that we will opt first for a purely algebraic approach! Let us see it by returning to the system with the slack variables and the economic function but slightly rearranged:

$$\begin{aligned}x_3&=160-(3x_1+4x_2)\\ x_4&=180-(6x_1+3x_2)\\ Z&=1200x_1+1000x_2\end{aligned} \tag{14.1.1322}$$

The $A1$ constraint then becomes:

$$x_3\geq 0 \tag{14.1.1323}$$

and the constraint $A2$ respectively becomes:

$$x_4\geq 0 \tag{14.1.1324}$$

Therefore, the problem consist to maximize Z with the constraints:

$$x_1, x_2, x_3, x_4 \geq 0 \tag{14.1.1325}$$

Let us start with an obvious feasible solution given the constraints that is trivially:

$$x_1 = x_2 = 0 \tag{14.1.1326}$$

Therefore with the system:

$$\begin{aligned} x_3 &= 160 - (3x_1 + 4x_2) \\ x_4 &= 180 - (6x_1 + 3x_2) \\ Z &= 1200x_1 + 1000x_2 \end{aligned} \tag{14.1.1327}$$

we find immediately:

$$\begin{aligned} x_3 &= 160 \\ x_4 &= 180 \\ Z &= 0 \end{aligned} \tag{14.1.1328}$$

The parameters in the actual state can be summarized as:

$$x_1 = 0, x_2 = 0, x_3 = 160, x_4 = 180 \tag{14.1.1329}$$

To go forward, the goal will be to make Z grow and for this purpose we will increase only one single variable, choosing the one with the largest coefficient (weight) in:

$$Z = 1200x_1 + 1000x_2 \tag{14.1.1330}$$

that is to say x_1 (because implicitly we think this is how the Z will increase the faster). We speak then of x_1 as the "pivot direction". We keep then $x_2 = 0$ and we increase x_1 with the system which then reduces to:

$$\begin{aligned} x_3 &= 160 - 3x_1 \\ x_4 &= 180 - 6x_1 \\ Z &= 1200x_1 \end{aligned} \tag{14.1.1331}$$

Therefore with $x_2 = 0$ and to begin $x_1 = 1$, we have:

$$\begin{aligned} x_3 &= 160 - 3 = 157 \\ x_4 &= 180 - 6 = 154 \\ Z &= 1200 \end{aligned} \tag{14.1.1332}$$

and we see that the constraints $x_1, x_2, x_3, x_4 \geq 0$ are still respected, it is the same if x_1 is equal to 2, 3, 4, 5, ... and this until 31, because after:

$$\begin{aligned} x_3 &= 160 - 3 \cdot 31 = 67 \\ x_4 &= 180 - 6 \cdot 31 = -6 \\ Z &= 1200 \cdot 31 = 37200 \end{aligned} \tag{14.1.1333}$$

and one of the slack variable has become negative, the constraints $x_1, x_2, x_3, x_4 \geq 0$ are not all met and therefore this solution is not feasible.

The question in the general case is to ask ourselves until what value (the most constraint value, verbatim the smallest) we can increase x_1 while maintaining the condition $x_1, x_2, x_3, x_4 \geq 0$ when $x_2 = 0$? And the answer is quite simple:

$$\begin{aligned} x_3 &= 160 - 3x_1 \Rightarrow x_1 = \frac{160}{3} \\ x_4 &= 180 - 6x_1 \Rightarrow x_1 = \frac{180}{6} = 30 \end{aligned} \tag{14.1.1334}$$

and therefore it is:

$$x_1 = 30 \tag{14.1.1335}$$

then we speak sometimes of the "pivot step". We then have the actual solution:

$$x_1 = 30, x_2 = 0, x_3 = 70, x_4 = 0 \tag{14.1.1336}$$

Which gives:

$$Z = 1200 \cdot 30 = 36000 \tag{14.1.1337}$$

Graphically, this is what we have just done:

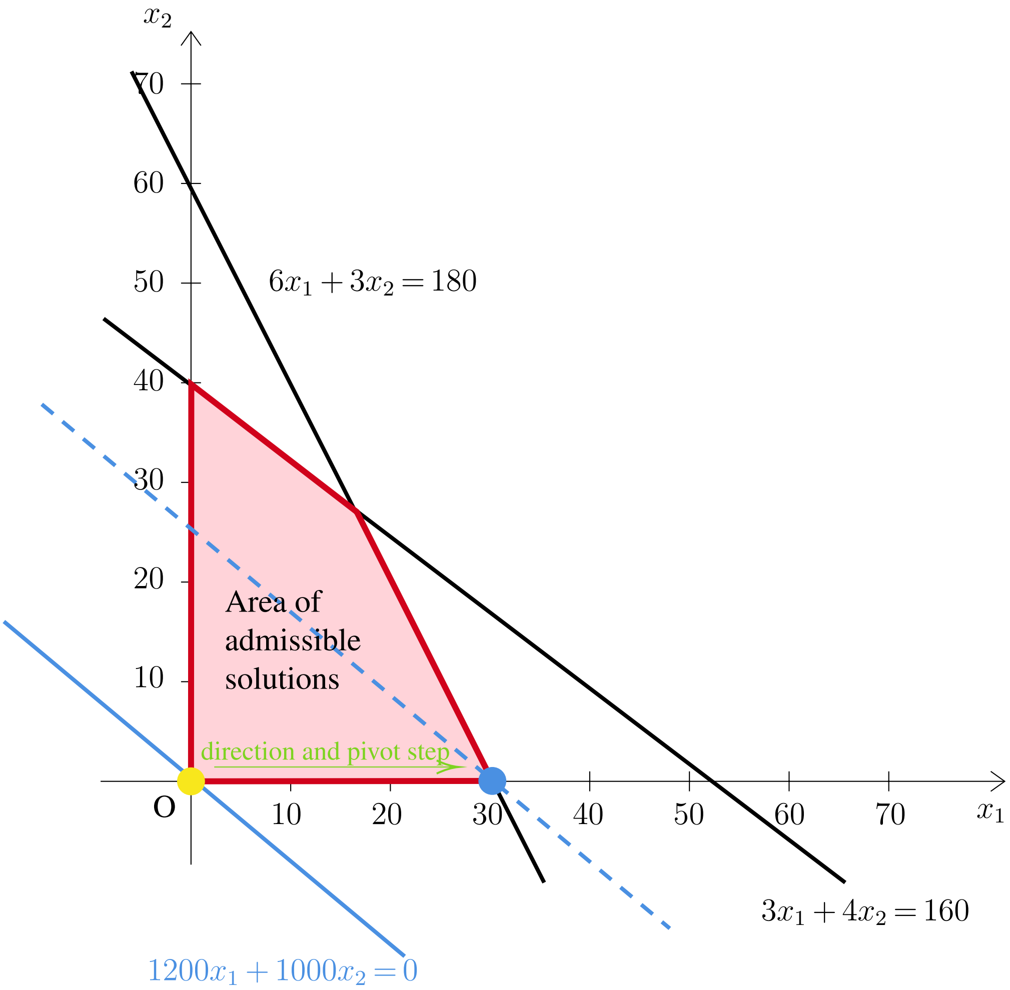


Figure 14.81: Direction of the pivot head and arrival at point $(30, 0)$

To continue to increase Z such simply (by increasing only one variable), we need a new system of equations similar to the original system:

$$\begin{aligned} x_3 &= 160-(3x_1+4x_2)\\ x_4 &= 180-(6x_1+3x_2)\\ Z &= 1200x_1+1000x_2 \end{aligned} \tag{14.1.1338}$$

where we have expressed the variables that takes non-zero value depending on others who take a zero value! That is to say x_3,x_4, in function of x_1,x_2 since we had for recall:

$$x_1=0, x_2=0, x_3=160, x_4=180 \tag{14.1.1339}$$

For the remaining part, we must express x_1,x_3 and also Z in function of x_2,x_4 since we just get for reminder:

$$x_1=30, x_2=0, x_3=70, x_4=0 \tag{14.1.1340}$$

Before getting the new system reaction function, making some algebraic manipulations:

$$\left.\begin{aligned} x_3 &= 160-(3x_1+4x_2)\\ x_4 &= 180-(6x_1+3x_2) \end{aligned}\right\} \Rightarrow 2x_3-x_4=(320-(6x_1+8x_2))-(180-(6x_1+3x_2)) \tag{14.1.1341}$$

which gives after simplification:

$$2x_3-x_4=140-5x_2 \Rightarrow 70-2.5x_2+0.5x_4 \tag{14.1.1342}$$

and therefore it comes:

$$\left.\begin{aligned} x_3 &= 160-(3x_1+4x_2)\\ x_3 &= 70-2.5x_2+0.5x_4 \end{aligned}\right\} \Rightarrow x_3-x_3=(160-(3x_1+4x_2))-(70-2.5x_2+0.5x_4) \tag{14.1.1343}$$

which gives after simplification:

$$x_1=30-0.5x_2-\frac{1}{6}x_4 \tag{14.1.1344}$$

and we have identically:

$$Z=1200x_1+1000x_2=1200\left(30-0.5x_2-\frac{1}{6}x_4\right)+1000x_2=36000+400x_2-200x_4 \tag{14.1.1345}$$

So finally the system is:

$$\begin{aligned} x_1 &= 30-0.5x_2-\frac{1}{6}x_4\\ x_3 &= 70-2.5x_2+0.5x_4\\ Z &= 36000+400x_2-200x_4 \end{aligned} \tag{14.1.1346}$$

from which we reiterate the process (we increase only one variable in Z keeping the other to 0). When we can not increase Z as all coefficients are negative, well it is that we are at a maximum (thank you convexity...). Let us see this...

In Z the biggest coefficient is now x_2 and so it leads us to put $x_4 = 0$. The most constraint value of x_2 that allows to always respect the constraints $x_1, x_2, x_3, x_4 \geq 0$ is therefore:

$$x_2 = 28 \tag{14.1.1347}$$

And for this value, we have:

$$x_1 = 30 - 0.5x_2 = 30 - 0.5 \cdot 28 = 16 \tag{14.1.1348}$$

The original economic function then takes the value:

$$1200x_1 + 1000x_2 = Z \Rightarrow 1200 \cdot 16 + 1000 \cdot 28 = 47200 \tag{14.1.1349}$$

which therefore corresponds graphically to:

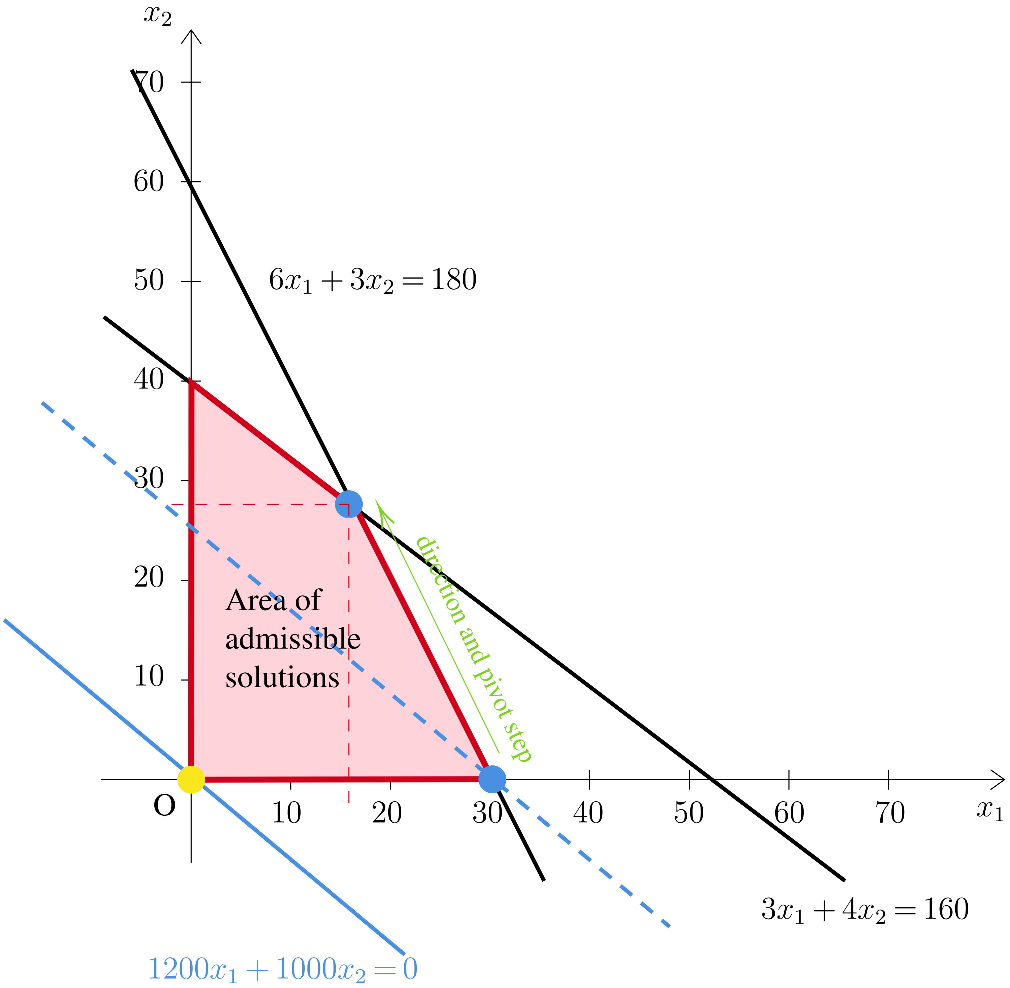


Figure 14.82: **Pivot direction with arrival point at** $(16, 28)$ **for the second iteration**

So we see above that we arrived at the optimum value visible on the graph given at the beginning of this example. But how do we know that we arrived at the final point if we do not have plots or if we work in higher dimensions?

In fact, the process is terminated either when all the coefficients of the economic function are negative or that the most constraint value that respects the constraints is equal to zero !!! Let's see if this is the case! We therefore have in our example:

$$x_1=16, x_2=28, x_3=0, x_4=0 \tag{14.1.1350}$$

So we will rewrite the system:

$$\begin{aligned}x_1&=30-0.5x_2-\frac{1}{6}x_4\\x_3&=70-2.5x_2+0.5x_4\\Z&=36000+400x_2-200x_4\end{aligned} \tag{14.1.1351}$$

with this time x_1,x_2 and Z dependant to x_3,x_4. We then have first:

$$x_1=30-0.5\left(\frac{70-x_3+0.5x_4}{2.5}\right)-\frac{1}{6}x_4=16+\frac{1}{5}x_3-\frac{4}{15}x_4 \tag{14.1.1352}$$

and we have:

$$\begin{aligned}x_1=30-0.5x_2-\frac{1}{6}x_4\Rightarrow x_2=60-2x_1-\frac{1}{3}x_4&=60-2\left(16+\frac{1}{5}-\frac{4}{15}x_4\right)-\frac{1}{3}x_4\\&=28-\frac{2}{5}x_3+\frac{1}{5}x_4\end{aligned} \tag{14.1.1353}$$

Therefore for Z we have (the coefficients are all negative so we guess what comes ...):

$$Z=3600+400\left(28-\frac{2}{5}x_3+\frac{1}{5}x_4\right)-200x_4=47200-160x_3-120x_2 \tag{14.1.1354}$$

We then have the new system:

$$\begin{aligned}x_1&=16+\frac{1}{5}x_3-\frac{4}{15}x_4\\x_2&=28-\frac{2}{5}x_3+\frac{1}{5}x_4\\Z&=47200-160x_3-120x_4\end{aligned} \tag{14.1.1355}$$

As all coefficients of Z are now negative, we're blocked because we would go in the wrong direction if we continue. We must therefore stop here and we adopt finally the solution:

$$\begin{aligned}x_1&=16\\x_2&=28\end{aligned} \tag{14.1.1356}$$

The method of resolution using tables that is often presented in the literature is only useful to write the coefficients of the variables of the system a table, but the changes that we made are exactly those we just made algebraically before (but a the opposite of the tables the method we used don't hide the logic of the method).

14.1.17.1.3 Simplex algorithm LP resolution

To implement the simplex algorithm, we must write the problem in a "standard" form and introduce the concept of "base program" that is the algebraic expression corresponding to the notion of "extreme point of the polyhedron of eligible programs" presented earlier above. Indeed, we will see that the solution of a problem of the linear programming type it exists, can still be obtained with a base program. The simplex method will therefore be to find a first base program and to build a following base programs constantly improving the economic function and thus leading to the optimum (this is what we name "dynamic programming").

An LP problem is said to be placed in its "standard form" if it involves the search for the minimum of the objective function, the latter being subject to constraints in the form of linear equations and conditions of non-negativity of the variables, that is, we can write it in the form seen earlier before:

$$\boxed{\begin{aligned} &\underset{\vec{x}}{\text{minimize}} \quad Z(\vec{x})=\sum_{j=1}^{n} c_i x_i \\ &\text{subject to} \quad \sum_{i=1}^{n} a_{ij}x_j=b_i \quad i=1\ldots m \\ &\qquad\qquad\qquad\qquad x_j\geq 0 \quad j=1\ldots n \end{aligned}} \tag{14.1.1357}$$

Or using matrix notation:

$$\boxed{\begin{aligned} &\underset{\vec{x}}{\text{minimize}} \quad \vec{c}^T\vec{x}=z \\ &\text{subject to} \quad A\vec{x}=\vec{b} \\ &\qquad\qquad\quad\; \vec{x}\geq\vec{0} \end{aligned}} \tag{14.1.1358}$$

where the matrices $C:n\times 1, A:m\times n, b:m\times 1$ respectively correspond to the activity coefficients of objective function, to the technical coefficients of activities and to second members of the constraints.

We will now see how a LP general problem can always be reduced to a standard form. As we guess it the concept of "slack variable" will be essential to perform this "reduction".

Find the maximum of a function $f(x)$ is equivalent to find the minimum of the opposite sign function $f(x)$. Moreover, a constraint which is presented as an inequality:

$$\sum_{j=1}^{p} a_{ij}x_j\leq b_i \tag{14.1.1359}$$

can be replaced by the following system:

$$\sum_{j=1}^{p} a_{ij}x_j+s_i=b_i \tag{14.1.1360}$$

where as we already know s_i is the slack variable constraint such that $s_i\geq 0$.

Of course, if the system is such that:

$$\sum_{j=1}^{p} a_{ij}x_j \geq b_i \tag{14.1.1361}$$

can be replaced by the system:

$$\sum_{j=1}^{p} a_{ij}x_j - e_i = b_i \tag{14.1.1362}$$

implying again to add a slack variable and always with the constraint that $e_i \geq 0$.

This work of putting in standard form work us to find a system of linear equations to solve system (we saw previously at the beginning of this section how to solve this kind of system with the pivot algorithm).

The matrix A representing the components of the system of equations can be, as we know, expressed in different ways depending on the chosen vector basis (see section Vector Calculus page 1809). We will introduce now the concept of "canonical associated usable form" by choosing a special base and show that this reformulation of constraint system will enable us to move towards the optimum.

The matrix A can, after introduction of the slack variables be decomposed into two sub-matrices $[D|B]$, one containing the initial variables D and the other with the slack variables B such that:

$$A = [D|B] \tag{14.1.1363}$$

Remark

The slack variables are... variable and not constant!! In a system where the variables are in quantity n and the equations in quantity m written as:

$$a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n \leq b_i \tag{14.1.1364}$$

We add a slack variable to change the latter into:

$$a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n + x_{n+1} = b_i \tag{14.1.1365}$$

where $x_{n+1} = e_i$ on each row m. The added slack variable being different of all variables already existing in the system! This is why we can decompose the initial matrix in two sub matrices.

The columns of the matrix B are obviously, by definition of the method, units columns, linearly independent. These columns form a basis of the vector space of columns of m elements (or dimensions) - the number of lines of the system. We name B the "base matrix".

Right now it can seem a little bit confusing. So let us continue the theory with a companion example as we have for habit to do it in this textbook.

The text below describing the simplex algorithm has been taken to Marcel Oliver (April 12, 12012 according to holocene calendar) and formalized by our own work.

1. Step 1: Write the linear programming problem in standard form

   Turning a problem into standard form involves the following steps.

   (a) Turn Maximization into minimization and write inequalities in standard order.

   This step is obvious. Multiply expressions, where appropriate, by -1.

   (b) Introduce slack variables to turn inequality constraints into equality constraints with non-negative unknowns.

   Any inequality of the form

$$a_1\,x_1+\cdots+a_n\,x_n\leq b \tag{14.1.1366}$$

   can be replaced by:

$$a_1\,x_1+\cdots+a_n\,x_n+s=b \tag{14.1.1367}$$

   with $s\geq 0$.

   (c) Replace variables which are not sign-constrained by differences.

   Any real number x can be written as the difference of non-negative numbers $x=u-v$ with $u,v\geq 0$.

   Consider the following example.

$$\max Z=x_1+2\,x_2+3\,x_3 \tag{14.1.1368}$$

   subject to:

$$\begin{aligned}x_1+x_2-x_3&=1\\-2\,x_1+x_2+2\,x_3&\geq -5\\x_1-x_2&\leq 4\\x_2+x_3&\leq 5\\x_1&\geq 0\\x_2&\geq 0\end{aligned} \tag{14.1.1369}$$

   Written in standard form, the problem becomes:

$$\min Z=-x_1-2\,x_2-3\,u+3\,v \tag{14.1.1370}$$

   subject to:

$$\begin{aligned}x_1+x_2-u+v&=1\\2\,x_1-x_2-2\,u+2\,v+s_1&=5\\x_1-x_2+s_2&=4\\x_2+u-v+s_3&=5\\x_1,x_2,u,v,s_1,s_2,s_3&\geq 0\end{aligned} \tag{14.1.1371}$$

2. Step 2: Write the coefficients of the problem into a "simplex tableau".

   The coefficients of the linear system are collected in an augmented matrix as known from Gaussian elimination for systems of linear equations; the coefficients of the objective function are written in a separate bottom row with a zero in the right hand column.

   For our example, the initial tableau reads:

x_1	x_2	u	v	s_1	s_2	s_3	
1	1	-1	1	0	0	0	1
2	-1	-2	2	1	0	0	5
1	-1	0	0	0	1	0	4
0	1	1	-1	0	0	1	5
-1	-2	-3	3	0	0	0	0

So following what we have above, we get for:

$$A=[X|S] \tag{14.1.1372}$$

that:

$$D=\begin{bmatrix}1 & 1 & -1 & 1\\ 2 & -1 & -2 & 2\\ 1 & -1 & 0 & 0\\ 0 & 1 & 1 & -1\end{bmatrix} \quad \text{and} \quad B=\begin{bmatrix}0 & 0 & 0\\ 1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix} \tag{14.1.1373}$$

The variables associated to the column components of the matrix S will now be named "bases variables". In our case the bases variables are then essentially the slack variables that we will write now $x_{n+1},x_{n+2},\ldots,x_{n+m}$. The variables associates to the column of the matrix X are named "off-base variables", these are the variables $x_1,x_2,\ldots,x_n$.

Remark

Let us recall that in the expression of the economic function, only the off-base variables appear.

In the following steps, we will act on the tableau by the rules of Gaussian elimination, where the pivots are always chosen from the columns corresponding to the bases variables.

Before proceeding, we need to choose an initial set of basic variables which corresponds to a point in the feasible region of the linear programming problem. Such a choice may be non-obvious, but we shall defer this discussion for now. In our example, x_1 and $s_1,\ldots,s_3$ shall be chosen as the initial bases variables, indicated by gray columns in the tableau above.

3. Step 3: Gaussian elimination

   For a given set of basic variables, we use Gaussian elimination (see page 6305) to reduce the corresponding columns to a permutation of the identity matrix. This amounts to solving $A\vec{x}=\vec{b}$ in such a way that the values of the non-basic variables are zero and the values for the basic variables are explicitly given by the entries in the right hand column of the fully reduced matrix. In addition, we eliminate the coefficients of the objective function below each pivot.

   Our initial tableau is thus reduced to:

x_1	x_2	$\boldsymbol{u}$	v	s_1	s_2	s_3	
1	1	$\mathbf{-1}$	1	0	0	0	1
0	-3	$\mathbf{0}$	0	1	0	0	3
0	-2	***1***	*– 1*	*0*	*1*	*0*	*3*
0	1	$\mathbf{1}$	-1	0	0	1	5
0	-1	$\mathbf{-4}$	4	0	0	0	1

The solution expressed by the tableau is only admissible if all basic variables are non-negative, i.e., if the right hand column of the reduced tableau is free of negative entries. This is the case in this example. At the initial stage, however, negative entries may come up; this indicates that different initial basic variables should have been chosen. At later stages in the process, the selection rules for the basic variables will guarantee that an initially feasible tableau will remain feasible throughout the process.

4. Step 4: Choose new basic variables

   If, at this stage, the objective function row has at least one negative entry, the cost can be lowered by making the corresponding variable basic. This new basic variable is named the "entering variable". Correspondingly, one formerly basic variable has then to become non-basic, this variable is named the "leaving variable". We use the following standard selection rules.

   (a) The entering variable shall correspond to the column which has the most negative entry in the cost function row. If all cost function coefficients are non-negative, the cost cannot be lowered and we have reached an optimum. The algorithm then terminates.

   (b) Once the entering variable is determined, the leaving variable shall be chosen as follows. Compute for each row the ratio of its right hand coefficient to the corresponding coefficient in the entering variable column. Select the row with the smallest finite positive ratio. The leaving variable is then determined by the column which currently owns the pivot in this row. If all coefficients in the entering variable column are non-positive, the cost can be lowered indefinitely, i.e., the linear programming problem does not have a finite solution. The algorithm then also terminates.

If entering and leaving variable can be found, go to Step 3 and iterate.

Note that choosing the most negative coefficient in rule (i) is only a heuristic for choosing a direction of fast decrease of the objective function. Rule (ii) ensures that the new set of basic variables remains feasible.

Let us see how this applies to our problem. The previous tableau holds the most negative cost function coefficient in column 3, thus u shall be the entering variable (marked in boldface). The smallest positive ratio of right hand column to entering variable column is in row 3, as $\frac{3}{1} < \frac{5}{1}$. The pivot in this row points to s_2 as the leaving variable. Thus, after going through the Gaussian elimination once more, we arrive at:

x_1	$\boldsymbol{x_2}$	u	v	s_1	s_2	s_3	
1	**−1**	0	0	0	1	0	4
0	**−3**	0	0	1	0	0	3
0	**−2**	1	−1	0	1	0	3
0	***3***	***0***	*0*	*0*	*−1*	*1*	*2*
0	**−9**	0	0	0	4	0	13

At this point, the new entering variable is x_2 corresponding to the only negative entry in the last row, the leaving variable is s_3. After Gaussian elimination, we find:

x_1	x_2	u	v	s_1	s_2	s_3	
1	0	0	0	0	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{14}{3}$
0	0	0	0	1	-1	1	5
0	0	1	-1	0	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{13}{3}$
0	1	0	0	0	$-\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{3}$
0	0	0	0	0	1	3	19

Since there is no more negative entry in the last row, the cost cannot be lowered by choosing a different set of basic variables; the termination condition applies.

5. Step 5: Read off the solution

   The solution represented by the final tableau has all non-basic variables set to zero, while the values for the basic variables can be can be read off the right hand column. The bottom right corner gives the negative of the objective function.

   In our example, the solution reads $x_1=\frac{14}{3}$, $x_2=\frac{2}{3}$, $x_3=u=\frac{13}{3}$, $s_1=5$, $v=s_2=s_3=0$, which corresponds to $\zeta=-19$, which can be independently checked by plugging the solution back into the objective function.

   As a further check, we notice that the solution must satisfy the initial equation and inequations. This can obviously be checked by direct computation.

In summary, any LP once put in standard form is such that:

- There is a square sub-matrix of matrix A, which is named the "base matrix" and is equal to the square unit matrix $\mathbb{1}$ of size m (indeed there are as many slack variables that lines in the original equations - at the number of m - and as many columns as each slack variable has a different index).
- The basic variables involved does not appear in the expression of the economic function.
- The second member of the constraints consists of non-negative values.

We say then that the problem is put under a "a canonical form associated with the basis B, corresponding to the basis variables $x_{n+1},x_{n+2},\ldots,x_{n+m}$".

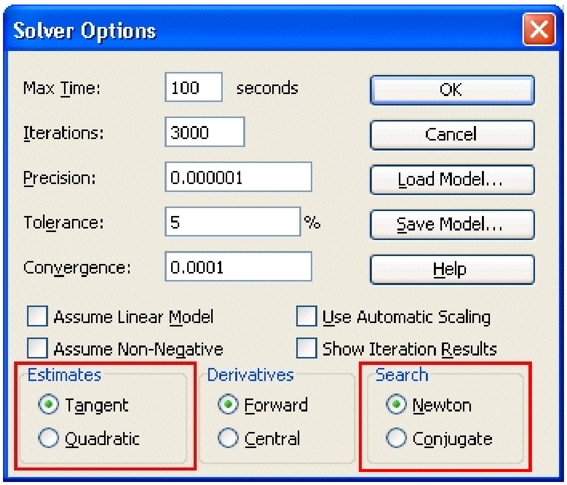
14.1.17.2 Nonlinear programming (Nonlinear optimization)

Nonlinear programming is the process of solving an optimization problem defined by a system of equalities and inequalities, collectively termed constraints, over a set of unknown real variables, along with an objective function to be maximized or minimized, where some of the constraints or the objective function are non-linear.

A non-linear optimization program (NLOP) is a generalization of linear programming (simplex algorithm) but about non-linear functions and can also include non-linear constraints and non-linear economic functions.

The purpose of what follows is to understand in outline but with an acceptable level of rigour the optimization tools that offer many spreadsheets softwares like the previous versions of

Microsoft® Excel® to the version 2007 (since the version 2007 we cannot make a fine-tuning of these options anymore):



Max Time: 100 seconds

Iterations: 3000

Precision: 0.000001

Tolerance: 5 %

Convergence: 0.0001

☐ Assume Linear Model ☐ Use Automatic Scaling

☐ Assume Non-Negative ☐ Show Iteration Results

Estimates: ● Tangent ○ Quadratic

Derivatives: ● Forward ○ Central

Search: ● Newton ○ Conjugate

Figure 14.83: **Microsoft® Excel® 2003 Solver options**

We will especially see now in what consist the *Newton* Search (meaning implicitly: "Gauss-Newton's method") with the *Tangent* and *Quadratic* estimates. After which we will study also the Conjugate Gradients Search also with the tangent and quadratic methods respectively.

Remark

We will stop at the study of the above cited models because there is an excessive quantity of empirical models such as for example the best known models (algorithms): substitution method, method of Lagrange multipliers, Nelder-Mead algorithm, Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, algorithm of simultaneous annihilation (SA), methods of interior points... and see Wikipedia for a more complete list (there are over a dozen of methods without taking into account the variations including empirical adjustments).

We will see it further below, but we already guess that the choice *Tangent* use a linear approximation (tangent) of the function to be optimized at the point considered when at the opposite the *Quadratic* option will make an estimation of a function of the second degree at the considered point (typically a parabola). If at the considered point, the function is well modelled by a quadric, then the *Quadratic* option can save time by choosing a better starting point that will require fewer steps on each additional research. If you have no idea of the behaviour of the a priori function, then the *Tangent* option is slower but safer.

A well known example in the literature to introduce the search of optimums of non-linear functions, before moving on to the part taking into account constraints on the system, is the "hump-back whale function" of that consist to find the minimum of:

$$x_1^2\left(4-2.1x^2+\frac{1}{3}x_1^4\right)+x_1x_2+x_2^2(-4+4x_2^2) \qquad (14.1.1374)$$

With the range constraints:

$$\begin{gathered}-2\geq x_1\geq 2\\ -1\geq x_2\geq 1\end{gathered} \qquad (14.1.1375)$$

what we can indeed check visually:

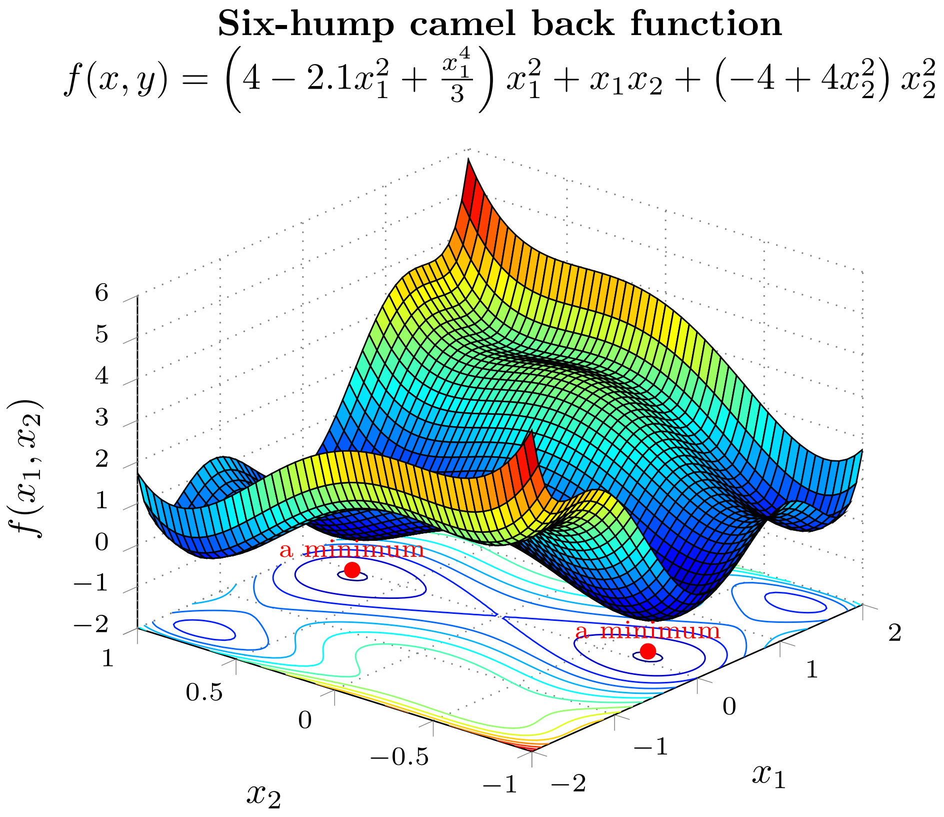
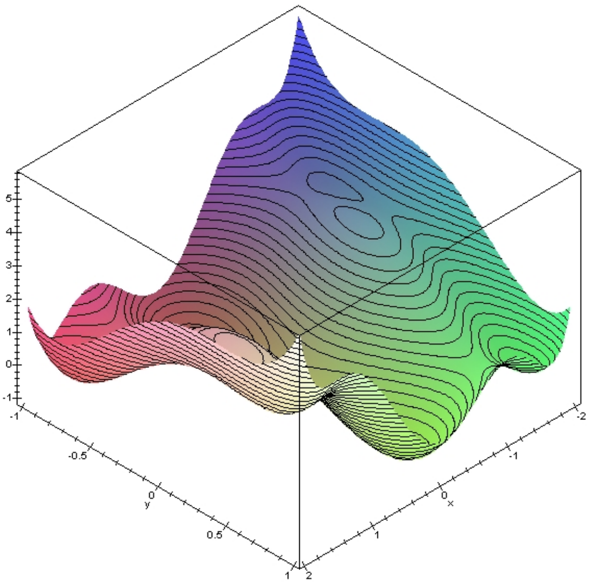


Figure 14.84: Plot of the humpback whale function with minima already visible made with TikZ/pgfplot/gnulplot (authors: Vincent Isoz, Jasper Habicht)

Either with Maple™ 4.00b:

```
>plot3d(x^2*(4-2.1*x^2+1/3*x^4)+x*y+y^2*(-4+4*y^2),
x=-2..2,y=-1..1,contours=20,style=patchcontour,axes=boxed);
```

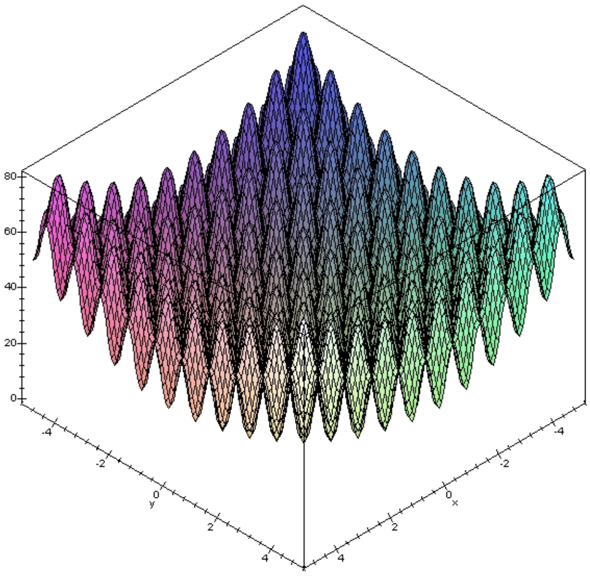
That gives:



As we can see, this function is a great example of multiple local minimum but there is also a must more vicious one that we will refer to when we will study evolutionary algorithms, the "Rastgrini's function":

```
>plot3d(20+x^2+y^2-10*(cos(2*Pi*x)+cos(2*Pi*y)),
x=-5..5,y=-5..5,contours=20,style=patch,axes=boxed,numpoints=10000);
```

That gives:



14.1.17.2.1 Substitution Method

The least complex method for solving a non-linear programming problem is named the "substitution method".

This method is restricted to models containing a single constraint and must be in addition to the equality type.

Let us consider a companion example by considering the following economic function to maximize:

$$\max Z = 4x_1 - 0.1x_1^2 + 5x_2 - 0.2x_2^2 \qquad (14.1.1376)$$

with the constraint (remember that it must be an equality !!!):

$$x_1 + 2x_2 = 40 \qquad (14.1.1377)$$

The first step is then to arbitrarily substitute:

$$x_1 = 40 - 2x_2 \qquad (14.1.1378)$$

into the economic function to get:

$$\begin{aligned}Z &= 4(40-2x_2) - 0.1(40-2x_2)^2 + 5x_2 - 0.2x_2^2\\ &= 160 - 8x_2 - 0.1(1600 - 160x_2 + 4x_2^2) + 5x_2 - 0.2x_2^2\\ &= 160 - 8x_2 - 160 + 16x_2 - 0.4x_2^2 + 5x_2 - 0.2x_2^2\\ &= 13x_2 - 0.6x_2^2\end{aligned} \qquad (14.1.1379)$$

This therefore brings us back to a function to be maximized which is unconstrained, so we can differentiate it and put it as zero. Then it comes:

$$\frac{\partial Z}{\partial x_2} = 13 - 1.2x_2 = 0 \qquad (14.1.1380)$$

that gives:

$$x_2 = 10.8 \qquad (14.1.1381)$$

From this we can deduce immediately that:

$$x_1 = 18.4 \qquad (14.1.1382)$$

By injecting those two values in the economic function, we then have:

$$Z = 4x_1 - 0.1x_1^2 + 5x_2 - 0.2x_2^2 = 70.42 \qquad (14.1.1383)$$

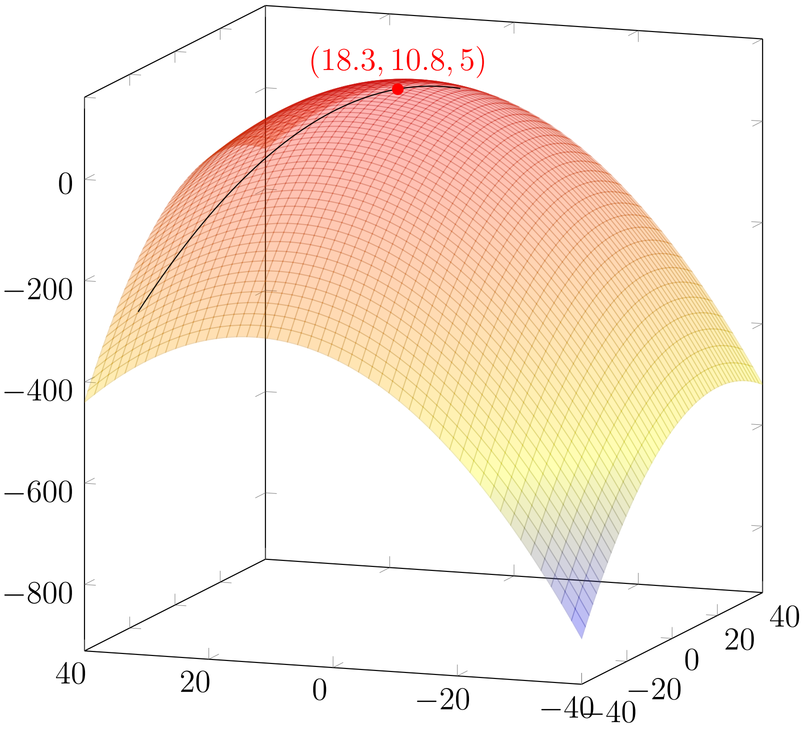


Figure 14.85: **Substitution method example plotted**

We see by this example very quickly the limitations of this technique. The first being that the economic function was not too complex, the final equation to solve was therefore not a problem, the second being that there were only two variables, the third being that the constraint must be an equality as we have already mentioned.

14.1.17.2.2 Lagrange Multipliers Method

The "Lagrange multipliers method" is a rather general technique for solving non-linear programming problems with one or more constraints with linear or non linear economic function with inequalities (rather than strict equations only) and with more than two variables (for other examples than those presented here the reader can go to the corresponding page of Wikipedia).

We will begin to introduce this technique with a simple case which consists of taking again the example used during our study of the method of substitution:

$$\begin{aligned} \underset{\vec{x}}{\text{maximise}} \quad & Z = 4x_1 - 0.1x_1^2 + 5x_2 - 0.2x_2^2 \\ \text{subject to} \quad & x_1 + 2x_2 = 40 \end{aligned} \tag{14.1.1384}$$

The first step is to write the constraint function in Lagrangian form as we do in Analytical Mechanics (see section Analytical Mechanics page 3162) at the difference that there are no general variables depending on the time:

$$\frac{\partial L}{\partial q_i} - \underbrace{\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}_i}}_{=0} = 0 \tag{14.1.1385}$$

For this, we write first:

$$x_1+2x_2-40=0 \tag{14.1.1386}$$

Then the idea is that since this expression is null, nothing prevents us from summing it or subtracting it from the economic function with why not an empirical multiplier that we will denote λ and which we will name the "Lagrange multiplier".

It comes then if we choose to subtract for example (in fact the choice of the subtraction is made by anticipation of an interpretation of the Lagrange multiplier that we will see immediately after):

$$L=4x_1-0.1x_1^2+5x_2-0.2x_2^2-\lambda(x_1+2x_2-40)=0 \tag{14.1.1387}$$

Which gives us our Lagrangian function. In generic form the latter is often written as following:

$$\boxed{L(\vec{x},\lambda)=f(\vec{x})+\lambda(g(\vec{x})-c)} \tag{14.1.1388}$$

Now the idea is to determine the values of the variables where the partial derivatives of the Lagrangian with respect to the variables vanish at the same time (corresponding in Analytical Mechanics as the sum of all the Lagrangians relatively to one variable to be equal to zero):

$$\begin{aligned}\frac{\partial L}{\partial x_1}&=4-0.2x_1-\lambda=0\\ \frac{\partial L}{\partial x_2}&=5-0.4x_2-2\lambda=0\\ \frac{\partial L}{\partial \lambda}&=-x_1-2x_2+40=0\end{aligned} \tag{14.1.1389}$$

What is generally written as:

$$\boxed{\vec{\nabla}_{\vec{x},\lambda}(L(\vec{x},\lambda))=\vec{0}} \tag{14.1.1390}$$

We therefore have a system of three equations with three unknowns which we know trivially how to solve (see section Linear Algebra page 1929). We then get as solutions:

$$\begin{aligned}x_1&=18.3\\ x_2&=10.8\\ \lambda&=0.33\end{aligned} \tag{14.1.1391}$$

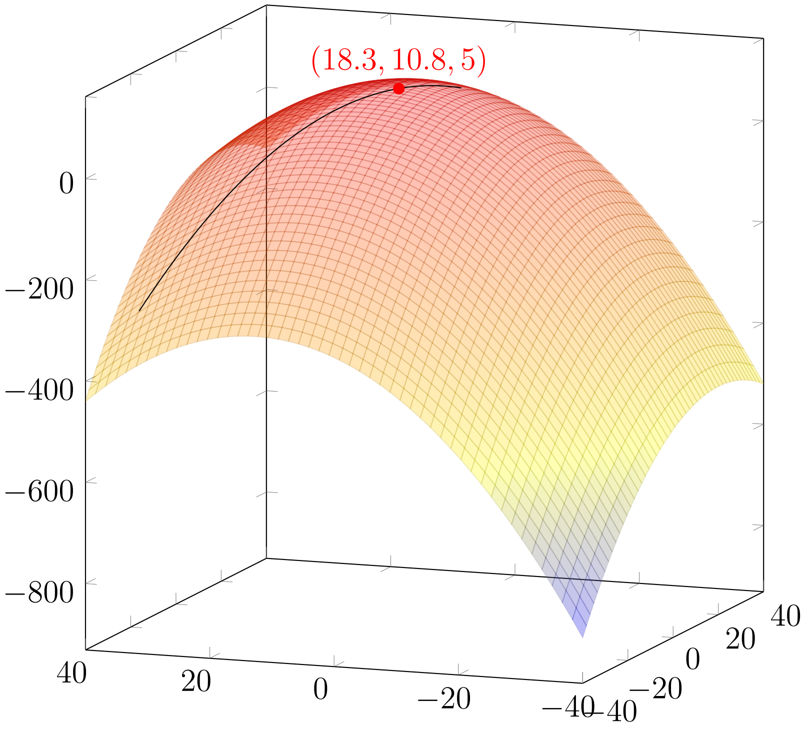


Figure 14.86: Lagrange multiplier example plotted

Notice that if we had not taken into consideration λ and therefore that implicitly that latter had been equal to 1 since the beginning, we would have had the following system of equations:

$$\begin{aligned}\frac{\partial L}{\partial x_1}&=4-0.2x_1^2-1=0\\ \frac{\partial L}{\partial x_2}&=5-0.4x_2-2=0\end{aligned}\tag{14.1.1392}$$

and therefore we would not have obtained the results of the substitution method seen previously ... hence the multiplicative factor!

The final value of Z is then the same as for the substitution method taking into account λ!!!

Although the Lagrange multiplier method is powerful, its increasing complexity with a large number of variables makes it a difficult tool to manipulate in practice.

We will now focus on the interpretation of λ. As we shall see, the latter represents the local variation rate per unit of positive variation of the constant of the constraint function. Thus, if the constraint function becomes (we have changed the 40 into a 41):

$$\begin{aligned}\underset{\vec{x}}{\text{maximise}}\quad & Z=4x_1-0.1x_1^2+5x_2-0.2x_2^2\\ \text{subject to}\quad & x_1+2x_2=41\end{aligned}\tag{14.1.1393}$$

By doing again the same calculations as before, we then get:

$$x_1=18.7\qquad x_2=11.2\qquad \lambda=0.27\qquad Z=70.75\tag{14.1.1394}$$

Therefore, by having increased of one unit the constant of the constrained function, we have:

$$\Delta Z = 0.33 = \lambda \tag{14.1.1395}$$

In general, if the Lagrange multiplier is positive, then the economic function will increase if the constant of the constraint is also incremented positively and vice versa.

Let us now consider a case much more elaborated and useful for our study of Economics (especial Portfolio Optimization). We want to solve by the method of the Lagrange multipliers the Markowitz portfolio problem (see section Economy page 7699) but without the risk free asset (we will do that later case in the section of Economy!):

$$\begin{aligned} \underset{\mathrm{V}(R_P)}{\text{minimize}} \quad & \mathrm{E}(R_P) = c^{te} \\ \text{subject to} \quad & \sum X_i = 1 \end{aligned} \tag{14.1.1396}$$

with for recall:

$$\mathrm{V}(R_P) = X^T c_{ij} X = \sum_{i=1}^{n} \sum_{j=1}^{n} X_i X_j \operatorname{cov}(R_i, R_j) \tag{14.1.1397}$$

To facilitate the developments that will follow we will write this system with the following quite ugly traditional notation (for the details on the notations see the section Economy):

$$\begin{gathered} \min \sigma_P = \sum_{i,j}^{n} w_i w_j \sigma_{ij} = \vec{w}^T \Sigma \vec{w} \\ \sum_{i=1}^{n} w_i \mu_i - \mu_P = \vec{w}^T \vec{\mu} - \mu_P = 0 \\ \sum_{i=1}^{n} w_i - 1 = \vec{w} \cdot \vec{1} - 1 = \vec{w} 1 - 1 = 0 \end{gathered} \tag{14.1.1398}$$

The Lagrangian function will therefore be written by adding the three constraints above:

$$L(\vec{w}, \lambda_1, \lambda_2) = \vec{w}^T \Sigma \vec{w} - \lambda_1 (\vec{w}^T \vec{\mu} - \mu_P) - \lambda_2 (\vec{w}^T 1 - 1) \tag{14.1.1399}$$

Therefore:

$$L(\vec{w}, \lambda_1, \lambda_2) = \vec{w}^T \Sigma \vec{w} + \lambda_1 (\mu_P - \vec{w}^T \vec{\mu}) - \lambda_2 (1 - \vec{w}^T 1) \tag{14.1.1400}$$

Now we calculate:

$$\vec{\nabla}_{\vec{w}, \lambda_1, \lambda_2} (L(\vec{w}, \lambda_1, \lambda_2)) = \vec{0} \tag{14.1.1401}$$

This gives us the system of three equations:

$$\begin{aligned} \frac{\partial L}{\partial \vec{w}} &= 2\Sigma \vec{w} - \lambda_1 \vec{\mu} - \lambda_2 1 = 0 \\ \frac{\partial L}{\partial \lambda_1} &= \mu_P = \vec{w}^T \vec{\mu} = 0 \\ \frac{\partial L}{\partial \lambda_2} &= 1 - \vec{w}^T 1 = 0 \end{aligned} \tag{14.1.1402}$$

By rearranging the first equation:

$$2\Sigma\vec{w}-\lambda_1\vec{\mu}-\lambda_1\vec{\mu}-\lambda_2 1=0 \tag{14.1.1403}$$

We have:

$$\vec{w}=\frac{1}{2}\left(\lambda_1\Sigma^{-1}\vec{\mu}+\lambda_2\Sigma^{-1}1\right) \tag{14.1.1404}$$

Let us now take the two equations:

$$\begin{gathered}\mu_P-\vec{w}^T\vec{\mu}=0\Leftrightarrow\vec{w}^T\vec{\mu}=\mu_P\\ 1-\vec{w}^T1=0\Leftrightarrow\vec{w}^T1=1\end{gathered} \tag{14.1.1405}$$

Thus in an equivalent way:

$$\begin{gathered}\vec{\mu}^T\vec{w}=\mu_P\\ 1^T\vec{w}=1\end{gathered} \tag{14.1.1406}$$

By injecting in it the explicit relation of the weights of the portfolio, we have:

$$\begin{gathered}\vec{\mu}^T\vec{w}=\vec{\mu}^T\left(\frac{1}{2}(\lambda_1\Sigma^{-1}\vec{\mu}+\lambda_2\Sigma^{-1}1)\right)=\frac{1}{2}\left(\lambda_1(\vec{\mu}^T\Sigma^{-1}\vec{\mu})+\lambda_2(\vec{\mu}^T\Sigma^{-1}1)\right)=\mu_P\\ 1^T\vec{w}=1^T\left(\frac{1}{2}(\lambda_1\Sigma^{-1}\vec{\mu}+\lambda_2\Sigma^{-1}1)\right)=\frac{1}{2}\left(\lambda_1(1^T\Sigma^{-1}\vec{\mu})+\lambda_2(1^T\Sigma^{-1}1)\right)=1\end{gathered} \tag{14.1.1407}$$

We can put this system in matrix form:

$$\frac{1}{2}\begin{bmatrix}\vec{\mu}^T\Sigma^{-1}\vec{\mu} & \vec{\mu}^T\Sigma^{-1}1\\ 1^T\Sigma^{-1}\vec{\mu} & 1^T\Sigma^{-1}1\end{bmatrix}\begin{pmatrix}\lambda_1\\ \lambda_2\end{pmatrix}=\begin{pmatrix}\mu_P\\ 1\end{pmatrix} \tag{14.1.1408}$$

Which gives us:

$$\begin{aligned}\begin{pmatrix}\lambda_1\\ \lambda_2\end{pmatrix}&=2\begin{bmatrix}\vec{\mu}^T\Sigma^{-1}\vec{\mu} & \vec{\mu}^T\Sigma^{-1}1\\ 1^T\Sigma^{-1}\vec{\mu} & 1^T\Sigma^{-1}1\end{bmatrix}^{-1}\begin{pmatrix}\mu_P\\ 1\end{pmatrix}\\ &=2\frac{1}{(\vec{\mu}^T\Sigma^{-1}1)-(\vec{\mu}\Sigma^{-1}1)^2)}\begin{bmatrix}\vec{\mu}^T\Sigma^{-1}1 & -1^T\Sigma^{-1}1\\ -1^T\Sigma^{-1}1 & \vec{\mu}\Sigma^{-1}\vec{\mu}\end{bmatrix}\begin{pmatrix}\mu_P\\ 1\end{pmatrix}\end{aligned} \tag{14.1.1409}$$

Therefore:

$$\begin{aligned}\lambda_1&=2\frac{\vec{\mu}^T\Sigma^{-1}1\mu_P-1^T\Sigma^{-1}1}{(\vec{\mu}^T\Sigma^{-1}\vec{\mu})(1^T\Sigma^{-1}1)-(\vec{\mu}^T\Sigma^{-1}1)^2}\\ \lambda_2&=2\frac{-1^T\Sigma^{-1}1\mu_P+\vec{\mu}^T\Sigma^{-1}\vec{\mu}}{(\vec{\mu}^T\Sigma^{-1}\vec{\mu})(1^T\Sigma^{-1}1)-(\vec{\mu}^T\Sigma^{-1}1)^2}\end{aligned} \tag{14.1.1410}$$

So once we have the values of these two Lagrange multipliers, we just have to inject them into:

$$\vec{w}=\frac{1}{2}\left(\lambda_1\Sigma^{-1}\vec{\mu}+\lambda_2\Sigma^{-1}1\right) \tag{14.1.1411}$$

to get the weights that minimize the portfolio variance:

$$\boxed{\vec{w}=\frac{\left(\left(\vec{\mu}^T\Sigma^{-1}1\mu_P-1^T\Sigma^{-1}1\right)\Sigma^{-1}\vec{\mu}+\left(-1^T\Sigma^{-1}1\mu_P+\vec{\mu}^T\Sigma^{-1}\vec{\mu}\right)\Sigma^{-1}1\right)}{(\vec{\mu}^T\Sigma^{-1}\vec{\mu})(1^T\Sigma^{-1}1)-(\vec{\mu}^T\Sigma^{-1}1)^2}} \tag{14.1.1412}$$

So the reader will see during our study of Modern Portfolio management that getting the optimal weights (see page 7699) is much more easy using the Lagrange multiplier method (and less time consuming) than using a software optimizer!

Remark

By the way the attentive reader will perhaps have noticed that finally making all the development with the factor $1/2$ is useless since during the final substitution, the latter is neutralized with itself since 2 is multiplied by $1/2$.

14.1.17.2.3 Newton-Raphson Method (Quadratic Newton)

The "Newton-Raphson method", also named "Newton's method for unconstrained optimization", is a technique for searching the extremum of a function or also, as we will see it when we will compare with a special example the difference between the Gauss-Newton's method with that of Newton, for non-linear regression.

The Newton-Raphson, who in earlier versions to Microsoft® Excel® 2007 was activated in the solver by selecting Newton and Quadratic option uses the second order Taylor approximation (i.e. with second order derivatives) to have a quadratic function (parabola) which converges if the origin point of the research is close to the optimum. This approximation is repeated to each iteration.

To start the formal approach let us recall that we have proved in the section Sequences and Series that a Taylor expansion for a function of two variables could be written in quadratic approximation by:

$$f(x_0+h,y_0+k)=f(x_0,y_0)+\frac{\partial f}{\partial x}(x_0,y_0)h+\frac{\partial f}{\partial y}(x_0,y_0)k+\frac{1}{2}(h \quad k)H\begin{pmatrix}h\\k\end{pmatrix} \tag{14.1.1413}$$

where for recall h and k are variables and x_0, y_0 are fixed and where we have the Hessian matrix (see section Sequences and Series page 1715):

$$\begin{pmatrix}\dfrac{\partial^2 f}{\partial x^2} & \dfrac{\partial^2 f}{\partial x\partial y}\\ \dfrac{\partial^2 f}{\partial x\partial y} & \dfrac{\partial^2 f}{\partial y^2}\end{pmatrix} \tag{14.1.1414}$$

that USA experts in the field have the habit (unfortunate in my opinion ...) to denote:

$$H=\left[\frac{\partial^2 f}{\partial x_i\partial x_j}\right]_{ij}=\left[\vec{\nabla}\right]=\vec{\nabla}^2 f \tag{14.1.1415}$$

the latter expression being the most common can be very misleading with the notation of the Laplacian...

In the field of numerical methods it is customary to write the Taylor series above with few notations changes by putting first:

$$\begin{aligned}&x_0+h=x\Rightarrow h=x-x_0\\&y_0+k=y\Rightarrow k=y-y_0\\&\begin{pmatrix}x_0\\y_0\end{pmatrix}+\begin{pmatrix}h\\k\end{pmatrix}=\begin{pmatrix}x_0+h\\y_0+k\end{pmatrix}=\vec{x}_0+\begin{pmatrix}h\\k\end{pmatrix}=\vec{x}\Rightarrow\begin{pmatrix}j\\k\end{pmatrix}=\vec{x}-\vec{x}_0\\&=\vec{\nabla}=\begin{pmatrix}\partial_x\\\partial_y\end{pmatrix}\end{aligned} \tag{14.1.1416}$$

This gives us a more condensed and technical form of the Taylor series around $\vec{x}$:

$$f(\vec{x})\cong f(\vec{x}_0)+\vec{\nabla}(f(\vec{x}_0))(\vec{x}-\vec{x}_0)+\frac{1}{2}(\vec{x}-\vec{x}_0)^T H(\vec{x}_0)(\vec{x}-\vec{x}_0) \tag{14.1.1417}$$

By changing again a little bit the notations:

$$f(\vec{x}) \cong f(\vec{x}_i) + \vec{\nabla}(f(\vec{x}_i))(\vec{x} - \vec{x}_i) + \frac{1}{2}(\vec{x} - \vec{x}_i)^T H(\vec{x}_i)(\vec{x} - \vec{x}_i) \qquad (14.1.1418)$$

We thus fall back on the usual expression of a function of $\mathbb{R}^2 \rightarrow \mathbb{R}$ evaluated in Taylor series centered on $\vec{x}$.

But if we seek for a local extrema (also sometimes named "critical point"), we will need in first time that the derivative of the whole Taylor series be equal to zero. That is to say:

$$\vec{\nabla}\left(f(\vec{x})\right) = \begin{pmatrix} \partial_x f(\vec{x}) \\ \partial_y f(\vec{x}) \end{pmatrix} \cong \vec{\nabla}\left(f(\vec{x}_i) + \vec{\nabla}(f(\vec{x}_i))(\vec{x} - \vec{x}_i) + \frac{1}{2}(\vec{x} - \vec{x}_i)^T H(\vec{x}_i)(\vec{x} - \vec{x}_i)\right) = \vec{0} \qquad (14.1.1419)$$

and that the determinant of the Hessian matrix is positive (see section Sequences and Series page 1715). And to know if we are on a local maximum or local minimum, we must look at the sign of $\partial_x^2 f(\vec{x})$.

Let us rewrite the above relation explicitly as we proved it in the section of Sequences and Series for pedagogical reasons:

$$\begin{aligned} f(x_0+h, y_0+k) = f(x_0, y_0) &+ \frac{1}{1!}\frac{\partial f}{\partial x}(x_0, y_0)h + \frac{1}{1!}\frac{\partial f}{\partial y}(x_0, y_0)k \\ &+ \frac{1}{2!}\left[\frac{\partial^2 f}{\partial x^2}(x_0, y_0)h^2 + 2\frac{\partial^2 f}{\partial x \partial y}(x_0, y_h)hk + \frac{\partial^2 f}{\partial y^2}(x_0, y_0)k^2\right] \end{aligned} \qquad (14.1.1420)$$

And let us recall that all terms x_0, y_0 are constants because it is either the function f evaluated the particular point (x_0, y_0), or the partial derivative evaluated at the same point, either the partial second derivative always evaluated at the same point, etc.

So finally the gradient will give:

$$\begin{aligned} \vec{\nabla}\left(f(x_0+h, y_0+h)\right) &= \begin{pmatrix} \partial_h f(x_0+h, y_0+h) \\ \partial_k f(x_0+h, y_0+h) \end{pmatrix} \\ &= \begin{pmatrix} \underbrace{\partial_h f(x_0, y_0)}_{=0} + \underbrace{\partial_h(\partial_x f(x_0, y_0)h)}_{=\partial_x f(x_0,y_0)} + \underbrace{\partial_h(\partial_y f(x_0, y_0)k)}_{=0} + \underbrace{\frac{1}{2}\partial_h(\partial_x^2 f(x_0, y_0)h^2)}_{=\partial_x^2 f(x_0,y_0)h} \\ + \underbrace{\frac{1}{2}\partial_h\left(\partial_{xy}^2 f(x_0, y_0)hk\right)}_{=\frac{1}{2}\partial_{xy}^2 f(x_0,y_0)k} + \underbrace{\frac{1}{2}\partial_h\left(\partial_y^2 f(x_0, y_0)k^2\right)}_{=0} \\ \underbrace{\partial_k f(x_0, y_0)}_{=0} + \underbrace{\partial_k(\partial_x f(x_0, y_0)h)}_{=0} + \underbrace{\partial_k(\partial_y f(x_0, y_0)k)}_{=\partial_y f(x_0,y_0)} + \underbrace{\frac{1}{2}\partial_k(\partial_x^2 f(x_0, y_0)h^2)}_{=0} \\ + \underbrace{\frac{1}{2}\partial_k\left(\partial_{xy}^2 f(x_0, y_0)hk\right)}_{=\frac{1}{2}\partial_{xy}^2 f(x_0,y_0)h} + \underbrace{\frac{1}{2}\partial_k\left(\partial_y^2 f(x_0, y_0)k^2\right)}_{=\partial_y^2 f(x_0,y_0)k} \end{pmatrix} \\ &= \begin{pmatrix} \partial_x f(x_0, y_0) + \partial_x^2 f(x_0, y_0)h + \frac{1}{2}\partial_{xy}^2 f(x_0, y_0)k \\ \partial_y f(x_0, y_0) + \partial_y^2 f(x_0, y_0)k + \frac{1}{2}\partial_{xy}^2 f(x_0, y_0)h \end{pmatrix} = \vec{\nabla}\left(f(x_0, y_0)\right) + H\begin{pmatrix} h \\ h \end{pmatrix} \end{aligned} \qquad (14.1.1421)$$

and returning traditional notations in the field of numerical methods, we have then:

$$\vec{\nabla}\left(f(\vec{x})\right)=\vec{\nabla}\left(f(\vec{x}_i)\right)+H(\vec{x}_i)(\vec{x}-\vec{x}_i) \tag{14.1.1422}$$

And so as the gradient has to be equal zero, we have:

$$\vec{\nabla}\left(f(\vec{x}_i)\right)+H(\vec{x}_i)(\vec{x}-\vec{x}_i)=\vec{0} \tag{14.1.1423}$$

and after a first rearrangement:

$$H^{-1}(\vec{x}_i)\vec{\nabla}\left(f(\vec{x}_i)\right)+\vec{x}-\vec{x}_i=\vec{0} \tag{14.1.1424}$$

and a second rearrangement:

$$\boxed{\vec{x}=\vec{x}_i-H^{-1}(\vec{x}_i)\vec{\nabla}\left(f(\vec{x}_i)\right)} \tag{14.1.1425}$$

which is often written:

$$\vec{x}_{i+1}=\vec{x}_i-H^{-1}(\vec{x}_i)\vec{\nabla}\left(f(\vec{x}_i)\right) \tag{14.1.1426}$$

and by USA specialists in the field:

$$\vec{x}_{i+1}=\vec{x}_i-\left[\vec{\nabla}^2 f(\vec{x}_i)\right]^{-1}(\vec{x}_i)\vec{\nabla}\left(f(\vec{x}_i)\right) \tag{14.1.1427}$$

Finally, before moving on to a concrete example it is important that the reader remembers the relation just seen above:

$$\vec{\nabla}\left(f(\vec{x}_i)\right)+H(\vec{x}_i)(\vec{x}-\vec{x}_i)=\vec{0}\Rightarrow-\vec{\nabla}\left(f(\vec{x}_i)\right)=H(\vec{x}_i)(\vec{x}-\vec{x}_i) \tag{14.1.1428}$$

✍**Example:**

We are seeking a local extremum of the "humpback whale" function shown earlier above:

$$f(x,y)=x^2\left(4-2.1x^2+\frac{1}{3}x^4\right)+xy+y^2\left(-4+4y^2\right) \tag{14.1.1429}$$

with the starting point (arbitrary):

$$\vec{x}_0=\begin{pmatrix}1\\1\end{pmatrix} \tag{14.1.1430}$$

To do the search, we calculate the gradient:

$$\vec{\nabla}(f(\vec{x}_i))=\begin{pmatrix}\partial_x f(\vec{x}_i)\\ \partial_y f(\vec{x}_i)\end{pmatrix}=\begin{pmatrix}8x_1-8.4x_i^3+2x_i^5+y_i\\ x_i-8y_i+16y_i^3\end{pmatrix} \tag{14.1.1431}$$

and the Hessian matrix:

$$H(\vec{x}_i)=\begin{bmatrix}\partial_x^2 f(x_i,y_i) & \partial_{xy}^2 f(x_i,y_i)\\ \partial_{xy}^2 f(x_i,y_i) & \partial_y^2 f(x_i,y_i)\end{bmatrix}=\begin{bmatrix}8-25.2x_i^2+10x_i^4 & 0\\ 0 & -8+48y_i^2\end{bmatrix} \tag{14.1.1432}$$

We then have:

$$\vec{\nabla}(f(\vec{x}_0))=\begin{pmatrix}\partial_x f(x_0)\\ \partial_y f(y_0)\end{pmatrix}=\begin{pmatrix}\partial_x f(1)\\ \partial_y f(1)\end{pmatrix}=\begin{pmatrix}9x_0-8.4x_0^3+2x_0^5+y_0\\ x_0-8y_0+16y_0^3\end{pmatrix}=\begin{pmatrix}2.60\\ 9.00\end{pmatrix} \tag{14.1.1433}$$

and:

$$H(\vec{x}_0)=\begin{bmatrix}\partial_x^2 f(x_0,y_0) & \partial_{xy}^2 f(x_0,y_0)\\ \partial_{xy}^2 f(x_0,y_0) & \partial_y^2 f(x_0,y_0)\end{bmatrix}=\begin{bmatrix}8-25.2x_0^2+10x_0^4 & 0\\ 0 & -8+48y_0^2\end{bmatrix}= \begin{bmatrix}-7.20 & 0\\ 0 & 40.00\end{bmatrix} \tag{14.1.1434}$$

and:

$$H^{-1}(\vec{x}_0)=\begin{bmatrix}-7.20 & 0\\ 0 & 40.00\end{bmatrix}^{-1}=\begin{bmatrix}-0.138 & 0\\ 0 & 0.025\end{bmatrix} \tag{14.1.1435}$$

and therefore:

$$\vec{x}_1=\vec{x}_0-H^{-1}\vec{\nabla}(f(\vec{x}_0))=\begin{pmatrix}1\\1\end{pmatrix}\begin{bmatrix}-0.138 & 0\\ 0 & 0.025\end{bmatrix}\begin{pmatrix}2.60\\ 9.00\end{pmatrix}=\begin{pmatrix}1.361\\ 0.775\end{pmatrix} \tag{14.1.1436}$$

and we start again (with less detail):

$$\vec{\nabla}(f(\vec{x}_2))=\begin{pmatrix}-0.01\\ 0.51\end{pmatrix} \tag{14.1.1437}$$

and:

$$\begin{aligned} H(\vec{x}_2) &= \begin{pmatrix} -5.52 & 0 \\ 0 & 12.26 \end{pmatrix} \\ H^{-1}(\vec{x}_2) &= \begin{pmatrix} -0.181 & 0 \\ 0 & 0.081 \end{pmatrix} \end{aligned} \tag{14.1.1438}$$

Therefore:

$$\vec{x}_3 = \vec{x}_2 - H^{-1}(\vec{x}_2)\vec{\nabla}(f(\vec{x}_2)) = \begin{pmatrix} 1.303 \\ 0.608 \end{pmatrix} \tag{14.1.1439}$$

and once again (with again less details):

$$\begin{aligned} \vec{\nabla}(f(\vec{x}_3)) &= \begin{pmatrix} -0.04 \\ 0.03 \end{pmatrix} \\ H^{-1}(\vec{x}_3) &= \begin{bmatrix} -0.167 & 0 \\ 0 & 0.102 \end{bmatrix} \end{aligned} \tag{14.1.1440}$$

and therefore:

$$\vec{x}_4 = \vec{x}_3 - H^{-1}(\vec{x}_3)\vec{\nabla}(f(\vec{x}_3)) = \begin{pmatrix} 1.296 \\ 0.604 \end{pmatrix} \tag{14.1.1441}$$

and again (with even less detail):

$$\vec{x}_5 = \vec{x}_4 - H^{-1}(\vec{x}_4)\vec{\nabla}(f(\vec{x}_4)) = \begin{pmatrix} 1.296 \\ 0.605 \end{pmatrix} \tag{14.1.1442}$$

and values will not move anymore. But if we look at the original graphic where we highlighted the convergence point by a red point:

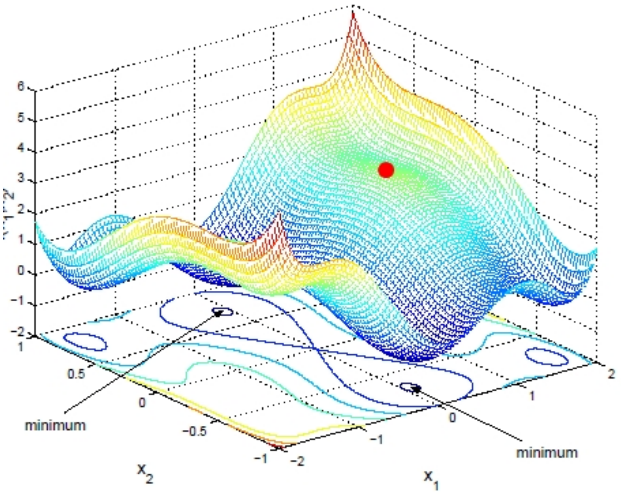


Figure 14.87: **Highlight of the convergence point in the humpback whale function**

we see that this system does not search a global extremum but a local extremum as we already specified it. In fact, as the reader may test itself, convergence is very sensitive to initial starting point.

Keep really in mind that this method (and also the next one) works well only for smooth convex optimization problem (obviously twice differentiable and the opposite of the next method that need to be only once differentiable):

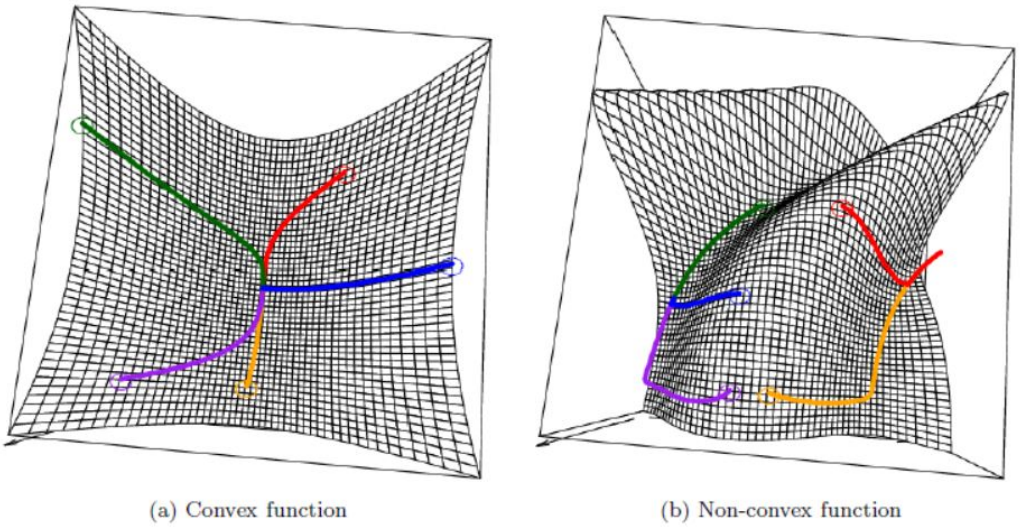

(a) Convex function (b) Non-convex function

Figure 14.88: Optimization paths with different starting points are illustrated in different colours. In the case of strictly convex function (Figure a.), the paths starting from any points all lead to the global optimum. Conversely, in the case of non-convex function, different paths may end up at different local optima!

Here is the corresponding pseudocode:

```
Algorithm 13: Newton-Raphson method (quadratic Newton)
Input: $f:\mathbb{R}^n \rightarrow \mathbb{R}$ a twice-differentiable function
       $\vec{x}_0 \in \mathbb{R}^n$ an initial solution
       $\varepsilon$ is a tolerance value
for $k$ to maxiter do
    Compute $\vec{\nabla} f\left(\vec{x}_k\right)$;
    Compute $H^{-1}\left(\vec{x}_k\right)$;
    Update $\vec{x}_{k+1} \leftarrow \vec{x}_k - H^{-1}\left(\vec{x}_k\right) \vec{\nabla} f\left(\vec{x}_k\right)$;
    if $\left\|f\left(\vec{x}_{k+1}\right) - f\left(\vec{x}_k\right)\right\| < \varepsilon$ or $\left\|\vec{\nabla} f\left(\vec{x}_k\right)\right\| < \varepsilon$ then
        Converged. Print $\vec{x}^* \leftarrow \vec{x}_{k+1}$ and $f\left(\vec{x}^*\right) \leftarrow f\left(\vec{x}_{k+1}\right)$;
```

14.1.17.2.4 Gauss-Newton method's (Tangent Newton)

The Gauss-Newton's method is a powerful approximation without the derivatives of the second order of the Newton-Raphson method that in the prior versions of Microsoft® Excel® 2007 was activated in the solver by selecting the option *Newton* and *Tangent*.

To study this method, let us do use a companion concrete example! Suppose we obtained the following data:

t_i	y_i
1	3.2939
2	4.2699
4	7.1749
5	9.3008
8	20.259

and we suppose "a priori" that the data follow the following theoretical model (we could also try any other function):

$$y_i \cong x_1 e^{x_2 t_i} \qquad (14.1.1443)$$

We look then for x_1, x_2 that minimize the sum of squares (the square of the euclidean distance - L^2 - that is obviously not the only choice of distance in practice!) between the experimental and theoretical values such that:

$$\min_{x_1,x_2} f(x_1,x_2) = \sum_{i=1}^{5} (x_1 e^{x_2 t_i} - y_i)^2 = \sum_{i=1}^{5} (y_i - x_1 e^{x_2 t_i})^2 \qquad (14.1.1444)$$

with therefore:

$$f_i(x_1,x_2) = f_i(\vec{x}) = x_1 e^{x_2 t_i} - y_i \qquad (14.1.1445)$$

The previous relation is also sometimes (especially in the field of Machine Learning) denoted as following (but with a notation quite generalized):

$$\boxed{J(\vec{\omega}) = \frac{1}{2}\sum_i \left(y_i(\vec{x}) - \phi_i(\vec{\omega},\vec{x})\right)^2} \qquad (14.1.1446)$$

and named the "cost function". The $1/2$ is here by anticipation to eliminate a useless factor 2 that will appear later below (as it change nothing to the result of minimization anyway!). The way to put a problem in this form and minimize the cost function is generally named a "least square optimization problem" or in finance and ecology a "model calibration" (that will result to a "calibrated model"...).

Remark

We can also use this method to just minimize a function:

$$\min_{\vec{x}} f(\vec{x}) \qquad (14.1.1447)$$

without using any sum of square. This depends on the problem! The methods remains exactly the same.

Let us write (following the tradition in the field of optimization):

$$F(\vec{x}) = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{pmatrix} \qquad (14.1.1448)$$

We then have the following common notation:

$$\min_{x_1,x_2} f(x_1,x_2)=\sum_i f_i^2 \Rightarrow \min_{\vec{x}} f(\vec{x})=F(\vec{x})^T F(\vec{x}) \qquad (14.1.1449)$$

Now, imagine that we found a bipoint $(x_1,x_2)=(\vec{x})$ that gives this minimum and let us denote it $(\vec{x}_*)$ and without forgetting that it will remain a local minimum and with luck a global one...!

Let us consider a special case that we will name "compatible solution" and define by the fact the bi-point that minimize the sum of squares of errors is also such that for all i we have:

$$f_i(\vec{x}_*)=0 \qquad (14.1.1450)$$

Therefore it is immediate that:

$$F(\vec{x}_*)=\vec{0} \qquad (14.1.1451)$$

Before going further, let us notice for example that for a component j (which corresponds in our case to each variable of the a priori supposed theoretical function of our model):

$$\begin{aligned}\vec{\nabla}\left(F(\vec{x})^T F(\vec{x})\right)_j &= \frac{\partial}{\partial x_j}\sum_i f_i^2=\sum_i f_i^2=\sum_i \frac{\partial f_i^2}{\partial x_j}=\sum_i 2 f_i \frac{\partial f_i}{\partial x_j}=2\sum_i f_i \frac{\partial f_i}{\partial x_j}\\ &=2\sum_i \frac{\partial f_i}{\partial x_j} f_i=2\left(\vec{\nabla}\left(F(\vec{x})^T\right)F(\vec{x})\right)_j\end{aligned} \qquad (14.1.1452)$$

where the last condensed equality is many times less obvious as it makes usage of the gradient of a vector field (see section Vector Calculus page 1878) that we see rarely in practice (note the factor 2 that we were speaking about earlier!). The reader that should be destabilized can refer directly to the numerical example further below to makes things more crystal clear.

✍**Example:**

In the field of Machine Learning, considering (this a liner weighted input of a neural network as we will see later during our study of these objects):

$$z=\omega_0 x_0+\omega_1 x_1+\omega_m x_m=\sum_{j=0}^{m} x_j\omega_j=\vec{\omega}^T\vec{x} \qquad (14.1.1453)$$

where x_0 refers something named a "bias unit" such that $x_0=1$.

Here, our cost function is also the following sum of squared errors (SSE) with the traditional notation of Machine Learning field that fits more when we do detailed developments:

$$J(\vec{w})=\frac{1}{2}\sum_i\left(y^{(i)}-\phi(z)^{(i)}\right)^2 \qquad (14.1.1454)$$

We then have for one component j:

$$\begin{aligned}\frac{\partial J}{\partial \omega_j}&=\frac{\partial}{\partial \omega_j}\frac{1}{2}\sum_i\left(y^{(i)}-\phi(z)^{(i)}\right)^2=\frac{1}{2}\frac{\partial}{\partial \omega_j}\sum_i\left(y^{(i)}-\phi(z)^{(i)}\right)^2\\&=\frac{1}{2}\sum_i 2\left(y^{(i)}-\phi(z)^{(i)}\right)\frac{\partial}{\partial \omega_j}\left(y^{(i)}-\phi(z)^{(i)}\right)=\sum_i\left(y^{(i)}-\phi(z)^{(i)}\right)\frac{\partial}{\partial \omega_j}\left(y^{(i)}-\sum_j(\omega_j^{(i)}x_j^{(i)})\right)\\&=\sum_i\left(y^{(i)}-\phi(z)^{(i)}\right)(-x_j^{(i)})=-\sum_i\left(y^{(i)}-\phi(z)^{(i)}\right)x_j^{(i)}\end{aligned} \tag{14.1.1455}$$

So we have the correspondence for this special (famous) case:

$$\vec{\nabla}\left(J(\vec{\omega})^T J(\vec{\omega})\right)_j=\frac{\partial J}{\partial \omega_j}=-\sum_i\left(y^{(i)}-\phi(z)^{(i)}\right)x_j^{(i)} \tag{14.1.1456}$$

So to continue... we deduce that:

$$\vec{\nabla}\left(\sum_i f_i^2\right)=\vec{\nabla}(f(\vec{x}))=2\vec{\nabla}\left(F(\vec{x})^T\right)F(\vec{x}) \tag{14.1.1457}$$

and the "compatible solution" brings us obviously to:

$$\vec{\nabla}\left(f(\vec{x}_*)\right)=2\vec{\nabla}\left(F(\vec{x}_*)^T\right)F(\vec{x}_*) \tag{14.1.1458}$$

Following the same step, we have:

$$\begin{aligned}\vec{\nabla}^2\left(F(\vec{x}^T)F(\vec{x})\right)=\vec{\nabla}\left(2\sum_i f_i\frac{\partial f_i}{\partial x_j}\right)&=2\vec{\nabla}\left(\sum_i f_i\frac{\partial f_i}{\partial x_j}\right)\\&=2\sum_{i,k}\left(\frac{\partial f_i}{\partial x_k}\frac{\partial f_i}{\partial x_j}+f_i\frac{\partial f_i}{\partial x_k\partial x_j}\right)=2\left(\sum_{i,k}\frac{\partial f_i}{\partial x_k}\frac{\partial f_i}{\partial x_j}+\sum_{i,k}f_i\frac{\partial f_i}{\partial x_k\partial x_j}\right)\\&=2\left(\vec{\nabla}\left(F(\vec{x})^T\right)\left(\vec{\nabla}F(\vec{x})\right)+\sum_i f_i\vec{\nabla}^2 f_i(\vec{x})\right)_j\end{aligned} \tag{14.1.1459}$$

So finally we have the following two relations:

$$\vec{\nabla}\left(f(\vec{x})\right)=2\vec{\nabla}\left(F(\vec{x})^T\right)F(\vec{x})\vec{\nabla}^2\left(F(\vec{x}^T F(\vec{x})\right)=2\left(\vec{\nabla}\left(F(\vec{x})^T\right)\left(\vec{\nabla}F(\vec{x})\right)+\sum_i f_i\vec{\nabla}^2 f_i(\vec{x})\right)_j \tag{14.1.1460}$$

Given that for the "compatible solution" we have:

$$f_i(\vec{x}_*)=0 \tag{14.1.1461}$$

it follows that in this case the second relation becomes:

$$\begin{aligned}\vec{\nabla}^2\left(F(\vec{x}_*)^T F(\vec{x}_*)\right)&\cong 2\left(\vec{\nabla}\left(F(\vec{x}_*)^T\right)\vec{\nabla}(F(\vec{x}_*)+\sum_i\underbrace{f_i(\vec{x}_*)}_{=0}\vec{\nabla}^2 f_i(\vec{x}_x)\right)\\&=2\vec{\nabla}\left(F(\vec{x}_*)^T\right)\vec{\nabla}\left(F(\vec{x}_*)\right)=2[1\quad 1]H\begin{bmatrix}1\\1\end{bmatrix}\end{aligned} \tag{14.1.1462}$$

where H is the Hessian matrix (see section Sequences and Series page 1715) what practitioners in the USA write simply:

$$\vec{\nabla}\left(F(\vec{x}_*)^T\right)\vec{\nabla}\left(F(\vec{x}_*)\right)\cong H(\vec{x}_*) \tag{14.1.1463}$$

So we can approximate in the case of the compatible solution, the Hessian that contains derivatives of the second order by derivatives of the first order and that the main difference between the Gauss-Newton and Newton-Raphson method!!!!

So we have finally in this special case the two relations that are the pillar of the Gauss-Newton's method:

$$\boxed{\begin{gathered}\vec{\nabla}(f(\vec{x}_*)=2\vec{\nabla}(F(\vec{x}_*)^T)F(\vec{x}_*)\\ \vec{\nabla}\left(F(\vec{x}_*)^T\right)\vec{\nabla}(F(\vec{x}_*)\cong H(\vec{x}_*)\end{gathered}} \tag{14.1.1464}$$

Now let us recall the basic relation of the Newton-Raphson method obtained earlier above:

$$-\vec{\nabla}(f(\vec{x}_i))=H(\vec{x}_i)(\vec{x}-\vec{x}_i) \tag{14.1.1465}$$

and for information, any mathematical technique (because they are many of them!) that simplifies the Hessian matrix (see section Sequences and Series page 1715) to the right of equality becomes part of the family named "quasi-Newton's methods".

Well the Gauss-Newton's method that interest us here and is therefore one of the techniques of the family of the "quasi-Newton's methods" consists simply in a first step in getting rid of the second derivatives of the Hessian (see section Sequences and Series page 1715) of the Newton-Raphson method at the right of the equality thanks to the previously established relations such that (attention to remember the abuse of writing!):

$$-\vec{\nabla}\left(f(\vec{x}_i)\right)\cong\vec{\nabla}\left(F(\vec{x}_{i*})^T\right)\vec{\nabla}\left(F(\vec{x}_{i*})\right)(\vec{x}-\vec{x}_{i*}) \tag{14.1.1466}$$

and in a second time rewrite the gradient on the left of the equality thanks also to the previously established relation. Which gives us:

$$\boxed{-2\vec{\nabla}\left(F(\vec{}_*)^T\right)F(\vec{x}_*)\cong\vec{\nabla}\left(F(\vec{x}_{i*})^T\right)\vec{\nabla}\left(F(\vec{x}_{i*})\right)(\vec{x}-\vec{x}_{i*})} \tag{14.1.1467}$$

The factor 2 above being not very aesthetic, almost all textbooks simplify such optimization problem with the following type of notation:

$$\min_{x_1,x_2} f(x_1,x_2)=\frac{1}{2}\sum_{i=1}^{5}\left(x_1e^{x_2t_i}-y_i\right)^2 \tag{14.1.1468}$$

So by simply multiplying by a factor $1/2$ (which does not change the result as we already know!) we have then:

$$\vec{\nabla}\left(F(\vec{}_*)^T\right)F(\vec{x}_*)\cong\vec{\nabla}\left(F(\vec{x}_{i*})^T\right)\vec{\nabla}\left(F(\vec{x}_{i*})\right)(\vec{x}-\vec{x}_{i*}) \tag{14.1.1469}$$

Let us recall again that a spreadsheet software like Microsoft® Excel® can not determine the derivatives it will calculate them using the numerical methods of right or centered derivatives as we have presented a earlier above.

Now let us come back to our example of the beginning! So we have:

$$t=\begin{pmatrix}1\\2\\4\\5\\8\end{pmatrix}\qquad y=\begin{pmatrix}3.2939\\4.2699\\7.1749\\9.3008\\20.259\end{pmatrix}\tag{14.1.1470}$$

We start with a bipoint that seems the closest to the desired solution:

$$x=\begin{pmatrix}x_1\\x_2\end{pmatrix}=\begin{pmatrix}x_{1,0*}\\x_{2,0*}\end{pmatrix}=\begin{pmatrix}2.5\\0.25\end{pmatrix}\tag{14.1.1471}$$

Then we have:

$$F(\vec{x})=\begin{pmatrix}f_1\\f_2\\f_3\\f_4\\f_5\end{pmatrix}=\begin{pmatrix}x_1e^{x_2t_1}-y_1\\x_1e^{x_2t_2}-y_2\\x_1e^{x_2t_3}-y_3\\x_1e^{x_2t_4}-y_4\\x_1e^{x_2t_5}-y_5\end{pmatrix}=\begin{pmatrix}2.5e^{0.25\cdot 1}-3.2939\\2.5e^{0.25\cdot 2}-4.2699\\2.5e^{0.25\cdot 4}-7.17149\\2.5e^{0.25\cdot 5}-9.3008\\2.5e^{0.25\cdot 8}-20.259\end{pmatrix}=\begin{pmatrix}-0.0838\\-0.1481\\-0.3792\\-0.5749\\-1.7864\end{pmatrix}\tag{14.1.1472}$$

and therefore we have:

$$\min_{x_1,x_2} f(x_1,x_2)=\frac{1}{2}\sum_{i=1}^{5}\left(x_1e^{x_2t_i}-y_i\right)^2=1.847\cong 2\tag{14.1.1473}$$

Then it comes:

$$\vec{\nabla}(f(\vec{x}))=\frac{1}{2}\begin{pmatrix}\displaystyle\sum_{i=1}^{5}2\left(x_1e^{x_2t_i}-y_i\right)e^{x_2t_i}\\\displaystyle\sum_{i=1}^{5}2\left(x_1e^{x_2t_i}-y_i\right)x_1t_ie^{x_2t_i}\end{pmatrix}=\begin{pmatrix}\displaystyle\sum_{i=1}^{5}\left(x_1e^{x_2t_i}-y_i\right)e^{x_2t_i}\\\displaystyle\sum_{i=1}^{5}\left(x_1e^{x_2t_i}-y_i\right)x_1t_ie^{x_2t_i}\end{pmatrix}\tag{14.1.1474}$$

What we can therefore rewrite as:

$$\begin{aligned}\vec{\nabla}(f(\vec{x}))&=\vec{\nabla}\left(F(\vec{x})^T\right)F(\vec{x})\\&=\begin{bmatrix}e^{x_2t_1}&e^{x_2t_2}&e^{x_2t_3}&e^{x_2t_4}&e^{x_2t_5}\\x_1t_1e^{x_2t_1}&x_1t_2e^{x_2t_2}&x_1t_3e^{x_2t_3}&x_1t_4e^{x_2t_4}&x_1t_5e^{x_2t_5}\end{bmatrix}\begin{pmatrix}x_1e^{x_2t_1}-y_1\\x_1e^{x_2t_2}-y_2\\x_1e^{x_2t_3}-y_3\\x_1e^{x_2t_4}-y_4\\x_1e^{x_2t_5}-y_5\end{pmatrix}\\&=\begin{bmatrix}1.2840&1.6847&2.718&3.4903&7.3891\\3.2101&8.2435&27.1828&43.6293&147.7811\end{bmatrix}\begin{pmatrix}-0.0383\\-0.1481\\-0.3792\\-0.5749\\-1.7864\end{pmatrix}=\begin{bmatrix}-16.5888\\-300.8722\end{bmatrix}\end{aligned}\tag{14.1.1475}$$

We also have by extension:

$$\begin{aligned}&\vec{\nabla}\left(F(\vec{x})^T\right)\vec{\nabla}(F(\vec{x}))\\&=\begin{bmatrix}e^{x_2t_1} & e^{x_2t_2} & e^{x_2t_3} & e^{x_2t_4} & e^{x_2t_5}\\ x_1t_1e^{x_2t_1} & x_1t_2e^{x_2t_2} & x_1t_3e^{x_2t_3} & x_1t_4e^{x_2t_4} & x_1t_5e^{x_2t_5}\end{bmatrix}\begin{pmatrix}e^{x_2t_1} & x_1t_1e^{x_2t_1}\\ e^{x_2t_2} & x_1t_2e^{x_2t_2}\\ e^{x_2t_3} & x_1t_3e^{x_2t_3}\\ e^{x_2t_4} & x_1t_4e^{x_2t_4}\\ e^{x_2t_5} & x_1t_5e^{x_2t_5}\end{pmatrix}\\&=\begin{bmatrix}78.5367 & 1335.8479\\ 1335.8479 & 24559.9419\end{bmatrix}\end{aligned}\tag{14.1.1476}$$

Then we apply the relation proved earlier above:

$$-\vec{\nabla}\left(F(\vec{x}_*)^T\right)F(\vec{x}_*)\cong\vec{\nabla}\left(F(\vec{x}_{i*})^T\right)(\vec{x}-\vec{x}_{i*})\tag{14.1.1477}$$

Therefore:

$$-\begin{pmatrix}-16.5888\\-300.8722\end{pmatrix}=\begin{bmatrix}78.5367 & 1335.8479\\ 1335.8479 & 24559.9419\end{bmatrix}\begin{pmatrix}x_1-x_{1,0*}\\ x_2-x_{2,0*}\end{pmatrix}\tag{14.1.1478}$$

After a minor simplification:

$$\begin{pmatrix}16.5888\\300.8722\end{pmatrix}=\begin{bmatrix}78.5367 & 1335.8479\\ 1335.8479 & 24559.9419\end{bmatrix}\begin{pmatrix}x_1-x_{1,0*}\\ x_2-x_{2,0*}\end{pmatrix}\tag{14.1.1479}$$

Therefore:

$$\begin{pmatrix}x_1-x_{1,0*}\\ x_2-x_{2,0*}\end{pmatrix}=\begin{pmatrix}x_1\\ x_2\end{pmatrix}-\begin{pmatrix}x_{1,0*}\\ x_{2,0*}\end{pmatrix}=\begin{bmatrix}78.5367 & 1335.8479\\ 1335.8379 & 24559.9419\end{bmatrix}^{-1}\begin{pmatrix}16.5888\\300.8722\end{pmatrix}=\begin{pmatrix}0.0381\\0.0120\end{pmatrix}\tag{14.1.1480}$$

and therefore the next bipoint for the iteration will be:

$$\begin{pmatrix}x_1\\ x_2\end{pmatrix}=\begin{pmatrix}x_{1,1*}\\ x_{2,1*}\end{pmatrix}=\begin{pmatrix}0.0381\\0.0102\end{pmatrix}+\begin{pmatrix}x_{1,0*}\\ x_{2,0*}\end{pmatrix}=\begin{pmatrix}0.0381\\0.0102\end{pmatrix}+\begin{pmatrix}2.5\\0.25\end{pmatrix}=\begin{pmatrix}0.0381\\0.0102\end{pmatrix}+\begin{pmatrix}2.5381\\0.2602\end{pmatrix}\tag{14.1.1481}$$

Hence formally:

$$\boxed{\begin{pmatrix}x_{1,1*}\\ x_{2,1*}\end{pmatrix}=\begin{pmatrix}x_{1,0*}\\ x_{2,0*}\end{pmatrix}-\left[\vec{\nabla}\left(F(\vec{x}_{i*})^T\right)\vec{\nabla}\left(F(\vec{x}_{i*})\right)\right]^{-1}\left(\vec{\nabla}\left(F(\vec{x}_*)^T\right)F(\vec{x}_*)\right)}\tag{14.1.1482}$$

Remark

In the field of Machine Learning, that latter relation is written as following:

$$\boxed{\vec{\omega}^+=\vec{\omega}+\Delta\omega=\vec{\omega}-\eta\vec{\nabla}\left(J(\vec{\omega}_*)^T\right)J(\vec{\omega}_*)=\vec{\omega}-\eta\vec{\nabla}J(\vec{\omega})}\tag{14.1.1483}$$

but where η is fixed (in the field of Machine Learning!) and named the "learning rate" and the whole relation (corresponding to a minimization problem for recall!) is named the "gradient descent method".

Obviously we see that the weight change is then defined by:

$$\boxed{\Delta\omega = -\eta\vec{\nabla}\left(J(\vec{\omega}_*)^T\right) J(\vec{\omega}_*) = -\eta\vec{\nabla}J(\vec{\omega})} \tag{14.1.1484}$$

In component notation, for the component j this gives:

$$\omega_j^+ = \omega_j + \Delta\omega = \omega_j - \eta\frac{\partial}{\partial\omega_j}J(\vec{\omega}) \tag{14.1.1485}$$

In numerical computing this is often denoted:

$$x^{(k)} = x^{(k-1)} - t_k\nabla f\left(x^{(k-1)}\right) \tag{14.1.1486}$$

where $k = 1, 2, \ldots$ is iteration number, t_k is step size (or step length) at iteration k.

Which corresponds well to the values of the first iteration:

$$\begin{pmatrix} x_{1,1*} \\ x_{2,2*} \end{pmatrix} = \begin{pmatrix} 2.5381 \\ 0.2602 \end{pmatrix} \tag{14.1.1487}$$

We will not do again explicitly the other iterations. So this is what we get in the end:

i	$\min_{x_1,x_2} f(x_1,x_2)$
0	$2\cdot 10^0$
1	$4\cdot 10^{-3}$
2	$2\cdot 10^{-8}$
3	$3\cdot 10^{-9}$
4	$3\cdot 10^{-9}$

Table 14.20: **Gauss-Newton iterations of our example**

with therefore for local solution at the 4th iteration:

$$\begin{pmatrix} x_{1,4*} \\ x_{2,4*} \end{pmatrix} = \begin{pmatrix} 2.5411 \\ 0.2595 \end{pmatrix} \tag{14.1.1488}$$

With Maple™ 4.00b we get:

```
>with(plots):
>points:=plot([[1,3.2939],[2,4.2699],[4,7.1749],[5,9.3008],[8,20.259]
],style=point,color=blue,symbol=circle):
>plot_GN:=plot(2.5411*exp(0.2595*x),x=0..8):
>display([pict1,pict2]);
```

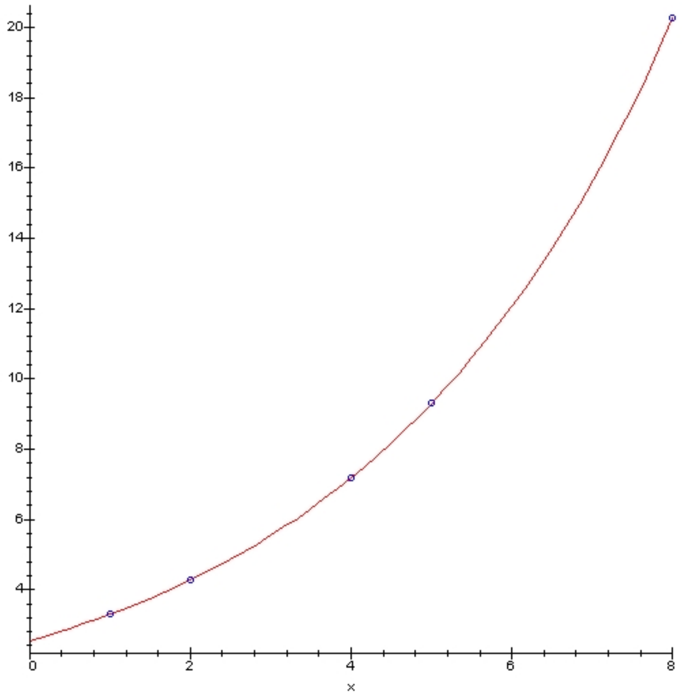


Figure 14.89: **Points and our interpolated Gauss-Newton function**

To close this subject let us do a comparison with the Newton-Raphson method for the first iteration using the same starting bipoint. Let us recall again that for this latter method, the iterations are based on the relation:

$$-\vec{\nabla}(f(\vec{x}))=H(\vec{x}_i)(\vec{x}-\vec{x}_i) \quad (14.1.1489)$$

and we will write the function as following for the Newton-Raphson method:

$$f(x_1,x_2)=f(\vec{x})=\frac{1}{2}\sum_{i=1}^{5}\left(x_1e^{x_2t_i}-y_i\right)^2 \quad (14.1.1490)$$

and:

$$\begin{aligned}H(\vec{x}_i)&=\begin{bmatrix}\partial^2_{x_1}f(x_i,y_i) & \partial_{x_1}\partial_{x_2}f(x_i,y_i)\\ \partial_{x_1}\partial_{x_2}f(x_i,y_i) & \partial_{x_2}\partial_{x_2}f(x_i,y_i)\end{bmatrix}\\
&=\begin{bmatrix}\partial^2_{x_1}\dfrac{1}{2}\displaystyle\sum_{i=1}^{5}\left(x_1e^{x_2t_i}-y_i\right)^2 & \partial_{x_1}\partial_{x_2}\dfrac{1}{2}\displaystyle\sum_{i=1}^{5}\left(x_1e^{x_2t_i}-y_i\right)^2\\ \partial_{x_1}\partial_{x_2}\dfrac{1}{2}\displaystyle\sum_{i=1}^{5}\left(x_1e^{x_2t_i}-y_i\right)^2 & \partial_{x_2}\partial_{x_2}\dfrac{1}{2}\displaystyle\sum_{i=1}^{5}\left(x_1e^{x_2t_i}-y_i\right)^2\end{bmatrix}\\
&=\begin{bmatrix}\partial_{x_1}\displaystyle\sum_{i=1}^{5}\left(x_1e^{x_2t_i}-y_i\right)e^{x_2t_i} & \partial_{x_1}\displaystyle\sum_{i=1}^{5}\left(x_1e^{x_2t_i}-y_i\right)x_1t_ie^{x_2t_i}\\ \partial_{x_1}\displaystyle\sum_{i=1}^{5}\left(x_1e^{x_2t_i}-y_i\right)x_1t_ie^{x_2t_1} & \partial_{x_2}\displaystyle\sum_{i=1}^{5}\left(x_1e^{x_2t_i}-y_i\right)x_1t_ie^{x_2t_i}\end{bmatrix}\\
&=\begin{bmatrix}\displaystyle\sum_{i=1}^{5}e^{2x_2t_i} & \displaystyle\sum_{i=1}^{5}2x_1t_ie^{x_2t_i}-y_it_ie^{x_2t_i}\\ \displaystyle\sum_{i=1}^{5}2x_1t_ie^{x_2t_i}-y_it_ie^{x_2t_i} & \displaystyle\sum_{i=1}^{5}2x_1^2t_ie^{2x_2t_i}-x_1y_it_i^2e^{x_2t_i}\end{bmatrix}\\
&=\begin{bmatrix}78.5367 & 1214.4991\\ 1215.4991 & 22278.6570\end{bmatrix}\end{aligned}\tag{14.1.1491}$$

Then we have:

$$-\vec{\nabla}(f(\vec{x}_i))=H(\vec{x}_i)(\vec{x}-\vec{x}_i)\tag{14.1.1492}$$

that becomes:

$$(16.5888300.8722)=\begin{bmatrix}78.5367 & 1215.4991\\ 1215.4991 & 22278.6570\end{bmatrix}\begin{pmatrix}x_1-x_{1,0*}\\ x_2-x_{2,0*}\end{pmatrix}\tag{14.1.1493}$$

Then we have:

$$-\vec{\nabla}(f(\vec{x}_i))=H(\vec{x}_i)(\vec{x}-\vec{x}_i)\tag{14.1.1494}$$

That becomes:

$$\begin{pmatrix}16.5888\\ 300.8722\end{pmatrix}\begin{bmatrix}78.5367 & 1215.4991\\ 1215.4991 & 22278.6570\end{bmatrix}\begin{pmatrix}x_1-x_{1,0*}\\ x_2-x_{2,0*}\end{pmatrix}\tag{14.1.1495}$$

Thus:

$$\begin{pmatrix}x_1-x_{1,0*}\\ x_2-x_{2,0*}\end{pmatrix}=\begin{pmatrix}x_1\\ x_2\end{pmatrix}-\begin{pmatrix}x_{1,0*}\\ x_{2,0*}\end{pmatrix}=\begin{pmatrix}0.0142\\ 0.0127\end{pmatrix}+\begin{pmatrix}2.5\\ 0.25\end{pmatrix}=\begin{pmatrix}2.5142\\ 0.2627\end{pmatrix}\tag{14.1.1496}$$

Therefore: and therefore the next bipoint for the iteration will be:

$$\begin{pmatrix}x_1\\ x_2\end{pmatrix}=\begin{pmatrix}x_{1,1}\\ x_{2,1}\end{pmatrix}=\begin{pmatrix}0.0381\\ 0.0102\end{pmatrix}+\begin{pmatrix}x_{1,0*}\\ x_{2,0*}\end{pmatrix}=\begin{pmatrix}0.0142\\ 0.0127\end{pmatrix}+\begin{pmatrix}2.5\\ 0.25\end{pmatrix}=\begin{pmatrix}2.5142\\ 0.2627\end{pmatrix}\tag{14.1.1497}$$

which corresponds obviously to the values of the first iteration:

$$\begin{pmatrix}x_{1,1*}\\ x_{2,2*}\end{pmatrix}=\begin{pmatrix}2.5381\\ 0.2602\end{pmatrix}\tag{14.1.1498}$$

We will not do again explicitly also other iterations. So this is what it gives finally about our function to be minimized:

i	$\min_{x_1,x_2} f(x_1,x_2)$
0	$2\cdot 10^{0}$
1	$1\cdot 10^{-1}$
2	$2\cdot 10^{-4}$
3	$5\cdot 10^{-9}$
4	$6\cdot 10^{-9}$
5	$3\cdot 10^{-9}$

Table 14.21: Gauss-Newton iterations of our example

So the Newton-Raphson method converges in this case slower than that of Gauss-Newton.

Here is the corresponding pseudocode:

```
Algorithm 14: Gauss-Newton method (tangent Newton)
Input: $y(\vec{x},t):\mathbb{R}^n \rightarrow \mathbb{R}$ a twice-differentiable a priori function to fit the data
       $\vec{y}\in\mathbb{R}^m$ the real data
       $\vec{x}_0\in\mathbb{R}^n$ an initial estimation
       $\varepsilon$ is a tolerance value
for $k$ to maxiter do
    for $i$ to $m$ do
        Assign $F_i = y(\vec{x}_k,t) - y_i$ ;                ▷ to get the vector $F(\vec{x})$
    ▷ to get the vector $\vec{\nabla}(f(\vec{x}_k)) = \vec{\nabla}\left(F(\vec{x}_k)^T\right) F(\vec{x}_k)$
    for $j$ to $n$ do
        Assign $\vec{\nabla}(f(\vec{x}_k))_j = \sum_{i=1}^{m} F_i \dfrac{\partial y(\vec{x}_k,t)}{\partial x_j}$
    ▷ to get the $n\times n$ matrix: $[\vec{\nabla}\left(F(\vec{x}_k)^T\right)\vec{\nabla}(F(\vec{x}_k))]$
    for $i$ to $n$ do
        for $j$ to $n$ do
            Assign $[\vec{\nabla}\left(F(\vec{x}_k)^T\right)\vec{\nabla}(F(\vec{x}_k))]_{ij} = \sum_{l=1}^{m} \dfrac{\partial y(\vec{x}_k,t)}{\partial x_i}\dfrac{\partial y(\vec{x}_k,t)}{\partial x_j}$
    Update $\vec{x}_{k+1} \leftarrow \vec{x}_k - \left[\vec{\nabla}\left(F\left(\vec{x}_k\right)^T\right)\vec{\nabla}\left(F\left(\vec{x}_k\right)\right)\right]^{-1}\left(\vec{\nabla}\left(F\left(\vec{x}_k\right)^T\right)F\left(\vec{x}_k\right)\right)$;
    if $\|f\left(\vec{x}_{k+1}\right) - f\left(\vec{x}_k\right)\| < \varepsilon$ or $\left\|\vec{\nabla} f\left(\vec{x}_k\right)\right\| < \varepsilon$ then
        Converged. Print $\vec{x}^* \leftarrow \vec{x}_{k+1}$ and $f\left(\vec{x}^*\right) \leftarrow f\left(\vec{x}_{k+1}\right)$;
```

Remark

For the square matrix inversion in the pseudocode above, one can use the determinant and cofactor method (see section Linear Algebra page 1980), QR decomposition (see section Linear Algebra page 2030), LU decomposition (see section Linear Algebra page 2024) or Gaussian Elimination with partial pivoting (GEPP).

So this was a very detailed step-by-step numerical example. For people interested in Machine Learning and for a more formal example (without numerical values) with a famous case (univariate linear regression) let us consider:

$$h_\theta(x)=\theta_0+\theta_1 x \tag{14.1.1499}$$

The cost function for any guess of θ_0,θ_1 can be computed as:

$$J(\vec{\theta})=J(\theta_0,\theta_1)=\frac{1}{2}\sum_{i=1}^{m}\left(h_\theta(x^{(i)})-y^{(i)}\right)^2 \tag{14.1.1500}$$

where $x^{(i)}$ and $y^{(i)}$ are the x and y values for the ith component in the learning set. If we substitute for $h_\theta(x)$:

$$J(\theta_0,\theta_1)=\frac{1}{2}\sum_{i=1}^{m}\left(\theta_0+\theta_1 x^{(i)}-y^{(i)}\right)^2 \tag{14.1.1501}$$

Then as we know, in Machine Learning, the goal of gradient descent can be expressed as:

$$\min_{\theta_0,\theta_1} J(\theta_0,\theta_1) \tag{14.1.1502}$$

Remark

At the opposite of the "gradient ascent" where we seek:

$$\max_{\theta_0,\theta_1} J(\theta_0,\theta_1) \tag{14.1.1503}$$

Finally, each step in the gradient descent can be described as we have seen earlier above for a component j:

$$\theta_j^{+}=\theta_j-\eta\frac{\partial}{\partial\theta_j}J(\theta_0,\theta_1) \tag{14.1.1504}$$

for $j=0$ and $j=1$ in this special case!

It is straightforward that:

$$\frac{\partial J(\theta_0,\theta_1)}{\partial\theta_0}=\frac{\partial}{\partial\theta_0}\frac{1}{2}\sum_{i=1}^{m}\left(\theta_0+\theta_1 x^{(i)}-y^{(i)}\right)^2=\sum_{i=1}^{m}\left(\theta_0+\theta_1 x^{(i)}-y^{(i)}\right) \tag{14.1.1505}$$

and:

$$\frac{\partial J(\theta_0,\theta_1)}{\partial\theta_1}=\frac{\partial}{\partial\theta_1}\frac{1}{2}\sum_{i=1}^{m}\left(\theta_0+\theta_1 x^{(i)}-y^{(i)}\right)^2=\sum_{i=1}^{m}\left(\theta_0+\theta_1 x^{(i)}-y^{(i)}\right)x^{(i)} \tag{14.1.1506}$$

Or in vector form:

$$\vec{\theta}^{+}=\vec{\theta}+\Delta\theta=\vec{\theta}-\eta\vec{\nabla}\left(J(\vec{\theta}_*)^T\right)J(\vec{\theta}_*)=\vec{\theta}-\eta\vec{\nabla}J(\vec{\theta})=\vec{\theta}-\eta\sum_{i=1}^{m}\left(\vec{\theta}^T\vec{x}^{(i)}-y^{(i)}\right)\vec{x}^{(i)} \tag{14.1.1507}$$

In order to minimize the above cost function $J(\theta_0,\theta_1)$, notice that the gradient is calculated from the whole training set (hence the sum!). This is why this approach of "gradient descent" is more specifically referred to as "batch gradient descent".

Remark

The use of the traditional complete method using the Hessian is in the early 121st century (holocene calendar) impractical for most deep learning applications because computing (and inverting) the Hessian in its explicit form is a very costly process in both space and time. For instance, a Neural Network with one million parameters would have a Hessian matrix of size $[1,000,000\times 1,000,000]$, occupying approximately $3,725$ gigabytes of RAM. Hence, a large variety of quasi-Newton methods have been developed that seek to approximate the inverse Hessian!

As we need to calculate the gradients for the whole dataset to perform just one update, batch gradient-descent can be very slow and is intractable for datasets that do not fit in memory. Batch gradient descent also does not allow us to update our model online, i.e. with new examples on-the-fly

So, in "stochastic gradient descent method" (SGD), also known as "incremental gradient descent", instead of updating the weights $\vec{\theta}$ based on the sum of the m accumulated errors over all samples $x^{(i)}$:

$$\theta_j^+=\theta_j-\eta\frac{\partial}{\partial\theta_j}J(\theta_0,\theta_1)=\theta_j-\eta\frac{\partial}{\partial\theta_j}\frac{1}{2}\sum_{i=1}^{m}\left(\theta_0+\theta_1x^{(i)}-y^{(i)}\right)^2 \tag{14.1.1508}$$

we can use the following update:

$$\theta_j^+=\theta_j-\eta\frac{\partial}{\partial\theta_j}J(\theta_0,\theta_1)=\theta_j-\eta\frac{\partial}{\partial\theta_j}\frac{1}{2}\left(\theta_0+\theta_1x^{(i)}-y^{(i)}\right)^2 \tag{14.1.1509}$$

Note that we now update the weights incrementally with a single training sample but not with the whole training set.

It is said to be "stochastic" because samples are selected randomly (or shuffled) instead of as a single group (as in standard gradient descent) or in the order they appear in the training set.

Example:

Let's suppose we want to fit a straight line $y_i=\theta_0+\theta_1x_i$ to a training set with observations $(x_1,x_2,\ldots,x_n)$ and corresponding estimated responses $(\hat{y}_1,\hat{y}_2,\ldots,\hat{y}_n)$ using least squares. The objective function to be minimized is:

$$J(\vec{\theta})=\sum_{i=1}^{m}J_i(\vec{\theta})=\sum_{i=1}^{m}\left(\hat{y}^{(i)}-y^{(i)}\right)^2=\sum_{i=1}^{m}\left(\theta_0+\theta_1x^{(i)}-y^{(i)}\right)^2 \tag{14.1.1510}$$

Then we will have using the stochastic gradient descent method:

$$\begin{bmatrix}\theta_0\\ \theta_1\end{bmatrix}^+:=\begin{bmatrix}\theta_0\\ \theta_1\end{bmatrix}-\eta\begin{bmatrix}\dfrac{\partial}{\partial\theta_0}(\theta_0+\theta_1x^{(i)}-^{(i)})^2\\ \dfrac{\partial}{\partial\theta_1}(\theta_0+\theta_1x^{(i)}-y^{(i)})^2\end{bmatrix}=\begin{bmatrix}\theta_0\\ \theta_1\end{bmatrix}-\eta\begin{bmatrix}2(\theta_0+\theta_1x^{(i)}-y^{(i)})\\ 2x_i(\theta_0+\theta_1x^{(i)}-y^{(i)})\end{bmatrix} \tag{14.1.1511}$$

Note again that in each iteration, the gradient is evaluated at a single point $x^{(i)}$ instead of being evaluated on the set of all samples.

The key difference compared to standard (Batch) Gradient Descent is that only one piece of data from the dataset is used to calculate the step, and the piece of data is picked randomly at each step. SGD does away with this redundancy by performing one update at a time. It is therefore usually much faster and can also be used to learn online. SGD performs frequent updates with a high variance that cause the objective function to fluctuate heavily.

While batch gradient descent converges to the minimum of the basin the parameters are placed in, SGD's fluctuation, on the one hand, enables it to jump to new and potentially better local minima. On the other hand, this ultimately complicates convergence to the exact minimum, as SGD will keep overshooting. However, it has been shown that when we slowly decrease the learning rate, SGD shows the same convergence behaviour as batch gradient descent, almost certainly converging to a local or the global minimum for non-convex and convex optimization respectively.

Remark

"mini-batch gradient descent" initially takes the best of both worlds and performs an update for every mini-batch of n training examples. Common mini-batch sizes range between 50 and 256, but can vary for different applications. Mini-batch gradient descent is typically the algorithm of choice when training a neural network and the term SGD usually is employed also when mini-batches are used

14.1.17.2.4.1 Gradient Descent/Ascent techniques summary

At this point it may be useful to also introduce the gradient descent and gradient ascent in a simpler way (if it can help!) as it is quite and important topic.

The reader may have already understand that the gradient descent/ascent is a general framework for solving optimization problems where we want to maximize or minimize functions of continuous (differentiable) parameters.

We will illustrate the technique here for maximization (ascent) in terms of a single free variable, but the generalization is quite simple and there is already plenty of example above!

Suppose we want to find a maximum of some function $y=f(x)$. The standard procedure is to find $f'(x)$, set it to zero and solve for x. But what if $f'(x)=0$ is a difficult equation to solve?

We can still find the maximum algorithmically!

Let $y=f(x)$ have a maximum at x_{max}. Pick an arbitrary value for x, say x_1. Compute $f'(x_1)$. If the slope of y is positive at x_1, i.e. $f'(x_1)>0$, then $x_{\text{max}}>x_1$ lies to the right of x_1. Likewise if $f'(x_1)<0$, then $x_{\text{max}}<x_1$ lies to its left!

Thus we know the direction in which x_1 should be updated in order to approach x_{max}. In fact this direction is given by $f'(x_1)$. So we can use the update rule:

$$x_1 \leftarrow x_1+\eta f'(x_1) \qquad (14.1.1512)$$

where η is a positive constant. If η is sufficiently small, and there is indeed a maximum for f, the above update rule will converge to it after a finite number of iterations (when $f'(x_1)=0$).

This is the so-named "gradient ascent technique" to determine maxima when actual solution of $f' = 0$ is difficult. To find minima, the corresponding gradient descent rule can set up to update x_1 away from $x_{\max}$.

For $y = f(x_1, \cdots, x_n)$ the partial derivatives with respect to each x_i indicate how rapidly y is changing along that axis and so the gradient, ∇y, in fact denotes the precise direction which promises maximum increase of y.

Collectively these, among others, are named "gradient techniques".

We will summarise now the common gradient descent optimisation algorithms used in popular deep learning frameworks in the early 121st century (holocene calendar).

Recall that the vanilla stochastic gradient descent (SGD) updates weights by subtracting the current weight by a factor (i.e. η, the learning rate) of its gradient:

$$w_{\text{new}} = w - \underbrace{\eta}_{\substack{\text{learning}\\\text{rate}\\\text{component}}} \underbrace{\frac{\partial L}{\partial w}}_{\substack{\text{gradient}\\\text{descent}}} \tag{14.1.1513}$$

Variations in this equation are commonly known as "stochastic gradient descent optimisers". There are 3 main ways how they differ:

1. Adapt the "gradient component" ($\frac{\partial L}{\partial w}$)

   Instead of using only one single gradient like in stochastic vanilla gradient descent to update the weight, take an aggregate of multiple gradients. Specifically, these optimisers use the exponential moving average of gradients.

2. Adapt the "learning rate component" (η)

   Instead of keeping a constant learning rate, adapt the learning rate according to the magnitude of the gradient(s).

3. Both (1) and (2)

   Adapt both the gradient component and the learning rate component.

As you will see later, these optimisers try to improve the amount of information used to update the weights, mainly through using previous (and future) gradients, instead of only the present available gradient.

Below is a table that summarises which components are being adapted (years are based on the holocene calendar):

Optimiser	Year	Learning Rate	Gradient
Momentum	11964		✓
AdaGrad	12011	✓	
RMSprop	12012	✓	
AdaDelta	12012	✓	
Nesterov	12013		✓
Adam	12014	✓	✓
AdaMax	12015	✓	✓
Nadam	12015	✓	✓
AMSGrad	12018	✓	✓

Table 14.22: **Gradient descent optimisers**

Here are the detailed expression of each of the empirical method visible in the above table (the knowledgeable reader may recognize some techniques inspired by time series forecasting techniques):

- Vanilla SGD:

$$w_{t+1}=w_t-\eta\frac{\partial L}{\partial w_t} \tag{14.1.1514}$$

- Momentum:

$$\begin{aligned}w_{t+1}&=w_t-\eta m_t\\ m_t&=\beta m_{t-1}+(1-\beta)\frac{\partial L}{\partial w_t}\end{aligned} \tag{14.1.1515}$$

- Adagrad:

$$\begin{aligned}w_{t+1}&=w_t-\frac{\eta}{\sqrt{v_t+\varepsilon}}\cdot\frac{\partial L}{\partial w_t}\\ v_t&=v_{t-1}+\left[\frac{\partial L}{\partial w_t}\right]^2\end{aligned} \tag{14.1.1516}$$

- RMSprop:

$$\begin{aligned}w_{t+1}&=w_t-\frac{\eta}{\sqrt{v_t+\varepsilon}}\cdot\frac{\partial L}{\partial w_t}\\ v_t&=\beta v_{t-1}+(1-\beta)\left[\frac{\partial L}{\partial w_t}\right]^2\end{aligned} \tag{14.1.1517}$$

- Adadelta:

$$\begin{aligned}w_{t+1}&=w_t-\frac{\sqrt{D_{t-1}+\varepsilon}}{\sqrt{v_t+\varepsilon}}\cdot\frac{\partial L}{\partial w_t}\\ D_t&=\beta D_{t-1}+(1-\beta)[\Delta w_t]^2\\ v_t&=\beta v_{t-1}+(1-\beta)\left[\frac{\partial L}{\partial w_t}\right]^2\end{aligned} \tag{14.1.1518}$$

- Nesterov:

$$\begin{aligned} w_{t+1} &= w_t - \eta m_t \\ m_t &= \beta m_{t-1} + (1-\beta)\frac{\partial L}{\partial w^*} \\ w^* &= w_t - \eta m_{t-1} \end{aligned} \tag{14.1.1519}$$

- Adam:

$$\begin{aligned} w_{t+1} &= w_t - \frac{\eta}{\sqrt{\hat{v}_t}+\varepsilon}\cdot \hat{m}_t \\ \hat{m}_t &= \frac{m_t}{1-\beta_1^t} \\ \hat{v}_t &= \frac{v_t}{1-\beta_2^t} \\ m_t &= \beta_1 m_{t-1} + (1-\beta_1)\frac{\partial L}{\partial w_t} \\ v_t &= \beta_2 v_{t-1} + (1-\beta_2)\left[\frac{\partial L}{\partial w_t}\right]^2 \end{aligned} \tag{14.1.1520}$$

- AdaMax:

$$\begin{aligned} w_{t+1} &= w_t - \frac{\eta}{v_t}\cdot \hat{m}_t \\ \hat{m}_t &= \frac{m_t}{1-\beta_1^t} \\ m_t &= \beta_1 m_{t-1} + (1-\beta_1)\frac{\partial L}{\partial w_t} \\ v_t &= \max\left(\beta_2 v_{t-1}, \left|\frac{\partial L}{\partial w_t}\right|\right) \end{aligned} \tag{14.1.1521}$$

- Nadam:

$$\begin{aligned} w_{t+1} &= w_t - \frac{\eta}{\sqrt{\hat{v}_t}+\varepsilon}\left(\beta_1 \hat{m}_t + \frac{1-\beta_1}{1-\beta_1^t}\cdot\frac{\partial L}{\partial w_t}\right) \\ \hat{m}_t &= \frac{m_t}{1-\beta_1^t} \\ \hat{v}_t &= \frac{v_t}{1-\beta_2^t} \\ m_t &= \beta_1 m_{t-1} + (1-\beta_1)\frac{\partial L}{\partial w_t} \\ v_t &= \beta_2 v_{t-1} + (1-\beta_2)\left[\frac{\partial L}{\partial w_t}\right]^2 \end{aligned} \tag{14.1.1522}$$

- AMSGrad:

$$\begin{aligned} w_{t+1} &= w_t - \frac{\eta}{\sqrt{\hat{v}_t}+\varepsilon}\cdot m_t \\ m_t &= \beta_1 m_{t-1} + (1-\beta_1)\frac{\partial L}{\partial w_t} \\ v_t &= \beta_2 v_{t-1} + (1-\beta_2)\left[\frac{\partial L}{\partial w_t}\right]^2 \\ \hat{v}_t &= \max(\hat{v}_{t-1}, v_t) \end{aligned} \tag{14.1.1523}$$

Below the reader can see animations (credit: Alec Radford) that may help to have a better intuition about the learning process dynamics (we also strongly recommend to read the paper *An overview of gradient descent optimization algorithms* to get more details on the underlying ideas [470]).

The first Flash animation below depicts the contours of a loss surface and time evolution of different optimization algorithms. Notice the "overshooting" behaviour of momentum-based methods, which make the optimization look like a ball rolling down the hill:

The animation above will run for people having a PDF reader with Adobe Flash player installed and activated (otherwise see here: `https://vimeo.com/575749745`).

The second and last Flash animation is a visualization of a saddle point in the optimization landscape, where the curvature along different dimension has different signs (one dimension curves up and another down). Notice that SGD has a very hard time breaking symmetry and gets stuck on the top. Conversely, algorithms such as RMSprop will see very low gradients in the saddle direction. Due to the denominator term in the RMSprop update, this will increase the effective learning rate along this direction, helping RMSProp proceed:

The animation above will run for people having a PDF reader with Adobe Flash player installed and activated (otherwise see here: https://vimeo.com/575752619).

14.1.17.3 Expectation-Maximization (EM) algorithm

The "Expectation-Maximization algorithm" is a stochastic optimizations algorithm often used in machine learning and very quite different of the methods that we have seen before. We will use it explicitly in this textbook later below for various applications like Factor Analysis (see page 6925) but also for the T-distributed Stochastic Neighbour Embedding (see page 6939) and others during our study of Data Mining.

The EM algorithm seems to have been introduced in the early 11950 (holocene calendar) by R. Ceppellini, M. Siniscalco and C.A. Smith in the context of gene frequency estimation. The expectation maximization algorithm was analysed more generally by H. Hartley and by L.E. Baum in the context of hidden Markov models, where it is commonly known as the "Baum-Welch algorithm".

Remark

The text that follows is inspired of the following paper [76] but also of the following web page `https://indowhiz.com/articles/en/the-simple-concept-of-expectation-maximization-em-algorithm` itself inspired of [151].

To summarize, EM algorithm is actually an iterative method, involving expectation (E-step) and maximization (M-step) to find the local maximum likelihood from the data. Commonly, EM algorithm is used on several distributions or statistical models, where there are one or more unknown variables named the "latent variables".

To easily understand EM algorithm, we will use first a companion example of the coin tosses distribution that is widely used in quite a few textbooks, articles and blogs and afterwards we will deal with the underlying mathematical formalism.

So let us consider that we have 2 coins: Coin A and Coin B, where both have a different head-up probability. We will randomly choose a coin 5 times, whether coin A or B. Then, each coin selection is followed by tossing it 10 times. Therefore, we have the following outcomes:

- Set 1 (Coin B): `H T T T H H T H T H` (5H5T)
- Set 2 (Coin A): `H H H H T H H H H H` (9H1T)
- Set 3 (Coin A): `H T H H H H H T H H` (8H2T)
- Set 4 (Coin B): `H T H T T T H H T T` (4H6T)
- Set 5 (Coin A): `T H H H T H H H T H` (7H3T)

where the image side of the coin will be denoted by `H` or Head-up, and the number side will be denoted by `T` or Tail-up. Then, the probability of a coin to land with head-up for each of these coins, will be denoted as θ.

The question is, how do we estimate θ for each coin?

Because we already have all the data we need, we can easily calculate the probability of getting head-up for each coin. Therefore, the calculation of θ will be done through the maximum likelihood estimation (MLE) of the proportion that can naively be obtained by:

- By knowing the result of coin $A=24\texttt{H}6\texttt{T}$, in 3 sets with a total of 30 tosses, then:

$$\hat{\theta}_A=\frac{24}{30}=0.80 \tag{14.1.1524}$$

- Similarly to the result of coin $B=9\texttt{H}11\texttt{T}$, in 2 sets with a total of 20 tosses, then:

$$\hat{\theta}_B=\frac{9}{20}=0.45 \tag{14.1.1525}$$

Before returning to the original case, let us explain briefly the different the stages of the EM algorithm as it will give us an overall picture of the EM algorithm flow and may help to understand what will follow:

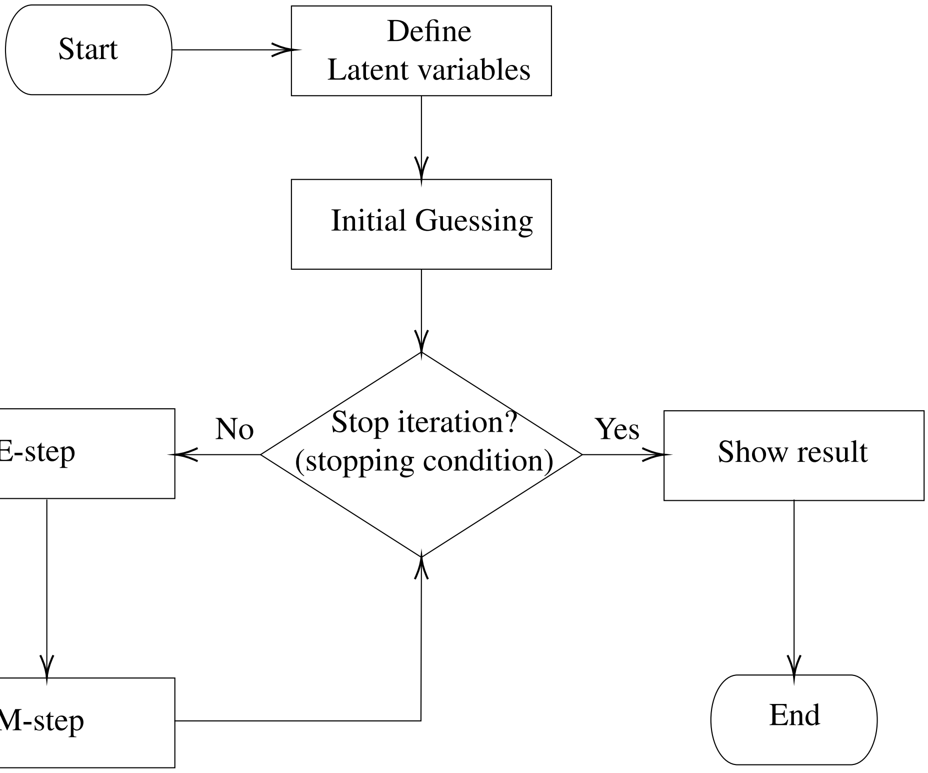


Figure 14.90: **ISO 5807:85 Flowchart of EM algorithm**

Actually, the main point of EM is the iteration between E-step and M-step, which could be seen in the figure above. The E-step will estimate your hidden variables, and the M-step will re-update the parameters, based on the estimation of the hidden variables. In other words, this iteration aims to re-improve the estimation of current parameters.

First, what we must define is what variables are required; but are not observed directly in the data.

The goal is to estimate the probability of getting heads-up for each coin. However, it cannot be calculated directly if we don't know the identity of the coin used in each set. Therefore, it is necessary to know which coin is used in each set. In other words, this coin identity will be our latent variable.

Right now, besides the 5 sets of outcomes above, we only know that we used two coins; which are coin A and coin B.

As stated before, the probability of getting head-up for each of these coins is denoted as θ. Currently, there are only coin A and coin B; with unknown parameter values of θ_A and θ_B.

Remark

By the way, please note that there is no relationship between parameters guessing; both θ_A and θ_B. For example, if you think that the sum of θ_A and θ_B must be $=1$, NO! This probability represents the individual value of getting heads-up on each coin.

Okay enough! Now, let's say we have randomly set both initial values, which are:

$$\hat{\theta}_A^{(0)} = 0.6 \quad \text{and} \quad \hat{\theta}_B^{(0)} = 0.8 \tag{14.1.1526}$$

Now we have the required variables. So, we can start estimating the identity of coin used in each set.

Each set, which contains the outcomes of heads and tails, can be denoted by E notation. Then, the probability of *using coin* A denoted as Z_A, and the probability of *using coin* B denoted as Z_B.

At the beginning of E-Step, we need to know the probability of a set using coins A or B. Take an example from the set 3 earlier above, where $E = \texttt{HTHHHHHTHH}$ (8H2T). If $\hat{\theta}_A = 0.6$, it means the probability of getting head is 0.6 (and tail 0.4) uses coin A. Then, how much probability of coin A will give $\texttt{8H2T}$ in 10 tosses (a set)? This is what we need to estimate first. Then, this probability can be denoted by $P(E|Z_A)$. Similarly, the probability of coin B giving $E = \texttt{8H2T}$, could be denoted by $P(E|Z_B)$.

This actually relates to the probability distribution of a binomial random variable, which definition is for recall:

$$P(E|Z_x) = \frac{n!}{h!\,(n-h)!}\,{\theta_x}^h\,(1-\theta_x)^{n-h} \tag{14.1.1527}$$

Or denoted:

$$P(E|Z_x,\theta) = \frac{n!}{h!\,(n-h)!}\,{\theta_{Z_x}}^h\,(1-\theta_{Z_x})^{n-h} \tag{14.1.1528}$$

Then, we can calculate the probability of getting E using coins A and B as follows:

$$\begin{aligned} P(E_{\text{8H2T}}|Z_A) &= \frac{10!}{8!\,2!}\,0.6^8\,0.4^2 = 0.121 \\ P(E_{\text{8H2T}}|Z_B) &= \frac{10!}{8!\,2!}\,0.5^8\,0.5^2 = 0.044 \end{aligned} \tag{14.1.1529}$$

Next, we can compare the probabilities of both coin A and coin B giving E. So, according to Bayes' theorem and the law of total probability, we can determine the ratio of probability (posterior value of latent variables) using the following relation:

$$P(Z_x|E)=\frac{P(E|Z_x)P(Z_x)}{P(E|Z_x)P(Z_x)+P(E|Z_y)P(Z_y)} \tag{14.1.1530}$$

where $P(Z_x|E)$ is the probability of coin x giving E (compared to coin y), $P(Z_x)$ is the probability of choosing coin x and $P(Z_x)$ is the probability of choosing coin y.

However, because our case is only using 2 coins, coins A and B, then the probability that we will choose one of them is $50/50$. Then $P(Z_A)=P(Z_B)=0.5$.

Accordingly, for coin A, we can simplify the relation above into:

$$P(Z_A|E)=\frac{P(E|Z_A)\cancel{P(Z_A)}}{P(E|Z_A)\cancel{P(Z_A)}+P(E|Z_B)\cancel{P(Z_B)}}=\frac{P(E|Z_A)}{P(E|Z_A)+P(E|Z_B)} \tag{14.1.1531}$$

Hence:

$$P(Z_A|E_{\text{8H2T}})=\frac{\cancel{\frac{10!}{8!\,2!}}\,0.6^8\,0.4^2}{\cancel{\frac{10!}{8!\,2!}}\,0.6^8\,0.4^2+\cancel{\frac{10!}{8!\,2!}}\,0.5^8\,0.5^2}=\frac{0.6^8\,0.4^2}{0.6^8\,0.4^2+0.5^8\,0.5^2}=0.73 \tag{14.1.1532}$$

Similarly, for coin B, we have:

$$P(Z_B|E_{\text{8H2T}})=\frac{\cancel{\frac{10!}{8!\,2!}}\,0.5^8\,0.5^2}{\cancel{\frac{10!}{8!\,2!}}\,0.6^8\,0.4^2+\cancel{\frac{10!}{8!\,2!}}\,0.5^8\,0.5^2}=\frac{0.5^8\,0.5^2}{0.6^8\,0.4^2+0.5^8\,0.5^2}=0.27 \tag{14.1.1533}$$

Previously, the ratio of coins A and B in giving each E needs to be calculated. In this case, it is calculated from set 1 to 5. Then the results could be seen in the following table:

Coin Tosses	E	**Coin A probability**	**Coin B probability**
HTTTHHTHTH	5H5T	0.45	0.55
HHHHTHHHHH	9H1T	0.80	0.20
HTHHHHHTHH	8H3T	0.73	0.27
HTHTTTHHTT	4H6T	0.35	0.65
THHHTHHHTH	7H3T	0.65	0.35

After that, we need to estimate the total number of `H` for each coin. It is calculated based on the coin ratio above. To calculate *total heads and tails* for coin x, it is similar to the complete data. The method is quite easy, we just need to multiply the ratio of each coin to the number of heads in each E, the results are shown in the following table:

Coin Tosses	E	**Estimated `H` for Coin A**	**Estimated `H` for Coin B**
HTTTHHTHTH	5H5T	$5\cdot0.45=2.25$	$5\cdot0.55=2.75$
HHHHTHHHHH	9H1T	$9\cdot0.8=7.2$	$9\cdot0.20=1.8$
HTHHHHHTHH	8H3T	$8\cdot0.73=5.84$	$8\cdot0.27=2.16$
HTHTTTHHTT	4H6T	$4\cdot0.35=1.4$	$4\cdot0.65=2.6$
THHHTHHHTH	7H3T	$7\cdot0.65=4.55$	$7\cdot0.35=2.45$

If we want, we can also calculate the tails for each coin as in the table above. But that is not necessary, because we can use another straightforward approach.

Until this step, we already have the E-Step calculations. Just a little bit more effort to finish our calculation in the M-Step.

The results from the E-step can be used to improve the $\hat{\theta}_x$ parameters. Here, we can use the maximum likelihood estimation (MLE) relation similar to the completed data.

As we said before, it is not necessary to calculate the tails for each coin; then we sum the heads and tails for each coin. Because each set contains 10 tosses, we just need to multiply the coin ratio with 10. That way, we could get the total estimated tosses from. Therefore (we will prove in details just further below where these both relations comes from):

$$\begin{aligned}\hat{\theta}_A^{(1)} &= \frac{2.25+7.2+5.84+1.4+4.55}{10(0.45+0.8+0.73+0.35+0.65)}=0.713\\ \hat{\theta}_B^{(1)} &= \frac{2.75+1.8+2.16+2.6+2.45}{10(0.55+0.2+0.27+0.65+0.35)}=0.581\end{aligned} \tag{14.1.1534}$$

Finally, the parameter of $\hat{\theta}_A$ and $\hat{\theta}_B$ for the first iteration have been improved. For the next iteration, the E-Step will use this new parameter value; and re-improved at next M-step. This iteration will always repeating the E-step and M-step, until it reaches any stop condition as illustrated below:

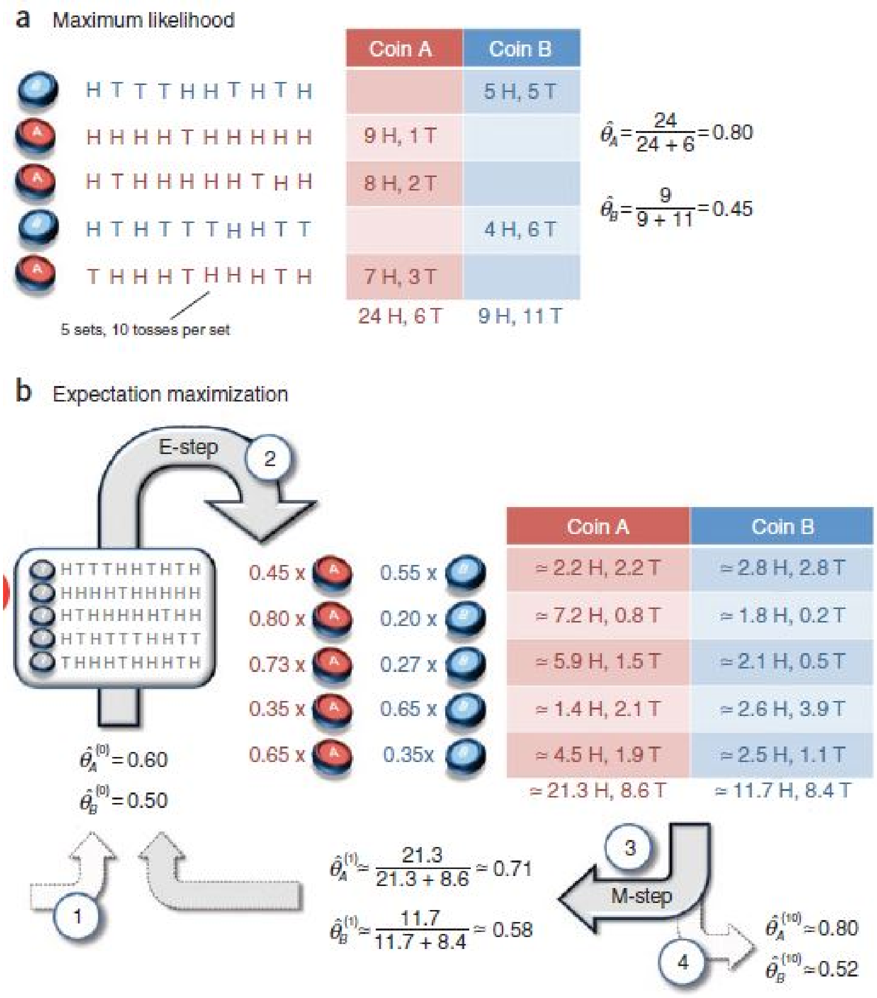

a Maximum likelihood
Coin A
Coin B
H T T T H H T H T H
5 H, 5 T
H H H H T H H H H H
9 H, 1 T
H T H H H H H T H H
8 H, 2 T
H T H T T T H H T T
4 H, 6 T
T H H H T H H H T H
7 H, 3 T
5 sets, 10 tosses per set
24 H, 6 T
9 H, 11 T
$\hat{\theta}_A = \frac{24}{24+6} = 0.80$
$\hat{\theta}_B = \frac{9}{9+11} = 0.45$
b Expectation maximization
E-step
2
0.45 x
0.55 x
0.80 x
0.20 x
0.73 x
0.27 x
0.35 x
0.65 x
0.65 x
0.35x
≈ 2.2 H, 2.2 T
≈ 2.8 H, 2.8 T
≈ 7.2 H, 0.8 T
≈ 1.8 H, 0.2 T
≈ 5.9 H, 1.5 T
≈ 2.1 H, 0.5 T
≈ 1.4 H, 2.1 T
≈ 2.6 H, 3.9 T
≈ 4.5 H, 1.9 T
≈ 2.5 H, 1.1 T
≈ 21.3 H, 8.6 T
≈ 11.7 H, 8.4 T
$\hat{\theta}_A^{(0)} = 0.60$
$\hat{\theta}_B^{(0)} = 0.50$
1
$\hat{\theta}_A^{(1)} \approx \frac{21.3}{21.3+8.6} \approx 0.71$
$\hat{\theta}_B^{(1)} \approx \frac{11.7}{11.7+8.4} \approx 0.58$
3
M-step
4
$\hat{\theta}_A^{(10)} \approx 0.80$
$\hat{\theta}_B^{(10)} \approx 0.52$

Figure 14.91: **EM parameter estimation for complete and incomplete data (source: Nature Biotechnology Volume 26 Number 8 August 12008)**

The iteration of the E-Step and M-Step, will be repeated until they meet the stopping condition. Commonly, the EM algorithm has two options of stopping condition, which are:

- Maximum iteration: means that, the EM Algorithm will stop if a certain number of iterations has been reached. For example the maximum iteration is set to 10 iterations, then the EM Algorithm will not be more than 10 iterations
- Convergence threshold: means that, the M-step gives no significant parameter improvement; compared to the improvement in the previous iteration. The changes are very small below our threshold

In the EM case example above, the parameter improvement in each iteration can be seen in the following table:

Iteration i	$\theta_A^{(i)}$	$\theta_B^{(i)}$	**Differences**
0	0.6	0.5	0.7180
1	0.713	0.581	0.1390
2	0.745	0.569	0.0342
3	0.768	0.550	0.0298
4	0.783	0.535	0.0212
5	0.791	0.526	0.0120
6	0.795	0.522	0.0057
7	0.796	0.521	0.0014
8	0.796	0.520	0.0010
9	0.796	0.520	0.0000

To calculate the differences or improvements in each iteration, we can use the Euclidean Distance such that:

$$d\left(\hat{\theta}^{(i)},\hat{\theta}^{(i-1)}\right)=\sqrt{\left(\hat{\theta}_A^{(i)}-\hat{\theta}_A^{(i-1)}\right)^2+\left(\hat{\theta}_A^{(i)}-\hat{\theta}_B^{(i)}\right)^2} \qquad (14.1.1535)$$

So for example:

$$d\left(\hat{\theta}^{(2)},\hat{\theta}^{(1)}\right)=\sqrt{(0.713-0.745)^2+(0.581-0.569)^2}=0.0342 \qquad (14.1.1536)$$

As many algorithm that one also has weaknesses, without too much doubts. Indeed, every iteration in the EM algorithm, in general, will always improve the parameter closer to the local maximum likelihood. In other words, the EM algorithm will guarantee convergence, but will not guarantee to give a global maximum likelihood. And also there is no guarantee that you will get the maximum likelihood estimation (MLE).

Now let us formalize what we have seen. For this purpose we will use the personal notes of Benjamin Bray [76] who authorized us to reproduce them below (we did some minor modifications).

Suppose we still have two coins, each with a different probability of heads, θ_A and θ_B, unknown to us. We collect data from a series of N trials in order to estimate the bias of each coin. Each trial k consists of flipping the same random coin Z_k a total of M times and recording only the total number X_k of heads.

Assuming all samples in the observed dataset $X=[\vec{x}_1,\ldots,\vec{x}_N]$ are independent and identically distributed (i.i.d.), we can find the "complete data likelihood" function of θ (using the Bayes' relation established in the section of Probabilities page 498):

$$L(\theta|\vec{X},\vec{z})=P(\vec{X},\vec{z}|\theta)=\prod_{n=1}^{N}P\left(\vec{x}_n,\vec{z}|\theta\right)=\prod_{n=1}^{N}P\left(\vec{x}_n|\vec{z},\theta\right)P(\vec{z}|\theta) \qquad (14.1.1537)$$

and the log-likelihood:

$$\ln\left(L(\theta|X,\vec{z})\right)=\ln\left(P(\vec{X},\vec{z}|\theta)\right)=\sum_{n=1}^{N}\left(\ln\left(P(\vec{x}_n|\vec{z},\theta)\right)\right)+\ln\left(\vec{z}|\theta\right) \qquad (14.1.1538)$$

Or in scalar form (for non-multivariate distributions):

$$\ln\left(L(\theta|X,z)\right)=\ln\left(P(X,z|\theta)\right)=\sum_{n=1}^{N}\left(\ln\left(P(x_n|z_n,\theta)\right)\right)+\ln\left(z_n|\theta\right) \tag{14.1.1539}$$

The second term can be dropped in our companion example as $P(z_n|\theta)$ is independent of the model parameters in θ and therefore irrelevant to the maximization of the log likelihood with respect to θ. Indeed, the reader must remember that in our companion example, z_n is the realization of the random (hidden) variable that describes the probability to deal with coin A either coin B. We assume that this probability is constant and equal to $1/2$, hence:

$$P(z_n|\theta)=P(z_n|\theta)=1/2=c^{te} \qquad \forall n \tag{14.1.1540}$$

The remaining term in our companion example is:

$$\begin{aligned}\ln\left(P(x_n|z_n,\theta)\right)&=\ln\binom{M}{x_n}\theta_{z_n}^{x_n}(1-\theta_{z_n})^{M-x_n}\\&=\ln\binom{M}{x_n}+x_n\ln(\theta_{z_n})+(M-x_n)\ln(1-\theta_{z_n})\end{aligned} \tag{14.1.1541}$$

Now that we have specified the probabilistic model and worked out all relevant probabilities, we are ready to derive an Expectation-Maximization algorithm.

Our general approach will be to reason about the hidden variables through a proxy distribution q, which we use to compute a lower-bound on the log-likelihood. This section is devoted to deriving one such bound, named the "Evidence Lower Bound" (ELBO). We can expand the data log-likelihood by marginalizing over the hidden variables:

$$L(\theta|\chi)=\ln\left(P(\chi|\theta)\right)=\ln\left(\sum_z P(\chi,z|\theta)\right) \tag{14.1.1542}$$

Through Jensen's inequality (see section Statistics page 576), we obtain the following bound:

$$\begin{aligned}L(\theta|\chi)&=\ln\left(\sum_z P(\chi,z|\theta)\right)\\&=\ln\left(\sum_z q(z)\frac{P(\chi,z|\theta)}{q(z)}\right)\\&\geq\sum_z q(z)\ln\left(\frac{P(\chi,z|\theta)}{q(z)}\right)\end{aligned} \tag{14.1.1543}$$

The lower bound can be rewritten as follows:

$$\begin{aligned}L(\theta|\chi)&\geq\sum_z q(z)\ln\left(\frac{P(\chi,z|\theta)}{q(z)}\right)\\&=\sum_z q(z)\ln\left(P(\chi,z|\theta)\right)-\sum_z q(z)\ln\left(q(z)\right)\\&=\mathrm{E}[\ln\left(P(\chi,z|\theta)\right)]-\mathrm{E}[\ln\left(q(z)\right)]\end{aligned} \tag{14.1.1544}$$

Where the last term is often a constant at each iteration that we can ignore for the maximization problem.

Now that we have this result, the **E-Step** is straightforward (calculation of the lower bound on the observed log-likelihood). The **M-Step** computes a new parameter estimate θ_{t+1} by optimizing over the lower bound found in the E-Step, as:

$$\begin{aligned}\theta_{t+1}=\arg\max_{\theta} L(\theta,\chi) &= \arg\max_{\theta} \mathrm{E}\left[\ln\left(P(\chi,z|\theta)\right)\right]\\ &= \arg\max_{\theta} \mathrm{E}\left[\ln\left(P(\chi|z,\theta)P(z)\right)\right]\\ &= \arg\max_{\theta} \mathrm{E}\left[\ln\left(P(\chi|z,\theta)\right)\right]+\ln\left(P(z)\right)\\ &= \arg\max_{\theta} \mathrm{E}\left[\ln P(\chi|z,\theta)\right]\end{aligned} \tag{14.1.1545}$$

Now, because each trial is conditionally independent of the others, given the parameters:

$$\begin{aligned}\mathrm{E}[\ln\left(P(\chi|z,\theta)\right)] = \mathrm{E}\left[\ln\left(\prod_{n=1}^{N}P(x_n|z_n,\theta)\right)\right] &= \sum_{n=1}^{N}\mathrm{E}[\ln\left(P(x_n|z_n,\theta)\right)]\\ &= \sum_{n=1}^{N}\mathrm{E}\Big[x_n\ln(\theta_{z_n})+(M-x_n)\ln(1-\theta_{z_n})\Big]+\sum_{n=1}^{N}\ln\binom{M}{x_n}\\ &= \sum_{n=1}^{N}\mathrm{E}\Big[x_n\ln(\theta_{z_n})+(M-x_n)\ln(1-\theta_{z_n})\Big]+\text{const. w.r.t. }\theta\\ &= \sum_{n=1}^{N}q(z_n=A)\Big[x_n\ln(\theta_A)+(M-x_n)\ln(\theta_A)\Big]\\ &+\sum_{n=1}^{N}q(z_n=B)\Big[x_n\ln(\theta_B)+(M-x_n)\ln(\theta_B)\Big]+\text{const. w.r.t. }\theta\end{aligned} \tag{14.1.1546}$$

Let $a_k=q(z_k=A)$ and $b_k=q(z_k=B)$. Note that:

$$\sum_{k=1}^{N}a_k=\sum_{k=1}^{N}b_k=1 \tag{14.1.1547}$$

To maximize the above expression with respect to the parameters, we take derivatives with respect to θ_A and θ_B and set to zero:

$$\begin{aligned}\frac{\partial}{\partial\theta_A}\Big[\mathrm{E}[\ln\left(P(\chi|z,\theta)\right)]\Big] &= \frac{1}{\theta_A}\sum_{n=1}^{N}a_nx_n+\frac{1}{1-\theta_A}\sum_{n=1}^{N}a_n(M-x_n)=0\\ \frac{\partial}{\partial\theta_B}\Big[\mathrm{E}[\ln\left(P(\chi|z,\theta)\right)]\Big] &= \frac{1}{\theta_B}\sum_{n=1}^{N}b_nx_n+\frac{1}{1-\theta_B}\sum_{n=1}^{N}b_n(M-x_n)=0\end{aligned} \tag{14.1.1548}$$

Solving for θ_A and θ_B, we get

$$\hat{\theta}_A=\frac{\displaystyle\sum_{n=1}^{N}a_nx_n}{\displaystyle M\sum_{n=1}^{N}a_n}\qquad \hat{\theta}_B=\frac{\displaystyle\sum_{n=1}^{N}b_nx_n}{\displaystyle M\sum_{n=1}^{N}b_n} \tag{14.1.1549}$$

These are the both relations we used earlier above in our companion example!

We will see more complicated and practical business applications of the EM algorithm further below like:

- For our study of Gaussian Mixture Classification (related to K-mean)
- For our study of Factor Analysis (Exploratory Factor Analysis (EFA))
- For our study of T-distributed Stochastic Neighbour Embedding (T-SNE)

14.1.18 Resampling statistics

Resampling statistics refers to the use of the observed data or of a data generating mechanism (such as a die) to produce new hypothetical samples (resamples) that mimic the underlying population, the results of which can then be analysed. With numerous cross-disciplinary applications especially in the sub-disciplines of the life science, resampling methods are widely used since they are options when parametric approaches are difficult to employ or otherwise do not apply.

Resampled data is derived using a manual mechanism to simulate many pseudo-trials. These approaches were difficult to utilize prior to 11980s (holocene calendar) since these methods require many repetitions. With the incorporation of computers, the trials can be simulated in a few minutes and is why these methods have become widely used. The methods that will be discussed are used to make many statistical inferences about the underlying population. The most practical use of resampling methods is to derive confidence intervals and test hypotheses. This is accomplished by drawing simulated samples from the data themselves (resamples) or from a reference distribution based on the data; afterwards, you are able to observe how the statistic of interest in these resamples behaves. Resampling approaches can be used to substitute for traditional statistical (formulaic) approaches or when a traditional approach is difficult to apply. These methods are widely used because their ease of use. They generally require minimal mathematical formulas, needing a small amount of mathematical (algebraic) knowledge. These methods are easy to understand and stray away from choosing an incorrect formula in your diagnostics.

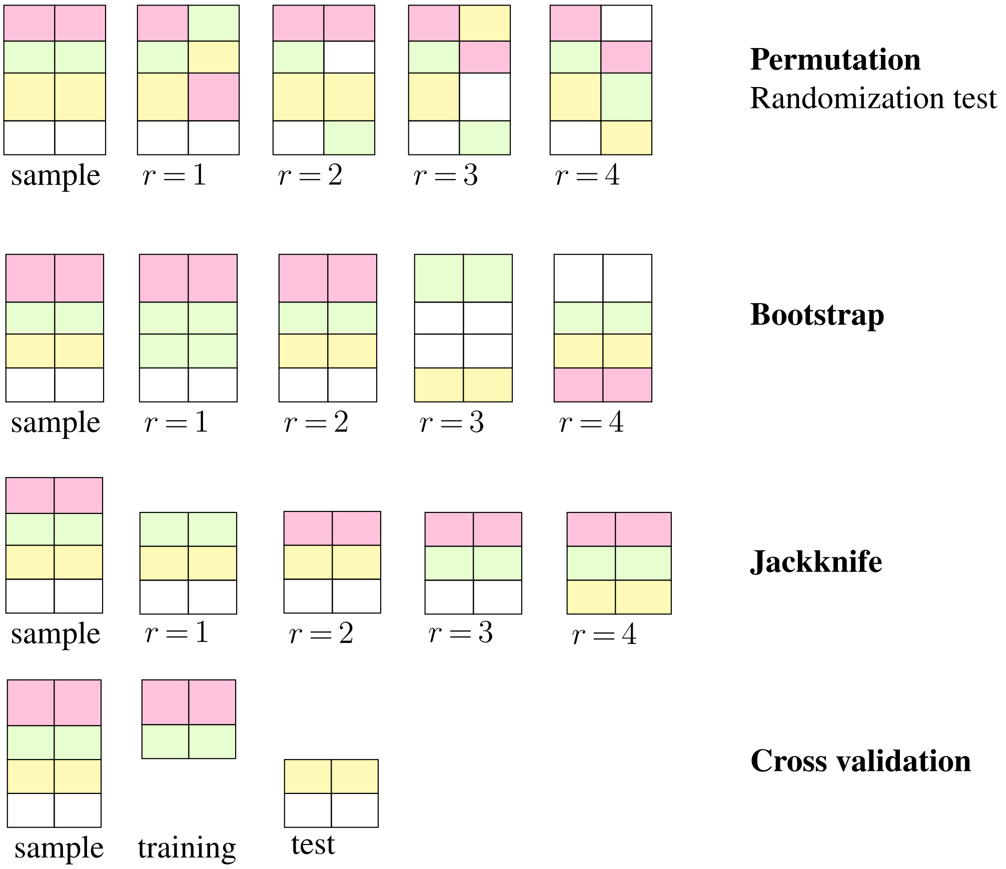


Figure 14.92: **Summary of resampling in different methods**

14.1.18.1 Monte Carlo Simulations

"Monte Carlo methods", also named "Monte Carlo experiments" or "Monte Carlo simulations" (MCS), are a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results. They are often used in physical and mathematical problems (finance, supply chain, decisioneering, quality, etc.) and are most useful as workaround when it is difficult or impossible to use other mathematical methods.

It finds applications in various fields including the following examples:

- Problems related to the neutron bomb (or any other problem of the same family: neutron diffusion)
- Calculations of integrals or various parameters of random variables (finance, insurance, risk, forecasting)
- Resolution of elliptic or parabolic equations
- Solving linear systems
- Optimization Problem Solving (operations research, project management, supply chain)
- Creation of statistical tests (Anderson-Darling, Kolmogorov, Levene, Brown-Forsythe, etc.)

Thus there are two types of problems that can be treated by the Monte Carlo probabilistic method:

1. Problems, which have a random behaviour
2. Deterministic problems, which do not have a random behaviour

About the probabilistic case, the idea is to observe the behaviour of a series of random numbers that simulates how the real problem behaves and derive statistical solutions/conclusions. We then speak of "Monte Carlo estimation".

In the deterministic case, the studied problem is completely defined and we can in principle predict its evolution, but some parameters of the problem can be treated as if it were random variables (this is typically the case in "vector regression" technique in Economy). The deterministic problem then becomes probabilistic and still solvable numerically. We then speak of "Monte Carlo elaborated estimation".

We will begin further below with the most used one in business: generating draws from a probability distribution. For this purpose we will need to introduce the "inverse transform sampling" (we will treat mostly only univariate case in this textbook!).

Remark

The name of "Monte Carlo method" dates around 11944 (holocene calendar). Isolated researchers have used however long before similar statistical methods: for example, Edwin Herbert Hall for the experimental determination of the speed of light (11873 according to holocene calendar), or Kelvin in a discussion of the Boltzmann equation (11901 according to holocene calendar), but the real use of Monte Carlo methods began with research on the atomic bomb.

During the immediate postwar period, John Von Neumann, Enrico Fermi and Stanislaw Ulam warned the scientific community of the applicability of the Monte Carlo methods (eg for the approximation of the eigenvalues of the Schrödinger equation). The systematic study was made by Harris and Herman Khan in 11948 (holocene calendar). After an eclipse caused by too intensive use during the 11950s (holocene calendar), the Monte Carlo method is back since almost everybody can run complex management or business strategic simulations on office computers (with simple softwares like @Risk or Crystal Ball). in short, wherever it is profitable to use simulation processes.

There are ways of using probabilities that are definitely not Monte Carlo simulations – for example, deterministic modelling using single-point estimates, commonly named "what if analysis". Each uncertain variable within a model is assigned a "best guess" estimate. Scenarios (such as best, worst, or most likely case) for each input variable are chosen and the results recorded. By contrast, Monte Carlo simulations sample from a probability distribution for each variable to produce hundreds, thousands or millions of possible outcomes. The results are analysed to get probabilities of different outcomes occurring. But what if analysis gives equal weight to all scenarios and that's their major issue!

Also the reader should know that Monte Carlo simulation (MCS) and historical simulation (HS) are both methods that can be used to determine the riskiness of a financial project. However, each method uses different assumptions and techniques to develop the probability distribution of possible outcomes!

"Historical simulation" (HS) involves, as indicated by its name..., the use of historical records to simulate the possible outcomes. The method assumes that past performance is an indication of future performance. Historical simulation uses actual past figures or variables that have been experienced before. Monte Carlo simulation comes with the advantage of incorporating a wider variety of scenarios than historical data, whose information scope is limited. In addition, Monte Carlo simulation answers the what if question, weighted by probabilities (!), which is not possible under historical simulation.

14.1.18.1.1 Inverse Transform Sampling

Inverse transform sampling (also known as inversion sampling, inverse probability integral transform, inverse transformation method, Smirnov transform, golden rule,) is a basic method for pseudo-random number sampling, i.e. for generating sample numbers at random from any probability distribution given its cumulative distribution function.

Inverse transformation sampling takes uniform samples of a number $\mathcal{U}$ between 0 and 1, interpreted as a probability, and then return the largest number x from the domain of the distribution $P(X)$ such that $P(-\infty < X < x) \leq \mathcal{U}$.

Computationally, this method involves computing the quantile function of the distribution — in other words, computing the cumulative distribution function (CDF) of the distribution (which maps a number in the domain to a probability between 0 and 1) and then inverting that function.

To use this we method we go through he "probability integral transform" that states that if X is a continuous random variable with cumulative distribution function F_X, then the random variable $Y = F_X(X)$ has a uniform distribution on $[0,1]$. The inverse probability integral transform is just the inverse of this: specifically, if Y has a uniform distribution on $[0,1]$ and if X has a cumulative distribution F_X, then the random variable $F_X^{-1}(Y)$ has the same distribution as X.

Theorem 14.6. Suppose that a random variable X has a distribution for which the cumulative distribution function (CDF) is F_X. Then the random variable Y defined as:

$$Y = F_X(X) \tag{14.1.1550}$$

has a uniform distribution.

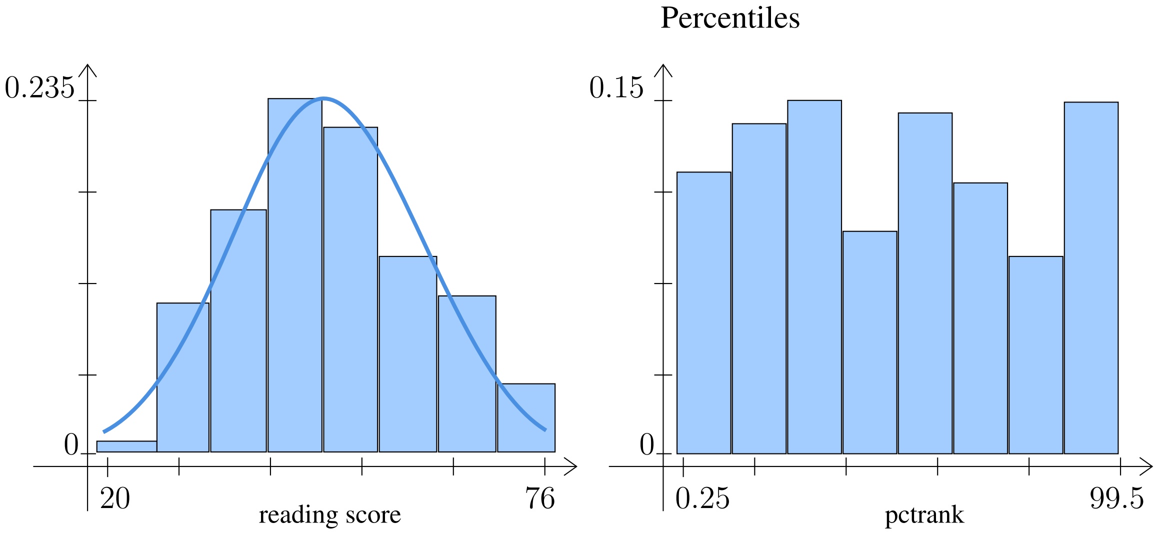
Proof 14.6.1. Given any random variable X, define $Y = F_X(X)$. Then (it is not always easy to read this proof through the first time even if afterwards it is obvious):

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(F_X(X) \leq y) \\ &= P(X \leq F_X^{-1}(y)) \\ &= F_X(F_X^{-1}(y)) \\ &= y \end{aligned} \tag{14.1.1551}$$

Therefore F_Y is just the CDF of the uniform random variable $\mathcal{U}[0,1]$. Thus, Y has a uniform distribution on the interval $[0,1]$ (or in other words, if X is a random variable with CDF $F_X(X)$ then $F_X(X) = \mathcal{U}[0,1]$).

■ Q.E.D.

In other words, the percentiles are uniformly distributed (i.e. the frequency plot of the percentiles is uniformly distributed) as illustrated below:



Caution! In NHST (Null Hypothesis Statistical Tests) the p-value itself is considered as a realization of random variable. Therefore if we replace the random variable X above by the random variable p, we get the famous result that p-value is uniformly distributed when the null hypothesis is true (indeed under the null H_0 there is no reason that the value of p is more located in given place than in another one by definition!).

The problem that the inverse transform sampling method solves is as follows:

- Let X be a random variable whose distribution can be described by the cumulative distribution function F_X.
- We want to generate values of X which are distributed according to this distribution.

The inverse transform sampling method works as follows:

1. Generate a random number $\mathcal{U}$ from the standard uniform distribution in the interval $[0,1]$.
2. Compute the value x such that $F_X(x)=\mathcal{U}$ (using $F_X^{-1}(\mathcal{U})$).
3. Take x to be the random number drawn from the distribution described by F_X.

Expressed differently, given a continuous uniform variable $\mathcal{U}$ in $[0,1]$ and an invertible cumulative distribution function F_X, the random variable $X=F_X^{-1}(\mathcal{U})$ has distribution F_X (or, X is distributed F_X).

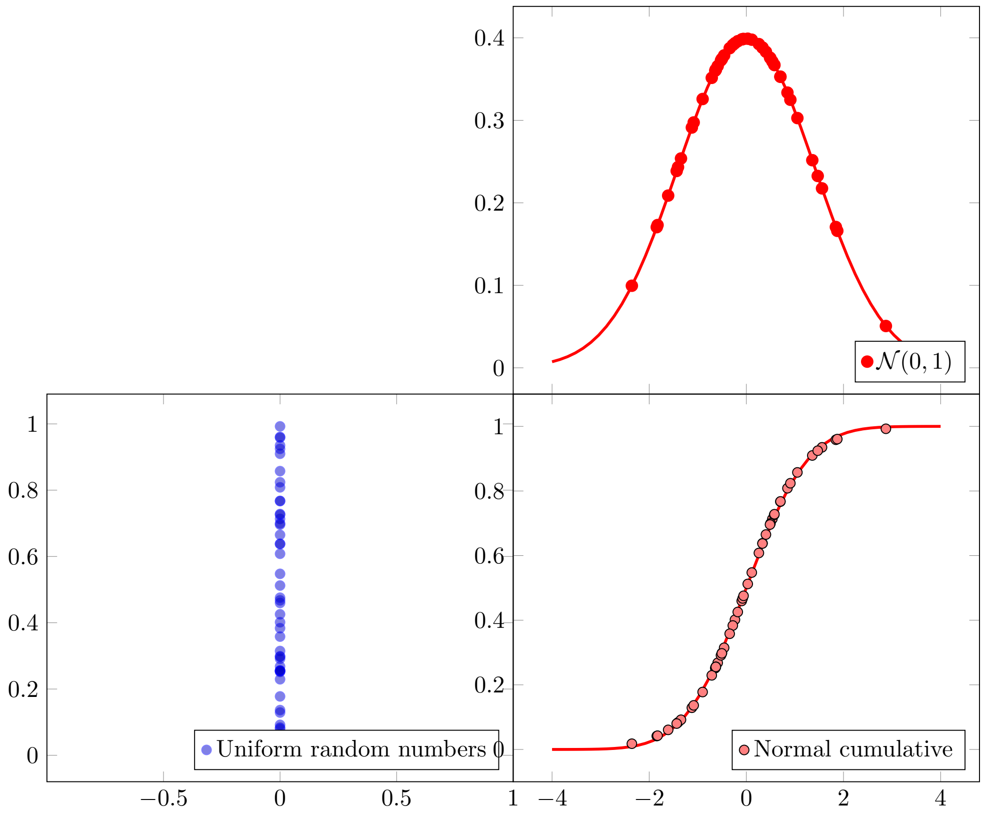


Figure 14.93: **Principle of the inverse transform method**

14.1.18.1.2 Random number generation

The best way to understand the method of Monte Carlo is to make examples (even small one should be enough). But for this, we must first have a good random number generator (which is quite difficult depending on the job). This is a very delicate and sensitive field for which an international standards is published (ISO 28640:2010 *Random variate generation methods*).

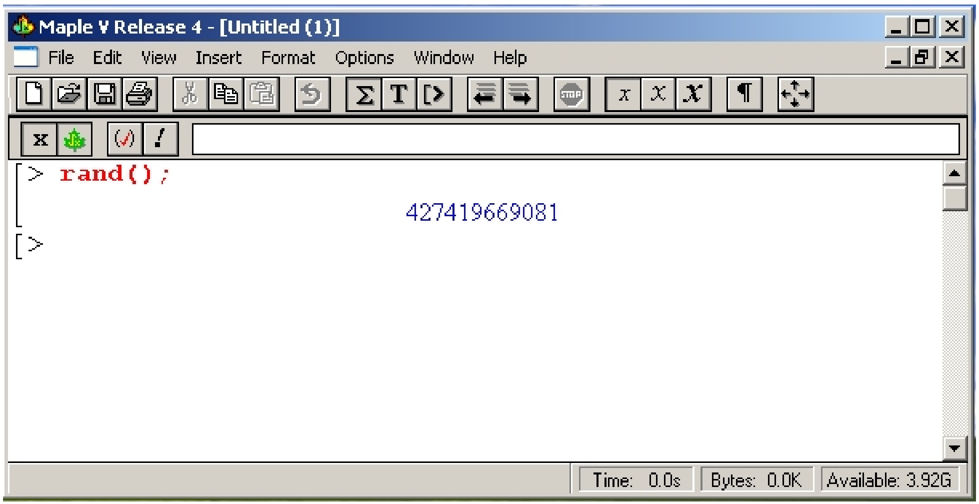
A random-number generator (RNG) is a computational or physical device designed to generate a sequence of numbers or symbols that cannot be reasonably predicted better than by a random chance.

Several computational methods for random-number generation exist. Many fall short of the goal of true randomness, although they may meet, with varying success, some of the statistical tests for randomness intended to measure how unpredictable their results are.

There are two principal methods used to generate random numbers. The first method measures some physical phenomenon that is expected to be random and then compensates for possible biases in the measurement process. Example sources include measuring atmospheric noise, thermal noise, and other external electromagnetic and quantum phenomena. For example, cosmic background radiation or radioactive decay as measured over short time-scales represent sources of natural entropy.

The second method uses computational algorithms that can produce long sequences of apparently random results, which are in fact completely determined by a shorter initial value, known as a seed value or key. As a result, the entire seemingly random sequence can be reproduced if the seed value is known. This type of random number generator is often named a "pseudo-random number generator". This type of generator typically does not rely on sources of naturally occurring entropy, though it may be periodically seeded by natural sources. This generator type is non-blocking, so they are not rate-limited by an external event, making large bulk reads a possibility.

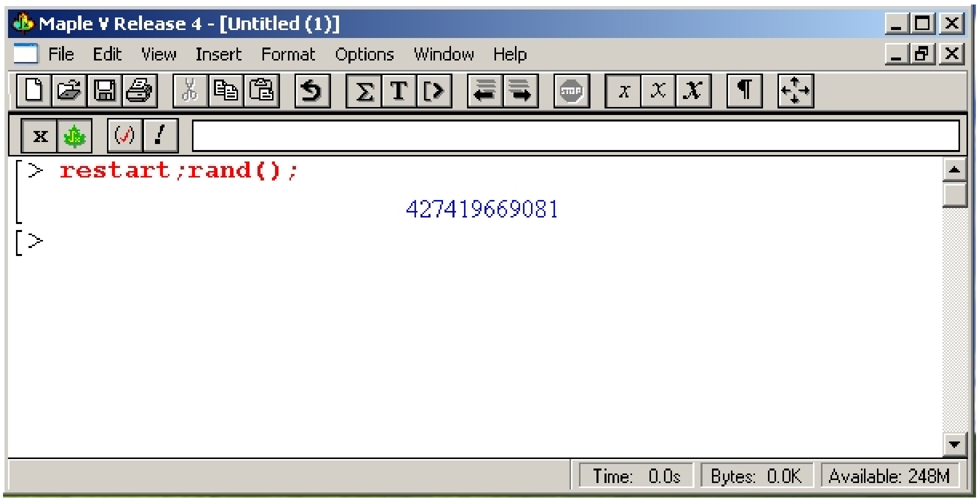
Let us take, to begin, an example with the Maple™ 4.00b random generator:


Maple V Release 4 - [Untitled (1)]
File Edit View Insert Format Options Window Help
Time: 0.0s
Bytes: 0.0K
Available: 3.92G

```
> rand();
                              427419669081
>
```

Figure 14.94: **Pseudo-random generator in Maple™ 4.00b**

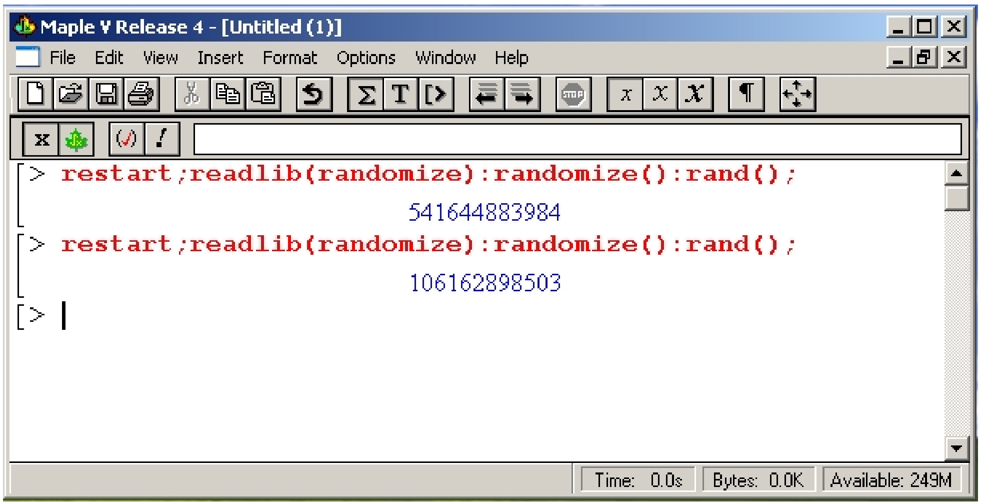
and:


Maple V Release 4 - [Untitled (1)]
File Edit View Insert Format Options Window Help
Time: 0.0s
Bytes: 0.0K
Available: 248M

```
> restart;rand();
                          427419669081
>
```

Figure 14.95: Pseudo-random generator restart with Maple™ 4.00b

So we see that the default random number generator used by default in Maple™ 4.00b should be used with extreme caution since a system reset is enough to find... equal random values! This is therefore as we already said "pseudo-random generator" that gives the possibility to make what we name "pseudo Monte Carlo method".


Maple V Release 4 - [Untitled (1)]
File Edit View Insert Format Options Window Help
Time: 0.0s
Bytes: 0.0K
Available: 249M

```
> restart;readlib(randomize):randomize():rand();
                          541644883984
> restart;readlib(randomize):randomize():rand();
                          106162898503
> |
```

Figure 14.96: Pseudo-random library with Maple™ 4.00b

The `RAND( )` and `RANDBETWEEN( )` functions of the of Microsoft® Excel® 14.0.6123 are also pseudo-random generators which here is a sample of 100 simulations (of course in Microsoft® Excel® the chart below will change each time you press on the keyboard button: F9):

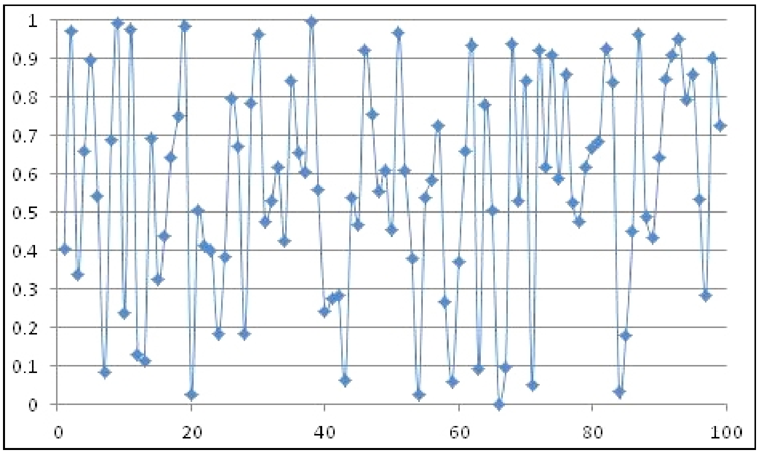


Figure 14.97: **Illustration of a sequence of pseudo-random numbers with Microsoft® Excel® 14.0.6123**

Unfortunately, it may happen with pseudo-random numbers that the numbers generated are presented in bunches, that is to say by sets of numbers close to each other, which reduces the effectiveness of the Monte Carlo simulation .

An empirical technique is to use then sequences of numbers generated by algorithms that scan almost surely the range $[0, 1]$. We speak then of "quasi-random numbers" to make simulations sometimes named "quasi-Monte Carlo". In almost all Microsoft® Excel®, you can create a Visual Basic Application function that will replace the pseudo-random generators that are the `RAND( )` or `RANDBETWEEN( )`.

Here is an example of such a V.B.A. function which generates quasi-random number named "Fauré random number sequence":

Listing 14.1: **VBA Fauré sequence code**

```
Function SequenceFaure(n) As Double
    Dim f As Double, sb As Double
    Dim i As Integer, n1 As Integer, n2 As Integer

    n1 = n
    sb = 1 / 2
    Do While n1 > 0
        n2 = Int(n1 / 2)
        i = n1 - n2 * 2
        f = f + sb * i
        sb = sb / 2
        n1 = n2
    Loop
    SequenceFaure = f
End Function
```

This will gives the following sequence for a sample of 100 simulations:

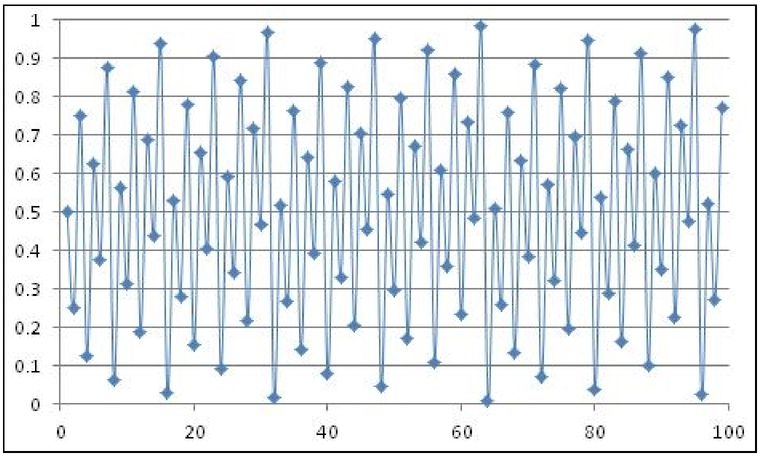


Figure 14.98: **Illustration of a Fauré sequence of pseudo-random numbers with Microsoft® Excel® 14.0.6123**

where we see well that the sequence covers well the whole area between 0 and 1 (we say then that: it covers faster the integration surface). This technique is sometimes preferred because it has the advantage of keeping the values off the simulation every time we restart the simulation (therefore in Microsoft® Excel® the chart above will not change when you press the keyboard button F9).

By cons the sequence generators have a great weakness: they are only applicable (to my knowledge at least) for problems of simulations with a single random variable (typically pricing single option strategy following Black & Scholes model). Indeed if we have several random variables (and this is the most common case!), then the variables are artificially correlated (correlation coefficient $= 1$) because they travel all the area between 0 and 1 in the same way! So a good simulation with several variables is a simulation including the treated variables have a correlation coefficient which approaches zero!!!!!!

In addition, sequence generators require algorithms that are very time consuming when there are many variables relatively to a pseudo-random generator, this is why in most situations we prefer the old methods.

Remark

As we have already mentioned it, engineers should refer to the international standard ISO 28640:2010 *Random variate generation methods* when they need to implement random number generators in their softwares.

Once the pseudo-random or random generator created and tested, we can see some application of the Monte Carlo method before continuing on performance tricks relatively to this method. Thus, in the calculation of integrals, this method is very useful and very fast in terms of convergence speed.

14.1.18.1.3 Monte Carlo integration

In numerical integration, methods such as the Trapezoidal rule use a deterministic approach as we already know. Monte Carlo integration, on the other hand, employs a non-deterministic approaches: each realization provides a different outcome. In Monte Carlo, the final outcome is an approximation of the correct value with respective error bars, and the correct value is within those error bars.

Consider for example the calculation of the following univariate defined integral of a function f and positive over the interval $[a,b]$:

$$I=\int\limits_a^b f(x)\mathrm{d}x \tag{14.1.1552}$$

Given:

$$m=\max(f(a),f(b)) \tag{14.1.1553}$$

the maximal value of the function f between the bounds $[a,b]$.

We consider the rectangle bounding function on the interval $[a,b]$ defined by vertices $\{(a,0),(b,0),(b,m),(a,m)\}$:

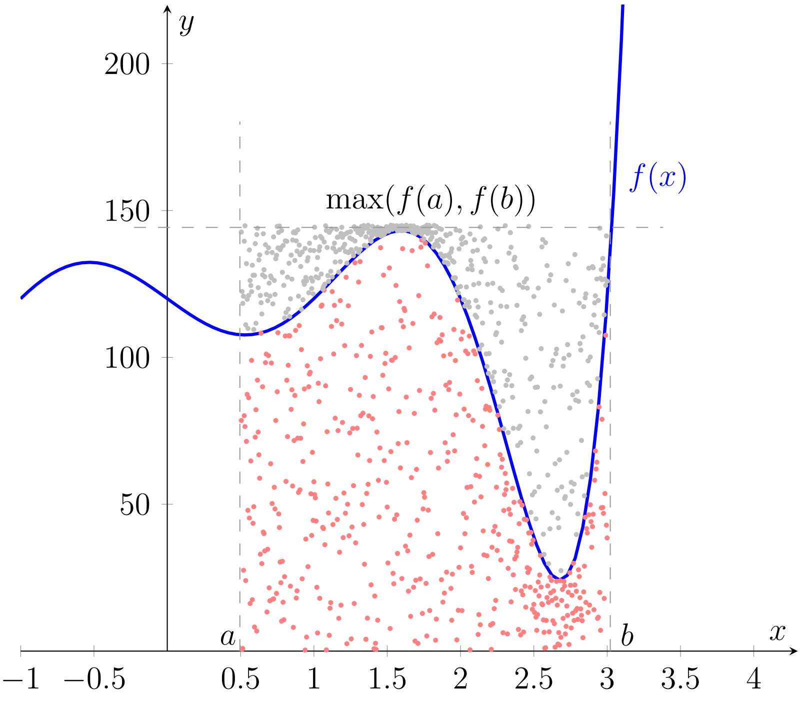


Figure 14.99: Basic principle of the univariate integral calculation with Monte Carlo

We draw a large number N of random points in this rectangle. For each point, $\xi_i=(x,y)$ we test if it is below the blue curve. Given P the proportion of points below this curve, we have:

$$\boxed{I_{\text{crude}}=\lim_{N\rightarrow+\infty}(b-a)m\frac{k}{N}} \tag{14.1.1554}$$

This approach seems to be sometimes named the "Monte Carlo Stochastic Point Method" or Monte Carlo hit-or-miss approach" or "Monte Carlo poor man's sampling".

The corresponding Maple™ 4.00b algorithm is given by:

```
>intmonte:=proc(f,a,b,N)
local i,al,bl,m,P,aleaabs,aleaord,isabove;
m:=round(max(f(a),f(b))*10^4);
al:=round(a*10^4);
bl:=round(b*10^4);
aleaabs:=rand(al..bl);
aleaord:=rand(0..m);
k:=0;
for i from 1 to N do
isabove:=(f(aleaabs()/10^4)-aleaord()/10^4)>=0;
if isabove then
k:=k+1;
fi
od:
RETURN((b-a)*max(f(a),f(b))*k/N)
end:
```

To call this procedure in Maple™, just write `>intmonte(f,a,b,N)` but replacing the first argument passed as a parameter with the expression of a function and the other arguments by numerical values (obviously!).

Notice that the relation above can be rewritten:

$$I_{\text{crude}} = \lim_{N\rightarrow +\infty} (b-a)m\hat{p} \tag{14.1.1555}$$

where $\hat{p}$ is the estimated parameter of a Bernoulli distribution and the corresponding random variable is ξ_i if we reject or accept the generated point:

$$\xi_i = \begin{cases} 1 & \text{if } y \in [0, f(x)] \\ 0 & \text{otherwise} \end{cases} \tag{14.1.1556}$$

Then $\{\xi_i\}$ are outcomes of independent duplicate trials and $\xi_1, \cdots, \xi_N \overset{i.i.d.}{\sim} \mathrm{B}(1,\hat{p})$. So if $\{\xi_i\}$ follows a Bernoulli distribution, then:

$$I_{\text{crude}} = (b-a)m\frac{P}{N} \tag{14.1.1557}$$

follows a binomial distribution $\mathcal{B}(N,k)$.

We know that the mean standard deviation of a variable (without factors) following a binomial distribution is:

$$\mu = np \qquad \sigma = \sqrt{npq} = \sqrt{np(1-p)} \tag{14.1.1558}$$

But here we have a factor $(b-a)m/N$, so that the mean becomes:

$$\mu=\frac{(b-a)m}{N}Np=(b-a)mp=(b-a)m\frac{k}{N}=I_{\text{crude}} \qquad (14.1.1559)$$

and the standard deviation will be:

$$\sigma_{I_{\text{crude}}}=\frac{(b-a)m}{N}\sqrt{Np(1-p)}=\sqrt{\frac{(b-a)^2m^2Np(1-p)}{N^2}}=\sqrt{\frac{(b-a)^2m^2p(1-p)}{N}} \qquad (14.1.1560)$$

The both latter relation are often written in textbooks by considering (implicitly) that the range of integration has been transformed to have an interval $a=0,b=1$ and the function transformed such that maximum $m=1$. Therefore:

$$\mu=\frac{1}{N}Np=p=I_{\text{crude}} \qquad \text{and} \qquad \sigma_{I_{\text{crude}}}=\sqrt{\frac{p(1-p)}{N}}=\sqrt{\frac{I_{\text{crude}}(1-I_{\text{crude}})}{N}} \qquad (14.1.1561)$$

To estimate the variance (or precision), we use the Central Limit Theorem and get[35] (don't forget that this relation is valid only if the integration has been transformed to have an interval $a=0,b=1$ and the function transformed such that maximum $m=1$):

$$\frac{\hat{I}_{\text{crude}}-I_{\text{crude}}}{\mathrm{V}(I_{\text{crude}})}=\frac{\hat{I}_{\text{crude}}-I}{\sqrt{\frac{I(1-I)}{N^3}}}\xrightarrow{D}\mathcal{N}(0,1) \qquad (14.1.1562)$$

That's a useful result to build confidence interval for that Monte Carlo method.

Let us consider another very common approach named "Monte Carlo mean value method" (m.v.) or "Monte Carlo uniform sampling method"!

Let $g(x)$ be a function and suppose that we want to compute:

$$I=\int_0^1 g(x)\mathrm{d}x \qquad (14.1.1563)$$

Recall that if X is a random variable with density $f(x)$, then the mathematical expectation of the random variable $Y=g(X)$ between 0 and 1 is given by:

$$\mathrm{E}(g(X)_{[0,1]})=\int_0^1 g(x)f(x)\mathrm{d}x \qquad (14.1.1564)$$

If a random sample is available from the distribution of X, an unbiased estimator of $\mathrm{E}(g(X))$ is the sample mean:

$$\mathrm{E}(g(X)_{[0,1]})\cong\hat{I}=\sum_{i=1}^{N}g(X_{[0,1]}^{(i)})f(X_{[0,1]}^{(i)}) \qquad (14.1.1565)$$

---

[35] Remember that the denominator is the standard error of the mean $\sigma_\mu=\sigma/\sqrt{N}$

If we take a distribution $f(X_{[0,1]}^{(i)})$ that is equally likely between 0 and 1 like the Uniform distribution $\mathcal{U}_{[0,1]}$, then the above relation reduces to:

$$I=\int\limits_0^1 g(x)\mathrm{d}x\cong\frac{1}{N}\sum_{i=1}^N g(\mathcal{U}_{[0,1]}) \qquad (14.1.1566)$$

To compute $\int_a^b g(t)\mathrm{d}t$ we make a change of variables so that the limits of integration are from 0 to 1. The linear transformation is $y=(t-a)/(b-a)$ and $\mathrm{d}t=(b-a)\mathrm{d}y$. Therefore:

$$I=\int\limits_a^b g(t)\mathrm{d}t=\int\limits_0^1 g(y(b-a)+a)(b-a)\mathrm{d}y=(b-a)\int\limits_0^1 g(y(b-a)+a)\mathrm{d}y\cong\frac{b-a}{N}\sum_{i=1}^N g(\mathcal{U}_{[a,b]}) \qquad (14.1.1567)$$

Hence:

$$\boxed{I_{\text{m.v.}}=\lim_{N\rightarrow+\infty}\frac{b-a}{N}\sum_{i=1}^N g(\mathcal{U}_{[a,b]})} \qquad (14.1.1568)$$

By the fundamental theorem of Monte Carlo (see section Statistics page 924) it converges with probability 1.

Notice also that by the Central Limit Theorem we have that the standard error of $I_{\text{m.v.}}$ is given by:

$$\sigma_{I_{\text{m.v.}}}=\frac{\sigma}{\sqrt{N}}=\frac{\mathrm{V}(g(X))}{\sqrt{N}}=\frac{\sqrt{\int_0^1(g(x)-I)^2\mathrm{d}t}}{\sqrt{N}}=\sqrt{\frac{\int_0^1(g(x)-I)^2\mathrm{d}t}{N}} \qquad (14.1.1569)$$

Or using Huygens relation, given for recall by $\mathrm{V}(X)=\mathrm{E}(X^2)-\mathrm{E}(X)^2$, we get:

$$\sigma_{I_{\text{m.v.}}}^2=\frac{1}{N}\mathrm{E}(g(x)^2)-\frac{I^2}{N} \qquad (14.1.1570)$$

This is useful to put confidence limits or error bounds on the Monte Carlo estimate of the integral, and check for convergence.

However, the m.v. method (Monte Carlo mean value method) has two major limitations:

- It does not apply to unbounded intervals!
- It can be inefficient to draw samples uniformly across the interval if the function $g(x)$ is not very uniform!

Remark

The reader can refer to our `R` or MATLAB® free companion books to see examples with the respective script language of these two softwares for the both method presented above.

Now let us just compare:

$$\sigma^2_{I_{\text{m.v.}}} - \sigma^2_{I_{\text{crude}}} = \left[\frac{1}{N}\mathrm{E}(g^2) - \frac{I^2}{N}\right] - \frac{I(1-I)}{N} = \frac{1}{N}\mathrm{E}(g^2) - \frac{I^2}{N} - \frac{I}{n} + \frac{I^2}{n} = \frac{1}{N}(\mathrm{E}(g^2) - I)$$
$$= \frac{1}{N}(\mathrm{E}(g^2) - \mathrm{E}(I)) = \frac{1}{N}\int_0^1 g(x)(f(x)-1)\mathrm{d}x \qquad (14.1.1571)$$

Moreover, many textbooks claim that:

$$\frac{1}{N}\int_0^1 g(x)(f(x)-1)\mathrm{d}x > 0 \qquad (14.1.1572)$$

But this is neither true nor false!

14.1.18.1.4 Monte Carlo estimation of π

For the calculation of π the principle is the same and therefore consist to use the proportion of the number of points in a quarter of circle area (this simplifies the algorithm by restricting the calculations to strictly positive coordinates) inscribed in a square of side 1 (so the radius of the circle is also equal to 1 obviously) relatively to the total number of points (to test if a point is outside the circle, we obviously use the Pythagorean theorem) such that:

$$\pi = \lim_{N\rightarrow+\infty} 4\frac{P}{N} \qquad (14.1.1573)$$

Because we have for the number of points P inside the quarter or circle that corresponds to a surface of $\pi R^2/4$ and the number of points N inside surface of the quarter of square that corresponds to a surface R^2. So the ratio of both should tend to $\pi/4$ when $N\rightarrow+\infty$. We just have to multiply by 4 to get a good approximation of π.

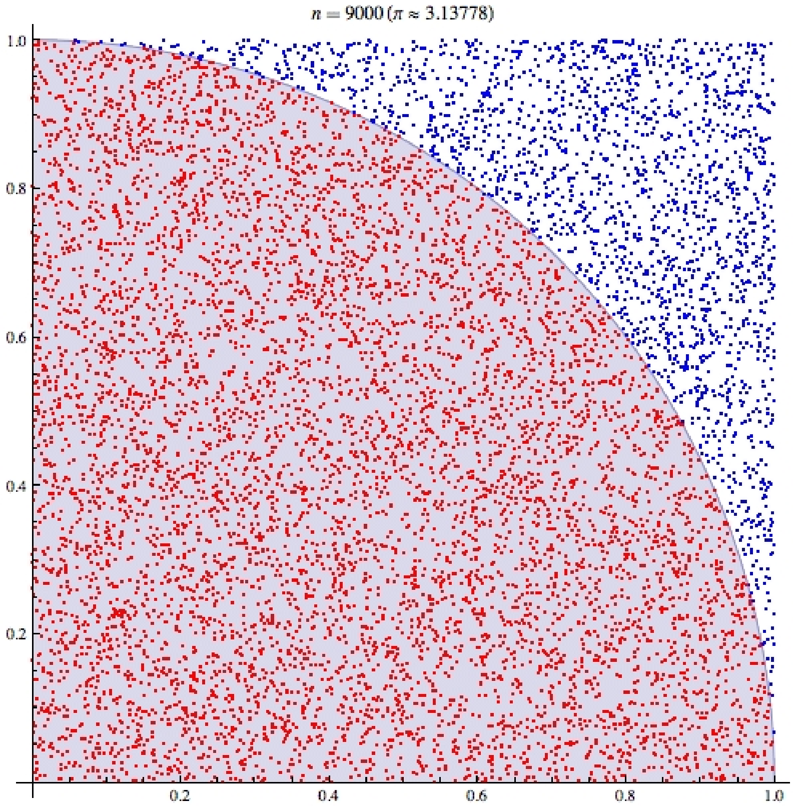


Figure 14.100: **Monte Carlo π estimate**

That method is equivalent to defining an indicator variable I as:

$$I = \begin{cases} 1 & \text{if } X^2 + Y^2 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

where $X, Y = \mathcal{U}_{[0,1]}$. Where that:

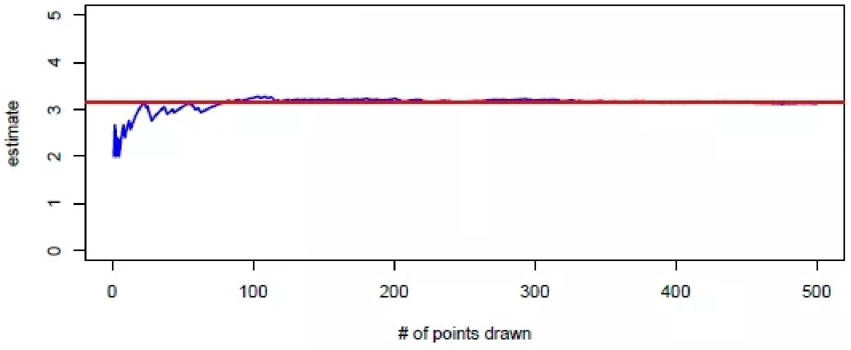
$$\mathrm{E}(I) = \frac{\pi}{4} \quad \text{and} \quad \mathrm{V}(I) = \frac{\pi}{4}\left(1 - \frac{\pi}{4}\right)$$

The corresponding Maple™ 4.00b algorithm is given by:

```
>isinside:=proc(x,y) x^2+y^2<1 end:
>calculatepi:=proc(N)
local i,P,abs,ord,alea;
alea:=rand(-10^4..10^4);
P:=0;
for i from 1 to N do
abs:=alea()/10^4;ord:=alea()/10^4;
if isinside(abs,ord) then
P:=P+1;
fi
od;
RETURN(4*P/N)
```

```
end:
>evalf(calculatepi(100));evalf(calculatepi(1000));
evalf(calculatepi(10000));evalf(calculatepi(100000));
```

In terms of convergence it looks typically as:



Remark

The reader can refer again to our `R` or MATLAB® free companion books to see examples with the respective script language of these two softwares for the example above.

14.1.18.1.5 Monte Carlo Modelling

The most common application of the Monte Carlo method in business and industry is certainly the study of random variables. Furthermore, this method is part of the ISO 31010 Risk Management standard under the name of "Monte Carlo analysis". Many cutting edge tech companies make Monte Carlo modelling with a spreadsheet softwares like Microsoft® Excel® (even multinationals!) and with a lesser extent with professional oriented softwares or add-ins such as @RISK, Crystal Ball, TreeAge, Isograph or MATLAB® (without taking into account C, C++, Java, `R`, and Python™ obviously).

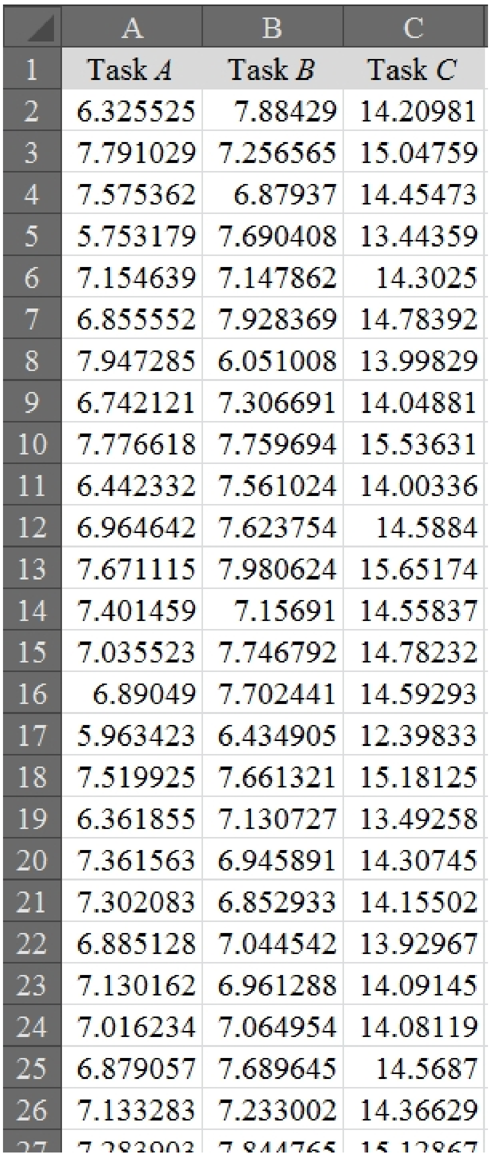
The advantages of this method in modelling random variables are:

- We can integrate any distribution including empirical one and not continuous one!
- Models are very easy to implement and can be expanded as needed without too much effort.
- All influences or relation occurring in reality (at least the identified one....) may be represented and implemented in the model.
- Sensitivity analysis (see section Quantitative Management Techniques page 8184) can be applied.
- The models are easily understandable and provide a measure of the accuracy of the result.

- Many inexpensive software are available (at least inexpensive in comparison to criticality of the business analysed that is most of time in the order of the billion of dollars).

Let us consider a simple but concrete case (widely used in business) that I like to use in my introduction training of a small project of two tasks denoted by A and B which follow each other without free margin (or free slack). Let us imagine that the duration of each task has been estimated in accordance with the recommendation of the Project Management Institute with a beta distribution (see section Statistics page 760) as learn it almost all good project managers in their training curriculum (see section Quantitative Management Techniques page 8074).

For this example, the task A has an optimistic duration of 5 days and a pessimistic duration of 8 days. Task B an optimistic duration of 1 day and a pessimistic duration of 4 days. We would like in the spreadsheet software Microsoft® Excel® using a pseudo Monte Carlo simulation (therefore necessarily based on a pseudo-random variable) introduce three traditional minimum information:



	A	B	C
1	Task *A*	Task *B*	Task *C*
2	6.325525	7.88429	14.20981
3	7.791029	7.256565	15.04759
4	7.575362	6.87937	14.45473
5	5.753179	7.690408	13.44359
6	7.154639	7.147862	14.3025
7	6.855552	7.928369	14.78392
8	7.947285	6.051008	13.99829
9	6.742121	7.306691	14.04881
10	7.776618	7.759694	15.53631
11	6.442332	7.561024	14.00336
12	6.964642	7.623754	14.5884
13	7.671115	7.980624	15.65174
14	7.401459	7.15691	14.55837
15	7.035523	7.746792	14.78232
16	6.89049	7.702441	14.59293
17	5.963423	6.434905	12.39833
18	7.519925	7.661321	15.18125
19	6.361855	7.130727	13.49258
20	7.361563	6.945891	14.30745
21	7.302083	6.852933	14.15502
22	6.885128	7.044542	13.92967
23	7.130162	6.961288	14.09145
24	7.016234	7.064954	14.08119
25	6.879057	7.689645	14.5687
26	7.133283	7.233002	14.36629
27	7.283903	7.844765	15.12867

- A table with 3 columns (duration of A, B and sum of both) and $10,000$ simulations (rows)

- The graphical distribution function of the sum of the two random variables
- The convergence of the 95th percentile on the 100 first simulations (useful for the subject further below).

We then construct the following table of $10,000$ row (the screenshot shows only the first rows...) where all cells from row 2 to row $10,000$ of column `A` contains the following function (Microsoft® Excel® 14.0.7166):

```
=BETA.INV(RAND(),3+SQRT(2),3-SQRT(2),5,8)
```

and for column `B` contains the following function:

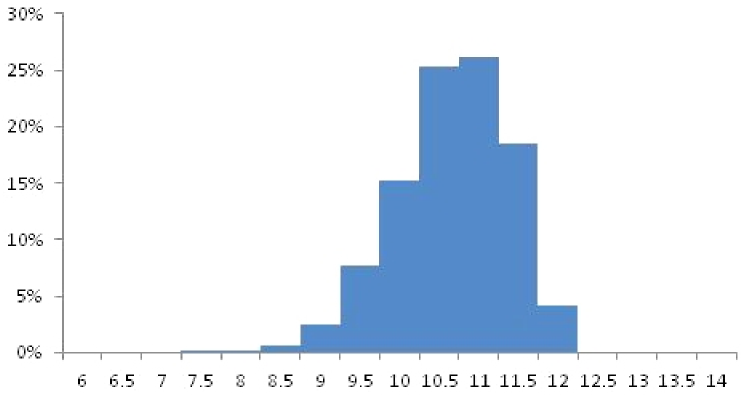
```
=BETA.INV(RAND(),3+SQRT(2),3-SQRT(2),1,4)
```

and finally the cell `C1` contains the following function that was pull down until cell `C10000`:

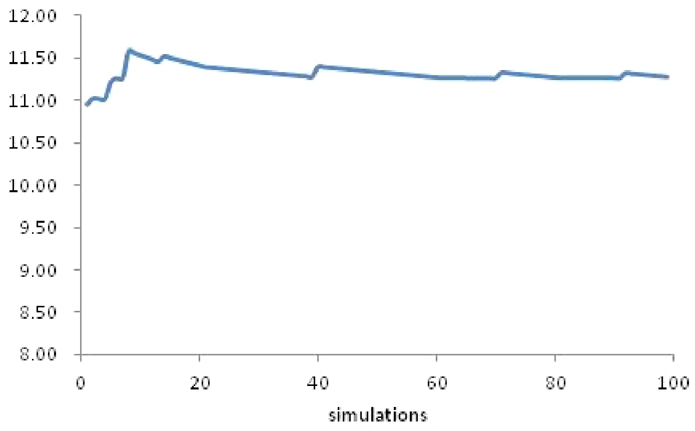
```
=A2+B
```

Obviously the values in Microsoft® Excel® 14.0.6123 will change each time you press the `F9` key on the keyboard.

Then this gives us the histogram still made with the same software version (we will not detail how to building such a chart it is a basic subject of Microsoft Office knowledge and has nothing to do in a scientific book) where the x-axes are the number of days:



and the convergence of the 95th percentile of the first 100 simulations (because as this example is simple, the system converges quickly enough so that we do not to need to take more than 100 simulations as example) where the y-axes is the number of days:



Obviously by default, in Microsoft® Excel® 14.0.6123 the chart above will change each time you press the `F9` key on the keyboard.

Remark

In the case of the simulations of random variables, we can in simple situations involving only sums or subtractions of random variables, as for the example above, determine the mean and the standard deviation of the results analytically using the property of linearity of the mean and variance (because normally for the variance of two independent random variables, the covariance is zero). By analysing the difference between the analytical value and that obtained by numerical simulation, the offset can be corrected certain other statistical indicators by simply adding or subtracting the differential. This is known as the technique of "control variables" that we will detail further below.

There are other variance reduction techniques (i.e. standard deviation) than control variables techniques to reduce the variance of the Monte Carlo estimators in specific conditions:

- One of these techniques is the use of "antithetic variables" which consists very simply (programming this technique is at the high school level as you can see it in the MATLAB® companion book) to decorrelate simulations to make the covariance between the variables negative and so reduce the global variance (such as we have seen in the section Statistics, the variance of the sum of two random variables make appear a covariance term). Unfortunately, this technique works satisfactorily with symmetric distributions this is why to my knowledge it is not implemented in simulation software available on the market.

- There is also the technique named "stratified sampling" that consist to cut the pre-image space of the random variable in regular intervals (the programming of this technique is also at the high school level as you can see in the MATLAB® companion book). This technique works very well when the number of simulations must be small but only in the case of a single variable. This also why, as far as we I know, it is not implemented in simulation softwares available on the market.

- There exist is a generalization of stratified sampling (the programming of this technique is also at the high school level) for simulations with multiple variables and that is named "Latin Hypercube" (abbreviated as "LHS" for Latin Hypercube Stratification). This technique ensures that each n-tuple of random variables (corresponding to a space of n dimensions) uses a unique pre-image at each iteration, hence the name of that technique (Latin: refers to magic squares where each value appears uniquely, Hypercube because is an n-dimensional generalization of a magic square). Some simulation software available on the market implement this technique (@RISK, Crystal Ball).

To summarize, whether that it's the technique of Faure sequence generators, of antithetic variables, of control variables, of stratified sampling or Latin Hypercube even if these techniques are all easy to program, the method using the pseudo-random variables is privileged because is the most suitable for the majority of common situations in the business and in non-cutting edge scientific applications.

Remark

For the reader interested to have a deeper overview, however with out too much mathematical details, we recommend the following actual reference: *Introducing Monte Carlo Methods with R* (see [464]).

14.1.18.1.6 Monte Carlo variance reduction

What we have seen so far is only what we may name the "crude Monte Carlo simulation". We will now introduce the problem of reducing the variance of the samples that are generated to estimate the mean value λ of some random variable X using Monte Carlo method. The main underlying idea is to reduce the standard error and increasing the confidence in the accuracy of the estimate (very important in speculative finance and nuclear or quantum physics!). We will see further below four variance-reduction methods (among seven existing one).

Suppose one wants to compute $z := \text{E}(Z)$ with the random variable Z. As we know, Monte Carlo does this by sampling i.i.d. copies $Z_1, \ldots, Z_R$ of Z and then to estimate z via the sample-mean estimator:

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} Z_i \tag{14.1.1574}$$

Under further mild conditions such as $\text{V}(Z) < +\infty$, a central limit theorem will apply such that for large $n \rightarrow +\infty$, the distribution of $\bar{z}$ converges to a Normal distribution with mean z and standard deviation $\sigma/\sqrt{n}$. Because the standard deviation only converges towards 0 at the rate $\sqrt{n}$. implying one needs to increase the number of simulations (n) by a factor of 10 to half the standard deviation of $\bar{z}$, variance reduction methods are often useful for obtaining more precise estimates for z without needing very large numbers of simulations.

There are commonly seven variance reduction methods:

1. Antithetics
2. Stratification
3. Quasi-random method

4. Conditioning (conditional sampling)
5. Control variates
6. Moment matching and re-weighting
7. Monte Carlo Markov Chains

Remark

For practical examples with explicit code (scripts) the reader can refer to our MATLAB® or `R` free companion book (actually the MATLAB® one has the most interesting examples).

As already mentioned, will focus here on four of them:

1. **Antithetic variables**: Instead of a single sample sequence $X_1,\ldots,X_n$ sampled from X, this method instead uses an additional variable Y for which $\text{E}(Y)=\text{E}(X)=\lambda$, but is negatively correlated with X. The ith sample now becomes $\frac{X_i+Y_i}{2}$.
2. **Control variables**: A control variable is a random variable Y for which $\text{E}[Y]=\mu$ is known, and for which sampling variable X is replaced with $X+c(Y-\mu)$, where c is chosen so as to minimize $\text{V}(X+c(Y-\mu))$.
3. **Conditional sampling**: Conditional sampling selects a suitable variable Y for which $\text{E}(X|Y)$ has a smaller variance than X. Y is chosen so that, when it is observed, the uncertainty (i.e. entropy) of X gets reduced.
4. **Stratified (Regional) sampling**: Stratified sampling is a divide-and-conquer strategy that partitions the domain of X into subdomains, and then a sampling plan is devised for each subdomain. Moreover, common sense suggests that the smaller the subdomain, the less variance that should exist in that region, and so a variance reduction may be achievable in each of the subdomains that comprise the original domain.

And let us recall that if X and Y are two random variables, then we have proved in the section Statistics (see page 597) that:

$$\text{V}(X+Y)=\text{V}(X)+\text{V}(Y)+2\text{cov}(X,Y) \tag{14.1.1575}$$

Ok let's start!

14.1.18.1.6.1 Method of Antithetic Variables

Suppose X and Y are random variables for which $\mathrm{E}(Y)=\mathrm{E}(X)=\lambda$. Then:

$$\mathrm{E}\left(\frac{X+Y}{2}\right)=\frac{1}{2}\mathrm{E}(X)+\frac{1}{2}\mathrm{E}(Y)=\lambda/2+\lambda/2=\lambda \tag{14.1.1576}$$

Moreover:

$$\mathrm{V}\left(\frac{X+Y}{2}\right)=\frac{1}{4}\left[\mathrm{V}(X)+\mathrm{V}(Y)+2\mathrm{cov}(X,Y)\right] \tag{14.1.1577}$$

Remark

In the special case of $X=Y$ we fall back obviously on:

$$\mathrm{E}\left(\frac{X+Y}{2}\right)=\mathrm{E}\left(\frac{X+X}{2}\right)=\mathrm{E}\left(X\right)=\lambda$$

and:

$$\mathrm{V}\left(\frac{X+X}{2}\right)=\mathrm{V}\left(X\right)=\frac{1}{4}\left[\mathrm{V}(X)+\mathrm{V}(X)+2\mathrm{V}(X)\right]$$

Notice that, if X and Y are dependent and have negative correlation, then $\mathrm{cov}(X,Y)<0$, which causes a reduction in variance (compared to the case where X and Y are uncorrelated)[36].

A typical example given in many textbooks is to compute the mean of $\mathcal{U}_{[0,1]}$ by using an antithetic variable $1-\mathcal{U}_{[0,1]}$ (it's indeed an important example as $\mathcal{U}_{[0,1]}$ is used for the inverse transform method!).

✍**Example:**

The simple example is if $X=\mathcal{U}$, then $Y=1-\mathcal{U}$. We have then:

$$\frac{X+Y}{2}=\frac{1}{2}$$

and:

$$\frac{X+Y}{2}=\frac{\mathcal{U}+(1-\mathcal{U})}{2}=0$$

ie perfect negative correlation!

Notice that this method can obviously be used to compute integrals by implementing it in the Monte Carlo Mean Value Method presented earlier above!

---

[36] Remember that by definition, the correlation between X and Y is defined as $\mathrm{cor}(X,Y)=\frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y}$

14.1.18.1.6.2 Method of Control Variables

Suppose we are interested in estimating $\text{E}(X)=\lambda$. Then instead of exclusively sampling X, we may also introduce another variable Y, named the "control variable", that is correlated with X, and sample $Z=X+c(Y-\mu)$, where c is a constant, and $\mu=\text{E}(Y)$. Notice that :

$$\text{E}(Z)=\text{E}(X)+c\text{E}(Y-\mu)=\lambda+c(\mu-\mu)=\lambda+c\cdot 0=\lambda \tag{14.1.1578}$$

and so sampling Z will lead to an unbiased estimator. Moreover:

$$\text{V}(Z)=\text{V}(X+c(Y-\mu))=\text{V}(X)+c^2\text{V}(Y)+2c\cdot\text{cov}(X,Y) \tag{14.1.1579}$$

which is minimized when:

$$\frac{\partial}{\partial c}\left(\text{V}(X)+c^2\text{V}(Y)+2c\cdot\text{cov}(X,Y)\right)=0 \tag{14.1.1580}$$

Therefore:

$$c=-\frac{\text{cov}(X,Y)}{\text{V}(Y)} \tag{14.1.1581}$$

(optimal c sometimes denoted c^*) and yields a variance equal to:

$$\text{V}(Z^*)=\text{V}(Z+c^*Y)=\text{V}(X)-\frac{\text{cov}^2(X,Y)}{\text{V}(Y)} \tag{14.1.1582}$$

Hence:

$$\text{V}(Z^*)<\text{V}(X)\Leftrightarrow\text{cov}^2(X,Y)<0 \tag{14.1.1583}$$

Therefore, the method of control variables will always reduce variance, so long as X and Y have non-zero correlation.

One issue faced by this method is that usually neither $\text{cov}(X,Y)$ nor $\text{V}(X)$ are known beforehand. However, as a preprocessing step, we may obtain the estimators:

$$\widehat{\text{cov}}(X,Y)=\frac{\sum_{i=1}^n(X_i-\hat{\lambda})(Y_i-\mu)}{n-1} \tag{14.1.1584}$$

and:

$$\hat{\sigma}_Y^2=\frac{\sum_{i=1}^n(Y_i-\mu)^2}{n-1} \tag{14.1.1585}$$

to obtain the estimate:

$$\hat{c}=\frac{\sum_{i=1}^n(X_i-\hat{\lambda})(Y_i-\mu)}{\sum_{i=1}^n(Y_i-\mu)^2} \tag{14.1.1586}$$

for c.

And finally:

$$\hat{\mu}_{N,c^*}=\frac{1}{N}\sum_{i=1}^N[X_i+c^*(Y_i-\text{E}(Y)]\cong\mu_X \tag{14.1.1587}$$

The real challenge is to choose the control variable Y that is easy to simulate and for which μ_Y is easy to find...

Notice that this method can also obviously be used to compute integrals by implementing it in the Monte Carlo Mean Value Method presented earlier above!

14.1.18.1.6.3 Method of Conditioning

There are times when knowledge of the value of a random variable Y can reduce the variance of what can be observed for X.

In this case $\text{E}(X|Y)$ may have less variance than X. Note also (see section Probabilities page 501) the iterated conditional mean $\text{E}(\text{E}(X|Y))=\text{E}(X)$, and so $\text{E}(X|Y)$ is an unbiased estimator of $\text{E}(X)$.

So in other words, the idea is that instead of computing $\mu_Y=\text{E}(X)$, for a given random variable Z, we put $V=\text{E}(X|Y)$ and we calculate instead $\mu_X=\text{E}(V)=\text{E}(\text{E}(X|Y))=\text{E}(X)$.

Moreover, the amount of reduction in variance is determined by the following proposition.

Let us prove first that:

$$\text{V}(X)=\text{E}(\text{V}(X|Y))+\text{V}(\text{E}(X|Y)) \tag{14.1.1588}$$

To do that we just write:

$$\begin{aligned}\text{E}[\text{V}(X|Y)]+\text{V}(\text{E}(X|Y))&=\text{E}[\text{E}(X^2|Y)-\text{E}^2(X|Y))+\text{E}(\text{E}^2(X|Y))-(\text{E}[\text{E}(X|Y)))^2]\\&=\text{E}(\text{E}(X^2|Y))-\text{E}(\text{E}^2(X|Y))+\text{E}(\text{E}^2(X|Y))+\text{E}^2(X)\\&=\text{E}(X^2)-\text{E}^2(X)=\text{V}(X)\end{aligned} \tag{14.1.1589}$$

We have for immediate corollary:

Corollary 14.6.1.

$$\text{V}(X)-\text{V}(\text{E}(X|Y))=\text{E}(\text{V}(X|Y))\Leftrightarrow\text{V}(X)=\text{V}(\text{E}(X|Y))+\text{E}(\text{V}(X|Y)) \tag{14.1.1590}$$

So it proves that:

$$\text{V}(\text{E}(X|Y))<\text{V}(X)\Leftrightarrow\frac{\text{V}(\text{E}(X|Y))}{\text{V}(X)}<1 \tag{14.1.1591}$$

Very likely for many people that is very abstract and counter intuitive. So let us see an example!

✍**Example:**

Let us recall the method of approximating π using Monte Carlo that was presented earlier above with an indicator variable I such that:

$$I=\begin{cases}1 & \text{if } X^2+Y^2\leq 1\\0 & \text{otherwise}\end{cases}$$

where $X,Y=\mathcal{U}_{[0,1]}$. Recall that (remember that's for the quarter of a circle!):

$$\text{E}(I)=\frac{\pi}{4}\quad\text{and}\quad\text{V}(I)=\frac{\pi}{4}\left(1-\frac{\pi}{4}\right)\cong 0.1685$$

We now show that $\text{E}(I|X)$ has a smaller variance. For that purpose we start by calculating:

$$\begin{aligned}\mathrm{E}(I|X=x)&=P(X^2+Y^2\leq 1|X=x)=P(Y^2\leq 1-x^2|X=x)\\&=P(Y^2\leq 1-x^2)=P(-(1-x^2)^{1/2}\leq Y\leq (1-x^2)^{1/2})\\&=\int\limits_{-(1-x^2)^{1/2}}^{(1-x^2)^{1/2}}\frac{1}{2}\mathrm{d}y=(1-x^2)^{1/2}\\&\Rightarrow \mathrm{E}(I|X)=(1-X^2)^{1/2}\end{aligned}$$

However, since X^2 has the same distribution as U^2, $U=\mathcal{U}_{[0,1]}$, it follows that we get the following:

$$\mathrm{E}(I|X)=(1-U^2)^{1/2}$$

Notice that the expected means remains a random variable (without surprise as it is a conditions expected mean) that we will denote $Z:=\mathrm{E}(I|X)=(1-U^2)^{1/2}$! Therefore using Huygens relation (see section Statistics page 593), $\mathrm{V}(Z)=\mathrm{E}(Z^2)-\mathrm{E}(Z)^2$ we get:

$$\begin{aligned}\mathrm{V}(\mathrm{E}(I|X))=\mathrm{V}\left((1-U^2)^{1/2}\right)&=\mathrm{E}\left(\left((1-U^2)^{1/2}\right)^2\right)-\mathrm{E}\left((1-U^2)^{1/2}\right)^2\\&=\mathrm{E}\left(1-U^2\right)-\mathrm{E}\left((1-U^2)^{1/2}\right)^2=1-\mathrm{E}\left(U^2\right)-\mathrm{E}\left((1-U^2)^{1/2}\right)^2\end{aligned}$$

Using the result derived in the section Statistics page 739, i.e. $\mathrm{E}(\mathcal{U}_{0,1}^2)=1/3$, we then have:

$$\mathrm{V}(\mathrm{E}(I|X))=\mathrm{V}\left((1-U^2)^{1/2}\right)=1-\frac{1}{3}-\mathrm{E}\left((1-U^2)^{1/2}\right)^2=\frac{2}{3}-\mathrm{E}\left((1-U^2)^{1/2}\right)^2$$

We just have left to compute $\mathrm{E}(\sqrt{1-U^2})=\int_0^1\sqrt{1-x^2}\mathrm{d}x$ and using the primitive derived in the section of Differential and Integral Calculus at page 1534, we get $\mathrm{E}(\sqrt{1-U^2})=\pi/4$. Therefore:

$$\mathrm{V}(\mathrm{E}(I|X))=\frac{2}{3}-\left(\frac{\pi}{2}\right)^2$$

Hence:

$$\frac{\mathrm{V}(\mathrm{E}(X|Y))}{\mathrm{V}(X)}\cong\frac{0.0498}{0.1685}\cong 0.295$$

yielding a 70.5% reduction in variance.

14.1.18.1.6.4 Method of Stratified Sampling

The reader remembers maybe that we have seen earlier above the method of Simple Random Sampling (SRS) in which realizations of a vector $\vec{x}$ (samples) are generated as independent and identically distributed realizations on a domain $\mathcal{S}$ with marginal cumulative distribution functions (CDFs) $F(x)$ by:

$$x_i = F_{X_i}^{-1}\left(\mathcal{U}_i\right); i = 1, 2, \ldots, n \qquad (14.1.1592)$$

where $\mathcal{U}_i$ are independent and identically distributed uniformly distributed samples on $[0,1]$. The realizations $\vec{x}$ are then applied to the system $\vec{y} = \vec{F}(\vec{x})$ and $\vec{y}$ is statistically evaluated.

"Stratified sampling" represents another conditioning approach to reduce variance. It aims to reduce the variance of the estimator by dividing the interval into strata and estimating the integral on each of the stratum with smaller variance.

Again, suppose the goal is to estimate $\lambda = \text{E}(X)$. Suppose Y is a finite random variable with $\text{dom}(Y) = \{y_1, \ldots, y_k\}$ and $p_i = P(y_i)$ is known for each $i = 1, \ldots, k$.

Then:

$$\text{E}(X) = \sum_{i=1}^{k} \text{E}(X|Y = y_i)p(y_i) \qquad (14.1.1593)$$

Then λ can be estimated by estimating each of $\text{E}(X|Y = y_i)$ with $\hat{\lambda}_i$ and producing the estimate:

$$\hat{\lambda} = \sum_{i=1}^{n} \hat{\lambda}_i p_i \qquad (14.1.1594)$$

where $\hat{\lambda}_i$ is obtained using n_i samples, and $n_1 + \cdots + n_k = n$ and $n_i = np_i$.

Notice that:

$$\text{V}(\hat{\lambda}) = \sum_{i=1}^{n} p_i^2 \text{V}(\hat{\lambda}_i) = \sum_{i=1}^{n} p_i^2 \frac{\text{V}(X|Y = y_i)}{np_i} = \frac{1}{n}\text{E}(\text{V}(X|Y)) \qquad (14.1.1595)$$

Thus:

$$\frac{1}{n}\text{V}(X) - \frac{1}{n}\text{E}(\text{V}(X|Y)) = \frac{1}{n}\text{V}(\text{V}(X|Y)) \qquad (14.1.1596)$$

Hence:

$$\frac{1}{n}\text{V}(\text{V}(X|Y)) < \frac{1}{n}\text{V}(X) \qquad (14.1.1597)$$

so that the reduction in variance depends on the amount of variance that occurs in $\text{E}(X|Y)$. After simplification and rearrangement we get obviously:

$$\frac{\text{V}(\text{V}(X|Y))}{\text{V}(X)} < 1 \qquad (14.1.1598)$$

The introduction above may be to abstract. So let us introduce it in another way! Consider we want to estimate an integral that consist a distribution p and a function of interest f

$$I = \int_{\chi} p(x) f(x) \mathrm{d}x \tag{14.1.1599}$$

via partitioning χ into smaller groups, named "strata". It is a type of quasi-Monte Carlo (QMC) method as it introduces deterministic procedures into the MC estimation. Same as crude Monte Carlo the estimation is still unbiased; however, the variance of the estimator can be smaller than crude Monte Carlo.

The (crude) Monte Carlo estimator of I is:

$$\hat{I} = \frac{1}{N} \sum_{i=1}^{N} f(x^i) \tag{14.1.1600}$$

where $x^i \sim p(x)$ and N is the number of Monte Carlo samples.

This estimator is an unbiased estimator of the original target I:

$$\mathrm{E}(\hat{I}) = \mathrm{E}\left(\frac{1}{N} \sum_{i=1}^{N} f(x^i)\right) = \frac{1}{N} \sum_{i=1}^{N} \mathrm{E}(f(x^i)) = \frac{1}{N} \sum_{i=1}^{N} I = I \tag{14.1.1601}$$

Now let's look at the variance of the estimator $\mathrm{V}(\hat{I})$. We first need to define:

$$\sigma^2 := \int_{\chi} p(x)(f(x) - I)^2 \mathrm{d}x \tag{14.1.1602}$$

which is the variance of $f(x)$ under the distribution $p(x)$. Now the variance of interest is

$$\mathrm{V}(\hat{I}) = \mathrm{V}\left(\frac{1}{N} \sum_{i=1}^{N} f(x^i)\right) = \frac{1}{N^2} \sum_{i=1}^{N} \mathrm{V}(f(x^i)) = \frac{1}{N^2} \sum_{i=1}^{N} \sigma^2 = \frac{\sigma^2}{N} \tag{14.1.1603}$$

As it can be seen, the variance drops with the order of $\mathcal{O}(\frac{1}{N})$ by using more samples.

Now, we look at how stratified sampling works and how the variance of the corresponding estimator changes depending on different stratifications. For the sake of simplicity, we consider the case where $p(x) = \mathcal{U}^D_{[0,1]}$ a D-dimensional uniform distribution for the integral. We first partition χ into H strata s.t. $\chi = \chi_1 \cup \ldots \cup \chi_H$. Then, as the simplest way of performing stratified sampling, we uniformly draw independent Monte Carlo samples for each stratum χ_h, and the overall estimator is given as below:

$$\hat{I}_{\text{strat}} = \sum_{h=1}^{H} \frac{|\chi_h|}{n_h} \sum_{i=1}^{n_h} f(x^i_h) \tag{14.1.1604}$$

where $\{x^i_h\}_{i=1}^{n_h}$ are samples that lie in χ_h and n_h is the number of MC samples in stratum h.

The mean of f within each stratum h is therefore:

$$\mu_h = \int_{\chi_h} \mathcal{U}(x; \chi_h) f(x) \mathrm{d}x = |\chi_h|^{-1} \int_{\chi_h} f(x) \mathrm{d}x \tag{14.1.1605}$$

and the corresponding variance is:

$$\sigma_h^2 = \int_{\chi_h} \mathcal{U}(x;\chi_h)(f(x)-\mu_h)^2\mathrm{d}x = |\chi_h|^{-1}\int_{\chi_h}(f(x)-\mu_h)^2\mathrm{d}x \tag{14.1.1606}$$

As it is mentioned in the beginning, $\hat{I}_{\text{strat}}$ is still an unbiased estimator of I. We can validate this by computing the mean of $\hat{I}_{\text{strat}}$:

$$\begin{aligned}
\mathrm{E}\left(\hat{I}_{\text{strat}}\right) &= \mathrm{E}\left(\sum_{h=1}^{H}\frac{|\chi_h|}{n_h}\sum_{i=1}^{n_h} f\left(x_h^i\right)\right) = \sum_{h=1}^{H}\frac{|\chi_h|}{n_h}\sum_{i=1}^{n_h}\mathrm{E}\left(f\left(x_h^i\right)\right)\\
&= \sum_{h=1}^{H}\frac{|\chi_h|}{n_h}\sum_{i=1}^{n_h}|\chi_h|^{-1}\int_{\chi_h} f(x)\mathrm{d}x = \sum_{h=1}^{H}\frac{1}{n_h}\sum_{i=1}^{n_h}\int_{\chi_h} f(x)\mathrm{d}x = \sum_{h=1}^{H}\int_{\chi_h} f(x)\mathrm{d}x\\
&= \int_{\chi_1\cup\ldots\cup\chi_H} f(x)\mathrm{d}x = \int_{\mathcal{X}} f(x)\mathrm{d}x = I
\end{aligned} \tag{14.1.1607}$$

The variance of $\hat{I}_{\text{strat}}$ is then:

$$\begin{aligned}
\mathrm{V}\left(\hat{I}_{\text{strat}}\right) &= \mathrm{V}\left(\sum_{h=1}^{H}\frac{|\chi_h|}{n_h}\sum_{i=1}^{n_h} f\left(x_h^i\right)\right) = \sum_{h=1}^{H}\mathrm{V}\left(\frac{|\chi_h|}{n_h}\sum_{i=1}^{n_h} f\left(x_h^i\right)\right) = \sum_{h=1}^{H}\frac{|\chi_h|^2}{n_h^2}\mathrm{V}\left(\sum_{i=1}^{n_h} f\left(x_h^i\right)\right)\\
&= \sum_{h=1}^{H}\frac{|\chi_h|^2}{n_h^2}\sum_{i=1}^{n_h}\mathrm{V}\left(f\left(x_h^i\right)\right) = \sum_{h=1}^{H}\frac{|\chi_h|^2}{n_h^2}\sum_{i=1}^{n_h}\sigma_h^2 = \sum_{h=1}^{H}\frac{|\chi_h|^2}{n_h}\sigma_h^2
\end{aligned} \tag{14.1.1608}$$

We are interested in comparing this estimator against (crude) Monte Carlo. In particular, we'd like to show that when n_h is allocated proportionally to $|\,\chi_h\,|$ the estimator can only decrease in variance as compared to (crude) Monte Carlo.

In order to show this, we need to rewrite the variance of crude Monte Carlo in terms of strata. Let's introduce $h(x)$ - the stratum that contains $x : x\in\mathcal{X}_{h(x)}, \forall x\in\mathcal{X}$. By using in a clever way the law of total variance (see section Statistics page 502) we can relate σ and σ_h as:

$$\sigma^2 = \mathrm{V}(f(x)) = \mathrm{E}\left(\mathrm{V}(f(x)\,|\,h(x))\right) + \mathrm{V}\left(\mathrm{E}(f(x)\,|\,h(x))\right) = \sum_{h=1}^{H}|\chi_h|\,\sigma_h^2 + \sum_{h=1}^{H}|\chi_h|\,(\mu_h - I)^2 \tag{14.1.1609}$$

Therefore:

$$\mathrm{V}(\hat{I}) = \frac{\sigma^2}{n} = \frac{1}{n}\sum_{h=1}^{H}|\mathcal{X}_h|\,\sigma_h^2 + \frac{1}{n}\sum_{h=1}^{H}|\mathcal{X}_h|\,(\mu_h - I)^2 \tag{14.1.1610}$$

Note that in the case of proportional allocation, because $n_h = n\,|\mathcal{X}_h|$, from the above deviation we get:

$$\mathrm{V}(\hat{I}_{\text{strat}}) = \sum_{h=1}^{H}\frac{|\chi_h|^2}{n_h}\sigma_h^2 = \frac{1}{n}\sum_{h=1}^{H}|\mathcal{X}_h|\,\sigma_h^2 \tag{14.1.1611}$$

So comparing:

$$\mathrm{V}(\hat{I})=\frac{1}{n}\sum_{h=1}^{H}\left|\mathcal{X}_h\right|\sigma_h^2+\frac{1}{n}\sum_{h=1}^{H}\left|\mathcal{X}_h\right|(\mu_h-I)^2 \quad \text{and} \quad \mathrm{V}(\hat{I}_{\text{strat}})=\frac{1}{n}\sum_{h=1}^{H}\left|\mathcal{X}_h\right|\sigma_h^2 \qquad (14.1.1612)$$

So we see immediately that $\mathrm{V}(\hat{I}_{\text{strat}})\leq\mathrm{V}(\hat{I})$, that is:

$$\frac{1}{n}\sum_{h=1}^{H}\left|\mathcal{X}_h\right|\sigma_h^2\leq\frac{1}{n}\sum_{h=1}^{H}\left|\mathcal{X}_h\right|\sigma_h^2+\frac{1}{n}\sum_{h=1}^{H}\left|\mathcal{X}_h\right|(\mu_h-I)^2 \qquad (14.1.1613)$$

Rearranged and simplified:

$$\frac{\sum_{h=1}^{H}\left|\mathcal{X}_h\right|\sigma_h^2}{\sum_{h=1}^{H}\left|\mathcal{X}_h\right|\sigma_h^2+\sum_{h=1}^{H}\left|\mathcal{X}_h\right|(\mu_h-I)^2}\leq 1 \qquad (14.1.1614)$$

which is smaller than $\mathrm{V}(\hat{I})$. Although it is not the optimal allocation, proportional allocation as applied to stratified sampling will never increase the variance over that of crude Monte Carlo. Also, note that a poor allocation can also increase the variance of the estimator in comparison to crude Monte Carlo!

So to summarize what we have seen here, in a more formal way, "stratified sampling" begins by dividing the sample domain $\mathcal{S}$ into a collection of M disjoint subsets (strata) $\mathcal{S}_k; k=1,2,\ldots,M$ such as $\cup_{k=1}^{M}\mathcal{S}_k=\mathcal{S}$ and $\mathcal{S}_p\cap\mathcal{S}_q=\varnothing; p\neq q$. Sample realizations from a given stratum k, $\vec{x}_k=\{x_{1k},x_{2k},\ldots,x_{Nk}\}$, are generated by randomly sampling the vector components according to:

$$x_{ik}=F_{X_i}^{-1}(\mathcal{U}_{ik})\,; i=1,2,\ldots,N \qquad (14.1.1615)$$

where $\mathcal{U}_{ik}$ are identically and independently uniformly distributed samples on $\left[\xi_{ik}^l,\xi_{ik}^u\right]$ with $\xi_{ik}^l=F_{X_i}\left(\zeta_{ik}^l\right)$ and $\xi_{ik}^u=F_{X_i}\left(\zeta_{ik}^u\right)$, and ζ_{ik}^l and ζ_{ik}^u denote the lower and upper bounds respectively of the i^{th} vector component of stratum $\mathcal{S}k$. Typically the stratification is performed directly in the probability space meaning that the strata are defined directly by the bounds ξ_{ik}^l and ξ_{ik}^u.

More explicitly, given a vector of uniform random numbers $U_1,U_2,\cdots,U_n$ on $[0,1]$. We divide the interval $[0,1]$ into $k\leq n$ equal subintervals. We can use the given uniform random numbers to generate uniform random in each of the k intervals by using the relation:

$$\hat{U}_i=\begin{cases}\frac{i-1+U_i}{k} & \text{if } i\leq k\\ \frac{(i-k)-1+U_i}{k} & \text{if } i>k\end{cases} \qquad (14.1.1616)$$

Thus, $\hat{U}_1$ is a random number in the interval $\left(0,\frac{1}{k}\right)$, $\hat{U}_2$ belongs to the interval $\left(\frac{1}{k},\frac{2}{k}\right)$ and so on. Moreover, the numbers $\hat{U}_1,\hat{U}_{1+k},\hat{U}_{1+2k},\cdots$ are uniformly distributed in the interval $\left(0,\frac{1}{k}\right)$, the numbers $\hat{U}_2,\hat{U}_{2+k},\hat{U}_{2+2k},\cdots$ are uniformly distributed in $\left(\frac{1}{k},\frac{2}{k}\right)$ and so on.

✍**Example:**

Consider the following 8 uniform random numbers in $[0,1]$:

i	1	2	3	4	5	6	7	8
U_i	0.4880	0.7894	0.8628	0.4482	0.3172	0.8944	0.5013	0.3015

To generate uniform random numbers in each quartile, i.e. $k=4$ we have:

$$\begin{aligned}
\hat{U}_1 &= \frac{1-1+U_1}{4} = 0.122\\
\hat{U}_2 &- \frac{2-1+U_2}{4} - 0.44735\\
\hat{U}_3 &= \frac{3-1+U_3}{4} = 0.7157\\
\hat{U}_4 &- \frac{4-1+U_4}{4} - 0.86205\\
\hat{U}_5 &= \frac{(5-4)-1+U_5}{4} = 0.0793\\
\hat{U}_6 &= \frac{(6-4)-1+U_6}{4} = 0.4736\\
\hat{U}_7 &= \frac{(7-4)-1+U_7}{4} = 0.625325\\
\hat{U}_8 &= \frac{(8-4)-1+U_8}{4} = 0.825375
\end{aligned}$$

14.1.18.1.6.5 Importance Sampling (IS)

In statistics, "importance sampling" is a general technique for estimating properties of a particular distribution, while only having samples generated from a different distribution than the distribution of interest!

The method is named "importance sampling" because it relies on so-named "importance functions", which are instrumental distributions, in lieu of the original distributions (the term "sampling" is arguably a misnomer since the method does not attempt to draw samples from any given distribution). In fact, an evaluation of:

$$\mathrm{E}(f(X)) = \int f(x)\phi(x)\mathrm{d}x \qquad (14.1.1617)$$

based on simulations from f is almost never optimal in the sense that using alternative distributions can improve the variance of the resulting estimator of $\mathrm{E}(f(X))$.

Generally it is a very common technique used to estimate the tail of some probability distribution and is in that special case denoted:

$$\theta = P(X \leq \text{ treshold }) = \mathrm{E}\left(1_{X \leq \text{treshold}}\right) \qquad (14.1.1618)$$

when the occurrence of $X \leq$ threshold is very small. Such situations occurs typically in the calculation of climatic, rail and air disasters, bankruptcies of large companies/states, major stock market cracks, portfolio risks, operational risks (legal obligation for banks or insurance companies).

To explain the idea consider we want to compute for a $\mathcal{N}(0,1)$ the following:

$$\theta = P(X \leq -10) = \mathrm{E}\left(1_{X \leq -10}\right) \tag{14.1.1619}$$

If we try to compute that with crude Monte Carlo method with N varying from $1,000$ too $10,000,000$ the conclusion is almost always the same: $\theta = 0$ and it's wrong as the correct value is known to be $\theta \cong 7.6199 \cdot 10^{-24}$. The necessary number of trials (N) should be of the order of 10^{25}... value that is hardly accessible to most common personal computer even in the early 121st century (holocene calendar).

The underling workaround idea of importance sampling is to use the following trick (i.e. change the law!):

$$\theta = \mathrm{E}(f(X)) = \int f(x)\phi(x)\mathrm{d}x = \int f(x)\frac{\phi(x)}{\psi(x)}\psi(x)\mathrm{d}x = \mathrm{E}(h(Y)) \tag{14.1.1620}$$

with:

$$h(x) = f(x)\frac{\phi(x)}{\psi(x)} \quad \text{and} \quad Y \sim \psi \tag{14.1.1621}$$

where ϕ is the density of X and ψ another density - an "importance function" - such as $\psi(x) \neq 0$ if $\phi(x) \neq 0$.

Hence:

$$\hat{\theta} = \frac{1}{N}\sum_{i=1}^{N} f(X_i)\frac{\phi(X_i)}{\psi(X_i)} \tag{14.1.1622}$$

To study the variance, let X be a random variable with probability density function p. Consider evaluating the following quantity:

$$I = \mathrm{E}(f(x)) = \int f(x)p(x)\mathrm{d}x \tag{14.1.1623}$$

According to what we have just seen, then the importance sampling estimator is:

$$\hat{I}_N = \frac{1}{N}\sum_{i=1}^{N} f\left(x_i\right) \cdot \frac{p\left(x_i\right)}{q\left(x_i\right)} \tag{14.1.1624}$$

When $p = q$, this reduces obviously to the simple estimator that uses sample means of $f(x)$ to estimate its expectation. Does this estimator a good estimator? Let's study its bias before we focus on the variance! For the bias (see page 6798):

$$\begin{aligned}\mathrm{Bias}(\hat{I}) &= \mathrm{E}\left(\hat{I}_N - I\right) = \mathrm{E}\left(\hat{I}_N\right) - I = \mathrm{E}\left(f(x) \cdot \frac{p(x)}{q(x)}\right) - I \\ &= \int f(x)\frac{p(x)}{q(x)}q(x)\mathrm{d}x - I = \int f(x)p(x)\mathrm{d}x - I = 0\end{aligned} \tag{14.1.1625}$$

Thus, it is an unbiased estimator! That lead us also to the result that (we will us it just below):

$$I=\mathrm{E}\left(f\left(x\right)\cdot\frac{p\left(x\right)}{q\left(x\right)}\right)\tag{14.1.1626}$$

And now let's see the variance, using the equality just above and Huygens relation (see section Statistics page 593), we get:

$$\begin{aligned}\mathrm{V}\left(\hat{I}\right)&=\mathrm{V}\left(f\left(x\right)\cdot\frac{p\left(x\right)}{q\left(x\right)}\right)=\mathrm{E}(\hat{I}^2)-\mathrm{E}(\hat{I})^2\\&=\mathrm{E}\left(f^2\left(x\right)\cdot\frac{p^2\left(x\right)}{q^2\left(x\right)}\right)-\underbrace{\mathrm{E}\left(f\left(x\right)\cdot\frac{p\left(x\right)}{q\left(x\right)}\right)^2}_{I^2}\\&=\int\frac{f^2(x)p^2(x)}{q^2(x)}q(x)\mathrm{d}x-I^2=\int\frac{f^2(x)p(x)}{q(x)}p(x)\mathrm{d}x-I^2\\&=\mathrm{E}\left(\frac{f^2(x)p(x)}{q^2(x)}\right)-\mathrm{E}(f(x))^2\end{aligned}\tag{14.1.1627}$$

The last equality has to be compared with:

$$\mathrm{V}\left(I\right)=\mathrm{V}\left(f(x)\right)=\mathrm{E}\left(f(x)^2\right)-\mathrm{E}(f(x))^2\tag{14.1.1628}$$

That is putting side by side:

$$\begin{aligned}\mathrm{V}\left(\hat{I}\right)&=\mathrm{E}\left(\frac{f^2(x)p(x)}{q^2(x)}\right)-\mathrm{E}(f(x))^2\\\mathrm{V}\left(I\right)&=\mathrm{E}\left(f(x)^2\right)-\mathrm{E}(f(x))^2\end{aligned}\tag{14.1.1629}$$

So we see that we have a variance reduction, if and only if:

$$\mathrm{E}\left(\frac{f^2(x)p(x)}{q^2(x)}\right)\leq\mathrm{E}\left(f(x)^2\right)\tag{14.1.1630}$$

how to build a density $q(x)$ having the above property? It seems that this question doesn't have an easy answer other than trial and error and return on experience (REX).

As it may be a quite confusing notation and concept for some readers, let us see a simple concrete example!

✍**Example:**

Consider that we have $X=\mathcal{N}(\mu,\sigma)=\mathcal{N}(0,1)$ and we want to compute $\theta=P(X\leq-10)$. We know already that this is an issue!

The idea is then to force the trial of a new variable Y to be around -10. So the idea would be to choose: $Y=\mathcal{N}(-10,1)$.

Notice that if $Y=\mathcal{N}\left(\mu,\sigma^2\right)$, then $P(Y\leq\mu)=1/2$.

Therefore:

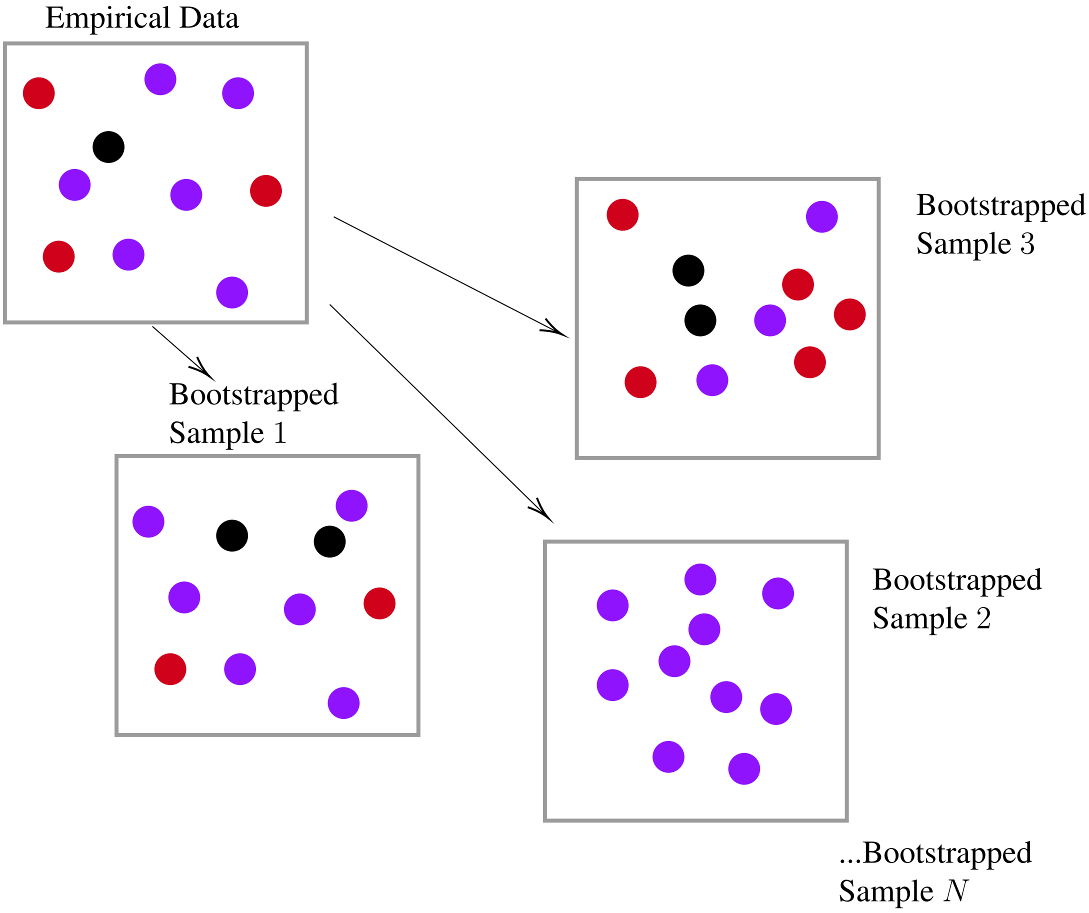
$$\begin{aligned}\theta=P(X\leq -10)&=\int\limits_{-\infty}^{-10}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)\mathrm{d}x\\&=\int\limits_{-\infty}^{-10}\left(\frac{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)}{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x+10)^2}{2}\right)}\right)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x+10)^2}{2}\right)\mathrm{d}x\\&=\int\limits_{-\infty}^{-10}\exp\left(10x+(-10)^2/2\right)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x+10)^2}{2}\right)\mathrm{d}x\\&=\int\limits_{-\infty}^{-10}\exp\left(10x+50\right)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x+10)^2}{2}\right)\mathrm{d}x\\&=\mathrm{E}\left(1_{Y\leq -10}\exp(10Y+50)\right)\end{aligned}$$

The trick works very well in this case as the reader can check it by itself in the corresponding implementation in `R` and MATLAB® as available in our corresponding companion free books.

In practice it is quite common to implement all the different methods of variance reduction seen so far and to compare which one converge the fastest after a given number of runs (or "epochs").

14.1.18.2 Bootstrapping

In statistics, "bootstrap techniques" can refer to any test or metric that relies on random sampling with replacement of a population of data to run statistical inference on small samples. This methods is relatively intensive in computations still for expensive office computers at this beginning of the 121st century (holocene calendar).



The goal of bootstrapping is to find some indication of a statistic: its estimate of course, but also its dispersion (variance, standard deviation), confidence intervals or hypothesis testing. This method is based on simulations, such as Monte Carlo methods, with the difference that the bootstrap does not require additional information than the one that is available already in the initial sample. In general, it is based on new samples obtained by sampling with replacement from the original sample (then we speak also of "resampling").

We distinguish generally two types of bootstrap:

1. The bootstraps which make no assumptions about the probability distribution of the data analysed. We then speak then as in statistics of "nonparametric bootstrap".
2. The bootstraps replacing each data measured by those corresponding to the analytical expression of the law of probability distribution assumed. We speak then of "parametric bootstrap". Once all the original values replaced, the process is exactly that of the nonparametric bootstrap.

Definition (#14.1.565): Given a set $\mathcal{D}$ of n observations, and a number B, we name "bootstrap" the procedure which consists in creating B samples $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_B$ from $\mathcal{D}$, each obtained by drawing n examples from $\mathcal{D}$ with replacement. Thus, each example may appear multiple times, or not at all, in $\mathcal{D}_b$.

We will illustrate the principle of the bootstrap on the example of the confidence interval of the mean μ of a random variable (this special case is named a "studentized bootstrap". For this example, the confidence interval for the mean of a random variable is completely determined from the mean and the variance calculated on the sample (see section Statistics page 818).

We consider a sample of the random variable composed of 10 estimates:

$$X=\{16,12,14,6,43,7,0,54,25,13\} \tag{14.1.1631}$$

The arithmetic average of this sample is:

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}x_i=\frac{1}{10}\sum_{i=1}^{10}x_i=19 \tag{14.1.1632}$$

and it standard deviation (maximum unbiased likelihood estimator of the standard deviation):

$$\sigma=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{X})^2}=\sqrt{\frac{1}{9}\sum_{i=1}^{10}(x_i-19)^2}=17.09 \tag{14.1.1633}$$

As we are in the situation of a known sample mean and an unknown sample variance, to do the calculation of a confidence interval, then we have proved in the section Statistics that we had to use:

$$\bar{X}-\frac{S}{\sqrt{n}}T_{\alpha/2}(n-1)\leq\mu\leq\bar{X}+\frac{S}{\sqrt{n}}T_{\alpha/2}(n-1) \tag{14.1.1634}$$

where S is for recall another traditional notation in some areas of statistics for the notation of the empirical standard deviation (see section Statistics page 824). We then have for the confidence interval at 95% of the mean:

$$19-\frac{17.09}{\sqrt{10}}T_{2.5\%}(9)\leq\mu\leq 19+\frac{17.09}{\sqrt{10}}T_{2.5\%}(9) \tag{14.1.1635}$$

Therefore:

$$19-\frac{17.09}{\sqrt{10}}2.26\leq\mu\leq 19+\frac{17.09}{\sqrt{10}}2.26 \tag{14.1.1636}$$

Which gives:

$$4.49\leq\mu\leq 33.51 \tag{14.1.1637}$$

The confidence interval can also be calculated by bootstrap (this is especially useful for complicate distributions that are not symmetric and when we focus on the median rather than on the mean). Then it is therefore obtained by the following algorithm (and abbreviated BCI for "bootstrapped confidence interval"):

1. From the initial sample, we simulate new samples of the same size, named "bootstrap replicates" of size n, by random draws with replacement (see figure above). For example with the previous series, we could get the following replicate:

$$X^*=\{54,0,16,7,43,54,0,25,25,6\} \tag{14.1.1638}$$

in which, by definition, some of the original sample values do not appear, and where others appear several times (yes it's a sampling with replacement therefore...). Several samples are simulated in this way. So we can form a number of replicas (arrangements with replacement) equal to (see section Probabilities page 523):

$$\bar{A}_n^n = n^n \tag{14.1.1639}$$

Therefore with 10 values we have 10, 000, 000, 000 possibilities...

2. For each simulated sample, an average[37] is calculated (so we will have several thousand of averages!).

3. The 95% confidence interval is then calculated on this set of averages (or any other estimator) by typically using the percentile calculation (through the functions of a spreadsheet software or a programming/scripting language). This grouping is named the "bagging" that stands for the abbreviation of "Bootstrap Aggregating".

Obviously for each set of bootstrap, the percentiles themselves will not be the same so it is even possible to create a confidence interval for the percentiles themselves and this is named "percentile bootstrap"! But one important assumption is that such a distribution is pivotal. This means that if the underlying parameter changes, the shape of the distribution is only shifted by a constant, and the scale does not necessarily change.

It is quite easy (just like the Monte Carlo methods) to create replicas with spreadsheet software like Microsoft® Excel® (at least for people that know a little bit how to use a spreadsheet software) without computer programming or scripting (see below an example with Microsoft® Excel®)! Furthermore, the bootstrap technique is very powerful because it does not use any assumptions about the underlying statistical distribution.

The most common field of application of bootstrapping in "direct" business (I don't mind about Data Mining for Marketing that is not what I mean about "direct business") that I know is in project management during meetings where a dozen people estimates the duration of a project task or project phase.

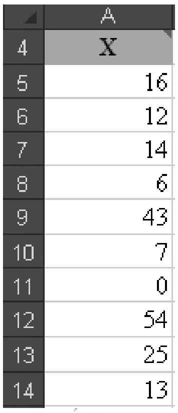
Bootstrapping can therefore be applied to any estimator other than the average, as the median, the correlation coefficient between two random variables or the principal eigenvalue of a variance-covariance matrix (for principal component analysis for example!), or the slope and intercept of a regression and this is its great strength!!! Indeed, for these estimators, there is no general mathematical relations that defines the standard error or confidence interval. The only methods applicable are resampling methods to which bootstrapping belongs to and this is intensively used since almost any home or office computer at the beginning of the 121st century (holocene calendar) is powerful enough to bootstrap small databases.

✍**Example:**

As example let us first use a spreadsheet software like Microsoft® Excel® table 14.0.6123 and taking the theoretical companion example above as practical software example (we prohibit ourselves of doing VBA programming). We then build a small table with the previous sample:

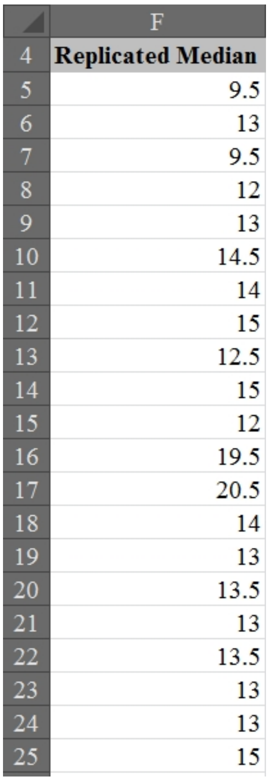
---

[37] in fact the process is the same for any estimate of any statistical indicator $\hat{\theta}$



	A
4	X
5	16
6	12
7	14
8	6
9	43
10	7
11	0
12	54
13	25
14	13

At the opposite of the companion theoretical example, we would like to be able to determine a confidence interval for the median instead of the arithmetic mean (we take on purpose a statistical indicator for which there is no simple analytical confidence interval). For this, we calculate the median of several thousand of replications in the column `F` (random choice!), where each replication corresponds to a row :



	F
4	**Replicated Median**
5	9.5
6	13
7	9.5
8	12
9	13
10	14.5
11	14
12	15
13	12.5
14	15
15	12
16	19.5
17	20.5
18	14
19	13
20	13.5
21	13
22	13.5
23	13
24	13
25	15

with the following quite long formula for Microsoft® Excel® Next 14.0.6123 to put in cell `F5` and then pull down to the end of the sheet:

```
=MEDIAN(INDEX(A5:A14,RANDBETWEEN(1,10),1),
INDEX(A5:A14,RANDBETWEEN(1,10),1),
INDEX(A5:A14,RANDBETWEEN(1,10),1),
INDEX(A5:A14,RANDBETWEEN(1,10),1),
INDEX(A5:A14,RANDBETWEEN(1,10),1),
INDEX(A5:A14,RANDBETWEEN(1,10),1),
INDEX(A5:A14,RANDBETWEEN(1,10),1),
INDEX(A5:A14,RANDBETWEEN(1,10),1),
INDEX(A5:A14,RANDBETWEEN(1,10),1),
INDEX(A5:A14,RANDBETWEEN(1,10),1))
```

So we can not have more that 10 billion more... corresponding to the $\bar{A}_n^n$ calculated above (Microsoft® Excel® 14.0.6123 and after is limited to $17,179,869,184$ cells...).

Then simply in a cell of our choice we write:

```
=PERCENTILE(F5:F2003,0.025)
```

and in another cell:

```
=PERCENTILE(F5:F2003,0.975)
```

which will give us $2,000$ replications respectively 7 and 29.5.

With basic knowledge of a spreadsheet software, it is possible to graphically plot the convergence of the median in function the number of replications (below we used only the first 100 replications):

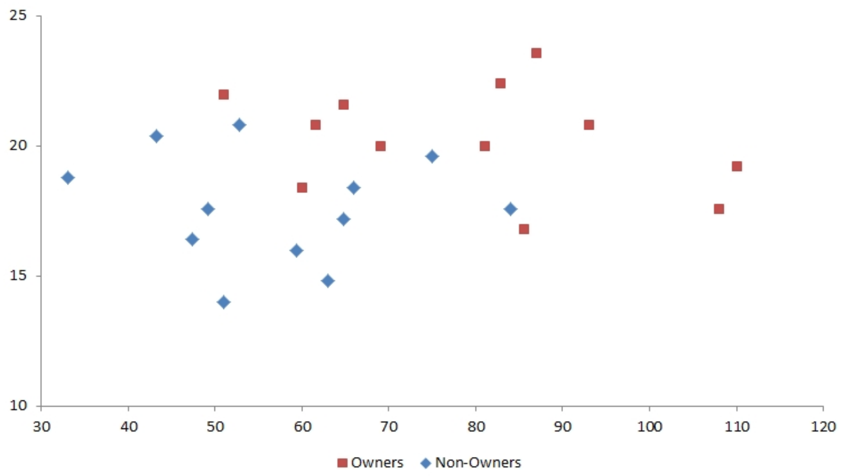


Figure 14.101: **Convergence of the median as function of the number of replications**

Obviously, this chart will look different every time we restart the simulation in Microsoft® Excel® 14.0.6123 by pressing the F9 key.

Finally let us indicate that after having studied earlier above many linear regression models this does avoid the fact that in many cases no theoretical model is adapted either to interpolate or verbatim to extrapolate some data. Therefore, if we have for each abscissa point (exogenous variable) a given quantity of values of the output (endogenous) variable, we can therefore use the bootstrapping method which will give us the bootstrapped regression coefficients and also bootstrapped interpolated or extrapolated values! This is an extremely interesting technique in practice of nonparametric regression! We can do such bootstrap regressions in softwares like SPSS (with additional module) or SAS but also in the free `R` software with the right packages (see the companion book on `R`).

Remark

Bootstrapping has enormous potential in statistics education and practice, but there are subtle issues and ways to go wrong. For example, the common combination of non-parametric bootstrapping and bootstrap percentile confidence intervals seems very likely to be less accurate than using T-intervals for small samples, though more accurate for larger samples. We therefore strongly recommend anyone to read the following paper to avoid some traps about the usage of bootstrapping: Hesterberg TC. *What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum.* The American Statistician. 12015 Oct 2;69(4):371-86 (holocene calendar).

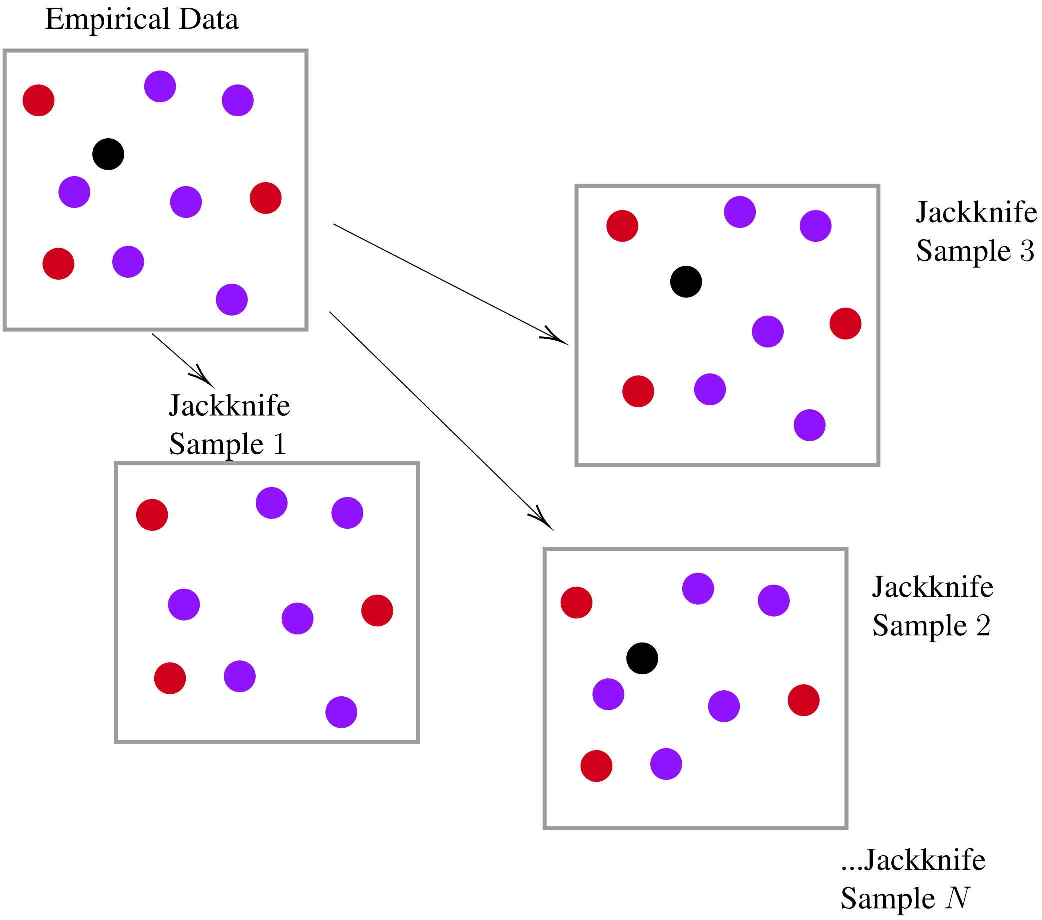
Caution!!!!!!!!! Use bootstrap mainly for confidence intervals or parameters estimation. Not for testing hypotheses (under the null), unless you really know what you are doing! For NHST you should prefer permutations techniques that we will introduce further below.

Indeed, we use bootstrap to generate data under the empirical distribution of the observed data. To get a p-value by bootstrap, we need to generate bootstraps under the null hypothesis!!!! So when conducting bootstrapped hypothesis tests like on the mean for example we must first subtract the empirical mean $\hat{\mu}$ of the sample and secondly add to it the mean under $H_0 = \mu$. Then we compare the original observe sample mean $\hat{\mu}$ to our bootstrapped null distribution $H_0 = \mu$ and see how many bootstrapped means are at least as extreme as $\hat{\mu}$ is. This gives us a p-value.

14.1.18.3 Jackknife Resampling

The jackknife was proposed by M.H. Quenouille in 11949 (holocene calendar) and later refined and given its current name by John Tukey in 11956 according to holocene calendar (it predates other common resampling methods such as the bootstrap). M.H. Quenouille originally developed the method as a procedure for correcting bias. Later, Tukey described its use in constructing confidence limits for a large class of estimators. It is similar to the bootstrap in that it involves resampling, but instead of sampling with replacement, the method samples without replacement.

So in statistics, the "jackknife" is a resampling technique especially useful for variance and bias estimation at least... when these estimators converge quick enough (the number of resampling is much more limited with jackknifing than with bootstrapping). The jackknife estimator of a parameter is found by systematically leaving out each observation from a dataset and calculating the estimate and then finding the average of these calculations. Given a sample of size N, the jackknife estimate is found by aggregating the estimates of each $N-1$ estimate in the sample.



This method is especially useful when:

1. The computer is not powerful enough to run a bootstrap
2. We don't trust the resampling method as not suited to the situation

Definition (#14.1.566): The "delete-1 Jackknife samples" are selected by taking the original data vector and deleting one observation from the set. Thus, there are n unique Jackknife samples, and the ith Jackknife sample vector is defined as:

$$\vec{X}_{[i]} = \{X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_{n-1}, X_n\} \tag{14.1.1640}$$

This procedure is obviously generalizable to k deletions.

The ith Jackknife replicate is defined as the value of the estimator $s(\cdot)$ evaluated at the ith Jackknife sample.

$$\hat{\theta}_{(i)} := s(\vec{X}_{[i]}) \tag{14.1.1641}$$

As we will prove it further below, the jackknife standard error of the estimator is (given typically by the bootstrap package of `R` as you can see it in the corresponding companion book):

$$\boxed{\text{SE}(\hat{\theta})_{\text{jack}} = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}\right)^2}} \tag{14.1.1642}$$

where $\hat{\theta}_{(\cdot)}$ is the empirical average of the Jackknife replicates:

$$\hat{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)} \tag{14.1.1643}$$

with:

$$\hat{\theta}_i = \bar{x}_i = \frac{\sum_j x_j - x_i}{n-1} = \frac{n\bar{x} - x_i}{n-1} \tag{14.1.1644}$$

For the proof let us consider the special case where the Jackknife estimator above is an unbiased estimator of the variance of the sample mean.

Proof 14.6.2. So to prove the previous relation with the sample mean we just need to prove that (see section Statistics page 611):

$$\frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_i - \hat{\theta}_{(\cdot)}\right)^2 \tag{14.1.1645}$$

is equal to:

$$\hat{V}_{\overline{X}} = \frac{\hat{\sigma}^2}{n} = \frac{\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \tag{14.1.1646}$$

And to prove that latter, we write it craftily:

$$\begin{aligned}\left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}\right) &= \left(\frac{n\bar{x} - x_i}{n-1}\right) - \frac{1}{n}\sum_{i=1}^{n}\bar{x}_{(i)} = \left(\frac{n\bar{x} - x_i}{n-1}\right) - \frac{n-1}{(n-1)n}\sum_{i=1}^{n}\bar{x}_{(i)}\\ &= \frac{1}{n-1}\left(n\bar{x} - x_i - \frac{1}{n}\sum_{i=1}^{n}n\bar{x}_{(i)} - \bar{x}_{(i)}\right) = \frac{1}{n-1}\left(n\bar{x} - x_i - \frac{1}{n}\sum_{i=1}^{n}n\bar{x} - x_i\right)\\ &= \frac{1}{n-1}(\bar{x} - x_i)\end{aligned} \tag{14.1.1647}$$

Once the term is squared, the equation is complete, and is identically equal to the right hand term above. Thus, in the case of the sample mean, the Jackknife estimate of the standard error reduces to the regular, unbiased estimator commonly used.

We say sometimes then that $n-1$ is the "standard error jackknife bias inflation factor".

■ Q.E.D.

As practical business oriented example of the use of Jackknife let me give the example of a customer (Fortune 500 company) that has to analyse worldwide counterfeiting of its products by sampling and controlling completely each year a given area of a given city chosen randomly in various countries and calculating the sum by country of counterfeit products as it has a major influence on the strategy of the company, national politics and borders controls (the counterfeiting being manly done by the mafia). As the geographical sampling error cannot be calculated as it is not guarantee that the chosen city has the same heterogeneity overall cities of the country it is as far as we know impossible to calculate with a closed form equation a tolerance interval for the real total counterfeit. So an easy way to get a tolerance interval is to make a Jackknife resampling as it is easily acceptable for the board committee to make an analysis of what would have been the sum if for example we removed 50% of the sampling (at the condition that the a posteriori power of the test is still important enough!).

14.1.18.4 Permutation Tests

A "permutation test" (also named a "randomization test" or "re-randomization test") is a type of statistical significance test in which the distribution of the test statistic under the null hypothesis is obtained by calculating all possible values of the test statistic under rearrangements of the labels on the observed data points.

To illustrate the basic idea of a permutation test, suppose we have two groups (we take the special case where the size of the groups are equal!):

$$\begin{aligned}\vec{A}&=\{a_1,a_2,\ldots,a_n\}\\ \vec{B}&=\{b_1,b_2,\ldots,b_n\}\end{aligned}\qquad(14.1.1648)$$

whose sample means are $\bar{x}_A$ and $\bar{x}_B$, and that we want to test, at 5% significance level, whether they come from the same distribution. Let n be the sample size corresponding to each group. The permutation test is designed to determine whether the observed difference between the sample means is large enough to reject the null hypothesis:

$$H_0: \quad \bar{x}_A=\bar{x}_B \qquad(14.1.1649)$$

that the two groups have identical probability distributions (i.e. identical mean).

The test proceeds as follows:

1. The difference in means between the two samples is calculated as we do normally. This lead to the observed value of the test statistic $T_n^{(\text{obs})}$.

2. Then the observations of groups $\vec{A}$ and $\vec{B}$ are pooled and mixed formally using a "random sampling operator" $\vec{\delta}_n^R$ defined by:

$$\boxed{\vec{\delta}_n^R = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{pmatrix}} \tag{14.1.1650}$$

   where r_i is a binary random variable defined by $r_i \in \mathbb{Z}_2$, where $\mathbb{Z}_2 = \{0, 1\}$ and we also define its binary negation by:

$$\bar{\vec{\delta}}_n^R = \begin{pmatrix} \bar{r}_1 \\ \bar{r}_2 \\ \vdots \\ \bar{r}_n \end{pmatrix} \tag{14.1.1651}$$

   Then we create a permutation of $\vec{A}$ denoted $\vec{A}\,'$ (same respectively for $\vec{B}$) by applying the random sampling operator:

$$\begin{aligned} \vec{A}\,' &= \vec{\delta}_n^R \vec{A} + \bar{\vec{\delta}}_n^R \vec{B} \\ \vec{B}\,' &= \bar{\vec{\delta}}_n^R \vec{A} + \vec{\delta}_n^R \vec{B} \end{aligned} \tag{14.1.1652}$$

3. Next, the difference in sample means is calculated and recorded for this new groups $\vec{A}'$ and $\vec{B}'$ and we get a new $T_n^{\prime(\text{obs})}$.

4. We repeat the above procedure for every possible way of combining the samples (in practice we do that completely randomly then some cases - pairs of samples - may repeat).

   The number of ways in which $m \cdot n$ different items can be divided equally into m groups ($m = 2$ in our example) each containing n objects where the order of the groups is important is given by:

$$\frac{(mn)!}{(n!)^m} \tag{14.1.1653}$$

   And the number of ways in which $m \cdot n$ different items can be divided equally into m groups, each containing n objects and where the order of the groups is NOT important is:

$$\frac{(mn)!}{(n!)^m m!} \tag{14.1.1654}$$

5. Finally the one-sided p-value of the test is calculated as the proportion of sampled permutations where the difference in means was greater than or equal to $T_n^{(\text{obs})}$. The two-sided p-value of the test is calculated as the proportion of sampled permutations where the absolute difference was greater than or equal to $|T_n^{\prime(\text{obs})}|$.

The major down-side to permutation tests are that they:

- Can be computationally intensive and may require "custom" code for difficult-to-calculate statistics. This must be rewritten for every case.
- Are primarily used to provide a p-value. The inversion of the test to get confidence regions/intervals requires even more computation.

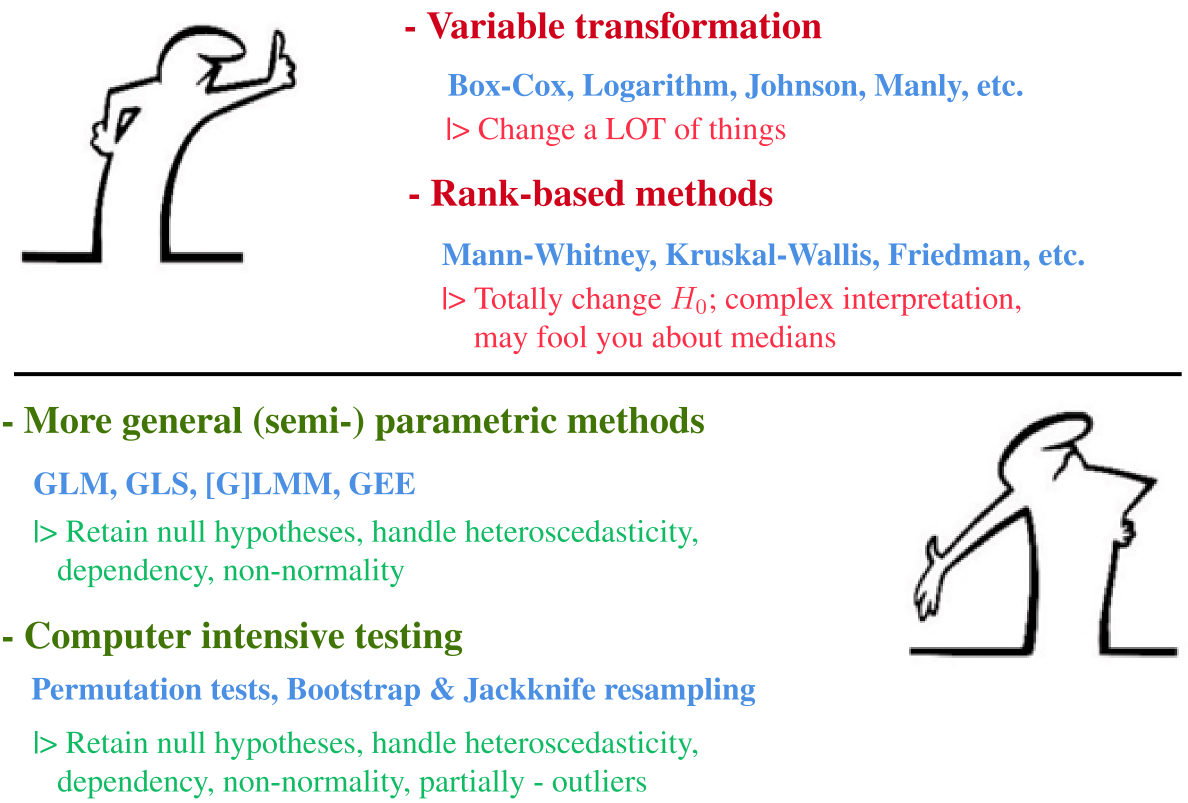
Hopefully it is quite easy to do that with modern software. You can see the `R` companion book for an example of a Student-T permutation test.

Keep in mind that as the assumption behind a permutation test is that the observations are exchangeable under the null hypothesis (assumption of "data exchangeability"), an important consequence of this assumption is that tests of difference in location (like a permutation T-test) require equal variance.

Remark

Permutation tests are a subset of nonparametric statistics as the basic premise is to use only the assumption that it is possible that all of the treatment groups are equivalent, and that every member of them is the same before sampling began (i.e. the slot that they fill is not differentiable from other slots before the slots are filled).

And keep in mind that excellent meme by Adrian Olszewski:



- Variable transformation

Box-Cox, Logarithm, Johnson, Manly, etc.

|> Change a LOT of things

- Rank-based methods

Mann-Whitney, Kruskal-Wallis, Friedman, etc.

|> Totally change H_0; complex interpretation, may fool you about medians

- More general (semi-) parametric methods

GLM, GLS, [G]LMM, GEE

|> Retain null hypotheses, handle heteroscedasticity, dependency, non-normality

- Computer intensive testing

Permutation tests, Bootstrap & Jackknife resampling

|> Retain null hypotheses, handle heteroscedasticity, dependency, non-normality, partially - outliers

14.1.19 Finite difference method (F.D.M.)

The "finite element method" (FEM) is a numerical technique for finding approximate solutions to boundary value problems for partial differential equations. It is also referred to as finite element analysis (FEA). FEM subdivides a large problem into smaller, simpler, parts, named "finite elements". The simple equations that model these finite elements are then assembled into a larger system of equations that models the entire problem. FEM then uses variational methods from the calculus of variations to approximate a solution by minimizing an associated error function.

Finite difference method helps to accelerate Research and Development in many fields like illustrated in the figure below as it helped for example to decrease significantly the number of combinatorics to find the optimal set of combinations of atoms to found the most promising material for new batteries:

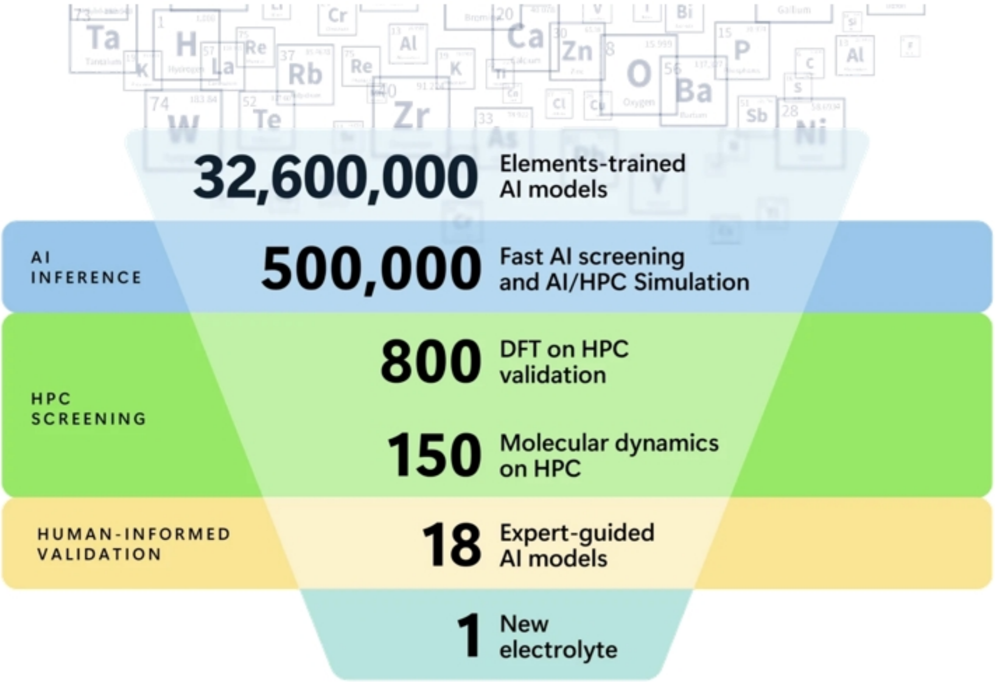

32,600,000 Elements-trained AI models
AI INFERENCE
500,000 Fast AI screening and AI/HPC Simulation
HPC SCREENING
800 DFT on HPC validation
150 Molecular dynamics on HPC
HUMAN-INFORMED VALIDATION
18 Expert-guided AI models
1 New electrolyte

Figure 14.102: Source Generative Chemistry and Accelerated DFT (source: Microsoft)

Relatively to the above figure DFT (Density Functional Theory) is based on an adaptive finite-element discretization that handles pseudopotential and all-electron calculations in the same framework. HPC is the acronym of High Performance Calculators.

14.1.19.1 One space dimension F.D.M.

Let us recall that we have proved in the section of Thermodynamics (see page 3684) the following heat diffusion equation (we present here the equation reduced to only one spatial dimension):

$$\frac{\partial T(x,t)}{\partial t}-\frac{\kappa}{\rho c_p}\frac{\partial^2 T(x,t)}{\partial x^2}=0 \qquad (14.1.1655)$$

and let us notice that this equation is not very general ... (it is not relativistic and does not take into account the heat generated in the form of radiation by the concerned material concerned or many other factors ...).

We can consider (see section Differential and Integral Calculus page 1450) that:

$$\frac{\partial T(x,t)}{\partial x}\cong\frac{T(x+h,t)-T(x,t)}{h} \qquad (14.1.1656)$$

and:

$$\begin{aligned}\frac{\partial^2 T(x,t)}{\partial^2 x}&\cong\frac{1}{h}\left(\frac{T(x+h,t)-T(x,t)}{h}-\frac{T(x,t)-T(x-h,t)}{h}\right)\\&=\frac{1}{h^2}\left(T(x+h,t)-2T(x,t)+T(x-h,t)\right)\end{aligned} \qquad (14.1.1657)$$

Also:

$$\frac{\partial T(x,t)}{\partial t}\cong\frac{T(x,t+k)-T(x,t)}{k} \qquad (14.1.1658)$$

Then the heat equation becomes:

$$\frac{T(x,t+k)-T(x,t)}{k}-\frac{\kappa}{\rho c_p}\frac{1}{h^2}\left(T(x+h,t)-2T(x,t)+T(x-h,t)\right)=0 \qquad (14.1.1659)$$

After rearranging, we have:

$$\boxed{\begin{aligned}T(x,t+k)&=k\frac{\kappa}{\rho c_p}\frac{1}{h^2}\left(T(x+h,t)-2T(x,t)+T(x-h,t)\right)+T(x,t)\\&=\left(1-\frac{2k\kappa}{\rho c_p h^2}\right)T(x,t)+k\frac{\kappa}{\rho c_p}\frac{1}{h^2}\left(T(x+h,t)+T(x-h,t)\right)\end{aligned}} \qquad (14.1.1660)$$

If we look at this relation more closely, we see that this is a simple recursion. We just need to know the initial distribution (initial conditions) $T(x,0)$ to determine the distribution then all other values as:

$$T(x,0+k)=\left(1-\frac{2k\kappa}{\rho c_p h^2}\right)T(x,0)+k\frac{\kappa}{\rho c_p}\frac{1}{h^2}\left(T(x+h,0)+T(x-h,0)\right) \qquad (14.1.1661)$$

and:

$$T(x,2k)=\left(1-\frac{2k\kappa}{\rho c_p h^2}\right)T(x,k)+k\frac{\kappa}{\rho c_p}\frac{1}{h^2}\left(T(x+h,k)+T(x-h,k)\right) \qquad (14.1.1662)$$

etc. It is possible to implement such a simulation with nothing but a small spreadsheet software and a little time as we will we see just after... (using a spreadsheet software to understand the mechanism is better than using a blackbox like Maple™ or MATLAB® by my experience).

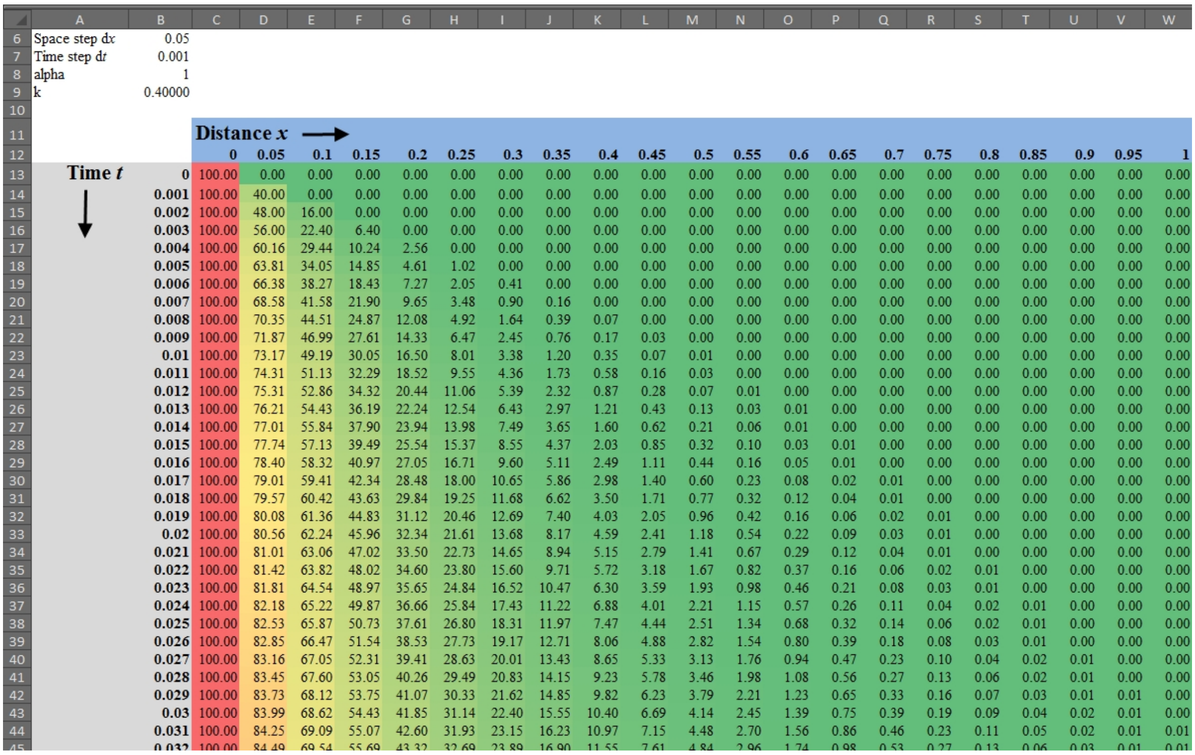
For information h and k are named then the "mesh step" or "space step" of the model.

Let us see an application example with a spreadsheet software like Microsoft® Excel® as it is quite a good practical exercise to understand how to implement the method.

So let us consider the following worksheet[38] where we consider a bar that is initially at a temperature of 0 [C]and that is heated on the left-hand side at a constant temperature of 100 [C] and where we use the relation proved earlier above:

$$\begin{aligned}T(x,t+k)&=T(x,t)+k\frac{\kappa}{\rho c_p}\frac{1}{h^2}\left(T(x+h,t)-2T(x,t)+T(x-h,t)\right)\\&=T(x,t)+\Delta t\frac{\kappa}{\rho c_p}\frac{1}{\Delta x^2}\left(T(x+\Delta x,t)-2T(x,t)+T(x-\Delta x,t)\right)\\&=T(x,t)+\alpha\frac{\Delta t}{\Delta x^2}\left(T(x+\Delta x,t)-2T(x,t)+T(x-\Delta x,t)\right)\end{aligned}\tag{14.1.1663}$$

So this gives (first rows only of $1,000$ rows):



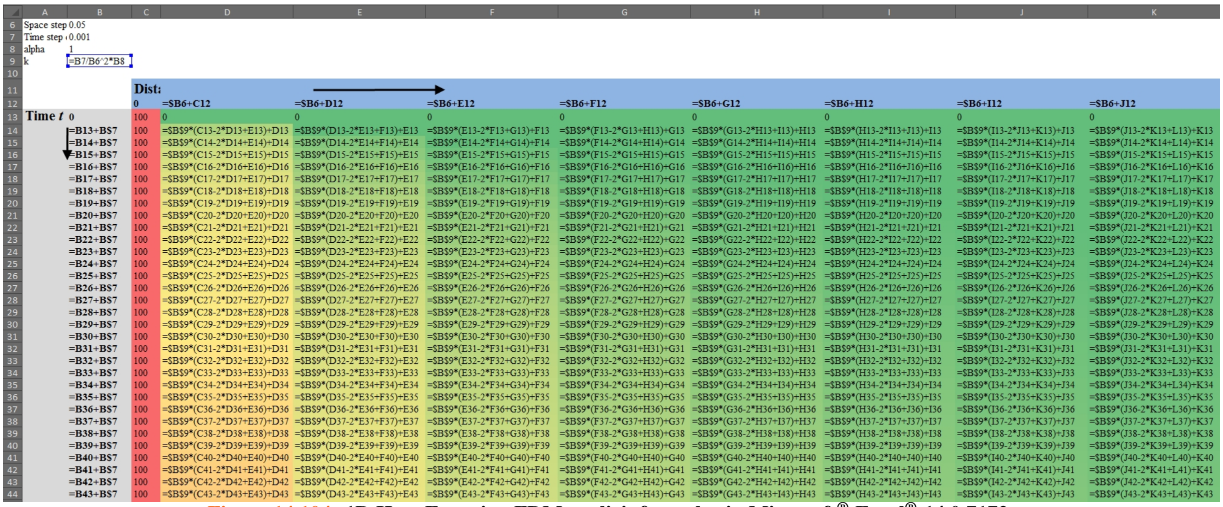
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
6	Space step *dx*	0.05																					
7	Time step *dt*	0.001																					
8	alpha	1																					
9	k	0.40000																					
10																							
11			**Distance *x* →**																				
12			**0**	**0.05**	**0.1**	**0.15**	**0.2**	**0.25**	**0.3**	**0.35**	**0.4**	**0.45**	**0.5**	**0.55**	**0.6**	**0.65**	**0.7**	**0.75**	**0.8**	**0.85**	**0.9**	**0.95**	**1**
13	**Time *t***	**0**	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
14	↓	**0.001**	100.00	40.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15		**0.002**	100.00	48.00	16.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
16		**0.003**	100.00	56.00	22.40	6.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
17		**0.004**	100.00	60.16	29.44	10.24	2.56	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
18		**0.005**	100.00	63.81	34.05	14.85	4.61	1.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
19		**0.006**	100.00	66.38	38.27	18.43	7.27	2.05	0.41	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
20		**0.007**	100.00	68.58	41.58	21.90	9.65	3.48	0.90	0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
21		**0.008**	100.00	70.35	44.51	24.87	12.08	4.92	1.64	0.39	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
22		**0.009**	100.00	71.87	46.99	27.61	14.33	6.47	2.45	0.76	0.17	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
23		**0.01**	100.00	73.17	49.19	30.05	16.50	8.01	3.38	1.20	0.35	0.07	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
24		**0.011**	100.00	74.31	51.13	32.29	18.52	9.55	4.36	1.73	0.58	0.16	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25		**0.012**	100.00	75.31	52.86	34.32	20.44	11.06	5.39	2.32	0.87	0.28	0.07	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
26		**0.013**	100.00	76.21	54.43	36.19	22.24	12.54	6.43	2.97	1.21	0.43	0.13	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
27		**0.014**	100.00	77.01	55.84	37.90	23.94	13.98	7.49	3.65	1.60	0.62	0.21	0.06	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
28		**0.015**	100.00	77.74	57.13	39.49	25.54	15.37	8.55	4.37	2.03	0.85	0.32	0.10	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
29		**0.016**	100.00	78.40	58.32	40.97	27.05	16.71	9.60	5.11	2.49	1.11	0.44	0.16	0.05	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
30		**0.017**	100.00	79.01	59.41	42.34	28.48	18.00	10.65	5.86	2.98	1.40	0.60	0.23	0.08	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00
31		**0.018**	100.00	79.57	60.42	43.63	29.84	19.25	11.68	6.62	3.50	1.71	0.77	0.32	0.12	0.04	0.01	0.00	0.00	0.00	0.00	0.00	0.00
32		**0.019**	100.00	80.08	61.36	44.83	31.12	20.46	12.69	7.40	4.03	2.05	0.96	0.42	0.16	0.06	0.02	0.01	0.00	0.00	0.00	0.00	0.00
33		**0.02**	100.00	80.56	62.24	45.96	32.34	21.61	13.68	8.17	4.59	2.41	1.18	0.54	0.22	0.09	0.03	0.01	0.00	0.00	0.00	0.00	0.00
34		**0.021**	100.00	81.01	63.06	47.02	33.50	22.73	14.65	8.94	5.15	2.79	1.41	0.67	0.29	0.12	0.04	0.01	0.00	0.00	0.00	0.00	0.00
35		**0.022**	100.00	81.42	63.82	48.02	34.60	23.80	15.60	9.71	5.72	3.18	1.67	0.82	0.37	0.16	0.06	0.02	0.01	0.00	0.00	0.00	0.00
36		**0.023**	100.00	81.81	64.54	48.97	35.65	24.84	16.52	10.47	6.30	3.59	1.93	0.98	0.46	0.21	0.08	0.03	0.01	0.00	0.00	0.00	0.00
37		**0.024**	100.00	82.18	65.22	49.87	36.66	25.84	17.43	11.22	6.88	4.01	2.21	1.15	0.57	0.26	0.11	0.04	0.02	0.01	0.00	0.00	0.00
38		**0.025**	100.00	82.53	65.87	50.73	37.61	26.80	18.31	11.97	7.47	4.44	2.51	1.34	0.68	0.32	0.14	0.06	0.02	0.01	0.00	0.00	0.00
39		**0.026**	100.00	82.85	66.47	51.54	38.53	27.73	19.17	12.71	8.06	4.88	2.82	1.54	0.80	0.39	0.18	0.08	0.03	0.01	0.00	0.00	0.00
40		**0.027**	100.00	83.16	67.05	52.31	39.41	28.63	20.01	13.43	8.65	5.33	3.13	1.76	0.94	0.47	0.23	0.10	0.04	0.02	0.01	0.00	0.00
41		**0.028**	100.00	83.45	67.60	53.05	40.26	29.49	20.83	14.15	9.23	5.78	3.46	1.98	1.08	0.56	0.27	0.13	0.06	0.02	0.01	0.00	0.00
42		**0.029**	100.00	83.73	68.12	53.75	41.07	30.33	21.62	14.85	9.82	6.23	3.79	2.21	1.23	0.65	0.33	0.16	0.07	0.03	0.01	0.01	0.00
43		**0.03**	100.00	83.99	68.62	54.43	41.85	31.14	22.40	15.55	10.40	6.69	4.14	2.45	1.39	0.75	0.39	0.19	0.09	0.04	0.02	0.01	0.00
44		**0.031**	100.00	84.25	69.09	55.07	42.60	31.93	23.15	16.23	10.97	7.15	4.48	2.70	1.56	0.86	0.46	0.23	0.11	0.05	0.02	0.01	0.01
45		**0.032**	100.00	84.49	69.54	55.69	43.32	32.69	23.89	16.90	11.55	7.61	4.84	2.96	1.74	0.98	0.53	0.27	0.13	0.06	0.03	0.01	0.01

Figure 14.103: **1D heat equation FDM calculations in Microsoft® Excel® 14.0.7172**

Explicitly for only a few rows a and few columns visible, we have:

---

[38] source: `http://excelcalculations.blogspot.co.at/2011/04/solving-1d-heat-equation-using-finite.html`



	A	B	C	D	E	F	G	H	I	J	K
6	Space step	0.05									
7	Time step (	0.001									
8	alpha	1									
9	k	=B7/B6^2*B8									
10											
11			**Dist:**								
12			**0**	**=$B6+C12**	**=$B6+D12**	**=$B6+E12**	**=$B6+F12**	**=$B6+G12**	**=$B6+H12**	**=$B6+I12**	**=$B6+J12**
13	**Time *t***	0	100	0	0	0	0	0	0	0	0
14		**=B13+BS7**	100	=$B$9*(C13-2*D13+E13)+D13	=$B$9*(D13-2*E13+F13)+E13	=$B$9*(E13-2*F13+G13)+F13	=$B$9*(F13-2*G13+H13)+G13	=$B$9*(G13-2*H13+I13)+H13	=$B$9*(H13-2*I13+J13)+I13	=$B$9*(I13-2*J13+K13)+J13	=$B$9*(J13-2*K13+L13)+K13
15		**=B14+BS7**	100	=$B$9*(C14-2*D14+E14)+D14	=$B$9*(D14-2*E14+F14)+E14	=$B$9*(E14-2*F14+G14)+F14	=$B$9*(F14-2*G14+H14)+G14	=$B$9*(G14-2*H14+I14)+H14	=$B$9*(H14-2*I14+J14)+I14	=$B$9*(I14-2*J14+K14)+J14	=$B$9*(J14-2*K14+L14)+K14
16		**=B15+BS7**	100	=$B$9*(C15-2*D15+E15)+D15	=$B$9*(D15-2*E15+F15)+E15	=$B$9*(E15-2*F15+G15)+F15	=$B$9*(F15-2*G15+H15)+G15	=$B$9*(G15-2*H15+I15)+H15	=$B$9*(H15-2*I15+J15)+I15	=$B$9*(I15-2*J15+K15)+J15	=$B$9*(J15-2*K15+L15)+K15
17		**=B16+BS7**	100	=$B$9*(C16-2*D16+E16)+D16	=$B$9*(D16-2*E16+F16)+E16	=$B$9*(E16-2*F16+G16)+F16	=$B$9*(F16-2*G16+H16)+G16	=$B$9*(G16-2*H16+I16)+H16	=$B$9*(H16-2*I16+J16)+I16	=$B$9*(I16-2*J16+K16)+J16	=$B$9*(J16-2*K16+L16)+K16
18		**=B17+BS7**	100	=$B$9*(C17-2*D17+E17)+D17	=$B$9*(D17-2*E17+F17)+E17	=$B$9*(E17-2*F17+G17)+F17	=$B$9*(F17-2*G17+H17)+G17	=$B$9*(G17-2*H17+I17)+H17	=$B$9*(H17-2*I17+J17)+I17	=$B$9*(I17-2*J17+K17)+J17	=$B$9*(J17-2*K17+L17)+K17
19		**=B18+BS7**	100	=$B$9*(C18-2*D18+E18)+D18	=$B$9*(D18-2*E18+F18)+E18	=$B$9*(E18-2*F18+G18)+F18	=$B$9*(F18-2*G18+H18)+G18	=$B$9*(G18-2*H18+I18)+H18	=$B$9*(H18-2*I18+J18)+I18	=$B$9*(I18-2*J18+K18)+J18	=$B$9*(J18-2*K18+L18)+K18
20		**=B19+BS7**	100	=$B$9*(C19-2*D19+E19)+D19	=$B$9*(D19-2*E19+F19)+E19	=$B$9*(E19-2*F19+G19)+F19	=$B$9*(F19-2*G19+H19)+G19	=$B$9*(G19-2*H19+I19)+H19	=$B$9*(H19-2*I19+J19)+I19	=$B$9*(I19-2*J19+K19)+J19	=$B$9*(J19-2*K19+L19)+K19
21		**=B20+BS7**	100	=$B$9*(C20-2*D20+E20)+D20	=$B$9*(D20-2*E20+F20)+E20	=$B$9*(E20-2*F20+G20)+F20	=$B$9*(F20-2*G20+H20)+G20	=$B$9*(G20-2*H20+I20)+H20	=$B$9*(H20-2*I20+J20)+I20	=$B$9*(I20-2*J20+K20)+J20	=$B$9*(J20-2*K20+L20)+K20
22		**=B21+BS7**	100	=$B$9*(C21-2*D21+E21)+D21	=$B$9*(D21-2*E21+F21)+E21	=$B$9*(E21-2*F21+G21)+F21	=$B$9*(F21-2*G21+H21)+G21	=$B$9*(G21-2*H21+I21)+H21	=$B$9*(H21-2*I21+J21)+I21	=$B$9*(I21-2*J21+K21)+J21	=$B$9*(J21-2*K21+L21)+K21
23		**=B22+BS7**	100	=$B$9*(C22-2*D22+E22)+D22	=$B$9*(D22-2*E22+F22)+E22	=$B$9*(E22-2*F22+G22)+F22	=$B$9*(F22-2*G22+H22)+G22	=$B$9*(G22-2*H22+I22)+H22	=$B$9*(H22-2*I22+J22)+I22	=$B$9*(I22-2*J22+K22)+J22	=$B$9*(J22-2*K22+L22)+K22
24		**=B23+BS7**	100	=$B$9*(C23-2*D23+E23)+D23	=$B$9*(D23-2*E23+F23)+E23	=$B$9*(E23-2*F23+G23)+F23	=$B$9*(F23-2*G23+H23)+G23	=$B$9*(G23-2*H23+I23)+H23	=$B$9*(H23-2*I23+J23)+I23	=$B$9*(I23-2*J23+K23)+J23	=$B$9*(J23-2*K23+L23)+K23
25		**=B24+BS7**	100	=$B$9*(C24-2*D24+E24)+D24	=$B$9*(D24-2*E24+F24)+E24	=$B$9*(E24-2*F24+G24)+F24	=$B$9*(F24-2*G24+H24)+G24	=$B$9*(G24-2*H24+I24)+H24	=$B$9*(H24-2*I24+J24)+I24	=$B$9*(I24-2*J24+K24)+J24	=$B$9*(J24-2*K24+L24)+K24
26		**=B25+BS7**	100	=$B$9*(C25-2*D25+E25)+D25	=$B$9*(D25-2*E25+F25)+E25	=$B$9*(E25-2*F25+G25)+F25	=$B$9*(F25-2*G25+H25)+G25	=$B$9*(G25-2*H25+I25)+H25	=$B$9*(H25-2*I25+J25)+I25	=$B$9*(I25-2*J25+K25)+J25	=$B$9*(J25-2*K25+L25)+K25
27		**=B26+BS7**	100	=$B$9*(C26-2*D26+E26)+D26	=$B$9*(D26-2*E26+F26)+E26	=$B$9*(E26-2*F26+G26)+F26	=$B$9*(F26-2*G26+H26)+G26	=$B$9*(G26-2*H26+I26)+H26	=$B$9*(H26-2*I26+J26)+I26	=$B$9*(I26-2*J26+K26)+J26	=$B$9*(J26-2*K26+L26)+K26
28		**=B27+BS7**	100	=$B$9*(C27-2*D27+E27)+D27	=$B$9*(D27-2*E27+F27)+E27	=$B$9*(E27-2*F27+G27)+F27	=$B$9*(F27-2*G27+H27)+G27	=$B$9*(G27-2*H27+I27)+H27	=$B$9*(H27-2*I27+J27)+I27	=$B$9*(I27-2*J27+K27)+J27	=$B$9*(J27-2*K27+L27)+K27
29		**=B28+BS7**	100	=$B$9*(C28-2*D28+E28)+D28	=$B$9*(D28-2*E28+F28)+E28	=$B$9*(E28-2*F28+G28)+F28	=$B$9*(F28-2*G28+H28)+G28	=$B$9*(G28-2*H28+I28)+H28	=$B$9*(H28-2*I28+J28)+I28	=$B$9*(I28-2*J28+K28)+J28	=$B$9*(J28-2*K28+L28)+K28
30		**=B29+BS7**	100	=$B$9*(C29-2*D29+E29)+D29	=$B$9*(D29-2*E29+F29)+E29	=$B$9*(E29-2*F29+G29)+F29	=$B$9*(F29-2*G29+H29)+G29	=$B$9*(G29-2*H29+I29)+H29	=$B$9*(H29-2*I29+J29)+I29	=$B$9*(I29-2*J29+K29)+J29	=$B$9*(J29-2*K29+L29)+K29
31		**=B30+BS7**	100	=$B$9*(C30-2*D30+E30)+D30	=$B$9*(D30-2*E30+F30)+E30	=$B$9*(E30-2*F30+G30)+F30	=$B$9*(F30-2*G30+H30)+G30	=$B$9*(G30-2*H30+I30)+H30	=$B$9*(H30-2*I30+J30)+I30	=$B$9*(I30-2*J30+K30)+J30	=$B$9*(J30-2*K30+L30)+K30
32		**=B31+BS7**	100	=$B$9*(C31-2*D31+E31)+D31	=$B$9*(D31-2*E31+F31)+E31	=$B$9*(E31-2*F31+G31)+F31	=$B$9*(F31-2*G31+H31)+G31	=$B$9*(G31-2*H31+I31)+H31	=$B$9*(H31-2*I31+J31)+I31	=$B$9*(I31-2*J31+K31)+J31	=$B$9*(J31-2*K31+L31)+K31
33		**=B32+BS7**	100	=$B$9*(C32-2*D32+E32)+D32	=$B$9*(D32-2*E32+F32)+E32	=$B$9*(E32-2*F32+G32)+F32	=$B$9*(F32-2*G32+H32)+G32	=$B$9*(G32-2*H32+I32)+H32	=$B$9*(H32-2*I32+J32)+I32	=$B$9*(I32-2*J32+K32)+J32	=$B$9*(J32-2*K32+L32)+K32
34		**=B33+BS7**	100	=$B$9*(C33-2*D33+E33)+D33	=$B$9*(D33-2*E33+F33)+E33	=$B$9*(E33-2*F33+G33)+F33	=$B$9*(F33-2*G33+H33)+G33	=$B$9*(G33-2*H33+I33)+H33	=$B$9*(H33-2*I33+J33)+I33	=$B$9*(I33-2*J33+K33)+J33	=$B$9*(J33-2*K33+L33)+K33
35		**=B34+BS7**	100	=$B$9*(C34-2*D34+E34)+D34	=$B$9*(D34-2*E34+F34)+E34	=$B$9*(E34-2*F34+G34)+F34	=$B$9*(F34-2*G34+H34)+G34	=$B$9*(G34-2*H34+I34)+H34	=$B$9*(H34-2*I34+J34)+I34	=$B$9*(I34-2*J34+K34)+J34	=$B$9*(J34-2*K34+L34)+K34
36		**=B35+BS7**	100	=$B$9*(C35-2*D35+E35)+D35	=$B$9*(D35-2*E35+F35)+E35	=$B$9*(E35-2*F35+G35)+F35	=$B$9*(F35-2*G35+H35)+G35	=$B$9*(G35-2*H35+I35)+H35	=$B$9*(H35-2*I35+J35)+I35	=$B$9*(I35-2*J35+K35)+J35	=$B$9*(J35-2*K35+L35)+K35
37		**=B36+BS7**	100	=$B$9*(C36-2*D36+E36)+D36	=$B$9*(D36-2*E36+F36)+E36	=$B$9*(E36-2*F36+G36)+F36	=$B$9*(F36-2*G36+H36)+G36	=$B$9*(G36-2*H36+I36)+H36	=$B$9*(H36-2*I36+J36)+I36	=$B$9*(I36-2*J36+K36)+J36	=$B$9*(J36-2*K36+L36)+K36
38		**=B37+BS7**	100	=$B$9*(C37-2*D37+E37)+D37	=$B$9*(D37-2*E37+F37)+E37	=$B$9*(E37-2*F37+G37)+F37	=$B$9*(F37-2*G37+H37)+G37	=$B$9*(G37-2*H37+I37)+H37	=$B$9*(H37-2*I37+J37)+I37	=$B$9*(I37-2*J37+K37)+J37	=$B$9*(J37-2*K37+L37)+K37
39		**=B38+BS7**	100	=$B$9*(C38-2*D38+E38)+D38	=$B$9*(D38-2*E38+F38)+E38	=$B$9*(E38-2*F38+G38)+F38	=$B$9*(F38-2*G38+H38)+G38	=$B$9*(G38-2*H38+I38)+H38	=$B$9*(H38-2*I38+J38)+I38	=$B$9*(I38-2*J38+K38)+J38	=$B$9*(J38-2*K38+L38)+K38
40		**=B39+BS7**	100	=$B$9*(C39-2*D39+E39)+D39	=$B$9*(D39-2*E39+F39)+E39	=$B$9*(E39-2*F39+G39)+F39	=$B$9*(F39-2*G39+H39)+G39	=$B$9*(G39-2*H39+I39)+H39	=$B$9*(H39-2*I39+J39)+I39	=$B$9*(I39-2*J39+K39)+J39	=$B$9*(J39-2*K39+L39)+K39
41		**=B40+BS7**	100	=$B$9*(C40-2*D40+E40)+D40	=$B$9*(D40-2*E40+F40)+E40	=$B$9*(E40-2*F40+G40)+F40	=$B$9*(F40-2*G40+H40)+G40	=$B$9*(G40-2*H40+I40)+H40	=$B$9*(H40-2*I40+J40)+I40	=$B$9*(I40-2*J40+K40)+J40	=$B$9*(J40-2*K40+L40)+K40
42		**=B41+BS7**	100	=$B$9*(C41-2*D41+E41)+D41	=$B$9*(D41-2*E41+F41)+E41	=$B$9*(E41-2*F41+G41)+F41	=$B$9*(F41-2*G41+H41)+G41	=$B$9*(G41-2*H41+I41)+H41	=$B$9*(H41-2*I41+J41)+I41	=$B$9*(I41-2*J41+K41)+J41	=$B$9*(J41-2*K41+L41)+K41
43		**=B42+BS7**	100	=$B$9*(C42-2*D42+E42)+D42	=$B$9*(D42-2*E42+F42)+E42	=$B$9*(E42-2*F42+G42)+F42	=$B$9*(F42-2*G42+H42)+G42	=$B$9*(G42-2*H42+I42)+H42	=$B$9*(H42-2*I42+J42)+I42	=$B$9*(I42-2*J42+K42)+J42	=$B$9*(J42-2*K42+L42)+K42
44		**=B43+BS7**	100	=$B$9*(C43-2*D43+E43)+D43	=$B$9*(D43-2*E43+F43)+E43	=$B$9*(E43-2*F43+G43)+F43	=$B$9*(F43-2*G43+H43)+G43	=$B$9*(G43-2*H43+I43)+H43	=$B$9*(H43-2*I43+J43)+I43	=$B$9*(I43-2*J43+K43)+J43	=$B$9*(J43-2*K43+L43)+K43

Figure 14.104: **1D Heat Equation FDM explicit formulas in Microsoft® Excel® 14.0.7172**

We have above taken for boundary conditions:

$$\begin{aligned} T(0,t) &= T_0 \\ T(1,t+\Delta t) &= T(1,t)+2\alpha\frac{\Delta t}{\Delta x^2}(T(1-\Delta x,t)-T(1,t)) \end{aligned} \tag{14.1.1664}$$

All this with a chart view (famous figure that is painful to obtain in a spreadsheet software...):

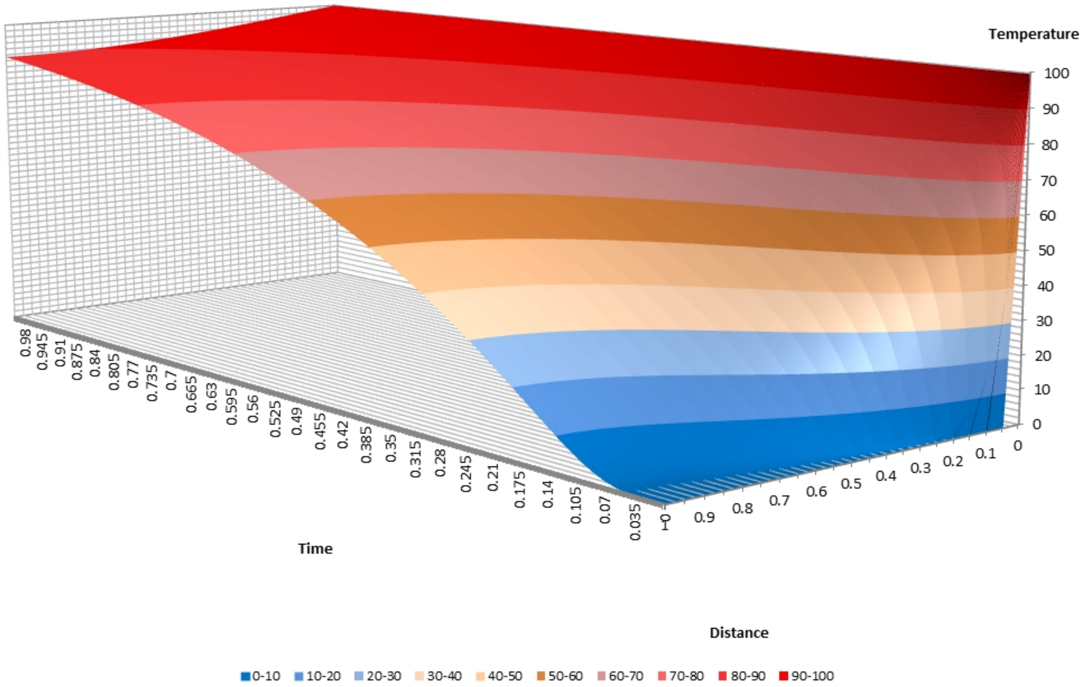


Figure 14.105: **1D Heat Equation FDM plot in Microsoft® Excel® 14.0.7172**

For readers wishing to practice with real values ... a longitudinal Iron bar of 1 [kg] has a specific heat capacity of $450\,[\mathrm{J}\cdot\mathrm{kg}^{-1}\cdot\mathrm{K}^{-1}]$, a density of almost $7.88\,[\mathrm{kg}\cdot\mathrm{m}^{-3}]$ and a thermal conductivity of $82\,[\mathrm{J}\cdot\mathrm{s}^{-1}\cdot\mathrm{m}^{-1}\cdot\mathrm{K}^{-1}]$.

However with such a file a above the reader will see that the FDM is stable after some trials and errors if and only if:

$$\alpha\frac{\Delta t}{\Delta x^2}\leq\frac{1}{2} \tag{14.1.1665}$$

This is what we will study now, first with a simple example, and afterwards with a more elaborated one:

14.1.19.1.1 Numerical instability

Numerical instability is a huge issues in numerical computing that occurs in many situation (the most know one being the finite elements method). To explain this let us consider the following companion example:

$$I_n=\int\limits_0^1\frac{x^n}{10+x}\mathrm{d}x \tag{14.1.1666}$$

with $n\in\mathbb{N}$.

An immediate calculation gives:

$$\begin{aligned}I_0&=\int\limits_0^1\frac{\mathrm{d}x}{10+x}=\ln(10+x)\Big|_0^1=\ln\left(\frac{11}{10}\right)\\I_n&=\int\limits_0^1\frac{x}{10+x}x^{n-1}\mathrm{d}x=\int\limits_0^1\left(1-\frac{10}{10+x}\right)x^{n-1}\mathrm{d}x\\&=\int\limits_0^1 x^{n-1}\mathrm{d}x-10\int\limits_0^1\frac{x^{n-1}}{10+x}\mathrm{d}x=\frac{1}{n}-10I_{n-1}\end{aligned} \tag{14.1.1667}$$

This allow us to calculate I_n by recurrence with:

$$\begin{cases}I_0=\ln\left(\dfrac{11}{10}\right)\\I_n=\dfrac{1}{n}-10I_{n-1}\end{cases} \tag{14.1.1668}$$

This problem apparently well posed mathematically leads digitally to catastrophic results. Indeed we have:

$$\Delta I_n\cong 10\Delta I_{n-1} \tag{14.1.1669}$$

even if we neglect the rounding error on $1/n$. The error on I_n explode exponentially, the initial error on I_0 being multiplied by 10^n at the step n.

14.1.19.1.2 von Neumann stability

In numerical analysis, "von Neumann stability analysis" (also known as "Fourier stability analysis") is a procedure used to check the stability of finite difference schemes as applied to linear partial differential equations. The analysis is based on the Fourier decomposition of numerical error and was developed at Los Alamos National Laboratory after having been briefly described in a 11947 (holocene calendar) article by British researchers Crank and Nicolson. This method is an example of explicit time integration where the function that defines governing equation is evaluated at the current time. Later, the method was given a more rigorous treatment in an article co-authored by John von Neumann.

The stability of numerical schemes is closely associated with numerical error. A finite difference scheme is stable if the errors made at one time step of the calculation do not cause the errors to be magnified as the computations are continued. A neutrally stable scheme is one in which errors remain constant as the computations are carried forward. If the errors decay and eventually damp out, the numerical scheme is said to be stable. If, on the contrary, the errors grow with time the numerical scheme is said to be unstable. The stability of numerical schemes can be investigated by performing von Neumann stability analysis. For time-dependent problems, stability guarantees that the numerical method produces a bounded solution whenever the solution of the exact differential equation is bounded. Stability, in general, can be difficult to investigate, especially when the equation under consideration is non-linear.

The von Neumann method is based on the decomposition of the errors into Fourier series. To illustrate the procedure, consider the one-dimensional heat equation:

$$\frac{\partial T(x,t)}{\partial t}=\alpha\frac{\partial^2 T(x,t)}{\partial x^2} \tag{14.1.1670}$$

defined on the spatial interval L, which can be discretized as we have just proved as (in a very condensed form):

$$T_j^{n+1}=T_j^n+r\left(T_{j+1}^n-2T_j^n+T_{j-1}^n\right) \tag{14.1.1671}$$

where as we have just proved earlier:

$$r=\frac{\alpha\Delta t}{\Delta x^2} \tag{14.1.1672}$$

and the solution T_j^n of the discrete equation approximates the analytical solution $T(x,t)$ of the PDE on the grid.

Let us define the round-off error ε_j^n as:

$$\varepsilon_j^n=N_j^n-T_j^n \tag{14.1.1673}$$

where T_j^n is the solution of the discretized PDE as we knot it that would be computed in the absence of round-off error, and N_j^n is the numerical solution obtained in finite precision arithmetic. Since the exact solution T_j^n must satisfy the discretized PDE exactly, the error ε_j^n must also satisfy the discretized equation (superposition principle). Here we assumed that N_j^n satisfies the PDE, too (this is only true in machine precision). Thus:

$$\varepsilon_j^{n+1}=\varepsilon_j^n+r\left(\varepsilon_{j+1}^n-2\varepsilon_j^n+\varepsilon_{j-1}^n\right) \tag{14.1.1674}$$

is a recurrence relation for the error.

Equations:

$$\begin{aligned}T_j^{n+1}&=T_j^n+r\left(T_{j+1}^n-2T_j^n+T_{j-1}^n\right)\\ \varepsilon_j^{n+1}&=\varepsilon_j^n+r\left(\varepsilon_{j+1}^n-2\varepsilon_j^n+\varepsilon_{j-1}^n\right)\end{aligned}\qquad(14.1.1675)$$

show that both the error and the numerical solution have the same growth or decay behaviour with respect to time. For linear differential equations with periodic boundary condition, the spatial variation of error may be expanded in a finite Fourier series, in the interval L, as: As we have proved it in the section Thermodynamics the solution of the PDE for the errors can be written in a condensed way for a given term:

$$\varepsilon(x,t)=\sum_{\lambda}\left(C'_1(\lambda)e^{\mathrm{i}\lambda x}+C'_2(\lambda)e^{-\mathrm{i}\lambda x}\right)e^{-\lambda^2 Dt}\qquad(14.1.1676)$$

Since the difference equation for error is linear (the behaviour of each term of the series is the same as series itself), it is enough to consider the growth of error of a typical term:

$$\varepsilon(x,t)=C'_1(\lambda)e^{\mathrm{i}\lambda x}\qquad(14.1.1677)$$

and as the reader will see with the next development we can simplify already the constant such that it remains:

$$\varepsilon(x,t)=e^{\mathrm{i}\lambda x}e^{-\lambda^2 Dt}\qquad(14.1.1678)$$

The stability characteristics can be studied using just this form for the error with no loss in generality. To find out how error varies in steps of time, substitute the relation above into:

$$\varepsilon_j^{n+1}=\varepsilon_j^n+r\left(\varepsilon_{j+1}^n-2\varepsilon_j^n+\varepsilon_{j-1}^n\right)\qquad(14.1.1679)$$

after noting that:

$$\begin{aligned}\varepsilon_j^n&=e^{\mathrm{i}\lambda x}e^{-\lambda Dt}\\ \varepsilon_j^{n+1}&=e^{\mathrm{i}\lambda x}e^{-\lambda D(t+\Delta t)}\\ \varepsilon_{j+1}^n&=e^{\mathrm{i}\lambda(x+\Delta x)}e^{-\lambda Dt}\\ \varepsilon_{j-1}^n&=e^{\mathrm{i}\lambda(x-\Delta x)}e^{-\lambda Dt}\end{aligned}\qquad(14.1.1680)$$

to yield (after simplification):

$$e^{-\lambda D\Delta t}=1+\frac{\alpha\Delta t}{\Delta x^2}\left(e^{\mathrm{i}\lambda\Delta x}+e^{-\mathrm{i}\lambda\Delta x}-2\right)\qquad(14.1.1681)$$

Using the identities (see section Trigonometry page 2532):

$$\cos(\lambda\Delta x)=\frac{e^{\mathrm{i}\lambda\Delta x}+e^{-\mathrm{i}\lambda\Delta x}}{2}\qquad\sin^2\left(\frac{\lambda\Delta x}{2}\right)=\frac{1-\cos(\lambda\Delta x)}{2}\qquad(14.1.1682)$$

Then the prior previous relation can the be written as:

$$e^{-\lambda D\Delta t}=1-\frac{4\alpha\Delta t}{\Delta x^2}\sin^2\left(\frac{\lambda\Delta x}{2}\right)\qquad(14.1.1683)$$

Let us now define the "amplification factor":

$$G = \frac{\varepsilon_j^{n+1}}{\varepsilon_j^n} \qquad (14.1.1684)$$

The necessary and sufficient condition for the error to remain bounded is that $|G| < 1$. However in our case:

$$\left|e^{-\lambda D\Delta t}\right| < 1 \Leftrightarrow \left|1 - \frac{4\alpha\Delta t}{\Delta x^2}\sin^2\left(\frac{\lambda\Delta x}{2}\right)\right| < 1 \qquad (14.1.1685)$$

this is the explicit condition for stability of the numerical scheme.

Note that the term:

$$\frac{4\alpha\Delta t}{\Delta x^2}\sin^2\left(\frac{\lambda\Delta x}{2}\right) \qquad (14.1.1686)$$

is always positive. Thus, to satisfy the prior previous relation we have:

$$\frac{4\alpha\Delta t}{\Delta x^2}\sin^2\left(\frac{\lambda\Delta x}{2}\right) < 2 \qquad (14.1.1687)$$

For the above condition to hold at all:

$$\sin^2\left(\frac{\lambda\Delta x}{2}\right) \qquad (14.1.1688)$$

we have:

$$\boxed{\frac{\alpha\Delta t}{\Delta x^2} < \frac{1}{2}} \qquad (14.1.1689)$$

It gives the stability requirement for the FTCS scheme as applied to one-dimensional heat equation. This is exactly the value we found when playing with the Microsoft® Excel® worksheet.

14.1.19.2 Space-time F.D.M (finite-volume method)

The finite-volume method (FVM) is a method for representing and evaluating partial differential equations in the form of algebraic equations. Similar to the finite difference method or finite element method, values are calculated at discrete places on a meshed geometry.

"Finite volume" refers to the small volume surrounding each node point on a mesh. In the finite volume method, volume integrals in a partial differential equation that contain a divergence term are converted to surface integrals, using the divergence theorem. These terms are then evaluated as fluxes at the surfaces of each finite volume. Because the flux entering a given volume is identical to that leaving the adjacent volume, these methods are conservative. Another advantage of the finite volume method is that it is easily formulated to allow for unstructured meshes. The method is used in many computational fluid dynamics computing packages as illustrated below:

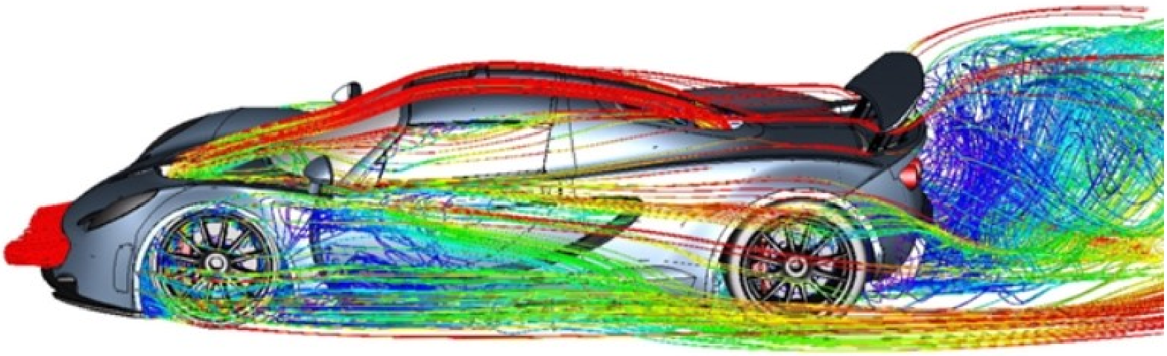


Figure 14.106: Space-time FDM for car C_x study

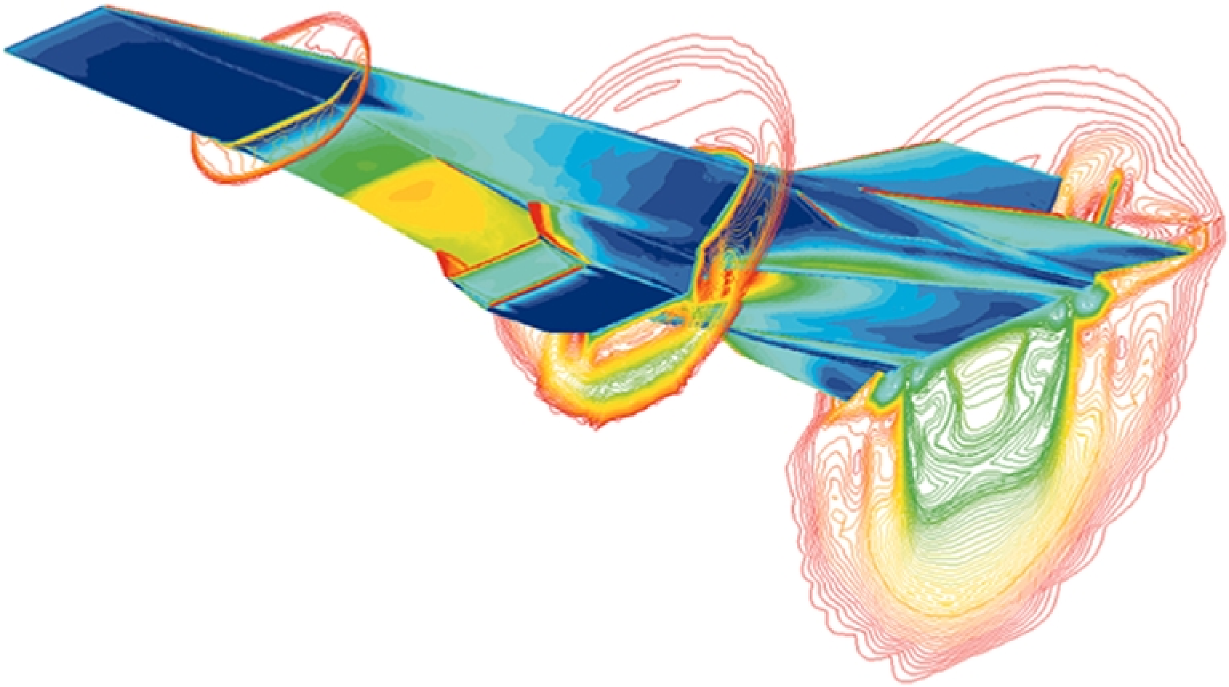


Figure 14.107: Space-time FDM for airplane wing profile study

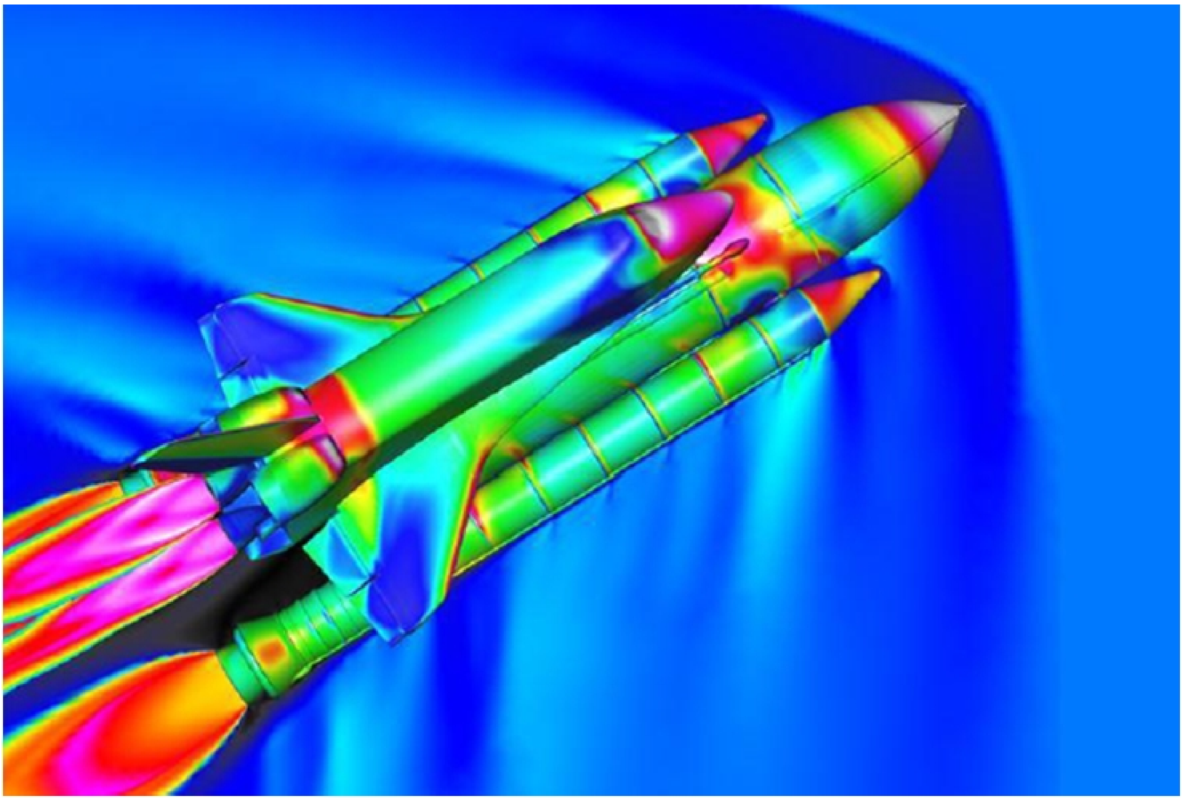


Figure 14.108: **Space-time FDM for NASA space shuttle launch (source: NASA)**

The F.D.M. is a therefore a veeeeery important numerical method in practice because it also gives the possibility to solve directly Maxwell's equations in the time domain and space (and also General Relativity situations). It is then the classified in the 3D (three-dimensional quantification of space) and temporal computational methods and finds its main industrial applications in the fields of design (antennas and circuits), of electromagnetic compatibility, of the diffraction and of the propagation and electromagnetic dosimetry (living beings and waves interactions).

We will discuss now the basics of the concept in a special case as in practice, programming the F.D.M. is a whole team job in itself (like the rest of this textbook obviously but sometimes it is useful to recall that). Indeed a loot of problems must be resolved when dealing with computer programs using F.D.M. (convergence criteria, meshing methods, boundary conditions, user input, programming language methods, etc.).

In a electrodynamics problem treated by F.D.M., the first necessary step is to define the volume V of the space and the time interval $I=[0,T]$ for which the resolution is desired (it is unrealistic at this day to hope to solve Maxwell's equations for an infinite space and for an unlimited period of time!). The volume of calculation contains the object (antenna circuit, ...) that it is desired to characterize, in response to a given excitation. Secondly, the space (meshing of V) and time should be discretized to allow a numerical implementation of the resolution (and in reality the meshing is not uniform...). The problem then becomes the one of determining the field at any point of the mesh for any discrete moment of the observation time interval. The spatial and temporal discretization will be specified in what will follow below and will naturally come from physics equations to solve. They obviously condition both the accuracy of the calculation results and the computing resources required to carry it out.

The structuring of the F.D.M. mesh and the resolution method directly result of the equations to solve.

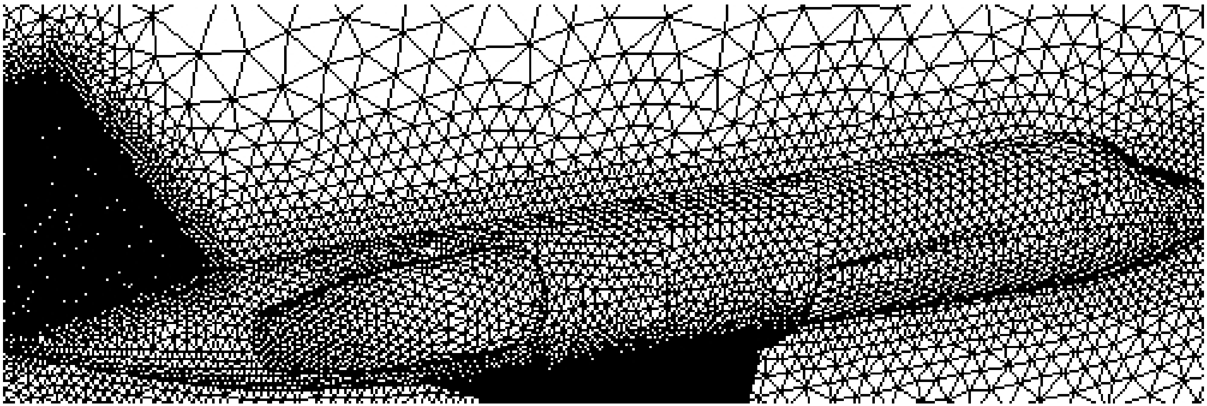


Figure 14.109: **Airplane typical FDM meshing**

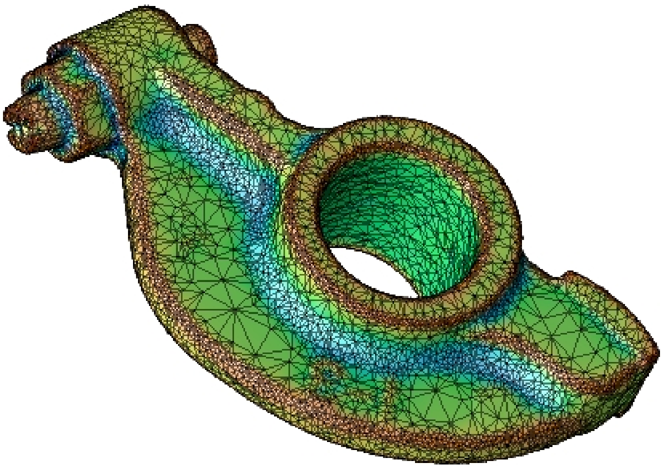


Figure 14.110: **Mechanical element FDM meshing**

In a linear, homogeneous, isotropic, non-dispersive and non-magnetic (...) material, the Maxwell equations will be written explicitly based on the third Maxwell's equation (see section Electrodynamics page 4101):

$$-\frac{\partial\vec{B}}{\partial t}=\vec{\nabla}\times\vec{E} \quad (14.1.1690)$$

Either explicitly with the negative sign put at the other side of the equality:

$$\begin{aligned}\frac{\partial B_x}{\partial t}&=-(\partial_y E_z-\partial_z E_y)\\ \frac{\partial B_y}{\partial t}&=-(\partial_z E_x-\partial_x E_z)\\ \frac{\partial B_z}{\partial t}&=-(\partial_x E_y-\partial_y E_x)\end{aligned} \quad (14.1.1691)$$

And we will also use the fourth Maxwell equation without sources:

$$\vec{\nabla}\times\vec{B}=\mu\varepsilon\frac{\partial\vec{E}}{\partial t} \quad (14.1.1692)$$

Thus explicitly and rearranged:

$$\begin{aligned}\frac{\partial E_x}{\partial t}&=\frac{1}{\mu\varepsilon}(\partial_y B_z-\partial_z B_y)\\ \frac{\partial E_y}{\partial t}&=\frac{1}{\mu\varepsilon}(\partial_z B_x-\partial_x B_z)\\ \frac{\partial E_z}{\partial t}&=\frac{1}{\mu\varepsilon}(\partial_x B_y-\partial_y B_x)\end{aligned} \tag{14.1.1693}$$

That is to say for summary:

$$\begin{aligned}&\begin{cases}\dfrac{\partial}{\partial t}B_x(\vec{r},t)=-(\partial_y E_z(\vec{r},t)-\partial_z E_y(\vec{r},t))\\ \dfrac{\partial}{\partial t}B_y(\vec{r},t)=-(\partial_z E_x(\vec{r},t)-\partial_x E_z(\vec{r},t))\\ \dfrac{\partial}{\partial t}B_z=-(\partial_x E_y(\vec{r},t)-\partial_y E_x(\vec{r},t))\end{cases}\\ &\begin{cases}\dfrac{\partial}{\partial t}E_x(\vec{r},t)=\dfrac{1}{\mu\varepsilon}(\partial_y B_z(\vec{r},t)-\partial_z B_y(\vec{r},t))\\ \dfrac{\partial}{\partial t}E_y(\vec{r},t)=\dfrac{1}{\mu\varepsilon}(\partial_z B_x(\vec{r},t)-\partial_x B_z(\vec{r},t))\\ \dfrac{\partial}{\partial t}E_z(\vec{r},t)=\dfrac{1}{\mu\varepsilon}(\partial_x B_y(\vec{r},t)-\partial_y B_x(\vec{r},t))\end{cases}\end{aligned} \tag{14.1.1694}$$

In what follows, we will only concerned with the first equation, the other leading to similar developments.

To allow a computer processing, the various derivatives present in the equation must be approximated numerically as we already know. To do this, we use the principle of centered finite difference which is based on the following Taylor series expansions for recall (see section Sequences and Series page 1696):

$$f(x)=\sum_{k=0}^{n}\frac{1}{k!}\frac{\mathrm{d}^k f(x_0)}{\mathrm{d}x^k}(x-x_0)^k=\sum_{k=0}^{n}\frac{\mathrm{d}^{(k)}f(x_0)}{k!}(x-x_0)^k \tag{14.1.1695}$$

We then have on the basis of this principle:

$$\begin{aligned}f\left(x_0+\frac{\mathrm{d}_x}{2}\right)&=f(x_0)+\frac{\mathrm{d}_x}{2}f'(x_0)+\frac{1}{2}\left(\frac{\mathrm{d}_x}{2}\right)^2 f''(x_0)+\ldots\\ f\left(x_0-\frac{\mathrm{d}_x}{2}\right)&=f(x_0)-\frac{\mathrm{d}_x}{2}f'(x_0)+\frac{1}{2}\left(\frac{\mathrm{d}_x}{2}\right)^2 f''(x_0)-\ldots\end{aligned} \tag{14.1.1696}$$

If we neglect the terms of the second order, it comes by subtracting the two series:

$$f'(x_0)=\frac{f\left(x_0+\dfrac{\mathrm{d}_x}{2}\right)-f\left(x_0-\dfrac{\mathrm{d}_x}{2}\right)}{\mathrm{d}_x}+\varepsilon(\mathrm{d}_x^2) \tag{14.1.1697}$$

where $\varepsilon(\mathrm{d}_x^2)$ is an error or order 2, neglected thereafter (we notice that this is the centering that, allowing compensation of second derivatives, minimizes the error in the approximation).

Applying this principle to temporal and spatial derivatives of:

$$\begin{aligned}\frac{\partial}{\partial t}B_x(\vec{r},t)&=-(\partial_y E_z(\vec{r},t)-\partial_z E_y(\vec{r},t))\\ \frac{\partial}{\partial t}B_y(\vec{r},t)&=-(\partial_z E_x(\vec{r},t)-\partial_x E_z(\vec{r},t))\\ \frac{\partial}{\partial t}B_z(\vec{r},t)&=-(\partial_x E_y(\vec{r},t)-\partial_y E_x(\vec{r},t))\end{aligned} \tag{14.1.1698}$$

it comes:

$$\begin{aligned}&\frac{B_x\left(x,y,z,t+\dfrac{\mathrm{d}_t}{2}\right)-B_x\left(x,y,z,t-\dfrac{\mathrm{d}_t}{2}\right)}{\mathrm{d}_t}\\ &=-\left(\frac{E_z\left(x,y+\dfrac{\mathrm{d}_y}{2},z,t\right)-E_z\left(x,y-\dfrac{\mathrm{d}_y}{2},z,t\right)}{\mathrm{d}_y}-\frac{E_y\left(x,y,z+\dfrac{\mathrm{d}_z}{2},t\right)-E_y\left(x,y,z-\dfrac{\mathrm{d}_z}{2},t\right)}{\mathrm{d}_z}\right)\end{aligned} \tag{14.1.1699}$$

or after rearrangement:

$$\begin{aligned}&B_x\left(x,y,z,t+\frac{\mathrm{d}_t}{2}\right)=B_x\left(x,y,z,t-\frac{\mathrm{d}_t}{2}\right)\\ &+\mathrm{d}_t\left(\frac{E_y\left(x,y,z+\dfrac{\mathrm{d}_z}{2},t\right)-E_y\left(x,y,z-\dfrac{\mathrm{d}_z}{2},t\right)}{\mathrm{d}_z}-\frac{E_z\left(x,y+\dfrac{\mathrm{d}_y}{2},z,t\right)-E_z\left(x,y-\dfrac{\mathrm{d}_y}{2},z,t\right)}{\mathrm{d}_y}\right)\end{aligned} \tag{14.1.1700}$$

This relation shows that if we know the components E_y, E_z of electric field at time t and the component B_x of the magnetic field at the earlier time $t-\mathrm{d}_t/2$, it is possible to determine B_x at the time $t+\mathrm{d}_t/2$. Obviously the process is exactly the same for all other components and shows the same time lag. This result suggests using an iterative numerical solution, in which the electric and magnetic fields are evaluated alternately, respectively at the discrete time $n\mathrm{d}_t$ and $(n+1/2)\mathrm{d}_t$, d_t being the time step (denoted Δt by the computer scientists). It is customary in the literature to denote by $B_x^{n+\frac{1}{2}}$ the component of the magnetic field at the time $(n+1/2)\mathrm{d}_t$.

The same analysis applies for the spatial distribution of the field on the observed points. Thus, evaluating B_x at the point (x,y,z) is based on the knowledge of E_y at the points $(x,y,z+\mathrm{d}_z/2)$ and $(x,y,z-\mathrm{d}_z/2)$ and of E_z at the points $(x,y+\mathrm{d}_y/2,z)$ and $(x,y-\mathrm{d}_y/2,z)$.

So we can summarize this geometrically in the following figure named "Yee cell":

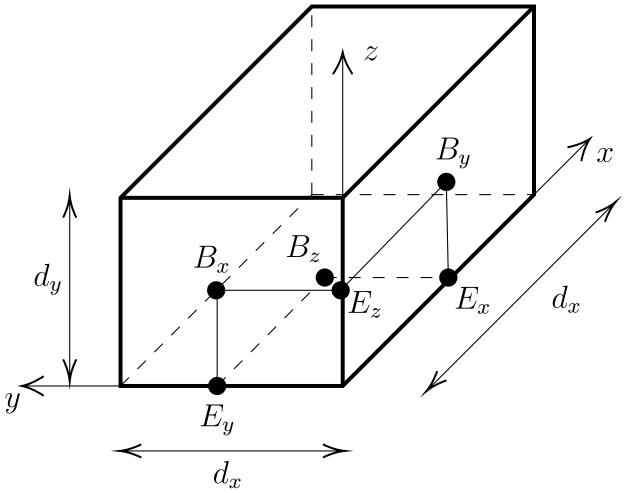


Figure 14.111: **Generic Yee cell in parallelepiped mesh element**

The electric field components are evaluated at the centers of the edges of the mesh and the components of the magnetic field at the centers of the faces so as to ensure the alternation imposed by the equations (as mentioned previously we name "Yee cell" the unit cell with this distribution of points).

In the special case of a propagating electromagnetic wave, the $\vec{E}$ and $\vec{B}$ fields are always perpendicular (see section Electrodynamics page 4138) in a homogeneous, linear, anisotropic medium (this type of media includes many things like air, water, glass without stress or tempering) and when the engineers prefers to represent the "real" magnetic field $\vec{H}$ instead of the magnetic excitation $\vec{H}$ we can find in the literature the following parallelepiped Yee cell figure:

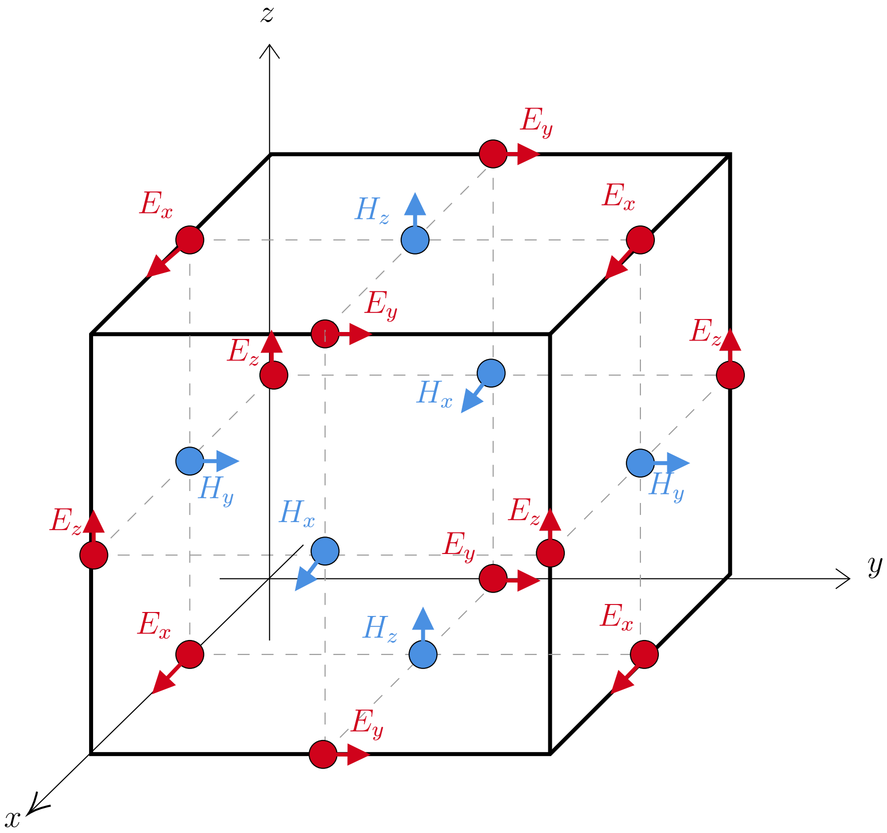


Figure 14.112: Yee cell in parallelepiped mesh element for EM wave propagating in vacuum

So globally in vacuum without the presence of any object we have a meshing of the space that can be represented as follows:

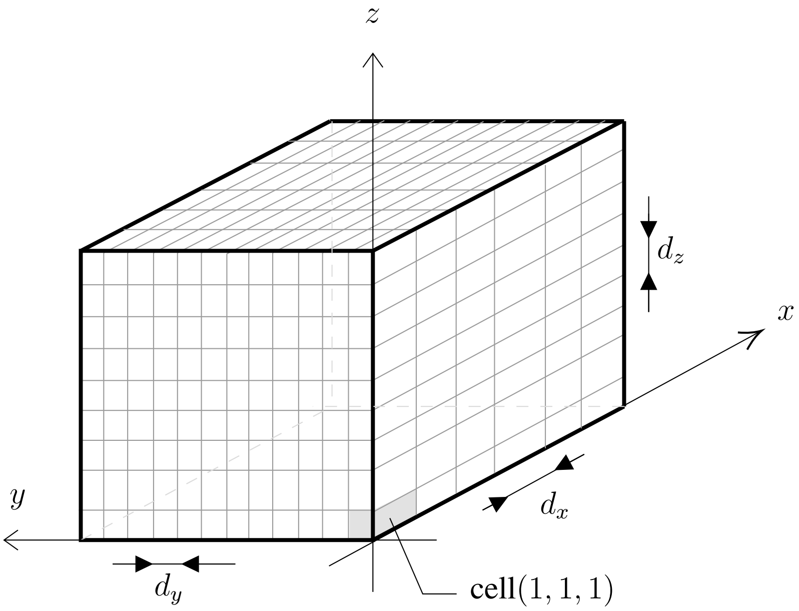


Figure 14.113: Typical simple meshing with multiple Yee cells

Obviously, we can perform the calculation routines with a value of the permittivity and permeability that are not necessarily equal in all the cells. Allowing in addition to model the propagation of electromagnetic waves in heterogeneous and non isotropic media.

Finally, it is important to notice that the spatial and temporal mesh step must be configured by the user running such simulations. This for computing resources reasons as well as accuracy goals. Indeed, we do not do the same simulations for a multi-physics system in the low frequency in vacuum than for non-isotropic material at high frequency and, either on a desktop computer or on a supercomputer.

14.1.20 Data Science & Data Mining

Definition (#14.1.567): "Data science" is an interdisciplinary business field that uses statistics, computer science, scientific methods, processes, algorithms, domain knowledge and systems to extract or extrapolate knowledge and insights from various sources of various types of datas.

Definition (#14.1.568): "Data Mining", also popularly referred to as "knowledge discovery from data" (KDD), is a subset of data science that focuses specifically on extracting patterns and knowledge from large datasets. It applies statistical and machine learning[39] algorithms to identify hidden relationships, trends, and anomalies in data. Data mining is often used for tasks such as customer segmentation, fraud detection, and predictive analytics.

Remark

Some statisticians argue that data science is not a new field, but rather another name for statistics. Others argue that data science is distinct from statistics because it focuses on problems and techniques unique to digital data. Some others argue that statistics emphasizes quantitative data and description. In contrast, data science deals with quantitative and qualitative data (e.g., from images, text, sensors, transactions, customer information, etc.) and emphasizes prediction and action. Some describe statistics as a non-essential part of data science and others that data science is not distinguished from statistics by the size of datasets or use of computing and that many graduate programs misleadingly advertise their analytics and statistics training as the essence of a data-science program. In this textbook we will consider Data Science as the set of Data investigations that don't use inferential statistics (null hypothesis testing).

quote 14.1:6768.#77

> *If you want to make big improvements in communication, my advice is—hire physicists, not communications people from normal companies...*
>
> Dominic Cummings

The following naive figure may help the reader to contextualize these both fields:

---

[39] As said by Arthur Samuel, a computer program is said to learn (i.e. machine learning) from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.

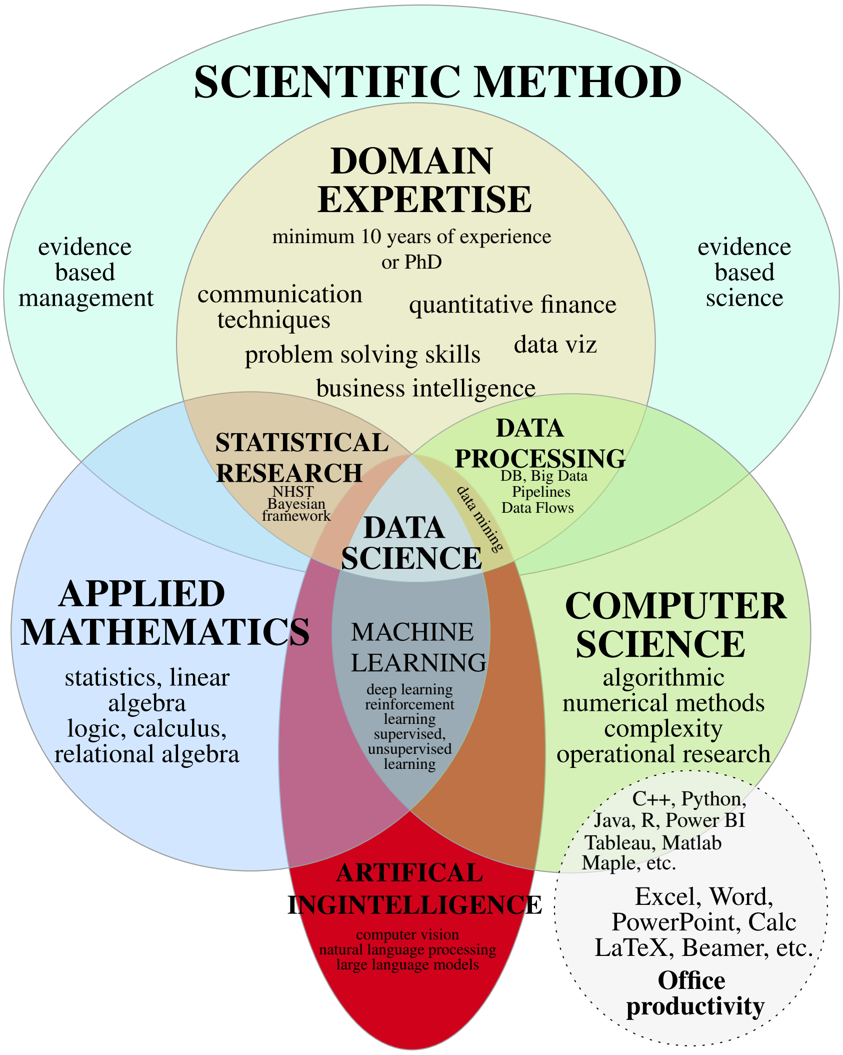

SCIENTIFIC METHOD
DOMAIN EXPERTISE
minimum 10 years of experience or PhD
communication techniques
quantitative finance
problem solving skills
data viz
business intelligence
evidence based management
evidence based science
STATISTICAL RESEARCH
NHST
Bayesian framework
DATA PROCESSING
DB, Big Data
Pipelines
Data Flows
data mining
DATA SCIENCE
APPLIED MATHEMATICS
statistics, linear algebra
logic, calculus,
relational algebra
MACHINE LEARNING
deep learning
reinforcement learning
supervised, unsupervised learning
COMPUTER SCIENCE
algorithmic
numerical methods
complexity
operational research
C++, Python,
Java, R, Power BI
Tableau, Matlab
Maple, etc.
ARTIFICAL INGINTELLIGENCE
computer vision
natural language processing
large language models
Excel, Word,
PowerPoint, Calc
LaTeX, Beamer, etc.
Office productivity

Figure 14.114: Oversimplified overview of Data Science

The main purpose of Data Mining is to ascend the following analytics levels:

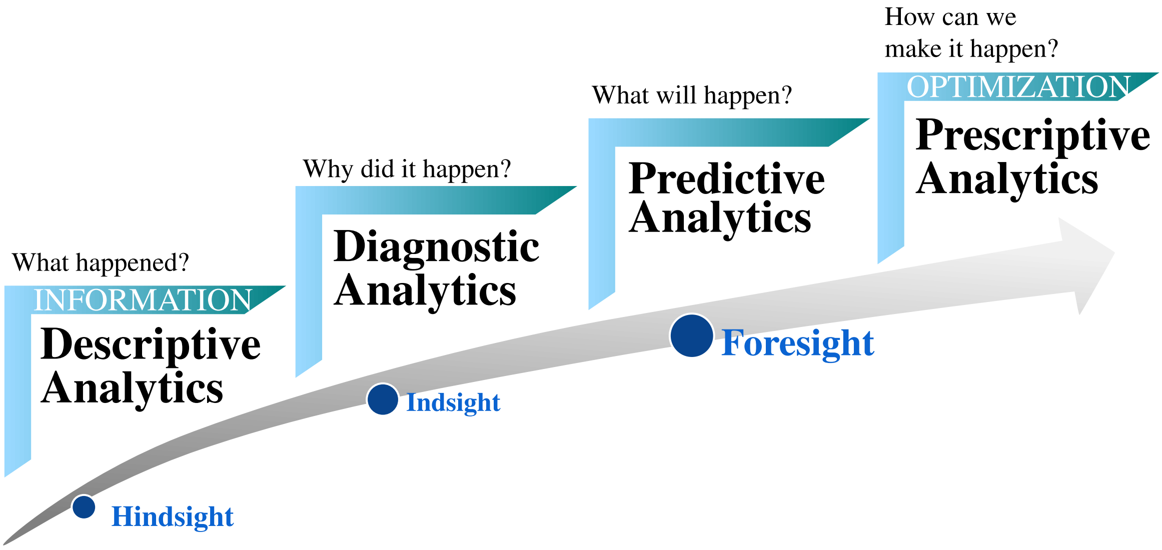

What happened?
INFORMATION
Descriptive Analytics
Why did it happen?
Diagnostic Analytics
What will happen?
Predictive Analytics
How can we make it happen?
OPTIMIZATION
Prescriptive Analytics
Hindsight
Indsight
Foresight

Figure 14.115: Descriptive, Diagnostic, Predictive and Prospective analytics

Or an extended and detailed version of the above figure but instead as a table and with a slightly different vocabulary:

Analytics Steps	Typical Questions Addressed
Descriptive analytics	•What is the current situation? What's happening? •What has change? What's new? •What should we focus on? What should we worry about? •Is this the right data to use for making the prediction? •How reliable is the data we are using here?
Causal analytics	•Diagnosis, explanation, and attribution: What explains the current situation? •What can we do about it? •How would different actions change the probabilities of different future outcomes? •What decisions will this be used to make? •What's the cost of getting those decisions wrong? •How much will decisions affect the targeted people (especially the vulnerable one!)?
Predictive analytics	•If we do not change what we are doing, what will (probably) happen next? When? How likely are the different possibilities? •Given observed (or assumed) values for some variables, what are the probabilities for values of other variables? •How well can some variables be predicted from others? •How well can future outcomes be predicted now? •Are we certain that's what we want to predict? •Are we really aware of the limits of the predictions?
Prescriptive analytics	•What should we do next? •What decisions and policies implemented now will most improve probabilities of future outcomes?
Evaluation analytics	•How much will it cost to deploy and maintain the models? •How well are our current decisions and policies working? •What effects have our decisions and policies actually caused? •How do different policies affect behaviour and outcomes for different people?
Learning analytics	•What decisions or policies might work better than our current ones? •How can we use data and experimentation to find out? •By how much do different items of information improve decisions? •What is the value of information for different measurements?
Collaborative analytics	•How can we best work together to improve probabilities of future outcomes? •Who should share what information with whom, how and when? •What actions should each division of an organization or each member of a team take?

Table 14.23: **Component of risk analytics**

A second level of granularity, with their relations but poorly readable and not error proof (...), would be (we will come back on a more precise definition further below):

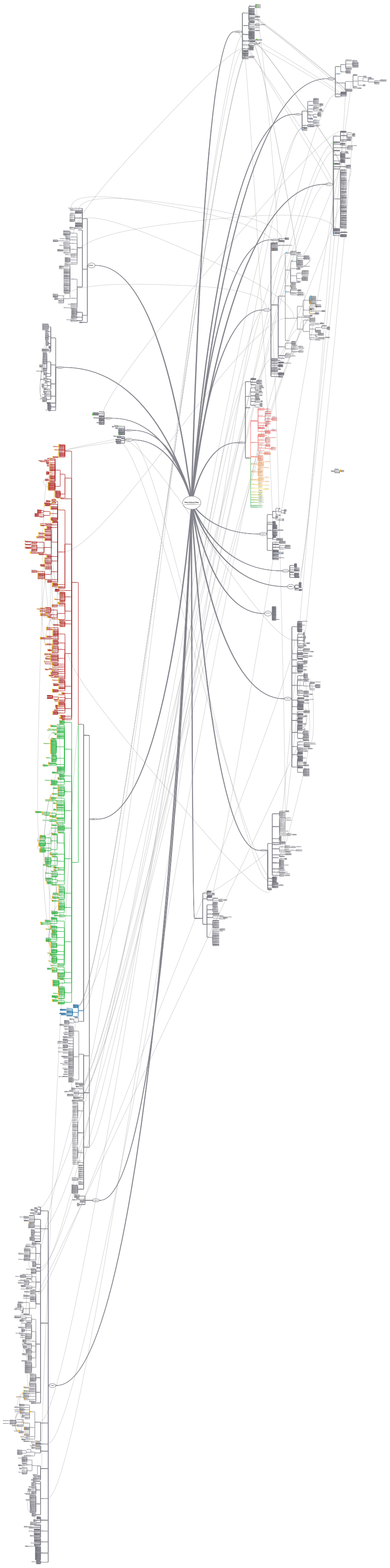


Figure 14.116: **Data Mining/Machine Learning non-exhaustive Mind Map techniques and tools**

and a third level of granularity would give[40] (sorry it's quite long but "data scientists", managers and IT staff in my teachings ask me many times to have an exhaustive one-place list):

- **Variables transformations:**
  1. Percentage ("normalization")
  2. Logarithmic transformation
  3. Center transformation (i.e. mean normalization)
  4. Z-transformation (centered-reduced / standardized)
  5. Rank transformation
  6. Percentile transformation (scaling)
  7. Range transformation (i.e. Min-Max transformation)
  8. Robust scaling
  9. Box-Cox transformation
  10. Modified Box-Cox transformation
  11. Johnson transformation (Yeo-Johnson)
  12. Variance stabilizing transformation
  13. Fisher transformation
  14. Lambert W x F transformation
  15. ORQ normalization
  16. Manly's exponential transformation
  17. John/Draper's modulus
  18. Bickel/Doksum's modified Box-Cox
  19. Whitening
  20. ...

- **Distance metrics:**
  1. Euclidean distance (special case of Hölder's distance)
  2. Mahalanobis distance
  3. Manhattan distance
  4. Canberra distance
  5. Chebyshev distance
  6. Earth mover's distance
  7. Hölder distance
  8. Cosine similarity
  9. Minkowski distance
  10. Haversine distance
  11. Hamming distance
  12. Jaccard distance
  13. Chi-squared distance
  14. Kendall rank correlation distance
  15. Heterogeneous euclidean overlap Metric (HEOM)
  16. ...

- **Sampling and validation techniques:**
  1. Simple random sampling with/without replacement (SRWR)
  2. Cluster sampling
  3. Systematic sampling
  4. Stratified random sampling
  5. Bootstrap sampling
  6. Conveniance sampling
  7. Judgemental or purposive sampling
  8. Snowball sampling
  9. Quota sampling
  10. Jackknife sampling
  11. Latin Hypercube sampling
  12. V-Fold cross-validation
  13. Leave-one-out cross-validation
  14. Monte Carlo cross-validation
  15. Group V-fold cross-validation

---

[40] this list is strongly inspired by Wikipedia, SPSS, SAS, R, Rapid Miner and Tanagra softwares options. For a neural networks list see the figure further below of Fjodor van Veen where 27 different neural networks type are given.

16. Rolling origin forecast resampling
17. Nested or double resampling
18. ...

- **Dimensionality reduction (feature extraction+feature selection):**

1. Correlation coefficients (Pearson's, Kendall, MIC, etc.)
2. Stepwise regression
3. Forward/Backward-logit
4. CFS (Correlation Feature Selection) filtering (Hall & Smith CFS)
5. FCBF filtering (Yiu & Liu Fast Correlation Based Filter)
6. mRMR (Minimum-redundancy-maximum-relevance) feature selection
7. Chi-squared feature ranking
8. Fisher Criterion Scoring
9. Battiti's MIFS (Mutual Information-based Feature Selection Method) feature filtering
10. NIMFS (Normalized mutual information feature) selection
11. MODTree (Multivalued Oblivious Decision) filtering (Lallich & Rakotomalala MODTree)
12. Non-negative matrix factorization dimension reduction
13. ReliefF (Kira & Rendell)
14. Runs filtering
15. Stepdisc (Wilk's partial lambda)
16. Oracle Minimum Description Length (MDL)
17. PCA (Principal Component Analysis)
18. SVD (Singular Value Decomposition)
19. CCA (Curvilinear Component Analysis)
20. ICA (Independent Component Analysis)/Kernel ICA
21. Parallel Factor Analysis (PARAFAC)
22. Tucker Decomposition (PARAFAC generalization)
23. Sammon mapping
24. Kernel PCA / SVD
25. Projection Pursuit (PP)
26. Laplacian eigenmaps
27. Mixture Discriminant Analysis (MDA)
28. Regularized Discriminant Analysis (RDA)
29. Flexible Discriminant Analysis (FDA)
30. Latent Dirichlet allocation
31. Isomap
32. Locally linear embedding
33. Maximum variance unfolding
34. Synthetic Minority Over-sampling Technique (SMOTE)
35. Adaptive Synthetic Sampling Approach (ADASYN)
36. Autoencoder
37. T-SNE (T-distributed stochastic neighbour embedding)
38. Generalized Low Rank Models (GLRM)
39. Multivariate (Soft) Self Modelling Curve Resolution (MCR)
40. Recursive Feature Selection
41. Uniform Manifold Approximation and Projection (UMAP)
42. ...

- **Statistical indicators (parametric and nonparametric[41]):**

1. Mean (arithmetic, geometric, harmonic, weighted, LS...)
2. Bias
3. M-Estimators$^{\text{np}}$ / L-Estimators$^{\text{np}}$ / W-estimators / S-estimators
4. Maximum, Minimum, Range$^{\text{np}}$, Mid-

---

[41] indicators with the exponent $^{\text{np}}$ are nonparametric

hinge, Trimean

5. Median^{np}, Pseudomedian^{np} (Hodges-Lehmann estimator)
6. Median absolute deviation (MAD)
7. Quantiles, Upper and Lower Hinge, Interquartile range^{np}
8. Mode^{np}
9. Variance, semivariance, standard deviation/MSE (biased or unbiased), circular variance
10. Fluctuation interval
11. Skewness (Pearson's, Bowleys's, Kelly's, Groeneveld & Meeden's, Medcouple)
12. Kurtosis
13. Bimodal coefficient
14. Pearson correlation, adjusted correlation, partial correlation, semi-partial correlation, biweight correlation
15. Blomqvist's correlation^{np}
16. Spearman^{np}, Kendall^{np} correlations
17. McFadden, Cox & Snell's, Nagelkerke, Efron, McKelvey & Zavoina, Count, Adjusted count pseudo-correlations[42]
18. Lin' correlation
19. Shepherd's Pi correlation
20. d-variable Hilbert-Schmidt independence criterion (dHSIC)
21. p-value
22. β of a NHST (power of test)
23. Pitman's ARE (asymptotic relative efficiency)
24. Coefficient of variation
25. Effect size (ω^2, partial ω^2,η^2, Cohen's f^2, Cohen's q, Cohen's w, Cohen's h, Glass' Δ, Ψ, Cliff's Δ, risk ratio, odds ratio, Hedges' g, etc.)
26. Cohen's κ
27. Yule Q coefficient
28. Gini impurity index (i.e. Somers' D)
29. Area Under the Curve (AUC)
30. Harrell's C-index
31. F_1 score, F_β score
32. Piotroski score
33. Intraclass correlation coefficient
34. Bangdiwala's B
35. Hartley entropy
36. Rényi entropy
37. Moran's I correlation
38. Fleiss Kappa
39. Pearson's measure of mean square contingency
40. Cramér'V (ϕ_c)
41. Tschuprow's T
42. Scott's π
43. ϕ-coefficient (Matthews correlation coefficient)
44. Point-Biserial correlation coefficient
45. Confusion matrix $\mathcal{C}$
46. Mallow's C_p
47. Akaike information criterion (AIC)
48. Corrected Akaike information criterion (AICc)
49. Bayesian information criterion (BIC)
50. Bayes factor (BF)
51. Deviance information criterion (DIC)
52. Hannan–Quinn information criterion (HQIC)
53. Focused information criterion (FIC)
54. Mutual information criterion (MIC)
55. Bayesian predictive information criterion (BPIC)
56. Covariance inflation criterion (CIC)
57. Risk inflation criterion (RIC)
58. Relative error, Absolute error, Final prediction error (FPE)
59. Widely applicable information criterion (WAIC)
60. Quasilikelihood information criteria (QIC)
61. Shannon entropy (+perplexity)
62. Specificity
63. Log-loss (LL)

---

[42] Pseudo-correlations are used in logistic regression

64. Sensitivity
65. Jensen–Shannon divergence
66. Kullback–Leibler divergence
67. Aggregation indices (econometric)
68. Helsel's coefficient
69. Minimum message length (MML)
70. Freeman's θ
71. ε-squared
72. Pearson contingency coefficient
73. Goodman's Gamma
74. Guttman monotonicity coefficient^{np}
75. Rand index
76. Concentration coefficient
77. Uncertainty coefficient
78. Brier score loss
79. NNS (Nonlinear Nonparametric Statistic) dependence
80. ...

- **Statistical tests[43] (parametric, semi-parametric and nonparametric[44]):**

1. (Brown-)Mood's test^{np}
2. (Siegel-)Tukeys HSD test^{np}
3. Abelson-Tukey score test
4. Ahsanullah test^{np}
5. Adjacency test^{np}
6. ADF-GLS test
7. 2-AFC/3-AFC (alternative force choice) tests
8. Agostino test of skewness
9. Agresti–Pendergrast test^{np}
10. (Hodges)-Ajne's test^{np}
11. Anscombe-Glynn test of kurtosis
12. Anderson-Darling's adequation test^{np}
13. ANCOVA test
14. ANOVA/MANOVA tests (+Welch's ANOVA, Friedman's ANOVA^{np}, Kruskal-Wallis ANOVA^{np}, Scheirer-Ray-Hare ANOVA^{np}, Van der Waerden ANOVA^{np})
15. Ansari-Bradley's test^{np}
16. Anscombe-Glynn test
17. Aroian test
18. (Aspin-)Welch test
19. Auto-Regressive Conditional Heteroskedasticity (ARCH) effect test^{np}
20. Baarda's w-test
21. Bai and Ng's test
22. Banerjee-Lumsdaine-Stock test
23. Baumgartner-Weiss-Schindler test^{np}
24. Baringhaus-Henze-Epps-Pulley (BHEP) test
25. Bartholomew's likelihood ratio test
26. Barnard's test^{np}
27. Bartels rank test^{np} of randomness
28. Bartlett-(Kendall) test for variances
29. Bartlett's test for sphericity
30. Bartlett-Diananda test
31. Bayesian A/B tests
32. Beale's F test
33. Behrens-Fisher test
34. Bentler and Yuan's scree test
35. Beran's tests
36. Begg test
37. Berk-Jones tests^{np}
38. Besag-Newell's R test
39. BDS (Brock-Dechert-Scheinkman) test for non-linear serial dependence in time series
40. Bhapkar's test^{np}
41. Bieren's test
42. Bimodality test
43. (Binomial) sign test^{np}
44. Binomial exact test^{np}
45. Bithell's linear rank score test^{np}
46. Birch test^{np}
47. Bivariate sign test^{np}
48. Blumen's test^{np}
49. Blum-Kiefer-Rosenblatt independence test
50. BMP (Boehmer, Musumeci and Poulsen) cross-sectional test
51. Bonett-Seier test of kurtosis
52. Bonferroni outlier test
53. Boschloo's test^{np}

---

[43] Some of these tests are proved in the section Statistics page 542. Some items in the lists below are not really tests but more "procedures" or "methods". And obviously some tests are just special cases of others tests!!!

[44] Tests with the exponent ^{sp} are semi-parametric, those with ^{np} are nonparametric

54. Bowker's test for symmetry^{np}
55. Bowman and Shenton combination tests
56. Box's (M) test
57. Box–Pierce test
58. Brant test^{np} (proportional odds test)
59. Breitung test^{np}
60. Breslow-Day test^{np}
61. Breusch-Pagan-Godfrey's test for homogeneity of variances
62. Breusch-Godfrey stationarity test
63. Breusch-Pagan stationarity test
64. Brown–Forsythe's test
65. Brunk's^{np} test
66. Brunner-Munzel generalized Wilcoxon test^{np}
67. Buishand range and U test^{np}
68. Busetti-Taylor test
69. Buy-and-hold abnormal return test (BHAR)
70. Butler-Smirnov test
71. Canova and Hansen test^{np}
72. Capon test^{np}
73. CAR (cumulative abnormal return) test
74. Cattell's scree test
75. Chacko's test^{np}
76. Change point mean stationarity test
77. Change point variance changes stationarity test
78. Chen-Deo test
79. Chen-Wolfe test^{np}
80. Cheng-Spiring test^{np}
81. Chi-squared test for association / independence or homogeneity^{np}
82. Chi-squared test for independence with/without Yates correction
83. Chi-squared test for outlier
84. Chi-squared test for adequation^{np}
85. Chow's test^{np}
86. Cho–Gaines's test^{np}
87. Chow-Denning test
88. Chakraborti-Desu test
89. Chakraborti-Desu test^{np}
90. Chebyshev distance test^{np}
91. Circular-cone test
92. Clark-Evans' test
93. Cliff-Ord tests
94. Cochran's C-test
95. Cochran's Q-test^{np}
96. Cochran-Armitage's test^{np} (aka Chi square test for trend)
97. Cochran-Mantel-Haenszel's test^{np}
98. Cohen's d test
99. Conover-Iman squared ranks pairwise comparison test^{np}
100. Corner test for association^{np}
101. Comparing correlation coefficients of overlapping (CCO) samples
102. Corrado rank test^{np}
103. Correlation for categorical variables test
104. Cowan generalized sign test^{np}
105. Cox's F-test^{np}
106. Cox and Oakes test^{np}
107. Cox–Small test
108. Cox-Stuart trend test^{np}
109. Craddock test^{np}
110. Cramer–von Mises' test^{np}
111. Cressie tests^{np}
112. Cross sectional T-test
113. Cusum test for structural change
114. Cuzick–Edwards test
115. Cuzick's trend test^{np}
116. d-variable Hilbert Schmidt independence criterion test^{np}
117. Daniel's test for trend^{np}
118. Darroch interaction test^{np}
119. Davidson-MacKinnon test for comparing nonnested models
120. David-Barton test
121. David-Hellwig test
122. Davis-Quade's test^{np}
123. Demsar test^{np}
124. Depaulis test^{np}
125. Deshpande test^{np}
126. Dickey–Fuller's test^{np}
127. Difference between two non-overlapping dependent correlation coefficients test
128. Diggle and Chetwynd's test
129. Dixon's test
130. Doornik-Hansen test
131. Duckworth's test^{np}
132. Dudewicz-van der Meulen test^{np}
133. Duncan's multiple range comparison (MRT) test
134. Dunn's test^{np}

135. Dunnett's T, C or $T3$ test
136. Duo-Trio test
137. Durbin's rank test^{np}
138. Durbin-Watson autocorrelation test
139. Durbin-Wu-Hausman test
140. Dwass-Steele-Critchlow paired test^{np}
141. Dzhaparidze-Nikulin test^{np}
142. Egger's test
143. Euclidean distance test^{np}
144. Elliott-Rothenberg stock test
145. Enders sequential test procedure for unit roots
146. Engelman-Hartigan test
147. Engle test
148. Engle-Granger cointegration test
149. Epstein's test
150. Epps and Pulley test
151. Epps-Singleton (ES) test
152. Epsilon squared test
153. Eta squared test
154. Ewens-Watterson test^{np}
155. Excess mass test^{np}
156. Extreme rank sum test^{np}
157. Fagerland-Hosmer-Bofin test^{np}
158. Fay tests^{np}
159. Ferguson test
160. Fieller's test (A/B test)
161. Filliben's probability plot correlation test
162. Finkelstein and Schoenfeld test^{np}
163. Fisher-ADF test^{np}
164. Fisher-PP^{np}
165. Fisher-Behrens test
166. Fisher's Exact (Fisher-Irwin) test^{np}
167. Fisher-Freeman-Halton Exact test^{np}
168. Fisher's LSD (Least Significant Difference) test
169. Fisher-Hayter's LSD test
170. Fisher-Pitman's permutation test^{np}
171. Fisher's G-test for periodicity
172. Fisher's variance test
173. Fisher's cumulant test
174. Fisher-Yates test
175. Fisher-Yates-Terry-Hoeffding test^{np}
176. Fligner(-Killeen)'s test for homogeneity of variances^{np}
177. Fligner–Wolfe's test^{np}
178. Fligner–Policello's test^{np}
179. Foutz' F_n test
180. Fung-Paul test
181. Freeman–Tukey's test
182. Freund-Ansari-Bradley test for scale^{np}
183. Friedman correlation test^{np}
184. Friedman–Rafsky test^{np}
185. Frosini test
186. Fu and Li's test^{np}
187. G-test
188. Gabriel's pairwise test
189. Galton's rank order test^{np}
190. Gart's test
191. Garud tests^{np}
192. Gel-Gastwirth test
193. Generalized extreme studentized deviate test^{np} (GESD)
194. Generalized rank T-test^{np}
195. Generalized rank Z-test^{np}
196. Generalized sequential probability ratio test^{np} (GSPR)
197. Generalised sign test^{np}
198. Games-Howell test^{np}
199. Geary's test
200. Gehan's generalized Wilcoxon test^{np}
201. Gini test (Gail and Gastwirth)
202. Glejser's test
203. Gnedenko F-test^{np}
204. Goldfeld–Quandt's test
205. Goodman-Kruskal's Gamma test^{np}
206. Goodman-Kruskal's Lambda test^{np}
207. Goodman-Kruskal's tau test^{np}
208. Goodman's interaction test^{np}
209. Gore's test^{np}
210. Grambsch-Therneau test of proportionality
211. Granger causality test
212. Grazzini stationarity test^{np}
213. Greenhouse-Geisser test
214. Greenwood's test^{np}
215. Greenwood-Quesenberry-Miller test^{np}
216. Gregory–Hansen test^{np}
217. Grubbs' test
218. Gupta's test^{np}
219. Hahn-Shapiro test
220. Hadri test
221. Harrington and Fleming's Gp tests
222. Harris-Gnedenko test^{np}
223. Harrison-McCabe test for heteroskedasticity

224. Harrison-Kanji-Gadsden test
225. Hartigan's Dip (HDS) test
226. Hartley's (F-max) test
227. Harbord's test
228. Harvey-Collier test for linearity
229. Hatemi–J test^{np}
230. Haugh's test
231. Hayter-Stone test^{np}
232. Hayter-Nasimoto-Wright test
233. Hegazy-Green test^{np}
234. Henze-Zirkler test
235. Hermans–Rasson test
236. Hettmansperger–McKean test^{np}
237. Hirsch-Slack correlation test^{np}
238. Hochberg's GT2 pairwise test^{np}
239. Hodges' bivariate sign test^{np}
240. Hodges-Lehmann test^{np}
241. Hoeffding's independence test^{np}
242. Hollander's bivariate symmetry test^{np}
243. Hollander's parallelism test^{np}
244. Hollander-Proshan test^{np}
245. Hopkins-Skellam test
246. Hosmer–Lemeshow test^{np}
247. Hosmer-Le Cessie test^{np}
248. Hosking test
249. Hotelling's T2-test
250. Hsieh test^{np}
251. Hsu's MCB (multiple comparisons with the best) test
252. Hudson-Martin-Arguadé test^{np} (HKA)
253. Huynh-Feldt test
254. Imam test^{np}
255. Im, Pesaran and Shin test
256. J-tests
257. Jacquez's k-nearest neighbours test
258. Jalal and Jamshidian test
259. Jarque-Bera test
260. Jennrich test of the equality of two matrices[45]
261. Johansen cointegration test
262. Joenssen's JP-square test^{np}
263. Jonckheere-Terpstra (JT) trend test^{np}
264. Johnson-Mehrotra test^{np}
265. Joint digit Benford test^{np}
266. Judge-Schechter Mean Deviation test^{np}
267. Inannan HCT test^{np}
268. Inner-Wedge test
269. K test (Kleibergen)
270. Kallenberg-Ledwina test^{np}
271. Kao's test^{np}
272. Kattumannil-Sreedevi test^{np}
273. Kaplan-Meier test^{np}
274. Kaiser-Meyer-Olkin test
275. Keenan's test^{np}
276. Kendall rank correlation test^{np}
277. Kendall's tau-b test^{np}
278. Kendall's tau-c test^{np}
279. Kendall's concordance W test^{np}
280. Kendall's tau test^{np}
281. Kenward-Roger tests
282. Kim's ratio test
283. Kimball test^{np}
284. Kimber-Michael test
285. KPSS (Kwiatkowski, Phillips, Schmidt and Shin) test^{np}
286. Klotz scale test^{np}
287. Knox's tests
288. Kochar test^{np}
289. Kochar and Gupta tests^{np}
290. Kolmogorov-Smirnov test^{np}
291. Kowalski's bivariate line test
292. Koziol-Nemec test^{np}
293. Kruskal-Wallace test
294. Kuiper's test^{np}
295. Kurtosis test
296. LaBreque's tests
297. Lagrange multiplier (LM) score test
298. Lanzante's test^{np}
299. Larocque and Labarre sign test^{np}
300. Lawless and Landau recurrence Mean Cumulated Function (MCF) test^{np}
301. Le's test^{np}
302. Lee-Wolfe test^{np}
303. Lehmacher test^{np}
304. Lehmann's test
305. Leybourne-McCabe test
306. Leybourne-Kim-Smith-Newbold test
307. Lepage test
308. Levene's test
309. Liddell's exact test^{np}
310. Li-Mak test

---

[45] Correlation matrices typically!

311. Likelihood ratio test (LRT)
312. Lilliefors test
313. Lim–Wolfe test^{np}
314. Linear Time MMD^{np}
315. Link–Wallace test
316. Lin-Mudholkar test
317. Lin-Wang test
318. Lin-Xu test
319. Lipsitz goodness of fit test^{np}
320. Little's test for data MCAR
321. Liu and Singh rank sum test^{np}
322. Ljung-Box test
323. Locally most powerful rank order test^{np}
324. Locally best invariant (LBI) test
325. Log-rank (Mantel-Cox) test^{np}
326. Locke and Spurrier tests
327. Loh test^{np}
328. Lo-MacKinlay test
329. Lumsdaine and Papell (LP) unit root test
330. Lorenz test^{np}
331. Lu-Smith Normal score test
332. Luukkonen test^{np}
333. Macaskill test
334. Mack–Wolfe's test^{np}
335. Maddala-Wu panel unit root test^{np}
336. Madhava Rao-Raghunath test^{np}
337. Mann-Kendall trend test^{np}
338. Mann-Whitney's test^{np}
339. Mantel-Haenszel log-rank test^{np}
340. Mantel's test
341. Mardia's test of multivariate normality
342. Mardia–Watson–Wheeler's test
343. Martinez-Iglewicz test
344. Mathisen's test^{np}
345. Mauchly's test
346. max-combo test
347. McLeod-Li test
348. McCabe–Tremayne's test
349. McDonald-Kreitman test^{np}
350. McNemar's test^{np}
351. McCulloch test^{np}
352. Mean slippage test (Schwager-Margolin)
353. Median test^{np}
354. Median crossing test^{np}
355. Miller test^{np}
356. Michael's test
357. Mood's scale test^{np}
358. Moon and Perron
359. Mojena's test
360. Moran test
361. Moran's I and II tests^{np}
362. Moses test^{np}
363. Mosteller's k-sample slippage test
364. Mudlarks-McDermott test of ordered variance
365. Mukerjee-Robertson-Wright multiple-contrast test
366. Multi-binomial test
367. Murphy's test
368. Murakami k-Sample BWS normal test^{np}
369. MZ test^{np}
370. Nagarwalla's scan statistic
371. Nearest distance test^{np} (Andrews-Bickel-Hampel-Huber-Rogers-Tukey)
372. Nemenyi-Damico-Wolfe test^{np}
373. (Student)-Newman–Keuls (SNK) test (contested)
374. Neuhauser's test (Murakami's B2 test)
375. Neumann trend test[46]
376. Neyman's smooth goodness of fit test
377. Ng and Perron (NP) tests^{np}
378. Nikulin-Rao-Robson test^{np}
379. Noether's test for cyclical trend
380. Norton's test^{np}
381. O'Brien-Flemings test
382. O'Brien's test for homogeneity of variance^{np}
383. O'Brien's test for association of a single variable with survival^{np}
384. Oja's tests
385. Omega square test
386. Orthogonal T-test
387. Osius-Rojek test^{np}
388. Page's test^{np} (paired and not paired versions)
389. Pardo test^{np}
390. Park test
391. Partial Pearson correlation trend test

---

[46] Also named "adjacency test" or "mean successive difference test".

392. Partial Spearman correlation trend test^{np}
393. Patell test
394. Partial Theil U^{np}
395. Pearson correlation T-test
396. Pearson's chi-squared test
397. Pearson-D'Agostino-Bowman test
398. Pedroni's test
399. Permutation tests
400. Perron-Vogelsang test
401. Pesaran's test
402. Peters test
403. Pettitt's test
404. Peto and Peto test (aka Peto-Prentice test)
405. Phillips-Perron (PP) test^{np}
406. Phillips-Ouliaris test^{np}
407. Pillai's trace test
408. Pitman-Morgan test
409. Pietra test
410. Pocock's test
411. Plackett's interaction test
412. Poisson tests (E-test, C-test)
413. Poisson recurrence Mean Cumulated Function (MCF) test^{np}
414. Pope's Tau-test
415. Pulkstenis-Robinson test^{np}
416. Potthoff and Whitlinghill's test
417. Potthoff's test^{np}
418. Precedence tests^{np}
419. Prentice test^{np}
420. Prentice-Wilcoxon test^{np}
421. Priestley's tests
422. Priestley-Subba Rao (PSR) test
423. Proportional Mass test^{np}
424. Puri's expected Normal scores test^{np}
425. Pustejovsky test
426. p (proportion/binomial)-test^{np}
427. Quade test^{np}
428. Quadratic Time MMD^{np}
429. Positive quadrant dependence test^{np}
430. Quenouille's test
431. Rainbow test for linearity
432. Ramsey Regression Equation Specification Error test (RESET)
433. Rank product test^{np}
434. Ranked von Neumann ratio test^{np}
435. Rao's score test
436. Rao's spacing test of uniformity^{np}
437. Rao's spacing test of homogeneity^{np}
438. Rayleigh's test of uniformity
439. Renyi test^{np}
440. RESET test (REgression Specification Error)
441. Robust rank-order test^{np}
442. Rohlf generalized gap test
443. Rosenbaum's test of uniformity^{np}
444. Rosenbaum's cross matching test^{np}
445. Rosner outlier test
446. Rossberg test^{np}
447. Royston's test
448. Roy's largest root test statistic
449. Roy and Kastenbaum test^{np}
450. Ryan-Einot-Gabriel-Welsch range (REGWQ) test
451. Ryan-Einot-Gabriel-Welsch F (REGWF) test
452. Runs test^{np}
453. Rücker's test^{np}
454. Sandvik-Olsson test
455. Sargan–Hansen test
456. Sarkadi-Kosik test
457. Savage's test
458. Schach's two-sample tests^{np}
459. Scheffé's test
460. Schuster's test^{np}
461. Scott-Knott test
462. Schmidt-Phillips test
463. Schwarzer's test^{np}
464. Seasonal Hybrid ESD test
465. Sen's slope test^{np}
466. Semi-partial correlation tests
467. Sequential test^{np}
468. Serial correlation test^{np}
469. Shapiro-Wilk adequation test
470. Shapiro–Francia test
471. Sherman's test^{np}
472. Shirley-Williams test^{np}
473. Siegel test^{np}
474. Silverman's bandwidth test^{np}
475. Skewness test
476. Skillings-Mack test^{np}
477. Skillings-Wolfe expected Normal scores test^{np}
478. Slatkin's exact test^{np}
479. Slivka's test
480. Smith quartile means test

481. Smith and Jain's test
482. Smith and Taylor ratio test$^{\text{np}}$
483. Spatial scan statistic$^{\text{np}}$
484. Spearman correlation test$^{\text{np}}$
485. Sobel test
486. Sobolev test$^{\text{np}}$
487. Somers's d test$^{\text{np}}$
488. Standard normal homogeneity test (SNHT)
489. Steel test$^{\text{np}}$
490. Stone's test$^{\text{np}}$
491. Stuart–Maxwell test$^{\text{np}}$
492. Stukel test$^{\text{np}}$
493. Subba Rao-Gabr's test$^{\text{np}}$
494. Sukhatme's test$^{\text{np}}$
495. sup-Wald test$^{\text{np}}$, sup-LM test$^{\text{np}}$, sup-LR test$^{\text{np}}$, sup-MZ test$^{\text{np}}$
496. Swartz' entropy test$^{\text{np}}$
497. Tarone's Z test
498. Taha's test$^{\text{np}}$
499. Tajima's D test$^{\text{np}}$
500. Tamhane's T2 test
501. Tamhane-Dunnett test
502. Tango's maximized excess events test$^{\text{np}}$ (MEET)
503. Tango's score test
504. Tarone-Ware test
505. TAR-LR test
506. Taylor robust stationarity test
507. Taylor variance ratio test
508. Tetrad test
509. Tetrachoric correlation coefficient test$^{\text{np}}$
510. Terasvirta neural network test$^{\text{np}}$
511. Theil's U test$^{\text{np}}$
512. Tietjen-Moore test
513. Tiku's test$^{\text{np}}$
514. Toda-Yamamoto test
515. Traditional HEGY (Hylleberg, Engle, Granger and Yoo) test$^{\text{np}}$
516. Triangle test
517. Triples test
518. T-test for coefficient slope
519. T-test (heteroscedastic or homoscedastic version - paired or unpaired version)
520. Trim & Fill test$^{\text{np}}$
521. Trimmed means T-test
522. Tryon T-test
523. Tsay's test$^{\text{np}}$
524. Two out of five test
525. Tsai-Koziol correlation test
526. Tukey's B test
527. Tukey's quick test$^{\text{np}}$
528. Tukey's test for non-additivity
529. Tukey-Kramer test
530. Turning point test$^{\text{np}}$
531. Ury-Wiggins-Hochberg test
532. Vasicek's sample entropy test
533. V-test (modified Rayleigh)
534. Von Neumann ratio test
535. Vuong's test$^{\text{np}}$
536. W/S-test
537. Waerden Normal-Scores test
538. Wald Chi-square test
539. Wald's F-test
540. Wald–Wolfowitz test$^{\text{np}}$ for continuous/dichotomous data
541. Walsh's outlier test$^{\text{np}}$
542. Walker's test$^{\text{np}}$
543. Wall test$^{\text{np}}$
544. Waller-Duncan k-ratio T-test
545. Waller and Lawson's score test
546. Wallis-Moore phase frequency test$^{\text{np}}$
547. Wang and Tsiatis' test$^{\text{np}}$
548. Wavelet stationarity test$^{\text{np}}$
549. Watson's U^2-test$^{\text{np}}$
550. Watson–Williams test$^{\text{np}}$
551. Wei–Lachin test
552. Weisberg–Bingham test
553. Welch(-Satterthwaite) T-test
554. Westerlund cointegration test$^{\text{np}}$
555. Jager-Wellner test$^{\text{np}}$
556. Westenberg's interquartile range test
557. Wheeler-Watson test$^{\text{np}}$
558. White neural network test for non-linearity$^{\text{np}}$
559. White test for homogeneity of variances
560. Whittemore's test
561. Wilcoxon's rank sum test$^{\text{np}}$
562. Wilcoxon's signed rank test$^{\text{np}}$
563. Wilk's multivariate outlier test
564. Wilk's lambda test
565. William's trend test
566. Woolf logit test$^{\text{np}}$
567. Worsley's likelihood ratio test
568. Wright tests
569. Wu test$^{\text{np}}$
570. Wu's H test$^{\text{np}}$

571. Yang test$^{\text{np}}$
572. Yuen's T-test
573. Yuen-Welch T-test
574. Z-test for the arithmetic or geometric mean
575. Z-test for the difference between independent correlations
576. Zelen's test$^{\text{np}}$
577. Zhang's test$^{\text{np}}$
578. Zivot-Andrews stationarity test
579. ...

- **Regression and regularization techniques[47]:**

1. Linear regression model (LSM)
2. Gaussian linear regression model
3. Non-linear regression models with binary or continuous variables
4. Polynomial regression model
5. Weighted least squares regression model
6. B-spline or of collocation polynomial regression
7. Logistic regression models (binomial, multinomial, ordinal)
8. Counting Poisson regression (Poisson MLE, PMLE, GLM, Poisson-quasi-Lindley)
9. Negative binomial (binomial MLE and QGPMLE) regression model
10. Beta regression
11. Cox proportional hazard regression
12. Threshold regression models
13. Total linear regression
14. Orthogonal linear regression model
15. Deming regression
16. Quantile regression model
17. Partial least squares regression (PLS)
18. Structural equation modelling (Two-Stage Least Squares/2SLS)
19. Moderated multiple regression
20. Mixed linear/non-linear or generalized models (nested or not, splitted or not)
21. Panel regression
22. LAD (Least Absolute Deviation) regression
23. LOESS (LOcally Estimated Scatterplot Smoothing) and LOWESS (LOcally Weighted Scatterplot Smoothing)
24. Least absolute shrinkage and selection operator (LASSO)
25. Multivariate adaptive regression splines (MARS)
26. Bayesian linear regression model
27. Logic regression
28. Ridge regression model
29. Bootstrap or Jackknife regression model
30. Backward elimination regressions
31. Forward entry regression
32. C-RT regression tree
33. DfBetas
34. Epsilon SVR (support vector regression)
35. Nu SVR
36. Least-angle regression (LARS)
37. Elastic net regularization
38. Simplex regression
39. Independent component regression
40. Theil-Sen estimator regression method
41. Isotonic regression
42. ...

- **Factor Analysis (FA):**

1. FADM (Mixed data factor analysis)
2. Bootstrap eigenvalues
3. Linear discriminant analysis (canonical LDA)/Kernel LDA
4. Correspondence analysis
5. Discriminant correspondence analysis

---

[47] see section Numerical Methods page 6316

6. Principal Component Analysis with/without Factor rotation (VariMax)
7. Harris component analysis
8. Multiple correspondence analysis
9. NIPALS (Non-linear Iterative Partial Least Squares)
10. Parallel analysis
11. Principal factor analysis
12. ...

- **Clustering (unsupervised):**

1. CatVARCHA (categorical variable hierarchical agglomerative clustering)
2. Clustering Tree (CT) with/without post-pruning
3. Fuzzy clustering tree
4. Kernel Density estimation (density based clustering)
5. Kernel PCA (Principal Component Analysis)
6. Singular Value Decomposition (SVD)
7. EM-Clustering (Expectation-Maximization)
8. HAC (Hierarchical Agglomerative Clustering)
9. k-means, k-Medians or k-Medoids clustering, Fuzzy clustering, k-Modes, k-Prototypes
10. Kernel k-means
11. Mean shift clustering
12. Spectral clustering
13. Multidimensional scaling (MDS)
14. Kohonen-SOM (Self Organization Map)
15. Kohonen-LVQ (Learning Vector Quantizer)
16. GMM (Gaussian Mixture Model)
17. VARCLUS (top down approach)
18. VARHCA (clustering variables using Hierarchical Cluster Analysis)
19. VARKMeans (clustering variable using k-means)
20. Density-based spatial clustering of applications with noise (DBSCAN)
21. Soft independent modelling of class analogies (SIMCA)
22. ...

- **Supervised (Spv) Learning:**

1. Binary logistic regression
2. BVM (Ball Vector Machine)
3. Iterative Dichotomiser 3 (ID3 Quinla algorithm)
4. C4.5 (Quinlan algorithm extension of ID3)
5. M5 (Quinlan algorithm)
6. Cubist (extension of M5)
7. C-PLS (PLS for classification)
8. C-RT (Regression Tree for classification)
9. CS-CRT (Cost Sensitive Classification RT)
10. C-MC4 (M-Estimates based and Laplace smoothed CS-CRT)
11. C-SVC (Continuous Supervised Classification)
12. CVM (Core Vector Machine)
13. Decision List (One-Rule, Zero-Rule, CN2)
14. Repeated Incremental Pruning to Produce Error Reduction (RIPPER)
15. CHAID (Chi-squared Automatic Interaction Detection)
16. Gradient Boosting
17. Gaussian processes
18. k-nn (Nearest Neigbhors)
19. Linear discriminant analysis (canonical LDA)/Kernel LDA
20. QDA (Quadratic Discriminant Analysis)
21. Log-Reg TRIRLS
22. Multilayer perceptron (MLP neural network)

23. Multinomial Logistic Regression
24. Naive bayes categorical variables classification
25. Naive bayes continuous variables classification
26. PLS-DA (PLS Discriminant Analysis)
27. PLS-LDS (PLS Linear DA)
28. Prototype-NN (Nearest Neighbours)
29. Radial basis function (RBF neural network)
30. Random Tree
31. Rule Induction, Fuzzy rule induction
32. Support Vector Machine (SVM)
33. Genetic classification
34. Locally Weighted Learning (LWL)
35. ...

- **Semi-Supervised learning:**
  1. Contrastive Pessimistic Likelihood Estimation (CPLE)
  2. semi-supervised vector machine (S3VM)
  3. Transductive Support Vector Machines (TSVM)
  4. Label propagation
  5. Label spreading
  6. ...
- **Meta Supervised learning (ensemble learning):**
  1. Random forests
  2. Extremely Randomized trees
  3. Rotation forests
  4. Mondrian forests
  5. Gradient Boosting Machines (GMB)
  6. Boosting
  7. Bootstrapped Aggregation (bagging)
  8. CatBoost, AdaBoost, XGBoost
  9. Light GBM
  10. Stacked Generalization (blending)
  11. Gradient Boosted Regression Trees (GBRT)
  12. Arcing (Arc-x4) bagging with weights
  13. Bagging with/without cost sensitivity
  14. Boosting with/without cost sensitivity
  15. MultiCost sensitive supervised learning
  16. Stacked generalization
  17. ...
- **Supervised learning assessment:**
  1. Bias-variance decomposition (Wolpert & Kohavi)
  2. Bootstrap
  3. Cross-validation
  4. Leave-One-Out
  5. Test set assessment
  6. Learning/Train assessment test
  7. ...
- **Scoring:**
  1. Lift curve
  2. ROC curve
  3. Precision-Recall curve
  4. Log-loss (and i.e. log likelihood)
  5. Log-pointwise predictive density (LPPD)
  6. Reliability diagram
  7. ...

- **Association:**

  1. A priori (MR version or not)
  2. A priori PT (Borgelt's algorithm)
  3. Assoc outlier (association rule mining principle)
  4. Frequent itemsets (Borgelt's algorithm)
  5. ECLAT (equivalence class transformation algorithm)
  6. FP-growth (frequent pattern growth)
  7. RElim (recursive elimination)
  8. SaM (Split and Merge)
  9. JIM (Jaccard Itemset Mining)
  10. Spv association rule
  11. Spv association tree
  12. ...

- **Anomaly detection:**

  1. Chauvenet's criterion
  2. Cochran C test
  3. Dixon's test
  4. Grubb test
  5. Cook's distance
  6. Peirce's criterion
  7. Studentized residual
  8. Isolation forest
  9. Control Charts[48] (P, NP, C, U, R-R, X-R, etc.)
  10. Markov modulated Poisson process (MMPP)
  11. Local outlier factor (FOL)
  12. Online Over-Sampling Principal Component Analysis
  13. ...

- **Reinforcement learning:**

  1. Markov Decisions Processes
  2. Case-Based Reasoning
  3. Genetic algorithm
  4. Asynchronous Advantage Actor Critic (A3C)
  5. State–Action–Reward–State–Action (SARSA)
  6. Q-learning
  7. TD(λ) algorithm
  8. TD(0) Actor-Critic
  9. ...

- **Imputations methods[49]:**

  1. Predictive mean matching
  2. Weighted predictive mean matching
  3. Random sample from observed values
  4. Classification and regression trees
  5. Hot or Cold deck imputation
  6. Random forest imputations
  7. Unconditional mean imputation
  8. Bayesian linear regression
  9. Linear regression ignoring model error
  10. Linear regression using bootstrap
  11. Linear regression, predicted values
  12. Imputation of quadratic terms
  13. Laplace smoothing
  14. Random indicator for non-ignorable data
  15. Logistic regression
  16. Logistic regression with bootstrap

---

[48] see section Industrial Engineering (page 8934) for the details on control charts.

[49] Be always careful when using imputation methods. For example impute missing values with mean or zero alters summary statistics significantly, changes the distribution significantly and inflate the presence of a specific value and therefore leads to inaccurate modelling and unlikely conclusions.

17. Proportional odds model
18. Polytomous logistic regression
19. Linear discriminant analysis (LDA)
20. Level-1 normal heteroscedastic
21. Level-1 logistic
22. Level-2 class mean/normal
23. Level-2 class predictive mean matching
24. Multivariate imputation by Chained Equations
25. k-NN imputer
26. ...

- **Forecastings:**
  1. Linear and polynomial regressions (see above)
  2. Moving Average
  3. Simple exponential smoothing
  4. Double exponential smoothing (Brown)
  5. Triple exponential smoothing (Holt & Winters)
  6. ETS (Error Trend Seasonal) models
  7. Logistic forecasts
  8. Croston's method / Adjusted Croston's method
  9. ARIMA[50] processes (AR, ARMA DARMA, INARMA, ARIMA, SARIMA, ARFIMA, SARFIMA, ARIMAX)
  10. GARCH (General-ARCH[51]) processes (GARCH-X, EGARCH, ARCH, sGARCH, RUGARCH, iGARCH, MGARCH)
  11. NARCH (Non-linear-ARCH) processes
  12. TARCH (Threshold-ARCH) processes
  13. Independent Component Analysis (ICA)
  14. ...

- **Sequence mining:**
  1. GSP (Generalized Sequential Patterns)
  2. SPADE (Sequential Pattern Discovery)
  3. FreeSpan
  4. HMM (hidden Markov models)
  5. ...

- **Neural Networks/Deep learning:**
  1. Perceptron
  2. Feed-Forward Neural Network (FFNN)
  3. Radial Basis Network (RBF)
  4. Recurrent Neural Network (RNN)
  5. Gated Recurrent Unit (GRU)
  6. Long Short-Term Memory (LSTM)
  7. Deep Boltzmann Machine (DBM)
  8. Deep Feed Forward Networks (DBN)
  9. Deep Belief Networks (DBN)
  10. Deep Belief Convolutional Network (DCN)
  11. Deep Residual Network (DRN)
  12. Differentiable Neural Computer (DNC)
  13. Neural Turing Machine (NTM)
  14. Capsule Network (CN)
  15. Kohonen Network (KN)
  16. Attention Network (AN)
  17. Deconvolutional Network (DN)
  18. Deep Convolutional Inverse Graphics Network (DCIGN)
  19. Liquid State Machine (LSM)
  20. Extreme Learning Machine (ELM)

---

[50] AutoRegressive Integrated Moving Average
[51] AutoRegressive Conditional Heteroskedasticity

21. Echo State Network (ESN)
22. Convolutional Neural Networks (CNN)
23. Fuzzy Neural Network (FNN)
24. Spectral-Residual CNN (SR-CNN)
25. Super Resolution CNN (SRCNN)
26. Stacked Auto-Encoders (SAE)
27. Variational Auto-Encoder (VAE)
28. Denoising Auto-Encoder (DAE)
29. Sparse Auto-Encoder (DAE)
30. Hopfield Network (HN)
31. Restricted Boltzmann Machine Network (RBM)
32. Adaptative Neuro-Fuzzy Inference System (ANFIS)
33. Generative Adversarial Networks (GAN)
34. Wasserstein Generative Adversarial Networks (WGAN)
35. Spectral Normalization Generative Adversarial Networks (SN-GAN)
36. Self-attention Generative Adversarial Networks (S-GAN)
37. Progressive Generative Adversarial Networks (PROGAN)
38. Contextual RNN-GANs (Context-RNN-GAN)
39. Continuous recurrent neural networks With Generative Adversarial Networks (C-RNN-GAN)
40. Conditional Sequence Generative Adversarial Networks (CS-GAN)
41. Fine-Grained Image Generation through Asymmetric Training (CVAE-GAN)
42. Cycle-consistent Adversarial Networks (CycleGAN)
43. Unsupervised Cross-Domain Image (DTN)
44. Unsupervised Learning With Deep Convolutional Generative Adversarial Networks (DCGAN)
45. Discover Cross-Domain Relations Generative Adversarial Networks (DiscoGAN)
46. Disentangled Representation Learning GAN for pose-Invariant Recognition (DR-GAN)
47. Unsupervised Dual Learning for Translation (DualGAN)
48. Energy-based Generative Adversarial Network (EBCAN)
49. Training Generative Neural Samplers Variational Divergence Minimization (f-GAN)
50. Face Frontalization Generative Adversarial Network (FF-GAN)
51. Generative Adversarial What-Where Network (GAWWN)
52. Learning Object Transfiguration and Attribute Subspacefrom unpaired Data (GeneGAN)
53. Geometric Generative Adversarial Networks (GGAN)
54. Generative Adversarial Networks With Maximum Margin Ranking (Hogan)
55. Towards Realistic Blending (GP-GAN)
56. Introspective Adversarial Networks (IAN)
57. Natural Image Manifold Generative Adversarial Networks (iGAN)
58. Invertible Conditional Generative Adversarial Networks (ICGAN)
59. Image De-raining using a Conditional Generative Adversarial Network (ID-CGAN)
60. Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets (InfoGAN)
61. Location-Aware Generative Adversarial Networkss (LAGAN)
62. Laplacian Pyramid Of Adversarial Networks (LAPGAN)
63. 3D Generative-Adversarial Modelling (3D-GAN)
64. Face Aging With Conditional Generative Adversarial Networks (acGAN)
65. Auxiliary Classifier Generative Adversarial Networks (AC-GAN)
66. Boosting Generative Models (AdaGAN)
67. Autoencoder based Generative Adversarial Networks (AEGAN)

68. Amortised MAP Inference for Image Super-resolution (AffGAN)
69. Attributes and Semantic Layouts Conditional Generative Adversarial Networks (AL-CGAN)
70. Adversarial Learned Inference (ALI)
71. Activation Maximization Generative Adversarial Networks (AM-GAN)
72. Anomaly Detection Generative Adversarial Networks (AnoGAN)
73. Artwork Synthesis Generative Adversarial Networks (ArtGAN)
74. Brainstorming Generative Adversarial Networks (b-GAN)
75. Bayesian GAN (BGAN)
76. Boundary Equilibrium Generative Adversarial Networks (BEGAN)
77. Bidirectional (Generative Adversarial Networks)
78. Boundary-Seeking Generative Adversarial Networks (BS-GAN)
79. Conditional Generative Adversarial Nets (CGAN)
80. Calorimeters With Generative Adversarial Networks (CaloGAN)
81. Context-Conditional Generative Adversarial Networks (CCCAN)
82. Categorical Generative Adversarial Networks (CatGAN)
83. Coupled Generative Adversarial Networks (COGAN)
84. ...

- **Cost functions:**

  1. Quadratic cost
  2. Cross-entropy cost
  3. Exponential cost
  4. Hellinger cost
  5. Kullback-Leibler cost
  6. Generalized Kullback-Leibler cost
  7. Bregman cost
  8. Jensen-Shannon cost
  9. Itakura-Saito cost
  10. ...

- **Calibration methods:**

  1. Histogram binning
  2. Isotonic regression
  3. Bayesian Binning into Quantiles (BBQ)
  4. Platt scaling
  5. Matrix and vector scaling
  6. ...

- **Text Mining:**

  1. Latent Semantic Analysis (LSA)
  2. Latent Semantic Indexing (LSI)
  3. Latent Dirichlet allocation (LDA)
  4. Probabilistic Latent Semantic Indexing (PLSI)
  5. Vector space model (term vector model)
  6. Sentiment Analysis
  7. Language detection
  8. Pattern detection/correlation
  9. ...

Remark

Keep in mind that Machine Learning cannot be used for everything related to statistics. Indeed, in drug assessment and approval we are limited by a set of statistical methods forming an industry standard and approved by regulatory agencies (FDA, EMA, etc.), by the accepted approach named "confirmatory data analysis" via inferential statistics, both frequentist and bayesian, by small and extremely small data (starting from 5 observations in early trial phases) by demand for interpretability (not rarely we prefer here simpler but more "transparent" methods rather than efficient ones). In clinical research (but not only!), every single possible aspect is regulated by guidelines, controlled by validation, every step of statistical analysis must be well enough justified under the threat of being rejected by regulatory or statistical reviewers and lead the whole trial to fail (that is why in practice we have to construct a "Statistical Analysis Plan (SAP)" before any well done experiment).

The reader must also keep in mind some possible limitations of predictive model based on data fitting:

- History cannot always predict future
- The issue of unknown unknowns
- Bad model choice
- Over-fitting model
- Overconfidence of humans on algorithms results
- ...

So we see through these lists that Data Mining and Machine Learning are child of statistics, computer science, and mathematical optimization. Along the way, it took inspiration from information theory, neural science, theoretical physics, and many other fields.

This was the scientist point of view of Data Mining... From the point of view of business, Data Mining is more considered as the following data life cycle:

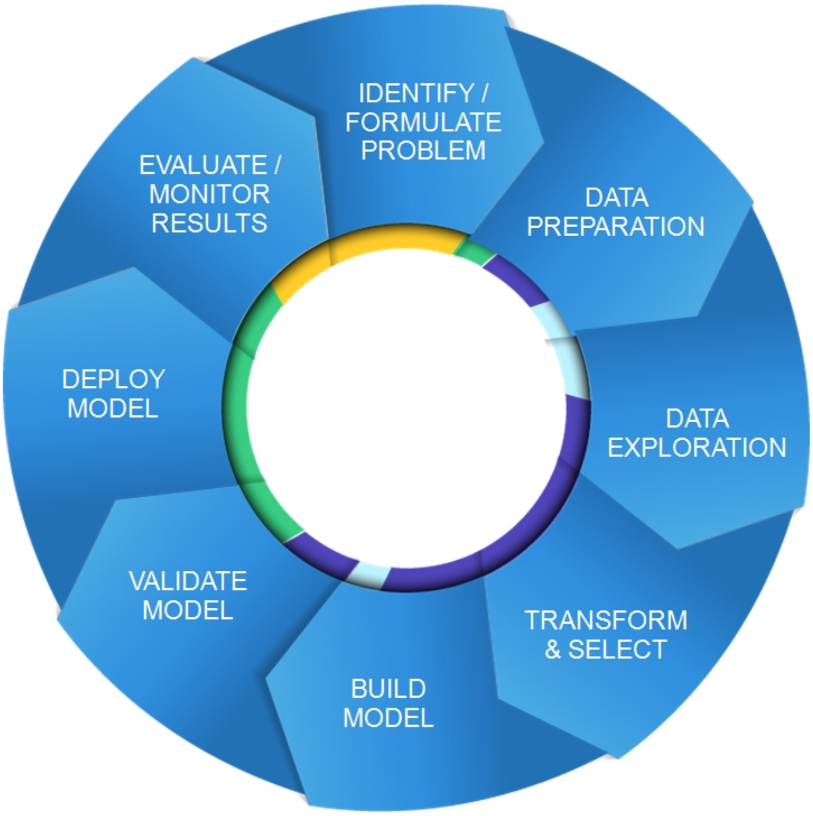

IDENTIFY / FORMULATE PROBLEM
DATA PREPARATION
DATA EXPLORATION
TRANSFORM & SELECT
BUILD MODEL
VALIDATE MODEL
DEPLOY MODEL
EVALUATE / MONITOR RESULTS

Figure 14.117: Data Mining life cycle as seen by SAS™ (source: SlideShare SAS)

Remember, as we already said it in the introduction chapter of this textbook, that statisticians do not in general exactly agree on how to analyse anything but the simplest of problems! The fact that statistical inference uses mathematics does not imply there is only one reasonable or useful way to conduct an analysis. Engineering uses math as well, but there are many ways to build a bridge. So trying multiple analyses given to multiple independent teams of researchers may be considered scientifically prudent to test the robustness of findings.

And also let us introduce the 4V's of Data Mining/Machine Learning that resume quite well the most common cases of data to which the Data Scientist is faced to:

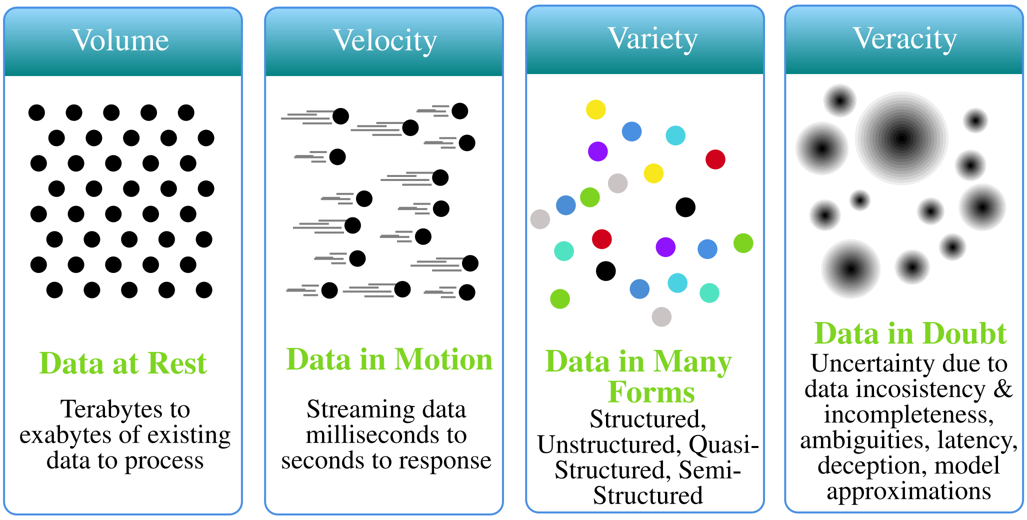

Volume
Data at Rest
Terabytes to exabytes of existing data to process
Velocity
Data in Motion
Streaming data milliseconds to seconds to response
Variety
Data in Many Forms
Structured, Unstructured, Quasi-Structured, Semi-Structured
Veracity
Data in Doubt
Uncertainty due to data incosistency & incompleteness, ambiguities, latency, deception, model approximations

Figure 14.118: Data Mining/Machine Learning 4V's

Remarks

R1. The data scientist must take care of not becoming a "Data Pusher" and not to do just "infotainement" (almost no-sense statistics) in order to remain employed. This is ethically non-scientific and as the reader of this textbook already knows it, as going at the opposite of the Archimedes Oath!

R2. Corporations and managers have to understand that a data steward cannot be a data analyst, that a data analyst that cannot be a data scientist and that latter is also most of time not at the level of statistician. Think the opposite show an evident lack of technical and scientific knowledge from the management, especially when the data scientist is supposed alone to install Big Data servers, clean data, develop new mathematical models, do programming for implementing the models and put them in production.

R3. Data is not a currency, most data is garbage. Actionable information is currency. Extracting information form modern big data sets requires the equivalent processing infrastructure and time of extracting a nugget of gold from a mountain of dirt.

Type of Data	Description	Examples
Unstructured	Data that has no inherent structure	PDF, Text documents (txt, doc, docx), Images, Videos, Slideshows
Quasi-structured	Textual data with erratic formats that can be formatted with effort and software tools	website pages, log files, social networks post, configuration files
Semi-structured	Textual data files with an apparent pattern enabling analysis (but with some pain)	spreadhsheet (xls, xlsx), Comma separated value (csv), JSON, XML, emails
Structured	Data having a defined data model, format, structure,	Parquet files, Relational Databases (i.e. semantic model)

Table 14.24: Types of data in Data Science

Small, medium and international companies must not be surprised to be disappointed after having hired a Data Scientist that theoretically (on the paper...) should "master" various programming languages (R, Python™, C++, Java), various tools (Microsoft® Excel®, Salesforce Tableau, Microsoft Power BI) data retrieval, data wrangling, data archiving, data governance, scientific methodology and advanced statistics as it is impossible for one person to master all these subjects at the same time (or even to have an undergraduate level at the same time of each of these topics)! Furthermore when considering the four types of data below where each one required very specific high level mathematical skills (at PhD level) to be able to correctly and efficiently deal with:

So don't be surprised if by hiring only one person in your company that should do all the following jobs:

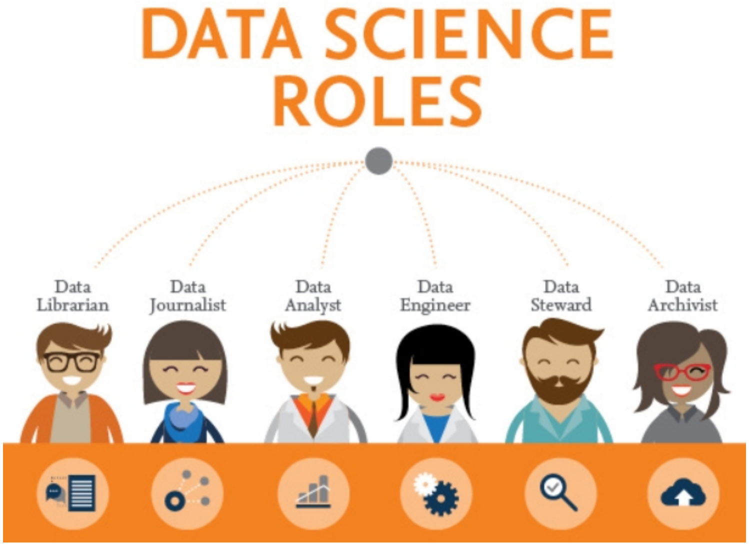
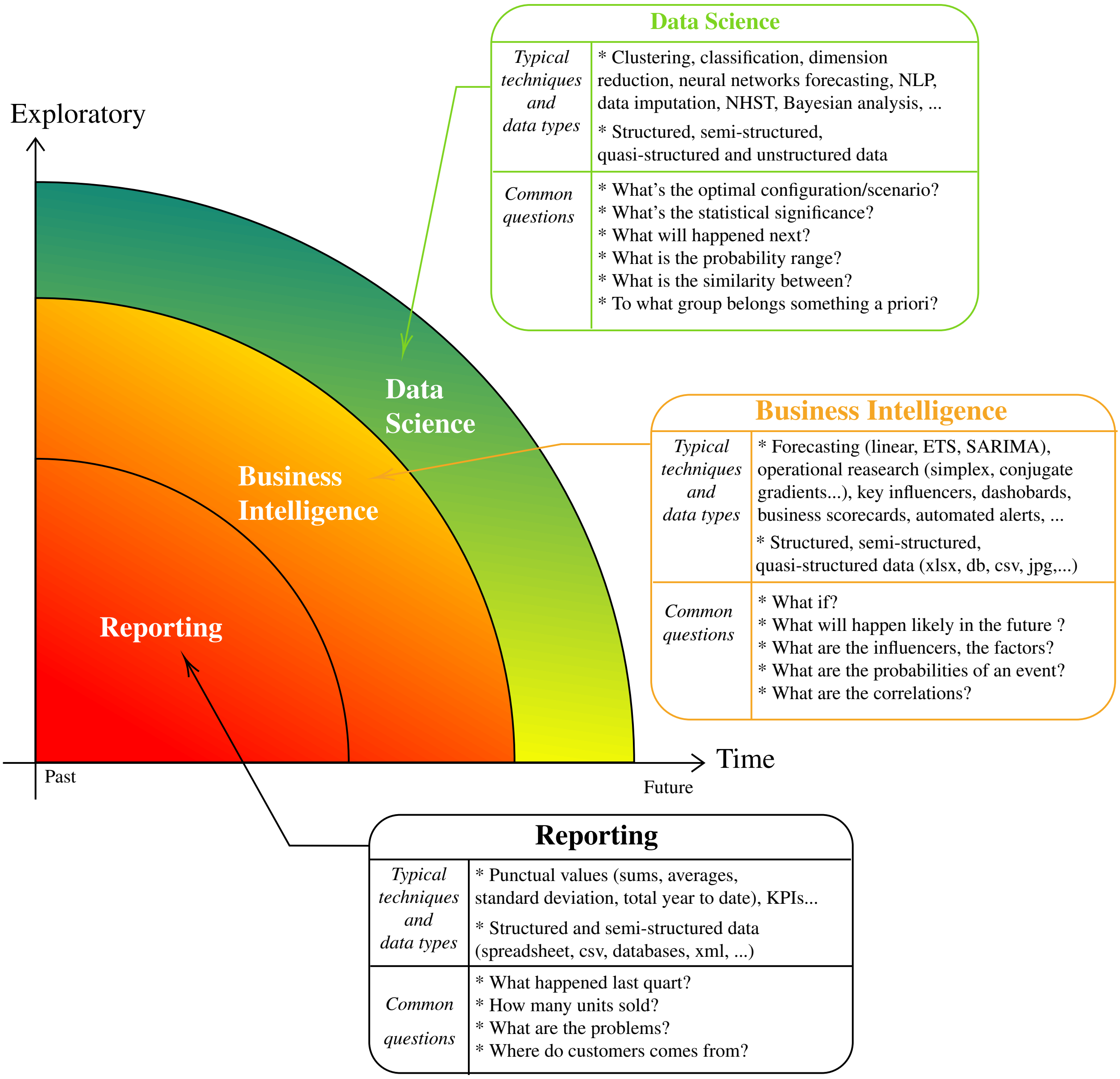


Figure 14.119: Data science roles (source: ?)

that you get garbage (unreliable) results without any strong scientific methodology.

It should be noticed that business intelligence (BI) is quite different from Data Science as illustrated below:



Data Science

Typical techniques and data types	* Clustering, classification, dimension reduction, neural networks forecasting, NLP, data imputation, NHST, Bayesian analysis, ... * Structured, semi-structured, quasi-structured and unstructured data
Common questions	* What's the optimal configuration/scenario? * What's the statistical significance? * What will happened next? * What is the probability range? * What is the similarity between? * To what group belongs something a priori?

Business Intelligence

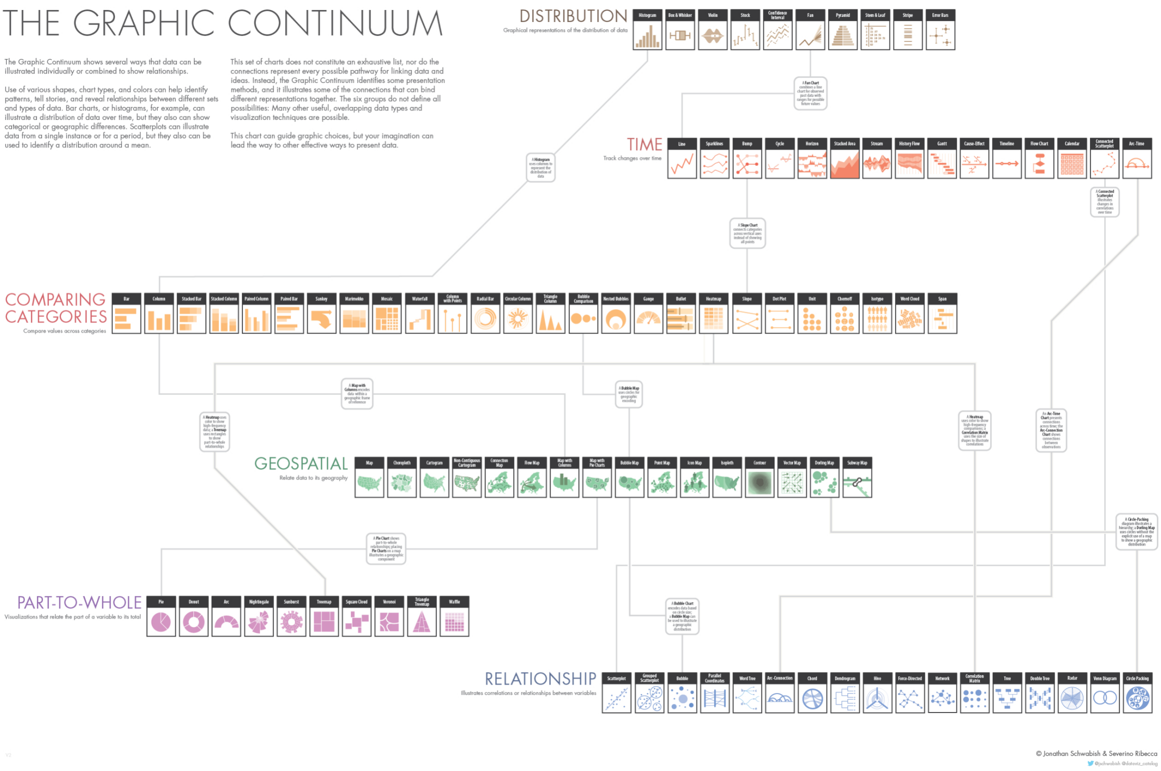
Typical techniques and data types	* Forecasting (linear, ETS, SARIMA), operational reasearch (simplex, conjugate gradients...), key influencers, dashobards, business scorecards, automated alerts, ... * Structured, semi-structured, quasi-structured data (xlsx, db, csv, jpg,...)
Common questions	* What if? * What will happen likely in the future ? * What are the influencers, the factors? * What are the probabilities of an event? * What are the correlations?

Reporting

Typical techniques and data types	* Punctual values (sums, averages, standard deviation, total year to date), KPIs... * Structured and semi-structured data (spreadsheet, csv, databases, xml, ...)
Common questions	* What happened last quart? * How many units sold? * What are the problems? * Where do customers comes from?

Figure 14.120: **Comparing Reporting and Business Intelligence with Data Science**

Finally, as requested by a customer, here is also a list of visuals (charts) used in Business Intelligence (for a more complete list including high level scientists and engineers charts, see page 2204):



THE GRAPHIC CONTINUUM

The Graphic Continuum shows several ways that data can be illustrated individually or combined to show relationships.

Use of various shapes, chart types, and colors can help identify patterns, tell stories, and reveal relationships between different sets and types of data. Bar charts, or histograms, for example, can illustrate a distribution of data over time, but they also can show categorical or geographic differences. Scatterplots can illustrate data from a single instance or for a period, but they also can be used to identify a distribution around a mean.

This set of charts does not constitute an exhaustive list, nor do the connections represent every possible pathway for linking data and ideas. Instead, the Graphic Continuum identifies some presentation methods, and it illustrates some of the connections that can bind different representations together. The six groups do not define all possibilities: Many other useful, overlapping data types and visualization techniques are possible.

This chart can guide graphic choices, but your imagination can lead the way to other effective ways to present data.

A **Histogram** uses columns to represent the distribution of data

A **Fan Chart** combines a line chart for observed past data with ranges for possible future values

A **Slope Chart** connects categories across vertical axes instead of showing all points

A **Connected Scatterplot** illustrates changes in correlations over time

A **Map with Columns** encodes data within a geographic frame of reference

A **Bubble Map** uses circles for geographic encoding

A **Heatmap** uses color to show high-frequency data; a **Treemap** uses rectangles to show part-to-whole relationships

A **Heatmap** uses color to show high-frequency comparisons; a **Correlation Matrix** uses the size of shapes to illustrate correlations

An **Arc-Time Chart** presents connections across time; the **Arc-Connection Chart** shows connections between observations

A **Circle-Packing** diagram illustrates a hierarchy; a **Dorling Map** uses circles without the explicit use of a map to show a geographic distribution

A **Pie Chart** shows part-to-whole relationships; placing **Pie Charts** on a map illustrates a geographic component

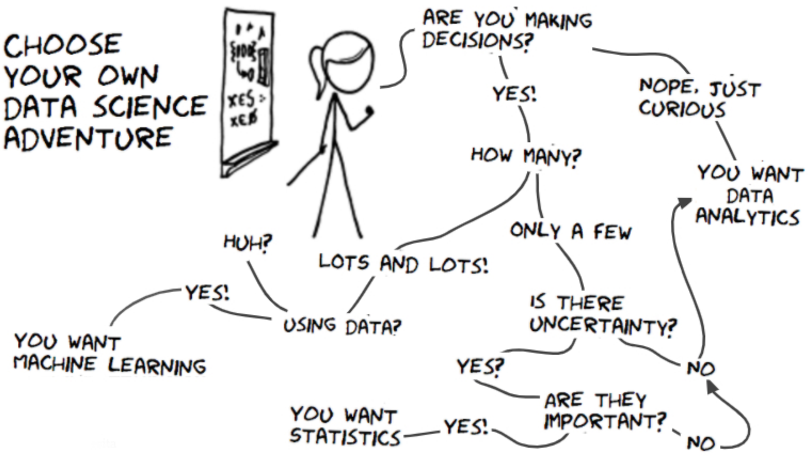
A **Bubble Chart** encodes data based on circle size; a **Bubble Map** can be used to illustrate a geographic distribution

DISTRIBUTION
Graphical representations of the distribution of data
Histogram
Box & Whisker
Violin
Stock
Confidence Interval
Fan
Pyramid
Stem & Leaf
Stripe
Error Bars
TIME
Track changes over time
Line
Sparklines
Bump
Cycle
Horizon
Stacked Area
Stream
History Flow
Gantt
Cause-Effect
Timeline
Flow Chart
Calendar
Connected Scatterplot
Arc-Time
COMPARING CATEGORIES
Compare values across categories
Bar
Column
Stacked Bar
Stacked Column
Paired Column
Paired Bar
Sankey
Marimekko
Mosaic
Waterfall
Column with Points
Radial Bar
Circular Column
Triangle Column
Bubble Comparison
Nested Bubbles
Gauge
Bullet
Heatmap
Slope
Dot Plot
Unit
Chernoff
Isotype
Word Cloud
Span
GEOSPATIAL
Relate data to its geography
Map
Choropleth
Cartogram
Non-Contiguous Cartogram
Connection Map
Flow Map
Map with Columns
Map with Pie Charts
Bubble Map
Point Map
Icon Map
Isopleth
Contour
Vector Map
Dorling Map
Subway Map
PART-TO-WHOLE
Visualizations that relate the part of a variable to its total
Pie
Donut
Arc
Nightingale
Sunburst
Treemap
Square Cloud
Voronoi
Triangle Treemap
Waffle
RELATIONSHIP
Illustrates correlations or relationships between variables
Scatterplot
Grouped Scatterplot
Bubble
Parallel Coordinates
Word Tree
Arc-Connection
Chord
Dendrogram
Hive
Force-Directed
Network
Correlation Matrix
Tree
Double Tree
Radar
Venn Diagram
Circle Packing

© Jonathan Schwabish & Severino Ribecca
@jschwabish @dataviz_catalog

Figure 14.121: Graphic Continuum (authors: Jon Schwabish, Severino Ribecca)

And keep in mind to avoid as Data Scientist, or... as a manager of a Data Scientist (!), to be careful with most common data fallacies (see section Statistics page 1342) and especially with the fact that an engineer is by default not a scientist (same applies for physicians: they are by default not scientists)!



CHOOSE YOUR OWN DATA SCIENCE ADVENTURE
ARE YOU MAKING DECISIONS?
YES!
NOPE, JUST CURIOUS
YOU WANT DATA ANALYTICS
HOW MANY?
ONLY A FEW
LOTS AND LOTS!
HUH?
YES!
USING DATA?
YOU WANT MACHINE LEARNING
IS THERE UNCERTAINTY?
YES?
NO
ARE THEY IMPORTANT?
NO
YES!
YOU WANT STATISTICS

Figure 14.122: source: www.xkcd.com (author: Randall Munroe)

14.1.20.1 Training vs Test Data

First, to define a problem of machine learning, we have to decide what type of modelization function we want to consider.

Definition (#14.1.569): We name "hypothesis space", the set of functions $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ describing the modelization functions that we will consider. This space is chosen depending on our convictions (return on experience) relatively to the problem we are faced to.

We typically split the input data into learning and testing datasets. We then run the Machine Learning algorithm on the learning dataset to generate the prediction model. Later, we use the test dataset to evaluate our model.

Definition (#14.1.570): Given a dataset $\mathcal{D}=\{(\vec{x}^i, y^i)\}_{i=1,\ldots,n}$, partitioned into two sets $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{te}}$, we name "training set" the set $\mathcal{D}_{\text{tr}}$ used to train a predictive model, and test set the set $\mathcal{D}_{\text{te}}$ used for its evaluation.

It is important that the test data is separate from the one used in training otherwise we will be kind of cheating because may for example the generated model memorizes the data and hence if the test data is also part of the training data then our evaluation scores of the model will be higher than they actually are.

The data is usually split 75% training and 25% data or $2/3$ training and $1/3$ testing. It is important to note that: the smaller the training set the more challenging it is for the algorithm to discover the rules.

In addition, when splitting the dataset, we need to maintaining class proportions and population statistics otherwise we will have some classes that are under represented in the training dataset and over represented in the test dataset.

A crucial step when building our Machine Learning model is to estimate its performance on that the model hadn't seen before. We want to make sure that the model generalizes well to new unseen data.

Now consider the situation where we want to choose between K models. We can then train each of the models on the training dataset, thus obtaining K decision functions $f_1, f_2, \ldots, f_K$, then calculate the error of each of these models on the test set. We can then choose as a model the one that has the smallest error on the test set:

$$\hat{f}=\underset{k=1,\ldots,K}{\arg\min}\frac{1}{|\mathcal{D}_{\text{te}}|}\sum_{\vec{x},y\in\mathcal{D}_{\text{te}}} L\left(y, f_k(\vec{x})\right) \qquad (14.1.1701)$$

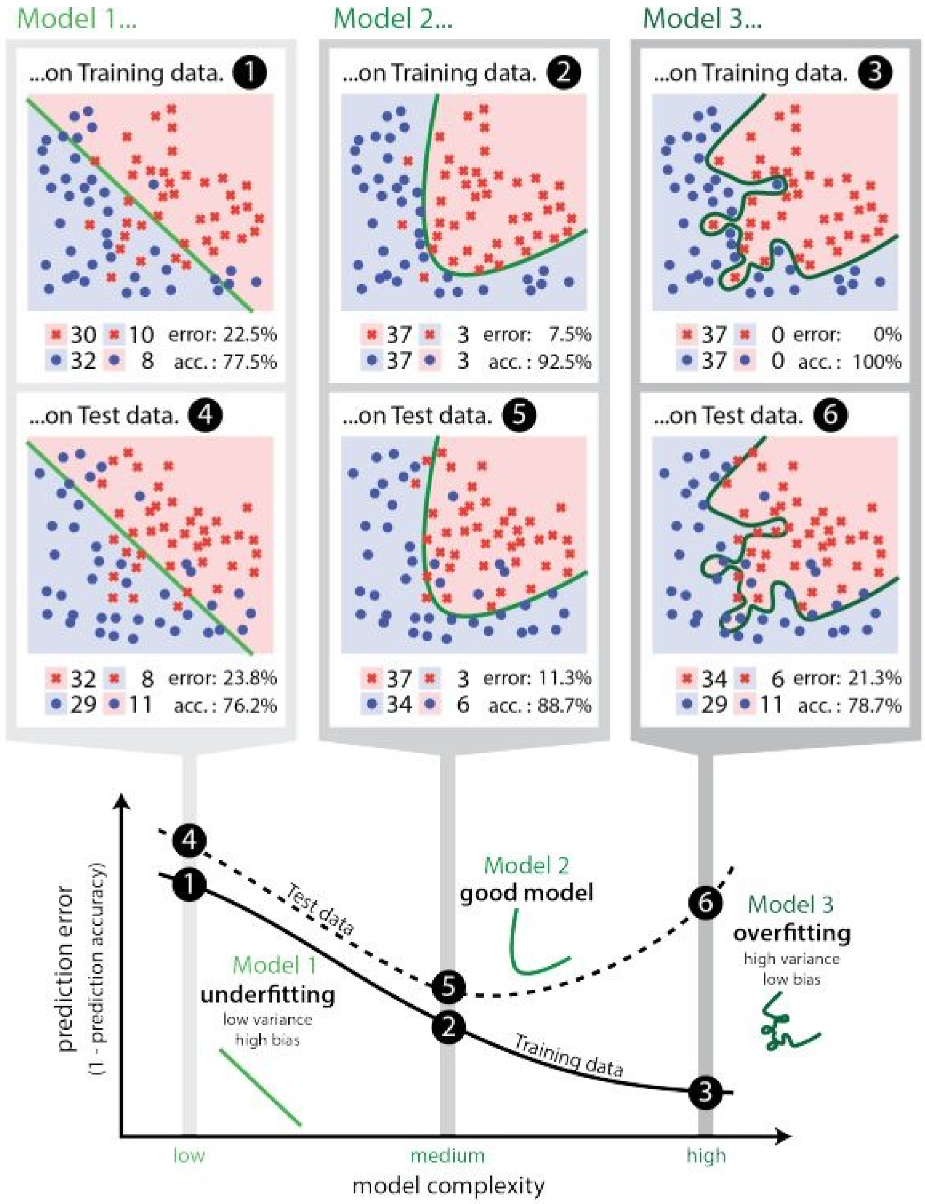


Figure 14.123: **Underfitting and overfitting errors**

In statistics and Machine Learning, the "bias–variance tradeoff" (or dilemma) is the problem of simultaneously minimizing two sources of error that prevent supervised learning algorithms from generalizing beyond their training set:

- The bias is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).
- The variance is an error from sensitivity to small fluctuations in the training set. High variance can cause an algorithm to model the random noise in the training data, rather than the intended outputs (overfitting).

This tradeoff applies to all forms of supervised learning: classification, regression (function fitting), and structured output learning. It may has also been invoked to explain the effectiveness of heuristics in human learning.

The bias-variance tradeoff is a central problem in supervised learning. Ideally, one wants to choose a model that both accurately capture the regularities in its training data, but also generalizes well to unseen data. Unfortunately, it is typically impossible to do both simultaneously. High-variance learning methods may be able to represent their training set well but are at risk of overfitting to noisy or unrepresentative training data. In contrast, algorithms with high bias typically produce simpler models that don't tend to overfit but may underfit their training data, failing to capture important regularities.

Models with low bias are usually more complex (e.g. higher-order regression polynomials), enabling them to represent the training set more accurately. In the process, however, they may also represent a large noise component in the training set, making their predictions less accurate - despite their added complexity. In contrast, models with higher bias tend to be relatively simple (low-order or even linear regression polynomials) but may produce lower variance predictions when applied beyond the training set.

Suppose that we have a training set consisting of a set of points $x_1,\ldots,x_n$ and real values y_i associated with each point x_i. We assume that there is a function with noise $y=f(x)+\varepsilon$, where the noise, ε, has zero mean and variance σ_ε^2.

We want to find a function $\hat{f}$, that approximates the true function $f(x)$ as well as possible, by means of some learning algorithm. We make "as well as possible" precise by measuring the mean squared error between y and $\hat{f}(x)$: we want $(y-\hat{f}(x))^2$ to be minimal, both for $x_1,\ldots,x_n$ and for points outside of our sample. Of course, we cannot hope to do so perfectly, since the y_i contain noise ε; this means we must be prepared to accept an irreducible error in any function we come up with.

Theorem 14.7. Finding an $\hat{f}$ that generalizes to points outside of the training set can be done with any of the countless algorithms used for supervised learning. It turns out that whichever function $\hat{f}$ we select, we can decompose its expected error on an unseen sample x as follows:

$$\text{PE}=\text{E}\left[\left(y-\hat{f}(x)\right)^2\right]=\sigma_\varepsilon^2+\underbrace{\text{V}\left[\hat{f}(x)\right]+\text{Bias}\left[\hat{f}(x)\right]^2}_{\text{model estimation error (to minimize)}} \tag{14.1.1702}$$

named "bias-variance tradeoff", where:

$$\text{Bias}\left[\hat{f}(x)\right]=\text{E}\left[\hat{f}(x)-f(x)\right] \tag{14.1.1703}$$

and:

$$\text{V}\left[\hat{f}(x)\right]=\text{E}\left[\hat{f}(x)^2\right]-\text{E}\left[\hat{f}(x)\right]^2 \tag{14.1.1704}$$

and where PE denotes the "prediction error" also sometimes named the "expected loss".

Proof 14.7.1. The derivation of the bias-variance decomposition for squared error proceeds as follows. For notational convenience, abbreviate $f=f(x)$ and $\hat{f}=\hat{f}(x)$. First, recall that, by definition, for any random variable X, we have:

$$\text{V}(X)=\text{E}(X^2)-\text{E}(X)^2 \tag{14.1.1705}$$

Rearranging, we get:

$$\mathrm{E}(X^2)=\mathrm{V}(X)+\mathrm{E}(X)^2 \tag{14.1.1706}$$

If we assume a model where f is deterministic:

$$\mathrm{E}(f)=f \tag{14.1.1707}$$

This, given $y=f+\varepsilon$ and $\mathrm{E}(\varepsilon)=0$, implies:

$$\mathrm{E}(y)=\mathrm{E}[f+\varepsilon]=\mathrm{E}(f)=f \tag{14.1.1708}$$

Also, since $\mathrm{V}(\varepsilon)=\sigma_\varepsilon^2$:

$$\mathrm{V}(y)=\mathrm{E}\left[(y-\mathrm{E}(y))^2\right]=\mathrm{E}\left[(y-f)^2\right]=\mathrm{E}\left[((f+\varepsilon)-f)^2\right]=\mathrm{E}(\varepsilon^2)=\mathrm{V}[\varepsilon]+\mathrm{E}[\varepsilon]^2=\sigma_\varepsilon^2 \tag{14.1.1709}$$

Thus, since ε and $\hat{f}$ are independent, we can write:

$$\begin{aligned}\mathrm{E}\left[(y-\hat{f})^2\right]&=\mathrm{E}\left(y^2+\hat{f}^2-2y\hat{f}\right)=\mathrm{E}(y^2)+\mathrm{E}(\hat{f}^2)-\mathrm{E}(2y\hat{f})\\&=\left(\mathrm{V}(y)+\mathrm{E}(y)^2\right)+\left(\mathrm{V}(\hat{f})+\mathrm{E}(\hat{f})^2\right)-2f\mathrm{E}(\hat{f})\\&=\mathrm{V}(y)+\mathrm{V}(\hat{f})+\left(f^2-2f\mathrm{E}(\hat{f})+\mathrm{E}(\hat{f})^2\right)\\&=\mathrm{V}(y)+\mathrm{V}(\hat{f})+\left(f-\mathrm{E}(\hat{f})\right)^2\\&=\sigma_\varepsilon^2+\mathrm{V}(\hat{f})+\mathrm{Bias}(\hat{f})^2=\text{noise}+\mathrm{V}(\hat{f})+\mathrm{Bias}(\hat{f})^2\end{aligned} \tag{14.1.1710}$$

So we get finally the bias-variance tradeoff relation:

$$\boxed{\mathrm{E}\left[(y-\hat{f})^2\right]=\text{noise}+\mathrm{V}(\hat{f})+\mathrm{Bias}(\hat{f})^2=\sigma_\varepsilon^2+\mathrm{E}\left[\hat{f}(x)^2\right]-\mathrm{E}\left[\hat{f}(x)\right]^2+\mathrm{E}\left[\hat{f}(x)-f(x)\right]} \tag{14.1.1711}$$

That latter is sometimes denoted as following in some textbooks (the letter D denotes the known data):

$$\mathrm{E}_{D,(y_*,x_*)}\left[(y_*-f(x_*\,|\,D))^2\right]=\underbrace{\mathrm{E}_{\varepsilon_*}\left[(y_*-h(x_*))^2\right]}_{\text{noise}}+\underbrace{\mathrm{E}_D\left[(f(x_*\,|\,D)-\mathrm{E}_D[f(x_*\,|\,D)])^2\right]}_{\text{variance}}+\underbrace{(h(x_*)-\mathrm{E}_D[f(x_*\,|\,D)])^2}_{\text{bias}^2} \tag{14.1.1712}$$

■ Q.E.D.

The trade-off between bias and variance is that:

- Simple Models: High Bias, Low Variance (over-fitting)
- Complex Models: Low Bias, High Variance (under-fitting)

We will derive later the bias and variance for the (vanilla) k nearest neighbours clustering but also for the bagging and bootstrap of random forests (when we will introduce these topics!). Let us however see right now as an example how to derive the prediction error for the linear regression!

✍**Example:**

Our goal is first to derive the bias of the ordinary least square regression:

$$f\left(\vec{x}_* \mid D\right) = \vec{x}_*^T \hat{\vec{\theta}}_{\text{MLE}} \tag{14.1.1713}$$

For that we assume obviously the true model to be:

$$h\left(\vec{x}_*\right) = \vec{x}_*^T \vec{\theta} \tag{14.1.1714}$$

Therefore:

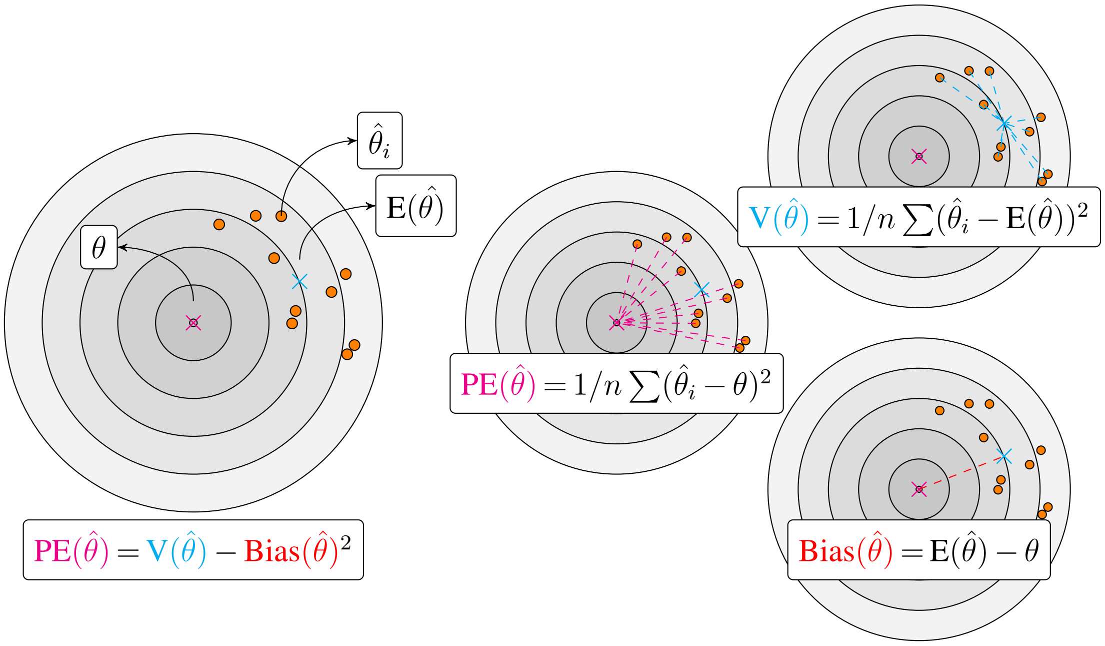
$$\begin{aligned}
\text{Bias} &= h\left(\vec{x}_*\right) - \mathrm{E}_D\left[f\left(\vec{x}_* \mid D\right)\right] = \vec{x}_*^T \vec{\theta} - \mathrm{E}_D\left[\vec{x}_*^T \hat{\vec{\theta}}_{\text{MLE}}\right] = \vec{x}_*^T \vec{\theta} - \mathrm{E}_D\left[\vec{x}_*^T \left(X^T X\right)^{-1} X^T Y\right] \\
&= \vec{x}_*^T \vec{\theta} - \mathrm{E}_D\left[\vec{x}_*^T \left(X^T X\right)^{-1} X^T (X\vec{\theta} + \vec{\varepsilon})\right] \\
&= \vec{x}_*^T \vec{\theta} - \mathrm{E}_D\left[\vec{x}_*^T \left(X^T X\right)^{-1} X^T X \vec{\theta} + \vec{x}_*^T \left(X^T X\right)^{-1} X^T \vec{\varepsilon}\right] \\
&= \vec{x}_*^T \vec{\theta} - \mathrm{E}_D\left[\vec{x}_*^T \vec{\theta} + \vec{x}_*^T \left(X^T X\right)^{-1} X^T \vec{\varepsilon}\right] \\
&= \vec{x}_*^T \vec{\theta} - \vec{x}_*^T \vec{\theta} + \vec{x}_*^T \left(X^T X\right)^{-1} X^T \mathrm{E}_D[\vec{\varepsilon}] = \vec{x}_*^T \vec{\theta} - \vec{x}_*^T \vec{\theta} = 0
\end{aligned} \tag{14.1.1715}$$

Hence $\hat{\theta}_{\text{MLE}}$ is unbiased! Let's compute the variance now:

$$\begin{aligned}
\text{Variance} &= \mathrm{E}\left[\left(f\left(\vec{x}_* \mid D\right) - \mathrm{E}_D\left[f\left(\vec{x}_* \mid D\right)\right]\right)^2\right] = \mathrm{E}\left[\left(\vec{x}_*^T \hat{\vec{\theta}}_{\text{MLE}} - \vec{x}_*^T \vec{\theta}\right)^2\right] \\
&= \mathrm{E}\left[\left(\vec{x}_*^T \left(X^T X\right)^{-1} X^T Y - \vec{x}_*^T \vec{\theta}\right)^2\right] = \mathrm{E}\left[\left(\vec{x}_*^T \left(X^T X\right)^{-1} X^T (X\vec{\theta} + \vec{\varepsilon}) - \vec{x}_*^T \vec{\theta}\right)^2\right] \\
&= \mathrm{E}\left[\left(\vec{x}_*^T \vec{\theta} + x_*^T \left(X^T X\right)^{-1} X^T \vec{\varepsilon} - \vec{x}_*^T \vec{\theta}\right)^2\right] = \mathrm{E}\left[\left(\vec{x}_*^T \left(X^T X\right)^{-1} X^T \vec{\varepsilon}\right)^2\right] \\
&= \mathrm{E}\left[\left(\vec{x}_*^T \left(X^T X\right)^{-1} X^T \vec{\varepsilon}\right)\left(\vec{x}_*^T \left(X^T X\right)^{-1} X^T \vec{\varepsilon}\right)^T\right] \\
&= \mathrm{E}\left[\vec{x}_*^T \left(X^T X\right)^{-1} X^T \vec{\varepsilon}\vec{\varepsilon}^T \left(\vec{x}_*^T \left(X^T X\right)^{-1} X^T\right)^T\right] \\
&= \vec{x}_*^T \left(X^T X\right)^{-1} X^T \mathrm{E}\left[\vec{\varepsilon}\vec{\varepsilon}^T\right] \left(\vec{x}_*^T \left(X^T X\right)^{-1} X^T\right)^T \\
&= \vec{x}_*^T \left(X^T X\right)^{-1} X^T \sigma_\varepsilon^2 I \left(\vec{x}_*^T \left(X^T X\right)^{-1} X^T\right)^T \\
&= \sigma_\varepsilon^2 \vec{x}_*^T \left(X^T X\right)^{-1} X^T X \left(\vec{x}_*^T \left(X^T X\right)^{-1}\right)^T \\
&= \sigma_\varepsilon^2 \vec{x}_*^T \left(\vec{x}_*^T \left(X^T X\right)^{-1}\right)^T = \sigma_\varepsilon^2 \vec{x}_*^T \left(X^T X\right)^{-1} \vec{x}_*
\end{aligned} \tag{14.1.1716}$$

With the above example we see therefore that:

$$\begin{aligned}\text{PE}_{\text{OLS}} &= \underbrace{\text{E}_{\varepsilon_*}\left[(\vec{y}_* - h(\vec{x}_*))^2\right]}_{\text{noise}} + \underbrace{\text{E}_D\left[(f(\vec{x}_* \mid D) - \text{E}_D[f(\vec{x}_* \mid D)])^2\right]}_{\text{variance}} + \underbrace{(h(\vec{x}_*) - \text{E}_D[f(\vec{x}_* \mid D)])^2}_{\text{bias}^2}\\ &= \sigma_\varepsilon^2 + \sigma_\varepsilon^2 \vec{x}_*^T \left(X^TX\right)^{-1} \vec{x}_* + 0 = \sigma_\varepsilon^2\left(1 + \vec{x}_*^T\left(X^TX\right)^{-1}\vec{x}_*\right)\end{aligned} \tag{14.1.1717}$$


$\hat{\theta}_i$
$\text{E}(\hat{\theta})$
θ
$\text{PE}(\hat{\theta}) = \text{V}(\hat{\theta}) - \text{Bias}(\hat{\theta})^2$
$\text{PE}(\hat{\theta}) = 1/n\sum(\hat{\theta}_i - \theta)^2$
$\text{V}(\hat{\theta}) = 1/n\sum(\hat{\theta}_i - \text{E}(\hat{\theta}))^2$
$\text{Bias}(\hat{\theta}) = \text{E}(\hat{\theta}) - \theta$

One case, the Machine Learning algorithm has different parameters and we want to tune these parameters to achieve the best performance (the parameters of the Machine Learning algorithm are named "hyperparameters" as we will see further below[52]). Another case, sometimes we want to try out different algorithms and choose the best performing one. Here is a typical technique:

We simply split the data into training and testing datasets. Then, the training data is further split into training and validation sets.

The training data is used to train different models. Then the validation data is used to compute performance of each of them and we select the best one. Finally, the model is then used for the test set to evaluate performance. The next figure illustrates this idea:

---

[52] The double problem in Data Mining to find the best algorithm + its better hyperparameters is named the "CASH problem", where CASH stands for "Combined Algorithm Selection and Hyperparameters"

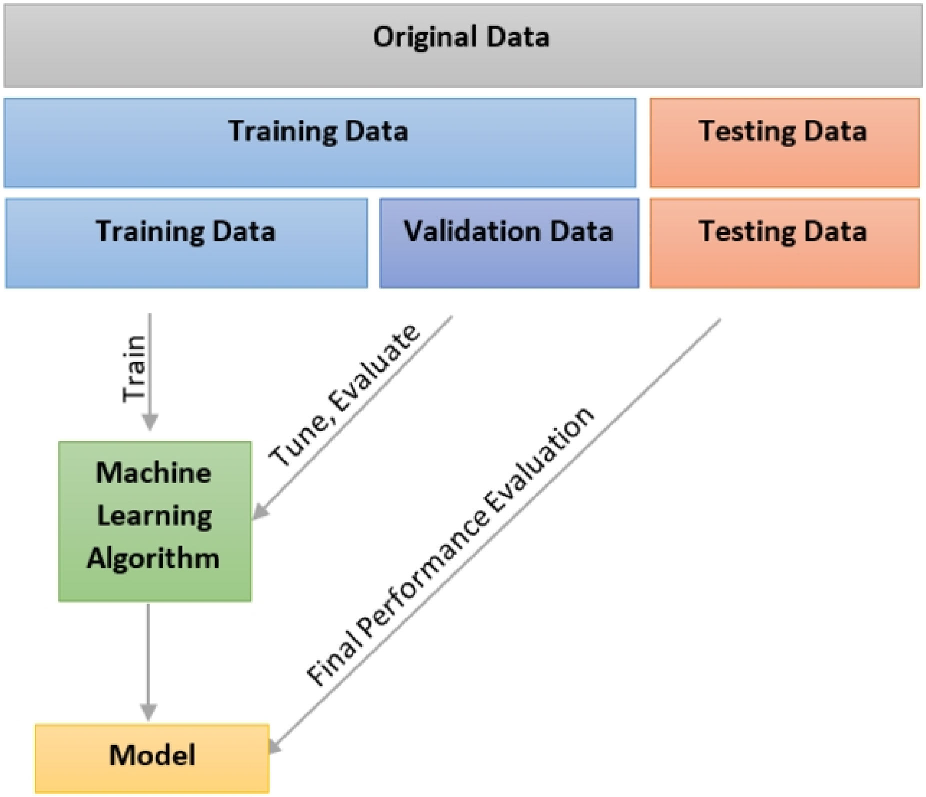


Figure 14.124: Holdout Validation in Machine learning

However, because we use the validation set multiple times, Holdout validation is sensitive to how we partition the data and that is what K-fold cross validation tries to solve.

Initially, we split the data into training and testing dataset. Furthermore, the training dataset is split into K chunks.

Suppose we will use 5-fold cross validation, the training data set is split into 5 chunks and the training phase will take place over 5 iterations. In each iteration we use one chunk as the validation dataset while the rest of the chunk are grouped together to form the training dataset.

This is very similar to Holdout validation except in each iteration the validation data is different and this removed the bias. Each iteration generates a score and the final score is the average score of all iteration. As before we select the best model and use the test data for the final performance evaluation.

It is interesting to know that this practice of splitting data seems to come first form an article of David Freedman [191] published in 11983 according to holocene calendar (hence sometimes the name "Freedman paradox"). Freedman shows in an impressive way the dangers of data reuse in statistical analysis. The potentially dangerous scenarios include those where the results of one statistical procedure performed on the data are fed into another procedure performed on the same data. As a concrete example Freedman considers the practice of performing variable selection first, and then fitting another model using only the identified variables on the same data that was used to identify them in the first place.

The figure below shows the distributions of the considered model statistics (variable selection and afterwards a regression) by David Freedman across the 1000 repetitions for model fits with and without data reuse (the code producing this figure is given at the bottom of this post):

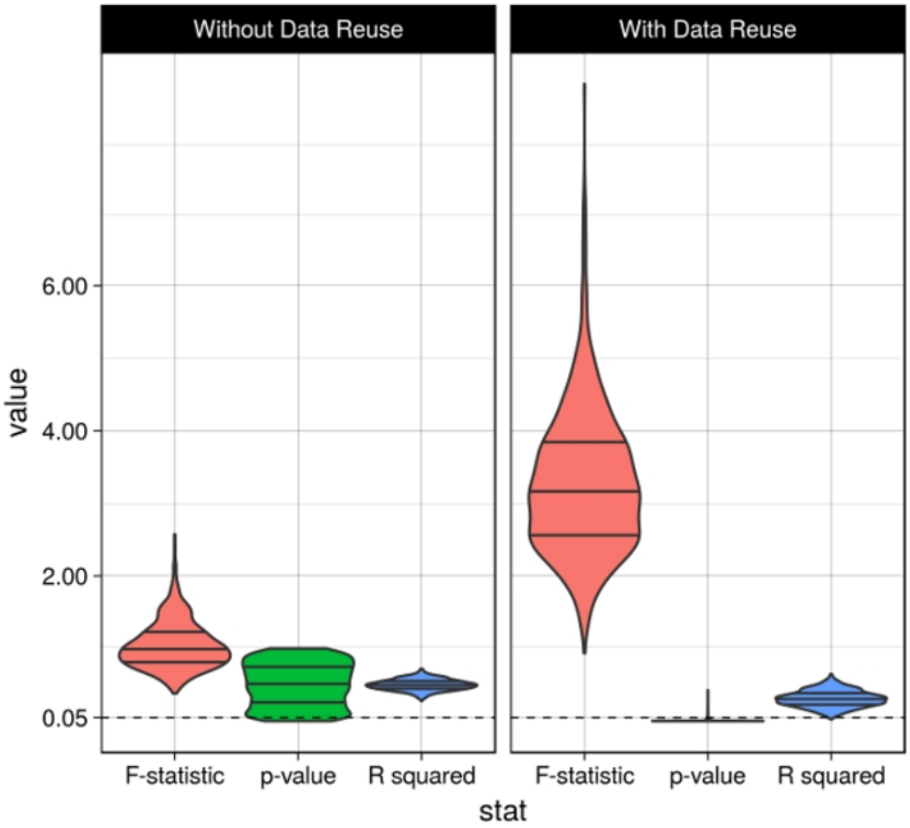


Figure 14.125: **Freedman paradox illustration**

Well, the R squared statistic shows a moderate change between models with or without data reuse (average of 0.3093018 vs. 0.5001641). The F test statistic however grows immensely to an average of 3.2806118 (from 1.0480097), and the p-values fall after data reuse to an average of 0.0112216 (from 0.5017696), below the widely used (but arbitrary) 5% significance level.

Obviously the model with data reuse is highly misleading here, because in fact there are absolutely no relationships between the predictor variables and the response (as per the data generation procedure).

OK! This done let us now begin first with the basics of clustering methods!

14.1.20.2 Hyperparameters vs Parameters

You may have heard of terms like "hyperparameter" search, autotuning (which is just a shorter way of saying hyperparameter search), or grid search (a possible method for hyperparameter search). Where do those terms fit in? To understand hyperparameter search, we have to talk about the difference between a model parameter and a hyperparameter. In brief, model parameters are the knobs that the training algorithm knows how to tweak; they are learned from data. Hyperparameters, on the other hand, are not learned by the training method, but they also need to be tuned.

A famous example are neural networks (see further below) where weights and bias are autotuned parameters, but the number of layers, the type of activation function of the bias values are hyperparameters:

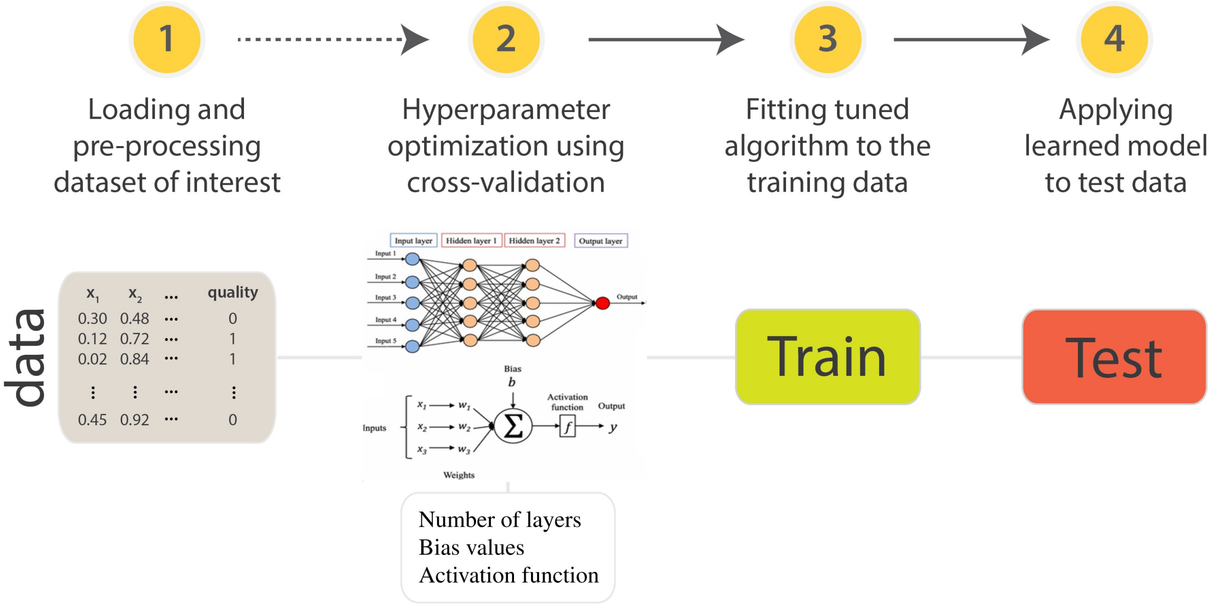

1
Loading and pre-processing dataset of interest
2
Hyperparameter optimization using cross-validation
3
Fitting tuned algorithm to the training data
4
Applying learned model to test data
data
x1 x2 ... quality
0.30 0.48 ... 0
0.12 0.72 ... 1
0.02 0.84 ... 1
0.45 0.92 ... 0
Input layer
Hidden layer 1
Hidden layer 2
Output layer
Input 1
Input 2
Input 3
Input 4
Input 5
Output
Bias
b
Inputs
Weights
Activation function
Output
y
Train
Test
Number of layers
Bias values
Activation function

Figure 14.126: **Parameters vs Hyperparameters**

"Bagging" and "Boosting" are both ensemble methods in Machine Learning that also use hyperparameters (number of iterations/trees, sampling with or without replacement, minimum leaf size, minim parent size, maximum number of decision splits, learning rate for shrinkage, etc).

Bagging and Boosting are similar in that they are both ensemble techniques, where a set of weak learners are combined to create a strong learner that obtains better performance than a single one. So, let's start from the beginning[53].

"Ensemble learning" is a Machine Learning concept in which the idea is to train multiple models using the same learning algorithm. The ensembles take part in a bigger group of methods, named "multiclassifiers", where a set of hundreds or thousands of learners with a common objective are fused together to solve the problem.

---

[53] The whole text introducing the difference between bagging and boosting in this textbook is taken form the QuantDare blog article `https://quantdare.com/what-is-the-difference-between-bagging-and-boosting/` written by Ana Porras Garrido who works for the company ETS Asset Management Factory (`https://www.etsfactory.com/`)

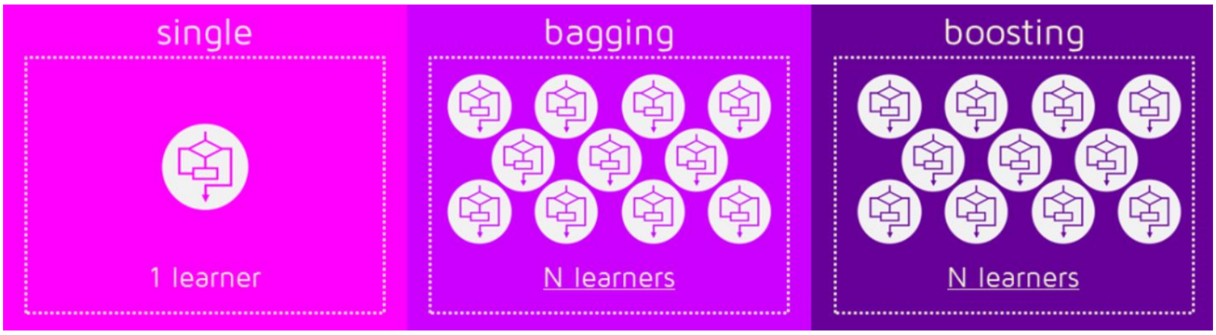
The second group of multiclassifiers contain the "hybrid methods ". They use a set of learners too, but they can be trained using different learning techniques.

The main causes of error in learning are due to noise, bias and variance. Ensemble helps to minimize these factors. These methods are designed to improve the stability and the accuracy of Machine Learning algorithms. Combinations of multiple classifiers decrease variance, especially in the case of unstable classifiers, and may produce a more reliable classification than a single classifier.

To use Bagging or Boosting you must select a base learner algorithm. For example, if we choose a classification tree, Bagging and Boosting would consist of a pool of trees as big as we want.

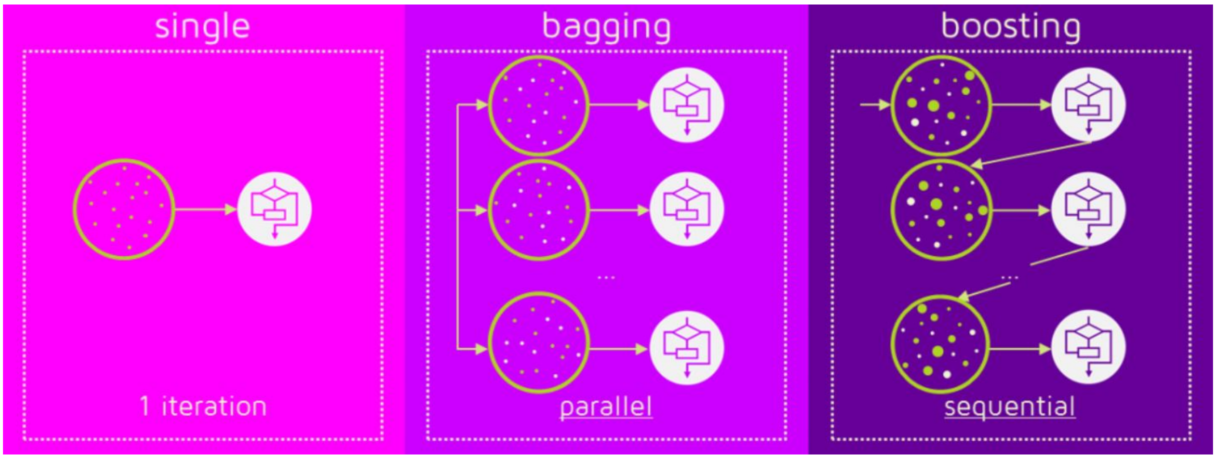
Bagging and Boosting get N learners by generating additional data in the training stage. N new training data sets are produced by random sampling with replacement from the original set. By sampling with replacement some observations may be repeated in each new training data set.

In the case of Bagging, any element has the same probability to appear in a new data set. However, for Boosting the observations are weighted and therefore some of them will take part in the new sets more often:

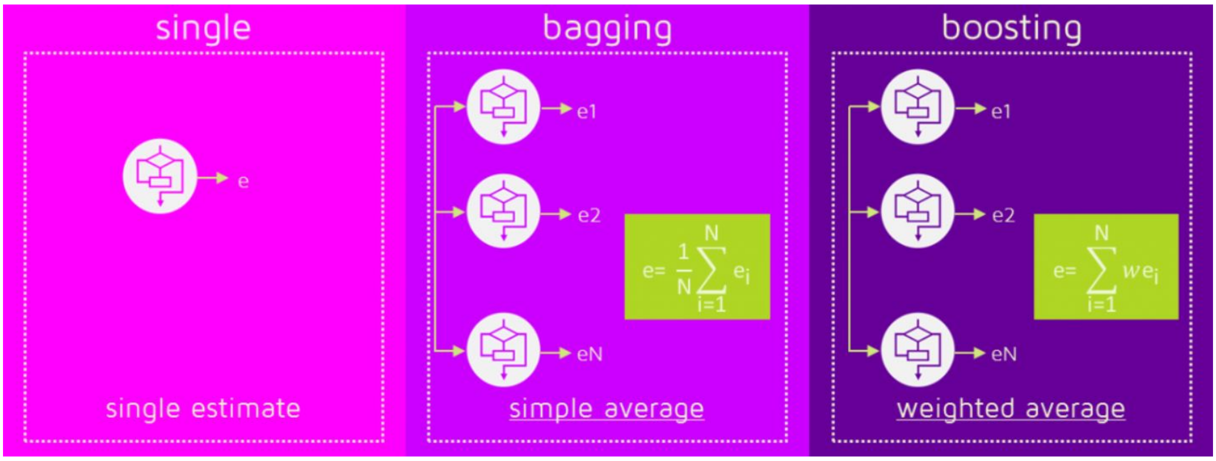


These multiple sets are used to train the same learner algorithm and therefore different classifiers are produced.

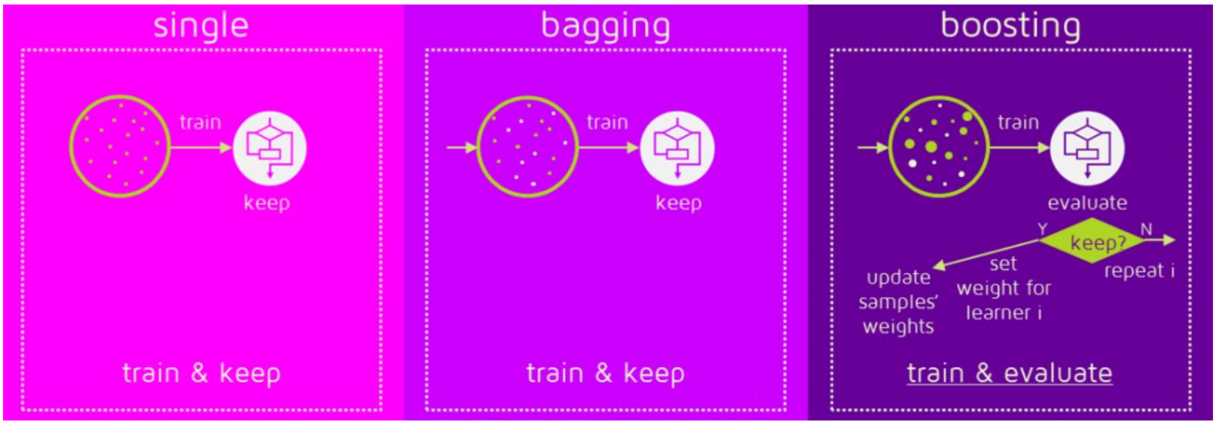
At this point, we begin to deal with the main difference between the two methods. While the training stage is parallel for Bagging (i.e., each model is built independently), Boosting builds the new learner in a sequential way:


single
1 iteration
bagging
...
parallel
boosting
...
sequential

To predict the class of new data we only need to apply the N learners to the new observations. In Bagging the result is obtained by averaging the responses of the N learners (or majority vote). However, Boosting assigns a second set of weights, this time for the N classifiers, in order to take a weighted average of their estimates.


single
e
single estimate
bagging
e1
e2
eN
$e=\frac{1}{N}\sum_{i=1}^{N} e_i$
simple average
boosting
e1
e2
eN
$e=\sum_{i=1}^{N} we_i$
weighted average

In the Boosting training stage, the algorithm allocates weights to each resulting model. A learner with good a classification result on the training data will be assigned a higher weight than a poor one. So when evaluating a new learner, Boosting needs to keep track of learners' errors, too. Let's see the differences in the procedures:


single
train
keep
train & keep
bagging
train
keep
train & keep
boosting
train
evaluate
keep?
Y
N
repeat i
set weight for learner i
update samples' weights
train & evaluate

Some of the Boosting techniques include an extra-condition (hyperparamater) to keep or discard a single learner. For example, in AdaBoost, the most renowned, an error less than 50% is required to maintain the model; otherwise, the iteration is repeated until achieving a learner better than a random guess.

The previous image shows the general process of a Boosting method, but several alternatives exist with different ways to determine the weights to use in the next training step and in the classification stage (AdaBoost, LPBoost, XGBoost, GradientBoost, BrownBoost).

We may ask ourselves what is the best, Bagging or Boosting?

There's not an outright winner; it depends on the data, the simulation and the circumstances. Bagging and Boosting decrease the variance of your single estimate as they combine several estimates from different models. So the result may be a model with higher stability.

Indeed, let us recall that the bias-variance tradeoff relation that we have derived at page 6798:

$$\text{PE}=\text{noise}+\text{V}(\hat{f})+\text{Bias}(\hat{f})^2=\sigma_\varepsilon^2+\text{E}\left[\hat{f}(x)^2\right]-\text{E}\left[\hat{f}(x)\right]^2+\text{E}\left[\hat{f}(x)-f(x)\right] \quad (14.1.1718)$$

As bagging consist to average the different models, we know that averaging reduces the variance according:

$$\text{V}(\bar{X})=\frac{\text{V}(X)}{n} \quad (14.1.1719)$$

So bagging B times leads to:

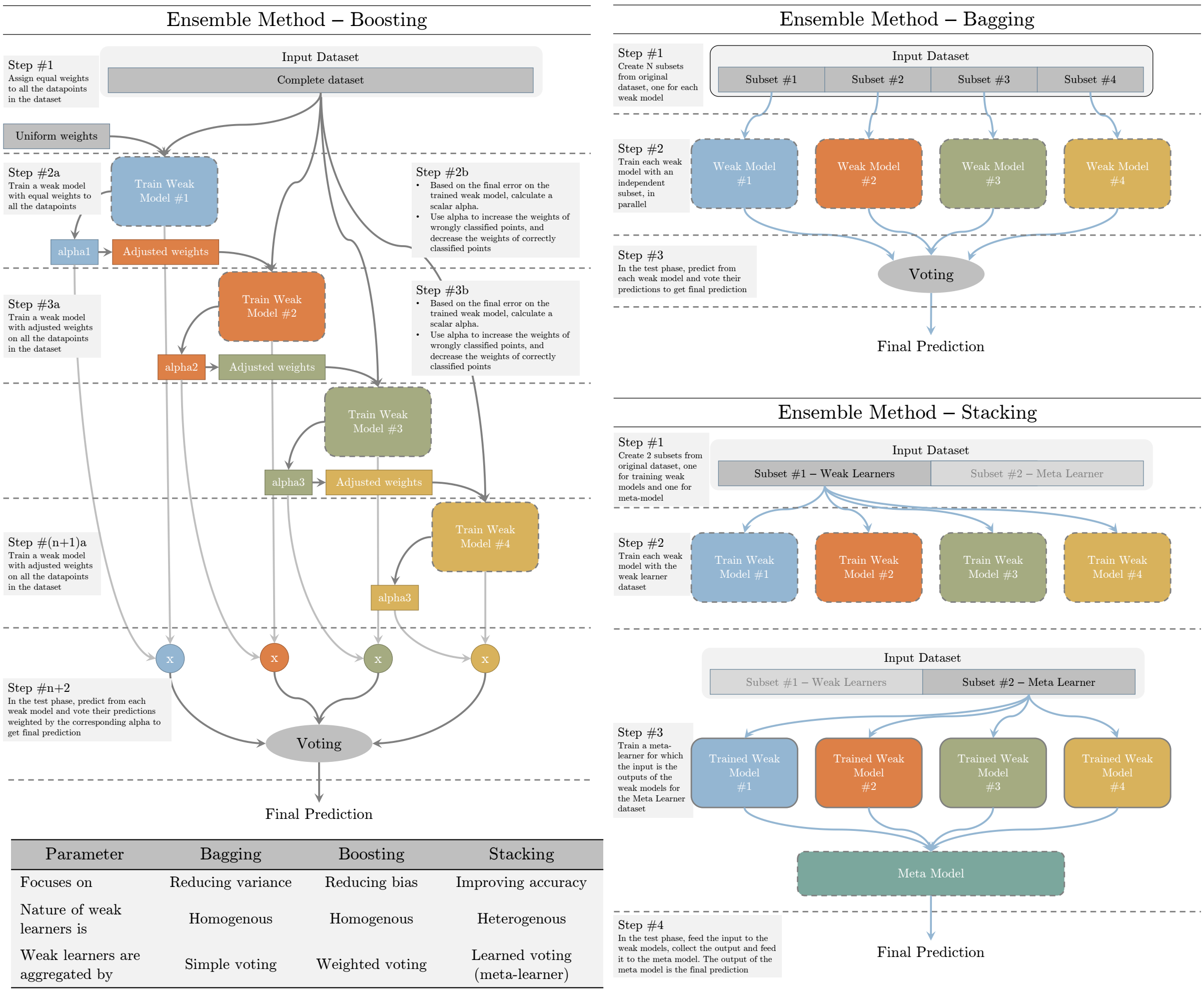
$$\text{PE}_{\text{bagging}}=\text{noise}+\frac{\text{V}(\hat{f})}{B}+\text{Bias}(\hat{f})^2 \quad (14.1.1720)$$

As we know decreasing the variance therefore increase the bias (as the sum of the both must always be equal). In practice the bagging models are correlated, and as we know it, correlation makes variance even smaller!

If the problem is that the single model gets a very low performance, Bagging will rarely get a better bias. However, Boosting could generate a combined model with lower errors as it optimises the advantages and reduces pitfalls of the single model.

By contrast, if the difficulty of the single model is over-fitting, then Bagging is the best option. Boosting for its part doesn't help to avoid over-fitting; in fact, this technique is faced with this problem itself. For this reason, Bagging is effective more often than Boosting.

Here is a very nice summary of the most three common ensemble learning method in one picture:



Ensemble Method – Boosting
Input Dataset
Complete dataset
Uniform weights
Train Weak Model #1
alpha1
Adjusted weights
Train Weak Model #2
alpha2
Adjusted weights
Train Weak Model #3
alpha3
Adjusted weights
Train Weak Model #4
alpha3
x
x
x
x
Voting
Final Prediction
Ensemble Method – Bagging
Input Dataset
Subset #1
Subset #2
Subset #3
Subset #4
Weak Model #1
Weak Model #2
Weak Model #3
Weak Model #4
Voting
Final Prediction
Ensemble Method – Stacking
Input Dataset
Subset #1 – Weak Learners
Subset #2 – Meta Learner
Train Weak Model #1
Train Weak Model #2
Train Weak Model #3
Train Weak Model #4
Input Dataset
Subset #1 – Weak Learners
Subset #2 – Meta Learner
Trained Weak Model #1
Trained Weak Model #2
Trained Weak Model #3
Trained Weak Model #4
Meta Model
Final Prediction

Ensemble Method – Boosting

Step #1
Assign equal weights to all the datapoints in the dataset

Step #2a
Train a weak model with equal weights to all the datapoints

Step #2b
- Based on the final error on the trained weak model, calculate a scalar alpha.
- Use alpha to increase the weights of wrongly classified points, and decrease the weights of correctly classified points

Step #3a
Train a weak model with adjusted weights on all the datapoints in the dataset

Step #3b
- Based on the final error on the trained weak model, calculate a scalar alpha.
- Use alpha to increase the weights of wrongly classified points, and decrease the weights of correctly classified points

Step #(n+1)a
Train a weak model with adjusted weights on all the datapoints in the dataset

Step #n+2
In the test phase, predict from each weak model and vote their predictions weighted by the corresponding alpha to get final prediction

Ensemble Method – Bagging

Step #1
Create N subsets from original dataset, one for each weak model

Step #2
Train each weak model with an independent subset, in parallel

Step #3
In the test phase, predict from each weak model and vote their predictions to get final prediction

Ensemble Method – Stacking

Step #1
Create 2 subsets from original dataset, one for training weak models and one for meta-model

Step #2
Train each weak model with the weak learner dataset

Step #3
Train a meta-learner for which the input is the outputs of the weak models for the Meta Learner dataset

Step #4
In the test phase, feed the input to the weak models, collect the output and feed it to the meta model. The output of the meta model is the final prediction

Parameter	Bagging	Boosting	Stacking
Focuses on	Reducing variance	Reducing bias	Improving accuracy
Nature of weak learners is	Homogenous	Homogenous	Heterogenous
Weak learners are aggregated by	Simple voting	Weighted voting	Learned voting (meta-learner)

Figure 14.127: Three most common ensemble learning methods (source: https://www.cheatsheets.aqeel-anwar.com, author: Aqeel Anwar)

14.1.20.3 Association Rules

A classic story tells about the increase in beer sales after a store decided to place the beer next to the diapers. What is the relationship between baby diapers and beers? We would be wondering...

Several explanations follow to understand this observation, but that's about the purchases made by the male parents. The latter, when they go to take diapers profit to buy a few beers. This little anecdote actually reveals the existence of a rule of association between these two products which at first sight have nothing in common.

Association rule learning is a unsupervised rule-based Machine Learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness.

For example, the rule $\{\text{diapeer}\}\Rightarrow\{\text{beer}\}$ or $\{\text{onions},\text{potatoes}\}\Rightarrow\{\text{burger}\}$ found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, they are likely to also buy hamburger meat. Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements. In addition to the above example from "market basket analysis" association rules are employed today in many application areas including Web usage mining, intrusion detection, continuous production, and bioinformatics. In contrast with sequence mining, association rule learning typically does not consider the order of items either within a transaction or across transactions.

Here are some possible questions that association rules can answer:

- Which products tend to be purchased together?
- Of those customers who are similar to this person, what products do they tend to buy?
- Of those customers who have purchased this product, what other similar products do they tend to view or purchase?

Some well-known algorithms are OneR (OneRule), Apriori, ECLAT (Equivalence Class Transformation) and FP-Growth (first pass), but they only do half the job, since they are algorithms for mining frequent itemsets. Another step needs to be done after to generate rules from frequent itemsets found in a database.

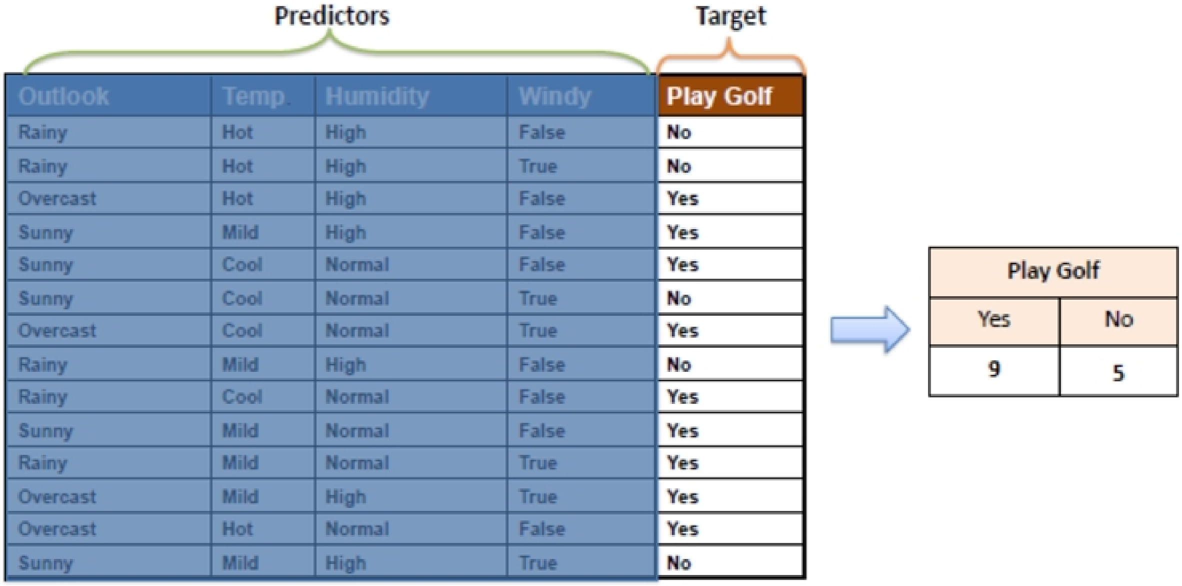
14.1.20.3.1 ZeroR

"ZeroR", short for "Zero Rule", is the simplest supervised classification method which relies on the target and ignores all predictors. ZeroR classifier simply predicts the majority category (class). Although there is no predictability power in ZeroR, it is useful for determining a baseline performance as a benchmark for other classification methods.

A nice story about ZeroR algorithm is that we use it unconsciously everyday when our brain chooses the most frequent seen item as the most probable event (ZeroR classification) or when we talk about averages (ZeroR Regression).

The ZeroR algorithm is (it's not a joke!): Construct a frequency table for the target and select its most frequent value.

The "$Play\ Golf = Yes$" in the example below[54] is the ZeroR rule for the learning table (14 rows) with an accuracy of $9/14 \cong 0.64$.


Predictors
Target

Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Play Golf	
Yes	No
9	5

Figure 14.128: **ZeroR Data Mining (association rule) example**

There is nothing to be said about the predictors contribution to the model because ZeroR does not use any of them.

The following confusion matrix C (more complete than the one we have introduce during our study of the binomial logistic regression):

---

[54] Thanks to Dr. Saed Sayad for having authorized us to reproduce its web page: `http://www.saedsayad.com/zeror.htm`

Confusion matrix		**Play Golf**			
		Yes	No		
ZeroR	Yes	a	b	Positive Predictive Value	$\frac{a}{a+b}$
	No	c	d	Negative Predictive Value	$\frac{d}{c+d}$
		Sensitivity $\frac{a}{a+c}$	Specificity $\frac{d}{b+d}$	**Accuracy** $=\frac{a+d}{a+b+c+d}$	

Table 14.25: **Full Confusion matrix example**

gives in the case of our example:

Confusion matrix		**Play Golf**			
		Yes	No		
ZeroR	Yes	9	5	Positive Predictive Value	0.64
	No	0	0	Negative Predictive Value	0.00
		Sensitivity 1.00	Specificity 0.00	**Accuracy** =0.65	

that ZeroR only predicts the majority class correctly. As mentioned before, ZeroR is only useful for determining a baseline performance for other classification methods.

14.1.20.3.2 OneR (classification by induction)

"OneR", short for "One Rule", is a simple, yet accurate, classification algorithm that generates one rule for each predictor in the data, then selects the rule with the smallest total error as its "one rule". To create a rule for a predictor p_i among the set of all predictor $\mathbb{P}$, we construct a frequency table for each predictor against each possible value t_i of the target attribute set $\mathbb{T}$.

The OneR algorithm is the following[55]:

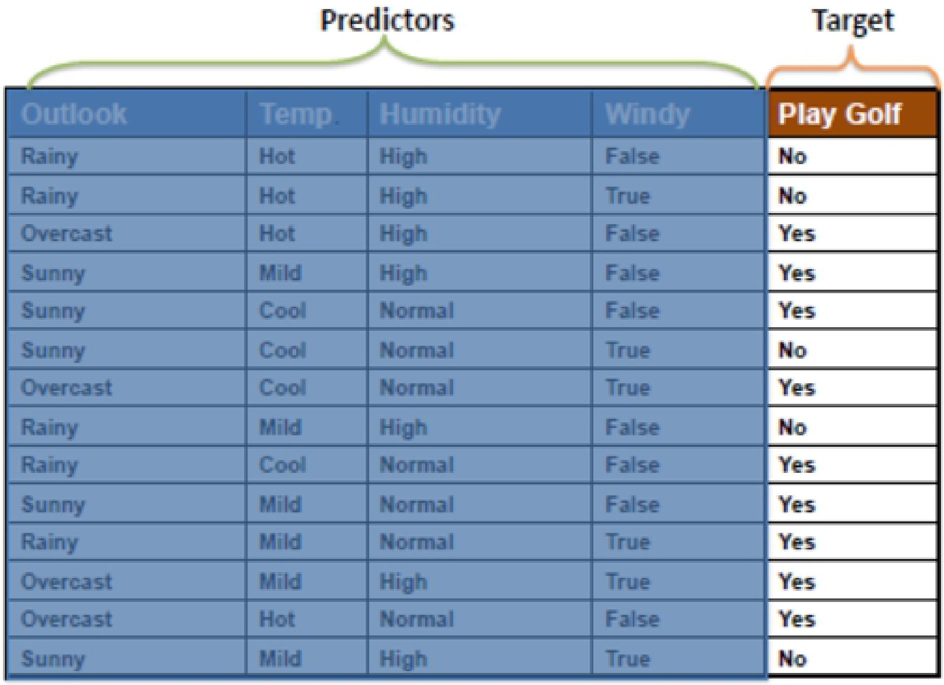
Algorithm 15: OneR (one-Rule) algorithm

```
Data: $p_i$ $a_{i,j}$,$r$, $T$
Result: $\min\{(p_i,\varepsilon_i\}$
initialization;
foreach predictor $p_i \in \mathcal{P}$ do
    foreach attribute value $a_{i,j}$ in $p_i$ do
        Count how often each value of target (class) $t_k \in \mathcal{T}$ appears in all rows $r$ for $p_i$;
        Find the most frequent class $t_k$ for $a_{i,j}$ ;
        Make the rule assign that class to this value of the predictor;
    Calculate the total error of the rules of each predictor $\varepsilon_i$
```

---

[55] Thanks again to Dr. Saed Sayad for having authorized us to reproduce its web page: `http://www.saedsayad.com/oner.htm`

As companion example let us consider again the following learning table (14 rows):


Predictors
Target

Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Figure 14.129: **OneR Data Mining (association rule) example**

Now we create the following frequency tables for each predictor and each attribute inside each predictor:

		Play Golf	
		Yes	*No*
Outlook	*Sunny*	3	2
	Overcast	4	0
	Rainy	2	3

		Play Golf	
		Yes	*No*
Temperature	*Hot*	2	2
	Mild	4	2
	Cold	3	1

		Play Golf	
		Yes	*No*
Humidity	*High*	3	4
	Normal	6	1

		Play Golf	
		Yes	*No*
Windy	*False*	6	2
	True	3	3

Now according to the algorithm, for each predictor we keep only the attribute that has the biggest frequency (in red below[56]):

		Play Golf	
		Yes	*No*
Outlook	*Sunny*	3	2
	Overcast	4	0
	Rainy	2	3

		Play Golf	
		Yes	*No*
Temperature	*Hot*	2	2
	Mild	4	2
	Cold	3	1

		Play Golf	
		Yes	*No*
Humidity	*High*	3	4
	Normal	6	1

		Play Golf	
		Yes	*No*
Windy	*False*	6	2
	True	3	3

This means that so far we would have the rules:

- For *Outlook*:

```
IF Outlook = Sunny THEN Play Golf = Yes
IF Outlook = Overcast THEN Play Golf = Yes
IF Outlook = Rainy THEN Play Golf = No
```

- For *Temperature*:

```
IF Temperature = Hot THEN Play Golf = Yes
IF Temperature = Mild THEN Play Golf = Yes
IF Temperature = Cool THEN Play Golf = Yes
```

- For *Humidity*:

```
IF Humidity = High THEN Play Golf = No
IF Humidity = Normal THEN Play Golf = Yes
```

---

[56] We can notice that in case of equality we take the target that occurs the more for the other attributes.

- For *Windy*:

```
IF Windy = False THEN Play Golf = Yes
IF Windy = True THEN Play Golf = Yes
```

Now if we count for each attribute according to the above rules the number of error of wrongly (false) classified items, we get:

- For *Outlook*: 4 wrongly classified items (10 well classified)
- For *Temperature*: 5 wrongly classified items (9 well classified)
- For *Humidity*: 4 wrongly classified items (10 well classified)
- For *Windy*: 5 wrongly classified items (9 well classified)

We see that only like this it is quite not easy to select the best rule. Therefore a possible solution is to use again a confusion matrix $\mathcal{C}$ for each attribute:

Confusion matrix		**Play Golf** Yes	 No		
OneR	Yes	a	b	Positive Predictive Value	$\frac{a}{a+b}$
	No	c	d	Negative Predictive Value	$\frac{d}{c+d}$
		Sensitivity $\frac{a}{a+c}$	Specificity $\frac{d}{b+d}$	**Accuracy =** $\frac{a+d}{a+b+c+d}$	

Therefore after building all the confusion matrices we notice that the one with the best accuracy if for the *Outlook* attribute:

Confusion matrix		**Play Golf** Yes	 No		
OneR	Yes	7	2	Positive Predictive Value	0.78
	No	2	3	Negative Predictive Value	0.60
		Sensitivity 0.78	Specificity 0.60	**Accuracy =**0.71	

Hence the final chosen association rule:

```
IF Outlook = Sunny THEN Play Golf = Yes
IF Outlook = Overcast THEN Play Golf = Yes
IF Outlook = Rainy THEN Play Golf = No
```

ToDo: Create the confusion matrices for the other attributes

14.1.20.3.3 Apriori

Before introducing the Apriori algorithm which is the second most popular algorithm of the Association Rules, it is important to define some key concepts specific to these particular methods of learning.

Here are some vocabulary definitions:

- A "transaction" is a basket or cart of one or more items purchased by consumers.

✍**Example:**

Here is a base of five transactions or baskets:

Transaction Number	Basket composition			
1	bread	juice	jam	sugar
2	milk	bread	sugar	
3	cheese	bread	butter	milk
4	cheese	milk	sugar	flour
5	chips	bread	jam	yoghurt

- An "item" designates an article. Example: juice, bread, milk ...
- An "itemset" is a collection of items or articles. A itemset containing k items or a "k-itemset" is written between brackets $\{\text{item}\,1, \text{item}\,2, \text{item}\,3, \ldots, \text{item}\,k\}$.

✍**Examples:**

- 1-itemset: $\{\text{milk}\}$ or $\{\text{bread}\}$
- 2-itemset: $\{\text{milk}, \text{bread}\}$ or $\{\text{bread}, \text{butter}\}$
- 3-itemset: $\{\text{bread}, \text{butter}, \text{sugar}\}$
- ...

- Given a k-itemset P, the "support" of P, is the percentage of transactions containing P. This is an index of reliability.

✍**Examples:**

E1. For $S=\{\text{bread},\text{milk}\}$ then the support according to the previous set of baskets is equal to $P=2/5=0.4$, hence 40%:

Transaction Number	Basket composition			
1	bread	juice	jam	sugar
2	milk	bread	sugar	
3	cheese	bread	butter	milk
4	cheese	milk	sugar	flour
5	chips	bread	jam	yoghurt

E2. For $L=\{\text{bread}\}$ then the support of L is $P=4/5=0.8$, hence 80% of the transactions contains L:

Transaction Number	Basket composition			
1	bread	juice	jam	sugar
2	milk	bread	sugar	
3	cheese	bread	butter	milk
4	cheese	milk	sugar	flour
5	chips	bread	jam	yoghurt

- The "minimum support" is the minimum threshold of from which one qualifies a k-itemset as frequent.
- One will say of a itemset that it is a "frequent k-itemset" if its support is greater or equal to the minimum support.

✍**Examples:**

If the minimum support is 0.5, we can say that L, from the previous example, is a frequent itemset but not S whose support is only 0.4.

"Apriori" is an algorithm that was proposed by Agrawal and Srikant in 11994 (holocene calendar) for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis.

The "a priori" algorithm is based on the so-named a priori property which is stated as follows: *«If one itemset is considered as frequent then any other sub itemset arising from the items of P is also frequent».*

Taking again the following table as companion example:

Transaction Number	**Basket composition**			
1	bread	juice	jam	sugar
2	milk	bread	sugar	
3	cheese	bread	butter	milk
4	cheese	milk	sugar	flour
5	chips	bread	jam	yoghurt

By setting the minimum support to P_{min} to 0.35, the previous itemset $S=\{\text{bread},\text{milk}\}$ with a support of 0.4 is therefore referred to as frequent. The previous itemset $L=\{\text{bread}\}$ is a sub-item of $L\subset S$, it is according to the apriori property also frequent (support of $L=0.8>0.35$) and if one considers another new sub-itemset $D=\{\text{milk}\}$ of S, one sees that the support the 1-itemset of D is $3/5=0.6>0.35$.

If we have n different items, then we have 2^n-1 possible different itemsets (the -1 is to remove the empty itemset). Therefore in our case, we could create $2^{10}-1=1023$ different possible itemsets with:

$$\{\text{bread},\text{juice},\text{jam},\text{sugar},\text{milk},\text{cheese},\text{butter},\text{flour},\text{chips},\text{yogourt}\} \tag{14.1.1721}$$

Thus, the number of itemsets to be analysed can be considerably reduced following the pruning of low-level non-frequent itemsets.

Given a minimal support that we will denote δ, the Apriori algorithm adopts an ascending iterative approach that determines all possible 1-itemset first and then those whose support is less than δ are subject to pruning and the rest, those that are frequent are stored in an L_1 itemset collections.

For the 5 transactions of the previous illustration in the 10 possible itemsets of type 1-itemset, for a minimum support of $\delta=0.35$, only the following itemsets will be part of L_1:

1-itemset	**Support**
$\{\text{bread}\}$	0.8
$\{\text{sugar}\}$	0.6
$\{\text{milk}\}$	0.6
$\{\text{jam}\}$	0.4
$\{\text{cheese}\}$	0.4

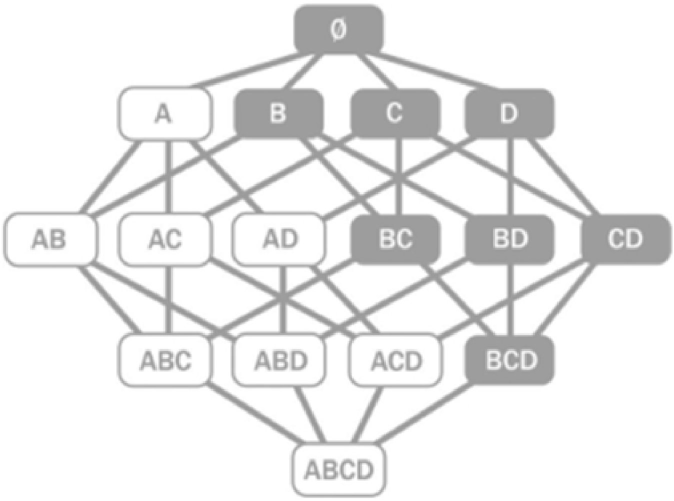
The next step consists of all possible 2-itemsets from the items contained in L_1. Again the 2-itemsets are tested, those whose support is less than δ are pruned and only the frequent 2-itemsets are kept in a collection of L_2 itemsets.

On the basis of the L_1 previously constituted, it is always possible with $\delta=0.35$ to obtain a collection L_2 consisting of:

2-itemset	**Support**
$\{\text{jam, bread}\}$	0.4
$\{\text{cheese, milk}\}$	0.6
$\{\text{sugar, milk}\}$	0.6
$\{\text{bread, sugar}\}$	0.4
$\{\text{milk, bread}\}$	0.4

This aggregation and pruning process is repeated for a collection L_k containing the possible itemsets of k-itemsets having validated the test of the minimum support δ.

So figuring the concept with transactions of a maximum of 4-itemsets $\{\mathrm{A},\mathrm{B},\mathrm{C},\mathrm{D}\}$ then with a total of $2^4-1=15$ possible combinations, this procedure would lead for example to:



The algorithm stops if for a new collection L_{k+1}, no $(k+1)$-itemsets validates the test of the minimum support δ (as is the case for the previous figure where one can not obtain 3-itemsets on the basis of L_2). Optionally, we can define an integer K which specifies either the maximum number of items that an itemset can contain or the maximum number of iterations to be performed by the algorithm.

Once the collections of itemsets are formed, it is possible to transcribe this in the form of candidate rules (for example $X\Rightarrow Y$ or: X implies Y). However, some criteria are defined to assess the relevance of an association rule:

- "Confidence" or "trust of a rule": It's the measure of certainty or confidence associated with an association rule. Formally, for a rule $X\Rightarrow Y$, it is the percentage of transaction containing both X and Y on the total of transactions containing X:

$$\text{Conf}(X\Rightarrow Y)=\frac{\text{Support}(X\cap Y)}{\text{Support}(X)}=\frac{S(X\cap Y)}{S(X)} \tag{14.1.1722}$$

✍**Example:**

What is the confidence index of the rule:

$$\{\text{milk},\text{sugar}\}\Rightarrow\text{flour}$$

As:

$$\text{Support}(\text{milk},\text{sugar})=0.4$$

and:

$$\text{Support}(\text{farine},\text{milk},\text{sugar})=0.2$$

Then:

$$\text{Conf}(\{\text{milk},\text{sugar}\}\Rightarrow\text{flour})=\frac{\{\text{milk},\text{sugar}\}}{\text{flour}}=\frac{0.2}{0.4}=0.5$$

thus a confidence index of 50%.

In practice, an acceptable threshold of confidence is defined. We are confident with regard to a rule when the index of confidence is large and moves away from the minimum.

- "Lift of an association rule": Assuming that X and Y are statistically independent (see section Probabilities page 480), the lift measures the percentage of the number of times it is more frequent to obtain X and Y:

$$\text{Lift}(X\Rightarrow Y)=\frac{\text{Support}(X\cap Y)}{\text{Support}(X)\cdot\text{Support}(Y)}=\frac{S(X\cap Y)}{S(X)\cdot S(Y)}\tag{14.1.1723}$$

Thus, if the lift is equal to 1 then X and Y are independent, which means that the rule is not at all relevant: a simple coincidence! While a lift greater than 1, shows how useful the rule is and a higher value shows how strong the rule is and therefore relevant, it is then a kind of correlation measure. A lift smaller than 1 indicates that X and Y appear less often together than expected, this means that the occurrence of X has a negative effect on the occurrence of Y or that X is negatively correlated with Y. Thus, lift is a value between 0 and $+\infty$.

Remark

We could also run (obviously) a complementary χ^2-test of independence of a 2×2 contingency table (having for entries X, Y, $\neg X$, $\neg Y$) to test if there is statistically or not independence (see section Statistics page 1238).

✍**Example:**

What is the lift of the rule:

$$\{\text{milk},\text{sugar}\}\Rightarrow\{\text{flour}\}$$

As:

$$\text{Support}(\{\text{milk}, \text{sugar}\}) = 0.4$$

and:

$$\text{Support}(\{\text{flour}, \text{milk}, \text{sugar}\}) = 0.2$$

and:

$$\text{Support}(\{\text{flour}\}) = 0.2$$

Then:

$$\text{Lift}(\{\text{milk}, \text{sugar}\} \Rightarrow \{\text{flour}\}) = \frac{\{\text{flour}, \text{milk}, \text{sugar}\}}{\{\text{milk}, \text{sugar}\} \cdot \{\text{flour}\}} = \frac{0.2}{0.4 \cdot 0.2} = 2.5$$

Of course, it is always necessary to compare these results with the values obtained with the other association rules before judging the relevance.

Lift and χ^2 are not "null-invariant". That means that they are not so good at evaluating data that contain too many null transactions.

We limited ourselves here to the two most widely used criteria for evaluating association rules, but there are other criteria (some of them are null-invariant) such as Leverage, Conviction index, Coverage, Jaccard, Cosine, Kulcynski, etc.

Measure	Definition	Range	Null-Invariant
$\chi^2(X,Y)$	$\displaystyle\sum_{i,j} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$	$[0,+\infty[$	No
$\text{Lift}(X,Y)$	$\dfrac{S(X\cap Y)}{S(X)\cdot S(Y)}$	$[0,+\infty[$	No
$\text{AllConf}(X,Y)$	$\dfrac{S(X\cap Y)}{\max\left(S(X),S(Y)\right)}$	$[0,1]$	Yes
$\text{Jaccard}(X,Y)$	$\dfrac{S(X\cap Y)}{S(X)+S(Y)-S(X\cap Y)}$	$[0,1]$	Yes
$\text{Cosine}(X,Y)$	$\dfrac{S(X\cap Y)}{\sqrt{S(X)\cdot S(Y)}}$	$[0,1]$	Yes
$\text{Kulczinsky}(X,Y)$	$\dfrac{1}{2}\left(\dfrac{S(X\cap Y)}{S(X)}+\dfrac{S(X\cap Y)}{S(Y)}\right)$	$[0,1]$	Yes
$\text{MaxConf}(X,Y)$	$\max\left(\dfrac{S(X)}{S(X\cap Y)},\dfrac{S(Y)}{S(X\cap Y)}\right)$	$[0,1]$	Yes
...	...	...	...

Table 14.26: **Some support-confidence interesting measures**

Finally, we can say that the association rules are very easily interpretable and their implementation simple and easy to transcribe into operational rules. However, a number of points are quite problematic:

- Runtime cost: Even if the Apriori algorithm eliminates several occurrences with the minimum support rule, it is easy to imagine that with a large database, occurrences that exceed the minimum support can be large.
- Fallacious or parasitic associations: This effect depends in particular on the minimum support defined, but in front of a large database, to obtain a sufficient quantity of rules it is sometimes useful to consider a low minimum support.

14.1.20.4 Equivalence CLAss Transformation (ECLAT)

The "ECLAT" algorithm, acronym for Equivalence CLAss Transformation, is an alternative way to the previous method to build the item-set that performs better in huge databases.

Let us use the companion example as before to introduce detail how ECLAT works:

Transaction Number	**Basket composition**			
1	bread	juice	jam	sugar
2	milk	bread	sugar	
3	cheese	bread	butter	milk
4	cheese	milk	sugar	flour
5	chips	bread	jam	yoghurt

The idea of ECLAT is for each item, to store a list of transaction ID's (abbreviated "tids" as acronym from "transaction ID's" or sometimes "tidlist"). It as vertical data layout at the opposite of the above one. Indeed, the transaction table above transformed into a tids will look like following for the 1-itemset:

Item Set	**TID Set**
bread	$1, 2, 3, 5$
jam	$1, 5$
cheese	$3, 4$
juice	1
milk	$2, 3, 4$
butter	3
sugar	$1, 2, 4$
flour	4
chips	5

Table 14.27: **ECLAT 1-itemset tidlist**

And the corresponding tidlist for 2-itemset:

Item Set	TID Set
{bread,jam}	1, 5
{bread,cheese}	3
{bread,juice}	1
{bread,milk}	2, 3
{bread,butter}	2, 3
{bread,sugar}	1, 2
{cheese,flour}	4
{breach, chips}	5
{jam, juice}	1
{jam, chips}	5
{cheese, milk}	4
{cheese, butter}	4
...	...

Table 14.28: **ECLAT 2-itemset tidlist**

And the corresponding tidlist for 3-itemset:

Item Set	TID Set
{bread,jam,juice}	1
{bread,jam,chips}	5
{bread,milk,sugar}	2
{bread,cheese,milk}	3
{bread,cheese,butter}	3
{jam,juice,sugar}	1
{cheese,milk,butter}	3
{cheese,milk,sugar}	4
...	...

Table 14.29: **ECLAT 3-itemset tidlist**

This process repeats, with k incremented by 1 each time, until no frequent items or no candidate itemsets can be found.

The pros are:

- Depth-first search reduces memory requirements
- Usually (considerably) faster than Apriori
- No need to scan the database to find the support of $(k+1)$-itemsets, for $k >\geq 1$

and the cons:

- The TID-sets can be quite long, hence expensive to manipulate

14.1.20.5 Clustering and Classification

In statistical data analysis, "clustering" describes empirical methods of data classification (hierarchical clustering method or data partitioning method). In other words clustering and classification are tasks for predicting the value of a categorical variable (target or class) by building a model based on one or more numerical and/or categorical variables (predictors or attributes).

These techniques typically enable the segmentation of all customers of a company based on their demography or buying patterns, to group documents for presentations, identify new animal or plant species, to group information by individuals or by interests.

Classification	**Clustering**
Use labelled as the input	Use unlabelled as the input
The output is known	The output is unknown
Uses supervised machine learning	Uses unsupervised machine learning
A training data set is provided and used to produce classifications	A training data set is not provided and used to produce clusters
Examples of algorithms: Decision-trees, Bayesian classifiers and Support Vector Machines (SVM)	Examples of algorithms: Partition-based clustering (k-means), Hierarchical clustering (agglomerative / divisive) and DBSCAN
Can be more complex than clustering	Can be less complex than clustering
Does not specify areas for improvement	Specify areas for improvement
Known number of classes	Unknown number of classes
Interpretability High	Interpretability Low

Table 14.30: **Typical Classification vs Clustering differences**

By using some mathematical tricks, some classification techniques can be transformed into clustering techniques. That's why sometimes all these different techniques are confusing!

We see in practice two major families of clustering techniques (careful even experts in the field are unable to agree on a common classification ...!):

1. The "non-hierarchical techniques": that is to say where the final number of classes (groups) is chosen in advance.
2. The "hierarchical techniques": that is to say where it leads to a classification by successive aggregations or divisions.

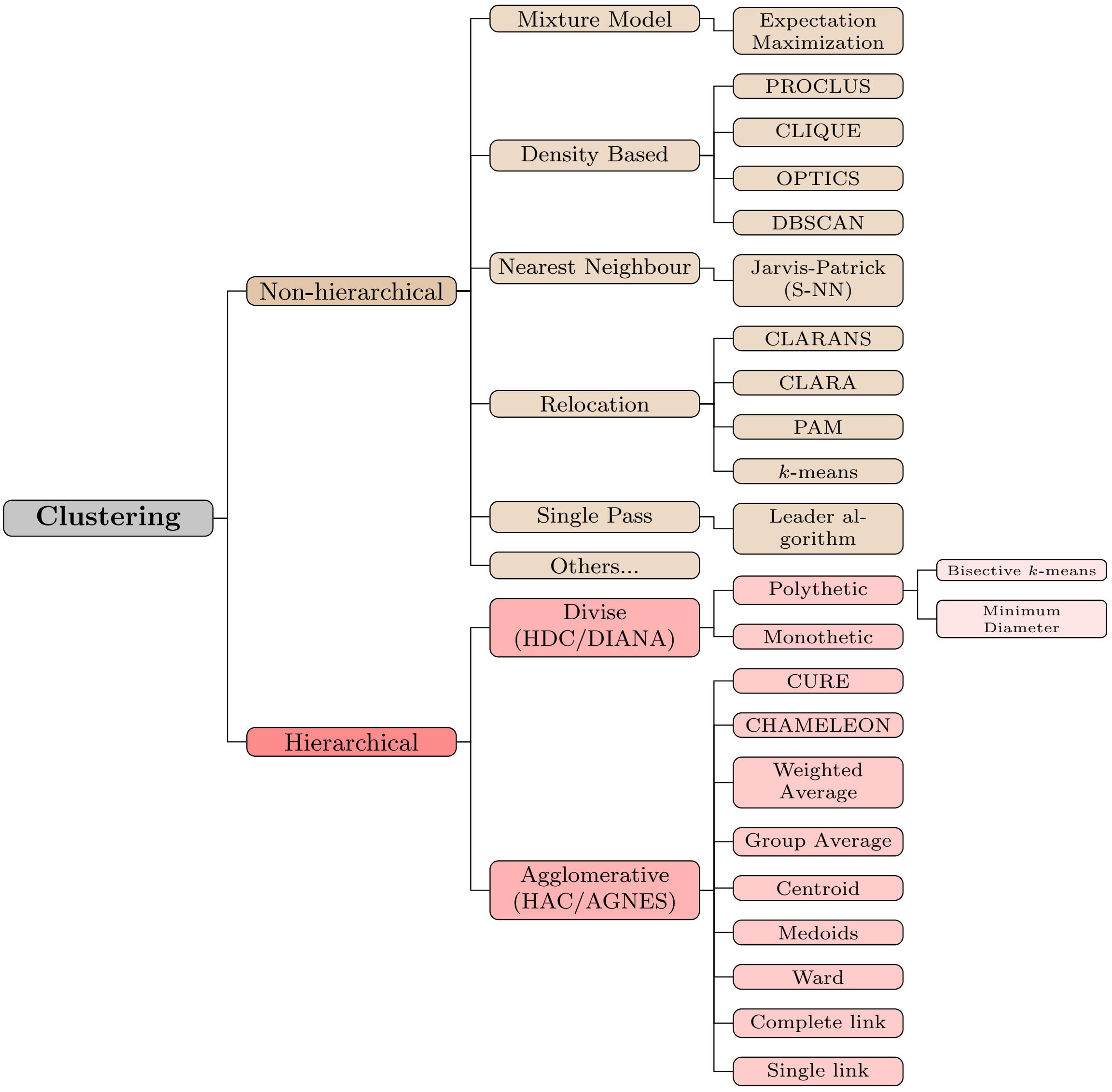

Clustering
Non-hierarchical
Hierarchical
Mixture Model
Expectation Maximization
Density Based
PROCLUS
CLIQUE
OPTICS
DBSCAN
Nearest Neighbour
Jarvis-Patrick (S-NN)
Relocation
CLARANS
CLARA
PAM
k-means
Single Pass
Leader algorithm
Others...
Divise (HDC/DIANA)
Polythetic
Bisective k-means
Minimum Diameter
Monothetic
Agglomerative (HAC/AGNES)
CURE
CHAMELEON
Weighted Average
Group Average
Centroid
Medoids
Ward
Complete link
Single link

Figure 14.130: **Non-exhaustive orgchart of clustering techniques**

Among these two families we distinguish three sub-families not highlighted in the orgchart above (this is the definition for THIS textbook - and that correspond almost to that of the SAS company - as they are no common accepted definitions until now between specialists working in this field):

1. The techniques that make use of data whose nominal classification property is known in advance to train a prediction model: "Machine Learning[57]" or "supervised learning". We find in this category the logistic regression techniques, CRT, ID3, discriminant analysis, Bayesian networks, decision lists, k-NN, neural networks, etc.

   As there are not today a consensus in the definitions, it is important that the reader also knows that "supervised" algorithms by are algorithms that have a correction mechanism of the model parameters based on generated errors.

---

[57] Oxford Dictionary definition: The capacity of a computer to learn from experience, i.e. to modify its processing on the basis of newly acquired information.

2. The techniques that make use of data whose no nominal classification is known in advance to suggest a classification and seek a possible classification (which at the business level is often more interesting): "data mining". We find in this category the, HAC, k-means, kohonen maps, K-modes, K-medoids, etc.

3. The techniques that make use of a mixture of classified and unclassified data and named "semi-supervised learning".

The first two families are quite well summarized by the following figure:

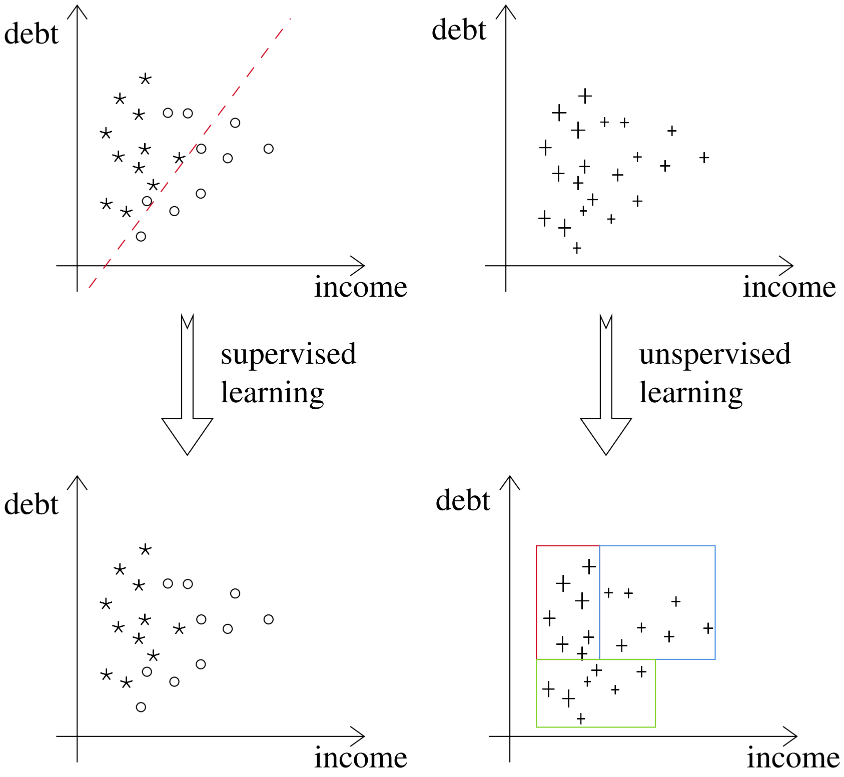


Figure 14.131: **Supervised vs Unsupervised learning idea**

are themselves divided into two sub-families: "kernel-based" (named also "soft clustering" or "fuzzy clustering") or non-kernel based (named also "hard clustering")... (that means for the latter we don't need to make any assumptions on the statistical distribution shape of the variable of interest and that an individual can only belong to one category[58]).

---

[58] Indeed, in fuzzy clustering technique like GMM (Gaussian Mixture Models) an individual has classification weights for each possible class!

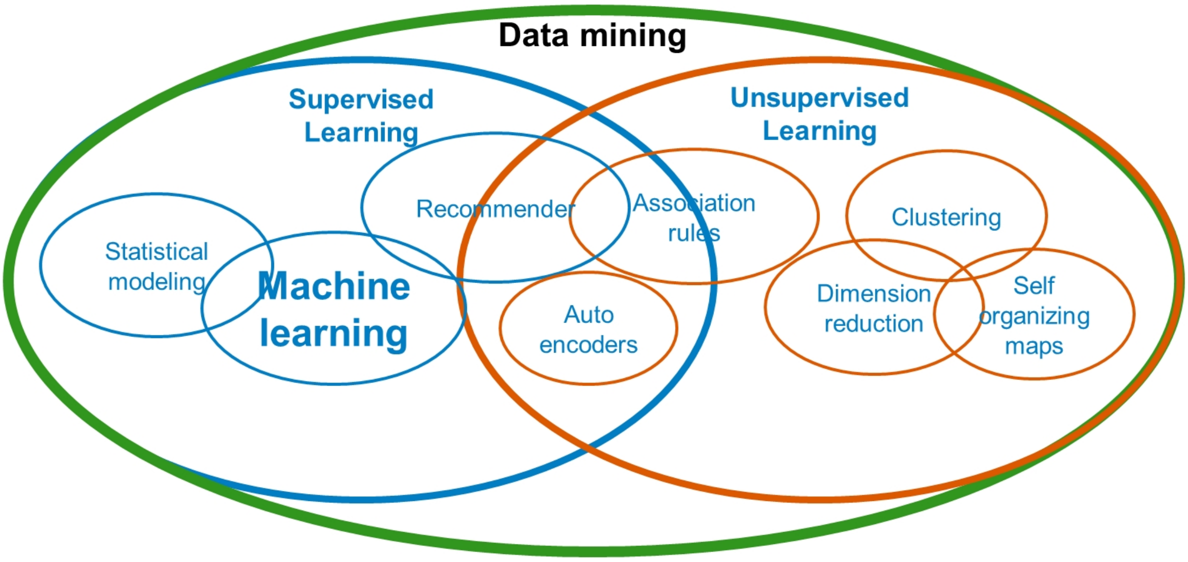


Figure 14.132: **Data Mining as seen by SAS™ (source: SlideShare SAS)**

The distinction between Machine Learning and data mining has become more and more blurred, and there is a great deal of "cross-fertilization".

The industrial or operational use of this knowledge in the professional world can solve very different problems, ranging from customer relationship management to preventive maintenance, through fraud detection and the optimization of websites, the supervision of financial markets, pro-active prospecting (customer consumption preferences), optimization of paths (analysis of road caps and left turns), or target selection (probability of acquiring a new given prospects) anticipate the identification of terrorists, pricing of products in comparison to equivalent one on the mark, etc.

Remark

Be careful to not to confuse strictly speaking the concepts of classification, segmentation and association! Although the first two are often confused (classification / segmentation) because many algorithms do both at once, classification is the prediction of one or more discrete variables based on the value of other fields in the data set as the segmentation divides the data into groups of items having the most identical properties as possible. About association it is clear that many segmentation algorithms shows what is associated with what but the idea strictly speaking of association is to quantified by a scalar the degree of association between two data sets.

We will see here some trivial techniques that we will complete with time. We tried to order them in the ascending order of pedagogical complexity (but that concept is quite subjective obviously that's why many people would argue that k-nn should be presented first)...

14.1.20.5.1 Naive Bayes classifiers

In Machine Learning, "naive Bayes classifiers" are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features (properties).

Naive Bayes has been studied extensively since the 11950s (holocene calendar). It was introduced under a different name into the text retrieval community in the early 11960s (holocene calendar), and remains a popular (baseline) method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features. With appropriate pre-processing, it is competitive in this domain with more advanced methods including support vector machines. It also finds application in automatic medical diagnosis.

14.1.20.5.1.1 Binomial Naive Bayes classifier

The "binomial naive Bayes classifier" consists by using Bayes probabilities to classify an item in two possible outcomes.

For example, the table below consisting of 4 e-mail, having 2 attributes and two possible classification outcomes that are $\{\text{No}-\text{Spam},\text{Spam}\}$:

Email	100% **Guaranteed**	**Meeting**	**Classification**
1	No	Yes	No – Spam
2	No	Yes	No – Spam
3	Yes	Yes	Spam
4	Yes	No	No – Spam

Table 14.31: **Naive Bayes spam-email companion example**

Abstractly, as already mentioned naive Bayes is a conditional probability model: given a problem instance to be classified, represented by a vector $\vec{x}=(x_1,\ldots,x_n)$ representing some n features (independent variables), it assigns a posteriori with probabilities $P(C_k/x_1,\ldots,x_n)$ to the instance C_k, or also written $P(C_k/\vec{x})$, for each of k possible outcomes or classes C_k.

Remark

Don't forget as we have seen in the section Probabilities (see section Probabilities page 497) that $P(C_k/x_1,\ldots,x_n)$ is just a convenient notation to write $P(C_k/x_1\cap\ldots\cap x_n)$. And therefore $P(\vec{x})$ is also a convenient notation for $P(x_1\cap\ldots\cap x_n)$.

With the example above it would look like this for the first e-mail (thus $n=2$):

$$P\left(C_k/\{100\%\text{ Guaranteed}\}=\text{No},\{100\%\text{ Meeting}\}=\text{Yes}\right) \qquad (14.1.1724)$$

The problem with the above formulation is that if the number of features n is large or if a feature can take on a large number of values, then basing such a model on probability tables is very hard. We therefore reformulate the model to make it more tractable. Using Bayes' theorem (see section Probabilities page 492), the conditional probability can be decomposed as:

$$P(C_k/\vec{x})=\frac{P(C_k)P(\vec{x}/C_k)}{P(\vec{x})} \tag{14.1.1725}$$

With for recall:

- $P(C_k/\vec{x})$ is the a posteriori probability to belong the class C_k knowing the features x_i (probability of the hypothetical class given the evidence ie. the features)
- $P(C_k)$ is the a priori probability (or "marginal probability") of the class C_k
- $P(\vec{x})$, the a priori probability of the feature (it is also a marginal probability)
- $P(\vec{x}/C_k)$, the a posteriori probability of the features x_i knowing the class C_k, it is also the likelihood of Y_k for $\vec{x}_i$ known (probability of evidence, i.e.. the feature, given that the hypothetical class holds).

In plain English, using Bayesian probability terminology, the above relation can be written for recall as:

$$\text{posterior}=\frac{\text{priori}\cdot\text{likelihood}}{\text{evidence}} \tag{14.1.1726}$$

In practice, there is interest only in the numerator of that fraction (equivalent to the joint probability), because the denominator does not depend on C_k and the values of the features x_i are given, so that the denominator is effectively constant whatever the k.

Hence the fact that we often can see in textbooks the following relation:

$$P(C_k/\vec{x})\propto P(C_k)P(\vec{x}/C_k) \tag{14.1.1727}$$

Theorem 14.8. A key idea in the NB model is the following assumption:

$$P(C=C_k/X_1=x_1,\ldots,X_n=x_n)\propto P(C=C_k)\prod_{i=1}^{n}P(X_i=x_i/C=C_k) \tag{14.1.1728}$$

Proof 14.8.1. This equality is derived using iteratively (don't forget that we don't care about the numerator!) the relation proved in the section of Probabilities but without the denominator (see section Probabilities page 492):

$$P(C_k/\vec{x})\propto P(C_k)P(\vec{x}/C_k) \tag{14.1.1729}$$

Indeed, let us rewrite this relation more explicitly:

$$P(C=C_k/X_1=x_1,\ldots,X_n=x_n)\propto P(C=C_k)P(X_1=x_1\ldots,X_n=x_n/C=C_k) \tag{14.1.1730}$$

Next we rewrite the second product term using the assumption of conditional independence:

$$P(X_1=x_1,\ldots,X_n=x_n/C=C_k)=P(X_1=x_1|C=C_k)\ldots P(X_n=x_n/C=C_k)=\prod_{i=1}^{n}P(X_i=x_i/C=C_k) \tag{14.1.1731}$$

Hence reinjecting in the prior-previous relation:

$$\begin{aligned}P(C=C_k/X_1=x_1,\ldots,X_n=x_n)&\propto P(C=C_k)P(X_1=x_1\ldots,X_n=x_n|C=C_k)\\&=P(C=C_k)\prod_{i=1}^{n}P(X_i=x_i/C=C_k)\end{aligned} \tag{14.1.1732}$$

In a condensed form the posterior probability of belonging to C_k knowing $\vec{x}$ is then given by:

$$\boxed{P(C_k/\vec{x})\propto P(C_k)\prod_{i=1}^{n}P(\vec{x}/C_k)} \tag{14.1.1733}$$

■ Q.E.D.

The idea now, when we have a training set of size N, is just to take for estimators:

$$\hat{P}(C_k)=\frac{\displaystyle\sum_{i=1}^{N}(\#C_i=C_k)}{N}=\frac{\text{count}(C_k)}{N} \tag{14.1.1734}$$

where obviously we define $C=C_k$ to be 1 if $C=C_k$, 0 otherwise. This estimator is simple the number of times that the label C_k is seen in the training set!

Similarly:

$$\hat{P}(\vec{x}/C_k)=\frac{\displaystyle\sum_{i=1}^{N}\#\left[(C=C_k)\cap(\vec{x}_i=\vec{x})\right]}{\displaystyle\sum_{i=1}^{N}(\#C_i=C_k)}=\frac{\text{count}(\vec{x}|C_k)}{\text{count}(C_k)} \tag{14.1.1735}$$

This is a very natural estimate: we simply count the number of times label C_k is seen in conjunction with $\vec{x}_i$ taking value $\vec{x}$ that we divide by the number of times the label C_k is seen in total.

✍**Example:**

We consider the following training set:

Email	100% **Guaranteed**	**Meeting**	**Classification**
1	No	Yes	No – Spam
2	No	Yes	No – Spam
3	Yes	Yes	Spam
4	Yes	No	No – Spam

A new e-mail arrives and contains the expression "100% *Guaranteed*" and the expression "*Meeting*". We want to predict if it is a Spam or not?

Email	100% Guaranteed	Meeting	Classification
5	Yes	Yes	?

For this purpose we will first simplify the notation by denoting the expression "100% *Guaranteed*" by the letter G and the expression "*Meeting*" by the letter M.

The posterior probabilities are in this simple case given by Bayes theorem:

$$\begin{aligned} P(\text{Spam}/G=\text{Yes},M=\text{Yes}) &= \frac{P(G=\text{Yes},M=\text{Yes}/\text{Spam})\cdot P(\text{Spam})}{P(G=\text{Yes},M=\text{Yes})} \\ P(\text{No}-\text{Spam}/G=\text{Yes},M=\text{Yes}) &= \frac{P(G=\text{Yes},M=\text{Yes}/\text{No}-\text{Spam})\cdot P(\text{No}-\text{Spam})}{P(G=\text{Yes},M=\text{Yes})} \end{aligned} \tag{14.1.1736}$$

Hence the ratio, that eliminates (as we know) the denominator that is then useless, that must be greater that 1 if the e-mail is a Spam:

$$\frac{P(G=\text{Yes},M=\text{Yes}/\text{Spam})\cdot P(\text{Spam})}{P(G=\text{Yes},M=\text{Yes}/\text{No}-\text{Spam})\cdot P(\text{No}-\text{Spam})}>1 \tag{14.1.1737}$$

According to the assumption of the independence of conditional probabilities we can write for the first factor of the numerator and respectively for that of the denominator:

$$P(G=\text{Yes},M=\text{Yes}/\text{Spam})=P(G=\text{Yes}/\text{Spam})\cdot P(M=\text{Yes}/\text{Spam}) \tag{14.1.1738}$$

$$P(G=\text{Yes},M=\text{Yes}/\text{No}-\text{Spam})=P(G=\text{Yes}/\text{No}-\text{Spam})\cdot P(M=\text{Yes}/\text{No}-\text{Spam}) \tag{14.1.1739}$$

So we can rewrite the ratio as:

$$\frac{P(G=\text{Yes}/\text{Spam})\cdot P(M=\text{Yes}/\text{Spam})\cdot P(\text{Spam})}{P(G=\text{Yes}/\text{No}-\text{Spam})\cdot P(M=\text{Yes}/\text{No}-\text{Spam})\cdot P(\text{No}-\text{Spam})}>1 \tag{14.1.1740}$$

And as we have in our training set:

$$\begin{aligned} P(G=\text{Yes}/\text{Spam}) &= \frac{1}{1} & P(G=\text{Yes}/\text{No}-\text{Spam}) &= \frac{1}{3} \\ P(M=\text{Yes}/\text{Spam}) &= \frac{1}{1} & P(M=\text{Yes}/\text{No}-\text{Spam}) &= \frac{2}{3} \\ P(\text{Spam}) &= \frac{1}{4} & P(\text{No}-\text{Spam}) &= \frac{3}{4} \end{aligned} \tag{14.1.1741}$$

Injecting in the previous ratio, we get:

$$\frac{\frac{1}{1}\cdot\frac{1}{1}\cdot\frac{1}{4}}{\frac{1}{3}\cdot\frac{2}{3}\cdot\frac{3}{4}}=1.5>1 \tag{14.1.1742}$$

That's all for this naive Bayesian example for document classification problem.

Statisticians are somewhat disturbed by use of the NBC (which they dub Idiot's Bayes) because the naive assumption of independence is almost always invalid in the real world.

However, the method has been shown to perform surprisingly well in a wide variety of contexts.

14.1.20.5.1.2 Gaussian Naive Bayes classifier

When dealing with continuous data, a typical assumption is that the continuous values associated with each class are distributed according to a Gaussian distribution. For example, suppose the training data contains a continuous attribute, x. We first segment the data by the class, and then compute the mean $\hat{\mu}$ and standard deviation $\hat{\sigma}$ of x for each class.

Let $\hat{\mu}_k$ be the mean of the values in x associated with class C_k, and let $\hat{\sigma}_k$ be the standard deviation of the values in x associated with class C_k.

Suppose we have collected some new observation value x'. Then, the probability density of x' given a class C_k, $P(x=x'/C_k)$, can be computed by plugging x' into the equation for a Normal distribution parametrized by $\hat{\mu}_k$ and $\hat{\sigma}_k^2$. That is:

$$P(x=x'/C_k)=\frac{1}{\sqrt{2\pi\hat{\sigma}_k^2}}e^{-\frac{(x'-\hat{\mu}_k)^2}{2\hat{\sigma}_k^2}} \tag{14.1.1743}$$

And after we apply again:

$$\boxed{P(C_k/\vec{x})\propto P(C_k)\prod_{i=1}^{n}P(\vec{x}/C_k)} \tag{14.1.1744}$$

✍**Example:**

We consider the following training set:

Person	Height [cm]	Weight [kg]	Foot size [cm]
male	182.88	81.64	30.48
male	180.44	86.18	27.94
male	170.07	77.11	30.48
male	180.44	74.84	25.4
female	152.4	45.35	15.24
female	167.64	68.03	20.32
female	165.20	58.96	17.78
female	175.26	68.03	22.86

The classifier created from the training set using a Gaussian distribution assumption would be (given variances are unbiased sample variances):

Person	Mean Height	Variance Height	Mean Weight	Variance Weight	Mean Foot size	Variance Foot size
male	178.45	32.59	79.94	25.28	28.58	5.91
female	165.13	90.32	60.09	114.88	19.05	10.75

We also know from the training set that we have equiprobable classes so:

$$P(\text{male})=P(\text{female})=0.5 \tag{14.1.1745}$$

Now consider a new sample to be classified as male or female:

Person	Height [cm]	Weight [kg]	Foot size [cm]
?	182.88	58.96	20.32

We wish to determine which posterior is greater, male or female. For the classification as male the posterior is given according to our boxed relation above:

$$P(C_k=\text{male}/\vec{x})\propto P(\text{male})\;P(\text{height}/\textbf{male})\;P(\text{weight}/\text{male})\;P(\text{footsize}/\text{male}) \tag{14.1.1746}$$

For the classification as female the posterior is given by:

$$P(C_k=\text{female}/\vec{x})\propto P(\text{female})\;P(\text{height}/\textbf{female})\;P(\text{weight}/\text{female})\;P(\text{footsize}/\text{female}) \tag{14.1.1747}$$

We now determine the probabilities distribution for the sex of the sample:

- $P(\text{male})=0.5$
- $P(\text{female})=0.5$
- $P(\text{height}/\text{male})=\dfrac{1}{\sqrt{2\pi\hat{\sigma}_{H,M}^2}}\exp\left(\frac{-(182.88-\hat{\mu}_{H,M})^2}{2\hat{\sigma}_{H,M}^2}\right)\cong 0.0517$
- $P(\text{height}/\text{female})=\dfrac{1}{\sqrt{2\pi\hat{\sigma}_{H,F}^2}}\exp\left(\frac{-(182.88-\hat{\mu}_{H,F})^2}{2\hat{\sigma}_{H,F}^2}\right)\cong 0.0073$
- $P(\text{weight}/\text{male})=\dfrac{1}{\sqrt{2\pi\hat{\sigma}_{W,M}^2}}\exp\left(\frac{-(58.96-\hat{\mu}_{W,M})^2}{2\hat{\sigma}_{W,M}^2}\right)\cong 1.3142\cdot 10^{-05}$
- $P(\text{weight}/\text{female})=\dfrac{1}{\sqrt{2\pi\hat{\sigma}_{W,M}^2}}\exp\left(\frac{-(58.96-\hat{\mu}_{W,M})^2}{2\hat{\sigma}_{W,M}^2}\right)\cong 0.0370$
- $P(\text{foot size}/\text{male})=\dfrac{1}{\sqrt{2\pi\hat{\sigma}_{FS,M}^2}}\exp\left(\frac{-(20.32-\hat{\mu}_{FS,M})^2}{2\hat{\sigma}_{FS,M}^2}\right)\cong 0.0005$

- $P(\text{foot size}/\text{female})=\dfrac{1}{\sqrt{2\pi\hat{\sigma}^2_{FS,F}}}\exp\left(\frac{-(20.32-\hat{\mu}_{FS,F})^2}{2\hat{\sigma}^2_{FS,F}}\right)\cong 0.1128$

Hence:

$$\begin{aligned}P(C_k=\text{male}/\vec{x})&\propto P(\text{male})\,P(\text{height}/\text{male})\,P(\text{weight}/\text{male})\,P(\text{footsize}/\text{male})\\&\cong 1.6986\cdot 10^{-10}\\P(C_k=\text{female}/\vec{x})&\propto P(\text{female})\,P(\text{height}/\text{female})\,P(\text{weight}/\text{female})\,P(\text{footsize}/\text{female})\\&\cong 1.5233\cdot 10^{-5}\end{aligned} \tag{14.1.1748}$$

Since posterior numerator is greater in the female case, we predict the new individual is female.

Most software multiplyes these two probabilities by a factor f such that:

$$f\cdot(1.6986\cdot 10^{-10}+1.5233\cdot 10^{-5})=1$$

It is easy to find that in this special case $f\cong 65646.21869$ and therefore we get:

$$P(C_k=\text{male}/\vec{x})\cong 1.11507\cdot 10^{-5}\qquad\text{and}\qquad P(C_k=\text{female}/\vec{x})\cong 0.999988849$$

14.1.20.5.2 k-nearest neighbours classifier (k-nn)

The "k-nn" is a quite simple supervised classification technique (do not confuse with the k-means that is a unsupervised clustering technique and that we will see later!). It takes a bunch of labelled points and uses them to learn how to label other points. To label a new point, it looks at the labelled points closest to that new point (those are its nearest neighbours), and has those neighbours vote, so whichever label the most of the neighbours have is the label for the new point (the "k" is the number of neighbours it checks).

In this framework, we have a learning database consisting of N input-output pairs. To estimate the output associated with a new input x, the k nearest neighbours method consists in taking into account (identically) the k learning samples whose input is closest to the new input x, according to a distance to be defined.

For example, in a classification problem, we will retain the class most represented among the k outputs associated with the k inputs closest to the new input x.

Given D the dataset composed of n pairs $(\vec{x},y)$, with $\vec{x}$ the description of an observation according to d quantitative variables, in the form of a real vector of dimensions d, and y the class of this observation. Consider the case where the descriptive variables are all quantitative and where the class variable have q modalities:

$$D=\{(\vec{x}_1,y_1),(\vec{x}_2,y_2),\ldots,(\vec{x}_n,y_n)\}\qquad\forall(x_i,y_i)_{1\leq i\leq n}:\vec{x}_i\in\mathbb{R}^d,y_i\in\{0,1,\ldots,q\} \tag{14.1.1749}$$

The space of representation of the observations must be provided with a distance necessary for the construction of the neighbourhood of a new observation. The distance is a function $d:\mathbb{R}^d\times\mathbb{R}^d\mapsto\mathbb{R}^+$, respecting the three axioms of the distance that we know well (see section Topology page 2450).

The most common distances for vectors of dimension n (corresponding therefore to the number the input attributes that must all belong to $\mathbb{R}$) in a table of size N (be careful! we know that computer scientists use some distances that doesn't respect the four axioms of distances and especially the triangular inequality as already discussed in the section Analysis at page 2454):

- Minkowski distance:

$$d_{\text{M},r}(\vec{x},\vec{y})=\sqrt[r]{\sum_{i=1}^{n}|x_i-y_i|^r} \tag{14.1.1750}$$

- Euclidean distance:

$$d_e(\vec{x},\vec{y})=d_{\text{M},2}(\vec{x},\vec{y})=\sqrt{\sum_{i=1}^{n}|x_i-y_i|^2} \tag{14.1.1751}$$

- Manhattan distance (city-block):

$$d_{\text{cb}}(\vec{x},\vec{y})=d_{\text{M},1}(\vec{x},\vec{y})=\sum_{i=1}^{n}|x_i-y_i| \tag{14.1.1752}$$

- Quadratic distance (σ is as we know a correlation matrix):

$$d_{\text{Q}}(\vec{x},\vec{y})=(\vec{x}-\vec{y})^T\Sigma(\vec{x}-\vec{y}) \tag{14.1.1753}$$

  as we can see it corresponds to the square of the Mahalanobis distance.

- Correlation distance:

$$d_\rho(\vec{x},\vec{y})=\frac{\displaystyle\sum_{i=1}^{n}(x_i-\bar{x}_i)(y_i-\bar{y}_i)}{\sqrt{\displaystyle\sum_{i=1}^{n}(x_i-\bar{x}_i)^2\sum_{i=1}^{n}(y_i-\bar{y}_i)^2}} \tag{14.1.1754}$$

- Chi-2 distance:

$$d_\chi(\vec{x},\vec{y})=\left(\sum_{i=1}^{n}\frac{1}{\displaystyle\sum_{i=1}^{n}(x_i+y_i)}\left(\frac{x_i}{\displaystyle\sum_{i=1}^{n}x_i}-\frac{y_i}{\displaystyle\sum_{i=1}^{n}y_i}\right)\right)^2 \tag{14.1.1755}$$

- Kendall rank correlation distance:

$$d_\kappa(\vec{x},\vec{y})=1-\frac{2}{n(n-1)}\sum_{i=1}^{m}\sum_{j=1}^{i-1}\text{sgn}(x_i-x_j)\text{sgn}(y_i-y_j) \tag{14.1.1756}$$

  with for recall:

$$\text{sgn}(x)=\begin{cases}-1 & x<0\\ 0 & x=0\\ +1 & x>0\end{cases} \tag{14.1.1757}$$

- Heterogeneous Euclidean Overlap Metric (HEOM) by focusing only on continuous attributes (case of the Tanagra software):

$$d_{\text{HEOM}}(\vec{x},\vec{y})=\sqrt{\sum_i\left(\frac{|x_i-y_i|}{\max_{i,a}-\min_{i,a}}\right)^2}=\sqrt{\sum_i\left(\frac{x_i-y_i}{\max_{i,a}-\min_{i,a}}\right)^2} \tag{14.1.1758}$$

  where $\max_{i,a}$ and $\min_{i,a}$ are respectively the largest and smallest value of the attribute i throughout the training sample so that the term in the parenthesis is always between 0 and 1 (normalized in other words...).

The figures below illustrate how the k nearest neighbours solves a binary supervised classification task ($y_i\in\{0,1\}$) in $\mathbb{R}^2$, depending on the Euclidean distance and for $k=3$. The first figure show (in three steps, from left to right) a new observation is attributed to class 1, its neighbour being composed of three observations belonging to class 1. The second figure shows how a new observation is attributed to class 1, its vicinity being composed of two observations belonging to the class 1 and of one observation belonging to the class 0.

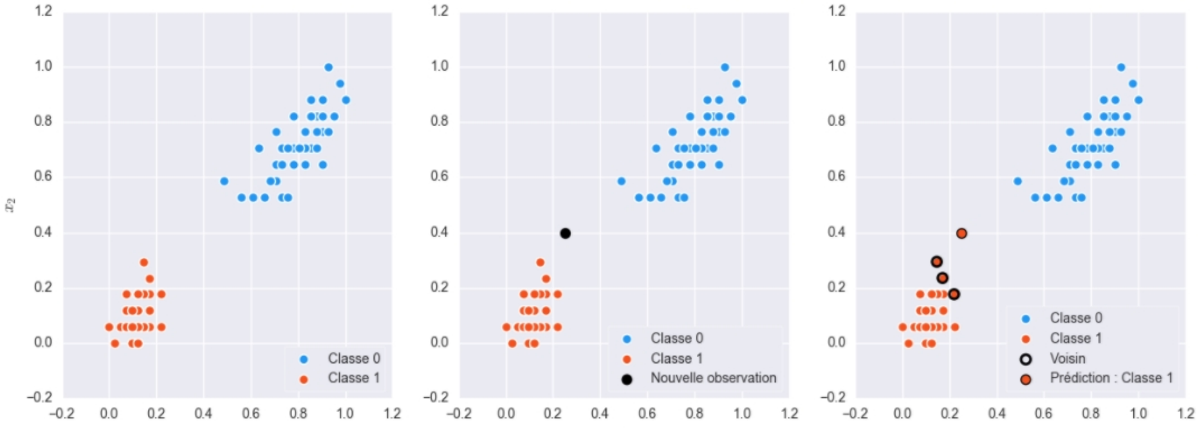


Figure 14.133: **Classification of a new observation with $k-nn$: the three nearest neighbours belong to class 1**

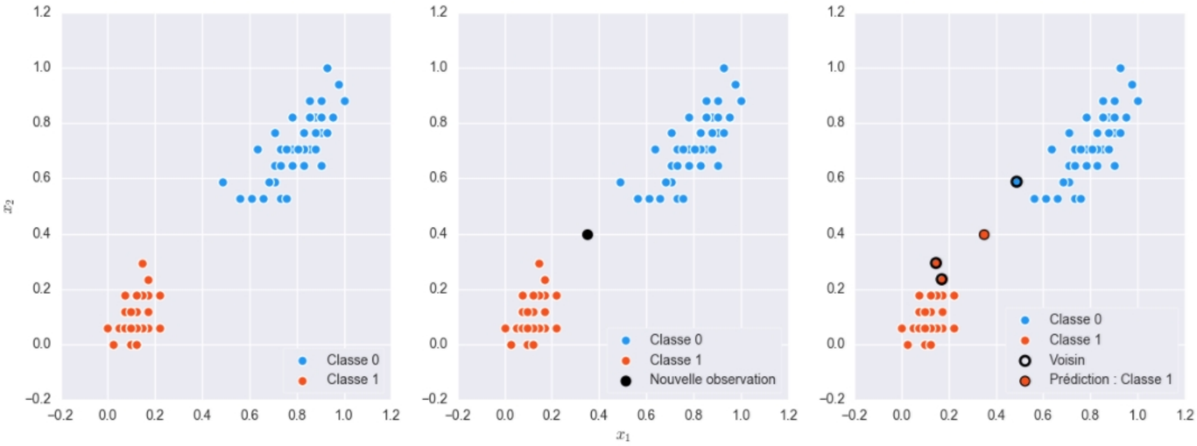


Figure 14.134: Classification of a new observation with $k-nn$: 2 of the 3 nearest neighbours belong to class 1

The distance to the kth nearest neighbour can also be seen as a local density estimate and thus is also a popular outlier score in anomaly detection. The larger the distance to the k-nn, the lower the local density, the more likely the query point is an outlier. Although quite simple, this outlier model, along with another classic data mining method, local outlier factor, works quite well also in comparison to more recent and more complex approaches, according to a large scale experimental analysis.

Remark

Obviously we may run the k-nn on the factorial space obtained after having run a PCA (Principal Component Analysis as seen on page 1213) on our data. We can also even reduce the dimensions using that latter if necessary!

14.1.20.5.3 Regression and classification trees

"Classification and Regression Tree" (CART) are a set of heuristic supervised algorithms introduced in 11984 (holocene calendar) by Leo Breiman, Jerome Friedman, Richard Olshen and Charles Stone as an umbrella term to refer to the following types of decision trees[59]":

- "Classification Trees": where the target variable is categorical and the tree is used to identify the "class" within which a target variable would likely fall into.
- "Regression Trees": where the target variable is continuous and tree is used to predict it's value.

Decision tree are non-linear, nonparametric and nonmetric classifier. Concretely, decision trees are predictive models that proceed through a tree structure of rules or a hierarchy of tests extracted from learning or training data, to partition the basic data into homogeneous subgroups from the point of view of the variable to predict. The generated rules tree easily lends itself to the human reasoning mode and it becomes more intuitive to interpret the tree and make predictions.

---

[59] A "decision stump" is a Machine Learning model consisting of a one-level decision tree

These trees are widely used in advanced marketing or social science to discriminate (categorize) a very large population. Obviously, these algorithms (that are part of hierarchical techniques) will never do better than a human being (at least in this first decade of the 121st century according to holocene calendar)... but also ask an employee to create groups in a population of 5 million customers on the basis of ten explanatory variables. You will have to wait for the response a quite long time...

While these automated classification techniques are very useful in the above situations, they nevertheless have a major problem that makes we will not focus too much on this subject:

- These techniques are very sensitive to the analysed population and give very different results.
- The various existing techniques of classifications give results that are completely different for the same population.
- Decision trees can also never predict values out of the range of the targets seen in the training dataset (i.e. classical decision trees don't extrapolate, they only retrodict).

It is better to be cautious about the conclusions that we can draw from these models and compare the results of several methods depending on the return on experience choose the one that seems the best suited.

Remark

In good practice situations, the construction of a tree is, as always in the field of Machine Learning (...), made from a set of data named the "learning sample". Once the tree is built, it is tested for its predictive power usually on a set of data named the "test sample".

14.1.20.5.3.1 Binary classification trees with quantitative variables

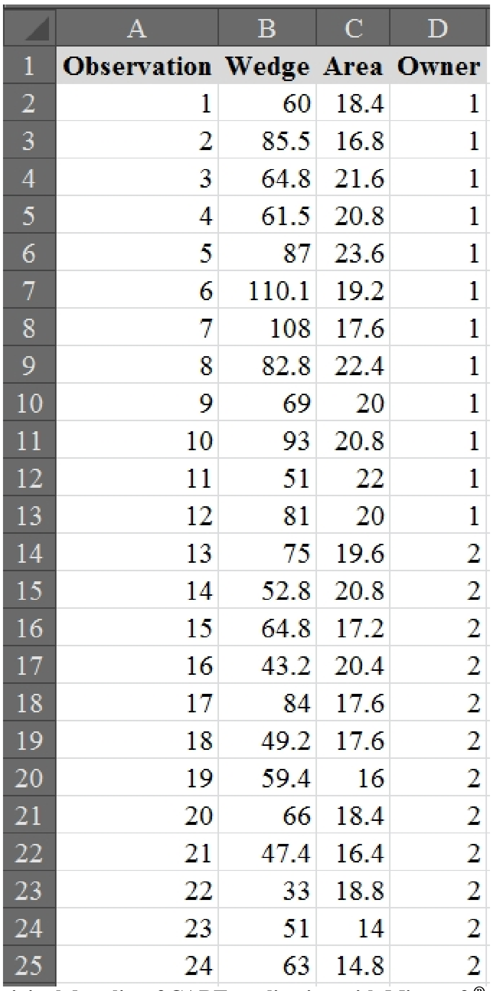
For companion example, let us consider a set of categorical variables $x_1,\ldots,x_p$. The recursive partitioning has for purpose to divide the p variables of the space into rectangles which do not overlap.

For example, consider the variable x_i and a value s_i of this variable, we find that the partitioning $x_i<s_i$ and $s_i<x_i$ separates well the data into two disjoint sets. Then one of the parts is in turn divided by a value x_i or by the value of another variable. We end then with three rectangles and so on...

The idea is to create n rectangles such that all data contained in a rectangle are homogeneous (that is to say contains only one family of points).

To address this issue, consider the following first practical case with only quantitative variables (it is therefore a binary regression tree):

A dealer would like for his city to find a way to classify the families that are able to buy a car (owners) and those that are not ready to buy a car (non-owners). A sample of 12 owners ("1" in the figure below) and 12 non-owners ("2" in the figure below) is selected. The two independent variables are x_1 (wage in kilo-dollars) and x_2 (area of their home). In other words, we have here a case where $x_i \in \mathbb{R}^+$ and $y_i\{1,2\}$.



	A	B	C	D
1	**Observation**	**Wedge**	**Area**	**Owner**
2	1	60	18.4	1
3	2	85.5	16.8	1
4	3	64.8	21.6	1
5	4	61.5	20.8	1
6	5	87	23.6	1
7	6	110.1	19.2	1
8	7	108	17.6	1
9	8	82.8	22.4	1
10	9	69	20	1
11	10	93	20.8	1
12	11	51	22	1
13	12	81	20	1
14	13	75	19.6	2
15	14	52.8	20.8	2
16	15	64.8	17.2	2
17	16	43.2	20.4	2
18	17	84	17.6	2
19	18	49.2	17.6	2
20	19	59.4	16	2
21	20	66	18.4	2
22	21	47.4	16.4	2
23	22	33	18.8	2
24	23	51	14	2
25	24	63	14.8	2

Figure 14.135: **Original data list of CART application with Microsoft® Excel® 14.0.7172**

We see that we have as many owners than non-owners (assumed equal appearance frequency in the whole population). Therefore the probability of belonging to a class is 50%.

Or graphically:

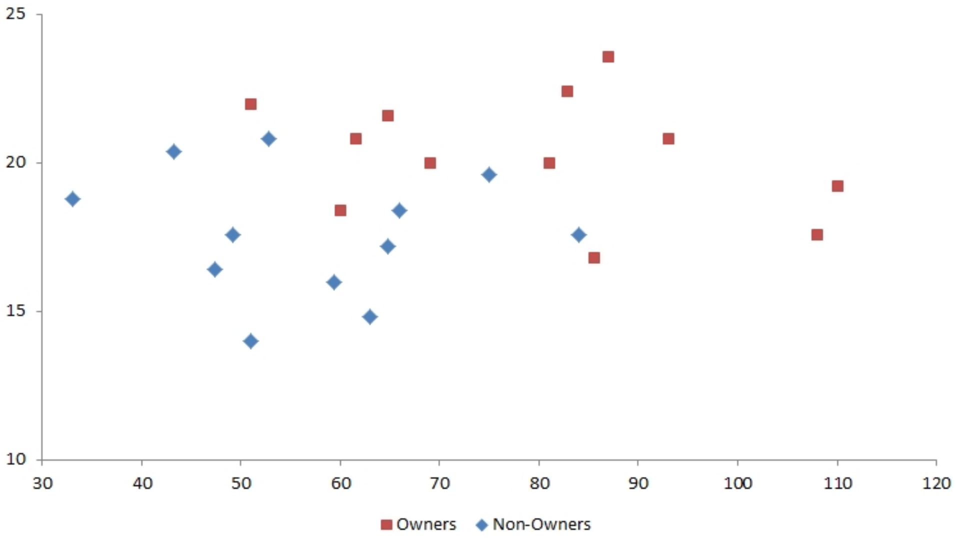


Figure 14.136: **Original plot CART application with Microsoft® Excel® 14.0.7172**

If we apply the CART algorithm on this data, we see that we must choose x_2 (Area) as the first choice of division with the division value of 19 (we will justify why!). The space (x_1, x_2) is now divided into two rectangles (it was easy to guess that discriminating step without even using mathematics):

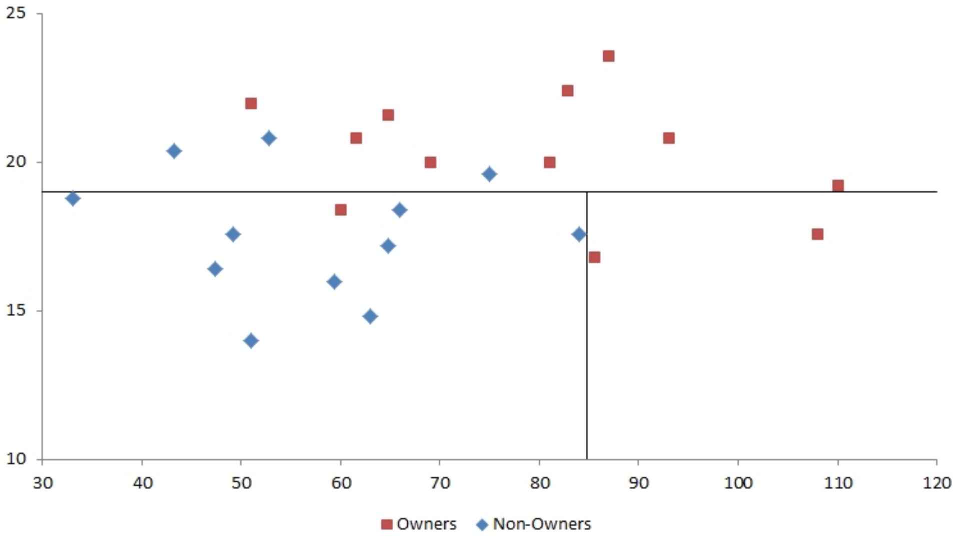


Figure 14.137: **First iteration of CART application with Microsoft® Excel® 14.0.7172**

Notice how the division into two rectangles created two zones (splits) which are more homogeneous than the original graph! The upper rectangle contains points which are more Owners while the lower rectangle contains more Non-owners.

To determine this division, the CART algorithm examines each variable and all possible values for each variable division in order to find the best division.

Thus, the possible points of division for x_1 are (notice that this is every time the average of the two related values in the table):

$$x_1 : \{38.1, 45.3, 50.1, \ldots, 109.5\}$$

and that for x_2 are:

$$x_2 : \{14.4, 15.4, 16.2, \ldots, 23\}$$

These points are ordered by the algorithm according to the way they reduce "impurity" (heterogeneous of composition) in the rectangle that generates the "split".

There are a large number of empirical ways to measure the impurity. But to start with our first example, let us use the easiest and most common "impurity measurement criterion" or "segmentation criteria" so far that is the use of an indicator inspired by the Gini coefficient (see section Quantitative Management Techniques page 8099). Thus, if we denote the classes by $k = 1, 2, 3, \ldots, C$ where C is the total number of classes to be predicted, the "Gini impurity index" for the A rectangle is defined by:

$$\boxed{I(A) = 1 - \sum_{k=1}^{C} p_k^2} \tag{14.1.1759}$$

where p_k is the fraction of observations in the rectangle A which belong to the class k.

In our example, we always have only two classes: Owners / Non-Owners.

Next, the global Gini index is defined as the weighted average of the Gini indices.

So in our example, we have two classes, therefore $C = 2$. Before the separation, we have:

$$I(A) = 1 - \sum_{k=1}^{C} p_k^2 = 1 - \left(\frac{12}{24}\right)^2 - \left(\frac{12}{24}\right)^2 = 1 - 2(0.5)^2 = 0.5 \tag{14.1.1760}$$

The separation found in 19 (see the second figure) gives for example for the top rectangle of the first subdivision:

$$\begin{aligned} I(A_{\text{sup}}) &= 1 - \sum_{k=1}^{2} p_k^2 = 1 - \left[\left(\frac{\text{blue points}}{n}\right)^2 + \left(\frac{\text{red points}}{n}\right)^2\right] \\ &= 1 - \left[\left(\frac{3}{12}\right)^2 - \left(\frac{9}{12}\right)^2\right] = 0.375 \end{aligned} \tag{14.1.1761}$$

For the inferior part:

$$\begin{aligned} I(A_{\text{inf}}) &= 1 - \sum_{k=1}^{2} p_k^2 = 1 - \left[\left(\frac{\text{blue points}}{n}\right)^2 + \left(\frac{\text{red points}}{n}\right)^2\right] \\ &= 1 - \left[\left(\frac{9}{12}\right)^2 - \left(\frac{3}{12}\right)^2\right] = 0.375 \end{aligned} \tag{14.1.1762}$$

By the hazard of the choice of this example, the impurity is the same for both rectangles (top and bottom).

The overall Gini index is then given by:

$$I(A) = \frac{12}{24}0.375 + \frac{12}{24}0.375 = 0.375 \tag{14.1.1763}$$

Notice before continuing that if the subdivision is perfect (only have one family of points in one of the boxes), then we have:

$$\begin{aligned}I(A_{\text{inf,sup}})&=1-\sum_{k=1}^{2}p_k^2=1-\left[\left(\frac{\text{blue points}}{n}\right)^2+\left(\frac{\text{red points}}{n}\right)^2\right]=1-\left[\left(\frac{12}{12}\right)^2+\left(\frac{0}{12}\right)^2\right]\\&=1-\left[\left(\frac{\text{blue points}}{n}\right)^2+\left(\frac{\text{red points}}{n}\right)^2\right]=1-\left[\left(\frac{0}{12}\right)^2+\left(\frac{12}{12}\right)^2\right]\\&=0\end{aligned}\tag{14.1.1764}$$

So the impurity is zero ... And if all the points appear in equal proportions in each of the rectangles (worst situation we could say), the value is then:

$$\begin{aligned}I(\text{A}_{\text{inf,sup}})&=1-\sum_{k=1}^{2}p_k^2\\&=1-\left[\left(\frac{\text{blue points}}{n}\right)^2+\left(\frac{\text{red points}}{n}\right)^2\right]=1-\left[\left(\frac{x^{2}}{n_1}+\frac{y}{n_2}\right)^2\right]\\&\underset{\substack{x=y\\n_1=n_2=n}}{=}1-\left[\left(\frac{x}{n}\right)^2+\left(\frac{x}{n}\right)^2\right]=1-2\left(\frac{x}{n}\right)^2\underset{x=\frac{1}{2}n}{=}=1-2\left(\frac{1}{2}\right)^2\\&=1-2\frac{1}{4}=1-\frac{1}{2}=0.5\end{aligned}\tag{14.1.1765}$$

If we generalize to C classes (C spatial dimensions), it comes immediately:

$$I(A_{\text{inf,sup}})=1-\sum_{k=1}^{C}p_k^2=1-\frac{1}{C}=\frac{C-1}{C}\tag{14.1.1766}$$

which is the maximum impurity.

So the impurity is always defined by a value in the range:

$$I(A)\in\left[0,\frac{C-1}{C}\right]\tag{14.1.1767}$$

Now, to continue with our example, even without using a computer algorithm, without even calculating the impurity, it is relatively easy to guess which will be the next discrimination: it will be $x_1=84.75$ (Income). Which will give:

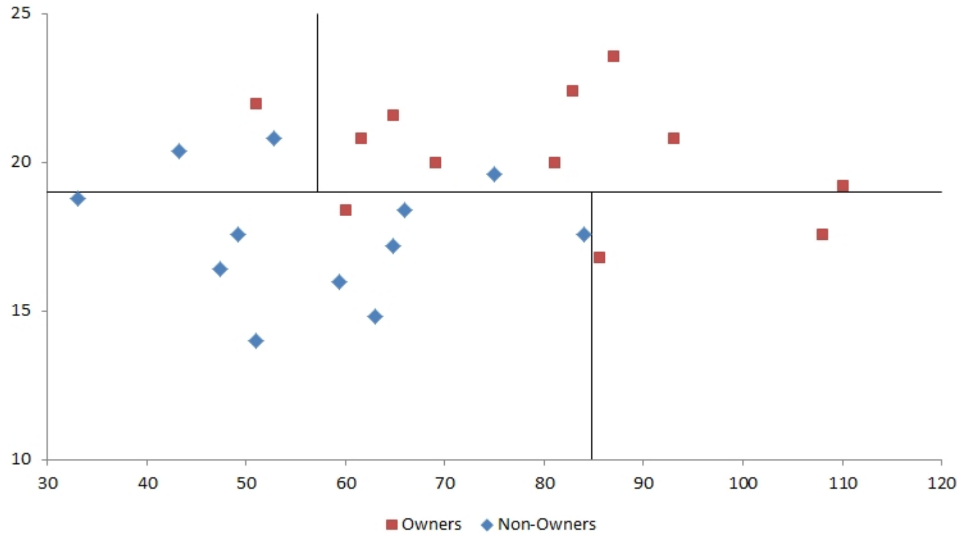


Figure 14.138: **Second iteration of CART application with Microsoft® Excel® 14.0.7172**

What was also easy to guess even without using the calculations (try your friends, you will see that very often they can find the first two discrimination).

The impurity will be calculated in the new discriminated area by:

$$\begin{aligned}I(A_{\text{left}})&=1-\sum_{k=1}^{2}p_k^2=1-\left[\left(\frac{\text{blue points}}{n}\right)^2+\left(\frac{\text{red points}}{n}\right)^2\right]\\&=1-\left[\left(\frac{9}{10}\right)^2+\left(\frac{1}{10}\right)^2\right]=0.18\\I(A_{\text{right}})&=1-\sum_{k=1}^{2}p_k^2=1-\left[\left(\frac{\text{blue points}}{n}\right)^2+\left(\frac{\text{red points}}{n}\right)^2\right]\\&=1-\left[\left(\frac{0}{2}\right)^2+\left(\frac{2}{2}\right)^2\right]=0\end{aligned} \tag{14.1.1768}$$

The overall Gini index is given by:

$$I(A)=\frac{10}{12}\cdot 0.18+\frac{2}{2}\cdot 0=0.15 \tag{14.1.1769}$$

We continue, but be aware that the result is less easy to guess.

The majority of individuals interviewed are wrong without using the mathematical definition of impurity and intuitively wrong to propose one or more of the following discrimination:

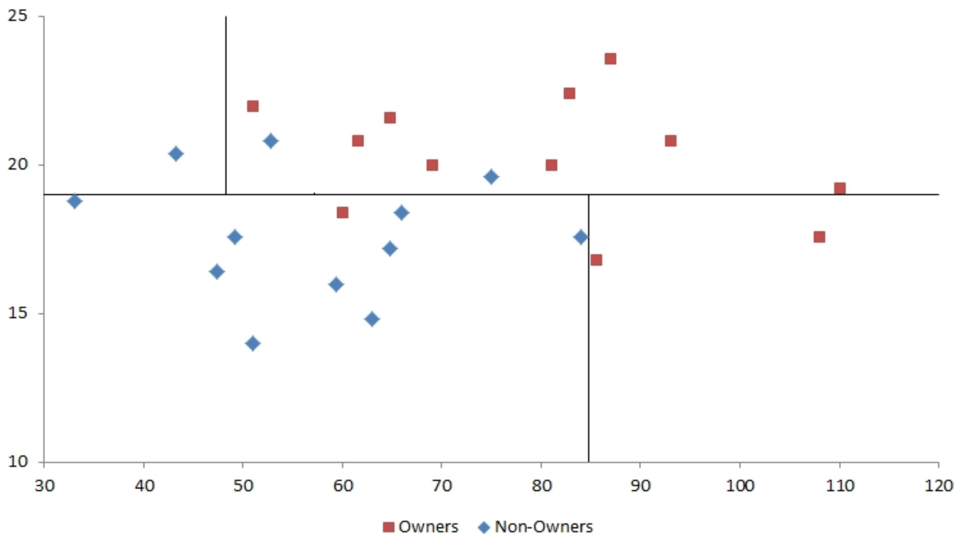


Figure 14.139: Third iteration of CART application with Microsoft® Excel® 14.0.7172

with (you can do the math) a total impurity of 0.2727. When in reality, the optimum discrimination is:

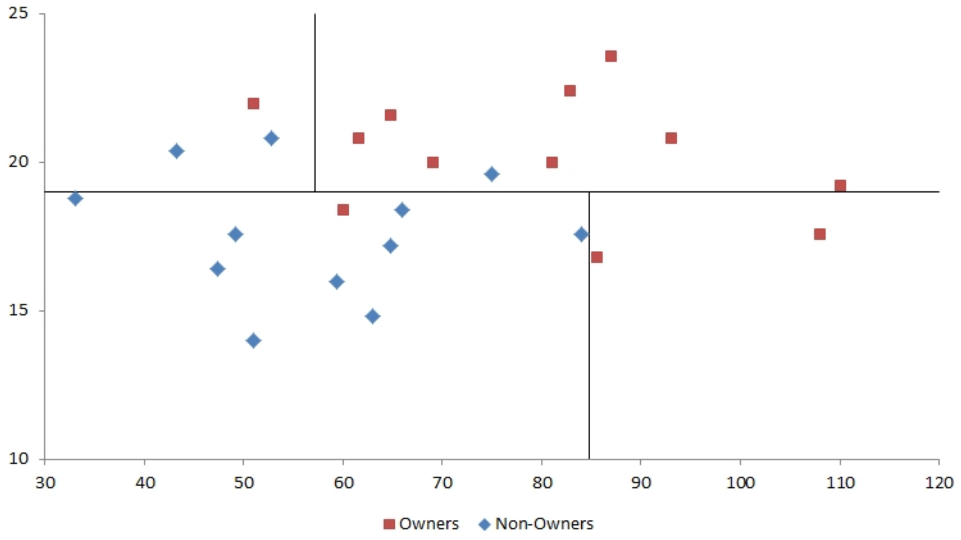


Figure 14.140: Third real iteration of CART application with Microsoft® Excel® 14.0.7172

with a total impurity of 0.2592. Indeed:

$$\begin{aligned}
I(A_{\text{left}})&=1-\sum_{k=1}^{2}p_k^2=1-\left[\left(\frac{\text{blue points}}{n}\right)^2+\left(\frac{\text{red points}}{n}\right)^2\right]=1-\left[\left(\frac{1}{3}\right)^2+\left(\frac{2}{3}\right)^2\right]=0.\bar{4}\\
I(A_{\text{right}})&=1-\sum_{k=1}^{2}p_k^2=1-\left[\left(\frac{\text{blue points}}{n}\right)^2+\left(\frac{\text{red points}}{n}\right)^2\right]=1-\left[\left(\frac{1}{9}\right)^2-\left(\frac{8}{9}\right)^2\right]\cong 0.197
\end{aligned} \tag{14.1.1770}$$

The overall Gini index is:

$$I(A)=\frac{3}{12}0.\bar{4}+\frac{9}{12}0.197\cong 0.2592 \tag{14.1.1771}$$

In the next step, we have:

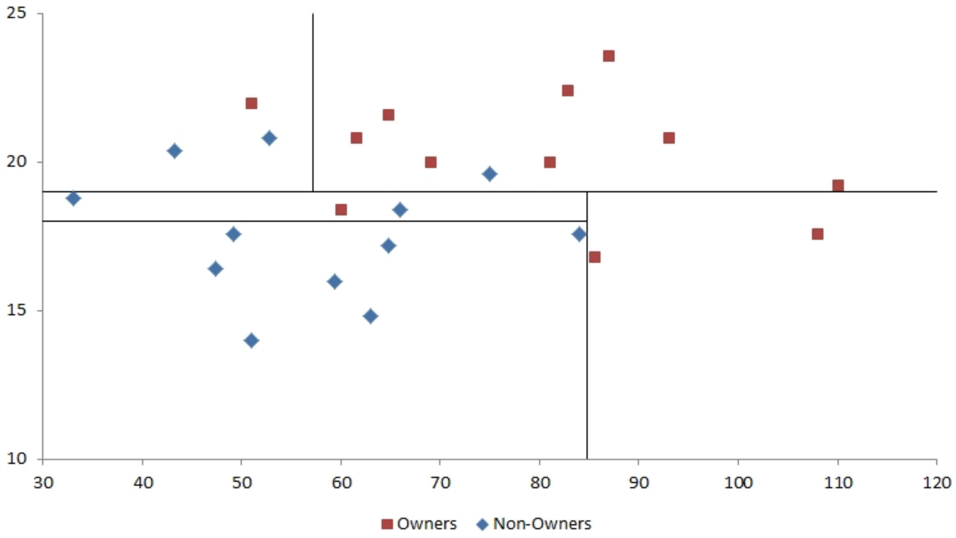


Figure 14.141: **Fourth iteration of CART application with Microsoft® Excel® 14.0.7172**

At the next step:

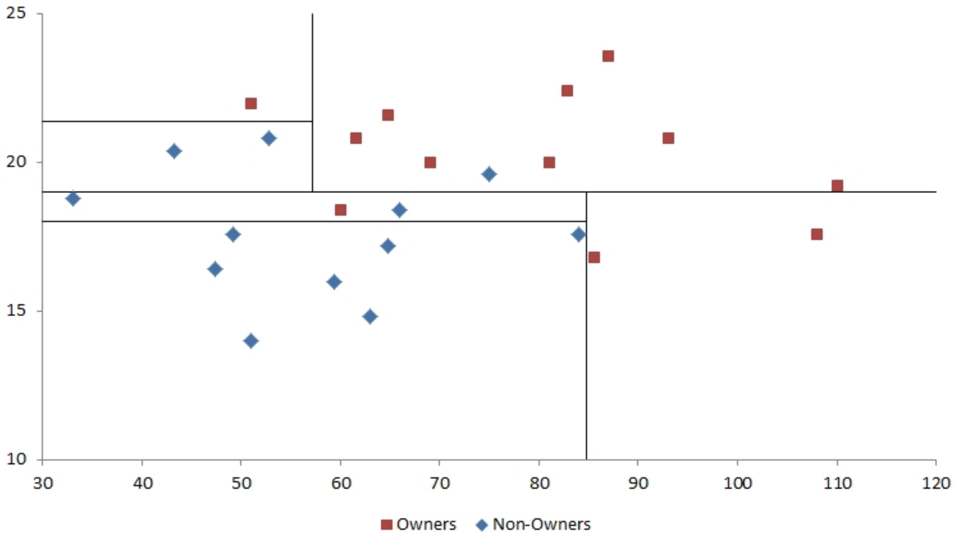


Figure 14.142: **Fifth iteration of CART application with Microsoft® Excel® 14.0.7172**

etc. until the end:

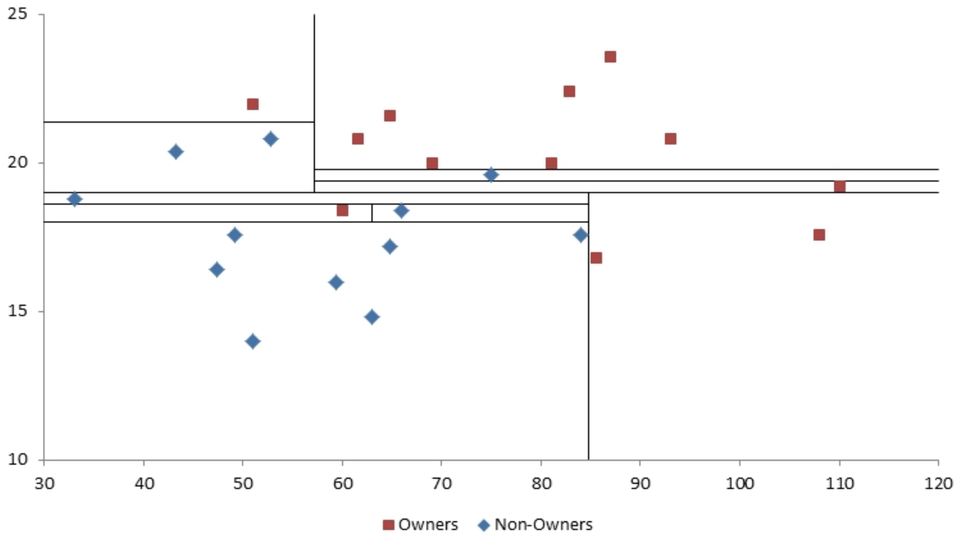


Figure 14.143: **Final iteration of CART application with Microsoft® Excel® 14.0.7172**

where each rectangle is pure (only contains data that one of the two classes).

The reason why the method is named "Classification And Regression Tree" algorithm is that each division can be represented as the division of a node into two successor nodes. The first division is shown as a branch of the tree root node. Here, for example are the first six iterations of the algorithm (only the first six one as the page size is to small to get them all):

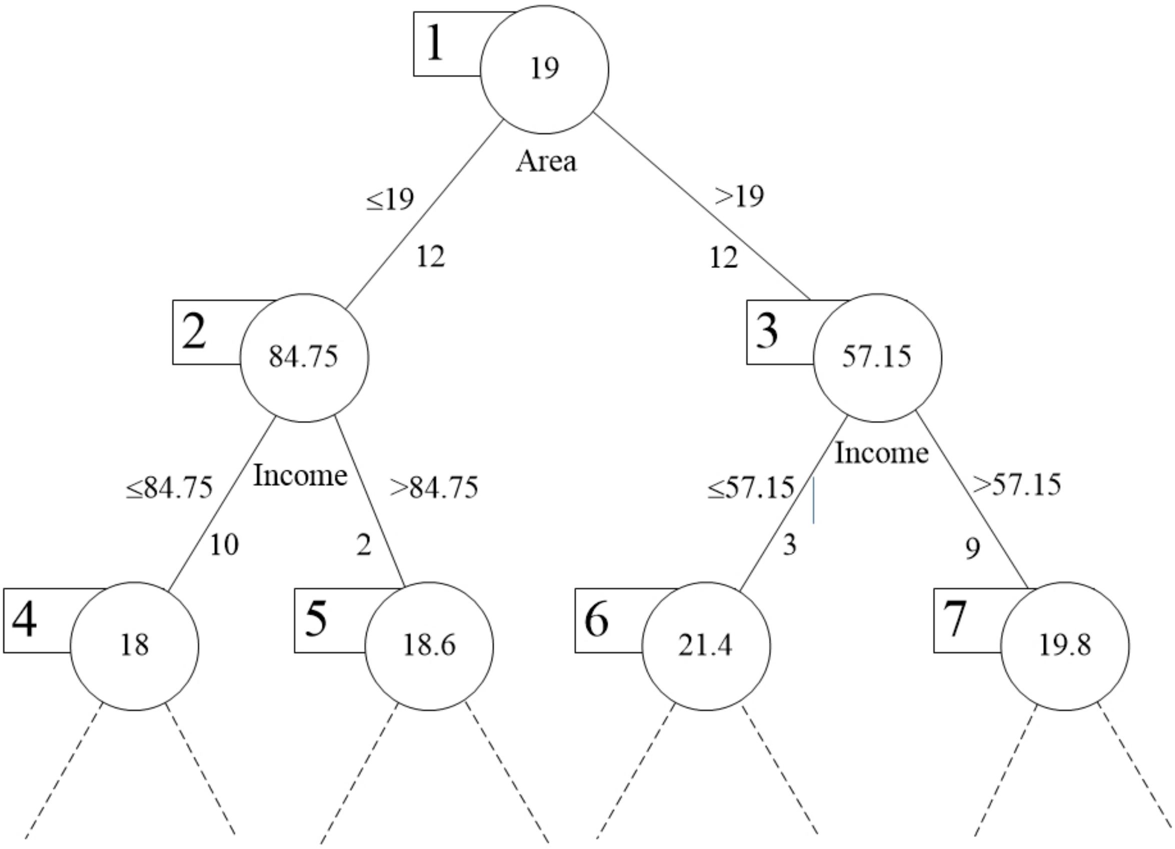

1
19
Area
≤19
>19
12
12
2
84.75
3
57.15
Income
Income
≤84.75
>84.75
≤57.15
>57.15
10
2
3
9
4
18
5
18.6
6
21.4
7
19.8

Figure 14.144: **CART Final result in the traditional tree form**

If you follow the detailed steps given in our `R` companion book you will see the corresponding result that is:

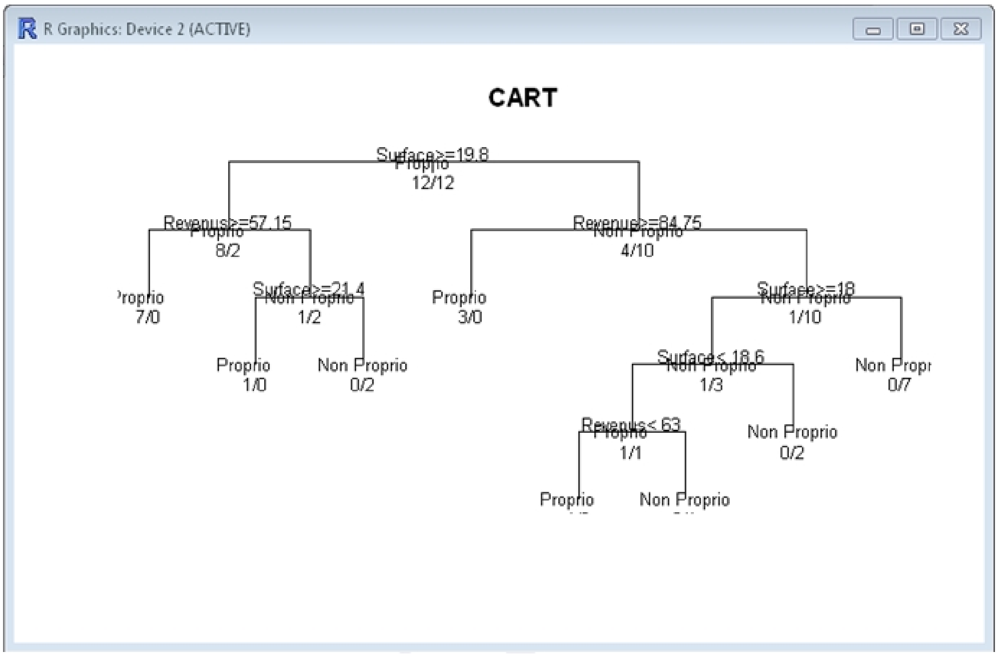
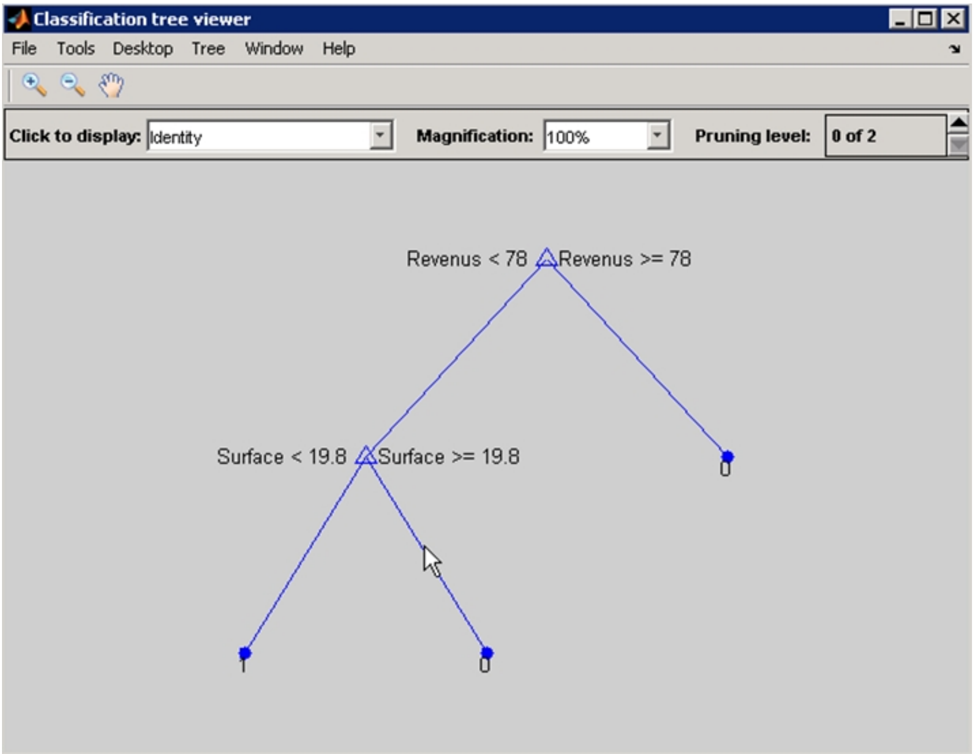

R Graphics: Device 2 (ACTIVE)
CART
Surface>=19.8
Proprio
12/12
Revenus>=57.15
Proprio
8/2
Revenus>=84.75
Non Proprio
4/10
Proprio
7/0
Surface>=21.4
Non Proprio
1/2
Proprio
3/0
Surface>=18
Non Proprio
1/10
Proprio
1/0
Non Proprio
0/2
Surface< 18.6
Non Proprio
1/3
Non Propr
0/7
Revenus< 63
Proprio
1/1
Non Proprio
0/2
Proprio
Non Proprio

Figure 14.145: **CART Final result in the traditional tree form with `R` 3.0.2**

And if you follow the detailed steps given in our MATLAB® companion book you will see the corresponding result that is:


Classification tree viewer
File
Tools
Desktop
Tree
Window
Help
Revenus < 78
Revenus >= 78
Surface < 19.8
Surface >= 19.8
0
1
0

Click to display: Identity
Magnification: 100%
Pruning level: 0 of 2

Figure 14.146: **CART Final result in the traditional tree form with MATLAB® 2013a**

We need to be careful to pick an appropriate tree depth! Indeed:

- If the tree is too deep, we may overfit
- If the tree is too shallow, we may underfit

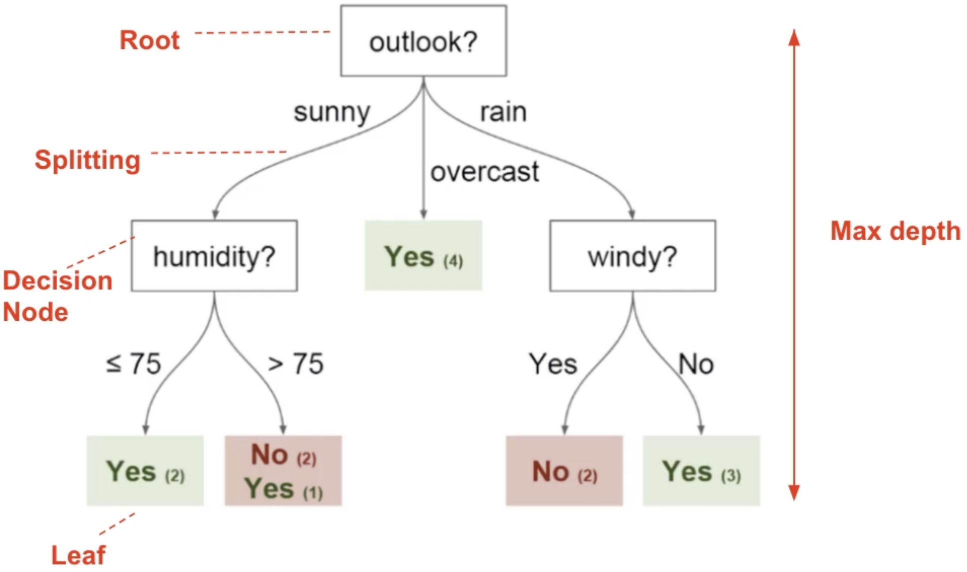


Figure 14.147: **Main vocabulary of decision trees**

Alternative strategy is to create a very deep tree, and then to prune it.

Remark

A common question for binary decision trees is: why not use logistic regression instead? The answer is quite simple! Decision trees helps you developing decision rules but in logistic regression you can not visualize a rule even though you can know which factor are more important! We will also use binary decision trees when there are too many variables, especially categorical ones and you have outliers in predictor variables and relationships are non-linear. And if we want probabilities for each case, we should go for logistic regression obviously!

14.1.20.5.3.2 Detection of significant splits

Quite often it is necessary to measure the significance of a split in a decision tree, especially when the information gain (see further below page 6851) is small.

Let N_A and N_B be the number of items of class A and class B in the parent node. Let N_{AL} represent the number of class A going to the left child node, N_{BL} represent the number of class B going to the left child node, N_{AR} represent the number of class B going to the right child node, and N_{BR} represent the number of class B going to the right child node.

Let p_L and p_B denote the proportion of data going to the left and right node, respectively:

$$\begin{aligned} p_L &= \frac{N_{AL}+N_{BL}}{N_A+N_B} \\ p_R &= \frac{N_{AR}+N_{BR}}{N_A+N_B} \end{aligned} \tag{14.1.1772}$$

The following measure computes the significance of a split. In other words, it measures how much the split deviates from what would be expected in the random data:

$$K=\frac{(N'_{AL}-N_{AL})^2}{N'_{AL}}+\frac{(N'_{BL}-N_{BL})^2}{N'_{BL}}+\frac{(N'_{AR}-N_{AR})^2}{N_{AR}}+\frac{(N'_{BR}-N_{BR})^2}{N'_{BR}} \qquad (14.1.1773)$$

where:

$$\begin{gathered}N'_{AL}=N_A\cdot p_L\\ N'_{BL}=N_B\cdot p_L\\ N'_{AR}=N_A\cdot p_R\\ N'_{BR}=N_B\cdot p_R\end{gathered} \qquad (14.1.1774)$$

We recognise here the shape of a χ^2 random variable with 4 degrees of freedom (χ^2_4).

If K is small, the information gain from the split is not significant. If K is big, it would suggest the information gain from the split is significant.

14.1.20.5.3.3 Classification (and clustering) validation

There exist many different clustering/classification methods, depending on the type of clusters/class sought and on the inherent data characteristics. Given the diversity of clustering/classification algorithms and their parameters it is important to develop objective approaches to assess clustering/classification results. Cluster/class validation and assessment encompasses three main tasks: "clustering/class evaluation" seeks to assess the goodness or quality of the clustering/classification, "clustering/class stability" seeks to understand the sensitivity of the clustering/class result to various algorithmic parameters, for example, the number of clusters/classes, and "clustering/class tendency" assesses the suitability of applying clustering/class in the first place, that is, whether the data has any inherent grouping structure. There are a number of validity measures and statistics that have been proposed for each of the aforementioned tasks, which can be divided into three main types:

- **External validation** measures employ criteria that are not inherent to the dataset. This can be in form of prior or expert-specified knowledge about the clusters/classes.
- **Internal validation** measures employ criteria that are derived from the data itself. For instance, we can use intracluster and intercluster distances to obtain measures of cluster compactness (e.g., how similar are the points in the same cluster?) and separation (e.g., how far apart are the points in different clusters?).
- **Relative validation** measures aim to directly compare different clusterings/classes, usually those obtained via different parameter settings for the same algorithm.

Remark

For the reader interested to further reading on the topic of clustering validation, we strongly recommend: *DATA MINING AND ANALYSIS: Fundamental Concepts and Algorithms* (see [592]), that has excellent detailed definitions with step by step manually calculated examples from page 425 to page 462 (all chapter 17).

To see now some other segmentation external criteria but now with qualitative variables let us consider the following first companion example based on the table below:

Color	**Shape**	**Size**	**Class**
red (R)	square (S)	big (B)	plus
blue (B)	square (S)	big (B)	plus
red (R)	round (R)	small (S)	minus
green (G)	square (S)	small (S)	minus
red (R)	round (R)	big (B)	plus
green (G)	square (S)	big (B)	minus

First let us do again the calculation with the Gini impurity index:

$$I(t)=1-\sum_{i=1}^{n}p(i/t)^2 \qquad (14.1.1775)$$

where $p(i/t)$ still represents the frequency or the probability of the i relatively to the partition t (again, a partition is "pure" if it's Gini impurity index is equal to zero).

As in the previous example we must afterwards calculate the overall Gini impurity index as the weighted average of the Gini impurity index of each partition following:

$$\boxed{I(C)=\sum_{t=1}^{n}f_t\cdot I(t)} \qquad (14.1.1776)$$

Therefore the overall Gini index for the *Color* attribute is given by (keep in mind that our dataset has 6 rows and the numerator of each fraction represent the number of occurrences of the concerned partition!):

$$I(\text{Color})=\frac{1}{6}I(B)+\frac{3}{6}I(R)+\frac{2}{6}I(G)=0.22$$

with (keep in mind that the denominator is always the number of rows of the corresponding partition, and the numerator correspond to how many of a given *Class* we have among all these rows):

- $I(B)=1-(0/1)^2-(1/1)^2=0.0$
- $I(R)=1-(1/3)^2-(2/3)^2=0.44$
- $I(G)=1-(2/2)^2-(0/2)^2=0.0$

And the overall Gini index for the *Shape* attribute is then given by:

$$I(\text{Shape})=\frac{4}{6}I(C)+\frac{2}{6}I(R)=0.5$$

with:

- $I(S)=1-(2/4)^2-(2/4)^2=0.50$
- $I(C)=1-(1/2)^2-(1/2)^2=0.50$

And finally the overall Gini index for the *Size* attribute:

$$I(\text{Size})=\frac{4}{6}I(B)+\frac{2}{6}I(S)=0.25$$

with:

- $I(B)=1-(1/4)^2-(3/4)^2=0.375$
- $I(S)=1-(2/2)^2-(0/2)^2=0.0$

As the variables with lower Gini impurity index are the one to split, then we will chose the *Color* attribute as first split in our example! So we see that in classification trees, the Gini Index can be used as "variable importance" metric for classification trees!

Now let us see a new segmentation criteria named the "information gain", used by the ID3, C4.5 and C5.0 tree-generation algorithms, and based on the measurement of information's entropy[60] from Statistical Mechanics (see page 3530).

For this purpose let us define the entropy for the partition t by the relation (named for recall "Shannon formula"):

$$S(t)=-\sum_{i=1}^{n}p(i/t)\log(p(i/t)) \tag{14.1.1777}$$

Also denoted in some textbooks as:

$$H(\mathcal{C})=-\sum_{i=1}^{r}p_{C_i}\log(p_{C_i})=-\sum_{i=1}^{r}\frac{n_i}{n}\log\left(\frac{n_i}{n}\right) \tag{14.1.1778}$$

where p_{C_i} is the probability of cluster C_i.

Thus, similarly to the Gini impurity index, the conditional entropy for a given attribute is given by the weighted information entropies:

$$\boxed{S(C,a)=\sum_{t=1}^{m}f_t\cdot S(t)=-\sum_{t=1}^{m}f_t\sum_{i=1}^{n}p(i/t)\log(p(i/t))} \tag{14.1.1779}$$

Candidate splits are determined by looking at each variable that makes up an object belonging in a given class. In the example above all objects can either be *plus* or *minus*.

Now for all possible split, we determine fist the entropy before the split which is found using the classification of each object (keep in mind that our dataset has 6 rows and the numerator of each fraction represent the number of occurrences of the concerned partition!):

$$S(\text{Class})=\frac{3}{6}S(\text{plus})+\frac{3}{6}S(\text{minus})=-\frac{3}{6}\log\left(\frac{3}{6}\right)-\frac{3}{6}\log\left(\frac{3}{6}\right)\cong 0.301$$

---

[60] ID3, C4.5 and C5.0 uses entropy information gain (IG) when CART uses the Gini impurity (see page 6839)

Now the conditional entropy for the *Class* given the *Color* attribute is given by :

$$S(C=\text{Class}, a=\text{Color}) = -\left(\frac{1}{6}S(B)+\frac{3}{6}S(R)+\frac{2}{6}S(G)\right)\cong 0.138$$

with (keep in mind that the denominator is always the number of rows of the corresponding partition, and the numerator correspond to how many of a given *Class* we have among all these rows):

- $S(B)=-(1/1)\cdot\log(1/1)=0$
- $S(R)=-(1/3)\cdot\log(1/3)-(2/3)\cdot\log(2/3)\cong 0.276$
- $S(G)=-(2/2)\cdot\log(2/2)=0$

The above definition of $S(C,a)$ is also sometimes defined is some textbooks with the following notation (where $\mathcal{T}$ is the partitioning):

$$\begin{aligned}H(\mathcal{T}\,|\,\mathcal{C}) &= \sum_{i=1}^{r}\frac{n_i}{n}H\left(\mathcal{T}\,|\,C_i\right) = -\sum_{i=1}^{r}\sum_{j=1}^{k}\frac{n_i}{n}\frac{n_{ij}}{n_i}\log\left(\frac{n_{ij}}{n_i}\right) = -\sum_{i=1}^{r}\sum_{j=1}^{k}\frac{n_{ij}}{n}\log\left(\frac{n_{ij}}{n_i}\right)\\ &= -\sum_{i=1}^{r}\sum_{j=1}^{k}p_{ij}\log\left(\frac{p_{ij}}{p_{C_i}}\right)\end{aligned}\tag{14.1.1780}$$

where $p_{ij}=\frac{n_{ij}}{n}$ is the probability that a point in cluster i also belongs to partition j. The more a cluster's members are split into different partitions, the higher the conditional entropy. For a perfect clustering, the conditional entropy value is zero, whereas the worst possible conditional entropy value is $\log(k)$. Further, expanding the previous relation, we can see that:

$$\begin{aligned}H(\mathcal{T}\,|\,\mathcal{C}) &= -\sum_{i=1}^{r}\sum_{j=1}^{k}p_{ij}\left[\log(p_{ij})-\log(p_{C_i})\right]\\ &= -\left(\sum_{i=1}^{r}\sum_{j=1}^{k}p_{ij}\log(p_{ij})\right)+\sum_{i=1}^{r}\left(\log(p_{C_i})\sum_{j=1}^{k}p_{ij}\right)\\ &= -\sum_{i=1}^{r}\sum_{j=1}^{k}p_{ij}\log(p_{ij})+\sum_{i=1}^{r}p_{C_i}\log(p_{C_i}) = H(\mathcal{C},\mathcal{T})-H(\mathcal{C})\end{aligned}\tag{14.1.1781}$$

where $H(\mathcal{C},\mathcal{T})=-\sum_{i=1}^{r}\sum_{j=1}^{k}p_{ij}\log(p_{ij})$ is the joint entropy of $\mathcal{C}$ and $\mathcal{T}$. The conditional entropy $H(\mathcal{T}\,|\,\mathcal{C})$ thus measures the remaining entropy of $\mathcal{T}$ given the clustering $\mathcal{C}$. In particular, $H(\mathcal{T}\,|\,\mathcal{C})=0$ if and only if $\mathcal{T}$ is completely determined by $\mathcal{C}$, corresponding to the ideal clustering. On the other hand, if $\mathcal{C}$ and $\mathcal{T}$ are independent of each other, then $H(\mathcal{T}\,|\,\mathcal{C})=H(\mathcal{T})$, which means that $\mathcal{C}$ provides no information about $\mathcal{T}$.

Remark

The "mutual information" tries to quantify the amount of shared information between the clustering $\mathcal{C}$ and partitioning $\mathcal{T}$, and it is defined as:

$$I(\mathcal{C},\mathcal{T})=\sum_{i=1}^{r}\sum_{j=1}^{k}p_{ij}\log\left(\frac{p_{ij}}{p_{C_i}\cdot p_{T_j}}\right) \tag{14.1.1782}$$

It measures the dependence between the observed joint probability p_{ij} of $\mathcal{C}$ and $\mathcal{T}$, and the expected joint probability $p_{C_i}\cdot p_{T_j}$ under the independence assumption. When $\mathcal{C}$ and $\mathcal{T}$ are independent then $p_{ij}=p_{C_i}\cdot p_{T_j}$, and thus $I(\mathcal{C},\mathcal{T})=0$. However, there is no upper bound on the mutual information.

Expanding this relation we observe that $I(\mathcal{C},\mathcal{T})=H(\mathcal{C})+H(\mathcal{T})-H(\mathcal{C},\mathcal{T})$. Using $H(\mathcal{T}\,|\,\mathcal{C})=H(\mathcal{C},\mathcal{T})-H(\mathcal{C})$ prove just earlier, we get the two equivalent expressions:

$$\begin{aligned}I(\mathcal{C},\mathcal{T})&=H(\mathcal{T})-H(\mathcal{T}\,|\,\mathcal{C})\\ I(\mathcal{C},\mathcal{T})&=H(\mathcal{C})-H(\mathcal{C}\,|\,\mathcal{T})\end{aligned} \tag{14.1.1783}$$

The "information gain" (IG) is then be defined by finding the difference in the prior entropy and the conditional entropy for the class C given a specific attribute variable a:

$$\boxed{\text{IG}(C,a)=S(C)-S(C,a)=\left(-\sum_{t=1}^{m}f_t\log(f_t)\right)-\left(-\sum_{t=1}^{m}f_t\sum_{i=1}^{n}p(i/t)\log(p(i/t))\right)} \tag{14.1.1784}$$

We also recognize here that the information gain is simply the relation derived in the remark box above:

$$I(\mathcal{C},\mathcal{T})=H(\mathcal{C})-H(\mathcal{C}\,|\,\mathcal{T}) \tag{14.1.1785}$$

Therefore in our example here the information gain for colors is:

$$\text{IG}(\text{Class},\text{Color})=S(\text{Class})-S(\text{Class},\text{Color})\cong 0.301-0.459=-0.158$$

And now the conditional entropy for the *Class* given the *Shape* attribute is then given by (we change to notation for the entropy from H to S in the second term to avoid having to write an ugly $S(S)$...):

$$S(\text{Class},\text{Shape})=\frac{4}{6}S(R)+\frac{2}{6}H(S)=1$$

with:

- $S(R)=-(2/4)\cdot\log(2/4)-(2/4)\cdot\log(2/4)$
- $S(S)=-(1/2)\cdot\log(1/2)-(1/2)\cdot\log(1/2)$

Therefore the information gain for *Shape* is:

$$\text{IG}(\text{Class},\text{Shape})=S(\text{Class})-S(\text{Class},\text{Shape})\cong 0.301-1=-0.699$$

And finally the conditional entropy for the *Class* given the *Size* attribute is then given by:

$$S(\text{Class, Size})=\frac{4}{6}S(B)+\frac{2}{6}H(S)=0.541$$

with:

- $S(B)=-(1/4)\cdot\log(1/4)-(3/4)\cdot\log(3/4)$
- $S(S)=-(2/2)\cdot\log(2/2)$

Therefore the information gain for sizes is:

$$\text{IG}(\text{Class, Size})=S(\text{Class})-S(\text{Class, Size})\cong 0.301-0.541=-0.24$$

As the variables with lower conditional entropy (respectively highest information gain) are the one to split, then we will chose the *Color* attribute as first split in our example! So here the split choice is the same as with the Gini impurity index. So we see that in classification trees, the overall information entropy can be used as "variable importance" metric for classification trees!

The problem with information gain as a measure to select the attribute for partition is that in the quest of pure partition, it may select attributes that are meaningless from the Machine Learning point of view. This drawback is due to the inherent deficiency in the measure information gain that gives preference to the attribute that can divide the parent dataset to datasets with the least amount of entropy.

Remark

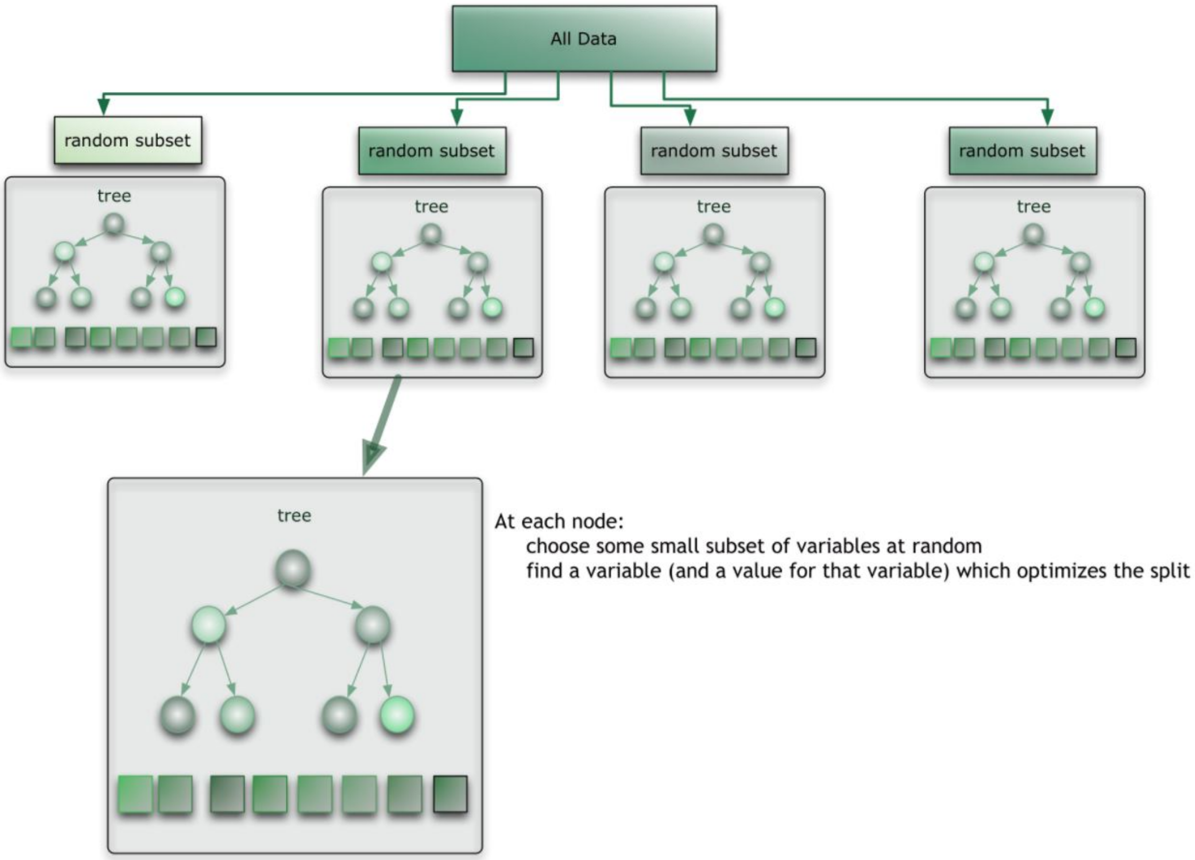
Notice that there are a lot of other segmentation criteria as the likelihood-ratio chi-squared statistic, the DKM criterion, the Twoing criterion, the Orthogonal criterion (ORT), the Kolmogorov-Smirnoff criterion, the AUC-Splitting criterion, ...

And as mentioned above, we will stop here about our study of CART because the study of the corresponding empirical variations techniques are a full time job as they are numerous (and almost none gives the same results...).

14.1.20.5.3.4 Random Forests

The random forest (see figure below) takes the notion of binary decision tree to the next level by combining trees with the notion of an ensemble. Thus, in ensemble terms, the standard decision trees are weak learners and the random forest is a strong learner (in the sense of a statistical crows sourcing).

In other words "random forests" or "random decision forests" are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.



At each node:
choose some small subset of variables at random
find a variable (and a value for that variable) which optimizes the split

Figure 14.148: Random Forster illustration (author: ?)

Here is how such a system is trained; for some number of trees:

1. Sample N cases at random with replacement to create a subset of the data (see top layer of figure above). The subset should be about 66% of the total set.
2. At each node:
   (a) For some number m, m predictor variables are selected at random from all the predictor variables.
   (b) The predictor variable that provides the best split, according to some objective function, is used to do a binary split on that node.
   (c) At the next node, choose another m variables at random from all predictor variables and do the same.

For prediction with a new entered input into the system, it is run down all of the trees. The result may either be an average or weighted average of all of the terminal nodes that are reached, or, in the case of categorical variables, a voting majority.

Note that:

- With a large number of predictors, the eligible predictor set will be quite different from node to node.

- The greater the inter-tree correlation, the greater the random forest error rate, so one pressure on the model is to have the trees as uncorrelated as possible.
- As m goes down, both inter-tree correlation and the strength of individual trees go down. So some optimal value of m must be discovered.

Random forest runtimes are quite fast, and they are able to deal with unbalanced and missing data. Random Forest weaknesses are that when used for regression they cannot predict beyond the range in the training data, and that they may over-fit data sets that are particularly noisy. Of course, the best test of any algorithm is how well it works upon your own data set.

To understand why random forest works quite good you must first remember the bias-variance tradeoff relation (see earlier above page 6798) that is given by:

$$\text{E}\left[(y-\hat{f})^2\right]=\sigma_\varepsilon^2+\text{V}(\hat{f})+\text{Bias}(\hat{f})^2 \tag{14.1.1786}$$

So first, if the trees are sufficiently deep, they have very small bias!

Now let us recall the variance of the mean of identically correlated i.i.d variables (see section Statistics 620):

$$\text{V}(\bar{X})=\rho\sigma^2+\frac{1-\rho}{B}\sigma^2 \tag{14.1.1787}$$

So to improve the variance over that of a single tree, we see that de-correlation gives better accuracy as a small ρ decrease the first term, but we also see that the second term decrease if the number of trees B increase (independently of ρ)! So the idea in random forests is to improve the variance reduction of bagging by reducing the correlation between the trees, without increasing the variance too much. This is achieved in the tree-growing process through random selection of the input variables.

Why would anyone use logistic regression when machine learning models like Random Forest often predict better?

Because sometimes, prediction isn't the main goal! If we want to assess whether a treatment or a biomarker has a real effect on a health outcome, we need to quantify the relationship and test whether it's statistically significant. That's where logistic regression becomes more relevant than Random Forests.

14.1.20.5.4 k-means clustering

The "k-means" algorithm or "mobile centers", also sometimes named "classification method with dynamic clouds" is in statistics and Machine Learning (specifically unsupervised learning), a data partitioning algorithm, i.e. a method that aims to divide observations in K clusters in which each observation belongs to the partition with the nearest average (mean).

The basic steps of the algorithm are as follows:

1. We choose a partitioning into K groups

2. We generate K averages (centers c_i) randomly
3. The data are assigned to the group whose center is closest to them
4. We calculated the average of each group using the affected data (new centers)
5. We return to step 3

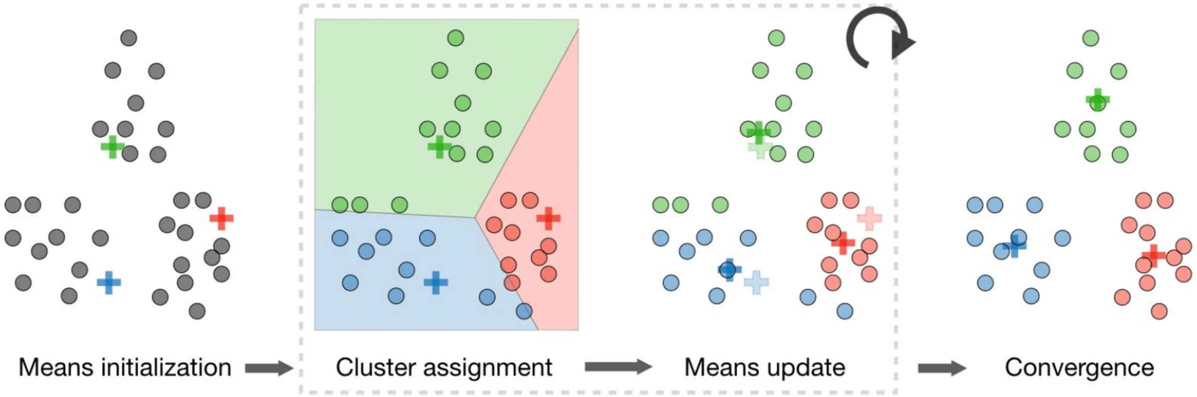


Figure 14.149: k-means iterative algorithm (authors: Afshine Amidi, Shervine Amidi)

To make more explicit the algorithm each step will be illustrated by means of the following small dataset:

Object	Weight	Length
a	12	10
b	15	25
c	30	55
d	50	100
e	35	70
f	45	70
g	35	60

we determine k temporary centers of k clusters generally by randomly dragging k individuals from the database. These k individuals form the center of the k clusters. To achieve the first clustering, we assign each individual to the cluster of which it is closest. Here the notion of proximity is understood in a measure of the Euclidean distance:

$$d(\vec{x}_i,\vec{c}_k)=\sqrt{\sum_{j=1}^{n}(x_{ij}-c_{kj})^2} \tag{14.1.1788}$$

For $k=2$, we will randomly select k individuals, here the object a and the object d for example:

Clusters	Objects	Centers
Cluster c_1	$c_1\{\text{a}\}$	$(12,10)$
Cluster c_2	$c_2\{\text{d}\}$	$(50,100)$

Now we must calculate the Euclidean distance of each individual with respect to the centers and assign each to the cluster that is closer to it:

Objects	**Distance Cluster 1**	**Distance Cluster 2**	**Affectation To**
a	0.00	97.69	Cluster c_1
b	15.30	82.76	Cluster c_1
c	48.47	49.24	Cluster c_1
d	97.69	0.00	Cluster c_2
e	64.26	33.54	Cluster c_2
f	68.48	30.41	Cluster c_2
g	55.04	42.72	Cluster c_2

Then, we again determine the k centers of the k last partitions performed. The distances of each individual with respect to these centers are recalculated and each is reallocated according to the cluster that is close to it.

Following the last clustering we obtain the center (point of gravity or barycenter) of each partitions like this:

Clusters	**Objects**	**Centers**
Cluster c_1	$c_1\{\mathrm{a, b, c}\}$	$(19, 30)$
Cluster c_2	$c_2\{\mathrm{d, e, f, g}\}$	$(41.25, 75)$

Again, it is necessary to calculate the Euclidean distance of each individual with respect to the centers and reassign each to the cluster that is closer to it:

Objects	**Distance Cluster 1**	**Distance Cluster 2**	**Affectation To**
a	21.19	71.28	Cluster c_1
b	6.40	56.47	Cluster c_1
c	27.31	22.95	Cluster c_2
d	76.56	26.49	Cluster c_2
e	43.08	8.00	Cluster c_2
f	47.71	6.25	Cluster c_2
g	34.00	16.25	Cluster c_2

We see at this level that the object c change the group (centroid) he belongs to before.

This last step is repeated as many times as necessary until the process is stabilized. The algorithm can be stopped either:

- Because we do not observe changes anymore in the composition of each cluster after two successive iterations
- Because an empirical control criterion is verified (number of iteration fixed for example)

In the previous illustration, it is rather the first reason that will decide the end of the clustering. Indeed, at the 3rd iteration there is no change in the composition of the clusters.

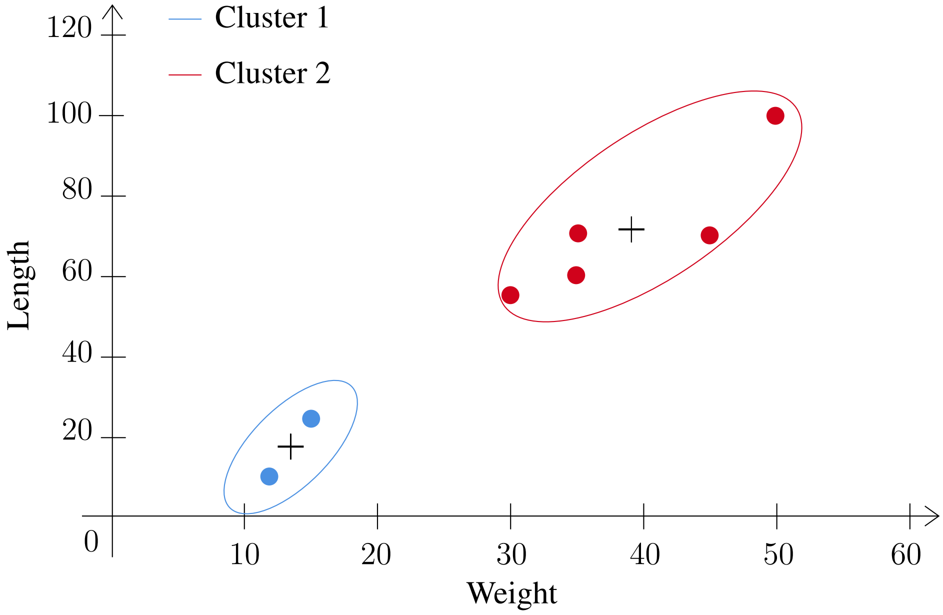
New centers following last assignment:

Clusters	**Objects**	**Centers**
Cluster c_1	$c_1\{\mathrm{a},\mathrm{b}\}$	$(13.5, 17.5)$
Cluster c_2	$c_2\{\mathrm{d},\mathrm{e},\mathrm{f},\mathrm{g}\}$	$(39, 71)$

Distance calculation and reallocation:

Objects	**Distance Cluster 1**	**Distance Cluster 2**	**Affectation To**
a	7.65	66.71	Cluster c_1
b	7.65	51.88	Cluster c_1
c	40.97	18.36	Cluster c_2
d	90.21	31.02	Cluster c_2
e	56.73	4.12	Cluster c_2
f	61.22	6.08	Cluster c_2
g	47.63	11.70	Cluster c_2

It is clear that the composition of the clusters does not change anymore and a projection of the clusters on a chart gives:



So we see with this small previous example that our problem is the same as minimizing the enlarged criterion:

$$K_{\text{means}} \min \sum_{k=1}^{K} \sum_{C(i)=k} ||\vec{x}_i - \vec{c}_k||^2 \qquad (14.1.1789)$$

over bother clusterings C and $c_1, \ldots, c_K \in \mathbb{R}^p$. That is to say, minimize the "within-cluster sum of squares" (WCSS).

Remark

k-means is as we have already mentioned it a hard assignment version of the soft assignment variant of EM (see above page 6697), with the assumptions that clusters are spherical (plus all variables must have same variance and the prior probability for all clusters are the same, i.e. each cluster has roughly equal number of observations). Here "spherical" means identical variance-covariance matrices for each cluster (assuming gaussian distribution), which is also known as "model-based clustering". Furthermore notice that there is no "k-means algorithm". There is MacQueens algorithm for k-means, the Lloyd/Forgy algorithm for k-means, the Hartigan-Wong method, ...

The procedure describe above is the Lloyds algorithm (the one assumed by default in most cheap books on Data Science, it is sometimes also referred to as "naive k-means", because there exist much faster alternatives) that consists of two step given a set of k-means $m_1^{(1)}, \ldots, m_k^{(1)}$ and therefore of $S = \{S_1, S_2, \ldots, S_k\}$ clusters:

- the E-step (assignment step), where each object is assigned to the centroid such that it is assigned to the most likely cluster using the shortest euclidean distance):

$$S_i^{(t)} = \{x_p : \left\|x_p - m_i^{(t)}\right\|^2 \leq \left\|x_p - m_j^{(t)}\right\|^2 \forall j, 1 \leq j \leq k\} \quad (14.1.1790)$$

- the M-step (update step), where the model (=centroids) are recomputed (= least squares optimization):

$$m_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{x_j \in S_i^{(t)}} x_j \quad (14.1.1791)$$

... iterating these two steps, as done by Lloyd, makes this effectively an instance of the general EM scheme.

We see also well through the previous step-by-step example that the k-means algorithm (which therefore belongs to non-hierarchical clustering techniques) however does not necessarily converge to an optimal solution. Let us recall that a global optimization calculation is inconsistent with the data volumes used, regardless of the power of computers. Thus, the k-means will use iterative algorithms to reach a local optimum. It is also an algorithm that will try to find the best K initial points.

Some software gives you the ability to set the initial values and this will affect the final quality of the typology, knowing that there is ONE no good initial choice. This varies depending on the configuration of data, on the return of experience (REX) and even the chance...

There are three common solutions:

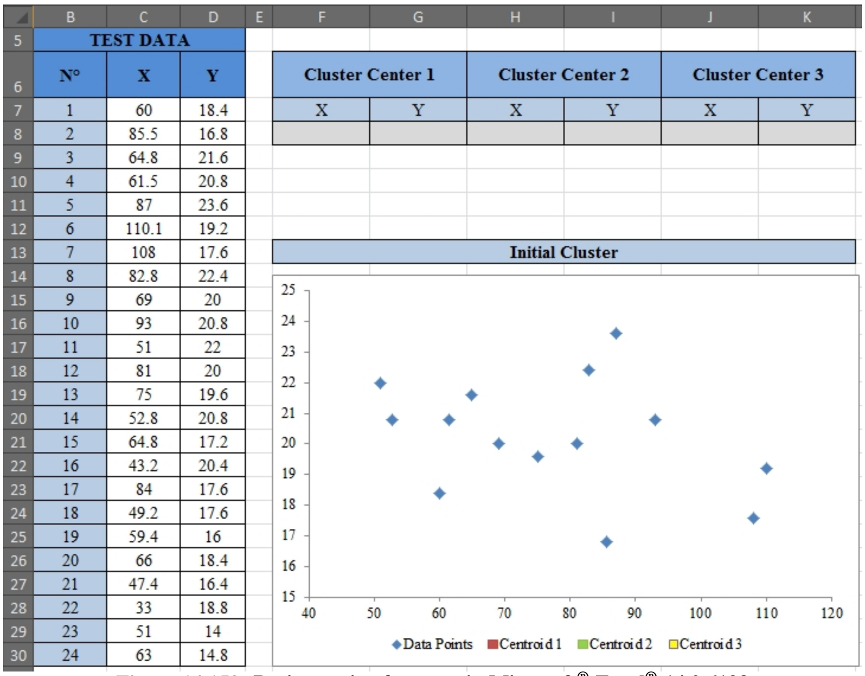
1. The software determines the K initial points randomly. It can perform a number of tests and he will choose the most conclusive one.

2. We use the expert opinion that suppose someone has a fairly good knowledge of the study population for attaching to each class an ideal type. This may or not be a real individual.

3. The software distribute the K initial points not randomly but according to some empirical algorithms.

Let us see a further step that is instructive to show how to implement this technique in a spreadsheet software like Microsoft® Excel® 14.0.6123 with genetic algorithms (see further below).

For this, we will first consider the following structure for which the idea is to establish three centers (so it is a 3-means):



TEST DATA		
N°	X	Y
1	60	18.4
2	85.5	16.8
3	64.8	21.6
4	61.5	20.8
5	87	23.6
6	110.1	19.2
7	108	17.6
8	82.8	22.4
9	69	20
10	93	20.8
11	51	22
12	81	20
13	75	19.6
14	52.8	20.8
15	64.8	17.2
16	43.2	20.4
17	84	17.6
18	49.2	17.6
19	59.4	16
20	66	18.4
21	47.4	16.4
22	33	18.8
23	51	14
24	63	14.8

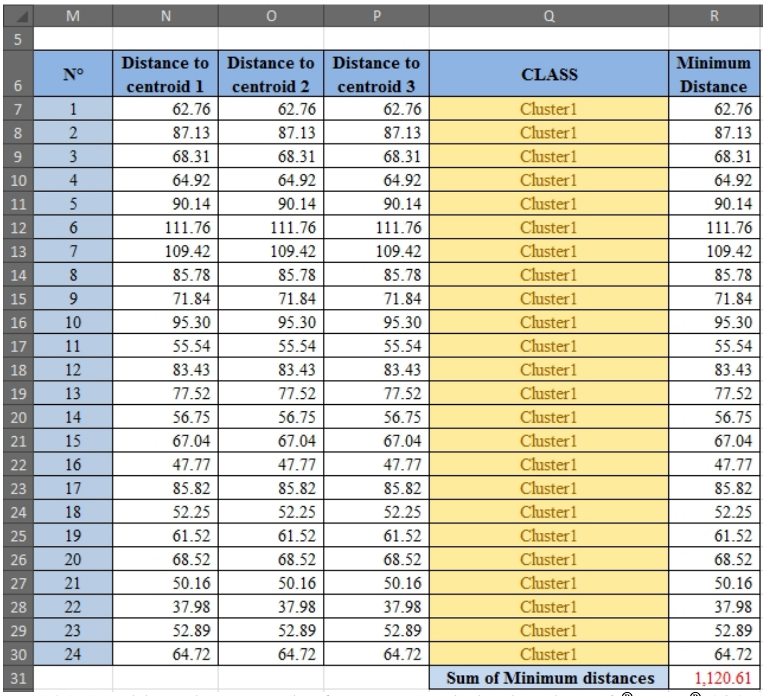
Cluster Center 1		Cluster Center 2		Cluster Center 3	
X	Y	X	Y	X	Y

Initial Cluster

B
C
D
E
F
G
H
I
J
K
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
25
24
23
22
21
20
19
18
17
16
15
40
50
60
70
80
90
100
110
120
◆Data Points
🟥Centroid 1
🟩Centroid 2
☐Centroid 3

Figure 14.150: **Basic starting k-means in Microsoft® Excel® 14.0.6123**

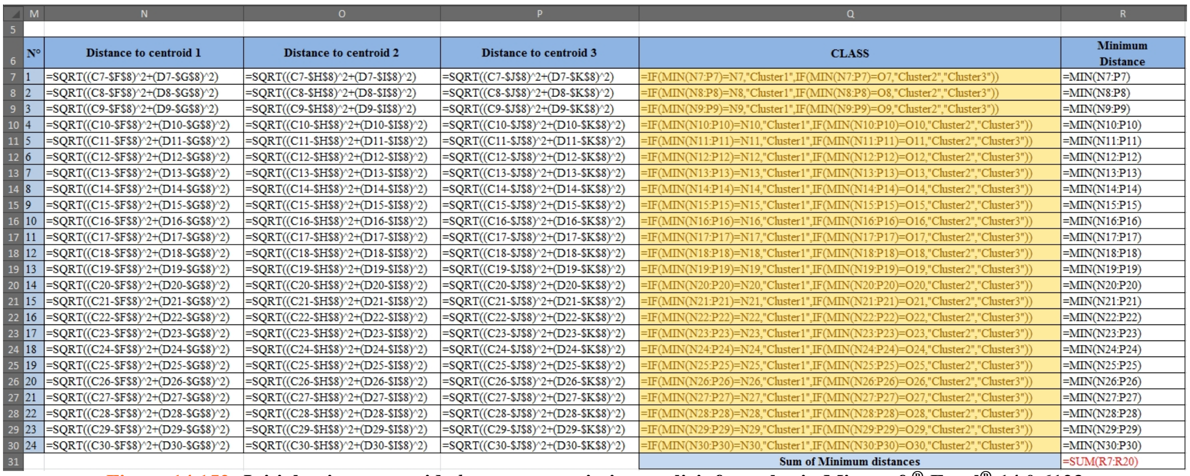
where we have on the left data from a population-based on the characteristics X and Y with a small table that will display the coordinates of the three centroid. We create on the same sheet the following table:



	M	N	O	P	Q	R
5						
6	N°	Distance to centroid 1	Distance to centroid 2	Distance to centroid 3	CLASS	Minimum Distance
7	1	62.76	62.76	62.76	Cluster1	62.76
8	2	87.13	87.13	87.13	Cluster1	87.13
9	3	68.31	68.31	68.31	Cluster1	68.31
10	4	64.92	64.92	64.92	Cluster1	64.92
11	5	90.14	90.14	90.14	Cluster1	90.14
12	6	111.76	111.76	111.76	Cluster1	111.76
13	7	109.42	109.42	109.42	Cluster1	109.42
14	8	85.78	85.78	85.78	Cluster1	85.78
15	9	71.84	71.84	71.84	Cluster1	71.84
16	10	95.30	95.30	95.30	Cluster1	95.30
17	11	55.54	55.54	55.54	Cluster1	55.54
18	12	83.43	83.43	83.43	Cluster1	83.43
19	13	77.52	77.52	77.52	Cluster1	77.52
20	14	56.75	56.75	56.75	Cluster1	56.75
21	15	67.04	67.04	67.04	Cluster1	67.04
22	16	47.77	47.77	47.77	Cluster1	47.77
23	17	85.82	85.82	85.82	Cluster1	85.82
24	18	52.25	52.25	52.25	Cluster1	52.25
25	19	61.52	61.52	61.52	Cluster1	61.52
26	20	68.52	68.52	68.52	Cluster1	68.52
27	21	50.16	50.16	50.16	Cluster1	50.16
28	22	37.98	37.98	37.98	Cluster1	37.98
29	23	52.89	52.89	52.89	Cluster1	52.89
30	24	64.72	64.72	64.72	Cluster1	64.72
31					Sum of Minimum distances	1,120.61

Figure 14.151: **Initial points-centroids k-means association in Microsoft® Excel® 14.0.6123**

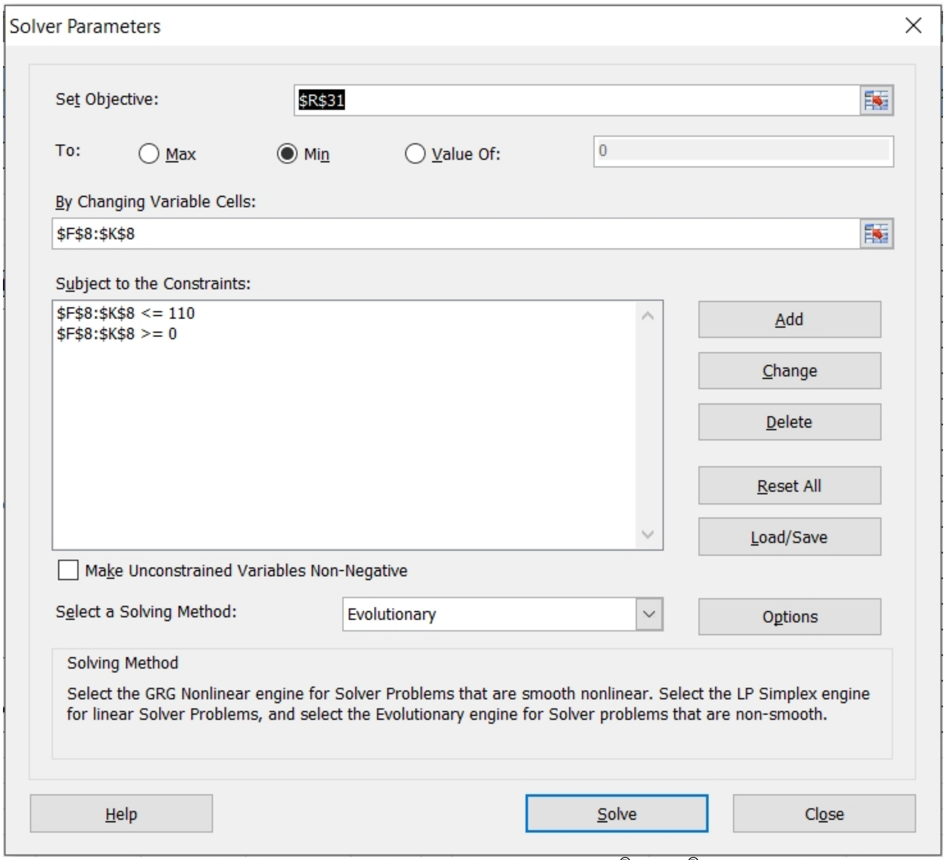
with trivial formulas for the three columns `N`, `O`, `P`, where we used the standard Euclidean distance:



	M	N	O	P	Q	R
5						
6	N°	Distance to centroid 1	Distance to centroid 2	Distance to centroid 3	CLASS	Minimum Distance
7	1	=SQRT((C7-$F$8)^2+(D7-$G$8)^2)	=SQRT((C7-$H$8)^2+(D7-$I$8)^2)	=SQRT((C7-$J$8)^2+(D7-$K$8)^2)	=IF(MIN(N7:P7)=N7,"Cluster1",IF(MIN(N7:P7)=O7,"Cluster2","Cluster3"))	=MIN(N7:P7)
8	2	=SQRT((C8-$F$8)^2+(D8-$G$8)^2)	=SQRT((C8-$H$8)^2+(D8-$I$8)^2)	=SQRT((C8-$J$8)^2+(D8-$K$8)^2)	=IF(MIN(N8:P8)=N8,"Cluster1",IF(MIN(N8:P8)=O8,"Cluster2","Cluster3"))	=MIN(N8:P8)
9	3	=SQRT((C9-$F$8)^2+(D9-$G$8)^2)	=SQRT((C9-$H$8)^2+(D9-$I$8)^2)	=SQRT((C9-$J$8)^2+(D9-$K$8)^2)	=IF(MIN(N9:P9)=N9,"Cluster1",IF(MIN(N9:P9)=O9,"Cluster2","Cluster3"))	=MIN(N9:P9)
10	4	=SQRT((C10-$F$8)^2+(D10-$G$8)^2)	=SQRT((C10-$H$8)^2+(D10-$I$8)^2)	=SQRT((C10-$J$8)^2+(D10-$K$8)^2)	=IF(MIN(N10:P10)=N10,"Cluster1",IF(MIN(N10:P10)=O10,"Cluster2","Cluster3"))	=MIN(N10:P10)
11	5	=SQRT((C11-$F$8)^2+(D11-$G$8)^2)	=SQRT((C11-$H$8)^2+(D11-$I$8)^2)	=SQRT((C11-$J$8)^2+(D11-$K$8)^2)	=IF(MIN(N11:P11)=N11,"Cluster1",IF(MIN(N11:P11)=O11,"Cluster2","Cluster3"))	=MIN(N11:P11)
12	6	=SQRT((C12-$F$8)^2+(D12-$G$8)^2)	=SQRT((C12-$H$8)^2+(D12-$I$8)^2)	=SQRT((C12-$J$8)^2+(D12-$K$8)^2)	=IF(MIN(N12:P12)=N12,"Cluster1",IF(MIN(N12:P12)=O12,"Cluster2","Cluster3"))	=MIN(N12:P12)
13	7	=SQRT((C13-$F$8)^2+(D13-$G$8)^2)	=SQRT((C13-$H$8)^2+(D13-$I$8)^2)	=SQRT((C13-$J$8)^2+(D13-$K$8)^2)	=IF(MIN(N13:P13)=N13,"Cluster1",IF(MIN(N13:P13)=O13,"Cluster2","Cluster3"))	=MIN(N13:P13)
14	8	=SQRT((C14-$F$8)^2+(D14-$G$8)^2)	=SQRT((C14-$H$8)^2+(D14-$I$8)^2)	=SQRT((C14-$J$8)^2+(D14-$K$8)^2)	=IF(MIN(N14:P14)=N14,"Cluster1",IF(MIN(N14:P14)=O14,"Cluster2","Cluster3"))	=MIN(N14:P14)
15	9	=SQRT((C15-$F$8)^2+(D15-$G$8)^2)	=SQRT((C15-$H$8)^2+(D15-$I$8)^2)	=SQRT((C15-$J$8)^2+(D15-$K$8)^2)	=IF(MIN(N15:P15)=N15,"Cluster1",IF(MIN(N15:P15)=O15,"Cluster2","Cluster3"))	=MIN(N15:P15)
16	10	=SQRT((C16-$F$8)^2+(D16-$G$8)^2)	=SQRT((C16-$H$8)^2+(D16-$I$8)^2)	=SQRT((C16-$J$8)^2+(D16-$K$8)^2)	=IF(MIN(N16:P16)=N16,"Cluster1",IF(MIN(N16:P16)=O16,"Cluster2","Cluster3"))	=MIN(N16:P16)
17	11	=SQRT((C17-$F$8)^2+(D17-$G$8)^2)	=SQRT((C17-$H$8)^2+(D17-$I$8)^2)	=SQRT((C17-$J$8)^2+(D17-$K$8)^2)	=IF(MIN(N17:P17)=N17,"Cluster1",IF(MIN(N17:P17)=O17,"Cluster2","Cluster3"))	=MIN(N17:P17)
18	12	=SQRT((C18-$F$8)^2+(D18-$G$8)^2)	=SQRT((C18-$H$8)^2+(D18-$I$8)^2)	=SQRT((C18-$J$8)^2+(D18-$K$8)^2)	=IF(MIN(N18:P18)=N18,"Cluster1",IF(MIN(N18:P18)=O18,"Cluster2","Cluster3"))	=MIN(N18:P18)
19	13	=SQRT((C19-$F$8)^2+(D19-$G$8)^2)	=SQRT((C19-$H$8)^2+(D19-$I$8)^2)	=SQRT((C19-$J$8)^2+(D19-$K$8)^2)	=IF(MIN(N19:P19)=N19,"Cluster1",IF(MIN(N19:P19)=O19,"Cluster2","Cluster3"))	=MIN(N19:P19)
20	14	=SQRT((C20-$F$8)^2+(D20-$G$8)^2)	=SQRT((C20-$H$8)^2+(D20-$I$8)^2)	=SQRT((C20-$J$8)^2+(D20-$K$8)^2)	=IF(MIN(N20:P20)=N20,"Cluster1",IF(MIN(N20:P20)=O20,"Cluster2","Cluster3"))	=MIN(N20:P20)
21	15	=SQRT((C21-$F$8)^2+(D21-$G$8)^2)	=SQRT((C21-$H$8)^2+(D21-$I$8)^2)	=SQRT((C21-$J$8)^2+(D21-$K$8)^2)	=IF(MIN(N21:P21)=N21,"Cluster1",IF(MIN(N21:P21)=O21,"Cluster2","Cluster3"))	=MIN(N21:P21)
22	16	=SQRT((C22-$F$8)^2+(D22-$G$8)^2)	=SQRT((C22-$H$8)^2+(D22-$I$8)^2)	=SQRT((C22-$J$8)^2+(D22-$K$8)^2)	=IF(MIN(N22:P22)=N22,"Cluster1",IF(MIN(N22:P22)=O22,"Cluster2","Cluster3"))	=MIN(N22:P22)
23	17	=SQRT((C23-$F$8)^2+(D23-$G$8)^2)	=SQRT((C23-$H$8)^2+(D23-$I$8)^2)	=SQRT((C23-$J$8)^2+(D23-$K$8)^2)	=IF(MIN(N23:P23)=N23,"Cluster1",IF(MIN(N23:P23)=O23,"Cluster2","Cluster3"))	=MIN(N23:P23)
24	18	=SQRT((C24-$F$8)^2+(D24-$G$8)^2)	=SQRT((C24-$H$8)^2+(D24-$I$8)^2)	=SQRT((C24-$J$8)^2+(D24-$K$8)^2)	=IF(MIN(N24:P24)=N24,"Cluster1",IF(MIN(N24:P24)=O24,"Cluster2","Cluster3"))	=MIN(N24:P24)
25	19	=SQRT((C25-$F$8)^2+(D25-$G$8)^2)	=SQRT((C25-$H$8)^2+(D25-$I$8)^2)	=SQRT((C25-$J$8)^2+(D25-$K$8)^2)	=IF(MIN(N25:P25)=N25,"Cluster1",IF(MIN(N25:P25)=O25,"Cluster2","Cluster3"))	=MIN(N25:P25)
26	20	=SQRT((C26-$F$8)^2+(D26-$G$8)^2)	=SQRT((C26-$H$8)^2+(D26-$I$8)^2)	=SQRT((C26-$J$8)^2+(D26-$K$8)^2)	=IF(MIN(N26:P26)=N26,"Cluster1",IF(MIN(N26:P26)=O26,"Cluster2","Cluster3"))	=MIN(N26:P26)
27	21	=SQRT((C27-$F$8)^2+(D27-$G$8)^2)	=SQRT((C27-$H$8)^2+(D27-$I$8)^2)	=SQRT((C27-$J$8)^2+(D27-$K$8)^2)	=IF(MIN(N27:P27)=N27,"Cluster1",IF(MIN(N27:P27)=O27,"Cluster2","Cluster3"))	=MIN(N27:P27)
28	22	=SQRT((C28-$F$8)^2+(D28-$G$8)^2)	=SQRT((C28-$H$8)^2+(D28-$I$8)^2)	=SQRT((C28-$J$8)^2+(D28-$K$8)^2)	=IF(MIN(N28:P28)=N28,"Cluster1",IF(MIN(N28:P28)=O28,"Cluster2","Cluster3"))	=MIN(N28:P28)
29	23	=SQRT((C29-$F$8)^2+(D29-$G$8)^2)	=SQRT((C29-$H$8)^2+(D29-$I$8)^2)	=SQRT((C29-$J$8)^2+(D29-$K$8)^2)	=IF(MIN(N29:P29)=N29,"Cluster1",IF(MIN(N29:P29)=O29,"Cluster2","Cluster3"))	=MIN(N29:P29)
30	24	=SQRT((C30-$F$8)^2+(D30-$G$8)^2)	=SQRT((C30-$H$8)^2+(D30-$I$8)^2)	=SQRT((C30-$J$8)^2+(D30-$K$8)^2)	=IF(MIN(N30:P30)=N30,"Cluster1",IF(MIN(N30:P30)=O30,"Cluster2","Cluster3"))	=MIN(N30:P30)
31					Sum of Minimum distances	=SUM(R7:R20)

Figure 14.152: **Initial points-centroids k-means association explicit formulas in Microsoft® Excel® 14.0.6123**

Then we launch the Microsoft® Excel® 14.0.6123 Solver with the following parameters being careful to take the Evolutionary algorithm option. Therefore we assume that at every run we could have a different results:



Solver Parameters

Set Objective: $R$31

To: ○ Max ● Min ○ Value Of: 0

By Changing Variable Cells: $F$8:$K$8

Subject to the Constraints:

$F$8:$K$8 <= 110
$F$8:$K$8 >= 0

☐ Make Unconstrained Variables Non-Negative

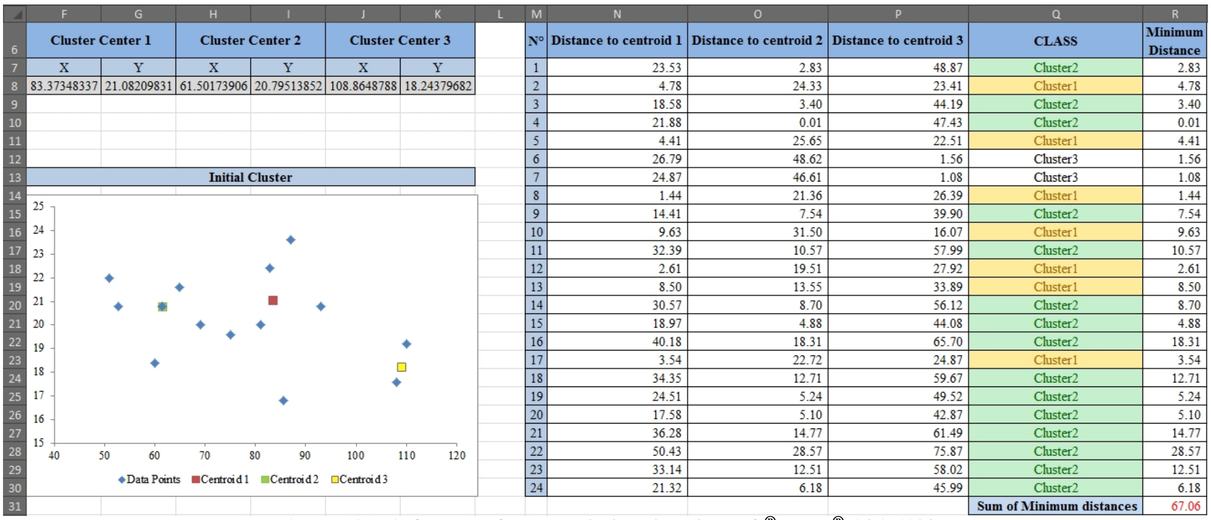
Select a Solving Method: Evolutionary

Solving Method

Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

Figure 14.153: k-means solver settings in Microsoft® Excel® 14.0.6123

and therefore we get for results:

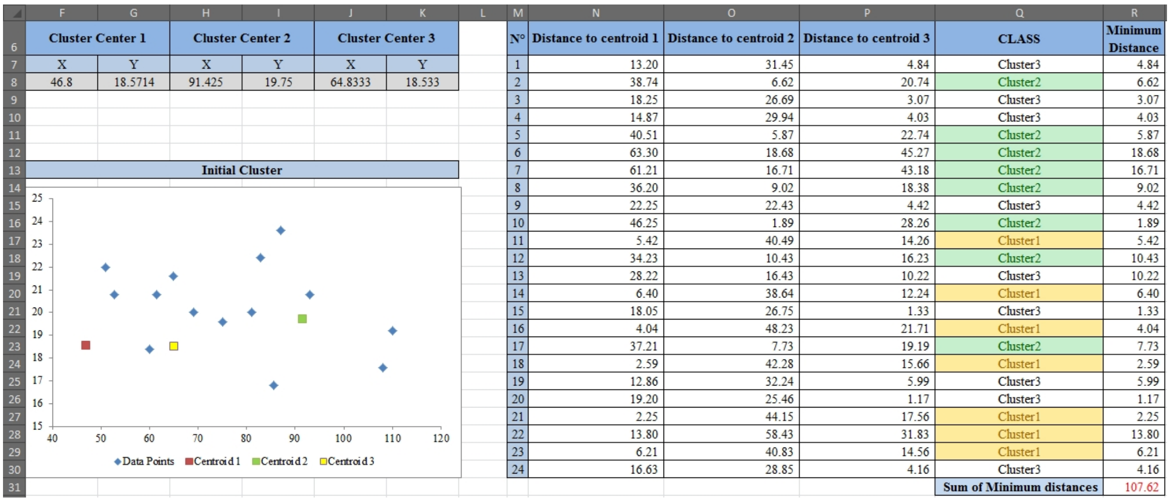


	F	G	H	I	J	K
6	Cluster Center 1		Cluster Center 2		Cluster Center 3	
7	X	Y	X	Y	X	Y
8	83.37348337	21.08209831	61.50173906	20.79513852	108.8648788	18.24379682
13	Initial Cluster					

N°	Distance to centroid 1	Distance to centroid 2	Distance to centroid 3	CLASS	Minimum Distance
1	23.53	2.83	48.87	Cluster2	2.83
2	4.78	24.33	23.41	Cluster1	4.78
3	18.58	3.40	44.19	Cluster2	3.40
4	21.88	0.01	47.43	Cluster2	0.01
5	4.41	25.65	22.51	Cluster1	4.41
6	26.79	48.62	1.56	Cluster3	1.56
7	24.87	46.61	1.08	Cluster3	1.08
8	1.44	21.36	26.39	Cluster1	1.44
9	14.41	7.54	39.90	Cluster2	7.54
10	9.63	31.50	16.07	Cluster1	9.63
11	32.39	10.57	57.99	Cluster2	10.57
12	2.61	19.51	27.92	Cluster1	2.61
13	8.50	13.55	33.89	Cluster1	8.50
14	30.57	8.70	56.12	Cluster2	8.70
15	18.97	4.88	44.08	Cluster2	4.88
16	40.18	18.31	65.70	Cluster2	18.31
17	3.54	22.72	24.87	Cluster1	3.54
18	34.35	12.71	59.67	Cluster2	12.71
19	24.51	5.24	49.52	Cluster2	5.24
20	17.58	5.10	42.87	Cluster2	5.10
21	36.28	14.77	61.49	Cluster2	14.77
22	50.43	28.57	75.87	Cluster2	28.57
23	33.14	12.51	58.02	Cluster2	12.51
24	21.32	6.18	45.99	Cluster2	6.18
				Sum of Minimum distances	67.06

Figure 14.154: k-means final associations in Microsoft® Excel® 14.0.6123

A software like Minitab 15.1.1 gives us other values for the centroids. As the latter don't give any plots let us see how the associations looks like if we write manually the centroid values given by Minitab in our Microsoft® Excel® sheet:



Cluster Center 1		Cluster Center 2		Cluster Center 3	
X	Y	X	Y	X	Y
46.8	18.5714	91.425	19.75	64.8333	18.533

Initial Cluster

N°	Distance to centroid 1	Distance to centroid 2	Distance to centroid 3	CLASS	Minimum Distance
1	13.20	31.45	4.84	Cluster3	4.84
2	38.74	6.62	20.74	Cluster2	6.62
3	18.25	26.69	3.07	Cluster3	3.07
4	14.87	29.94	4.03	Cluster3	4.03
5	40.51	5.87	22.74	Cluster2	5.87
6	63.30	18.68	45.27	Cluster2	18.68
7	61.21	16.71	43.18	Cluster2	16.71
8	36.20	9.02	18.38	Cluster2	9.02
9	22.25	22.43	4.42	Cluster3	4.42
10	46.25	1.89	28.26	Cluster2	1.89
11	5.42	40.49	14.26	Cluster1	5.42
12	34.23	10.43	16.23	Cluster2	10.43
13	28.22	16.43	10.22	Cluster3	10.22
14	6.40	38.64	12.24	Cluster1	6.40
15	18.05	26.75	1.33	Cluster3	1.33
16	4.04	48.23	21.71	Cluster1	4.04
17	37.21	7.73	19.19	Cluster2	7.73
18	2.59	42.28	15.66	Cluster1	2.59
19	12.86	32.24	5.99	Cluster3	5.99
20	19.20	25.46	1.17	Cluster3	1.17
21	2.25	44.15	17.56	Cluster1	2.25
22	13.80	58.43	31.83	Cluster1	13.80
23	6.21	40.83	14.56	Cluster1	6.21
24	16.63	28.85	4.16	Cluster3	4.16
				Sum of Minimum distances	107.62

Figure 14.155: k**-means final associations with Minitab 15.1.1 values**

The huge difference between Microsoft® Excel® and any other statistical software is quite simple to explain! Software such as Microsoft® Excel® 14.0.6123 minimizes the distance points to the centers but is unable simultaneously to maximize the distance between the centers. Again the statistical software have algorithms implemented for this purpose. More in details the idea is the following:

The quality/assessment of a clustering[61] is many times measured as following:

- We use the "cluster cohesion", also named "with sum of squares", already introduced earlier (we just generalize the notation without expliciting the type of distance/metric):

$$\text{WSS}=\sum_{k=1}^{K}\sum_{C(i)=k} d^2(\vec{x}_i-\vec{c}_k) \qquad (14.1.1792)$$

  In practice the purpose is to minimize it.

- We introduce the "cluster separation", also named "between sum of squares", defined by:

$$\text{BSS}=\sum_{k=1}^{K} n_k d^2(\vec{c}_k,\vec{C}) \qquad (14.1.1793)$$

  where n_k is the number of points in cluster number k and C is the barycenter of all clusters. In practice the purpose is to maximize it.

Using the properties of the barycenter (see section Geometry page 2635), we can then write the total sum of squares:

$$\boxed{\underbrace{\sum_{i=1}^{n} d^2(\vec{x}_i,C)}_{\text{TSS}}=\underbrace{\sum_{k=1}^{K}\sum_{C(i)=k} d^2(\vec{x}_i-\vec{c}_k)}_{\text{WSS}}+\underbrace{\sum_{k=1}^{K} n_k d^2(\vec{c}_k,\vec{C})}_{\text{BSS}}} \qquad (14.1.1794)$$

---

[61] In an unsupervised learning setting, it is often hard to assess the performance of a model since we don't have the ground truth labels as was the case in the supervised learning setting.

That latter relation is also often written as:

$$\underbrace{\sum_{q=1}^{Q}\sum_{i=1}^{I}(x_{iq}-\bar{x})^2}_{\text{Total inertia}}=\underbrace{\sum_{q=1}^{Q}\sum_{i=1}^{I}(x_{iq}-\bar{x}_q)^2}_{\text{Intra inertia}}+\underbrace{\sum_{q=1}^{Q}\sum_{i=1}^{I}(\bar{x}_q-\bar{x})^2}_{\text{Inter inertia}} \tag{14.1.1795}$$

with the following companion illustration:

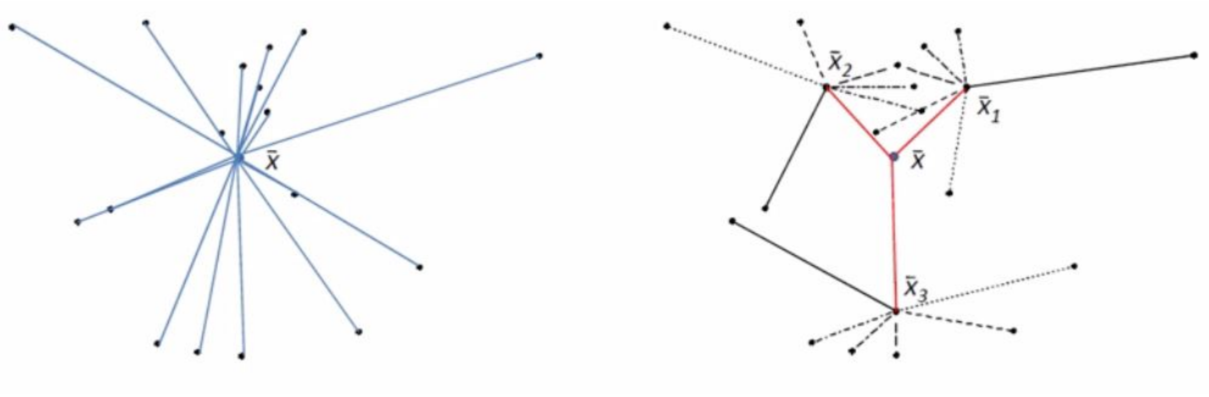

$\bar{x}$
$\bar{x}_2$
$\bar{x}_1$
$\bar{x}$
$\bar{x}_3$

Figure 14.156: k-means BSS and WSS

So in practice, the k-means method is fully encompassed based on the above fundamental relation. It is then quite easy to understand that if Intra-cluster inertia decreases the inter-cluster inertia increases and vice versa.

Therefore with Tanagra 1.4.44 we get for example:

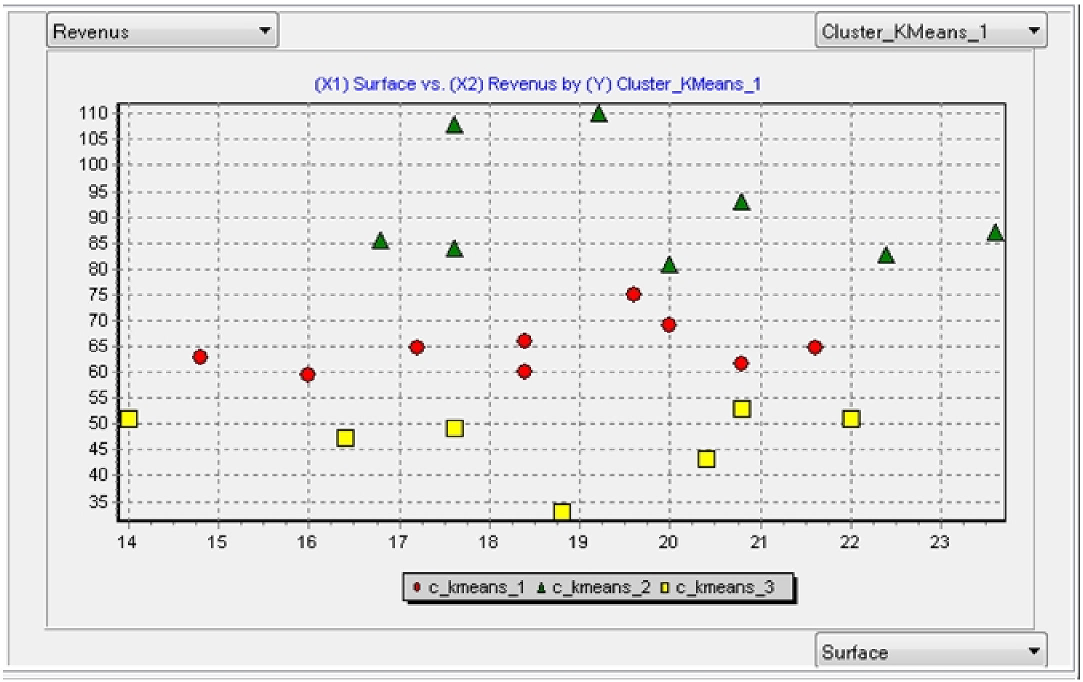
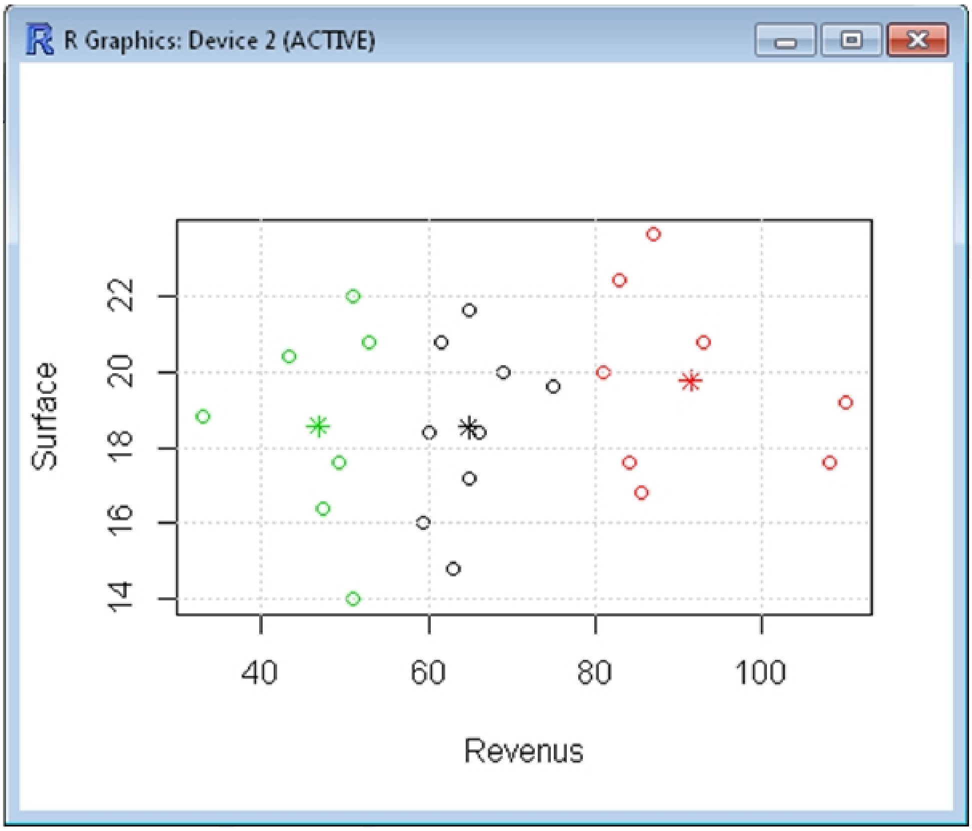
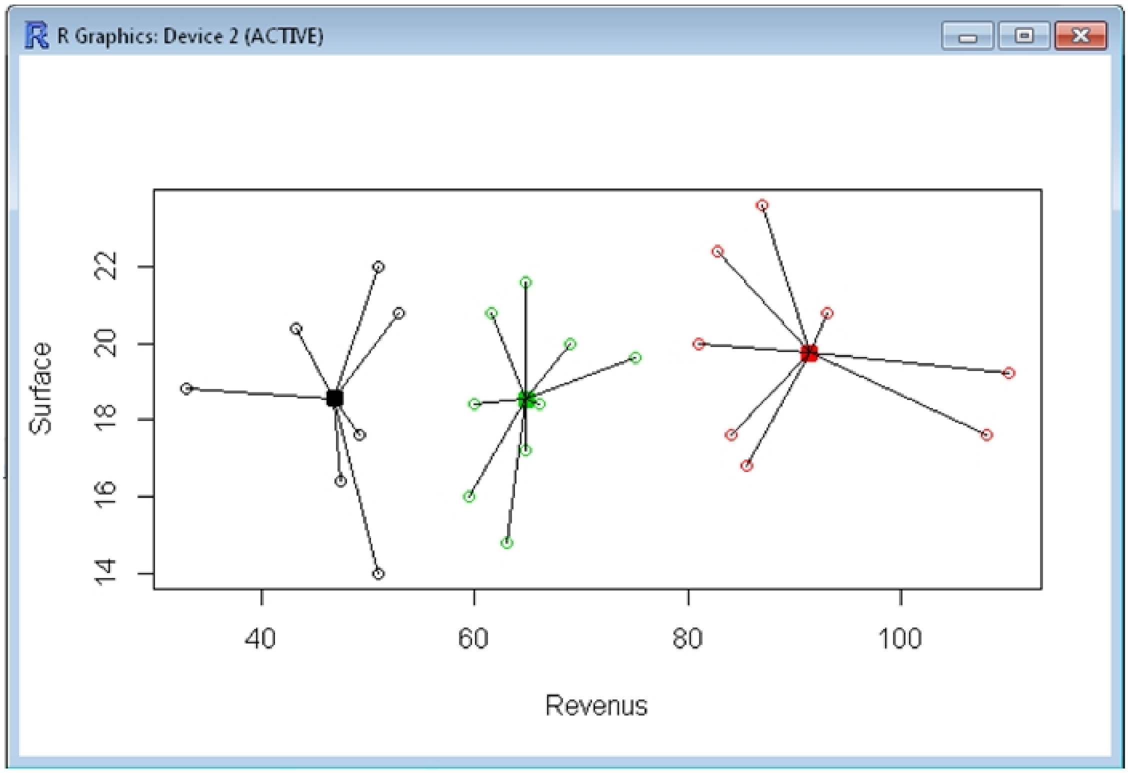

Revenus
Cluster_KMeans_1
(X1) Surface vs. (X2) Revenus by (Y) Cluster_KMeans_1
110
105
100
95
90
85
80
75
70
65
60
55
50
45
40
35
14
15
16
17
18
19
20
21
22
23
● c_kmeans_1 ▲ c_kmeans_2 □ c_kmeans_3
Surface

Figure 14.157: k-means final associations in Tanagra 1.4.44

and using the detailed steps given in our `R` companion book we get:


R Graphics: Device 2 (ACTIVE)
Surface
14
16
18
20
22
40
60
80
100
Revenus


R Graphics: Device 2 (ACTIVE)
Surface
14
16
18
20
22
40
60
80
100
Revenus

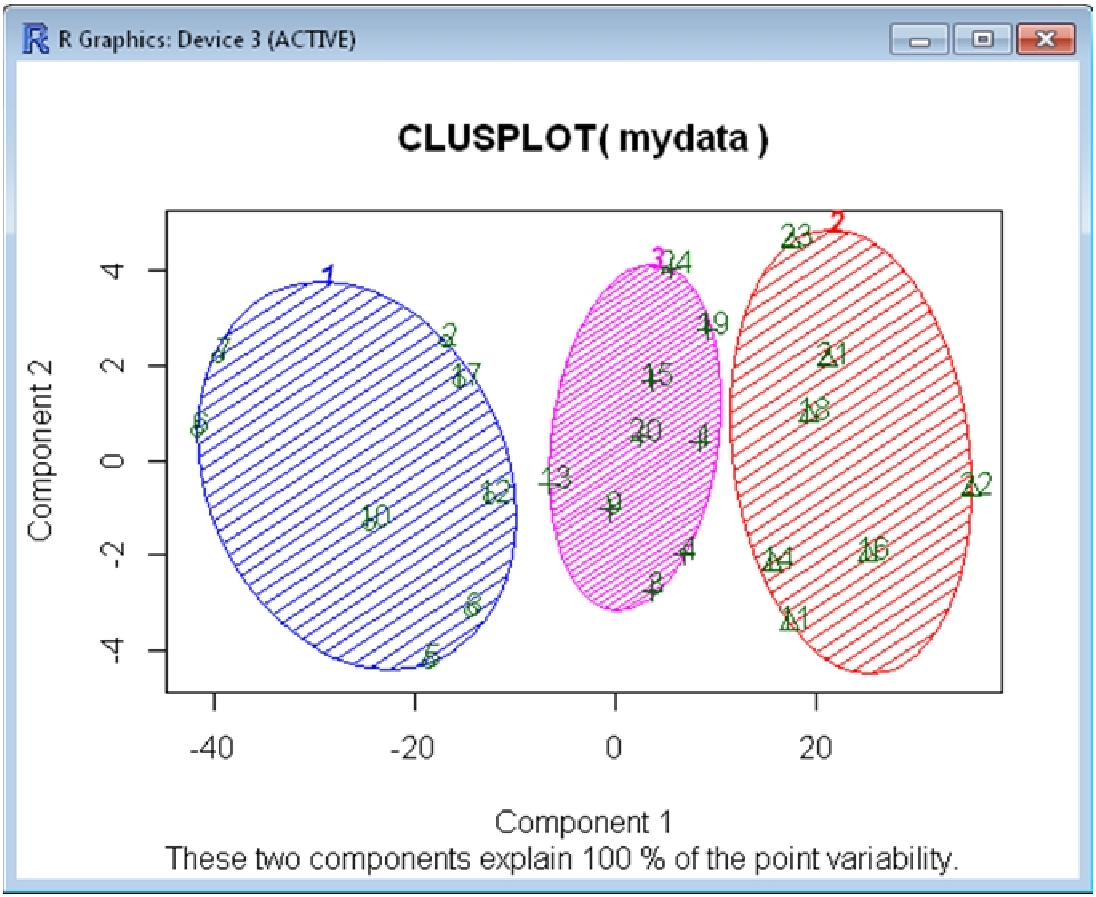
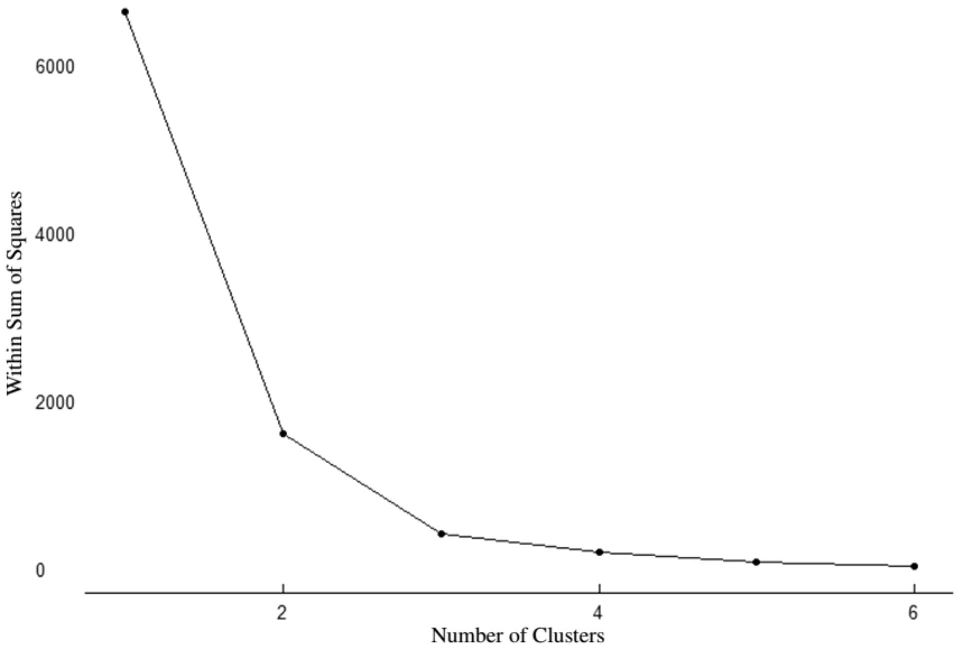


Figure 14.158: k-means final associations in `R` 3.0.2

A major difficulty with the k-means is to choose the number of clusters. For this let us recall the fundamental relation of k-means:

$$\underbrace{\sum_{i=1}^{n} d^2(x_i,C)}_{\text{TSS}}=\underbrace{\sum_{k=1}^{K}\sum_{C(i)=k} d^2(\vec{X}_i-\vec{c}_k)}_{\text{WSS}}+\underbrace{\sum_{k=1}^{K} n_k d^2(\vec{c}_k,\vec{C})}_{\text{BSS}} \tag{14.1.1796}$$

The strategy commonly chosen to choose the right K is to make vary WSS as a function of k and to take the last of k which induces a considerable reduction of WSS. Geometrically speaking, at this optimal point, we observe a kind of "elbow" (or "knee") forming on the plot of $\text{WSS}=f(K)$:



In the case of our first k-means clustering, the partition in $K=2$ is more interesting, because not only does it induce a considerable gain in information (according to the figure above), thus a significant loss or reduction of WSS, but also it offers clusters having more than 2 individuals. Beyond this, for example, for $K=3$, the gain in information is no longer significant and we will obtain a cluster with only 1 individual (not very relevant for a group analysis...).

In addition to homogeneity of the observations within each group, we also seek heterogeneity of the groups; hence, intuitively, an F -like statistic as a ratio of the between-group dissimilarity to the within-group dissimilarity could be used to indicate the goodness of a given clustering. We define a "pseudo F":

$$\widetilde{F}_K=\dfrac{\text{BSS}/(K-1)}{\text{WSS}/(n-K)} \tag{14.1.1797}$$

This measure is also named the "Calinski-Harabasz index". The larger that ratio, the better the clustering at any fixed number of clusters.

A practical approach to the problem is to vary K, as $K_0,K_0+1,\ldots$, computing $\widetilde{F}_K$, which will initially increase fairly rapidly, and to choose the value of K as the point at which $\widetilde{F}_{K+1}-\widetilde{F}_K$ is relatively small. This is similar to the type of approach in a different context; that is, choosing the number of principal components.

Remarks

R1. In some applications we want each center to be one of the point itself. This is where "K-medoids" (partition around medoids) comes in an algorithm similar to the k-means algorithm, except when fitting the centers $c_1,\ldots,c_K$, we restrict our attention to the point themselves. Generally K-medoids obviously return a higher value of K_{means}. There is also on example in our R companion book.

R2. Obviously as for the k-nn, we may run the k-means on the factorial space obtained after having run a PCA (Principal Component Analysis as seen on page 1213) on our data. We can also even reduce the dimensions using that latter if necessary!

A very common simple assessment metric is the "silhouette coefficient" given as graphical output in most statistical software and defined as given an observation $\vec{x} \in \mathcal{D}$ by:

$$\boxed{s(\vec{x}) = \frac{b(\vec{x}) - a(\vec{x})}{\max(a(\vec{x}), b(\vec{x}))}} \tag{14.1.1798}$$

where $a(\vec{x})$ is the average distance of $\vec{x}$ to all other elements in the cluster it belongs to and $b(\vec{x})$ is the smallest value that could take $a(\vec{x})$, if a belonged to another cluster:

$$a(\vec{x}) = \frac{1}{\left|\mathcal{C}_{k(\vec{x})}\right| - 1} \sum_{\vec{u} \in \mathcal{C}_{k(\vec{x})}, \vec{u} \neq \vec{x}} d(\vec{u}, \vec{x}) \qquad b(\vec{x}) = \min_{l \neq k(\vec{x})} \frac{1}{|\mathcal{C}_l|} \sum_{\vec{u} \in \mathcal{C}_l} d(\vec{u}, \vec{x}) \tag{14.1.1799}$$

The silhouette coefficient of $\vec{x}$ is all the closer to 1 as its assignment to the cluster $\mathcal{C}_{k(\vec{x})}$ is satisfactory.

The "global silhouette coefficient" of the clustering is the average of all silhouette coefficients:

$$s = \frac{1}{n} \sum_{i=1}^{n} s\left(\vec{x}^i\right) \tag{14.1.1800}$$

14.1.20.5.5 Support Vector Machines (SVM) classifier

"Support vector machines" (SVM) is learning method used for binary classification[62] that searches for so-named "support vectors" which are observations that are found to lie at the edge of an area in space which presents a boundary between one of these classes of observations (e.g., the squares) and another class of observations (e.g., the circles). SVM seems to have been introduced by Vladimir Vapnik and colleagues. The earliest mention was in (Vapnik, 11979 according to holocene calendar), but the first main paper seems to be (Vapnik, 11995 according to holocene calendar). In the terminology of SVM we talk about the space between these two regions as the "margin" between the classes. Each region contains observations with the same value for the target variable (i.e., the class). The support vectors, and only the support vectors, are used to identify a hyperplane (a straight line in two dimensions) that separates the classes. The maximum margin between the separable classes is sought. This then represents the model.

---

[62] In its most simple type, SVM doesn't support multiclass classification natively. It supports binary classification and separating data points into two classes. For multiclass classification, the same principle is utilized after breaking down the multiclassification problem into multiple binary classification problems.

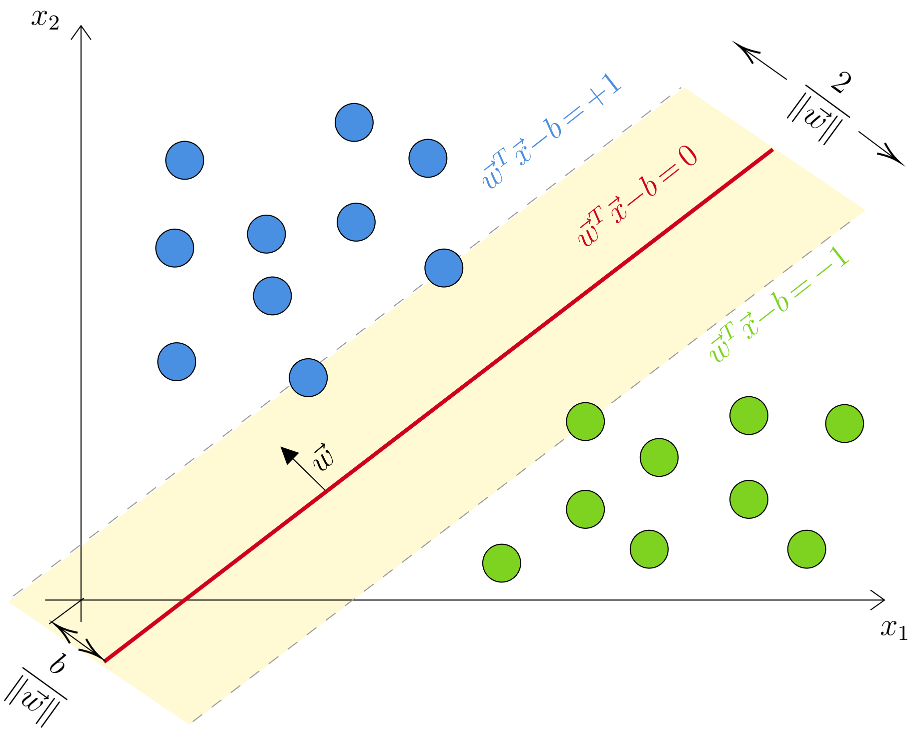


Figure 14.159: SVM idea (source: Wikipedia)

It is usually quite rare that we can separate the data with a straight line (or a hyperplane when we have more than two input variables). That is, the data is not usually distributed in such a way that it is "linearly separable". When this is the case, a technique is used to combine (or remap) the data in different ways, creating new variables so that the classes are then more likely to become linearly separable by a hyperplane (i.e., so that with the new dimensional data there is a gap between observations in the two classes). We can use the model we have built to score new observations by mapping the data in the same way as when the model was built, and then decide on which side of the hyperplane the observation lies and hence the decision associated with it.

Support vector machines have been found to perform well on problems that are non-linear, sparse, and high-dimensional. A disadvantage is that the algorithm is sensitive to the choice of tuning option (e.g., the type of transformations to perform), making it harder to use and time consuming to identify the best model. Another disadvantage is that the transformations performed can be computationally expensive and are performed both whilst building the model and when scoring new data.

An advantage of the method is that the modelling only deals with these support vectors rather than the whole training dataset, and so the size of the training set is not usually an issue. Also, as a consequence of only using the support vectors to build a model, the model is less affected by outliers.

Ok let us first introduce the idea in a gently way...:

Let $\mathcal{D}=\{(\vec{x}_i,y_i)\}_{i=1}^n$ be a classification dataset, with n points in a d-dimensional space. Further, let us assume that there are only two class labels, that is, $y_i\in\{+1,-1\}$ denoting the positive and negative classes.

A hyperplane in d dimensions is given as the set of all points $\vec{x}\in\mathbb{R}^d$ that satisfy the equation $h(\vec{x})=0$, where $h(\vec{x})$ is the hyperplane function, defined as follows (see section Analytical Geometry page 2800):

$$h(\vec{x})=\vec{w}^T\vec{x}+b=w_1x_1+w_2x_2+\ldots+w_dx_d+b \tag{14.1.1801}$$

Here, $\vec{w}$ is a dimensional weight vector and b is a scalar, named without surprise the "bias". For points that lie on the hyperplane, we have:

$$h(\vec{x})=\vec{w}^T\vec{x}+b=0 \tag{14.1.1802}$$

The hyperplane is thus defined as the set of all points such that $\vec{w}^T\vec{x}=-b$. To see the role played by b, assuming that $w_1\neq 0$, and setting $x_i=0$ for all $i>1$, we can obtain the offset where the hyperplane intersects the first axis. Therefore we have obviously:

$$w_1x_1=-b \quad \text{or} \quad x_1=\frac{-b}{w_1} \tag{14.1.1803}$$

In other words, the point:

$$\left(\frac{-b}{w_1},0,\ldots,0\right) \tag{14.1.1804}$$

lies on the hyperplane (and it's module is equal to the perpendicular distance from the hyperplane to the origin!). In a similar manner, we can obtain the offset where the hyperplane intersects each of the axes, which is given as $-b/w_i$ (provided $w_i\neq 0$).

A hyperplane splits the original d -dimensional space into two half-spaces (remember the first figure above depicting the idea!). A dataset is said to be "linearly separable" if each half-space has points only from a single class. If the input dataset is linearly separable, then we can find a separating hyperplane $h(\vec{x})=0$, such that for all points labelled $y_i=-1$, we have $h\left(\vec{x}_i\right)<0$, and for all points labelled $y_i=+1$, we have $h\left(\vec{x}_i\right)>0$. In fact, the hyperplane function $h(\vec{x})$ serves as a linear classifier or a linear discriminant, which predicts the class y for any given point x, according to the decision rule:

$$y=\begin{cases}+1 & \text{if } h(\vec{x})>0\\ -1 & \text{if } h(\vec{x})<0\end{cases} \tag{14.1.1805}$$

Let $\vec{a}_1$ and $\vec{a}_2$ be two arbitrary points that lie on the hyperplane. From the hyperplane equation we have:

$$\begin{aligned}h\left(\vec{a}_1\right)=\vec{w}^T\vec{a}_1+b=0\\ h\left(\vec{a}_2\right)=\vec{w}^T\vec{a}_2+b=0\end{aligned} \tag{14.1.1806}$$

Subtracting one from the other we obtain:

$$\vec{w}^T\left(\vec{a}_1-\vec{a}_2\right)=0 \tag{14.1.1807}$$

This means without surprise (as we also proved it in the section of Analytical Geometry) that the weight vector $\vec{w}$ is orthogonal to the hyperplane because it is orthogonal to any arbitrary vector $\left(\vec{a}_1-\vec{a}_2\right)$ on the hyperplane. In other words, the weight vector $\vec{w}$ specifies the direction that is normal to the hyperplane, which fixes the orientation of the hyperplane, whereas the bias b fixes the offset of the hyperplane in the d-dimensional space. Because both $\vec{w}$ and $-\vec{w}$ are normal to the hyperplane, we remove this ambiguity by requiring that $h(\vec{x}_i)>0$ when $y_i=1$, and $h(\vec{x}_i)<0$ when $y_i=-1$.

Ok now that we have finished with the gentle introduction let's go more in the maths with a text completely inspired from [185]!

We are given l training examples $\{\vec{x}_i,y_i\},\; i=1,\ldots,L$, where each example has d inputs ($\vec{x}_i\in\mathbb{R}^d$), and a class label with one of two values $y_i\in\{-1,1\}$.

Now, all hyperplanes in $\mathbb{R}^d$ are parametrized by a vector ($\vec{w}$), and a constant (b), expressed in the equation:

$$\vec{w}\circ\vec{x}+b=0 \tag{14.1.1808}$$

(recall again that $\vec{w}$ is in fact the vector orthogonal to the hyperplane). Given such a hyperplane ($\vec{w}$,b) that separates the data, this gives the function:

$$f(\vec{x})=\text{sgn}(\vec{w}\circ\vec{x}+b) \tag{14.1.1809}$$

which correctly classifies the training data (and hopefully other "testing" data it hasn't seen yet).

However, a given hyperplane represented by ($\vec{w}$,b) is equally expressed by all pairs $\{\lambda\vec{w},\lambda b\}$ for $\lambda\in\mathbb{R}^+$. So we define the "canonical hyperplane" to be that which separates the data from the hyperplane by a "distance" of at least[63] 1. That is, we consider those that satisfy:

$$\begin{aligned}\vec{x}_i\circ\vec{w}+b\geq+1 &\quad\text{when}\quad y_i=+1\\ \vec{x}_i\circ\vec{w}+b\leq-1 &\quad\text{when}\quad y_i=-1\end{aligned} \tag{14.1.1810}$$

or more compactly:

$$y_i(\vec{x}_i\circ\vec{w}+b)\geq 1 \qquad \forall i \tag{14.1.1811}$$

That we often write as:

$$y_i(\vec{x}_i\circ\vec{w}+b)-1\geq 0 \qquad \forall i \tag{14.1.1812}$$

All such hyperplanes have a "functional distance" ≥ 1 (quite literally, the function's value is ≥ 1). This shouldn't be confused with the "geometric" or "Euclidean distance" (also known as the "margin"). For a given hyperplane ($\vec{w}$,b), all pairs $\{\lambda\vec{w},\lambda b\}$ define the exact same hyperplane, but each has a different functional distance to a given data point.

To obtain the geometric distance from the hyperplane to a data point, we must normalize by the magnitude of $\vec{w}$. This distance is simply:

$$d\Big((\vec{w},b)\;,\;\vec{x}_i\Big)=\frac{y_i(\vec{x}_i\circ\vec{w}+b)}{\|\vec{w}\|}\geq\frac{1}{\|\vec{w}\|} \tag{14.1.1813}$$

Intuitively, we want the hyperplane that maximizes the geometric distance to the closest data points.

---

[63] In fact, we require that at least one example on both sides has a distance of *exactly* 1. Thus, for a given hyperplane, the scaling (the λ) is implicitly set.

As simple vector geometry shows that the margin is equal to $1/\|\vec{w}\|$ and maximizing it subject to the constraint in:

$$y_i(\vec{x}_i \circ \vec{w} + b) - 1 \geq 0 \quad \forall i \tag{14.1.1814}$$

is equivalent to find:

$$\min \|\vec{w}\| \quad \text{such that} \quad y_i\left(\vec{x}_i \circ \vec{w} + b\right) - 1 \geq 0 \quad \forall_i \tag{14.1.1815}$$

Minimize $\|\vec{w}\|$ is equivalent to minimize $\frac{1}{2}\|\vec{w}\|^2$ and the use of this term makes it possible to perform Nonlinear Programming (see page 6662) optimization later on. We therefore need to find:

$$\min \frac{1}{2}\|\vec{w}\|^2 \quad \text{s.t.} \quad y_i\left(\vec{x}_i \circ \vec{w} + b\right) - 1 \geq 0 \quad \forall_i \tag{14.1.1816}$$

In order to cater for the constraints in this minimization, we need to allocate them Lagrange multipliers (see page 6667) α, where[64] $\alpha_i \geq 0\,\forall_i$:

$$\begin{aligned} L_P &= \frac{1}{2}\|\vec{w}\|^2 - \alpha\left[y_i\left(\vec{x}_i \circ \vec{w} + b\right) - 1\right] \quad \forall_i \\ &= \frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{L} \alpha_i\left[y_i\left(\vec{x}_i \circ \vec{w} + b\right) - 1\right] \\ &= \frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{L} \alpha_i y_i\left(\vec{x}_i \circ \vec{w} + b\right) + \sum_{i=1}^{L} \alpha_i \end{aligned} \tag{14.1.1817}$$

We wish to find the $\vec{w}$ and b which minimizes, and the $\vec{\alpha}$ which maximizes the above relation (whilst keeping $\alpha_i \geq 0\,\forall_i$). We can do this by differentiating L_P with respect to $\vec{w}$ and b and setting the derivatives to zero:

$$\begin{aligned} \frac{\partial L_P}{\partial \vec{w}} = 0 \Rightarrow \vec{w} = \sum_{i=1}^{L} \alpha_i y_i \vec{x}_i \\ \frac{\partial L_P}{\partial b} = 0 \Rightarrow \sum_{i=1}^{L} \alpha_i y_i = 0 \end{aligned} \tag{14.1.1818}$$

Substituting these two relations into:

$$L_P = \frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{L} \alpha_i y_i\left(\vec{x}_i \circ \vec{w} + b\right) + \sum_{i=1}^{L} \alpha_i \tag{14.1.1819}$$

---

[64] These conditions are necessary for the primal optimization problem L_P to be strictly equal to the dual problem L_D, such that $L_P = L_D$ and not just equal or less such that $L_D \leq L_P$.

gives a new formulation which, being dependent on $\vec{\alpha}$, we need to maximize:

$$\begin{aligned}
L_D &= \frac{1}{2}\left(\sum_{i=1}^{L}\alpha_i y_i \vec{x}_i\right)\left(\sum_{j=1}^{L}\alpha_j y_j \vec{x}_j\right) - \sum_{i=1}^{L}\alpha_i y_i\left(\vec{x}_i \circ \sum_{j=1}^{L}\alpha_j y_j \vec{x}_j + b\right) + \sum_{i=1}^{L}\alpha_i \\
&= \frac{1}{2}\left(\sum_{i=1}^{L}\alpha_i y_i \vec{x}_i\right)\left(\sum_{j=1}^{L}\alpha_j y_j \vec{x}_j\right) - \sum_{i=1}^{L}\alpha_i y_i \vec{x}_i \circ \sum_{j=1}^{L}\alpha_j y_j \vec{x}_j - b\underbrace{\sum_{i=1}^{L}\alpha_i y_i}_{=0} + \sum_{i=1}^{L}\alpha_i \\
&= \sum_{i=1}^{L}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \vec{x}_i \circ \vec{x}_j \qquad \text{s.t. } \alpha_i \geq 0\,\forall_i, \sum_{i=1}^{L}\alpha_i y_i = 0 \\
&= \sum_{i=1}^{L}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i H_{ij}\alpha_j \qquad \text{where: } H_{ij} = y_i y_j \vec{x}_i \circ \vec{x}_j \\
&= \sum_{i=1}^{L}\alpha_i - \frac{1}{2}\vec{\alpha}^T H \vec{\alpha} \qquad \text{s.t. } \alpha_i \geq 0\,\forall_i, \sum_{i=1}^{L}\alpha_i y_i = 0
\end{aligned} \tag{14.1.1820}$$

This new formulation L_D is referred to as we know to as the "Dual form" of the Primary L_P. And the corresponding constraint (conditions) are named the "Karush-Kuhn-Tucker conditions" (there exist other possible conditions but there are out of the scope of this textbook). It is worth noting that the Dual form requires only the dot product of each input vector $\vec{x}_i$ to be calculated, this is important for the Kernel Trick described further below.

Remark

The matrix H above is often written $H_{ij} \equiv y_i y_j k(\vec{x}_i, \vec{x}_j)$ where $k(\vec{x}_i, \vec{x}_j)$ is an example of a family of functions named "kernel functions". The special case $k(\vec{x}_i, \vec{x}_j) = \vec{x}_i \circ \vec{x}_j$ is known as a the "linear kernel".

The reason that this kernel trick is useful is that there are many classification/regression problems that are not linearly separable/regressable in the space of the inputs $\vec{x}$, which might be in a higher dimensionality feature space given a suitable mapping $\vec{x} \mapsto \Phi(\vec{x})$. For example the "radial kernel":

$$k\left(\vec{x}_i, \vec{x}_j\right) = e^{-\frac{\left\|\vec{x}_i - \vec{x}_j\right\|^2}{2\sigma^2}}$$

Other popular kernels (among many others) for classification and regression are the "polynomial kernel":

$$k\left(\vec{x}_i, \vec{x}_j\right) = \left(\vec{x}_i \circ \vec{x}_j + a\right)^b$$

and the "sigmoidal kernel":

$$k\left(\vec{x}_i, \vec{x}_j\right) = \tanh\left(a\vec{x}_i \circ \vec{x}_j - b\right)$$

where a and b are parameters defining the kernel's behaviour.

Having moved from minimizing L_P to maximizing L_D, we need to find:

$$\max_{\alpha}\left[\sum_{i=1}^{L}\alpha_i-\frac{1}{2}\vec{\alpha}^T H\vec{\alpha}\right] \quad \text{s.t.} \quad \alpha_i\geq 0 \quad \forall_i \quad \text{and} \quad \sum_{i=1}^{L}\alpha_i y_i=0 \tag{14.1.1821}$$

This is a convex quadratic optimization problem, and we run a solver which will return $\vec{\alpha}$ and from:

$$\vec{w}=\sum_{i=1}^{L}\alpha_i y_i\vec{x}_i \tag{14.1.1822}$$

will give us $\vec{w}$. What remains is to calculate b.

Any data point satisfying:

$$\sum_{i=1}^{L}\alpha_i y_i=0 \tag{14.1.1823}$$

which is a support vector denoted $\vec{x}_s$ (samples that are on the gutter) will have the form:

$$y_s\left(\vec{x}_s\circ\vec{w}+b\right)=1 \tag{14.1.1824}$$

If we inject into it the previous derived relation:

$$\vec{w}=\sum_{i=1}^{L}\alpha_i y_i\vec{x}_i \tag{14.1.1825}$$

we get:

$$y_s\left(\sum_{m\in S}\alpha_m y_m\vec{x}_m\circ\vec{x}_s+b\right)=1 \tag{14.1.1826}$$

Where S denotes the set of indices of the support vectors. S is determined by finding the indices i where $\alpha_i>0$. Multiplying through by y_s and then using $y_s^2=1$ we have :

$$y_s^2\left(\sum_{m\in S}\alpha_m y_m\vec{x}_m\circ\vec{x}_s+b\right)=y_s \tag{14.1.1827}$$

Therefore:

$$b=y_s-\sum_{m\in S}\alpha_m y_m\vec{x}_m\circ\vec{x}_s \tag{14.1.1828}$$

Instead of using an arbitrary support vector x_s, it is better to take an average over all of the support vectors in S :

$$b=\frac{1}{N_s}\sum_{s\in S}\left(y_s-\sum_{m\in S}\alpha_m y_m\vec{x}_m\circ\vec{x}_s\right) \tag{14.1.1829}$$

We now have the variables $\vec{w}$ and b that define our separating hyperplane's optimal orientation and hence our Support Vector Machines.

So to summarize... In order to use a SVM to solve a linearly separable, binary classification problem we need to:

- Create the matrix H, where $H_{ij}=y_iy_j\vec{x}_i\circ\vec{x}_j$
- Find $\vec{\alpha}$ so that:

  $$\sum_{i=1}^{L}\alpha_i-\frac{1}{2}\vec{\alpha}^TH\vec{\alpha} \tag{14.1.1830}$$

  is maximized, subject to the constraints:

  $$\alpha_i\geq 0\,\forall_i \quad \text{and} \quad \sum_{i=1}^{L}\alpha_iy_i=0 \tag{14.1.1831}$$

  This is done using a nonlinear programming solver.
- Calculate:

  $$\vec{w}=\sum_{i=1}^{L}\alpha_iy_i\vec{x}_i \tag{14.1.1832}$$

- Determine the set of support vectors S by finding the indices such that $\alpha_i>0$
- Calculate:

  $$b=\frac{1}{N_s}\sum_{s\in S}\left(y_s-\sum_{m\in S}\alpha_my_m\vec{x}_m\circ\vec{x}_s\right) \tag{14.1.1833}$$

- Each new point $\vec{x}'$ is classified by evaluating:

  $$y'=\text{sgn}\left(\vec{w}\circ\vec{x}'+b\right) \tag{14.1.1834}$$

Remark

We recommend the reader interested in delving deeper into the subject to read the article *Support Vector Machines Explained* from Tristan Fletcher [185] which was used to write the above presentation. You will then see how we can extend the SVM methodology to handle data that is not fully linearly separable or what are the maths behind regression SVM.

14.1.20.5.6 Gaussian Mixture Model (GMM) clustering

A more formal approach to clustering can be achieved if we assume underlying distributions of the observations. This is also known as "finite mixture model clustering". Observations arise from a distribution that is a mixture of two or more components, or clusters. Each cluster is described by a density and has an associated probability or weight in the mixture.

We can think of building a "Gaussian Mixture Model" as a type of clustering algorithm. Using again the expectation-maximization algorithm seen earlier above (see page 6697), the process and result is very similar to k-means clustering. The difference is that the clusters are assumed to each have an independent Gaussian distribution, each with their own mean and covariance matrix.

When performing k-means clustering, we assign points to clusters using the straight Euclidean distance. The Euclidean distance is a poor metric, however, when the cluster contains significant covariance! The Gaussian Mixture Models approach will take cluster covariance into account when forming the clusters.

Another important difference with k-means is that standard k-means performs a hard assignment of data points to clusters, i.e. each point is assigned to the closest cluster. With Gaussian Mixture Models, what we will end up is a collection of independent multivariate Gaussian distributions, and so for each data point, we will have a probability that it belongs to each of these distributions / clusters.

Suppose we are given a data set $D=\{\vec{x}_1,\ldots,\vec{x}_N\}$ where $\vec{x}_i$ is a d-dimensional vector measurement. Let us assume that the points are generated in an independent and identically distributed way from an underlying density $P(\vec{x})$.

We further assume that $P(\vec{x})$ is defined as a finite mixture model with K components:

$$P(\vec{x}|\Theta)=\sum_{k=1}^{K}\alpha_k P_k(\vec{x}|z_k,\theta_k) \tag{14.1.1835}$$

where:

- The $P_k(\vec{x}|z_k,\theta_k)$ are mixture components, $1\leq k\leq K$. Each is a density or distribution defined over $P(\vec{x})$, with parameters θ_k.
- $\vec{z}=(z_1,\ldots,z_K)$ is a vector of K binary indicator variables that are mutually exclusive and exhaustive (i.e., one and only one of the z_k's is equal to 1, and the others are 0). $\vec{z}$ is a K-ary random variable representing the identity of the mixture component that generated $\vec{x}$. It is convenient for mixture models to represent z as a vector of K indicator variables.
- The $\alpha_k=P(z_k)$ are the mixture weights, representing the probability that a randomly selected $\vec{x}$ was generated by component k, where $\sum_{k=1}^{K}\alpha_k=1$.

The complete set of parameters for a mixture model with K components is:

$$\Theta=\{\alpha_1,\ldots,\alpha_K,\theta_1,\ldots,\theta_K\} \tag{14.1.1836}$$

We can compute the "membership weight" (i.e. hidden posterior) of data point $\vec{x}_i$ in cluster k, given parameters Θ as:

$$w_{i,k}=P(z_{i,k}=1|\vec{x}_i,\Theta)\;=\;\frac{P_k(\vec{x}_i|z_k,\theta_k)\cdot\alpha_k}{\displaystyle\sum_{m=1}^{K}P_m(\vec{x}_i|z_m,\theta_m)\cdot\alpha_m},\qquad 1\leq k\leq K,\quad 1\leq i\leq N \tag{14.1.1837}$$

This follows from a direct application of Bayes rule often written as:

$$w_{i,k}=P(z_{i,k}=k|\vec{x}_i,\Theta)=\frac{P(z_{i,k}=k|\Theta)P(\vec{x}_i|z_{i,k}=k,\Theta)}{P(\vec{x}_N|\Theta)}=\frac{\mathcal{N}(\vec{x}_i|\mu_k,\Sigma_k)\alpha_k}{\displaystyle\sum_{m=1}^{K}\mathcal{N}(\vec{x}_i|\mu_m,\Sigma_m)\alpha_m} \tag{14.1.1838}$$

The membership weights above reflect our uncertainty, given $\vec{x}_i$ and Θ, about which of the K components generated vector $\vec{x}_i$. Note that we are assuming in our generative mixture model that each $\vec{x}_i$ was generated by a single component so these probabilities reflect our uncertainty about which component $\vec{x}_i$ came from, not any "mixing" in the generative process.

For $\vec{x}\in\mathbb{R}^d$ we can define a Gaussian mixture model by making each of the K components a Gaussian density with parameters $\vec{\mu}_k$ and Σ_k.

Each component (cluster) is a multivariate Gaussian density given for recall by:

$$P_k(\vec{x}|\theta_k)\;=\;\frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}}e^{-\frac{1}{2}(\vec{x}-\vec{\mu}_k)^T\Sigma_k^{-1}(\vec{x}-\vec{\mu}_k)} \tag{14.1.1839}$$

with its own parameters $\theta_k=\{\vec{\mu}_k,\Sigma_k\}$.

We use the Expectation-Maximization algorithm for Gaussian mixtures as follows: The algorithm is an iterative algorithm that starts from some initial estimate of Θ (e.g., random), and then proceeds to iteratively update Θ until convergence is detected. Each iteration consists of an E-step and a M-step.

To kickstart the EM algorithm, we randomly select data points to use as the initial means, and we set the covariance matrix for each cluster to be equal to the covariance of the full training set. Also, we give each cluster equal prior probability. A cluster's prior probability is just the fraction of the dataset that belongs to each cluster. We start by assuming the dataset is equally divided between the clusters.

In other words:

- In the "Expectation" step (E-step), we will calculate the (posterior) probability that each data point $\vec{x^i}$ belongs to each cluster k (using our current estimated mean vectors and covariance matrices) following:

  $$P_k(\vec{x}_i|\theta_k)=\frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}}e^{-\frac{1}{2}(\vec{x}_i-\vec{\mu}_k)^T\Sigma_k^{-1}(\vec{x}_i-\vec{\mu}_k)} \tag{14.1.1840}$$

  This seems analogous to the cluster assignment step in k-means. Technically this step denote the current parameter values as Θ. It computes the $w_{i,k}$ (using the equation above for membership weights) for all data points $\vec{x}_i, 1\leq i\leq N$ and all mixture components $1\leq k\leq K$. Note that for each data point $\vec{x}_i$ the membership weights are defined such that $\sum_{k=1}^{K}w_{i,k}=1$. This yields an $N\times K$ matrix of membership weights, where each of the rows sum to 1.

- In the "Maximization" step (M-step), we will re-calculate the cluster means and covariances based on the probabilities calculated in the expectation step (i.e. we use the membership weights and the data to calculate new parameter values). This seems analogous to the cluster movement step in k-means.

Let $N_k = \sum_{i=1}^{N} w_{i,k}$, i.e., the sum of the membership weights for the kth component—this is the effective number of data points assigned to component k.

Specifically:

$$\alpha_k^{\text{new}} = \frac{N_k}{N} \qquad 1 \leq k \leq K \tag{14.1.1841}$$

These are the new mixture weights:

$$\vec{\mu}_k^{\text{new}} = \left(\frac{1}{N_k}\right) \sum_{i=1}^{N} w_{i,k}\vec{x}_i \qquad 1 \leq k \leq K \tag{14.1.1842}$$

The updated mean is calculated in a manner similar to how we could compute a standard empirical average, except that the ith data vector $\vec{x}_i$ has a fractional weight $w_{i,k}$.

Note that this is a vector equation since $\vec{\mu}_k^{\text{new}}$ and $\vec{x}_i$ are both d-dimensional vectors:

$$\Sigma_k^{\text{new}} = \left(\frac{1}{N_k}\right) \sum_{i=1}^{N} w_{ik} \cdot (\vec{x}_i - \vec{\mu}_k^{\text{new}})(\vec{x}_i - \vec{\mu}_k^{\text{new}})^T \qquad 1 \leq k \leq K \tag{14.1.1843}$$

Again we get an equation that is similar in shape to how we would normally compute an empirical covariance matrix, except that the contribution of each data point is weighted by $w_{i,k}$. Note that this is a matrix equation of dimensionality $d \times d$ on each side.

The equations in the M-step need to be computed in this order, i.e., first compute the K new α's, then the K new $\vec{\mu}_k$'s, and finally the K new Σ_k's.

After we have computed all of the new parameters, the M-step is complete and we can now go back and recompute the membership weights in the E-step, then recompute the parameters again in the E-step, and continue updating the parameters in this manner. Each pair of E and M steps is considered to be one iteration.

The EM algorithm can be started by either initializing the algorithm with a set of initial parameters and then conducting an E-step, or by starting with a set of initial weights and then doing a first M-step. The initial parameters or weights can be chosen randomly (e.g. select K random data points as initial means and select the covariance matrix of the whole data set for each of the initial K covariance matrices) or could be chosen via some heuristic method (such as by using the k-means algorithm to cluster the data first and then defining weights based on k-means memberships).

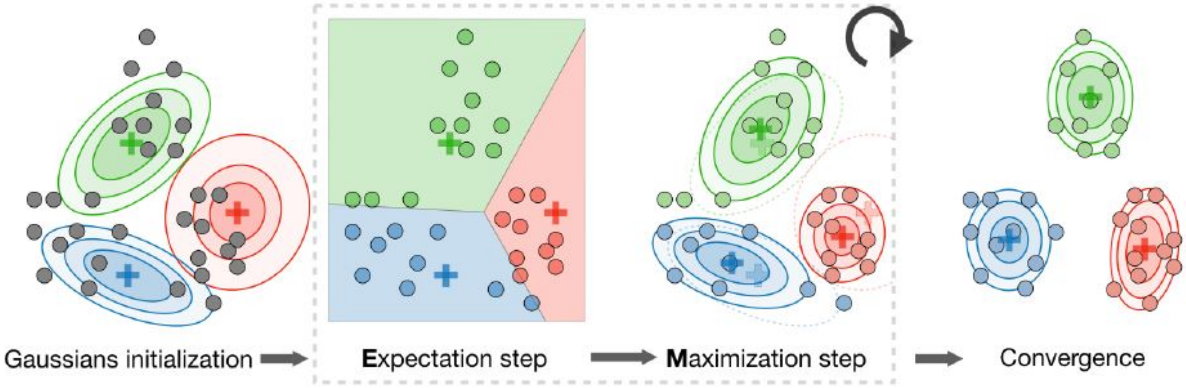


Figure 14.160: Expectation-Maximization iterative algorithm (authors: Afshine Amidi, Shervine Amidi)

Convergence is generally detected by computing the value of the log-likelihood after each iteration and halting when it appears not to be changing in a significant manner from one iteration to the next. Note that the log-likelihood (under the IID assumption) is defined as follows for this case:

$$\ln\left(L(\Theta)\right)=\ln\left(\prod_{i=1}^{N}P(\vec{x}_i|\Theta)\right)=\sum_{i=1}^{N}\ln\left(P(\vec{x}_i|\Theta)\right)=\sum_{i=1}^{N}\ln\left(\prod_{k=1}^{K}\alpha_k P_k(\vec{x}_i|z_k,\theta_k)\right) \tag{14.1.1844}$$

where $P_k(\vec{x}_i|z_k,\theta_k)$ is the Gaussian density for the kth mixture component.

Explicitly the complete data log-likelihood over all points $\{(\vec{x}_n,\vec{z}_n)\}_{n=1}^{N}$ is:

$$\begin{aligned}\ln\left(L(\Theta)\right)&=\sum_{i=1}^{N}\ln\left(\prod_{k=1}^{K}\alpha_k P_k(\vec{x}_i|z_k,\theta_k)\right)=\sum_{i=1}^{N}\ln\left(\prod_{k=1}^{K}\alpha_k\mathcal{N}(\vec{x}_n\,|\,\mu_k,\Sigma_k)^{\mathbb{I}(\vec{z}_n=k)}\right)\\&=\sum_{i=1}^{N}\sum_{k=1}^{K}\mathbb{I}(\vec{z}_n=k)\ln\left(\alpha_k\mathcal{N}(\vec{x}_n\,|\,\mu_k,\Sigma_k)\right)\end{aligned} \tag{14.1.1845}$$

We can now detail the Expectation-Maximization algorithm for Gaussian mixtures:

- **E-Step:**

  Before the E-step, we have an estimate θ_t of the parameters, and seek to compute a new lower bound on the observed log-likelihood. Earlier, we showed that the optimal lower bound is (see page 6704):

  $$L(q_{\theta_t},\theta)=\mathrm{E}_q[\ln\left(p(\vec{x},\vec{z}|\theta)\right)]+c^{te} \tag{14.1.1846}$$

  where $q_{\theta_t}(z)\equiv p(\vec{z}|\vec{x},\theta_t)$ and the second term is constant with respect to θ. The E-Step requires us to derive an expression for the first term. Using the relation above:

  $$\ln\left(L(\Theta)\right)=\sum_{i=1}^{N}\sum_{k=1}^{K}\mathbb{I}(\vec{z}_n=k)\ln\left(\alpha_k\mathcal{N}(\vec{x}_n\,|\,\mu_k,\Sigma_k)\right) \tag{14.1.1847}$$

the expected complete data log-likelihood is given by:

$$\begin{aligned}
Q(\theta_t,\theta)=\mathrm{E}_q[\ln p(\vec{x},\vec{z}|\theta)] &= \sum_{n=1}^{N}\sum_{k=1}^{K}\mathrm{E}_q\left[\mathbb{I}(z_n=k)\ln\left(\alpha_k\mathcal{N}(x_n\,|\,\mu_k,\Sigma_k)\right)\right]\\
&= \sum_{n=1}^{N}\sum_{k=1}^{K}\mathrm{E}_q\left[\mathbb{I}(z_n=k)\right]\ln\left(\alpha_k\mathcal{N}(x_n\,|\,\mu_k,\Sigma_k)\right)\\
&= \sum_{n=1}^{N}\sum_{k=1}^{K}P(z_n=k\,|\,x_n,\theta_t)\ln\left(\alpha_k\mathcal{N}(x_n\,|\,\mu_k,\Sigma_k)\right)\\
&= \sum_{n=1}^{N}\sum_{k=1}^{K}r_{nk}\ln(\alpha_k)+\sum_{n=1}^{N}\sum_{k=1}^{K}r_{nk}\ln\left(\mathcal{N}(x_n\,|\,\mu_k,\Sigma_k)\right)
\end{aligned} \tag{14.1.1848}$$

where $r_{nk}:=P(z_n=k\,|\,x_n,\theta_t)$ is named the "responsability" that cluster k takes for data point x_n after step t. During the E-Step, we compute these values explicitly with the relation seen earlier above:

$$w_{n,k}=P(z_{n,k}=k|\vec{x}_n,\Theta)=\frac{P(z_{n,k}=k|\Theta)P(\vec{x}_i|z_{i,k}=k,\Theta)}{P(\vec{x}_N|\Theta)}=\frac{\mathcal{N}(\vec{x}_n|\mu_k,\Sigma_k)\alpha_k}{\displaystyle\sum_{m=1}^{K}\mathcal{N}(\vec{x}_n|\mu_m,\Sigma_m)\alpha_m} \tag{14.1.1849}$$

- **M-Step:**

  During the M-Step, we optimize our lower bound with respect to the parameters $\theta=(\vec{\alpha},\vec{\mu},\vec{\Sigma})$. For the mixing weights $\vec{\pi}$, we use Lagrange multipliers to maximize the evidence lower bound (see page 6704) subject to the constraint $\sum_{k=1}^{K}\pi_k=1$. The Lagrangian (see page 6667) is:

$$L(\alpha,\lambda)=Q(\theta_t,\theta)+\lambda\left(\sum_{k=1}^{K}\alpha_k-1\right) \tag{14.1.1850}$$

  with for recall:

$$\begin{gathered}
Q(\theta_t,\theta)=\sum_{n=1}^{N}\sum_{k=1}^{K}r_{nk}\ln\left(\alpha_k\mathcal{N}(x_n\,|\,\mu_k,\Sigma_k)\right)\\
r_{nk}:=P(z_n=k\,|\,x_n,\theta_t)
\end{gathered} \tag{14.1.1851}$$

  So let us derive the four corresponding results... In order to resolve this problem, we have to find the parameters for which the partial derivatives are null:

$$\max_{\Theta,\lambda}L(\theta,r_{nk},\lambda)\Leftrightarrow\nabla_{\theta,\lambda}L(\theta_t,r_{nk},\lambda)=0\Leftrightarrow\begin{cases}\dfrac{\partial L(\theta_t,r_{nk},\lambda)}{\partial\alpha_k}=0\\ \dfrac{\partial L(\theta_t,r_{nk},\lambda)}{\partial\mu_k}=0\\ \dfrac{\partial L(\theta_t,r_{nk},\lambda)}{\partial\Sigma_k}=0\\ \dfrac{\partial L(\theta_t,r_{nk},\lambda)}{\partial\lambda}=0\end{cases} \tag{14.1.1852}$$

First, let's start with the equation with respect to λ:

$$\begin{aligned}\frac{\partial \mathcal{L}(\theta, r_{nk}, \lambda)}{\partial \lambda} = 0 \Leftrightarrow & \frac{\partial}{\partial \lambda}\left[\left(\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\ln\left(\alpha_k \mathcal{N}\left(x_n \,|\, \mu_k, \Sigma_k\right)\right)\right) - \lambda\left(\sum_{k=1}^{K}\alpha_k - 1\right)\right] = 0\\ & \frac{\partial}{\partial \lambda}\left[-\lambda\sum_{k=1}^{K}\alpha_k + \lambda\right] = 0\\ & -\sum_{k=1}^{K}\alpha_k + 1 = 0\\ & \sum_{k=1}^{K}\alpha_k = 1\end{aligned} \tag{14.1.1853}$$

Ok, so we are back to the definition of the constraint. No surprise here but not very useful.

Now let's move on with the equation with respect to α_k. Note that in the following derivations, we use the fact that deriving an α_j with j different than j results in a constant (this basically means that the summation over the K clusters can be ignored):

$$\begin{aligned}\frac{\partial L(\theta_t, r_{nk}, \lambda)}{\partial \alpha_k} = 0 \quad \Leftrightarrow & \frac{\partial}{\partial \alpha_k}\left[\left(\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\ln\left(\alpha_k \mathcal{N}\left(x_n \,|\, \mu_k, \Sigma_k\right)\right)\right) - \lambda\left(\sum_{k=1}^{K}\alpha_k - 1\right)\right] = 0\\ & \frac{\partial}{\partial \alpha_k}\left[\left(\sum_{n=1}^{K} r_{nk}\ln\left(\alpha_k \mathcal{N}\left(x_n \,|\, \mu_k, \Sigma_k\right)\right)\right) - \lambda\alpha_k\right] = 0\\ & \frac{\partial}{\partial \alpha_k}\left[\left(\sum_{n=1}^{N} r_{nk}\ln\left(\alpha_k\right) + \sum_{n=1}^{N} r_{nk}\ln\left(\mathcal{N}\left(x_n \,|\, \mu_k, \Sigma_k\right)\right)\right] - \lambda\alpha_k\right] = 0\\ & \frac{\partial}{\partial \alpha_k}\left[\sum_{n=1}^{N} r_{nk}\ln\left(\alpha_k\right) - \lambda\alpha_k\right] = 0\\ & \frac{\sum_{n=1}^{N} r_{nk}}{\alpha_k} - \lambda = 0\\ & \alpha_k = \frac{\sum_{n=1}^{N} r_{nk}}{\lambda}\end{aligned} \tag{14.1.1854}$$

Now in order to get rid of the λ, we use the definition of the constraint regarding the mixtures weights:

$$\begin{aligned}\sum_{k=1}^{K}\alpha_k = 1 \Leftrightarrow & \sum_{k=1}^{K}\frac{\sum_{n=1}^{N} r_{nk}}{\lambda} = 1\\ & \frac{\sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk}}{\lambda} = 1\\ & \lambda = \sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk}\end{aligned} \tag{14.1.1855}$$

And this gives us the final closed-form expressions for the mixture weights:

$$\alpha_k = \frac{\sum_{n=1}^{N} r_{nk}}{\sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk}} = \frac{\sum_{n=1}^{N} r_{nk}}{N} \tag{14.1.1856}$$

This tells us that for each class, after the M-Step, the mixture weight will be the sum of all the individual weights for that class normalized by the sum of all the individual weights for all the classes. And this makes perfect sense, if all the observations put a little weight on a specific class compared to the other classes than this class will have a small overall weight, and vice-versa.

Now let's move on the derivation of the mean vector μ. This one is a little trickier because, this time, we are computing the partial derivative for vectors and matrices:

$$\begin{aligned}\frac{\partial L(\theta_t,r_{nk},\lambda)}{\partial\mu_k}=0\Leftrightarrow&\frac{\partial}{\partial\mu_k}\left[\left(\sum_{n=1}^{N}\sum_{k=1}^{K}r_{nk}\ln\left(\alpha_k\mathcal{N}\left(x_n\,|\,\mu_k,\Sigma_k\right)\right)\right)-\lambda\left(\sum_{k=1}^{K}\alpha_k-1\right)\right]=0\\&\frac{\partial}{\partial\mu_k}\left[\sum_{n=1}^{N}r_{nk}\ln\left(\alpha_k\right)+\sum_{n=1}^{N}r_{nk}\ln\left(\mathcal{N}\left(x_n\,|\,\mu_k,\Sigma_k\right)\right)\right]=0\\&\frac{\partial}{\partial\mu_k}\Bigg[\sum_{n=1}^{N}r_{nk}\ln\left(\frac{1}{(2\pi)^{(N/2)}\left|\Sigma_k\right|^{1/2}}\exp\left(-\frac{1}{2}\left(x_n-\mu_k\right)^T\Sigma_k^{-1}\left(x_n-\mu_k\right)\right)\right)\Bigg]\Bigg]=0\\&\frac{\partial}{\partial\mu_k}\sum_{n=1}^{N}r_{nk}\ln\left(\frac{1}{(2\pi)^{(N/2)}\left|\Sigma_k\right|^{1/2}}\right)+\sum_{n=1}^{N}r_{nk}\ln\left(\exp\left(-\frac{1}{2}\left(x_n-\mu_k\right)^T\Sigma_k^{-1}\left(x_n-\mu_k\right)\right)\right)]=0\\&\frac{\partial}{\partial\mu_k}\left[\sum_{n=1}^{N}r_{nk}\left(-\frac{1}{2}\left(x_n-\mu_k\right)^T\Sigma_k^{-1}\left(x_n-\mu_k\right)\right)\right]=0\end{aligned}\tag{14.1.1857}$$

Hence:

$$\begin{aligned}&\sum_{n=1}^{N}r_{nk}\left(-\frac{1}{2}\left(-2\Sigma_k^{-1}\left(x_n-\mu_k\right)\right)\right)=0\\&\sum_{n=1}^{N}r_{nk}\Sigma_k^{-1}\left(x_n-\mu_k\right)=0\\&\sum_{n=1}^{N}r_{nk}\Sigma_k^{-1}x_n-\sum_{n=1}^{N}r_{nk}\Sigma_k^{-1}\mu_k=0\end{aligned}\tag{14.1.1858}$$

So finally:

$$\mu_k=\frac{\sum_{n=1}^{N}r_{nk}\Sigma_k^{-1}x_n}{\sum_i r_{nk}\Sigma_{jk}^{-1}}=\frac{\sum_{n=1}^{N}r_{nk}x_n}{\sum_{n=1}^{N}r_{nk}}\tag{14.1.1859}$$

This looks perfectly right. Out of the M-Step, the mean value for a specific Gaussian distribution will be the expected value of the data points with respect to the variational distribution r_{nk} normalized by the sum of all the weights.

Finally, let's derive the analytical update expression for the covariance matrix Σ (we use in this development derivatives of matrices as seen in the section of Linear Algebra page

1976):

$$\begin{aligned}\frac{\partial L(\theta_t, r_{nk}, \lambda)}{\partial \Sigma_k} = 0 \quad \Leftrightarrow & \frac{\partial}{\partial \Sigma_k}\left[\left(\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk} \ln\left(\alpha_k \mathcal{N}\left(x_n \,|\, \mu_k, \Sigma_k\right)\right)\right) - \lambda\left(\sum_{k=1}^{K}\alpha_k - 1\right)\right] = \mathbb{O}\\ & \frac{\partial}{\partial \Sigma_k}\left[\sum_{n=1}^{N} r_{nk} \ln\left(\mathcal{N}\left(x_n \,|\, \mu_k, \Sigma_k\right)\right)\right] = \mathbb{O}\\ & \frac{\partial}{\partial \Sigma_k}\left[\sum_{n=1}^{N} r_{nk} \ln\left(\frac{1}{(2\pi)^{(N/2)}\left|\Sigma_k\right|^{1/2}}\exp\left(-\frac{1}{2}\left(x_n - \mu_k\right)^T \Sigma_k^{-1}\left(x_n - \mu_k\right)\right)\right]\right] = \mathbb{O}\\ & \frac{\partial}{\partial \Sigma_k}\left[\sum_{n=1}^{N} r_{nk}\left(\ln\left(\frac{1}{(2\pi)^{(N/2)}}\right) + \frac{1}{2}\ln\left(\frac{1}{\left|\Sigma_k\right|}\right) + \ln\left(\exp\left(-\frac{1}{2}\left(x_n - \mu_k\right)^T \Sigma_k^{-1}\left(x_n - \mu_k\right)\right)\right)\right)\right] = \mathbb{O}\\ & \frac{\partial}{\partial \Sigma_k}\left[\sum_{n=1}^{N} r_{nk}\left(-\frac{1}{2}\ln\left(\left|\Sigma_k\right|\right) - \frac{1}{2}\left(x_n - \mu_k\right)^T \Sigma_k^{-1}\left(x_n - \mu_k\right)\right)\right] = \mathbb{O}\\ & \frac{\partial}{\partial \Sigma_k}\left[\sum_{n=1}^{N} r_{nk}\left(\ln\left(\left|\Sigma_k\right|\right) + \left(x_n - \mu_k\right)^T \Sigma_k^{-1}\left(x_n - \mu_k\right)\right)\right] = \mathbb{O}\end{aligned} \tag{14.1.1860}$$

Therefore:

$$\frac{\partial}{\partial \Sigma_k^{-1}}\left[\sum_{n=1}^{N} r_{nk}\left(\ln\left(\left|\Sigma_k\right|\right) + \left(x_n - \mu_k\right)^T \Sigma_k^{-1}\left(x_n - \mu_k\right)\right)\right] = \mathbb{O} \tag{14.1.1861}$$

Hence:

$$\sum_{n=1}^{N} r_{nk}\left(-\Sigma_k^T + \left(x_n - \mu_k\right)\left(x_n - \mu_k\right)^T\right) = \mathbb{O} \tag{14.1.1862}$$

Finally:

$$\Sigma_k^T = \frac{\sum_{n=1}^{N} r_{nk}\left(x_n - \mu_k\right)\left(x_n - \mu_k\right)^T}{\sum_{n=1}^{N} r_{nk}} \tag{14.1.1863}$$

And this looks pretty dam right! The covariance matrix after the M-Step is a re-weighted version of the previous covariance matrix.

Most algorithms searches over a range of different types of Gaussians and n (named sometimes the "number of components") to find the best model by the Bayesian information criterion (BIC). The model is often referred to a three letter code nomenclature from the variance-covariance matrix describing the "shape" (relative magnitude of eigenvalues categorized into 2 different possibilities), "volume" (absolute magnitude of eigenvalues categorized into 3 different possibilities) and "orientation" (orientation of the eigenvectors categorized into 4 different possibilities) of the clusters. So there is a total of $2 \cdot 3 \cdot 4 = 24$ possibilities, here is a non-exhaustive list:

- `EII`: equal volume, round shape (spherical covariance)
- `VII`: varying volume, round shape (spherical covariance)
- `EEI`: equal volume, equal shape, axis parallel orientation (diagonal covariance)
- `VEI`: varying volume, equal shape, axis parallel orientation (diagonal covariance)

- `EVI`: equal volume, varying shape, axis parallel orientation (diagonal covariance)
- `VVI`: varying volume, varying shape, equal orientation (diagonal covariance)
- `EEE`: equal volume, equal shape, equal orientation (ellipsoidal covariance)
- `EEV`: equal volume, equal shape, varying orientation (ellipsoidal covariance)
- `VEV`: varying volume, equal shape, varying orientation (ellipsoidal covariance)
- `VVV`: varying volume, varying shape, varying orientation (ellipsoidal covariance)

Some of the above combinations are illustrated below:

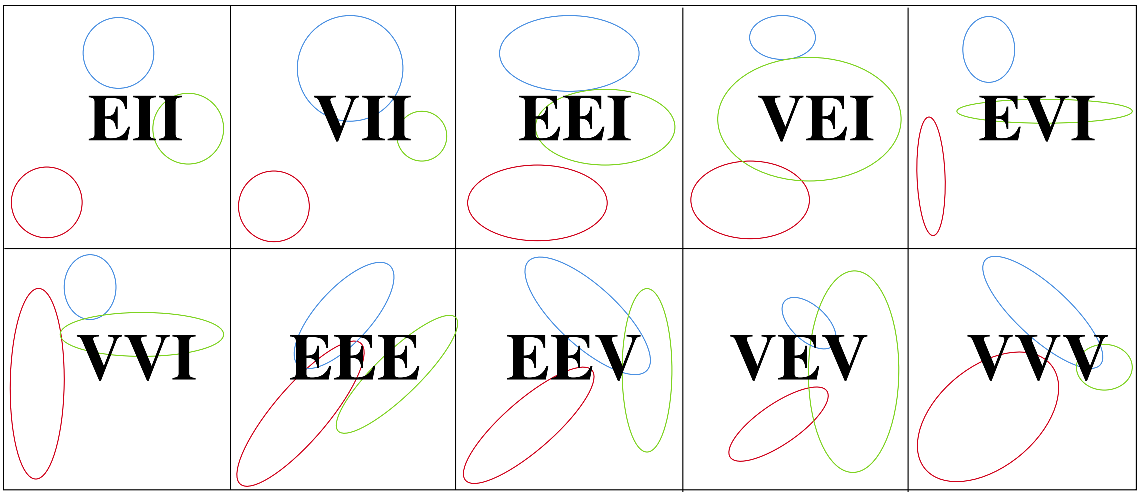


Figure 14.161: **Variance-Covariance matrix typical structures**

14.1.20.5.7 Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

"Density-based spatial clustering of applications with noise" (DBSCAN) is a data clustering algorithm proposed by Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu in 11996 (holocene calendar). It is a density-based clustering non-parametric (unsupervised) algorithm: given a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbors), marking as outliers points that lie alone in low-density regions (whose nearest neighbors are too far away). DBSCAN is one of the most common, and most commonly cited, clustering algorithms iN the early 121st century (holocene calendar).

Consider a set of points in some space to be clustered. Let ε be a parameter specifying the radius of a neighborhood with respect to some point. For the purpose of DBSCAN clustering, the points are classified as:

- core points
- (directly-) reachable points
- outliers

as follows:

Definitions (#14.1.571):

D1. A point p is a "core point" if at least minPts[65] points are within distance ε of it (including p)

D2. A point q is "directly reachable" from p if point q is within distance ε from core point p. Points are only said to be directly reachable from core points.

An object is said to be in a "ε-neighbourhood" if the object is within a radius of ε from an object as:

$$N_\varepsilon(p):\{q\,|\,d(p,q)\leq\varepsilon\}$$

D3. A point q is "reachable" from p if there is a path $p_1,\ldots,p_n$ with $p_1=p$ and $p_n=q$, where each p_{i+1} is directly reachable from p_i. Note that this implies that the initial point and all points on the path must be core points, with the possible exception of q.

D4. All points not reachable from any other point are "outliers" or "noise".

Now if p is a core point, then it forms a cluster together with all points (core or non-core) that are reachable from it. Each cluster contains at least one core point; non-core points can be part of a cluster, but they form its "edge", since they cannot be used to reach more points.

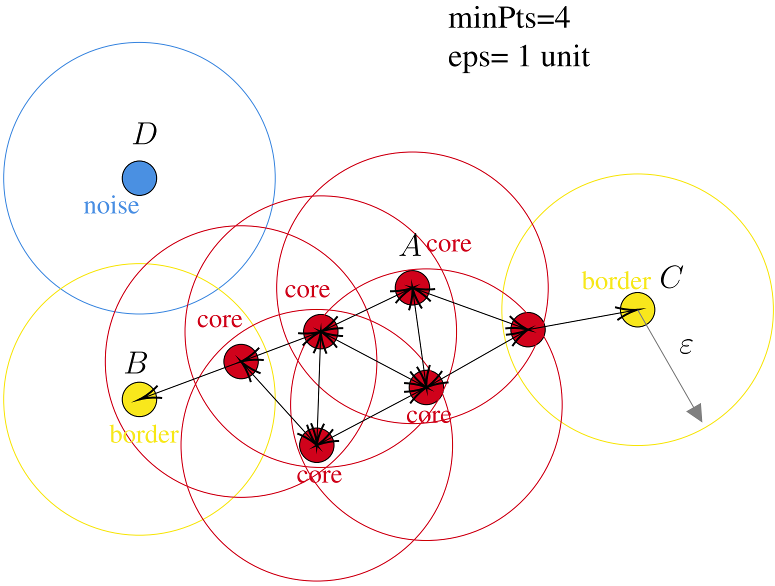
Remark

Reachability is not a symmetric relation: by definition, only core points can reach non-core points. The opposite is not true, so a non-core point may be reachable, but nothing can be reached from it.

In the figure below $\text{minPts}=4$ the point A and the other red points are "core points", because the area surrounding these points in an ε radius contain at least 4 points (including the point itself). Because they are all reachable from one another, they form a single cluster. Points B and C are not core points, but are reachable from A (via other core points) and thus belong to the cluster as well. Point N is a noise point that is neither a core point nor directly-reachable:

---

[65] While minPts intuitively is the minimum cluster size, in some cases DBSCAN can produce smaller clusters. A DBSCAN cluster consists of at least one core point. As other points may be border points to more than one cluster, there is no guarantee that at least minPts points are included in every cluster.



DBSCAN requires two parameters: ε (eps) and the minimum number of points required to form a dense region (minPts). It starts with an arbitrary starting point that has not been visited. This point's ε-neighborhood is retrieved, and if it contains sufficiently many points, a cluster is started. Otherwise, the point is labeled as noise. Note that this point might later be found in a sufficiently sized ε-environment of a different point and hence be made part of a cluster.

If the number of points is more than minPts, the current point and its nearby points from a cluster, and the starting point is marked as visited. Then, all the points in the cluster are not marked as visited that are processed in the same way recursively, to expand the cluster. Otherwise, the point is temporarily marked as a noise point. If the cluster is fully expanded, that is, all points in the cluster are marked as visited, and then the same algorithm is used to process the non-visited points. Until all objects are marked as a certain cluster or noise, the clustering process ends.

DBSCAN can be used with any distance function (as well as similarity functions or other predicates). The distance function (dist) can therefore be seen as an additional parameter.

Here is a general idea of the DBSCAN algorithm:

```
Input: DB (database), distFunc, $\varepsilon$, minPts
foreach $\text{point}\ (P) \in \text{database DB}$ do
    if $\text{point}\ (P)$ is not yet classified then
        $\text{collect all objects density} - \text{reachable from}\ (P)$ and assign them to a new cluster;
    else
        $\text{assign}\ (P)\ \text{to noise}$;
return cluster, label
```

Or with more details:

Algorithm 16: DBSCAN Algorithm

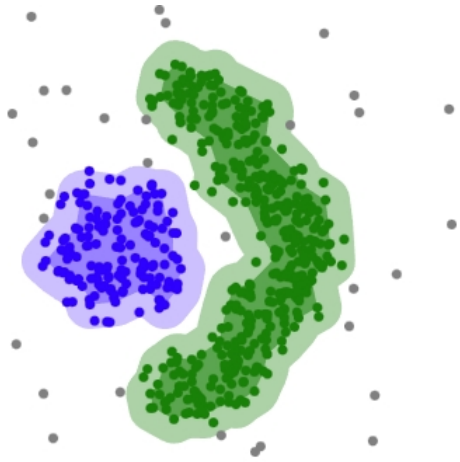
```
  Input: DB (database), distFunc, $\varepsilon$, minPts
Function DBSCAN(DB,distFunc,$\varepsilon$,minPts):
    $C := 0$ ;                          ▷ we initialize the number of clusters
    foreach point $p \in$ database DB do
        if label$(P) \neq$ undefined then
            Next ;                      ▷ we jump to the next iteration
        Neighbors $N :=$ RangeQuery(DB, distFunc, P, $\varepsilon$);
        if Card$(N) <$ minPts then
            label$(P) =$ noise;
            Next ;                      ▷ we jump to the next iteration
        $C := C + 1$;
        label$(P) := C$;
        SeedSet $S := N \setminus \{P\}$;   ▷ random point of $C$ excepted the core one
        foreach point $Q \in S$ do
            if label$(Q) =$ noise then
                label$(Q) := C$;
            if label$(Q) \neq$ undefined then
                Next ;                  ▷ we jump to the next iteration
            label$(Q) := C$;
            Neighbors $N :=$ RangeQuery(DB, distFunc, Q, $\varepsilon$);
            if Card$(N) \geq$ minPts then
                $S := S \cup N$
    return cluster, label
Function RangeQuery(DB,distFunc,Q,$\varepsilon$):
    Neighbors $N :=$ empty list;
    foreach point $P \in$ database DB do
        if distFun$(Q, P) \leq \varepsilon$ then
            $N := N \cup \{P\}$;
    return $N$
```

As we can see from the above pseudocode DBSCAN has a worst-case of $\mathcal{O}(n^2)$.

DBSCAN has the following advantages:

- DBSCAN does not require one to specify the number of clusters in the data a priori, as opposed to k-means.
- DBSCAN can find arbitrarily-shaped clusters. It can even find a cluster completely surrounded by (but not connected to) a different cluster. Due to the MinPts parameter, the so-called single-link effect (different clusters being connected by a thin line of points) is reduced.

  DBSCAN can find non-linearly separable clusters. The dataset below cannot be adequately clustered with k-means or Gaussian Mixture EM clustering:



- DBSCAN has a notion of noise, and is robust to outliers.
- DBSCAN requires just two parameters and is mostly insensitive to the ordering of the points in the database.

and the following disadvantages:

- DBSCAN is not entirely deterministic: border points that are reachable from more than one cluster can be part of either cluster, depending on the order the data are processed.
- The quality of DBSCAN depends on the distance measure used in the function $\text{regionQuery}(P, \varepsilon)$. The most common distance metric used is Euclidean distance.
- DBSCAN cannot cluster data sets well with large differences in densities, since the $\text{minPts} - \varepsilon$ combination cannot then be chosen appropriately for all clusters.
- If the data and scale are not well understood, choosing a meaningful distance threshold ε can be difficult.

Remark

Generalized DBSCAN (GDBSCAN) is a generalization by the same authors to arbitrary "neighborhood" and "dense" predicates. HDBSCAN is a hierarchical version of DBSCAN.

14.1.20.5.8 Mean shift clustering

"Mean shift" is a powerful and versatile non parametric iterative algorithm that can be used for lot of purposes like finding modes, clustering, tracking images etc. Mean Shift was introduced by Keinosuke Fukunaga and Larry D. Hostetler in 11975 (holocene calendar) and has been extended to be applicable in other fields like Computer Vision. The text that follows will provide a discussion of Mean Shift , prove its convergence and slightly discuss its important applications.

This text provides an intuitive idea of Mean shift and the later sections will expand the idea. Mean shift considers feature space as a empirical probability density function. If the input is a set of points then Mean shift considers them as sampled from the underlying probability density function. If dense regions (or clusters) are present in the feature space , then they correspond to the mode (or local maxima) of the probability density function. We can also identify clusters associated with the given mode using Mean Shift.

For each data point, Mean shift associates it with the nearby peak of the dataset's probability density function. For each data point, Mean shift defines a window around it and computes the mean of the data point. Then it shifts the center of the window to the mean and repeats the algorithm till it converges. After each iteration, we can consider that the window shifts to a more denser region of the dataset.

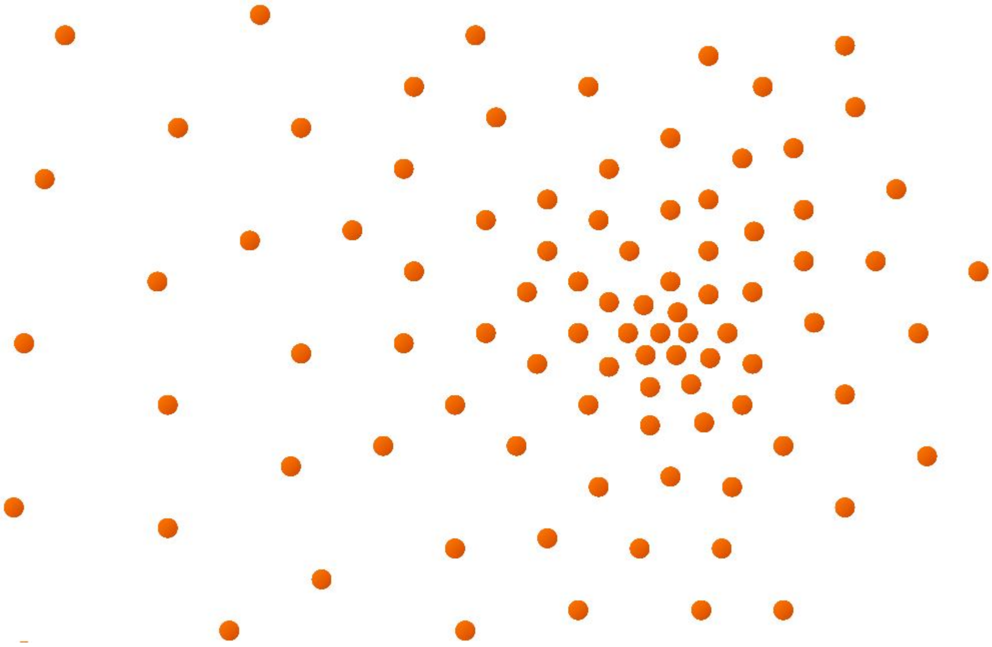
At the high level, we can specify Mean Shift as follows :

1. Choose a kernel K
2. Fix a window size h[66]
3. Randomly segment the domain with the kernel K (of size h)
4. Compute the mean of data (barycenter) weighted within the window h of each kernel
5. Shift the window h of each kernel K to the mean (barycenter) and repeat till convergence
6. Calculate the probability each point has to belong to one of the kernel K
7. Cluster each point to the kernel K have the highest probability

Before dealing with the mathematical aspect, let us consider the graphic and playful aspect! The purpose of the Mean Shift algorithm is to find the maximum density region(s) per iteration. So consider for example the following bi-varied case with some points:

---

[66] Some algorithms have a dynamic window



where we place randomly one or many circles (in d-dimensional case it will be d-dimensions sphere instead of simple circles) of a given radius:

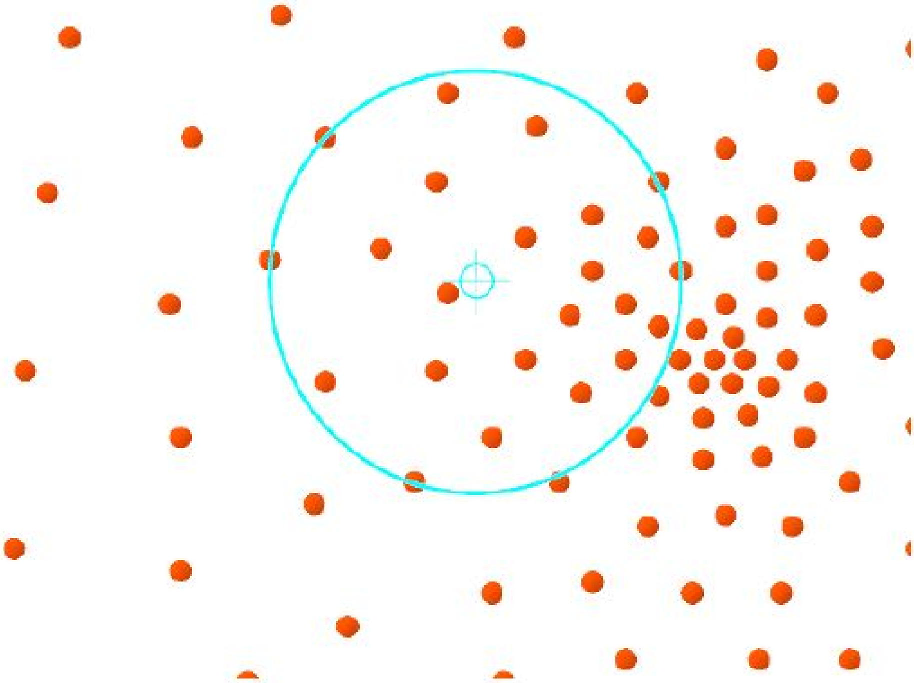
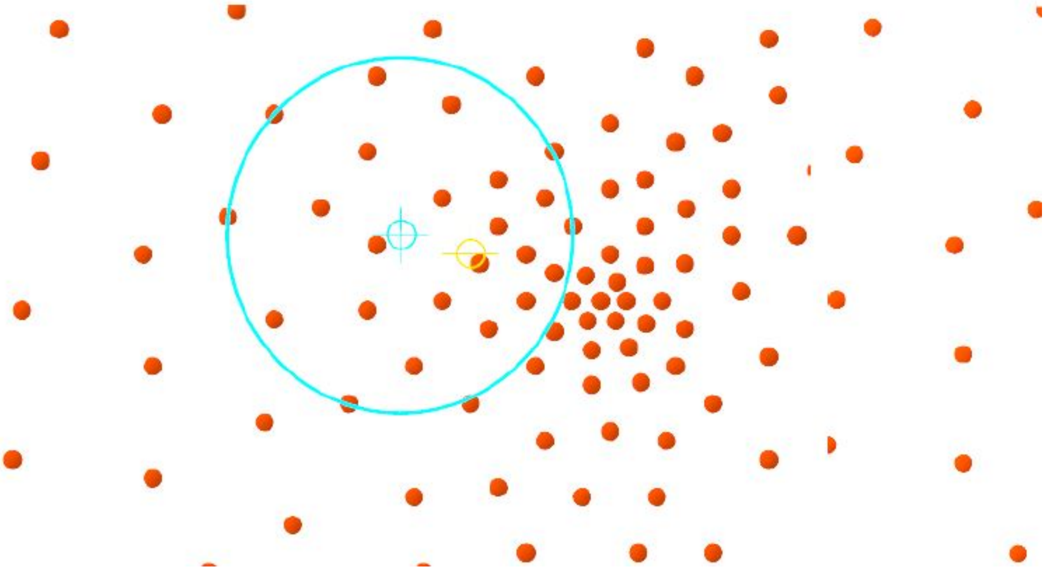
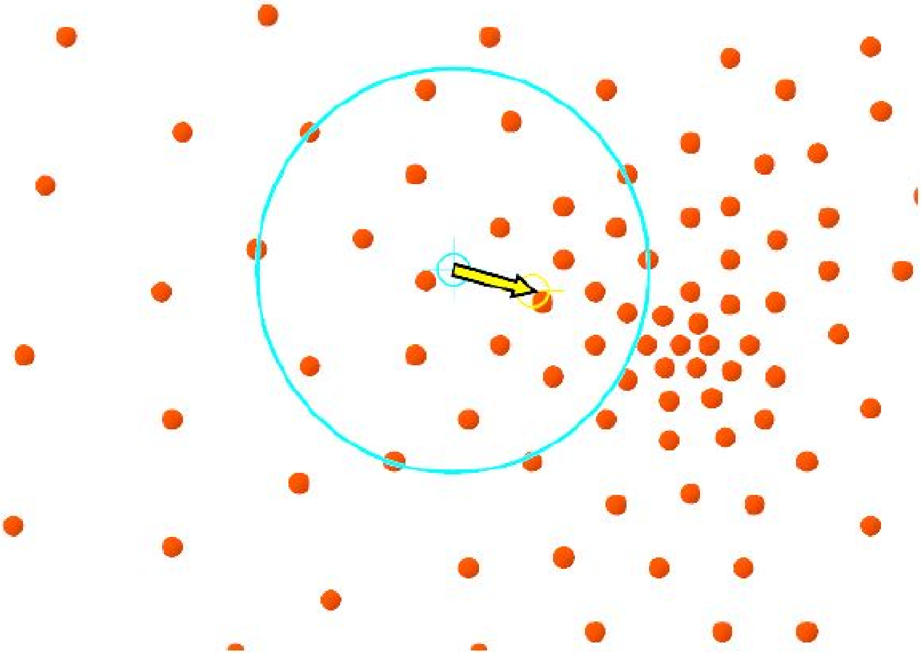


Figure 14.162: **Initial configuration of a bi-dimensional mean shift application**

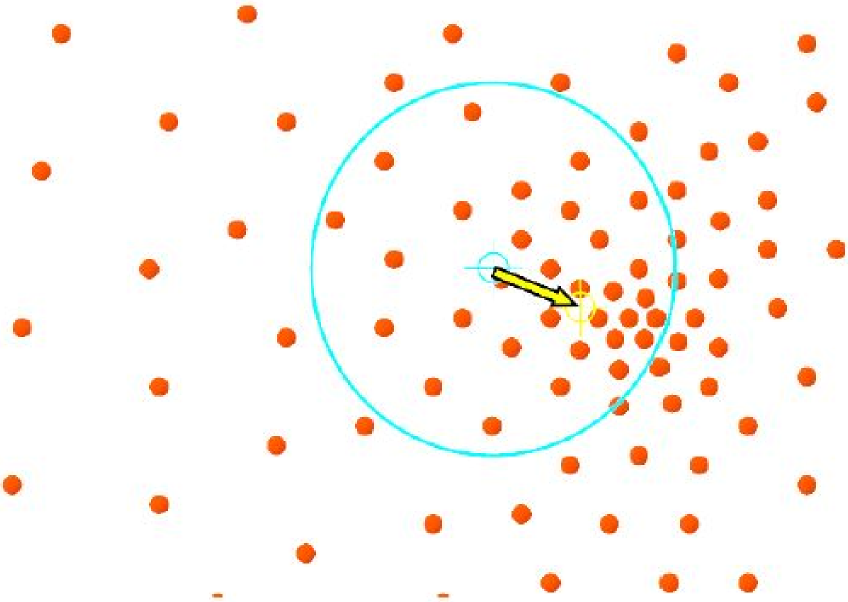
Then more or less empirically (there are different techniques!) we calculate the center of mass inside this circle:



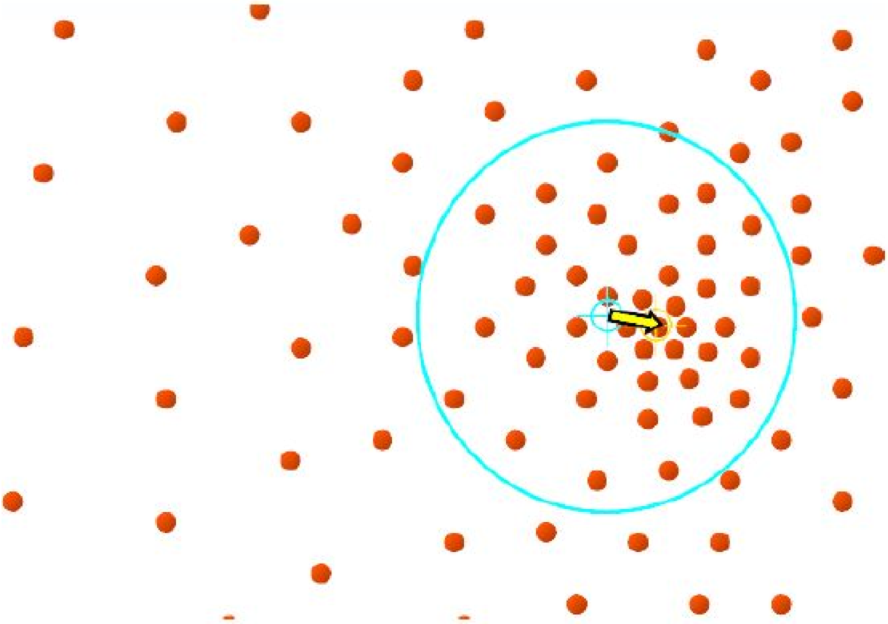
Then we move the circle on the center of gravity (barycenter) to get:



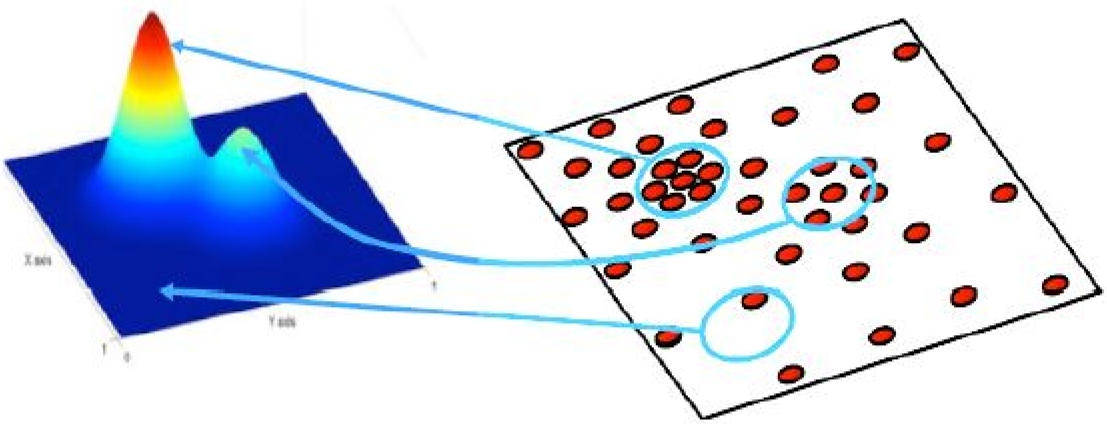
and we repeat the procedure:



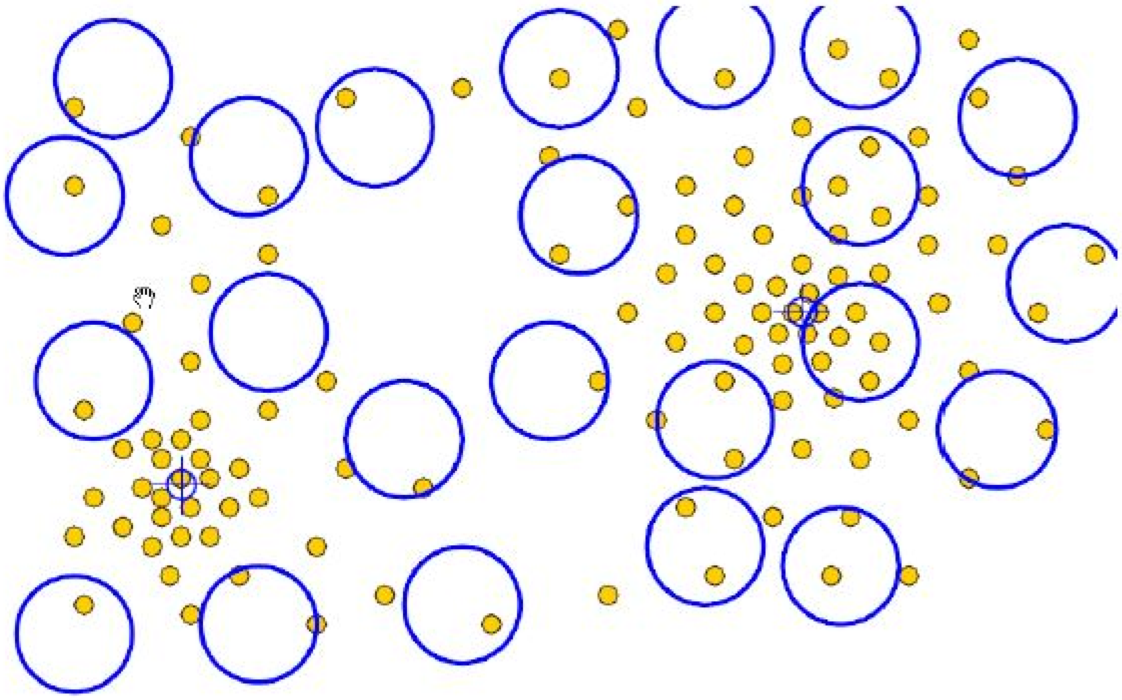
And so on:



That is to say that if we have for example a distribution of points such as below:



In reality we will randomly segment the domain and run the algorithm iteratively in parallel for each sphere (circle in the case of dimension 2):



Remark

The reader will be able to find a practical example in our `R` companion book!

As the reader can guess intuitively, this method converges securely (it can be proved algebraically but it's quite boring!), on the other hand it converges infinitely so we must set a convergence terminal for the algorithm to stop its iterations.

Let's move on to the computational aspect! We will introduce two different and very common approaches.

For the both approach, the first idea of the Mean Shift algorithm is to consider the workspace as a probability density (or multiple one!). Then each point of the workspace is identified by a coordinate vector $\vec{p}_i$ with a "mass" (corresponding mathematically to it's probability density function: the kernel!) given by the kernel:

$$K(\vec{p}_0-\vec{p}_i) \tag{14.1.1864}$$

centered on $\vec{p}_0$. We have already meet such Kernels[67] and we know what should be their properties. We won't come back here on these aspects! If the workspace contains a set of points, the latter are considered as a sample of an underlying probability density function and the dense regions (local maximum) as the modal value of the function of density of interest!

Obvious a very common Kernel is the multivariate Gaussian one given for recall by (where the lowercase k means that at each iteration k the variance-covariance matrix is recomputed):

$$K(\vec{p}_0-\vec{p}_i)=\sum_{i=1}^{N}\frac{1}{(2\pi)^{N/2}|\Sigma_k|^{1/2}}e^{-\frac{1}{2}(\vec{p}_0-\vec{p}_i)^T\Sigma_k^{-1}(\vec{p}_0-\vec{p}_i)} \tag{14.1.1865}$$

In practice some algorithms introduce a constant h such that the latter is written for a d-dimensional case:

$$K(\vec{p}_0-\vec{p}_i)=\frac{1}{h^d}\sum_{i=1}^{N}e^{-\frac{\|\vec{p}_0-\vec{p}_i\|^2}{h^2}}=\frac{1}{h^d}\sum_{i=1}^{N}e^{-\left\|\frac{\vec{p}_0-\vec{p}_i}{h}\right\|^2}=\frac{1}{h^d}\sum_{i=1}^{N}K\left(\left\|\frac{\vec{p}_0-\vec{p}_i}{h}\right\|^2\right) \tag{14.1.1866}$$

The first typical implementation of the Mean Shift clustering algorithm is based on the barycenter in Geometry (or center of gravity in physics). For this let us recall that we have proved in the section of Euclidean Geometry (page 2635), that the barycenter was given by:

$$\overrightarrow{\mathrm{O}G}=\frac{\displaystyle\sum_{i=1}^{N}\alpha_i\overrightarrow{\mathrm{O}A_i}}{\displaystyle\sum_{i=1}^{N}\alpha_i} \tag{14.1.1867}$$

This will be written in the context of Mean Shift with Gaussian kernel:

$$\boxed{\vec{p}_0^{[k+1]}=\frac{\displaystyle\sum_{i=1}^{N}K\left(\left\|\frac{\vec{p}_0^{[k]}-\vec{p}_i}{h}\right\|^2\right)\vec{p}_i}{\displaystyle\sum_{i=1}^{N}K\left(\left\|\frac{\vec{p}_0^{[k]}-\vec{p}_i}{h}\right\|^2\right)}} \tag{14.1.1868}$$

The difference:

$$m(\vec{p}_0^{[k+1]})=\vec{p}_0^{[k+1]}-\vec{p}_0^{[k]} \tag{14.1.1869}$$

is named the "mean shift vector" in the paper of Fukunaga and Hostetler. The mean-shift algorithm sets afterwards $\vec{p}_0^{[k]}\leftarrow m(\vec{p}_0)$, and repeats the estimation k time $m(\vec{p}_0)^{[k]}$ converges.

---

[67] Several types of kernel functions are commonly used: uniform, triangle, Epanechnikov, quartic (biweight), tricube, triweight, Gaussian, quadratic and cosine.

For the second typical possible implementation, consider again that $\{\vec{p}_i\}_{i=1\ldots N}$ is the position of the points and K is a given kernel function (that weights the points) for a given iteration, then for each of the N measurement points we have, we calculate the mean relatively to a reference point $\vec{p}_0$ (the center point of the kernel):

$$f(\vec{p}_0)=\frac{1}{N}\sum_{i=1}^{N}K\left(\left\|\frac{\vec{p}_0-\vec{p}_i}{h}\right\|^2\right) \tag{14.1.1870}$$

The first step in this second method with this underlying density $f(\vec{p}_0,h)$ is to find the mode of this density! The mode is located at the zero of the gradient $\vec{\nabla}f(\vec{p}_0,h)=\vec{0}$ and this second method is an elegant way to locate this zero!

Indeed, let us compute the gradient:

$$\begin{aligned}\vec{\nabla}f(\vec{p}_0,h)=\vec{\nabla}\left(\frac{1}{Nh^d}\sum_{i=1}^{N}K\left(\left\|\frac{\vec{p}_0-\vec{p}_i}{h}\right\|^2\right)\right)&=\frac{1}{Nh^d}\sum_{i=1}^{N}K'\left(\left\|\frac{\vec{p}_0-\vec{p}_i}{h}\right\|^2\right)\frac{2}{h^2}(\vec{p}_0-\vec{p}_i)\\&=\frac{2}{Nh^{2+d}}\sum_{i=1}^{N}K'\left(\left\|\frac{\vec{p}_0-\vec{p}_i}{h}\right\|^2\right)(\vec{p}_0-\vec{p}_i)=\vec{0}\end{aligned} \tag{14.1.1871}$$

Now a traditional idea is to replace $\vec{p}_0$ in the last parenthesis thanks to gradient itself. Indeed, that latter leads us to:

$$\frac{2}{Nh^{2+d}}\sum_{i=1}^{N}K'\left(\left\|\frac{\vec{p}_0-\vec{p}_i}{h}\right\|^2\right)\vec{p}_0-\frac{2}{Nh^{2+d}}\sum_{n=1}^{N}K'\left(\left\|\frac{\vec{p}_0-\vec{p}_i}{h}\right\|^2\right)\vec{p}_i=\vec{0} \tag{14.1.1872}$$

Hence after rearranging:

$$\vec{p}_0=\frac{\displaystyle\sum_{i=1}^{N}K'\left(\left\|\frac{\vec{p}_0-\vec{p}_i}{h}\right\|^2\right)\vec{p}_i}{\displaystyle\sum_{n'=1}^{N}K'\left(\left\|\frac{\vec{p}_0-\vec{p}_i}{h}\right\|^2\right)} \tag{14.1.1873}$$

So we get a fixed-point iteration, that brings us to write:

$$\boxed{\vec{p}_0^{[k+1]}=\frac{\displaystyle\sum_{i=1}^{N}K'\left(\left\|\frac{\vec{p}_0^{[k]}-\vec{p}_i}{h}\right\|^2\right)\vec{p}_i}{\displaystyle\sum_{n'=1}^{N}K'\left(\left\|\frac{\vec{p}_0^{[k]}-\vec{p}_i}{h}\right\|^2\right)}} \tag{14.1.1874}$$

Remark

As in the case of Gaussian Kernel we have $K'=K$ we then fall back on the result of the first method:

$$\vec{p}_0^{[k+1]}=\frac{\sum_{i=1}^{N}K\left(\left\|\frac{\vec{p}_0^{[k]}-\vec{p}_i}{h}\right\|^2\right)\vec{p}_i}{\sum_{n'=1}^{N}K\left(\left\|\frac{\vec{p}_0^{[k]}-\vec{p}_i}{h}\right\|^2\right)} \tag{14.1.1875}$$

What interest us is again the "mean shift vector" defined this time by:

$$m(\vec{p}_0^{[k+1]})=\vec{p}_0^{[k+1]}-\vec{p}_0^{[k]}=\frac{\sum_{i=1}^{N}\vec{p}_iK'\left(\left\|\frac{\vec{p}_0^{[k]}-\vec{p}_i}{h}\right\|^2\right)}{\sum_{n'=1}^{N}K'\left(\left\|\frac{\vec{p}_0^{[k]}-\vec{p}_i}{h}\right\|^2\right)}-\vec{p}_0^{[k]} \tag{14.1.1876}$$

Whatever the method, the choice of bandwidth parameter h or Kernel are critical. A large h might result in incorrect clustering and might merge distinct clusters. A very small h might result in too many clusters.

Mean-shift looks very similar to k-means, they both move the point closer to the cluster centroids. One may wonder: How is this different from k-means? The key difference is that Mean shift does not require the user to specify the number of clusters. In some cases, it is not straightforward to guess the right number of clusters to use. In k-means, the output may end up having too few clusters or too many clusters to be useful. At the cost of larger time complexity, Mean shift determines the number of clusters suitable to the dataset provided.

Another commonly cited difference is that k-means can only learn circle or ellipsoidal clusters. However, this is not true. The reason that Mean shift can learn arbitrary shapes is because the features are mapped to another higher dimensional feature space through the kernel. The arbitrary shapes are due to the algorithm finding circle or ellipsoidal clusters in higher dimensional feature space. When the features are mapped back to 1D/2D/3D, the resulting clusters look like strange shapes. This is also the trick as used in Support Vector Machines.

A traditional k-means does not use kernels, but Kernel k-means is available. Kernel k-means is useful if:

1. The number of clusters is known or can be reasonably estimated
2. Dataset needs learning non-ellipsoidal cluster shapes

So, you can enjoy the better runtime complexity of k-means and learn arbitrary clusters if you can determine the number of clusters to use.

14.1.20.5.9 Hierarchical Ascendant Clustering (HAC)

In data mining and statistics, "hierarchical clustering", also named "hierarchical cluster analysis" (HCA) or "hierarchical agglomerative clustering" is an unsupervised method of cluster analysis which seeks to build a hierarchy of clusters in a predefined order. Let us remind that strategies for hierarchical clustering generally fall into two types:

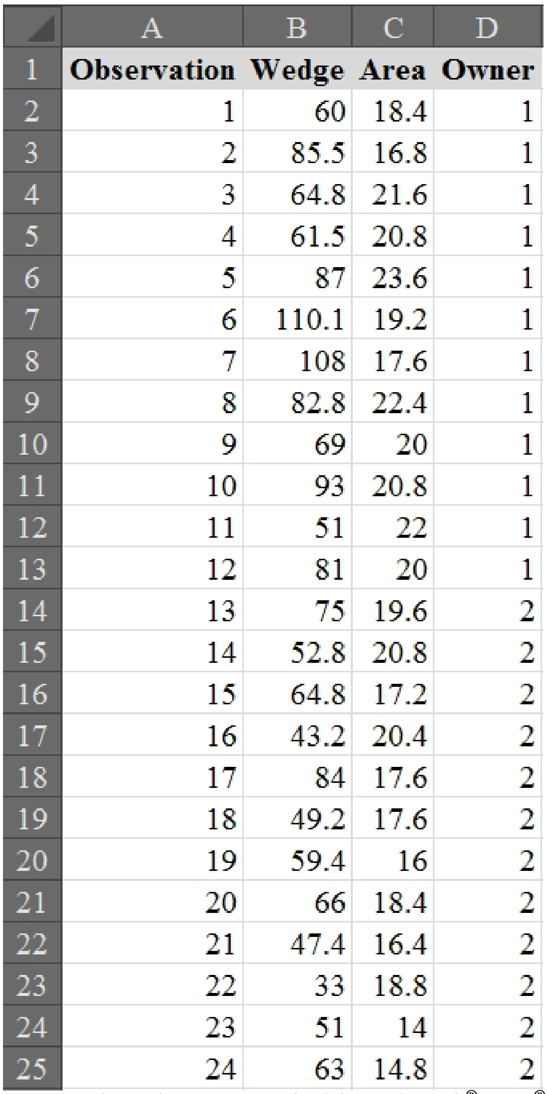
- **Agglomerative/Ascendant:** This is a "bottom-up" approach! each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. We speak then of "hierarchical ascendant clustering" (HAC), also known as "Hierarchical Agglomerative Clustering" (HAC) or "AGNES" (acronym for "AGglomerative NESting")
- **Divisive/Descendant:** This is a "top-down" approach! all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy. We speak then of "hierarchical descendant clustering" (HDC), also known as "DIANA" (acronym for "DivisiveANAly").

In order to decide which clusters should be combined (for agglomerative), or where a cluster should be split (for divisive), a measure of dissimilarity between sets of observations is required. In most methods of hierarchical clustering, this is achieved by use of an appropriate metric (a measure of distance between pairs of observations), and a linkage criterion which specifies the dissimilarity of sets as a function of the pairwise distances of observations in the sets.

Remark

"Non hierarchical clustering" involves formation of new clusters by merging or splitting the clusters. It does not follow a tree like structure like hierarchical clustering. This technique groups the data in order to maximize or minimize some evaluation criteria. k-means clustering is an effective way of non hierarchical clustering (see page 6854). In this method the partitions are made such that non-overlapping groups having no hierarchical relationships between themselves.

To introduce this clustering technique let us consider the following companion example based on the following list of data in Microsoft® Excel®:



	A	B	C	D
1	Observation	Wedge	Area	Owner
2	1	60	18.4	1
3	2	85.5	16.8	1
4	3	64.8	21.6	1
5	4	61.5	20.8	1
6	5	87	23.6	1
7	6	110.1	19.2	1
8	7	108	17.6	1
9	8	82.8	22.4	1
10	9	69	20	1
11	10	93	20.8	1
12	11	51	22	1
13	12	81	20	1
14	13	75	19.6	2
15	14	52.8	20.8	2
16	15	64.8	17.2	2
17	16	43.2	20.4	2
18	17	84	17.6	2
19	18	49.2	17.6	2
20	19	59.4	16	2
21	20	66	18.4	2
22	21	47.4	16.4	2
23	22	33	18.8	2
24	23	51	14	2
25	24	63	14.8	2

Figure 14.163: List of data for our study of HAC in Microsoft® Excel® 14.0.6123

We wish to have a hierarchical organization of likeness of individuals based on their income (*Wage*) and their living space (*Area*).

One possible technique is to define for this measurement a distance of similarity. For example, the Euclidean distance:

$$d_{ij}(x,y)=\sqrt{(x_i-x_j)^2+(y_i-y_j)^2} \tag{14.1.1877}$$

is a special choice that will associate two individuals whose distance is minimal. We speak then in the area of clustering "single linkage".

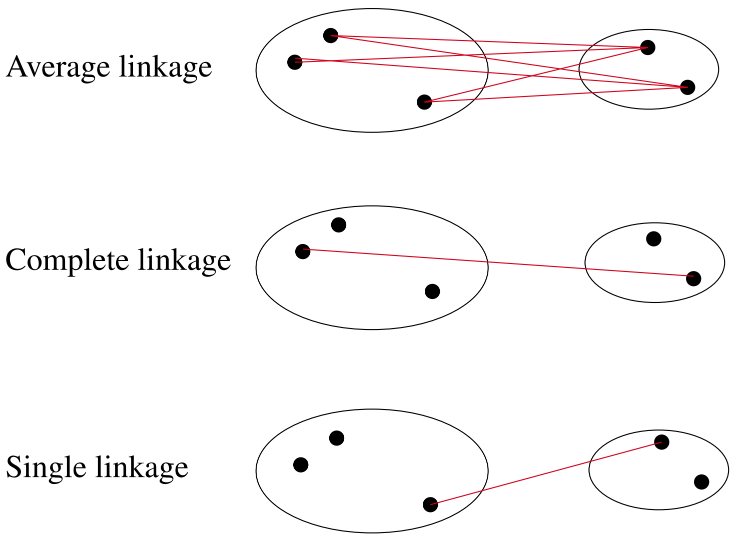


Figure 14.164: **Type of links in Hierarchical Clustering**

Formally:

- Maximum or complete linkage clustering:

$$\max\{d(a,b):a\in A,b\in B\}\tag{14.1.1878}$$

- Minimum or single linkage clustering:

$$\min\{d(a,b):a\in A,b\in B\}\tag{14.1.1879}$$

- Average linkage:

$$\frac{1}{\text{Card}(A\cdot B)}\sum_{a\in A}\sum_{b\in B}d(a,b)\tag{14.1.1880}$$

✍**Example:**

How to measure the distance between two individuals considering:

Id	Firstname	Age	Children	Shoe size	Height
1	Alain	45	3	45	182
2	Martine	28	1	36	165
3	Pierre	22	0	43	172

Who is Martine closest to? We will use the Euclidean distance!

- Distance between Alain and Martine:

$$d_{1*2}=\sqrt{(45-28)^2+(3-1)^2+(45-36)^2+(182-165)^2}$$

- Distance between Martine and Pierre:

$$d_{2*3}=\sqrt{(28-22)^2+(1-0)^2+(36-43)^2+(165-172)^2}$$

- Distance between Alain and Pierre :

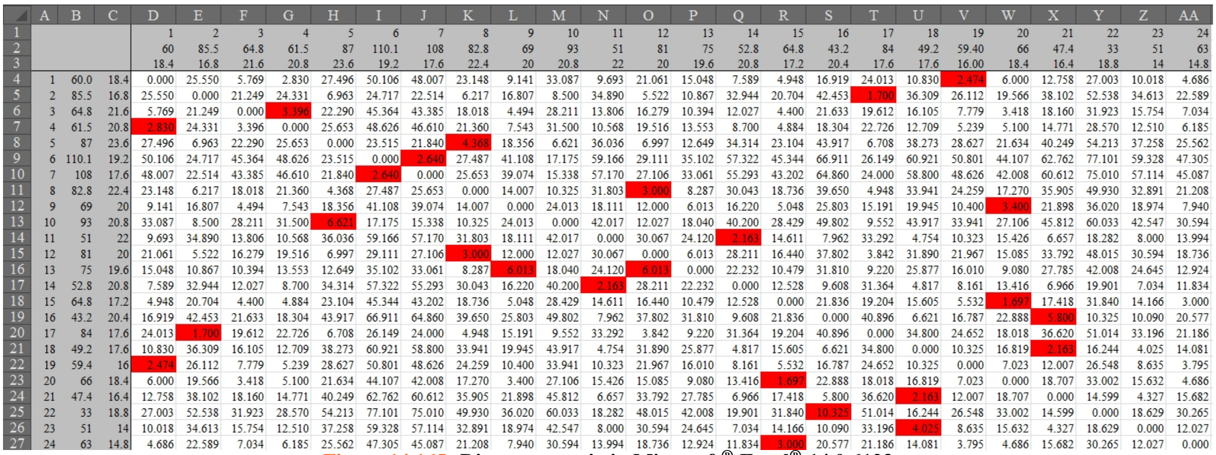
$$d_{1*3}=\sqrt{(45-22)^2+(3-0)^2+(45-43)^2+(182-172)^2}$$

Table of distances:

	Alain	Martine	Pierre
Alain	0		
Martine	25.74	0	
Pierre	25.33	11.61	0

We notice that the two variables *Size* and *Children* have little weight in the distance calculation. To remedy this, we must center and reduce the data normally before computing distances!

So we can easily, using a spreadsheet software like Microsoft® Excel® 14.0.6123, create a "distance matrix" or "proximity matrix" which is a symmetric matrix with zero in diagonal and that relatively to the list above will give us:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1				1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
2				60	85.5	64.8	61.5	87	110.1	108	82.8	69	93	51	81	75	52.8	64.8	43.2	84	49.2	59.40	66	47.4	33	51	63
3				18.4	16.8	21.6	20.8	23.6	19.2	17.6	22.4	20	20.8	22	20	19.6	20.8	17.2	20.4	17.6	17.6	16.00	18.4	16.4	18.8	14	14.8
4	1	60.0	18.4	0.000	25.550	5.769	2.830	27.496	50.106	48.007	23.148	9.141	33.087	9.693	21.061	15.048	7.589	4.948	16.919	24.013	10.830	2.474	6.000	12.758	27.003	10.018	4.686
5	2	85.5	16.8	25.550	0.000	21.249	24.331	6.963	24.717	22.514	6.217	16.807	8.500	34.890	5.522	10.867	32.944	20.704	42.453	1.700	36.309	26.112	19.566	38.102	52.538	34.613	22.589
6	3	64.8	21.6	5.769	21.249	0.000	3.396	22.290	45.364	43.385	18.018	4.494	28.211	13.806	16.279	10.394	12.027	4.400	21.633	19.612	16.105	7.779	3.418	18.160	31.923	15.754	7.034
7	4	61.5	20.8	2.830	24.331	3.396	0.000	25.653	48.626	46.610	21.360	7.543	31.500	10.568	19.516	13.553	8.700	4.884	18.304	22.726	12.709	5.239	5.100	14.771	28.570	12.510	6.185
8	5	87	23.6	27.496	6.963	22.290	25.653	0.000	23.515	21.840	4.368	18.356	6.621	36.036	6.997	12.649	34.314	23.104	43.917	6.708	38.273	28.627	21.634	40.249	54.213	37.258	25.562
9	6	110.1	19.2	50.106	24.717	45.364	48.626	23.515	0.000	2.640	27.487	41.108	17.175	59.166	29.111	35.102	57.322	45.344	66.911	26.149	60.921	50.801	44.107	62.762	77.101	59.328	47.305
10	7	108	17.6	48.007	22.514	43.385	46.610	21.840	2.640	0.000	25.653	39.074	15.338	57.170	27.106	33.061	55.293	43.202	64.860	24.000	58.800	48.626	42.008	60.612	75.010	57.114	45.087
11	8	82.8	22.4	23.148	6.217	18.018	21.360	4.368	27.487	25.653	0.000	14.007	10.325	31.803	3.000	8.287	30.043	18.736	39.650	4.948	33.941	24.259	17.270	35.905	49.930	32.891	21.208
12	9	69	20	9.141	16.807	4.494	7.543	18.356	41.108	39.074	14.007	0.000	24.013	18.111	12.000	6.013	16.220	5.048	25.803	15.191	19.945	10.400	3.400	21.898	36.020	18.974	7.940
13	10	93	20.8	33.087	8.500	28.211	31.500	6.621	17.175	15.338	10.325	24.013	0.000	42.017	12.027	18.040	40.200	28.429	49.802	9.552	43.917	33.941	27.106	45.812	60.033	42.547	30.594
14	11	51	22	9.693	34.890	13.806	10.568	36.036	59.166	57.170	31.803	18.111	42.017	0.000	30.067	24.120	2.163	14.611	7.962	33.292	4.754	10.323	15.426	6.657	18.282	8.000	13.994
15	12	81	20	21.061	5.522	16.279	19.516	6.997	29.111	27.106	3.000	12.000	12.027	30.067	0.000	6.013	28.211	16.440	37.802	3.842	31.890	21.967	15.085	33.792	48.015	30.594	18.736
16	13	75	19.6	15.048	10.867	10.394	13.553	12.649	35.102	33.061	8.287	6.013	18.040	24.120	6.013	0.000	22.232	10.479	31.810	9.220	25.877	16.010	9.080	27.785	42.008	24.645	12.924
17	14	52.8	20.8	7.589	32.944	12.027	8.700	34.314	57.322	55.293	30.043	16.220	40.200	2.163	28.211	22.232	0.000	12.528	9.608	31.364	4.817	8.161	13.416	6.966	19.901	7.034	11.834
18	15	64.8	17.2	4.948	20.704	4.400	4.884	23.104	45.344	43.202	18.736	5.048	28.429	14.611	16.440	10.479	12.528	0.000	21.836	19.204	15.605	5.532	1.697	17.418	31.840	14.166	3.000
19	16	43.2	20.4	16.919	42.453	21.633	18.304	43.917	66.911	64.860	39.650	25.803	49.802	7.962	37.802	31.810	9.608	21.836	0.000	40.896	6.621	16.787	22.888	5.800	10.325	10.090	20.577
20	17	84	17.6	24.013	1.700	19.612	22.726	6.708	26.149	24.000	4.948	15.191	9.552	33.292	3.842	9.220	31.364	19.204	40.896	0.000	34.800	24.652	18.018	36.620	51.014	33.196	21.186
21	18	49.2	17.6	10.830	36.309	16.105	12.709	38.273	60.921	58.800	33.941	19.945	43.917	4.754	31.890	25.877	4.817	15.605	6.621	34.800	0.000	10.325	16.819	2.163	16.244	4.025	14.081
22	19	59.4	16	2.474	26.112	7.779	5.239	28.627	50.801	48.626	24.259	10.400	33.941	10.323	21.967	16.010	8.161	5.532	16.787	24.652	10.325	0.000	7.023	12.007	26.548	8.635	3.795
23	20	66	18.4	6.000	19.566	3.418	5.100	21.634	44.107	42.008	17.270	3.400	27.106	15.426	15.085	9.080	13.416	1.697	22.888	18.018	16.819	7.023	0.000	18.707	33.002	15.632	4.686
24	21	47.4	16.4	12.758	38.102	18.160	14.771	40.249	62.762	60.612	35.905	21.898	45.812	6.657	33.792	27.785	6.966	17.418	5.800	36.620	2.163	12.007	18.707	0.000	14.599	4.327	15.682
25	22	33	18.8	27.003	52.538	31.923	28.570	54.213	77.101	75.010	49.930	36.020	60.033	18.282	48.015	42.008	19.901	31.840	10.325	51.014	16.244	26.548	33.002	14.599	0.000	18.629	30.265
26	23	51	14	10.018	34.613	15.754	12.510	37.258	59.328	57.114	32.891	18.974	42.547	8.000	30.594	24.645	7.034	14.166	10.090	33.196	4.025	8.635	15.632	4.327	18.629	0.000	12.027
27	24	63	14.8	4.686	22.589	7.034	6.185	25.562	47.305	45.087	21.208	7.940	30.594	13.994	18.736	12.924	11.834	3.000	20.577	21.186	14.081	3.795	4.686	15.682	30.265	12.027	0.000

Figure 14.165: **Distance matrix in Microsoft® Excel® 14.0.6123**

where we put in the cell `D4` the following Euclidean distance formula:

```
=SQRT((B4-D2)^2+(C4-D3)^2)
```

we then drag this formula for the rest of the matrix to the cell `AA27`.

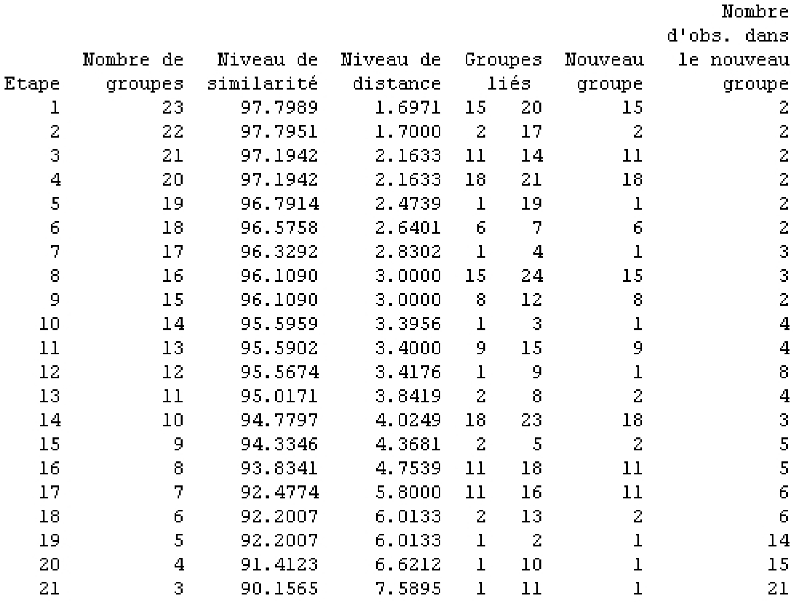
Then we use the bottom-up method (ascendant agglomeration) where we combine the groups until there is only one group remaining (containing all data) and this is a very boring work to describe and to do manually in spreadsheet software with a table of the size given above. We will give a step by step small example later below (see page 6904).

Remark

Let us remind that in their implementation, hierarchical clustering algorithms can adopt one of two following methods:

- The "top-down" ("descending") or "dividing method" consists in starting with a large cluster consisting of the whole data set and then in each step, the successive splitting of the generated clusters is carried out until each cluster consists of only one individual.
- The "bottom-up" ("ascending") or "agglomerative method" reverses the previous logic, it starts with clusters each composed of an individual and then at each step, the clusters are associated in pairs to form other clusters which are in turn joined two to two until a cluster is obtained.

This work will give us in a tabular form following the detailed steps given in our Minitab 15.1.1.0 companion book (values are not rounded to hundredths unlike the small matrix given in the figure above):



```
                                                                   Nombre
                                                                d'obs. dans
        Nombre de   Niveau de   Niveau de   Groupes   Nouveau   le nouveau
Etape     groupes  similarité    distance     liés     groupe       groupe
    1          23     97.7989      1.6971   15   20        15            2
    2          22     97.7951      1.7000    2   17         2            2
    3          21     97.1942      2.1633   11   14        11            2
    4          20     97.1942      2.1633   18   21        18            2
    5          19     96.7914      2.4739    1   19         1            2
    6          18     96.5758      2.6401    6    7         6            2
    7          17     96.3292      2.8302    1    4         1            3
    8          16     96.1090      3.0000   15   24        15            3
    9          15     96.1090      3.0000    8   12         8            2
   10          14     95.5959      3.3956    1    3         1            4
   11          13     95.5902      3.4000    9   15         9            4
   12          12     95.5674      3.4176    1    9         1            8
   13          11     95.0171      3.8419    2    8         2            4
   14          10     94.7797      4.0249   18   23        18            3
   15           9     94.3346      4.3681    2    5         2            5
   16           8     93.8341      4.7539   11   18        11            5
   17           7     92.4774      5.8000   11   16        11            6
   18           6     92.2007      6.0133    2   13         2            6
   19           5     92.2007      6.0133    1    2         1           14
   20           4     91.4123      6.6212    1   10         1           15
   21           3     90.1565      7.5895    1   11         1           21
```

Figure 14.166: HAC distance summary list in Minitab 15.1.1.0

where the level of similarity of the linked group i,j is defined empirically by:

$$S_{i,j}=1-\frac{d_{i,j}(x,y)}{\max(d_{i,j}(x,y))} \qquad (14.1.1881)$$

Thus, for the first row we have for example:

$$S_{15,20}=1-\frac{d_{15,20}(x,y)}{\max(d_{15,20}(x,y))}=1-\frac{1.6971}{77.1010}=97.7989\% \qquad (14.1.1882)$$

The preceding list is more pleasant to analyse if, as it is customary, we represent it is a "dendrogram" (from Greek dendro "tree" and gramma "drawing"), i.e. a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering, as given by Minitab 15.1.1.0 below:

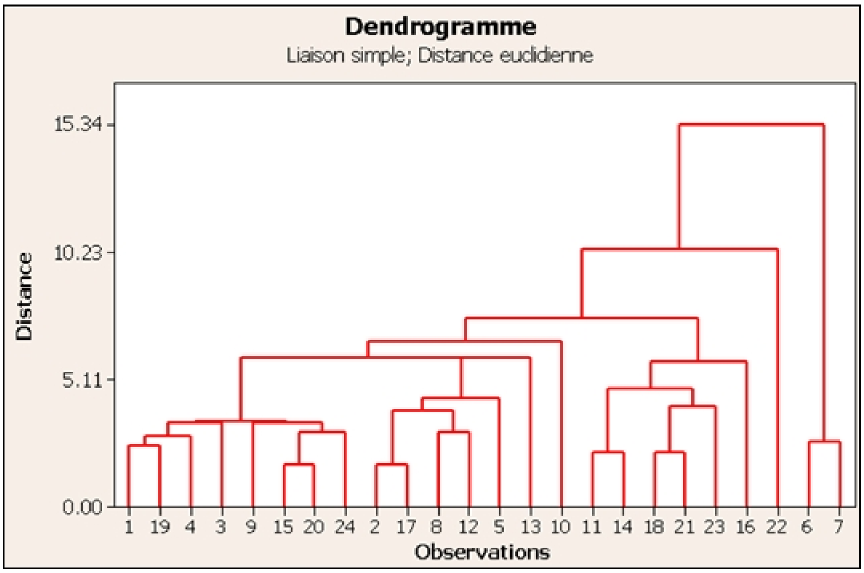


Figure 14.167: **Dendrograms plot in Minitab 15.1.1.0**

If you follow the detailed steps given in our MATLAB® 2013a companion book you will get:

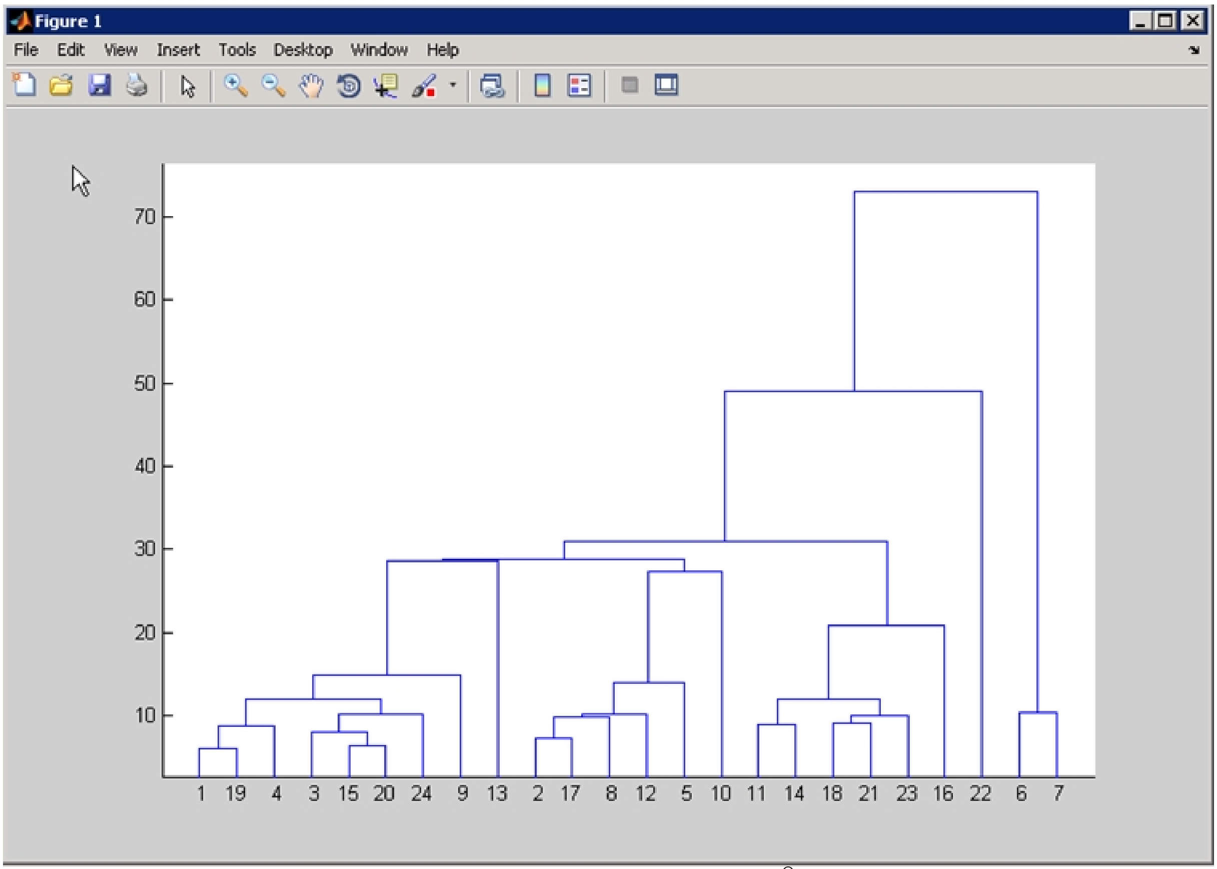


Figure 1
File Edit View Insert Tools Desktop Window Help
70
60
50
40
30
20
10
1 19 4 3 15 20 24 9 13 2 17 8 12 5 10 11 14 18 21 23 16 22 6 7

Figure 14.168: **Dendrograms plot in MATLAB® 2013a**

If you follow the detailed steps given in our `R` companion book you will get:

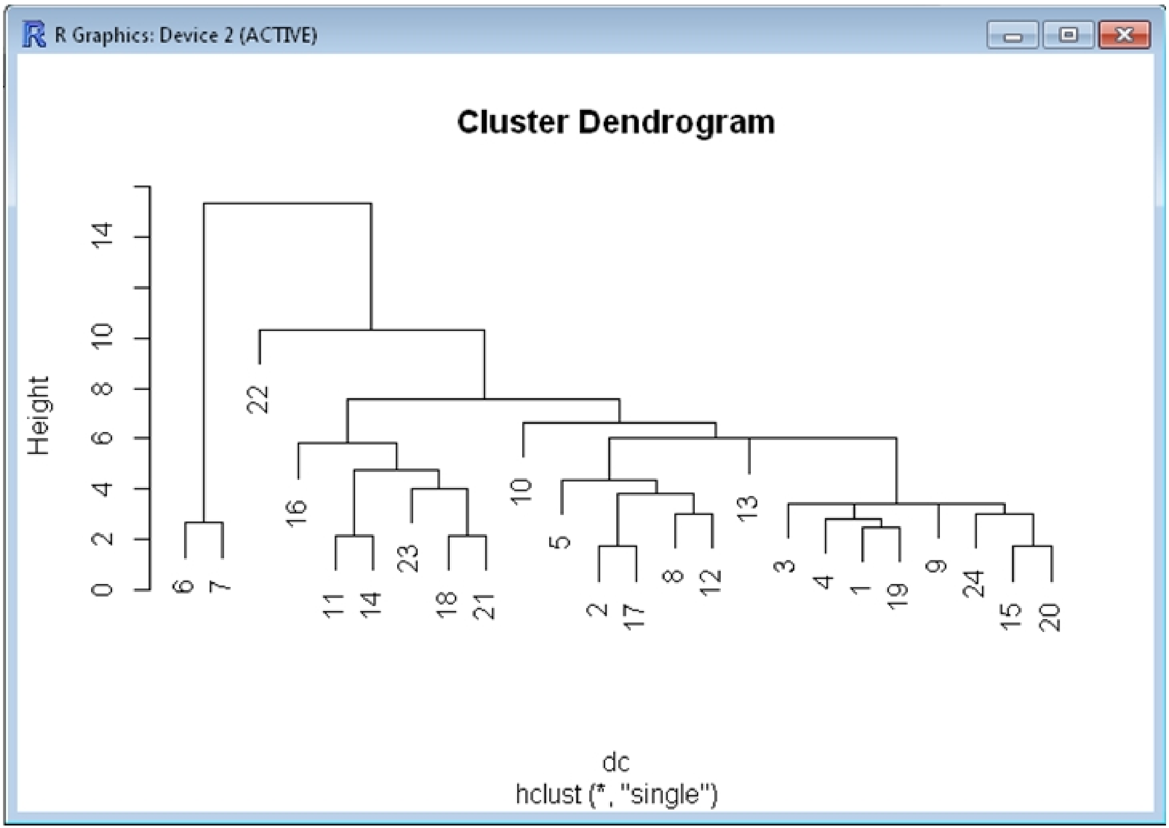
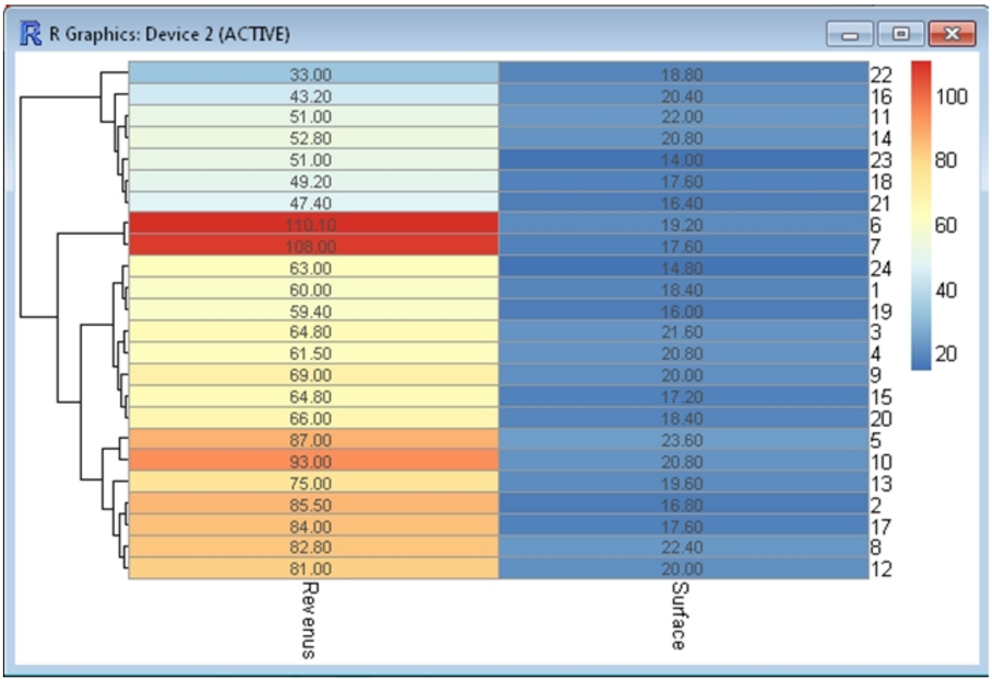

R Graphics: Device 2 (ACTIVE)
Cluster Dendrogram
Height
0
2
4
6
8
10
14
6
7
22
16
11
14
23
18
21
10
5
2
17
8
12
13
3
4
1
19
9
24
15
20
dc
hclust (*, "single")

Figure 14.169: Dendrograms plot in R 3.0.2

and still with R (see the corresponding companion book) for the same data:


R Graphics: Device 2 (ACTIVE)
100
80
60
40
20

	Revenus	Surface
22	33.00	18.80
16	43.20	20.40
11	51.00	22.00
14	52.80	20.80
23	51.00	14.00
18	49.20	17.60
21	47.40	16.40
6	110.10	19.20
7	108.00	17.60
24	63.00	14.80
1	60.00	18.40
19	59.40	16.00
3	64.80	21.60
4	61.50	20.80
9	69.00	20.00
15	64.80	17.20
20	66.00	18.40
5	87.00	23.60
10	93.00	20.80
13	75.00	19.60
2	85.50	16.80
17	84.00	17.60
8	82.80	22.40
12	81.00	20.00

Figure 14.170: Dendrograms heatmap plot in R 3.0.2

Dendrograms and HAC are also used in biostatistics as shown below (still with R but not described in the companion book yet):

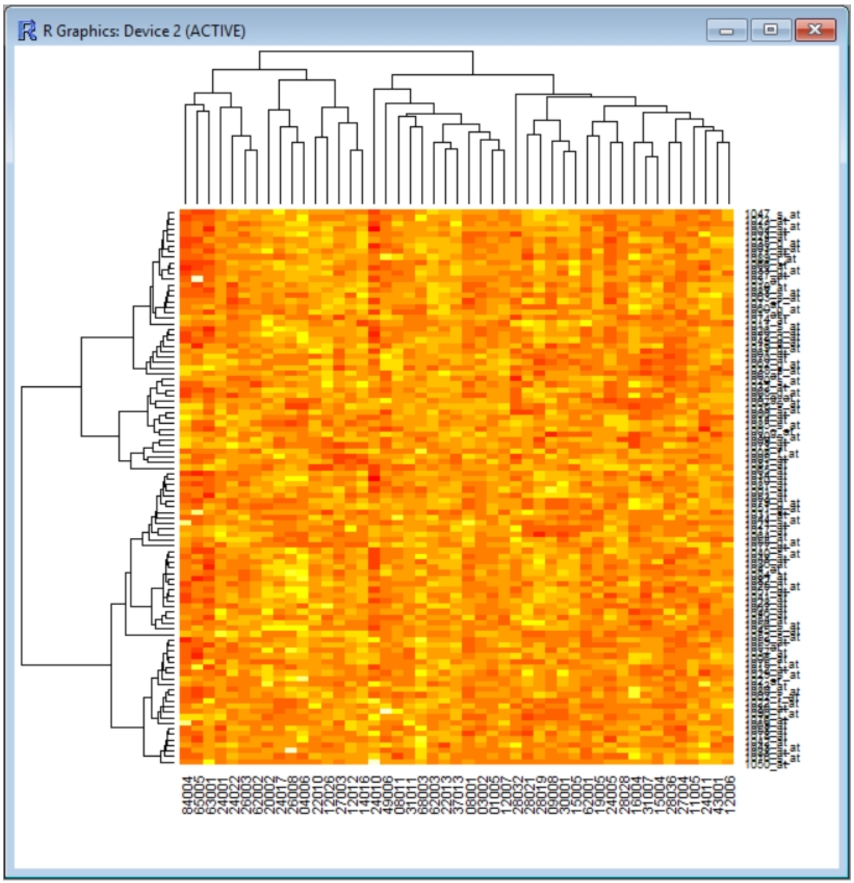


Figure 14.171: Dendrograms heatmap microarray plot in `R` 3.0.2

Or in financial engineering to group similar times series (still with `R` and detailed steps given in the companion book):

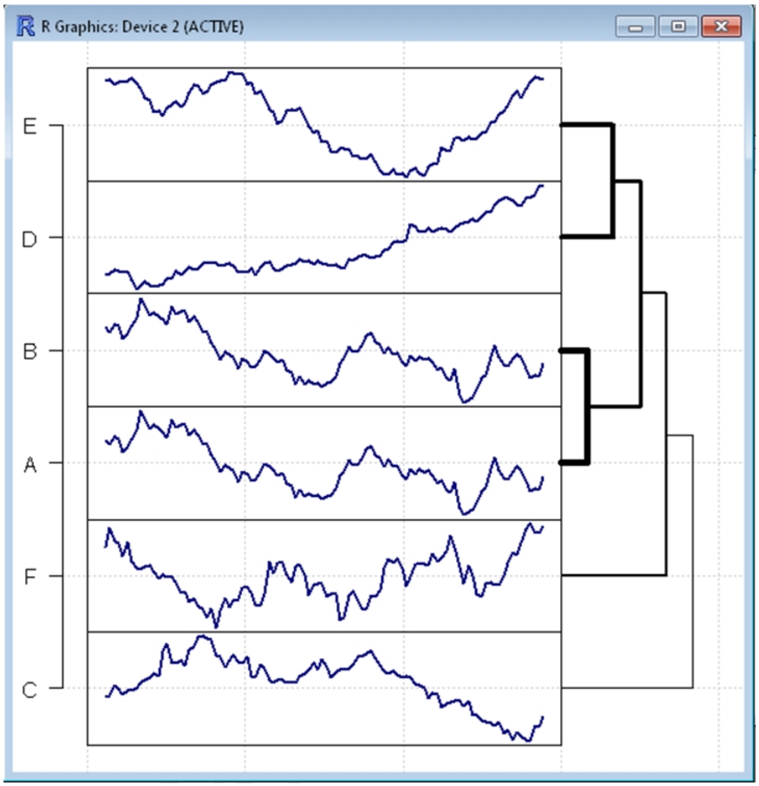


Figure 14.172: **Dendrograms TSA microarray plot in R 3.0.2**

Let's go now for a detailed step by step example! For this consider the following table:

Object	**Weight**	**Length**
a	12	10
b	15	25
c	30	55
d	50	100
e	35	70
f	45	70
g	35	60

All that remains is to determine the matrix of Euclidean distances. The following distance matrix is then obtained with the data set presented above:

	a	b	c	d	e	f	g
a	-	15.30	48.47	97.69	64.26	68.48	55.04
b	15.30	-	33.54	82.76	49.24	54.08	40.31
c	48.47	33.54	-	49.24	15.81	21.21	7.07
d	97.69	82.76	49.24	-	33.54	30.41	42.72
e	64.26	49.24	15.81	33.54	-	10.00	10.00
f	68.48	54.08	21.21	30.41	10.00	-	14.14
g	55.04	40.31	7.07	42.72	10.00	14.14	-

Obviously, the distance between a point and itself is 0 for example $d(a,a)=0$ and the distance:

$$d(a,c)=d(c,a)=\sqrt{(12-30)^2+(10-55)^2}=48.47 \tag{14.1.1883}$$

For the next step, the closest individuals are detected and aggregated into a new cluster and the distance matrix is again calculated. This last one is based on the old one except that the distance between this new cluster formed and the other objects or individuals (or clusters since after 2 iterations, it is possible that clusters already aggregate between them) are recalculated according to the defined method (Single linkage or Complete linkage, ... etc).

On the basis of the previous distance matrix, it can be seen that c and g are the closest (7.07), then they are aggregated. The distance between the cluster $\{\mathrm{c},\mathrm{g}\}$ and the other objects is updated according to the Complete Linkage method (another method could have been used as we have already mention it):

	a	b	{c, g}	d	e	f
a	-	15.30	55.04	97.69	64.26	68.48
b	15.30	-	40.31	82.76	49.24	54.08
{c, g}	55.04	40.31	-	49.24	15.81	21.21
d	97.69	82.76	49.24	-	33.54	30.41
e	64.26	49.24	15.81	33.54	-	10.00
f	68.48	54.08	21.21	30.41	10.00	-

As the method of the complete linkage or the maximum linkage defines the distance between cluster based on the most distant points, then one can calculate:

$$\begin{aligned} d(\mathrm{a},\{\mathrm{c},\mathrm{g}\})&=\max(d(\mathrm{a},\mathrm{c}),d(\mathrm{a},\mathrm{g}))=\max(48.47,55.04)=55.04\\ d(\mathrm{f},\{\mathrm{c},\mathrm{g}\})&=\max(d(\mathrm{f},\mathrm{c}),d(\mathrm{f},\mathrm{g}))=\max(21.21,14.14)=21.21 \end{aligned} \tag{14.1.1884}$$

On the basis of the updated matrix above, a close distance between f and e is detected since their distance is the smallest and equal to 10. Again, a cluster $\{\mathrm{e},\mathrm{f}\}$ is constructed and their distance from the others will also be readjusted according to maximum linkage:

	a	b	{c, g}	d	{e, f}
a	-	15.30	55.04	97.69	68.48
b	15.30	-	40.31	82.76	54.08
{c, g}	55.04	40.31	-	49.24	21.21
d	97.69	82.76	49.24	-	33.54
{e, f}	68.48	54.08	21.21	33.54	-

As the method of the Complete linkage or the maximum linkage defines the distance between cluster based on the most distant points, then one can calculate:

$$\begin{aligned} d(\mathrm{a},\{\mathrm{e},\mathrm{f}\})&=\max(d(\mathrm{a},\mathrm{e}),d(\mathrm{a},\mathrm{f}))=\max(64.26,68.48)=68.48\\ d(\{\mathrm{c},\mathrm{g}\},\{\mathrm{e},\mathrm{f}\})&=\max(d(\{\mathrm{c},\mathrm{g}\},\mathrm{c}),d(\{\mathrm{c},\mathrm{g}\},\mathrm{f}))=\max(15.81,21.21)=21.21 \end{aligned} \tag{14.1.1885}$$

This time it is a and b that will be clustered at a distance of 15.3 and the matrix after updating the distances becomes this:

	$\{a,b\}$	$\{c,g\}$	d	$\{e,f\}$
$\{a,b\}$	-	55.04	97.69	68.48
$\{c,g\}$	55.04	-	49.24	21.21
d	97.69	49.24	-	33.54
$\{e,f\}$	68.48	21.21	33.54	-

The following matrix will give what we have for distances after that $\{\mathrm{c},\mathrm{g}\}$ and $\{\mathrm{e},\mathrm{f}\}$ are grouped to another cluster at a distance of 21.21:

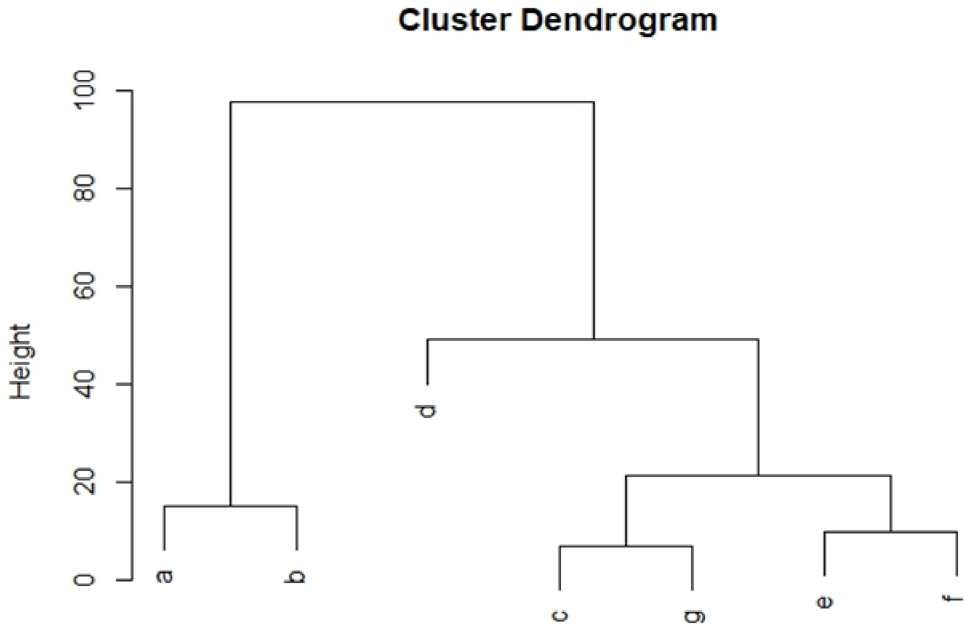
	$\{a,b\}$	$\{\{c,g\},\{e,f\}\}$	d
$\{a,b\}$	-	55.04	97.69
$\{\{c,g\},\{e,f\}\}$	55.04	-	49.24
d	97.69	49.24	-

And then it is the cluster group $\{\{\mathrm{c},\mathrm{g}\},\{\mathrm{e},\mathrm{f}\}\}$ and d that will be grouped at a distance of 49.24, updating the new distances gives:

	$\{a,b\}$	$\{\{c,g\},\{e,f\},d\}$
$\{a,b\}$	-	97.69
$\{\{c,g\},\{e,f\},d\}$	97.69	-

The algorithm ends at the next iteration when $\{\mathrm{a},\mathrm{b}\}$ and the cluster group $\{\{\mathrm{c},\mathrm{g}\},\{\mathrm{e},\mathrm{f}\},\mathrm{d}\}$ will form the root cluster at a distance of 97.69.

This gives us finally the following hierarchical agglomerative ascendant clustering:



The vertical axis *Height* above corresponds to the following calculations we did above:

- $\{\mathrm{c},\mathrm{g}\}$: 7.07
- $\{\mathrm{e},\mathrm{f}\}$: 10.00
- $\{\mathrm{a},\mathrm{b}\}$: 15.3
- $\{\{\mathrm{c},\mathrm{g}\},\{\mathrm{e},\mathrm{f}\}\}$: 21.1
- $\{\{\mathrm{c},\mathrm{g}\},\{\mathrm{e},\mathrm{f}\},d\}$: 97.69

Hierarchical clustering offers the advantage of presenting hierarchical groups so that the resulting ramifications are easy to read. In contrast to the k-means, it does not require the definition of the number of clusters in advance and it is reproducible in the sense that the choice of the initial cluster centers is not random. But in return it is more expensive in time when it comes to implementing it for a large database.

This is why in practice, when dealing with huge databases we use mixed (or hybrid) methods, i.e. before doing a HAC, we first run a k-means and after we do a HAC on the resulting k-means (therefore the bottom of the dendrogram tree will be obtain with a k-means and the upper part with the HAC). This process is named "hybrid hierarchical k-means clustering" (hkmeans) and the reader can see a practical example in our `R` companion book.

Obviously as for the k-nn and the k-means, we may run the HAC on the factorial space obtained after having run a PCA (Principal Component Analysis as seen on page 1213) on our data (see our `R` companion book for an application example). We can also even reduce the dimensions using that latter if necessary!

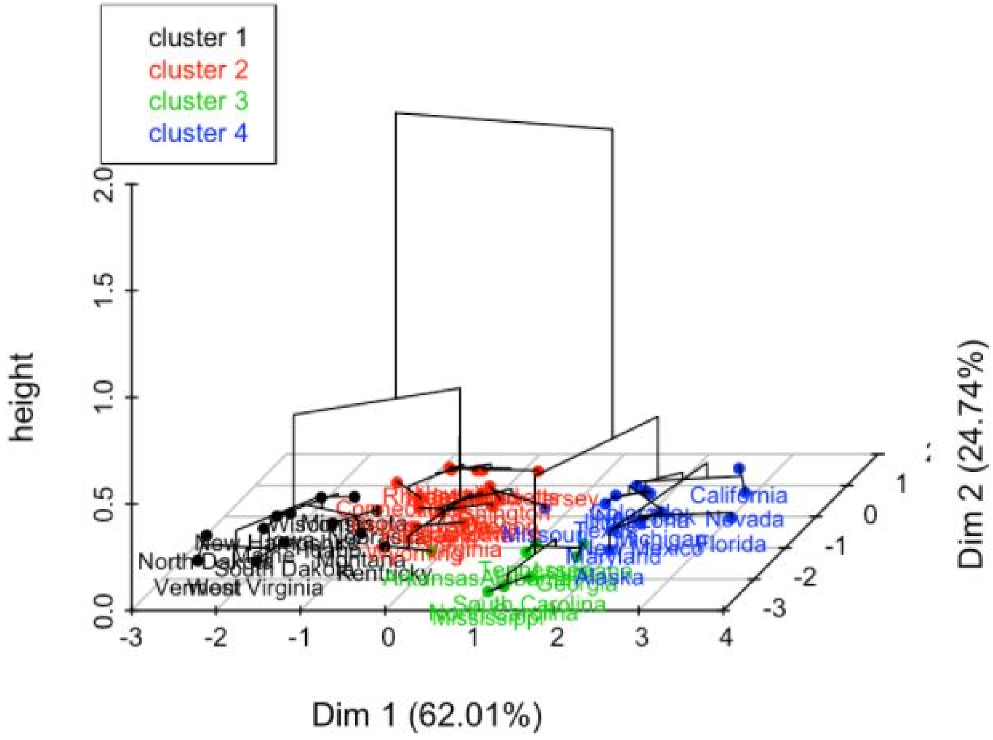


Figure 14.173: HAC on a factor map with R

Remark

To do a classification on qualitative data we will use typically a Multiple Correspondence Analysis (MCA).

14.1.20.5.10 Linear Discriminant Analysis

"Fisher Linear Discriminant Analysis" is a method used in statistics, pattern recognition and machine learning to find a linear combination of features which characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction before later classification.

There are commonly three linear discriminant methods for this technique:

- Maximum likelihood: Assigns $\vec{x}$ to the group that maximizes population (group) density.
- Bayes Discriminant Rule: Assigns $\vec{x}$ to the group that maximizes $\pi_i f_i(\vec{x})$ where π_i represents the prior probability of that classification, and $f_i(x)$ represents the population density.
- Fisher's linear discriminant rule. We will focus here on that method which is the easiest (so see details below!)

LDA is closely related to PCA, for both of them are based on linear, i.e. matrix multiplication, transformations. For the case of PCA, the transformation is based on minimizing mean square error between original data vectors and data vectors that can be estimated fro the reduced dimensionality data vectors. And the PCA does not take into account any difference in class. But for the case of LDA, the transformation is based on maximizing a ratio of “between-class variance" to "within-class variance" with the goal of reducing data variation in the same class and increasing the separation between classes. Let's see an example of LDA.

The idea is seeking to obtain a scalar y (in the special two classes case it's a scalar otherwise it's a vector!) by projecting the samples X onto a line:

$$y = \vec{\theta}^T X \qquad (14.1.1886)$$

Then we try to find the $\vec{\theta}^*$ to maximize the ratio of "between-class variance" to "within-class variance". Next, we will introduce how to use mathematics way to present this problem.

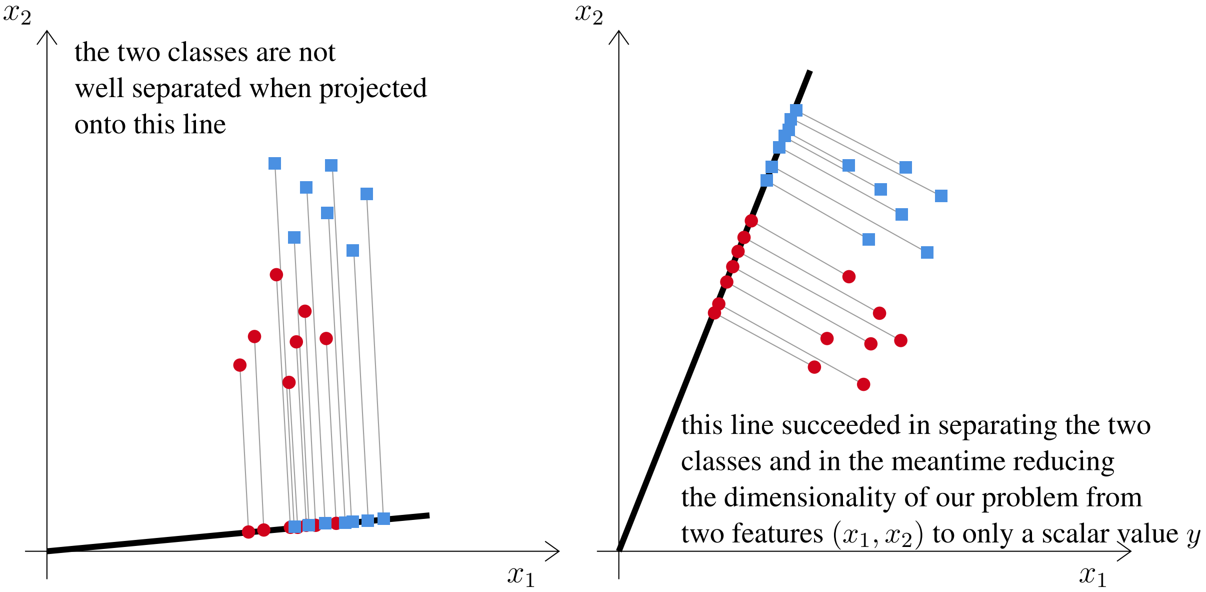


Figure 14.174: Underlying idea of linear discriminant analysis

To figure out the LDA, first we need know how to translate "between-class variance" and "within-class variance" to mathematics language. Then we try to maximize the ratio between these two. To simplify the problem, we start with two classes problem.

14.1.20.5.10.1 Two classes problem

Assume we have a set of D-dimensional samples $X = \{\vec{x}^{(1)}, \vec{x}^{(2)}, \ldots, \vec{x}^{(m)}\}$, n_1 of which belong to class $\mathcal{C}_1$, and n_2 of which belong to class $\mathcal{C}_2$.

We also assume the mean vector of two classes in X-space (if needed, there is a practical detailed numerical example just a little bit further below that can help the reader to understand what follows!):

$$\vec{u}_k = \frac{1}{n_k} \sum_{i \in \mathcal{C}_k} \vec{x}^{(i)} \quad \text{where} \quad k = 1, 2 \tag{14.1.1887}$$

and in y-space:

$$\hat{u}_k = \frac{1}{n_k} \sum_{i \in \mathcal{C}_k} y^{(i)} = \frac{1}{n_k} \sum_{i \in \mathcal{C}_k} \vec{\theta}^T \vec{x}^{(i)} = \vec{\theta}^T \vec{u}_k \quad \text{where} \quad k = 1, 2 \tag{14.1.1888}$$

One way to define a measure of separation between two classes is to choose the distance between the projected means, which is in y-space, so the "between-class variance" is:

$$\hat{u}_2 - \hat{u}_1 = \vec{\theta}^T (\vec{u}_2 - \vec{u}_1) \tag{14.1.1889}$$

Also, we can define the "within-class variance" for each class $\mathcal{C}_k$ is:

$$\hat{s}_k^2 = \sum_{i \in C_k} (\vec{y}^{(i)} - \hat{u}_k)^2 \quad \text{where} \quad k = 1, 2 \tag{14.1.1890}$$

Then, we get the between-class variance and within-class variance, we can define our objective function $J(\theta)$ as:

$$J(\theta)=\frac{(\hat{u}_2-\hat{u}_1)^2}{\hat{s}_1^2+\hat{s}_2^2} \tag{14.1.1891}$$

In fact, if maximizing the objective function J, we are looking for a projection where examples from the class are projected very close to each other and at the same time, the projected means are as farther apart as possible.

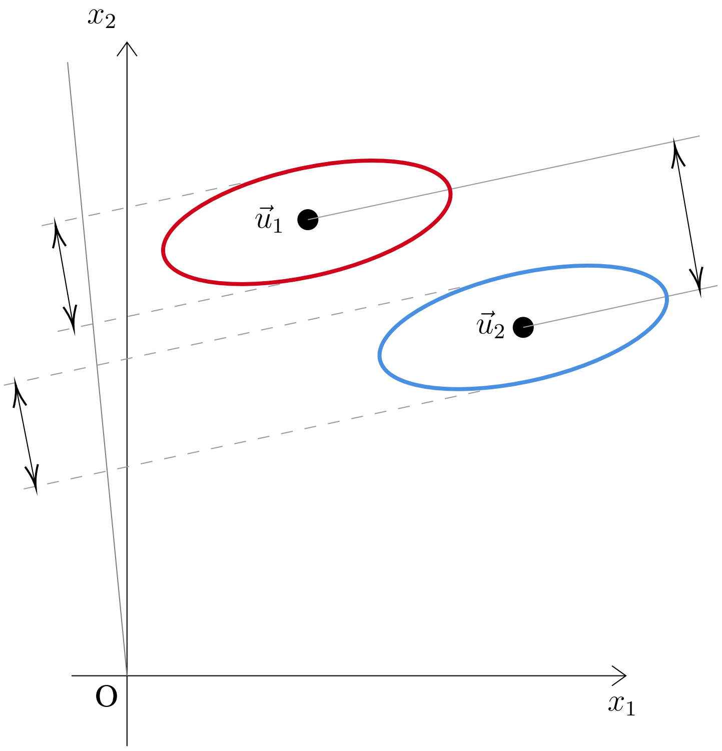


Figure 14.175: **Expected configuration for linear Discriminant analysis**

To find the optimum θ^*, we must express $J(\theta)$ as a function of θ. Before the optimum,we need introduce the "scatter" matrix instead of variance.

We define some measures of the scatter as following:

- The scatter matrix in feature space-x:

$$\vec{S}_k=\sum_{i\in\mathcal{C}_k}(\vec{x}^{(i)}-\vec{u}_k)(\vec{x}^{(i)}-\vec{u}_k)^T \tag{14.1.1892}$$

- The within-class scatter matrix:

$$S_W=S_1+S_2 \tag{14.1.1893}$$

- The between-class scatter matrix:

$$S_B=(\vec{u}_2-\vec{u}_1)(\vec{u}_2-\vec{u}_1)^T \tag{14.1.1894}$$

Let's see $J(\theta)$ again:

$$J(\theta)=\frac{(\hat{u}_2-\hat{u}_1)^2}{\hat{s}_1^2+\hat{s}_2^2} \tag{14.1.1895}$$

The scatter of the projection y can then be expressed as a function of the scatter matrix in feature space X:

$$\begin{aligned}\hat{s}_k^2 &=\sum_{i\in\mathcal{C}_k}(\vec{y}^{(i)}-\hat{u}_k)^2=\sum_{i\in\mathcal{C}_k}(\vec{\theta}^T\vec{x}^{(i)}-\vec{\theta}^T\vec{u}_k)^2=\sum_{i\in\mathcal{C}_k}\vec{\theta}^T(\vec{x}^{(i)}-\vec{u}_k)(\vec{x}^{(i)}-\vec{u}_k)^T\vec{\theta}\\ &=\vec{\theta}^T S_k\vec{\theta}\end{aligned} \tag{14.1.1896}$$

So we can get:

$$\hat{s}_1^2+\hat{s}_2^2=\vec{\theta}^T S_1\theta+\vec{\theta}^T S_2\vec{\theta}=\vec{\theta}^T S_W\vec{\theta} \tag{14.1.1897}$$

Similarly, the difference between the projected means can be expressed in terms of the means in the original feature space:

$$(\hat{u}_2-\hat{u}_1)^2=(\vec{\theta}^T\vec{u}_2-\vec{\theta}^T\vec{u}_1)^2=\vec{\theta}^T(\vec{u}_2-\vec{u}_1)(\vec{u}_2-\vec{u}_1)^T\vec{\theta}=\vec{\theta}^T S_B\vec{\theta} \tag{14.1.1898}$$

We can finally express the "Fisher criterion" in terms of S_W and S_B as:

$$\boxed{J(\vec{\theta})=\frac{\vec{\theta}^T S_B\vec{\theta}}{\vec{\theta}^T S_W\vec{\theta}}} \tag{14.1.1899}$$

Next, we will maximize this objective function.

The easiest way to maximize the object function J is to derive it and set it to zero:

$$\begin{aligned}\frac{\partial J(\vec{\theta})}{\partial\vec{\theta}} &=\frac{\partial}{\partial\vec{\theta}}\left(\frac{\vec{\theta}^T S_B\vec{\theta}}{\vec{\theta}^T S_W\vec{\theta}}\right)\\ &=(\vec{\theta}^T S_W\vec{\theta})\frac{\partial(\vec{\theta}^T S_B\vec{\theta})}{\partial\vec{\theta}}-(\theta^T S_B\vec{\theta})\frac{\partial(\vec{\theta}^T S_W\vec{\theta})}{\partial\vec{\theta}}=\vec{0}\\ \Longrightarrow &=(\vec{\theta}^T S_W\vec{\theta})2S_B\vec{\theta}-(\vec{\theta}^T S_B\vec{\theta})2S_W\vec{\theta}=\vec{0}\end{aligned} \tag{14.1.1900}$$

Divided by $\vec{\theta}^T S_W\vec{\theta}$:

$$\begin{aligned}&\Longrightarrow\left(\frac{\vec{\theta}^T S_W\vec{\theta}}{\vec{\theta}^T S_W\vec{\theta}}\right)S_B\vec{\theta}-\left(\frac{\vec{\theta}^T S_B\vec{\theta}}{\vec{\theta}^T S_W\vec{\theta}}\right)S_W\vec{\theta}=\vec{0}\\ &\Longrightarrow S_B\vec{\theta}-JS_W\vec{\theta}=\vec{0}\\ &\Longrightarrow S_W^{-1}S_B\vec{\theta}-J\vec{\theta}=\vec{0}\\ &\Longrightarrow J\vec{\theta}=S_W^{-1}S_B\vec{\theta}\\ &\Longrightarrow J\vec{\theta}=S_W^{-1}(\vec{u}_2-\vec{u}_1)(\vec{u}_2-\vec{u}_1)^T\vec{\theta}\\ &\Longrightarrow J\vec{\theta}=S_W^{-1}(\vec{u}_2-\vec{u}_1)(\underbrace{(\vec{u}_2-\vec{u}_1)^T\vec{\theta}}_{c^{te}\in\mathbb{R}})\\ &\Longrightarrow J\vec{\theta}=c^{te}S_W^{-1}(\vec{u}_2-\vec{u}_1)\\ &\Longrightarrow\vec{\theta}=\frac{c^{te}}{J}S_W^{-1}(\vec{u}_2-\vec{u}_1)\end{aligned} \tag{14.1.1901}$$

Notice that using the third row above, we have a classical generalized eigenvalue problem:

$$\left(S_W^{-1}S_B - J\right)\vec{\theta} = \vec{0} \qquad (14.1.1902)$$

Often denoted as ((see section Linear Algebra page 1998):

$$\det\left(S_W^{-1}S_B - \lambda I\right) = 0 \qquad (14.1.1903)$$

For now, the problem has been solved and we just want to get the direction of the $\vec{\theta}$, which is the optimum $\vec{\theta}^*$:

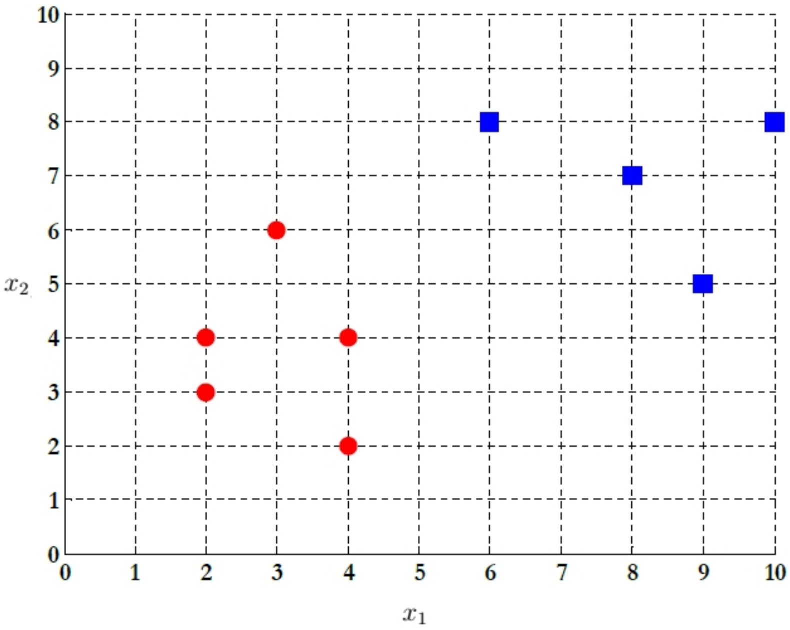
$$\vec{\theta}^* \propto S_W^{-1}(\vec{u}_2 - \vec{u}_1) \qquad (14.1.1904)$$

This is known as "Fisher's linear discriminant" (11936 according to holocene calendar), although it is not a discriminant but rather a specific choice of direction for the projection of the data down to one dimension, which is $y = \vec{\theta}^{*T}X$.

✍**Example:**

We want to compute the Linear Discriminant projection for the following two-dimensional dataset:

- Samples for class $\mathcal{C}_1 : X_1 = (x_1, x_2) = \{(4,2),(2,4),(2,3),(3,6),(4,4)\}$
- Sample for class $\mathcal{C}_2 : X_2 = (x_1, x_2) = \{(9,10),(6,8),(9,5),(8,7),(10,8)\}$



The classes means are :

$$\vec{u}_1 = \frac{1}{n_1}\sum_{\vec{x}\in\mathcal{C}_1}\vec{x}^{(i)} = \frac{1}{5}\left[\begin{pmatrix}4\\2\end{pmatrix}+\begin{pmatrix}2\\4\end{pmatrix}+\begin{pmatrix}2\\3\end{pmatrix}+\begin{pmatrix}3\\6\end{pmatrix}+\begin{pmatrix}4\\4\end{pmatrix}\right] = \begin{pmatrix}3\\3.8\end{pmatrix}$$

$$\vec{u}_2 = \frac{1}{n_2}\sum_{\vec{x}\in\mathcal{C}_2}\vec{x}^{(i)} = \frac{1}{5}\left[\begin{pmatrix}9\\10\end{pmatrix}+\begin{pmatrix}6\\8\end{pmatrix}+\begin{pmatrix}9\\5\end{pmatrix}+\begin{pmatrix}8\\7\end{pmatrix}+\begin{pmatrix}10\\8\end{pmatrix}\right] = \begin{pmatrix}8.4\\7.6\end{pmatrix}$$

The covariance matrix of the first class is:

$$\begin{aligned}
S_1 = \sum_{\vec{x}\in\mathcal{C}_1} (\vec{x}^{(i)} - \vec{u}_1)\,(\vec{x}^{(i)} - \vec{u}_1)^T &= \left[\begin{pmatrix} 4 \\ 2 \end{pmatrix} - \begin{pmatrix} 3 \\ 3.8 \end{pmatrix}\right]^2 + \left[\begin{pmatrix} 2 \\ 4 \end{pmatrix} - \begin{pmatrix} 3 \\ 3.8 \end{pmatrix}\right]^2 \\
&+ \left[\begin{pmatrix} 2 \\ 3 \end{pmatrix} - \begin{pmatrix} 3 \\ 3.8 \end{pmatrix}\right]^2 + \left[\begin{pmatrix} 3 \\ 6 \end{pmatrix} - \begin{pmatrix} 3 \\ 3.8 \end{pmatrix}\right]^2 + \left[\begin{pmatrix} 4 \\ 4 \end{pmatrix} - \begin{pmatrix} 3 \\ 3.8 \end{pmatrix}\right]^2 \\
&= \begin{pmatrix} 1 & -0.25 \\ -0.25 & 2.2 \end{pmatrix}
\end{aligned}$$

Covariance matrix of the second class:

$$\begin{aligned}
S_2 = \sum_{\vec{x}\in\mathcal{C}_2} (\vec{x}^{(i)} - \vec{u}_2)\,(\vec{x}^{(i)} - \vec{u}_2)^T &= \left[\begin{pmatrix} 9 \\ 10 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix}\right]^2 + \left[\begin{pmatrix} 6 \\ 8 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix}\right]^2 \\
&+ \left[\begin{pmatrix} 9 \\ 5 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix}\right]^2 + \left[\begin{pmatrix} 8 \\ 7 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix}\right]^2 + \left[\begin{pmatrix} 10 \\ 8 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix}\right]^2 \\
&= \begin{pmatrix} 2.3 & -0.05 \\ -0.05 & 3.3 \end{pmatrix}
\end{aligned}$$

Therefore we have the within-class scatter matrix:

$$S_W = S_1 + S_2 = \begin{pmatrix} 1 & -0.25 \\ -0.25 & 2.2 \end{pmatrix} + \begin{pmatrix} 2.3 & -0.05 \\ -0.05 & 3.3 \end{pmatrix} = \begin{pmatrix} 3.3 & -0.3 \\ -0.3 & 5.5 \end{pmatrix}$$

Between-class scatter matrix:

$$\begin{aligned}
S_B = (\vec{u}_1 - \vec{u}_2)\,(\vec{u}_1 - \vec{u}_2)^T &= \left[\begin{pmatrix} 3 \\ 3.8 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix}\right]\left[\begin{pmatrix} 3 \\ 3.8 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix}\right]^T \\
&= \begin{pmatrix} -5.4 \\ -3.8 \end{pmatrix} \begin{pmatrix} -5.4 & -3.8 \end{pmatrix} = \begin{pmatrix} 29.16 & 20.52 \\ 20.52 & 14.44 \end{pmatrix}
\end{aligned}$$

The LDA projection is then obtained as the solution of the generalized eigen value problem:

$$\begin{aligned}
&S_W^{-1} S_B \vec{\theta} = \lambda \vec{\theta} \\
&\qquad \det\left(S_W^{-1} S_B - \lambda I\right) = 0 \\
&\qquad \Rightarrow \left| \begin{pmatrix} 3.3 & -0.3 \\ -0.3 & 5.5 \end{pmatrix}^{-1} \begin{pmatrix} 29.16 & 20.52 \\ 20.52 & 14.44 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right| = 0 \\
&\qquad \Rightarrow \left| \begin{pmatrix} 0.3045 & 0.0166 \\ 0.0166 & 0.1827 \end{pmatrix} \begin{pmatrix} 29.16 & 20.52 \\ 20.52 & 14.44 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right| = 0 \\
&\qquad \Rightarrow \left| \begin{pmatrix} 9.2213 - \lambda & 6.489 \\ 4.2339 & 2.9794 - \lambda \end{pmatrix} \right| \\
&\qquad = \left(9.2213 - \lambda \right)(2.9794 - \lambda) - 6.489 \times 4.2339 = 0 \\
&\qquad \Rightarrow \lambda^2 - 12.2007\lambda = 0 \Rightarrow \lambda(\lambda - 12.2007) = 0 \\
&\qquad \Rightarrow \lambda_1 = 0, \lambda_2 = 12.2007
\end{aligned}$$

We have therefore also two eigenvectors. One is the zero vector so it has no interest. It remains only the second that takes us to write:

$$S_W^{-1}S_B\vec{\theta}_2=\lambda_2\vec{\theta}_2 \Leftrightarrow \begin{pmatrix}9.2213 & 6.489\\ 4.2339 & 2.9794\end{pmatrix}\begin{pmatrix}\theta_1^{(2)}\\ \theta_2^{(2)}\end{pmatrix}=\underbrace{12.2007}_{\lambda_2}\begin{pmatrix}\theta_1^{(2)}\\ \theta_2^{(2)}\end{pmatrix}$$

Therefore we just have to solve the following system:

$$\begin{aligned}9.2213\theta_1^{(2)}+6.489\theta_2^{(2)}&=12.2007\theta_1^{(2)}\\ 4.2339\theta_1^{(2)}+2.979\theta_2^{(2)}&=12.2007\theta_2^{(2)}\end{aligned}$$

Hence:

$$\vec{\theta}^*=\vec{\theta}_2=\begin{pmatrix}\theta_1^{(2)}\\ \theta_2^{(2)}\end{pmatrix}=\begin{pmatrix}0.9088\\ 0.4173\end{pmatrix}$$

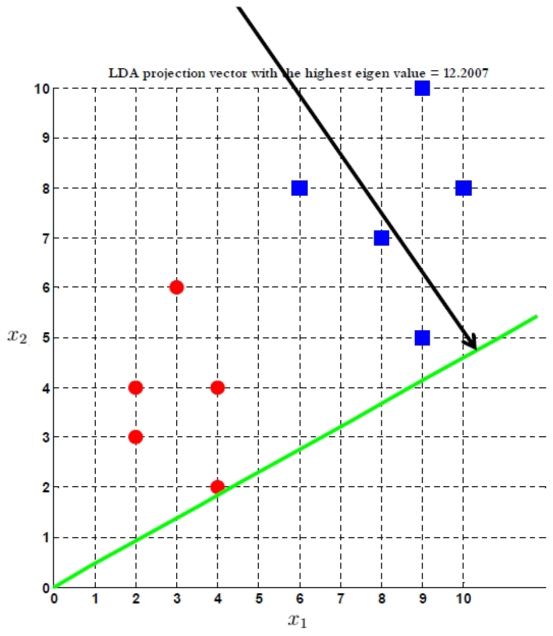
We can get the same result more directly using the relation derived also earlier above:

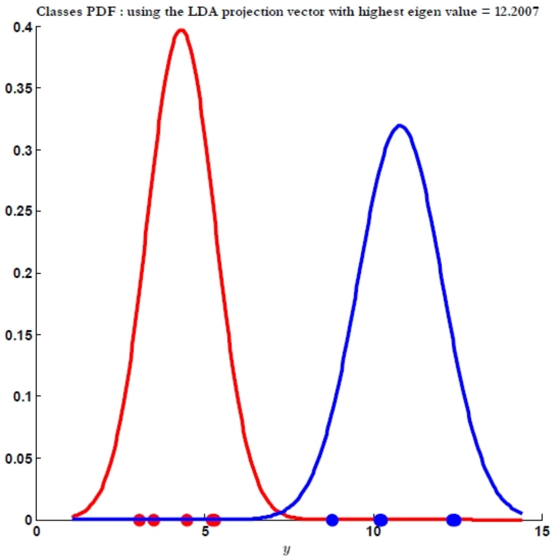
$$\vec{\theta}^*\propto S_W^{-1}(\vec{u}_2-\vec{u}_1)$$

Hence:

$$\begin{aligned}\vec{\theta}^*=S_W^{-1}(\mu_1-\mu_2)&=\begin{pmatrix}3.3 & -0.3\\ -0.3 & 5.5\end{pmatrix}^{-1}\left[\begin{pmatrix}3\\ 3.8\end{pmatrix}-\begin{pmatrix}8.4\\ 7.6\end{pmatrix}\right]\\ &=\begin{pmatrix}0.3045 & 0.0166\\ 0.0166 & 0.1827\end{pmatrix}\begin{pmatrix}-5.4\\ -3.8\end{pmatrix}=\begin{pmatrix}0.9088\\ 0.4173\end{pmatrix}\end{aligned}$$

The projection vector corresponding to the **highest** eigen value





Using this vector leads to **good separability** between the two classes

Notice that LDF will fail when the discriminatory information is not in the mean but rather in the variance of the data as illustrated below:

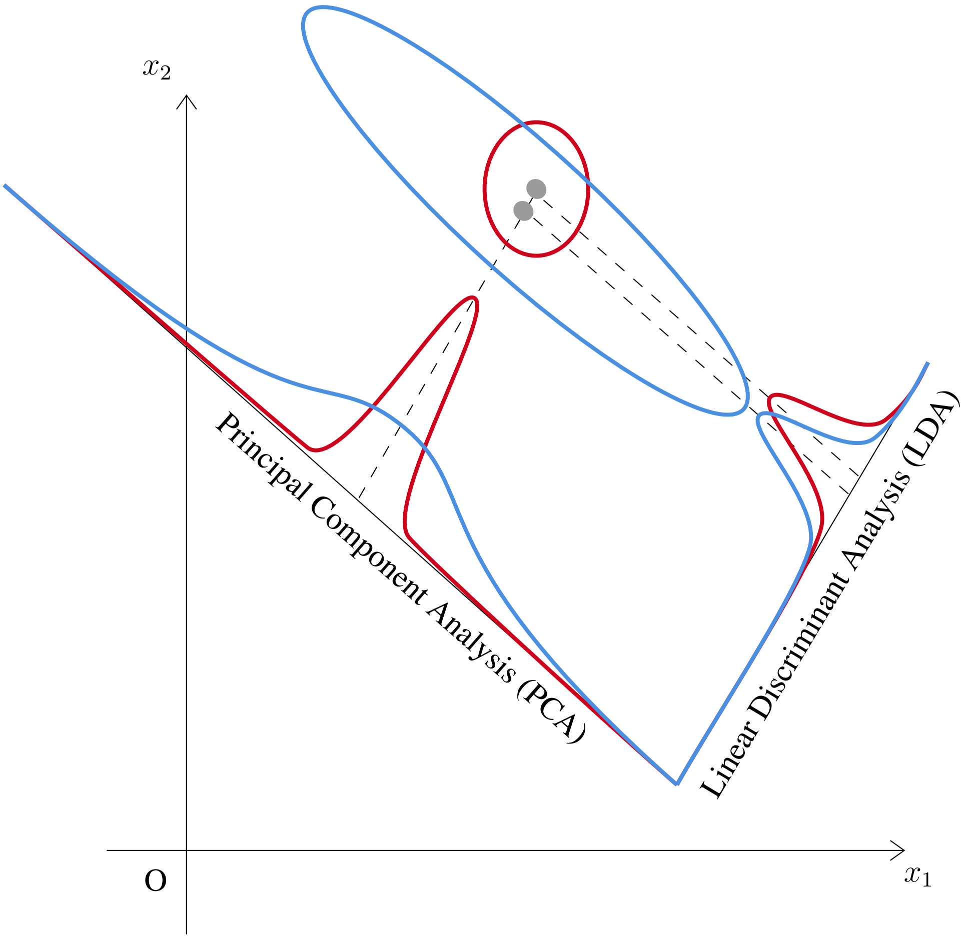

x_2
x_1
O
Principal Component Analysis (PCA)
Linear Discriminant Analysis (LDA)

Figure 14.176: Typical case where LDA fail and PCA don't

Through this example with see better why LDA is not only a classification technique. It is a composite, two-stage technique: first it reduce dimensionality, then it classifies.

Classification of new data that weren't in the training set is essentially done with the class that belongs the closest class mean[68].

Remark

The reader can refer to our R or MATLAB® free companion books to see examples with the respective script language of these two softwares for the same set of numerical data.

---

[68] And in the case of the Bayesian approach of LDA a prior term is added to shift the mean closer or further from the new point.

14.1.20.5.10.2 Multiclass problem

Based on two classes problem, we can see that the Fisher's LDA generalizes gracefully for multiple classes problem (we speak then of "Multiple Discriminant Analysis"). Assume we still have a set of D-dimensional samples $X=\{\vec{x}^{(1)},\vec{x}^{(2)},\ldots,\vec{x}^{(m)}\}$, and there are totally C classes. Instead of one projection y, mentioned above, we now will seek $(\mathcal{C}-1)$ projections $[y_1,y_2,\ldots y_{\mathcal{C}-1}]$ by means of $(\mathcal{C}-1)$ projection vectors $\vec{\theta}_i$ arranged by columns into a projection matrix $\Theta=[\vec{\theta}_1|\vec{\theta}_2|\ldots|\vec{\theta}_{C-1}]$, where:

$$y_i=\vec{\theta}_i^T X \Longrightarrow \vec{y}=\Theta^T X \tag{14.1.1905}$$

First we will use the scatters in space-x as following:

- Within-class scatter matrix:

$$S_W=\sum_{i=1}^{\mathcal{C}} S_i \quad \text{where} \quad S_i=\sum_{i\in\mathcal{C}_i}(\vec{x}^{(i)}-\vec{u}_i)(\vec{x}^{(i)}-\vec{u}_i)^T \text{ and } \vec{u}_i=\frac{1}{n_i}\sum_{i\in\mathcal{C}_i}\vec{x}^{(i)} \tag{14.1.1906}$$

- Between-class scatter matrix:

$$S_B=\sum_{i=1}^{\mathcal{C}} n_i(\vec{u}_i-\vec{u})(\vec{u}_i-\vec{u})^T \quad \text{where} \quad \vec{u}=\frac{1}{m}\sum_{i=1}^{m}\vec{x}^{(i)}=\frac{1}{m}\sum_{i=1}^{\mathcal{C}} n_i\vec{u}_i \tag{14.1.1907}$$

- Total scatter matrix:

$$S_T=S_B+S_W \tag{14.1.1908}$$

Similarly, we define the mean vector and scatter matrices for the projected samples as:

- $\vec{\hat{u}}_i=\frac{1}{n_i}\sum_{i\in\mathcal{C}_i}\vec{y}^{(i)}$
- $\vec{\hat{u}}=\frac{1}{N}\sum_{i=1}^{m}\vec{y}^{(i)}$
- $\hat{S}_W=\sum_{i=1}^{\mathcal{C}}\sum_{\vec{y}\in\mathcal{C}_i}(\vec{y}-\vec{\hat{u}}_i)(\vec{y}-\vec{\hat{u}}_i)^T$
- $\hat{S}_B=\sum_{i=1}^{\mathcal{C}} n_i(\vec{\hat{u}}_i-\vec{\hat{u}})(\vec{\hat{u}}_i-\vec{\hat{u}})^T$

From our derivation for the two-class problem, we can get:

$$\begin{gathered}\hat{S}_W=\Theta^T S_W\Theta\\ \hat{S}_B=\Theta^T S_B\Theta\end{gathered} \tag{14.1.1909}$$

Recall that we are looking for a projection that maximizes the ratio of between-class to within-class scatter. Since the projection is no longer a scalar (it has $\mathcal{C}-1$ dimensions), we use the determinant of the scatter matrices to obtain a scalar objective function:

$$J(\Theta)=\frac{|\hat{S}_B|}{|\hat{S}_W|}=\frac{|\Theta^T S_B\Theta|}{|\Theta^T S_W\Theta|} \tag{14.1.1910}$$

And now, our job is to seek the projection matrix Θ^* that maximize this ratio. We will not give the derivation process. But we know that the optimal projection matrix Θ^* is the one whose columns are the eigenvectors corresponding to the largest eigenvalues of the following generalized eigenvalue problem:

$$\Theta^* = [\theta_1^*|\theta_2^*|\ldots|\theta_{C-1}^*] = \text{argmax}\frac{|\Theta^T S_B \Theta|}{|\Theta^T S_W \Theta|} \tag{14.1.1911}$$

Hence:

$$(S_B - \lambda_i S_W)\Theta_i^* = 0 \tag{14.1.1912}$$

Thus, if S_W is a non-singular matrix, and can be inverted, then the Fisher's criterion is maximized when the projection matrix Θ^* is composed of the eigenvectors of:

$$S_W^{-1} S_B \tag{14.1.1913}$$

Noticed that, there will be at most $\mathcal{C}-1$ eigenvectors with non-zero real corresponding eigenvalues λ_i. This is because S_B is of rank $(\mathcal{C}-1)$ or less. So we can see that LDA can represent a massive reduction in the dimensionality of the problem. In face recognition for example there may be several thousand variables, but only a few hundred classes.

Remark

The reader must keep in mind that what we have seen here is a naive approach of LDA. Investigating further and using Bayesian analysis, we can prove that the above results are just a special case of that latter. Maybe (or maybe not) we will provide the detailed bayesian derivation in the 5th edition of the present book.

14.1.20.6 Dimensionality Reduction

"Dimensionality reduction" or "dimension reduction" is the process of reducing the number of random variables under consideration.

More technically, dimensionality reduction is typically choosing a basis or mathematical representation within which you can describe most but not all of the variance within your data, thereby retaining the relevant information, while reducing the amount of information necessary to represent it.

Approaches can be divided into "feature selection" and "feature extraction".

Definitions (#14.1.572):

D1. "Feature selection" approaches try to find a subset of the known input variables (also naed "features" or "attributes"). The three strategies are: the filter strategy (e.g. information gain), the wrapper strategy (e.g. search guided by accuracy), and the embedded strategy (selected features are added or are removed while building the model based on prediction errors).

Typical feature selection methods are:

- Missing Value Ratio
- Low Variance Filter
- High Correlation Filter
- Random Forests
- Backward Feature Selection
- Forward Feature Selection
- ...

D2. "Feature extraction" approaches transform the data from a well known interpretable high-dimensional space to a space of smaller abstract dimensional space. The data transformation may be linear, as in Principal Component Analysis (we have already study quite in detail the naive version of Principal Component Analysis in the section Statistics page 1213), but many non-linear dimensionality reduction techniques also exist.

Typical feature extraction methods are:

- Components/Factor based:
  - Factor Analysis
  - Principal Component Analysis
  - Independent Component Analysis
  - ...
- Projection based:
  - ISOMAP
  - T-SNE
  - UMAP

- ∗ Linear discriminant analysis (LDA)
- ∗ ...

For a more exhaustive list of techniques the reader can refer as always to the Data Science Mind Map at page 6770.

14.1.20.6.1 Principal Component Analysis

As already mentioned earlier, PCA unsupervised feature selection dimensionality reduction is already treated quite with details in its naive version in the section of Statistics at page 1213.

14.1.20.6.2 Linear Discriminant Analysis (LDA)

The LDA supervised projection dimensionality reduction method has also been treated earlier quite in details in it's naive version at page 6908 because it is also a classification technique as we saw it!

So let us focus on the second most common component/factor based dimensionality reduction technique that is MDS:

14.1.20.6.3 MultiDimensional Scaling (MDS)

The goal of "multidimensional scaling" (MDS) is to visualize a set of N objects based on their similarities measured in n different aspects, and the "classical MDS" (cMDS), also known as "principal coordinates analysis" (PCoA), is one of the methods for MDS but that adds the possibility to reduce the space dimension of the original set.

Let us introduce the first aspect of MDS. Being known the level of similarity of individual cases of a dataset (without knowing their intrinsic characteristics!), the purpose of MDS is to find a way of visualizing the original intrinsic unknown characteristics. In other words, MDS is used to translate information about the pairwise distances among a set of N objects or individuals into a configuration of N points mapped into an abstract Cartesian space.

Formally, given a pairwise similarity matrix $D_x = [d_{ij}^x]_{N\times N}$ (see section Linear Algebra page 1959) of a set of N objects (N data points in an n-dimensional space) $X = [\vec{x}_1,\ldots,\vec{x}_N]$ but of unknown components, where the ij-th component d_{ij}^x is some measurement of the similarity between $\vec{x}_i$ and $\vec{x}_j$ such as the Euclidean distance $d_{ij}^x = ||\vec{x}_i - \vec{x}_j||_2$ between $\vec{x}_i$ and $\vec{x}_j$, we want to be able to construct a map $Y = [\vec{y}_1,\ldots,\vec{y}_N]$ in an n dimensional space that explains well enough the original D_x by reversing the original unknown components of the $X = [\vec{x}_1,\ldots,\vec{x}_N]$. When $n = 2$ or 3, the resulting dataset Y can be visualized and the corresponding chart is then name a "perceptual plot".

So the idea is given the only known dissimilarity matrix $D_x = [d_{ij}^x]_{N\times N}$ to find a best estimate of the original unknown $X = [\vec{x}_1,\ldots,\vec{x}_N]$, that we will denote $Y = [\vec{y}_1,\ldots,\vec{y}_N]$. These estimates will obviously lead to a dissimilarity matrix that we will denote $D_y = [d_{ij}^y]_{N\times N}$.

Our goal again is then to find $Y = [\vec{y}_1,\ldots,\vec{y}_N]$ so that its pairwise similarity matrix D_y matches D_x optimally, in the sense that the following objective function, named sometimes "strain" in the MDS study field, is minimized:

$$o(Y) = \left\|D_y^2 - D_x^2\right\| \tag{14.1.1914}$$

with obviously all d_{ij}^x who are given, but where the given estimated pairwise similarity of Euclidean distance:

$$d_{ij}^y = \|\vec{y}_i - \vec{y}_j\| = \sqrt{(\vec{y}_i - \vec{y}_j)^T(\vec{y}_i - \vec{y}_j)} \tag{14.1.1915}$$

have to be determined.

Remark

This optimization problem is also often denoted:

$$o(\hat{X}) = \left\|D_{\hat{x}}^2 - D_x^2\right\| \tag{14.1.1916}$$

with D_x^2 given and $D_{\hat{x}}$ to be found when calculated as:

$$d_{ij}^{\hat{x}} = \left\|\vec{\hat{x}}_i - \vec{\hat{x}}_j\right\| = \sqrt{\left(\vec{\hat{x}}_i - \vec{\hat{x}}_j\right)^T\left(\vec{\hat{x}}_i - \vec{\hat{x}}_j\right)} \tag{14.1.1917}$$

The optimisation problem above is often written explicitly as (but there are other empirical choices as we already know that leads to different final results):

$$\arg\min_{Y} \sum_{i,j} \left(d_{ij}^{y} - d_{ij}^{x}\right)^2 \tag{14.1.1918}$$

otherwise it doesn't make really sense.

This relation is often normalized (to have values that have non physical units) and root squared (it's more convenient to deal with distance)and a special name is given to it, the "Kruskal's Stress coefficient" and denoted:

$$\arg\min_{Y} \text{Stress}_D \sqrt{\frac{\sum_{i,j} \left(d_{ij}^{y} - d_{ij}^{x}\right)^2}{\sum_{i,j} d_{ij}^{x}}} \tag{14.1.1919}$$

Remark

Some computer softwares don't normalize, nor take the square root, that's why the objective final value function may differ between softwares.

14.1.20.6.3.1 Classical MDS (cMDS)

Given again a pairwise similarity matrix $D_x = [d_{ij}^x]_{N\times N}$ of a set of N objects (N data points in an n dimensional space) $X = [\vec{x}_1, \ldots, \vec{x}_N]$, where the ij-th component d_{ij}^x is some measurement of the similarity between $\vec{x}_i$ and $\vec{x}_j$ such as the Euclidean distance $d_{ij}^x = ||\vec{x}_i - \vec{x}_j||_2$ between $\vec{x}_i$ and $\vec{x}_j$, we can construct a map $Y = [\vec{y}_1, \ldots, \vec{y}_N]$ in an m dimensional space ($m < n$) so that its pairwise distance matrix $D_y = [d_{ij}^y]_{N\times N} = ||\vec{y}_i - \vec{y}_j||_2$ explains well enough the original D_x. As $m < n$, cMDS can be considered as a method for dimension reduction.

In some cases only the similarities d_{ij}^x $(i, j = 1, \ldots, N)$ are given, while the N objects in X are not explicitly given, and the dimensionality n may not even be known!

We first assume the given pairwise similarity is the Euclidean distance $d_{ij} = ||\vec{x}_i - \vec{x}_j||_2$:

$$d_{ij}^2 = \|\vec{x}_i - \vec{x}_j\|^2 = (\vec{x}_i - \vec{x}_j)^T(\vec{x}_i - \vec{x}_j) = \|\vec{x}_i\|^2 - 2\vec{x}_i^T\vec{x}_j + \|\vec{x}_j\|^2 \tag{14.1.1920}$$

and we build the following similarity matrix:

$$D_x^2 = \left[d_{ij}^2\right]_{N\times N} = \left[\|\vec{x}_i\|^2 - 2\vec{x}_i^T\vec{x}_j + \|\vec{x}_j\|^2\right]_{N\times N} = X_r - 2X^TX + X_c \tag{14.1.1921}$$

where we have defined the following three matrices to simplify the notations:

$$\begin{gathered}
X_r = \begin{bmatrix} \|\vec{x}_1\|^2 & \cdots & \|\vec{x}_1\|^2 \\ \vdots & \ddots & \vdots \\ \|\vec{x}_N\|^2 & \cdots & \|\vec{x}_N\|^2 \end{bmatrix} \\
X_c = X_r^T = \begin{bmatrix} \|\vec{x}_1\|^2 & \cdots & \|\vec{x}_N\|^2 \\ \vdots & \ddots & \vdots \\ \|\vec{x}_1\|^2 & \cdots & \|\vec{x}_N\|^2 \end{bmatrix} \\
X^TX = \begin{bmatrix} \vec{x}_1^T \\ \vdots \\ \vec{x}_N^T \end{bmatrix} [\vec{x}_1, \cdots, \vec{x}_N] = \begin{bmatrix} \vec{x}_1^T\vec{x}_1 & \cdots & \vec{x}_1^T\vec{x}_N \\ \vdots & \ddots & \vdots \\ \vec{x}_N^T\vec{x}_1 & \ldots & \vec{x}_N^T\vec{x}_N \end{bmatrix}
\end{gathered} \tag{14.1.1922}$$

We then introduce a centering matrix (for recall $\mathbb{J}$ is a matrix of ones as already defined in the section of Linear Algebra):

$$C_N = \mathbb{1}_N - \frac{1}{N}\mathbb{J}_N \tag{14.1.1923}$$

which is an $N \times N$ and symmetric $\text{C}_N^T = \text{C}_N$.

Pre-multiplying C_N by a column vector $\vec{a} = [a_1, \ldots, a_N]^T$ removes the mean $\bar{a} = \sum_{i=1}^N a_i/N$ of from each component of $\vec{a}$:

$$\begin{aligned} C_N\vec{a} = \mathbb{1}\vec{a} - \frac{1}{N}\mathbb{J}_N\vec{a} &= \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}\begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix} - \frac{1}{N}\begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}\begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix} \\ &= \begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix} - \begin{bmatrix} \bar{a} \\ \vdots \\ \bar{a} \end{bmatrix} = \begin{bmatrix} a_1 - \bar{a} \\ \vdots \\ a_N - \bar{a} \end{bmatrix} \end{aligned} \tag{14.1.1924}$$

Taking transpose on both sides, we see that post-multiplying C_N by a row vector $\vec{a}^T$ removes the mean of $\vec{a}^T$ from each component:

$$\vec{a}^T C_N = \vec{a}^T\mathbb{1} - \frac{1}{N}\vec{a}^T\mathbb{J} = [a_1 - \bar{a}, \ldots, a_N - \bar{a}] \tag{14.1.1925}$$

We now apply double centering to D_x^2 by pre and post multiplying C_N to D_x^2:

$$\begin{aligned} C_N D_x^2 C_N &= C_N\left(X_r - 2X^TX + X_c\right)C_N = C_N X_r C_N - 2C_N\left(X^TX\right)C_N + C_N X_c C_N \\ &= -2C_N\left(X^TX\right)C_N = -2\left(XC_N\right)^T\left(XC_N\right) = -2\overline{X}^T\overline{X} \end{aligned} \tag{14.1.1926}$$

where $X_rC_N = \mathbb{0}$ as all components of each row of X_r are the same as their mean, and removing the mean results in a zero row vector. Similarly, $C_NX_c = \mathbb{0}$, and we have also defined just above:

$$\overline{X} = XC_N = \left[\overline{\vec{x}}_1, \ldots, \overline{\vec{x}}_N\right] \tag{14.1.1927}$$

with the mean of each row of X removed, i.e., the ith column is:

$$\overline{\vec{x}}_i = \vec{x}_i - \overline{\vec{x}}, \quad \text{where} \quad \overline{\vec{x}} = \frac{1}{N}\sum_{i=1}^N \vec{x}_i \tag{14.1.1928}$$

We therefore see that $\sum_{i=1}^N \bar{x}_i = 0$, i.e.., all vectors in $\bar{X} = [\bar{x}_1, \ldots, \bar{x}_N]$ are those in $X = [\vec{x}_1, \ldots, \vec{x}_N]$ shifted by the mean vector $\bar{X}$. Now the points in $\bar{X}$ are centralized, but their pairwise similarities remain the same as the given d_{ij}^x.

The relation:

$$C_N D_x^2 C_N = -2\overline{X}^T\overline{X} \tag{14.1.1929}$$

that we have just proved earlier can be rearranged to get:

$$B_x=\overline{X}^T\overline{X}=-\frac{1}{2}C_ND_x^2C_N \tag{14.1.1930}$$

that is Gram matrix (see section Linear Algebra page 2014) of $\overline{X}$.

Our goal is to find $Y=[y_1,\cdots,y_N]$ so that its pairwise similarity matrix D_y matches D_x optimally, in the sense that the following objective function is minimized:

$$o(Y)=\left\|D_y^2-D_x^2\right\| \tag{14.1.1931}$$

Assuming Y is also a double centered matrix as well as X, we get its Gram matrix:

$$B_y=\overline{Y}^T\overline{Y}=-\frac{1}{2}C_ND_y^2C_N \tag{14.1.1932}$$

and the objective function above can be redefined as:

$$o(Y)=\|B_y-B_x\|=\left\|\overline{Y}^T\overline{Y}-\overline{X}^TX\right\| \tag{14.1.1933}$$

We now carry out eigenvalue decomposition (see section Linear Algebra page 2023) of the $N\times N$ symmetric matrix B_x to find its real eigenvalue matrix Λ and the orthogonal eigenvector matrix V satisfying $B_xV=\Lambda V$, i.e.:

$$\begin{aligned}B_x&=V\Lambda V^{-1}=V\Lambda V^T=[\vec{v}_1,\ldots,\vec{v}_N]\begin{bmatrix}\lambda_1&\cdots&0\\\vdots&\ddots&\vdots\\0&\cdots&\lambda_N\end{bmatrix}\begin{bmatrix}\vec{v}_1^T\\\vdots\\\vec{v}_N^T\end{bmatrix}\\&=\left(V\Lambda^{1/2}\right)\left(\Lambda^{1/2}V^T\right)=\left(\Lambda^{1/2}V^T\right)^T\left(\Lambda^{1/2}V^T\right)\end{aligned} \tag{14.1.1934}$$

To minimize the objective function $o(Y)=\|B_y-B_x\|$, we let:

$$Y=\Lambda^{1/2}V^T=\begin{bmatrix}\sqrt{\lambda_1}&\cdots&0\\\vdots&\ddots&\vdots\\0&\cdots&\sqrt{\lambda_N}\end{bmatrix}\begin{bmatrix}\vec{v}_1^T\\\vdots\\\vec{v}_N^T\end{bmatrix}=\begin{bmatrix}\sqrt{\lambda_1}\vec{v}_1^T\\\vdots\\\sqrt{\lambda_N}\vec{v}_N^T\end{bmatrix} \tag{14.1.1935}$$

Each column $\vec{y}_i$ in $Y=[\vec{y}_1,\ldots,\vec{y}_N]$ is an n dimension vector. To reduce the dimensionality from n to a given chosen m, we take the first m rows of Y corresponding to the m greatest eigenvalues of $\lambda_1\geq\ldots\geq\lambda_m\geq\ldots\geq\lambda_N$ and their corresponding eigenvectors $\vec{v}_1,\ldots,\vec{v}_m$ as we do in Principal Component Analysis (see section Statistics page 1213).

Remark

In practice it happens sometimes that only D_x^2 is given without the original coordinates of the N data points $X=[\vec{x}_1,\ldots,\vec{x}_N]$. Also the objective function may have different empirical expression and the distance may be another one than the classic Euclidean.

In summary, here are the different steps:

- Given the pairwise similarity, an $N\times N$ matrix $D_x=[d_{ij}]$, construct the squared proximity matrix $D_x^2=[d_{ij}^2]$

- Apply double centering to get $B_x = -C_N D_x^2 C_N$;
- Find the m greatest eigenvalues $\lambda_1, \ldots, \lambda_m$ of B_x, and the corresponding eigenvectors $\vec{v}_1, \ldots, \vec{v}_m$;
- Get the map:

$$Y_{m\times N} = \Lambda_m^{1/2} V_m = \begin{bmatrix} \sqrt{\lambda_1 \vec{v}_1^T} \\ \vdots \\ \sqrt{\lambda_m \vec{v}_m^T} \end{bmatrix} \tag{14.1.1936}$$

Remark

For a practical application the reader can take a look to our `R` or MATLAB® companion books.

So let us focus on the third most common component/factor based dimensionality reduction technique:

14.1.20.6.4 Factor Analysis (Exploratory Factor Analysis (EFA))

The method of "Factor Analysis" (FA), also named "Exploratory Factor Analysis" (EFA), not to be confuse with the method of Chi-square Correspondence Factor Analysis (see section Statistics page 1233), models a set of D observed manifest variables in $\vec{x}=[x_1,\ldots,x_D]^T$ as a linear combination of a set of $d<D$ unobserved hidden "latent variables" or "common factors" in $\vec{z}=[z_1,\ldots,z_d]^T$, to explain and reveal the variability and dependency among the D observed variables, typically correlated, in terms of the latent variables, assumed to be independent and therefore uncorrelated. The method of FA is therefore considered as a means for dimensionality reduction[69]!

Specifically, we assume each of the observed variables in $\vec{x}$ is a linear combination of the d factors in $\vec{z}$:

$$x_i=\sum_{j=1}^d w_{ij}z_j+\varepsilon_i=[w_{i1},\ldots,w_{id}]\begin{bmatrix} z_1\\ \vdots\\ z_d\end{bmatrix} \tag{14.1.1937}$$

with $i=1,\ldots,D$ or in matrix form:

$$\begin{aligned}\vec{x}=\begin{pmatrix} x_1\\ \vdots\\ \vdots\\ x_D\end{pmatrix}&=\begin{bmatrix} w_{11} & \ldots & w_{1d}\\ \vdots & \ddots & \vdots\\ \vdots & \ddots & \vdots\\ w_{D1} & \ldots & w_{Dd}\end{bmatrix}\begin{pmatrix} z_1\\ \vdots\\ z_d\end{pmatrix}+\begin{pmatrix} \varepsilon_1\\ \vdots\\ \vdots\\ \varepsilon_D\end{pmatrix}\\ &=[\vec{w}_1,\ldots,\vec{w}_d]\begin{bmatrix} z_1\\ \vdots\\ z_d\end{bmatrix}+\vec{\varepsilon}=\sum_{i=1}^d z_i\vec{w}_i+\vec{\varepsilon}=W\vec{z}+\vec{\varepsilon}\end{aligned} \tag{14.1.1938}$$

where:

$$W=[\vec{w}_1,\ldots,\vec{w}_d]=\begin{bmatrix} w_{11} & \ldots & w_{1d}\\ \vdots & \ddots & \vdots\\ w_{D1} & \ldots & w_{Dd}\end{bmatrix} \tag{14.1.1939}$$

is a $D\times d$ factor loading matrix, and ε_i is the noise associated with x_i.

Remark

In many textbooks the reader can find the following notation:

$$x_i-\mu_i=\sum_{j=1}^d l_{ij}F_j+\varepsilon_i=L\vec{F}+\vec{\varepsilon} \tag{14.1.1940}$$

where the l_{ij} are named the "loadings" and the F_j the "latent factors".

---

[69] The text that follows is completely inspired from the lecture notes of the professor Ruye Wang (`http://fourier.eng.hmc.edu/e176/lectures/index_e176.html`)

✍**Example:**

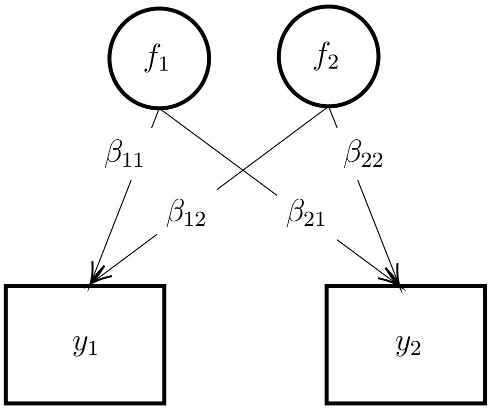
Let's approach this technique by going directly with an example. For this, consider that 5 candidates have passed 3 exams in the field of Finance, Statistics and Standards (Norms) to validate their entry into a MBA course. Each of the corrected exams is marked on a scale ranging from 0 to 10 points corresponding to the variables to be explained. The data is summarized in the table below:

Candidate n°	Finance y_1	Statistics y_2	Norms y_3
1	3	6	5
2	7	3	3
3	10	9	8
4	3	9	7
5	10	6	5

It was suggested by an expert committee that the results obtained are dependent on two explanatory variables not directly observable, f_1 and f_2, which are respectively judged as being the *Analytical Capacity* and the *Memorization Capacity* of the candidates. It is accepted that each of the explained variables depends linearly on these two factors such as (this is a notation common in french textbooks for EFA):

$$\begin{aligned}y_1&=\beta_{10}+\beta_{11}f_1+\beta_{12}f_2+\varepsilon_1\\y_2&=\beta_{20}+\beta_{21}f_1+\beta_{22}f_2+\varepsilon_2\\y_3&=\beta_{30}+\beta_{31}f_1+\beta_{32}f_2+\varepsilon_3\end{aligned}$$

For information, this type of model is represented in the framework of the Structural Equation Modelling (we will introduce this type of model later) in the following way (disregarding the constant and the error terms):



With the following assumptions:

$$\mathrm{E}(\varepsilon_i)=0,\mathrm{V}(\varepsilon_i)=\sigma^2_{\varepsilon_i},\mathrm{cov}(\varepsilon_i,\varepsilon_j)=0\,\forall i\neq j$$

and:

$$\mathrm{E}(f_i)=0,\mathrm{V}(f_i)=1,\mathrm{cov}(f_i,f_j)=0\,\forall i\neq j$$

The model is generally written in french textbooks in the following matrix form, using our particular example:

$$\bar{Y}=\bar{\beta}_0+L\bar{f}+\bar{\varepsilon}=\begin{pmatrix}\beta_{10}\\ \beta_{20}\\ \beta_{30}\end{pmatrix}+\begin{bmatrix}\beta_{11} & \beta_{12} & \beta_{13}\\ \beta_{21} & \beta_{22} & \beta_{23}\\ \beta_{31} & \beta_{32} & \beta_{33}\end{bmatrix}\begin{pmatrix}f_1\\ f_2\\ f_3=0\end{pmatrix}+\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \varepsilon_3\end{pmatrix}$$

Also, for simplicity and without loss of generality, we assume the dataset has a zero mean. If W were available, the d factors in $\vec{z}$ can be found by solving this over-determined linear equation system of D equations but $d<D$ unknowns by the least-square method as seen at page 6381 (with minimum squared error $||\vec{\varepsilon}||^2$):

$$\hat{\vec{z}}=W^{-}\vec{x}=(W^TW)^{-1}W^T\vec{x} \tag{14.1.1941}$$

where for recall $W^{-}=(W^TW)^{-1}W^T$ is the left pseudo-inverse of W.

However, as neither W nor $\vec{z}$ is available, we need to estimate both of them at the same time, based on a set of N observed data points $X=[\vec{x}_1,\ldots,\vec{x}_N]$, typically $N\gg d$, by the method of expectation-maximization (EM), which in general is an iterative algorithm to find the maximum likelihood (ML) or maximum a posteriori (MAP) estimates of some model parameters.

We treat both $\vec{z}$ and $\vec{\varepsilon}$ as random vectors, and make the following assumptions:

- The latent variables in $\vec{z}$ are of zero mean, independent of each other, and with unity variance:

  $$\vec{\mu}_z=\mathrm{E}(\vec{z})=\vec{0},\qquad \Sigma_z=\mathbb{1} \tag{14.1.1942}$$

  and they are Normally distributed:

  $$P(\vec{z})=\mathcal{N}(\vec{0},\mathbb{1}) \tag{14.1.1943}$$

- The noise components in $\vec{\varepsilon}$ are of zero mean, independent of each other, and with a diagonal covariance matrix:

  $$\vec{\mu}_\varepsilon=\mathrm{E}(\vec{\varepsilon})=\vec{0}\qquad \mathbf{\Sigma}_\varepsilon=\Psi=\begin{bmatrix}\psi_1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \psi_D\end{bmatrix} \tag{14.1.1944}$$

  and they are also normally distributed:

  $$P(\vec{\varepsilon})=\mathcal{N}(\vec{0},\Psi) \tag{14.1.1945}$$

- The latent variables and the noise are independent of each other:

  $$\Sigma_{z\varepsilon}=\mathrm{cov}(\vec{z},\vec{\varepsilon})=\mathrm{E}\left[(\vec{z}-\vec{\mu}_z)(\vec{\varepsilon}-\vec{\mu}_\varepsilon)^T\right]=\mathrm{E}(\vec{z}\vec{\varepsilon}^T)=\mathbb{0} \tag{14.1.1946}$$

The two matrices W and Ψ defined above are the parameters of the FA model, generally denoted by $\theta=\{W,\Psi\}$.

As a linear combination of the two Gaussian vectors $\vec{z}$ and $\vec{\varepsilon}$ then:

$$\vec{x}=W\vec{z}+\vec{\varepsilon} \tag{14.1.1947}$$

also has a Gaussian distribution. We first find its mean $\mathrm{E}(\vec{x})$:

$$\Sigma_x=\vec{\mu}_x=\mathrm{E}(\vec{x})=\mathrm{E}(W\vec{z}+\vec{\varepsilon})=W\mathrm{E}(\vec{z})+\mathrm{E}(\vec{\varepsilon})=\vec{0} \tag{14.1.1948}$$

and variance-covariance matrix Σ_x (see page 720):

$$\begin{aligned}\Sigma_x&=\mathrm{cov}(\vec{x})=\mathrm{E}\left[(\vec{x}-\vec{\mu}_x)(\vec{x}-\vec{\mu}_x)^T\right]=\mathrm{E}(\vec{x}\vec{x}^T)=\mathrm{E}\left[(W\vec{z}+\vec{\varepsilon})(W\vec{z}+\vec{\varepsilon})^T\right]\\&=W\mathrm{E}(\vec{z}\vec{z}^T)W^T-\mathrm{E}(\vec{\varepsilon}\vec{z}^T)W^T-W\mathrm{E}(\vec{z}\vec{\varepsilon}^T)+\mathrm{E}(\vec{\varepsilon}\vec{\varepsilon}^T)\\&=WW^T+\Psi\end{aligned} \tag{14.1.1949}$$

Remark

That last relation is often denoted in textbooks as:

$$\mathrm{cov}(\vec{x})=LL^T+\Psi \tag{14.1.1950}$$

with therefore:

$$\begin{aligned}\mathrm{V}(x_i)&=l_{i1}^2+\ldots+l_{id}^2+\psi_i\\\mathrm{cov}(x_i,x_k)&=l_{i1}l_{k1}+\ldots+l_{id}l_{kd}\end{aligned} \tag{14.1.1951}$$

The portion of the variance of the ith variable contributed by the m common factors is named the "ith communality". That portion of $\mathrm{V}(x_i)=\sigma_{ii}$ due to the specific factor is often named the "uniqueness", or "specific variance". Denoting the ith communality by h_i^2, we see that:

$$\mathrm{V}(x_i)=\sigma_{ii}=h_i^2+\psi_i \tag{14.1.1952}$$

where:

$$h_i^2=l_{i1}^2+l_{i2}^2+\ldots+l_{id}^2 \tag{14.1.1953}$$

✍**Example:**

Let us observe what are the implications of the aforementioned assumptions. To do this, let's write the general form of each of the linear relations from our previous example:

$$y_i=\beta_{i0}+\beta_{i1}f_1+\beta_{i2}f_2+\varepsilon_i \tag{14.1.1954}$$

If we calculate the variance, then we have using its properties:

$$\begin{aligned}&\mathrm{V}(y_i)=\mathrm{V}(\beta_{i0})+\mathrm{V}(\beta_{11}f_1)+\mathrm{V}(\beta_{i2}f_2)+\mathrm{V}(\varepsilon_i)\\&=\mathrm{V}(\beta_{i1}f_1)+\mathrm{V}(\beta_{i2}f_2)+\sigma_i^2=\beta_{i1}^2\mathrm{V}(f_1)+\beta_{i2}^2V(f_2)+\sigma_{\varepsilon_i}^2\\&=\underbrace{\beta_{i1}^2+\beta_{i2}^2}_{\text{communality}}+\sigma_{\varepsilon_i}^2\end{aligned} \tag{14.1.1955}$$

Hence in the general case:

$$\mathrm{V}\left(y_i\right)=\underbrace{\sum_{j=1}\beta_{i1}^2}_{\text{communality}}+\sigma_{\varepsilon_i}^2 \tag{14.1.1956}$$

The "communality" is then the part of the variance of the variable which is explained purely by the factors of the explanatory variables of the model. The greater the communality, the better the predictive power of the factors !!

Obviously, if the factors are perfect predictors, we have the specific variance $\sigma_{\varepsilon_i}^2$ which would then be zero.

And then get the Normal distribution of $\vec{x}$ given $\vec{\theta}$:

$$P(\vec{x}|\vec{\theta})=\mathcal{N}(\vec{\mu}_x,\Sigma_x)=\mathcal{N}(\vec{0},WW^T+\Psi) \tag{14.1.1957}$$

We can further find the joint distribution $P(\vec{z},\vec{x})$, also a Gaussian, which has a zero mean:

$$\vec{\mu}=\mathrm{E}\left(\begin{bmatrix}\vec{z}\\\vec{x}\end{bmatrix}\right)=\begin{bmatrix}\vec{\mu}_z\\\vec{\mu}_x\end{bmatrix}=\begin{bmatrix}\vec{0}\\\vec{0}\end{bmatrix} \tag{14.1.1958}$$

and a covariance matrix Σ composed of four submatrices:

$$\Sigma=\begin{bmatrix}\Sigma_{zz} & \Sigma_{zx}\\ \Sigma_{xz} & \Sigma_{xx}\end{bmatrix} \tag{14.1.1959}$$

where he have immediately:

$$\begin{aligned}
\Sigma_{zz}&=\mathbb{1}\\
\Sigma_{xx}&=WW^T+\Psi\\
\Sigma_{zx}&=\Sigma_{zx}^T=\mathrm{E}(\vec{z}-\vec{\mu}_z)(\vec{x}-\vec{\mu}_x)^T=\mathrm{E}(\vec{z}\vec{x}^T)=E[\vec{z}(W\vec{z}+\vec{\varepsilon})^T]\\
&=\mathrm{E}(\vec{z}\vec{z}^T)W^T+\mathrm{E}(\vec{z}\vec{\varepsilon}^T)=\mathbb{1}W^T+\mathbb{O}=W^T
\end{aligned} \tag{14.1.1960}$$

Now the Normal distribution $P(\vec{z},\vec{x}|\vec{\theta})$ given model parameter $\vec{\theta}=\{W,\Psi\}$ can be expressed as:

$$\begin{aligned}
P(\vec{x},\vec{z}|\vec{\theta})=P\left(\begin{bmatrix}\vec{z}\\\vec{x}\end{bmatrix}\right)&=\mathcal{N}(\vec{\mu},\Sigma)=\mathcal{N}\left(\begin{bmatrix}\vec{0}\\\vec{0}\end{bmatrix},\begin{bmatrix}\Sigma_{zz} & \Sigma_{zx}\\ \Sigma_{xz} & \Sigma_{xx}\end{bmatrix}\right)\\
&=\mathcal{N}\left(\begin{bmatrix}\vec{0}\\\vec{0}\end{bmatrix},\begin{bmatrix}\mathbb{1} & W^T\\ W & WW^T+\Psi\end{bmatrix}\right)
\end{aligned} \tag{14.1.1961}$$

✍**Example:**

Let us now calculate the covariance of a couple of variables y_i,y_j from our previous example by first writing:

$$\begin{aligned}
y_i&=\beta_{i0}+\beta_{i1}f_1+\beta_{i2}f_2+\varepsilon_i\\
y_j&=\beta_{j0}+\beta_{j1}f_1+\beta_{j2}f_2+\varepsilon_j
\end{aligned} \tag{14.1.1962}$$

We then have:

$$\text{cov}\left(y_i,y_j\right)=\text{cov}\left(\beta_{i0}+\beta_{i1}f_1+\beta_{i2}f_2+\varepsilon_i,\beta_{j0}+\beta_{j1}f_1+\beta_{j2}f_2+\varepsilon_j\right) \quad (14.1.1963)$$

From the definition of covariance we know that the constants will vanish. So we can focus on:

$$\text{cov}\left(y_i,y_j\right)=\text{cov}\left(\beta_{i1}f_1+\beta_{i2}f_2+\varepsilon_i,\beta_{j1}f_1+\beta_{j2}f_2+\varepsilon_j\right) \quad (14.1.1964)$$

We saw just previously that in the simple case where $i=j$, we have in our particular case:

$$\text{cov}\left(y_i,y_j\right)_{i=j}=V\left(y_i\right)=\beta_{i1}^2+\beta_{i2}^2+\sigma_{\varepsilon_i}^2 \quad (14.1.1965)$$

Now let's deal with the case where $i\neq j$. We then have:

$$\begin{aligned}
&\text{cov}\left(y_i,y_j\right)=\text{cov}\left(\beta_{i1}f_1+\beta_{i2}f_2+\varepsilon_i,\beta_{j1}f_1+\beta_{j2}f_2+\varepsilon_j\right)\\
&=\beta_{j1}\,\text{cov}\left(\beta_{i1}f_1+\beta_{i2}f_2+\varepsilon_i,f_1\right)+\beta_{j2}\,\text{cov}\left(\beta_{i1}f_1+\beta_{i2}f_2+\varepsilon_i,f_2\right)\\
&+\text{cov}\left(\beta_{i1}f_1+\beta_{i2}f_2+\varepsilon_i,\varepsilon_j\right)\\
&=\beta_{j1}\left(\beta_{i1}\,\text{cov}\left(f_1,f_1\right)+\beta_{i2}\,\text{cov}\left(f_2,f_1\right)+\text{cov}\left(\varepsilon_i,f_1\right)\right)\\
&+\beta_{j2}\left(\beta_{i1}\,\text{cov}\left(f_1,f_2\right)+\beta_{i2}\,\text{cov}\left(f_2,f_2\right)+\text{cov}\left(\varepsilon_i,f_2\right)\right)\\
&+1\cdot\left(\left(\beta_{i1}\,\text{cov}\left(f_1,\varepsilon_j\right)+\beta_{i2}\,\text{cov}\left(f_2,\varepsilon_j\right)+\text{cov}\left(\varepsilon_i,\varepsilon_j\right)\right)\right)\\
&=\beta_{j1}\left(\beta_{i1}\text{V}\left(f_1\right)+\beta_{i2}\,\text{cov}\left(f_2,f_1\right)+\text{cov}\left(\varepsilon_i,f_1\right)\right)\\
&+\beta_{j2}\left(\beta_{i1}\,\text{cov}\left(f_1,f_2\right)+\beta_{i2}\text{V}\left(f_2\right)+\text{cov}\left(\varepsilon_i,f_2\right)\right)\\
&+1\cdot\left(\left(\beta_{i1}\,\text{cov}\left(f_1,\varepsilon_j\right)+\beta_{i2}\,\text{cov}\left(f_2,\varepsilon_j\right)+\text{cov}\left(\varepsilon_i,\varepsilon_j\right)\right)\right)\\
&=\beta_{j1}\left(\beta_{i1}+\beta_{i2}\,\text{cov}\left(f_2,f_1\right)+\text{cov}\left(\varepsilon_i,f_1\right)\right)\\
&+\beta_{j2}\left(\beta_{i1}\,\text{cov}\left(f_1,f_2\right)+\beta_{i2}+\text{cov}\left(\varepsilon_i,f_2\right)\right)\\
&+1\cdot\left(\left(\beta_{i1}\,\text{cov}\left(f_1,\varepsilon_j\right)+\beta_{i2}\,\text{cov}\left(f_2,\varepsilon_j\right)+\text{cov}\left(\varepsilon_i,\varepsilon_j\right)\right)\right)
\end{aligned} \quad (14.1.1966)$$

Since the factors are assumed to be independent random variables by assumption in the model, their covariance is zero. It is the same for the terms of errors. It remains then:

$$\begin{aligned}
&\text{cov}\left(y_i,y_j\right)=\beta_{j1}\left(\beta_{11}+\text{cov}\left(\varepsilon_i,f_1\right)\right)+\beta_{j2}\left(\beta_{i2}+\text{cov}\left(\varepsilon_i,f_2\right)\right)\\
&+1\cdot\left(\left(\beta_{i1}\,\text{cov}\left(f_1,\varepsilon_j\right)+\beta_{i2}\,\text{cov}\left(f_2,\varepsilon_j\right)\right)\right)
\end{aligned} \quad (14.1.1967)$$

We will make the assumption that the errors are random variables independent of the factors. So at the level of the covariances for the moment all the assumptions are then:

$$\begin{aligned}
\text{cov}\left(f_i,f_j\right)_{i\neq j}&=0\\
\text{cov}\left(\varepsilon_i,f_j\right)_{i=j}&=0\\
\text{cov}\left(\varepsilon_i,\varepsilon_j\right)_{i\neq j}&=0
\end{aligned} \quad (14.1.1968)$$

So their covariance also becomes zero and it remains:

$$\text{cov}\left(y_i,y_j\right)_{i\neq j}=\beta_{j1}\beta_{i1}+\beta_{j2}\beta_{i2} \quad (14.1.1969)$$

So in the general case where $i\neq j$ we finally have:

$$\text{cov}\left(y_i,y_j\right)_{i\neq j}=\sum_{k=1}^{m}\beta_{jk}\beta_{ik} \quad (14.1.1970)$$

We can then build the "theoretical matrix of the variances-covariances of the factorial model" in the particular case which serves as an example, we have:

$$\begin{aligned}C&=\begin{bmatrix}\beta_{11}^2+\beta_{12}^2+\sigma_{\varepsilon_1}^2 & \beta_{21}\beta_{11}+\beta_{22}\beta_{12} & \beta_{31}\beta_{11}+\beta_{32}\beta_{21}\\ \beta_{11}\beta_{21}+\beta_{12}\beta_{22} & \beta_{11}^2+\beta_{22}^2+\sigma_{\varepsilon_2}^2 & \beta_{21}\beta_{31}+\beta_{22}\beta_{32}\\ \beta_{11}\beta_{31}+\beta_{12}\beta_{32} & \beta_{21}\beta_{31}+\beta_{22}\beta_{32} & \beta_{31}^2+\beta_{32}^2+\sigma_{\varepsilon_3}^2\end{bmatrix}\\ &=L\cdot L^T+\begin{bmatrix}\sigma_{\varepsilon_1}^2 & 0 & 0\\ 0 & \sigma_{\varepsilon_2}^2 & 0\\ 0 & 0 & \sigma_{\varepsilon_3}^2\end{bmatrix}=L\cdot L^T+\Psi\end{aligned}\tag{14.1.1971}$$

Notice that if we disregard the error term, it comes in our particular case:

$$\begin{aligned}C&=\begin{bmatrix}\beta_{11}^2+\beta_{12}^2 & \beta_{21}\beta_{11}+\beta_{22}\beta_{12} & \beta_{31}\beta_{11}+\beta_{32}\beta_{21}\\ \beta_{11}\beta_{21}+\beta_{12}\beta_{22} & \beta_{11}^2+\beta_{22}^2 & \beta_{21}\beta_{31}+\beta_{22}\beta_{32}\\ \beta_{11}\beta_{31}+\beta_{12}\beta_{32} & \beta_{21}\beta_{31}+\beta_{22}\beta_{32} & \beta_{31}^2+\beta_{32}^2\end{bmatrix}\\ &=\begin{bmatrix}\beta_{11} & \beta_{12} & \beta_{13}\\ \beta_{21} & \beta_{22} & 0\\ \beta_{31} & \beta_{32} & 0\end{bmatrix}\begin{bmatrix}\beta_{11} & \beta_{12} & \beta_{13}\\ \beta_{21} & \beta_{22} & 0\\ \beta_{31} & \beta_{32} & 0\end{bmatrix}^T\\ &=\begin{bmatrix}\beta_{11} & \beta_{12} & \beta_{13}\\ \beta_{21} & \beta_{22} & 0\\ \beta_{31} & \beta_{32} & 0\end{bmatrix}\begin{bmatrix}\beta_{11} & \beta_{21} & \beta_{31}\\ \beta_{12} & \beta_{22} & \beta_{32}\\ \beta_{13} & 0 & 0\end{bmatrix}\end{aligned}\tag{14.1.1972}$$

In practice, we can easily calculate the "experimental variance-covariance matrix" of the measured data. We usually note it in the case of three variables to be explained:

$$\hat{C}=\begin{bmatrix}S_1^2 & S_{12} & S_{13}\\ S_{21} & S_2^2 & S_{23}\\ S_{31} & S_{32} & S_3^2\end{bmatrix}=L\cdot L^T+\Psi\tag{14.1.1973}$$

Then we just have to put the two matrices in correspondence (well normally it is customary to denote by b_{ij} the estimated values of the β_{ij} but we will suppose that the reader will be able to easily differentiate the moment when we work with the estimators or the actual theoretical values):

$$\begin{bmatrix}S_1^2 & S_{12} & S_{13}\\ S_{21} & S_2^2 & S_{23}\\ S_{31} & S_{32} & S_3^2\end{bmatrix}=\begin{bmatrix}\beta_{11}^2+\beta_{12}^2+\sigma_{\varepsilon_1}^2 & \beta_{21}\beta_{11}+\beta_{22}\beta_{12} & \beta_{31}\beta_{11}+\beta_{32}\beta_{21}\\ \beta_{11}\beta_{21}+\beta_{12}\beta_{22} & \beta_{11}^2+\beta_{22}^2+\sigma_{\varepsilon_2}^2 & \beta_{21}\beta_{31}+\beta_{22}\beta_{32}\\ \beta_{11}\beta_{31}+\beta_{12}\beta_{32} & \beta_{21}\beta_{31}+\beta_{22}\beta_{32} & \beta_{31}^2+\beta_{32}^2+\sigma_{\varepsilon_3}^2\end{bmatrix}\tag{14.1.1974}$$

If we go back to our original example, then we have:

$$\begin{bmatrix}1 & -0.051 & 0.804\\ -0.051 & 1 & 0.981\\ 0.804 & 0.981 & 1\end{bmatrix}=\begin{bmatrix}\beta_{11}^2+\beta_{12}^2+\sigma_{\varepsilon_1}^2 & \beta_{21}\beta_{11}+\beta_{22}\beta_{12} & \beta_{31}\beta_{11}+\beta_{32}\beta_{21}\\ \beta_{11}\beta_{21}+\beta_{12}\beta_{22} & \beta_{11}^2+\beta_{22}^2+\sigma_{\varepsilon_2}^2 & \beta_{21}\beta_{31}+\beta_{22}\beta_{32}\\ \beta_{11}\beta_{31}+\beta_{12}\beta_{32} & \beta_{21}\beta_{31}+\beta_{22}\beta_{32} & \beta_{31}^2+\beta_{32}^2+\sigma_{\varepsilon_3}^2\end{bmatrix}\tag{14.1.1975}$$

We will then see a little further that we can deduce all the saturations (coefficients of the model) from this positive definite matrix!

Now we have two typical ways to determine the parameters of interest! First an iterative way based on the Expectation-Maximization (EM) algorithm that gives us the possibility to determine W and Ψ and another one, named "Principal Factor Analysis Extraction", based on the eigenvalue spectral decomposition (same idea as Principal Component Analysis) to determine only W.

14.1.20.6.4.1 Expectation-Maximization (EM) algorithm solution

Based on this joint Normal distribution, we can further find the following marginal distributions (see properties of Normal distributions at page 723):

- First:

$$P(\vec{x}|\vec{z},\theta)=\mathcal{N}\left(\vec{\mu}_{x|z},\Sigma_{x|z}\right) \tag{14.1.1976}$$

  with:

$$\begin{cases}\vec{\mu}_{x|z}=\vec{\mu}_x+\Sigma_{xz}\Sigma_{zz}^{-1}\left(\vec{z}-\vec{\mu}_z\right)=W\vec{z}\\ \Sigma_{x|z}=\Sigma_{xx}-\Sigma_{xz}^{-1}z_{zx}=WW^T+\Psi-WW^T=\Psi\end{cases} \tag{14.1.1977}$$

- Secondly:

$$P(\vec{z}|\vec{x},\theta)=\mathcal{N}\left(\vec{\mu}_{z|x},\Sigma_{z|x}\right) \tag{14.1.1978}$$

  with:

$$\begin{cases}\vec{\mu}_{z|x}=\vec{\mu}_z+\Sigma_{zx}\Sigma_{xx}^{-1}\left(\vec{x}-\vec{\mu}_x\right)=W^T\left(WW^T+\Psi\right)^{-1}\vec{x}=B\vec{x}\\ \Sigma_{z|x}=\Sigma_{zz}-\Sigma_{zx}\Sigma_{xx}^{-1}\Sigma_{xz}=\mathbb{1}-W^T\left(WW^T+\Psi\right)^{-1}W=\mathbb{1}-BW\end{cases} \tag{14.1.1979}$$

  where we have defined:

$$B=W^T(WW^T+\Psi)^{-1} \tag{14.1.1980}$$

Note that while $P(\vec{z})=\mathcal{N}(\vec{0},\mathbb{1})$ has zero mean and diagonal covariance, the conditional distribution $P(\vec{z}|\vec{x},\vec{\theta})=\mathcal{N}(\vec{\mu}_{z|x},\Sigma_{z|x})$ has non-zero mean $\vec{\mu}_{z|x}$ and non-diagonal covariance $\Sigma_{z|x}$!

The computational complexity for the inversion of the $N\times N$ matrix $WW^T+\Psi$ is $\mathcal{O}(N^3)$. However, by applying the Woodbury matrix identity (see section Linear Algebra page 2035):

$$(\Psi+WW^T)^{-1}=\Psi^{-1}-\Psi^{-1}W(\mathbb{1}+W^T\Psi^{-1}W)^{-1}W^T\Psi^{-1} \tag{14.1.1981}$$

where Ψ as a diagonal matrix can be easily inverted, and $\mathbb{1}+W^T\Psi^{-1}W$ is an $d\times d$ matrix that can be inverted with complexity $\mathcal{O}(d^3)\ll\mathcal{O}(N^3)$.

The model parameters in $\theta=\{W,\Psi\}$ can be estimated based on an observed dataset X by the EM algorithm (see page 6697), an iterative process of the following two steps:

- The E-step:

  Find the expectation of the log likelihood function of the model parameters in θ, to be maximized in the following M-step.

  Assuming all samples in the observed dataset $X=[\vec{x}_1,\ldots,\vec{x}_N]$ are independent and identically distributed (i.i.d.), we can find the "complete data likelihood" function of θ (using the Bayes' relation established in the section of Probabilities page 498):

$$L(\theta|\vec{X},\vec{z})=P(\vec{X},\vec{z}|\theta)=\prod_{n=1}^{N}P\left(\vec{x}_n,\vec{z}|\theta\right)=\prod_{n=1}^{N}P\left(\vec{x}_n|\vec{z},\theta\right)P(\vec{z}|\theta) \tag{14.1.1982}$$

  and the log-likelihood:

$$\ln\left(L(\theta|X,\vec{z})\right)=\ln\left(P(\vec{X},\vec{z}|\theta)\right)=\sum_{n=1}^{N}\left(\ln\left(P(\vec{x}_n|\vec{z},\theta)\right)+\ln\left(P(\vec{z}|\theta)\right)\right) \tag{14.1.1983}$$

  The second term can be dropped as $P(\vec{z}|\theta)=\mathcal{N}(\vec{0},\mathbb{1})$ is independent of the model parameters in θ and therefore irrelevant to the maximization of the log likelihood with respect to θ.

  The expectation of the log-likelihood function above (aka "expected complete data log-likelihood"), denoted by Q, with respect to the latent variable $\vec{z}$ is:

$$\begin{aligned}Q&=\mathrm{E}_{\vec{z}|\vec{x}}\left[\ln\left(L(\theta|\vec{X},\vec{z})\right)\right]=\sum_{n=1}^{N}\mathrm{E}_{z|x_n}\left[\ln P\left(\vec{x}_n|\vec{z},\theta\right)\right]\\&=\sum_{n=1}^{N}\mathrm{E}_{z|x_n}\left(\ln\left[\frac{1}{(2\pi)^{D/2}\left|\Sigma_{x|z}\right|^{1/2}}\exp\left(-\frac{1}{2}\left(\vec{x}_n-\vec{\mu}_{x|z}\right)^T\Sigma_{x|z}^{-1}\left(\vec{x}_n-\vec{\mu}_{x|z}\right)\right)\right]\right)\\&=\sum_{n=1}^{N}\mathrm{E}_{z|x_n}\left[-\frac{D}{2}\ln(2\pi)-\frac{1}{2}\ln|\Psi|-\frac{1}{2}\left(\vec{x}_n-W\vec{z}\right)^T\Psi^{-1}\left(\vec{x}_n-W\vec{z}\right)\right]\\&=-\frac{DN}{2}\ln(2\pi)-\frac{N}{2}\ln|\Psi|-\frac{1}{2}\sum_{n=1}^{N}\mathrm{E}_{z|x_n}\left(\left(\vec{x}_n-W\vec{z}\right)^T\Psi^{-1}\left(\vec{x}_n-W\vec{z}\right)\right)\\&=C-\frac{N}{2}\ln|\Psi|-\frac{1}{2}\sum_{n=1}^{N}\mathrm{E}_{z|x_n}\left(\vec{x}_n^T\Psi^{-1}\vec{x}_n-2\vec{x}_n^T\Psi^{-1}W\vec{z}+\vec{z}^TW^T\Psi^{-1}W\vec{z}\right)\end{aligned} \tag{14.1.1984}$$

  Note that Ψ is diagonal and therefore also symmetric, so is its own inverse Ψ^{-1}, we have:

$$\vec{x}_n^T\Psi^{-1}(W\vec{z})=(W\vec{z})^T\Psi^{-1}\vec{x}_n \tag{14.1.1985}$$

  Also, note that $C=-DN\ln(2\pi)/2$ is a constant independent of $\vec{\theta}$ and is therefore dropped.

- The M-step:

  Find the optimal model parameters in $\theta=\{W,\Psi\}$ that maximize the expectation of the log-likelihood Q obtained in the E-step.

  This is done by setting to zero the derivative of Q with respective each of the two parameters in $\theta=\{W,\Psi\}$:

- The first derivative is:

$$\begin{aligned}\frac{\partial Q}{\partial W}&=\frac{\partial}{\partial W}\sum_{n=1}^{N}\mathrm{E}_{z|x_n}\left(-\vec{x}_n^T\Psi^{-1}\vec{x}_n+2\vec{x}_n^T\Psi^{-1}W\vec{z}-\vec{z}^TW^T\Psi^{-1}W\vec{z}\right)\\&=\sum_{n=1}^{N}\mathrm{E}_{z|x_n}\left[\frac{\partial}{\partial W}\left(2\vec{x}_n^T\Psi^{-1}W\vec{z}-\vec{z}^TW^T\Psi^{-1}W\vec{z}\right)\right]\\&=\sum_{n=1}^{N}\mathrm{E}_{z|x_n}\left(2\Psi^{-1}\vec{x}_n\vec{z}^T-2\Psi^{-1}W\vec{z}\vec{z}^T\right)\\&=2\Psi^{-1}\sum_{n=1}^{N}\left(\vec{x}_n\mathrm{E}_{z|x_n}\left(\vec{z}^T\right)-W\mathrm{E}_{z|x_n}\left(\vec{z}\vec{z}^T\right)\right)=0\end{aligned} \tag{14.1.1986}$$

Solving for W we get:

$$W=\left(\sum_{n=1}^{N}\vec{x}_n\mathrm{E}_{z|x_n}(\vec{z}^T)\right)\left(\sum_{n=1}^{N}\mathrm{E}_{z|x_n}(\vec{z}\vec{z}^T)\right)^{-1} \tag{14.1.1987}$$

where $\mathrm{E}_{z|x_n}(\vec{z})$ and $\mathrm{E}_{z|x_n}(\vec{z}\vec{z}^T)$ can be found in (see previously):

$$\begin{cases}\vec{\mu}_{z|x}=\vec{\mu}_z+\Sigma_{zx}\Sigma_{xx}^{-1}\left(\vec{x}-\vec{\mu}_x\right)=W^T\left(WW^T+\Psi\right)^{-1}\vec{x}=B\vec{x}\\ \Sigma_{z|x}=\Sigma_{zz}-\Sigma_{zx}\Sigma_{xx}^{-1}\Sigma_{xz}=\mathbb{1}-W^T\left(WW^T+\Psi\right)^{-1}W=\mathbb{1}-BW\end{cases} \tag{14.1.1988}$$

Therefore:

$$\begin{cases}\mathrm{E}_{z|x_n}(\vec{z}) & =\vec{\mu}_{z|x_n}=B\vec{x}_n\\ \mathrm{E}_{z|x_n}\left(\vec{z}\vec{z}^T\right) & =\Sigma_{z|x_n}+\vec{\mu}_{z|x_n}\vec{\mu}_{z|x_n}^T=\mathbb{1}-BW+B\vec{x}_n\vec{x}_n^TB^T\end{cases} \tag{14.1.1989}$$

The second equation is due to the fact that $\Sigma_z=\mathrm{E}(\vec{z}\vec{z}^T)-\vec{\mu}_z\vec{\mu}_z^T$. Here $\mathrm{E}_{z|x_n}(\vec{z})$ can be considered as the estimation of the $\vec{z}$, while $\mathrm{E}_{z|x_n}(\vec{z}\vec{z}^T)$ the uncertainty of the estimation.

- The second derivative (we use in this development derivatives of matrices as seen in the section of Linear Algebra page 1976):

$$\begin{aligned}\frac{\partial Q}{\partial\Psi^{-1}}&=\frac{\partial}{\partial\Psi^{-1}}\left[N\ln|\Psi|+\sum_{n=1}^{N}\mathrm{E}_{z|x_n}\left(\vec{x}_n^T\Psi^{-1}\vec{x}_n-2\vec{x}_n^T\Psi^{-1}W\vec{z}+\vec{z}^TW^T\Psi^{-1}W\vec{z}\right)\right]\\&=-N\Psi+\sum_{n=1}^{N}\mathrm{E}_{z|x_n}\left(\vec{x}_n\vec{x}_n^T-2\vec{x}_n\vec{z}^TW^T+W\left(\vec{z}\vec{z}^T\right)W^T\right)\\&=-N\Psi+\sum_{n=1}^{N}\vec{x}_n\vec{x}_n^T-2\sum_{n=1}^{N}\vec{x}_n\mathrm{E}_{z|x_n}(\vec{z})^TW^T+W\sum_{n=1}^{N}\mathrm{E}_{z|x}\left(\vec{z}\vec{z}^T\right)W^T=\mathbb{0}\end{aligned} \tag{14.1.1990}$$

Solving for Ψ we get:

$$\Psi=\frac{1}{N}\operatorname{diag}\left(\sum_{n=1}^{N}\vec{x}_n\vec{x}_n^T-2\sum_{n=1}^{N}\vec{x}_n\mathrm{E}_{z|x_n}(\vec{z})^TW^T+W\sum_{n=1}^{N}\mathrm{E}_{z|x_n}\left(\vec{z}\vec{z}^T\right)W^T\right) \tag{14.1.1991}$$

where diag represents the operation that sets all off-diagonal elements of the right-hand side to zero, so that the resulting matrix Ψ is guaranteed to be diagonal as required. We further replace W in front of the last term above by that in:

$$W=\left(\sum_{n=1}^{N}\vec{x}_n\mathrm{E}_{z|x_n}(\vec{z}^T)\right)\left(\sum_{n=1}^{N}\mathrm{E}_{z|x_n}(\vec{z}\vec{z}^T)\right)^{-1} \tag{14.1.1992}$$

and get

$$\Psi=\frac{1}{N}\mathrm{diag}\left(\sum_{n=1}^{N}\vec{x}_n\vec{x}_n^T-\sum_{n=1}^{N}\vec{x}_n\mathrm{E}_{z|x_n}(\vec{z})^TW^T\right) \tag{14.1.1993}$$

In summary, here is the list of steps in the EM algorithm for FA:

1. Initialize parameters $\theta_{\text{old}}=\{W_{\text{old}},\Psi_{\text{old}}\}$
2. E-step: Based on θ_{old}, find $\vec{\mu}_{z|x_n}$ and $\Sigma_{z|x_n}$ in:

   $$\begin{cases}\vec{\mu}_{z|x}=\vec{\mu}_z+\Sigma_{zx}\Sigma_{xx}^{-1}(\vec{x}-\vec{\mu}_x)=W^T\left(WW^T+\Psi\right)^{-1}\vec{x}=B\vec{x}\\ \Sigma_{z|x}=\Sigma_{zz}-\Sigma_{zx}\Sigma_{xx}^{-1}\Sigma_{xz}=\mathbb{1}-W^T\left(WW^T+\Psi\right)^{-1}W=\mathbb{1}-BW\end{cases} \tag{14.1.1994}$$

   and then $\mathrm{E}_{z|x}(\vec{z})$ and $\mathrm{E}_{z|x}(\vec{z}\vec{z}^T)$ in:

   $$\begin{cases}\mathrm{E}_{z|x_n}(\vec{z}) & =\vec{\mu}_{z|x_n}=B\vec{x}_n\\ \mathrm{E}_{z|x_n}\left(\vec{z}\vec{z}^T\right) & =\Sigma_{z|x_n}+\vec{\mu}_{z|x_n}\vec{\mu}_{z|x_n}^T=\mathbb{1}-BW+B\vec{x}_n\vec{x}_n^TB^T\end{cases} \tag{14.1.1995}$$

3. M-step: Find $\theta_{\text{new}}=\{W_{\text{new}},\Psi_{\text{new}}\}$ by evaluating:

   $$W=\left(\sum_{n=1}^{N}\vec{x}_n\mathrm{E}_{z|x_n}(\vec{z}^T)\right)\left(\sum_{n=1}^{N}\mathrm{E}_{z|x_n}(\vec{z}\vec{z}^T)\right)^{-1} \tag{14.1.1996}$$

   and:

   $$\Psi=\frac{1}{N}\mathrm{diag}\left(\sum_{n=1}^{N}\vec{x}_n\vec{x}_n^T-\sum_{n=1}^{N}\vec{x}_n\mathrm{E}_{z|x_n}(\vec{z})^TW^T\right) \tag{14.1.1997}$$

4. Terminate if convergence criterion is satisfied, otherwise replace θ_{old} by θ_{new} and return to step 2.

14.1.20.6.4.2 Principal Factor Analysis Extraction solution

Now, if we center and reduce the starting vectors containing the measured values, the variance-covariance matrix becomes the correlation matrix as we have proved it during our study of Principal Component Analysis (PCA)! This is why the PCA is the basis of the Factorial Analysis (exploratory) and we then say that we are doing a "Principal Factor Analysis Extraction" (PFAE).

To understand the idea let us recall that we can write the Factor Analysis model in the following matrix form:

$$\bar{Y}=\bar{\beta}_0+L\bar{f}+\bar{\varepsilon} \tag{14.1.1998}$$

If the data are centered and reduced, we will write:

$$\bar{Z}=L\bar{f}+\bar{\varepsilon} \tag{14.1.1999}$$

Now let's play cleverly with an orthogonal matrix S:

$$\bar{Z}=L\bar{f}+\bar{\varepsilon}=L\left(S^TS\right)\bar{f}+\bar{\varepsilon}=\left(LS^T\right)\left(S\bar{f}\right)+\bar{\varepsilon}=L^*\bar{f}^*+\bar{\varepsilon} \tag{14.1.2000}$$

It is thus an alternative notation for the initial model but as there is an infinity of orthogonal matrices, there is an infinity of alternative models.

We denoted during our study of the Principal Component Analysis the matrix of the eigenvalues arranged in decreasing order in the diagonal by Λ and by S the matrix of the eigenvectors. This then gives us using the results of our study of the Principal Component Analysis and from what has just been seen:

$$R=L\cdot L^T=S\Lambda S^T=(S\sqrt{\Lambda})(\sqrt{\Lambda}S^T) \tag{14.1.2001}$$

Using now the property of the transpose seen during our study of Linear Algebra, we have:

$$\sqrt{\Lambda}S^T=(S\sqrt{\Lambda})^T \tag{14.1.2002}$$

Therefore:

$$S\Lambda S^T=(S\sqrt{\Lambda})(S\sqrt{\Lambda})^T \tag{14.1.2003}$$

The matrix:

$$L=S\sqrt{\Lambda} \tag{14.1.2004}$$

therefore does the trick and it is customary to name the components of the matrix L the "factor saturations" and it therefore contains the coefficients of the factors of the implicit linear model when the variables to be explained are centered and reduced. This is enough to give a qualitative idea of the weight of each of the factors.

Some statistical software return what is are named the "loadings" or "factor score" and which corresponds to the expression:

$$\ell=\left((S\sqrt{\Lambda})^T\right)^{-1} \tag{14.1.2005}$$

Whatever the chosen above solution, the factor model assumes that the $D(D+1)/2$ components of its covariance-variance matrix for $\vec{x}$ can be reproduced from the $D\cdot d$ factor loading (weights) w_{ij} and the D specific variances ψ_D. Because we have proved earlier above that:

$$\Sigma_x=WW^T+\Psi \tag{14.1.2006}$$

then when $D=d$ any covariance matrix Σ_x can be reproduced exactly as WW^T, so Ψ can be the zero matrix. It is when d is small relative to D that factor analysis is more useful. In this case, the factor model provides a simple explanation of the covariation in $\vec{x}$ with fewer parameters than $D(D+1)/2$ in Σ_x.

Unfortunately for the factor analyst it is not easy to choose the number of latent variables. If the number of common factors is not determined by a priori considerations, such as by theory or the work of other researchers, the choice of d can be based on the estimated eigenvalues of the principal component analysis.

Given observations $x_1, x_2, \ldots, x_D$ on d generally correlated variables, factor analysis seeks to answer the question: Does the factor model (linear in our case), with a small number of factors, adequately represent the data? In essence, we tackle this statistical model-building problem by trying to verify the covariance relations:

$$\begin{aligned} \text{V}(x_i) &= l_{i1}^2 + \ldots + l_{id}^2 + \psi_i \\ \text{cov}(x_i, x_k) &= l_{i1}l_{k1} + \ldots + l_{id}l_{kd} \\ \Sigma_{zx} &= L^T \end{aligned} \tag{14.1.2007}$$

If the off-diagonal elements of the estimator of Σ_x are small, the variables are not related, and a factor analysis will not provide any useful evidence. If Σ_x appears to deviate significantly from a diagonal matrix, then a factor model can be entertained, and the initial problem is one of estimating the factor loadings w_{ij} and specific variables ψ_i.

It should be also noticed that there is an infinity of solutions. To understand why consider the following model (we use here again the traditional notation in french textbooks):

$$\begin{aligned} y_1 &= \beta_{11}f_1 + \beta_{12}f_2 + \varepsilon_1 = 0.5f_1 + 0.5f_2 + \varepsilon_1 \\ y_2 &= \beta_{21}f_1 + \beta_{22}f_2 + \varepsilon_2 = 0.3f_1 + 0.3f_2 + \varepsilon_2 \\ y_3 &= \beta_{31}f_1 + \beta_{32}f_2 + \varepsilon_3 = 0.5f_1 - 0.5f_2 + \varepsilon_3 \end{aligned} \tag{14.1.2008}$$

The variance-covariance matrix of the above factorial model is then:

$$\text{cov}\,(y_i, y_j) = \begin{bmatrix} 0.5 + \sigma_{\varepsilon_1}^2 & 0.3 & 0 \\ 0.3 & 0.18 + \sigma_{\varepsilon_2}^2 & 0 \\ 0 & 0 & 0.5 + \sigma_{\varepsilon_3}^2 \end{bmatrix} \tag{14.1.2009}$$

But the following model:

$$\begin{aligned} y_1 &= \frac{\sqrt{2}}{2}f_1 + 0 \cdot f_2 + \varepsilon_1 \\ y_2 &= 0.3\sqrt{2}f_1 + 0 \cdot f_2 + \varepsilon_2 \\ y_3 &= 0 \cdot f_1 - \frac{\sqrt{2}}{2}f_2 + \varepsilon_3 \end{aligned} \tag{14.1.2010}$$

also has exactly the same variance-covariance matrix of the factor model as the reader can quickly verify and is therefore also a solution! The reader will be able to verify that the application:

$$\begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} \beta_{11} \\ \beta_{12} \end{bmatrix} = \begin{bmatrix} \beta_{11}\cos(\theta) - \beta_{12}\sin(\theta) \\ \beta_{11}\sin(\theta) + \beta_{12}\cos(\theta) \end{bmatrix} = \begin{bmatrix} \beta'_{11} \\ \beta'_{12} \end{bmatrix} \tag{14.1.2011}$$

for any angle always gives the same variance-covariance matrix (reason for which there is an infinity of possible models for the same theoretical matrix of variances-covariance of the factorial model). So, in the example above, we simply took an angle of $\theta = -\pi/4$ which gives for the first line of the model:

$$\begin{bmatrix} 0.5\cos\left(\frac{\pi}{4}\right)-0.5\sin\left(\frac{\pi}{4}\right) \\ 0.5\sin\left(\frac{\pi}{4}\right)+0.5\cos\left(\frac{\pi}{4}\right) \end{bmatrix} = \begin{bmatrix} 0.5\frac{\sqrt{2}}{2}-0.5\left(-\frac{\sqrt{2}}{2}\right) \\ 0.5\left(-\frac{\sqrt{2}}{2}\right)+0.5\frac{\sqrt{2}}{2} \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ 0 \end{bmatrix} \qquad (14.1.2012)$$

and so on line by line with each pair of coefficients and with the same rotation matrix.

Although factor analysis and principal component analysis are typically labelled as data-reduction techniques, there are significant differences between these two. The objective of principal component analysis is to reduce the number of variables to a few components such that each component forms a new variable and the retained components explain the maximum amount of variance in the data. The objective of factor analysis, on the other hand, is to search or identify the underlying factor(s) or latent constructs that can explain the intercorrelation among the variables.

14.1.20.6.5 T-distributed Stochastic Neighbor Embedding (T-SNE)

"T-distributed Stochastic Neighbor Embedding" (T-SNE) is a machine learning algorithm for visualization developed by Laurens van der Maaten and Geoffrey Hinton (betwen 12002 and 12008 according to holocene calendar). It is a non-linear dimensionality reduction technique well-suited for embedding high-dimensional data for visualization in a low-dimensional space of two or three dimensions. Specifically, it models each high-dimensional object by a two- or three-dimensional point in such a way that similar objects are modelled by nearby points and dissimilar objects are modelled by distant points with high probability.

Remark

It is very important that the reader keep in mind that is designed only for visualizing a dataset in a low (2 or 3) dimension space. We give it all the data we want to visualize all at once. It is not a general purpose dimensionality reduction tool!

The T-SNE algorithm comprises two main stages. First, T-SNE constructs a probability distribution over pairs of high-dimensional objects in such a way that similar objects have a high probability of being picked while dissimilar points have an extremely small probability of being picked. Second, T-SNE defines a similar probability distribution over the points in the low-dimensional map, and it minimizes the Kullback–Leibler divergence (see section Statistical Mechanics page 3542) between the two distributions with respect to the locations of the points in the map. Note that while the original algorithm uses the Euclidean distance between objects as the base of its similarity metric, this should be changed as appropriate.

T-SNE has been used for visualization in a wide range of applications, including computer security research, music analysis, cancer research, bioinformatics, and biomedical signal processing. It is often used to visualize high-level representations learned by an artificial neural network. Indeed, as the reader may already know, deep CNN (convolutional neural networks) are basically black boxes. There is no way to really interpret what's on deeper levels in the network. A common explanation is that deeper levels contain information about more complex objects. But that's not completely true, you can interpret it like that but data itself is just a high-dimensional noise for humans. But, with the help of T-SNE you can create maps to display which input data seams similar for the network.

While T-SNE plots often seem to display clusters, the visual clusters can be influenced strongly by the chosen parametrization and therefore a good understanding of the parameters for T-SNE is necessary. Such "clusters" can be shown to even appear in non-clustered data, and thus may be false findings! Interactive exploration may thus be necessary to choose parameters and validate results. It has been demonstrated that T-SNE is often able to recover well-separated clusters, and with special parameter choices, approximates a simple form of spectral clustering.

Given a set of N high-dimensional objects $\vec{x}_1,\ldots,\vec{x}_N$, T-SNE first computes probabilities p_{ij} that are proportional to the similarity of objects $\vec{x}_i$ and $\vec{x}_j$, defined as follows:

$$p_{j|i}=\frac{\exp(-\|\vec{x}_i-\vec{x}_j\|^2/2\sigma_i^2)}{\sum\limits_{k\neq i}\exp(-\|\vec{x}_i-\vec{x}_k\|^2/2\sigma_i^2)} \qquad (14.1.2013)$$

Note that we have obviously $\sum_j p_{j|i}=1$ for all i.

Remark

It seems that in general all SNE models are written in the form:

$$p_{j|i}=\frac{g(x_i-\vec{x}_j)}{\sum\limits_{k\neq i} g(\vec{x}_i-\vec{x}_k)} \tag{14.1.2014}$$

where g is a given function.

As Van der Maaten and Hinton explained: "The similarity of datapoint $\vec{x}_j$ to datapoint $\vec{x}_i$ is the conditional probability, $p_{j|i}$, that $\vec{x}_i$ would pick $\vec{x}_j$ as its neighbour if neighbours were picked in proportion to their probability density under a Gaussian centered at $\vec{x}_i$."

Now we define:

$$p_{ij}=\frac{p_{j|i}+p_{i|j}}{2N} \tag{14.1.2015}$$

and note that $p_{ij}=p_{ji}$, $p_{ii}=0$ and $\sum_{i,j} p_{ij}=1$.

The bandwidth of the Gaussian kernels σ_i is set in such a way that the perplexity of the conditional distribution equals a predefined perplexity using the bisection method to match a pre-specified perplexity value (Perp). The perplexity is $\text{Perp}(P_j)=2^{H(p_j)}$, where $H(P_j)=-\sum_j p_{i|j}\log(p_{i|j})$, and σ_j is selected so that $\text{Perp}(P_j)=\text{Perp}$. As a result, the bandwidth is adapted to the density of the data: smaller values of σ_i are used in denser parts of the data space.

T-SNE is actually very sensitive to the Perplexity value. Let us see the following value:

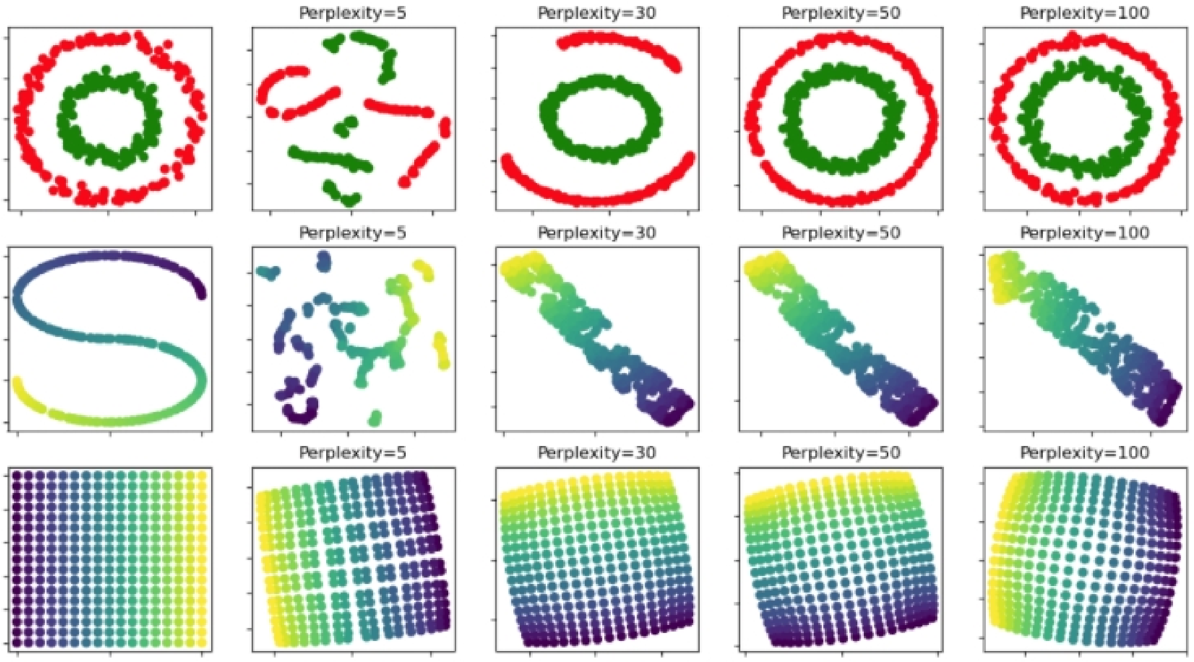


Figure 14.177: T-SNE perplexity sensitivity

Remark

There seems to be no standard way to choose the correct perplexity aside looking at the produced reduced dimension dataset and then assessing if it is meaningful. There are some general facts, eg. distances between clusters are mostly meaningless, small perplexity values encourage small clot-like structures but that's about it. A rule of thumb is to set the perplexity to 5% of the dataset size.

Since the Gaussian kernel uses the Euclidean distance $\|\vec{x}_i - \vec{x}_j\|$, it is affected by the curse of dimensionality, and in high dimensional data when distances lose the ability to discriminate, the p_{ij} become too similar (asymptotically, they would converge to a constant). It has been proposed to adjust the distances with a power transform, based on the intrinsic dimension of each point, to alleviate this.

In other words, in many of the dimensionality reduction techniques (e.g. T-SNE), we try to map the data points from a higher dimension to a lower dimension. While doing so a fundamental problem occurs, the distance between the points at a higher dimension can't be preserved at a lower dimension. This results in a overlapping or erroneous distance between the points at a lower dimension. This problem is named the "Crowding Problem". The T-distribution being heavier tailed helps in this regard.

T-SNE aims to learn a d-dimensional map $\vec{y}_1, \ldots, \vec{y}_N$ (with $\vec{y}_i \in \mathbb{R}^d$ that reflects the similarities $p_{i|j}$ as well as possible. To this end, it measures similarities q_{ij} between two points in the map $\vec{y}_i$ and $\vec{y}_j$, using a very similar approach. Specifically, $q_{j|i}$ is defined as:

$$q_{j|i}^{\text{TSNE}} = \frac{(1 + \|\vec{y}_i - \vec{y}_j\|^2)^{-1}}{\sum_k \sum_{k \neq i} (1 + \|\vec{y}_k - \vec{y}_i\|^2)^{-1}} \tag{14.1.2016}$$

Herein a heavy-tailed Student T-distribution (with one-degree of freedom, which is the same as a Cauchy distribution) is used to measure similarities between low-dimensional points in order to allow dissimilar objects to be modelled far apart in the map. Note that also in this case we set $q_{ii} = 0$.

The locations of the points $\vec{y}_i$ in the map are determined by minimizing the (non-symmetric) Kullback–Leibler divergence (see section Statistical Mechanics page 3542) of the distribution Q from the distribution P, that is:

$$C = \sum_i D_{\text{KL}}(P_i, Q_i) = \sum_j \sum_i p_{j|i} \log \left(\frac{p_{j|i}}{q_{j|i}} \right) \tag{14.1.2017}$$

in which P_i is the conditional probability distribution over all other datapoints given data-point $\vec{x}_i$, and Q_i represents the conditional probability distribution over all other map points given map point $\vec{y}_i$.

The minimization of the Kullback–Leibler divergence with respect to the points $\vec{y}_i$ is performed using gradient descent. The result of this optimization should be a map that reflects the similarities between the high-dimensional inputs well.

Let us see now how to derive the gradient of the KL divergence Loss function for the standard SNE and the T-SNE:

14.1.20.6.5.1 KL gradient of Stochastic Neighbour Embedding (SNE)

The definition of $q_{j|i}$ for SNE is different from that of the T-SNE (thanks to Federico Errica for providing the detailed developments):

$$q_{j|i}^{\text{SNE}}=\frac{e^{-\|y_i-y_j\|^2}}{\sum\limits_{k\neq l}e^{-\|y_k-y_l\|^2}}=\frac{E_{ij}}{\sum\limits_{k\neq l}E_{kl}}=\frac{E_{ij}}{Z_l} \tag{14.1.2018}$$

For what follows we will treat the one-dimensional case to simplify the notations (and we will also omit the SNE exponent).

Notice that $E_{ij}=E_{ji}$. The loss function is defined as:

$$\begin{aligned}C&=\sum_k\sum_{l\neq k}p_{l|k}\log\left(\frac{p_{l|k}}{q_{l|k}}\right)=\sum_{k,l\neq k}p_{l|k}\log\left(\frac{p_{l|k}}{q_{l|k}}\right)=\sum_{k,l\neq k}p_{l|k}\log(p_{l|k})-p_{l|k}\log(q_{l|k})\\&=\sum_{k,l\neq k}p_{l|k}\log(p_{l|k})-p_{l|k}\log(E_{kl})+p_{l|k}\log(Z_k)\end{aligned} \tag{14.1.2019}$$

We derive with respect to a y_i (notice that the first term vanish a it depends only on x_i or x_j but not on any y_i or y_j. To make the derivation less cluttered, we will omit the ∂y_i term at the denominator:

$$\frac{\partial C}{\partial y_i}=\sum_{k,l\neq k}-p_{l|k}\partial\log(E_{kl})+\sum_{k,l\neq k}p_{l|k}\partial\log(Z_k) \tag{14.1.2020}$$

We start with the first term, noting that the derivative are obviously non-zero if and only if each term of the sums contains at least a i. Therefore we can reduce the two sums to only one with two terms (by symmetry):

$$\sum_{k,l\neq k}-p_{l|k}\partial\log(E_{kl})=\sum_j-p_{j|i}\partial\log(E_{ji})=\sum_j-p_{j|i}\partial\log(E_{ij})-p_{i|j}\partial\log(E_{ji}) \tag{14.1.2021}$$

Since $\partial E_{ij}=E_{ij}(-2(y_i-y_j))$ we have:

$$\sum_j-p_{j|i}\cancel{\frac{E_{ij}}{E_{ij}}}\left(-2\left(y_i-y_j\right)\right)-p_{i|j}\cancel{\frac{E_{ji}}{E_{ji}}}\left(2\left(y_j-y_i\right)\right)=2\sum_j\left(p_{j|i}+p_{i|j}\right)\left(y_i-y_j\right) \tag{14.1.2022}$$

We conclude with the second term:

$$\sum_{k,l\neq k}p_{l|k}\partial\log(Z_k)=\sum_k\sum_{l\neq k}p_{l|k}\partial\log(Z_k)=\sum_k\partial\log(Z_k)\sum_{l\neq k}p_{l|k} \tag{14.1.2023}$$

Since $\sum_{l\neq j}p_{l|j}=1$ and Z_j does not depend on k, we can write (changing variable from l to j to make it more similar to the already computed terms):

$$\sum_{j,l\neq k}p_{l|j}\partial\log(Z_j)=\sum_j\partial\log(Z_j)=\partial\log\left(\sum_{k\neq l}E_{kl}\right)=\partial\log\left(\sum_{k\neq l}e^{-\|y_k-y_l\|^2}\right) \tag{14.1.2024}$$

The derivative is non-zero only when $k=i$ or $j=i$ (also, in the latter case we can move Z_i inside the summation because constant) therefore:

$$\begin{aligned}\sum_{j,l\neq k} p_{l|j}\partial\log(Z_j) &= \sum_j \frac{1}{Z_j}\sum_{k\neq j}\partial E_{jk} = \sum_j \frac{E_{ji}}{Z_j}\left(2\left(y_j-y_i\right)\right)+\sum_j \frac{E_{ij}}{Z_i}\left(-2\left(y_i-y_j\right)\right)\\ &= 2\sum_j\left(-q_{j|i}-qi|j\right)\left(y_i-y_j\right)\end{aligned} \tag{14.1.2025}$$

Combining the previous and prior-previous relations we arrive at the final result:

$$\frac{\partial C_{\text{SNE}}}{\partial y_i} = 2\sum_j\left(p_{j|i}-q_{j|i}+p_{i|j}-q_{i|j}\right)\left(y_i-y_j\right) \tag{14.1.2026}$$

14.1.20.6.5.2 KL gradient of T-distributed Stochastic Neighbour Embedding (T-SNE)

Let us do now the same job for T-SNE and for that the reader must remember:

$$q_{j|i}^{\text{TSNE}} = q_{i|j}^{\text{TSNE}} = \frac{\left(1+\|y_i-y_j\|^2\right)^{-1}}{\sum\limits_{k,l\neq}\left(1+\|y_k-y_l\|^2\right)^{-1}} = \frac{E_{ij}^{-1}}{\sum\limits_{k,l\neq} E_{kl}^{-1}} = \frac{E_{ij}^{-1}}{Z} \tag{14.1.2027}$$

For what follows we will treat again the one-dimensional case to simplify the notations (and we will also omit the TSNE exponent).

Notice that $E_{ij}=E_{ji}$. The loss function is defined as:

$$\begin{aligned}C_{\text{TSNE}} &= \sum_{k,l\neq k} p_{l|k}\log\left(\frac{p_{l|k}}{q_{l|k}}\right) = \sum_{k,l\neq k} p_{l|k}\log(p_{l|k})-p_{l|k}\log(q_{l|k})\\ &= \sum_{k,l\neq k} p_{l|k}\log(p_{l|k})-p_{l|k}\log(E_{kl}^{-1})+p_{l|k}\log(Z)\end{aligned} \tag{14.1.2028}$$

We derive with respect to a y_i. To make the derivation less cluttered, we will again omit the ∂y_i term at the denominator:

$$\frac{\partial C_{\text{TSNE}}}{\partial y_i} = \sum_{k,l\neq k} -p_{l|k}\partial\log(E_{kl}^{-1})+\sum_{k,l\neq k} p_{l|k}\partial\log(Z) \tag{14.1.2029}$$

We start with the first term, noting that the derivative is non-zero only when $\forall j$, $k=i$ or $l=i$, and also remembering that $p_{j|i}=p_{i|j}$ and $E_{ji}=E_{ij}$:

$$\sum_{k,l\neq k} -p_{l|k}\partial\log(E_{kl}^{-1}) = -2\sum_j p_{j|i}\partial\log(E_{ij}^{-1}) \tag{14.1.2030}$$

since $\partial E_{ij}^{-1}=E_{ij}^{-2}\left(-2\left(y_i-y_j\right)\right)$ we have:

$$-2\sum_j p_{j|i}\frac{E_{ij}^{-2}}{E_{ij}^{-1}}\left(-2\left(y_i-y_j\right)\right) = 4\sum_j p_{j|i}E_{ij}^{-1}\left(y_i-y_j\right) \tag{14.1.2031}$$

We conclude with the second term. Using the fact that $\sum_{k,l\neq k} p_{kl} = 1$ and that Z does not depend on k or l:

$$\sum_{k,l\neq k} p_{l|k}\partial \log Z = \frac{1}{Z}\sum_{k',l'\neq k'} \partial E_{kl}^{-1} = 2\sum_{j} \frac{E_{ji}^{-2}}{Z}(-2(y_j - y_i)) = -4\sum_{j} q_{i|j}E_{ji}^{-1}(y_i - y_j) \tag{14.1.2032}$$

Combining we arrive at the final result:

$$\frac{\partial C_{\text{TSNE}}}{\partial y_i} = 4\sum_{j}\left(p_{j|i} - q_{j|i}\right)E_{ji}^{-1}\left(y_i - y_j\right) = 4\sum_{j}\left(p_{j|i} - q_{j|i}\right)\left(1 + \|y_i - y_j\|^2\right)^{-1}\left(y_i - y_j\right) \tag{14.1.2033}$$

Keep in mind that T-SNE learns a nonparametric mapping, which means that it does not learn an explicit function that maps data from the input space to the map!!! Therefore, it is not possible to embed test points in an existing map (although we could re-run T-SNE on the full dataset). A potential approach to deal with this would be to train a multivariate regressor to predict the map location from the input data. But if we are trying to apply T-SNE to "new" data, we are probably not thinking about our problem correctly, or perhaps simply we did not understand the purpose of T-SNE.

14.1.20.7 Gradient Boosting

In many supervised learning problems one has an output variable y and a vector of input variables $\vec{x}$ described via a joint probability distribution $P(\vec{x},y)$ (at least for the purposes of theoretical analysis!). Using a training set $\{(x_1,y_1),\ldots,(x_n,y_n)\}$ of known values of $\vec{x}$ and corresponding values of y, the goal is to find an approximation $\hat{F}(x)$ to a function $F(x)$ that minimizes the expected value of some specified loss function $L(y,F(x))$:

$$\hat{F}=\min_{F}\mathrm{E}_{x,y}[L(y,F(x))] \qquad (14.1.2034)$$

The "gradient boosting method" (that personally the author of these lines like better to name "incremental iterative learning") assumes a real-valued y and seeks an approximation $\hat{F}(x)$ in the form of a weighted sum of functions $h_i(x)$ from some class $\mathcal{H}$ of weak learner:

$$\hat{F}(x)=\sum_{i=1}^{M}\gamma_i h_i(x)+c^{te} \qquad (14.1.2035)$$

Keep in mind that the underlying idea is that instead of creating a single model, boosting combines multiple simple models into a single composite model. The idea is that, as we introduce more and more simple models, the overall model becomes stronger and stronger. And for reminder, in boosting terminology, the simple models are named "weak models" or "weak learners".

In accordance with the empirical risk minimization principle, the method tries to find an approximation $\hat{F}(x)$ that minimizes the average value of the loss function on the training set, i.e., minimizes the empirical risk. It does so by starting with a model, consisting of a constant function $F_0(x)$, and incrementally expands it in a greedy fashion:

$$\begin{aligned}F_0(x)&=\min_{\gamma}\sum_{i=1}^{n}L(y_i,\gamma)\\ F_m(x)&=F_{m-1}(x)+\min_{h_m\in\mathcal{H}}\left[\sum_{i=1}^{n}L(y_i,F_{m-1}(x_i)+h_m(x_i))\right]\end{aligned} \qquad (14.1.2036)$$

where $h_m\in\mathcal{H}$ is a base learner function.

Unfortunately, choosing the best function h_i at each step for an arbitrary loss function L is a computationally infeasible optimization problem in general (as least as far as we know for the moment...). Therefore, we restrict our approach to a simplified version of the problem.

The idea is to apply a steepest descent step to this minimization problem. If we consider the continuous case, i.e. where $\mathcal{H}$ is the set of arbitrary differentiable functions on $\mathbb{R}$, we would update the model in accordance with the following equations:

$$\begin{aligned}F_m(x)&=F_{m-1}(x)-\gamma_m\sum_{i=1}^{n}\nabla_{F_{m-1}}L(y_i,F_{m-1}(x_i))\\ \gamma_m&=\min_{\gamma}\sum_{i=1}^{n}L\left(y_i,F_{m-1}(x_i)-\gamma\nabla_{F_{m-1}}L(y_i,F_{m-1}(x_i))\right)\end{aligned} \qquad (14.1.2037)$$

where the derivatives are taken with respect to the functions F_i for $i \in \{1,..,m\}$. In the discrete case however, i.e. when the set $\mathcal{H}$ is finite, we choose the candidate function h closest to the gradient of L for which the coefficient γ may then be calculated with the aid of line search on the above equations. Note that this approach is a heuristic and therefore doesn't yield an exact solution to the given problem, but rather an approximation.

In pseudocode, the generic gradient boosting method is:

1. Training set $\{(x_i, y_i)\}_{i=1}^n$, a differentiable loss function $L(y, F(x))$, number of iterations M.
2. For $m = 1$ to M
   - 2.1. Compute so-named pseudo-residuals:

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \ldots, n. \tag{14.1.2038}$$

   - 2.2. Fit a base learner (e.g. tree) $h_m(x)$ to pseudo-residuals, i.e. train it using the training set $\{(x_i, r_{im})\}_{i=1}^n$.
   - 2.3. Compute multiplier γ_m by solving the following one-dimensional optimization problem:

$$\gamma_m = \min_\gamma \sum_{i=1}^n L\left(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\right) \tag{14.1.2039}$$

   - 2.4. Update the model:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{14.1.2040}$$

   - 2.5. Output $F_M(x)$

In pseudocode (non-unique and not optimized):

Algorithm 17: Generic gradient boosting tree pseudocode

```
Data: $\{\vec{x}_i, y_i\}_{i=1}^n$
Result: $F_M(\vec{x})$
initialization;
$F_0(\vec{x}) = \arg\min_\gamma \sum_{i=1}^n L(y_i, \gamma)$;
for $m = 1$ to $M$ do
    for $i = 1$ to $n$ do
        $r_{im} = -\left[\frac{\partial L(y_i, F(\vec{x}_i))}{\partial F(\vec{x}_i)}\right]_{F(x)=F_{m-1}(\vec{x}_i)}$;
    tree($h_m(\{\vec{x}_i, r_{im}\}_{i=1}^n)$);
    $\gamma_m = \min_\gamma \sum_{i=1}^n L(y_i, F_{m-1}(\vec{x}_i) + \gamma h_m(\vec{x}_i))$;
    $F_m(\vec{x}_i) = F_{m-1}(\vec{x}_i) + \gamma_m h_m(\vec{x}_i)$;
```

Here is a very nice detailed example developed by antonioACR1 (we only have his pseudonyme sadly) on Stackexchange that is categorized as a special case of gradient boosting named: "least-squares boosting algorithm[70]" (LS Boost).

The example aims to predict salary per month (in dollars) based on whether or not the observation has own house, own car and own family/children. Suppose we have a dataset of three observations where the first variable is *have own house*, the second is *have own car* and the third variable is *have family/children*, and target is *salary per month*. The observations are:

- {Yes,Yes,Yes,1000}
- {No,No,No,25}
- {Yes,No,No,5000}

Choose a number M of boosting stages, say $M=1$. The first step of gradient boosting algorithm is to start with an initial model F_0. This model is a constant defined by $\min\limits_{\gamma}\sum_{i=1}^{3} L(y_i,\lambda)$ in our case, where L is the loss function. Suppose that we are working with the usual loss function:

$$L(y_i,\gamma)=\frac{1}{2}(y_i-\gamma)^2 \tag{14.1.2041}$$

When this is the case, this constant is equal to the mean of the outputs y_i, so in our case:

$$F_0(x_i)=\frac{10000+25+5000}{3}=5008.3 \tag{14.1.2042}$$

So our initial model is $F_0(x)=5008.3$ (which maps every observation x (e.g. {No,Yes,No}) to 5008.3.

Next we should create a new dataset, which is the previous dataset but instead of y_i we take the residuals:

$$r_{i0}=-\frac{\partial L(y_i,F_0(x_i))}{\partial F_0(x_i)} \tag{14.1.2043}$$

In our case, we have for the "alignment accuracy residuals":

$$r_{i0}=y_i-F_0(x_i)=y_i-5008.3 \tag{14.1.2044}$$

So our dataset becomes:

- {Yes,Yes,Yes,4991.6}
- {No,No,No,-4983.3}
- {Yes,No,No,-8.3}

---

[70] As it is based on the usual squared error loss function!

The next step is to fit a base learner h to this new dataset. Usually the base learner is a decision tree, so we use this.

Now assume that we constructed the following decision tree h. We constructed this tree using entropy and information gain formulas (but maybe there is some mistake as the calculations were done by hand, however for our purposes we can assume it's correct!):

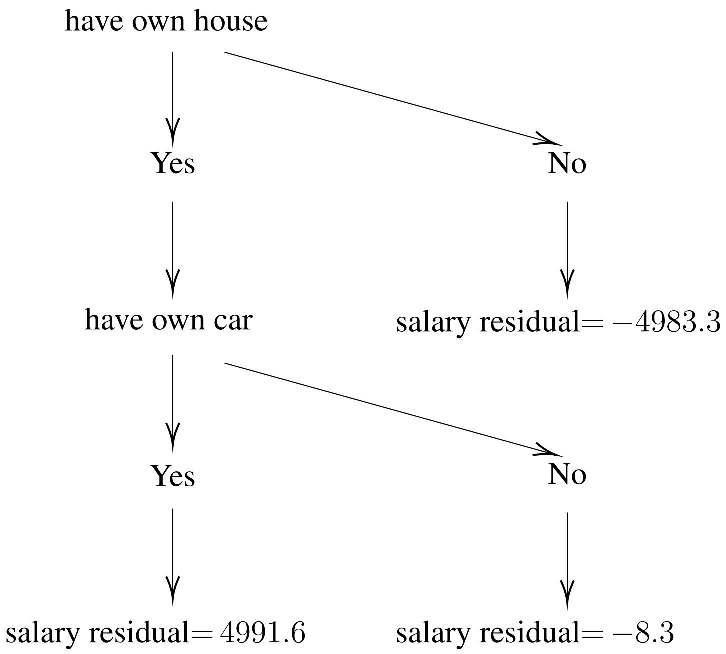


Figure 14.178: **Decision tree base learn for gradient boosting**

Let us denote this decision tree h_0. The next step is to find a constant:

$$\lambda_0 = \min_{\gamma} \sum_{i=1}^{3} L\left(y_i, F_0(x_i) + \lambda h_0(x_i)\right) \tag{14.1.2045}$$

Therefore, we want a constant λ minimizing:

$$\begin{aligned} C = &\frac{1}{2}(10000 - (5008.3 + \lambda \cdot 4991.6))^2 + \frac{1}{2}(25 - (5008.3 + \lambda(-4983.3)))^2 \\ &+ \frac{1}{2}(5000 - (5008.3 + \lambda(-8.3)))^2 \end{aligned} \tag{14.1.2046}$$

This is where gradient descent comes in handy!

Suppose that we start at $P_0 = 0$. We choose the learning rate equal to $\eta = 0.01$. We have:

$$\begin{aligned} \frac{\partial C}{\partial \lambda} = &(10000 - (5008.3 + \lambda \cdot 4991.6))(-4991.6) + (25 - (5008.3 + \lambda(-4983.3))) \cdot 4983.3 \\ &+ (5000 - (5008.3 + \lambda(-8.3))) \cdot 8.3 \end{aligned} \tag{14.1.2047}$$

Then our next value P_1 is given by:

$$P_1 = 0 - \eta \frac{\partial C}{\partial \lambda}(0) = 0 - .01(-4991.6 \cdot 4991.7 + 4983.4 \cdot (-4983.3) + (-8.3) \cdot 8.3) \tag{14.1.2048}$$

We repeat this step N times, and suppose that the last value is P_N. If N is sufficiently large and η is sufficiently small then $\lambda := P_N$ should be the value where:

$$\sum_{i=1}^{3} L(y_i, F_0(x_i) + \lambda h_0(x_i)) \tag{14.1.2049}$$

is minimized. If this is the case, then our λ_0 will be equal to P_N. Just for the sake of it, suppose that $P_N = 0.5$ (so that $\sum_{i=1}^{3} L(y_i, F_0(x_i) + \lambda h_0(x_i)))$ is minimized at $\lambda := 0.5$). Therefore, $\lambda_0 = 0.5$.

The next step is to update our initial model F_0 by $F_1(x) := F_0(x) + \lambda_0 h_0(x)$. Since our number of boosting stages is just one, then this is our final model F_1.

Now suppose that we want to predict a new observation $x = \{\text{Yes}, \text{Yes}, \text{No}\}$ (so this person does have own house and own car but no children). What is the salary per month of this person? We just compute $F_1(x) = F_0(x) + \lambda_0 h_0(x) = 5008.3 + 0.5 \cdot 4991.6 = 7504.1$. So this person earns 7504.1 per month according to our model.

Here is an interesting abstract visualization that may help to understand this incremental iterative learning process (we don't know who is the author of this figure for sure but maybe it's Terence Parr):

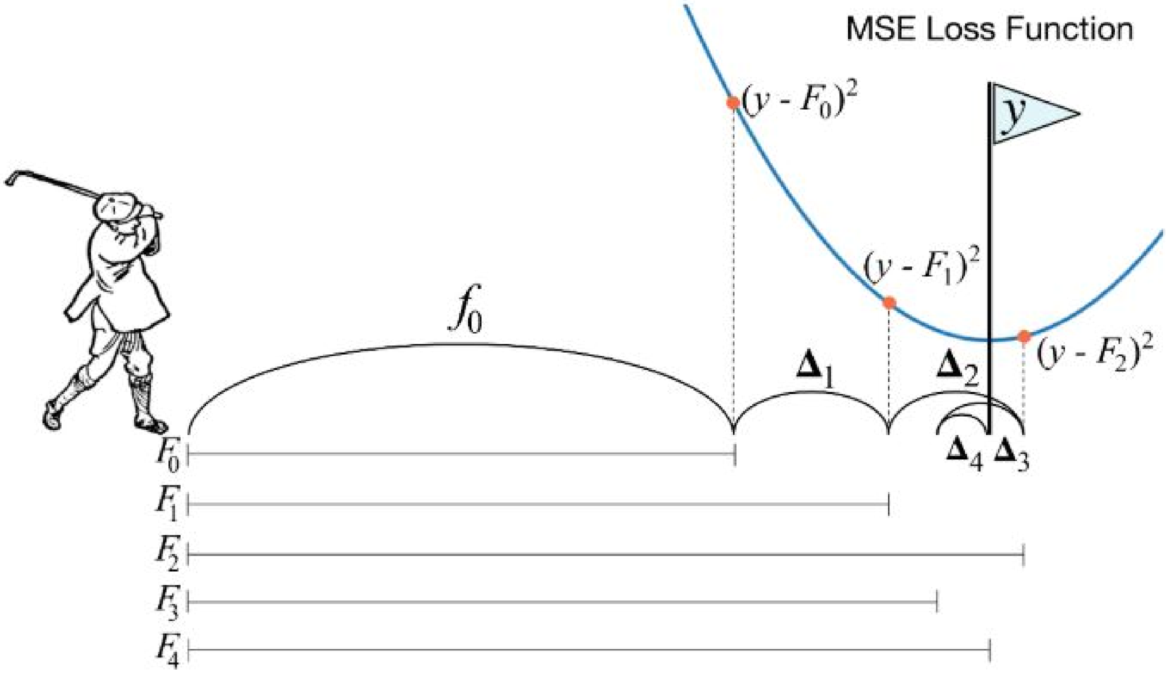


Figure 14.179: **Illustrate visualization of the iterative concept of Gradient Boosting**

Or a more explicit one would be:

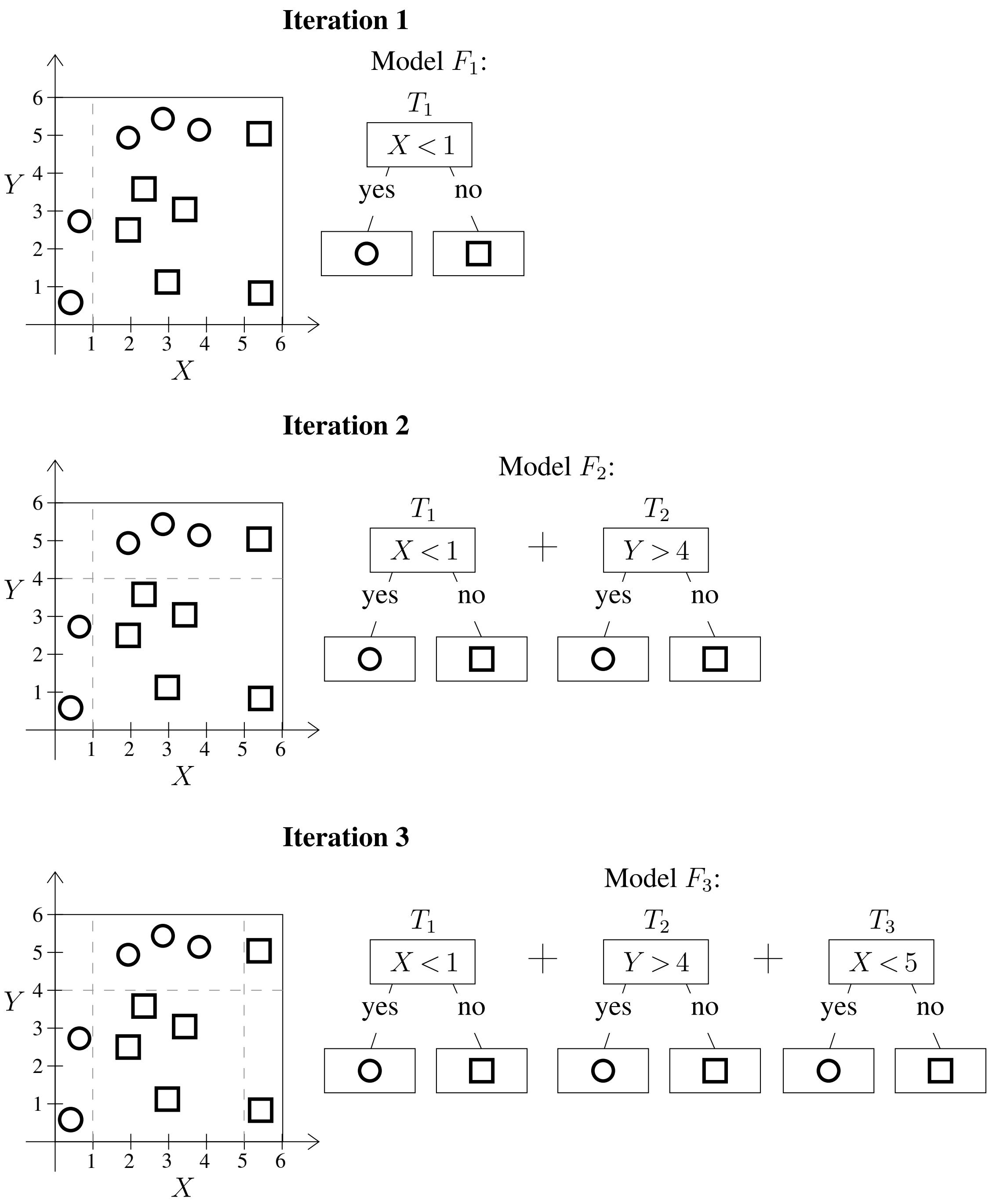

Iteration 1
Model F_1:
T_1
$X<1$
yes
no
Iteration 2
Model F_2:
T_1
$X<1$
yes
no
+
T_2
$Y>4$
yes
no
Iteration 3
Model F_3:
T_1
$X<1$
yes
no
+
T_2
$Y>4$
yes
no
+
T_3
$X<5$
yes
no
Y
X
1
2
3
4
5
6

Figure 14.180: **Illustrated Gradient Boosting with Trees**

Remark

It should be quite obvious at this level of the book that take a linear regression model as weak learner is a non-sense. Indeed, the minimization of the linear regression Loss function is a closed form problem whose solution is perfectly known and that gives us the best linear unbiased estimator $\vec{\beta}$ out of all possibilities. However, dropping unbiasedness may allow to do a bit better particularly under high multicollinearity.

14.1.20.8 Neural networks

Neural networks made from artificial cell structures are an approach for addressing from a new angle of perspective the problems of perception, memory, learning and non-linear reasoning (in other words ... artificial intelligence, abbreviated "IA") as well as genetic algorithms (see further below). They have also proved to be very promising alternatives to bypass some of the limitations of conventional numerical methods (see autodriven cars, autodriven airplanes, automatic trading systems). Thanks to their parallel processing of information and mechanisms inspired of nerve cells (neurons), they infer emergent properties to solve problems once referred to be as highly complex. Sometimes they result of artificial neurons is so bluffy that even best engineers have difficulties to explain how the neural network could arrive to a given observed performance (an typical non-business well known example is the Google Deep-learning dreaming machine).

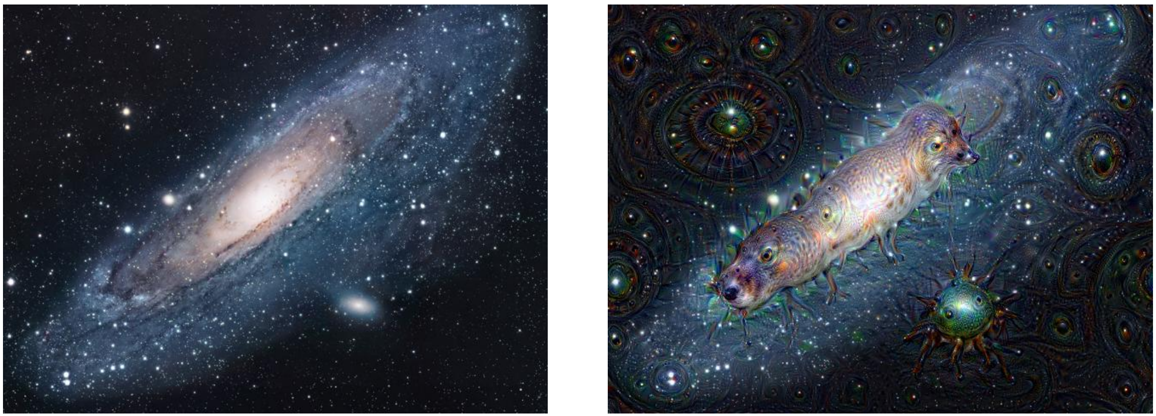


Figure 14.181: **Google Dream Machine test obtained the 12016-09-27 T20:42GMT (source: `http://psychic-vr-lab.com/deepdream`)**

However, the major problem of artificial neurons (as far as we know...) is that they are not able to self-organize, nor to self-structured intelligently by themselves, they can only change weights and change some layer parameters in a range given previously by a human. So we need at this date to proceed heuristically to find the best neural network structure adapted to a problem and this is their large current actual weakness (either using brute force via a database containing millions of models or genetic algorithms that we will see a bit further below).

We will discuss here the main architectures of neural networks. The purpose is not to study them all because they are too many variations (see figure on the next page), but rather to understand the basic internal mechanisms and how and when to use them with a minimal companion example with a common spreadsheet software. We will also discuss some concepts on fuzzy sets and logic (see section Logical Systems page 7136) in the idea that these later are incorporated into some neural network architectures that we will study.

The human brain is said ton contains about 100 billion neurons. These neurons enable us among others to read a text while maintaining regular breathing to oxygenate our blood, activating our heart which ensures efficient circulation of the blood to nourish our cells, etc. After a very long learning path and trial and errors they also enable us to read books, innovate, create, copy concepts and gentle with other people of the same species (but the learning path doesn't work for all human being for all humans machine as some reject respect of life and of differences...).

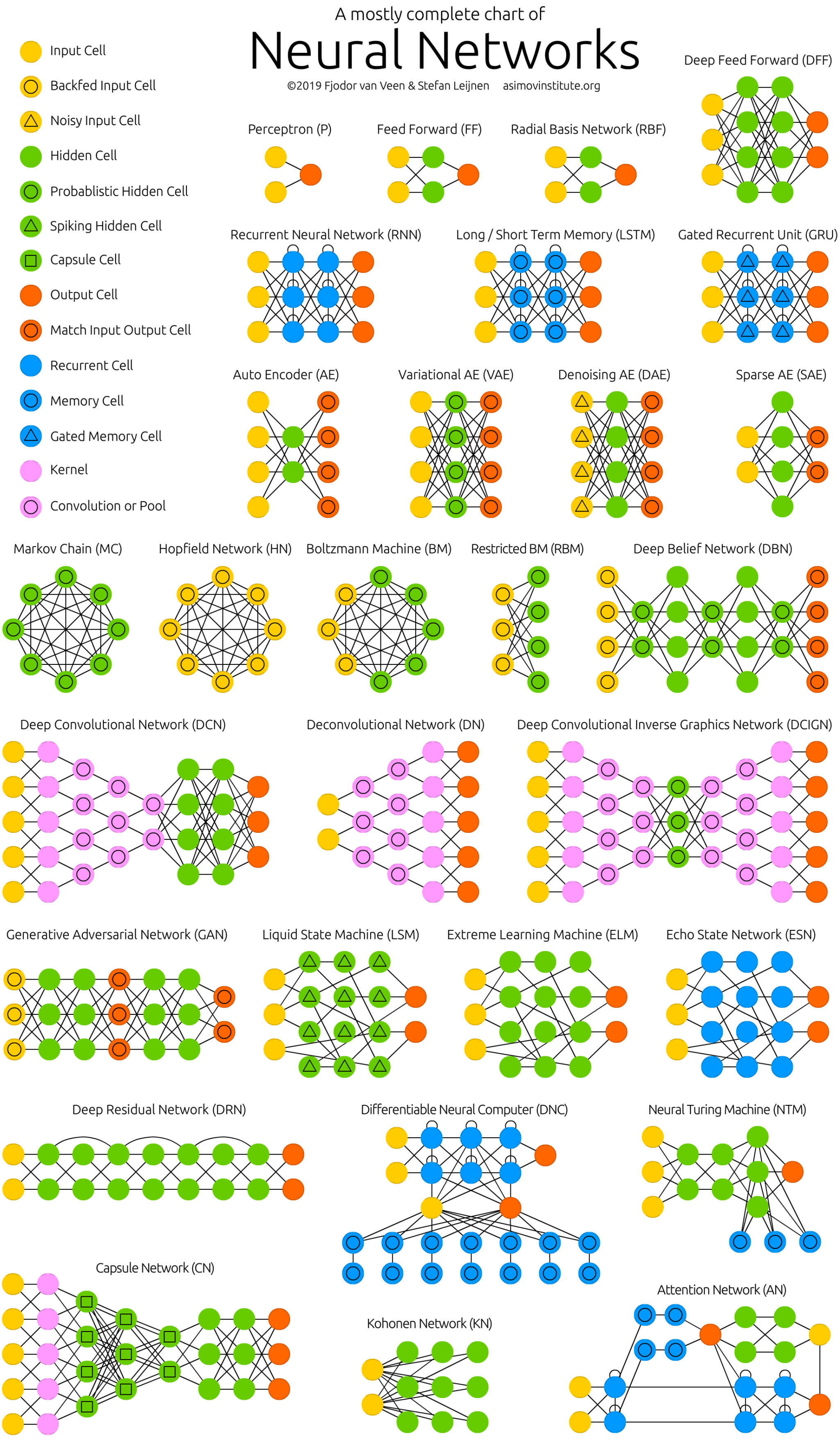

A mostly complete chart of
Neural Networks
©2019 Fjodor van Veen & Stefan Leijnen asimovinstitute.org
Input Cell
Backfed Input Cell
Noisy Input Cell
Hidden Cell
Probablistic Hidden Cell
Spiking Hidden Cell
Capsule Cell
Output Cell
Match Input Output Cell
Recurrent Cell
Memory Cell
Gated Memory Cell
Kernel
Convolution or Pool
Perceptron (P)
Feed Forward (FF)
Radial Basis Network (RBF)
Deep Feed Forward (DFF)
Recurrent Neural Network (RNN)
Long / Short Term Memory (LSTM)
Gated Recurrent Unit (GRU)
Auto Encoder (AE)
Variational AE (VAE)
Denoising AE (DAE)
Sparse AE (SAE)
Markov Chain (MC)
Hopfield Network (HN)
Boltzmann Machine (BM)
Restricted BM (RBM)
Deep Belief Network (DBN)
Deep Convolutional Network (DCN)
Deconvolutional Network (DN)
Deep Convolutional Inverse Graphics Network (DCIGN)
Generative Adversarial Network (GAN)
Liquid State Machine (LSM)
Extreme Learning Machine (ELM)
Echo State Network (ESN)
Deep Residual Network (DRN)
Differentiable Neural Computer (DNC)
Neural Turing Machine (NTM)
Capsule Network (CN)
Kohonen Network (KN)
Attention Network (AN)

Figure 14.182: **Neural Networks complete chart (source: `http://www.asimovinstitute.org`, author: Fjodor van Veen)**

Each of these neurons is also quit complex. Essentially, it is living tissue and chemistry and application of physics law (as the rest of our body). Neuro-physicists are just beginning to understand some of their internal mechanisms and are also able to influence them since a few decades. We usually think that their different neuronal functions, including memory, is stored at the connections (synapses) between neurons. It is this kind of theory that has inspired most of the artificial neural network architectures (say to be "formal neural networks"). Learning is then the process that consist either to establish new connections, or modify existing one by trial and error or by an external reference (we will focus especially on the latter option).

This brings us to a fundamental question, based on our current knowledge: can we build approximate mathematical models of neurons and make them to possibly perform useful tasks? Well, the short answer is: yes, even if the networks that we develop have only a tiny fraction of the power of the human brain actually (year 12003 according to holocene calendar), and that's the goal here to show how we can do it formally (because technically it is obvious that more powerful are the computer, more impressive will be the results)!

Neural networks are now used in all kinds of application in various fields. For example, there are neural networks developed for: aircraft autopilots, car autopilot, automotive control systems automatic reading of bank checks, automatic reading of postal addresses, signal processing, ballistic missile autopilots, voice synthesis, personal assistant, computer vision systems, market predictions, financial risks assessments, various manufacturing processes, medical diagnosis, oil and gas exploration, robotics, telecommunications, management decisioneering systems, classification and many others. In short, neural networks today have a significant impact and there is a safe bet that their importance will continue to grow in the future.

14.1.20.8.1 Neuron model

The mathematical model of an artificial neuron, or "perceptron" is shown in the figure below (also in the previous figure with the overall summary of classical neural networks). A neuron consists essentially of an "integrator" which performs the weighted sum of its inputs (such as statistical mean ponderated by the inverse number of items!). The result n of that is then transformed by a transfer function f that produces the output a of the formal neuron.

The R neuron inputs correspond to vector correspond traditionally denoted:

$$\vec{p}=\begin{pmatrix}p_1\\p_2\\\vdots\\p_R\end{pmatrix} \tag{14.1.2050}$$

while:

$$\vec{w}=\begin{pmatrix}w_{1,1}\\w_{1,2}\\\vdots\\w_{1,R}\end{pmatrix} \tag{14.1.2051}$$

represents the neuron's vector weight (we distinguish them to prepare ourselves to the study of multiple layers formal neurons... also named "deep learning" systems):

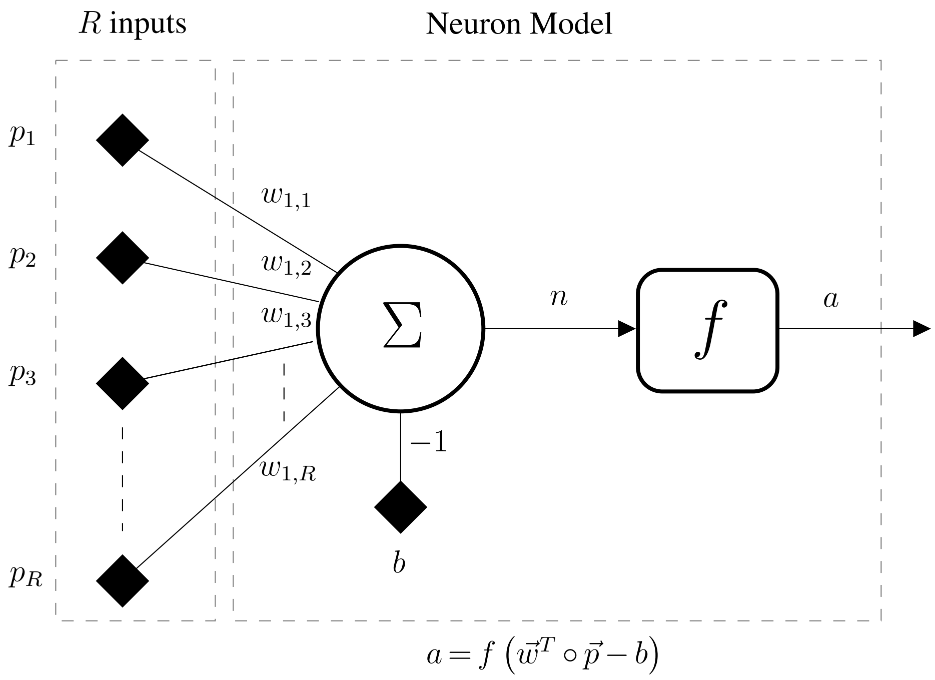


Figure 14.183: **Example of one-layer formal neuron with the input vector and scalar output**

The output of the integrator is defined (as it is an engineering technique) by the following relation (weighted sum minus the bias):

$$n=\sum_{j=1}^{R}w_{1,j}p_j-b=w_{1,1}p_1+w_{1,2}p_2+\ldots+w_{1,R}p_R-b \qquad (14.1.2052)$$

that we can also write in vector form (it could also be written as tensor but ...):

$$n=\sum_{j=1}^{R}w_{1,j}p_j-b=\vec{w}^T\vec{p}-b \qquad (14.1.2053)$$

Such that finally:

$$\boxed{a=f(n)=f\left(\sum_{j=1}^{R}w_{1,j}p_j-b\right)=f\left(\vec{w}^T\vec{p}-b\right)} \qquad (14.1.2054)$$

This output is obviously a weighted sum of the weights and inputs less what we name the "bias b of the neuron" (correction factor determined by trial and error and often null in practice). The weighted sum is named "activation level of the neuron". The bias b is also named "activation threshold of the neuron". When the activation level reaches or exceeds the threshold b, then n, the argument of f, named "activation function", becomes zero or positive of course. Otherwise it is negative.

That latter relation is also sometimes denoted:

$$a=\sigma(n)=\sigma\left(\vec{w}^T\vec{p}-b\right) \qquad (14.1.2055)$$

Remark

There's absolutely no reason for the weights in the linear layer (a.k.a. dense or fully-connected layer) to sum up to anything specific, such as 1.0. They are usually initialized with small random numbers (so initial sum is unlikely to be 1.0).

We can draw a parallel between this mathematical model and some information that we know (or think we know) about the biological neuron. That latter has three main components: the dendrites, the cell body and axon:

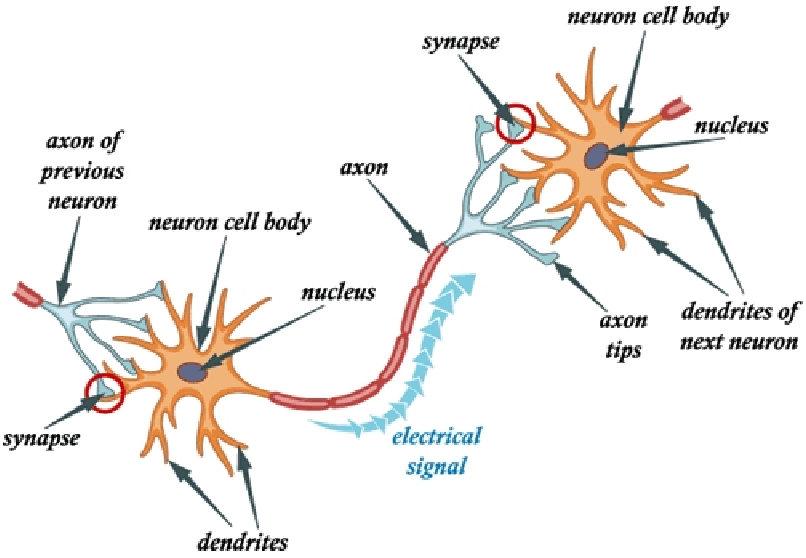


Figure 14.184: Simplified representation of the vocabulary of human neurons

The dendrites form a network of nerve receptors that are used to route to the body of the neuron the electrical signals from other neurons. It acts as like an integrator, accumulating electric charges. When the neuron becomes sufficiently excited (when the accumulated charge exceeds a certain threshold), by an electrochemical process, it generates an electric potential that spreads through its axon to eventually excite other neurons. The point of contact between the axon of one neuron and the dendrite of another neuron is named the "synapse". It seems that this is the spatial arrangement of neurons and their axons and also their quantity ($86,000,000,000$ neurons and $1.5\cdot 10^{14}$ synapses for the average human brain), and the quality of individual synaptic connections that determine the precise function of a biological neural network. This is based on this knowledge that the mathematical model described above has been defined (this is "Biomimicry Engineering" or "Biomimestim Engineering"). The detailed circuitry of neural pathways within it is known as a "connectome".

A weight of an artificial neuron represents somehow the effectiveness of a synaptic connection. A negative weight inhibits in way the entrance, while a positive weight increase its effect. It is important to remember that this is a rough approximation of a real synapse resulting in fact of a quite complex chemical process and dependent on many external factors still unclear. We must understand that our artificial neuron is a pragmatic model which, as we shall see later, will help us to accomplish interesting tasks and more neurons we have more the tasks can be complex. The biological plausibility of this model is not important to us. What counts is the result that this model will allow us to achieve.

Another limiting factor in the model we set ourselves is regarding its discreet nature. Indeed, in order to simulate a neural network, we will make the time discrete in our equations. In other words, we assume that all neurons are synchronous, that is to say that at each time t, they will simultaneously calculate the weighted sum and produce an output $a(t)=f(n(t))$. In biological networks, all neurons are actually asynchronous...

So let us come back to our model as formulated by the previous relation and by doing a change of notation $\vec{w}^T$:

$$\boxed{a=f(n)=f\left(\vec{w}^T\vec{p}-b\right)} \tag{14.1.2056}$$

This equation leads us to introduce a new more formal scheme of our neural network (FNW) or perceptron:

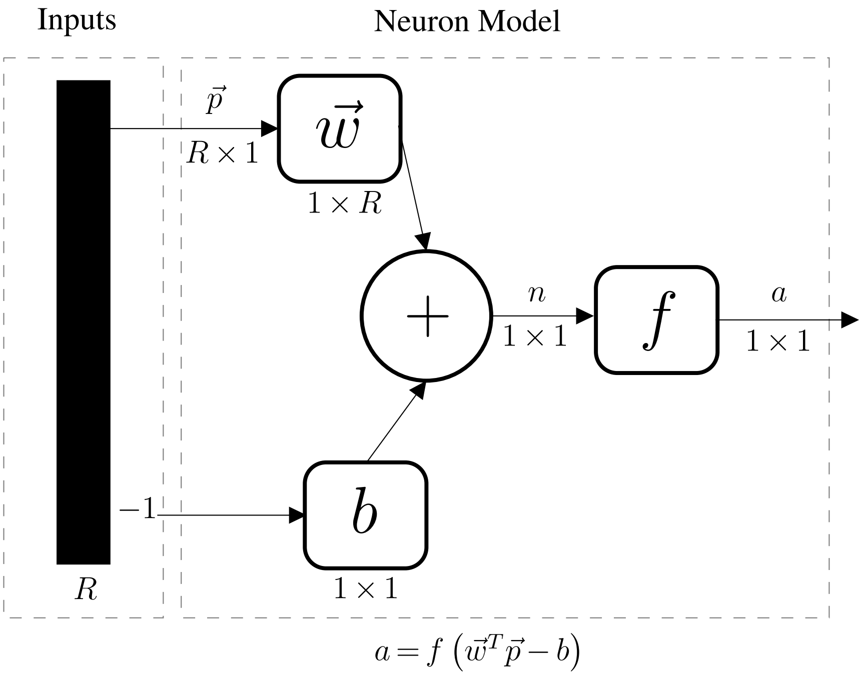


Figure 14.185: **Vector written example of one-layer formal neuron with the input vector and scalar output**

We represent the R inputs as a black rectangle (the number of entries is indicated below the rectangle). From this rectangle result a vector $\vec{p}$ whose dimensions are $R\times 1$. This vector is multiplied by a vector $\vec{w}$ that contains the weights (synaptic) neuron. In the case of a single neuron, this vector has dimension $1\times R$. The result of the multiplication is the "activation level" (scalar) which is then compared to the threshold b (a scalar) by subtraction. Finally, the output of the neuron is calculated by the function f. The output of a single neuron is then always a scalar in this special case.

To find the components of the matrix $\vec{w}$ (neuron input weight), and the bias b we use operational research techniques (simplex method, conjugate gradient method, evolutionary algorithms, etc.) on a sample of data sample of the company of size n in order to "train the model of the neuron". The purpose will then be to find weights $\vec{w}$ that minimize the quadratic cost error function given by:

$$\boxed{\min_{\vec{w},b} \sum_{k=1}^{n}(a_k - y_k)^2} \tag{14.1.2057}$$

where a_k is the model output for the vector $\vec{r}_k$ and y_k is the real value (the expected answer) corresponding to the $\vec{r}_k$.

After we check the results on a sample test before putting the neural network in production for not yet existing data (we will do a detailed example with Microsoft® Excel® just after the presentations of the transfer functions).

14.1.20.8.2 Transfer functions

So far, we have not specified the nature of the activation function $a=f(n)$ of our model. It turns out that several possibilities exist and these are empirical and must adapt to different situations (and the adaptation is sometimes also dynamic). The most common and most cited in the literature are listed in the figure below:

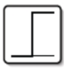
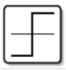
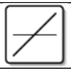
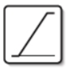
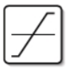
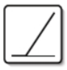
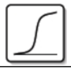
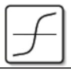
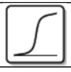
Activation function name f	**Input/Output relation**	**Profile plot**
Threshold	$a=0$ if $n<0$ $a=1$ if $n\geq 0$	
Symmetric threshold	$a=-1$ if $n<0$ $a=1$ if $n\geq 0$	
Linear	$a=n$	
Saturated linear	$a=0$ if $n<0$ $a=n$ if $0\geq n\geq 1$ $a=1$ if $n>1$	
Symmetric saturated linear	$a=-1$ if $n<-1$ $a=n$ if $-1\geq n\geq 1$ $a=1$ if $n>1$	
Rectified Linear Unit Function (RELU)	$a=0$ if $n<0$ $a=n$ if $n\geq 0$	
Sigmoid	$a=\dfrac{1}{1+e^{-n}}$	
Hyperbolic tangent (TANH)	$a=\dfrac{e^{n}-e^{-n}}{e^{n}+e^{-n}}$	
Competitive	$a=1$ if n is maximum $a=0$ otherwise	**C**
Gompertz	$a=\alpha e^{-\beta e^{\gamma n}}$	
...	...	...

Table 14.32: **Types of transfer functions for FNN**

The three most used in the field of engineering are the functions "threshold" (I) , "linear" (II) and "sigmoid" (III) as shown in details below:

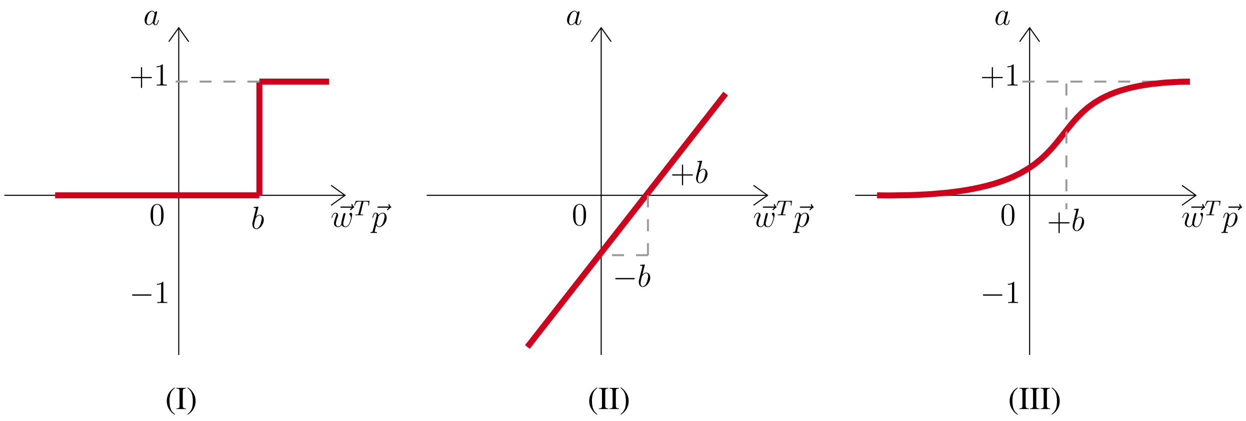


Figure 14.186: **Most used transfer functions for neural networks in the 120th century**

As its name suggests it, the threshold function applies a threshold on its input. Specifically, a negative input does not pass the threshold, the function returns then a 0 (false), while a positive or zero input exceeds the threshold, and the function returns then a 1 (true). It is obvious that this kind of feature is to make binary decisions (this function can also be assimilated to the Heaviside function for those who know it...).

The linear function is itself very simple, it directly associate its input to an output according to the relation $a=f(n)=n$. It is then evident that the output of the neuron corresponds to its activation level for which the zero value (the ordinate at the origin) occurs when $\vec{w}^T\vec{p}=b$.

The sigmoid transfer function is itself defined by the mathematical relation:

$$a=\frac{1}{1+e^{-n}} \tag{14.1.2058}$$

It looks like two threshold function, either the linear function, as we are far or near to b respectively. The threshold function is very non-linear because there is a discontinuity when $\vec{w}^T\vec{p}=b$. For its part, the linear function is entirely linear. It has no change in slope. The sigmoid is an interesting compromise between the two. Finally notice that the hyperbolic tangent (TANH) function is a symmetric version of the sigmoid.

It may be quite obvious to the reader that if the activation function is linear, then the neural network is a regression model! If it is a logistic function, then the neural network is a binary classification model!

We have introduced during our study of logistics regressions the softmax function (see above page 6479). The softmax function is a typical activation function for a multi-class neural network classifier. That is a neural network with multiple input but also multiple outputs. If we denote the weights $\vec{v}$ and we index each different output by the letter j, the such a neural network will be represented by:

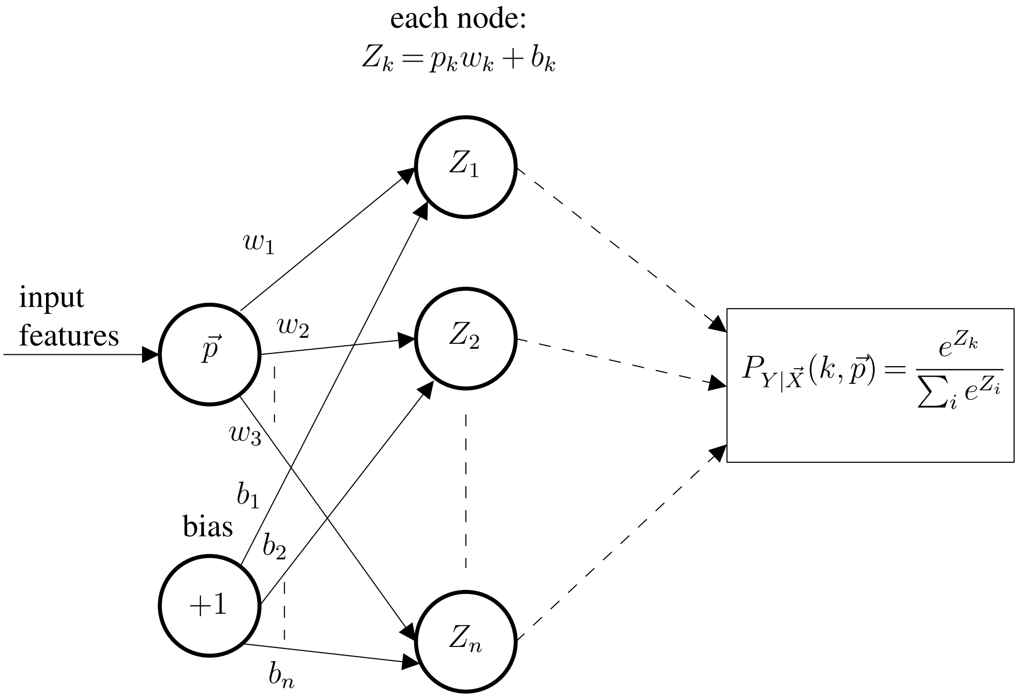

each node:
$Z_k = p_k w_k + b_k$
input features
$\vec{p}$
w_1
w_2
w_3
bias
$+1$
b_1
b_2
b_n
Z_1
Z_2
Z_n
$P_{Y|\vec{X}}(k,\vec{p}) = \frac{e^{Z_k}}{\sum_i e^{Z_i}}$

Figure 14.187: Multi-class softmax neural network

14.1.20.8.3 Various cost functions

Relatively to Neural Networks, as we already know, a cost function is a measure of how good a neural network did with respect to it's given training sample and the expected output. It also may depend on variables such as weights and biases.

Remember also that a cost function is a single value (scalar), not a vector, because it rates how good the neural network did as a whole.

Specifically, a cost function is of the form:

$$J(W,\vec{b},\vec{y},\vec{\phi}) \tag{14.1.2059}$$

where W is our neural network's weights, $\vec{b}$ is our neural network's biases, $\vec{x}$ is the input of a single training sample, $\vec{y}$ is the desired output (i.e. expected output) of that training sample and ϕ_i is the activation function.

Here are some common empirical cost functions:

- The "Quadratic cost" also known as mean squared error, maximum likelihood, and sum squared error, is defined as:

  $$J_{\text{MST}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\frac{1}{2}\sum_{i=1}^{n}\left(\phi_i(z)-y_i\right)^2 \tag{14.1.2060}$$

  The gradient of this cost function with respect to the output of a neural network is:

  $$\nabla_{\phi}J_{\text{MST}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\sum_{i=1}^{n}\left(\phi_i(z)-y_i\right) \tag{14.1.2061}$$

- The "Cross-entropy cost" also known as "Bernoulli negative log-likelihood" and "Binary Cross-Entropy", is defined as:

  $$J_{\text{CE}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=-\sum_{i=1}^{n}\left[y_i\ln(\phi(z))+(1-y_i)\ln(1-y_i)\right] \tag{14.1.2062}$$

  The gradient of this cost function with respect to the output of a neural network is:

  $$\nabla_{\phi}J_{\text{CE}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\sum_{i=1}^{n}\frac{\phi_i(z)-y_i}{(1-\phi_i(z))\phi_i(z)} \tag{14.1.2063}$$

- The "Exponential cost" requires choosing some parameter τ that you think will give you the behaviour you want, typically you'll just need to play with this until things work good, and is it is defined by:

  $$J_{\exp}\left(W,\vec{b},\vec{y},\vec{\phi},\tau\right)=\tau\exp\left(\frac{1}{\tau}\sum_{i=1}^{n}\left(\phi_i(z)-y_i\right)^2\right) \tag{14.1.2064}$$

  The gradient of this cost function with respect to the output of a neural network is:

  $$\nabla_{\phi}J_{\exp}\left(W,\vec{b},\vec{y},\vec{\phi},\tau\right)=\frac{2}{\tau}J_{\exp}\left(W,\vec{b},\vec{y},\vec{\phi},\tau\right)\sum_{i=1}^{n}\phi_i(z)-y_i \tag{14.1.2065}$$

- The "Hellinger distance" is defined as:

$$J_{\text{HD}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\frac{1}{\sqrt{2}}\sum_{i=1}^{n}\left(\sqrt{\phi_i(z)}-\sqrt{y_i}\right)^2 \tag{14.1.2066}$$

  This needs to have positive values, and ideally values between 0 and 1. The same is true for the following divergences.

  The gradient of this cost function with respect to the output of a neural network is:

$$\nabla_{\phi}J_{\text{exp}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\frac{1}{\sqrt{2}}\sum_{i=1}^{n}\frac{\sqrt{\phi_i(z)}-\sqrt{y_i}}{\sqrt{\phi_i(z)}} \tag{14.1.2067}$$

- The "Kullback–Leibler cost" is defined as (see section Statistical Mechanics page 3542):

$$J_{\text{KL}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\sum_{i=1}^{n}y_i\log\left(\frac{y_i}{\phi_i(z)}\right) \tag{14.1.2068}$$

  The gradient of this cost function with respect to the output of a neural network is:

$$\nabla_{\phi}J_{\text{KL}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=-\sum_{i=1}^{n}\frac{y_i}{\phi_i(z)} \tag{14.1.2069}$$

- The "Generalized Kullback–Leibler cost", a special case of the "Bregman divergence[71]", is defined as:

$$J_{\text{GKL}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\sum_{i=1}^{n}y_i\log\left(\frac{y_i}{\phi_i(z)}\right)-\sum_{i=1}^{n}y_i+\sum_{i=1}^{n}\phi_i(z) \tag{14.1.2070}$$

  The gradient of this cost function with respect to the output of a neural network is:

$$\nabla_{\phi}J_{\text{GKL}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\sum_{i=1}^{n}\frac{\phi_i(z)-y_i}{\phi_i(z)} \tag{14.1.2071}$$

- The "Itakura–Saito distance", also a special case of the Bregman divergence, is defined as:

$$J_{\text{IS}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\sum_{i=1}^{n}\left(\frac{y_i}{\phi_i(z)}-\log\left(\frac{y_i}{\phi_i(z)}\right)-1\right) \tag{14.1.2072}$$

  The gradient of this cost function with respect to the output of a neural network is:

$$\nabla_{\phi}J_{\text{IS}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\sum_{i=1}^{n}\frac{\phi_i(z)-y_i}{\phi_i^2(z)} \tag{14.1.2073}$$

- And very likely many others...

---

[71] It seems that in machine learning, Bregman divergences are used to calculate the bi-tempered logistic loss, performing better than the softmax function with noisy datasets

To summarize:

Cost function name	**Mathematical expression**	**Gradient**
Quadratic cost	$J_{\text{MST}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\dfrac{1}{2}\sum\limits_{i=1}^{n}(\phi_i(z)-y_i)^2$	$\nabla_{\phi}J_{\text{MST}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\sum\limits_{i=1}^{n}(\phi_i(z)-y_i)$
Cross-entropy cost	$J_{\text{CE}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=-\sum\limits_{i=1}^{n}\left[y_i\ln(\phi(z))+(1-y_i)\ln(1-y_i)\right]$	$\nabla_{\phi}J_{\text{CE}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\sum\limits_{i=1}^{n}\dfrac{\phi_i(z)-y_i}{(1-\phi_i(z))\phi_i(z)}$
Exponential cost	$J_{\exp}\left(W,\vec{b},\vec{y},\vec{\phi},\tau\right)=\tau\exp\left(\dfrac{1}{\tau}\sum\limits_{i=1}^{n}(\phi_i(z)-y_i)^2\right)$	$\nabla_{\phi}J_{\exp}\left(W,\vec{b},\vec{y},\vec{\phi},\tau\right)=\dfrac{2}{\tau}J_{\exp}\left(W,\vec{b},\vec{y},\vec{\phi},\tau\right)\sum\limits_{i=1}^{n}\phi_i(z)-y_i$
Hellinger distance	$J_{\text{HD}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\dfrac{1}{\sqrt{2}}\sum\limits_{i=1}^{n}\left(\sqrt{\phi_i(z)}-\sqrt{y_i}\right)^2$	$\nabla_{\phi}J_{\exp}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\dfrac{1}{\sqrt{2}}\sum\limits_{i=1}^{n}\dfrac{\sqrt{\phi_i(z)}-\sqrt{y_i}}{\sqrt{\phi_i(z)}}$
Kullback-Leibler (KL) cost	$J_{\text{KL}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\sum\limits_{i=1}^{n}y_i\log\left(\dfrac{y_i}{\phi_i(z)}\right)$	$\nabla_{\phi}J_{\text{GKL}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\sum\limits_{i=1}^{n}\dfrac{\phi_i(z)-y_i}{\phi_i(z)}$
Generalized KL cost	$J_{\text{GKL}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\sum\limits_{i=1}^{n}y_i\log\left(\dfrac{y_i}{\phi_i(z)}\right)-\sum\limits_{i=1}^{n}y_i+\sum\limits_{i=1}^{n}\phi_i(z)$	$\nabla_{\phi}J_{\text{GKL}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\sum\limits_{i=1}^{n}\dfrac{\phi_i(z)-y_i}{\phi_i(z)}$
Itakura-Saito distance cost	$J_{\text{IS}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\sum\limits_{i=1}^{n}\left(\dfrac{y_i}{\phi_i(z)}-\log\left(\dfrac{y_i}{\phi_i(z)}\right)-1\right)$	$\nabla_{\phi}J_{\text{IS}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\sum\limits_{i=1}^{n}\dfrac{\phi_i(z)-y_i}{\phi_i^2(z)}$
...	...	...

14.1.20.8.4 Network Architecture

By definition, a "neural network" (we will see the formal mathematical definition further below at page 6977) is a network of several neurons, usually organized in layers. To build one layer of neurons S, we simply need to assemble them as in the figure below (R inputs with S hidden neurons and S outputs):

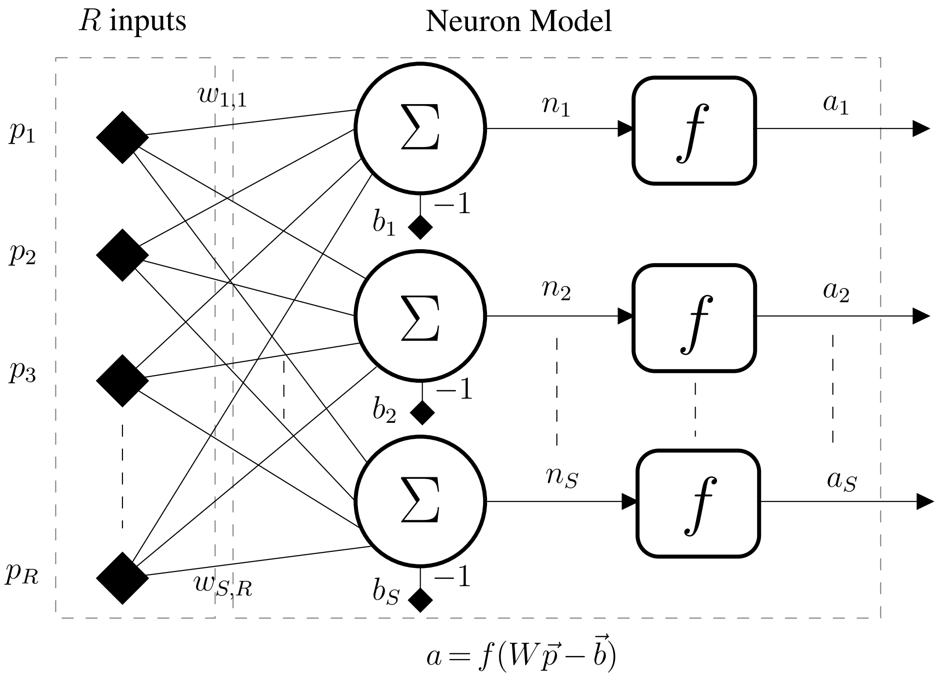


Figure 14.188: Example of one-layer formal neural network with the input vector and output vector

The S neurons of a same layer are all connected to the R inputs in the figure above. We then say that the layer is "fully connected". But this is a special case and not a generality. Often the inputs of a neuron are different from those of another neuron, etc.

Remark

If a neural network that has for purpose to learn that can learn a probability distribution over its set of inputs has only one hidden layer of S neurons not interconnected between them (as is the case in the figure above), that is to say independent, we speak then of "restricted Boltzmann machine RBM ".

A weight $w_{i,j}$ is associated with each connection. We will always denote the first index by i and the second by j. The first index (row) always means the neuron number on the layer, while the second index (column) specifies the number of the input. Thus, $w_{i,j}$ denotes the weight of the connection that connects the neuron i to its input j. All the weights of a neuron lays thus forms a matrix W of dimension $S \times R$:

$$W=\begin{bmatrix} w_{1,1} & w_{1,2} & \ldots & w_{1,R} \\ w_{2,1} & w_{2,2} & \ldots & w_{1,R} \\ \vdots & \vdots & \ddots & \vdots \\ w_{S,1} & w_{S,2} & \ldots & w_{S,R} \end{bmatrix} \tag{14.1.2074}$$

We must of course take into account that dimensionally we have not necessarily $S=R$ in the general case (the numbers of neurons and inputs are independent). If we consider that the S neurons form a vector, then we can create the vectors:

$$\vec{b}=\begin{pmatrix}b_1\\b_2\\\vdots\\b_S\end{pmatrix}\qquad\vec{n}=\begin{pmatrix}n_1\\n_2\\\vdots\\n_S\end{pmatrix}\qquad\vec{a}=\begin{pmatrix}a_1\\a_2\\\vdots\\a_S\end{pmatrix}\tag{14.1.2075}$$

This brings us to the simplified representation illustrated below:

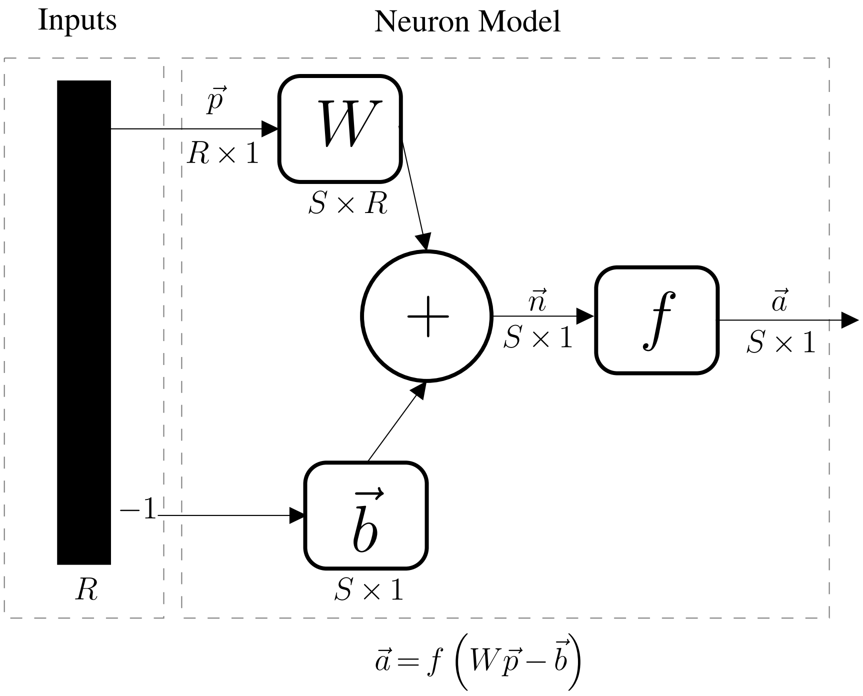

Inputs
Neuron Model
$\vec{p}$
$R\times 1$
W
$S\times R$
$+$
$\vec{n}$
$S\times 1$
f
$\vec{a}$
$S\times 1$
-1
$\vec{b}$
$S\times 1$
R
$\vec{a}=f\left(W\vec{p}-\vec{b}\right)$

Figure 14.189: Vector written example of one-layer formal neural network with the input vector and output vector

Finally, to build a neural network (or MLP for "Multi-Layer Perceptron"), it is just sufficient to combine layers as below:

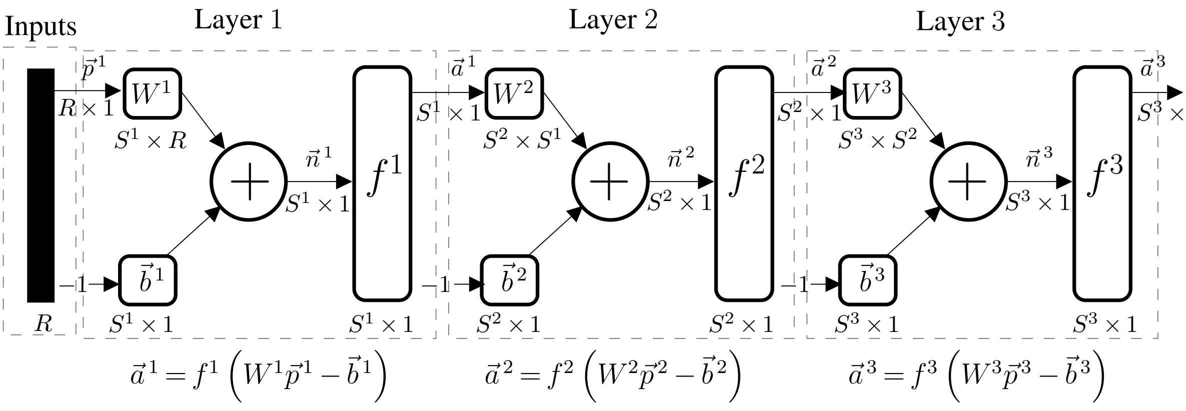

Inputs
Layer 1
Layer 2
Layer 3
$\vec{p}^1$
$R\times 1$
W^1
$S^1\times R$
$\vec{n}^1$
$S^1\times 1$
f^1
-1
$\vec{b}^1$
$S^1\times 1$
R
$S^1\times 1$
$\vec{a}^1$
$S^1\times 1$
W^2
$S^2\times S^1$
$\vec{n}^2$
$S^2\times 1$
f^2
-1
$\vec{b}^2$
$S^2\times 1$
$S^2\times 1$
$\vec{a}^2$
$S^2\times 1$
W^3
$S^3\times S^2$
$\vec{n}^3$
$S^3\times 1$
f^3
-1
$\vec{b}^3$
$S^3\times 1$
$S^3\times 1$
$\vec{a}^3$
$S^3\times 1$
$\vec{a}^1=f^1\left(W^1\vec{p}^1-\vec{b}^1\right)$
$\vec{a}^2=f^2\left(W^2\vec{p}^2-\vec{b}^2\right)$
$\vec{a}^3=f^3\left(W^3\vec{p}^3-\vec{b}^3\right)$

Figure 14.190: Principle of construction of a Multi-Layer Perceptron

This example includes R inputs and three layers of neurons having respectively S^1, S^2, S^3 neurons. In the general case, again these numbers are not necessarily equal. Each layer also has its own weight matrix W^k, where k is the layer index. In the context of vectors and matrices relatively to one a layer, we always will use an exponent to describe this index. Thus, the vectors $\vec{b}^k$, $\vec{n}^k$, $\vec{a}^k$ are also associated with the layer A.

Remark

Multilayer perceptrons are also named "vanilla feed-forward neural networks[a]". A perceptron is always feedforward, that is, all the arrows are going in the direction of the output. Neural networks in general might have loops, and if so, are often named "recurrent neural networks" (RNN). A superposition of at least three MLP is what seems to be usually named a "deep beliefs net" (DLN).

---

[a] "vanilla"= the basic version.

It should be noticed in this example that the layers that follows the first has as input the output of the previous layer. So we can put on as many layers as we want, at least in theory. We can set any number of neurons of each layer. In practice, we will see later however it is not desirable to use too many neurons. Note also that nothing prevents us from changing transfer function from one layer to another. Thus, in the general case we have not necessarily $f^1 = f^2 = f^3$.

The first last layer is obviously named "input layer", and the last layer the "output layer". The layers between the output and input layers are commonly named "hidden layers".

Not all textbooks and softwares count the number of layers in the same way. Here is a typical example where the first layer doesn't have any activation function, but is however counted as one layer (and the notation also differs slightly):

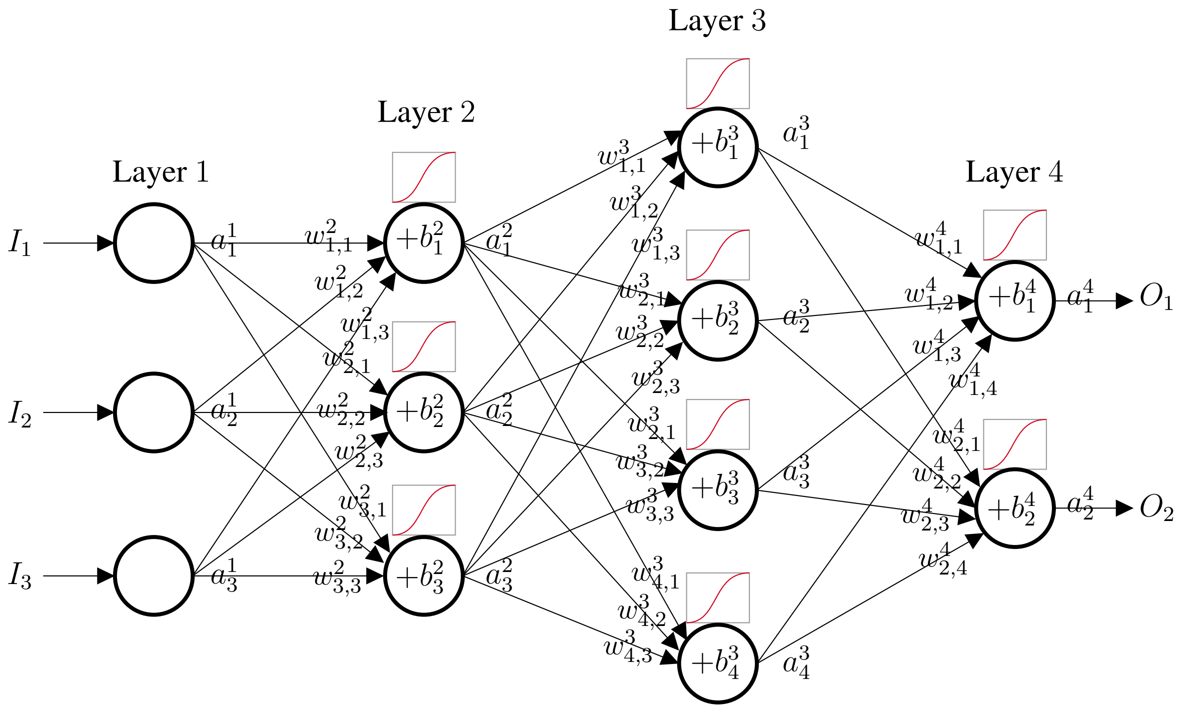

Layer 1
Layer 2
Layer 3
Layer 4
I_1
I_2
I_3
a_1^1
a_2^1
a_3^1
$w_{1,1}^2$
$w_{1,2}^2$
$w_{1,3}^2$
$w_{2,1}^2$
$w_{2,2}^2$
$w_{2,3}^2$
$w_{3,1}^2$
$w_{3,2}^2$
$w_{3,3}^2$
$+b_1^2$
$+b_2^2$
$+b_3^2$
a_1^2
a_2^2
a_3^2
$w_{1,1}^3$
$w_{1,2}^3$
$w_{1,3}^3$
$w_{2,1}^3$
$w_{2,2}^3$
$w_{2,3}^3$
$w_{3,1}^3$
$w_{3,2}^3$
$w_{3,3}^3$
$w_{4,1}^3$
$w_{4,2}^3$
$w_{4,3}^3$
$+b_1^3$
$+b_2^3$
$+b_3^3$
$+b_4^3$
a_1^3
a_2^3
a_3^3
a_4^3
$w_{1,1}^4$
$w_{1,2}^4$
$w_{1,3}^4$
$w_{1,4}^4$
$w_{2,1}^4$
$w_{2,2}^4$
$w_{2,3}^4$
$w_{2,4}^4$
$+b_1^4$
$+b_2^4$
a_1^4
a_2^4
O_1
O_2

Figure 14.191: **Explicit construction of a Multi-layer Perceptron**

Then we can relate the next layer's input to it's previous via the following relation:

$$a_j^i = f\left(\sum_k (w_{jk}^i \cdot a_k^{i-1}) + b_j^i\right) \tag{14.1.2076}$$

where:

- f is the activation function
- w_{jk}^i is the weight from the k-th neuron in the $(i-1)$-th layer to the j-th neuron in the i-th layer
- b_j^i is the bias of the j-th neuron in the i-th layer
- a_j^i represents the activation value of the j-th neuron in the i-th layer

Remark

Multilayer neural networks are more powerful than simple single layer neural networks of course. Using two layers, provided you use a sigmoid activation function on the hidden layer we can "train" a network to produce an approximation of most functions with arbitrary precision. Except in rare cases, artificial neural networks use two or three layers.

"Train" a neural network means changing the value of its weight matrices and its so that it realizes the desired input/output function (I / O). We will study in detail various algorithms and methods of heuristics approach to achieve it in different contexts.

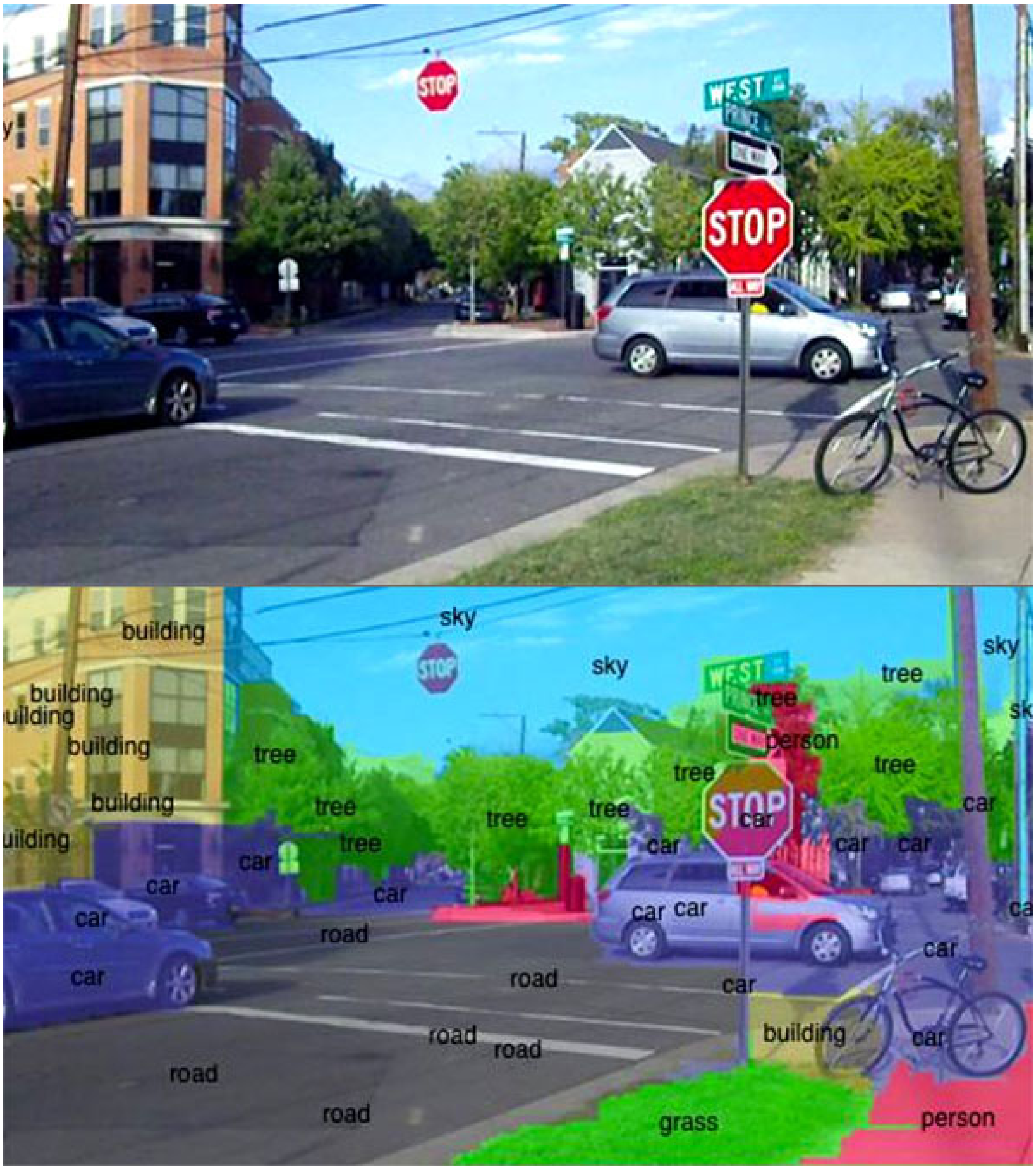

STOP
WEST
STOP
building
building
uilding
building
building
uilding
sky
sky
sky
tree
tree
tree
tree
tree
tree
tree
tree
tree
person
car
car
car
car
car
car
car
car
car
car
car
car
car
road
road
road
road
road
road
building
grass
person

Figure 14.192: **Neural Information Processing (source: Purdue University image/e-Lab)**

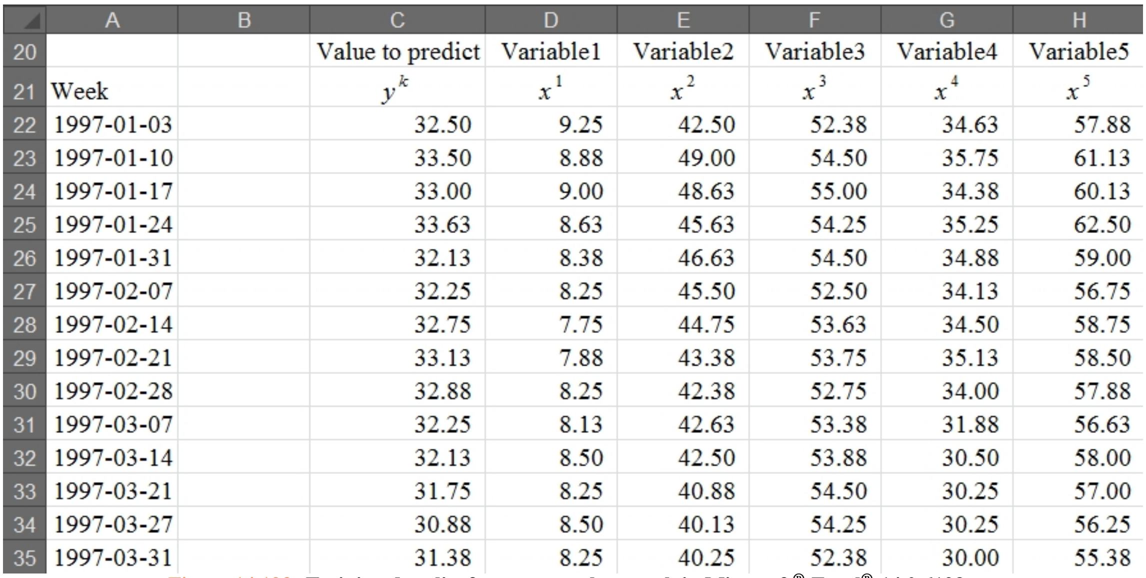
Remark

In the early 121st century (holocene calendar) there seems to no exist any standard nor accepted method for selecting the number of layers, and the number of nodes in each layer, in a feed-forward neural network. It's more trial and error.

Let us now see an easy companion example (originally developed by Joe Breedlove) as always in this textbook first done with a spreadsheet software like Microsoft® Excel®. Afterwards we will show the same output result with R and MATLAB® for which you can find the detailed procedure in the corresponding companion books.

✍**Example:**

A company has measured during 14 weeks its actual sales (Column: *Value to predict*) in function of forecast sales of five of its branches (*Variable1*, *Variable2*, etc.) and has reproduced them in Microsoft® Excel® 14.0.6123:

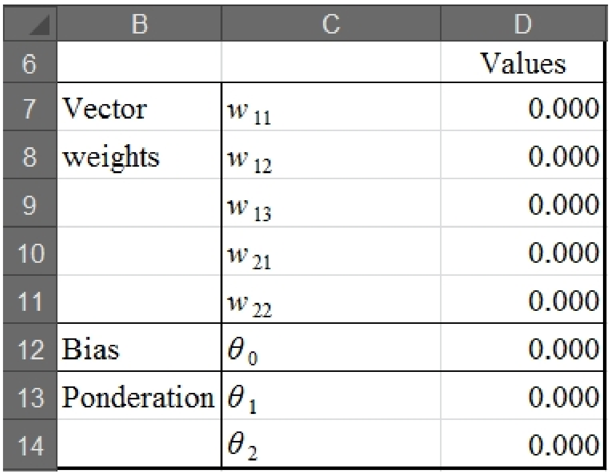


	A	B	C	D	E	F	G	H
20			Value to predict	Variable1	Variable2	Variable3	Variable4	Variable5
21	Week		y^k	x^1	x^2	x^3	x^4	x^5
22	1997-01-03		32.50	9.25	42.50	52.38	34.63	57.88
23	1997-01-10		33.50	8.88	49.00	54.50	35.75	61.13
24	1997-01-17		33.00	9.00	48.63	55.00	34.38	60.13
25	1997-01-24		33.63	8.63	45.63	54.25	35.25	62.50
26	1997-01-31		32.13	8.38	46.63	54.50	34.88	59.00
27	1997-02-07		32.25	8.25	45.50	52.50	34.13	56.75
28	1997-02-14		32.75	7.75	44.75	53.63	34.50	58.75
29	1997-02-21		33.13	7.88	43.38	53.75	35.13	58.50
30	1997-02-28		32.88	8.25	42.38	52.75	34.00	57.88
31	1997-03-07		32.25	8.13	42.63	53.38	31.88	56.63
32	1997-03-14		32.13	8.50	42.50	53.88	30.50	58.00
33	1997-03-21		31.75	8.25	40.88	54.50	30.25	57.00
34	1997-03-27		30.88	8.50	40.13	54.25	30.25	56.25
35	1997-03-31		31.38	8.25	40.25	52.38	30.00	55.38

Figure 14.193: Training data list for our neural network in Microsoft® Excel® 14.0.6123

Notice that there is absolutely no formula in the list above! The return on experience (especially with Microsoft® Excel®...) tell us it would be better to do a network architecture with two neurons, first with branches $\{1,2,3\}$ based on a sigmoid and a second with branches $\{4.5\}$ also based on a sigmoid. In addition, all should have a single bias and the both neurons should have a specific weight in comparison with the one and the other.

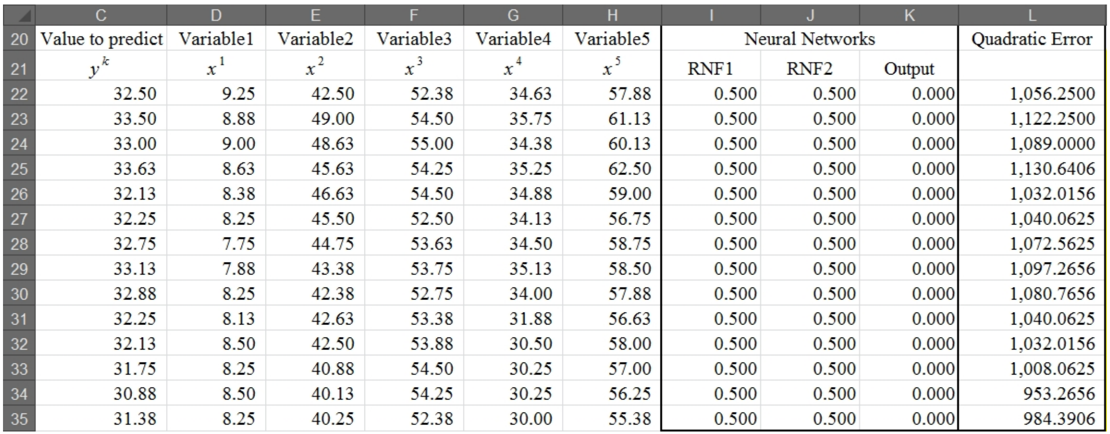
We then prepare the following table:



	B	C	D
6			Values
7	Vector	w_{11}	0.000
8	weights	w_{12}	0.000
9		w_{13}	0.000
10		w_{21}	0.000
11		w_{22}	0.000
12	Bias	θ_0	0.000
13	Ponderation	θ_1	0.000
14		θ_2	0.000

Figure 14.194: Weight bias and weights to determined for our neural network

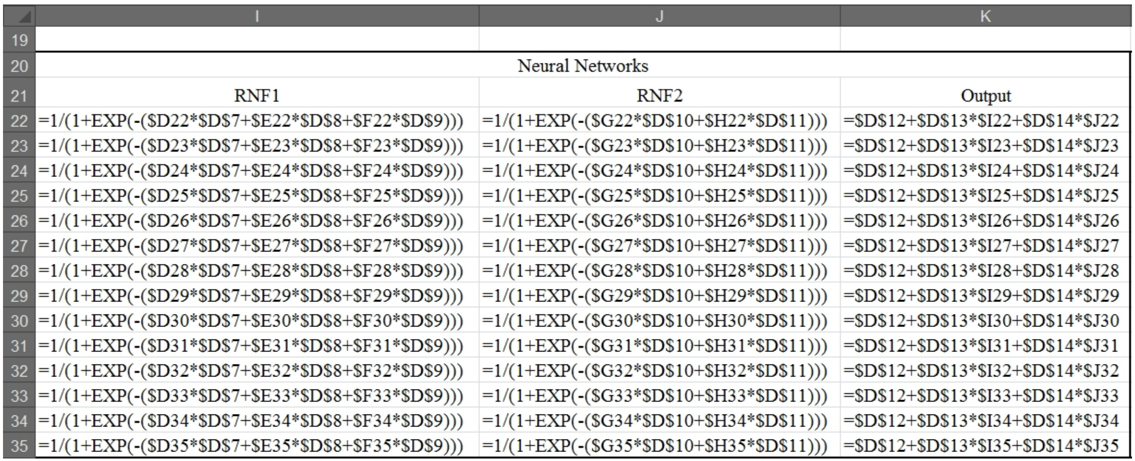
Once the table of weights, bias and ponderations built, we write our two neurons network with the sigmoid function, for example, right next to the training sample data (which will facilitate the comparison):



	C	D	E	F	G	H	I	J	K	L
20	Value to predict	Variable1	Variable2	Variable3	Variable4	Variable5	Neural Networks			Quadratic Error
21	y^k	x^1	x^2	x^3	x^4	x^5	RNF1	RNF2	Output	
22	32.50	9.25	42.50	52.38	34.63	57.88	0.500	0.500	0.000	1,056.2500
23	33.50	8.88	49.00	54.50	35.75	61.13	0.500	0.500	0.000	1,122.2500
24	33.00	9.00	48.63	55.00	34.38	60.13	0.500	0.500	0.000	1,089.0000
25	33.63	8.63	45.63	54.25	35.25	62.50	0.500	0.500	0.000	1,130.6406
26	32.13	8.38	46.63	54.50	34.88	59.00	0.500	0.500	0.000	1,032.0156
27	32.25	8.25	45.50	52.50	34.13	56.75	0.500	0.500	0.000	1,040.0625
28	32.75	7.75	44.75	53.63	34.50	58.75	0.500	0.500	0.000	1,072.5625
29	33.13	7.88	43.38	53.75	35.13	58.50	0.500	0.500	0.000	1,097.2656
30	32.88	8.25	42.38	52.75	34.00	57.88	0.500	0.500	0.000	1,080.7656
31	32.25	8.13	42.63	53.38	31.88	56.63	0.500	0.500	0.000	1,040.0625
32	32.13	8.50	42.50	53.88	30.50	58.00	0.500	0.500	0.000	1,032.0156
33	31.75	8.25	40.88	54.50	30.25	57.00	0.500	0.500	0.000	1,008.0625
34	30.88	8.50	40.13	54.25	30.25	56.25	0.500	0.500	0.000	953.2656
35	31.38	8.25	40.25	52.38	30.00	55.38	0.500	0.500	0.000	984.3906

Figure 14.195: List of sample training data with neural network cells

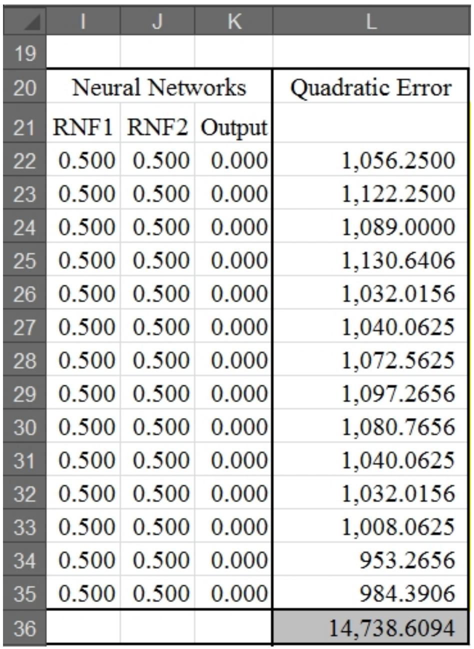
Or with the explicit formulas for the last three columns of interest:



	I	J	K
19			
20	Neural Networks		
21	RNF1	RNF2	Output
22	=1/(1+EXP(-($D22*$D$7+$E22*$D$8+$F22*$D$9)))	=1/(1+EXP(-($G22*$D$10+$H22*$D$11)))	=$D$12+$D$13*$I22+$D$14*$J22
23	=1/(1+EXP(-($D23*$D$7+$E23*$D$8+$F23*$D$9)))	=1/(1+EXP(-($G23*$D$10+$H23*$D$11)))	=$D$12+$D$13*$I23+$D$14*$J23
24	=1/(1+EXP(-($D24*$D$7+$E24*$D$8+$F24*$D$9)))	=1/(1+EXP(-($G24*$D$10+$H24*$D$11)))	=$D$12+$D$13*$I24+$D$14*$J24
25	=1/(1+EXP(-($D25*$D$7+$E25*$D$8+$F25*$D$9)))	=1/(1+EXP(-($G25*$D$10+$H25*$D$11)))	=$D$12+$D$13*$I25+$D$14*$J25
26	=1/(1+EXP(-($D26*$D$7+$E26*$D$8+$F26*$D$9)))	=1/(1+EXP(-($G26*$D$10+$H26*$D$11)))	=$D$12+$D$13*$I26+$D$14*$J26
27	=1/(1+EXP(-($D27*$D$7+$E27*$D$8+$F27*$D$9)))	=1/(1+EXP(-($G27*$D$10+$H27*$D$11)))	=$D$12+$D$13*$I27+$D$14*$J27
28	=1/(1+EXP(-($D28*$D$7+$E28*$D$8+$F28*$D$9)))	=1/(1+EXP(-($G28*$D$10+$H28*$D$11)))	=$D$12+$D$13*$I28+$D$14*$J28
29	=1/(1+EXP(-($D29*$D$7+$E29*$D$8+$F29*$D$9)))	=1/(1+EXP(-($G29*$D$10+$H29*$D$11)))	=$D$12+$D$13*$I29+$D$14*$J29
30	=1/(1+EXP(-($D30*$D$7+$E30*$D$8+$F30*$D$9)))	=1/(1+EXP(-($G30*$D$10+$H30*$D$11)))	=$D$12+$D$13*$I30+$D$14*$J30
31	=1/(1+EXP(-($D31*$D$7+$E31*$D$8+$F31*$D$9)))	=1/(1+EXP(-($G31*$D$10+$H31*$D$11)))	=$D$12+$D$13*$I31+$D$14*$J31
32	=1/(1+EXP(-($D32*$D$7+$E32*$D$8+$F32*$D$9)))	=1/(1+EXP(-($G32*$D$10+$H32*$D$11)))	=$D$12+$D$13*$I32+$D$14*$J32
33	=1/(1+EXP(-($D33*$D$7+$E33*$D$8+$F33*$D$9)))	=1/(1+EXP(-($G33*$D$10+$H33*$D$11)))	=$D$12+$D$13*$I33+$D$14*$J33
34	=1/(1+EXP(-($D34*$D$7+$E34*$D$8+$F34*$D$9)))	=1/(1+EXP(-($G34*$D$10+$H34*$D$11)))	=$D$12+$D$13*$I34+$D$14*$J34
35	=1/(1+EXP(-($D35*$D$7+$E35*$D$8+$F35*$D$9)))	=1/(1+EXP(-($G35*$D$10+$H35*$D$11)))	=$D$12+$D$13*$I35+$D$14*$J35

Figure 14.196: List of sample training data with neural network cells

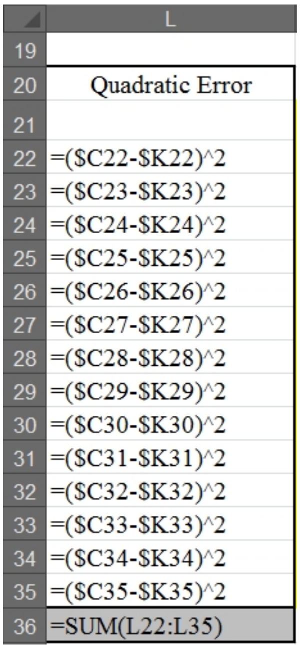
To apply operational research techniques, we need to minimize or maximize something. Therefore, we will seek to minimize the sum of squared errors by creating the following column:



	I	J	K	L
19				
20	Neural Networks			Quadratic Error
21	RNF1	RNF2	Output	
22	0.500	0.500	0.000	1,056.2500
23	0.500	0.500	0.000	1,122.2500
24	0.500	0.500	0.000	1,089.0000
25	0.500	0.500	0.000	1,130.6406
26	0.500	0.500	0.000	1,032.0156
27	0.500	0.500	0.000	1,040.0625
28	0.500	0.500	0.000	1,072.5625
29	0.500	0.500	0.000	1,097.2656
30	0.500	0.500	0.000	1,080.7656
31	0.500	0.500	0.000	1,040.0625
32	0.500	0.500	0.000	1,032.0156
33	0.500	0.500	0.000	1,008.0625
34	0.500	0.500	0.000	953.2656
35	0.500	0.500	0.000	984.3906
36				14,738.6094

Figure 14.197: **List of sample training data with neural network cells and quadratic error minimization**

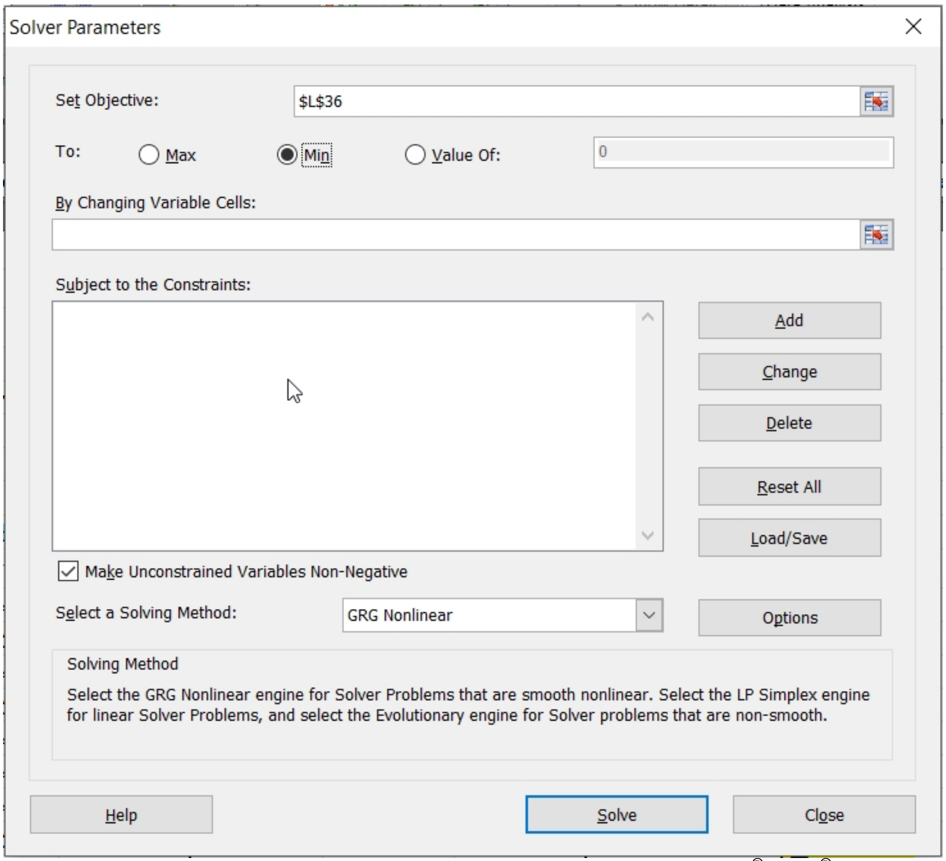
Or with the explicit formulas (we see well that this corresponds indeed to the square of the difference between the measurements and model):



	L
19	
20	Quadratic Error
21	
22	=($C22-$K22)^2
23	=($C23-$K23)^2
24	=($C24-$K24)^2
25	=($C25-$K25)^2
26	=($C26-$K26)^2
27	=($C27-$K27)^2
28	=($C28-$K28)^2
29	=($C29-$K29)^2
30	=($C30-$K30)^2
31	=($C31-$K31)^2
32	=($C32-$K32)^2
33	=($C33-$K33)^2
34	=($C34-$K34)^2
35	=($C35-$K35)^2
36	=SUM(L22:L35)

Figure 14.198: **List of sample training data with neural network cells and quadratic error minimization**

Now with the solver of Microsoft® Excel® 14.0.6123 we minimize the content of the cell L36:


Solver Parameters
Add
Change
Delete
Reset All
Load/Save
Options
Help
Solve
Close

Set Objective: $L$36

To: ○ Max ● Min ○ Value Of: 0

By Changing Variable Cells:

Subject to the Constraints:

☑ Make Unconstrained Variables Non-Negative

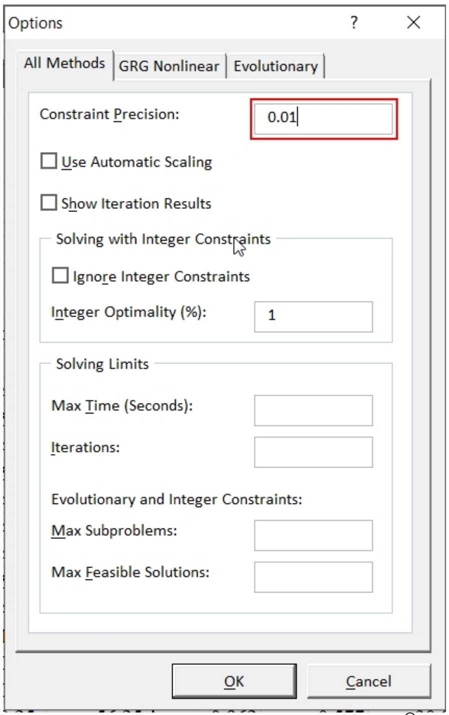
Select a Solving Method: GRG Nonlinear

Solving Method

Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

Figure 14.199: Neural network solver quadratic error minimization in Microsoft® Excel® 14.0.6123

we should not be too focused about the accuracy of the constraints for this case and therefore we have to play a little with this setting to get a satisfactory result:


Options
?
All Methods
GRG Nonlinear
Evolutionary
OK
Cancel

Constraint Precision: 0.01

☐ Use Automatic Scaling

☐ Show Iteration Results

Solving with Integer Constraints

☐ Ignore Integer Constraints

Integer Optimality (%): 1

Solving Limits

Max Time (Seconds):

Iterations:

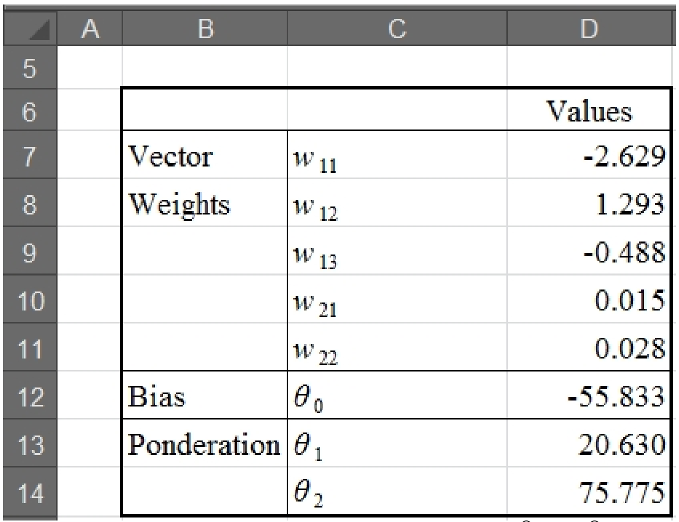
Evolutionary and Integer Constraints:

Max Subproblems:

Max Feasible Solutions:

Figure 14.200: Neural network solver settings in Microsoft® Excel® 14.0.6123

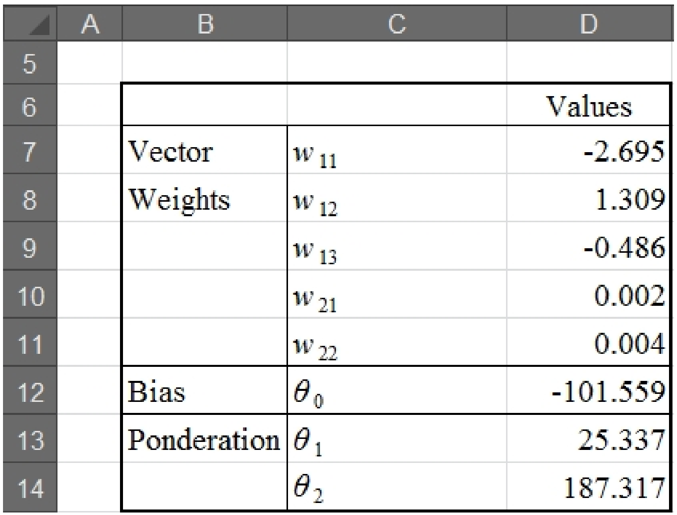
To get a satisfactory result, it will be necessary in this case to request an accuracy of 0.001. Which will give after the execution of the search by the solver a total square error of 0.8479 (cell `L36`) and for the parameters of the neural network:



	A	B	C	D
5				
6				Values
7		Vector	w_{11}	-2.629
8		Weights	w_{12}	1.293
9			w_{13}	-0.488
10			w_{21}	0.015
11			w_{22}	0.028
12		Bias	θ_0	-55.833
13		Ponderation	θ_1	20.630
14			θ_2	75.775

Figure 14.201: **Neural network solver solution in Microsoft® Excel® 14.0.6123**

Specialized software will do better with a total square error of 0.8405 (still for cell `L36`) and for the parameters of the neural network:



	A	B	C	D
5				
6				Values
7		Vector	w_{11}	-2.695
8		Weights	w_{12}	1.309
9			w_{13}	-0.486
10			w_{21}	0.002
11			w_{22}	0.004
12		Bias	θ_0	-101.559
13		Ponderation	θ_1	25.337
14			θ_2	187.317

Figure 14.202: **Neural network optimal values with specialized software**

We can graphically compare the measurements used to train the neural network and the result of the neural network model itself. Then we have:

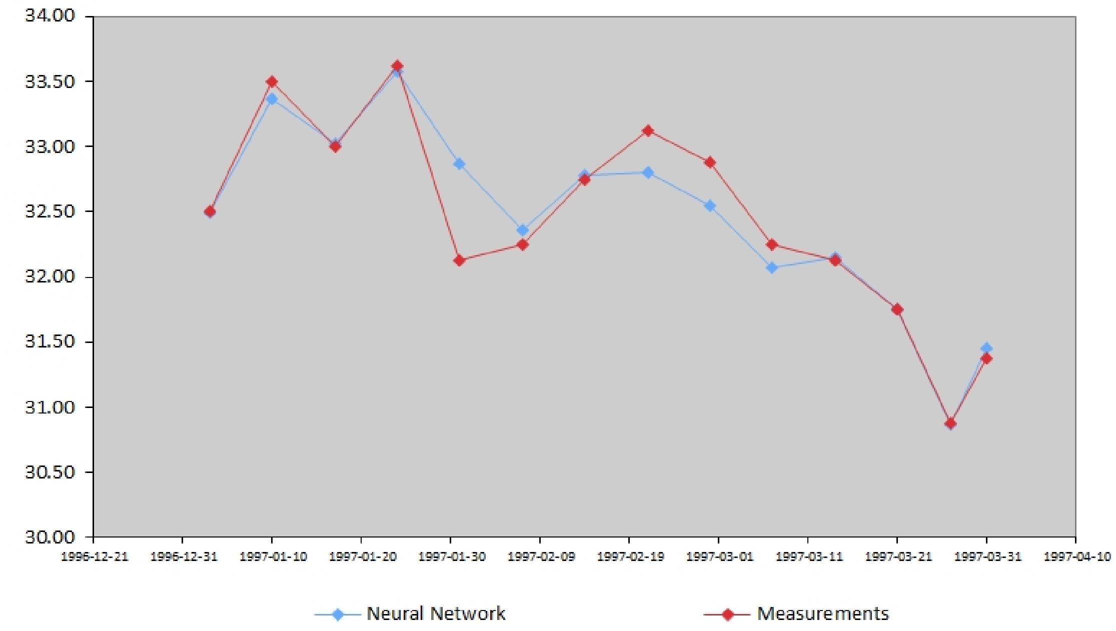
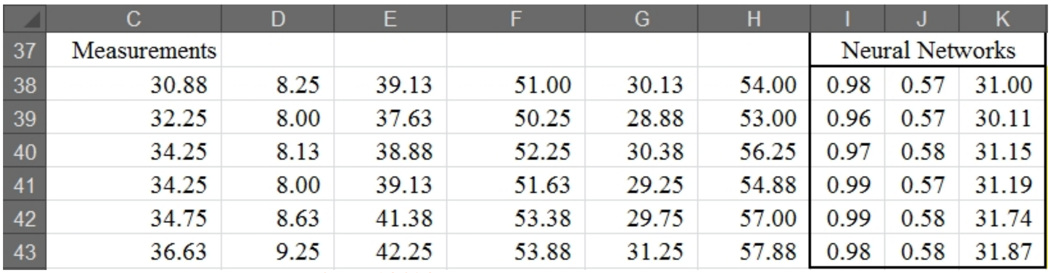


Figure 14.203: **Neural network model vs Measurements**

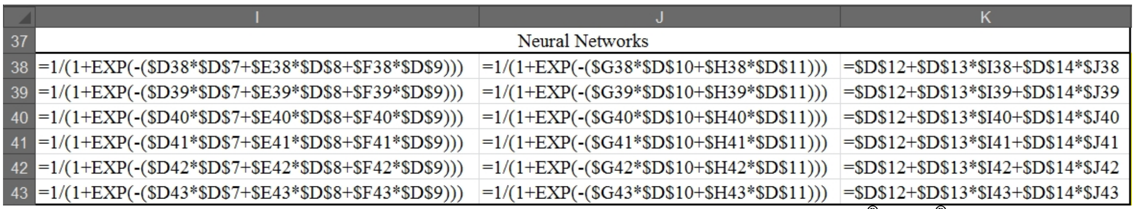
Which seems not bad for a non-linear model! But once the model trained, we must always see if it applies to other data (test sample). Therefore let us consider:



	C	D	E	F	G	H	I	J	K
37	Measurements						Neural Networks		
38	30.88	8.25	39.13	51.00	30.13	54.00	0.98	0.57	31.00
39	32.25	8.00	37.63	50.25	28.88	53.00	0.96	0.57	30.11
40	34.25	8.13	38.88	52.25	30.38	56.25	0.97	0.58	31.15
41	34.25	8.00	39.13	51.63	29.25	54.88	0.99	0.57	31.19
42	34.75	8.63	41.38	53.38	29.75	57.00	0.99	0.58	31.74
43	36.63	9.25	42.25	53.88	31.25	57.88	0.98	0.58	31.87

Figure 14.204: **Neural network test data sample**

Always with the same formulas:



	I	J	K
37	Neural Networks		
38	=1/(1+EXP(-($D38*$D$7+$E38*$D$8+$F38*$D$9)))	=1/(1+EXP(-($G38*$D$10+$H38*$D$11)))	=$D$12+$D$13*$I38+$D$14*$J38
39	=1/(1+EXP(-($D39*$D$7+$E39*$D$8+$F39*$D$9)))	=1/(1+EXP(-($G39*$D$10+$H39*$D$11)))	=$D$12+$D$13*$I39+$D$14*$J39
40	=1/(1+EXP(-($D40*$D$7+$E40*$D$8+$F40*$D$9)))	=1/(1+EXP(-($G40*$D$10+$H40*$D$11)))	=$D$12+$D$13*$I40+$D$14*$J40
41	=1/(1+EXP(-($D41*$D$7+$E41*$D$8+$F41*$D$9)))	=1/(1+EXP(-($G41*$D$10+$H41*$D$11)))	=$D$12+$D$13*$I41+$D$14*$J41
42	=1/(1+EXP(-($D42*$D$7+$E42*$D$8+$F42*$D$9)))	=1/(1+EXP(-($G42*$D$10+$H42*$D$11)))	=$D$12+$D$13*$I42+$D$14*$J42
43	=1/(1+EXP(-($D43*$D$7+$E43*$D$8+$F43*$D$9)))	=1/(1+EXP(-($G43*$D$10+$H43*$D$11)))	=$D$12+$D$13*$I43+$D$14*$J43

Figure 14.205: **Neural network test data sample neural network formula in Microsoft® Excel® 14.0.6123**

And if we also graphically compare the real data and the modelled data, we get:

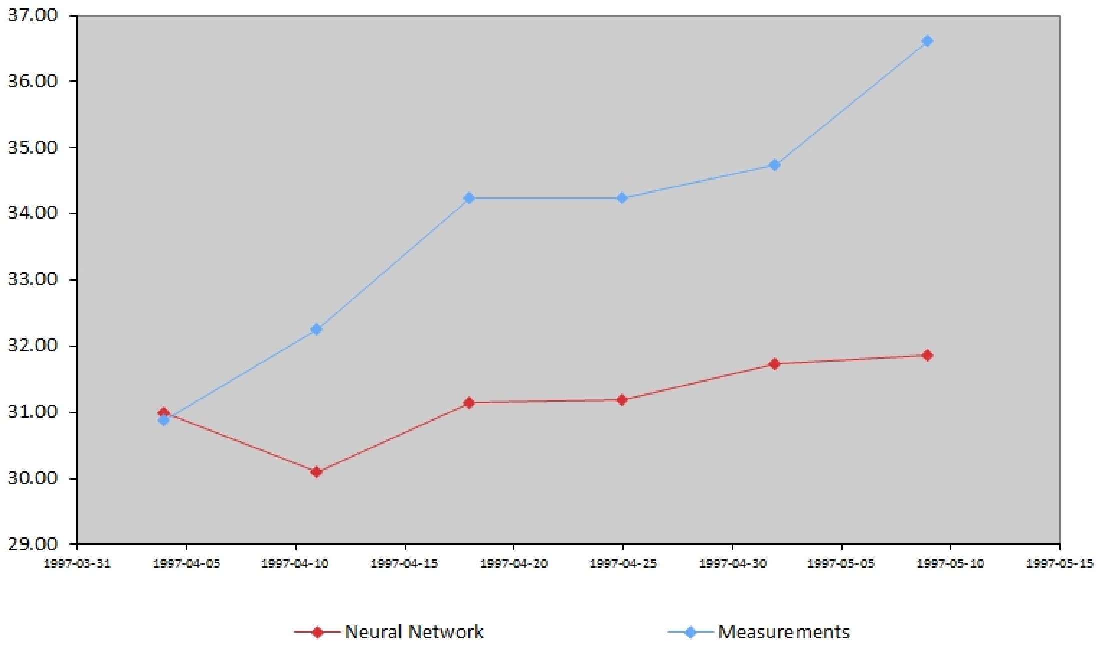


Figure 14.206: **Neural network test data sample neural network formula in Microsoft® Excel® 14.0.6123**

and here we see that the model is significantly worse. But it is so! The predictive science is not an exact science but a heuristic...

In the MATLAB® companion book, working with the same data and building a similar neural network:

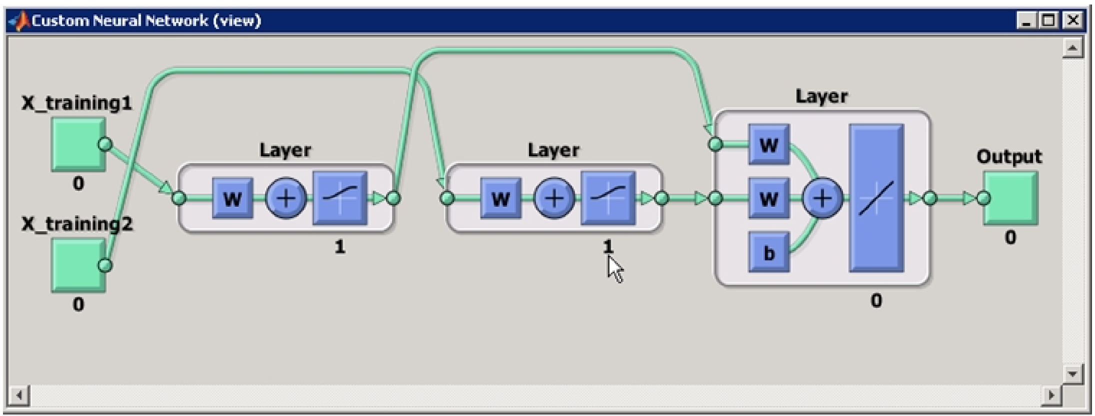

Custom Neural Network (view)
X_training1
0
X_training2
0
Layer
W
1
Layer
W
1
Layer
W
W
b
0
Output
0

Figure 14.207: **Neural network in MATLAB® 2013a**

we get the following model fitting plot:

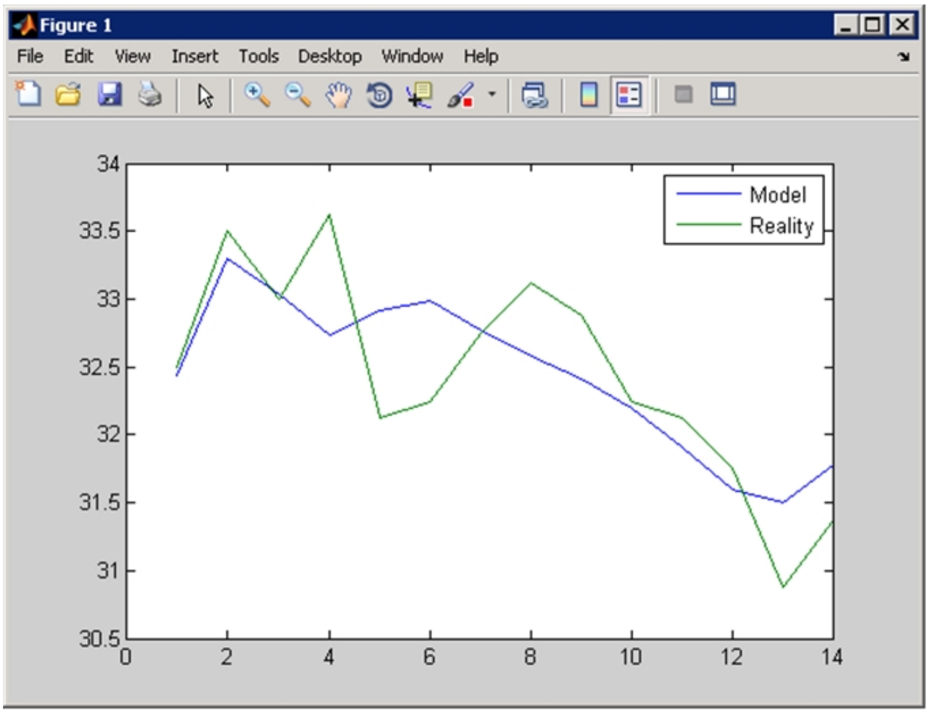


Figure 14.208: **Neural network model vs Measurements in MATLAB® 2013a**

and test sample plot (MATLAB® performs better than Microsoft® Excel® solver in this special case):

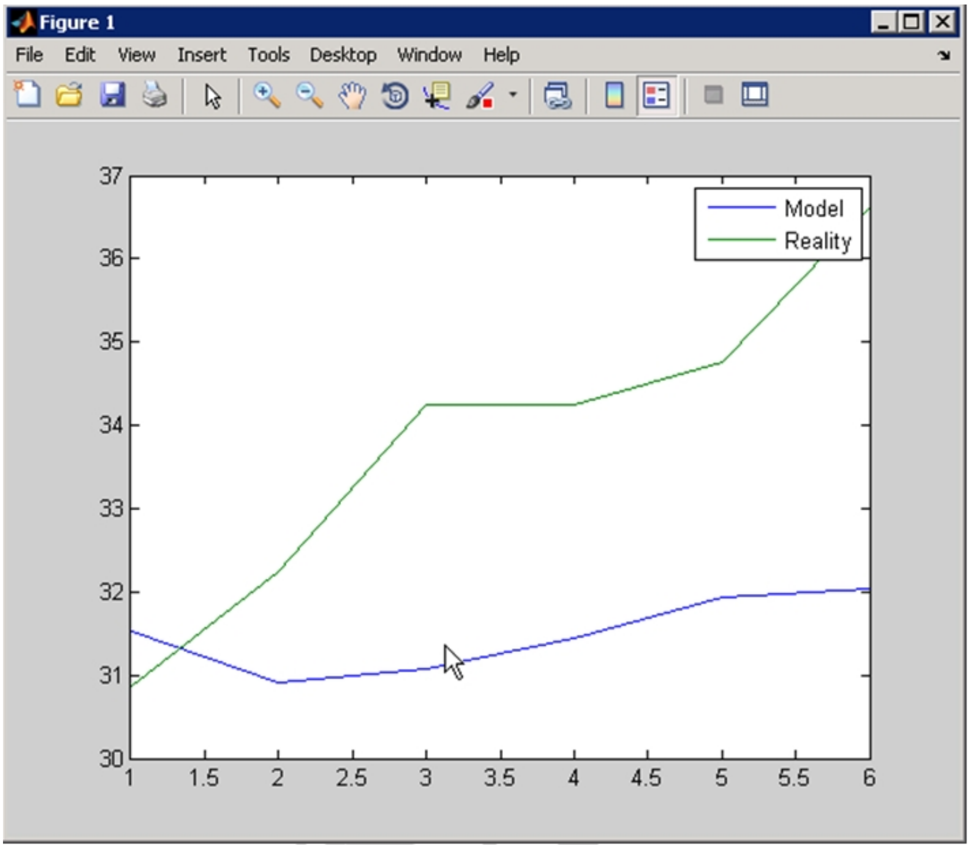


Figure 14.209: **Neural network model test sample plot in MATLAB® 2013a**

Now that we have some basic knowledge of perceptrons and vanilla neural networks we can introduce the mathematical definition of a "neural network" as it should be crystal clear enough for the reader!

Definition (#14.1.573): Let $d, L \in \mathbb{N}$, $N = (N_0, N_\ell, \cdots, N_L) \in \mathbb{N}^{L+1}$ and $\sigma : \mathbb{R} \to \mathbb{R}$. We say that σ is an "activation function", L is the number of layers, and $N_0, N_L, N_\ell, \ell \in [L-1]$ as number of neurons in the input, output, and ℓ-th hidden layer, respectively. Let $\theta = \left(\theta^{(\ell)}\right)_\ell$ be the parameters such that:

$$\theta^{(\ell)} = \left(W^{(\ell)}, \vec{b}^{(\ell)}\right)_\ell \tag{14.1.2077}$$

We denote the number of parameters by $P_N = \sum_{\ell=1}^{L} N_\ell N_{\ell-1} + N_\ell$. Define the corresponding realization function $\vec{\Phi}_\alpha : \mathbb{R}^{N_0} \times \mathbb{R}^{P_N} \to \mathbb{R}^{N_L}$ which satisfies, for any input $\vec{x} \in \mathbb{R}^{N_0}$ and parameters θ, we set $\vec{\Phi}_\alpha = \vec{\Phi}^{(L)}$ where $\alpha = (N, \sigma)$, then:

$$\begin{aligned} \vec{\Phi}^{(1)}(\vec{x}, \theta) &= W^{(1)} \circ \vec{x} + \vec{b}^{(1)} \\ \vec{\bar{\Phi}}^{(\ell)}(\vec{x}, \theta) &= \sigma\left(\vec{\Phi}^{(\ell)}(\vec{x}, \theta)\right), \text{ for } \ell \in [L-1] \\ \vec{\Phi}^{(\ell+1)}(\vec{x}, \theta) &= W^{(\ell+1)} \circ \vec{\bar{\Phi}}^{(\ell)}(\vec{x}, \theta) + \vec{b}^{(\ell+1)} \end{aligned} \tag{14.1.2078}$$

and σ applied component-wise. As we already know, we refer W as the "weight matrix" and $\vec{b}$ as the "bias vector".

Note that $W^{(\ell)} \in \mathbb{R}^{N_\ell} \times \mathbb{R}^{N_{\ell-1}}$ are matrices, so they represent a linear transformation from $\mathbb{R}^{N_\ell}$ to $\mathbb{R}^{N_{\ell-1}}$. Therefore we may see the Neural Networks as successive composition of affine linear transformations, that is:

$$\vec{x} \mapsto W^{(\ell)} \circ \vec{x} + \vec{b}^{(\ell)} \tag{14.1.2079}$$

Now we can also provide the formal definition of a set of set neural networks.

Definition (#14.1.574): Let $\alpha = (N, \sigma)$ be a neural network with input $N_0 = d$ and $N_L = 1$, and activation function σ. The "set of Neural Networks: is defined by:

$$\mathcal{F}_\alpha = \mathcal{F}_{(N,\sigma)} = \left\{\vec{\Phi}_\alpha(\cdot, \theta) : \theta \in \mathbb{R}^{P_N}\right\} \tag{14.1.2080}$$

There is a more simple way to write the two previous definitions, in terms of activation function and affine transformations. Since the realization of a neural network is given by recursively applying σ we may write as:

$$\vec{x} \mapsto F(x) := \vec{T}_L\left(\sigma\left(\vec{T}_{L-1}\left(\cdots\left(\sigma\left(\vec{T}_1(x)\right)\cdots\right)\right)\right)\right) \tag{14.1.2081}$$

where $\vec{T}_L$ is the corresponding affine transformation, i.e. $\vec{T}_\ell = W^{(\ell)} \circ \vec{x} + \vec{b}^{(\ell)}$. Then previous definition of a set of neural network basically becomes the set of all functions $F(x)$ of the form which is described above, that is $F \in \mathcal{F}_{(N,\sigma)}$.

14.1.20.8.5 Backpropagation

Gradient descent (see earlier above page 6678) is a good technique to minimize a cost function. But if we cannot analytically compute the function local derivatives, we will instead have to compute the slope manually by evaluating the function around a neighbourhood. Translated to the example of descending a mountain, it is like either walking with eyes wide open (analytical way), or being blindfolded and having to walk a few steps in each direction, each time we want to assess the slope around us (manual way). This manual option is very time consuming, especially if you have many parameters—which is the case with neural networks.

Hopefully, a technique was found to analytically compute the local derivatives in a neural network: back-propagation. The idea is that a neural network can be used in two directions:

1. Forward propagation where the neurons propagate information through the network by applying their activation functions to their weighted inputs and biases.
2. Backpropagation where the neurons propagate their local derivatives (that is the impact of changing weights, input values, and biases). In other words, each neuron will propagate how much a change in one of its inputs (weights, input values, bias) will impact its own output.

During the training phase of a neural network, forward propagation and back-propagation will be performed in turns. Forward propagation will be used to generate predictions (as data flow through the network). Backpropagation will be used to tune the weights based on the latest errors.

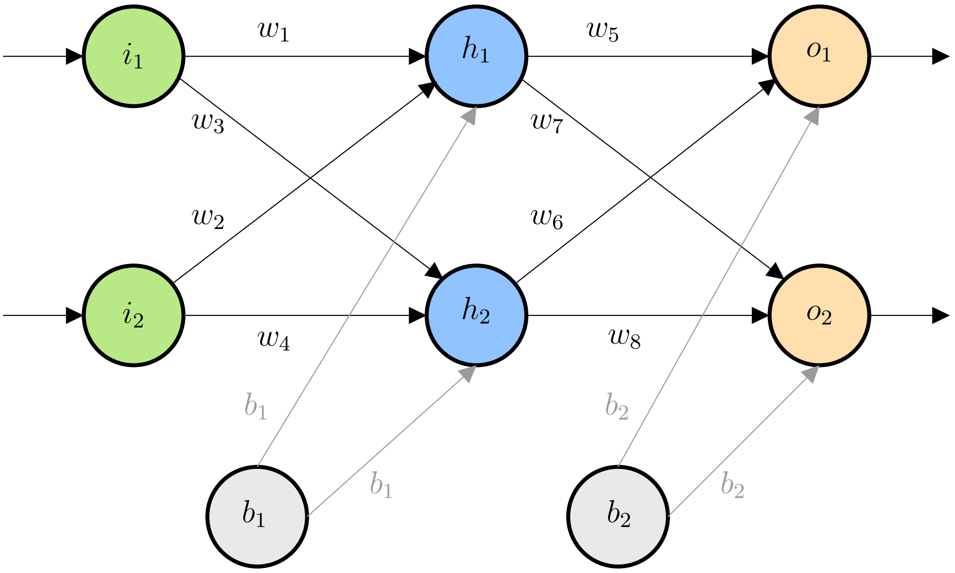
Therefore in the field of neural networks "back-propagation" is a method to calculate the gradient of the loss function with respect to the weights in an artificial neural network. It is commonly used as a part of algorithms that optimize the performance of the network by adjusting the weights, for example in the gradient descent algorithm (see page 6678). It is also named "backward propagation of errors". It seems that there is no shortage of papers online that attempt to explain how back-propagation works, but few that include an example with actual numbers. The text and example below is an attempt of Matt Mazur[72] to explain how it works with a concrete example that everybody can compare their own calculations to in order to ensure they understand backpropagation correctly.

To explain backpropagation, we're going to use a neural network with two inputs, two hidden neurons, two output neurons. Additionally, the hidden and output neurons will include a bias.

Here's the basic structure:

---

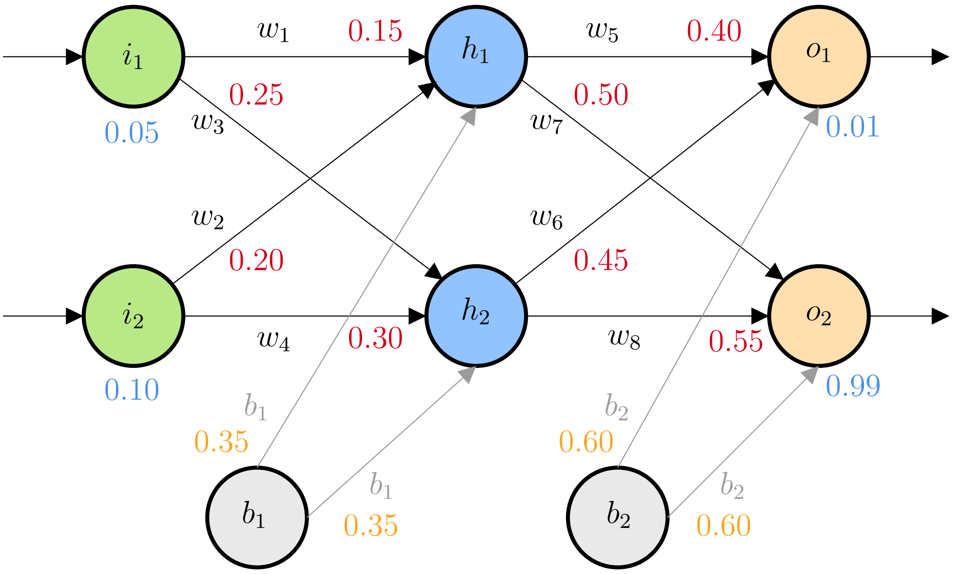
[72] Reproduced from `https://mattmazur.com/2015/03/17/a-step-by-step-backpropagation-example/` with its authorization



i_1
i_2
h_1
h_2
o_1
o_2
b_1
b_2
w_1
w_2
w_3
w_4
w_5
w_6
w_7
w_8

Figure 14.210: **Basis neural network for backpropagation study**

In order to have some numbers to work with, here are the initial weights (in red), the biases (in orange), and training inputs/outputs (in blue):



i_1
0.05
i_2
0.10
h_1
h_2
o_1
0.01
o_2
0.99
w_1 0.15
w_3 0.25
w_2 0.20
w_4 0.30
w_5 0.40
w_7 0.50
w_6 0.45
w_8 0.55
b_1 0.35
b_2 0.60

The goal of backpropagation is to optimize the weights so that the neural network can learn how to correctly map arbitrary inputs to outputs.

For the rest of this tutorial we're going to work with a single training set: given inputs 0.05 and 0.10 and $b_1 = 1$ and $b_2 = 2$, we want the neural network to output 0.01 and 0.99.

Let us decompose the method:

1. The Forward Pass:

   To begin, lets see what the neural network currently predicts given the weights and biases above and inputs of 0.05 and 0.10. To do this we will feed those inputs forward though the network.

We figure out the total net input[73] to each hidden layer neuron, squash the total net input using an activation function (here we use the logistic function), then repeat the process with the output layer neurons.

Here is how we calculate the total net input for h_1:

$$\text{net}_{h_1}=w_1 i_1+w_2 i_2+b_1\cdot 1=0.15\cdot 0.05+0.2\cdot 0.1+0.35\cdot 1=0.3375 \tag{14.1.2082}$$

We then squash it using (arbitrarily) using the logistic function to get the output of h_1:

$$\text{out}_{h_1}=\frac{1}{1+e^{\text{net}_{h_1}}}=\frac{1}{1+e^{-0.3775}}=0.593269992 \tag{14.1.2083}$$

Carrying out the same process for h_2 we get:

$$\text{out}_{h_2}=0.596884378 \tag{14.1.2084}$$

We repeat this process for the output layer neurons, using the output from the hidden layer neurons as inputs.

Here's the output for o_1:

$$\begin{aligned}\text{net}_{o_1}&=w_5\text{out}_{h_1}+w_6\text{out}_{h_2}+b_2\cdot 1\\&=0.4\cdot 0.593269992+0.45\cdot 0.596884378+0.6\cdot 1\\&=1.105905967\end{aligned} \tag{14.1.2085}$$

Hence:

$$\text{out}_{o_1}=\frac{1}{1+e^{-\text{net}_{o_1}}}=\frac{1}{1+e^{-1.105905967}}=0.75136507 \tag{14.1.2086}$$

And carrying out the same process for o_2 we get:

$$\text{out}_{o_2}=0.772928465 \tag{14.1.2087}$$

We can now calculate the error for each output neuron using the squared error function (a loss function) and sum them to get the total error (using the traditional notations in the field...), i.e. a cost function:

$$E_{\text{tot}}=\sum\frac{1}{2}(\text{target}-\text{output})^2 \tag{14.1.2088}$$

Or if written in another common way:

$$J(\vec{w},\vec{b})=\frac{1}{2}\sum_i(y_i-\phi_i(z))^2=\frac{1}{2}\sum_i\left(y_i-\phi_i(\vec{w}^T\vec{x}+\vec{b})\right)^2 \tag{14.1.2089}$$

Since backpropagation uses the gradient descent method (see page 6678), the factor of $\frac{1}{2}$ is included to cancel the exponent when differentiating. Later, the expression will be multiplied with an arbitrary learning rate, so that it doesn't matter if a constant coefficient is introduced now.

---

[73] The "Total net input" is also referred to as just "net input" by some sources.

Remark

Some textbooks refer to the "target" as the "ideal" and the "output" as the "actual".

For example, the target output for o_1 is 0.01 but the neural network output 0.75136507, therefore its error is:

$$E_{o_1}=\frac{1}{2}(\text{target}_{o_1}-\text{out}_{o_1})^2=\frac{1}{2}(0.01-0.75136507)^2=0.274811083 \qquad (14.1.2090)$$

Repeating this process for o_2 (remembering that the target is 0.99) we get:

$$E_{o_2}=\frac{1}{2}(\text{target}_{o_2}-\text{out}_{o_2})^2=0.023560026 \qquad (14.1.2091)$$

The total error for the neural network is the sum of these errors:

$$E_{\text{tot}}=E_{o_1}+E_{o_2}=0.274811083+0.023560026=0.298371109 \qquad (14.1.2092)$$

2. The Backwards Pass: Our goal with backpropagation is to update each of the weights in the network so that they cause the actual output to be closer the target output, thereby minimizing the error for each output neuron and the network as a whole.

   Consider w_5. We want to know how much a change in w_5 affects the total error, aka $\frac{\partial E_{\text{tot}}}{\partial w_5}$.

   By applying the chain rule twice (see section Differential and Integral Calculus page 1483) we know that:

$$\frac{\partial E_{\text{tot}}}{\partial w_5}=\frac{\partial E_{\text{tot}}}{\partial \text{out}_{o_1}}\frac{\partial \text{out}_{o_1}}{\partial \text{net}_{o_1}}\frac{\partial \text{net}_{o_1}}{\partial w_5} \qquad (14.1.2093)$$

   We need to figure out each piece in this equation.

   - First, how much does the total error change with respect to the output? As:

$$E_{\text{tot}}=\frac{1}{2}(\text{target}_{o_1}-\text{out}_{o_1})^2+\frac{1}{2}(\text{target}_{o_2}-\text{out}_{o_2})^2 \qquad (14.1.2094)$$

     Therefore:

$$\begin{aligned}\frac{\partial E_{\text{tot}}}{\partial \text{out}_{o_1}}&=2\cdot\frac{1}{2}(\text{target}_{o_1}-\text{out}_{o_1})^{2-1}\cdot(-1)+0=-(\text{target}_{o_1}-\text{out}_{o_1})\\&=-(0.01-0.75136507)=0.74136507\end{aligned} \qquad (14.1.2095)$$

   - Next, how much does the output of o_1 change with respect to its total net input? As:

$$\text{out}_{o_1}=\frac{1}{1+e^{-\text{net}_{o_1}}} \qquad (14.1.2096)$$

     Therefore:

$$\begin{aligned}\frac{\partial \text{out}_{o_1}}{\partial \text{net}_{o_1}}&=-\frac{e^{-\text{out}_{o_1}}\ln(e)}{(1-e^{-\text{out}_{o_1}})^2}=-\frac{e^{-\text{out}_{o_1}}}{(1-e^{-\text{out}_{o_1}})^2}=\text{out}_{o_1}(1-\text{out}_{o_1})\\&=0.75136507(1-0.75136507)=0.186815602\end{aligned} \qquad (14.1.2097)$$

- Finally, how much does the total net input of o_1 change with respect to w_5?

  As:

$$\text{net}_{o_1} = w_5 \text{out}_{h_1} + w_6 \text{out}_{h2} + b_2 \cdot 1 \tag{14.1.2098}$$

  Therefore:

$$\frac{\partial \text{net}_{o_1}}{\partial w_5} = 1 \cdot \text{out}_{h_1} \cdot w_5^{(1-1)} + 0 + 0 = \text{out}_{h_1} = 0.593269992 \tag{14.1.2099}$$

Putting it all together:

$$\begin{aligned}\frac{\partial E_{\text{tot}}}{\partial w_5} &= \frac{\partial E_{\text{tot}}}{\partial \text{out}_{o_1}} \frac{\partial \text{out}_{o_1}}{\partial \text{net}_{o_1}} \frac{\partial \text{net}_{o_1}}{\partial w_5} = \frac{\partial E_{\text{tot}}}{\partial w_5} = -(\text{target}_{o_1} - \text{out}_{o_1})\text{out}_{o_1}(1-\text{out}_{o_1})\text{out}_{h_1} \\ &= 0.74136507 \cdot 0.186815602 \cdot 0.593269992 = 0.082167041\end{aligned} \tag{14.1.2100}$$

To decrease the error, we then subtract this value from the current weight, by applying the relation derived earlier above (see page 6678):

$$\vec{\omega}^+ = \vec{\omega} + \Delta\omega = \vec{\omega} - \eta \vec{\nabla} E_{\text{tot}}(\vec{\omega}) \tag{14.1.2101}$$

optionally multiplied by some learning rate, η, which we will set to 0.5:

$$w_5^+ = w_5 - \eta \frac{\partial E_{\text{tot}}}{\partial w_5} = 0.4 - 0.5 \cdot 0.082167041 = 0.35891648 \tag{14.1.2102}$$

Remark

Some textbooks use the notation α to represent the learning rate, others use η, and others even use ε...

We can repeat this process to get the new weights w_6, w_7, and w_8:

$$w_6^+ = 0.408666186 \qquad w_7^+ = 0.511301270 \qquad w_8^+ = 0.561370121 \tag{14.1.2103}$$

We perform the actual updates in the neural network after we have the new weights leading into the hidden layer neurons (i.e., we use the original weights, not the updated weights, when we continue the backpropagation algorithm below)!

Next, we will continue the backwards pass by calculating new values for w_1, w_2, w_3, and w_4 of the hidden layer.

Here is what we need to figure out:

$$\frac{\partial E_{\text{tot}}}{\partial w_1} = \frac{\partial E_{\text{tot}}}{\partial \text{out}_{h_1}} \frac{\partial \text{out}_{h_1}}{\partial \text{net}_{h_1}} \frac{\partial \text{net}_{h_1}}{\partial w_1} \tag{14.1.2104}$$

We are going to use a similar process as we did for the output layer, but slightly different to account for the fact that the output of each hidden layer neuron contributes to the output (and therefore error) of multiple output neurons.

We know that out_{h_1} affects both out_{o_1} and out_{o_2} therefore the $\frac{\partial E_{\text{tot}}}{\partial \text{out}_{h_1}}$ needs to take into consideration its effect on the both output neurons:

$$\frac{\partial E_{\text{tot}}}{\partial \text{out}_{h_1}} = \frac{\partial E_{o_1}}{\partial \text{out}_{h_1}} + \frac{\partial E_{o_2}}{\partial \text{out}_{h_1}} \tag{14.1.2105}$$

Starting with $\frac{\partial E_{o_1}}{\partial \text{out}_{h_1}}$:

$$\frac{\partial E_{o_1}}{\partial \text{out}_{h_1}} = \frac{\partial E_{o_1}}{\partial \text{net}_{o_1}} \frac{\partial \text{net}_{o_1}}{\partial \text{out}_{h_1}} \tag{14.1.2106}$$

We can calculate $\frac{\partial E_{o_1}}{\partial \text{net}_{o_1}}$ using values we calculated earlier:

$$\frac{\partial E_{o_1}}{\partial \text{net}_{o_1}} = \frac{\partial E_{o_1}}{\partial \text{out}_{o_1}} \frac{\partial \text{out}_{o_1}}{\partial \text{net}_{o_1}} = 0.74136507 \cdot 0.186815602 = 0.138498562 \tag{14.1.2107}$$

And $\frac{\partial \text{net}_{o_1}}{\partial \text{out}_{h_1}}$ is equal to w_5. Indeed:

$$\text{net}_{o_1} = w_5 \text{out}_{h_1} + w_6 \text{out}_{h_2} + b_2 \cdot 1 \tag{14.1.2108}$$

Therefore:

$$\frac{\partial \text{net}_{o_1}}{\partial \text{out}_{h_1}} = w_5 = 0.40 \tag{14.1.2109}$$

Plugging them in:

$$\frac{\partial E_{o_1}}{\partial \text{out}_{h_1}} = \frac{\partial E_{o_1}}{\partial \text{net}_{o_1}} \frac{\partial \text{net}_{o_1}}{\partial \text{out}_{h_1}} = 0.138498562 \cdot 0.40 = 0.055399425 \tag{14.1.2110}$$

Following the same process for $\frac{\partial E_{o_2}}{\partial \text{out}_{h_1}}$, we get:

$$\frac{\partial E_{o_2}}{\partial \text{out}_{h_1}} = -0.019049119 \tag{14.1.2111}$$

Therefore:

$$\begin{aligned}\frac{\partial E_{\text{tot}}}{\partial \text{out}_{h_1}} &= \frac{\partial E_{o_1}}{\partial \text{out}_{h_1}} + \frac{\partial E_{o_2}}{\partial \text{out}_{h_1}} = \frac{\partial E_{o_1}}{\partial \text{net}_{o_1}} \frac{\partial \text{net}_{o_1}}{\partial \text{out}_{h_1}} + \frac{\partial E_{o_2}}{\partial \text{net}_{o_2}} \frac{\partial \text{net}_{o_2}}{\partial \text{out}_{h_1}} \\ &= 0.055399425 + (-0.019049119) = 0.036350306\end{aligned} \tag{14.1.2112}$$

Now that we have $\frac{\partial E_{\text{tot}}}{\partial \text{out}_{h1}}$, we need to figure out $\frac{\partial \text{out}_{h_1}}{\partial \text{net}_{h1}}$ and then $\frac{\partial \text{net}_{h_1}}{\partial w_1}$.

As:

$$\text{out}_{h_1} = \frac{1}{1 + e^{-\text{net}_{h_1}}} \tag{14.1.2113}$$

Therefore identically as earlier above:

$$\frac{\partial \text{out}_{h_1}}{\partial \text{net}_{h_1}} = \text{out}_{h_1}(1 - \text{out}_{h_1}) = 0.59326999(1 - 0.59326999) = 0.241300709 \tag{14.1.2114}$$

And as:

$$\text{net}_{h_1} = w_1 i_1 + w_2 i_2 + b_1 \cdot 1 \tag{14.1.2115}$$

Therefore:

$$\frac{\partial \text{net}_{h_1}}{\partial w_1} = i_1 = 0.05 \tag{14.1.2116}$$

Putting it all together (and writing all the steps explicitly):

$$\begin{aligned}\frac{\partial E_{\text{tot}}}{\partial w_1}&=\frac{\partial E_{\text{tot}}}{\partial \text{out}_{h_1}}\frac{\partial \text{out}_{h_1}}{\partial \text{net}_{h_1}}\frac{\partial \text{net}_{h_1}}{\partial w_1}=\left(\frac{\partial E_{o_1}}{\partial \text{out}_{h_1}}+\frac{\partial E_{o_2}}{\partial \text{out}_{h_1}}\right)\frac{\partial \text{out}_{h_1}}{\partial \text{net}_{h_1}}\frac{\partial \text{net}_{h_1}}{\partial w_1}\\&=\left(\frac{\partial E_{o_1}}{\partial \text{net}_{o_1}}\frac{\partial \text{net}_{o_1}}{\partial \text{out}_{h_1}}+\frac{\partial E_{o_2}}{\partial \text{net}_{o_2}}\frac{\partial \text{net}_{o_2}}{\partial \text{out}_{h_1}}\right)\frac{\partial \text{out}_{h_1}}{\partial \text{net}_{h_1}}\frac{\partial \text{net}_{h_1}}{\partial w_1}\\&=0.036350306\cdot 0.241300709\cdot 0.05=0.000438568\end{aligned}\tag{14.1.2117}$$

We can now update w_1:

$$w_1^+=w_1-\eta\frac{\partial E_{\text{tot}}}{\partial w_1}=0.15-0.5\cdot 0.000438568=0.149780716\tag{14.1.2118}$$

Repeating this for w_2, w_3, and w_4:

$$w_2^+=0.19956143\qquad w_3^+=0.24975114\qquad w_4^+=0.29950229\tag{14.1.2119}$$

Finally, we have updated all of our weights! When we fed forward the 0.05 and 0.1 inputs originally, the error on the network was 0.298371109. After this first round of backpropagation, the total error is now down to 0.291027924. It might not seem like much, but after repeating this process $10,000$ times, for example, the error plummets to 0.000035085. At this point, when we feed forward 0.05 and 0.1, the two outputs neurons generate 0.015912196 (vs 0.01 target) and 0.984065734 (vs 0.99 target).

Therefore we have seen that this optimization algorithm repeats a two phase cycle, propagation and weight update. When an input vector is presented to the network, it is propagated forward through the network, layer by layer, until it reaches the output layer. The output of the network is then compared to the desired output, using a loss function, and an error value is calculated for each of the neurons in the output layer. The error values are then propagated backwards, starting from the output, until each neuron has an associated error value which roughly represents its contribution to the original output. Backpropagation uses these error values to calculate the gradient of the loss function. In the second phase, this gradient is fed to the optimization method, which in turn uses it to update the weights, in an attempt to minimize the loss function.

We have also seen that the backpropagation requires that the activation function used by the artificial neurons (or "nodes") to be differentiable.

Remarks

R1. Backpropagation requires a known, desired output for each input value in order to calculate the loss function gradient – it is therefore usually considered to be a supervised learning method.

R2. The choice of learning rate η is important for the method, since a high value can cause too strong a change, causing the minimum to be missed, while a too low learning rate slows the training unnecessarily. In order to avoid oscillation inside the network, such as alternating connection weights, and to improve the rate of convergence, there are refinements of this algorithm that use an adaptive learning rate.

R3. Backpropagation learning does not require normalization of input vectors; however, normalization could improve performance.

R4. Over-fitting is a major-concern in deep learning (and to humans also...) since large networks can have hundreds of millions of weights. In image recognition, the number of training images can be significantly increased by random jittering of the images. Another technique named dropout randomly deletes a fraction of the weights at each training iteration. Regularization (see page 6435) is used to assign a cost to the size of weights and many other ideas are being explored.

For information... when the net is passed to the activation (transfer) function and the function's output is used for adjusting the weights we then speak of "McCulloch–Pitts perceptron" or just simply of a "(feed-forward) perceptron". If the weights are adjusted only according to the weighted sum of the inputs (that is without passing through the activation function!), we then speak of "adaptive linear neuron" (ADALINE) or of "feed-forward adaptive linear neuron[74]".

Finally, don't forget (as we have seen earlier) that gradient descent with backpropagation is not guaranteed to find the global minimum of the error function, but only a local minimum!

As the number of layers grew, we encounter mainly two calculatory difficulties (aside computer memory issues): "exploding gradient" and "vanishing gradient".

While the problem of exploding gradient can be treated quite well by applying simple techniques like gradient clipping and L_1 or L_2 regularization (see earlier above page 6441), the problem of vanishing gradient remained intractable for decades.

What is vanishing gradient and why does it arise? To update the values of the parameters in neural networks, we have seen that the back-propagation algorithm was typically used and is based on the chain rule to calculated partial derivatives of some more or less complex functions. During gradient descent, the neural network's parameters receive an update proportional to the derivative of the cost function with respect to the current parameter in each iteration of training. The problem is that in some cases, the gradient will be vanishingly small, effectively preventing some parameters from chaining their value. In the worst case, this may completely stop the neural network from further training.

Traditional activation functions, such as the hyperbolic tangent function i mentioned above, have gradients in the range $[0, 1]$, and back-propagation computes gradients by the chain rule. That has the effect of multiplying n of these small numbers to compute gradients of the earlier (leftmost) layers in an n-layer network, meaning that the gradient decreases exponentially with n. That results in the effect that the earlier layers train very slowly, if at all.

Remark

At this point the reader may have noticed that technically, logistic regression is equivalent to a single layer neural network with a sigmoid transfer function (and vice-versa).

---

[74] A multilayer network of ADALINE units is known as a MADALINE.

14.1.20.8.6 Recursive Neural Networks (RNN)

Sometimes we want to remember an input for later use. There are many examples of such a situation, such as the stock market. To make a good investment judgement, we have to at least look at the stock data from a time window.

The naive way to let neural network accept a time series data is connecting several neural networks together. Each of the neural networks handles one time step. Instead of feeding the data at each individual time step, you provide data at all time steps within a window, or a context, to the neural network.

A "recurrent neural network" (RNN) is a class of artificial neural networks where connections between nodes form a directed graph along a temporal sequence. This allows it to exhibit temporal dynamic behaviour. Derived from feed-forward neural networks, RNNs can use their internal state (memory) to process variable length sequences of inputs.

There are three common type of recursive neural network[75] (recurrent neural network, long short term memory unit and gated recurrent unit):

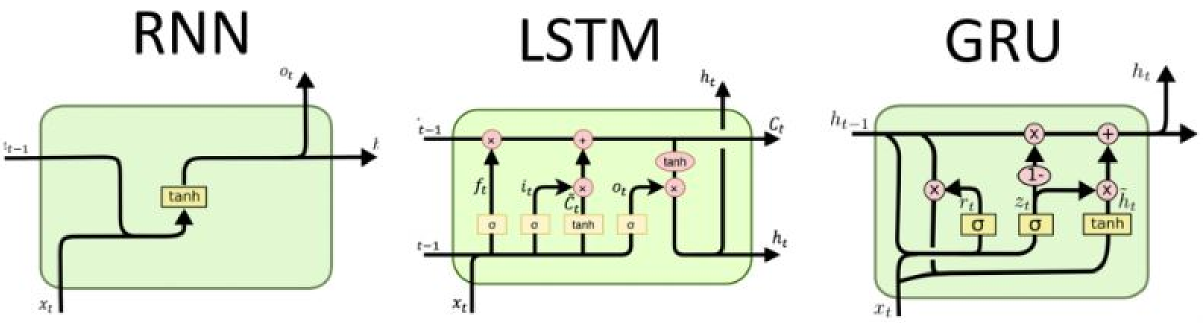


Figure 14.211: **Three common type of recursive neural networks**

As we will provide further below the equations of LSTM (Long-Short Term Memory) and GRU (Gated Recurrent Unit), the reader will be able to notice that GRU (introduced in year 12014 according to holocene calendar) are a special case of LSTM (introduced in year 11997 according to holocene calendar)!

Remark

The GRU is an alternative to the LSTM which is similarly difficult to justify. It seems that comparison of the GRU to the LSTM and its variants provides evidence that the GRU outperforms the LSTM on nearly all tasks except language modelling.

A lot of times, you need to process data that has periodic patterns (LSTM at the day we write these lines seems to be efficient only for such patterns!). As a silly example, suppose you want to predict Christmas tree sales. This is a very seasonal thing and likely to peak only once a year. So a good strategy to predict Christmas tree sale is looking at the data from exactly a year back. For this kind of problems, you either need to have a big context to include ancient data points, or you have a good memory. You know what data is valuable to remember for later use and what needs to be forgotten when it is useless.

---

[75] Huge thanks to Shi Yan, senior software engineer at NVIDIA, for having authorized us to use and slightly modify his LSTM illustrations!

Theoretically the naively connected neural network, so named "recurrent neural network":

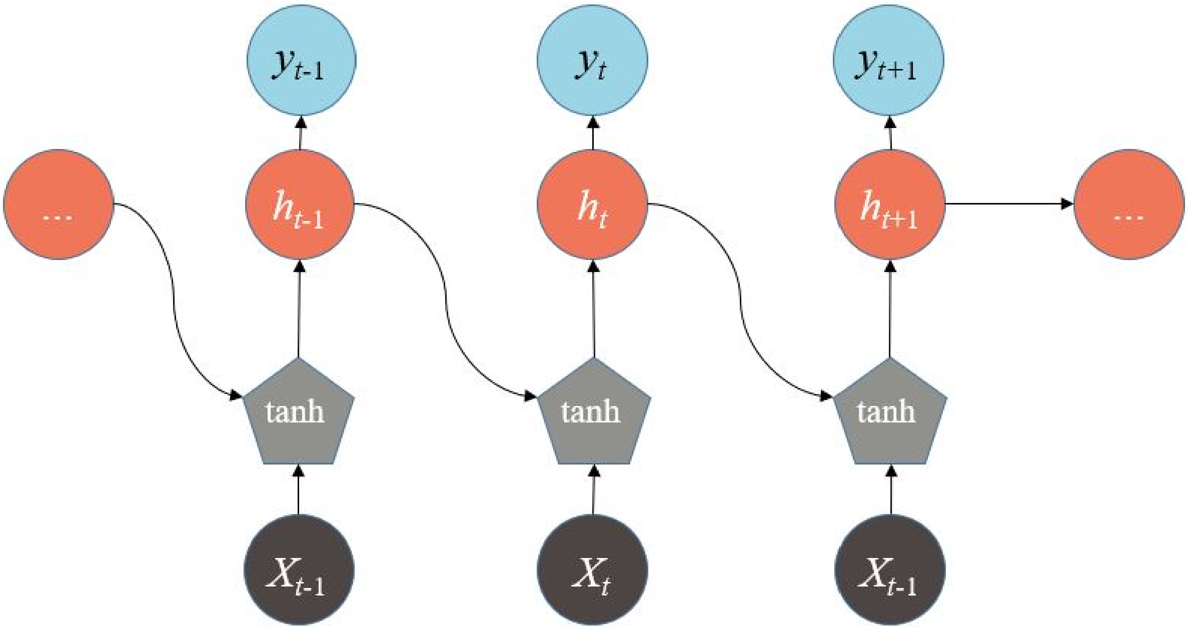


Figure 14.212: **Recurrent neural network**

can work. But in practice, it suffers from two problems: vanishing gradient and exploding gradient, which make it unusable.

Then later,"Long-Short Term memory Neural Networks" (LSTM) were introduced by Hochreiter & Schmidhuber (11997 according to holocene calendar) [255] to solve this issue by explicitly introducing a memory unit, named the "cell" into the network. This is the diagram of a LSTM building block:

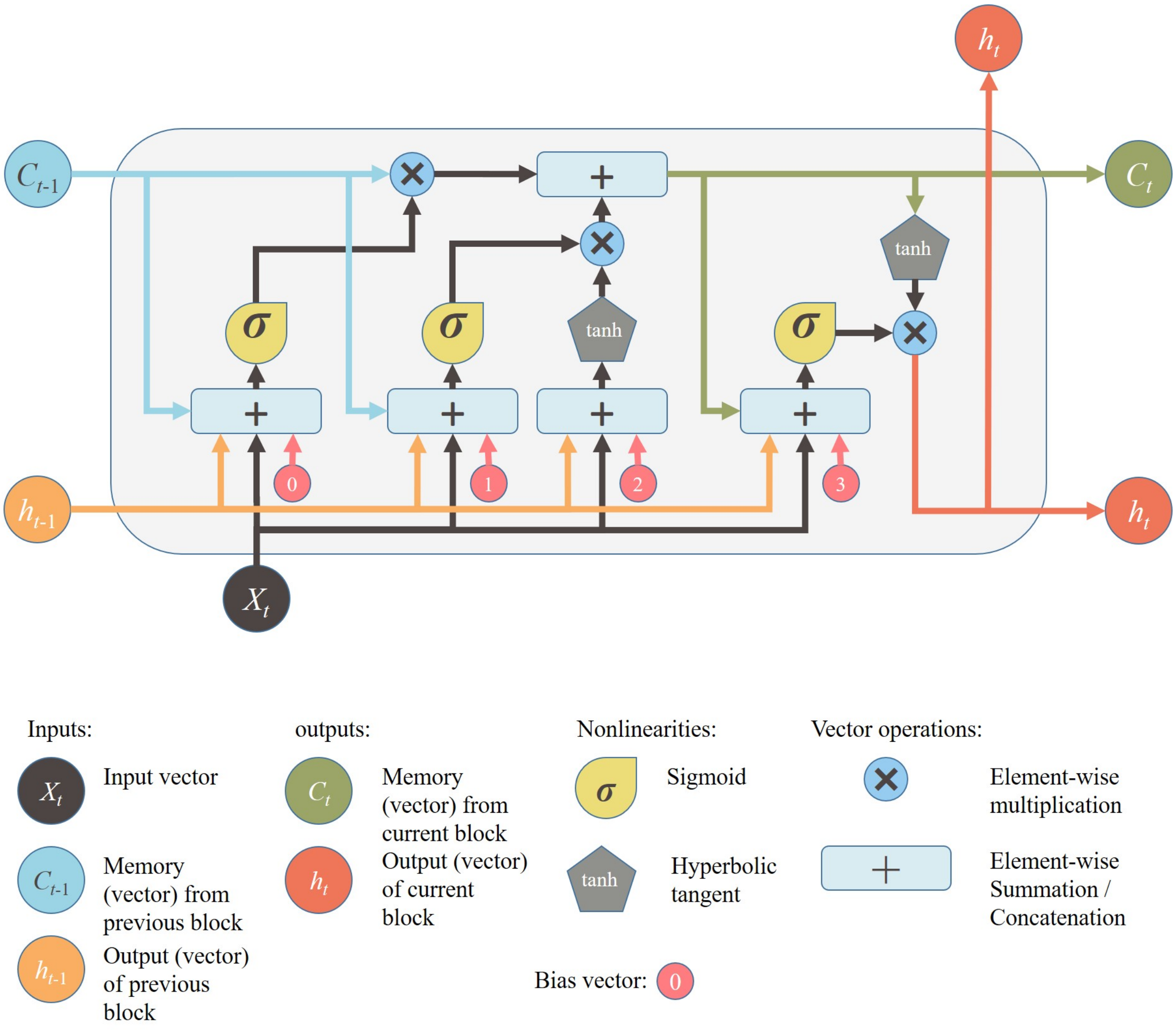

h_t
C_{t-1}
C_t
σ
σ
tanh
σ
tanh
+
+
+
+
+
0
1
2
3
h_{t-1}
h_t
X_t
Inputs:
X_t
Input vector
C_{t-1}
Memory (vector) from previous block
h_{t-1}
Output (vector) of previous block
outputs:
C_t
Memory (vector) from current block
h_t
Output (vector) of current block
Nonlinearities:
σ
Sigmoid
tanh
Hyperbolic tangent
Vector operations:
Element-wise multiplication
+
Element-wise Summation / Concatenation
Bias vector:
0

Figure 14.213: LSTM (long short term memory) building block

Obviously in the general case most variables are vectors. But in most typical univariate forecasting techniques, most vectors (excepted $\vec{X}_t$) are one-dimensional vectors: i.e. scalars!

Remark

LSTM networks are a type of RNN that uses special units in addition to standard units. Again (!), standard RNNs (Recurrent Neural Networks) suffer from vanishing and exploding gradient problems. LSTMs (Long Short Term Memory) deal with these problems by introducing new gates, such as input and forget gates, which allow for a better control over the gradient flow and enable better preservation of long-range dependencies.

At a first sight, this looks intimidating. Let's ignore the internals, but only look at the inputs and outputs of the unit. The network takes three inputs. X_t is the input of the current time step. h_{t-1} is the output from the previous LSTM unit and C_{t-1} is the "memory" of the previous unit, which may be seen as the most important input. As for outputs, h_t is the output of the current network. C_t is the memory of the current unit.

Therefore, this single unit makes decision by considering the current input, previous output and previous memory. And it generates a new output and alters its memory.

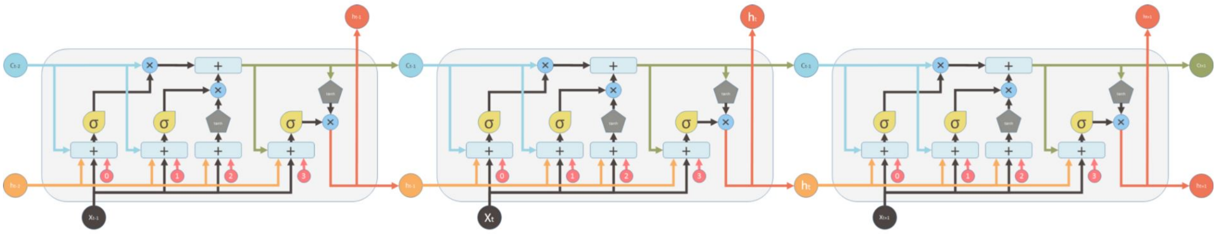


Figure 14.214: **Typical LSTM overall structure**

The way its internal memory C_t changes is pretty similar to piping water through a pipe. Assuming the memory is water, it flows into a pipe. You want to change this memory flow along the way and this change is controlled by two valves:

1. The first valve is named the "forget valve". If you shut it, no old memory will be kept. If you fully open this valve, all old memory will pass through.

2. The second valve is the new "memory valve". New memory will come in through a T shaped joint like below and merge with the old memory. Exactly how much new memory should come in is controlled by the second valve.

On the LSTM diagram below the red area is the memory pipe (do not confuse it with the "memory valve"!). The input is the old memory (a vector). The first cross × it passes through is the forget valve. It is actually an element-wise multiplication operation (i.e. "Hadamard product"). So if you multiply the old memory C_{t-1} with a vector that is close to 0, that means you want to forget most of the old memory. You let the old memory goes through, if you forget valve equals 1.

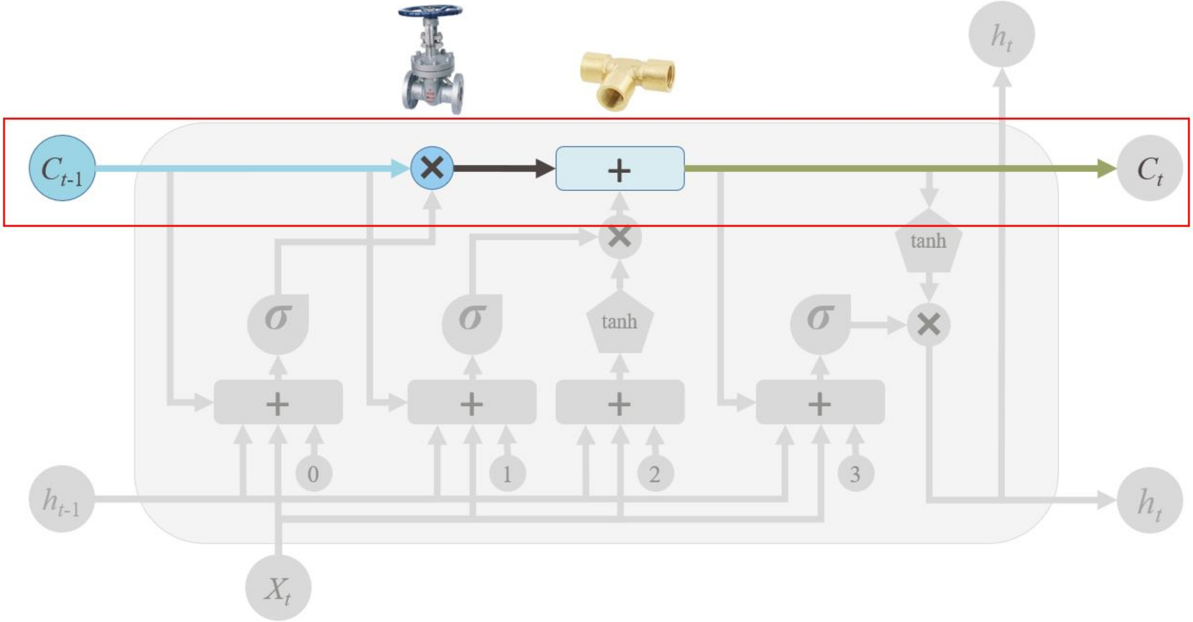


Figure 14.215: **Memory pipe with memory valve and forget valve of a LSTM**

Then the second operation the memory flow will go through is this $+$ operator. This operator means piece-wise summation. It resembles the T shape joint pipe. New memory and the old memory will merge by this operation.

After these two operations, you have the old memory C_{t-1} changed to the new memory C_t.

Now let us look at the input of both valves more in details.

We begin with the "forget valve"! It is controlled by a simple feed-forward one layer neural network. The inputs of the neural network is h_{t-1}, the output of the previous LSTM block, X_t, the input for the current LSTM block, C_{t-1}, the memory of the previous block and finally a bias vector b_0. This neural network has a sigmoid function as activation, and it's output vector is the forget valve, which will applied to the old memory C_{t-1} by element-wise multiplication.

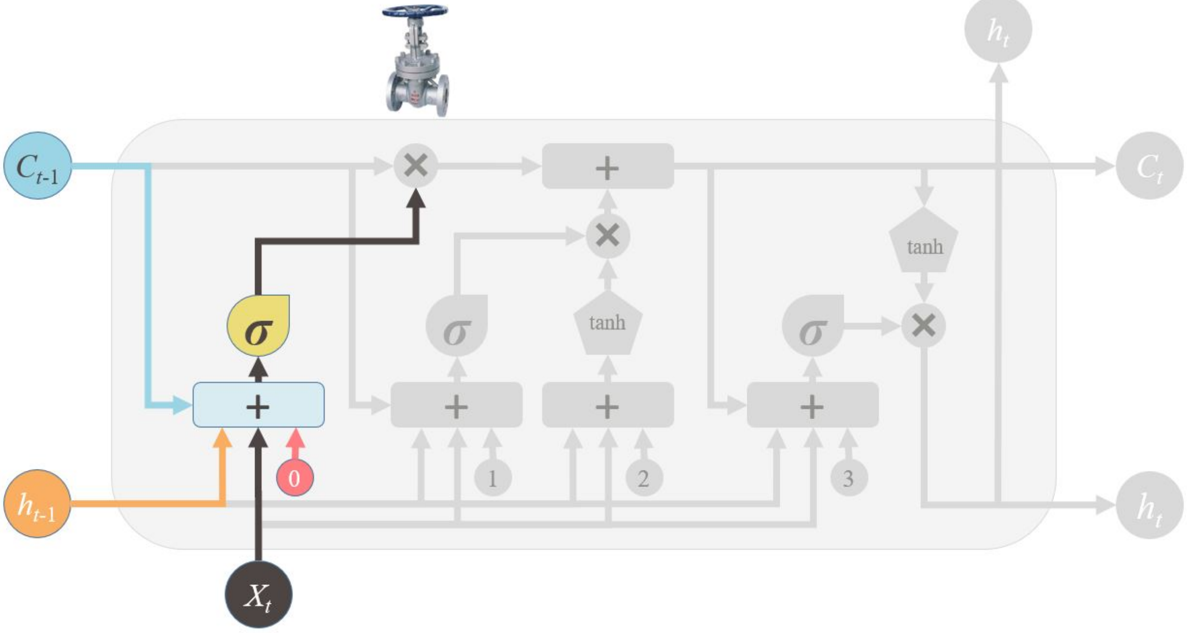

h_t
$C_{t\text{-}1}$
C_t
tanh
σ
tanh
σ
σ
+
+
+
+
+
0
1
2
3
$h_{t\text{-}1}$
h_t
X_t

Figure 14.216: **Details of the forget valve on the memory pipe**

Now the second valve named "memory valve"! Again, it is a one layer simple neural network that takes the same inputs as the forget valve. This valve controls how much the new memory should influence the old memory:

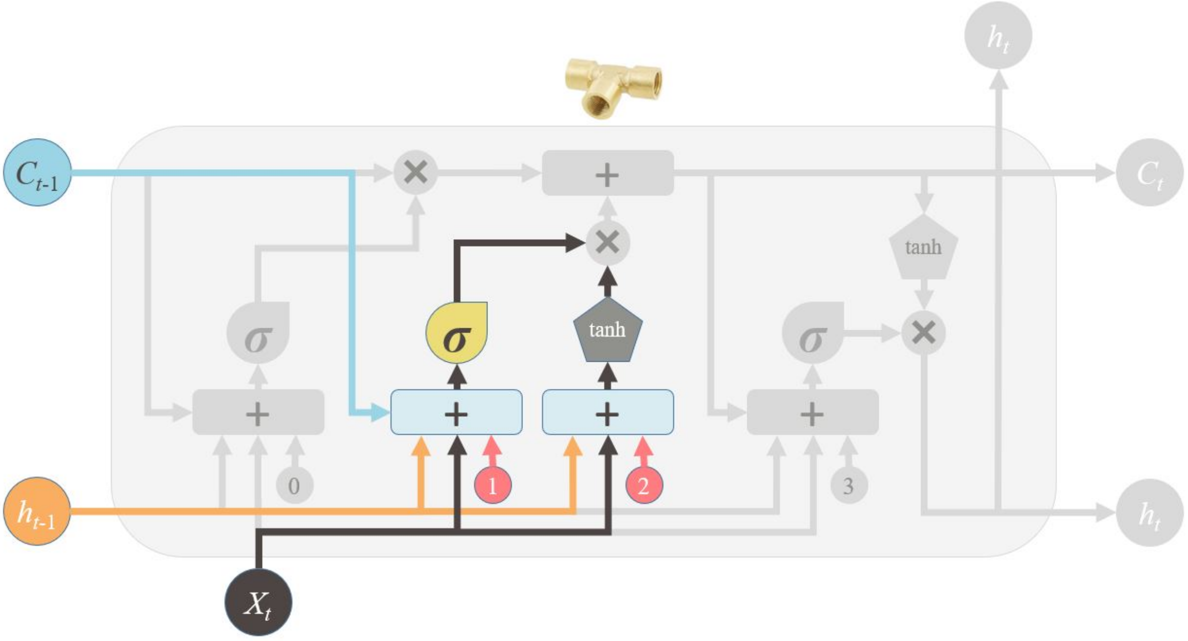
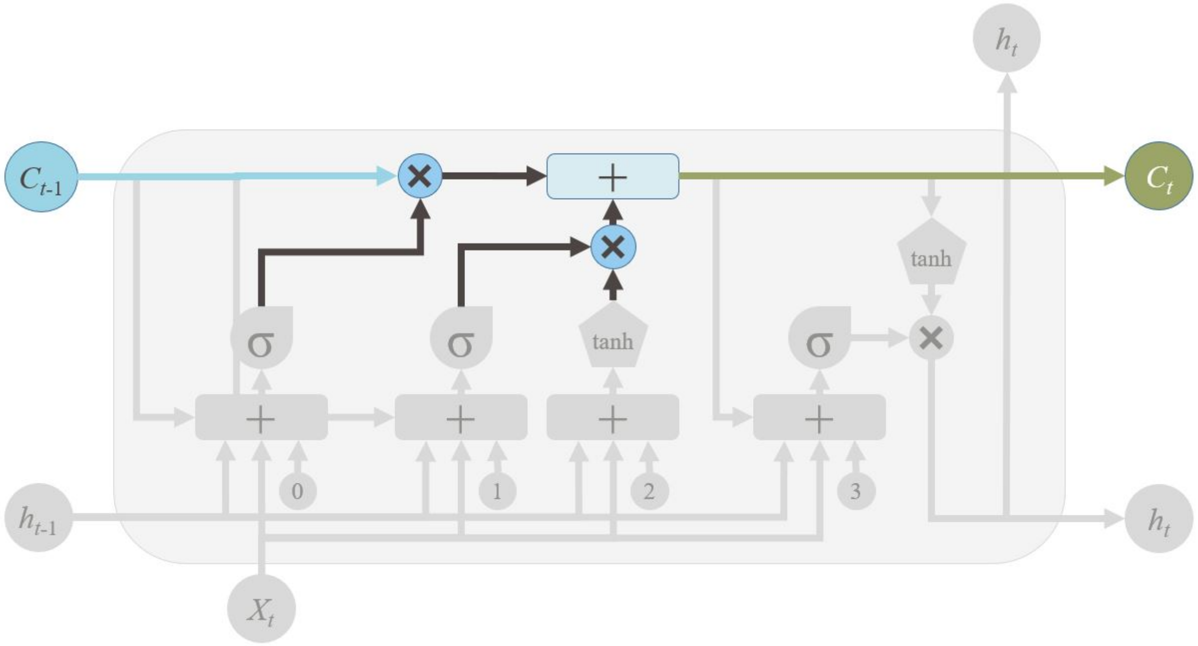
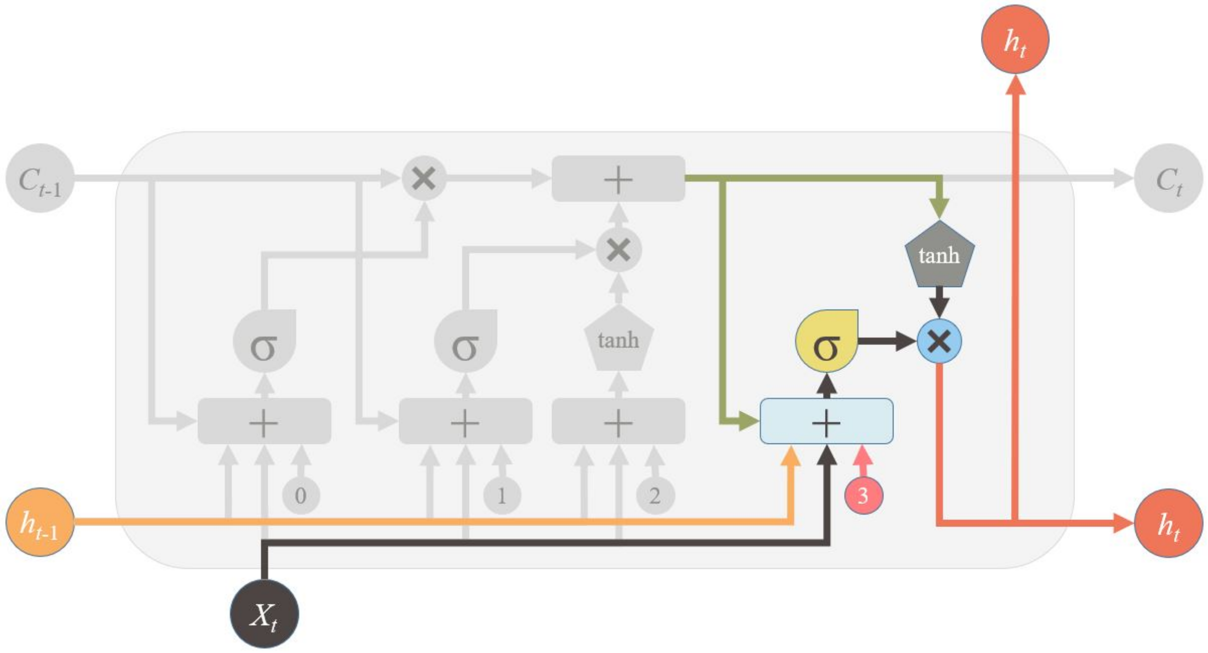

h_t
$C_{t\text{-}1}$
C_t
tanh
σ
σ
tanh
σ
+
+
+
+
+
0
1
2
3
$h_{t\text{-}1}$
h_t
X_t

Figure 14.217: **Details of the memory valve below the memory pipe**

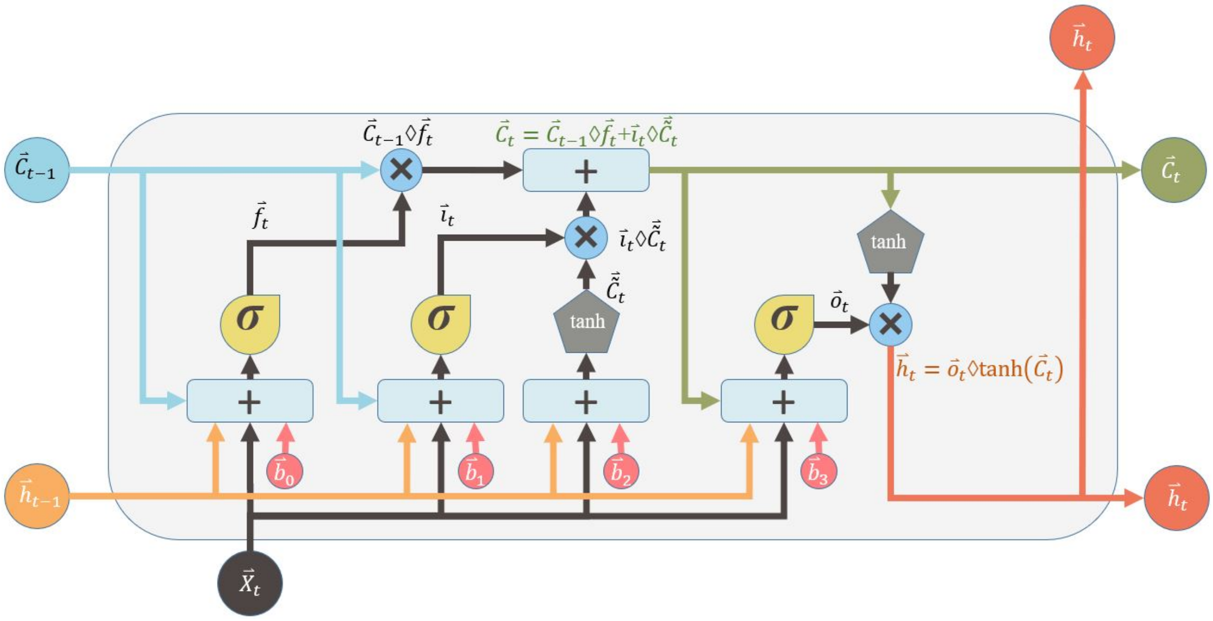
The output of this network will element-wise multiple the new memory valve, and add to the old memory to form the new memory:


h_t
C_{t-1}
C_t
tanh
σ
σ
tanh
σ
+
+
+
+
+
0
1
2
3
h_{t-1}
h_t
X_t

And finally, we need to generate the output for this LSTM unit. This step has an output valve that is controlled by the new memory, the previous output h_{t-1}, the input X_t and a bias vector. This valve controls how much new memory should output to the next LSTM unit:


h_t
C_{t-1}
C_t
tanh
σ
σ
tanh
σ
+
+
+
+
+
0
1
2
3
h_{t-1}
h_t
X_t

Let us see now the mathematical point of view. For this consider again the LSTM block but with more explicit notations (for recall, $\diamond$ is the vector element-wise multiplication):


$\vec{h}_t$
$\vec{C}_{t-1}$
$\vec{C}_{t-1}\Diamond\vec{f}_t$
$\vec{C}_t = \vec{C}_{t-1}\Diamond\vec{f}_t+\vec{\imath}_t\Diamond\vec{\vec{C}}_t$
$\vec{C}_t$
$\vec{f}_t$
$\vec{\imath}_t$
$\vec{\imath}_t\Diamond\vec{\vec{C}}_t$
tanh
$\vec{\vec{C}}_t$
σ
σ
tanh
σ
$\vec{o}_t$
$\vec{h}_t = \vec{o}_t\Diamond\tanh(\vec{C}_t)$
$\vec{h}_{t-1}$
$\vec{b}_0$
$\vec{b}_1$
$\vec{b}_2$
$\vec{b}_3$
$\vec{h}_t$
$\vec{X}_t$

And here are the corresponding relations:

- Gating variables (respectively $\vec{f}_t$ forget gate, $\vec{\imath}_t$ update gate, $\vec{o}_t$ output gate):

$$\begin{aligned}\vec{f}_t&=\sigma\left(W_f\circ[\vec{C}_{t-1},\vec{h}_{t-1},\vec{X}_t]+\vec{b}_0\right)\\ \vec{\imath}_t&=\sigma\left(W_i\circ[\vec{C}_{t-1},\vec{h}_{t-1},\vec{X}_t]+\vec{b}_1\right)\\ \vec{o}_t&=\sigma\left(W_o\circ[\vec{C}_t,\vec{h}_{t-1},\vec{X}_t]+\vec{b}_3\right)\end{aligned}\tag{14.1.2120}$$

- Candidate (memory) cell state:

$$\vec{\vec{C}}_t=\tanh\left(W_c\circ[\vec{h}_{t-1},\vec{X}_t]+\vec{b}_2\right)\tag{14.1.2121}$$

- Cell and hidden state:

$$\begin{aligned}\vec{C}_t&=\vec{C}_{t-1}\diamond\vec{f}_t+\vec{\imath}_t\diamond\vec{\vec{C}}_t\\ &=\vec{C}_{t-1}\diamond\left[\sigma\left(W_f\circ[\vec{C}_{t-1},\vec{h}_{t-1},\vec{X}_t]+\vec{b}_0\right)\right]\\ &+\left[\sigma\left(W_i\circ[\vec{C}_{t-1},\vec{h}_{t-1},\vec{X}_t]+\vec{b}_1\right)\right]\diamond\tanh\left(W_c\circ[\vec{h}_{t-1},\vec{X}_t]+\vec{b}_2\right)\\ \vec{h}_t&=\vec{o}_t\diamond\tanh\left(\vec{C}_t\right)=\vec{o}_t\diamond\tanh\left(\vec{C}_{t-1}\diamond\vec{f}_t+\vec{\imath}_t\diamond\vec{\vec{C}}_t\right)\end{aligned}\tag{14.1.2122}$$

Remark

For a "Gated Recurrent Unit Neural Network" (GRU), only two relations change:

$$\begin{aligned}\vec{C}_t&=\vec{C}_{t-1}\diamond(1-\vec{\imath}_t)+\vec{\imath}_t\diamond\vec{\vec{C}}_t\\ \vec{h}_t&=\vec{C}_t\end{aligned}\tag{14.1.2123}$$

hence the fact that GRU is a special case of LSTM!

The core reason that recurrent nets are more exciting is that they allow us to operate over sequences of vectors: Sequences in the input, the output, or in the most general case both. A few examples may make this more concrete:

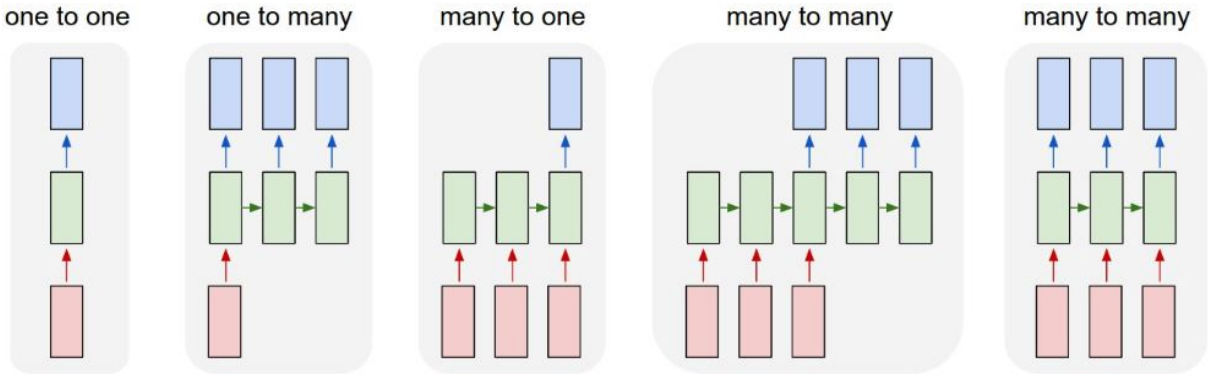


Figure 14.218: **Each rectangle is a vector and arrows represent functions (e.g. matrix multiply). Input vectors are in red, output vectors are in blue and green vectors hold the RNN's state (more on this soon). From left to right: (1) Vanilla mode of processing without RNN, from fixed-sized input to fixed-sized output (e.g. image classification). (2) Sequence output (e.g. image captioning takes an image and outputs a sentence of words). (3) Sequence input (e.g. sentiment analysis where a given sentence is classified as expressing positive or negative sentiment). (4) Sequence input and sequence output (e.g. Machine Translation: a RNN reads a sentence in English and then outputs a sentence in French). (5) Synced sequence input and output (e.g. video classification where we wish to label each frame of the video). Notice that in every case are no pre-specified constraints on the lengths sequences because the recurrent transformation (green) is fixed and can be applied as many times as we like.**

RNNs are often used in text processing because sentences and texts are naturally sequences of either words/punctuation marks or sequences of characters. For the same reason, recurrent neural networks are also used in speech processing.

In the case of a recurrent neural network, the loss function L of all time:

$$L_{\text{RNN}}(\hat{y},y)=\sum_{t=1}^{T}L(\hat{y}_t,y_t) \tag{14.1.2124}$$

And as always we do first a forward pass (first estimation based on initial chosen values) and afterwards we correct the weights with a backward pass done at each point in time (BPTT: Back-propagation Through Time). At time-step T the derivative of the loss L_{RNN} with respect to the different weight matrices W and biases is expressed as follows:

$$\frac{\partial L^{(T)}}{\partial W}=\sum_{t=1}^{T}\left.\frac{\partial L^{(T)}}{\partial W}\right|_{(t)} \qquad \frac{\partial L^{(T)}}{\partial b}=\sum_{t=1}^{T}\left.\frac{\partial L^{(T)}}{\partial b}\right|_{(t)} \tag{14.1.2125}$$

Remember that during our example introducing back-propagation earlier above, we used the chain rule twice. The ideal here would be to write something similar:

$$\frac{\partial L_t^T}{\partial W}=\frac{\partial L_t^T}{\partial W}\frac{\partial \widetilde{C}_T}{\partial h_t}\frac{\partial h_t}{\partial W} \tag{14.1.2126}$$

However this is incomplete as h_t depends on previous $h_{t-\ldots}$. Therefore we should write instead:

$$\begin{aligned}\frac{\partial L_t^T}{\partial W}&=\frac{\partial L_t^T}{\partial W}\frac{\partial \widetilde{C}_T}{\partial h_t}\left(\frac{\partial h_t}{\partial W}+\frac{\partial h_t}{\partial h_{t-1}}\frac{\partial h_{t-1}}{\partial W}+\frac{\partial h_t}{\partial h_{t-1}}\frac{\partial h_{t-2}}{\partial h_{t-2}}\frac{\partial h_{t-2}}{\partial W}+\ldots\right)\\&=\frac{\partial L_t^T}{\partial W}\frac{\partial \widetilde{C}_T}{\partial h_t}\left[\sum_{k=0}^{t}\left(\prod_{i=k+1}^{t}\frac{\partial h_i}{\partial h_{i-1}}\right)\frac{\partial h_k}{\partial W}\right]\end{aligned} \tag{14.1.2127}$$

Remark

A recursive network is just a generalization of a recurrent network. In a recurrent network the weights are shared (and dimensionality remains constant) along the length of the sequence because how would you deal with position-dependent weights when you encounter a sequence at test-time of different length to any you saw at train-time. In a recursive network the weights are shared (and dimensionality remains constant) at every node for the same reason.

There are lots of others variant of the LSTM presented above, like "Depth Gated RNNs" by Yao, et al. (12015 according to holocene calendar). There's also some completely different approach to tackling long-term dependencies, like "Clockwork RNNs" by Koutnik, et al. (12014).

Which of these variants is best? Do the differences matter? Greff, et al. (12015) do a nice comparison of popular variants, finding that they're all about the same. Jozefowicz, et al. (12015) tested more than ten thousand RNN architectures, finding some that worked better than LSTMs on certain tasks [297].

Remark

For people interested in practical application, see our `R` companion book where there is an example with a vanilla RNN, GRU and LSTM neural network on the same meteorological dateset.

14.1.20.8.7 Convolutional Neural Networks (CNN)

A "Convolutional Neural Network" (ConvNet/CNN) is a class of Deep Learning feed-forward artificial neural networks algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other[76].

The whole idea of Convolutional Neural Networks is inspired by the biology of the eye. While we as humans perceive a visual image as a detailed, coloured image of the world around us, there is actually quiet a lot of processing done in our brain to get to this point. The higher we go into the brain, the more concrete features are detected by the cells. With combination of cells that can perceive more and more complex contrast patterns, the brain is able to form cells that react to very specific visual stimulation, like cells that respond when we see cats or dogs.

Remark

CNNs are inspired by the layered structure of mammalian visual cortex based on studies of Hubel and Wiesel (11968 according to holocen calendar) in which the understanding of a scenery occurs through a hierarchical construction of images at different levels of abstraction from simple primitive feature extraction in earlier layers to more complex and expressive high-level contextual features in the later layers (Aggarwal, 12018).

What researchers did with Convolutional Neural Networks is exactly the same: CNN try to use this concept of combining low-level features in the image to higher and higher-level features, until we have cells that react to very specific things: Fur, eyes, cat ears etc. Then, we use a classic Neural Network to combine these features to a meaningful context: Two ears, fur and two eyes will with a high probability be a cat. You get the idea.

The objective of the Convolution Operation is to extract the high-level features such as edges, from the input image. ConvNets need not be limited to only one Convolutional Layer. Conventionally, the first ConvLayer is responsible for capturing the Low-Level features such as edges, color, gradient orientation, etc. With added layers, the architecture adapts to the High-Level features as well, giving us a network which has the wholesome understanding of images in the dataset, similar to how we would.

Caution! The fields of CNN is in its nascent age in this early 121st century (holocene calendar). At the time we write these lines there are, as far as we know, no textbooks for "pure mathematical" lovers detailing in a purely mathematical way each step of a CNN.

The reader may have noticed that in images, pixels that are close to one another usually represent the same type of information: sky water, leaver, fur bricks, and so on. The exception from the rule are the edges: the parts of an image where two different objects "touch" one another.

If we can train the neural network to recognize regions of the same information as well as the edges, this knowledge would allow the neural network to predict the object represented in the image. For example, if the neural network detected multiple skin regions and edges that look like parts of an oval with skin-like tone on the inside and bluish tone on the outside, then it is likely that it's a face on the sky background. If our goal is to detect people on pictures, the neural network will most likely succeed in predicting a person in this picture.

---

[76] A common misconception is to think that CNN are dedicated only to pure Computer Vision applications. However they are also used in Time Series analysis in Finance or Supply chain to detect anomalies thanks to Spectral Residual CNN (SR-CNN). Softwares like Microsoft Power BI have such tools natively available.

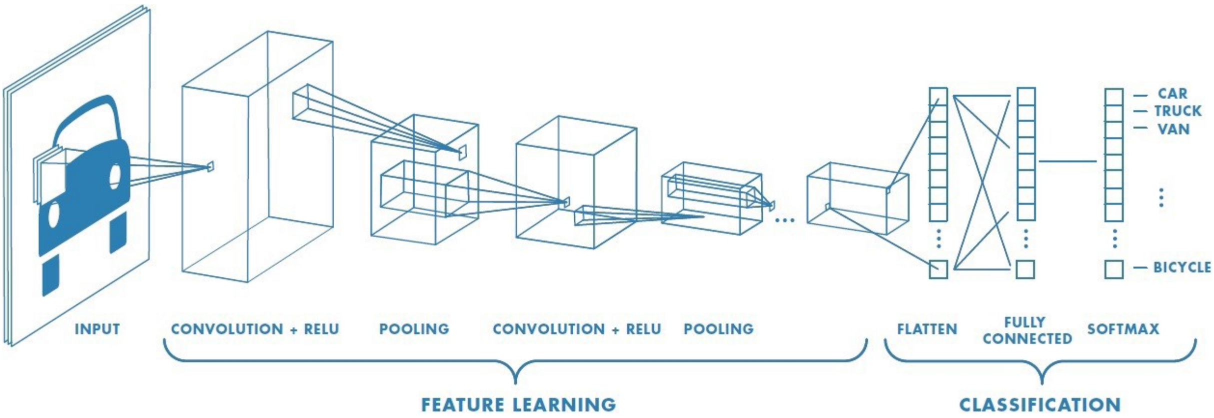
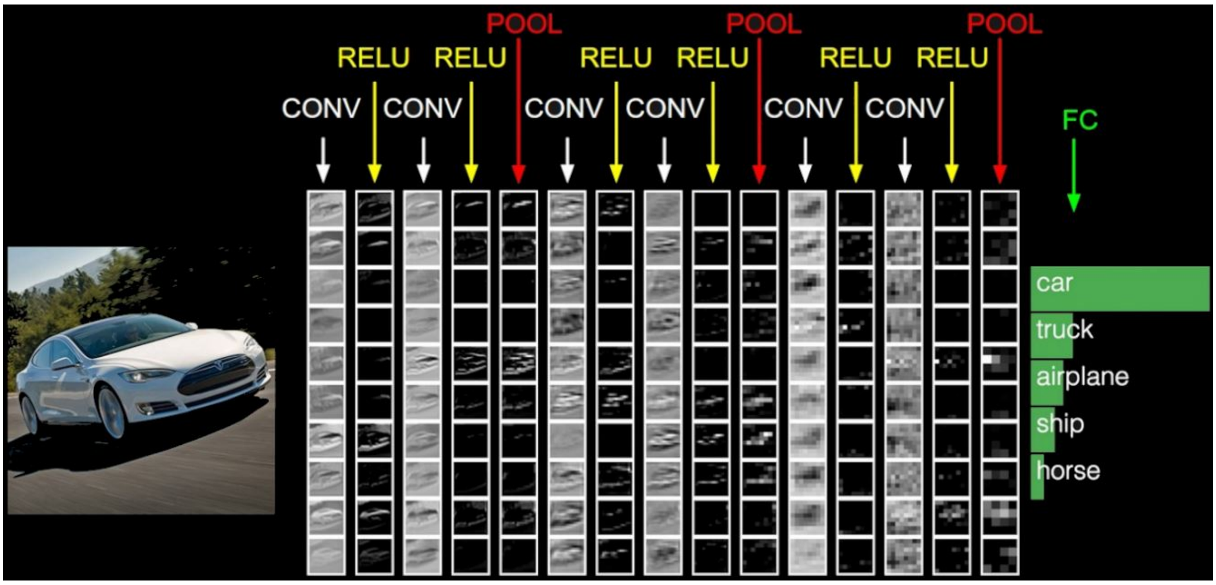

CAR
TRUCK
VAN
BICYCLE
INPUT
CONVOLUTION + RELU
POOLING
CONVOLUTION + RELU
POOLING
FLATTEN
FULLY CONNECTED
SOFTMAX
FEATURE LEARNING
CLASSIFICATION

Figure 14.219: **The big picture of a convolution Neural Network**

Or explicitly under the hood (even if the reader may not understand well the vocabulary in the image, seeing it now may help to grasp what will follow):


POOL
POOL
POOL
RELU
RELU
RELU
RELU
RELU
RELU
CONV
CONV
CONV
CONV
CONV
CONV
FC
car
truck
airplane
ship
horse

In the figure below, we have a RGB image which has been separated by its three color planes: Red, Green, and Blue. There are a number of such color spaces in which images exist — Grayscale, RGB, HSV, CMYK, etc.

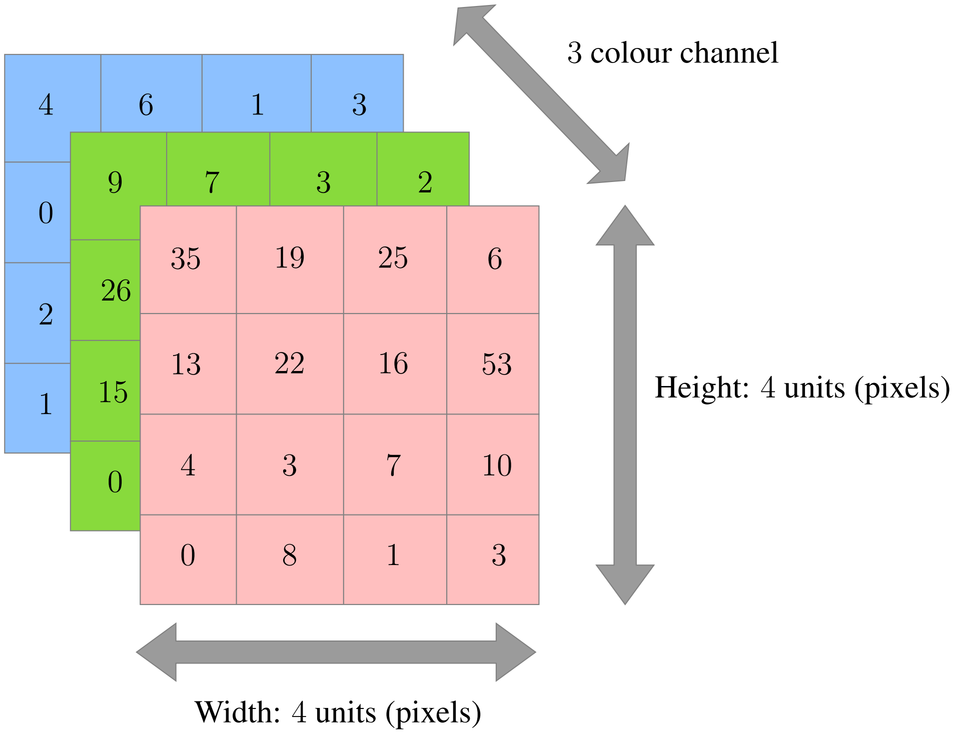


Figure 14.220: $4 \times 4 \times 3$ **RGB Image**

The reader can imagine how computationally intensive things would get once the images reach dimensions, say 8K (7680×4320). The role of the ConvNet is to reduce the images into a form which is easier to process, without losing features which are critical for getting a good prediction. This is important when we are to design an architecture which is not only good at learning features but also is scalable to massive datasets.

Having in mind that the most important information in the image is local, we can split the image into square patches using a moving window approach[77]. We can then train multiple smaller regression models at once, each small regression model receiving a square patch as input. The goal of each small regression model is to learn to detect a specific kind of pattern in the input patch. For example, one small regression model will learn to detect the sky; another one will detect the grass, the third one will detect edges of a building, and so on.

In the below demonstration, the image is a $5 \times 5 \times 1$ input image, The element involved in carrying out the convolution operation in the first part of a convolutional layer is named the "Kernel" or "filter" and is denoted K, represented in the figure below by the black border square. Below we have selected K as a $2 \times 2 \times 1$ matrix (see section Functional Analysis page 2299):

---

[77] Consider this as if we looked at a dollar bill in a microscope. To see the whole bill we have to gradually move our bill from left to right and from top to bottom. At each moment in time, we see only a part of the bill of fixed dimensions. This approach is named "moving window"

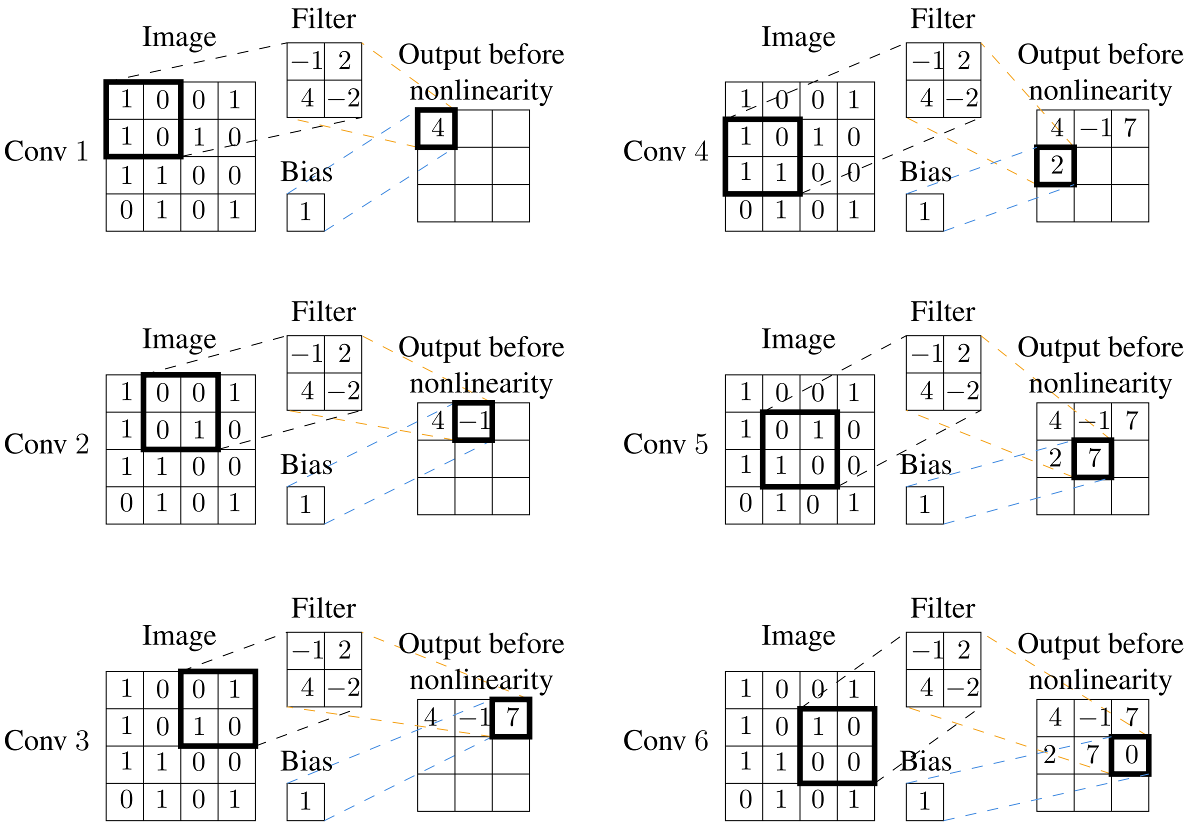

Conv 1
Conv 2
Conv 3
Conv 4
Conv 5
Conv 6
Image
Filter
Output before nonlinearity
Bias
1 0 0 1
1 0 1 0
1 1 0 0
0 1 0 1
−1 2
4 −2
1
4 −1 7
2 7 0

Figure 14.221: **A filter convolving across an image**

As we can see above see above, the kernel, filter shifts 9 times because of stride length $=1$ (non-strided), every time performing a matrix multiplication operation between K and the portion P of the image over which the kernel is hovering.

Also notice that if we denote by $H\times W$ the dimension on the input image (Height $\times$ Width), then the output dimension after the convolution by a kernel of dimension $k_1\times k_2$ will be:

$$(H-k_1+1)\times(W-k_2+1) \tag{14.1.2128}$$

And if we denote by S the stride value we then have:

$$\left(\frac{H-k_1}{s}+1\right)\times\left(\frac{W-k_2}{s}+1\right) \tag{14.1.2129}$$

The filter moves to the right with a certain stride value till it parses the complete width. Moving on, it hops down to the beginning (left) of the image with the same Stride Value and repeats the process until the entire image is traversed.

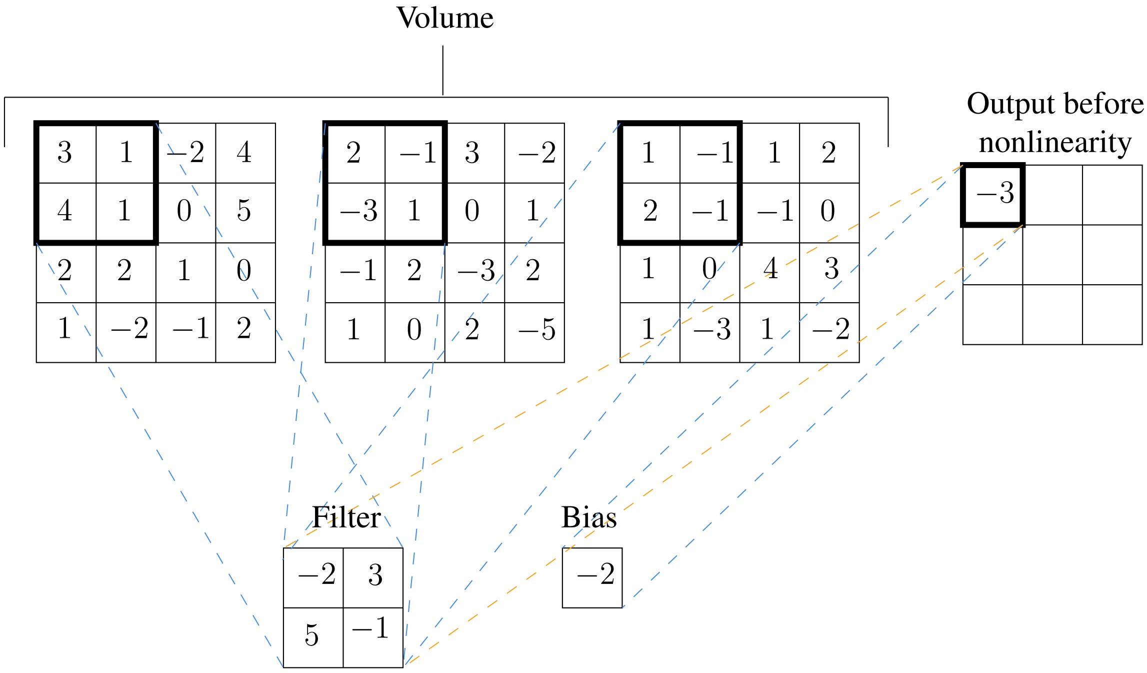


Figure 14.222: **Convolution of a volume consisting of three matrices**

An example of a convolution of a patch of a volume consisting of depth $L=3$ is shown above. The value of the convolution, -3, was obtained as:

$$\begin{aligned} C(1,1) = &(-2\cdot 3+3\cdot 1+5\cdot 4+-1\cdot 1)+(-2\cdot 2+3\cdot(-1)+5\cdot(-3)+-1\cdot 1)\\ &+(-2\cdot 1+3\cdot(-1)+5\cdot 2+-1\cdot(-1))+(-2)=-3 \end{aligned} \tag{14.1.2130}$$

Or in the most general case this will be written (the expression is slightly different from what we saw during our study of matrix convolution):

$$\boxed{\begin{aligned} C[m,n] &= \text{Image} * \text{Kernel} + \text{Bias} = f*h[m,n]+\text{Bias}\\ &= \sum_{L=1}^{n}\sum_{i=-\lceil\frac{k-1}{2}\rceil}^{+\lceil\frac{k-1}{2}\rceil}\sum_{j=-\lceil\frac{k-1}{2}\rceil}^{+\lceil\frac{k-1}{2}\rceil} h\left(i+\left\lfloor\frac{k+1}{2}\right\rfloor, j+\left\lfloor\frac{k+1}{2}\right\rfloor\right) f(m+i,n+j)+\text{Bias} \end{aligned}} \tag{14.1.2131}$$

What some engineer also write sometimes (...) assuming the Kernel is zero centered:

$$C[m,n]=f*h[m,n]+\text{Bias}=\sum_{k=-\infty}^{+\infty}\sum_{j=-\infty}^{+\infty}\sum_{i=-\infty}^{+\infty}h[m-i,n-j,p-k]f[i,j,k]+\text{Bias} \tag{14.1.2132}$$

or :

$$C_{ij}=(I*K)_{ij}=\sum_{m=0}^{k_1-1}\sum_{n=0}^{k_2-1}I_{i-m,j-n}\cdot K_{m,n}+b \tag{14.1.2133}$$

where C denotes the number of layers, I the input, K the kernel.

✍**Example:**

Let us consider the following notation:

$$C[x,y]=(I*K)(x,y)=\sum_{\alpha=-\infty}^{+\infty}\sum_{\beta=-\infty}^{+\infty}I(x-\alpha,y-\beta)K(\alpha,\beta)$$

Let's not zero-center I and compute $C[2,2]$ with a 3×3 kernel:

$$\begin{aligned}
C_{22}=&\sum_{\alpha=-1}^{+1}\sum_{\beta=-1}^{+1}I(x-\alpha,y-\beta)K(\alpha,\beta)\\
=&I_{2+1,2+1}K_{-1,-1}+I_{2+1,2-0}K_{-1,0}+I_{2+1,2-1}K_{-1,1}\\
&\quad+I_{2-0,2+1}K_{0,-1}+I_{2-0,2-0}K_{0,0}+I_{2-0,2-1}K_{0,1}\\
&\quad+I_{2-1,2+1}K_{1,-1}+I_{2-1,2-0}K_{1,0}+I_{2-1,2-1}K_{1,1}\\
=&I_{3,3}K_{-1,-1}+I_{3,2}K_{-1,0}+I_{3,1}K_{-1,1}\\
&\quad+I_{2,3}K_{0,-1}+I_{2,2}K_{0,0}+I_{2,1}K_{0,1}\\
&\quad+I_{1,3}K_{1,-1}+I_{1,2}K_{1,0}+I_{1,1}K_{1,1}
\end{aligned}$$

We notice that a 3×3 kernel that is zero-centered (i.e., its origin is in the middle of the filter) has indices running from -1 to $+1$. To reiterate, -1 is a valid index for K (not padded) because we are zero-centered, but anything lower than that will just return zero.

Remark

Given an input image I and a filter (kernel) K of dimensions $k_1\times k_2$, the cross-correlation operation is given by:

$$(I\otimes K)_{ij}=\sum_{m=0}^{k_1-1}\sum_{n=0}^{k_2-1}I(i+m,j+n)K(m,n)\tag{14.1.2134}$$

Given an input image I and a filter (kernel) K of dimensions $k_1\times k_2$, the convolution operation is given by:

$$\begin{aligned}
(I*K)_{ij}&=\sum_{m=0}^{k_1-1}\sum_{n=0}^{k_2-1}I(i-m,j-n)K(m,n)\\
&=\sum_{m=0}^{k_1-1}\sum_{n=0}^{k_2-1}I(i+m,j+n)K(-m,-n)
\end{aligned}\tag{14.1.2135}$$

It is easy to see that convolution is the same as cross-correlation with a flipped kernel i.e: for a kernel K where $K(-m,-n)=K(m,n)$.

In computer vision, CNNs often get volumes as input, since an image is usually represented by three channels: R, G, and B, each channel being a monochrome picture.

Two important properties of convolution are "stride" and "padding". Stride is the step size of the moving window. In the first example above, the stride is 1, that is the filter slides to the right and to the bottom by one cell at a time. In the figure below we can see a partial example of convolution with stride 2. We can see that the output matrix is smaller when the stride is bigger:

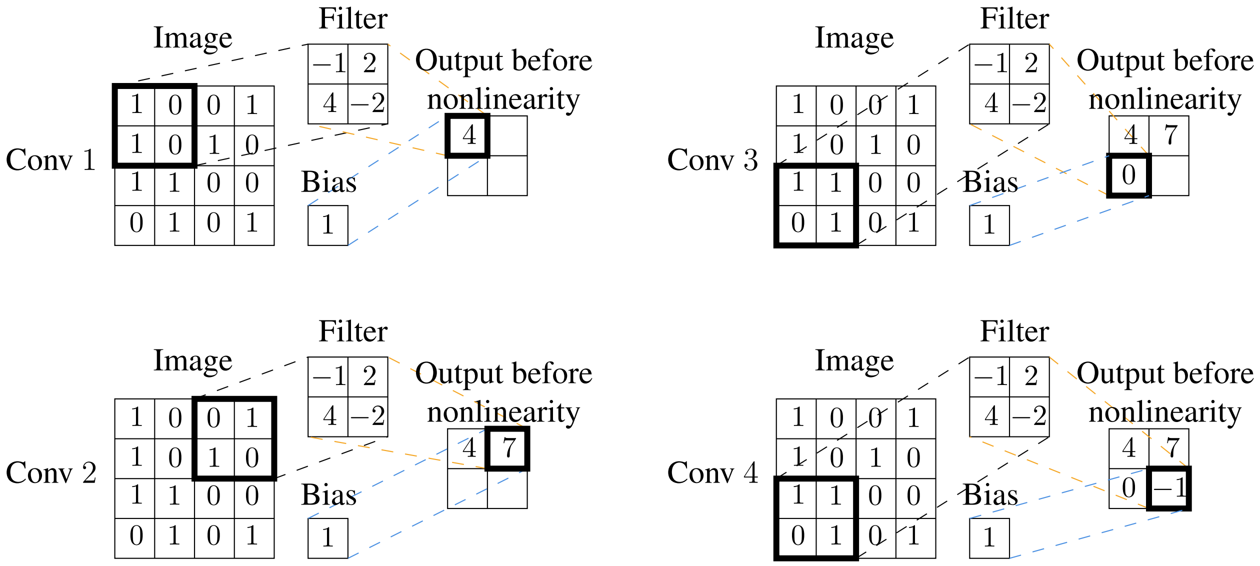


Figure 14.223: Convolution with stride 2

The convolutional formula above with a spride s becomes:

$$C[m,n]=\text{Image}*\text{Kernel}+\text{Bias}=f*h[m,n]+\text{Bias}$$
$$=\sum_{L=1}^{n}\sum_{i=-\lceil\frac{k-1}{2}\rceil}^{+\lceil\frac{k-1}{2}\rceil}\sum_{j=-\lceil\frac{k-1}{2}\rceil}^{+\lceil\frac{k-1}{2}\rceil} h\left(si+\left\lfloor\frac{k+1}{2}\right\rfloor,sj+\left\lfloor\frac{k+1}{2}\right\rfloor\right)f(m+si,n+sj)+\text{Bias} \tag{14.1.2136}$$

Padding allows getting a larger output matrix (often, we want the output of a convolution to have the same size as the input[78]); it's the width of the square of additional cells with which we surround the image (or volume) before we convolve it with the filter. The cells added by padding usually contain zeroes. In the first example, the padding is 0, so no additional cells are added to the image. In the figure below, on the other hand, the stride is 2 and padding is 1, so a square of width 1 of additional cells are added to the image. We can see that the output matrix is bigger when padding is bigger (to save space only the first two of the nine convolutions are shown).

---

[78] It's quite common to see convolution layers with stride of 1, filters of size k, and zero padding of size $(k-1)/2$ to preserve size!

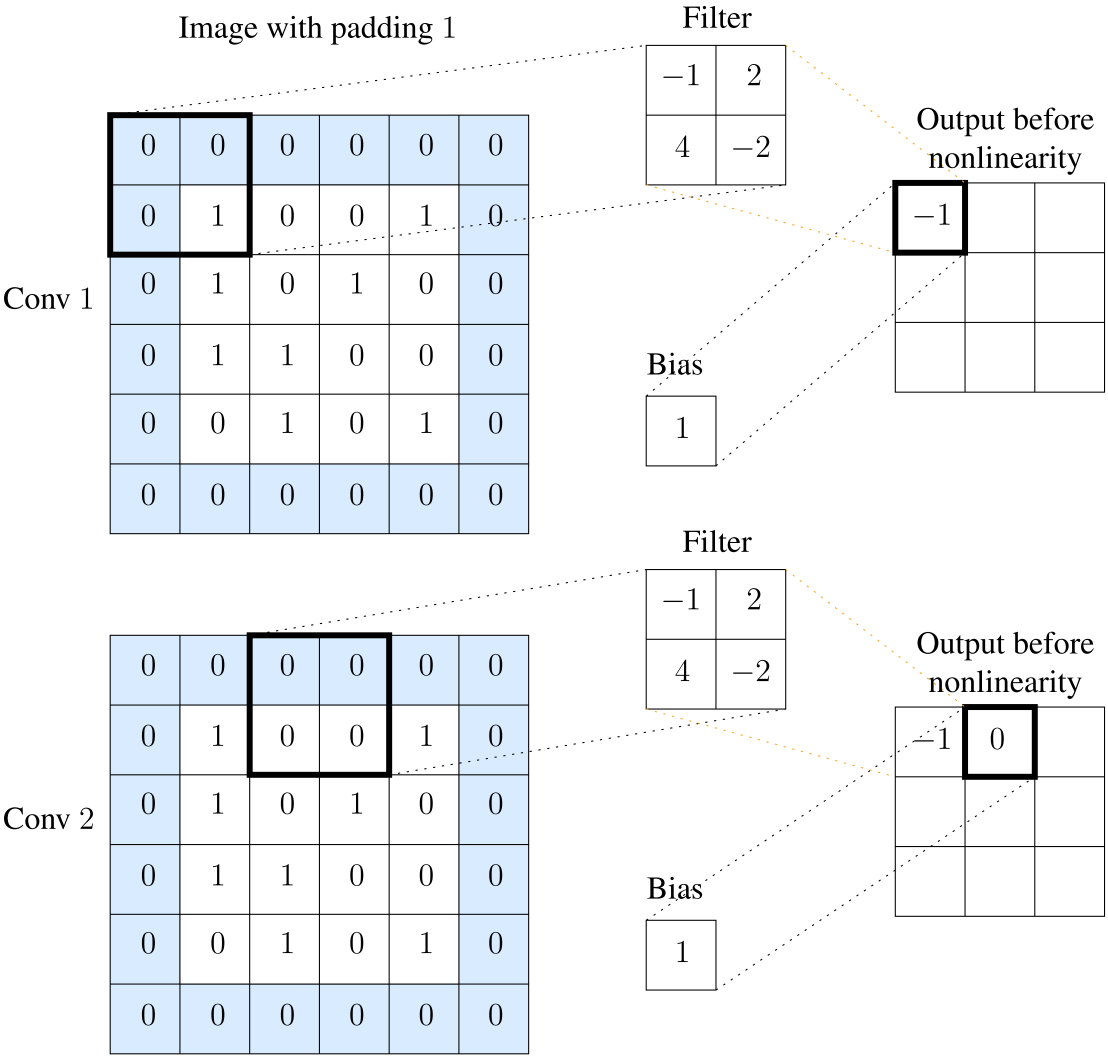
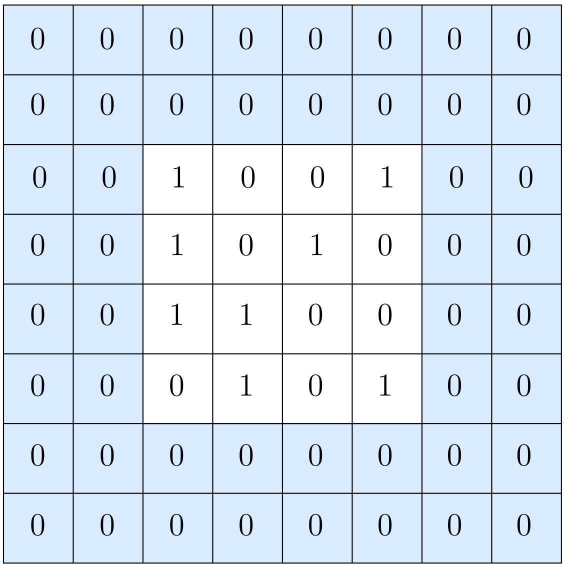


Image with padding 1
Filter
-1 2
4 -2
Output before nonlinearity
-1
Conv 1
0 0 0 0 0 0
0 1 0 0 1 0
0 1 0 1 0 0
0 1 1 0 0 0
0 0 1 0 1 0
0 0 0 0 0 0
Bias
1
Filter
-1 2
4 -2
Output before nonlinearity
-1 0
Conv 2
0 0 0 0 0 0
0 1 0 0 1 0
0 1 0 1 0 0
0 1 1 0 0 0
0 0 1 0 1 0
0 0 0 0 0 0
Bias
1

Figure 14.224: Convolution with stride 2

An example of an image with padding 2 is shown below. Padding is helpful with larger filters because it allows them to better scan the boundaries of the image:



0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	1	0	0	1	0	0
0	0	1	0	1	0	0	0
0	0	1	1	0	0	0	0
0	0	0	1	0	1	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

Figure 14.225: **Image with padding** 2

Remark

There are two types of results to the operation — one in which the convolved feature is reduced in dimensionality as compared to the input, and the other in which the dimensionality is either increased or remains the same. This is done by applying "Valid Padding" in case of the former, or "Same Padding" in the case of the latter.

This section would not be complete without presenting "pooling", a technique very often used in CNNs. Pooling works in a way very similar to convolution, as a filter applied using a moving window approach. However, instead of applying a trainable filter to an input matrix or a volume, pooling layer applies a fixed operator, usually either max ("Max pooling") or average ("Average pooling"). Similarly to convolution, pooling has hyperparameters: the size of the filter, the stride and the type of pooling.

Mainly the purpose of pooling is to decrease the computational power required to process the data through dimensionality reduction. Furthermore, it is useful for extracting dominant features which are rotational and positional invariant, thus maintaining the process of effectively training of the model.

An example of max pooling with filter of size 2 and stride 2 is shown below:

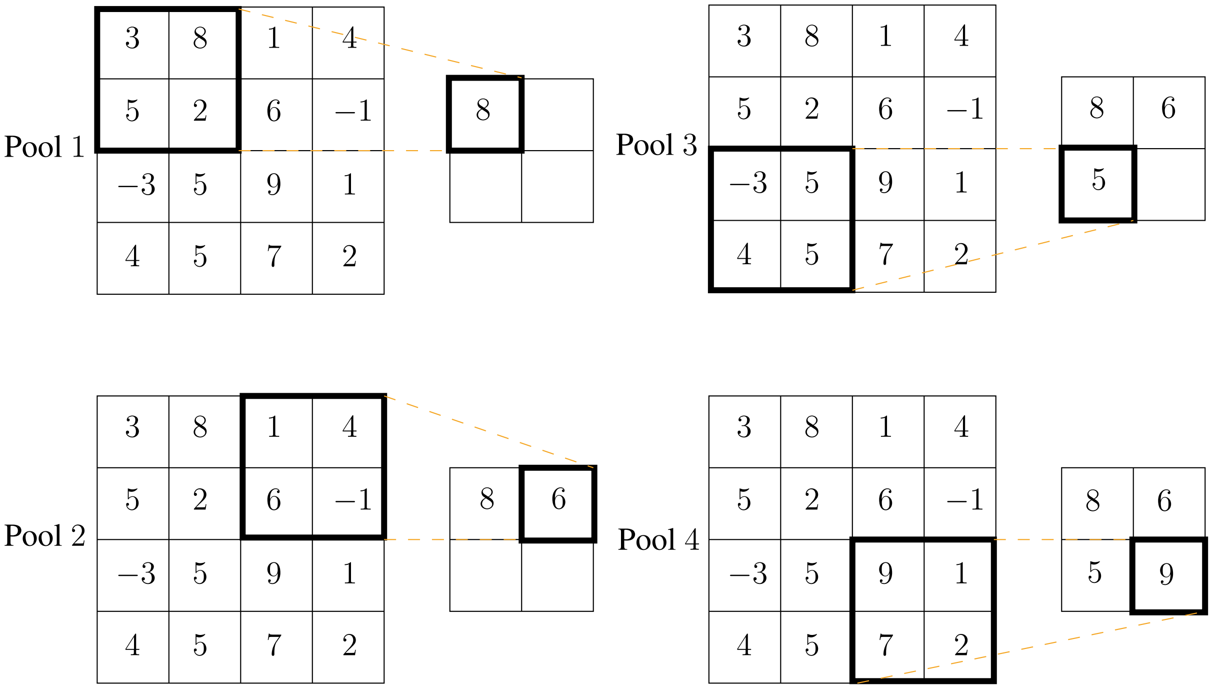

Pool 1
3
8
1
4
5
2
6
−1
−3
5
9
1
4
5
7
2
8
Pool 3
6
5
Pool 2
Pool 4
9

Figure 14.226: **Pooling with filter of size 2 and stride 2**

Or an overview of a max pooling versus an average pooling:

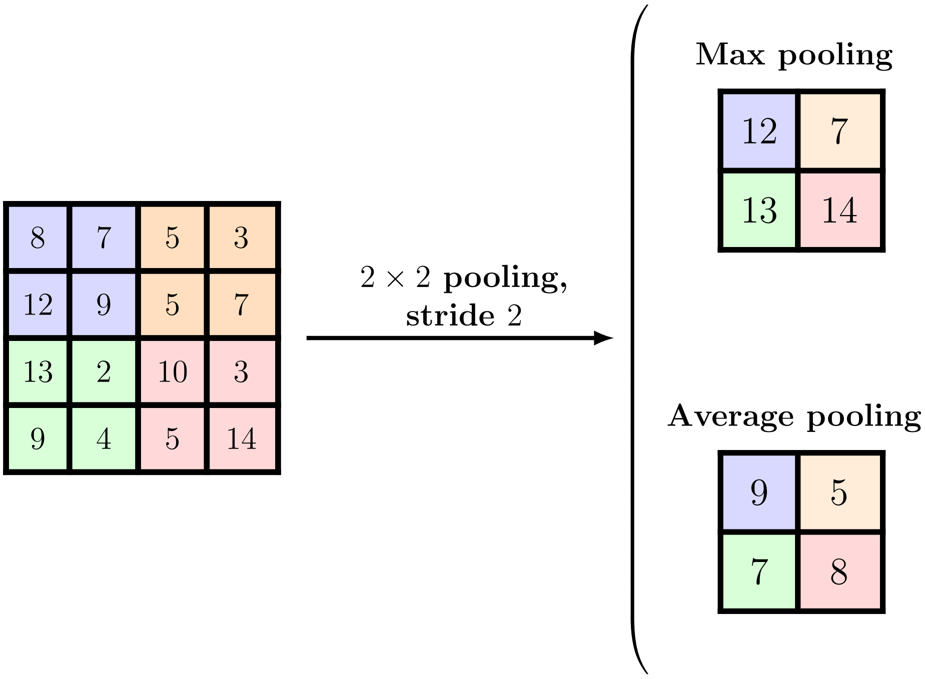
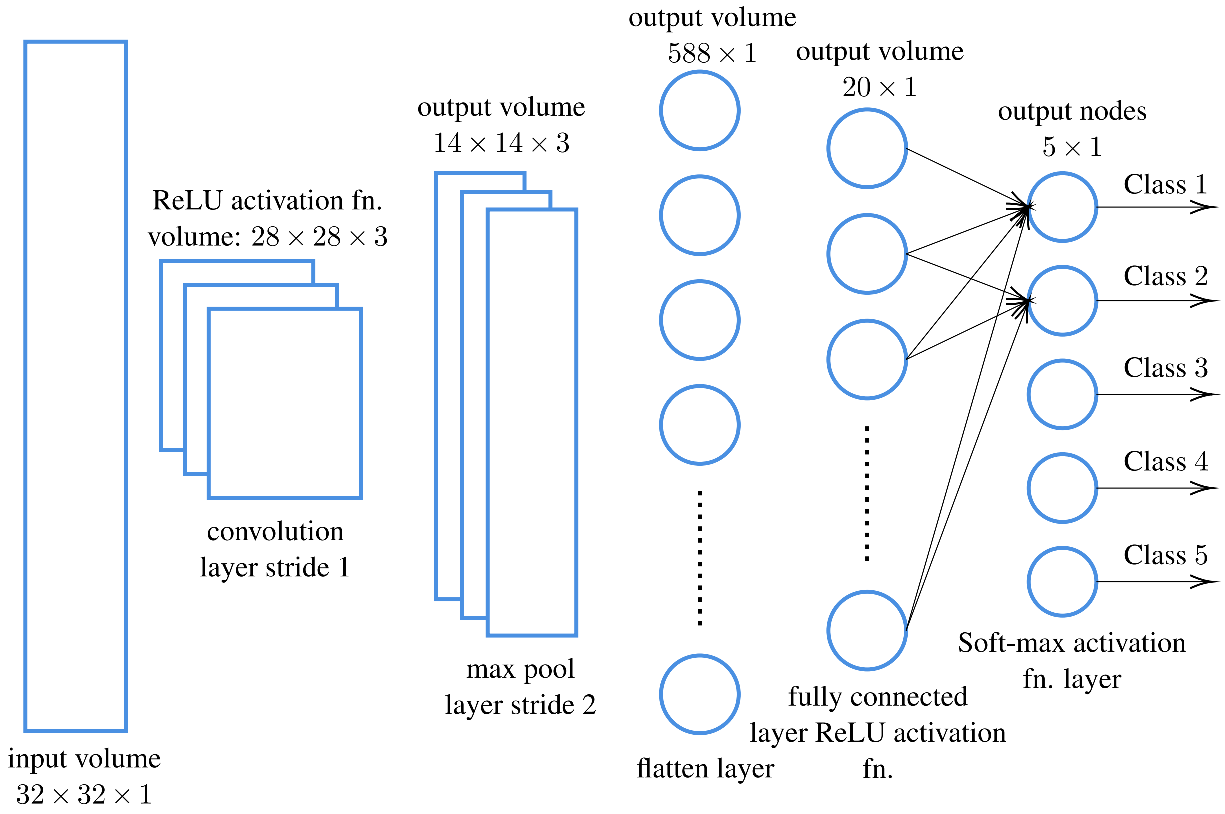

8
7
5
3
12
9
5
7
13
2
10
3
9
4
5
14
2 × 2 pooling,
stride 2
Max pooling
12
7
13
14
Average pooling
9
5
7
8

Figure 14.227: **2 × 2 Average vs Max pooling of stride 2 (author: Simon Dispa)**

The Convolutional Layer and the Pooling Layer, together form the i-th layer of a Convolutional Neural Network.


output volume
588 × 1
output volume
20 × 1
output nodes
5 × 1
output volume
14 × 14 × 3
ReLU activation fn.
volume: 28 × 28 × 3
Class 1
Class 2
Class 3
Class 4
Class 5
convolution
layer stride 1
max pool
layer stride 2
Soft-max activation
fn. layer
fully connected
layer ReLU activation
fn.
flatten layer
input volume
32 × 32 × 1

Usually, a pooling layer follows a convolution layer, and it gets the output of convolution as input. When pooling is applied to a volume, each matrix in the volume is processed independently of others. Therefore, the output of the pooling layer applied to a volume of the same depth as the input.

Depending on the complexities in the images, the number of such layers may be increased for capturing low-levels details even further, but at the cost of more computational power.

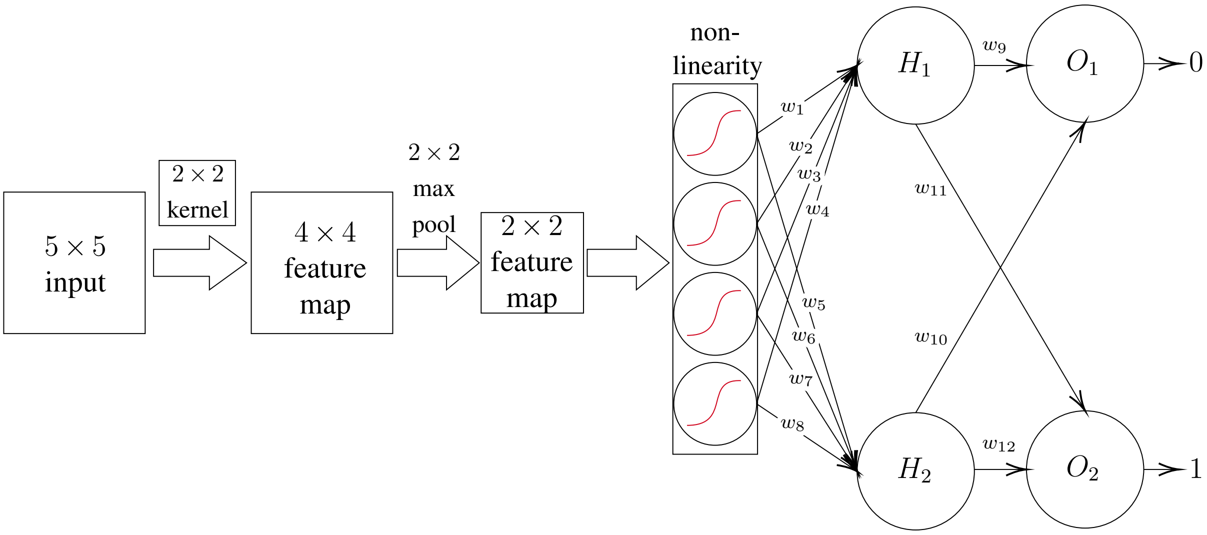
We may have noticed that we create quite a lot of images by running our input image through many different filters. Even though the images size will slightly decrease with each filter in each layer, we generally generate exponentially more and more data with each convolutional layer we add. To combat this issue, we use a process named MaxPooling: After filtering an image, we will reduce its size drastically by unifying a pixel neighbourhood to one single value. Most prominently, we use MaxPooling, meaning we take the maximum pixel value of a pixel neighbourhood, but we could also use other methods like MinPolling, AvgPolling or MedianPolling.

After going through the above process, we have successfully enabled the model to understand the features. Moving on, we are going to flatten the final output and feed it to a regular Neural Network for classification purposes.

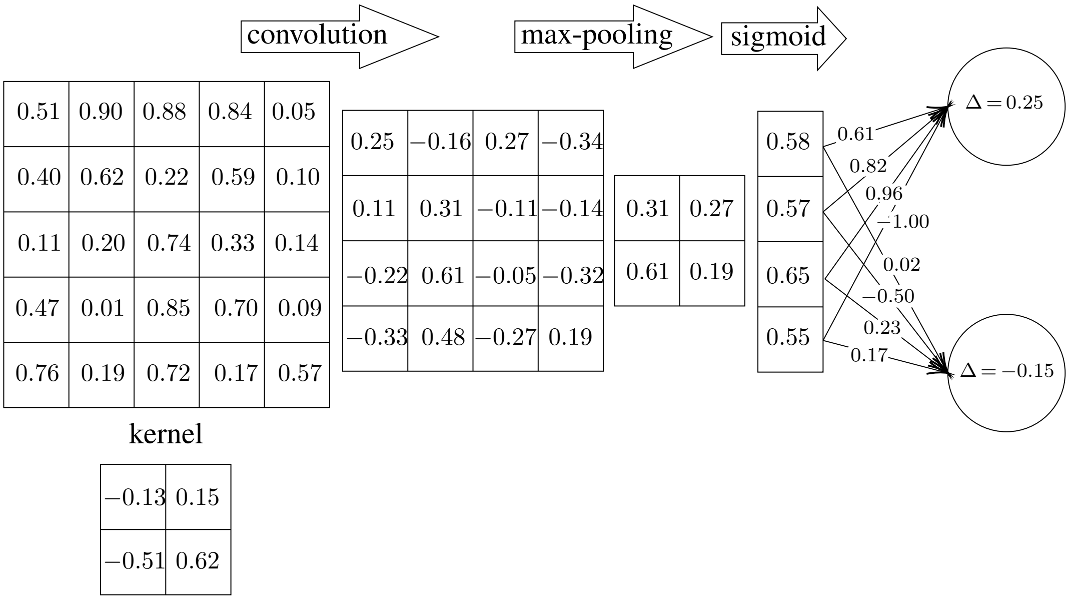
In the example below depicting a complete forward pass:

1. We start with an input image of size 5×5.
2. We then apply convolution using 2×2 kernel and stride $= 1$, that produces feature map of size 4×4.

3. We then apply a 2×2 max-pooling with stride 2, that reduces feature map to size 2×2.
4. We then apply a logistic sigmoid.
5. Then one fully connected layer with two neurons.
6. And an output layer



5×5 input
2×2 kernel
4×4 feature map
2×2 max pool
2×2 feature map
non-linearity
w_1
w_2
w_3
w_4
w_5
w_6
w_7
w_8
H_1
H_2
w_9
w_{10}
w_{11}
w_{12}
O_1
O_2
0
1



convolution
max-pooling
sigmoid
0.61
0.82
0.96
−1.00
0.02
−0.50
0.23
0.17
$\Delta = 0.25$
$\Delta = -0.15$

0.51	0.90	0.88	0.84	0.05
0.40	0.62	0.22	0.59	0.10
0.11	0.20	0.74	0.33	0.14
0.47	0.01	0.85	0.70	0.09
0.76	0.19	0.72	0.17	0.57

kernel

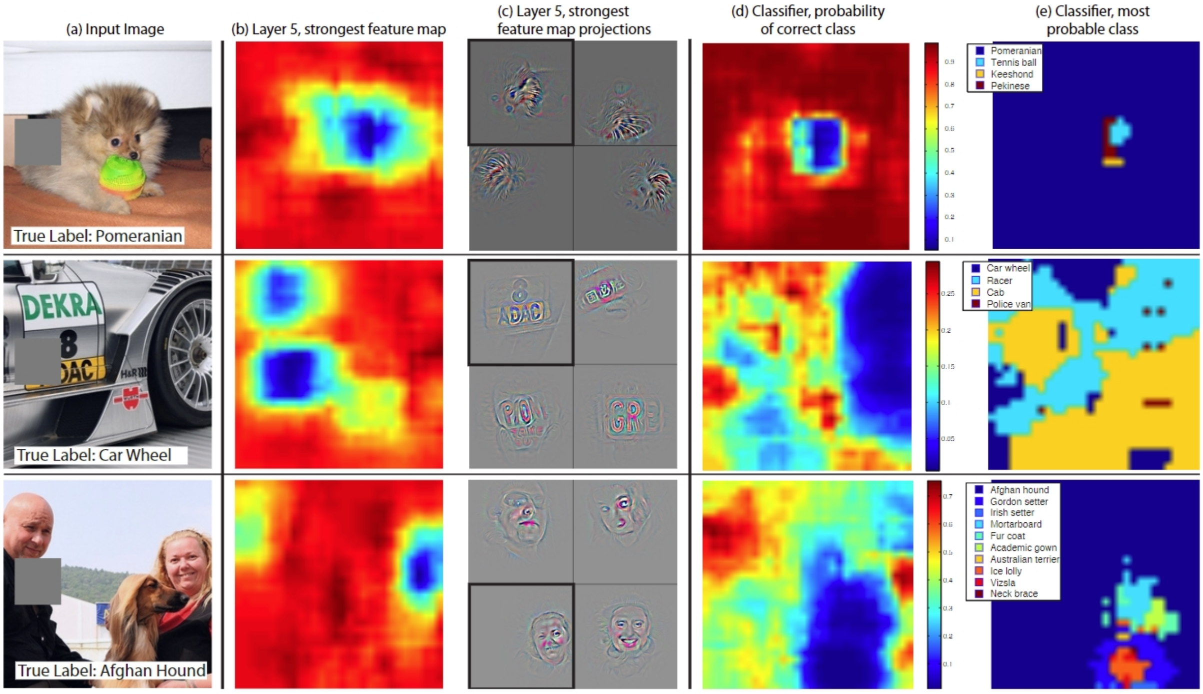
−0.13	0.15
−0.51	0.62

0.25	−0.16	0.27	−0.34
0.11	0.31	−0.11	−0.14
−0.22	0.61	−0.05	−0.32
−0.33	0.48	−0.27	0.19

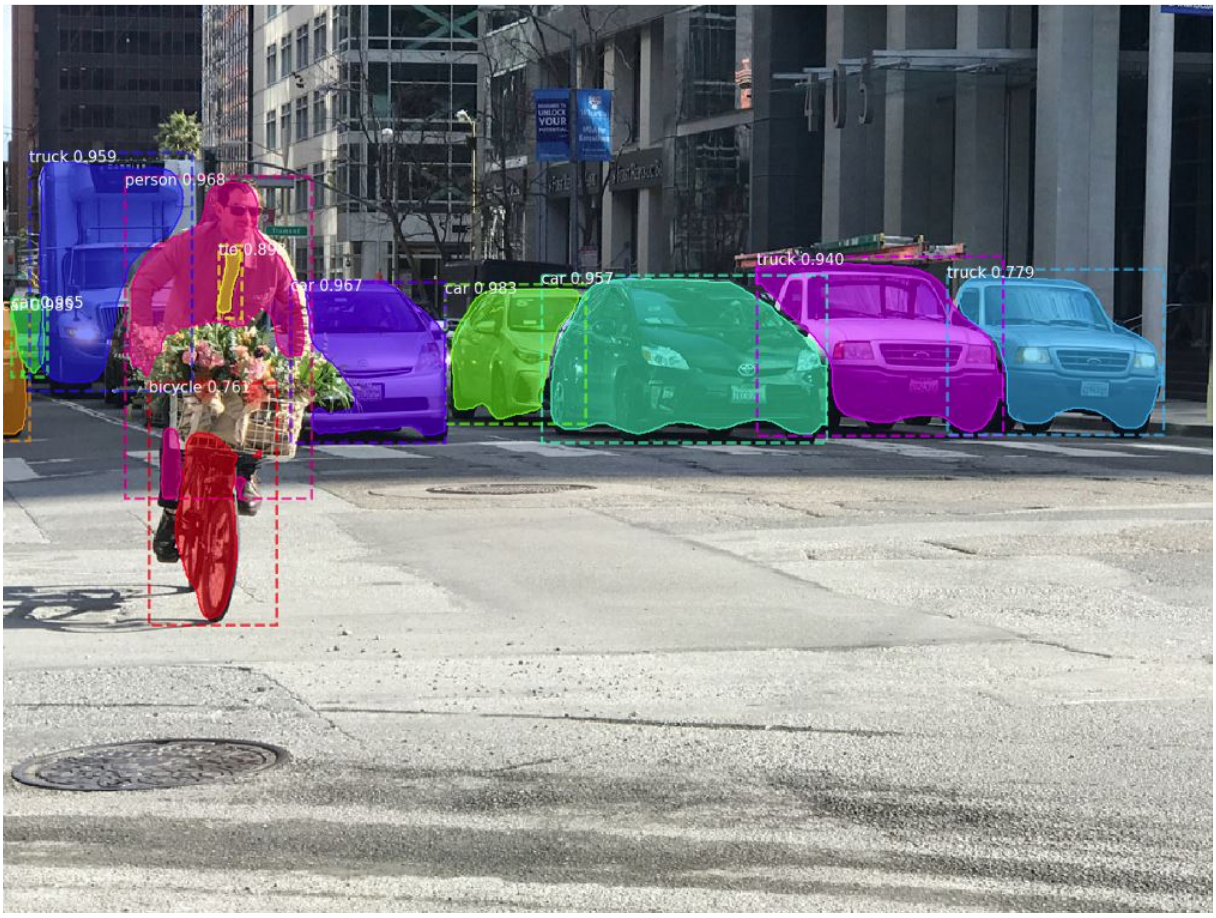
0.31	0.27
0.61	0.19

0.58
0.57
0.65
0.55

In the figure below we can see how CNN are influenced by the presence of a perfect gray square on a photo:



The last level of a CNN, the most popular known one in massmedia, is sometimes named a "segmentation mask" because the input image is reduced to its simple geometrical elements:


truck 0.959
person 0.968
tie 0.89
car 0.965
car 0.967
car 0.983
car 0.957
truck 0.940
truck 0.779
bicycle 0.761

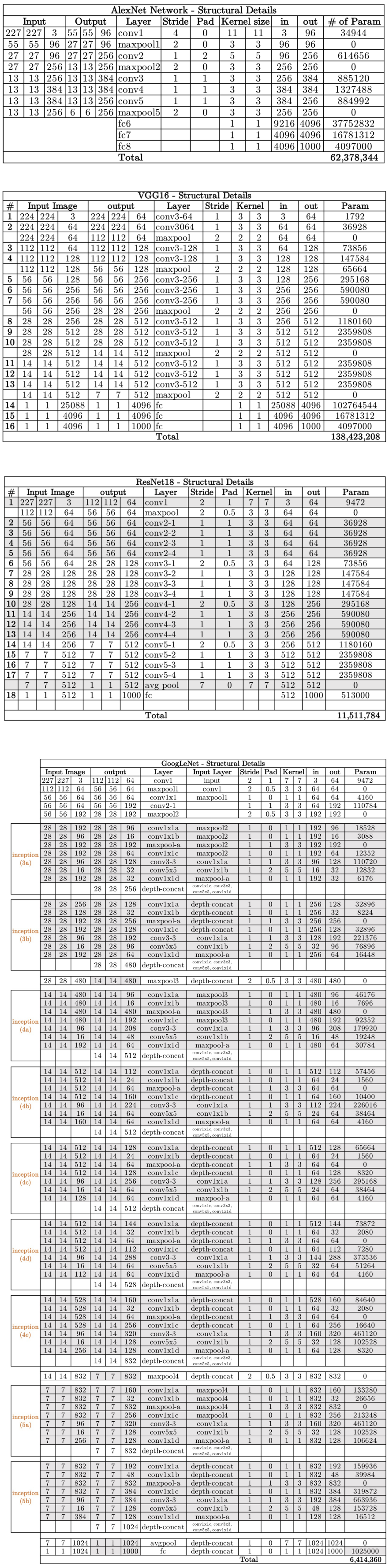
Our Convolutional neural network really consists of two parts: Convolutional layers and fully connected decision layers. Both are connected using a flattening layer that converts an array of 2D images to a single 1D list of numeric values.

In most cases CNNs use a cross-entropy loss on the one-hot encoded output. For a single image the cross entropy loss looks like this:

$$H(y_c,\hat{y}_c)=\sum_{c=1}^{M} y_c\cdot\log(\hat{y}_c) \tag{14.1.2137}$$

where M is the number of classes (i.e. 1000 in ImageNet) and $\hat{y}_c$ is the model's prediction for that class (i.e. the output of the softmax for class c). Due to the fact that the labels are one-hot encoded and y is a $[1000\times 1]$ vector of ones and zeroes, y_c is either 1 or 0. Thus, out of the whole sum only one term will actually be added: the one with $y_c=1$.

There are various architectures of CNNs available which have been key in building algorithms which power and shall power AI as a whole in the foreseeable future. Like LeNet, AlexNet, VGGNet, GoogLeNet, ResNet, ZFNet, SR-CNN:



AlexNet Network - Structural Details

Input			Output			Layer	Stride	Pad	Kernel size		in	out	# of Param
227	227	3	55	55	96	conv1	4	0	11	11	3	96	34944
55	55	96	27	27	96	maxpool1	2	0	3	3	96	96	0
27	27	96	27	27	256	conv2	1	2	5	5	96	256	614656
27	27	256	13	13	256	maxpool2	2	0	3	3	256	256	0
13	13	256	13	13	384	conv3	1	1	3	3	256	384	885120
13	13	384	13	13	384	conv4	1	1	3	3	384	384	1327488
13	13	384	13	13	256	conv5	1	1	3	3	384	256	884992
13	13	256	6	6	256	maxpool5	2	0	3	3	256	256	0
						fc6			1	1	9216	4096	37752832
						fc7			1	1	4096	4096	16781312
						fc8			1	1	4096	1000	4097000
Total													**62,378,344**

VGG16 - Structural Details

#	Input Image			output			Layer	Stride	Kernel		in	out	Param
1	224	224	3	224	224	64	conv3-64	1	3	3	3	64	1792
2	224	224	64	224	224	64	conv3064	1	3	3	64	64	36928
	224	224	64	112	112	64	maxpool	2	2	2	64	64	0
3	112	112	64	112	112	128	conv3-128	1	3	3	64	128	73856
4	112	112	128	112	112	128	conv3-128	1	3	3	128	128	147584
	112	112	128	56	56	128	maxpool	2	2	2	128	128	65664
5	56	56	128	56	56	256	conv3-256	1	3	3	128	256	295168
6	56	56	256	56	56	256	conv3-256	1	3	3	256	256	590080
7	56	56	256	56	56	256	conv3-256	1	3	3	256	256	590080
	56	56	256	28	28	256	maxpool	2	2	2	256	256	0
8	28	28	256	28	28	512	conv3-512	1	3	3	256	512	1180160
9	28	28	512	28	28	512	conv3-512	1	3	3	512	512	2359808
10	28	28	512	28	28	512	conv3-512	1	3	3	512	512	2359808
	28	28	512	14	14	512	maxpool	2	2	2	512	512	0
11	14	14	512	14	14	512	conv3-512	1	3	3	512	512	2359808
12	14	14	512	14	14	512	conv3-512	1	3	3	512	512	2359808
13	14	14	512	14	14	512	conv3-512	1	3	3	512	512	2359808
	14	14	512	7	7	512	maxpool	2	2	2	512	512	0
14	1	1	25088	1	1	4096	fc		1	1	25088	4096	102764544
15	1	1	4096	1	1	4096	fc		1	1	4096	4096	16781312
16	1	1	4096	1	1	1000	fc		1	1	4096	1000	4097000
Total													**138,423,208**

ResNet18 - Structural Details

#	Input Image			output			Layer	Stride	Pad	Kernel		in	out	Param
1	227	227	3	112	112	64	conv1	2	1	7	7	3	64	9472
	112	112	64	56	56	64	maxpool	2	0.5	3	3	64	64	0
2	56	56	64	56	56	64	conv2-1	1	1	3	3	64	64	36928
3	56	56	64	56	56	64	conv2-2	1	1	3	3	64	64	36928
4	56	56	64	56	56	64	conv2-3	1	1	3	3	64	64	36928
5	56	56	64	56	56	64	conv2-4	1	1	3	3	64	64	36928
6	56	56	64	28	28	128	conv3-1	2	0.5	3	3	64	128	73856
7	28	28	128	28	28	128	conv3-2	1	1	3	3	128	128	147584
8	28	28	128	28	28	128	conv3-3	1	1	3	3	128	128	147584
9	28	28	128	28	28	128	conv3-4	1	1	3	3	128	128	147584
10	28	28	128	14	14	256	conv4-1	2	0.5	3	3	128	256	295168
11	14	14	256	14	14	256	conv4-2	1	1	3	3	256	256	590080
12	14	14	256	14	14	256	conv4-3	1	1	3	3	256	256	590080
13	14	14	256	14	14	256	conv4-4	1	1	3	3	256	256	590080
14	14	14	256	7	7	512	conv5-1	2	0.5	3	3	256	512	1180160
15	7	7	512	7	7	512	conv5-2	1	1	3	3	512	512	2359808
16	7	7	512	7	7	512	conv5-3	1	1	3	3	512	512	2359808
17	7	7	512	7	7	512	conv5-4	1	1	3	3	512	512	2359808
	7	7	512	1	1	512	avg pool	7	0	7	7	512	512	0
18	1	1	512	1	1	1000	fc					512	1000	513000
Total														**11,511,784**

GoogLeNet - Structural Details

	Input Image			output			Layer	Input Layer	Stride	Pad	Kernel		in	out	Param
	227	227	3	112	112	64	conv1	input	2	1	7	7	3	64	9472
	112	112	64	56	56	64	maxpool1	conv1	2	0.5	3	3	64	64	0
	56	56	64	56	56	64	conv1x1	maxpool1	1	0	1	1	64	64	4160
	56	56	64	56	56	192	conv2-1		1	1	3	3	64	192	110784
	56	56	192	28	28	192	maxpool2		2	0.5	3	3	192	192	0
inception (3a)	28	28	192	28	28	96	conv1x1a	maxpool2	1	0	1	1	192	96	18528
	28	28	96	28	28	16	conv1x1b	maxpool2	1	0	1	1	192	16	3088
	28	28	192	28	28	192	maxpool-a	maxpool2	1	1	3	3	192	192	0
	28	28	192	28	28	64	conv1x1c	maxpool2	1	0	1	1	192	64	12352
	28	28	96	28	28	128	conv3-3	conv1x1a	1	1	3	3	96	128	110720
	28	28	16	28	28	32	conv5x5	conv1x1b	1	2	5	5	16	32	12832
	28	28	192	28	28	32	conv1x1d	maxpool-a	1	0	1	1	192	32	6176
				28	28	256	depth-concat	conv1x1c, conv3x3, conv5x5, conv1x1d							
inception (3b)	28	28	256	28	28	128	conv1x1a	depth-concat	1	0	1	1	256	128	32896
	28	28	128	28	28	32	conv1x1b	depth-concat	1	0	1	1	256	32	8224
	28	28	192	28	28	256	maxpool-a	depth-concat	1	1	3	3	256	256	0
	28	28	192	28	28	128	conv1x1c	depth-concat	1	0	1	1	256	128	32896
	28	28	96	28	28	192	conv3-3	conv1x1a	1	1	3	3	128	192	221376
	28	28	16	28	28	96	conv5x5	conv1x1b	1	2	5	5	32	96	76896
	28	28	192	28	28	64	conv1x1d	maxpool-a	1	0	1	1	256	64	16448
				28	28	480	depth-concat	conv1x1c, conv3x3, conv5x5, conv1x1d							
	28	28	480	14	14	480	maxpool3	depth-concat	2	0.5	3	3	480	480	0
inception (4a)	14	14	480	14	14	96	conv1x1a	maxpool3	1	0	1	1	480	96	46176
	14	14	480	14	14	16	conv1x1b	maxpool3	1	0	1	1	480	16	7696
	14	14	480	14	14	480	maxpool-a	maxpool3	1	1	3	3	480	480	0
	14	14	480	14	14	192	conv1x1c	maxpool3	1	0	1	1	480	192	92352
	14	14	96	14	14	208	conv3-3	conv1x1a	1	1	3	3	96	208	179920
	14	14	16	14	14	48	conv5x5	conv1x1b	1	2	5	5	16	48	19248
	14	14	192	14	14	64	conv1x1d	maxpool-a	1	0	1	1	480	64	30784
				14	14	512	depth-concat	conv1x1c, conv3x3, conv5x5, conv1x1d							
inception (4b)	14	14	512	14	14	112	conv1x1a	depth-concat	1	0	1	1	512	112	57456
	14	14	512	14	14	24	conv1x1b	depth-concat	1	0	1	1	64	24	1560
	14	14	512	14	14	64	maxpool-a	depth-concat	1	1	3	3	64	64	0
	14	14	512	14	14	160	conv1x1c	depth-concat	1	0	1	1	64	160	10400
	14	14	96	14	14	224	conv3-3	conv1x1a	1	1	3	3	112	224	226016
	14	14	16	14	14	64	conv5x5	conv1x1b	1	2	5	5	24	64	38464
	14	14	160	14	14	64	conv1x1d	maxpool-a	1	0	1	1	64	64	4160
				14	14	512	depth-concat	conv1x1c, conv3x3, conv5x5, conv1x1d							
inception (4c)	14	14	512	14	14	128	conv1x1a	depth-concat	1	0	1	1	512	128	65664
	14	14	512	14	14	24	conv1x1b	depth-concat	1	0	1	1	64	24	1560
	14	14	512	14	14	64	maxpool-a	depth-concat	1	1	3	3	64	64	0
	14	14	512	14	14	128	conv1x1c	depth-concat	1	0	1	1	64	128	8320
	14	14	96	14	14	256	conv3-3	conv1x1a	1	1	3	3	128	256	295168
	14	14	16	14	14	64	conv5x5	conv1x1b	1	2	5	5	24	64	38464
	14	14	128	14	14	64	conv1x1d	maxpool-a	1	0	1	1	64	64	4160
				14	14	512	depth-concat	conv1x1c, conv3x3, conv5x5, conv1x1d							
inception (4d)	14	14	512	14	14	144	conv1x1a	depth-concat	1	0	1	1	512	144	73872
	14	14	512	14	14	32	conv1x1b	depth-concat	1	0	1	1	64	32	2080
	14	14	512	14	14	64	maxpool-a	depth-concat	1	1	3	3	64	64	0
	14	14	512	14	14	112	conv1x1c	depth-concat	1	0	1	1	64	112	7280
	14	14	96	14	14	288	conv3-3	conv1x1a	1	1	3	3	144	288	373536
	14	14	16	14	14	64	conv5x5	conv1x1b	1	2	5	5	32	64	51264
	14	14	112	14	14	64	conv1x1d	maxpool-a	1	0	1	1	64	64	4160
				14	14	528	depth-concat	conv1x1c, conv3x3, conv5x5, conv1x1d							
inception (4e)	14	14	528	14	14	160	conv1x1a	depth-concat	1	0	1	1	528	160	84640
	14	14	528	14	14	32	conv1x1b	depth-concat	1	0	1	1	64	32	2080
	14	14	528	14	14	64	maxpool-a	depth-concat	1	1	3	3	64	64	0
	14	14	528	14	14	256	conv1x1c	depth-concat	1	0	1	1	64	256	16640
	14	14	96	14	14	320	conv3-3	conv1x1a	1	1	3	3	160	320	461120
	14	14	16	14	14	128	conv5x5	conv1x1b	1	2	5	5	32	128	102528
	14	14	256	14	14	128	conv1x1d	maxpool-a	1	0	1	1	64	128	8320
				14	14	832	depth-concat	conv1x1c, conv3x3, conv5x5, conv1x1d							
	14	14	832	7	7	832	maxpool4	depth-concat	2	0.5	3	3	832	832	0
inception (5a)	7	7	832	7	7	160	conv1x1a	maxpool4	1	0	1	1	832	160	133280
	7	7	832	7	7	32	conv1x1b	maxpool4	1	0	1	1	832	32	26656
	7	7	832	7	7	832	maxpool-a	maxpool4	1	1	3	3	832	832	0
	7	7	832	7	7	256	conv1x1c	maxpool4	1	0	1	1	832	256	213248
	7	7	96	7	7	320	conv3-3	conv1x1a	1	1	3	3	160	320	461120
	7	7	16	7	7	128	conv5x5	conv1x1b	1	2	5	5	32	128	102528
	7	7	256	7	7	128	conv1x1d	maxpool-a	1	0	1	1	832	128	106624
				7	7	832	depth-concat	conv1x1c, conv3x3, conv5x5, conv1x1d							
inception (5b)	7	7	832	7	7	192	conv1x1a	depth-concat	1	0	1	1	832	192	159936
	7	7	832	7	7	48	conv1x1b	depth-concat	1	0	1	1	832	48	39984
	7	7	832	7	7	832	maxpool-a	depth-concat	1	1	3	3	832	832	0
	7	7	832	7	7	384	conv1x1c	depth-concat	1	0	1	1	832	384	319872
	7	7	96	7	7	384	conv3-3	conv1x1a	1	1	3	3	192	384	663936
	7	7	16	7	7	128	conv5x5	conv1x1b	1	2	5	5	48	128	153728
	7	7	384	7	7	128	conv1x1d	maxpool-a	1	0	1	1	128	128	16512
				7	7	1024	depth-concat	conv1x1c, conv3x3, conv5x5, conv1x1d							
	7	7	1024	1	1	1024	avgpool	depth-concat	1	0	7	7	1024	1024	0
	1	1	1024	1	1	1000	fc	depth-concat	1	0	1	1	1024	1000	1025000
									Total						**6,414,360**

Figure 14.228: **Various type of CNN internal structures (author: Aqeel Anwar)**

14.1.20.8.8 Generative Adversarial Network (GAN)

A "generative adversarial network" (GAN) is a class of machine learning systems invented by Ian Goodfellow and his colleagues in 12014 (holocene calendar). Two neural networks contest with each other in a game (in the sense of game theory, often but not always in the form of a zero-sum game). Given a training set, this technique learns to generate new data with the same statistics as the training set. For example, a GAN trained on photographs can generate new photographs that look at least superficially authentic to human observers, having many realistic characteristics. Though originally proposed as a form of generative model for unsupervised learning, GANs seems also proven useful for semi-supervised learning, fully supervised learning and reinforcement learning.

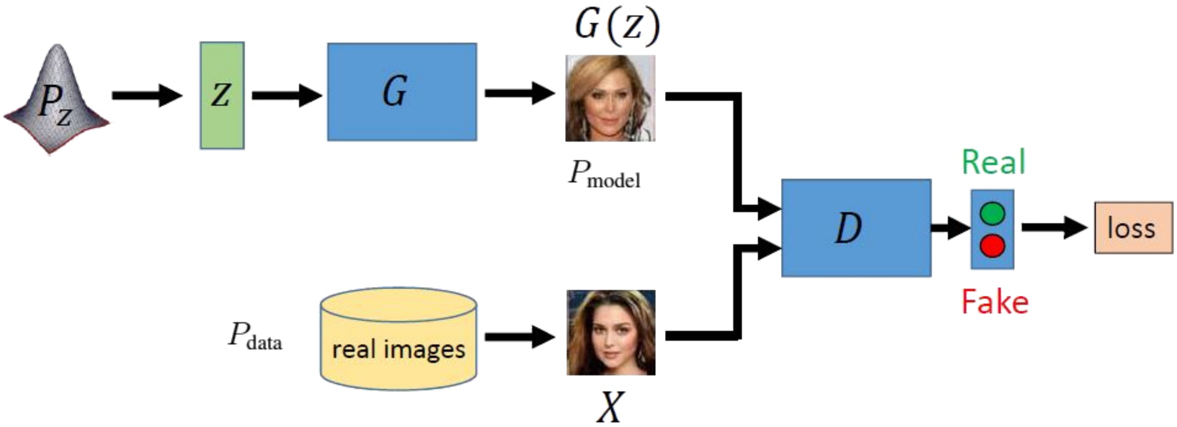


Figure 14.229: **General idea of a GAN**

The generative network generates candidates while the discriminative network evaluates them. The contest operates in terms of data distributions. Typically, the generative network learns to map from a latent space to a data distribution of interest, while the discriminative D network distinguishes candidates produced by the generator G from the true data distribution. The generative network's training objective is to increase the error rate of the discriminative network (i.e., "fool" the discriminator network by producing novel candidates that the discriminator thinks are not synthesized (are part of the true data distribution)).

A known dataset serves as the initial training data for the discriminator D. Training it involves presenting it with samples from the training dataset, until it achieves acceptable accuracy. The generator trains based on whether it succeeds in fooling the discriminator. Typically the generator is seeded with randomized input that is sampled from a predefined latent space (e.g. a multivariate Normal distribution). Thereafter, candidates synthesized by the generator are evaluated by the discriminator. Back-propagation is applied in both networks so that the generator produces better images, while the discriminator becomes more skilled at flagging synthetic images. The generator is typically a deconvolutional neural network, and the discriminator is a convolutional neural network.

Training a vanilla GAN is like a 2-players game, where:

- The discriminator is trying to maximize the cost function $L(D, G)$

- The generator is trying to minimize the cost function $L(D,G)$

The discriminator is the generator's opponent, and performs a mapping $D(x) \in [0,1]$. Its goal is to look at sample images (that could be real or synthetic from the generator), and determine if they are real samples ($D(x)$ closer to 1) or synthetic samples from the generator ($D(x)$ closer to 0). $D(x)$ can be interpreted as the probability that the image is a real training example.

The generator, $G(z)$, has parameters $\theta^{(G)}$, and the discriminator, $D(x)$, has parameters $\theta^{(D)}$. The generator can only control $\theta^{(G)}$, while the discriminator can only control $\theta^{(D)}$.

Remark

Keep in mind that $D(x)$ is simply a function that maps to $[0,1]$. So the possible values of $D(G(z))+D(x)$ range from 0 to 2.

In addition to this, the discriminator and generator have different cost functions they wish to optimize:

- This ought to be intuitive as the generator and discriminator have different goals.
- We denote the discriminator's cost function as $L^{(D)}\left(\theta^{(D)},\theta^{(G)}\right)$. For convenience, we will sometimes denote this as $L(D)$.
- We denote the generator's cost function as $L^{(G)}\left(\theta^{(D)},\theta^{(G)}\right)$. For convenience, we will sometimes denote this as $L(G)$

What is the solution?

- The discriminator wishes to minimize $L^{(D)}$, but can only do so by changing $\theta^{(D)}$
- The generator wishes to minimize $L^{(G)}$, but can only do so by changing $\theta^{(G)}$

This is slightly different from the optimization problems we've described thus far, where we have one set of parameters to minimize one cost function L.

Instead of treating this as an optimization problem, we treat this as a game between two players. The solution to a game is named a "Nash equilibrium". For GANs, a Nash equilibrium is a tuple, $(\theta^{(D)},\theta^{(G)})$ that is:

- A local minimum of $L(D)$ with respect to $\theta^{(D)}$
- A local minimum of $L(G)$ with respect to $\theta^{(G)}$

The simplest type of game to analyse is a zero-sum game, in which the sum of the generator's loss and the discriminator's loss is always zero. In a zero-sum game, the generator's loss is:

$$L(D)+L(G)=0 \Leftrightarrow L(D)=-L(G) \qquad (14.1.2138)$$

The solution for a zero-sum game is named a minimax solution, where the goal is to minimize the maximum loss. Since the game is zero-sum, we can summarize (see proof just below) the entire game by stating that the loss function is explicitly (based on binary cross-entropy as introduced at page 6463):

$$\max_D \min_G \left\{ \mathrm{E}_{x\sim p_{\text{data}}}\left[\underbrace{\log(D(x))}_{\substack{\text{output of the}\\ \text{discriminator for}\\ \text{real data}}}\right] + \mathrm{E}_{z\sim p_z}\left[\log\left(1-\underbrace{D(G(z))}_{\substack{\text{output of the}\\ \text{discriminatorfor}\\ \text{artificial data}}}\right)\right]\right\} \tag{14.1.2139}$$

Often condensed in the following form:

$$(D^*,G^*)=\max_D \min_G L(D,G) \tag{14.1.2140}$$

with obviously:

$$L(D,G)=\{\mathrm{E}_{x\sim p_{\text{data}}}[\log(D(x))]+\mathrm{E}_{z\sim p_z}[\log(1-D(G(z)))]\} \tag{14.1.2141}$$

- The discriminator wants to maximize the objective (i.e., its payoff) such that $D(x)$ is close to 1 and $D(G(z))$ is close to zero.
- The generator wants to minimize the objective (i.e., its loss) so that $D(G(z))$ is close to 1.

The Nash equilibrium (see section Games and Decision Theory page 7395) of this particular game is achieved at:

$$p_{\text{data}}=p_{\text{model}} \tag{14.1.2142}$$

Proof 14.8.2. Let us recall the definition of cross-entropy (see page 6463):

$$H(p,q)=\sum_i p_i\log(q_i)=\mathrm{E}(\log(q_i)) \tag{14.1.2143}$$

Let us rewrite it $H(y,D(x))$ where H remains the cross-entropy x is sampled either from p_{data} or from p_{model} with a probability of 50%. More formally:

$$x\sim\begin{cases} p_{\text{data}}\rightarrow y=1, & \text{with prob 0.5}\\ p_{\text{model}}\rightarrow y=0, & \text{otherwise}\end{cases} \tag{14.1.2144}$$

We consider y to be 1 if x is sampled from the real distribution and 0 if it is sampled from the fake one. Finally, $D(x)$ represents the probability with which D thinks that x belongs to p_{data}. By writing the cross-entropy formula we get:

$$H(y,D(x))=\mathrm{E}_y[-\log(D(x))]=\frac{1}{N}\sum_{i=1}^{N} y_i\log(D(x_i)) \tag{14.1.2145}$$

where N is the size of the dataset. Since each class has $N/2$ samples we can split this sum into two parts:

$$H(y, D(x)) = -\left[\frac{1}{N}\sum_{i=1}^{N/2} y_i \log(D(x_i)) + \frac{1}{N}\sum_{i=N/2}^{N}(1-y_i)\log((1-D(x_i)))\right] \quad (14.1.2146)$$

The first of the two terms represents the samples from the p_{data} distribution, while the second one the samples from the p_{model} distribution. Since all y_i are equally likely to occur, we can convert the sums into expectations:

$$H(y, D(x)) = -\left[\frac{1}{2}\mathrm{E}_{x\sim p_{\text{data}}}[\log(D(x))] + \frac{1}{2}\mathrm{E}_{x\sim p_{\text{model}}}[\log(1-D(x))]\right] \quad (14.1.2147)$$

At this point, we will ignore $1/2$ from the equations since it's constant and thus irrelevant when optimizing this equation. Now, remember that samples that were drawn from p_{model} were actually outputs from the generator (obviously this affects only the second term). If we substitute $D(x), x\sim p_{\text{model}}$ with $D(G(z)), z\sim p_z$ we will get:

$$H(y, D(x)) = -\left[\mathrm{E}_{x\sim p_{\text{data}}}[\log(D(x))] + \mathrm{E}_{z\sim p_z}[\log(1-D(G(z)))]\right] \quad (14.1.2148)$$

This is the final form of the discriminator loss.

■ Q.E.D.

Remark

This whole endeavour was to provide a mathematical formulation to training GANs. In practice there are maaaany tricks that are invoked to effectively train a GAN, that are not depicted in the above equations.

For GANs, in practice this game is implemented in an iterative numerical approach. It involves two steps:

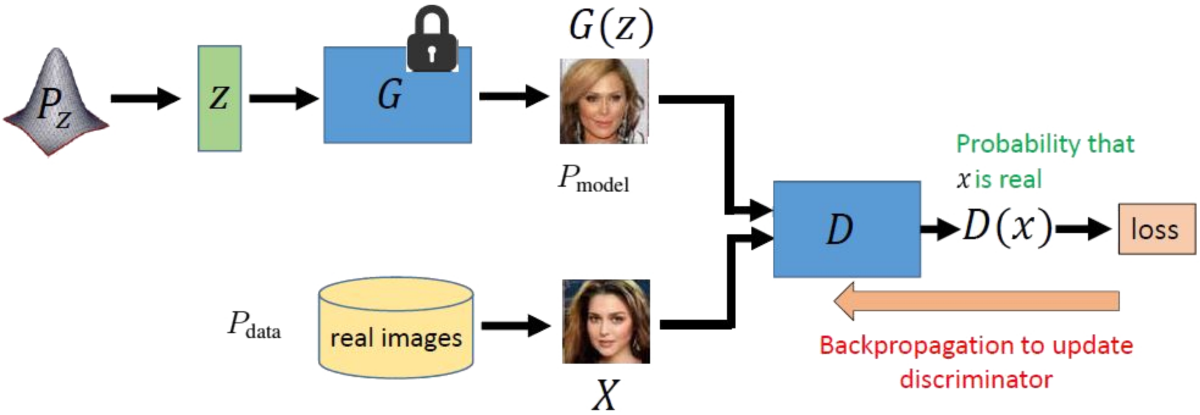


Figure 14.230: General idea of a GAN Discriminator training

Gradient ascent for the discriminator. We modify $\theta^{(D)}$ to maximize the minimax objective:

$$\max_D \left\{ \mathrm{E}_{x\sim p_{\text{data}}} \left[\underbrace{\log(D(x))}_{\substack{\text{output of the}\\ \text{discriminator for}\\ \text{real data}}} \right] + \mathrm{E}_{z\sim p_z}\left[\log\left(1-\underbrace{D(G(z))}_{\substack{\text{output of the}\\ \text{discriminatorfor}\\ \text{artificial data}}}\right)\right]\right\} \tag{14.1.2149}$$

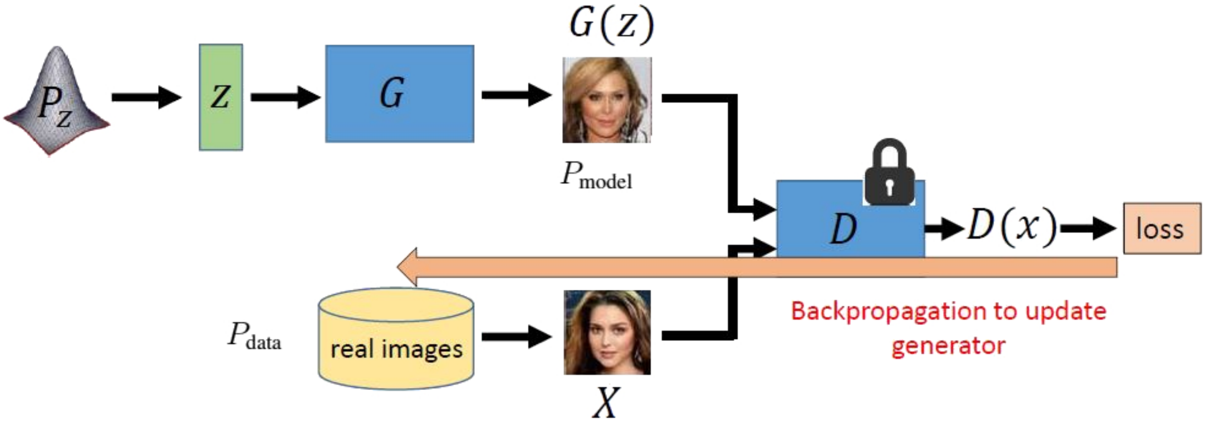


Figure 14.231: **General idea of a GAN Generator training**

Gradient descent on the discriminator. We modify $\theta^{(G)}$ to minimize the minimax objective:

$$\min_G \left\{ \mathrm{E}_{x\sim p_{\text{data}}} \left[\underbrace{\log(D(x))}_{\substack{\text{output of the}\\ \text{discriminator for}\\ \text{real data}}} \right] + \mathrm{E}_{z\sim p_z}\left[\log\left(1-\underbrace{D(G(z))}_{\substack{\text{output of the}\\ \text{discriminatorfor}\\ \text{artificial data}}}\right)\right]\right\} \tag{14.1.2150}$$

For the sake of intuition, consider that instead of optimizing with respect to $\theta^{(D)}$, we get to optimize $D(x)$ for every value of x. Further, assume that p_{data} and p_{model} are non-zero everywhere. What would be the optimal strategy for $D(x)$?

To answer this question, we differentiate with respect to $D(x)$ and set the derivative equal to zero. In particular, we start with:

$$L^{(D)} = \int p_{\text{data}}(x)\log(D(x))\,\mathrm{d}x + \int p_{\text{model}}(x)\log(1-D(x))\,\mathrm{d}x \tag{14.1.2151}$$

Differentiating with respect to $D(x)$, we get:

$$\frac{\partial L^{(D)}}{\partial D(x)} = -\int p_{\text{data}}(x)\frac{1}{D(x)}\mathrm{d}x - \int p_{\text{model}}(x)\frac{1}{1-D(x)})\mathrm{d}x \tag{14.1.2152}$$

We can now set the derivative equal to zero to find the optimal $D(x)$. This leads to:

$$\int p_{\text{data}}(x)\frac{1}{D(x)}\mathrm{d}x = \int p_{\text{model}}(x)\frac{1}{1-D(x)}\mathrm{d}x \tag{14.1.2153}$$

A solution occurs when the integrands are equal for all x, i.e.:

$$p_{\text{data}}(x)\frac{1}{D(x)}=p_{\text{model}}(x)\frac{1}{1-D(x)} \tag{14.1.2154}$$

Rearranging terms, this gives:

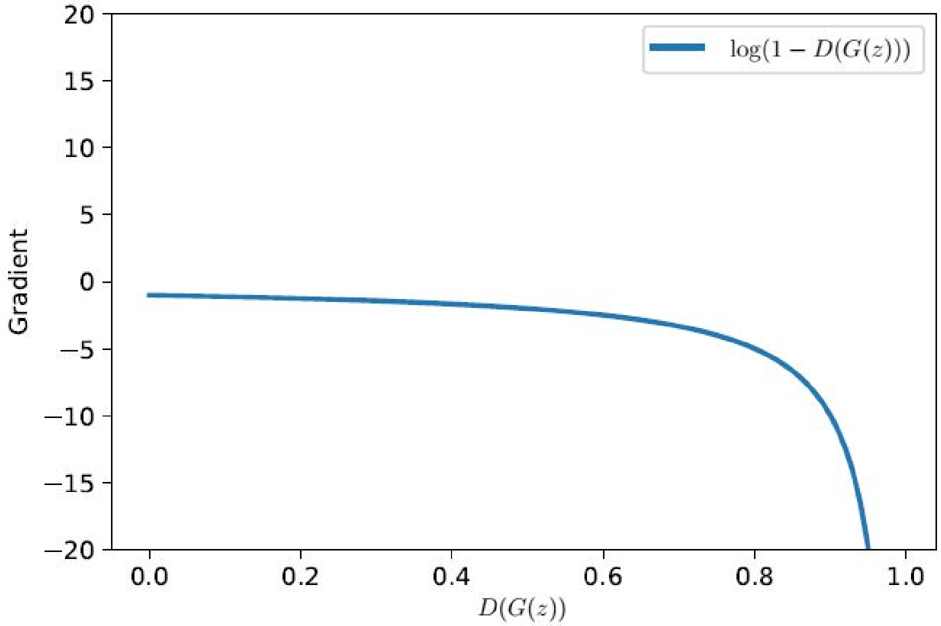
$$D(x)=\frac{p_{\text{data}}(x)}{p_{\text{data}}(x)+p_{\text{model}}(x)} \tag{14.1.2155}$$

If the generator has high enough capacity, it will then move to set $p_{\text{model}}(x)=p_{\text{data}}(x)$ for all x. This results in the output:

$$D(x)=\frac{p_{\text{data}}(x)}{p_{\text{data}}(x)+p_{\text{model}}(x)}=\frac{1}{2} \tag{14.1.2156}$$

This is the Nash equilibrium!

As stated, the current training paradigm has an important limitation. Note that the gradient of $\log(1-D(G(z)))$ is $-\frac{1}{1-D(G(z))}$, and thus has the following gradient as function of $D(G(z))$:



Therefore, if $D(G(z))\cong 0$, as may happen early on in training when the discriminator can tell the difference between real and synthetic examples, the gradient is close to zero. This results in little learning for $\theta^{(G)}$, and thus in practice the generator cost function:

$$\mathrm{E}_{z\sim p_z}\left[\log\left(1-D(G(z))\right)\right] \tag{14.1.2157}$$

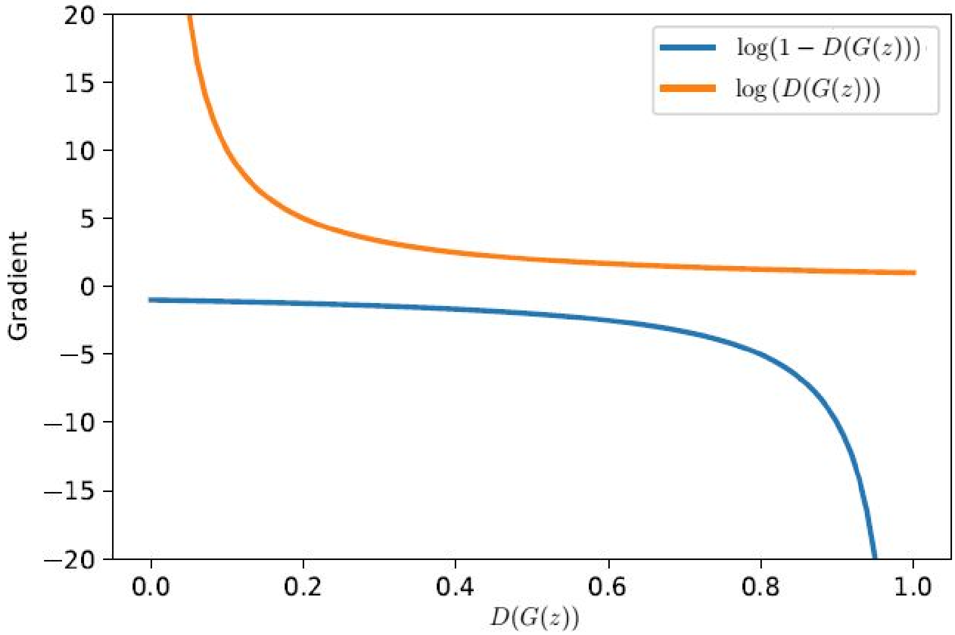
is rarely ever used.

Instead, we opt for a cost function that has a large gradient when $D(G(z))\cong 0$, so that the generator is encouraged to learn much more early in training. This is the cost function:

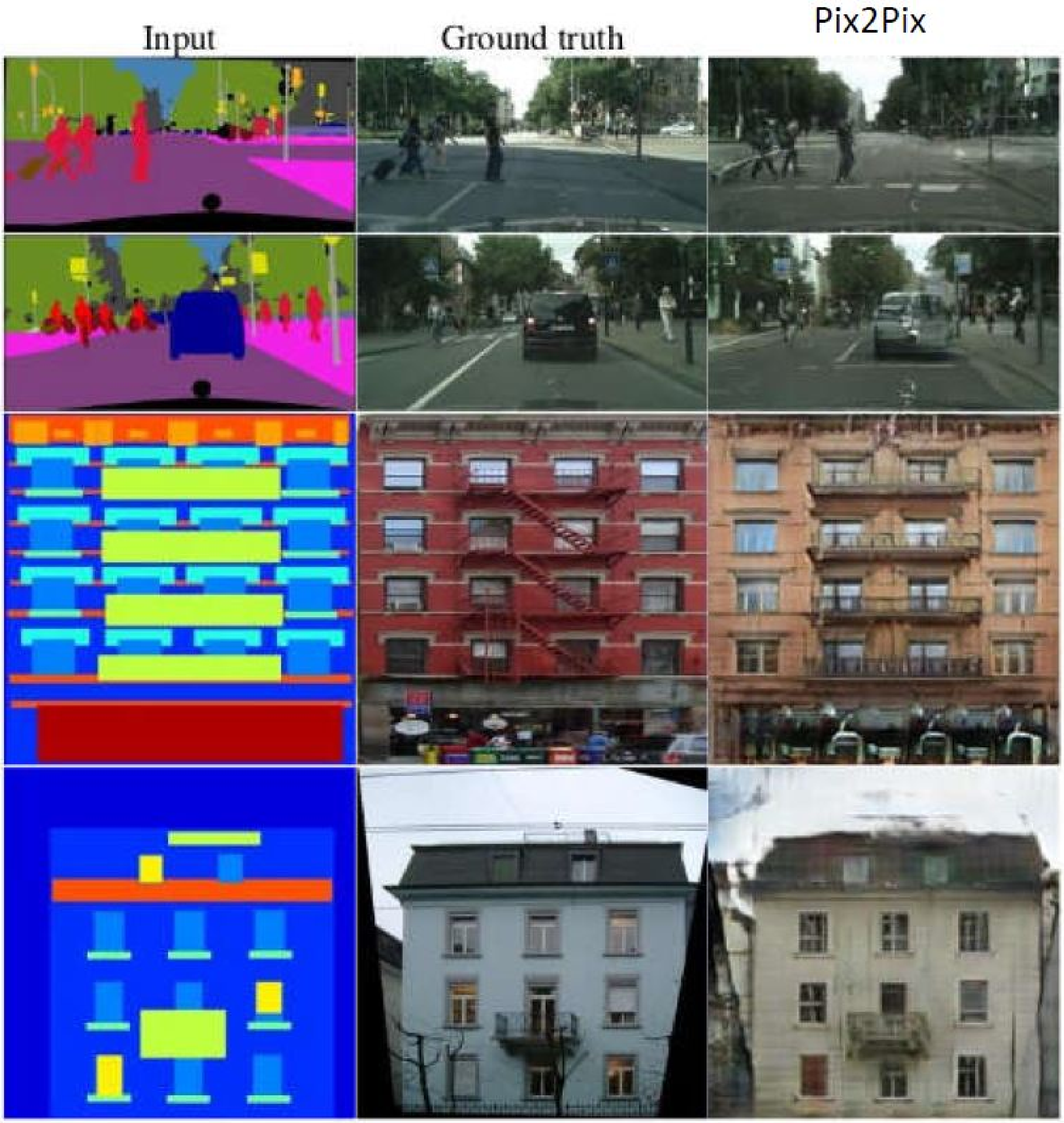
$$-\mathrm{E}_{z\sim p_z}\left[\log\left(D(G(z))\right)\right] \tag{14.1.2158}$$

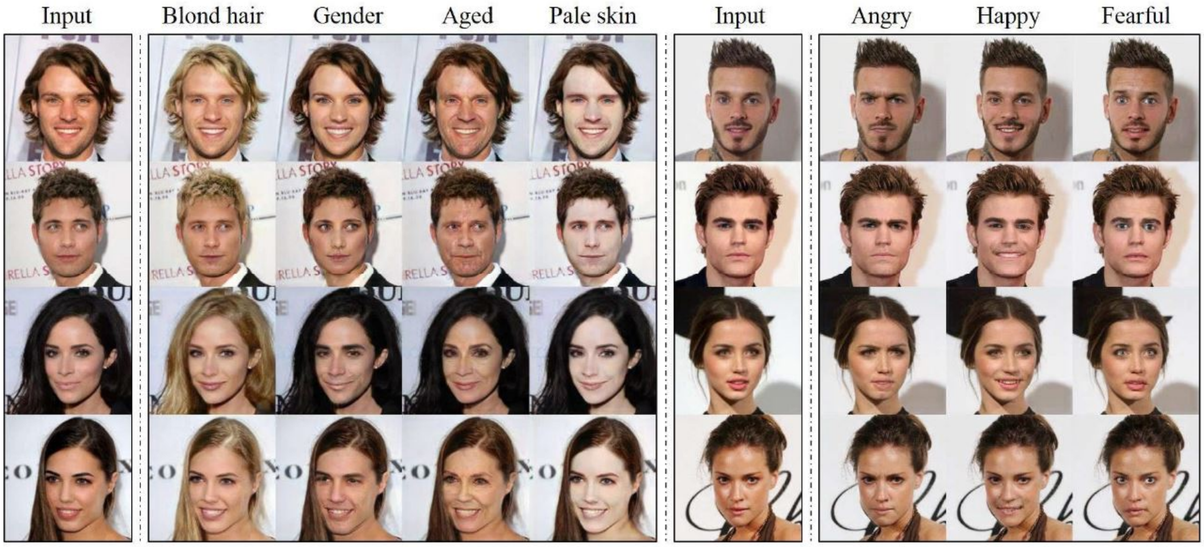
This still obtains the same overall goal of being minimized when $D(G(z)) = 1$, but now admits far more learning when the generator performs poorly.

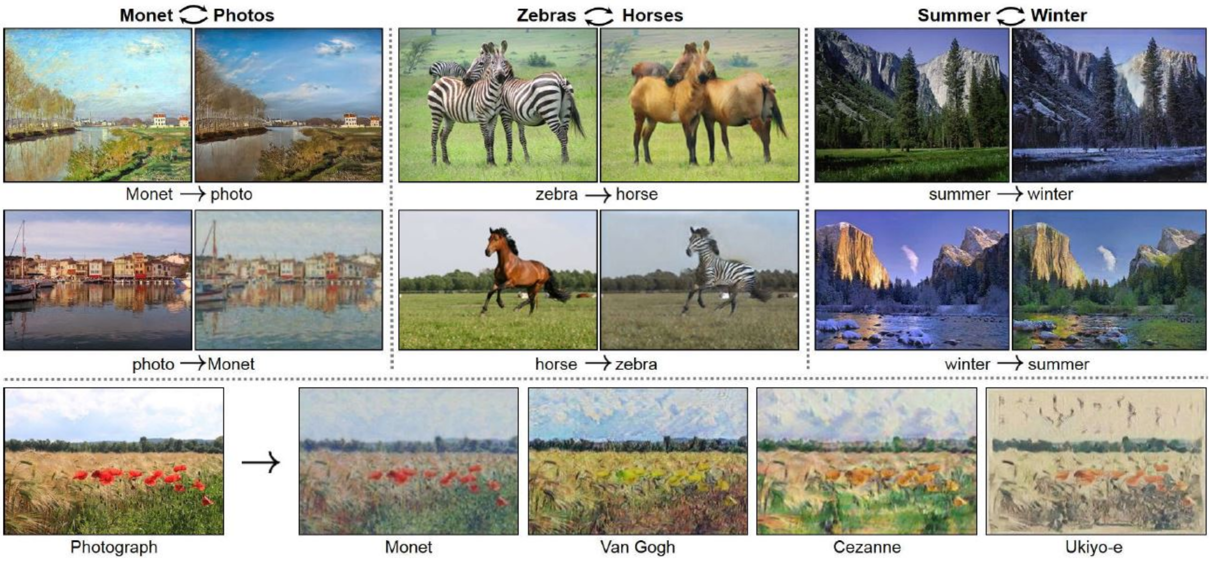
The gradient of this new cost for the generator encourages more learning when the generator performs poorly.



Note that by changing the cost function for the generator network, the game is no longer zero-sum. This is a heuristic change made to the game to solve the practical problem of saturating gradients when the generator isn't doing well!


Input
Ground truth
Pix2Pix


Input
Blond hair
Gender
Aged
Pale skin
Input
Angry
Happy
Fearful


Monet ⇄ Photos
Zebras ⇄ Horses
Summer ⇄ Winter
Monet → photo
zebra → horse
summer → winter
photo → Monet
horse → zebra
winter → summer
Photograph
Monet
Van Gogh
Cezanne
Ukiyo-e

Remark

There are largely empirical good practices to train a GAN network like the fact to normalize the images between -1 and $+1$, and use $\tanh$ as the output of the generator model, let the prior on z be Gaussian rather than uniform, avoid sparse gradients, avoid using ReLU or maxpool and use LeakyReLU instead and to downsample, increase the stride, etc.

14.1.20.8.9 Objective, cost and loss function

In machine learning (i.e. Statistics...), people talk about objective function, cost function, loss function[79]. These are not very strict terms and they are highly related. However:

- "Loss function", often denoted L, is usually a function defined on a single data point, prediction and label, and measures the penalty. For example:
  - Square loss (also named "quadratic loss", "L_2 loss" or MSE):

    $$L(f(x_i|\theta),y_i)=(f(x_i|\theta)-y_i)^2 \tag{14.1.2159}$$

    used in linear regression.
  - Mean Absolute Error (also named "L_2 loss" or MAE):

    $$L(f(x_i|\theta),y_i)=|f(x_i|\theta)-y_i| \tag{14.1.2160}$$

    used in linear regression.
  - Cross entropy loss:

    $$L(f(x_i|\theta),y_i)=-(y_i\log(p_i)+(1-y_i)\log(1-p_i)) \tag{14.1.2161}$$

    used in binary classification.
  - Hinge loss:

    $$L(f(x_i|\theta),y_i)=\max(0,1-f(x_i|\theta)) \tag{14.1.2162}$$

    used in SVM.
  - Huber loss:

    $$L_\delta=\begin{cases}\frac{1}{2}(y-\hat{y})^2 & \text{if } |(y-\hat{y})|<\delta\\ \delta((y-\hat{y})-\frac{1}{2}\delta) & \text{otherwise}\end{cases} \tag{14.1.2163}$$

    Typically used for regression. It's less sensitive to outliers than the MSE as it treats error as square only inside an interval. There is also a "pseudo-Huber loss" function and a "modified Huber loss".
  - $0/1$ loss:

    $$L(f(x_i|\theta),y_i)=1\Leftrightarrow f(x_i|\theta)\neq y_i \tag{14.1.2164}$$

    used in theoretical analysis.
  - ...
- "Cost function", often denoted J, is usually more general. It might be a sum of loss functions over your training set plus some model complexity penalty (regularization). For example:

---

[79] Not to be confused with the "activation functions" used in neural networks!

- Mean Squared Error:

$$\text{MSE}(\theta)=\frac{1}{N}\sum_{i=1}^{N}(f(x_i|\theta)-y_i)^2 \tag{14.1.2165}$$

- SVM cost function:

$$\text{SVM}(\theta)=||\theta||^2+C\sum_{i=1}^{N}\xi_i \tag{14.1.2166}$$

- ...

Or as we have seen earlier typically for neural networks the following cost functions:

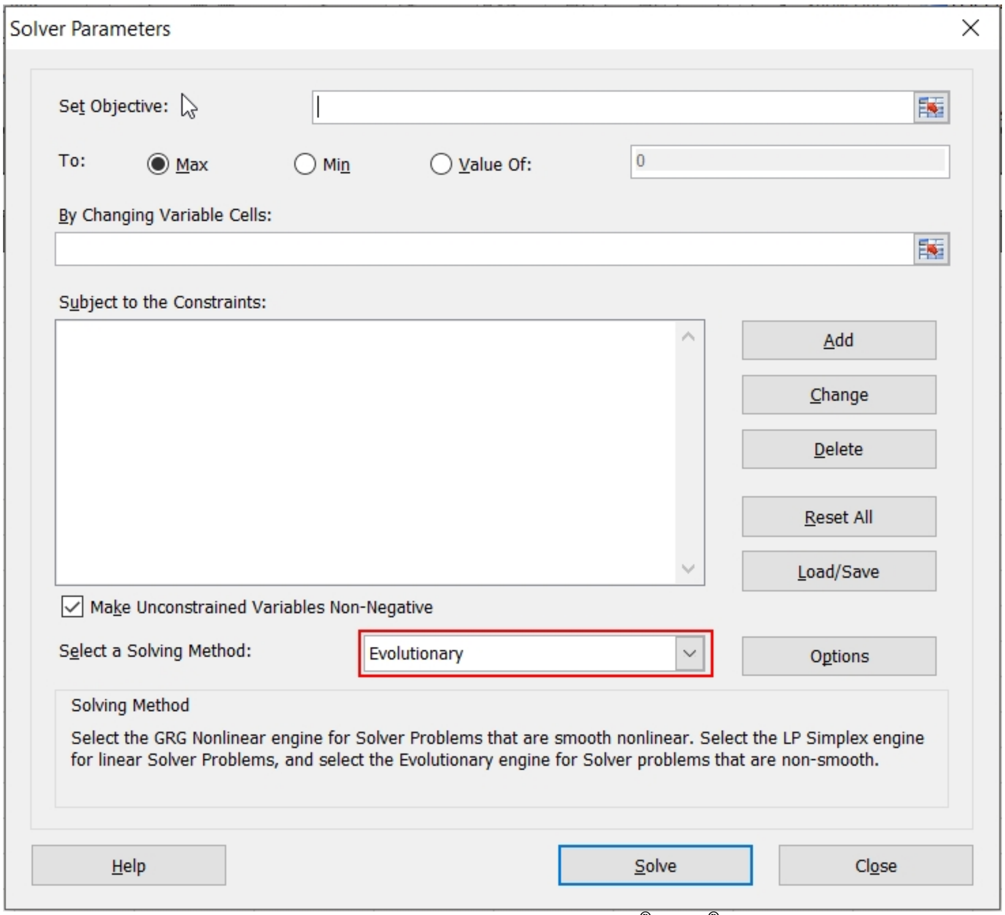
Cost function name	**Mathematical expression**	**Gradient**
Quadratic cost	$J_{\text{MST}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\frac{1}{2}\sum_{i=1}^{n}(\phi_i(z)-y_i)^2$	$\nabla_\phi J_{\text{MST}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\sum_{i=1}^{n}(\phi_i(z)-y_i)$
Cross-entropy cost	$J_{\text{CE}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=-\sum_{i=1}^{n}[y_i\ln(\phi(z))+(1-y_i)\ln(1-y_i)]$	$\nabla_\phi J_{\text{CE}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\sum_{i=1}^{n}\frac{\phi_i(z)-y_i}{(1-\phi_i(z))\phi_i(z)}$
Exponential cost	$J_{\text{exp}}\left(W,\vec{b},\vec{y},\vec{\phi},\tau\right)=\tau\exp\left(\frac{1}{\tau}\sum_{i=1}^{n}(\phi_i(z)-y_i)^2\right)$	$\nabla_\phi J_{\text{exp}}\left(W,\vec{b},\vec{y},\vec{\phi},\tau\right)=\frac{2}{\tau}J_{\text{exp}}\left(W,\vec{b},\vec{y},\vec{\phi},\tau\right)\sum_{i=1}^{n}\phi_i(z)-y_i$
Hellinger distance	$J_{\text{HD}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\frac{1}{\sqrt{2}}\sum_{i=1}^{n}\left(\sqrt{\phi_i(z)}-\sqrt{y_i}\right)^2$	$\nabla_\phi J_{\text{exp}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\frac{1}{\sqrt{2}}\sum_{i=1}^{n}\frac{\sqrt{\phi_i(z)}-\sqrt{y_i}}{\sqrt{\phi_i(z)}}$
Kullback-Leibler (KL) cost	$J_{\text{KL}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\sum_{i=1}^{n}y_i\log\left(\frac{y_i}{\phi_i(z)}\right)$	$\nabla_\phi J_{\text{GKL}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\sum_{i=1}^{n}\frac{\phi_i(z)-y_i}{\phi_i(z)}$
Generalized KL cost	$J_{\text{GKL}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\sum_{i=1}^{n}y_i\log\left(\frac{y_i}{\phi_i(z)}\right)-\sum_{i=1}^{n}y_i+\sum_{i=1}^{n}\phi_i(z)$	$\nabla_\phi J_{\text{GKL}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\sum_{i=1}^{n}\frac{\phi_i(z)-y_i}{\phi_i(z)}$
Itakura-Saito distance cost	$J_{\text{IS}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\sum_{i=1}^{n}\left(\frac{y_i}{\phi_i(z)}-\log\left(\frac{y_i}{\phi_i(z)}\right)-1\right)$	$\nabla_\phi J_{\text{IS}}\left(W,\vec{b},\vec{y},\vec{\phi}\right)=\sum_{i=1}^{n}\frac{\phi_i(z)-y_i}{\phi_i^2(z)}$
...	...	...

- "Objective function" is the most general term for any function that you optimize during training. For example, a probability of generating training set in maximum likelihood approach is a well defined objective function, but it is not a loss function nor cost function (however you could define an equivalent cost function). For example:
  - MLE (Maximum Likelihood Estimator) is a type of objective function (which you maximize)
  - Divergence between classes can be an objective function but it is barely a cost function, unless you define something artificial

Long story short, we may say that: A loss function is a part of a cost function which is a type of an objective function.

14.1.20.9 Genetic Algorithms

Genetic algorithms (GAs) are iterated stochastic optimization algorithms based on the mechanisms of natural selection and genetics belonging to the family of "evolutionary algorithms". This is an optimization technique that has spread widely since the beginning of the 121st century (holocene calendar) through the version 14.0.6123 of Microsoft® Excel® wherein the solver incorporates an evolutionary algorithm by default as shown by the screenshot below:



Solver Parameters

Set Objective:

To: ● Max ○ Min ○ Value Of: 0

By Changing Variable Cells:

Subject to the Constraints:

☑ Make Unconstrained Variables Non-Negative

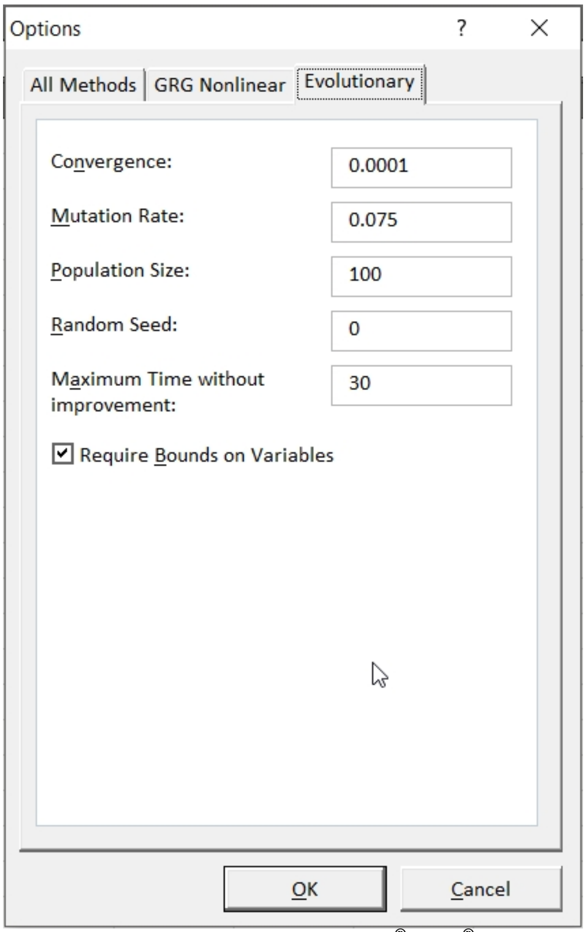
Select a Solving Method: Evolutionary

Solving Method

Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

Figure 14.232: **Screenshot of the evolutionary Microsoft® Excel® 14.0.6123 solver**

with the corresponding options:



Convergence: 0.0001

Mutation Rate: 0.075

Population Size: 100

Random Seed: 0

Maximum Time without improvement: 30

☑ Require Bounds on Variables

Figure 14.233: **Screenshot of the evolutionary Microsoft® Excel® 14.0.6123 solver options**

The process of the genetic algorithm is quite simple:

1. We start with an initial population of potential solutions (chromosomes) arbitrarily selected
2. We evaluate their relative performance (fitness)
3. Based on this performance, we create a new population of potential solutions using simple evolutionary operators: selection, crossover and mutation
4. We start this cycle until we find a satisfactory solution

Genetic Algorithms were originally developed by John Holland (11975 according to holocene calendar). This is the book of David E. Goldberg (11989 according to holocene calendar) that we own their popularization. Their fields of application are widespread. Besides the economy (portfolio risk minimization), they are used for optimization functions in finance, in optimal control theory (operational research), in the theory of repetitive and differentials games (namely: in evolutionary games and the prisoner's dilemma) and information retrieval (Google) and search for shortest path in graph theory (Internet routing or GPS). The reason for the large number of applications is clear: simplicity and efficiency. Of course, other stochastic exploration techniques exist, the Monte Carlo can be regarded as a similar concept.

To summarize, Lerman and Ngouenet (11995 according to holocene calendar) identified four main properties that make the fundamental difference between these algorithms and other methods:

1. The genetic algorithms use a coding of the input parameters, not the parameters themselves
2. Genetic algorithms work on a population of points, instead of a single point
3. Genetic algorithms use only the values of the function considered, not its derivative, or other auxiliary knowledge
4. The algorithms use probabilistic transition rules, not deterministic one

The simplicity of their mechanisms, their ease of implementation and efficiency even for complex problems led to a growing number of publication this recent years in the scientific community.

Definitions (#14.1.575):

D1. A "genetic algorithm" is defined by an individual / chromosome / sequence and a potential solution to the given problem.

D2. A "population" is a set of chromosomes or points of the search space

D3. The "environment" is assimilated with the search space

D4. The function that we seek to maximize is named "fitness function"

Before going further, we need to define more formally the above concepts but under the particular case of Binary coding!

The organisms in competition are the "individuals". Given an alphabet $A=\{a_1,a_2,\ldots,a_n\}$, we assume that each individual can be represented by a word of fixed-length l caught in the in A^*.

The word associated with an individual of the population will be named a "chromosome" or "sequence" (the term is not quite equivalent to its biological namesake, however, it is common practice to use the term here too) and thus given by A of the length $l(A)$ with $\forall i\in[1,l]:a_i\in A=\{0,1\}$ (reason: assumption of binary coding).

If there is no risk of confusion, we will identify the terms of "individual" and "chromosome".

The individuals form a population P of size P, denoted by:

$$P=\{A_i\} \tag{14.1.2167}$$

with $i=1\ldots N$. We will make another important statement, that is to say, there is a function f from one sequence with positive values which we denote $f(A)$, named "fitness function" that to any A_i associates real number such that for $i\neq j$:

$$f(A_i)>f(A_j) \tag{14.1.2168}$$

if and only if A_i is better suited to the environment than A_j.

Notice that the term "appropriate" is not defined. For this, we would characterize the environment in which the individuals evolve, what we will not do. In fact, since we assume the existence of such a function and we put it in equivalence to the degree of adaptation, it is automatically set by the definition of f.

We will name "generation" a population at time t, what must be put in relation with the notion of lifetime or age. However, we place ourselves here in the particular case where each individual has a life equal to 1, so the generation $(t+1)$ consists of different individuals from the generationt, we name them obviously the "descendants". Conversely, individuals of generation t are the "ancestors" of the individuals of the generation $(t+1)$. We denote the generation at time t by $P(t)$, thus the population at time t.

Thus, a chromosome is seen as a bit sequence in binary code known as "bit string". In the case of a non-binary coding, such as the real number encoding for example, then the sequence A contains only one point, we have then $A=\{a\}$ with $a\in\mathbb{R}$.

Remark

The fitness (effectiveness) is given by a function with real positive values. In the case of binary encoding, we will often use a function of decoding d that will gives the possibility to transform a binary string to a real number:

$$d:\{0,1\}^l\rightarrow\mathbb{R} \qquad (14.1.2169)$$

afterwards the fitness function is chosen such that it transforms this value into a positive value:

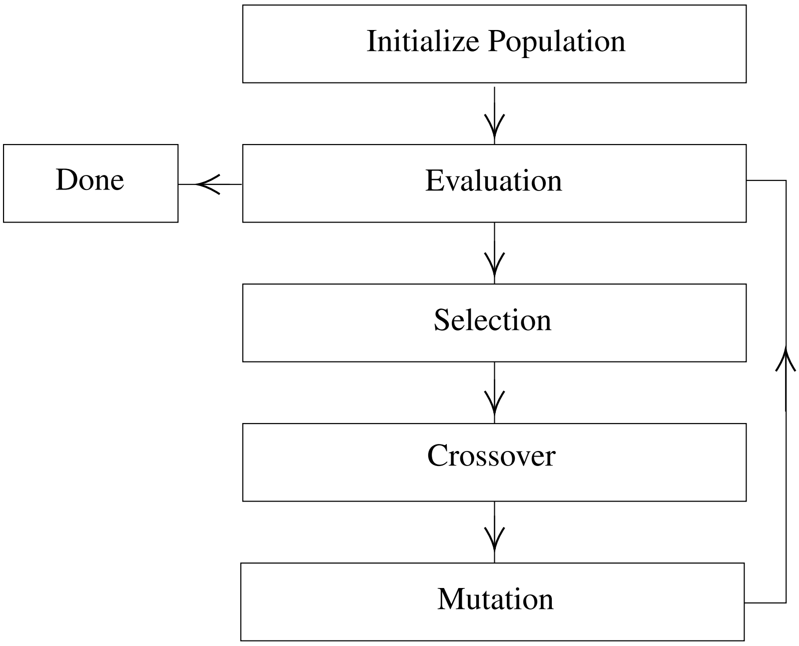
$$f:d\{0,1\}^l\rightarrow\mathbb{R}_+^* \qquad (14.1.2170)$$

The purpose of a genetic algorithm is then simply to find the string that maximizes this function f. Of course, each individual problem will require its own functions d and f.

GAs are then based roughly on the following phases:

1. Initialization: an initial population of N chromosomes is randomly chosen
2. Evaluation: each chromosome is decoded and evaluated
3. Selection: creation of a new population of N chromosomes based on previous step by using an appropriate method of selection.
4. Reproduction: possibility of crossover and mutation in the new population
5. Return to the Evaluation phase until the stop of the algorithm

Or for people that are more visual:



14.1.20.9.1 Encoding and Initial population

There exist three main types of coding:

1. Binary
2. Gray
3. Real

We can easily move from one encoding to another. Some authors do not hesitate, moreover, to draw parallels with biology, by speaking of "genotype" (see section Population Dynamics page 7361) regarding the binary representation of an individual, and "phenotype" (see section Population Dynamics page 7367) with respect to its corresponding real value in the search space.

Let us recall that the simplest transformation (decoding function d) of a binary string A into an integer x is done by the following rule (see section Numbers page 211 or this section page 6274):

$$x = d(A) = \sum_{i=0}^{l} a_i 2^i \qquad (14.1.2171)$$

where l is the number of digits of the string minus 1.

Therefore the chromosome $A = \{1, 0, 1, 1\}$ has trivially for value:

$$1 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0 = 8 + 2 + 1 = 11 \qquad (14.1.2172)$$

Obviously, the function needs to be adapted (by trial and error!) depending on the problem. Thus, if we seek to maximize a function $f : [0, 1] \to [0, 1]$ a possible method would be as follows (the size of the chromosome of course dependent on the desired accuracy):

$$x = d(A) = \sum_{i=1}^{l+1} a_i 2^{-i} \tag{14.1.2173}$$

This can be assimilated to a harmonic series (see section Sequences and Series page 1665). For accuracy to the fifth decimal place, we will put $l = 17 - 1$ since:

$$d\left(\left\{\underbrace{1,\ldots,1,\ldots,1}_{=17}\right\}\right) \cong 0.999992 \to 1 \tag{14.1.2174}$$

Again another way to do would be to choose d such as:

$$x = d(A) = \sum_{i=0}^{l} \frac{a_i 2^i}{2^{l+1}-1} \tag{14.1.2175}$$

Let us give an explanation of this choice:

$$\begin{aligned} &x^n - 1 = (x-1)(1 + x + x^2 + \ldots + x^{n-1}) \\ &\frac{x^n-1}{x-1} = 1 + x + x^2 + \ldots + x^{n-1} = \sum_{i=0}^{n-1} x^i \\ &2^n - 1 = \sum_{i=0}^{n-1} 2^i \Leftrightarrow 1 = \sum_{i=0}^{n-1} \frac{2^i}{2^n-1} \end{aligned} \tag{14.1.2176}$$

Let us put $l = n - 1$:

$$1 = \sum_{i=0}^{l} \frac{2^i}{2^{l+1}-1} \tag{14.1.2177}$$

So, with $l = 16$ we have $2^{17} - 1 = 131071$ and:

$$d\left(\left\{\underbrace{1,\ldots,1,\ldots,1}_{=17}\right\}\right) = \frac{131071}{131071} = 1 \tag{14.1.2178}$$

This last rule can be generalized. Thus, suppose that we seek to maximize ("normalize" would be a more suited term perhaps...) f according to a real variable x. Given $D = [x_{\text{min}}, x_{\text{max}}]$, with $D \subset \mathbb{R}$, the allowed search space with x_{min} and x_{max} the lower and upper bounds of this space. Given prec the precision (decimal) with which we seek x. Given:

$$R = x_{\text{max}} - x_{\text{min}} \tag{14.1.2179}$$

the length (range) of the interval D. We then have to divide this interval at worst in:

$$n = R \cdot 10^{\text{prec}} \tag{14.1.2180}$$

equal sub-intervals to meet accuracy expectations. For example, given $D = [-1, 2]$ so we have $R = 3$, if we wanted a precision $\text{prec} = 6$, then we must divide this interval in $n = 3,000,000$ sub-intervals.

Let k denote the natural integer such that $2^k > n$, which in our example involves $k = 22$ as:

$$2^{21} = 2,097,152 < 3,000,000 < 2^{22} = 4,194,304 \tag{14.1.2181}$$

the transformation of a binary string $A = \{a_1, \ldots, a_l\}$ in a real number x can then be run in three stages:

1. Conversion (base 2 into base 10):

$$x' = \sum_{i=0}^{l} a_i 2^i \tag{14.1.2182}$$

2. Normalization:

$$x'' = x' \frac{x_{\max} - x_{\min}}{2^{l+1} - 1} = \frac{x'R}{2^{l+1} - 1} \tag{14.1.2183}$$

3. Maximization:

$$x = x_{\min} + x'' \tag{14.1.2184}$$

Or what remains the same directly in one step by using:

$$x = x_{\min} + \sum_{i=0}^{l} \frac{a_i 2^i R}{2^{l+1} - 1} \tag{14.1.2185}$$

Therefore for $f : [0,1] \to [0,1]$ and $\forall i, a_i = 1$ we fall back well on:

$$x = 0 + \sum_{i=0}^{l} \frac{2^i \cdot 1}{2^{l+1} - 1} = \sum_{i=0}^{l} \frac{2^i}{2^{l+1} - 1} \tag{14.1.2186}$$

About the initialization phase, the procedure is quite simple. It consists of a random selection of N individuals in the space of allowed individuals. In binary coding, according to the size l of the string, we do for a chromosome l sampling in $\{0,1\}$ with equal probability.

14.1.20.9.2 Operators

Operators play a key role in the possible success of a GA. We number three main one:

1. the selection operator
2. the crossover operator
3. the mutation operator

If the principle of each of these operators is easy to understand, it is difficult to explain the isolated importance of each of these operators in the success of the AG. This is partly due to the fact that each of these operators acts according to various criteria that depends on its own characteristics (fitness of individuals, likelihood of activation of the operator, etc.).

14.1.20.9.2.1 Operator of selection

This operator may be the most important since it allows individuals in a population to survive, reproduce or die. Generally, the probability of survival of an individual will be directly connected to its relative effectiveness in the population. The basic part of the selection process is to stochastically select from one generation to create the basis of the next generation.

There are several methods for reproduction. The most known and used method is undoubtedly the biased David E. Goldberg's (11989 according to holocene calendar) lottery wheel (roulette wheel). According to this method, each chromosome is duplicated in a new population in proportion to its adaptive value. We perform in some way, as many sampling that there are individuals in the population. Thus, in the case of a binary coding, the fitness of a particular chromosome being $f(d(A))$, the probability with which it will be reintroduced into the new population of size N is given by the relative fitness:

$$\frac{f(d(A))}{\displaystyle\sum_{i=1}^{N} f(d(A))} \tag{14.1.2187}$$

Individuals with high fitness value thus have more chance of being selected by the wheel. We speak then of "proportional selection":

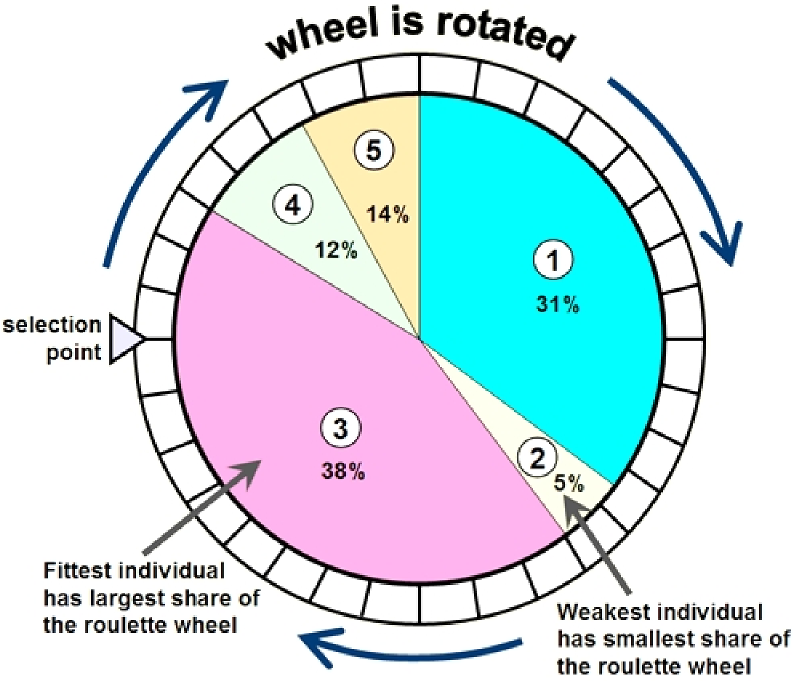

wheel is rotated
selection point
5
14%
4
12%
1
31%
3
38%
2
5%
Fittest individual has largest share of the roulette wheel
Weakest individual has smallest share of the roulette wheel

Figure 14.234: **Example of David E. Goldberg's wheel with five individuals with their respective relative fitness**

Obviously the number of times the roulette wheel is spun is equal to the size of the new population.

Each time the wheel stops this gives the fitter individuals the greatest chance of being selected for the next generation and subsequent: "mating pool".

The major drawback of this method lies in the fact that an individual that is not the best may still dominate the selection (imagine the search for maxima of a function in $\mathbb{R}^2$, there may be several of them - maxima - and therefore we could get a wrong selection ...), we will speak rightly then of "premature convergence" and this is one of the most common problems when using genetic algorithms. It can therefore also result in a loss of diversity by the domination of a super-individual. Another drawback is its poor performance towards the end when all individuals are alike.

One solution to this problem lies not in the use of another method of selection but the use of a modified fitness function. So we can use a scaling to decrease or increase artificially the relative difference between the fitness of individuals.

Briefly, there are other methods, the best known being that of the tournament (tournament selection) we draw two random individuals in the population and reproduce the best of both in the new population. We repeat this procedure until the new population P is complete. This method gives good results. However, as important as the selection phase, it does not create new individuals in the population. This is the role of crossover and mutation operators.

14.1.20.9.2.2 Crossover operator

The crossover operator allows the creation of new individuals in a very simple process. It allows the exchange of information between chromosomes (individuals). First, two individuals, which then form a couple, are sample in the new population issued from the selection (or reproduction). Then one (or potentially many) crossing site is randomly draw (number between 1 and $l-1$). Finally, according to a probability P_c that the crossing is done, the end segments (in the case of a single crossing site) of both parents are then exchanged around this site:

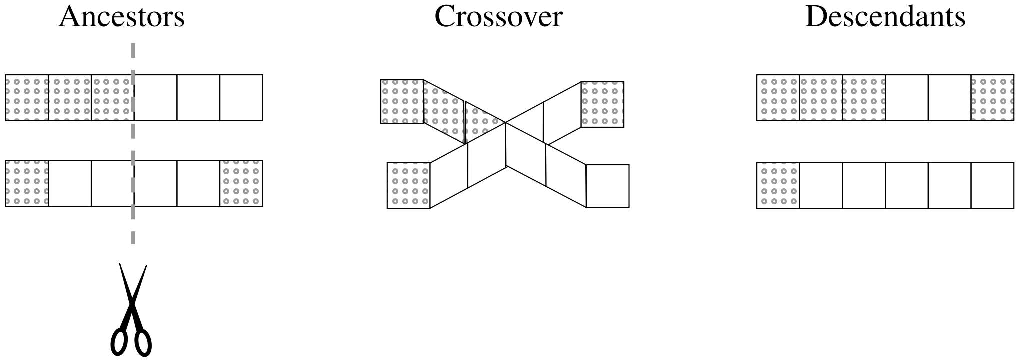


Figure 14.235: **Illustrative example of crossover**

This operator allows the creation of two new individuals. However, an individual selected in the reproduction (selection) is not necessarily subjected to a crossover. The latter is carried out with a certain probability P_c. The more this probability is high and the more the population will undergo a crossover modification.

Anyway, it is possible that the joint action of reproduction and the crossing is insufficient to ensure the success of the GA. Thus, in the case of binary encoding we have chosen so far, some information (i.e. the characters of the alphabet) may disappear from the population. Thus if no individual of the initial population contains a 1 in the last position of the string and that we know a priori that this 1 in the last position is part of the optimal string to find, all possible crosses will never show this 1 initially unknown. In real number coding, such a situation can happen when using a simple crossover operator, it was such that the initial population was between 0 and 40 and that the optimal value was 50. All possible combinations of convex digits belonging to the range $[0, 40]$ will never allow to reach a the number of 50. This is to address, among others, this problem that the mutation operator is used.

As usually the crossover operation uses 2 individuals. So, if you have 20 individuals, we choose 10 pairs to cross, and with a probability of $P_c = 80\%$, on average we're going to cross only 8 pairs.

14.1.20.9.2.3 Mutation operator

The purpose of this operator is to change randomly, with some probability, the value of a component of the individual. In the case of binary encoding, each bit $a_i \in \{0, 1\}$ is replaced following a probability P_m by its inverse $a'_i = 1 - a_i$. This is what is shown in the figure below. Like many crossover positions may be possible, we can very well admit that a same string can undergo several mutations.

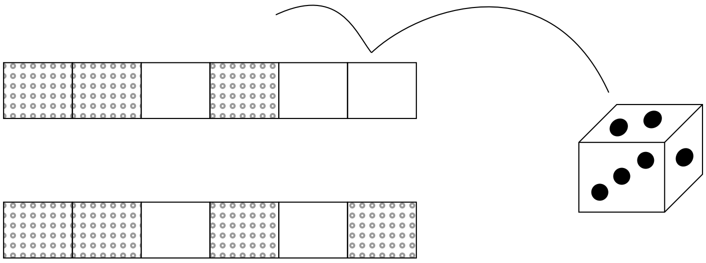


Figure 14.236: Illustrative example of mutation

The mutation is traditionally considered a marginal operator even if somehow it gives to genetic algorithms the ergodic property (i.e. all points of the search space can be achieved). However operator is of great importance. It performs a dual role: perform a local search and / or out a global search (remote search).

The operators of the genetic algorithm are guided by a number of parameters set in advance. The value of these parameters affect the success of failure of a genetic algorithm. These parameters are (the reader can compare this list with the parameters available in the Microsoft® Excel® solver screenshot given earlier above):

- The size of the initial population N, and the coding length l of each individual (in the case of binary encoding). If N is too large, the computation time of the algorithm can be very important, and if N is too small, it may converge too quickly to the wrong chromosome.
- The crossover probability P_c, that depends on the form of the fitness function. Its choice is general heuristic (just like P_m). The higher it is, the more the initial population obviously undergoes significant changes. The generally accepted values are between 0.5 and 0.9.

- The probability of mutation P_m is generally small since a high rate may lead to a suboptimal solution.

  Rather than reducing P_m, another way to avoid the best individuals to be altered is to use explicit report of the elite individuals in a certain proportion. So often, the top 5%, for example, of the population is directly reused directly, the operator of reproduction (selection) or mutation operating then only on the remaining 95%. This is named an "elitist strategy".

Let us now see an example of GA (example of David E. Goldberg - 11989 according to holocene calendar).

✍**Example:**

We want to find the maximum of the function $(f)=x$ on the interval $[0,31]$ where x is an integer. The first step consist in coding the function. For example, we use a binary coding of x, the sequence (chromosome) containing a maximum of 5 bits. Thus, we have $x=2\rightarrow\{0,0,0,1,0\}$, and also $x=31\rightarrow\{1,1,1,1,1\}$. We are therefore seeking maximum of a fitness function (we will choose $f(x)$ itself in this simple example) in a space of $2^5=32$ possible values of x.

1. Sampling and evaluation of the initial population

   We set the size of the population to $N=4$. We draw randomly 4 chromosomes knowing that a chromosome consists of 5 bits, and each bit has a 50% probability of having a value of 0 or 1. The maximum, (randomly) 16 is reached by the second sequence. Let us see how the algorithm will try to improve this result.

   First, we get the following table:

N°	Chromosome	Value	Fitness	$P_i\%$
1	00101	5	5	14.3
2	10000	16	16	45.7
3	00010	2	2	5.7
4	01100	12	12	24.2
Total			**35**	**100**

Table 14.33: **Evolution (mutation) of chromosomes**

   We turn again the David E. Goldberg's wheel 4 times to obtain the following sequence:

Selection (sampling)	Chromosome
1	10000
2	01100
3	00101
4	00101

Table 14.34: **Sampling sequence of chromosomes**

We see here well the risk that we would have to lose the Sequence N°2 from the start... that's the problem with this method. It can converge more slowly than others. However, the reader will notice that we have lost the sequence N°3.

We now turn to the crossover part: the ancestors are randomly selected. We randomly draw a crossover location ("site" or "loci") in the sequence. The crossing then operates at this location with a probability P_c. The table below shows the consequences of this operator assuming chromosomes 1 and 3, afterwards 2 and 4 are paired, and each time the crossing takes place (e.g. with $P_c = 1$):

	$l=2$	$l=3$
Original Sequences	100|00	01|100
	001|01	10|000
Crossed Sequences	10001	01000
	00100	10100

Table 14.35: **Chromosome crossing**

We now turn to the mutation part: in this binary coding example, the mutation is the occasional random modification (low probability) of the value of a bit (bit reversal). We thus draw for each bit a random number between 0 and 1 and if this digit is less than P_m then the mutation takes place. The table below with $P_m = 0.05$ highlights this process:

Old chromosome	Random drawing	New bit	New chromosome
10001	15 25 36 *04* 12	1	10011
00100	26 89 13 48 59	–	00100
01000	32 45 87 22 65	–	01000
10100	47 *01* 85 62 35	1	11100

Table 14.36: **Mutation of chromosomes**

Now that the new population is fully created, we can evaluate it again:

N°	Chromosome	Value	Fitness $f(x)$	P_i%
1	10011	19	19	32.2
2	00100	4	4	6.8
3	01000	8	8	13.5
4	11100	28	28	47.5
Total			**59**	**100**

Table 14.37: **Evaluation of the mutation of chromosomes**

The maximum is now 28 (N°4). So we went from 16 to 28 after a single generation. Of course, we must repeat the procedure from the selection stage until the overall maximum, 31, is obtained, or until that a stop criterion has been satisfied.

Remark

It is possible to prove mathematically, what's remarkable (!!!), that the portions of chromosomes that are found in the best individuals will tend to reproduce ...

The reader interested can also take a look to the MATLAB® companion book where we use GAs to find the optimum of the Rastriging function:

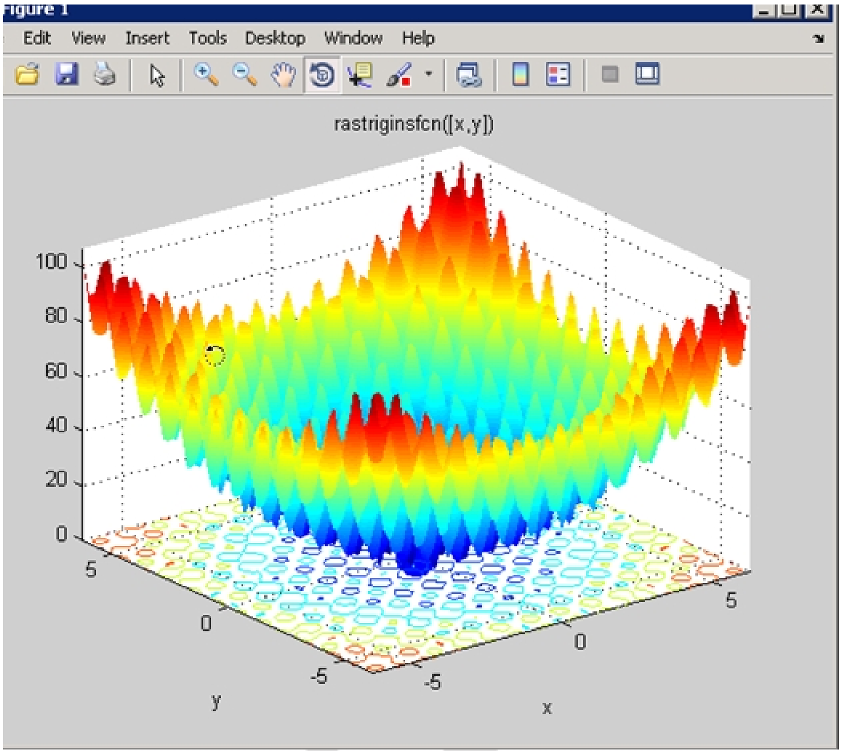


Figure 1
Edit View Insert Tools Desktop Window Help
rastriginsfcn([x,y])
100
80
60
40
20
0
5
0
-5
y
-5
0
5
x

Figure 14.237: **Rastriging function in MATLAB® 2013a**

where it works quite well as shown it the result below:

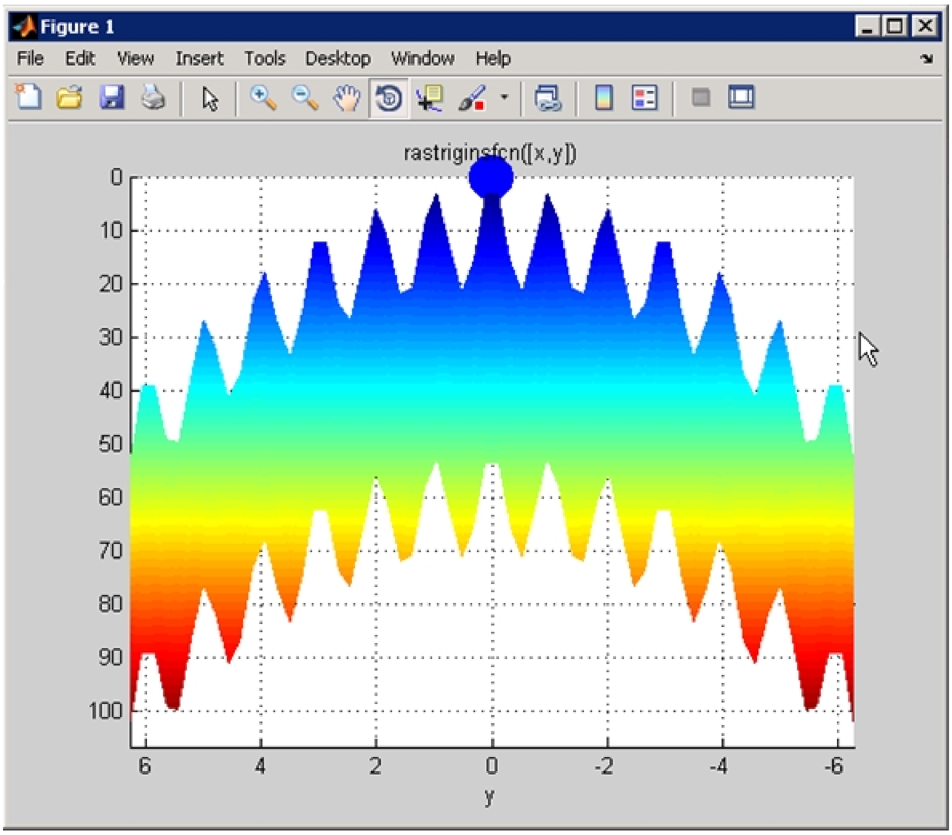


Figure 1
File Edit View Insert Tools Desktop Window Help
rastriginsfcn([x,y])
0
10
20
30
40
50
60
70
80
90
100
6
4
2
0
-2
-4
-6
y

and also for neural networks optimization.

14.1.20.10 Total Unduplicated Reach and Frequency Analysis (TURF)

"TURF Analysis", an acronym for "Total Unduplicated Reach and Frequency", is a type of statistical analysis used for providing estimates of media or market potential and devising optimal communication and placement strategies given limited resources. TURF analysis identifies the number of users reached by a communication, and how often they are reached.

Although originally used by media schedulers to maximize reach and frequency of media spending across different items (print, broadcast, etc.), TURF is also now used to provide estimates of market potential. For example, if a company plans to market a new yoghurt, they may consider launching 10 possible flavours, but in reality, only three might be purchased in large quantities. The TURF algorithm identifies the optimal product line to maximize the total number of consumers who will purchase at least one SKU[80] (Stock Keeping Unit). Typically, when TURF is undertaken for optimizing a product range, the analysis only looks at the reach of the product range (ignoring the Frequency component of TURF).

In order to obtain data on the items being evaluated, ratings/choices may be obtained via quantitative marketing research (such as a survey).

Let us give a companion example! Consider we want to sell three new ice flavours. We did a non-representative survey on 10 people that lead us to the following table:

Individual	**Weight**	**Chocolate (A)**	**Vanilla (B)**	**Pistachio (C)**
1	1	1	0	0
2	1	1	0	0
3	1	1	0	0
4	1	0	1	0
5	1	0	1	0
6	1	0	1	0
7	1	1	1	0
8	1	1	1	0
9	1	0	1	1
10	1	1	1	1

Three possible ice cream flavours available. After a survey, we get:

- Flavour A was chosen by 60% (6/10) of the people
  - 50% (3/6) chose A exclusively
  - $33.\bar{3}\%$ (2/6) chose also B
  - 0% (0/6) chose also C
  - $16.\bar{6}\%$ (1/6) chose also B plus C
- Flavour B was chosen by 70% (7/10) of the people

---

[80] A SKU is a distinct string of letters and numbers that helps retailers identify every product in their inventory and each product's specific traits, like its manufacturer, brand, price, style, color, and size.

  - 42.85% (3/7) chose B exclusively
  - 28.57% (2/7) chose also A
  - 14.29% (1/7) chose also C
  - 14.29% (1/7) chose also A plus C
- Flavour C was chosen by 20% (2/10) of the people
  - 0% (0/2) chose C exclusively
  - 0% (0/2) chose also A
  - 50% (1/2) chose also B
  - 50% (1/2) chose also A plus B

An obvious "macro-result" if we would have to choose only 2 flavour, would be to choose flavours A and B.

However we must also focus on the combinations focus on all combinations of 2 among 3 flavours to check if it match our previous naive macro analysis (and that's may not be always the case!!!).

We then have:

$$C_k^n = \frac{n!}{k!(n-k)!} = \frac{3!}{2!(3-2)!} = 3 \qquad (14.1.2188)$$

to analyse! It's easy to do by hand and lead us to the following reaching rates:

- 100% Would chose A alone, B alone or A and B together
- 70% Would chose A alone, C alone or A and C together
- 70% Would chose B alone, C alone or B and C together

So in this special case. The TURF analysis confirms the naive macro analysis!

Remark

The reader can find this example in our `R` companion book.

TURF analysis allows us to see the effect of the combinations of answers that people make that we may not see in simple percentage choices.

Making decisions based on simple percentage choice can be wrong. TURF analysis then allows us to take into account the complex combinations of choices that people make when selecting products or services.

Obviously, TURF analysis, has many flaws, among other the fact that is purely deterministic. But this can be quite simply resolved using bootstrapping.

~60 % ★★★☆☆
23 votes, 58.26%

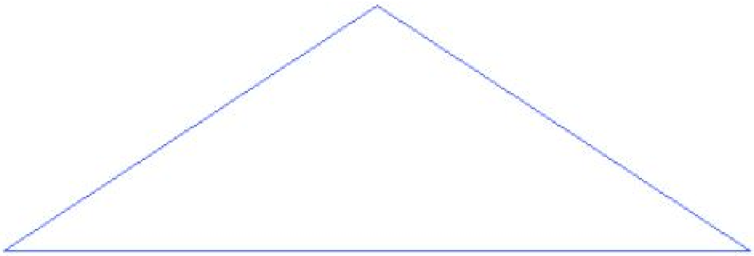
14.2. Fractals

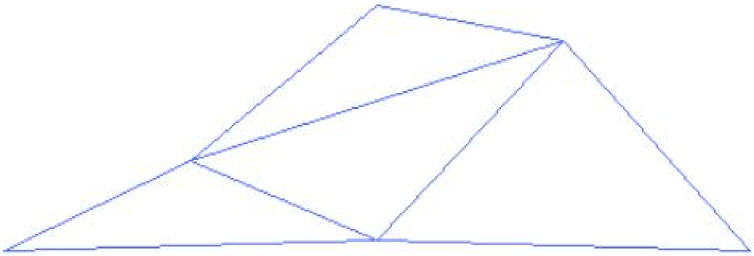
FRACTALS are figures invariant by scale change (we also talk about "self-similar structures") and are the graphic representation of contractant recurrent sequences (for IFS fractals that we will see later) or not divergent (for escape-time fractals as we will see further below).

The basic idea - simple and great ... at the same time - often involves taking a starting point, to build its image through a particular mathematical function, to take the image of the image and so on. The goal is to study how the successive points are allocate in the global target set of the defined function, if they are approaching a limit or if they roam between different values that can we explain, if more points in part of the set than another?

The advantage of this type of questions concerns both the study of the evolution of biological populations than the future of the solar system, 3D computing (the origin being the generation of mountains for 3D landscapes) changes in stock prices or random number generation in particular fields, or even medical diagnostic (especially for brain or heart).

See below a simple example image by image:





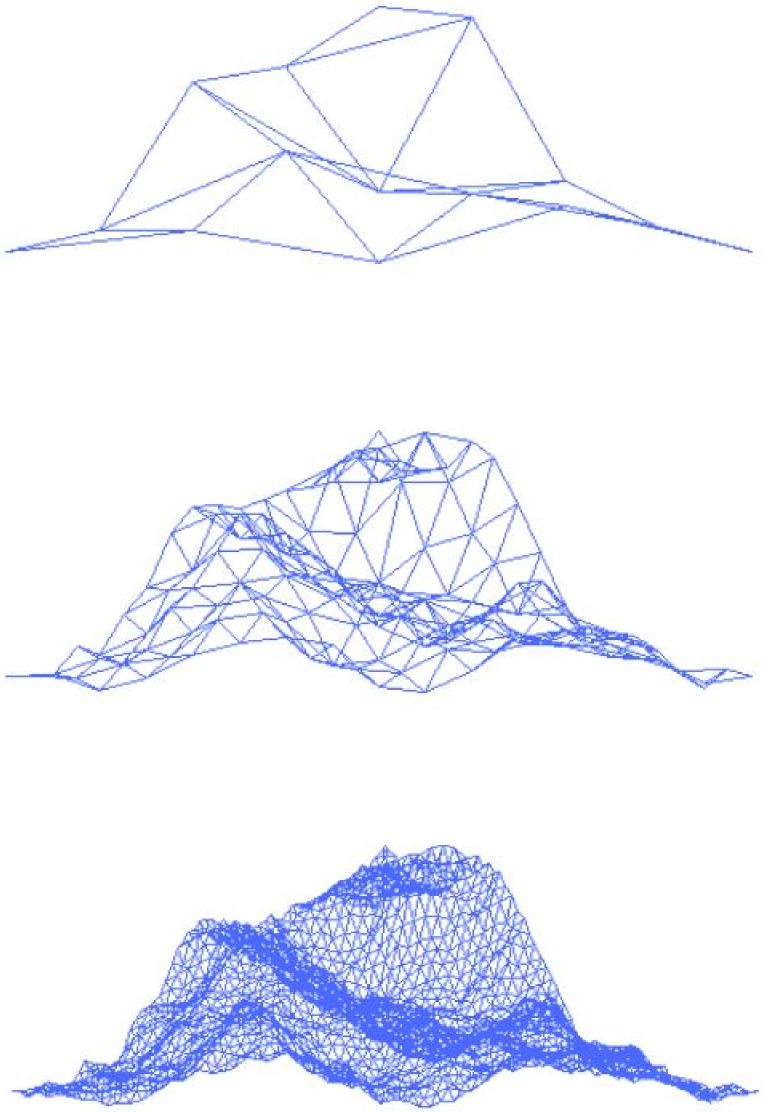


Figure 14.238: **Pseudo-mountains generation from a random fractal (probabilistic fractal)**

For the average person, fractals are used to look pretty. But they have far more serious applications: for example we already saw in this textbook that some of these "attractive" images reproduced physical phenomena (population dynamics for the Feigenbaum's Fractal, turbulence in a fluid with the Lorentz attractor, distribution of galaxies, L-Fractals, clusters and supercluster of galaxies, ...). Fractals have also found applications in music (with software generating fractal music) and movies (3D to generate mountains, fire, grass) productions. In the field of computer graphics, fractals are used to compress images very effectively, with consistent quality regardless of the zoom, they help to create realistic textures, and can afford to dither an image with good results. Fractals are also used to reduce the size of the receive antennas and to extend their effectiveness frequency spectrum (some of our cell phones of the early 121st century (holocene calendar) have fractal receptor of the type "Sierpinski carpet" - see below - because of all types of frequencies they need to manage!). In civil engineering fractals are used for building some sound absorbers walls. And many other things... Finally, in cosmology, fractals are also used to attempt to give a maximum structure size to check if the cosmological principle really holds in our observable universe (see [583]).

This fractal geometry differs from the Euclidean geometry first by its definition: the figures of Euclidean geometry are generally determined by algebraic relations, while the fractal curves are defined recursively as we have already mentioned. Fractals also have fractional dimensions (we have already discussed this topic in the section of Euclidean geometry when defining the concept of dimension). On the other hand, we must not neglect their autosimilar appearance: each part of a fractal can be observed at any scale: each part is (essentially) a copy of the whole.

Remark

The developments that follow could easily have been put in the section of Sequences and Series or even of Functional Analysis or seen as a special case of the section Topology reduces to the Euclidean space (this is why you will find here also many references to the topology section). Our choice is pedagogical as well as for the section of Cryptography, in the sense that it is much more interesting for a high-school student to see an application of abstract concepts of topology in a practical framework (and furthermore aesthetics) where they are absolutely necessary for a proper understanding of the subject rather than in a framework where we can escape them very well without too suffer. The reader will find here some developments and theorems proposed elsewhere in this textbook and this only in the order to avoid having too "turn pages" too much.

Natural fractals are named "non-deterministic objects" because the dynamic process that allows their creation itself varies with time randomly (see the section of Population Dynamics for an excellent example). Nevertheless, we can try to model dynamic systems that lead to fractal objects under a rigorous mathematical form (that is still a good example of the way in which the mathematicians manage to make a simple and intuitive concept into a concrete abstract mathematical model and somewhat confusing as for Knot Theory).

In this section, we will consider the study of two families of fractals that will be in order:

1. Deterministic fractals based on iterated function that are strictly self-similar. They are generated, as we shall see, by the recursive application of contracting functions on subsets of a metric space. The fixed point theorem will guarantee (as we shall also see!) the existence and uniqueness of a "fixed subset" of the metric space, towards which every subset converge.

2. Escape-time fractals (also known as "fractals by induction") that are not strictly self-similar: They are generated as we shall see later by recurring non-divergent sequences. The fixed point theorem serving as guarantee for the non-divergence of the function with respect to the chosen starting points.

14.2.1 IFS Fractals

Let us start by looking at the first family of fractals discovered by Michael Barnsley in 11987 (holocene calendar): the "deterministic iterated function systems IFS".

Of all fractals, figures only those built by iterated function systems usually shows the self-similarity property, meaning that their complexity is invariant under change of scale.

Let us start by "bounding" the thing...

We take an initial geometry E_0 of space E, a function f from E to E such that:

$$f(E_0)\subset E_0 \tag{14.2.1}$$

(which requires that the initial object can not leave its own definition domain through the iteration of the function f) and we create the discrete dynamical system defined by:

$$S_D=\begin{cases} E_0 \\ E_{n+1}=f(E_n) \end{cases} \tag{14.2.2}$$

Under certain conditions we will now see, the sequence of geometric objects (E_0) "tends" to a limit, which is often a fractal object (we will see further below some famous examples).

Naturally, there is a rigorous mathematical framework in which the mentioned conditions and the verb "tends" have a precise definition. In particular, the objects E_n are all compacts of E, that is to say bounded subsets (that we can include in a segment if E is a straight line, in a disk if E is a plane or a ball if E is the three-dimensional space) and closed (every convergent sequence of E_n has its limit in E). We place ourselves then in the compact metric space, equipped with the Hausdorff distance (see below for definition) for which we will show that it is complete when we work with compact sets the plane and of space, and we will check that f is an "Hutchinson operator", i.e. a contraction application from the space of the compact in itself for that distance. It then will then just remain to apply the fixed point theorem.

Dynamic systems of this type are said to be "deterministic", and therefore named "IFS" (iterated function systems deterministic). Let us precise that the limit of the IFS is named the "attractor of the IFS". We can show that under the conditions mentioned above, this attractor does not depend on the shape of the original geometric object (we will see practical examples further below).

Initially, we will limit our study to $\mathbb{R}$ (the general case is given in section Topology) knowing anyway that a generalization to the two-dimensional Euclidean space does not require too big and intellectual work and that the whole complex is isomorphic to it.

Definition (#14.2.576): To enable us to define the boundaries of our fractal functions let us consider $X\subseteq\mathbb{R}$. We say that ξ is the "supremum" of X and denote it by:

$$\xi=\sup(X) \tag{14.2.3}$$

if ξ is the smallest "upper bound" of X (an upper bound of X is a number a that satisfies $\forall x\in X, x\leq a$).

Similarly, we say that ξ is an "infimum" X and denote it by:

$$\xi=\inf(X) \tag{14.2.4}$$

if ξ is the biggest "lower bound" of X (a lower bound of X is a number a that satisfies $\forall x \in X, a \leq x$).

Remark

We often use the following characterization of the supremum:

$$\xi=\sup(X) \tag{14.2.5}$$

if and only if:

$$\forall \varepsilon>0, \exists x \in X: \quad \xi-\varepsilon \leq x \leq \xi \tag{14.2.6}$$

which is almost obvious because we can approach as close as we want of ξ with elements of X (think with small ε). For information, we then also have in the same idea:

$$\xi=\inf(X) \tag{14.2.7}$$

if and only if:

$$\forall \varepsilon>0, \exists x \in X: \quad \xi \leq x \leq \xi+\varepsilon \tag{14.2.8}$$

We consider as intuitive that if $X \subseteq \mathbb{R}$ has an upper bound, that is to say if there exists $a \in \mathbb{R}$ as $\forall x \in X, x \leq a$ (respectively lower bounded), then X has a supremum (respectively infimum).

We will see later that it is this property that will give us the possibility to prove later $\mathbb{R}$ a "complete metric space"!

Remark

By the way, let us underline the importance of taking $\mathbb{R}$ as a definition for metric space for this property to be satisfied. We can in fact notice that it is not verified in the set $\mathbb{Q}$ of rational numbers with the following simple example:

$$X=\left\{X \in \mathbb{Q} \mid x^{2} \leq 2\right\} \tag{14.2.9}$$

which is majorated but has no supremum in $\mathbb{Q}$ because this supremum is in $\mathbb{R}$ as:

$$x \leq \sqrt{2} \tag{14.2.10}$$

Therefore:

$$\sup(X)=\sqrt{2} \tag{14.2.11}$$

This is what makes $\mathbb{Q}$ is not "complete".

Definition (#14.2.577): We say that $X \subseteq \mathbb{R}$ is "bounded" if X is minorated and majorated.

From the definition it follows immediately that X is bounded if and only if there exists a, b with such that $X \subseteq [a, b]$.

Now that the concept of bound is relatively well defined, let see how a sequence can behave near from it.

Definition (#14.2.578): We say that $(a_n)_{\mathbb{N}}$ of $\mathbb{R}$ is an "increasing sequence" ("decreasing" respectively) if:

$$\forall n \in \mathbb{N}: \quad a_n \leq a_{n+1} \tag{14.2.12}$$

respectively:

$$\forall n \in \mathbb{N}: \quad a_n \geq a_{n+1} \tag{14.2.13}$$

We say the sequence $(a_n)_{\mathbb{N}}$ is "monotone" if it is increasing or decreasing as we have already seen in the section Sequences and Series.

Definition (#14.2.579): Given $T = \{n_0, n_1, n_2, \ldots\}$ infinite subset of $\mathbb{N}$ with $n_0 < n_1 < n_2 < \ldots$. We say the sequence $(a_{n_i})_{i \in \mathbb{N}}$ is a "subsequence" of the sequence $(a_n)_{\mathbb{N}}$.

Theorem 14.9. Let us now prove that every sequence in $\mathbb{R}$ admits a monotone subsequence (it's a bit the idea of a fractal!)

Proof 14.9.1. We say that a_m is a "peak" of the sequence if:

$$\forall n \geq m: \quad a_m \geq a_n \tag{14.2.14}$$

Consider the set P of peaks of the sequence $(a_n)_{\mathbb{N}}$.

- If P is infinite then the subsequence $(a_n)_{n \in P}$ is monotone since decreasing.
- If P is finite or empty:

$$m_1 = \max(P) + 1 \tag{14.2.15}$$

(if $P = \varnothing$ we choose any m_1). a_{m_1} is therefore not by construction not a peak, so there exists $m_2 \geq m_1$ such as $a_{m_1} \leq a_{m_2}$. In turn a_{m_2} is not a peak, so there exist $m_3 \geq m_2$ such as $a_{m_2} \leq a_{m_3}$ etc. We see that we define thus as an increasing subsequence.

Definition (#14.2.580): We say that the sequence $(a_n)_{\mathbb{N}}$ "converge" into $a \in \mathbb{R}$ and we note this:

$$\lim_{n \to +\infty} a_n = a \tag{14.2.16}$$

if:

$$\boxed{\forall \varepsilon > 0, \exists n \in \mathbb{N}: \quad n \geq N \Rightarrow |a_n - a| < \varepsilon} \tag{14.2.17}$$

In this case we say that a is the "limit of the sequence" $(a_n)_{\mathbb{N}}$.

■ Q.E.D.

Example:

In the example of the figure below where the sequence seems to converge towards the value 1.13, we observe that for a particular non-zero positive ε, there exists a particular n which we will denote N (which value is 17 in the below example) from which the sequence converges.

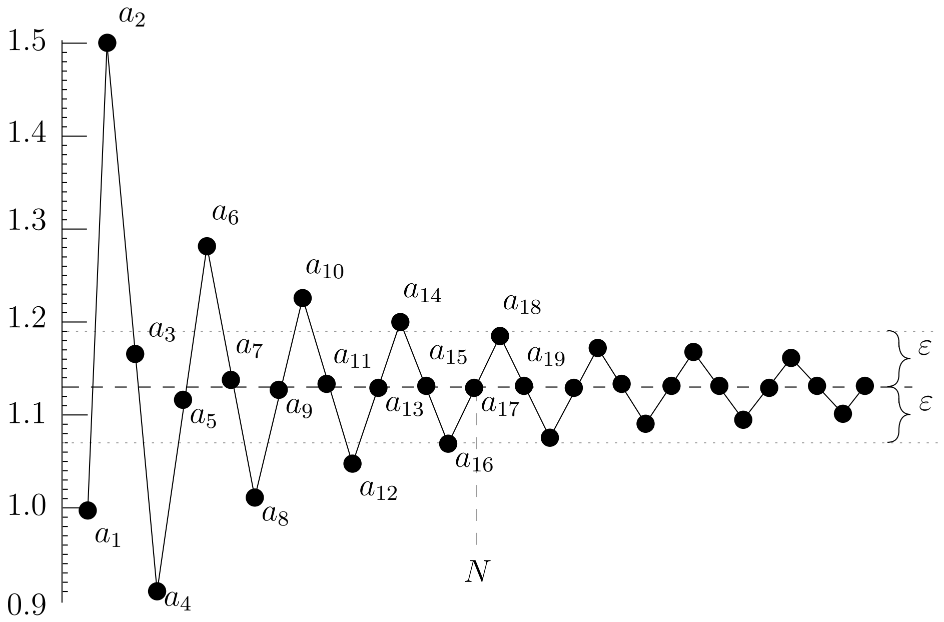


Figure 14.239: **Illustration of the principle of convergence of a sequence**

If there is no a (respectively N) for which the previous relation is true, then we say that the sequence "diverge".

Theorem 14.10. Let us prove now that every increasing sequence $(a_n)_{\mathbb{N}}$ (respectively decreasing) and majorated (resp. minorated) converges.

In other words, we seek to prove that any monotonous and bounded sequence $(a_n)_{\mathbb{N}}$ converges (obviously ... by construction).

Remark

If it does not converge, we could not easily find out what is its lower bound and upper bound ... hence the fact that the need of this theorem becomes trivial.

Proof 14.10.1. This theorem is actually quite intuitive. Consider for this an increasing sequence. We suspect that:

$$a = \sup\{a_0, a_1, a_2, \ldots\} \tag{14.2.18}$$

is the limit of this sequence. Note first of all that $a = \sup\{a_0, a_1, a_2, \ldots\}$ exists because $(a_n)_{\mathbb{N}}$ is majorated (see theorem proved previously).

Given $\varepsilon > 0$. It exists an a_N such tat $a - \varepsilon \leq a_N \leq a$. But in this case as the sequence is increasing, we have $\forall \geq N, a - \varepsilon \leq a_n \leq a$. That is to say $|a_n - a| \leq \varepsilon$. In the case where the sequence is decreasing by proceeding in the same way we prove that $a = \inf a_0, a_1, a_2, \ldots$ is the limit of this sequence.

■ Q.E.D.

Theorem 14.11. And now the important theorem to remember after all this: Every bounded sequence of real numbers has a convergent subsequence (that is intuitive ... but again ... when formalized it becomes sometimes less intuitive...).

This is what mathematicians name the "Bolzano-Weierstrass theorem" and it is extremely important in many areas of mathematics:

Proof 14.11.1. Given $(a_n)_{\mathbb{N}}$ such a sequence. By a previous proposal we know there is a monotonic subsequence which we denote $(b_n)_{\mathbb{N}}$. $(b_n)_{\mathbb{N}}$ is therefore a monotone and bounded sequence and by the previous theorem, $(b_n)_{\mathbb{N}}$ converges.

So if we are unable to determine whether the subsequence converges nor his exact limit (which in practice is often very difficult), we only need to know that the subsequence is monotone and bounded to ensure that it converges (which is most of time much easier).

■ Q.E.D.

Theorem 14.12. Remember that we saw in the section of Sequences and Series that a Cauchy sequence is a sequence $(a_n)_{\mathbb{N}}$ that verifies (we restrict ourselves to the special case of Euclidean distance):

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}: \quad n,m \geq N \Rightarrow |a_n - a_m| \leq \varepsilon \tag{14.2.19}$$

The difference between the two terms of a Cauchy sequence can be made arbitrarily small provided that the indices of these terms are big enough.

We have also proved (again in the section of Sequences and Series) that in the case of a distance in the general topological sense any convergent sequence is a Cauchy sequence (by cons the reciprocal is not always true at the condition that do not complete the set... otherwise the reciprocal is always true). For example, a sequence of rational numbers that converges to a real number is not a Cauchy sequence, except if we complete the set of rationals to get the set of real numbers.

Let us redo the proof restricted to Euclidean distance (the method is exactly the same as the reader will notice):

Proof 14.12.1. Given ε, we must show that there is:

$$N \in \mathbb{N} \quad \text{such that} \quad n,m \geq N \Rightarrow |a_n - a_m| \leq \varepsilon \tag{14.2.20}$$

But $(a_n)_{\mathbb{N}}$ tends to a therefore it exists a $N \in \mathbb{N}$ such that $n \geq N \Rightarrow |a_n - a| \leq \varepsilon/2$. For $n,m \geq N$ we therefore have:

$$|a_n - a_m| \leq |a_n - a| + |a - a_m| \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \tag{14.2.21}$$

■ Q.E.D.

Theorem 14.13. Let us now prove that every Cauchy sequence is bounded (we never talked about this until now anywhere in this textbook therefore we need to do the proof). Since currently we have just proved that every convergent sequence is a Cauchy sequence...

Proof 14.13.1. If $(a_n)_\mathbb{N}$ is a Cauchy sequence then particularly for $\varepsilon=1$ (randomly) we know that there exist $N\in\mathbb{N}$ such that $n,m\geq N\Rightarrow|a_n-a_m|\leq 1$. So if we fix m, we get:

$$\forall n\geq N:\quad |a_n|\leq 1+|a_m| \tag{14.2.22}$$

■ Q.E.D.

Let us see now the fundamental theorem (it is at this level that there is a huge impact on the understanding of what is actually a fractal!) that can be deduced from the previous lines.

Theorem 14.14. We will show that every Cauchy sequence of real numbers is convergent (by construction ...). We say then that the metric space $\mathbb{R}$ provided with the Euclidean distance (absolute value) is a "complete space".

Remarks

R1. The completeness property is related to the metric (hence this theorem could equally have its place in the section of Topology!): The same space can be complete for a given distance and incomplete for another one. It is therefore important to always specify the distance that we take when we speak of complete space.

R2. Intuitively, a space is complete if it has no holes. The set of rational numbers $\mathbb{Q}$ is by example complete if the real numbers are added to it.

Theorem 14.15. Consider first $(a_n)_\mathbb{N}$ a Cauchy sequence. We have seen just before that $(a_n)_\mathbb{N}$ is bounded and then that by the Bolzano-Weierstrass theorem, there exists a convergent subsequence $(a_{n_i})_{i\in\mathbb{N}}$. Let us denote by a the limit of the subsequence $(a_{n_i})_{i\in\mathbb{N}}$. We will now prove that the sequence $(a_n)_\mathbb{N}$ is convergent of limit a.

Proof 14.15.1. Given $\varepsilon>0$, there exist a $N\in\mathbb{N}$ such as (application of the definition of convergence for a subsequence):

$$i\geq N\Rightarrow|a_{n_i}-a|\leq\frac{\varepsilon}{2} \tag{14.2.23}$$

For this same ε it exists $M\geq 0$ (application of the definition of convergence for a Cauchy sequence):

$$M\geq 0,p,q\geq M\Rightarrow|a_p-a_q|\leq\frac{\varepsilon}{2} \tag{14.2.24}$$

Given $C=N+M$. We choose $i\geq C$. We then have $|a_{n_i}-a|\leq\varepsilon/2$ and for $n,i>C$:

$$|a_n-a_{n_i}|\leq\frac{\varepsilon}{2} \tag{14.2.25}$$

So by the triangle inequality (see section Vector Calculus page 1828), for any $n\geq C$:

$$|a_n-a|\leq|a_n-a_{n_i}|+|a_{n_i}-a|\leq\frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\varepsilon \tag{14.2.26}$$

This means precisely that $(a_n)_\mathbb{N}$ converges to a.

■ Q.E.D.

Basically it is an intuitive result but at the time when real numbers were not known or not rigorously defined it was a different story! In fact, it is simply enough complete any set by the real numbers to get a complete space. Moreover, some mathematicians define the set of real number saying that it is the set for which every Cauchy sequence converges...

Definition (naive #581): An "accumulation point" or "cluster point" is a point which we can approach as much as we want thanks to elements of a given set X (we will approach it for example with a sequence). However, this accumulation point can be both inside and outside of X (all items within X are obviously limit points). A good image is to see a series that approach this accumulation point and define circles around it that are becoming smaller and smaller containing elements of the sequence.

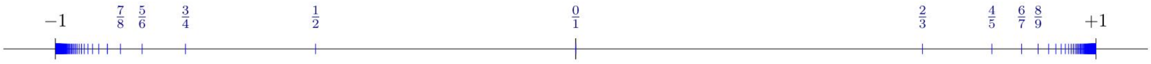
We can imagine as an example a sequence defined by the set X of rational numbers $\mathbb{Q}$ which tends to an irrational or to a transcendental number (these two points being elements not belonging to the set of rational $\mathbb{Q}$). So in this case, the accumulation point is outside X (the set of rational $\mathbb{Q}$). By cons, any accumulation point that would be a rational number for a sequence of rational number will necessarily ... in X (that is to say $\mathbb{Q}$).

✍**Example:**

With respect to the usual Euclidean topology, the sequence of rational numbers:

$$x_n = (-1)^n \frac{n}{n+1} \tag{14.2.27}$$

has no limit (i.e. does not converge) when $n \rightarrow \pm\infty$, but has two accumulation points (which are considered limit points here), that are -1 and $+1$:



Therefore, comes the following definition (for more details see the section Topology page 2446):

Definition (formal #582): Given $X \subseteq \mathbb{R}^n$. We say that $x \in \mathbb{R}^n$ is an "accumulation point" to X if for all ball $\mathcal{B}(x,r)$ of radius r center on x we have (for more details see the section Topology):

$$X \cap B(x,r) = \varnothing \tag{14.2.28}$$

The set of all accumulation points (of a sequence) to X is the "limit set" of X and denoted by $\bar{X}$. We have obviously (it suffice to conceptualize it in an abstract way for all possible ball) $X \subseteq \bar{X}$.

✍**Example:**

Let us consider the interval $]0,1]$ with the ball $B(0,1)$. The intersection between the ball and the interval is not zero, we can say that 0 is an accumulation point! But now let us take a sequence $1/n$ for example in the interval $]0,1]$. This sequence tends to zero but 0 is not in interval. It is a good example of $X \subseteq \bar{X}$.

We can therefore make the proposition:

Theorem 14.16. Let us prove now that $x \in \mathbb{R}^n$ is adherent to X if and only if there is a sequence $(u_n)_{\mathbb{N}}$ in X converging to x (note that the previous example shows that x is not necessarily in X).

In fact, we will instead prove (if we can say it is a proof..) that if we choose an accumulation point x then we can always find a sequence inX converging to x.

Proof 14.16.1. If $x \in \mathbb{R}^n$ is adherent to X then let us consider the following concentric balls $B\left(x,\dfrac{1}{n}\right)$ with $n \geq 1$ such as:

$$X \cap B\left(x,\frac{1}{n}\right) = \varnothing \tag{14.2.29}$$

and then there are always elements u_n that satisfy:

$$u_n \in \left(X \cap B\left(x,\frac{1}{n}\right)\right) \tag{14.2.30}$$

with whom we can create a sequence by the infinity of existing sequences.

■ Q.E.D.

Definition (#14.2.583): We say that $X \subseteq \mathbb{R}^n$ is a "closed space" if $X = \bar{X}$.

From the previous proposals it follows that in any closed space F, a sequence $(x_n)_{\mathbb{N}}$ that converges has its limit in F.

We consider as if trivial that if $(F_i)_I$ is a family of closed indexed spaces on any set I, then $\bigcap_I F_i$ is closed.

Definition (#14.2.584): $X \subseteq \mathbb{R}^n$ is a "compact space" if X is closed and bounded.

The following theorem gives a characterization of the compact from the sequences:

Theorem 14.17. $X \subseteq \mathbb{R}^n$ is compact if and only if any sequence $(x_n)_{\mathbb{N}}$ of X possess a subsequence that converges in X.

Proof 14.17.1. First let us prove that X is closed:

If X is compact and $(x_n)_{\mathbb{N}}$ is a sequence of X then by the Bolzano-Weierstrass theorem, $(x_n)_{\mathbb{N}}$ has a convergent subsequence of limit $x \in \mathbb{R}^n$. But since X is closed, we have $x \in X$. Conversely, let us assume that any sequence $(x_n)_{\mathbb{N}}$ of X has a subsequence which converges in X.

Then X is closed because if $x \in \bar{X}$ there is a sequence $(x_n)_{\mathbb{N}}$ of X which tends to x. By assumption, $(x_n)_{\mathbb{N}}$ has a subsequence that converges $y \in X$. The sequence $(x_n)_{\mathbb{N}}$ being convergent all the sub-sequences converge to the same value, therefore $x = y \in X$ (is this not great?!!). Thus $X = \bar{X}$ that is to say X is closed!

Let us now prove that X is bounded:

Let us suppose the opposite. So there exists a sequence $(x_n)_{\mathbb{N}}$ of X such that $||x_n|| \geq n$. But in this case, no subsequence of $(x_n)_{\mathbb{N}}$ is convergent, which is a contradiction! So X is bounded. Finally, X is compact!!!

■ Q.E.D.

A property of compacts is that if we consider $(A_n)_{\mathbb{N}}$ a decreasing sequence of non-empty compacts, that is to say $A_{n+1} \subseteq A_n$, then $\bigcap_{\in \mathbb{N}} A_n$ is a non-empty compact. We will pass trough the proof that is relatively trivial by the definition of the concept of Adherence Set that oblige that compacts are by construction non-empty...!

✍**Example:**

We build the Cantor set C as follows:

We begin by considering the closed bounded interval $C_0 = [0, 1]$ of $\mathbb{R}$ which is therefore a compact space (bounded and closed set). We split C_0 into three equal parts and we remove the middle interval. This gives us all the set:

$$C_1 = [0, 1/3] \cup [2/3, 1] \tag{14.2.31}$$

which can be also considered also as the application of a contracting scaling factor of $1/3$ on the closed bounded interval of departure of which we translate the center of homothety.

We start again with the two intervals $[0, 1/3]$, $[2/3, 1]$ for:

$$C_2 = [0, 1/9] \cup [2/9, 1/3] \cup [2/3, 7/9] \cup [8/9, 1] \tag{14.2.32}$$

disjoint union of 4 intervals. And so on... So we get a decreasing sequence C_n of compact. We define:

$$C = \bigcap_{n \geq 20} C_n \tag{14.2.33}$$

Thanks to the previous proposal, we know that C is not empty and is compact which shows that the compacts are not all "trivial" as intervals. The Cantor set (because he had played by doing the drawing below starting from the bottom) is an example of (compact) fractal:

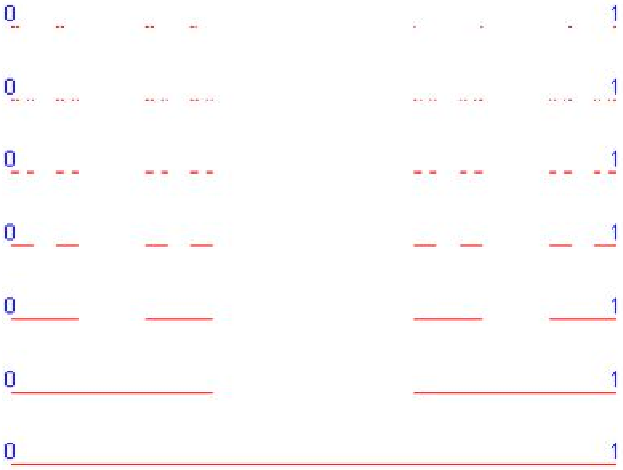


Figure 14.240: **Cantor set with Maple™ 4.00b**

that is possible to get with the following Maple™ 4.00b code:

```
 >with(plots):
line := proc(a::  list, b::  list)
local plotoptionen, n;
if nargs > 2 then
plotoptionen := seq(args[n], n=3 ..  nargs)
else
plotoptionen := NULL
fi;
plot([a, b], style=line, plotoptionen);
end:

cree_segment := (a,b,h) -> line([a,h],[b,h],color=black):
f1:=x->x/3:  f2:=x->(x+2)/3:

f := s -> s union map(f1, s) union map(f2, s):

sequence_de_segments := proc(l,h)
local accu, i;
accu := NULL;
for i to nops(l) by 2 do
accu := accu,cree_segment(l[i], l[i+1], h) od;
accu
end:
 >Cantor:= proc(n) local s, i;
>option remember;
>s:=sequence_de_segments([0,1], 1);
>for i from 1 to n do
>s:=sequence_de_segments(sort([op((f@@i)(0,1))]), (1-i/n)), >s;
>od;
```

```
>display(sunionseq(textplot([[0,(i+1/2)/n, '0'], [1, (i+1/2)/n,
'1']]
), i=0 ..  n), color=blue,axes=NONE,thickness=7)
>end:

>Cantor(7);
```

It is very interesting to notice that we converges to the Cantor fractal (in terms of geometry but also in term of values!) whatever the chosen starting compact we start from (the closed bounded interval) and also ... whatever the chosen contractor factor!

Benoît Mandelbrot also observed this type of self similar structure in the analysis of electrical signals transmitted at his (junior) time when working for IBM on copper cables (IBM had transmission information loss problems).

Now let us look for finish how behave compactis vis-à-vis continuous applications (we need this to prove how to determine the distance from a point to a set which we will need absolutely after to determine the properties of the Hausdorff distance).

Let us recall (see section Topology page 2460) that an application $f: X \mapsto \mathbb{R}^m$ whatever $X \subseteq \mathbb{R}^n$ is delta-epsilon continuous on a point $x \in X$ if:

$$\forall \varepsilon > 0, \exists \delta > 0, \forall y \in X: \quad d(x,y) \leq \delta \Rightarrow d(f(x), f(y)) \leq \varepsilon \tag{14.2.34}$$

This reflects the fact that for y close enough to x, $f(y)$ is arbitrarily close to $f(x)$. We also say that f is continuous on X if it is continuous at each point of X.

Theorem 14.18. Given $f: X \mapsto \mathbb{R}^m$ a delta-epsilon continuous application on $x \in X$ and $(x_n)_{\mathbb{N}}$ a sequence of X with:

$$\lim_{n \to +\infty} x_n = x \in X \tag{14.2.35}$$

Then the sequence $f(x_n)$ converges and (this proposal is very important!):

$$\boxed{\lim_{n \to +\infty} f(x_n) = f(x)} \tag{14.2.36}$$

In other words, if we use as a set of starting values of a convergent sequence, then the function that take as input the values of this sequence will converge too!

Proof 14.18.1. Given $\varepsilon > 0$. f is delta-epsilon continuous on x, so there exists $\delta > 0$ such that:

$$d(x_n, x) \leq \delta \Rightarrow d(f(x_n), f(x)) \leq \varepsilon \tag{14.2.37}$$

The sequence $(x_n)_{\mathbb{N}}$ tends to x therefore it exists $N \in \mathbb{N}$ such that:

$$n \leq N \Rightarrow d(x, x_n) \leq \delta \tag{14.2.38}$$

If follows that $n \geq N$, we have:

$$d(f(x_n), f(x)) \leq \varepsilon \tag{14.2.39}$$

■ Q.E.D.

If we now consider a compact $X \subseteq \mathbb{R}^n$ and $f : X \mapsto \mathbb{R}^n$ a continuous application. The application $f(X)$ is compact. In particular $\sup(f)$ and $\inf(f)$ will be reached by definition and by construction of a compact (closed and bounded set) which is equal to its adherence.

Theorem 14.19. In other words, a continuous real-valued function on a compact always reaches its supremum or infimum.

Proof 14.19.1. We will do the proof in two steps:

- Let us prove that $f(X)$ is closed.

  Indeed, given $f(x_n)$ a sequence that tends to $y \in \overline{f(X)}$ (we take theadherence to hope to prove that it is equal to the set itself) then X being compact, then $(x_n)_\mathbb{N}$ has a convergent subsequence $(x_{n_i})_{i\in\mathbb{N}}$. Let us put:

  $$\lim_{i\rightarrow+\infty} x_{n_i} = x \in X \qquad (14.2.40)$$

  The function f is delta-epsilon continuous, therefore:

  $$\lim_{i\rightarrow+\infty} f(x_{n_i}) = f(x) \qquad (14.2.41)$$

  But as:

  $$y = \lim_{n\rightarrow+\infty} f(x_n) = \lim_{i\rightarrow+\infty} f(x_{n_i}) \qquad (14.2.42)$$

  we have:

  $$y = f(x) \in f(X) \qquad (14.2.43)$$

  This proves that:

  $$f(X) = \overline{f(X)} \qquad (14.2.44)$$

  and therefore that $f(X)$ is closed.

- Let us now prove that $f(X)$ is bounded.

  For this let us suppose the contrary. There is therefore a sequence $f(x_n)$ such that:

  $$||f(x_n)|| \geq n \qquad (14.2.45)$$

  for every natural integer n (precisely because it is assumed unbounded). Given $(x_{n_i})_{i\in\mathbb{N}}$ a convergent subsequence of $(x_n)_\mathbb{N}$ with:

  $$\lim_{i\rightarrow+\infty} x_{n_i} = x \in X \qquad (14.2.46)$$

  Then:

  $$\lim_{i\rightarrow+\infty} f(x_{n_i}) = f(x) \qquad (14.2.47)$$

and it follows:

$$\lim_{i \to +\infty} ||f(x_{n_i})|| = ||f(x)|| \tag{14.2.48}$$

but this is in contradiction with:

$$\lim_{i \to +\infty} ||f(x_{n_i})|| = \lim_{i \to +\infty} n_i = +\infty \tag{14.2.49}$$

Therefore $f(X)$ is bounded. Then $f(X)$ being closed is bounded and therefore is compact.

■ Q.E.D.

Now let apply this result (because this is what interests us in fractal spaces) to calculate the distance from a point to a set:

Theorem 14.20. Given $x \in \mathbb{R}^n$, the application $f : \mathbb{R}^n \mapsto \mathbb{R}$ defined by $f(y) = d(x, y)$ is continuous.

Proof 14.20.1. For $(y, z) \in \mathbb{R}^2$, the triangle inequality gives us:

$$d(x,y) \leq d(x,z) + d(z,y) \Rightarrow d(x,y) - d(x,z) \leq d(z,y) \tag{14.2.50}$$

By changing the roles of y, z we get:

$$d(x,z) - d(x,y) \leq d(z,y) \tag{14.2.51}$$

and therefore:

$$|d(x,z) - d(x,y)| \leq d(z,y) \tag{14.2.52}$$

Therefore for a given $\varepsilon > 0$, $d(z,y) \leq \varepsilon$ implies:

$$|d(x,z) - d(x,y)| \leq \varepsilon \tag{14.2.53}$$

That is to say:

$$|f(z) - f(y)| \leq \varepsilon \tag{14.2.54}$$

and f is therefore continuous on y.

■ Q.E.D.

Definition (#14.2.585): For $x \in \mathbb{R}^n$ and $A \subseteq \mathbb{R}^n$ we define the distance x to A as being the value:

$$d(x,A) = \inf\{d(x,a) | a \in A\} \tag{14.2.55}$$

Theorem 14.21. If $x \in A$ then $d(x,A) = 0$ (should me almost trivial). The reciprocal is not true. Indeed in the case $x = 0$ and $A =]0,1]$ we have indeed $d(x,A) = 0$ but $x \in A$. So we have the following important proposal:

$$d(x,A) = 0 \Leftrightarrow x \in \bar{A} \tag{14.2.56}$$

Proof 14.21.1. The fact that $d(x,A)=0$ implies the existence of a sequence $(a_n)_{\mathbb{N}}$ of elements of A such that:

$$\lim_{n\rightarrow+\infty} d(x,a_n)=0 \tag{14.2.57}$$

which means

$$\lim_{n+\infty} a_n=x \tag{14.2.58}$$

therefore $x\in\bar{A}$ (see previous developments).

Conversely, if $x\in\bar{A}$ then for every $\varepsilon>0$ there exists $a\in A$ such as $d(x,a)\leq\varepsilon$. But $d(x,A)\leq d(x,a)\leq\varepsilon$. Thus for any $\varepsilon>0$, $d(x,A)\leq\varepsilon$. That is to say:

$$d(x,A)=0 \tag{14.2.59}$$

■ Q.E.D.

In general the distance from x to A is not reached. That is to say that there is no $a\in A$ such that $d(x,A)=d(x,a)$. To check this, it is sufficient to consider the example $x=-1$ and $A=]0,1]$. We have in this example $d(x,A)=1$ but for any $a\in A$, $d(x,a)>1$. If A is compact, the situation is obviously different according to the following proposal (the most important for the Hausdorff distance in our point of view):

Theorem 14.22. If $A\subseteq\mathbb{R}^n$ is compact, there exist $a\in A$ such that $d(x,A)=d(x,a)$. Therefore:

$$d(x,A)=\min\{d(x,y)|y\in A\} \tag{14.2.60}$$

Proof 14.22.1. The application $f:A\mapsto\mathbb{R}$ defined by $f(a)=d(x,a)$ is continuous as previously proved. Therefore $f(A)$ is compact (see a previous proposal). Thus, f reaches its bounds, that is to say, there is $a\int A$ such that $f(a)=\inf(f(A))$. Therefore:

$$d(x,a)=f(a)=\inf(f(A))=d(x,A) \tag{14.2.61}$$

■ Q.E.D.

Remark

This previous proposition does not say that a is unique, in general in fact it is not!

14.2.1.1 Fractals Metric Space

Fractals are often perceived by people as pretty drawings on paper, but when we look in detail the geometry of fractal, we need a particular space to study them, much like the biologist who puts small worms on a wafer to observe the worms in detail to the microscope. We will do the same for our fractals by placing them in a place they like...

This place is likely to be a subspace of $\mathbb{R}^2$ or $\mathbb{R}^3$, since in the end it will produce drawings... And to illustrate this we often will place ourselves in $\mathbb{R}^2$ (with the Euclidean metric), and unless otherwise stated, we always consider the case where (X,d) is a complete metric space.

Let us collect different items in order to construct this space:

Definition (#14.2.586): We define $\mathcal{H}(X)$ as the space whose points are the compacts subsets of X other than X itself. From now, we will name "fractal" any element of $\mathcal{H}(X)$.

✍**Example:**

It is immediate that if $x,y\in\mathcal{H}$, then $x\cup y\in\mathcal{H}$ but $x\cap y$ is not necessarily in $\mathcal{H}$. It is sufficient to see the figure below with the two compact sets of $\mathbb{R}^2$ (closed and bounded therefore) below. There are therefore two points of $\mathcal{H}$.. Their union is still a compact, and therefore:

$$z=x\cup y\in\mathcal{H}(\mathbb{R}^2) \tag{14.2.62}$$

By cons, if the sets are disjoint (as here), $x\cap y=\varnothing$ and therefore x, y are not a point of $\mathcal{H}(\mathbb{R}^2)$ (see previous theory).

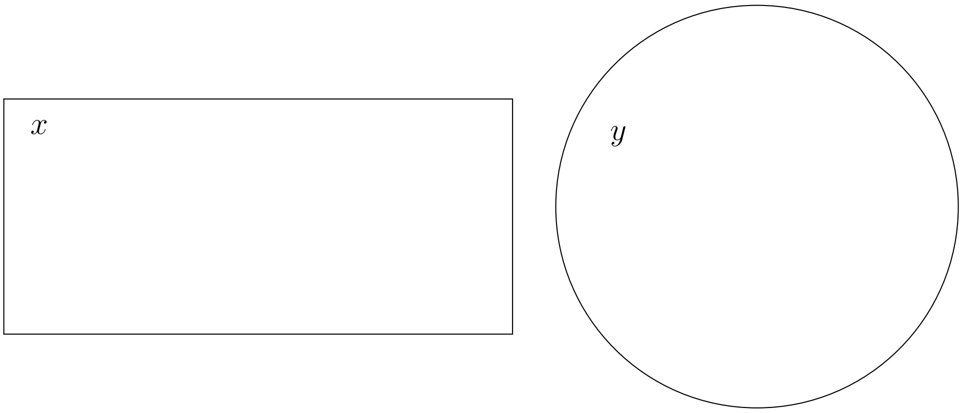


Figure 14.241: **Source: IFS and L-System V. Rezzonico, C. Hebeisen**

Another example involves taking the Cantor fractal...

Definition (#14.2.587): Given $x\in X$ and $B\in\mathcal{H}(X)$, we define the distance $d(x,B)$ of a point x to a set B by:

$$d(x,B)=\min\{d(x,y)|y\in B\} \tag{14.2.63}$$

Remarks

R1. This definition is quite general and applies to any non-empty subset of X, by replacing $\min$ by $\inf$. But in our specific case, we are interested in taking precisely $\mathcal{H}(X)$ as a subspace.

R2. This distance is well defined (is exists) as B is non-empty and compact.

R3. It is trivial to see that if this distance is zero, then $x\in\bar{B}$.

Example:

Illustration in the case where $X=\mathbb{R}^2$:

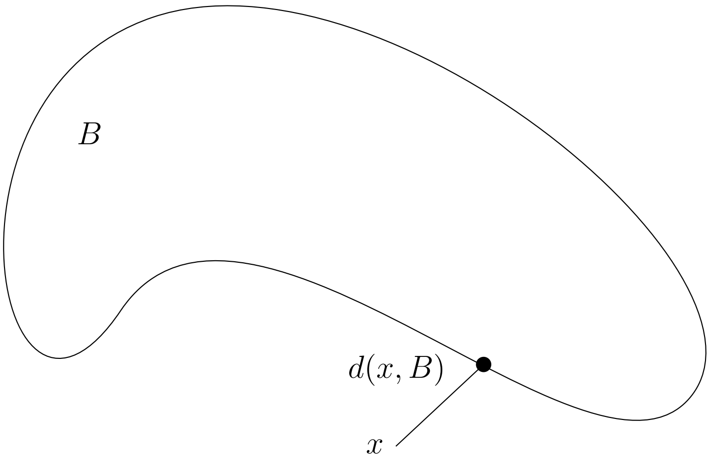


Figure 14.242: **Source: IFS and L-System V. Rezzonico, C. Hebeisen**

Definition (#14.2.588): Given $A,B\in\mathcal{H}(X)$. We define and denote the distance from A to B by:

$$d(A,B)=\max\{d(x,B)|x\in A\} \tag{14.2.64}$$

Remarks

R1. This definition is quite general and applies to any non-empty subset of X, by replacing $\min$ by $\inf$. But in our specific case, we are interested in taking precisely $\mathcal{H}(X)$ as a subspace.

R2. We notice that this distance does not provide a $\mathcal{H}(X)$ metric: indeed, $d(A,B)\neq d(B,A)$ in general (take for example the Cantor fractal where some compact we have $A\subset B$ with $A\neq B$, then we have $d(A,B)=0$ but $d(B,A)>0$).

Definition (#14.2.589): Given $A,B\in\mathcal{H}(X)$. We define and denote "Hausdorff distance" between two sets $A,B\in\mathcal{H}(X)$ by:

$$h(A,B)=\max\{d(A,B),d(B,A)\} \tag{14.2.65}$$

This time, by this definition, we have well a metric on $\mathcal{H}(X)$.

Indeed, let us check that the five properties of a distance are satisifed (see section Topology page 2450):

Given $A,B,C\in\mathcal{H}(X)$. Clearly we have without proof[81] (symmetry, nullity and separation on the diagonal):

---

[81] But let us know as always if you want we put the proofs

$$\begin{gathered}h(A,B)=h(B,A)\\ h(A,A)=d(A,A)=0\\ h(A,B)=0\Leftrightarrow A=B\end{gathered}\tag{14.2.66}$$

Moreover, since A and B are compact, $h(A,B)=d(A,B)$ (see on of the previous proposals) for a given $a\in A$ and a given $b\in B$. But, since $d(a,b)>0$ by definition, we have (property of positivity) finally $h(A,B)>0$ such that $a\in A,A\notin B$:

$$h(A,B)=\max\{d(A,B),d(B,A)\}\geq d(A,B)=\max\{d(a,B)|a\in A\}\geq d(a,B)>0\tag{14.2.67}$$

as B is closed.

Finally, since $h(A,B)=d(a,b)$ (see the extension of one of previous proposals), the triangle inequality is necessarily respected and then:

$$h(A,B)\leq h(A,B)+h(C,B)\tag{14.2.68}$$

So h is indeed a metric of $\mathcal{X}$, which makes $(\mathcal{H}(X),h)$ a metric space. This is a first step in the desired direction, we now have the tools to compare two sets belonging to $\mathcal{X}(X)$ by the Hausdorff distance between them. If the two are not "too different", so intuitively that distance should be fairly small.

If we choose a strictly contracting function $f:X\mapsto X$ of constant λ (see section Topology page 2459). Then, the application:

$$T_\lambda:\mathcal{H}(X)\mapsto\mathcal{H}(X)\tag{14.2.69}$$

defined by:

$$T_\lambda(A)=f(A)\tag{14.2.70}$$

is by onstruction also strictly contracting of constant λ.

Given $f_i:\mathbb{R}^2\mapsto\mathbb{R}^n$, $i\leq i\leq k$ strictly contracting applications of contraction constant $0\leq\lambda_i<1$. Then, there exists a unique compact $A\in\mathcal{H}(\mathbb{R}^n)$ such that:

$$A=T_{\lambda_1,\ldots,\lambda_k}(A)=f_1(A)\cup\ldots\cup f_k(A)\tag{14.2.71}$$

(A is the unique fixe point of $T_{\lambda_1,\ldots,\lambda_k}$) and for any compact B, we have:

$$\lim_{m\to+\infty}T_{\lambda_1,\ldots,\lambda_k}(B)=A\tag{14.2.72}$$

where $T^m_{\lambda_1,\ldots,\lambda_k}(B)$ is the m-th iteration of B by $T_{\lambda_1,\ldots,\lambda_k}$.

This result derives from the fixed-point theorem (see section Sequences and Series page 1781) applied to the space $\mathcal{H}(\mathbb{R}^n)$ that is complete.

With the same notation, we say that $f_1,\ldots,f_k$ is an IFS (Iterated Function Systems) coding of the compact A. Thus the functions $f_1,\ldots,f_k$ define the compact A. What is surprising, as we will see in the following examples, is that the $f_1,\ldots,f_k$ are usually quite simple (as homotheties of the plan) while the compact A is in many cases relatively or very "complicated" visually speaking.

If case $k=1$ is without interest, we would have $A=(x)$ where x is the fixed point of f_1. With $k=2$ we already obtain nontrivial results.

Remark

When the iterative contracting functions are all homotheties we speak then of "Sierpinski's fractal". Thus, the Cantor fractal belongs to the family of Sierpinski's fractals.

A frequently used method for to generate IFS fractals (as it will be the case below with Maple™ 4.00b[82]) is to consider a point in the plane (x_n, y_n) on which we can without conditions or constraints apply an affine transformation to get a new point such that:

$$\begin{aligned} x_{n+1} &= ax_n + by_n + e \\ y_{n+1} &= cx_n + dy_n + f \end{aligned} \tag{14.2.73}$$

where a, b, c, d, e and f are any constants, and (x_0, y_0) is given (chosen).

We can therefore consider an application T that describes our transformation, and in matrix form we can write the previous system as follows:

$$T\begin{pmatrix} x \\ y \end{pmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{pmatrix} e \\ f \end{pmatrix} \tag{14.2.74}$$

or even:

$$W(\bar{x}) = A\vec{x} + \vec{b} \tag{14.2.75}$$

So in general, the vector $\vec{b}$ above simply describes a translation, and the matrix A is the composition of a rotations and a scaling (see section Euclidean Geometry page 2640). Computer programs (as it will be the case in the examples below), thus often require that the knowledge of the six parameters a, b, c, d, e and f that can for majority be equal to zero.

---

[82] But if some readers have reproduced all the example below in C++ they are welcome to share their code

14.2.2 Fractals Visualization

14.2.2.1 Cantor's Fractal (Cantor Set)

Let us come back now on Cantor's Fractal but see now from the point of view of the application of two iterative contracting functions f_1, f_2 (thus corresponding to $k=2$).

So we start from the following closed bounded set (then a compact as it is bounded):

$$[[0,1]] \tag{14.2.76}$$

Therefore:

Figure 14.243: **Start set of Cantor's Fractal**

So now by definition we split in three equal parts and we remove the middle interval. This gives us the set:

$$[[0,1/3],[2/3,1]]=[0,1/3]\cup[2/3,1] \tag{14.2.77}$$

Therefore:

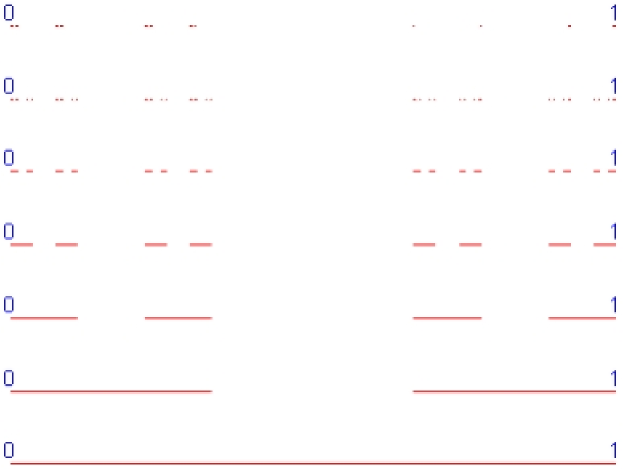


Figure 14.244: **First iteration of Cantor's Fractal**

We can see that $[0,1/3]$ can be obtained by the following homothety (scaling) application of factor $1/3$ centered at $(0,0)$:

$$f_1=h_{(x,y)}^{\lambda}=h_{(0,0)}^{1/3} \tag{14.2.78}$$

and that $[2/3,1]$ can be obtained by the following homothety (scaling) of factor $1/3$ centered at $(1,0)$:

$$f_2=h_{(x,y)}^{\lambda}=h_{(1,0)}^{1/3} \tag{14.2.79}$$

and so on, and we get as we already know (see Maple™ 4.00b code already given above):

Figure 14.245: Cantor attractor after 6 iterations

Which corresponds using the formalism seen earlier to:

$$\begin{aligned}&A_0=[0,1]\\&A_1 f_1([0,1])\cup f_2([0,1])=[[0,1/3],[2/3,1]]\\&A_2=f_1(A_1)\cup f_2(A_1)=\ldots\\&\quad\ldots\end{aligned} \tag{14.2.80}$$

This process is continued ad infinitum, where the n-th set is

$$C_n=\frac{C_{n-1}}{3}\cup\left(\frac{2}{3}+\frac{C_{n-1}}{3}\right) \tag{14.2.81}$$

with obviously $C_0=[0,1]$.

But let us see that it works with any compact of $\mathbb{R}$ as a square for example with the following Maple™ 4.00b code (we always show all the details of Maple™ 4.00b, because nothing says that readers have the software or that the software will still exist in 50 years for reproductibility purposes...).

Remark

It seems that there may be sometimes issues by copying and pasting the below code into Maple™ 4.00b. If it's the case one your computer, just rewrite it from scratch directly in the Maple™ console.

```
>transforme_point := proc(t, p)
[t[1]*p[1]+t[2]*p[2]+t[5], t[3]*p[1]+t[4]*p[2]+t[6]]
end:


>IFSS := proc(n, liste_de_transformations,col)
local i, j, k, s, seq_square:
seq_square :=[[0,0],[1,0],[1,1],[0,1]];
for j to n do
s := NULL;
for i to nops(liste_de_transformations) do
s := s,
seq(transform_square(liste_de_transformations[i],
op(k, [seq_square])),
k=1 .. nops([seq_square]))
od;
seq_square := s
od;
plots[polygonplot]([seq_square], axes=none, color=col,
scaling=constrained)
end:
```

```
>cantor:=[[evalf(1/3),0,0,evalf(1/3),0,0],[evalf(1/3),0,0,evalf(1/3),
evalf(2/3),0]]:
```

```
>IFSS(1, cantor,blue);
```

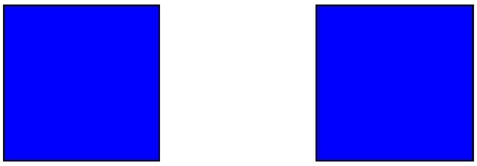


Figure 14.246: **First iteration on the Cantor set with squares**

```
>IFSS(2, cantor,blue);
```

Figure 14.247: **Second iteration on the Cantor set with squares**

```
>IFSS(3, cantor,blue);
```

Figure 14.248: **Third iteration on the Cantor set with squares**

etc.

So, whatever the starting set, the sequence of compact obtained by successive application of these two plane homotheties always converge (in the sense of the Hausdorff distance) to the same compact/attractor (assimilated to the fixed of the Fixed point theorem...) named Cantor's fractal or Cantor set (thus belonging to the family of Serpienski fractals). The latest figure above is a good approximation of this set.

The Cantor set being self-similar, consisting of $N=2$ congruent subsets, each when magnified by a factor of $M=3$ yields the original set. Hence the fractal dimension (see section Euclidean Geometry page 2555) of the Cantor set is:

$$d=\frac{\log(2^n)}{\log(3^n)}=\frac{n\log(2)}{n\log(3)}\cong 0.631 \tag{14.2.82}$$

14.2.2.2 Triangle Sierpinski Fractal

To build the Sierpinski fractal (which can be found as curiosity sometimes on the seashell Cymbiola innexa REEVE), based on three iterative contracting functions f_1, f_2, f_3 (thus corresponding to $k=3$), we assume for example three following points of $\mathbb{R}^2$:

$$[[0,0],[1,0],[0.5,1]] \tag{14.2.83}$$

Which gives with Maple™ 4.00b:

The Sierkpinski Triangle consists of 3^n subsets with magnification factor 2^n. So the fractal dimension is:

$$d=\frac{\log(3^n)}{\log(2^n)}=\frac{n\log(3)}{n\log(3)}\cong 1.585 \tag{14.2.84}$$

```
>plots[polygonplot]( [[0, 0], [1, 0], [0.5, 1]],axes=none,color=black,
scaling=constrained);
```

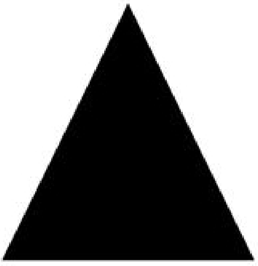


Figure 14.249: Start set of triangle Sierpinski fractal

This is a triangle, but we could start from any shape and we always arrive at the same result as we will see later.

We apply on each set a contracting function of factor 0.5, this gives the triangle:

$$[[0,0],[0.5,0],[0.25,0.5]] \tag{14.2.85}$$

and we denote that this homothety (scaling) of factor 0.5 and center $(0,0)$ on the original triangle by:

$$f_1=h_{(x,y)}^{\lambda}=h_{(0,0)}^{1/2} \tag{14.2.86}$$

We do now on this triangle a translation of 0.5 in the direction of the x-axis, which gives the triangle:

$$[[0.5,0],[1,0],[0.75,0.5]] \tag{14.2.87}$$

which corresponds to a scaling factor of 0.5 and center (1.0) on the original triangle:

$$f_2=h_{(x,y)}^{\lambda}=h_{(1,0)}^{1/2} \tag{14.2.88}$$

We now translate $[[0,0],[0.5,0],[0.25,0.5]]$ of 0.25 along the x-axis and of 0.5 according to the y-axis to have:

$$[[0.25,0.5],[0.75,0.5],[0.5,1]] \tag{14.2.89}$$

which corresponds to a scaling factor of 0.5 of center $(0.5,0.75)$ on the original triangle:

$$f_3=h_{(x,y)}^{\lambda}=h_{(0.5,0.75)}^{1/2} \tag{14.2.90}$$

With Maple™ 4.00b it now gives for the three triangles:

```
>plots[polygonplot]([[[0,0],[0.5,0],[0.25,0.5]],[[0.5,0],[1,0],[.75,0.5]],
[[0.25,0.5],[0.75,0.5],[0.5,1]]], axes=none,color=black,
scaling=constrained);
```

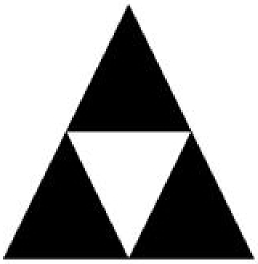


Figure 14.250: First iteration on the Sierpinksi triangle

and so on...:

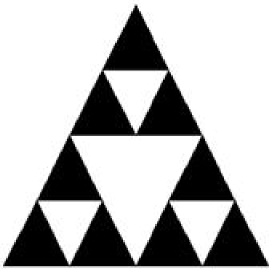


Figure 14.251: Second iteration on the Sierpinksi triangle

and so on...:

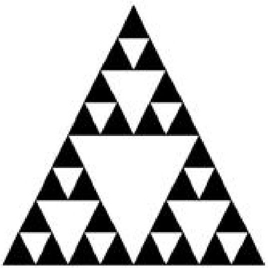


Figure 14.252: Third iteration on the Sierpinksi triangle

and so on...:

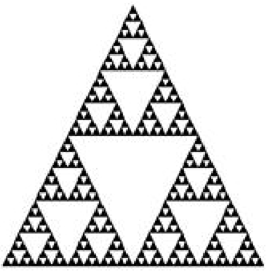


Figure 14.253: **Fourth iteration on the Sierpinksi triangle**

and so on (the triangles begins to be quite small to see a difference without zoom)...:

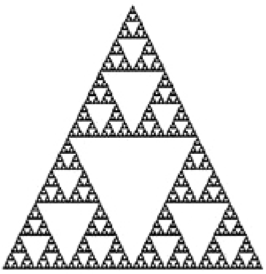


Figure 14.254: **Sixth iteration on the Sierpinksi triangle**

Which corresponds by taking the formalism seen previously:

$$\begin{aligned} A_0 &= [[0,0],[1,0],[0.5,1]] \\ A_1 &= f_1([0,0]) \cup f_2([1,0]) \cup f_3([0.5,1]) \end{aligned} \quad (14.2.91)$$

We can make the same remark as when we have study the Cantor's Fractal: whatever the starting set, the sequence of compact obtained by successive application of these three plane homotheties always converge (in the sense of the Hausdorff distance) to the same compact/attractor (assimilated to the fixed of the Fixed point theorem...) named Sierpinski fractal. The latest figure above is a good approximation of this set.

Let's see this with a Maple™ 4.00b code (once again if the copy/paste from the book in Maple™ 4.00b does not work, simply rewrite the code in the Maple™ console):

```
>transforme_triangle := proc(t, triangle)
local i;
[seq(transforme_point(t, triangle[i]), i=1 ..  3)] end:


>IFS := proc(n, liste_de_transformations,col)
local i, j, k, s, sequence_de_triangles:
options ‘Copyright by Alain Schauber, 1996‘;
sequence_de_triangles := [[0, 0], [1, 0], [0.5, 1]];
for j to n do
s := NULL;
for i to nops(liste_de_transformations) do
s:= s,
seq(transforme_triangle(liste_de_transformations[i],
op(k, [sequence_de_triangles])),
k=1 .. nops([sequence_de_triangles]))
od;
sequence_de_triangles := s
od;
plots[polygonplot]([sequence_de_triangles], axes=none, color=col,
scaling=constrained)
end:
```

```
>triangle_de_Sierpinski:=[[0.5,0,0,0.5,0,0],[0.5,0,0,0.5,0.5,0],
[0.5,0,0,0.5,0.25,0.5]]:
```

```
>IFS(6, triangle_de_Sierpinski,blue);
```

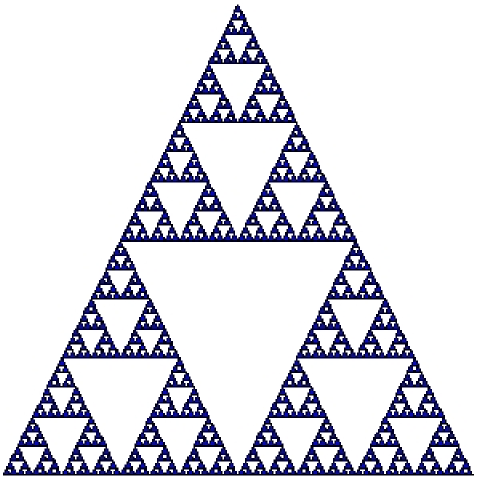


Figure 14.255: **Sierpinski triangle attractor**

And this time we don't start from a triangle but from a square (IFS Square) with the following Maple™ 4.00b code:

```
>transforme_square := proc(t, square)
local i;
[seq(transforme_point(t, square[i]), i=1 .. 4)] end:
```

```
>IFS := proc(n, liste_de_transformations,col)
local i, j, k, s, seq_square:
seq_square := [[0, 0], [1, 0], [1, 1],[0, 1]];
for j to n do
s := NULL;
for i to nops(liste_de_transformations) do
s:= s,
seq(transform_square(liste_de_transformations[i],
op(k, [seq_square])),
k=1 .. nops([seq_square]))
od;
```

```
seq_square := s
od;
plots[polygonplot]([seq_square], axes=none, color=col,
scaling=constrained)
end:
```

```
>square_Sierpinski:=[[0.5,0,0,0.5,0,0],[0.5,0,0,0.5,0.5,0],
[0.5,0,0,0.5,0.25,0.5]]:
```

```
>IFS(1, square_Sierpinski,green);
```

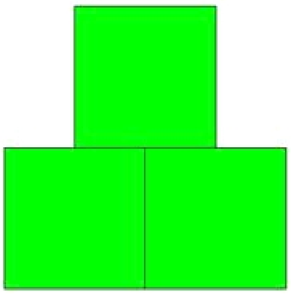


Figure 14.256: **First iteration on the Sierpinksi square**

```
>IFS(2, square_Sierpinski,green);
```

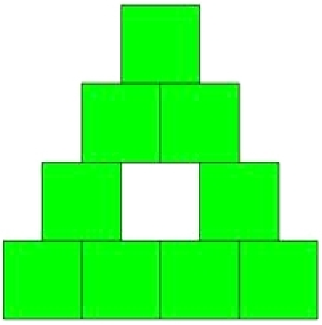


Figure 14.257: **Second iteration on the Sierpinksi square**

```
>IFS(3, square_Sierpinski,green);
```

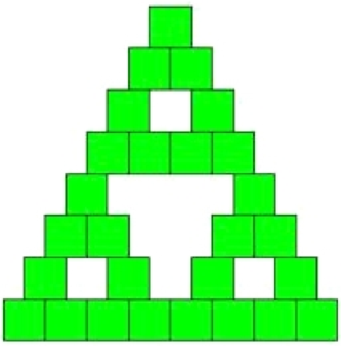


Figure 14.258: **Third iteration on the Sierpinksi square**

and so on until...:

```
>IFS(6, square_Sierpinski,green);
```

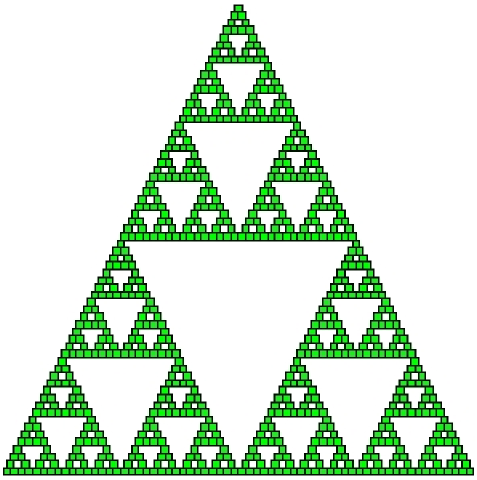


Figure 14.259: Sixth iteration on the Sierpinksi square

Basically, the Sierpinski fractal can obviously be also seen as a triangle to which the middle of the triangle is removed and where for each of the remaining triangles, we restart the process!

14.2.2.3 Sierpinski carpet fractal

The Sierpinski carpet is the attractor of eight contracting iterative functions of homothety of ratio $1/3$ centered at the vertices and sides of a square in which can be any put any plane geometric shape.

This time in $\mathbb{R}^2$ we consider the eight homotheties (h):

$$h_{(0,0)}^{1/3}, h_{(0,0.5)}^{1/3}, h_{(0,1)}^{1/3}, h_{(0.5,1)}^{1/3}, h_{(1,1)}^{1/3}, h_{(1,0.5)}^{1/3}, h_{(1,0)}^{1/3}, h_{(0.5,0)}^{1/3}, \tag{14.2.92}$$

and we start for example from the four following points:

$$[[0,0],[1,0],[1,1],[0,1]] \tag{14.2.93}$$

that corresponds to a filled square (but we can choose anything else!):

Figure 14.260: Start set of Sierpinski carpet fractal

After application of the eight homotheties functions (we leave to the reader the manual calculations in the same we way we have already do it for the triangle), we get the following form of eight squares:

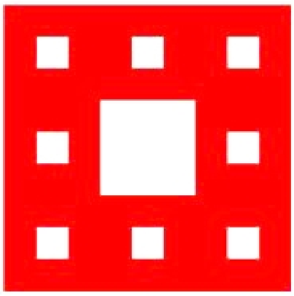


Figure 14.261: **Second iteration of Sierpinski carpet fractal**

and applying again the eight homotheties (fortunately there are computers...):

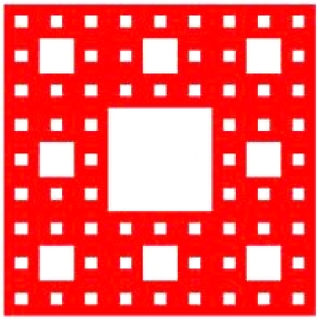


Figure 14.262: **Third iteration of Sierpinski carpet fractal**

and again:

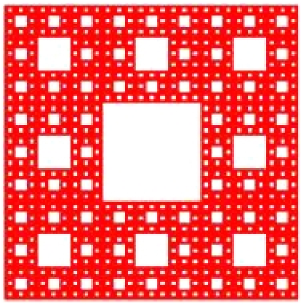


Figure 14.263: **Fourth iteration of Sierpinski carpet fractal**

etc.

The resulting fixed point (attractor) is obviously named the "Sierpinski's carpet" and this is the shape of the receiving antenna of the majority of our cell phones in the early 121st century (holocene calendar).

The above figures can be obtained successively with the following Maple™ 4.00b code (if the copy/paste form the book in Maple™ 4.00b does not work, simply rewrite the code in the Maple™ console):

```
>transforme_point := proc(t, p)
[t[1]*p[1]+t[2]*p[2]+t[5], t[3]*p[1]+t[4]*p[2]+t[6]]
end:
```

```
>transform_square := proc(t, square)
local i;
[seq(transforme_point(t, square[i]), i=1 ..  4)]
end:
```

```
>IFSS := proc(n, liste_de _transformations,col)
local i, j, k, s, seq_square:
seq_square :=[[0,0],[1,0],[1,1],[0,1]];
for j to n do
s := NULL;
for i to nops(liste_de_transformations) do
s := s,
seq(transform_square(liste_de _transformations[i],
op(k, [seq_square])),
k=1 ..  nops([seq_square]))
od;
seq_square := s
od;
plots[polygonplot]([seq_square], axes=none, color=col,
scaling=constrained)
end:
```

```
> dywan:= [[evalf(1/3),0,0,evalf(1/3),0,0],[evalf(1/3),0,0,evalf(1/3),
evalf(1/3),0],[evalf(1/3),0,0,evalf(1/3),evalf(2/3),0],
[evalf(1/3),0,0,evalf(1/3),0,evalf(2/3)], [evalf(1/3),0,0,evalf(1/3),evalf(1/3),
evalf(2/3)],[evalf(1/3),0,0,evalf(1/3), evalf(2/3),evalf(2/3)],
[evalf(1/3),0,0,evalf(1/3),0,evalf(1/3)], [evalf(1/3),0,0,evalf(1/3),evalf(2/3),eva
```

```
>IFSS(0, dywan, blue);
```

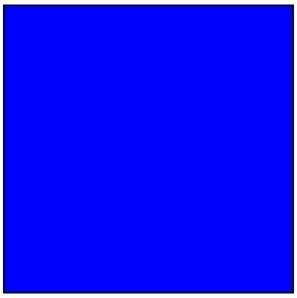


Figure 14.264: **Start set of Sierpinski carpet fractal**

```
>IFSS(1, dywan, blue);
```

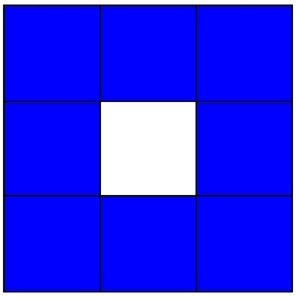


Figure 14.265: First iteration of Sierpinski carpet fractal

```
>IFSS(2, dywan, blue);
```

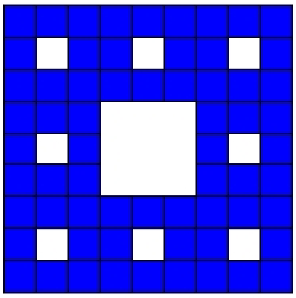


Figure 14.266: Second iteration of Sierpinski carpet fractal

```
>IFSS(3, dywan, blue);
```

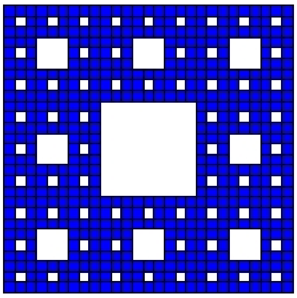


Figure 14.267: Third iteration of Sierpinski carpet fractal

```
>IFSS(4, dywan, blue);
```

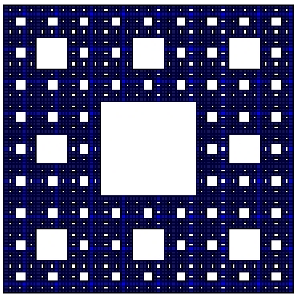


Figure 14.268: Fourth iteration of Sierpinski carpet fractal

14.2.2.4 Fractal spirals

We just saw two fractals of the Sierpinski's fractal family therefore based solely on contracting homotheties. Let us now see a fractal that combines rotation and scaling contraction.

In $\mathbb{R}^2$ we consider the two applications of homotheties (h) and rotations (R) as follows:

$$h_{(0.25,0.5)} \circ R^{5\pi/6}_{(0.25,0.5)} \qquad h_{(0.5,0.5)} \circ R^{\pi/6}_{(0.5,0.5)} \tag{14.2.94}$$

With a triangle and always with Maple™ 4.00b, this gives us (if the copy/paste form the book in Maple™ 4.00b does not work, simply rewrite the code in the Maple™ console):

```
>transforme_triangle := proc(t, triangle)
local i;
[seq(transforme_point(t, triangle[i]), i=1 ..  3)]
end:
```

```
>IFS := proc(n, liste_de_transformations,col)
local i, j, k, s, sequence_de_triangles:
options ‘Copyright by Alain Schauber, 1996‘;
sequence_de_triangles := [[0, 0], [1, 0], [0.5, 1]];
for j to n do
s := NULL;
for i to nops(liste_de_transformations) do
s := s,
seq(transforme_triangle(liste_de_transformations[i],
op(k, [sequence_de_triangles])),
k=1 ..  nops([sequence_de_triangles]))
od;
sequence_de_triangles := s
od;
plots[polygonplot]([sequence_de_triangles], axes=none, color=col,
scaling=constrained)
end:
```

```
>a:=evalf(5*Pi/6);b:=evalf(Pi/6);
>c1x:=0.25;c1y:=0.5;c2x:=0.5;c2y:=0.5;
>h1:=0.2;h2:=0.95;
>spirale:=[[h1*cos(a),-h1*sin(a),h1*sin(a),h1*cos(a),(1-h1*cos(a))*c1x
+h1*sin(a)*c1y,-h1*sin(a)*c1x+(1-h1*cos(a))*c1y],[h2*cos(b),-h2*sin(b),h2*sin(b),
h2*cos(b),(1-h2*cos(b))*c2x+h2*sin(b)*c2y,-h2*sin(b)*c2x+(1-h2*cos(b))*c2y]]:
```

```
>IFS(1,spirale,blue);
```

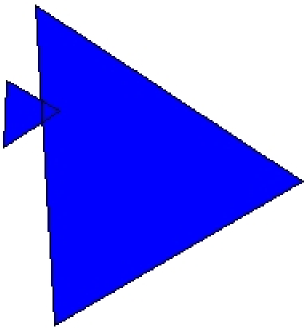


Figure 14.269: **First iteration of spiral fractal**

and as the convergence is very slow, we will give the results by step of 5 iterations.

```
>IFS(6,spirale,blue);
```

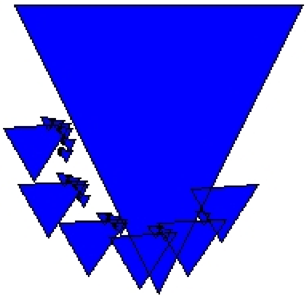


Figure 14.270: **Sixth iteration of spiral fractal**

```
>IFS(11,spirale,blue);
```

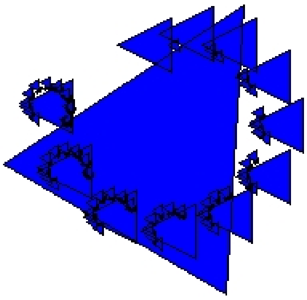


Figure 14.271: **Eleventh iteration of spiral fractal**

```
>IFS(16,spirale,blue);
```

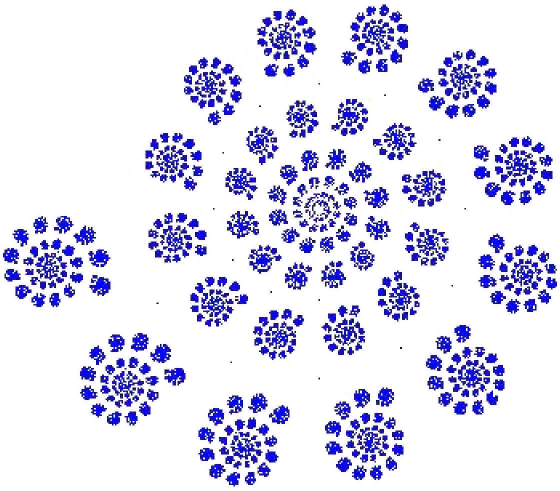


Figure 14.272: Sixteenth iteration of spiral fractal

14.2.2.5 Von Koch fractal (Koch snowflake)

Still, in fractal obtained by contracting homotethies (h) and rotation (R) but to which we add now a translation (T), the Von Koch curve is a well known fractal, it can be obtained by the following applications (we already met this fractal in the section of Euclidean Geometry when we have introduced the concept of fractal dimension):

$$h^{1/3}_{(0,0)} \qquad h^{1/3}_{(1,0)} \qquad T_{(1/3,0)}\circ h^{1/3}_{(0,0)}\circ R^{\pi/3}_{(0,0)} \qquad T_{(-1/3,0)}\circ h^{1/3}_{(1,0)}\circ R^{-\pi/3}_{(1,0)} \tag{14.2.95}$$

Here is the corresponding Maple™ 4.00b code (if the copy/paste form the book in Maple™ 4.00b does not work, simply rewrite the code in the Maple™ console):

```
>koch := proc(p::  numeric)
local m, n, k, l, s, h, x, y, pts, t, i;
h := 3^(-p);
pts := table([]):  # [0, 0];
pts[0]:=[0,0];
x := 0; y := 0;
for n from 0 to (4^p) do
m := n;
s := 0;
for l from 0 to p-1 do
t := irem(m, 4);
m := iquo(m, 4);
s := s+irem((t+1), 3) - 1
od; # end of for l
x := evalhf(x+cos(Pi*s/3)*h);
y := evalhf(y+sin(Pi*s/3)*h);
pts[n+1] := [x, y];
od;
[seq(pts[i], i=0 ..  n-1)];
end:
```

```
>plot(koch(0), scaling=constrained, style=LINE, axes=NONE,
color=blue,thickness=2);
```

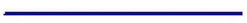


Figure 14.273: **Start set of Von Koch fractal**

```
>plot(koch(1), scaling=constrained, style=LINE, axes=NONE,
color=blue,thickness=2);
```

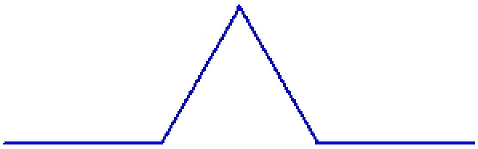


Figure 14.274: **First iteration of Von Koch fractal**

```
>plot(koch(2), scaling=constrained, style=LINE, axes=NONE,
color=blue,thickness=2);
```

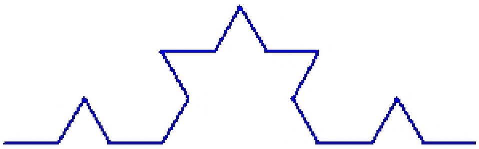


Figure 14.275: **Second iteration of Von Koch fractal**

```
>plot(koch(3), scaling=constrained, style=LINE, axes=NONE,
color=blue,thickness=2);
```

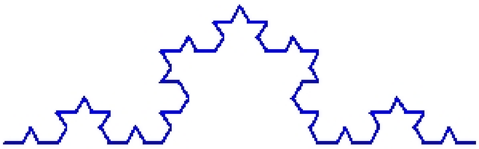


Figure 14.276: **Third iteration of Von Koch fractal**

etc. etc. Until the following attractor:

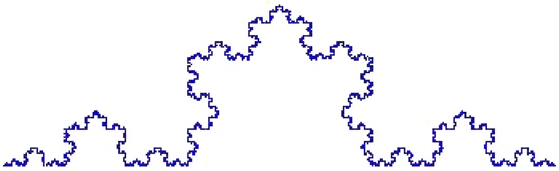


Figure 14.277: Fourth iteration of Von Koch fractal

The Koch snowflake (see figure below) is a variant of the Von Koch line and thab can be constructed by starting with an equilateral triangle, then recursively altering each line segment as follows:

1. Divide the line segment into three segments of equal length.
2. Draw an equilateral triangle that has the middle segment from step 1 as its base and points outward.
3. Remove the line segment that is the base of the triangle from step 2.

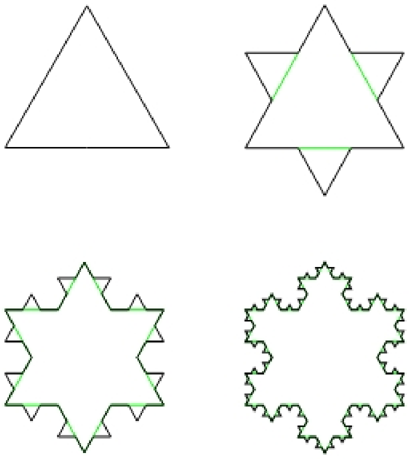


Figure 14.278: Von Koch snowflake (source: Wikipedia)

After each iteration, the number of sides of the Koch snowflake increases by a factor of 4, so the number of sides after n iterations is given by:

$$N_n = N_{n-1} \cdot 4 = 3 \cdot 4^n \tag{14.2.96}$$

If the original equilateral triangle has sides of length s, the length of each side of the snowflake after n iterations is:

$$S_n = \frac{S_{n-1}}{3} = \frac{s}{3^n} \tag{14.2.97}$$

the perimeter of the snowflake after n iterations is therefore of:

$$P_n = N_n \cdot S_n = 3 \cdot s \cdot \left(\frac{4}{3}\right)^n \tag{14.2.98}$$

The Koch curve has an infinite length because the total length of the curve increases by one third with each iteration. Thas is to say:

$$\lim_{n\to+\infty} P_n=\lim_{n\to+\infty} 3\cdot s\cdot\left(\frac{4}{3}\right)^n=+\infty \tag{14.2.99}$$

The funny thing is that the area is finite... when the perimeter is infinite...

Indeed, in each iteration a new triangle is added on each side of the previous iteration, so the number of new triangles added in iteration n is:

$$T_n=3\cdot 4^{n-1}=\frac{3}{4}\cdot 4^n \tag{14.2.100}$$

The area of each new equilateral triangle added in an iteration is one ninth of the area of each equilateral triangle added in the previous iteration, so the area of each equilateral triangle added in iteration n is:

$$a_n=\frac{a_{n-1}}{9}=\frac{a_0}{9^n} \tag{14.2.101}$$

where a_0 is the area of the original equilateral triangle. The total new area added in iteration n is therefore:

$$b_n=T_n\cdot a_n=\frac{3}{4}\cdot\left(\frac{4}{9}\right)^n\cdot a_0 \tag{14.2.102}$$

The total area of the snowflake after n iterations is then obviously:

$$A_n=a_0+\sum_{k=1}^{n} b_k=a_0\left(1+\frac{3}{4}\sum_{k=1}^{n}\left(\frac{4}{9}\right)^k\right)=a_0\left(1+\frac{1}{3}\sum_{k=0}^{n-1}\left(\frac{4}{9}\right)^k\right) \tag{14.2.103}$$

Collapsing the geometric sum of the type $\sum x^n$ using (see section Sequences and Series page 1656):

$$S_n=\sum u_0q^n=\frac{1-q^{n+1}}{1-q} \tag{14.2.104}$$

we get:

$$A_n=a_0\left(1+\frac{3}{5}\left(1-\left(\frac{4}{9}\right)^n\right)\right)=\frac{a_0}{5}\left(8-3\left(\frac{4}{9}\right)^n\right) \tag{14.2.105}$$

The limit of the area is then immediate:

$$\lim_{n\to+\infty} A_n=\lim_{n\to+\infty}\frac{a_0}{5}\cdot\left(8-3\left(\frac{4}{9}\right)^n\right)=\frac{8}{5}\cdot a_0 \tag{14.2.106}$$

So the area of the Koch snowflake is $8/5$ of the area of the original triangle.

What is so disturbing with Von Koch fractal is that we start from a line of finite length, to arrive at the end to a line of infinite length if we reiterate infinitly structure but it has a finished surface ... it's a "pathological" curve as the mathematicians say sometimes.

The Koch curve is legendary because it was used by Mandelbrot to write an article about the problem of measuring the length of the coasts of sea coasts (because the most the basic unit of measurement taken was small, more the perimeter ot the coast was great). He proposed to consider the coasts as fractals for which it is impossible to measure the perimeter but "fractal tree" or in other words: the fractal dimension.

The Koch Curve consists of 4^n subsets with magnification factor 3^n. So the fractal dimension is:

$$d=\frac{\log(4^n)}{\log(3^n)}=\frac{n\log(3)}{n\log(4)}\cong 1.2619 \tag{14.2.107}$$

14.2.2.5.1 Coastline paradox

It is now the right moment in our point of view to speak about the "coastline paradox" that is the counterintuitive observation that the coastline of a landmass does not have a well-defined length. This results from the fractal-like properties of coastlines. The first recorded observation of this phenomenon was by Lewis Fry Richardson and it was expanded by Benoit Mandelbrot.

More concretely, the length of the coastline depends on the method used to measure it. Since a landmass has features at all scales, from hundreds of kilometers in size to tiny fractions of a millimeter and below, there is no obvious size of the smallest feature that should be measured around, and hence no single well-defined perimeter to the landmass. Various approximations exist when specific assumptions are made about minimum feature size.

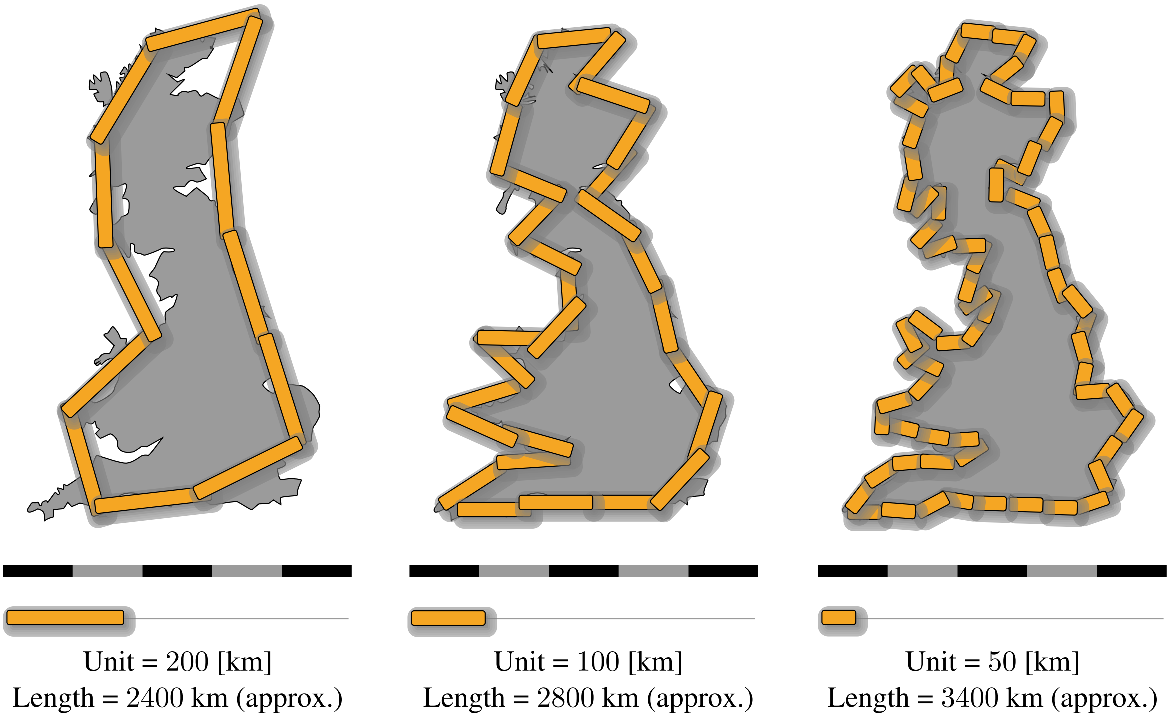

Unit = 200 [km]
Length = 2400 km (approx.)
Unit = 100 [km]
Length = 2800 km (approx.)
Unit = 50 [km]
Length = 3400 km (approx.)

Figure 14.279: Coastline paradox

The length of "true fractal" therefore always diverges to infinity, as if one were to measure a coastline with infinite, or near-infinite resolution, the length of the infinitely smaller bends of the coastline would add up to infinity. However, this figure relies on the assumption that space can be subdivided indefinitely. The truth value of this assumption - which underlies Euclidean geometry and serves as a useful model in everyday measurement - is a matter of philosophical speculation, and may or may not reflect the changing realities of 'space' and 'distance' on the atomic level (approximately the scale of a nanometer). The Planck length, many orders of magnitude smaller than an atom, is proposed as the smallest measurable unit possible in the universe.

In reality, permanent features of the coastline of order of size 1 cm or less do not exist, because of erosion and other action of the sea. In most places the minimum size is much larger than this. Thus the concept of an infinite fractal is not applicable to the coastline. For practical considerations, an appropriate choice of minimum feature size is on the order of the units being used to measure. If a coastline is measured in kilometers, then small variations much smaller than one kilometer are easily ignored.

14.2.2.6 Natural fractals

Besides the purely mathematical aspect of fractals, we can find, via heuristics, contracting applications for fractal shapes similar to that we can find in nature. Let us see some examples with Maple™ 4.00b always taking first for common basis of all fractals that follow, the following procedures (if the copy/paste form the book in Maple™ 4.00b does not work, simply rewrite the code in the Maple™ console):

```
>transforme_point := proc(t, p)
[t[1]*p[1]+t[2]*p[2]+t[5], t[3]*p[1]+t[4]*p[2]+t[6]]
end:

>transforme_triangle := proc(t, triangle)
local i;
[seq(transforme_point(t, triangle[i]), i=1 ..  3)]
end:

>transform_square := proc(t, square)
local i;
[seq(transforme_point(t, square[i]), i=1 ..  4)]
end:

>IFS := proc(n, liste_de_transformations,col)
local i, j, k, s, sequence_de_triangles:
options `Copyright by Alain Schauber, 1996`;
sequence_de_triangles := [[0, 0], [1, 0], [0.5, 1]];
for j to n do
s := NULL;
for i to nops(liste_de_transformations) do
s := s,
seq(transforme_triangle(liste_de_transformations[i],
op(k, [sequence_de_triangles])),
k=1 ..  nops([sequence_de_triangles]))
od;
sequence_de_triangles := s
od;
plots[polygonplot]([sequence_de_triangles], axes=none, color=col,
scaling=constrained)
end:

>IFSS := proc(n, liste_de_transformations,col)
local i, j, k, s, seq_square:
seq_square :=[[0,0],[1,0],[1,1],[0,1]];
for j to n do
s := NULL;
for i to nops(liste_de_transformations) do
s := s,
seq(transform_square(liste_de_transformations[i],
op(k, [seq_square])),
k=1 ..  nops([seq_square]))
od;
seq_square := s
od;
plots[polygonplot]([seq_square], axes=none, color=col,
scaling=constrained)
end:
```

14.2.2.6.1 Branch

We start from:

```
>rameau:=[[.387,.430,.430,-.387,.2560,.5220],
[.441,-.091,-.009,-.322,.4219,.5059], [-.468,.020,-.113,.015,.4,.4]]:
```

And we get:

```
> IFSS(0,rameau,green);
```

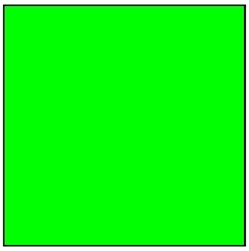


Figure 14.280: **Start set for generic branch fractal**

```
> IFSS(1,rameau,green);
```

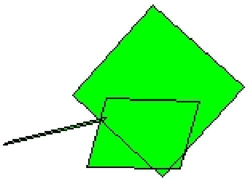


Figure 14.281: **First iteration set for generic branch fractal**

```
> IFSS(2,rameau,green);
```

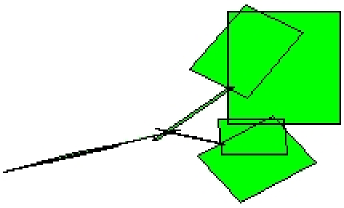


Figure 14.282: **Second iteration set for generic branch fractal**

```
> IFSS(3,rameau,green);
```

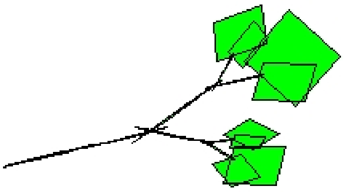


Figure 14.283: **Third iteration set for generic branch fractal**

```
> IFSS(4,rameau,green);
```

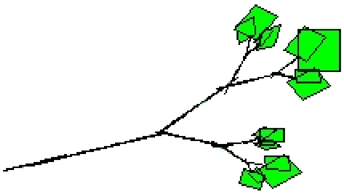


Figure 14.284: **Fourth iteration set for generic branch fractal**

```
> IFSS(5,rameau,green);
```

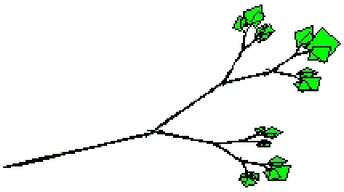


Figure 14.285: Fifth iteration set for generic branch fractal

```
> IFSS(6,rameau,green);
```

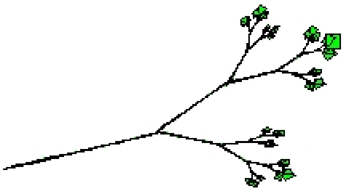


Figure 14.286: Sixth iteration set for generic branch fractal

14.2.2.6.2 Snowflake

We start from:

```
>cristal:=[[.255,0,0,.255,.3726,.6714],[.255,0,0,.255,.1146,.2232],
[.255,0,0,.255,.6306,.2232],[.37,-.642,.642,.37,.6356,-.0061]]:
```

And we get:

```
>IFSS(0, cristal,green);
```

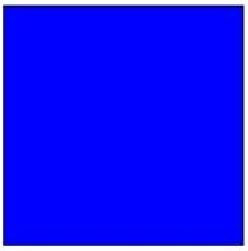


Figure 14.287: Start set for snowflake fractal

```
> IFSS(1, cristal,green);
```

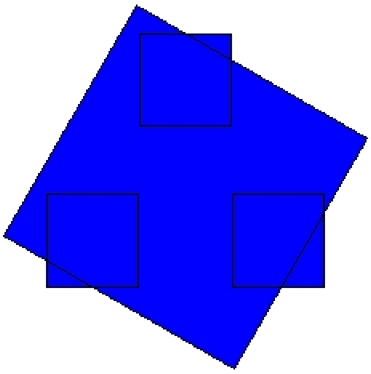


Figure 14.288: **First iteration set for snowflake fractal**

```
> IFSS(2, cristal,green);
```

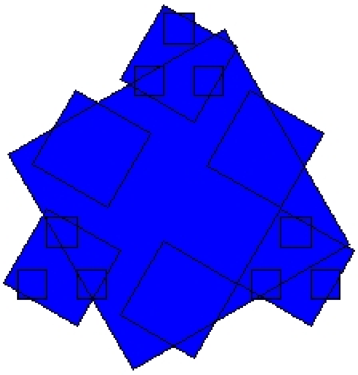


Figure 14.289: **Second iteration set for snowflake fractal**

```
> IFSS(3, cristal,green);
```

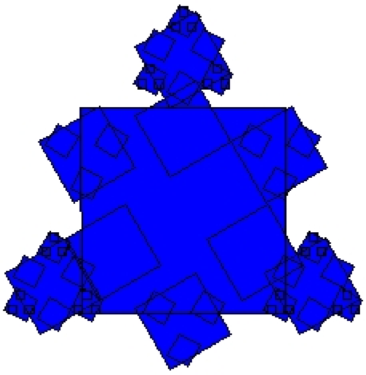


Figure 14.290: **Third iteration set for snowflake fractal**

```
> IFSS(4, cristal,green);
```

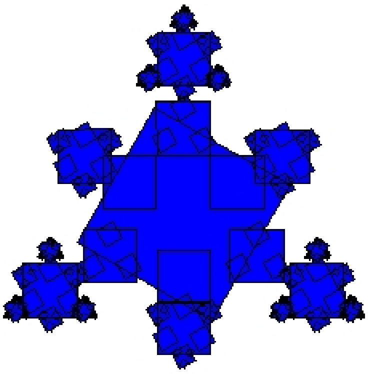


Figure 14.291: **Fourth iteration set for snowflake fractal**

```
> IFSS(5, cristal,green);
```

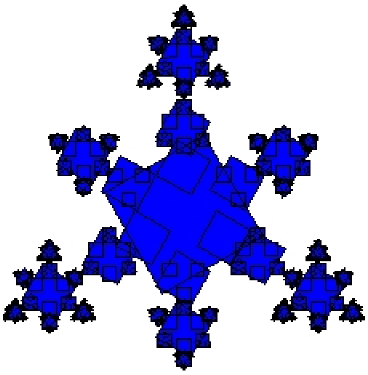


Figure 14.292: **Fifth iteration set for snowflake fractal**

```
> IFSS(6, cristal,green);
```

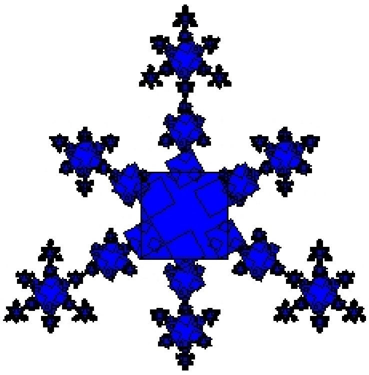


Figure 14.293: **Sixth iteration set for snowflake fractal**

```
> IFSS(7, cristal,green);
```

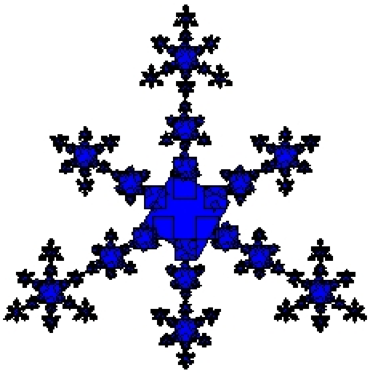


Figure 14.294: Seventh iteration set for snowflake fractal

14.2.2.6.3 Tree

We start from:

```
>tree := [[-0.04, 0, -0.23, -0.65, -0.08, 0.26], [0.61, 0, 0, 0.31, 0.07,
2.5],        [0.65, 0.29, -0.3, 0.48, 0.54, 0.39], [0.64, -0.3, 0.16, 0.56,
-0.56, 0.4]]:
```

And we get:

```
>IFS(0, tree, green);
```

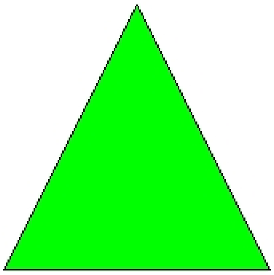


Figure 14.295: Start set for tree fractal

```
> IFS(1, tree, green);
```

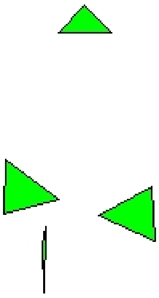


Figure 14.296: First iteration set for tree fractal

```
> IFS(2, tree, green);
```

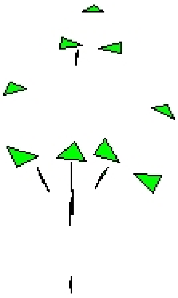


Figure 14.297: Second iteration set for tree fractal

```
> IFS(3, tree, green);
```

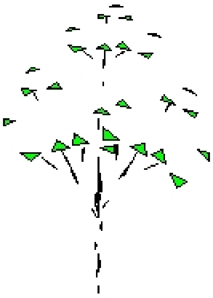


Figure 14.298: **Third iteration set for tree fractal**

```
> IFS(4, tree, green);
```

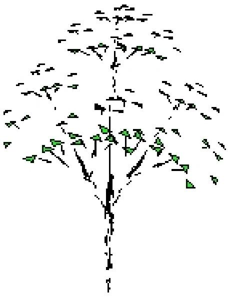


Figure 14.299: **Fourth iteration set for tree fractal**

```
> IFS(5, tree, green);
```

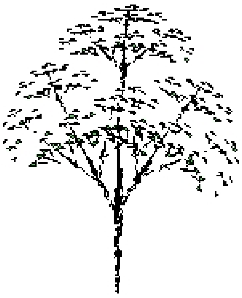


Figure 14.300: **Fifth iteration set for tree fractal**

14.2.2.6.4 Fern

We start from:

```
>fougere:=[[0,0,0,0.16,0,0],[0.2,-0.26,0.23,0.22,0,1.6],
[-0.15,0.28,0.26,0.24,0,0.44],[0.85,0.04,-0.04,0.85,0,1.6]]:
```

And we get:

```
>IFS(0, fougere, blue);
```

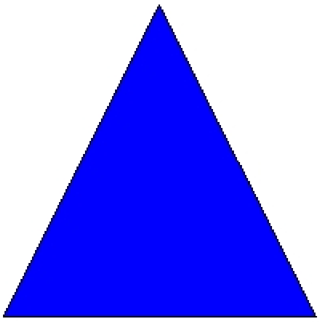


Figure 14.301: Start set for fern fractal

```
>IFS(1, fougere, blue);
```

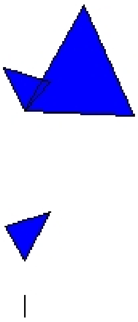


Figure 14.302: First iteration set for fern fractal

```
>IFS(2, fougere, blue);
```

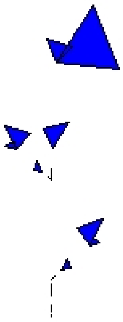


Figure 14.303: Second iteration set for fern fractal

```
>IFS(3, fougere, blue);
```

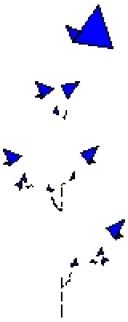


Figure 14.304: **Third iteration set for fern fractal**

```
>IFS(4, fougere, blue);
```

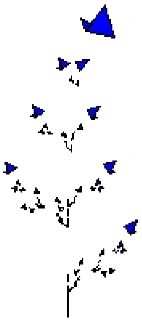


Figure 14.305: **Fourth iteration set for fern fractal**

```
>IFS(5, fougere, blue);
```

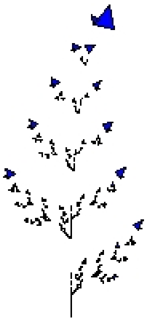


Figure 14.306: **Fifth iteration set for fern fractal**

```
>IFS(6, fougere, blue);
```

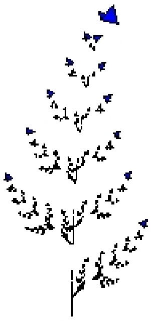


Figure 14.307: Sixth iteration set for fern fractal

```
>IFS(7, fougere, blue);
```

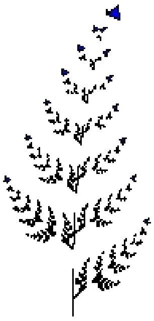


Figure 14.308: Eighth iteration set for fern fractal

etc. Unitl we get:

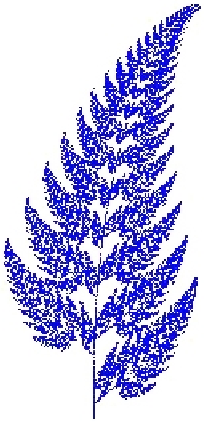
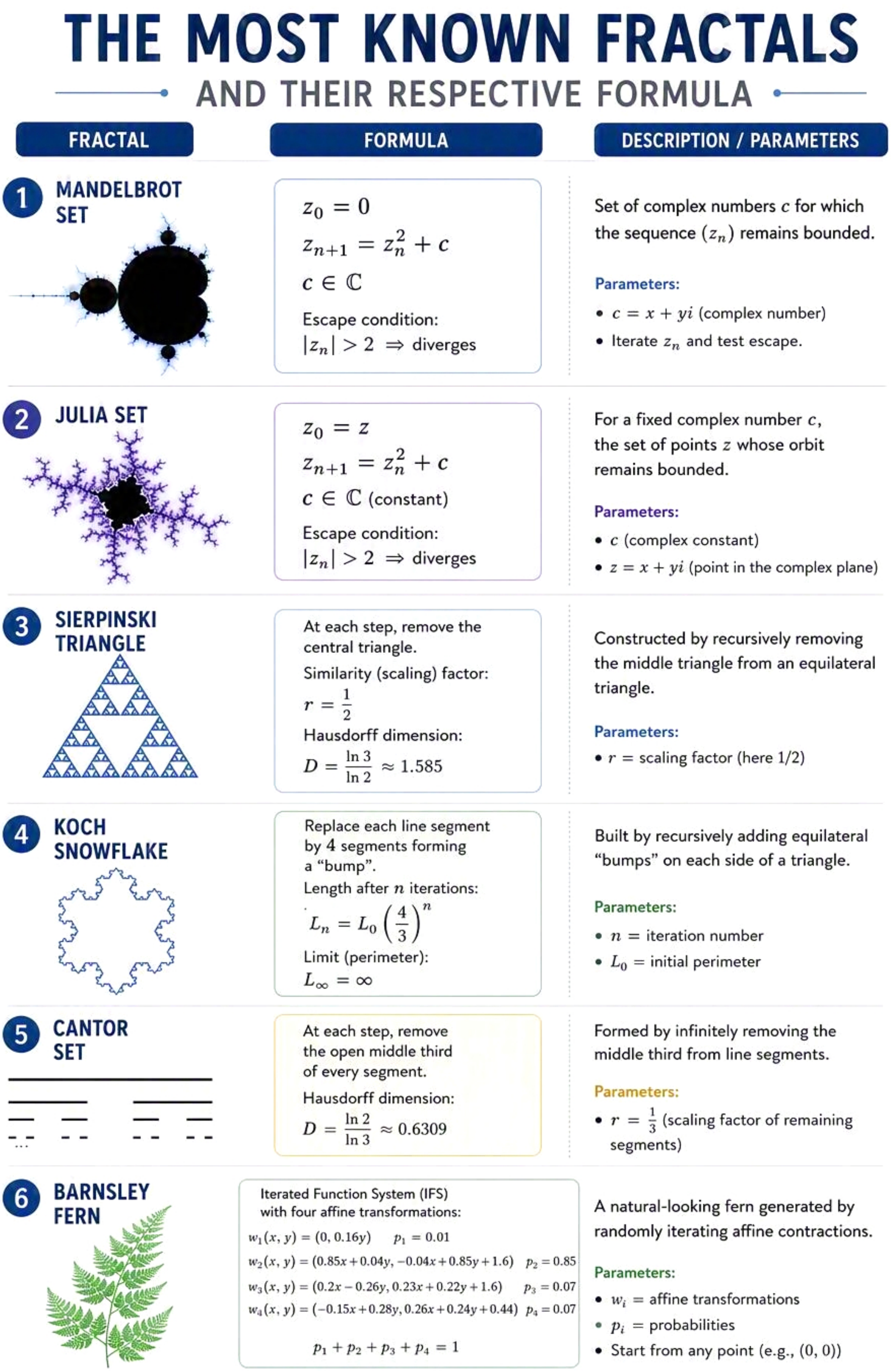


Figure 14.309: Attractor set for fern fractal

and we will stop here because the examples of IFS and natural fractals are uncountable...



THE MOST KNOWN FRACTALS

AND THEIR RESPECTIVE FORMULA

FRACTAL	FORMULA	DESCRIPTION / PARAMETERS
1 MANDELBROT SET	$z_0 = 0$ $z_{n+1} = z_n^2 + c$ $c \in \mathbb{C}$ Escape condition: $\lvert z_n \rvert > 2 \Rightarrow$ diverges	Set of complex numbers c for which the sequence (z_n) remains bounded. Parameters: • $c = x + yi$ (complex number) • Iterate z_n and test escape.
2 JULIA SET	$z_0 = z$ $z_{n+1} = z_n^2 + c$ $c \in \mathbb{C}$ (constant) Escape condition: $\lvert z_n \rvert > 2 \Rightarrow$ diverges	For a fixed complex number c, the set of points z whose orbit remains bounded. Parameters: • c (complex constant) • $z = x + yi$ (point in the complex plane)
3 SIERPINSKI TRIANGLE	At each step, remove the central triangle. Similarity (scaling) factor: $r = \frac{1}{2}$ Hausdorff dimension: $D = \frac{\ln 3}{\ln 2} \approx 1.585$	Constructed by recursively removing the middle triangle from an equilateral triangle. Parameters: • r = scaling factor (here 1/2)
4 KOCH SNOWFLAKE	Replace each line segment by 4 segments forming a "bump". Length after n iterations: $L_n = L_0 \left(\frac{4}{3}\right)^n$ Limit (perimeter): $L_\infty = \infty$	Built by recursively adding equilateral "bumps" on each side of a triangle. Parameters: • n = iteration number • L_0 = initial perimeter
5 CANTOR SET	At each step, remove the open middle third of every segment. Hausdorff dimension: $D = \frac{\ln 2}{\ln 3} \approx 0.6309$	Formed by infinitely removing the middle third from line segments. Parameters: • $r = \frac{1}{3}$ (scaling factor of remaining segments)
6 BARNSLEY FERN	Iterated Function System (IFS) with four affine transformations: $w_1(x,y) = (0, 0.16y) \quad p_1 = 0.01$ $w_2(x,y) = (0.85x + 0.04y, -0.04x + 0.85y + 1.6) \quad p_2 = 0.85$ $w_3(x,y) = (0.2x - 0.26y, 0.23x + 0.22y + 1.6) \quad p_3 = 0.07$ $w_4(x,y) = (-0.15x + 0.28y, 0.26x + 0.24y + 0.44) \quad p_4 = 0.07$ $p_1 + p_2 + p_3 + p_4 = 1$	A natural-looking fern generated by randomly iterating affine contractions. Parameters: • w_i = affine transformations • p_i = probabilities • Start from any point (e.g., (0, 0))

Figure 14.310: **Summary of most well knows basic fractals**

14.2.3 Escape Time Algorithm Fractals

Several methods have been proposed to construct fractal images as we mentioned at the beginning of this section. The are generally three way of generating fractals that are well known:

- Iterated function systems (seen previously)
- Escape time fractals (we will see now)
- Random fractals

So we will now focus on the methods named "escape time methods".

For this, we place ourselves in the complex plane $\mathbb{C}$ consisting of the points $M(x,y)$ of affix:

$$z=x+\mathrm{i}y \tag{14.2.108}$$

Afterwards we consider a complex sequence defined by:

$$u(0)=z_0 \tag{14.2.109}$$

and:

$$u(n+1)=f(u(n)) \tag{14.2.110}$$

the function f being a complex continuous function. We will assumed that f is build in such a way that is has a fixed point x_0, that is to say that there exists x_0 such as:

$$f(x_0)=x_0 \tag{14.2.111}$$

It is therefore simply the application of fixed-point theorem already mentioned several times before. Under some conditions on f and equation, we find that the following points does not diverge (which means it is not interested in points that converge but to those who do not diverge!). This method is the basis for the construction of Mandelbrot and Julia Sets.

Building a fractal image from such a set of defined sequences, is equivalent to study for each pair (x,y) of the plane the behaviour of the sequence itself. We then associates a color to each result (that is to say, each pair (x,y)) representing the "speed" of divergence of the sequence.

To study the convergence of a series, we look at his first n elements, if we detect that the divergence conditions are satisfied then we can say that this sequence diverges, otherwise this result is potentially convergent. We notice that more n is big, the more accurate are the results (but the computing time will be great).

The simplest algorithm for generating a representation of an escape time fractal (set) consist in repeating a calculation performed for each x, y point in the plot area and based on the behaviour of that calculation, a color is chosen for that pixel.

The x and y locations of each point are used as starting values in a repeating, or iterating calculation (described in detail below). The result of each iteration is used as the starting values for the next. The values are checked during each iteration to see whether they have reached a critical "escape" condition. If that condition is reached, the calculation is stopped, the pixel is drawn, and the next x, y point is examined. For some starting values, escape occurs quickly, after only a small number of iterations. For starting values very close to but not in the set, it may take hundreds or thousands of iterations to escape. For values within the Mandelbrot Set, escape will never occur. The programmer or user must choose how much iteration, or "depth", they wish to examine. The higher the maximal number of iterations, the more detail and subtlety emerge in the final image, but the longer time it will take to calculate the fractal image.

Escape conditions can be simple or complex. Because no complex number with a real or imaginary part greater than 2 can be part of the set, a common bailout is to escape when either coefficient exceeds2. A more computationally complex method that detects escapes sooner, is to compute distance from the origin using the Pythagorean theorem, i.e., to determine the absolute value, or modulus, of the complex number. If this value exceeds 2, the point has reached escape. More computationally intensive rendering variations include the Buddhabrot method, which finds escaping points and plots their iterated coordinates.

The color of each point represents how quickly the values reached the escape point. Often black is used to show values that fail to escape before the iteration limit, and gradually brighter colors are used for points that escape. This gives a visual representation of how many cycles were required before reaching the escape condition.

To render such an image, the region of the complex plane we are considering is subdivided into a certain number of pixels. To color any such pixel, let c be the midpoint of that pixel. We now iterate the critical point under the chosen function, checking at each step whether the orbit point has modulus larger than R the convergence radius. When this is the case, we know that c does not belong to the escape time fractal, and we color our pixel according to the number of iterations used to find out. Otherwise, we keep iterating up to a fixed number of steps, after which we decide that our parameter is "probably" in the escape time fractal, or at least very close to it, and color the pixel black.

In pseudocode, this algorithm would look as follows. The algorithm does not use complex numbers and manually simulates complex-number operations using two real numbers, for those who do not have a complex data type. The program may be simplified if the programming language includes complex-data-type operations.

Algorithm 18: Escape Time Fractal pseudocode algorithm

```
Data: I,R
for each pixel $(P_x, P_y)$ on the screen do
    $x_0 =$ scaled $x$ coordinate of pixel ;
    $y_0 =$ scaled $y$ coordinate of pixel ;
    $x = 0.0$;
    $y = 0.0$;
    iteration= $0$;
    max_iteration= $I$;
    while $x^2 + y^2 < R^2$ AND $i <$ max_iteration do
        $x = \Re(z(x,y)) + x_0$;
        $y = \Im(z(x,y)) + y_0$;
        $i := i + 1$;
    color := palette[iteration];
plot $(P_x, P_y$, colors);
```

14.2.3.1 Mandelbrot Set

We construct the Mandelbrot Set through iterations in the complex plane (this is named also "holomorphic dynamics"). The function is of the form:

$$\boxed{f(z) = z^2 + c} \tag{14.2.112}$$

where c a constant parameter such that $c \in \mathbb{C}$ (so we double the argument of the initial complex number we squared its norm!).

The first term of the sequence is zero. So we have the following defined by:

$$z_0 = 0 \quad \text{and} \quad z_{n+1} = z_n^2 + c \tag{14.2.113}$$

Why do we start with $z_0 = 0$?: Because zero is the critical point of $z^2 + c$, that is to say the point satisfying the extremum:

$$\frac{\mathrm{d}(z^2 + c)}{\mathrm{d}z} = 0 \tag{14.2.114}$$

For each point of affix $x + \mathrm{i}y$ of the plane, we study the above sequence for $c = x + \mathrm{i}y$. If the sequence diverges, we say that the tested point does not belong to the Mandelbrot Set, if the sequence converges, we say that the point belongs to the Mandelbrot Set defined then by (the definition and name is due to Adrien Douady, in tribute to Benoît Mandelbrot):

$$\boxed{\begin{cases} z_n = z_{n-1}^2 + c \\ z_0 = 0, c \in \mathbb{C} \end{cases}} \tag{14.2.115}$$

To reproduce a basic representation of the Mandelbrot fractal, we associate to c complex values of the plane. It is generally considered the portion of the complex plane having as real part, values between -2.5 and 1.5, and as imaginary part, values between -1.5 and 1.5. This portion of the complex plane is divided so to form a grid whose elements will be associated with values of c. For each value of c, we get a result that modules can converge (bounded sequence) or diverge (non-bounded sequence).

In practice, it is considered that sequence of modules converges if the first 30 modules are less than 2 ($R=2$ in the previous pseudocode), that is to say $|z_{30}|\leq 30$. When the sequence of the modules converges (bounded sequence), we color in black the grid point. After considering all points of the grid, we get a set of blackened points named: "Mandelbrot Set" or "Mandelbrot fractal" as we already know and denoted by $\mathcal{M}$. What constitutes a remarkably curious result!

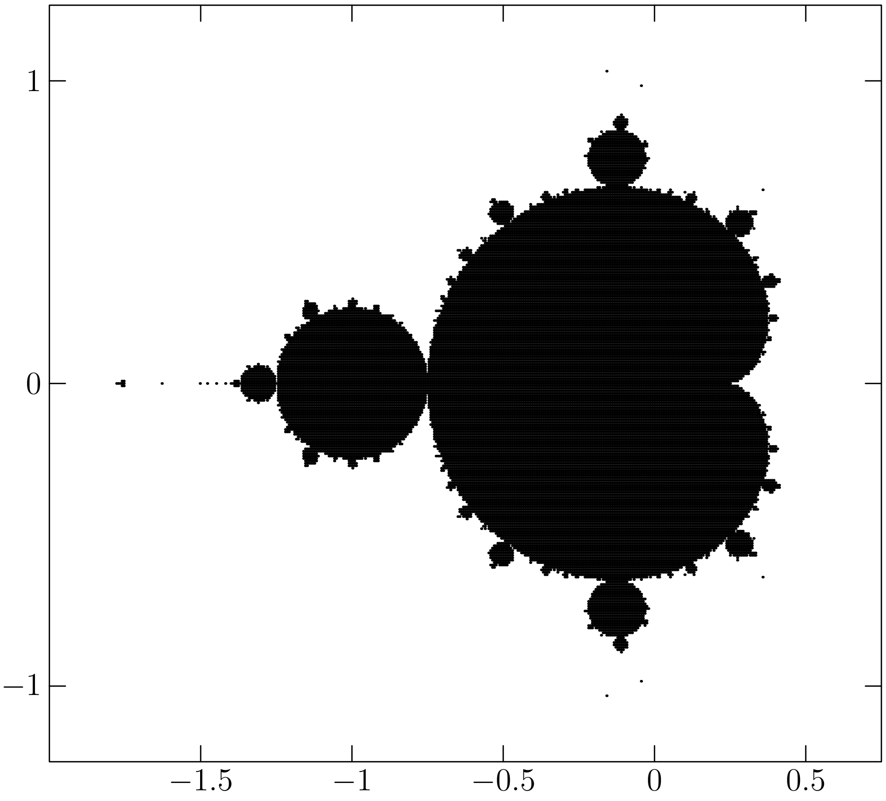


Figure 14.311: **Bichromic Mandelbrot Fractal**

We can also color the points outside the Mandelbrot Set using colors that depend on the number of terms calculated before obtaining a module greater than or equal to 2. The Module points of the same color can be interpreted as points away at the same speed of the Mandelbrot Set.

Remark

The list of z_i generated by the iteration is named the "orbit" of z_0.

We can also discover the Mandelbrot fractal in-deep using the following Maple™ 4.00b code derived from the earlier pseudocode (available usually in high-school). The reader just have to copy the program below on a Maple™ worksheet and indicate instead of -2..1..$1.5-1.5$ of the last line, the range of the real and imaginary parts of c he wants to discover:

```
>restart:  with(plots):
>couleur:=proc(a,b)
local x,y,xi,yi,n;
x:=a;
y:=b;
for n from 0 to 30 while evalf(x^2+y^2) < 4 do;
xi:=evalf(x^2-y^2+a);
yi:=evalf(2*x*y+b);
x:=xi;
y:=yi;
od;
n
end:
```

You will therefore get the following result (the equivalent code for MATLAB® and `R` are given in the companion books about MATLAB® and `R`):

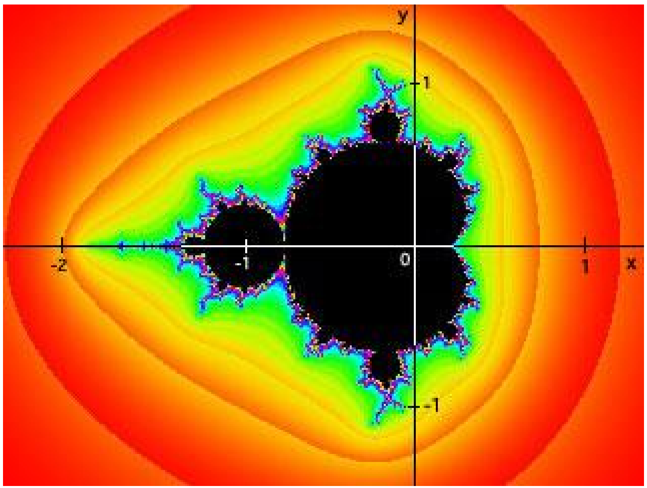


Figure 14.312: **Mandelbrot Fractal with Maple™ 4.00b**

The Mandelbrot Set is self-similar in the vicinity of points named "Misiurewicz points":

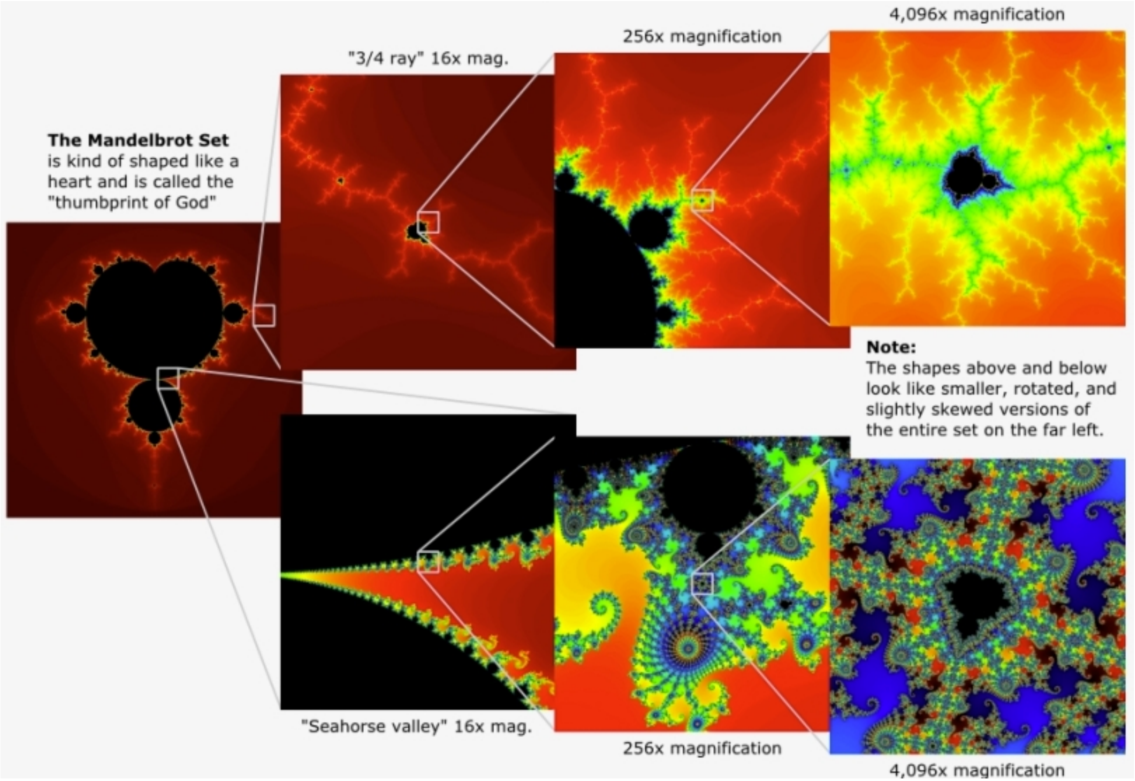

The Mandelbrot Set
is kind of shaped like a
heart and is called the
"thumbprint of God"
"3/4 ray" 16x mag.
256x magnification
4,096x magnification
Note:
The shapes above and below
look like smaller, rotated, and
slightly skewed versions of
the entire set on the far left.
"Seahorse valley" 16x mag.
256x magnification
4,096x magnification

Figure 14.313: **Self-Similarity of Mandelbrot Fractal**

It seems also that it is possible to prove (we are still searching for the proof...) that the Mandelbrot fractal real axes can be put in correspondence with the logistic bifurcation diagram that we have studied in the section of Population (see page 7327) Dynamics such that (it's a perfect example of chaos producing order!):

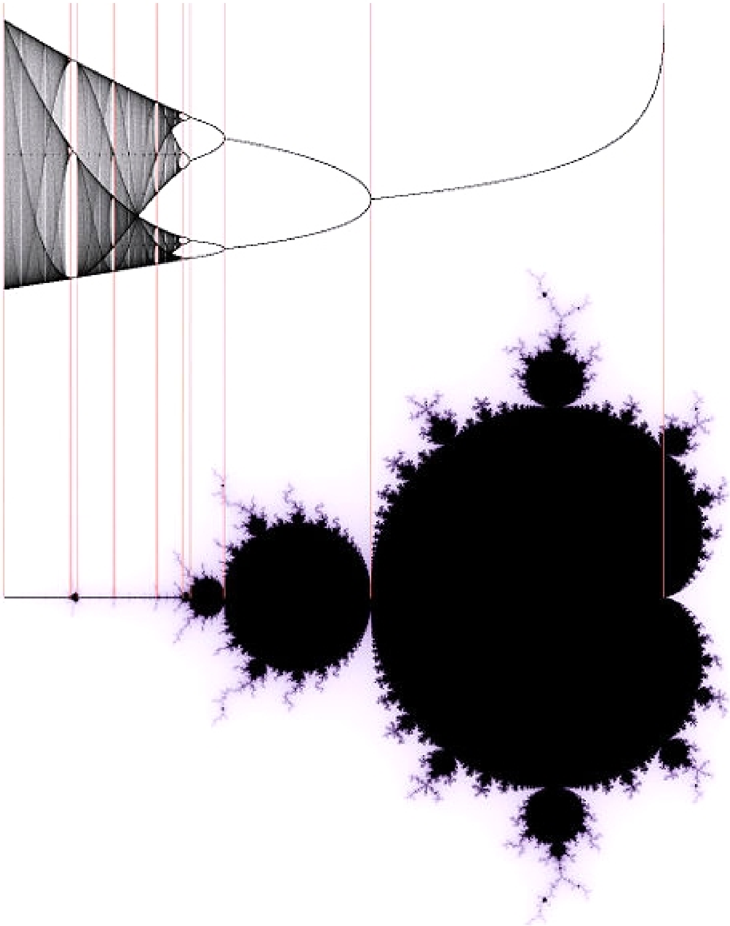


Figure 14.314: **Mandelbrot-Logistic bifurcation correspondence (source: Wikipedia)**

The functions are obviously both quadratic. In fact, the Mandelbrot Set can be recoded into the form of the Logistic map (and vice versa).

Indeed, the Mandelbrot Fractal is obtained by iterating:

$$z_{n+1}=z_n^2+c \tag{14.2.116}$$

and we know that the logistic bifurcation is obtained by iterating:

$$x_{n+1}=rx_n(1-x_n)=rx_n-rx_n^2 \tag{14.2.117}$$

So let us put in the Mandelbrot function the (anticipated) change of variable such that z_n is dependent of x_n and r but c must as constant also be only dependent of the constant r:

$$z_n=\mathrm{i}r\left(x_n-\frac{1}{2}\right)\qquad\text{and}\qquad c=\frac{1}{4}r \tag{14.2.118}$$

Then:

$$\begin{aligned}z_{n+1}=z_n^2+c&=\left(\mathrm{i}r\left(x_n-\frac{1}{2}\right)\right)^2+\frac{1}{4}r=-r\left(x_n-\frac{1}{2}\right)^2+\frac{1}{4}r\\&=-r\left(x_n^2-x_n+\frac{1}{4}\right)+\frac{1}{4}r=-rx_n^2+rx_n-r\frac{1}{4}+r\frac{1}{4}=rx_n-rx_n^2\end{aligned} \tag{14.2.119}$$

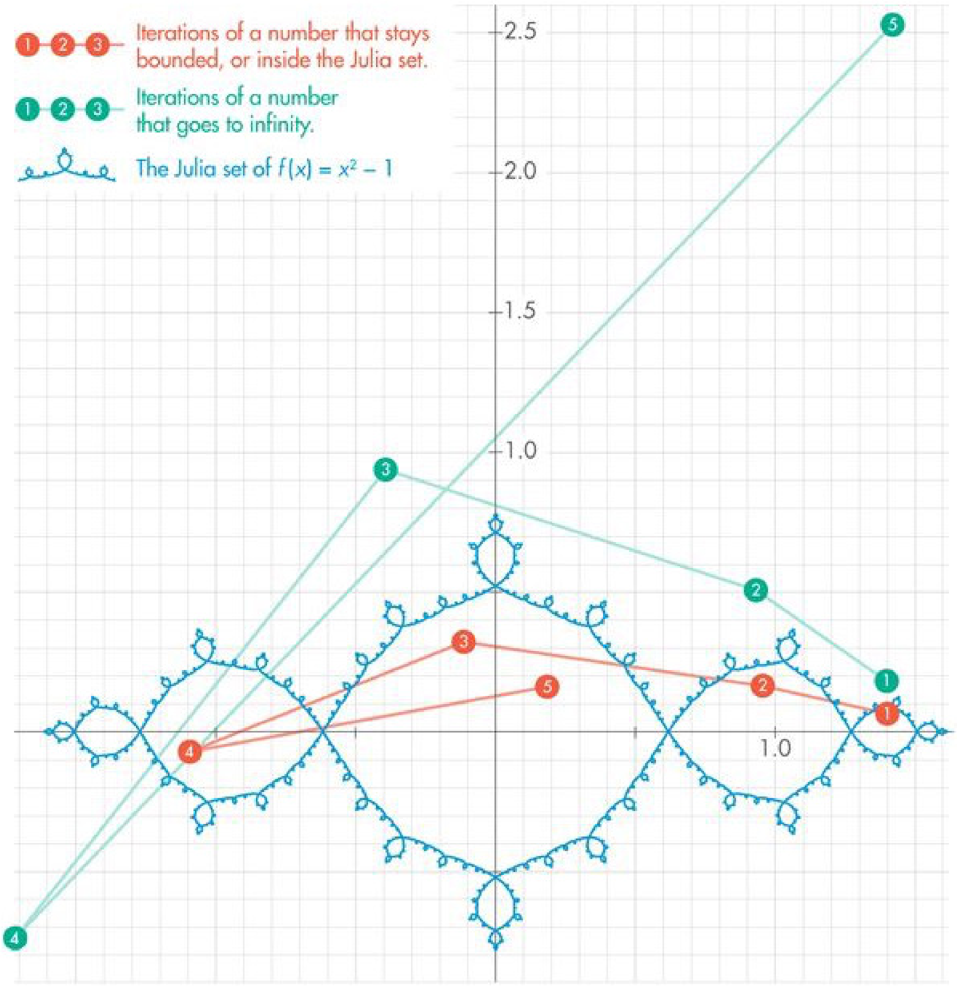
14.2.3.2 Julia Set

The Julia Set is built almost in the same way that the Mandelbrot Set (since the Julia Set is actually a subset of it after investigation!). In the Mandelbrot Set, c scans the plane. For the Julia Set, c is fixed during the computation of the image. To each c corresponds a particular set that will be denoted $J(c)$. What varies this time is z_0, which takes the value of the point to test. It is therefore z_0 that scans the plan.

A point of initial coordinates (x_0, y_0) and affix $x+\mathrm{i}y$ belongs to $J(c)$ if and only if the sequence defined by:

$$z_0 = x+\mathrm{i}y \quad \text{and} \quad z_{n+1} = z_n^2 + c \tag{14.2.120}$$

converge (is bounded).



① ② ③ Iterations of a number that stays bounded, or inside the Julia set.
① ② ③ Iterations of a number that goes to infinity.
The Julia set of f(x) = x² − 1
2.5
2.0
1.5
1.0
1.0
1
2
3
4
5

Figure 14.315: **Makingofa Julia Set**

In fact, the Mandelbrot Set is the set of points c such that the Julia Set of parameter c is connex to, that is to say we can always found a c such that when starting with z_0 in the Mandelbrot Set is equivalent after a few iterations as starting with a fixed $z_0 \neq 0$ and a given c (thus the Mandelbrot Set generalizes all the Julia Sets !!!). So the figure of the Mandelbrot Set contains figures all the Julia Sets, which is remarkable (but logic ...!):

If again we adapt the pseudocode given earlier, we get to a given scale factor the fractal shown below (obtained through the small Maple™ 4.00b code below and already used earlier for the Mandelbrot fractal):

```
>restart; with(plots):
>julia:= proc(c,x, y)local z, m;
z:= evalf(x+y*I);
for m from 0 to 30 while abs(z) < 3 do
z:= z^2 + c
od;
m
end:

>J:= proc(d)
global phonyvar;
phonyvar:= d;
(x, y) -> julia(phonyvar, x, y)
end:

>plot3d(0, -2..2, -1.3..1.3, style=patchnogrid,orientation=[-90,0],
grid=[270, 270],scaling=constrained, color=J(-1.25));
```

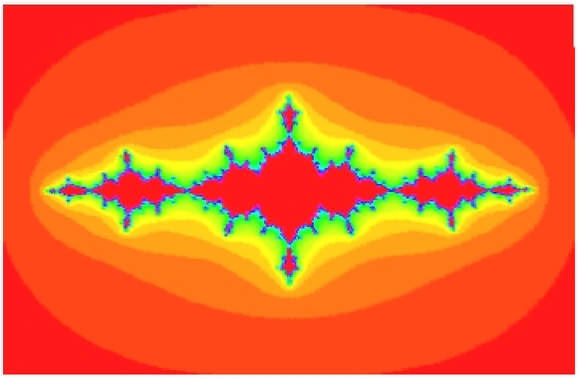


Figure 14.316: **Julia Fractal with Maple™ 4.00b**

and to show that the Mandelbrot Set contains all the Julia Sets we have qualitatively:

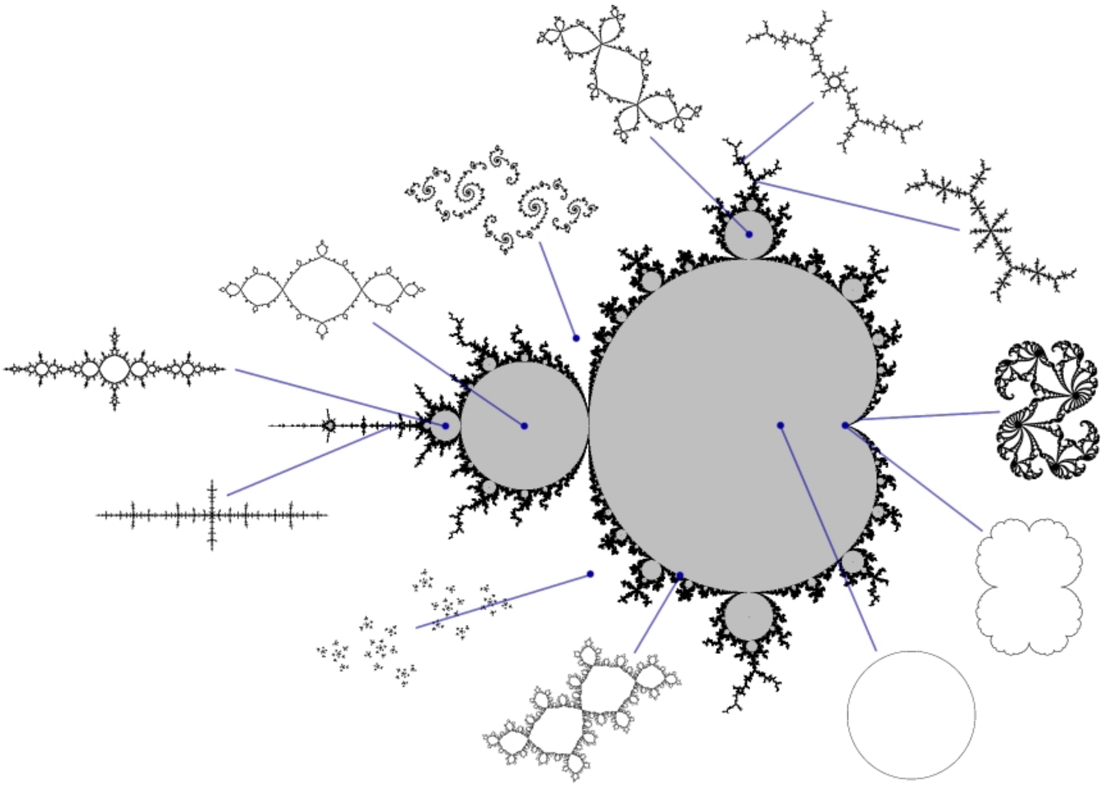
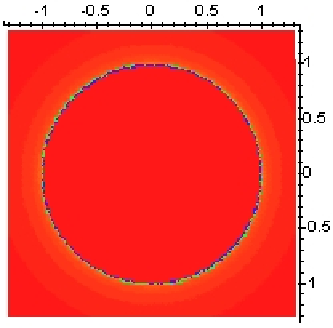
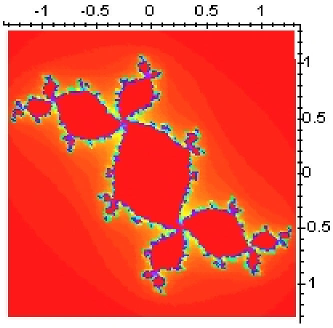


Figure 14.317: Illustration of "fatherhood" ... of the Mandelbrot Set

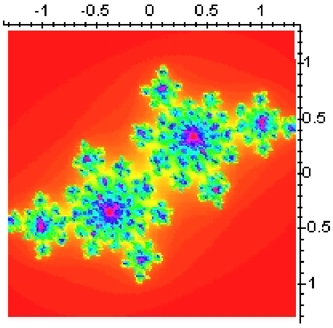
So we must be able to write a single algorithm (see below) that achieves all of Julia fractals by simply selecting a good starting point as shown in the following figures (we can see on top right the "Douady's Rabbit Fractal" also named "dragon fractal":



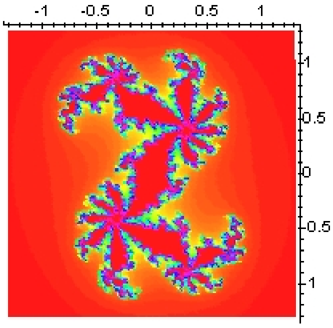
$z_0 = 0 + i0$



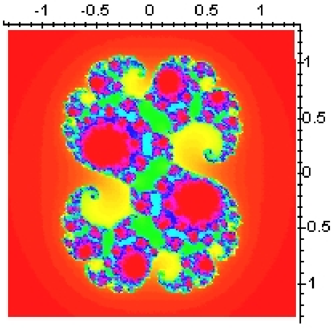
$z_0 = -0.123 + i0.745$



$z_0 = -0.414 - i0.612$



$z_0 = 0.382 + i0.147$



and so on...

$z_0 = 0.284 - i0.0122$

That we get with the following Maple™ 4.00b code:

```
>couleur:=proc(a,b)
local x,y,xi,yi,n;
global reel,imaginaire;
x:=a;
y:=b;
for n from 0 to 100 while evalf(x^2+y^2)<4 do;
xi:=evalf(x^2-y^2+reel);
yi:=evalf(2*x*y+imaginaire);
x:=xi;
y:=yi;
od;
n;
```

```
end:

>reel:=-0.181;
>imaginaire:=-0.667;

>plot3d(0,(-13/10)..(13/10),(-13/10)..(13/10),orientation=[-90,0],
style=patchnogrid,scaling=constrained,axes=framed,numpoints=20000
,color=couleur);
```

14.2.3.3 Newton Set

Newton Sets are also so named because they arise from the resolution of the problem looking for zeros of a function by the Newton's method (see section Numerical Methods page 6612).

Given f a function with values in $\mathbb{C}$ and differentiable in $\mathbb{C}$, we take z_0 in $\mathbb{C}$ such that:

$$\boxed{z_{n+1}=z_n\frac{f(z_i)}{f'(z_i)}} \tag{14.2.121}$$

There are then two ways to proceed:

1. Either we focus on $|z_0-z_i|$ and then we do same as before.
2. Either we wonder to which zero r_k the sequence converges and we focus on $|z_i-z_k|$.

If again we translate our pseudocode algorithm, into Maple™ language we get to a given scale factor the fractal shown below with Maple™ 4.00b:

```
>restart:
>newton:= proc(x, y)
local z, m;
z:= evalf(x+y*I);
for m from 0 to 50 while abs(z^3-1) >= 0.001
do
z:= z - (z^3-1)/(3*z^2)
od;
m
end:

>plot3d(0,-2..2,-1.5..1.5,orientation=[-90,0],grid=[250, 250],
style=patchnogrid,scaling=constrained,color=newton);
```

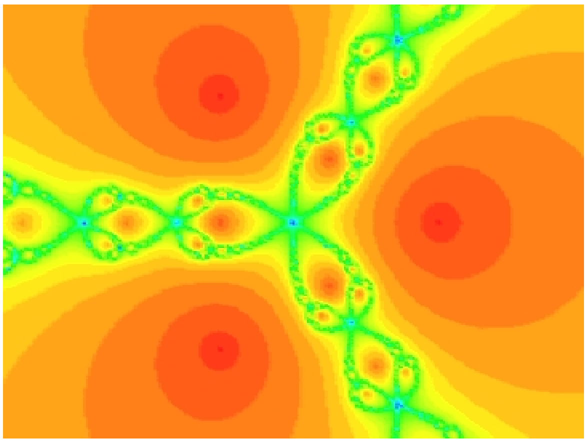


Figure 14.318: **Newton Set fractal with Maple™ 4.00b**

~70 %

★★★☆☆

20 votes, 66.00%

14.3. Logical Systems

THE reader familiar with the purpose of this textbook should not expect to see here any schemes of buttons, switches, timing diagrams or MIL standard wiring diagrams. We will remain in a purely formal framework of logical systems and their tools.

Definitions (#14.3.590):

D1. We speak of "asynchronous logic model" (commonly named "sequential logic model") when the outputs of a system depends on the chronological order in which the entries will succeed.

D2. We speak of "combinatorial logic model" when the outputs of a system depend only on the combination of the input variables.

> **Remark**
>
> We differentiate the "strict logic" of "fuzzy logic" that will both be defined in the details later.

14.3.1 Strict Logic

Consider first a set which we will denote by $\mathcal{B}$ with two elements (more formally denoted by $\perp, \top$).

Definitions (#14.3.591):

D1. A "strict logic variable" or "boolean variable" is an element of $\mathcal{B}$ that has only two states 0 and 1 (as opposed to a fuzzy variable whose value can be between 0 and 1). It is represented by Latin uppercase letters or lowercase Latin letters (depending on your choice).

D2. A multivariate "logic function F" of n variables applies $\mathcal{B}^n$ in $\mathcal{B}$ such that:

$$F : \mathcal{B}^n \mapsto \mathcal{B} \tag{14.3.1}$$

It combines to a n-tuple of logical variables $(b_0, b_1, \ldots, b_{n-1})$ a value $F(b_0, b_1, \ldots, b_{n-1})$.

D3. There are different ways of expressing a logic function ("Boolean function"). A function of n variables is fully described by stating the values of this function for the set (or the subset of definition) of the combinations of the n-tuple variables:

$$F(0, \ldots, 0, 0), F(0, \ldots, 0, 1), F(0, \ldots, 1, 0), F(1, \ldots, 1, 1) \tag{14.3.2}$$

This statement usually takes the form of a table with $n+1$ columns and no more than 2^n lines, each line exposing a combination of variables and the corresponding value of the function. The following table gives the general form of a "truth table" function of three variables completely defined through a function F (we already saw some simple examples in the section of Proof Theory at page 131):

A	B	C	$F(A,B,C)$
0	0	0	$F(0,0,0)$
0	0	1	$F(0,0,1)$
0	1	0	$F(0,1,0)$
0	1	1	$F(0,1,1)$
1	0	0	$F(1,0,0)$
1	0	1	$F(1,0,1)$
1	1	0	$F(1,1,0)$
1	1	1	$F(1,1,1)$

Table 14.38: **Generic truth table**

The elements of input of the systems will be considered as Boolean variables on which we can build a ring structure set, that by adding a particular axiom, we can bring to an algebra (in the computational sense and the set one!) commonly named "Boolean Algebra" as we will see now.

So Boolean Algebra is an algebra on itself (with a ring structure as we will define it rigorously later) proposing to translate signals with a value of the type $0/1$ (assimilated to: True/False) in mathematical expressions. For this, we define each elementary signal by "logical variables" and their treatment by "logical functions". Methods ("truth tables") exists to define the operations that we want to achieve, and to transcribe the result into an algebraic expression. Thanks the rules we will see later, these expressions can be simplified. This will allow to represent with simple symbols a logic circuit capable of performing basic arithmetic operations, that is to say a circuit that design the core components (at logic level) regardless of the realization through transistors (physical level).

14.3.1.1 Boolean Algebra

Boolean Algebra (or "Boolean ring" to a given axiom...) is therefore a structure which is most often used in electronic (or microelectronics/optoelectronics) this is why some people name it sometimes "Switching Algebra" or "logic gates". Therefore, a processor is composed of transistors for performing functions on digital signals. These transistors assembled together form components for performing simple functions. From these components it is possible to create circuits performing fairly complex operations. Boolean Algebra (named after the English mathematician George Boole 11815-11864 according to holocene calendar) is a mean to create more or less easily such circuits.

Remark

It would be better before you start reading this chapter, to read at least in diagonal the subsection about Logic in the section of Proof Theory (see page 131) and on algebraic structures in the section of Set Theory (see page 378).

It is necessary for a rigorous definition of a Boolean Algebra to give it in terms of abstract algebra.

Reminder: A "Boolean Algebra" $(\mathcal{B}, \vee, \wedge)$ is a set containing two particular elements $\perp, \top$, (abstract forms the 0 and 1) and has two internal composition laws, $\vee, \wedge$ (AND and logical OR) and verifies the following axioms to form a ring structure such that $\forall a, b, c \in \mathcal{B}$:

A1. Associativity: $(a \vee b) \vee c = a \vee (b \vee c)$ and $(a \wedge b) \wedge c = a \wedge (b \wedge c)$

A2. Commutativity: $a \vee b = b \vee a$ and $a \wedge b = b \wedge a$

A3. Absorption: $a \wedge (a \vee b) = a$ and $a \vee (a \wedge b) = a$

A4. Distributivity: $(a \vee b) \wedge c = (a \wedge c) \vee b \wedge c$ and $(a \wedge b) \vee c = (a \vee c) \wedge (b \vee c)$

A5. Idempotence: $a \vee a = a$ and $a \wedge a = a$

A6. Completation (or inversion): a has a complement (negation) denoted by $\neg a$ or $\bar{a}$ (NOT) such as: $a \wedge \neg a = \top$ and $a \vee \neg a = \perp$

Remark

The first four axioms establish a ring structure. The fifth axiom (idempotence) added to the first four defines the concept of "Boolean Algebra".

Strictly speaking to form a Boolean Algebra we required a symmetrical element (see section Set Theory page 417) with one of the two fundamental operators and we cannot do this directly with the two previous operators previously defined. That is why the real operators of Boolean Algebra are normally the $\wedge$ (AND) and the Δ (symmetric difference), denoted in Boolean Algebra by the symbol $\oplus$, the latter one being given by the logic operation:

$$x \Delta y = (x \wedge \bar{y}) \vee (y \wedge \bar{x}) \tag{14.3.3}$$

but to simplify, in the early school grades, it is common that we do implicitly reference to it without going into details.

It follows that the binary set $(\mathcal{B}, 0, 1)$ is therefore relatively to the laws $\vee, \wedge$ an "Abelian group". (see section Set Theory page 421) Therefore, $(\mathcal{B}, \vee)$ being an Abelian group, the law $\wedge$ being associative and distributive with respect to $\vee$, $(\mathcal{B}, \vee, \wedge)$ is therefore a "commutative ring with unit" (see section Set Theory page 426) since $\mathcal{B}$ has a neutral element relatively to the law $\vee$.

Remarks

R1. Thus, the operations $\vee, \wedge$ admits each a neutral element such that the value 1 is the neutral element of $\wedge$ and 0 the neutral element of $\vee$.

R2. The two operations that we usually use to form a Boolean Algebra are the "inclusive OR" rigorously denoted $\vee$ but more frequently denoted by the addition sign "$+$" and the "inclusive AND" rigorously denoted by $\wedge$ but more frequently denotey by the multiplication sign "$\cdot$".

The preceding axioms may, however been proved from the "axioms of the definition":

A1. Negation: $\top\neg\bot$ and $\bot\neg\top$

A2. Double complentation: $\neg\neg a=\bar{\bar{a}}=a$

A3. Neutral element 1: $\top$ is the neutral element of $\wedge$

A4. Neutral element 2: $\bot$ is the neutral element of $\vee$

A5. De Morgan theorem: $\neg(a\vee b)=\neg a\wedge\neg b$ and $\neg(a\wedge b)=\neg a\vee\neg b$

Remark

De Morgan's theorem can be proved using a simple truth table or algebraically as we will see just a little further below.

It thus follows the following dual expressions:

$$\left.\begin{array}{l}0\cdot 0=0\\1+1=1\end{array}\right\}\text{dual expression}$$

$$\left.\begin{array}{l}1\cdot 1=1\\0+0=0\end{array}\right\}\text{dual expression} \tag{14.3.4}$$

$$\left.\begin{array}{l}1\cdot 0=0\\0+1=1\end{array}\right\}\text{dual expression}$$

$$\left.\begin{array}{l}\bar{0}=1\\\bar{1}=0\end{array}\right\}\text{dual expression}$$

We name these expressions "dual expressions" because by replacing in one equation logic equation, the 0 by the 1, the $\cdot$ by $+$ and inversely, by this same equation remains verified.

Theorem 14.23. Let us see now what we name the "constants theorem" that consists to prove that:

$$\begin{aligned}a\cdot 0&=0\\a\cdot 1&=a\\a+0&=a\\a+1&=1\end{aligned} \tag{14.3.5}$$

Proof 14.23.1. The proof is trivial (if necessary the reader can quickly do a truth table) as it comes from the same property of the concept of "Boolean ring" and the identity element 1 relative to the $\wedge$ and its neutral element 0 relative to $\vee$.

■ Q.E.D.

Theorem 14.24. We have:

$$a+\bar{a}b=a+b \tag{14.3.6}$$

Proof 14.24.1. The distributivity brings us to write:

$$a+\bar{a}b=(a+\bar{a})(b+\bar{b}) \tag{14.3.7}$$

and applying the complementation:

$$(a+\bar{a})(a+b)=1(a+b) \tag{14.3.8}$$

applying commutativity:

$$1(a+b)=(a+b)1 \tag{14.3.9}$$

and finally by applying the theorem of constants:

$$(a+b)1=a+b \tag{14.3.10}$$

■ Q.E.D.

This proof will allow us to prove the famous "consensus theorem":

Theorem 14.25. In Boolean Algebra, the consensus theorem or rule of consensus is the identity:

$$\begin{gathered} ax+b\bar{x}+ab=ax+bx \\ (a+x)(b+\bar{x})(a+b)=(a+x)(b+\bar{x}) \end{gathered} \tag{14.3.11}$$

Proof 14.25.1. To verify the consensus theorem relative to the logical product:

$$ax+b\bar{x}+ab=ax+bx \tag{14.3.12}$$

we can make use of a Venn diagram where we can see trivially that the therm term ab is contained in the other two:

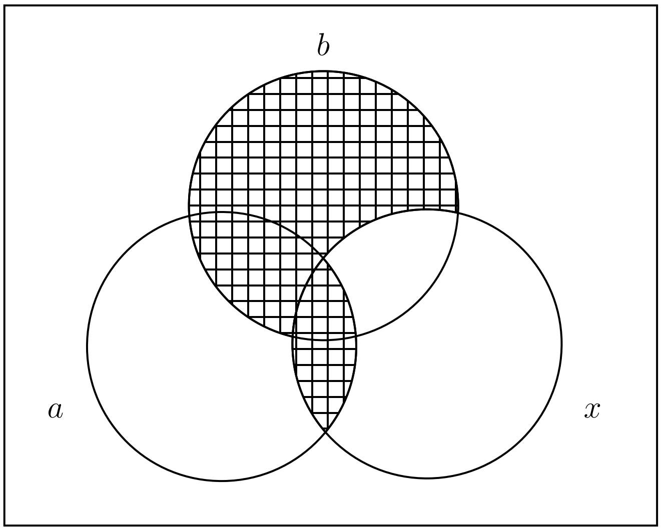


Figure 14.319: **Venn diagram of the consensus theorem**

Or more formally with other notations:

$$\begin{aligned}xy\vee\bar{x}z\vee yz&=xy\vee\bar{x}z\vee(x\vee\bar{x})yz\\&=xy\vee\bar{x}z\vee xyz\vee\bar{x}yz\\&=xy\vee xyz\vee\bar{x}z\vee\bar{x}yz\\&=xy(1\vee z)\vee\bar{x}z(1\vee y)\\&=xy\vee\bar{x}z=xy\vee\bar{x}z\end{aligned}\tag{14.3.13}$$

Proceeding the request of a reader we can also build a truth table:

$\boldsymbol{x}$	$\boldsymbol{y}$	$\boldsymbol{z}$	$\boldsymbol{xy\vee\bar{x}z\vee yz}$	$\boldsymbol{xy\vee\bar{x}z}$
0	0	0	0	0
0	0	1	1	0
0	1	0	0	0
0	1	1	1	0
1	0	0	0	0
1	0	1	0	0
1	1	0	1	0
1	1	1	1	0

Table 14.39: **Consensus theorem truth table**

Proceeding also with a Venn diagram, the reader will see without problem that we also have:

$$(a+x)(b+\bar{x})(a+b)=(a+x)(b+\bar{x})\tag{14.3.14}$$

■ Q.E.D.

Theorem 14.26. And finally the very famous "Shannon's theorem" (not to be confused with the Shannon theorem in signal theory!):

$$\begin{aligned}a+b&=ab+a\bar{b}+\bar{a}b\\ab&=(a+b)(a+\bar{b})(\bar{a}+b)\end{aligned}\tag{14.3.15}$$

Proof 14.26.1. We begin with the first relation:

$$\begin{aligned}a+b&=ab+a\bar{b}+\bar{a}b\\&=ab+a\bar{b}+\bar{a}b+ab\\&=a(b+\bar{b})+b(\bar{a}+a)\\&=a1+b1\\&=a+b\end{aligned}\tag{14.3.16}$$

and for the second relation:

$$\begin{aligned}ab&=(a+b)(a+\bar{b})(\bar{a}+b)\\&=(aa+a\bar{b}+ab+b\bar{b})(\bar{a}+b)\\&=(a+a(\bar{b}+b)+0)(\bar{a}+b)\\&=(a+a1+0)(\bar{a}+b)\\&=a(\bar{a}+b)=a\bar{a}+ab=ab\end{aligned}\tag{14.3.17}$$

■ Q.E.D.

Theorem 14.27. Now let us come back on the De Morgan theorems previously presented as axioms:

$$\overline{(a+b)}=\bar{a}\bar{b} \qquad \overline{(ab)}=\bar{a}+\bar{b} \tag{14.3.18}$$

These two relations therefore express that the inverse (or opposite) of product (or respectively the sum) of two variables is equal to the sum (respectively the product) of the inverse of these same variables.

Proof 14.27.1. Suppose $\overline{(a+b)}=\bar{a}\bar{b}$ is true. So under the relations $\bar{a}+a=1$ and $\bar{a}a=1$ (axiom of complementation) we must have:

$$\begin{aligned}(a+b)+\bar{a}\bar{b}&=1\\(a+b)\bar{a}\bar{b}&=0\end{aligned} \tag{14.3.19}$$

So we need to prove that these relations are true:

$$\begin{aligned}(a+b)+\bar{a}\bar{b}&=(a+b+\bar{a})(a+b+\bar{b})=(a+\bar{a}+\bar{b})(a+b+\bar{b})\\&=(1+b)(a+1)=1\end{aligned} \tag{14.3.20}$$

and:

$$(a+b)\bar{a}\bar{b}=a\bar{a}\bar{b}+\bar{a}b\bar{b}=0 \tag{14.3.21}$$

The second De Morgan theorem can be proven in the same way (we can put the details on request).

■ Q.E.D.

Remark

These two theorems can be extended to as many variables as we want.

Corollary 14.27.1. As immediate corollary we have:

- $ab=\overline{\bar{a}+\bar{b}}$
- $a+b=\overline{\bar{a}\bar{b}}$
- $ab=\overline{\overline{ab}}$
- $a+b=\overline{\overline{a+b}}$
- $\overline{a+b+c+\ldots}=\bar{a}\bar{b}\bar{c}\ldots$
- $\overline{abc\ldots}=\bar{a}+\bar{b}+\bar{c}+\ldots$

The logical expressions, as we have seen it until now thanks to the previous axioms, properties and theorems, can always be writtent into two different forms (by paling also with the negations $\neg$):

1. Under the form of a sum of logical products, also named "normal disjcontive form NDF", such as for example:

$$F=\sum_i\prod_i(b_i)=\bar{a}\bar{c}\bar{d}+\bar{a}c\bar{d} \tag{14.3.22}$$

   The constitutive terms of this polynomial are in this example the monomials $\bar{a}\bar{c}d, \bar{a}c\bar{d}$. The variables or complementary variables of the monomials are the "letters": $\bar{a}, \bar{c}, d, c\bar{d}$.

   Remark

   If each (all) of the product contains all the input variables in a direct or complementary form, then the form is named "first canonical form" or "disjonctive canonical form". Each of the products is therefore named the "minterm" (therefore in the previous relation, there are 2 minterms).

   Obviously if we consider that each term and its negation is equivalent to 0 and 1 if we have one variable, the first canonical form has two terms (2^1), if we have two variables, the first canonical form has four terms (2^n) and if n variables we have 2^n terms.

2. In the form of a product of logical sum, also named "normal conjunctive form NCF":

$$F=\prod_i\sum_i(b_i)=(\bar{a}+\bar{c}+d)(\bar{a}+c+\bar{d}) \tag{14.3.23}$$

   Remark

   If each (all) of the sums contains all the input variables in a direct or complementary form, the form is named "second canonical form" or "conjunctive canonical form". Each of the sum is therefore named the "maxterm".

   Obviously if we consider that each term and it negation is equivalent to 0 and 1 if we have one variable, the second canonical form has two terms (2^1), if we have two variables, the second canonical form has four terms (2^n) and if n variables we have 2^n terms.

   Therefore, in other words, a normal disjonctive form is either a litteral and its complementary (one letter) or a disjonction of formulas written as conjonction of litterals. A conjonctive form is either a litteral (one letter) and its complementary, or a conjonction of formulas written as disjonction of litterals.

The simplification methods we will see later aim to minimize the number of letters of the expressions so as to reduce the number of inputs of our logic system logic and therefore also the number of its components.

Remark

The algebraic simplification of an expression is to transform it so as to minimize the number of letters by applying theorems proven previously.

To simplify expressions (or identify them) a known technique is therefore to use the "Karnaugh tables" that we will see further below in details.

14.3.1.2 Logical Functions (Boolean operators)

So when we talk about Boolean Algebra unless other indication, we refer to the three basic Boolean operations (AND, OR, NOT) and some other logic functions arising from them for which we have the following symbols as defined in circuit theory (MIL norm representation if no error...):

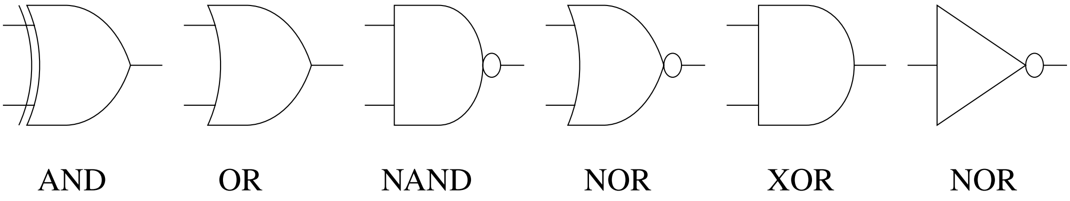


Figure 14.320: **Logic MIL Gates**

and their respective "truth tables":

AND (∧) Truth Table		
AND	**0**	**1**
0	0	0
1	0	1

Table 14.40: **AND (∧) Truth table**

OR (∨) Truth Table		
OR	**0**	**1**
0	0	1
1	1	1

Table 14.41: **OR (∨) truth table**

NOT (¬) Truth Table	
NOT	**-**
0	0
1	1

Table 14.42: **NOT (¬) truth table**

All others known (common) "logic functions" can be composed of these two basic operators. Such that by definition (given with their standard definition in the first line and with their different algebraic forms under their respective truth table):

$$\neg(a\wedge b)=\overline{a\cdot b}=\bar{a}+\bar{b}$$

NOT-AND (NAND)**:** NOT (a AND b)		
NAND	**0**	**1**
0	1	0
1	1	0

Table 14.43: **NOT-AND truth table**

$$\neg(a\vee b)=\overline{a+b}=\bar{a}\cdot\bar{b}$$

NOT-OR (NOR)**:** NOT (a OR b)		
NAND	**0**	**1**
0	1	0
1	0	0

Table 14.44: **NOT-OR truth table**

$$a\oplus b=(a\vee b)\wedge\neg(a\wedge b)=(a+b)\cdot\overline{(a\cdot b)}$$
$$a\oplus b=a\wedge\neg b+b\wedge\neg q=a\cdot\bar{b}+b\cdot\bar{a}$$
$$a\oplus b=\neg(a\wedge b+\neg a\wedge\neg b)=\overline{a\cdot b+\bar{a}\cdot\bar{b}}$$

EXCLUSIVE OR (XOR)**:** [a OR b] AND [NOT (a AND b]		
XOR	**0**	**1**
0	0	1
1	1	0

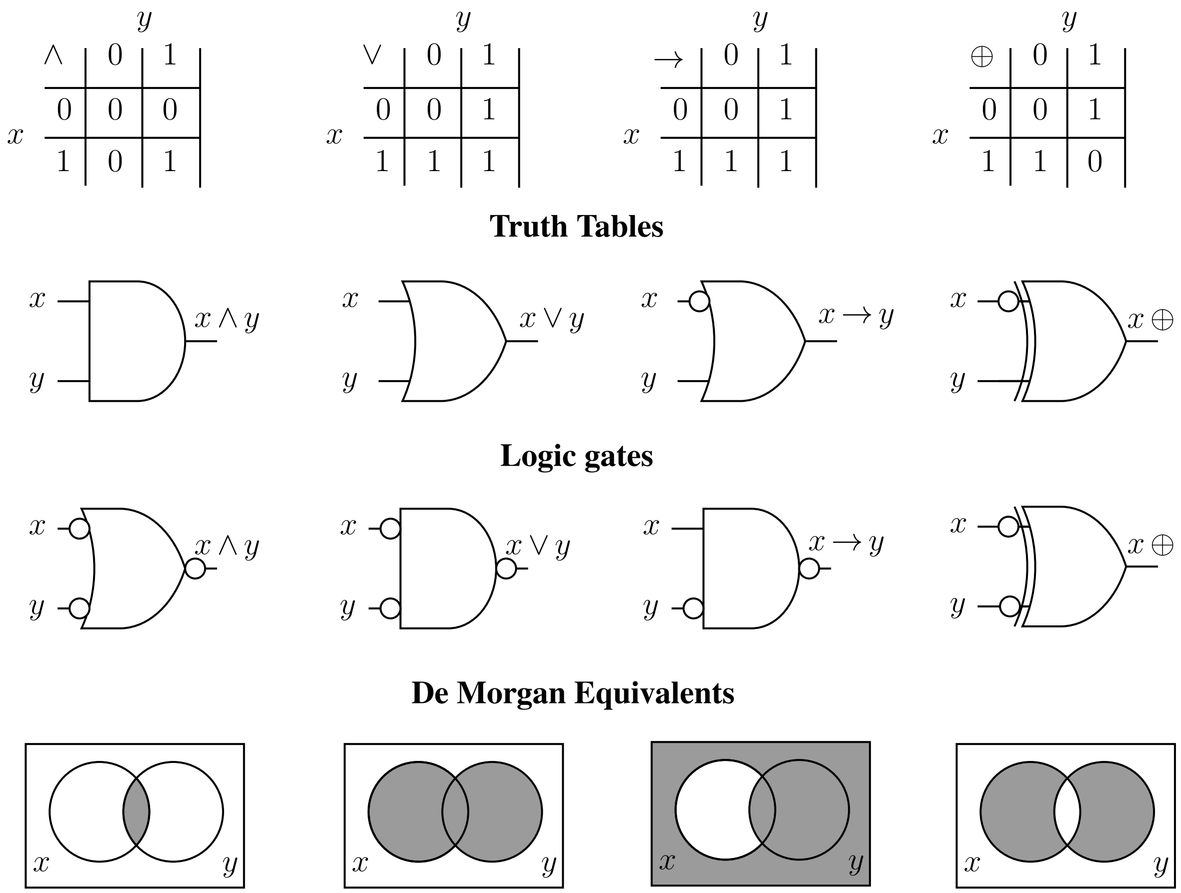
Table 14.45: **EXCLUSIVE OR $\oplus$ Truth table**

where a and b are, as you will have understood, variables (or "bit" of Binary Digit) that can arbitrarily take the binary values 0 or 1.

Remark

The XOR logic function is often denoted in the literature by the operator $\oplus$ and we will consider as obvious that the XOR is also a group law and thus allows to construct an abelian commutative group. This property of the XOR is particularly used in cryptography.

All the tables above can be sum up in the following figure with the corresponding MIL circuit theory symbols:



Truth Tables

$\wedge$	0	1
0	0	0
1	0	1

$\vee$	0	1
0	0	1
1	1	1

$\rightarrow$	0	1
0	0	1
1	1	1

$\oplus$	0	1
0	0	1
1	1	0

Logic gates

De Morgan Equivalents

Figure 14.321: **Summary of basic Boole algebra**

The reader can easily check by himself that with a NAND gate (or an AND gate and an INVERTER) you've got all the parts you need to build a modern computer (classical or quantum one!):

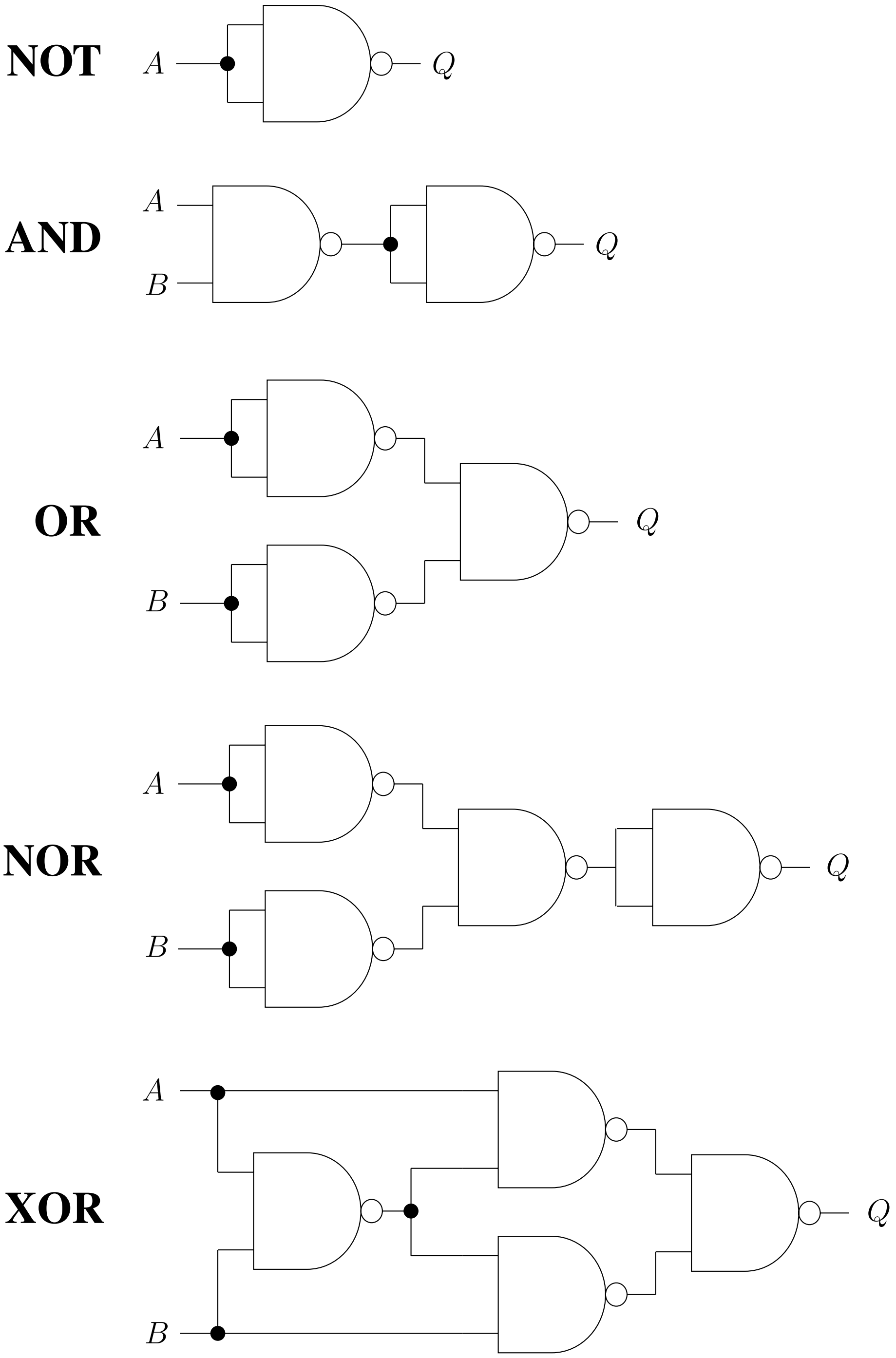

NOT
A
Q
AND
A
B
Q
OR
A
B
Q
NOR
A
B
Q
XOR
A
B
Q

Figure 14.322: **Construction of all gates from NAND gates**

14.3.1.3 Karnaugh maps

The "Karnaugh map", also known as the "K-map", is a method to simplify Boolean Algebra expressions. Maurice Karnaugh introduced it in 11953 (holocene calendar). The Karnaugh map reduces the need for extensive calculations by taking advantage of humans' pattern-recognition capability[83].

The required boolean results are transferred from a truth table onto a two-dimensional grid where the cells are ordered in Gray code, and each cell position represents one combination of input conditions, while each cell value represents the corresponding output value. Optimal groups of 1s or 0s are identified, which represent the terms of a canonical form of the logic in the original truth table. These terms can be used to write a minimal boolean expression representing the required logic.

Karnaugh maps are used to simplify real-world logic requirements so that they can be implemented using a minimum number of physical logic gates. A sum-of-products expression can always be implemented using AND gates feeding into an OR gate, and a product-of-sums expression leads to OR gates feeding an AND gate. Karnaugh maps can also be used to simplify logic expressions in software design. Boolean conditions, as used for example in conditional statements, can get very complicated, which makes the code difficult to read and to maintain. Once minimized, canonical sum-of-products and product-of-sums expressions can be implemented directly using AND and OR logic operators.

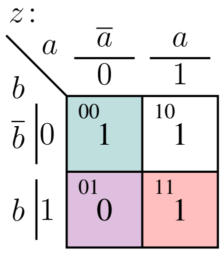
Let us consider for example the function:

$$z(a,b)=ab+\bar{a}b+\bar{a}\bar{b} \tag{14.3.24}$$

(in disjunctive normal form) and its respective truth table:

b	a	$z(a,b)$
0	0	1
0	1	0
1	0	1
1	1	1

The Karnaugh map is defined by a representation like the one below following the `karnaugh-map` package of LaTeX (as it does sadly not exist at this day any international norm on how to represent Karnaugh map):



---

[83] The reader can find easily on the Internet many K-map generators that simplifies expressions automatically!

The colored items are those that we need to keep. That is: $\bar{a}\bar{b}$, $\bar{a}b$, ab.

The Karnaugh table of a logic function thus has as many cells as possible combinations of variables which compose it, i.e. four cells for a function with two variables, and 2^n cells for a function with n variables. Each cell, which is at the intersection of a row and column of the Karnaugh table, has the state 0 or 1 that the function $z(a,b)$ takes for the corresponding logical product of the variables (minterms).

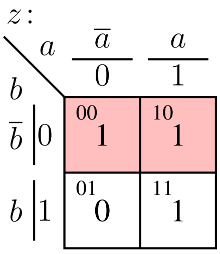
In the preceding example, however, we can see something interesting, the function $z(a,b)$, as we see very well, can be simplified in two ways:

$$z=b+\bar{a}\bar{b} \tag{14.3.25}$$

or also:

$$z=ab+\bar{a} \tag{14.3.26}$$

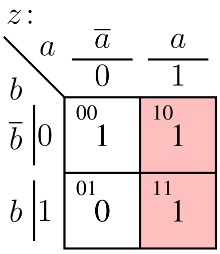
This possible simplification is always done with two adjacent minterms in the Karnaugh table such as for the first solution $z=b+\bar{a}\bar{b}$:



$z:$

b \ a	$\bar{a}$ 0	a 1
$\bar{b}$ 0	00 1	10 1
b 1	01 0	11 1

and for $z=ab+\bar{a}$:



$z:$

b \ a	$\bar{a}$ 0	a 1
$\bar{b}$ 0	00 1	10 1
b 1	01 0	11 1

We see that indeed the first grouping / simplification (horizontal) is done on the row $\bar{a}$ and the second grouping / simplification (vertical) is done on column b both results of the algebraic simplification of the function as following:

$$z=\bar{a}\bar{b}+\bar{a}b+ab=\bar{a}\bar{b}+(\bar{a}+a)b=\bar{a}\bar{b}+1\cdot b=\bar{a}\bar{b}+b \tag{14.3.27}$$

and:

$$z=\bar{a}\bar{b}+\bar{a}b+ab=\bar{a}(\bar{b}+b)+ab=\bar{a}\cdot 1+ab=\bar{a}+ab \tag{14.3.28}$$

So we could make the hypothesize that the Karnaugh table has for properties:

P1. To give us the normal disjunctive form of a function.

P2. That all adjoining cells with a value of 1 can be simplified in the respective symbol (letter) of their union

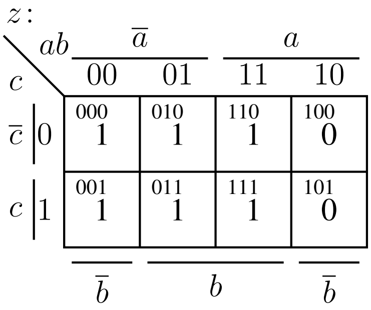
It is therefore an extremely powerful tool (algorithm) for simplifying and determining logical functions.

Let's look at an example with three variables!

First we give the truth table that can help for a better understanding:

a	b	c	$z(a,b,c)$
0	0	0	1
0	0	1	1
0	1	0	1
0	1	1	1
1	1	0	0
1	1	1	0
1	0	0	1
1	0	1	1

and the corresponding Karnaugh map:

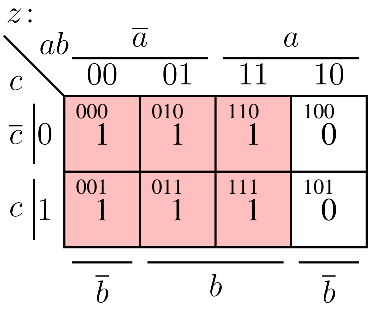


$z:$	$\bar{a}$		a	
c \ ab	00	01	11	10
$\bar{c}$ 0	000 1	010 1	110 1	100 0
c 1	001 1	011 1	111 1	101 0
	$\bar{b}$	b		$\bar{b}$

The normal disjonctive form is the given by all the cells that are equal to 1:

$$z(a,b,c)=\bar{a}\bar{b}\bar{c}+\bar{a}\bar{b}c+\bar{a}b\bar{c}+\bar{a}bc+ab\bar{c}+abc \tag{14.3.29}$$

This corresponds to the following colored Karnaugh map:

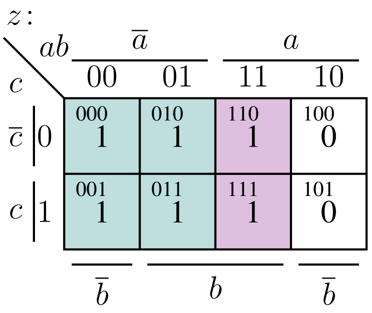


$z:$	$\bar{a}$		a	
c \ ab	00	01	11	10
$\bar{c}$ 0	000 1	010 1	110 1	100 0
c 1	001 1	011 1	111 1	101 0
	$\bar{b}$	b		$\bar{b}$

We see quickly that the previous normal disjonctive form can be simplified as:

$$z(a,b,c)=\bar{a}\bar{b}\bar{c}+\bar{a}\bar{b}c+\bar{a}b\bar{c}+\bar{a}bc+ab\bar{c}+abc=\bar{a}\bar{b}+\bar{a}b+ab=\bar{a}+ab \tag{14.3.30}$$

That gives the following Karnaugh map:

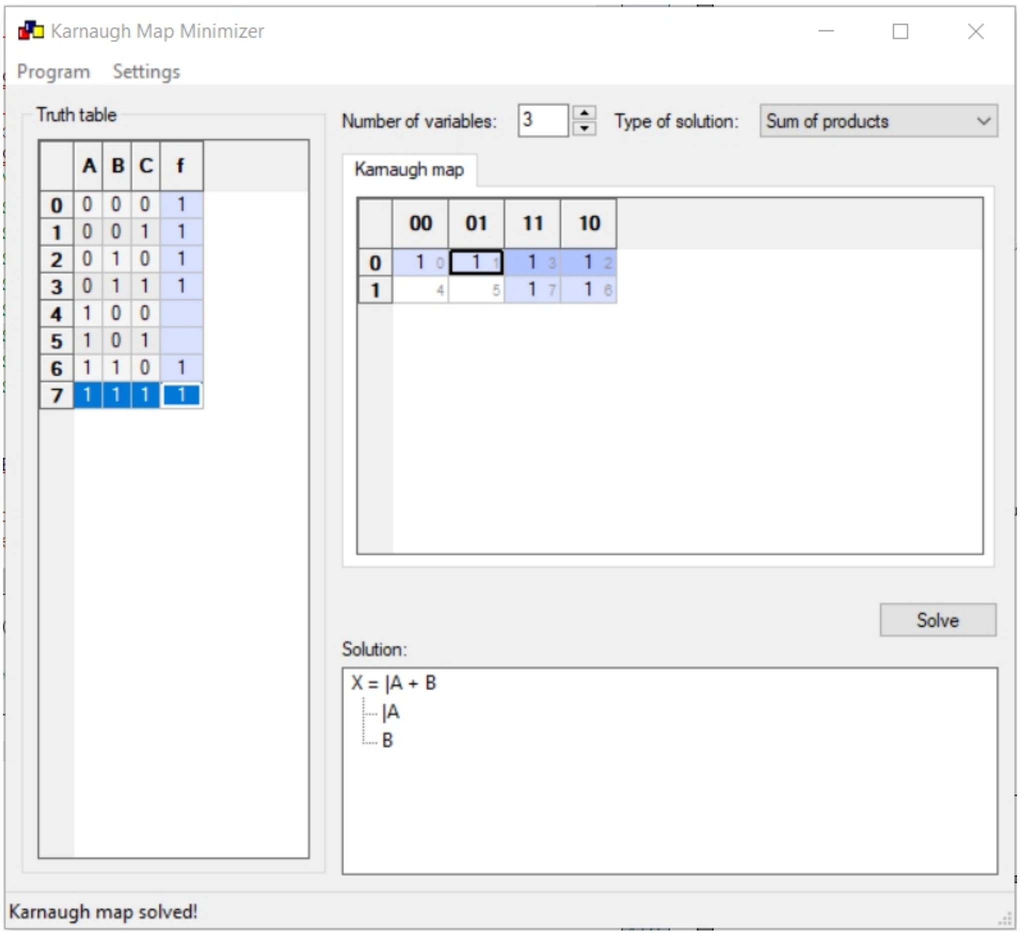


	00	01	11	10
0	000 1	010 1	110 1	100 0
1	001 1	011 1	111 1	101 0

But it is less obvious with a Karnaugh map to see that all previous relations can be simplified into (using binary addition rules):

$$z(a,b,c)=\bar{a}+b \tag{14.3.31}$$

This latter result can be seen with the available Karnaugh map minimizer freeware that we can find nowadays on the Internet. For example the freeware Karnaugh Map Minimizer that gives us for the previous example (we also see at the same time why it is boring - as always also in other field of engineering a science - that there are no ISO norm to standardize Karnaugh map representation at the days we write these lines):



Number of variables: 3 Type of solution: Sum of products

	A	B	C	f
0	0	0	0	1
1	0	0	1	1
2	0	1	0	1
3	0	1	1	1
4	1	0	0	
5	1	0	1	
6	1	1	0	1
7	1	1	1	1

	00	01	11	10
0	1 0	1 1	1 3	1 2
1	4	5	1 7	1 6

Solution:

```
X = |A + B
  |A
  B
```

Karnaugh map solved!

Figure 14.323: **Karnaugh Map Minimizer 0.4**

A difficulty remains however sometimes with this technique: how to choose the best construction of the table (disposition of letters)?

In fact, there is a specific way of associating the Boolean Algebra complement rule with what we name the "Gray code".

Definition (#14.3.592): The "reflected binary code (RBC)", also known as "Gray code" after Frank Gray, is a binary numeral system where two successive values differ in only one bit (binary digit). The reflected binary code was originally designed to prevent spurious output from electromechanical switches[84] and is useful to optimally build Karnaugh maps.

Here is an example of the Gray code for decimal 0 to 15 (the reader can notice that only one bit/switch at a time change at each row):

Decimal	Binary	Gray
0	0000	0000
1	0001	0001
2	0010	0011
3	0011	0010
4	0100	0110
5	0101	0111
6	0110	0101
7	0111	0100
8	1000	1100
9	1001	1101
10	1010	1111
11	1011	1110
12	1100	1010
13	1101	1011
14	1110	1001
15	1111	1000

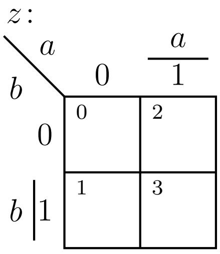
Table 14.46: **Gray code of 0 to 15**

Using Gray code we can create optimal Karnaugh tables. The reason is simple, the Gray code changes only one bit at a time at each increment as we have just seen. In practice this means that for two consecutive values, 1 and 2 for example, one of the two variables will be the opposite of the other one.

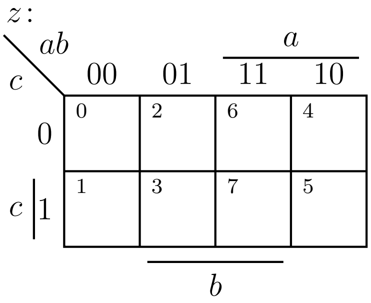
We can see this very well with the following structures:

---

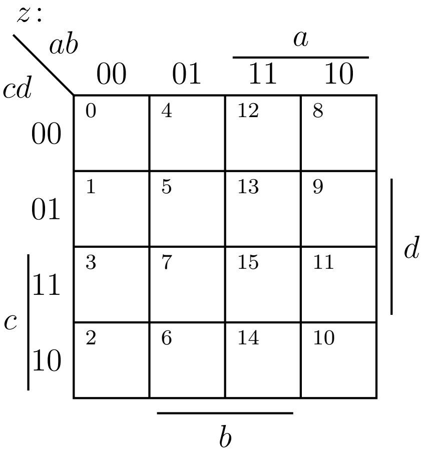
[84] The problem with natural binary codes is that physical switches are not ideal: it is very unlikely that physical switches will change states exactly in synchrony. In the transition between the two states shown above, all three switches change state.


z:
a
b
0
1
0
b
1
0
2
1
3

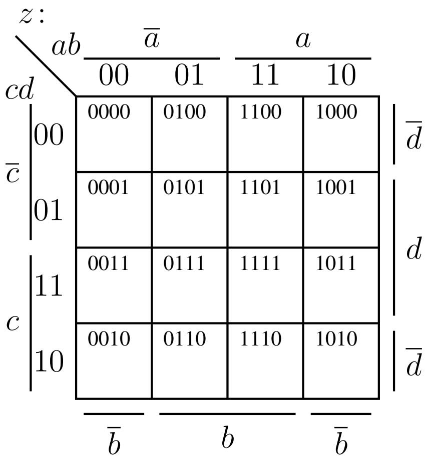
but more especially with greater tables (the order of the cell is not obvious):


z:
ab
a
c
00
01
11
10
0
c
1
0
2
6
4
1
3
7
5
b

or:


z:
ab
a
cd
00
01
11
10
00
01
11
10
c
d
0
4
12
8
1
5
13
9
3
7
15
11
2
6
14
10
b

or more explicitly for that latter (you can therefore compare with the above Gray code table with 4 digits to understand from where the internal numbering comes from):


z :
ab
$\bar{a}$
a
00 01 11 10
cd
$\bar{c}$
c
00
01
11
10
0000 0100 1100 1000
0001 0101 1101 1001
0011 0111 1111 1011
0010 0110 1110 1010
$\bar{d}$
d
$\bar{d}$
$\bar{b}$
b
$\bar{b}$

✍**Example:**

Given 01 corresponding to $\bar{b}a$ and 11 corresponding to ba, the sum (disjunctive form) would give us:

$$z(a,b)=01+11=\bar{b}a+ba \tag{14.3.32}$$

which is reduced by using the complementation rule directly to:

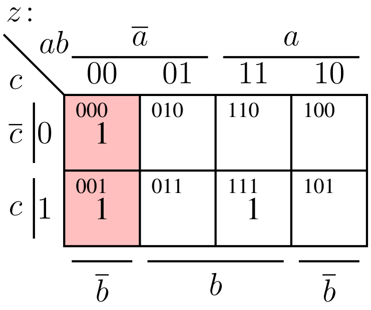
$$z(a,b)=\bar{b}a+ba=(\bar{b}+b)a \tag{14.3.33}$$

hence the advantage to represent them next to each other in a Karnaugh table.

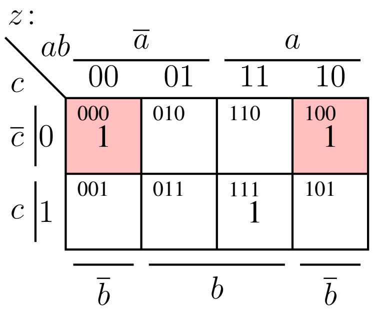
All this to say that when two formulas are found side by side in a Karnaugh map, we retain the similar elements only.

The rules are such that we can reduce therefore when (see previous concrete example):

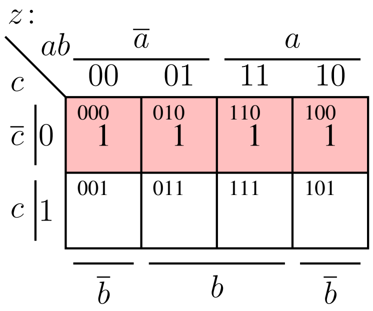
R1. Two 1 are juxtaposed in the table (here the $\bar{c}+c$ will be simplified in the disjunctive form):


z :
ab
$\bar{a}$
a
00 01 11 10
c
$\bar{c}$ 0
c 1
000 1
010
110
100
001 1
011
111 1
101
$\bar{b}$
b
$\bar{b}$

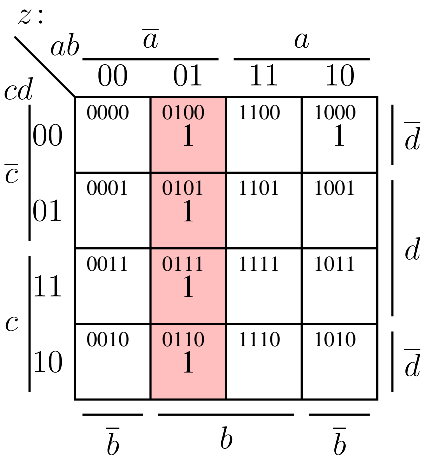
R2. When two 1 are at the extremities of the table (here the $\bar{b}+b$ will be simplified in the disjunctive form:


$z:$
ab
$\overline{a}$
a
00
01
11
10
c
$\overline{c}$ 0
c 1
000
1
010
110
100
1
001
011
111
1
101
$\overline{b}$
b
$\overline{b}$

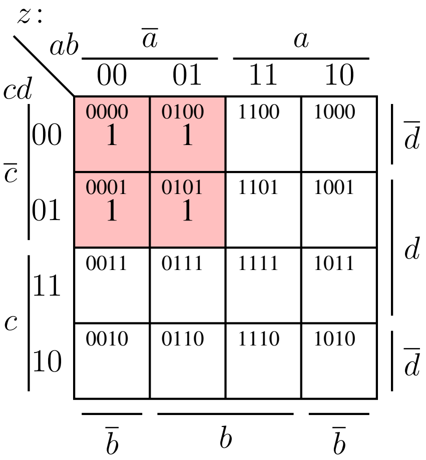
R3. A whole row is full of 1 (in this case the both variables ab disappear from the disjonctive form):


$z:$
ab
$\overline{a}$
a
00
01
11
10
c
$\overline{c}$ 0
c 1
000
1
010
1
110
1
100
1
001
011
111
101
$\overline{b}$
b
$\overline{b}$

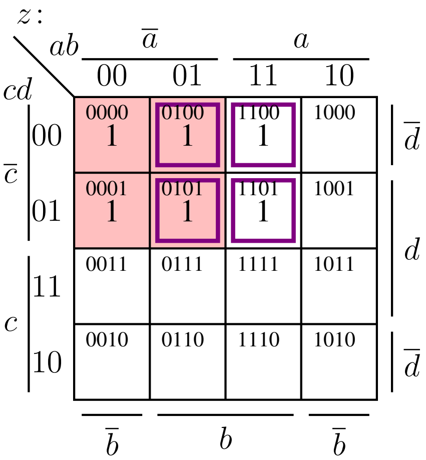
R4. A whole column is full of 1 (in this case the both variables cd disappear from the disjonctive form):


$z:$
ab
$\overline{a}$
a
00
01
11
10
cd
$\overline{c}$
c
00
01
11
10
0000
0100
1
1100
1000
1
0001
0101
1
1101
1001
0011
0111
1
1111
1011
0010
0110
1
1110
1010
$\overline{d}$
d
$\overline{d}$
$\overline{b}$
b
$\overline{b}$

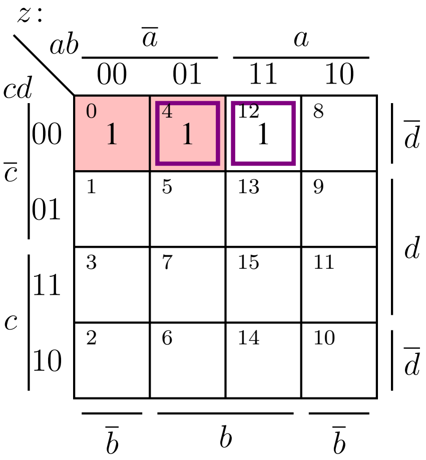
R5. Four adjacents cells are full of 1 (in this case b and d disappear):


z:
ab
cd
$\overline{a}$
a
00
01
11
10
$\overline{c}$
c
$\overline{d}$
d
$\overline{d}$
$\overline{b}$
b
$\overline{b}$
0000 1
0100 1
1100
1000
0001 1
0101 1
1101
1001
0011
0111
1111
1011
0010
0110
1110
1010

R6. The same cells can be used for two reductions:


z:
ab
cd
$\overline{a}$
a
00
01
11
10
$\overline{c}$
c
$\overline{d}$
d
$\overline{d}$
$\overline{b}$
b
$\overline{b}$
0000 1
0100 1
1100 1
1000
0001 1
0101 1
1101 1
1001
0011
0111
1111
1011
0010
0110
1110
1010

R7. The same box can be used for two reductions:


z:
ab
cd
$\overline{a}$
a
00
01
11
10
$\overline{c}$
c
$\overline{d}$
d
$\overline{d}$
$\overline{b}$
b
$\overline{b}$
0 1
4 1
12 1
8
1
5
13
9
3
7
15
11
2
6
14
10

and without errors ... that's all but it's already not bad!

14.3.1.4 Arithmetic Boolean (binary) operations

Using all the elements demonstrated and given previously, we are now able to rigorously determine the logic function allowing Boolean addition and subtraction. Let us also recall that this being done, we can construct multiplication and division using respectively addition and subtraction.

However, we can not with formal digital systems build elements for integration and differentiation. For this we refer the reader to the section of Electrokinetics where it is shown how to use inductors and capacitors to perform such operations with signals.

Remark

We will work on integers but the reader must remember that rational numbers can always be increased in power to be represented in an integral way (it remains to perform the inverse operation if necessary).

The sum of two bytes will be denoted S, the retention C_s (outgoing retention, also often denoted C_{out}) and the deferred retention C_e (inward retention, also often denoted C_{in}).

The truth table will be build with the "tip" that the system inputs (a,b,C_e) take all possible values on 3 bits (three letters) thus $2^3=8$ rows that we have represented in the following table:

C_e	a	b
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

Table 14.47: **Identification of the reported retentions for the binary sum**

And now the idea consists in adding the column constituted by the sum:

$$S=C_e+a+b \tag{14.3.34}$$

row by row (without thinking to the outgoing retention C_s that we will see a little bit further below):

$\boldsymbol{C_e}$	$\boldsymbol{a}$	$\boldsymbol{b}$	$\boldsymbol{S}$
0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	0
1	0	0	1
1	0	1	0
1	1	0	0
1	1	1	1

Table 14.48: **Sum for the binary sum**

Now, row by row, we add the outgoing retention C_s (which is none other than the value that is sent to the incoming retention of the next row) of the sum S:

$\boldsymbol{C_e}$	$\boldsymbol{a}$	$\boldsymbol{b}$	$\boldsymbol{S}$	$\boldsymbol{C_s}$	**Minterms**
0	0	0	0	0	$\bar{C}_e\bar{a}\bar{b}$
0	0	1	1	0	$\bar{C}_e\bar{a}b$
0	1	0	1	0	$\bar{C}_e a\bar{b}$
0	1	1	0	1	$\bar{C}_e ab$
1	0	0	1	0	$C_e\bar{a}\bar{b}$
1	0	1	0	1	$C_e\bar{a}b$
1	1	0	0	1	$C_e a\bar{b}$
1	1	1	1	1	$C_e ab$

Table 14.49: **Retentions reported for the binary sum**

Therefore we have 4 minterms (that is, the terms for which S is non-zero in the rows 2, 3, 5 and 8) such that the normal disjonctive form is written:

$$S=\bar{C}_e\bar{a}b+\bar{C}_e a\bar{b}+C_e\bar{a}\bar{b}+C_e ab \tag{14.3.35}$$

A possible simplification is:

$$S=C_e(\bar{a}\bar{b}+ab)+\bar{C}_e(a\bar{b}+\bar{a}b)=C_e\bar{\alpha}+\bar{C}_e\alpha=C_e\oplus\alpha=C_e\oplus a\oplus b \tag{14.3.36}$$

It also comes for the outgoing retention the following minterms:

$$C_s=\bar{C}_e ab+C_e\bar{a}b+C_e a\bar{b}+C_e ab=ab+C_e(\bar{a}b+a\bar{b})=ab+C_e a\oplus b \tag{14.3.37}$$

So finally we have:

$$\boxed{\begin{gathered}S=C_e\oplus a\oplus b\\ C_s=ab+C_e a\oplus b\end{gathered}} \tag{14.3.38}$$

To build physically with fundamental logic gates this addition, it is useful first to introduce an intermediary logic circuit (but it's not obliged, it's just for pedagogical reasons and by educational tradition!).

Let us for this purpose consider the truth table of the addition without incoming retention, named a "half adder", and represented by the following "technical" drawing:

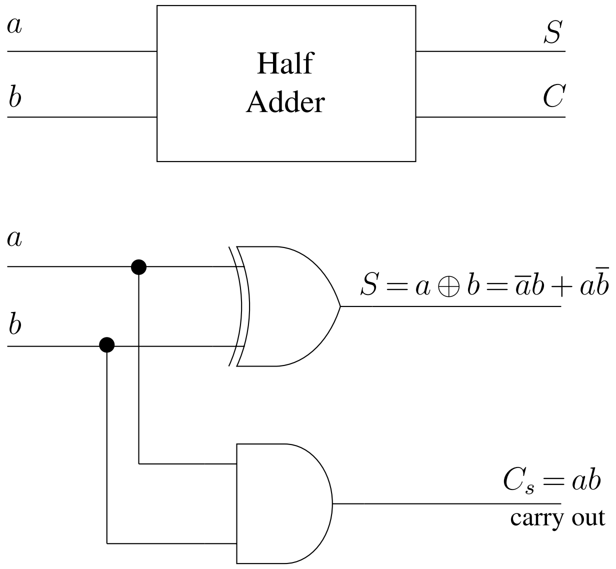

a
b
Half Adder
S
C
a
b
$S = a \oplus b = \overline{a}b + a\overline{b}$
$C_s = ab$
carry out

Figure 14.324: **Half adder logic diagram**

a	b	S	C_s
0	0	0	0
0	1	1	0
1	0	1	0
1	1	0	1

And now we can introduce the "full adder" that adds binary numbers and accounts for values carried in as well as out following the relation proved earlier above and is made of two half-adder. So here is for recall the minterms and we introduce the corresponding logical circuit:

$$\begin{gathered} S = C_e \oplus a \oplus b \\ C_s = ab + C_e a \oplus b \end{gathered}$$

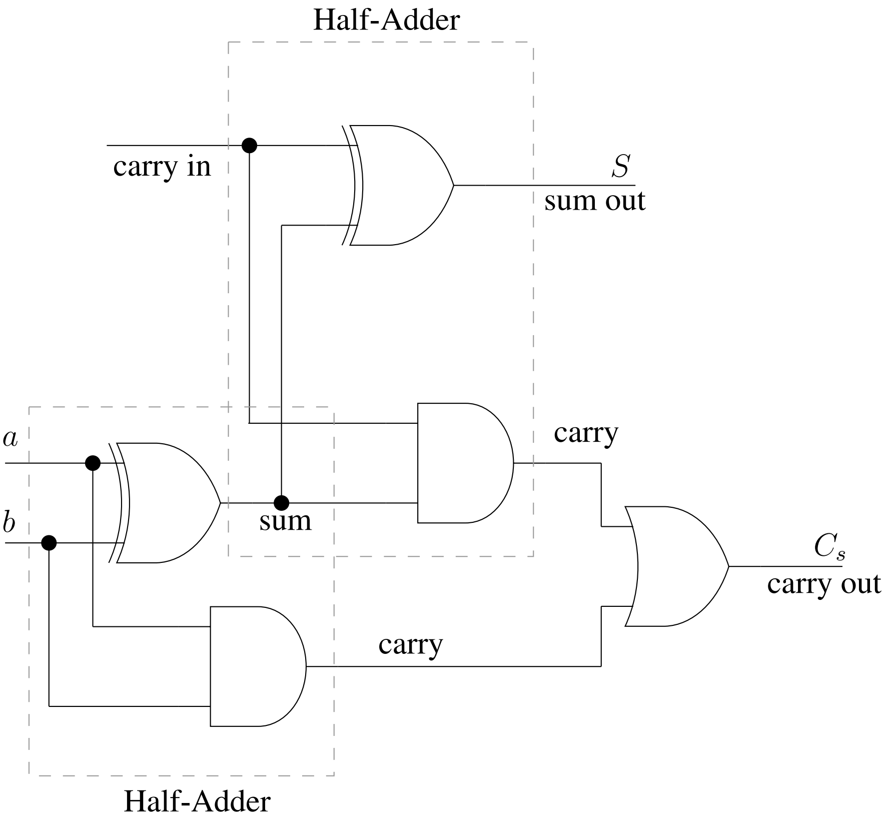

Half-Adder
carry in
S
sum out
a
b
sum
carry
carry
C_s
carry out
Half-Adder

Figure 14.325: **Full adder logic diagram**

C_e	a	b	S	C_s
0	0	0	0	0
0	0	1	1	0
0	1	0	1	0
0	1	1	0	1
1	0	0	1	0
1	0	1	0	1
1	1	0	0	1
1	1	1	1	1

The subtraction (difference) of two bytes will be denoted D, the borrowing B_s (outgoing borrowing, also often denoted B_{out}) and the reported borrowing B_e (inbound borrowing, also often denoted B_{in}). The truth table will first be build as for the addition. That is, the system inputs (a, b, e_e) take all possible values on 3-bit (three-letter) the $2^3 = 8$ rows. Therefore:

B_e	a	b
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

Table 14.50: **Identification of the reported borrowing for the binary subtraction**

But we will a little subtlety. Rather than bore us to calculate $D = a - b - B_e$, we will calculate $D = a + (-b) + (-B_e)$ in the purpose to be able to work with the following truth table:

$-B_e$	a	$-b$
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

Table 14.51: **Inversion of the reported borrowing for the binary subtraction**

and now the idea is to add the difference column $D = a + (-b) + (-B_e)$ row by row (without thinking about the borrowing B_s) which will be strictly identical to the truth table of the sum:

$-B_e$	a	$-b$	D
0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	0
1	0	0	1
1	0	1	0
1	1	0	0
1	1	1	1

Table 14.52: **Reported borrowing for the binary subtraction**

Now, row by row, we add the outgoing borrowing B_s of the difference $D=a+(-b)+(-B_e)$ which written so, then becomes a sum S:

$-B_e$	a	$-b$	$S(D)$	B_s	**minterms**
0	0	0	0	0	$\bar{B}_e\bar{a}\bar{b}$
0	0	1	1	1	$\bar{B}_e\bar{a}b$
0	1	0	1	0	$\bar{B}_e a\bar{b}$
0	1	1	0	0	$\bar{B}_e ab$
1	0	0	1	1	$B_e\bar{a}\bar{b}$
1	0	1	0	1	$B_e\bar{a}b$
1	1	0	0	0	$B_e a\bar{b}$
1	1	1	1	1	$B_e ab$

Table 14.53: **Identification of subtraction minterms**

Therefore it comes 4 minterms (that is, the terms for which $S(D)$ is non-zero at row 2, 3, 5, 8) such that the normal disjonctive form of the subtraction can be written:

$$S(D)=D=\bar{B}_e\bar{a}b+\bar{B}_e a\bar{b}+B_e\bar{a}\bar{b}+B_e ab \tag{14.3.39}$$

A trivial possible simplification is:

$$D=\bar{B}_e(\bar{a}b+a\bar{b})+B_e(\bar{a}\bar{b}+ab)=B_e\oplus a\oplus b \tag{14.3.40}$$

It also comes for the outgoing borrowing the following minterms:

$$B_s=\bar{B}_e\bar{a}b+B_e\bar{a}\bar{b}+B_e\bar{a}b+B_e ab=\bar{a}b+B_e(\overline{a\oplus b}) \tag{14.3.41}$$

So finally:

$$\boxed{\begin{aligned} D&=B_e\oplus a\oplus b\\ B_s&=\bar{a}b+B_e(\overline{a\oplus b})\end{aligned}} \tag{14.3.42}$$

To build physically with fundamental logic gates this subtraction, it is useful first to introduce an intermediary logic circuit (but it's not obliged, it's just for pedagogical reasons and by educational tradition!).

Let us for this purpose consider the truth table of the subtraction without incoming borrowing, named a "half subtracter", and represented by the following "technical" drawing (don't forget that b means in fact $-b$!):

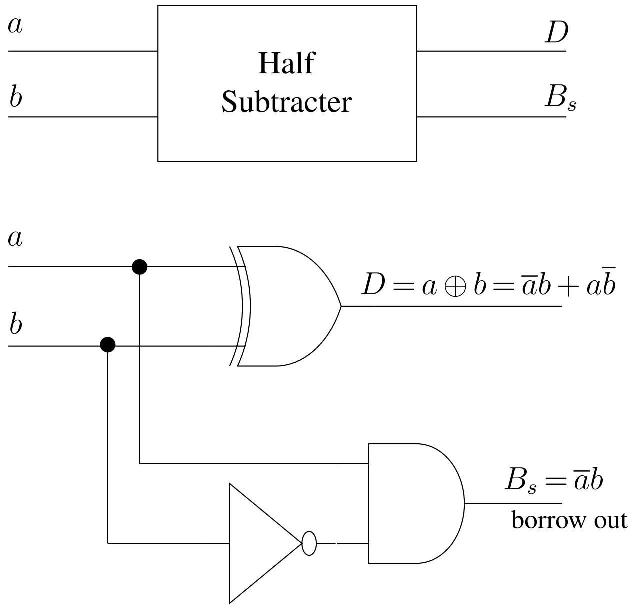

a
b
Half Subtracter
D
B_s
a
b
$D = a \oplus b = \bar{a}b + a\bar{b}$
$B_s = \bar{a}b$
borrow out

Figure 14.326: **Half subtracter logic diagram**

a	$-b$	$S(D)$	B_s
0	0	0	0
0	1	1	1
1	0	1	0
1	1	0	0

And now we can introduce the "full subtracter" that substracts binary numbers and accounts for values carried in as well as out following the relation proved earlier above and is made of two half-adder. So here is for recall the minterms and we introduce the corresponding logical circuit:

$$D = B_e \oplus a \oplus b$$
$$B_s = \bar{a}b + B_e(\overline{a \oplus b})$$

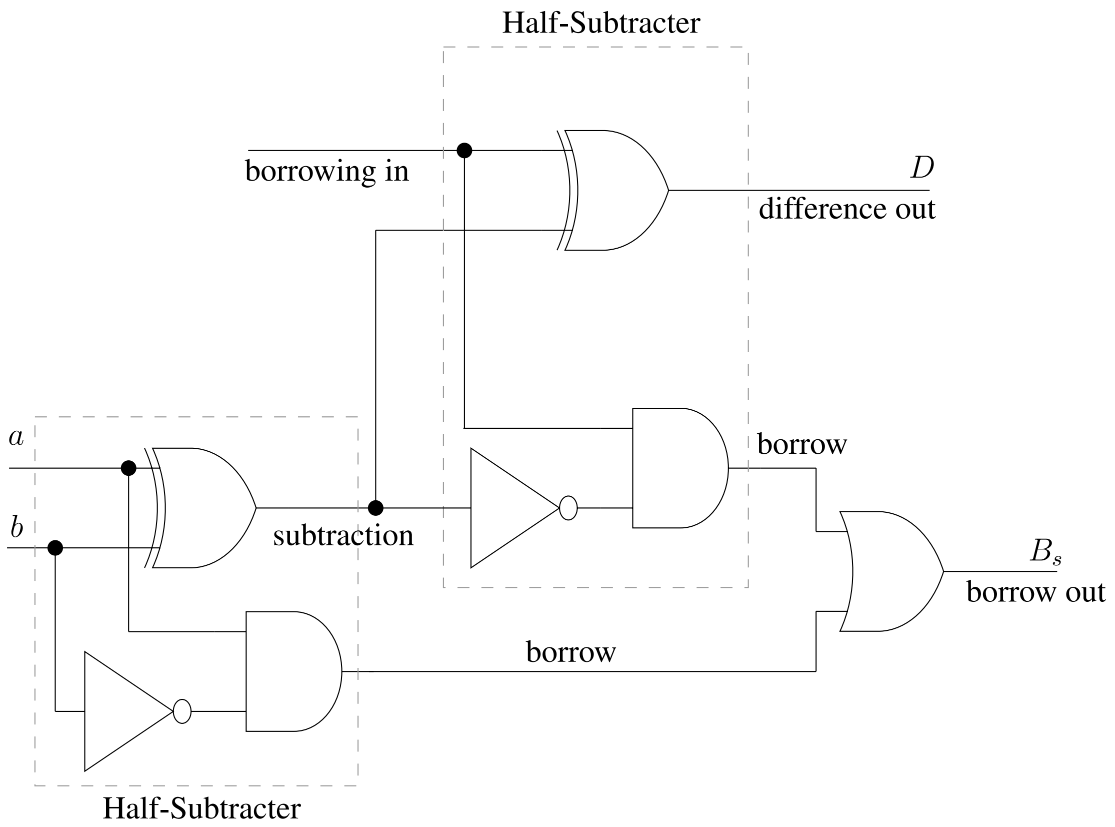

Half-Subtracter
borrowing in
D
difference out
a
b
subtraction
borrow
B_s
borrow out
borrow
Half-Subtracter

Figure 14.327: **Full subtracter logic diagram**

$-B_e$	a	$-b$	$S(D)$	B_s
0	0	0	0	0
0	0	1	1	1
0	1	0	1	0
0	1	1	0	0
1	0	0	1	1
1	0	1	0	1
1	1	0	0	0
1	1	1	1	1

14.3.1.4.1 Prefix codes

"Binary comparators" are logic gates circuit used to compare two binary inputs. There are two types of binary comparators:

- "Equality comparators" are used to check if the two binary inputs (A and B) are equal or not.
- "Magnitude comparators" are used to fully compare two binary inputs A and B and produce three possible outpus if $A > B$, $A == B$ or $A < B$

We will focus in what follows on Magnitude Operators symbolized by:

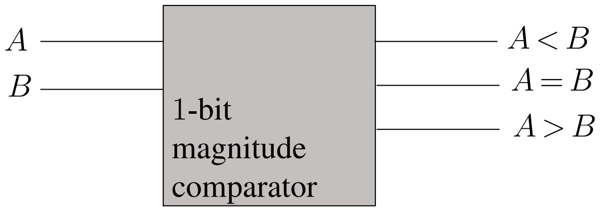


Figure 14.328: 1-bit magnitude comparator

With the following truth table:

INPUT		OUTPUT		
A	B	$A < B$	$A = B$	$A > B$
0	0	0	**1**	0
1	0	0	0	**1**
0	1	**1**	0	0
1	1	0	**1**	0

Table 14.54: 1-bit magnitude comparator truth table

And here is the logic gates diagram for this circuit:

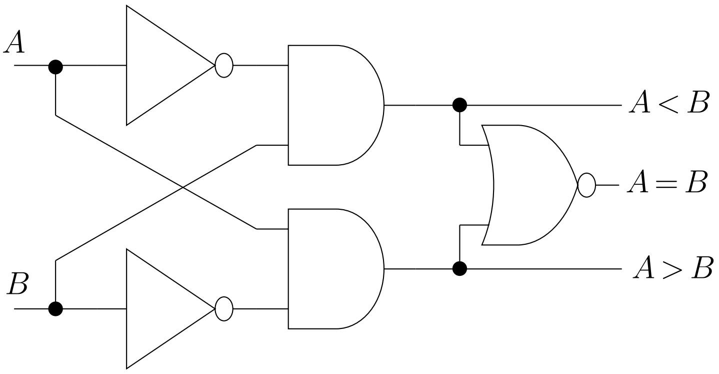


Figure 14.329: 1-bit magnitude comparator

We can now combine several of the above diagrams to create a 4-bit magnitude comparator as follows (thanks to Karnaugh table simplification process):

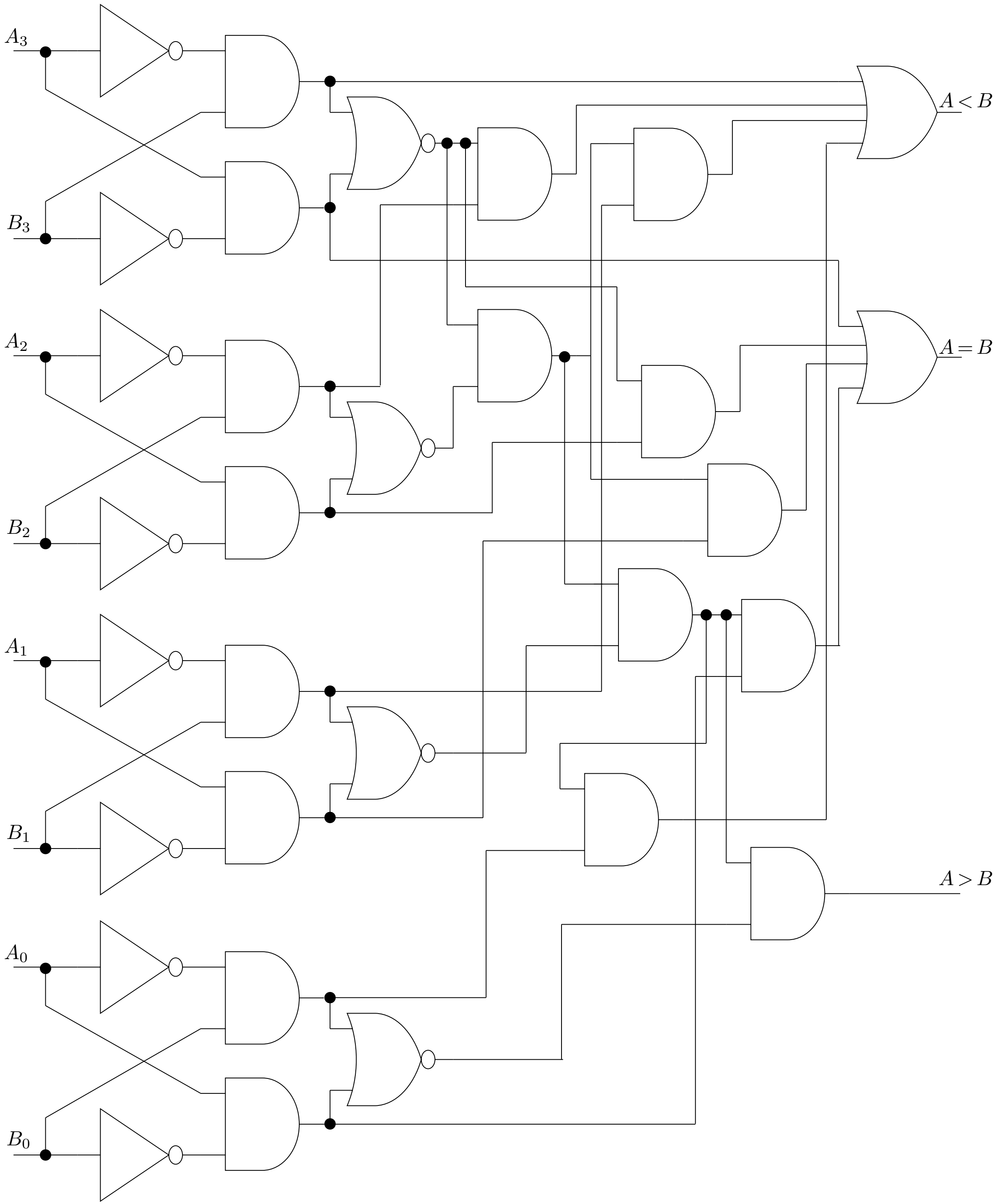


Figure 14.330: 4-bits magnitude comparator

Notice that as far as we know it doesn't seem to exist a simple reducible case to any number of bits. That's why computer systems have a limit depending on their bus width (system that transfers data between components inside a computer) like those in the early 121st century (holocene calendar) have a 32 bits or 64 bits bus and hence a 32 bits or 64 bits magnitude comparator.

The generic method for implementing if-else statement:

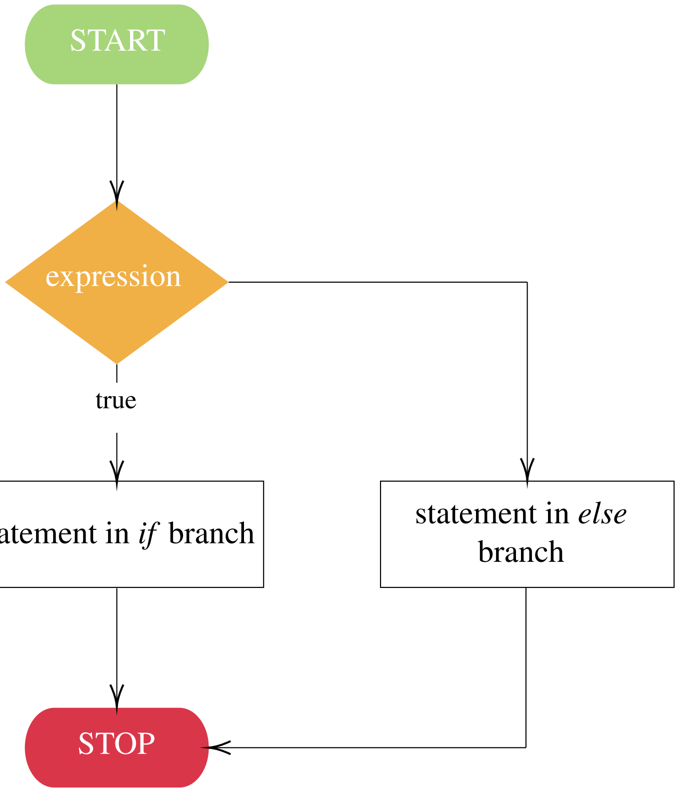

START
expression
true
statement in *if* branch
statement in *else* branch
STOP

Figure 14.331: **if-else statement flowchat**

in hardware (using gates,mux-demux, flip flops) is given in Boolean Algebra by:

$$Q = (A \cdot S) + \left(B \cdot \bar{S}\right) \quad (14.3.43)$$

where:

- S is the input fed by the *if* condition
- A is the input fed by the *then* subexpression
- B is the input fed by the *else* subexpression, and
- Q is the output of the expression.

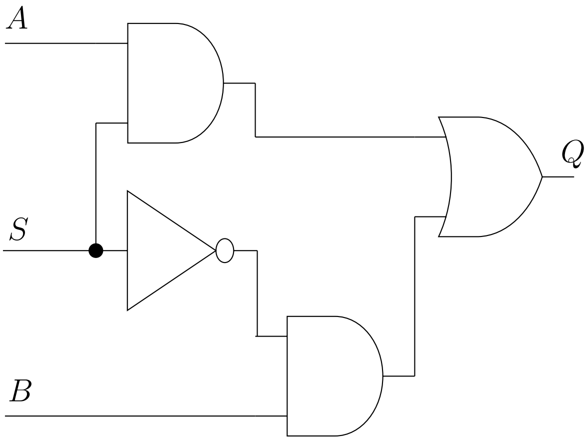


Figure 14.332: **If-else statement in hardware based on boolean gates**

14.3.2 Fuzzy logic

Classical logic only permits conclusions which are either true or false. However, there are also propositions with variable answers, such as one might find when asking a group of people to identify a color. In such instances, the truth appears as the result of reasoning from inexact or partial knowledge in which the sampled answers are mapped on a spectrum.

Humans and animals often operate using fuzzy evaluations in many everyday situations. In the case where someone is tossing an object into a container from a distance, the person does not compute exact values for the object weight, density, distance, direction, container height and width, and air resistance to determine the force and angle to toss the object. Instead the person instinctively applies quick "fuzzy" estimates, based upon previous experience, to determine what output values of force, direction and vertical angle to use to make the toss. Both degrees of truth and probabilities range between 0 and 1 and hence may seem similar at first, but fuzzy logic uses degrees of truth as a mathematical model of vagueness, while probability is a mathematical model of ignorance.

Take, for example, the concepts of "empty" and "full". The meaning of each of them can be represented by a certain fuzzy set. The concept of emptiness would be subjective and thus would depend on the observer or designer. A 100 [ml] glass containing 30 [ml] of water may be defined as being 0.7 empty and 0.3 full, but another designer might, equally well, design a set membership function where the glass would be considered full for all values down to 50 [ml].

Most of the problems encountered are certainly mathematically modelizable. But these models often require overly restrictive assumptions, making application to the real world tricky. The problems of this world must take into account imprecise, uncertain information. Let us take the example of air conditioning: if we want to obtain a cool temperature, we can ask ourselves what temperature range will be appropriate (the demand is imprecise); Furthermore the reliability of the sensors comes into play (the measurement of the ambient temperature is uncertain). We see the difficulty of interpreting the linguistic variables as fresh, hot, ... and the processing of these uncertainties.

An approach was developed from 11965 (holocene calendar) by Loft. A. Zadeh, a professor at the University of California at Berkeley, based on the theory of fuzzy sets, generalizing the theory of classical sets. In the new Zadeh theory, an element can more or less belong to a certain set. Inaccuracies and uncertainties can thus be modeled, and reasonings acquire a flexibility that is not allowed by classical logic: "fuzzy logic" was born. Many applications have developed in various domains, where no deterministic model exists or is practically implementable, as well as in situations where data imprecision makes control by conventional methods impossible.

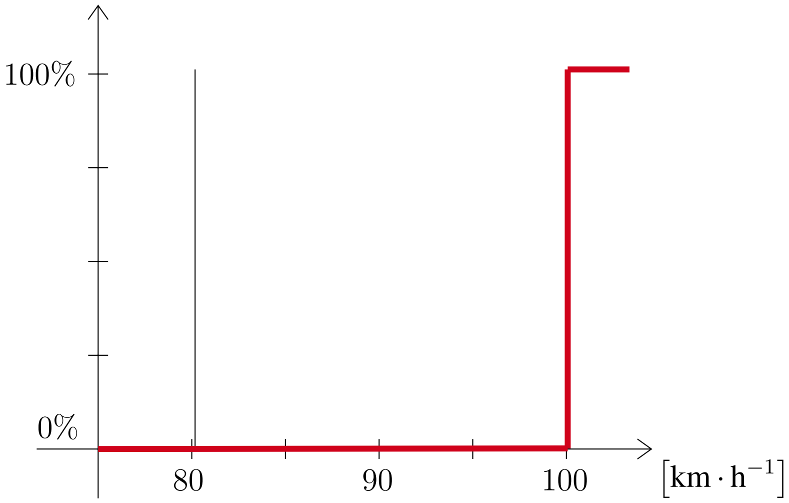
In the following, we will first develop the basics of the theory of fuzzy subsets, then we will clarify the reasoning in fuzzy logic, we will examine the methods of exploitation of the results obtained, and finally we will see an effective application (and if possible an application with MATLAB® and/or `R`).

Before turning to the formal side of the thing (mathematically speaking), it may be preferable (since it is still a technique of the engineer mainly) to briefly present the concepts of fuzzy logic in a pictorial way.

Remark

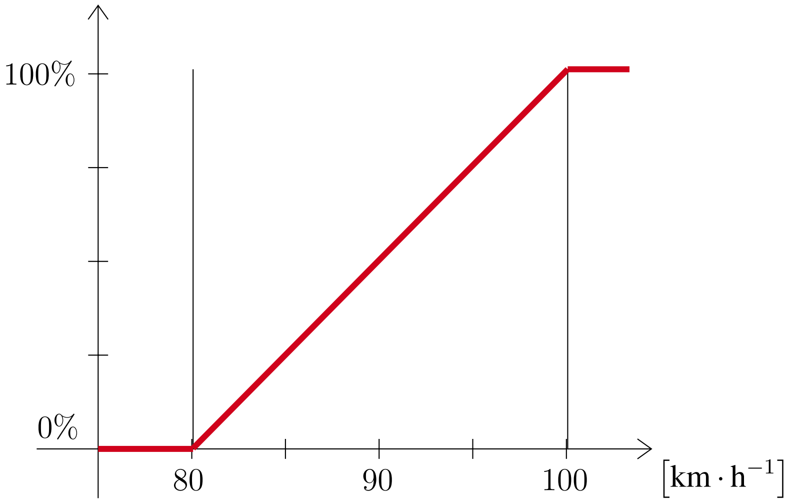
The fuzzy logic is a technique used for real in fields as varied as automatism (ABS brakes), robotics (pattern recognition), road traffic management (red lights), air traffic control, environment (meteorology, climatology, seismology), medicine (diagnostic aid), psychology, data mining, Machine Learning, and many others.

Consider, for example, the speed of a vehicle on a national highway. The normal speed is 90 $[\text{km}\cdot\text{h}^{-1}]$. A speed can be considered as high above 100 $[\text{km}\cdot\text{h}^{-1}]$, and as low below 80 $[\text{km}\cdot\text{h}^{-1}]$. Boolean logic would look at a thing like this:



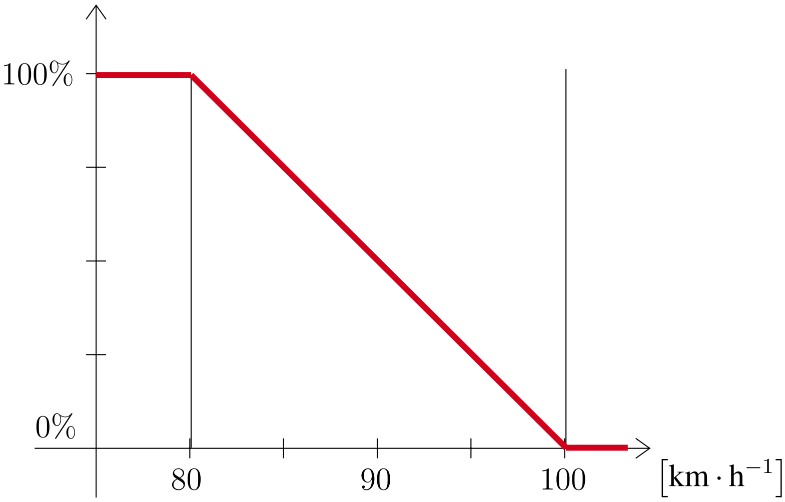
We see above that therefore the speed is considered 100% as high starting from 100 $[\text{km}\cdot\text{h}^{-1}]$, and 0% below.

Fuzzy logic, on the other hand, allows degrees of verification of the condition "Is speed high?" according to:



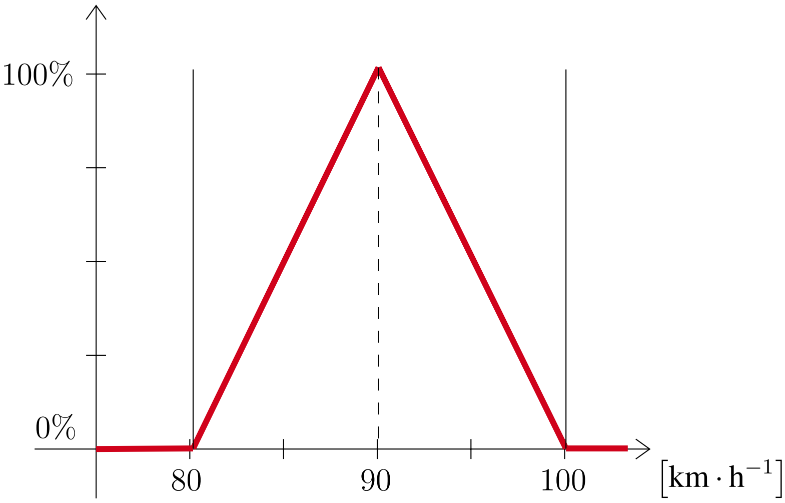
The situation is better here as the speed is considered as not high at all below 80 $[\text{km}\cdot\text{h}^{-1}]$. We can therefore say that below 80 $[\text{km}\cdot\text{h}^{-1}]$, the speed is high at 0%. The speed is considered to be high above $100\,[\text{km}\cdot\text{h}^{-1}]$. The speed is therefore high at 100% above $100\,[\text{km}\cdot\text{h}^{-1}]$. The speed is thus high at 50% when at $90\,[\text{km}\cdot\text{h}^{-1}]$ and high at 25% at $85\,[\text{km}\cdot\text{h}^{-1}]$.

Similarly, the function "Is the speed low?" Will be addressed typically as following by most humans:



Therefore following the above figure the speed is considered low under $80\,[\text{km}\cdot\text{h}^{-1}]$. It is therefore 100% low. The speed is considered not at all low above $100\,[\text{km}\cdot\text{h}^{-1}]$. It is therefore 0% low. The speed is thus a 50% (bit low) when at $90\,[\text{km}\cdot\text{h}^{-1}]$ and at 75% (quite low) when at $85\,[\text{km}\cdot\text{h}^{-1}]$.

We can also define a function "Is the speed average?" by:

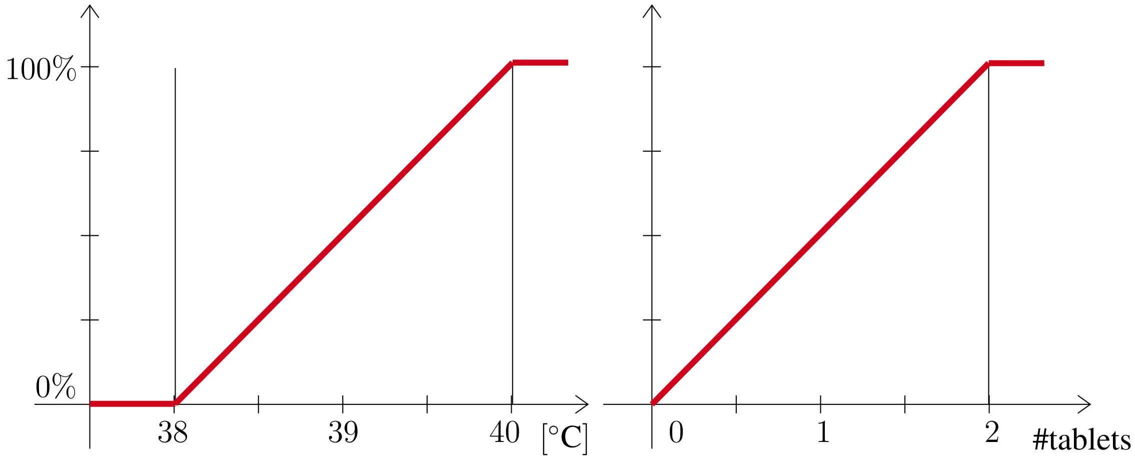


Once the input value evaluated ("Is the speed high?"), a value can be determined for an output function. Consider the function "If the fever is strong, then administer aspirin". Such a function is named "fuzzy control". It is composed of two parts:

1. An input: "Is the fever strong?". We consider that a fever is not strong below $38\,[^\circ\text{C}]$, and that it is high above $40\,[^\circ\text{C}]$.

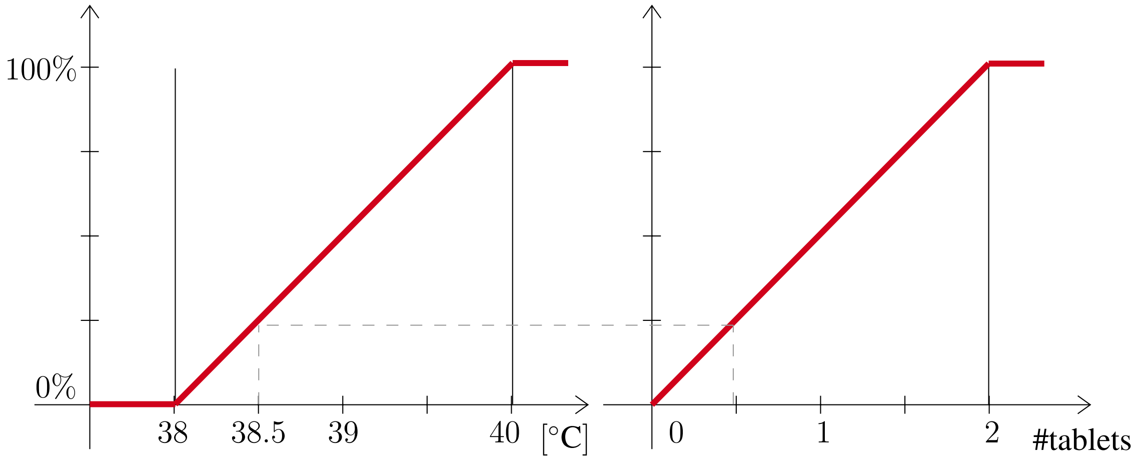
2. An output: "Administer a given number of aspirin tablets!"

These two parts are related. We can represent them together as below:

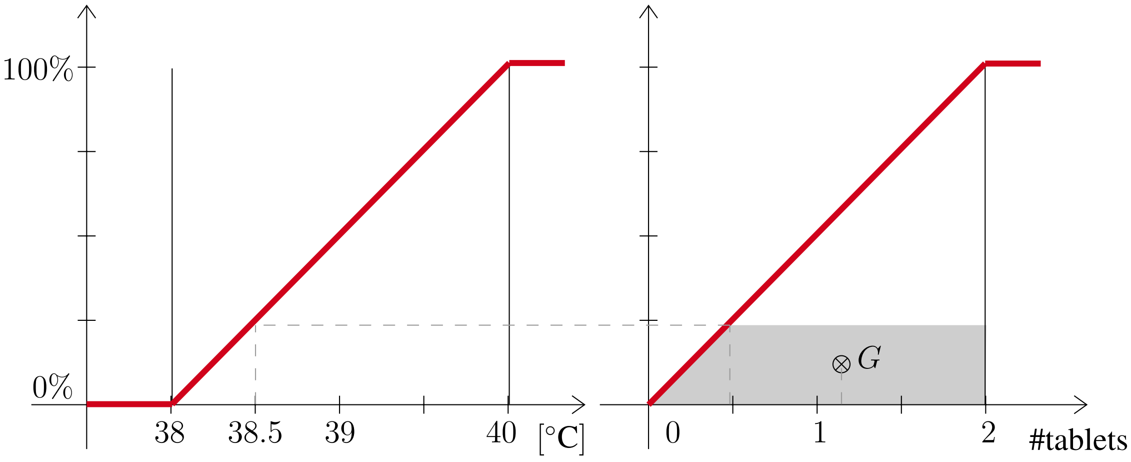


There are several empirical techniques for determining the output value (in the example: the number of aspirin tablets to be administered):

An example consists in taking the horizontal passing through the corresponding ordinate point on the starting curve at the abscissa of the value of the input and of looking at where this horizontal section intersects the output curve. The abscissa of this point of intersection is a possible output value as shown below:



A second empirical possible choice consists in taking as output value of the center of gravity of the gray trapezoid delimited by the horizontal and the output curve as shown in the figure below:



From these two non-software oriented examples, we see that we are at the frontier of pure science and engineering frontier since there is a technical and/or statistical choice to be made in the method to be chosen.

Let us now do a computer aided example as engineers in practice use such tools to speed up their research and development! The case below in inspired by a textbook and many softwares take it now as basis example named the "Basic Tipping Problem".

✍**Example:**

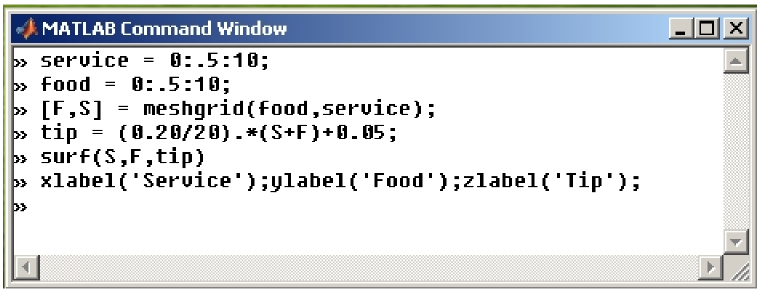
Given a number between 0 and 10 that represents the quality of service and food at a restaurant (where 10 is excellent), what should the tip be?

This problem is based on tipping as it is typically practiced in the United States. An average tip for a meal in the U.S. is 15%, though the actual amount may vary depending on the quality of the service provided. But because service and food is rated on a scale of 0 to 10, we might have the tip go linearly from 5% if the service is bad to 25% if the service is excellent.

If we denote S for the service and F for food quality, we will chose the tip as being given by:

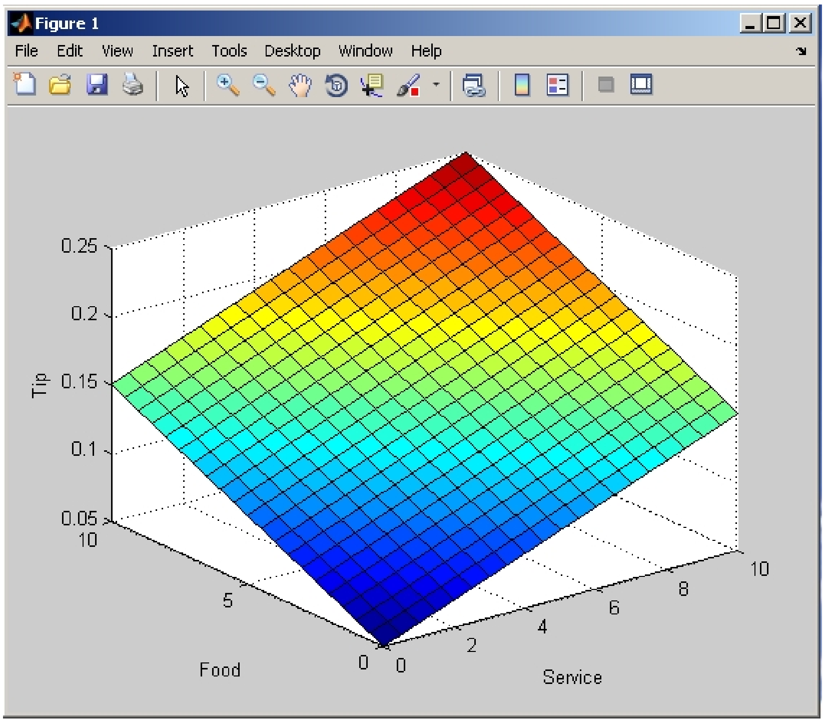
$$\text{tip} = \frac{1}{100}(S+F)+0.05 \tag{14.3.44}$$

Here is a copy/paste of the MATLAB® 2013a script for the original url[a]:



```
>> service = 0:.5:10;
>> food = 0:.5:10;
>> [F,S] = meshgrid(food,service);
>> tip = (0.20/20).*(S+F)+0.05;
>> surf(S,F,tip)
>> xlabel('Service');ylabel('Food');zlabel('Tip');
>>
```

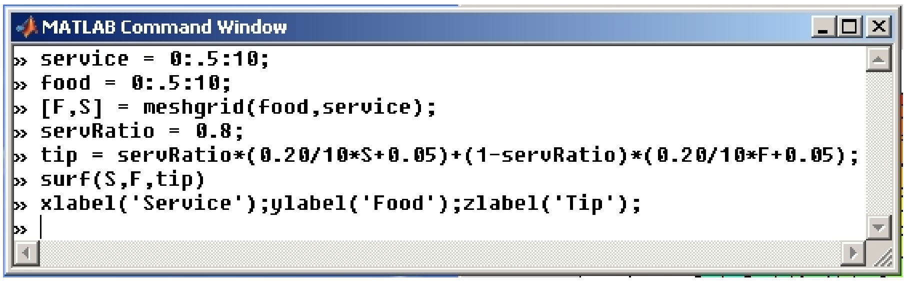
In this case, the results look satisfactory, but when you look at them closely, they do not seem quite right:



Suppose you want the service to be a more important factor than the food quality. Specify that service accounts for 80% of the overall tipping grade and the food makes up the other 20%. Try this equation:

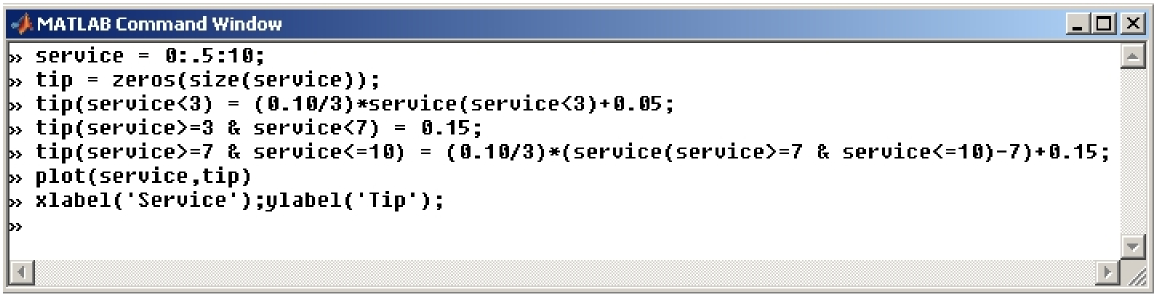
$$\text{tip} = 0.8\left(\frac{0.2}{10}S + 0.05\right) + (1-0.8)\left(\frac{0.2}{10}F + 0.05\right) \tag{14.3.45}$$

Thus in MATLAB® 2013a:



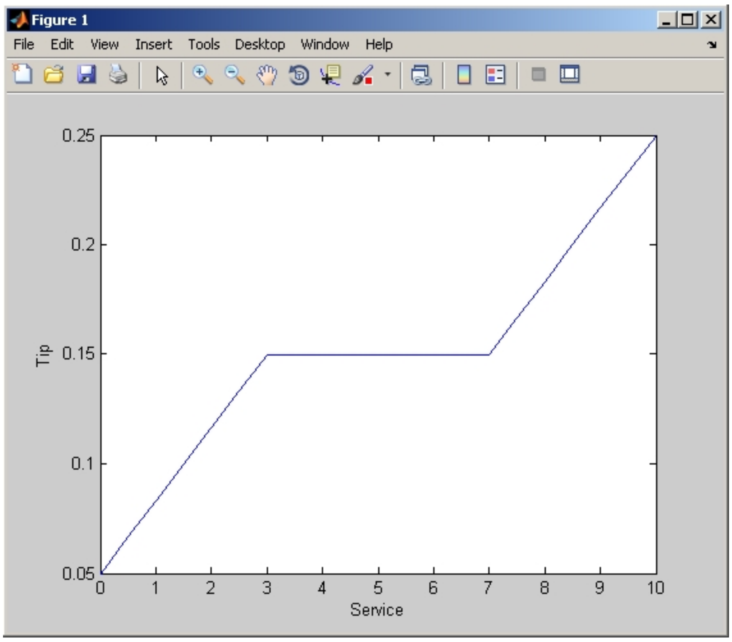
```
>> service = 0:.5:10;
>> food = 0:.5:10;
>> [F,S] = meshgrid(food,service);
>> servRatio = 0.8;
>> tip = servRatio*(0.20/10*S+0.05)+(1-servRatio)*(0.20/10*F+0.05);
>> surf(S,F,tip)
>> xlabel('Service');ylabel('Food');zlabel('Tip');
>> |
```

The response is still some how too uniformly linear. Suppose you want more of a flat response in the middle, i.e., you want to give a 15% tip in general, but want to also specify a variation if the service is exceptionally good or bad. This factor, in turn, means that the previous linear mappings no longer apply. You can still use the linear calculation with a piecewise linear construction. Now, return to the one-dimensional problem of just considering the service. You can create a simple conditional tip assignment using logical indexing:

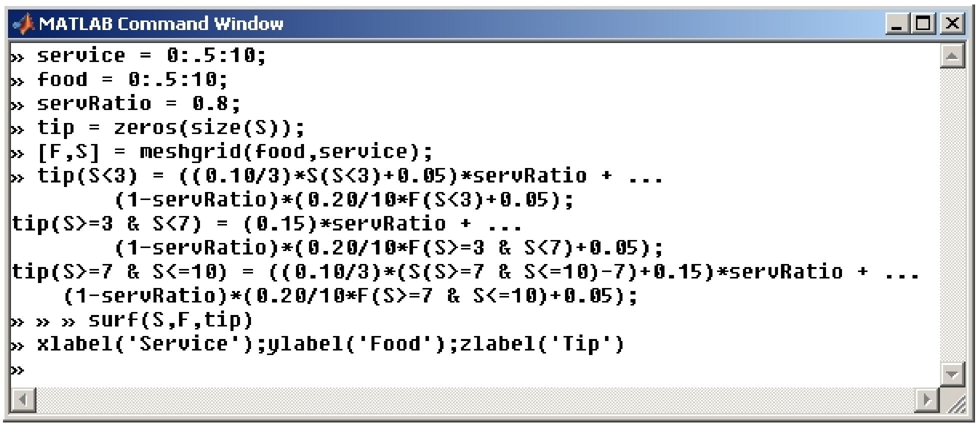


```
» service = 0:.5:10;
» tip = zeros(size(service));
» tip(service<3) = (0.10/3)*service(service<3)+0.05;
» tip(service>=3 & service<7) = 0.15;
» tip(service>=7 & service<=10) = (0.10/3)*(service(service>=7 & service<=10)-7)+0.15;
» plot(service,tip)
» xlabel('Service');ylabel('Tip');
»
```

that gives the following plot:

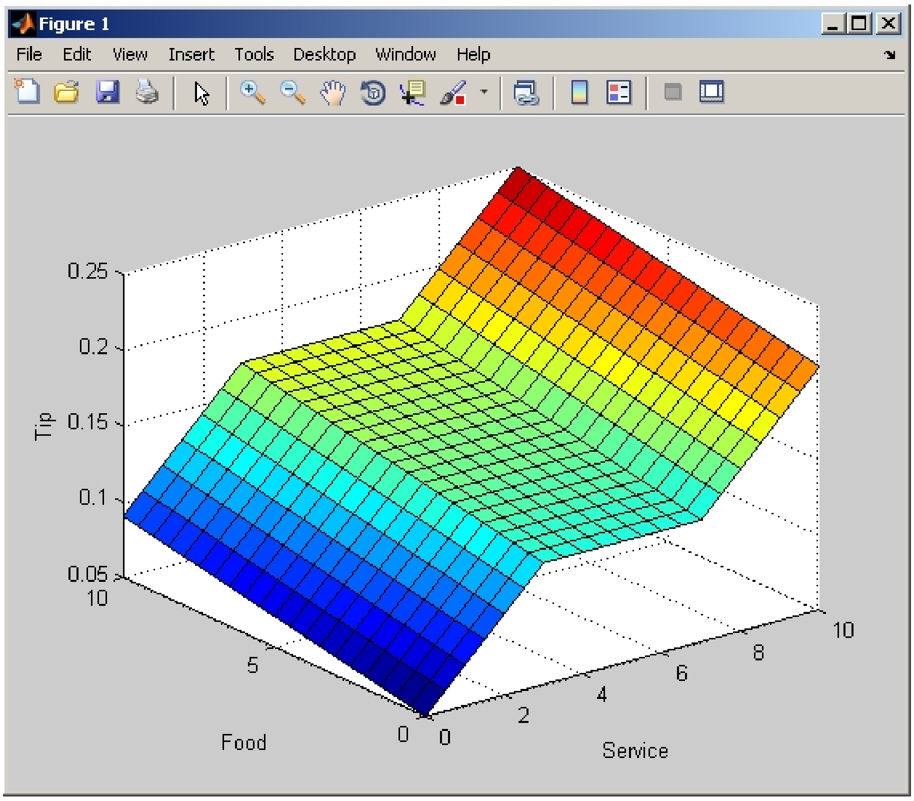


Suppose you extend this to two dimensions, where we take food into account again:



```
» service = 0:.5:10;
» food = 0:.5:10;
» servRatio = 0.8;
» tip = zeros(size(S));
» [F,S] = meshgrid(food,service);
» tip(S<3) = ((0.10/3)*S(S<3)+0.05)*servRatio + ...
        (1-servRatio)*(0.20/10*F(S<3)+0.05);
tip(S>=3 & S<7) = (0.15)*servRatio + ...
        (1-servRatio)*(0.20/10*F(S>=3 & S<7)+0.05);
tip(S>=7 & S<=10) = ((0.10/3)*(S(S>=7 & S<=10)-7)+0.15)*servRatio + ...
    (1-servRatio)*(0.20/10*F(S>=7 & S<=10)+0.05);
» » » surf(S,F,tip)
» xlabel('Service');ylabel('Food');zlabel('Tip')
»
```

that gives the following plot:



The plot looks good, but the function is surprisingly complicated. It was a little difficult to code this correctly, and it is definitely not easy to modify this code in the future. Moreover, it is even less apparent how the algorithm works to someone who did not see the original design process.

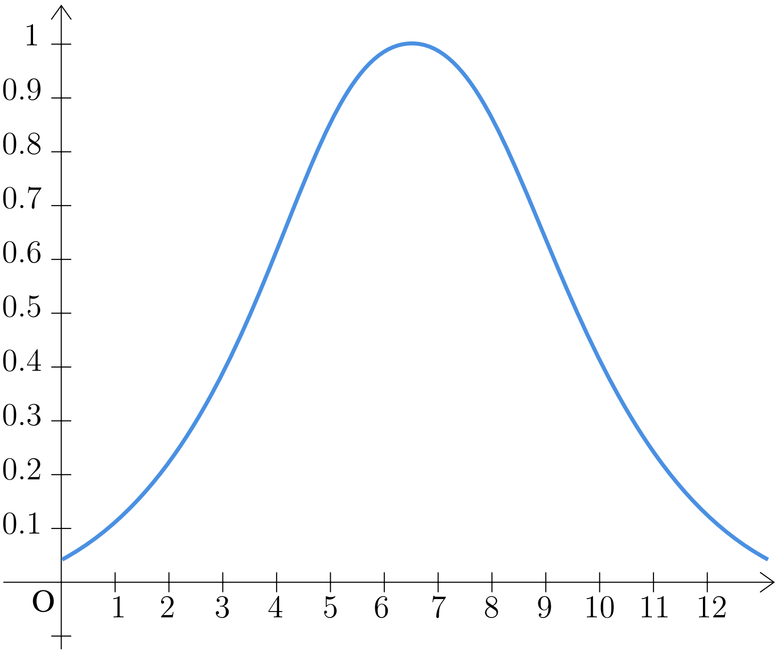
---

[a] See https://ch.mathworks.com/help/fuzzy/an-introductory-example-fuzzy-versus-nonfuzzy-logic.html

In practice we don't use a fuzzy function like the one above as there are sharp angles (imagine a car having speed settings wit abrupt changes as above...).

Therefore rather than using linear functions a input we can choose among a class of continuous and smooth functions. For example the "Gaussian curve membership function" defined by:

$$\text{gaussmf}(x,\sigma,c)=e^{-\frac{(x-c)^2}{2\sigma^2}} \tag{14.3.46}$$

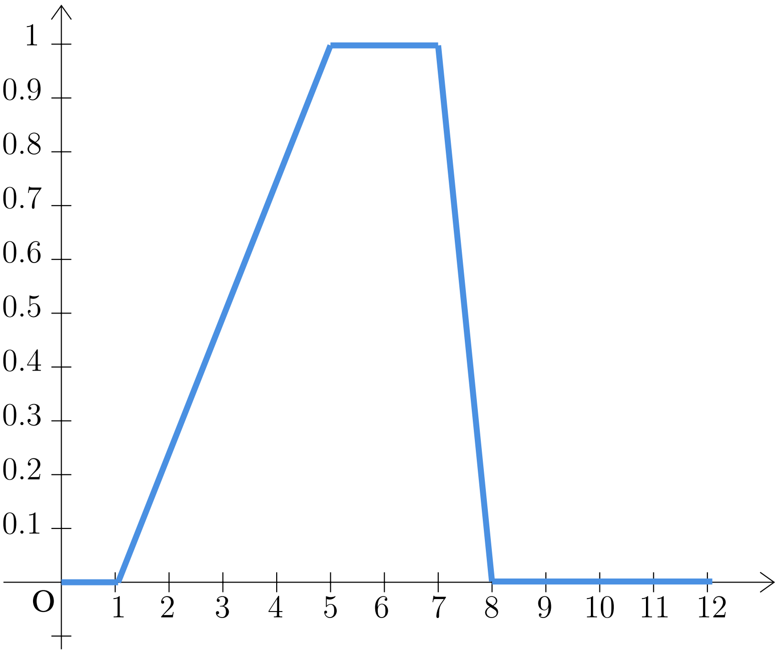


gaussmf, P=[2,6.5]

or the "Trapezoidal-shaped membership function" defined by:

$$\text{trimf}(x,a,b,c,d)=\begin{cases}0 & x\leq a\\ \dfrac{x-a}{b-a} & a\leq x\leq b\\ 1 & b\leq x\leq c\\ \dfrac{d-x}{d-c} & c\leq x\leq d\\ 0 & d\leq x\end{cases}=\max\left(\min\left(\frac{x-a}{b-a},1,\frac{d-x}{d-c}\right),0\right) \tag{14.3.47}$$

The parameters a and d locate the "feet" of the trapezoid and the parameters b and c locate the "shoulders."

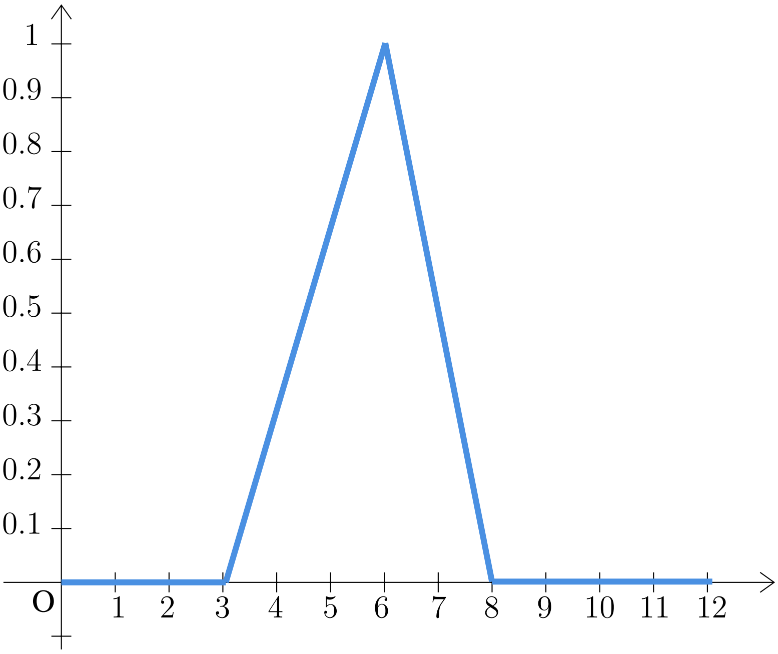


trapmf, P=[1,5,7,8]

Or also the "Triangular-shaped membership function" defined by:

$$\text{trimf}(x,a,b,c)=\begin{cases}0 & x\leq 0\\ \dfrac{x-a}{b-a} & a\leq x\leq b\\ \dfrac{c-x}{c-b} & b\leq x\leq c\\ 0 & c\leq x\end{cases}=\max\left(\min\left(\frac{x-a}{b-a},\frac{c-x}{c-b}\right),0\right) \tag{14.3.48}$$

The parameters a and c locate the "feet" of the triangle and the parameter b locates the peak:

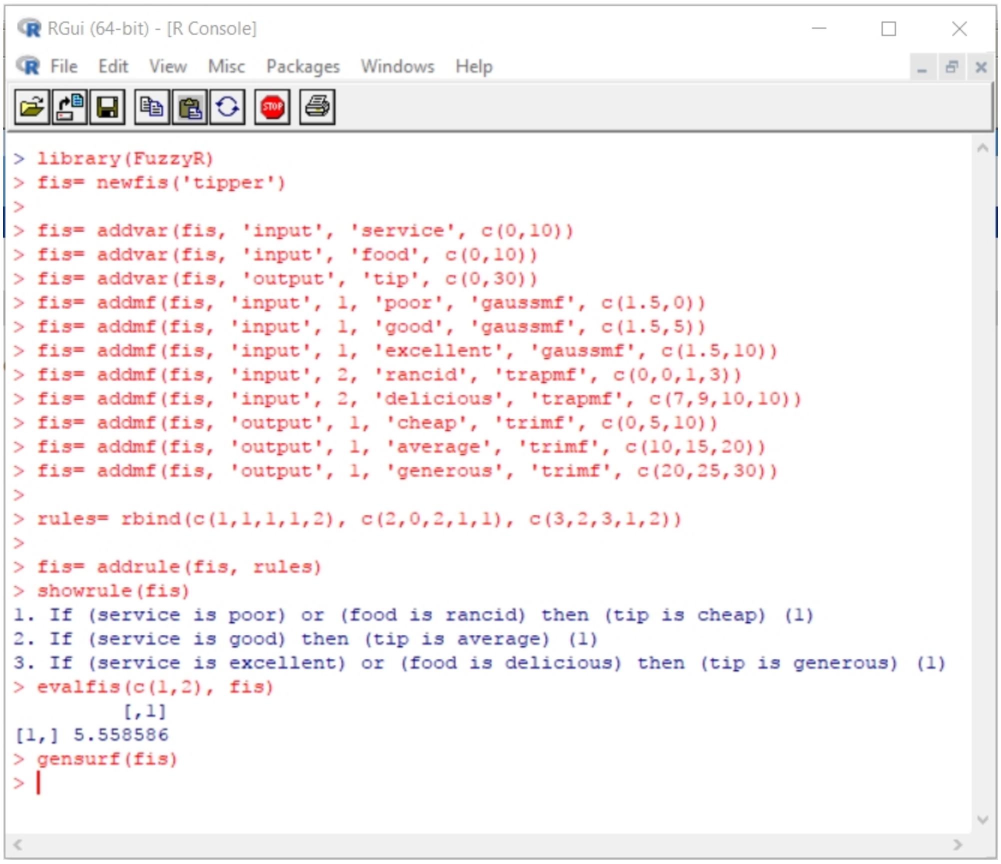


trimf, P=[3,6,8]

Let us see with R the tipper problem managed with the fuzzy functions above (for more details see the R companion book!).

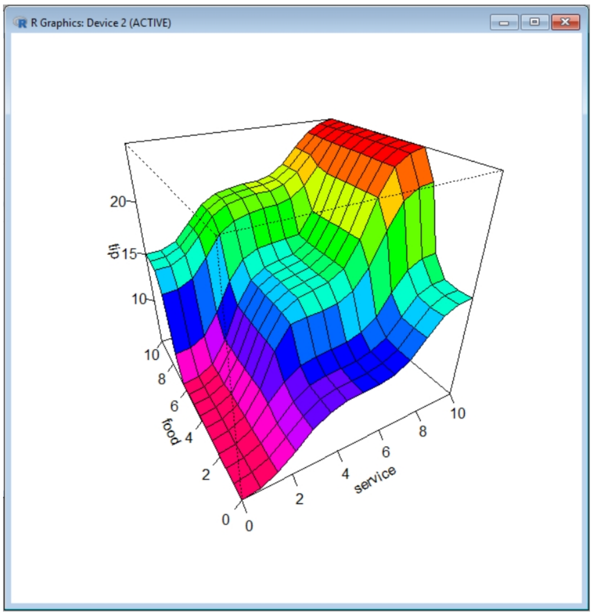
✍**Example:**

Here is a copy/paste of the R script given in [564] with detailed explanations:



```
> library(FuzzyR)
> fis= newfis('tipper')
>
> fis= addvar(fis, 'input', 'service', c(0,10))
> fis= addvar(fis, 'input', 'food', c(0,10))
> fis= addvar(fis, 'output', 'tip', c(0,30))
> fis= addmf(fis, 'input', 1, 'poor', 'gaussmf', c(1.5,0))
> fis= addmf(fis, 'input', 1, 'good', 'gaussmf', c(1.5,5))
> fis= addmf(fis, 'input', 1, 'excellent', 'gaussmf', c(1.5,10))
> fis= addmf(fis, 'input', 2, 'rancid', 'trapmf', c(0,0,1,3))
> fis= addmf(fis, 'input', 2, 'delicious', 'trapmf', c(7,9,10,10))
> fis= addmf(fis, 'output', 1, 'cheap', 'trimf', c(0,5,10))
> fis= addmf(fis, 'output', 1, 'average', 'trimf', c(10,15,20))
> fis= addmf(fis, 'output', 1, 'generous', 'trimf', c(20,25,30))
>
> rules= rbind(c(1,1,1,1,2), c(2,0,2,1,1), c(3,2,3,1,2))
>
> fis= addrule(fis, rules)
> showrule(fis)
1. If (service is poor) or (food is rancid) then (tip is cheap) (1)
2. If (service is good) then (tip is average) (1)
3. If (service is excellent) or (food is delicious) then (tip is generous) (1)
> evalfis(c(1,2), fis)
          [,1]
[1,] 5.558586
> gensurf(fis)
>
```

That gives:



14.3.2.1 Fuzzy set

Definition (#14.3.593): Given X a set. A "fuzzy subset" A of X is defined by a belonging function f_A on X with values in the interval $[0, 1]$.

Remark

The belonging function f_A can be set arbitrarily. A practical application problem is for the engineer to define these functions (we usually use statistical data or the opinion of an expert to make the least worst choice...).

The notion of fuzzy subset encompasses that of classical subset for which f_A is the indicator function given for recall by:

Definition (#14.3.594): If A and B are two sets, such that A is included in B (i.e. $A \subset B$), we name "indicator function" of A (relatively to B), the function 1_A defined in $\{0, 1\}$, and such that:

$$\begin{aligned} 1_A(x) &= 1 \quad \text{if } x \text{ is in } A \\ 1_A(x) &= 0 \quad \text{if } x \text{ is not in } A \end{aligned} \tag{14.3.49}$$

This is for a classical set. Obviously such indicator functions are often very practical technical intermediaries!

But for fuzzy logic we have scnerio such like this:

✍**Example:**

A possible characteristic function to define the fuzzy subset A "to be twenty years" on the set X of the real positive numbers:

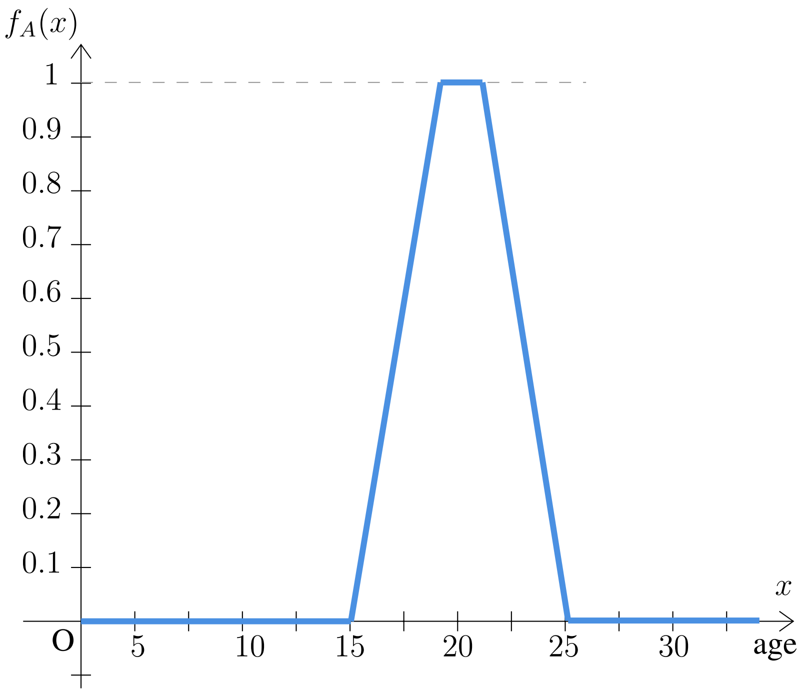


Figure 14.333: **Centered linear fuzzy function**

The following concepts are characteristic of A in the field of fuzzy sets:

Definitions (#14.3.595):

D1. Support of A:

$$\text{supp}(A)=\{x\in X, f_A(x)\neq 0\} \tag{14.3.50}$$

D2. Height of A:

$$h(A)=\sup_{x\in X} f_A(x) \tag{14.3.51}$$

D3. A is said to be normalized if $h(A)=1$

✍**Example:**

A possible characteristic function to define the fuzzy subset A "to be twenty years" on the set X of the real positive numbers:

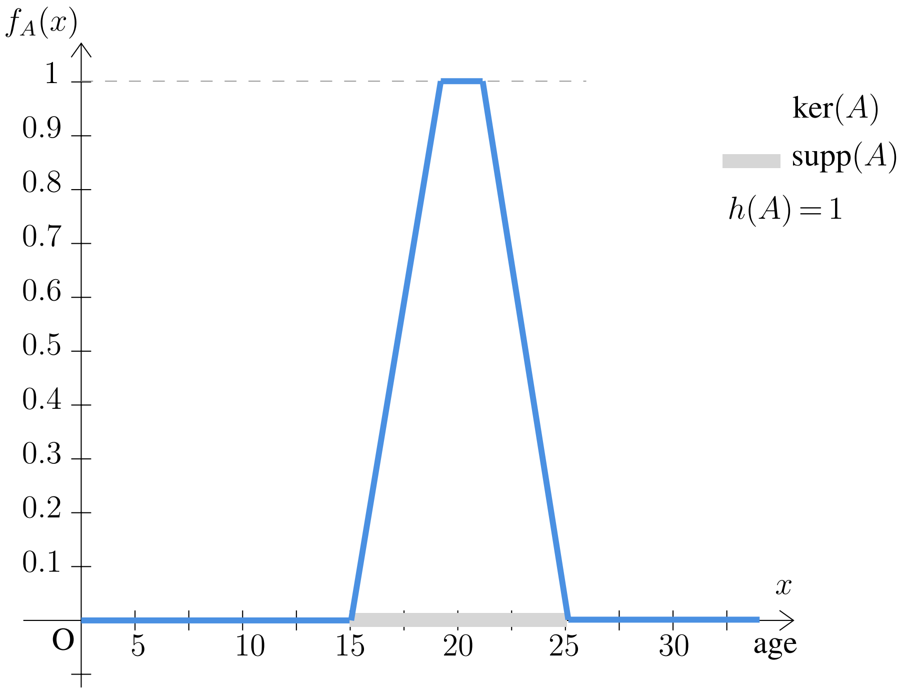


Figure 14.334: **Centered linear fuzzy function**

Remark

The fuzzy subsets considered will all be assumed normalized, in extenso of height equal to 1.

D4. The kernel of A:

$$\text{ker}(A)=\{x\in X, f_A(x)=1\} \tag{14.3.52}$$

D5. Cardinality of A:

$$\text{Card}(A)=\sum_{n\in X} f_A(x) \tag{14.3.53}$$

D6. A is "more specific" than B if:

$$\text{ker}(A)\not\subseteq\text{ker}(B) \qquad \text{and} \qquad \text{supp}(A)\subseteq\text{supp}(B) \tag{14.3.54}$$

D7. A is "more precise" than B if:

$$\text{ker}(A)=\text{ker}(B) \qquad \text{and} \qquad \text{supp}(A)\subseteq\text{supp}(B) \tag{14.3.55}$$

D8. There is equality between two fuzzy subsets if and only if:

D9. There is equality between two fuzzy subsets if and only if:

$$\forall x \in X: \quad f_A(x) = f_B(x) \tag{14.3.56}$$

D10. There is inclusion between two fuzzy subsets if and only if:

$$\forall x \in X: \quad f_A(x) \leq f_B(x) \tag{14.3.57}$$

D11. The intersection $A \cap B$ is defined by:

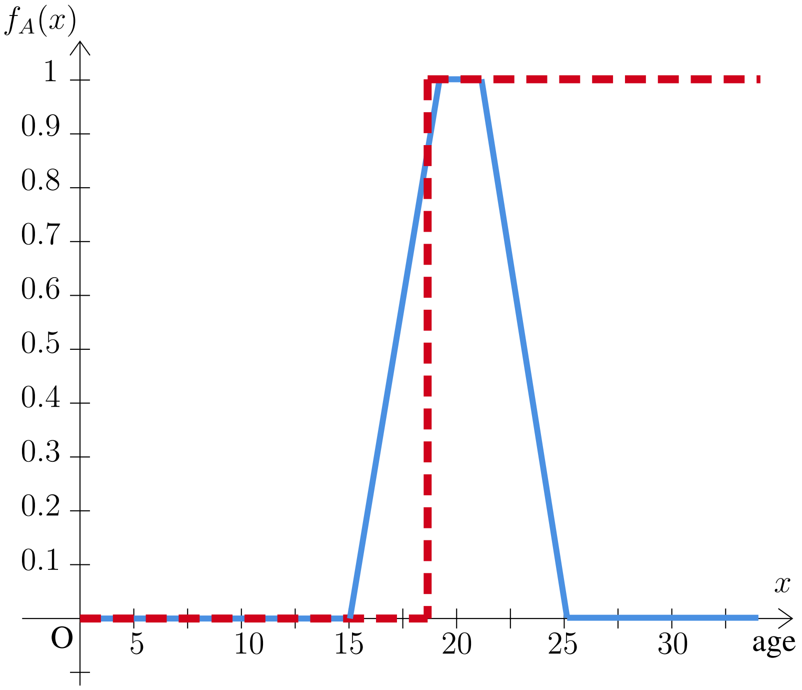
$$f_{A \cap B} : x \rightarrow \min(f_A(x), f_B(x)) \tag{14.3.58}$$

D12. The union $A \cup B$ is defined by:

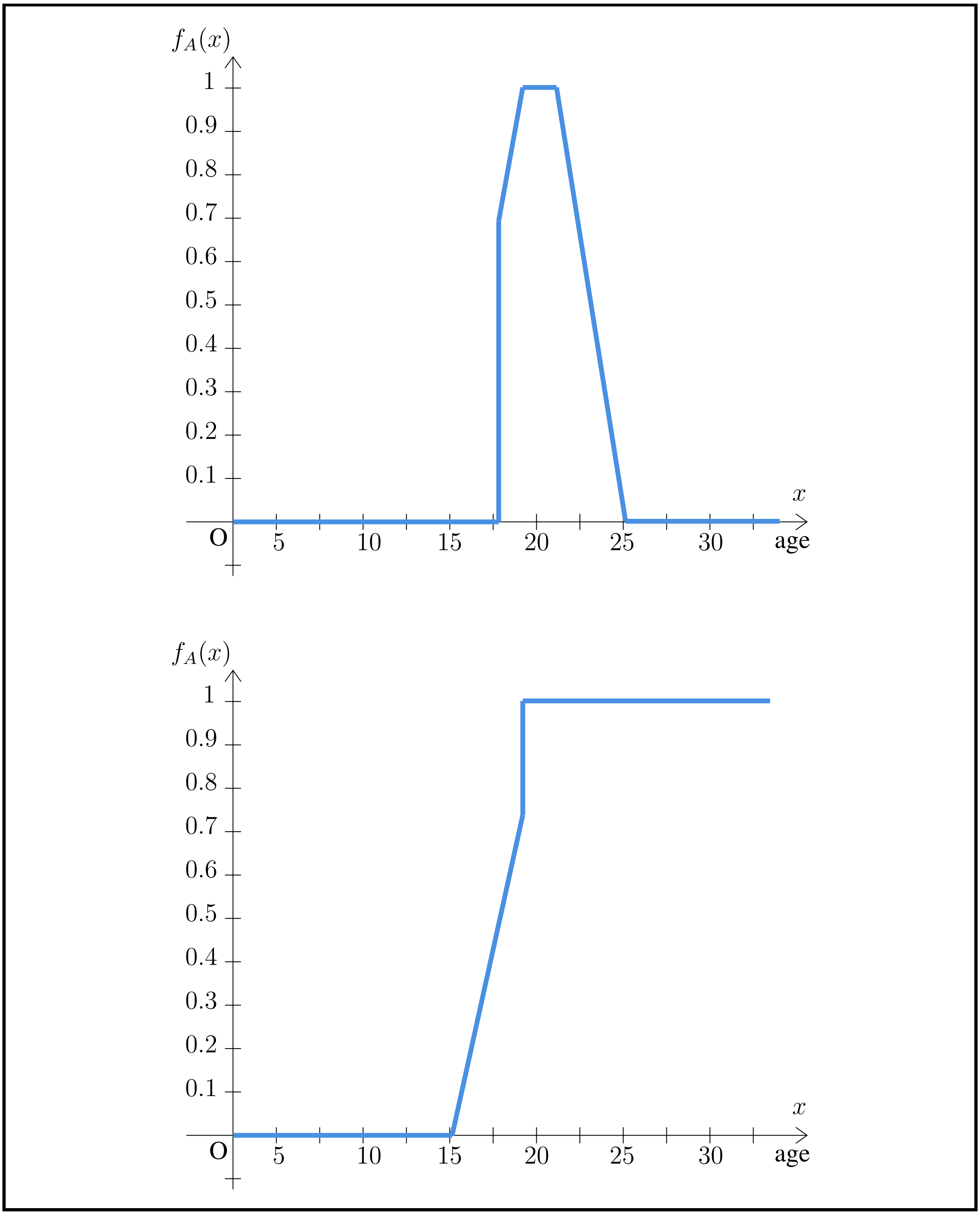
$$f_{A \cup B} : x \rightarrow \max(f_A(x), f_B(x)) \tag{14.3.59}$$

✍**Example:**

Let us return to the case already envisaged. We consider the "twenty-year-old" people and those with "being in age" (i.e., those who are in age to vote/drink alcohol) (dotted in the figure: we consider it as a non-fuzzy sub-set!):



According on the definitions of the intersection ("logical AND" or logical multiplication according to the Boolean Algebra) and the union ("logical OR" or logical addition according to Boolean Algebra), we can characterize the subsets (first figure below), as well as those "being in their twenties or being in age" (second figure below):



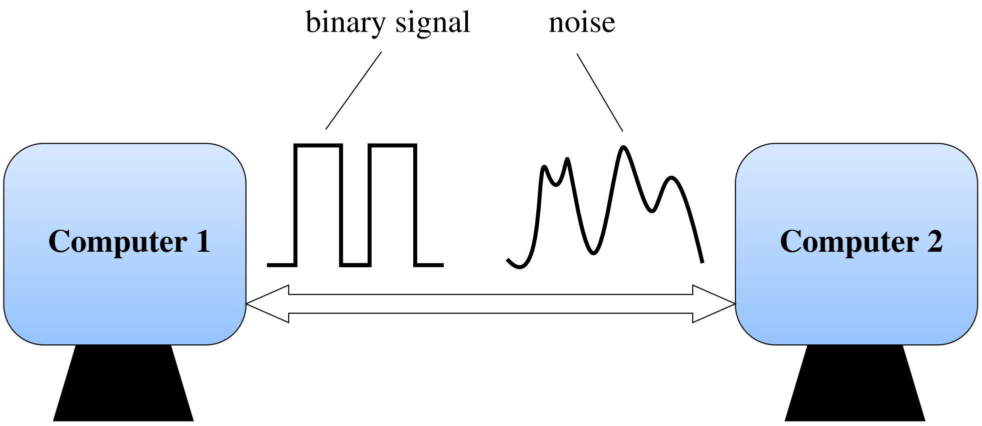
~80 %

★★★☆☆

10 votes, 66.00%

14.4. Error-Correcting Codes

If the first half of the 120th century (holocene calendar) was that of the analogue revolution by radio and television, the second half of this century is that of the digital revolution and the systematic use of algebra in the data transmission. It is also the emergence automated error handling, where an "Error" is a condition when the output information does not match with the input information. During transmission, digital signals suffer from noise that can introduce errors in the binary bits travelling from one system to other.



The "error correction codes" (ECC) are used to add redundancy to the data to make it tolerant to the transmission errors (at least to a given degree). Basically, the idea is to encode one way or another an information sequence and add encoding the original data as a control data integrity. So even if some of the information is corrupted, but not too much... redundancy will identify the incorrect parts of the message.

Remark

ECC are not only used for data transmission but also to give unique identifiers to bills in some countries and also in some factories to build a nomenclature system (naming) of pieces and to control afterwards the input of employees in computer systems! For example, bar-codes and QR-codes contains ECC.

Thus, after the appearance of audio CD in the early 11980s (holocene calendar), we must take into account the development of broadcasting satellites and new means of communication such as the fax, the Minitel, Internet or digital phone using any correcting codes errors (CCE). Even photography, radio and books are becoming digital (and will probably only be in this format in a hundred years).

Image or sound reproduction techniques are related to the transmission and to the correct reading of many digital messages, also known as "words". A message consists of words themselves made up of symbols (a particular example being the "bit" which for reminder is contraction of "BInary digiT"), taken from an alphabet. If the alphabet is binary so each symbol will be a bit.

Let us take the message 00101 formed of 5 bits each worth 0 or 1. If we send the message as is, a transmission error or of reading can take place and make the message unintelligible (or the bill/piece unique identifier). Let us decide to repeat the message three times and send it:

$$001010010100101 \tag{14.4.1}$$

If the received message contains an error, this error can be corrected. If there are two errors, the receiver is able to detect that there was a mistake but can not always recover the original message. Finally, if it occurs more than two errors during transmission, the receiver can not detect them.

We have seen just now a first example of ECC, named "repeatedly code". This code, which corrects errors and detects two, was used in some Audio CD player having three heads. The signal 0 or 1 is read independently by each of these three heads to give a word of three digits, and a reading error can be sometimes corrected.

Note that it is natural to extend a message to protect it. Let us consider the words of a language. They are usually very far from each other, two words differs according to their lengths and in the letters and syllables used. Thus it will be difficult to confound the words "library" and "cabinet" although these words are mispronounced or misheard and we will naturally reconstitute the message in a conversation even when some letters would be deleted or distorted. The military meanwhile spell some information by saying "Alpha Zulu" for "AZ" to avoid errors...

A second example widely used in computing science of error detection is the addition of a "parity bit". Let 00101 be the original message and add it a last bit obtained by adding five bits of departure modulo 2. The message is 001010 and we can therefore detect errors but can not correct it.... For this, we make the sum of all the bits to obtain 0 if there is no error, and 1 otherwise. This code named "parity code" is used everywhere: in the social security numbers where we add a key, in those of bank accounts or in bar-code of supermarkets where it is the 13th digit that is the control key (the space probe Voyager II is one of the many users of parity codes to communicate almost in a reliable way and also the 8th bit in the ASCII system which is used as a parity bit).

For many years, the DRAM sockets managed words with one bit only; it was then necessary to put the 8 memory card sockets for working with bytes (8 bits). But at this time many cards included not 8 but 9 sockets! The ninth socket was designed to store a parity bit at each start of a memory byte. When reading a byte, we checked, if between the time of writing and that of reading, parity had not been changed (due to a parasite, for example).

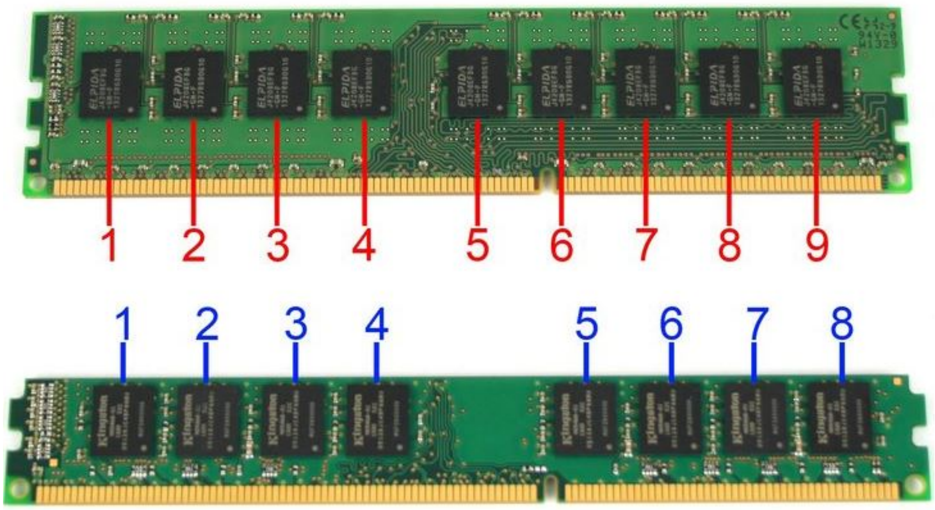


Figure 14.335: **ECC RAM vs Non-ECC RAM**

Remark

We would like to point out that ECC memory is considered server-grade. They are sold to a market which pays much closer attention to reliability than the consumer market. As a result, ECC RAM is usually subjected to much stricter testing and validation before shipping. This is a big factor in why ECC modules see much lower failure rates.

Finally, let us see a third example used on some computer servers that use in RAID4 or RAID6 parallel disks, this latter using Hamming codes that we will see later (see page 7167).

Suppose we have 3 hard drives, and the content of the first byte of each disc is the following:

$$\begin{gathered}\text{D1:11110000}\\ \text{D2:10101010}\\ \text{D3:00110011}\end{gathered}\tag{14.4.2}$$

So it is sufficient take each column and count the number p of 1 in the column. The value of the parity bit is then p modulo 2. We have for the first column in the example above p which is 2. Therefore, the parity bit is equal to 0, etc. Then we have on the control disc (CD):

$$\text{CD:011001001}\tag{14.4.3}$$

These three examples are fundamental to the basic coding theory and show that we can control the appearance of error by deliberately lengthening the message before transmission or reading. More sophisticated algebraic techniques are then used to improve the performance of coding, thanks to:

1. Know if any errors occurred (detection problems)
2. Find the initial correct message from the message received (correction of problem)
3. Correct the most possible mistakes while using the least possible additional bits (the problem of performance encoding)

From the mathematical point of view, one of the interests of coding theory is to show that algebra applies once again well in our everyday life when we listen to music, or we settle in front of our televisions, and that such abstract notions as those of vector spaces or polynomials over finite fields allow us to read messages, listen to music or watch movies in optimum conditions!

We distinguish the following two classes of ECC:

1. Block codes
2. Treillis codes

The figure below provides a simple overview of the error correcting codes family[85]. In the first class (right in the figure), we have the most popular codes such as BCH codes, Reed-Solomon and Goppa, Golay and Hamming. The second class (left in the figure) is less rich in variety but has a lot more flexibility, especially in the choice of parameters and decoding algorithms available. Let us cite for example the binary systematic recursive convolutional codes widely used in coded modulation and the parallel concatenated codes (Turbo Codes).

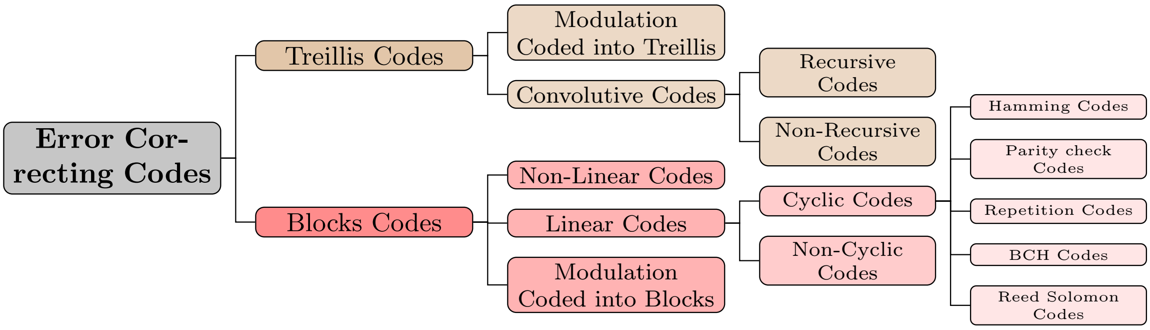

Error Correcting Codes
Treillis Codes
Blocks Codes
Modulation Coded into Treillis
Convolutive Codes
Recursive Codes
Non-Recursive Codes
Non-Linear Codes
Linear Codes
Modulation Coded into Blocks
Cyclic Codes
Non-Cyclic Codes
Hamming Codes
Parity check Codes
Repetition Codes
BCH Codes
Reed Solomon Codes

Figure 14.336: **Non-exhaustive orgchart of correcting codes**

Remark

To introduce the foundations of the theory of error correcting codes, we recommend strongly the reader to have read through first the section of Set Theory (see page 378), after this of Statistical Mechanics (where the information theory is at page 3525) after of Numerical Methods (see page 6272) and finally of Topology (see page 2446).

When possible we will show the reader how to put quickly in practice ECC using native MATLAB® Communication Toolbox functions.

---

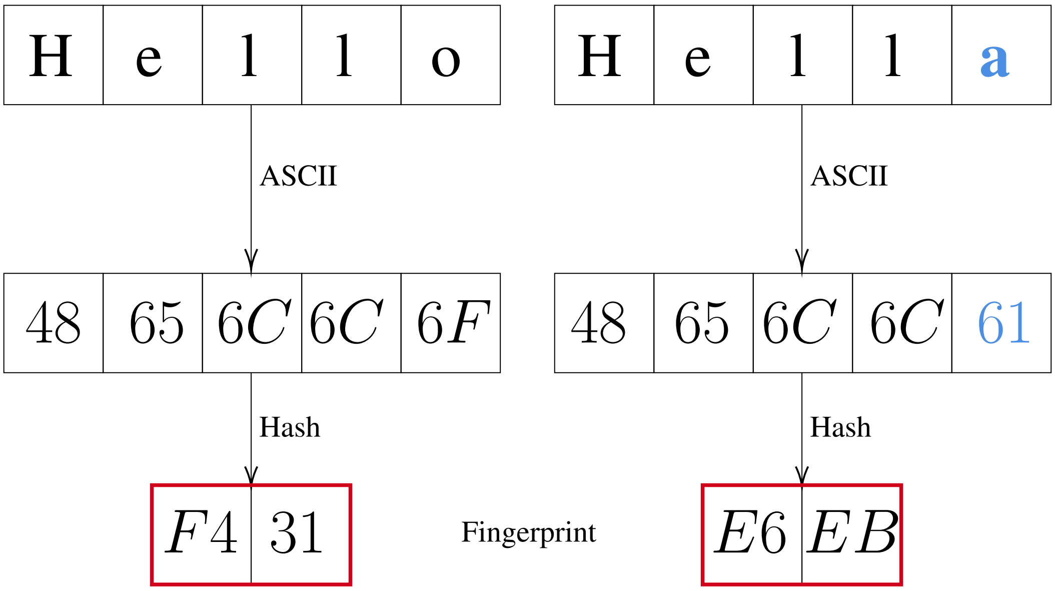
[85] For example "Quantum Error Correction" is not indicated

14.4.1 CheckSum

Before starting the part of pure mathematics, we would like to make a small introduction to the "checksum" (control sum) that is a tool frequently used in the business when exchanging files over several Giga Bytes between two computers or when downloading on the Internet.

The checksum, also sometimes named "fingerprint" is a basic concept of coding theory used for correcting codes. It corresponds to a particular case of "redundancy control code". It is widely used in computer and digital telecommunications as already said.

One basic technique (among a dozens more or less sophisticated) is to take the sum of a given length of bits (byte, word, or other ...) and calculate the modulo 255 (FF in hexadecimal). For example, if we take two words and we rely on their hexadecimal ASCII code (you can find ASCII tables almost everywhere on the internet):



$$48+65+6C+6C+6F=F4 \bmod FF$$
$$48\cdot 1+65\cdot 2+6C\cdot 3+6C\cdot 4+6F\cdot 5=31 \bmod FF$$

$$48+65+6C+6C+6F=F4 \bmod FF$$
$$48\cdot 1+65\cdot 2+6C\cdot 3+6C\cdot 4+61\cdot 5=EB \bmod FF$$

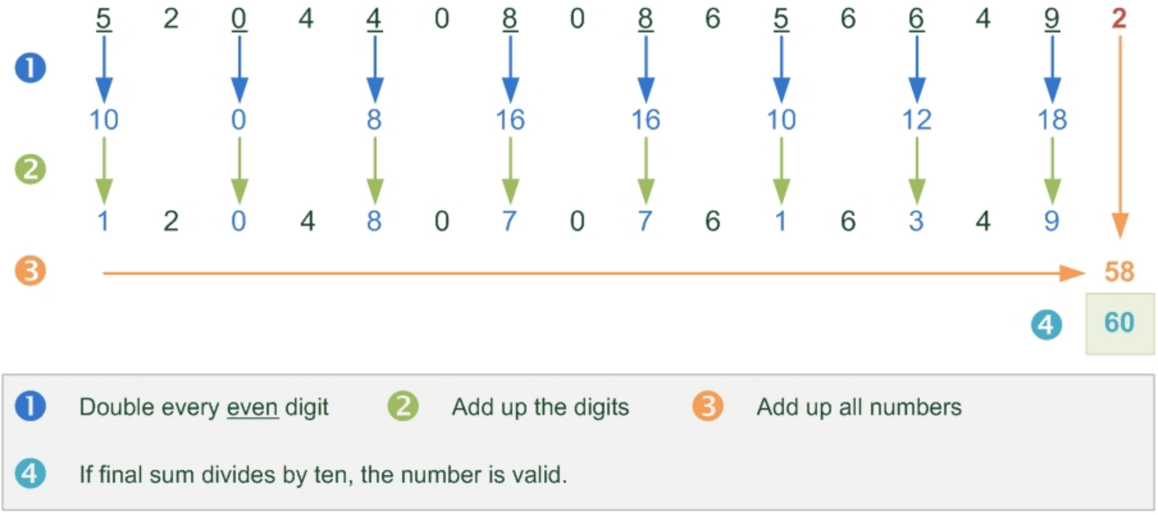
Figure 14.337: Basic Checksum principle

Some tools also use the MD5 algorithm (Message Digest 5) or SHA (Secure Hash Algorithm) to have a checksum of a message (see section Cryptography page 7249 and page 7253).

14.4.1.1 Luhn algorithm

The "Luhn algorithm" or "Luhn formula", also known as the "modulus 10" or "mod 10" algorithm, is a simple checksum formula used to validate a variety of identification numbers, such as Credit card numbers, IMEI numbers, National Provider Identifier numbers in the US, and Canadian Social Insurance Numbers (see examples further below). It was created by IBM scientist Hans Peter Luhn and described in U.S. Patent No. 2,950,048, filed on January 6, 11954 (holocene calendar), and granted on August 23, 11960 (holocene calendar).

The algorithm is in the public domain and is in wide use today. It is specified in ISO/IEC 7812-1 and is completely described in the following figure:

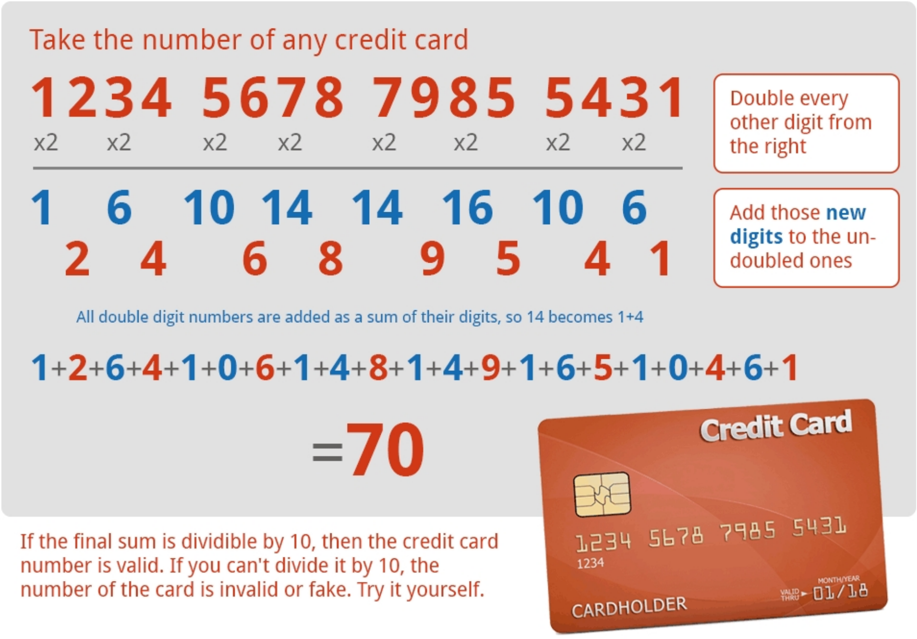


5 2 0 4 4 0 8 0 8 6 5 6 6 4 9 2
①
10 0 8 16 16 10 12 18
②
1 2 0 4 8 0 7 0 7 6 1 6 3 4 9
③
58
④
60
① Double every even digit
② Add up the digits
③ Add up all numbers
④ If final sum divides by ten, the number is valid.

Figure 14.338: **Luhn algorithm for checksum principle**

The Luhn algorithm will detect any single-digit error, as well as almost all transpositions of adjacent digits. It will not, however, detect transposition of the two-digit sequence 09 to 90 (or vice versa). It will detect 7 of the 10 possible twin errors (it will not detect 22 ↔ 55, 33 ↔ 66 or 44 ↔ 77).

A picture is worth a thousand words let us see how this apply to VISA cards:


1234 5678 7985 5431
x2 x2 x2 x2 x2 x2 x2 x2
1 6 10 14 14 16 10 6
2 4 6 8 9 5 4 1
1+2+6+4+1+0+6+1+4+8+1+4+9+1+6+5+1+0+4+6+1
=70
Credit Card
1234 5678 7985 5431
1234
CARDHOLDER
VALID THRU
MONTH/YEAR
01/18

Take the number of any credit card

Double every other digit from the right

Add those **new digits** to the un-doubled ones

All double digit numbers are added as a sum of their digits, so 14 becomes 1+4

If the final sum is dividible by 10, then the credit card number is valid. If you can't divide it by 10, the number of the card is invalid or fake. Try it yourself.

Other, more complex check-digit algorithms (such as the Verhoeff algorithm and the Damm algorithm) can detect more transcription errors.

14.4.2 Check Digit

Also before starting the part of pure mathematics, we would like to make a small introduction to some common check digits (in fact any professionally driven government or company should have all objects or individuals identifiers with at least one check digit). We will only focus on example that have been asked to me by my students or people that contacted me on the Internet (otherwise I can dedicate a whole book only for examples on check digits...)!

Definition (#14.4.596): A "check digit" is a form of redundancy check used for error detection on identification numbers, such as bank account numbers, which are used in an application where they will at least sometimes be input manually. It is analogous to a binary parity bit used to check for errors in computer-generated data. It consists of one or more digits computed by an algorithm from the other digits (or letters) in the sequence input.

14.4.2.1 European Article Numbering (EAN-13)

Let us see first the check digit of EAN-13 bar codes (Cyclic Redundancy Check[86] (CRC) ECC type):

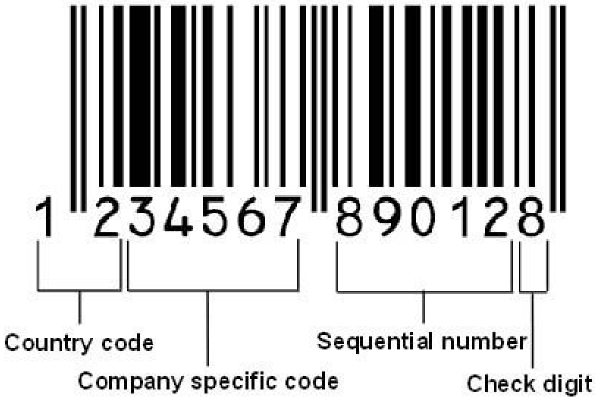


Figure 14.339: **EAN-13 bar-code example**

There are various different errors that can occur when numbers are written, printed or transferred in any manner. Different methods of assigning check digits are better at detecting certain kinds of errors than others. The most common types of errors that occur in practice and their frequencies, according to one study, are as follows:

Error type	Form	Relative frequency
Single error	a replaced by b	79.1%
Transposition of adjacent digits	ab replaced by ba	10.2%
Jump transposition	abc replaced by cba	0.8%
Twin error	aa replaced by bb	0.5%
Phonetic error	$a0$ swapped with $1a$, $a=2,\ldots,9$	0.5%
Jump twin error	aca replaced by bcb	0.3%

Table 14.55: **Various costs**

---

[86] type of checksum, specifically a position dependent checksum algorithm

Another common type of error not mentioned here is accidental insertion or deletion of characters. In the cases we will consider, the number will have a fixed length, so insertions and deletions will be automatically detected.

The EAN-13 format uses a modulus 10 scheme, with check digit (a_c) defined by:

$$a_c = -(a_1, a_2, \ldots, a_6, a_7) \cdot (3,1,3,1,3,1,3) \mod 10 \tag{14.4.4}$$

For example, if we start with the number 1234567 in the EAN-8 scheme, then our check digit is:

$$\begin{aligned} a_c &= -(1,2,3,4,5,6,7) \cdot (3,1,3,1,3,1,3) \mod 10 \\ &= -(1 \cdot 3 + 2 \cdot 1 + 3 \cdot 3 + 4 \cdot 1 + 5 \cdot 3 + 6\dot{1} + 7 \cdot 3) \mod 10 \\ &= -60 \mod 10 = 0 \end{aligned} \tag{14.4.5}$$

which makes the full bar code number 12345670.

When dealing with ECC we must worry about how effective a scheme is in detecting errors!

Let us consider only two cases:

1. Single error detection rate:

   If a digit d whose weight is 1 is changed to c, the weighted sum will change by $d-c$. The error will go undetected only if $d-c=0 \mod 10$. But this happens only when $d=c$, in which case there has not been an error after all, so all errors of this kind are caught.

   What if the weight were 3? Then the error would be undetected if $3(d-c)=0 \mod 10$. But again, this cannot happen unless $d=c$. Thus, this method has a 100% single error detection rate (SEDR).

2. Transposition of adjacent digits detection rate:

   Suppose two adjacent digits, cd, are transposed to dc. If c's weight is 3 (hence d's weight is 1), the weighted sum is changed by:

   $$(3c+d)-(c+3d)=2(c-d) \tag{14.4.6}$$

   which will be detected unless $2(c-d)=0 \mod 10$, which can happen only if c and d differ by 5. The same would have applied if c had been weighted by 1 and d by 3.

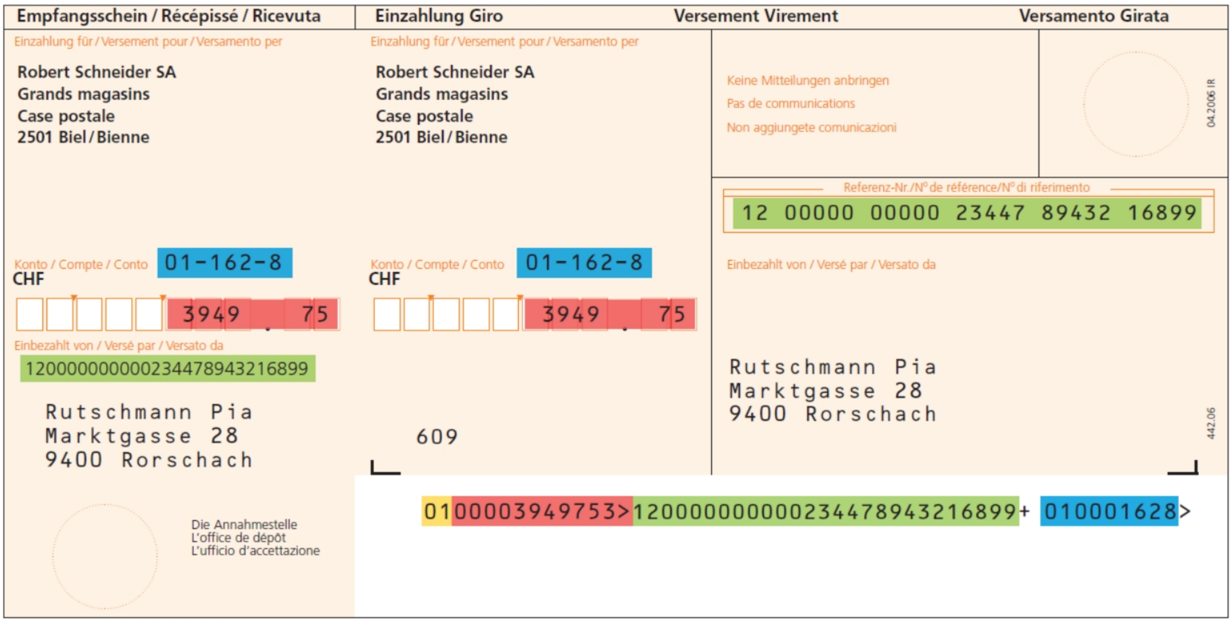
   As a result, the transpositions that will go undetected must involve $0 \leftrightarrow 5, 1 \leftrightarrow 6, 2 \leftrightarrow 7, 3 \leftrightarrow 8$ and $4 \leftrightarrow 9$. So, 10 transpositions are undetectable.

   There are 100 possibilities for each pairing, and the transposition of 90 of these would result in an error. Therefore the detection rate is $80/90 = 88.9\%$ transposition error detection rate (TEDR).

14.4.2.2 Swiss Post Payment slip

Another famous example for my country is the Swiss Post payment slip where numerous numbers have a check digit.

Here is a sample preview of such a payment slip:



Empfangsschein / Récépissé / Ricevuta

Einzahlung für / Versement pour / Versamento per

Robert Schneider SA
Grands magasins
Case postale
2501 Biel / Bienne

Konto / Compte / Conto 01-162-8
CHF
3949 . 75

Einbezahlt von / Versé par / Versato da
120000000000234478943216899

Rutschmann Pia
Marktgasse 28
9400 Rorschach

Die Annahmestelle
L'office de dépôt
L'ufficio d'accettazione

Einzahlung Giro **Versement Virement** **Versamento Girata**

Einzahlung für / Versement pour / Versamento per

Robert Schneider SA
Grands magasins
Case postale
2501 Biel / Bienne

Konto / Compte / Conto 01-162-8
CHF
3949 . 75

609

Keine Mitteilungen anbringen
Pas de communications
Non aggiungete comunicazioni

04.2006 IR

Referenz-Nr./N° de référence/N° di riferimento
12 00000 00000 23447 89432 16899

Einbezahlt von / Versé par / Versato da

Rutschmann Pia
Marktgasse 28
9400 Rorschach

442.06

0100003949753>120000000000234478943216899+ 010001628>

Figure 14.340: Swiss payment slip

and below is the technical description of the bottom right number with in red the check digit on which we will focus here:

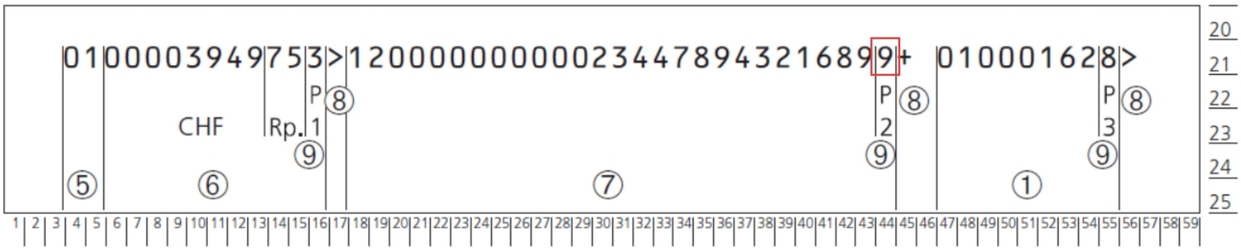

0100003949753>120000000000234478943216899+ 010001628>
P
⑧
CHF
Rp.
1
⑨
⑤
⑥
⑦
P
⑧
2
⑨
①
P
⑧
3
⑨
20
21
22
23
24
25
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59

Figure 14.341: Swiss payment slip check digit for Reference Number

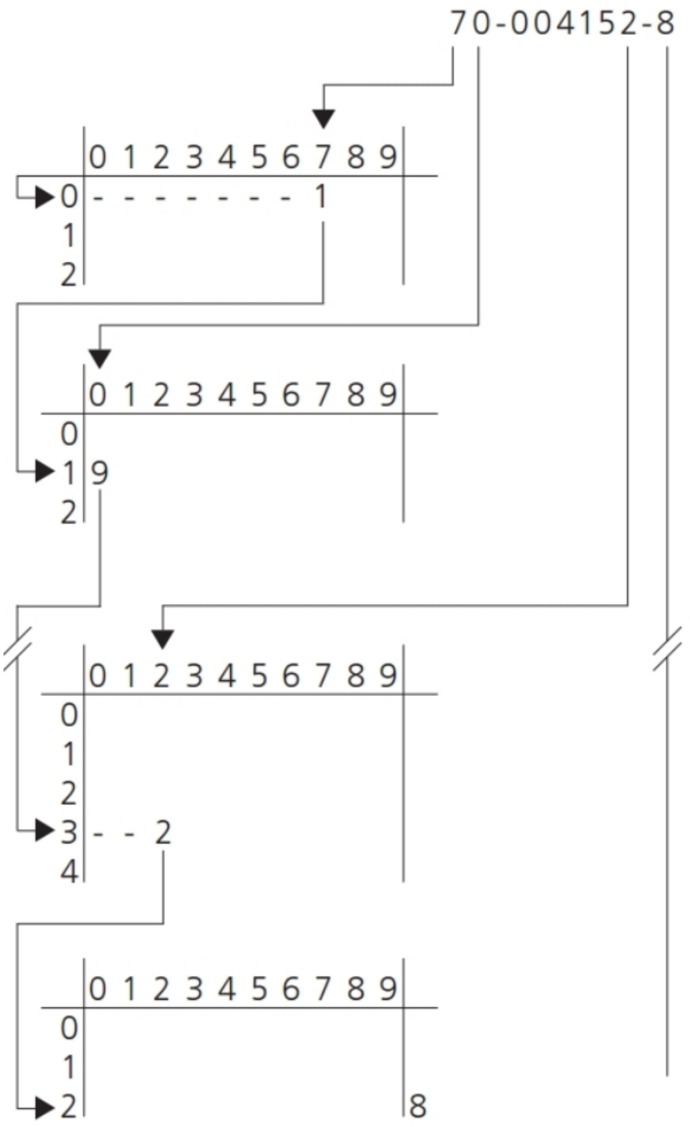
All the digit check in this Swiss payment slip are based on a recursive modulo 10 computation.

Let us give a small example on how to calculate the check digit based on the Swiss method. For this let us consider first the following matrix:

Remainder	Digit in the number										Check Digit
	0	1	2	3	4	5	6	7	8	9	
0	0	9	4	6	8	2	7	1	3	5	0
1	9	4	6	8	2	7	1	3	5	0	9
2	4	6	8	2	7	1	3	5	0	9	8
3	6	8	2	7	1	3	5	0	9	4	7
4	8	2	7	1	3	5	0	9	4	6	6
5	2	7	1	3	5	0	9	4	6	8	5
6	7	1	3	5	0	9	4	6	8	2	4
7	1	3	5	0	9	4	6	8	2	7	3
8	3	5	0	9	4	6	8	2	7	1	2
9	5	0	9	4	6	8	2	7	1	3	1

So what would be the check digit of the number 70004152 (it's a typical 8 digit Swiss Post account)?

The process to determine that latter is simply the following:



Règles

- Begin by the remainder "0" → combine it with the first digit of the example: "7". The cross value in the matrix to take as new remainder is "1"
- Combine the previous remainder "1" with the second digit of the example "0" to get the new matrix row and column an we get as new remainder "9"
- etc.
- Combine the previous remainder "3" with the last digit of the example "2" to get the new matrix row and column an we get as final remainder "2"
- The digit in the last column at the row number corresponding to last remainder "2" is the check digit in this case given by the value "8"

14.4.2.3 International Bank Account Number (IBAN)

The International Bank Account Number (IBAN) is an internationally agreed system of identifying bank accounts across national borders to facilitate the communication and processing of cross border transactions with a reduced risk of transcription errors. It was originally adopted by the European Committee for Banking Standards (ECBS), and later as an international standard under ISO 13616.

The IBAN consists of up to 34 alphanumeric characters comprising: a country code; two check digits; and a number that includes the domestic bank account number, branch identifier, and potential routing information. The check digits enable a sanity check of the bank account number to confirm its integrity before submitting a transaction.

Before IBAN errors of transcription were not detectable and it was not possible for a sending bank to validate the routing information prior to submitting the payment. Routing errors caused delayed payments and incurred extra costs to the sending and receiving banks and often to intermediate routing banks.

The IBAN should not contain spaces when transmitted electronically. When printed it is expressed in groups of four characters separated by a single space, the last group being of variable length. Here is an example of how typically Switzerland write IBAN numbers:

```
CH93 0076 2011 6238 5295 7
```

Permitted IBAN characters are the digits 0 to 9 and the 26 upper-case Latin alphabetic characters A to Z. This applies even in countries (e.g., Thailand) where these characters are not used in the national language.

An IBAN is validated by converting it into an integer and performing a basic modulo 97 operation (as described in ISO 7064: *Security techniques - Check character systems*) on it. If the IBAN is valid, the remainder equals 1.

The procedure is a follows:

- Check that the total IBAN length is correct as per the country. If not, the IBAN is invalid;
- Move the four initial characters to the end of the string;
- Replace each letter in the string with two digits, thereby expanding the string, where A $=10$, B $=11$, ..., Z $=35$;
- Interpret the string as a decimal integer and compute the remainder of that number on division by 97.

If the remainder is 1, the check digit test is passed and the IBAN might be valid.

✍**Example:**

Given a fictitious United Kingdom bank, sort code `12-34-56`, account number `98765432`. For the sanity check we follow the above procedure:

- The IBAN is therefore:

  `GB82 WEST12345698765432`

- We rearrange:

  `WEST12345698765432GB82`

- We convert to integer:

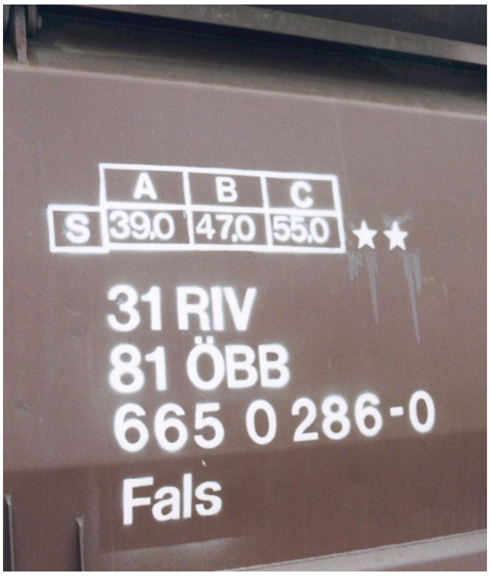
  `3214282912345698765432161182`

- We compute the modulo 97:

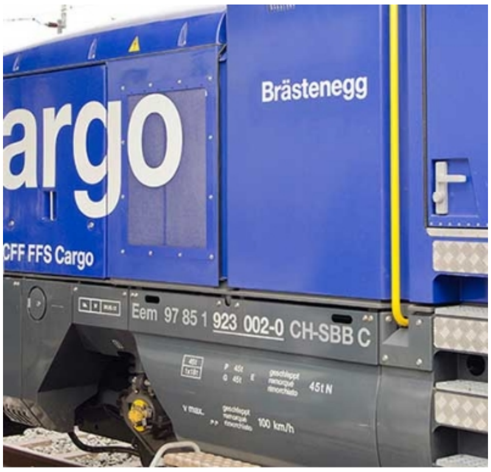
$$\texttt{3214282912345698765432161182} \mod 97 = 1$$

14.4.2.4 UIC wagon numbers

Wagon numbers (or coach numbers) are key data for railway operations. They enable a railway wagon or coach to be positively identified and form a common language between railway operators, infrastructure companies and the state authorities. The system of wagon numbering has been laid down by the International Union of Railways (Union Internationale des Chemins de fer or UIC) and is similar to that used for the locomotives and multiple units.


A B C
S 39.0 47.0 55.0
31 RIV
81 ÖBB
665 0 286-0
Fals

(a) Slovak UIC wagon number


Brästenegg
argo
CFF FFS Cargo
Eem 97 85 1 923 002-0 CH-SBB C
45 t N
100 km/h

(b) Swiss UIC wagon number

The complete wagon number comprises 12 digits. The individual digits have the following meaning:

- Digit 1-2: Type of vehicle and indication of the interoperability capacity
- Digit 3-4: Country Code (Switzerland (CH) = 85)
- Digit 5-8: Vehicle type information
- Digit 9-11: Individual running number (serial number)
- Digit 12: Self-check digit

The digits are multiplied individually from right to left alternately by 2 and 1, and digit summed. The difference between this sum and the next multiple of ten is the check digit, placed after the eleventh digit, separated by a dash.

✍**Example:**

Given a fictitious Wagon UIC `21-81-24 7121 7`. For the sanity check we follow the above procedure:

- The UIC is therefore:

  `2 1 8 1 2 4 7 1 2 1 7`

- We multiply:

  $$2\cdot 2 \quad 1\cdot 1 \quad 8\cdot 2 \quad 1\cdot 1 \quad 2\cdot 2 \quad 4\cdot 1 \quad 7\cdot 2 \quad 1\cdot 1 \quad 2\cdot 2 \quad 1\cdot 1 \quad 7\cdot 2$$

  It gives:

  $$4\;1\;16\;1\;4\;4\;14\;1\;4\;1\;14$$

- We sum up the digits using:

  $$4+1+1+6+1+4+4+1+4+1+4+1+1+4=37$$

- We take the next multiple of 10 that 40 so the check digit is equal to three 3.
- Finally we get:

  `21-81-24 7121 7-3`

Remark

The swiss "Unternehmens-Identifikationsnummer" (company identification number) also has a check digit and so on...

14.4.3 Permutations

Also before starting the part of pure mathematics, we would make a small introduction to the check digit using permutations in VISA card numbers.

Credit cards use an error-detecting scheme that was developed by IBM. It uses the permutation:

$$\sigma=(0)(124875)(36)(9) \tag{14.4.7}$$

In other words (for more details see the section of Set Algebra page 1417) $\sigma(0)=0$, $\sigma(1)=2$, $\sigma(2)=4$, etc.

Notice also that:

$$\sigma(x)=2x \mod 9 \tag{14.4.8}$$

In a 16 digit credit card number, the final digit is the check digit. Let the credit card number be $(a_1,a_2,\ldots,a_{15},a_{16})$, with a_{16} being the check digit. Then:

$$a_{16}=-[\sigma(a_1)+a_2+\sigma(a_3)+a_4+\ldots+a_{14}+\sigma(a_{15})] \mod 10 \tag{14.4.9}$$

Note that in this example the permutation was applied to a_i where i is odd (a_1, a_3, etc), because there is an odd number of digits excluding the check digit. Had this scheme been used on a number with an even number of digits excluding the check digit, the permutation would have been applied to a_i where i was even.

As we know form the previous example, dealing we ECC must worry about how effective a scheme is in detecting errors!

Let us consider again only two cases:

1. Single error detection rate:

   This scheme catch all single-digit errors (100% SEDR). Indeed, for example if digit a_i is changed from c to d, and i is even, the remainder will change by $c-d$, which is non-zero (and is, of course, smaller than the modulus $N=10$). If i is odd, it will change by $\sigma(c)-\sigma(d)$. This is again non-zero: $\sigma(c)$ cannot be equal to $\sigma(d)$ if σ is a permutation.

2. Transposition of adjacent digits detection rate:

   If two adjacent digits c and d are transposed, one of them must have the permutation applied - say c. The remainder will be unchanged only if $\sigma(c)+d=\sigma(d)+c$. Since $\sigma(x)=2x \mod 9$, this happens only when $c=d \mod 9$, that is, only when c and d are 0 and 9 (in either order).

   Therefore, for each pair of adjacent digits, of the 90 possible transposition errors, two will be undetectable. So the detection rate for transpositions is $88/90=97.8\%$ (TEDR).

14.4.4 Encoders

Given Q a finite set of q elements (bits, alphabets). Given k and n two non-zero integers with $k \leq n$. The set of messages will be a part E of Q^k and we introduce a bijective application (at least that is the goal!):

$$\begin{gathered} f : E^k \leftrightarrow Q^n \\ a = (a_1, a_2, \ldots, a_k) \leftrightarrow c = (c1, c_2, \ldots, c_n) \end{gathered} \tag{14.4.10}$$

named "encoding application" or "encoder". The message or word is an element of E that is to say Q^k. It is modified to provide the word:

$$f(a) = c \in Q^n \tag{14.4.11}$$

It is the word c that will be transmitted and read by any system to give a message received $x = (x_1, \ldots, x_n)$ possibly flawed.

Let us now denote $C = f(E)$ the image of f. Since f is subjective by definition it is also injective, f realized a bijection of E on C and C can be considered as the set of all possible error coding messages. C is named the "code of length n", and the elements of C are named the "words" of the code. The cardinal of the code is by definition that of C that is to say Card(C). To measure the degree of difference between two words x and y of Q^n, we use the "Hamming distance" d_H defined by:

$$\boxed{\forall x, y \in Q^n : \quad d_H(x, y) = \text{Card}\,\{i \in \mathbb{N} | x_i \neq y_i\}} \tag{14.4.12}$$

For binary strings, the Hamming distance is equal to the number of ones in XOR (exclusive OR) operation on the strings.

Theorem 14.28. On any set Q, we therefore define the application $d : Q^n \times Q^n \rightarrow \mathbb{R}$ by:

$$d_H(x, y) = \text{Card}\,\{i \in \mathbb{N} | x_i \neq y_i\} \tag{14.4.13}$$

If we denote by δ_x the characteristic function of x:

$$\delta_x(y) = \begin{cases} 1 & x \neq y \\ 0 & \text{otherwise} \end{cases} \tag{14.4.14}$$

then:

$$d_H(x, y) = \sum_{i=1}^{n} (1 - \delta_{x_i}(y_i)) \tag{14.4.15}$$

is a distance.

Let us now prove that following the topological axioms of a distance (see section Topology page 2450) that this is really a distance:

Proof 14.28.1. Ok let us prove the five axioms of a distance:

1. We will suppose that for the reader:

   $$\forall x,y\in Q^n:\quad d(x,y)\geq 0 \tag{14.4.16}$$

   is obvious (if not send us a request).

2. We will also suppose that:

   $$\forall x,y\in Q^n:\quad d_H(x,y)=d_H(y,x) \tag{14.4.17}$$

   is obvious (if not send us a request).

3. We will also suppose that:

   $$\forall x\in Q^n:\quad d_H(x,x)=0 \tag{14.4.18}$$

   is obvious (if not send us a request).

4. We will also suppose that:

   $$\forall x,y\in Q^n:\quad d_H(x,y)=0\Rightarrow x=y \tag{14.4.19}$$

   where $d_H(x,y)=0$ mean that $x_i=y_i$ for $i=1\dots n$ and therefore that $x=y$.

5. Finally:

   $$\forall x,y\in Q^n:\quad d_H(x,z)\leq d_H(x,y)+d_H(y,z) \tag{14.4.20}$$

   Indeed:

   $$d_H(x,z)=\sum_{i=1}^{n}(1-\delta_{x_i}(z_i)) \tag{14.4.21}$$

   but as:

   $$(1-\delta_{x_i}(z_i))\leq(1-\delta_{x_i}(y_i))+(1-\delta_{y_i}(z_i)) \tag{14.4.22}$$

   as $(1-\delta_{x_i}(z_i))$ is equal to 1 if $x_i\neq z_i$ and 0 otherwise, then:

   $$d_H(x,z)=\sum_{i=1}^{n}(1-\delta_{x_i}(z_i))\leq\sum_{i=1}^{n}(1-\delta_{x_i}(y_i))+(1-\delta_{y_i}(z_i))=d_H(x,y)+d_H(y,z) \tag{14.4.23}$$

■ Q.E.D.

✍**Example:**

The Hamming distance between the word "**ramer**" and "**cases**" or between "**0100**" and "**1001**" is equal to 3.

Caution!!! Vectors will be denoted without the arrow in respect for tradition for this study field.

The Hamming distance d_H is therefore indeed a metric (see section Topology page 2447) as we just proved above and then we name "Hamming space" on Q the set Q^n equipped with the metric d_H.

Definition (#14.4.597): If Q is a group, the "Hamming weight" $w_H(x)$ of a word $x \in Q^n$ is the number of non-zero components:

$$w_H(x) = d_H(x,0) \tag{14.4.24}$$

where 0 is the word (vector) of Q^n with all its components equal to the neutral element of Q. Furthermore, we have the following trivial property:

$$d_H(x,y) = d_H(x-y,0) = w(x-y) \tag{14.4.25}$$

Remark

When $Q = \{0,1\}$ we will talk about "binary code" (we will see soon later an another form of writing for this binary set) of dimension n equal to 2.

The "minimum distance" of the code C is the minimum distance between two distinct words of this code. We denote that integer by $d(C)$ or simply d and therefore:

$$d(C) = d = \min\{d_H(x,y) | x,y \in C \text{ and } x \neq y\} \tag{14.4.26}$$

or using the property of Hamming weight $d_H(x,y) = w(x-y) = w(x)$:

$$d(C) = d = \min\{w(z) | z \in C^*\} \tag{14.4.27}$$

✍**Example:**

Let us consider the following redundant code denoted $(5,4)$ for 4 coded words of length 5:

Original Word	Coded Word	Weight	Code Identifier
00	**00**000	0	C_1
01	**01**110	3	C_2
10	**10**011	3	C_3
11	**11**101	4	C_4

The Hamming distance of each of the pairs of code are:

$$\begin{aligned} d(C_1,C_2)&=3 \qquad & d(C_2,C_3)&=4\\ d(C_1,C_3)&=3 & d(C_2,C_4)&=3\\ d(C_1,C_4)&=4 & d(C_3,C_4)&=3 \end{aligned} \tag{14.4.28}$$

The smallest non-zero minimum weight is therefore 3 and the smallest Hamming distance is 3.

Definitions (#14.4.598):

D1. A code C of length n, of cardinal M and of minimum distance d is named a "(n, M, d) code". The numbers n, M, d are the "code parameters". Thus, the code $(7, 4, 3)$ is a code of length seven, that is to say that the receiver receives seven bits, of length four, that is to say that once decoded, the message contains four symbols (letters) and the minimum distance between each codeword is three.

D2. We name "minimum weight" of a C code the integer:

$$\min\{w_H(x)|x\in C, x\neq 0\} \tag{14.4.29}$$

The parameter d plays an important role because it is closely related to the number of errors that can be corrected. Suppose that the encoded message is $c=(c_1,\ldots,c_n)$ and that there were at least e errors of transmission or reading. The resulting message obtained $x=(x_1,\ldots,x_n)$ satisfies $d(x, c)\leq e$. We can fall back on c from x if, and only if, there exists a single code word located at a distance of x less than or equal to e (i.e. the center-to-center distance between two balls is equal to $2e$). In other words, it is necessary and sufficient that the closed balls of radius e and centered on the elements of the code C are disjoint. A code will correct e errors if this condition is true.

Therefore, a code C of minimum distance d corrects at most:

$$e=\left[\frac{d-1}{2}\right] \tag{14.4.30}$$

where [] represents the integer part of a real number.

Indeed, if a message of the code is at $d/2$ we will not be able to know to which message of the code (center of ball) it belongs since being (in a pictorial way) at the tangent of two balls. This is the reason why we will take $\dfrac{d-1}{2}$ which is then the "safe distance" to correct as much as possible an erroneous message of the code. Moreover, since the number of errors is an integer, it comes the previous notation with the square brackets.

It is clear that the code can detect at most $d-1$ errors. Indeed, however how to distinguish an incorrect code from a correct code (code = coded message)? Apart from the fact that each element of the code is different (injective application of the set of messages in the set of encoded messages), it is also necessary to be able to differentiate among them those which are erroneous codes from those which do not Are not. Hence the $d-1$!!!

✍**Example:**

Let us consider the following redundant code denoted $(5,4)$ for 4 coded words of length 5 whose minimal distance was therefore $d=3$:

Original Word	Coded Word	Weight	Code Identifier
00	**00**000	0	C_1
01	**01**110	3	C_2
10	**10**011	3	C_3
11	**11**101	4	C_4

This code therefore allows and make it possible to detect at most:

$$d=3-1=2 \tag{14.4.31}$$

errors and to correct at most:

$$e=\left[\frac{d-1}{2}\right]=\left[\frac{3-1}{2}\right]=\left[\frac{2}{2}\right]=1 \tag{14.4.32}$$

of them.

14.4.4.1 Block code

In coding theory, a block code is any member of the large and important family of error-correcting codes that encode data in blocks. There is a vast number of examples for block codes, many of which have a wide range of practical applications. Block codes are conceptually useful because they allow coding theorists, mathematicians, and computer scientists to study the limitations of all block codes in a unified way. Such limitations often take the form of bounds that relate different parameters of the block code to each other, such as its rate and its ability to detect and correct errors.

Definition (#14.4.599): A "block code" of size M and length n defined on an alphabet of q symbols (1 and 0 for the binary language for example) is a set of M vectors named the "code words". The idea is that each information word composed of k symbols is associated with a single codeword composed of n symbols. The vectors are therefore of length $n\geq k$ and their components are q-ary (thus "2-ary" in the case of the binary language).

Examples of block codes are Reed–Solomon codes, Hamming codes, Hadamard codes, Expander codes, Golay codes, and Reed–Muller codes. These examples also belong to the class of linear codes, and hence they are named "linear block codes". More particularly, these codes are known as algebraic block codes, or cyclic block codes, because they can be generated using boolean polynomials.

Definition (#14.4.600): A "linear coding" means a set of code in which any linear combination (modular 2 sum most of time) of two codes within the set results in a code which also belong to the original set. Let's assume that you have a set of codes as in the figure below:

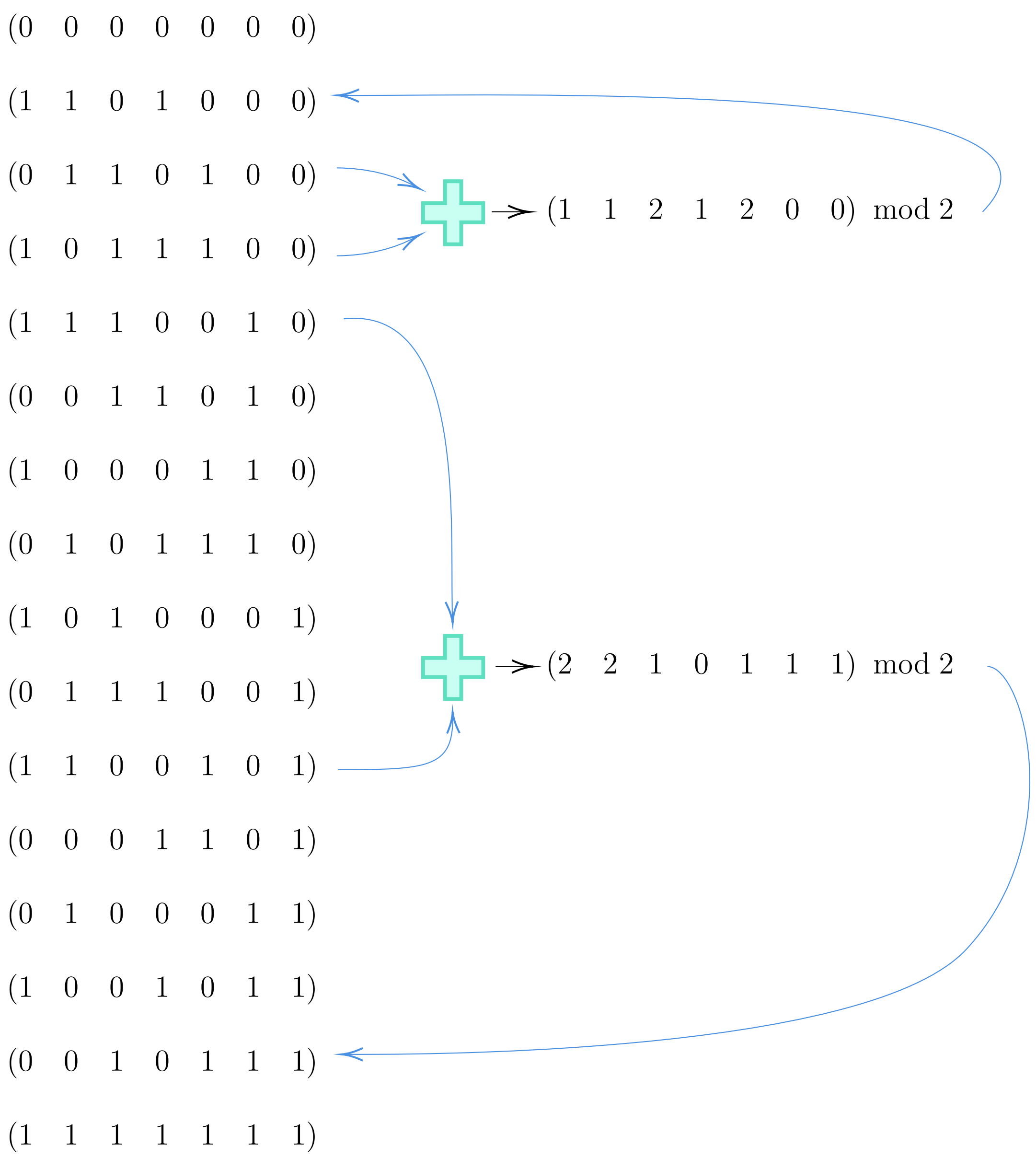

(0 0 0 0 0 0 0)
(1 1 0 1 0 0 0)
(0 1 1 0 1 0 0)
(1 0 1 1 1 0 0)
(1 1 2 1 2 0 0) mod 2
(1 1 1 0 0 1 0)
(0 0 1 1 0 1 0)
(1 0 0 0 1 1 0)
(0 1 0 1 1 1 0)
(1 0 1 0 0 0 1)
(0 1 1 1 0 0 1)
(2 2 1 0 1 1 1) mod 2
(1 1 0 0 1 0 1)
(0 0 0 1 1 0 1)
(0 1 0 0 0 1 1)
(1 0 0 1 0 1 1)
(0 0 1 0 1 1 1)
(1 1 1 1 1 1 1)

Figure 14.343: **Principle of a Linear code**

Take out any two codes from the set and take modular 2 sum of them. The result is also a member of the set as shown below. Take any other two codes and try yourself.

Remark

The linearity of the block codes also mean that the n symbols of the code word are obtained by a linear combination of the k symbols of the information word.

Let us see an example!

✍**Example:**

Let us start from the following M vectors based on $q=3$ binary symbols (hence $k=2$). In this case, we have $M=q^k$:

Vectors
000
100
010
110
001
101
011
111

We then choose n as equal to 6 and we define a bijective mapping such that:

Vectors	**Code Vectors**
000	000**000**
100	110**100**
010	011**010**
110	101**110**
001	101**001**
101	011**101**
011	110**011**
111	000**111**

As shows this example, of the particular block code conventionally denoted $(n,k)_q=(6,3)_2$, the code has no particular structure. The decoding operation involves making an exhaustive comparison of the word received at the output of the channel with all the code modes before determining the most likely code word. This simple and stupid approach explains why many times this error correcting code is faster than many others...

Thus, according to the above definition, a block code C is the result of a injective application which associates with each vector formed by k q-ary symbols (k information symbols), an image vector of length n with components in the same alphabet (n encoded symbols):

The encoding adds to the initial information $n-k$ additional symbols. The quantity:

$$R=\frac{k}{n} \tag{14.4.33}$$

is named the "rate of C", or "coding rate". The block encoding operation is "without memory", in extenso the blocks are coded independently without any correlation between two consecutive blocks.

Now it is convenient to come back a little bit on Boole's Algebras (see section Logical Systems page 7105). To the 5 axioms which define a Boolean Algebra let us add a sixth one which gives it a structure of a field:

A6. The Boolean Algebra (extension of a unitary ring by an axiom) with the law $*$ (or $\wedge$) is a field.

Remark

Let us recall that a field is a non-zero ring in which every non-zero element is invertible.

If we take the Boolean Algebra formed by the $q=2$ elementary elements $\{0,1\}$ forming a binary set (alphabet), we actually have 1 which is invertible since there exists x such that:

$$1*x=1 \tag{14.4.34}$$

which is 1 itself!

This field is denoted $\mathbb{F}_q=\mathbb{F}_2$. In the area of error correcting codes, we often work in $\mathbb{F}_2$ (single field with two elements) where for recall the addition is defined by:

$$0+0=0 \qquad 1+1=0\,(\text{therefore } 1=-1) \qquad 1+0=1 \tag{14.4.35}$$

The multiplication being defined by:

$$1\cdot 1=1 \qquad 1\cdot 0=0 \qquad 0\cdot 0=0 \tag{14.4.36}$$

$$C:\mathcal{A}^k \mapsto \mathcal{A}^n \tag{14.4.37}$$

To return to our theory of codes: the set of messages $E=\mathbb{F}_q^k$ can be equipped with a vector space structure of dimension k on $\mathbb{F}_q$ (see section Set Theory page 1805). Indeed, it suffices for this that $(E,+)$ to an Abelian group and $*$ an external law defined by $\mathbb{F}_q^k\times\mathbb{F}_q\mapsto\mathbb{F}_q^k$. If we decide to use only encoders that are linear (applications), the code $C=f(\mathbb{F}_q^k)$ becomes a vector subspace of $\mathbb{F}_q^n$ (because even if the application is bijective, since the body of the coded messages is finite, we necessarily have vector subspace of the vector space of all possible encoded messages).

Definition (#14.4.601): A "linear code" of dimension k and length n is a vector subspace of dimension k of $\mathbb{F}_q^n$ (it is the way this is said..). If the minimum distance of C is d, we say that C is a "$[n,k,d]_q$ code" or more simply "code $[n,k]$".

Remark

Linear codes are therefore a special case of block codes as shown in the hierarchical scheme at the beginning of this section.

The addition of the linearity constraint could undermine the quality of the code sought, but fortunately the performance study shows that the linear codes are very close to the best Block codes. Thus, linearity facilitates the study of block codes and allows the use of very powerful algebraic tools without reducing the class of linear blocks to an inefficient class.

Let us denote by G the matrix of the linear application $f:\mathbb{F}_q^k\mapsto\mathbb{F}_q^n$. G is a matrix of obviously dimension $n\times k$ type and every word c of C is obtained from every word x of E by:

$$c=f(x)=Gx \tag{14.4.38}$$

where $c=(c_1,\ldots,c_n)\in\mathbb{F}_q^n$ and $x=(x_1,\ldots,x_k)\in\mathbb{F}_q^k$ are line vectors with always $n\geq k$. Thus $f(\mathbb{F}_q^k)=C$.

Remark

The bases of $\mathbb{F}_2^k$, $\mathbb{F}_2^n$ are the common canonical bases (those we have often used in the section of Vector Calculus).

Definition (#14.4.602): Let C be a linear code $[n,k]$ and given $\{g_1,g_2,\ldots,g_k\}$ the basis of C. A "generating matrix" G of C is therefore a matrix $n\times k$ whose columns are formed by the vectors g_i of the basis (see the example further below).

Given $u=(u_1,\ldots,u_k)$ the information word, in extenso the vector containing the k information symbols. Then we can write the matrix relation linking the code word c and the information word u by:

$$c=Gu \tag{14.4.39}$$

Definition (#14.4.603): Let C be a block code $[n,k]$. This code is named "systematic code", if the set of code words contains the k unmodified symbols of the original information (we will return on this type of code further below). The remaining $n-k$ symbols are named "parity symbols".

Remark

The "Hamming Code" is such a code! In addition, systematic codes are special cases of block codes and we will return to their study further below.

Definition (#14.4.604): Let H be an $(n-k)\times n$ matrix with elements in $\mathbb{F}_q$, which satisfies $Hc=0$ for every word c of a linear code C (in other words: whose kernel is C). Then, H is named the "control matrix" of the code C. Conversely, c belongs to the code if and only if $Hc=0$. Otherwise there is a mistake!

Remark

It is easy to find H because it is "orthogonal" to G since the above definition implies:

$$Hc=H(Gu)=(GH)u=0\Leftrightarrow GH=0 \tag{14.4.40}$$

of course we must not take $H=0$ in practice...

Let's see a companion example of all this with the Hamming code which is a systematic block code (caution!! it seem that there exist several definitions of a "Hamming code!"):

This method consists of doubling the information, sending as many parity bits as data bits. A first matrix for this purpose is:

$$G=\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\\1&1&0&0\\0&0&1&1\\1&0&1&0\\0&1&0&1\end{bmatrix} \tag{14.4.41}$$

The coding matrix G above is of dimension $n\times k$, where n is the number of bits received per packet, and k is the number of bits per message containing the information (here $n=8$ and $k=4$). It automatically generates the parity bits specific to a message. For example, in order to send the message 1101, in order to respect the matrix multiplication rule, consider this quartet as a column vector:

$$u=\begin{pmatrix}1\\1\\0\\1\end{pmatrix} \tag{14.4.42}$$

So by multiplying, we get:

$$c=Gu=\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\\1&1&0&0\\0&0&1&1\\1&0&1&0\\0&1&0&1\end{bmatrix}\begin{pmatrix}1\\1\\0\\1\end{pmatrix}=\begin{pmatrix}1\\1\\0\\1\\0\\1\\1\\0\end{pmatrix} \tag{14.4.43}$$

We will therefore send a byte 11010110, whose the first four bits form the message u and the last four bits the parity bits, which are used to check the veracity / integrity of the message.

The corresponding control matrix H is:

$$H=\begin{bmatrix}1&1&0&0&1&0&0&0\\0&0&1&1&0&1&0&0\\1&0&1&0&0&0&1&0\\0&1&0&1&0&0&0&1\end{bmatrix} \tag{14.4.44}$$

Thus, when the receiver receives the byte 11000110 instead of 11010110, the decoding gives as "syndrome":

$$s=Hc=\begin{bmatrix}1&1&0&0&1&0&0&0\\0&0&1&1&0&1&0&0\\1&0&1&0&0&0&1&0\\0&1&0&1&0&0&0&1\end{bmatrix}\begin{pmatrix}1\\1\\0\\1\\0\\1\\1\\0\end{pmatrix}=\begin{pmatrix}0\\1\\0\\1\end{pmatrix} \tag{14.4.45}$$

The resulting column vector is therefore not zero. So there is an error! With the control matrix, the theory (see proof below) makes it possible to assert that as the vector obtained is the same as that which is in fourth position in the decoding matrix, the error is due to the fourth bit. As we are in base 2, it is enough to change the 0 into a 1. This coding of the information is expensive, because it occupies twice as much bandwidth. However this is one of the most effective ways to secure information.

Theorem 14.29. The syndrome s of a Hamming code corresponds to one of the column of the control matrix H.

Proof 14.29.1. To show that the syndrome of a Hamming code corresponds to one of the columns of the control matrix, we denote by e_i the vectors-columns of the canonical basis on $\mathbb{F}_2^n$, $e_i=(0,\ldots,1,0,1,\ldots,0)$ with 1 in the i-th place. Given c a code word. We thus have by the definition of H: $Hc=0$. Let us suppose that the received word, which we will denote by $\tilde{c}$, is tainted by a single error and that this error is on the j-th bit. Therefore we have:

$$c+\tilde{c}=e_j \tag{14.4.46}$$

and:

$$He_j=H(c+\tilde{c})=Hc+H\tilde{c}=H\tilde{c} \tag{14.4.47}$$

Therefore it comes that:

$$He_j=H\tilde{c} \tag{14.4.48}$$

but He_j is the j-th column vector of the matrix H.

This shows us that when we $\tilde{c}$ and we compute $H\tilde{c}$ we get the column vector of the matrix H located exactly at the location of the error (in this case j).

■ Q.E.D.

Remark

A null syndrome does not mean the absence of error(s). There are therefore undetectable error configurations!!!

Let us now write:

$$A=\begin{bmatrix}1&1&0&0\\0&0&1&1\\1&0&1&0\\0&1&0&1\end{bmatrix} \qquad \text{and} \qquad \mathbb{1}_4=\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix} \tag{14.4.49}$$

Then we will notice that the matrices G and H of our companion example above are formed by the blocks $\mathbb{1}_4$ and A in the following way:

$$G=(\mathbb{1}_4|A)=(\mathbb{1}_4|A) \tag{14.4.50}$$

named the "(canonical) generator matrix of a linear (n,k) code", and:

$$H=(A \quad \mathbb{1}_k)=(A \quad \mathbb{1}_4) \tag{14.4.51}$$

The "parity-check matrix".

Therefore:

$$HG=(A \quad \mathbb{1}_k)\begin{pmatrix}\mathbb{1}_k\\A\end{pmatrix}=A\mathbb{1}_k+\mathbb{I}_kA=A+A=A(1+1)=0 \tag{14.4.52}$$

Because $1+1=0$ in $\mathbb{F}_2$.

In general, if we work with the alphabet $\mathbb{F}_2$ and if $G=\begin{pmatrix}\mathbb{1}_k\\A\end{pmatrix}$ where A is $(n-k)\times k$ matrix then $H=(A \quad -\mathbb{1}_k$ is also a control matrix because again:

$$HG=(A \quad -\mathbb{1}_k)\begin{pmatrix}\mathbb{1}_k\\A\end{pmatrix}=A\mathbb{1}_k-\mathbb{1}_kA=0 \tag{14.4.53}$$

Remark

In $\mathbb{F}_2$, we have $\mathbb{1}_k=-\mathbb{1}_k$, since $1=-1$ this is why we wrote $H=(A \quad -\mathbb{1}_k)$ in the previous companion example.

14.4.4.2 Systematic codes

As we have mention it earlier above, let us come back on "systematic codes" that we have already defined.

Constructing a systematic code consists as we already know in adding to each word $x=(x_1,\ldots,x_k)$ of the message $n-k$ symbols $(c_{k+1},\ldots,c_{k+n})$ linearly depending of the x_i to get the code word $c=f(x)$.

We know already that symbols are named "control bits" and (we will see another example just below):

$$c=(x_1,\ldots,x_k,c_{k+1},\ldots,c_{k+n})=(\mathbb{1}_k|A)(x_1,\ldots,x_k) \tag{14.4.54}$$

where, for recall, $(\mathbb{1}_k|A)$ denotes the matrix $n\times k$ obtained by writing one below the other, the identity matrix $\mathbb{1}_k$ of size k and any matrix A.

We will say that a code C is "systematic" if it has a generating matrix of the form $G=(\mathbb{1}_k|A)$.

✍**Example:**

We propose to construct a systematic linear code with $n=k=3$ as example. We will denote by a_1,a_2,a_3 the information bits. The control bits a_4,a_5,a_6 will be defined by:

$$\begin{aligned}a_4&=a_1+a_2\\ a_5&=a_2+a_3\\ a_6&=a_1+a_3\end{aligned} \tag{14.4.55}$$

The generating matrix G is such that its upper part is the identity matrix of dimension 3 (we had the same thing for the Hamming code). The first line (110) of the matrix A corresponds to the expression of the control bit a_4:

$$a_4=1\cdot a_1+1\cdot a_2+0\cdot a_3 \tag{14.4.56}$$

etc. For each control bit.

The generating matrix G is then written:

$$G=\begin{bmatrix}\mathbb{1}_k\\ A\end{bmatrix}=\begin{bmatrix}\mathbb{1}_3\\ A\end{bmatrix}=\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\\1&1&0\\0&1&1\\1&0&1\end{bmatrix} \tag{14.4.57}$$

By multiplying this matrix by the $2^3 = 8$ possible vectors (the words consisting of three bits of information), we get the following code words:

a_1	a_2	a_3	a_4	a_5	a_6
0	0	0	0	0	0
0	0	1	0	1	1
0	1	0	1	1	0
1	0	0	1	0	1
1	0	1	1	1	0
1	1	0	0	1	1
1	1	1	0	0	0

We thus find that the minimum weight of the code words is 3. Therefore the code detects $3-1=2$ errors and can correct of them $\left[\dfrac{3-1}{2}\right]=1$.

The reader interested can refer to out MATLAB® companion book to see how to handle or generate error correcting codes.

~60 %

★★★★★

11 votes, 58.18%

14.5. Automata Theory

THE purpose of this section is to study the theoretical aspect of the computer/machine concept. We are located here at the level of mathematics and logic, regardless of any reference to a real specific computer/machine (or software). We will look at how this theoretical machine will make acquisition of digital data, of whatever nature, to do treatment on it or to solve a general problem. We will then be taken to see that, from this point of view, any theoretical machine is reducible in its operating principle, to an ideal machine. Thus, we can say that all computers, or all programs are equivalent to each other, as the purpose of computer, in its theoretical definition, is universal, that is to say the capacity to treat all actually treatable problems.

Modern computing is the result of research undertaken in the early 120th century (holocene calendar) by Bertrand Russell and Alfred North Whitehead to constitute a formal mathematical system where any proposal could be proved by a logical calculation see section Proof Theory page 131. David Hilbert and Kurt Gödel accomplished decisive steps in the exploration of this program. In 11931 (holocene calendar) Gödel proved that (recall):

1. It may be that in some cases we can prove one thing and its opposite (inconsistency).
2. In any formal mathematical system there are mathematical truths that can not be proved (incompleteness)

Gödel's theorem thus ruin the dream to make from mathematics a perfectly coherent deductive system, but from the intellectual activity around the *Principia Mathematica* project of Russel and Whitehead will release the founding ideas of computer science. This brings Alan Turing in 11936 (holocene calendar), after Gödel, to tackle the problem of decidability.

Definition (#14.5.605): A system is named "decidable system" if there is an effective procedure for distinguishing provable proposals of others. To define more rigorously the concept of effective procedure, Alan Turing developed the concept of "automata", hereinafter named "Turing machine" (see example below), which allows him to clarify the concept of implementation of an "algorithm" (see section Numerical Methods page 6273).

Definition (#14.5.606): A "Turing Complete language" is a language with at least a conditional and a while-loop construct - such a language can be used to implement a Turing machine that is "powerful" enough to perform any realizable algorithm.

Inventing effective procedures (algorithms) is to determine a sequence of elementary operations that perform the calculations/treatments necessary to solve problems for which there are computable solutions (there are unsolved problems and incalculable solutions as we have seen during our study of complexity in the section of Theoretical Computing). Turing also proved that its calculation model is universal, that is to say that all Turing machines are equivalent (we will prove this below). It makes the assumption that any algorithm can be computed by a Turing machine. These ideas underpin the theory of computer programming.

These ideas underlying the theory of computer programming were one important factor of winning the second World War thanks to the machine developed by Turing to uncrypt the Nazi Enigma cypher machine (a romance movie has been made about this subject in 12015 (holocene calendar) with Alan Turing as major role).

Remark

This section should have normally been placed at the first position of this chapter but it seemed wiser to do first hand on concrete examples of theoretical computing before moving to the abstract formalism of their executions. This is one of the reasons why we will return here briefly on the concepts of algorithms, complexity, formal logical systems, proof theory and information (see all sections with the corresponding name). Furthermore, for this section, an experience in the development of computer software is a big plus for understanding certain concepts (or to imagine practical applications).

Before we begin, it should be useful for the reader to have a non-exhaustive overview of the applications of Language and Automata Theory: specification of programming languages, compilation, pattern matching (in a text, in a database on the web, in the genes .. .), text compression, program verification, electronic computers, encoding for transmission, encryption, decoding of the genome, language, cognitive science, etc.

14.5.1 Von Neumann machine

We due to John von Neumann to conceive in 11945 (holocene calendar) the general architecture of the concrete apparatuses which will carry out the calculations according to the Turing model, an architecture so efficient and elegant that the computers this early 121st century (holocene calendar) are still constructed on these principles.

Remark

In a way, we can say that this decade between 11936 and 11945, according to holocene calendar (corresponding well to the Second World War), saw the birth of computer science, which went from the mathematical and logical intellectual construction stage to the application of these ideas to the realization of concrete physical systems.

Here is the diagram of the of a von Neumann architecture:

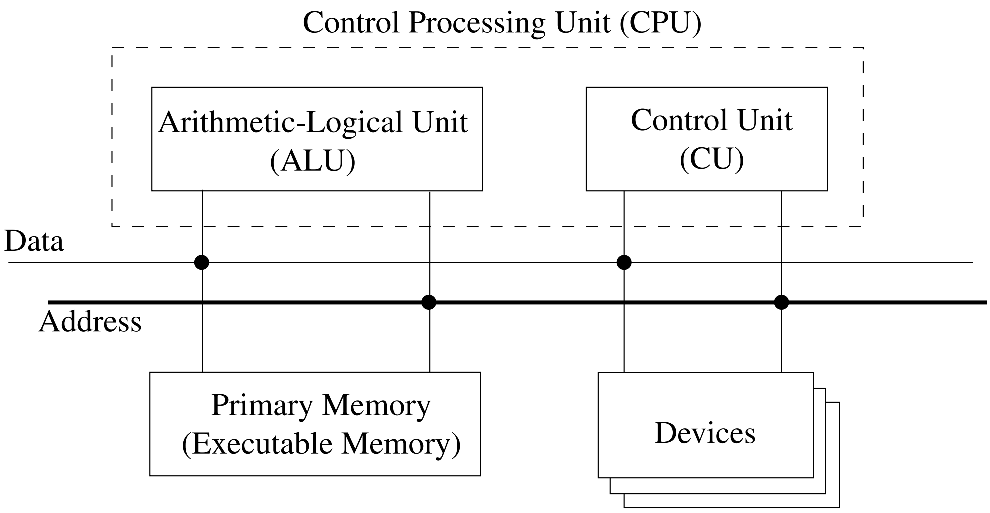

Control Processing Unit (CPU)
Arithmetic-Logical Unit (ALU)
Control Unit (CU)
Data
Address
Primary Memory (Executable Memory)
Devices

Figure 14.344: **Principle of the von Neumann machine**

The Control Units, Arithmetic-Logic Unit (ALU), and Primary Memory constitute all three the "Central Unit", or "Processor", of the computer. The processor consists of electronic circuits that can perform actions. The set of actions wired in the processor constitutes the instruction set of the processor and determines the basic language of its use, named "machine language".

The role of the Control Unit is to enable the desired action (instruction) to be triggered at the desired moment. This instruction can belong to the Arithmetic-Logic Unit, to the Memory Unit or to the Control Unit itself. An instruction can also consult the contents of the Primary Memory unit (the "read") or modify the contents of that memory (the "write"). In general, an action consists in either consulting or modifying the state of the memory or one of the registers (which are special memory elements incorporated in the central processing unit), or triggering an input-output operation (communication with the outside world and in particular a human user).

The memory is made up of elements that can take states. A basic element of the memory can take two distinct states and can be used to represent an elementary data item, or "bit" (see section Logical Systems 6275). This representation of a data element by a memory element is named a "code". A memory with many bits allows the coding of complex data, within the limit of the size of the memory.

The way in which the central unit, memory and I/O devices (input/output) communicate is generically named a "bus" (it is, in a way, the highway where data flows from one point to another Using addresses). In a somewhat formal way, a bus is a complete connected graph (see section Graph Theory page 3010), which means in common language that all the elements connected to the bus can communicate with one another.

Remark

The "code" makes bits and groups of symbols match together. The simplest symbols are numbers and letters. To represent complex data, you can define methods, rules for grouping symbols, and associate a data element with a symbol group constructed according to the rules.

Definition (#14.5.607): We will name "language" a set of symbols or groups of symbols, constructed according to certain rules, and which are the "words" of that language. A "language syntax" is the set of construction rules for language words.

The memory of the computer (this is the basic idea of von Neumann) contains information of two types: Programs and Data. The programs and the data are represented with the same symbols, only the semantics allows to interpret their respective texts. Moreover, the text of a program can sometimes be considered as data for another program, for example a program of translation from one language to another.

quote 14.5:7184.#78

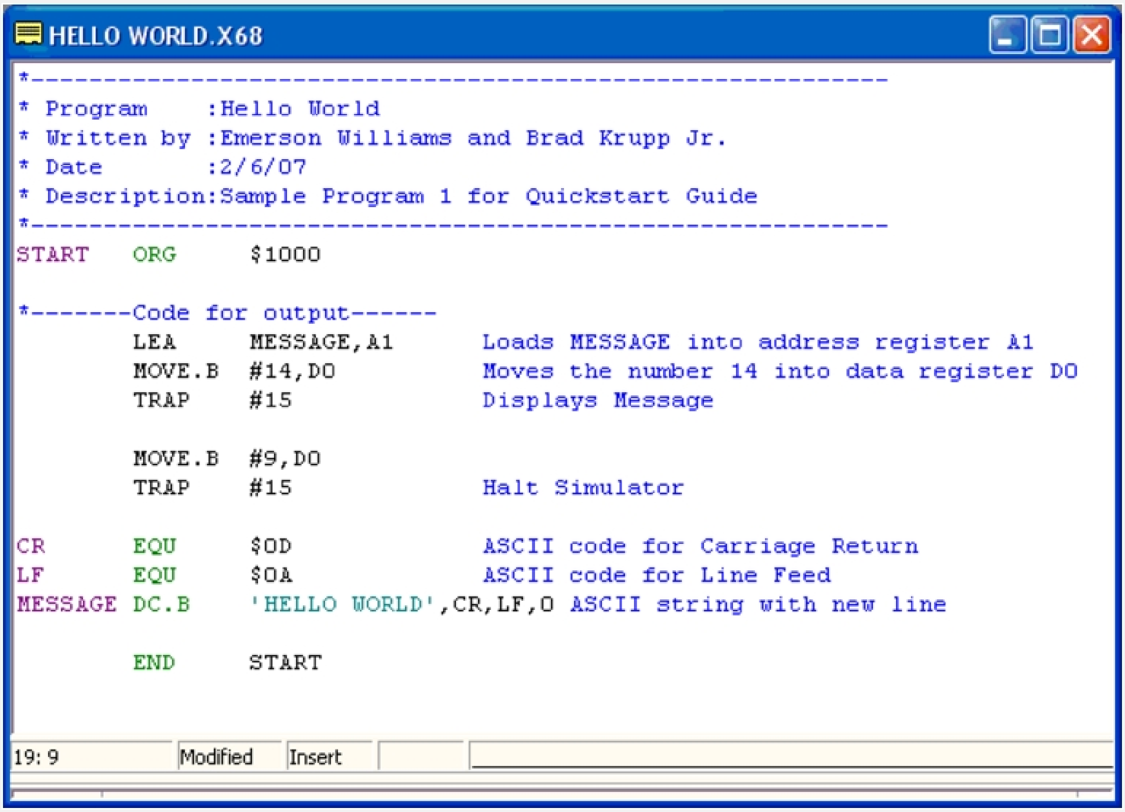
> *Computers are like humans - they do everything except think.*
>
> John von Neumann

14.5.2 Turing machine

It is important to be convinced (it will perhaps not be done in one day...) that all the programs we can write in different languages are equivalent!!! The Turing machine is a model of an automata whose description is very low-level (before going on to a much more formal definition). The von Neumann architecture, designed to efficiently perform the processes described by a Turing machine, generates imperative languages (see definition in remark R1 below). Any program, functional or imperative, intended to be executed, will be translated into an imperative language, the "machine language" of the computer used. The coherence of computer science and the semantic equivalence of programs written in various languages which ensure the validity of this operation are the result not of chance but of a common original theoretical conception. Gödel, Church, von Neumann and Turing were all in Princeton in 11936 (holocene calendar)...!

Remarks

R1. The first evolved languages that have appeared are so-named "imperative languages", based on the notion of the state of memory (it is the "assembly language" by the way!). These languages, inspired by John von Neumann's model, include, like machine language, instructions that produce changes in memory (assignment instruction). The writing of a program in imperative language consists in writing the sequence of instructions which will cause the successive states by which the memory will have to pass so that, starting from an initial state allowing the initialization of the program, it arrives in a state providing final results.



```
*-----------------------------------------------------------
* Program    :Hello World
* Written by :Emerson Williams and Brad Krupp Jr.
* Date       :2/6/07
* Description:Sample Program 1 for Quickstart Guide
*-----------------------------------------------------------
START   ORG     $1000

*-------Code for output------
        LEA     MESSAGE,A1      Loads MESSAGE into address register A1
        MOVE.B  #14,D0          Moves the number 14 into data register D0
        TRAP    #15             Displays Message

        MOVE.B  #9,D0
        TRAP    #15             Halt Simulator

CR      EQU     $0D             ASCII code for Carriage Return
LF      EQU     $0A             ASCII code for Line Feed
MESSAGE DC.B    'HELLO WORLD',CR,LF,0 ASCII string with new line

        END     START
```

Figure 14.345: **Assembly language "Hello World" demo**

R2. In addition to "computational languages", we distinguish in computer science "sequential languages", "interpreted languages", "description languages", "functional languages" and "compiled languages".

A formal model for an effective procedure (to describe an algorithm) must possess certain properties. First, each procedure must receive a finite definition. Secondly, the procedure must consist of separate steps, each of which must be capable of being accomplished mechanically. In its simplicity, the Turing machine composed of the following elements answers to this program:

1. An infinite memory represented by a ribbon divided into boxes. Each square of the ribbon may be given a symbol of the alphabet defined for the machine;
2. A reading head capable of traversing the tape in both directions;
3. A finite set of states among which we distinguish an initial state and the other states, named "accepting states"
4. A transition function which, for each state of the machine and each symbol under the read head, specifies: the next state, the character that will be written on the ribbon instead of the one that was under the head, the direction of the next playback of the playback head.

One can equip his Turing Machine with the finite alphabet of his choice. Its ribbon can be infinite in either direction or in one. It may even have several ribbons. It is shown that these various machines are equivalent.

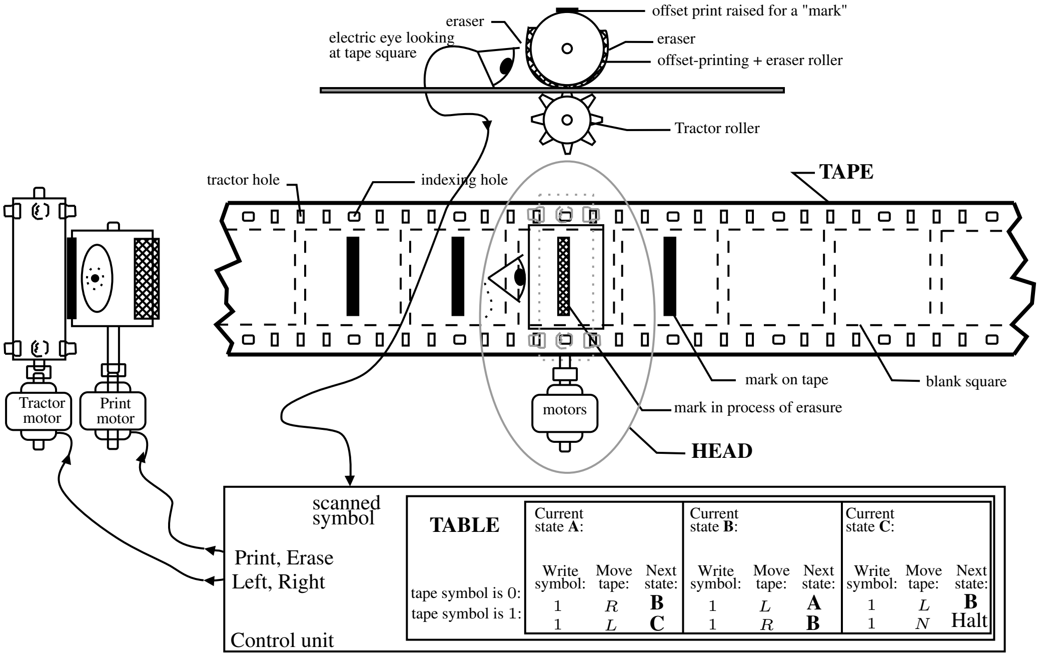

offset print raised for a "mark"
eraser
electric eye looking at tape square
eraser
offset-printing + eraser roller
Tractor roller
tractor hole
indexing hole
TAPE
Tractor motor
Print motor
motors
mark on tape
blank square
mark in process of erasure
HEAD
scanned symbol
Print, Erase
Left, Right
Control unit
TABLE

TABLE	Current state **A**:			Current state **B**:			Current state **C**:		
	Write symbol:	Move tape:	Next state:	Write symbol:	Move tape:	Next state:	Write symbol:	Move tape:	Next state:
tape symbol is 0:	1	R	**B**	1	L	**A**	1	L	**B**
tape symbol is 1:	1	L	**C**	1	R	**B**	1	N	Halt

A fanciful mechanical Turing machine's **TAPE** and **HEAD**. The **TABLE** instructions might be on another "read only" tape, or perhaps on punch-cards. Usually a "finite state machine" is the model for the **TABLE**.

Figure 14.346: **Turing Machine principle**

We are then led to the following simplistic definition:

Definition (#14.5.608): A "finite automate" is a mathematical model of systems having a finite number of states and that actions (external or internal) can move from one state to another. The external actions are represented by the symbols of an alphabet $\mathcal{A}$; The internal actions (invisible, silent, or spontaneous) are represented by a symbol not belonging to the aforementioned alphabet.

Caution! For example, any real computer is finite so it isn't an abstract model like a Turing Machine. The finite version of the Turing Machine has a confusing name: "Linear Bounded Automaton". This is basically what real computers are. However, LBA are not seen or used in discussion nearly as much as Turing Machine.

An automate is represented by a graph (see section Graph Theory page 2999) whose vertices are states and with each arc is associated the recognition of one or more letters.

Finite automata are used to model and control finite-state systems and to solve common problems: lexical analysis, search of patterns in text, genome analysis, etc.

✍**Examples:**

E1. A finite and deterministic automate which recognizes all integers whose writing is normalized (regular language), that is to say not starting with 0 (the numbers in the circles are just there to describe the order in which the controller performs the operation):

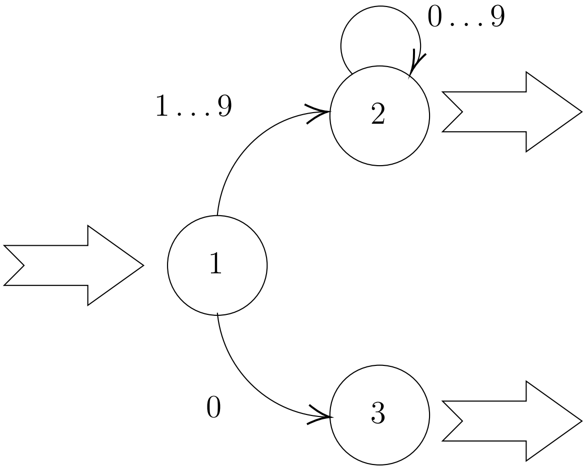


Figure 14.347: **First example of a finite automata (pedagogical version)**

Description: The automata receives an integer in input ①, it looks at whether this number starts with a 0 or it is a number between 1 and 9. If the number starts with zero, the controller exits and stop at ③. Otherwise, the controller goes to ② and analyses the numbers one after the other until it reaches the end and then stops and goes out at ③.

Here is a version of the same automata but respecting the traditional design and notation in academics:

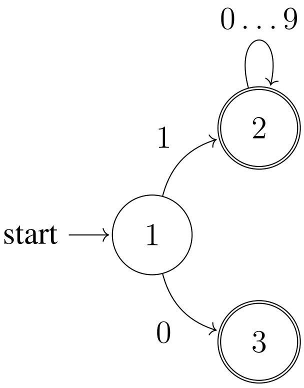


Figure 14.348: **Finite automata with academic official notation and design**

E2. A finite and deterministic automata which recognizes a numerical input in a regular language spreadsheet (for example: $+12,3$ or 08 or -15 or $5\text{E}12$ or $14\text{E}-3$):

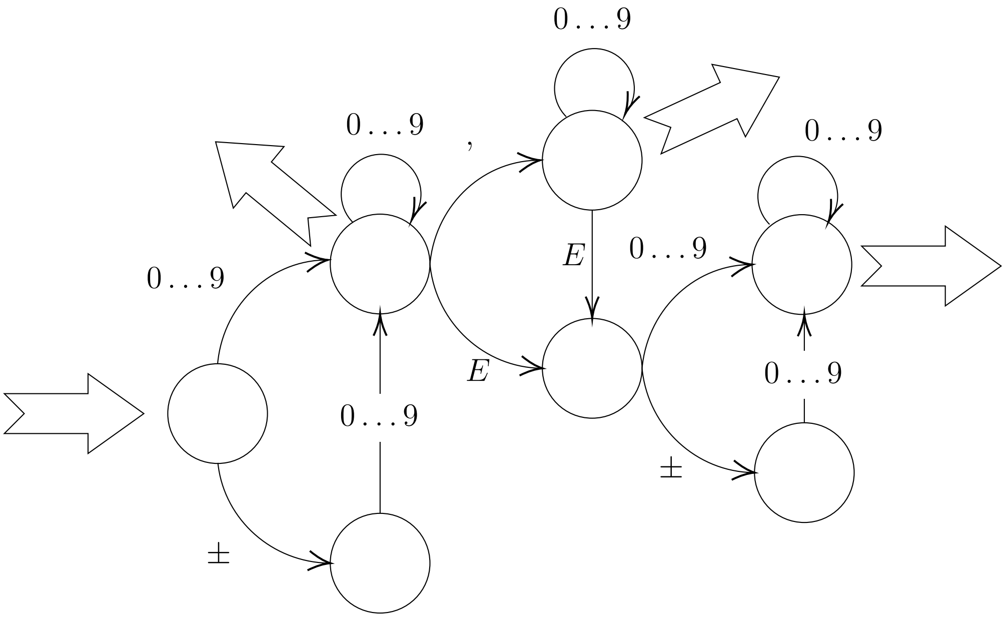


Figure 14.349: **Second example of a finite automata**

In other words, it is enough to recognize a language of the form:

$$\{-,+,\varepsilon\}\mathcal{A}^*,\mathcal{A}^*\mathrm{E}\{-,+,\varepsilon\}\mathcal{A}^* \tag{14.5.1}$$

which is indeed regular, where ε is the empty word, $\mathcal{A}$ is the alphabet $\{0,1,\ldots,9\}$ and equation the set of words (in extenso of numbers) that can be written with $\mathcal{A}$.

E3. A finite and deterministic automata recognizing all the multiples of 3, regular language type (in other words if such a multiple is found, the automata gives an output, otherwise nothing):

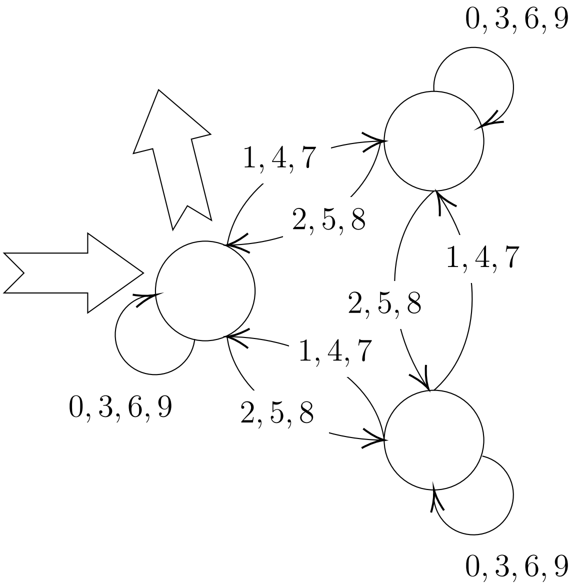


Figure 14.350: **Third example of a finite automata**

I strongly recommend any reader that want the be familiar with complete Turing Machines to practice the challenge that Google made with the Turing Doodle (it is probably "complete" but not sure):

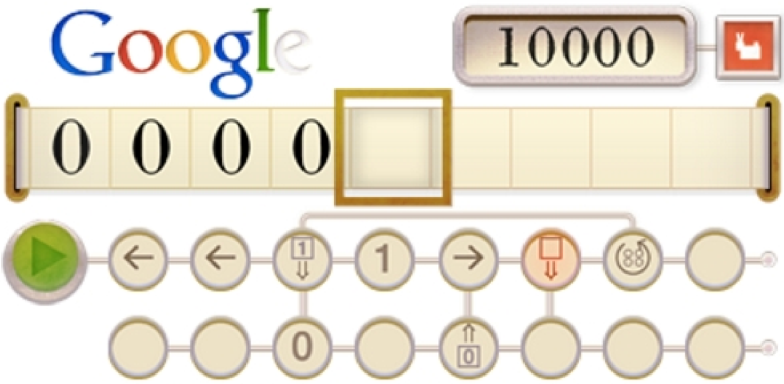


Figure 14.351: **Turing Doodle (source: Google)**

available here:

https://doodles.google/doodle/alan-turings-100th-birthday/

The purpose of that Doodle is to change the algorithm so that when it is executed and the machine stops the content of the tape is compared to the content of the display on the right top corner. When the comparison is successful the player will go to the next level with another algorithm to change... and so on...

Some passionate of Turing Machines have build real life small and beautiful electronic Turing machines as the one visible in the picture below (perhaps they build it as a hobby or sell it to schools for education purposes???):

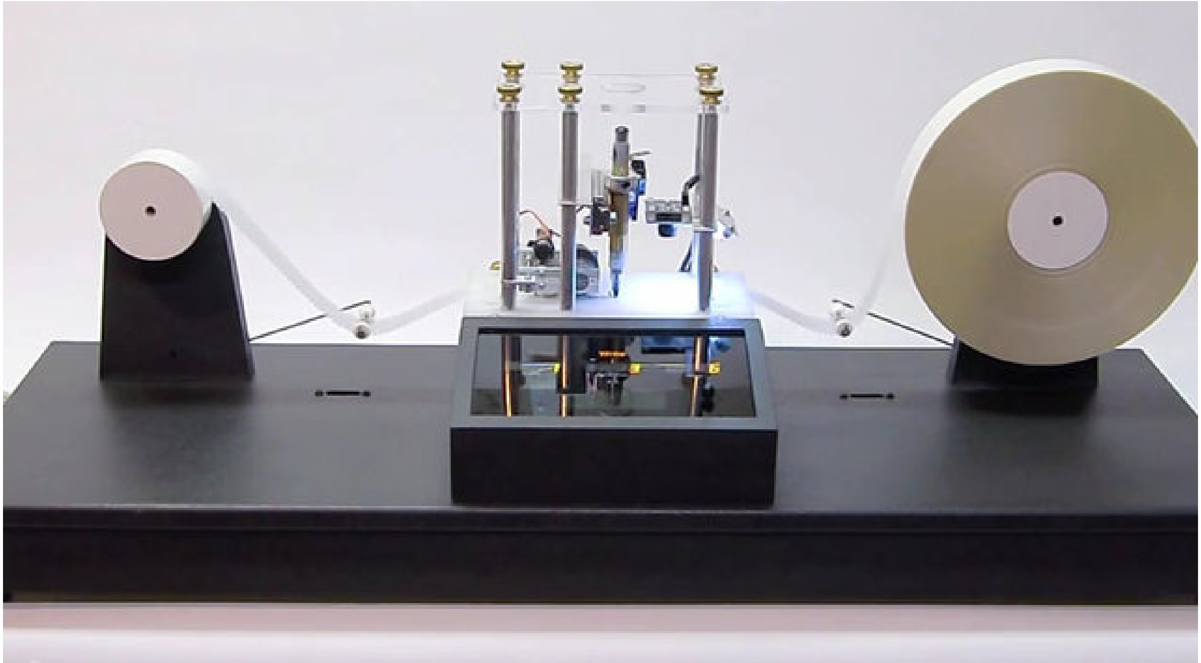


Figure 14.352: **Turing Machine (source: ?)**

Remarks

R1. Relatively to a common question...: YES! Artificial intelligence are Turing Machines. However at the day we write these lines (early 121st century according to holocene calendar), we still have no evidence based likelihood to claim that Homo Sapiens brain is a Linear Bounded Automaton or not (we let this work to future generations!).

R2. Markov chains can be represented by finite states machines. The main thing to keep in mind is that the transitions in a Markov chain are probabilistic rather than deterministic. which means that you can't always say with perfect certainty what will happen at time $t+1$. The brain is closer to a Markov chain than a Turing machine.

Because of a common request, let us provide a definition:

Definition (#14.5.609): "Consciousness" is the apparent (see [507]) fabricated awareness of internal and external existence by stimulus with the environment. Consciousness is tested scientifically by different means like the "Subject-Object Model test" and "Cellular/Basis for Consciousness" (for bacterias that are bio-sensitive and any bigger animal), the "Mirror test" (for animals having a vision suitable for such a test) or "Turing test" (for humans or machines like computers). General anaesthesia can switch off temporarily consciousness in most animals and death switch it off definitely.

The *Cambridge Declaration on Consciousness* - a scientific statement publicly proclaimed on July 7, 12012 (holocene calendar), at Churchill College, University of Cambridge, in which a group of leading neuroscientists - affirms that many non-human animals—including all mammals, birds, and several invertebrates such as octopuses—possess the neuroanatomical, neurochemical, and neurophysiological substrates of conscious states, meaning they are sentient and capable of experiencing positive and negative mental states.

14.5.3 Chomsky hierarchy

The "Chomsky hierarchy" is a classification of the languages described by the formal grammars proposed in 11956 (holocene calendar) by the linguist Noam Chomsky. It is widely used today in computing, especially for the design of interpreters or compilers, or for the analysis of natural languages.

It is necessary before to define certain concepts!

14.5.3.1 Formal language

Definition (naive version #610): In a broad range of contexts (scientific, legal, etc.) we designate naively by "formal language $\mathcal{L}$" a more formalized and more precise form of expression than the everyday natural language (the two do not necessarily go together).

In mathematics, logic and computer science, a formal language is formed by:

1. A set of words obeying to strict logical rules (formal grammar or syntax).
2. Possibly of an underlying semantics (the strength of formal languages is to be able to disregard such semantics, which makes theories reusable in several models).

Remark

Thus, while a particular payroll or inverse matrix calculation will always remain a payroll or inverse matrix calculation, a group theorem will apply both to the set of integers and to the transformations of the Rubik cube.

The formal language of a scientific discipline is therefore actually a language obeying strict formal syntax and which will serve to expose statements precisely, if possible concisely and without ambiguity, and is in opposition to natural language.

Formal language has the advantage of making easy the manipulation and transformations these statements. Indeed, we will generally have precise transformation rules (development of logical formulas, normal forms, contrapositions, commutativity, associativity, etc.) that can be applied without even knowing the meaning of the statements to be transformed or the meaning of the transformation. It is therefore a powerful exploration tool, and it is the only language that allows machines to "do mathematics".

The disadvantage is obvious: not knowing the meaning of the statement prevents us from knowing which are the relevant transformations and hurts the intuition of the reasoning. Thus, it is good to know how to quickly read a statement in formal language and to translate it just as quickly into one or more natural language statements.

This is where the limit of what we name "proof-aid software" lies: of course, the computer has (for now...) no intuition. The skill of the designer of such a program is to find ways for the computer to understand.

Giving relevant meaning to a programming language in order to run its programs is relatively easy, because these formal languages have been designed to mean sequences of elementary actions of the machine. To prove a program (to prove that the algorithm ends in a finite number of times) or a mathematical theorem (which is almost the same thing), there is, on the other hand, no infallible method, the correction of a program being an undecidable decision problem. Thus, the prover must simply apply certain heuristics (a technique consisting in learning little by little taking into account what has been done beforehand) and often calling for help to the human user (same as humans do in fact...!). However, thanks to its heuristics and computing power, the computer explores thousands of ways that the human user would not have been able to test in several years, thus accelerating the work of the mathematician, physicist or engineer.

Definition (#14.5.611): As an object of study, a "formal language $\mathcal{L}$" is defined as a set $\mathcal{W}$ of finite-length words w_i (i.e. strings) deduced from a certain finite alphabet $\mathcal{A}$, that is to say a free monoid (the set of words formed on an alphabet, provided with the internal law of concatenation - which is a law of composition - is a monoid which we name "free monoid", which empty word is the neutral element) on this alphabet.

Definition (#14.5.612): A "word" is a sound, sign, letter, or group of sounds, signs, letters, used to designate objects, to express actions, feelings, ideas, as well as their relationships.

Remark

Thus, while a particular payroll or inverse matrix calculation will always remain a payroll or inverse matrix calculation, a group theorem will apply both to the set of integers and to the transformations of the Rubik cube.

Definition (#14.5.613): A "rich language" is a language with a vast lexicon (almost a bijective relationship between all words and all the concepts). That means a language where each word or statement language minimize the number of possible contradictory or fuzzy interpretations. An "exact language" is the extreme case of a rich language where each word can have one and only one meaning (i.e. there is a bijective relationship between the set of words and the set of descriptions).

Some languages and their cardinality (source `https://www.sketchengine.eu/documentation/tenten-corpora/`):

Corpus	Language	Year	Size (tokens)	Number of unique words
SketchEngine enTenTen21	English	12021	$61,585,997,113$	$131,222,868$
SketchEngine deTenTen23	German	12023	$19,993,027,721$	$105,954,872$
SketchEngine ruTenTen20	Russian	12020	$23,968,214,880$	$88,999,479$
SketchEngine esTenTen23	Spanish	12023	$33,135,276,527$	$64,959,073$
SketchEngine frTenTen23	French	12023	$27,111,801,657$	$60,213,633$
SketchEngine prTenTen23	Portuguese	12023	$20,044,773,021$	$43,305,805$
SketchEngine itTenTen20	Italian	12020	$14,514,566,714$	$33,582,719$
SketchEngine arTenTen24	Arabic	12024	$7,592,287,486$	$27,380,266$
SketchEngine jaTenTen11	Japenese	12011	$10,321,875,664$	$13,884,199$
SketchEngine hiTenTen21	Hindi	12021	$901,352,786$	$4,287,476$

Table 14.56: **Some languages in descending order of richness**

In the table above the years are given on the basis of the holocene calendar as always in this textbook.

14.5.3.2 Syntax

Definition (#14.5.614): The "syntax" is the branch of linguistics that studies the way in which "free morphemes" (words) combine to form "syntagmas" (nominal or verbal) that can lead to propositions that can combine in turn to form statements.

✍**Example:**

The syntagma (sentence): *a modest house of red bricks* is encompassed in the upper syntagma, that is, the complete sentence. But this same sentence *a modest house of red bricks* includes among its elements, the lower syntagma *of red bricks*, complement of the name house.

Definitions (#14.5.615):

D1. In grammar school, a "proposition" is a syntagma articulated around a verb. This notion is mainly used in language learning.

D2. A "statement" in linguistics is everything that is pronounced by a speaker between two breaks. Syntactically, the statement can therefore extend from the simple word to the length of a sentence (even to a discourse), through the syntagma.

The term "syntax" is also used in computer science, where its definition is similar, modulo a different terminology... Thus the syntax is the respect, or the non-respect, of the formal grammar of a computer language, that is to say of the rules of arrangement of the lexemes (which in computer science are only lexical entities) in more complex terms, often: "programs". In the theory of formal languages, what plays the role of lexeme is usually named "letter" or "symbol", and the product terms are named "words".

From a purely grammatical point of view, the study of syntax concerns three kinds of units:

- The "sentence", which is the upper limit of the syntax.
- The "word", which is its basic constituent, sometimes named "terminal element"
- The "syntagma", which is its intermediate unit

The syntactic relations between these different units can be of two kinds:

- The "coordination" when the elements are of the same status
- The "subordination" in the opposite case (when there is subordination, the subordinate element fulfils a syntactic function determined with respect to the higher level unit)

The study of syntax will take into account, in particular, the nature (or category or species) of the words, their form (morphology) and their function. It is why we speak more generally of "morphosyntactic relations".

14.5.3.3 Grammar

Definition (naive version #616): A "formal grammar $\mathcal{G}$" is a formalism used to define a syntax and therefore a formal language $\mathcal{L}$, that is to say a set $\mathcal{W}$ of words w_i on a given alphabet $\mathcal{A}$.

The concept of formal grammar is particularly used in the following fields:

- Programs compilation (syntactic analysis)
- The analysis and processing of natural languages
- Calculation models (automata, circuits, Turing machines, etc.)

To define a grammar, we need (see the example below to understand):

1. An alphabet of non-terminal items
2. An alphabet of terminals item;
3. An initial symbol (the axiom) taken among the non-terminals items;
4. A set of production rules.

✍**Examples:**

E1. We can define arithmetic expressions in the following way (writings that we often find in the Proof theory) where | is the symbol commonly use for the logical "OR":

$$\texttt{exp}\rightarrow\texttt{exp+exp}|\texttt{exp*exp}|\texttt{(exp)}|\texttt{num}$$

where `exp` means "expression" or:

$$\texttt{num}\rightarrow 0\texttt{num}|1\texttt{num}|2\texttt{num}|\ldots|9\texttt{num}|1|2|\ldots|9$$

The non-terminals here are explicitly `exp` and `num`, the terminals are `+`, `*`, `(`, `)` and the digits. The axiom is `exp`.

E2. The syntax of classical propositional logic can be defined in the following way (see section Proof Theory page 168):

$$\mathcal{F}=\text{Atom}|F\vee F|F\wedge F|F\rightarrow|F\neg F|\exists xF|\forall xF$$

The most commonly used types of grammars are:

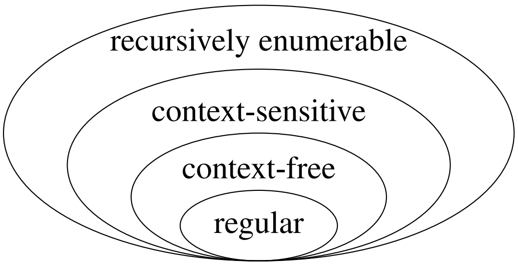
1. Left linear grammars that produce the same languages as regular expressions (this is what we are interested to in this section)
2. Context-free grammar (example above)
3. Contextual grammars (this type of grammar requires a mathematical formalism and can not be defined without it)

A language is therefore a set of words $\mathcal{W}$, which are simply sequences of symbols chosen from a finite alphabet set $\mathcal{A}$. The languages of the Chomsky's hierarchy consist of words that respect a particular formal grammar. What distinguishes them within the framework of classification is the nature of the grammar.

Remark

Most often, the symbols that are considered are formed of several characters, so that they correspond rather to what "words" in the current language. When there is ambiguity, for example in lexical analysis (vocabulary) and syntactic analysis (part of the grammar that deals with the function and the disposition of words and propositions in the sentence), we speak of "characters" for the symbols of the alphabet used to encode the information, and of "lexemes" for the symbols of the abstract alphabet, which are the basic units of the language. Similarly, the "words" of the language correspond rather to "sentences" or "texts".

The Chomsky hierarchy consists of the following 4 levels, from the most restrictive to the broadest one:



- L1. The "languages of type 3" or "regular language": these are the languages defined by a regular grammar or a regular expression, or the languages recognizable by a finite-state automata.
- L2. The "languages of type 2" or "algebraic languages" also named "context-free languages": these are the languages defined by a context-free grammar, or the languages recognizable by a non-deterministic stack automata. In this category are for example the computer programming languages.
- L3. The "languages of type 1" or "context sensitive languages": these are the languages defined by a contextual grammar, or the languages recognizable by a non-deterministic Turing machine with a length bounded by a fixed multiple of the word length's (these types of languages require a mathematical formalism and can not be defined without it).

L4. The "languages of type 0", or "recursively enumerable languages": This set includes all languages defined by a formal grammar. It is also the set of languages acceptable by a Turing machine (which is allowed to loop on a word that is not of the language).

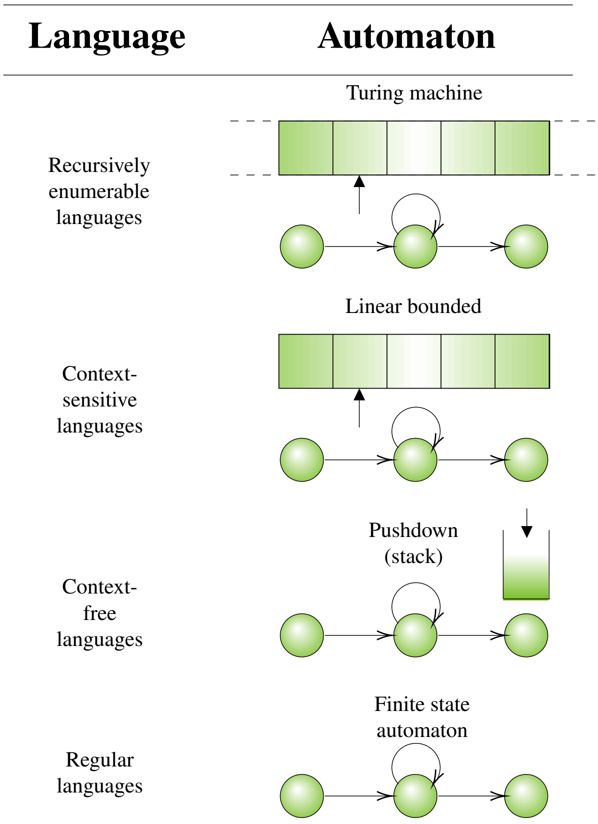

Language
Automaton
Turing machine
Recursively enumerable languages
Linear bounded
Context-sensitive languages
Pushdown (stack)
Context-free languages
Finite state automaton
Regular languages

Figure 14.353: **Chomsky hierarchy**

Remarks

R1. In addition to the 4 types of the Chomsky hierarchy, there are remarkable intermediate classes! For example between types 3 and 2: deterministic non-contextual languages, recognizable by a deterministic stack automata and languages between levels 1 and 0: recursive languages, that is to say, recognizable by a Turing machine (the latter must refuse words which are not in the language).

R2. The 4 types of languages and 2 of intermediate languages above are strictly included in each other.

A parser for a formal language is a computer program that decides whether a given input word belongs or not to the language, and possibly constructs a derivation of it.

There are systematic methods for writing type 2 or 3 language analysis programs (parsers). Interpreters or compilers almost always include a phase of lexical analysis, which consists of recognizing type 3 languages, followed by a phase of syntactic analysis that is a in fact a type 2 language analysis.

We can now finally in a vulgarize way (always with the aim of paving the way) define what an automata is in the Chomsky hierarchy.

14.5.3.4 Associated automata

Definition (naive version #617): In the field of theoretical computing, an "automata" is a machine to process information by a formal model (a Turing machine) on a given language. So:

- On a "finite language" (language containing a finite number of words), the associated automata is a machine comparing a text with that which is stored in a read-only memory. The grammar associated with a finite language is a list of the words of the language.
- On a "regular language" (language where the syntactic correction is verified by storing only a finite number of information), the associated automata is the "deterministic finite automaton" (that is, for each word entered, there is only one possible path of the graph) or the "non-deterministic finite automata". The grammar associated with a regular language is a left linear grammar.
- On an "algebraic language" (language where the main syntactic constraint are the parenthesis), the associated automata is the "pushdown (stack) non-deterministic automata". The associated grammar is the algebraic grammar.
- On a "bounded language" (description requiring a mathematical formalism), the associated automata is the "linearly bounded automata". The associated grammar is the contextual grammar.
- On a "decidable language" (an intelligent being manages to find a process to know whether or not one he is in the language), the associated automata is a Turing machine that stops on all data. There is no grammar associated to it.
- On a "semi-decidable language" (an intelligent being manages to find a process to know whether or not one he is in the language), the associated automata is the Turing machine (thus contains conditional structures and loops). The associated grammars are the "semi-Turing grammar", the "de Vangarden grammar" or the "affixed grammars".

14.5.4 Terminology

The automata therefore work mainly on letters, words, sentences and languages. In order to construct rigorous and optimal analysis methods and treatment on the subject it is interesting to formalize the treated objects. This is what we will do initially by defining these latter and their mathematical properties (which are very intuitive).

14.5.4.1 Words

Definitions (#14.5.618):

D1. An "alphabet $\mathcal{A}$" is a set whose elements are the "letter ℓ". The alphabets are always supposed to be finished.

D2. A "word w_i" is a finite sequence of "letters" which we denote by juxtaposition:

$$w_i = \ell_1 \ell_2 \ldots \ell_n \qquad \ell_i \in \mathcal{A} \tag{14.5.2}$$

D3. The "empty word", denoted ε, is the only word composed of no letters.

D4. The "length" of a word w is the number of letters that compose it, and is denoted $|w|$ (the empty word ε is the only word of length 0).

The "concatenation product" of two words $w_1 = a_1 a_2 \ldots a_n$ and $w_i = b_1 b_2 \ldots b_m$ is the word $w_1 w_2$ obtained by juxtaposition (concatenation):

$$w_1 w_2 = a_1 a_2 \ldots a_n b_1 b_2 \ldots b_m \tag{14.5.3}$$

Of course (trivial), we have:

$$|w_1 w_2| = |w_1| + |w_2| \tag{14.5.4}$$

We denote by $\mathcal{A}^*$ the set of words on $\mathcal{A}$.

✍**Example:**

Genes are words on the ACGT alphabet, proteins are words on a 20-letter alphabet. The natural integers, written in base 10, are words on the alphabet of the decimal digits...

Let $\mathcal{A}$ be an alphabet. Let $\mathcal{B}$ be a subset of $\mathcal{A}$. For any word $w \in \mathcal{A}$, the length in $\mathcal{B}$ of w is the number of occurrences of letters of $\mathcal{B}$ in the word w. This number will be denoted $|w|_{\mathcal{B}}$.

Remark

In particular, we have trivially $|w| = |w|_{\mathcal{A}}$.

For every letter $\ell \in \mathcal{A}$, $|w|_\ell$ is the number of occurrences of ℓ in w. We have:

$$|w|_{\mathcal{B}} = \sum_{b \in \mathcal{B}} |w|_b \tag{14.5.5}$$

D5. Given $w=\ell_1\ell_2\dots\ell_n$, with $\ell_1\ell_2\dots\ell_n\in\mathcal{A}$. The "mirror word" of w is the word denoted $\widetilde{w}$ defined by:

$$\widetilde{w}=\ell_n\dots\ell_2\ell_1 \tag{14.5.6}$$

Obviously:

$$\widetilde{(uv)}=\widetilde{v}\widetilde{u} \qquad \text{and} \qquad \widetilde{(\widetilde{w})}=w \tag{14.5.7}$$

D6. A word u is a "prefix" or "left factor" of a word v if there is a word x such that $ux=v$. The word u is moreover a "strict prefix" or "eigen-prefix" if $u\neq v$. Symmetrically, u is a "suffix" or "right factor" of v if $xu=v$ a word x. If $u\neq v$, then u is "strict suffix" or "eigen-suffix". The number of prefixes of a non-empty word v is $1+|v|$ (the empty word always being a prefix, we always have any non-empty word that has at least the empty word as a prefix).

✍**Example:**

The word $w=aabab$ on $\mathcal{A}=\{a,b\}$ has 12 different possible factors:

$$\{\varepsilon,a,b,aa,ab,ba,aab,aba,bab,aaba,abab,aabab\} \tag{14.5.8}$$

Lemma 14.29.1 (Levy's lemma)**.** Given x, y, z, t be words such as $xy=zt$. Then there exists a word w such that:

$$xw=z \tag{14.5.9}$$

with obviously:

$$y=wt \tag{14.5.10}$$

or:

$$x=zw \tag{14.5.11}$$

with also by extension:

$$wy=t \tag{14.5.12}$$

It results logically in particular that if $|x|=|y|$, the word w is empty, and therefore $x=z$ and $y=t$. In other words:

Theorem 14.30. A free monoid (see the reminder below) can be simplified on the left and on the right.

Proof 14.30.1. Let us put:

$$x=a_1\dots a_n \qquad y=a_{n+1}\dots a_m \tag{14.5.13}$$

with $a_i\in\mathcal{A}$, similarly:

$$z=b_1\dots b_p \qquad t=b_{p+1}\dots b_q \tag{14.5.14}$$

with $b_i \in \mathcal{A}$.

As:

$$xy = zt \tag{14.5.15}$$

We have:

$$m = q \tag{14.5.16}$$

(but not necessarily $n = p$) and:

$$a_i = b_i \tag{14.5.17}$$

for $i = 1, \ldots, m$ so that:

$$z = a_1 \ldots a_p \qquad \text{and} \qquad t = a_{p+1} \ldots a_m \tag{14.5.18}$$

If $|z| = p \leq n = |x|$, let us put $w = x_{p+1} \ldots x_n$. Therefore:

$$x = zw \qquad \text{and} \qquad wy = t \tag{14.5.19}$$

If $|z| > |x|$, let us put $w = x_{n+1} \ldots x_p$. Therefore:

$$xw = z \qquad \text{and} \qquad y = wt \tag{14.5.20}$$

■ Q.E.D.

Let us just do a recall of a parallel of the section of Set Theory... In the framework of the study automata a "free monoid" is a set $\mathcal{A}$ (the alphabet), whose elements are the letters ℓ_i. Therefore the composition law $\cdot$ is denoted:

$$(\mathcal{A}, \cdot) \text{ is a free monoid if: } \begin{cases} \cdot \text{ is associative (concatenation)} \\ \cdot \text{ has a neutral element } \varepsilon \text{ in } \mathcal{A} \end{cases} \tag{14.5.21}$$

In the Set Theory section, we were simply talking about the concept of "monoid". The monoid $(\mathcal{A}, \cdot) = (\mathcal{A}, \text{concatenation})$.

14.5.4.2 Languages

The subsets of $\mathcal{A}$ are named "formal languages". For example, for $\mathcal{A}=\{a,b\}$, the set $\mathcal{A}^+=\{a^n b^n | n \geq 0\}$ is a language.

We define on the languages several operations. The set operations are the union, intersection, complementarity and the resulting difference (see section Set Theory page 393). If X and Y are two parts of $\mathcal{A}^*$ then for recall each of this operation is defined by:

- Union:

$$X \cup Y = \{z \in \mathcal{A} | z \in \mathcal{A}^* | z \in X \text{ or } z \in Y\} \tag{14.5.22}$$

- Intersection:

$$X \cap Y = \{z \in \mathcal{A} | z \in \mathcal{A}^* | z \in X \text{ and } z \in Y\} \tag{14.5.23}$$

- Complementarity:

$$X \backslash Y = X \cap Y^C = \{z \in \mathcal{A}^* | z \in X \text{ and } z \notin Y\} \tag{14.5.24}$$

- Difference:

$$X \backslash Y = X \cap Y^C = \{z \in \mathcal{A}^* | z \in X \text{ and } z \notin Y\} \tag{14.5.25}$$

The also have for operation the product (of concatenation) of two languages X and Y is the language:

$$XY = \{xy | x \in X \text{ and } y \in Y\} \tag{14.5.26}$$

and we have for recall:

$$X\{\varepsilon\} = \{\varepsilon\}X = X \tag{14.5.27}$$

and also the operation of left quotient of Y by X:

$$X^{-1}Y = \{w | \exists u \in X \text{ and } v \in Y | v = uw\} \tag{14.5.28}$$

✍**Example:**

Let us consider three languages:

$$X = \{0,1,2,3,4\} \qquad Y = \{5,6,7,8,9\} \qquad Z = \{0,1\} \tag{14.5.29}$$

Then we have for the union:

$$X \cup Y = \{0,1,2,3,4,5,6,7,8,9\} \tag{14.5.30}$$

for the concatenation:

$$XY = \{05,06,07,08,09,15,16,17,18,19,25,26,27,28,29,35,\ldots\} \tag{14.5.31}$$

for a given quotient:

$$Z^{-1}(XY)=\{5,6,7,8,9\}=Y \tag{14.5.32}$$

another quotient:

$$Z^{-1}X=\{\varepsilon\} \tag{14.5.33}$$

and a last quotient:

$$Y^{-1}X=\varnothing \tag{14.5.34}$$

and a stupid difference example:

$$(X\cup Y)\backslash\{\varepsilon\} \text{ is the set of all decimal numbers} \tag{14.5.35}$$

Following the request of a reader we will detail:

$$Z^{-1}(XY)=\{5,6,7,8,9\}=Y \tag{14.5.36}$$

by recalling the definition:

$$X^{-1}Y=\{w|\exists u\in X \text{ and } v\in Y|v=uw\} \tag{14.5.37}$$

We then have explicitly:

$$\{0,1\}^{-1}\cdot\{05,06,07,08,09,15,16,17,18,19,25,26,\ldots\} \tag{14.5.38}$$

and in the concatenation product of the two languages X and Y, the only words where we find the elements $\{0,1\}$ of the language Z as a prefix are:

$$\{05,06,07,08,09,15,16,17,18,19\} \tag{14.5.39}$$

and as by definition $Z^{-1}(XY)$ is the unique set of terms w which follow the prefixes that constitute the terms of Z, then there remains only:

$$\{5,6,7,8,9\} \tag{14.5.40}$$

We have the following properties:

P1. Obviously:

$$X(Y\cup Z)=XY\cup XZ \tag{14.5.41}$$

P2. Less obvious:

$$X(Y\cap Z)\subset(XY\cap XZ) \tag{14.5.42}$$

where the inclusion is generally strict. To conceptualize this property, we must not forget that X is a set of words and that Y, Z do not necessarily have words of the same length!

The powers of X are defined by:

$$X^0=\{\varepsilon\}, X^1=X \quad \text{and} \quad X^{n+1}=X^n \tag{14.5.43}$$

for $n\geq 1$

In particular, if $\mathcal{A}$ is an alphabet, $\mathcal{A}^n$ is the set of words of length n.

Definitions (#14.5.619):

D1. The "Kleene star" (or "Kleene operator" or "Kleene closure") of X is the set:

$$X^*=\bigcup_{n\geq 0} X^n=\{x_1\ldots x_n|n\geq 0, x_1\ldots x_n\in X\} \tag{14.5.44}$$

✍**Example:**

Given $X=\{a,ba\}$. The words of X^*, classified by length are:

0	ε
1	a
2	aa, ba
3	aaa, aba, baa
4	$aaaa, aaba, abaa, baaa, baba$
...	

D2. The operator "$+$" is defined similarly:

$$X^+=\bigcup_{n>0} X^n=\{x_1\ldots x_n|n>0, x_1\ldots x_n\in X\} \tag{14.5.45}$$

14.5.4.3 Equations

First let us see a little something we will need later: given u and v two non-empty words. The three following conditions are equivalent (without proof because quite trivial):

C1. $uv=vu$

C2. $\exists n,m>1: \quad u^n=v^m$

C3. $\exists w\neq\varepsilon, k,l\geq 1: \quad u=w^k, v=w^l$

Remark

Let us recall again that we do not necessarily have $|u|=|v|$ but that we can very well have $|u|>|v|$.

Let us now turn to interesting things (some fuzzy points of the section of Proof Theory can be clarified here sometimes...)!

Definitions (#14.5.620):

D1. Let $\mathcal{V}$ and $\mathcal{A}$ be two disjoint alphabets (you can imagine them as the set of variables and respectively of the constants for example...). An "equation in words" with constants on $\mathcal{A}$ is a couple $e=(\alpha,\beta)$ of words $(\mathcal{V}\cup\mathcal{A})^*$. Such an equation is represented by $\alpha=\beta$. It is therefore necessary to see the two chosen words as the left and right members respectively of an equation.

D2. An equation is said to be "non-trivial equation" if $\alpha\neq\beta$.

✍**Example:**

Given $\mathcal{V}=\{x\}$ and $\mathcal{A}=\{a\}$ and let us define:

$$\alpha=ax \qquad \beta=xa \tag{14.5.46}$$

then we have the following equation in words:

$$xa=ax\Leftrightarrow\alpha=\beta \tag{14.5.47}$$

D3. An equation e is said to be an "equation without constant" if $\alpha,\beta\in\mathcal{V}^*$.

D4. A "solution" of the equation e is a monoid homomorphism (see section Set Theory page 439):

$$h:\,(\mathcal{V}\cup\mathcal{A})^*\mapsto\mathcal{A}^* \tag{14.5.48}$$

invariant (because every letter on $\mathcal{A}$ is sent on $\mathcal{A}^*$ and therefore every word on $\mathcal{A}$ is sent on $\mathcal{A}$) on $\mathcal{A}$ such that:

$$h(\alpha)=h(\beta) \tag{14.5.49}$$

Remark

Let us recall that the definition of the homomorphism is such that if $\alpha=xy$ then:

$$h(\alpha)=h(xy)=h(x)h(y) \tag{14.5.50}$$

✍**Example:**

Given $\mathcal{A}=\{a,b\}$ and $\mathcal{V}=\{x,y\}$. Let us wow consider the following words:

$$\alpha=xab \qquad \beta=byx \tag{14.5.51}$$

let us define h such that it sends x on b, y on a, a on a, b on b. Therefore we have well:

$$h(\alpha)=h(xab)=bab \qquad h(\beta)=h(byx)=bab \tag{14.5.52}$$

and we will always have for every couple:

$$(\alpha,\beta)\rightarrow h(\alpha)=h(\beta) \tag{14.5.53}$$

D5. A solution h is said to be a "cyclic solution" if there exists a word w (belonging to $\mathcal{A}$) such that $h(x)\in w^E$ (considering the word itself as an alphabet therefore) for any variable x.

14.5.4.4 Codes

Definition (#14.5.621): We name "code" any part $\mathcal{C}$ of a free monoid $\mathcal{A}^*$ that satisfies the following condition for any (word) $x_1,\ldots,x_n,y_1,\ldots,y_m\in\mathcal{C}$:

$$x_1x_2\cdots x_n=y_1y_2\cdots y_m \Longrightarrow n=m \quad \text{and} \quad x_i=y_i \text{ for all } 1\leq i\leq n \tag{14.5.54}$$

In other words, $\mathcal{C}$ is a code if every word of $\mathcal{C}^*$ (word composed of words) uniquely factorize into a product words of $\mathcal{C}$. When a set is not a code, general we see it quite easily.

✍**Examples:**

E1. The set (language) $\{a,ab,ba\}$ is not a code since the word aba can written both as product $a\cdot ba$ and as product $ab\cdot a$.

Remark

The simplest codes are the "uniform codes". These are sets whose all words have the same length (which means that since each word is different, the combination of words can hardly differ).

E2. The set $\mathcal{A}^*$ of the words of length n is a code, if $n\geq 1$. The ASCII code that associates to some characters binary words of length 7 (see ASCII table) with some characters is an example of uniform code.

E3. The base of the DNA $\{A,T,C,G\}$ is a uniform code of length 1. However what defines DNA are the sequences of pairs, and the only canonical pairs of DNA should be $\{AT,TA,GC,CG\}$ that is also a uniform code of length 2. However no human DNA ever has been observed as perfect, incorrect bases are always present so that we can observe the following pairs in reality $\{AT, TA, GC, CG, AA, GG, AG, GA, CC, TT, CT, TC, AC, CA, GT, TG\}$ (i.e. all combinations), that is a uniform code of length 2.

During protein synthesis (translation), the cell reads messenger RNA (mRNA) in non-overlapping sequences of three nucleotide bases at a time, known as "codons". With 4 available bases and a codeword length of 3, the total set of codewords contains $4^3 = 64$ distinct triplets. It is observed experimentally that different codons give the same amino acids making the code degenerate or non-injective!

Therefore DNA is not a "code" but a "degenerate code"!

Furthermore, because the DNA sequence lacks explicit boundary delimiters between words, decoding depends strictly on a fixed reading frame anchored by an initiation and stop signal. Inserting or deleting a single base shifts the frame and changes every downstream codeword (a frameshift mutation). As DNA transcription doesn't possess any error correcting code (see page 7152) - like we would have expect if an all knowing deity had created life (...) - this can lead sometimes to serious malformations.

14.5.4.4.1 Prefix codes

Definition (#14.5.622): A "prefix code" is a type of code system (typically a variable-length code) distinguished by its possession of the "prefix property", which requires that there is no whole code word in the system that is a prefix (initial segment) of any other code word in the system.

Prefix codes are also known as "prefix-free codes", "prefix condition codes" and "instantaneous codes". Although Huffman coding is just one of many algorithms for deriving prefix codes, prefix codes are also widely referred to as "Huffman codes" (see further below), even when the code was not produced by a Huffman algorithm.

Using prefix codes, a message can be transmitted as a sequence of concatenated code words, without any out-of-band markers or (alternatively) special markers between words to frame the words in the message. The recipient can decode the message unambiguously, by repeatedly finding and removing sequences that form valid code words. This is not generally possible with codes that lack the prefix property

Remarks

R1. The variable-length Huffman codes, country calling codes, the country and publisher parts of ISBNs, the Secondary Synchronization Codes used in the UMTS W-CDMA 3G Wireless Standard, and the instruction sets (machine language) of most computer microarchitectures are prefix codes.

R2. Prefix codes are not error-correcting codes. In practice, a message might first be compressed with a prefix code, and then encoded again with channel coding (including error correction) before transmission.

The Morse codes encodes the letter `E`, the most frequent, with a '.' And the letter `Y`, more rare, by '-.–': this is an example of a variable length code, which makes it possible to represent the most frequent letters or words by shorter words.

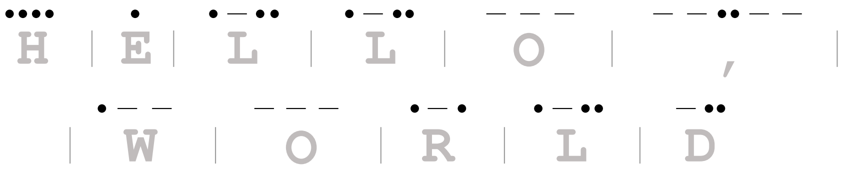


Figure 14.354: 'Hello World' in Morse code

An important property is the uniqueness of the decoding (injective application), a problem that does not arise for codes of constant length. It can be solved, but too costly, when a special symbol separates two successive words of the code (the "blank" in the case of the Morse code). We can therefore not use such a non-constant length code if no word is the prefix of another code word. And as the reader has probably understand it now, a code with this property is named a "prefix code".

✍**Example:**

Suppose we decide on a variable-size code convention, which matches, among other things, the following values:

Character	**Code**
0	11
2	11010
12	00
127	0111100
255	0100

Let us suppose that we have to decode the sequence:

$$1101000111100$$

Several interpretations (factorization) are then possible:

$$1101000111100 = 11\;0100\;0111100 = 0\;255\;127$$

or:

$$1101000111100 = 11010\;00\;11\;11\;00 = 2\;12\;0\;0\;12$$

And now we are very embarrassed! With several equivalent possibilities between which one can not decide, one is incapable of retranscribing the initial code.

The problem that has arisen here is that some codes are the beginning of other codes. Here, "11" is the code of the number 0, but it is also the beginning of "11010", code of the number 2. Hence the ambiguity!

We then better understand the purpose and the name of "prefix codes". Thus, in order for there to be no ambiguity at the time of the decoding, we must absolutely have a prefix code if the code is not of constant length.

✍**Example:**

The set:

$$X = \{ab, ababa, baa\}$$

is a code. Here, knowledge of the beginning of a possible code $ababa$ does not yet make it possible to know whether the decomposition begins by $ab \cdot ab \cdot ab \cdot a$ or by $ababa \cdot ba$. It is only after reading the next letter (not indicated in this example) that we know if the decomposition starts with ab (if the letter is b) or by $ababa$ (if the letter is a)

Definition (#14.5.623): A code is with "finite decryption delay" if there exists an integer d such that, whenever a message w begins with $p=x_1x_2\ldots x_{d+1}$ with $x_1,x_2,\ldots,x_{d+1}\in X$ then the complete factorization of w begins with x_d. It is therefore after a "delay" of d code words that we can affirm that the first word found is the right one (as in the example above that is a 8 finite decryption delay code).

✍**Examples:**

E1. The code:

$$X=\{a,ba,bb\} \tag{14.5.55}$$

has therefore a delay $d=0$.

E2. The code:

$$X=\{ab,ababa,baa\} \tag{14.5.56}$$

has therefore a delay $d=2$.

14.5.5 Linguistic algorithms

Let us put into practice what has already been seen so far in order to support us a little on "useful" concrete stuff!

14.5.5.1 Huffman algorithm

Let us first recall that in computing, we decide to encode an integer that has a value between 0 and 255 by a sequence of 8 binary digits (or "bits" in English, valued 0 or 1), also named "byte" (whose maximum value is equal to 2^8).

Even if there is a mathematical logic in the way of associating an 8-bit binary number to an integer between 0 and 255 (see section Numerical Methods page 6274) we can imagine any coding of the type:

Integer	Corresponding unique byte
...	...
138	10001010
139	10001011
...	...

in fact the correspondence can be any one, dictated by our imagination, as long as each integer between 0 and 255 is assigned to a fixed length binary code and only one. Once a correspondence is fixed, it is enough to take it as a convention.

The byte defined according to this convention is the basic unit of data storage. Any computer file is a sequence of bytes arranged in a defined order. The size of the file is simply the number of bytes that constitute it. The kilobyte (KB) corresponds to 1024 (not 1000) bytes, the megabyte (MB) to 1024×1024 bytes.

It should be noticed that this representation in base 2 is only a convention! Other conventions are possible, which would be equally appropriate if everyone agrees to use the same convention.

The problem we are asking ourselves is: would there be another way of coding the numbers, perhaps less logical but more judicious, in such a way that the size of the same file rewritten according to the new convention would be smaller?

The binary encoding convention is ultimately very democratic: whether you are a 0 or a 255, we allocate you 8 bits anyway to be able to code you. In other words, each possible input (a number between 0 and 255) is encoded on 8 bits. This is a fixed size encoding.

From the point of view of our problem (data compression), it would not matter if each of the possible values (0...255) were represented as frequently as the others. But in general this is not the case.

For example, see below the parsing of Wordpad.exe file (see your Accessories folder in your Microsoft Windows operating system[87]...). In this plot, on the abscissa as on the ordinate are the possible values of a given octet (hence 0 ... 255). On the diagonal, at the abscissa and corresponding ordinate, the size of the circle is proportional to the number of bytes having this value in the file:

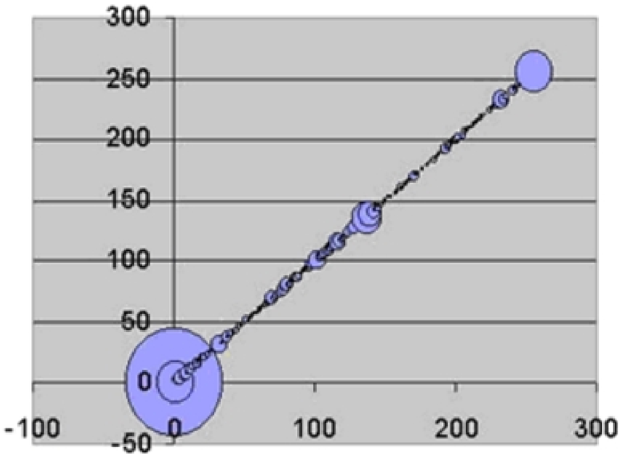


Figure 14.355: **Analysis of the byte distribution of a the Wordpad.exe file**

We see clearly that the values 0, 128 and 255 are much more frequent than the others!

As an indication, here are some values:

---

[87] This plot dates from they year 12001 (holocene calendar)... so its rendering quality is very bad. We will do it again when we will have the time

Value	**Count**	**Frequency**
0	71891	34.4%
2	1119	0.53%
128	1968	0.94%
130	79	0.038%
255	10422	4.99%

We will now decide of a variable-sized coding convention, which represents a value that is frequent by a small number of bits, and an uncommon value by a large number of bits.

We will now decide on a variable-sized coding convention, which represents a value that is frequent by a small number of bits, and an uncommon value by a large number of bits.

For example, 0 will now be represented by the sequence "11" (when before it was "00000000"), 128 by "1011010" (when before it was "10000000"), 255 by "0100" (when before it was "11111111"), etc.

Given that 0 represents almost one third of the file, we have gained a considerable place by coding it on two bits instead of eight! And same for the other frequent values...

Therefore "Huffman algorithm" is a recipe for generating a variable-length prefix code from the frequency table of a sequence of values. So, if you have followed the theory so far, it is a solution to our problem.

Suppose that our file is extremely simple, consisting of a single french word (*unconstitutionally*):

`anticonstitutionnellement`

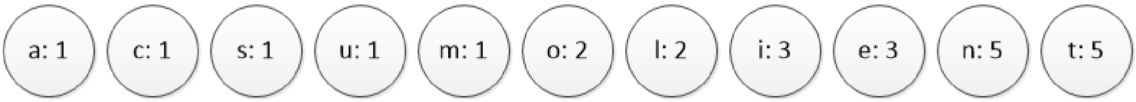
There are 25 characters in this file. Each character being encoded by an 8-bit octet (ASCII encoding), this means 25 bytes, or 200 bits! Let's see what we can do with that.

First let us render the table of frequencies:

Letter	**Count**
a	1
c	1
s	1
u	1
m	1
o	2
l	2
i	3
e	3
n	5
t	5

All other bytes (strings) have a null frequency: they are not represented in the file.

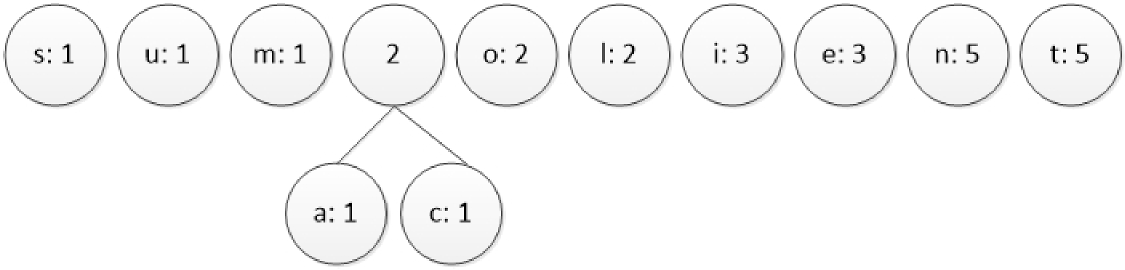
Now we create a "terminal node" for each entry of the array:


a: 1
c: 1
s: 1
u: 1
m: 1
o: 2
l: 2
i: 3
e: 3
n: 5
t: 5

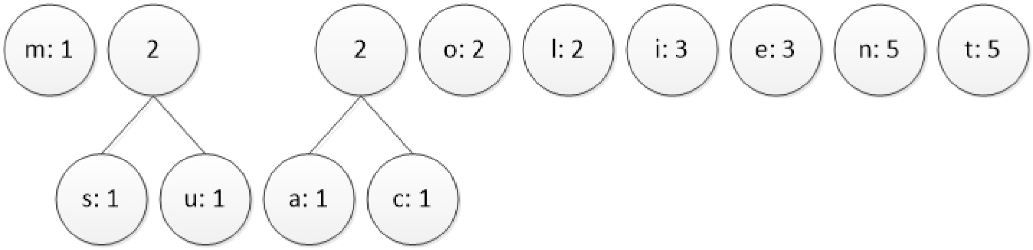
What makes for us now 11 trees containing only one knot each.

We now start an iteration: each time we delete the two trees on the left and replace them with a "sum tree". The new tree is inserted in the list in ascending order, and is repeated until there is only one tree left. Therefore we get:

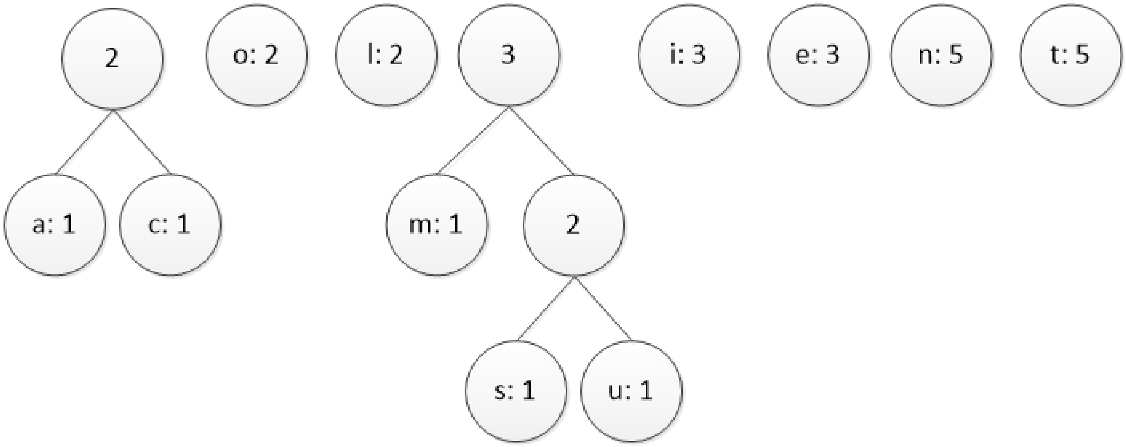
- First iteration:


s: 1
u: 1
m: 1
2
o: 2
l: 2
i: 3
e: 3
n: 5
t: 5
a: 1
c: 1

- Second iteration:

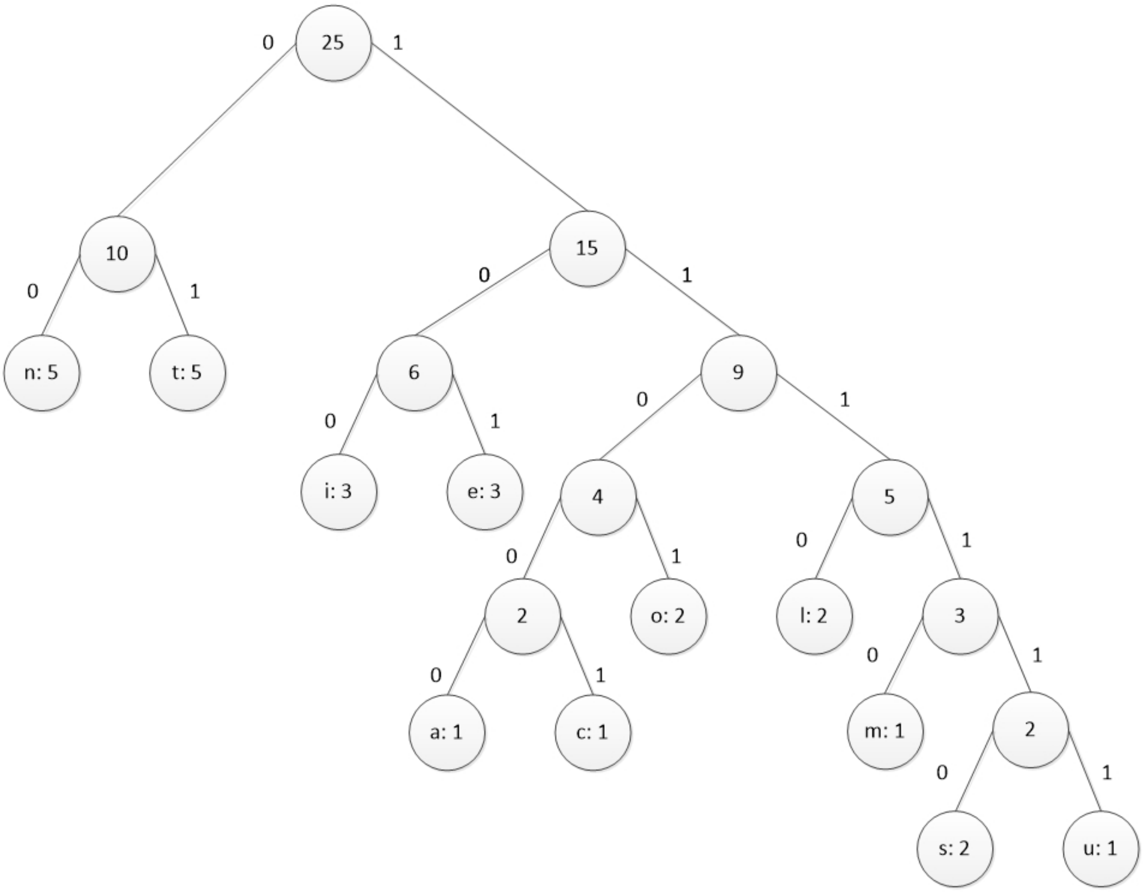

m: 1
2
2
o: 2
l: 2
i: 3
e: 3
n: 5
t: 5
s: 1
u: 1
a: 1
c: 1

- Third iteration:


2
o: 2
l: 2
3
i: 3
e: 3
n: 5
t: 5
a: 1
c: 1
m: 1
2
s: 1
u: 1

- ....

...and the final tree is:


25
0
1
10
15
0
1
n: 5
t: 5
0
1
6
9
0
1
i: 3
e: 3
0
1
4
5
0
1
2
o: 2
0
1
l: 2
3
0
1
a: 1
c: 1
0
1
m: 1
2
0
1
s: 2
u: 1

And that's it!

Now, the associated code to each letter is none other than the path to the corresponding terminal node from the root, noting 0 for each left branch and 1 for each right branch.

Finally:

Letter	Huffman binary code
`n`	00
`t`	01
`i`	100
`e`	101
`a`	11000
`c`	11001
`o`	1101
`l`	1110
`m`	11110
`s`	111110
`u`	111111

And here is now, transcribed with our new code, the starting word:

`110000001100110011101001111100110001111111011001101000010111101110101111101010001`

which makes 81 bits, instead of 200 at the beginning! This corresponds to a compression ratio of almost 60%.

The fact of having generated code using a binary tree ensures that no code can be the prefix of another one. You can verify that using the encoding table, there is no ambiguity possible to decode our compressed word! This is why the Huffman algorithm is in many compressed formats (for example the famous MP3 for audio encoding).

14.5.5.2 Sardinas and Patterson algorithm

When we have to deal with long codes, the difficulty is to check whether the code is really one... In order to do this, we can use the Sardinas and Patterson algorithm (the proof of this algorithm will be done during the next update of this section of the book).

In coding theory, the Sardinas–Patterson algorithm is a classical algorithm for determining in polynomial time whether a given variable-length code is uniquely decodable, named after August Albert Sardinas and George W. Patterson, who published it in 11953 (holocene calendar).

To do this check, we construct a graph $G(X)=(P,U)$, where P is the set of non-empty prefixes (according to the definition of "prefixes" given earlier above) of words of X, and U the set of pairs (u,v) such that one of the following possibilities is met:

P1. $uv\in X$ (by eliminating duplicate pairs if necessary)

P2. $v\notin X$ and it exist $x\in X$ such that $ux=v$

✍**Example:**

For $X=\{a,bb,abbba,babab\}$, the set P contains, in addition to X (which are prefixes of ε), the words $\{b,ab,abb,abbb,b,ba,bab,baba\}$ (respectively prefixes of $\{b,bba,ba,a,abab,bab,ab,b\}$).

First we see immediately that the set of $X=\{a,bb,abbba,babab\}$ is not a code because:

$$a\cdot bb\cdot babab\cdot abbba=abbba\cdot babab\cdot bb\cdot a \tag{14.5.57}$$

Now the pairs of U are for the first possibility P1:

$$(b,b),(abbb,a),(abb,ba),(ba,bab),(bab,ab),(baba,b) \tag{14.5.58}$$

and for the second possibility P2:

$$(a,abb),(ab,abbb),(b,ba),(bab,baba) \tag{14.5.59}$$

for which the x which is used to form the v is respectively bb, bb, a, a.

The Sardinas and Patterson graph will therefore be:

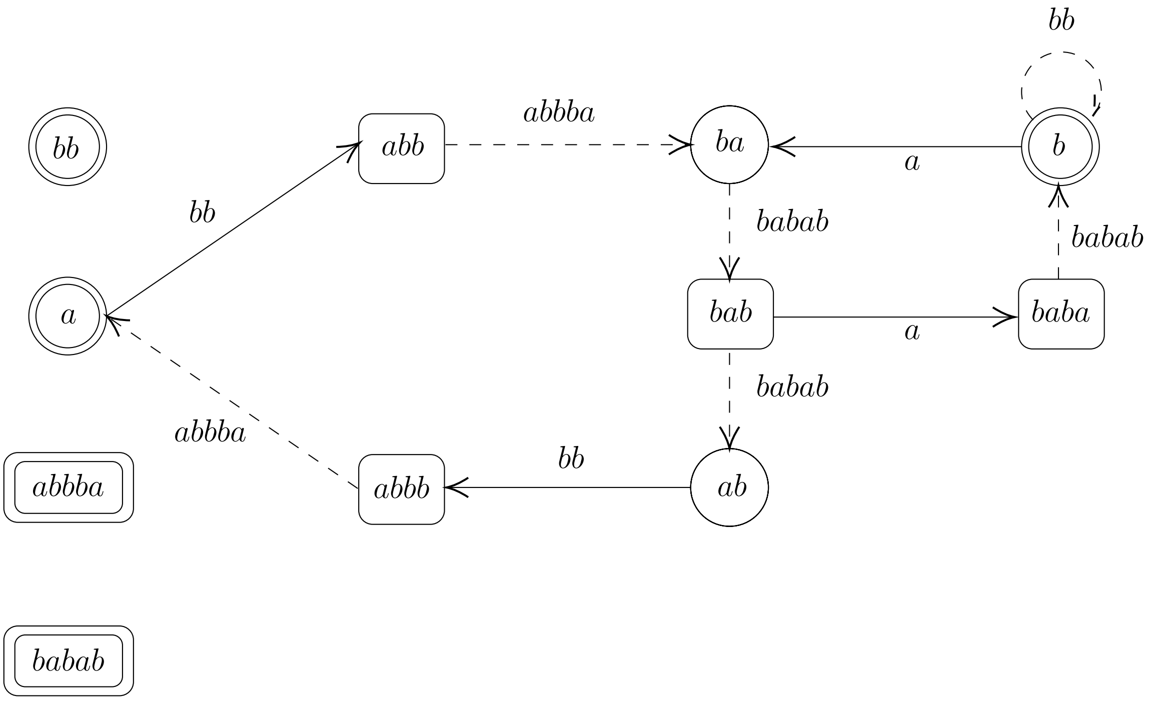


Figure 14.356: **Sardinas and Patterson example graph**

where the vertices corresponding to the words of X are doubly circled. The label of each arc is a word of the set X. The "crossed arcs" are traced in dotted lines and the "front arcs" in solid lines. If the arc (u,v) is crossed, then the label is uv, otherwise it is the word x such that $ux=v$.

In our example above, there is a path from a to a. By virtue of the Sardinas and Patterson theorem (which we will proved in the next update of this section), the set X is not a code (a single and unique path between any two vertices of X is enough).

The set X is a code if and only if there is no non trivial path in $G(x)$ of a vertex of X to a vertex of X (in other words, a set X is a code if and only if there is only the one and only trivial path leading from one vertex of X to another one - this vertex may be the same as in the previous example).

~60 %

★★☆☆☆

10 votes, 46.00%

14.6. Cryptography

CRYPTOGRAPHY is one of the disciplines of cryptology endeavouring to protect messages (ensuring confidentiality and / or authenticity) that two people wish to share through an insecure channel often thanks to secrets or keys.

The history of cryptography is already long and exciting (since it is a kind of "game"). We report its first use in Egypt 4,000 years ago. However, for centuries, the methods used were often remained very primitive. Moreover, its implementation was limited to the needs of the army and diplomacy. Thus, encryption methods and cryptanalysis (code breaking) experienced an important development during the Second World War and had a profound influence on the course of it.

At the end (especially!) of the 120th century (holocene calendar), with the proliferation of computers and electronic communications media, it became increasingly important to use secret codes for transmitting data between the military or private organizations. Thus, engineers have had to look at this same time solid numerical methods whose implementation and use was within reach of almost everyone (nation, enterprise and individual) while ensuring that external attacks required tools out of reach of an individual or group of individuals equipped with standard and high-performance IT tools (in computing power). Engineers and researchers then plunged into the mathematical tools to search for satisfying these specifications and the most known systems, mathematical theories which were adopted had over 200 years old (apart quantum cryptography).

The growth of cryptographic technology has raised a number of legal issues in the information age. Cryptography's potential for use as a tool for espionage and sedition has led many governments to classify it as a weapon and to limit or even prohibit its use and export. In some jurisdictions where the use of cryptography is legal, laws permit investigators to compel the disclosure of encryption keys for documents relevant to an investigation. Cryptography also plays a major role in digital rights management and piracy of digital media.

Steganography techniques (art of concealing a message in another one or in an image) however must be preserved because nothing tells us that computing power will still be available in times of war. It should be noted also that the steganography deployed wealth of imagination. Note for example: the permutations of letters, special and subtle formatting of characters, use of synonyms, hidden messages in text or comma behind a stamp, inside shots chess (hence the fact that these games were banned by the USA for some years after the attack on Pearl Harbor), in pictures / drawings, musical scores, etc. All of these techniques make that during the Second World War, the office of censorship in the United States occupied 10,000 full-time employees that analysed the mail of citizens, classified ads, radio text, etc.

Remarks

R1. To address the foundations of the theory of cryptography, we advise the reader to have read at first and at least in diagonal the section on Number Theory, on Set Theory, on Numerical Methods (especially the subsection on computational complexity), on Numerical Systems , on Statistical Mechanics (where information theory can be found) and for the part about quantum cryptography: the section of on Quantum Computing.

R2. We must remain aware that cryptography is more an engineer science than physicist science (except with quantum cryptography) and we must then not be so surprised to see some algorithms like fallen just from nowhere and adopted by industry because they just work almost well... Furthermore, it is also certain that only a few years after writing this text it will already be considered as obsolete (that is the art of engineering ... planned obsolescence).

14.6.1 Cryptographic systems

Definitions (#14.6.624):

A "cryptographic system" is composed of:

D1. A finite set P named "space of clear texts".

D2. A finite set C named "space of encrypted texts".

D3. A finite set K named the "space of keys".

For each key k, we seek an encryption function e_k:

$$e_k : P \mapsto C \tag{14.6.1}$$

and a deciphering (decryption) function d_k:

$$d_k : C \mapsto P \tag{14.6.2}$$

such as (see section Set Theory page 182):

$$\boxed{e_k \circ d_k = \mathrm{Id}_P} \tag{14.6.3}$$

In other words, these two functions must be injective!

To achieve this, two types of cryptographic techniques are mainly distinguished, encompassing almost all known modern encryption methods of the 120th century (for mathematical details see below):

1. The first concern cryptosystems with "symmetrical secret key".

Remark

Public keys often refer to the DES protocol (see below) for: Data Encryption System.

2. The second concerning encryption systems with "asymmetric public key".

Remark

This type of key often refers for example to the RSA protocol, the names of those to whom we awarded the Development: Rivest, Shamir and Adleman. They are widely used thanks to their rapid time encryption and decryption as well as their high entropy (see definition below).

By nature, these two types of keys are very different. Let us try to understand the reasons:

A symmetric encryption means a system where the key used in the encryption operation is that used in the deciphering operation. In this case, during a secure exchange (assumed to be as), both sides of the correspondence must share a same secret: the used key or "session key".

An asymmetric encryption designates an encryption system in which the key used for encryption (private key of the sender) differs from that used for decryption (recipient's private key). The only exchange that exists between members of the group is the public key, which allows each member to adjust its encryption based on the private key of the other members (among the many who have asymmetric systems been proposed, one of the most widespread in the early 121st century is the RSA).

The symmetric key ciphers are traditionally classified into two groups: "stream ciphers" and "block ciphers".

Definitions (#14.6.625):

D1. A "block cipher" is an encryption algorithm that encrypts a fixed size of n-bits of data - known as a block - at one time. The usual sizes of each block are 64 bits, 128 bits, and 256 bits. So for example, a 64-bit block cipher will take in 64 bits of plaintext and encrypt it into 64 bits of ciphertext. In cases where bits of plaintext is shorter than the block size, padding schemes are called into play. Majority of the symmetric ciphers used today are actually block ciphers. DES, Triple DES, AES, IDEA, and Blowfish are some of the commonly used encryption algorithms that fall under this group.

D2. A "stream cipher" is an encryption algorithm that encrypts 1 bit or byte of plaintext at a time. It uses an infinite stream of pseudorandom bits as the key. For a stream cipher implementation to remain secure, its pseudorandom generator should be unpredictable and the key should never be reused. Stream ciphers are designed to approximate an idealized cipher, known as the One-Time Pad. RC4, which stands for Rivest Cipher 4, is the most widely used of all stream ciphers, particularly in software. The cypher Engima machine of the second World Was is also a famous application of stream cipher.

✍**Example:**

Without going in the mechanical and electrical description of Enignma, the reader has just to know that the second version of the Enigma cypher machine had first a box 5 rotors with 26 start positions! The user had to choose 3 of theses 5 rotors and put them in a given position (position order has an importance!) in the machine.

So we have:

$$A_5^3=\frac{5!}{2!}=5\cdot 4\cdot 3=60 \tag{14.6.4}$$

combinations to put 3 rotors in a given order choosing among 5 rotors (5 rotors for the first position, multiplied the 4 remaining rotors for the second position and so on...).

After the Enigma user had to choose among on of the 26 position of each rotor. Then the number of starting positions possibilities is equal to:

$$26\cdot 26\cdot 26=17,576 \tag{14.6.5}$$

Finally the business and military version of the Enigma machine had something extra: plugboard.

This plug-board has 10 wires that connect two letter together among 26 letters (there $2\cdot 20$ letters combine together). The combination is therefore:

$$\frac{26!}{6!10!2^{10}}=150,738,274,937,250 \tag{14.6.6}$$

Indeed, we have $26!$ combinations of letters, but as there are 10 cables for therefore 20 letters we don't care about the combinations of 6 of them. Hence the division by $6!$. We divide by $10!$ as for all the remaining combination of letters we can use only $10!$ of them. Since we don't care about the direction of the cables (from A to B or B to C) and that we have 10 cables, we must divide 10 times by 2 and this is equivalent as dividing by 2^{10}.

So finally the total is:

$$60\cdot 17,576\cdot 150,738,274,937,250=158,962,555,217,826,360,000 \tag{14.6.7}$$

This is the total number of ways you can set the Enigma machine...

Remarks

R1. In 12001 (holocene calendar), Microsoft Internet Explorer (Microsoft's web browser in this time) worked with a 1024-bit asynchronous system certified by a synchronous system and Adobe Acrobat (PDF) in 12004 (holocene calendar) with an AES (Advanced Encryption System) of 128 bits for the low protection as well in the years 12010-12015 (holocene calendar) the iPhone 4S and 5.

R2. Microsoft Windows Enterprise and its E.F.S. system (Encrypting File System) uses a symmetric key (to encrypt the file) named "File Encryption Key" and asymmetric cryptography to encrypt the symmetric key in the file header as shown below (the key being updated regularly via Windows Update root certificates):

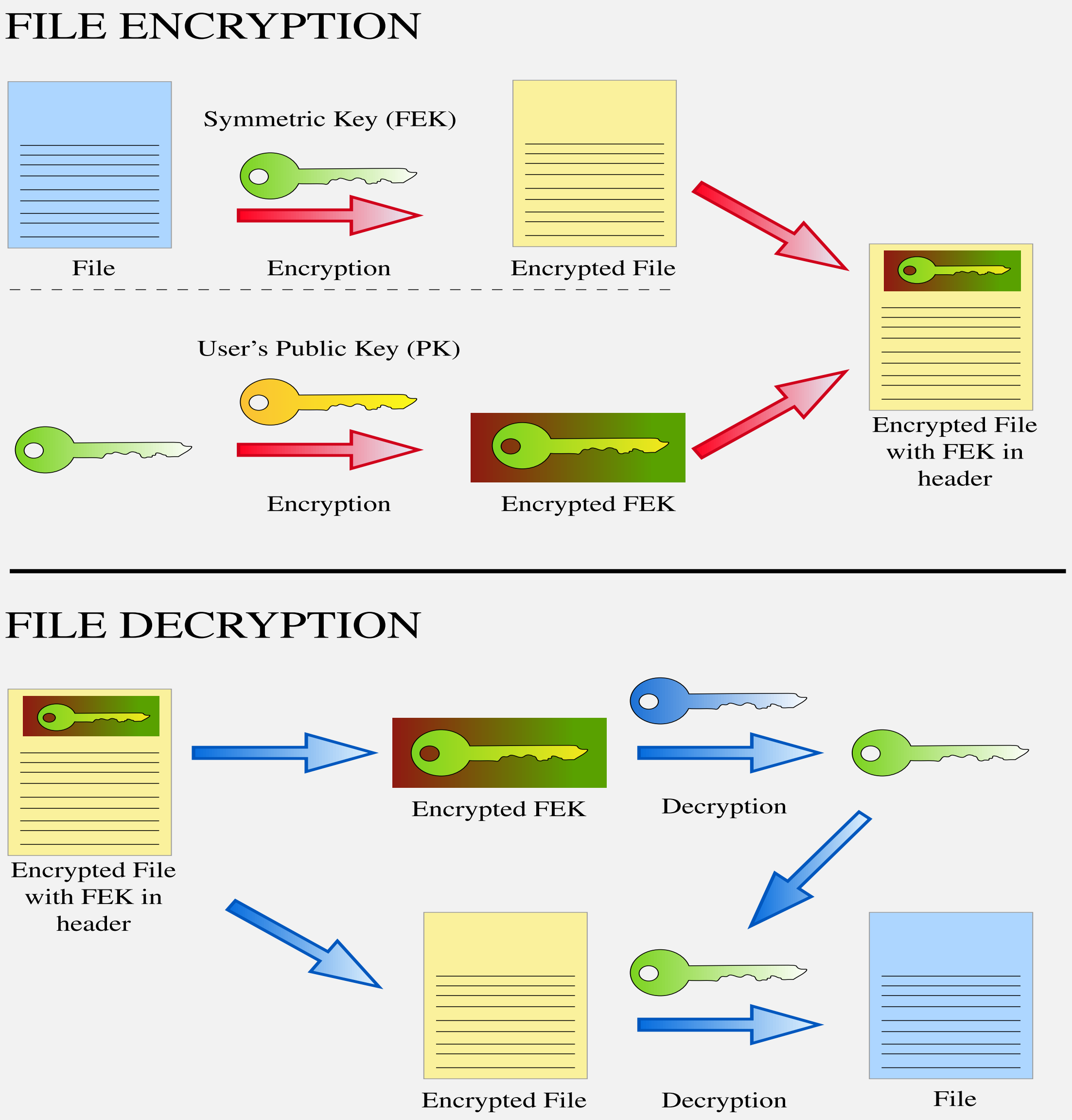

FILE ENCRYPTION
Symmetric Key (FEK)
File
Encryption
Encrypted File
User's Public Key (PK)
Encryption
Encrypted FEK
Encrypted File with FEK in header
FILE DECRYPTION
Encrypted File with FEK in header
Encrypted FEK
Decryption
Encrypted File
Decryption
File

Figure 14.357: **Principle of F.E.K. in Microsoft Windows XP (source: Wikipedia)**

However, the cryptography keys for EFS are in practice protected by the user account password, and are therefore susceptible to most password attacks. In other words, encryption of files is only as strong as the password to unlock the decryption key.

These methods are still decipherable, provided that the interceptor has enough time and paper/money (excepted at this date for quantum encryption).

Here is a small summary table of broken keys and their respective size for both conventional systems (the years are given based on the holocene calendar):

Secret key (symmetric system) exhaustive search		**RSA public key (asymmetric system)** exhaustive search	
Number of bits	Year	Number of bits	Year
40	Broken in 11995	256	Broken in 11985
56	Broken in 11998	512	Broken in 11999
64	Brokable	1024	Broken in 12010
128	Brokable in ∼12100	2048	Brokable in ∼2100
256	?	4096	?

Table 14.57: **Key systems and recent broke**

14.6.1.1 Kerckhoffs' principle

The primary function of cryptography is therefore to ensure the confidentiality of information exchange. Two parts of a confidential exchange will first agree on a secret convention to write their messages, and if they have carefully chosen, no one else should be able to enter their exchange.

If the secrecy of such agreements is possible from a few isolated individuals for a limited period, it is inconceivable at large scale and for a fairly long period. This is what Auguste Kerckhoffs understood when establishing the basic principles of practical cryptography which requires a fundamental principle encryption system "that does not require secrecy, and which can conveniently fall into the hands the enemy".

The six postulates of Kerchoffs are:

1. The system must be practically, if not mathematically, indecipherable;
2. It should not require secrecy, and it should not be a problem if it falls into enemy hands;
3. It must be possible to communicate and remember the key without using written notes, and correspondents must be able to change or modify it at will;
4. It must be applicable to telegraph communications;
5. It must be portable, and should not require several persons to handle or operate;
6. Lastly, given the circumstances in which it is to be used, the system must be easy to use and should not be stressful to use or require its users to know and comply with a long list of rules.

The second postulate, known today as the "Kerckhoffs principle" states that the security of an encryption system is not based on the secrecy of the procedure, but only on one parameter used when its implemented: the key. This key is the only secret of the Exchange Agreement.

This principle, however, was reformulated by Claude Shannon: "the enemy knows the system". This formulation is known as the "Shannon's maxim". This is the principle usually adopted by cryptologists, as opposed to the security through obscurity.

14.6.2 Traps

Sometimes there are what we name "trap doors" in public and secret keys. This is because when generating the key, which has to be done randomly within certain predefined theoretical constraints, the random generator may have an issue (the issue is sometimes voluntary on the part of the supplier of the material for spy purpose...).

In the secret keys, the traps are located at the level of the "key's entropy" (see section Statistical Mechanics page 3658), directly linked to the entropy of the random generator. We can simplistically define the entropy of a key generator by the average optimal binary questions (that is to say giving rise to the type of answers Yes / No) that we need to ask someone knowing a key produced by this generator to determine it. More the entropy of a key generator is high, higher is the number of questions we need to determine the key. Conversely, the smaller is the entropy, the lower are the questions, so that the search of a key is facilitated.

The introduction of traps in the asymmetric key systems is much more difficult, since this type of key already has intrinsic mathematical structure: their construction is not due to chance but is the result of mathematical rules. Chance is here in the choice of large prime numbers used. The fact that asymmetric systems can be easily calculated, but they are difficult to reverse are sometimes named "trapdoor functions".

✍**Example:**

An example of a simple mathematical trapdoor is "6895601 is the product of two prime numbers. What are those numbers?" A typical solution would be to try dividing 6895601 by several prime numbers until finding the answer. However, if one is told that 1931 is one of the numbers, one can find the answer by entering "6895601 ÷ 1931" into any calculator. This example is not a sturdy trapdoor function – modern computers can guess all of the possible answers within a second – but this sample problem could be improved by using the product of two much larger primes.

Therefore if a random generator that generates prime numbers is biased (see section Statistics page 818), this bias will facilitate the research of a trapdoor.

Definition (#14.6.626): A "trapdoor function" is a function that is easy to compute in one direction, yet difficult to compute in the opposite direction (finding its inverse) without special information.

Remarks

R1. Functions related to the hardness of the discrete logarithm problem (either modulo a prime or in a group defined over an elliptic curve) are not known to be trapdoor functions, because there is no known "trapdoor" information about the group that enables the efficient computation of discrete logarithms.

R2. Trapdoor must not to be confused with a "backdoor" as this latter is a deliberate mechanism that is added to a cryptographic algorithm or operating system, for example, that permits one or more unauthorized parties to bypass or subvert the security of the system in some fashion.

14.6.3 Secret-key encryption system

Definition (#14.6.627): The "single-use encryption" is a secret key encryption algorithm proved unconditionally secure. Properly used (and that's an important point), it provides an unbreakable encryption in reasonable time.

The theoretical basis of this encryption system are:

Given a message M in binary form to be transmitted between people A (creator and originator of the message M) and B (reader and receiver). We generate a large amount of bits if possible "truly randomly" forming a secret key K of same size as the message to be transmitted (computer programs, deterministic by nature, can not generate truly random bits).

This key will be sent to B by a channel supposedly safe ... A given time after the transmission of this key A will encode his message into C by performing the operation:

$$C = K \star M \tag{14.6.8}$$

where $\star$ is an operator that must satisfy to a group law (see section Set Theory page 421) on a finite set (that contains a limited number of items or "letters").

The idea in computing science is to use the XOR law (exclusive OR) denoted $\oplus$ for what will follow (see section Logical Systems page 7112) as it is enough as a group law (remember that a group is the smallest structure having an opposite - and is associative and having neutral element - that gives therefore the possibility to reverse the encoding process). Therefore:

$$C = K \oplus M \tag{14.6.9}$$

Finally, the sender A transmits the encrypted version of his message C by a route not necessarily secure. B can read the original message M by using the inverse operator $\oplus^{-1}$ (the XOR operator is its own inverse as we have proved it thanks to its truth table in the section of Logical Systems!!!!). So receiver B will do the following:

$$M = C \oplus K \tag{14.6.10}$$

Provided that the K has been generated totally randomly and that each bit of it has been used only once to encrypt the message, an interceptor gets no information about the clear message M if he intercepts C. Indeed, in these conditions, we can not establish any correlation between M and C without the knowledge of K.

Even with future ultra-powerful quantum computers, the problem is insoluble, because nothing connects the information which is available and the problem to solve. Consequently, the "single-use encoding" is an encryption algorithm "unconditionally secure". The proof of its security does not rely on unproven mathematical conjectures and decryption attempts of an interceptor with infinite computing power are futile.

However, each stage of encryption is a source of possible errors. Indeed, the key K may have been poorly developed. The slightest statistical deviation of K compared to the "real" random provides information on the clear message M from its encrypted version. This is why the K bits are to be used only once if possible.

Indeed, suppose that same key is used to encrypt messages of French language M_1 and M_2 an attacker manages to intercept the two corresponding encrypted messages. From C_1 and C_2 the interceptor and can easily obtain information about K and this because of language peculiarities (same for English). Indeed, since:

$$C_1 = M_1 \oplus K \qquad C_2 = M_2 \oplus K \tag{14.6.11}$$

then the interceptor knows a simple result that involves M_1 and M_2 without the key K:

$$C_1 \oplus C_2 = M_1 \oplus M_2 \tag{14.6.12}$$

as:

$$K \oplus K = 0 \tag{14.6.13}$$

(if necessary make the truth table to be convinced of this relation). Now, if M_1 and M_2 are in the same language, we will know, usually due to language redundancies (e.g. the letter "e" often appears in French), found from $C_1 \oplus C_2$ each of the two original messages (the work is though laborious without statistical automated tools).

✍**Example:**

Imagine that we want to send a little message M binary coded by 1101 and we generated a random key K that gave 0101.

Then we have:

$$C = M \oplus K = 1101 \oplus 0101 = 1000 \tag{14.6.14}$$

and therefore:

$$M = C \oplus K = 1000 \oplus 0101 = 1101 \tag{14.6.15}$$

Obviously in this kind of small situations we can guess M without much difficulty just by having C if there such like here only a single encryption step. This is why there are encoding patterns as we shall see now.

The main problem with this technique is the creation of a key as random as possible. To overcome this, mathematicians do pass the key through a series of nested functions, the result after many iterations, becomes "pseudo-random".

Building a pseudo-random iteration is one thing, building a pseudo-random bijection is yet another!!! Indeed, we need to decrypt the message later, which is why we absolutely need a bijective system (which has everything arrival element - encrypted message - matches a single starting element - decrypted message - and vice versa).

14.6.3.1 Feistel Schemes

Even if encryption algorithms, in this late 120th century (holocene calendar) and early 121st century (holocene calendar), are considered sufficient with a key having a finite number of bits, the goal remains the development that from a message M and a random sequence of digits, or at least that looks like, to build a key K to send an encrypted message C that can be decipher easily only by people knowing the key. Specifically, this target application is to construct or identify a function which, firstly, do correspond to each digit of M a digit C that seems to look random (but whose value depends in reality on the deterministic key) and, secondly, authorizing the reverse path (inverse function by the property of bijection), that is to say that from a digit C, we can uniquely trace back to the corresponding digit of M. We therefore would like to find a pseudo-random bijection function.

In the years 11950s (holocene calendar), the mathematician Horst Feistel has shown that a pseudo-random function transformed itself, by a relatively simple method, in bijection function. Today, the "Feistel cypher" is most commonly used in the secret key encryption systems and is also the basis of the DES (Data Encryption System). How does it work?

Here is the principle:

The initial message to be encrypted has a size of $2n$ bits. The split the original message M into two blocks (thus the Feistel Schemes belongs ton the family of "block cipher"), G and D, of equal length (G includes the first n bits and D the following) and we build the transformation φ that associates to G and D the numbers T and S such as:

$$S=D \qquad \text{and} \qquad T=G\oplus f_1(D) \tag{14.6.16}$$

where for reminder the $\oplus$ still represents the bit by bit XOR operation and where f_1 is any function, non-necessarily bijective, from n bits to n bit using the secret key K.

The transformation $\varphi(G,D)=(S,T)$ is indeed bijective, as we can go back in a unambiguous way starting form S and T to G and D by the operations:

$$D=S \qquad \text{and} \qquad G=D\oplus f_1(S) \tag{14.6.17}$$

Obviously we must not stop here, since the right side of the message, D, has not been encrypted, it is simply passed to the left. However, as φ is bijective, we can repeat the process. A Feistel scheme where we apply n times the function φ is named a "n-step pattern".

✍**Example:**

We will encrypt by the a two-step Feistel cypher a message consisting of 4 bits (thus $2^4=16$ possibilities of messages), what is equivalent to building a bijection from 4 bits to $4$bits from two functions f_1,f_2 of two bits to two bits. The functions f_1,f_2 have in input both: the message to encrypt and the secret key. We will assume that for some input key, these functions are:

Input	f_1	**Output**
00	$\rightarrow$	01
01	$\rightarrow$	11
10	$\rightarrow$	10
11	$\rightarrow$	01

Input	f_2	**Output**
00	$\rightarrow$	11
01	$\rightarrow$	00
10	$\rightarrow$	00
11	$\rightarrow$	01

Table 14.58: **Input/Output key matches by functions**

Let us notice that neither f_1 nor f_2 are bijections ($f_1(00)=f_1(11)=01, f_2(01)=f_2(10)=00$). For example, encrypt the message 1101. G designates the left part of the message to be encrypted, D the right part:

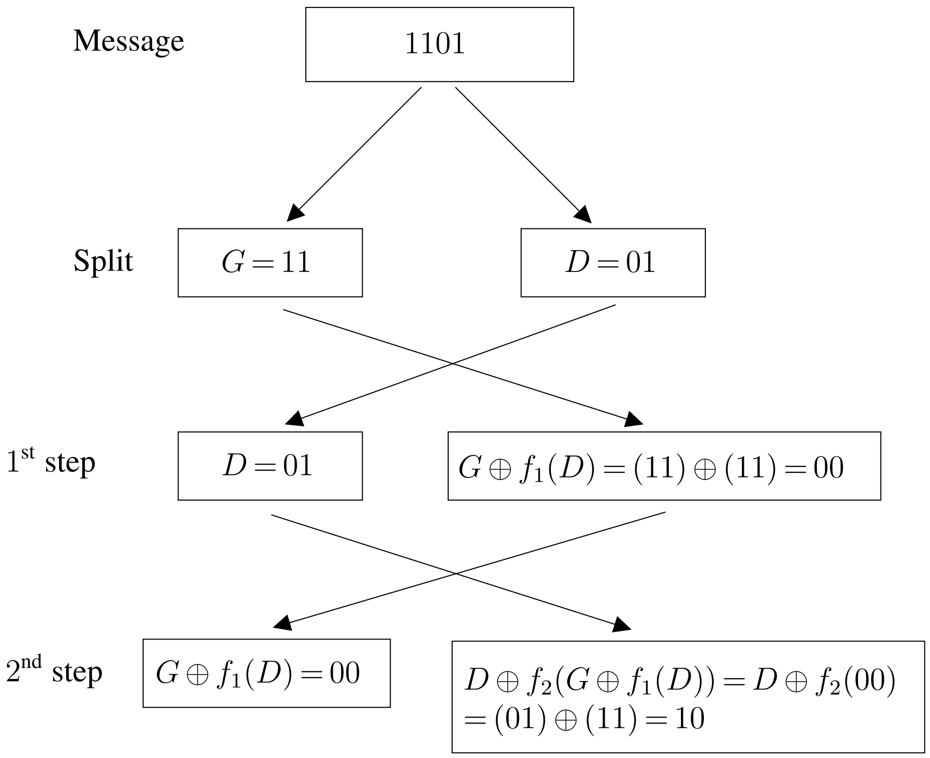


Figure 14.358: **Encryption of 1101 using the Feistel method**

The result is 0010. We will compute the image of the other 15 other possible messages and verify that there is an unambiguous correspondence between each message and its image by the Feistel scheme: we have constructed a bijection from two functions that are not bijective.

Quite complex theoretical results guarantee the cryptographic security of Feistel schemes starting from 4 steps when n is large enough and when the functions f_i are indistinguishable from truly random functions. In practice, rather than using 4 steps and functions f_i that look like random, it is generally preferred to use more steps and more simple functions f_i. After a few steps, the obtained bijection often becomes very difficult to distinguish from random bijections. And for parameters well chosen, we no longer know at all how to distinguish them from truly random bijections!!!

Most of the secret key encryption algorithms currently in this end of 120th century (holocene calendar) used in the civilian world are Feistel schemas. In particular, the DES (Data Encryption

System) algorithm which is a 16-step Feistel scheme as shown in the figure below and the Triple DES (TDES) algorithm which is a 48-step Feistal scheme and the Blowfish algorithm that will not be discussed here).

Remark

For example, there are, in some bank cards (at least in the beginning of this 121st century... according to holocene calendar), a DES key (or TDES since October 12001 according to holocene calendar) which provides proof of the legitimacy of the card between the bank's control center and the merchant's terminal in addition to Public part of a RSA key to make sure the user code is entered (control done by an internal chip on the card, which must then be manufactured in very secure premises).

Rigorously the Feistel scheme is a bit different because it involves keys, which we did not use in the example presented before. Here is a more detailed figure in what this Feistel scheme consists of (see figures below).

Principle of the diagram: A message to be encrypted is divided into blocks of 64 bits, each of which is divided into two 32-bit sub-blocks, the left block (G) and the right-hand block (D). At each iteration, the old right block becomes the new left block and the new right block results from the XOR operation of the old right block, whose bits are mixed by a confusion function, and of the previous left block. The iteration is repeated 16 times.

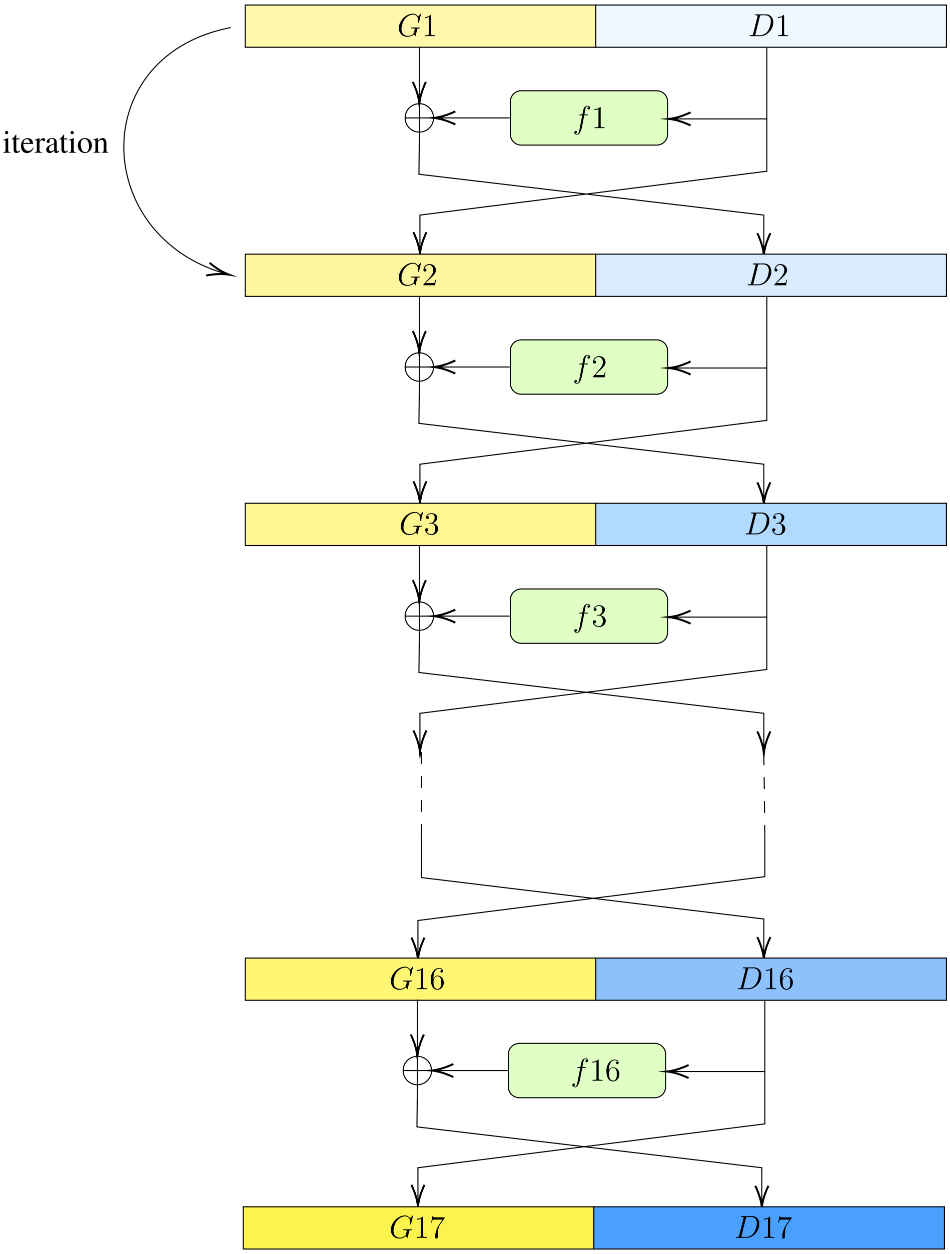


Figure 14.359: Feistel's scheme a little more realistic (source: ?)

The confusion function (**f**), which acts on the 32-bit blocks, mixes the bits according to the following processes (see the figure below):

- First, it transforms the 32-bit block into a 48-bit block by duplicating certain bits ("expansion").
- Then, it adds to this block a 48-bit ("token key") subkey extracted from the 56-bit secret key
- And then transforms each 6-bit set into 4 bits by local transformations (**S** transform)

The result is a 32-bit block which is finally mixed according to a fixed permutation.

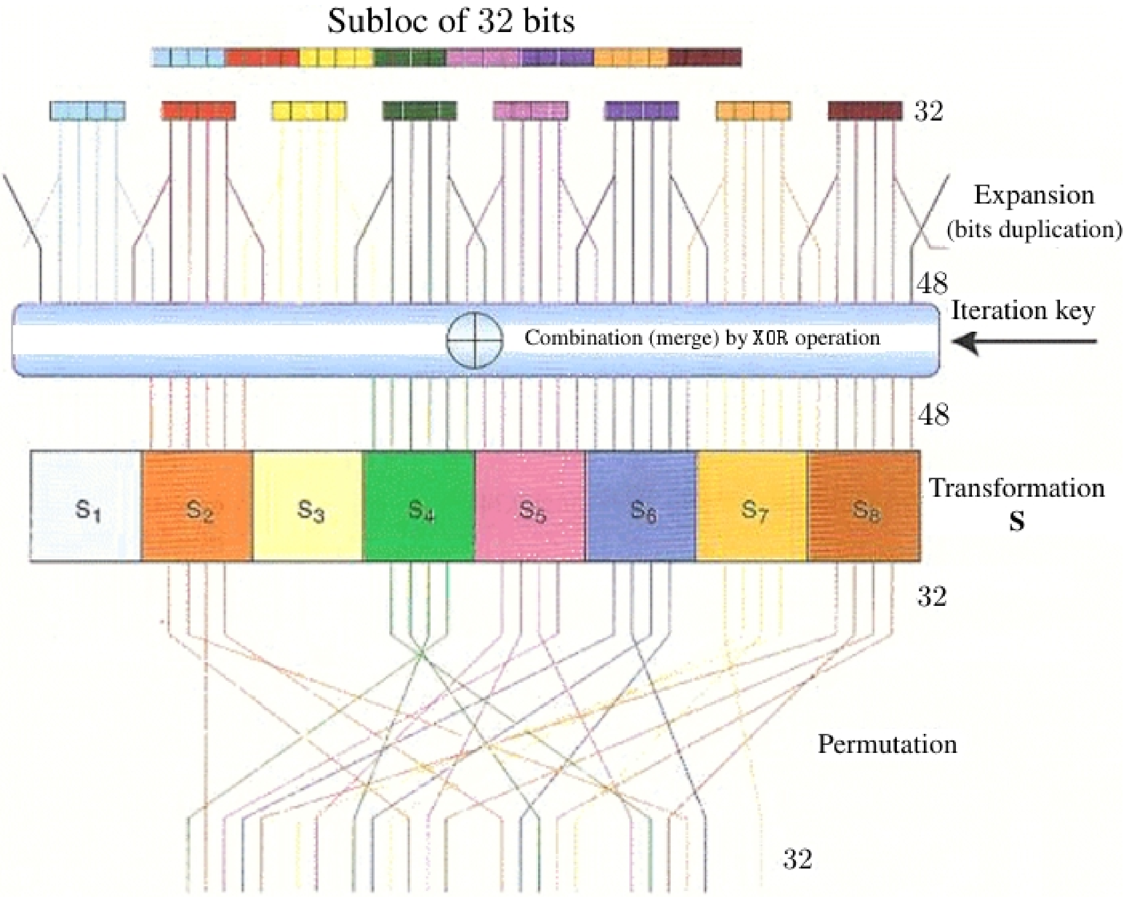

Subloc of 32 bits
32
Expansion
(bits duplication)
48
Iteration key
Combination (merge) by XOR operation
48
S_1
S_2
S_3
S_4
S_5
S_6
S_7
S_8
Transformation
S
32
Permutation
32

Figure 14.360: Feistel's scheme confusion function

14.6.4 Public key encryption

In 11975 (holocene calendar), W. Diffie and M. E. Hellman revolutionized the science of cryptography by proving the existence of a protocol that could not be deciphered by an interceptor unless the interceptor had large computer resources. The most fascinating in their method - the principle of which is still in use in this early 121st century (holocene calendar) - is that the code used does not require to hide the chosen method and can be applied repeatedly without any modification (Kerckhoffs principle). At their time, they simply created the concept of "public-key cryptography", or "asymmetric cryptography" (which we mentioned earlier in this section), an invention that sparked the emergence of a dynamic academic and industrial community.

Remark

Contrary to what one might think, public key cryptography has not relegated secret key cryptography to oblivion, on the contrary: these two types of cryptography are most often used in hybrid cryptosystems where the authentication of published keys is performed by a "certification authority".

Before describing the Diffie-Hellman protocol in detail, let us recall that the protocol of exchange of the "secret keys" was not reliable at that time (and is still not today) as it was transiting between the interlocutors, the element making it possible to encrypt and therefore decrypt the messages. In addition, even if only one key were to travel, anyone with sufficient computing power could break the code. Hence the need to change (misfortune more!) Periodically the keys (cryptoperiod). At least two solution are therefore available to us:

1. Do not exchange any key (it is possible but it is quite long as we will see in the figure below)

2. Exchange a secret key using a non-invertible mathematical function or at least very difficult to inverse (this is the Diffie-Hellman protocol that we will also see in a figure below).

Therefore Public key cryptography systems often rely on cryptographic algorithms based on mathematical problems that currently admit no efficient solution—particularly those inherent in certain integer factorization, discrete logarithm, and elliptic curve relationships. Public key algorithms, unlike symmetric key algorithms, do not require a secure channel for the initial exchange of one (or more) secret keys between the parties.

Because of the computational complexity of asymmetric encryption, it is usually used only for small blocks of data, typically the transfer of a symmetric encryption key (e.g. a session key). This symmetric key is then used to encrypt the rest of the potentially long message sequence. The symmetric encryption/decryption is based on simpler algorithms and is much faster.

Let us see what the first solution is and its blatant disadvantage:

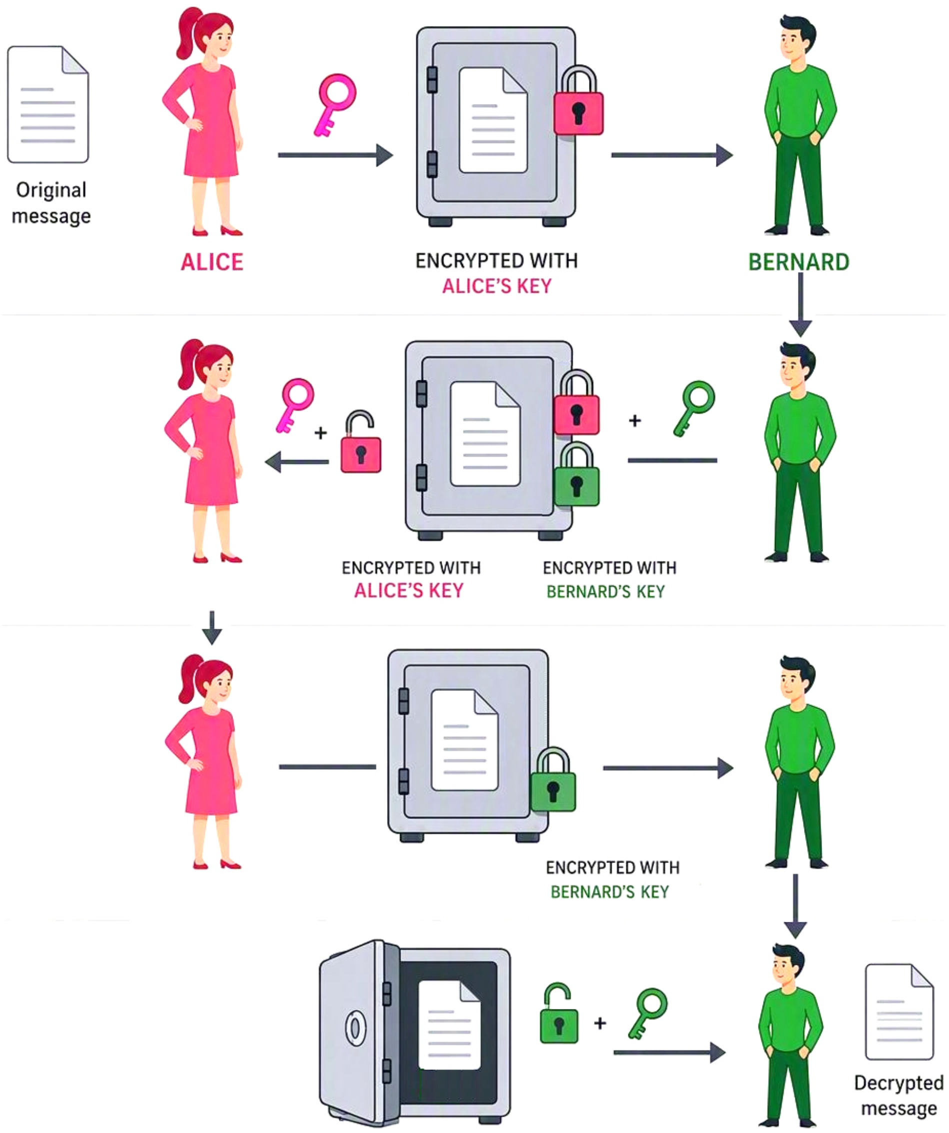

Original message
ALICE
ENCRYPTED WITH
ALICE'S KEY
BERNARD
+
+
ENCRYPTED WITH
ALICE'S KEY
ENCRYPTED WITH
BERNARD'S KEY
ENCRYPTED WITH
BERNARD'S KEY
+
Decrypted message

Figure 14.361: **Principle of public key encryption**

Explanation: Alice and Bernard want to transmit a message on an unsecured line and without exchanging keys. To do this, Alice puts her letter in a chest that she closes with her key and sends it to Bernard. The latter returns the chest to Alice where he added his own padlock which he closed with his own key. When Alice receives the chest, she takes off her padlock and sends Bernard a chest that no longer includes her padlock but. The latter then only has to open the chest with his key to read the letter. This operation is safe and does not require exchange of keys. On the other hand, it requires several paths.

The principle of the public key must allow secure exchanges, without a secret key, in a single path. Bernard distributes widely copies of his public padlock. Alice gets one, but anyone could do the same. Alice places the message in the trunk and closes it with the Bernard code lock, then sends it the trunk (represented by the 5th transaction in the figure above). On receiving the trunk, Bernard can open the trunk, since he alone holds the key that opens this lock. The transfer is safe in one trip. In cryptography, the public key is equivalent to the code lock, which is available for example in directories, while the key that opens this lock is the private key, owned solely by their owner and never disclosed. The private and public keys (the so-named "key trousseau") are constructed from a supposed "one-way" mathematical function.

Let's now see the second solution making use of public key according to the Diffie-Hellman protocol:

14.6.4.1 Diffie-Hellman protocol

As the name implies, a one-way function gives easily a result, but the reverse operation is very difficult. Finding such functions in the mathematical world seemed very arduous to mathematicians. How to imagine a function that is one-way for everyone, except for its creator who can reverse it through the knowledge of a particular information. Thus, W. Diffie and Hellman were the first to publicly propose a one-way function to solve the problem of agreeing on a common secret. The basic idea is to calculate values of the type:

$$\alpha^a \mod p \tag{14.6.18}$$

where α and a are imposed as being integers and p is a prime number.

Mathematicians name this kind of operation a "modular exponentiation" or "discrete exponential" and it is customary to denote the finite field of integers modulo p (where p is a prime number) by GF(p) in honour of Évariste Galois.

To explicate such a calculation (as a reminder of what was seen in the section of Number Theory ...), we raise a number α to the power of a, and then divide the result by a large prime number p and we keep finally the remainder of this division (operation modulo p). If this remainder is denoted r then we write this:

$$\boxed{\alpha^a \equiv r \mod p} \tag{14.6.19}$$

Modular exponentiation similar to the one described above are considered easy to compute, even when the numbers involved are enormous. On the other hand, computing the "discrete logarithm" – that is, the task of finding the exponent a when given α, p, and r (i.e. $\alpha^a \mod p$) – is believed to be difficult. This one-way function behaviour makes modular exponentiation a candidate for use in cryptographic algorithms! In addition, one-way functions such as the one above from the modular arithmetic behave very irregularly as is shown in the table with the particular example below:

a	$\alpha^a = 3^a$	$\alpha^3 \mod p = 3^a \mod 7$
0	1	1
1	3	3
2	9	2
3	27	6
4	81	4
5	243	5
6	729	1
7	217	3
8	6561	2

Table 14.59: **Examples of modular exponentiation applications**

So even if it is easy to compute a discrete exponential, it is almost impossible to find the starting number a from the result, especially when this modular function is applied to very large primes p.

The reader can check this by playing with Maple™ 4.00b that can calculate the discrete logarithm as following:

```
>with(numtheory):
>mlog(r,alpha,p);
```

Ok we know how to crypt message... But now how can we communicate messages to someone that should be able to uncrypt them? This is named the "Diffie–Hellman key exchange" that is a specific method of securely exchanging cryptographic keys over a public channel and was one of the first public-key protocols as originally conceptualized by Ralph Merkle and named after Whitfield Diffie and Martin Hellman. D–H is one of the earliest practical examples of public key exchange implemented within the field of cryptography. Here is the idea of the protocol:

ALICE	**Public (Internet)**	**BERNARD**
We choose an arbitrary common prime number $p=419$ and a common random number smaller than p: $\alpha=7$. These two values are assumed to be secret.		
Alice chooses a secret random number: $a=178$		Bernard chooses a secret random number: $b=344$
With the number a Alice generates the public element: $k_a=\alpha^a \mod (p)=181$		With the number b Bob generates the public element: $k_b=\alpha^b \mod (p)=351$
The result is sent to Bernard: $k_a=181$	$\longrightarrow$	$k_a=181$
$k_b=351$	$\longleftarrow$	The result is sent to Alice: $k_b=351$
The shared secret is then: $K=(k_b)^a \mod (p)=493$		The shared secret is then: $K=(k_a)^b \mod (p)=493$
$\leftarrow$ The exchanges are then encrypted with the secret key K $\rightarrow$		

Table 14.60: **Example of key exchange following Diffie-Hellmann protocol**

The security of this protocol is computational. It is based on the assumption that with limited computing power and time, an opponent (spy) can not reverse the modular exponential function (by making use of the properties of the logarithms with the exponential functions as we saw in the section of Functional Analysis) and therefore can not find the secret a from the exchanged elements. This computational difficulty is due to the fact that the computation time necessary for the inversion of a one-way function does not have an algorithmic complexity (see section of Numerical Methods page 6284) polynomial but exponential with p.

Alice and Bernard have calculated the same common secret: 493. Then 493 is used to encrypt the exchanged data (in practice, much larger numbers are used). The spy is supposed to be able to intervene only after the exchange of the common choice of p and α (no man-in-the-middle attack!).

Remark

This protocol is vulnerable to a "man-in-the-middle attack" that is an attack where the attacker secretly relays and possibly alters the communication between two parties who believe they are directly communicating with each other. One example of man-in-the-middle attacks is active eavesdropping, in which the attacker makes independent connections with the victims and relays messages between them to make them believe they are talking directly to each other over a private connection, when in fact the entire conversation is controlled by the attacker. The attacker must be able to intercept all relevant messages passing between the two victims and inject new ones.

The key K is obtained by the fact that the power operation is compatible with the relation of equivalence modulo p (see section Number Theory page 356) such that:

$$x \mod (y) = z \Leftrightarrow (x \mod (y))^w = (x^w \mod (y)) = z \qquad (14.6.20)$$

✍**Example:**

We have:

$$5 \mod (3) = 2 \qquad (14.6.21)$$

when with $w = 2$ we have:

$$5^2 \mod (3^2) \neq 2^2 \qquad (14.6.22)$$

but:

$$5^2 \mod (3) = 2^2 \qquad (14.6.23)$$

If it is not clear let us write it differently:

$$5 \equiv 2 \mod (3) \Leftrightarrow 25 \equiv 4 \mod (3) \qquad (14.6.24)$$

Indeed, for the first one: $5 - 2 = 3$ can be divided by 3 and for the second one $25 - 4 = 21$ can be divided by 3.

Thus, since $x < p$, the second modulo below has no meaning, so we can write:

$$(k_b)^a \mod (p) = (\alpha^b \mod (p))^a \mod (p) = (\alpha^b \mod (p))^a = \alpha^{b+a} \mod (p) \qquad (14.6.25)$$

identically:

$$(k_a)^b \mod (p) = (\alpha^a \mod (p))^b \mod (p) = (\alpha^a \mod (p))^b = \alpha^{a+b} \mod (p) \qquad (14.6.26)$$

and therefore:

$$(k_b)^a \mod (p) = (k_a)^b \mod (p) = K \qquad (14.6.27)$$

Diffie-Hellman is a cornerstone of modern cryptography used for VPNs, HTTPS websites, email, and many other protocols. Bad implementation choices combined with advances in number theory mean real-world users of Diffie-Hellman are likely vulnerable to state-level attackers.

Despite these precautions, experts established a record at the beginning of the 121st century (holocene calendar) using a new algorithm, they succeeded in reversing the modular exponential function for a p-number of 120 digits (about 400 bits), using a Computer with four 525 [MHz] processors. This record shows that the security of the protocol depends greatly on the constant progress made in the field of algorithmic complexity. Researchers estimate that breaking a single, common 1024-bit prime would allow NSA (USA National Security Agency) to passively decrypt connections to two-thirds of VPNs and a quarter of all SSH servers globally. Breaking a second 1024-bit prime would allow passive eavesdropping on connections to nearly 20% of the top million HTTPS websites. In other words, a one-time colossal investment in power-lifting computation would make it possible to eavesdrop on trillions of encrypted connections.

The clever schema of Diffie-Hellman remains a schema of principle. Its main disadvantage is that it does not make it possible to provide the traditional security services: authentication of the two interveners, control of the integrity of the key and anti-replay (verification that information already transmitted is not re-transmitted). It follows that an attacker can, for example, impersonate Alice by replacing the public element of Alice with her own public element. To overcome this disadvantage, secure versions of this generic protocol have been published, for example a protocol named "STS" (Station To Station), which uses, in particular, the electronic signature to ensure the authentication of the interveners (see below). This policy is the basis of the secured Internet connection (IPSec).

The Diffie-Hellman protocol paved the way for a whole series of algorithms, that of "public key encryption" being the first. The idea was to break the symmetry of encryption and decryption by using one-way functions.

14.6.4.2 R.S.A system

Curiously, the first "R.S.A. encryption system" is conceptually quite different from the Diffie-Hellman protocol: it does not use the discrete exponential, but the factorization of large numbers. This public key system was invented in 11977 (holocene calendar) by Ron Rivest, Adi Shamir, and Leonard Adleman (hence the abbreviation "R.S.A."). Having quickly become an international standard, the R.S.A. technique has been marketed by more than 400 companies and we estimate that more than 400 million softwares use it. It is implemented in web browsers, such as Netscape Navigator, Microsoft Internet Explorer, or some bank smart cards, such as VISA cards.

The R.S.A. system is based on the difficulty of factorizing large numbers and the one-way function used is a "power" function. The R.S.A. encryption protocol is divided into three phases:

1. Creation of keys (public and private)
2. Encryption using the recipient's public key
3. Decryption using the private key

Its concept is based on a famous theorem named "Euler's theorem" (nothing to do with the theorem of the same name seen in the section of Graph Theory or in the section of Geometric Shapes). Let's see what it is (be careful it is relatively long!).

14.6.4.2.1 Euler's theorem

Before we see what Euler's theorem consists of, we must define two elements that are included in it. Apart the concept of congruence which we have already studied in the section of Number Theory (see page 365), there remains a special function named the "Euler indicator" or also named "totient function" and defined in general by:

$$\boxed{\phi(m)=\sum_{k=1,(k,m)=1}^{m} 1} \tag{14.6.28}$$

In other words, the function ϕ of the integer m results in a number n strictly less than m, given by the number of elements between 1 and m whose greatest common divisor (see section Number Theory page 340) with m is 1. We have already given a practical example of the utility of this indicator function in the section of Number Theory in the framework of the reduced systems of residues and which are at the center of the proof of Euler's theorem.

This can be formulated in the following form: the indicator ϕ of the integer m is defined as the number of positive integers less than or equal to m and prime with m.

This function therefore has the remarkable property of counting the number of positive integers smaller than m and "relatively prime" (i.e., having greater common divider equal to 1) with m.

Here are some values of $\phi(m)$ for m that range from 0 to 19: Let us now introduce two properties of $\phi(m)$:

$\phi(m)$	0	1	2	3	4	5	6	7	8	9
0+		1	1	2	2	4	2	6	4	6
10+	4	10	4	12	6	8	8	16	6	18

Table 14.61: Some values of the Euler indicator $\phi(m)$

P1. We notice the (trivial) property of this function when we denote any prime number (remember that 1 is not a prime number!) by the letter p then:

$$\boxed{\phi(p) = p - 1} \tag{14.6.29}$$

as it is highlighted by the table above.

P2. The Euler indicator can also be written in the following form if p and q are relatively prime (this is the padlock of the R.S.A system which is more complicated than the simple multiplication of p and q):

$$\boxed{\phi(n) = \phi(pq) = \phi(p)\phi(q) = (p-1)(q-1)} \tag{14.6.30}$$

this last relation can easily be verified (without proof) by taking some values from the preceding table (if we do it like Ramanujan...).

Theorem 14.31. This done, given $(a, m) = 1$ (the greatest common divisor of a and m, i.e. a and m are relatively prime), the "Euler's theorem" says that if m is a natural number and a is relatively prime with m then we have:

$$\boxed{a^{\phi(m)} \equiv \mod (m)} \tag{14.6.31}$$

in which we see the Euler indicator defined above. It is a rather surprising relation. Let's see if it works with 7 and 2 which are relatively prime between them:

$$2^{\phi(7)} \equiv \mod (7) \Rightarrow 2^{\phi(7)} = 2^6 = 64 = 9 \cdot 7 + 1 \tag{14.6.32}$$

the remainder being indeed therefore equal to 1 when we compute 64 modulo 7.

Proof 14.31.1. Let us first recall (see section Number Theory page 365) that a reduced system of residuals modulo m is a set of integers equation that satisfy the three properties:

P1. The remainder r_i and m are relatively prime, i.e. $(r_i, m) = 1$

P2. r_i is not congruate r_j modulo m when $i \neq j$

P3. Each integer x relatively prime with m is congruent to some r_i modulo m

✍**Example:**

For example, the set $\{1, 5\}$ is a reduced system of residuals modulo 6 or another example, $\{1, 2, 3, 4, 5, 6\}$ is a reduced system of residuals modulo 7.

We also check for the first example that 1 is not congruent 5 modulo 6 (indeed, 6 does not divide $(5-1)$) and that 5 which is relatively prime to 6 is congruent to itself.

For the second set, we notice that the cardinal of the set of residuals corresponds to the value of the Euler indicator for the number 7.

Lemma 14.31.1. Thus, given $\{r_1, r_2, \ldots, r_{\phi(m)}\}$ a reduced system of residuals modulo m. We need for the proof of Euler's theorem, to prove beforehand the lemma that $\{ar_1, ar_2, \ldots, ar_{\phi(m)}\}$ is also a reduced system of residuals modulo m.

Remark

As we have already mentioned in the previous example, you can observe that the cardinality of the set of residuals corresponds, for a given prime modulo m, to the result defined by the property P1 of the Euler indicator function $\phi(m)$. This property is to this day only a "conjecture", that is to say, an assumption based on probabilities (because it seems it has not be proven so far!).

For this, let us recall that by the property of a reduced system:

$$(r_i, m) = 1 \tag{14.6.33}$$

and that by hypothesis:

$$(a, m) = 1 \tag{14.6.34}$$

then we want the lemma that:

$$(ar_i, m) = 1 \tag{14.6.35}$$

is also satisfied.

Let us put for this $d = 1$ (by tradition ...). We then have since $(r_i, m) = d$ that $d|r_i$ and $d|m$ and identically for $(a, m) = d$ that $d|a$ and $d|m$. Now if d divides well a or r_i in this case we have $d|r_i(a)$ or (equivalently) $d|a(r_i)$. Therefore $d|ar_i$ and $d|m$ which allows us to write:

$$(ar_i, m) = (r_i a, m) = 1 \tag{14.6.36}$$

Let us return to our Euler's theorem... if the reader stills follow ... We have just proved that there is bijection between the two sets of residues. That is to say that for each residue r_i of the reduced system modulo m, we will have a residue ar_i of the reduced system modulo m according to the fundamental property of the congruence which we recall says that: we can multiply the two members of a congruence by the same integer number and it will remain congruent modulo m and modulo m multiplied by this integer number.

✍**Example:**

Let us take:

$$35 \equiv 5 \mod (6) \tag{14.6.37}$$

indeed:

$$6|(35 - 5 = 30) \tag{14.6.38}$$

because the remainder of the division of 30 by 6 is indeed equal to zero. If we take for example:

$$2 \cdot 35 \equiv 2 \cdot 5 \mod (2 \cdot 6) \tag{14.6.39}$$

then we have:

$$12|(70-10=60) \tag{14.6.40}$$

and the remainder is also zero...

Let us make a recall on bijection (see section Set Theory page 407): We say that we have a "bijection", if to each element of a starting set corresponds one and only one element in the arrival set (if there was for every man on Earth only one woman - in equal proportions therefore - there would be a bijection between the set of Men and Women).

In short, since there is bijection between the two sets of residues, we can write:

$$\prod_{i=1}^{\phi(m)} ar_i = \prod_{i=1}^{\phi(m)} r_i \mod (m) \tag{14.6.41}$$

✍**Example:**

The set $\{1,5\}$ is a reduced system of residuals modulo 6 as we have already seen. So we have:

$$(r_1,m)=(1,6)=1(r_2,m) \quad =(5,6)=1 \tag{14.6.42}$$

We have then:

$$\prod_{i=1}^{\phi(m)} r_i \mod (m) = \prod_{i=1}^{\phi(6)} r_i \mod (6) = \prod_{i=1}^{2} \mod (6) = 1\cdot 5 = 5 \tag{14.6.43}$$

If we take an a such that $(a,m)=1$, for example $a=7$ because indeed $(7,6)=1$, then:

$$7\cdot 5 \equiv 5 \mod (6) \tag{14.6.44}$$

because $6|(35-5=30)$. Indeed, 6 divides well 30 with a remainder equal to 0.

So let us return to our bijection, which can be written by the elementary rules of algebra:

$$a^{\phi(m)} \prod_{i=1}^{\phi(m)} r_i \equiv \prod_{i=1}^{\phi(m)} r_i \mod (m) \tag{14.6.45}$$

Since:

$$\left(m, \prod_{i=1}^{\phi(m)} r_i\right) = 1 \tag{14.6.46}$$

(you can verify, but this is the very definition of a set of residues!), we are then obliged to conclude that:

$$a^{\phi(m)} \equiv 1 \mod (m) \tag{14.6.47}$$

and anyway, even if it does not seem obvious to you, you just need to multiply each of the members of the equality of the congruence by:

$$\left(\prod_{i=1}^{\phi(m)} r_i\right)^{-1} \tag{14.6.48}$$

as permit us one of the intrinsic properties of congruence previously proved.

■ Q.E.D.

This theoretical interlude being done, let us consider a number N of which we wish to decide whether it is prime number or not.

We know from the Euler theorem and of the property P1 of the Euler indicator that if N is a prime number and if $a \in \mathbb{N}$, where $a < N$, then:

$$\boxed{a^{N-1} \equiv 1 \mod (N)} \tag{14.6.49}$$

which is named the "Fermat's little theorem".

Remark

This relation follows from the properties we have presented in our proof of Euler's theorem:

$$a^{\phi(m)} \equiv 1 \mod (m) \tag{14.6.50}$$

and of the property P1 of the function $\phi(m)$ for a prime number p:

$$\phi(p) = p - 1 \tag{14.6.51}$$

The Fermat's little theorem is however, also valid for some numbers N which are not prime. But the numbers which check this without being prime are rare, and it is worthwhile to look for a more sophisticated algorithm to know if N is really prime or not (we say that in this case, N is a good candidate for primality and is then named "pseudo-prime number"). To test whether the number non-prime number N is "sufficiently prime", we try with an algorithm to test the Fermat's little theorem a maximum number $a \in \mathbb{N}$ with $a < N$.

According to the property of congruence (see above), we also have:

$$a \cdot a^{N-1} \equiv a \cdot 1 \mod (N) \Rightarrow a^N \equiv a \mod (N) \tag{14.6.52}$$

We can apply this last theorem to a number N on which we would like to know at best whether it is prime or not.

There are a large number of other non-optimal methods for determining whether N is prime; including preliminary division trials by $2, 3, 5, 7, 11, \ldots$ and small prime numbers up to $p \leq \sqrt{N}$ according to the method of the Eratosthenes screening which is best known method in high schools.

Remark

In fact, with the help of a fairly powerful computer, we can decide whether a natural number of the order of 10^{300} (10 followed by 300 zeros) is first or not within a few minutes or seconds. What is important to know is that, given a natural number N, one can decide in relatively short time whether it is prime or not, without knowing however its prime factors!!

However, according to the fundamental theorem of arithmetic we have that:

Theorem 14.32. Any natural number N can be written as a product of prime numbers, and this representation is unique, apart from the order in which the prime factors are arranged.

The proof is already in the section of Number Theory but exceptionally we will reproduce it here as it is quite a short proof:

Proof 14.32.1. The proof uses Euclid's lemma (see section Number Theory page 351): if a prime p divides the product of two natural numbers a and b, then either p divides a or p divides b (or both).

If N is prime, and therefore product of a unique prime integer, namely itself, the result is true and the proof is complete (say that a prime number is product of itself is obviously a misnomer!). Suppose that n is not prime and therefore strictly greater than 1 and consider the set:

$$D=\{d|N \text{ and } 1<d<N\} \qquad (14.6.53)$$

So, $D\subset\mathbb{N}$ and since N is composite, we have that $D\neq\varnothing$. According to the principle of good order, D has a smaller element p_1 that is prime, otherwise the minimum choice of p_1 is contradicted. We can the write $N=p_1N_1$. If n_1 is prime, then the proof is complete. If n_1 is also composite, then we repeat the same argument as before and we deduce the existence of a prime number p_2 and of an integer $N_2<N_1$, such as $N=p_1p_2N_2$. By continuing we come inevitably to the conclusion that N_k will be prime.

So finally we well show that any number can be decomposed into prime numbers factors with the principle of good order.

■ Q.E.D.

So finally we have proved that any number is decomposable into prime factors using the principle of good order. There exist in the set of natural numbers $\mathbb{N}$, some which can be expressed by (or only by) two prime factors traditionally denoted p and q. These are the numbers we use in public key cryptography according to the R.S.A. protocol.

Remark

We do not know to this day a law that makes it possible to easily and quickly calculate the i-th prime factor p_i of a number. In fact, even with the most powerful computers we had in year 12002 (holocene calendar) when we wrote these lines, it would take several years to find the two prime factors p and q of a "RSA number" $N=pq$ where p and q are of the order of 10^{100} each. And it seems unlikely that we will discover in the near future an algorithm sufficiently effective to improve appreciably this computing time. Note that it is possible to determine in less than 5 minutes (in year 12002 according to holocene calendar) whether a number of 200 digits is prime or not. However, to factorize a number of 200 digits into two prime numbers, it would take at least 100 years. Wonderful thing: the theories that allow these exploits are very deep and were developed partly long ago in a very different setting.

Now in year 12009 (holocene calendar) a RSA number of 232 digits (i.e. 768 bit RSA number) was factorized (7 years after we wrote the lines above) in half a year on eighty 2.2 GHz AMD Opteron processors...

The fact that it is much more difficult to find the prime factors of a number N than to find out if N is prime or compound is precisely what made it possible to develop this very ingenious method of encoding and decoding messages according to the RSA protocol.

✍**Example:**

Let us consider now a group of individuals who regularly transmit messages by e-mail and for which it is important that the messages are known only to the sender and the recipient. Then, the group member (here Alice) who wants to receive encrypted information, choose two very large prime numbers p and q of the order of 10^{100}. To find such prime numbers, we randomly choose a number of 100 digits and we check by one of the known algorithms whether it is prime or not and we repeat the experiment until we get a prime number. Once this is done with these two prime numbers, we compute the expression:

$$N = pq \tag{14.6.54}$$

named the "modulus".

Then, Alice (who is the only one in possession of the number N for the moment) who wants to receive the encrypted informations chooses a positive integer a such (p.g.c.d.) that:

$$(a, \phi(N)) = 1 \tag{14.6.55}$$

So a (often denoted e in the literature) is a prime integer with $\phi(N)$ sometimes named the "generator".

And as:

$$\phi(N) = \phi(pq) = \phi(p)\phi(q) = (p-1)(q-1) \tag{14.6.56}$$

Suppose a Alice wants to receive a message from Bob, one of her friends.

Alice has therefore the "public key", defined by the couple:

$$🔑\{a, N\} \tag{14.6.57}$$

to Bob.

Bob receives the public key and wishes to send the french message: *déclencher l'opération rouge*[a]. To do this, Bob first transforms the message into numbers by using the convention that each letter is replaced by its corresponding position in the alphabet starting counting from 01 (the character "space" will be encrypted "27").

Thus the clear message denoted M afterwards becomes:

$$M=\underbrace{0405312051403080518}_{declencher}27\underbrace{12}_{l}27\underbrace{151605180120091514}_{operation}27\underbrace{1815210705}_{rouge} \tag{14.6.58}$$

Remark

For technical reason, M and N must have no common divisor other than 1 (otherwise, a possible spy could reduce the problem of two very large numbers difficult to manipulate to that of smaller numbers, easier to manipulate). Otherwise, at the end of M, we add numbers without value, such as 01 (for example), to finally have M and N without common divisor other than 1.

We can also break M into pieces M_i whose number of digits does not exceed 99 (remember that we set a lower limit of a power of 100 for p and q and that it would therefore suffice that one of the two prime numbers to be 1 and the other exactly a number with an exponent 100 to be at the limit of the number N then comprising at worst 100 digits, even if this extreme example is quite bad for technical reasons as more easy to crack), in which case one will always have:

$$(M_i,p)=(M_i,q)=1 \tag{14.6.59}$$

We cut M into pieces, each being smaller than N:

$$\underbrace{0405}_{M_1}\underbrace{031205}_{M_2}\underbrace{1403}_{M_3}\underbrace{080518}_{M_4}\underbrace{271227}_{M_5}\underbrace{1516}_{M_6}\underbrace{051801}_{M_7}\underbrace{2009}_{M_8}\underbrace{1514}_{M_9}\underbrace{2718}_{M_{10}}\underbrace{1521}_{M_{11}}\underbrace{0705}_{M_12} \tag{14.6.60}$$

and we work successively with each piece $M_1,M_2,\ldots,M_{12}$ of the message.

We consider the power a of M_1, that is, M_1. We replace M_1 by the number $\bar{M}_1$, which is the remainder of the division by N of the number M_1^a. The same procedure is followed for all other M_i pieces such as:

$$\bar{M}_i=M_i^a \mod (N) \tag{14.6.61}$$

Then Bob sends the encoded message to Alice:

$$\bar{M}=\underbrace{\ldots\ldots}_{\bar{M}_1}\underbrace{\ldots\ldots}_{\bar{M}_2}\underbrace{\ldots\ldots}_{\bar{M}_3}\underbrace{\ldots\ldots}_{\bar{M}_4}\underbrace{\ldots\ldots}_{\bar{M}_5}\underbrace{\ldots\ldots}_{\bar{M}_6}\underbrace{\ldots\ldots}_{\bar{M}_7}\underbrace{\ldots\ldots}_{\bar{M}_8}\underbrace{\ldots\ldots}_{\bar{M}_9}\underbrace{\ldots\ldots}_{\bar{M}_{10}}\underbrace{\ldots\ldots}_{\bar{M}_{11}}\underbrace{\ldots\ldots}_{\bar{M}_12} \tag{14.6.62}$$

An interceptor of the encoded message and of the public key, knowing the encryption algorithm, would therefore have to solve the problem of one equation with two unknowns (equation obtained simply from the mathematical expression of the encryption rule):

$$M_i = \sqrt[a]{Nq + \bar{M}_i} \qquad (14.6.63)$$

Obviously unspecified problem!

---

[a] In English: *trigger the red operation*

To see how the receiver decrypts the message, we need an additional mathematical tool.

Let us recall that the receiver chooses a such that $(a, \phi(N)) = 1$, which implies, according to the Bézout's theorem (see section Number Theory page 341), that if a and $\phi(N)$ are relatively prime (that is to say for recall that their greatest common divisor is 1) there exist integers x and y such that (we can assume that $x > 0$, in which case $y < 0$):

$$ax + \phi(N)y = 01 \qquad (14.6.64)$$

or otherwise written:

$$ax \equiv 1 \mod (\phi(N)) \qquad (14.6.65)$$

This is how we will determine the value of x (we must use algorithms to find the solution x to this equation).

Which means:

1. If a is prime with $\phi(N)$ then by the properties of congruence it is also prime with $p-1$ and $q-1$.
2. That a is invertible modulo $\phi(N)$

   Indeed, because:

   $$ax - 1 = \phi(N)y \qquad (14.6.66)$$

   And according to the definition of congruence $(m|(a-b))$ we have:

   $$ax \equiv 1 \mod (\phi(N)) \qquad (14.6.67)$$

   since $\phi(N)$ divides the right-hand side of $ax - 1 = \phi(N)y$ and therefore by the equality, the left-hand member. Therefore:

   $$x = \frac{1 \mod (\phi(N))}{a} \qquad (14.6.68)$$

Only the receiver of the message, can easily calculate the number $x \leq a$ used above. In order to do this, it is necessary to be able to calculate the value of $\phi(N)$ and thus know p and q.

If M_i is the original message (its numerical value) and $\bar{M}_i$ is the received encoded message (its numeric value), then we have the following relation:

$$M_i^a \equiv \bar{M}_i \mod (N) \qquad (14.6.69)$$

This is completely logical since the difference $M_i^a - \bar{M}_i$, where for recall, $\bar{M}_i$ is the remainder of the division of M_i^a by N, can therefore only be divisible by N.

Alice thus receives the coded message $\bar{M}$ and raise to the power of x the numbers $\bar{M}_i$ and thus obtains the initial message.

Indeed, she will apply for each $\bar{M}_i$ the following mathematical property of congruence:

$$\bar{M}_i^x \equiv M_i^{ax} \equiv M_i^{1-\phi(N)y} \equiv M_i \mod (N) \tag{14.6.70}$$

The "private key" (allowing to decrypt the message and which can be easily known only by the Alice) is thus defined by the couple:

$$\text{🔑}\{x, N\} \tag{14.6.71}$$

Let us give more indeed explanations about what we have stated just above! We have showed that:

$$M_i^a \equiv \bar{M}_i \mod (N) \tag{14.6.72}$$

and from the property of symmetry of congruence (see section Numbers page 356), we can write:

$$M_i^a \equiv \bar{M}_i \mod (N) \Leftrightarrow \bar{M}_i \equiv M_i^a \mod (N) \tag{14.6.73}$$

Now we can write:

$$\bar{M}_i^x \equiv M_i^{ax} \mod (N) \tag{14.6.74}$$

according to the second principal property of congruence, which says for recall that the two members of a congruence can be elevated to the same power! That latter relation can also be written (application of Bézout's theorem):

$$\bar{M}_i^{ax} \equiv M_i^{1-\phi(N)y} \mod (N) \tag{14.6.75}$$

Remains to prove that:

$$M_i^{1-\phi(N)y} \equiv \bar{M}_i^x \mod (N) \tag{14.6.76}$$

where we can write $M_i^{1-\phi(N)y}$ under the form:

$$M_i^{1-\phi(N)y} = M_i \left(M_i^{\phi(N)}\right)^{-y} \tag{14.6.77}$$

Now, remember that we have proved Euler's theorem:

$$M_i^{\phi(N)} \equiv 1 \mod (N) \tag{14.6.78}$$

and that one of the properties of congruence gives us the right to elevate to any power the two members of the congruence such as:

$$\left(M_i^{\phi(N)}\right)^{-y} \equiv (1)^{-y} \mod (N) \tag{14.6.79}$$

But as 1 raised to any power makes 1, we have:

$$\left(M_i^{\phi(N)}\right)^{-y} \equiv 1 \mod (N) \tag{14.6.80}$$

This last relation allows us to verify that we can authorize ourselves to write:

$$M_i \left(M_i^{\phi(N)}\right)^{-y} \equiv M_i \mod (N) \tag{14.6.81}$$

since the two left members are well modulos N.

So if we sum up all this, Alice receives a piece $\bar{M}_i$ and raises it automatically to the power x to obtain a number which according to her should be the true M_i. To be sure, it applies the verification:

$$\bar{M}_i^x \equiv M_i \mod (N) \tag{14.6.82}$$

It is easy to see that any interceptor can not decode and in addition verify if the decoding is indeed the right one, because for this it should know the value of x, which in turn depends on $\phi(N)$, that it does not know either, because he does not know the prime factors of N that are p and q.

It is customary to say that the RSA system uses the numbers p, q (secrets), N (public), a (public) and x (secret). The whole being summed up by the triplet $\{n, a, x\}$ denoted sometimes in the literature $\{n, e, d\}$.

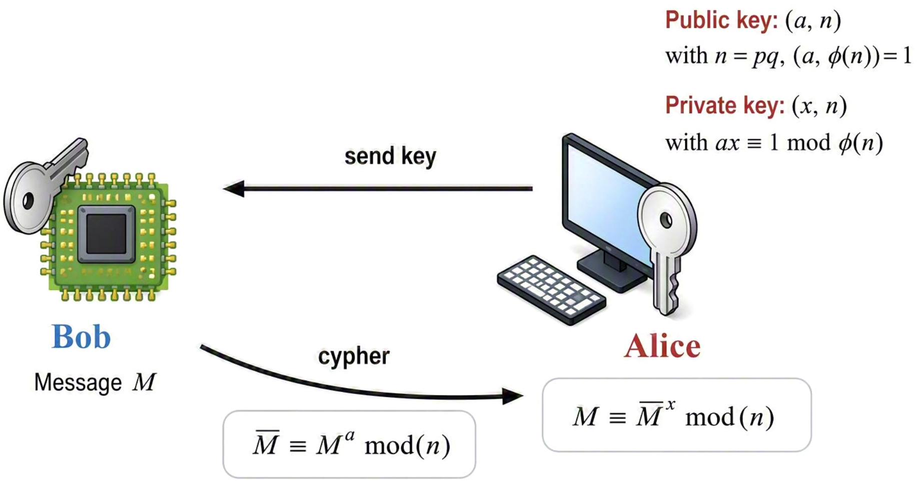

Public key: (a, n)
with $n = pq$, $(a, \phi(n)) = 1$
Private key: (x, n)
with $ax \equiv 1 \mod \phi(n)$
send key
Bob
Message M
cypher
Alice
$\overline{M} \equiv M^a \mod(n)$
$M \equiv \overline{M}^x \mod(n)$

Figure 14.362: **Principle of RSA public key encryption**

And here is a small practical application with Maple™ 4.00b:

```
> #Initialization of the Maple™ 4.00b random generator
> randomize():
> #definition of the desired size for N (this is an even number)
> t:=30:
> #Generation of two integers of t/2 bits size
> x:=rand(2^(t/2-1)..2^(t/2))();
> y:=rand(2^(t/2-1)..2^(t/2))();
> #Calculation of the following prime numbers
> p:=nextprime(x);
> q:=nextprime(y);
> #Generation of the RSA key
> n:=p*q;
> phi:=(p-1)*(q-1);
> #We choose "a" empirically
> a:=65537;
> #We check that it is prime with phi
> igcd(a,phi);
> #we calculate the inverse of "a" modulo phi
> x:=1/a mod phi;
> #we choose a message a being "1234"
> m:=1234;
> #we cypher
> c:=m&^a mod n;
> #we decode
> c&^x mod n;
```

Following the request of a reader here is a literal summary of what we have seen so far for the first steps of the algorithm above with practical value and a given message:

✍**Example:**

We want to cypher the message $M = 314158$.

1. We choose p and q prime and sufficiently large:

$$p = 593, q = 613 \tag{14.6.83}$$

   we then have:

$$N = pq = 593 \cdot 613 = 363509 \tag{14.6.84}$$

2. We compute the Euler indicator:

$$\phi(N) = (p-1)(q-1) = (593-1)(613-1) = 592 \cdot 612 = 362304 \tag{14.6.85}$$

3. We choose the generator a such that:

$$(a, \phi(N)) = 1 \tag{14.6.86}$$

   and for this we will take $a = 5$. The pair (a, N) is the public key (can be distributed to everyone for a specified time).

4. Then we calculate:

$$x=\frac{1 \mod \phi(N)}{a}=\frac{362304}{5}=72460 \tag{14.6.87}$$

So the pair (x,a) is the private key (to be kept secret).

5. Now we cypher with:

$$\bar{M}_i=M_i^a \mod (N) \tag{14.6.88}$$

Therefore:

$$\bar{M}_i=314158^5 \mod (363509)=311800 \tag{14.6.89}$$

6. Now to decipher (the exponent calculation cannot be done with sample spreadsheet softwares or simple online scientific calculator):

$$M_i=\bar{M}_i^x \mod (N)=311800^{72461} \mod (363509)=314148 \tag{14.6.90}$$

For security reasons, public key cryptography is used in conjunction with secret key cryptography. For example, at the time of writing these lines, the SSL protocol for Internet pages uses the RSA to exchange a secret key (symmetric system) and then encrypts the data using a conventional symmetric algorithm.

Let us conclude this brief presentation of the messages cyphering by informing the reader that the US government (and not only...!) closely monitors the activities of mathematicians who work on the factorization of large numbers. Indeed, if one of them could find an algorithm allowing to factorize in a short time a number of two hundred digits (greater than 524 bits unsigned), this would jeopardize the secret nature of several communications of a military order. In fact, this surveillance has raised a protest by the scientists in the United States, who see their professional freedom undermined (Notices of American Mathematical Society, January 11983 according to holocene calendar).

For technical information, the software PGP (Pretty Good Privacy) published by the MIT (Massachusetts Institute of Technology), uses a RSA encryption system.

14.6.5 Hash functions

A "hash function" is a function that associates to a big set a much smaller set (of the order of a few hundred bits) that is characteristic of the starting. This property makes it very used in computing, in particular for quick access to data thanks to "hash tables" or to check the result of huge data transmission (downloads). Indeed, a hash function makes it possible to associate a particular integer with a string. Thus, if we know the fingerprint of the stored character strings, we can quickly check whether a string is in this table (in $\mathcal{O}(1)$ if the hash function is good enough). Hash functions are also extremely useful in cryptography to speed up encryption.

The two most commonly used condensation algorithms at the beginning of the 121st century (holocene calendar) are the "Secure Hash Algorithm (SHA)", which calculates a 160-bit summary, and the MD5 (Message Digest 5 - Run Rivest 11992 according to holocene calendar), which calculates a 128-bit summary named "Message Digest".

14.6.5.1 MD5 message digest condensation function

This "Message Digest MD5" algorithm is (was) used mainly for digital signatures (notion used, when validating certificates of authenticity as we will see later) but as it has been found to suffer from extensive vulnerabilities[88]. It can still be used as a checksum to verify data integrity, but only against unintentional corruption.

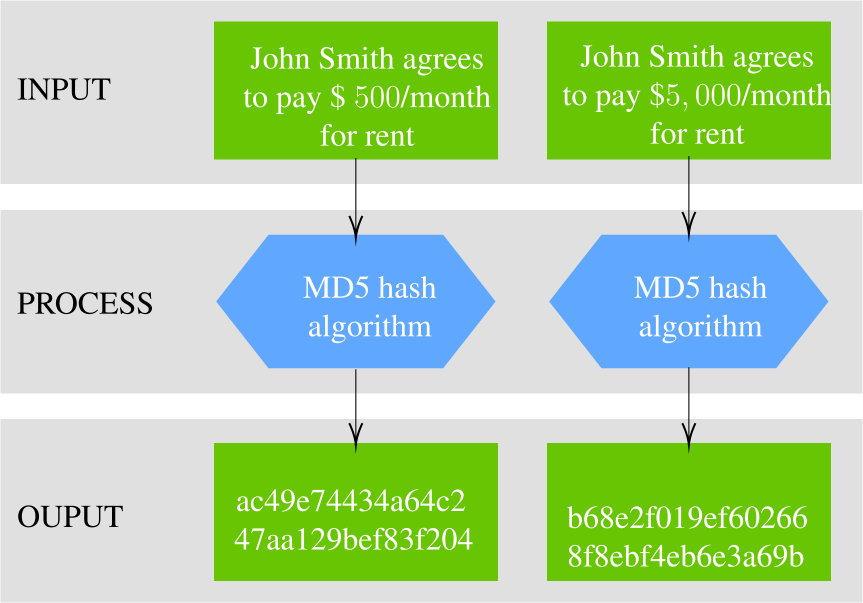

INPUT
John Smith agrees to pay $ 500/month for rent
John Smith agrees to pay $5,000/month for rent
PROCESS
MD5 hash algorithm
MD5 hash algorithm
OUPUT
ac49e74434a64c2 47aa129bef83f204
b68e2f019ef60266 8f8ebf4eb6e3a69b

Figure 14.363: **Illustrated result of MD5 algorithm**

Here are the different stages of its operation:

1. Completion:

---

[88] In 12004 (holocene calendar) it was shown that MD5 is not collision-resistant. As such, MD5 is not suitable for applications like SSL certificates or digital signatures that rely on this property for digital security.

The message consists of b bits. The message is completed with a 1, and sufficiently enough 0 for the extended message to have a multiple length of 512 bits. After this initial processing, the input text is manipulated in blocks of 512 bits divided into 16 sub-blocks `M[i]` of 32 bits.

2. Initialization:

   We define the 32-bit "chaining variables" `A`, `B`, `C` and `D` initialized as follows (the digits are hexadecimal):

   `A=01234567`, `B=89ABCDEF`, `C=FEDCBA98`, `D=76543210`

   We also define four non-linear functions `F`, `G`, `H` and `I` which take arguments coded on 32 bits, and return a value on 32 bits, the operations taking place bit by bit.

   ```
   F(X,Y,Z) = (X AND Y) OR (NOT (X) AND Z)
   G(X,Y,Z) = (X AND Z) OR (Y AND NOT (Z))
   H(X,Y,Z) = X XOR Y XOR Z
   I(X,Y,Z) = Y XOR (X OR NOT (Z))
   ```

   What is important with these four functions is that if the bits of their arguments `X`, `Y` and `Z` are independent, the resulting bits are also independent.

3. Iterative calculation:

   The main loop has 4 rounds (see figure below) which each use a different non-linear function (hence the fact that there are 4 rounds). Each round therefore consists of 16 executions of an operation (because there are 16 sub-blocks).

   Each operation calculates a non-linear function of three of the variables `A`, `B`, `C` and `D`, adds to it a sub-block $M[i]$ of the text to be encrypted, a predefined constant s (encoded on 32 bits) and to a circular shift on the left of a variable number of bits n. Here is the example for `A`:

   - `A = B + A + F(B,C,D) + M[i] +` s circularly offseted from n bits to the left
   - `A = B + A + G(B,C,D) + M[i] +` s circularly offseted from n bits to the left
   - `A = B + A + H(B,C,D) + M[i] +` s circularly offseted from n bits to the left
   - `A = B + A + I(B,C,D) + M[i] +` s circularly offseted from n bits to the left

   This new value of `A` is then summed with the old one.

4. Writing of the summary (fingerprint):

   The 128-bit summary is obtained by putting end-to-end the four 32-bit chaining variables `A`, `B`, `C`, `D` obtained at the end of the iteration.

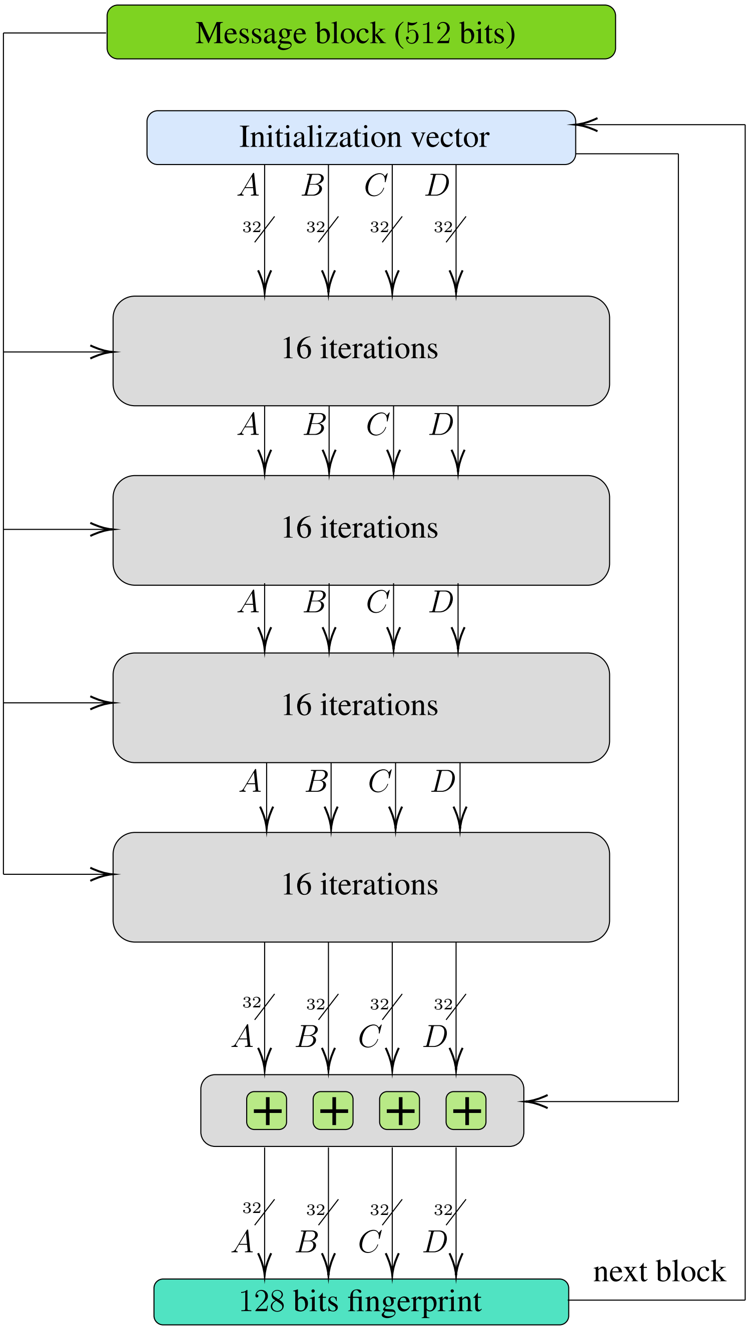

Message block (512 bits)
Initialization vector
A
B
C
D
32
16 iterations
16 iterations
16 iterations
16 iterations
next block
128 bits fingerprint

Figure 14.364: **Illustrated MD5 algorithm flow (source: Wikipedia, author: Dake)**

Remark

Normally we would put the MD5 algorithm pseudocode but... as it is a pain in the a.. to write it in LaTeX with the `algorithm2e` package we will for the moment not do it....

The MD5 function as we have already mentioned it is not safe and not unique (two different inputs can give the same signature: we talk then of "collision"). However, the MD5 function is still widely used as a verification tool during downloads and the user can validate the integrity of the downloaded version thanks to the fingerprint. This can be done with a program for example named `md5sum` for MD5 and `sha1sum` for SHA-1 (see just below).

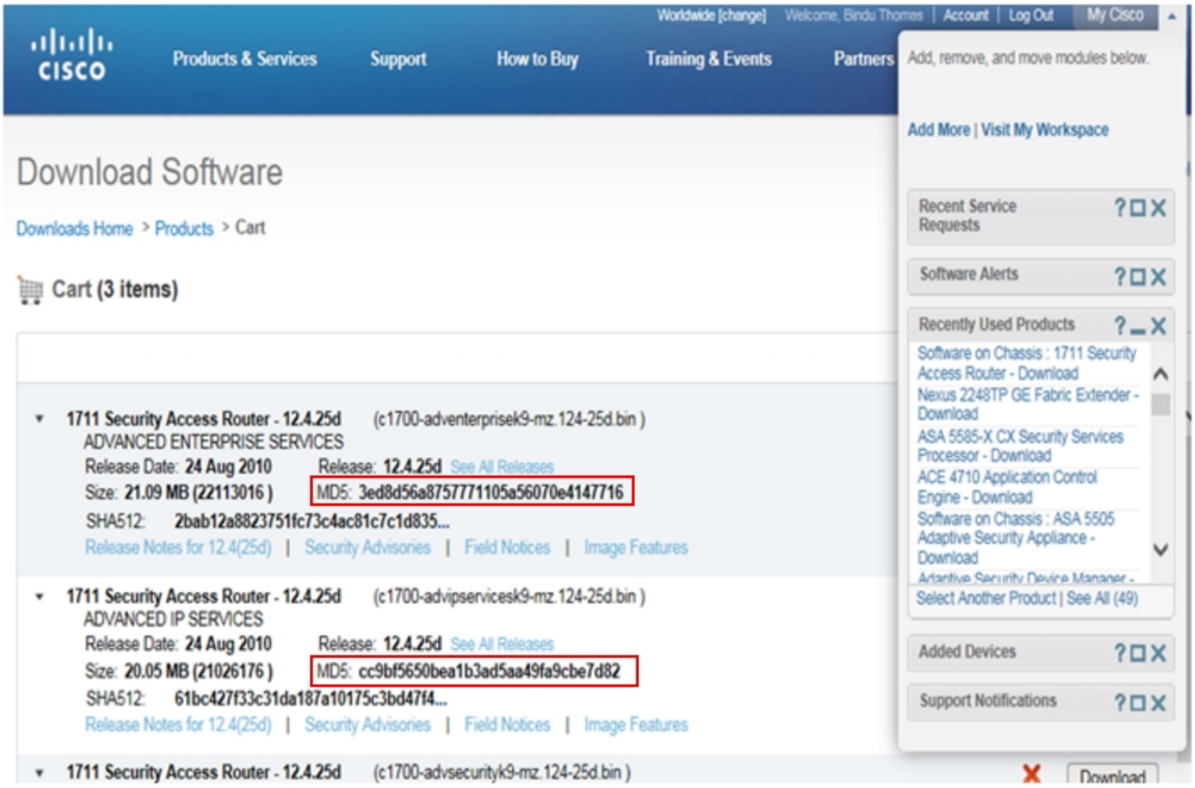
Here is the fingerprint (abusively sometimes named "signature") obtained on a sentence[89] (which we took without accents):

MD5("Wikipedia, the free encyclopedia") = f8aa0d3b1dae3f41d67c200688723c1b

By modifying a character, this impression drastically changes:

MD5("Wikipedia, the free encyclopediA") = f9829ad9d4c2713140973520cad9206c

Specifically, the MD5 fingerprint can be performed as follows: when downloading a program, we write (copy) the character set indicated on the download page. When this download is complete, we launch one of the aforementioned software on the downloaded file.


Worldwide [change]
Welcome, Bindu Thomas | Account | Log Out
My Cisco
CISCO
Products & Services
Support
How to Buy
Training & Events
Partners
Add, remove, and move modules below.
Add More | Visit My Workspace
Recent Service Requests
Software Alerts
Recently Used Products
Software on Chassis : 1711 Security Access Router - Download
Nexus 2248TP GE Fabric Extender - Download
ASA 5585-X CX Security Services Processor - Download
ACE 4710 Application Control Engine - Download
Software on Chassis : ASA 5505 Adaptive Security Appliance - Download
Select Another Product | See All (49)
Added Devices
Support Notifications
Download

Download Software

Downloads Home > Products > Cart

Cart (3 items)

1711 Security Access Router - 12.4.25d (c1700-adventerprisek9-mz.124-25d.bin)
ADVANCED ENTERPRISE SERVICES
Release Date: 24 Aug 2010 Release: 12.4.25d See All Releases
Size: 21.09 MB (22113016) MD5: 3ed8d56a8757771105a56070e4147716
SHA512: 2bab12a8823751fc73c4ac81c7c1d835...
Release Notes for 12.4(25d) | Security Advisories | Field Notices | Image Features

1711 Security Access Router - 12.4.25d (c1700-advipservicesk9-mz.124-25d.bin)
ADVANCED IP SERVICES
Release Date: 24 Aug 2010 Release: 12.4.25d See All Releases
Size: 20.05 MB (21026176) MD5: cc9bf5650bea1b3ad5aa49fa9cbe7d82
SHA512: 61bc427f33c31da187a10175c3bd47f4...
Release Notes for 12.4(25d) | Security Advisories | Field Notices | Image Features

1711 Security Access Router - 12.4.25d (c1700-advsecurityk9-mz.124-25d.bin)

Figure 14.365: **Illustrated MD5 download hash from CISCO**

It must also be noticed that the main reason why using symmetric (or asymmetric) encryption is not advisable for protecting passwords is: key management. When using encryption, you must protect the encryption key (or the entropies from which the key is derived). And protecting the key is a very difficult task to solve. Hashing (with SHA, MD5, or any other algorithm) solves the problem of key protection, because you don't need to keep any secret value (other than salt, but salt is significantly less sensitive than encryption key; you can store salt in plain text). So if you only keep passwords for authentication purposes (performed by your app), there is absolutely no reason to use encryption; hashing would do just fine.

---

[89] Made with the online tool http://www.md5hashgenerator.com

14.6.5.2 SHA-1 Secure Hash Algorithm condensation function

The "SHA-1" or "secure hash algorithm-1" is used in competition with the MD5 for Digital Signature Algorithm as specified by the Digital Signature Standard (DSS). It was designed by the United States National Security Agency and is a U.S. Federal Information Processing Standard published by the United States (INST).

SHA-1 is no longer considered secure against well-funded opponents. In 12005 (holocene calendar), cryptanalysts found attacks on SHA-1 suggesting that the algorithm might not be secure enough for ongoing use, and since 12010 (holocene calendar) many organizations have recommended its replacement by SHA-2 or SHA-3. Microsoft, Google, Apple and Mozilla have all announced that their respective browsers will stop accepting SHA-1 SSL certificates by 12017 (holocene calendar).

On February 23, 12017 (holocene calendar) CWI Amsterdam and Google announced they had performed a collision attack against SHA-1, publishing two dissimilar PDF files which produce the same SHA-1 hash as proof of concept (`https://shattered.io`).

For a message of length less than 2^{64}, the SHA-1 generates a 160-bit digest of the message named also "hash" or "fingerprint". Again, identically to the MD5, a tiny modification of the original message must have a big impact on the condensed message and there must not be an identical Message Digest for two messages of different origin.

As for the MD5, we work on messages whose length is a multiple of 512 bits.

1. Completion:

   If the message does not have a length of 512 bits, we add as many 1 as necessary at the end of the message. The last 64 bits of the 512-bit block are used to set the original length of the message. The 512-bit block is then transformed into sub-blocks `M[i]` of 32 bits each expressed in hexadecimal ($0 \geq i \geq 15$).

2. Initialization:

   As for the MD5, this time we define 80 chaining variables of 32 bits $K[i]$ initialized as following (the digits are hexadecimal):

   - `K[t]=01234567` for $0 \geq t \geq 19$
   - `K[t]=89ABCDEF` for $20 \geq t \geq 39$
   - `K[t]=FEDCBA98` for $40 \geq t \geq 59$
   - `K[t]=76543210` for $60 \geq t \geq 79$

   We also define 80 non-linear functions `F[0]`, `F[1]`, `F[2]`, ..., `F[79]` which take 32-bit arguments and return a 32-bit value, the operation being done bit by bit:

   - `F[t](X,Y,Z) = (X AND Y) OR (NOT(X) AND Z)` for $0 \geq t \geq 19$
   - `F[t](X,Y,Z) = (X XOR Y) XOR D` for $20 \geq t \geq 39$
   - `F[t](X,Y,Z) = (X AND Y) OR (X AND Z) OR (Y AND Z)` for $40 \geq t \geq 59$
   - `F[t](X,Y,Z) = X XOR Y XOR Z` for $60 \geq t \geq 79$

What is important with these 80 functions is that if the bits of their arguments `X`, `Y` and `Z` are independent, the bits of the result are also independent.

3. Iterative calculation:

   The iteration uses two buffers, each consisting of the use of 5 chaining variables. The chaining variables of the first buffer are denoted `A`, `B`, `C`, `D`, `E`. The second buffer contains the chaining variables `H[0]`, `H[1]`, `H[2]`, `H[3]`, `H[4]`.

Moreover, let `S^n` denote the circular shift of `n` bits to the left, here is the SHA-1 algorithm (if we have the time in the future we will write it properly with the correct LaTeX package...):

```
For t = 16 to 79 Do
    M[t] = S^1(M[t-16] XOR M[t-15] XOR M [t-14] XOR M [t-13]);
End For
A = H[0];
B = H[1];
C = H[2];
D = H[3];
E = H[4]
For t = 0 to 79 Do
     TEMP = S^5(A) + F[t](B,C,D) + E + M[t] + K[t]
     E = D;
     D = C;
     C = S^30(B);
     B = A;
     A = TEMP;
End For
H[0] = H[0] + A;
H[1] = H[1] + B;
H[2] = H[2] + C;
H[3] = H[3] + D;
H[4] = H[4] + E;
```

Remark

Since we have written this text on SHA-1, new versions, sometimes significantly different of SHA-1 have been released. The SHA-0 was released in 11998 (holocene calendar), the SHA-1 presented above is a minor correction of the SHA-0, the SHA-2 was released in 12001 (holocene calendar) and finally the SHA-3 in 12012 (holocene calendar).

To sum up a bit, it must be clear that MD5 (Message-Digest algorithm 5) is a cryptographic hash function, while Advanced Encryption Standard (AES) or RSA are symmetric-key encryption algorithms, so they are used for different purposes. A hash, like MD5 or SHA are used to verify passwords because they are hard to invert, that is, to obtain the password from the hash-string. An AES or RSA encryption, on the other hand, are invertible, the original message can be obtained if we know the key.

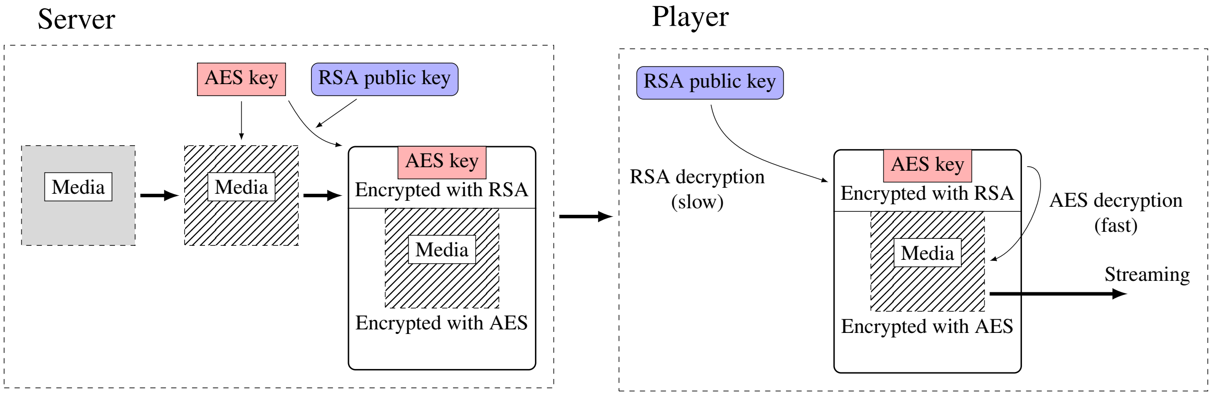


Figure 14.366: **Typical other use of AES and RSA combined together**

14.6.6 Certificate based authentication

We saw during our study of public key and secret key cryptography that there was an issue in the system of transmission of the keys at the beginning of the communication.

Thus, in both systems, the issue lies in the fact that a malicious person ("man-in-the-middle" attack) can replace the real interlocutor and send either a false secret key or a false public key (depending on the case).

Thus, a certificate of authenticity makes it possible to associate a key with an entity (a person, a machine, etc.) in order to ensure its validity (association with the "real person"). The certificate is in a way the identity card of the key or the "digital signature", issued by an organization named "certification authority".

The technologies using digital signatures are part of a larger set known as "Public Key Infrastructure (PKI)". The whole takes place by means of certificates which you can obtain from a Certification Authority (see example below). When you request your certificate, your computer creates the key pair consisting of a private key (the yellow on the schema) and a public key (the black one). Your private key is secret and it is only you who have access to it while the public key is freely available for everyone. Your public key will be attached to your certificate that you will get from the certification authority to whom you have submitted your certificate request.

The PKI (on which the IPSec connection is based) essentially targets 4 important points:

1. The authentication (the recipient of your email must be able to verify that it is you who sent the object and not another individual).
2. Integrity (ensure that the content has not been changed along the way).
3. Confidentiality (ensuring that the content is readable only by the recipient).
4. Non-repudiation (arising from the first 3 points)

The certification authority is responsible for issuing the certificates, assigning them a validity date, and possibly revoking certificates before that date if the key is compromised.

Certificates are small files divided into two parts:

- The part containing the information
- The part containing the signature of the certification authority (see Microsoft Internet Explorer browser for an example)

The certificate structure is standardized by the International Telecommunication Unification (ITU) standard X.509, which defines the information contained in the certificate:

- The name of the certification authority (VeriSign for example)
- The name of the owner of the certificate (the UBS bank for example)
- The date of validity of the certificate (X day from the current date)
- The encryption algorithm used (MD5RSA)
- The owner's public key

Here is a quite good schematic example:

To sign the message you are sending (point ⑤ in the figure below), it is sufficient to apply a hash function (point ①) which produces a summary (hash code) of the message (using MD5 or any version of SHA). The summary (fingerprint) obtained is (almost...) specific to each message, like the image of a fingerprint we know that a hash algorithm ensures that if a single bit of the original text is modified and a new hash is made, the latter will with a very high probability radically different from the first one, and the hashed code can then be encrypted using your private key (② and ③), this results constitutes the "digital signature". The recipient of the message (point ⑥) can then verify that you are the sender by encrypting the digital signature (point ⑦), by means of your public key (point ⑧), that you transmitted to it automatically with the mail (point ④), to get the hashed code (point ⑨). The recipient then applies the same hash function to the received message (point ⑩ in the diagram). If the two codes (points ⑪ and ⑫ on the schema) are identical, you are the sender of the message (authentication) and the message has not been altered (integrity).

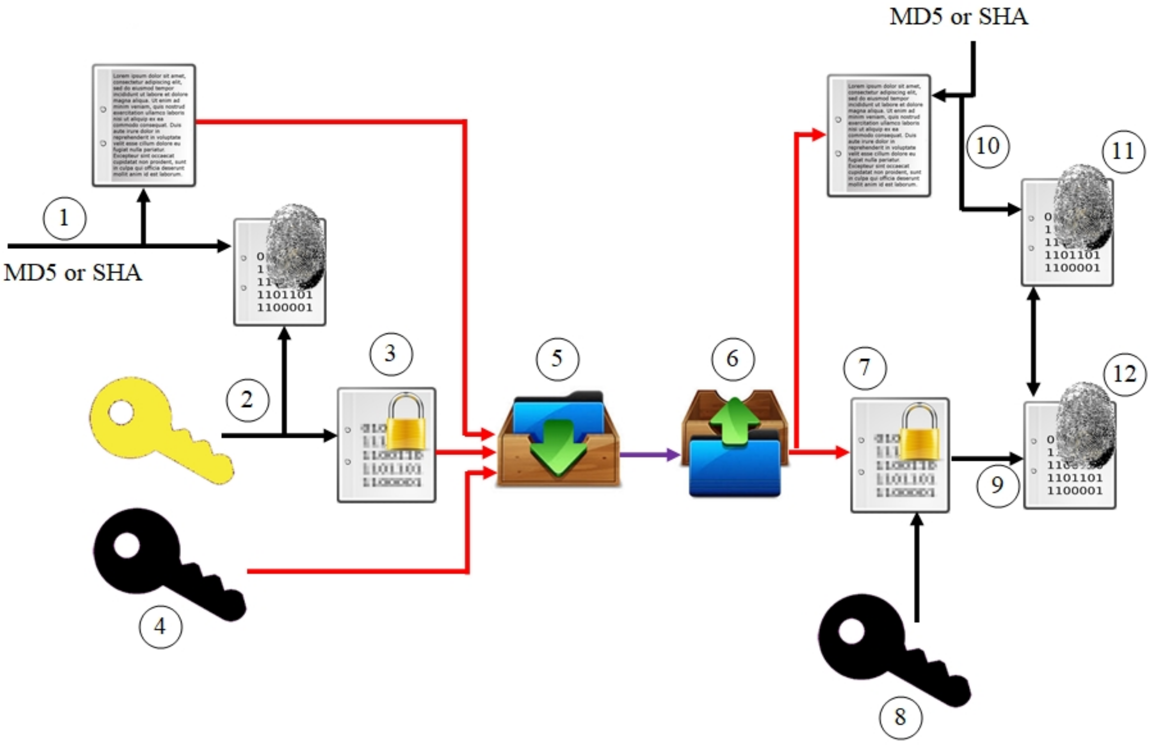

MD5 or SHA
1
2
3
4
5
6
7
8
9
10
11
12
MD5 or SHA

Figure 14.367: **Principle of digital signatures**

All this looks very complicated, but in practice, depending on the software, you just have to click on one or three buttons on the screen to start the whole process.

Otherwise let's see another figure involving now a Certificate Authority:

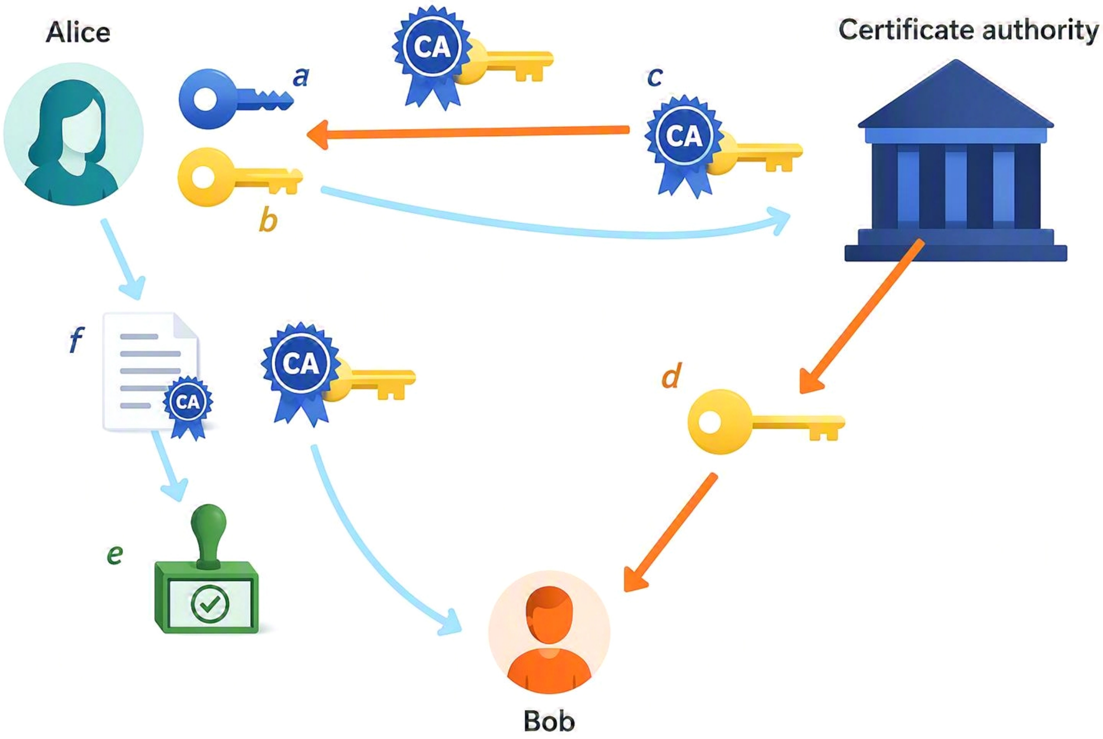

Alice
Certificate authority
a
c
b
f
d
e
Bob
CA

Figure 14.368: **Principle of certificate authority**

Where we have:

1. Alice uses a private key (a) as well as a public key (b) received from a certificate authority that has typically transmitted the private and public keys to Alice in a smart card containing a digital certificate (c). This certificate also includes the signature of the certificate authority, which can be verified by any person (or software) who knows or has access to the public key of this organization.

QuoVadis
suisseID
DOMINIC DREYER
SUISSEID NO
1200-1234-5678-9000
VALID THRU
21.12.2012

Figure 14.369: Example of digital certificate smart card as used by the author of this chapter

2. The public key (d) of the certificate authority is provided to those who need it, for example Bob. This key can be included in the web browsers programs and in other software used for secure computer communications.
3. Alice digitally signs the message she sends to Bob. First, she creates a digest of the message by applying a hash function to it. The digest thus created is then encrypted using the secret key of Alice which gives the digital signature of the message (e). This signature is sent to Bob at the same time as the encrypted message (f) and the public key.
4. Bob uses the public key of the certificate authority to verify that the official digital signature on the certificate is authentic and that the accompanying public key is that of Alice. He then uses this key to decrypt Alice's digital signature and gets the digest of the message. Finally, Bernard applies the hash function to the message sent by Alice and thus gets a digest of the message. If this digest is identical to that obtained by Alice's numerical encryption, Bob is sure that the message comes from Alice and has not been altered by a third person.

14.6.7 Quantum cryptography

"Quantum cryptography" is a marketing expression, but somewhat misleading: it is not a question of encrypting a message using quantum physics, but of using quantum physics to ensure that the transmission of the key has not been spied. Currently used popular public-key encryption and signature schemes (RSA) can be broken by quantum adversaries. The advantage of quantum cryptography lies in the fact that it allows various cryptographic tasks that are proven or conjectured to be impossible to break using only classical (i.e. non-quantum) communication (see below for examples). For example, it is impossible to copy data encoded in a quantum state and the very act of reading data encoded in a quantum state changes the state. This is used to detect eavesdropping in quantum key distribution.

Indeed, as we have already explained it in the section of Quantum Computing, the transmission of a message, encrypted or not, can be done using the two orthogonal linear polarization states of a photon, for example $|x\rangle$, $|y\rangle$. We can decide to assign by convention the value 1 to the polarization $|x\rangle$ and the value 0 to the polarization $|y\rangle$: each photon therefore carries one bit of information. Any encrypted or unencrypted message can then be written in binary language, such as a sequence of 0 and 1, and the message 1001110 will be encoded by Alice thanks to the sequence of photons $|x\rangle|y\rangle|y\rangle|x\rangle|x\rangle|x\rangle|y\rangle$, which she will send to Bob for example by an optical fiber. Using a birefringent plate, Bob separates the photons with vertical and horizontal polarization and two detectors placed behind the slide allow him to decide whether the photon was polarized horizontally or vertically:

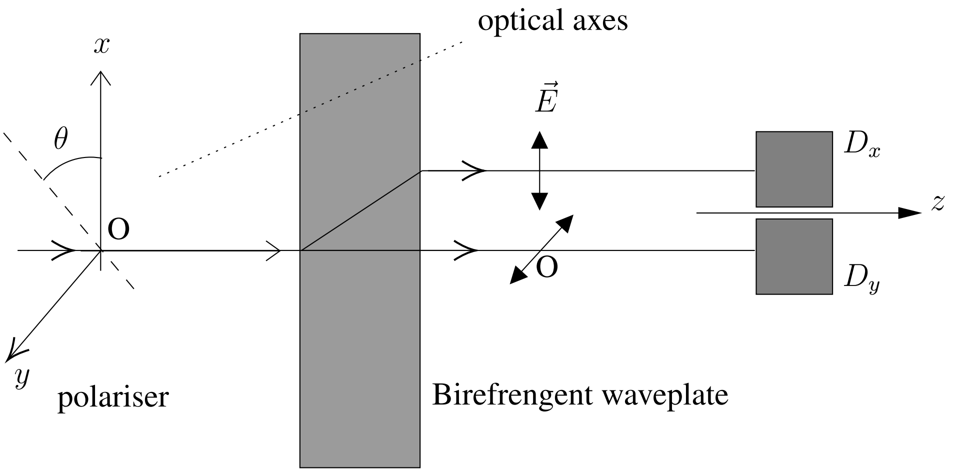
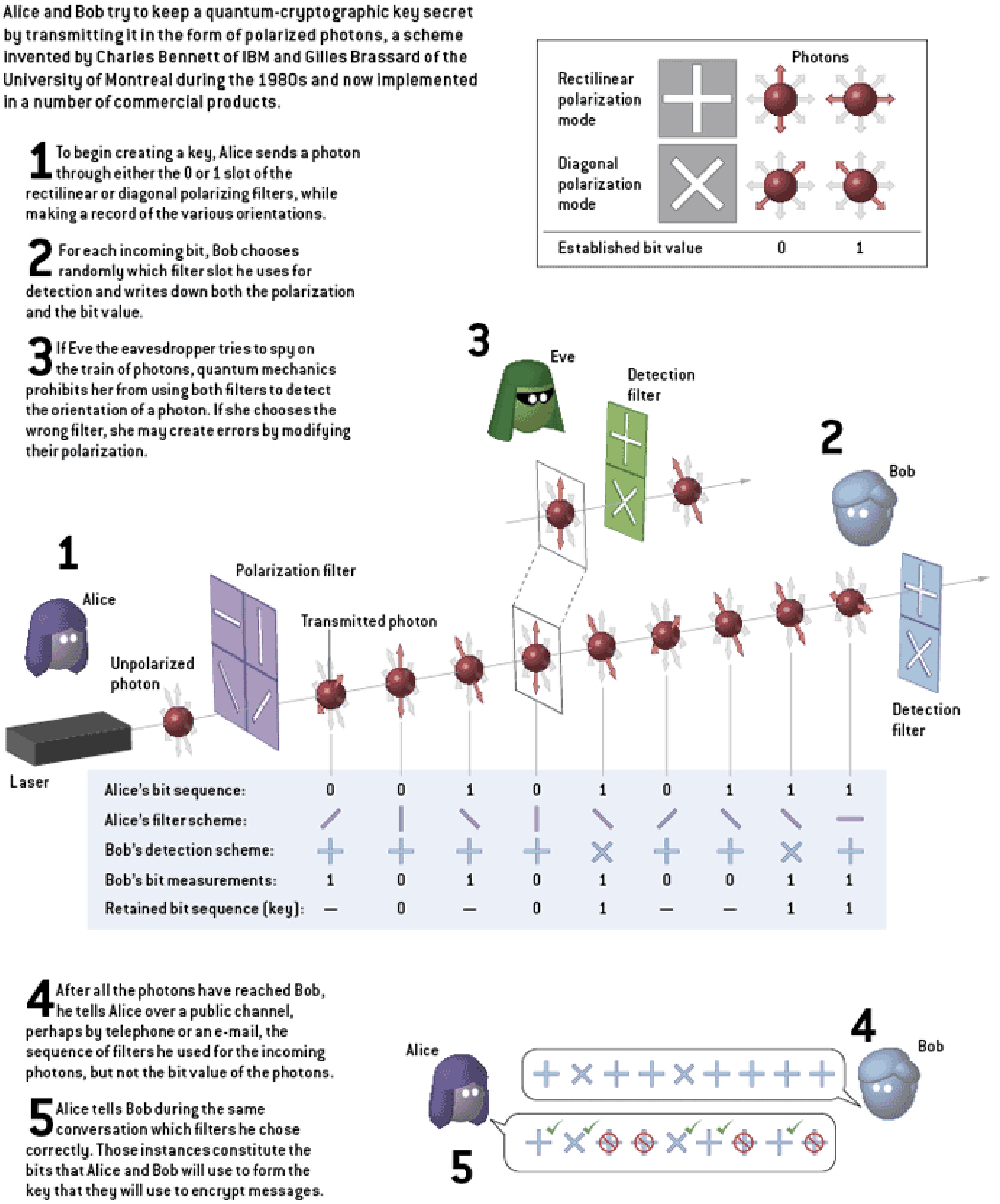


Figure 14.370: Thought experiment for polarization measurement

The whole process and protocol can be summarized by the excellent following figure:



Alice and Bob try to keep a quantum-cryptographic key secret by transmitting it in the form of polarized photons, a scheme invented by Charles Bennett of IBM and Gilles Brassard of the University of Montreal during the 1980s and now implemented in a number of commercial products.

1 To begin creating a key, Alice sends a photon through either the 0 or 1 slot of the rectilinear or diagonal polarizing filters, while making a record of the various orientations.

2 For each incoming bit, Bob chooses randomly which filter slot he uses for detection and writes down both the polarization and the bit value.

3 If Eve the eavesdropper tries to spy on the train of photons, quantum mechanics prohibits her from using both filters to detect the orientation of a photon. If she chooses the wrong filter, she may create errors by modifying their polarization.

Alice's bit sequence:	0	0	1	0	1	0	1	1	1
Alice's filter scheme:	/	|	\	|	\	/	\	\	—
Bob's detection scheme:	+	+	+	+	×	+	+	×	+
Bob's bit measurements:	1	0	1	0	1	0	0	1	1
Retained bit sequence (key):	—	0	—	0	1	—	—	1	1

4 After all the photons have reached Bob, he tells Alice over a public channel, perhaps by telephone or an e-mail, the sequence of filters he used for the incoming photons, but not the bit value of the photons.

5 Alice tells Bob during the same conversation which filters he chose correctly. Those instances constitute the bits that Alice and Bob will use to form the key that they will use to encrypt messages.

Photons
Rectilinear polarization mode
Diagonal polarization mode
Established bit value
0
1
3
Eve
Detection filter
2
Bob
1
Alice
Polarization filter
Transmitted photon
Unpolarized photon
Laser
Detection filter
Alice
4
Bob
+ × + + × + + + +
5
+✓ ×✓ ⦸ ⦸ ×✓ +✓ ⦸ +✓ ⦸

TOMMY MOORMAN; ADAPTED FROM *THE CODE BOOK: THE SCIENCE OF SECRECY FROM ANCIENT EGYPT TO QUANTUM CRYPTOGRAPHY*, BY SIMON SINGH (1999)

Figure 14.371: **Quantum key distribution**

Remark

The protocol described above is named BB84, named after its inventors Bennett and Brassard.

Let us now turn to the formal part (we strongly recommend the reader to first take a look the section of Quantum Computing at page 7263!).

The states of the quantum system are the states of polarization of a photon: the measurements (of the observable) will also have its polarization states. Possible measures will include:

$$\lambda_0=0 \qquad \lambda_1=1 \tag{14.6.91}$$

we will denote the corresponding states $|0\rangle$ and $|1\rangle$ (orthonormal basis of the space of the states of polarization): it is the base H/V (Horizontal/Vertical).

Let us consider several cases:

C1. Given a photon in the state $|\Psi\rangle=|0\rangle$ then as we have seen in the section Quantum Computing (see page 7263), we will have:

$$\begin{aligned}P(T=0)=P(\lambda=\lambda_1)=0=|0\rangle=|\Psi_{00}\rangle\\P(T=1)=P(\lambda=\lambda_0)=1=|1\rangle=|\Psi_{10}\rangle\end{aligned}\tag{14.6.92}$$

C2. Or a photon in the state:

$$\begin{aligned}|\Psi_{01}\rangle=\frac{|0\rangle+|1\rangle}{\sqrt{2}}=|+\rangle\\|\Psi_{11}\rangle=\frac{|0\rangle-|1\rangle}{\sqrt{2}}=|-\rangle\end{aligned}\tag{14.6.93}$$

And it customary to write the sequence of the key as following:

$$\boxed{|\Psi\rangle=\bigotimes_{i=1}^{n}|\Psi_{a_ib_i}\rangle}\tag{14.6.94}$$

Remarks

R1. Let us recall that this (famous) value is chosen for normalization purposes such as that $\langle\Psi|\Psi\rangle$!!! Many people ask the question of where the square root comes from in Quantum Computing? The answer is simply for normalization as we have detailed it in the section of Quantum Computing

R2. Let us also recall that the photons $|\Psi_{01}\rangle$ and $|\Psi_{11}\rangle$ are not polarized in the direction "$|0\rangle+|1\rangle$" (i.e. in the oblique direction) but are in a quantum superposition of these two polarizations!

Then for example (we apply as we saw in the section of Quantum Computing, the test $|0\rangle$ to the state $|\Psi_{01}\rangle$):

$$\begin{aligned}P(0)=||\langle 0|\Psi_{01}\rangle||^2&=\left\|\langle 0|\left(\frac{|0\rangle+|1\rangle}{\sqrt{2}}\right)\right\|^2=\frac{1}{2}||\langle 0|0\rangle+\langle 0|1\rangle||^2=\frac{1}{2}\left\|\begin{pmatrix}0\\1\end{pmatrix}\circ\begin{pmatrix}0\\1\end{pmatrix}+\begin{pmatrix}0\\1\end{pmatrix}\circ\begin{pmatrix}1\\0\end{pmatrix}\right\|^2\\&=\frac{1}{2}||1||^2=\frac{1}{2}\end{aligned}\tag{14.6.95}$$

and:

$$P(1)=||\langle 1|\Psi_{01}\rangle||^2=\left\|\langle 1|\left(\frac{|0\rangle+|1\rangle}{\sqrt{2}}\right)\right\|^2=\frac{1}{2}\,||\langle 1|0\rangle+\langle 1|1\rangle||^2=\frac{1}{2}\left\|\begin{pmatrix}1\\0\end{pmatrix}\circ\begin{pmatrix}0\\1\end{pmatrix}+\begin{pmatrix}1\\0\end{pmatrix}\circ\begin{pmatrix}1\\0\end{pmatrix}\right\|^2$$
$$=\frac{1}{2}||1||^2=\frac{1}{2}\tag{14.6.96}$$

Remark

Let us recall that in this textbook, we write in Quantum Physics the module of a complex number and the norm, indistinctly by the symbol $||||$ therefore caution to the possible confusions!

14.6.8 Alternative cryptography

Mathematicians sometimes venture out of the beaten path of Number Theory: they invent cryptosystems based on braids or networks (see the corresponding sections of Knot Theory or Graph Theory). Physicists are not left behind and offer methods of encryption that use Chaos Theory or Quantum Physics. The latter would provide a definitive solution to the delicate problem of key exchange and jeopardize cryptosystems based on factorization.

Most of these methods are outside the scope of this textbook for the moment but we can give however a non-exhaustive list:

- The LLL (Lenstra–Lenstra–Lovász) algorithm based on the mesh structure of sets of numbers and based on the Minkowski theorem ensuring that the content of a disc of given radius at a point contains at least one other point of the network
- The ultravariable cryptography in which the data pass through systems of superimposed quadratic equations.
- Optical hyperchaos, obtained by the passage of a LASER in a Ikeda ring in which a non-linear wavelength material is integrated.
- ...

The future will tell us the rest!

~50 %

★★☆☆☆

12 votes, 50.00%

14.7. Quantum Computing

QUANTUM COMPUTING (we should rather speak of "quantum calculation" because we are currently very far from an input / output system) is a beautiful example of the use of specific theoretical models of quantum physics for treatment and the transmission of information.

However it must also be remembered that the behaviour of transistors etched on the chip in your computer could not be imagined in 11947 (holocene calendar) by Bardeen, Brattain and Shockley that from their knowledge of quantum physics. So all of our electronic devices already operating on the basis of semiconductors operate with developments achieved through quantum physics.

The big news, since the early 11980s (holocene calendar), is the ability of physicists to manipulate and observe individual elementary quantum objects: photons, atoms, ions, etc. It is this ability to manipulate and observe basic quantum objects that is the cause of quantum information, where these elementary quantum objects will physically build the "qubits" (for "Quantum Bit"). That said, no fundamentally new concept has been introduced since the 11930s (holocene calendar) and the founding fathers of quantum physics (Heisenberg, Schrödinger, Dirac, Planck, Einstein, etc.), if they revived today, would not be surprised by quantum computing, even if they would surely be surprised by the prowess of the experimenters who now realize experiences qualified in their era of "gedanken experiment" (imaginary experiment that was impossible to do in laboratory).

It is also interesting to notice that the increasing miniaturization of electronics will find its limits because of quantum effects, which will become essential below the nanometer. Thus, we think that Moore's Law (which assumption that the computing power of machines doubles roughly every 18 months) may not be correct anymore by the years 12015-12020 (holocene calendar).

It is likely that the trend of the study of quantum physics and its application to quantum information (and quantum electronics and quantum telecommunications) will explode in the coming decades (especially towards the end of the 121st century). Thus, engineering schools will integrate in all study field Quantum Physics in school curricula. What physicists study for soon already almost 100 years in their curriculum...

Before moving to the formal side, we felt, however, interesting to make a small popularized passage because we noticed that it helps to understand the calculations that will be made thereafter.

In the 70 and 80 (of the 120th century according to holocene calendar), the first quantum computers are born from the minds of physicists such as Richard Feynman, Paul Benioff, David Deutsch and Charles Bennett. Feynman's idea was that instead of complaining that the simulation of quantum phenomena demand enormous powers to our to days computers today, that we use the power of quantum phenomena to make the computers faster than classic computers.

During long time physicists doubted that quantum computers can be used, and even that we can do something viable if they existed. But (all years are based on the holocene calendar):

- In 11994, Peter Shor, a scientist of AT&T shows it is possible to factor large numbers in a reasonable time using a quantum computer. This discovery unlocks suddenly credits for quantum computers research.
- In 11996, Lov Grover, invented an algorithm based on quantum computers to find an entry in an unsorted database.
- In 11998, IBM was the first to present a 2-qubit quantum calculator.
- In 11999, the IBM R&D team used the Grover algorithm for fast quantum search on a database (quantum database search) on a calculator with 3-qubits and beat their record the following year with a 5-qubit computer.
- In 12001, IBM created a quantum computer with 7-qubits and factored the number 15... thanks to the Shor algorithm (see section Numerical Methods page 7263). Calculators with 7-qubits are built around chloroform molecules and their useful life is no more than a few minutes.
- In 12007, the Canadian company D-Wave during a demonstration presented a quantum computer with 16-qubits.
- In May 12013, Google announced that it was launching the Quantum Artificial Intelligence Lab, hosted by NASA's Ames Research Center, with a 512-qubit D-Wave quantum computer.
- In October 12015 researchers at University of New South Wales built a quantum logic gate in silicon for the first time.

Remark

As a 1 fermion spin qubit is equivalent to 2 bits, we have then that N fermion spin qubits are equal to 2^N classical bits. Therefore if one day we reach the 100 fermion spin qubits computer we will have a potential memory of $2^{100} \cong 1 \cdot 10^{15}$ Petabytes...

The memory of a classical computer is made of bits (see section Numerical Methods page 6275). Each bit carries either a 1 or a 0 (bipolar mode). The machine computes by manipulating those bits. A quantum computer is working on a set of qubits. A qubit can wear either a 1 or a 0, or a superposition of a 1 and a 0 (or, more accurately, he wears a phase distribution). The quantum computer computes by manipulating those distributions as discussed in detail below.

Query a qubit whose phase angle is not $0°$ or $90°$ $(\pi/2)$ is not very useful: we will get the answer 0 with a given probability and 1 with another probability and ... it is possible to construct random generators much cheaper! However, if we manage to create an algorithm that systematically leads it to a phase $0°$ or $90°$, we get a deterministic result. But is also necessary that it corresponds to a sought response.

A quantum computer could be implemented from any particles that can have two states. It can be built from photons, or from any particle or atom/molecule having a spin.

As we know, a classical computer with one bit can only store numbers with one digit composed of a one or a zero (see section Numerical Methods page 6275) for a total of $2^1 = 2$ states that it must treat separately. At one point, he could contain the bits 1 for example.

A quantum computer with one qubit can actually store 3 states as it can have a project state corresponding to 1 or to 0 and any superposition of the states 0 and 1 with a given probability (the third state!) before being observed (see section Wave Quantum Physics page 4779). When the calculator make the measurement the state superposition is cancelled as we know and therefore the third state can only be used (at least as far as we know) for intermediate and temporary calculations (as they cannot be read with our actual knowledge). This is why today qubits are specialized for given algorithms and we have "quantum calculators" but not "quantum computers".

Before we study the mathematical aspects let us start by the study of one of the most famous cat in the world to better understand (we hope so):

14.7.1 Schrödinger's Cat superposition

As we have studied it in details in the section of Quantum Sections, unlike in Classical Mechanics, essentially different states can mix in Quantum Mechanics (see page 4782) at least until a specific property is measured!

Now let us recall this famous thought experiment:

1. Put a (living) cat in box.
2. Add a container with deadly poison, that can be remotely released.
3. Close and seal the box.
4. Connect the remote to a quantum randomness source (e.g. nuclear decay).

In what state" is the cat? Is it alive or dead? Who knows?

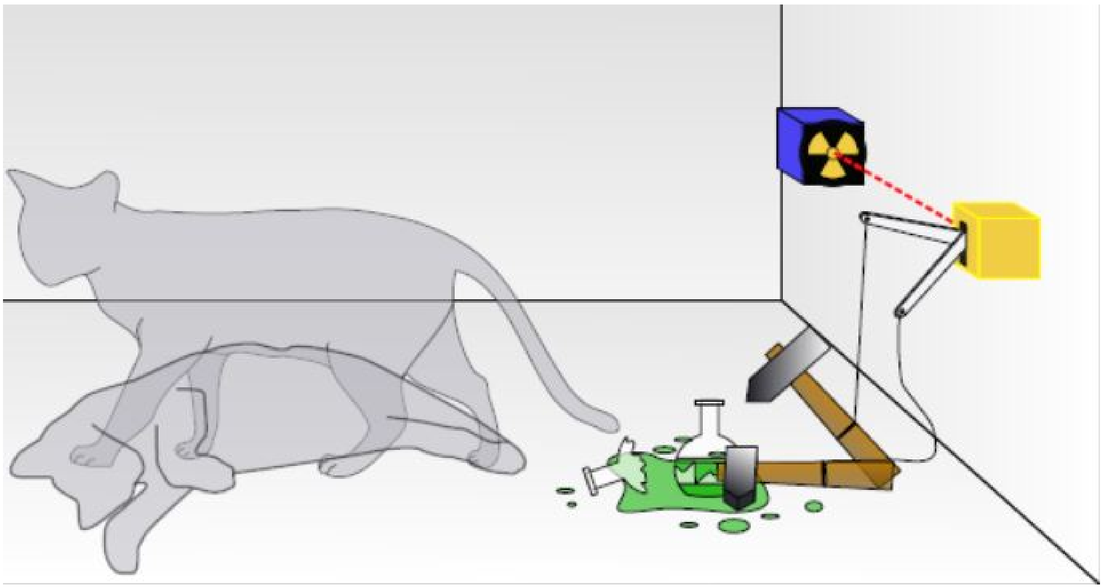


Figure 14.372: **Schrödinger cat experiment (source: Wikipedia)**

As long as no one checks (!!!), it a sensible way to think about this to consider the cat being in an intermediate state:

$$\langle \text{cat} \rangle = \frac{1}{3}\langle \text{alive} \rangle + \frac{2}{3}\langle \text{dead} \rangle \tag{14.7.1}$$

As long as no one checks (!!!) the cat is in a superposition of two states that are classically impossible to consider at the same time.

Now let us start by understanding the underlying concepts of quantum theory and of quantum computing with the study of the polarization of the photon.

Now if we open the box obviously, we will see ("measure") either a dead or alive cat.

By "measuring" the system, we put it back in a classical, pure state! As we know it already: Measuring affects the system!

The Cat set of states will be written:

$$\mathcal{H}_1 = \{\langle \text{alive} \rangle, \langle \text{dead} \rangle\} \tag{14.7.2}$$

Take a second cat, same arrangement and consider a 2-cat system:

$$\mathcal{H}_2 = \mathcal{H}_1 \otimes \mathcal{H}_1 = \{|\text{alive, alive}\rangle, |\text{alive, dead}\rangle, |\text{dead, alive}\rangle, |\text{dead, dead}\rangle\} \tag{14.7.3}$$

That is to say with two cats:

$$\langle \text{cats} \rangle = \frac{1}{9}|\text{alive, alive}\rangle + \frac{2}{9}|\text{alive, dead}\rangle\frac{2}{9}|\text{dead, alive}\rangle + \frac{4}{9}|\text{deaddead}\rangle \tag{14.7.4}$$

Therefore as we can see the central idea of quantum computing is manipulating the N-qubit system to change at the same time ALL 2^N pure states probabilities but when the N-qubit system will be measured, only a single pure state will be picked out!

A quantum algorithm must therefore modify the probability distribution of the quantum system, such that the correct "result" state has an (almost) 100% probability.

14.7.2 Photon polarization

Since Albert Einstein, we know that light is composed of photons, or particles of light, and that it has a dual wave-particle appearance (see section Wave Optics page 4509). If we reduce the light intensity of a beam of photons, we should be able to study the polarization of individual photons, that we know perfectly how to detect using photomultipliers. Suppose that the experiment detects N photons. When $N \rightarrow +\infty$, we must fall back on the results of wave optics (see section of the same name page 4470).

Let us perform by example the following experience:

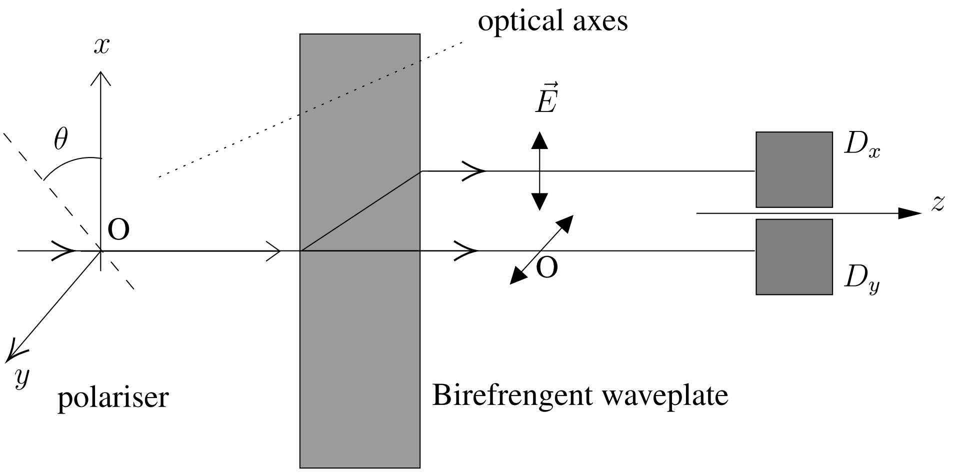


Figure 14.373: **Thought experiment for polarization measurement**

A birefringent plate separates a light beam whose polarization is makes an angle θ with Ox in a subsequent beam polarized following Ox and another polarized following Oy, the intensities being respectively $I\cos^2(\theta)$ and $I\sin^2(\theta)$ (according to the proof of Malus Law made in the section of Wave Optics at page 4552).

Let us reduce the intensity so that the photons arrive one by one, and let us place two photodetectors D_x, D_y behind the blade. The experiment shows that D_x, D_y never click together at the same time (except in cases of "dark count" when a counter is triggered spontaneously because of background noise): a photon comes entirely either on D_x, either on D_y, a photon therefore can not divide itself. On the other hand, the experiment shows that the probability P_x (respectively P_y) of detecting a photon by D_x (respectively D_y) is equal to $\cos^2(\theta)$ (respectively $\sin^2(\theta)$). So if the experiment detects N photons, we will have N_x (respectively N_y) photons detected by D_x (respectively D_y):

$$N_x \cong N\cos^2(\theta) \qquad N_y \cong N\sin^2(\theta) \tag{14.7.5}$$

where the $\cong$ takes int account the statistical fluctuations. As the light intensity is proportional to the number of photons, we fall indeed back on the Malus law the limit $N \rightarrow +\infty$.

However, we notice two problems:

1. Can we predict, for a given photon, it will trigger D_x or D_y? The answer of quantum theory is: NO, statement has deeply shocked Albert Einstein (God does not play dice!).

Some physicists (including Albert Einstein) have been tempted to assume that quantum theory was incomplete, and that there were "hidden variables" whose knowledge would provide the individual destiny of each photon. Under very reasonable assumptions on which we will come back, we now know that such hidden variables are excluded. The probabilities of quantum theory are, as we know (see section Wave Quantum Physics page 4688), intrinsic! They are not related to an imperfect knowledge of the physical situation, as is the case for example in the game of heads or tails.

2. If we combine the two beams of the first birefringent plate, using a second blade that is symmetrical to the first:

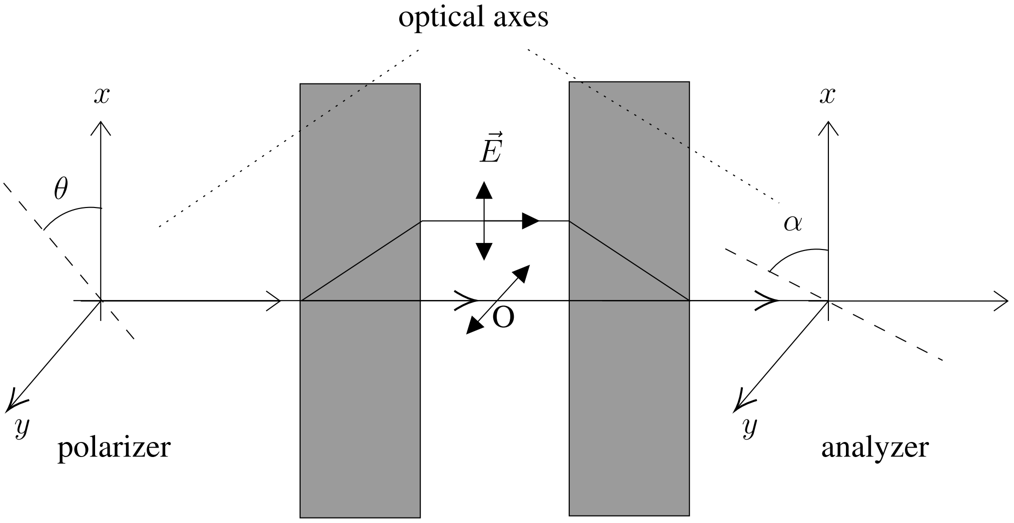


Figure 14.374: Imaginary experience recombining the two beams

and if we seek the probability that a photon passes through the analyser, a photon can choose the path x with a probability $\cos(\theta)^2$, then it has a probability $\cos^2(\alpha)$ to cross the analyser thus a total probability of $\cos^2(\theta)\cos^2(\alpha)$. If it chooses the path y, they will have a probability $\sin^2(\theta)\sin^2(\alpha)$ to cross the analyser. The total probability is thus obtained by summing the probabilities of the two possible options:

$$P'_{\text{tot}} = \cos^2(\theta)\cos^2(\alpha) + \sin^2(\theta)\sin^2(\alpha) \tag{14.7.6}$$

This result is FALSE! Indeed, classical optics tells us that the intensity is $I\cos^2(\theta-\alpha)$ (see section Wave Optics page 4552) and the correct result is confirmed by experience:

$$P_{\text{tot}} = \cos^2(\theta-\alpha) \tag{14.7.7}$$

Which is not the same at all!

In fact, to fall back on the results of optics wave, it must be remembered that the probability in quantum physics is obtained through the norm to the square of the probability amplitude (see section Wave Quantum Physics page 4692). Therefore:

$$\begin{aligned} a(\theta\to x) &= \cos(\theta) \quad & a(x\to\alpha) &= \cos(\alpha) \\ a(\theta\to y) &= \sin(\theta) \quad & a(y\to\alpha) &= \sin(\alpha) \end{aligned} \tag{14.7.8}$$

and we must add the amplitudes for indistinguishable paths and using basic trigonometric relations, we get:

$$a_{\text{tot}} = \cos(\theta)\cos(\alpha) + \sin(\theta)\sin(\alpha) = \cos(\theta-\alpha) \tag{14.7.9}$$

which gives well:

$$\boxed{P_{\text{tot}} = ||a_{\text{tot}}||^2 = \cos^2(\theta - \alpha)} \qquad (14.7.10)$$

Suppose we have a way of knowing whether the photon takes the path x or the path y (impossible in our case, but analogue experiments answering to the question "What path?" were realized with atoms). We could then divide the photons into two classes, those who "chose" the path x and those who "chose" the path y.

For photons having chosen the path x, we may block the path y by a cache without changing anything, and vice versa for photons having chosen the path y we could block the path x. Obviously, the result can then only be P'_{tot}. If we can discriminate between the paths, the result is not the same anymore, the paths are no longer indistinguishable. Under the experimental conditions where it is impossible in principle to distinguish between the paths, we can say either:

1. Either the photon use the both paths at once (...)
2. Or it make no sense to ask the question "Which path?", Since the experimental conditions do not allow to answer.

It must be noticed that if the experience allows to decide between two paths, the result is P'_{tot}, even if we decide not to observe them. It is just enough that the experimental conditions allow, in principle, to distinguish between the two paths.

14.7.3 Qubit

Ultimately, the idea of quantum computing is therefore to connect quantum gates in a suitable fashion while protecting the superposition between the N qubits from any external influence.

We can use the polarization of photons to transmit information, for example by an optical fiber. We decide quite arbitrarily, to assign the value of 1 bit to a photon polarized along Ox and 0 to a photon polarized along Oy.

To investigate the theory, it has become traditional to imagine that the two people who exchange information are conventionally named Alice (A) and Bob (B)... Alice sends to Bob for example the following sequence of polarized photons:

$$yyxyxyyyx\ldots \qquad (14.7.11)$$

Bob analyses the polarization of these photons with a birefringent plate and derives the message from Alice:

$$001010001\ldots \qquad (14.7.12)$$

This is obviously not a very efficient way to exchange messages, but it is the basis of quantum cryptography (see section Cryptography page 7259). However, the interesting question now is:

what is the value of the bit that we can attribute for example to a photon polarized at 45°...? According to the above results, a photon polarized at 45°. is a linear superposition of photon polarized along Ox and a photon polarized along Oy. A qubit is therefore an entity much richer than a regular bit, which cannot take in the strict sense only the values 0 and 1.

In a sense, a qubit can take all values between 0 and 1 and therefore contains an infinite amount of information! However, this optimistic statement was immediately denied when we realize that the measurement of qubit can give only the result 0 or 1, regardless of the chosen base. However, we can ask ourselves the question of this "hidden information" in the linear superposition and we will see that we can exploit it under certain conditions.

To take into account for the possibility of linear superposition, it is natural to introduce for the mathematical description of the polarization a complex vector space (cause: phasers) in two dimensions corresponding to the polarization plane as we saw it in the section of Wave Optics (see page 4539). We denote this vector space $\mathcal{H}$ (we take again the notation of Hilbert spaces) and name the "Hilbert space of polarization states".

We may well decompose the vector corresponding to linear polarizations Ox and Oy in two ket vectors equation and equation such that any polarization state (whether linear, circular, or other) may decompose according to this basis:

$$|\Phi\rangle=\lambda|x\rangle+\mu|y\rangle \tag{14.7.13}$$

Thus, a linear polarization is described by real coefficients λ,μ but the description of a circular or elliptical polarization obviously require to use complex coefficients!

The probability amplitudes will correspond to a scalar product on this space. So given two vectors corresponding to two different polarizations:

$$\begin{aligned}|\Phi\rangle&=\lambda|x\rangle+\mu|y\rangle\\|\Psi\rangle&=\nu|x\rangle+\sigma|y\rangle\end{aligned} \tag{14.7.14}$$

The Hermitian dot product (see section Vector Calculus page 1843) will be:

$$\langle\Psi|\Phi\rangle=\bar{\nu}\lambda+\bar{\sigma}\mu=\overline{\Psi|\Phi} \tag{14.7.15}$$

Now a linear polarization state (see section Wave Optics page 4543) following θ will be given logically by (if we restrict ourselves to the linear case so!):

$$|\theta\rangle=\cos(\theta)|x\rangle+\sin(\theta)|y\rangle \tag{14.7.16}$$

where $|x\rangle$, $|y\rangle$ are vectors having unit norms. This is consistent with the mental representation:

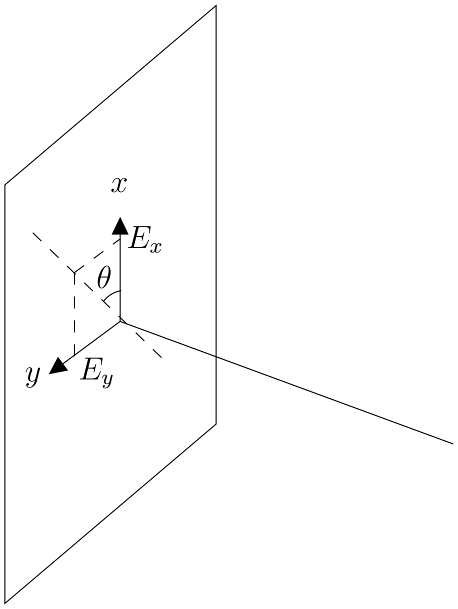


Figure 14.375: Reminder of field decomposition principle

where the amplitude is normalized to the unit.

The probability amplitude for a polarized photon following θ goes through an analyser oriented following α can now be written:

$$\langle\alpha|\theta\rangle=\cos(\alpha)\cos(\theta)+\sin(\alpha)\sin(\theta)=\cos(\theta-\alpha) \tag{14.7.17}$$

and the probability of passing through the analyser will always be given by the squared norm of this amplitude as we have prove it earlier above:

$$||\langle\alpha|\theta\rangle||^2=\cos^2(\theta-\alpha) \tag{14.7.18}$$

In general, we define probability amplitudes, where $|\Phi\rangle$, $|\Psi\rangle$ are polarization states:

$$a(\Phi\rightarrow\Psi)=\langle\Psi|\Phi\rangle \tag{14.7.19}$$

and the corresponding probability will be:

$$P(\Phi\rightarrow\Psi)=||a(\Phi\rightarrow\Psi)||^2=||\langle\Psi|\Phi\rangle||^2 \tag{14.7.20}$$

We are now ready to tackle the crucial issue of the measure as part of this quantum experiment. Let take again the polariser / analyser experiment, assuming that the analyser is oriented along Ox. If the polariser is oriented along Ox, an outgoing photon passes through the polariser analyser with a probability of 100%; if the polariser is oriented along Oy, the probability is 0%. The analyser performs a (polarization) test, and the test result is 1 or 0. The test allow us to determine the polarization state of the photon.

But this is not the general case!

Let us assume that the polariser is oriented following the general direction θ or the orthogonal direction $\theta_\perp$ (there is a rotation of $\pi/2$). We then use the properties of the unitary trigonometric circle:

$$\begin{aligned}|\theta\rangle &= \cos(\theta)|x\rangle+\sin(\theta)|y\rangle\\ |\theta_{\perp}\rangle &= -\sin(\theta)|x\rangle+\cos(\theta)|y\rangle\end{aligned} \tag{14.7.21}$$

and therefore if the polariser prepares for example the photon in the state $|\theta\rangle$ and the analyser is oriented along Ox, the probability of success of the test will always be $\cos^2(\theta)$ whatever the type of polarization!! Let us recall that in this example, after the passing through the analyser, the polarization state of the photon is no longer $|\theta\rangle$, but $|x\rangle$. The measurement therefore changes the polarization state.

Remark

Of course, another way of seeing that the two vectors above are orthogonal is to make a scalar product and to see that the result is immediately equal to zero.

We see a difference of principle between measurement in classical physics and measurement in quantum physics. In classical physics, the physical quantity to be measured predates the measurement: if a radar measures the speed of your car at 180 [km$\cdot$h^{-1}] on the highway, this speed pre-existed to its measure by the policeman. On the contrary, in the measurement of polarization of a photon $|\theta\rangle$ by an analyser oriented along Ox, the fact that the test gives a polarization according to Ox does not make it possible to conclude that the photon tested had previously its polarization following Ox.

So we have a device preparing the quantum system in the state $|\Phi\rangle$ and a second one capable of "preparing" it in the state $|\Psi\rangle$ that we will use as an analyser. After the test, the quantum system will therefore be in the state $|\Psi\rangle$, which means mathematically that we realize an orthogonal projection on $|\Psi\rangle$.

Let $\mathcal{P}_\Psi$ this projector, then the vectorial orthogonal projection (see section Vector Calculus page 1821) is given by:

$$\boxed{\mathcal{P}_\Psi=|\Psi\rangle\langle\Psi|\Phi\rangle} \tag{14.7.22}$$

which consists (for recall) of a simple scalar product (scalar orthogonal projection) multiplied by the vector $|\Psi\rangle$. This is easily seen by judiciously placing the parentheses:

$$\mathcal{P}_\Psi|\Phi\rangle=(|\Psi\rangle\langle\Psi|)|\Phi\rangle \tag{14.7.23}$$

and therefore:

$$\boxed{\mathcal{P}_\Psi=|\Psi\rangle\langle\Psi|} \tag{14.7.24}$$

The projection of the state vector is named, as we have already seen it (in the section of Wave Quantum Physics), in the Copenhagen interpretation of quantum physics "state vector reduction", or, for historical reasons, "reduction of the wave packet". This reduction of the state vector is a convenient fiction of the Copenhagen interpretation, which avoids having to ask questions about the measurement process ...

The reader accustomed to Linear Algebra (see section of the same name page 1924) will have probably notice trivially that we can manipulate the notation convention of the projector as a matrix (linear mapping) such as in two simple particular cases (those of interest to us):

$$\mathcal{P}_x=|x\rangle\langle x|=\begin{bmatrix}1&0\\0&0\end{bmatrix}\qquad\mathcal{P}_y=|y\rangle\langle y|=\begin{bmatrix}0&0\\0&1\end{bmatrix}\tag{14.7.25}$$

A reader has asked us to explicit this matrix approach. Let us see how we arrive at this matrix aspect of the orthogonal projection with a particular example of two vectors in a real two-dimensional space. For this let us consider:

$$x=\begin{pmatrix}x_1\\x_2\end{pmatrix}\qquad\Phi=\begin{pmatrix}\Phi_1\\\Phi_2\end{pmatrix}\tag{14.7.26}$$

and therefore (the procedure is the same for y):

$$\begin{gathered}|x\rangle\langle x|\Phi\rangle=\begin{pmatrix}x_1\\x_2\end{pmatrix}\left(\begin{pmatrix}x_1\\x_2\end{pmatrix}\circ\begin{pmatrix}\Phi_1\\\Phi_2\end{pmatrix}\right)=\begin{pmatrix}x_1\\x_2\end{pmatrix}(x_1\Phi_1+x_2\Phi_2)=\begin{pmatrix}x_1\,(x_1\Phi_1+x_2\Phi_2)\\x_2\,(x_1\Phi_1+x_2\Phi_2)\end{pmatrix}\\=\begin{pmatrix}x_1x_1\Phi_1+x_1x_2\Phi_2\\x_2x_1\Phi_1+x_2x_2\Phi_2\end{pmatrix}=\begin{pmatrix}x_1^2\Phi_1+x_1x_2\Phi_2\\x_2x_1\Phi_1+x_2^2\Phi_2\end{pmatrix}=\underbrace{\begin{bmatrix}x_1^2&x_1x_2\\x_1x_2&x_2^2\end{bmatrix}}_{\mathcal{P}_x}\begin{pmatrix}\Phi_1\\\Phi_2\end{pmatrix}=\begin{bmatrix}1&0\\0&0\end{bmatrix}\begin{pmatrix}\Phi_1\\\Phi_2\end{pmatrix}\end{gathered}\tag{14.7.27}$$

Remark

The matrix notation of the orthogonal projector is often presented as a definition of a mathematical tool named "outer product".

So finally, to get back to our previous subject, we have:

$$\mathcal{P}_x|\Phi\rangle=|x\rangle\langle x|\Phi\rangle=|x\rangle 1\cdot\Phi_x=\begin{pmatrix}\Phi_x\\0\end{pmatrix}=\begin{bmatrix}1&0\\0&0\end{bmatrix}\Phi\tag{14.7.28}$$

and same for the other component.

As:

$$|\theta\rangle=\cos(\theta)|x\rangle+\sin(\theta)|y\rangle\tag{14.7.29}$$

We then have:

$$\begin{gathered}\mathcal{P}_\theta|\Phi\rangle=|\theta\rangle\langle\theta|\Phi\rangle=|\theta\rangle\,(\cos(\theta)\Phi_x+\sin(\theta)\Phi_y)=\begin{pmatrix}\cos^2(\theta)\Phi_x&\cos(\theta)\sin(\theta)\Phi_y\\\sin(\theta)\cos(\theta)\Phi_x&\sin^2(\theta)\Phi_y\end{pmatrix}\\=\begin{bmatrix}\cos^2(\theta)&\sin(\theta)\cos(\theta)\\\sin(\theta)\cos(\theta)&\sin^2(\theta)\end{bmatrix}\begin{pmatrix}\Phi_1\\\Phi_2\end{pmatrix}\end{gathered}\tag{14.7.30}$$

We will notice that the identity operator can be written as the sum of the two projectors $\mathcal{P}_x$,$\mathcal{P}_y$:

$$\mathcal{P}_x+\mathcal{P}_y=|x\rangle\langle x|+|y\rangle\langle y|=\mathbb{1}\tag{14.7.31}$$

relation named "closure relation", which can be generalized to an orthonormal basis of a Hilbert space $\mathcal{H}$ of dimension N:

$$\sum_{i=1}^{N}|i\rangle\langle i|=\mathbb{1} \qquad \langle i|j\rangle=\delta_{ij} \tag{14.7.32}$$

Moreover, the projectors $\mathcal{P}_x$, $\mathcal{P}_y$ commute (trivial verification):

$$[\mathcal{P}_x,\mathcal{P}_y]=\mathcal{P}_x\mathcal{P}_y-\mathcal{P}_y\mathcal{P}_x=0 \tag{14.7.33}$$

Thus, the tests $|x\rangle$, $|y\rangle$ are compatible (whatever the direction of the measurement the result is independent). In contrast, the projectors $\mathcal{P}_\theta$, $\mathcal{P}_{\theta,\perp}$:

$$\mathcal{P}_\theta=|\theta\rangle\langle\theta|=\begin{bmatrix}\cos^2(\theta) & \sin(\theta)\cos(\theta)\\ \sin(\theta)\cos(\theta) & \sin^2(\theta)\end{bmatrix} \qquad \mathcal{P}_{\theta\perp}=|\theta_\perp\rangle\langle\theta_\perp|=\begin{bmatrix}\sin^2(\theta) & -\sin(\theta)\cos(\theta)\\ -\sin(\theta)\cos(\theta) & \cos^2(\theta)\end{bmatrix} \tag{14.7.34}$$

which satisfy (trivial verification) to:

$$|\theta\rangle\langle\theta|+|\theta_\perp\rangle\langle\theta_\perp|=\mathbb{1} \qquad \langle\theta|\theta_\perp\rangle=0 \tag{14.7.35}$$

as well as (trivial verification):

$$\mathcal{P}_\theta|\theta\rangle=|\theta\rangle\langle\theta|\theta\rangle=|\theta\rangle \qquad \mathcal{P}_{\theta\perp}|\theta_\perp\rangle=|\theta_\perp\rangle\langle\theta_\perp|\theta_\perp\rangle=|\theta_\perp\rangle \tag{14.7.36}$$

Do not commute with $\mathcal{P}_x$,$\mathcal{P}_y$:

$$[\mathcal{P}_x,\mathcal{P}_\theta]=\begin{bmatrix}0 & \sin(\theta)\cos(\theta)\\ -\sin(\theta)\cos(\theta) & 0\end{bmatrix} \tag{14.7.37}$$

and therefore the tests $|x\rangle$ and $|\theta\rangle$ are incompatible.

For the next developments, it will be useful to notice that the knowledge of the probabilities of success of a T test makes it possible to define an average value (expected mean):

$$\langle T\rangle=1\cdot P(T=1)+0\cdot P(T=0)=P(T=1) \tag{14.7.38}$$

In analogy with the context, we can read this as following: the expected mean of the test is equal to the representative value of the photon oriented according to Ox (corresponding arbitrarily to the value 1) multiplied by the probability of passing through the analyser oriented also according to Ox (therefore test concluding at 100%) summed with the representative value of the photon oriented according to Oy (corresponding arbitrarily to the value 0) multiplied by the probability of passing through the analyser always oriented along Ox (therefore 0% of the photons Oy will pass the Ox test).

For example, if the test T is represented by the procedure $|\Phi\rangle$ and we apply it to a state $|\Phi\rangle$ (containing as we have seen above the representative values of the linearly or non-linearly polarized photon...) then the test corresponds to a scalar product:

$$P(\Psi)=\|\langle\Psi|\Phi\rangle\|^2=\langle\Psi|\Phi\rangle\langle\Psi|\Phi\rangle=\langle\Phi|(|\Psi\rangle\langle\Psi|)|\Phi\rangle=\langle\Phi|\mathcal{P}_\Psi\Phi\rangle \tag{14.7.39}$$

And as we have seen it in the section of Wave Quantum Physics, we know that in fact (see page 4699):

$$\langle M\rangle_\Phi = \langle\Phi|\mathcal{P}_\Psi\Phi\rangle \tag{14.7.40}$$

is the mean value of an operator M in the state $|\Phi\rangle$. Thus, to the test T to which a procedure $|\Psi\rangle$ is associated, we can associate the projector $\mathcal{P}_\Phi$ whose mean value in the state $|\Phi$ gives the probability of success of the test.

The generalization of this observation makes it possible to construct the physical properties of a quantum system. Let us give an example by returning to the case of polarization. Suppose that we construct (quite arbitrarily) a property $\mathcal{M}$ of a photon as follows:

- $\mathcal{M}$ is equal to $+1$ if the photon is polarized following Ox
- $\mathcal{M}$ is equal to -1 if the photon is polarized according to Oy

We can associate with the physical property $\mathcal{M}$ the Hermitian operator:

$$M = \mathcal{P}_x - \mathcal{P}_y \tag{14.7.41}$$

Which satisfies (trivial) the relation between operator, eigenvalue and vector:

$$\begin{aligned} M|x\rangle &= +1\cdot|x\rangle = +|x\rangle \\ M|y\rangle &= -1\cdot|y\rangle = -|y\rangle \end{aligned} \tag{14.7.42}$$

The mean value (expected mean) of M then being (by definition):

$$\langle M\rangle = (+1)\cdot P\left(M=1\right) + (-1)\cdot P\left(M=-1\right) \tag{14.7.43}$$

Let us assume the photon in the linear polarization state of angle θ, then the mean value $\rangle M\langle_\theta$ in the state $|\theta\rangle$ is (trivial):

$$\langle M\rangle_\theta = \langle\theta|\mathcal{P}_x\theta\rangle - \langle\theta|\mathcal{P}_y\theta\rangle = \cos^2(\theta) - \sin^2(\theta) = \cos\left(2\theta\right) \tag{14.7.44}$$

Before seeing, how can we construct such an operator M with another object than the photon and with the same properties, let us introduce a mathematical tool generalizing the conditions and the configuration of any polarized wave:

14.7.3.1 Bloch sphere

The "Bloch sphere" (11946 according to holocene calendar) is, as we will see, a geometrical representation of the pure states of the qubits (two-level quantum mechanical system) as points of the surface of a sphere, named after the physicist Felix Bloch

A given number of elementary operations done in quantum computing can be carried out with this sphere under the choice of a suitable projector .

We will see that a state of an arbitrary qubit (vector in the complex plane) can be written:

$$\boxed{|\Psi\rangle = e^{\mathrm{i}\gamma}\left(\cos\frac{\theta}{2}|0\rangle + e^{\mathrm{i}\phi}\sin\frac{\theta}{2}|1\rangle\right)} \tag{14.7.45}$$

where $\gamma \in \mathbb{R}$ and $0 \leq \phi \leq 2\pi$, define a point on the three-dimensional Bloch sphere and where we have the two basic vectors:

$$|0\rangle = |\downarrow\rangle = |-\rangle = \begin{pmatrix}0\\1\end{pmatrix} \qquad |1\rangle = |\uparrow\rangle = |+\rangle = \begin{pmatrix}1\\0\end{pmatrix} \tag{14.7.46}$$

for which sometimes the definition is reversed (but it does not matter as long as it forms an orthogonal basis!).

The qubits represented by arbitrary values γ (global gauge invariance according to $U(1)$ as seen in the section of Set Algebra page 1417) are all represented by the same point on the Bloch sphere because we will show that the factor has no observable effect and that we can then write without losing in generality:

$$|\Psi\rangle = \cos\left(\frac{\theta}{2}\right)|0\rangle + e^{\mathrm{i}\phi}\sin\left(\frac{\theta}{2}\right)|1\rangle \tag{14.7.47}$$

which is represented as we will justify later by the figure below:

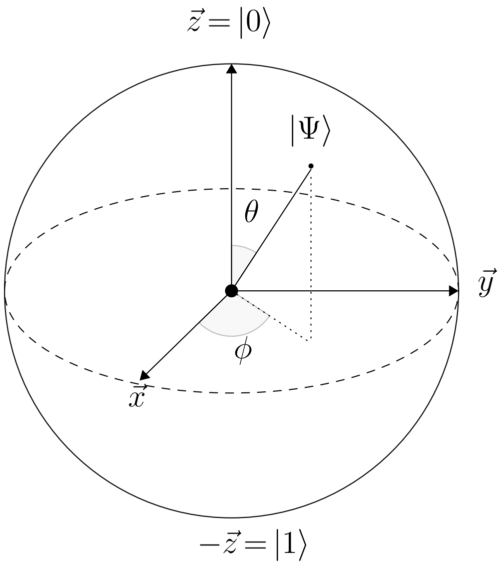


Figure 14.376: Bloch-shpere (two dimensional one)

Remark

Let us assume that we have a Quantum System with k qubits:

$$|0\rangle, |1\rangle, |2\rangle \ldots |k-1\rangle \tag{14.7.48}$$

The qubit will be in a superimposed state:

$$\Psi = \alpha_0|0\rangle + \alpha_1|1\rangle + \alpha_2|2\rangle + \ldots + \alpha_{\mathrm{k}-1}|k-1\rangle \tag{14.7.49}$$

where α_i is a complex vector with the property $\sum \alpha_i = 1$.

The Bloch sphere is a generalization of the representation of a complex number z (see section Numbers page 243) with $|z|^2 = 1$ as point of the circle in the (Gauss-) plane.

We also saw in this same section that a complex number could be represented by a complex exponential such that:

$$z = x + \mathrm{i}y = re^{\mathrm{i}\theta} \tag{14.7.50}$$

and if the circle were unitary:

$$z = e^{\mathrm{i}\theta} \tag{14.7.51}$$

Let us notice that the constraint $|z|^2 = 1$ eliminates a degree of freedom.

We will now notice the decomposition of a state of polarization in the form:

$$|\Psi\rangle = \alpha|x\rangle + \beta|y\rangle \tag{14.7.52}$$

And this one in a more traditional form in quantum computing (logic when we see the bases ...):

$$|\Psi\rangle = \alpha|0\rangle + \beta|1\rangle \tag{14.7.53}$$

where $\alpha, \beta \in \mathbb{C}$ (yes! indeed we are not anymore in the simple case of a wave linearly polarized now ...!) without forgetting the condition of normalization:

$$\langle\Psi|\Psi\rangle = \Psi\overline{\Psi} = |\alpha|^2 + |\beta|^2 = 1 \tag{14.7.54}$$

We can therefore write the qubit in the form:

$$|\Psi\rangle = r_\alpha e^{\mathrm{i}\phi_\alpha}|0\rangle + r_\beta e^{\mathrm{i}\phi_\beta}|1\rangle \tag{14.7.55}$$

Caution! It is very important to have read the part dealing with the polarization of light in the section of Wave Optics (see page 4538) to understand that it does not come from nowhere!!! With the difference that we do not work here with phasors because the solution of Schrödinger's equation of evolution involves complex exponentials as we saw in the context of the resolution of the latter for the eigen-mode of free particle.

Adding an overall phase factor should have no influence on the coefficients α, β because:

$$|e^{\mathrm{i}\gamma}\alpha|^2 = \left(e^{\mathrm{i}\gamma}\alpha\right)\overline{\left(e^{\mathrm{i}\gamma}\alpha\right)} = \left(e^{\mathrm{i}\gamma}\alpha\right)\left(e^{-\mathrm{i}\gamma}\bar{\alpha}\right) = \alpha\bar{\alpha} = |\alpha|^2 \tag{14.7.56}$$

and similarly for $|\beta|^2$. Thus, we are free to multiply our polarized and standardized qubit:

$$|\Psi\rangle = r_\alpha e^{\mathrm{i}\phi_\alpha}|0\rangle + r_\beta e^{\mathrm{i}\phi_\beta}|1\rangle \tag{14.7.57}$$

by the global phase $e^{-\mathrm{i}\phi_\alpha}$ which gives:

$$|\Psi'\rangle = r_\alpha e^{-\mathrm{i}\phi_\alpha} e^{i\phi_\alpha}|0\rangle + r_\beta e^{-\mathrm{i}\phi_\alpha} e^{\mathrm{i}\phi_\beta}|1\rangle = r_\alpha|0\rangle + r_\beta e^{\mathrm{i}\phi_\beta - \mathrm{i}\phi_\alpha}|1\rangle = r_\alpha|0\rangle + r_\beta e^{\mathrm{i}\phi}|1\rangle \tag{14.7.58}$$

In addition, we always have the condition of normalization $\langle\Psi'|\Psi\rangle$ to respect (impose).

Returning to the Cartesian coordinates, we have:

$$|\Psi'\rangle = r_\alpha|0\rangle + r_\beta e^{\mathrm{i}\phi}|1\rangle = r_\alpha|0\rangle + (x+\mathrm{i}y)\,|1\rangle \tag{14.7.59}$$

And the normalization constraint then gives:

$$\langle\Psi'|\Psi'\rangle = |r_\alpha|^2 + |x+iy|^2 = r_\alpha^2 + (x+\mathrm{i}y)\overline{(x+\mathrm{i}y)} = r_\alpha^2 + x^2 + y^2 = 1 \tag{14.7.60}$$

what is the equation of a unit sphere in $\mathbb{R}^2$ space with the Cartesian coordinates (x,y,r_α). Hence the quantum origin of the Bloch sphere!

We know (see section Vector Calculus page 1849) that the Cartesian coordinates are connected to the spherical coordinates by the relations:

$$x = r\sin(\theta)\cos(\phi) \qquad y = r\sin(\theta)\sin(\phi) \qquad z = r\cos(\theta) \tag{14.7.61}$$

therefore by putting $z := r_\alpha$ and remembering that $r = 1$, we can write:

$$|\Psi'\rangle = z|0\rangle + (x+\mathrm{i}y)|1\rangle = \cos(\theta)|0\rangle + \sin(\theta)\left(\cos(\phi)+\mathrm{i}\sin(\phi)\right)|1\rangle = \cos(\theta)|0\rangle + \sin(\theta)e^{\mathrm{i}\phi}|1\rangle \tag{14.7.62}$$

We now have only two useful parameters to know to define our point on the unit sphere (and always to the arbitrariness of phase).

The reader will notice that unlike the linearly polarized qubit, the general case above adds a complex term and that inversely by removing this additional term, we fall back on the relation of the linearly polarized wave seen at the beginning of this section. However, notice (for general culture) that if we add the same complex term to the first term, then we also have the very common following representation of the linearly polarized wave whose writing is named "Jones vector":

$$|\Psi'\rangle = \cos(\theta)e^{\mathrm{i}\phi}|0\rangle + \sin(\theta)e^{\mathrm{i}\phi}|1\rangle = \cos(\theta)e^{\mathrm{i}\phi}\begin{pmatrix}0\\1\end{pmatrix} + \sin(\theta)e^{\mathrm{i}\phi}\begin{pmatrix}1\\0\end{pmatrix} = \begin{pmatrix}\sin(\theta)\\ \cos(\theta)\end{pmatrix}e^{\mathrm{i}\phi} \tag{14.7.63}$$

But we have not finished yet!

Let us return to (we remove the apostrophe for the state of polarization):

$$|\Psi\rangle = \cos(\theta)|0\rangle + \sin(\theta)e^{\mathrm{i}\phi}|1\rangle \tag{14.7.64}$$

and notice that:

$$\theta = 0 \Rightarrow |\Psi\rangle = |0\rangle \tag{14.7.65}$$

and:

$$\theta = \frac{\pi}{2} \Rightarrow |\Psi\rangle = e^{\mathrm{i}\phi}|1\rangle \tag{14.7.66}$$

without forgetting that $\phi\in\mathbb{R}$ and that in this case the exponential factor in front of the $|1\rangle$ is a global phase change without influence.

All this suggests that $0\leq\theta\leq\pi/2$ is sufficient to describe any state of polarization and hence all points of the Bloch sphere.

On the other hand, we can see that in the system (r,θ,ϕ) the point of coordinates $(1,\pi-\theta,\phi+\pi)$ is the point opposite to that of coordinates $(1,\theta,\pi)$:

$$\begin{aligned}|\Psi\rangle&=\cos(\pi-\theta)|0\rangle+\sin(\pi-\theta)e^{\mathrm{i}(\phi+\pi)}|1\rangle\\&=-\cos(\theta)|0\rangle+\sin(\theta)e^{\mathrm{i}\pi}e^{\mathrm{i}\phi}|1\rangle\\&=-\cos(\theta)|0\rangle-\sin(\theta)e^{\mathrm{i}\phi}|1\rangle=-|\Psi\rangle=|\Psi'\rangle\end{aligned}\tag{14.7.67}$$

We also have (this is quite immediate but we can detail on request):

$$\langle\Psi|\Psi'\rangle=0\tag{14.7.68}$$

and therefore opposite points on the Bloch's sphere correspond orthogonal qubits (states of polarization)!

Thus we can consider only the upper hemisphere of the Bloch sphere since the opposite points differ only from a phase factor -1 and are therefore equivalent in the representation of the Bloch sphere.

Thus, the relation:

$$\boxed{|\Psi\rangle=\cos\left(\frac{\theta}{2}\right)|0\rangle+e^{\mathrm{i}\phi}\sin\left(\frac{\theta}{2}\right)|1\rangle}\tag{14.7.69}$$

is sufficient to describe the whole Bloch sphere in a complex space of dimension 2.

By construction, each point given by the previous relation of dimension 2 contains a double representation of a rotation in the real space of dimension 3.

We have also seen in the chapter of Spinor Calculus (see page 2145) that a rotation could be written in the form:

$$R\left(\vec{L},\frac{\theta}{2}\right)=\cos\left(\frac{\theta}{2}\right)\sigma_0-\mathrm{i}\sin\left(\frac{\theta}{2}\right)\vec{L}\circ\vec{\sigma}\tag{14.7.70}$$

With for recall the Pauli Matrices (see section Relativistic Quantum Physics page 4937):

$$\sigma_0=\begin{bmatrix}1&0\\0&1\end{bmatrix},\sigma_1=\begin{bmatrix}0&1\\1&0\end{bmatrix},\sigma_2=\begin{bmatrix}0&-\mathrm{i}\\\mathrm{i}&0\end{bmatrix},\sigma_3=\begin{bmatrix}1&0\\0&-1\end{bmatrix}\tag{14.7.71}$$

Either with the usual (traditional) writing of the field of Quantum Computing:

$$\begin{aligned}R_x(\theta)&=\cos\left(\frac{\theta}{2}\right)I-\mathrm{i}\sin\left(\frac{\theta}{2}\right)\sigma_1=\begin{bmatrix}\cos\left(\frac{\theta}{2}\right)&-\mathrm{i}\sin\left(\frac{\theta}{2}\right)\\-\sin\left(\frac{\theta}{2}\right)&\cos\left(\frac{\theta}{2}\right)\end{bmatrix}\\R_y(\theta)&=\cos\left(\frac{\theta}{2}\right)I-\mathrm{i}\sin\left(\frac{\theta}{2}\right)\sigma_2=\begin{bmatrix}\cos\left(\frac{\theta}{2}\right)&-\sin\left(\frac{\theta}{2}\right)\\\sin\left(\frac{\theta}{2}\right)&\cos\left(\frac{\theta}{2}\right)\end{bmatrix}\\R_z(\theta)&=\cos\left(\frac{\theta}{2}\right)I-\mathrm{i}\sin\left(\frac{\theta}{2}\right)\sigma_3=\begin{bmatrix}\cos\left(\frac{\theta}{2}\right)-\mathrm{i}\sin\left(\frac{\theta}{2}\right)&0\\0&\cos\left(\frac{\theta}{2}\right)+\mathrm{i}\sin\left(\frac{\theta}{2}\right)\end{bmatrix}\end{aligned}\tag{14.7.72}$$

Either after simplifying the last matrix:

$$\begin{aligned}R_x(\theta)&=\begin{bmatrix}\cos\left(\dfrac{\theta}{2}\right) & -\mathrm{i}\sin\left(\dfrac{\theta}{2}\right)\\ -\mathrm{i}\sin\left(\dfrac{\theta}{2}\right) & \cos\left(\dfrac{\theta}{2}\right)\end{bmatrix}\\ R_y(\theta)&=\begin{bmatrix}\cos\left(\dfrac{\theta}{2}\right) & -\sin\left(\dfrac{\theta}{2}\right)\\ \sin\left(\dfrac{\theta}{2}\right) & \cos\left(\dfrac{\theta}{2}\right)\end{bmatrix}\\ R_z(\theta)&=\begin{bmatrix}e^{-\mathrm{i}\theta/2} & 0\\ 0 & e^{\mathrm{i}\theta/2}\end{bmatrix}\end{aligned} \tag{14.7.73}$$

Now let us consider the rotation of our polarization state vector (due to a projector):

$$R_z(\alpha)|\Psi\rangle=\begin{bmatrix}e^{-\mathrm{i}\theta/2} & 0\\ 0 & e^{\mathrm{i}\theta/2}\end{bmatrix}\begin{pmatrix}\cos\left(\dfrac{\theta}{2}\right)\\ e^{\mathrm{i}\phi}\sin\left(\dfrac{\theta}{2}\right)\end{pmatrix}=\begin{pmatrix}e^{-\mathrm{i}\theta/2}\cos\left(\dfrac{\theta}{2}\right)\\ e^{\mathrm{i}\theta/2}e^{\mathrm{i}\phi}\sin\left(\dfrac{\theta}{2}\right)\end{pmatrix} \tag{14.7.74}$$

To get a coefficient of $|0\rangle$ that is real (in order to have an observable in the projection of an axis), we multiply by a phase factor $e^{\mathrm{i}\alpha/2}$ giving:

So to get a rotation around the z-axis it enough to change $\phi\rightarrow\phi+\alpha$.

So if we come back to:

$$R\left(\vec{L},\theta\right)=\cos\left(\dfrac{\theta}{2}\right)I-\mathrm{i}\sin\left(\dfrac{\theta}{2}\right)\vec{L}\circ\vec{\sigma} \tag{14.7.75}$$

it can be shown in the same way that in a general framework a unitary qubit operator can be written in the observable form:

$$U=e^{\mathrm{i}\alpha}R\left(\vec{L},\theta\right) \tag{14.7.76}$$

It is then necessary to choose the angles and the axis of rotation to define the operator completely.

14.7.3.1.1 Qubit of polarization

We will come back here on the case of polarization of the photon but this time we will be able to generalize thanks to the formulation of the Bloch's sphere to any type of polarization.

Let us consider for this a polariser that only passes photons polarized vertically followed by a photodetector, which do a "click" if a photon is detected and nothing else. This device allows us to detect vertically polarized photons.

Let us translate this in the language of quantum mechanics: The states of the system are therefore the states of polarization of a photon. The measurements of the observable will also have its states of polarization.

The possible measures are:

$$\begin{aligned}\lambda_0 &= 0 = \text{'click'}\\ \lambda_1 &= 1 = \text{'no click'}\end{aligned} \tag{14.7.77}$$

We will denote the corresponding states $|x\rangle$, $|y\rangle$. In our configuration, it is then obvious that the couple (λ_0, λ_1) represents the eigenvalues and $|x\rangle$, $|y\rangle$ the eigenvectors of an operator (which we do not know for know) and that we will therefore denote $\mathcal{P}$.

As we know, $|x\rangle$, $|y\rangle$ is an orthonormal basis of the space of states (of polarization). This is the base named "H/V base" for "Horizontal/Vertical" and is denoted normally:

$$|0\rangle, |1\rangle \tag{14.7.78}$$

Let us now take several cases:

1. Given a photon in the sate:

   $$|\Psi\rangle = |0\rangle \tag{14.7.79}$$

   then:

   $$\begin{aligned}\mathcal{P}(\lambda=\lambda_0) = \mathcal{P}(0) = \langle 0|0\rangle = 1\\ \mathcal{P}(\lambda=\lambda_1) = \mathcal{P}(1) = \langle 1|0\rangle = 0\end{aligned} \tag{14.7.80}$$

2. Let a photon be in the semi-vertical / horizontal state, that is to say oblique (which can be assimilated to the quantum superposition of its two polarizations):

   $$|\Psi\rangle = \frac{|0\rangle + |1\rangle}{|||0\rangle + |1\rangle||} = \frac{|0\rangle + |1\rangle}{\sqrt{2}} \tag{14.7.81}$$

   where the root is just there to ensure the normalization condition: $\langle\Psi|\Psi\rangle = 1$. Indeed:

   $$\langle\Psi|\Psi\rangle = |\overline{\Psi}\Psi|^2 = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix} \circ \frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix} = 1 \tag{14.7.82}$$

   Then since there is superposition (i.e. half of each in the total wave):

   $$\begin{aligned}\mathcal{P}(\lambda=\lambda_0) = \mathcal{P}(0) = \langle 0|\Psi\rangle = \frac{1}{2}\\ \mathcal{P}(\lambda=\lambda_1) = \mathcal{P}(1) = \langle 1|\Psi\rangle = \frac{1}{2}\end{aligned} \tag{14.7.83}$$

3. Let us now take any polarization (and this is what we did not have before!):

   $$|\Psi\rangle = \cos\left(\frac{\theta}{2}\right)|0\rangle + e^{\mathrm{i}\phi}\sin\left(\frac{\theta}{2}\right)|1\rangle \tag{14.7.84}$$

   which is well-normalized as we know it. So:

   $$\begin{aligned}\mathcal{P}(\lambda=\lambda_0) = \mathcal{P}(0) = \langle 0|\Psi\rangle = \cos^2\left(\frac{1}{2}\right)\\ \mathcal{P}(\lambda=\lambda_1) = \mathcal{P}(1) = \langle 1|\Psi\rangle = \sin^2\left(\frac{1}{2}\right)\end{aligned} \tag{14.7.85}$$

   The sum of the two probabilities giving indeed 1!

Now, let's imagine that we turn the polariser of $\pi/4$. We will denote the new basis of this polariser $|0'\rangle$, $|1'\rangle$ determined by a rotation of angle $\pi/4$ with by construction:

$$\begin{aligned}|0'\rangle&=\frac{|0\rangle+|1\rangle}{\sqrt{2}}\\|1'\rangle&=\frac{|0\rangle-|1\rangle}{\sqrt{2}}\end{aligned}\tag{14.7.86}$$

where the first base therefore corresponds to the diagonal polarization and the second is named "antidiagonal polarization". It is therefore the "D/A base" (Diagonal / Antidiagonal). In the form of the Jones vectors the last two relations are written:

$$\begin{aligned}|0'\rangle&=\frac{|0\rangle+|1\rangle}{\sqrt{2}}=\frac{\begin{pmatrix}0\\1\end{pmatrix}+\begin{pmatrix}1\\0\end{pmatrix}}{\sqrt{2}}=\frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix}\\|1'\rangle&=\frac{|0\rangle-|1\rangle}{\sqrt{2}}=\frac{\begin{pmatrix}0\\1\end{pmatrix}-\begin{pmatrix}1\\0\end{pmatrix}}{\sqrt{2}}=\frac{1}{\sqrt{2}}\begin{pmatrix}-1\\1\end{pmatrix}\end{aligned}\tag{14.7.87}$$

If we imagine that we have two polarisers that follow one another. The first having the D/A base and the second the H/V base, the first will prepare the photon polarized generally in a particular state (polarization) which will be by construction be the oblique state for the base H/V. Thus our second polariser will have only situations of the type:

$$\begin{aligned}\mathcal{P}(\lambda=\lambda_0)=\mathcal{P}(0)=\langle 0|\Psi\rangle=\frac{1}{2}\\\mathcal{P}(\lambda=\lambda_1)=\mathcal{P}(1)=\langle 1|\Psi\rangle=\frac{1}{2}\end{aligned}\tag{14.7.88}$$

Thus, this shows that any measure obviously disturbs the state of polarization of the photon and thus disrupts the state of the system. This last result is used in quantum cryptography!

We notice by the way that by using the spin operator introduced during our study of the section on Wave Quantum Physics (see page 4903) we have:

$$\hat{S}^2|0\rangle=\frac{3}{4}\hbar^2\begin{bmatrix}1&0\\0&1\end{bmatrix}\frac{\begin{pmatrix}0\\1\end{pmatrix}+\begin{pmatrix}1\\0\end{pmatrix}}{\sqrt{2}}=\frac{3}{4}\hbar^2\begin{bmatrix}1&0\\0&1\end{bmatrix}\frac{\begin{pmatrix}1\\1\end{pmatrix}}{\sqrt{2}}=\frac{3}{4}\hbar^2\frac{\begin{pmatrix}1\\1\end{pmatrix}}{\sqrt{2}}=\frac{3}{4}\hbar^2|0\rangle\tag{14.7.89}$$

which is of the same form as the following relation (relation linking eigenvector and eigenvalue) obtained in the in the section of Wave Quantum Physics:

$$\hat{S}^2\left|\frac{1}{2},\pm\frac{1}{2}\right\rangle=\frac{3}{4}\hbar^2\left|\frac{1}{2},\pm\frac{1}{2}\right\rangle\tag{14.7.90}$$

And that indicates indeed that the state:

$$|\Psi\rangle=\frac{|0\rangle+|1\rangle}{\sqrt{2}}\tag{14.7.91}$$

seems to be necessarily associated with a $1/2$ spin particle. We also notice that this state is an eigenvector of the operator S_x associated with the eigenvalue $\hbar/2$. Indeed:

$$\hat{S}_x|0\rangle=\frac{\hbar}{2}\begin{bmatrix}0&1\\1&0\end{bmatrix}\frac{\begin{pmatrix}0\\1\end{pmatrix}+\begin{pmatrix}1\\0\end{pmatrix}}{\sqrt{2}}=\frac{\hbar}{2}\begin{bmatrix}0&1\\1&0\end{bmatrix}\frac{\begin{pmatrix}1\\1\end{pmatrix}}{\sqrt{2}}=\frac{\hbar}{2}\frac{\begin{pmatrix}1\\1\end{pmatrix}}{\sqrt{2}}=\frac{\hbar}{2}|0\rangle\tag{14.7.92}$$

Since in the present case with the operator S_x we have immediately the eigenvalue and the associated eigenvector without making any intermediate calculations, it comes that the probability of measuring this eigenvalue (refer for recall to the 4th postulate of Wave Quantum Physics which associates operator to a measurement through the eigenvalue) is easy to calculate because the eigenvector is equal to the state vector in this particular case. Therefore:

$$\mathcal{P}\left(+\frac{1}{2}\hbar\right)=\langle\psi|0\rangle=\left\langle\frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix}\middle|\frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix}\right\rangle=1 \qquad (14.7.93)$$

14.7.3.1.2 $1/2$ spin Qubit

We will see here how to build a qubit based on a particle with a $1/2$ spin.

In the study of the Bloch sphere, we have examined a qubit at a given instant and we have seen that in a Hilbert space H, this qubit is described (by choice) by a unit vector:

$$|\Psi\rangle=\alpha|0\rangle+\beta|1\rangle \qquad |\alpha|^2+|\beta|^2=1 \qquad (14.7.94)$$

decomposed into an orthonormal basis $(|0\rangle,|1\rangle)$.

Let us consider the most general and minimum initial state corresponding to an arbitrary orientation of a spin:

$$|\Psi\rangle=\cos\left(\frac{\theta}{2}\right)|0\rangle+e^{\mathrm{i}\phi}\sin\left(\frac{\theta}{2}\right)|1\rangle \qquad (14.7.95)$$

Which corresponds, as we know, to two opposite (and superposed) states on the Bloch sphere.

Let us notice that we have indeed a normalized probability of the form:

$$\langle\Psi|\Psi\rangle=\cos^2\left(\frac{\theta}{2}\right)+\sin^2\left(\frac{\theta}{2}\right)=1 \qquad (14.7.96)$$

We have also seen that the projection along z by the operator $R_z(\alpha)$ of the state Ψ is given to the phase arbitrariness by:

$$e^{\mathrm{i}\alpha/2}R_z(\alpha)|\Psi\rangle=\begin{pmatrix}\cos\left(\dfrac{\theta}{2}\right)\\ e^{\mathrm{i}\phi}\sin\left(\dfrac{\theta}{2}\right)\end{pmatrix} \qquad (14.7.97)$$

Now let u recall our example:

$$\begin{aligned}M|x\rangle&=+1\cdot|x\rangle=+|x\rangle\\ M|y\rangle&=-1\cdot|y\rangle=-|y\rangle\end{aligned} \qquad (14.7.98)$$

where M was a Hermitian operator. Now the Pauli matrices are simple Hermitian operators. Moreover, as we have demonstrated in the section of Spinor Calculus, the Hermitian operator σ_z (assimilated to M) has by chance the same eigenvalues and eigenvectors corresponding to the two relations above. But it is then written in a traditional way as we saw it in the section of Spinor Calculus:

$$\begin{aligned}\sigma_z|+\rangle &= +|+\rangle\\ \sigma_z|-\rangle &= +|-\rangle\end{aligned} \tag{14.7.99}$$

or even more condensed:

$$\sigma_z|\pm\rangle = \pm|\pm\rangle \tag{14.7.100}$$

Moreover, this operator also satisfies the relation:

$$\sigma_z = M = \mathcal{P}_x - \mathcal{P}_y = \begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix} - \begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix} = \begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix} \tag{14.7.101}$$

And what is the physical property associated with σ_z? Well this is the spin (!) and we will come back to it a bit further below because that means we can use the spin $1/2$ as a qubit.

Now let us introduce the evolution of the system on this projection because it is not static (but this would not change this particular case to consider it static).

We have seen in the section of Wave Quantum Physics that this operation consisted in a simple case (as here) of introducing a phase term dependent on time of the type (see page 4770):

$$e^{-\frac{\mathrm{i}t}{\hbar}H} \tag{14.7.102}$$

Therefore:

$$\begin{aligned}|\Psi(\alpha,t)\rangle &= \begin{pmatrix} e^{-\frac{\mathrm{i}t}{\hbar}H}\cos\left(\dfrac{\theta}{2}\right)\\ e^{-\frac{\mathrm{i}t}{\hbar}H}e^{\mathrm{i}\phi}\sin\left(\dfrac{\theta}{2}\right)\end{pmatrix} = \begin{pmatrix} e^{-\frac{\mathrm{i}t}{\hbar}h\nu}\cos\left(\dfrac{\theta}{2}\right)\\ e^{-\frac{\mathrm{i}t}{\hbar}h\nu}e^{\mathrm{i}\phi}\sin\left(\dfrac{\theta}{2}\right)\end{pmatrix} = \begin{pmatrix} e^{-2\pi\mathrm{i}t\nu}\cos\left(\dfrac{\theta}{2}\right)\\ e^{-2\pi\mathrm{i}t\nu}e^{\mathrm{i}\phi}\sin\left(\dfrac{\theta}{2}\right)\end{pmatrix}\\ &= \begin{pmatrix} e^{-\omega\mathrm{i}t}\cos\left(\dfrac{\theta}{2}\right)\\ e^{-\omega\mathrm{i}t}e^{\mathrm{i}\phi}\sin\left(\dfrac{\theta}{2}\right)\end{pmatrix}\end{aligned} \tag{14.7.103}$$

What is written more soberly:

$$|\Psi(t)\rangle = e^{-\omega\mathrm{i}t}\begin{bmatrix}\cos\left(\dfrac{\theta}{2}\right)\\ e^{\mathrm{i}\phi}\sin\left(\dfrac{\theta}{2}\right)\end{bmatrix} \tag{14.7.104}$$

Let us recall that the state of a $1/2$ spin particle is two-dimensional and described by the state matrix (see section Relativistic Quantum Physics page 4985):

$$\mathcal{M} = S = \frac{1}{2}\hbar\sigma \tag{14.7.105}$$

with:

$$\begin{aligned}S_x&=\frac{\hbar}{2}\sigma_1=\frac{\hbar}{2}\begin{bmatrix}0&1\\1&0\end{bmatrix}\\S_y&=\frac{\hbar}{2}\sigma_2=\frac{\hbar}{2}\begin{bmatrix}0&-\mathrm{i}\\\mathrm{i}&0\end{bmatrix}\\S_z&=\frac{\hbar}{2}\sigma_3=\frac{\hbar}{2}\begin{bmatrix}1&0\\0&-1\end{bmatrix}\end{aligned}\tag{14.7.106}$$

If we want to calculate the expected mean of the observable (physical property) along each axis, then we use the 5th postulate (see page 4699) in the case of the x-axis:

$$\begin{aligned}&\langle S_x\rangle=\langle\Psi(t)|S_x|\Psi(t)\rangle=e^{+\mathrm{i}\omega t}\begin{pmatrix}\cos\left(\dfrac{\theta}{2}\right)&e^{-\mathrm{i}\phi}\sin\left(\dfrac{\theta}{2}\right)\end{pmatrix}\frac{\hbar}{2}\begin{bmatrix}0&1\\1&0\end{bmatrix}e^{-\mathrm{i}\omega t}\begin{pmatrix}\cos\left(\dfrac{\theta}{2}\right)\\e^{\mathrm{i}\phi}\sin\left(\dfrac{\theta}{2}\right)\end{pmatrix}\\&=\frac{\hbar}{2}\begin{pmatrix}\cos\left(\frac{\theta}{2}\right)&e^{-\mathrm{i}\phi}\sin\left(\frac{\theta}{2}\right)\end{pmatrix}\begin{bmatrix}0&1\\1&0\end{bmatrix}\begin{pmatrix}\cos\left(\dfrac{\theta}{2}\right)\\e^{\mathrm{i}\phi}\sin\left(\dfrac{\theta}{2}\right)\end{pmatrix}=\frac{\hbar}{2}\begin{pmatrix}\cos\left(\frac{\theta}{2}\right)&e^{-\mathrm{i}\phi}\sin\left(\frac{\theta}{2}\right)\end{pmatrix}\begin{pmatrix}e^{\mathrm{i}\phi}\sin\left(\dfrac{\theta}{2}\right)\\\cos\left(\dfrac{\theta}{2}\right)\end{pmatrix}\\&=\frac{\hbar}{2}\left(\cos\left(\frac{\theta}{2}\right)e^{\mathrm{i}\phi}\sin\left(\frac{\theta}{2}\right)+e^{-\mathrm{i}\phi}\sin\left(\frac{\theta}{2}\right)\cos\left(\frac{\theta}{2}\right)\right)=\frac{\hbar}{2}\sin\left(\frac{\theta}{2}\right)\cos\left(\frac{\theta}{2}\right)\left(e^{\mathrm{i}\phi}+e^{-\mathrm{i}\phi}\right)\\&=\frac{\hbar}{2}\sin\left(\frac{\theta}{2}\right)\cos\left(\frac{\theta}{2}\right)(\cos(\phi)+\mathrm{i}\sin(\phi)+\cos(\phi)-\mathrm{i}\sin(\phi))=\hbar\sin\left(\frac{\theta}{2}\right)\cos\left(\frac{\theta}{2}\right)\cos(\phi)\\&=\frac{\hbar}{2}\sin(\theta)\cos(\phi)\end{aligned}\tag{14.7.107}$$

Follows the y-axis:

$$\begin{aligned}&\langle S_y\rangle=\langle\Psi(t)|S_y|\Psi(t)\rangle=e^{+\mathrm{i}\omega t}\begin{pmatrix}\cos\left(\dfrac{\theta}{2}\right)&e^{-\mathrm{i}\phi}\sin\left(\dfrac{\theta}{2}\right)\end{pmatrix}\frac{\hbar}{2}\begin{bmatrix}0&-\mathrm{i}\\\mathrm{i}&0\end{bmatrix}e^{-\mathrm{i}\omega t}\begin{pmatrix}\cos\left(\dfrac{\theta}{2}\right)\\e^{\mathrm{i}\phi}\sin\left(\dfrac{\theta}{2}\right)\end{pmatrix}\\&=\frac{\hbar}{2}\begin{pmatrix}\cos\left(\dfrac{\theta}{2}\right)&e^{-\mathrm{i}\phi}\sin\left(\dfrac{\theta}{2}\right)\end{pmatrix}\begin{bmatrix}0&-\mathrm{i}\\\mathrm{i}&0\end{bmatrix}\begin{pmatrix}\cos\left(\dfrac{\theta}{2}\right)\\e^{\mathrm{i}\phi}\sin\left(\dfrac{\theta}{2}\right)\end{pmatrix}=\frac{\hbar}{2}\begin{pmatrix}\cos\left(\dfrac{\theta}{2}\right)&e^{-\mathrm{i}\phi}\sin\left(\dfrac{\theta}{2}\right)\end{pmatrix}\begin{pmatrix}-\mathrm{i}e^{\mathrm{i}\phi}\sin\left(\dfrac{\theta}{2}\right)\\\mathrm{i}\cos\left(\dfrac{\theta}{2}\right)\end{pmatrix}\\&=\frac{\hbar}{2}\left(-\mathrm{i}\cos\left(\frac{\theta}{2}\right)e^{\mathrm{i}\phi}\sin\left(\frac{\theta}{2}\right)+\mathrm{i}e^{-\mathrm{i}\phi}\sin\left(\frac{\theta}{2}\right)\cos\left(\frac{\theta}{2}\right)\right)=\frac{\hbar}{2}\sin\left(\frac{\theta}{2}\right)\cos\left(\frac{\theta}{2}\right)\left(-\mathrm{i}e^{\mathrm{i}\phi}+\mathrm{i}e^{-\mathrm{i}\phi}\right)\\&=\frac{\hbar}{2}\sin\left(\frac{\theta}{2}\right)\cos\left(\frac{\theta}{2}\right)(-\mathrm{i}\cos(\phi)+\sin(\phi)+\mathrm{i}\cos(\phi)+\sin(\phi))=\hbar\sin\left(\frac{\theta}{2}\right)\cos\left(\frac{\theta}{2}\right)\sin(\phi)\\&=\frac{\hbar}{2}\sin(\theta)\sin(\phi)\end{aligned}\tag{14.7.108}$$

and finally following the z-axis:

$$
\begin{aligned}
&\langle S_z\rangle=\langle\Psi(t)|S_z|\Psi(t)\rangle=e^{+\mathrm{i}\omega t}\begin{pmatrix}\cos\left(\dfrac{\theta}{2}\right) & e^{-\mathrm{i}\phi}\left(\dfrac{\theta}{2}\right)\end{pmatrix}\frac{\hbar}{2}\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}e^{-\mathrm{i}\omega t}\begin{pmatrix}\cos\left(\dfrac{\theta}{2}\right)\\ e^{\mathrm{i}\phi}\sin\left(\dfrac{\theta}{2}\right)\end{pmatrix}\\
&=\frac{\hbar}{2}\begin{pmatrix}\cos\left(\dfrac{\theta}{2}\right) & e^{-\mathrm{i}\phi}\sin\left(\dfrac{\theta}{2}\right)\end{pmatrix}\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}\begin{pmatrix}\cos\left(\dfrac{\theta}{2}\right)\\ e^{\mathrm{i}\phi}\sin\left(\dfrac{\theta}{2}\right)\end{pmatrix}=\frac{\hbar}{2}\begin{pmatrix}\cos\left(\dfrac{\theta}{2}\right) & e^{-\mathrm{i}\phi}\sin\left(\dfrac{\theta}{2}\right)\end{pmatrix}\begin{pmatrix}\cos\left(\dfrac{\theta}{2}\right)\\ -e^{\mathrm{i}\phi}\sin\left(\dfrac{\theta}{2}\right)\end{pmatrix}\\
&=\frac{\hbar}{2}\left(\cos^2\left(\frac{\theta}{2}\right)-\sin^2\left(\frac{\theta}{2}\right)\right)=\frac{\hbar}{2}\cos(\theta)
\end{aligned}
\tag{14.7.109}
$$

For thus fall back well for the component z (because it is the only one that interests us here), the result which was imposed above in the form:

$$\langle M\rangle_\theta=\langle\theta|\mathcal{P}_x\theta\rangle-\langle\theta|\mathcal{P}_y\theta\rangle=\cos^2(\theta)-\sin^2(\theta)=\cos(2\theta)\tag{14.7.110}$$

with an angle difference which is a matter of substitution and an amplitude which makes it possible to match the particularity of the configuration of the system. We therefore have mathematical relations similar in all respects to the manipulation of qubits of oriented spins or qubits of polarized photons.

The question we can ask ourselves about the spin is how to prepare it in the state $|\Psi\rangle$ In fact, we can do this with a magnetic field by copying the Stern-Gerlach experiment, which allows us to separate a beam of spin $1/2$ particles into two distinct beam.

Since we now know all the eigenvalues and eigenvectors of the spin operator S, we can then determine the general form of the spin operator in any orientation (!):

$$
\begin{aligned}
S_n&=S_x\sin(\theta)\cos(\phi)-S_y\sin(\theta)\sin(\phi)+S_z\cos(\theta)\\
&=\frac{\hbar}{2}\left(\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}\sin(\theta)\cos(\phi)+\begin{bmatrix}0 & -\mathrm{i}\\ \mathrm{i} & 0\end{bmatrix}\sin(\theta)\cos(\phi)+\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}\cos(\theta)\right)\\
&=\frac{\hbar}{2}\begin{bmatrix}\cos(\theta) & \sin(\theta)\cos(\phi)-\mathrm{i}\sin(\theta)\sin(\phi)\\ \sin(\theta)\cos(\phi)+\mathrm{i}\sin(\theta)\sin(\phi) & -\cos(\theta)\end{bmatrix}\\
&=\frac{\hbar}{2}\begin{bmatrix}\cos(\theta) & \sin(\theta)(\cos(\phi)-\mathrm{i}\sin(\phi))\\ \sin(\theta)(\cos(\phi)+\mathrm{i}\sin(\phi)) & -\cos(\theta)\end{bmatrix}=\frac{\hbar}{2}\begin{bmatrix}\cos(\theta) & e^{-\mathrm{i}\phi}\sin(\theta)\\ e^{\mathrm{i}\phi}\sin(\theta) & -\cos(\theta)\end{bmatrix}
\end{aligned}
\tag{14.7.111}
$$

We have also:

$$\begin{aligned}S_n|\Psi\rangle &= S_n\left(\cos\left(\frac{\theta}{2}\right)|0\rangle+e^{\mathrm{i}\phi}\sin\left(\frac{\theta}{2}\right)|1\rangle\right)=\frac{\hbar}{2}\begin{bmatrix}\cos(\theta) & e^{-\mathrm{i}\phi}\sin(\theta)\\ e^{\mathrm{i}\phi}\sin(\theta) & -\cos(\theta)\end{bmatrix}\begin{pmatrix}\cos\left(\dfrac{\theta}{2}\right)\\ e^{\mathrm{i}\phi}\sin\left(\dfrac{\theta}{2}\right)\end{pmatrix}\\ &=\frac{\hbar}{2}\begin{pmatrix}\cos(\theta)\cos\left(\dfrac{\theta}{2}\right)+\sin(\theta)\sin\left(\dfrac{\theta}{2}\right)\\ e^{i\phi}\sin(\theta)\cos\left(\dfrac{\theta}{2}\right)-\cos(\theta)e^{\mathrm{i}\phi}\sin\left(\dfrac{\theta}{2}\right)\end{pmatrix}=\frac{\hbar}{2}\begin{pmatrix}\cos\left(\dfrac{\theta}{2}\right)\\ e^{\mathrm{i}\phi}\left(\sin\theta\cos\left(\dfrac{\theta}{2}\right)-\cos(\theta)\sin\left(\dfrac{\theta}{2}\right)\right)\end{pmatrix}\\ &=\frac{\hbar}{2}\begin{pmatrix}\cos\left(\dfrac{\theta}{2}\right)\\ e^{i\phi}\left(\sin(\theta)\cos\left(\dfrac{\theta}{2}\right)-\cos(\theta)\sin\left(\dfrac{\theta}{2}\right)\right)\end{pmatrix}=\frac{\hbar}{2}\begin{pmatrix}\cos\left(\dfrac{\theta}{2}\right)\\ e^{\mathrm{i}\phi}\sin\left(\dfrac{\theta}{2}\right)\end{pmatrix}=\frac{\hbar}{2}|\Psi\rangle\end{aligned} \tag{14.7.112}$$

We have the chosen $|\Psi\rangle$ which is, in addition to being a state vector, an eigenvector of S_n with the eigenvalue $\hbar/2$. Therefore the probability of measuring this particular state vector is equal to 1.

Now we know by using the 5th postulate that the probability of finding the eigenvalue $\hbar/2$(of the operator S_n), in a measurement of the property S along the z-axis performed at time t on the quantum system prepared in the state Ψ, is given by the square of the module of the projection of the function or state vector Ψ on the vector or eigenvector φ associated with the eigenvalue $\hbar/2$ (and its operator along this axis).

But the operator according to z is:

$$\sigma_z=\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix} \tag{14.7.113}$$

For the same operator, we have seen in the section of Spinor Calculus that it had as eigenvectors (see page 2145):

$$|+\rangle=\begin{pmatrix}1\\0\end{pmatrix},|-\rangle=\begin{pmatrix}0\\1\end{pmatrix} \tag{14.7.114}$$

Let us take the first eigenvector oriented according to z-axis on the Bloch sphere. We have then:

$$P=|\langle+|\Psi\rangle|^2=\left|\begin{pmatrix}1 & 0\end{pmatrix}\begin{pmatrix}\cos\left(\dfrac{\theta}{2}\right)\\ e^{i\phi}\sin\left(\dfrac{\theta}{2}\right)\end{pmatrix}\right|^2=\left|\cos\left(\frac{\theta}{2}\right)\right|^2=\cos^2\left(\frac{\theta}{2}\right) \tag{14.7.115}$$

And the probability according to the other eigenvector would give the same expression but with a sinus. The sum of the two probabilities would then give us well 1!

We thus notice that the relations are very similar between the photon and the spin in our case of study. This is normal since both are two-level systems, resulting in similar results.

14.7.4 Entangled qubit system

A 2 qubit system is a superposition of all possible 2 qubit states. A 2 qubit system and can be represented as:

$$\Psi=\alpha_{00}|00\rangle+\alpha_{01}|01\rangle+\alpha_{10}|10\rangle+\alpha_{11}|11\rangle \tag{14.7.116}$$

Measuring the 2 qubit system, as earlier, results in the collapse of the superposition and the result is one of the 4 qubit states. The probability of the measure state is the square of the amplitude $|\alpha_{ij}|^2$.

More specifically a 2 qubit system in which we have:

$$\Psi=\alpha_0|0\rangle+\alpha_1|1\rangle \qquad \Phi=\beta_0|0\rangle+\beta_1|1\rangle \tag{14.7.117}$$

the tensor product of the 2 qubits is:

$$\begin{aligned}\xi=\Psi\otimes\Phi&=(\alpha_0|0\rangle+\alpha_1|1\rangle)\otimes(\beta_0|0\rangle+\beta_1|1\rangle)\\&=\alpha_0\beta_0|0\rangle|0\rangle+\alpha_0\beta_1|0\rangle|1\rangle+\alpha_1\beta_0|1\rangle|0\rangle+\alpha_1\beta_1|1\rangle|1\rangle\\&=\alpha_0\beta_0|00\rangle+\alpha_0\beta_1|01\rangle+\alpha_1\beta_0|01\rangle+\alpha_1\beta_1|11\rangle\end{aligned} \tag{14.7.118}$$

where (notice that we have still a set of orthonormal vectors):

$$\begin{aligned}|00\rangle&=\begin{bmatrix}1\\0\\0\\0\end{bmatrix} \qquad |01\rangle=\begin{bmatrix}0\\1\\0\\0\end{bmatrix}\\|10\rangle&=\begin{bmatrix}0\\0\\1\\0\end{bmatrix} \qquad |11\rangle=\begin{bmatrix}0\\0\\0\\1\end{bmatrix}\end{aligned} \tag{14.7.119}$$

Notice that the above reasoning is sometimes stated as one of the postulates of quantum physics as following:

The state space of a composite physical system is the tensor product of the state spaces of the component physical systems. Moreover, if we have systems numbered 1 through n, and system number i is prepared in the state $|\psi_i\rangle$, then the joint state of the total system is $|\psi_1\rangle\otimes|\psi_2\rangle\otimes\cdots\otimes|\psi_n\rangle$.

14.7.5 Quantum logic gates

In quantum computing and specifically the quantum circuit model of computation, a quantum gate (or quantum logic gate) is a basic quantum circuit operating on a small number of qubits. They are the building blocks of quantum circuits, like classical logic gates are for conventional digital circuits.

Let us recall that we have stated above that the construction of a quantum qubit was done on the vector basis:

$$|0\rangle=|\downarrow\rangle=|-\rangle=\begin{pmatrix}0\\1\end{pmatrix} \qquad |1\rangle=|\uparrow\rangle=|+\rangle=\begin{pmatrix}1\\0\end{pmatrix} \tag{14.7.120}$$

or in extenso, by linearity (and this is all elementary trick!), any transformation that acts on these basic vectors, will therefore act on all qubit of the complex plane.

Let us consider the particular case for this introduction the logic gate which modifies the state of the 1-qubit, that is to say qubit which are collinear to one of the basis vectors, into the opposite state. That is to say the "quantum inverter":

$$X : |0\rangle \rightarrow |1\rangle, |1\rangle \rightarrow |0\rangle \tag{14.7.121}$$

We quickly guess that the matrix that satisfies this relation (caution! the reader will notice that it is not a special case of the rotation matrix in the plane as seen in the section of Euclidean Geometry!):

$$X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{14.7.122}$$

And which is sometimes written by specialists:

$$X = |1\rangle\langle 0| + |0\rangle\langle 1| = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \end{pmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{14.7.123}$$

And we see that this logical gate of negation of the 1-qubit is nothing else than the first matrix of Pauli. So we can also write:

$$\boxed{X = |1\rangle\langle 0| + |0\rangle\langle 1| = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \sigma_1 = \sigma_x} \tag{14.7.124}$$

Negation which is sometimes named "X-Pauli quantum gate" or "NOT quantum gate" and represented by the following symbol in "quantum circuits":

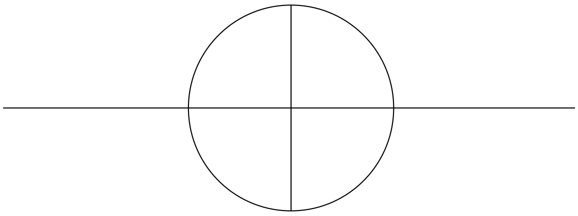


Figure 14.377: NOT Quantum gate

Now let us look for the logic gate making the following orthogonal clockwise transformation:

$$Y : |0\rangle \rightarrow -|1\rangle, |1\rangle \rightarrow |0\rangle \tag{14.7.125}$$

We guess quite quick that the matrix which satisfies this relation (caution! the reader will notice that this time is a special case of the rotation matrix in the plane as seen in the section of Euclidean Geometry!):

$$Y = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \tag{14.7.126}$$

And which is sometimes written by specialists (in a somewhat unfortunate way):

$$Y = |0\rangle\langle 1| - |1\rangle\langle 0| = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \tag{14.7.127}$$

And we see that this logical transformation gate in the clockwise direction of the 1-qubit involves the second Pauli matrix. So we can also write:

$$\boxed{Y=|0\rangle\langle 1|-|1\rangle\langle 0|=\begin{bmatrix}0 & -1\\ 1 & 0\end{bmatrix}=\mathrm{i}\sigma_2=\mathrm{i}\sigma_y} \tag{14.7.128}$$

Negation which is sometimes named "Y-Pauli quantum gate" and which has no classical equivalent.

Now let us seek for the logical gate making the following strictly counter-clockwise orthogonal transformation:

$$Z:|0\rangle\rightarrow|0\rangle,|1\rangle\rightarrow-|1\rangle \tag{14.7.129}$$

We guess quite quick that the matrix that satisfies this relation (caution! the reader will notice that it is not a special case of the rotation matrix in the plane seen in the section of Euclidean Geometry!):

$$Z=\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix} \tag{14.7.130}$$

And which is sometimes written by specialists (in a somewhat unfortunate way):

$$Z=|0\rangle\langle 0|-|1\rangle\langle 1|=\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix} \tag{14.7.131}$$

And we see that this logical transformation gate in the clockwise direction of 1-qubit involves the second Pauli matrix. So we can also write:

$$\boxed{Z=|0\rangle\langle 0|-|1\rangle\langle 1|=\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}=\sigma_z} \tag{14.7.132}$$

Negation then we sometimes name "Z-Pauli quantum gate".

When describing a quantum gate on an individual qubit, any dynamical operation, G, is a member of the unitary group $\mathrm{U}(2)$ (see section Set Algebra page 1417), which consists of all 2×2 matrices where $G^{\dagger}=G^{-1}$. Up to a global (and unphysical) phase factor, any single qubit operation can be expressed as a linear combination of the generator of $\mathrm{SU}(2)$ that are the Pauli matrices!

Now let us return to the following transformation which we have deal with earlier:

$$\begin{aligned}|0'\rangle&=\frac{|0\rangle+|1\rangle}{\sqrt{2}}=\frac{\begin{pmatrix}0\\1\end{pmatrix}+\begin{pmatrix}1\\0\end{pmatrix}}{\sqrt{2}}=\frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix}\\ |1'\rangle&=\frac{|0\rangle-|1\rangle}{\sqrt{2}}=\frac{\begin{pmatrix}0\\1\end{pmatrix}-\begin{pmatrix}1\\0\end{pmatrix}}{\sqrt{2}}=\frac{1}{\sqrt{2}}\begin{pmatrix}-1\\1\end{pmatrix}\end{aligned} \tag{14.7.133}$$

In other words, it is the transformation (diagonal for recall) that makes to qubit equiprobable[90]:

$$H:|0\rangle\rightarrow|0'\rangle=\frac{|0\rangle+|1\rangle}{\sqrt{2}}\qquad |1\rangle\rightarrow|1'\rangle=\frac{|0\rangle-|1\rangle}{\sqrt{2}} \tag{14.7.134}$$

---

[90] As the sum of square of the amplitudes is indeed equal to: $(1/\sqrt{2})^2+(1/\sqrt{2})^2=0.5+0.5=1$

We quickly notice that the corresponding matrix is then:

$$H=\frac{1}{\sqrt{2}}\begin{bmatrix}-1 & 1\\ 1 & 1\end{bmatrix} \tag{14.7.135}$$

which is named "Hadamard quantum gate" and thus corresponds to a counter-clockwise rotation of $\pi/4$. This gate is represented by the following symbol in quantum circuits:

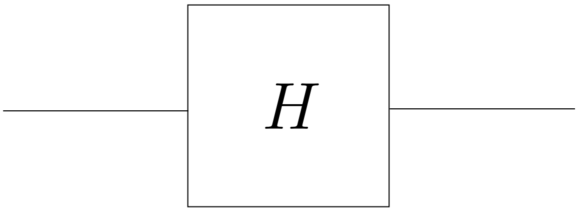


Figure 14.378: **Hadamard Quantum gate**

And we can continue like this for a long time to create empirical quantum logic gates ... so we will stop here.

Now remember that as we have seen in the section of Boolean logic (see page 7114), with a NAND gate (or an AND gate and an INVERTER) we've got all the parts we need to build a modern computer!

We have already seen the quantum inverter earlier above and its corresponding matrix. We need now and AND gate, however that latter has two input, i.e. entangles quantum bits.

We built a "quantum AND gate" as following:

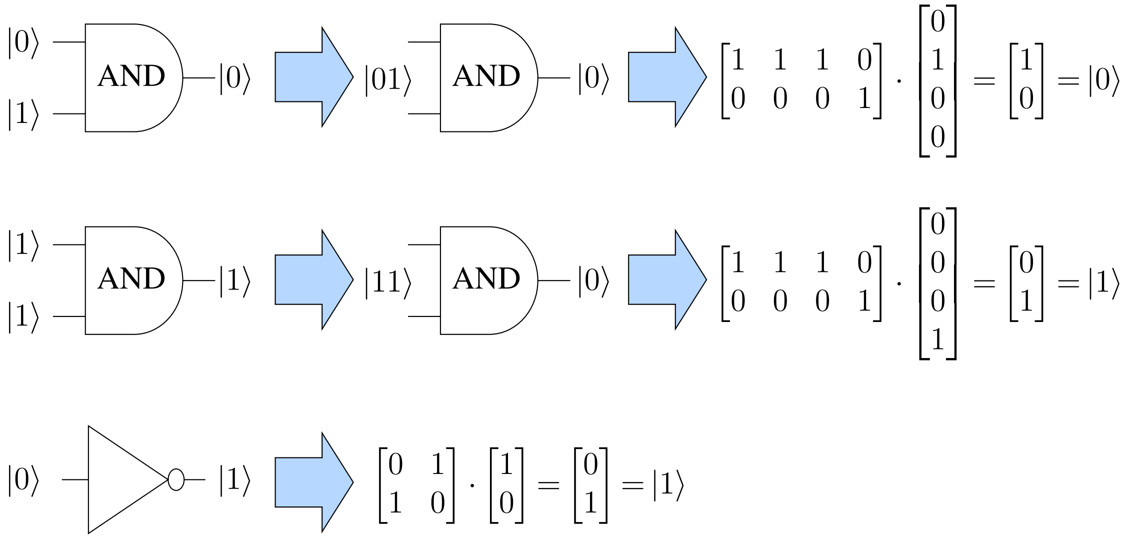

|0⟩
|1⟩
AND
|0⟩
|01⟩
AND
|0⟩
$\begin{bmatrix}1&1&1&0\\0&0&0&1\end{bmatrix}\cdot\begin{bmatrix}0\\1\\0\\0\end{bmatrix}=\begin{bmatrix}1\\0\end{bmatrix}=|0\rangle$
|1⟩
|1⟩
AND
|1⟩
|11⟩
AND
|0⟩
$\begin{bmatrix}1&1&1&0\\0&0&0&1\end{bmatrix}\cdot\begin{bmatrix}0\\0\\0\\1\end{bmatrix}=\begin{bmatrix}0\\1\end{bmatrix}=|1\rangle$
|0⟩
|1⟩
$\begin{bmatrix}0&1\\1&0\end{bmatrix}\cdot\begin{bmatrix}1\\0\end{bmatrix}=\begin{bmatrix}0\\1\end{bmatrix}=|1\rangle$

Figure 14.379: **NAND Quantum Gate**

~60 %

★★★★★

22 votes, 53.64%

15

Social Sciences

Social mathematics are the analysis and formal modeling tools of the behaviour and management of a population and its trade in goods in all its activities, in interaction with its environment. (Sciences.ch)

Contents

15.1 Population Dynamics . . . 7297
- 15.1.1 Birth rate and mortality tables (biometric features) . . . 7298
  - 15.1.1.1 Population Renewal . . . 7308
- 15.1.2 Population Models . . . 7310
  - 15.1.2.1 Exponential disease model . . . 7310
  - 15.1.2.2 Deterministic Logistic Model (Verlhust) . . . 7313
  - 15.1.2.3 Chaotic Logistic Model . . . 7317
  - 15.1.2.4 Malthusian Growth Law . . . 7332
  - 15.1.2.5 Leslie model . . . 7333
  - 15.1.2.6 Epidemic models for spread of disease . . . 7335
  - 15.1.2.7 Lotka–Volterra predator–prey model . . . 7350
- 15.1.3 Schaefer's optimal capture model . . . 7359
- 15.1.4 Hardy-Weinberg model . . . 7361
- 15.1.5 Mendel's law . . . 7371
- 15.1.6 Growth rate with temperature . . . 7375

15.2 Game and Decision Theory . . . 7376
- 15.2.1 Behavioural decision bias (cognitive bias) . . . 7381
  - 15.2.1.1 Sunk Cost . . . 7386
  - 15.2.1.2 Anchoring Bias . . . 7388
  - 15.2.1.3 Wisdom of Crowds . . . 7389
  - 15.2.1.4 Delphi Method . . . 7392

15.2.2 Utility 7394
15.2.2.1 Pareto Optimum 7394
15.2.2.2 Nash Equilibrium 7395
15.2.3 Games Representations 7398
15.2.3.1 Extensive representation of a Game 7399
15.2.3.2 Extensive representation of a Decision 7400
15.2.3.3 Normal representation of a Game 7410
15.2.3.4 Set representation of a Game 7420
15.2.3.5 Graphical representation of a Game 7424
15.2.4 Expected Utility 7428
15.2.4.1 Hurwitz Criterion 7429
15.2.4.2 Laplace Criterion 7436
15.2.4.3 Kelly criterion 7437
15.2.5 Evolutionary Games 7444
15.2.5.1 Dove & Hawk game in pure strategy (without probabilities) 7448
15.2.5.2 Dove & Hawk game in stable evolutionary strategy (with probabilities) 7449
15.2.6 Cournot Competition 7452
15.2.7 Markov Decision Processes (MDPs) 7456
15.2.8 Multi-Criteria Decision Making (MCDM) 7462
15.2.8.1 Analytic Hierarchy Process (AHP) 7462
15.2.9 Electoral systems 7473
15.2.10 Ethics and Moral 7478
15.2.10.1 Golden Rule of morality (Tit for Tat) 7479
15.3 Economy 7485
15.3.1 Concepts 7486
15.3.1.1 Microeconomics 7488
15.3.1.2 Macroeconomics 7508
15.3.2 Monetary Model 7515
15.3.2.1 Monetary base conservation 7517
15.3.2.2 Walras' law 7520
15.3.3 Price Index and GDP 7525
15.3.3.1 Paasche and Laspeyres price indices 7526
15.3.3.2 Fisher index and Marshall–Edgeworth index 7527
15.3.3.3 Gross domestic product (GDP) 7528
15.3.4 Supply and Demand Theory 7530
15.3.4.1 Expected utility theory 7530
15.3.5 Net gain/loss opposite feedback model 7538

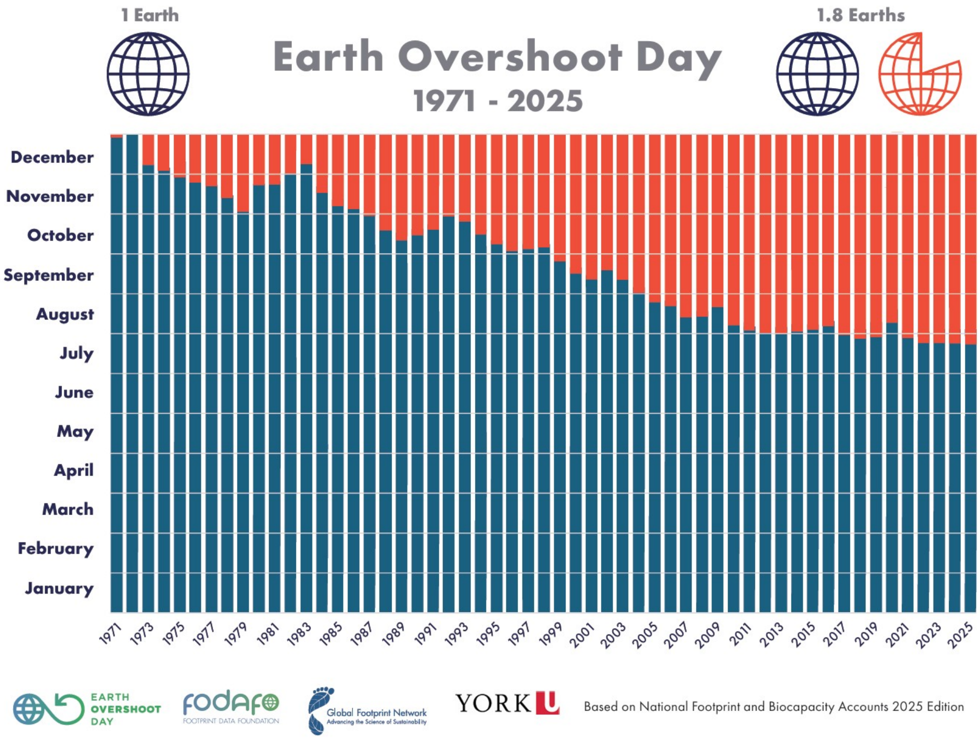
15.3.6 Capitalization and Actuarial . . . 7547
15.3.6.1 Dates Intervals . . . 7549
15.3.6.2 Rates Equivalence . . . 7552
15.3.6.3 Simple Interest . . . 7553
15.3.6.4 Compound Interest . . . 7557
15.3.6.5 Continuous Interest . . . 7560
15.3.7 Progressive interest (annuities) . . . 7562
15.3.7.1 Postnumerando annuities . . . 7563
15.3.7.2 Praenumerando annuities . . . 7565
15.3.8 Rounding . . . 7567
15.3.9 Loans Amortization/Repayments . . . 7570
15.3.9.1 Fixed-Term Loan . . . 7572
15.3.9.2 Loan with constant amortization . . . 7573
15.3.9.3 Loan with constant annuity . . . 7573
15.3.10 Modern Portfolio Theory . . . 7581
15.3.10.1 No Arbitrage Opportunity (N.A.O.) . . . 7585
15.3.10.2 Liquidity and Illiquidity . . . 7590
15.3.10.3 Portfolios . . . 7599
15.3.11 Exchange rate . . . 7880
15.3.12 Credit default Risk . . . 7884
15.3.12.1 Credit spread risk . . . 7885
15.3.12.2 Value At Risk equity coverage rating . . . 7887
15.3.12.3 Altman Z-score . . . 7889
15.3.12.4 Merton's Distance to Default and KMV Expected Default Frequency . . . 7890
15.3.13 Time Series Analysis . . . 7896
15.3.13.1 Type of Errors . . . 7903
15.3.13.2 Decompositions . . . 7907
15.3.13.3 Types of Forecasting Models . . . 7914
15.3.13.4 Autoregressive Models . . . 7972
15.3.13.5 Forecasting Stationary Processes . . . 8005
15.3.13.6 TBATS . . . 8012
15.3.13.7 Recipes for modelling time series data . . . 8013
15.4 Quantitative Management . . . 8021
15.4.1 Corporate and Government Finance Management . . . 8024
15.4.1.1 Basic Accounting Equation . . . 8026
15.4.1.2 Ratio Analysis . . . 8027
15.4.1.3 Weighted Average Cost of Capital (WACC) . . . 8044

15.4.1.4 Break-even Point Analysis (BEPA) . . . 8047
15.4.1.5 Investment Strategies . . . 8049
15.4.1.6 Company Valuation Methods . . . 8060
15.4.1.7 Capital Goods . . . 8070
15.4.2 Project Management . . . 8074
15.4.2.1 Probabilistic PERT (Beta-PERT Distribution) . . . 8074
15.4.2.2 Project planning variance reduction . . . 8083
15.4.2.3 Process Reliability . . . 8086
15.4.2.4 Divide and rule? . . . 8088
15.4.3 Lean Management (Six Sigma Process) . . . 8092
15.4.3.1 Pareto Analysis . . . 8095
15.4.3.2 Weighted Ishikawa Diagram . . . 8102
15.4.4 Supply Chain Management . . . 8106
15.4.4.1 Supply Chain Management in uncertain future . . . 8108
15.4.4.2 Optimal initial stock management with zero rotation . . . 8112
15.4.4.3 Wilson's Models . . . 8119
15.4.5 Queueing Theory . . . 8141
15.4.5.1 $M/M/\ldots$ arrival times modelisation . . . 8145
15.4.5.2 $M/M/\ldots$ service times modelisation . . . 8150
15.4.5.3 Kendall queues notation . . . 8152
15.4.5.4 Modelling of arrivals and departures $M/M/1$ . . . 8154
15.4.5.5 Probability of standby in a $M/M/k/k$ queue (Erlang-B formula) . . . 8162
15.4.5.6 Probability $M/M/K/+\infty$ of standby (Erlang-C formula) . . . 8165
15.4.6 Insurance . . . 8170
15.4.6.1 Premium pricing . . . 8172
15.4.6.2 Taking account of the return on experience . . . 8175
15.4.6.3 Discounting factor of a retirement insurance . . . 8177
15.4.6.4 Pension insurance (life annuities) . . . 8180
15.4.7 Sensitivity Analysis . . . 8184
15.4.7.1 Direct Bias Method . . . 8186
15.4.7.2 Correlation Method . . . 8190
15.5 Music Mathematics (physics of hearing) . . . 8195
15.5.1 Longitudinal Sound Waves . . . 8195
15.5.1.1 Power carried by a sound wave . . . 8203
15.5.1.2 Measuring the intensity of a sound . . . 8207
15.5.2 Spherical Sound Waves . . . 8210
15.5.3 Doppler effect . . . 8212

15.5.3.1 Fixed source-Moving observer 8214
15.5.3.2 Moving source-Fixed observer 8215
15.5.3.3 Moving source and observer 8217
15.5.4 Shock waves 8218
15.5.5 Music Scales 8220
15.5.6 Harmonic Oscillator 8224
15.5.6.1 Damping and Resonance 8227

15.1. Population Dynamics

It is said that Albert Einstein at the end of his life, used to evoke the three explosions that would soon threaten humanity: the explosion of nuclear bombs, the explosion of our knowledge (robotics), the explosion in the number of humans on Earth. The role of the scientist at this level is not only to improve their knowledge but to share and especially to disseminate the awareness they have gained on the consequences they can discern of the explosion of the number of humans on Earth. Einstein was keen to play this role. Beyond the immediate vicissitudes, he knew to see the long-term developments. Among these developments, the number of humans appeared to him clearly as that of the greatest dangers (already introduced by the initiators of dynamic population science in the early 118th century according to holocene calendar!!!). Today, at the beginning of the 121st century (holocene calendar), we can see their insight. Humanity is actually taken by the throat by the increase of its actual effective without forcing it down (because politicians favour the infinite growth and participation to social helps which do impose an endless increase of the population by mathematical construction!) as human do with the regulation of the number of animal in some protected areas. The major (naive) problems for politicians being to ensure full employment (avoid "technological unemployment" at some level) and the fact that the increase in population naturally enriches the owners and impoverishes workers (among many other problems...).



Based on National Footprint and Biocapacity Accounts 2025 Edition

Figure 15.1: Earth overshoot day evolution

The chart above, that sadly doesn't have confidence intervals, shows that if we assume that the amount of ecological resources of Earth must regenerate in the same year, the human population of 12025 (holocene calendar) should be of approximately 5.4 billions individuals (https://overshoot.footprintnetwork.org) to be sustainable! However if each individual consumes like the average occidental style of life as in the early 121st (holocene calendar) century, under the same mentioned assumption, Earth could sustained only (under other assumptions, nobody really knows what the planet's human carrying capacity is!) around 2 billion individuals (source: Australian Academy of Science, peer-reviewed by Professor Stephen Dovers Fenner School of Environment and Society, Australian National University; Professor Colin Butler Public Health, University of Canberra).

Until the politicians of all major countries understand and show the example by limiting the population of their respective country (and therefore the number of births) and by proposing the same constraints for the whole world and the mentality of the population is open to type of proposal (it is it that elects politicians...), the mathematical theoretical study of the subject can help open some perspectives... regarding the renewal and the cycle of the resources of our unique planet... (because we only have one... and for the moment there is almost not good plan B).

To introduce social mathematics, we must first determine the characteristics that describe the dynamics of the number of individuals before formalizing the unique properties that characterize them.

15.1.1 Birth rate and mortality tables (biometric features)

No one knows for sure the day nor the hour of his death with certitude (excepted for some special cases...). This individual trivial fact is no longer relevant if we are interested not only in one person, but a numerous collectivity (a population in other words). Then we have a play between those who succumb from premature accidents and those who escape almost miraculously to the worst dangers. We can then describe how the way that is paid globally the tribute to death by considering a large number of children born the same year and specifying, thanks to the civil status data (when the exist and are robust), how their numbers gradually decrease for a day cancel completely.

Such a set of people is named by demographers a "cohort". So consider the cohort of French males born in 11900 (holocene calendar). Each year, we can, by gathering the indications of the government, calculate the number of those who died in the cohort.

Let us represent by d_a the ratio the number of deaths between birthdays a and $a+1$ by the initial size of the cohort.

Remark

We will in what follows define the basis of probabilities required for actuarial calculation. Indeed, actuarial mathematics merge the financial calculation and the calculation of probabilities. The payment of a capital is not certain and depends, for example on the survival of a person (see the section of Quantitative Management Techniques for Actuarial Calculation at page 7547).

The sequence of these numbers contains all the necessary information to study mortality in this cohort.

We can infer the proportion of them surviving at age a:

$$S_a = 1 - d_0 - d_1 - d_2 - \ldots - d_{a-1} \tag{15.1.1}$$

We can also characterize the intensity of the mortality at each age by dividing the number of deaths between ages a and $a+1$ by the number of survivors. This number is then named the "mortality rate":

$$q_a = \frac{S_a - S_{a+1}}{S_a} = \frac{D_a}{S_a} \tag{15.1.2}$$

The list age by age of these three parameters d, S and q is the "mortality table" or "life table" of the studied cohort. The table below provides a summary for ages multiple of 5.

Age a	**Survivals** S_a	**Death from age** a to $a+5$ $S_a - S_{a+5}$	**Probability of death** $q_a = \dfrac{S_a - S_{a+5}}{S_a}$
0	1	0.228	0.228
5	0.772	0.013	0.017
10	0.759	0.009	0.012
15	0.750	0.023	0.031
20	0.727	0.023	0.032
25	0.704	0.020	0.028
30	0.684	0.021	0.031
35	0.663	0.025	0.038
40	0.638	0.029	0.045
45	0.609	0.026	0.043
50	0.583	0.036	0.062
55	0.547	0.047	0.086
60	0.500	0.062	0.124
65	0.438	0.078	0.178
70	0.360	0.360	

Table 15.1: **Mortality table example**

Typical public and governmental mortality tables that can be found for free on Internet include only gender and age as cofactors. But mortality tables in insurance companies contains much more cofactors as ethnic origin, religion, number of children, weight, height, wage, country, district, sport activity, driving license, known disease, sexual orientation, education level and many other variables that are kept confidential (data is the core of insurance companies)!

Remark

Follow a real generation of individuals throughout its existence or a specific period of time is named "longitudinal analysis" in contrast to "cross-sectional analysis", which is to study the characteristics of a population at a given time. The table above is thus an example of a longitudinal analysis.

Let us randomly select in the table the name of an unknown child in the list of births in the year 11900 (holocene calendar). The big question at this birth was: how many years will he live? Today we are able to retroactively answer this question, at least evoking probabilities because we know the mortality table of his cohort.

If the only information we have about this child today is the fact that he was born in 11900 (holocene calendar), we can state that the probability that he was still alive at the age of 5 years is equal to 0.772, at the age of 50 years of 0.583... In other words the probability that he died before 5 years is equal to 0.228.

If we now choose today an unknown individual included on the list of those who were included in the cohort during the year 11920 (holocene calendar) at the age of 20, we can still calculate the probabilities of the various periods of his life but we have additional information: he was still alive at age 20, he avoided the risk of death before this age. The probability that it then reaches the age of 50 years has become:

$$_{30|}p_{20}=\frac{S_{50}}{S_{20}}=\frac{0.583}{0.727}=0.802 \tag{15.1.3}$$

Therefore the probability for a person of age a to be alive at age $a+n$ is equal to:

$$\boxed{_{n|}p_a=\frac{S_{a+n}}{S_a}} \tag{15.1.4}$$

It is of course possible to make this calculation also for many individuals simultaneously. Indeed, in the field of insurance sector it is common for insurance to be a couple (husband and wife) where the age of each differs. Therefore, the joint probability of survival will logically be given by (see section Probabilities page 480):

$$_{n|}p_{ab}={}_{n|}p_a\cdot{}_{n|}p_b \tag{15.1.5}$$

Or if we are interested in the probability that one of the people is still alive (two incompatible events not necessarily disjoint as see in the section of Probabilities):

$$_{n|}p_{\overline{ab}}={}_{n|}p_a+{}_{n|}p_b-{}_{n|}p_a\cdot{}_{n|}p_b={}_{n|}p_a+{}_{n|}p_b-{}_{n|}p_{ab} \tag{15.1.6}$$

The probability of a person of age to die between age a and $a+n$ is logically given by:

$$_{n|}q_a=\frac{S_a-S_{a+n}}{S_a}=\frac{D_{a+n}}{S_a} \tag{15.1.7}$$

Thus at each age, we can give the law of the variable "time to stay alive". This law can be summarized by indicating its expected mean (see section Statistics page 634). An immediate calculation makes it possible to give the value for each age in function of the mortality table.

This can be exciting for a historian, but it gives an answer to a question asked long ago and since forgotten. What interests us is the present. This child that is just born, what is his life expectancy? To answer, we must know the mortality table of his generation, but the gift of premonition does not exist (as far as we know...). The probabilistic approach circumvents this difficulty as long as you specify the underlying assumptions.

So, to answer the question: what is the life expectancy of newborns in 11990 (holocene calendar), then we assume, completely free, they will encounter at every age, in the future, the conditions that 'encountered in 11990 (holocene calendar) people of these ages: in 12000 (holocene calendar) they will suffer the same mortality than that experienced in 11990 (holocene calendar) by those born in 11980 (holocene calendar). Of course, no one would imagine that the reality will be consistent with this hypothesis, but the calculation it allows provides a synthetic picture of current conditions of the struggle against death.

It is this assumption that life expectancy sometimes rejuvenates the old people (the life expectancy tends to increase as they become older thanks to advances in science...).

Let us take for the calculation of life expectancy, the mortality table for men in Switzerland in 11983-11993 (according to holocene calendar):

Age	**Survivals**	**Death Life Expectancy**	**Age**	**Survivals**	**Death Life Expectancy**
n	S_a	E(a)	n	S_a	E(a)
1	1.000000	73.688230	55	0.902240	22.810930
2	0.992460	73.248060	56	0.895810	21.974660
3	0.991830	72.294590	57	0.888750	21.149220
4	0.991480	71.320110	58	0.880990	20.335510
5	0.991170	70.342410	59	0.872470	19.534090
6	0.990900	69.361580	60	0.863120	18.745700
7	0.990660	68.378380	61	0.852880	17.970770
8	0.990440	67.393570	62	0.841690	17.209690
9	0.990220	66.408550	63	0.829480	16.463010
10	0.990010	65.422630	64	0.816180	15.731280
11	0.989710	64.442460	65	0.801740	15.014620
12	0.989600	63.449630	66	0.786090	14.313540
13	0.989380	62.463730	67	0.769180	13.628210
14	0.989150	61.478260	68	0.750960	12.958870
15	0.988890	60.494420	69	0.731380	12.305790
16	0.988600	59.512170	70	0.710400	11.669210
17	0.988230	58.534450	71	0.687980	11.049490
18	0.987700	57.565860	72	0.664090	10.446990
19	0.986920	56.611360	73	0.638730	9.861773
20	0.985810	55.675100	74	0.611900	9.294182
21	0.984390	54.755410	75	0.583620	8.744543
22	0.982850	53.841210	76	0.553930	8.213240
23	0.981310	52.925700	77	0.522910	7.700465

Table 15.2: **Mortality table in Switzerland in 11983-11993 (part 1)**

Age n	Survivals S_a	Death Life Expectancy E(a)	Age n	Survivals S_a	Death Life Expectancy E(a)
24	0.979750	52.009970	78	0.490660	7.206599
25	0.978150	51.095050	79	0.457330	6.731813
26	0.976530	50.179810	80	0.423090	6.276608
27	0.974900	49.263710	81	0.388190	5.840903
28	0.973280	48.345710	82	0.352880	5.425357
29	0.971660	47.426310	83	0.317480	5.030301
30	0.970070	46.504050	84	0.282340	4.656372
31	0.968500	45.579430	85	0.247850	4.304337
32	0.966940	44.652970	86	0.214460	3.974494
33	0.965410	43.723730	87	0.182630	3.667196
34	0.963880	42.793140	88	0.152810	3.382828
35	0.962360	41.860730	89	0.125460	3.120277
36	0.960820	40.927820	90	0.100910	2.879397
37	0.959270	39.993950	91	0.079420	2.658524
38	0.957680	39.060350	92	0.061110	2.455081
39	0.956040	38.127360	93	0.045930	2.266492
40	0.954340	37.195280	94	0.033730	2.086273
41	0.952570	36.264390	95	0.024140	1.915079
42	0.950710	35.335340	96	0.016800	1.751786
43	0.948740	34.408710	97	0.011340	1.595238
44	0.946650	33.484680	98	0.007390	1.447903
45	0.944410	32.564100	99	0.004640	1.306034
46	0.942010	31.647060	100	0.002790	1.172043
47	0.939420	30.734310	101	0.001600	1.043750
48	0.936620	29.826190	102	0.000870	0.919540
49	0.933560	28.923960	103	0.000450	0.777778
50	0.930230	28.027500	104	0.000210	0.666667
51	0.926590	27.137600	105	0.000090	0.555556
52	0.922590	26.255260	106	0.000040	0.250000
53	0.918210	25.380500	107	0.000010	0
54	0.913380	24.514710	108	0	
55	0.908080	23.657790			

Table 15.3: **Mortality table in Switzerland in 11983-11993 (part 2)**

Whose corresponding graph, named "the order of living" is:

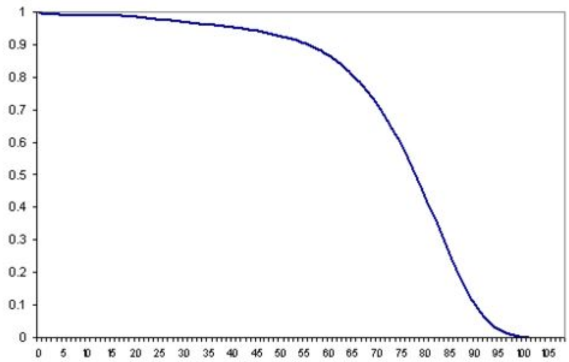


Figure 15.2: **Cumulative probability of survival as a function of age**

Let us see how to calculate life expectancy. For this, consider a human alive when he his a-th birthday. The number of years left to live is a random variable which we can calculate the expected value (see section Statistics page 585). Neglecting fractional years, that expected mean can be written:

$$\begin{aligned}\mathrm{E}(a)&=0\cdot {}_{0|}q_a+1\cdot {}_{1|}q_a+2\cdot {}_{2|}q_a+\ldots\\&=\frac{1}{S_a}\left(0D_a+1D_{a+1}+2D_{a+2}+\ldots\right)\\&=\frac{1}{S_a}\left(0(S_a-S_{a+1})+1(S_{a+1}-S_{a+2})+2(S_{a+2}-S_{a+3})+\ldots\right)\\&=\frac{1}{S_a}\left(S_{a+1}-S_{a+2}+2S_{a+2}-2S_{a+3}+\ldots\right)\\&=\frac{1}{S_a}\left(S_{a+1}+S_{a+2}+S_{a+3}+\ldots\right)\\&=\frac{1}{S_a}\sum_{i=1}^{\dagger}S_{a+n}\end{aligned}\tag{15.1.8}$$

If a is taken as zero, demographers speak of LEAB (Life Expectancy At Birth).

Here is for information the increasing life expectancy for men (source: French National Institute for Demographic Studies) from 11996 to 12006 (holocene calendar):

Year (holocene calendar)	Life expectancy
11996	74.1
11997	74.5
11998	74.8
11999	75.0
12000	75.3
12001	75.5
12002	75.7
12003	75.9
12004	76.7
12005	76.8
12006	77.2

Table 15.4: **Increased life expectancy at birth in France 11996-12006**

We can therefore observe using the above table that life expectancy increases by a little over one year every four years for more than 50 years (and until when ...???).

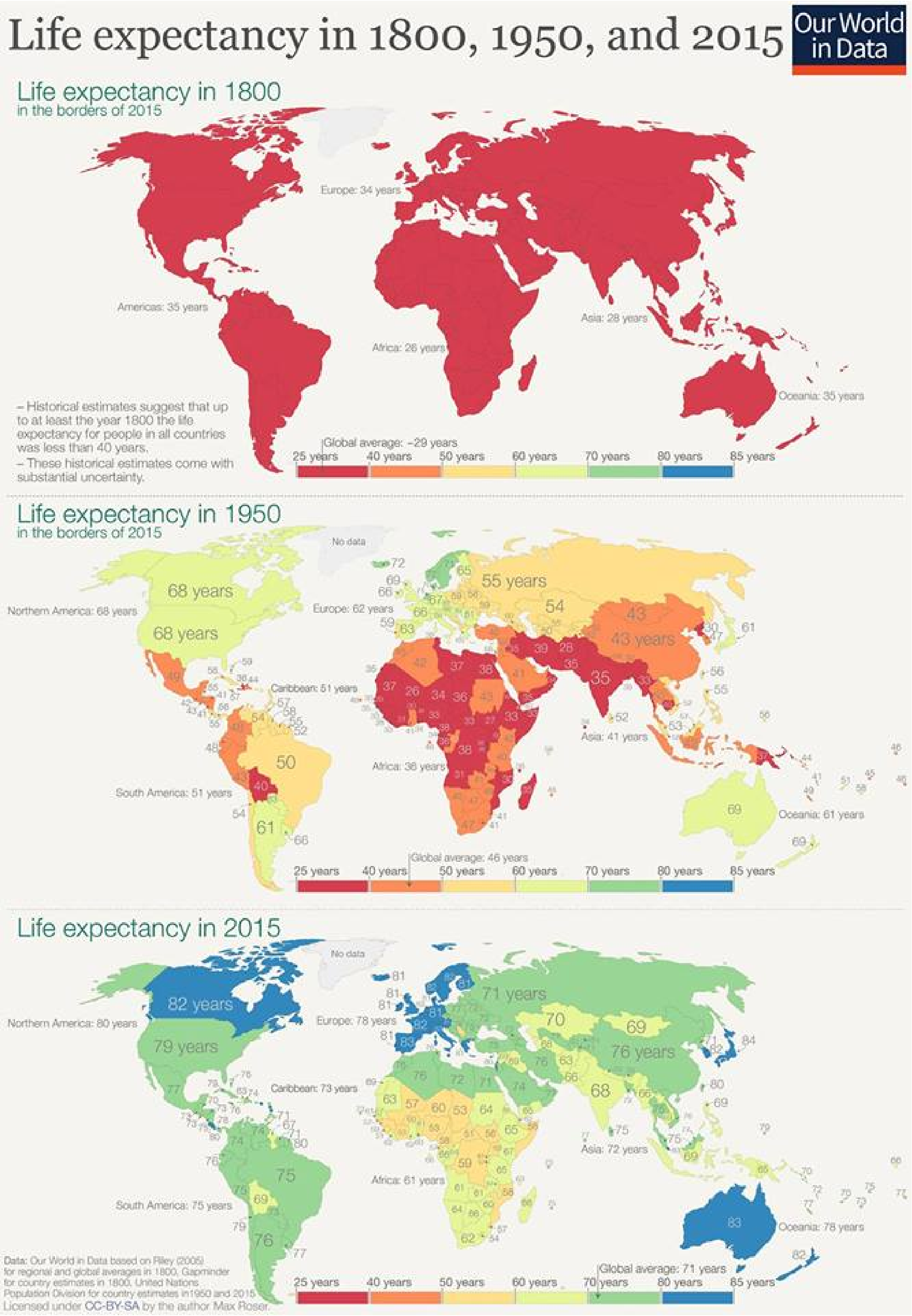

Life expectancy in 1800, 1950, and 2015
Our World in Data
Life expectancy in 1800
in the borders of 2015
Europe: 34 years
Americas: 35 years
Africa: 26 years
Asia: 28 years
Oceania: 35 years
– Historical estimates suggest that up to at least the year 1800 the life expectancy for people in all countries was less than 40 years.
– These historical estimates come with substantial uncertainty.
Global average: ~29 years
25 years
40 years
50 years
60 years
70 years
80 years
85 years
Life expectancy in 1950
in the borders of 2015
No data
Northern America: 68 years
68 years
68 years
Europe: 62 years
55 years
54
43
43 years
35
Caribbean: 51 years
South America: 51 years
50
61
Africa: 38 years
Asia: 41 years
Oceania: 61 years
69
Global average: 46 years
Life expectancy in 2015
No data
Northern America: 80 years
82 years
79 years
Europe: 78 years
71 years
70
69
76 years
68
Caribbean: 73 years
South America: 75 years
75
76
Africa: 61 years
Asia: 72 years
Oceania: 78 years
83
Global average: 71 years
Data: Our World in Data based on Riley (2005) for regional and global averages in 1800, Gapminder for country estimates in 1800, United Nations Population Division for country estimates in1950 and 2015
Licensed under CC-BY-SA by the author Max Roser.

Figure 15.3: **Life expectancy evolution (source: Max Roser)**

Or the life expectancy evolution in France since 11740 (holocene calendar):

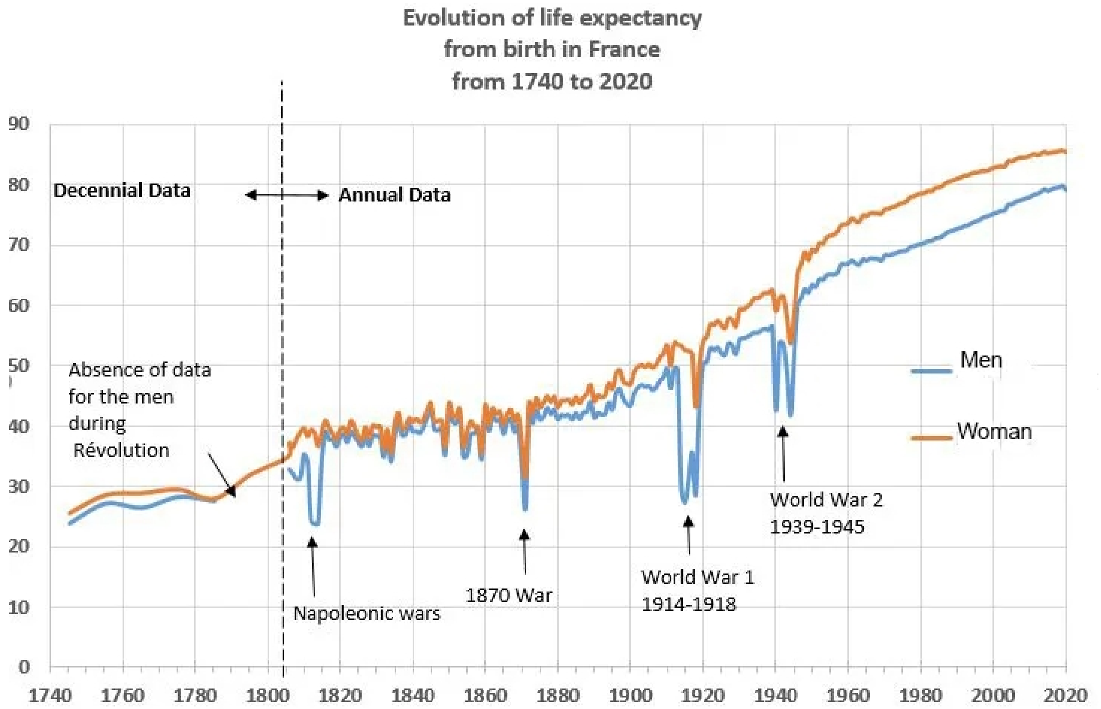
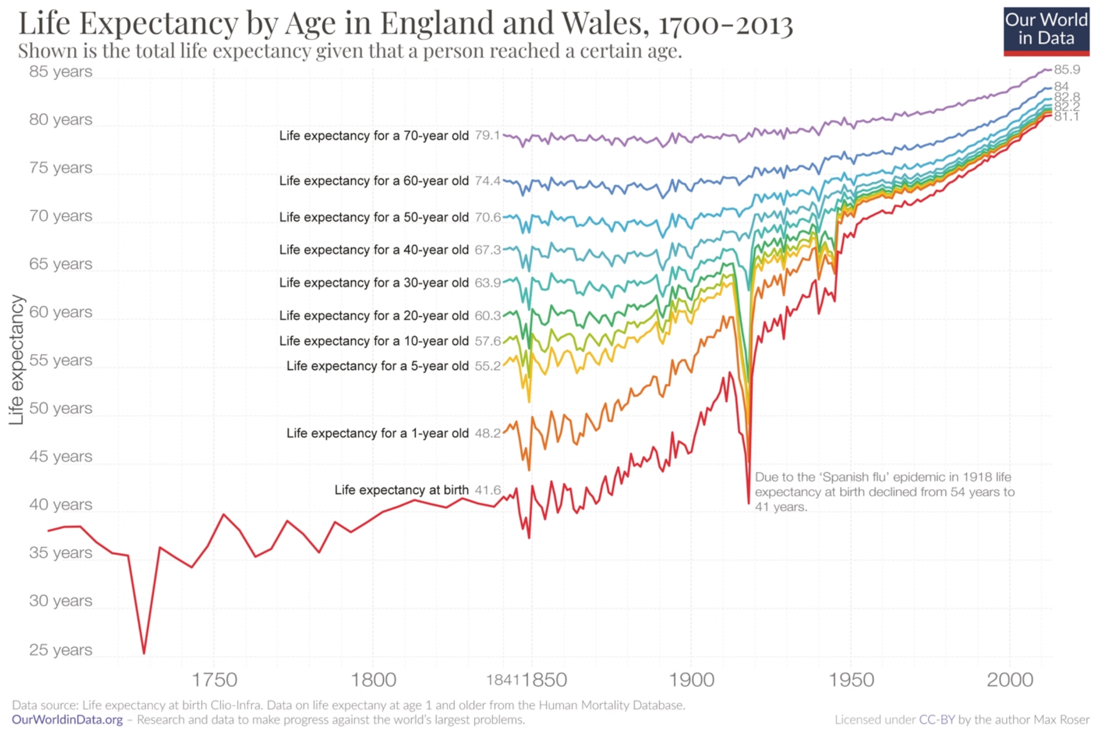


Figure 15.4: **Life expectancy evolution from birth in France since 11740 (source: INED)**

and an another one:



Data source: Life expectancy at birth Clio-Infra. Data on life expectancy at age 1 and older from the Human Mortality Database.
OurWorldinData.org – Research and data to make progress against the world's largest problems.
Licensed under CC-BY by the author Max Roser

Figure 15.5: **Life expectancy by age in England and Wales (11700-12013) (source: OurWorldinData)**

When it comes to transforming yearly mortality rates into monthly ones (the same reasoning applies for daily ones!), actuaries commonly employ two empirical methods when the raw data aren't available for such granularity of time:

- **Constant rate method:** The constant rate method is the simplest. It assumes that the survival rate is consistent for each month:

  $$ {}_{\frac{1}{12}}p_{x+\frac{m}{12}}=(p_x)^{\frac{1}{12}} \tag{15.1.9} $$

  For example, using the mortality table for Switzerland presented earlier above, the monthly probability of death for people of 25 year olds is:

  $$ {}_{\frac{1}{12}}p_{25+\frac{m}{12}}=1-(p_x)^{\frac{1}{12}}=1-\left(\frac{S_{26}}{S_{25}}\right)^{\frac{1}{12}}\cong 0.01381\% \tag{15.1.10} $$

  Or based on daily rate for a year with an average of 365 days:

  $$ {}_{\frac{1}{365}}p_{25+\frac{y}{365}}=1-(p_x)^{\frac{1}{365}}=1-\left(\frac{S_{26}}{S_{25}}\right)^{\frac{1}{365}}\cong 0.00045 \tag{15.1.11} $$

  Hence approximately one chance on $225,000$ to die every day between the age of 25 and the age of 26.

- **Uniform distribution of deaths:** Uniform distribution of deaths assumes that the number of deaths in each month is uniform. Under uniform distribution of deaths approach:

  $$ {}_{\frac{1}{12}}q_{x+\frac{m}{12}}=\frac{\frac{1}{12}\cdot q_x}{\left(1-\frac{m}{12}\cdot q_x\right)} \tag{15.1.12} $$

Finally, it must be known that actuaries define conditional life expectancy as the sum of:

$$ e(a,s)=a+\text{E}(a) \tag{15.1.13} $$

at the age a for the gender s. These calculations will be useful to us later for calculating pension insurance premiums in the section of Quantitative Management.

> Caution! Many debates (especially on social networks) confuses average life expectancy and expected maximum life span. The average has risen dramatically - in large part because fewer small children die. The maximum has not changed much. Pharaoh Pepi II might have been the longest reigning monarch.

Let's move to another topic. We know that to date, to make a child, we need most of time two people of opposite sex, certainly demographers know it too. But this double source of each of the newborn poses such problems in the description and analysis of fertility that in general they prefer to ignore it. Their attitude is justified by the fact that only the design requires the intervention of two people. At the time of birth, the mother acts alone. But demography is not interested in the designs, inaccessible to observation, but only to births.

We have seen how we can track a sample of men or women born in a given year and record the successive deaths, allowing us to establish this cohort mortality table. Similarly, we can note the numbers of children they give birth every year. We obtain then the "fertility table".

We just need to follow a cohort of women between 15 to 50 years to have an almost complete description of its reproductive behaviour. Data given by the government gives us the possibility to calculate each year the number of children that women have given birth in this cohort, grouped by age or age groups. Dividing by the rate of women surviving at this age, we get the "fertility rate" F_a.

If we add all these rate, we get the number of children that would have, on average, the women of this cohort if their mortality was zero. That was obviously not the case! To characterize how they ensured the renewal of their generation, we have to sum the real average of births, product of rate F_a by the survival rate S_a.

The set of these data is presented in a fertility table where we have an example where below where $n(a,a+5)$ is the number of births among survivors women survivors in this age group and where F_a is the number of births of $1,000$ women of this age group.

French women born around 11830 (holocene calendar):

Age n	**Survival rate** S_a	**Number of births** $n(a,a+5)$	**Fecundity rate** F_a
15	0.672	91	135
20	0.645	464	720
25	0.616	589	955
30	0.587	475	810
35	0.558	328	588
40	0.528	153	290
Total		2,100	3,498

Table 15.5: **Fertility table in France in 11830**

For example, on $1,000$ girls born in 11830 (holocene calendar), 645 have reached the age of 20 years and had 464 children between the age of 20 and 25 years. The intensity of fertility is measured by the number of births that would have $1,000$ women of this age:

$$\frac{464}{645}\cdot 1000=720 \tag{15.1.14}$$

And the average of births:

Age n	**Survivals**	**Number of births** $n(a,a+5)$
15	672	91
20	645	464
25	616	589
30	587	475
35	558	328
40	528	153
Total		2,100

Table 15.6: **Average of births in 11830 in France**

We can observe that in 11830 (holocene calendar), from 20 to 30 years old women gave birth to an average of:

$$\frac{464+589}{616}=\frac{1053}{616}=1.709 \tag{15.1.15}$$

this value being assimilated by the general public as the "fertility rate" (take care not to confuse with F_a).

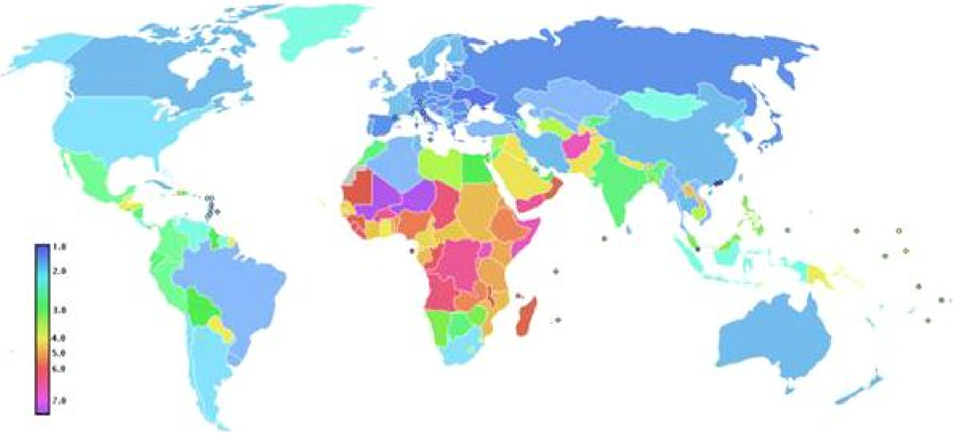


Figure 15.6: **Fertility rates in the early 121st century (source: Wikipedia)**

15.1.1.1 Population Renewal

The essential question for a set of human (or animals) in permanent renewal due to inflows and outflows that are births and deaths is: is the effective increasing or decreasing?

The female birth table gives the possibility to answer this question, thanks to the ratio of boys at birth to the number of girls. Thus, in the early 121st century (holocene calendar): 105 boys born for every 100 girls in the world. Thus, the proportion of girls is therefore:

$$\frac{105}{205}=0.49 \tag{15.1.16}$$

And the sex ratio is therefore:

$$S_r=\frac{105}{100} \tag{15.1.17}$$

Around forty years (49 in France according to INSEE), the balance is reversed and the number of women generally outweighs the number of men, despite significant regional disparities.

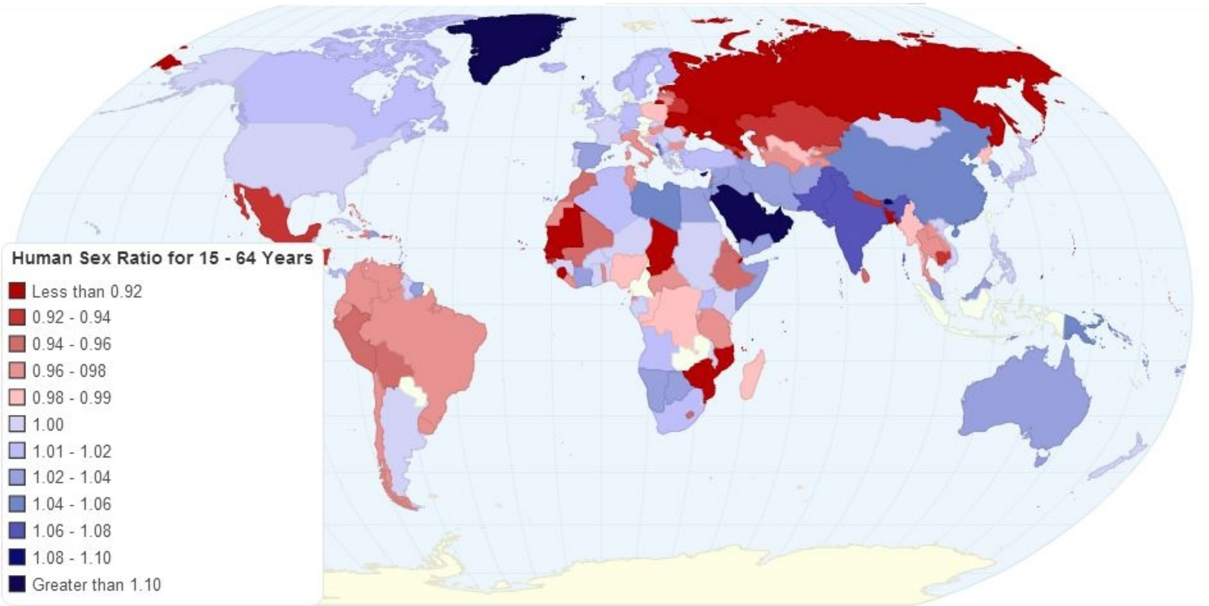


Figure 15.7: **Sex ratio trend in the early 121st century (source: chartsbin.com)**

Assuming the proportion was the same at the time, the fertility table in 11830 (holocene calendar) therefore shows that on average a woman in the 11830 (holocene calendar) cohort spawned:

$$\frac{2,100}{1,000}\cdot 0.49=2.1\cdot 0.49=1.03 \tag{15.1.18}$$

girl. That is to say and increase which we will denote k (in analogy with the exponential population model that we will see later) of 3%. Some politicians argue therefore that the value 2.1 is the necessary fertility rate ... it takes to ensure the renewal of generations and we have just seen that is not entirely accurate.

The female effective was therefore increasing. Therefore the population could ensure its renewal (unless the ratio is greater than 1).

The resulting number is the "net reproduction rate" (NRR). This rate is normally constituted by the ratio between the number of girls born for 100 women, this ratio is adjusted by the planned mortality between the birth of these girls and the average age of reproduction, because some of the girls don't reach not the age of reproduction, given the deaths among those between birth and age at childbirth. The average age at reproduction is given by the average age of mothers at birth.

An NRR of 1 means that each generation of mothers is having exactly enough daughters to replace themselves in the population. If the NRR is less than one, the reproductive performance of the population is below replacement level.

15.1.2 Population Models

Population dynamics is the branch of life sciences that studies the size and age composition of populations as dynamic systems, and the biological and environmental processes driving them (such as birth and death rates, and by immigration and emigration).

There are many mathematical models for studying the growth of a population as there all not all yet merged together in a single beautiful theory. Obviously the term "population" is used herein in the broadest sense - it can be a population of humans, animals, plants, individuals infected with a virus, etc.

As we know, to build a mathematical model, it is necessary to make assumptions. These assumptions play two roles: to preserve some essential features of reality and simplify this reality sufficiently so it can be studied by mathematics.

15.1.2.1 Exponential disease model

In this population dynamics model (one of the simplest), the event will be: the rate of change of the population is proportional, at any time t, to the population $P(t)$ at time t.

We can think, a priori, that this assumption is reasonable for a lot of situations. For example, more the human population is big, the greater the rate of change of this population will be, expressed in number of people added (or subtracted) per unit time, will be great. Similarly, the more people infected with a virus and more in the coming weeks, there will be new cases of people infected (if we don't take into consideration the limitness of the population and the methods to cure of a virus).

Mathematically, this assumption can be translated using the differential equation (see section Differential and Integral Calculus page 1605):

$$\frac{\mathrm{d}P(t)}{\mathrm{d}t}=kP(t) \tag{15.1.19}$$

So this differential equation is a mathematical model representing a situation where the growth rate of the population is proportional of a factor k to the size of the population at any time t. In this case, k is a constant named "growth rate" and we will see later how we can determine it. In some situations, the value of k is negative indicating that the population decreases with time rather than increasing. It is clear that a solution to this differential equation (see section Differential and Integral Calculus page 1605) is:

$$P(t)=c^{te}e^{kt} \tag{15.1.20}$$

First we will determine the value of the constant k from demographic data for the year 11965 (holocene calendar). At that time, there were almost 3 billion people on planet Earth. Moreover, at that time, the population grew by almost 54 million per year.

Therefore, in 11965 (holocene calendar):

$$\frac{\mathrm{d}P(t)}{\mathrm{d}t}=54\cdot 10^6=kP=k\cdot 3\cdot 10^9 \tag{15.1.21}$$

This give us:

$$k = 0.018 = 1.8\% \tag{15.1.22}$$

The differential equation is then:

$$P(t) = c^{te} e^{0.018t} \tag{15.1.23}$$

To determine the multiplicative constant, we simply have to put $t = 0$ and choose accordingly (since it corresponds to the initial condition). Thus, in 11965 (holocene calendar) we had:

$$P(t) = c_{11965}^{te} e^{0.018t} = 30 \cdot 10^9 e^{0.018t} \tag{15.1.24}$$

with t being the number of years after 11965 (holocene calendar).

If this mathematical model is consistent with reality, the solution found should allows us to estimate the population of humans on Earth to later times than 11965 (holocene calendar). Here is the graph of the function:

```
>plot(3*10^9*exp(0.018*t),t=1..50,labels=[years,population],
title='PopulationGrowth');
```

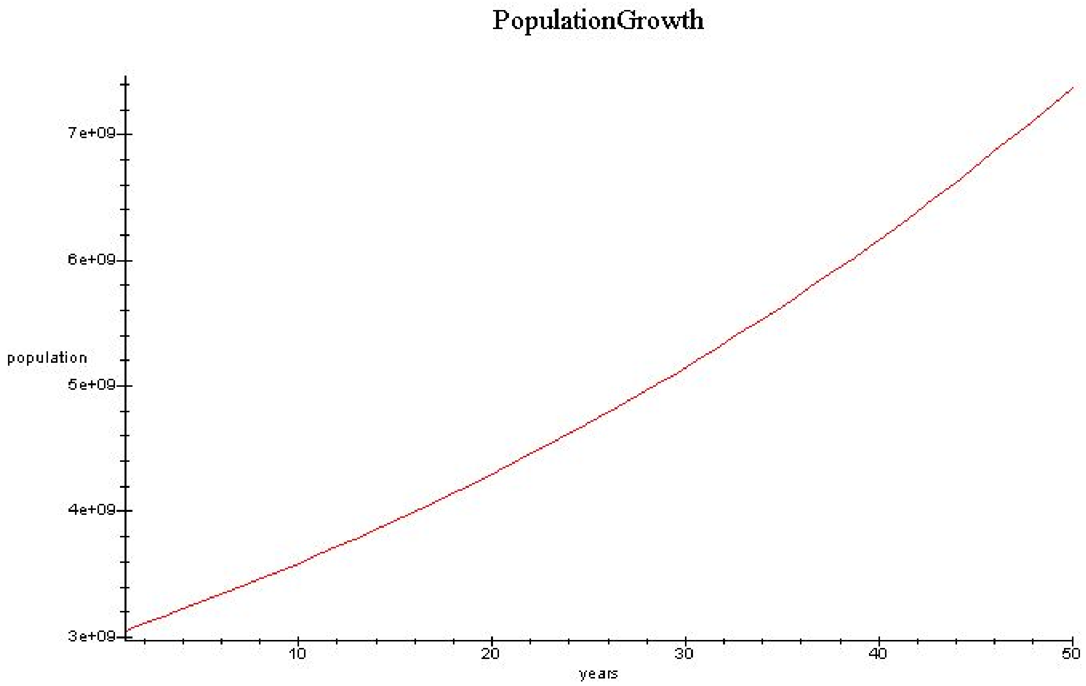


Figure 15.8: Evolution of the world population since 11965 in Maple™ 4.00b

If we evaluate $P(37)$, this will provide us with the prediction for the population in 12002 (holocene calendar). We find 5.84 billion which is pretty close to reality.

The readers can also easily check the following table according to the values of k:

Growing rate k per year	Doubling time in years
0.5%	139
1%	69
1.5%	46
2%	35
2.5%	28
3%	23
3.5%	20
4%	17

Table 15.7: Various costs

We then see immediately, that according to the exponential model, a growth that may seem slow, on the order of 3% per year, is truly made an explosive results since doubling every 23 years. That is to say a multiplication over 17 in a century. Or a century passed quickly (at least for physicists because politicians seems to not be able to think on more than 50 years)!

This simple observation shows how false are in this model the arguments that make us seek the solution of an economic or social problem in growth, and even, as it is often admitted in sustainable growth!!! It is clear that no growth can truly last, it is only a transitory episode, necessarily followed by a plateau or even a decrease. Resolve an economic issue by the growth or its stability is thereof transfer the problem to later, at a period at which it will be necessary to find ways both: to stop the growth or to resolve it in another way to solve a problem hidden by politicians not listening to scientists since to long time (more than 150 years!!!).

The growth model we have just see right now may obviously not be valid over very long periods of time. Indeed, if we calculated, using the above equation, the population in 7 centuries the result would be that on every square meter of land, excluding water, there would, on average, be ten human! Similarly, a population of people infected with a virus can also not really be described by such a model.

These results tell us that if we want to develop a model of the growth of a population, that is more in line with reality, we have to change our initial assumptions. This is what we will see after with deterministic logistic models.

But first let us give a nice example where an exponential explosion also occurs. Every individual normally has on Earth, as we know, two parents, four grandparents and eight great-grandparents and so on in every generation, if we go back in time, the number of ancestors doubles. But it can not double indefinitely and exponential growth must stop at some point or another! If the number of ancestors doubles with each generation and that we go back only 2,000 years ago, we would individually have about 1023 ancestors considering only 25 years for a generation ... Two thousand years before, which is not so distant, all the ancestor of all our actual population would then represent roughly the equivalent of the total mass of the Earth if this mass was made only of human bodies! The answer to this apparent paradox is of course that there are a lot (really lot!) of repetitions in each family tree (parents with several children ... and fathers with several women... and vice versa) and so we are all in a relatively short period cousins...

We also have the following data on the evolution of the world population (all the years are given relatively to the holocene calendar):

Year	World Population
-90000	0.5 million
0	1 to 10 million
4500	5 to 10 million
5000	5 to 10 million
9800	150 to 20 million
10000	170 to 400 million
10200	190 to 206 million
10400	190 to 206 million
10500	190 to 206 million
10600	200 to 206 million
10700	207 to 210 million
10800	220 to 224 million
10900	226 to 240 million
11000	254 to 345 million
11100	301 to 320 million
11200	360 to 450 million
11250	400 to 416 million
11300	360 to 432 million
11400	350 to 374 million
11500	425 to 540 million
11600	545 to 579 million
11650	470 to 545 million
11700	600 to 679 million
11800	0.813 to 1.125 million
11850	1.128 to 1.402 million
11900	1.550 to 1.762 million

Year	World Population
11910	1.750 billion
11920	1.860 billion
11930	2.07 billion
11940	2.3 billion
11950	2.519 billion
11960	3.023 billion
11965	3.337 billion
11970	3.696 billion
11975	4.073 billion
11980	4.442 billion
11985	4.843 billion
11990	5.279 billion
11995	5.692 billion
12000	6.085 billion
12005	6.541 billion
12010	6.956 billion
12015	7.379 billion
12020	7.795 billion

Table 15.8: **Evolution of the world population (source: Wikipedia)**

We see well an exponential growth for now ... but still not with the values corresponding to the small example calculation of our family tree even if we take the cumulative value!

15.1.2.2 Deterministic Logistic Model (Verlhust)

Now we will focus on a different type of model than the exponential one (where the population explodes) which has the advantage of having an asymptotic behaviour rather than a diverging one.

This type of behaviour is interesting because resources are usually limited and there is a competition between individuals. The logistic model, also named "Verhulst model" allows to account for this relatively well.

Let $N(t)$ be the population at time t. Let us write:

$$\frac{\mathrm{d}N}{\mathrm{d}t}=rN(t) \tag{15.1.25}$$

where r is the rate of increase that this time will not be constant and is defined as being:

$$r=r_0\left(1-\frac{N(t)}{K}\right) \tag{15.1.26}$$

where the idea is that K is the maximum capacity of the medium (parameter sometimes named "carrying capacity"). We see that if K is infinite we fall immediately on the exponential model and if $N(t)$ equals K then r is zero and therefore the model is asymptotic. The constant r_0 is determined by adjusting the model to real datas (likelihood method or conjugate gradient method like seen in the section of Numerical Methods).

Finally we have:

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t}=r_0\left(1-\frac{N(t)}{K}\right)N(t) \tag{15.1.27}$$

which gives after rearrangement:

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t}=r_0N(t)\left(1-\frac{N(t)}{K}\right) \tag{15.1.28}$$

Or mathematically:

$$y'=r_0y\left(-\frac{y}{K}\right) \tag{15.1.29}$$

with as initial condition that $y(0)=N_0$.

We have now to solve this differential equation (See Differential and Integral Calculus):

$$y'=r_0y\left(1-\frac{y}{K}\right) \tag{15.1.30}$$

We put $y=1/f(x):=1/f$. The differential equation then becomes:

$$-\frac{f'}{f^2}=r_0\frac{1}{f}\left(1-\frac{1}{Kf}\right) \tag{15.1.31}$$

which simplifies to:

$$-f'=r_0f\left(1-\frac{1}{Kf}\right)=r_0f-\frac{r_0}{K} \tag{15.1.32}$$

which gives after rearrangement:

$$f'+r_0f-\frac{r_0}{K}=0 \tag{15.1.33}$$

This differential equation is solved as usual using conventional techniques studied in the Differential and Integral Calculus section. We first write the homogeneous equation:

$$f'+r_0f=0 \tag{15.1.34}$$

An immediate special solution is:

$$f(x)=Ae^{-r_0x} \tag{15.1.35}$$

where A is a constant. By injecting this solution in the initial differential equation this gives:

$$-r_0Ae^{-r_0x}+r_0Ae^{-r_0x}-\frac{r_0}{K}=-\frac{r_0}{K}=0 \tag{15.1.36}$$

We see immediately that for equality to be satisfied our special solution becomes the general solution if we write it in the form:

$$f(x)=Ae^{-r_0x}+\frac{1}{K} \tag{15.1.37}$$

The solution of the differential equation is therefore:

$$y=\frac{1}{f}=\frac{1}{Ae^{-r_0x}+\dfrac{1}{K}}=\frac{K}{KAe^{-r_0x}+1} \tag{15.1.38}$$

The initial condition:

$$y(0)=N_0 \tag{15.1.39}$$

Takes us to:

$$y(0)=\frac{K}{KAe^{-r_0x}+1}=\frac{K}{KAe^{-r_00}+1}=\frac{K}{KA+1}=N_0 \tag{15.1.40}$$

Therefore:

$$K=N_0(KA+1)\Rightarrow A=\frac{K-N_0}{KN_0} \tag{15.1.41}$$

This gives us:

$$\begin{aligned}y&=\frac{K}{KAe^{-r_0x}+1}=\frac{K}{K\dfrac{K-N_0}{KN_0}e^{-r_0x}+1}=\frac{K}{\dfrac{K-N_0}{N_0}e^{-r_0x}+1}=\frac{K}{1+\left(\dfrac{K}{y_0}-1\right)e^{-r_0x}}\\&=\frac{y_0}{y_0}\frac{K}{1+\left(\dfrac{K}{y_0}-1\right)e^{-r_0x}}=\frac{y_0K}{y_0+(K-y_0)\,e^{-r_0x}}\end{aligned} \tag{15.1.42}$$

Either in traditional form in the field of population dynamics:

$$\boxed{N(t)=\frac{N_0K}{N_0+(K-N_0)\,e^{-r_0t}}} \tag{15.1.43}$$

which is therefore the final expression of logistic deterministic model. We see easily that:

1. If t tends to infinity then the horizontal asymptote is K.
2. At time $t=0$, the initial population is equal to $N(0)=N_0$.

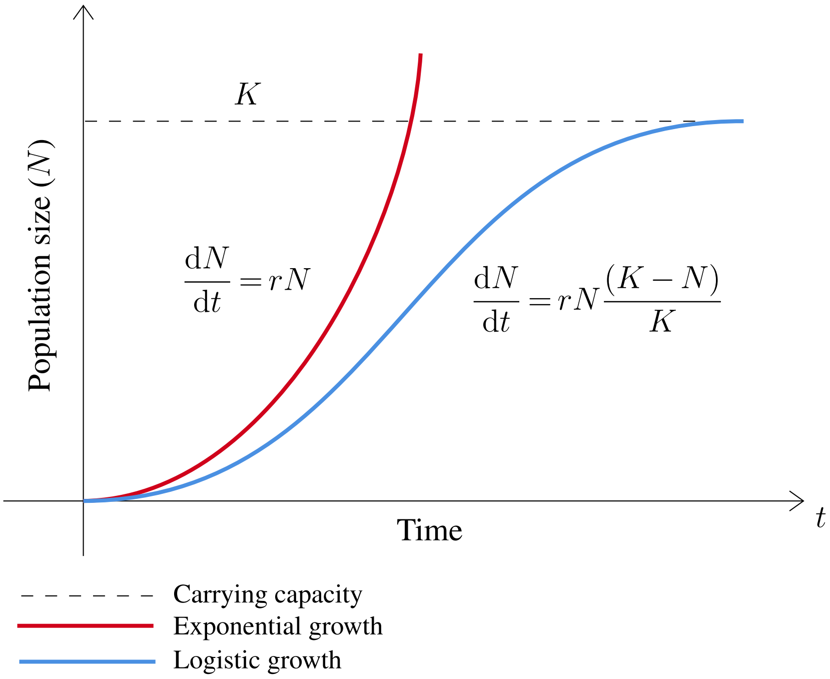


Figure 15.9: **Logistic model vs Exponential model**

To finish by returning to the previous form of our development:

$$y=\frac{K}{1+\left(\frac{K}{y_0}-1\right)e^{-r_0t}} \tag{15.1.44}$$

If we rewrite with a small change of variable we get:

$$y=\frac{K}{1+ae^{-rt}} \tag{15.1.45}$$

we obtain what is commonly named by its form a "logistic function".

15.1.2.3 Chaotic Logistic Model

We will still assume now that each generation is proportional to the previous one. We will suppose from one period to the next that the evolution of the population can then be translated into a sequence of the type:

$$x_{t+1} = Kx_t \tag{15.1.46}$$

where x_t is the population at time t, x_{t+1} the population at time $t+1$ and K the reproduction ratio.

As the previous model this suppose obviously that the initial population size is not zero...

But we quickly realize that such a model is unrealistic: the Earth would have since long time be submerged by a human tide. Indeed, a power model such this one ignores important factors like the fact that a priori resources are not unlimited.

To establish a more satisfactory model we must take into account (or by prudence have the pessimistic point of view) that the available resources are limited and that, for any country, there is a maximum population beyond which the population decreases, regardless of the species. To find a function that translates more realistically the evolution of a population, let us enumerate, simplifying a bit, the properties that must satisfied this evolution:

1. This is an iterative phenomenon: If x_{t+1} is the size of the population of the period $t+1$, it depends on that of the previous period t.
2. A population can not grow indefinitely in a defined territory: there is a maximum after which it must decrease. We must plan a retroactive factor limiting the population increase when its density becomes too high. To simplify the calculations, x will not represent anymore the population in absolute number but a percentage of that maximum value corresponding to a given territory. The idea is than that x_t can fluctuate between only between 0 and 1.
3. K remain the actual growth rate from one period to the next.

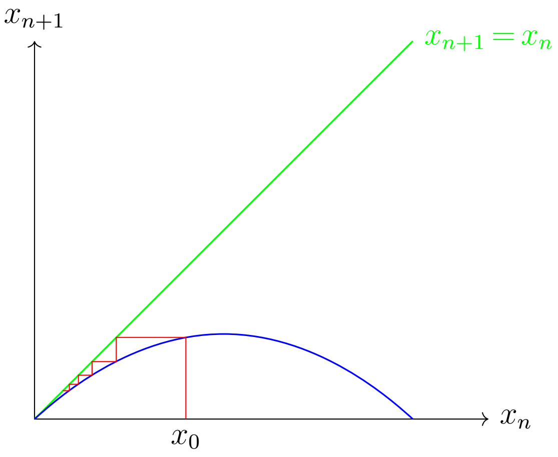
The size of the population of the period $t+1$ will be the value x_{t+1}, expressed as a percentage of the maximum population that can accommodate the given territory, and will be obtained from the size of the population x_t of the previous period t by the supposed relation:

$$\boxed{x_{t+1} = Kx_t(1-x_t)} \tag{15.1.47}$$

the factor $(1-x_t)$ represents obviously the retroactive effect. This suite is often named "logistic sequence", "logistic equation" or "logistic map" and we also find it in many other areas of non-linear physics.

Indeed, when the population density is high, close to saturation, the x_t is close to 1 and therefore $1-x_t$ will be close to 0. So this retroactive factor will tend to minimize the population x_{t+1}.

Plot of the case where $K<1$ (deaths exceed births and the population decays to zero):



We see that $x_n = 0$ is a fixed point. In general, to find fixed points, we solve for $x_{n+1} = x_n$, i.e. $f(x_n) = x_n$. For the logistic map, we have:

$$K x_n (1 - x_n) = x_n \tag{15.1.48}$$

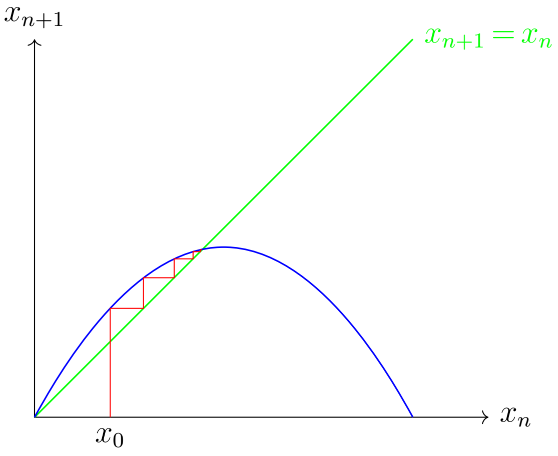
Therefore:

$$x_n [1 - K(1 - x_n)] = 0 \tag{15.1.49}$$

Hence:

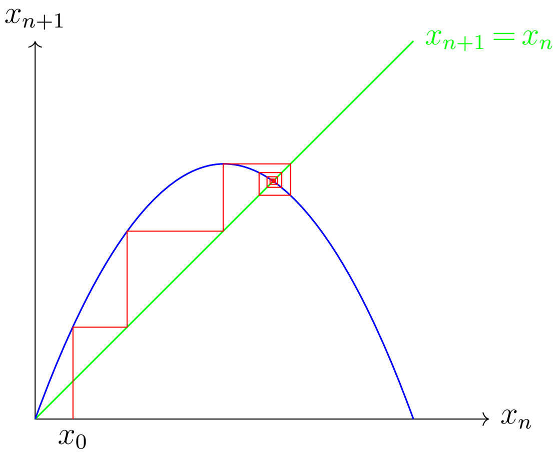
$$x_n = 0 \text{ or } x_n = 1 - \frac{1}{K} \tag{15.1.50}$$

When $1 < \lambda < 2$, we have:



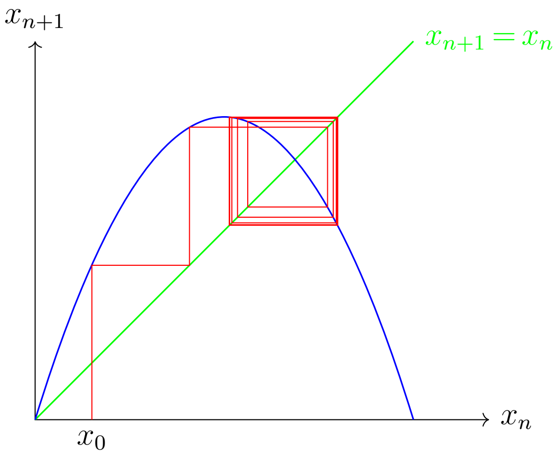
We see that $x_n = 0$ is an unstable fixed point and $x_n = 1 - \frac{1}{K}$ is a stable fixed point.

When $2 < K < 3$, we have:



There is an oscillatory convergence to $x_n=1-\frac{1}{K}$.

When $K>3$, we have a limit cycle, in which x_n oscillates between two values, i.e. $x_{n+2}=x_n$. When $K=1+\sqrt{6}\cong 3.449$, we have a 4-cycle, and so on:



Thus, for certain animal species, it is normal that population change regularly (oscillates), while for others it is normal to move towards an equilibrium situation. This variety, "chaos" in some cases, is even linked to the mathematical properties of the previous relation.

We will see that depending on the value of its actual growth rate k, an animal population (including therefore humans), may tend toward a state of equilibrium, or fluctuate between two or four or eight values, or change completely pseudo-randomly.

The complexity becomes the rule and not the exception, and what appears "chaotic" follows very specific properties from a very specific closed form function.

It is these results that we will see. For this, we will study the evolution of animal populations (this includes obviously therefore humans) whose numbers growth rate K varies from 1 to 4. We will study these variations by the calculation and by studying the graph of the logistic function.

Let us see now some special cases:

1. If the initial value x_0 is equal to 0.1 and $K = 1$ the evolution of the population for the first forty periods is represented by the graph below:

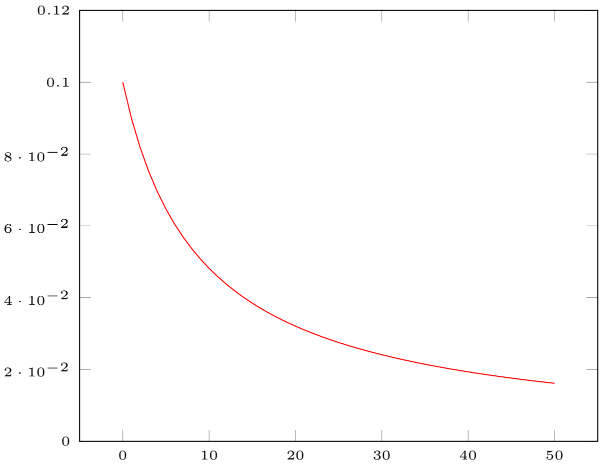


Figure 15.10: Plot of the logistic equation with $x_0 = 0.1$ and $K = 1$

We note that for this growth rate and this initial condition, the population will decrease and approach zero.

2. If the initial value x_0 is equal to 0.1 and $K = 2$ the evolution of the population for the first forty periods is represented by the graph below:

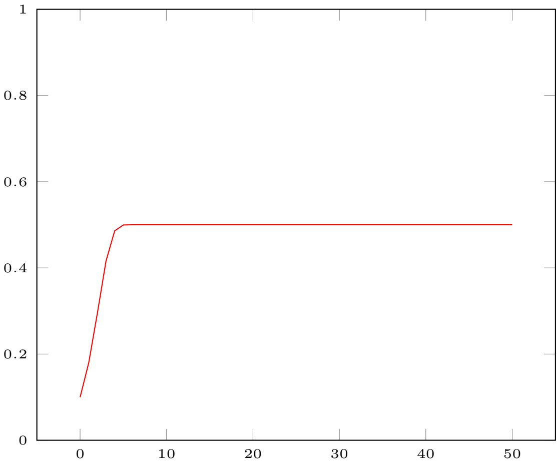


Figure 15.11: Plot of the logistic equation with $x_0 = 0.1$ and $k = 2$

We note that for this growth rate, the population will stabilize around a number corresponding to half of the population that the territory could accommodate (remember that $1 = 100\%$ on the y-axis).

3. If the initial value x_0 is equal to 0.1 and $K = 3$ the evolution of the population for the first forty periods is represented by the graph below:

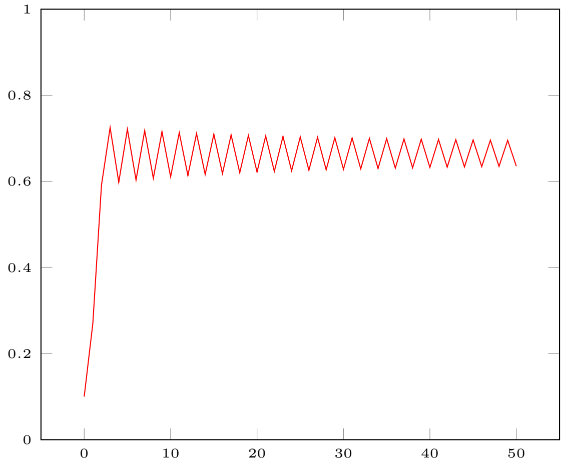


Figure 15.12: **Plot of the logistic equation with $x_0 = 0.1$ and $k = 3$**

With a growing rate of 3, the number of individuals of this population starts to oscillate between two values that can be calculated using the computer. We then obtain the values 0.64 and 0.68.

4. If the initial value x_0 is equal to 0.1 and $K = 3.5$ the evolution of the population for the first forty periods is represented by the graph below:

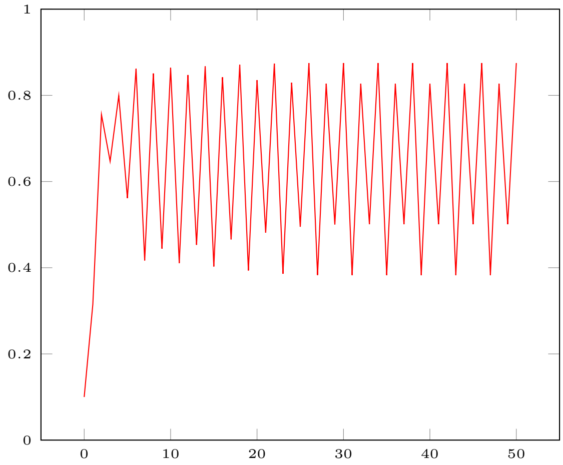


Figure 15.13: **Plot of the logistic equation with $x_0 = 0.1$ and $K = 3.5$**

With an effective growth rate of 3.5 the population oscillates between four values: 0.39 and 0.83, then 0.49, and finally 0.87. The evolution of a population that has such actual growth rate is clearly cyclical.

5. If the initial value x_0 is equal to 0.1 and $K = 4.5$ the evolution of the population for the first forty periods is represented by the graph below:

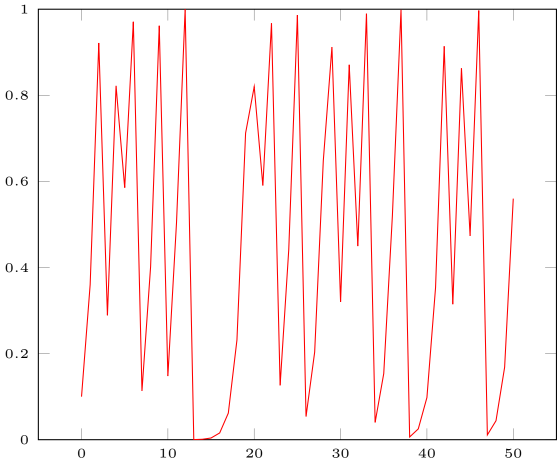


Figure 15.14: Plot of the logistic equation with $x_0 = 0.1$ and $K = 4.5$

We notice that when the growth rate is equal to 4.5, the number of individuals of seems to oscillate irregularly in a "chaotic" way between two extremes: the saturation as x approaches 1 and extinction as x approaches 0.

The question now is what is the behaviour of this function for other initial values of population?

To find out why, we take the above calculations with an initial population representing 0.8 (80%) of the maximum population of 1 (100%) for a given area and compare the results.

The population growth is always given by the equation:

$$x_{t+1} = Kx_t(1 - x_t) \tag{15.1.51}$$

1. Depending on to the initial population is $x_0 = 0.1$ or $x_0 = 0.8$ we get for $K = 1$:

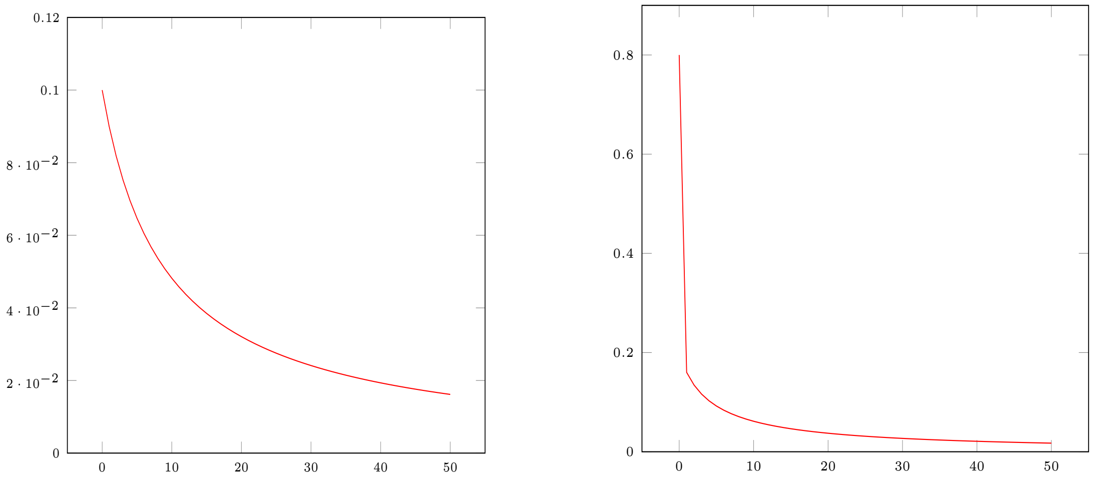


Figure 15.15: Small k and $x_0 = 0.1$

Figure 15.16: Small k and $x_0 = 0.8$

Figure 15.17: Comparison plots of logistic equations for small k

2. Depending on to the initial population is $x_0 = 0.1$ or $x_0 = 0.8$ we get for $K = 2$:

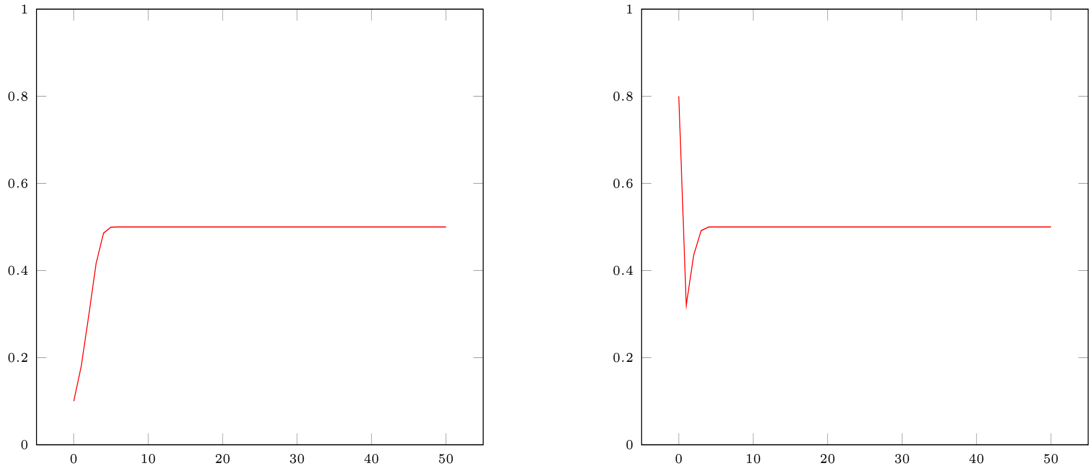
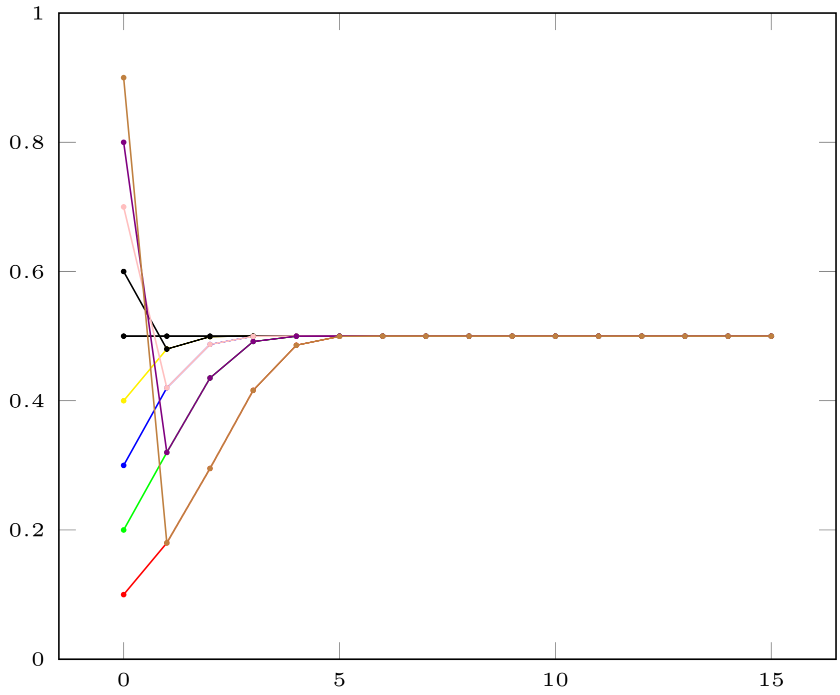


Figure 15.18: Medium k and $x_0 = 0.1$

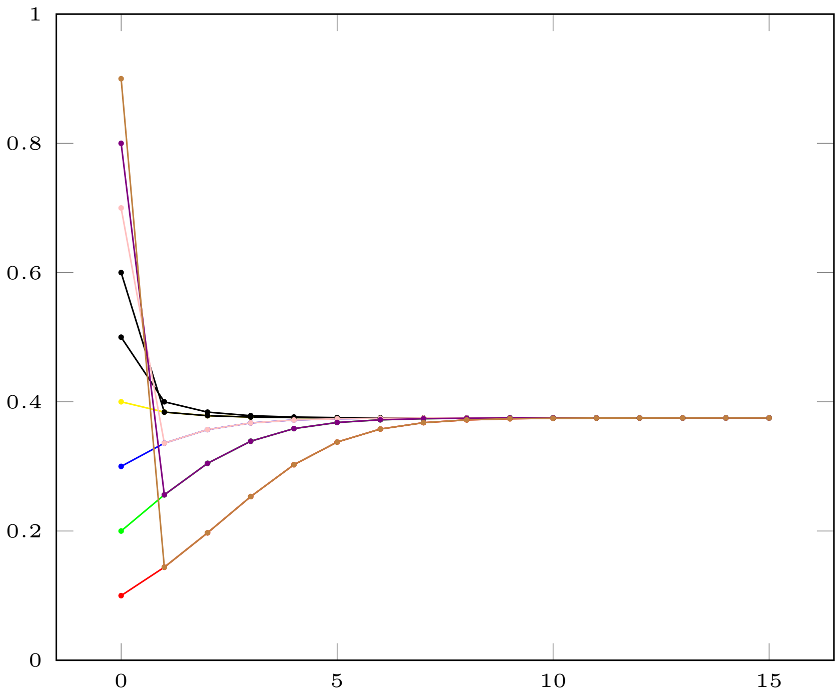
Figure 15.19: Medium k and $x_0 = 0.8$

Figure 15.20: Comparison plots of logistic equations for medium k

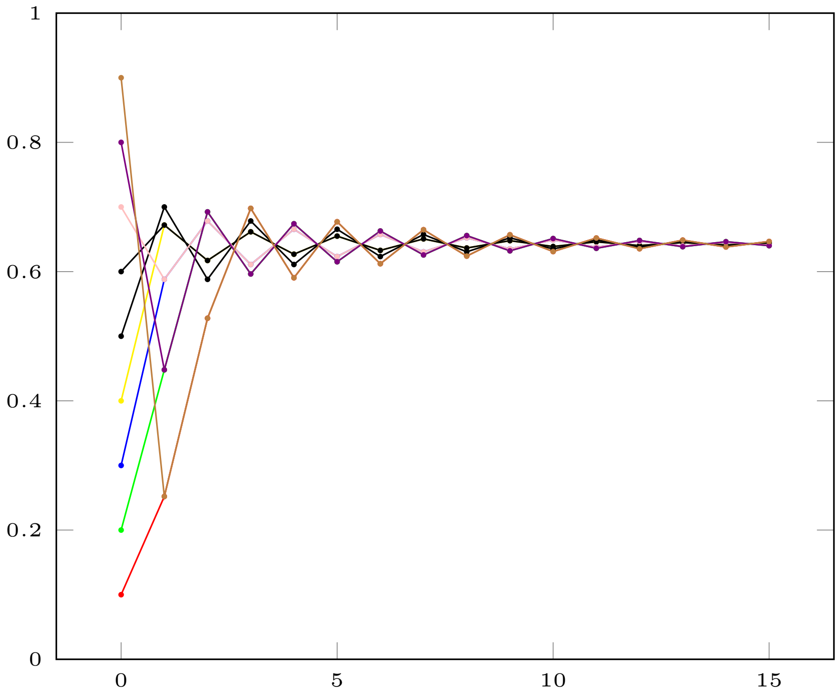
We see that regardless of the initial population, it tends to a stabilization (equal to 0.5) after a few iterations named an "attractor". This is even more visible with the following plot:



or with $K = 1.6$ we see that the attractor is at almost 0.38:



or with $K = 2.8$ we see that the attractor is at almost 0.64:



3. Depending on to the initial population is $x_0 = 0.1$ or $x_0 = 0.8$ we get for $k = 3.8$:

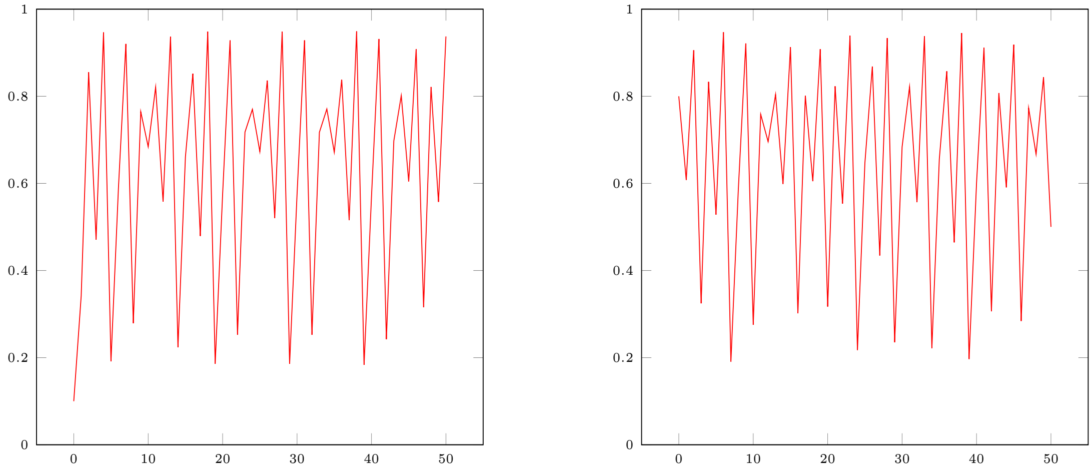


Figure 15.21: **Large k and $x_0 = 0.1$**

Figure 15.22: **Large k and $x_0 = 0.8$**

Figure 15.23: **Comparison plots of logistic equations for large k**

We can notice here that the evolution of populations is very different. We took remote initial populations. If we had very similar initial populations, what would have been the evolution of the two populations?

4. Depending on to the initial population is $x_0 = 0.8$ or $x_0 = 0.8001$ we get for $k = 3.8$:

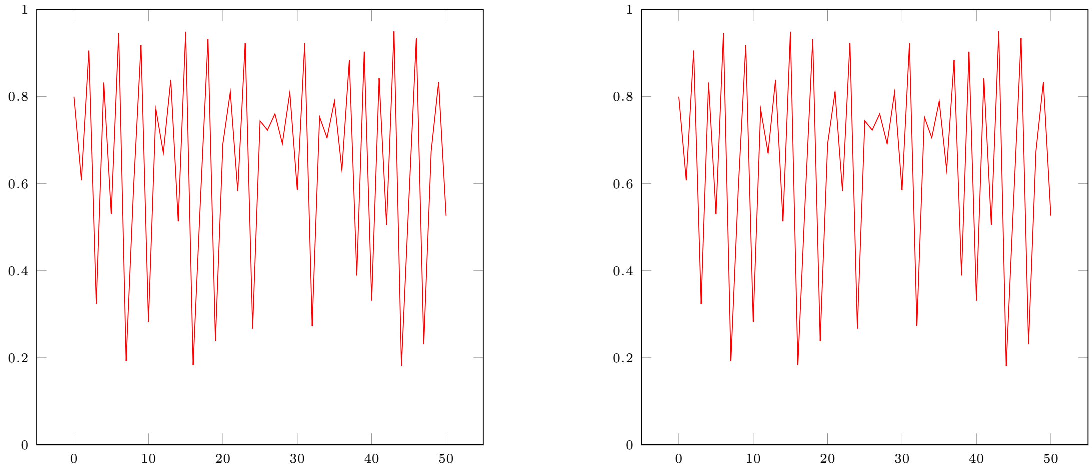


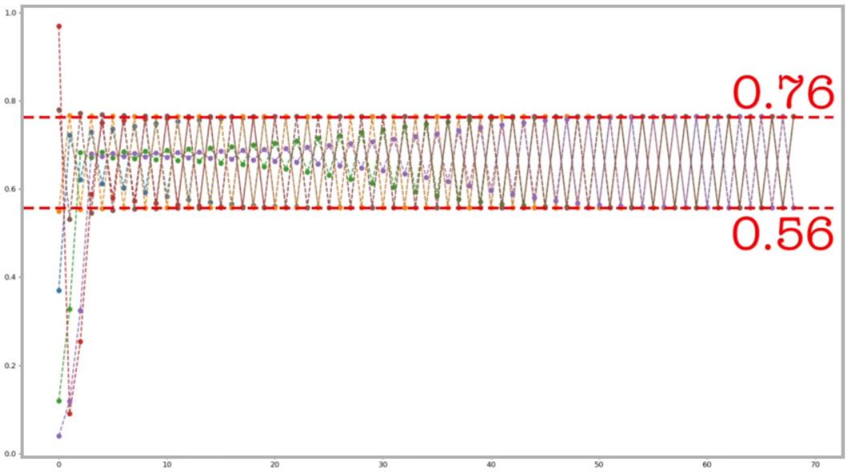
Figure 15.24: **Large k and $x_0 = 0.8$**

Figure 15.25: **Large k and $x_0 = 0.8001$**

Figure 15.26: **Comparison plots of logistic equations for one large k and small δk**

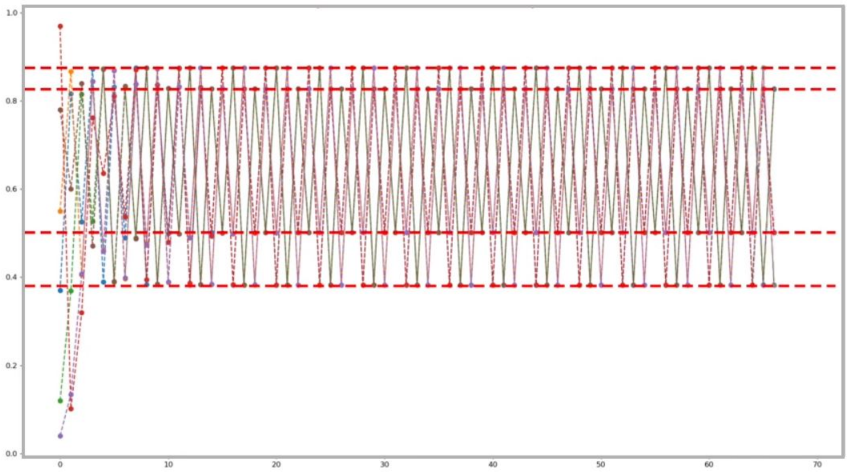
We can notice here that evolution of the populations are very different even though they were very close initially and even if the evolution function of these populations is a very simple expression. If we had taken as initial population $x_0 = 8.00001$ we could have also found a very different evolution of population trends.

This oscillation is even more visible with the following plot:

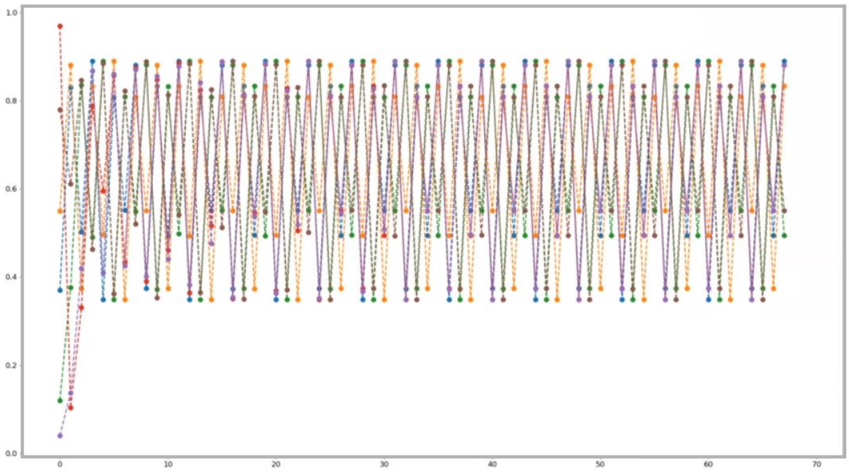


We see above that the points oscillates between two bounds. This is not an attractor with a unique point, but an attractor with two points. The attractor is then named a "periodic attractor".

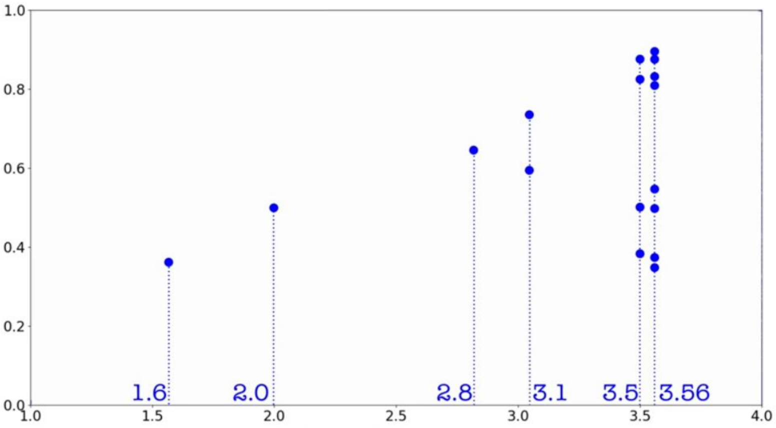
With $K = 3.5$ there are 4 periodic attractors:



With $K = 3.56$ there are 8 periodic attractors:



Now if we plot the attractors value in function of the iterations for difference values of the K above we get the following dot point plot:



The conclusion of this model is that outside of the chaotic phase, the initial value is not important, but in this chaotic phase, otherwise, the smallest variation in initial value change completely the following values. This is named the "butterfly effect" (for more details about the Butterfly Effect see the mathematical developments of the section of Marine & Weather engineering).

15.1.2.3.1 Feigenbaum's Bifurcation Diagram

To understand the evolution of a population with logistic model, we have previously showed the evolution over time, forty periods, of a population corresponding to a specific initial value and a constant k determined. We have seen, by taking some special values of k, that for these different values, the population had a different pattern. So we will study now the behaviour of the logistics by taking k as variable.

Giving to k values between 0 and 4 with a step of 0.02, we will, for each of these values, calculate what will be the population for each period between the 30th and 130th.

With Maple™ 4.00b we get (in the `R` companion book the equivalent script is also given in `R` language):

```
 >with(plots):  with(plottools);
>feigenbaum:=proc(start,end,step) local k,iteration,a,b,s;
s:=; a:=stat;
while a<=end do iteration:=0.1;
for k to 50 do iteration:=a*iteration*(1-iteration) od;
for k to 100 do iteration:=a*iteration*(1-iteration);
s:=s union [a,evalf(iteration,4)];
od;
a:=a+step
od;
plot([op(s)],'a'=start..end,style=POINT,symbol=POINT)
```

```
end:
>feigenbaum(1,4,0.01);
```

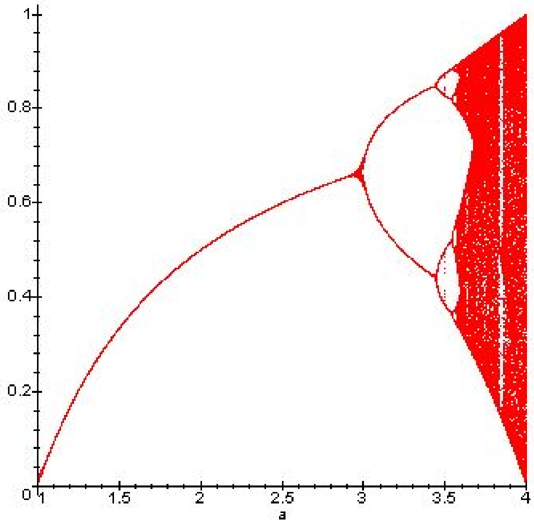
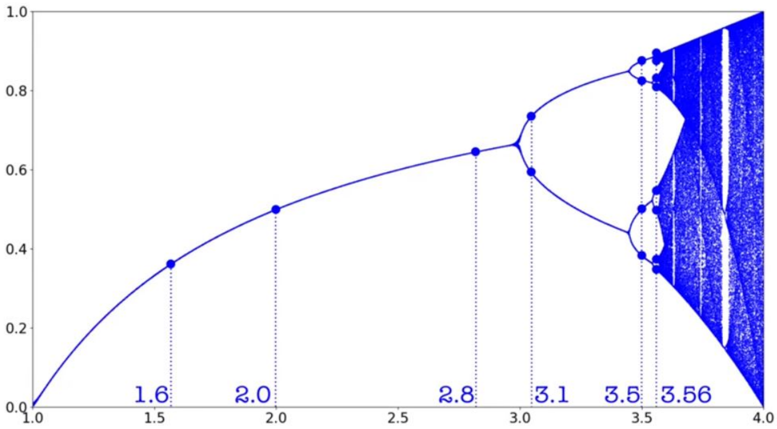


Figure 15.27: **Feigenbaum diagram in Maple™ 4.00b**

This diagram (where we can observe bifurcations), is named "Feigenbaum diagram", named after the physicist Mitchell Feigenbaum who studied it in depth and showed that we found him in many natural phenomena.

We can identify in the diagram above the different attractors hat we have seen just earlier:



We will now have a qualitative approach of two remarkable properties of this diagram: the doubling period and its fractal dimension!!

Going back to our example where the logistics function gives the evolution of a population according to its actual growth rate k, the Feigenbaum diagram indicates that, when the rate is less than 3, the system tends towards a stable final state. This generally corresponds to our intuition influenced by our unconscious desire for order and simplicity.

But starting from 3, the things become little bit more complicated: the number of individuals per generation starts oscillating between 2 and 4, then 8 values ... to finally get into a "chaotic" area where all values seem possible. There is already something fascinating. But if you observe more closely this doubling of interval before the chaotic area, we will discover even more curious results.

So we first take the sequence of the three first three duplications with 2, then 4, then 8 branches, for an effective growth rate k on an interval $k = [3, 3.564407]$ and we get the following diagram:

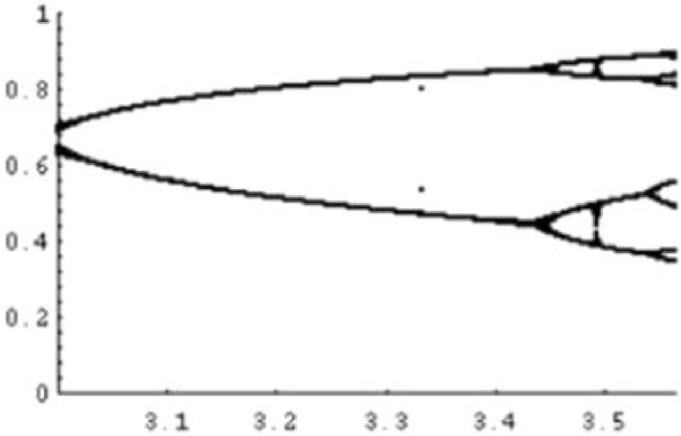


Figure 15.28: **Bifurcations in the Feigenbaum diagram**

It is remarkable to observe in the figure above that there are two point that are totally stable (the both just a little bit after 3.3)!

If we take the interval $k = [3.5, 3.569891]$ we get the following Feigenbaum diagram:

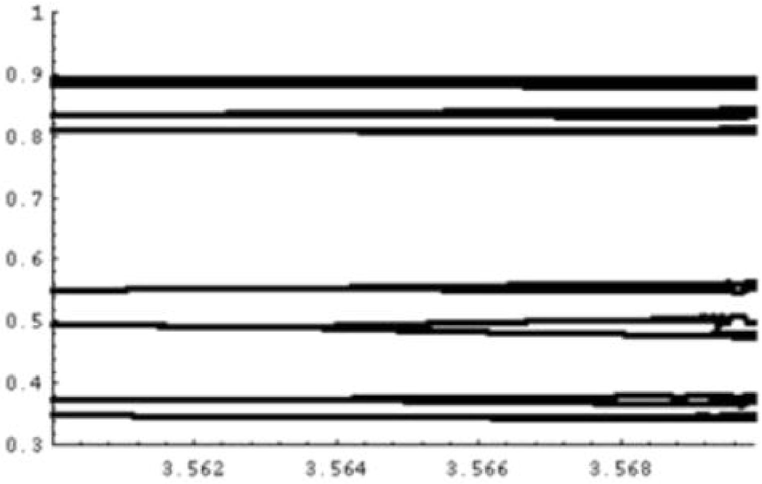


Figure 15.29: **Multiple bifurcations in the Feigenbaum diagram**

We can notice in the figure above that the bifurcation multiplies from points that are always more and more closer and on intervals always shorter. The physicist Feigenbaum has therefore proven (by trial and error) two strange results:

1. The bifurcations will multiply to infinity (!) on a range which does not exceed the point 3.5699456 of the x-axis named the "Feigenbaum point" or "gateway to chaos", because,

after this point, the system becomes chaotic. It begins to fluctuate between unpredictable values and becomes extremely sensitive to initial conditions.

2. The lengths of intervals specific to different classes of bifurcations $(2, 4, 8, 16, \ldots)$ decreases in a constant ratio of $4.6692\ldots$ named, off course, the "Feigenbaum constant".

Remark

Heinz-Otto Peitgen has observed that these bifurcations, Feigenbaum points and constant, are not only found in the case of the logistic function studied by Robert May (see below), but also in many physical phenomena such as hydrodynamics, electronics, LASERs or acoustics.

Finally, we can notice a curious phenomenon about these stable windows! If we examine more closely the chaotic area between the Feigenbaum point and $k = 4$, we see that there are areas where the logistic function starts to oscillate between a finite number of bifurcations before plunging into chaos again:

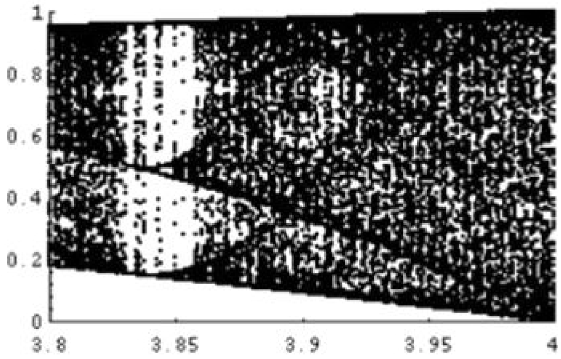


Figure 15.30: Viewing the background changes and the stabilization windows

We can notice that in this windows there are bifurcations that look like the Feigenbaum diagram itself if we zoom on it.

We could also take the same calculations between $k = 3.848$ and 3.850 and we would have found we find that we fall back on the same figure, and if we zoom on it to infinity, we always found the same figure ever and ever (!) in an auto-similarity process:

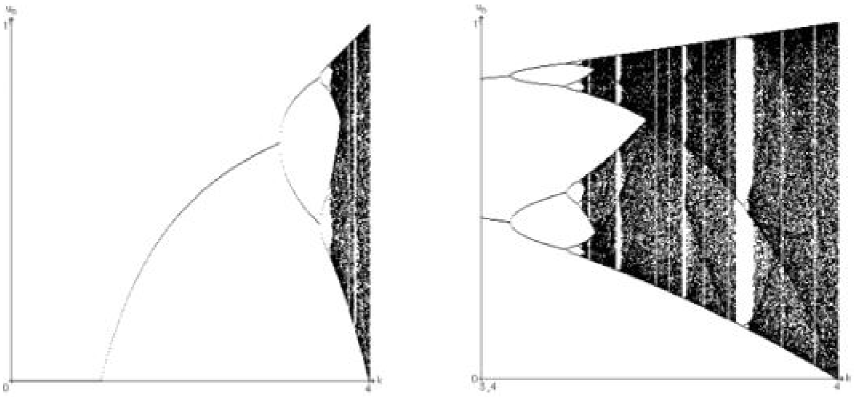


Figure 15.31: Auto-similarity of the Feigenbaum diagram

We have therefore just discovered a fractal figure (see the section Fractals page 7038 for more details): named also "strange attractor of the logistics function" also named "bifurcation diagram of the logistic equation" or simply "Feigenbaum diagram".

We will see (without proof) in the section Fractals that the Feigenbaum diagram above is closely related to the Mandelbrot fractal (without proof as we never used this result for engineering applications).

Thanks to the mathematics of Chaos Theory applied to population dynamics, ecology received a decisive stimulation. Until the early 11960s (holocene calendar), the debate on population dynamics opposed the proponents of a deterministic theory, seeing regular changes in populations undergoing exceptionally sudden changes, to those believing these developments as purely random. But many facts remain poorly explained. Especially cyclic explosions of some populations and their strange periodicity did not fall into either of the two explanations. By showing that deterministic models can give rise to random behaviour, Robert May, reconciles those views from a deeper theory. What appears to a given level of apprehension as a general instability can appear to another level as a stable chaos. A chaotic mathematical situation can be stable ecologically.

If, because of the increasing power of computers, Chaos Theory has led to major advances in different disciplines, it raises a major debate, that of determinism. How does science can predict the future? For some, the results thanks to Chaos Theory prove the importance of initial conditions. For these people, deterministic equations have limited scope and the future remains unpredictable. For others however, the results show that we can find order and laws in what may seem chaotic. These laws are simply more complex.

Here is a non-exhaustive list of chaotic natural phenomena:

- Atmospheric dynamics: Winds, storms, gradient of temperature and pressure.
- Evolutionary processes: Evolutionary processes are often nonlinear, meaning that small changes in conditions can lead to disproportionately large effects.
- Volcanic storms: Lightning occurs in plumes of volcanic ash.
- Supercell storms: Massive thunderstorms that can produce tornadoes.
- Catatumbo storms: A region in Venezuela where lightning occurs almost continuously.
- Rogue waves: Extremely large and unpredictable ocean waves.
- Earthquakes: Sudden shaking of the Earth's crust.
- Volcanic eruptions: Explosions of lava, ash, and gas from volcanoes.
- Tsunamis: Giant waves caused by underwater earthquakes.
- Landslides: Rapid movements of rock and earth down slopes.
- Tropical cyclones: Powerful storms with high winds and torrential rain.

- Solar storms: Generates high energy particle winds that can have significant effects on living things.
- Avalanches: Rapid descents of snow and ice down mountainsides.
- Chemical reactions: Belousov-Zhabotinsky or Briggs-Rauscher oscillating reaction
- Social Dynamics: Populations behaviours
- ...

It is an open and lively debate that goes far beyond the scientific community.

15.1.2.4 Malthusian Growth Law

We have already seen some deterministic and chaotic models using growth rates for the simulation. Now introduce another type of model using the number of children, women and their respective fertility.

For this model, we state the following hypothesis (assumptions):

H1. At the beginning, N people where $N/2$ are females.

H2. The growing rate r is supposed to be constant because of a fecundity rate f that is also constant.

H3. The population is growing.

So we have for data: the population at time t denoted $N(t)$, the female population denoted $N_F(t)$ and $N_E(t)$ the number of children.

With the following relations:

$$\begin{aligned} N_F(t) &= \frac{N(t)}{2} \\ N_E(t) &= fN_F(t) \end{aligned} \qquad (15.1.52)$$

and therefore:

$$N(t+1)=N(t)+N_E(t)=N(t)+fN_F(t)=N(t)+f\frac{N(t)}{2}=N(t)\left(1+\frac{f}{2}\right)=rN(t) \qquad (15.1.53)$$

then the growth rate is:

$$r=1+\frac{f}{2} \qquad (15.1.54)$$

Finally we get the Malthus law:

$$\boxed{N(t)=r^tN(0)} \qquad (15.1.55)$$

this model is therefore continuous, takes into account the fecundity but diverge.

15.1.2.5 Leslie model

The Leslie model is a bit more advanced than the others deterministic models but just as empirical (it is possible, as in any field of science to construct valid theoretical models always more complete and complex).

Besides the fertility and mortality rates, this model can take into account the ages groups of the population and some of their properties in relation to the two aforementioned factors (fertility, mortality). The disadvantage of this model is, however, the too many parameters to be determined for all age groups...

The system is based on the cutting of the population in age groups such as for example:

- $N_1(t)$: Is the number of people of age 1
- $N_2(t)$: Is the number of people of age 2
- ...
- $N_i(t)$: Is the number of individuals of age i.
- ...
- $N_T(t)$: Is the number of individuals of age T.

Then the idea is that the evolution of an age group depends on the other age groups. For example, births are given by the reproduction rate r summed over all age groups (of course for some of them the rate is zero...) such as:

$$N_1(t+1)=\sum_{i=2}^{T} r_i N_i(t) \tag{15.1.56}$$

It is customary and reasonable for humans being to accept that normally only age groups for which $i\geq 11$ and $T=55$ are taken into account regarding to the reproduction which implies that normally the r_i are non-zero only for this interval.

Then, aging and mortality m will be taken into account by the relations (we take again all the age groups of the entire population because obviously there is not only the individuals between 11 and 55 that die ...):

$$N_i(t+1)=N_{i-1}(t)-m_{i-1}N_{i-1}(t) \qquad \forall i\in\mathbb{N}, i=2\ldots T \tag{15.1.57}$$

It is relatively easy to see that these equations can be put in a matrix form (see section Linear Algebra page 1924) as follows:

$$\begin{pmatrix} N_1(t+1)\\ N_2(t+1)\\ \vdots\\ N_i(t+1)\\ \vdots\\ N_T(t+1)\end{pmatrix}=\begin{bmatrix} 0 & r_2\ldots & r_i & \ldots & r_T & \\ 1-m_1 & 0 & 0 & 0 & 0 & 0\\ 0 & 1-m_2 & 0 & 0 & 0 & 0\\ \vdots & 0 & \ddots & 0 & 0 & \vdots\\ \vdots & 0 & 0 & 1-m_i & 0 & \vdots\\ 0 & 0 & 0 & 0 & 1-m_{T-1} & 0\end{bmatrix}\begin{pmatrix} N_1(t)\\ N_2(t)\\ \vdots\\ N_i(t)\\ \vdots\\ N_T(t)\end{pmatrix} \tag{15.1.58}$$

where the matrix containing the coefficients of mortality and reproduction is named "Leslie matrix".

More compactly it is written:

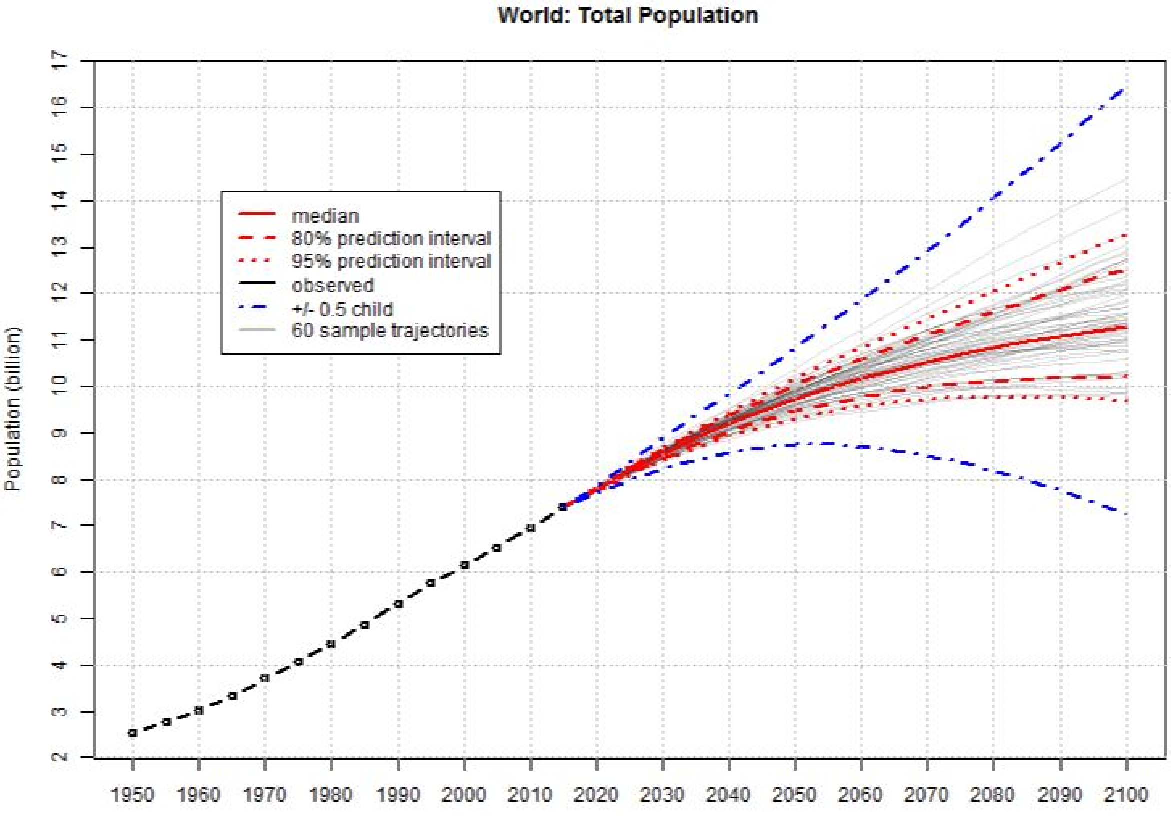
$$N(t+1)=LN(t) \tag{15.1.59}$$

or when starting for the initial population:

$$N(t)=LN(0) \tag{15.1.60}$$

It is possible to make interesting analyses on this model with respect to the age of having children and the consequences relating thereto. This model seems to be relatively widely used in marine biology.

We will stop here about population models. However in a few decades or centuries it will be interesting for the reader to see how accurate was the following projection United Nations:



Source: United Nations, Department of Economic and Social Affairs, Population Division (2017). *World Population Prospects: The 2017 Revision. http://esa.un.org/unpd/wpp/*

15.1.2.6 Epidemic models for spread of disease

The "theory of epidemics" provides many intuitive systems of differential equations or with partial derivatives in discrete or continuous time having analytical deterministic solutions in which intervene contamination phenomena, diffusion phenomena, etc.

We will see below a few naive deterministic models to simulate epidemics. They are "naive" because they don't take into considerations the behaviours of the individuals (their movements, their social interactions, their respect of governmental rules, the answer of the medical system, the topology of the geographical area, the pyramid of ages, newborns and deaths, etc.) and are completely deterministic. Models used in real life in the 121st century (holocene calendar) use models based on "Discrete Event Analysis" also named "Discrete Event Simulation" (with Monte Carlo simulations and Bayesian methods including forecasting technique like VAR and GARCH) that runs only on supercomputers and include a few hundred to a few thousands of stochastic interacting or non-interacting variables.

Any idea to cure the plague?
Pray to god
Kill the infected
SCIENCE

15.1.2.6.1 Exponential Model

Let us consider a model where out of N people, there are those infected I and those susceptible S with the assumption that eventually all susceptible can contracts the disease, we have:

$$\begin{aligned}\frac{\Delta S(t)}{\Delta t}&=rI(t)\\ S(t)&=N(t)-I(t)\end{aligned} \tag{15.1.61}$$

That model is named obviously the "exponential disease model" as the differential equation:

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t}=rI(t) \tag{15.1.62}$$

concerning the infected constitutes an exponential growth:

$$I(t)=I_0e^{rt} \tag{15.1.63}$$

This model is very flawed because in real life, not all contacted by the infected contracts the disease, and as the community runs out of people to get infected, the rate at which the disease spreads slows:

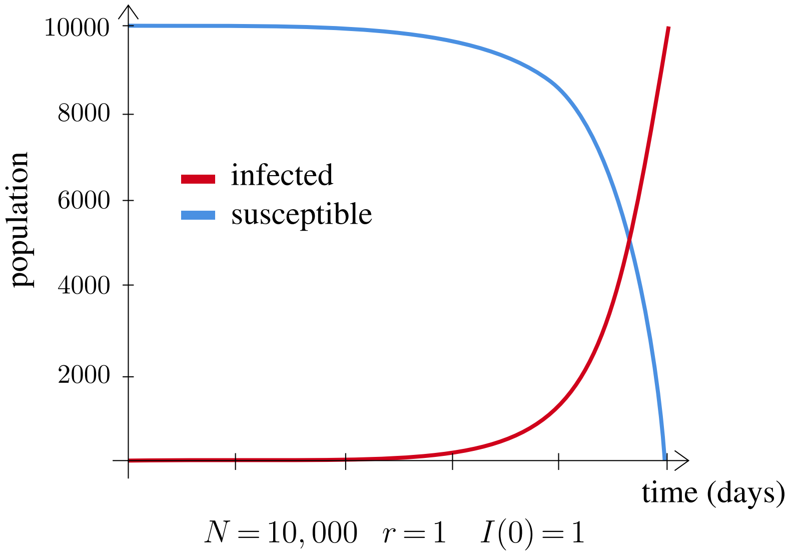


Figure 15.32: **Exponential epidemic model**

15.1.2.6.2 SI Model (Susceptible-Infected)

With what we have from the exponential model, let us introduce a new parameter β: the probability for the susceptible to contract the disease upon contact.

Then we put by construction:

$$\frac{\Delta S(t)}{\Delta t}+\frac{\Delta I(t)}{\Delta t}=0 \tag{15.1.64}$$

and:

$$\begin{aligned}\frac{\Delta I(t)}{\Delta t}&=r\beta S(t)I(t)\\ \frac{\Delta S(t)}{\Delta t}&=-r\beta S(t)I(t)\end{aligned}\qquad(15.1.65)$$

Adding the equations, we get:

$$\frac{\Delta I(t)}{\Delta t}+\frac{\Delta S(t)}{\Delta t}=r\beta S(t)I(t)-r\beta S(t)I(t)=0\qquad(15.1.66)$$

which is no surprise since $I(t)+S(t)=N$ is constant. Therefore, we only consider the first equation. After substituting $S(t)=N-I(t)$, we get a single differential equation about the infected population $I(t)$:

$$\frac{\Delta I(t)}{\Delta t}=r\beta I(t)(N-I(t))\qquad(15.1.67)$$

Hence at the limit and after rearranging:

$$\frac{\mathrm{d}I(t)}{\mathrm{d}t}=r\beta I(t)N\left(1-\frac{I(t)}{N}\right)\qquad(15.1.68)$$

We have seen during our study of the logistic model that the solution to this differential equation is the logistic function:

$$I(t)=\frac{NI_0}{I_0+(N-I_0)e^{-r\beta t}}\qquad(15.1.69)$$

That is:

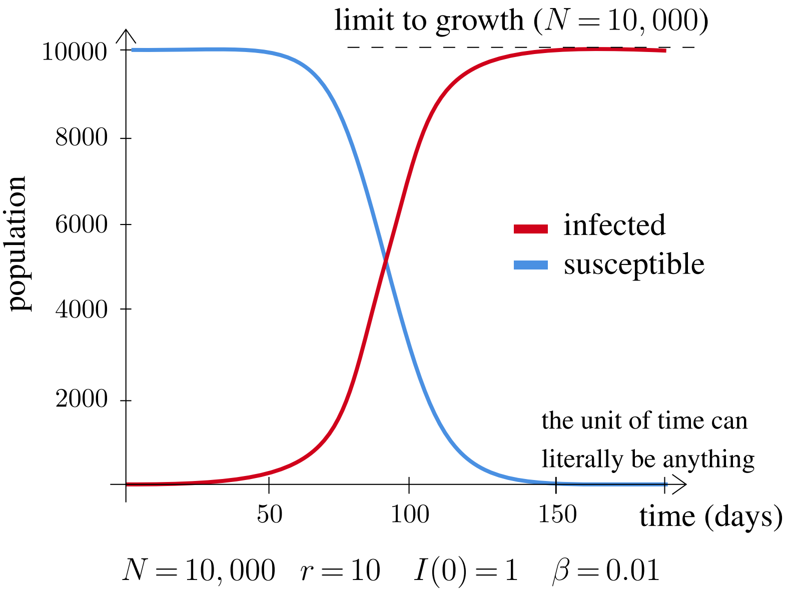


Figure 15.33: **SI epidemic model**

The number of infected people increases monotonically but stops as it approach the total population N. At $I=N/2$, we have:

$$I''(t)|_{I=\frac{N}{2}}=0\qquad(15.1.70)$$

so I grows the fastest. After this point, the growth rate of I declines.

15.1.2.6.3 SIS Model (Susceptible-Infected-Susceptible)

The idea of the SIS model is that susceptible individuals who were infected may become susceptible individuals again. That is:

$$S \underset{\gamma}{\overset{r\beta}{\rightleftharpoons}} I \tag{15.1.71}$$

Then we introduce a new parameter γ that takes care of the fact that individuals who recover from the disease may not gain full immunity.

The set of equations is very similar to the SI model but just with a supplementary term:

$$\begin{aligned}\frac{\Delta I(t)}{\Delta t} &= r\beta S(t)I(t)-\gamma I(t)\\ \frac{\Delta S(t)}{\Delta t} &= -r\beta S(t)I(t)+\gamma I(t)\end{aligned} \tag{15.1.72}$$

Adding the equations, we also get:

$$\frac{\Delta I(t)}{\Delta t}+\frac{\Delta S(t)}{\Delta t}=(r\beta S(t)I(t)-\gamma I)+(-r\beta S(t)I(t)+\gamma I)=0 \tag{15.1.73}$$

which again is no surprise since $I(t)+S(t)=N$ is constant. Therefore, we only consider the first equation. After substituting $S(t)=N-I(t)$, we get a single differential equation about the infected population $I(t)$:

$$\frac{\Delta I(t)}{\Delta t}=r\beta I(t)(N-I(t))-\gamma I(t) \tag{15.1.74}$$

Rearranging and taking the limit:

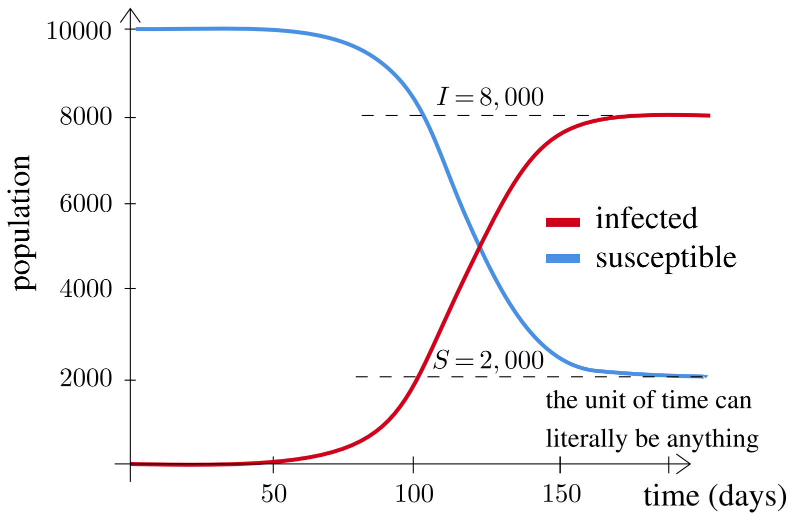
$$\frac{\mathrm{d}I(t)}{\mathrm{d}t}=(r\beta N-\gamma)I(t)-r\beta I^2(t) \tag{15.1.75}$$

If we put $A:=r\beta N-\gamma$ the above differential equation can be rewritten as:

$$\frac{\mathrm{d}I(t)}{\mathrm{d}t}=(A-r\beta I(t))I(t)=AI(t)\left(1-\frac{r\beta I(t)}{A}\right) \tag{15.1.76}$$

And we recognize here a differential equation of the same shape as the one from the previous model! So the solution is also a logistic function given by:

$$I(t)=\frac{AI_0}{I_0+(A-I_0)e^{-r\beta t}}=\frac{(r\beta N-\gamma)I_0}{I_0+(r\beta N-\gamma-I_0)e^{-r\beta t}} \tag{15.1.77}$$



$$N = 10,000 \quad r = 10 \quad I(0) = 1 \quad \beta = 0.01 \quad \gamma = 0.02$$

Figure 15.34: **SIS epidemic model**

15.1.2.6.4 SIR Model (Susceptible-Infected-Recovered)

Here we will take the example of the spread of rabies in the fox population, and present a fairly simple model where we consider that individuals (foxes) can be in only three different states:

1. There are healthy foxes in quantity S (susceptible)
2. There are infected foxes in quantity I
3. There are death foxes in quantity R

This assumption of classification into three categories make that this model of spread of epidemics is often named the "SIR model".

First, we assume that the overall (closed[1]) population is:

$$S + I + R = 1 \qquad (15.1.78)$$

so that S, I, and R are proportions (and obviously are function of time). We further assume that the overall population does not change, that is:

$$\frac{\Delta(S+I+R)}{\Delta t} = 1 \Leftrightarrow \frac{\Delta I}{\Delta t} = -\frac{\Delta S}{\Delta t} - \frac{\Delta R}{\Delta t} \qquad (15.1.79)$$

For more optimistic people, we can also use the model that assumes that some individuals are immune and can therefore not be infected. Finally, with three different states, we can link the three proper equation of developments such that:

---

[1] *Closed* means that there is no immigration or emigration!

$$\begin{aligned}\frac{\Delta S}{\Delta t}&=S'=-r\beta IS\\ \frac{\Delta R}{\Delta t}&=R'=\gamma I\\ \frac{\Delta I}{\Delta t}&=I'=-\frac{\Delta S}{\Delta t}-\frac{\Delta R}{\Delta t}=r\beta IS-\gamma I\end{aligned}\tag{15.1.80}$$

where $r>0$ is the rate of infection, $0<\beta\leq 1$ the probability of infection and $\gamma>0$ is the recovery rate (percentage of people who recover from the disease but don't gain full immunity) and with:

$$\begin{aligned}S_{t+1}&=S+S'\Delta t\\ I_{t+1}&=I+I'\Delta t\\ R_{t+1}&=R+R'\Delta t\end{aligned}\tag{15.1.81}$$

where the S,I,R are given in $\%$ of the whole population. This is the simplest of the SIR models and named the "Kermack-McKendrick model".

In this model, the first equation corresponds to the phenomenon of contamination: when coexist healthy individuals and of infected individuals, a number of healthy individuals will be infected. It is also natural to consider this term as proportional to the product IS. Indeed, the quantity of microbes in the defined medium (thus the factor r for a healthy individual to be infected in the presence of an infected individual) is proportional to I according to the factor. We then have to multiply this probability by the number of healthy individuals, that is to say, by S. The negative sign is present to indicate the reduction of the healthy population.

Regarding the variations of I, the first term corresponds to the infected individuals (which increase I). The second term in $-\gamma I$, consists of individuals who die.

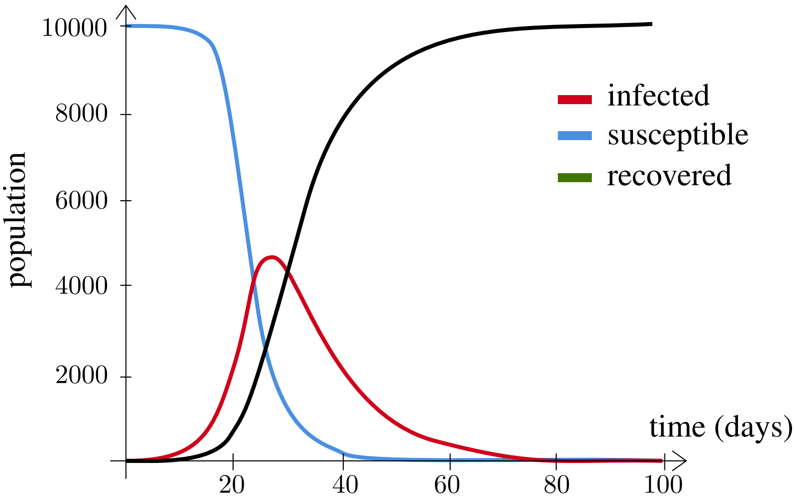
Observe that since we assumed that the overall population does not change since:

$$R(t)=1-S(t)-I(t)\tag{15.1.82}$$

then the system is therefore fully characterized by:

$$\boxed{\begin{aligned}\frac{\Delta S}{\Delta t}&=S'=-r\beta IS\\ \frac{\Delta I}{\Delta t}&=I'=r\beta IS-\gamma I\end{aligned}}\tag{15.1.83}$$

The plot below shows a calculation of the evolution of the populations S, I, R as a function of time (one second is considered as a loop of the algorithm). According to the values of parameters γ, β and r the behaviours are completely different: if γ is too high, for example, infected individuals die almost immediately and do not have time to infect many others (...):



$$N = 10,000 \quad r = 10 \quad I(0) = 1 \quad \beta = 0.01 \quad \gamma = 0.02$$

Figure 15.35: **SIR epidemic model**

According to this model, during the beginning phase of a pandemic, there would be a surge of confirmed cases. The as time goes on, a fraction of the individuals gradually gain immunity to the disease, thus the number of infected people goes down.

Definition (#15.1.628): A "pandemic" is an epidemic of an infectious disease that has spread across a large region, for instance multiple continents or worldwide, affecting a substantial number of individuals. A widespread endemic disease with a stable number of infected individuals is not a pandemic.

Let us show now that we can get an analytic solution for $I_{\max}$.

Dividing the two previous relation we get:

$$\frac{\frac{\mathrm{d}S}{\mathrm{d}t}}{\frac{\mathrm{d}I}{\mathrm{d}t}} = \frac{\mathrm{d}S}{\mathrm{d}I} = \frac{-r\beta IS}{r\beta IS - \gamma I} \tag{15.1.84}$$

We can rewrite that latter relation as:

$$\frac{r\beta S - \gamma}{r\beta S}\mathrm{d}S = -\mathrm{d}I \tag{15.1.85}$$

That we can obviously integrate both sides:

$$\int \frac{r\beta S - \gamma}{r\beta S}\mathrm{d}S = -\int \mathrm{d}I \tag{15.1.86}$$

That lead us to:

$$S - \frac{\gamma}{r\beta}\log(S) = -I - c^{te} \tag{15.1.87}$$

After rearranging (we put the negative sign in the constant):

$$I + S - \frac{\gamma}{r\beta}\log(S) = c^{te} \tag{15.1.88}$$

In other words, for every $t \geq 0$:

$$I(t)+S(t)-\frac{\gamma}{r\beta}\log(S(t))=I(0)+S(0)-\frac{\gamma}{r\beta}\log(S(0)) \tag{15.1.89}$$

I_{max} occurs when $\mathrm{d}I/\mathrm{d}t=0$, i.e.:

$$\frac{\mathrm{d}I}{\mathrm{d}t}=r\beta IS-\gamma I=0 \tag{15.1.90}$$

Therefore this occurs when $S=\gamma/(r\beta)$. Applying in our previous equality, we get:

$$I_{\text{max}}+\frac{\gamma}{r\beta}-\frac{\gamma}{r\beta}\log\left(\frac{\gamma}{r\beta}\right)=I(0)+S(0)-\frac{\gamma}{r\beta}\log(S(0)) \tag{15.1.91}$$

or:

$$I_{\text{max}}=I(0)+S(0)-\frac{\gamma}{r\beta}\log(S(0))-\frac{\gamma}{r\beta}+\frac{\gamma}{r\beta}\log\left(\frac{\gamma}{r\beta}\right) \tag{15.1.92}$$

Condensed it leads to:

$$\boxed{I_{\text{max}}=I_0+S_0+\frac{\gamma}{r\beta}\left[\log\left(\frac{\gamma}{r\beta S_0}\right)-1\right]} \tag{15.1.93}$$

The implementation algorithm looks like this in pseudocode (for a more formal approach see our MATLAB® or R companion book).

In pseudocode (non-unique and not optimized):

Algorithm 19: SIR Model pseudocode algorithm

```
Data: $r$ $\beta$,$\gamma$, $\max_t$, $\Delta_t$
Result: $\bar{x}$
initialization;
for $i=1$ to $E(\max_t/\Delta_t)$ do
    $\mathrm{d}S=-r\beta S(i)I(i)$;
    $\mathrm{d}I=r\beta S(i)I(i)-\gamma I(i)$;
    $\mathrm{d}R=\gamma I(i)$;
    $\mathrm{d}R=\gamma I(i)$;
    $S(i+1)=S(i)+\mathrm{dtd}S$;
    $I(i+1)=I(i)+\mathrm{dtd}I$;
    $R(i+1)=R(i)+\mathrm{dtd}R$;
plot $(I,t),(S,t),(R,t)$;
```

A key parameter in epidemiology is the "basic reproductive ratio R_0". It is defined as the average number of secondary cases transmitted by a single infected individual that is placed into a fully susceptible population. In other words, R_0 tells us about the initial rate of spread of the disease. Hence, if $R_0>1$, there will be an epidemic, and if $R_0<1$, the introduced infected will recover (or die) without being able to replace themselves by new infections. In the special case of the model above, it is pretty easy to derive R_0. The disease-free state corresponds to:

$$S=N \qquad I=0 \qquad R=0 \tag{15.1.94}$$

where don't work in percentages any more but rather in total population size N such that:

$$S+I+R=N \tag{15.1.95}$$

If one infected individual appears in the population, there will be an epidemic if and only if:

$$\frac{\Delta I}{\Delta t}>0 \tag{15.1.96}$$

By replacing S with N in the third relation of:

$$\begin{aligned}\frac{\Delta S}{\Delta t}&=S'=-r\beta IS\\ \frac{\Delta R}{\Delta t}&=R'=\gamma I\\ \frac{\Delta I}{\Delta t}&=I'=-\frac{\Delta S}{\Delta t}-\frac{\Delta R}{\Delta t}=r\beta IS-\gamma I\end{aligned} \tag{15.1.97}$$

This leads to if we put $\beta=1$:

$$\frac{rN}{\gamma}>1 \tag{15.1.98}$$

That is:

$$R_0=\frac{rN}{\gamma} \tag{15.1.99}$$

To avoid any epidemic of a disease whose basic reproduction rate is R_0, it is therefore necessary that the proportion of susceptible is less than $1/R_0$, in other words that the proportion of immune is greater than minimum vaccination cover rate given by (this result has nothing to do specifically with the SIR model, it's just applied common sense!):

$$\text{MVCR}=\frac{R_0-1}{R_0} \tag{15.1.100}$$

Thus, if $R_0=2$ (seasonal influenza reproduction rate), at least half of the population must be vaccinated to prevent or slowdown epidemics. On the other hand, if $R_0=15$ to 20 (measles reproduction rate), herd immunity is only guaranteed if at least 93 to 95% of the population is vaccinated!

We can of course come to many variations of this model depending on the assumptions we make about the disease under study. It will often be useful to define more populations categories. Thus, for S.T.D. (Sexually Transmitted Diseases), we will differentiate the populations according to sex (we should also take the age into account), which double the number of possibilities. To define the underlying differential system, it becomes better sometimes to model the process by a diagram.

Definition (#15.1.629): A "vaccine" is a preparation which must be ingested or injected once or multiple times in a single lifetime that provides a statistically significant active acquired immunity to a particular infectious disease in the purpose to avoid completely or partially its effects[2]. A vaccine must have a statistical interesting benefice/risk ratio (better than a placebo or standard of care in a control group). The purpose of a vaccine is not to eliminate the transmission of its targeted virus but that may be an optional beneficial side effect.

---

[2] At the difference of a "treatment" that is used only once the infectious disease is already acquired by the individual.

Definition (#15.1.630): In medicine, a "treatment", also named "medical treatment", "therapeutic treatment", "therapy" or more generally "therapeutic", is a set of measures applied by a health professional (or therapist) to a person vis-à-vis a disease in order to help that person to recover or to relieve from symptoms of the disease with a given likelihood and with an efficiency statistically better than a placebo or standard of care.

To study naively an infection like AIDS (Acquired Immunodeficiency Syndrome), we can make such a model in which we consider four population groups: healthy individuals X, infected individuals Y, the healthy carriers Z and finally the sick individuals A. In the diagram below, the arrows between the different states represent possible transitions for a given individual with the corresponding probabilities:

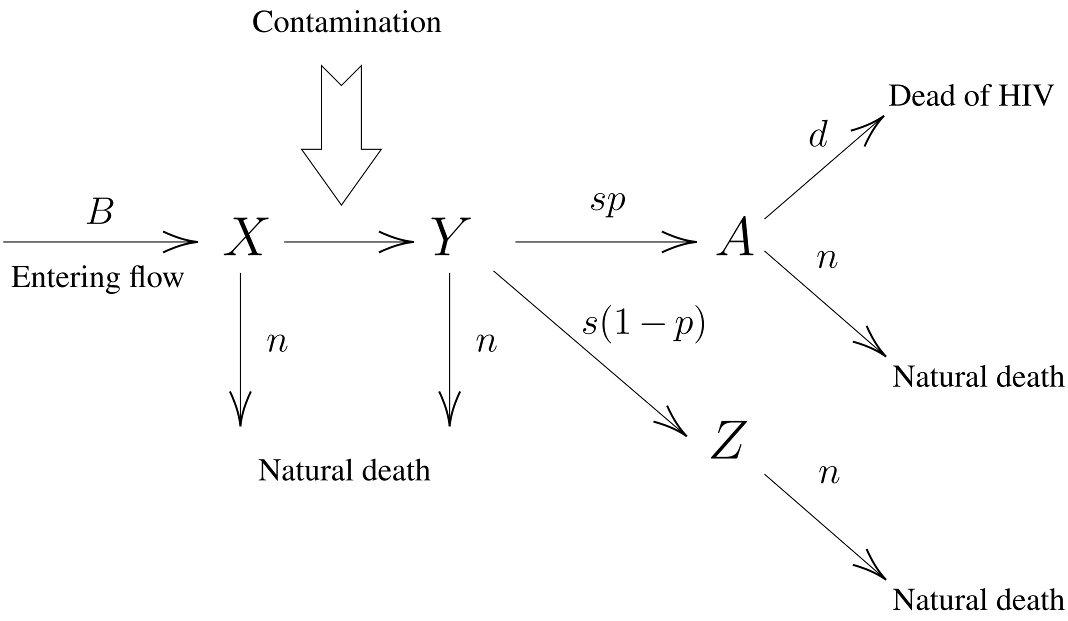


Figure 15.36: **Markov chain of deadly disease**

Indeed, from the perspective of the study of AIDS, a given individual, whatever his condition, may die of something else than AIDS, this is what we put into the category "natural death", We consider that there is a fixed mortality rate, denoted n, for all populations considered (healthy, infected, sick or healthy carriers). On the other hand, a healthy individual X may also be contaminated, in which case it becomes an infected individual Y. An infected individual can change into to two possible stages: sick A or healthy carrier Z.

The number s is the speed of evolution of the disease: bigger is this rate, the more the infected individuals evolve into one of the next stages. The number p, which value is between 0 and 1, represents the probability that this change is done into the direction of the sick state. Of course, if p decreases, the amount of healthy carriers, immunized against the disease increases. The disease is then globally less dangerous. Finally, a patient may die of his disease, with a mortality rate d considered as fixed, while a healthy carrier, immunized, can not change anymore of the point of view of this disease. All these assumptions lead to put a system of equations of the form:

$$\begin{aligned}X' &= B - nX - \lambda_c X\\ Y' &= \lambda_c X - (n+s)Y\\ A' &= spY - (d+n)A\\ Z' &= s(1-p)Y - nZ\end{aligned} \tag{15.1.101}$$

where:

$$\lambda_c = \frac{\beta Y}{X+Y+Z+A} \tag{15.1.102}$$

Remark

The SIR model assumes that once infected people are immune to the disease forever, and so any disease occurs only once and then never comes back. More interesting dynamics occur when we allow for the reinfection of recovered people; we can then ask, for example, under what circumstances the disease becomes established in the population. The SIRS model extends the SIR model, allowing the recovered population to become susceptible again (hence the extra 'S'). It assumes that the susceptible population increases proportional to the recovered population such that:

$$\begin{aligned}\frac{\Delta S}{\Delta t} &= S' = -r\beta IS + \mu R\\ \frac{\Delta R}{\Delta t} &= R' = \gamma I - \mu R\\ \frac{\Delta I}{\Delta t} &= I' = -\frac{\Delta S}{\Delta t} - \frac{\Delta R}{\Delta t} = r\beta IS - \gamma I\end{aligned} \tag{15.1.103}$$

where, since we added μR to the change in the proportion of susceptible people, we had to subtract μR from the change in the proportion of recovered people.

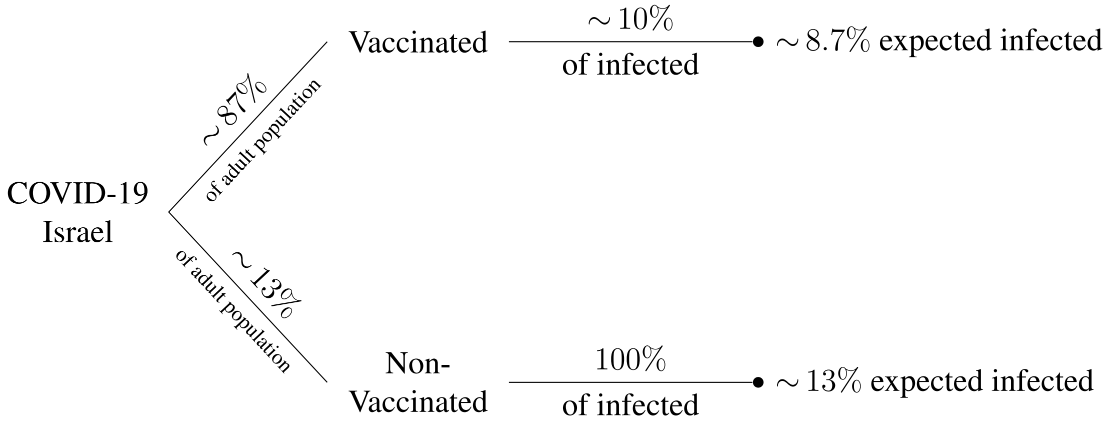
Now let us quote the CNews article from the 12021-06-21 (holocene calendar) about the COVID-19 written by Clémence Barbier (Master in Journalism, Bachelor in History, Catholic and obviously ignorant in basics scientific calculations...) to prove how mass-media articles titles may sometimes be misleading (naive human error[3]? on purpose click-bait title? catholic bigot bias?):

"Despite significant vaccination coverage in Israel, 40% of new infections are from vaccinated people"

It seems that this journalist, above any reasonable doubt, added the word "despite" from the original quote from Gabi Barbash (director general at Tel Aviv Sourasky Medical Center.). This is obviously dumb (as shown in the calculation below) because this is exactly what we expected from a 90% efficient vaccine in the adult Israelite population. Her article was quite massively shared on social networks by scientifically illiterate anti-vax people afterwards (especially by other bigots...).

---

[3] Notice that the day she write that paper this young journalist had only one year of experience and seems to not have followed any training on scientific communication nor be informed about the process of being proofread by an internal scientific consultant at CNews.



$$\frac{8.7}{13+8.7}\cong 40\%$$

So as proven thanks to the decision tree above, 40% is exactly what is expected and is a very good result. So adding the word "despite" was a double mistake!!! Notice that one week after the title of the article still wasn't corrected by the redaction team of CNews...

Notice that a minority of people don't understand - because of a lack of Statistical education or just because of a difficulty to apprehend the concept of proportions - why when a huge majority of people are vaccinated then the majority of people who are sick will be... vaccinated people. For example if you take the extreme case where 100% of the population is vaccinated than obviously you will have at the emergencies 100% of vaccinated people... But in absolute value that will still be less people than if no one was vaccinated.

Let us denote as always the efficiency of a vaccine by it's Odds Ratio O.R. (see section Numerical Methods page 6496). Then $1-\text{O.R.}$ is obviously its ineffectiveness. Denoting by HR the hospitalisation rate of people who catches the illness, P_{V} the population of vaccinated people and P_{NV} the population of non-vaccinated people then the number of hospitalized are respectively for each population:

$$P_{\text{V}}\cdot\text{HR}\cdot(1-\text{O.R.})\qquad\text{and}\qquad P_{\text{NV}}\cdot\text{HR}\tag{15.1.104}$$

Equating to get the threshold of ratio of population at which the number of hospitalized vaccinated people is equal to that of non-vaccinated we get immediately after rearranging:

$$\frac{P_{\text{V}}}{P_{\text{NV}}}=\frac{1}{1-\text{O.R.}}\tag{15.1.105}$$

So for example with a vaccine having an efficiency of $\text{O.R.}=95\%$. We will get the same number of hospitalized vaccinated and non-vaccinated people when the ratio in the whole population is of $1/(1-95\%)=20$ vaccinated for 1 vaccinated. If we want to transform this value into a whole population limit ratio (WPLR) then we just calculate:

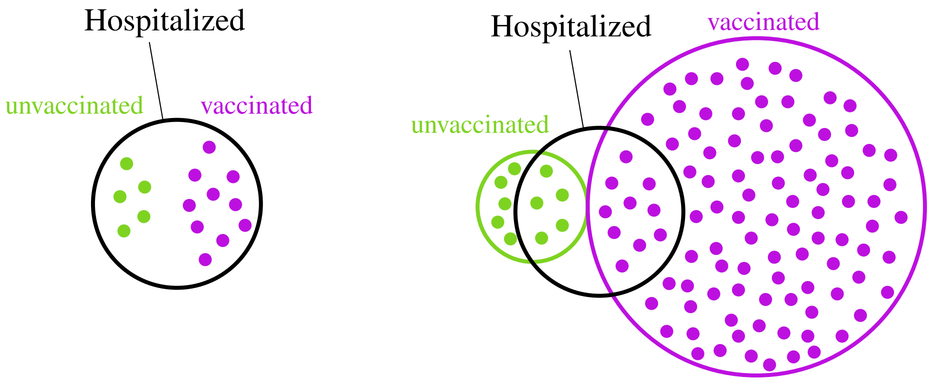
$$\text{WPLR}=\frac{\dfrac{P_{\text{V}}}{P_{\text{NV}}}}{\dfrac{P_{\text{V}}}{P_{\text{NV}}}+1}=\frac{\dfrac{1}{1\text{O.R.}}}{\dfrac{1}{1-\text{O.R.}}+1}=\frac{1}{2-\text{O.R.}} \qquad (15.1.106)$$

Applying this to our example we get:

$$\text{WPLR}=\frac{20}{20+1}=\frac{1}{2-0.95}\cong 95.238\% \qquad (15.1.107)$$

That means we reach an equality of hospitalized vaccinated and non-vaccinated people when 95.238% of the whole population is vaccinated for a vaccine with 95% efficiency.

Such a poor comprehension of proportions by some uneducated people can be well illustrated by the following figure:



Remark

As always, there exist much more elaborated models like for example the SEIR model Susceptible - Exposed - Infectious - Recovered - Susceptible). In this category of models, individuals experience a long incubation duration (the "exposed" category), such that the individual is infected but not yet infectious! There is also SEIRS (Susceptible - Exposed - Infectious - Recovered - Susceptible) model, where recovered people may become susceptible again (recovery does not confer lifelong immunity). But also the SIRD model Susceptible-Infectious-Recovered-Deceased, the SIRV model Susceptible-Infectious-Recovered-Vaccinated model, the MSIR model Measles-Infectious-Recovered, etc. For more see [244].

The CORANA Virus (COVID-19) pandemia which lasts from 12019 to 12022 (holocene calendar) has taught us:

- Prayers doesn't work
- No faith healing works
- There are no miracles
- Religious people rely on Science in critical times

- Jesus, Allah & Hindu gods and others are as real as the Greek gods
- A significant percentage of scientifically illiterate people can't grasp complexity
- Our hope is critical thinking & Science education

15.1.2.6.5 Incidence rate

In epidemiology, "incidence" is a measure of the probability of occurrence of a given medical condition in a population within a specified period of time. Although sometimes loosely expressed and understood - that's why we treat it shortly here - simply as the number of new cases during some time period, it is better expressed as a proportion or a rate with a denominator.

The numerical portion of an incidence rate has a lower bound of zero but has no upper bound; it has the mathematical range for the ratio of two non-negative quantities, in this case the number of events in the numerator and the person-time in the denominator.

At first, it may seem surprising that an incidence rate can exceed the value of 1.0, which would seem to indicate that more than 100% of a population is affected. It is true that at most, only 100% of persons in a population can get a disease, but the incidence rate does not measure the proportion of a population with illness and in fact is not a proportion at all.

Recall that incidence rate is measured in units of the reciprocal of time. Among 100 people, no more than 100 deaths can occur, but those 100 deaths can occur in $10,000$ person-years, in $1,000$ person-years, in 100 person-years, or even in 1 person-year (if the 100 deaths occur after an average of 3.65 days each).

An incidence rate of 100 cases (or deaths) per 1 person-year might be expressed as:

$$100\,[\text{cases}\cdot\text{person-year}^{-1}]$$

It might also be expressed as:

$$10,000\,[\text{cases}\cdot\text{person-century}^{-1}]$$

or ($100/12=8.\bar{3}$):

$$8.\bar{3}\,[\text{cases}\cdot\text{person-month}^{-1}]$$

or ($100/52\cong 1.92$):

$$1.92\,[\text{cases}\cdot\text{person-week}^{-1}]$$

or ($100/365\cong 0.27$):

$$0.27\,[\text{cases}\cdot\text{person-day}^{-1}]$$

The numerical value of an incidence rate in itself has no interpretability because it depends on the arbitrary selection of the time unit! It is thus essential in presenting incidence rates to give the appropriate time units, either as in the examples given above or as in $8.33\,[\text{month}^{-1}]$ or $1.92\,[\text{week}^{-1}]$. Although the measure of time in the denominator of an incidence rate is often taken in terms of years, one can have units of years in the denominator regardless of whether the observations were collected over 1 year of time, over 2 week of time, or over 10 years of time.

Unfortunately, the interpretation of incidence rates as the inverse of the average "waiting time" will usually not be valid unless the incidence rate is calculated for a stationary population with no migration (no immigration or emigration) or a closed population with complete follow-up. For example, the death rate for the United States in 11977 (holocene calendar) was 0.0088 year^{-1}; in a steady state, this rate would correspond to a mean life span, or expectation of life, of 114 years. Other analyses, however, indicate that the actual expectation of life in 11977 (holocene calendar) was 73 years. The discrepancy is due to immigration and to the lack of a steady state. Note that the no-migration assumption cannot hold within specific age groups, for people are always "migrating" in and out of age groups as they age.

Remark

Incidence should not be confused with "prevalence", which is the proportion of cases in the population at a given time rather than rate of occurrence of new cases. Thus, incidence conveys information about the risk of contracting the disease, whereas prevalence indicates how widespread a disease is.

15.1.2.7 Lotka–Volterra predator–prey model

This predator-prey interaction model was proposed by Volterra after the First World War. The purpose was to explain the dynamics of populations of sardines and sharks in the Adriatic (and hop! a bit of Oceanography in this textbook...!). Especially to explain in why the quantities of sardines fished after the interruption due to the war were not as important as before (which may seem counter-intuitive) and why at the resumption of fishing the observed proportion of sharks increased.

Taking again what we saw at the beginning of this section by writing $N(t)$ the number of preys and $P(t)$ the number of predators (that is to say implicitly a growth following a power law of the number of preys, in absence of fishing):

$$\frac{\mathrm{d}N}{\mathrm{d}t}=rN(t)\Rightarrow N(t)=c_1^{te}e^{rt} \tag{15.1.108}$$

and in the absence of preys (in the case of fishing), the number of predators decrease also following a power law, we have respectively:

$$\frac{\mathrm{d}P}{\mathrm{d}t}=-mP(t)\Rightarrow O(t)=c_2^{te}e^{-mt} \tag{15.1.109}$$

At this point of the speech, we must consider two species (preys and predators) that are obviously not isolated but in interaction. To quantify the contribution of the interaction between species, we will consider only predation, by assuming that is value or intensity (of interaction) is a function of the probability of prey-predators to meet together that will be supposed proportional to the product $N \cdot P$ of the percentages of the two populations.

These meetings do not have the same effects on both species. First, of course, each prey eaten by a predator is a net gain for the population of the latter and a net loss for the first. Thus, if the effect of interactions is accepted as being proportional to $N \cdot P$, the signs of the influence of interaction differ for the two species so that our equations with interaction this time become:

$$\boxed{\begin{aligned}\frac{\mathrm{d}N}{\mathrm{d}t}&=rN-cNP\\ \frac{\mathrm{d}P}{\mathrm{d}t}&=-mP+bNP\end{aligned}} \tag{15.1.110}$$

This system of equation is named the "Lotka-Volterra model".

Before going further, let us seek for the values for which the derivatives vanish (which will give us in fact the equilibrium point of the system):

$$\begin{aligned}rN'-cN'P'&=0\\ -mP'+bN'P'&=0\end{aligned} \tag{15.1.111}$$

where N' and P' are the values that cancel the derivatives (not to be confused this time with the notation sometimes used to condense the writing of a derivative !!!).

Thus:

$$\begin{aligned}(r-cP')N'&=0\\ (bN'-m)P'&=0\end{aligned} \tag{15.1.112}$$

A trivial solution is the "extinction solution" (also sometimes named "critical solution") given by:

$$N'_e = P'_e = 0 \tag{15.1.113}$$

Otherwise, we also have as a possible solution:

$$\begin{aligned} r = CP' \rightarrow P' = \frac{r}{c} \\ bN' = m \rightarrow N' = \frac{m}{b} \end{aligned} \tag{15.1.114}$$

Now we normalize these equations by writing (so they are dimensionless):

$$\begin{aligned} x = \frac{N}{N'} = \frac{b}{m}\frac{1}{N'} := \frac{b}{m}N \rightarrow N = \frac{m}{b}x \\ y = \frac{P}{P'} = \frac{c}{r}\frac{1}{P'} := \frac{c}{r}P \rightarrow P = \frac{r}{c}y \end{aligned} \tag{15.1.115}$$

with this normalization, the model can be rewritten:

$$\begin{aligned} \frac{m}{b}\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{mr}{b}x - \frac{cmr}{xy} \\ \frac{r}{c}\frac{\mathrm{d}y}{\mathrm{d}t} = \frac{bmr}{bc}xy - \frac{mr}{c}y \end{aligned} \tag{15.1.116}$$

By rearranging the coefficients, the system can finally be written (excluding the solution of inexistence, i.e. non-existence):

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} = rx(1-y) \\ \frac{\mathrm{d}y}{\mathrm{d}t} = my(x-1) \end{aligned} \tag{15.1.117}$$

for which derivatives vanish at the point $(1,1)$.

The discreet plot of this system of equations (in which we recognize a logistic term as seen in the section of Populations Dynamics) gives us with $m = r = b = c = 1$ and initial conditions $(x_0, y_0) = (4, 1)$:

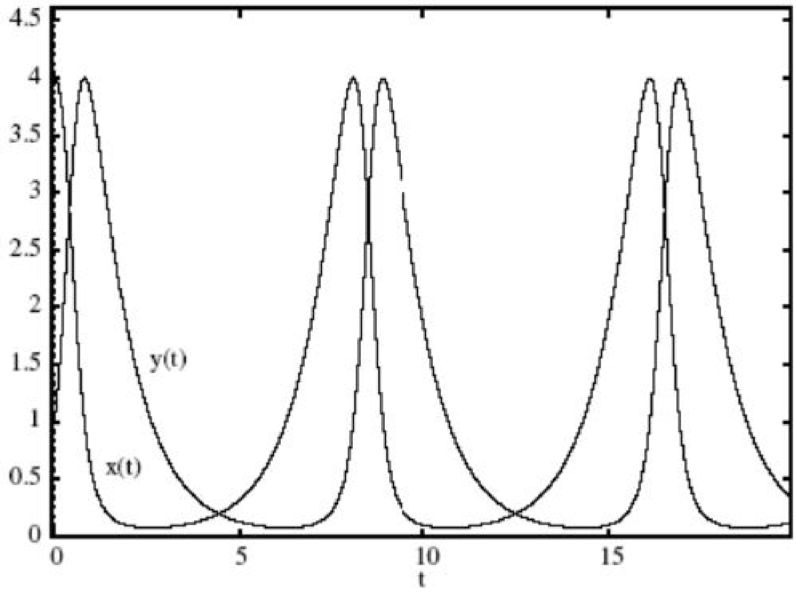


Figure 15.37: Theoretical fluctuation of population

In comparison here is a practical (real) example of prey-predators measurements (Hares-Lynx) by the Hudson Bay-Company (when we seek for data corresponding to a model we will finish soon or late to find data fitting the model...):

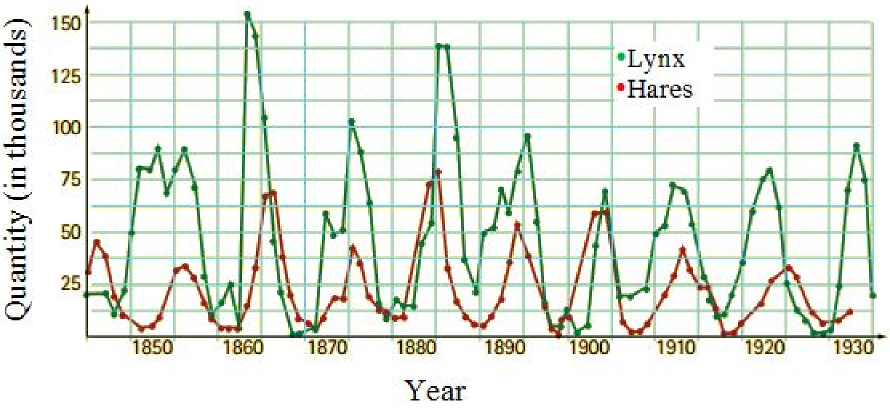


Figure 15.38: Hudson Bay-Company real Lotka-Volterra measurements

The two previous figures represent the variables x and y as a function of time. However, what can be interesting for a scientist (or an economist) is the representation of y in terms of x and vice versa. Thus, we obtain for the same initial conditions m, r, b, c and for the various initial values of (x_0, y_0) (the Predators is in ordinate and the Preys on the abscissa):

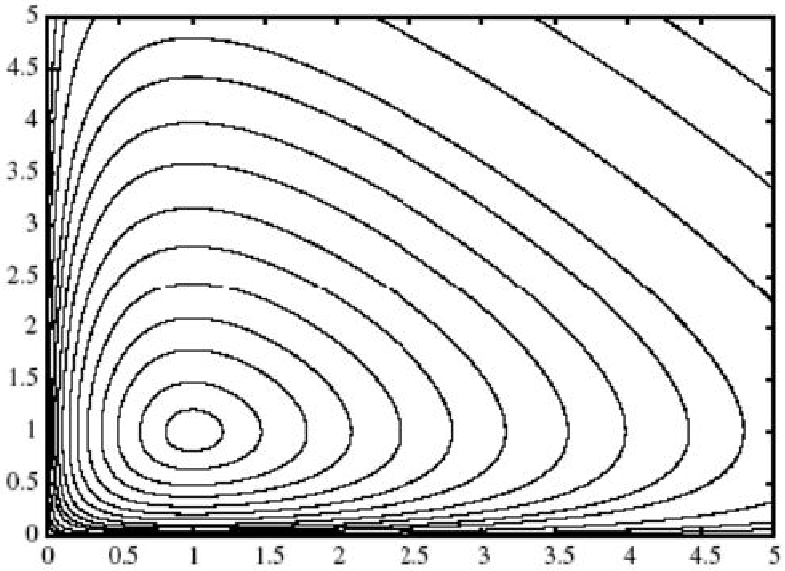


Figure 15.39: **Representation in phase space of the prey/predator cycle**

The above figure is very interesting to interpret when you follow a path in the counter-clockwise direction.

Thus we see (in the phase space representation) that for fixed initial conditions, the system is periodic and has trivially an equilibrium point at:

$$(x,y)=(1,1) \tag{15.1.118}$$

named "fixed points" that correspond to the points where:

$$\frac{\mathrm{d}x}{\mathrm{d}t}=\frac{\mathrm{d}y}{\mathrm{d}t}=0 \tag{15.1.119}$$

Finally, we have two pairs of equilibrium points (that is almost trivial, looking at the system of equations):

$$(x_1^e,y_1^e)=(0,0) \qquad (x_2^e,y_2^e)=(1,1) \tag{15.1.120}$$

The question that will naturally arise is the "real" direction of rotation (representation) of the phase plane. Thus, representing the directions using a vector field, we get the representation which is effectively counter-clockwise (more interesting representation to understand what happened after 1st World War):

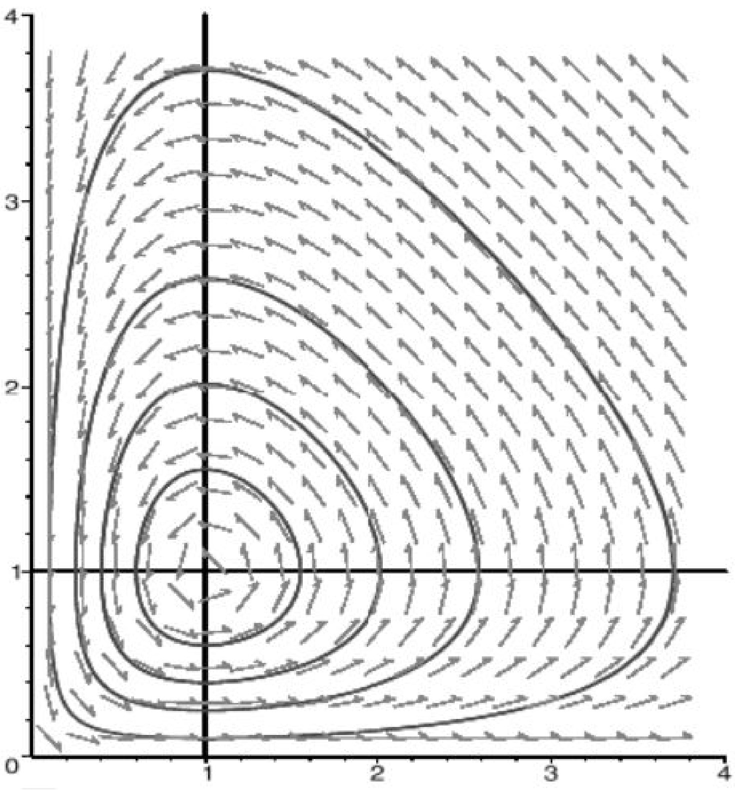


Figure 15.40: Rotation direction offer/demand phase space

To know in which direction we are going in the phase space at a given time, it suffices to know the derivative $\mathrm{d}y/\mathrm{d}x$ (or vice versa $\mathrm{d}x/\mathrm{d}y$). Therefore we have:

$$\frac{\mathrm{d}y}{\mathrm{d}x}=\frac{\mathrm{d}y}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}x}=\frac{\mathrm{d}y}{\mathrm{d}t}\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^{-1}=\frac{my(x-1)}{rx(1-y)} \tag{15.1.121}$$

That said, we see well on the phase diagram with vector field that it comes a time in the cycle of this model where the offer is very high for low demand. So the mathematical model (theoretical) clearly explains what can be a priori counter-intuitive to many humans (offer creates the demands).

However, we can (must) ask ourselves the question of what happens after a small perturbation around the equilibrium point.

So we have the following Lotka-Volterra system in equilibrium:

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}t}&=rx(1-y)\\ \frac{\mathrm{d}y}{\mathrm{d}t}&=my(x-1)\end{aligned} \tag{15.1.122}$$

By putting an infinitely small perturbation, it is will be written:

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}t}&=r(x+x^*)(1-(y+y^*))\\ \frac{\mathrm{d}y}{\mathrm{d}t}&=m(y+y^*)((x+x^*)-1)\end{aligned} \tag{15.1.123}$$

Neglecting the quadratic xy terms, we obtain:

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}t} &\cong r\left[(1-y^*)x+(1-y-y^*)x^*\right]\\ \frac{\mathrm{d}y}{\mathrm{d}t} &\cong -m\left[(1-x^*)y+(1-x-x^*)y^*\right]\end{aligned} \tag{15.1.124}$$

1. We focus first on the study near the point of extinction $(0,0)$, this is why we can neglect the quadratic terms xy but the expression remain too complicated. So, always considering we are near the point of extinction $(0,0)$, we consider a bit unfairly but cleverly (otherwise we still could not solve the problem analytically) the following approximations:

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}t} &\cong r\left[x-\underbrace{y^*x}_{=0}+x^*-\underbrace{yx^*}_{=0}-\underbrace{y^*x^*}_{=0}\right]\underset{x\gg x^*}{\cong} r(x+x^*)\cong rx\\ \frac{\mathrm{d}y}{\mathrm{d}t} &\cong -m\left[y-\underbrace{x^*y}_{=0}+y^*-\underbrace{xy^*}_{=0}-\underbrace{x^*y^*}_{=0}\right]\cong -m(y+y^*)\underset{y\gg y^*}{\cong} -my\end{aligned} \tag{15.1.125}$$

Therefore:

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}t} &\cong rx=rx(t)\\ \frac{\mathrm{d}y}{\mathrm{d}t} &\cong -my=-my(t)\end{aligned} \tag{15.1.126}$$

This shows us for the system in equilibrium, close to the inexistence point, the prays decreases exponentially (power law) as predators increases exponentially:

$$\begin{aligned}x(t)&=c_1^{te}e^{rt}\\ y(t)&=c_2^{te}e^{-mt}\end{aligned} \tag{15.1.127}$$

This shows us for the equilibrium system, near the extinction point, that the number of individuals predators decreases exponentially while the prey are increasing exponentially. This has a biological sense: when there are few individuals predators, then prey multiply and at the same time the number of prey increases, predators multiply and concentrate increasingly on their prey; respectively when there are few prey only, the number of predator decreases as the number of prey increases (ahhh nature...).

Remark

In the literature we find sometimes the "-" sign in the top or bottom in the previous equations. In reality, the position of the "-" sign is not important because it is just the starting choice in the system dynamics.

2. Close to the equilibrium point $(1,1)$ we start from:

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}t}&=rx(1-y)\\ \frac{\mathrm{d}y}{\mathrm{d}t}&=my(x-1)\end{aligned} \quad (15.1.128)$$

And we put $x:=1+x, y:=1+y$ where x, y are now small perturbations near the point $(1,1)$. Therefore we have:

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}t}&=r(1+x)(1-(1+y))=r(1+x)y\cong ry\\ \frac{\mathrm{d}y}{\mathrm{d}t}&=m(1+y)((x+1)-1)=m(1+y)x\cong mx\end{aligned} \quad (15.1.129)$$

Now we change the position of the "-" sign on purpose to show that this is only a starting choice:

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}t}&\cong -ry\\ \frac{\mathrm{d}y}{\mathrm{d}t}&\cong mx\end{aligned} \quad (15.1.130)$$

To solve this system, let us differentiate the first equation once again:

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2}\cong -r\frac{\mathrm{d}y}{\mathrm{d}t} \quad (15.1.131)$$

and by injecting in it $\mathrm{d}y/\mathrm{d}t$:

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2}\cong -rmx \quad (15.1.132)$$

So we get a small second order differential equation (see section Differential and Integral Calculus page 1610). Whose typical simple solution is:

$$x=x_0e^{-\lambda t} \quad (15.1.133)$$

By injecting this solution into the differential equation, we obtain after simplification of exponentials a simple polynomial of the second degree (see section Calculus page 1368):

$$\lambda^2x_0+rmx_0=0 \quad (15.1.134)$$

Whose solution is trivial:

$$\lambda_{1,2}=\pm\mathrm{i}\sqrt{rm} \quad (15.1.135)$$

Thus, the general solution of the differential equation is the linear combination of the two special solutions we get:

$$x_(t)=x_0\left(e^{+\mathrm{i}\sqrt{rm}}+e^{-\mathrm{i}\sqrt{rm}}\right) \quad (15.1.136)$$

But we therefore have:

$$\frac{\mathrm{d}x}{\mathrm{d}t}\cong -ry \Leftrightarrow y(t)=-\frac{1}{r}\frac{\mathrm{d}x}{\mathrm{d}t} \tag{15.1.137}$$

Therefore, knowing $x(t)$ we obtain easily:

$$x(t)=-x_0\sqrt{\frac{m}{r}}\left(\mathrm{i}e^{+\mathrm{i}\sqrt{rm}}-\mathrm{i}e^{-\mathrm{i}\sqrt{rm}}\right) \tag{15.1.138}$$

Now let us use the Euler formula (see section Numbers page 251):

$$e^{\mathrm{i}x}=\cos(x)+\mathrm{i}\sin(x) \tag{15.1.139}$$

Thus we have:

$$x(t)=x_0\left[\cos\left(-\sqrt{rm}t\right)+\mathrm{i}\sin\left(-\sqrt{rm}t\right)+\cos\left(\sqrt{rm}t\right)+\mathrm{i}\sin\left(\sqrt{rm}t\right)\right] \tag{15.1.140}$$

and as (see section Trigonometry page 2514) $\cos(x)=\cos(-x), -\sin(x)=\sin(-x)$ then we have:

$$x(t)=2x_0\cos(\sqrt{rm}t) \tag{15.1.141}$$

and similarly, we obtain:

$$\begin{aligned}y(t)&=-x_0\sqrt{\frac{m}{r}}\mathrm{i}\left[-\left(\cos\left(-\sqrt{rm}t\right)+\mathrm{i}\sin\left(-\sqrt{rm}t\right)\right)+\cos\left(\sqrt{rm}t\right)+\mathrm{i}\sin\left(\sqrt{rm}t\right)\right]\\&=-x_0\sqrt{\frac{m}{r}}\mathrm{i}\left[-\left(\cos\left(\sqrt{rm}t\right)-\mathrm{i}\sin\left(\sqrt{rm}t\right)\right)+\cos\left(\sqrt{rm}t\right)+\mathrm{i}\sin\left(\sqrt{rm}t\right)\right]\\&=-x_0\sqrt{\frac{m}{r}}\mathrm{i}\left[-\cos\left(\sqrt{rm}t\right)+\mathrm{i}\sin\left(\sqrt{rm}t\right)+\cos\left(\sqrt{rm}t\right)+\mathrm{i}\sin\left(\sqrt{rm}t\right)\right]\\&=-x_0\sqrt{\frac{m}{r}}\mathrm{i}2\mathrm{i}\sin\left(\sqrt{rm}t\right)=2x_0\sqrt{\frac{m}{r}}\sin\left(\sqrt{rm}t\right)\end{aligned} \tag{15.1.142}$$

Thus, around the equilibrium point $(1,1)$ with sufficiently small perturbations to validate linearisation the system oscillate as ellipses (or circles) whose axes are defined by the two equations above.

We can get the graphs above with Maple™ 4.00b (this is a nice example of application of this software):

```
>restart:  with(plots):  with(DEtools):
>rate_eqn1:= diff(h(t),t)=(0.1)*h-(0.005)*h*(1/60)*u;
rate_eqn2:=diff(u(t),t)=(0.00004)*h*u-(0.04)*u;vars:= [h(t), u(t)];
>init1:=[h(0)=2000,u(0)=600]; init2:=[h(0)=2000,u(0)=1200];
init3:=[h(0)=2000, u(0)=3000];domain := 0 ..  320;
>L:= DEplot(rate_eqn1, rate_eqn2, vars, domain,init1, stepsize=0.5,
scene=[t, u], arrows=NONE):
>H:= DEplot(rate_eqn1, rate_eqn2, vars, domain,init1 , stepsize=0.5,
scene=[t, h], arrows=NONE):
```

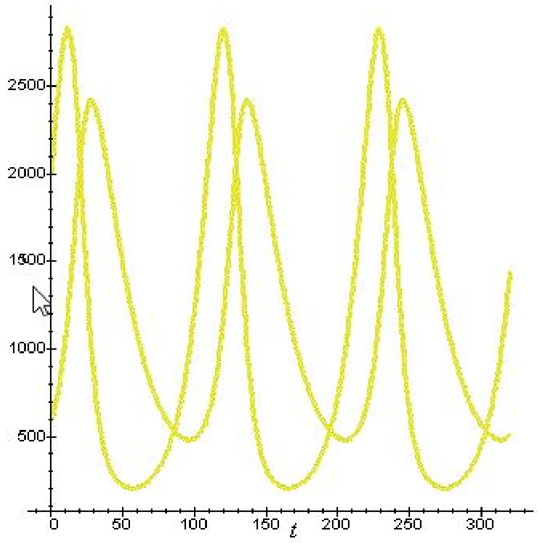


Figure 15.41: Lotka-Volterra offer/demand model periodic plot in Maple™ 4.00b

```
>DEplot(rate_eqn1, rate_eqn2, vars, t= 0 ..  160, init1, init2, init3,
stepsize=0.5, scene=[h,u], title='Demand vs.  60 * Offer for t = 0 ..
160', arrows=slim);
```

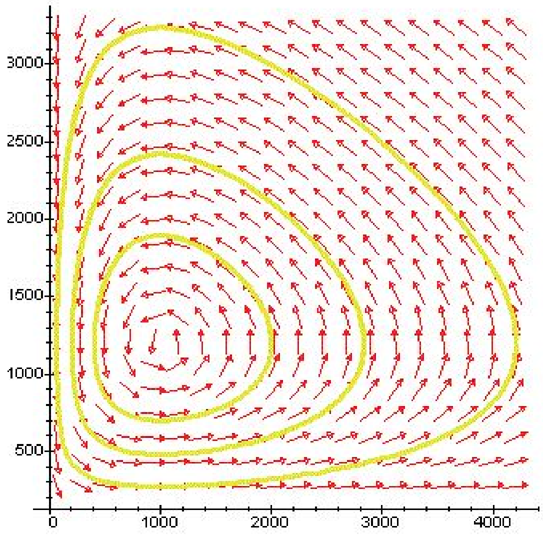


Figure 15.42: Lotka-Volterra offer/demand phase space model plot in Maple™ 4.00b

15.1.3 Schaefer's optimal capture model

A useful principle for the management of fishing, hunting, wild collection quotas or exploitation of raw materials is the "maximum sustainable yield" which is directly linked to the good management of the exploited dynamics of the concerned population.

Definition (#15.1.631): The "maximum sustainable yield" is the maximum amount of fish, deer, birds or fungi, etc. which can be killed / harvested in a population without jeopardizing its survival.

Determine this maximum sustainable yield give the possibility the get to the goal of fishing, hunting, gathering as much as possible, without falling into the overconsumption that potentially lead to the disappearance of the population.

The distinction between a balanced consumption of natural resources and an exaggerated one is often hard to do. Mathematical and statistical models are then always useful and necessary to establish that frontier.

The calculations are involved in the determination of quotas even still in the early 121st century (holocene calendar). The simplistic model that we will present is used by some large States (but it will in time disappear to make way for a more sophisticated model with probabilistic parameters).

We are therefore interested in the biomass of a population of a consumable. The model put forward by Schaefer in 11954 (holocene calendar) involves only two parameters:

1. The maximum biomass M that can live constantly in a system
2. The reproduction rate r

We start from the assumption (to check in practice because of the Lotka-Volterra model) that at the year zero (before to the consumption of the resource), the corresponding biomass is:

$$X_0 = M \tag{15.1.143}$$

The empirical model is:

$$\boxed{X_{n+1} = X_n + rX_n\left(1-\frac{X_n}{M}\right) - C} \tag{15.1.144}$$

To understand this model, let assume first that the population is not exploited, that is to say that $C = 0$. At the year n, the population then increases of:

$$r\left(1-\frac{X_n}{M}\right) \tag{15.1.145}$$

Thus, in the absence of fishing, if X_0 is smaller than M, the population will grow slowly until it reaches its maximum capacity M.

In the extreme case, where consumption C is equal to M, we have:

$$X_{0+1}=X_{+}rX_0\left(1-\frac{X_0}{M}\right)-C=M+rM\left(1-\frac{M}{M}\right)=M+0-M=0 \tag{15.1.146}$$

which makes sense ... (there is a total extinction!).

Now rewrite the Schaefer model as follows:

$$X_{n+1}-X_n=\Delta X=rX_n\left(1-\frac{X_n}{M}\right)-C \tag{15.1.147}$$

We would like to know for what value of X_n, we would have the biomass which no longer varies. That is to say:

$$\Delta X=rX_n\left(1-\frac{X_n}{M}\right)-C=0 \tag{15.1.148}$$

Therefore:

$$rX_n-r\frac{X_n^2}{M}-C=\frac{r}{M}X_n^2-rX_n+C=0 \tag{15.1.149}$$

So it is a simple equation of the second degree! It has a real solution only if the discriminant (see section Calculus page 1383) is positive or zero. Therefore, knowing that r is positive or zero by construction:

$$r^2-4\frac{r}{M}C\geq 0\Leftrightarrow r\left(r-\frac{4C}{M}\right)\geq 0\Leftrightarrow r-\frac{4C}{M}\geq 0\Leftrightarrow C\leq\frac{rM}{4} \tag{15.1.150}$$

Thus, the maximum possible value of C is $rM/4$. When we replace C by this value in the original equation:

$$\frac{r}{M}X_n^2-rX_n+C=0 \tag{15.1.151}$$

we get:

$$\frac{r}{M}X_n^2-rX_n+\frac{rM}{4}=0 \tag{15.1.152}$$

Hence:

$$\begin{aligned}&\frac{r}{M}\left(X_n^2-MX_n+\frac{M^2}{4}\right)=0\\&X_n^2-MX_n+\frac{M^2}{4}=0\\&\left(X_n-\frac{M}{2}\right)^2=0\end{aligned} \tag{15.1.153}$$

Therefore the solution is:

$$X_n=\frac{M}{2} \tag{15.1.154}$$

Thus, there exists a time n (infinite) for which the population tends asymptotically to $M/2$ at the condition that $X_0\geq\dfrac{M}{2}$.

Still its simplicity and its deterministic aspect the Schaefer model is still used in the early 121st century (holocene calendar) in fisheries management.

15.1.4 Hardy-Weinberg model

The Hardy-Weinberg model (developed in 11908 according to holocene calendar) describes the diversity (rather than evolution) of living in the point of view of population genetics. However, this model still allows to understand the impact of a mutation agent on a population of individuals and therefore to solve problems related to the refusal of the Natural selection model (evolution) of Darwin. Indeed, Darwin's work focused on characters that varied continuously. But he could not explain how individuals transmit these changes (this is actually quite annoying). The Mendel's genetics, meanwhile imposed that only discontinuous traits were hereditary.

So there was a conflict between the continuous model (Darwin) and the discontinuous one (Mendel) and it was genetics that in the 120th century (holocene calendar) helped to establish that fact that it is the mix of genes that creates diversity and verbatim natural selection.

Remark

We will prove in the section of Numerical Methods (see page 7022) as part of the study of Genetic Algorithms that the chromosome having the biggest function value of "fitness" will be the one most likely to reproduce! This is part of the theorem named the "Schema Theorem" ...

Conclusion: The evolution is not made at the level of the individual but of the population (this is a quite good lesson to remember...).

Definitions (#15.1.632):

D1. A gene or "genotype" is composed of "alleles" (usually 2, so we will consider individuals as being "*diploid*")

So the genotype is the genetic makeup of an organism. It's basically which alleles an organism has for a specific gene or trait. This is what causes genetic variations.

D2. A "gene pool" is a set of genes present in a population.

D3. A "micro-evolution" is a change in the frequency of alleles of a gene pool of a population.

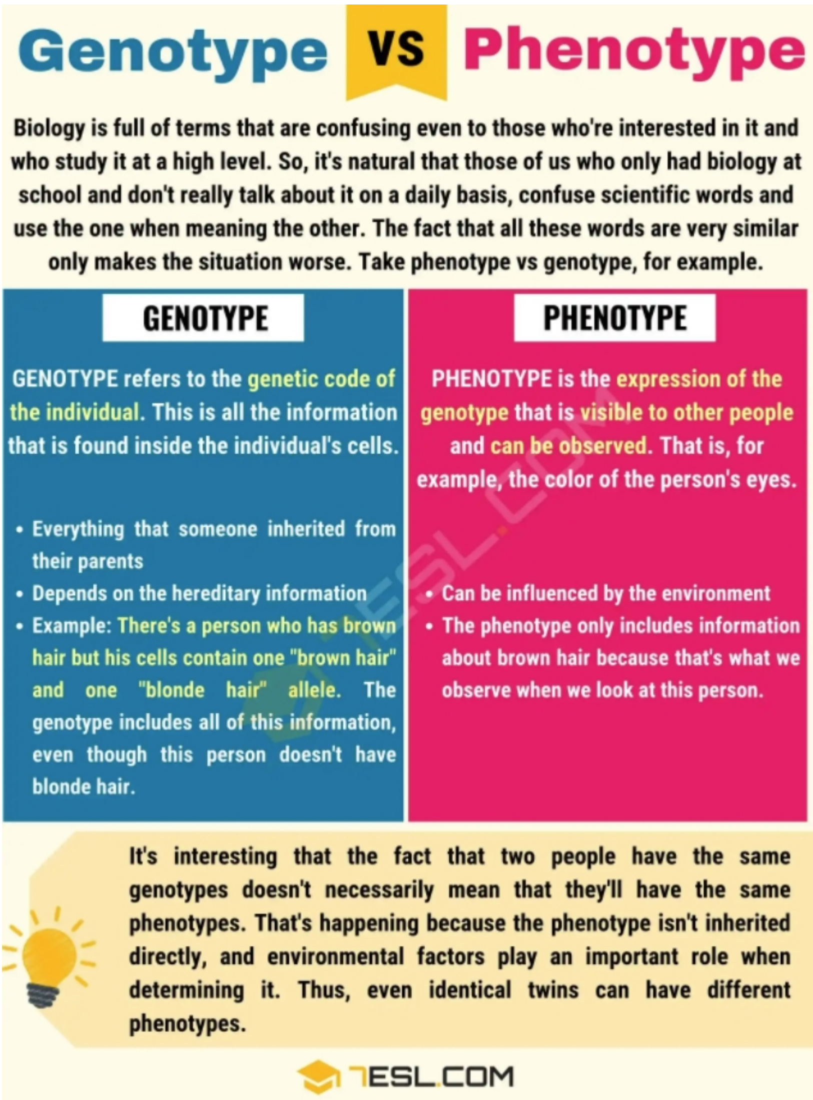
D4. An allele is named "fixed allele" when all members of a population have two identical alleles.

D5. A "speciation" is a long-term occurrence of a micro-evolution, making appear a new species.

D6. The "mutation" is an evolutionary agent that alters the gene pool of a population by producing new genes (we therefore consider that there is no evolutionary agents at work in the population in this model).

Remark

Mutation is a rare event and usually harmful. Its quantitative effects are greater in short generation time organisms (bacteria and viruses), the effect is much less pronounced among organisms with long generation time (animals and plants).



Genotype VS Phenotype

Biology is full of terms that are confusing even to those who're interested in it and who study it at a high level. So, it's natural that those of us who only had biology at school and don't really talk about it on a daily basis, confuse scientific words and use the one when meaning the other. The fact that all these words are very similar only makes the situation worse. Take phenotype vs genotype, for example.

GENOTYPE	PHENOTYPE
GENOTYPE refers to the genetic code of the individual. This is all the information that is found inside the individual's cells.	PHENOTYPE is the expression of the genotype that is visible to other people and can be observed. That is, for example, the color of the person's eyes.
• Everything that someone inherited from their parents • Depends on the hereditary information • Example: There's a person who has brown hair but his cells contain one "brown hair" and one "blonde hair" allele. The genotype includes all of this information, even though this person doesn't have blonde hair.	• Can be influenced by the environment • The phenotype only includes information about brown hair because that's what we observe when we look at this person.

It's interesting that the fact that two people have the same genotypes doesn't necessarily mean that they'll have the same phenotypes. That's happening because the phenotype isn't inherited directly, and environmental factors play an important role when determining it. Thus, even identical twins can have different phenotypes.

7ESL.COM

Figure 15.43: **Genotype vs Phenotype (source: TESL.com)**

Let us now consider a population consisting of 13 individuals (1 gene, each consisting of two alleles selected among 3):

$$P=\left\{\begin{matrix}\text{X1X1,X2X1,X3X2,X1X1,X1X1,X1X1,X2X3,}\\\text{X1X3,X3X2,X1X1,X1X3,X3X3,X3X3}\end{matrix}\right\}$$

Therefore we have the following relative allele frequencies:

$$p_{X1}=\frac{13}{26}=0.5 \qquad p_{X2}=\frac{4}{26}=0.154 \qquad p_{X3}=\frac{9}{26}=0.346 \tag{15.1.155}$$

with obviously:

$$p_{X1}+p_{X2}+p_{X3}=1 \tag{15.1.156}$$

And the following relative genotype frequencies:

$$p_{X1X1}=\frac{5}{13}=0.384 \quad p_{X2X1}=\frac{1}{13}=0.077 \quad p_{X3X2}=\frac{3}{13}=0.231 \quad p_{X1X3}=\frac{2}{13}=0.154 \quad p_{X3X3}=\frac{2}{13} \tag{15.1.157}$$

with obviously:

$$p_{X1X1}+p_{X2X1}+p_{X3X2}+p_{X1X3}+p_{X3X3}=1 \tag{15.1.158}$$

To summarize, if we now consider a population of individuals possessing genes made of only two types of alleles, we have therefore (more concrete case relatively to the actual living biology on Earth):

$$BB, Bb, bb \tag{15.1.159}$$

So we have for the relative frequency of the number of alleles B:

$$p=\frac{2(\text{number of individuals } BB)+1(\text{number of individuals } Bb)}{\text{total number of individuals}}=\frac{3}{6} \tag{15.1.160}$$

and for the relative frequency of allele b:

$$q=\frac{2(\text{number of individuals } bb)+1(\text{number of individuals } Bb)}{\text{total number of individuals}}=\frac{3}{6} \tag{15.1.161}$$

With obviously:

$$p+q=1 \tag{15.1.162}$$

This can be illustrated (by changing the letter B by Y and b by y):

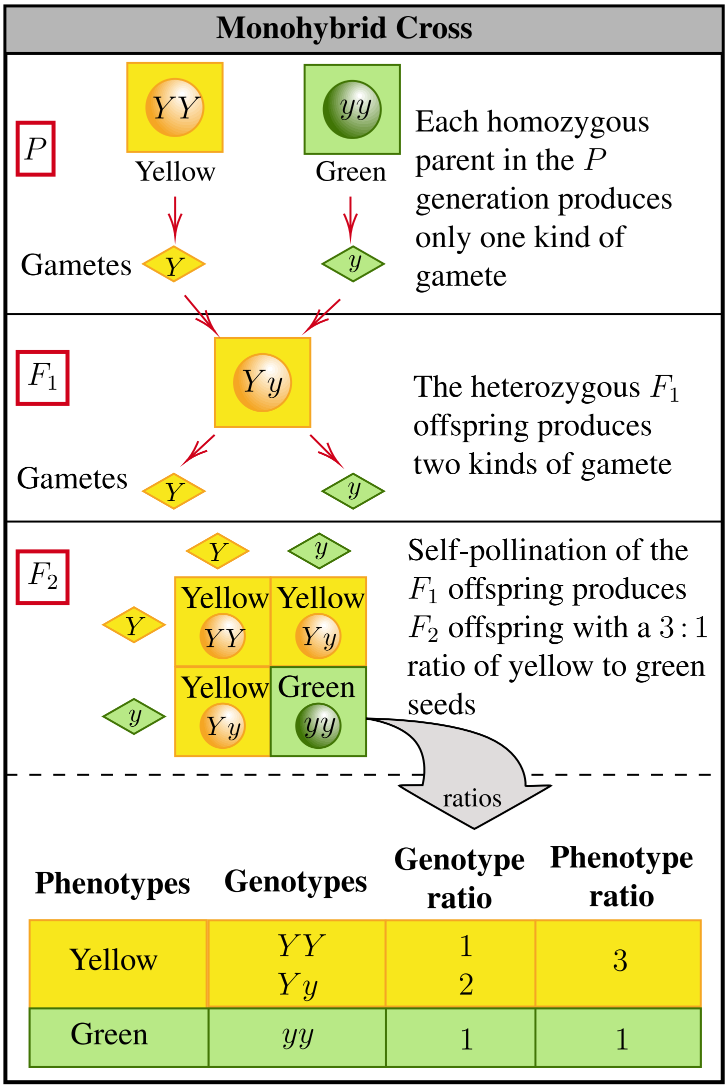

Monohybrid Cross
P
YY
Yellow
yy
Green
Each homozygous parent in the P generation produces only one kind of gamete
Gametes
Y
y
F1
Yy
The heterozygous F1 offspring produces two kinds of gamete
Gametes
Y
y
F2
Y
y
Y
y
Yellow YY
Yellow Yy
Yellow Yy
Green yy
Self-pollination of the F1 offspring produces F2 offspring with a 3 : 1 ratio of yellow to green seeds
ratios
Phenotypes
Genotypes
Genotype ratio
Phenotype ratio
Yellow
YY
Yy
1
2
3
Green
yy
1
1

Figure 15.44: **Monohybrid cross**

We can see in the figure above that the allele relative frequencies are $50\%/50\%$, the genotype relative frequencies are $25\%, 50\%, 25\%$ (also often denoted $1:2:1$) and the phenotype relative frequency $66.\bar{6}\%, 33.\bar{3}\%$ (also often denoted $3:1$).

We can make a few observations about the way of constructing this model. Indeed, the frequency of alleles (genetic equilibrium) remains in a population approximately constant generation after generation if:

1. The population is very large (a tiny disturbance changes only a little bit the frequency of the types of genotypes).
2. No emigration or immigration (no new types of alleles in the population).
3. No mutations altering the alleles (this contains the point number 2).
4. Random couplings (not influenced by the type of allele studied).
5. No natural selection (this contains the point number 4).

It may be relevant (for general knowledge) to highlight the fact that we distinguish two types of non-random accouplement:

1. Inbreeding: crossing between individuals of the neighbourhood, thus having family links; in species dispersing only a little bit. This apply especially to vegetables.
2. Homogamy: free choose of partners who have some similarities for given physical characters. This apply especially among animals.

The inbreeding and homogamy accouplement is not a direct cause of micro-evolution since they do not alter the gene pool of a population. However, if some individuals coming from this non-random accouplement are more likely to have accouplement with others (natural selection), it will follow a change in the allele frequency in the downward population and therefore a micro-evolution. As a result, non-random accouplement are a potential factor of micro-evolution.

Finally, to return to our model, the probability that an individual has one of three possible genotypes (relatively to a genotype) in the population (relative to our example) is:

1. Probability that an individual is BB:

$$BB = pp = p^2 \tag{15.1.163}$$

2. Probability that an individual is bb:

$$bb = qq = q^2 \tag{15.1.164}$$

3. Probability that an individual is Bb or bB:

$$P(bB, Bb) = P(bB \vee Bb) = pq + qp = 2pq \tag{15.1.165}$$

Finally, we get the "Hardy-Weinberg equation" (for the special case of two alleles):

$$\boxed{p^2+2pq+q^2=(p+q)^2=1} \tag{15.1.166}$$

That we can represent the graph:

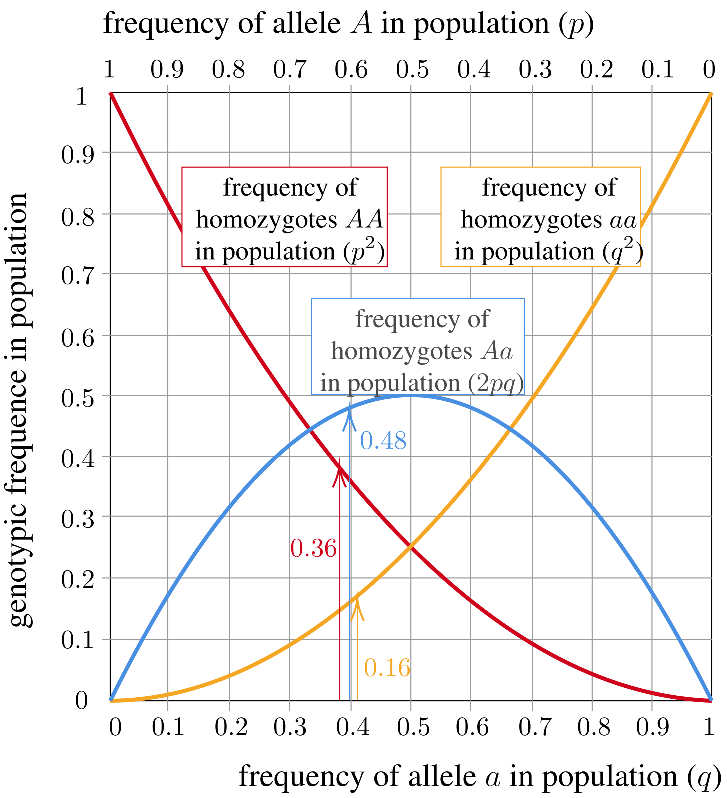


Figure 15.45: **Plot of the Hardy-Weinberg equation (author: Antonín Šípek)**

A population that is not changing genetically is in Hardy-Weinberg equilibrium if these five assumptions are correct:

- Random mating (panmixia)
- Large population size (N approaching infinity)
- No migration between populations
- No (or negligible) mutations
- Natural selection does not affect alleles being considered

If these assumptions are true, it follows that:

- Allele frequencies remain constant from one generation to the next
- After one (or more) generations of random mating (breeding), the genotype frequencies (for a 2-allele gene with allele frequencies p,q) are in the proportions: $p^2_{(AA)}:2pq_{(Aa)}:q^2_{(aa)}$, and population will be in H-W equilibrium.

- For a population to be in Hardy Weinberg equilibrium, the observed genotype frequencies must match those predicted by the equation $p^2+2pq+q^2$.

The Hardy–Weinberg principle may also be generalized to polyploid[4] systems, that is, for organisms that have more than two copies of each chromosome. Consider again only two alleles. Multinomial expansion for two alleles A and a with frequencies p and $qp^2+2pq+q^2$ is a binomial expansion of $(p+q)^2$. That is:

$$p^2_{(AA)}+2pq_{(Aa)}+q^2_{(aa)}=(p+q)^2=(1)^2=1 \quad (15.1.167)$$

For three alleles A, B and C with frequencies p, q and r, the multinomial expansion is $(p+q+r)^2$ which expands into:

$$p^2+q^2+r^2+2pq+2pr+2qr \quad (15.1.168)$$

where the first three terms being homozygotes and the remaining three heterozygotes. Therefore the three alleles Hardy-Weinberg equation is given by:

$$p+q+r=p^2+q^2+r^2+2pq+2pr+2qr=1 \quad (15.1.169)$$

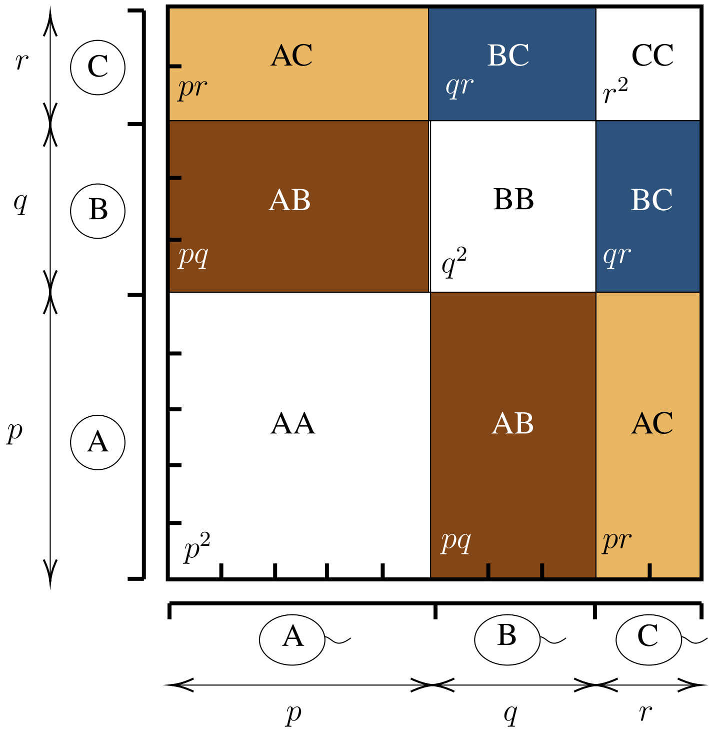


Figure 15.46: **Punnett square for three-allele case (source: Wikipedia)**

therefore the polyploid case is the polynomial expansion of $(p+q)^c$, where c is the ploidy, for example with tetraploid $(c=4)$:

---

[4] Humans are named "diploid organisms" because they have two alleles at each genetic locus, with one allele inherited from each parent.

Genotype	Frequency
$AAAA$	p^4
$AAAa$	$4p^3q$
$AAaa$	$6p^2q^2$
$Aaaa$	$4pq^3$
$aaaa$	q^4

For n distinct alleles in c-ploids, the genotype frequencies in the Hardy-Weinberg equilibrium are given by individual terms in the multinomial expansion of $(p_1+\cdots+p_n)^c$:

$$(p_1+\cdots+p_n)^c=\sum_{k_1,\ldots,k_n\in\mathbb{N}:k_1+\cdots+k_n=c}\binom{c}{k_1,\ldots,k_n}p_1^{k_1}\cdots p_n^{k_n} \tag{15.1.170}$$

After this study, we can finally (re)defined the "Darwin's law" so it gives us the possibility to eliminate the problem of the big question (problem) of continuous or discrete changes:

Definition (#15.1.633): "Natural selection" is an evolutionary agent that alters the gene pool of a population by increasing the frequency of alleles producing "phenotypes" (genotypes defines the behaviour of an individual) best suited to the medium (environment). Thus, the individuals that are the best adapted to their environment reproduce more than others and contribute more to the genetic heritage of the descendants. The frequency of good genes gradually increases in the population from generation to generation. Thus, natural selection directs the adaptation of a population to its environment by accumulating genotypes that promote survival in the medium (environment). For the definition of (biological) Evolution the reader can go see the page 459.

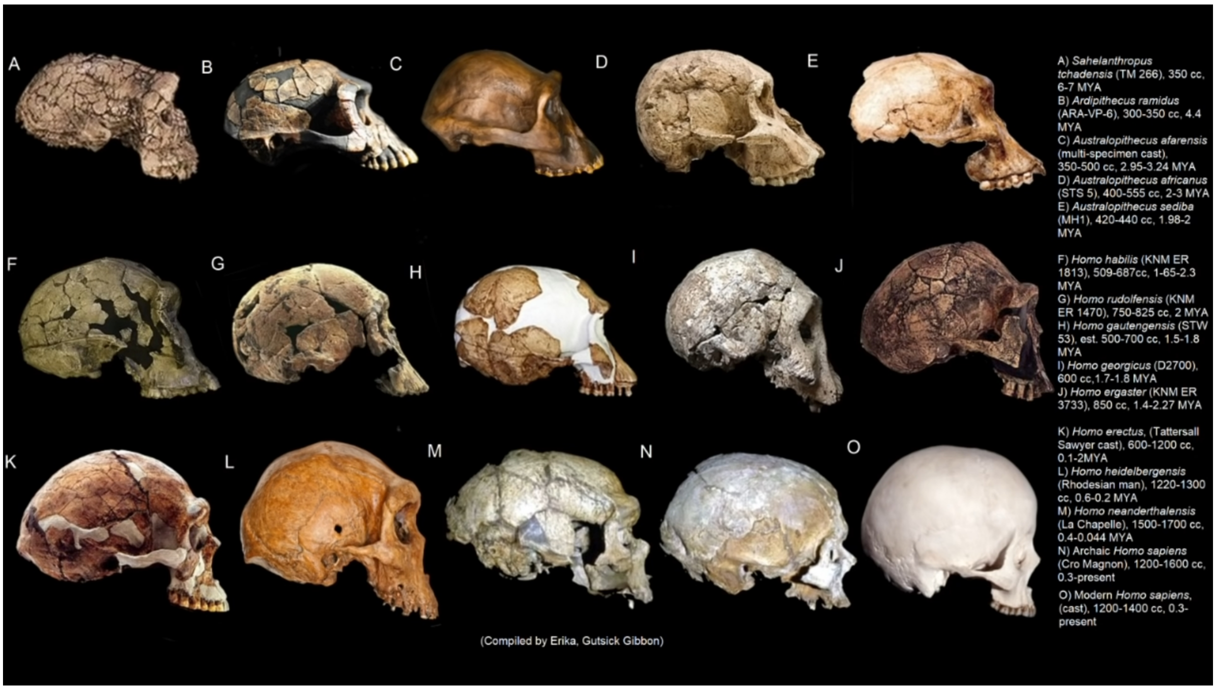

A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
A) Sahelanthropus tchadensis (TM 266), 350 cc, 6-7 MYA
B) Ardipithecus ramidus (ARA-VP-6), 300-350 cc, 4.4 MYA
C) Australopithecus afarensis (multi-specimen cast), 350-500 cc, 2.95-3.24 MYA
D) Australopithecus africanus (STS 5), 400-555 cc, 2-3 MYA
E) Australopithecus sediba (MH1), 420-440 cc, 1.98-2 MYA
F) Homo habilis (KNM ER 1813), 509-687cc, 1-65-2.3 MYA
G) Homo rudolfensis (KNM ER 1470), 750-825 cc, 2 MYA
H) Homo gautengensis (STW 53), est. 500-700 cc, 1.5-1.8 MYA
I) Homo georgicus (D2700), 600 cc,1.7-1.8 MYA
J) Homo ergaster (KNM ER 3733), 850 cc, 1.4-2.27 MYA
K) Homo erectus, (Tattersall Sawyer cast), 600-1200 cc, 0.1-2MYA
L) Homo heidelbergensis (Rhodesian man), 1220-1300 cc, 0.6-0.2 MYA
M) Homo neanderthalensis (La Chapelle), 1500-1700 cc, 0.4-0.044 MYA
N) Archaic Homo sapiens (Cro Magnon), 1200-1600 cc, 0.3-present
O) Modern Homo sapiens, (cast), 1200-1400 cc, 0.3-present
(Compiled by Erika, Gutsick Gibbon)

Figure 15.47: **Evidence of the smooth evolution of hominini crane through six to seven million years (author: Erika/Gutsick Gibbon)**

A study [406] provides some evidence that average mutation rate can be estimated to approximately 175 mutations per diploid genome per generation ($\sim 2.5\cdot 10^{-8}$ per base per generation).

The highest per base pair per generation mutation rates are found in viruses [157], which can have either RNA or DNA genomes. DNA viruses have mutation rates between 10^{-6} to 10^{-8} mutations per base per generation, and RNA viruses (whose genomes contain ca. 10^4 bases) have mutation rates between 10^{-3} to 10^{-5} per base per generation.

The high-resolution figure below (you can deeply zoom in for the details!) represents a robust evolutionary timescale for all approximately 6,000 living species of mammals [550] in the shape of a circular phylogenetic tree:

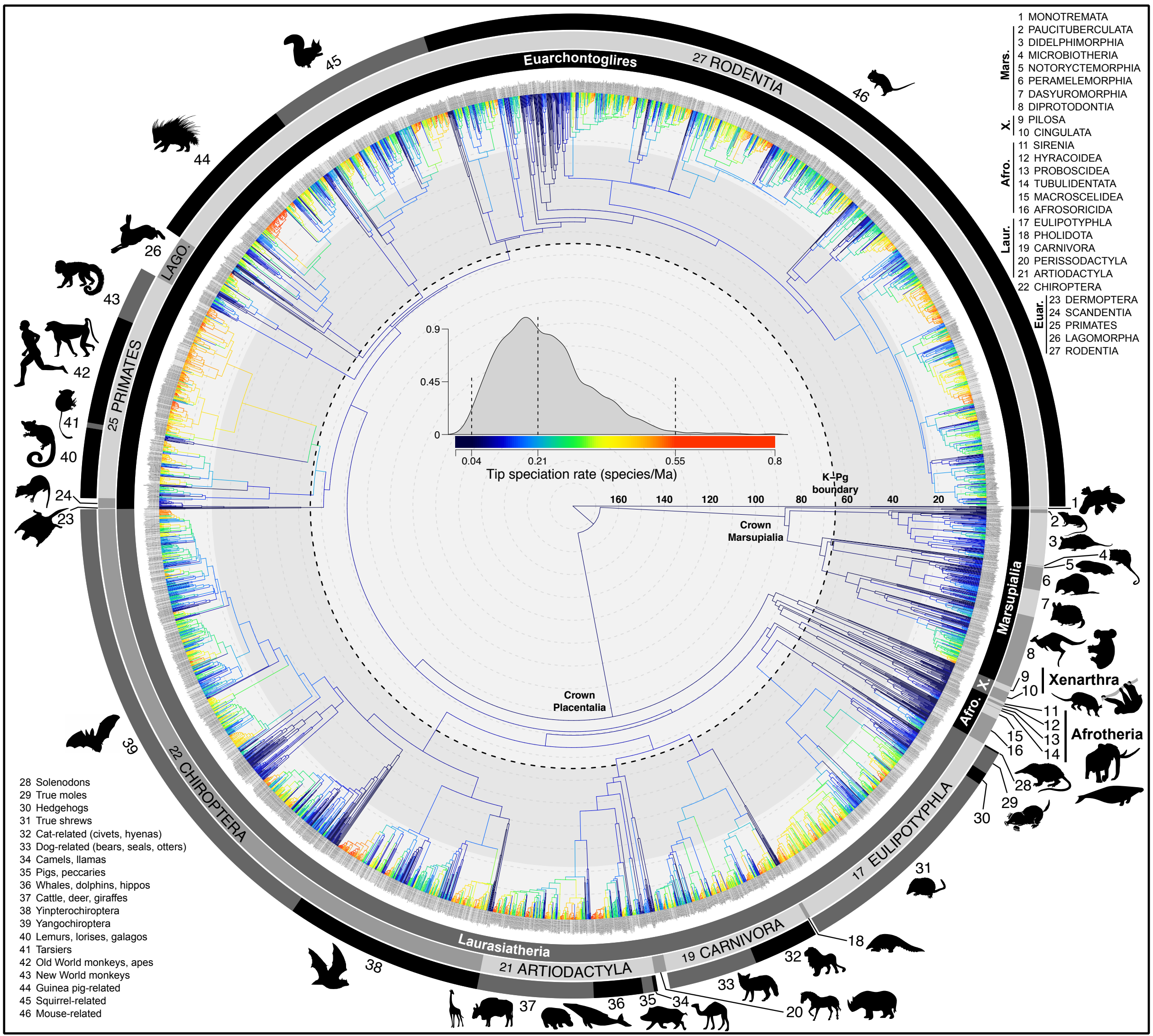
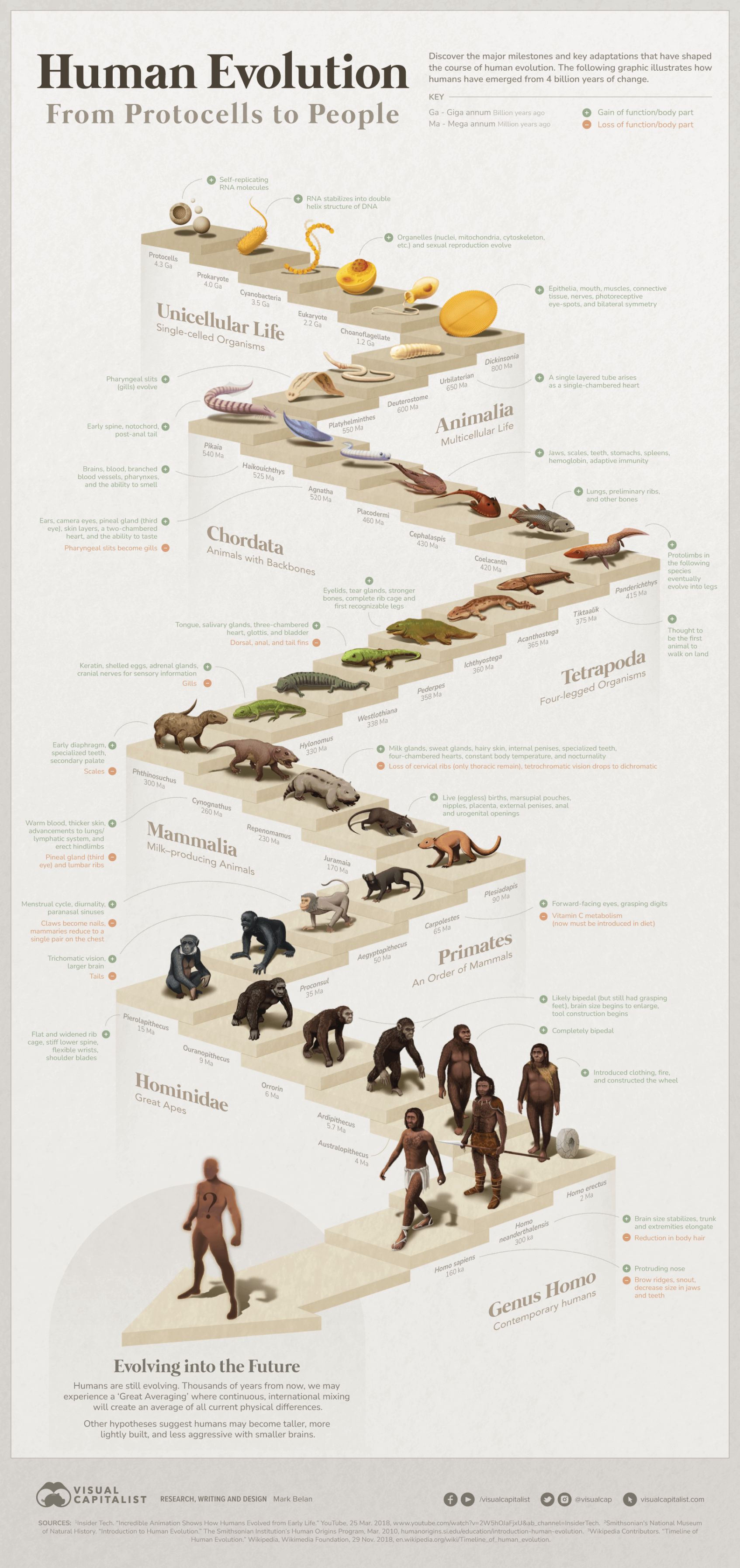

1 MONOTREMATA
2 PAUCITUBERCULATA
3 DIDELPHIMORPHIA
4 MICROBIOTHERIA
5 NOTORYCTEMORPHIA
6 PERAMELEMORPHIA
7 DASYUROMORPHIA
8 DIPROTODONTIA
Mars.
9 PILOSA
10 CINGULATA
X.
11 SIRENIA
12 HYRACOIDEA
13 PROBOSCIDEA
14 TUBULIDENTATA
15 MACROSCELIDEA
16 AFROSORICIDA
Afro.
17 EULIPOTYPHLA
18 PHOLIDOTA
19 CARNIVORA
20 PERISSODACTYLA
21 ARTIODACTYLA
22 CHIROPTERA
Laur.
23 DERMOPTERA
24 SCANDENTIA
25 PRIMATES
26 LAGOMORPHA
27 RODENTIA
Euar.
28 Solenodons
29 True moles
30 Hedgehogs
31 True shrews
32 Cat-related (civets, hyenas)
33 Dog-related (bears, seals, otters)
34 Camels, llamas
35 Pigs, peccaries
36 Whales, dolphins, hippos
37 Cattle, deer, giraffes
38 Yinpterochiroptera
39 Yangochiroptera
40 Lemurs, lorises, galagos
41 Tarsiers
42 Old World monkeys, apes
43 New World monkeys
44 Guinea pig-related
45 Squirrel-related
46 Mouse-related
Euarchontoglires
27 RODENTIA
LAGO.
25 PRIMATES
22 CHIROPTERA
Laurasiatheria
21 ARTIODACTYLA
19 CARNIVORA
17 EULIPOTYPHLA
Afro.
X.
Marsupialia
Xenarthra
Afrotheria
0.9
0.45
0
0.04
0.21
0.55
0.8
Tip speciation rate (species/Ma)
K-Pg boundary
160
140
120
100
80
60
40
20
Crown Marsupialia
Crown Placentalia

Figure 15.48: Species-level relationships and tempo of diversification across mammals (philogenetic tree)

In a philogenetic tree we have abductive evidence that from the egg and the chicken (i.e. Gallus gallus deomesticus) the egg existed first as the chicken is a descendant of the Gallus gallus (5 millions years ago) that already made eggs and itself is also a descendant of the Basai Phasianid (30 millions years ago) that also already made eggs before, etc.

And here is a quite common linear perspective (therefore wrong for pedagogical reasons) of Human Evolution:



Human Evolution

From Protocells to People

Discover the major milestones and key adaptations that have shaped the course of human evolution. The following graphic illustrates how humans have emerged from 4 billion years of change.

Evolving into the Future

Humans are still evolving. Thousands of years from now, we may experience a 'Great Averaging' where continuous, international mixing will create an average of all current physical differences.

Other hypotheses suggest humans may become taller, more lightly built, and less aggressive with smaller brains.

RESEARCH, WRITING AND DESIGN Mark Belan

SOURCES: [1]Insider Tech. "Incredible Animation Shows How Humans Evolved from Early Life." YouTube, 25 Mar. 2018, www.youtube.com/watch?v=2W5hOJaFjxU&ab_channel=InsiderTech. [2]Smithsonian's National Museum of Natural History. "Introduction to Human Evolution." The Smithsonian Institution's Human Origins Program, Mar. 2010, humanorigins.si.edu/education/introduction-human-evolution. [3]Wikipedia Contributors. "Timeline of Human Evolution." Wikipedia, Wikimedia Foundation, 29 Nov. 2018, en.wikipedia.org/wiki/Timeline_of_human_evolution.

KEY
Ga - Giga annum Billion years ago
Ma - Mega annum Million years ago
+ Gain of function/body part
− Loss of function/body part
Self-replicating RNA molecules
RNA stabilizes into double helix structure of DNA
Organelles (nuclei, mitochondria, cytoskeleton, etc.) and sexual reproduction evolve
Epithelia, mouth, muscles, connective tissue, nerves, photoreceptive eye-spots, and bilateral symmetry
Protocells 4.3 Ga
Prokaryote 4.0 Ga
Cyanobacteria 3.5 Ga
Eukaryote 2.2 Ga
Choanoflagellate 1.2 Ga
Unicellular Life
Single-celled Organisms
Dickinsonia 800 Ma
Urbilaterian 650 Ma
Deuterostome 600 Ma
Platyhelminthes 550 Ma
Animalia
Multicellular Life
A single layered tube arises as a single-chambered heart
Pharyngeal slits (gills) evolve
Early spine, notochord, post-anal tail
Brains, blood, branched blood vessels, pharynxes, and the ability to smell
Ears, camera eyes, pineal gland (third eye), skin layers, a two-chambered heart, and the ability to taste
Pharyngeal slits become gills
Jaws, scales, teeth, stomachs, spleens, hemoglobin, adaptive immunity
Lungs, preliminary ribs, and other bones
Pikaia 540 Ma
Haikouichthys 525 Ma
Agnatha 520 Ma
Placodermi 460 Ma
Cephalaspis 430 Ma
Coelacanth 420 Ma
Chordata
Animals with Backbones
Protolimbs in the following species eventually evolve into legs
Eyelids, tear glands, stronger bones, complete rib cage and first recognizable legs
Tongue, salivary glands, three-chambered heart, glottis, and bladder
Dorsal, anal, and tail fins
Keratin, shelled eggs, adrenal glands, cranial nerves for sensory information
Gills
Panderichthys 415 Ma
Tiktaalik 375 Ma
Acanthostega 365 Ma
Ichthyostega 360 Ma
Pederpes 358 Ma
Westlothiana 338 Ma
Thought to be the first animal to walk on land
Tetrapoda
Four-legged Organisms
Early diaphragm, specialized teeth, secondary palate
Scales
Hylonomus 330 Ma
Milk glands, sweat glands, hairy skin, internal penises, specialized teeth, four-chambered hearts, constant body temperature, and nocturnality
Loss of cervical ribs (only thoracic remain), tetrochromatic vision drops to dichromatic
Phthinosuchus 300 Ma
Cynognathus 260 Ma
Repenomamus 230 Ma
Juramaia 170 Ma
Live (eggless) births, marsupial pouches, nipples, placenta, external penises, anal and urogenital openings
Warm blood, thicker skin, advancements to lungs/ lymphatic system, and erect hindlimbs
Pineal gland (third eye) and lumbar ribs
Mammalia
Milk–producing Animals
Menstrual cycle, diurnality, paranasal sinuses
Claws become nails, mammaries reduce to a single pair on the chest
Trichomatic vision, larger brain
Tails
Plesiadapis 90 Ma
Carpolestes 65 Ma
Aegyptopithecus 50 Ma
Proconsul 35 Ma
Forward-facing eyes, grasping digits
Vitamin C metabolism (now must be introduced in diet)
Primates
An Order of Mammals
Likely bipedal (but still had grasping feet), brain size begins to enlarge, tool construction begins
Completely bipedal
Introduced clothing, fire, and constructed the wheel
Flat and widened rib cage, stiff lower spine, flexible wrists, shoulder blades
Pierolapithecus 15 Ma
Ouranopithecus 9 Ma
Orrorin 6 Ma
Ardipithecus 5.7 Ma
Australopithecus 4 Ma
Hominidae
Great Apes
Homo erectus 2 Ma
Homo neanderthalensis 300 ka
Homo sapiens 160 ka
Brain size stabilizes, trunk and extremities elongate
Reduction in body hair
Protruding nose
Brow ridges, snout, decrease size in jaws and teeth
Genus Homo
Contemporary humans
?
VISUAL CAPITALIST
/visualcapitalist
@visualcap
visualcapitalist.com

Figure 15.49: Human linear evolution perspective (credit: Visual Capitalist, https://www.visualcapitalist.com/path-of-human-evolution/)

Remarks

R1. The Hardy-Weinberg model is very unlikely in nature. Indeed, the probabilities that all the necessary assumptions are present simultaneously is quite low. However, this is a theoretical model that evaluates if there is a micro-evolution in a population. Thus, we measure the frequency of alleles in the parent population and then we apply the Hardy-Weinberg equation. If the genetic structure of the descendant population differs from that predicted by the law, then we know that at least one agent of evolution is at work.

R2. This theoretical model can help to predict a punctual estimator of the social costs of genetic diseases. Indeed, if we know the percentage of individuals with recessive disease in a population, we can apply the Hardy-Weinberg equation and evaluate the number of healthy transmitters and then we can predict the likelihood of the disease in future generations. If it is large (this reasoning is valid only in a liberal and capitalist framework), we may decide to invest in research or providing funds to treat the sick people.

R3. For people interested in exploring the subject further, we highly recommend reading [118] and [226], which are two masterpiece in this field!

✍**Examples:**

E1. Given a genotype of possible sequences $\{BB, bb, Bb\}$ and let consider that $p_B = 0.6, p_b = 0.4$. We then have:

$$p^2 = 0.36 \quad 2py = 0.48 \quad q^2 = 0.16 \tag{15.1.171}$$

Therefore, in a population of $1,000,000$ individuals, we should have:

$$\begin{aligned} 0.36 \cdot 1 \cdot 10^6 &= 360,000 BB \\ 0.48 \cdot 1 \cdot 10^6 &= 480,000 Bb \\ 0.16 \cdot 1 \cdot 10^6 &= 160,000 BB \end{aligned} \tag{15.1.172}$$

E2. If in Quebec (6 million inhabitant), one in five has blue eyes ($bb \Rightarrow q^2 = 1/5$), how many people are of genotype BB and how many of type Bb?

$$q = \sqrt{1/5} = 0.44721 \tag{15.1.173}$$

$$p = (1 - q) = 0.5528 \tag{15.1.174}$$

Therefore:

$$p^2 = 0.3056 \quad 2pq = 0.4944 \quad q^2 = 0.2 \tag{15.1.175}$$

Therefore we have:

$$\begin{aligned} 0.3056 \cdot 6 \cdot 10^6 &= 1,833,600 BB \\ 0.4944 \cdot 6 \cdot 10^6 &= 2,966,400 Bb \\ 0.2 \cdot 6 \cdot 10^6 &= 1,200,000 bb = \frac{1}{5} \cdot 6 \cdot 10^6 \end{aligned} \tag{15.1.176}$$

E3. We introduce $1,000$ spotted frogs in a pond (homozygous for this characteristic: TT) and 250 frogs without spots (homozygous for this trait: tt). If we allow the frogs to reproduce for several years, assuming that the population remains stable, what number of frogs TT, Tt, tt should we then observe?

So we have a population of $1,250$ individuals in which we $1,000TT$ ($2,000$ allele T) and $250tt$ (500 alleles t). So:

$$\left.\begin{aligned} p(T)&=\frac{2,000}{2,500}=0.8\\ q(t)&=\frac{500}{2,500}=0.2\end{aligned}\right\}\Rightarrow\begin{cases}p^2=0.64\\ 2pq=0.32\\ q^2=0.04\end{cases} \tag{15.1.177}$$

If the population is of $1,250$ people, we have:

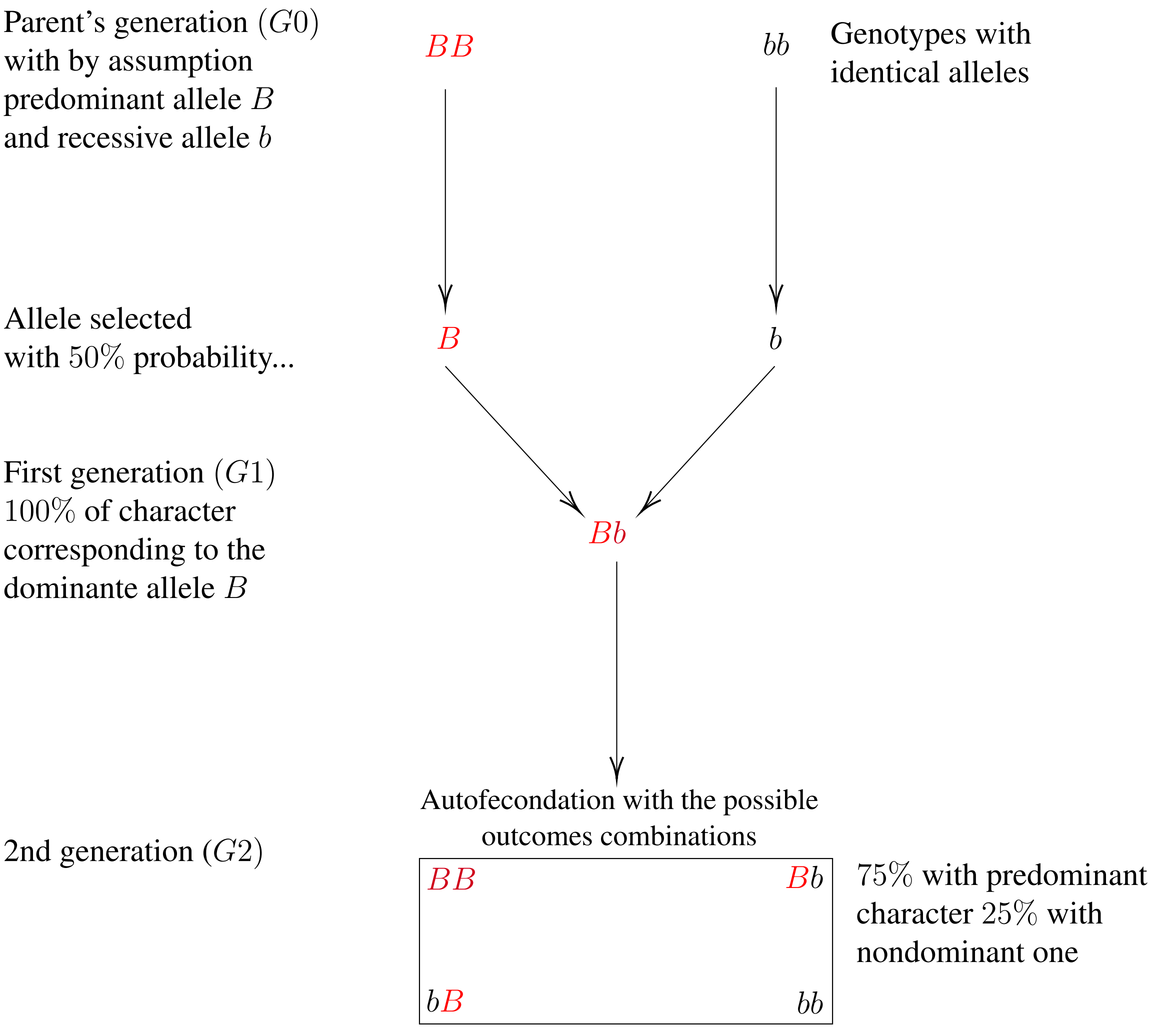
$$\begin{aligned}0.64\cdot 1,250&=800\text{ individuals }TT\\ 0.32\cdot 1,250&=400\text{ individuals }Tt\\ 0.04\cdot 1,250&=50\text{ individuals }tt\end{aligned} \tag{15.1.178}$$

Caution! As always... reality is more complex and chaotic than theory. A lot of people still believe that the human is perfect and the world too (we know that Earth is far from being perfect after our study of the section of Astrophysics)! However these people don't know that there are ~ 150 types of cancer (a few hundreds of them are described in [343]), a few thousands of mental disorders (a few hundreds of them are described in [532]) and $\sim 4,200$ different of genetic diseases (half a thousand are described in [410]). For example, the Swyer syndrome is a form of intersex linked to hypogonadism where an individual with a female phenotype has a karyotype with XY type sex chromosomes (in other words the individual has female organs but is born with male chromosomes) but where the SRY gene, responsible for the differentiation of the undifferentiated embryo into a male individual at the sixth week of embryonic development, is mutated (i.e. has undergone a mutation).

15.1.5 Mendel's law

During methodical hybridization of many species and varieties of various plants, it has been observed, in general, the uniformity and character of the first hybrid generation ($G1$) intermediate between the parents. So the inheritance is totally taken form the dominant character in the first generation (father or mother but of the mix). At the second generation ($G2$) occurs polymorphism showing various combinations of characters of the initial generation ($G0$). Thus we find in the second generation 25% of individuals with non-dominant character of the initial generation and 75% with the dominant character. The character that was masked in,the first generation (G1) is then named "recessive character".

This fact can be summarized with the following structure:


Parent's generation $(G0)$ with by assumption predominant allele B and recessive allele b
BB
bb
Genotypes with identical alleles
Allele selected with 50% probability...
B
b
First generation $(G1)$ 100% of character corresponding to the dominante allele B
Bb
2nd generation $(G2)$
Autofecondation with the possible outcomes combinations
BB
Bb
bB
bb
75% with predominant character 25% with nondominant one

Mendel then made the assumption that the gametes of hybrids of the first generation $(G1)$ must be in equal numerical proportions, of the type of one or other of the parents. This is the "law of purity of gametes".

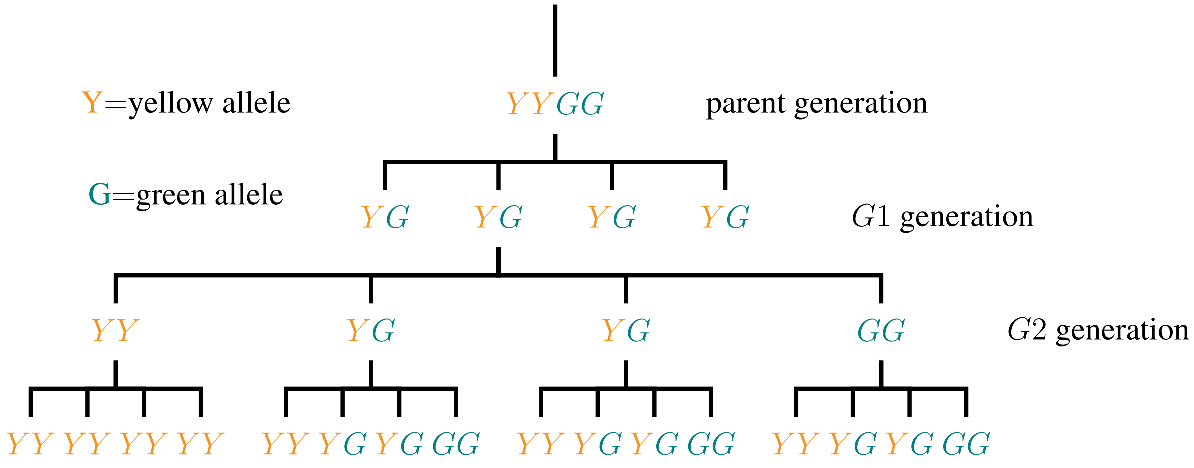
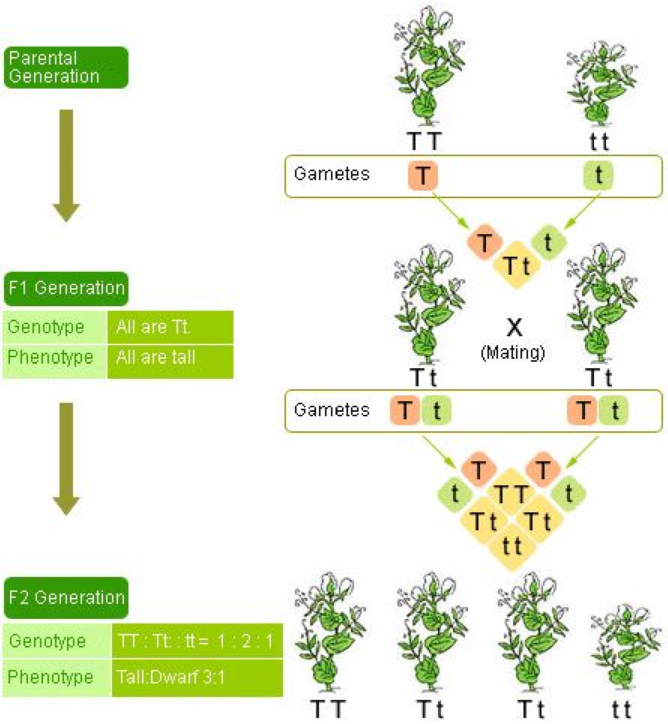

Y=yellow allele
G=green allele
$YYGG$
parent generation
YG YG YG YG
$G1$ generation
YY
YG
YG
GG
$G2$ generation
YY YY YY YY
YY YG YG GG
YY YG YG GG
YY YG YG GG

Figure 15.50: **Mendelian Inheritence Chart**

It has been established that human blood groups also obey to the rules of Mendelian inheritance. Thus, A and B blood groups are dominant and can appear in children if and only if they are present by the parents.

In a more pedagogical way, the previous figure can be illustrated as following:


Parental Generation
TT
tt
Gametes
T
t
T
t
Tt
F1 Generation
Genotype
All are Tt
Phenotype
All are tall
Tt
X
(Mating)
Tt
Gametes
T
t
T
t
T
T
t
t
TT
Tt
Tt
tt
F2 Generation
Genotype
TT : Tt : tt = 1 : 2 : 1
Phenotype
Tall:Dwarf 3:1
TT
Tt
Tt
tt

Mendel's laws are stated as following:

L1. "Law of dominance and uniformity": Some alleles are dominant while others are recessive; an organism with at least one dominant allele will display the effect of the dominant allele.

L2. "Law of segregation": During gamete formation, the alleles for each gene segregate from each other so that each gamete carries only one allele for each gene.

L3. "Law of independent assortment": Genes of different traits can segregate independently during the formation of gametes.

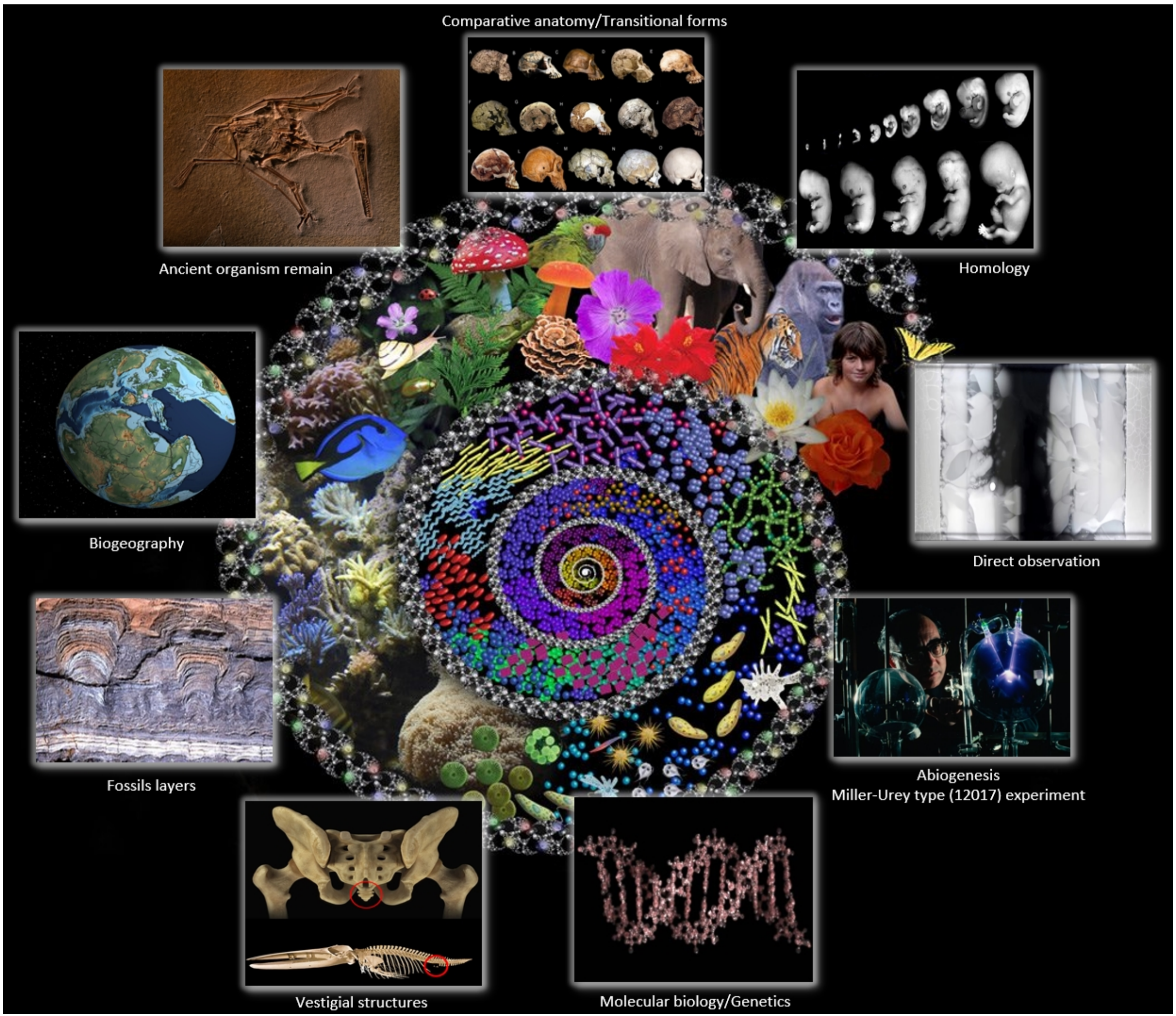

Comparative anatomy/Transitional forms
Ancient organism remain
Homology
Biogeography
Direct observation
Fossils layers
Abiogenesis
Miller-Urey type (12017) experiment
Vestigial structures
Molecular biology/Genetics

Figure 15.51: **Evolution and Original of Life experimental evidence as in the beginning of the 121st century**

Remark

After Mendel's studies and discoveries more and more new discoveries about genetics were made. Mendel himself has said that the regularities he discovered apply only to the organisms and characteristics he consciously chose for his experiments. Mendel explained inheritance in terms of discrete factors - genes - that are passed along from generation to generation according to the rules of probability. Mendel's laws are valid for all sexually reproducing organisms, including garden peas and human beings. However, Mendel's laws stop short of explaining some patterns of genetic inheritance. For most sexually reproducing organisms, cases where Mendel's laws can strictly account for all patterns of inheritance are relatively rare. Often the inheritance patterns are more complex.

In cases of codominance the phenotypes produced by both alleles are clearly expressed. Mendel chose genetic traits in plants that are determined by only two alleles, such as "A" and "a". In nature, genes often exist in several different forms with multiple alleles. Furthermore, many traits are produced by the interaction of several genes. Traits controlled by two or more genes are said to be polygenic traits.

15.1.6 Growth rate with temperature

For many species (mammals, fish, micro-organisms), it is reasonable to consider as a first approximation, the relation of the population growth rate with the temperature of the environment is a second degree polynomial:

$$\mu=aT^2+bt+c \tag{15.1.179}$$

where the values a,b,c will depend of the considered species.

We know also that there is for most individuals a given optimum growth rate at a given temperature denoted T_{opt}, as well as minimum T_{min} and maximum T_{max} temperatures below and above which there is no more growth.

Therefore, the parameters a,b,c must satisfy the following relations:

$$\frac{\mathrm{d}\mu(T_{\text{opt}})}{\mathrm{d}T}=0 \tag{15.1.180}$$

$$\mu(T_{\text{min}})=\mu(T_{\text{max}})=0 \tag{15.1.181}$$

This implies for the three temperatures $T_{\text{opt}},T_{\text{min}},T_{\text{max}}$ to respect the relations proved in the context of the study of polynomials of the second degree in the section Calculus (see page 1382):

$$\begin{aligned}T_{\text{opt}}&=-\frac{b}{2a}\\T_{\text{min}}&=\frac{-b-\sqrt{b^2-4ac}}{2a}=-\frac{b}{2a}-\frac{\sqrt{b^2-4ac}}{2a}=T_{\text{opt}}-\frac{\sqrt{b^2-4ac}}{2a}\\T_{\text{max}}&=\frac{-b+\sqrt{b^2-4ac}}{2a}=-\frac{b}{2a}+\frac{\sqrt{b^2-4ac}}{2a}=T_{\text{opt}}+\frac{\sqrt{b^2-4ac}}{2a}\end{aligned} \tag{15.1.182}$$

So if we know for a living entity T_{opt} and T_{min} it is easy to deduce theoretically an approximation of T_{max} in the context of the main assumption...

Remark

Again, for people interested in exploring the subject further, we highly recommend reading [118], which is a masterpiece in this field!

~90 %

★★☆☆☆

57 votes, 60.00%

15.2. Game and Decision Theory

THE Decision and Game Theory goes far beyond the narrow confines of social games, even if they constituted the first object of study and gave it its name in most commercially available books. Furthermore, the two theories are very close to each other from which the fact that they are very often non-differentiated in the literature.

Definitions (#15.2.634):

D1. "Game theory" is the study of decision making models in uncertain future with unknown probabilities.

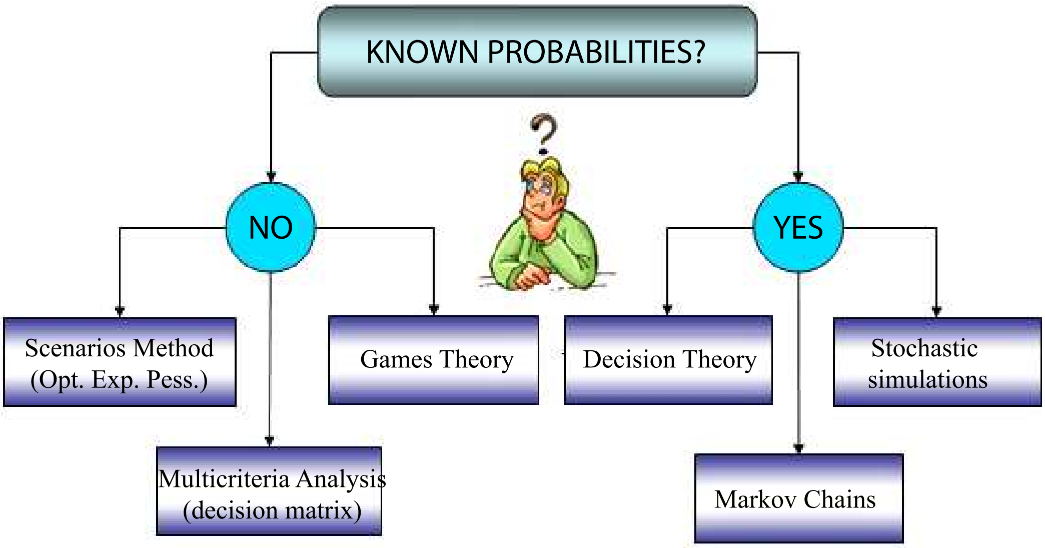
D2. "Decision theory" (sometimes also named "decision analysis") is the study of decision-making models in an uncertain probabilistic (objectively or subjectively) future.

D3. "Data drive decision making" is the practice of basing decisions on the analysis of data ("Machine Learning" or "Data Mining"), rather than purely on intuition (see section Numerical Methods page 6768).

D4. "Evidence based management" is the practice of taking decisions based on best available scientific evidence in peer-reviewed research papers (see [41] for more details).

Each of the methods of analysis of these two theories is mainly made in tabular (table) form or as a vertical or horizontal tree.

Here is a fairly familiar pattern for project manager who sums up the whole situation (data driven decision is not included as treated in the Theoretical Computing section during the study of decision trees):



KNOWN PROBABILITIES?
NO
YES
Scenarios Method (Opt. Exp. Pess.)
Games Theory
Decision Theory
Stochastic simulations
Multicriteria Analysis (decision matrix)
Markov Chains

Figure 15.52: **Elementary classification of decision techniques**

To summarize, "game theory" is the study of interdependent choice and action. It includes the study of "strategic decision making", the analysis of how the choices and decisions of a rational agent depend on (or should be influenced by) the choices of other agents, as well as the study of "group dynamics", the analysis of how the distribution of strategies in a population evolves in various contexts and how these distributions impact the outcomes of individual interactions. It should be distinguished from "decision theory", the study of individual choice in contexts where the agent is choosing independently of other agents, and from social choice theory, the study of collective decision making.

These tools aim to try to formalize to decide which configuration or decision is better than another[5]. We will look for this purpose to find the optimum of certain parameters that quantify the quality of a strategic situation. We must also determine which conditions lead to a configuration that is considered optimal.

Game theory and the decision is now quite common and used in academic circles, not only in economy (particularly corporate finance), but also by a whole class of other science in which the study on conflictual situations is relevant: sociology, biology, evolution, computer (video games), marketing, etc.

Remark

In the world of corporations, and especially in Europe, decisions techniques are unknown to almost all of the leaders whose choices are often more qualitative and instinctive than scientific and thus less accurate... That's why "Evidence based Management" has been developed. Most management decisions are not based on the best available evidence. Indeed some practitioners often prefer to base decisions solely on their judgement based on personal experience. However, personal judgement alone is not a very reliable source of evidence because it is highly susceptible to systematic errors.

We will try, as always in this textbook, to minimize as possible the number of definitions and concepts in order not to drown the rigour of mathematical analysis in the chaos of useless vocabulary and not necessary for such analysis (and in the field of game theory it is otherwise a bit like in graph theory... a real nightmare of definitions!).

Definition (naive #635): A "game" is a situation where the players are driven to make strategic choices among a number of possible actions, within the framework defined in advance by the "rules of the game", the result of these choices constituting an "outcome of the game", which is associated with a "gain" (or payment), positive or negative, for each player.

Remark

A player may be a person, a group of people, a society, a region, a political party, a country or even mother Nature ...

Assumptions (we will find them again in the Econometry section):

A1. The market is governed by competition and cooperation

---

[5] For example, a survey of 950 US HR practitioners showed large discrepancies between what practitioners think is effective and what the current scientific research shows [474]. This study has been repeated in other countries with similar findings.

A2. The behaviour of economic agents are rational (...)

A3. It is possible to formalize competitive behaviour

A4. All competitive phenomena have a utilitarian dimension

We differentiate and define mainly six types of games which we will formalize further below.

Definitions (#15.2.636):

D1. The "cooperative or non-cooperative games": a game is said to be cooperative when players can freely communicate and make agreements (e.g. in the form of a contract). They thus form a coalition and seek the general interest followed by a partition of earnings between all players. In a non-cooperative game, players (who do not communicate or can not communicate with one another) act according to the principle of economic rationality: each seeks to make the best decisions for themselves (i.e. seeks to maximize selfishly its individual earnings). This type of game involves probabilities.

A first possible approach (without using math at first) of this game spirit is accessible to young children (without they know it!). Indeed, let us consider the following example:

✍**Example:**

Imagine two children, both gourmands in the presence of a homogeneous cake, perfectly divisible (and very good ...). If mom does two parts, there will inevitably disputes, each one finding the part of the other bigger. The only way (expected order) to avoid any argument is for the mother to impose the following rule: one of the children makes the parts, and the other one chooses first the part. The one who cuts can therefore not reason anymore taking into account only his own preferences, which would push him to cut a big part. He knows that the other will choose it. If then he cut a larger part than the other, it could find the latter in the neighbour's plate. He will therefore try to cut as evenly as possible in his point of view. Thus, whatever the choice of the other, he will think that the game is strategically equilibrated. It is this anticipation of the choice of the other decision which makes the originality of the theory of decision and also of cooperation!

D2. The "simultaneous or sequential move games": Simultaneous games are the one in which the move of two players (the strategy adopted by two players) is simultaneous. In simultaneous move, players do not have knowledge about the move of other players. On the contrary, sequential games are the one in which players are aware about the moves of players who have already adopted a strategy. However, in sequential games, the players do not have a deep knowledge about the strategies of other players. For example, a player has knowledge that the other player would not use a single strategy, but he/she is not sure about the number of strategies the other player may use.

D3. The "zero-sum or non-zero game": a game is said to be "zero sum" when the sum of the earnings of players is constant (or by the subtle choice of a utility function may be constant...) or in other words: what one wins is necessarily lost by another (chess, poker and some say that the stock market is a zero sum game but in reality it is false when you think about it globally ...). The social games are often zero-sum games but the real situations are often better described by the non-cooperative games with non-zero sum because some issues are profitable for all, or harmful for all (political, business situations...). For example slavery is a non-zero sum game.

Remarks

R1. Some theorists criticize zero-sum games, at least in the area of economic situation, on the basis that economic exchange is in principle mutually beneficial and that the zero-sum games would be totally unrealistic.

R2. Zero sum games are sometimes named "antagonistic games".

R3. Since the invention of the atomic bomb, the balance of terror is based on the doctrine of offensive deterrence. Opposed by the reciprocal ability to inflict massive damage, the respective nuclear arsenals self-cancelling in a zero-sum game, by a principle of mutually assured destruction.

D4. The "with and without balance games": a cooperative non zero sum game is said with "Nash equilibrium" if there are a couple of strategies (in the case of a two-player game) such that no player has the interest to unilaterally change his strategy to ensure the maximum minimum (the "maximin") earnings.

D5. The "competitive or non-competitive games": a non-competitive game is the opposite of a competitive game such as by definition, when any pair of strategies (in the case of a two-player game) is such that it makes win or lose all players simultaneously a given gain (i.e. when I lose something, you lose something, when I win something you also win something).

D6. The "symmetric and asymmetric games": in symmetric games. strategies adopted by all players are same. Symmetry can exist in short-term games only because in long-term games the number of options with a player increases. The decisions in a symmetric game depend on the strategies used, not on the players of the game. Even in case of interchanging players, the decisions remain the same in symmetric games. Example of symmetric games is prisoner's dilemma. On the other hand, asymmetric games are the one in which strategies adopted by players are different. In asymmetric games, the strategy that provides benefit to one player may not be equally beneficial for the other player.

Most of these types of games are interlaced and therefore it is quite hard to created an orgchart of the existing ones. A quite good illustration that summarize the six of them and the fact that they are interconnect is given by the following smart art:

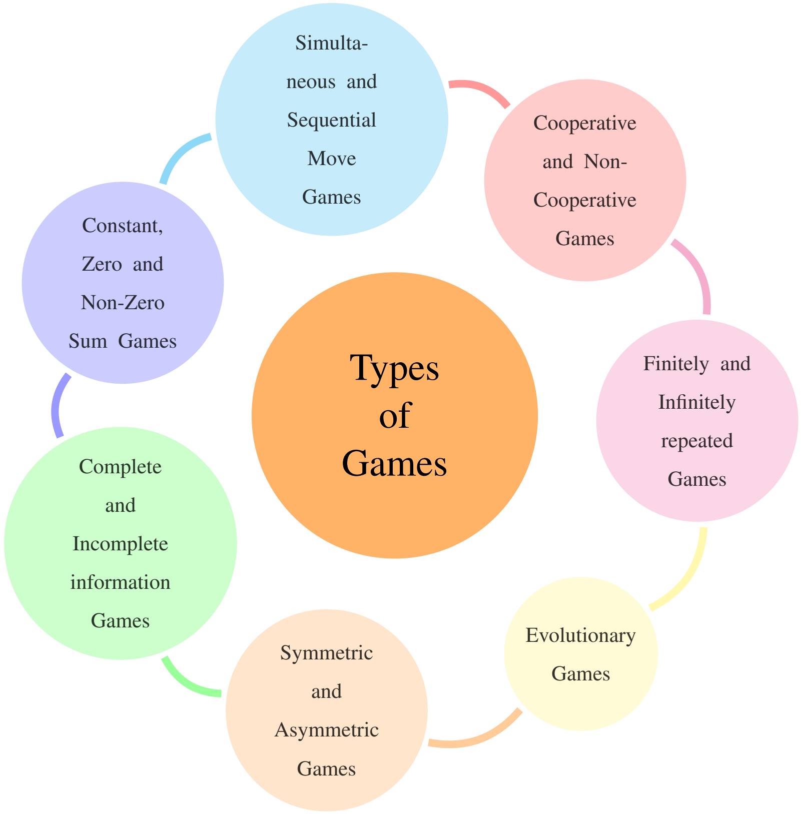

Simultaneous and Sequential Move Games
Cooperative and Non-Cooperative Games
Constant, Zero and Non-Zero Sum Games
Types of Games
Finitely and Infinitely repeated Games
Complete and Incomplete information Games
Evolutionary Games
Symmetric and Asymmetric Games

Figure 15.53: **Common types of games in Game Theory**

Scientists (and afterwards politicians, based on scientists recommendations) have to take sometimes decisions that are no popular! Indeed, a non-negligible percentage of humans are sadly totally irrationals (see for example the toilet paper shortage from year 12020 - according to holocene calendar - during the COVID-19 pandemic). That's why sometimes scientists have to recommend the retirement of some freely accessible medications because in a crisis situation a fraction of the population may create an irrational shortage of a product and even worst... use that same product in a way that may do more harm than good!

quote 15.2:7380.#79

> *Sooner or later the moment arrives when a decision must be made which is neither certain, neither suitable, nor popular, but it must be taken, because it is right.*
>
> Martin Luther King

quote 15.2:7380.#80

> *A good compromise is when both parties are dissatisfied.*
>
> Larry David

15.2.1 Behavioural decision bias (cognitive bias)

Before starting with technical and mathematics techniques it is important for the manager, top manager, CEO/CFO and any other CXO (or even the student in Economy) to understand why it is more important to rely in mathematical results rather than to trust on human intuition and to try not forget what follows as most of time I see huge error of evaluation in Fortune 500 companies!

First, for information, the study of human intuition in economy is named "behavioural economics". And in this study field "cognitive biases" are tendencies by humans to think in certain ways that can lead to systematic deviations from a standard of rationality or good judgement, and are often studied in psychology and behavioural economics.

In the early 11970s (holocene calendar), Amos Tversky and Daniel Kahneman introduced the term "cognitive bias" to describe people's systematic but purportedly flawed patterns of responses to judgement and decision problems (managers, students, economists, and so on...). Indeed, they have showed that people routinely employ heuristics - rules of thumb, or mental shortcuts - to simplify and, worse, oversimplify decisions under uncertainty. Moreover, they have shown time and again that our choices are frequently skewed by an array of cognitive biases.

Most work before the 11970 (holocene calendar) were done in behavioural finance and focused on asset pricing and the behaviour of investors. But increasingly, attention is being paid to decision-making in the corporate realm. Because of their training and experience, managers might be presumed to be less likely to use mental shortcuts, and less vulnerable to cognitive biases. True or not, consultants in decision analysis have made a good living by showing managers how they fall into decision traps, and professors have delighted in showing their executive MBA students just how flawed their judgement can be.

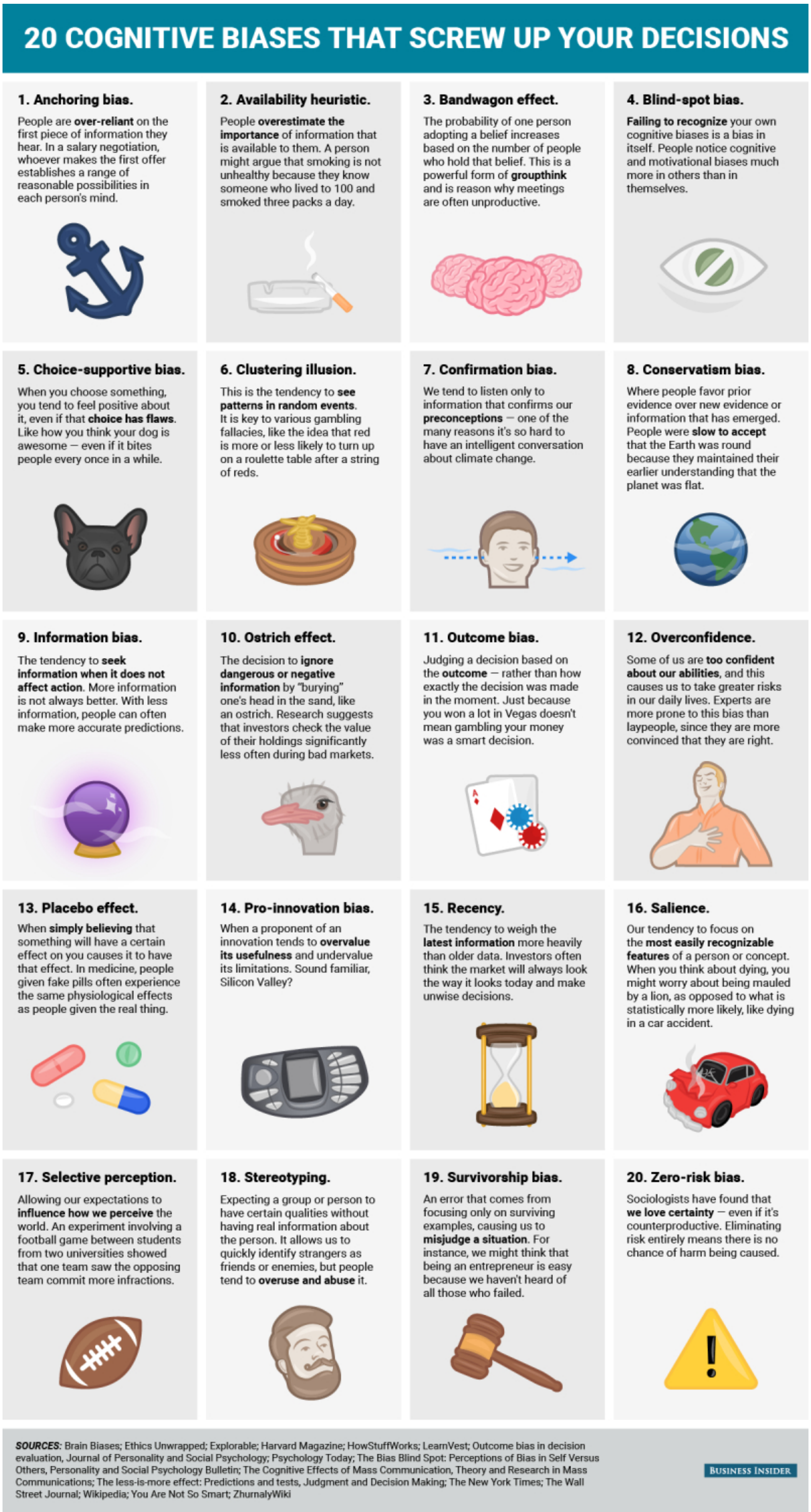
On the way it is also important to distinguish between "cognitive biases" and "logical fallacies". A logical fallacy is an error in logical argumentation (e.g. ad hominem attacks, slippery slopes, circular arguments, appeal to force, etc.). A cognitive bias, on the other hand, is a genuine deficiency or limitation in our thinking — a flaw in judgement that arises from errors of memory, social attribution, and miscalculations (such as statistical errors or a false sense of probability[6]).

Although the reality of these biases is confirmed by replicable research, there are often controversies about how to classify these biases or how to explain them. Some are effects of information-processing rules (i.e., mental shortcuts), named "heuristics", that the brain uses to produce decisions or judgements. Such effects are named "cognitive biases". Biases have a variety of forms and appear as cognitive ("cold") bias, such as mental noise, or motivational ("hot") bias, such as when beliefs are distorted by wishful thinking. Both effects can be present at the same time.

Over time, the behaviourists have compiled a long list of biases and heuristics. No one can say with certainty which of these inflict the most harm, but financial managers would do well to watch out for five: anchoring and adjustment, framing, optimism, overconfidence, and self-serving bias. In the figure below the reader can see a quite good summary illustration (the original author is sadly unknown) of the most common well known cognitive biases:

---

[6] These latter being often the most difficult point during debates when people don't understand that when they use arguments, they have the **burden of proof**, and even more difficult when they don't understand the concept of "proof"!!!!



20 COGNITIVE BIASES THAT SCREW UP YOUR DECISIONS

1. Anchoring bias.
People are **over-reliant** on the first piece of information they hear. In a salary negotiation, whoever makes the first offer establishes a range of reasonable possibilities in each person's mind.

2. Availability heuristic.
People **overestimate the importance** of information that is available to them. A person might argue that smoking is not unhealthy because they know someone who lived to 100 and smoked three packs a day.

3. Bandwagon effect.
The probability of one person adopting a belief increases based on the number of people who hold that belief. This is a powerful form of **groupthink** and is reason why meetings are often unproductive.

4. Blind-spot bias.
Failing to recognize your own cognitive biases is a bias in itself. People notice cognitive and motivational biases much more in others than in themselves.

5. Choice-supportive bias.
When you choose something, you tend to feel positive about it, even if that **choice has flaws**. Like how you think your dog is awesome — even if it bites people every once in a while.

6. Clustering illusion.
This is the tendency to **see patterns in random events**. It is key to various gambling fallacies, like the idea that red is more or less likely to turn up on a roulette table after a string of reds.

7. Confirmation bias.
We tend to listen only to information that confirms our **preconceptions** — one of the many reasons it's so hard to have an intelligent conversation about climate change.

8. Conservatism bias.
Where people favor prior evidence over new evidence or information that has emerged. People were **slow to accept** that the Earth was round because they maintained their earlier understanding that the planet was flat.

9. Information bias.
The tendency to **seek information when it does not affect action**. More information is not always better. With less information, people can often make more accurate predictions.

10. Ostrich effect.
The decision to **ignore dangerous or negative information** by "burying" one's head in the sand, like an ostrich. Research suggests that investors check the value of their holdings significantly less often during bad markets.

11. Outcome bias.
Judging a decision based on **the outcome** — rather than how exactly the decision was made in the moment. Just because you won a lot in Vegas doesn't mean gambling your money was a smart decision.

12. Overconfidence.
Some of us are **too confident about our abilities**, and this causes us to take greater risks in our daily lives. Experts are more prone to this bias than laypeople, since they are more convinced that they are right.

13. Placebo effect.
When **simply believing** that something will have a certain effect on you causes it to have that effect. In medicine, people given fake pills often experience the same physiological effects as people given the real thing.

14. Pro-innovation bias.
When a proponent of an innovation tends to **overvalue its usefulness** and undervalue its limitations. Sound familiar, Silicon Valley?

15. Recency.
The tendency to weigh the **latest information** more heavily than older data. Investors often think the market will always look the way it looks today and make unwise decisions.

16. Salience.
Our tendency to focus on the **most easily recognizable features** of a person or concept. When you think about dying, you might worry about being mauled by a lion, as opposed to what is statistically more likely, like dying in a car accident.

17. Selective perception.
Allowing our expectations to **influence how we perceive** the world. An experiment involving a football game between students from two universities showed that one team saw the opposing team commit more infractions.

18. Stereotyping.
Expecting a group or person to have certain qualities without having real information about the person. It allows us to quickly identify strangers as friends or enemies, but people tend to **overuse and abuse** it.

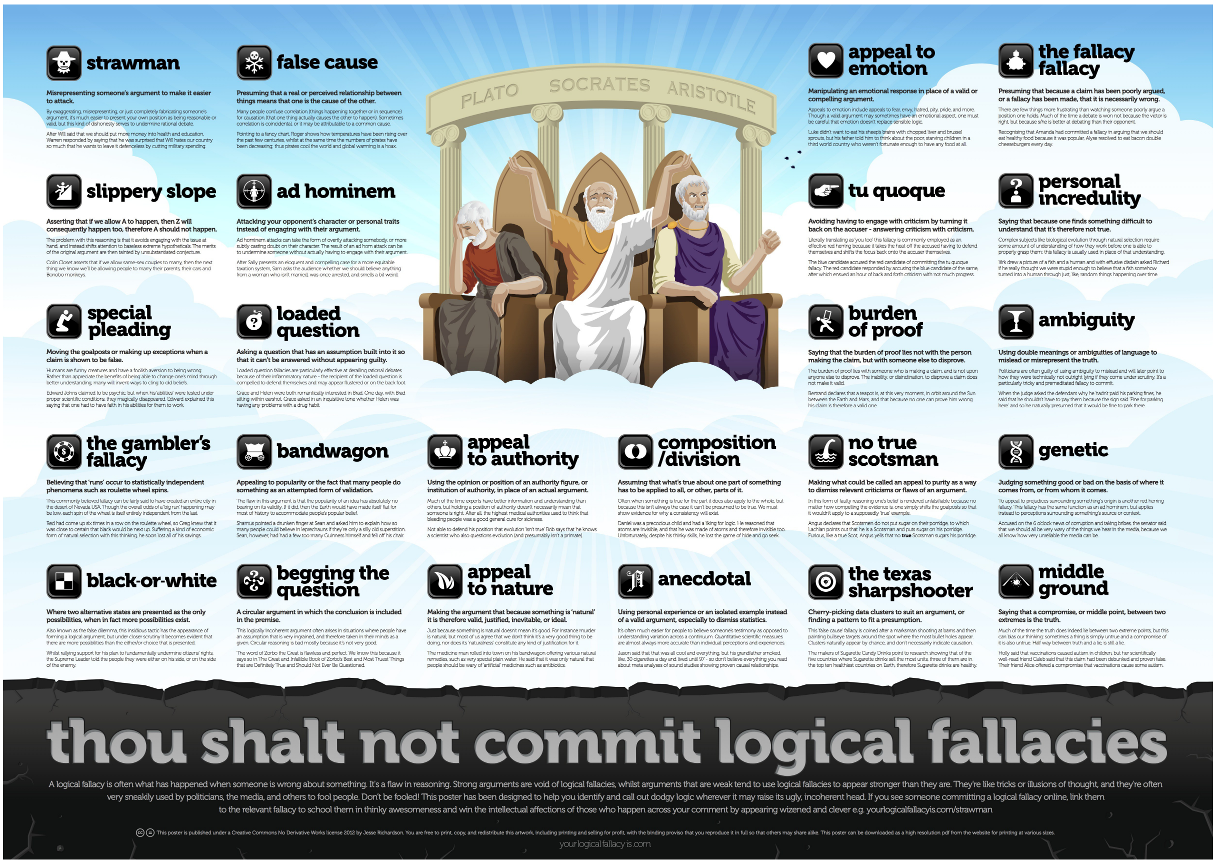
19. Survivorship bias.
An error that comes from focusing only on surviving examples, causing us to **misjudge a situation**. For instance, we might think that being an entrepreneur is easy because we haven't heard of all those who failed.

20. Zero-risk bias.
Sociologists have found that **we love certainty** — even if it's counterproductive. Eliminating risk entirely means there is no chance of harm being caused.

SOURCES: Brain Biases; Ethics Unwrapped; Explorable; Harvard Magazine; HowStuffWorks; LearnVest; Outcome bias in decision evaluation, Journal of Personality and Social Psychology; Psychology Today; The Bias Blind Spot: Perceptions of Bias in Self Versus Others, Personality and Social Psychology Bulletin; The Cognitive Effects of Mass Communication, Theory and Research in Mass Communications; The less-is-more effect: Predictions and tests, Judgment and Decision Making; The New York Times; The Wall Street Journal; Wikipedia; You Are Not So Smart; ZhurnalyWiki

BUSINESS INSIDER

Also here is another useful and nice complementary poster that will be appreciated by any scientist or any person that has for purpose to reason following some good practices:



strawman

Misrepresenting someone's argument to make it easier to attack.

By exaggerating, misrepresenting, or just completely fabricating someone's argument, it's much easier to present your own position as being reasonable or valid, but this kind of dishonesty serves to undermine rational debate.

After Will said that we should put more money into health and education, Warren responded by saying that he was surprised that Will hates our country so much that he wants to leave it defenceless by cutting military spending.

false cause

Presuming that a real or perceived relationship between things means that one is the cause of the other.

Many people confuse correlation (things happening together or in sequence) for causation (that one thing actually causes the other to happen). Sometimes correlation is coincidental, or it may be attributable to a common cause.

Pointing to a fancy chart, Roger shows how temperatures have been rising over the past few centuries, whilst at the same time the numbers of pirates have been decreasing; thus pirates cool the world and global warming is a hoax.

appeal to emotion

Manipulating an emotional response in place of a valid or compelling argument.

Appeals to emotion include appeals to fear, envy, hatred, pity, pride, and more. Though a valid argument may sometimes have an emotional aspect, one must be careful that emotion doesn't replace sensible logic.

Luke didn't want to eat his sheep's brains with chopped liver and brussel sprouts, but his father told him to think about the poor, starving children in a third world country who weren't fortunate enough to have any food at all.

the fallacy fallacy

Presuming that because a claim has been poorly argued, or a fallacy has been made, that it is necessarily wrong.

There are few things more frustrating than watching someone poorly argue a position one holds. Much of the time a debate is won not because the victor is right, but because s/he is better at debating than their opponent.

Recognising that Amanda had committed a fallacy in arguing that we should eat healthy food because it was popular, Alyse resolved to eat bacon double cheeseburgers every day.

slippery slope

Asserting that if we allow A to happen, then Z will consequently happen too, therefore A should not happen.

The problem with this reasoning is that it avoids engaging with the issue at hand, and instead shifts attention to baseless extreme hypotheticals. The merits of the original argument are then tainted by unsubstantiated conjecture.

Colin Closet asserts that if we allow same-sex couples to marry, then the next thing we know we'll be allowing people to marry their parents, their cars and Bonobo monkeys.

ad hominem

Attacking your opponent's character or personal traits instead of engaging with their argument.

Ad hominem attacks can take the form of overtly attacking somebody, or more subtly casting doubt on their character. The result of an ad hom attack can be to undermine someone without actually having to engage with their argument.

After Sally presents an eloquent and compelling case for a more equitable taxation system, Sam asks the audience whether we should believe anything from a woman who isn't married, was once arrested, and smells a bit weird.

tu quoque

Avoiding having to engage with criticism by turning it back on the accuser - answering criticism with criticism.

Literally translating as 'you too' this fallacy is commonly employed as an effective red herring because it takes the heat off the accused having to defend themselves and shifts the focus back onto the accuser themselves.

The blue candidate accused the red candidate of committing the tu quoque fallacy. The red candidate responded by accusing the blue candidate of the same, after which ensued an hour of back and forth criticism with not much progress.

personal incredulity

Saying that because one finds something difficult to understand that it's therefore not true.

Complex subjects like biological evolution through natural selection require some amount of understanding of how they work before one is able to properly grasp them; this fallacy is usually used in place of that understanding.

Kirk drew a picture of a fish and a human and with effusive disdain asked Richard if he really thought we were stupid enough to believe that a fish somehow turned into a human through just, like, random things happening over time.

special pleading

Moving the goalposts or making up exceptions when a claim is shown to be false.

Humans are funny creatures and have a foolish aversion to being wrong. Rather than appreciate the benefits of being able to change one's mind through better understanding, many will invent ways to cling to old beliefs.

Edward Johns claimed to be psychic, but when his 'abilities' were tested under proper scientific conditions, they magically disappeared. Edward explained this saying that one had to have faith in his abilities for them to work.

loaded question

Asking a question that has an assumption built into it so that it can't be answered without appearing guilty.

Loaded question fallacies are particularly effective at derailing rational debates because of their inflammatory nature - the recipient of the loaded question is compelled to defend themselves and may appear flustered or on the back foot.

Grace and Helen were both romantically interested in Brad. One day, with Brad sitting within earshot, Grace asked in an inquisitive tone whether Helen was having any problems with a drug habit.

burden of proof

Saying that the burden of proof lies not with the person making the claim, but with someone else to disprove.

The burden of proof lies with someone who is making a claim, and is not upon anyone else to disprove. The inability, or disinclination, to disprove a claim does not make it valid.

Bertrand declares that a teapot is, at this very moment, in orbit around the Sun between the Earth and Mars, and that because no one can prove him wrong his claim is therefore a valid one.

ambiguity

Using double meanings or ambiguities of language to mislead or misrepresent the truth.

Politicians are often guilty of using ambiguity to mislead and will later point to how they were technically not outright lying if they come under scrutiny. It's a particularly tricky and premeditated fallacy to commit.

When the judge asked the defendant why he hadn't paid his parking fines, he said that he shouldn't have to pay them because the sign said 'Fine for parking here' and so he naturally presumed that it would be fine to park there.

the gambler's fallacy

Believing that 'runs' occur to statistically independent phenomena such as roulette wheel spins.

This commonly believed fallacy can be fairly said to have created an entire city in the desert of Nevada USA. Though the overall odds of a 'big run' happening may be low, each spin of the wheel is itself entirely independent from the last.

Red had come up six times in a row on the roulette wheel, so Greg knew that it was close to certain that black would be next up. Suffering a kind of economic form of natural selection with this thinking, he soon lost all of his savings.

bandwagon

Appealing to popularity or the fact that many people do something as an attempted form of validation.

The flaw in this argument is that the popularity of an idea has absolutely no bearing on its validity. If it did, then the Earth would have made itself flat for most of history to accommodate people's popular belief.

Shamus pointed a drunken finger at Sean and asked him to explain how so many people could believe in leprechauns if they're only a silly old superstition. Sean, however, had had a few too many Guinness himself and fell off his chair.

appeal to authority

Using the opinion or position of an authority figure, or institution of authority, in place of an actual argument.

Much of the time experts have better information and understanding than others, but holding a position of authority doesn't necessarily mean that someone is right. After all, the highest medical authorities used to think that bleeding people was a good general cure for sickness.

Not able to defend his position that evolution 'isn't true' Bob says that he knows a scientist who also questions evolution (and presumably isn't a primate).

composition /division

Assuming that what's true about one part of something has to be applied to all, or other, parts of it.

Often when something is true for the part it does also apply to the whole, but because this isn't always the case it can't be presumed to be true. We must show evidence for why a consistency will exist.

Daniel was a precocious child and had a liking for logic. He reasoned that atoms are invisible, and that he was made of atoms and therefore invisible too. Unfortunately, despite his thinky skills, he lost the game of hide and go seek.

no true scotsman

Making what could be called an appeal to purity as a way to dismiss relevant criticisms or flaws of an argument.

In this form of faulty reasoning one's belief is rendered unfalsifiable because no matter how compelling the evidence is, one simply shifts the goalposts so that it wouldn't apply to a supposedly 'true' example.

Angus declares that Scotsmen do not put sugar on their porridge, to which Lachlan points out that he is a Scotsman and puts sugar on his porridge. Furious, like a true Scot, Angus yells that no **true** Scotsman sugars his porridge.

genetic

Judging something good or bad on the basis of where it comes from, or from whom it comes.

To appeal to prejudices surrounding something's origin is another red herring fallacy. This fallacy has the same function as an ad hominem, but applies instead to perceptions surrounding something's source or context.

Accused on the 6 o'clock news of corruption and taking bribes, the senator said that we should all be very wary of the things we hear in the media, because we all know how very unreliable the media can be.

black-or-white

Where two alternative states are presented as the only possibilities, when in fact more possibilities exist.

Also known as the false dilemma, this insidious tactic has the appearance of forming a logical argument, but under closer scrutiny it becomes evident that there are more possibilities than the either/or choice that is presented.

Whilst rallying support for his plan to fundamentally undermine citizens' rights, the Supreme Leader told the people they were either on his side, or on the side of the enemy.

begging the question

A circular argument in which the conclusion is included in the premise.

This logically incoherent argument often arises in situations where people have an assumption that is very ingrained, and therefore taken in their minds as a given. Circular reasoning is bad mostly because it's not very good.

The word of Zorbo the Great is flawless and perfect. We know this because it says so in The Great and Infallible Book of Zorbo's Best and Most Truest Things that are Definitely True and Should Not Ever Be Questioned.

appeal to nature

Making the argument that because something is 'natural' it is therefore valid, justified, inevitable, or ideal.

Just because something is natural doesn't mean it's good. For instance murder is natural, but most of us agree that we don't think it's a very good thing to be doing, nor does its 'naturalness' constitute any kind of justification for it.

The medicine man rolled into town on his bandwagon offering various natural remedies, such as very special plain water. He said that it was only natural that people should be wary of 'artificial' medicines such as antibiotics.

anecdotal

Using personal experience or an isolated example instead of a valid argument, especially to dismiss statistics.

It's often much easier for people to believe someone's testimony as opposed to understanding variation across a continuum. Quantitative scientific measures are almost always more accurate than individual perceptions and experiences.

Jason said that that was all cool and everything, but his grandfather smoked, like, 30 cigarettes a day and lived until 97 - so don't believe everything you read about meta analyses of sound studies showing proven causal relationships.

the texas sharpshooter

Cherry-picking data clusters to suit an argument, or finding a pattern to fit a presumption.

This 'false cause' fallacy is coined after a marksman shooting at barns and then painting bullseye targets around the spot where the most bullet holes appear. Clusters naturally appear by chance, and don't necessarily indicate causation.

The makers of Sugarette Candy Drinks point to research showing that of the five countries where Sugarette drinks sell the most units, three of them are in the top ten healthiest countries on Earth, therefore Sugarette drinks are healthy.

middle ground

Saying that a compromise, or middle point, between two extremes is the truth.

Much of the time the truth does indeed lie between two extreme points, but this can bias our thinking: sometimes a thing is simply untrue and a compromise of it is also untrue. Half way between truth and a lie, is still a lie.

Holly said that vaccinations caused autism in children, but her scientifically well-read friend Caleb said that this claim had been debunked and proven false. Their friend Alice offered a compromise that vaccinations cause some autism.

thou shalt not commit logical fallacies

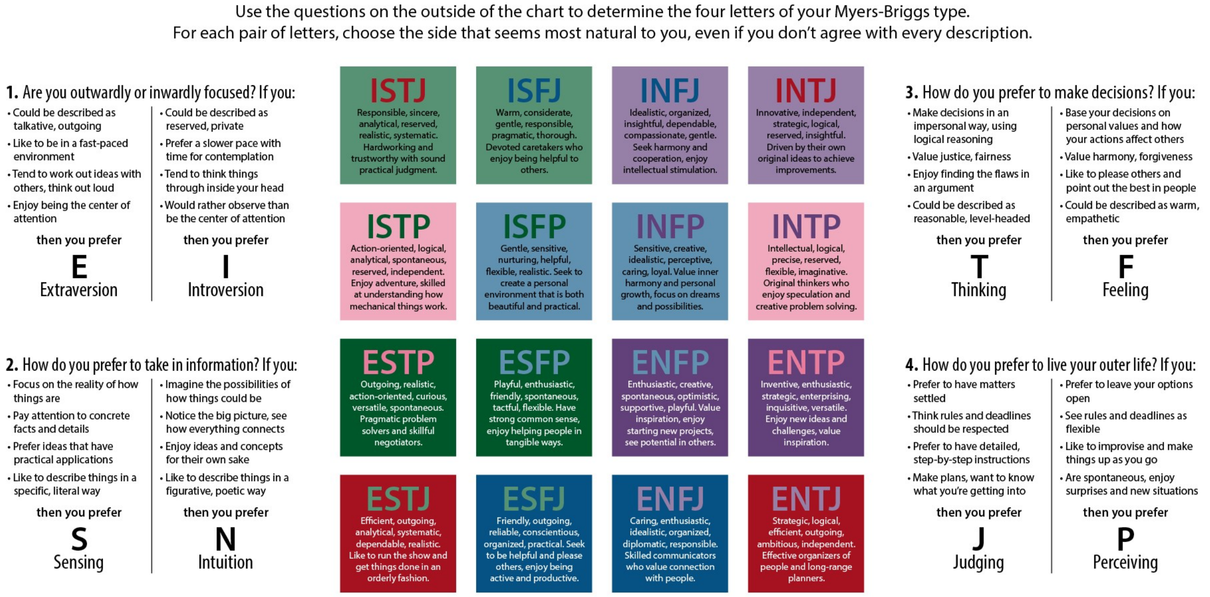
A logical fallacy is often what has happened when someone is wrong about something. It's a flaw in reasoning. Strong arguments are void of logical fallacies, whilst arguments that are weak tend to use logical fallacies to appear stronger than they are. They're like tricks or illusions of thought, and they're often very sneakily used by politicians, the media, and others to fool people. Don't be fooled! This poster has been designed to help you identify and call out dodgy logic wherever it may raise its ugly, incoherent head. If you see someone committing a logical fallacy online, link them to the relevant fallacy to school them in thinky awesomeness and win the intellectual affections of those who happen across your comment by appearing wizened and clever e.g. yourlogicalfallacyis.com/strawman

This poster is published under a Creative Commons No Derivative Works license 2012 by Jesse Richardson. You are free to print, copy, and redistribute this artwork, including printing and selling for profit, with the binding proviso that you reproduce it in full so that others may share alike. This poster can be downloaded as a high resolution pdf from the website for printing at various sizes.

yourlogicalfallacyis.com

Figure 15.54: Some famous logical fallacies

And apart from the fact that you can detect logical fallacies used by people, you can also categorize them quite well using the Myers-Briggs classification model:



Use the questions on the outside of the chart to determine the four letters of your Myers-Briggs type.
For each pair of letters, choose the side that seems most natural to you, even if you don't agree with every description.

1. Are you outwardly or inwardly focused? If you:

• Could be described as talkative, outgoing • Like to be in a fast-paced environment • Tend to work out ideas with others, think out loud • Enjoy being the center of attention	• Could be described as reserved, private • Prefer a slower pace with time for contemplation • Tend to think things through inside your head • Would rather observe than be the center of attention
then you prefer **E** Extraversion	**then you prefer** **I** Introversion

2. How do you prefer to take in information? If you:

• Focus on the reality of how things are • Pay attention to concrete facts and details • Prefer ideas that have practical applications • Like to describe things in a specific, literal way	• Imagine the possibilities of how things could be • Notice the big picture, see how everything connects • Enjoy ideas and concepts for their own sake • Like to describe things in a figurative, poetic way
then you prefer **S** Sensing	**then you prefer** **N** Intuition

ISTJ Responsible, sincere, analytical, reserved, realistic, systematic. Hardworking and trustworthy with sound practical judgment.	**ISFJ** Warm, considerate, gentle, responsible, pragmatic, thorough. Devoted caretakers who enjoy being helpful to others.	**INFJ** Idealistic, organized, insightful, dependable, compassionate, gentle. Seek harmony and cooperation, enjoy intellectual stimulation.	**INTJ** Innovative, independent, strategic, logical, reserved, insightful. Driven by their own original ideas to achieve improvements.
ISTP Action-oriented, logical, analytical, spontaneous, reserved, independent. Enjoy adventure, skilled at understanding how mechanical things work.	**ISFP** Gentle, sensitive, nurturing, helpful, flexible, realistic. Seek to create a personal environment that is both beautiful and practical.	**INFP** Sensitive, creative, idealistic, perceptive, caring, loyal. Value inner harmony and personal growth, focus on dreams and possibilities.	**INTP** Intellectual, logical, precise, reserved, flexible, imaginative. Original thinkers who enjoy speculation and creative problem solving.
ESTP Outgoing, realistic, action-oriented, curious, versatile, spontaneous. Pragmatic problem solvers and skillful negotiators.	**ESFP** Playful, enthusiastic, friendly, spontaneous, tactful, flexible. Have strong common sense, enjoy helping people in tangible ways.	**ENFP** Enthusiastic, creative, spontaneous, optimistic, supportive, playful. Value inspiration, enjoy starting new projects, see potential in others.	**ENTP** Inventive, enthusiastic, strategic, enterprising, inquisitive, versatile. Enjoy new ideas and challenges, value inspiration.
ESTJ Efficient, outgoing, analytical, systematic, dependable, realistic. Like to run the show and get things done in an orderly fashion.	**ESFJ** Friendly, outgoing, reliable, conscientious, organized, practical. Seek to be helpful and please others, enjoy being active and productive.	**ENFJ** Caring, enthusiastic, idealistic, organized, diplomatic, responsible. Skilled communicators who value connection with people.	**ENTJ** Strategic, logical, efficient, outgoing, ambitious, independent. Effective organizers of people and long-range planners.

3. How do you prefer to make decisions? If you:

• Make decisions in an impersonal way, using logical reasoning • Value justice, fairness • Enjoy finding the flaws in an argument • Could be described as reasonable, level-headed	• Base your decisions on personal values and how your actions affect others • Value harmony, forgiveness • Like to please others and point out the best in people • Could be described as warm, empathetic
then you prefer **T** Thinking	**then you prefer** **F** Feeling

4. How do you prefer to live your outer life? If you:

• Prefer to have matters settled • Think rules and deadlines should be respected • Prefer to have detailed, step-by-step instructions • Make plans, want to know what you're getting into	• Prefer to leave your options open • See rules and deadlines as flexible • Like to improvise and make things up as you go • Are spontaneous, enjoy surprises and new situations
then you prefer **J** Judging	**then you prefer** **P** Perceiving

Figure 15.55: Myers-Briggs personality type classification

So regarding everything we have seen so far in this textbook, we can give an attempt of definition of what may be considered as "intelligence".

Definition (#15.2.637): "Intelligence" is the general mental state that involves the aptitude to reason, plan, create, solve problems, think abstractly, comprehend complex idea and new informations, evaluate the relative value of things, identify what is morally rational and responsible and continuously learn from experience recognizing the different levels of expertise (knowledge/ignorance) in a field and being aware of cognitive biases while remaining humble and likely neutral. Intelligent individuals are recognized by being able to accept to change their opinions based on factual new proofs/evidences (and admit that there were wrong), they can also distinguish rigorously the concepts and different levels of proof/evidence/claim and handle incertitudes and unknowns quantitatively by rejecting beliefs, they also have the characteristic to be cautious about all their opinions as they are aware of the Dunning–Kruger bias. It is not merely book learning, a narrow academic skill, or test-taking smarts. Rather, it reflects a broader and deeper capability for comprehending surroundings - "catching on", "making sense" of things, or "figuring out" what to do and to say.

> ❞ *The whole trouble with the world is that the stupid are cocksure, and the intelligent are full of doubt.* ❞
>
> Bertrand Russell

quote 15.2:7385.#81

The paper [381] shows how cooperation may have created evolutionary selection pressure for greater intelligence.

Notice that physicists are necessary because they have learned during their curriculum to distrust any argument based on common sense and intuition. "Common sense" is an individual perception based on personal experience and frequently invalid when applied outside the scope of the personal history. People that claim having common sense often have the least of it! Common sense being an excuse to not thing similar to the one of the "god of the gaps". "Common sense" and "everybody knows" are a couple of phrases for people with integrity to be wary of especially when coming from HIPPO (the Highest Paid Person's Opinion in an organization)...

> ❞ *We measure the degree of intelligence of an individual by the quantity of incertitudes he can handle.* ❞
>
> Emmanuel Kant

quote 15.2:7385.#82

Remark

And keep in mind.... the fact that you (as reader of this textbook) or other people are unable to grasp science and maths is obviously not a valid argument against it.

quote 15.2:7386.#83

> ❞ *Intelligence is not what we know but what we do when we don't know.* ❞
>
> Jean Piaget

Notice that an important another bias in modern physics (years 11920 to this actual early 121st century according to holocene calendar) is the general topic of naturalness and beauty of physics equations. This conflation of beauty between arts and science probably helps to sell books. And it's not so good to seek for beauty in science because we have not yet any evidence that sciences must be beautiful (there are plenty of equations that are awful anyway!). So this is very unfortunate because it's far away from reality, the reality of a decent, honest physicist.

quote 15.2:7386.#84

> ❞ *Not doubt, certainty is what drives one insane.* ❞
>
> Friedrich Nietzsche

Here we will focus however only on two important bias for business: the "sunk cost" bias and the "anchor bias".

15.2.1.1 Sunk Cost

Definition (#15.2.638): A "sunk cost" is a cost that has already been incurred and thus cannot be recovered. A sunk cost differs from future costs that a business may face, such as decisions about inventory purchase costs or product pricing. Sunk costs (past costs) are excluded from future business decisions, because the cost will be the same regardless of the outcome of a decision.

The sunk cost fallacy is in game theory sometimes known as the "Concorde Fallacy", referring to the fact that the British and French governments continued to fund the joint development of Concorde even after it became apparent that there was no longer an economic case for the aircraft. The project was regarded privately by the British government as a "commercial disaster" which should never have been started and was almost cancelled, but political and legal issues had ultimately made it impossible for either government to pull out.

The sunk cost fallacy is the theory that continuing to put money into a project or other investment that is failing is worthwhile because of the expense that has already been spent on that investment. This is a problem because each injection of capital or investment money should be judged in terms of the likely returns on that money, not on costs previously accrued that may already be lost.

The sunk cost dilemma with its sequence of good decisions should not be confused with the sunk cost fallacy, where a misconception of sunk costs can lead to bad decisions. Sunk-cost fallacy occurs when people make decisions about a current situation based on what they have previously invested in the situation. For example, spending $100 on a concert and on the day you find that it's cold and rainy. You feel that if you don't go you would've wasted the money and the time you spent in line to get that ticket and feel obligated to follow through even if you don't want to.

When making business decisions, organizations consider relevant costs, which include the future costs and revenue of one choice compared with another. To make an informed decision, a business only considers the costs and revenue that will change as a result of the decision; sunk costs that do not change are not considered.

✍**Example:**

In a study of 96 business students in 11976[a] (sample size that seems to me far too small for such a study after having done the sample size calculation by hand...) had a choice between making an R&D investment either in an underperforming company department, or in other sections of the hypothetical company. The participants were divided into two groups: a low responsibility condition and a high responsibility condition. In the high responsibility condition, the participants were told that they, as manager, had made an earlier, disappointing R&D investment. In the low responsibility condition, subjects were told that a former manager had made a previous R&D investment in the underperforming division and were given the same profit data as the other group. In both cases subjects were then asked to make a new $ 20 million investment. There was a significant interaction between assumed responsibility and average investment, with the high responsibility condition averaging $ 12.97 million and the low condition averaging $ 9.43 million.

Similar results seems to have been obtained in earlier studies (all years are given according to holocene calendar) by Staw, Arkes and Blumer (11985[b]) and Whyte (11986[c]) but as it quite difficult to get the source data and scientific protocol it difficult to trust these results.

---

[a] Staw, Barry; Blumer, Catherine (11976). *Knee Deep in the Big Muddy*. Organizational Behavior and Human Decision Process. 35: 124–140. doi:10.1016/0749-5978(85)90049-4

[b] Arkes, Hal; Blumer, Catherine (11985). *The Psychology of Sunk Cost*. Organizational Behavior and Human Decision Process. 35: 124–140. doi:10.1016/0749-5978(85)90049-4

[c] Whyte, Glen (11986). *Escalating Commitment to a Course of Action: A Reinterpretation*. The Academy of Management Review. 11 (2): 311. doi:10.2307/258462. ISSN 0363-7425

Many people have strong misgivings about "wasting" resources (loss aversion). In the above example involving a non-refundable project, many people, for example, would feel obliged to go to continue the project despite not really wanting to, because doing otherwise would be wasting the project price; they feel they've passed the point of no return. This is sometimes referred to as the "sunk cost fallacy". Economists would label this behaviour "irrational": it is inefficient because it misallocates resources by depending on information that is irrelevant to the decision being made.

15.2.1.2 Anchoring Bias

Definition (#15.2.639): "Anchoring" or "focalism" is a cognitive bias that describes the common human tendency to rely too heavily on the first piece of information offered (the "anchor") when making decisions.

When people are trying to make a decision, they often use an anchor or focal point as a reference or starting point. Psychologists have found that people have a tendency to rely too heavily on the very first piece of information they learn, which can have a serious impact on the decision they end up making. In psychology, this type of cognitive bias is known as the anchoring bias or anchoring effect.

For example, the initial price offered for a used car sets the standard for the rest of the negotiations, so that prices lower than the initial price seem more reasonable even if they are still higher than what the car is really worth.

✍**Example:**

The anchoring and adjustment heuristic seems to have been first "theorized" by Amos Tversky and Daniel Kahneman (11979 according to holocene calendar). In one of their first studies, participants (in a unknown sample size to us so quite difficult to say if the result is accurate or not) were asked to compute, within 5 seconds, the product of the numbers one through eight, either as $1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7 \cdot 8$ or reversed as $8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$. Because participants did not have enough time to calculate the full answer, they had to make an estimate after their first few multiplications. When these first multiplications gave a small answer – because the sequence started with small numbers – the median estimate was 512; when the sequence started with the larger numbers, the median estimate was $2,250$ (the correct answer was $40,320$).

In another study by Tversky and Kahneman (11974), participants (again in a unknown sample size to us...) observed a roulette wheel that was predetermined to stop on either 10 or 65. Participants were then asked to guess the percentage of the United Nations that were African nations. Participants whose wheel stopped on 10 guessed lower values (25% on average) than participants whose wheel stopped at 65 (45% on average). The pattern has held in other experiments for a wide variety of different subjects of estimation[a].

---

[a] Tversky, A.; Kahneman, D. (11974 according to holocene calendar). *Judgement under Uncertainty: Heuristics and Biases* (PDF). Science 185 (4157): 1124–1131. doi:10.1126/science.185.4157.1124. PMID 17835457

This is why in business whoever makes that first offer has the edge since the anchoring effect will essentially make that number the starting point for all further negotiations. Not only that, it will bias those negotiations in your favour. That first offer helps establish a range of acceptable counteroffers, and any future offers will use that initial number as an anchor or focal point. One study even found that starting with an overly high salary request actually resulted in higher resulting salary offers.

15.2.1.3 Wisdom of Crowds

Way back in 11906 (holocene calendar), the English polymath Francis Galton visited a country fair in which 800 people took part in a contest to guess the weight of a slaughtered ox. After the fair, he collected the guesses and calculated their average which turned out to be 547 [kg]. To Galton's surprise, this was within 1% of the true weight of 543 [kg].

This is one of the earliest examples of a phenomenon that has come to be known as the "Wisdom of the crowd" (even if it is anecdotal, not scientific!). The idea is that the collective opinion of a group of individuals can be better (outperform) than a single expert opinion.

> *We know very well the wisdom of crowds! It was the one who elected Hitler in 11933 (holocene calendar) and Trump in 12016 (holocene calendar).*

quote
15.2:7389.#85

The popular interpretation of the Wisdom of Crowds phenomenon is that each participant brings a certain amount of information to the result, and a certain amount of noise. Over a large enough sample size, the noise (divergent) cancels itself out if the estimates are independent and non-emotional (there is no influence between the people), while the information converges on a value which, in the absence of systematic bias, should be proximate to the true value. This works quite well on very simple questions that doesn't involve mathematical competences.

This is result is related to the central limit theorem (CLT) and the law of large numbers (LoLN) that are cousins. Basically, the main difference is that central limit theorem is concerned with the distribution of the empirical mean and the law of large numbers is concerned with how close the empirical mean approaches the expected value:

- Law of large numbers (LoLN): As the number of samples grows large, the empirical mean approaches the expected value with probability 1, for recall (see section Statistics page 920):

$$\forall\varepsilon>0:\quad\left(\left(\frac{1}{n}\sum_{i=1}^{n}X_i-\mu\right)^2\geqslant\varepsilon^2\right)\leqslant\frac{\sigma^2}{n\varepsilon^2}\underset{n\rightarrow+\infty}{\rightarrow}0 \qquad (15.2.1)$$

- Central limit theorem: As the number of samples grows large, the difference between the empirical mean and the true mean becomes Normally distributed. The mean of the Normal distribution is that specified by the LoLN. Also, the standard deviation shrinks as the number of samples grows so, for a very large number of samples, the normal distribution becomes eventually a single number and the two results become equivalent.

$$\sqrt{n}\left(\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)-\mu\right)\xrightarrow{d}\mathcal{N}\left(0,\sigma^2\right) \qquad (15.2.2)$$

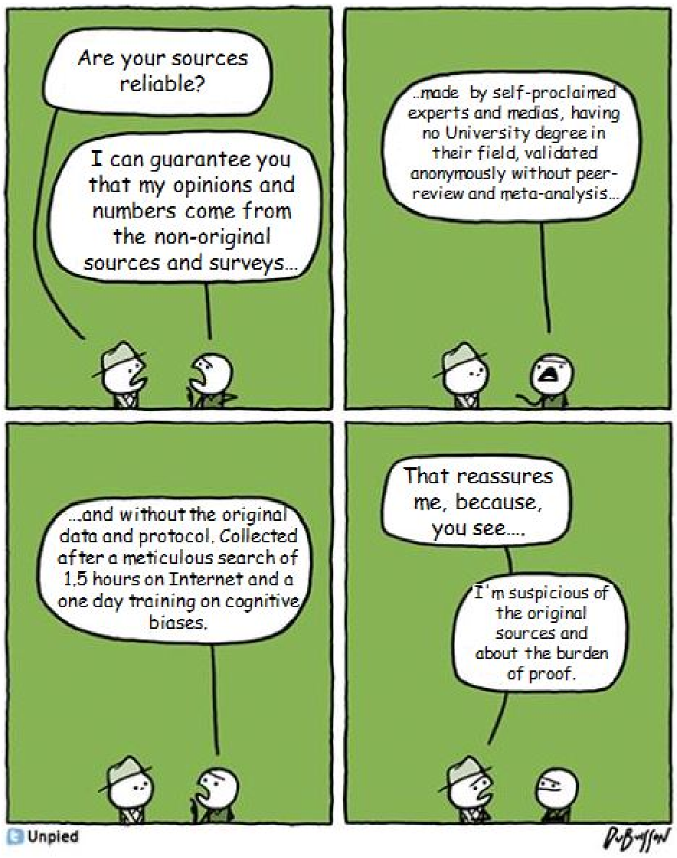
Obviously there as statistical situations where the crowd produces very bad judgement, and argues that in these types of situations their cognition or cooperation failed because (in one way or another) the members of the crowd were too conscious of the opinions of others and began to emulate each other and conform rather than think differently.

The most common application is the prediction market or project management task duration estimation.

To make it work well the following practical assumptions need to be meet:

- The individuals must be isolated to not be influenced by other people
- The individuals must come from different localizations to avoid cultural bias
- The individuals must have a given degree of expertise in the field related to the question
- The reasoning methods of individuals must be checked (to avoid false experts and emotion bias)
- The individuals must have a reward if they are near from the real value (otherwise some will troll the results)
- Individuals that may have a benefit for one reason or another that their estimated value is below or above the real one must be eliminated (risk of conflict of interest)
- Extreme values must be eliminated (as sadly there are still people that like to play to trolls)

and a good practice (as always in management!) is to ask for a three point estimates: optimistic, modal (most likely), pessimistic.


Are your sources reliable?
I can guarantee you that my opinions and numbers come from the non-original sources and surveys...
..made by self-proclaimed experts and medias, having no University degree in their field, validated anonymously without peer-review and meta-analysis...
...and without the original data and protocol. Collected after a meticulous search of 1,5 hours on Internet and a one day training on cognitive biases.
That reassures me, because, you see....
I'm suspicious of the original sources and about the burden of proof.
Unpied
DuBuisson

15.2.1.4 Delphi Method

The "Delphi method" or "Delphi technique", developed at the beginning of the Cold War (11959 according to holocene calendar) to forecast the impact of technology on warfare, is a structured empirical communication technique or method, originally developed as a systematic, interactive forecasting method which relies on a panel of experts (the name Delphi derives from the Oracle of Delphi).

Delphi is based on the speculation that forecasts (or decisions) from a structured group of individuals are more accurate than those from unstructured groups (the Delphi method has still 70 years after its creation never provided even weak scientific evidence to be better than no method at all). The experts answer questionnaires in two or more rounds. After each round, a facilitator or change agent provides an anonymised summary of the experts' forecasts from the previous round as well as the reasons they provided for their judgements. Thus, experts are encouraged to revise their earlier answers in light of the replies of other members of their panel. It is believed that during this process the range of the answers will decrease and the group will converge towards the "correct" answer. Finally, the process is stopped after a predefined stop criterion (e.g., number of rounds, achievement of consensus, stability of results), and the mean or median scores of the final rounds determine the results.

The Delphi method is used when different forecasting methods, such as theoretical approach, quantitative models or trend extrapolation, quickly became inefficient in areas where precise scientific laws have not been established yet. To combat these shortcomings, the Delphi method was developed by Project RAND during the in 11959 (holocene calendar) by Olaf Helmer, Norman Dalkey, and Nicholas Rescher. It has been used ever since, together with various modifications and reformulations, such as the Imen-Delphi procedure but still without being able to given statistically evidence of its efficiency (...).

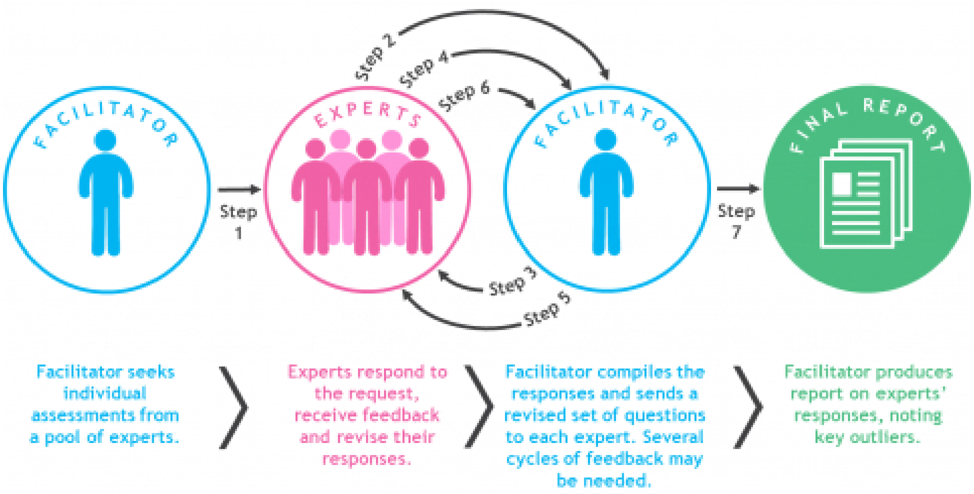
There are many variations of the Delphi method. Here is a classic variant, with four steps:

- Step 1: define with rigour and precision the object on which the Delphi method will relate. The object corresponds to the problem that will have to examine the experts and the major questions linked to this problem. It is important to spend time defining the object otherwise it can lead the experts into an uncertain process.
- Step 2: Choose the experts. This choice is made according to various criteria, in particular their independence and their excellent knowledge of the subject. It is recommended that the final number of experts not be less than 25. This means that a larger number should be planned at the start to take account of refusals and abandonments.
- Step 3: Develop a questionnaire. The questions must be targeted, precise and possibly quantifiable.
- Step 4: Administer the questionnaire and process the responses. The first questionnaire, which serves as a basis, will just be enriched at each round by results and comments generated previously.

It is a process which today, is done quite quickly thanks to the intervention of information systems.

In the second round of the questionnaire, the experts receive the results of the first round and must again give their opinion on the questionnaire, now knowing the group's opinion. The fourth round will give the final answers, that is to say the median consensus opinions and the dispersion of opinions around this median. At the end of the Delphi method, the analysts write a summary report.

The purpose of the first questionnaire is to identify the median and the interquartile range. Then, the second questionnaire (Q2) aims to reduce the contradictory positions (i.e. the interval Q1-Q3). Thirdly, a new questionnaire (Q3) is carried out. It aims to oppose the extreme answers by reconciling their arguments. It is the fourth questionnaire which gives the definitive answer.


Step 2
Step 4
Step 6
FACILITATOR
EXPERTS
FACILITATOR
FINAL REPORT
Step 1
Step 7
Step 3
Step 5
Facilitator seeks individual assessments from a pool of experts.
Experts respond to the request, receive feedback and revise their responses.
Facilitator compiles the responses and sends a revised set of questions to each expert. Several cycles of feedback may be needed.
Facilitator produces report on experts' responses, noting key outliers.

The reader can take a look to the well written paper [533] published in 11993 (holocene calendar) that systematically reviews empirical studies looking at the effectiveness of the Delphi technique, and provides a qualitative critique of this research. This paper also provides a very detailed six step practical, systematic approach to the design and delivery of a Delphi survey (but don't expect to find any scientific evidence or mathematics... it a purely qualitative paper based on beliefs written by psychologists...). Findings at the time of publication of that paper suggest that Delphi groups outperform statistical groups (by twelve studies to two with two 'ties') and standard interacting groups (by five studies to one with two 'ties'), although there is no consistent evidence that the technique outperforms other structured group procedures and that such small sample size cannot lead to any serious conclusion (see page 876). However, important differences exist between the typical laboratory version of the technique and the original concept of Delphi, which make generalisations about 'Delphi' per se difficult. These differences derive from a lack of control of important group, task, and technique characteristics (such as the relative level of panellist expertise and the nature of feedback used). Indeed, according to the authors of this paper there may be theoretical and empirical reasons to believe that a Delphi conducted according to 'ideal' specifications might perform better than the standard laboratory interpretations (...). It is concluded that a different focus of research is required to answer questions on Delphi effectiveness, focusing on an analysis of the process of judgement change within nominal groups (and obviously on much larger meta-sample sizes).

We also strongly recommend the user to read [579] to know more about the Delphi method and the human cognitive biases.

15.2.2 Utility

As we will see more in details in the section Economy, "utility" is a measure of preferences over some set of goods and services. The concept is an important underpinning of rational choice in economics and game theory, because it represents satisfaction experienced by the consumer of a good. A good is something that satisfies human wants. Since one cannot directly measure benefit, satisfaction or happiness from a good or service, economists instead have devised ways of representing and measuring utility in terms of economic choices that can be measured. Economists have attempted to perfect highly abstract methods of comparing utilities by observing and calculating economic choices. In the simplest sense, economists consider utility to be revealed in people's willingness to pay different amounts for different goods.

Definition (#15.2.640): A "utility function" or "payoff function" is a function of all earnings (gains) of the set of n-players to $\mathbb{R}^n$ (most of time corresponding to monetary value) that associates utilities withdrawn by each player at the issue of the game. Formally if a game has a set of strategies S_i for a given player i we denote it by:

$$U : S_i \mapsto \mathbb{R} \tag{15.2.3}$$

If U is a utility function, we will denote U_i the function of the set of all outgoings of a game to $\mathbb{R}$ of player i. Such a function will be said to be "representative of the preference $\succcurlyeq$" if for any outgoing $x, y \in I$ (any outgoing of the set of outgoings I) we have:

$$\boxed{x \succcurlyeq y \Leftrightarrow U_i(x) \succcurlyeq U_i(y)} \tag{15.2.4}$$

The theory of utility, that game theory use, axiomatize the fact that only this concept of preference should be important. Briefly, we will say that only the preference order of the utility of the outgoings of a game is important, the values of the gains relatively to each outgoing being without importance (however reality is more complexed as discussed in [89]).

15.2.2.1 Pareto Optimum

A first criterion that comes to mind when we speak about Utility, and that is due to the Italian sociologist Vilfredo Pareto (11915 according to holocene calendar), is the optimality of the same name[7].

If in a game, a couple of strategies (tactics/choices) is such that it is impossible to improve the score of one of the two players without decreasing the score of the other, we say that these outcome is "Pareto-optimal" or "Pareto-efficient".

More formally:

Let us consider two issues (choices) x and y, both belonging to the set I of issues, and suppose that for each individual i, we have the following situation:

$$\forall i \quad U_i(y) \geq U_i(x) \tag{15.2.5}$$

---

[7] Not to be confused with the "Pareto" completely empirical concept in economics that the most distributions are in the ratio $20/80\%$ (see section Quantitative Management page 8095).

In other words, no individual would be a priori prejudiced (compared to the other) if we have substituted for each state y by the state x. Let us assume moreover, that there is at least one individual j who prefers strictly y to x as:

$$U_j(y) > U_j(x) \tag{15.2.6}$$

In these circumstances, we do not really see why player should choose y rather than x.

Definition (#15.2.641): A feasible outcome i that admits no possible improvement is named a "Pareto optimum" and is strictly formally defined as:

$$\boxed{\forall j \in I \setminus \{i\}, \exists n \in N: \quad U_n(i) > U_n(j)} \tag{15.2.7}$$

The existence of Pareto optimality must be understood as a prerequisite, a "minimum minimorum" without which the concept of a cooperative game solution that we seek to develop should automatically be rejected!!!

Remark

This result form what we have already written earlier in this section. That is to say that if in a game, a couple of outcomes is such that it is impossible to improve the score of one of the two players without decreasing the score of the other, we say that these outcomes are "Pareto-optimal" or "Pareto-efficient".

15.2.2.2 Nash Equilibrium

Definition (#15.2.642): A "Nash equilibrium" (or "equilibrium" to make short) thus described game in issue in which no player has interest to change its strategy unilaterally, given the strategies of other players. In other words, if each player has chosen a strategy and no player can benefit by changing strategies while the other players keep theirs unchanged, then the current set of strategy choices and the corresponding payoffs constitutes a Nash equilibrium.

Game theorists use the Nash equilibrium concept to analyse the outcome of the strategic interaction of several decision makers. In other words, it provides a way of predicting what will happen if several people or several institutions are making decisions at the same time, and if the outcome depends on the decisions of the others. The simple insight underlying John Nash's idea is that one cannot predict the result of the choices of multiple decision makers if one analyses those decisions in isolation. Instead, one must ask what each player would do, taking into account the decision-making of the others.

Given G a game with n players, and:

$$s^* = (s_1^*, \ldots, s_n^*) \tag{15.2.8}$$

a combination of strategic choices of these n players is the best strategic choice of player i with $s_i^* \in S_i$, the set of all feasible strategies by player i.

Given $U(s_1^*, \ldots, s_n^*)$ the utility of player i when s^* is selected. A combination of strategic choices is a Nash equilibrium if and only if:

$$\boxed{U_i(s_1^*, \ldots, s_i^*, \ldots, s_n^*) \geq U_i(s_1^*, \ldots, s_i, \ldots, s_n^*)} \tag{15.2.9}$$

for any s_i in S_i and any i.

Interpretation: No player may make a profit of a deviation of s_i^*, whatever the strategy he chooses in the set S_i. Thus, no player has an interest to deviate, and s_i^* participates to an equilibrium.

Remark

It can happen that a Pareto optimum is not distinguishable form Nash equilibrium but this is not always the case (therefore a Nash equilibrium is not always Pareto optimal)!

Definition (#15.2.643): When the strategy of a player is the best response to all possible strategies of the rivals, we speak then of "dominant strategy" (this strategy dominates all other strategies of the player). The equilibrium of this game is the named "equilibrium in dominant strategy".

Verbatim, a strategy is "dominated" if it offers the player always lower earnings than those associated with at least one other of its strategies.

Remark

We can ask ourselves if in a non-cooperative game a Nash equilibrium (if it exists) is not such that it cause anyway like an implicit cooperation? In fact, this is not the case (and it's a very important result) as we will discussed it later in the study of the famous "prisoner's dilemma", that is a game where the Nash equilibrium is ensured by such individualistic and rational choices they are uncooperative !!! So it will be an extremely important example in the context of the market economy.

Method: One way of determining the equilibrium of the game is to first eliminate all dominated strategies and to seek the equilibrium in the reduced corresponding game.

✍**Example:**

Consider the following game in a table form:

J_2 J_1	S_1	S_2	S_3
S_1	$5,2$	$4,4$	$6,4$
S_2	$3,1$	$2,0$	$5,2$

Table 15.9: **Payoff matrix with Nash Equilibrium**

But the next game at the opposite, does not include a Nash equilibrium. Indeed, whatever the couple of planned strategies, one player always gets more by modifying his choice:

J_2 J_1	S_1	S_2
S_1	$1,0$	$0,1$
S_2	$0,1$	$1,0$

Table 15.10: **Payoff matrix without Nash equilibrium**

However, currently it seems at least premature to prescribe to the players the choice of an equilibrium. Indeed if it is chosen, the situation is relatively stable, but there are three problems:

1. We are not sure of the existence of a couple of tactical equilibrium (conjunction of prudent tactics)
2. Even in case of existence, we are not sure of the uniqueness of a couple of tactics in equilibrium
3. Even in case of existence and uniqueness, we can prescribe another choice (!!!!)

15.2.3 Games Representations

There are different ways to formalize Games and Decision theory and especially depending the type of situations in question. Thus, we distinguish:

1. The "extensive forms" which are synoptic forms (tree, branch, leaf) useful to a simple understanding of the possible strategies and where the outcome of a game is represented by a leaf in which we find the vector of gains (or "earnings") of the respective players. This kind of representation becomes complicated (hard to draw) with repetitive games.

   When an extensive form use probabilities, we then refer to a "decision tree" because, as we mentioned at the beginning, which says known probabilities says complete apart theory: decision theory.

2. The "normal forms" that can significantly reduce the size and time of graphical representation of a game as a table (matrix) of gains (or "earnings") but are inappropriate for repetitive games.

   At least two main sub-categories can be distinguished:

   - The "normal forms of zero-sum games" (strictly competitive games) where according to an appropriate choice, it is possible to simplify the matrix representation (or "bimatrix") in a half-matrix since earnings are equal and opposite for players for each particular strategy.
   - The "normal forms of non-zero-sum games" (competitive games).

   Remark

   Each cell of the table/matrix therefore contains a "vector" whose components are the respective gains of the players. If the game is a zero-sum one each cell contains a single value as what is earned by a player is lost by the other. We will see many such examples soon.

3. The "set-forms" that have a probability set-oriented approach that will allow us to study the last form below.

4. The "graphic forms" that are pleasant to watch and that we will introduce as a complementary approach as using operations research (see section Numerical Methods page 6645).

Formally a game specifies:

1. Players: A set of players (agents who play the game) $N = \{1, \ldots, n\}$ with typical element (player) $i \in N$.

2. Strategies: For each player i a non-empty set of feasible strategies S_i with typical element is $s_i \in S_i$.

3. Payoffs: For each player $i \in N$ a payoff (utility) function:

$$u_i : S_i \mapsto \mathbb{R} \tag{15.2.10}$$

A formal way to write down the normal-form of a game is:

$$G^N = \left[N, \{S_i\}_{i \in N}, \{u_i\}_{i \in N}\right] \tag{15.2.11}$$

15.2.3.1 Extensive representation of a Game

The rules of a game of strategy and associated payoffs with it can thus be represented in a more extensive form commonly named by specialists "Kuhn tree".

As example consider two computers firms named *MBI* and *Pear* that have to do the choice of an operating system CAM or MAC. The compatibility between the systems would be socially preferable, but for reasons related to the history of the two firms, each would prefer that it be the other who make the effort to adapt. If both firms choose CAM, *MBI* (abbreviated J_1) wins \$600 million and *Pear* (abbreviated J_2) wins \$200 million. If they choose MAC, *Pear* wins \$600 millions and *MBI* \$200 million. If they are not compatible, then each earn \$100 million.

> **Remark**
>
> We name this type of game, a "coordination game". For example, the choice of television standards or Mac and PC drive match this type of games. Every manufacturer wants to impose its own standard but in case of disagreement, consumers may refuse to buy the product.

Firms sequentially play thus the game can be represented as a decision tree:

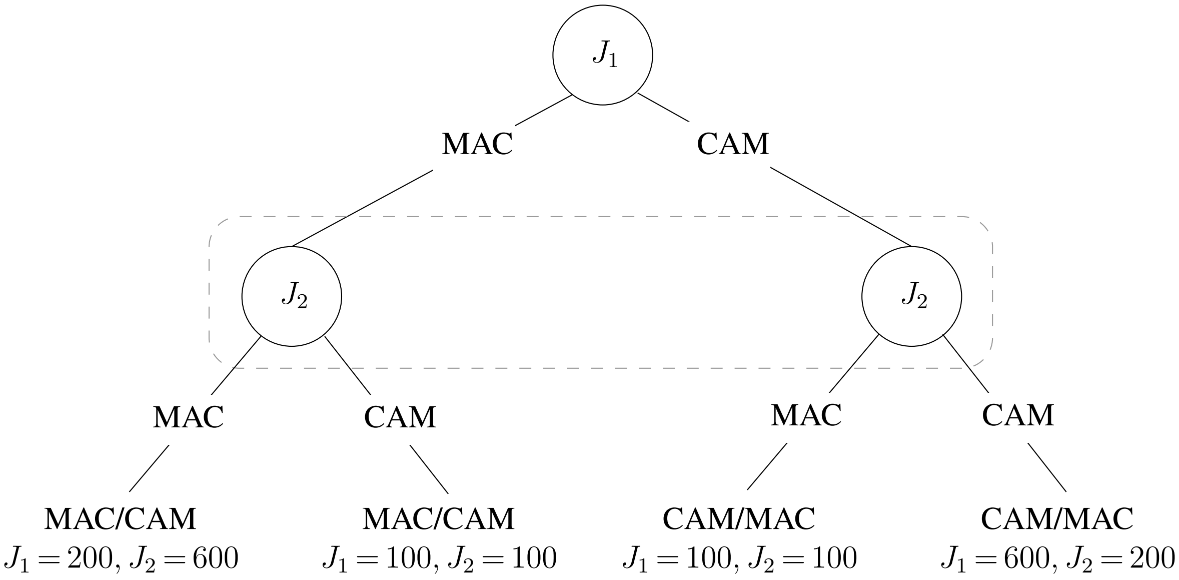


Figure 15.56: Sequential game in the form of a decision tree (or Kuhn's tree)

Remarks

R1. The information structure highlighted above refers to the information available to each player at each node making the game.

R2. MAC/CAM is a "perfect information game" in the sense that players know exactly their range of strategies and those of their opponent and the precise consequences of these strategies. Thus, each node of the extensive form is visible by the players (we will define the concept of perfect information formally a little bit later).

A simple analysis of the best strategy in the context of a game is to jump to the normal form as we shall see a little further (but this normal form is not suitable for an extensive form of a decision).

15.2.3.2 Extensive representation of a Decision

As we mentioned at the beginning of this section, the theories of games and decision are differentiated by the fact that the data of the first are totally deterministic universe while for the second, they are completely probabilistic. This last case is so important in the industry that there are, as we will see later, well-know softwares (@RISK, Isograph, TreeAge) specialized for the management of extensive forms (the latter being methods included in the Risk Management ISO 31010 norm).

The simplest case of decision in the industry decision is the "event tree analysis" which uses (when the probabilities are fixed) only the basic axioms of probability (see section Probabilities page 476). If the probabilities are not fixed (which is common in reality), it will be necessary to use softwares incorporating the Monte Carlo methods (see section Numerical Methods page 6708).

The event tree analysis is a graphical technique for representing sequences of mutually exclusive events following an initiating event depending on the operational/non-operational state of various systems designed to limit its consequences.

Below you can see a simple example of event tree whose calculations are made automatically with the Microsoft Office Visio software (but same can be done with Microsoft Office Excel spreadsheet software):

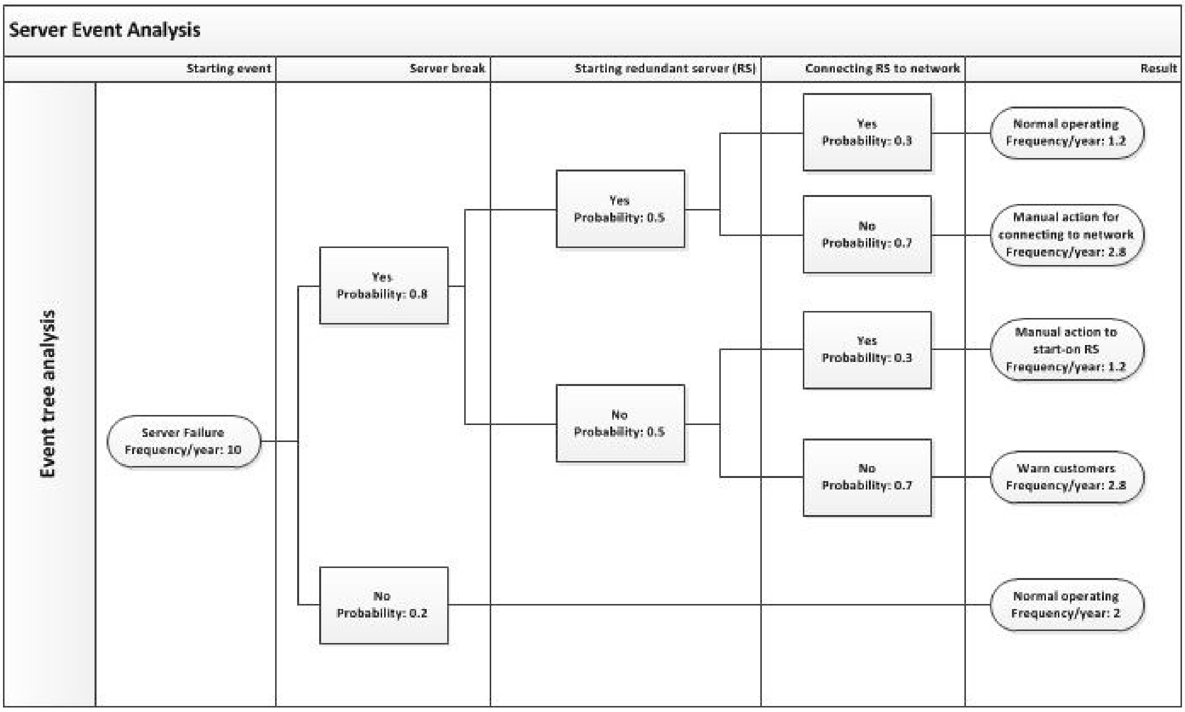

Server Event Analysis
Starting event
Server break
Starting redundant server (RS)
Connecting RS to network
Result
Event tree analysis
Server Failure
Frequency/year: 10
Yes
Probability: 0.8
No
Probability: 0.2
Yes
Probability: 0.5
No
Probability: 0.5
Yes
Probability: 0.3
No
Probability: 0.7
Yes
Probability: 0.3
No
Probability: 0.7
Normal operating
Frequency/year: 1.2
Manual action for
connecting to network
Frequency/year: 2.8
Manual action to
start-on RS
Frequency/year: 1.2
Warn customers
Frequency/year: 2.8
Normal operating
Frequency/year: 2

Figure 15.57: **Event tree analysis with fixed probabilities in Microsoft Office Visio**

The annual frequency in the column located at the extreme right of the table (the "leafs" as say practitioners or "resulting frequencies") is simply equal to the estimated annual frequency of the initiating event multiplied by the product of the probabilities of a branch such that:

$$f_r = f_0 \prod_i P_i \tag{15.2.12}$$

and we must of course be careful that the sum of the probabilities in each column is equal to 100%.

Another typical example of extensive form are "decision trees".

To see what it is let us imagine an IT company B in potential competition with another company A (the latter can be seen as a set of competitors too!) for an international IT migration for a firm X.

Simplifying a little bit, however without being out of reality, consider that two options are open to B: target "high prices" or target "low prices".

Suppose we also know that in the past B has submitted a proposal for each contract of this type, while A has done it only in 60% of cases (no probability distribution function in our scenario yet but only punctual estimations!).

We also know that:

1. If B submits a high price and is the only one to submit a proposal, its expected profit is 22 million.
2. If B submits a higher price but is in competition with A, it will get the contract on the

price level requested by the group A. In this case, he knows he will get an average of 1 million.

3. If B submits a low price, it is sure to get the contract and make a profit of 10 million.

So in the case where the project is chosen only by its price (at the detriment of quality as often in reality...) the questions are then the following:

Q1. What should do B, if no further information can be obtained?

This is a situation of the type: "decision without information".

Q2. Assuming that a spy within the group A can inform B if the group A will submit an offer or not, what would be the price of this information for B?

This is a situation of the type: "decision with perfect information".

Q3. A consulting firm may advise us, but his expertise, expensive, amounts to 1 million per study. To consider the use of its services, we know that in the past, of the 30 times that Group A had in fact submitted an offer, the consulting firm had anticipated it 24 times. And, of the 20 times he had not submitted an offer, the consulting firm had anticipated it 17 times. Should B order a study (anyway in reality this kind of information is almost very hard to obtain...)?

This is a situation of the type: "decision with imperfect information".

Let us see now the solution (S) for each question (Q):

S1. To answer the first question (Q1), we first represent the problem to solve in the graphic form of a decision tree (which is for the moment quite simple to build also as a table) with the TreeAge software for example:

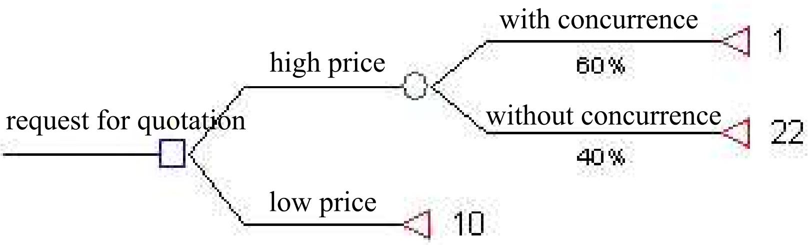


Figure 15.58: Decision tree without information in TreeAge

Remark

We would like to indicate that it is possible to make the same type of trees with Microsoft Office Visio but Monte Carlo modelling is not incorporated in this software and creating formulas consumes a lot of time (count a time factor of 10 to 20 compared to TreeAge, Isograph or @RISK).

Then, launching the calculation of the expected mean at each branch, named "expected monetary value" (E.M.V.) TreeAge gives us just (this software has an option to make Monte Carlo modelling but the example here being with fixed probabilities this option is not necessary at this level):

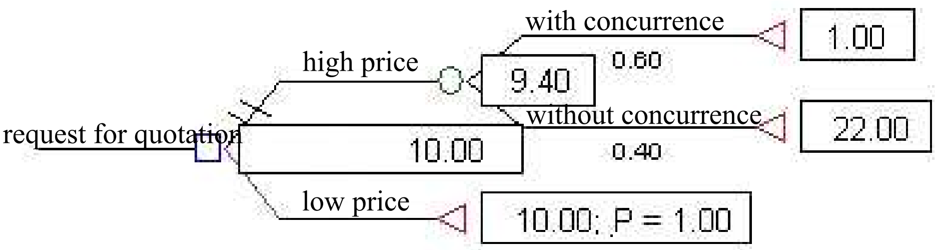


Figure 15.59: **Decision tree without information and expected mean in TreeAge**

Thus, the answer to the first question is that the strategy giving the greatest expected mean is the "low price" strategy because there is an expected gain of 10 million.

With the first decision (high price) we would have an expected mean of:

$$22\cdot 0.4+1\cdot 0.6=9.4 \tag{15.2.13}$$

Remark

In decision trees a basic check rule is to have all probabilities of a given branch that sum up to 1!

S2. To answer the second question (Q2) which is to know the financial value of the information given by the spy, we must first build the tree (in facts the example is so simple here that it is not really necessary but...) of a so-named competitive situation with "perfect information" (as the spy can give us an information completely sure).

The tree is with this small example very easy to build. If the spy tells us that the competitor A will make an offer, then we will have to propose the cheapest offer. Otherwise, we will propose the highest offer. The scenario is as follows:

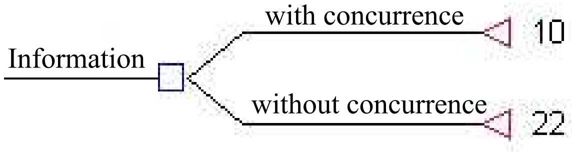


Figure 15.60: **Decision tree with perfect information in TreeAge**

The probability that there is competition is 60% and 40% that there is none. So the "expected mean of monetary value of perfect information" (E.M.V.P.I.) is in a situation of perfect information:

$$22\cdot 0.4+10\cdot 0.6=14.8 \tag{15.2.14}$$

Therefore compared to the best previous situation we have a positive difference of 4.8 million. So this is the price value of the perfect information given by the spy.

S3. About the third question (Q3) of determining the value of imperfect information provided by the consulting firm, the only certainty we have is that this information can't have a value greater than that of the perfect information. Thus it will have a value between 0 and 4.8 million.

To start, remember that according to the statement, we believe that for the current request for proposal, there is 60% probability that there is competition and the consulting company in the past had 80% of the time right (24 times out of 30) when it said there would be competition (and thus 20% of other times it was wrong...).

Respectively, we believe for the current request proposal that there is 40% probability that there is no competition $(1-60\%)$ and the consulting firm in the past was right 85% of the time (17 times out of 20) when it said there would be no competition (and thus 15% of other times wrong...).

What can be summarized in the form of a table as following:

Previsions / **Reality**	**Probability**	**With Competition**	**Without Competition**
With Competition	60%	80%	20%
Without Competition	40%	15%	85%

Table 15.11: **Decision with imperfect information (original form)**

We would like now:

(a) Calculate the probability that there is really competition AND that the consulting firm has planned a scenario with competition.

(b) Calculate the probability that there is really no competition AND that the consulting firm has planned a scenario with competition.

(c) Calculate the probability that there is really no competition AND that the consulting firm has planned a scenario without competition.

(d) Calculate the probability that there is really competition AND that the consulting firm has planned a scenario without competition.

To calculate these probabilities, we will use Bayes' formula (see section Probabilities page 492). As a reminder, the posterior and a priori probabilities are given by:

$$P(B/A)=\frac{P(A\cap B)}{P(A)} \quad \text{and} \quad P(A/B)=\frac{P(A\cap B}{P(B)} \tag{15.2.15}$$

thus (we know a posteriori probabilities in our example):

$$P(A\cap B)=P(B/A)P(A) \tag{15.2.16}$$

We can now:

(a) Calculate the probability that there is really competition AND that the consulting firm has planned a scenario with competition. We then use the previous table:

$$P(A\cap B)=P(B/A)P(A)=80\%\cdot 60\%=48\% \tag{15.2.17}$$

where in this situation B is the event "there was really competition" and A is the event "competition scenario planned by consulting firm". B/A is thus the a posteriori event "there was really competition when competition planned by consulting firm ".

(b) Calculate the probability that there is really no competition AND that the consulting firm has planned a scenario with competition. We then use the previous table:

$$P(A\cap B)=P(B/A)P(A)=15\%40\%=6\% \tag{15.2.18}$$

where in this situation B is the event "there was really no competition" and A is the event "competition scenario planned by consulting firm". B/A is thus the a posteriori event "there was really no competition when competition planned by consulting firm".

(c) Calculate the probability that there is really no competition AND that the consulting firm has planned a scenario without competition. We then use the previous table:

$$P(A\cap B)=P(B/A)P(A)=85\%40\%=34\% \tag{15.2.19}$$

where in this situation B is the event "there was really no competition" and A is the event "no competition scenario planned by the consulting firm". B/A is thus the a posteriori event "there was really no competition when no competition was planned by consulting firm".

(d) Calculate the probability that there is really competition AND that the consulting firm has planned a scenario without competition. We then use the previous table:

$$P(A\cap B)=P(B/A)P(A)=20\%60\%=12\% \tag{15.2.20}$$

where in this situation B is the event "there was really competition" and A is the event "no competition scenario planned by the consulting firm". B/A is thus the a posteriori event "there was really competition when no competition was planned by consulting firm".

We then have the following table which summarizes in a more or less usable way all possible scenarios:

Previsions. / **Reality**	**With Competition**	**Without Competition**
With Competition	48%	12%
Without Competition	6%	34%
Total	54%	46%

Table 15.12: **Decision with imperfect information (tabular form)**

Using this table, we can easily check that the sum of columns equal 100% (that is to say all eventualities) and thus that the calculations are correct.

So we see that 54% of the time the consultancy firm plans a competition scenario (whatever will be the reality) and 46% of the time no competition (whatever will be the reality).

We therefore get the following decision tree with the TreeAge software:

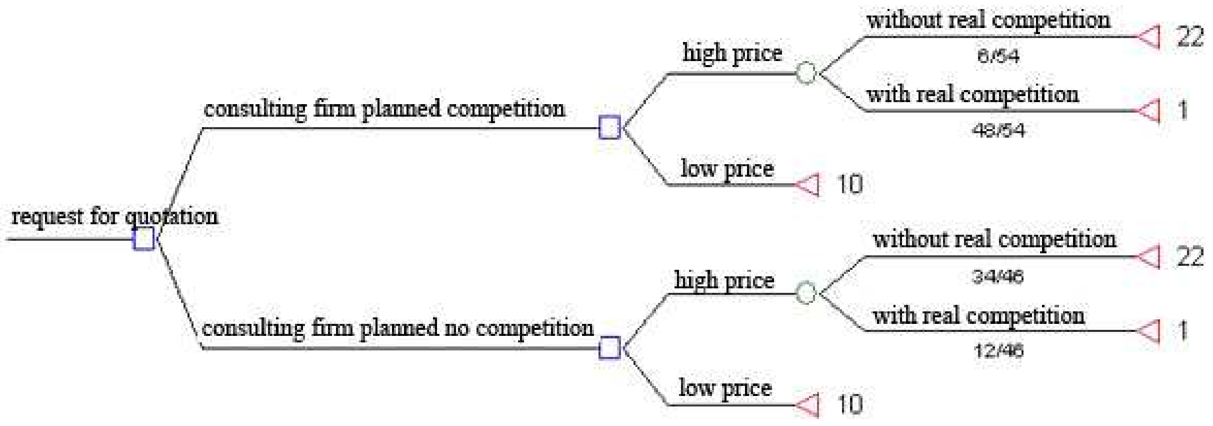
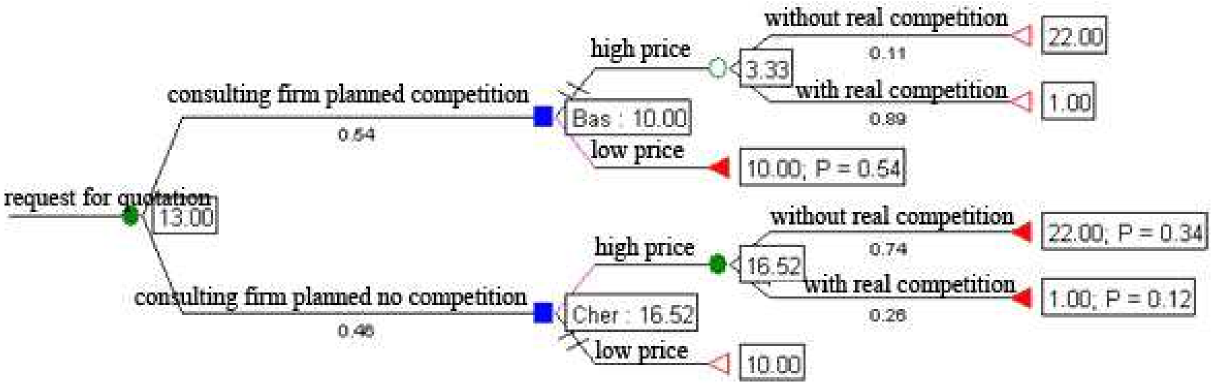

request for quotation
consulting firm planned competition
high price
without real competition
6/54
22
with real competition
48/54
1
low price
10
consulting firm planned no competition
high price
without real competition
34/46
22
with real competition
12/46
1
low price
10

Figure 15.61: **Decision tree with imperfect information in TreeAge**

Giving after calculations (always in TreeAge):


request for quotation
13.00
consulting firm planned competition
0.54
Bas : 10.00
high price
3.33
without real competition
0.11
22.00
with real competition
0.89
1.00
low price
10.00; P = 0.54
consulting firm planned no competition
0.46
Cher : 16.52
high price
16.52
without real competition
0.74
22.00; P = 0.34
with real competition
0.26
1.00; P = 0.12
low price
10.00

We find ourselves with an expected gain of 13 million (visible on the root of the above tree) minus the 1 million for the payment for the consulting firm, resulting finally in a net gain of 12 million.

Thus, the imperfect information the value is 2 million (remember that with low price strategy we are sure to get the contract and have a profit of 10 million) and must be compared to 4.8 million of the case with perfect information. This result is quite logic.

Remarks

R1. This type of tree is often used in pharmaco-economics (Cost–Benefit analysis). Thus, the root of the tree is an infection X for which there are several antibiotics (branches A/B) and each has two issues (successful/failed treatment) with two possible outcomes (side effects yes/no). A probability is associated with each node and for the terminal nodes the treatment costs. Thus, by calculating the expectation, it is possible to determine the best economic antibiotic choice for the Medical Institute (obviously this is ethically useful when the success rate of both antibiotics A/B is close and that the rate of side effects yes/no is also close... otherwise this method would be a scandal!).

R2. In industry, many companies use these trees with probability distributions defined on each node. They then make a Monte Carlo simulation on the whole and make a sensitivity analysis (Tornado graph) with business intelligence softwares such as @RISK from Palisade.

15.2.3.2.1 Real Options

When using trees for investment decisions in projects (we speak then about "investment analysis with real options") we must take into account that each series of parallel branches is a project period (month, quarter, semester or year). We must therefore not forget then to update the values to the risk free rate of return of the market (see section Economy page 7548) for each period. Also such a tree gives the possibility calculate the investment option value, which is required by the executive board in high level companies.

Consider as illustration case the following situation with a single time period:

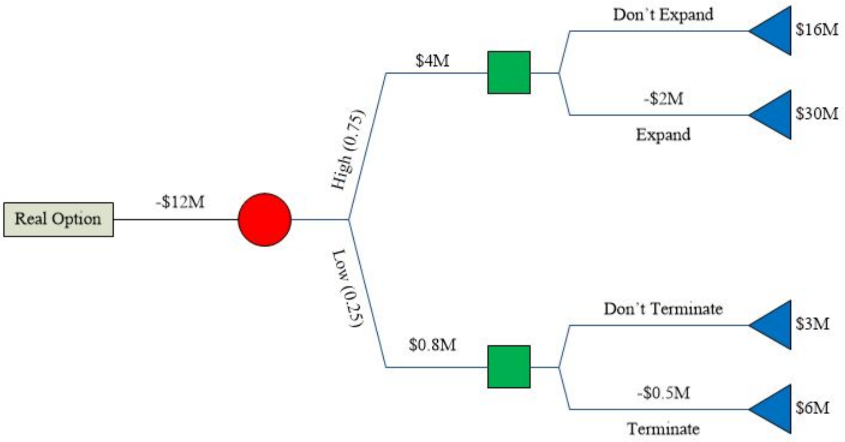

Real Option
-$12M
High (0.75)
Low (0.25)
$4M
$0.8M
Don't Expand
$16M
-$2M
Expand
$30M
Don't Terminate
$3M
-$0.5M
Terminate
$6M

Figure 15.62: **Complete investment Tree (unreduced)**

In a software like Palisade @Risk this will give typically (unfortunately the options costs are always hidden in the total of the branches in specialized software, which is why we prefer most of time to draw our own trees in Microsoft® Excel®):

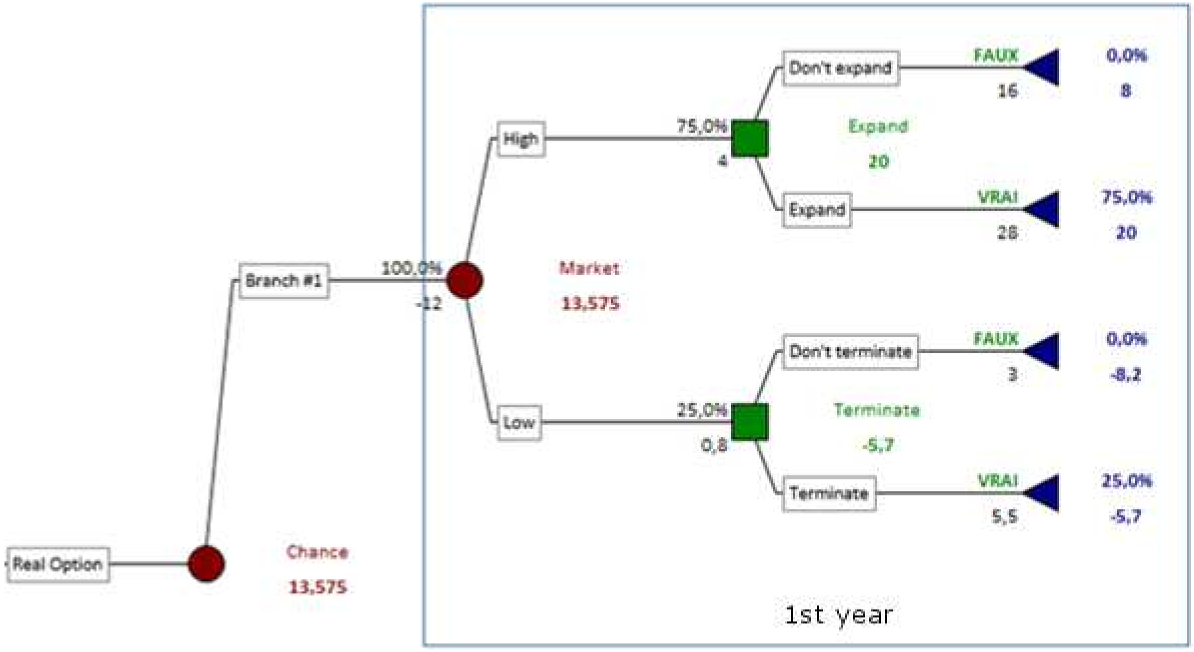

Real Option
Chance
13,575
Branch #1
100,0%
-12
Market
13,575
High
75,0%
4
Low
25,0%
0,8
Don't expand
FAUX
16
0,0%
8
Expand
20
Expand
VRAI
28
75,0%
20
Don't terminate
FAUX
3
0,0%
-8,2
Terminate
-5,7
Terminate
VRAI
5,5
25,0%
-5,7
1st year

Figure 15.63: **Previous tree build in the software Precision Tree 6.3 of Palisade**

Thus we have obviously:

$$\text{E(Payoff)}=-12+0.75(4+30-2)+0.25(0.8+6-0.5)=-12+0.75\cdot 32+0.25\cdot 6.3=13.575 \qquad (15.2.21)$$

Therefore the tree can be reduced to:

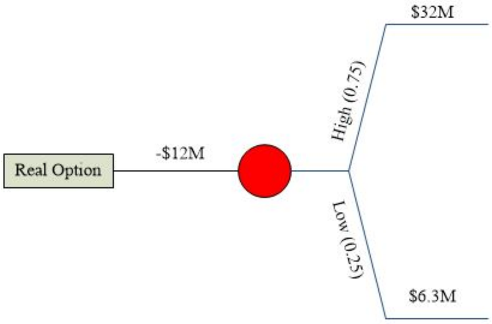


Figure 15.64: **Optimistic investment reduced tree**

But what is the expected net present value (eNPV)? Well this type of softwares allows us easy to integrate this type of calculation with or without a Monte Carlo simulation. Thus, if the market risk free rate is 10% over a year, then we have (see the section on Economy at page 8050 for the detailed study of Net Present Value):

$$\text{eNPV}=-12+\frac{0.75\cdot 32+0.25\cdot 6.3}{(1+10\%)^1}=11.25 \qquad (15.2.22)$$

Now it might be interesting (even requested by executive board!) to calculate the "price of the option to invest". In the pessimistic configuration our tree becomes:

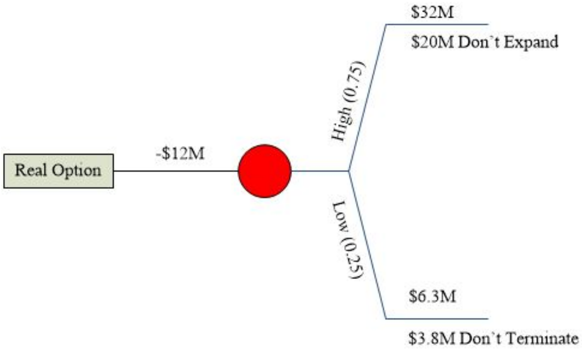


Figure 15.65: **Pessimistic investment reduced tree**

We then for this configuration:

$$\text{eNPV}=-12+\frac{0.75\cdot 20+0.25\cdot 3.8}{(1+10\%)^1}=3.95 \qquad (15.2.23)$$

The price of real investment option (expected Option Value) is therefore:

$$\text{eOV}=11.25-3.95=7.2\,[\text{M\$}] \qquad (15.2.24)$$

15.2.3.3 Normal representation of a Game

To move to the "normal representation" or "strategic representation", we define a strategy as a comprehensive action plan for each player, which specifies a choice for each node of the tree and therefore for every situation that may occur during the game. the "payoff matrix" represents the strategic situation of the players and the payoffs they receive for each strategy.

We take again the previous example MAC/CAM:

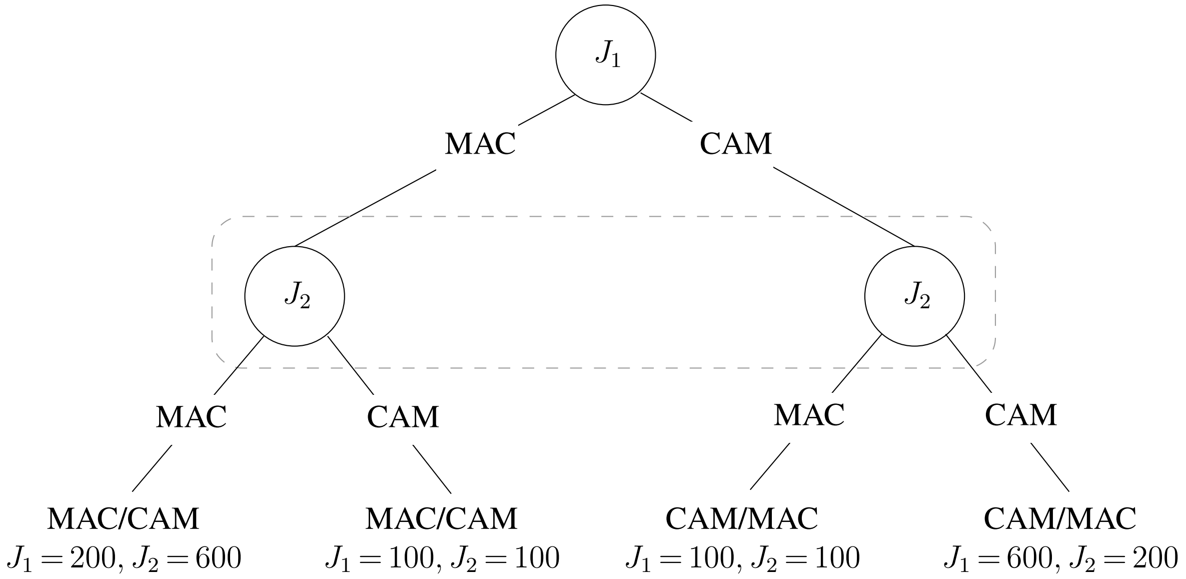


Figure 15.66: Sequential game in the form of a decision tree (or Kuhn's tree)

And we get for the strategic representation:

J_2 J_1	**CAM**	**MAC**
CAM	$600, 200$	$100, 100$
MAC	$100, 100$	$200, 600$

Table 15.13: Payoff matrix of a non-zero sum game

So this is a simple tabular representation of a game.

Remarks

R1. We see in this matrix that the interests of the two companies are not completely opposed, they progress each time in the same direction when the strategies are opposed (if one lose, the other lose too and vice versa). Thus, the MAC/CAM game is a game whose payoffs are not growing in opposite directions (strategies). We speak then of "not strictly competitive game" (we will define this concept formally a little bit further).

R2. We also see that regardless of the strategy chosen by one player, each possible choice by the other player always bring to equivalent payoff. Therefore, we say that it is a "no prudent tactical game".

Definition (#15.2.644): A strategy is said to be a "prudent tactic" (that is the choice of the row number for the row player or the number of the column for the column player) when the gain of one of the players is such that when compared to a selected strategy, all the choice of its competitor provides a maximum gain to the latter. The minimum insured gain of (for example) J_1 is named the "security level" of J_1.

$J_2(B)$ $J_1(A)$	b_1	b_2	b_3	b_4
a_1	5 , 5	6 , 4	0 , 10	4 , 6
a_2	1 , 9	7 , 3	5 , 5	6 , 4
a_3	6 , 4	7 , 3	7 , 3	8 , 1
a_4	4 , 6	8 , 1	0 , 10	2 , 8
a_5	3 , 7	5 , 5	9 , 0	0 , 10

Table 15.14: **Payoff of a game with potential prudent tactic**

Player $A(J_1)$ may think that player $B(J_2)$ is very insightful, or very lucky, and so is able to choose the best possible response to any tactic of A.

Therefore:

- If A choose a_1, B guess that and choose b_3 and A would won 0 (while B will won 10)
- If A choose a_2, B guess that and choose b_1 and A would won 1 (while B will won 9)
- If A choose a_3, B guess that and choose b_1 and A would won 6 (while B will won 4)
- If A choose a_4, B guess that and choose b_3 and A would won 0 (while B will won 10)
- If A choose a_5, B guess that and choose b_4 and A would won 0 (while B will won 10)

The careful choice of player A (J_1) is a_3, which ensures him to earn at least 6 (maximin). This minimum guaranteed gain is the security level. By doing the same the player B (J_2), if he fears the extreme insight of the player A (J_1), will choose b_1. This tactic ensures a gain of 4 (maximin), which is also its security level.

Remark

The response (strategy) of a player i to the responses (strategies) of all other player is denoted by $r_i(\sigma_{-i})$ and therefore the best response to maximize the utility (payoff) U of player i is denoted by:

$$r_i(\sigma_{-i}) = \arg\max_{\sigma_i} U_i(\sigma_i, \sigma_{-i}) \qquad (15.2.25)$$

Therefore in the example above (a_3, b_1) is the best response of each player to the others one and therefore we write this:

$$r = (r_i(\sigma_{-i}), r_{-i}(\sigma_i)) = (a_3, b_1) \qquad (15.2.26)$$

If we study the game MAC/CAM by its first payoff matrix:

J_1 \ J_2	**CAM**	**MAC**
CAM	$600, 200$	$100, 100$
MAC	$100, 100$	$200, 600$

Table 15.15: **Payoff matrix of a non-zero sum game**

we can realize that there are two outstanding issues where the gain of both companies is maximum compared to other strategies (if we exclude the case taking into account probabilities and therefore "mixed strategies" that we will see much further below). These two issues are of interest for many reasons!

Indeed, the two companies have no regrets about their choice of strategy. If they consider the strategy of their opponent as inevitable, their own strategy of choice is the best possible. Remember that we say therefore that the two issues are a "Nash equilibrium". The Nash equilibrium somehow characterizes individual rationality!

The MAC/CAM game table above having two Nash equilibrium we are therefore not able, without any further information, to predict exactly what will be the outcome of the game. The two results are also equally likely.

This is how game theory reveals the most favourable social strategy to both players: that is to say both players have at least to adopt the same strategy. All the problem that remains is to know which strategy will be chosen... therefore the game will have to be a cooperative game if possible.

In the previous MAC/CAM game, the conjunction of tactics (a_3, b_1) is also a Nash equilibrium in the way that each player has no reason to unilaterally change his strategy if he wants to preserve the minimum gain. This is due to a peculiarity of this specific game! In other words there are games:

1. That do not have equilibrium.
2. That have equilibrium not corresponding to the conjunction of prudent tactics.

✍**Example:**

In this game, two players J_1, J_2 compete to stone, scissors, paper (SCP game...). In general, stone beats scissors (by blunting them), scissors beat paper (by cutting them), paper beats stone (by wrapping them). So every strategy beats another strategy, a draw against the second (equivalent) and is beaten by the third (so there is no winning strategy and there is verbatim no Nash equilibrium).

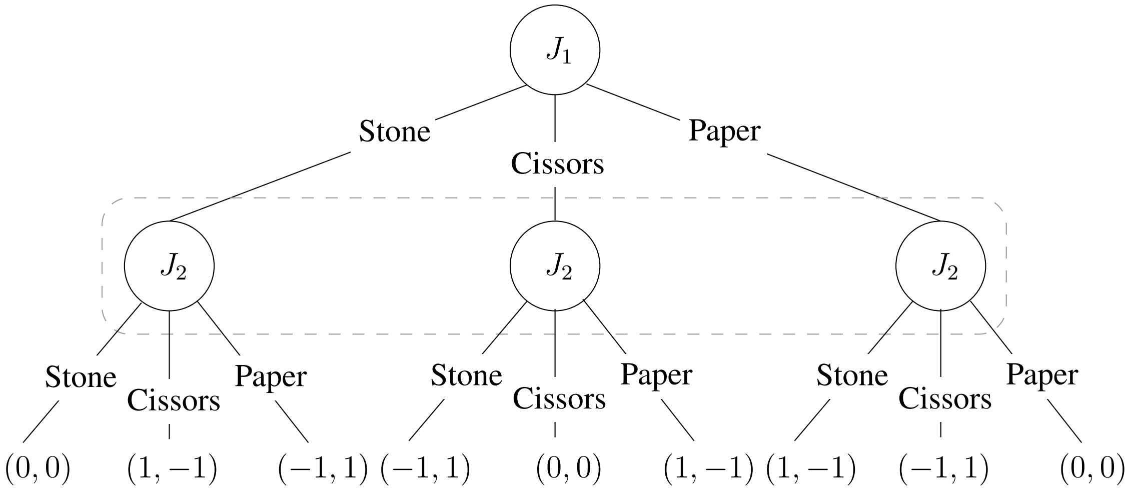


Figure 15.67: **Extensive form of the stone-scissors-paper game (Kuhn's tree)**

To reveal the simultaneity of the game on the above representation, we surrounded the sets of information. J_1 knows that J_2 has chosen one of the three items, but he does not know which one, so he does not know the exact node where his own choice will intervene, and therefore he is unable to determine the outcome of the game. The game is therefore said to have an "imperfect information".

In a normal representation form, we have:

J_2 J_1	**Stone**	**Scissors**	**Paper**
Stone	$0,0$	$1,-1$	$-1,1$
Scissors	$-1,1$	$0,0$	$1,-1$
Paper	$1,-1$	$-1,1$	$0,0$

Table 15.16: **Payoff matrix of the stone-scissors-paper zero sum game**

As we know this game is a "zero sum game" in the sense that whatever is gained by one player is lost by the other. In other words, we have already seen that we could therefore speak about a "strictly competitive game".

Because in a "zero sum game" whatever is gained by one player is lost by the other. we obviously see that we can represent the payoff matrix by a semi-matrix (therefore relatively to only one player):

J_2 J_1	**Stone**	**Scissors**	**Paper**
Stone	0	1	-1
Scissors	-1	0	1
Paper	1	-1	0

Table 15.17: **Payoff semi-matrix of the stone-scissors-paper zero sum game**

If necessary, if the respective game gains and losses does not have the same "delta", we can just define a suitable utility function for the other player such that it is always possible for any strictly competitive game where the gains are not equal and opposite to be put in the form of a half-matrix. We will prove later that there is such an utility function.

Remark

On the half-matrix of a zero-sum game, it is very easy to recognize if there is a Nash equilibrium or not. For example:

2	0
1	3

In this game, the second row is the prudent tactic of line player and column player will choose the first column as prudent tactic in which the greatest loss is not located. Therefore, the line player will have interest to move to the first row and therefore tactics are not in balanced and hence there is no Nash equilibrium!!!

In the tactical duel thus defined, the expected gain of the line player is the maximum of the minimum of the rows, that is to say the "maximin" while the expected gain of the column player is the minimum of the maximum of that is to say the "minimax".

Definitions:

D1. The "maximin", also sometimes named "Wald criterion" is a pessimistic criterion. It is indeed the purpose of this criterion to maximize the minimum result. That is to say formally in a 2-person game:

$$\mathcal{W}=\max_{d\in D}\min_{s\in S(d)} f(d,s) \tag{15.2.27}$$

where D denotes the decision space, $S(d)$ denotes the set of states associated with decision d and $f(d,s)$ denotes the payoff (outcome) associated with decision d and state s.

To implement it, it is convenient in a zero sum game and with perfect information:

1. For each decision (or strategy), to retain the minimum expected result.
2. Among the worst results, choose the higher of the worst results of the various strategies.

D2. The "maximax" according to the same logic as the previous criterion is to retain the best results of the possible strategies, therefore it is an optimistic criterion:

$$\mathcal{M}=\max_{d\in D}\max_{s\in S(d)} f(d,s) \tag{15.2.28}$$

To implement it, it is convenient in a zero sum game and with perfect information:

1. For each decision (or strategy), to retain the most highest expected result.
2. Among the best results, choose the highest of the best results of the various strategies.

D3. The "minimax" criterion is also named the "Savage criterion" and only sometimes "Von Neumann criterion" and the idea is to retain the smaller of the best results of the possible strategies:

$$\mathcal{S} = \min_{d \in D} \max_{s \in S(d)} f(d,s) \tag{15.2.29}$$

To implement it, it is convenient in a zero sum game with perfect information:

1. For each decision (or strategy), to retain the lowest expected result.
2. Among the best results, choose the highest of the best results of the various strategies.

If and only if the maximin equals minimax, their common value, which is the common expected gain of both opponents (players), is named the "game value" (we will prove this just a few lines below), and any couple formed by a such prudent tactic of the row player and prudent tactic of the column player define an equilibrium (for this reason the above example has not such equilibrium).

✍**Example:**

Consider the following semi-matrix of a zero-sum game with a Nash equilibrium:

2	4
1	3
0	4

In this game, the line 1 is the best prudent tactic of the row player and the column player will choose column 1 as prudent tactic in which the smallest losses are. Therefore, the top left cell corresponds to joint prudent tactics and corresponds as we can see it to a Nash equilibrium.

Definition (#15.2.645): In a zero-sum game we name "top" the utility (seen in the point of view of the gain or loss), which is both a minimum in its row and a maximum in its in its column (this is the case in the previous example where the equilibrium is a "top").

Theorem 15.1. Let us prove now that in any zero-sum game, if and only if the security levels of both players are opposed (the minimax equals the maximin), the combination of prudent tactics is always an equilibrium.

Proof 15.1.1. Let us take again definition of a couple formed:

- by a prudent tactic a_p for the player A, assuring him of gain at least of v (maximin).
- by a prudent tactic b_p for the player B, assuring him of gain at least of $-w$ (maximin).

In the case of a zero-sum game, we can always redefine the utility function of one of the players to obtain $v = w$ as we have seen before in order to write the half-matrix game form. Therefore, let us observe what happens (remembering that in such a game, the gain is equivalent to the loss, therefore by extension when the gain is minimal loss for the first player then the loss is minimal for the other player):

The couple (a_p, b_p), like any couple that contains a_p ensures A to earn at least v and verbatim B to earn at least $-v$ (since $v=w$).

A has therefore has no interest to deviate unilaterally from the strategy a_p, since B has guaranteed himself to lose at maximum v in the strategy of A. Similarly, B has no interest to unilaterally deviate from the strategy b_p since A has guaranteed himself to earn at least v.

Therefore, in the case where security levels of the two players are equal and opposite, the combination is a prudent tactical balance.

We have already seen previously an example where the levels were not exactly opposed.

■ Q.E.D.

Let us now see a famous example that summarize some concepts see so far and that has some important implication in Economy named the "prisoner's dilemna".

✍**Example:**

The prisoner's dilemma is a standard example of a game that shows why two completely "rational" individuals might not cooperate, even if it appears that it is in their best interests to do so. The prisoner's dilemma is therefore a scenario in which the gains from cooperation are larger than the rewards from pursuing self-interest. It applies well to concept of "oligopoly" that we will introduce in the section of Economy (see page 7492).

The story behind the prisoner's dilemma goes like this:

Two co-conspiratorial criminals are arrested. When they are taken to the police station, they refuse to say anything and are put in separate interrogation rooms. Eventually, a police officer enters the room where Prisoner A is being held and says: «*You know what? Your partner in the other room is confessing. So your partner is going to get a light prison sentence of just one year, and because you're remaining silent, the judge is going to stick you with* 8 *years in prison. Why don't you get smart? If you confess, too, we'll cut your jail time down to* 5 *years, and your partner will get* 5 *years, also*». Over in the next room, another police officer is giving exactly the same speech to Prisoner B. What the police officers do not say is that if both prisoners remain silent, the evidence against them is not especially strong, and the prisoners will end up with only 2 years in jail each.

To understand the dilemma, first consider the choices from Prisoner A's point of view. If A believes that B will confess, then A ought to confess, too, so as to not get stuck with the 8 years in prison. But if A believes that B will not confess, then A will be tempted to act selfishly and confess, so as to serve only 1 year. The key point is that A has an incentive to confess regardless of what choice B makes! B faces the same set of choices, and thus will have an incentive to confess regardless of what choice A makes. Confess is considered the dominant strategy or the strategy an individual (or firm) will pursue regardless of the other individual's (or firm's) decision. The result is that if prisoners pursue their own self-interest, both are likely to confess, and end up doing a total of 10 years of jail time between them.

It is implied that the prisoners will have no opportunity to reward or punish their partner other than the prison sentences they get, and that their decision will not affect their reputation in the future and they have no loyalty to each other.

Because betraying a partner offers a greater reward than cooperating with them, all purely rational self-interested prisoners would betray the other, and so the only possible outcome for two purely rational prisoners is for them to betray each other. The interesting part of this result is that pursuing individual reward logically leads both of the prisoners to betray, when they would get a better reward if they both kept silent.

For summary, the prisoner's dilemma is a two-player non-zero-sum game (so cooperation may be worthwhile), but does not have perfect information (so cooperation is tricky). It has a Nash equilibrium (mutual defection) that is not Pareto optimal. Defection is always a dominant strategy.

What is important to remember in this game is that the members of an oligopoly (see section Economy page 7492) can face a prisoner's dilemma, also! If each of the oligopolists cooperates then high monopoly profits are possible!!!

15.2.3.3.1 Repetitive Games

Suppose that a man J_2 and a woman J_1 go see a movie. Once there, they must choose between going to see a documentary or a comedy. One of them prefers documentaries and the other comedies, but both prefer to see a movie together rather than separately: it's... the war of the sexes (WoS...).

The strategies available to each player, considering that they make their choices simultaneously (which is unlikely in a real case, gallantry forcing to desynchronise the game in favour of women ☺) are then:

1. Go see a documentary, what that we will denote by Doc
2. Go see a comedy, what we will denote by Com

The payoff matrix will then be:

J_2 J_1	Doc	Com
Doc	$2,3$	$1,1$
Com	$1,1$	$3,2$

Table 15.18: **Payoff matrix of the WoS repetitive game**

That can be rewritten in the following form:

J_2 / J_1	Doc	Com
Doc	$1,2$	$0,0$
Com	$0,0$	$2,1$

Table 15.19: Payoff equilibrated matrix of the WoS repetitive game

First point, we can notice that the WoF is not a strictly competitive game (therefore we don't need to try to represent it in the form of a half-matrix) and it is a coordination game. Second point, we note that the two issues with the maximum gain are Nash equilibrium (so we can not predict the outcome of the game). Therefore it is a game with multiple Nash equilibrium.

Suppose now that the couple returns to the cinema the following week, and they must make this choice again. We can represent this new situation with a game, which is actually a repetition of WoS, let us denote it by WoS2.

If we consider that during the second stage each player knows what the other chose in the first step, the available strategies are now conditional strategies: the new strategy can take into account of moves played by the opponent during the preceding game!!!

The description of these strategies follows the following pattern: if we play S_1 the first time, then if the other chose the documentary Dom at the first movie, then we play S_{2A}, otherwise S_{2B} with S_1, S_{2A}, S_{2B} taking their value in the set $\{Doc, Com\}$. We denote these strategies by:

$$(S_1, S_{2A}|\,[S_1, Doc]\,;S_{2B}|\,[S_1, Com]) \tag{15.2.30}$$

We can read this notation as follows: we play S_1, then if we find ourselves in the situation $[S_1, Doc]$, then we play S_{2A}, or if we find ourselves in the situation $[S_2, Com]$, then we play S_{2B}. In this case WoS2 we have obviously 8 strategies:

1. $(Doc, Doc|\,[Doc, Doc]\,;Doc|\,[Doc, Com])$: We always choose the documentary (aggressive strategy).
2. $(Doc, Doc|\,[Doc, Doc]\,;Com|\,[Doc, Com])$: We always choose the documentary, excepted if the first time we found ourselves alone (then to be nice we choose the comedy).
3. $(Doc, Com|\,[Doc, Doc]\,;Doc|\,[Doc, Com])$: We choose always the documentary, except if the first time we've both have chosen the documentary (logic strategy).
4. $(Doc, Com|\,[Doc, Doc]\,;Com|\,[Doc, Com])$: The first time we choose the documentary and the second time the comedy (win-win strategy).
5. $(Com, Doc|\,[Com, Doc]\,;Doc|\,[Com, Com])$: The first time we choose the comedy and the second time the documentary (win-win strategy).
6. $(Com, Doc|\,[Com, Doc]\,;Com|\,[Com, Com])$: We always choose the comedy, excepted if the first time we found ourselves alone (then to be nice we choose the documentary).

7. $(Com, Com| [Com, Doc]\,; Doc| [Com, Com])$: we choose the comedy unless the first time, we both chose the comedy.

8. $(Com, Com| [Com, Doc]\,; Com| [Com, Com])$: We always choose the comedy (aggressive strategy).

For each end of WoS2, utility vectors are determined by the sum of vectors obtained for each step considered as end of WoS. We say that WoS is a "super game" which WoS is the "constitutive game".

Definition (#15.2.646): A "subgame perfect equilibrium" corresponds to a strategic combination of actions such that every subgame are Nash equilibrium.

Remark

Obviously a "subgame" is simply a subtree of the game tree.

Let us now see all these concepts in an ensemblist (Set Theory) way (hang on well!!!).

15.2.3.4 Set representation of a Game

So far we have seen that there are a given number of elements that make up a game: the players, actions and strategies of the players, the sequences and steps of the game, game results and information available by the players at each choice of action.

So far we have seen that there are a number of elements that make up a game: the players, the actions and strategies of the players, the sequences and stages of the game, the game results and the information available to the players each time they choice an action.

D1. The rules of a game indicate or must indicate:

- The succession of stages of the game, and the order in which the players are involved.
- The actions that are allowed at each step.
- The information available to the player everytime it needs to make a decision.

We have seen so far that there are two forms of possible representations for a game. One of them uses a tree (an extensive form) and the other a table (normal form). Under a formal expression it gives:

D2. We have already defined previously the Set point of view of a normal form of a finite (non-probabilistic) game. Remember it was defined formally by (we change the notation a little bit):

$$\boxed{G^N = \left[J, \{S_i\}_{i\in J}, \{U_i\}_{i\in N}\right]} \tag{15.2.31}$$

with for recall:

(a) Players: A set of players (agents who play the game) $J=\{1,\ldots,n\}$ with typical element (player) $i\in J$.

(b) Strategies: For each player i a non-empty set of feasible strategies S_i with typical element is $s_i\in S_i$.

(c) Payoffs: For each player $i\in J$ a payoff (utility) function:

$$U_i: \quad S_i \mapsto \mathbb{R} \tag{15.2.32}$$

Then for a finite game G^N of length K (the number of finite steps so that the game is surely finished for one player) and players $n=2$, a strategy for player 1 is a function:

$$\sigma : \{s\in\{S_1\} : |s| \text{ is even}\} \tag{15.2.33}$$

a strategy for player 2 is function:

$$\tau : \{s\in\{S_2\} : |s| \text{ is odd}\} \tag{15.2.34}$$

Notice that it is Player 1 turn to move if and only if the sequence of preceding moves is even, and Player 2 turn if and only if it is odd.

Remark

If it may help, a "finite game":

- Has a beginning and an end
- Is played with the goal of winning (so there is always a winner)
- Rules exist to ensure the game is finite

and a "infinite game":

- Do not necessarily have a knowable end point (no definite winner)
- Repeated interaction continue for an infinite amount of time
- The rational long term behaviour is affected by endless moves of the players
- The players and the rules are not fixed

D3. Formally a game tree (extensive form game) can be defined in multiples ways as there are not yet a convention between all practitioners. In our point of view the most complete one for finite and non-probabilistic games requires first the following definitions:

(a) A finite set of nodes N.

(b) A unique initial node $n^0 \in N$.

(c) A set of terminal nodes $T \subset N$.

(d) A set of decision nodes such as $D = N \setminus T$ that is to say $D \cap N = \varnothing$.

(e) An immediate predecessor function p such that $p : N \mapsto D$.

(f) A set of players (agents who play the game) $J = \{1, \ldots, n\}$ with typical element (player) $i \in J$.

(g) For each player i a non-empty set of feasible strategies S_i with typical element is $s_i \in S_i$.

(h) For each player $i \in J$ a payoff (utility) function U_i.

And therefore we can define a game tree as:

$$\boxed{G^T = \left[\left\{N, n^0, T, D, p\right\}, J, \left\{S_i\right\}_{i \in J}, \left\{U_i\right\}_{i \in N}\right] = \left[\mathcal{T}, J, \left\{S_i\right\}_{i \in J}, \left\{U_i\right\}_{i \in N}\right]} \tag{15.2.35}$$

D4. A "pure strategy" s for a player is an application of the whole set of strategies S_i of this player to the set S_i of all decisions of the game such that:

$$\forall s_i \in S_i, p(s(s_i)) = s_i \tag{15.2.36}$$

Simply said, a "pure strategy" is a strategy that avoids any form of hazard, and is therefore completely deterministic.

A pure strategy therefore provides a complete definition of how a player will play a game. In particular, it determines the move a player will make for any situation he or she could face. A player's strategy set is the set of pure strategies available to that player.

D5. A "mixed strategy" for a player is a probability distribution $(P_1, P_2, \ldots, P_n)$ with:

$$\sum_i P_i = 1 \tag{15.2.37}$$

on the set of all pure strategies $S_i = (s_1, s_2, \ldots, s_n)$ of player $j = i$.

✍**Example:**

Penalty kicks are a form of mixed-strategy game. Indeed, the goalkeeper has to anticipate the shot and can almost not analyse it quick enough. He must randomly choose if it will remain in the middle, or go to the left or right. Same thing for the attacker (usually the guardian should start just when the attacker kick) that can not know where the goalkeeper will jump will kick randomly.

Obviously when a player no dominant strategy, he should consider playing a mixed strategy. We will see further below an application of such a situation involving probabilities (be patient!).

Remarks

R1. A pure strategy can obviously be regarded as a strategy that gives the probability 1 at the strategy s_i and 0 to all others.

R2. In our definition of the set of strategies S_i of player $j = i$, there is a finite number of strategies for each agent but in economics, sets of strategies are often continuous and contain an infinite number of possible strategies (choice of quantity, price, etc.).

Naturally, the result obtained by the player can therefore not be guarantee anymore with certainty, since the selection process of the decision involves probabilities. We can only get an expected mean!

A pure strategy is thus a strategy by choosing among of all mixed strategies and that will be used thereof for all the duration of the game. A player using a mixed strategy against a player using a pure strategy will use (or will be forced) therefore to use also a pure strategy, but will not always use the same pure strategy every time they meet on the same game.

D6. A "strategy profile" (sometimes named a "strategy combination") is a set of strategies for all players which fully specifies all actions in a game. A strategy profile must include one and only one strategy for every player.

While a mixed strategy assigns a probability distribution over pure strategies, a behaviour strategy assigns at each information set a probability distribution over the set of possible actions. While the two concepts are very closely related in the context of normal form games, they have very different implications for extensive form games. Roughly, a mixed strategy randomly chooses a deterministic path through the game tree, while a behaviour strategy can be seen as a stochastic path.

(a) A game is a "purely concurential" or "strictly competitive" if:

$$\boxed{\forall i, j \in I \times I \exists m, n \in N \times N : \quad (U_m(j) - U_m(i))(U_n(j) - U_n(i)) < 0} \tag{15.2.38}$$

Therefore a game is strictly competitive if for a set of pairs of outgoings, the gains of at least one of the player decrease whatever the strategy. If the both player have for a pair of outgoing, their respective gains that increase or decrease, then we have:

$$\underbrace{(U_m(j)-U_m(i))}_{>0}\underbrace{(U_n(j)-U_n(i))}_{>0}>0 \quad (15.2.39)$$

(b) A strictly competitive game is a "zero-sum game" if:

$$\boxed{\forall i\in I: \quad \sum_{n\in N} U_n(i)=0} \quad (15.2.40)$$

As we already know, we are dealing with a zero sum when the players' interests are diametrically opposed. In a two-player zero-sum game, for example, what is gained by one is lost by the other. This term has for original parlour games like poker where a player who wants to make money should do it at the expense of the others. Chess is also a zero sum game.

D8. A "super game" is defined by:

$$G^S=\{G^N,T,\Omega\} \quad (15.2.41)$$

and is made by a constitutive game $G^N(J,\{S_i\}_{i\in J},\{U_i\}_{i\in N})$, and a number of iterations T and of the vector $\Omega=(\omega_1,\ldots,\omega_n)$ of the discount rate of utility, ω_i being the discount rate of the player i (most of times taken as being equal to 1).

Thus, as we have already mentioned it during our study of the WoS2 repetitive game, we consider that at a stage t the choice dictated by a strategic combination s to the player n is denoted by $s_{n,t}$ and that the utility, for this same player, obtained at this stage of the game, that is to say the utility corresponding to the constituting game is denoted by $U_n(s_{n,t})$, then the utility associated with the outcome of the super game is given by:

$$U_n=\sum_{t=0}^{t=T-1}\omega_n^t U_n(s_{n,t}) \quad (15.2.42)$$

15.2.3.5 Graphical representation of a Game

We now have cumulated enough evidence to have a probabilistic and operational approach relatively to simple zero-sum games.

As it is still relatively difficult not to be too theoretical for this field of mathematics, to remain understandable we will study the theory through a unique "big" example.

Let us consider two companies that we will name respectively $S1$ and $S2$ which are specialized in large-scale sale of a certain product and form a bilateral oligopoly in perfect competition (see section Economy page 7492). The company $S1$ decided to invest in a new market, consisting of a set of areas of comparable importance.

The penetration in different regions is done through the installation of a stand in chains stores $C1$ or $C2$ in each region. To better motivate its retailers, the company $S1$ will choose at the final stage only one distribution channel ($C1$ or $C2$) by region to sell its products.

Following a market survey (we must get at least some number to do some math ...), the company $S1$ learn that its gains relatively to that of its competitor would be those as shown in the table below:

$S2$ / $S1$	**C1**	**C2**	**NP**
C1	0	2	4
C2	6	-3	8
NP	-3	-5	0

Table 15.20: **Game preparation for graphical analysis**

The company $S2$ arrives at the same result following a market study (we simplify thanks to this hypothesis the analysis of the problem).

Remarks

R1. Since everything that wins a competitor would be lost by the other, the game is zero sum game (hence the fact that there is only one value in each cell)!

R2. We will assume that the two companies can not and do not want to communicate with each other, in other words it is a non-cooperative game.

Let us start by analysing what are the strategies that have no interest for the both companies.

For that, let us look if there is a strategy that will never be chosen by $S1$ whatever the strategy of $S2$:

1. If $S2$ choose $C1$ then $S1$ will have for best interest to choose $C2$.
2. If $S2$ chooses $C2$ then $S1$ will have best interest to choose $C1$.
3. If $S2$ chooses NP then $S1$ will have for best interest to choose $C2$.

We see here that whatever is the choice of $S2$, the company $S1$ will never choose NP. Therefore the strategy NP for $S1$ is totally dominated and can be therefore eliminated.

Similarly, let us look if there is a strategy that will never be chosen by $S2$ whatever the strategy $S1$:

1. If $S2$ choose $C1$ then $S1$ will have for best interest to choose $C2$.
2. If $S2$ chooses $C2$ then $S1$ will have best interest to choose $C1$.
3. If $S2$ chooses NP then $S1$ will have for best interest to choose $C2$.

We see here that whatever is the choice of $S1$, the company $S2$ will never choose NP. Therefore the strategy NP for $S2$ is totally dominated and can be therefore eliminated.

The table can therefore be simplified in the following way:

$S2$ $S1$	**C1**	**C2**
C1	0	2
C2	6	-3

Table 15.21: **Simplification of the game for graphical analysis**

Moreover, this game contains no Nash equilibrium (thus no pure strategy is advantageous). This game is therefore without equilibrium. Indeed, if $S1$ choose $S1$, the $S2$ has better for interest to choose $C1$. But then $S1$ has better interest to play $C2$. But $S2$ has now better interest to choose rather $C2$. Which gives $S1$ again the desire to choose $C1$...

For this purpose let us denote by p and q the frequencies with which the companies $S1$ and $S2$ choose the store chain $C1$:

$S2$ $S1$	**C1**	**C2**	
C1	0	2	p
C2	6	-3	$1-p$
	q	$1-q$	

Table 15.22: **Parametric form of the game for graphical analysis**

These probabilities must be interpreted as follows:

1. If p and q are equal to 1 then for all regions, it will be the store $C1$ that will get the business market.
2. If p and q are equals for example to $9/11$ and $5/11$ it would mean that the company $S1$ will give the right to sales to the store chain $C1$ in 9 regions on 11 (the remaining 2 being for the store chain $C2$) and respectively the company $S2$ will give the right to sell the store chain $C1$ in 5 regions on 11 (the remaining 6 being for $C2$).

So let us begin our study. We will put us in an analysis point of view in which the company $S1$ is looking for a mixed strategy to maximize its gain (or utility) that we will denote by v and to know the mixed strategy of the company $S2$ so that it minimizes its loss v (since it is a zero sum game all what one wins another one will loose it).

The system of equations will therefore be naturally for the company $S1$:

$$\begin{gathered}\max(v)\\ 6\cdot(1-p)+0\cdot p\geq v\\ 2\cdot p-3\cdot(1-p)\geq v\\ 0\leq p\leq 1\end{gathered}\tag{15.2.43}$$

and for the company $S2$:

$$\begin{gathered}\min(v)\\ 2\cdot(1-q)+0\cdot q\leq v\\ 6\cdot q-3\cdot(1-q)\leq v\\ 0\leq 1\leq 1\end{gathered}\tag{15.2.44}$$

Now, we fall back here on a remarkable situation. Indeed, we have only here two standard forms of linear programming (see section Numerical Methods page 6646). We saw in our study of the latter that when there is only one unknown by system then it is possible, to pass a to graphical resolution without using the simplex algorithm (graphical resolution is easy to use in 2D or 3D situations).

After simplification we get:

$S1$	$S2$
$\max(v)$	$\min(v)$
$v\leq -6p+6$	$v\geq -2q+2$
$v\leq 5p-3$	$v\geq 9q-3$
$0\leq p\leq 1$	$0\leq 1\leq 1$

and the corresponding graphical representations of v as a function of p and respectively q:

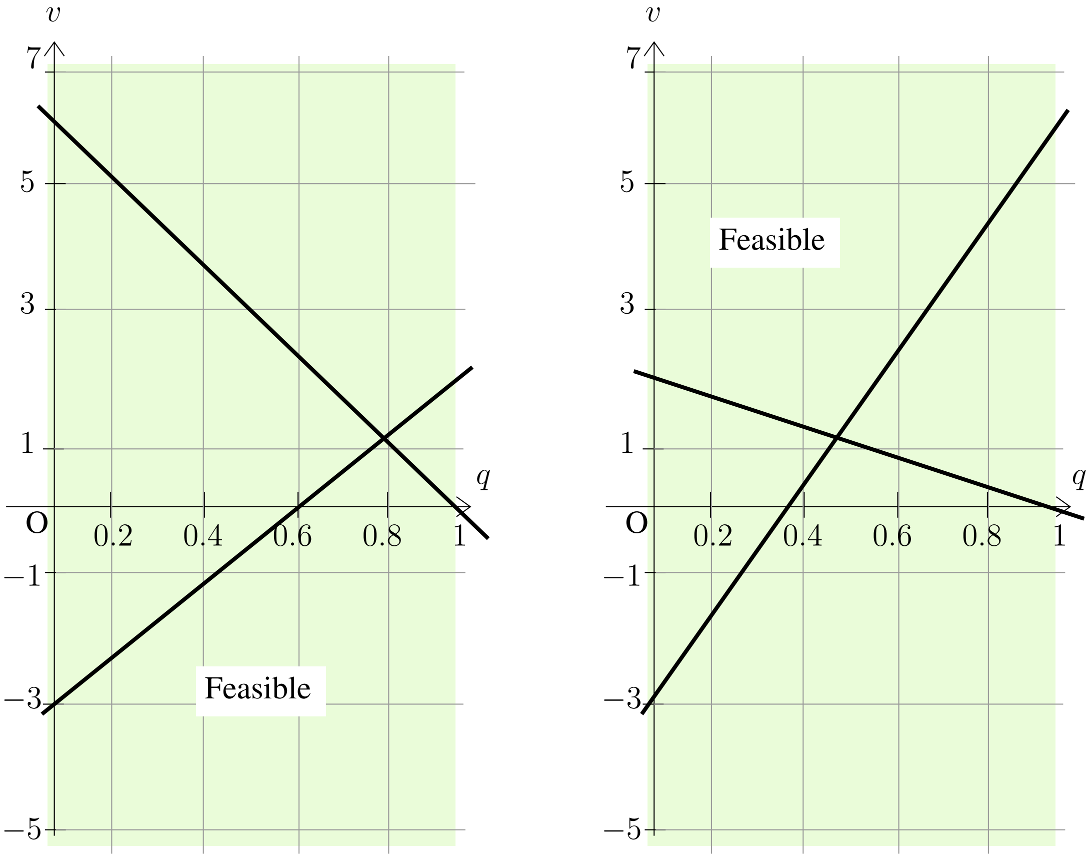


Figure 15.68: **Correspondence of the inequalities of the optimization problem**

By solving the simplex algorithm (see section Numerical Methods page 6648) we have for optimal values for the two respective systems (it is also possible to read the approximate value on the graphs above but nevermind):

$S1$	$S2$
$v=12/11$ $p=9/11$	$v=12/11$ $q=5/11$

The company $S1$ can therefore guarantee an average gain v (we should talk about "mean" rather to be rigorous) of $12/11$. Indeed:

$$v=\mathrm{E}(U_i)=\begin{cases}=6(1-p)+0\cdot p=6\left(1-\dfrac{9}{11}\right)=\dfrac{12}{11}\\ =2p-3(1-p)=2\dfrac{9}{11}-3\left(1-\dfrac{9}{11}\right)=\dfrac{12}{11}\end{cases}\tag{15.2.45}$$

and the probability p is giving in fact the distribution between the store chain $C1$ that will have $9/11$ of the market of $C2$ the rest $2/11$ (the sum being of course equal to 1).

The company $S2$ may therefore also ensure an average gain v of $12/11$. Indeed:

$$v=\mathrm{E}(U_i)=\begin{cases}=2(1-1)+0\cdot q=2\left(1-\dfrac{5}{11}\right)=\dfrac{12}{11}\\[2ex] =6q-3(1-q)=6\dfrac{5}{11}-3\left(1-\dfrac{5}{11}\right)=\dfrac{12}{11}\end{cases} \tag{15.2.46}$$

and the probability q is giving in fact the distribution between the store chain $C1$ that will have $5/11$ of the market of $C2$ the rest $6/11$ (the sum being of course always equal to 1).

15.2.4 Expected Utility

To introduce expected utility that is used a lot by undergraduate project managers let us consider the following null-sum non-cooperative game:

J_1 \ J_2	S_1	S_2
S_1	0	2
S_2	3	1

Table 15.23: **Payoff matrix without Nash Equilibrium**

that does not have any equilibrium as we have seen earlier above. In this kind of game, any recommendation to a player to choose one tactic over another can harm him, since the opponent is informed, or can guess that recommendation.

Indeed, if J_1 think that J_2 will choose the strategy 1, he has better to chose the strategy 2 (utility 3 against 0). But then, if J_2 think that J_1 will choose the strategy 2, he has better to chose the strategy 2 (loss of 1 instead of 3). Then, if J_1 think that J_2 will choose the strategy 2, he has better to chose the strategy 1 (utility 2 against 1). But then, if J_2 think that J_1 will choose the strategy 1, he had better to chose the strategy 1 (loss of 0 instead of 3). And the circle is completed...

Ultimately, the thing that matters the most in a non-cooperative game is that the strategy of a player can not be guessed by his opponent. Like any reasoning eventually be anticipated, the opponents being perfectly rational and informed, the only conceivable solution is to rely on a specific process, relied on probabilities assigned to various possible tactics. Thus, as we have defined it earlier above, this game has a "mixed strategy".

Naturally, the result obtained by the player can not be guaranteed as the process of choice of the decision involves probabilities. Compare results therefore is like comparing lotteries. We imagine the situation of an admiral that has to answer to a military court for the loss of a ship, and explaining that he made his decision in playing dice (assuming a non-cooperative battle without Nash equilibrium): even if it is full compliant with the requirements of game theory, this explanation will be low efficient argument!

15.2.4.1 Hurwitz Criterion

So we need to introduce a probabilistic utility that we also sometimes name the "Hurwitz criterion". Let us consider a two-strategies S_1, S_2 game and let us note the respective utility by:

$$U[S_1P, S_1(1-P)] \tag{15.2.47}$$

which allows to get S_1 with a probability P and S_2 with a probability $1-P$. This relation is written with obvious notation (see section Statistics page 585):

$$\text{E} = U[S_1P, S_2(1-P)] = PU(S_1) + (1-P)U(S_2) \tag{15.2.48}$$

where E is obviously the "expected utility" (in similarity with the concept of expected mean studied in the section Statistics) or "expected early gain".

We can already notice that, if there is such an utility (expected), there are an infinite to a given arbitration, obtained from U by an affine transformation strictly increasing, that is to say, a relation of the form:

$$V = aU + b \tag{15.2.49}$$

with $a > 0$. Indeed, the prior-previous relation gives for $a > 0$:

$$a\text{E} = aU[S_1P, S_2(1-P)] = PaU(S_1) + (1-P)aU(S_2) \tag{15.2.50}$$

which, added term by term to the obvious relation (required):

$$b = Pb + (1-P)b \tag{15.2.51}$$

leads to (do not confuse with the notation of the variance!):

$$V[S_1P, S_2(1-P)] = PV(S_1) + (1-P)V(S_2) \tag{15.2.52}$$

This proves among other what we claimed earlier: we can always choose a utility function (even in a perspective of pure strategy where $P = 0$ or $P = 1$) such that the delta of gains of the players in the zero-sum games are equal and opposite.

Remark

The expected utility (or "Hurwitz criterion") coincides with the criterion of the maximin when $P = 0$ and maximax when $P = 1$ (see below).

Let us see first an academic example by considering the following zero-sum game:

J_1 \ J_2	b_1	b_2
a_1	5	2
a_2	3	4

We see in this game that there is no Nash equilibrium. Indeed, if J_2 think that J_1 will decide a_1, he has better to chose b_2 (loss of 2 instead of 5). But J_1 understanding this will change for a_2

(gain of 4 instead of 2). But J_2 guessing this will change to b_1 (loss of 3 instead of 4), and J_1 that has understand everything will go back to a_1 (gain of 5 instead of 3).

Let us now consider that the player J_1 will choose a number x between 0 and 1, and thus take the decision a_1 with the probability x and a_2 with the probability $1-x$. Similarly, the player J_2 will choose a number y between 0 and 1 and thus take the decision b_1 with the probability y and b_2 with the probability $1-y$.

The results of these joint decisions are then:

- The gain of 5, resulting from the combination of the decisions (a_1, b_1), is obtained with the probability xy (decisions are independent for both players!).
- The gain of 2, resulting from the combination of the decisions (a_1, b_2), is obtained with the probability $x(1-y)$ (decisions are still independent for both players!).
- The gain of 3, resulting from the combination of the decisions (a_2, b_1), is obtained with the probability $(1-x)y$ (decisions are still independent for both players!).
- The gain of 4, resulting from the combination of the decisions (a_2, b_2), is obtained with the probability $(1-x)(1-y)$ (decisions are still independent for both players!).

The expected gain (or utility) of the player J_1 is therefore:

$$\mathrm{E}(J_1)=5xy+2x(1-y)+3y(1-x)+4(1-x)(1-y)=x(4y-2)+4-y \tag{15.2.53}$$

Remark

We see well that if $x=0$ (and $y=1$) then we fall on the criterion of Minimax (the maximum gain of the more pessimistic strategies) that is to say $\mathrm{E}(J_1)$ is equal to 3. Similarly if $x=1$ (and $y=1$) then we fall on the criterion of Maximax (the maximum gain the more optimistic strategies).

If there is an equilibrium between the probabilistic strategies. J_1 will have no reason to change the value of x in the hope of increasing $\mathrm{E}(J_1)$. Therefore, the derivative with respect to x must be zero as (maxima):

$$\frac{\mathrm{dE}(J_1)}{\mathrm{d}x}=4y-2=0\Rightarrow y=\frac{1}{2} \tag{15.2.54}$$

In these conditions:

$$\mathrm{E}(J_1)=\frac{7}{2} \tag{15.2.55}$$

To examine what is offer to J_2 as opportunities, whose expected mean, let us recall it, will be in a zero-sum game necessarily opposed to that of J_1, we write:

$$\mathrm{E}(J_2)=-\mathrm{E}(J_2)=-x(4y-2)-4+y=y(1-4x)-4+2x \tag{15.2.56}$$

Applying the same reasoning (but implicitly for the minima):

$$\frac{\mathrm{dE}(J_2)}{\mathrm{d}x}=1-4x=0\Rightarrow x=\frac{1}{4} \tag{15.2.57}$$

In this case:

$$\mathrm{E}(J_2)=-\frac{7}{2} \tag{15.2.58}$$

Thus, we have determined the probabilities of the strategies that maximize the expected gains of this non-cooperative game! By adopting them J_1 is certain of an expected gain of at $7/2$ (since J_2 has nothing to gain to change its strategy) and J_2 is certain go get a gain of at least$|-7/2|$. The value $7/2$ (absolute value) is the "value of the game".

Definition (#15.2.647): If the value of a non-cooperative game with mixed strategy is the same for both players, then we say that this is a "mixed strategy equilibrium" (neither player interest to deviate unilaterally).

This result is certainly the most remarkable so far in out study of decision theory, as the non-cooperative games are the largest on the market.

15.2.4.1.1 Pascal's wager

The "Pascal's wager" is an old argument in philosophy presented by the seventeenth-century French philosopher, mathematician and physicist, Blaise Pascal (11623–11662 according to holocene calendar). It posits that humans bet with their lives that the christian god either exists or does not.

Pascal argues that a rational person should live as though that christian god exists (though similar arguments have occurred in other religious traditions) and seek to believe in this christian god. If the christian god does not actually exist, such a person will have only a finite loss (some pleasures, luxury, etc.), whereas if the christian god does exist, he stands to receive infinite gains (as represented by eternity in Heaven) and avoid infinite losses (eternity in Hell). We may generalize the wager by arguing that: whatever belief makes you avoid an infinite loss, believe in it (we see then by generalizing that the wager is very childish).

Historically, Pascal's wager was groundbreaking because it charted new territory in probability theory, marked the first formal use of decision theory, existentialism, pragmatism, and voluntarism. However, without any assumption about the probability assignment to the christian god's existence, the argument is invalid. Rationality does not require anyone to wager for the christian god if we assign probability 0 to the christian god existing, as a strict atheist might[8]. And Pascal does not explicitly rule this possibility out until a later passage, when he assumes that we assign positive probability to the christian god's existence; yet this argument is presented as if it is self-contained. His claim that «*reason can decide nothing here*» may suggest that Pascal regards this as a decision under uncertainty, which is to assume that we do not assign probability at all to the christian god's existence. If that is a further premise, then the argument is apparently valid; but that premise contradicts his subsequent assumption that we assign positive probability (it's again an evidence that logic and philosophy are not robust way to do science!).

Since there have been many religions throughout history and there are still many ($4,200$ active religions in the early 121st century (holocene calendar) and $8,000$ gods/deities), and therefore many very different conceptions of God (or gods), some assert that all of them need to be factored into the wager, in an argumentation known as the "argument from inconsistent revelations". This, as Pascal proponents argue, would lead to a high probability of believing in "the wrong god", which, they claim, eliminates the mathematical advantage Pascal claimed with his wager.

---

[8] The argument assumes that the same decision matrix applies to everybody. However, it is sure that the relevant rewards are different for different people.

Furthermore for every religion that promulgates rules, there exists another religion that has rules quite of the opposite kind (as we have already pointed out in the introduction of this textbook, this is one think that science can use to test religion using the scientific method!). If a certain action leads one closer to salvation in the former religion, it leads one further away from it in the latter. Therefore, the expected value of following a certain religion could be negative. Or, one could also argue that there are an infinite number of mutually exclusive religions (which is a subset of the set of all possible religions), and that the probability of any one of them being true is zero; therefore, the expected value of following a certain religion is zero.

quote
15.2:7432.#86

> *Religion is an insult to human dignity. With or without it, you'd have good people doing good things and evil people doing evil things. But for good people to do evil things, it takes religion!*
>
> Steven Weinberg

The possibilities defined by Pascal's wager can be thought of as a decision under uncertainty with the values of the following decision matrix:

	God exists (G)	**God does not exist** $(\neg G)$
Belief (B)	$+\infty$	$-c_1^{te}$
Disbelief $(\neg B)$	$-\infty$	-$+c_2^{te}$

Table 15.24: **Pascal's wager original matrix**

Given these values, the option of living as if one god exists (B) dominates the option of living as if one god does not exist $(\neg B)$, as long as one assumes a positive probability that god exists (therefore it's Pareto Optimum). In other words, the expected value gained by choosing (B) is greater than or equal to that of choosing $(\neg B)$.

Following McClennen (11994 according to holocene calendar), Pascal's naive argument seems to be best captured as presenting the following decision matrix:

	God exists (G)	**God does not exist** $(\neg G)$
Wager for God	Gain all	Status quo
Wager against God	Misery	Status quo

In fact, according to decision theory, the only value that matters in the above matrix is the $+\infty$ (infinitely positive). Any matrix of the following type (where u_1, u_2, and u_3 are all negative or finite positive numbers) results in (B) as being the only rational decision:

	God exists (G)	**God does not exist** ($\neg G$)
Belief (B)	$+\infty$	u_1
Disbelief ($\neg B$)	u_2	-u_3

The first conclusion seems to follow from the usual calculations of expected utility (where P is our positive, non-infinitesimal probability for God's existence):

$$\text{E}(\text{wager for God})=+\infty\cdot P+u_1\cdot(1-P)=+\infty \tag{15.2.59}$$

That is, our expected utility of belief in the christian god is infinite - as Pascal puts it, «*our proposition is of infinite force*». On the other hand, our expected utility of wagering against the christian god is:

$$\text{E}(\text{wager against God})=u_2\cdot P+u_3\cdot(1-P) \tag{15.2.60}$$

This is finite. By the premise of rationality requires us to perform the act of maximum expected utility (when there is one). Therefore, probabilities requires we wage for the christian god.

However the Pascal wage fails for atheist. Indeed, for a majority of these latter: $P=0$. Therefore:

$$\text{E}(\text{wager for God})=+\infty\cdot P+u_1\cdot(1-P)=+\infty\cdot 0+u_1\cdot(1-0)=\text{UNDEF} \tag{15.2.61}$$

As we cannot do arithmetic with infinity (we have provided multiple examples in previous chapters and section), we are stucked!

Remark

But let's take a limit and see if it is true:

$$\lim_{x\rightarrow+\infty}f(x)=+\infty,\quad \lim_{x\rightarrow+\infty}g(x)=0,\quad \lim_{x\rightarrow+\infty}f(x)g(x)=? \tag{15.2.62}$$

We know two such function are $f(x)=x$ and $g(x)=1/x$. But the limit is then 1 and not 0, and hence it is not necessarily 0. We could make this limit any value that we wish, which is why $+\infty\cdot 0$ is undefined.

And:

$$\text{E}(\text{wager against God})=u_2\cdot P+u_3\cdot(1-P)=u_2\cdot 0+u_3\cdot(1-0)=u_3 \tag{15.2.63}$$

And nothing in the argument implies that $u_1>u_3$ (indeed, this inequality is questionable, as even Pascal seems to allow). In short, Pascal's wager has no pull on strict atheists!

From the experimental point of view (evidence based), we have evidence that there are a few thousands religions on Earth with a few thousands gods. Further scientists are quite confident that the Universe is infinite, therefore with infinite worlds...

So whatever we have evidence that $P\rightarrow 0_+$. Therefore:

$$\text{E}(\text{wager for the right God})=\lim_{P\rightarrow 0_+}(+\infty\cdot P+u_1\cdot(1-P))=+\infty\cdot 0_+ +u_1\cdot(1-0_+)=\text{UNDEF} \tag{15.2.64}$$

So as we see the result is undefined. However we may argue that the "infinite reward" isn't infinite. Some individual may argue that an infinite reward would be to become a God. Assuming that god wouldn't give us give such a reward, then the reward is finite and not infinite. Therefore, experimental evidence, and the previous argument leads us to:

$$\text{E(wager for the right God)} = \lim_{P \to 0_+} \left(u_4 \cdot P + u_1 \cdot (1-P)\right) = +\infty \cdot u_4 + u_1 \cdot (1-0_+) = u_1 \tag{15.2.65}$$

So the expected utility for a theist will be the same as for an atheist under the above assumptions.

Remark

There are some religions and philosophical ways of life where there are no special rewards ($u_4 = u_1$). For example: Satanism, Taoism, Confucianism, Cheondoism, Tenrikyo, Messianism, etc.

Here is a picture that seems to summarize quite well the actual status of some religions, beliefs and philosophical positions in the world in the early 121st century (holocene calendar):

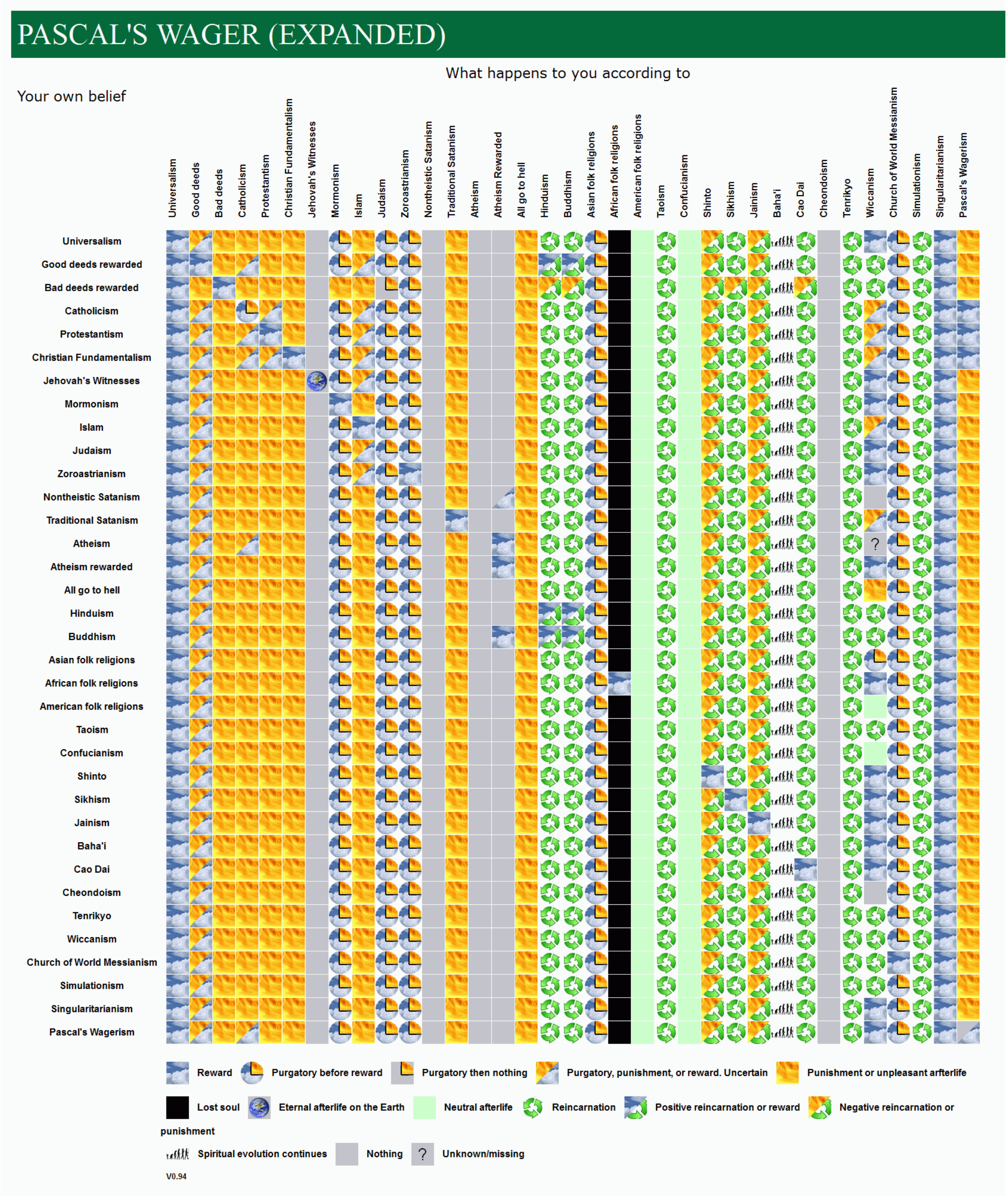
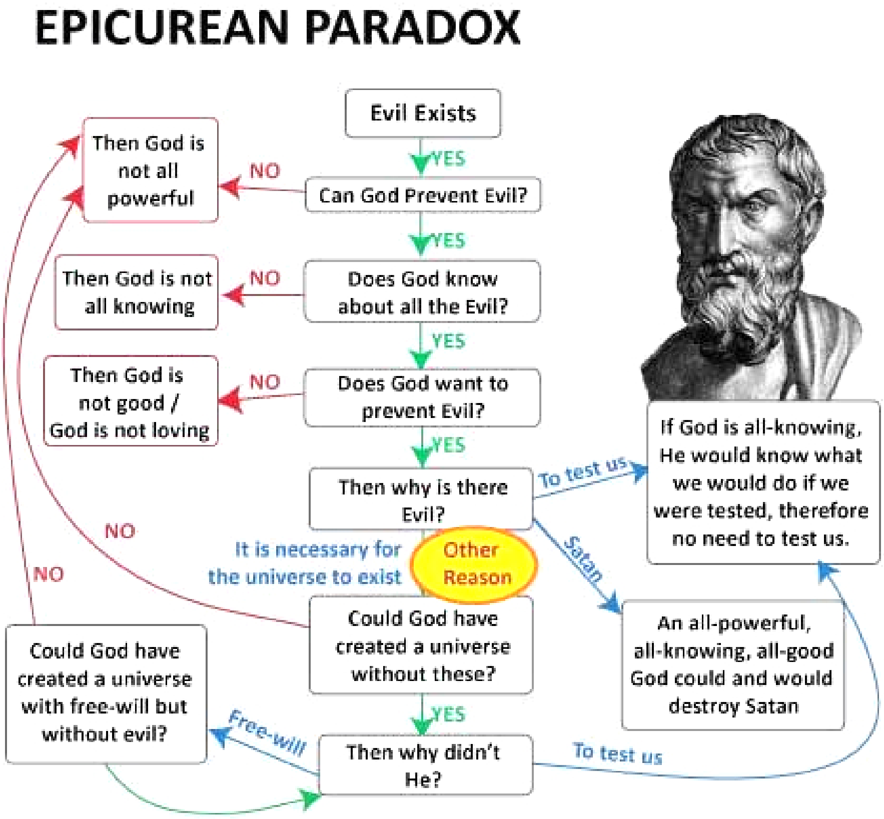

PASCAL'S WAGER (EXPANDED)
What happens to you according to
Your own belief
Universalism
Good deeds rewarded
Bad deeds rewarded
Catholicism
Protestantism
Christian Fundamentalism
Jehovah's Witnesses
Mormonism
Islam
Judaism
Zoroastrianism
Nontheistic Satanism
Traditional Satanism
Atheism
Atheism rewarded
All go to hell
Hinduism
Buddhism
Asian folk religions
African folk religions
American folk religions
Taoism
Confucianism
Shinto
Sikhism
Jainism
Baha'i
Cao Dai
Cheondoism
Tenrikyo
Wiccanism
Church of World Messianism
Simulationism
Singularitarianism
Pascal's Wagerism
Reward
Purgatory before reward
Purgatory then nothing
Purgatory, punishment, or reward. Uncertain
Punishment or unpleasant arfterlife
Lost soul
Eternal afterlife on the Earth
Neutral afterlife
Reincarnation
Positive reincarnation or reward
Negative reincarnation or punishment
Spiritual evolution continues
Nothing
?
Unknown/missing
V0.94

Figure 15.69: **Pascal's wager 121st century religions and beliefs outcomes**

We may also point out (for some religions or variations of a same religion) the "Problem of Evil" to debunk especially the Christian case of Pascal's wager. Epicurus is generally credited (even if there are not strong evidence) with first expounding the problem of evil, and it is sometimes named the "Epicurean paradox" (so a few thousands of years before Pascal's wager!). This paradox can be illustrated and expanded by the following illustration:


EPICUREAN PARADOX
Evil Exists
YES
Can God Prevent Evil?
NO
Then God is not all powerful
YES
Does God know about all the Evil?
NO
Then God is not all knowing
YES
Does God want to prevent Evil?
NO
Then God is not good / God is not loving
YES
Then why is there Evil?
To test us
If God is all-knowing, He would know what we would do if we were tested, therefore no need to test us.
Satan
An all-powerful, all-knowing, all-good God could and would destroy Satan
It is necessary for the universe to exist
Other Reason
NO
NO
Could God have created a universe without these?
YES
Then why didn't He?
Free-will
To test us
Could God have created a universe with free-will but without evil?

quote 15.2:7436.#87

> *Men never do evil so completely and cheerfully as when they do it from religious conviction!*
>
> Blaise Pascal

15.2.4.2 Laplace Criterion

The "Laplace criterion" is a criterion that assigns the same probability in the absence of information for each decision (equal probability). The idea is then also to calculate an expected gain for each decision given the assigned probability.

In other words, the Laplace criterion purpose is to determine for each strategy the expected value by assigning the same probability to each state of nature and keeping the one whose expected utility is the highest.

Let s see on an example by considering again the next zero-sum game:

J_2 J_1	b_1	b_2
a_1	5	2
a_2	3	4

By applying the equiprobability, we have the following table:

J_2 J_1	$\mathbf{E}(b_1)$	$\mathbf{E}(b_2)$
$\mathbf{E}(a_1)$	$5\frac{1}{2}+2\frac{1}{2}=3.5, 5\frac{1}{2}+3\frac{1}{2}=4$	$5\frac{1}{2}+2\frac{1}{2}=3.5, 2\frac{1}{2}+4\frac{1}{2}=3$
$\mathbf{E}(a_2)$	$3\frac{1}{2}+4\frac{1}{2}=3.5, 5\frac{1}{2}+3\frac{1}{2}=4$	$3\frac{1}{2}+4\frac{1}{2}=3.5, 2\frac{1}{2}+4\frac{1}{2}=3$

Table 15.25: **Detailed application of Laplace criterion**

The game then becomes:

J_2 J_1	$\mathbf{E}(b_1)$	$\mathbf{E}(b_2)$
$\mathbf{E}(a_1)$	$3.5, 4$	$3.5, 3$
$\mathbf{E}(a_2)$	$3.5, 4$	$3.5, 3$

Table 15.26: **Laplace criterion simplified table**

In this example, where the expected utilities is always equal for the player J_1 whatever his strategy, the player J_2 will choose the strategy where his loss is the lowest either b_2. So here we have a Nash equilibrium (without Pareto optimality).

15.2.4.3 Kelly criterion

So we have found an investment we really like. That's great, but what do we do next? In other words, after deciding what to invest in, what's the next question you should ask? If we're thinking of asking exactly how much money you should invest, then we're exactly right! Luckily, the mathematics of betting has an incredibly useful idea that we can use: the "Kelly Criterion". That latter maximises growth when the wealth compounds multiplicatively. It makes no assumption about the shape of utility! In other words, it's prescriptive, not descriptive. The Kelly criterion says that if you want to maximise growth, you should adopt log utility.

The Kelly Criterion was developed by John L. Kelly Jr. in 11956 (holocene calendar), who was a scientist at Bell Labs who developed this method through research into game theory and information theory combined with trials in Las Vegas betting and investments in the stock market.

When we're putting money in some effort to receive more money back, i.e. a bet, we will need to find out how much of your money you should wager. We wouldn't bet all our money unless we are 100% sure that you will win a bet and you would not bet any of your money unless we are 100% sure we will lose the bet. Intuitively, we would only bet some of ou money — and what *some of your money* amounts to depends on how likely we think the winning outcome is.

If we think the winning outcome is more likely to occur, we will bet more of our money, and if we think a losing outcome is more likely, we will bet less of our money. Think of the lottery — we never bet all our money on one lottery ticket because the chance of winning is infinitesimal. Using this intuition of betting more or less depending on the likelihood of a profitable outcome, we can begin understanding the exact amount to bet.

The math behind the Kelly Criterion is based on simple probability and manipulation. What is important to understand is the compounding nature of bets it assumes. What that means is that each bet and its profits feed into the next bet. If you bet 100.- and get 110.- back, that 110.- is used entirely in the next bet you place. This compounding principle is essentially the same one that compound interest uses.

The following hypothesis are required to successfully implement the Kelly criterion. They are:

H1. The expected rate of return must be greater than 0.

H2. It must be possible to reinvest the cash flows coming from investment.

H3. The existence of an independent relationship between the final cash flows must be held.

H4. The cash flows of investment must satisfy the conditions of a Bernoulli process[9].

H5. The investing activities must be carried out infinitely.

H6. It must be possible to dissemiate an investment capital in an infinitesimal size.

H7. The winning probability of p must be kept constant to investors.

Because the Kelly Criterion leads to higher wealth than any other strategy in the long run (i.e., the theoretical maximum return as the number of bets goes to infinity), it is a scientific gambling method.

The practical use of the formula has been demonstrated for gambling and the same idea was used to explain diversification in investment management. In the 12000s (holocene calendar), Kelly-style analysis became a part of mainstream investment theory and the claim has been made that well-known successful investors including Warren Buffett and Bill Gross use Kelly methods.

---

[9] This can be tested with the well-known nonparametric Pearson's χ^2 goodness of fit test!

15.2.4.3.1 Discrete time Kelly criterion

Theorem 15.2. The Kelly bet size, i.e. optimal theoretical size for a bet, is found by maximizing the expected value of the logarithm of wealth (i.e. the point of the Kelly Criterion is to maximize the expected geometric growth rate - not the expected arithmetic return and not the expected terminal wealth!!!!), which is equivalent to maximizing the expected geometric growth rate is equal to:

$$f^*=p-\frac{q}{b}=p+\frac{p-1}{b}=\frac{p(b+1)-1}{b} \tag{15.2.66}$$

where:

- f^* is the fraction of the current bankroll to wager
- p is the probability of a win
- q is the probability of a loss ($q=1-p$)
- b is the proportion of the bet gained with a win[10].

That is a special case of the following more general case:

$$f^*=\frac{p}{a}-\frac{q}{b} \tag{15.2.67}$$

where a is the fraction that is lost in a negative outcome.

Proof 15.2.1. Heuristic proofs of the Kelly criterion are straightforward (for a rigorous and general proof, see Kelly's original paper!). The Kelly criterion maximizes the expected value of the logarithm of wealth (the expectation value of a function is given by the sum, over all possible outcomes, of the probability of each particular outcome multiplied by the value of the function in the event of that outcome). We start with 1 unit of wealth and bet a fraction f of that wealth on an outcome that occurs with probability p and offers odds of b.

The probability of winning is p, and in that case the resulting wealth is equal to:

$$1+f\cdot b \tag{15.2.68}$$

Every time we win, our money grows by this amount. Since we have a p percent chance of winning on the long run, it makes sense to mathematically model the overall percentage growth from winning bets as follows (the detailed explanation if this is given in a remark box just below):

$$(1+f\cdot b)^p \tag{15.2.69}$$

So because our investment is compounded so we get exponential growth (or decay as we will see just below)!

---

[10] For example, if betting 10.- on a 2-to-1 odds bet, (upon win you are returned 30, winning you 20), then $b=\$20/\$10=2.0$

The probability of losing is $q=1-p$, and in that case the resulting wealth is equal to:

$$1-f\cdot a \tag{15.2.70}$$

Every time we loss, our money falls by this amount. Since we have a q percent chance of losing on the long run, it makes sense to mathematically model the overall percentage growth from winning bets as follows:

$$(1-f\cdot a)^q \tag{15.2.71}$$

Therefore, the geometric growth rate is given by:

$$r=(1+f\cdot b)^p\cdot(1-f\cdot a)^q \tag{15.2.72}$$

and the logarithmic geometric growth rate is then given by:

$$\log(r)=p\log(1+f\cdot b)+q\log(1-f\cdot a) \tag{15.2.73}$$

To quote from Kelly's original paper "*The reason (for the log rule) has nothing to do with the value which (the gambler) attached to his money, but merely with the fact that it is the logarithm which is additive in repeated bets*".

Remark

Our current wealth at any stage $t=T$ will be given by the number of winning N and the number of looses M (where $N+M=T$). Now let try to calculate the equivalent Compound Growth rate τ for the process at stage T:

$$\begin{aligned} r=(1+\tau)^T &= (1+f\cdot b)^N\cdot(1-f\cdot a)^M \\ &= \left((1+f\cdot b)^{\frac{N}{T}}\cdot(1-f\cdot a)^{\frac{M}{T}}\right)^T \\ &= \left((1+f\cdot b)^{\frac{N}{N+M}}\cdot(1-f\cdot a)^{\frac{M}{N+M}}\right)^T \\ &\overset{\text{behave as when }T\to\infty}{\approx} \left((1+f\cdot b)^p\cdot(1-f\cdot a)^q\right)^T \end{aligned} \tag{15.2.74}$$

Taking the derivative to find the optimum point:

$$\left.\frac{\mathrm{d}\log(r)}{\mathrm{d}f}\right|_{f=f^*}=\frac{p\cdot b}{1+f^*\cdot b}+\frac{-qa}{1-f^*\cdot a}=0 \tag{15.2.75}$$

Rearranging this equation to solve for the value of f^* gives the "discrete Kelly criterion":

$$\boxed{f^*=\frac{p}{a}-\frac{q}{b}} \tag{15.2.76}$$

Notice that this expression reduces to the simple "gambling formula" when $a=1=100\%$, when a loss results in full loss of the wager:

$$\boxed{f^*=p-\frac{q}{b}} \tag{15.2.77}$$

The latter is often written as:

$$f^*=p-\frac{q}{b}=\frac{p\cdot b-q}{b}=\frac{\text{edge}}{\text{odds}} \tag{15.2.78}$$

■ Q.E.D.

✍**Examples:**

E1. If a gamble has a 60% chance of winning ($p=0.6$, $q=1-p=0.4$), and the gambler receives 1-to-1 odds on a winning bet (i.e. $b=1$), then the gambler should bet 20% of the bankroll at each opportunity as:

$$f^*=0.6-\frac{0.4}{1}=0.2$$

in order to maximize the long-run growth rate of the bankroll.

E2. If the gambler has zero edge, i.e. if $b=q/p$ (typically the case of a coin flip game), then the criterion recommends for the gambler to bet nothing as:

$$f^*=p-\frac{q}{b}=p-\frac{q}{q/p}=0$$

However, the major difficulty in applying the criterion is that it assumes that the true probabilities p of events occurring is known to the bettor.

In reality, these are often unknown and the probabilities have to be estimated by the bettor. Unfortunately, any estimate is, by definition, likely to be wrong since it's just a guess. The criterion is sensitive to the estimated probabilities, and since the criterion maximizes the wealth doubling exponent, mistakes made in the estimated probabilities can easily ruin the bettor over time.

Now let us see an example on how to apply the above results in a naive way on a simple trading case.

✍**Example:**

We have the history of 75 trades. From these trades we have the following informations:

- Number of trades that gained money: 66
- Number of trades that lost money: 9
- Average gain per winning trade: 659.12.-
- Average loss per losing trade: $1,799.06$.-

We calculate the following needed for edge over odds:

- The win probability:

$$p=\frac{66}{75}=0.88$$

- The gain per unit bet:

$$b=\frac{659.12}{1799.06}=0.366$$

Hence:

$$f^*=p-\frac{q}{b}=p-\frac{1-p}{b}=0.88-\frac{0.88}{0.366}\cong 0.552$$

15.2.4.3.2 Continuous time Kelly criterion

The formula we have derived above is in fact the "discrete Kelly criterion". That's not so useful in trading, where the outcomes are continuous! So let us see how to derive it for the continuous case!

We have strategy with a return stream r_i, where r_i are independent identically distributed random variables with such properties:

$$\begin{aligned}-1\leq r_i &\leq +1\\ \mathrm{E}\left(r_i\right)&=\mu_r\\ \mathrm{V}\left(r_i\right)&=\sigma_r^2\end{aligned} \tag{15.2.79}$$

C_0 is the total initial capital, C_t is the total capital after t trades, f is the fraction of capital we invest in every trade:

$$\begin{aligned}&C_1=C_0\left(1+f\cdot r_1\right)\\ &C_2=C_1\left(1+f\cdot r_2\right)=C_0\left(1+f\cdot r_1\right)\left(1+f\cdot r_2\right)\\ &\ldots\\ &C_t=C_0\left(1+f\cdot r_1\right)\left(1+f\cdot r_2\right)\ldots\left(1+f\cdot r_t\right)=C_0\prod_{i=1}^{t}\left(1+f\cdot r_i\right)\end{aligned} \tag{15.2.80}$$

At period t we define the "gain" G as the ratio between the capital C_t and the initial capital C_0 :

$$G=\frac{C_t}{C_0}=\prod_{i=1}^{t}\left(1+f\cdot r_i\right) \tag{15.2.81}$$

Since C_t depends on the random variables r_i, it is itself a random variable.

So we want to find the fraction f^* that maximises the expected value of G, i.e.:

$$\begin{aligned}\mathrm{E}(G)&=\mathrm{E}\left(\prod_{i=1}^{t}\left(1+f\cdot r_i\right)\right)\\ f^*&=\mathrm{argmax}_f\mathrm{E}(G)\end{aligned} \tag{15.2.82}$$

This is equivalent to the easier task of maximizing $\mathrm{E}\left(\log(G)\right)$:

$$\log(G)=\log\left(\prod_{i=1}^{t}\left(1+f\cdot r_i\right)\right)=\sum_{i=1}^{t}\log\left(1+f\cdot r_i\right) \tag{15.2.83}$$

Therefore:

$$\mathrm{E}\left(\log(G)\right)=\mathrm{E}\left(\sum_{i=1}^{t}\log\left(1+f\cdot r_i\right)\right)=\sum_{i=1}^{t}\mathrm{E}\left(\log\left(1+f\cdot r_i\right)\right)=t\mathrm{E}[\log(1+f\cdot r)] \tag{15.2.84}$$

So:

$$f^{*}=\operatorname{argmax}_f \mathrm{E}\left(\log(G)\right)\equiv\operatorname{argmax}_f t\mathrm{E}[\log(1+f\cdot r)]=\operatorname{argmax}_f\mathrm{E}\left(\log(1+f\cdot r)\right) \tag{15.2.85}$$

So to find the optimum we seek the derivative such that:

$$\frac{\mathrm{d}}{\mathrm{d}f}\mathrm{E}\left(\log(1+f\cdot r)\right)=0 \tag{15.2.86}$$

For an easier differentiation, we use a well known second order Taylor series expansion (see section Sequences and Series page 1710):

$$\log(1+x)\cong x-\frac{x^2}{2}\Rightarrow\log(1+f\cdot r)\cong(fr)-\frac{(f\cdot r)^2}{2} \tag{15.2.87}$$

Therefore:

$$\begin{aligned}\mathrm{E}\left(\log(1+f\cdot r)\right)\cong\mathrm{E}\left((f\cdot r)-\frac{(f\cdot r)^2}{2}\right)&=\frac{\mathrm{d}}{\mathrm{d}f}\mathrm{E}\left(f\cdot r\right)-\frac{\mathrm{d}}{\mathrm{d}f}\mathrm{E}\left(\frac{f^2r^2}{2}\right)\\&=\mathrm{E}(r)\frac{\mathrm{d}}{\mathrm{d}f}f-\mathrm{E}\left(r^2\right)\frac{\mathrm{d}}{\mathrm{d}f}\frac{f^2}{2}=\mathrm{E}(r)-\mathrm{E}\left(r^2\right)f\end{aligned} \tag{15.2.88}$$

We equate to zero:

$$\frac{\mathrm{d}}{\mathrm{d}f}\mathrm{E}\left(\log(1+fr)\right)\cong\mathrm{E}(r)-\mathrm{E}\left(r^2\right)f=0 \tag{15.2.89}$$

Therefore:

$$\mathrm{E}(r)\cong\mathrm{E}\left(r^2\right)f \tag{15.2.90}$$

By rearranging:

$$f^{*}\cong\frac{\mathrm{E}(r)}{\mathrm{E}\left(r^2\right)} \tag{15.2.91}$$

But as $\mathrm{E}(r)=\mu_r$ and using Huygens relation (see section Statistics page 593) $\mathrm{E}\left(r^2\right)=\sigma_r^2+\mu_r^2$, we finally have:

$$f^{*}\cong\frac{\mu_r}{\sigma_r^2+{\mu_r}^2}\cong\frac{\mu_r}{\sigma_r^2}\qquad\text{if }\mu_r\ll\sigma_r \tag{15.2.92}$$

The average return stream μ_r is the superposition of the (average) rate of the risky bet (i.e. asset) and the average return of the non-risky bet (i.e. asset). We have to focus on the risky part! Therefore, denoting as always by r_f the return without risk and r_A the risky asset return such that $\mu_r=r_A+r_f$, we have then to replace μ_r by μ_r-r_f such that:

$$\boxed{f^{*}\cong\frac{\mu_r-r_f}{\sigma_r^2+(\mu_r-r_f)^2}\cong\frac{\mu_r-r_f}{\sigma_r^2}}\qquad\text{if }\mu_r\ll\sigma_r \tag{15.2.93}$$

We don't need to correct the variance and the denominator by including the variance of the free risk rate return as by definition that latter is supposed to be zero! For single assets (stock, index fund, etc.), and a risk-free rate r, the optimal fraction to invest through geometric Brownian motion $\mathrm{d}S_t/S_t = \mu \mathrm{d}t + \sigma \mathrm{d}W_t$ lead to exactly the same result as above!

That is sometimes rewritten using the Sharpe ratio (see page 7723) as following:

$$f^* \cong \frac{S_P}{\sigma_r} \tag{15.2.94}$$

Remark

There are other ways to get the same result but some well known other derivations have some strange counter-intuitive assumptions that we wanted to avoid here.

This is how $\mathrm{E}\left(\log(G)\right)$ and $\mathrm{E}(G)$ vary with f :

$$\begin{aligned}
\mathrm{E}\left(\log(G)\right) &= t\mathrm{E}\left(\log(1+f\cdot r)\right) \cong t\mathrm{E}\left((f\cdot r)-\frac{(f\cdot r)^2}{2}\right) = t\left(f\mathrm{E}(r)-\frac{f^2}{2}\mathrm{E}\left(r^2\right)\right)\\
&= t\left(f\cdot\mu_r - f^2\frac{(\sigma_r^2+\mu_r^2)}{2}\right) \cong t\left(f\cdot\mu_r-\frac{1}{2}f^2\sigma_r^2\right) \text{ if } \mu_r \ll \sigma_r \\
\mathrm{E}(G) &= \mathrm{E}\left(e^{\log(G)}\right) = e^{\mathrm{E}(\log(G))} \cong e^t\left(f\cdot\mu_r - f^2\frac{\sigma_r^2}{2}\right)
\end{aligned} \tag{15.2.95}$$

This is the max value of $\mathrm{E}[\log G]$ and $\mathrm{E}[G]$ when $f = f^*$:

$$\begin{aligned}
\max\mathrm{E}\left(\log(G)\right) &= t\left(\frac{\mu_r}{\sigma_r^2+\mu_r^2}\mu_r - \left(\frac{\mu_r}{\sigma_r^2+\mu_r^2}\right)^2\frac{(\sigma_r^2+\mu_r^2)}{2}\right) = t\left(\frac{\mu_r^2}{\sigma_r^2+\mu_r^2}-\frac{1}{2}\frac{\mu_r^2}{\sigma_r^2+\mu_r^2}\right)\\
&= \frac{t}{2}\frac{\mu_r^2}{\sigma_r^2+\mu_r^2} \cong \frac{t}{2}\frac{\mu_r^2}{\sigma_r^2} \text{ if } \mu_r \ll \sigma_r \\
\max\mathrm{E}(G) &= \max e^{\mathrm{E}(\log(G))} = e^{\max\mathrm{E}(\log(G))} \cong e^{\frac{t\mu_r^2}{2\sigma_r^2}}
\end{aligned} \tag{15.2.96}$$

Moreover, no money management system is perfect. This system will help you to diversify your portfolio efficiently, but there are many things that it can't do. It cannot pick winning stocks for you or predict sudden market crashes (although it can lighten the blow). There is always a certain amount of "luck" or randomness in the markets which can alter your returns.

15.2.5 Evolutionary Games

The strategies of biological evolution, as we have mentioned it earlier in this section, can be modelled using game theory. In this context, the biologist is bring to define remarkable relations defining a given development strategy (dominance, stagnation, suicide or other). Evolutionary Games originated in 11973 (holocene calendar) with John Maynard Smith and George R. Price's formalisation of contests, analysed as strategies, and the mathematical criteria that can be used to predict the results of competing strategies (hence 23 years after the publication by Nash of its famous paper just to put into context...).

Definition (#15.2.648): An "evolutionary steady strategy" (ESS) is a strategy adopted by the majority and preventing a population is invaded by a mutant that would use a different strategy.

This strategy is written in the form of a stable condition such that given two strategies S_1, S_2 of two players we have (explanations are given just after):

$$\boxed{U(S_1,S_1)>U(S_2,S_1)} \tag{15.2.97}$$

or (if it does not occur) by the simultaneity of the two strategies of non-selection and suicide:

$$\boxed{U(S_1,S_1)=U(S_2,S_1) \qquad \text{and} \qquad U(S_1,S_2)=U(S_2,S_2)} \tag{15.2.98}$$

1. The first relation means that in no case an individual has to change from strategy to defend against a mutant evolution with the same strategy than him, because any other strategy would be unfavourable.

2. The second relation means that whatever the strategy adopted against a mutant strategy, there will be stagnation.

3. The third relation means that against S_2 any strategy different of S_2 is preferable to counteract S_2 itself. In other words, apply a strategy S_2 different from S_1 is suicidal (the opposite is therefore not!).

To introduce such decision situation let use again a long companion example: the Hawk (bad) against Doves (good) game.

The H/D game has for purpose to model the relations between individuals competing for a rare resource (food, water, area, companion, contract, etc.) which degree of adaptation will be changed both by obtaining this resource and the violence they suffer or inflict to them to get it.

In their competitive interactions, biological organisms use two types of strategies (behaviours): the strategy of the hawk or that of the dove (as you can notice the "neutral" one is not present as it a suicide strategy).

The falcon intensify the conflict until he is injured or until the other run away. The dove withdrew after a first fight if the opponent chooses to escalate the violence of the conflict. When two hawks meet, one is injured and the other carries the resource. If a hawk and a dove fight, the hawk seized the safe resource to be injured and the dove gets no advantage or damage (at least on the short term and if there is not implicit trap). Finally, two doves share equally the resource.

We also put the following hypothesis (assumptions):

H1. Fight take place one by one

H2. The population can be considerate as infinite

H3. The encounters are random

H4. The battles are balanced (in the sense that neither age nor size nor experience influence the outcome of the battle)

H5. It is impossible to know before the start of a conflict which strategy will be adopted by one of the player

Based on these rules of interaction (where some are quite far from reality ...), it is possible to construct the tabular form of the game that allow us to calculate the advantages or disadvantages of the various strategies depending on the circumstances.

Therefore the tabular form of the game is (explanations of the variables are given after):

J_2 / J_1	H **(Hawk)**	D **(Dove)**
H **(Hawk)**	$\dfrac{V-C}{2},\dfrac{V-C}{2}$	$V,0$
D **(Dove)**	$0,V$	$\dfrac{V}{2},\dfrac{V}{2}$

Table 15.27: Hawk vs Dove evolutionary game matrix

As we can see the above game is a zero sum game, thus we can simplify the above table:

J_2 / J_1	H **(Hawk)**	D **(Dove)**
H **(Hawk)**	$\dfrac{V-C}{2}$	V
D **(Dove)**	0	$\dfrac{V}{2}$

Table 15.28: Simplified Hawk vs Dove evolutionary game matrix

We denote here by V the advantage that a player withdraws from obtaining the resource. V does not designate the resource itself, but the increased degree of adaptation that it provides to the player that gets it. C represents the cost paid, endangerment or injury, to acquire the resource.

First let us make explicit how this table should be read technically:

1. For the strategy (D,D) - everyone is nice to everyone - the total gain of the two individuals is:

$$\frac{V}{2}+\frac{V}{2}=V \tag{15.2.99}$$

   The population will therefore remain constant (it's the stagnation). In short, by their behaviour doves amicably share the resource value.

2. For the strategies (H,D), (D,H) the "doves" D always lose (they do not progress in their evolution). Their gain is zero while the "hawks" have eliminated a number V of doves (hence the gain).

3. For the (H,H) strategy the "hawks" support a loss $(V-C)/2$ where C is a constant and such that the sum of gains of the hawks is normally less than or equal to V. In other words:

$$\frac{V-C}{2}+\frac{V-C}{2}=V-C\leq V \tag{15.2.100}$$

In short, when a hawk fight another hawk, he gets on average 50% of the time the value of the resource reduced of the cost price to get it.

In summary to simplify:

1. Hawks (aggressive):
   - Always attack other individuals, taking the resource if the win
   - If a hawk encounters another hawk, it will win only half the time
   - When it loses, it will suffer and injury cost
2. Doves (non-aggressive):
   - When a dove encounters an opponent, it may put on an aggressive display, but it does not fight
   - When a dove encounters a dove, it will win the resource one half the time
   - When a dove encounters a hawk, the hawk will always win the resource
   - But since they don't fight, they don't incur any injury cost

Or in the form of a table:

J_1 \ J_2	H **(Hawk)**	D **(Dove)**
H **(Hawk)**	Hawk wins 50% of fights; is injured in 50% of fights.	Hawk always wins; dove flees.
D **(Dove)**	Dove never wins; is never injured.	Dove wins 50% of fights; is never injured.

Table 15.29: **Hawk vs Dove evolutionary game matrix explicit example**

Remark

This game can be seen as a war game between two players (or ideologies)... the interpretation of this results is therefore more than relevant. It is also an excellent comparison example between the competition that may exist between government offices (dowes) and private offices (hawks) providing the same type of product/service.

We must now consider two strategies:

1. The study of the game using a pure strategic game (without probabilities)
2. The study of the game using a mixed strategy (involving probabilities)

15.2.5.1 Dove & Hawk game in pure strategy (without probabilities)

For this type of strategy we will consider three cases:

$$V<C \qquad V=C \qquad V>C \tag{15.2.101}$$

1. If $V>C$ (higher benefit than the cost of the strategy), choosing (H,H) (which is therefore the strict Nash equilibrium of the game) we can see that the game will be a stable evolutionary strategy (SES). Indeed, we fall back on the relation:

   $$\left[U(H,H)=\frac{V-C}{2}\right]>[U(D,H)=0] \tag{15.2.102}$$

   corresponding well to:

   $$U(S_1,S_2)>U(S_2,S_1) \tag{15.2.103}$$

   and we can therefore also observe that there exist is also a weakly dominant strategy (no natural selection) in:

   $$[U(H,D)=V]>\left[U(D,D)=\frac{V}{2}\right] \tag{15.2.104}$$

   corresponding well to:

   $$U(S_1,S_2)>U(S_2,S_2) \tag{15.2.105}$$

   But that latter will not be adopted as weaker than the Nash equilibrium.

2. If $V=C$ (benefit equal to the cost), the game is also of the SES kind. Indeed, (H,H) becomes a weakly dominant strategy:

   $$[U(H,H)=0]=[U(D,H)=0] \tag{15.2.106}$$

   corresponding well to:

   $$U(S_1,S_1)=U(S_2,S_1) \tag{15.2.107}$$

   and there is therefore no evolution and we can also observe that there is also:

   $$[U(H,D)=V]=\left[U(D,D)=\frac{V}{2}\right] \tag{15.2.108}$$

   corresponding to:

   $$U(S_1,S_2)>U(S_2,S_2) \tag{15.2.109}$$

   Since we have simultaneously:

   $$U(S_1,S_1)=U(S_2,S_1) \qquad \text{and} \qquad U(S_1,S_2)>U(S_2,S_2) \tag{15.2.110}$$

   when $V=C$ the game is a SES.

3. If $V<C$ (benefit less than the cost), nor H, nor D are the dominant strategies and we do not have a SES:

$$[U(D,H)=0]>\left[U(H,H)=\frac{V-C}{2}\right] \tag{15.2.111}$$

and:

$$[U(H,D)=V]>\left[U(D,D)=\frac{V}{2}\right] \tag{15.2.112}$$

the latter two relations corresponding both to:

$$U(S_1,S_2)>U(S_2,S_2) \tag{15.2.113}$$

It is rather annoying... it's a kind of collective suicide...

Remark

The last two relations lead us to observe that the hawks will not want to necessarily reveal to others their strategy of doves predators, since: any strategy is better to be countered by another strategy rather than by itself. They may prefer to discuss among themselves which leads to the fact that the game therefore cooperative.

15.2.5.2 Dove & Hawk game in stable evolutionary strategy (with probabilities)

Let us now using the study in mixed strategy what we could do to bring the previous configuration to a SES (thus relatively to the last configuration $V<C$ to see more closely what we can do to avoid this) :

We consider a population of individuals who therefore play a mixed strategy which we denote for each:

$$\sigma(P,1-P) \tag{15.2.114}$$

with $P=P(H)$ and:

$$1-P=P(D) \tag{15.2.115}$$

If $P=1$ (pure strategy), we will therefore have:

$$\sigma(P,1-P)=\sigma(1,0)=\sigma_H \tag{15.2.116}$$

Let us now study the three cases:

$$V<C \qquad V=C \qquad V>C \tag{15.2.117}$$

- If $V>C$ with σ_H and $1>P\geq 0$ we always have:

  $$U(\sigma_H,\sigma_H)>U(\sigma,\sigma_H) \tag{15.2.118}$$

  in other words, the strategy will be always of the kind stable evolutionary (SES) if σ_H is a pure strategy and this even if σ can vary and tend to σ_H.

- If $V=C$ with σ_H and $1>P\geq 0$, we always have:

  $$U(\sigma_H,\sigma_H)=U(\sigma,\sigma_H) \qquad \text{and} \qquad U(\sigma_H,\sigma)>U(\sigma,\sigma)=\frac{V}{2} \tag{15.2.119}$$

  in other words: the strategy will always be of the non-selective type if σ_H is a pure strategy even if σ can vary and tend to σ_H.

- If $V<C$ and $1>P\geq 0$, we drop σ_H to focus only to the generalization $\sigma(P,1-P)$. Then we have:

  $$U(\sigma,D)>U(D,D) \tag{15.2.120}$$

  Indeed:

  $$\left[U(\sigma,D)=PV+(1-P)\frac{V}{2}=(1+P)\frac{V}{2}\right]>\left[U(D,D)=\frac{V}{2}\right] \tag{15.2.121}$$

  Also:

  $$U(\sigma,H)>U(H,H) \tag{15.2.122}$$

  Indeed:

  $$\left[U(\sigma,H)=P\frac{V-C}{2}+(1-P)\cdot 0=P\frac{V-C}{2}\right]>\left[U(H,H)=\frac{V-C}{2}\right] \tag{15.2.123}$$

  and we would to arrive at a SES with a mixed strategy. Is this possible?

  As part of a mixed strategy, we have proved during our study of a zero-sum game that mixed equilibrium was given by:

  $$\mathrm{E}(J_1)=-\mathrm{E}(J_2) \tag{15.2.124}$$

  It is therefore quite obvious that for a game that is not a zero-sum game, we have the mixed equilibrium given by:

  $$\mathrm{E}(J_1)=\mathrm{E}(J_2) \tag{15.2.125}$$

  Therefore we seek the relation between P, C and V such that this equilibrium is reached. So in cascade we have:

  $$\mathrm{E}(J_1)=\mathrm{E}(J_2) \tag{15.2.126}$$

  that can be written:

  $$\mathrm{E}(H)=\mathrm{E}(D) \tag{15.2.127}$$

  Which lead us to write:

  $$U(H,\sigma)=U(D,\sigma)\Leftrightarrow U(S_1,S_1)=U(S_2,S_1) \tag{15.2.128}$$

  Therefore:

  $$P\cdot U(H,H)+(1-P)\cdot U(H,D)=P\cdot U(D,H)+(1-P)\cdot U(D,D) \tag{15.2.129}$$

Knowing the utilities:

$$P\cdot\frac{V-C}{2}+(1-P)\cdot V=P\cdot 0+(1-P)\cdot\frac{V}{2} \tag{15.2.130}$$

from which we derive that the equilibrium in mixed strategy is given by:

$$P=\frac{V}{C} \tag{15.2.131}$$

and therefore that the equilibrium is given by the mixed strategy:

$$\sigma=\left(\frac{V}{C},1-\frac{V}{C}\right) \tag{15.2.132}$$

To what will this strategy lead to in practice? Well, simply in the suicidal case this mixed strategy is the best answer against itself (this is what is best done in what is worst) because it leads to the two conditions that satisfy a SES.

Remark

We strongly recommend the reader to take a look to the paper *Social penalty promotes cooperation in a cooperative society* (see [279]) that as the title says provides evidence in the case of the evolutionary Hawk-Dove game that introducing a social penalty (and even punishments) on the aggressive individual is likely to promote cooperation and never reduce cooperation. The reader can also take a look to the paper *Culture and the evolution of human cooperation* (see [71]) that explains how cultural evolution, social learning, punishment, and group norms support large-scale cooperation.

15.2.6 Cournot Competition

Cournot competition (11838 according to holocene calendar) is an economic model used to describe an industry structure in which companies compete on the amount of output they will produce, which they decide on independently of each other and at the same time.

Let us imagine two companies M and N, and two sources whose qualities are identical and that are positioned to concurrently supply the same market so that the total quantity delivered to shops consists of the sum of the quantities m, n delivered by each company at a price that is necessarily the same for each of them as there is no reason to prefer one source to another. This price is determined when the sum of the quantities m, n is completely known, because of the link between offer and demand. Let us suppose the company N has set arbitrarily, without regard to the price, quantity he n he intends to deliver, then the company M will indirectly fix the sale price on the market, that is to say the total available production (consisting of the sum of the quantities m and n for recall...), that is to say, its own production quantity m to obtain the greatest possible income.

In practice, a series of trial and error will bring the two companies in their equilibrium position, and the theory shows that this equilibrium is stable: that is to say that if one or the other owners depart momentarily from this equilibrium, he will be taken back by a series of oscillations of the kind that had originally led to constitute the balance.

Definition (#15.2.649): A "Cournot equilibrium" is therefore related to the fact that when a firm raises price,the others will stay put. If a firm reduces price,the other firms will do the same.

> Remark
>
> The Nash equilibrium is the generalization of the Cournot equilibrium. A Nash equilibrium can be interpreted as a pair of expectations about each player's choice such that, when the other person's choice is revealed, neither individual wants to change his behaviour.

OK! Let us now focus on the maths part! But for this let us first gives the hypothesis (assumptions) of the mode:

H1. There is more than one firm and all firms produce a homogeneous product, i.e. there is no product differentiation;

H2. Firms do not cooperate, i.e. there is no collusion;

H3. Firms have market power, i.e. each firm's output decision affects the good's price; The number of firms is fixed;

H4. Firms compete in quantities, and choose quantities simultaneously;

H5. The firms are economically rational and act strategically, usually seeking to maximize profit given their competitors' decisions.

We will put ourselves in the situation of a game with two plays. We will put that the price per unit P of the product is a linear function of the total quantity produced Q such that:

$$P(Q)=P(m+n)=\alpha(m+n) \tag{15.2.133}$$

where α a normalization constant of units: [price · piece^{-2}]. As it represents a price per unit divided be a global quantity.

Remark

We will see that this choice we will lead to absurd result as in reality the price per unit decrease in function of the quantity and don't increase.

We assume the marginal production costs equal and fixed and represents by the variable C, and the fixed costs as zero such that the production cost will be written respectively for the both companies:

$$C_M=C_N=C \tag{15.2.134}$$

To recognize a game in normal form, it remains for us to identify the benefit of each opponent for any pair (m,n) of strategies such that, if desired, to build the matrix of gains.

Let us denote π_i the margin of competitor i. Then we have respectively:

$$\begin{gathered}\pi_M(m,n)=mP(m+n)-mC=m\left[\alpha(m+n)-C\right]\\ \pi_N(m,n)=nP(m+n)-nC=n\left[\alpha(m+n)-C\right]\end{gathered} \tag{15.2.135}$$

The search for a Nash equilibrium leads each company to choose a production quantity that maximize profits and minimize its storage costs (see the Wilson's model in the section of Quantitative Management Techniques at page 8119), the production of his concurrent being assumed to be known.

For this purpose, we cancels the derivative of the previous two functions:

$$\begin{gathered}\frac{\partial\pi_M}{\partial m}=2\alpha m+\alpha n-C=0\\ \frac{\partial\pi_N}{\partial n}=\alpha m+2\alpha n=0\end{gathered} \tag{15.2.136}$$

System whose resolution brings leads us easily leads to the obvious solution that:

$$m_{\text{Nash}}=n_{\text{Nash}}=\frac{1}{3}\frac{C}{\alpha} \tag{15.2.137}$$

It remains to verify that these solutions are maximums, by controlling the second-order derivatives, and not minimums.

The equilibrium position of the duopoly intervenes therefore when each of the two companies produces the same quantity... (not surprising result). The is then the Nash equilibrium production quantity.

The sale price unit under a Nash equilibrium would be:

$$\boxed{P_{\text{Nash}}=\alpha(m+n)=\frac{2}{3}C} \tag{15.2.138}$$

Either a value below the marginal cost C which obviously sounds not good as the margin is then negative. Indeed: Therefore the profit is given by:

$$\begin{aligned}\max(\pi_{M,\text{Nash}})=m\left[\alpha(m+n)-C\right]=-m\frac{1}{3}C=-\frac{2}{9}\frac{C^2}{\alpha}\\ \max(\pi_{N,\text{Nash}})=n\left[\alpha(m+n)-C\right]=-n\frac{2}{3}C-mC=-\frac{2}{9}\frac{C^2}{\alpha}\end{aligned} \tag{15.2.139}$$

So it would be better that the price is not an affine function of the quantity...

These calculations are to be compared with the purely economic reasoning, for which each company would like to be alone on the market (monopoly). The profit of the company M in monopoly position would then be in the assumptions above given by:

$$\pi_M(m,0)=m\left[\alpha m-C\right]=\alpha m^2-mC \tag{15.2.140}$$

$$m_{\text{monopoly}}=\frac{1}{2}\frac{C}{\alpha} \tag{15.2.141}$$

So we see obviously that:

$$\boxed{m_{\text{monopolay}}>m_{\text{Nash}}} \tag{15.2.142}$$

and therefore that the quantity produced in case of monopoly is greater in case of duopoly (logical!) and the profit and the prices can then be higher.

The sale price unit under a monopoly would then be:

$$\boxed{P_{\text{monopoly}}=\alpha(m+n)=\alpha m=\frac{1}{2}C} \tag{15.2.143}$$

and therefore:

$$\max(\pi_{M,\text{monompoly}})=m\left[\alpha(m+n)-C\right]=-m\frac{1}{2}C=-\frac{1}{4}\frac{C^2}{\alpha} \tag{15.2.144}$$

Therefore the benefit is still negative with this affine model (...) and even worst... the benefit in case of monopoly is small that in case of duopole...

The idea would be now, if we go back to our two companies, that they establish and agreement (case named "oligopoly agreement" against the competition ... which is forbidden by the commercial laws in most countries!), that makes them share a potential majorited profit. The perfect symmetry of the situations would lead to the division by half of the market. But the difficulty is that the decision to produce:

$$m_{\text{agreement}}=n_{\text{agreement}}=\frac{1}{2}\left(\frac{1}{2}\frac{C}{\alpha}\right) \tag{15.2.145}$$

is not the best answer because it encourages to betray the agreement with the other. Thus, the best balance is that of imposing Nash equilibrium already seen above.

During the development of an agreement or a cartel, one can distinguish several levels of quantities that depend on the degree of precision of the rules defined by the companies belonging to this cartel.

The first case is that we can name a "perfect agreement"; This is the agreement that maximizes the total profit of all the companies at the same time.

The maximization of the total margin (profit) a set of companies will obviously be given by:

$$\boxed{\pi_{\text{cartel}}(q_1,q_2,\ldots,q_n)=QP(Q)-\sum_{i=1}^{n}C_iq_i} \tag{15.2.146}$$

where for recall we have the conditions:

$$C_i\geq 0 \qquad q_i\geq 0 \qquad Q=\sum_{i=1}^{n}q_i \tag{15.2.147}$$

This profit is maximized when obviously all the partial derivatives of order 1 are zero and maximum, that is:

$$\dfrac{\partial\pi_{\text{cartel}}}{\partial q_i}=0\Leftrightarrow\dfrac{\partial\left(\displaystyle\sum_{i=1}^{n}q_iP\left(\sum_{i=1}^{n}q_i\right)\right)}{\partial q_i}-\dfrac{\partial\left(C_iq_i\right)}{\partial q_i}=0 \tag{15.2.148}$$

Where the term:

$$\dfrac{\partial\left(\displaystyle\sum_{i=1}^{n}q_iP\left(\sum_{i=1}^{n}q_i\right)\right)}{\partial q_i} \tag{15.2.149}$$

expresses the change in total revenue caused by a small change in the quantity q_i produced by the manufacturer i, and:

$$\dfrac{\partial\left(C_iq_i\right)}{\partial q_i} \tag{15.2.150}$$

expresses the change in cost caused by the same variation of q_i (producer i marginal cost).

The marginal revenue caused by a given change in output Q is the same, regardless of the manufacturer which changed its production. Indeed, the influence of an additional production on the total supply and the price is the same, that this additional production comes from a producer or another one.

15.2.7 Markov Decision Processes (MDPs)

As we mentioned in the section Probability, Markov chains are also used in the field of decision making techniques. We speak then of "Markov decision processes (MDPs)" (denomination that can be found in the ISO 31010 Risk Management standard) that provide a mathematical framework for modelling decision making in situations where outcomes are partly random and partly under the control of a decision maker (MDPs are known at least as early as the 11950 according to holocene calendar). They are used in a wide area of disciplines, including robotics, automated control, economics, and manufacturing.

More precisely, a Markov Decision Process is a discrete time stochastic control process. At each time step, the process is in some state s, and the hazard may choose any action a a that is available in state s. The process responds at the next time step by randomly moving into a new state s'. The finally issue led to the possibility to help to take better decisions.

Remark

Strictly speaking, Markov decision processes are an extension of Markov chains; the difference is the addition of actions (allowing choice) and rewards (giving motivation). Conversely, if only one action exists for each state (e.g. "wait") and all rewards are the same (e.g. "zero"), a Markov decision process reduces to a Markov chain.

For example, the following simple Markov chain (state diagram) on a particular medical symptom, has a step say to "absorbing state" known by all that it is not possible to survive to this day...:

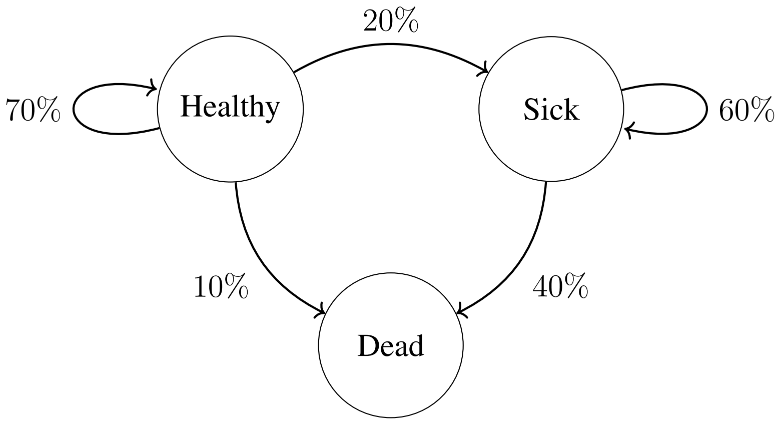
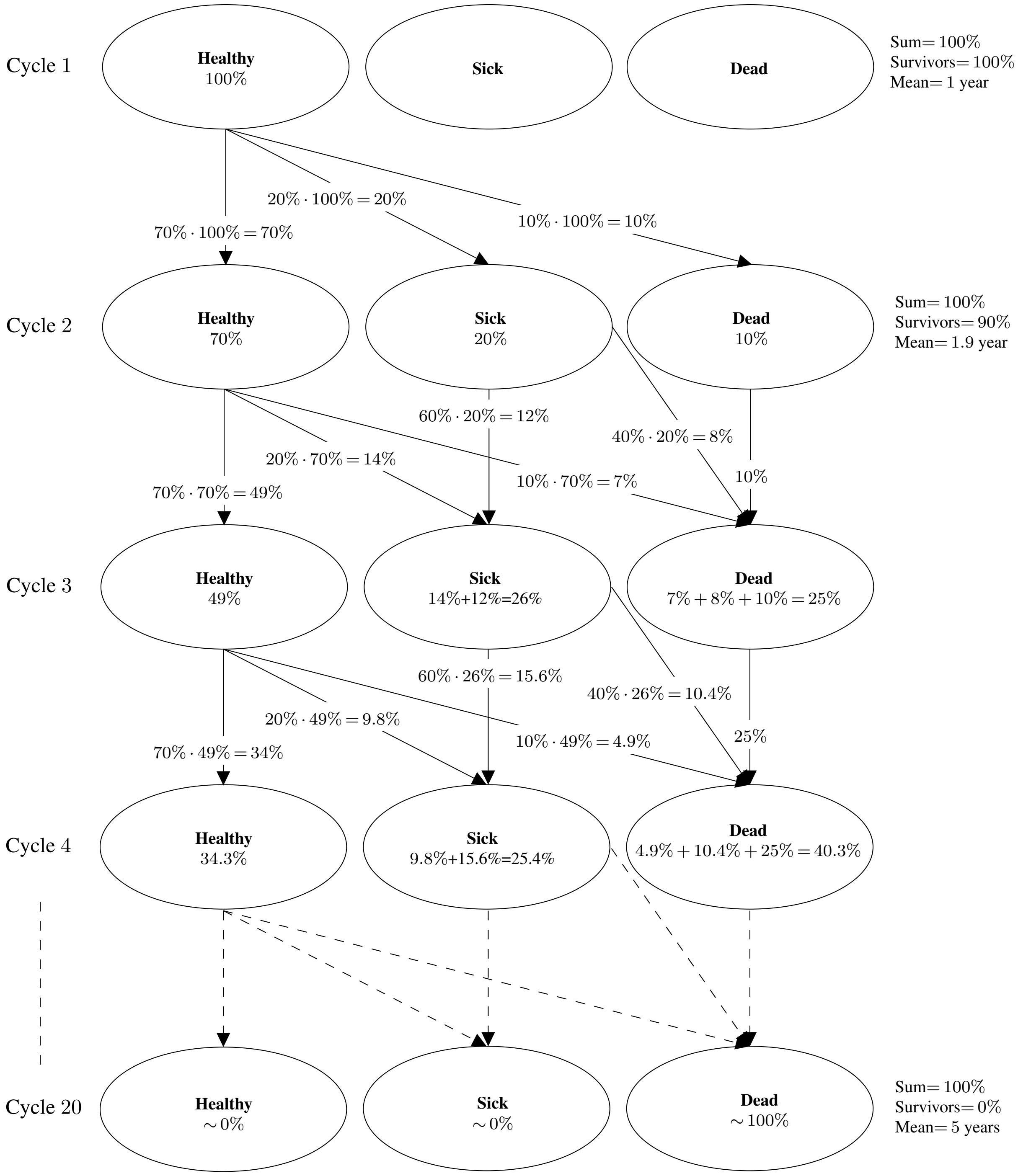


Figure 15.70: **Medical state diagram**

And here is the same Markov chain decomposed over 20 cycles as traditionally presented in the medical field (this assumes that the probabilities do not change over time...):


Cycle 1
Healthy
100%
Sick
Dead
Sum= 100%
Survivors= 100%
Mean= 1 year
20% · 100% = 20%
10% · 100% = 10%
70% · 100% = 70%
Cycle 2
Healthy
70%
Sick
20%
Dead
10%
Sum= 100%
Survivors= 90%
Mean= 1.9 year
60% · 20% = 12%
40% · 20% = 8%
20% · 70% = 14%
10%
10% · 70% = 7%
70% · 70% = 49%
Cycle 3
Healthy
49%
Sick
14%+12%=26%
Dead
7% + 8% + 10% = 25%
60% · 26% = 15.6%
40% · 26% = 10.4%
20% · 49% = 9.8%
25%
10% · 49% = 4.9%
70% · 49% = 34%
Cycle 4
Healthy
34.3%
Sick
9.8%+15.6%=25.4%
Dead
4.9% + 10.4% + 25% = 40.3%
Cycle 20
Healthy
∼ 0%
Sick
∼ 0%
Dead
∼ 100%
Sum= 100%
Survivors= 0%
Mean= 5 years

We can also consider the following example applied to finance:

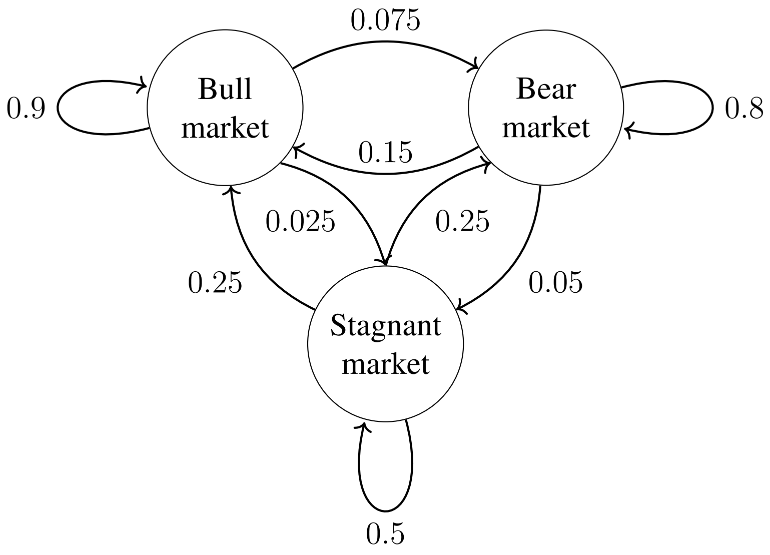


Figure 15.71: Financial state diagram (source: Wikipedia, author: Gareth Jones)

where a bull week[11] is followed by another bull week 90% of the time, a bear week 7.5% of the time, and a stagnant week the other 2.5% of the time. Labelling the state space 1 = bull, 2 = bear, 3 = stagnant the transition matrix (see section Probabilities page 538) for this example is:

$$P=\begin{bmatrix}0.9 & 0.075 & 0.025\\ 0.15 & 0.8 & 0.05\\ 0.25 & 0.25 & 0.5\end{bmatrix} \tag{15.2.151}$$

The distribution over states can be written as a stochastic row vector p with the relation:

$$x(n+1)=x(n)P \tag{15.2.152}$$

So if at time n the system is in state $p(n)$, then three time periods later, at time $p+3$ the distribution is:

$$\begin{aligned}p^{(n+3)}&=p^{(n+2)}P=\left(x^{(n+1)}P\right)\,P\\ &=p^{(n+1)}P^2=\left(p^{(n)}P\right)\,P^2\\ &=p^{(n)}P^3\end{aligned} \tag{15.2.153}$$

In particular, if at time n the system is in state 2 (bear), then at time $n+3$ the distribution is:

$$\begin{aligned}p^{(n+3)}&=\begin{bmatrix}0 & 1 & 0\end{bmatrix}\begin{bmatrix}0.9 & 0.075 & 0.025\\ 0.15 & 0.8 & 0.05\\ 0.25 & 0.25 & 0.5\end{bmatrix}^3\\ &=\begin{bmatrix}0 & 1 & 0\end{bmatrix}\begin{bmatrix}0.7745 & 0.17875 & 0.04675\\ 0.3575 & 0.56825 & 0.07425\\ 0.4675 & 0.37125 & 0.16125\end{bmatrix}\\ &=\begin{bmatrix}0.3575 & 0.56825 & 0.07425\end{bmatrix}\end{aligned} \tag{15.2.154}$$

---

[11] "Bearish" and "Bullish" are simply terms used to characterize trends in the currency, commodity or stock markets. If prices tend to be moving upward, it is a bull market. If prices are moving downward, it is a bear market.

Using the transition matrix it is possible to calculate, for example, the long-term fraction of weeks during which the market is stagnant, or the average number of weeks it will take to go from a stagnant to a bull market. Using the transition probabilities, the steady-state probabilities indicate that 62.5% of weeks will be in a bull market, 31.25% of weeks will be in a bear market and 6.25% of weeks will be stagnant, since:

$$\lim_{N\to+\infty} P^N = \begin{bmatrix} 0.625 & 0.3125 & 0.0625 \\ 0.625 & 0.3125 & 0.0625 \\ 0.625 & 0.3125 & 0.0625 \end{bmatrix} \qquad (15.2.155)$$

Obviously it is possible to perform the same calculation directly in matrix form... again the transition matrix (stochastic matrix) is easy to identify. It is (see section Probabilities page 538):

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix} = \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0 & 0.6 & 0.4 \\ 0 & 0 & 1 \end{bmatrix} \qquad (15.2.156)$$

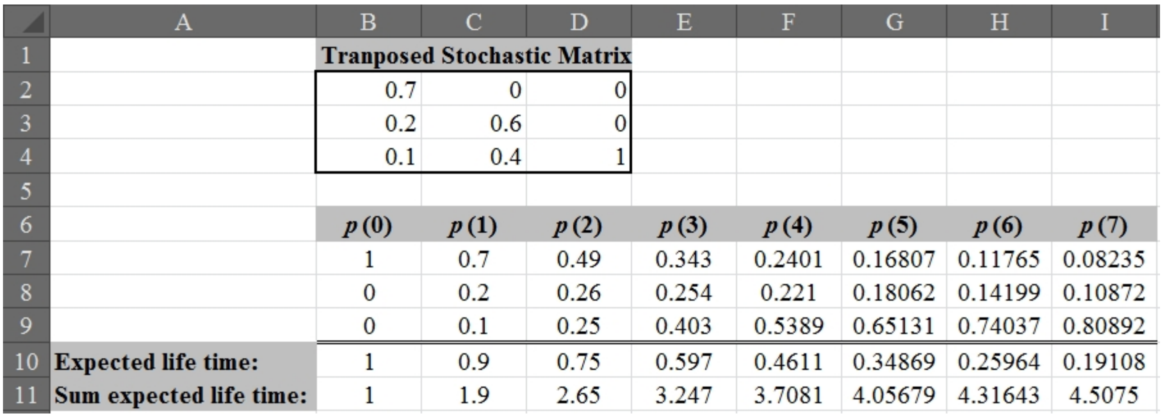
The initial probability vector $p(0)$ is obviously in the most common cases of disease ...:

$$p(0) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \qquad (15.2.157)$$

Every time we multiply the transpose of the transition matrix by the vector of initial probabilities, we then get the probability of being in a given state at a given cycle !:

$$p(n+1) = P^T p(n) \qquad (15.2.158)$$

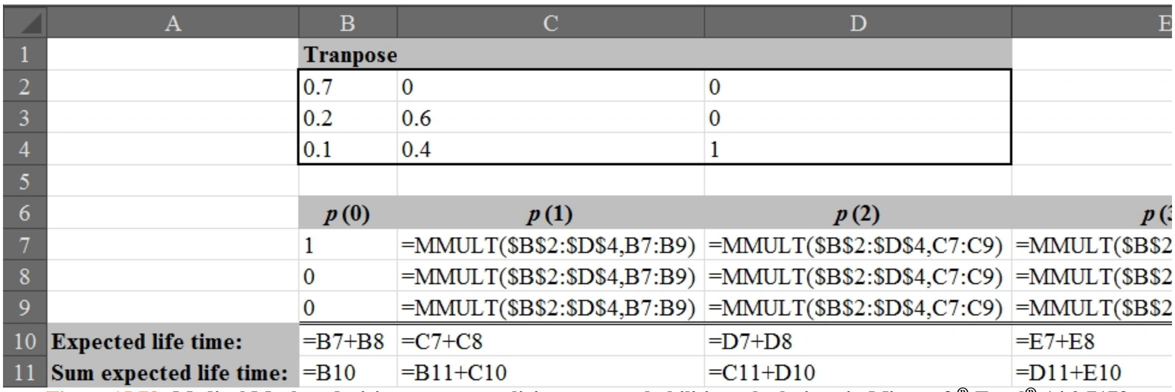
With the Microsoft® Excel® 14.0.7173 the modelling is quite simple to reproduce:



	A	B	C	D	E	F	G	H	I
1		Tranposed Stochastic Matrix							
2		0.7	0	0					
3		0.2	0.6	0					
4		0.1	0.4	1					
5									
6		*p* (0)	*p* (1)	*p* (2)	*p* (3)	*p* (4)	*p* (5)	*p* (6)	*p* (7)
7		1	0.7	0.49	0.343	0.2401	0.16807	0.11765	0.08235
8		0	0.2	0.26	0.254	0.221	0.18062	0.14199	0.10872
9		0	0.1	0.25	0.403	0.5389	0.65131	0.74037	0.80892
10	**Expected life time:**	1	0.9	0.75	0.597	0.4611	0.34869	0.25964	0.19108
11	**Sum expected life time:**	1	1.9	2.65	3.247	3.7081	4.05679	4.31643	4.5075

Figure 15.72: **Medical Markov decision process example of vector probabilities calculations in Microsoft® Excel® 14.0.7173**

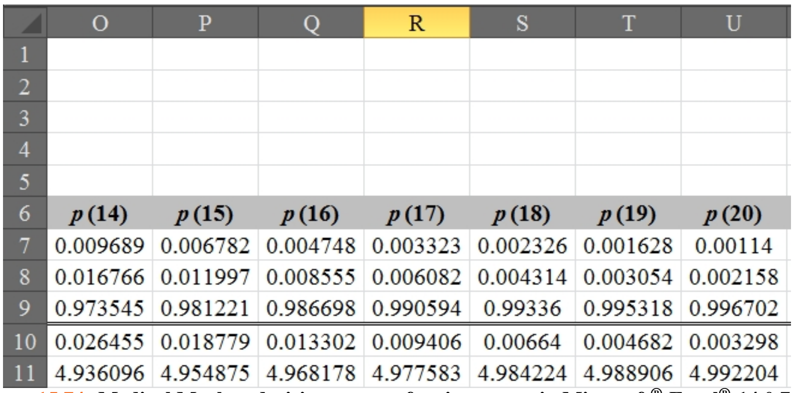
Explicitly for the first probability vectors:



	A	B	C	D	E
1		**Tranpose**			
2		0.7	0	0	
3		0.2	0.6	0	
4		0.1	0.4	1	
5					
6		**p (0)**	**p (1)**	**p (2)**	**p (**
7		1	=MMULT($B$2:$D$4,B7:B9)	=MMULT($B$2:$D$4,C7:C9)	=MMULT($B$2
8		0	=MMULT($B$2:$D$4,B7:B9)	=MMULT($B$2:$D$4,C7:C9)	=MMULT($B$2
9		0	=MMULT($B$2:$D$4,B7:B9)	=MMULT($B$2:$D$4,C7:C9)	=MMULT($B$2
10	**Expected life time:**	=B7+B8	=C7+C8	=D7+D8	=E7+E8
11	**Sum expected life time:**	=B10	=B11+C10	=C11+D10	=D11+E10

Figure 15.73: **Medical Markov decision process explicit vector probabilities calculations in Microsoft® Excel® 14.0.7173**

If we continue up to the 20th cycle:



	O	P	Q	R	S	T	U
1							
2							
3							
4							
5							
6	**p (14)**	**p (15)**	**p (16)**	**p (17)**	**p (18)**	**p (19)**	**p (20)**
7	0.009689	0.006782	0.004748	0.003323	0.002326	0.001628	0.00114
8	0.016766	0.011997	0.008555	0.006082	0.004314	0.003054	0.002158
9	0.973545	0.981221	0.986698	0.990594	0.99336	0.995318	0.996702
10	0.026455	0.018779	0.013302	0.009406	0.00664	0.004682	0.003298
11	4.936096	4.954875	4.968178	4.977583	4.984224	4.988906	4.992204

Figure 15.74: **Medical Markov decision process far time states in Microsoft® Excel® 14.0.7173**

It is very common in companies to synthesize the evolution graphically (more meaningful to management...):

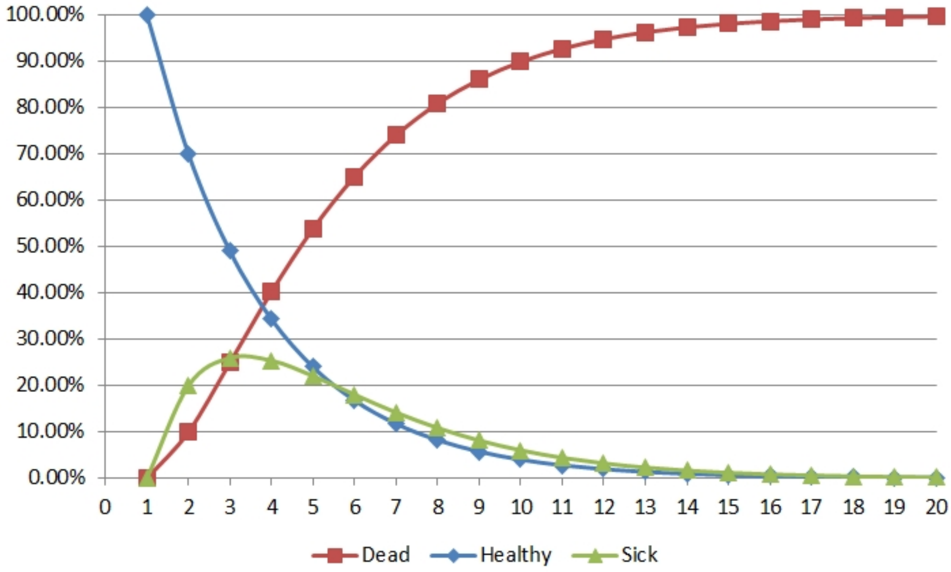


Figure 15.75: **Medical Markov decision process far time states in Microsoft® Excel® 14.0.7173**

So we see that the state "death" is an absorbing state because all probabilities converges to it (unfortunately ...). We also see that the total life expectancy if of 4.99 cycles. If we assimilate a cycle to one year, then life expectancy is 4.99 years. We immediately see that the stationary measure (see section Probabilities page 540) of the chain is:

$$\pi=\begin{pmatrix}0\\0\\1\end{pmatrix} \tag{15.2.159}$$

15.2.8 Multi-Criteria Decision Making (MCDM)

The field of "multi-criteria decision making" (MCDM) studies (among other topics) the problem of how to select the best alternative among a number of competing alternatives. This is an important task in decision making. In such a setting each alternative is described in terms of a set of evaluative criteria. These criteria are associated with weights of importance. Intuitively, one may think that the larger the weight for a criterion is, the more critical that criterion should be. However, this may not be the case. It is important to distinguish here the notion of "criticality" with that of "importance". By "critical", we mean that a criterion with small change (as a percentage) in its weight, may cause a significant change of the final solution. It is possible criteria with rather small weights of importance (i.e., ones that are not so important in that respect) to be much more critical in a given situation than ones with larger weights.

In our daily lives, we usually weigh multiple criteria implicitly and we may be comfortable with the consequences of such decisions that are made based on only intuition. On the other hand, when stakes are high, it is important to properly structure the problem and explicitly evaluate multiple criteria. In making the decision of whether to build a nuclear power plant or not, and where to build it, there are not only very complex issues involving multiple criteria, but there are also multiple parties who are deeply affected from the consequences.

The are dozens of MCDM models but we will here focus only on the most widely used one and also available in specialized decision softwares: the AHP.

MCDM has been an active area of research since the 11970s (holocene calendar). There are several MCDM-related organizations including the International Society on Multi-criteria Decision Making, Euro Working Group on MCDA and INFORMS Section on MCDM

15.2.8.1 Analytic Hierarchy Process (AHP)

The "analytic hierarchy process" (AHP) is a MCDM structured technique for organizing and analysing complex decisions, based on mathematics and psychology. It was developed by Thomas L. Saaty in the 11970s (holocene calendar) and has been extensively studied and refined since then. It is used almost exclusively in large US multinationals and in governments where politician know well the issue of human decision bias to make complex deterministic decisions.

In the large majority of books, AHP is wrongly introduced (also in Wikipedia still in 12015... according to holocene calendar). So let us see exactly what are the mathematical concepts behind (you can check and apply this model with a free software like `R` - see our companion book on this software that has more details).

Remark

In the AHP, each element in the hierarchy is considered to be independent of all the others at the opposite of the ANP (Analytic Network Process).

We know that traditional voting or score / rank systems can not be used in multi-criteria and multi-level complex decisions because they do not trivially take into account the relative importance of components. Moreover, even if a weight was affected the problem lies in the choice of an empirical value of that latter which makes the approach totally unscientific and opens the door to endless debates in practice.

To introduce the technique of the AHP it seems pertinent to study this technique with a companion example of naive analysis of a sum of scores:

Alternative / **Criteria**	**Choice X**	**Choice Y**	**Choice Z**	**Range**
Factor A	1	4	5	0 to +5
Factor B	20	70	50	+1 to +100
Factor C	-2	0	1	-2 to $+2$
Factor D	0.4	0.75	0.4	0 to +1
Sum	19.4	74.75	56.4	
Normalized score	12.9%	49.7%	37.5%	

Table 15.30: **Naive ranged scoring decision analysis approach**

However, the reader will notice that the range of values for each factor of the product / project is not the same. It is then quite unfair to summarize all the values of several criteria and compare the result. It is clear that factor B is dominant because the range has a higher value. To be fair, we can propose at least two trivial solutions:

- Instead of using arbitrary values for each factor, we position just the choice for each factor: the smaller rank value is more preferable than the higher rank value (this is psychology in relation to the daily use of The majority of humans).
- We transform the score value of each factor as a function of the range value so that each factor will have the same extent (normalization of the scores).

Let us try the two solutions, starting with the first:

Alternative / **Criteria**	**Choice X**	**Choice Y**	**Choice Z**
Factor A	3	2	1
Factor B	3	1	2
Factor C	3	2	1
Factor D	2	1	2
Sum	11	6	6
Normalized score	26.09%	36.96%	36.96%

Table 15.31: **Naive ranged scoring decision analysis approach**

Where since the smallest rank value is the best, we have renormalized the codes in the following logical way:

1. Since the sum of all ranks is always equal, subtracting from the total of the ranks R_T the sum of the ranks of a column automatically has the effect of giving it the inverse

importance (indeed if the total sum of the ranks is equal to 23 as above, then $23-6=17$ whereas $23-11=12$ so we have the desired effect). We then have the following sum of ranks (SR) for each criterion (C):

$$SR_{N_i}=R_T-R_{C_i} \tag{15.2.160}$$

2. We want a percentage, so it is natural to write:

$$\%SR_{N_i}=\frac{R_T-R_{C_i}}{R_T}=1-\frac{R_{C_i}}{R_T} \tag{15.2.161}$$

3. The problem of this last calculation is that it gives systematically, summed for all the columns, the value of 200%. It must then be multiplied by 0.5 such that in the end:

$$\%SR_{N_i}=0.5\left(1-\frac{R_{C_i}}{R_T}\right) \tag{15.2.162}$$

Now it is important to notice that working with the ranks no longer gives the same result as in the antecedent table. Indeed, the choices Y and Z become indifferent and so we lost information!

Reason why it is recommended to work on scores rather than ranks because not only working on the ranks can lead to equalities that did not need have to exist to be but in some cases it even reverses the final rank of some elements of a choice (we speak then of "inversion effect of rank by the scores").

Now let us see what happen if we transform the score value of each factor in such a way such that all factors have the same range value. Say, we choose all factors to have range to be 0 to 1. To convert linearly the score of each factor from the table below into the next one, we use the following formula which is based on simple geometric of a line segment:

$$\begin{aligned}\text{New Score}&=\frac{\text{New Upper Bound}-\text{New Lower Bound}}{\text{Original Upper Bound}-\text{Original Lower Bound}}(\text{Original Score}-\text{Original Lower Bound})+\text{New Lower Bound}\\&=\frac{\text{NUB}-\text{NLB}}{\text{OUB}-\text{OLB}}(\text{OS}-\text{OLB})+\text{NLB}\end{aligned} \tag{15.2.163}$$

Alternative **Criteria**	**Choice X**	**Choice Y**	**Choice Z**	**Range**
Factor A	1	4	5	0 to +5
Factor B	20	70	50	+1 to +100
Factor C	−2	0	1	−2 to +2
Factor D	0.4	0.75	0.4	0 to +1
Sum	19.4	74.75	56.4	
Normalized score	12.9%	49.7%	37.5%	

For instance, Factor A has originally range 0 to 5. To make score of choice Y from 4 into a range of 0 to 1 we have OLB $=0$, OUB $=5$, NLB $=0$, NUB $=5$ and Original Score $=5$ thus:

$$\text{New Score}=\frac{1-0}{5-0}(4-0)+0=\frac{4}{5}=0.8 \tag{15.2.164}$$

Another example, for choice X in factor B has original score of 20 and original range 1 to 100. Thus we have OLB $=1$, OUB $=100$, NLB $=0$, NUB $=1$ and Original Score $=20$ thus:

$$\text{New Score} = \frac{1-0}{100-1}(20-1)+0 = \frac{19}{99} = 0.192 \qquad (15.2.165)$$

Therefore becomes:

Criteria \ **Alternative**	**Choice X**	**Choice Y**	**Choice Z**
Factor A	0.109	0.436	0.545
Factor B	0.035	0.127	0.090
Factor C	0.000	0.091	0.136
Factor D	0.036	0.068	0.036

Clearly the transformation of score value is a little bit more complicated than rank but we get better results.

Now let us see a naturally more advanced version inspired by financial portfolios including weights to criteria. Obviously the result of the example that follows will no longer be comparable to the previous ones.

So consider the following table of weights:

	Factor A	**Factor B**	**Factor C**	**Factor D**	**Sum**
Weight	6	2	2	1	11
Relative Weight	54.5%	18.2%	18.2%	9.1%	100%

Our table of normalized scores therefore becomes:

Criteria \ **Alternative**	**Weight**	**Choice X**	**Choice Y**	**Choice Z**
Factor A	54.5%	0.109	0.436	0.545
Factor B	18.2%	0.035	0.127	0.090
Factor C	18.2%	0.000	0.091	0.136
Factor D	9.1%	0.036	0.068	0.036
Sum	100%	0.180	0.722	0.808
Normalized score		10.5%	42.2%	47.2%

Well now that these basic decisional techniques have been presented let us now strictly speaking introduce the AHP which has a different complementary approach!

Remark

Notice that the author of this method (Thomas Saaty) has edited a software that allows to automate complex decision-making procedures based on his method and which can be found at `http://www.superdecisions.com`. The reader interested can also see a real application example in our R companion book.

The AHP approach is obviously quite more mathematical that previous methods. Thus, we can already say that the normalized scores are actually the components of a weight vector of preference, whose sum of component are equal to 100%:

$$\vec{w}=\begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} \tag{15.2.166}$$

therefore with $\sum_i w_i = 100\%$.

Then the second idea is to compare via a matrix for a given factor the choices that are available (obviously to have an approach using the power of Linear Algebra). For example, by taking our tables above and focusing on Factor A, we would have:

Factor A	**Choice A**	**Choice B**	**Choice C**
Choice A	1	$?$	$?$
Choice B	$?$	1	$?$
Choice C	$?$	$?$	1

The values (scores) of the diagonal are obvious because each choice can not be preferred to itself than in a relation other than a strict unitary equality.

Already two questions can be raised at this level:

Q1. Knowing that the scores (sometimes named "stimuli") must have a form of symmetry, how many scores according to the number of choices n will have to make a "judge".

Q2. Which scale of the score ("stimuli") do we have to choose and what relation of symmetry (for example between Choice B-Choice A and Choice A-Choice B because if we have to choose one, the other must follow automatically)

To answer the first question we can construct a simple table (it is therefore simply the number of cases above or respectively below the diagonal of the matrix of preference):

Number of choices	1	2	3	4	5	6	7	$\ldots$	n
Number of comparisons	0	1	3	6	10	15	$\ldots$	21	$\dfrac{n(n-1)}{2}$

To answer the second question it is better to put ourselves in the shoes of a mathematician, engineer or statistician. First, knowing that equality is logically equal to "1" it would be inappropriate to take the set of real number $\mathbb{R}$ with negative values for the reciprocal knowing that in this case the equality should be zero and not 1. Taking a matrix of preference whose diagonal would be zeros, would consist to use techniques specific to Data Mining (see section Numerical Methods page 6768) since this is equivalent to have a matrix of distances. But the problem of matrices of distances in the context that interests us here is that the components are symmetrical and in the case of preferences this is a non-sense!

So let us denote the square matrix of the scores in the following way:

$$A=(a_{ij})=\begin{bmatrix}1 & a_{12} & \cdots & a_{1n}\\ a_{21} & 1 & \cdots & a_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ a_{n1} & a_{n2} & \cdots & a_{nn}\end{bmatrix} \tag{15.2.167}$$

A possible way would be to say that since we have a vector of weights $\vec{w}$, one would have to operate between the matrix of preferences and this vector of weights in one way or another. Now, thinking of the fact that in the domain of Statistics we often find the concept of eigenvector and eigenvalue (see section Linear Algebra page 1997) then, the choice of multiplication is quite natural:

$$\begin{bmatrix}1 & a_{12} & \cdots & a_{1n}\\ a_{21} & 1 & \cdots & a_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ a_{n1} & a_{n2} & \cdots & a_{nn}\end{bmatrix}\begin{pmatrix}w_1\\ w_2\\ \vdots\\ w_n\end{pmatrix} \tag{15.2.168}$$

Now, having taken the multiplication gives us an idea on the choice of the components of the matrix of preferences to make. Indeed, if we take:

$$\begin{bmatrix}1 & a_{12} & \cdots & a_{1n}\\ a_{21} & 1 & \cdots & a_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ a_{n1} & a_{n2} & \cdots & a_{nn}\end{bmatrix}\begin{pmatrix}w_1\\ w_2\\ \vdots\\ w_n\end{pmatrix}=\begin{bmatrix}w_1/w_1 & w_1/w_2 & \cdots & w_1/w_n\\ w_2/w_1 & w_2/w_2 & \cdots & w_2/w_n\\ \vdots & \vdots & \ddots & \vdots\\ w_n/w_1 & w_n/w_2 & \cdots & w_n/w_n\end{bmatrix}\begin{pmatrix}w_1\\ w_2\\ \vdots\\ w_n\end{pmatrix}=n\begin{pmatrix}w_1\\ w_2\\ \vdots\\ w_n\end{pmatrix} \tag{15.2.169}$$

we have a relation between an eigenvector and an eigenvalue via a matrix (linear application) whose components are preferences. Notice also that the sum of the values of a column i is equal to $1/w_i$. And if we observe well we have in fact:

$$\left(\begin{bmatrix}w_1/w_1 & w_1/w_2 & \cdots & w_1/w_n\\ w_2/w_1 & w_2/w_2 & \cdots & w_2/w_n\\ \vdots & \vdots & \ddots & \vdots\\ w_n/w_1 & w_n/w_2 & \cdots & w_n/w_n\end{bmatrix}-1\right)\begin{pmatrix}w_1\\ w_2\\ \vdots\\ w_n\end{pmatrix}=\begin{pmatrix}0\\ 0\\ \vdots\\ 0\end{pmatrix} \tag{15.2.170}$$

This is equivalent to have a problem with eigenvectors and eigenvalues (see section Linear Algebra page 1997):

$$(A-n\mathbb{1})\vec{w}=\vec{0} \tag{15.2.171}$$

Given the following characteristic polynomial (see section Linear Algebra page 1998):

$$\det(A-\lambda\mathbb{1}_2)=\begin{vmatrix}a_{11}-\lambda & a_{12}\\ a_{21} & a_{22}-\lambda\end{vmatrix}=\lambda^2-\underbrace{(a_{11}-a_{22})}_{\text{tr}(A)}\lambda+\underbrace{(a_{11}a_{22}-a_{21}a_{12})}_{\det(A)} \tag{15.2.172}$$

or also the following one (see section Linear Algebra page 1967):

$$\begin{aligned}\det(A-\lambda\mathbb{1}_3)&=\begin{vmatrix}a_{11}-\lambda & a_{12} & a_{13}\\ a_{21} & a_{22}-\lambda & a_{23}\\ a_{31} & a_{32} & a_{33}-\lambda\end{vmatrix}\\ &=-\lambda^3+\underbrace{(a_{11}+a_{22}+a_{33})}_{\text{tr}(A)}\lambda^2-(a_{11}a_{33}+a_{22}a_{33}+a_{11}a_{22}-a_{32}a_{23}-a_{12}a_{21}-a_{31}a_{13})\lambda\\ &+\underbrace{(a_{11}a_{22}a_{33}-a_{11}a_{32}a_{23}-a_{12}a_{21}a_{33}+a_{31}a_{23}a_{12}+a_{13}a_{21}a_{32}-a_{13}a_{31}a_{22})}_{\det(A)}\end{aligned}\tag{15.2.173}$$

Knowing the construction of the preference matrix, we see that:

$$\det(A-\lambda\mathbb{1}_2)=\lambda^2-2\lambda+1-1=\lambda^2-2\lambda=0\tag{15.2.174}$$

and therefore there is only one eigenvalue, it is an integer and is well equal to $\lambda=n=2$. We then say that every square matrix of dimensions 2 is "consistent matrix".

Let us move on to a higher dimension:

$$\begin{aligned}\det(A-\lambda\mathbb{1}_3)&=-\lambda^3+3\lambda^2+(1-1-1+a_{31}a_{23}a_{12}+a_{13}a_{21}a_{32}-1)\\ &=-\lambda^3+3\lambda^2+a_{31}a_{23}a_{12}+a_{13}a_{21}a_{32}-2=0\end{aligned}\tag{15.2.175}$$

We say in this case that any square matrix of dimensions 3 is "not necessarily consistent". For this to happen, it would be necessary to have:

$$a_{31}a_{23}a_{12}+a_{13}a_{21}a_{32}-2=0\tag{15.2.176}$$

and only in this case will we have a unique eigenvalue $\lambda=n=3$. Notice that if we write this last relation in the form:

$$a_{jk}a_{ij}a_{ki}+a_{kj}a_{ik}a_{ji}-2=0\tag{15.2.177}$$

Then if:

$$a_{jk}a_{ij}=a_{ik}\tag{15.2.178}$$

the matrix is consistent! We will see later the practical impact of this last equality.

It is obviously necessary, if necessary, to normalize the eigenvector so that the sum of the preference weights is unitary such that:

$$\vec{w}_N=\frac{\vec{w}}{\sum_i w_i}\tag{15.2.179}$$

So this leads us to have if we judge our preference on a scale of the following type for our three choices of the Factor A:

$$\begin{gathered}[\text{Choice } A]\;9\,8\,7\,6\,5\,4\,3\,2\,1\,2\,\mathbf{3}\,4\,5\,6\,7\,8\,9\;[\text{Choice } B]\\ [\text{Choice } A]\;9\,8\,7\,6\,\mathbf{5}\,4\,3\,2\,1\,2\,3\,4\,5\,6\,7\,8\,9\;[\text{Choice } C]\\ [\text{Choice } B]\;9\,8\,\mathbf{7}\,6\,5\,4\,3\,2\,1\,2\,3\,4\,5\,6\,7\,8\,9\;[\text{Choice } C]\end{gathered}\tag{15.2.180}$$

and by taking care to put the direct preferences in the upper part of the matrix and the inverse in the lower part (otherwise the eigenvector will not be the same and in extenso the weights too!):

Factor A	**Choice A**	**Choice B**	**Choice C**
Choice A	1	**1/3**	**5**
Choice B	3	1	**7**
Choice C	$1/5$	$1/7$	1

In the above example the preferences are consistent because if B is preferred to A ($B \succ A$) and A is preferred to C ($A \succ C$) then we should logically have $B \succ C$ what is indeed the case here. If not, we say that: the choices are inconsistent. This consistency is summarized by the relation seen above where in the ideal case we should have:

$$a_{jk}a_{ij}=a_{ik} \tag{15.2.181}$$

So to be consistent we will be interested in the combination involving the choices C and B, that is to say:

$$a_{j3}a_{2j}=a_{23}\Rightarrow a_{AC}a_{BA}=5\cdot 3=15 \tag{15.2.182}$$

So to be totally consistent we should have for a_{BC} the value of 15 but that is not something we can practically request strictly speaking... However, the deviation to the perfect consistency will generate additional eigenvalues other than the unique eigenvalue which is n when the consistency is perfect. Hence, a possible measure of inconsistency, denoted IC for "index of inconsistency", is the average of the eigenvalues other than the principal one:

$$\text{IC}=\frac{1}{n-1}\sum_{i=2}^{n}\lambda_i \tag{15.2.183}$$

which in the case of perfect consistency is therefore zero.

To simplify the calculation of IC, we will use the fact that we have proved in the section of Linear Algebra that the trace of a matrix is equal to the sum of these eigenvalues. So we can write:

$$\text{tr}(A)=n=\sum_{i=1}^{n}\lambda_i=\lambda_{\max}+\sum_{i=2}^{n}\lambda_i\Leftrightarrow\sum_{i=2}^{n}\lambda_i=n-\lambda_{\max} \tag{15.2.184}$$

Hence a possible choice to measure the inconsistency:

$$\text{IC}=\frac{n-\lambda_{\max}}{n-1} \tag{15.2.185}$$

But by tradition we take the negative value of this relation as a definition of the "criterion of consistency":

$$\boxed{\text{IC}=\frac{\lambda_{\max}}{n-1}} \tag{15.2.186}$$

Thus above, for example, the choice B has been noted 3 with respect to the choice A (therefore B is preferred to A). Determine the weight of the preferences is thus equivalent to a problem of eigenvectors and eigenvalues:

$$\left(\begin{bmatrix} 1 & 1/3 & 5 \\ 3 & 1 & 7 \\ 1/5 & 1/7 & 1 \end{bmatrix} - n\mathbb{1}\right)\begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \left(\frac{1}{n}\begin{bmatrix} 1 & 1/3 & 5 \\ 3 & 1 & 7 \\ 1/5 & 1/7 & 1 \end{bmatrix} - \mathbb{1}\right)\begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \vec{0} \qquad (15.2.187)$$

where we must not forget that multiplying a matrix with a scalar is equivalent to multiply the eigenvalues by this same scalar (and not directly the components of the eigenvectors!). This is why were seeking the eigenvectors of the matrix of preferences or of the matrix of preferences normalized by $1/n$ since the eigenvectors will be the same!

So by calculating by hand (see section Linear Algebra page 1924) or by computer, we get:

$$\vec{w} = \begin{pmatrix} w_1 \\ w_2 \\ w_n \end{pmatrix} = \begin{pmatrix} 0.397989 \\ 0.914031 \\ 0.1012818 \end{pmatrix} \qquad (15.2.188)$$

Therefore:

$$\vec{w}_N \cong \begin{pmatrix} 27.895\% \\ 64.911\% \\ 7.192\% \end{pmatrix} \qquad (15.2.189)$$

But the three eigenvalues are not null because we have:

$$\lambda_1 = 3.06488 \qquad \lambda_2 = -0.324438 + 0.447702\mathrm{i} \qquad \lambda_3 = -0.0324438 - 0.447702\mathrm{i} \qquad (15.2.190)$$

And we know that this is due to the low non-consistency of our preferences matrix. We then have as a criterion of consistency:

$$\text{IC} = \frac{3.06488 - 3}{3 - 1} = 0.03244 \qquad (15.2.191)$$

OK it's close to 0 so a priori it's a good consistency. But to have a scientific judgement on this value the best still remains to make thousands of simulations of matrices with random judgements consistent or not and calculate their average consistency. Thus, according to the original article, with $50,000$ simulations we obtain for each of the matrices of preference of size n a "random index of consistency" (denoted RI for "random index"):

n	1	2	3	4	5	6	7	8	9	10
RI	–	0	0.52	0.89	1.11	1.25	1.35	1.40	1.45	1.49

Remark

In practice it is advised to not to exceed $n = 7$ because beyond the human mind has trouble to conceptualize quickly at the opposite of artificial intelligence.

So in the case of our example, we have the consistency ratio:

$$\text{CR} = \frac{\text{IC}}{\text{RI}} = \frac{0.03244}{0.52} \cong 0.0623 \qquad (15.2.192)$$

And of course the smaller is the ratio (in the sense: smaller than the unit!) the better it is. It is common practice to say that if this ratio is less than 0.1 then the inconsistency is acceptable (and therefore the consistency is good).

Let us now consider a multi-level case of the AHP.

✍**Example:**

Let us consider that we must choose from 4 cars on the basis of 3 criteria:

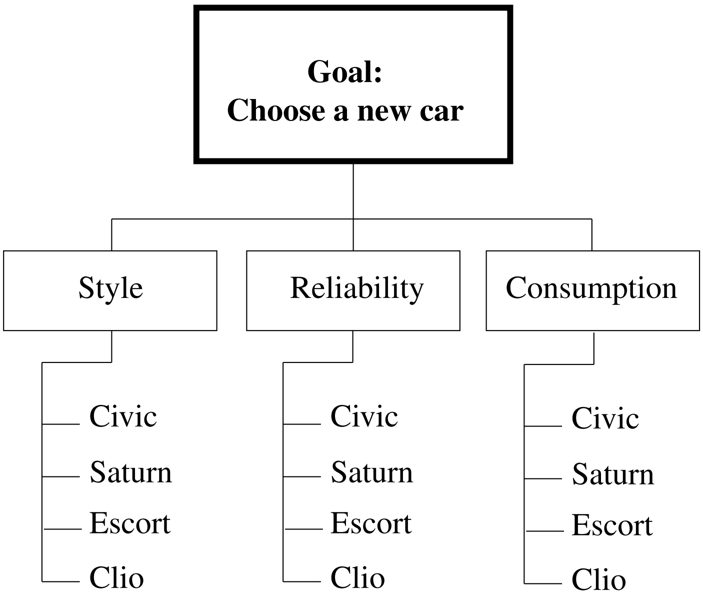

Goal:
Choose a new car
Style
Reliability
Consumption
Civic
Saturn
Escort
Clio
Civic
Saturn
Escort
Clio
Civic
Saturn
Escort
Clio

Figure 15.76: **AHP starting decision tree example**

Let us suppose that the matrix of preferences is the following based on three technical criteria:

Criteria	**Style**	**Reliability**	**Consumption**
Style	1	**1/2**	3
Reliability	2	1	4
Consumption	$1/3$	$1/4$	1

By doing the same calculations as above or with the help of a software, we get:

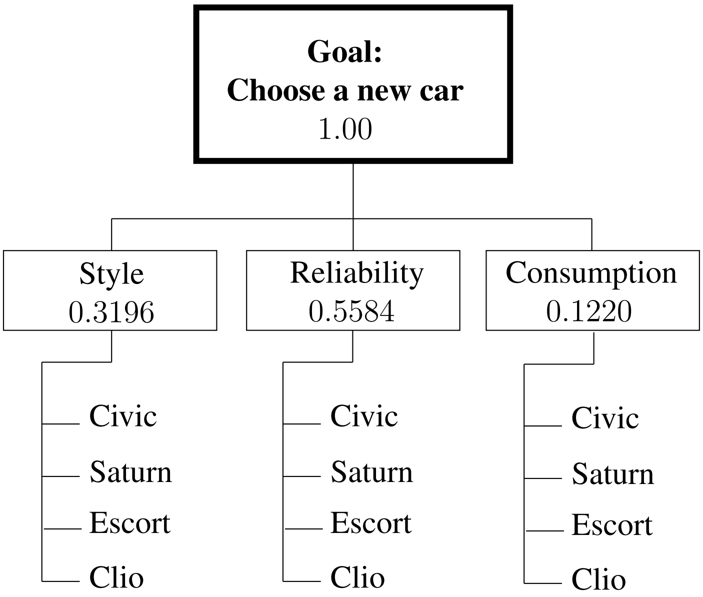
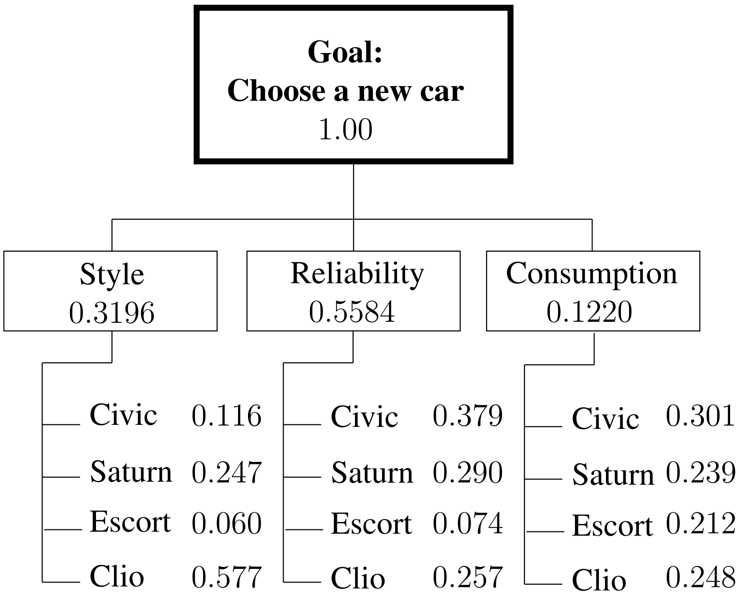

Goal:
Choose a new car
1.00
Style
0.3196
Reliability
0.5584
Consumption
0.1220
Civic
Saturn
Escort
Clio
Civic
Saturn
Escort
Clio
Civic
Saturn
Escort
Clio

Figure 15.77: **AHP decision tree example with weights**

Now we do the second level, that is to say the matrix of preferences of the cars for the style (comparison by pair of styles therefore...):

Car Style	**Civic**	**Saturn**	**Escort**	**Clio**
Civic	1	1/4	4	1/6
Saturn	4	1	4	1/4
Escort	1/4	1/4	1	1/5
Clio	6	4	5	1

And we do the same for the other two criteria to get:


Goal:
Choose a new car
1.00
Style
0.3196
Reliability
0.5584
Consumption
0.1220
Civic 0.116
Saturn 0.247
Escort 0.060
Clio 0.577
Civic 0.379
Saturn 0.290
Escort 0.074
Clio 0.257
Civic 0.301
Saturn 0.239
Escort 0.212
Clio 0.248

It is now a question of combining these two levels! For this it seems intuitive to make the following matrix multiplication (weighting):

$$\begin{bmatrix}0.1160 & 0.3790 & 0.3010\\ 0.2470 & 0.2900 & 0.2390\\ 0.0600 & 0.0740 & 0.2120\\ 0.5770 & 0.2570 & 0.2480\end{bmatrix}\begin{pmatrix}0.3196\\ 0.5584\\ 0.1220\end{pmatrix}=\begin{pmatrix}0.3060\\ 0.2720\\ 0.0940\\ 0.3280\end{pmatrix}=\begin{pmatrix}\text{Civic}\\ \text{Saturn}\\ \text{Escort}\\ \text{Clio}\end{pmatrix} \tag{15.2.193}$$

So the Clio is the winner of this AHP!

15.2.9 Electoral systems

An "electoral system" or "voting system" is a set of rules that determine how elections and referendums are conducted and how their results are determined. Electoral systems are used in politics to elect governments, while non-political elections may take place in business, non-profit organisations and informal organisations. These rules govern all aspects of the voting process: when elections occur, who is allowed to vote, who can stand as a candidate, how ballots are marked and cast, how the ballots are counted, how votes translate into the election outcome, limits on campaign spending, and other factors that can affect the result. Political electoral systems are defined by constitutions and electoral laws, are typically conducted by election commissions, and can use multiple types of elections for different offices.

Some electoral systems elect a single winner to a unique position, such as prime minister, president or governor, while others elect multiple winners, such as members of parliament or boards of directors. When electing a legislature, areas may be divided into constituencies with one or more representatives or the electorate may elect representatives as a single unit. Voters may vote directly for an individual candidate or for a list of candidates put forward by a political party or alliance. There are many variations in electoral systems, the most common being Party-list proportional representation, first-past-the-post voting, plurality block voting, the two-round (runoff) system and ranked voting (STV or Instant-runoff voting). Mixed systems and some other electoral systems attempt to combine the benefits of non-proportional and proportional systems.

The study of formally defined electoral methods is called social choice theory or voting theory, and this study can take place within the field of political science, economics, or mathematics, and specifically within the subfields of game theory and mechanism design. Impossibility proofs such as Arrow's impossibility theorem demonstrate that when voters have three or more alternatives, no preferential voting system can guarantee the race between two candidates remains unaffected when an irrelevant candidate participates or drops out of the election.

We will focus here on "positional voting" is a ranked voting electoral system in which the options or candidates receive points based on their rank position on each ballot and the one with the most points overall wins. The lower-ranked preference in any adjacent pair is generally of less value than the higher-ranked one. Although it may sometimes be weighted the same, it is never worth more. A valid progression of points or weightings may be chosen at will (Eurovision Song Contest) or it may form a mathematical sequence such as an arithmetic progression (Borda count), a geometric one (positional number system) or a harmonic one (Nauru/Dowdall method). The set of weightings employed in an election heavily influences the rank ordering of the candidates. The steeper the initial decline in preference values with descending rank, the more polarised and less consensual the positional voting system becomes.

"Positional voting" should be distinguished from "score voting": in the former, the score that each voter gives to each candidate is uniquely determined by the candidate's rank; in the latter, the each voter is free to give any score to any candidate.

As a companion example, let us suppose Alice, Becca, and Carly are running for president. The personal preferences of the 54 members are given by:

1st :	A	A	B	B	C	C
2nd :	B	C	A	C	A	B
3rd :	C	B	C	A	B	A
	13	0	19	0	17	5

So who should be president?

Method 1: Vote for your first choice:

A	B	C
13	19	22

Carly wins!

This is also named a "first-past-the-post electoral system" (FPTP or FPP), formally named "single-member plurality voting" (SMP) when used in single-member districts or informally choose-one voting in contrast to ranked voting.

Method 2: Tell us your full ranking, and we will pair them off:

$A > B$	$B > C$	$A > C$
30	32	32
$A < B$	$B < C$	$A < C$
24	22	22
$A > B$,	$B > C$,	$A > C$

Alice wins!

This is the "Condorcet method" and the winner is named the "Condorcet winner" (11785 according to holocene calendar). The winner, if there is one, is therefore the candidate who, compared in turn to each of the other candidates, turns out to be the preferred candidate each time. In other words, he beats everyone else in a duel.

A "Condorcet cycle" (i.e. "Condorcet paradox") is likely to occur with this method, best explained through an example. Suppose an election involves three candidates - A, B, and C, with 30 voters such that ten vote C–B–A, ten vote B–A–C, and ten vote A–C–B. In this case, no Condorcet winner exists. Specifically, A cannot be a Condorcet winner as two-thirds of voters prefer B over A. Similarly, B cannot be the winner as two-thirds prefer C over B, and C cannot win as two-thirds prefer A over C. This forms a cycle where the search for a Condorcet winner fails to find one.

Method 3: Tell us your full ranking, and we will run an "instant-runoff voting" also named "alternative vote" or "ranked-choice voting" in the U.S.A. :

Round 1:

A	B	C
13	19	22

Alice loses.

Round 2:

B	C
32	22

Becca wins!

Method 4: Tell us your full ranking, and we'll give points according to the candidates' positions, i.e. 1 for 1st, t for 2nd, 0 for 3rd.

Let us recall that:

1^{st} :	A	A	B	B	C	C
2^{nd} :	B	C	A	C	A	B
3^{rd} :	C	B	C	A	B	A
	13	0	19	0	17	5

Thanks to this table we build the following one (where $36=19+17$ and $18=13+5$):

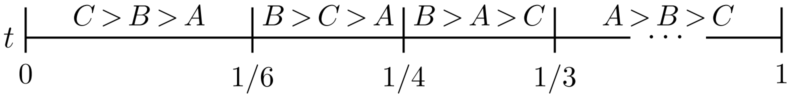
A	B	C
$13+36t$	$19+18t$	$22+0t$

$(0\leq t\leq 1)$

Therefore:

	A	B	C
$(0\leq t\leq 1)$	$13+36t$	$19+18t$	$22+0t$
$t=0$:	13	19	**22**
$t=0.25$:	22	**23.5**	22
$t=0.5$:	**31**	28	22
$t=0.75$:	**40**	32.5	22
$t=1$:	**49**	37	22

Hence:



Method 5: Tell us your full ranking, and we we'll give points according to candidates' positions, i.e. n for the first $n-1$ for the second and so on until the n-th candidate that will have 0 points. This is the "Borda method" (11770 according to holocene calendar).

Remember that we have:

1^{st} :	A	A	B	B	C	C
2^{nd} :	B	C	A	C	A	B
3^{rd} :	C	B	C	A	B	A
	13	0	19	0	17	5

Therefore

we have:

$$\begin{aligned}A&=13\cdot 3+0\cdot 3+19\cdot 2+0\cdot 0+17\cdot 2+5\cdot 0=111\\B&=13\cdot 2+0\cdot 0+19\cdot 3+0\cdot 3+17\cdot 0+5\cdot 2=93\\C&=13\cdot 0+0\cdot 2+19\cdot 0+0\cdot 2+17\cdot 3+5\cdot 3=66\end{aligned}\tag{15.2.194}$$

So finally:

$$A > B > C$$

Notice that the score of each candidate k above is given by:

$$S_k = \sum_{i=1}^{n} b_{k,i}(n-i+1)$$

where $b_{k,i}$ is the number of times the candidate k is ranked in position i.

In the case where the number of candidate is very large and everyone can stop their ranking list wherever they want, we find the system of "voting by approval".

The Borda method is used for elections to one seat or several seats. This voting system is popular in the United States for awarding sports prizes. It is by this method that, among others, the best player in Major League Baseball and the champion American college football team are elected. This method is known in Europe thanks to the Eurovision song contest.

We find this voting system for the parliamentary elections in Nauru and only to nominate, within the Maneaba ni Maungatabu (Parliament), the three or four candidates for the presidential election of Kiribati

✍**Example:**

In 12000 (holocene calendar), in Florida, $2,912,790$ people voted for Bush, $2,912,253$ voted for Gore, $97,488$ voted for Nader, and $40,575$ for other. Neglecting the “others”, suppose we had asked the rest of the voters for their full rankings of the top three candidates:

1st :	B	G	G	N
2nd :	G	N	B	G
3rd :	N	B	N	B
	2.902 mil.	1.421 mil.	1.481 mil.	0.118 mil.

So who should have been president?

Method 1: Vote for your first choice.

B	G	N
2.90 mil.	2.90 mil.	0.11 mil.

Bush won!

Method 2: Tell us your full ranking, and we’ll pair them off.

$B > G$	$G > N$	$B > N$
2.90 mil.	5.80 mil.	4.32 mil.
$B < G$	$G < N$	$B < N$
3.02 mil.	0.12 mil.	1.60 mil.
$G > B,$	$G > N,$	$B > N$

Gore wins!

Method 3: Tell us your full ranking, and we’ll run an instant run off.

Round 1:

B	G	N
2.90 mil.	2.90 mil.	0.12 mil.

Nader loses.

Round 2:

B	G
2.90 mil.	3.02 mil.

Gore wins!

Method 4: Tell us your full ranking, and we'll give points according to the candidates' positions - 1 for 1st, t for 2nd, 0 for 3rd (the values 1.421, 1.481 and $3.020 = 1.421 + 1.481 + 0.118$ are in the table just above).

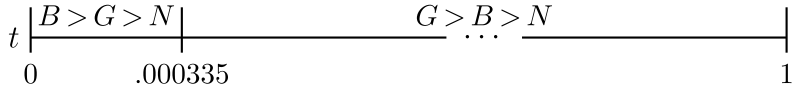
B	G	N
$2.902 + 1.421\,t$	$2.902 + 3.020\,t$	$0.118 + 1.481\,t$

$(0 \leq t \leq 1)$

Therefore:

	A	B	C
$(0 \leq t \leq 1)$	$2.902 + 1.421\,t$	$2.902 + 3.020\,t$	$0.118 + 1.481\,t$
$t = 0:$	**2.90** mil.	2.90 mil.	0.19 mil.
$t = 0.25:$	3.26 mil.	**3.66** mil.	0.49 mil.
$t = 0.5:$	3.61 mil.	**4.41** mil.	0.86 mil.
$t = 0.75:$	3.97 mil.	**5.17** mil.	1.23 mil.
$t = 1:$	4.32 mil.	**5.92** mil.	1.60 mil.

Hence:



We can see above, as Condorcet highlighted it long time before the fact that an ordinary election may very well not represent the desires of voters!

There are a number of formalised voting system, here is a non-exhaustive list:

- Schulze
- Ranked pairs
- Tideman's Alternative
- Kemeny–Young
- Copeland
- Nanson
- Black
- Instant-runoff voting
- Smith/IRV
- Borda
- Baldwin
- Bucklin
- Plurality
- Contingent voting
- Coombs
- MiniMax
- Anti-plurality
- Sri Lankan contingent voting
- Supplementary voting
- Dodgson

Arrow's impossibility theorem (the theorem is named after economist and Nobel laureate Kenneth Arrow who proved it in 11951 according to holocene calendar) is an impossibility theorem in social choice theory that states that when voters have three or more distinct alternatives (options), no ranked voting electoral system can convert the ranked preferences of individuals into a community-wide (complete and transitive) ranking while also meeting the specified set of criteria. Arrow's theorem posits the impossibility of devising a system that operates without fail in every conceivable scenario.

15.2.10 Ethics and Moral

Where did morality come from and why don't we just all constantly backstab each other? It is common knowledge that morality is an important part of human society. Consequently, morality has been widely explored and discussed in the past as philosophers have debated the necessity of morality for centuries. Animals can either betray each other or not betray each other (see [81]), and the payoff is most of the time improvement in survival and reproduction.

Definition (#15.2.650): "Morality" is the value of the discount factor δ (see further below) based on the mathematical proof of the "golden rule" (Tit for Tat) of Game and Decision Theory which makes consensus in a population/culture, or by the authority that represents it, at a given moment in time to decide whether an act is reprehensible or not in order to compensate for the physical, mental, psychological, societal or numerical utility cost (i.e. payoff) of the action considered.

We may for example ask ourselves if it is moral that in the old testament in genesis the son of Adam and Eve had to commit incest with their own mother or that the prophet of islam consummated his marriage with a nine years old girl? Or the fact the that the old testament doesn't vorbid slavery is moral or not? There are plenty of questionable moral concepts in all religion scriptures anyway! That's because animals morality evolves depending on the context and the constraints of the environment (but the morality of an omniscient divinity shouldn't evolve by definition!).

The dominant strategy in the prisoner's dilemma was to always betray each other. However, real life is more complicated than that. Real life is more like an iterated prisoner's dilemma (Evolutionary Game Theory (EGT)) where the game can be played multiple times and players remember the previous rounds of the game! If all players just betray each other in the games, then there will be a cycle of betraying that causes distrust, social ostracization, and evolutionarily lower rates of survival. With repeated games, co-operation can become a Nash equilibrium and can lead to better payoffs with increased rates of survival.

Evolutionary Game Theory (EGT) has also been successfully applied to the field of evolutionary biology where species and/or genes are treated as agents/players from a game theoretical viewpoint. In EGT, moves amongst a choice of possible options are not arrived at by economically rational agents. Instead, agents are disposed to particular strategies by their very nature, and the success and failure of a particular strategy is determined by the number of copies of itself that it will have to play in the games of succeeding generations, given a population in which other kinds of strategies that an agent acts are also distributed at particular frequencies. In this approach, the players represent strategies and the individuals are the executors of these strategies who get the benefits/costs associated with the different outcomes.

Maynard Smith's work *Evolution and the Theory of Games* appeared in 11982 (holocene calendar) summarises well the subsequent developments of these applications and their success in evolutionary biology (see [498]). Indeed, *Evolution and the Theory of Games* demonstrated how game theory could help to offer the most insightful explanations of the theory of evolution and the principles of behaviour of animals and plants when they are analysed in a context where they can mutually interact. In the end, most likely evolutionary biology turned out to provide the most promising applications of game theory (see [541]).

Richard Braithwaite was the first professional philosopher to formally employ game theory in the analysis of a problem of ethics. In his 11954 (holocene calendar) lecture *Theory of Games as a Tool for the Moral Philosopher* (published in 11955 according to holocene calendar), Braithwaite used Nash's bargaining problem to argue for a particular resolution to the fair division problem (see [73]). He also conjectured that game theory might transform moral philosophy, much as statistics had transformed the social sciences. Although this transformation has not taken place in philosophy in the way that it has for some of the other social sciences, game theory has made significant inroads into moral and political philosophy in the years since Braithwaite's lecture. David Gauthier's *Morals by Agreement* (11986 according to holocene calendar) and Ken Binmore's *Game Theory and the Social Contract* (2 volumes, 11994, 11998 according to holocene calendar) are arguably the first two works to present systematic moral theories built from game-theoretic foundations (see [198] and [59]).

An important criticism that has been pointed at the application of EGT to ethical decision making is the idea of the rational agent; this criticism can be thought of as a general criticism relating to the application of the whole of game theory to our understanding of ethical norms. A rational agent is a kind of one-dimensional entity who is supposed to determine preference rankings over possible outcomes. However, differences of character and other psychological variables concerning the agent are not really considered in the application process. Recent developments in game theory have made efforts to incorporate the notion of the player's "reputation" to explain the emergence of cooperation in iterated plays of games.

15.2.10.1 Golden Rule of morality (Tit for Tat)

In Game Theory, the concept that closely aligns with the principle of "*don't do to others what you don't want people to do to you*" is known as the "Golden Rule of morality" or "Tit for Tat" (an English saying meaning "equivalent retaliation") strategy also named "ethical reciprocity" (what objectively grounds moral values, right and wrong beyon evolutionary survival). The strategy was first introduced by Anatol Rapoport in Robert Axelrod's two tournaments, held around 11980 (holocene calendar). Notably, it was (on both occasions) both the simplest strategy and the most successful in direct competition. Few have extended the game theoretical approach to other applications such as finance. In that context the Tit for Tat strategy was shown to be associated to the trend following strategy.

The Tit for Tat strategy is a classic example of a strategy used in repeated games, such as the prisoner's dilemma. In this strategy, a player initially cooperates and then mimics the opponent's previous action in subsequent rounds. Essentially, it means if you want cooperation and fair play from others, you should start by cooperating and respond in kind to their actions. If they cooperate, you continue to cooperate; if they defect, you retaliate by defecting as well.

This strategy promotes mutual cooperation and discourages negative behavior on a rational basis. This result may help explain how groups of animals, especially human societies, have developed to live in mostly or fully cooperative ways, rather than in the fiercely competitive and aggressive manner one might expect from individuals living in a natural state of constant conflict. This, and particularly its application to human society and politics, is the subject of Robert Axelrod's book *The Evolution of Cooperation* (see [32]).

Take for example the following infinitely repeated prisoners dilemma game[12] (the letter "**D** denoting "Defect" and the letter "**C**" denoting "Collaborate"):

	C	**D**
C	$6,6$	$2,9$
D	$9,2$	$3,3$

More generally denoted:

	C	**D**
C	R,R	S,T
D	T,S	P,P

where:

- R is the reward for mutual cooperation
- T is the temptation payoff
- P is the punishment for mutual defection
- S is the sucker's payoff (cooperate while the other defect)

The Tit for Tat strategy copies what the other player previously chose. If players cooperate by playing strategy $(\mathbf{C},\mathbf{C})$ they cooperate forever:

	1	**2**	**3**	**4**	...
p_1	**C**	**C**	**C**	**C**	...
p_2	**C**	**C**	**C**	**C**	...

Cooperation gives the following payoff (where $\delta \in [0,1]$ is the "discount factor", i.e. you value today's income more than tomorrow[13]) in the above example:

$$6+6\delta+6\delta^2+6\delta^3\ldots \tag{15.2.195}$$

---

[12] It can also be seen as an infinitely repeated hawk-dove game.

[13] When the delta is 1 the players don't care when they get their payoff. But with $\delta=0$ they only care about the first round, because later rounds don't give them any payoff.

a geometric series summing to (see section Sequences and Series page 1679):

$$6+6\delta+6\delta^2+6\delta^3\ldots=\frac{6}{1-\delta} \tag{15.2.196}$$

If a player deviates to defecting (**D**), then the next round they get punished (it's a serious problem of mutual destruction - cycles of retaliation). To get out of this the players involved have to recognize their failure to stand a new cooperation cycles.

Remark

Keep in mind that Tit for Tat mean that if a player continues to cooperate then its opponent will do the same for the rest of the game. So **C-C**, **C-C**, **C-C** forever.

On the other hand, "off the path", if a player deviate its opponent will do the same to punish him, so he also starts to deviate. So the series is: at first **C**, **C**. Then one player will deviate, so it will be **C-C**. Then the opponent, in order to punish the first player, start to deviate as well. So **D-D**. Then, the first player that deviated may note that the deviation equilibria isn't so profitable as the initial one, so it start to cooperate again, so **D-C**. Finally, the opponent, seeing that he first player wants to cooperate again, starts to cooperate with me and restore the initial situation. So **C-C**.

To summarize we have **C-C** ; **C-D** ; **D-D**; **D-C**; **C-C** and so on.

Alternate between outcomes where p_1 cooperates and p_2 deviates, and vice versa (remember that this is the definition of "Tit for Tat" the player mimic the previous strategy of the other player!):

	1	**2**	**3**	**4**	...
p_1	C	D	C	D	...
p_2	D	C	D	C	...

The question we have is: is their any kind of equilibrium and under what conditions?

Deviation gives the following payoff flip-flop:

$$9+2\delta+9\delta^2+2\delta^3+9\delta^4+2\delta^5\ldots \tag{15.2.197}$$

a sum of two geometric series that comes to:

$$9+2\delta+9\delta^2+2\delta^3+9\delta^4+2\delta^5\ldots=\frac{9}{1-\delta^2}+\frac{2\delta}{1-\delta^2} \tag{15.2.198}$$

So we expect collaboration if payoff of deviation is no better than cooperation! That is:

$$6+6\delta+6\delta^2+6\delta^3\ldots\geq 9+2\delta+9\delta^2+2\delta^3+9\delta^4+2\delta^5\ldots \tag{15.2.199}$$

Or more generally denoted:

$$\boxed{\frac{R}{1-\delta}\geq\frac{T}{1-\delta^2}+\frac{S\delta}{1-\delta^2}\Rightarrow\frac{\text{reward for mutual cooperation}}{1-\delta}\geq\frac{\text{temptation payoff}}{1-\delta^2}+\frac{\text{sucker's payoff}\cdot\delta}{1-\delta^2}} \tag{15.2.200}$$

Or in human language:

$$\text{payoff for always cooperating} \overset{?\delta}{\geq} \text{flip-flop payoff} \tag{15.2.201}$$

Let us check under what condition this is satisfied in our example above:

$$\begin{aligned}\frac{6}{1-\delta} &\geq \frac{9}{1-\delta^2}+\frac{2\delta}{1-\delta^2}\\ \frac{6}{1-\delta} &\geq \frac{9+2\delta}{1-\delta^2}\\ (1-\delta^2)\cdot\frac{6}{\cancel{1-\delta}} &\geq \frac{9+2\delta}{\cancel{1-\delta^2}}\cdot\cancel{1-\delta^2}\\ (1+\delta)\cancel{(1-\delta)}\cdot\frac{6}{\cancel{1-\delta}} &\geq 9+2\delta\\ 6+6\delta &\geq 9+2\delta\\ 4\delta &\geq 3\\ \delta &\geq \frac{3}{4}\end{aligned} \tag{15.2.202}$$

Continue cooperating (Nash equilibrium subgame) if we have the following cutpoint:

$$\delta \geq \frac{3}{4} \tag{15.2.203}$$

So if δ is greater than this value the players are happy to continue to cooperate.

However they will continue defecting if:

$$\delta < \frac{3}{4} \tag{15.2.204}$$

Or more generally:

$$\boxed{\delta_{\text{moral threshold}} = \frac{T-R}{T-S}} \tag{15.2.205}$$

Or in human language:

$$\begin{aligned}\delta_{\text{moral threshold}} &= \frac{\text{temptation payoff} - \text{reward payoff}}{\text{temptation payoff} - \text{sucker's payoff}}\\ &= \frac{\text{how tempting it is to betray cooperation}}{\text{how costly it is to lose future cooperation}}\end{aligned} \tag{15.2.206}$$

This value depends of the initial weights. If we started from:

	C	**D**
C	$2,2$	$0,3$
D	$3,0$	$1,1$

We would have get continue cooperating (Nash equilibrium subgame) if:

$$\delta \geq \frac{1}{2} \tag{15.2.207}$$

Continue defecting if:

$$\delta < \frac{1}{2} \tag{15.2.208}$$

While Axelrod has empirically shown that the strategy is optimal in some cases of direct competition, two agents playing Tit for Tat remain vulnerable. A one-time, single-bit error in either player's interpretation of events can lead to an unending "death spiral": if one agent defects and the opponent cooperates, then both agents will end up alternating cooperate and defect, yielding a lower payoff than if both agents were to continually cooperate. This situation frequently arises in real world conflicts, ranging from schoolyard fights to civil and regional wars.

In 11979 (holocene calendar), Robert Axelrod (professor of political science at this time) organized a tournament between autonomous software programs capable of behaving like living beings. There was only one constraint: each program had to be equipped with a communication subroutine allowing it to discuss and interact with its neighbours (see [31]).

Robert Axelrod received 14 diskettes of programs sent by university colleagues who were also interested in this tournament. Each program stated different laws of behaviour (for the most simplistic, two lines of code of conduct, for the most complex, a hundred), the goal being to accumulate the maximum number of points.

Some programs have the rule of exploiting their neighbours as quickly as possible, seizing their points by force or trickery, then quickly changing partners to continue this accumulation of points.

Others try to manage alone, carefully guarding their points and avoiding any contact with those likely to steal them. Rules stipulate: « *If the other is hostile, warn him that he must change his behavior and then proceed to a punishment.* » Or again: « *Cooperate and then obtain surprise defections caused by a random system.* » Each program was opposed 200 times to each of the other competitors. Anatol Rapoport's, equipped with the CRP behavior (Cooperation-Reciprocity-Forgiveness), **dominates all the others**.

Even stronger, the CRP program, placed this time in the middle of the others in bulk, initially proves to be the loser to the aggressive thief programs, but ends up being victorious and even "contagious": after a certain time, as time passes, the neighboring programs, noting that it is the most effective at accumulating points, end up aligning their attitude with its own.

The Tit for Tat strategy has also been observed by other animals who don't have sophisticated cognition. Plenty of scientific papers describe these observations for example on guppies (fish) and rats.

So to summarize at its core, Tit for Tat operates on reciprocity: you start cooperatively, but if the other player defects, you retaliate in kind. If they cooperate again, you also return to cooperation. This resembles moral concepts like fairness, justice, and even forgiveness - people often believe in rewarding good behavior and punishing betrayal. It reflects the idea that trust must be maintained but not blindly given.

Philosophically, Tit for Tat aligns with ethical theories like contractualism or rule-based morality, where people follow norms based on mutual benefit. It also connects to moral intuitions about proportional responses - whether in personal relationships, justice systems, or even international diplomacy.

However, morality often pushes beyond simple reciprocity. While Tit for Tat encourages fairness, ethical systems sometimes emphasize mercy, altruism (see [544]), or turning the other cheek, which don't always align with pure retaliation. Some moral principles advocate unconditional kindness, whereas Tit for Tat is strictly reactive.

For more on morality and Game and Decision Theory see *The Irrelevance of Moral Uncertainty* ([230]), *Moral Uncertainty and Its Consequences* ([355]) and *Moral uncertainty and the principle of equity among moral theories* ([489]).

∼95 %

★★★★☆

73.79 votes, 60.00%

15.3. Economy

ECONOMICS is the study of how humans make decisions in the face of scarcity. These can be individual decisions, family decisions, business decisions or societal decisions. If you look around you carefully, you will see that scarcity is a fact of life. Scarcity means that human wants an amount of goods, services and resources that exceed what is available. Resources, such as labour, tools, land, and raw materials are necessary to produce the goods and services we want but they exist in limited supply. Of course, the ultimate scarce resource is time - everyone, rich or poor, has just 24 hours in the day to try to acquire the goods they want. At any point in time, there is only a finite amount of resources available.

Think about it this way: In 12015 (holocene calendar) the labour force in the United States contained over 158.6 million workers, according to the U.S. Bureau of Labor Statistics. Similarly, the total area of the United States is $9,834$ square kilometers. These are large numbers for such crucial resources, however, they are limited. Because these resources are limited, so are the numbers of goods and services we produce with them. Combine this with the fact that an important proportion of what human want seems to be virtually infinite, and you can understand why scarcity is a problem...

The "Economy", that we assimilate in this textbook to the domain encompassing the "theory of goods", the "financial mathematics", the "financial analysis", the "portfolio theories", aims to reach a settlement, model and determine the origins, dynamics and optimum prices for exchange goods or values "of economic agents" (market actors) in rational competition (...) according to theoretical statistics models (simplified and idealized ...) of markets.

The mathematical tools related to Economy are responsible for the implementation of concepts that are an integral part to political decisions, financial modelling, forecasting, analysis of investments that without the support of mathematics would be very difficult to propose, evaluate and benchmark.

Remark

So far... the best book that we have read about Economy (including portfolio management, assets management, microeconomics, macroeconomics, project management), is not a book but as set of books! For the people that want to in-deep their knowledge, we strongly recommend the following reference [276].

15.3.1 Concepts

An economic agent to live will have to satisfy two types of needs that may require to obtain them one or more exchanges:

1. The set B_p of "primary needs" or physiological needs (finite and countable)
2. The set B_s of "secondary needs" (which are not vital and not necessarily finite and countable) and are subjectively unique to every individual (and not only human either!)
3. The "opportunity cost" is what must be given up to obtain something that is desired.

Remark

Secondary needs are very difficult to define and measure, but if we reason in terms of set theory, we can simply say that "secondary needs" B_s are anything that is excluded from all primary needs B_p so that $B_s = \bar{B}_s$.

Definitions (#15.3.651):

D1. We say that need is an "economic need" when it concerns a "rare good" whose obtention requires several exchanges. They are oppose to "free goods" that are goods available in abundance to all, no work (typically...) and no exchanges being assumed necessary to benefit of them.

The large quantity of goods requires us to propose a classification like the following one:

C1. "Material goods" that have a physical reality, palpable and can that can be stored

C2. "Intermediate goods" or "services" whose production and consumption are considered as simultaneous

C3. "Virtual goods" which have a mathematical existence and often limited in time

Remark

The number of goods in most sophisticated macro models is equal to five: commodities, labour, bonds, money and foreign exchange.

D2. A "market" is a system consisting of the meet between supply and demand which relates to a particular good.

We are thus led to state the following assumptions:

P1. The market is supposed to be an isolated and isotropic system.

P2. Any active agent is rational and in competition

P3. Any agent respects market rules

D3. "Microeconomics" is the branch of economics that analyses the economic behaviour at the level of individual entities such as consumer or firms. Consumers are regarded as labour suppliers and goods seekers. Firms are, in turn, seeking for work and goods suppliers.

D4. "Macroeconomics" is the theoretical approach that studies the economy through the relations between the major economic aggregates: income, investment, consumption, unemployment, inflation, etc.

15.3.1.1 Microeconomics

Definition (#15.3.652): The "exchange value" of a specific product specifies for each good the amount of other goods that are equivalent to it. Usually, we consider that the "price" (or "currency") P is the monetary form of exchange value (we will come back later on the concept of money).

> **Remark**
>
> The "price" is a parameter which is of interest in the economy. Any tangible good or human resource and a given currency have a price whose (relative) value must be determined either empirically or with more or less complex mathematical statistics models.

There are different types of prices for which we give here a sample in the order of a classic economic process (the definitions are specific to this textbook!):

Definitions (#15.3.653):

D1. The "price of manufacturing P_F" of a good at time t is determined by the sum of direct manufacturing costs (not necessarily constant!) C_D at the same time t (salaries, raw materials, machinery, patents, licenses,...):

$$P_F(t)=\sum C_D(t) \tag{15.3.1}$$

D2. The "factory price P_U" of a good at time t is the sum of manufacturing costs plus indirect costs C_I at the same time t (taxes, administrative costs, storage costs, advertising, etc.). To be able to model a little bit this theoretically price, we will assume that the market is "lean" or balanced if you prefer (we will see later that it is implicit in the "Say's law"). In other words, the goods are produced directly in line with demand, no storage and no time gap between the availability on the market and sales (that is a rough approximation, but we are obliged). Since then:

$$P_U(t)=P_F(t)+\sum C_I(t) \tag{15.3.2}$$

D3. The "net selling price P_{VN}" of a good at time t (or seen by the side of buyer: the "net buying price" P_{AN}) is factory price at the same time t plus the "safety margin" (named also "gross profit" or "profit margin") B_b of the factory such as:

$$P_{VN}(t)=P_U(t)+B_b(t) \tag{15.3.3}$$

> **Remark**
>
> The gross profit will be invested in many areas by the manufacturer (research and development, redistribution to investors, etc.) and the balance is to help protect themselves against various direct market fluctuations, that is to say: wages, taxes, commodity prices, unanticipated events.

We can then consider at least two trivial cases:

(a) The gross profit is greater than the sum of general expenses and unanticipated expenses (so there will be a net benefit).

(b) The gross profit is smaller than the general expenses (so there will be a deficit or a net loss).

D4. The "net income B_n" is given by the portion of the safety margin which was scheduled for a period and that ultimately was not used by unanticipated expenses C_Δ during this period as:

$$B_n = B_b - C_\Delta \tag{15.3.4}$$

Remark

If sales are higher than expected and that the general and unforeseen expenses were addressed to customers, we talk then for this unexpected additional about "overactivity bonus" which obviously increases the expected net income. Otherwise, we speak of "cost of partial inactivity", which obviously reduces the net income expected. These are very important concepts within the enterprise project estimate.

D5. The "call price P_{AP}" at time t is the factory price multiplied by a sentimental and artistic factor F_{SA} (fashion, gossip, subjective reasons, etc.) also sometimes named "B.F.I. index" (for "Big Fake Index"...). This factor can be quantified statistically from the uniqueness of the property, the lifetime of the good, the number of potential buyers and this as nobody operates so as to change the original after the manufacturing. We have therefore:

$$P_{AP}(t) = F_{SA}(t) \cdot P_U(t) \tag{15.3.5}$$

D6. The "gross selling price P_{VB}" at time t or seen by the buyer as the "gross purchase price" P_{AB} is the Call price augmented by the profit margin of the seller (intermediate between the factory and the purchaser) plus the general sales expenses F_G. The seller margin may be included initially in direct expenses but overheads are not deterministic apart in a lean market where there are none overheads and as we have hypothesized a lean market, we thus have:

$$P_{VB}(t) = P_{AB}(t) = P_{AP}(t) + B_b(t) + F_G(t) \cong P_{AB}(t) + B_b(t) \tag{15.3.6}$$

Remark

The gross purchase price is also sometimes named "catalogue price".

D7. The "cost price P_R " at time t is the gross sales price (or purchase price depending from the point of view...) minus various possible or obligatory deductions D (being a negative value) made by the seller such that:

$$P_R(t) \cong P_{VB}(t) + D(t) \tag{15.3.7}$$

The agents of goods market exchange sometimes accept a discount on the catalogue price. Reductions exist in two main know forms:

1. The "remission R" which is a price bonification (negative value) given to an agent who buys goods in large quantities N or to a retailer which is charged a branded item sale price imposed by the manufacturer (strategic factor business, promotion, etc.). The discount depends on the time t and the amount n bought or ordered.
2. The "prompt payment discount E" which is a deduction at time t granted to the buyer agent for cash payment or prepayment or for payment for an agreed period (we will formally come back on this topic later below during our study of simple interest in Actuarial Calculation).

In the most general case we talk at a given time t of "exercise price" (or "invoice price") to which the good can be bought or sold as:

$$\boxed{P_E(t)=R(t,N)+E(t)+n\left\{F_{SA}(t)\left[\sum C_D(t)+\sum C_I(t)+B_b(t)+F_G(t)+D(t)\right]\right\}} \tag{15.3.8}$$

Remark

All the terms of these relations usually take their values in $\mathbb{R}$...

The factors to be considered when developing a pricing policy are synthesized without being exhaustive in the following diagram:

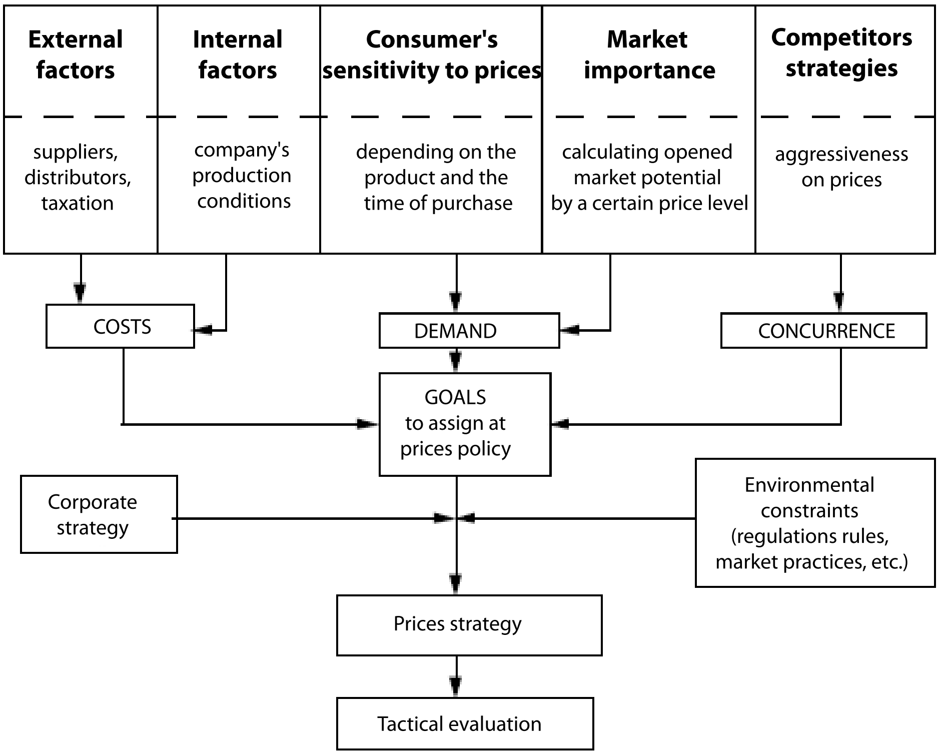

External factors
suppliers, distributors, taxation
Internal factors
company's production conditions
Consumer's sensitivity to prices
depending on the product and the time of purchase
Market importance
calculating opened market potential by a certain price level
Competitors strategies
aggressiveness on prices
COSTS
DEMAND
CONCURRENCE
GOALS to assign at prices policy
Corporate strategy
Environmental constraints (regulations rules, market practices, etc.)
Prices strategy
Tactical evaluation

Figure 15.78: **Simplified schematic summary to develop a pricing policy**

Definitions (#15.3.654):

D1. The "propensity to consume P_C" equation is the part of the income R of an agent that is devoted to consumption of an amount C (primary and secondary consumption):

$$P_C=\frac{C}{R} \tag{15.3.9}$$

D2. The difference between consumption expenditure and the income is defined as "savings E" while contributions and revenues on services represent "social transfers":

$$E=R-C \tag{15.3.10}$$

D3. The "income elasticity" is the ratio of the change in consumption on the change in income:

$$E_r=\frac{\Delta C}{\Delta R} \tag{15.3.11}$$

The concept of income elasticity propose to classify goods in supplementary categories:

1. "Inferior goods" are consumer goods whose income elasticity is negative and therefore whose consumption decreases with increasing income as $E_r<0$ (bread, flour, etc.).
2. "Superior goods" which are luxury consumer goods whose income elasticity is positive and therefore of which the consumption increases with an increase in income as $E_r>0$ (health, leisure, etc.).
3. "Normal goods": which are goods that are neutral and with a coefficient of elasticity with respect to income is a little different from 0 such that $E_r\cong 0$.

D4. The "price elasticity" is the ratio of the change in the amount of demand $D(P)$ for a good on the change in its price P, both assumed non-negative, continuous and differentiable, and is given by:

$$E_r=\frac{\Delta D}{\Delta P} \tag{15.3.12}$$

or at the limit, written:

$$E_r=\frac{\mathrm{d}D}{\mathrm{d}P}=D'(P) \tag{15.3.13}$$

and "arc-elasticity" is defined by:

$$e_D=-\frac{\%D}{\%P}=-\frac{\Delta D/\bar{D}}{\Delta P/\bar{P}} \tag{15.3.14}$$

or at the limit, written:

$$e_D=\frac{\mathrm{d}D/D}{\mathrm{d}P/P}=\frac{P\mathrm{d}D}{D\mathrm{d}P}=D'(P)\frac{P}{D} \tag{15.3.15}$$

The arc-elasticity gives the value of elasticity at the midpoint over a range of change.

Remarks

R1. A demand is named "price sensitive" when the percentage change in quantity demanded is greater than the percentage change in price. If we talk about "rigid price" demand.

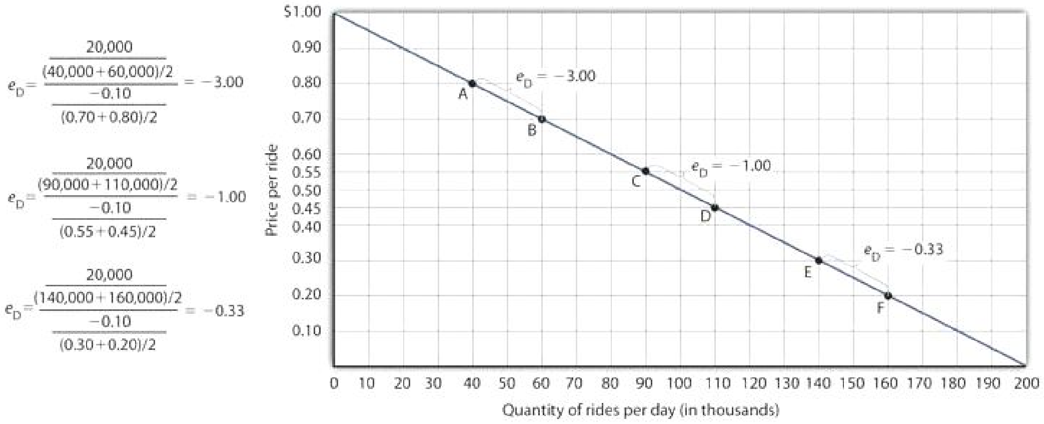
R2. Notice (useful for later) that as:

$$\frac{\mathrm{d}\ln(D)}{\mathrm{d}D}=\frac{1}{D}\Leftrightarrow\frac{\mathrm{d}D}{D}=\ln(D)\qquad\text{and}\qquad\frac{\mathrm{d}\ln(P)}{\mathrm{d}P}=\frac{1}{P}\Leftrightarrow\frac{\mathrm{d}P}{P}=\ln(P)\tag{15.3.16}$$

Therefore:

$$e_D=-\frac{\mathrm{d}D/D}{\mathrm{d}P/P}=\frac{\ln(D)}{\ln(P)}\tag{15.3.17}$$

What happens to price elasticity of demand when we travel along a demand curve depends on the nature of the demand curve itself. On a linear demand curve, such as the one in the figure below elasticity becomes smaller (in absolute value) as we travel downward and to the right:



D5. An "investment I" is the operation performed by an economic agent whose objective is to obtain produced goods in exchange.

D6. A "transaction T" is the exchange of a quantity of goods at a specified price between a "seller" and "buyer". It concludes on the market whose form is determined by the number of economic agents who are involved which determines the "concurrence".

The table below shows the different types of contracts:

	Suppliers		
Seekers	multitude	some	only one
multitude	perfect competition	oligopoly	monopoly
some	oligopsony	bilateral oligopoly	upset monopoly
only one	monopsony	upset monopsony	bilateral monopoly

Table 15.32: **Different forms of market competition**

When the actors of an ogilopoly work together to fix prices, we speak then of "cartel".

✍**Examples:**

E1. Lysine, a $600 million-a-year industry, is an amino acid used by farmers as a feed additive to ensure the proper growth of swine and poultry. The primary U.S. producer of lysine is Archer Daniels Midland (ADM), but several other large European and Japanese firms are also in this market. For a time in the first half of the 11990s (holocene calendar), the world's major lysine producers met together in hotel conference rooms and decided exactly how much each firm would sell and what it would charge. The U.S. Federal Bureau of Investigation (FBI), however, had learned of the cartel and placed wire taps on a number of their phone calls and meetings. From FBI surveillance tapes, following is a comment that Terry Wilson, president of the corn processing division at ADM, made to the other lysine producers at a 11994 (holocene calendar) meeting in Mona, Hawaii:

«I wanna go back and I wanna say something very simple. If we're going to trust each other, okay, and if I'm assured that I'm gonna get 67,000 tons by the year's end, we're gonna sell it at the prices we agreed to... The only thing we need to talk about there because we are gonna get manipulated by these [expletive] buyers-they can be smarter than us if we let them be smarter. ...They [the customers] are not your friend. They are not my friend. And we gotta have 'em, but they are not my friends. You are my friend. I wanna be closer to you than I am to any customer. Cause you can make us ... money. ... And all I wanna tell you again is let's - let's put the prices on the board. Let's all agree that's what we're gonna do and then walk out of here and do it.»

The price of lysine doubled while the cartel was in effect. Confronted by the FBI tapes, Archer Daniels Midland pled guilty in 11996 (holocene calendar) and paid a fine of $100 million. A number of top executives, both at ADM and other firms, later paid fines of up to $350, 000 and were sentenced to 24-30 months in prison.

In another one of the FBI recordings, the president of ADM told stated the slogan of its company this way: «*Our competitors are our friends. Our customers are the enemy*».

E2. French antitrust authorities in 12014 (holocene calendar) issued two fines of a cumulative amount of some 951 million euros against 13 of the main manufacturers of the maintenance sector, hygiene and beauty for having coordinated their marketing policy with the mass retailers and to have agreed on the price increases between 12003 and 12006 (holocene calendar). The first fine of concerns Colgate-Palmolive, Henkel, Unilever, Procter & Gamble, the second one concerns the same + Gillette and L'Oréal.

This case was revealed 3 years after another important decision of the Authority on the laundry sector, already containing some of the same companies. The four major manufacturers of laundry detergents - Unilever, Procter & Gamble, Colgate-

Palmolive and Henkel - secretly agreed during 6 years on their prices and promotions in France, resulting in a 4% surcharge of 6% for consumers. These manufacturers were then sentenced to pay only 361 million euros in fines...

A competition is qualified of "pure competition" (P.P.C.: pure and perfect competition) or "perfect competition" if it meets the next five following hypothesis:

H1. Atomicity: Buyers and sellers are numerous to the point that no one alone can influence the prices.

H2. Homogeneity: The goods traded are identical and interchangeable with each other. They can satisfy the same need.

H3. Free entry: There is no barrier to entry and exit of new economic agents.

H4. Free movement: Economic agents can move freely (from a market to the other one).

H5. Perfect information: Everyone knows at the same time and for free all quantities supplied and demanded by all the agents at different prices (see section Game and Decision Theory page 7400).

D7. The "intermediate management balances" are parts of the overall result of market activity period that are significant to the financial analyst. There are multiple definitions which arise from elementary algebraic operations on concepts defined above as (non exhaustive list!):

- The "gross margin" or "gross profit" is the difference between the proceeds of sales of goods and the purchase cost of goods sold (the sales margin is specific to trading activities, that is to say to companies with a distribution activity).
- The "operating income" is the difference between the gross margin less the operating expenses.
- The "net income" (also named "earning before interest and taxes" abbreviated EBIT[14]) is the difference between the operating income and other expenses (interest expenses) and taxes.
- The "production of the fiscal period" which is the sum of production sold, stored and immobilized (production of exercise is specific to production activities, that is to say companies with an industrial activity).
- The "turnover", which is the sum of revenue from sales of goods and sales of goods and services.
- The "added value" which is defined as the difference between production for the fiscal period and the intermediate consumption by the economic agents (the manager considers it created wealth resulting from the real business of the company and the added value also of national importance because it is an aggregate).
- The "degree of operating leverage" (DOL) is a numerical measure that helps an analyst compare the situation of two firms or of one firm over a period of time. When a proportionate change in EBIT, due to a change in sales, is more than the proportionate change in sales, operating leverage exists. Degree of Operating Leverage is the

---

[14] There is also the EBITDA that stands for the earnings before interest, depreciation, taxes, and amortization...

magnification factor by which % change in EBIT exceeded the % change in sales. It is defined by:

$$\text{DOL} = \frac{\text{\% change in Operating Income}}{\text{\% Change in Sales}} \tag{15.3.18}$$

- The "gross operating surplus" (G.O.S.) is the result of the current activity of the company and is defined as:

$$\begin{aligned}\text{G.O.S.} =&\text{Added Value} + \text{Amortization \& Provisions}+\\ &\text{Other Products} - \text{Other Expenses} + \text{Charges transfer}\end{aligned} \tag{15.3.19}$$

✍**Examples:**

E1. Consider the following non-standardized and naive income statement:

Imaginary Corp & Co.	
Income statement	
Sales	\$105,000
Less: Cost of Sales	\$20,000
Gross Profit	\$85,000
Less: Operating Expenses	\$15,000
Operating Income	\$70,000
Less: Other Expenses and Taxes	\$20,000
Net Income	\$50,000

E2. To introduce the DOL (degree of operating leverage), let us consider the two following income statements (the years are given based on the holocene calendar):

Imaginary Corp & Co.		
Income Statement	**11998**	**11999**
Sales	\$85,000	\$105,000
Cost of Operations	\$45,000	\$35,000
Operating Income	\$40,000	\$70,000

The percentage change in sales and in operating income is:

$$\begin{aligned}\text{\% change in sales} &= \frac{105,000-85,000}{85,000} \cong 23.5\%\\ \text{\% change in operating income} &= \frac{70,000-40,000}{40,000} \cong 75\%\end{aligned} \tag{15.3.20}$$

We can see that a 23.5% increase in sales results in a 75% increase in operating income. Using the definition above, the Degree Of Operating Leverage is:

$$\text{DOL} = \frac{\text{\% change in Operating Income}}{\text{\% Change in Sales}} = \frac{75\%}{23.5\%} = 3.19 \tag{15.3.21}$$

15.3.1.1.1 Willingness-To-Pay

The "willingness-to-pay" (W.T.P.) approach assumes that each potential customer has a maximum willingness to pay (sometimes named a "reservation price") for a product or service. A customer will purchase if and only if the price is less than her maximum willingness to pay (we will use willingness to pay, sometimes to mean "maximum willingness to pay"). For example, a customer with a willingness to pay of \$ 253 for an airline ticket from New York to Miami will purchase the ticket if the price is less than or equal to \$ 253 but not if it is \$ 253.01 or more. In this case, $D(253)$ equals the number of customers whose maximum willingness to pay is at least \$ 253. A customer with a maximum willingness to pay of \$ 0 (or less) will not buy at any price.

Define the function $w(x)$ as the W.T.P distribution across the population. Then, for any values $0 \leq P_1 \leq P_2$:

$$\int\limits_{P_1}^{P_2} w(x)\mathrm{d}x = \text{fraction of the population that has W.T.P. between } P_1 \text{ and } P_2 \tag{15.3.22}$$

We note that $0 \leq w(x) \leq 1$ for all non-negative values of x, therefore it is quite obvious that $w(x)$ is just a probability density function! Let:

$$D_{\max} = D(0) \tag{15.3.23}$$

be the maximum demand achievable. Then we can derive $D(P)$ from the W.T.P distribution from:

$$\boxed{D(P) = D_{\max} \int\limits_{P}^{+\infty} w(x)\mathrm{d}x} \tag{15.3.24}$$

Denoting the primitive of $w(x)$ by $W(x)$ that latter can be written:

$$D(P) = D_{\max}\left(W(+\infty) - W(P)\right) \tag{15.3.25}$$

Assuming:

$$w(+\infty) \rightarrow 0 \tag{15.3.26}$$

We can take the derivative of the corresponding price-response function to obtain:

$$D'(P) = D_{\max}\left(W(+\infty) - W(P)\right)' = D_{\max}\left(w(+\infty) - w(P)\right)' - D_{\max}w(P) \tag{15.3.27}$$

which is non-positive, as required by the downward-sloping demand curve property. Conversely, we can derive the willingness-to-pay distribution from the price-response function using:

$$w(P) = -\frac{D'(P)}{D_{\max}} = -\frac{D'(P)}{D(0)} \tag{15.3.28}$$

We will see further below what is the purpose to build such a W.T.P. function!

15.3.1.1.2 Pricing optimization

The scope of "pricing and revenue optimization" (PRO) is to set and update the prices for each combination of product, customer segment and channel. Due to market condition change over time, the PRO is responsible to update the prices. The basic element of a PRO is the price-response function or the price-response curve given by $D(P)$.

We are going here to compare three different types of demand models for homogeneous products and how to find optimal prices for each one of them.

For the linear model, demand is given by:

$$D(P)=\alpha P+\beta \tag{15.3.29}$$

and as we have seen earlier, the arc elasticity goes for infinity to zero in absolute value.

Remark

It is easy to verify that for this model:

$$D(0)=\alpha\cdot 0+\beta=\beta \tag{15.3.30}$$

And therefore:

$$w(x)=-\frac{D'(P)}{D(0)}=-\frac{\alpha}{\beta}=c^{te} \tag{15.3.31}$$

ie a uniform distribution!

Another very common demand model is the "constant arc-elasticity model" (CAEM), defined by:

$$\ln\left(D(P)\right)=\alpha\ln(P)+\beta \tag{15.3.32}$$

or if we take the exponential:

$$D(P)=e^{\beta}P^{\alpha}=C^{te}P^{\alpha} \tag{15.3.33}$$

Proof 15.2.2. Let us prove that the arc-elasticity is indeed constant for the specific model:

$$\begin{aligned}e_D&=\frac{\mathrm{d}D/D}{\mathrm{d}P/P}=\frac{\mathrm{d}D}{\mathrm{d}P}\frac{P}{D}=\left[\frac{\mathrm{d}}{\mathrm{d}P}\left(C^{te}P^{\alpha}\right)\right]\frac{P}{D}\\&=C^{te}\alpha P^{\alpha-1}\frac{P}{D}=C^{te}\alpha P^{\alpha-1}\frac{P}{C^{te}P^{\alpha}}=C^{te}P^{\alpha-1}\frac{P}{P^{\alpha}}\\&=C^{te}P^{\alpha-1}PP^{-\alpha}=C^{te}P^{\alpha-1}P^{-\alpha+1}=C^{te}\end{aligned} \tag{15.3.34}$$

As e_D should be positive or null, this assumes that $C^{te}\leq 0$ (once again this will make sense a bit later).

■ Q.E.D.

Remark

It is easy to verify that for this model:

$$D(0)=C^{te}\cdot 0^{\alpha}=0 \tag{15.3.35}$$

And therefore:

$$w(x)=-\frac{D'(P)}{D(0)}=+\infty \tag{15.3.36}$$

So this is a quite unrealistic probability density function...

A much more interesting demand curve is given by the logistic/sigmoid function (i.e. the logit price-response function):

$$D(P)=C^{te}\frac{e^{\alpha P+\beta}}{1+e^{\alpha P+\beta}}=\frac{C^{te}}{1+e^{-\alpha P-\beta}}=\frac{C^{te}}{1+e^{-\alpha(P+\beta/\alpha)}}=\frac{C^{te}}{1+e^{-\alpha(P-P_0)}} \tag{15.3.37}$$

Where:

$$P_0=-\frac{\beta}{\alpha} \tag{15.3.38}$$

is the inflection point of the demand and C^{te} is a scale factor and α measures price sensitivity.

Proof 15.2.3. Let us prove the value of the inflection point. For this we determine the first derivative before:

$$D'(P)=\left(C^{te}\left(1+e^{-\alpha P-\beta}\right)^{-1}\right)'=-C^{te}\left(1+e^{-\alpha P-\beta}\right)^{-2}\left(-\alpha e^{-\alpha P-\beta}\right) \tag{15.3.39}$$

Then the second derivative is then quite immediate:

$$D''(P)=2C^{te}\left(1+e^{-\alpha P-\beta}\right)^{-3}\left(-\alpha e^{-\alpha P-\beta}\right)^{2}-C^{te}\left(1+e^{-\alpha P-\beta}\right)^{-2}\left(\alpha^{2}e^{-\alpha P-\beta}\right) \tag{15.3.40}$$

to find the inflection point we therefore only have to solve (see section Differential and Integral Calculus page 1454):

$$2C^{te}\left(1+e^{-\alpha P-\beta}\right)^{-3}\left(-\alpha e^{-\alpha P-\beta}\right)^{2}-C^{te}\left(1+e^{-\alpha P-\beta}\right)^{-2}\left(\alpha^{2}e^{-\alpha P-\beta}\right)=0 \tag{15.3.41}$$

Very simple algebraic simplification then gives (abscissa value of the inflection point):

$$P=-\frac{\beta}{\alpha} \tag{15.3.42}$$

Hence the ordinate is given by:

$$D\left(-\frac{\beta}{\alpha}\right)=\frac{C^{te}}{1+e^{-\alpha(-\beta/\alpha)-\beta}}=\frac{C^{te}}{1+e^{0}}=\frac{C^{te}}{2} \tag{15.3.43}$$

■ Q.E.D.

Remark

It is easy to verify that for this model:

$$D(0)=\frac{C^{te}e^{\beta}}{1+e^{\beta}}=\frac{C^{te}}{1+e^{-\beta}} \tag{15.3.44}$$

And therefore:

$$\begin{aligned}w(x)&=-\frac{D'(P)}{D(0)}=-\frac{-C^{te}\left(1+e^{-\alpha P-\beta}\right)^{-2}\left(-\alpha e^{-\alpha P-\beta}\right)}{\dfrac{C^{te}}{1+e^{-\beta}}}=\frac{C^{te}\left(1+e^{-\alpha P-\beta}\right)^{-2}\left(-\alpha e^{-\alpha P-\beta}\right)}{\dfrac{C^{te}}{1+e^{-\beta}}}\\&=\frac{-\alpha e^{-\alpha P-\beta}\left(1+e^{-\beta}\right)}{\left(1+e^{-\alpha P-\beta}\right)^2}=K\frac{e^{-\alpha P-\beta}}{\left(1+e^{-\alpha P-\beta}\right)^2}\end{aligned} \tag{15.3.45}$$

This is a far more realistic W.T.P. distribution than those associated with either the constant-elasticity or linear price-response functions. For that reason, the logit is usually preferred to linear or constant-elasticity price-response functions.

It is important to notice that the price sensitivity α is almost always negative (excepted for luxury on limited edition products and some other obvious special cases) for all these three models:

$$\begin{aligned}D(P)&=\alpha P+\beta\\D(P)&=e^{\beta}P^{\alpha}\\D(P)&=\frac{C^{te}}{1+e^{-\alpha(P-P_0)}}\end{aligned} \tag{15.3.46}$$

In a quite general setting, one have for the total profit function:

$$L(P)=D(P)(P-C) \tag{15.3.47}$$

where L gives the profit, D is still the demand function that depends of the price and C is the marginal cost. Taking the derivative with respect to price we have:

$$\begin{aligned}L'(P)&=(D(P)(P-C))'=D'(P)(P-C)+D(P)(P-C)'=D'(P)(P-C)+D(P)P'\\&=D'(P)(P-C)+D(P)\end{aligned} \tag{15.3.48}$$

Making $L'(P)=0$ to calculate the optimum price P^* (first order condition), we have:

$$D'(P^*)(P^*-C)+D(P^*)=0 \tag{15.3.49}$$

Hence:

$$D'(P^*)P^*+D(P^*)=D'(P^*)C \tag{15.3.50}$$

Now for the linear model introduced just earlier above, we have obviously:

$$D'(P^*_{\text{lin}})=\alpha \tag{15.3.51}$$

Hence:

$$D'(P_{\text{lin}}^*)P_{\text{lin}}^* + D(P_{\text{lin}}^*) = D'(P_{\text{lin}}^*)C \rightarrow \alpha P + (\alpha P_{\text{lin}}^* + \beta) = \alpha C \tag{15.3.52}$$

Therefore the optimal price is given by:

$$P_{\text{lin}}^* = \frac{\alpha C - \beta}{2\alpha} \tag{15.3.53}$$

To deal now with the constant arc-elasticity model, let us first recall that:

$$D'(P^*)(P^* - C) + D(P^*) = 0 \tag{15.3.54}$$

Hence:

$$-\frac{D'(P^*)}{D(P^*)} = \frac{1}{P^* - C} \tag{15.3.55}$$

Multiplying both sides by P^*, we get:

$$-\frac{D'(P^*)P^*}{D(P^*)} = \frac{P^*}{P^* - C} \tag{15.3.56}$$

And according to our definition of arc-elasticity:

$$e_D(P^*) = \frac{P^*}{P^* - C} \tag{15.3.57}$$

Hence after rearranging:

$$P^* = \frac{e_D(P^*)C}{1 - e_D(P^*)} = \frac{C}{1 - 1/e_D(P^*)} \tag{15.3.58}$$

And as:

$$e_D = \frac{\mathrm{d}D/D}{\mathrm{d}P/P} = \frac{\ln(D)}{\ln(P)} = \frac{\alpha \ln(P) + \beta}{\ln(P)} = \alpha \tag{15.3.59}$$

Therefore:

$$P_{\text{CAEM}}^* = \frac{C}{1 - 1/e_D(P^*)} = \frac{C}{1 - 1/\alpha} \tag{15.3.60}$$

It is interesting to note that one needs $|\alpha| > 1$, otherwise the profit function will be convex with respect to price and the optimal price will be $+\infty$. If one have a monopolistic market, normally this assumption holds.

For the logistic model. we have for recall:

$$D'(P) = C^{te}\left(1 + e^{-\alpha P - \beta}\right)^{-2}\left(\alpha e^{-\alpha P - \beta}\right) \tag{15.3.61}$$

And if we put it into:

$$D'(P^*)(P^* - C) + D(P^*) = 0 \tag{15.3.62}$$

We get:

$$C^{te}\left(1+e^{-\alpha P^*-\beta}\right)^{-2}\left(\alpha e^{-\alpha P^*-\beta}\right)(P^*-C)+\frac{C^{te}}{1+e^{-\alpha P^*-\beta}}=0 \qquad (15.3.63)$$

Simplifying a bit we get:

$$\frac{C^{te}}{1+e^{-\alpha P^*-\beta}}\left(\alpha e^{-\alpha P^*-\beta}(P^*-C)+1\right)=0 \qquad (15.3.64)$$

With a further simple simplification:

$$\alpha e^{-\alpha P^*-\beta}(P^*-C)+1=0 \qquad (15.3.65)$$

Since the last equation above does not seems to have an analytical solution, one can easily find the result with a Newton-step algorithm or minimization problem (see section of Numerical Methods page 6662):

$$\min_{P^*\in\mathbb{R}}\left(\alpha e^{-\alpha P^*-\beta}(P^*-C)+1\right)^2 \qquad (15.3.66)$$

Remark

To see how to implement this on a computer software the reader can refer to the `R` companion book.

15.3.1.1.3 Average & Marginal Cost/Revenue

Let us introduce the subject directly by an example:

Suppose an amateur cook (and economist) invites his friends to his table and proposes to do their a tomato salad. It evaluates the work he has to do this job and figure it in monetary value. It considers that a minute of his time spent in the preparation of the salad corresponds to an expense of 1.-.

So the data are:

1. Each minute of work value is 1.-
2. Each tomato is 1.-/unit
3. Prepare the salad takes 15 minutes

If each of his friends is satisfied with a single tomato, prepare dinner for five friends (the cook does not eat) will cost in total:

$$C_T = 5.- + 15.- = 20.- \tag{15.3.67}$$

Mathematically, the "total cost function" will looks like following if we write q as the number of friends (in practice it is often a strictly increasing polynomial function):

$$C_T(q) = (1.-)\, q + 15.- \tag{15.3.68}$$

Definition (#15.3.655): The "average cost" or "average cost function" for each guest is of 20.- divided by 5 thus 4.-. This corresponds to:

$$\boxed{C_m(q) = \frac{C_T(q)}{q}} \tag{15.3.69}$$

which is not necessarily an increasing function!

Theorem 15.3. We can see that as q tends to infinity, under the assumption of constant production effort, the average cost C_m tends to the unit cost C_U. That is so to say:

$$\lim_{q \to +\infty} C_m(q) = C_U \tag{15.3.70}$$

Proof 15.3.1. Indeed, if we denote C_F the assumed fixed cost and C_U the unit cost, we have:

$$C_m(q) = \frac{C_T(q)}{q} = \frac{C_U q + C_F}{q} \tag{15.3.71}$$

and when $q \to +\infty$ then as $C_U q \gg C_F$ we can write:

$$C_m(q) = \frac{C_U q + C_F}{q} \cong \frac{C_U q}{q} = C_U \tag{15.3.72}$$

■ Q.E.D.

Remark

According to the books (and authors), the average cost is denoted by a lowercase m or uppercase M and the marginal cost (see further below) too.

If our cook invites a sixth friends, the total cost will be 21.-. In fact the preparation time will remain, at least we assume ... constant! In this case, the marginal cost of the sixth guest is:

$$(21.\text{-}) - (20.\text{-}) = 1.\text{-} \qquad (15.3.73)$$

while the average cost for all guests is now:

$$\frac{21.\text{-}}{6} = 3.5.\text{-} \qquad (15.3.74)$$

We notice in this situation that the average cost goes down of 0.5.- indefinitely as the number of guests increases. This quantity of 0.5.- is the marginal revenue!

This example illustrates the return scale and shows that we often have an incentive to increase production to reduce the average cost of production (but this does not mean that our amateur cook will appreciate the reasoning when reaching its physical limit...).

Moreover, it this not however a general rule! While the bowl of our economist can contain only 6 tomatoes, the 7th guest will force him to prepare a second bowl. In this case, the marginal variation will be greater than the previous one.

We must also be aware that the total cost function is in fact rarely a continuous function (because you have to buy machinery or engage staff in stepwise quantities) and even more it is the market that dictates the quantities to a factory and not the inverse.

Definition (#15.3.656): Mathematically, the "marginal cost" is defined by the variation of the total cost function $C_T(q)$, relative to the quantity produced q:

$$C_M = \frac{\Delta C_T(q)}{\Delta q} \qquad (15.3.75)$$

or if the total cost function is differentiable (anyway in economics, even if it is not continuous we makes abstraction of this fact...):

$$\boxed{C_M = \frac{\mathrm{d}C_T(q)}{\mathrm{d}q}} \qquad (15.3.76)$$

The marginal cost corresponds thus to the cost of producing one more or less unit (in the example above it is equal to 1.-). In practice, we are more interested in the cost of an additional series.

In addition, the reader will notice that as the total cost is a strictly increasing function, the marginal cost is also a constant or increasing function.

Remark

If the marginal cost increases as the quantity q increases, we say that the returns are diminishing. In contrast, they are increasing if the marginal cost is decreasing as the amount increases.

The marginal cost corresponds thus to the cost of producing one more or less unit. In practice, we are more interested in the cost of an additional series.

In addition, the reader will notice that as the total cost is a strictly increasing function, the marginal cost is also a constant or increasing function.

Theorem 15.4. Let us show now that if the average cost goes through an extremal value, the marginal cost is equal to it at that point. We name this particular situation a "technical optimum".

Recall previously that if a continuous and differentiable function $f(x)$ has a minimum (or maximum), its derivative at that point is equal to zero (see section Differential and Integral Calculus page 1453).

Proof 15.4.1. Apply this property to the average cost function. But recall that previously (see section Differential and Integral Calculus page 1475):

$$\left(\frac{u}{v}\right)'=\frac{u'v-uv'}{v^2} \tag{15.3.77}$$

So we have if the derivative of the average cost is zero:

$$\frac{\mathrm{d}C_m}{\mathrm{d}q}=0\Rightarrow\frac{\mathrm{d}}{\mathrm{d}q}\left(\frac{C_T}{q}\right)=\left(\frac{C_T}{q}\right)'=0 \tag{15.3.78}$$

So this is the derivative of the ratio of two functions. therefore:

$$\left(\frac{C_T}{q}\right)'=\frac{\dfrac{\mathrm{d}C_T}{\mathrm{d}q}q-C_T\dfrac{\mathrm{d}q}{\mathrm{d}q}}{q^2}=0\Rightarrow\frac{\mathrm{d}C_T}{\mathrm{d}q}q-C_T=C_T'q-C_T=0 \tag{15.3.79}$$

Hence we deduce:

$$C_T'q=C_T \tag{15.3.80}$$

Thus:

$$C_T'=\frac{C_T}{q} \tag{15.3.81}$$

■ Q.E.D.

This result tell us that in this situation, the marginal cost is equal to average cost (we find the definition of each cost on the to left or on the right respectively of equality).

In other words, where the average cost reaches an optimum (minimum in our case), then the marginal cost equals average cost.

Definition (#15.3.657): Mathematically, the "marginal revenue" is defined by the additional margin relative to the quantity produced q (it is the additional revenue that will be generated by increasing product sales by one unit):

$$\boxed{C_R = \Delta C_m(q) = \frac{C_T(q)}{q} - \frac{C_T(q-1)}{q-1}} \tag{15.3.82}$$

Theorem 15.5. As $q \rightarrow +\infty$ the marginal revenue C_R tends to zero. In other words, more we produce a given product the smaller is the margin variation.

Proof 15.5.1. As we have:

$$C_R = \frac{C_T(q)}{q} - \frac{C_T(q-1)}{q-1} = \frac{C_F + qC_U}{q} - \frac{C_F + (q-1)C_U}{q-1} \tag{15.3.83}$$

Then when $q \rightarrow +\infty$ we have:

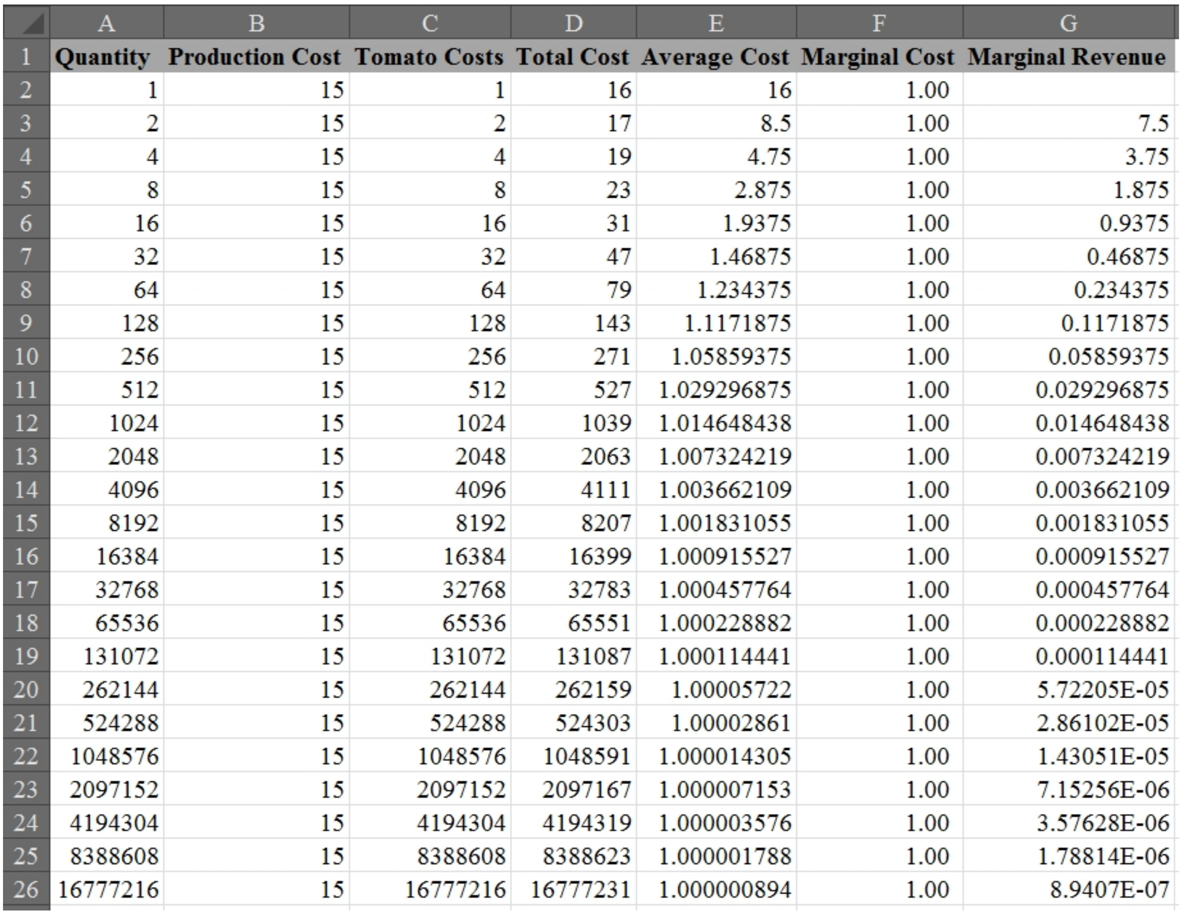
$$q \cong q-1 \tag{15.3.84}$$

and therefore:

$$C_R = \frac{C_F + qC_U}{q} - \frac{C_F + (q-1)C_U}{(q-1)} \cong \frac{C_F + qC_U}{q} - \frac{C_F + qC_U}{q} = 0 \tag{15.3.85}$$

■ Q.E.D.

For our above example the tabulated values give in a spreadsheet software:



	A	B	C	D	E	F	G
1	Quantity	Production Cost	Tomato Costs	Total Cost	Average Cost	Marginal Cost	Marginal Revenue
2	1	15	1	16	16	1.00	
3	2	15	2	17	8.5	1.00	7.5
4	4	15	4	19	4.75	1.00	3.75
5	8	15	8	23	2.875	1.00	1.875
6	16	15	16	31	1.9375	1.00	0.9375
7	32	15	32	47	1.46875	1.00	0.46875
8	64	15	64	79	1.234375	1.00	0.234375
9	128	15	128	143	1.1171875	1.00	0.1171875
10	256	15	256	271	1.05859375	1.00	0.05859375
11	512	15	512	527	1.029296875	1.00	0.029296875
12	1024	15	1024	1039	1.014648438	1.00	0.014648438
13	2048	15	2048	2063	1.007324219	1.00	0.007324219
14	4096	15	4096	4111	1.003662109	1.00	0.003662109
15	8192	15	8192	8207	1.001831055	1.00	0.001831055
16	16384	15	16384	16399	1.000915527	1.00	0.000915527
17	32768	15	32768	32783	1.000457764	1.00	0.000457764
18	65536	15	65536	65551	1.000228882	1.00	0.000228882
19	131072	15	131072	131087	1.000114441	1.00	0.000114441
20	262144	15	262144	262159	1.00005722	1.00	5.72205E-05
21	524288	15	524288	524303	1.00002861	1.00	2.86102E-05
22	1048576	15	1048576	1048591	1.000014305	1.00	1.43051E-05
23	2097152	15	2097152	2097167	1.000007153	1.00	7.15256E-06
24	4194304	15	4194304	4194319	1.000003576	1.00	3.57628E-06
25	8388608	15	8388608	8388623	1.000001788	1.00	1.78814E-06
26	16777216	15	16777216	16777231	1.000000894	1.00	8.9407E-07

✍**Example:**

A factory has calculated that the cost of producing q units of a product can be modelled by the following function within the yield capacity limits of a machine and manpower it has at its disposal:

$$C_T(q)=2,600+2q+0.001q^2 \tag{15.3.86}$$

Let us calculate the cost, the average cost and marginal cost of the production of $1,000$, $2,000$ and $3,000$ units. We then have the following information:

$$\begin{aligned}C_T(q)&=2,600+2q+0.001q^2\\C_m(q)&=\frac{C_T(q)}{q}=\frac{2,600}{q}+2+0.001q\\C_M(q)&=\frac{\mathrm{d}C_T(q)}{\mathrm{d}q}=2+0.002q\end{aligned} \tag{15.3.87}$$

Then we have:

Quantity	**Total Cost**	**Average Cost**	**Marginal Cost**
$1,000$	$5,600$.-	5.60.-	4.00.-
$2,000$	$10,600$.-	5.30.-	6.00.-
$3,000$	$17,600$.-	5.87.-	8.00.-

Table 15.33: **Various costs**

And for the average cost to be as low as possible, it is necessary that the marginal cost is equal to average cost:

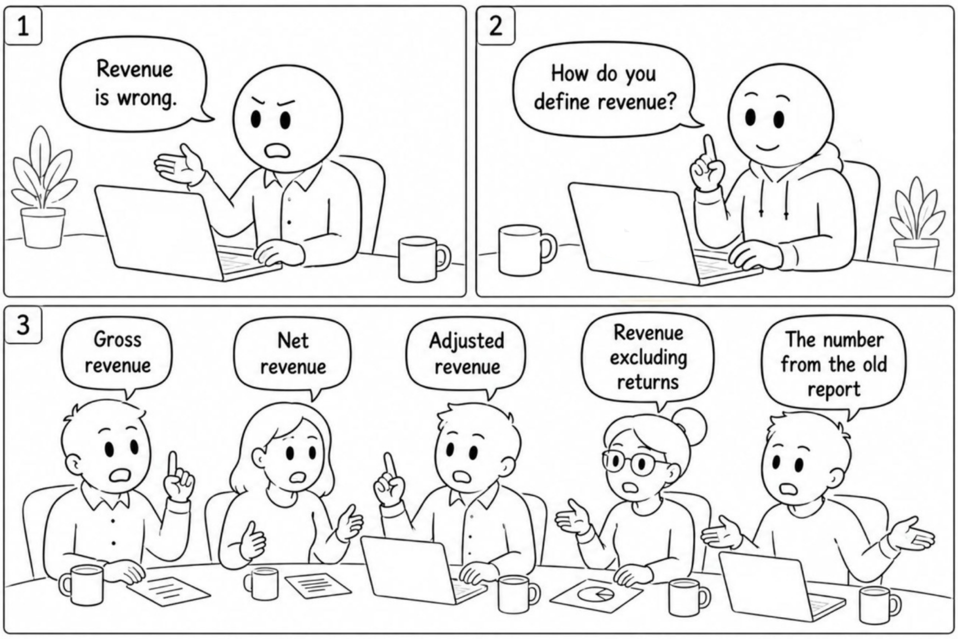
$$\frac{\mathrm{d}C_T(q)}{\mathrm{d}q}=\frac{C_T(q)}{q}\Rightarrow\frac{2,600}{q}+2+0.001q=2+0.001q \tag{15.3.88}$$

We conclude:

$$q=\sqrt{\frac{2,600}{0.001}}\cong 1,612 \tag{15.3.89}$$

By injecting this value in the total cost function, we deduce that the minimum average cost is 5.22.-.

In reality, we make Monte Carlo simulations that gives not only the possibility to work in an uncertain and quantities, but that an also take into account non-continuous functions! For more information on Monte Carlo modelling, the reader can refer to the section of Numerical Methods at page 6708.


1
Revenue is wrong.
2
How do you define revenue?
3
Gross revenue
Net revenue
Adjusted revenue
Revenue excluding returns
The number from the old report

15.3.1.2 Macroeconomics

"Aggregates" are synthetic variables developed by nations for their national accounting and measuring the result of their entire economy. Some of the main aggregates are defined by:

Definitions (#15.3.658):

D1. The "gross national product" (G.N.P.) whose purpose is to measure domestic production (considered as isolated), that is to say all the monetary values of goods and services produced in a given period (the term "Gross" indicates that the value of G.N.P. is not deducted from different existing taxes on production).

D2. The "gross domestic product" (G.D.P.) whose role is to measure national production (such as G.N.P.) and take into account the income of the rest of the world. In other words, G.D.P. is the G.N.P. which we sum the capital coming from outside and which is subtracted from capital going outside.

D3. The "gross fixed capital formation" (G.F.C.F.) measures the value of acquisitions of new or existing fixed assets by the business sector, governments and "pure" households (excluding their unincorporated enterprises) less disposals of fixed assets and it is a component of the expenditure on gross domestic product (G.D.P.), and thus shows something about how much of the new value added in the economy is invested rather than consumed.

D4. The "national income" (N.I.) whose role is to measure all the incomes received by economic agents.

D5. The "consumption" whose role is to represent the value of goods and services used for the direct needs of economic agents.

D6. The "consumer price index" (C.P.I.) measures changes in prices of goods and services representative of the consumption of private households. It indicates how much consumers have to increase or decrease their expenses to maintain the same volume of consumption.

D7. The "inflation" (I) is defined as the increase in prices or the purchasing power of money the most common way to calculate the inflation rate is by recording the prices of goods and services over the years, take a base year and then determine the percentage rate changes of those prices over the years. For example with C.P.I. we get:

$$\mathrm{I}=\frac{\mathrm{CPI}_y-\mathrm{CPI}_{y-1}}{\mathrm{CPI}_{y-1}} \tag{15.3.90}$$

D8. The "purchasing power standard" (P.P.S.) is an artificial common reference currency unit used in the European Union which eliminates the differences of price levels between countries. So, a P.P.S. allows to buy the same volume of goods and services in all the countries. This unit allows significant comparisons in volume of economic indicators between countries.

D9. The "purchasing power parity" (P.P.P.) is the exchange rate that equalizes the prices of internationally traded goods across countries. A group of economists at the International Comparison Program, rub by the World Bank, have calculated the PPP exchange rate for all countries, based on detailed studies of the prices and quantities of internationally tradable goods. The PPP is often used in any international comparison involving prices or amounts of money!!!

D10. The "added value" (A.V.) of a company whose role is to represent the difference between the value of goods and services produced by it with the value of goods and services used to produce these goods and services.

and so on...

Remark

Never forget that when you read macro-economics indicators in papers that first there is a margin error on them (even if it is not indicated) and second that these indicator are given with a time lag of months or years sometimes. So when you see a good macroeconomic indicator this year it is in fact the macroeconomics wealth of previous year and furthermore because sells of major Fortune 500 are done on customers orders of prior-previous year (typical for machines and some food stuff) it is in fact with two years late!

15.3.1.2.1 Cobb-Douglas Model

In 11928 (holocene calendar), Charles Cobb and Paul Douglas published a study in which appeared the modelling of the growth of the US economy between 11899 (holocene calendar) and 11922 (holocene calendar). They had adopted a simplified view of the economy in that the quantity produced of goods is based that the amount of work done and the amount of capital invested.

Although many other factors affect economic performance, their model has shown to be remarkably accurate. The function they used to model production was of the form:

$$P(L,K)=bL^{\alpha}K^{1-\alpha} \tag{15.3.91}$$

where P is the total production (the monetary value of all goods produced in one year), L the amount of work (the total number of hours worked in a year) and K the amount of investment (the monetary value of all machinery, equipment and buildings). The factor b is a "collector" constant that represents technology, efficiency, and other things not accounted for by K and L.

Although this model has been applied primarily to economy, we can find it in many scientific articles in biology or human resources. It seemed appropriate to show in detail (since it is the purpose of this textbook to see things in details!) where does this strange function comes from because the approach is relatively great.

First, if we have a function of two variables $P(L,K)$, the partial derivative $\partial P/\partial L$ indicates the rate of change in production with respect to the amount of labour only. This is what economists name the "marginal production in relation to labour" or "marginal productivity of labour". Similarly, the partial derivative $\partial P/\partial K$ indicates the rate of change in production relative to capital and is named the "marginal productivity of capital".

If we increase the number of machines K, will we be able to produce more cookies? Of course! What if we increase the amount of labour L? Same thing. We can observe this mathematically as well:

$$\begin{aligned}\frac{\partial Y}{\partial K}=b(\alpha K^{\alpha-1})L^{1-\alpha}>0\\ \frac{\partial Y}{\partial K}=b(\alpha K^{\alpha-1})L^{1-\alpha}>0\end{aligned} \tag{15.3.92}$$

In these terms, the hypothesis (assumptions) of Cobb and Douglas (to confront to historical measures and adapt to these measures in individual cases) are as follows:

- H1. Without labour L or without capital K no production.
- H2. The marginal productivity of labour is proportional to the amount produced P per unit of labour L.
- H3. The marginal productivity of capital is proportional to the amount produced P per unit of capital K

Given that the production per work unit is the P/L, hypothesis 2 can be written as:

$$\frac{\partial P}{\partial L}=\alpha\frac{P}{L} \tag{15.3.93}$$

for a given constant α.

If the case where the invested capital K is constant ($K=K_0$), the partial differential equation becomes an ordinary differential equation:

$$\frac{\mathrm{d}P}{\mathrm{d}L}=\alpha\frac{P}{L} \tag{15.3.94}$$

This gives us:

$$\frac{1}{P}\mathrm{d}P=\alpha\frac{1}{L}\mathrm{d}L \tag{15.3.95}$$

So solving this differential equation gives:

$$\ln(P)=\alpha\ln(L)+c^{te}\Rightarrow e^{\ln(P)}=e^{\alpha\ln(L)}e^{c^te}=C^{te}e^{\ln(L^\alpha)} \tag{15.3.96}$$

Therefore:

$$P(L,K_0)=C^{te}(K_0)L^\alpha \tag{15.3.97}$$

because as we fixed K it is obvious that the constant depends on K_0.

By doing exactly the same for the third hypothesis:

$$\frac{\partial P}{\partial K}=\beta\frac{P}{K} \tag{15.3.98}$$

Therefore:

$$P(L_0,K)=C^{te}(L_0)K^{\beta} \tag{15.3.99}$$

So we have at the end:

$$\begin{gathered}P(L,K_0)=C^{te}(K_0)L^{\alpha}\\P(L_0,K)=C^{te}(L_0)K^{\beta}\end{gathered} \tag{15.3.100}$$

and identifying term by term, we get immediately:

$$P(L,K)=c^{te}K^{\beta}L^{\alpha} \tag{15.3.101}$$

Let us observe the implications of this last relation if capital and labour are both multiplied by a constant factor m:

$$P(mL,mK)=c^{te}(mK)^{\beta}(mL)^{\alpha}=c^{te}m^{\alpha+\beta}K^{\beta}L^{\alpha}=c^{te}m^{\alpha+\beta}P(K,L) \tag{15.3.102}$$

If we impose:

$$\alpha+\beta=1 \tag{15.3.103}$$

then the production is multiplied by the same factor m. And we are in a proportionality assumption of production which seemed was consistent Cobb and Douglas assumption. since then:

$$\boxed{P(L,K)=c^{te}K^{\beta}L^{\alpha}} \tag{15.3.104}$$

The economic data used by Cobb and Douglas are those of the following table published by the U.S.A. Government (the years in the table below are given according to the holocene calendar):

Year	P	L	K
11899	100	100	100
11900	101	105	107
11901	112	110	114
11902	122	117	122
11903	124	122	131
11904	122	121	138
11905	143	125	149
11906	152	134	163
11907	151	140	176
11908	126	123	176
11909	155	143	198
11910	159	147	208
11911	153	148	216
11912	177	155	226
11913	184	156	236
11914	169	152	244
11915	189	156	236
11916	225	183	298
11917	227	198	335
11918	223	201	366
11919	218	196	387
11920	231	194	407
11921	179	146	417
11922	240	161	431

Table 15.34: **U.S. data in % used by Cobb and Douglas**

They deliberately took 11899 (holocene calendar) as a base, that is to say, they attributed the level 100 to each factor and expressed the values of other years as a percentage of that year.

To determine the Cobb-Douglas coefficients, we just have to make a linear regression by the least squares method (see section Numerical Methods page 6324). If we take the logarithm:

$$\begin{aligned}\ln(P)&=\ln\left(c^{te}K^{1-\alpha}L^{\alpha}\right)=\ln\left(c^{te}K^{1}K^{-\alpha}L^{\alpha}\right)=\ln\left(c^{te}K\left(\frac{L}{K}\right)^{\alpha}\right)\\&=\ln\left(c^{te}\right)+\ln(K)+\ln\left(\left(\frac{L}{K}\right)^{\alpha}\right)=\ln\left(c^{te}\right)+\ln(K)+\alpha\ln\left(\frac{L}{K}\right)\end{aligned} \tag{15.3.105}$$

and after rearranging, still using the properties of logarithms (see section Functional Analysis page 2284):

$$\ln\left(\frac{P}{K}\right)=\ln(c^{te})+\alpha\ln\left(\frac{L}{K}\right) \tag{15.3.106}$$

We put now to simplify:

$$x=\ln\left(\frac{L}{K}\right)\qquad y=\ln\left(\frac{P}{K}\right) \tag{15.3.107}$$

Therefore we just end up with a linear function:

$$y=\alpha x+\ln\left(c^{te}\right) \tag{15.3.108}$$

We then just adapt the table to inject it into a spreadsheet or statistical software to obtain (see section Numerical Methods page 6324):

$$\alpha\cong 0.75 \qquad c^{te}\cong 1.01 \tag{15.3.109}$$

Therefore:

$$P(L,K)\cong 1.01K^{0.25}L^{0.75} \tag{15.3.110}$$

By testing compared to the data table we get:

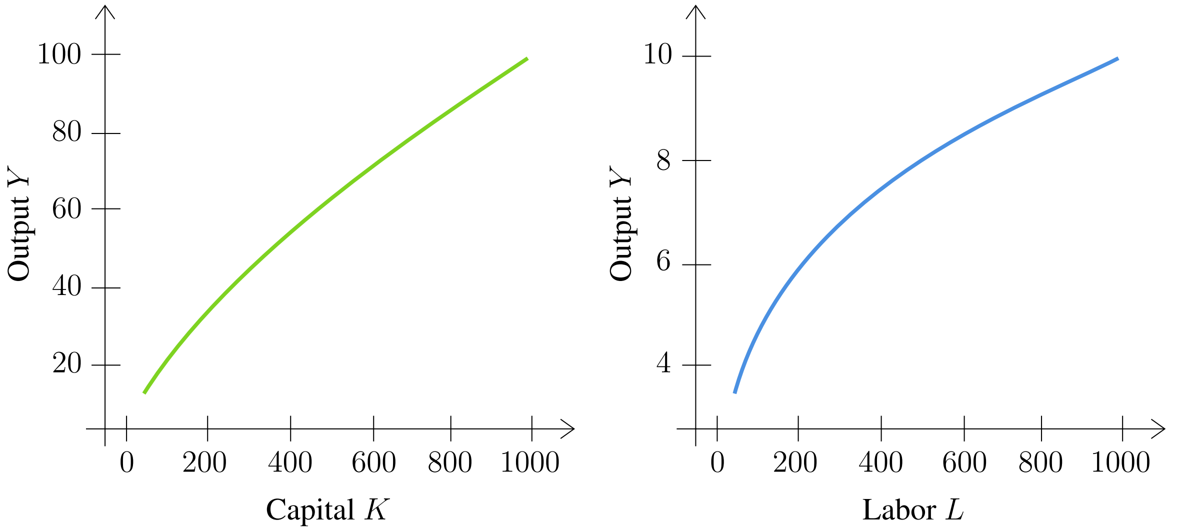
$$\begin{gathered}P_{11908}(123,185)\cong 137\\ P_{11910}(147,208)\cong 161\\ P_{11920}(194,407)\cong 235\\ P_{11922}(161,431)\cong 207\end{gathered} \tag{15.3.111}$$

So we see that the model... is a model... with some errors as every model. For example a typical "evolutionary" error (that a lot of economists still ignore and also politicians) of this model can illustrated by the following assumption:

If we keep increasing the number of machines K and hold labour L constant, economists say that we will produce less and less cookies per machine added. Same goes for labour if you hold capital constant. Why? Diminishing marginal returns! If we have too many machines but not enough workers, the machines will be idle. If we have too many workers for the machines, they overcrowd. Simple as that. But economists and politicians forget that now machine are "intelligent" and can manage and learn by themselves!

How can we prove mathematically this previous fact (that marginal returns diminish when increasing of the variable)?

Mathematically, we say that Y is increasing in K, but at a declining rate. Y is also increasing L, but at a declining rate. Indeed here is a plot of this:



(a) Plot of Y against K holding L constant (a) Plot of Y against L holding K constant

Figure 15.79: Y **Increasing in** K **and** L **at a declining rate**

This requires us to find the second derivatives:

$$\begin{aligned}\frac{\partial^2 Y}{\partial K^2} &= b(\alpha(\alpha-1)K^{\alpha-2})L^{1-\alpha}<0\\ \frac{\partial^2 Y}{\partial L^2} &= bK^{\alpha}((1-\alpha)(-\alpha)L^{-\alpha-1})<0\end{aligned} \quad (15.3.112)$$

as $(\alpha-1)$ is negative, since we stipulated that $0<\alpha<1$.

The assumptions H2 and H3 implies what economists name the "constant return to scale". In other words, we are taking our original K and L, and multiplying them both by the same proportion c^{te}. Then we investigate what happens to output.

So originally, we have:

$$Y_0 = bK^{\alpha}L^{1-\alpha} \quad (15.3.113)$$

After increases in input by a constant c^{te}, we have:

$$Y_1 = b(c^{te}K)^{\alpha}(c^{te}L)^{1-\alpha} = c^{te}Y_0 \quad (15.3.114)$$

We have just shown that if we increase both K and L by a same constant, then Y gets increased by the same factor. In fact we know it is much more interesting to invest in machines rather than in humans. So this model and through the assumptions H2 and H3 was valid only at the time computers (and defacto machines) were not able to do the work of humans.

15.3.2 Monetary Model

It is funny for an engineer to see at least once in his life how an idealistic macroeconomic Monetary Model can be build. We can see that the whole is just a summation of obvious assumptions that has for only purpose to try to have the most complicated mathematical formula at the end to makes it look like a pure science...

To build a macroeconomic monetary model in a pure and perfect competition market (remember above the five assumptions of such a market) we will assume that the "monetary utility" can be defined a priori by four properties:

- P1. It is a counting unit
- P2. It is a medium of payment (via trade)
- P3. It is a reserve value
- P4. It is in relative value a conservative quantity

This description approach is however insufficient to mathematical analysis: we need a comprehensive explanatory system because here we only formalized facts. We must therefore establish the link between money and monetary theory.

Apart from the value that represents currency, it derives its utility of the property it provides exchanges of goods. This is what we name the "derived utility".

Let us note O_u the available money on the market. It therefore depends on the existing total quantity of money Q_u minus the "cash" e kept by economic agents (who exchanged goods against money). We can then write the following relation named "money supply according to Walras":

$$\boxed{O_u = Q_u - e \Leftrightarrow Q_u = O_u + e} \tag{15.3.115}$$

The cash e is also in some way that of households and is a real demand for goods, which can be expressed necessarily in monetary form.

Remark

If all people really knew how modern money creation process works (the one used since the majority of states around the world allow banks to create "debt money" - that is to say, to create the money from a debt repayment promise) the system would probably not last long. Moreover, the contemporary system is relatively fragile and vicious in reality as we will see later (it will probably crash down soon or later).

Sales agents, when they sale their goods, a priori desire a given sum of monetary value against the sale of these goods, value named "desired monetary amount" and denoted by D_u. The problem is that D_u can be expressed as working hours, gold, air, animals, vegetables, etc. This is why it is important to convert this value into common artificial price measure of the monetary value.

We express this desired monetary amount in "cash" and for this we introduce a price to monetary value (because there exist different way to pay a cash value of a good having a given desired monetary amount). The desired monetary amount is then written over all market balances. The cash serves to express the relative prices for the overall balance. There is a cash desired by the agents for the realization of general equilibrium. These are actually real goods in monetary form:

$$\boxed{e = p_u D_u} \tag{15.3.116}$$

where p_u is the "monetary cash price" (factor variable over time and that brings in a market that is not "flow" to speculate). In a flow-market p_u will always assumed to be equal to unity. We can then write for different currencies in a global heterogeneous market:

$$\boxed{p_u D_u = \sum_{i=1}^{n} p_i D_i} \tag{15.3.117}$$

where p_i is the relative monetary cash price of currency i relatively to p_u

Remarks

R1. In an isotropic single currency market this relations would not need to be written.

R2. "Parity" is the term used when we look for equivalence of foreign currencies exchange rate and it is dependent (among others) of time and it is important to consider its variations in the context of goods market where the currency is not unique and payments not immediate.

The desired monetary amount by agents can then be define using the relation:

$$D_u = \frac{1}{p_u} \sum_{i=1}^{n} p_i D_i \tag{15.3.118}$$

Returning back to O_u but this time in the point of view of the companies (firms). They need money to make payments and to operate (wages, investments, etc.) and the desired monetary amount D'_u of all these companies at a given moment in time is in an ideal case necessarily equal to the total money available on the market such that:

$$D'_u := \frac{1}{p_u} O_u \tag{15.3.119}$$

since companies (firms) sell goods on the quantity of money Q_u of agents of the economic market minus the cash margins of these agent e.

The last relation implies that:

1. The aggregate demand necessarily creates an equal quantity of aggregate production ("Say Law").
2. The selling price of goods tends to be equal to their cost price (this is for simplification purposes and also a condition for what will follow!).

Remark

This last relation also means that the whole offer is satisfied only by the demand from the economic agents and the aforementioned amount of monetary value consists only of goods outside the companies (producers/firms/manufactures) and that they therefore have no stocks!

This also corresponds to a certain quantity of goods since the idea is to propose goods to get money (seen of the point of view of companies/firms/manufactures!). Therefore we can write:

$$D'_u := \frac{1}{p_u} O_u = \frac{1}{p_u}\left[\sum_{j=1}^{n} p_j D'_j\right] \tag{15.3.120}$$

But as the goods already in possession by the economic agents will also have to be renewed, the companies (firms) have finally as potential total amount of money available in the market:

$$Q_u = D'_u + D_u = \frac{1}{p_u}\left[\sum_{j=1}^{n} p_j D'_j + \sum_{i=1}^{n} p_i D_i\right] = O_u + e \tag{15.3.121}$$

The sum in brackets corresponds to the total money available in the market in the form of household assets and potential cash in the restriction of goods with a common currency cash price p_u. This is named the "money supply". It is restrictive as a model but sufficient in the context of determining the price of a given type good.

We write then by definition for the money supply according to Walras:

$$Q_u = D'_u + D_u = \frac{1}{p_u} H_u = O_u + e \tag{15.3.122}$$

This last relation implies that economic agents may not collectively choose to increase the amount of money e they hold...

In view of the foregoing, the reader will have noticed that this model considers that money is neutral in the sense that the total amount of money in circulation exerts no influence on the relative prices of products from each to others, nor on the level of supply and demand for products. The currency is not desired for itself...

15.3.2.1 Monetary base conservation

In economics, the "monetary base" in a country is defined as the portion of a commercial bank's reserves that consist of the commercial bank's accounts with its central bank plus the total currency circulating in the public, plus the currency, also known as vault cash, that is physically held in the bank's vault.

Remark

The monetary base should not be confused with the "money supply" which consists of the total currency circulating in the public plus the non-bank deposits with commercial banks.

The previous relations that we have obtained assume therefore that the money supply is constant. What is interesting to know then is that when a central bank print out money at time t, since it keep this money inside the bank and don't tell anyone it print it out, the relative value of money will remain constant such that:

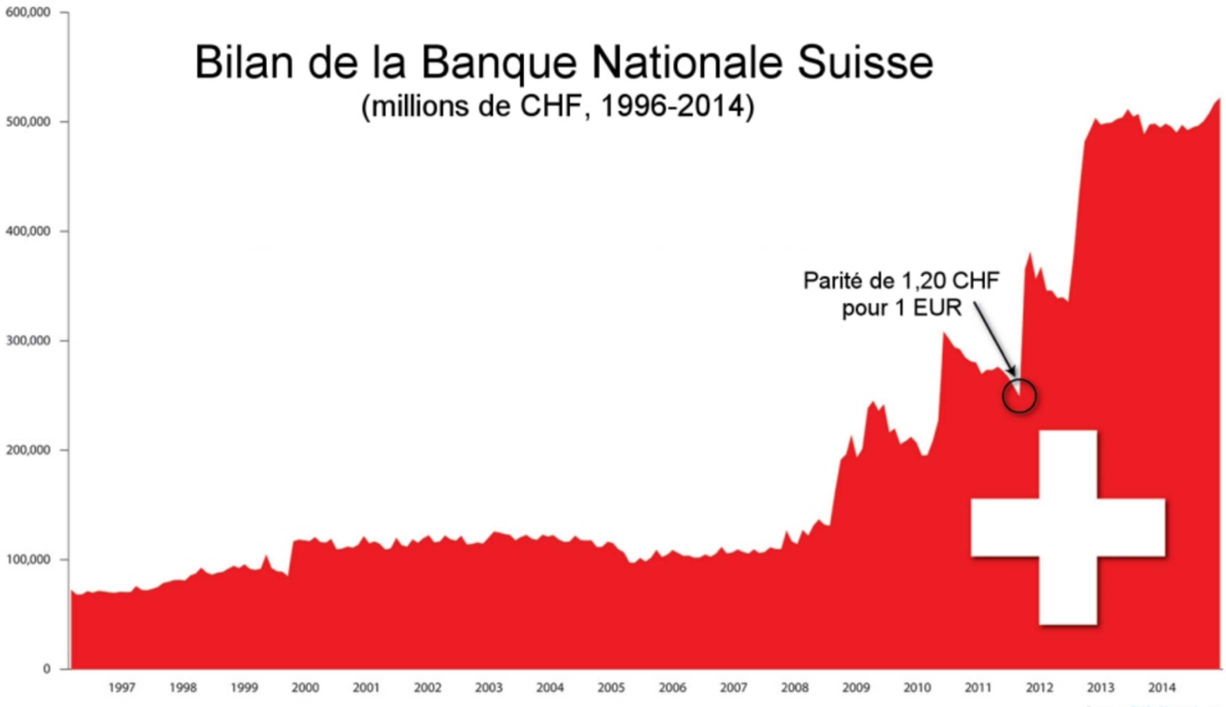
$$p_i(t)=p_i(t-\Delta t)\Leftrightarrow\frac{p_i(t-\Delta t)}{p_i(t)}=1 \tag{15.3.123}$$

But if the central bank put the new quantity of money on the market, then by conservation of the value p_i should be re-evaluated. That means that all the prices, wages, etc. in the country where the bank has printed out the money should be multiplied by a factor:

$$\boxed{\frac{p_i(t-\Delta t)}{p_i(t)}=\frac{Q_u(t)}{Q_u(t-\Delta t)}} \tag{15.3.124}$$

So there is no "law of conservation of money" or "law of monetary base conservation" on the long time term at work but.... rather a "law of relative money price value to the monetary base"!

However it is funny to notice that some Governments, hands in the hands with the Central bank ("monetary policy") don't do such corrections depending on the situations and therefore violates the monetary base conservation principle. For example in Switzerland, the monetary base in 12012 (holocene calendar) had increased by a factor 6, from 100 billion to 600 billions (we speak then of "money dilution"), without saying to the banks, shops and individuals that everything should have therefore been multiplied by a factor 6 (the amount in the bank accounts, the wages, GDP, and so on...) since the value of the Swiss money therefore lost a value of a factor 6 in an international point of view (deflation).



Source: SNB, Bloomberg

Figure 15.80: **Swiss National Banks reserve (source: Bloomberg)**

But in this case the Swiss Central Bank has argued during a conference (using random bullshit language...) that this was:

1. Because the Swiss money was over-evaluated of the point of view of the international market.
2. The wages, and individual bank account are not included in the measurement of the monetary base (this last statement being given without surprise with no mathematical proof...)

It is common in Governments to split the monetary base in three "monetary aggregates" (some split it in four):

M1. Money (banknotes and coins)
M2. Household savings (private saving accounts, deposits <2 years)
M3. Corporates savings (debts <2 years, bonds, shares)

And we have (the M_1 and M_2 as previously mentioned are sometimes ignored and the argument - when the mathematical proof shows the opposite... - is that they can be neglected):

$$M_1 \subset M_2 \subset M_3 \quad (15.3.125)$$

Here are some plots showing a monetary policy for a country (still Switzerland) for M_1, M_2 and M_3:

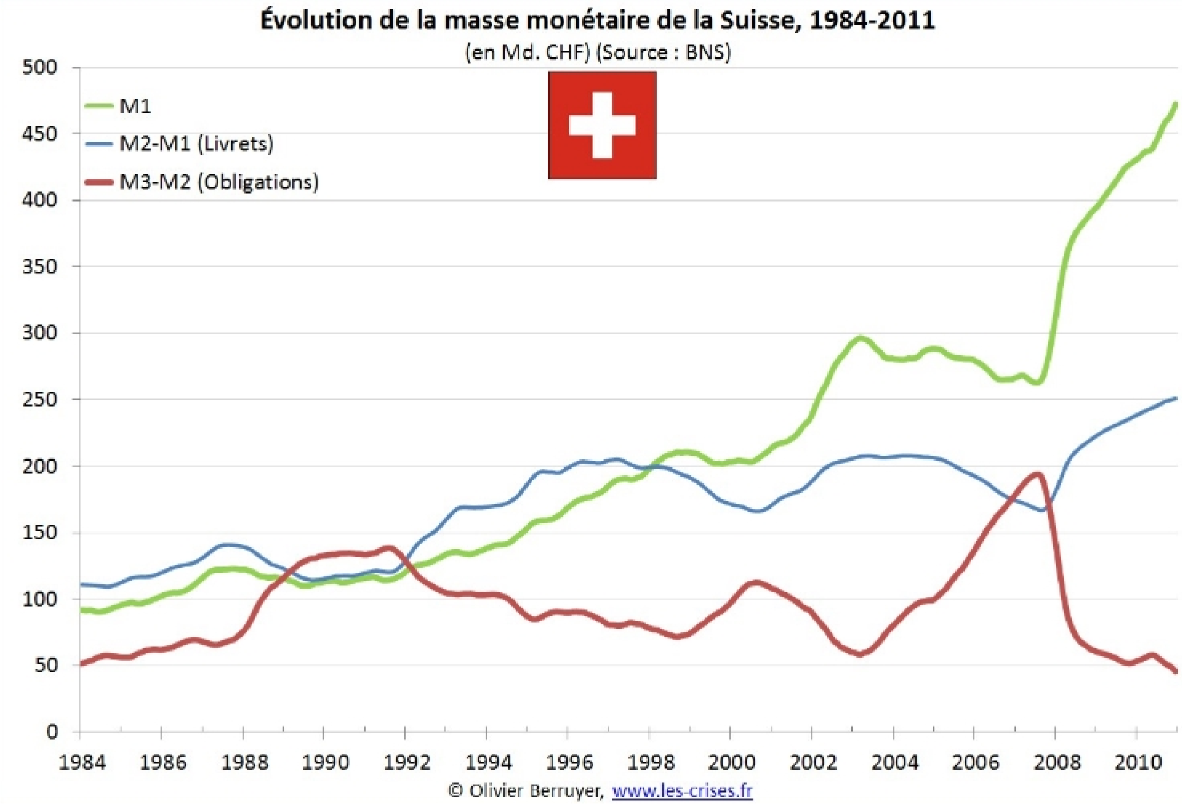


Figure 15.81: **Evolution of monetary base in Switzerland (source: `www.les-crises.fr`, author: Olivier Berruyer)**

and in relative value:

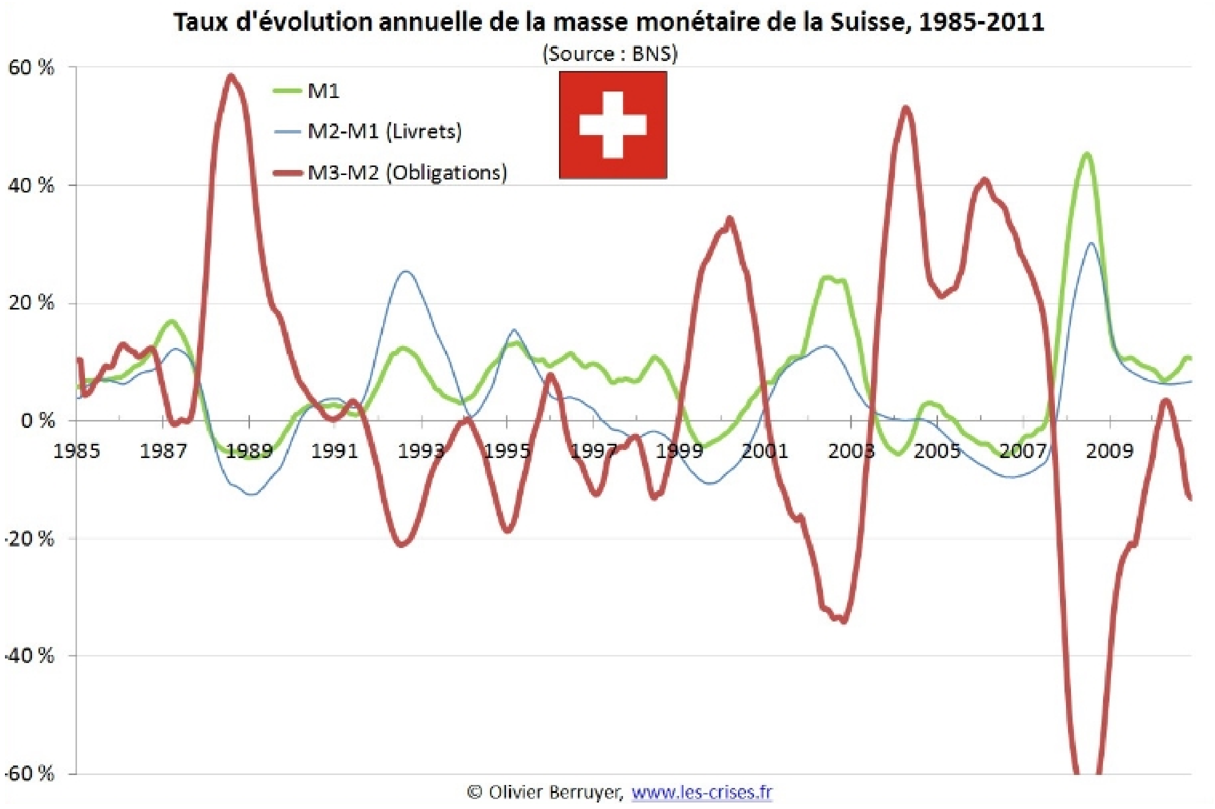


Figure 15.82: **Relative evolution of monetary base in Switzerland (source: `www.les-crises.fr`, author: Olivier Berruyer)**

15.3.2.2 Walras' law

Walras' Law is a principle in general equilibrium theory asserting that budget constraints imply that the values of excess market demands (or, conversely, excess market supplies) must sum to zero in a pure and perfect competition market where economic agents do not increase the amount of money they hold and where the value of money is equal...

Moving from the particular to the general, the money value of what the j-th agent transactor plans to purchase (Demand) can be written symbolically as:

$$p_1 D_{1j} + p_2 D_{2j} + \ldots + p_n D_{nj} = \sum_{i=1}^{n} P_i D_{ij} \tag{15.3.126}$$

where $p_1, \ldots, p_n$ are the prices of the n goods, and $D_{1j}, \ldots, D_{nj}$ are the quantities of those goods that the j-th agent plans to purchase.

Similarly, the money value of what the j'th individual plans to Sell can be written symbolically as:

$$p_1 S_{1j} + p_2 S_{2j} + \ldots + p_n S_{nj} = \sum_{i=1}^{n} P_i S_{ij} \tag{15.3.127}$$

where $p_1, \ldots, p_n$ are the prices of the n goods, and $S_{1j}, \ldots, S_{nj}$ are the quantities of those goods that the j'th agent plans to sell.

Now let us adopt Say's Law that is to say that the money value of all the goods that the j-th individual plans to buy must always be equal to the money value of all the goods that individual plans to sell, we may write:

$$\sum_{i=1}^{n} p_i D_{ij} \equiv \sum_{i=1}^{n} p_i S_{ij} \tag{15.3.128}$$

This condition is obviously written as an identity (remember that an identity is a proposition which is true by definition and is usually written with a $\equiv$ rather than an $=$ sign in it), since we assume that no individual transactor in our model will be so misguided as to suppose that he or she can acquire something for nothing (tacitly, our model assumes that transactors are not thieves, extortioners, embezzlers - or philanthropists!).

Remark

The changes in the currency prices do not affect the actual balance if and only if all these relative prices change in the same proportion (and thus excess demand should no increase or decrease).

Granted that each individual's planned market transactions in goods satisfy condition the previous identity, it follows as a matter of simple arithmetic that the money value of the quantities demanded by all individuals must be equal to the money value of the quantities offered for sale by all individuals. Therefore, summing previous identity over all individuals (as supposed to be independent!) we obtain:

$$\sum_{j=1}^{m} \sum_{i=1}^{n} p_i D_{ij} \equiv \sum_{j=1}^{m} \sum_{i=1}^{n} p_i S_{ij} \tag{15.3.129}$$

when we assume that there are m individual transactors.

Factoring out the price variables from this expression yields:

$$\sum_{i=1}^{n} p_i \left(\sum_{j=1}^{m} D_{ij} \right) \equiv \sum_{i=1}^{n} p_i \left(\sum_{j=1}^{m} S_{ij} \right) \tag{15.3.130}$$

However, the expression in parentheses on the left-hand side is simply the total market demand for the i-th good, since it is the sum of the individual transactors' demands for that good. We will write this total market demand as D_i. Similarly, the expression in parentheses on the right-hand side is simply the total market supply of the i'th good, since it is the sum of the individual transactors' supplies of that good. We will write this total market supply as S_i.

Thus, we arrive at the conclusion:

$$\boxed{\sum_{i=1}^{n} p_i D_i \equiv \sum_{i=1}^{n} p_i S_i} \tag{15.3.131}$$

Which is a proposition known as "Walras' identity". It states that the money value of all planned market purchases are identically equal to the money value of all planned market sales and it is valid whether or not market prices equate demand with supply for each individual type of good.

We can find also sometimes the identity with the following rearrangement:

$$\sum_{i=1}^{n} p_i D_i - \sum_{i=1}^{n} p_i S_i \equiv 0 \tag{15.3.132}$$

But this result is trivial and is not what interest in reality! What interest us is two very important and obvious implications that can be written in a mathematical way (do not forget that it is the main purpose of economists even if the level is a pre-school one). One implication relates to the generality of equilibrium. The other refers to states of dis-equilibrium. We will deal with each in turn using maths (even if it is trivial by the Say's Law).

Theorem 15.6. We will prove that if all but one of the markets in an economy are in equilibrium (+some other assumptions already communicated before!), then that other market also must be in equilibrium.

Proof 15.6.1. Assume that a set of prices has been established which will equate demand with supply in every market except the n-th market. Since all $n-1$ markets are in equilibrium, then if the market is big enough we can always find and indexation of goods such that:

$$D_1 = S_1, D_2 = S_2, \ldots, D_{n-1} = S_{n-1} \tag{15.3.133}$$

Next, multiply through by the set of prices that put these $n-1$ markets in equilibrium. Then:

$$p_1 D_1 = p_1 S_1, p_2 D_2 = p_2 S_2, \ldots, p_{n-1} D_{n-1} = p_{n-1} S_{n-1} \tag{15.3.134}$$

And we apply Say's law we get:

$$\sum_{i=1}^{n-1} p_i D_i \equiv \sum_{i=1}^{n-1} p_i S_i \tag{15.3.135}$$

If this is now subtracted from Walras' Identity we obtain:

$$p_n D_n \equiv p_n S_n \tag{15.3.136}$$

Which implies immediately that the n-th market is also in equilibrium. It also follows in an isotropic single currency market that:

$$D_n \equiv S_n \tag{15.3.137}$$

To recapitulate verbally, we have shown that if all but one of the markets in an economy are in equilibrium (+some other assumptions already communicated before!), then that other market also must be in equilibrium. This result is known as "Walras' law":

$$\boxed{\sum_{i=1}^{n-1} p_i D_i \equiv \sum_{i=1}^{n-1} p_i S_i \Rightarrow p_n D_n = p_n D_n} \tag{15.3.138}$$

■ Q.E.D.

We now look at some implications of Walras' Identity for dis-equilibrium.

Corollary 15.6.1. Assume that one market (the n-th market) is in dis-equilibrium. This may take the form of (positive) excess demand (where $p_nD_n > p_nS_n$) or excess supply, also known as "negative excess demand" (where $p_nD_n < p_nS_n$).

Proof 15.6.2. It is an implication of Walras' Identity that for all markets taken as a whole there can be neither excess supply nor excess demand when we sum over all markets. We can see this by rearranging Walras's Identity as follows (without forgetting assumptions!):

$$\sum_{i=1}^{n}(p_iD_{ij}-p_iS_{ij})\equiv 0 \tag{15.3.139}$$

In order for this condition to be satisfied in the presence of disequilibrium in the n-th market, it must be the case that there is on off-setting dis-equilibrium in at least one other market. Form this result emerge the basic idea of "demand & supply equilibrium".

■ Q.E.D.

To resume this pseudo-proof, Walras' and Say's laws are like the conservation energy in physics in a closed thermodynamic system but where mass are goods and the constant normalizing factor to conversion in energy is the money...

Corollary 15.6.2. Let us see now another pseudo-proof that shows that equilibrium (Say's Law) is not affected by the change in price if we suppose that the demand and supply are proportional (or inversely proportional) to the price. That is to say that the demand/supply is a homogeneous function of degree 1 of the price such that:

Proof 15.6.3.

$$\begin{aligned}D_i(\alpha p_1,\alpha p_2,\ldots,\alpha p_n)=\alpha^1D_i(p_1,p_2,\ldots,p_n)=\alpha D_i(p_1,p_2,\ldots,p_n)\\ S_i(\alpha p_1,\alpha p_2,\ldots,\alpha p_n)=\alpha^1S_i(p_1,p_2,\ldots,p_n)=\alpha S_i(p_1,p_2,\ldots,p_n)\end{aligned} \tag{15.3.140}$$

Study this fact is interesting only to have once again at look on how economists try to formalize idealistic concepts using scientific words and mathematical methods. For this purpose we start from Walras's identity:

$$\sum_{i=1}^{n}p_iD_i\equiv\sum_{i=1}^{n}p_iS_i \tag{15.3.141}$$

That we rewrite to indicate that D_i and P_i depends on prices $p_1,\ldots,p_n$ of their own components:

$$\sum_{i=1}^{n}p_iD_i(p_1,\ldots,p_n)=\sum_{i=1}^{n}p_iS_i(p_1,\ldots,p_n) \tag{15.3.142}$$

Now consider a change α on prices of components:

$$\sum_{i=1}^{n}p_iD_i(\alpha p_1,\ldots,\alpha p_n)=\sum_{i=1}^{n}p_iS_i(\alpha p_1,\ldots,\alpha p_n) \tag{15.3.143}$$

Using the assume property of homogeneity we can write this:

$$\sum_{i=1}^{n} p_i \alpha D_i(p_1,\ldots,p_n) = \sum_{i=1}^{n} p_i \alpha S_i(p_1,\ldots,p_n) \tag{15.3.144}$$

This result for the moment just means that a change in prices of the components of a goods is impacted on the global price of the good (trivial!). Therefore:

$$\alpha \sum_{i=1}^{n} p_i D_i(p_1,\ldots,p_n) = \alpha \sum_{i=1}^{n} p_i S_i(p_1,\ldots,p_n) \tag{15.3.145}$$

And obviously we can eliminate α from the two side of the equality (identity). This finish the proof.

■ Q.E.D.

Therefore the equilibrium is not affected by movements in prices under the assumptions given before. We can therefore understand why in such a simple market wages that are the monetary price of work decrease in implicit value in proportion to the increase of supplied goods (at least in theory...) because everything should be in equilibrium and work being a good like any other... (if something increase in monetary value, something must decrease in monetary value).

Here, relations are based on equations. Walras, however, distinguish two procedures to ensure balance between supply and demand:

1. A theoretical algebraic method. But... we can not easily determine individual needs in advance to know when there will be demand and prepare to build the offer. This system only works if and only if the economic agents are reasonable and agree to wait or that we have a huge database with all information about all products and all economic agents supply and offers.

2. An empirical method that searches the solution by tests/errors operations. There is the presence of a kind of market secretary, "auctioneer". This last announce the prizes for each type of goods that could exist. The economic agents react to this prices: they ask and they offer based on that prices. For the good i, there is a price p_i, we therefore have D_i, S_i. We then compare the supply and demand. In case of equality, the price is an equilibrium price. In case of non-equilibrium, the auctioneer starts the procedure again and so on until there is equilibrium. That is basically the procedure that is used in the stock market and is often named the "invisible hand of the market" !!!

However, the equations show us that we need the price of currency to measure supply and demand and it should be remembered that we have considered money as a given quantity fixed merchandise because the system is in equilibrium between supply and demand. But precisely, agents can not indefinitely divide the total amount of money if the number of goods is increasing. Therefore, for the process of offer and demand to be possible, we must be prepared to inject money in the market (otherwise it could become immobile which is probably not good on the long term...). We must of course also be prepared to withdraw also money and this is where the government intervenes in the economy to regulate the quantity of money in every possible way (through taxes for example).

Thus, according to the Walras model, the quantity of money available in the market is only depending on the number of economic agents. But should we therefore develop a new model for a more general framework of money demand?

In fact, it is not necessary. We know that if there is a general equilibrium for n goods, there is general equilibrium for $n+1$ goods (and recursively for $n-1$ as well); the last market being simply that of the currency. The Walras model explains therefore why for a given level of quantity of currency there is a corresponding monetary value to goods!

15.3.3 Price Index and GDP

Many people and also sometimes junior politicians when speaking about retirement pension amount forget to take into account that the value of money is not the same across time. Indeed (!) 1.- today is not equal to 1.- tomorrow because of inflation (variation of price of goods). The same occur with many project manager that in their calculation of the Net Present Value (see section Quantitative Management page 8050) forget to take this into account for big project whose development range on 10-20 years or sometimes even more (30 years!).

So a useful (but not unique) way to take this into account is the use of price index that is most of times a normalized average (typically a weighted average) of price relatives for a given class of goods or services in a given region, during a given interval of time.

Also when as consultant and trainer for business analysts of Fortune 500 companies that make market analysis (of consumption, of counterfeits or others) or appoint auditing companies like KPMG or Ernst & Young in the purpose to compare year by year indicators I see always the common undergraduate error that they don't take into account the National Price Index (NPI) or Gross Domestic Product (GDP) to make things really comparable (among other awful errors)...

Definition (#15.3.659): A "price index" is a statistic designed to help to compare how the prices relatives, taken as a whole, differ between time periods or geographical locations. The index can be said to measure the economy's general price level or a cost of living. More narrow price indices help producers with business plans and pricing and are always useful in helping to guide investment.

A naive approach is to consider a $\mathcal{C}$ of goods and services, the total market value of transactions in $\mathcal{C}$ in some period T would be:

$$\sum_{c\in\mathcal{C}} p_{c,T}q_{c,T} \qquad (15.3.146)$$

where $p_{c,T}$ represents the prevailing price of c in period T and $q_{c,T}$ represents the quantity of c sold in period T.

If, across two periods T_0 and T_n, the same quantities of each good or service were sold, but under different prices, then:

$$q_{c,T_n} = q_c = q_{c,T_0} \qquad \forall c \qquad (15.3.147)$$

and:

$$P = \frac{\sum p_{c,T_n} q_c}{\sum p_{c,T_0} q_c} \qquad (15.3.148)$$

would be a reasonable measure of the price of the set in one period relative to that in the other, and would provide an index measuring relative prices overall, weighted by quantities sold.

Of course, for any practical purpose, quantities purchased are rarely if ever identical across any two periods. As such, this is not a very practical index formula.

One might be tempted to modify the formula slightly to:

$$P = \frac{\sum p_{c,T_n} q_{c,T_n}}{\sum p_{c,T_0} q_{c,T_0}} \qquad (15.3.149)$$

This new index, however, doesn't do anything to distinguish growth or reduction in quantities sold from price changes. To see that this is so, consider what happens if all the prices double between T_0 and T_n while quantities stay the same: P will double. Now consider what happens if all the quantities double between T_0 and T_n while all the prices stay the same: P will double. In either case the change in P is identical. Various indices have therefore been constructed in an attempt to compensate for this difficulty.

Of course, for any practical purpose, quantities purchased are rarely if ever identical across any two periods. As such, this is not a very practical index formula.

15.3.3.1 Paasche and Laspeyres price indices

The two most basic formulae used to calculate price indices are the "Paasche index" (after the economist Hermann Paasche) and the "Laspeyres index" (after the economist Etienne Laspeyres).

The Paasche index is computed as:

$$\boxed{P_\text{P} = \frac{\sum p_{c,T_n} q_{c,T_n}}{\sum p_{c,T_0} q_{c,T_n}}} \qquad (15.3.150)$$

while the Laspeyres index is computed as:

$$\boxed{P_\text{L} = \frac{\sum p_{c,T_n} q_{c,T_0}}{\sum p_{c,T_0} q_{c,T_0}}} \qquad (15.3.151)$$

When applied to bundles of individual consumers, a Laspeyres index equal to 1 would state that an agent in the current period can afford to buy the same bundle as he consumed in the previous period, given that income has not changed; a Paasche index of 1 would state that an agent could have consumed the same bundle in the base period as he is consuming in the current period under the same assumption!!!

The Laspeyres index tends to overstate inflation (in a cost of living framework), while the Paasche index tends to understate it, because the indices do not account for the fact that consumers typically react to price changes by changing the quantities that they buy.

As can be seen from the definitions above, if one already has price and quantity data (or, alternatively, price and expenditure data) for the base period, then calculating the Laspeyres index for a new period requires only new price data. In contrast, calculating many other indices (e.g., the Paasche index) for a new period requires both new price data and new quantity data (or, alternatively, both new price data and new expenditure data) for each new period. Collecting only new price data is often easier than collecting both new price data and new quantity data, so calculating the Laspeyres index for a new period tends to require less time and effort than calculating these other indices for a new period.

In practice, price indexes regularly compiled and released by national statistical agencies are of the Laspeyres type, due to the above-mentioned difficulties in obtaining current-period quantity or expenditure data.

15.3.3.2 Fisher index and Marshall–Edgeworth index

A third index, the "Marshall–Edgeworth index" (named for economists Alfred Marshall and Francis Ysidro Edgeworth), tries to overcome these problems of under- and overstatement by using the arithmetic means of the quantities:

$$\boxed{P_{\text{ME}}=\frac{\sum\left[p_{c,T_n}\cdot\dfrac{1}{2}\cdot(q_{c,T_0}+q_{c,T_n})\right]}{\sum\left[p_{c,T_0}\cdot\dfrac{1}{2}\cdot(q_{c,T_0}+q_{c,T_n})\right]}=\frac{\sum[p_{c,T_n}\cdot(q_{c,T_0}+q_{c,T_n})]}{\sum[p_{c,T_0}\cdot(q_{c,T_0}+q_{c,T_n})]}} \tag{15.3.152}$$

A fourth, the "Fisher index" (after the American economist Irving Fisher and not the statistician!!!), is calculated as the geometric mean of P_{P} and P_{L}:

$$\boxed{P_{\text{F}}=\sqrt{P_{\text{P}}P_{\text{L}}}} \tag{15.3.153}$$

However, there is no guarantee with either the Marshall–Edgeworth index or the Fisher index that the overstatement and understatement will exactly cancel the other.

While these indices were introduced to provide overall measurement of relative prices, there is ultimately no way - as far as we know - of measuring the imperfections of any of these indices (Paasche, Laspeyres, Fisher, or Marshall–Edgeworth) against reality.

The Fisher index is used by the Canada statistics since 12001 (holocene calendar) instead of the Laspeyres formula to calculate the estimate of real GDP (Gross Domestic Product) in terms expenditure. The Switzerland OFS use the Laspeyres index (but they don't say since how many years...).

15.3.3.3 Gross domestic product (GDP)

Definition (#15.3.660): The "gross domestic product" (GDP) is a monetary measure of the market value of all final goods and services produced in a period (quarterly or yearly). Nominal GDP estimates are commonly used to determine the economic performance of a whole country or region, and to make international comparisons. Nominal GDP, however, does not reflect differences in the cost of living and the inflation rates of the countries; therefore using a "gross domestic product (at purchasing power parity) per capita" (GDP PPP). Such calculations are prepared by various organizations, including the International Monetary Fund and the World Bank. As estimates and assumptions have to be made, the results produced by different organizations for the same country are not hard facts and tend to differ, sometimes substantially, so they should be used with caution (some companies use the McDonald burger index[15] as element of comparison!).

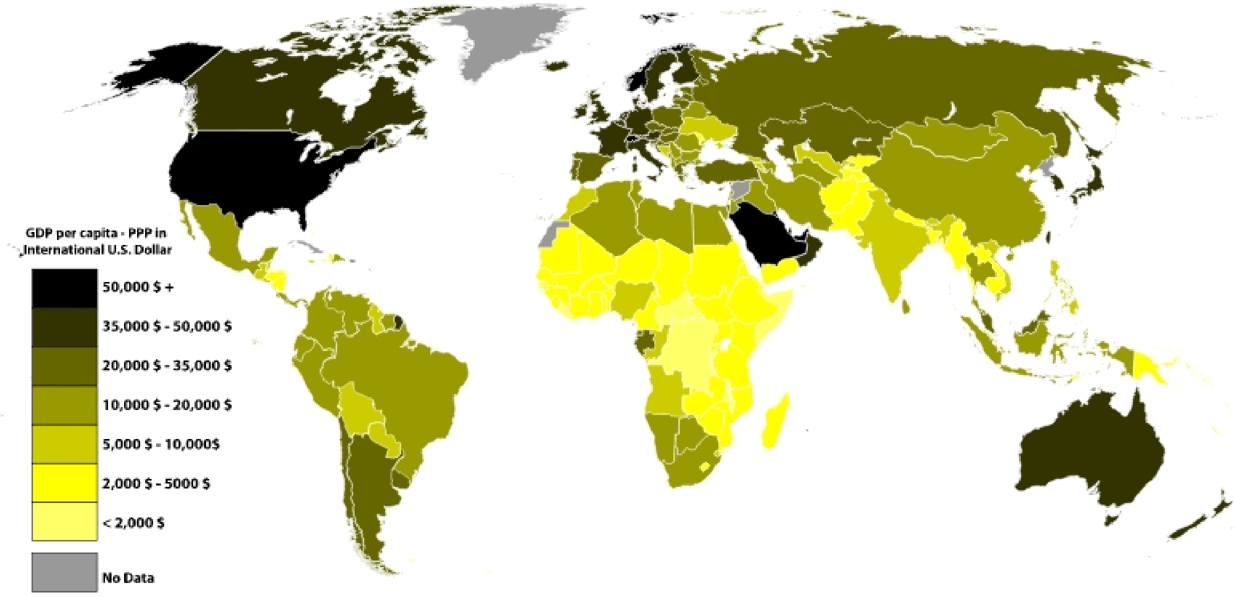


Figure 15.83: **Countries by GDP (PPP) Per Capita in 12015 (source: Wikipedia)**

The basic formula for calculating GDP is:

$$\boxed{\text{GDP} = C + I + E + G} \tag{15.3.154}$$

This formula is almost self-evident (if you take time to think about it)!

GDP is a measure of all the goods and services produced domestically. Therefore, to calculate the GDP, one only needs to add together the various components of the economy that are a measure of all the goods and services produced.

Many of the goods and services produced are purchased by consumers. So, what consumers spend on them (C) is a measure of that component.

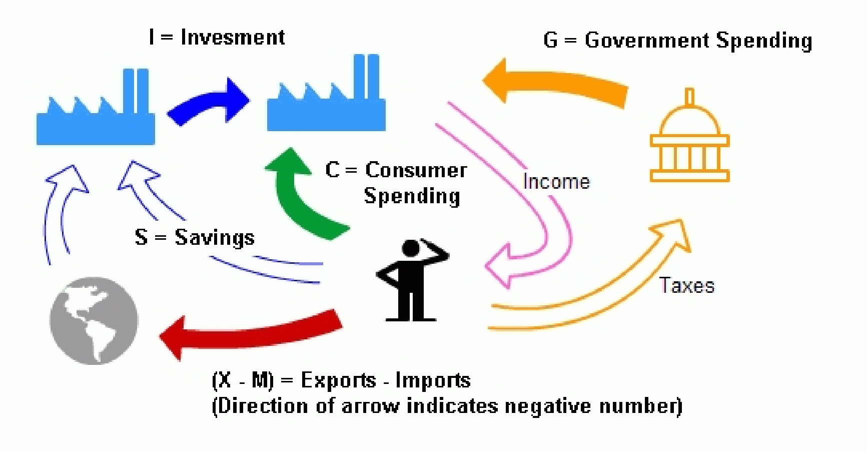
The next component is the somewhat mysterious quantity "I," or investment made by industry. However, this quantity is mysterious only because investment does not have its ordinary meaning. When calculating the GDP, investment does NOT mean what we normally think of in the case of individuals. It does not mean buying stocks and bonds or putting money in a savings account (S in the diagram). When calculating the GDP, investment means the purchases made by industry in new productive facilities, or, the process of "buying new capital and putting it to use". This includes, for example, buying a new truck, building a new factory, or purchasing new software. This is indicated in the diagram by an arrow pointing from one factory (enterprise) to another. In essence, it shows the factory "reproducing itself" by buying new goods and services that will produce still more new goods and services.

---

[15] published by The Economist as an informal way of measuring the purchasing power parity

The next component is E, or the difference between the value of all exports and the value of all imports. If Exports exceeds imports, it adds to the GDP. If not, it subtracts from the GDP. Thus, even if a nation's people work very hard to produce products for exports, but still import more than they export, the nation's GDP will be negatively impacted. This is one of the reasons trade deficits are frequently a political target. Because the balance of trade can be either positive or negative, we can rewrite the equation, showing the components of E, using X for Exports and M for Imports:

$$\boxed{\text{GDP} = C + I + (X - M) + G} \tag{15.3.155}$$



The practitioner mus obviously never that all these fours variables are associates with an estimation error that must be taken into account in any micro or macroeconomics model (for example Switzerland say that the error is only about $\pm 0.5\%$... but this statement should be check by a private independent company as some areas in the country say they have an error of $\pm 2.5\%$).

15.3.4 Supply and Demand Theory

The theory of supply and demand as presented below requires a major revision because highly incomplete. Therefore the ideas presented below are to this day to be taken lightly and only as a simple introduction to the underlying concepts of prices adjustment of goods.

In our human society where there is a exchange (reference) money and goods, one major problems remains that is the of determining the monetary value of a good (material or immaterial). To determine this, we need at least to know the evolution of supply and demand. This is on what we will focus now by starting with simplistic models and complicating them increasingly!

Remark

In our point of view and based on our consulting experience the determination of a product price belongs more to the field of corporate financial engineering than to Economy or Econometrics. Indeed, determining the product of an insurance or Call option doesn't use at all the same techniques level as determining the price of a manufacturing price.

15.3.4.1 Expected utility theory

Before going trough the lecture of a highly complex model on supply and demand, it is necessary to identify what motivates economic agents in their consumption choices and to model their behaviour based on the fundamental principle of rationality.

The economic agent will be perceived as a unique individual with a choice that he seeks to maximize the satisfaction. His tastes are subjective even though they depend on certain objective characteristics such as age, the level of culture, his fortune, etc. The level of satisfaction will be defined from a utility function that we will see the basic principles and the maximization under constraints.

Several principles based utility of goods and lead to the concept of "marginal utility", a central concept in the theory of the preference of the economic agent. According to Aristotle (being at the origin of the concept of value-utility), the utility of goods derives from the satisfaction of needs. Étienne Bonnot de Condillac states: "the value of things is based on their usage". This idea of value based on utility, fundamental for marginalist economists, opposes therefore to the current theoretical labour-value based on the amount of work, direct and indirect, incorporated in the manufacture of the good (Adam Smith, Karl Marx).

However we should, preferably, consider an important assumption in this model: There is a certain satiation of needs, but it is almost never total.

Thus, for a given good marginal utility of the last unit consumed is lower than that of previous units but not null and always positive. This is the "principle of diminishing marginal utility" on the additional unit consumed. Or in other words this law states that as additional units of a good are consumed, while holding then consumption of all other goods constant, the resulting increments in utility will diminish.

Thus, under the multiple consumption $n \in \mathbb{N}$ of a single good of given nominal value, the total utility U_T (sum of marginal utilities $U_M(n)$) is a curve of the type:

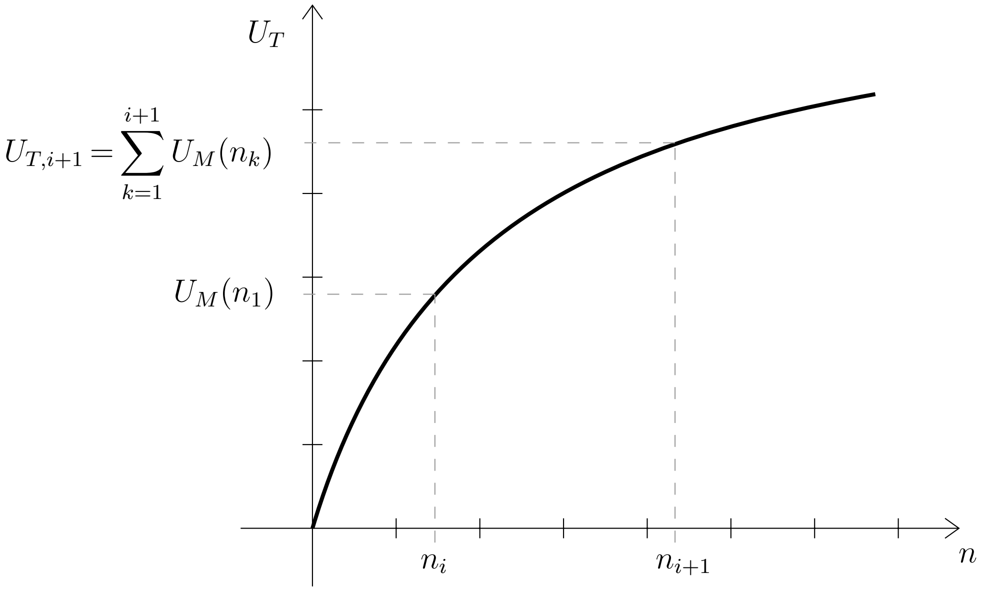


Figure 15.84: **Asymptotic utility**

and therefore the marginal utility is of the type:

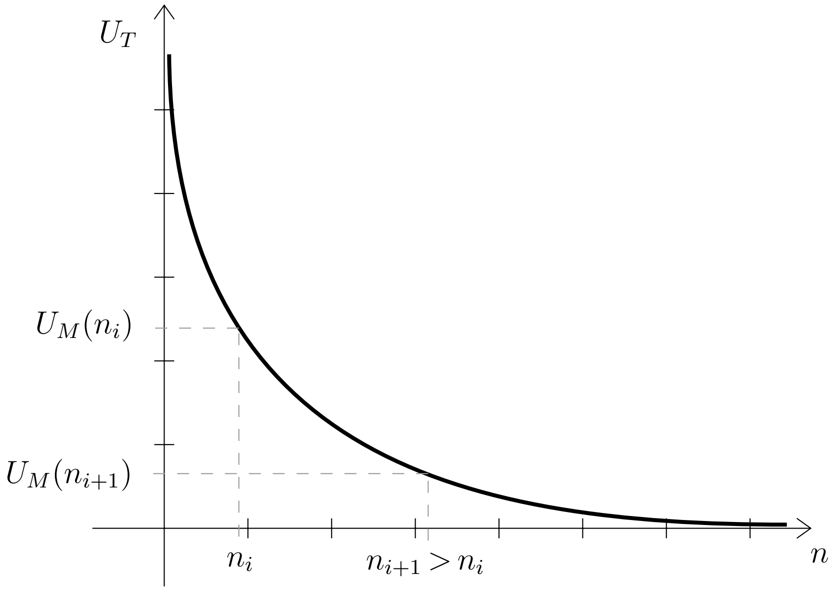


Figure 15.85: **Corresponding decreasing marginal utility**

So, faced with a price for each good, the economic agent compares this price with the marginal utilities that he could successively use from their consumption. He buys them as long as their value exceeds the price (surplus related to their purchase) and ceases to buy as soon as the marginal utility falls below the price of the good. His interest is therefore to buy other products for which there is a positive surplus (marginal utility that is higher to their prices).

This example, relative to one good, should now be extended to a basket of goods to determine the overall utility of this basket.

Consider for this an economic agent i in an economy that has $I \in \mathbb{N}$ goods. It can therefore buy at most I goods. A basket of consumption can therefore corresponds to a vector $\vec{n}_i = (n_1, n_2, \ldots, n_k)$ of goods: where the n_j may represent null quantities purchased by the consumer. The utility $U(\vec{n}_i)$ of this basket will be assumed as additive such that:

$$U(\vec{n}_i) = \sum_j U_{T,j}(n_j) \tag{15.3.156}$$

that is to say the sum of total of utilities relative to the quantities consumed of each good. Let us now consider a basket with two goods, we can without much error hypothesize that these goods can be divided into small fractions x_i as small as we want from other goods/components. So roughly, we no longer work $\mathbb{N}$ but in $\mathbb{N}_+$. Therefore if we consider a basket $\mathbb{R}^2$ of an economic agent, we will assume that it is such that its exact total differential (see section Differential and Integral Calculus page 1466) is zero such that:

$$\mathrm{d}U = \frac{\partial U}{\partial x_1}\mathrm{d}x_1 + \frac{\partial U}{\partial x_2}\mathrm{d}x_2 \tag{15.3.157}$$

The ratio:

$$\text{MRS} = -\frac{\mathrm{d}x_1}{\mathrm{d}x_2} = \frac{\dfrac{\partial U}{\partial x_2}}{\dfrac{\partial U}{\partial x_1}} \tag{15.3.158}$$

is defined as the "marginal rate of substitution" (MRS) between the two elementary goods $1, 2$. That is to say the additional quantity of good 1 we need to provide to the economic agents to accurately compensate for a decrease of one unit of good j under the hypothesis that they are always compensated (fact of the null total exact differential).

The behaviour attributed to the economic agent is to classify all the baskets of possible goods (vectors) using a scale of utility without it necessarily corresponds to a quantified assessment. This classification capability corresponds to the concept of "ordinal utility" (which can - by its name - therefore be ordered...) and to the use of a preference relation, denoted by $\succeq$ (preferred or indifferent to) that satisfies the following trivial properties:

P1. Reflexivity: For every $\vec{n}_1$, For every $\vec{n}_1 \succeq \vec{n}_1$.

P2. Completeness: For every $\vec{n}_1$ and $\vec{n}_1$ either $\vec{n}_1 \succeq \vec{n}_2$ or $\vec{n}_1 \preceq \vec{n}_2$.

P3. Transitivity: $\vec{n}_1 \succeq \vec{n}_2 \wedge \vec{n}_2 \succeq \vec{n}_3 \Leftrightarrow \vec{n}_1 \succeq \vec{n}_3$ (consistency of successive rankings).

P4. Continuity: Let $\vec{n}_1$, $\vec{n}_2$ and $\vec{n}_3$ be such that $\vec{n}_1 \succeq \vec{n}_2 \succeq \vec{n}_3$ then there exists a proportion/fraction $p \in [0,1]$ (that can be assimilated to a probability) such that $\vec{n}_2$ is equally good as $\vec{n}_2 \sim p\vec{n}_1 + (1-p)\vec{n}_3$.

This order relation in the mathematical sense, is used in most of today's presentations of preference theory. This order is complete if it permits to compare two baskets of goods in the set I of goods.

Such a complete order relations permits to define an equivalence relation on the set of goods and a strict order, and to represent preferences from utility functions:

$$\begin{gathered}\vec{n}_1 \succeq \vec{n}_2 \Leftrightarrow U(\vec{n}_1) \geq U(\vec{n}_2)\\ \vec{n}_1 \succ \vec{n}_2 \Leftrightarrow U(\vec{n}_1) > U(\vec{n}_2)\\ \vec{n}_1 \sim \vec{n}_2 \Leftrightarrow U(\vec{n}_1) = U(\vec{n}_2)\\ \vec{n}_1 \sim \vec{n}_2 \sim \vec{n}_3 \Leftrightarrow \forall p \in [0,1]: \quad U(\vec{n}_2) = pU(\vec{n}_1) + (1-p)U(\vec{n}_3)\end{gathered} \tag{15.3.159}$$

If the function U is defined by a number, it does not reflect anymore an assessment of the utility, but since only the possibility to compare the order of the utilities, relatives to baskets with any goods.

The ability to order different baskets of goods of $\mathbb{R}^I$ permits to define surfaces with isolines (see section Functional Analysis page 2857) where the utility is constant, named "indifference curves" or "iso-utility curves".

The following figures give a good representation of these curves in $\mathbb{R}^2$ (basket of two type of goods) and their main properties that will help us latter to better understand the indifference curved used to build the Capital Asset Pricing Model (CAPM) for the study of modern portfolio theory (see further below page 7740).

Thus, two baskets such that $\vec{A} \sim \vec{B}$ in $\mathbb{R}^2$ is graphically traduced in a family of curves, where each is constantly decreasing such that:

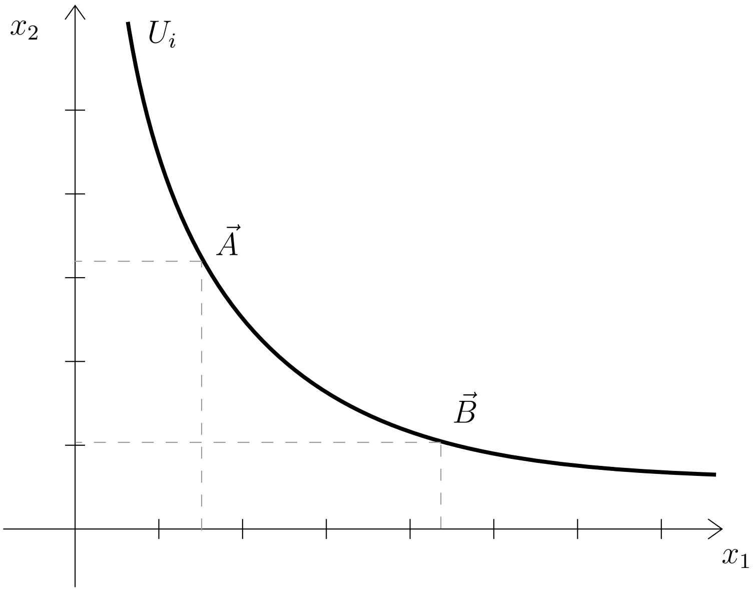


Figure 15.86: **Example of iso-utility (indifference) curve**

Why do not we have straight lines plots or other things? The reason is relatively simple and the following chart explains that trivially. Consider the iso-utility (indifference) curve below that is a projection in the plane of the $U_{\vec{B}}(x_1,x_2): \mathbb{R}^2 \mapsto \mathbb{R}^3$:

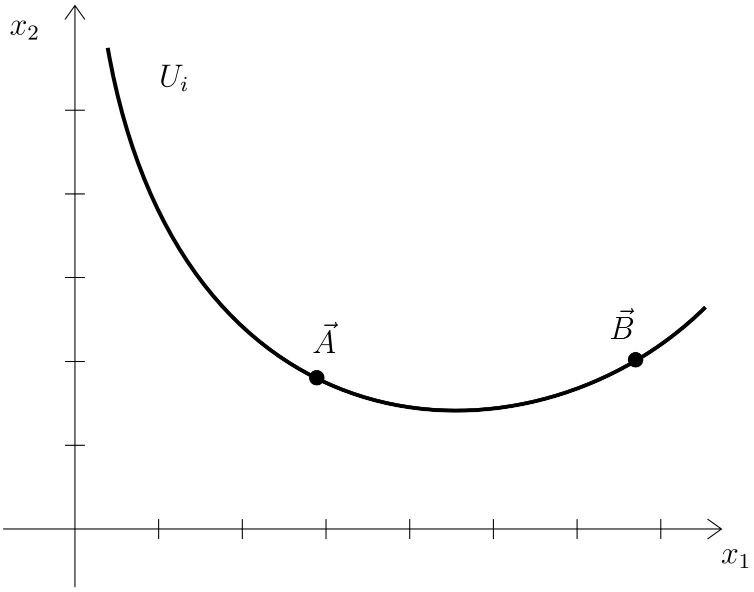


Figure 15.87: **Counter example of iso-utility (indifference) curve**

Above $\vec{B}$ dominates $\vec{A}$ but its basket of goods has more x_1 AND x_2 than $\vec{A}$ and then $U_{\vec{B}}(x_1,x_2)>U_{\vec{A}}(x_1,x_2)$. These two points can therefore not be on the same indifference curve (iso-line) and impose an indifference curve to be decreasing i.e. to be convex!

The assumption that $\vec{A}\sim\vec{B}$ also implies directly that the total exact differential is zero.

The iso-utility curves (indifference curves) can not cut themselves (may not intersect). Even if it is obvious if you did forget that $U_{\vec{B}}(x_1,x_2):\mathbb{R}^2\mapsto\mathbb{R}^3$, consider the two iso-utility curves (indifference curves) below:

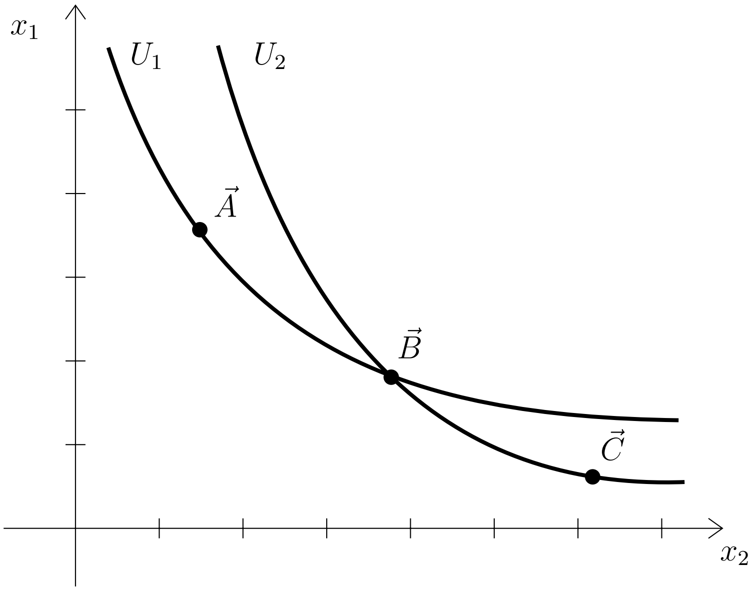


Figure 15.88: **Iso-utility intersection impossibility**

The baskets $\vec{A}$ and $\vec{B}$ are located on the same iso-utility curve U_1 while $\vec{B}$ and $\vec{C}$ are themselves located on the same curve U_2. Thus, we can write $\vec{A}\sim\vec{B}$ and that $\vec{B}\sim\vec{C}$. According the property of transitivity of the utilities, we should then have $\vec{A} \sim \vec{C}$. These two baskets are not equivalent since $\vec{A}$ and $\vec{C}$ are not located on the same curve. Therefore two indifference curves can not intersect.

We are therefore aware that there are special relations between the goods that will change our consumption attitudes. This is especially true for complementary and substitutions goods:

Definitions (#15.3.661):

D1. Two goods are said to be "complementary goods" if the possession of one and the other provides superior satisfaction to the total satisfaction of both goods if taken isolated ("super-additive"). Thus, there is complementarity between skis and a package on the ski lifts, between a car and gasoline. This can be interpreted by the following indifference curve:

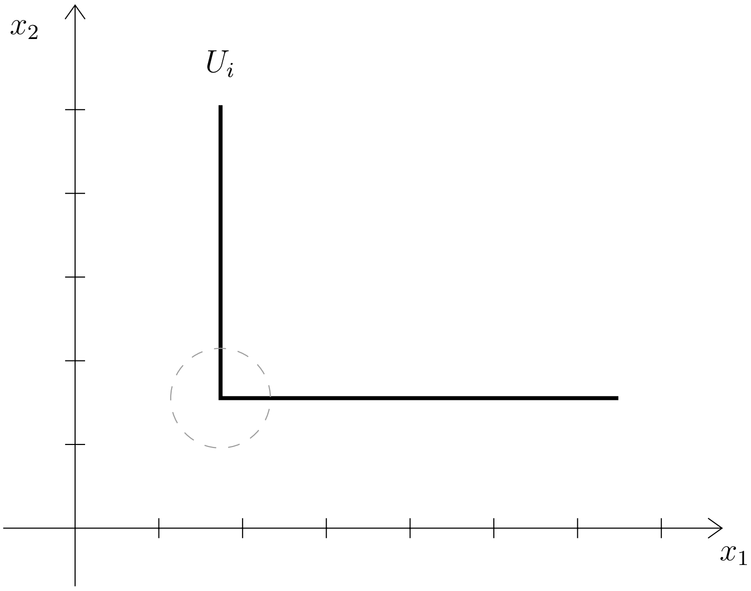


Figure 15.89: **Utility super-additivity visual representation**

Indeed, for the couple $\{\text{car}, \text{petrol}\}$ each curve has respectively a minimum below which we can not go down so that the couple maintains its consumer interest as such (it is not worthwhile to buy a car if the utility of gasoline tends to zero).

D2. Two goods are said to be "substitutable goods" if we can easily replace one by the other, for example in the case of or prices increasing or decreasing. Tea and coffee are substitutable goods because without one, we often defer to the other one. This is even more true for two brands of the same drink (Pepsi® and Coca-Cola®). The mad cow crisis is also a good indicator of the substitutability of meat products, with a postponing consumption on poultry and lamb. This can be interpreted by the following indifference curve:

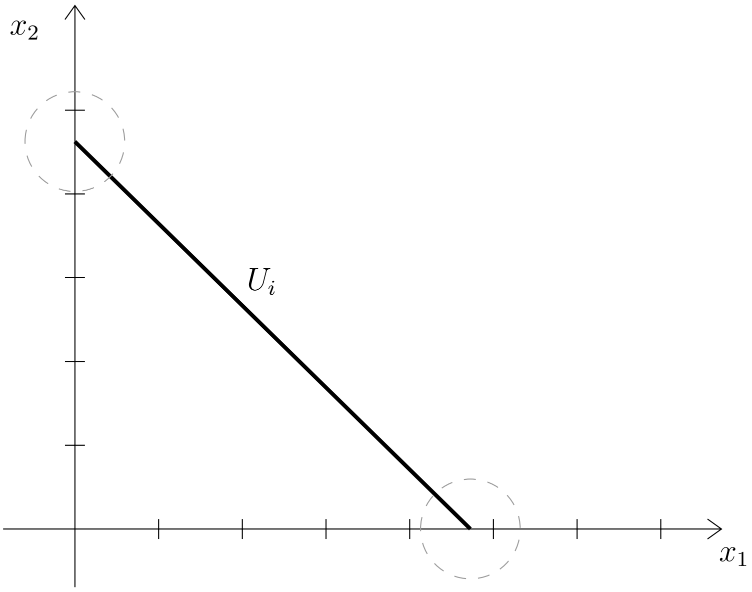


Figure 15.90: **Visual representation of substitutable goods utility**

Indeed, the intersection with the respective axes indicates precisely the total possible substitution of one good by the other in the basket.

✍**Example:**

Consider we want to calculate the MRS along the indifference curve of $U=x_1x_2=100$ (this is a hyperbolic function). We have proved that MRS was given by under a given assumption that:

$$\text{MRS}=-\frac{\mathrm{d}x_1}{\mathrm{d}x_2}=\frac{\dfrac{\partial U}{\partial x_2}}{\dfrac{\partial U}{\partial x_1}}=\frac{\dfrac{\partial x_1x_2}{\partial x_2}}{\dfrac{\partial x_1x_2}{\partial x_1}}=\frac{x_2}{x_1} \tag{15.3.160}$$

Therefore, for the three points A,B,C of respective coordinates:

$$A(5,20);B(10,10);C(20,5) \tag{15.3.161}$$

we found the respective MRS values:

$$\text{MRS}_A=4,\text{MRS}_B=1,\text{MRS}_C=0.25 \tag{15.3.162}$$

These values express the equivalences between the goods 2 and 1 for marginal changes in the quantities of these goods. Thus at point A, to keep the utility level of 100, the consumer is willing to abandon the good 2 for 1 to increase its consumption of good 1 by a factor of 4. At point B the equivalence between the two goods is in a ratio of 1 to 1, etc.

Remark

The concept of indifference curves was developed by Vilfredo Pareto and others in the first part of the 120th century (holocene calendar). The use of this concept has allowed the economic analysis to use this concept of preferences in determining the choice rather than the concept of utility that has the problem of not being able to be measured objectively.

15.3.5 Net gain/loss opposite feedback model

Let us now consider, regardless of the preference theory, a monopoly model with perfect information for a primary need. We denote by $D(t)$ the demand in the market and $O(t)$ the offer. We assume then a infinitely small change in demand depending on the time proportional to the demand (in absence of offer) that is to say the demand follows an exponential law:

$$\frac{\mathrm{d}D}{\mathrm{d}t}=rD(t)\Rightarrow D(t)=c_1^{te}e^{rt} \tag{15.3.163}$$

and for the offer, in the absence of demand (!), is a decreasing exponential law:

$$\frac{\mathrm{d}O}{\mathrm{d}t}=-mO(t)\Rightarrow O(t)=c_2^{te}e^{-mt} \tag{15.3.164}$$

Customers (demand) and suppliers (offers) are interacting. To quantify the contribution between groups, we consider the offer on the assumption that its value or intensity is a function of probability of meeting between the consumer and the supplier and will be proportional to the multiplication of the percentage of supply and demand $O\cdot D$.

The fact of discovering a new product does not have the same effects on both groups (consumers/suppliers). First, of course, each offer acquired by a supplier is a net gain for the second (supplier) and will be assumed as a net loss for the first (consumer). Therefore, if the effect of interactions is assumed as being proportional to $O\cdot D$ the signs of influence of interaction differ by:

$$\begin{aligned}\frac{\mathrm{d}D}{\mathrm{d}t}&=rD-cDO\\ \frac{\mathrm{d}O}{\mathrm{d}t}&=-mO+bDO\end{aligned} \tag{15.3.165}$$

As we can see this model is strongly inspired by the famous "Lotka-Voltera model" (see section Population Dynamics page 7350).

Therefore this model could be used to explain economics crash or the fact that some trends product have a cyclic demand (for sure with periods that need to be determined) or that we can see in supply chain management small periodic fluctuations.

Before going further, let us look to the values for which the derivatives vanish (which will give us in fact the equilibrium point between supply and demand because the two relations can therefore be written as equal):

$$\begin{aligned}r'D'-cD'O'&=0\\ -mO'+bD'O'&=0\end{aligned} \tag{15.3.166}$$

Thus:

$$\begin{aligned}(r'-cO')D'&=0\\ (bD'-m)O'&=0\end{aligned} \tag{15.3.167}$$

A trivial solution is the "solution of inexistence" (also sometimes named "critical solution") given by:

$$D_e' = O_e' = 0 \tag{15.3.168}$$

Otherwise, we also have as a possible solution:

$$\begin{aligned} r = CO' \rightarrow O' = \frac{r}{c} \\ bD' = m \rightarrow D' = \frac{m}{b} \end{aligned} \tag{15.3.169}$$

Now we normalize these equations by writing (so they are dimensionless):

$$\begin{aligned} x = \frac{D}{D'} = \frac{b}{m}\frac{1}{D'} := \frac{b}{m}D \rightarrow D = \frac{m}{b}x \\ y = \frac{O}{O'} = \frac{c}{r}\frac{1}{O'} := \frac{c}{r}O \rightarrow O = \frac{r}{c}y \end{aligned} \tag{15.3.170}$$

with this normalization, the model can be rewritten:

$$\begin{aligned} \frac{m}{b}\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{mr}{b}x - \frac{cmr}{xy} \\ \frac{r}{c}\frac{\mathrm{d}y}{\mathrm{d}t} = \frac{bmr}{bc}xy - \frac{mr}{c}y \end{aligned} \tag{15.3.171}$$

By rearranging the coefficients, the system can finally be written (excluding the solution of inexistence):

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} = rx(1-y) \\ \frac{\mathrm{d}y}{\mathrm{d}t} = my(x-1) \end{aligned} \tag{15.3.172}$$

for which derivatives vanish at the point $(1,1)$, which can be assimilated to the "Say equilibrium".

The discreet plot of this system of equations (in which we recognize a logistic term as seen in the section of Populations Dynamics) gives us with $m = r = b = c = 1$ and initial conditions $(x_0, y_0) = (4,1)$:

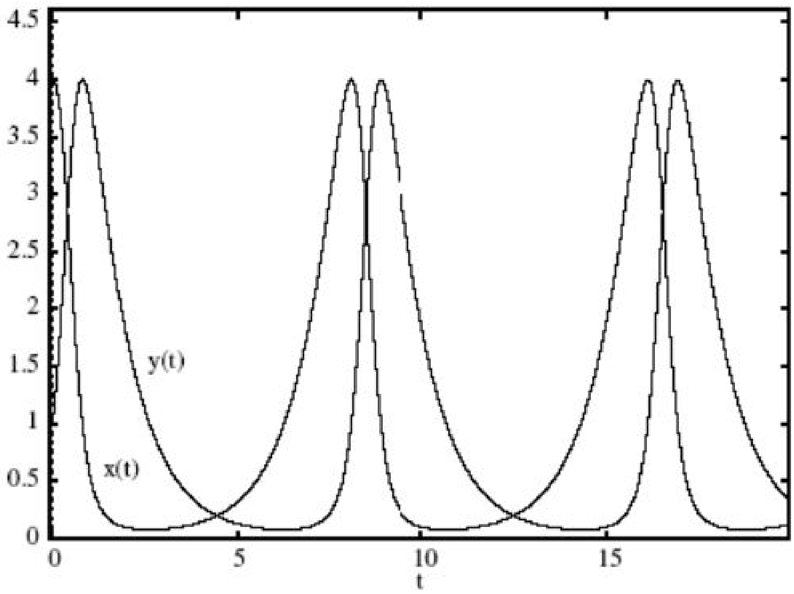


Figure 15.91: **Cycle of offer and demand**

We see as the real market seems show to us, the cycles of supply/demand (some outdated products become fashion again) for which, using statistics, we must determine the initial conditions in order to know the period of time. We note also on that figure that offers is always a little behind demand in this model (which is not always true because we know that many times that offer created demands). Unfortunately this model does not have a damping or multiplier factor suggesting a potential for improvement.

The previous figure represent the variables x and y as a function of time. However, what can be interesting for a scientist (or an economist) is the representation of y in terms of x and vice versa. Thus, we obtain for the same initial conditions m, r, b, c and for the various initial values of (x_0, y_0) (the Offer is in ordinate and the Demand on the abscissa):

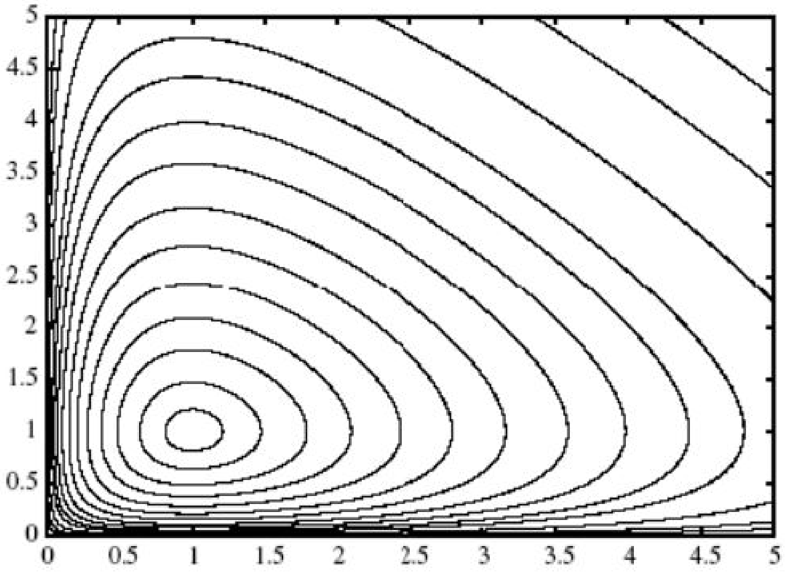


Figure 15.92: **Representation in phase space offer/demand cycle**

The above figure is very interesting to interpret when you follow a path in the counter-clockwise direction.

Thus we see (in the phase space representation) that for fixed initial conditions, the system is periodic and has trivially an equilibrium point at:

$$(x,y)=(1,1) \tag{15.3.173}$$

that correspond to the points where:

$$\frac{\mathrm{d}x}{\mathrm{d}t}=\frac{\mathrm{d}y}{\mathrm{d}t}=0 \tag{15.3.174}$$

Finally, we have two pairs of equilibrium points (that is almost trivial, looking at the system of equations):

$$(x_1^e,y_1^e)=(0,0) \qquad (x_2^e,y_2^e)=(1,1) \tag{15.3.175}$$

The question that will naturally arise is the "real" direction of rotation (representation) of the phase plane. Thus, representing the directions using a vector field, we get the representation which is effectively counter-clockwise:

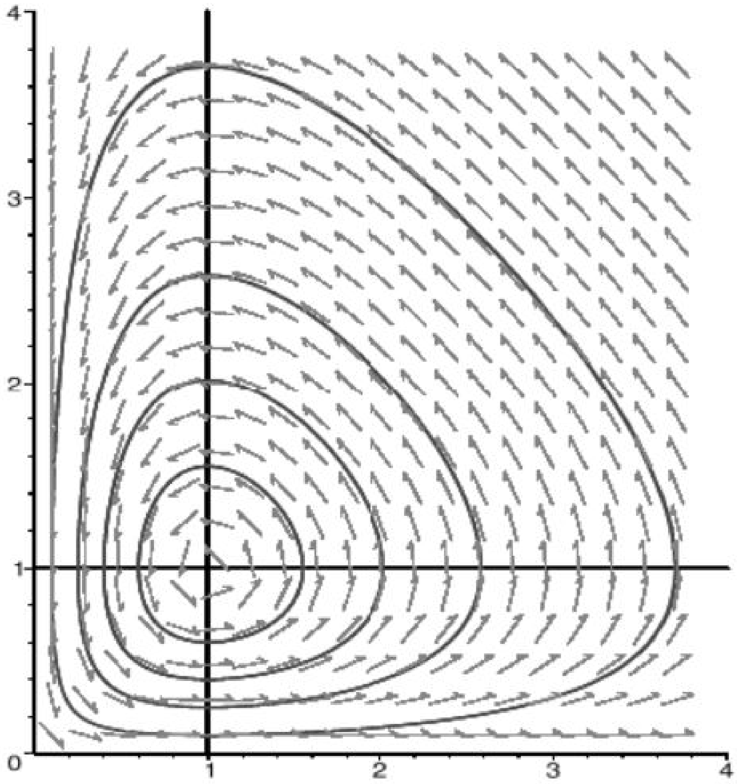


Figure 15.93: Rotation direction offer/demand phase space

To know in which direction we are going in the phase space at a given time, it suffices to know the derivative $\mathrm{d}y/\mathrm{d}x$ (or vice versa $\mathrm{d}x/\mathrm{d}y$). Therefore we have:

$$\frac{\mathrm{d}y}{\mathrm{d}x}=\frac{\mathrm{d}y}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}x}=\frac{\mathrm{d}y}{\mathrm{d}t}\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^{-1}=\frac{my(x-1)}{rx(1-y)} \tag{15.3.176}$$

That said, we see well on the phase diagram with vector field as vectors that it comes a time in the cycle of this model where the offer is very high for low demand. So the mathematical model (theoretical) clearly explains what can be a priori counter-intuitive to many humans (offer creates the demands).

However, we can (must) ask ourselves the question of what happens after a small perturbation around the equilibrium point (which is of utmost importance in economy).

So we have the following Lotka-Volterra system in equilibrium:

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}t}&=rx(1-y)\\ \frac{\mathrm{d}y}{\mathrm{d}t}&=my(x-1)\end{aligned} \tag{15.3.177}$$

By putting an infinitely small perturbation, it is will be written:

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}t}&=r(x+x^*)(1-(y+y^*))\\ \frac{\mathrm{d}y}{\mathrm{d}t}&=m(y+y^*)((x+x^*)-1)\end{aligned} \tag{15.3.178}$$

Neglecting the quadratic xy terms, we obtain:

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}t}&\cong r\left[(1-y^*)x+(1-y-y^*)x^*\right]\\ \frac{\mathrm{d}y}{\mathrm{d}t}&\cong -m\left[(1-x^*)y+(1-x-x^*)y^*\right]\end{aligned} \tag{15.3.179}$$

1. We focus first on the study near the point of extinction $(0,0)$, this is why we can neglect the quadratic terms xy but the expression remain too complicated. So, always considering we are near the point of extinction $(0,0)$, we consider a bit unfairly but cleverly (otherwise we still could not solve the problem analytically) the following approximations:

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}t}&\cong r\left[x-\underbrace{y^*x}_{=0}+x^*-\underbrace{yx^*}_{=0}-\underbrace{y^*x^*}_{=0}\right]\underset{x\gg x^*}{\cong} r(x+x^*)\cong rx\\ \frac{\mathrm{d}y}{\mathrm{d}t}&\cong -m\left[y-\underbrace{x^*y}_{=0}+y^*-\underbrace{xy^*}_{=0}-\underbrace{x^*y^*}_{=0}\right]\cong -m(y+y^*)\underset{y\gg y^*}{\cong} -my\end{aligned} \tag{15.3.180}$$

   Therefore:

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}t}&\cong rx=rx(t)\\ \frac{\mathrm{d}y}{\mathrm{d}t}&\cong -my=-my(t)\end{aligned} \tag{15.3.181}$$

This shows us for the system in equilibrium, close to the inexistence point, the offer (supply) decreases exponentially (power law) as demand increases exponentially:

$$\begin{aligned} x(t) &= c_1^{te} e^{rt} \\ y(t) &= c_2^{te} e^{-mt} \end{aligned} \qquad (15.3.182)$$

This makes economic sense: when there is little offer (respectively demand), the demand (respectively offer) is growing when as far as demand increases, the offer grows and converge increasingly near to the demand (nature...).

Remark

In the literature we find sometimes the "$-$" sign in the top or bottom in the previous equations. In reality, the position of the "$-$" sign is not important because it is just the starting choice in the system dynamics.

2. Close to the equilibrium point $(1,1)$ we start from:

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= rx(1-y) \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= my(x-1) \end{aligned} \qquad (15.3.183)$$

And we put $x := 1+x, y := 1+y$ where x, y are now small perturbations near the point $(1,1)$. Therefore we have:

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= r(1+x)(1-(1+y)) = -r(1+x)y \cong -ry \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= m(1+y)((x+1)-1) = m(1+y)x \cong mx \end{aligned} \qquad (15.3.184)$$

(hop! we change the position of the "-" sign on purpose to show that this is only a starting choice !!!):

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &\cong -ry \\ \frac{\mathrm{d}y}{\mathrm{d}t} &\cong mx \end{aligned} \qquad (15.3.185)$$

To solve this system, let us differentiate the first equation once again:

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} \cong -r\frac{\mathrm{d}y}{\mathrm{d}t} \qquad (15.3.186)$$

and by injecting in it $\mathrm{d}y/\mathrm{d}t$:

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} \cong -rmx \qquad (15.3.187)$$

So we get a small second order differential equation (see section Differential and Integral Calculus page 1610). Whose typical simple solution is:

$$x = x_0 e^{-\lambda t} \tag{15.3.188}$$

By injecting this solution into the differential equation, we obtain after simplification of exponentials a simple polynomial of the second degree (see section Calculus page 1368):

$$\lambda^2 x_0 + rmx_0 = 0 \tag{15.3.189}$$

Whose solution is trivial:

$$\lambda_{1,2} = \pm \mathrm{i}\sqrt{rm} \tag{15.3.190}$$

Thus, the general solution of the differential equation is the linear combination of the two special solutions we get:

$$x(t) = x_0 \left(e^{+\mathrm{i}\sqrt{rm}} + e^{-\mathrm{i}\sqrt{rm}} \right) \tag{15.3.191}$$

But we therefore have:

$$\frac{\mathrm{d}x}{\mathrm{d}t} \cong -ry \Leftrightarrow y(t) = -\frac{1}{r}\frac{\mathrm{d}x}{\mathrm{d}t} \tag{15.3.192}$$

Therefore, knowing $x(t)$ we obtain easily:

$$x(t) = -x_0 \sqrt{\frac{m}{r}} \left(\mathrm{i}e^{+\mathrm{i}\sqrt{rm}} - \mathrm{i}e^{-\mathrm{i}\sqrt{rm}} \right) \tag{15.3.193}$$

Now let us use the Euler formula (see section Numbers page 251):

$$e^{\mathrm{i}x} = \cos(x) + \mathrm{i}\sin(x) \tag{15.3.194}$$

Thus we have:

$$x(t) = x_0 \left[\cos\left(-\sqrt{rm}t\right) + \mathrm{i}\sin\left(-\sqrt{rm}t\right) + \cos\left(\sqrt{rm}t\right) + \mathrm{i}\sin\left(\sqrt{rm}t\right) \right] \tag{15.3.195}$$

and as (see section Trigonometry page 2514) $\cos(x) = \cos(-x), -\sin(x) = \sin(-x)$ then we have:

$$x(t) = 2x_0 \cos(\sqrt{rm}t) \tag{15.3.196}$$

and similarly, we obtain:

$$\begin{aligned} y(t) &= -x_0 \sqrt{\frac{m}{r}}\mathrm{i} \left[-\left(\cos\left(-\sqrt{rm}t\right) + \mathrm{i}\sin\left(-\sqrt{rm}t\right)\right) + \cos\left(\sqrt{rm}t\right) + \mathrm{i}\sin\left(\sqrt{rm}t\right) \right] \\ &= -x_0 \sqrt{\frac{m}{r}}\mathrm{i} \left[-\left(\cos\left(\sqrt{rm}t\right) - \mathrm{i}\sin\left(\sqrt{rm}t\right)\right) + \cos\left(\sqrt{rm}t\right) + \mathrm{i}\sin\left(\sqrt{rm}t\right) \right] \\ &= -x_0 \sqrt{\frac{m}{r}}\mathrm{i} \left[-\cos\left(\sqrt{rm}t\right) + \mathrm{i}\sin\left(\sqrt{rm}t\right) + \cos\left(\sqrt{rm}t\right) + \mathrm{i}\sin\left(\sqrt{rm}t\right) \right] \\ &= -x_0 \sqrt{\frac{m}{r}}\mathrm{i}2\mathrm{i}\sin\left(\sqrt{rm}t\right) = 2x_0\sqrt{\frac{m}{r}}\sin\left(\sqrt{rm}t\right) \end{aligned} \tag{15.3.197}$$

Thus, around the equilibrium point $(1, 1)$ with sufficiently small perturbations to validate linearisation the system oscillate as ellipses (or circles) whose axes are defined by the two equations above.

We can get the graphs above with Maple™ 4.00b (this is a nice example of application of this software):

```
>restart:  with(plots):  with(DEtools):
>rate_eqn1:= diff(h(t),t)=(0.1)*h-(0.005)*h*(1/60)*u;
rate_eqn2:=diff(u(t),t)=(0.00004)*h*u-(0.04)*u;vars:= [h(t), u(t)];
>init1:=[h(0)=2000,u(0)=600]; init2:=[h(0)=2000,u(0)=1200];
init3:=[h(0)=2000, u(0)=3000];domain := 0 ..  320;
>L:= DEplot(rate_eqn1, rate_eqn2, vars, domain,init1, stepsize=0.5,
scene=[t, u], arrows=NONE):
>H:= DEplot(rate_eqn1, rate_eqn2, vars, domain,init1 , stepsize=0.5,
scene=[t, h], arrows=NONE):
```

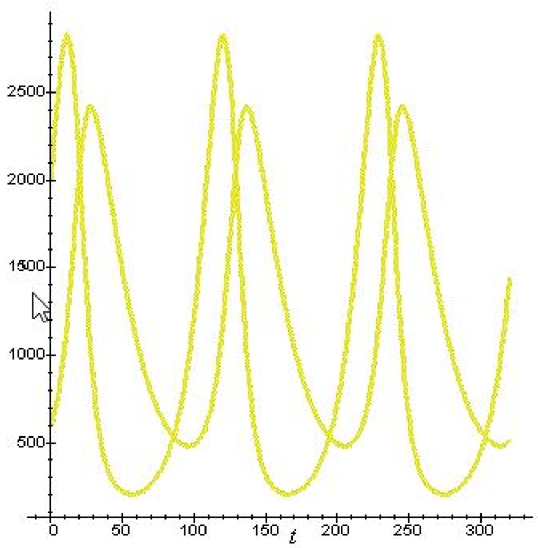


Figure 15.94: **Lotka-Volterra offer/demand model periodic plot in Maple™ 4.00b**

```
>DEplot(rate_eqn1, rate_eqn2, vars, t= 0 ..  160, init1, init2, init3,
stepsize=0.5, scene=[h,u], title='Demande vs.  60 * Offer for t = 0 ..
160', arrows=slim);
```

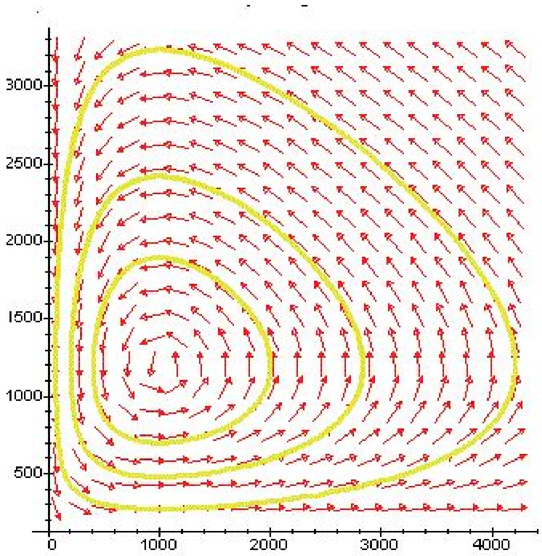


Figure 15.95: **Lotka-Volterra offer/demand phase space model plot in Maple™ 4.00b**

This model, however is imperfect because it takes into account only a disgruntled monopoly with net loss and perfect information. The fact of considering the population constant is not too bothersome but strictly speaking we should add a logistic term in the initial equations. There is still work to do...

15.3.6 Capitalization and Actuarial

Definition (#15.3.662): The "capitalization" is the field of financial mathematics which calculates future values based on present values, while the "actuarial calculation" consist to determine how much should be loaned to get a given amount fixed in advance (the opposite of capitalization).

In a dynamic market, economic agents can lend or borrow capital in return for which they receive or respectively pay a periodic interest. This interest is justified by the risk-taking that takes the creditor (the lender's of capital) in case of non-repayment of all or a share of the initial capital that must repay the debtor (the one who must repay the borrowed capital). Alternatively, in the point of view of economic market, loans allow some economic agents to set up goods by betting on the fact that either they will create the offer or that the offer will come by itself but wishing be ahead of the concurrence.

Remark

When credit is contracted with an economic agent that is non solvable financial analysts speak about "NINJA" loan for "No Income, No Job and no Assets".

When a capital is lent (or borrowed, it depends from the point of view...) in order to increase market dynamics (the amount of goods in circulation on a given duration) then we talk of "financial asset", this to make clear that capital participates in the activities of the economy.

Definitions (#15.3.663):

D1. We name "yield of a loaned financial asset" the relative progress ratio given by:

$$r_f\% = \frac{A_{t+\Delta t} - A_t}{A_t} \cdot 100 \tag{15.3.198}$$

D2. We name "arithmetic return on investment" the relative progress ratio:

$$\text{ROI}^{\text{Arith.}} = \frac{V_f - V_i}{V_i} \tag{15.3.199}$$

where V_i is the initial value of the investment and V_f its final value.

It follows from this latter definition that if an investment has reported 5% the first year and carried a net loss of 2% the second year, the "arithmetic return on investment" is then:

$$\overline{\text{ROI}}^{\text{Arith.}} = \frac{1}{n}\sum_{k=1}^{n} \text{ROI}_k^{\text{Arith.}} = \frac{\text{ROI}_1^{\text{Arith.}} + \text{ROI}_2^{\text{Arith.}}}{2} = \frac{5\% + (-2\%)}{2} = 1.50\% \tag{15.3.200}$$

But it is wrong to use the arithmetic average for this type of situation because the final sum obtained after two years would be mathematically:

$$V_n^f = V_0 \prod_{k=1}^{n} (1 + \text{ROI}^{\text{Arith.}}) \tag{15.3.201}$$

which then gives the previous example:

$$V_2^f = V_0(1+5\%)(1+(-2\%)) \tag{15.3.202}$$

So the real average return is by definition the "geometric return on investment" as:

$$V_2^f = V_0(1+5\%)(1+(-2\%)) = V_0(1+\text{ROI}^{\text{Geom.}}) \tag{15.3.203}$$

that is to say, it is simply a geometric mean (see section Statistics page 565) . It comes then:

$$\text{ROI}^{\text{Geom.}} = (1+5\%)(1+(-2\%)) - 1 = 2.90\% \tag{15.3.204}$$

which is obviously much different from the ROI arithmetic average obtained above!

Remark

We say of an asset that it has a "risk-free return" or "risk-free rate of return" if its future value (respectively it's corresponding rate) is perfectly known.

Given an asset that can have the future optimistic return $r_{f,1}$ with probability p_1 and the pessimistic value $r_{f,2}$ with a probability p_2 or other values r_{fn} with probabilities p_n then expected mean of return is obviously given by:

$$\text{E}(r_f) = \sum_{i=1}^{n} r_{fi}p_i \tag{15.3.205}$$

that the monetary sum is of the type active/passive, the types of applicable yields can be identical or variable. There exists, however, classic cases that are well known and non-stochastic. For their study, let us define some variables:

- C_0 represents the initial capital or more technically the "actual value" (AV) or "present value" (PV).
- C_n represents the final capital or "capitalized value" (CV) or "future value" (FV) after n time periods.
- $t\%$ is the rate most technically named "effective rate".
- I_n represents the "interest" obtained after n periods (horizon) by the current value

Let us add also the relation:

$$f_c = 1 + t\% \tag{15.3.206}$$

named "capitalization factor".

Definition (#15.3.664): We define the "interest" as the remuneration of a capital (the amount of money) loaned or invested for a given time. The interest can be paid at once or periodically if the duration of the loan or investment is for a long time. The interest may be payable in advance (praenumerando) or at the end of the period (postnumerando). The interest is based on the duration of the loan (or investment), of the borrowed (or paid) capital that we also name the "principal" and also from the "rate" of interest charged. The period over which the interest is calculated is in general the year, but it can be shorter: semester, quarter, month, day, hour, minute, seconds, etc.

Remark

In a text, the interest rate is normally expressed in percentage % but in financial calculations, it is customary to calculate or write it in decimal form.

The reader should also be aware the in the finance industry, we use many times the "basis points" (BPS) as common unit of measure for interest rates and other percentages. One basis point is equal to $1/100$th of 1%, or 0.01% (0.0001), and is used to denote the percentage change in a financial instrument. The relation between percentage changes and basis points can be summarized as follows: 1% change $=$ 100 basis points, and $0.01\% = 1$ basis point.

Basis Points	**Percentage Terms**
1	0.01%
10	0.1%
50	0.5%
100	1%
1000	10%
$10,000$	100%

Table 15.35: **Basis points (BPS) and percentage equivalence**

15.3.6.1 Dates Intervals

In determining the amount of interest of a loan (or investment), it is first necessary to know the duration of the latter or the dates defining the periods of payment of an obligation, i.e. the "settlement date".

The calculation of dates and durations is the first step to learn in actuarial mathematics. If some software used in the calculation of the period the calendar with the civil year (365 days according to the Gregorian calendar), others are based on the commercial year (360 days), which was the case of most European banks (it is to their advantage financially to make the choice of this latter...) before the arrival of the calendar for the Eurozone.

Remarks

R1. In financial markets, there is as far as we know only one single convention for determining a date range to calculate a duration: the first day (departure date) is included in the period. The last day (end date or settlement date) is excluded from the period. Thus the period from 15 to 25 June has 10 days.

R2. In the context of this textbook, who wish to have the more rigorous possible approach on treated topics, we will not dwell on the aberrant German, European or American 30/360, methods (we can then do every country on the planet then... and refer to the calculations of dates functions of a spreadsheet software like Microsoft® Excel®...) to focus only on 365 days method (the exact base system) that is, and remains, the most natural counting system to use as it includes the months with 28, 29, 30 or 31 days.

R3. Note that in regard to savings accounts, banks are based on a system of "fortnight" (two weeks/half a month), and they therefore estimate that there are 24 fortnights a year.

We must therefore in this system of the exact basis know how to calculate the number of days between two dates D_1, D_2 given by the calculation $D_2 - D_1$ each such date being encrypted in standardized format $yyyy - mm - dd$ (year-month-day compliant with ISO ISO 8601) or under non-standardized $d.m.y$ format (day.month.year).

Definitions (#15.3.665):

D1. The Gregorian Calendar is defined as having 12 months.

D2. The months:

$$\text{january}(1), \text{march}(3), \text{may}(5), \text{july}(7), \text{august}(8), \text{october}(10), \text{december}(12) \quad (15.3.207)$$

are 31-day months and the months of:

$$\text{april}(4), \text{june}(6), \text{september}(10), \text{november}(11) \quad (15.3.208)$$

have 30 days.

D3. February is a special case to correct the fact that the civil year calendar of 365 days does not correspond quite the orbital period of the Earth around the Sun (actually a G2V-type star, aka yellow dwarf) that is about 365.25 days... Thus, all the years that are multiples of 4 or 400 are leap years (February has 29 days instead of 28) but the years that are divisible by 100 are not leap years.

✍**Examples:**

E1. 11992, 11996, 12004, 12008 are leap (holocene) years.

E2. The years 11900, 12100, 12200, 12300 are not leap (holocene) years (because divisible by 100).

E3. The years 11600, 12000, 12400, 12800 are leap (holocene) years because even if they are divisible by 100, they are multiples of 400.

These definitions and examples being given, imagine now date given following the above rules. The number of days since year 1 is therefore (as seen in the section of Numerical Methods $[x]$ will be the integer part of x):

$$\boxed{D(d,m,y) = 365(y-1) + \left[\frac{y-1}{4}\right] - \left[\frac{y-1}{100}\right] + \left[\frac{y-1}{400}\right] + 31(m-1) + d} \quad (15.3.209)$$

This relation is logically deduced as follows for the dates where $m \leq 2$ (that is to say only for the first two months of the year including the problematic of February):

1. We have $365(y-1)$ because given a the number of calendar days since the year 1 is 365 times the number of years y has decreased by one unit as the current year is not over.
2. The same remark applies for months with $31(m-1)$: considering months of 31 days, the current month is not taken into account (this is why we have $m-1$). The correction for the months with 28 or 30 days is taken into account in the case where $m>2$ that we will see later below.
3. Logically, we add d (which contains all the information as to know whether the year y is leap year or not) to the sum of the previous two terms.
4. The terms:
$$\left[\frac{y-1}{4}\right]-\left[\frac{y-1}{100}\right]+\left[\frac{y-1}{400}\right]$$
give the number of 29th February between year 1 and taking into account all leap years that occur all the multiples of 4 and 400 years except for years that are multiples of 100.

If $m>2$, we must use the following relation:

$$\boxed{D(d.m.y)=365(y-1)+\left[\frac{y}{4}\right]-\left[\frac{y}{100}\right]+\left[\frac{y}{400}\right]+31(m-1)+d-[0.42M+2+0.5]} \tag{15.3.210}$$

This relation is deducted in the same way as the previous one except that certain terms in the numerator are not reduced by one unit because having $m>2$, it is necessary, for these terms, to consider the current year in the calculation.

The last term $[2+0.5+0.42M]$ is here to fix the fact that every month does not have 31 days (at least for the Gregorian calendar...). To get it, we build the following table (the third column gives the offset in days compared to the case where all the months have 31 days):

Month	Month number n	Offset Δ
march	3	3
april	4	4
may	5	4
june	6	5
july	7	5
august	8	5
september	9	6
october	10	6
november	11	7
december	12	7

Table 15.36: **Monthly offset Δ in days**

A simple linear regression (see section Numerical Methods page 6324) gives:

$$\Delta(n)=0.4242n+2.018 \tag{15.3.211}$$

By taking the integer value and verifying that the selected function is correct, we finally get (taking an accuracy of two decimal places):

Month	Month number n	Offset Δ	$\Delta(n)$	Offset $[\Delta(n)]$
march	3	3	3.26	3
april	4	4	3.68	4
may	5	4	4.1	4
june	6	5	4.52	5
july	7	5	4.94	5
august	8	5	5.36	5
september	9	6	5.78	6
october	10	6	6.2	6
november	11	7	6.62	7
december	12	7	7.04	7

Table 15.37: **Comparison of the regression function with the target**

15.3.6.2 Rates Equivalence

Let us now briefly focus on rates calculation before directly addressing the various calculations of interests types.

Definition (#15.3.666): The "proportional rate $t_p\%$" is bringing to a same capital, during the same period, the same "simple interest" (see the definition of the simple interest further below), and given by the relation:

$$\boxed{t_p\%=\frac{t\%}{n}} \tag{15.3.212}$$

If the proportional rate is calculated on the basis of a year, then we speak of "annualized yield rate"; if calculated on the basis of a month, then we speak of "monthly-rate of return".

✍**Example:**

Calculate the monthly proportional rate $t_p\%$ (i.e.: the monthly rate of return) of an annual rate $t\%$ of 12%:

$$t_p\%=\frac{t\%}{n}=\frac{12\%}{12\,\text{months}}=1\% \tag{15.3.213}$$

Definition (#15.3.667): The "equivalent rate $t_e\%$" is bringing to a same capital, during the same period, the same "compound interest" (see definition of compound interest below), and given by the relation:

$$\begin{aligned}&(1+t_e\%)^m=(1+t\%)^n\\ &\Rightarrow(1+t_e\%)=\sqrt[m]{(1+t\%)^n}=(1+t\%)^{\frac{n}{m}}\\ &\Rightarrow t_e\%=(1+t\%)^{\frac{n}{m}}-1\end{aligned} \tag{15.3.214}$$

and vice versa:

$$\boxed{t\% = (1 + t_e\%)^{\frac{m}{n}} - 1} \tag{15.3.215}$$

✍**Example:**

Monthly equivalent rate $t_e\%$ to an annual rate of 12% (result truncated at ten-thousandth):

$$t_e\% = (1 + 12\%)^{\frac{1}{12}} - 1 \cong 0.9488\% \tag{15.3.216}$$

the reverse procedure would therefore be to calculate the annualized rate and then we see that a monthly rate of 1% annualized would be not exactly 12%!!

15.3.6.3 Simple Interest

Definition (#15.3.668): The "simple interest" is defined by the relation (see above for the definition of notations):

$$I_n = C_0 t\% n \tag{15.3.217}$$

implying a capitalization (final value):

$$\boxed{C_n = C_0 + I_n = C_0(1 + nt\%)} \tag{15.3.218}$$

It is simply the interest that is calculated each period on the sole basis of the capital loaned or borrowed originally.

From the latter we get obviously the also following famous relation:

$$C_0 = \frac{C_n}{1 + nt\%} \tag{15.3.219}$$

And another one that is famous is to determine I_n if we don't know C_0. To get this we do a little bit of elementary algebra:

$$I_n = C_n - C_0 = C_n - \frac{C_n}{1 + nt\%} = C_n\left(1 - \frac{1}{1 + nt\%}\right) = C_n\frac{nt\%}{1 + nt\%} \tag{15.3.220}$$

In the field of real estate investments, with $n = 1$ we have the interest rate:

$$t\% = \frac{I_n}{C_o} \tag{15.3.221}$$

that is named quite obviously "capitalization rate" or simply "cap-rate", and I_n that is abbreviated NOI for "net operating income" (the NOI of a rental property is its rent minus the expenses paid for its maintenance like property management, property taxes, insurance, etc.).

Remarks

R1. It is very easy from the knowledge of three of the four parameters of the above relation to find the fourth one. As this is a simple equation of the first degree, we will not dwell on this kind of elementary algebra exercise.

R2. A specificity of the simple interest is to be proportional to the duration of the investment. If interest e.g. over a year is 12%, the "equivalent rate $t_e\%$" of an identical placement for 12 months will be 1% per month. This property is not true for compound interest as we will see afterwards.

R3. For savings accounts, we have already mentioned that some financial institutions use as fifteen time period (24 periods in the year consisting of fictive 30-day months). So to calculate the annual interest at each fortnight, they take the lowest balance on the account during the fortnight and calculate simple interest on a reported rate of 24 fortnights a year and defer the result obtained in annual closure of the account at the end of the year (they are not crazy...).

Moreover, if more simple interests investments are made simultaneously for different periods and at different rates, we may calculate the average T of all of these investments.

If we note C_t the investment number k, I_k the interest rate of investment number k, n_k the duration (number of periods) of the investment and N the number of investments, we obtain the following weighted arithmetic mean (see section Statistics page 548) as follows:

$$T=\frac{C_1I_1n_1+C_2I_2n_2+\ldots+C_NI_Nn_N}{C_1n_1+C_2n_2+\ldots+C_Nn_N}=\frac{\displaystyle\sum_{k=1}^{N}C_kI_kn_k}{\displaystyle\sum_{k=1}^{N}C_kn_k} \tag{15.3.222}$$

✍**Example:**

We seek the average annual rate of three placements $\{1000.\text{-}; 90 \text{ days}; 3\%\}$, $\{2000.\text{-}; 120 \text{ days}; 4\%\}$, $\{3000.\text{-}; 170 \text{ days}; 5\%\}$. Then we have:

$$T=\frac{1000\cdot 3\%\dfrac{90}{360}+2000\cdot 4\%\dfrac{120}{360}+3000\cdot 5\%\dfrac{170}{360}}{1000\dfrac{90}{360}+2000\dfrac{120}{360}+3000\dfrac{170}{360}}=0.045=4.5\% \tag{15.3.223}$$

15.3.6.3.1 Discounts

Still regarding to the simple interest, we can come back to a concept we talked about at the beginning of this section: the discount. This concept is used both in the traditional trade as in the financial markets for Treasury bills (see further below).

Let us recall that the discount is a rebating granted to a buyer by a seller in order to induce him to pay quickly before n units (periods) of time (that is the number of units n that matters!). A buyer should in principle take advantage of this discount (which can be seen as a credit). Otherwise, it is as if the purchaser implicitly borrowed from a bank for a given period (n units).

Historically, the discount is the withholding that does a banker on a promise of payment that has a merchant (promise signed by a buyer for a certain date), that the banker pays before maturity (thus as an advance).

Let us see this:

Let us note C_0 the present value discount included, C_n the amount without discount named "nominal value", n the duration relative to the time scale of the discount rate, $t\%$ the discount rate and I the implicit simple equivalent interest in case of renunciation of the discount.

We now have the following trivial relations:

$$C_0 = C_n - t\%C_n = C_n - I = C_n(1-t\%) \tag{15.3.224}$$

And nothing prevents us from writing that the amount without discount can be obtained by simple interest of the form:

$$C_n = C_0(1+nI) \tag{15.3.225}$$

from the value with discount. The simple implicit interest on the present value being then trivially:

$$I = \frac{C_n - C_0}{nC_0} \tag{15.3.226}$$

Therefore, it comes by substitution:

$$I = \frac{C_n - C_0}{nC_0} = \frac{C_n - C_n(1-t\%)}{nC_n(1-t\%)} = \frac{1-(1-t\%)}{n(1-t\%)} = \frac{t\%}{n(1-t\%)} \tag{15.3.227}$$

We notice then that it is sufficient to know the discount rate $t\%$ granted (often annualized...) and the renunciation period n to determine the simple equivalent rate of the granted credit.

If the rate $t\%$ communicated from the start is not the rate corresponding to the entire duration of n (as is the case above) but the rate of the corresponding time unit, then we have the following relation:

$$I = \frac{t\%}{n(1-t\%)} = \frac{nt'\%}{n(1-nt'\%)} = \frac{t'\%}{(1-nt'\%)} \tag{15.3.228}$$

✍**Example:**

Let us calculate the implicit simple rate i of a discount of 1% to 10 days or net at 30 days:

$$I=\frac{0.01}{20(1-0.01)}\cong 0.0505\% \qquad (15.3.229)$$

Thus, the discount if it is not considered, can be seen as being a credit to of 0.0505% daily during 20 days on the sum with discount! This corresponds on a month of 30 days to a simple interest rate of 1.515%.

This calculation method is named "commercial discount" because the calculations are based on the nominal value and not of the current value. However, historically, the calculations were rather similar to the following:

✍**Example:**

On March 19, a banker discount of $t\%=6\%$ (annualized) a buyer contract to a trader of $240,000$.- payable on May 31 (so there are 73 calendar days between the two dates). The discount value is then (on an annual basis of 360 days):

$$\text{Discount}=240,000\cdot\frac{6\%}{360}\cdot 73=2,920.\text{-} \qquad (15.3.230)$$

So the present value of the contract will be $237,080$. The simple daily implicit rate will be given by:

$$C_n=C_0(1+ni)=(C_n-C_nnt\%)(1+ni)\Rightarrow i=\frac{\dfrac{1}{1-nt\%}-1}{n}=\frac{t\%}{1-nt\%} \qquad (15.3.231)$$

Therefore:

$$I=\frac{t\%}{1-nt\%}=\frac{\dfrac{6\%}{360}}{1-\dfrac{6\%}{360}-1}=0.0169\% \qquad (15.3.232)$$

Let's see a last example involving a Treasury bill:

✍**Example:**

A dealer purchase a US Treasury bill with 40 day to maturity at a quote discount rate of 5.5%. The face value is $100,000$.-, we then get for the settlement value:

$$\text{settlement value}=100,000\left(1-5.5\%\frac{40}{360}\right)=99,389.\text{-} \qquad (15.3.233)$$

If necessary here is a summary of some concepts (if it may be useful to the reader):

Treasury Bills	Treasury Bonds	Treasury Notes
	Maturities	
Up to 1 year	20 and 30 years	2 to 10 years
	Investment Term	
Short-term	Long-term	Intermediate-term
	Yield and Risk	
Lowest yield, Least risk	Highest yield, Greatest risk	Balanced yield and risk
	Liquidity	
Highly liquid	Highly liquid	Highly liquid
	Pricing Mechanism	
Sold at discount, No coupon payments	Sold at par, Discount or premium, Make semi-annual interest payments	Sold at par, Discount, or premium, Make semi-annual interest payments

Table 15.38: **Treasury Bills vs Treasury Bonds vs Treasury Notes**

15.3.6.4 Compound Interest

Definition (#15.3.669): The "compound interest" is defined by the relation:

$$I_n = C_0 f_c^n - C_0 = C_0(f_c - 1) \tag{15.3.234}$$

and implies that the final value is given by:

$$\boxed{C_n = C_0 f_c^n = C_0(1+t\%)^n} \tag{15.3.235}$$

We say that the interest rate is "compound" when at the end of each period the interest is added to capital for the calculation of the next period. This relation obviously implies that there is no withdrawal or deposit in all periods.

We also deduce the following trivial relations (see section Calculus page 1350 and Functional Analysis page 2284):

$$C_0 = \frac{C_n}{(1+t\%)^n} \qquad n = \frac{\ln\left(\dfrac{C_n}{C_0}\right)}{\ln(1+t\%)} \qquad t\% = \sqrt[n]{\frac{C_n}{C_0}} - 1 \tag{15.3.236}$$

where the first relation gives the initial value if the final value is known and that there is no withdrawal or deposit during all periods.

Remark

The equivalent function in Microsoft Office Excel 11.8346 are respectively to find C_n, C_0, n, i the function `FV()`, `PV()`, `NPER()`, `RATE()` where the abbreviation `NPER` means "Number of PERiodic payments".

In the context of cumulative interest (compounds), two important concepts are the "present value" (PV) and "future value" (FV) of an acquired a capital.

By answering the question: «*What capital C_n do we get after some time today placing an amount C_0 in a savings account?*», we make a research of future value of capital. We speak then of "capitalization process".

By cons, if we ask, «*What capital C_0 should we place today on a savings account to get after some time a capital C_n?*», we make a research of present value of a capital. We speak then about "discount operation" (that is characteristic of "actuarial calculations").

Definition (#15.3.670): We name the f_c the "capitalization factor" and f_e the "discount factor" defined by the relations:

$$f_c = 1 + t\% \qquad f_e = \frac{1}{1+t\%} \tag{15.3.237}$$

also sometimes denoted by DF.

This brings us also to have:

$$f_c = f_e^{-1} \tag{15.3.238}$$

Starting from the compound capitalization relation give above but with the condensed notation we get:

$$C_n = C_0 f_c^n \tag{15.3.239}$$

The initial capital (present value) can be therefore expressed with the discount factor:

$$C_0 = \frac{C_n}{f_c^n} = C_n f_e^n \tag{15.3.240}$$

This then makes very simple the calculation of the actualization or capitalization as it is to multiply the final or initial capital by the discount factor or capitalization factor to the n-th power. So as we already know we talk about calculate the "future value" (FV) or "present value" (PV) when we invest a certain amount for a certain time at a certain rate and we speak of "present value" when we calculate what should be the initial value investing to get a given amount after a given time at a given rate.

Let us now recall the relation that we have obtained in our initial presentation of equivalent rates:

$$t_e\% = (1+t\%)^{\frac{n}{m}} - 1 \tag{15.3.241}$$

Often in order to simplify the calculations, the person seeking the equivalent rate will simply ask (normalized) $n = 1$. This led him to write:

$$t_e\% = (1+t\%)^{\frac{1}{m}} - 1 \qquad t\% = (1+t_e\%)^m - 1 \tag{15.3.242}$$

Then comes a small trick that use credit analysts in their sales processes that are the concepts of "effective rate" (already seen above!) and "nominal rate". The nominal rate is always lower than the actual rate, it allows the issuer of the debt to show a lower rate than what it actually is (incidentally this practice is prohibited by law in some countries!).

✍**Example:**

Suppose that the conditions for a loan are as follows: annual interest $t_n = 12\%$ (nominal rate written in small in the contract or in the add) payable by monthly increments of $t_e\% = t_n\%/12 = 1\%$. A careful individual realizes that paying 1% monthly interest in a compound system does not give an annual interest of 12% but:

$$t\% = (1+t_p\%)^n - 1 = \left(1+\frac{t_p\%}{p}\right)^p - 1 = \left(1+\frac{t_p\%}{12}\right)^{12} - 1 \cong 12.682\% \tag{15.3.243}$$

which is the effective rate $t_e\%$. In this special case $t_e\%$ is often named "effective annual rate" (EAR) and $t_p\%$ is named "annual percentage rate" (APR)!

Remark

Warning! Never forget that the nominal rate $t_n\%$ includes the rate of inflation $t_i\%$. So the "real rate" $t_r\%$ is equal in reality to:

$$(1+t_n\%) = (1+t_i\%)(1+t_r\%) \tag{15.3.244}$$

For example, if the nominal interest rate for one-year money is 10% with annual compounding and the assumed rate of inflation over the period is 5%, then the real interest rate can be calculated as follows:

$$t_r\% = \frac{1.10}{1.05} - 1 = 4.76\%$$

Now, if more investments are made simultaneously for different durations and at different rates, we may calculate the average compound rate T of all of these investments.

If we denote by C_t the investment number t, $f_{c,t}$ the capitalization factor of investment umber t, n_t the duration of investment number t, k the number of investments and finally T the average compound rate of all investments we can with the help of algebra rule until fourth degree (see section Calculus) or using numerical method (taking the solver or target tool of Microsoft® Excel® for example), solve the following equation:

$$C_1 f_{c,1}^{n_1} + C_2 f_{c,2}^{n_2} + \ldots + C_k f_{c,k}^{n_k} = C_1(1+T)^{n_1} + C_2(1+T)^{n_2} + \ldots + C_k(1+T)^{n_k} \tag{15.3.245}$$

If we make a change of variables $x = 1+T$ we then solve the unknown equation in x (all other terms are normally known in the problem statement):

$$C_1 x^{n_1} + C_2 x^{n_2} + \ldots + C_k x^{n_k} - C_1 f_{c,1}^{n_1} + C_2 f_{c,2}^{n_2} + \ldots + C_k f_{c,k}^{n_k} = 0 \tag{15.3.246}$$

15.3.6.5 Continuous Interest

Let us recall that the compound interest is defined by (with a different common notation for f_c):

$$C_n = C_0 r^n = C_0 (1 + t_n\%)^n \tag{15.3.247}$$

For n periods on a given time basis. Now consider an equivalent rate slitted in smaller periods p such that:

$$C_n = C_0 \left(1 + \left[\left(1 + \frac{t_n\%}{p}\right)^p - 1\right]\right)^n \tag{15.3.248}$$

We can now ask ourselves what would happen to the rate $t_n\%$ if interest was paid not monthly, not daily, but continuously, in an instantaneous way (or almost instantaneous). We write therefore (see section Functional Analysis page 2289):

$$\lim_{p \to +\infty} \left(1 + \frac{t_n\%}{p}\right)^p - 1 = e^{t_n\%} - 1 \tag{15.3.249}$$

Thus, in case of continuous capitalization, the function of capitalization can be finally written:

$$C_n = \lim_{p \to +\infty} C_0 \left(1 + \left[\left(1 + \frac{t_n\%}{p}\right)^p - 1\right]\right)^n = C_0 \left(1 + \left[e^{t_n\%} - 1\right]\right)^n = C_0 \left(e^{t_n\%}\right)^n = C_0 e^{n t_n\%} \tag{15.3.250}$$

And as:

$$\frac{C_n}{C_0} = (1 + t\%)^n \cong e^{n t_n\%} \Rightarrow \frac{C_1}{C_0} = 1 + t\% \cong e^{t_n\%} \tag{15.3.251}$$

We then have (relatively important relation in finance markets) the "logarithmic returns" or just "log-returns":

$$\boxed{t_n\% = \ln\left(\frac{C_n}{C_0}\right) \cong \ln(1 + t\%)} \tag{15.3.252}$$

Therefore the logarithmic return is defined as the Napierian logarithm of the ratio of two values distant in time by a very small interval.

In practice we can use this approximation for large sums of money if the rate is relatively low (under 10%) and that the number of investment period not exceed the ten...

A useful property and widely used in trading of the continuous interest is that it is "time consistent". To see what it is, consider the following three values of any portfolio of equity invested:

Capital	Value	Simple return rate	Logarithmic return rate
C_{n-2}	10.00.-	-	-
C_{n-1}	14.00.-	$=(14-10)/10 = 40\%$	$=\ln\left(\dfrac{14}{10}\right) \cong 33.65\%$
C_n	9.00.-	$=(9-14)/14 = -35.71\%$	$=\ln\left(\dfrac{9}{14}\right) \cong -44.18\%$
Sum:		$=4.29\%$	$=-10.54\%$
Comparison:		$=(9-10)/10 = 10\%$	$=\ln\left(\dfrac{9}{10}\right) \cong 10.54\%$

Table 15.39: Comparison between exact rate and approximated continuous

So as we can see, unlike the simple interest rate, the continuous rate of return is time-consistent because we can sum the different rates for the total variation (this is a very important property used a lot in financial market softwares and especially for performance back-testing!). This result simply comes from the following property:

$$C_n=\left(C_{n-2}e^{t_1\%}\right)e^{t_2\%}=C_{n-2}e^{t_1\%}e^{t_2\%}=C_{n-2}e^{t_1\%+t_2\%}\Rightarrow\ln\left(\frac{C_n}{C_{n-2}}\right)=t_1\%+t_2\% \quad (15.3.253)$$

Or more generally and traditionally:

$$\ln\left(\frac{C_n}{C_0}\right)=t_n\%-t_0\% \quad (15.3.254)$$

Compounding logarithmic returns over small intervals characterized by time points 1, 2, 3 and up to T, each equal interval results in the compounded return from 0 to T. The compounded continuous return collapses into a simple formula:

$$t_n\%=\ln\left(\frac{C_T}{C_0}\right)=\ln\left[\frac{C_1}{C_0}\frac{C_2}{C_1}\frac{C_3}{C_2}\ldots\frac{C_T}{C_{T-1}}\right] \quad (15.3.255)$$

The logarithmic return over several sub-periods is the sum of intermediate logarithmic returns because the logarithm of a product is the sum of logarithms:

$$t_n\%=\ln\left(\frac{C_n}{C_0}\right)=\ln\left(\frac{C_1}{C_0}\right)+\ln\left(\frac{C_2}{C_1}\right)+\ln\left(\frac{C_3}{C_2}\right)+\ldots+\ln\left(\frac{C_T}{C_{T-1}}\right) \quad (15.3.256)$$

Therefore:

$$t_n\%=\ln\left(\frac{C_T}{C_0}\right)=\sum_{i=0}^{T-1}\ln\left(\frac{C_{i+1}}{C_i}\right) \quad (15.3.257)$$

That's why you can find in many financial computer code the following expression:

$$\frac{C_T}{C_0}=e^{\sum_{i=0}^{T-1}\ln\left(\frac{C_{i+1}}{C_i}\right)} \quad (15.3.258)$$

To prior-previous relation can be compared with the traditional discrete compound interest:

$$t_n\%=\frac{C_T}{C_0}=\prod_{i=0}^{T-1}(1+t\%) \quad (15.3.259)$$

Or if we denote the return by the letter R we also found the later relation in the following form:

$$R^{\ln}_{\text{long-period}}=\sum R^{\ln}_{\text{short-period}} \quad (15.3.260)$$

For a log annual return, we could apply the following relation:

$$\boxed{R^{\ln}_{\text{annual}}=\sum R^{\ln}_{\text{monthly}}=\sum R^{\ln}_{\text{daily}}=\ldots} \quad (15.3.261)$$

15.3.7 Progressive interest (annuities)

Definition (#15.3.671): A "pension" or "annuity" is a series of periodic payments at regular time intervals and for a period fixed in advance with compound Interest (typical of the second or third pillar in Switzerland).

Then it is enough to apply the relation (see above proof) $C_0 = C_n f_e^n$ to each annuity term paid if we wish to know the present value of the annuity.

By cons, if we wish to obtain the final value of an annuity, we will apply to each term the following relation (see just earlier above for the proof):

$$C_n = C_0 f_c^n \tag{15.3.262}$$

Definition (#15.3.672): If the annuity is payable at the period end, it is named "postnumerando annuity". By cons, if payable a the beginning of period, it is named "praenumerando annuity".

Remarks

R1. Annuities that are always paid are named "certain annuities" and when the duration is predetermined, we speak of "temporary annuities".

R2. Annuities based on the life of an individual are named "life annuities".

Since the terms are often assumed to be constant, we have the habit of basing the calculations on the value of a currency unit (1$). Thus, we note (the adopted notation is that we find in the literature and as they are not very practical, they are original and pretty to look at ...):

- $a_{\overline{n}|}$ the present value of an annuity of a monetary unit payable postnumerando (at deadline) for a duration of n periods
- $\ddot{a}_{\overline{n}|}$ the present value of an annuity of a monetary unit payable praenumerando (in advance) for a duration of n periods
- $s_{\overline{n}|}$ the future value of an annuity of a monetary unit payable postnumerando (at deadline) for a duration of n periods
- $\ddot{s}_{\overline{n}|}$ the future value of an annuity of a monetary unit payable praenumerando (in advance) for a duration of n periods

The relations that determine these values make use of the properties of geometric series and their partial sum (see section Sequences and Series page 1656 and page 1664).

15.3.7.1 Postnumerando annuities

At constant term, to calculate the final value of an annuity at maturity/postnumerando, we can therefore only work on the capitalization factor multiplied by the final annuity amount.

✍**Example:**

We want to calculate the final value of a postnumerando annuity of $3,500$.- paid during 10 periods and calculated with the periodic effective interest rate of 6%.

Payments held on dates $1,2$, and 10. The payment $C_0=3,500$.- of date 1 has for acquired value on the date 10: $C_0(1+t\%)^9$. Similarly, the payment at date 2 earns interest for 8 years. Its value gained at date 10 is therefore : $C_0(1+t\%)^8$, etc. Finally, at the end, the payment at date 10 (which we have just put in the bank) has the value C_0. The acquired value of the 10 payments is therefore, setting $f_c=1+t\%$ (we will prove the simplifications right after):

$$\begin{aligned}C_0(f_c^9+f_c^8+\ldots+f_c^1+f_c^0)=C_0(f_c^9+f_c^8+\ldots+f_c^1+1)=C_0\frac{1-f_c^{10}}{1-f_c}=C_0\frac{f_c^{10}-1}{f_c-1}\\=C_0\frac{f_c^{10}-1}{t\%}=46,132.80.\text{-}\end{aligned} \tag{15.3.263}$$

In Microsoft® Excel® 14.0.7106 we simply write:

```
=3500*FV(6%,10,-1,0,0)=46,132.80.-
```

Therefore the postnumerando annuity is a payment with constant terms and at constant rates over a number of given periods leading to a simple geometric sequence.

Let us recall (again!) that:

$$C_n=C_0f_c^n \tag{15.3.264}$$

For the postnumerando annuity with constant terms, we then have the general form:

$$C_n=C_0f_c^0+C_0f_c^1+C_0f_c^2+\ldots+C_0f_c^{n-1} \tag{15.3.265}$$

Which is written:

$$C_n=C_0(f_c^0+f_c^1+f_c^2+\ldots+f_c^{n-1})=C_0(1+f_c^1+f_c^2+\ldots+f_c^{n-1})=C_0s_{\overline{n|}} \tag{15.3.266}$$

Therefore:

$$s_{\overline{n|}}=\sum_{t=0}^{n-1}f_c^t \tag{15.3.267}$$

In fact, we have a geometric series of reason q (see section Sequences and Series page 1656). Since then:

$$\begin{aligned}S_n&=a_0+a_1+\ldots+a_{n-1}\\-qS_n&=qa_0+qa_1+\ldots+qa_{n-1}=a_1+a_2+\ldots+a_n\\\hline S_n-qS_n&=a_0-a_n\Rightarrow S_n=\frac{a_0-a_n}{1-q}=a_0\frac{1-q^n}{1-q}\end{aligned} \tag{15.3.268}$$

Therefore:

$$s_{\overline{n}|}=\sum_{t=0}^{n-1} f_c^t = f_c^0\frac{1-f_c^n}{1-f_c}=1\frac{1-f_c^n}{1-f_c}=\frac{f_c^n-1}{f_c-1} \tag{15.3.269}$$

Finally:

$$\boxed{s_{\overline{n}|}=\frac{f_c^n-1}{t\%}} \tag{15.3.270}$$

✍**Example:**

So we have for your example ten periods (and therefore ten terms with $n=10$):

$$C_{10}=3,500s_{\overline{n}|}=3,500\frac{f_c^n-1}{t\%}=3,500\frac{1.06^{10}-1}{0.06}=46,132.60.\text{-} \tag{15.3.271}$$

This capital corresponds then to the sum gained after ten periods.

The method of calculating the "present value of an annuity at maturity/postnumerando" works on the same principle but in reverse according to the relation proved above:

$$C_0=C_nf_e^n \tag{15.3.272}$$

So if the terms (amounts paid) are constant we can write:

$$C_0=C_n(f_e^1+f_e^2+\ldots+f_e^n)=C_na_{\overline{n}|} \tag{15.3.273}$$

Therefore:

$$a_{\overline{n}|}=\sum_{t=1}^{n} f_e^t \tag{15.3.274}$$

Or:

$$\begin{array}{r}
S_n=a_0+a_1+\ldots+a_n\\
-qS_n=qa_0+qa_1+\ldots+qa_n=a_2+a_3+\ldots+a_{n+1}\\
\hline
S_n-qS_n=a_0-a_{n+1}\Rightarrow S_n=\dfrac{a_1-a_{n+1}}{1-q}=a_1\dfrac{1-q^n}{1-q}
\end{array} \tag{15.3.275}$$

Therefore:

$$a_{\overline{n}|}=\sum_{t=1}^{n} f_e^t=f_e^1\frac{1-f_e^n}{1-f_e}=f_e\frac{1-f_e^n}{1-\dfrac{1}{f_c}}=\frac{1}{f_c}\frac{1-f_e^n}{\dfrac{f_c-1}{f_c}}=\frac{1-f_e^n}{f_c-1}=\frac{1-f_e^n}{t\%} \tag{15.3.276}$$

Finally:

$$\boxed{a_{\overline{n}|}=\frac{1-f_e^n}{t\%}} \tag{15.3.277}$$

Remark

The value $C_0 = C_n a_{\overline{n}|}$ correspond therefore to the amount we would need to put on a savings account at the rate of $t\%$to be able to make a constant periodic removal C_n during n periods and thereby settle the account.

✍**Example:**

Let us calculate the present value of a postnumerando annuity of $3,500$.- paid during 10 years at the rate of 6% interest:

$$C_0 = C_n\left(\frac{1-\left(\dfrac{1}{1+t\%}\right)^n}{t\%}\right) = 3,500\left(\frac{1-\left(\dfrac{1}{1+6\%}\right)^{10}}{6\%}\right) = 25,760.30.\text{-} \qquad (15.3.278)$$

Or in Microsoft® Excel® 14.0.1706:

```
=3500*PV(6%,10,-1,0)=25,760.30.-
```

We also have the following relations between actual and final postnumerando rents:

$$\boxed{\left.\begin{aligned} a_{\overline{n}|} &= \frac{1-f_e^n}{t\%} \\ s_{\overline{n}|} &= \frac{f_c^n-1}{t\%} \end{aligned}\right\} \Rightarrow a_{\overline{n}|} = f_e^n s_{\overline{n}|}} \qquad (15.3.279)$$

We also have the following chain rule operations:

$$\begin{aligned} a_{\overline{n+k}|} &= a_{\overline{n}|} + f_e^n a_{\overline{k}|} = f_e^k a_{\overline{n}|} + a_{\overline{k}|} \\ s_{\overline{n+k}|} &= s_{\overline{n}|} + f_c^n s_{\overline{k}|} = f_c^k s_{\overline{n}|} + s_{\overline{k}|} \end{aligned} \qquad (15.3.280)$$

Remark

It is clear that since known $a_{\overline{n}|}$, f_e^n, $t\%$ that $n = \dfrac{\ln(1-t\% a_{\overline{n}|})}{\ln(f_e)}$ and so on for all other types of annuity.

15.3.7.2 Praenumerando annuities

The method of calculating the "present value of an annuity in advance/praenumerando" works on the same principle as the last one still using the relation $C_0 = C_n f_e^n$, but this time the terms of the geometric sequence change since the payment is done in advance:

$$C_0 = C_n\left(f_e^0 + f_e^1 + f_e^2 + \ldots + f_e^{n-1}\right) = C_n \ddot{a}_{\overline{n}|} \qquad (15.3.281)$$

Therefore:

$$\ddot{a}_{\overline{n}|}=\sum_{t=0}^{n} f_e^t \tag{15.3.282}$$

But:

$$\begin{array}{r}
S_n=a_0+a_1+\ldots+a_{n-1}\\
-qS_n=qa_0+qa_1+\ldots+qa_{n-1}=a_1+a_2+\ldots+a_n\\
\hline
S_n-qS_n=a_0-a_n\Rightarrow S_n=\dfrac{a_0-a_n}{1-q}=a_0\dfrac{1-q^n}{1-q}
\end{array} \tag{15.3.283}$$

Therefore:

$$\ddot{a}_{\overline{n}|}=\sum_{t=0}^{n-1} f^t=f_e^0\frac{1-f_e^n}{1-f_e}=1\frac{1-f_e^n}{1-\dfrac{1}{f_c}}=\frac{1-f_e^n}{\dfrac{f_c-1}{f_c}} \tag{15.3.284}$$

Finally:

$$\boxed{\ddot{a}_{\overline{n}|}=\frac{1-f_e^n}{t\% f_e}} \tag{15.3.285}$$

Remark

For the same number of periods and the same interest rate, we obviously have $\ddot{a}_{\overline{n}|}\geq a_{\overline{n}|}$ as $\ddot{a}_{\overline{n}|}=f_e a_{\overline{n}|}$.

✍**Example:**

Let us calculate the present value of a praenumerando annuity of $3,500$.- paid during 10 years at the rate of 6% interest:

$$C_0=C_n\left(\frac{1-\left(\dfrac{1}{1+t\%}\right)^n}{t\%\left(\dfrac{1}{1+t\%}\right)}\right)=3,500\left(\frac{1-\left(\dfrac{1}{1+6\%}\right)^{10}}{6\%\left(\dfrac{1}{1+6\%}\right)}\right)=27,305.92.\text{-} \tag{15.3.286}$$

Or in Microsoft® Excel® 14.0.1706:

```
=3500*PV(6%,10,-1,0,1)=27,305.92.-
```

The method of calculating the "final value of a rent in advance/praenumerando" works on the same principle as the last one using the relation $C_n=C_0 f_c^n$ but this time the terms of the geometric sequence change since the payment is done in advance:

$$C_n=C_0\left(f_c^1+f_c^2+\ldots+f_c^n\right)=C_0\ddot{s}_{\overline{n}|} \tag{15.3.287}$$

Therefore:

$$\ddot{s}_{\overline{n}|}=\sum_{t=1}^{n} f_c^t \tag{15.3.288}$$

As we have proved in the section Sequences and Series (page 1679) during our study of geometric series that:

$$S_n = u_0 \frac{(1-q^{n+1})}{(1-q)} \tag{15.3.289}$$

Therefore:

$$\ddot{s}_{\overline{n}|} = \sum_{t=1}^{n} f_c^t = f_c^1 \frac{1-f_c^n}{1-f_c} = \frac{f_c^n - 1}{\dfrac{f_c - 1}{f_c}} = \frac{f_c^n - 1}{\dfrac{t\%}{1-t\%}} \tag{15.3.290}$$

finally noting $d = \dfrac{t\%}{1-t\%}$ we have:

$$\boxed{\ddot{s}_{\overline{n}|} = \frac{f_c^n - 1}{d}} \tag{15.3.291}$$

Remarks

R1. With the same notation we have also the present value of the praenumerando annuity that is written $\ddot{a}_{\overline{n}|} = \dfrac{1-f_e^n}{d}$.

R2. For the same number of periods and the same interest rate, we obviously have $\ddot{s}_{\overline{n}|} \geq s_{\overline{n}|}$ as $\ddot{s}_{\overline{n}|} = f_c s_{\overline{n}|}$.

✍**Example:**

Let us calculate the final value of a praenumerando annuity of $3,500$.- paid during 10 years at the rate of 6% interest:

$$C_0 = C_n \left(\frac{1+t\%)^n - 1}{t\% \left(\dfrac{1}{1+t\%} \right)} \right) = 3,500 \left(\frac{(1+6\%)^{10} - 1}{6\% \left(\dfrac{1}{1+6\%} \right)} \right) = 25,760.30.\text{-} \tag{15.3.292}$$

Or in Microsoft® Excel® 14.0.1706:

```
=3500*FV(6%,10,-1,0,1)=48,900.75.-
```

15.3.8 Rounding

Rounding a numerical value means replacing it by another value that is approximately equal but has a shorter, simpler, or more explicit representation. Rounding is often done to obtain a value that is easier to report and communicate than the original. Rounding can also be important to avoid misleadingly precise reporting of a computed number, measurement or estimate; for example, a quantity that was computed as $123,456$ but is known to be accurate only to within a few hundred units is better stated as "about $123,500$".

On the other hand, rounding of exact numbers will introduce some round-off error in the reported result. Rounding is almost unavoidable when reporting many computations — especially when dividing two numbers in integer or fixed-point arithmetic; when computing mathematical functions such as square roots, logarithms, and sines; or when using a floating point representation with a fixed number of significant digits. In a sequence of calculations, these rounding errors generally accumulate, and in certain ill-conditioned cases they may make the result meaningless.

To round a number x to the nearest multiple of $1/n$ the relation to be used is as follows:

$$f(x,n)=\frac{\mathrm{E}[nx+0.5]}{n} \tag{15.3.293}$$

The proof is almost intuitive. Just imagine the real axis and cut it into $1/n$ small intervals. Then given x, the number of these intervals will be given by:

$$\frac{x}{1/n}=nx \tag{15.3.294}$$

Finally, to know the number strictly less than the desired multiple, we take the integer value of the last relation and multiply it by $1/n$ such that:

$$f_{\text{inf}}(x,n)=\lfloor x\rfloor=\mathrm{E}[nx]\frac{1}{n}=\frac{\mathrm{E}[nx]}{n} \tag{15.3.295}$$

Notice that the famous "round down" function above, formally named "floor function" is alos sometimes defined as a function $\lfloor x\rfloor:\mathbb{R}\rightarrow\mathbb{Z}$ with the following formula:

$$\lfloor x\rfloor:=\max\{k\in\mathbb{Z}\,|\,k\leq x\} \tag{15.3.296}$$

If, however, we wish to have the number rounded to the nearest multiple, then we see with a linear regression that it is necessary to add 0.5 as:

$$\boxed{f(x,n)=\frac{\mathrm{E}[nx+0.5]}{n}} \tag{15.3.297}$$

To round to the upper multiple we add one such that:

$$f_{\text{sup}}(x,n)=\lceil x\rceil=\frac{\mathrm{E}[nx+1]}{n} \tag{15.3.298}$$

OK... this seems simple but there's a lot more to rounding than might at first meet the eye...:

- Round-Toward-Nearest
- Round-Half-Up (Arithmetic Rounding)
- Round-Half-Down
- Round-Half-Even (Banker's Rounding)
- Round-Half-Odd
- Round-Ceiling (Positive Infinity)
- Round-Floor (Negative Infinity)
- Round-Toward-Zero
- Round-Away From-Zero
- Round-Up

- Round-Down
- Truncation (Chopping)
- Round-Alternate
- Round-Random (Stochastic Rounding)
- Sign-Magnitude Binary Values
- Rounding Signed Binary Values

For business training purpose, among this list, let us just introduce the "Banker's rounding":

If values are always rounded in the same direction (for example, if $+5.5$ rounds up to $+6$ and $+6.5$ rounds up to $+7$, as is the case with the "Round-Half-Up" algorithm presented earlier), the result can be a bias that grows as more and more rounding operations are performed. One solution toward minimizing this bias is to sometimes round up and sometimes round down.

In the case of the round-half-even algorithm (which is often referred to as "Bankers Rounding" because it is commonly used in financial calculations), "half-way" values are rounded toward the nearest even number. Thus, $+5.5$ will round up to $+6$ and $+6.5$ will round down to $+6$.

This algorithm is, by definition, symmetric for positive and negative values, so both -5.5 and -6.5 will round to the nearest even value, which is -6.

In the case of data sets that feature a relatively large number of "half-way" values (financial records often provide a good example of this), the round-half-even algorithm performs significantly better than the Round-Half-Up scheme in terms of total bias (it is for this reason that the use of round-half-even is a legal requirement for many financial calculations around the world!).

As surprising as it can be, a spreadsheet software like Microsoft® Excel® seems to not implement by default a simple function to do such a rounding... (when we know that this a legal requirement).

Remark

It turns out that IBM SAS implements round to the nearest number, and, if equidistant, round away from zero (RA0) with the functions `round()` and `put()`. That is the values of $-2.5, -1.5, 1.5, 2.5$ are rounded to $-3, -2, 2, 3$. In the `R` software however, the `base::round()` function uses the IEEE 754 standard, differing from RA0 only for the equidistant case. `R` rounds to even (R2E) and $-2.5, 2.5$ will round to -2 and 2, respectively.

15.3.9 Loans Amortization/Repayments

Individuals and companies often use loans (credit) as a financial way to find money. Here we will define the main types of loans encountered in practice and the mathematical relations that characterize them.

Definitions (#15.3.673):

D1. In finance, a "loan" is a debt provided by an entity (organization or individual) to another entity at an interest rate, and evidenced by a note which specifies, among other things, the principal amount, interest rate, and date of repayment. A loan entails the reallocation of the subject asset(s) for a period of time, between the lender and the borrower.

D2. We name "indivis loan", a loan with only one lender, usually a financial institution.

D3. As for equity management (see further below) a "guarantor" is the party exposed to the loan risk, but he don not present credit risk itself. We thus avoid the problem of two-sided credit risk.

Remark

When talking in financial terms, a loan is usually followed by an interest on the amount unlike a credit, which might or might not be followed by an interest. As they seems not to be a consensus on the difference between loan and credit some people say that credit grants them the ability to borrow, while a loan is actually the act of borrowing. If a bank gives me a 500.- line of credit, I can, at some date in the future when I want to, borrow money from them. If a bank gives me a 500.- loan, I take possession of the money today.

In a loan, the borrower initially receives or borrows an amount of money, named as we already know the "principal", from the lender, and is obligated to pay back or repay an equal amount of money to the lender at a later time.

The major variables in a mortgage calculation include loan principal, balance, periodic compound interest rate, number of payments per year, total number of payments and the regular payment amount. More complex calculators can take into account other costs associated with a mortgage, such as local and state taxes, and insurance.

When purchasing a new home, most buyers choose to finance a portion of the purchase price via the use of a mortgage. Prior to the wide availability of mortgage calculators, those wishing to understand the financial implications of changes to the five main variables in a mortgage transaction were forced to use compound interest rate tables. These tables generally required a working understanding of compound interest mathematics for proper use. In contrast, mortgage calculators make answers to questions regarding the impact of changes in mortgage variables available to everyone.

Mortgage calculators can be used to answer such questions as:

If one borrows $250,000$.- at a 7% annual interest rate and pays the loan back over thirty years, with $3,000$.- annual property tax payment, $1,500$.- annual property insurance cost and 0.5%

annual private mortgage insurance payment, what will the monthly payment be? The answer is $2,142.42$.-.

Mainly we consider two type of loans:

D1. A "secured loan" is a loan in which the borrower pledges some asset (e.g. a car or property) as collateral.

D2. An "unsecured loan" is monetary loan that is not secured against the borrower's assets. These may be available from financial institutions under many different guises or marketing packages (credit card, peer-to-peer landing, etc.).

The main points of interest on loans are:

- Know the state of the debt at any time
- Know the amount repayable in each period
- Know the interest due each period

Definition (#15.3.674): We name "annuities", payments made in the context of the clearance of a loan. An annuity includes a part of refund R also named "financial depreciation" and part I of interest according to the relation:

$$\begin{gathered}\text{Annuity}=\text{Refund}+\text{Interest}\\ A=R+I\end{gathered}\tag{15.3.299}$$

The decomposition of the annuity in a depreciation part and interest is an important concept not only in finance but also in accounting. Indeed, the part of financial amortization corresponds to a debt repayment at the difference of the interest that is a financial burden.

We will study here three types of loans:

1. Loan with fixed maturity
2. Loans with constant repayments
3. Loans with constant annuities (the most practised)

Remarks

R1. We consider here periodic loans. The transition from one temporal period to another and the calculation of an equivalent rate will be based on relation already proved before.

R2. Stable monthly annuities are named obviously "monthly payments".

15.3.9.1 Fixed-Term Loan

Definition (#15.3.675): We speak about "fixed-term loan" or "in fine loan" when at each period the annuity includes only the part of interest! The last period the annuity includes interest and all (!) the repayment of the loan.

Remark

This model of amortization is particularly used in bonds that we will study later.

The following relations make possible to establish any element of the "amortization table".

Thus, the status of debt (loan) C on n years is at the beginning of the year k is always:

$$C_k = C \quad \forall k < n \in \mathbb{N}^* \tag{15.3.300}$$

The payment (amortization) R_k performed at the en of the year k is equal to the accumulated depreciation S_k at the end of the year k and as it is only made last year n such that:

$$R_k = S_k = \begin{cases} 0 \text{ if } k < n \\ C \text{ if } k = n \end{cases} \tag{15.3.301}$$

The paid I_k will be constant throughout the loan term according to an interest rate $t\%$ on borrowed capital C as (simple interest):

$$I = I_k = t\% C_k = t\% C \quad \forall k \tag{15.3.302}$$

The annuity becomes therefore:

$$A_k = R_k + I_k \tag{15.3.303}$$

✍**Example:**

Let us see the amortization table of a loan of $1,000$.- at 10% year repaid at maturity after 4 years. The corresponding amortization table will be:

Period k	**Debt** C_k	**Amort.** R_k	**Cumulated Amort.** S_k	**Interest** I_k	**Annuity** A_k
1	$1,000$	0	0	100	100
2	$1,000$	0	0	100	100
3	$1,000$	0	0	100	100
4	1.000	$1,000$	$1,000$	100	$1,100$

Table 15.40: **Fixed-Term loan amortization table**

The cost of credit is the sum of interest that is 400.-.

15.3.9.2 Loan with constant amortization

Definition (#15.3.676): We speak of a "constant amortization loan" when the annual amount repaid (amortization) is constant, that is to say identical from year to year (intuitive case).

The following relation make possible to establish any element of this amortization table (no more details are needed because it is very trivial):

$$\begin{aligned}C_K &= C-k\frac{C}{n}+\frac{C}{n}=\frac{C}{n}(1+n-k)\\ R_k &= R=\frac{C}{n}\\ S_k &= k\frac{C}{n}\\ I_k &= t\%C_k\\ A_k &= R_k+I_k=R+I_k\end{aligned} \tag{15.3.304}$$

✍**Example:**

A loan of $1,000$.- at 10% per year is reimbursed by constant amortization in 4 years. The corresponding amortization table will be:

Period k	**Debt** C_k	**Amort.** R_k	**Cumulated Amort.** S_k	**Interest** I_k	**Annuity** A_k
1	1000	250	250	100	350
2	750	250	500	75	325
3	500	250	750	50	300
4	250	250	$1,000$	25	275

Table 15.41: **Constant amortization loan table**

The cost of credit is the sum of interest that is 250.-. So we pay less than with the previous system.

15.3.9.3 Loan with constant annuity

This is the most common case (the definition is in the title itself). It is used by most small credit and leasing institutions. The borrower knows in advance the amount he will have to pay each year. In other words, it is as if it were a capital C that must be settle each period by a constant withdrawal A: which consists to determine the present value of a postnumerando annuity such that:

$$C=Aa_{\overline{n}|} \tag{15.3.305}$$

The following relations are then used to establish any element of the amortization table:

$$A=A_k=\frac{C}{a_{\overline{n}|}} \tag{15.3.306}$$

and because $C=Aa_{\overline{n}|}$, then:

$$C_k=Aa_{\overline{n-k+1}|} \tag{15.3.307}$$

therefore, when $k=1$, we have in accordance with what we expect $C_1=Aa_{\overline{n}|}$.

And therefore the amortization is equal to:

$$R_k=A_k-I_k=A-I_k=A-t\%C_k \tag{15.3.308}$$

The cumulated depreciation is a little less obvious to find with common sense, let us take for the proof a depreciation A with a rate $t\%$ on n periods. Then we have by definition:

$$S_k=\sum_{i=1}^{k}R_i \tag{15.3.309}$$

with $k=2$ and $n=3$ we get:

$$\begin{aligned}R_1&=A-t\%C_1\\ R_2&=A-t\%C_2\end{aligned} \tag{15.3.310}$$

Therefore:

$$\begin{aligned}S_2&=2A-t\%(C_1+C_2)=2A-t\%(Aa_{\overline{3}|}+Aa_{\overline{2}|})=A\left[2-t\%(a_{\overline{3}|}+a_{\overline{2}|})\right]\\ &=A\left[2-t\%(f_e^1+f_e^2+f_e^3+f_e^1+f_e^2)\right]=A\left[2-t\%(2f_e^1+2f_e^2+f_e^3)\right]\\ &=A\left[2-t\%(2\frac{1}{(1+t\%)^1}+2\frac{1}{(1+t\%)^2}+\frac{1}{(1+t\%)^3})\right]=A\left[2-t\%\left(\frac{2(1+t\%)^2+2(1+t\%)+1}{(1+t\%)^3}\right)\right]\\ &=A\frac{1}{(1+t\%)^3}\left[2(1+t\%)^3-t\%(2(1+t\%)^2+2(1+t\%)+1)\right]\\ &=A\frac{1}{(1+t\%)^3}\left[2(1+3t\%+3t\%^2+t\%^3)-t\%(5+6t\%+2t\%^2)\right]\\ &=A\frac{1}{(1+t\%)^3}\left[2+t\%\right]=Af_e^3\left[\frac{t\%}{t\%}2+t\%\right]=Af_e^3\left[\frac{2t\%+t\%^2}{t\%}\right]\\ &=Af_e^3\left[\frac{1+2t\%+t\%^2-1}{t\%}\right]=Af_e^3\left[\frac{(1+t\%)^2-1}{t\%}\right]=Af_e^3s_{\overline{2}|}\end{aligned} \tag{15.3.311}$$

Therefore:

$$S_k=Af_e^n s_{\overline{k}|} \tag{15.3.312}$$

and also:

$$I_k=t\%C_k \tag{15.3.313}$$

✍**Example:**

A loan of $1,000$.- at 10% per year is reimbursed by constant annuity in 4 years. The corresponding amortization table is:

Period k	**Debt** C_k	**Amort.** R_k	**Cumulated Amort.** S_k	**Interest** I_k	**Annuity** A_k
1	1000	215	215	100	315
2	785	237	452	78	325
3	548	261	713	55	325
4	287	287	$1,000$	29	325

Table 15.42: **Constant annuity loan table**

The cost of credit is the sum of interest that is 262.-. This result could be obtained by $nA-C$.

Remark

Financial institutions add different charges to credit as filing fees, insurance costs, guarantee fees, etc. Adding these fees has the effect of increasing the interest rate whose final real value (all inclusive) is named "global effective rate" (GER). Because each financial institution made its small recipes or that each country imposes a particular calculation method (the obligation to communicate the GER and the associated methodology is specified in many countries by legislation to ensure that consumers are not deceived) we did not want to develop the subject in this textbook. However, the reader interested may refer to Econometrics exercises available on the associated website of this textbook where he will be able to find a concrete special case example!

15.3.9.3.1 Equivalent Monthly Installment (EMI)

The "Equivalent Monthly Instalment" (EMI) is the same thing as the loan with constant annuity but just based on a different word used by professional of actuarial science.

The EMI is however well known because the main idea is to have a unique simple formula to get the cost of the credit in one shot. That means the value of 262.- of the previous example given in the table above (remember that this value is rounded as 261.88.-).

Let's say, an amount P is borrowed on which interest is payable at the rate of i per installment period. This amount is to be paid back along with interest in n equal installments.

Let these installments be $I_1, I_2, I_3, \ldots, I_n$ respectively. Since they are all equal, we can say:

$$I_1 = I_2 = I_3 = \cdots = I_n = I \tag{15.3.314}$$

Each installment comprises some part of the principal P that is paid back, the rest is interest on the amount that was outstanding for that installment period. Let the part of the principal that is paid back with $I_1, I_2, I_3, \ldots, I_n$ respectively be $P_1, P_2, P_3, \ldots, P_n$. These principal payments must add up to the original amount P that was borrowed:

$$P_1 + P_2 + P_3 + \cdots + P_n = P \tag{15.3.315}$$

Now let's talk about interest. Don't forget that the idea is that the interest payable in this regard is $P \cdot i$ which will be paid along with a given P_1 to give the I_1 to pay, i.e. $I_1 = P_1 + P \times i$. At the end of the second installment period, only the amount $P - P_1$ will have been outstanding for the same period. Therefore, only $[P - P_1] \cdot i$ is payable as interest for the second installment period which will be paid along with P_2 when I_2 is paid. Similarly, at the end of the third installment period, interest will be paid only on $[P - (P_1 + P_2)]$ for one installment period along with P_3 when I_3 is paid.

Making the above argument for all installments, we get:

$$\begin{gathered}
I_1 = P_1 + P \cdot i \\
I_2 = P_2 + [P - P_1] \cdot i \\
I_3 = P_3 + [P - (P_1 + P_2)] \cdot i \\
\vdots \\
I_n = P_n + [P - (P_1 + P_2 + \cdots + P_{n-1})] \cdot i
\end{gathered} \tag{15.3.316}$$

If we add up the above equations, we get:

$$\begin{gathered}
I_1 + I_2 + I_3 + \cdots + I_n = [P_1 + P \cdot i] + [P_2 + [P - P_1] \cdot i] + [P_3 + [P - (P_1 + P_2)] \cdot i] \\
+ \cdots + [P_n + [P - (P_1 + P_2 + \cdots + P_{n-1})] \cdot i] \\
I_1 + I_2 + I_3 + \cdots + I_n = [P_1 + P_2 + P_3 + \cdots + P_n] + P \cdot i + [P - P_1] \cdot i \\
+ [P - (P_1 + P_2)] \cdot i + \cdots + [P - (P_1 + P_2 + \cdots + P_n)] \cdot i \\
I_1 + I_2 + I_3 + \cdots + I_n = [P_1 + P_2 + P_3 + \cdots + P_n] + P \cdot i + P \cdot i - P_1 \cdot i + P \cdot i \\
- (P_1 + P_2) \cdot i + \cdots + P \cdot i - (P_1 + P_2 + \cdots + P_n) \cdot i \\
I_1 + I_2 + I_3 + \cdots + I_n = [P_1 + P_2 + P_3 + \cdots + P_n] + n \cdot P \cdot i \\
- [P_1 + (P_1 + P_2) + \cdots + (P_1 + P_2 + \cdots + P_{n-1})] \cdot i
\end{gathered} \tag{15.3.317}$$

Remember that $I_1 = I_2 = I_3 = \cdots = I_n = I$ and $P_1 + P_2 + P_3 + \cdots + P_n = P$:

$$n \cdot I = P + n \cdot P \cdot i - [P_1 + (P_1 + P_2) + \cdots + (P_1 + P_2 + \cdots + P_{n-1})] \cdot i \tag{15.3.318}$$

Now, let j be a natural number such that $j \in [2, n-1]$. Then I_j and I_{j+1} can be written as follows:

$$\begin{gathered}
I_j = P_j + \left[P - \sum_{k=1}^{j-1} P_k \right] \cdot i \\
I_{j+1} = P_{j+1} + \left[P - \sum_{k=1}^{j} P_k \right] \cdot i
\end{gathered} \tag{15.3.319}$$

Consider $I_{j+1}-I_j$:

$$\begin{aligned}
I_{j+1}-I_j &= P_{j+1}+\left[P-\sum_{k=1}^{j}P_k\right]\cdot i-\left(P_j+\left[P-\sum_{k=1}^{j-1}P_k\right]\cdot i\right)\\
&= P_{j+1}+P\cdot i-\sum_{k=1}^{j}P_k\cdot i-P_j-\left[P-\sum_{k=1}^{j-1}P_k\right]\cdot i\\
&= P_{j+1}+\cancel{P\cdot i}-\sum_{k=1}^{j}P_k\cdot i-P_j-\cancel{P\cdot i}+\sum_{k=1}^{j-1}P_k\cdot i\\
&= P_{j+1}-\left(\sum_{k=1}^{j-1}P_k\cdot i+P_j\cdot i\right)-P_j+\sum_{k=1}^{j-1}P_k\cdot i\\
&= P_{j+1}-\cancel{\sum_{k=1}^{j-1}P_k}\cdot i-P_j\cdot i-P_j+\cancel{\sum_{k=1}^{j-1}P_k}\cdot i P_k\cdot i\\
&= P_{j+1}-P_j\cdot i-P_j
\end{aligned} \tag{15.3.320}$$

From $I_1=I_2=I_3=\cdots=I_n=I$ we then obviously have $I_j=I_{j+1}=I$. Therefore the last equality :

$$I_{j+1}-I_j=P_{j+1}-P_j\cdot i-P_j \tag{15.3.321}$$

Can be simplified to:

$$0=P_{j+1}-P_j\cdot i-P_j \tag{15.3.322}$$

And rearranged into (with $j\in[2,n-1]$):

$$P_{j+1}=P_j\cdot(i+1) \tag{15.3.323}$$

Again, since $I_1=I_2=I$:

$$\begin{aligned}
\cancel{I_2}-\cancel{I_1}&=P_2-P_1\cdot i-P_1\\
0&=P_2-P_1\cdot i-P_1\\
P_2&=P_1\cdot(i+1)
\end{aligned} \tag{15.3.324}$$

If we expand out $P_{j+1}=P_j\cdot(i+1); j\in[2,n-1]$ and add $P_2=P_1\cdot(i+1)$ to the resulting set of equations, we get:

$$\begin{aligned}
P_1&=P_1\\
P_2&=P_1\cdot(i+1)\\
P_3&=P_2\cdot(i+1)\\
&\vdots\\
P_j&=P_{j-1}\cdot(i+1)\\
&\vdots\\
P_{n-1}&=P_{n-2}\cdot(i+1)\\
P_n&=P_{n-1}\cdot(i+1)
\end{aligned} \tag{15.3.325}$$

Multiplying the above equations gives us:

$$\begin{aligned}
P_1\cdot P_2\cdot P_3\cdot\ldots\cdot P_{n-1}\cdot P_n &= P_1\cdot[P_1\cdot(i+1)]\cdot[P_2\cdot(i+1)]\cdot\ldots\\
&\cdot[P_{n-2}\cdot(i+1)]\cdot[P_{n-1}\cdot(i+1)]\\
\cancel{P_1}\cdot\cancel{P_2}\cdot\cancel{P_3}\cdot\ldots\cdot\cancel{P_{n-1}}\cdot P_n &= P_1\cdot\left[\cancel{P_1}\cdot(i+1)\right]\cdot\left[\cancel{P_2}\cdot(i+1)\right]\cdot\ldots\\
&\cdot[\cancel{P_{n-2}}\cdot(i+1)]\cdot[\cancel{P_{n-1}}\cdot(i+1)]\\
P_n &= P_1\cdot[(i+1)\cdot(i+1)\cdot\ldots\cdot(n-1)\text{ times }]\\
P_n &= P_1\cdot(i+1)^{n-1}
\end{aligned} \tag{15.3.326}$$

Now we would like to have an expression for $\sum_{k=1}^{j} P_k; j\in[1,n]$. The same set of equations that helped us figure the last relation will be useful in this regard:

$$\begin{aligned}
\sum_{k=1}^{j} P_k &= P_1+P_2+P_3+\ldots+P_j\\
\sum_{k=1}^{j} P_k &= P_1+P_1\cdot(i+1)+P_2\cdot(i+1)+\cdots+P_{j-1}\cdot(i+1)\\
\sum_{k=1}^{j} P_k &= P_1+[P_1+P_2+\cdots+P_{j-1}]\cdot(i+1)\\
\sum_{k=1}^{j} P_k &= P_1+[P_1+P_2+\cdots+P_{j-1}+P_j-P_j]\cdot(i+1)\\
\sum_{k=1}^{j} P_k &= P_1+\left[\sum_{k=1}^{j} P_k-P_j\right]\cdot(i+1)\\
\sum_{k=1}^{j} P_k &= P_1+\left(\sum_{k=1}^{j} P_k\right)\cdot(i+1)-P_j\cdot(i+1)\\
\cancel{\sum_{k=1}^{j} P_k} &= P_1+i\cdot\sum_{k=1}^{j} P_k+\cancel{\sum_{k=1}^{j} P_k}-P_j\cdot(i+1)\\
0 &= P_1+i\cdot\sum_{k=1}^{j} P_k-P_j\cdot(i+1)
\end{aligned} \tag{15.3.327}$$

Hence after rearranging:

$$\sum_{k=1}^{j} P_k=\frac{P_j\cdot(i+1)-P_1}{i}; j\in[1,n] \tag{15.3.328}$$

We get the following relation by using the relation above:

$$\begin{aligned} P_1 &= \frac{P_1\cdot(i+1)-P_1}{i} \\ P_1+P_2 &= \frac{P_2\cdot(i+1)-P_1}{i} \\ P_1+P_2+P_3 &= \frac{P_3\cdot(i+1)-P_1}{i} \\ &\vdots \\ P_1+P_2+P_3+\cdots+P_{n-1} &= \frac{P_{n-1}\cdot(i+1)-P_1}{i} \end{aligned} \tag{15.3.329}$$

Let us add these up to get the following:

$$\begin{gathered} P_1+(P_1+P_2)+(P_1+P_2+P_3)+\cdots+(P_1+P_2+P_3+\cdots+P_{n-1}) \\ = \frac{P_1\cdot(i+1)-P_1}{i}+\frac{P_2\cdot(i+1)-P_1}{i}+\frac{P_3\cdot(i+1)-P_1}{i}+\cdots \\ +\frac{P_{n-1}\cdot(i+1)-P_1}{i} \\ P_1+(P_1+P_2)+(P_1+P_2+P_3)+\cdots+(P_1+P_2+P_3+\cdots+P_{n-1}) \\ = \frac{(P_1\cdot(i+1)-P_1)+(P_2\cdot(i+1)-P_1)+(P_3\cdot(i+1)-P_1)+\cdots+(P_{n-1}\cdot(i+1)-P_1)}{i} \\ P_1+(P_1+P_2)+(P_1+P_2+P_3)+\cdots+(P_1+P_2+P_3+\cdots+P_{n-1}) \\ = \frac{(P_1+P_2+P_3+\cdots+P_{n-1})\cdot(i+1)-P_1\cdot(n-1)}{i} \\ P_1+(P_1+P_2)+(P_1+P_2+P_3)+\cdots+(P_1+P_2+P_3+\cdots+P_{n-1}) \\ = \frac{(P_1+P_2+P_3+\cdots+P_{n-1}+P_n-P_n)\cdot(i+1)-P_1\cdot(n-1)}{i} \end{gathered} \tag{15.3.330}$$

Let us substitute the value of $P_1+P_2+P_3+\cdots+P_{n-1}+P_n=P$ into the last relation above. Then we get:

$$\begin{gathered} P_1+(P_1+P_2)+(P_1+P_2+P_3)+\cdots+(P_1+P_2+P_3+\cdots+P_{n-1}) \\ = \frac{(P-P_n)\cdot(i+1)-P_1\cdot(n-1)}{i} \end{gathered} \tag{15.3.331}$$

Let us substitute into this relation the values of:

$$P_n = P_1\cdot(i+1)^{n-1} \tag{15.3.332}$$

derived earlier above. Then we get:

$$\begin{gathered} P_1+(P_1+P_2)+(P_1+P_2+P_3)+\cdots+(P_1+P_2+P_3+\cdots+P_{n-1}) \\ = \frac{(P-P_1\cdot(i+1)^{n-1})\cdot(i+1)-P_1\cdot(n-1)}{i} \\ P_1+(P_1+P_2)+(P_1+P_2+P_3)+\cdots+(P_1+P_2+P_3+\cdots+P_{n-1}) \\ = \frac{P\cdot(i+1)-P_1\cdot(i+1)^n-P_1\cdot(n-1)}{i} \\ P_1+(P_1+P_2)+(P_1+P_2+P_3)+\cdots+(P_1+P_2+P_3+\cdots+P_{n-1}) \\ = \frac{P\cdot(i+1)-P_1\cdot((i+1)^n+(n-1))}{i} \end{gathered} \tag{15.3.333}$$

Let us now substitute the value of $P_1+(P_1+P_2)+(P_1+P_2+P_3)+\cdots+(P_1+P_2+P_3+\cdots+P_{n-1})$ derived just above, into the following relation derived at the beginning:

$$n\cdot I=P+n\cdot P\cdot i-[P_1+(P_1+P_2)+\cdots+(P_1+P_2+\cdots+P_{n-1})]\cdot i \tag{15.3.334}$$

Therefore:

$$\begin{aligned}n\cdot I&=P+n\cdot P\cdot i-\left(\frac{P\cdot(i+1)-P_1\cdot((i+1)^n+(n-1))}{\cancel{i}}\right)\cdot\cancel{i}\\&=P+n\cdot P\cdot i-(P\cdot(i+1)-P_1\cdot((i+1)^n+(n-1)))\\&=P+n\cdot P\cdot i-P\cdot(i+1)+P_1\cdot((i+1)^n+(n-1))\\&=\cancel{P}+n\cdot P\cdot i-P\cdot i-\cancel{P}+P_1\cdot((i+1)^n+(n-1))\\&=n\cdot P\cdot i-P\cdot i+P_1\cdot((i+1)^n+(n-1))\end{aligned} \tag{15.3.335}$$

After rearranging we get:

$$P_1=\frac{n\cdot I+P\cdot i-n\cdot P\cdot i}{(i+1)^n+(n-1)} \tag{15.3.336}$$

Substituting the value of P_1 above into the following relation defined at the beginning:

$$I_1=P_1+P\cdot i \tag{15.3.337}$$

We get:

$$I_1=\frac{n\cdot I+P\cdot i-n\cdot P\cdot i}{(i+1)^n+(n-1)}+P\cdot i \tag{15.3.338}$$

We know that by definition $I_1=I$. Hence:

$$I=\frac{n\cdot I+P\cdot i-n\cdot P\cdot i}{(i+1)^n+(n-1)}+P\cdot i \tag{15.3.339}$$

Now let us rearrange and simplify:

$$\begin{aligned}I\cdot((i+1)^n+(n-1))=\left(\frac{n\cdot I+P\cdot i-n\cdot P\cdot i}{\cancel{(i+1)^n+(n-1)}}\right)\cdot\cancel{((i+1)^n+(n-1))}+P\\\cdot i\cdot((i+1)^n+(n-1))\\I\cdot((i+1)^n+(n-1))-n\cdot I=P\cdot i-n\cdot P\cdot i+P\cdot i\cdot((i+1)^n+(n-1))\\I\cdot((i+1)^n+(\cancel{n}-1)-\cancel{n})=P\cdot i\cdot(1-n)+P\cdot i\cdot(i+1)^n+P\cdot i\cdot(n-1)\\I\cdot((i+1)^n-1)=-\cancel{P\cdot i\cdot(n-1)}+P\cdot i\cdot(i+1)^n+\cancel{P\cdot i\cdot(n-1)}\\I\cdot((i+1)^n-1)=P\cdot i\cdot(i+1)^n\end{aligned} \tag{15.3.340}$$

So finally:

$$\boxed{I=\frac{P\cdot i\cdot(i+1)^n}{(i+1)^n-1}} \tag{15.3.341}$$

Applied to the previous example given earlier above, we get:

$$\text{EMI}=I=\frac{P\cdot i\cdot(i+1)^n}{(i+1)^n-1}=\frac{1000\cdot 10\%\cdot(10\%+1)^4}{(10\%+1)^4-1} \tag{15.3.342}$$

Hence $1261.88-1000=261.88$.- so we fall back on the same cost of credit.

15.3.10 Modern Portfolio Theory

Warning! Topics related to modern portfolio theory are relatively to the number of definitions almost as worst as Thermodynamics. It is the same for the number of assumptions of the theoretical models used therein. Let us also tell you that they are often several vocabulary terms (and also mathematical notations) to describe the same thing in the trade finance (we have no knowledge of an ISO standard in finance vocabulary although one exists to standardize the symbols of financial instruments) and that happens quite regularly that the banks themselves may correct mathematical definitions in their brochures and softwares to align with the majority opinion of the moment... (sign convention, symbols convention, convention of what should be divided by what, convention of what should be subtracted from what, etc.). We have also, in consistence with the other chapters of this textbook, tried to simplify to the maximum the mathematical proofs (and therefore limiting ourselves to the simplest theoretical bachelor level model...) even taking sometimes shortcuts that will make grind some teeth.... and we have also tried to give each time the different vocabulary terms of use and existing multiple mathematical notations for the same theoretical concept. So good read and good luck!

The theory of stock market also named "modern portfolio theory" is the mathematical theory that deals with the prices, selection, management and exchange transactions, loans and capital through time (in fact it is more related to engineering as this theories are build by trial and error and should be named "financial engineering" and be but in the corresponding chapter!). This theory is strongly makes a intensive usage of to statistical models and so it is important to have read and understood the sections of Probabilites and Statistics of this textbook before. We will see that the majority of models are based on a probabilistic representation and the problem becomes a determination of a price at a future date and this is equivalent to determine the probability distribution of asset prices, these latter being seen as random variables.

It should however be known that in practice in the private or public financial institutions, only a tiny minority of market players knows, understands and applies mathematical models and for the others who obtained certification or continuing education qualifications (CFA, FRM , CAIA, etc.), the mathematical level is often very very low. Indeed the vast majority of traders (market operators) are limited to the graphical analysis of candles stock type diagrams using analysis of moving averages type at the limit based on control charts (see section Industrial Engineering page 8934) with Bollinger bands or other USL/LSL, linear regressions and almost other similar techniques (the funniest being empirical index with scientific names like "stochastic index", which are actually just relative variations...). Financial management is therefore ultimately often only the application of common sense (at least at the end of 120th century... according to holocene calendar) on the variation of prices based on quantities... analysing graphs and knowing when to sell or buy at the right moment.

This situation can be easily explained: the current theoretical models of the 120th century and early 121st century (holocene calendar) are unable to date to take into account the real complexity and interaction of our modern world. You will see in this section that actually the majority of mathematical models studied in major universities involve simplified and idealized cases (independence, unimodal distributions, finite variance, choice of empirical statistical indicators, etc.). So in the current situation it is often better to listen to a trader who is informed of Government and Corporates strategies than to a mathematician for who the world is resumed to a fairytale (but nevertheless allows him to sell consulting hours at a very high rate and reassure his customers who are desperate to predictions that virtually almost make no sense...).

The reader must also be aware that a number of people are convinced that everything is written somewhere, than a kind of abstract reality exists outside our physical world and that if we were smart enough, we could mathematically formalize and provide future developments in the long term. In fact the scientist knows that in this kind of field we have to deal with the deterministic chaos of the market and that the only way to manage this is to correct with thumbs with a vague idea what will happen. In economics some specialists speak about the "market dictatorship", but this is recognized in a sense that we know we can predict almost nothing! Obviously some in the complexity business sell to bankers and others the idea that they will predict the fluctuations of the stock market... but it is sufficient to observe the past to see that no modern model could have predicted it. The only thing that math can do in financial management is to analyse the behaviour of an ideal financial asset in a idealized context and this is already not bad and forces a little common sense as decision support system (D.S.S.)... (for those who can do math which is very far from the case of 99% of people working in finance).

Buy the Rumor Sell the News

In finance, mathematical models therefore are used mainly to quantify and minimize the investment risk. As such, they play the role of decision support system for managers, investors and regulators. But, with rare exceptions, a bank or an investment fund does not base a major investment decision on a mathematical equation. The decision, for investment banks, is often motivated by the search for ever greater returns and for that they are not based on mathematical models. Moreover, the ruling people of private banks or government are often people from the world of politics, law or business management with little skill to really understand how markets work (see the Master Graduate program trading in many countries, for example). Also beware of companies - especially multinationals - seeking financial specialists mastering Microsoft® Excel®. VBA, Microsoft Access, C#... Because it means that they use non-professional tools to do a job which ought to be with the appropriate tools (and Microsoft® Excel® or Microsoft Access are not and C++ is much more faster than C#)!!! So in terms of internal organization, you can ensure that these companies organize and analyse anything, anyhow, with no suitable tools and therefore that there is internally a general mess.

Remark

Many of the techniques presented below are available as complete code in the C++ book [423] that we strongly recommend to the reader.

This doesn't stop however those that consider mathematics as an Art and think that financial theoretical models have a certain elegance. Furthermore these models allow to thoroughly understand the working mechanism of financial engineering and these models have sometimes interesting connections with other fields of Applied Mathematics.

Furthermore, increasingly used by financial institutions, automated trading bots account in 12016 (holocene calendar) for nearly 60% of global daily transactions, and almost 80% just on Wall Street. The high frequencies trading bots do transaction in 5 microseconds and this is so quick that even visualization tools as used by humans are not adapted anymore as the human brain needs at most a 5 seconds zoom out to be able to understand and analyse what happens. What happened is finance was quite easily predictable and probably the same will happen in a near of far future to all qualified jobs that can be quantified (Management, Project Management, Quality Management etc.).

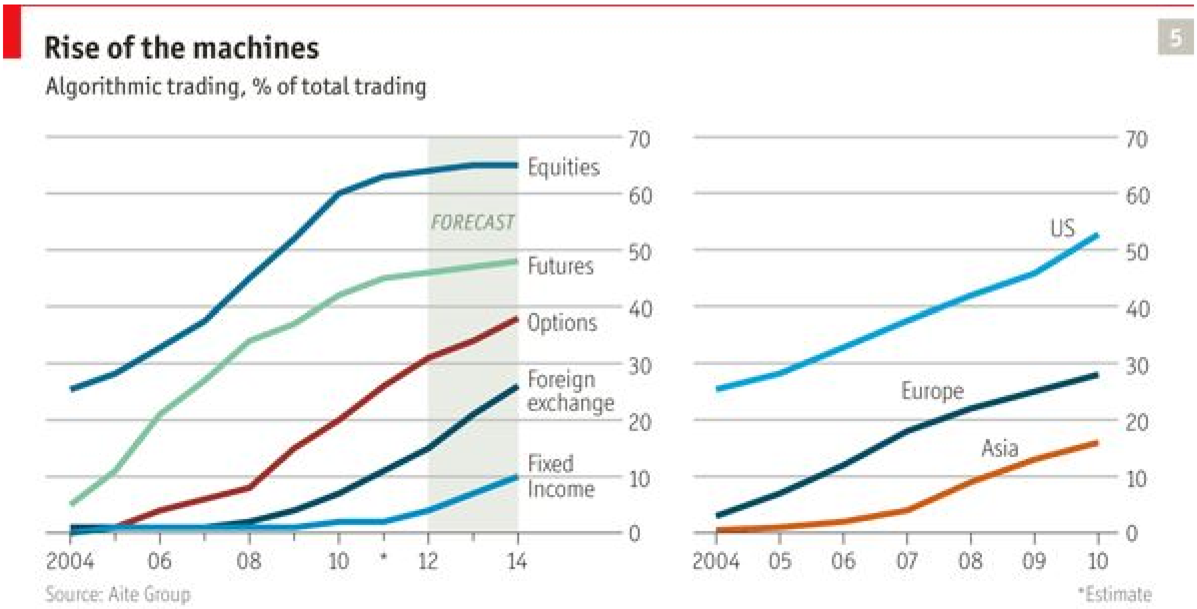


Figure 15.96: **Rise of the machines in algorithmic trading (source: Financial Times)**

It might also be useful to make liable this business (activity) by implementing at least the following measures:

- Require to make public the technical documentation of components (underlyings), algorithms and mathematical models of financial products.
- Require a dynamical level of own funds (fractional reserve) to financial institutions based on their market positions and risk classes of assets under management.
- Forcing actors of the market to operate in an area of activity with a given predefined level of connectivity with the class of managed assets.
- Limit the number of transactions for a financial asset not limited in time and also impose for this same family of asset a minimum time of non-transaction.

- Avoid front running or spoofing - form of insider trading[16] consisting in pretending buying or selling a security with the intention to cancel the transaction at the last minute - by requiring automated transactions (bypassing humans (brocker / trader)) and if necessary prohibit practising for life to people cheating and withdraw their academic degrees.
- Forcing people active in this business to take high level exams regularly to check that they understand what they do and the assumptions relating thereto (as for actuaries in certain insurance)
- Prohibit deposit banks to do also trading.
- Closely monitor all certified "risk manager" coming out of training centers for the past 30 years because given the events during those years, it would appear that they do not do their job or if they do, that their recommendations are not implemented at the highest level of management of the banks (and thus their work is totally counterproductive).

There would be a number of proposals to be implemented to restore its original purpose and ... ethics to the field of financial management.

Definitions (#15.3.677):

D1. "Stock Exchange" is the public market organized and in theory... regulated... where securities are exchanged (shares, bonds, options, etc.) whose value will fluctuate relatively to the "fundamental value" (basic value calculated using theoretical models) depending on the offer ("ask") and demand ("bid"). When a security is asked too much, the price goes up, and vice versa, when nobody wants it. The difference between the bid and the ask is named the "spread".

The stock exchange is a structure that allows:

(a) For companies that want to invest (and therefore increase their capital) to get funds to meet potential or proven demand.

(b) Make the economy more stable by making it more dynamic and fluid as possible (but still under control ...) in order that it regulates itself.

The above mentioned system is working if and only if it is transparent, rational, efficient, self-regulating and balanced!

Remark

We talk about "speculative bubble" when prices observed on a stock market deviate too much from the fundamental value of traded goods.

D2. We name "over-the-counter market" (OTC) a transaction between two parties that are free (directly between the seller and the buyer) to contract and normally out of the Stock Exchange and that is therefore private, non-organized and unregulated (or in a very flexible way...). Sometimes a broker acts as an intermediary in OTC, but this latter is not a counterpart: it will not intervene in the settlement rules of the transaction. Sometimes a bank offers itself this type of transaction and ensures itself the counterpart, but often by covering their risk in another existing market.

---

[16] The FINRA (Financial Industry Regulatory Authority) has a website tool named "BrokerCheck" where investors can obtain more information about any FINRA-registered broker or brokerage firm that have been barred from the securities industry for insider trading!

For example, the ForEx currency market (Foreign Exchange) or CDS (Credit Default Swap) are essentially OTC market (daily turnover of several thousand billion!). For example in the context of currencies, a company or a bank that wishes to make a foreign exchange transaction will be put in direct contact with another bank. However, there exist an organized market of currencies.

D3. An investment is named "liquid investment" if the investment involves financial instruments that one can buy or sell at any time respectively. Verbatim a portfolio is named "liquid portfolio" if it contains a majority of liquid instruments.

D4. In a financial institution we speak of "front office" when we mean traders who execute transactions, take positions, etc. We talk about "middle office" when we refer to risk managers that follow the risks taken by the front office. Finally, we talk about "back office" for any remaining administrative part (management of entries, accounting, administration, etc.).

D5. Investors are heterogeneous in terms of fortune. There are generally small savers ("retail), customers of private banks ("private banking" and "family office") and very fortunated investors ("high net worth").

Before starting to tackle pure and hard mathematics, we will need once again to give many definitions to be fluent with the vocabulary in use by financial analysts and financial engineers (caution! the vocabulary list is relatively long...).

15.3.10.1 No Arbitrage Opportunity (N.A.O.)

One of the fundamental assumptions of the usual financial models is that there is no financial strategy to, for a zero initial cost, acquire a certain fortune in a future date. This hypothesis is named "no arbitrage opportunities" (N.A.O.). It is theoretically justified by the uniqueness of prices that characterize a market in perfect competition (see section Game Theory page 7400).

In practice, there are arbitrage opportunities but that disappear very quickly due to the existence of arbitrage-seeker, these latter are market agents whose role is to detect this type of opportunities and use them. They then create a force that tends to change the price of the asset to its no-arbitrage price (the "real" price).

Thus, if multiple assets of same risk offer different returns, investors seeking new opportunities will logically turn their purchases to those whose performance is the highest, so this behaviour leads to a decline in the performance of these assets (because they become more expensive).

The financial mathematics based on the N.A.O. let these people make money and neglects these opportunities appearances which in any case will always exist but assumed on a very short time (this type of strategy is leveraged today with computers that can give orders sales and purchasing order at the millisecond and even faster).

A remarkable example to illustrate this is to use a simplified version of the model developed by Cox, Ross and Rubinstein as it explicitly reflects the concept of the N.A.O. and the importance of probabilistic models.

The example is based on the assumption that the market is composed of one risky asset and a (non-risky) constant investment rate r. For example, a sum of 1.- today placed at rate r, generates a guaranteed income $1+r$ at time 1 regardless of the future market developments in chosen example (by definition of riskless/risk-free investment rate...).

We first study this market over a single period of time such that the initial time will be noted $t=0$ and the final instant $t=1$ (we name this situation a "monoperiodic market"). We assume also that we have a perfect knowledge of the market at the initial time. In our context this means that the price of the risky asset is $S_0>0$ and fixed and the non-risky asset is determined by its performance $r>0$.

And about the risky asset, its value at time $t=1$ is not known in advance. To decrease the field of possibilities, we will assume that the performance of this asset can take only two values b (bottom) and h (high) with:

$$-1\leq b<h \tag{15.3.343}$$

Thus, the risky asset at time $t=1$ can take only two positive values. The low value:

$$S_1=S_0(1+b) \tag{15.3.344}$$

other the high value:

$$S_1=S_0(1+h) \tag{15.3.345}$$

hence the name "(naive) binomial model"...

An investor can then buy a quantity δ of risky asset and place (lend or borrow!) an - unlimited - amount n_0 of money at the assumed risk-free rate r. The value V_0 at time $t=0$ of the portfolio of composition (δ,n_0) is therefore:

$$V_0=\delta S_0+n_0 \tag{15.3.346}$$

At the final moment, we will have:

$$V_1=\delta S_1+n_0(1+r) \tag{15.3.347}$$

As we explained above, S_1 can take two values, it is therefore the same for V_1. This means that the income of the portfolio is also uncertain.

Now, to show the concept of N.O.A. let us do a numerical example considering the special situation where it is more advantageous, and without fail, to invest in the risky asset rather that the riskless (risk-free) one.

Imagine for this purpose that we borrowed 100.- (hence n_0 is negative!) to a bank to a risk-free rate of 5% and that the risky asset unit we want to acquire with this money is traded today at $S_0=100$.- and that this latter can take two future values:

$$t=1:\begin{cases}S_1=105\rightarrow b=5\%\\S_2=108\rightarrow h=8\%\end{cases} \tag{15.3.348}$$

We then have for our portfolio at the initial time:

$$V_0=\delta S_0+n_0=1\cdot 100-100=0.\text{-} \tag{15.3.349}$$

and at the final time two possible cases:

$$\begin{aligned}V_{1,b}&=\delta S_{1,b}+n_0(1+r)=1\cdot 105-100(1+5\%)=0.\text{-}\\V_{1,h}&=\delta S_{1,h}+n_0(1+r)=1\cdot 108-100(1+5\%)=3.\text{-}\end{aligned} \tag{15.3.350}$$

We then see trivially that if $r<b$ there is an arbitrage opportunity as it becomes possible to make surely money without spending any! To avoid a N.O.A. in this situation, it is necessary that the market is balanced and that there is:

$$b\leq r \tag{15.3.351}$$

Conversely, if it is surely more lucrative to invest in risk-free asset rather than in the risky asset, the market must organized itself to avoid any arbitrage opportunity to ensure that:

$$r<h \tag{15.3.352}$$

Thus, in both cases, we must avoid at all times that in the (naive) binomial market we have a N.O.A. and this will be possible if:

$$b\leq r<h \tag{15.3.353}$$

The absence of arbitrage opportunity has two major simple involvement (there are less simple one as we shall see later...). Consider the case where we have two assets of performance (yield/return) respectively i and j and that we know that the performance of asset i will not default in the payment of dividends. Instead the asset of performance j may fail with probability $1-p$. So, by the N.O.A. we must have:

$$p(1+j)\leq 1+i \tag{15.3.354}$$

So we can infer the probability of non-default of the asset of performance j (given the two performances at the same time):

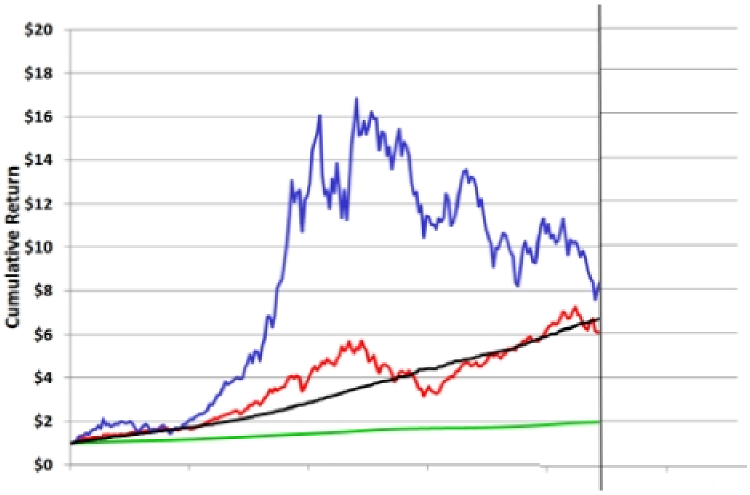
$$p\leq\frac{1+i}{1+j} \tag{15.3.355}$$

and therefore the probability of default:

$$1-p\geq 1-\frac{1+i}{1+j} \tag{15.3.356}$$

This type of reasoning thus also allows to require a performance j of an issuer of assets (typically bonds type) knowing by expertise/audit the probability of default of payment/repayment relatively to a 100% safe assets of performance i.

Such a reasoning can help to avoid misunderstanding and bad long term strategies investments. Indeed, here are four assets, real data (no information here about time, but it's the same scale for the four of them):



A little survey on the social network Twitter was made by Arthur Charpentier on this chart to know if you could invest in one, and only one, asset which one will you pick ?

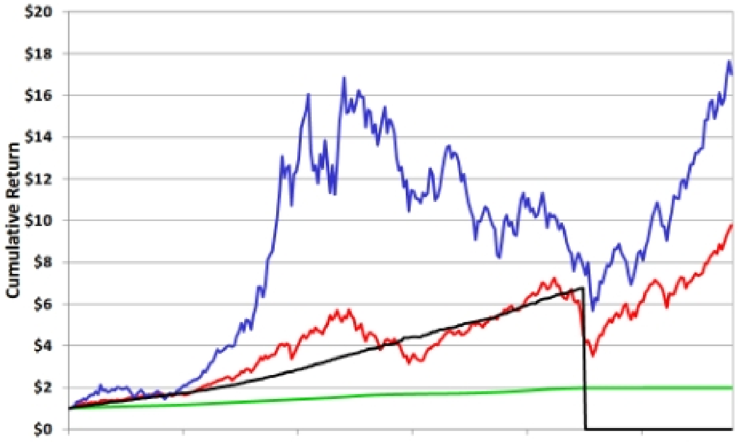
The answers were: 15% blue one, 58% black one, 21% red one, 6% green one.

The assets were here, respectively:

- green = U.S. Treasury Bills
- red = Stock Market (SP500)
- black = Fairfield Sentry
- blue = Pfizer (single stock)

The black curve was for a firm that had among the largest exposures to the Bernard Madoff fraud. And here are the curves (again, real data!).

The more complete graph after the cut is:



Now, if you think about it, it is kind of odd that almost 60% people pick the black one. In a real market, how can we get at the same time the black and the green curve after what we just studied??? They both had no risk, and very different return. The black curve has the same return as the SP500 index, without any risk. That should be seen as odd... Of course, it's easy to comment, once you see the data, but still.

Caution! The reader has to be very careful if he wants to play on the stock market because there exist a fraudulent practice quite common on the Internet and particularly on social networks and named "scalping", or "pump-and-dump". This system of "scalping" generally consists of the following (we take a real case): McAfee, Watson, and other associates bought large amounts of altcoins (publicly traded cryptocurrencies), which were referred to as commodities or securities at low market prices.

Parameters	**Scalping**	**Day Trading**	**Swing Trading**	**Position Trading**
Trading Time Frame	Short Term	Same Day	Medium Term	Long Term
Trading Signals	Many Daily	Few Daily	Few Weekly	Few Monthly or Yearly
Typical Holding Time	Minutes	Hours	Days or Weeks	Months or Years
Chart Interval Analysis	1 min., 5 min., 15 min.	15 min., 30 min., 1 h.	1 h., 4 h., 1 day	Days, Weeks, Months
Trade Monitoring	Very Active	Active	Moderate Active	Rarely Active
Mathematical Models	No Intuition based	College Level Maths	Undergraduate Level Maths	Graduate to Postgraduate Level Maths

Table 15.43: **Most common types of trading**

Then in agreement with the creators of altcoins McAfee endorses their cryptocurrency via social networks, recommending these altcoins to members of the investing public in order to artificially inflate (or "inflate") their market prices. We also find this practice among individuals on social networks who encourage people to invest but in reality in the sole interest of inflating their own investment (just before selling of course...).

15.3.10.2 Liquidity and Illiquidity

Asset managers and ordinary investors care about "liquidity" (the ability to turn an asset into cash), insofar as it affects the return on their investments, because illiquid securities cost more to buy and sell. "Illiquidity", which is opposite to liquidity, eats into an investor's return. Another important aspect of liquidity is its effect on a portfolio evaluation: portfolio liquidation, including illiquid assets, may reduce the value of a portfolio significantly. In other words, market liquidity is a market's feature whereby an individual or firm can quickly purchase or sell an asset without causing a drastic change in the asset's price.

Market-makers provide liquidity by taking the opposite side of a transaction. If an investor wants to buy, the market-maker sells and vice versa.

In exchange for this service, market-makers buy at a low bid price P^b and sell at a higher ask price P^a : This ability insures that the market-makers makes profits.

Definitions (#15.3.678):

D1. The difference $P^a - P^b$ is named the "bid-ask spread" (the cost of making transactions without delay):

$$\text{bid-ask spread} = P^a - P^b \tag{15.3.357}$$

It's a trading cost. We associate lower bid-ask spreads with more liquid securities. At the opposite, high trading costs (commissions, fees, opportunity costs, bid-ask spreads, etc.) are linked to less liquid securities. If the spread is 0 then we speak of a "frictionless asset".

Tinic (11972 according to holocene calendar) and Benston and Hagerman (11974 according to holocene calendar) seem to have find a positive relation between trading activity and liquidity and a negative relation between trading activity and volatility.

D2. We also define the "quoted spread" as percent of the midpoint (or also named the "midquote"), that is, the average between the lowest ask and highest bid (when expressed in percentage!):

$$\text{quoted spread} = \frac{P^a - P^b}{\text{midpoint}} \tag{15.3.358}$$

See the numerical example just further below if necessary!

D3. Quoted spreads often over-state the spreads finally paid by traders, due to "price improvement", that is, a dealer offering a better price than the quotes, also known as "trading inside the spread". Effective spreads account for this issue by using trade prices, and are typically defined as (when expressed in percentage!):

$$\text{effective spread} = 2 \cdot \frac{|\text{trade price} - \text{midpoint}|}{\text{midpoint}} \tag{15.3.359}$$

This definition embeds the assumption that trades above the midpoint are buys and trades below the midpoint are sales!

✍**Example:**

Suppose the quoted bid and ask prices were 25.45.- and 25.50.-, respectively. Also suppose the ask price at order execution was 25.48.-. Then, the quoted bid-ask spread is:

$$\text{bid-ask spread} = P^a - P^b = 25.50 - 25.45 = 0.05.\text{-}$$

The midpoint is:

$$\text{midpoint} = \frac{P^a + P^b}{2} = \frac{25.50 + 25.45}{2} = 25.475.\text{-}$$

Therefore:

$$\text{effective spread} = 2 \cdot |\text{trade price} - \text{midpoint}| = 2 \cdot |20.48 - 25.475| = 0.01.\text{-}$$

The effective spread is less than the quoted spread by $0.05 - 0.01 = 0.04$.-. This observation is known as the "price improvement" because there existed an effective spread that is narrower than the quoted spread.

D4. Quoted and effective spreads represent costs incurred by traders. This cost includes both a cost of asymmetric information, that is, a loss to traders that are more informed, as well as a cost of immediacy, that is, a cost for having a trade being executed by an intermediary. The "realized spread" isolates the cost of immediacy, also known as the "real cost". This spread is defined as:

$$\text{realized spread}_k = 2 \cdot \frac{|\text{trade price}_k - \text{midpoint}_{k+1}|}{\text{midpoint}_k} \quad (15.3.360)$$

The intuition for why this spread measures the cost of immediacy is that, after each trade, the dealer adjusts quotes to reflect the information in the trade (and inventory effects).

D5. Usually, in high-frequency data researchers use a volume-weighted spread. The "daily-trading volume weighted spread" (DTVWS) is given by:

$$\text{DTVWS} = \sum_i w_i (P_i^a - P_i^b) \quad \text{with: } w_i = \frac{V_i}{\sum_i V_i} \quad (15.3.361)$$

where DTVWS is the spread, w_i is the weight for the ith trade, and V_i is its corresponding trading volume.

The information needed to calculate transaction costs is often unavailable. Thus, a wide range of measures are used to evaluate liquidity:

- Volume related (scaled or unscaled): Indirect but widely used measure. It is simple and available (fact: more active markets tend to be more liquid). But, trading volume is correlated to volatility, which can impede market liquidity. Average daily volume (ADV) is widely used by practical traders to estimate how difficult is to trade a particular stock:

the higher the ADV the more liquid the stock and the easier it is to execute large orders without moving the market. Y. Amihud (see [16]) illiquidity ratio, ILLIQ, is one of the most widely used in the industry and is the daily ratio of absolute stock return to its dollar volume averaged over some period. It is used by regulators to estimate liquidity trends.

- Trading frequency related: Number of trades executed within a specified interval, without regard to trade size. High trading frequency may also indicate a more liquid market, but it too can be associated with volatility and hence lower liquidity.
- The bid-ask spread (level or %) or spread-related: It measures the cost of executing a small trade. Usually calculated as the difference between the bid or offer price and the bid-ask midpoint. It can be calculated quickly, with data widely available in real time. But, bid and offer quotes are good only for limited quantities and time periods; the spread just measures the cost of executing a single trade of a certain size.
- Quote size: Quantity of securities treadable at the bid and offer prices. It accounts for market depth and complements the bid-ask spread. Market makers often do not reveal the full quantities they will transact at a given price, so the measured depth underestimates the true depth.
- Trade size: Quantity of securities traded at the bid and offer prices, reflecting any negotiation over quantity. Alternative depth measure. Trade size also underestimates market depth, because the quantity traded is often less than the quantity that could have been traded at a given price.
- Price impact coefficient (Temporary or Permanent): It considers the rise (fall) in price that typically occurs with a buyer-initiated (seller-initiated) trade. Useful for large trades or a series of trades. Together with the bid-ask spread and depth measures provides a good picture of liquidity. A drawback is the difficulty of obtaining the data required for estimation, particularly on a real-time basis.

From a trader's perspective, the definition of liquidity could be formulated in the form of an answer to the following question: What amount of money can one invest without moving the market? This is difficult to answer as it depends on which strategy the trader uses to fill the order, how much time the trader has to execute it, how big the order is compared to ADV, etc. However, refining the above question into: What amount of money is needed to create a daily single unit price fluctuation of the stock? Leads to the following, more precise definition of liquidity:

$$\text{Liquidity} \cong \frac{\text{Consideration}}{\text{Price Range}} = \frac{\text{Volume} \cdot \text{Price}}{\text{High} - \text{Low}} \tag{15.3.362}$$

The liquidity measure calculated in the above formula would range from thousands to billions, so to reduce this range into more manageable numbers some practitioners take logarithm of this indicator.

Among others empirical liquidity indicators, we have the LIX that is a model-free indicator introduced by F. Guillaume in 12013 (holocene calendar):

$$\text{LIX}_t \cong \frac{V_t P_t}{P_{\text{High},t} - P_{\text{Low},t}} \tag{15.3.363}$$

We also have the Hui-Heubel liquidity ratio, as it is referenced in Sarr and Lybek (see [477]) and described by the following relation:

$$L_{hh}=\frac{(P_{\text{High}}-P_{\text{Low}})/P_{\text{Low}}}{\$V/(M\cdot \text{E}(P))} \tag{15.3.364}$$

where P_{High} is the is the highest daily price over the last five days, P_{Low} is the lowest daily price over the last five days $\$V$ is the total dollar volume traded over the last five days, M is the number of shares outstanding and $\text{E}(P)$ is the average closing price of the instrument over a five day period. Noticing that the total dollar volume is approximately the average price multiplied by the total volume V:

$$\$V=\sum_j P_j V_i \cong \text{E}(P)V \tag{15.3.365}$$

Amihud (see [16]) illiquidity measure is given by the formula:

$$\text{ILLIQ}=\frac{1}{N}\sum_{i=1}^{N}\text{ILLIQ}_i=\frac{1}{N}\sum_{i=1}^{N}\frac{|r_i|}{\$V_i} \tag{15.3.366}$$

where r_i is the stock return on day i and $\$V_i$ is the dollar volume on day i, whereas N is the number of days.

As far as we know there is a few dozens of liquidity measures that are all empiricals like LDV model, Lesmond-Ogden and Trzcinka's (11999) measure, Brennan and Subrahmanyam (11996), Datar-Naik and Radcliffe (11998), Elewswarapu (11997), Stoll and Whaley measure (11983), Roll's c (11984), Glosten and Harris's (11988) λ and Φ, Hasbrouk's (12005) c^{Gibbs}, Rabinovitch and al. (12003), Piqueira (12004).

15.3.10.2.1 Bias ratio

The "bias ratio" is an empirical indicator used in finance to analyse the returns of investment portfolios in the purpose to detect valuation bias or deliberate price manipulation of portfolio assets by a manager of a hedge fund, mutual fund or similar investment vehicle, without requiring disclosure (transparency) of the actual holdings. This metric measures in a naive way abnormalities in the distribution of returns that indicate the presence of bias in subjective pricing. The formulation of the Bias Ratio stems from an insight into the behaviour of asset managers as they address the expectations of investors with the valuation of assets that determine their performance.

The bias ratio measures how far the returns from an investment portfolio – e.g. one managed by a hedge fund – are from an unbiased distribution. Thus the bias ratio of a pure equity index will usually be close to 1. However, if a fund smooths its returns using subjective pricing of illiquid assets the bias ratio will be higher. As such, it can help identify the presence of illiquid securities where they are not expected.

The bias ratio was first defined by Adil Abdulali, a risk manager at the investment firm Protégé Partners. The concepts behind the bias ratio were formulated between 12001 and 12003 (holocene calendar) and privately used to screen money managers. The first public discussions on the subject took place in 12004 (holocene calendar) at New York University's Courant Institute and in 12006 (holocene calendar) at Columbia University.

Let:

- $[0,+\hat{\sigma}]=$ the closed interval from the mean to $+1$ standard deviation of returns (including 0)
- $[-\hat{\sigma},0)=$ the half open interval from -1 standard deviation of returns to the mean return (including $-\hat{\sigma}$ and excluding 0)
- $r_i=$ being the return in month i, $1\leq i\leq n$, and n being the number of monthly returns

Then:

$$\boxed{\text{Bias Ratio}=\text{BR}=\frac{\#\left(r_i\,|\,r_i\in[0,+\hat{\sigma}]\right)}{1+\#\left(r_i\,|\,r_i\in[-\hat{\sigma},0)\right)}} \tag{15.3.367}$$

The bias ratio roughly approximates the ratio between the area under the return histogram near zero in the first quadrant and the similar area in the second quadrant. It holds the following properties:

- $0\leq\text{BR}\leq n$
- If $r_i<0,\forall i$, then $\text{BR}=0$
- If $\forall r_i$ such that $r_i>0, r_i>\hat{\sigma}$ then $\text{BR}=0$
- If the distribution r_i is Normal with mean $\hat{\mu}=0$, then BR approaches 1 as n goes to infinity.

The bias ratio has since been used by a number of Risk Management professionals to spot suspicious funds that subsequently turned out to be frauds. The most spectacular example of this was reported in the Financial Times on 22 January 12009 (holocene calendar) titled *Bias ratio seen to unmask Madoff*!

Remark

The above definition is the most common one as the huge majority of financial papers use the latter to compare funds BR. However some softwares packages, like `R` (see our companion book on that software for a computational example of the Bias ratio), use the following definition instead:

$$\text{Bias Ratio}=\text{BR}=\frac{\#\left(r_i\,|\,r_i\in[\hat{\mu},\hat{\mu}+\hat{\sigma}]\right)}{1+\#\left(r_i\,|\,r_i\in[\hat{\mu}-\hat{\sigma},\hat{\mu})\right)} \tag{15.3.368}$$

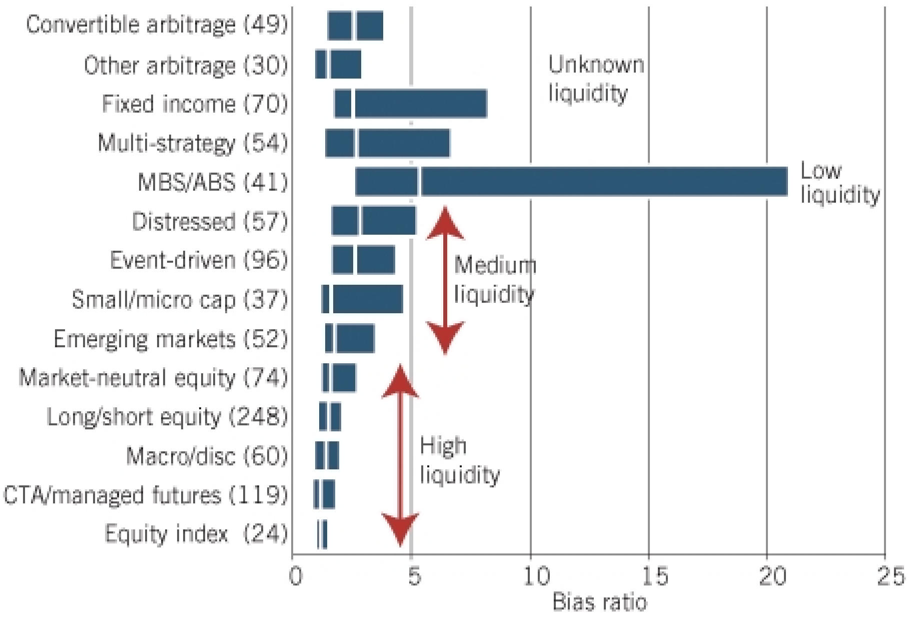


Figure 15.97: **Typical bias ratio of some funds**

Consider for example the following returns (the years are based on the holocene calendar):

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
12008	0.63	0.06	0.18	0.93	0.81	−0.06	0.72	0.71	0.50	−0.06	na	na
12007	0.29	−0.11	1.64	0.98	0.81	0.34	0.17	0.31	0.97	0.46	1.04	0.23
12006	0.70	0.20	1.31	0.94	0.70	0.51	1.06	0.77	0.68	0.42	0.86	0.86
12005	0.51	0.37	0.85	0.14	0.63	0.46	0.13	0.16	0.89	1.61	0.75	0.54
12004	0.88	0.44	−0.01	0.37	0.59	1.21	0.02	1.26	0.46	0.03	0.79	0.24
12003	−0.35	−0.05	1.85	0.03	0.9	0.93	1.37	0.16	0.86	1.26	−0.14	0.25
12002	−0.04	0.53	0.39	1.09	2.05	0.19	3.29	−0.13	0.06	0.66	0.09	0.00
12001	2.14	0.08	1.07	1.26	0.26	0.17	0.38	0.94	0.66	1.22	1.14	0.12
12000	2.14	0.13	1.77	0.27	1.30	0.73	0.58	1.26	0.18	0.86	0.62	0.36
11999	1.99	0.11	2.22	0.29	1.45	1.70	0.36	0.87	0.66	1.05	1.54	0.32
11998	0.85	1.23	1.68	0.36	1.69	1.22	0.76	0.21	0.98	1.86	0.78	0.26
11997	2.38	0.67	0.80	1.10	0.57	1.28	0.68	0.28	2.32	0.49	1.49	0.36
11996	1.42	0.66	1.16	0.57	1.34	0.15	1.86	0.20	1.16	1.03	1.51	0.41
11995	0.85	0.69	0.78	1.62	1.65	0.43	1.02	−0.24	1.63	1.53	0.44	1.03
11994	2.11	−0.44	1.45	1.75	0.44	0.23	1.71	0.35	0.75	1.81	−0.64	0.60
11993	−0.09	1.86	1.79	−0.01	1.65	0.79	0.02	1.71	0.28	1.71	0.19	0.39
11992	0.42	2.72	0.94	2.79	−0.27	1.22	−0.09	0.85	0.33	1.33	1.35	1.36
11991	3.01	1.40	0.52	1.32	1.82	0.30	1.98	1.00	0.73	2.75	0.01	1.56
11990												2.77

Table 15.44: **Fairfield Sentry monthly returns form Dec. 11999 to Oct. 12008**

We get:

$$\text{Bias Ratio} = \text{BR} = \frac{\#\left(r_i \,|\, r_i \in [0, +\hat{\sigma}]\right)}{1 + \#\left(r_i \,|\, r_i \in [-\hat{\sigma}, 0)\right)} = \frac{88}{1+16} \cong 5.18 \tag{15.3.369}$$

As a reference, let's note the Bias Ratio of two other recent frauds[17]:

---

[17] Conducting comparisons using the Sharpe ratio, it seems that the latter is far less discriminating as a fraud detector than the Bias Ratio.

1. Bayou Superfund was supposed to apply a high turnover, intraday stock trading strategy and admitted in 12005 (holocene calendar) that it made up returns. It had a Bias Ratio of 6.
2. Beacon Hill's Safe Harbour fund was trading mortgage-backed securities until fraud was discovered in 12002 (holocene calendar). It had a Bias Ratio of 7.
3. Maddoff Securities was also fabricating returns. It had a Bias Ratio of 8.09
4. Fairfield Sentry returns (a Madoff feeder fund) had a Bias Ratio between 6 and 7

15.3.10.2.2 Bid-Ask (spread) Roll Model

Let us see now a quite simple theoretical model of ask/bid prices: the "Roll model" (for illiquidity). Richard Roll developed the model in *A simple implicit measure of the elective bid-ask spread* (11984 according to holocene calendar).

The Roll Model is a method for inferring the effective bid-ask spread directly from a time series of market prices. The method requires no data other than the prices themselves.

We start by making the following simplifying assumptions:

- All trading in the market is conducted through dealers. These dealers will post ask/bid prices at each point in time as a_t and b_t.
- Dealers are competing with each other so that the spread between the ask/bid prices is equivalent to the dealers' cost c per trade.
- We are on a perfect information market (i.e. symmetrical ask-bid spread and hence no arbitrage opportunities!) that behaves according to a brownian motion.

We consider the bid/ask by considering an arithmetic random walk as the mid-point of the bid-ask:

$$m_t = m_{t-1}\varepsilon_t \tag{15.3.370}$$

with $\varepsilon_t = \mathcal{N}(0,\sigma_\varepsilon^2)$. Then the bid b_t and ask a_t are given by:

$$a_t = m_t + c \qquad b_t = m_t - c \tag{15.3.371}$$

We can infer the trade price p_t as a function of the mid-quote price m_t, half of the bid/ask cost c, and the sign of the trade q_t ($+1$ for buy with probability 50% and -1 for sell also with probability 50%):

$$p_t = m_t + cq_t \tag{15.3.372}$$

We can model this simply in R (see our companion book on R for an example).

We can also use this model in the opposite direction, to infer transaction costs from actual prices. The Roll model has two parameters that can be estimated, c and σ_ε^2.

This is very useful as it implies that we can estimate transaction costs by only using transaction data (i.e. without knowing quoted spreads). Here, transaction costs are inferred from serial covariance of daily returns:

$$\Delta p_t = -cq_{t-1} + cq_t + \varepsilon_t = c\Delta q_t + \varepsilon_t \tag{15.3.373}$$

We define the variance and covariance as:

$$\gamma_0 = \text{V}(\Delta p_t) \tag{15.3.374}$$

and (first-order serial covariance):

$$\gamma_1 = \text{cov}(\Delta p_t, \Delta p_{t-1}) \tag{15.3.375}$$

Let us calculate first the variance of q_t (we will admit that the mean $\text{E}(q_t)$ of q_t is obviously 0). We have:

$$\text{V}(q_t) = \text{E}(q_t^2) - \text{E}(q_t)^2 = \big(P(q_t = -1)\cdot(-1)^2 + P(q_t = +1)(+1)^2\big) - 0^2 = 1 \tag{15.3.376}$$

For the covariance we use the explicit definition of that latter:

$$\text{cov}(X,Y) = \text{E}(XY) - \text{E}(X)\text{E}(Y) \tag{15.3.377}$$

With in our case here:

$$\begin{aligned}
&\text{cov}(\Delta p_t, \Delta p_{t-1}) = \text{E}(\Delta p_t \Delta p_{t-1}) - \text{E}(\Delta p_t)\text{E}(\Delta p_{t-1}) = \text{E}(\Delta p_t \Delta p_{t-1}) - 0\cdot 0\\
&= \text{E}(\Delta p_t \Delta p_{t-1}) = \text{E}((c\Delta q_t + \varepsilon_t)(c\Delta q_{t-1} + \varepsilon_{t-1}))\\
&= \text{E}\left((-cq_{t-1} + cq_t + \varepsilon_t)(-cq_{t-2} + cq_{t-1} + \varepsilon_{t-1})\right)\\
&= \text{E}\left(c^2 q_{t-1}q_{t-2} - c^2 q_{t-1}^2 - cq_{t-1}\varepsilon_{t-1} - c^2 q_t q_{t-2} + c^2 q_t q_{t-1} + cq_t\varepsilon_{t-1}\right.\\
&\left.-c\varepsilon_t q_{t-2} + c\varepsilon_t q_{t-1} + \varepsilon_t\varepsilon_{t-1}\right)
\end{aligned} \tag{15.3.378}$$

We have (it is quite straightforward):

$$\begin{aligned}
\text{E}(c^2 q_{t-1}q_{t-2}) = 0 \quad \text{E}(-cq_{t-1}\varepsilon_{t-1}) = 0 \quad \text{E}(-c^2 q_{t-1}q_{t-2}) = 0 \quad \text{E}(c^2 q_t q_{t-1}) = 0\\
\text{E}(cq_t\varepsilon_{t-1}) = 0 \quad \text{E}(-c\varepsilon_t q_{t-2}) = 0 \quad \text{E}(c\varepsilon_t q_{t-1}) = 0
\end{aligned} \tag{15.3.379}$$

So it remains (keep in mind that when two random variables X and Y are independent, we have $\text{E}(XY) = \text{E}(X)\text{E}(Y)$):

$$\text{cov}(\Delta p_t, \Delta p_{t-1}) = \text{E}\left(-c^2 q_{t-1}^2 + \varepsilon_t\varepsilon_{t-1}\right) = -c^2\text{E}(q_{t-1}^2) + \text{E}(\varepsilon_t)\text{E}(\varepsilon_{t-1}) = -c^2 \tag{15.3.380}$$

Hence:

$$\gamma_1 = \text{cov}(\Delta p_t, \Delta p_{t-1}) = -c^2 \tag{15.3.381}$$

Therefore to summarize we have:

$$\gamma_0 = 2c^2 + \sigma_\varepsilon^2 \qquad \gamma_1 = -c^2 \tag{15.3.382}$$

So we can now solve for c and σ_ε^2 as:

$$\boxed{c = \sqrt{-\gamma_1} = \sqrt{-\text{cov}(\Delta p_t, \Delta p_{t-1})} \qquad \sigma_\varepsilon^2 = \gamma_0 + 2\gamma_1 = \text{V}(\Delta p_t) + 2\text{cov}(\Delta p_t, \Delta p_{t-1})} \tag{15.3.383}$$

So we just need to compute the variance and covariance of returns in order to estimate the expected cost. An example of estimating these values in R is given in our companion book on that software!

Remark

In the original paper, $c = s/2$, where s denotes the "spread". Therefore we have:

$$s = 2\sqrt{-\text{cov}(\Delta p_t, \Delta p_{t-1})} \tag{15.3.384}$$

Where in the case the spread should be expressed as percent, replace 2 with 200.

The cost variable might now be compared against actual bid/ask data whether our assumptions are reasonable. There are some obvious aspects of this model that don't match reality:

- Ordinarily bid and ask prices stay at the same level for different periods of time. This model doesn't include any sense of size of orders, and hence doesn't capture this temporal aspect.
- The bid/ask spread is not a constant in actual trading. It also follows clear intraday patterns where it varies between the open, midday, and close times.
- There is a strong relationship between trades and changes in quotes, this dynamic is completely ignored in the Roll model. A quote level may change when market makers change their prices or someone trades and moves from one level of an order book to another.
- Trades do not occur with every quote. There can be extended periods of time where markets are being quoted without any trading taking place.

Several existing empirical studies attempt to test the Roll model on real data and attempt to verify the validity of its estimates (if a researcher has trade prices but not quotes). Harris (11990 according to holocene calendar) *Statistical Properties of the Roll Serial Covariance Bid/Ask Spread Estimator* provides a nice summary of the basic flaws!

Remark

In practice we can find $\text{cov}(\Delta p_t, \Delta p_{t-1}) > 0$, to avoid this problem, Roll (11984 according to holocene calendar) defines the spread as $c = -\sqrt{\text{cov}(\Delta p_t, \Delta p_{t-1})}$. But there are also other empirical choices (treating the value as a missing value, or defining it as e equal to zero).

While the basic Roll model is not empirically successful, it is still appealing as a first model due to its simplicity, and it can be easily extended into more a realistic model (as a recent example, see Zhang and Hodges (12011 according to holocene calendar)). We will revisit extensions in later posts, but our next model will move on to consider market maker inventory models, starting with Garman (11976 according to holocene calendar).

15.3.10.3 Portfolios

Definitions (#15.3.679):

D1. "Portfolio management" is the art and science of making decisions about investment mix and policy, matching investments to objectives, asset allocation for individuals and institutions, and balancing risk against performance. Portfolio management is all about determining strengths, weaknesses, opportunities and threats in the choice of debt vs. equity, domestic vs. international, growth vs. safety, and many other trade-offs encountered in the attempt to maximize return at a given appetite for risk.

D2. An "Investment fund" or more simply "portfolio" is an investment vehicle (portfolio of securities, shares or bonds, for example) that financial institutions offer their customers and that is structured and maintained to match the investment objectives stated in its prospectus. In theory, if the stock market is failing, the investment fund should not loose money.

Even if an investment funds contains various financial assets, the customers can buy the issued units to a quite low price compared to buying individual assets. Each unit theoretically contains a proportion of each asset being in the fund. They guarantee a right to participate in the overall fund's assets without giving the rights on the companies included in the fund.

Some people estimate in the early 121st century (holocene calendar) that there is nearly $30,000$ investment funds worldwide.

D3. A "mutual fund" is an investment vehicle that is made up of a pool of funds collected from many investors for the purpose of investing in securities such as stocks, bonds, money market instruments and similar assets. Mutual funds are operated by money managers, who invest the fund's capital and attempt to produce capital gains and income for the fund's investors. A mutual fund's portfolio is structured and maintained to match the investment objectives stated in its prospectus.

One of the main advantages of mutual funds is that they give small investors access to professionally managed, diversified portfolios of equities, bonds and other securities, which would be quite difficult (if not impossible) to create with a small amount of capital. Each shareholder participates proportionally in the gain or loss of the fund. Mutual funds then offers to small investors the possibility to enjoy of a not too bad risk diversification to and also to get discount prices on transactions made by the fund managers.

D4. A "hedge fund" is an investment fund that pools capital from a limited number of accredited sophisticated individual (the criteria for becoming one are lengthy and restrictive) or institutional investors and invests in a variety of assets, often with complex portfolio construction and risk management techniques. It is administered by a professional management firm, and often structured as a limited partnership, limited liability company, or similar vehicle. Hedge funds are generally distinct from mutual funds as their use of leverage is not capped by regulators and distinct from private equity funds as the majority of hedge funds invest in relatively liquid assets such as options and to do aggressive investment (short sell). This will typically increase the leverage - and thus the risk - of the fund. This also means that it's possible for hedge funds to make money when the market is falling. Mutual funds, on the other hand, are not permitted to take these highly leveraged positions and are typically safer as a result.

The majority of stock transactions concern the content of "security portfolio", which are the set of titles that a market player may hold. Manage a portfolio is therefore (most typically...) for a manager to seek to maximize a return on investment (ROI) for the customer while minimizing the risks (at least on long term investments...).

Remark

The ROI is also sometimes named "yield" or "performance" or "profit rate" and thus refers to a financial ratio that measures the amount of money gained or lost relative to the amount originally invested (often on the basis of an annual or semi-annual period). Typically, this ratio is expressed as a percentage rather than as a decimal.

The "securities" (financial securities) are derived in the form of shares, bonds, foreign exchange options and commodities, all more generally named "financial products" or "financial assets" and whose definitions (non exhaustive) are given below. When some of these products are mixed, we then speak of "structured products" (typically the combination of an underlying with an option).

Issuing securities to the public isn't free, and the costs of different methods are important determinants of which is used. These costs associated with floating a new issue are generically named "flotation costs".

✍**Example:**

On March 17, 12011 (holocene calendar), Cornerstone, the software company, went public via an IPO (Initial Public Offering). Cornerstone issued 7.5 million shares of stock at a price of $\$13$ each. The lead underwriters on the IPO were Goldman, Sachs & Co. and Barclays Capital, assisted by a syndicate made up of 4 other investment banks.

Even though the IPO raised a gross sum of $\$97.5$ million, Cornerstone got to keep only $\$90,538,500$ after expenses. The biggest expense was the 7.14% underwriter spread, which is ordinary for an offering of this size. Cornerstone sold each of the 7.5 million shares to the underwriters for $\$12.0718$, and the underwriters in turn sold the shares to the public for $\$13.00$ each. But wait, there's more! Cornerstone spent $\$15,421$ in SEC registration fees, $\$13,783$ in other filing fees, and $\$150,000$ to be listed on the NASDAQ Global Market. The company also spent $\$2.76$ million in legal fees, $\$1,012,700$ on accounting to obtain the necessary audits, $\$9,350$ for a transfer agent to physically transfer the shares and maintain a list of shareholders, $\$130,000$ for printing and engraving expenses, $\$5,000$ in certain legal compliance fees and expenses, and finally, $\$3,746$ in miscellaneous expenses.

As this example show, an IPO can be a costly undertaking! In the end, expenses totaled $\$11,061,500$, of which $\$6,961,500$ went to the underwriters and $\$4,100,000$ went to other parties. All told, the total cost was 12.2% of the issue proceeds.

Definitions (#15.3.680):

D1. To measure the overall evolution of the stock market, we use "market indices" reflecting the arithmetic mean or the weighted average prices (values) of a number of representative assets. This allows economic agents to know the overall performance a reference basket for a group of technologies or for a country.

A price index (there exist also volatility indices) try to provides a summary of the overall price market by tracking some of the top stocks within that market[18] or small, medium, and large companies, while other indexes represent only the largest companies. Some indexes also tend to track companies within a certain sector, like technology, while other indices are more broad.

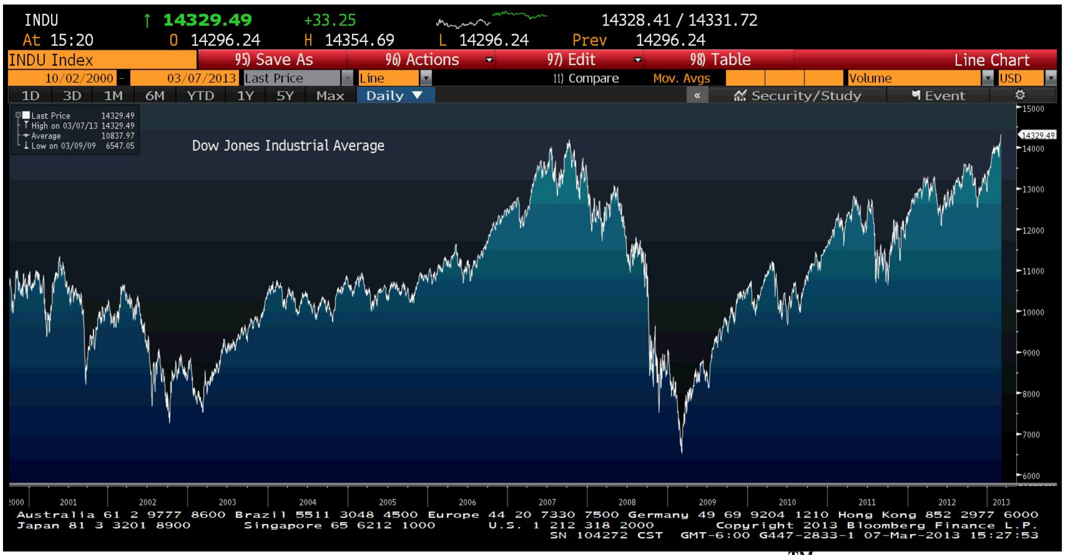

INDU
↑ 14329.49
+33.25
14328.41 / 14331.72
At 15:20
O 14296.24
H 14354.69
L 14296.24
Prev 14296.24
INDU Index
95) Save As
96) Actions
97) Edit
98) Table
Line Chart
10/02/2000 - 03/07/2013
Last Price
Line
11) Compare
Mov. Avgs
Volume
USD
1D 3D 1M 6M YTD 1Y 5Y Max
Daily
«
Security/Study
Event
Last Price 14329.49
High on 03/07/13 14329.49
Average 10837.97
Low on 03/09/09 6547.05
Dow Jones Industrial Average
15000
14329.49
14000
13000
12000
11000
10000
9000
8000
7000
6000
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013
Australia 61 2 9777 8600 Brazil 5511 3048 4500 Europe 44 20 7330 7500 Germany 49 69 9204 1210 Hong Kong 852 2977 6000
Japan 81 3 3201 8900 Singapore 65 6212 1000 U.S. 1 212 318 2000 Copyright 2013 Bloomberg Finance L.P.
SN 104272 CST GMT-6:00 G447-2833-1 07-Mar-2013 15:27:53

Figure 15.98: **Down Jones Industrial Index Monitoring in Bloomberg™ Terminal**

Here are a few of the most known worldwide market indices:

- The "S&P 500" tracks 500 large U.S. companies (selected by committee) across a wide span of industries and sectors. The stocks in the S&P 500 represent roughly 70% percent of all the stocks that are publicly traded.

  Standard & Poor is a company that doles out financial information and analysis and that was founded in 11860 (holocene calendar) by Henry Varnum Poor. In 11941 (holocene calendar) Poor's Publishing (Henry Varnum Poor's original company) merged with Standard Statistics (founded in 11906 - according to holocene calendar - as the Standard Statistics Bureau) and therein assumed the name Standard and Poor's Corporation.

  In order to keep the S&P 500 Index consistent over time, it is adjusted to capture corporate actions which affect market capitalization, such as additional share issuance, dividends and restructuring events such as mergers or spin-offs. Additionally, to remain indicative of the U.S. stock market, the constituent stocks are changed from time to time. Between 12005 and 12014 (holocene calendar), 188 index components were replaced by other components.

  To calculate the value of the S&P 500 Index, the sum of the adjusted market capitalization of all 500 stocks is divided by a factor, usually referred to as the Divisor d. For example, if the total adjusted market cap of the 500 component stocks is US$13 trillion and the Divisor is set at 8.933 billion, then the S&P 500 Index value would be $1,455.28$.

---

[18] In fact they are all bad indicators as they do not incorporate implicitly the market volatility and history trend

The formula to calculate the S&P 500 Index value is:

$$\boxed{\text{S\&P500} = \frac{\sum\limits_i n_i p_i}{d}} \tag{15.3.385}$$

where p_i is the price of each stock i in the index and n_i is the number of shares publicly available for each stock and d the "S&P 500 Divisor".

Companies can be listed in more than one index. Some of the largest companies within the S&P 500 are also in the Dow Jones Industrial Average.

- The "Dow Jones Industrial Average" index (created in 11896 - according to holocene calendar - by Charles Dow and Edward Jones) tracks the 30 largest U.S. companies. This means it represents "large-cap" companies, which is the industry term for "very big companies".

  Although the companies within the Dow Jones represent only about 25 percent of all stocks, the DJIA is widely accepted as the leading indicator of market health.

  To calculate the DJIA, the sum of the prices of all 30 stocks is divided by a divisor, the Dow Divisor. The divisor is adjusted in case of stock splits, spinoffs or similar structural changes, to ensure that such events do not in themselves alter the numerical value of the DJIA. Early on, the initial divisor was composed of the original number of component companies; which made the DJIA at first, a simple arithmetic average. The present divisor, after many adjustments, is less than one (meaning the index is larger than the sum of the prices of the components). That is:

$$\boxed{\text{DJIA} = \frac{\sum\limits_i p_i}{d}} \tag{15.3.386}$$

  where p_i are the prices of the component stocks and d is the "Dow Divisor".

  Events such as stock splits or changes in the list of the companies composing the index alter the sum of the component prices. In these cases, in order to avoid discontinuity in the index, the Dow Divisor is updated so that the quotations right before and after the event coincide:

$$\text{DJIA} = \frac{\sum\limits_i p_{i,\text{old}}}{d_{\text{old}}} = \frac{\sum\limits_i p_{i,\text{bew}}}{d_{\text{new}}} \tag{15.3.387}$$

  The Dow Divisor was 0.14602128057775 on December 24, 12015 (holocene calendar) such that every $\$1$ change in price in a particular stock j within the average, equates to a movement of:

$$\text{DJIA} = \frac{\sum\limits_i (p_i + 1_{i=j})}{d} = \frac{\sum\limits_i p_i}{d} + \frac{1}{d} = \frac{\sum\limits_i p_i}{d} + 6.87\% \tag{15.3.388}$$

  that is to say a 6.87 point movement.

- The "NASDAQ" market index (for "National Association of Securities Dealers Automated Quotations"), which is known as the "NASDAQ Composite", tracks the roughly $3,000$ companies that are traded on the NASDAQ Exchange and began in 11971 (holocene calendar) at a Base Value of 100.00. This is unusual, because no other exchange has its own popular index. The nightly news doesn't read statistics from the "New York Stock Exchange Composite".

  The NASDAQ Composite has grown popular because it's commonly accepted as a shorthand indicator of how tech-sector and innovative companies – both big and small – are faring.

  The NASDAQ assigns every company on the exchange a "share weight" based on the company's market capitalization, or the total value of its shares. The larger the company, the greater the weight and the greater the effect a change in its share price will have on the NASDAQ Composite Index. Every company's share weight is multiplied by its share price, and all companies' results are added together. That sum is then divided by a number named, appropriately enough, the "divisor". The purpose of the divisor is the same as for previous defined indexes.

  Therefore:

$$\boxed{\text{NASDAQ} = \frac{\sum\limits_i w_i p_i}{d} = \frac{\sum\limits_i \left(\dfrac{n_i p_i}{\sum\limits_j n_j p_j} \right) p_i}{d}} \qquad (15.3.389)$$

  where p_i is the price of each stock i in the index and n_i is the number of shares publicly available for each stock and d the "NASDAQ Divisor".

- and many other index per country... (like SMI fort Swiss Market Index in Switzerland or CAC40 in France)

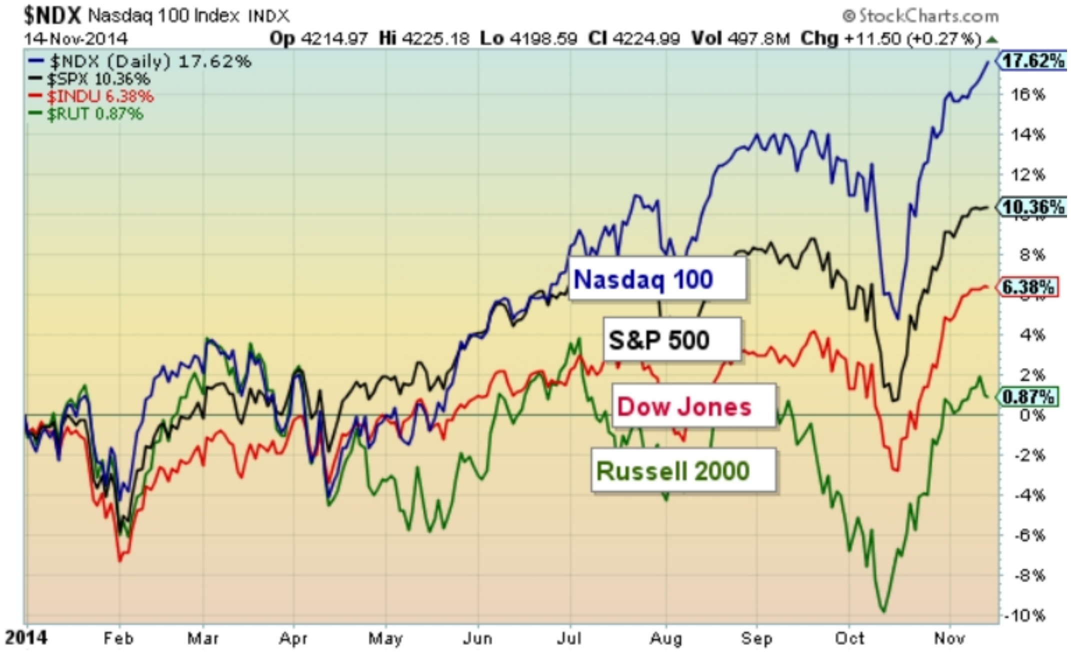


Figure 15.99: **Comparison of most well known market indices (source: StockCharts.com)**

Investment funds monitored on the basis of market indices are named "ETF" for "exchange traded funds" and that's just funds that replicate the performance of well known indices and that do not try (most of time) to outperform them.

D2. A "derivative" is a product / financial instrument, which is also bought and sold, and is still built on the basis of a financial security/asset. This latter is then named "underlying" or "support" of the derivative. These underlying can therefore be stocks, bonds, currencies and ... even derivatives... The danger with derivatives is that when nobody stop superimposing them that nobody knows anymore exactly what were underlying, which is why they are sometimes referred to as "massive damage weapons" in finance (some leaders have eliminated derivatives from their portfolios but this can the opposite effect: increase the risk since the derivatives have been created at the origin to hedge risks... in short... it is not easy to find to good mix!).

D3. "Private equity" is equity capital (remember that in accounting: Equity = Assets - Liabilities) that is not quoted on a public exchange. Private equity consists of investors and funds that make investments directly into private companies.

D4. The "volatility" measure the amplification of the variation of a trade. In other words, a financial security/asset whose volatility is high sees its trade vary greatly, even sometimes exaggerated over time. Conversely, an asset whose volatility is low will see its trade vary slightly and/or fairly consistently. The volatility is often expressed as a percentage in simple mathematical models (because there are several definitions which we will see later). Thus, the volatility of an asset over a specified period is basically defined by:

$$V=\frac{\text{highest trade}-\text{lowest trade}}{\text{average trade}}=\frac{H-L}{\mu} \qquad (15.3.390)$$

There is a lot of other historical volatility indicators (close-to-close volatility, Garman and Klass volatility, Parkinson volatility, Rogers and Satchell volatility, Garman and Klass - Yang and Zhang volatility, Yang and Zhang volatility). We will see them further below.

D5. The "ticker" is the abbreviated name on the markets of any financial instrument (or company name). The most prevalent of these is the "International Securities Identifying Number" (ISIN). An ISIN uniquely identifies a security and its structure is defined in ISO 6166. Securities for which ISINs are issued include bonds, commercial paper, stocks, and warrants. The ISIN code is a 12-character alpha-numerical code that does not contain information characterizing financial instruments, but serves for uniform identification of a security at trading and settlement.

In more complex situations, volatility is often assimilated to the standard deviation (i.e. variance) or semivariance and we will see that later.

We will try here to give an exhaustive view of the following marketable assets:

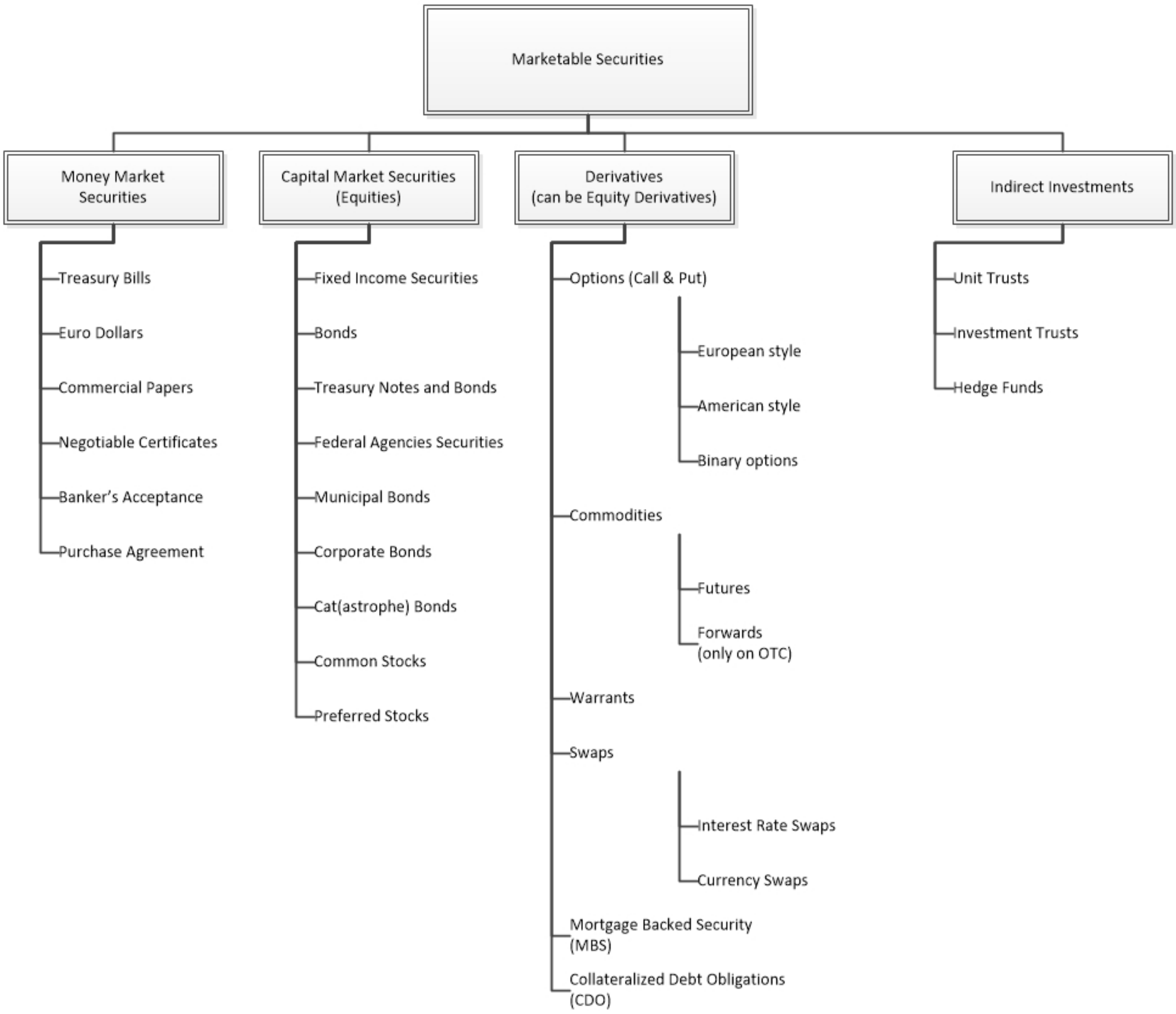

Marketable Securities
Money Market Securities
Capital Market Securities (Equities)
Derivatives (can be Equity Derivatives)
Indirect Investments
Treasury Bills
Euro Dollars
Commercial Papers
Negotiable Certificates
Banker's Acceptance
Purchase Agreement
Fixed Income Securities
Bonds
Treasury Notes and Bonds
Federal Agencies Securities
Municipal Bonds
Corporate Bonds
Cat(astrophe) Bonds
Common Stocks
Preferred Stocks
Options (Call & Put)
European style
American style
Binary options
Commodities
Futures
Forwards (only on OTC)
Warrants
Swaps
Interest Rate Swaps
Currency Swaps
Mortgage Backed Security (MBS)
Collateralized Debt Obligations (CDO)
Unit Trusts
Investment Trusts
Hedge Funds

Figure 15.100: **Marketable Securities (non-exhaustive)**

Fixed Income Securities market is a large part of the financial industry, and it presents unique challenges and opportunities for its practitioners. A large amount of money managed by pension funds and other institution funds is allocated to fixed income investments. Because fixed income has a quite predictable income stream, conservative money managers view it as a safer investment option when compared to socks and more exotic derivatives. This is true because, for equity investments (stocks), for example, it is practically impossible to determine how much money a company will make in a few years from now. With a fixed income investment such as a bond, however, we have a contract that guarantees the return on the investment for a specified period of time.

Clearly, there are also risks in such fixed income investments. A well-known risk is that of default on which we will come back later with mathematical tools. The second big risk, which is frequently overlooked by investors, is that the rate of return will not be able to cope with inflation during the period of investment. This all shows that analysing fixed income investments is not as easy as it initially sound and software are typical tools that help investors in this purpose.

But we will not study non-marketable assets:

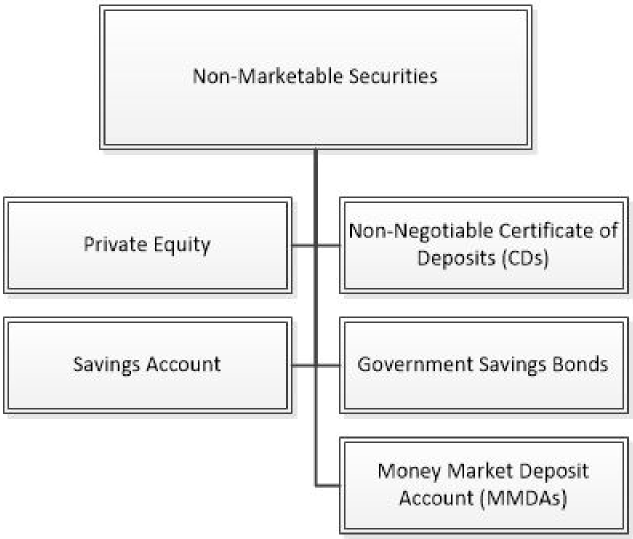

Non-Marketable Securities
Private Equity
Non-Negotiable Certificate of Deposits (CDs)
Savings Account
Government Savings Bonds
Money Market Deposit Account (MMDAs)

Figure 15.101: **Non-Marketable Securities (non-exhaustive)**

We also differentiate a set of various portfolio or funds management strategies:

Definitions (#15.3.681):

D1. "Passive management" also named "Index management" is an investing strategy that tracks a market-weighted index or portfolio. The most popular method is to mimic the performance of an externally specified index and doing this by buying one or more index funds and this is what we name "physical replication". When doing index replication through derivatives we speak about "synthetic replication".

D2. "Active management" refers to a portfolio management strategy where the manager makes specific investments with the goal of outperforming an investment benchmark index.

D3. "Diversified management" refers to a portfolio management strategy where the manager mix obligations and shares in various weights and the latter having different risks profiles in the purpose to maximize the return on investment.

D4. "Alternative management", that is most of time associated hedge funds, is a management strategy without constraints using therefore not long only investments and using strongly leverage effects.

D5. "Overlay management" is an overarching canopy of quantitative analyses and systems that operates an array of investments to achieve a desired net effect for an investor. Overlay management uses software to track an investor's combined position from the separate accounts. Any possible portfolio adjustments will be analysed by the overlay system, which ensures the overall portfolio will remain in balance and prevent any inefficient transactions from occurring.

D6. "Structured management" is a technique of portfolio management based on portfolio insurance techniques. Many of these portfolios are build on thanks to a unique set of formulas.

15.3.10.3.1 Stocks (shares of stocks)/Equities

Owning shares of company profits is one of the most common ways to invest and generate wealth. A large number of people who have made a fortune have achieved it by creating or buying an equity stake in a successful corporation.

Definitions (#15.3.682):

D1. "Stocks" are values papers acknowledging by contract property rights in the capital of an entity known as "public limited-liability company" (PLLC). This contract has a price and is exchanged on the market.

D2. The "capital stock" of a corporation constitutes the equity stake of its owners. It represents the residual assets of the company that would be due to stockholders after discharge of all senior claims such as secured and unsecured debt. Stockholders' equity cannot be withdrawn from the company in a way that is intended to be detrimental to the company's creditors.

Therefore the stock of a corporation is partitioned into shares, the total of which are stated at the time of business formation. Additional shares may subsequently be authorized by the existing shareholders and issued by the company. In some jurisdictions, each share of stock has a certain declared par value, which is a nominal accounting value used to represent the equity on the balance sheet of the corporation. In other jurisdictions, however, shares of stock may be issued without associated par value.

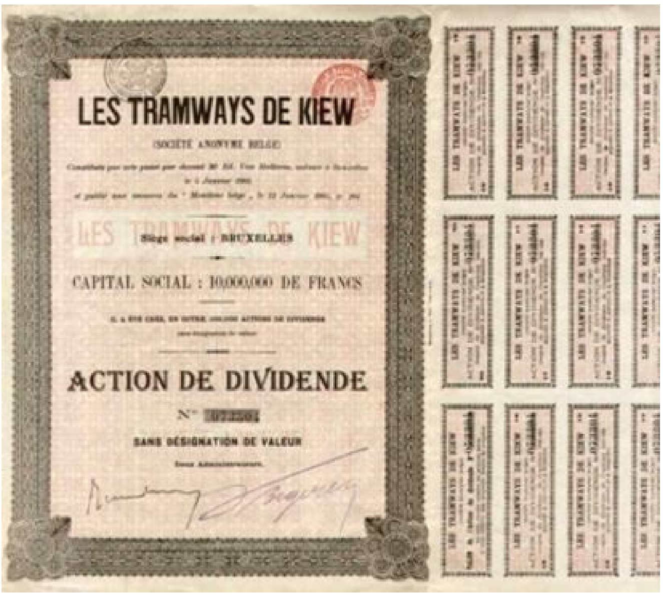
Shares represent a fraction of ownership in a business. A business may declare different types (classes) of shares, each having distinctive ownership rules, privileges, or share values. Ownership of shares may be documented by issuance of a stock certificate. A stock certificate is a legal document that specifies the amount of shares owned by the shareholder, and other specifics of the shares, such as the par value, if any, or the class of the shares.

Remark

"Rule 144 Stock" is a common name given to shares of stock subject to SEC Rule 144: Selling Restricted and Control Securities. Under Rule 144, restricted and controlled securities are acquired in unregistered form. Investors either purchase or take ownership of these securities through private sales from the issuing company or from an affiliate of the issuer. Investors wishing to sell these securities are subject to different rules than those selling traditional common stock. These individuals will only be allowed to liquidate their securities after meeting the specific conditions set forth by SEC Rule 144.

We distinguish in general the following stock shares:

1. The "classic share" or "common share" that gives a vote right at general meetings, the right to information (...) and the right to dividends.



LES TRAMWAYS DE KIEW

(SOCIÉTÉ ANONYME BELGE)

Siège social : BRUXELLES

CAPITAL SOCIAL : 10,000,000 DE FRANCS

ACTION DE DIVIDENDE

N°

SANS DÉSIGNATION DE VALEUR

Figure 15.102: **Example of an old French common stock share paper**

2. The "privileged share" that provides a privilege that can be a priority when voting at general meetings or a priority in the distribution of the dividend.
3. The "preferred share" that only gives privileged access to dividends and no voting rights!.
4. The "share warrant" which are shares that entitle the holder to subscribe for new shares at a given date.

Remark

We also differentiate "bearer stocks" negotiable without restrictions on the stock exchange from "registered stocks" whose value must be negotiated with more or less complex legal restrictions as they contain the name of the stockholder that must be registered into the stockholders register (most of times because of government taxes).

Something very tricky with shares is the "dilution effect". Indeed, when a company issues additional shares (and therefore its stock-capital), this reduces an existing investor's proportional ownership in that company. This is a risk of investing in stocks that investors must be aware of. The new stocks will be (normally... at the opposite of what happened for example at Facebook with Eduardo Saverin) offered to the stockholders of the company at a fixed rate and in proportion of stocks they already own ("subscription right") in order not to penalize against a discount of the value because of the new issued stocks. This will allow them to maintain the percentage of their share capital, and the weight of their voting rights. Dilution can also happen when there is a warrant plan (abusively named "stock-options plan" as we will see further below) as a company therefore grant to employees a large number of optionable securities and these latter exercise the options, common shareholders may be significantly diluted.

✍**Example:**

Assume that a simple business has 10 shareholders, and that each shareholder owns one share, or 10% of the company. If each investor receives voting rights for company decisions based on share ownership, every shareholder has 10% control.

Suppose that the company then issues 10 new shares and that a single investor buys them all up. There are now 20 total shares outstanding, and the new investor owns 50% of the company.

Meanwhile, each original investor now owns just 5% of the company (1 share out of 20 outstanding), because their ownership has been diluted by the new shares.

Because the earnings power of every share is reduced when convertible shares are executed, investors may want to know what the value of their shares would be if all convertible securities were executed.

Definition (#15.3.683): The "diluted earnings per share (EpS)" is calculated by firms and reported in their financial statements and is the value of earnings per share if equity warrants (see definition further below), stock options (see definition further below) and convertible bonds were converted to common shares. In a given point of the EpS is like a pessimistic value of a share.

15.3.10.3.1.1 Dividend Yield

There are several types of "dividend yield" based on the context that are intuitive for some and for others quite complex! Here are four of the most common one (we will see other more complex one later...) to our knowledge when introducing for the first time financial mathematics:

Definitions (#15.3.684):

D1. There is the ratio, expressed as a percentage, named "yield" or "rate of return" (not to be confused with "yield to maturity" that we will see further below), between the dividend per share distributed by a company and the share price on the stock markets of the company at the time the dividend is paid (some practitioners sometimes take the arithmetic average of the dividends paid over several periods):

$$R_B = \frac{\text{Dividend paid in average}}{\text{Share stock market price}} = \frac{\bar{D}}{P_{SS}} \qquad (15.3.391)$$

D2. There is the ratio, expressed in cash value by year and named "share stock", of the difference between the sale trading value (bid) of the stock with its dividends values added and the buy trading value (ask) of stock on the number of periods (mainly expressed in years):

$$R_B = \frac{\left(\text{Bid} + \sum \text{Dividends}\right) - \text{Ask}}{\text{Number of periods}} = \frac{B - A + \sum D}{N} \qquad (15.3.392)$$

Obviously if we divide the result of this latter annual cash yield by the initial capital invested, we get the return in percentage.

D3. There is the "implicit rate of return" (IRR) which is the ratio between the net profit of the company and its stock market capitalization value (the implicit rate or return abbreviation can be confusing with the "internal rate of return" that we have seen in the section of Quantitative Management Techniques this is why we avoid to use its abbreviation in practice):

$$\text{IRR}=\frac{\text{Net profit}}{\text{Stock Market capitalization value}}=\frac{\text{NP}}{S_{\mathcal{MV}}} \tag{15.3.393}$$

The IRR is a kind of performance yield between the potential income of a stock and its price (caution! it would be a performance yield only if we took into account the ratio dividends/stock price that can be very different). We assume, rightly or wrongly, that the company has a better long-term use in making profits than distribute them as dividends...

D4. The "Price Earnings Ratio" (PER) is simply define as at a time t by:

$$\text{PER}_t=\frac{\text{Market price per equity share}}{\text{Earning per share}}=\frac{P_t}{\bar{D}_{12}} \tag{15.3.394}$$

Therefore the PER also gives the number of years needed to pay back the share only with the dividends (obviously under the assumption of constant dividends). So strictly speaking, the ratio is measured in years, since the price is measured in dollars and earnings are measured in dollars per year.

The PER is generally quoted in the press is the one calculated on the last published annual profit of a company. However analysts are often based their judgement on the expected profit for the current year.

Here is a small table that can illustrated a possible interpretation of the PER:

PER	Quality
0 to 10	The share is undervalued or the profits of the company are expected to soon to be in decline
10 to 17	For most companies, a ratio that fall within that range is considered good
17 to 25	The share is overvalued or there is growth in profits for the latest announcements
>25	It is likely that high profits are expected in the future, or the share is the subject of a speculative bubble.

Table 15.45: **Qualitative judgements of PERs commonly accepted**

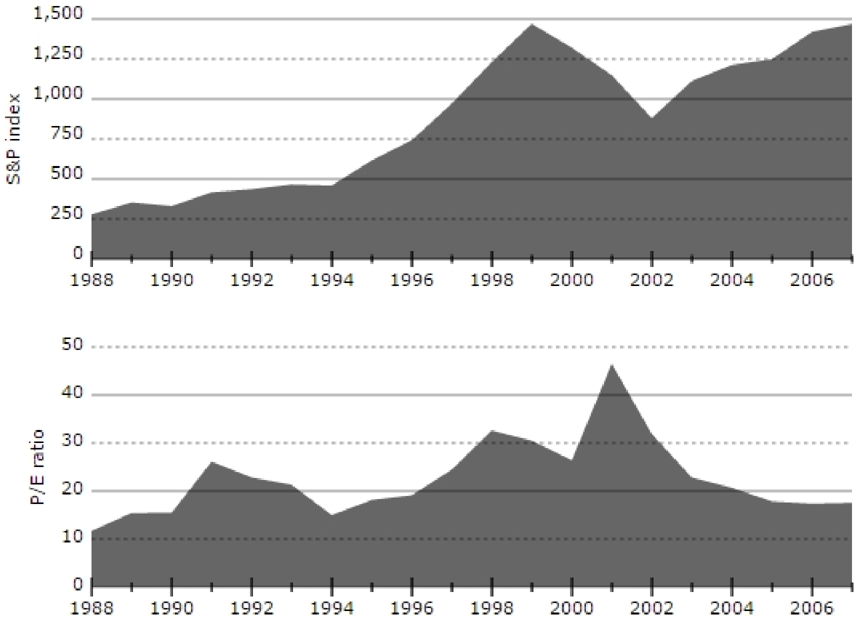


Figure 15.103: **Parallel graphs of the S&P index and PER for the same period (source: Wikipedia)**

Remark

Once a customer asked me a list of common financial indicators. In fact you can have an empirical choice with a poor scientific point of view of such a list by the UBS (Swiss Banks Union) named: *The* 100 *most important financial performance indicators* that you can order on the web (see [549]).

Examples:

E1. Consider a repayable stock purchased six years before for the price of 10.- (invested capital). The investor sell it the 12.50.-. The investor received before three times a dividend: three times 2.20.- and one time 1.-. Both yields then give respectively:

$$
\begin{aligned}
R_B &= \frac{\bar{D}}{P_{SS}} = \frac{\dfrac{7.60}{4}}{12.5} \\
R_B &= \frac{B - A + \sum D}{N} = \frac{12.5 - 10 + 7.6}{6} = 1.68.\text{-[year}^{-1}]
\end{aligned}
\qquad (15.3.395)
$$

and if we divide this latter by the initial invested capital (10.-) we get 16.8%.

E2. The maximum price (including expenses) we can put up to buy a stock of 500.- with a relating dividend of 12% in comparison to a similar investment of the same price with a yield of 5% with same periodicity is $1,200$.-. Indeed, the latter stock would bring 60.- dividend at each period (12% of 500.-). The sum that should be set to have the same interest at a rate of 5% is $1'200$.-. Indeed 5% of $1,200$.- being equal also to 60.-.

E3. The stock issued by a company (whose capital consists of 10 million shares) is valuated at 10.-, bringing the stock market value of the company at 1 billion. The planned net profit is 50 million for the current fiscal year, that is to say 5.- per stock (under the assumption that all profits are distributed into dividends...). The division of the net profit by the stock market value gives us a price earning ratio (PER) of 20.

15.3.10.3.1.2 Shares Benchmark Indices

We consider an index composed of n shares. Given $P_i(t)$ the price of the share i and $R_i(t)$ the corresponding yield for the period $[t-1,t]$:

$$R_i(t)=\frac{P_i(t)}{P_i(t-1)}-1 \tag{15.3.396}$$

Remark

That latter relation must not be confused with the common definition of "rate of change" in many statistical softwares and defined by:

$$\text{ROC}_i(t)=\frac{P_i(t-1)}{P_i(t)}-1=\frac{P_i(t-1)-P_i(t)}{P_i(t)} \tag{15.3.397}$$

The value of the index (or "benchmark") $B(t)$ is defined for $t\geq 1$ by:

$$B(t)=B(t-1)\sum_{i=1}^{n}w_i(t)(1+R_i(T)) \tag{15.3.398}$$

with $B(0)=1$, where $w_i(t)$ is the weight of the share i in the index such that $\sum_i w_i(t)=1$. The value of $B(t)$ is generally calculated on market closure.

The majority of market indices use obviously a weight based on the traded capitalization:

$$w_i(t)=\frac{N_i(t)P_i(t)}{\displaystyle\sum_{j=1}^{n}N_j(t)P_j(t-1)} \tag{15.3.399}$$

where $N_i(t)$ is the number of shares i issued (some indices as the Nikkei use weight based on the prices and not on the quantity.

15.3.10.3.1.3 Durand Model

To conclude this small introduction on the stocks, note that a theoretical model of current evaluation ("current" meaning "at this date") of the stock is named the "Durand model". The underlying idea is quite simple: the price of a stock P_0 today is equal to the sum of the cash flows (and therefore of its dividends D_t) actualized using market geometric rates $t_i\%$ (or expected/required by the stockholder) of the corresponding period i paid at each period k (therefore we assumed we are in a non-probabilistic scenario...) as well as its future resale price P_T also actualized.

That is to say formally:

$$P_0=\frac{P_T}{\prod\limits_{i=1}^{T}(1+t_i\%)}+\sum_{t=1}^{T}\frac{D_t}{\prod\limits_{i=1}^{k}(1+t_i\%)} \qquad (15.3.400)$$

If we make the usual high-school level assumption that the rate is always constant (if not, we will use a geometric average of the rates on the totality of the time or better... use a Monte Carlo simulation), we then have:

$$\boxed{P_0=\frac{P_T}{(1+t\%)^T}+\sum_{t=1}^{T}\frac{D_t}{(1+t\%)^t}} \qquad (15.3.401)$$

this relation is known under the name "Irving Fisher fundamental value".

But the resale price at time T will be equal to this same relation and so on to infinity, because a stock is not intended to be repaid! Then we have:

$$P_0=\frac{P_T}{(1+t_i\%)^{+\infty}}+\sum_{t=1}^{+\infty}\frac{D_t}{(1+t\%)^t} \qquad (15.3.402)$$

and as:

$$\frac{P_T}{(1+t_i\%)^{+\infty}}\rightarrow 0 \qquad (15.3.403)$$

Then it remains:

$$\boxed{P_0=\sum_{t=1}^{+\infty}\frac{D_t}{(1+t\%)^t}} \qquad (15.3.404)$$

that is also sometimes written:

$$P_t=\sum_{t_m\geq t}^{+\infty}\frac{D_{t+m}}{(1+r)^m} \qquad (15.3.405)$$

and is named the "discount cash flows discount model".

Now let us consider a simplification of the Durand model named the "Gordon-Shapiro model" or "Dividend Discount Model" (DDM). This latter model considers that at each period, dividends are growing at a same rate denoted $t_{cd}\%$ by:

$$D_t=D_0(1+t_{cd}\%)^t=D_1(1+t_{cd}\%)^{t-1} \qquad (15.3.406)$$

By applying what we have just proved in the Durand's model:

$$\begin{aligned}P_0&=\sum_{t=1}^{+\infty}\frac{D_t}{(1+t\%)^t}=\sum_{t=1}^{+\infty}\frac{D_1(1+t_{cd}\%)^{t-1}}{(1+t\%)^t}=\frac{D_1}{1+t\%}\sum_{t=1}^{+\infty}\frac{(1+t_{cd}\%)^{t-1}}{(1+t\%)^{t-1}}\\&=\frac{D_1}{1+t\%}\sum_{t=1}^{+\infty}\left(\frac{1+t_{cd}\%}{1+t\%}\right)^{t-1}\end{aligned} \qquad (15.3.407)$$

Let us put:

$$q=\frac{1+t_{cd}\%}{1+t\%} \quad \text{and} \quad t-1=n \tag{15.3.408}$$

Then we have:

$$P_0=\frac{D_1}{1+t\%}\sum_{t=1}^{+\infty}\left(\frac{1+t_{cd}\%}{1+t\%}\right)^{t-1}=\frac{D_1}{1+t\%}\sum_{t=1}^{+\infty}q^n \tag{15.3.409}$$

We have proved in the section Sequences and Series (see page 1679) that any geometric sequence of reason x was given for recall by:

$$S_n=\sum_{i=0}^{n}q^n=\frac{1-q^{n+1}}{1-q} \tag{15.3.410}$$

Then if n approaches infinity and under the assumption that:

$$q=\frac{1+t_{cd}\%}{1+t\%}<1 \tag{15.3.411}$$

that is to say, the rate of return expected by stockholders exceeds the dividend growth rate, then we have:

$$S_{+\infty}=\sum_{i=0}^{+\infty}q^n=\frac{1-q^{+\infty}}{1-q} \tag{15.3.412}$$

Therefore:

$$P_0=\frac{D_1}{1+t\%}\sum_{t=1}^{+\infty}q^n=\frac{D_1}{1+t\%}\frac{1}{1-q}=\frac{D_1}{1+t\%}\frac{1}{1-\dfrac{1+t_{cd}\%}{1+t\%}}=\frac{D_1}{1+t\%}\frac{1}{\dfrac{1+t\%-1-t_{cd}\%}{1+t\%}} \tag{15.3.413}$$

Hence in the end:

$$\boxed{P_0=\frac{D_1}{t\%-t_{cd}\%}} \tag{15.3.414}$$

The dividend growing rate $t_{cd}\%$ is determined either from historical data of the stock, either from the forecasts of analysts on future dividends.

Note that we find often this latter relation in the following forms in the literature:

$$\boxed{P_0=\frac{D_0(1+t_{cd}\%)}{t\%-t_{cd}\%}=\frac{D_0(1+g)}{t\%-g}=\frac{\text{DIV}_0(1+g)}{r-g}=\frac{\text{DPS}_1}{r-g}} \tag{15.3.415}$$

Practitioners sometimes use the last expression in the form:

$$t\%:=r_{\mathcal{E}}=\frac{D_0(1+g)}{P_0}=\frac{D_1}{P_0}+g=\frac{D_1}{P_0}+t_{cd}\% \tag{15.3.416}$$

to compare the yield $t\%$ of various stock knowing their dividend, their current price and forecasted expected dividend growth rate.

In the last relation the $\mathcal{E}$ in subscript of the r stands for "Equity" (see section Quantitative Management page 7604).

Because $r_{\mathcal{E}}$ is the return that the shareholders require on the stock, it can be interpreted as the "firm's cost of equity capital".

In the case where $t_{cd}\%$, that is to say in the case of "preferred stocks", that are stocks whose holders do not benefit from increases in the firms earnings and they generally cannot vote in corporate elections, the latter relation reduces obviously to:

$$P_0=\frac{D_0}{t\%} \tag{15.3.417}$$

✍**Example:**

We want to valorize a stock that pays a first dividend of 5.- and whose supposed growth is assumed constant (to infinity...) with a yield (geometric mean one) of 14.87% compared to an average geometric return of the market of 20%. We then have the price of this action that can be estimated by:

$$P_0=\frac{D_1}{t\%-t_{cd}\%}=\frac{5}{0.2+0.1487}\cong 14.339.\text{-} \tag{15.3.418}$$

Let us notice, to close this topic about the valuation of stocks, an important property (that also applies to obligations that we will see just further below) and that is the "risk of credit default". This risk, which must be quantified at best in terms of probabilities, simply consist in the fact that the issuer of such stocks may not meet its obligations and pay the money that we are entitled to expect of him simply because it is bankrupt or that the economic system does not lend itself (major economic crisis for example). The Governments normally guarantee only a fraction of the expected gain (guarantee that must also mathematically be actualized). Therefore, you should know that:

1. The real actualized price is still slightly lower than the simple model above, since it is related to an event with probability not equal to 100% (but less!).
2. The actualized price is in reality a weighted mean by the probabilities of the economic scenarios.
3. All algebraic economic models are idealized and you should never forget that (they do not into account macro-economic phenomena or when they do they actual do it bad most of time)!

The price-to-earning ratio (P/E) is perhaps the most widely recognized valuation indicator, familiar to readers of newspaper financial tables and institutional research reports. Using the Gordon growth model, an expression for P/E in terms of the fundamentals can be developed.

If we divide the Irving Fisher fundamental value by the current-year's earnings per share, E_0, we get:

$$\frac{P_0}{E_0}=\frac{D_0(1+g)/E_0}{t\%-g}=\frac{D_1/E_0}{t\%-t_{cd}\%} \tag{15.3.419}$$

This expression is for "trailing price-to-earning ratio", which is current price divided by trailing (current year) earnings.

✍**Example:**

We want to valorize a stock that pays a first dividend of 5.- and whose supposed growth is assumed constant (to infinity...) with a yield (geometric mean one) of 14.87% compared to an average geometric return of the market of 20%. The trailing annual earning per share is assumed to be 2.-. We then have the price of this action that can be estimated by:

$$\frac{P_0}{E_0}=\frac{D_1/E_0}{t\%-t_{cd}\%}=\frac{5/2}{0.2+0.1487}\cong 7.169 \tag{15.3.420}$$

15.3.10.3.2 Obligations (Bonds)

Unlike the individual loan (indivis loan), the loan named "bond issue" involves many lenders, named "subscribers", who receive in exchange for money lent, securities named "obligations".

Definition (#15.3.685): "Obligations", also named "bonds" or "financial commitment", are values papers (debt securities of an issuer) establishing through a contract debt securities (capital loaned) to an investor and that pay a fixed interest (usually in the form of annual coupons like dividends to) to the holder during a defined period of time (the initial amount invested is repaid to a deadline specified in the contract). This contract has a price (depending on the date!). It is most of time redeemable on the market and the debtor is obliged to pay interest (remuneration by coupons). Moreover, if the obligation is a "convertible bond" it entitles the creditor to obtain either a refund of the duty, or conversion into stocks, following procedures set in advance in the contract.

Stocks and Bonds are very different in what they represent. While the stock (auction) means a title related to the share capital of a company and vote right, coted or not (negotiable or not) without expiration and depending on company performance (long term security), the bond is a debt security (also negotiable or not depending on contract terms) limited in time and whose interest is considered a cost doing business and therefore tax deductible (at the opposite of stocks). The bond is based on the debt of a company, state, or local community while the stock (auction) is part of the equity of a corporation. Any company composed of stocks is not necessarily coted on the stock market, and any company is not necessarily a stock company. The issuance of a bond allows the issuer to diversify its sources of loans, and for a stock, to diversify its sources of funding. Also in case of bankruptcy of the issuer the holder of an obligation has priority over stockholders and can make a legal recourse in case if payments are missed (stockholders cannot make legal recourse!) then risk is less.

Remark

Some bonds have their coupons which are not disclosed in cash but are based on the level of economic indices as we will see later.

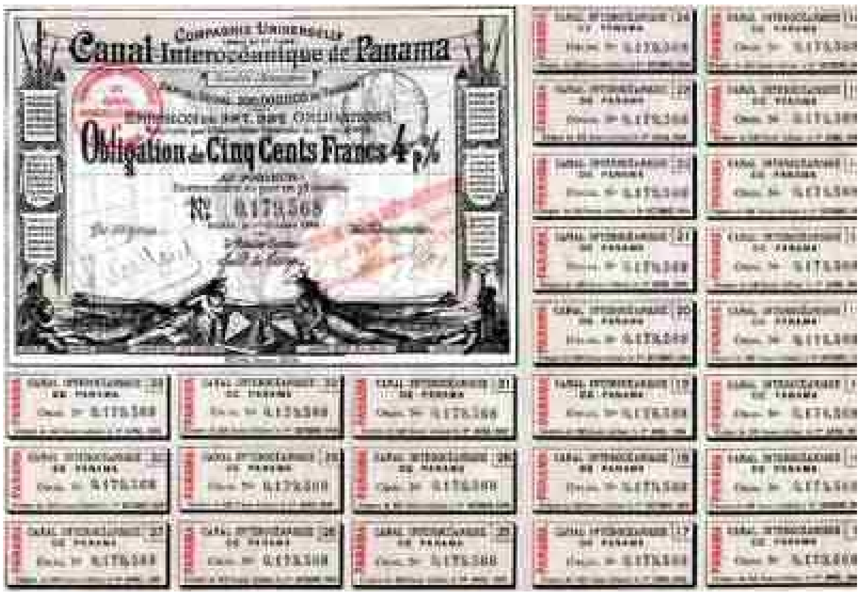
We distinguish three main types of bonds (actually we practitioners differentiate more than 20 types of bonds but that are most of times special cases of the three main one such as some that we already mentioned above such as: convertible bonds, bearer bonds, exchangeable bonds, perpetual bonds...):

1. "Fixed rate bonds" or "common obligation" which is the traditional bonds (it represents in the early 12000s (holocene calendar) about 85% of the bond market). It provides a cash flow definitely fixed of interest when issued (coupons) in a predefined intervals until maturity (which is safe) and whose mathematical interest rate corresponding to this flow is named "yield to maturity[19]" (YTM). Financial practitioners often refer to as "plain vanilla bond". When the duration of the flow can be considered infinite, we speak of "perpetual

---

[19] So in other words, the YTM measures the interest rate, as implied by the bond, that takes into account the present values of all the future coupon payments and the principal. It is assumed that bond holders can invest received coupons at the YTM until the maturity of the bond; according to risk-neutral expectations, the payments received should be the same as the price paid for the bond.

obligation". This is however not a safe investment as we will see in a simple numerical example further below.



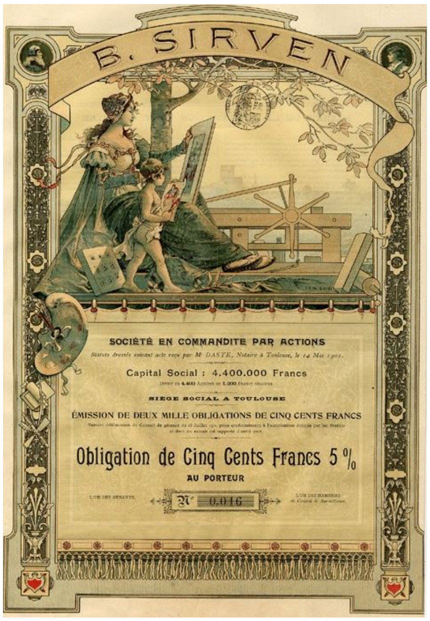
Compagnie Universelle

Canal Interocéanique de Panama

Obligation de Cinq Cents Francs 4 p%

Figure 15.104: **Example of bond with fixed rate with fixed number of coupons (at the time of paper...)**

When the duration of the cash flow can be considered infinite or really is unlimited, we speak of "perpetual obligation". This is however obviously not a safe investment that we need to quantify as we will also see in a simple numerical example further below.



B. SIRVEN

SOCIÉTÉ EN COMMANDITE PAR ACTIONS

Statuts dressés suivant acte reçu par Me DASTE, Notaire à Toulouse, le 14 Mai 1901.

Capital Social : 4.400.000 Francs

SIÈGE SOCIAL A TOULOUSE

ÉMISSION DE DEUX MILLE OBLIGATIONS DE CINQ CENTS FRANCS

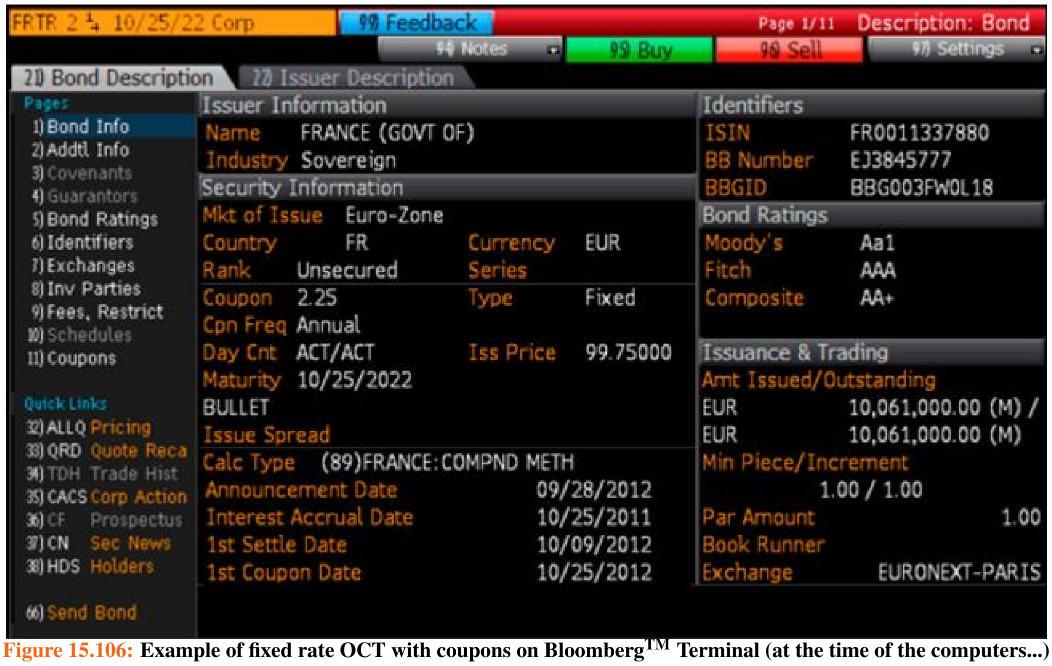
Obligation de Cinq Cents Francs 5 %

AU PORTEUR

No 0.016

Figure 15.105: **Example of perpetual bond with fixed rate (at the time of paper...)**

and in computer versions with the Bloomberg$^{\text{TM}}$ interface for an OCT bond (Obligation Comparable to a Treasury Bond):


FRTR 2 ¼ 10/25/22 Corp
99) Feedback
Page 1/11 Description: Bond
94) Notes
95) Buy
96) Sell
97) Settings
21) Bond Description
22) Issuer Description
Pages
1) Bond Info
2) Addtl Info
3) Covenants
4) Guarantors
5) Bond Ratings
6) Identifiers
7) Exchanges
8) Inv Parties
9) Fees, Restrict
10) Schedules
11) Coupons
Quick Links
32) ALLQ Pricing
33) QRD Quote Reca
34) TDH Trade Hist
35) CACS Corp Action
36) CF Prospectus
37) CN Sec News
38) HDS Holders
66) Send Bond

Issuer Information
Name FRANCE (GOVT OF)
Industry Sovereign

Security Information
Mkt of Issue Euro-Zone
Country FR
Currency EUR
Rank Unsecured
Series
Coupon 2.25
Type Fixed
Cpn Freq Annual
Day Cnt ACT/ACT
Iss Price 99.75000
Maturity 10/25/2022
BULLET
Issue Spread
Calc Type (89)FRANCE:COMPND METH
Announcement Date 09/28/2012
Interest Accrual Date 10/25/2011
1st Settle Date 10/09/2012
1st Coupon Date 10/25/2012

Identifiers
ISIN FR0011337880
BB Number EJ3845777
BBGID BBG003FW0L18

Bond Ratings
Moody's Aa1
Fitch AAA
Composite AA+

Issuance & Trading
Amt Issued/Outstanding
EUR 10,061,000.00 (M) /
EUR 10,061,000.00 (M)
Min Piece/Increment
1.00 / 1.00
Par Amount 1.00
Book Runner
Exchange EURONEXT-PARIS

Figure 15.106: Example of fixed rate OCT with coupons on Bloomberg$^{\text{TM}}$ Terminal (at the time of the computers...)

The fixed rate bond is typically quoted in prices or rates. The fixed rate bond is valued by actualizing future cash flows it delivers [20].

2. "Floating rate bonds" whose interest cash flows, but not the redemption price, are indexed to a benchmark rate such a as the policy rate of a central bank, the results of a company, or any other. The risk associated with such variable rate is named "interest rate risk". For the holder of a bond portfolio that wishes to protect its capital, it is sufficient to immunize the portfolio against fluctuations in the rate. This is what we name "duration hedging" and that we will detail further below.

3. "Zero-coupon bonds" (also sometimes named "discount bonds") that has only two financial flows: an initial flow (purchase) and a final flow (the unique coupon and the Nominal), without any interim payment (hence its name because it does not pay coupon meantime or coupons to 0% ... and only nominal is paid at maturity) and therefore the valuation calculation thereof require to know only the simple interest. This is most of time the least risky of all obligations since it pays one coupon and therefore its effective yield is equal to its original actuarial rate (since there can be reinvested in the meantime). The buyer purchased the bond at a value less than the face value, price which is paid at maturity of the contract. The zero-coupon is usually indexed on the inflation index (in this special case it belongs to the floating rate bonds family).

Remark

A short-term bond with a maturity of less than one year is named "treasury bond" ("T-bill" in the U.S.A. or "BTF" in France). On the same model as the zero-coupon bonds the T-Bill do not pay interest before maturity but instead are sold at a discount to their face value which allows the subscriber to obtain a benefit at maturity. Most of time T-Bill are reserved to bonds emitted by governments.

---

[20] Unlike the NYSE (New York Stock Exchange), where many stocks are traded, there is no particular physical locations where bonds are traded. Instead, they are traded electronically. The Financial Industry Regulatory Authority (FINRA) has made an effort to make bond prices more widely available. Their website, `www.finra.org/marketdata`, allows anyone to search for the most recent trades and quotes for bonds.

Bonds are characterized by several properties:

P1. The name of the issuer with its abbreviated name (ticker).

P2. Their "basis currency" that could also fluctuate on a global market.

P3. Their "deadline" or "maturity date" that will give the possibility, depending on their date of issue and the type of calendar, to know the actualized estimated value of the obligation at any time.

P4. Their "face value", named also the "pair", designates the value used to calculate the interests (but most of time it is implicit as all calculations are normalised to a face value of 100.-).

P5. Their "coupon rate" or "nominal rate" associated with their "frequency" (often annual or semi-annual) that gives the possibility to define the interest named the "coupon" or "dividend coupon" applied on the nominal value of a bond that will be paid to the subscriber to the date named "date of record". Normally the calculation method of the interest rate must be reported.

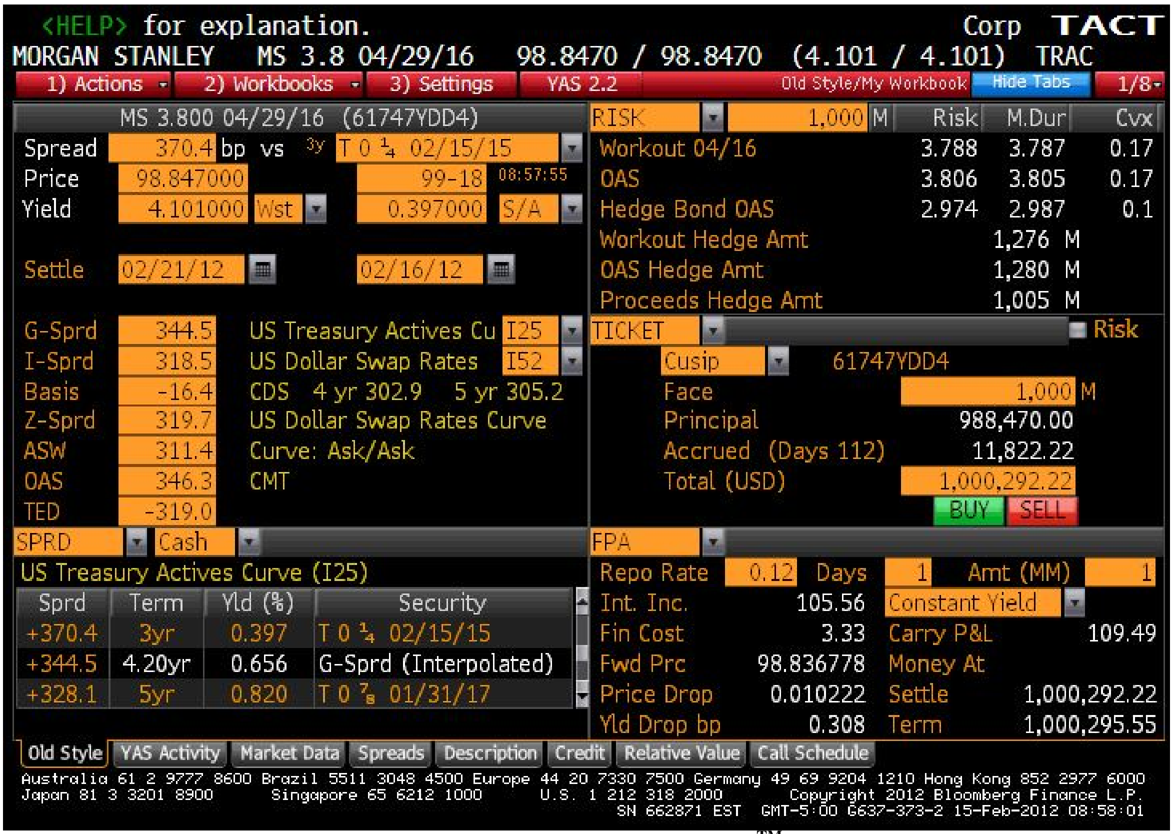
P6. Their "issue price" or "subscription price" or just "price" or even "last price" is the amount to pay today by the subscriber (investor) for owning a bond. The issue of bonds is therefore "at par" if the nominal value is equal to the amount requested for acquisition. The subscription price is below par if the requested amount is less than the nominal (most common), but the price can also sometimes be above par! The difference between nominal value and subscription price is named the "share premium".

The issue price is not always equal to the nominal value in order to make investor to want buy without necessarily offer him a nominal interest rates too high. Thus, the issue of a bond is a subtle game for the issuer between the value of nominal interest rate and the issue price.

P7. Their "redemption price" or "redemption value" is the amount really paid to the borrower on repayment of the bond at maturity. The reimbursement may be made at par or sometimes above when at maturity (in fine), in installments or never (perpetual bonds). The difference between the redemption value and the nominal is named the "redemption premium".

Bond investors prefer in general short maturities because the rate of return reflects more accurately the investor enrichment if he can reinvest each coupon at the same rate and retains the obligation upon maturity. This scenario is obviously more difficult to guarantee on the long term. In addition there is the "risk of default", that is to say, the risk of bankruptcy of the issuer.

Conversely and obviously, issuers have usually a preference for longer maturities, which allow them to spread their debts over time. The divergence between demand (investors) and supply (issuers) results in yields generally lower on the short-term than for medium and long term (this is trivial as the risk is lower for short-term bonds). The "yield curve" which is a function at a given date and for each maturity on the x-axis shows the level of interest rates associated with the orderly financial futures and therefore typically has a growing form.



MORGAN STANLEY MS 3.8 04/29/16 98.8470 / 98.8470 (4.101 / 4.101) TRAC

MS 3.800 04/29/16 (61747YDD4)
Spread 370.4 bp vs 3y T 0 ¼ 02/15/15
Price 98.847000 99-18 08:57:55
Yield 4.101000 Wst 0.397000 S/A
Settle 02/21/12 02/16/12

G-Sprd 344.5 US Treasury Actives Cu I25
I-Sprd 318.5 US Dollar Swap Rates I52
Basis -16.4 CDS 4 yr 302.9 5 yr 305.2
Z-Sprd 319.7 US Dollar Swap Rates Curve
ASW 311.4 Curve: Ask/Ask
OAS 346.3 CMT
TED -319.0

RISK 1,000 M

	Risk	M.Dur	Cvx
Workout 04/16	3.788	3.787	0.17
OAS	3.806	3.805	0.17
Hedge Bond OAS	2.974	2.987	0.1

Workout Hedge Amt 1,276 M
OAS Hedge Amt 1,280 M
Proceeds Hedge Amt 1,005 M

TICKET Risk
Cusip 61747YDD4
Face 1,000 M
Principal 988,470.00
Accrued (Days 112) 11,822.22
Total (USD) 1,000,292.22

SPRD Cash
US Treasury Actives Curve (I25)

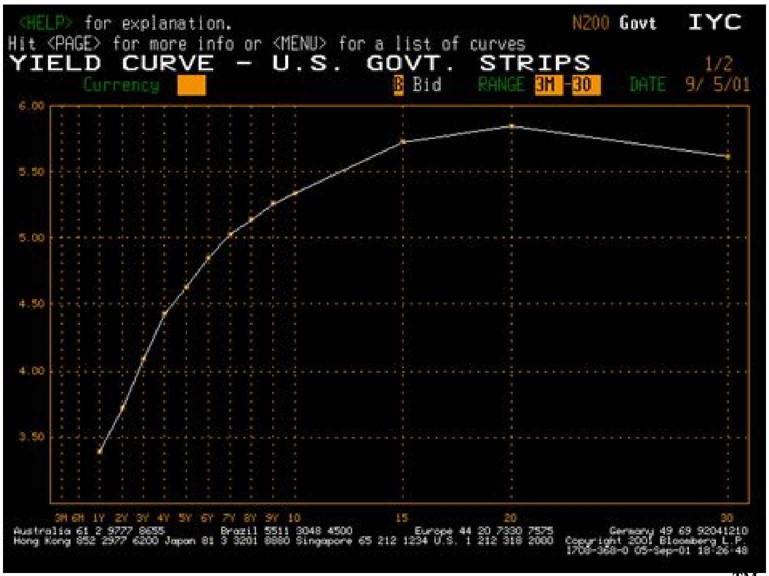
Sprd	Term	Yld (%)	Security
+370.4	3yr	0.397	T 0 ¼ 02/15/15
+344.5	4.20yr	0.656	G-Sprd (Interpolated)
+328.1	5yr	0.820	T 0 ⅞ 01/31/17

FPA
Repo Rate 0.12 Days 1 Amt (MM) 1
Int. Inc. 105.56 Constant Yield
Fin Cost 3.33 Carry P&L 109.49
Fwd Prc 98.836778 Money At
Price Drop 0.010222 Settle 1,000,292.22
Yld Drop bp 0.308 Term 1,000,295.55

Australia 61 2 9777 8600 Brazil 5511 3048 4500 Europe 44 20 7330 7500 Germany 49 69 9204 1210 Hong Kong 852 2977 6000
Japan 81 3 3201 8900 Singapore 65 6212 1000 U.S. 1 212 318 2000 Copyright 2012 Bloomberg Finance L.P.
SN 662871 EST GMT-5:00 G637-373-2 15-Feb-2012 08:58:01

Figure 15.107: Typical Bond details on Bloomberg$^{\text{TM}}$ Terminal

For example the zero-coupon rate curve of U.S.A. bonds as to the 5 September 12001 (holocene calendar) below comparing to zero-coupon U.S.A. Treasury bonds (as the reader can see it!) where there is no interpolation - or "stripping" as say sometimes financial practitioners... - but only a simple straight line between points while some programs offer various stripping based on polynomials or splines or more complicate models that we will see later):



```
<HELP> for explanation.                                    N200 Govt  IYC
Hit <PAGE> for more info or <MENU> for a list of curves
YIELD CURVE - U.S. GOVT. STRIPS                                   1/2
        Currency                       B Bid   RANGE 3M -30   DATE 9/ 5/01

Australia 61 2 9777 8655        Brazil 5511 3048 4500        Europe 44 20 7330 7575        Germany 49 69 92041210
Hong Kong 852 2977 6200 Japan 81 3 3201 8880 Singapore 65 212 1234 U.S. 1 212 318 2000  Copyright 2001 Bloomberg L.P.
                                                                                          1708-368-0 05-Sep-01 18:26:48
```

Figure 15.108: Example of yield curve for zero-coupon USA bonds (T-Bills) in Bloomberg$^{\text{TM}}$ Terminal

At a given date, in a given country or in an unified economic zone, there is a multitude of interest rate curves!

When the yield curve is flat, we speak then logically of "flat curve" when it is growing of "upward sloping curve" and when it is decreasing of "downward sloping curve".

Note that the "periodic zero yield", also sometimes named "zero coupon yield to maturity" (or "yield to maturity" as we have already mentioned it), at maturity of n units of period of time for a zero-coupon bond with nominal C and issue price P can easily be obtained from the relation of the compound interest in the deterministic case:

$$C=(1+t_0\%)^n \tag{15.3.421}$$

Relation that we can frequently found in the literature in the following forms:

$$\begin{gathered}C_T=\text{PV}(1+y)^T\\ \text{FV}=\text{PV}(1+y)^T\end{gathered} \tag{15.3.422}$$

with PC meaning as we already know "Present Value" and FV "Future Value" and where the rate is denoted y to mean "yield"... (obviously when we reverse the parenthesis we talk as usual of "actualization factor" or " discount factor" but this is not new for us as we saw in our study of annuities and loans earlier).

Therefore:

$$t_0\%=\left(\frac{C}{P}\right)^{1/n}-1 \tag{15.3.423}$$

What we sometimes find in the literature rather in the following traditional form (long live to the lack of standards for mathematical notations in finance...!):

$$\boxed{z(T)=\left(\frac{N}{P(T)}\right)^{1/T}-1} \tag{15.3.424}$$

✍**Example:**

The zero-coupon yield to 4 years .... corresponding to a nominal zero-coupon of 100.- of issue price 90.- (under par price) is:

$$t_0\% = \left(\frac{C}{P}\right)^{1/n} - 1 = \left(\frac{100}{90}\right)^{1/4} - 1 \cong 2.67\% \tag{15.3.425}$$

But let us consider a more explicit and realistic case with the following bond rates (remember nominal rate is 100.-):

Bond	Coupon	Maturity	Price
Bond 1	5.0	1	101.00.-
Bond 2	5.5	2	101.50.-
Bond 3	5.0	3	99.00.-
Bond 4	6.0	4	100.00.-

The zero coupon bond prices that we will denote here $B(0,t)$ can be found with the following system of equations:

$$\begin{aligned}101 &= 105B(0,1)\\ 101.5 &= 5.5B(0,1) + 105.5B(0,2)\\ 99 &= 5B(0,1) + 5B(0,2) + 105B(0,3)\\ 100 &= 6B(0,1) + 6B(0,2) + 6B(0,3) + 106B(0,4)\end{aligned} \tag{15.3.426}$$

Then we get:

$$B(0,1) = 0.9619, B(0,2) = 0.9114, B(0,3) = 0.85363, B(0,4) = 0.789 \tag{15.3.427}$$

Their corresponding zero coupon yield rates are given by the relation:

$$B_{(0,t)} = \frac{1}{(1 + r_{(0,t)})^t} \tag{15.3.428}$$

Thus:

$$r_{(0,1)} = 3.96\%, r_{(0,2)} = 4.717\%, r_{(0,3)} = 5.417\%, r_{(0,4)} = 6.103\% \tag{15.3.429}$$

Remark

Among the bonds, only fixed-rate zero-coupon bonds allow to eliminate almost surely the risk rate between two dates (but not risk of bankruptcy). A conventional fixed rate obligation actually generates as much additional rate risk that it has intermediate financial cash flows: the reinvestment rate of each of the coupons from the date of payment and the final repayment date is in fact unknown even if it is implicit in the price of the bond.

There is another common and very convenient way to write the current value of a fixed rate zero-coupon bond! Recall that we have proved earlier above under certain conditions that:

$$C_n = C_0(1 + t_0\%)^n = C_0 e^{nt_n\%} \tag{15.3.430}$$

Therefore:

$$P=\frac{C}{(1+t_0\%)^n}=Ce^{-nt_n\%} \tag{15.3.431}$$

✍**Example:**

The zero-coupon yield to 10 years .... corresponding to a nominal zero-coupon of 100.- of issue price 55.3895.- (under par price) is:

$$t_0\%=\left(\frac{C}{P}\right)^{1/n}-1=\left(\frac{100}{55.3895}\right)^{1/10}-1\cong 6\% \tag{15.3.432}$$

In continuous interest rate it gives (as proved already earlier above) the "continuous actuarial rate zero coupon":

$$t_n\%=-\frac{1}{n}\ln\left(\frac{P}{C}\right)=-\frac{1}{10}\ln\left(\frac{55.3895}{100}\right)\cong 5.908\% \tag{15.3.433}$$

Remark

Investors should pay particular attention to the indication "contingent" on its bond paper, which means that in case of bankruptcy of the debtor (likened to the "issuer risk"), the holder of the obligation will be repaid only after all other creditors... The issuer risk can be avoided by choosing (very) safe bonds as government bonds or renowned companies. The downside is then the low rates offered that we must furthermore put in contrast with inflation (on a rate of 3% over 10 years of a government bond that undergoes an inflation of 2% there remains only 1% of remuneration for example).

Let us indicate that multiple coupons bonds are trendy thanks to the tax system of the States and also by the cost of paperwork they generate. Hence the fact that the zero-coupon are preferred.

✍**Examples:**

E1. Consider a bond of $3,000,000$.- divided into 300 bonds each of nominal $10,000$.- issued in June 12004 (holocene calendar) for a period of 10 years. Subscription: 99.5% of the par value. Redemption at par at maturity. Fixed annual interest: 4.5%.

The values defined above are then expressed as follows:

The nominal value C of the obligation is to $10,000$.-. The number N of bonds is 300. The duration n of the loan is 10 years and the nominal rate is 4.5% per year. The issue price is 99.5% of $10,000$.- that is to say $E=9,950$.- (below par!). The coupon therefore has a value $c=450$.- and the repayment R is made at par and so is equal to $10,000-4,500=5,500$.-.

E2. Given fixed-rate bond, issued at a price of $1,000$.-, and paying an annual coupon of 100.-. The rate used is $100/1,000=10\%$.

Suppose the market rates go to 15%. This means that a new obligation, which is issued for a price of $1,000$.-, serves a coupon of 150.- (because $150/1,000=15\%$).

The new bond is more interesting than the old one, and everyone will want to sell the old to buy new. That is why the price of the old bond will implicitly decrease, until it matches that of a financial product providing 15% or here 666.-. Then, we will have as expected $100/666=15\%$.

Similarly, if market rates fall to 5%, this means that a new obligation, which is issued to a price of $1,000$.- price, will serve a coupon of 50.- (because $50/1,000=5\%$).

The new bond is the less interesting than the old one, and almost no one will buy it. That is why the price of the old obligation will implicitly go up until it match that of a financial product providing 5% rate that is to say here $2,000$.-. So, we will have as expected $100/2,000=5\%$.

Thus, the price of a fixed rate bond decreases when market interest rates rise and rises when market interest rates fall. This is why an investment in bonds is not without risks: one can lose part of the capital. In fact, the only (almost) safe strategy is to buy bonds at the time of issuance, and keep them until maturity.

At any time, the market value of a fixed rate bond or variable rate one must therefore be equal to the sum of the present values of the coupons and of the repayment which it gives the right. The present value is calculated at the bond market rate for bonds of the same type and same duration.

Thus, the present value of a fixed-rate bond (the case being easily generalized to a variable rate) should be seen as an initial capital of which is removed during n remaining periods a certain fixed sum corresponding to the amount of the price of the coupon (facial interest rate multiplied by the nominal):

$$c=t\%C \tag{15.3.434}$$

with C the nominal value of the obligation and total cumulated being periodically subject to market interest rate $t_M\%$ (thus the "rate to maturity") that is considered as constant in the context of a certain (deterministic) future.

Thus, the present value of a fixed-rate bond is initially formed as the present value of the remaining future coupons (often also named "cash flows") during n periods such that:

$$P_c=c\cdot a_{\overline{n}|}=\sum_{i=1}^{n}\frac{c}{(1+t_M\%)^i}=c\sum_{i=1}^{n}\frac{1}{(1+t_M\%)^i}=t\%C\sum_{i=1}^{n}\frac{1}{(1+t_M\%)^i} \tag{15.3.435}$$

relation that we found many times in the specialized literature in the following form:

$$P=cF\sum_{i=1}^{n}\frac{1}{(1+y)^i} \tag{15.3.436}$$

This part of the price of the value of the bond corresponds to the total amount required as we can settle the amount P_c after removing n times (the number of periods remaining) the value c at an interest rate $t_M\%$.

Remark

Let us recall that we have proved in the Section Sequences and Series (see page 1656), that if $t_M\%$ is less than 1 and the sum tends to infinity, then :

$$P_c=\frac{c}{t_M\%}=\frac{t\%C}{t_M\%} \tag{15.3.437}$$

Then the bond is also made of the value of the reimbursement R. Although it is repaid at maturity, it can be seen as a savings capital at a rate equal to that of the market $t_M\%$ as:

$$R=C_0(1+t_M\%)^n=C_0f_c^n\Leftrightarrow C_0=Rf_e^n \tag{15.3.438}$$

The present value of the bond concerning the reimbursement is then:

$$P_R=Rf_e^n \tag{15.3.439}$$

which is the capital for repayment R remaining after the n remaining periods.

Thus, the total price of a fixed-rate bond, also named "non-arbitrage bond price" is:

$$\boxed{P=P_c+P_R=ca_{\overline{n}|}+Rf_e^n=c\sum_{i=1}^{n}\frac{1}{(1+t_M\%)^i}+\frac{R}{(1+t_M\%)^n}} \tag{15.3.440}$$

that is to say, the present value of future coupons and also the present value of the in fine reimbursement. That latter relation is also denoted (CF stands as always for the abbreviation of "cash flow"):

$$P=\sum_{t=1}^{T}\frac{\text{CF}_t}{(1+y)^t} \tag{15.3.441}$$

This relation is important in deterministic quantitative finance, it is necessary to remember it!! In the context above, be aware that the rate equation is often named "yield to maturity at par".

Remark

We can also easily write now the general equation for the price of a bond (aka plain vanilla bonds), often denoted:

$$P=\sum_{t=1}^{T}\frac{C_t}{(1+\text{YTM})^t}+\frac{\text{Par}_T}{(1+\text{YTM})^T} \tag{15.3.442}$$

A bond that has all coupon payments that are constant is named a "straight bond".

The value of a bond, in the point of view of its trading price, may differ from its nominal value set at the issue if interest rates change in the market hence the purpose to calculate its current value.

✍**Example:**

Consider we want to calculate the current price of a bond, with annual coupon of 450.-, with a reimbursement at par in 5 years for $10,000$.-.

The present value P for a market rate between 0% and 100% has the following characteristics:

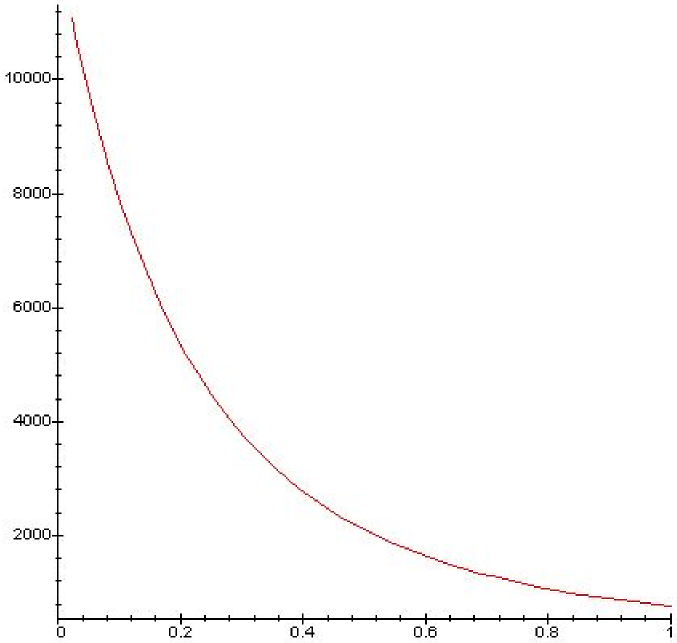


Figure 15.109: **Current value of a bond based on the market rates (yield)**

Thus, a fundamental property of the price of a fixed rate bond is that this is a strictly decreasing function of market rate of return. The finance practitioners (and mathematicians) say than the function is convex: that is to say, when market rates fall, the price accelerates upwards and vice versa when the market interest rates rise, the price is decelerating downwards.

We also guess that at the view of the relation of the non-arbitrage bond price P, that bond prices variations increase with the increase in maturity and with the value of coupons.

✍**Example:**

Today, we buy a fixed rate bond with of 3 years maturity for a nominal price 100.-, the coupon rates are of 5% and the fixed market rate 10%. The cash flows collected are then of 5.-, 5.- and 105.- after respectively 1 year, 2 years and 3 years. The price of this bond is then equal to:

$$P_c = 5\sum_{i=1}^{2}\frac{1}{(1+10\%)^i}+\frac{105}{(1+10\%)^3}\cong 87.57.\text{-} \tag{15.3.443}$$

That can get directly with Microsoft® Excel® 14.0.6129:

```
=PRICE("2013-01-01","2017-01-01",5%,10%,105,1,3)=87.577
```

Evaluate a bond is therefore equivalent to find what should its value in principle in the actual conditions of market, that is to say its potential trading value, by a mathematical operation named "operation of actualization" determining its theoretical present value. It is therefore, as we already know, an actuarial calculation.

The bond investor will of course have for aim to seek the market rate that allows to make the investment a profitable action. Thus, we define the "yield to maturity" x as the interest of the market that can satisfy the following relations, depending on the remaining time periods n of the bond among N periods.

Thus, when the bond is issued:

$$E = ca_{\overline{N}|} + Rf_e^N = \frac{c}{1+x} + \frac{c}{(1+x)^2} + \ldots + \frac{c}{(1+x)^N} + \frac{R}{(1+x)^N} \tag{15.3.444}$$

or on any date $n < N$:

$$P = ca_{\overline{n}|} + Rf_e^n = \frac{c}{1+x} + \frac{c}{(1+x)^2} + \ldots + \frac{c}{(1+x)^n} + \frac{R}{(1+x)^n} \tag{15.3.445}$$

The yield to maturity of a bond is then the rate x that cancels the difference between the value of the issue price E and the present value of future cash flows it generates. This rate is calculated at the date of settlement and must appear in the emission brochures. For the buyer of the obligation, the actuarial rate represents the rate of return achieved if keeping the bond until its repayment and reinvesting the interests at the same actuarial interest rate.

Remark

We have seen above that for a straight bond:

$$P = \sum_{t=1}^{T} \frac{C}{(1+\text{YTM})^t} + \frac{\text{Par}_T}{(1+\text{YTM}^T)}$$

Using the relation proved at page 7564:

$$a_{\overline{n}|} = \sum_{t=1}^{n} f_e^t = f_e^1 \frac{1-f_e^n}{1-f_e} = f_e \frac{1-f_e^n}{1-\dfrac{1}{f_c}} = \frac{1}{f_c} \frac{1-f_e^n}{\dfrac{f_c-1}{f_c}} = \frac{1-f_e^n}{f_c-1} = \frac{1-f_e^n}{t\%}$$

The latter can be rewritten:

$$P = \frac{C}{\text{YTM}} \left(1 - \frac{1}{(1+\text{YTM})^T}\right) + \frac{\text{Par}_T}{(1+\text{YTM})^T}$$

And as for a perpetual bond $T \rightarrow +\infty$, then we get the pricing formula for a perpetual bond:

$$\boxed{P_{\text{perp.}} = \frac{C}{\text{YTM}}}$$

Let us see some other useful definitions of bonds:

Definitions (#15.3.686):

D1. The "expired coupon" (EC), also named "net coupon", of a bond is paid to the owner after deduction of the anticipated taxes AT% (AT being the abbreviation for: Anticipated Taxes). Thus the calculation of the annual net coupon of bonds traded at P_c (monetary value) with a yield of $t_0\%$ is trivially given by:

$$\text{EC} = P_c t_0\%(1 - \text{AI}\%) \tag{15.3.446}$$

D2. The "accrued interest" (AI) is the amount of interest (fraction) that has been accumulated since the last date of payment of interest, but which is not yet due. It is won by a bond since its last term and is determined on a sale or inventory. Or in other words it is method when interest that is either payable or receivable has been recognized, but not yet paid or received. Accrued interest occurs as a result of the difference in timing of cash flows and the measurement of these cash flows.

Its calculation is trivially given by an application of rules of simple interest such as:

$$\text{AI} = P t_0\% \frac{\Delta d}{\Delta T} \tag{15.3.447}$$

where Δd is the time since last coupon payment and ΔT the period of coupons (both must be obviously in the same units).

So to get the effective value of a bond (the "present value" as we know), we add to its traded value, named "clean price", the accrued interest (in green in the figure below) since the last coupon payment.

Change in Present Value through Coupon Date

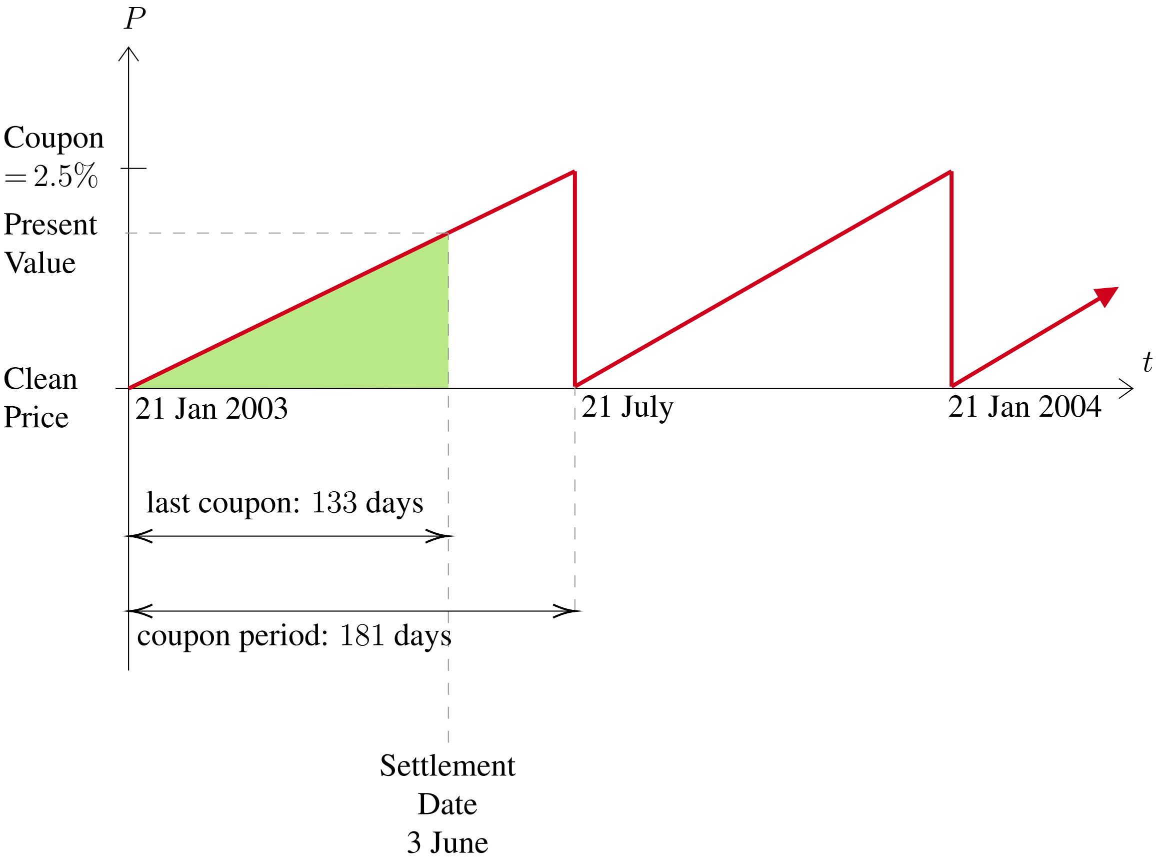


Figure 15.110: Schematic principle behind the bond accrued interest

D3. By extension, if we seek to calculate the net value of n (parallel) coupons with a same yield of $t_0\%$ of a bond which nominal value is P with an anticipated tax of AI%, we then can calculate the "net annual coupon at maturity" (NEC) by the following trivial relation:

$$\text{NEC} = nPt_0\%(1-\text{AI}\%) \tag{15.3.448}$$

D4. Suppose now we have two bonds with different maturities (say t and T, with $T > t$) and "spot interest rates" (actual rate) $r_{(0,t)}$ and $r_{(0,T)}$. The additional rate between t and the maturity $r_{(t,T)}$ is named the "forward interest rate" and is traditionally denoted as $F_{(0,t,T-t)}$.

Obviously the latter can be written:

$$(1+r_{(0,T)})^T = (1+r_{(0,t)})^t(1+F_{(0,t,T-t)})^{T-t} \tag{15.3.449}$$

and rearranged into:

$$\boxed{F_{(0,t,T-t)} = \left(\frac{(1+r_{(0,T)})^T}{(1+r_{(0,t)})^t}\right)^{1/(T-t)} - 1} \tag{15.3.450}$$

The concepts of "spot rate" and "forward rate" are now known to us, but we must be aware that the vocabulary and notation in the field of modern portfolio theory is a bit different and it is important to learn why these two rates are so important for the pricing of financial tools like Futures and Swaps!

As companion example, let us consider the final capital C_n and a risk-free market yield of $t\%$ for a given periodic basis. We then know that the present value is:

$$C_0 = \frac{C_n}{(1+t\%)^n} \tag{15.3.451}$$

Now if the rate varies from period to period, assuming these rates are evenly spaced, we have:

$$C_0 = \frac{C_n}{(1+t_1\%)(1+t_2\%)\ldots(1+t_n\%)} \tag{15.3.452}$$

Traditionally in the financial markets (but not always in textbooks!), the rate that is closest to today's date is named the "spot rate" and other future rates are named "forward rates" such as:

$$C_0 = \frac{C_n}{(1+s_1\%)(1+f_2\%)\ldots(1+f_n\%)} \tag{15.3.453}$$

We can then finally give the usual definition: A "forward rate" is a rate that discounts a future amount to another point in the future while the "spot rate" is a rate that discounts a future amount to the daily value.

However, in the financial markets, the spot rate and forward rates are not necessarily paid at regular intervals (yes ... it would be too easy otherwise...!).

Consider for example an investment over 5 periods with a rate market offering a spot rate for the first 4 periods and a forward rate during the last period starting from the 4th. We have then according to usual notations:

$$C_0 = \frac{C_5}{(1+s_4\%)^4(1+{}_1f_4)} \tag{15.3.454}$$

Or we can have three different rates offered on the market such as:

$$C_0=\frac{C_5}{(1+s_1\%)(1+{}_2f_1)^2(1+{}_2f_3)^2} \tag{15.3.455}$$

which means that we will first have a first spot rate s_1 during the first period (from 0 to 1), then we will have a particular forward rate for 2 periods (1 to 2 and 2 to 3) after this first period of ${}_2f_1$ and that finally we will have another particular forward rate for 2 periods (3 to 4 and 4 to 5) after the period 3. We understand via this last example much better this usual notation.

Now what is very important to understand is that in the markets, in the absence of arbitration opportunity, we should have for equal instruments the following equality if we take again the two examples above :

$$(1+s_1)(1+{}_2f_1)^2(1+{}_2f_3)^2=(1+s_4)^4(1+{}_1f_4) \tag{15.3.456}$$

The financial analysts are then interested in using all the market rates (using regression or interpolation techniques as seen in the section on Numerical Methods or other empirical techniques like the Nelson-Siegel model or other...) to construct a curve of spot rates (the "spot curve") for each future period and obviously this rate curve also corresponds to the rate that would pay a zero-coupon bond hence the reason for which the two terms "curve of spot rates" and "curve of zero-coupon rates "are equivalent!

In the case above, we will thus have for the first period the two points:

$$\{(1,1+s_1),(1,1+s_4)\} \tag{15.3.457}$$

for the second period:

$$\left\{(2,(1+s_1)(1+{}_2f_1))\,,\left(2,(1+s_4)^2\right)\right\} \tag{15.3.458}$$

For the third period:

$$\left\{\left(3,(1+s_1)(1+{}_2f_1)^2\right),\left(3,(1+s_4)^3\right)\right\} \tag{15.3.459}$$

For the fourth period:

$$\left\{\left(4,(1+s_1)(1+{}_2f_1)^2(1+{}_2f_3)\right),\left(4,(1+s_4)^4\right)\right\} \tag{15.3.460}$$

and finally for the fifth:

$$\left\{\left(5,(1+s_1)(1+{}_2f_1)^2(1+{}_2f_3)^2\right),\left(5,(1+s_4)^4(1+{}_1f_4)\right)\right\} \tag{15.3.461}$$

It is therefore easy to understand why this curve is increasing in a risk-free universe since the terms of ordinates that multiply each other are greater than 1. Indeed, in the latter case (without risk), we simply have an exponential curve of the type of traditional compound interest. On the other hand, in a universe with risk, which corresponds to the reality of the market, from a certain time horizon the risk premiums are such that the yield curve has a horizontal asymptote because the risk cancels the gain. And at a very distant horizon, the yield curve even decreases (the risk being higher than the interest rates!).

Thus forward rates can be calculated using spot rates of the spot rate curve and vice versa. This is why financial analysts sometimes talk about "spot rate curve" or "forward rate curve" and for the latter we sometimes talk about "implicit forward rates". For example in a simple case with a continuous rate we must have:

$$e^{s_2T_2}=e^{s_1T_1}e^{{}_1f_1(T_2-T_1)} \tag{15.3.462}$$

Therefore, the forward rate in 1 month is trivially given by taking the logarithm of both sides of the equality that are well known to us and that lead to a famous relation available in quite a significant percentage of textbooks (but without proof or explanations!):

$$\boxed{{}_1f_1=\frac{s_2T_2-s_1T_1}{T_2-T_1}} \tag{15.3.463}$$

Remark

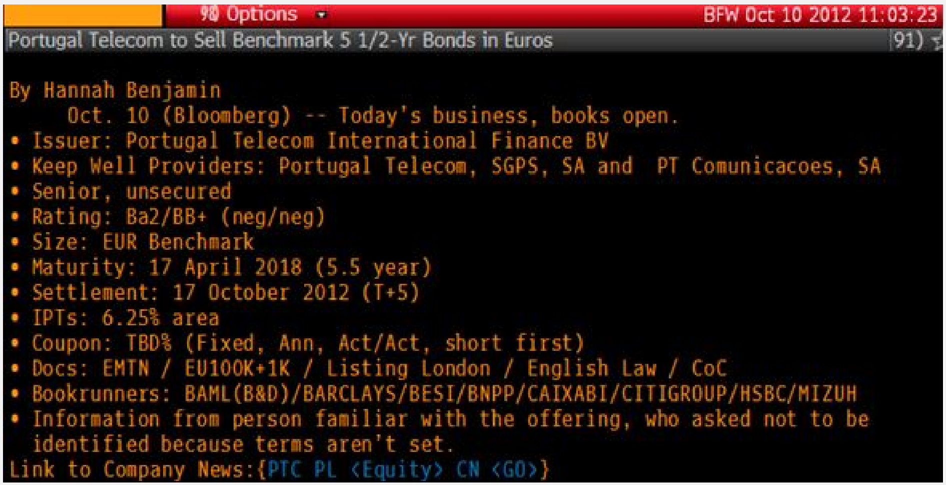
National banks generally communicate the mathematical model used to communicate the interest rate yield curve (the last well-implemented model is named the "Svensson model").

At the opposite of to the calculation of accrued interest of the coupons of a bond, the calculation of the accrued dividend of a stock is veeeery difficult to estimate. The stock price, however, is influenced by the more or less near the date of the dividend payment.

Remark

Let us indicate that the market in which the issuers selling their bonds (new) by auction, syndication or by direct placement to investors is named "primary market". The market in which investors exchange among themselves bonds (sometimes the same obligations which are offered at different prices) already in circulation (second hand market) is named "secondary market".

Below an example of announcement of bond issuance by syndication on Bloomberg™ through a news:



```
Portugal Telecom to Sell Benchmark 5 1/2-Yr Bonds in Euros

By Hannah Benjamin
     Oct. 10 (Bloomberg) -- Today's business, books open.
• Issuer: Portugal Telecom International Finance BV
• Keep Well Providers: Portugal Telecom, SGPS, SA and  PT Comunicacoes, SA
• Senior, unsecured
• Rating: Ba2/BB+ (neg/neg)
• Size: EUR Benchmark
• Maturity: 17 April 2018 (5.5 year)
• Settlement: 17 October 2012 (T+5)
• IPTs: 6.25% area
• Coupon: TBD% (Fixed, Ann, Act/Act, short first)
• Docs: EMTN / EU100K+1K / Listing London / English Law / CoC
• Bookrunners: BAML(B&D)/BARCLAYS/BESI/BNPP/CAIXABI/CITIGROUP/HSBC/MIZUH
• Information from person familiar with the offering, who asked not to be
  identified because terms aren't set.
Link to Company News:{PTC PL <Equity> CN <GO>}
```

Figure 15.111: **Current value of a bond based on the market rates (yield)**

On the secondary market bonds are quoted as a percentage of their face value (i.e. when they are traded on 100% of the nominal we, as we already know, that they are "au pair") and the quoted price market (traded price) is sometimes named as we also already know the "clean price", to get the overall price ("dirty price") of the transaction (purchase or sale) we must added the accrued interest of the coupon and other taxes and trading costs.

15.3.10.3.2.1 MacCaulay Duration, Modified Duration and Convexity

The "modified duration" is a measure (rate of change) of the price sensitivity of an asset in relation to the change in one unit of the market rate (unfortunately, contrary to what its name might suggest it does not give a time duration contrary to the Macaulay duration that we will see further below). Obviously, the higher the duration of a bond, the more sensitive it is to yield changes. Conversely, the lower the duration of a bond, the less sensitive it is to yield changes. In general, we speak of the calculation of "interest rate risk".

The modify duration can also be reported in dollars when multiplied by the spot price of the concerned asset, and then we refer to the term "dollar duration" (D.D.), or "Bloomberg risk", and we denote it DV01 (we will see further below that DV01 corresponds in the fields of derivatives to the "delta").

The concept of modified duration can of course be applied to any instrument sensitive to interest rates with fixed or non-fixed cash flows.

There are several ways of introducing this indicator - which is available in several variants - which mainly allows measuring the sensitivity in percent of a bond's clean price to changes in the risk-free return of the market or as an empirical indicator of any obligation with its zero-coupon theoretical equivalent (more generally bonds paying coupons at a different pace but with identical rates).

A first approach is to go through an elegant mathematical method that uses Taylor's serial developments and which shows that the "Duration" can also be used to measure the convexity of bond prices (or, if you prefer, their sensitivity to changes in interest rates).

Let us first recall that:

$$P=\sum_{t=1}^{T}\frac{\mathrm{CF}_t}{(1+y)^t} \qquad (15.3.464)$$

Remark

Obviously in the case of a zero-coupon bond (i.e. a single payment for recall) we can move approximately to a continuous rate such that $P=\mathrm{CF}e^{-yT}$, which simplifies the calculations considerably but is then only a special case.

In the case of a Taylor series development (see section Sequences and Series page 1696) we have for any continuous function in the discrete case:

$$\Delta P=\frac{\mathrm{d}P}{\mathrm{d}y}\Delta y+\frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}y^2}(\Delta y)^2+\ldots \qquad (15.3.465)$$

If the variation is really small compared to the reference rate y, we will use the first term (linear approximation), if the variation is considered to be relatively large we will take both (quadratic approximation) otherwise the approximation is bad as soon as we move away from the y value:

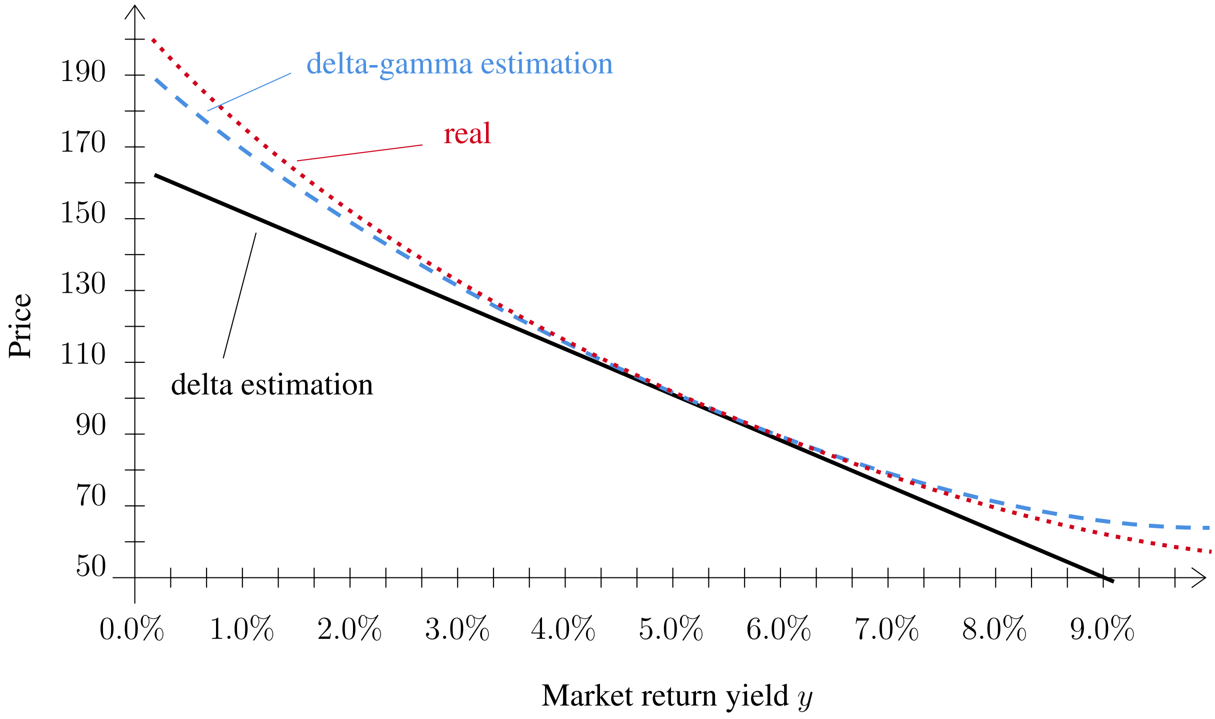


Figure 15.112: **Delta and Delta-Gamma Modified duration approach**

The relative variation of the price in percentages is then given by:

$$\boxed{\frac{\Delta P}{P}\cong\frac{1}{P}\frac{\mathrm{d}P}{\mathrm{d}y}\Delta y+\frac{1}{2}\frac{1}{P}\frac{\mathrm{d}^2P}{\mathrm{d}y^2}(\Delta y)^2}\qquad(15.3.466)$$

Remark

Because in the case of the derivatives that we will discuss below, some of the above factors reveal what we name the "delta" and "gamma" if we change the rate y above by the value of the underlying S of the derivative, we name this method (even when not applied to derivatives): "delta-gamma method".

In finance we write this relation in several forms:

$$\begin{aligned}\frac{\Delta P}{P}&=-\frac{1}{P}\text{DD}\Delta y+\frac{1}{2}\text{Convexity}(\Delta y)^2\\&=-D^*\Delta y+\frac{1}{2}\text{Convexity}(\Delta y)^2\\&=-\frac{1}{P}\text{DVBP}+\frac{1}{2}\text{Convexity}(\Delta y)^2\\&=-\frac{D}{1+y}\Delta y+\frac{1}{2}\text{CVX}(\Delta y)^2\\&=-\frac{D}{1+y}\Delta y+\frac{1}{2}\frac{1}{P}\text{DC}(\Delta y)^2\end{aligned}\qquad(15.3.467)$$

where DD is the "dollar duration" (already mentioned just before), D^* the "modified duration", DVBP the "dollar value by basis point", D is the "compound duration" named more often the "Macaulay duration" (also sometimes denoted MacD) and finally DC the "convexity dollar" (for CVX it's trivial!).

In short, to sum up we have for the first term where there are the more variations of writing in practice (because the units are not the same!):

$$\boxed{\begin{aligned}\text{DD}=\text{Dollar Duration}&=-\frac{\mathrm{d}P}{\mathrm{d}y}\\ D^*=\text{Modified Duration}&=-\frac{-\frac{\mathrm{d}P}{P}}{\mathrm{d}y}=-\frac{1}{P}\frac{\mathrm{d}P}{\mathrm{d}y}\\ \text{DVBP}=\text{Dollar value by basis point}&=-\frac{\mathrm{d}P}{\mathrm{d}y}\Delta y\\ D=\text{Macaulay Duration}&=-\frac{1}{P}\frac{\mathrm{d}P}{\mathrm{d}y}(1+y)=D^*(1+y)\end{aligned}} \tag{15.3.468}$$

Let us open a parenthesis with the expression:

$$\frac{\mathrm{d}P}{P}=-D^*\mathrm{d}y \tag{15.3.469}$$

In practice, assets are often given as yield volatility rather than price volatility. The previous relation will allow us to get one from the other. For this we will demonstrate further that under the hypothesis of a log-normal distribution of the rates we have:

$$\mathrm{d}X_t=\mu X_t\mathrm{d}t+\sigma X_t\mathrm{d}W_t \tag{15.3.470}$$

And under the assumption that what we name the "drift" (μ) is zero (which makes sense since the rate of assets like bonds or securities is fixed in advance!!!), then we have with our notations:

$$\frac{\mathrm{d}y}{y}=\sigma\mathrm{d}W\Leftrightarrow\mathrm{d}y=y\sigma\mathrm{d}W \tag{15.3.471}$$

Then we have:

$$\frac{\mathrm{d}P}{P}=-D^*\mathrm{d}y=-D^*y\sigma\mathrm{d}W \tag{15.3.472}$$

Then by analogy, we define the volatility of assets rates by:

$$\sigma_P=|-D^*y\sigma| \tag{15.3.473}$$

Or to avoid any notation confusion:

$$\sigma_P=|-D^*y\sigma_y| \tag{15.3.474}$$

For what follows, let us take one of these 4 indicators randomly (in fact it is not really random as it is the only one that really gives a duration in the "temporal" sense of the term):

$$D=-\frac{1}{P}\frac{\mathrm{d}P}{\mathrm{d}y}(1+y)=\sum_{t=1}^{T}t\frac{\text{CF}_t/(1+y)^t}{P}=\frac{1}{P}\sum_{t=1}^{T}t\frac{\text{CF}_t}{(1+y)^t}=\frac{1}{\displaystyle\sum_{t=1}^{T}\frac{\text{CF}_t}{(1+y)^t}}\sum_{t=1}^{T}t\frac{\text{CF}_t}{(1+y)^t} \tag{15.3.475}$$

Remark

In many textbooks the previous relation is denoted:

$$\text{MacD}=\frac{1}{\text{PV}}\sum_{t=1}^{T}t\frac{\text{CF}_t}{(1+\text{YTM})^t} \tag{15.3.476}$$

where for recall PV is the abbreviation of Present Value and YTM of Year To Maturity.

For a zero-coupon bond we then have the logical result (reason why this time indicator has a particular name!):

$$D=\frac{1}{P}t\frac{\text{CF}_1}{(1+y)}=\frac{1}{P}tP=t \tag{15.3.477}$$

We can give a simplified generalized economic interpretation of the Macauley's duration indicator by manipulating a little the relation obtained above:

$$D=-\frac{1}{P}\frac{\mathrm{d}P}{\mathrm{d}y}(1+y)=\sum_{t=1}^{T}\frac{\text{CF}_t/(1+y)^t}{P}=\sum_{t=1}^{T}\frac{\text{CF}_t/(1+y)^t}{\displaystyle\sum_{t=1}^{T}\frac{\text{CF}_t}{(1+y)^t}}=\sum_{t=1}^{T}tw_t \tag{15.3.478}$$

where the weights w_t represent the ratio of the discounted value of each cash flow (CF) relative to the total cash flow (thus normalization to the unit). It is therefore a weighted average! So the Macaulay's duration represents the weighted average of the time to wait to get all the cash flows.

The figure below lays out the present value of the cash flows of a 6% coupon, 10-year bond. Given a duration of 7.80 years, this coupon-paying bond is equivalent to a zero-coupon bond maturing in exactly 7.80 years:

For bonds with fixed coupons, duration is less than maturity. With a zero coupon, Macaulay duration is equal to maturity. Higher coupons place more weight on prior payments and therefore reduce duration.

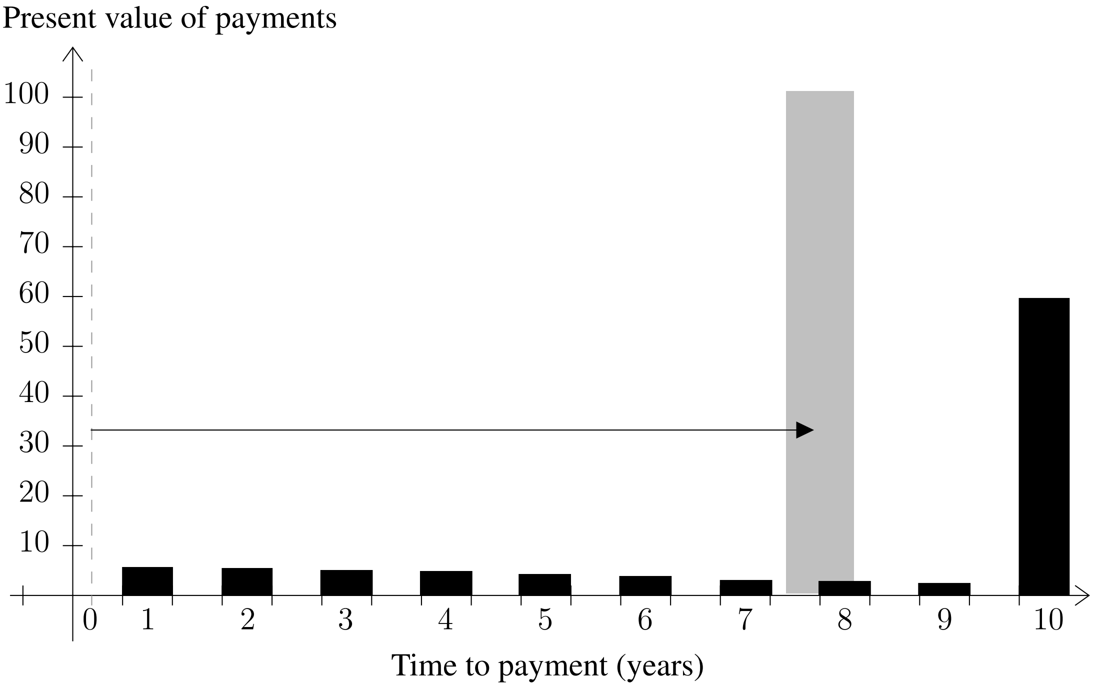


Figure 15.113: **Duration as the Maturity of a Zero-Coupon Bond**

Obviously, the longer the Macaulay's duration, the more the bond is sensitive to changes in interest rates (the longer the time at maturity, the longer the duration).

Remark

The modified duration is not additive (hence we have not been able to prove the opposite), it is then quite natural to define the modified duration of a portfolio as the weighted sum of durations of the assets that compose it.

We can also introduce the convexity duration. The calculation is defined as following:

$$\begin{aligned}\mathrm{CVX}&=\frac{1}{P}\frac{\mathrm{d}^2P}{\mathrm{d}y^2}=\frac{1}{P}\sum_{t=1}^{T}\frac{t\mathrm{CF}_t}{(1+y)^{t+1}}=\frac{1}{P}\sum_{t=1}^{T}\frac{\mathrm{CF}_t t(t+1)}{(1+y)^{t+2}}\\&=\frac{1}{\displaystyle\sum_{t=1}^{T}\frac{\mathrm{CF}_t}{(1+y)^t}}\sum_{t=1}^{T}\frac{\mathrm{CF}_t t(t+1)}{(1+y)^{t+2}}=\frac{1}{\displaystyle\sum_{t=1}^{T}\frac{\mathrm{CF}_t}{(1+y)^t}}\sum_{t=1}^{T}\frac{t(t+1)}{(1+y)^2}\frac{\mathrm{CF}_t}{(1+y)^t}\\&=\sum_{t=1}^{T}\frac{t(t+1)}{(1+y)^2}\frac{\mathrm{CF}_t/(1+y)^t}{\displaystyle\sum_{t=1}^{T}\frac{\mathrm{CF}_t}{(1+y)^t}}=\sum_{t=1}^{T}\frac{t(t+1)}{(1+y)^2}w_t\end{aligned}\tag{15.3.479}$$

✍**Examples:**

E1. Let us recall that for a zero-coupon bond we have proved (see page 7624) the following present value:

$$P = Ce^{-nt_n\%} \tag{15.3.480}$$

Adapting to the notation above we rewrite this relation as:

$$P = Ce^{-Ty} \tag{15.3.481}$$

Then:

$$D^* = -\frac{1}{P}\frac{\mathrm{d}P}{\mathrm{d}y} = -\frac{1}{Ce^{-Ty}}\frac{\mathrm{d}\left(Ce^{-Ty}\right)}{\mathrm{d}y} = -\frac{1}{Ce^{-Ty}}(-T)Ce^{-Ty} = T \tag{15.3.482}$$

The duration of a zero-coupon bound equals then the negative of its years to maturity. As:

$$\frac{\Delta P}{P} \cong -D^*\delta y \tag{15.3.483}$$

then an increase in the bond's yield will cause its value to fall approximately.

The convexity of the zero-coupon bond is therefore equal to:

$$\text{CVX} = \frac{1}{P}\frac{\mathrm{d}^2P}{\mathrm{d}y^2} = \frac{T^2Ce^{-Ty}}{Ce^{-yT}} = T^2 \tag{15.3.484}$$

E2. Consider a 2-year bond with face value of $\$100$, a 20% semi-annual coupon, and a yield of 4% semi-annually compounded. The total PV will be:

$$\begin{aligned} P &= \sum_{i=1}^{n}\text{PV}_i = \sum_{i=1}^{n}\frac{\text{CF}_i}{(1+y/k)^{k\cdot t_i}} = \sum_{t=1}^{4}\frac{10}{(1+0.04/2)^t} + \frac{100}{(1+0.04/2)^4} \\ &= 9.804 + 9.612 + 9.423 + 9.238 + 92.385 = 130.462 \end{aligned} \tag{15.3.485}$$

The Macaulay duration is then:

$$\begin{aligned} D &= \frac{1}{P}\sum_{t=1}^{T} t\frac{\text{CF}_t}{(1+y/k)^t} \\ &= 0.5\cdot\frac{9.804}{130.462} + 1.0\cdot\frac{9.612}{130.462} + 1.5\cdot\frac{9.423}{130.462} + 2.0\cdot\frac{9.238}{130.462} + 2.0\cdot\frac{92.385}{130.462} \\ &= 1.777\ \text{[years]} \end{aligned} \tag{15.3.486}$$

The modified duration, measured as percentage change in price per one percentage point change in yield, is:

$$D^* = \frac{D}{(1+y/k)} = \frac{1.777}{(1+0.04/2)} = 1.742\% \tag{15.3.487}$$

The DV01, measured as dollar change in price for a $\$100$ nominal bond for a one percentage point change in yield, is:

$$\text{DV01} = \frac{D^*\cdot P}{100} = \frac{1.742\%\cdot 130.462}{100} = 2.27 \tag{15.3.488}$$

where the division by 100 is because modified duration is the percentage change.

And finally the convexity is given by:

$$\begin{aligned}\text{CV}&=\sum_{t=1}^{T}\frac{t(t+1)}{(1+y)^2}\frac{\text{CF}_t/(1+y)^t}{\displaystyle\sum_{t=1}^{T}\frac{\text{CF}_t}{(1+y)^t}}=\sum_{t=1}^{4}\frac{t(t+1)}{(1+0.04/2)^{2\cdot 2}}\frac{20/(1+0.04/2)^{2t}}{130.462}\\&\cong 0.283+0.785+1.450+2.233=4.751\end{aligned}\tag{15.3.489}$$

The duration of a portfolio is the weighted average duration of the bonds in the portfolio. The weights are calculated by the market value (dirty price) of each bond divided by the total market of the portfolio.

Then for a portfolio of n assets we have:

$$D_p=\sum_{i=1}^{n}w_iD_i\tag{15.3.490}$$

with obviously $\sum_{i=1}^{n}w_i=1$. The same way we have the definitions:

$$\begin{aligned}D_p^*&=\sum_{i=1}^{n}w_iD_i^*\\\text{CVX}_p&=\sum_{i=1}^{n}w_i\text{CVX}_i\end{aligned}\tag{15.3.491}$$

But let us explicit D_p^* assuming for a portfolio of N assets, each asset represent a portion π_j (in currency amount!) of the portfolio:

$$\begin{aligned}D_p^*&=\sum_{i=1}^{n}w_iD_i^*=\sum_{i=1}^{n}w_i\frac{D_i}{(1+y_i/k_i)}=\sum_{i=1}^{n}w_i\frac{\displaystyle\frac{1}{P_i}\sum_{t=1}^{T_i}t\frac{\text{CF}_t}{(1+y_i/k_i)^t}}{(1+y_i/k_i)}\\&=\sum_{i=1}^{n}\left[\left(\frac{\pi_iP_i}{\displaystyle\sum_{j=1}^{N}\pi_jP_j}\right)\frac{\displaystyle\frac{1}{P_i}\sum_{t=1}^{T_i}t\frac{\text{CF}_t}{(1+y_i/k_i)^t}}{(1+y_i/k_i)}\right]\\&=\sum_{i=1}^{n}\left[\left(\frac{\displaystyle\pi_i\sum_{t=1}^{T_i}\frac{\text{CF}_t}{(1+y_i/k_i)^t}}{\displaystyle\sum_{j=1}^{N}\left(\pi_j\sum_{t=1}^{T_j}\frac{\text{CF}_t}{(1+y_j/k_j)^t}\right)}\right)\frac{\displaystyle\frac{1}{\displaystyle\sum_{t=1}^{T_i}\frac{\text{CF}_t}{(1+y_i/k_i)^t}}\sum_{t=1}^{T_i}t\frac{\text{CF}_t}{(1+y_i/k_i)^t}}{(1+y_i/k_i)}\right]\end{aligned}\tag{15.3.492}$$

A recurring wish of bond portfolio managers is to modify the allocation of a portfolio so that it has a given modified duration and a given convexity. We then used the definition of modified duration and the convexity of a portfolio given for recall by:

$$\begin{aligned} D_p^* &= \sum_{i=1}^{n} w_i D_i^* \\ \text{CVX}_p &= \sum_{i=1}^{n} w_i \text{CVX}_i \sum_{i=1}^{n} w_i \quad = 1 \end{aligned} \tag{15.3.493}$$

Put in the form of a linear system that gives:

$$\begin{bmatrix} D_1^* & \dots & D_n^* \\ \text{CVX}_1 & \dots & \text{CVX}_n \\ 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix} = \begin{bmatrix} D_p^* \\ \text{CVX}_p \\ 1 \end{bmatrix} \tag{15.3.494}$$

So the system is extremely easy to solve to find the corresponding weights:

$$\begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix} = \begin{bmatrix} D_1^* & \dots & D_n^* \\ \text{CVX}_1 & \dots & \text{CVX}_n \\ 1 & \dots & 1 \end{bmatrix}^{-1} \begin{bmatrix} D_p^* \\ \text{CVX}_p \\ 1 \end{bmatrix} \tag{15.3.495}$$

15.3.10.3.2.2 Equivalence of a Bond rate for a Treasury bill

Some instruments are traditionally traded not by interest rate (yield) but by discount rate. In the context of financial markets this means that it is traded on the interest rate earned on the final value of the instrument (therefore the gain is calculated in $\%$ of the final value) and not on the initial price of the instrument.

Thus, consider a treasury bill bought 900.- which generated 180.- of profit (cash flow). Consequently its corresponding simple interest is trivially 20% of the initial value according to the trivial relation and the notations of use:

$$I = \frac{\text{CF}}{\text{TB}} \tag{15.3.496}$$

On the other hand as the final value is 1080.- $(900+180)$ then 180.- represents 16.66% of the final value according to the trivial relation:

$$t\% = \frac{\text{CF}}{\text{TB}+\text{CF}} \tag{15.3.497}$$

In general, the simple interest rate equivalent to the discount rate is then obviously higher than the latter.

We therefore deduce the relation between simple interest rate and discount rate:

$$I = \frac{\text{CF}}{\text{TB}} = \frac{\dfrac{t\%\text{TB}}{1-t\%}}{\text{TB}} = \frac{t\%}{1-t\%} \tag{15.3.498}$$

So here we fall back on a simplified version of the commercial discount relationship that we demonstrated during our study of simple and compound interest earlier above (see page 7555).

The discount rate for a treasury bill that has a maturity of 6 months or less (in the USA 6-month treasury bills is 182 days) is frequently compared to the simple interest rate for a bond that will also soon reach maturity and therefore pay its last coupon (the rate corresponding to the last coupon being then also named "discount rate").

The objective is then to transform the discount rate of a treasury bill on a 365 day time base into a simple interest rate on a bond whose time base for the last coupon is $30/360$ at Wall Street (known as "discount base").

During our study of discounts we demonstrated that if the rate $t\%$ communicated from the start is not the rate corresponding to the entire duration of n but the rate of the corresponding time unit, then we have the following relation:

$$I=\frac{t\%}{n(1-t\%)}=\frac{nt'\%}{n(1-nt'\%)}=\frac{t'\%}{(1-nt'\%)} \tag{15.3.499}$$

So as in finance the treasury bills have a discount rate given in annual and that n is in days, we have on a $30/360$ basis the corresponding rate as simple daily interest rate below (the second equality is simply a traditional notation...):

$$I_d=\frac{\dfrac{t'\%}{360}}{1-n\dfrac{t'\%}{360}}=\frac{t\%}{1-n\dfrac{t\%}{360}}\frac{1}{360} \tag{15.3.500}$$

Then, as it is a simple interest and traditionally the rates of the treasury bills are given annually over a 365-day year, we then have:

$$I_y=365\frac{t\%}{1-n\dfrac{t\%}{360}}\frac{1}{360}=\frac{t\%}{1-n\dfrac{t\%}{360}}\frac{365}{360} \tag{15.3.501}$$

We fall back on the relation given in Microsoft English for the function named `T.BILLEQ()`.

✍**Example:**

Consider a treasury bill emitted at the price of 98.- the 11999-03-30 (holocene calendar) and with maturity 11999-06-01 (hence 63 days) and with maturity price of 100 (hence an increase of 2.-) what is the bond equivalent discount rate of this treasury bill?

First we can easily calculate:

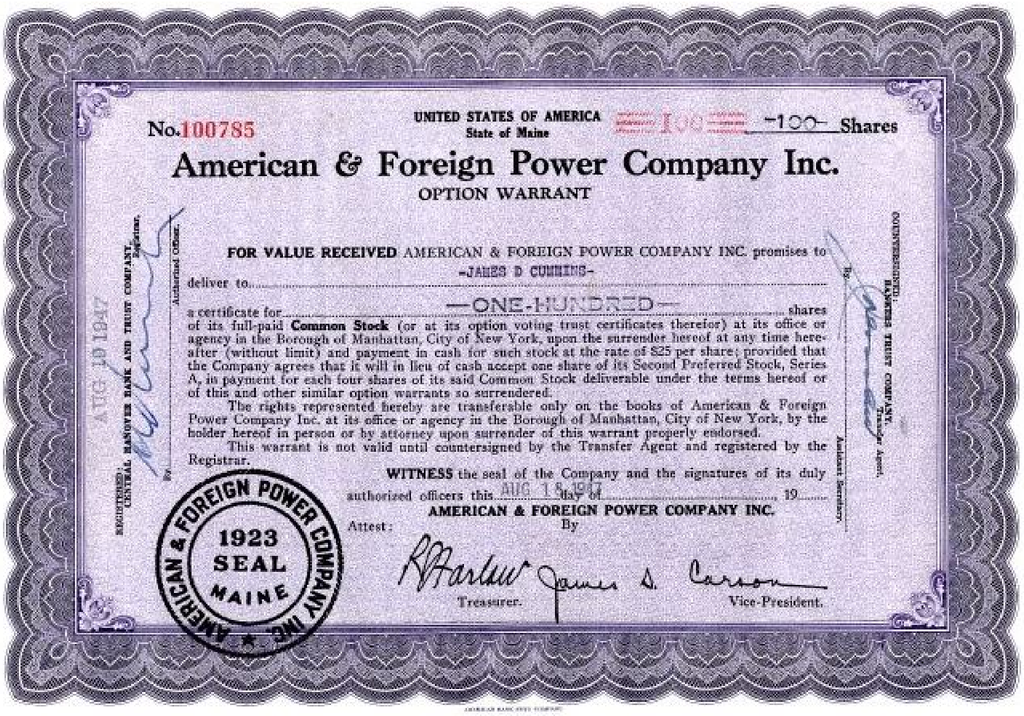
$$t\%=\frac{\text{CF}}{\text{TB}_A+\text{CF}}=\frac{2}{98+2}=20\% \tag{15.3.502}$$

and now we compute:

$$\begin{aligned} i_{\text{a}} &= \frac{t\%}{1-n\dfrac{t\%}{360}}\frac{365}{360} = \frac{20\%}{1-63\dfrac{20\%}{360}}\frac{365}{360} \\ &= \texttt{TBILLEQ(11999-03-30,11999-06-01,20\%)} = 21.01324\% \end{aligned} \tag{15.3.503}$$

15.3.10.3.3 Warrants

Definition (#15.3.687): Basically a "warrant" or "option warrant" ("stock-warrant" or "equity warrant"... as their are many type of warrants it is important to precise!) is the privilege sold by one party to another, that gives the buyer the right, but not the obligation, to buy (Call) or sell (Put) at an agreed-upon price (often an average of stock prices before the issuance of the warrants) within a certain period or on a specific date another financial underlying security that has to be issued (share, bond or any other good...). As these both financial instrument are linked to an underlying they therefore belong both to the category of derivatives instruments.



No. 100785

UNITED STATES OF AMERICA
State of Maine

100 -100- Shares

American & Foreign Power Company Inc.

OPTION WARRANT

REGISTERED:
CENTRAL HANOVER BANK AND TRUST COMPANY, Registrar.
By Authorized Officer.
AUG 19 1947

FOR VALUE RECEIVED AMERICAN & FOREIGN POWER COMPANY INC. promises to deliver to -JAMES D CUMMINS- a certificate for —ONE-HUNDRED— shares of its full-paid **Common Stock** (or at its option voting trust certificates therefor) at its office or agency in the Borough of Manhattan, City of New York, upon the surrender hereof at any time hereafter (without limit) and payment in cash for such stock at the rate of $25 per share; provided that the Company agrees that it will in lieu of cash accept one share of its Second Preferred Stock, Series A, in payment for each four shares of its said Common Stock deliverable under the terms hereof or of this and other similar option warrants so surrendered.

The rights represented hereby are transferable only on the books of American & Foreign Power Company Inc. at its office or agency in the Borough of Manhattan, City of New York, by the holder hereof in person or by attorney upon surrender of this warrant properly endorsed.

This warrant is not valid until countersigned by the Transfer Agent and registered by the Registrar.

WITNESS the seal of the Company and the signatures of its duly authorized officers this AUG 18 1947 day of ...., 19......

AMERICAN & FOREIGN POWER COMPANY INC.

Attest: By

Treasurer. James D. Carson Vice-President.

COUNTERSIGNED:
BANKERS TRUST COMPANY, Transfer Agent.
By Assistant Secretary.

1923
SEAL
MAINE

Figure 15.114: Warrant old style paper version

✍Example:

The warrant of a given company X gives to possibility to subscribe for one share of the company at a price of 500.- until April 30, 12004 (holocene calendar). If the share of X exceeds the level of 525.-, the warrant that allows in this situation the right to purchase one share at a lower cost to the stock market proves a winning placement. If the action is so for example traded at 525.- in April 30, 12004 (holocene calendar), the gain will be worth 25.-.

In addition, Warrants (also Call Options), are financial assets with low risk since there is no obligation to apply them and they were offered... it could also be noted that many companies cancel employee Warrants when they leave the company too early...

Even if we don't have seen yet what are Call Options (see further below) it seems to us important to make already the DEFAULT difference between Warrants and Call Options as they are very similar. For instance, many warrants confer the same rights as equity options and warrants often can be traded in secondary markets like options. However, there also are several key differences between warrants and stock options (or more generally equity options):

- The main default difference is straightforward and we already know it. It's the "dilutive effect" (see definition during our study of shares). Indeed an option is the right to purchase an existing share of a company's stock from the company at a specific price (usually the fair market value of that share on the date of issue), whereas a warrant gives the holder the right to purchase a share that will need to be issued (that is, "created") in the future. Therefore there is a boosting int the total share count.
- Warrants are issued by private parties, typically the corporation on which a warrant is based, rather than a public options exchange.
- Warrants are considered over the counter instruments and thus are usually only traded by financial institutions with the capacity to settle and clear these types of transactions.
- A warrant's lifetime is measured in years (as long as 15 years), while options are typically measured in months. Even LEAPS (Long-Term Equity Anticipation Securities), the longest stock options available, tend to expire in two or three years.
- Warrants are not standardized like exchange-listed options.
- The tax rules governing options and warrants are completely different.

Remark

The big problem with warrants is that because the supply is limited, brokers can do all sorts of nasty things to manipulate the price of the warrants. Because anyone can issue an option, there are fewer nasty things that you can to do manipulate their price.

The warrant (as Call Options) allows to be interested in the rise or fall of a security without having to devote the same amount of money than buying shares directly. Therefore, on acquisition, if the underlying security has a higher value than on the warrant, the buyer will make a profit named "surplus value acquisition". After the buyer that now owns the underlying securities may well sell them when the price is higher than when it made the acquisition and then it creates a (pseudo) second benefit named "cession plusvalue gain".

Remark

So, one warrant can be attached to the issue of a share or obligation. Therefore, depending on the case, we speak about "shares warrants to purchase shares" (SWPS) or "bonds warrants to purchases shares" (BWPS), but also of "bonds warrants to purchase bonds" (BWPB) or about "shares warrants to purchase bonds" (SWPB).

Since the issuance of these composed instruments, all split into parts: the stocks or the bonds become again classic titles and warrants has its own life. They are traded separately after the issuance.

For the mathematical point of view the pricing future value of warrants is not easy as we can not apply the Black & Scholes equation that we will prove later as:

- A warrant has a long life (2-4 years) and makes it difficult to accept the consistency of the hypothesis of constant interest rates used by the Black & Scholes model.

- Any operation of the issuer that changes the underlying security value affects the value of the warrant. Indeed, companies have the legal right to issue a new contract for the Warrants and to change their value and the time period of their validity!

- If the underlying warrant is a bond, the price of this latter changing over time and knowing that the more a bond approaches maturity, the more its value tends toward its redemption price. Its volatility is gradually reduced, which renders inapplicable the Black & Scholes model which postulates the constancy of volatility in time!

Therefore if Black & Scholes assumptions are not valid the only one simple thing we know so far is to build some naive mathematical model to computer their actual value.

In order to see how this works we need some simple notation (all values are those actual time t):

1. E is the value of the firm's equity at time t.
2. S is the common share price of the outstanding shares at time t.
3. N is the number of old common shares outstanding prior to any exercise of the warrants.
4. W is the value of the warrant at time t.
5. M is the actual number of warrants.
6. X is the exercise price of the warrants.

The choice of whether to exercise a warrant at maturity can be based on the following criteria. We exercise if:

$$X<\frac{E+MX}{N+M}\Leftrightarrow X<\frac{E}{N} \tag{15.3.504}$$

That is to say that we exercise the warrant if the average value per share of the company is greater than the actual value of the warrant (obvious!).

If the stock price after all the warrants are exercised is greater than the warrant's exercise price, all holders of the warrants will exercise. Hence the payoff of a warrant at maturity is:

$$\max\left(\frac{E+MX}{N+M}-X,0\right)\equiv\left(\frac{E+MX}{N+M}-X,0\right)^{+}\equiv\left(\frac{E}{N}-X,0\right)^{+} \tag{15.3.505}$$

This is quite different than the payoff of a regular Call Option as we will see later in that the payoff depends on the number of warrants outstanding.

Remark

The development of liquidity in the equity and bond markets has prompted financial institutions to issue warrants to acquire securities (underlying) existing independently of the financial operations of the concerned company. These warrants relate only to investors and are issued only by banks between them and do not allow the business financing (so this is pure speculation!). These warrants (also traded) are named "covered warrants" because on issue, the financial institution covered itself by purchasing shares on the market.

If necessary here is a small summary of main differences between options and warrants:

OPTIONS vs WARRANTS	
Options	**Warrants**
Options are exchange-traded products. They are regulated by the exchange, and features of the different options are also fixed by the exchange	Warrants are generally over-the-counter (OTC) products. A few companies may list them on exchange as well though in order to publicize the underlying stock for marketing purposes, but mostly they are traded by broking firms
Options have a variety of underlying products like, bonds, equities, interest, rates, currency or even commodity, and sometimes secondary derivatives having swaps as underlying (in which they are named "swaptions")	Warrants generally have underlying as stocks of the company which issues warrants
The seller of an option can be any party trading on the market, need not necessarily be the issuer of the underlying security	In case of warrants, the seller is the company itself which issues the underlying stocks that buyer can buy on a future date
As the product is exchange-traded, price of the option can be discovered and tracked	Price of the warrant is difficult to be discovered as it is an OTC product, and may be based on discretion of the seller..
Options may be traded for shorter period of time, as they are very volatile products	Warrants are generally traded for long periods (in years)
Options are standardized products since they are exchange-traded, and the features are set by exchange	Warrants are not standardized - features depend upon the issuer or the seller, and accordingly profits to be made are determined by buyer
There is no intention by the seller to market the underlying by selling the option. It may just be for the purpose of hedging their risk, or speculation or even arbitrage	Upon selling a warrant, the issuing company may link it to their bonds issued or to publicize selling of the underlying share which is to be introduced into markets in future
Options can be easily traded by approaching the exchange or hiring a broker/intermediary party. They can be traded in volumes with a purpose of hedging risk/speculation/arbitration, and which may also lead to a bubble in economy	Issuance of warrants is very much controlled by the issuing entity, and does not lead to uncontrollable trading

Table 15.46: **Differences between Options and Warrants**

15.3.10.3.4 Futures & Forwards

When we think about selling and buying commodities/currencies we target two types of markets:

1. An immediate physical cash settlement market also named "spot market" involving producers and consumers.
2. A physical market at maturity also named "Forward/Future market" involving producers and that have for desire to hedge price risks consumers and speculators that want to play with the trend of the underlying to make money.

Definition (#15.3.688): A "Forward contract" (or "commodity Forward") and a "Future contract" (or "commodity Future") are both contracts to buy or sale a financial product/underlying (therefore they belong both to the category of derivatives instruments), passed between two counterparties, all of whose characteristics are fixed in advance: settlement date Forward prices, etc. The price entered is named "Forward price" or respectively "Future price" or simply "fair value" for the both, and the exchange and also the payment will be mandatory at that price regardless of the current market price underlying (also named "spot price") on the date of delivery (i.e. maturity date)!

Fundamentally, Forwards and Futures contracts have the same function: both types of contracts allow people to buy or sell a specific type of asset at a specific time at a given price. The purpose of Futures and Forwards for stakeholders is obviously to freeze the prices over the future: it is in this case a hedge (hedging) in the point of view of the seller.

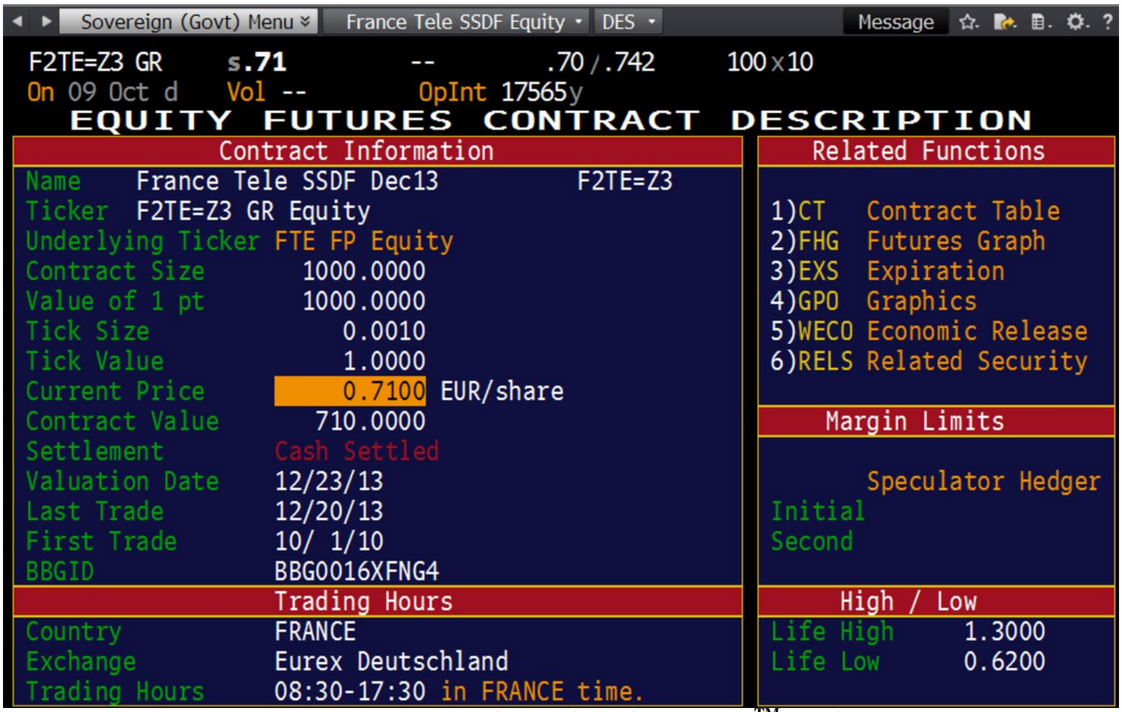
However, it is in the specific details that these contracts differ:

1. Futures contracts are exchange-traded and, therefore, are standardized contracts. Forward contracts, on the other hand, are private agreements between two parties and are not as rigid in their stated terms and conditions. Because Forward contracts are private agreements, there is always a chance that a party may default on its side of the agreement. Futures contracts have clearing houses that guarantee the transactions, which drastically lowers the probability of default to almost never.
2. Specific details concerning settlement and delivery are quite distinct. For Forward contracts, settlement of the contract occurs at the end of the contract. Futures contracts are marked-to-market daily, which means that daily changes are settled day by day until the end of the contract. Furthermore, settlement for Future contracts can occur over a range of dates. Forward contracts, on the other hand, only possess one settlement date.
3. Lastly, because Future contracts are quite frequently employed by speculators, who bet on the direction in which an asset's price will move, they are usually closed out prior to maturity and delivery usually never happens (closed out prior is not authorized with Forwards). On the other hand, Forward contracts are mostly used by hedgers that want to eliminate the volatility of an asset's price, and delivery of the asset or cash settlement will usually take place. Today, around 97% of Futures trading is done by speculators.

4. Forward don't have brokers margins. Indeed, margin are set by the futures exchanges to lower the risk exposure. Therefore the higher is a future margin the higher is the volatility of the underlying. As distinction is made between:
   - Initial Margin: That is the amount of money that is required to open a buy or sell position on a Future contract.
   - Margin Maintenance: That is the amount of money where a loss on your Futures position requires you to allocate more funds to bring the margin back to the initial margin level.
   - Margin Calls: A margin call on Future contracts is triggered when the value of your account drops below the maintenance level. The investor has then the option of taking and opposite positions to eliminate the margin call to pay the difference. In any case if the buyer don't pay the broker bears the cost. If the broker then defaults the clearing house bears the cost!

To summarize, the different types of futures/forwards contracts are generally divided into two main groups that are commodities futures/forwards and financial futures/forwards. Commodity futures/forwards cover traditional agricultural, metal and petroleum products. Finance futures/forwards include stocks, bonds and currencies and the main difference are:

FORWARD:	**FUTURES**:
Contracts executed by banks	Contracts executed by brokerage houses
Private contract (unregulated)	Traded on organized exchanges
Not standardized contract	Standardized contract (regulated)
High counterparty risk	Low counterparty risk
No initial payment required	Initial margin payment required
Normally one specified delivery date	Range of delivery date
Settle and the end of maturity	Daily settled
Traded 24 hours a day	Traded $4-8$ hours a day
No cash exchange prior to maturity	Profit/Loss are paid in cash
More than 90% of contract settled by actual delivery of assets Delivery or final cash settlement usually takes place	Not more than 5% of contract settled by actual delivery of assets
	Contract normally closed out prior to the delivery



```
F2TE=Z3 GR     s.71        --        .70 / .742     100 x 10
On 09 Oct d    Vol --         OpInt 17565y
  EQUITY  FUTURES  CONTRACT  DESCRIPTION
            Contract Information                   Related Functions
Name      France Tele SSDF Dec13          F2TE=Z3
Ticker    F2TE=Z3 GR Equity                        1)CT    Contract Table
Underlying Ticker FTE FP Equity                    2)FHG   Futures Graph
Contract Size          1000.0000                   3)EXS   Expiration
Value of 1 pt          1000.0000                   4)GPO   Graphics
Tick Size                 0.0010                   5)WECO  Economic Release
Tick Value                1.0000                   6)RELS  Related Security
Current Price             0.7100 EUR/share
Contract Value          710.0000                       Margin Limits
Settlement           Cash Settled
Valuation Date       12/23/13                              Speculator Hedger
Last Trade           12/20/13                      Initial
First Trade          10/ 1/10                      Second
BBGID                BBG0016XFNG4
                     Trading Hours                       High / Low
Country              FRANCE                        Life High      1.3000
Exchange             Eurex Deutschland             Life Low       0.6200
Trading Hours        08:30-17:30 in FRANCE time.
```

Figure 15.115: Future Equity description in Bloomberg™ Terminal

In 12003 (holocene calendar), there would have been $2,848$ million Forward contracts exchanged on markets respectively 2.06% for exchange rates, 17.08% for raw materials (43.9% for food, 10% for metals, 45.8% for fuels and 0.3% other), 25.48% for shares and bonds and 55.37% on interest rates.

✍**Example:**

A Swiss industrialist knows that he must receive an important amount of money in euros in six months. To hedge against a falling of euro (...), he buys a Forward contract of sale, with maturity six months on euro, in swiss franc (if financial operators also know that the euro will fall, the contract Forward sales will have a high enough cost so that the purchase contract is not interesting). Note that this currency hedging operation may be adversely to him if in six months, the contract gives less than the future spot exchange rate.

As we have already mention it, the major difference between Futures and Forwards is that the gains and losses in relation to underlying fluctuations are paid to the counterparty on a daily basis for Futures ("daily mark to market margin" or "daily mark to market settlement") !!! This means that the gain or loss is already almost entirely / disbursed the day of the settlement date (only remains to adjust the difference of the last day).

✍**Example:**

Consider a contract about $1,000$ tons of apples for delivery on a given settlement date. A future contract is available for a future price of 200.- by $1,000$ tons. The initial margin is of 20.- and the maintenance margin of 10.-.

Let's say that on first trading day the traded value of the future value is 195.-. Obviously the buyer of the 200.- contract will feel sad and the seller will feel smart! But the principle behind the mark to market is the margin will help the seller to avoid that the buyer close the position. Indeed, as there is 20.- margin, the 5.- loss can be taken on this margin and theoretically nothing is lost and the 200.- Future will be reset to 195.- because remember that if the seller don't do that the buyer will close is positions. Then it remains 15.- margin.

Let's say that on second trading day the traded value of the Future value is 188.-. Obviously the buyer of the now 195.- contract will feel once again sad and the seller will feel smart! But the principle behind the mark to market is the margin will help again the seller to avoid that the buyer close the position. Indeed, as there is 15.- margin remaining, the 7.- loss can be taken on this margin and theoretically nothing is lost and the 195.-. Future will be reset to 188.- because remember that if the seller don't do that the buyer will close is positions. Then it remains 8.- margin. But now as the remaining margin is less than the maintenance margin of 10.-, through a margin call, the buyer has the option to put 12.- again to be at the initial value of 20.- or to close the position.

If the buyer refuses to add the extra 15.- dollar. The contract is unbound. So the seller can sell his contract to someone else. The loss of the buyer is 20.-. In a "good" contract those 20.- aren't lost by the buying or selling party. The margin is just there as an assurance to both parties that ensure they will keep their position until the end of the deal (or lose 20.-).

At the opposite if the spot price goes up, then the seller has to transfer that difference by which the price increased to the buyer in order to cancel the changes in price. The seller also has a margin not only the buyer!

It must be understood that if the price of the contract goes down to 100.-, then the buyer only pays 100.-, but then adds on an extra 100.- in the margin account. If the final value of the contract is 200.- then the seller and the buyer take back the 20.- initial margin.

As we can not cancel ("opt-out") a Future contract, "closing out" a Future contract means that you enter into a contract that is opposite to the one you initially entered. In this example, the losing party (the buyer) would sell a contract for delivery at 185.-.

Thanks to this example we can see that the margins reduce volatility by eliminating counter-party risk, that is to say the risk that the exchange will end up buying the product because the original buyer disappears without having paid anything.

We see with the above example that the Future contract is something very safety for the seller of the commodity (farmer typically), assimilated to the "hedger", as he will have anyway at least his 200.- if the market price decrease (but if the market price raise he will virtually loose money as he will not get the positive difference, therefore pricing Future correctly is important to be at the best right price!!).

But how do speculators make money with Futures because of this mark to market margin????? Simple! Imagine they buy at day d the Future contract for $200.-$ plus the margin hoping that the price of commodity will increase. If n days after, the price of the commodity increase of δ (the seller has to give us the δ day by day thanks to the mark to market margin) so will do the old contracts and new investors coming in the market will have the choice to buy a completely new Future contract (issue by the farmer) with the price $200+\delta$ and a given margin or an old contract in possession of a speculator who's spot price is also equal to $200+\delta$ plus the margin. And therefore the speculator has make a profit of δ!!! In fact the speculator wins and keeps the day to day cumulative margin!!!!!!!

Two types of executions may occur:

1. The "physical settlement" the underlying is really exchanged (which is rare in the virtual economy, but in the real economy should be a reality...).
2. The "cash settlement": if the price of the underlying is below the set price, the buyer supplied himself on the market he wants and pays the difference to the seller and vice versa.

In practice, speculators, named "index managers", for which the raw material representing the underlying matter is of no interest, have the mission to reposition their portfolios before settlement to avoid physical delivery. This procedure, named "roll-over". The roll-over is normally done on the last day or last week of the month preceding the start of the period of delivery period.

A Future or Forward contract must refer to the role of each participant (buyer or seller), an official reference market - underlying - (stocks, indexes, bonds, currency market, interest rates, etc.) - whose spot price at time t will be denoted S_t, the settlement date reference denoted T, the nominal price K, the notional amount N to which the transaction of the Futures of Forwards is targeted (which is therefore a quantity for physical settlement and that is mathematically almost always reduced to unity), payment instructions (physical or cash settlement) and for Futures the various requires margins M_i (initial margin), and M_m (maintenance margin).

At maturity T, the gain or "payoff" of the contract is in both cases as we will see in the table below (whether for a Forward or Future):

$$\boxed{F_T=S_T-K} \tag{15.3.506}$$

This value[21] simply refers to losses or realized gains (we will reverse the sign depending on whether the trader is a buyer or seller). Indeed, it can be understood better by seeing the cash flow amount in the table below (this also helps to explain why the losses or gains are daily for Futures and that if in the case of gains we can reinvest them quickly on the market):

---

[21] Future prices are sometimes denoted $\mathcal{F}_t$ in textbooks

Time	Forward	Future
0	0	0
1	0	$S_1 - K$
2	0	$S_2 - (K - (K - S_1)) = S_2 - S_1$
3	0	$S_3 - (S_2 - (S_2 - S_1)) = S_3 - S_2$
4	0	$S_4 - S_3$
5	0	$S_5 - S_4$
$\ldots$	0	$\ldots$
$\ldots$	0	$\ldots$
$\ldots$	0	$\ldots$
$T-1$	0	$S_{T-1} - S_{T-2}$
T	$S_T - K$	$S_T - S_{T-1}$
Total:	$S_T - K$	$S_T - K$

Table 15.47: **Cash Flow of Forward and Future of the point of view of a buyer**

15.3.10.3.4.1 Futures & Forwards naive pricing

The pricing of Forward or Future contracts as part of a naive approach is quite simple if the underlying yield is deterministic and the average geometric return of the market as often assumed to be known .... But we must be honestly agree on one thing ...: the final price of a Forward contract or its Future equivalent (at least in the conditions seen before...) must be constituted at final only of the risk premium that will take upon him the seller of the contract.

For the pricing let start with the first element: there is a first information needed from the perspective of contract vendor, it will already actualize K (fixed term value of the contract) to the free risk rate of the to already know the lower limit of price that it will ask today to get K when the time comes (if the seller does not ... other vendors will offer lower prices to the buyer - competition does its job - or the buyer will do the calculation work himself). That is to say:

$$\frac{K}{(1+t\%)^T} \tag{15.3.507}$$

And of course if the sale of the contract is done later than the date of issue, we have:

$$\frac{K}{(1+t\%)^{T-t}} = K(1+t\%)^{-(T-t)} := K(1+t\%)^{-\tau} \tag{15.3.508}$$

What will be written in the continuous case (and adapting traditional notation that of continuous rate of return):

$$Ke^{-r\tau} \tag{15.3.509}$$

Therefore this is the lower limit to apply at any time t for a potential contract buyer (waiting the maturity of this amount of money will be placed in an investment considered as safe). However, this is not strictly speaking the risk premium and thus, the contract price in the sense we usually hear it. Indeed, the seller of the contract will have to buy the underlying of contract. If for example the sale of the contract is made just at its issue (that is to say at time T before maturity), then we have the actualized price of the assumed (projected) value of the underlying at maturity knowing it will have himself intrinsic yield (that the seller will take in the worst case: that is to say a positive rate) which we will denote y (the "spot-rate") and in non-stochastic framework we can do this calculation only based on something known which is the underlying spot price:

$$S_0 e^{-yT} \tag{15.3.510}$$

That is to say for any time t:

$$S_t e^{-y(T-t)} = S_t e^{-y\tau} \tag{15.3.511}$$

We then have the actualized payoff of a Forward or Future corresponding to the risk premium which is given in any time under the assumptions underlying the previous relations by the difference (obviously buyer point of view!):

$$S_t e^{-y\tau} - K e^{-r\tau} \tag{15.3.512}$$

which is written by tradition in the purpose to make appear the Future or Forward price at each day t:

$$S_t e^{-y\tau} - K e^{-r\tau} = S_t e^{-y\tau} e^{r\tau} e^{-r\tau} - K e^{-r\tau} = (S_t e^{-y\tau} e^{r\tau} - K) e^{-r\tau} \tag{15.3.513}$$

The factor on the left is the risk free actualization factor. Therefore we guess that we can focus only on the parenthesis: $(S_t e^{-y\tau} e^{r\tau} - K)$ that is the non-actualized risk premium. Fro this we conclude that:

$$S_t e^{-y\tau} e^{r\tau} - K = F_t - K \tag{15.3.514}$$

and therefore:

$$\boxed{F_t = S_t e^{-y\tau} e^{r\tau} = S_t e^{(r-y)\tau}} \tag{15.3.515}$$

It is interesting to notice that y is subtracted of the risk free rate r and therefore makes the contract price lower. This is obviously valid if be in possessions of underlying can makes us win money, but if this make us loose money (for example the cost of storage can be very expensive), then y will be negative and will increase the Forward or Future price.

If the underlying pays dividends, then obviously the value above is overestimated. We have logically then to actualize dividends and to subtract it to have the value of the Future or Forward equal to:

$$\boxed{F_0 = S_0 e^{(r-y)T} - \sum_{i=1}^{n} D_i e^{-r(T-t_i)}} \tag{15.3.516}$$

Definitions (#15.3.689):

D1. As for any derivative, we say that the market is "contango" if the option price (in our case: the Future or derivative) is greater than the spot price of the underlying.

D2. As for any derivative, we say that the market is "backwardation" if the option price (in our case: the Future or derivative) is lower than the spot price of the underlying.

Therefore, more explicitly, contango is a situation where the Futures price (or Forward price) of a commodity is higher than the expected spot price. In a contango situation, hedgers (commodity producers and commodity users) or arbitrageurs/speculators (non-commercial investors), are "willing to pay more (now) for a commodity at some point in the future than the actual expected price of the commodity (at that future point). This may be due to people's desire to pay a premium to have the commodity in the future rather than paying the costs of storage and carry costs of buying the commodity today". And a market is "in backwardation" when the Futures price is below the expected future spot price for a particular commodity. This is favourable for investors who have long positions since they want the futures price to rise.

A contango is normal for a non-perishable commodity that has a cost of carry. Such costs include warehousing fees and interest forgone on money tied up (or the time-value-of money, etc.), less income from leasing out the commodity if possible (e.g. gold).

The figure below depicts how the price of a single Forward contract will behave through time in relation to the expected Future price at any point time:

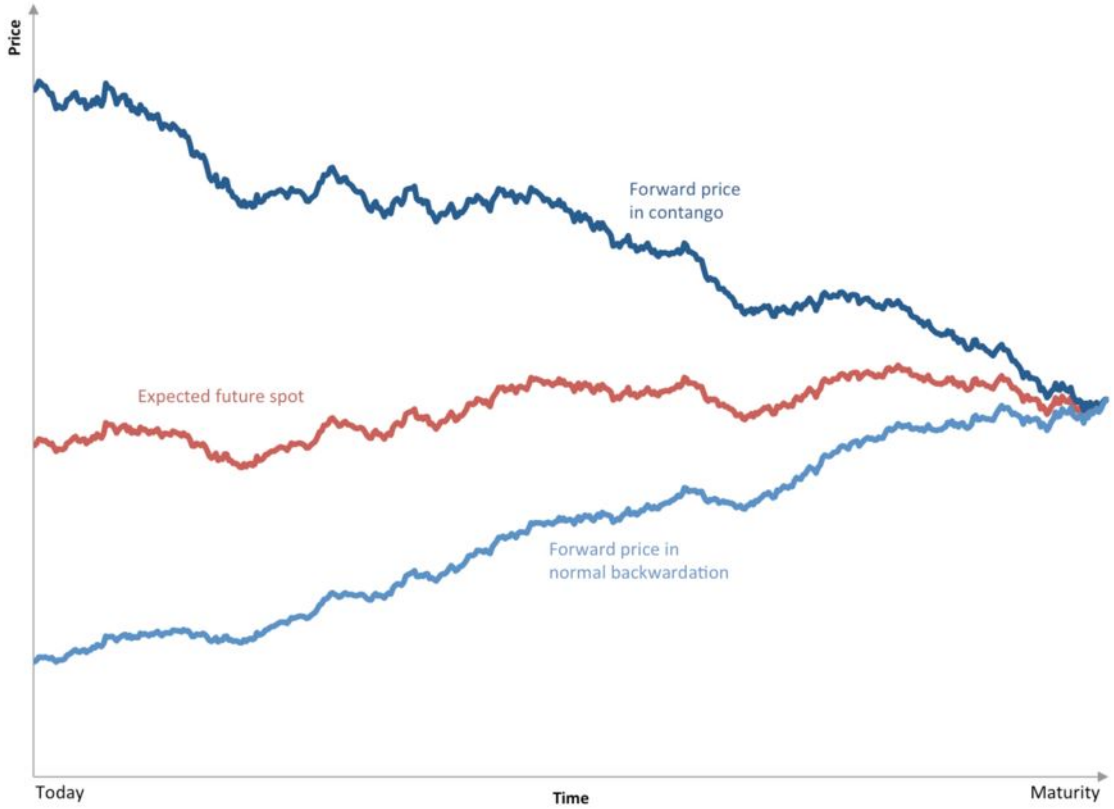


Figure 15.116: **Typical spot price and Future/Forward price dynamics**

✍**Example:**

A customer wants to buy to us in 6 months US dollars with Euros (or in other words he pays in Euros in 6 months a machine that he bought today and we are working in US dollars). We would then protect us already against changes in the raise of the monetary value (bid/ask) of the US dollar (because if the US dollar rises then we lose verbatim money since then 1 Euro is less than 1 US dollar). However, there is available on the market only Futures issued already 6 months before with maturity (cash settlement) in 12 months. Knowing that the market risk-free rate is 3% (in continuous rate...), that the dollar rate is assumed at worst at $+1.5\%$ (also in continuous rate...) and that we want to know the contract price in monetary unit and that the contracts emitted 6 months ago with maturity in 6 months had an exercise price of €1.00.- for \$$1.50$.- and that today the cost of €1.00.- is of \$$1.60$.-, the risk premium that we have to pay is therefore:

$$S_t e^{-y\tau} - Ke^{-r\tau} = S_{0.5}e^{-1.5\%(1-0.5)} - Ke^{-3\%(1-0.5)} = 1.6e^{-0.75\%} - 1.5e^{-1.5\%} \cong 0.11 \tag{15.3.517}$$

dollars per unit of currency in desired Euros (so we pay 11% as a risk premium which is a typical value of what is practised empirically on markets: added to 5% to 15%). That sounds like a lot but it limits our risk and transfers it on the seller of the contract! Moreover, buying the machine we could also see first the risk premiums and adapt price negotiations accordingly.

The spot price of the Future contract is therefore obviously:

$$F_{0.5} = S_{0.5}e^{-y\tau} = S_{0.5}e^{-1.5\%(1-0.5)} = 1.6e^{-0.75\%} \cong 1.612 \tag{15.3.518}$$

that is to say \$$1.612$.- for €$1.00$.- (which corresponds to the result returned by the GUIDE `R` package). So we cannot obviously get \$$1.50$.- for €$1.00$.- and this difference is explained by the risk taken by the seller of the Future contract.

So the total price of the Future is finally $1.612 + 0.11 = \$1.722$.- while the spot price is \$$1.60$.-. The difference, omitting the risk premium, is of \$$-0.012$ and is named the "basis" or "cost of carry" and it indicates to a future buyer that market agents are planning an increase if the basis is negative ("under") and a decrease if positive ("over").

Remark

In practice, as with many financial instruments, the risk premium is included in a broader terminology fees we name "margin money" which includes transaction costs, risk premium, insurance costs for bankruptcy of the buyer of the Future/Forward contract, etc. Some institutions, however, use the term "cost of carry" for all types of costs and then the risk premium alone is named "net cost of carry".

We will not look now the mathematical tools for the pricing of Futures/Forward contracts using stochastic return, because as they are "only" a special case of the pricing of options that we will see later it would be redundant to present every theorems twice while there is only a small change to make for in the Black & Scholes model to get one from the other.

Indeed, as we will see an option is a derivative that does not necessarily result in an execution due to its optional character in relation to a Future/Forward contract where the settlement date is fixed in advance.

15.3.10.3.4.2 Cox-Ingersoll-Ross equality of the Future/Forward price

Under a specific condition and a given strategy it is possible to demonstrate that the price of a Future is equal to that of a Forward at maturity (we own this proof to J.C. Cox, J.E. Ingersoll and S.A. Ross in their article of 11981 according to holocene calendar).

Let us suppose that the market risk-free rate is r and that the latter is constant and that each day i we apply the very special strategy of selling Futures in a quantity e^{-ri} in the so-named "CIR conditions" (these are the initials of the names of the authors of this strategy).

So in the beginning we sell Forward but we have to actualize the price, so the actualized price is then:

$$e^{-r}(F_1-K)e^{-r(T-1)}=(F_1-K)e^{-rT} \tag{15.3.519}$$

On the second day, we sell contracts that must also be actualized:

$$e^{-2r}(F_2-F_1)e^{-r(T-2)}=(F_2-F_1)e^{-rT} \tag{15.3.520}$$

And we quickly see that we will have in total:

$$(F_1-K)e^{-rT}+(F_2-F_1)e^{-rT}+(F_3-F_2)e^{-rT}+\ldots+(S_T-F_{T-1})e^{-rT} \tag{15.3.521}$$

This is simplified immediately to:

$$(S_T-K)e^{-rT} \tag{15.3.522}$$

We thus fall back on the actualized of the payoff of the forward according to the table seen earlier above. Obviously, this equality is valid only if the return of the risk-free market is constant and in the strategy mentioned above, which is not the case (realistic) the majority of time!

As far as we knew, even if this theoretical result is interesting, practical implications, however, of this observation are minimal.

15.3.10.3.4.3 Futures & Forwards commodity hedging

A company will prefer Future/Forwards to Options if its purpose is to know its costs in advance so that it is 100% certain to consume the underlying! Even if the purchase of an Option requires obviously the payment of a risk premium that generates a cost of opportunity. In the case where the company is not sure to use the underlying, it will prefer the Option that can be executed or not, which means that the management reserves a good result whatever happened but they will pay in counterpart the option risk premium.

Let us study now an interesting case of risk hedging of raw materials. Imagine that a company needs to buy for a total amount S a raw material but for which there is sadly no Forwards or Futures contracts available (and no partner agree to make an OTC contract). This company wish to hedge against an increase in the price of the raw material that it is interested for and therefore at least two typical strategies are available to it:

1. Found Future or Forward contracts of a commodity positively correlated to the raw material of interest. Therefore, hoping that this commodity replicates well the behaviour of the raw material of interest and that when its price will increase the other one will also probably do same (and if possible in the same proportions)... and the idea is to have a buyer (long) position of contract to buy at maturity the commodity to a value less than the spot price and hope to find someone to sell the latter to people interesting to buy quickly and use the payoff to buy the raw material of interest with a minimum loss (an in the ideal case with a loss equal to zero or even a positive remainder).

2. Found Future or Forward contracts of a commodity negatively correlated to the raw material of interest. Therefore, hoping that this commodity replicates well the behaviour of the raw material of interest and that when its price will decrease the other one will probably increase (and if possible in the same proportions)... and the idea is to have a seller (short) position of contract to sell at maturity the commodity to a value more than the spot price and use the payoff to buy the raw material of interest with a minimum loss (an in the ideal case with a loss equal to zero or even a positive remainder).

Mathematically the strategy consists to write in the both cases (we will see later what sign corresponds to which strategy):

$$\boxed{\Delta V=\Delta S\pm N\Delta F} \tag{15.3.523}$$

where ΔS is the variation of the raw material price on the market and ΔF is the payoff of the contracts that replicates positively or negatively the variation (increase) of the price S. We have ΔV that is the difference of the good or bad replication and that must in the ideal case tends to zero. Another way to see the context, more common as it encompass the point that makes problem (that is to say: the volatility!!!), is to write this strategy with the information available on the market such that we take the variance of the previous relation:

$$\sigma_{\Delta V}^2=\sigma_{\Delta S}^2+N^2\sigma_{\Delta F}^2\pm 2N\sigma_{\Delta S,\Delta V} \tag{15.3.524}$$

and to search the value N that minimize the variance ΔV in function of the correlation of the both commodities (that we will have to make appear) and of their volatility (containing implicitly the quantity as we will see). We have then by derivating relatively to N:

$$\frac{\partial\sigma_{\Delta V}^2}{\partial N}=2N\sigma_{\Delta F}^2\pm 2\sigma_{\Delta S,\Delta V}=0 \tag{15.3.525}$$

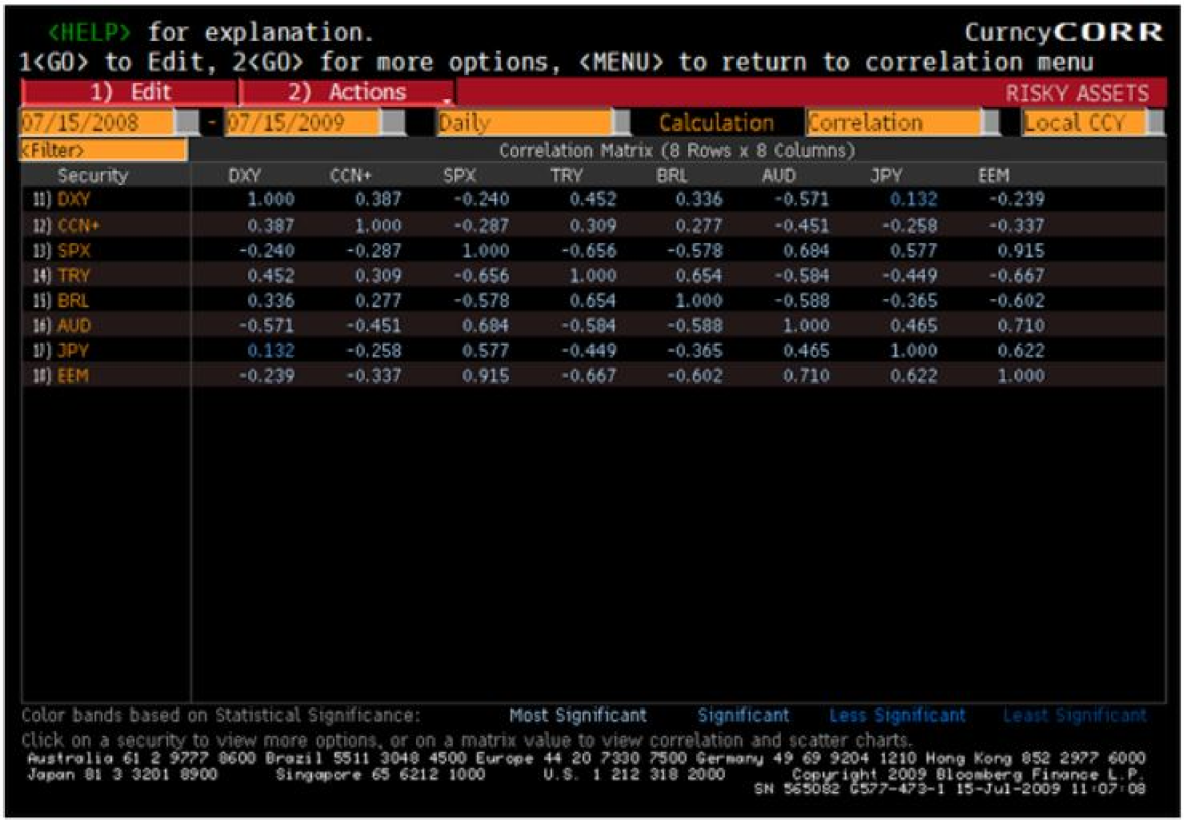
Which gives using the relation between the covariance and the linear correlation coefficient proved in the section Statistics (see page 597):

$$N^*=\pm\frac{\sigma_{\Delta S,\Delta V}}{\sigma_{\Delta F}^2}=\pm\frac{\text{cov}_{\Delta S,\Delta V}}{\sigma_{\Delta F}^2}=\pm\frac{R_{\Delta S,\Delta V}\sigma_{\Delta S}\sigma_{\Delta F}}{\sigma_{\Delta F}^2}=\pm\frac{R_{\Delta S,\Delta V}\sigma_{\Delta S}}{\sigma_{\Delta F}} \tag{15.3.526}$$

As the standard deviations are positive and N must be positive (and be rounded to the nearest integer to be a realistic physical solution), it is obvious that the correlation coefficient should always be of opposite sign to the type of the chosen strategy. This is why we have finally for tradition to take (as it still remains the same) the positive sign such as:

$$\boxed{N^*=\frac{R_{\Delta S,\Delta V}\sigma_{\Delta S}}{\sigma_{\Delta F}}} \tag{15.3.527}$$

and always take the correlation coefficient as positive (absolute value). Therefore N^* is the number of Futures or Forward contracts that whatever the strategy, will minimize the total variance.



<HELP> for explanation. CurncyCORR

1<GO> to Edit, 2<GO> for more options, <MENU> to return to correlation menu

07/15/2008 - 07/15/2009 Daily Calculation Correlation Local CCY

<Filter>

Correlation Matrix (8 Rows x 8 Columns)

Security	DXY	CCN+	SPX	TRY	BRL	AUD	JPY	EEM
11) DXY	1.000	0.387	-0.240	0.452	0.336	-0.571	0.132	-0.239
12) CCN+	0.387	1.000	-0.287	0.309	0.277	-0.451	-0.258	-0.337
13) SPX	-0.240	-0.287	1.000	-0.656	-0.578	0.684	0.577	0.915
14) TRY	0.452	0.309	-0.656	1.000	0.654	-0.584	-0.449	-0.667
15) BRL	0.336	0.277	-0.578	0.654	1.000	-0.588	-0.365	-0.602
16) AUD	-0.571	-0.451	0.684	-0.584	-0.588	1.000	0.465	0.710
17) JPY	0.132	-0.258	0.577	-0.449	-0.365	0.465	1.000	0.622
18) EEM	-0.239	-0.337	0.915	-0.667	-0.602	0.710	0.622	1.000

Color bands based on Statistical Significance: Most Significant Significant Less Significant Least Significant

Click on a security to view more options, or on a matrix value to view correlation and scatter charts.

Australia 61 2 9777 8600 Brazil 5511 3048 4500 Europe 44 20 7330 7500 Germany 49 69 9204 1210 Hong Kong 852 2977 6000
Japan 81 3 3201 8900 Singapore 65 6212 1000 U.S. 1 212 318 2000 Copyright 2009 Bloomberg Finance L.P.
SN 565082 G577-473-1 15-Jul-2009 11:07:08

Figure 15.117: **Correlation matrix for various currencies in Bloomberg™ Terminal**

However, it is customary to bring this expression such that we have not the variance of the price change, but in the change in rates especially in the purpose to make appear the quantity!!! Thus, as we have:

$$\begin{gathered}\underbrace{S}_{\substack{\text{total price of}\\ \text{raw material}}}=\underbrace{Q_S}_{\substack{\text{quantity of}\\ \text{raw material}}}\cdot\underbrace{s}_{\substack{\text{unit price of}\\ \text{raw material}}}\\ \underbrace{F}_{\substack{\text{cost price of}\\ \text{Future contracts}}}=\underbrace{Q_F}_{\substack{\text{quantity of}\\ \text{Future contracts}}}\cdot\underbrace{f}_{\substack{\text{unit price of}\\ \text{a Future contract}}}\end{gathered}\tag{15.3.528}$$

Therefore it comes:

$$\begin{gathered}\sigma_{\Delta S}=\sigma(\Delta S)=\sigma(\Delta(Q_S\cdot s))=Q_S\sigma(\Delta s)=sQ_S\sigma\left(\dfrac{\Delta S}{s}\right)\\ \sigma_{\Delta F}=\sigma(\Delta F)=\sigma(\Delta(Q_F\cdot f))=Q_F\sigma(\Delta f)=fQ_F\sigma\left(\dfrac{\Delta F}{f}\right)\end{gathered}\tag{15.3.529}$$

Hence:

$$N^*=R_{\Delta S,\Delta F}\frac{\sigma_{\Delta S}}{\sigma_{\Delta F}}=R_{\Delta S,\Delta F}\frac{sQ_S\sigma\left(\dfrac{\Delta S}{s}\right)}{fQ_F\sigma\left(\dfrac{\Delta F}{f}\right)}=\left(R_{\Delta S,\Delta F}\frac{\sigma\left(\dfrac{\Delta S}{s}\right)}{\sigma\left(\dfrac{\Delta F}{f}\right)}\right)\frac{sQ_S}{fQ_F}\tag{15.3.530}$$

That it is customary to write in the following condensed form:

$$\boxed{N^*=\beta_{\Delta S,\Delta F}\frac{sQ_S}{fQ_F}=\beta_{SF}\frac{sQ_S}{fQ_F}} \tag{15.3.531}$$

The variance V therefore becomes using this optimum (we take again the double sign ± that will be useful to us!):

$$\begin{aligned}\sigma^{*2}_{\Delta V}&=\sigma_{\Delta S}+N^{*2}\sigma^2_{\Delta F}+2N^*\sigma_{\Delta S,\Delta F}=\sigma^2_{\Delta S}+\left(\pm\frac{\sigma_{\Delta S,\Delta F}}{\sigma^2_{\Delta F}}\right)^2\sigma^2_{\Delta F}+2\left(\pm\frac{\sigma_{\Delta S,\Delta F}}{\sigma^2_{\Delta F}}\right)\sigma_{\Delta S,\Delta F}\\&=\sigma^2_{\Delta S}+\frac{\sigma^2_{\Delta S,\Delta F}}{\sigma^2_{\Delta F}}\pm 2\frac{\sigma_{\Delta S,\Delta F}}{\sigma^2_{\Delta F}}\end{aligned} \tag{15.3.532}$$

Obviously, the only one sign that seems to have a physical meaning is the "$-$" sign that will decrease the global variance (as it is the purpose of the both strategies). Then we have:

$$\begin{aligned}\sigma^{*2}_{\Delta V}&=\sigma^{*2}_{\Delta S}-\frac{\sigma^{*2}_{\Delta S,\Delta F}}{\sigma^{*2}_{\Delta F}}=\sigma^{*2}_{\Delta S}-\frac{(R_{\Delta S,\Delta F}\sigma_{\Delta S}\sigma_{\Delta F})^2}{\sigma^{*2}_{\Delta F}}=\sigma^{*2}_{\Delta S}-R^2_{\Delta S,\Delta F}\sigma^2_{\Delta S}\\&=\left(\sigma\left(\frac{\Delta S}{s}\right)sQ_S\right)^2-R^2_{\Delta S,\Delta F}\left(\sigma\left(\frac{\Delta S}{s}\right)sQ_S\right)^2=\left(\sigma\left(\frac{\Delta S}{s}\right)sQ_S\right)^2(1-R^2_{\Delta S,\Delta F})\end{aligned} \tag{15.3.533}$$

Thus Finally:

$$\boxed{\sigma^*_{\Delta V}=\sigma\left(\frac{\Delta S}{s}\right)sQ_S\sqrt{1-R^2_{\Delta S,\Delta F}}} \tag{15.3.534}$$

✍**Example:**

Let us consider that an airline company needs $10,000$ tons of fuel (which the current price is 277.- / ton) for its aircraft fleet in 3 months and wants to hedge against possible price increases of this transformed raw material and that there are no direct Future contracts directly fuel.. The airline company then wants to hedge risk by using another specific commodity for which there exists Future contracts which underlying quantity is $42,000$ tonnes (0.6903.- / ton). We would like to calculate the decrease of the fuel variance price by taking a purchase position (Call) on this Futures (so we do a down bet for the replicating commodity) knowing that the variance at 3 months of fuel is estimated to be 21.17% and that of the replicating commodity being of 18.59% over obviously the same period and that their correlation is in absolute value of equal to 0.8242.

Then we have:

$$N^{*}=\left(R_{\Delta S,\Delta F}\frac{\sigma\left(\dfrac{\Delta S}{s}\right)}{\sigma\left(\dfrac{\Delta F}{f}\right)}\right)\frac{sQ_S}{fQ_F}=\left(0.8243\frac{0.2117}{0.1859}\right)\frac{277\cdot 10,000}{0.6902\cdot 42,000}\cong 89.7\cong 90 \tag{15.3.535}$$

As we have for the volatility of the fuel in money:

$$\sigma_{\Delta S}=sQ_S\sigma\left(\frac{\Delta S}{s}\right)=277\cdot 10,000\cdot 0.2117=586,409.- \tag{15.3.536}$$

And once covered by the replicating commodity, it becomes:

$$\sigma_{\Delta V}=\sigma\left(\frac{\Delta S}{s}\right)sQ_s\sqrt{1-R_{\Delta S,\Delta F}^2}=0.2117\cdot 277\cdot 10,000\sqrt{1-0.8243^2}\cong 331,997.- \tag{15.3.537}$$

The hedging by Future contracts reduce the volatility of cash of a factor around 43.38%.

Remark

Some practitioners use the following empirical indicator named "Ederington efficiency" as hedging quality:

$$E=\frac{\sigma_{\Delta S}^2-\sigma_{\Delta V}^{*2}}{\sigma_{\Delta S}} \tag{15.3.538}$$

Which is nothing more than the ratio of the difference of the non-hedged variance with the hedged one by the non-hedged variance. In the example above its value is 67.49%. More this ratio is close to unity, the better it is!

When the relation:

$$N^{*}=\beta_{\Delta S,\Delta F}\frac{sQ_S}{fQ_F}=\beta_{SF}\frac{sQ_S}{fQ_F} \tag{15.3.539}$$

is applied not with a numerator that corresponds to a raw material but to stock portfolio, this technique is named "stock portfolio hedging by replication" or more simply "equity portfolio hedging" and is then rather denoted as follows:

$$N^{*}=\beta_{AB}\frac{V_A}{V_B} \tag{15.3.540}$$

where V stands for "value" (absolutely no relation with the notation of the variance!!!). Or even in the following form (P being the value of the portfolio and F that of Futures):

$$N^{*}=\beta\frac{P}{F} \tag{15.3.541}$$

✍**Example:**

A trader has a portfolio of $2,000,000$.- of IBM stocks. He would like to hedge this portfolio thinking that the market will go down on the S&P 500. Therefore he sells ("short") Futures for an amount of $225,000$.- and he know that the beta correlation between IBM and S&P 500 is equal to 1.1 (a beta less than 1 indicates that the investment is less volatile than the market, while a beta more than 1 indicates that the investment is more volatile than the market). The number of Futures to sale is then (therefore if the portfolio value goes down with the same magnitude as the S&P 500 the trader will lose nor gain anything):

$$N^*=\beta_{AB}\frac{V_A}{V_B}=1.1\frac{2,000,000}{225,000}=9.78\cong 10 \tag{15.3.542}$$

If the trader has purchased Futures to protect himself from the decrease he expected to happen (bet down price and therefore "short" strategy for recall...) then if IBM shares lost 10%, but the index S&P 500 index lost more than what the beta indicated such for example 15%, so even if the IBM portfolio had lost $200,000$.- of its value, the fact that the S&P 500 has lost 15%, thanks to the Futures, the trader will have made an indirect gain of:

$$225,000\cdot 15\%\cdot 10=337,500\text{.-} \tag{15.3.543}$$

So ultimately a gain for the trader of $137,500$.- (like what... sometimes... bet down is good).

Let us recall now that during our previous study of modified duration (see page 7633) we have define that latter, denoted as D^* as being:

$$\frac{\mathrm{d}P}{P}-D^*\mathrm{d}y \tag{15.3.544}$$

Hence:

$$\mathrm{d}P=-D^*P\mathrm{d}y \tag{15.3.545}$$

That can also obviously be written:

$$\Delta P=-D^*P\Delta y \tag{15.3.546}$$

Similarly let us then write:

$$\Delta S=(-D_S^*S)\Delta y \qquad \Delta F=(-D_F^*F)\Delta y \tag{15.3.547}$$

where D_S^* and D_F^* are the modified duration of S and F respectively.

So we can rewrite:

$$\Delta V=\Delta S\pm N\Delta F=(-D_S^*S)\Delta y+N(-D_F^*F)\Delta y=-[(D_S^*S)+N(D_F^*F)]\Delta y \tag{15.3.548}$$

Which is zero when the net exposure, represented by the term between bracket is zero, that is:

$$(D_S^*S)+N(D_F^*F)=0 \tag{15.3.549}$$

Hence:

$$N^*=-\frac{D_S^*S}{D_F^*F} \tag{15.3.550}$$

In other words, the optimal hedge ratio is simply minus the ratio of the dollar duration of cash relative to the dollar duration of the hedge.

More generally, we can use N as a tool to modify the total duration of the portfolio. If we have a target duration of D_V^*, this can be achieved simple by writing:

$$\Delta V=-D_V^*V\Delta y=-[(D_S^*S)+N(D_F^*F)]\Delta y \tag{15.3.551}$$

Therefore after rearranging:

$$N=\frac{D_V^*V-D_S^*S}{D_F^*F} \tag{15.3.552}$$

We can see that the previous relation with N^* is just a special case of that latter relation!

Remark

We will see further below on page 7757 how to hedge a portfolio with Commodities and Futures but based on the CAPM (Capital Asset Pricing Model).

15.3.10.3.4.4 Options

Options are "contingent asset" (also named "contingent claim[22]"), that is to say, a particular form of a title (contract) giving the holder in counterpart of payment the right, but not the obligation to buy or sell a given amount of a financial asset (stock, bond or foreign exchange), at a given date or up to a given date (expiry or maturity) to a price fixed in advance.

It is primarily as Futures and Forwards (also derivatives!) that options permits to hedge the risk of market fluctuations and to speculate. For example, if Boeing sells aircraft in euros but produces in the dollar zone. The sale price is fixed today, but the sale is made on delivery! Boeing must therefore protect itself against the risk of the exchange rate (which can sometimes be of a magnitude 100% in a few years). Generally, most big companies protect themselves from these risks by purchasing in banks derivatives such as options.

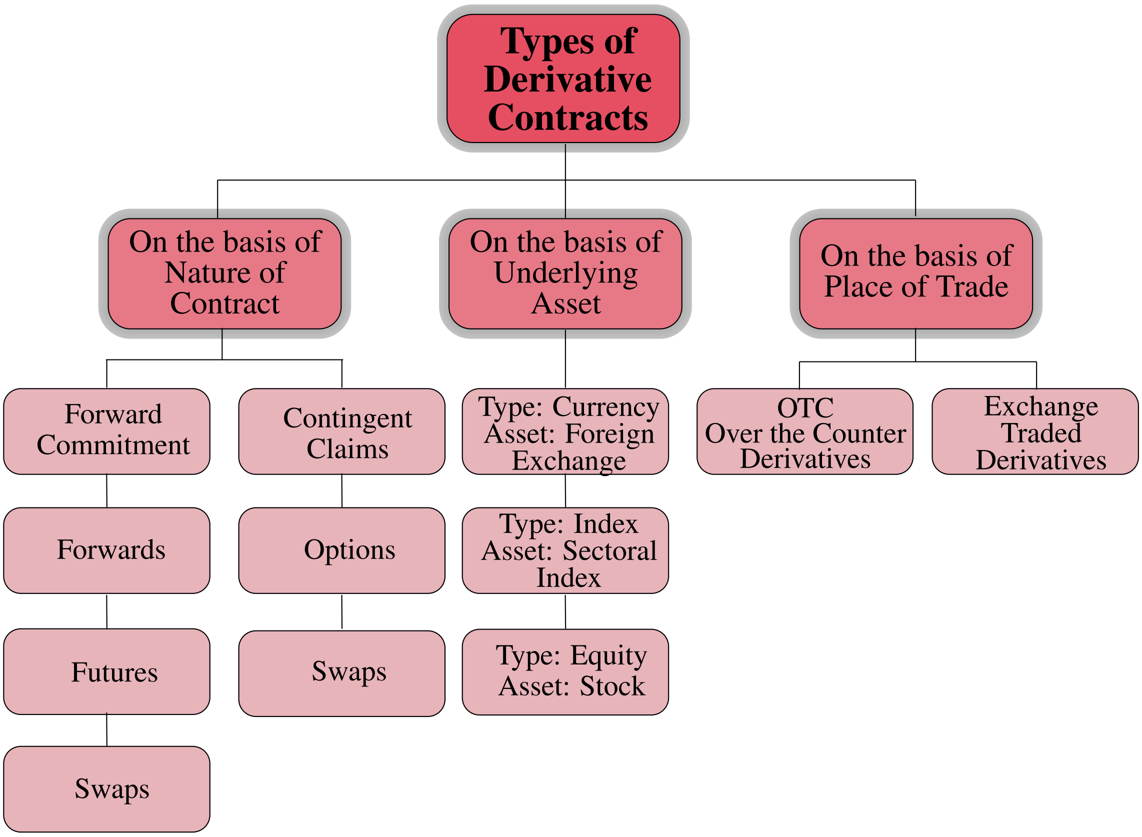

Types of Derivative Contracts
On the basis of Nature of Contract
On the basis of Underlying Asset
On the basis of Place of Trade
Forward Commitment
Contingent Claims
Type: Currency Asset: Foreign Exchange
OTC Over the Counter Derivatives
Exchange Traded Derivatives
Forwards
Options
Type: Index Asset: Sectoral Index
Futures
Swaps
Type: Equity Asset: Stock
Swaps

Figure 15.118: **Derivatives contracts viewed on various basis**

In 12003 (holocene calendar) there would have been $5,210$ million of options traded on the markets and respectively 0.98% for raw materials, 0.28% for exchange rates, 5.80% for the interest rates and 92.94% in securities and bonds.

---

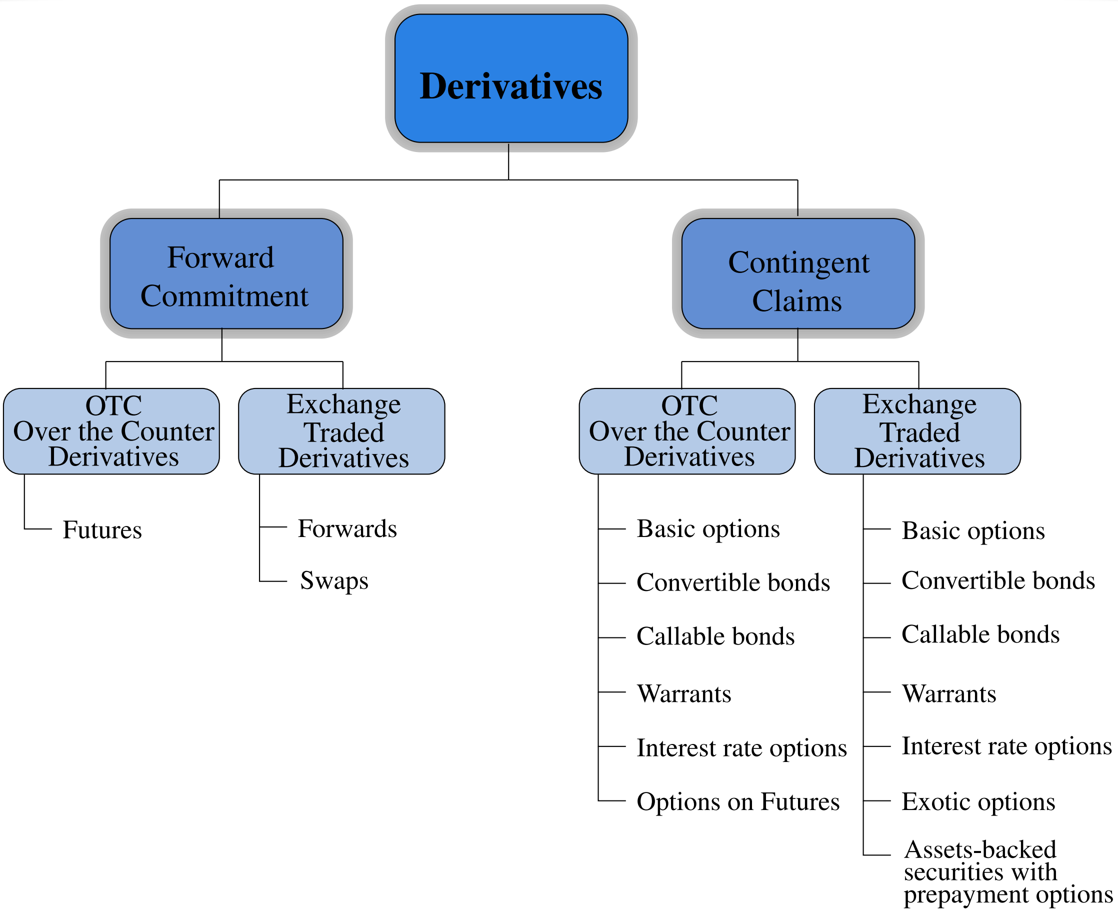
[22] In finance, a "contingent claim" is a derivative whose future payoff depends on the value of another underlying asset, or more generally, that is dependent on the realization of some uncertain future event. These are so named, since there is only a payoff under certain contingencies. Any derivative instrument that is not a contingent claim is named a "forward commitment".

Remarks

R1. We will return further below in detail on the mathematics of the options that are important products as also many other derivatives relies on the same mathematical properties.

R2. Options involve zero-sum game (see section Game Theory page 7378) in the sense that for every seller there is a buyer and what one wins, the other lose it (and vice versa).

R3. The main difference between options and Futures lies in the fact that options represent a right to buy or sell at the contract maturity while Future represent the obligation to exercise the contract at maturity.



Derivatives
Forward Commitment
Contingent Claims
OTC Over the Counter Derivatives
Exchange Traded Derivatives
OTC Over the Counter Derivatives
Exchange Traded Derivatives
Futures
Forwards
Swaps
Basic options
Convertible bonds
Callable bonds
Warrants
Interest rate options
Options on Futures
Basic options
Convertible bonds
Callable bonds
Warrants
Interest rate options
Exotic options
Assets-backed securities with prepayment options

Figure 15.119: **Forward commitment vs Contingent claims**

Definitions (#15.3.690):

D1. An "option" is a derivative product that gives the right, but not the obligation, to buy ("Call option", also just "Call") or sell ("Put option", also just "Put") a specified amount of an underlying asset S (share, bond, stock index, currency, row material, another derivative product, etc.) to a predetermined price named "strike price" denoted by K or "exercise price" and therefore denoted by E and during (until to) a given time named "maturity" denoted by T in exchange of a "risk premium" depending (denoted by C or P for the Call to Put) on the intrinsic value at maturity of the option named "flow" or more often "terminal payoff" (and by some practitioners: "stochastic target"). The determination of the premium with mathematical models is what we name the "pricing of the option". An option is therefore a kind of "insurance policy" to protect against variations in costs by transferring the risk to a person willing to sell the option in exchange of a risk premium ... this latter being also often named "insurance premium".

D2. We talk about "spot price" or just "spot" to designate the current market price S of the underlying asset at the time of an immediate transaction of the option (Call or Put). If the underlying is a foreign currency exchange rate, then we speak of "cross over" or simply "cross".

D3. If the intrinsic value[23] of an option according to some hypothesis) of an option is positive with respect to the spot price, it is said to be "in-the-money". In the case of the purchase of a Call for example, it means that the exercise price of the Call is below the spot price. It is possible therefore to buy cheaper than the traded value at the execution date (maturity most of time) of the option (which is a date between the option issue date and maturity date of the option for recall!!!).

D4. If the intrinsic value of an option is not advantageous compared to the spot price, it is said to be "out-the-money". In this case, the exercise price is above the current spot price for a Call for example. It would therefore be unwise to exercise the Call at the due date (either before), because that would be more expensive to apply the Call than to buy at the spot price on that date!

D5. If the intrinsic value at a given date of the underlying is equal to its current spot at maturity, the intrinsic value is zero and the value of the option is said to be "at-the-money" and abbreviated in many softwares obviously: ATM (we will see later the mathematical implications of this).

D6. If the intrinsic value of an option is very far form the stock price it is then referred as "deep in-the-money" or "deep out-the-money".

Given the above definitions and the example about Boeing the options can be naively summarized by the following figure:

---

[23] An option's premium is the sum of "intrinsic value" and "extrinsic value" (also known as "time value"). Intrinsic value is reflective of the actual value of the strike price versus the current market price. Extrinsic value is made up of time until expiration, implied volatility, dividends and interest rate risks.

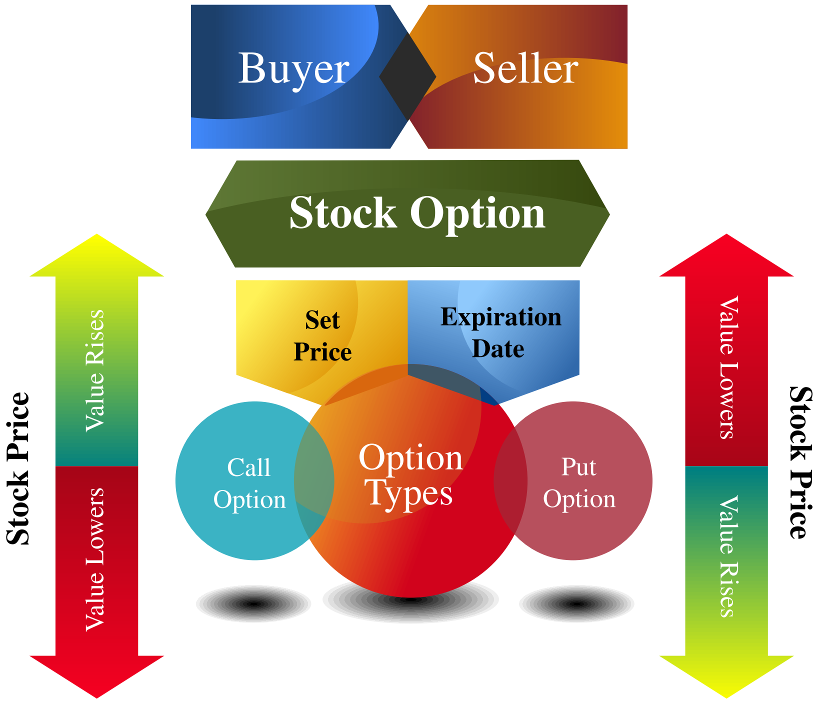


Figure 15.120: **Options graphical concept summary**

Remarks

R1. The usefulness of the existence of Options are the same as for the Futures and Forwards: they can be viewed as an artificial way to create assets (in this case: derivatives) to increase financial volatility (standard deviation or "loss / gain deviation") of market and thus guarantee his equilibrium.

R2. The holder or buyer of an option contract is said to be in a "long position" while its counterpart, the issuer or seller of the contract is said to be in a "short position".

R3. If the option can be exercised at any time prior to maturity, we talk about "American option", if the option can be exercised at maturity, we talk about "European option". One unexercised option is considered as "abandoned" (lost). These two families of options in their Call/Put version are in practice grouped under the term "plain vanilla options" as these are the most common (such as ice cream with vanilla... that are anything but exotic...).

Alongside the classic options (vanilla options), like:

- "Stock options" are options on individual stocks, also named "equities options". They are among the most popular type of options.
- "Exchange-listed options" are available widely on traded stocks and an option on any stock can potentially be created on the over-the-counter market.
- "Index options" on stock market indices (S&P 500, Down Jones, NASDAQ, etc.) are well known, not only in the investment community but also among many individuals who are not even directly investing in the market.

- "Bond options" are found almost exclusively in the over-the-counter market and are almost always options on government bonds.
- "Interest rate options" are options on interest rates like the LIBOR for example.
- "Currency options" in the same idea of interest rate options allows the holder to buy (if a call) or sell (if a put), an underlying currency at a fixed exercise rate, expressed as an exchange rate.
- "Future options" gives the holder the right to enter (call) or get out (put) into/from a long/short futures contract at a fixed future price.
- "Commodity options" are options on which the underlying asset is a commodity, such as oil, gold, wheat, or soybeans. There are exchange-trader as well over-the-counter versions.

appear since the years 11990s (holocene calendar), on the markets, options named "exotic options" sometimes characterized by the name of the place where they were created and the manner of determining the exercise price at maturity (so there exists at most as many mathematical models as there are types of options...) as the:

- "Asian options" whose exercise price is based on the average of the underlying price during the life of the option
- "Parisian options" that can be activated or cancelled if the underlying traded price remains above a given threshold value during a given time range
- "Russian options" which are perpetual American options that guarantee to get the maximum margin observed between the issue and the exercise date
- "Binary/Digital options" that gives the right to the amount fixed in advance if the underlying at maturity exceeds the exercise price
- "Lookback options" that give the right at maturity to the difference between the value of the traded underlying at maturity and the value of the min or max during the life of the option
- "Quantile options" that are a refinement of lookback options because the rule of the game is based on a quantile, not on the max or min
- "Barrier options (knock out, knock in)" whose exercise is permitted if the traded value of the underlying cross or doesn't cross a certain threshold
- "Quanto options" that are options on foreign underlying, but paid in local currency (then you should take into account the exchange rate)
- "Ratchet options" that gives the buyer the right to block its gains realized on the underlying during determined intervals during the life of the option
- "Compound options" or "chooser options" that are options on options, etc.

Remark

In the general case (more advanced models, other types of options), the price of an option can not be obtained from an analytical closed formula. Then we need to make use of numerical methods to price the option. Common numerical methods to price are: Binomial method (see further below), Monte Carlo method (see the `R` companion book), Finite difference method, Radial basis function method, Finite element method, Fourier method, Least squares calibration method, etc.

Here is a small summary of all derivatives that we have mentioned so far (maybe be useful before we continue):

FORWARD:

- Contracts executed by banks
- Private contract (unregulated)
- Not standardized contract
- High counterparty risk
- Unlimited profit for a buyer with limited loss, unlimited loss for a seller with limited profit
- No initial payment required
- Normally one specified delivery date
- Settle and the end of maturity
- Traded 24 hours a day
- No cash exchange prior to maturity
- More than 90% of contract settled by actual delivery of assets
- Delivery or final cash settlement usually takes place

FUTURES:

- Contracts executed by brokerage houses
- Traded on organized exchanges
- Standardized contract (regulated)
- Low counterparty shared risk
- Unlimited profit for a buyer with limited loss, unlimited loss for a seller with limited profit
- Initial margin payment required
- Range of delivery date
- Daily settled
- Traded $4-8$ hours a day
- Profit/Loss are paid in cash
- Not more than 5% of contract settled by actual delivery of assets
- Contract normally closed out prior to the delivery
- Contract seller has obligation to sell/buy if buyer exercise right
- Preferred by speculators or arbitrageurs

OPTIONS:

- Contracts executed by brokerage houses or private ("dealer options")
- Traded on organized exchanges or private
- Standardized or non-standardized contract
- High counterparty risk borne by one party only
- Unlimited profit for a buyer with limited loss, unlimited loss for a seller with limited profit
- Initial margin payment required (but less than Futures)
- Execution date fixed or variable
- Settled maturity of range of date
- Traded 24 hours a day
- Profit/Loss are paid in cash
- Contract normally closed at maturity
- Contract seller has obligation to sell/buy if buyer exercise right
- Preferred by hedgers

Here is a summary table between Call and Put:

Call			Put	
Buyer of a Call (Long Call)	Seller of a Call (Short Call)		Buyer of a Put (Long Put)	Seller of a Put (Short Put)
Has the right, but not the obligation, to buy the underlying stock at the predetermined price until the due date.	Has the obligation to sell the underlying asset at the predetermined price if the Call is exercised the due date.		Has the right, but not the obligation, to buy the underlying stock at the predetermined price until the due date.	Has the obligation to buy the underlying stock at the predetermined price if the Put is exercised

Table 15.48: **Differences between Put and Call**

So there is obviously an enormous important mathematical difference between Options and Stocks/Bonds. Indeed, Options have an exercise date, their price dynamic can be statistically predictable and this prediction is even better when we are close to their date of exercise and this is not applicable for the Stocks/Bonds because we never know at the strategic level when they will be sold or respectively purchased.

While the naive theoretical models for valuation of options (such as Black & Scholes) assume a constant historical volatility whatever the price of the underlying. In reality the graphical representation of the price of an Option based on their underlying price is not a horizontal line but a curve that is usually named a "volatility smile". We will come back later in detail on this concept for which we have an illustration below:

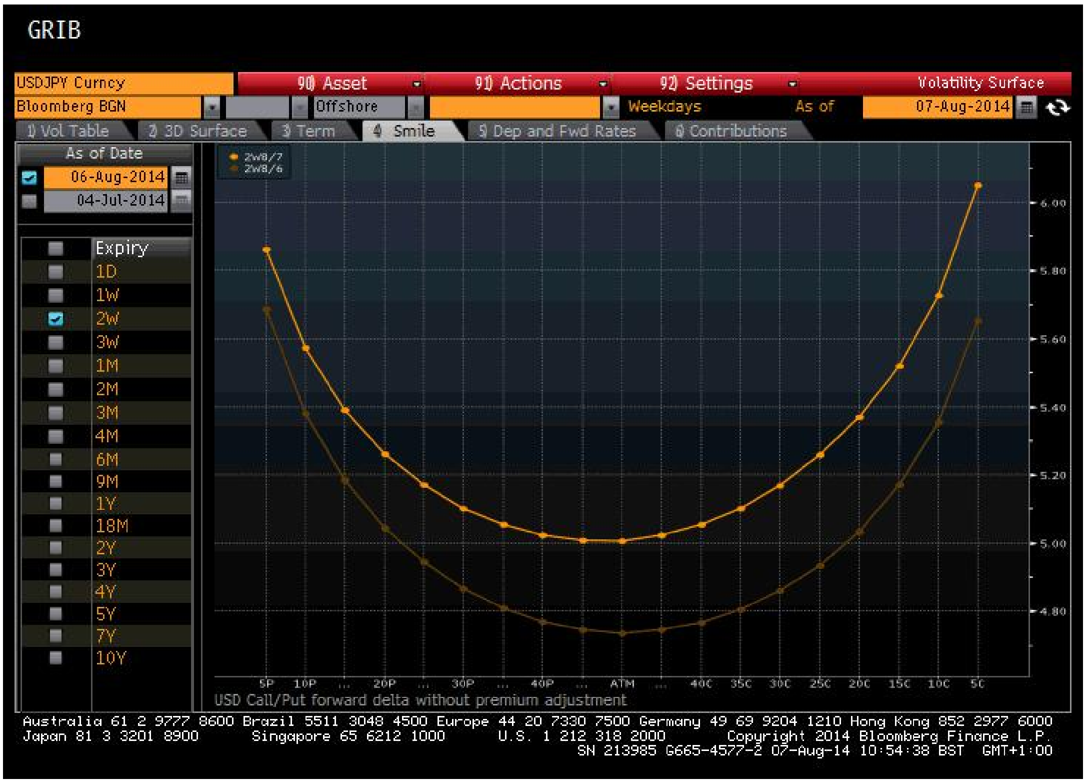

GRIB
USDJPY Curncy
90) Asset
91) Actions
92) Settings
Volatility Surface
Bloomberg BGN
Offshore
Weekdays
As of
07-Aug-2014
1) Vol Table
2) 3D Surface
3) Term
4) Smile
5) Dep and Fwd Rates
6) Contributions
As of Date
☑ 06-Aug-2014
☐ 04-Jul-2014
Expiry
1D
1W
2W
3W
1M
2M
3M
4M
6M
9M
1Y
18M
2Y
3Y
4Y
5Y
7Y
10Y
2W8/7
2W8/6
6.00
5.80
5.60
5.40
5.20
5.00
4.80
5P 10P ... 20P ... 30P ... 40P ... ATM ... 40C 35C 30C 25C 20C 15C 10C 5C
USD Call/Put forward delta without premium adjustment
Australia 61 2 9777 8600 Brazil 5511 3048 4500 Europe 44 20 7330 7500 Germany 49 69 9204 1210 Hong Kong 852 2977 6000
Japan 81 3 3201 8900 Singapore 65 6212 1000 U.S. 1 212 318 2000 Copyright 2014 Bloomberg Finance L.P.
SN 213985 G665-4577-2 07-Aug-14 10:54:38 BST GMT+1:00

Figure 15.121: **Typical volatility smile**

Remark

As we will prove why further the Options that are said to be "non-path-dependent options" (NPDO) have their price which depends (among others...) only on the final price at maturity of the underlying K, while others Options that are said to be "path-dependent options" (PDO) have their price which depends on all the values of the underlying during the term of the Option contract.

Let us formalize things now! But without going into too much detail at first time (we keep the details for the study of the Black & Scholes model later that consists to determine the premium price[24] of an option according to some hypothesis.). We consider first for simplicity options on a single underlying not paying dividends.

We will denote the price of the underlying asset of the concerned option by S_t (trade/exchange rate value) at time t and maturity T and will disregard the difference between continental Put and Calls (American and European).

Let us imagine a Call, which gives the holder the right (but not the obligation) to buy the underlying asset at any time between today $t=0$ and the $t=T$ at the (maturity) exercise price K set in advance. Let us take the practical common case of a Call option that protects a business against the rise of the euro/dollar exchange rate for example. Acquired today by the firm, it will therefore give it the right (but not the obligation!) to buy \$1 in exchange of K euros (the strike price at maturity K is a fixed characteristic of the contract) to the future date T (maturity date).

If the exchange rate has for value S_t at time t (that is to say today $\$1=S_t$ EUR), this insurance comes from the perspective of the firm to save an amount (a payoff Call from the perspective of buyer):

$$\max(S_T-K,0) \tag{15.3.553}$$

also denoted by practitioners:

$$(S_T-K)^+ \tag{15.3.554}$$

in euros at maturity T (the traded price/rate being a maturity denoted by S_T obviously...). At any time, two major cases occur once for our Call buyer:

1. $S_t<K$: in this case, the Call gives the right to buy the underlying at the price K that we could buy cheaper on the market. In the case of our Call for the dollar/euro it would be better for the buyer to apply the exchange market rates rather than exercise him Call since otherwise he will have less euros for one dollar (that's why a Call has a zero intrinsic value in this situation at maturity). But we will lose the risk premium we paid to purchase Call and this risk premium has also to be taken into account before not exercising the Call.

   The reader most not forget that for European Options we only exercise at maturity. Therefore we don't write the condition $S_t<K$ but instead $S_T<K$.

---

[24] Remember that an option's premium is the sum of "intrinsic value" and "extrinsic value" (also known as "time value"). Intrinsic value is reflective of the actual value of the strike price versus the current market price. Extrinsic value is made up of time until expiration, implied volatility, dividends and interest rate risks.

2. $S_t > K$: in this situation, the Call gives us the possibility to purchase the underlying cheaper than the market price/rate. We probably will exercise that opportunity (the profit being the difference between the two prices plus we have to take into consideration the paid risk premium). In the case of our Call for the dollar/euro exchange we will exercise our right at the lower rates guaranteed by the Option contract ($\$1 = K$ EUR).

From the perspective of the counterparty (seller of the Call), in the situation (1) it does not pay anything to the buyer, and all is forgotten (the contract expires, any contractual relationship between the two parties disappear). In case (2), the vendor is assigned, it must sell to his counterparty the underlying at price K. If it does not hold this underlying, it must first buy it at a most expensive price on market (at the price $S_t > K$). However, in both cases the counterparty (seller of the Call) has get the cash of the risk premium per unit of Call (cash that will perhaps avoid it to loose to much money).

Thus, in the first case, the buyer and seller do not pay anything. In the second case, it is as if the buyer of the Call buy the underlying on the market and received at the same time the amount $S_t - K$ (for the seller it is obviously the reverse). So with these derivatives is the seller of the Call (or Put) which assumes almost all the market risk and of course the interest is quite high to neutralize this risk by using a mathematical formalism to determine the risk premium price (using typically the Black & Scholes model).

Let us see now an example of the investment perspective (risk-taking is evident in this example):

✍**Example:**

Imagine the case of a share which is currently worth $1,000$.- (regardless of the currency) and it is expected to increase by 12% in a year (typically what people expect on stock options plans).

Let us imagine also that an investor has the alternative of buying the stock at $1,000$.- or buy the corresponding Call option at exercise (strike) price of $1,000$.- (therefore assumed to be equal to the actual price of the share, which is not necessarily always the case in practice obviously...) for a premium of 40.- (remember that we will see later how to calculate risk premiums). Obviously, the investor can then for $1,000$.- buy 25 Call options rather than a single stock option if he wants but after he will have to find the money to but the 25 underlyings... (or found someone else that will buy them).

The question is to find the most interesting investment: Thus, an increase of 120.- on one year in the case of buying a share represents a return on investment of 12% per year, while the purchase of a Call option will have a return on investment 80.- (120.- gains on the sale price less the premium paid of 40.-) hence 200% of return!!!

It is clear from this example that the purchase of a Call with a same investment has a profitability significantly higher than the purchase of the underlying at least since the option risk premium does not exceed a certain threshold.

Now lets discuss in detail and with an example another major concept that we have already implicitly presented in the preceding paragraphs and that requires our attention because it is often mentioned by analysts! This is the "leverage effect" of Options that is a double-edged weapon or a portfolio mass destruction weapon (if leverage is not calculated in advance and even approximately)!!!

When we talk about the options, we often retain only the right to buy or sell a financial instrument (at a fixed price and during a given time), neglecting the obligation by the seller of the option. However, the leverage that characterizes these financial instruments can make this a devastating obligation for the seller.

Ok let us see more in detail what we are speaking about:

- First with the case of a Call:

  The buyer of a Call on a share (for example) limit its risk to the option risk premium price for unlimited earning potential. The seller of a Call is in exactly the opposite position: he is collecting the option risk premium price amount, but takes unlimited risks.

  As theoretical introduction... let us take the example of a share traded 350.- at the mid-October. An investor bets on the rising of the share and buys 12.50.- a Call option on it with a risk premium with a maturity to next year in January at the exercise price of 380.-. A graphical representation allows to easily show the relation between the evolution of the share price (on the x-axis) and its effect on the buyer or seller of the Call option.

  - Let us consider first the buyer of the Call:

    As long as the share price remains below 380.- ("leverage value"), exercise price, the buyer of the Call will have no interest in exercising his option, which is said as we already know "out of the money" (OTM). By cons, if the share price increased and exceeds the exercise price, the option is said to be as we also know "in the money" (ITM) and it becomes interesting to exercise the option. When the option exercise price is equal to the underlying price of the stock exchange, we say that the option is "at the money" (ATM) as we also already know. As soon as the share price exceeds 392.50.-, hence the sum of the exercise price and the option risk premium at the mid-October ($380+12.50$), the holder of the Call begins to earn money on his initial investment. If the share price rises suddenly to 500.-, an increase of just over 30%, the gain will be much more than proportional: for 12.50.- invested, the purchaser will realize a profit of 107.50.- a gain of 860%: this is the famous "leverage effect".

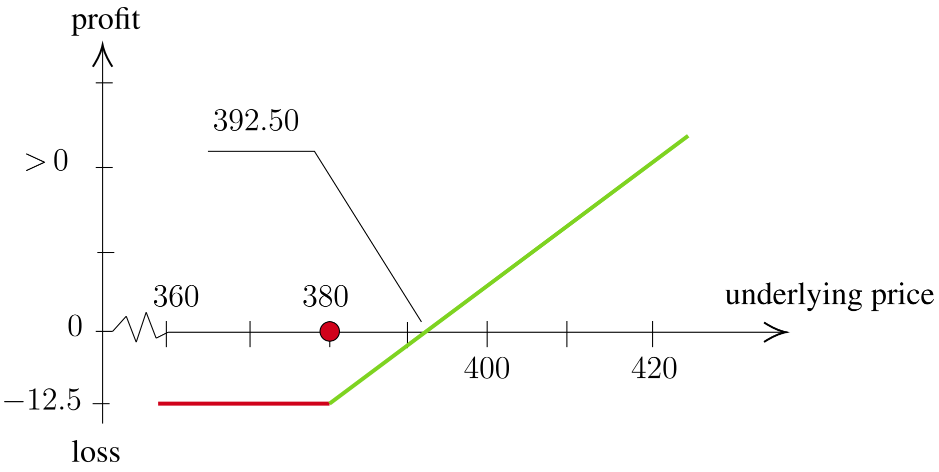


Figure 15.122: **Leverage effect for the buyer of a Call (simple Call payoff diagram)**

We can see above that the payoff of a Call from the buyer's point of view is a convex function (see section Functional Analysis page 2280).

- Let us consider now the seller of the Call:

  As the long as the share remains below 380.- ("leverage value" for recall...), the Call seller makes a profit of 12.50.- representing the option risk premium. Starting from 380.-, the seller may be obliged to deliver the share at the strike price, that is to say 380.-. Starting from 392.50.-, it starts to lose money on the transaction, since the share that he will undoubtedly have to deliver will worth be more expensive than the sum of the initial exercise price and of the option risk premium received. If unfortunately for him the share actually goes up to 500.- and he does not have the underlying, he will have to go buy it on stock market to honour the exercise request from the holder of the Call, loosing at the same time 107.50.- on the deal, thus more than eight times the premium received!!!!! With the leverage effect, it is possible to lose more than the risk premium received. We then say that we are experiencing a "margin call" and we better understand the importance to find an approximate martingale to evaluate the risk premium at its best.

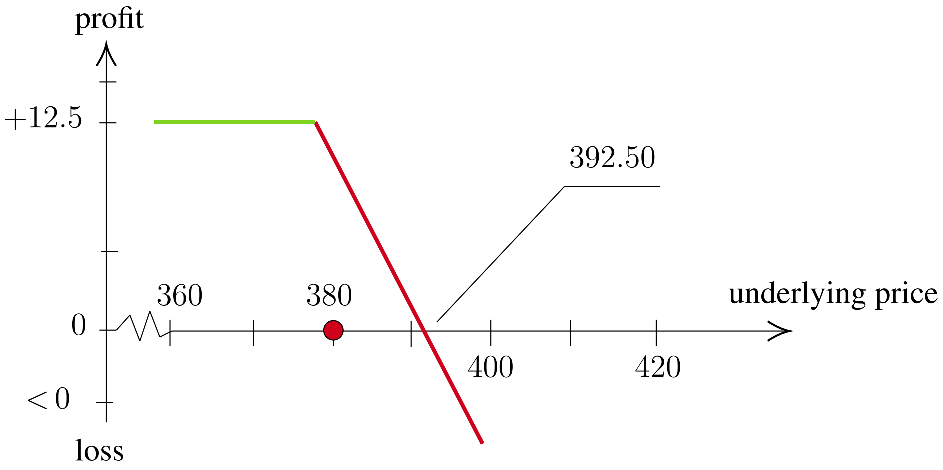


Figure 15.123: Leverage effect for the seller of a Call (simple Call payoff diagram)

We can see above that the payoff of a Call from the seller's point of view is a concave function (see section Functional Analysis page 2280).

Thus, if the underlying price increases of 3.29%, the seller of the Call loose 100%. We speak then of a leverage effect of 3.29 for 100 or more simply of a leverage corresponding to a ratio $100/3.29$ thus a leverage of about 30.39 (it is the same of course for the gain!).

- Second (and last) with the case of a Put:

  - Let us consider the buyer of the Put: The buyer of a Put limits its exposure to that of the cost of the option premium for a potential gain far more important. In front of him, the seller of the Put is in exactly the opposite position: he takes the option premium but takes a much greater risk. If we take the same share X traded at 350.- at mid-October, we are now with an investor that bets on the fall of the share price. So he buys for 49.50.- ("premium" price) the Put of maturity in December at an exercise price of 390.-.

The buyer of the Put begins to make a profit if the share price falls below 340.50.-, that is to say the strike price minus the option price $(390 - 49.50)$. Between 340.50.- and 390.- the option is not profitable but permits to reduce the loss. Above the at-the-money price (390.-) the exercise of the Put really offers no interest anymore and we say when then as we know that the Put option is out-the-money.

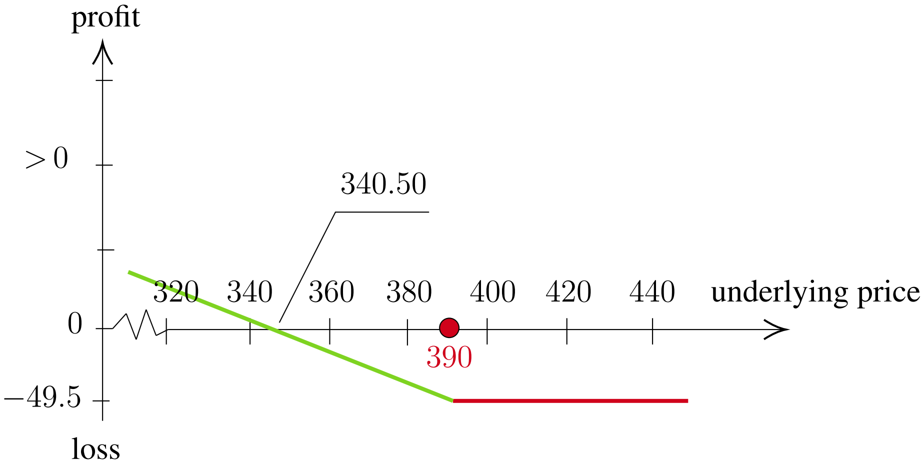


Figure 15.124: **Leverage effect for the buyer of a Put (simple Call payoff diagram)**

- Let us consider now the seller of the Put:

  The seller of the Put first get the cash of option premium that is to say 49.50.-. As long as the traded price remains above 390.- he is a winner. If the traded price of the option is between 340.50.- and 390.- and it loses some of its risk premium cash but he is still winning. Below 340.50.- the seller of the Put will be forced upon maturity to pay 390.- to the buyer of the Put (by selling the underlying and by paying the difference in one way or another). Obviously if the underlying price falls to zero, the seller of the Put may lose up 340.50.- of capital.

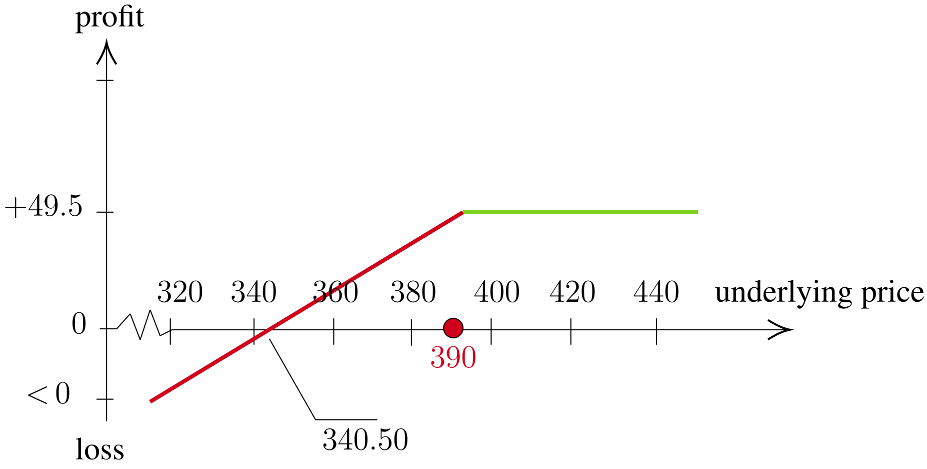
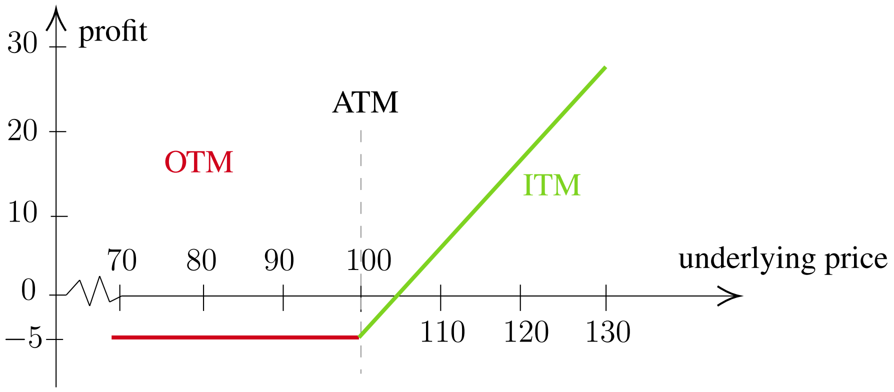
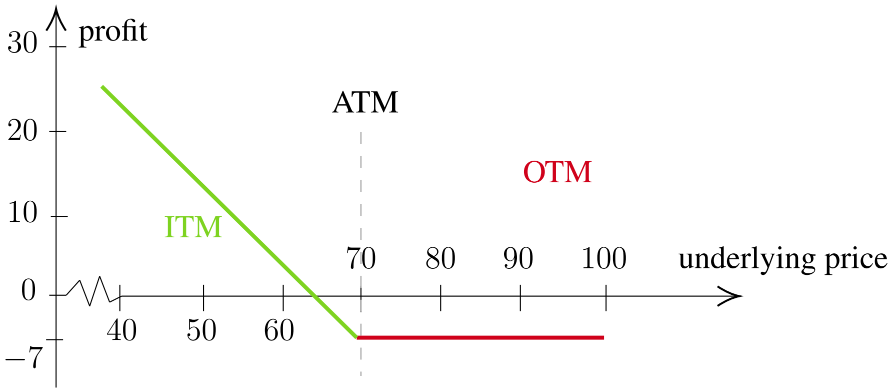


Figure 15.125: **Leverage effect for the seller of a Put (simple Call payoff diagram)**

At the vocabulary level, here below is an excellent summary diagram (top buyer of a Call, down buyer of a Put) knowing that the "moneyness" is defined as the ratio between the price of the underlying S and the strike price K:





Moneyness	Call	Put
< 0.94	deep ITM	deep OTM
$0.94 - 0.97$	ITM	OTM
$0.97 - 1.00$	ATM	ATM
$1.00 - 1.03$	ATM	ATM
$1.03 - 1.06$	OTM	ITM
> 1.06	deep OTM	deep ITM

Table 15.49: **Basic vocabulary of a situation of option payoff (source:Fast Calibration in the Heston Model, BAUER Rudolf, p. 14)**

If we think for a moment, a possible strategy is to buy Put and Call of the same underlying with the same strike price and the same expiration date (maturity). This type of strategy is named a "straddle" or "straddle purchase" or "bottom straddle". We will study options strategies mixtures much further below in details.

Thus, if the traded price of the underlying is near the strike price K at the expiration date T then the straddle generates a net loss (as we will not use either the Put or Call and we will have lost the amount equivalent to the options premium). However, if the underlying price is quite far from the strike at maturity the profit can be very important!

This is pure speculation of investment banks derivatives strategies and this should be objectively prohibited... but let see anyway the principle through an example...

✍**Example:**

Let us consider an underlying that is traded actually 69.- will shift violently within three months. We can set up a straddle strategy by buying both a Call and a Put with strike of 70.- and maturity in 3 months. Let us suppose that the Call cost 4.- and the Put 3.-.

If at maturity, the underlying value is that of strike K, for the speculator it is the worst situation as it will have lost the sum of the two risk premiums, that is to say 7.- because he did not have any interest in exercising the Call using its Put (since there is no difference between the both).

If the underlying traded value remains at 69.- we lost 6.- because the Call has no value anymore at maturity (at least as it makes lose money it would be unwise for a pure speculator to exercise it) and the Put is only worth 1.- (as we will be able to sell at 70.- something that has a value of $69.-$ and it will make a small gain of 1.-). Therefore:

$$(-4+(-3))+1=-7+1=-6.- \tag{15.3.555}$$

If the price of the underlying rises to 90.- (or down to 50.-), we will realize a profit of 13.- thanks to the Put as:

$$(-4-3)+(90-70)=13.- \tag{15.3.556}$$

It is hoped that the speculator will bet correctly on a strong variation ignored by most of the market players, otherwise the option risk premiums will be obviously high enough that his gain will be anyway equal to zero! Graphically ... the Straddle strategy is often represented as follows (the vertical axis is the gain, the horizontal axis is the underlying price):

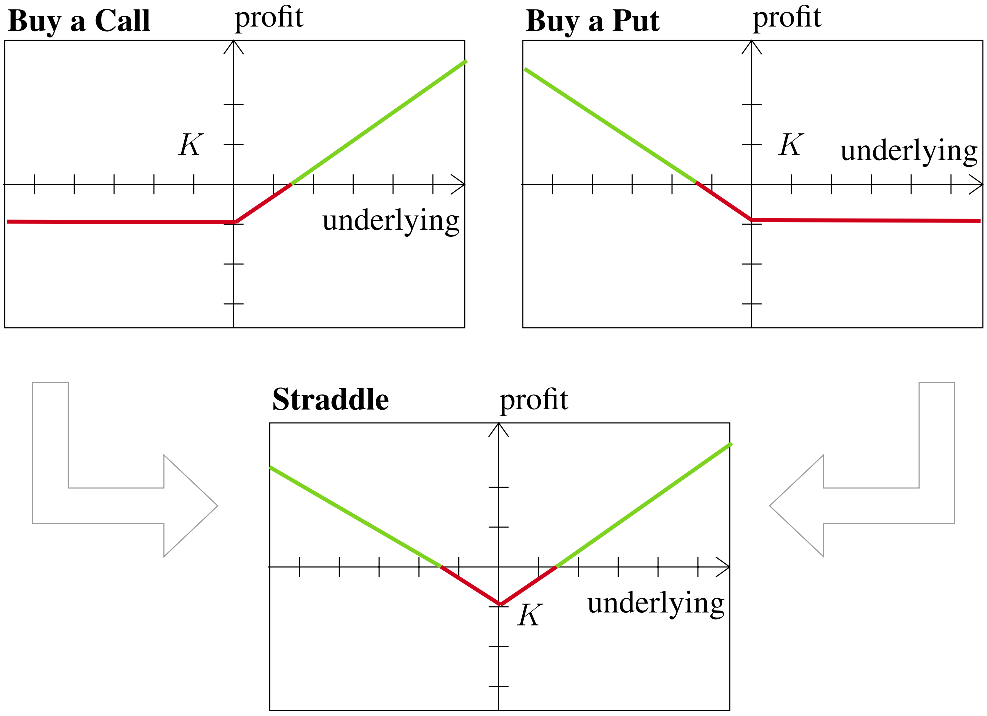


Figure 15.126: **Payoff diagram of a straddle strategy**

We can summarize these properties under the form of the following table:

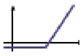
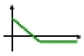
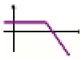
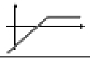
Strategy	Trend anticipation	Potential Gain	Potential Loss	Profit curve
Buy a Call	upward trend (bullish trend)	"Illimited"	Limited	
Buy a Put	falling trend (bearish trend)	Limited	Limited	
Sell a Call	Stable or weakly bearish	Limited	"Illimited"	
Sell a Put	Stable or weakly bullish	Limited	Limited	

Table 15.50: **Different elementary Call-Put strategies**

The "option plans", better known as the "stock option plans" or "employees stock option plan (ESOP)" are warrants (or Call Options) issued by packages (nominatives) for good employees of a company and have for purpose to strengthen the partnership of development between the same company and its employees. Thus, the latter on the acquisition of securities will be full-fledged stockholders receiving dividends and may participate in stockholders meetings. What is supposed to increase the motivation of the employees (...). This motivation is mainly done by the fact that the Warrants (also Call Options) that are given to employees will be very interesting to exercise if the company performs with them, and therefore the underlying price far exceeds the Exercise price of the Warrant. Thus, employees will exercise their Warrant and selling the underlying to get a profit (it is a common practice in USA for start-ups with limited financial resources in the beginning to hire specialists by offering warrants - like 10,000 warrants were distributed in the case of Microsoft later these employees became millionaires after the exercise of the Call they owned).

As we will see a little bit further below there is a huge difference between stock options and warrants relatively to the property of "dilution effect" and then we say that the warrant is not a "compensatory vehicles"!!!

15.3.10.3.5 Returns and Investments rates

To define the objective pursued by the owner of financial assets, we will refer to the economic motivation of any act of investing. This practice is to make currently an expense by expecting an increase of the asset in the future.

Of two or more investment strategies, the best at the individual level is the one that maximizes the final capital invested.

Then there exist different types of return rates on investment following the object of study. Thus, we differentiate in finance (before seeing the details):

1. The financial asset returns on a given economic horizon (return on investment) and their yield (rate of return, continuous or not, nominal or not).
2. The returns on investment compared to an average geometric rate of the market and the limit of the corresponding rate of return (internal rate of return).

Then we have to consider other rate of return approaches. Besides the two above-mentioned two other classics are portfolio management (before we see the details further below):

1. The "money weighted rate or return (M.W.R.R.)" for invested capital which has the advantage over the internal rate of return to take into account investments made outside the classical periods of time.
2. The "time of weighted rate of return (T.W.R.R.)" which is a useful tool for measuring the performance of fund managers because it does not take into account the flows (withdrawals or investment) of investors that are uncontrollable most of time.

Remark

We strongly recommend the reader to come back on the Capitalization and Actualization (see page 7547) subsection where we treat in details of: interest rates, nominal rates effective rates.

Let us see a bit all this in details:

15.3.10.3.5.1 Return on Investment (R.O.I.)

In practice, we will define the objective of the investor as of maximizing the growth of its initial fortune, whatever the modalities of this increase. This increase is named commonly "return on investment" (ROI) or, more briefly, just "return R" at time t and is defined by the (logical) following logical relation in the field of asset management by:

$$\boxed{R_t = (P_t - P_{t-1}) + C_t} \qquad (15.3.557)$$

where R_t is the return of the financial asset for the period (ending time) t, P_t the market price at time t of the financial asset and C_t the cash income attached to the holding of the financial assets during that same period (typically dividends). Obviously the practitioner must take care to take into account the various government taxes or trading fees when calculating this return, this is why when fees and taxes are not removed we speak rather of "total return"!

The income C_t is assumed to be received at time t, or, if it is perceived between time $t-1$ and t it is not supposed to be reinvested before time t (otherwise we must calculate it's future value). The market price P at time $t-1$ is a value "ex coupon" that is to say a value recorded immediately after the perception of the dividends. Empirically, the assumption of non-reinvestment until the elementary time period used is short (maximum one month) to avoid excessive statistical distortions in the treatment of historical data.

To facilitate comparisons between investment, we use a measure expressed in relative terms (as always) the "(short) rate of return" logically defined by:

$$\boxed{r_t = \frac{(P_t - P_{t-1}) + C_t}{P_{t-1}}} \qquad (15.3.558)$$

where r_t is the rate of return for period t.

The latter relation can also be found in the following form:

$$\text{TSR} = \frac{(P_1 - P_0) + D}{P_0} \qquad (15.3.559)$$

name the "total stock return" where P_1 is the initial stock price, P_0 the ending stock price (period 1), and D the dividends.

Never forget that we have proved in the section of Statistics that:

$$\mu_g < \mu_a \qquad (15.3.560)$$

ie the average long term period of return over-estimates the "real" average return given by the geometric mean (see section Statistics page 565)!

We will come back during our study of the mathematical model of asset pricing on these tools.

15.3.10.3.5.2 Internal Rate of Return (I.R.R.)

The use of a financial capital to enable the realization of real economic transactions (that is to say, devoting directly or indirectly, the financial capital for the acquisition or to build production tools, in the broadest sense of the term) can therefore produce over time positive or negative cash flows.

The actuarial calculus gives us the possibility to build a decision criterion to compare investments with different returns. Indeed, we define (logically but without being completely realistic) the risk-taking by the "net present value" as proved in the section of Quantitative Management (see page 8050):

$$\text{NPV}=\sum_{i=1}^{n}\frac{C_i}{(1+t\%)^i}+\frac{V_n}{(1+t\%)^n}-V_0 \tag{15.3.561}$$

where $t\%$ is supposed to be the risk free rate!

Obviously if there is no residual value this reduce to:

$$\boxed{\text{NPV}=\sum_{i=1}^{n}\frac{C_i}{(1+t\%)^i}-V_0} \tag{15.3.562}$$

where the C_i can sometimes be negative (intermediate investments). So the three decision criteria are for recall:

- $\text{NPV}<0$ corresponds to an investment that should be probably avoided
- $\text{NPV}=0$ then we are in front of an undecidable investment
- $\text{NPV}>0$ corresponds then probably to a good investment

Most of time the investor is interested to know in comparison to all its periodic investments and withdrawals (gains) what is the rate that cancels the NPV and after to compare this rate to that of the market risk free rate or some other portfolios or hedge funds rates. We know from our study on the section of Quantitative Management that this rate named "Internal Rate of Return" (IRR) is the average geometric rate (discount rate) $t\%$ such that it is the smallest positive root of the polynomial that satisfied:

$$\boxed{\text{NPV}=\sum_{i=1}^{n}\frac{C_i}{(1+t\%)^i}-V_0=0} \tag{15.3.563}$$

So if we consider that the smallest positive root of this polynomial equation is denoted IRR than for an initial investment V_0 and some wanted cash flows C_i, then if a banks offers us a given

TSR investment strategy (Total Stock Return) such that TSR $\geq$ IRR then we should probably consider this offer seriously.

We will see further below during our study of the Black & Scholes model much more sophisticated ways to consider investment strategies that are not purely deterministic as before but stochastic.

A firm's required rate of return on investments is often referred to as its "hurdle rate" because all projects must earn a rate of return high enough to clear this rate. Otherwise, a project will not cover its cost of financing, thereby reducing shareholder wealth.

15.3.10.3.5.3 Modified Internal Rate of Return (M.I.R.R.)

The "Modified Internal Rate of Return" (MIRR) is an empirical indicator for comparing financial flows with reinvestment. Even if the name seems to have a connection with the simple internal rate of return (IRR) seen earlier above it has almost nothing to do with it! The approach is totally different and the calculations of IRR and MIRR can give very different results.

Rather than theoretically speaking of this concept, a companion example is better than a thousand words.

Consider that a project is defined by the following flows:

Period	Cash Flow
0	$-1,000$
1	$-4,000$
2	$5,000$
3	$2,000$

To calculate the MIRR, we will assume that the borrowing rate (credit) is 10% and the risk-free rate of return of the market (potential reinvestment rate) is 12%.

We can then calculate what is the negative present value of the project:

$$\text{PV}_{\text{neg CF}}=-1000+\frac{-4000}{(1+10\%)^1}=-4,636.36.\text{-} \tag{15.3.564}$$

So far there is no difference with the usual implicit calculation in the calculation of the IRR through the NPV (Net Present Value). The previous sum of money therefore corresponds initially to the sum of money actually spent under the assumption that the first amount of $1,000$.- was already available in the form of a cash reserve and the second is a loan of $4,000$.- initiated at the beginning of the project and therefore the repayment is done at the end of the period relating to this loan (we can already see that it is therefore an extremely simplified scenario!).

On the other hand, for positive cash flows, the approach will be radically different to that of the NPV! We will not calculate their present value but their future value! It comes then (on the assumption that the analysis is done on the duration of the project only):

$$\text{FV}_{\text{pos CF}}=5000(1+12\%)^1+2000=7,600.00.\text{-} \tag{15.3.565}$$

This amount of money corresponds to the gain obtained if you had placed the positive cash flows in an investments with a return rate of 12% without risk (again it is super simplified because normally it is necessary to involve the probabilities and a given number of additional mathematical parameters...).

Once we have these two amounts of money $textPV_{\text{neg CF}}$ and $\text{FV}_{\text{pos CF}}$, then the idea of the MIRR, which then differs again considerably from that of the IRR, is to consider that the MIRR rate is empirically defined indirectly by:

$$\text{FV}_{\text{pos CF}} = |\text{PV}_{\text{neg CF}}|(1+\text{MIRR})^3 \tag{15.3.566}$$

Like the internal rate of return, the modified internal rate of return cannot be validly used to rank-order projects of different sizes, because a larger project with a smaller modified internal rate of return may have a higher present value. However, there exist variants of the modified internal rate of return which can be used for such comparisons.

Therefore with our companion example we get:

$$\text{MIRR} = \sqrt[3]{\frac{\text{FV}_{\text{pos CF}}}{|\text{PV}_{\text{neg CF}}|}} - 1 = \sqrt[3]{\frac{7,600}{|-4,636.36|}} - 1 \cong 17.91\% \tag{15.3.567}$$

Hence in Microsoft® Excel® 14.0.7106:

```
=MIRR({-1000,-4000,5000,2000},10%,12%)=17.91%
```

The general expression is:

$$\boxed{\text{MIRR} = \sqrt[n]{\frac{\text{FV}_{\text{pos CF}}}{|\text{PV}_{\text{neg CF}}|}} - 1} \tag{15.3.568}$$

Also denoted in different ways:

$$\text{MIRR} = \sqrt[n]{\frac{\text{FV}_{\text{CF+}}}{|\text{PV}_{\text{CF-}}|}} - 1 = \sqrt[n]{\frac{\text{Cash Inflows}}{|\text{Cash Outflows}|}} - 1 = \sqrt[n]{\frac{\text{Terminal cash flow}}{|\text{Outlay}|}} - 1 \tag{15.3.569}$$

Or even more explicit:

$$\text{MIRR} = \sqrt[n]{\frac{\text{FV(positive CF,reinvstment rate)}}{|\text{PV(negative CF, finance rate)}|}} - 1 = \sqrt[n]{\frac{\text{FV(positive CF,reinvstment rate)}}{-\text{PV(negative CF, finance rate)}}} - 1 \tag{15.3.570}$$

So in other words, the MIRR is the rate that corresponds to a capitalization during the periodic period of the project of the discounted (real) investment capital of the negative flows and which would yield the equivalent of the positive flows of the project.

15.3.10.3.5.4 Money Weighted Rate of Return (M.W.R.R.)

We will now introduce a kind of internal rate of return different the one related to NPV and that applies best to portfolio management (dynamics) than the internal rate of return see just before (for which we recall that one of the assumption is that cash flows are at periodic intervals).

Let us consider a fund F and the following information:

1. The value of the fund $F_0 \in \mathbb{R}^+$ just before time 0.
2. The value of the fund $F_1 \in \mathbb{R}^+$ just after the time 1.
3. A total net cash value $N \in \mathbb{R}$ invested during the period $[0,1]$ paid in two halves at the beginning and at end of period (to simplify the example ...).

The data that will interest us are:

1. The value:
$$F_0 + N/2$$
that represents the total value of the fund and part of the investment at time 0.
2. The value:
$$N/2(1+t\%)^{-1}$$
that represents the capital that we should have been gathering at time 0 to arrive at end of period to the value $N/2$ when the market free-risk rate is at $t\%$.
3. The value:
$$(1+t\%)^{-1}F_1$$
that represents at time 0 the value of the fund we would like to reach so we have at the end of period the value F_1 when the market free-risk rate is at $t\%$.

The difference:

$$(1+t\%)^{-1}F_1-(1+t\%)^{-1}\frac{N}{2}=(1+t\%)^{-1}\left(F_1-\frac{N}{2}\right) \tag{15.3.571}$$

gives the value that should have been capitalized to the get the sum $\left(F_1-\dfrac{N}{2}\right)$.

What is trivially interesting for an investor is then to know the rate $t\%$ such as the first investment equalizes its annual target. That is to say:

$$F_0+\frac{N}{2}=(1+t\%)^{-1}\left(F_1-\frac{N}{2}\right) \tag{15.3.572}$$

Hence after rearranging:

$$\boxed{F_0+\frac{N}{2}-(1+t\%)^{-1}\left(F_1-\frac{N}{2}\right)=0} \tag{15.3.573}$$

relationship is named "Hardy's relation".

If this relationship holds for F_0, N known and given F_1, an investor will have nothing to gain or to lose by investing in the fund or to capitalize at the market rate $t\%$.

If Hardy's relation is not equal to zero but positive then the investment in the fund is probably not interesting. If it is negative, then it is probably better to invest in the fund.

Using elementary algebra, this leads us to the relation:

$$F_0+\frac{N}{2}-(1+t\%)^{-1}\left(F_1-\frac{N}{2}\right)=0\Rightarrow t\%=\frac{2I}{F_0+F_1-I} \tag{15.3.574}$$

with:

$$I=F_1-F_0-N \tag{15.3.575}$$

Indeed:

$$t\%=\frac{F_1-\dfrac{N}{2}}{F_0+\dfrac{N}{2}}-1=\frac{F_1-\dfrac{N}{2}-F_0-\dfrac{N}{2}}{F_0+\dfrac{N}{2}}=\frac{2(F_1-F_0-N)}{2F_0+N}=\frac{2(F_1-F_0-N)}{F_0+F_1-(F_1-F_0-N)} \tag{15.3.576}$$

The rate $t\%$ is often named in fund management the "Money Weighted Rate of Return" (M.W.R.R.) or "Weighted Return Rate by Invested Capital" (W.R.R.I.C) and therefore represents the performance of a fund (portfolio) with consideration of inputs and outputs during the evaluation period.

Therefore the relation of money weighted rate of return that we got above:

$$\text{MWWR}=t\%=\frac{2I}{F_0+F_1-I} \tag{15.3.577}$$

can obviously also be written:

$$\text{MWWR}=\frac{F_1-F_0-N}{N/2} \tag{15.3.578}$$

And by changing the notation, as in some textbooks, this is rewritten:

$$\boxed{\text{MWWR}=\frac{V(T)-V(t_0)-F}{\bar{V}(t_0,T)}} \tag{15.3.579}$$

where F represents the net cash transactions recorded during the related time period (t_0,T); $V(t_0)$ the initial value of the fund; $V(T)$ the value of the fund at the end of the period of observation; $\bar{V}(t_0,T)$ the average capital the client has invested.

✍**Example:**

A fund had the following income during 12006 (holocene calendar):

- Value at 1 January 12006 (holocene calendar): 30 M$
- Investment in the fund during the year: 18 M$
- Withdrawals from the fund: 30 M$
- Value of the fund at 31 December 12006 (holocene calendar): 21 M$

and we want to know what is the effective rate (M.W.R.R.) of this fund in year 12006 (holocene calendar)?

We then as initial data $F_0 = 30$, $F_1 = 21$, $N = 18 - 30 = -12$ which gives if we accept the assumptions about N:

$$I = 21 - 30 - (-12) = 3 \tag{15.3.580}$$

and therefore:

$$\text{MWRR} = \frac{2 \cdot 3}{21 + 30 - 3} \cong 12.5\% \tag{15.3.581}$$

Let us consider now that we know that investments took place the $3/8$th part of the year and the withdrawals the $3/4$th part of the year.

The M.W.R.R. is then the rate of the following cash flows sequence ($C_i = (t, V_i)$):

$$\text{CF} = \{(0,30),(3/8,18),(3/4,-30),(1,21)\} \tag{15.3.582}$$

We must then find $t\%$ as:

$$30 = 21\left((1+t\%)^{-1} - 18(1+t\%)^{-3/8} + 30(1+t\%)^{-3/4}\right) \tag{15.3.583}$$

Solving this equation with Maple™ 4.00b gives:

$$\text{MWRR} = t\% = 8.875 \tag{15.3.584}$$

With a spreadsheet software like Microsoft® Excel® 14.0.7173 this is given by:

```
=XIRR(-30,-18,30,21,0,136.88,273.75,365)
```

We see that in considering the cash flow and the fractional times when they take place (thus more accurate and rigorous analysis) the M.W.R.R. is smaller. Moreover, the last calculation is more rigorous than the first, it is the latter that the investor want to know at the end of the year.

We see that from the above example in the more accurate M.W.R.R. example it comes that finally the M.W.R.R. is given by the $t\%$ that satisfies:

$$\boxed{F_0 = \sum_i \frac{\text{CF}_i}{(1+t\%)^{t/\Delta t}}} \tag{15.3.585}$$

This rate is therefore an effective measure of the increase funds rate, giving the impact of cash flow weight on the fund's value. It is therefore just a generalization of the I.R.R. (Internal Rate of Return).

Remark

In Switzerland, some insurance foundations of the LPP (Law of retirement provisions) communicate the annual or quarterly M.W.R.R. on the 5 to 10 past years.

15.3.10.3.5.5 Time Weighted Rate of Return (T.W.R.R.)

We will now discuss another financial tool for portfolio management also used to judge the performance of an investment.

Let us consider a fund such as (the years in the table below are given according to the holocene calendar):

	December 31 12000	**T1 12001**	**T2 12001**	**T3 12001**	**T4 12001**
Starting value of the fund		$1,000$.-	370.-	81.-	7.8.-
Profit or (loss) for the quarter in $\%$		10%	3%	-4%	6%
Profit or (loss) for the quarter (in .-)		100.-	11.1.-	-3.2.-	0.5.-
Quarterly cash flow Inflows/Outflows		-730.-	-300.-	-70.-	0.-
Fund value	$1,000$.-	370.-	81.1.-	7.8.-	8.3.-

Table 15.51: Time Weighted Rate Of Return

The 31 December 12000, the fund has a value of $1,000$.- value. During the first quarter T1 of 12001 it has a $+10\%$ return but we imagine that this value is far from what was expected, then the investor withdraws 730.- of the fund (portfolio based on the funds). During the second quarter T2, the fund gained $+3\%$ and an additional amount of 300.- were withdrawn of the fund by the investor. In the third quarter T3 the fund lost -4% and 70.- were removed from the fund. The last quarter T4, the fund gained $+6\%$ and no funds have been withdrawn.

We then have the overall increased (return) on the whole period (year) that is given by:

$$\text{TWRR} = (1+0.1)(1+0.03)(1-0.4)(1+0.06)-1 = 15.3\% \tag{15.3.586}$$

We see well that this value is independent of the cash flows of the investor's portfolio. We name the value of 15.3% the "Time Weighted Rate of Return" (T.W.R.R) on a quarterly basis.

In other words, the T.W.R.R. simply indicates the performance of a portfolio without taking into account the inflow and outflows during the evaluation period.

This specific case can be written generally as:

$$\boxed{\text{TWRR} = \prod_{i=1}^{n}(1+r_i)-1} \tag{15.3.587}$$

It should be remembered that if we wanted to calculate the average fund performance by quarter, we would have simply used the geometric mean (see section Statistics page 565)!

The T.W.R.R. is a practical tool to measure the performance of fund managers because it does not take into account the investors cash flows that are most of time uncontrollable. Thus, we have a measure of the quality of the dynamics of the fund independent of the choice of the investors that could consider withdrawals or investments and that would be used to calculate an I.R.R. which would more or less be meaningless compared to the dynamics of the fund.

Remark

In Switzerland, some insurance foundations of the LPP (Law of retirement provisions) communicate the annual or quarterly T.W.R.R. on the 5 to 10 past years but on a daily basis.

15.3.10.3.6 Theory of Speculation

After these many contextual definitions, the goal now is to introduce basic stochastic speculative mathematical techniques used in finance. Indeed, finance has become over time a domain increasingly competitive, the margins on standard products tend to be reduced, the premium is on innovation and automation. This has led to an increasing sophistication of financial products, thus making use of advanced mathematical concepts, primarily based on probability models introduced by Louis Bachelier in his *Theory of Speculation* (11900 according to holocene calendar) but really used massively in practice only since 11973 (holocene calendar) through different work of Black & Scholes and Merton (who have earned the Nobel Prize in economics).

Let us look to start what are the developments proposed by Louis Bachelier in his thesis to determine the predictive mathematical expectation and the estimated standard deviation of a financial asset (result that we will use as part of our study of the Black & Scholes evaluation model).

Let us denote by $p_{x,t}$ the probability density function of the price of an asset is traded x at time t. Therefore, the elementary cumulative probability that the current value lies within the elementary interval $[x, x + \mathrm{d}x]$ at time t is of the form:

$$p_{x,t}\mathrm{d}x \tag{15.3.588}$$

which integral on the whole domain of definition will have obviously to be equal to 1.

Under the fourth axiom of Probabilities (see section of the same name page 476), the elementary cumulative probability that the price moves of a certain value to another one (temporal Markov chain in continuous time), will be equal to the product of the density probability that value is traded x in a given time interval at the time t_1 that is to say:

$$p_{x,t_1}\mathrm{d}x \tag{15.3.589}$$

multiplied by the elementary cumulative probability that the price being treated at the value x at the time t_1, is traded z during a time interval $t_1 + t_1$, that is to say, multiplied by:

$$p_{z-x,t_2}\mathrm{d}z \tag{15.3.590}$$

The desired elementary cumulative probability is then obviously:

$$p_{x,t_1}p_{z-x,t_2}\mathrm{d}x\mathrm{d}z \tag{15.3.591}$$

This notation supposes that trading values are independent random variables...

As the traded value can be found at time t_1 in all the intervals $\mathrm{d}x$ included between $\pm\infty$, the cumulative probability that the traded value is z at the time $t_1 + t_2$ will be (the integration is relatively to $\mathrm{d}x$):

$$\int\limits_{-\infty}^{+\infty} p_{x,t_1}p_{z-x,t_2}\mathrm{d}z\mathrm{d}x \tag{15.3.592}$$

As the probability that the value z in the time interval t_1+t_2 has for expression:

$$p_{z,t_1+t_2}\mathrm{d}z \tag{15.3.593}$$

Then we have:

$$p_{z,t_1+t_2}\mathrm{d}z=\int\limits_{-\infty}^{+\infty} p_{x,t_1}p_{z-x,t_2}\mathrm{d}z\mathrm{d}x \tag{15.3.594}$$

or after simplification:

$$p_{z,t_1+t_2}=\int\limits_{-\infty}^{+\infty} p_{x,t_1}p_{z-x,t_2}\mathrm{d}x \tag{15.3.595}$$

Such is the equation that must satisfy the probability distribution function p that we seek. This equation is true, as we will see it, under some strong assumptions, by the function:

$$p(x,t)=Ae^{-B^2x^2} \tag{15.3.596}$$

But do not forget that this is a special solution (this why this model is sometimes named "Bachelier's Gaussian model") ... and nothing guarantees us that two independent random variables follow the same probability distribution...

The both assumptions to build the model seen so far (independence and identical distribution) are often indicated in finance under the name of "independence and stationary assumptions" and this is why the time variable does not appear anymore!

That said, we have then of course to require (probability axioms oblige!):

$$\int\limits_{-\infty}^{+\infty} p\mathrm{d}x=A\int\limits_{-\infty}^{+\infty} e^{-B^2x^2}\mathrm{d}x=1 \tag{15.3.597}$$

We recognize in the second term above the Gauss Integral that has for value (see section Statistics page 695):

$$A\int\limits_{-\infty}^{+\infty} e^{-B^2x^2}\mathrm{d}x=A\frac{\sqrt{\pi}}{B}=1 \tag{15.3.598}$$

We must therefore necessarily have for the normalization:

$$B=A\sqrt{\pi} \tag{15.3.599}$$

It results:

$$p(x,t)=Ae^{-\pi A^2x^2} \tag{15.3.600}$$

By putting $x=0$, we get $A=p_0$, that is to say that A equals the actual traded price.

We must now check that relation:

$$p(x,t)=p_0e^{-\pi p_0^2x^2} \tag{15.3.601}$$

satisfies well:

$$p_{z,t_1+t_2}=\int\limits_{-\infty}^{+\infty} p_{x,t_1}p_{z-x,t_2}\mathrm{d}x \tag{15.3.602}$$

under the given assumptions!!!!

Given the quantities p_1,p_2 the quantities corresponding to p_0 and relative to times t_1,t_2 then we must prove that the expression:

$$\int\limits_{-\infty}^{+\infty} p_1e^{-\pi p_1^2x^2}p_2e^{\pi p_2^2(z-x)^2}\mathrm{d}x \tag{15.3.603}$$

can be put under the form $Ae^{-B^2z^2}$ where A, B depends of the time!

This integral becomes by noting that z is not an integration variable (we assume that it is independent of x as you have understand from the beginning!):

$$p_1p_2e^{-\pi p_2^2z^2}\int\limits_{-\infty}^{+\infty} e^{-z(p_1^2+p_2^2)x^2+2\pi p_2^2zx}\mathrm{d}x \tag{15.3.604}$$

Now we will change the shape of the integral (we also change the exponential notation otherwise it becomes unreadable):

$$p_1p_2\exp\left(-\pi p_2^2z^2+\frac{\pi p_2^4z^2}{p_1^2+p_2^2}\right)\int\limits_{-\infty}^{+\infty}\exp\left(\pi\left(x\sqrt{p_1^2+p_2^2}-\frac{p_2^2z}{\sqrt{p_1^2+p_2^2}}\right)^2\right)\mathrm{d}x \tag{15.3.605}$$

and let us put:

$$x\sqrt{p_1^2+p_2^2}-\frac{p_2^2z}{\sqrt{p_1^2+p_2^2}}=u \tag{15.3.606}$$

Then we will have:

$$\frac{p_1p_2\exp\left(-\pi p_2^2z^2+\dfrac{\pi p_2^4z^2}{p_1^2+p_2^2}\right)}{\sqrt{p_1^2+p_2^2}}\int\limits_{-\infty}^{+\infty}e^{-\pi u^2}\mathrm{d}u \tag{15.3.607}$$

The integral:

$$\int\limits_{-\infty}^{+\infty}e^{-\pi u^2}\mathrm{d}u=2\int\limits_{0}^{+\infty}e^{-\pi u^2}\mathrm{d}u \tag{15.3.608}$$

having for value 1 we get finally:

$$\frac{p_1p_2\exp\left(-\pi p_2^2z^2+\dfrac{\pi p_2^4z^2}{p_1^2+p_2^2}\right)}{\sqrt{p_1^2+p_2^2}}=\frac{p_1p_2}{\sqrt{p_1^2+p_2^2}}\exp\left(\frac{-\pi p_1^2p_2^2}{p_1^2+p_2^2}z^2\right) \tag{15.3.609}$$

This expression has the desired shape since assuming $p_1 > 0, p_2 > 0$ a numerical integration gives:

$$\int_{-\infty}^{+\infty} \frac{p_1 p_2}{\sqrt{p_1^2+p_2^2}} \exp\left(\frac{-\pi p_1^2 p_2^2}{p_1^2+p_2^2} z^2\right) \mathrm{d}z = 1 \tag{15.3.610}$$

This relation expresses the fact that the total cumulative probability that the random variable z can take any value is equal to unity.

We must conclude that the probability that the asset is traded z at time $t_1 + t_2$ is expressed by the relation:

$$p = p_0 e^{-\pi p_0^2 x^2} = \frac{p_1 p_2}{\sqrt{p_1^2+p_2^2}} \exp\left(\frac{-\pi p_1^2 p_2^2}{p_1^2+p_2^2} z^2\right) \tag{15.3.611}$$

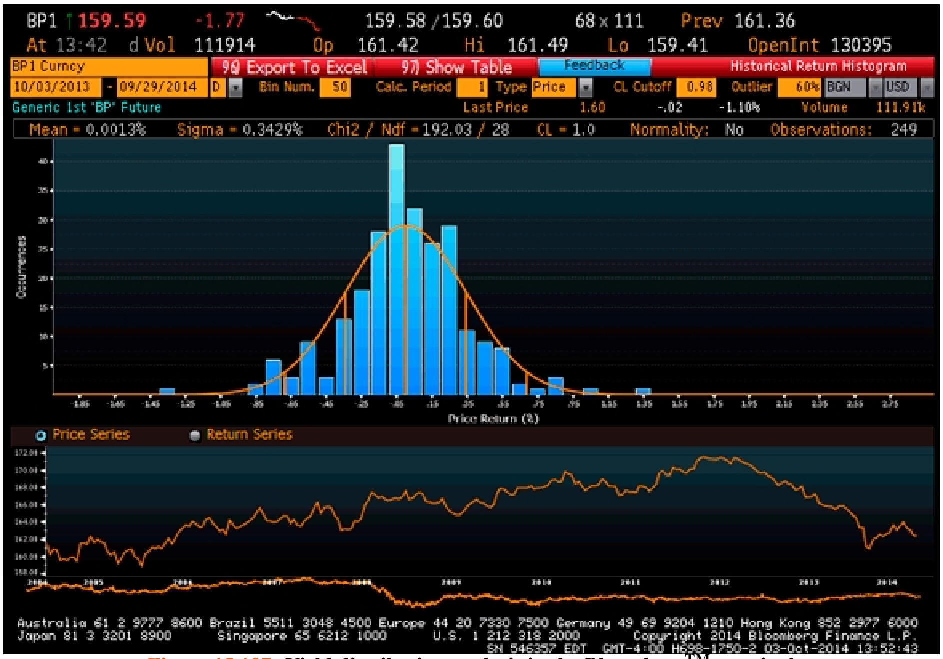
That can be written:

$$p = p_0 e^{-\pi p_0^2 x^2} = \frac{p_0(t_1) p_0(t_2)}{\sqrt{p_0^2(t_1)+p_0^2(t_2)}} \exp\left(\frac{-\pi p_0^2(t_1) p_0^2(t_2)}{p_0^2(t_1)+p_0^2(t_2)} z^2\right) \tag{15.3.612}$$

and this proves our initial hypothesis.

We see that the probability is then still governed by a Normal centered reduced distribution type (see section Statistics page 692)! This is a remarkable result obtained by Louis Bachelier in 11900 (holocene calendar) and which had already been speculated by Jules Regnault in the mid 119th century (holocene calendar)!

Indeed, Jules Regnault compared speculation to a heads or tail game in which the two sides of the coin correspond to the two possibilities, increase or decrease. Assuming that at any time whatsoever, there are never more chance for benefits than for loose. In other words, at each quote value, price has 50% chance of increasing and 50% of decreasing. But every trader has their opinion on the issue... Without this diversity of opinions, there would therefore no trades or price changes. So operators are divided into two groups (bullish, bearish) that make subjective assessments of the future value of the course which necessarily include a margin of error. However, to Jules Regnault, errors speculators are not any, they follow a Normal distribution. Indeed, as shown by Pierre Simon de Laplace, if the probability of error is small and there are numerous independent errors then the results follow a Normal distribution (see section Statistics page 692).



BP1 159.59 -1.77 159.58 / 159.60 68 x 111 Prev 161.36
At 13:42 d Vol 111914 Op 161.42 Hi 161.49 Lo 159.41 OpenInt 130395
BP1 Curncy
10/03/2013 - 09/29/2014 D Bin Num. 50 Calc. Period 1 Type Price CL Cutoff 0.98 Outlier 60% BGN USD
Generic 1st 'BP' Future Last Price 1.60 -.02 -1.10% Volume 111.91k
Mean = 0.0013% Sigma = 0.3429% Chi2 / Ndf = 192.03 / 28 CL = 1.0 Normality: No Observations: 249
Australia 61 2 9777 8600 Brazil 5511 3048 4500 Europe 44 20 7330 7500 Germany 49 69 9204 1210 Hong Kong 852 2977 6000
Japan 81 3 3201 8900 Singapore 65 6212 1000 U.S. 1 212 318 2000 Copyright 2014 Bloomberg Finance L.P.
SN 546357 EDT GMT-4:00 H698-1750-2 03-Oct-2014 13:52:43

Figure 15.127: Yield distribution analysis in the Bloomberg$^{\text{TM}}$ terminal

The prior previous relations shows that the parameters $p_0=f(t)$ satisfy the functional relation:

$$f^2(t_1+t_2)=\frac{f^2(t_1)f^2(t_2)}{f^2(t_1)+f^2(t_2)} \qquad (15.3.613)$$

Let us differentiate relatively to t_1 and after relatively to t_2. The first member having the same shape in both cases, we get:

$$\begin{aligned}2f'(t_1)&=\mathrm{d}_{t_1}\left(\frac{f^2(t_1)f^2(t_2)}{f^2(t_1)+f^2(t_2)}\right)=\left(\frac{2f(t_1)}{f^2(t_1)+f^2(t_2)}-\frac{2f^3(t_1)}{(f^2(t_1)+f^2(t_2))^2}\right)f^2(t_2)\\ 2f'(t_2)&=\mathrm{d}_{t_2}\left(\frac{f^2(t_1)f^2(t_2)}{f^2(t_1)+f^2(t_2)}\right)=\left(\frac{2f(t_2)}{f^2(t_1)+f^2(t_2)}-\frac{2f^3(t_2)}{(f^2(t_1)+f^2(t_2))^2}\right)f^2(t_1)\end{aligned} \qquad (15.3.614)$$

So after simplification:

$$\left.\begin{aligned}f'(t_1)&=\frac{f(t_1)f^4(t_2)}{(f^2(t_1)+f^2(t_2))'2}\\ f'(t_2)&=\frac{f(t_2)f^4(t_1)}{(f^2(t_1)+f^2(t_2))'2}\end{aligned}\right\}\Rightarrow\frac{f(t_1)f^4(t_2)}{f'(t_1)}=\frac{f(t_2)f^4(t_1)}{f'(t_2)} \qquad (15.3.615)$$

Which finally gives:

$$\frac{f^3(t_2)}{f'(t_1)}=\frac{f^3(t_1)}{f'(t_2)} \qquad (15.3.616)$$

This relation being verified regardless the values of t_1,t_2, the common value of the two ratio is constant and then we have:

$$f'(t)=c^{te}f^3(t) \qquad (15.3.617)$$

A function that satisfies this relation exists and is:

$$f(t) = p_0 = \frac{H}{\sqrt{t}} \tag{15.3.618}$$

Where H denotes a constant or an independent function of time.

Verification:

$$f'(t) = -\frac{1}{2}\frac{H}{t^{3/2}} \qquad f^3(t) = \frac{H^3}{t^{3/2}} \tag{15.3.619}$$

therefore:

$$c^{te} = -\frac{1}{2H^2} \tag{15.3.620}$$

thus we have the final expression for the probability density function of the current traded value x under the above assumptions:

$$\boxed{p = \frac{H}{\sqrt{t}}e^{-\frac{\pi H^2 x^2}{t}} = \mathcal{N}\left(0, \frac{\sqrt{t}}{H\sqrt{2\pi}}\right) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}} = \mathcal{N}(0,\sigma)} \tag{15.3.621}$$

with x that is for recall greater than or equal to 0.

The reader will have noticed perhaps that for a given value of H and t fixed we always have here as a centered Normal distribution (see section Statistics page 692)!! Financial analysts then say that we are dealing with a "wise chance", this implied that the changes are small and regular.

Let us now obviously calculate the mean and variance of this important distribution function (notice that these two moments should be dependent of time!).

As the traded value should not be negative (in most cases... because in practice we know that it happens very rarely!), we restrict the calculation to the positive mean as being (see section Statistics page 634):

$$\begin{aligned}\mu &= \int\limits_0^{+\infty} \frac{Hx}{\sqrt{t}}e^{-\frac{\pi H^2 x^2}{t}}\mathrm{d}x = \frac{H}{\sqrt{t}}\int\limits_0^{+\infty} xe^{-\frac{\pi H^2 x^2}{t}}\mathrm{d}x = \frac{H}{\sqrt{t}}\left(-\frac{t}{2\pi H^2}e^{\frac{\pi H^2 x^2}{t}}\right)\Bigg|_0^{+\infty} \\ &= \frac{H}{\sqrt{t}}\left(\left[0 - \left(-\frac{t}{2\pi H^2}\right)\right]\right) = \frac{t}{2\pi H^2}\frac{H}{\sqrt{t}} = \frac{\sqrt{t}}{2\pi H}\end{aligned} \tag{15.3.622}$$

by defining:

$$k = \frac{1}{2\pi H} \tag{15.3.623}$$

the "coefficient of instability" (of which we know nothing). So finally we have the positive mean of the traded value that is:

$$\boxed{\mu^+(t) = \int\limits_0^{+\infty} px\mathrm{d}x = k\sqrt{t}} \tag{15.3.624}$$

the expected mean of the traded value is therefore proportional to the square root of time like is the Brownian motion that we studied in the section of Statistical Mechanics (see page 3582)!!

It also follows immediately from this result that the average deviation of the current value at two consecutive times is also proportional to the square root of time between the two times!

We also notice that at the instant where $t=0$, the positive expected mean of the gain is zero, because the value is known with certitude (it is how it should be interpreted).

Let us also calculate the positive variance:

$$\mathrm{V}^{+}(x)=\int\limits_{0}^{+\infty}\frac{Hx^2}{\sqrt{t}}e^{-\frac{\pi H^2x^2}{t}}\mathrm{d}x=\frac{H}{\sqrt{t}}\int\limits_{0}^{+\infty}x^2e^{-\frac{\pi H^2x^2}{t}}\mathrm{d}x \tag{15.3.625}$$

We put:

$$z=\sqrt{\pi}tHx \tag{15.3.626}$$

Therefore:

$$x=\frac{1}{H}\sqrt{\frac{t}{\pi}}\Rightarrow x^2=z^2\frac{t}{H^2\pi} \tag{15.3.627}$$

with:

$$\frac{\mathrm{d}z}{\mathrm{d}x}=\sqrt{\frac{\pi}{t}}H \tag{15.3.628}$$

Therefore it comes:

$$\mathrm{V}^{+}(x)=\int\limits_{0}^{+\infty}\frac{Hx^2}{\sqrt{t}}e^{-\frac{\pi H^2x^2}{t}}\mathrm{d}x=\frac{H}{\sqrt{t}}\frac{1}{H}\sqrt{\frac{t}{\pi}}\frac{t}{H^2\pi}\int\limits_{0}^{+\infty}z^2e^{-z^2}\mathrm{d}z=\frac{t}{H^2\pi\sqrt{\pi}}\int\limits_{0}^{+\infty}z^2e^{-z^2}\mathrm{d}z \tag{15.3.629}$$

But in the section Statistics, we have proved by integration by parts (see section Differential and Integral Calculus page 1532) that:

$$\int\limits_{0}^{+\infty}z^2e^{-z^2}\mathrm{d}z=\left.\frac{-ze^{-z^2}}{2}\right|_{0}^{+\infty}+\int\limits_{0}^{+\infty}\frac{e^{-z^2}}{2}\mathrm{d}z=\frac{1}{2}\int\limits_{0}^{+\infty}e^{-z^2}\mathrm{d}z=\frac{1}{2}\sqrt{\frac{\pi}{4}}=\frac{\sqrt{\pi}}{4} \tag{15.3.630}$$

Thus:

$$\mathrm{V}^{+}(x)=\frac{t}{2H^2\pi\sqrt{\pi}}\frac{\sqrt{\pi}}{4}=\frac{t}{4H^2\pi} \tag{15.3.631}$$

So if we put:

$$K=\frac{1}{2H\sqrt{\pi}} \tag{15.3.632}$$

We have finally:

$$\boxed{\sigma^+ = \sqrt{\int\limits_0^{+\infty} px^2 \mathrm{d}x} = K\sqrt{t}} \tag{15.3.633}$$

Thus the standard deviation is also proportional to the square root of time (and the result would be the same if we had calculated the global standard deviation)!

Remark

The Brownian motion (and its variations) is heavily used by professionals, since many annualized volatility calculations (in % / year) which we find the results in any financial page of the daily press, are only the square root conversions time of periodic volatility calculations (in % / month or % / week) used as the basis of estimation.

Multiplying monthly Standard Deviation by the $\sqrt{12}$ is an industry standard method of approximating annualized Standard Deviations of Monthly Returns.

Therefore by the stability of the Normal law (see section Statistics page 710) we have:

$$\Delta p = \mathcal{N}\left(\mu_2 - \mu_1, \sqrt{\sigma_2^2 - \sigma_1^2}\right) = \mathcal{N}\left(0, \sqrt{\frac{t_2}{4\pi H^2} - \frac{t_1}{4\pi H^2}}\right) \tag{15.3.634}$$

It follows immediately that:

$$\sigma_{\Delta p} = \frac{1}{2H\sqrt{\pi}}\sqrt{\Delta t} \tag{15.3.635}$$

(the variance $\mathrm{V}(\Delta p) \propto \Delta t$ is sometimes named "quadratic variation" in stochastic finance) and:

$$\mu_{\Delta p} = \frac{\sqrt{t_2 - t_1}}{2\pi H} = \frac{1}{2\pi H}\sqrt{\Delta t} \tag{15.3.636}$$

So the traded price variations of a financial asset between two successive times have then obviously a probability distribution described by a Normal centered distribution (thus resulting by the stability property of this law) that is characterized by a positive expected mean and a positive standard deviation proportional to the square root of time!

This mathematically proved result was measured by Jules Regnault fifty years before Bachelier (~11850 according to holocene calendar) by observing that the standard deviation of French bonds was proportional to the square root of time.

That said we have to accept the limitations of this approach. Let us take for example the daily returns of the Dow Jones in 12008 and 12009 (holocene calendar). According to experts having the detailed of this data, it would follow rather follow a Student's T-distribution with 3 degrees of freedom rather than a Normal distribution...!

To give a blatant comparison of the limits of these approach let us remind (see section Statistics page 692) that the cumulative probability of a random variable following a Normal distribution beyond 4σ is $1-99.99366\%$, that is 0.00634%. This means, if the stock market has 252 open workdays, we have certainty of having a large deviation every:

$$\frac{1}{\dfrac{0.00634\%}{252}} \cong 63\,[\text{years}] \tag{15.3.637}$$

we therefore consider (see section Probabilities page 453) pairwise disjoint events.

Finally three major results are to be retained here under the strong assumptions of centered Normality and independence:

1. That the probability distribution function that the price of a financial asset to be x at time t follows a Normal centered distribution... (and hence can the price can go freely negative)!!
2. That positive expected mean and the positive standard-deviation of the value of a financial asset are proportional to the square root of time with a factor about which we don't know anything!!
3. That the expected mean gain is generally zero (which makes it unrealistic model but already provides a working basis...).

Much later below we will compare the Bachelier Call pricing Model and the Black & Scholes Call pricing model. But for this we need to know an assumption that did Bachelier crystallized in his famous dictum:

L'ésperance mathématique du spéculateur est nul

i.e. « *the mathematical expectation of a speculator is zero* ». Therefore as we assume a Normally distribution of the price and the standard deviation and the mean are proportional to the square root of time but that the mean is assumed as zero, we get:

$$\boxed{p=\mathcal{N}\left(0,\sigma\sqrt{t}\right)=\frac{1}{\sigma\sqrt{2\pi t}}e^{-\frac{x^2}{2\sigma^2 t}}} \tag{15.3.638}$$

and this is the relation we will have to use much later below.

Finally, it should be noted that this is a theoretical model...! It must therefore always be compared to real practical values to see if it is valid or not. In this case the observation of financial markets shows that it is only outside of the periods of speculative bubbles that the variations can be modelled by a Brownian motion. We must therefore look for more powerful models and we will see one day (...) that the Brownian movement (also named "Brownian process") that is smooth (continuous) and therefore without sudden jumps (therefore unable to model some sudden event markets like the crash or corrections) is a special case of "Lévy processes".

15.3.10.3.7 Portfolio efficient diversification models

Diversification is a well known way of reduce common known investment risks under the assumptions that the risk must be minimized and a given portfolio construction target is reached.

The most common efficient diversification refers to the organizing principle of portfolio theory, which attempts to maximize portfolio expected return for a given level of portfolio risk. In other words, having enough diversification in an investment portfolio to earn money, while still keeping risks within reasonable bounds.

Here is the definition given by the NASDAQ (`http://marketbusinessnews.com/financial-glossary/efficient-diversification/`) (it is not the best definition as we will see later with the various existing models that are not based only on highest return):

Definition (#15.3.691): "Efficient diversification" is the organizing principle of portfolio theory, which maintains that any risk-averse investor will search for the highest expected return for any particular level of portfolio risk.

There is quite a large number of others empirical models of portfolios diversification strategies. Among the best known (we tried to put them in the order in which they are normally taught at University:

1. Overall minimum variance portfolio (PVGM) also named Markowitz portfolio (see page 7699)
2. Tangent portfolio (TP) (see page 7755)
3. Minimum overall robust variance portfolio (MORVP)
4. Minimum overall robust semivariance portfolio (MORSVP)
5. Maximum Sharpe ratio portfolio (MSRP) (see page 7713)
6. Minimum Tracking error portfolio (MTEP) (see page 7733)
7. Identical weight portfolio (P1/N)
8. Treynor-Black weight portfolio (TBP) (see page 7735)
9. Zero Index portfolio (ZIP)
10. Neutral-Dollar portfolio (sum of weight zero) (NDP)
11. Single Average portfolio (SAP)
12. Time-Invariant Portfolio Protection (TIPP)
13. Zero Beta portfolio (ZBP)
14. Constant Proportion portfolio Insurance (CPPI)

15. Maximum utility portfolio (MUP)
16. Maximum diversification portfolio (MDP) (see page 7734)
17. Black-Litterman portfolio (BLP)
18. Draw-down portfolio (DDP)
19. Equal Risk Contribution optimal weights portfolio (ERC) (see page 7732)
20. Risk Parity portfolio (RPP) (see page 7738)
21. Minimum α-VaR portfolio (MVaRP) (see page 7872)
22. Minimum α-CVaR portfolio (MCVaRP)
23. ...

A number of practical and theoretical problems have limited the use of portfolio diversification models by investment managers. For example, it is often difficult to obtain sufficient high-quality historical data for thorough analysis. In addition, the efficient frontier where optimal portfolios lie tends to shift over time, quickly making these portfolios suboptimal.

For each of the model above we will start with the theoretical aspect and give afterwards a detailed example with Microsoft® Excel®, `R` or MATLAB® depending of what software is the more suited to the corresponding.

15.3.10.3.7.1 Overall minimum variance portfolio (Markowitz portfolio)

They are different approaches for introducing the Markowitz portfolio diversification model. We will start with the one using quadratic optimization.

Given R_p the performance of a portfolio of n assets characterized by their respective performance (return) $R_1, R_2, \ldots, R_n$. We put, as constraint, that each asset i enters the portfolio P in a proportion X_i such that:

$$\sum_{i=1}^{n} X_i = 1 \tag{15.3.639}$$

Remark

A part X_i of an asset may also be negative ... Hold a negative share of an asset is what is already known as being "short-selling" (short sale). This technique consists for recall to borrow a lot of assets (assumed overvalued on the market) to a bank, sell them to make the price of the asset go lower, make a profit by buying them cheaper to give them back to the bank (roughly it can be quite more complex and subtle in the reality but some organized group of people used this strategy to make some companies go bankrupt...).

Therefore the variance of the portfolio is given by (see section Statistics page 597) using the standard variance (we will see later that there is a model using "semivariance"):

$$\text{V}(R_P)=\sum_{i=1}^{n} X_i^2 \text{V}(R_i)+2\sum_{i,j|i<j} X_i X_j \text{cov}(R_i,R_j) \tag{15.3.640}$$

Before going further, let us precise (because it is important in practice!) that we can also write the last relation in matrix form (the reader can easily verify that by taking such two assets that both formulation give an identical result) if we denote by $\vec{X}$ the vector or the portions of assets and $\vec{X}^T$ the correspond transposed vector:

$$\vec{X}=\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \qquad X^T=(X_1\, X_2\, \ldots X_n) \tag{15.3.641}$$

and finally c_{ij} is the covariance matrix (remember, that there exist a simple method to move from the covariance matrix to the correlation matrix and vice versa as seen in the section Statistics at page 604):

$$c_{ij}=\begin{bmatrix} \text{cov}(R_1,R_1) & \text{cov}(R_1,R_2) & \ldots & \text{cov}(R_1,R_n) \\ \text{cov}(R_2,R_1) & \text{cov}(R_2,R_2) & & \vdots \\ \vdots & & \ddots & \\ \text{cov}(R_n,R_1) & \ldots & & \text{cov}(R_n,R_n) \end{bmatrix} \tag{15.3.642}$$

Matrix that as we know (see section Statistics page 604) can be simplified directly in:

$$c_{ij}=\begin{bmatrix} \sigma_1^2 & \text{cov}(R_1,R_2) & \ldots & \text{cov}(R_1,R_n) \\ \text{cov}(R_2,R_1) & \sigma_2^2 & & \vdots \\ \vdots & & \ddots & \\ \text{cov}(R_n,R_1) & \ldots & & \sigma_n^2 \end{bmatrix} \tag{15.3.643}$$

we finally get the relation of the variance in condensed matrix form:

$$\boxed{\text{V}(R_P)=\vec{X}^T(c_{ij}\vec{X})} \tag{15.3.644}$$

as we often see it in the literature.

To come back to the algebraic form of the model, since the covariance is symmetric (see section Statistics page 604):

$$\text{cov}(R_i,R_j)=\text{cov}(R_j,R_i) \tag{15.3.645}$$

and that:

$$\text{cov}(R_j,R_j)=\text{V}(R_j) \tag{15.3.646}$$

We can simplify and write the variance:

$$\text{V}(R_P)=\sum_{i=1}^{n} X_i^2 \text{V}(R_i)+2\sum_{i,j|i<j} X_i X_j \text{cov}(R_i,R_j) \tag{15.3.647}$$

as following:

$$\boxed{\text{V}(R_P)=\sum_{i=1}^{n}\sum_{j=1}^{n}X_iX_j\text{cov}(R_i,R_j)} \qquad (15.3.648)$$

as we often see it in the literature of a few decades old...

Select a Markowitz portfolio is therefore equivalent as solving the following constraints under maximization problem:

$$\boxed{\begin{array}{ll}\underset{\text{V}(R_P)}{\text{minimize}} & \text{E}(R_P)=c^{te}\\ \text{subject to} & \sum X_i=1\end{array}} \qquad (15.3.649)$$

using quadratic programming (see section Numerical Methods page 6673) or the Lagrange multiplier method where we have detailed how to resolve this specific case (see section Numerical Methods page 6670).

Remark

The above formulation is also often denoted ($\vec{w}$ is the vector of weights and Σ is the variance-covariance matrix):

$$\begin{cases}\text{minimize} & \sigma_P^2(\vec{w}):\vec{w}^T\Sigma\vec{w}\\ \text{under const.} & \vec{w}^T\vec{\mu}=\mu_0\\ & \sum_i w_i=1\end{cases} \qquad (15.3.650)$$

also sometimes denoted:

$$\begin{cases}\text{minimize} & \sigma_P^2(\vec{w}):\vec{w}^T\Sigma\vec{w}\\ \text{under const.} & \vec{w}^T\vec{R}=R_p\\ & \mathbb{1}^T\vec{w}=1\end{cases} \qquad (15.3.651)$$

either that the investor seeks the vector $\vec{w}$ which minimizes his risk for a given expected return μ_0 (i.e. R_p).

Either reciprocally, opt for the search for $\vec{w}$ which maximizes the expected return of the portfolio given a maximum level of risk σ_0:

$$\begin{cases}\text{maximize} & \mu_P^2(\vec{w}):\vec{w}^T\vec{\mu}\\ \text{under const.} & \vec{w}^T\Sigma\vec{w}=\sigma_0\\ & \sum_i w_i=1\end{cases} \qquad (15.3.652)$$

In practice, we seek not one but all portfolios that for a given expected return minimize the variance! We then obtain a function of the mean return based on the variance for the optimal portfolios. If we plot it on a graph (see below). This function is often assimilated by financial practitioners (rightly!) to a frontier as specified in the following definition.

Definition (#15.3.692): The frontier that characterizes the polygon or the curve of constraints in this situation is named the "(Markowitz) efficient frontier" and in the polygon / curve occur all portfolios to reject that are so named "dominated portfolios". Another way of stating this is to say that the combinations of (performance, risk) of this frontier form an optimum set of Pareto (see section Game and Decision Theory page 7394), that is to say, if one of the elements increases, the other must also increase.

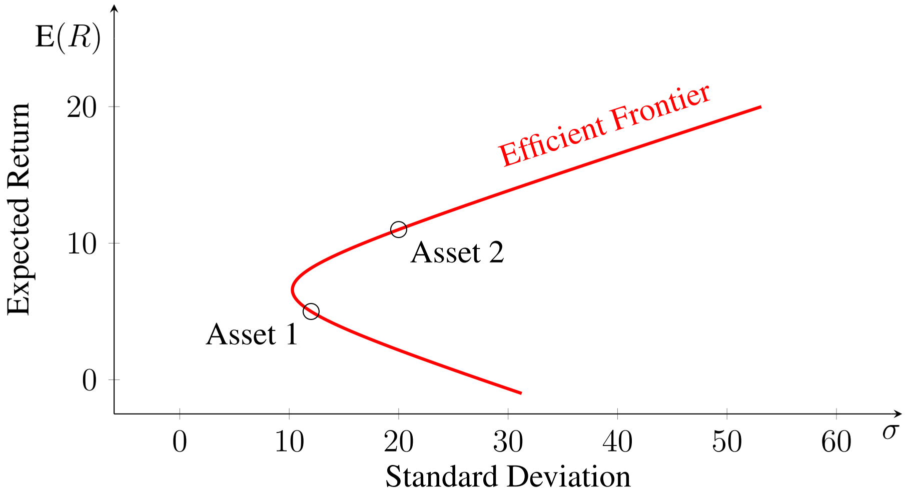


Figure 15.128: **Efficient Frontier**

Now let us formalize the optimization in the way that was done in the days when people had to further develop the algorithms themselves...

Given Z the aforementioned economic function:

$$Z = \Phi \mathrm{E}(R_P) - \mathrm{V}(R_P) \tag{15.3.653}$$

which should be maximized under the constraint $\sum_{i=1}^{n} X_i = 1$ and where Φ is a parameter that represents the degree of risk aversion of investors (in the standardize the relation units...).

The problem of maximization under constraints to determine the maximum of the economic function Z is defined by:

$$Z = \Phi \sum_{i=1}^{n} X_i \mathrm{E}(R_i) - \sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j \mathrm{cov}(R_i, R_j) + \lambda \left(1 - \sum_{i=1}^{n} X_i\right) \tag{15.3.654}$$

This function of $n+1$ variables $(X_1, X_2, \ldots, X_n, \lambda)$ is maximized if its derivative (partial derivative) from each of these variables is zero, which is to put the following system:

$$\begin{cases} \dfrac{\partial Z}{\partial X_1} = \Phi \mathrm{E}(R_1) - 2X_1\mathrm{cov}(R_1, R_1) - 2X_2\mathrm{cov}(R_1, R_2) - \ldots - 2X_n cov(R_1, R_n) - \lambda = 0 \\ \dfrac{\partial Z}{\partial X_2} = \Phi \mathrm{E}(R_2) - 2X_2\mathrm{cov}(R_2, R_1) - 2X_2\mathrm{cov}(R_2, R_2) - \ldots - 2X_n\mathrm{cov}(R_2, R_n) - \lambda = 0 \\ \ldots \\ \dfrac{\partial Z}{\partial X_n} = \Phi \mathrm{E}(R_n) - 2X_1\mathrm{cov}(R_n, R_1) - 2X_2\mathrm{cov}(R_n, R_2) - \ldots - 2X_n\mathrm{cov}(R_n, R_n) - \lambda = 0 \\ \dfrac{\partial Z}{\partial \lambda} = 1 - X_1 - X_2 - \ldots - X_n = 0 \end{cases} \tag{15.3.655}$$

Let us put:

$$\text{cov}(R_i,R_j)=\sigma_{ij} \tag{15.3.656}$$

Then we can write:

$$\begin{cases} 2X_1\sigma_{11}+2X_2\sigma_{12}+\ldots+2X_n\sigma_{1n}+\lambda=\Phi\text{E}(R_1) \\ 2X_2\sigma_{21}+2X_2\sigma_{22}+\ldots+2X_n\sigma_{2n}+\lambda=\Phi\text{E}(R_2) \\ \ldots \\ 2X_1\sigma_{n1}+2X_2\sigma_{n2}+\ldots+2X_n\sigma_{nn}+\lambda=\Phi\text{E}(R_n) \\ X_1+X_2+\ldots+X_n=1 \end{cases} \tag{15.3.657}$$

or in matrix form:

$$\begin{bmatrix} 2\sigma_{11} & 2\sigma_{12} & \ldots & 2\sigma_{1n} & 1 \\ 2\sigma_{21} & 2\sigma_{22} & \ldots & 2\sigma_{2n} & 1 \\ \ldots & & \ldots & \ldots & 1 \\ 2\sigma_{n1} & 2\sigma_{n2} & \ldots & 2\sigma_{nn} & 1 \\ 1 & 1 & \ldots & 1 & 0 \end{bmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \ldots \\ X_n \\ \lambda \end{pmatrix} = \begin{pmatrix} \Phi\text{E}(R_1) \\ \Phi\text{E}(R_2) \\ \ldots \\ \Phi\text{E}(R_n) \\ 1 \end{pmatrix} \tag{15.3.658}$$

And we will put now:

$$A=\begin{bmatrix} 2\sigma_{11} & 2\sigma_{12} & \ldots & 2\sigma_{1n} & 1 \\ 2\sigma_{21} & 2\sigma_{22} & \ldots & 2\sigma_{2n} & 1 \\ \ldots & & \ldots & \ldots & 1 \\ 2\sigma_{n1} & 2\sigma_{n2} & \ldots & 2\sigma_{nn} & 1 \\ 1 & 1 & \ldots & 1 & 0 \end{bmatrix} \qquad B=\begin{pmatrix} \Phi\text{E}(R_1) \\ \Phi\text{E}(R_2) \\ \ldots \\ \Phi\text{E}(R_n) \\ 1 \end{pmatrix} \qquad X=\begin{pmatrix} X_1 \\ X_2 \\ \ldots \\ X_n \\ \lambda \end{pmatrix} \tag{15.3.659}$$

In this case, the system of equations to solve can be summarized in matrix form as:

$$AX=B \tag{15.3.660}$$

Therefore:

$$X=A^{-1}B \tag{15.3.661}$$

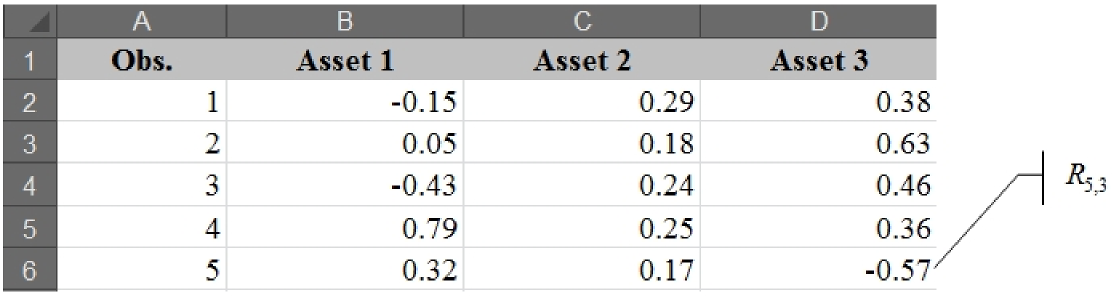
Determining the weight of each of the n assets that enter into the composition of a portfolio thus requires inversion of a square matrix of $n+1$ rows and $n+1$ columns having $(n^2-n)/2$ covariances (the diagonal having variances only and the matrix being symmetrical!). What is relatively long to calculate for large portfolios.

However, even once the weighting of assets is completed, the problem itself is not completely finished. Indeed, we can find the efficient frontier but the customer (investor) will impose a logic constraint relatively to the zero risk level of its portfolio and of the ratio return/maximum risk.

Given the heavy calculations required for the inversion of the matrix A, William F. has proposed a simplified model that we will see after a practical example of the Markowitz model.

✍**Example:**

Let us consider three assets entering a portfolio composition at the same time in equal proportions and n of their return observations (without missing values) R_{ij} entered in Microsoft® Excel® 14.0.7173 (the component j can be seen as a time period):



	A	B	C	D
1	Obs.	Asset 1	Asset 2	Asset 3
2	1	-0.15	0.29	0.38
3	2	0.05	0.18	0.63
4	3	-0.43	0.24	0.46
5	4	0.79	0.25	0.36
6	5	0.32	0.17	-0.57

Figure 15.129: Initial observed values of our portfolio assets

The goal will be now using the native tools of this spreadsheet software to determine the Markowitz portfolio efficiency frontier model and the C.M.L. (see definition and mathematical details further below) and the corresponding weight of assets (optimal weights) that minimizes the variance for a given expected mean of the return knowing the risk-free market rate R_f is equal to 0.22.

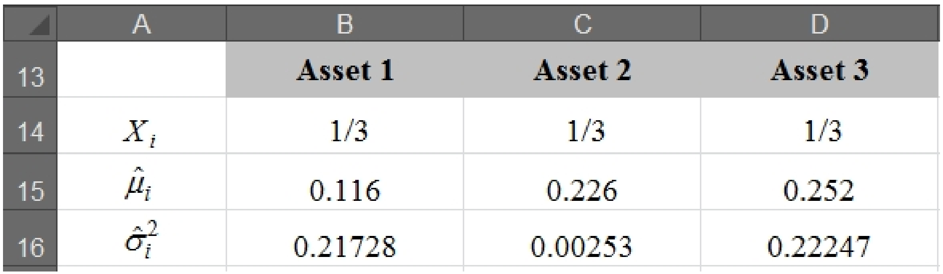
Below the table given above we will create in Microsoft® Excel® the list containing the proportions X_i of the assets (which we assume equidistributed at the beginning such that each weight is equal to $1/3$), we will display the average short return μ_i of each asset calculated using the estimator[a]:

$$\text{E}(R_i)=\hat{\mu}_i=\frac{\displaystyle\sum_{j=1}^{n}R_{ij}}{n}=\begin{cases}\texttt{=AVERAGE(B2:B6)}\\ \texttt{=AVERAGE(C2:C6)}\\ \texttt{=AVERAGE(D2:D6)}\end{cases}\tag{15.3.662}$$

and the variance σ_i^2 calculated for each asset of the estimator (we could also take the semivariance but this is another subject...):

$$\text{V}(R_i)=\hat{\sigma}_i^2=\frac{\displaystyle\sum_{j=1}^{n}(R_{ij}-\hat{\mu}_i)^2}{n-1}=\begin{cases}\texttt{=VAR(B2:B6)}\\ \texttt{=VAR(C2:C6)}\\ \texttt{=VAR(D2:D6)}\end{cases}\tag{15.3.663}$$

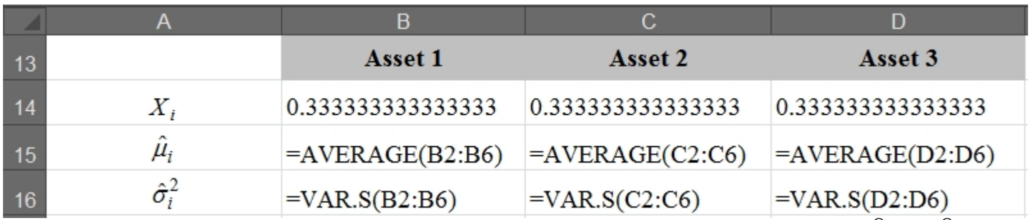
This gives us the following list in Microsoft® Excel® 14.07.7173:



	A	B	C	D
13		**Asset 1**	**Asset 2**	**Asset 3**
14	X_i	1/3	1/3	1/3
15	$\hat{\mu}_i$	0.116	0.226	0.252
16	$\hat{\sigma}_i^2$	0.21728	0.00253	0.22247

Figure 15.130: **Proportions and expected mean/variance estimators of 3 assets**

Thus explicitly:



	A	B	C	D
13		**Asset 1**	**Asset 2**	**Asset 3**
14	X_i	0.333333333333333	0.333333333333333	0.333333333333333
15	$\hat{\mu}_i$	=AVERAGE(B2:B6)	=AVERAGE(C2:C6)	=AVERAGE(D2:D6)
16	$\hat{\sigma}_i^2$	=VAR.S(B2:B6)	=VAR.S(C2:C6)	=VAR.S(D2:D6)

Figure 15.131: **Explicit Proportions and expected mean/variance estimators of 3 assets in Microsoft® Excel® 14.07.7173**

We must now calculate the average return of the portfolio as given by:

$$\text{E}(R_P)=\sum_{i=1}^{n} X_i R_i = X_1\hat{\mu}_1+X_2\hat{\mu}_2+X_3\hat{\mu}_3 \tag{15.3.664}$$

This relation is a bit long to enter in a spreadsheet software as it is, and will be even more if we have a much larger number of assets.

In our case, it is the sum of the products term by term of two cell ranges (X_i and $\hat{\mu}_i$) having the same dimensions (same number of rows and the same number of columns). We can then use the following function in the in Microsoft® Excel® 14.0.7173:

`=SOMMEPROD(B14:D14,B15:D15)`

For the portfolio the variance is a little bit more complicated since we have to calculate:

$$\begin{aligned}\text{V}(R_P)=\sigma^2(R_P)&=\sum_{i=1}^{n} X_i^2\text{V}(R_i)+2\sum_{i,j\mid i<j} X_iX_j\text{cov}(R_i,R_j)\\ &=\sum_{i=1}^{n} X_i^2\sigma_i^2+2\sum_{i,j\mid i<j} X_iX_j\text{cov}(R_i,R_j)\end{aligned} \tag{15.3.665}$$

The relation developed in our special case gives:

$$\begin{aligned}\sigma^2(R_P)=X_1^2\sigma_1^2+X_2^2\sigma_2^2+X_3^2\sigma_3^2+2X_1X_2\text{cov}(R_1,R_2)\\ +2X_1X_3\text{cov}(R_1,R_3)+2X_2X_3\text{cov}(R_2,R_3)\end{aligned} \tag{15.3.666}$$

The trick to apply this in a spreadsheet like Microsoft® Excel® is to use linear algebra and write this relationship in matrix form as we have proved it earlier above:

$$\sigma^2(R_P) = X^T c_{ij} X \tag{15.3.667}$$

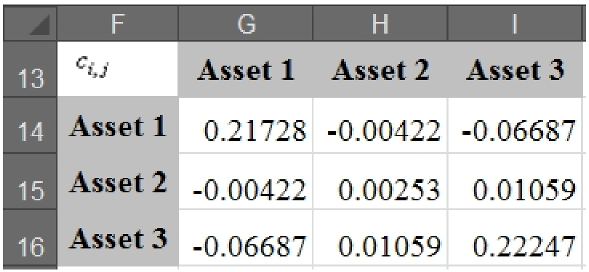
Which is equivalent to write in Microsoft® Excel® 14.0.:

`=SUMPROD(MMULT(B14:D14,G14:I16),B14:D14)`

Either in explicit matrix form:

$$\sigma^2(R_P) = (X_1\; X_2\; X_3) \begin{bmatrix} \sigma_1^2 & \text{cov}(R_2,R_1) & \text{cov}(R_3,R_1) \\ \text{cov}(R_1,R_2) & \sigma_2^2 & \text{cov}(R_3,R_2) \\ \text{cov}(R_1,R_3) & \text{cov}(R_2,R_3) & \sigma_3^2 \end{bmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \tag{15.3.668}$$

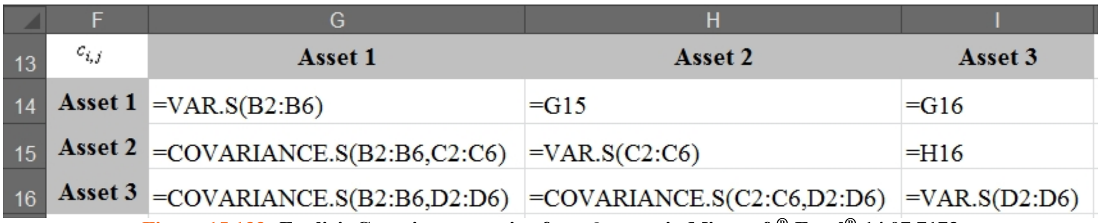
Based on the above tables, it is simple in Microsoft® Excel® to obtain the covariance matrix (in practice it is often considered as difficult to get a robust estimate of the covariance matrix and there are various techniques for this purpose):



	F	G	H	I
13	$c_{i,j}$	Asset 1	Asset 2	Asset 3
14	Asset 1	0.21728	-0.00422	-0.06687
15	Asset 2	-0.00422	0.00253	0.01059
16	Asset 3	-0.06687	0.01059	0.22247

Figure 15.132: Covariance matrix of our 3 assets

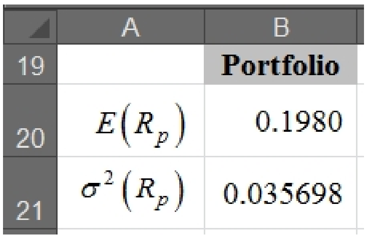
Thus explicitly (the covariance matrix is symmetrical for recall... as seen in the section Statistics at page 604):



	F	G	H	I
13	$c_{i,j}$	Asset 1	Asset 2	Asset 3
14	Asset 1	=VAR.S(B2:B6)	=G15	=G16
15	Asset 2	=COVARIANCE.S(B2:B6,C2:C6)	=VAR.S(C2:C6)	=H16
16	Asset 3	=COVARIANCE.S(B2:B6,D2:D6)	=COVARIANCE.S(C2:C6,D2:D6)	=VAR.S(D2:D6)

Figure 15.133: Explicit Covariance matrix of our 3 assets in Microsoft® Excel® 14.07.7173

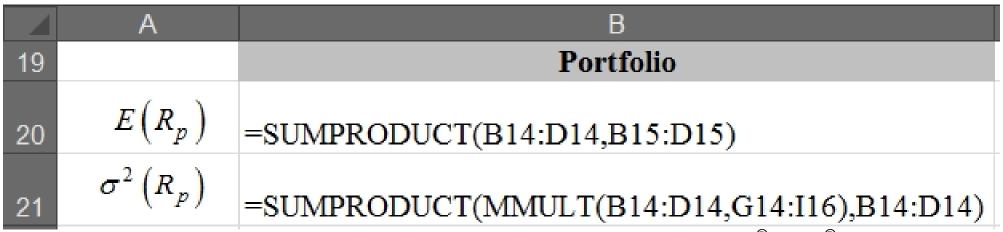
And for the expected mean and variance of the portfolio we have the following table:



	A	B
19		Portfolio
20	$E(R_p)$	0.1980
21	$\sigma^2(R_p)$	0.035698

Figure 15.134: Portfolio Mean and Variance estimators

thus applying the above relations:

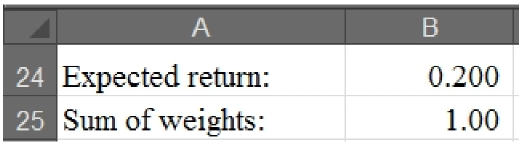


	A	B
19		**Portfolio**
20	$E(R_p)$	=SUMPRODUCT(B14:D14,B15:D15)
21	$\sigma^2(R_p)$	=SUMPRODUCT(MMULT(B14:D14,G14:I16),B14:D14)

Figure 15.135: Explicit Portfolio Mean and Variance estimators in Microsoft® Excel® 14.07.7173

The problem now is to determine for a given fixed return of the portfolio (cell `B20`), the different proportions X_i of the asset that minimize the risk.

After having added the both cells `B24` (expected return of the portfolio) and `B25` (total number of proportions of assets):

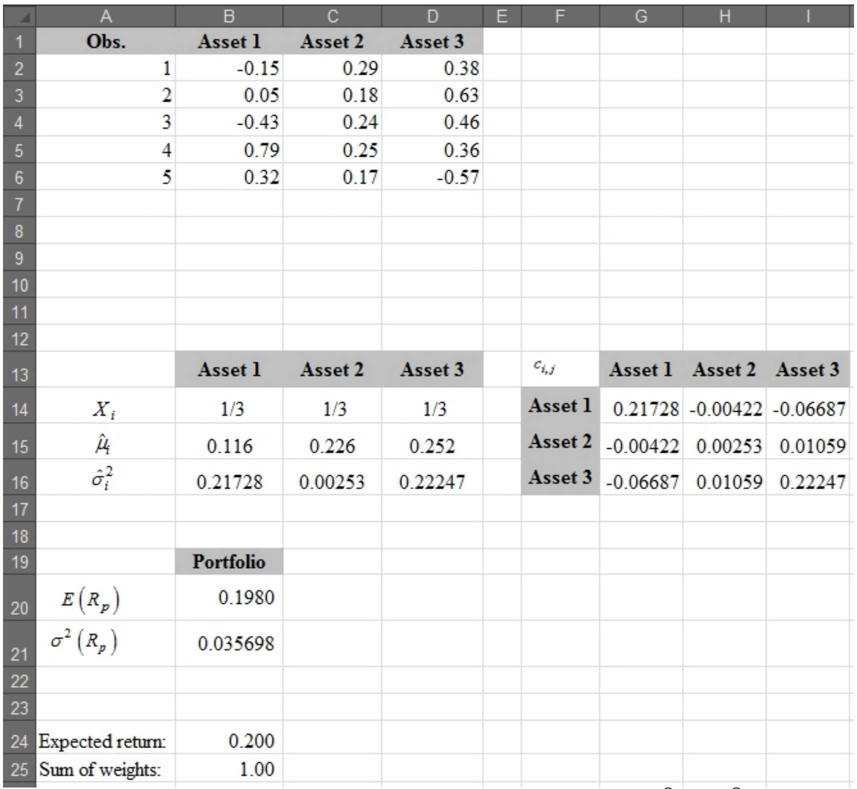


	A	B
24	Expected return:	0.200
25	Sum of weights:	1.00

We must now solve the non-linear optimization problem:

$$\begin{aligned} &\underset{\sigma^2(R_P)}{\text{minimize}} && \mathrm{E}(R_P)=0.2 \\ &\text{subject to} && \sum X_i = 1 \end{aligned} \tag{15.3.669}$$

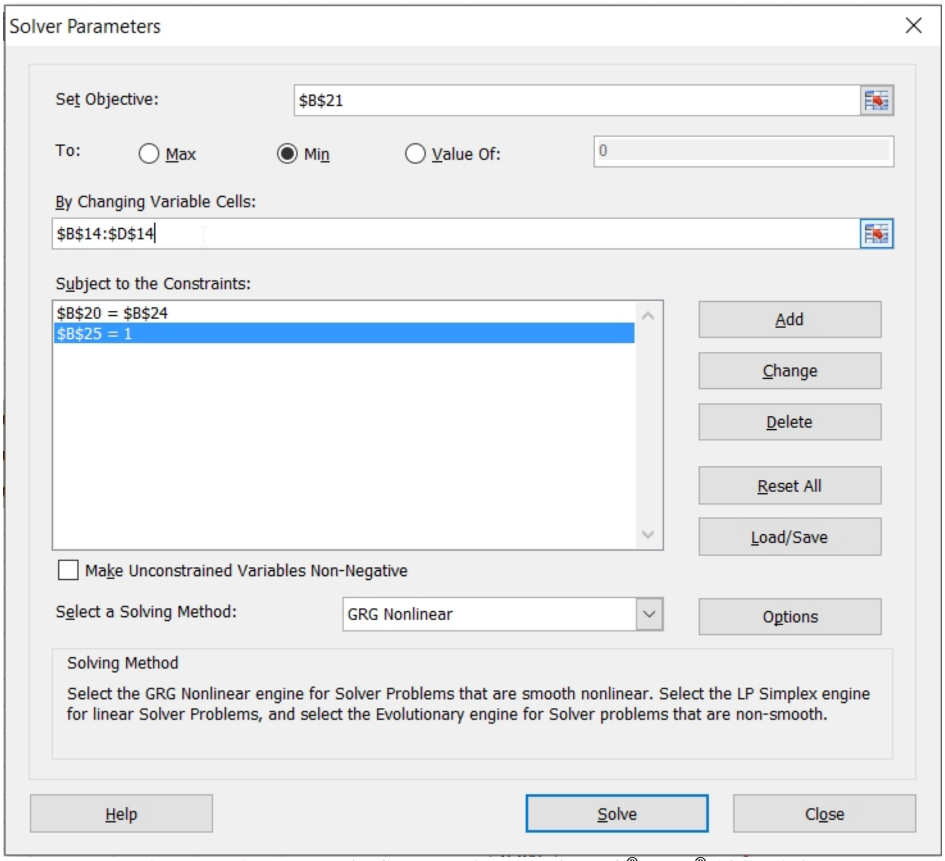
A global view of our sheet gives so far:



	A	B	C	D	E	F	G	H	I
1	Obs.	Asset 1	Asset 2	Asset 3					
2	1	-0.15	0.29	0.38					
3	2	0.05	0.18	0.63					
4	3	-0.43	0.24	0.46					
5	4	0.79	0.25	0.36					
6	5	0.32	0.17	-0.57					
7									
8									
9									
10									
11									
12									
13		Asset 1	Asset 2	Asset 3		$c_{i,j}$	Asset 1	Asset 2	Asset 3
14	X_i	1/3	1/3	1/3		Asset 1	0.21728	-0.00422	-0.06687
15	$\hat{\mu}_i$	0.116	0.226	0.252		Asset 2	-0.00422	0.00253	0.01059
16	$\hat{\sigma}_i^2$	0.21728	0.00253	0.22247		Asset 3	-0.06687	0.01059	0.22247
17									
18									
19		Portfolio							
20	$E(R_p)$	0.1980							
21	$\sigma^2(R_p)$	0.035698							
22									
23									
24	Expected return:	0.200							
25	Sum of weights:	1.00							

Figure 15.136: Markowitz manual resolution configuration in Microsoft® Excel® 14.07.7173

And now we launch the solver:



Solver Parameters

Set Objective: `$B$21`

To: ○ Max ● Min ○ Value Of: 0

By Changing Variable Cells: `$B$14:$D$14`

Subject to the Constraints:

`$B$20 = $B$24`
`$B$25 = 1`

☐ Make Unconstrained Variables Non-Negative

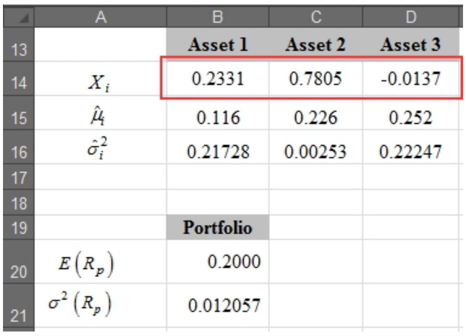
Select a Solving Method: GRG Nonlinear

Solving Method

Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

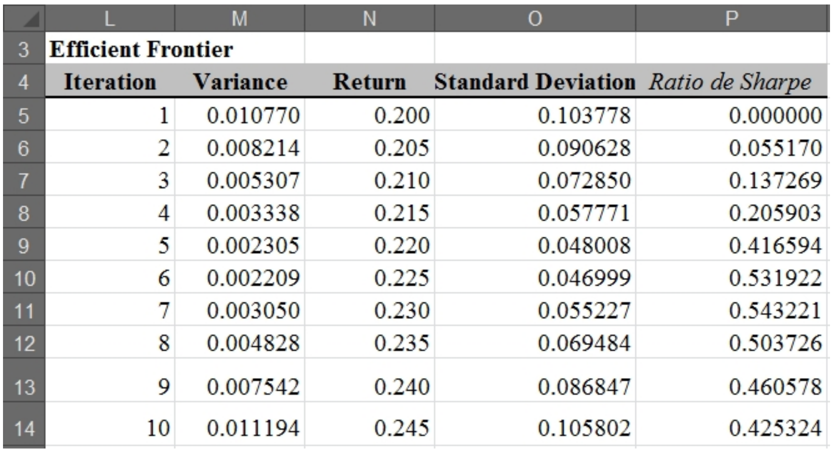
Figure 15.137: Markowitz Solver settings in Microsoft® Excel® 14.07.7173

This gives us straight for solution (the same as the one given by R 3.0.2 as the reader can check it in the corresponding companion book):



	A	B	C	D
13		Asset 1	Asset 2	Asset 3
14	X_i	0.2331	0.7805	-0.0137
15	$\hat{\mu}_i$	0.116	0.226	0.252
16	$\hat{\sigma}_i^2$	0.21728	0.00253	0.22247
17				
18				
19		Portfolio		
20	$E(R_p)$	0.2000		
21	$\sigma^2(R_p)$	0.012057		

Now to determine the efficient Markowitz frontier we report manually (or with VBA) for different values of the target mean, the corresponding minimum variance. That gives us the following list:



	L	M	N	O	P
3	**Efficient Frontier**				
4	**Iteration**	**Variance**	**Return**	**Standard Deviation**	***Ratio de Sharpe***
5	1	0.010770	0.200	0.103778	0.000000
6	2	0.008214	0.205	0.090628	0.055170
7	3	0.005307	0.210	0.072850	0.137269
8	4	0.003338	0.215	0.057771	0.205903
9	5	0.002305	0.220	0.048008	0.416594
10	6	0.002209	0.225	0.046999	0.531922
11	7	0.003050	0.230	0.055227	0.543221
12	8	0.004828	0.235	0.069484	0.503726
13	9	0.007542	0.240	0.086847	0.460578
14	10	0.011194	0.245	0.105802	0.425324

Figure 15.138: Markowitz efficient frontier values samples

This gives us the following $\text{E}(R_P)=f(\sigma_{R_p})$ plot:

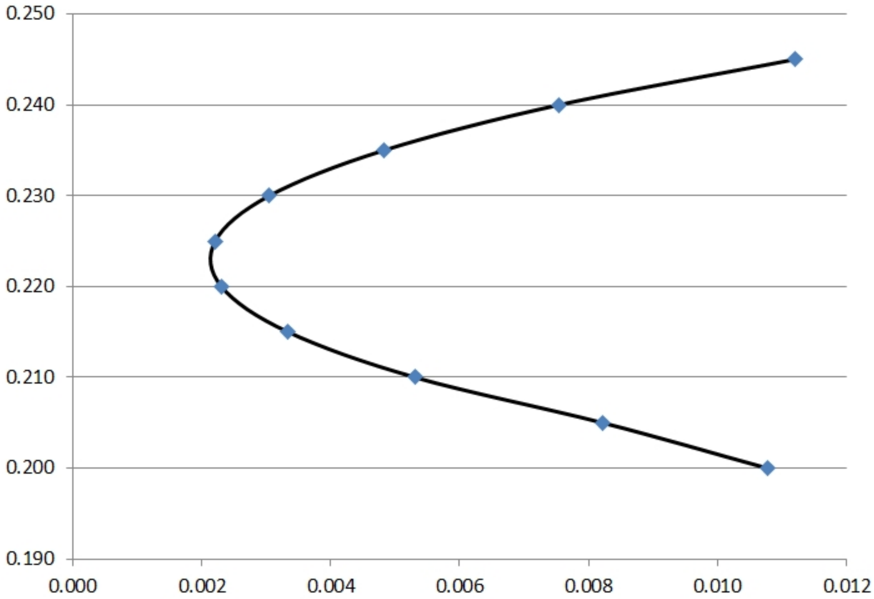


Figure 15.139: Markowitz efficient frontier plot in Microsoft® Excel® 14.07.7173

If the reader refers to our `R` companions book, the equivalent result for the efficient frontier (EWP: Equally Weighted Portfolio) is:

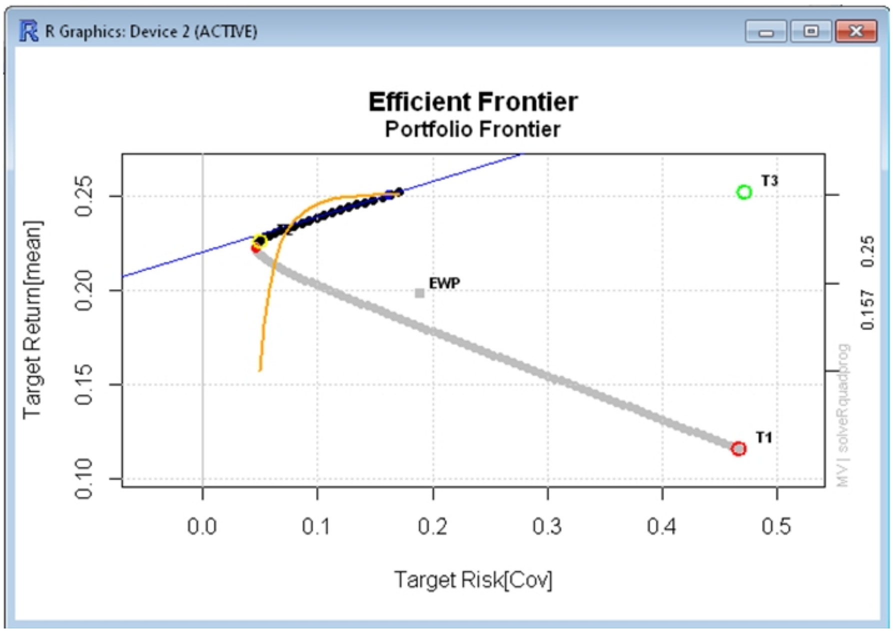


Figure 15.140: Markowitz efficient frontier plot in R 3.0.2

Or still with R always by following the step given in the companion book we can plot the evolution of the return and risk in function of the weights:

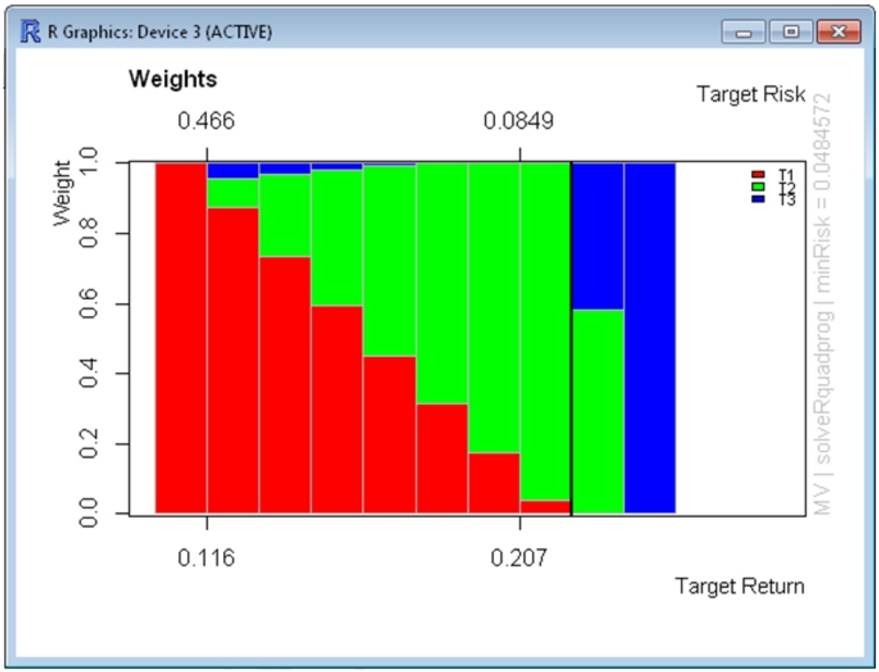


Figure 15.141: Markowitz asset weight plot in R 3.0.2

We will not this time give the results obtained with MATLAB® as the output of that latter are plotted in quite an awful way... (but the reader can see the corresponding companion book where there are given!).

---

[a] Obviously we know that this is not the best way as the arithmetic mean overestimated the real mean compared to geometric mean as $\mu_a > \mu_g$

So this is a nice little application example in a software accessible to almost everyone but that is relatively false in reality (but at least it is educational)!

However remember the 4 construction assumptions of this model that makes its use quite limited in practice...:

H1. The mean-variance optimization model is equivalent to model the assets of the portfolio by random variables for which these two moments exist. But in reality, nothing says that this is the case.

H2. The mean-variance optimization model assumes that only the mean and variance are important for an investor. But in reality we should take into account the skewness (or even the kurtosis) and maybe use the semi-variance instead of the variance.

H3. The model is build on a posteriori information without taking into account how the future could be. So an a priori information of the market should complete this model.

H4. The model is based on average return estimated for an allocation period that may change in the short-term. But a very small change in average return can have disproportionate impact on the proportions.

Also in practice we must be careful to one thing also.... Rebalancing a portfolio following an optimization calculation can be very expensive in transaction fees (a portfolio rebalancing computer model that doesn't take into account transaction fees is said to be a "frictionless portfolio model"). A case often cited in university is that of a US pension fund consisting of more than 2, 000 different types of assets whose rebalancing in October 12000 (holocene calendar) implied 40 portfolio managers, 500 million of assets for a total of 17.5 billion dollars!!! The legends says that the cost of this rebalancing in terms of transactions cost was not far from 120 million dollars...!

Remark

A big issue to the Markowitz model is the important sensitivity of the resulting portfolios if the parameters or inputs to the models undergo small changes. Multiple independent investigations has been able to provide strong evidence that two efficient frontiers whose parameters or inputs are extracted from the same distribution of returns but simply where the sampling period is different we get clearly statistically different efficient frontiers. R. O. Michaud et al. in their textbook *A practical guide to stock portfolio optimization and asset allocation* claim that « *In practice, the most important limitations of MV optimization are instability and ambiguity. [...] Small changes in input assumptions often imply large changes in the optimized portfolio. Consequently, portfolio optimality is often not well defined... The procedure overuses statistically estimated information and magnifies the impact of estimation errors. The result is that optimized portfolios are "error maximized" and often have little, if any, reliable investment value. Indeed, an equally weighted portfolio may often be substantially closer to true MV optimality than an optimized portfolio.* ». That's because of these limitations of the MV portfolio that they created a resampling portfolio method (see the practical example in our `R` companion book!).

Now based on exactly the same method we used in the section of Theoretical Computing with the Lagrange multipliers (see section Numerical Methods page 6670), we now introduce a risk free asset into the framework. Let R_f and w_f represent the return and weight of the risk free asset, respectively. Let $\vec{Z}$ represent the vector of excess return over the risk free rate. then using exactly the same reasoning we get:

$$\begin{aligned}&R_P=\vec{\mu}^T\vec{w}+\left(1-\vec{1}^T\vec{w}\right)R_f=R_f+\left(\vec{\mu}^T-R_f\vec{1}^T\right)\vec{w}=R_f+\vec{Z}^T\vec{w}\\&\sigma_P^2=\vec{w}^T\Sigma\vec{w}\end{aligned}\quad(15.3.670)$$

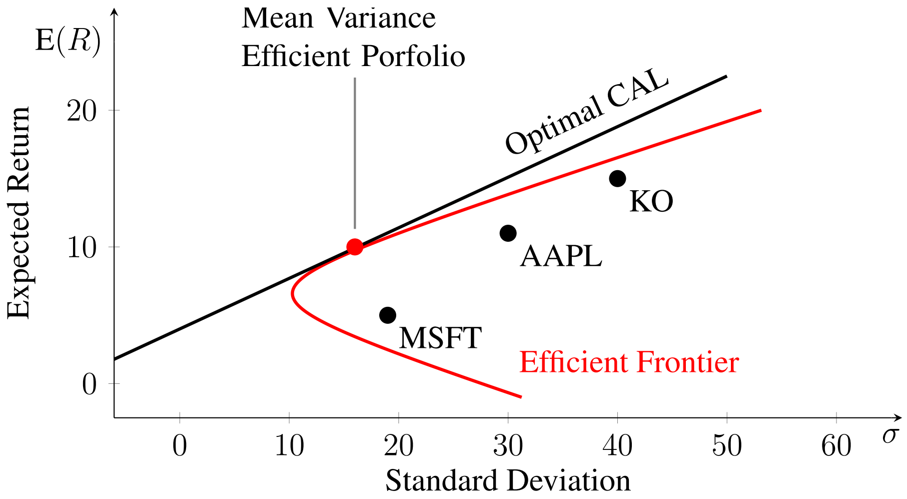


Figure 15.142: **Mean-Variance efficient portfolio**

The portfolio optimisation problem is then again to minimise the variance of the portfolio for a given level of return but there is no budget constraints as the presence of risk free assets allows unlimited borrowing and lending (we already put a factor $1/2$ as we know that an appearing factor of 2 will anyway cancel it later):

$$L(\vec{w},\lambda)=\frac{1}{2}\vec{w}^T\Sigma\vec{w}+\lambda\left(R_P-R_f-\vec{Z}^T\vec{w}\right)\quad(15.3.671)$$

Differentiating and setting the derivatives equal to zero, we get:

$$\begin{aligned}\frac{\partial L(\vec{w},\lambda)}{\partial\vec{w}}&=\Sigma\vec{w}-\lambda\vec{Z}=\vec{0}\\\frac{\partial L(\vec{w},\lambda)}{\partial\lambda}&=R_P-R_f-\vec{Z}^T\vec{w}=0\end{aligned}\quad(15.3.672)$$

We need to solve the two equations for the two unknowns, with $\vec{w}$ being the ultimate price. Solving the first equation for $\vec{w}$, we get:

$$\begin{aligned}&\Sigma\vec{w}-\lambda\vec{Z}=\vec{0}\\&\vec{w}=\lambda\Sigma^{-1}\vec{w}\end{aligned}\quad(15.3.673)$$

And substituting into the second equation:

$$\begin{aligned}&R_P-R_f-\vec{Z}^T\vec{w}=\vec{0}\\&\Rightarrow R_P-r_f-\lambda\vec{Z}^T\Sigma^{-1}\vec{Z}=0\\&\Rightarrow\lambda=\frac{R_P-R_f}{\vec{Z}^T\Sigma^{-1}\vec{Z}}\end{aligned}\quad(15.3.674)$$

Thus:

$$\vec{w}_{\text{MP}} = \lambda\Sigma^{-1}\vec{Z} = \frac{R_P - R_f}{\vec{Z}^T\Sigma^{-1}\vec{Z}}\Sigma^{-1}\vec{Z} = \frac{R_P - R_f}{\left(\vec{\mu}^T - R_f\vec{1}^T\right)^T\Sigma^{-1}\left(\vec{\mu}^T - R_f\vec{1}^T\right)}\Sigma^{-1}\left(\vec{\mu}^T - R_f\vec{1}^T\right) \tag{15.3.675}$$

This portfolio made of $\vec{w}_{\text{MP}}$ risked assets is named "market portfolio".

15.3.10.3.7.2 Overall maximum Sharpe portfolio

The use of the Markowitz model, as proposed in his 11959 (holocene calendar) book, raised many problems when it came to use algorithms for a basic list with a large number of values. These problems were of two orders:

1. The size of matrices required at that time a huge computer capacity and a fairly long period of calculation!
2. The use of the base model required that we know in its entirety the covariance matrix. The main problem about it is in the number of estimates to provide to build it and also the difficulty of making accurate and consistent estimates.

If we want the approach proposed by Markowitz can enter the field of application, we must obviously find a way to significantly simplify the procedure while losing as little as possible of the rigour of the method.

In 11963 (holocene calendar), William Sharpe proposed a solution whose essential characteristic consists in the assumption that the returns of the various values are only linked by their common relation with an underlying basic factor (stock index typically) which allows determining a coefficient named the "beta" (correlation like factor between the return of a security and that of a market portfolio).

This purely empirical hypothesis named "one index model" (or "unifactor model", "monofactorial model") has later take considerable importance because it was, as we shall see in subsequent developments, at the basis of the theory of pricing for some financial assets in an uncertain Universe.

Remark

Again, the developments that follow should look abstract but ... we'll see how to solve the previous above example with Microsoft® Excel® for the Markowitz model but applying the Sharpe's model and we will then even be able to visually compare the results of both methods.

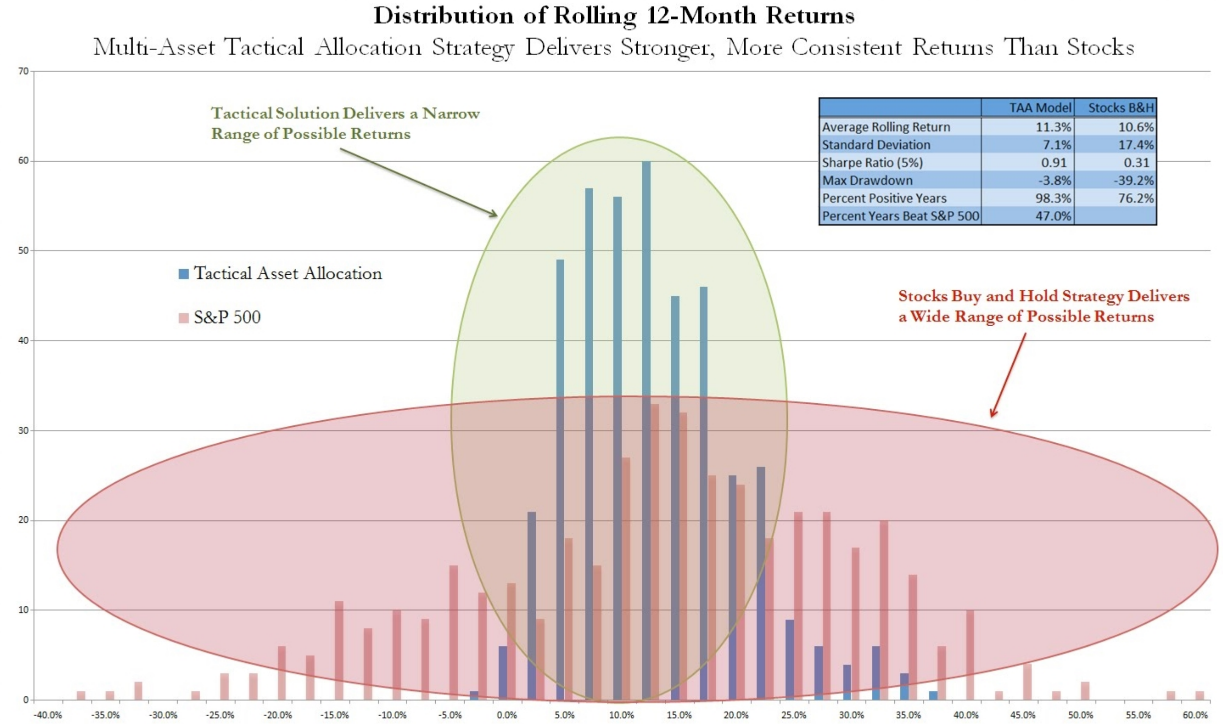
The term "unifactor model" comes from the fact that the basic purpose of the Sharpe model is to define the performance of a financial investment in terms of its non-diversifiable risk, assimilated to that of the market risk (or "systematic risk") given by a number named the "beta" as previously mentioned.

Investors and managers distinguish three main kinds of risks:

1. The "specific risk" relative (implicit) to the asset himself (its variance) also named "unsystematic risk" or "idiosyncratic risk".
2. The "systematic/nondiversifiable risk" relatively to the economy / market in the broadest sense of the term (variance of the market reference portfolio).
3. The "global risk" which is somehow the sum of both (it's a little more subtle than a simple sum if fact...).

As you have probably guessed, the risk factor is difficult to quantify. The element that will help determine the variation is the performance of the asset compared to the change of market performance as a whole. Financial assets whose prices change frequently and whose volatility is high present therefore a high risk if the market is almost stable.

Here are for example distributions of two investment funds:



	TAA Model	Stocks B&H
Average Rolling Return	11.3%	10.6%
Standard Deviation	7.1%	17.4%
Sharpe Ratio (5%)	0.91	0.31
Max Drawdown	-3.8%	-39.2%
Percent Positive Years	98.3%	76.2%
Percent Years Beat S&P 500	47.0%	

Figure 15.143: Multi-Asset moving average strategy VS One index portfolio (source: Faber (12009), Butler|Philbrick & Associates (12010))

Here, is another example of return distributions of the second retirement fund (LPP) in Switzerland depending on the portfolio diversification strategy (the values after the abbreviation LPP represent the % of shares in the portfolio):

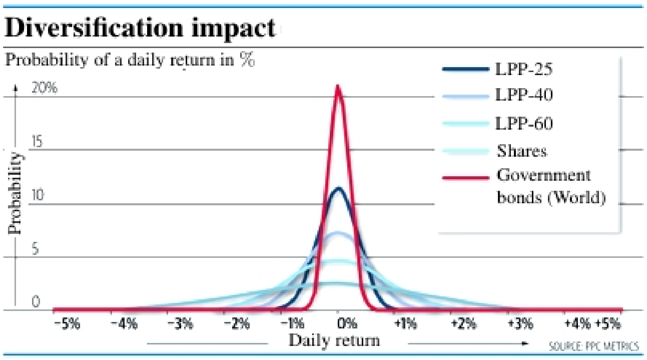


Figure 15.144: Swiss LPP various diversification returns

Definition (naive version #693): The "beta" measures the dependence (ratio) of the performance of a portfolio or of a financial asset relatively to the performance of a benchmark index and is the slope of a line named "security characteristic line" (SCL):

$$\beta = \frac{\text{expected variation of portfolio's}}{\text{variation of index level}} \tag{15.3.676}$$

This ratio is of course especially accurate when the past data are on the long run and that the sampling frequency is small. This ratio is also sometimes named "relative volatility".

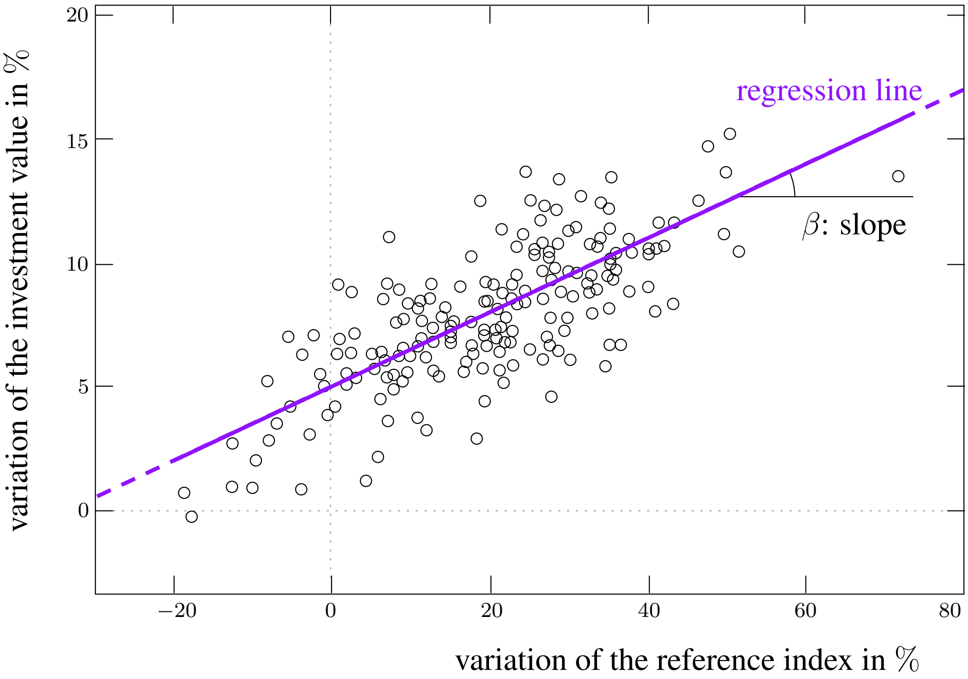
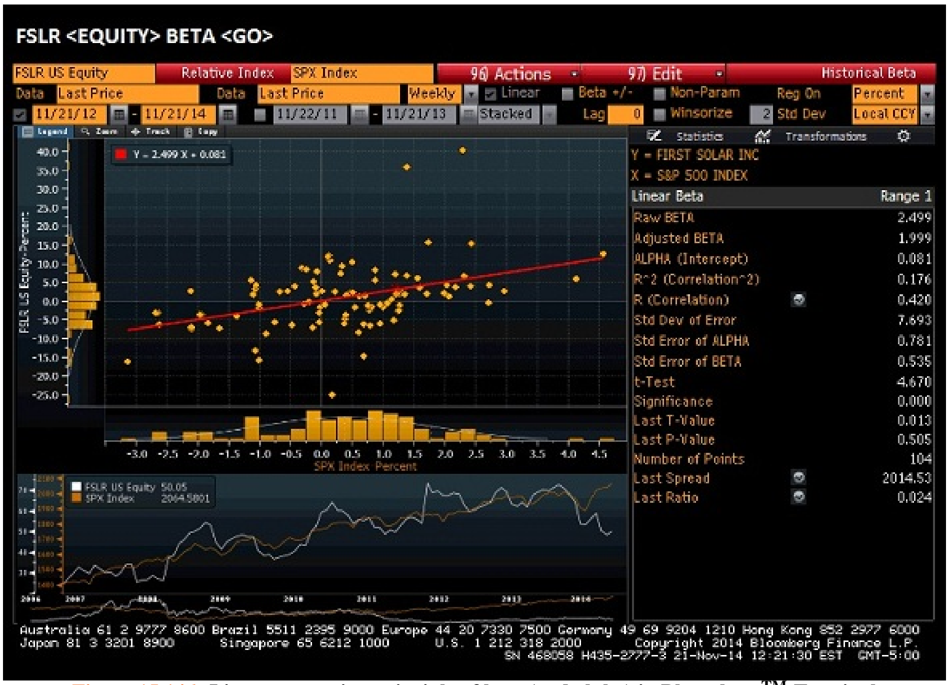


Figure 15.145: Idea of the linear regression behind the β coefficient

Remark

The benchmark index is chosen in the most relevant possible way with everything that this implies... If possible when the return of the reference index is zero, the change in value of the portfolio or asset should be zero.

Or a more realistic example (some informations visible in this screenshot will be discussed later):


FSLR US Equity
Relative Index
SPX Index
96) Actions
97) Edit
Historical Beta
Data
Last Price
Data
Last Price
Weekly
Linear
Beta +/-
Non-Param
Reg On
Percent
11/21/12 - 11/21/14
11/22/11 - 11/21/13
Stacked
Lag 0
Winsorize
2 Std Dev
Local CCY
Legend
Zoom
Track
Copy
Statistics
Transformations
Y = 2.499 X + 0.081
FSLR US Equity-Percent
SPX Index Percent
FSLR US Equity 50.05
SPX Index 2064.5801

FSLR <EQUITY> BETA <GO>

Y = FIRST SOLAR INC
X = S&P 500 INDEX

Linear Beta	Range 1
Raw BETA	2.499
Adjusted BETA	1.999
ALPHA (Intercept)	0.081
R^2 (Correlation^2)	0.176
R (Correlation)	0.420
Std Dev of Error	7.693
Std Error of ALPHA	0.781
Std Error of BETA	0.535
t-Test	4.670
Significance	0.000
Last T-Value	0.013
Last P-Value	0.505
Number of Points	104
Last Spread	2014.53
Last Ratio	0.024

Australia 61 2 9777 8600 Brazil 5511 2395 9000 Europe 44 20 7330 7500 Germany 49 69 9204 1210 Hong Kong 852 2977 6000
Japan 81 3 3201 8900 Singapore 65 6212 1000 U.S. 1 212 318 2000 Copyright 2014 Bloomberg Finance L.P.
SN 468058 H435-2777-3 21-Nov-14 12:21:30 EST GMT-5:00

Figure 15.146: **Linear regression principle of beta (and alpha) in Bloomberg$^{\text{TM}}$ Terminal**

Definitions (#15.3.694): The beta of a stock can be presented as either an Adjusted Beta or a Raw Beta:

D1. A "raw beta" is obtained from the linear regression of a stock's historical data. Raw beta, also known as "historical beta", is based on the observed relationship between the security's return and the returns of an index.

D2. The "adjusted beta" is an estimate of a security's future beta. Adjusted Beta is initially derived from historical data, but modified by the assumption that a security's true beta will move towards the market average, of 1, over time. The formula used to adjust the beta is according to Blume's Method (used by Bloomberg Terminal):

$$\beta_{\text{adj.}}=\frac{2}{3}\beta+\frac{1}{3} \qquad (15.3.677)$$

A simple analysis of the graph above (i.e. elementary functional analysis) therefore shows that a beta equal to 1 for a given asset means that an increase (respectively decrease) of 10% of the market reference index for a certain period will result in an increase (respectively decrease) of 10% on average performance of this asset. So the volatility of the asset is equal to that of the index.

A beta greater than 1 means that the evolution of the return of the financial asset is more volatile (or rather, was volatile, as this coefficient usually refers to a past period) than the return of the market reference index, while a beta less than 1 indicates the opposite. Thus, a fund with a beta of 1.15 is 15% more volatile than the index. Conversely, a fund with a beta of 0.70 is 30% less volatile than the index.

So to summarize:

- An investment with no risk (relatively to the benchmark) therefore displays a beta of 1.
- A beta of less than 1 indicates that if the market (benchmark) is down (bearish), the investment will likely to decline less than the market.
- A beta greater than 1 indicates that if the market (benchmark) is on the rise (bullish), the investment will likely follow a slower upward trend.

The concept of beta having been naively introduced, let us move now to the theory of the model that aims to simplify the Markowitz one using this famous coefficient!

First, by choice, we take the overall beta of a portfolio as being determined from the respective weighted betas of each of the asset or underlying betas that it is make up such that:

$$\boxed{\beta_P=\sum_{i=1} nX_i\beta_i} \tag{15.3.678}$$

with β_P being the beta of the overall portfolio, X_i the proportion of security i in portfolio P, β_i the beta of asset i and n is obviously the number of financial assets in the portfolio (the beta is then additive when weighted).

Sharpe suggested (in 11963 according to holocene calendar) that the return R_i of each asset i at time t is given by the linear regression (see section Numerical Methods page 6324) determining the "security characteristic line" view above (so it is a linear regression model with one explanatory factor but in practice there are obviously multiple linear or non-linear models):

$$\boxed{R_i=\alpha_i+\beta_i I+\varepsilon_i} \tag{15.3.679}$$

named also sometimes "single-factor benchmark of returns", where:

- I is the performance of a given reference economic index ... (stock index, index of gross national product, price levels or even portfolio market performance itself...) at time t and is the dependent variable of the univariate regression (in the terminology used in the section of Numerical Methods) considered as a random variable[25] (...).
- α_i,β_i are unbiased estimators (see section Statistics page 818) of the parameters specific to that value. The first term in finance is named "alpha coefficient" and is simply the intercept of the regression (the return of the assets when the performance of the benchmark is zero or when the market has zero returns often also named "residual return") and the second parameter is simply to recall the "beta" of the risky asset i. The alpha is not a source of risk, it is often ignored in the calculations.
- ε_i is a random variable assumed to be characterized by a zero mean, and variance equal to a constant (see section Numerical Methods page 6348) and the different ε_i are assumed uncorrelated between them (zero covariance).

---

[25] We know that is this case we should rather use the Deming regression seen also in the section of Numerical Methods

About the level of the index I, it will be characterized by the relation (to simplify later some developments):

$$I = \alpha_{n+1} + \upsilon_{n+1} \tag{15.3.680}$$

where α_{n+1} is an additional unbiased parameter to characterize index I and υ_{n+1} a random variable characterized by a zero mean and variance equal to a constant.

Notice first that (by adopting on the way the traditional notation for the performance of the market index) for an asset:

$$\begin{aligned} &R_i = \alpha + \beta_i I + \varepsilon = \alpha + \beta_i R_M + \varepsilon_i \\ &\Rightarrow \mathrm{V}(R_i) = \beta_i^2 \mathrm{V}(R_m) + \mathrm{V}(\varepsilon_i) \\ &\Rightarrow \underbrace{\sigma_i}_{\text{total risk}} = \sqrt{\underbrace{\beta_i^2\sigma_M^2}_{(\text{systematic risk})^2} + \underbrace{\sigma_\varepsilon^2}_{(\text{specific risk})^2}} \end{aligned} \tag{15.3.681}$$

This is easily generalized to a sum of weighted assets for a portfolio and where we get also as a standard deviation of the portfolio is the sum of two terms, i.e. systematic risk and specific risk (diversifiable):

$$\boxed{\sigma_P = \sqrt{\sum_i \sum_j w_i w_j \beta_i \beta_j \sigma_M^2 + \sum_i \sum_j w_i w_j \sigma_{\varepsilon_{ij}}}} \tag{15.3.682}$$

Of course if the number of assets approaches infinity then the weights approach zero and therefore the total risk also tends to zero and this is the trick of multi-asset portfolio management!!

Remark

The SML is sometimes confused with the "Capital Market Line" (CML). The SML is derived from the CML as it is a special case of it obviously! While the CML shows the rates of return for a specific portfolio, the SML represents the market's risk and return at a given time, and shows the expected returns of individual assets.

Optimal portfolio allocation lies on the Capital Allocation Line (CAL) such that it maximizes the utility of the investor. Thus investor's utility curve is tangent on that point to the CAL:

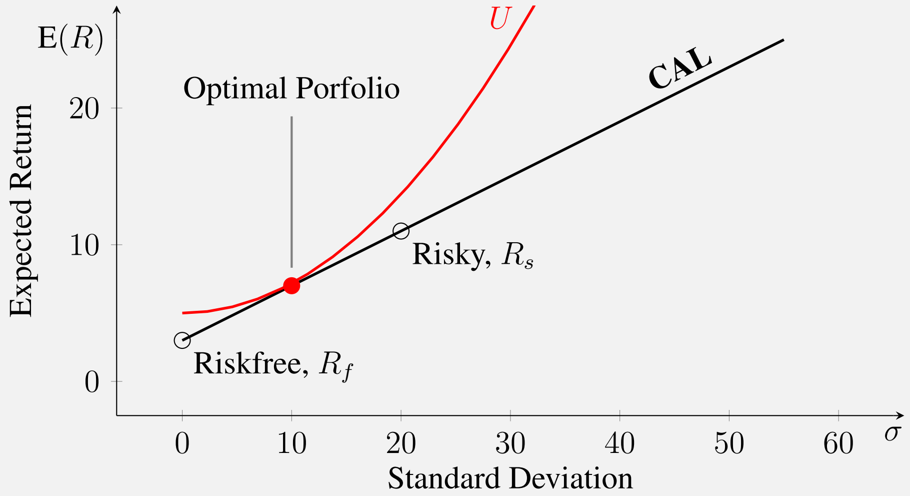


Figure 15.147: **Optimal portfolio allocation and Capital Allocation Line**

Notice that in econometrics the above relation is sometimes rewritten as:

$$\sigma_P^2=\underbrace{\left(\underbrace{\sum_{i=1}^n w_i^2\beta_i^2\sigma_M^2}_{\text{variance}}+\underbrace{\sum_{i=1}^n\sum_{\substack{j=1\\j\neq 1}}^n w_iw_j\beta_i\beta_j\sigma_M^2}_{\text{covariance}}\right)}_{\text{Market proportional}}+\underbrace{\left(\underbrace{\sum_{i=1}^n w_i^2\sigma_{\varepsilon_i}^2}_{\text{variance}}\right)+\sum_{i=1}^n\sum_{\substack{j=1\\j\neq 1}}^n w_iw_j\sigma_{\varepsilon_{ij}}}_{\text{Firm specific}} \quad (15.3.683)$$

The firm specific covariance is sometimes assumed zero. So we can write:

$$\sigma_P^2=\sum_{i=1}^n w_i^2(\beta_i^2\sigma_M^2+\sigma_{\varepsilon_i}^2)+\sum_{i=1}^n\sum_{\substack{j=1\\j\neq 1}}^n w_iw_j\beta_i\beta_j\sigma_M^2 \quad (15.3.684)$$

To summarize the main points, the simple linear regression model of financial asset yield is based on the following key assumptions:

H1. The return model is written in general (univariate case):

$$R_i=\alpha_i+\beta_iI+\varepsilon_i \quad (15.3.685)$$

assuming that we don't do any errors on the linear assumption of the model, nor the list of regressors (see other models below taking into account other factors).

H2. We assume that the perturbation ε satisfies the same assumption as the usual general linear models (Gaussian or others) as the zero mean, zero covariance, fixed variance over time which has to be said is a quite dangerous practice but useful assumption to be usable (and understandable) by practitioners of finance...

H3. For any sample of size n, we use the maximum likelihood estimators (see section Statistics page 818) for the mean and variance of returns on financial assets of the reference portfolio (index):

$$\hat{\sigma}_j=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(R_{ij}-\hat{\mu})^2}\qquad \hat{\mu}_j=\frac{1}{n}\sum_{i=1}^{n}R_{ij}\tag{15.3.686}$$

These assumptions now stated, we also use the results obtained in the section of Numerical Methods on the linear regression to get the beta. We have proved that there are several ways of doing a linear regression, one is to use the covariance and mean. By adopting the notation of econometrics, the slope of the regression (beta) can then be written:

$$\boxed{\beta_i=\frac{\text{cov}(R_i,R_I)}{\text{V}(R_I)}=\frac{\text{cov}(R_i,R_I)}{\sigma_I^2}}\tag{15.3.687}$$

giving the rigorous definition of the beta coefficient according to the Sharpe model where R_i is the return of the financial asset and R_I that of the market reference index (or market reference portfolio).

Now, considering the same assumption as in the Markowitz model, the return R_P of portfolio is defined again quite logically by:

$$R_P=\sum_{i=1}^{n}X_iR_i\tag{15.3.688}$$

where for recall X_i is the proportion of the asset i in the multi-asset portfolio P.

If the return are not explicitly known in practice, then we use the linear model:

$$R_P=\sum_{i=1}^{n}X_iR_i=\sum_{i=1}^{n}X_i(\alpha_i+\beta_iI+\varepsilon_i)=\sum_{i=1}^{n}X_i(\alpha_i+\varepsilon_i)+\sum_{i=1}^{n}X_i\beta_iI\tag{15.3.689}$$

Therefore by using the properties of the mean:

$$\text{E}(R_P)=\text{E}\left[\sum_{i=1}^{n}X_i(\alpha_i+\varepsilon_i)+\sum_{i=1}^{n}X_i\beta_iI\right]=\sum_{i=1}^{n}X_i\text{E}(\alpha_i)+\sum_{i=1}^{n}X_i\text{E}(\varepsilon_i)+\sum_{i=1}^{n}X_i\beta_i\text{E}(\alpha_{n+1}+\nu_{n+1})\tag{15.3.690}$$

Let us put to simplify the notation that:

$$X_{n+1}=\sum_{i=1}^{n}X_i\beta_i\tag{15.3.691}$$

In this case, by the two of the various hypothesis of the model ($\text{E}(\varepsilon_i)=0$ and $\text{E}(\nu_{n+1})=0$):

$$\text{E}(R_P)=\sum_{i=1}^{n}X_i\alpha_i+0+X_{n+1}\text{E}(\alpha_{n+1}+\nu_{n+1})=\sum_{i=1}^{n}X_i\alpha_i+0+X_{n+1}\alpha_{n+1}=\sum_{i=1}^{n+1}X_i\alpha_i\tag{15.3.692}$$

Finally:

$$\boxed{\mathrm{E}(R_P)=\sum_{i=1}^{n+1}X_i\alpha_i} \tag{15.3.693}$$

If the returns are explicitly given and therefore known, the expected mean is calculated with:

$$\boxed{\mathrm{E}(R_P)=\mathrm{E}\left(\sum_{i=1}^{n}X_iR_i\right)=\sum_{i=1}^{n}X_i\mathrm{E}(R_i)} \tag{15.3.694}$$

This relation is often named "weighted average return" or "weighted average yield".

As the customer (investor) will often seek to maximize the expected mean return while minimizing the variance (risk) it remains to determine the latter. Since now we assume explicitly known the returns of the financial assets and the market portfolio returns (index) we have:

$$\begin{aligned}\mathrm{V}(R_P)=\mathrm{V}\left[\sum_{i=1}^{n}X_iR_i\right]&=\sum_{i=1}^{n}X_i^2\mathrm{V}(R_i)+2\sum_{i,j|i<j}X_iX_j\mathrm{cov}(R_i,R_j)\\&=\sum_{i=1}^{n}X_i^2\mathrm{V}(R_i)+2\sum_{i,j|i<j}X_iX_j\mathrm{cov}(\alpha_i+\beta_iI,\alpha_j+\beta_jI)\end{aligned} \tag{15.3.695}$$

Hypothesis: If the index I is selected correctly, when $I=0$ we must have $R_i=0$ implying $\alpha_i=0$ (this is a strong assumption that leads to have an approximation!).

Therefore:

$$\begin{aligned}\mathrm{V}(R_P)&=\sum_{i=1}^{n}X_i^2\mathrm{V}(R_i)+2\sum_{i,j|i<j}X_iX_j\mathrm{cov}(\beta_jI,\beta_jI)\\&=\sum_{i=1}^{n}X_i^2\mathrm{V}(R_i)+2\sum_{i,j|i<j}X_iX_j\beta_i\beta_j\mathrm{cov}(I,I)\\&=\sum_{i=1}^{n}X_i^2\mathrm{V}(R_i)+2\sum_{i,j|i<j}X_iX_j\beta_i\beta_j\mathrm{V}(I)\end{aligned} \tag{15.3.696}$$

Finally:

$$\boxed{\mathrm{V}(R_P)=\sum_{i=1}^{n}X_i^2\sigma_i^2+2\sigma_I^2\sum_{i,j|i<j}X_iX_j\beta_i\beta_j} \tag{15.3.697}$$

So what gives for a portfolio with two assets:

$$\mathrm{V}(R_P)=\sum_{i=1}^{2}X_i^2\sigma_i^2+2\sigma_I^2\sum_{i,j|i<j}X_iX_j\beta_i\beta_j=X_1^2\sigma_1^2+X_2^2\sigma_2^2+2\sigma_I^2(X_1X_2\beta_1\beta_2) \tag{15.3.698}$$

We can condense the notation of the variance using as always the matrix formulation by first noting respectively the transposed vector and the vector-column of the weights of the portfolio by:

$$X^T=[X_1\,X_2\,\ldots\,X_n] \qquad X=\begin{pmatrix}X_1\\X_2\\\vdots\\X_n\end{pmatrix} \tag{15.3.699}$$

and by defining the array of betas:

$$B_{ij}=\begin{bmatrix}\dfrac{\sigma_1^2}{\sigma_I^2} & \beta_1\beta_2 & \ldots & \beta_1\beta_n\\ \beta_2\beta_1 & \dfrac{\sigma_2^2}{\sigma_I^2} & & \vdots\\ \vdots & & \ddots & \\ \beta_n\beta_1 & \ldots & & \dfrac{\sigma_n^2}{\sigma_I^2}\end{bmatrix} \tag{15.3.700}$$

Which finally gives us:

$$\boxed{\text{V}(R_P)=\sigma_I^2X^T(B_{ij}X)} \tag{15.3.701}$$

For a portfolio of two assets this last relationship is reduced to:

$$\begin{aligned}\text{V}(R_P)&=\sigma_I^2[X_1\,X_2]\left(\begin{bmatrix}\dfrac{\sigma_1^2}{\sigma_I^2} & \beta_1\beta_2\\ \beta_2\beta_1 & \dfrac{\sigma_2^2}{\sigma_I^2}\end{bmatrix}\right)=\sigma_I^2[X_1\,X_2]\begin{bmatrix}\dfrac{\sigma_1^2}{\sigma_I^2}X_1+\beta_1\beta_2X_2\\ \beta_2\beta_1X_1+\dfrac{\sigma_2^2}{\sigma_I^2}X_2\end{bmatrix}\\ &=\sigma_I^2\left(\frac{\sigma_1^2}{\sigma_I^2}X_1^2+\beta_1\beta_2X_2X_1+\beta_2\beta_1X_1X_2+\frac{\sigma_2^2}{\sigma_I^2}X_2^2\right)=\sigma_1^2X_1^2+\sigma_2^2X_2^2+2\sigma_I^2X_1X_2\beta_1\beta-2\end{aligned} \tag{15.3.702}$$

We thus find much the same as the algebraic form!

If we do not know explicitly the returns, the study of the variance is a little trickier. Then we must use the linear model such as:

$$\text{V}(R_P)=\text{V}\left[\sum_{i=1}^n X_i(\alpha_i+\varepsilon_i)+\sum_{i=1}^n X_i\beta_iI\right]=\sum_{i=1}^n X_i^2\text{V}(\alpha_i)+\sum_{i=1}^n X_i^2\text{V}(\varepsilon_i)+\sum_{i=1}^n X_i^2\beta_i^2\text{V}(\alpha_{n+1}+\nu_{n+1}) \tag{15.3.703}$$

And always under the assumption already specified above as that if the index I is properly chosen it implies $\alpha_i=0$.

Also, let us put:

$$Q_i=\text{V}(\varepsilon_i) \tag{15.3.704}$$

Moreover, we know that:

$$Q_{n+1}=\text{V}(\nu_{n+1}) \tag{15.3.705}$$

Therefore ignoring the variance of the intercept:

$$\text{V}(R_P)=\sum_{i=1}^{n} X_i^2 Q_i + X_{n+1}^2 Q_{n+1} \tag{15.3.706}$$

because:

$$\text{V}(\alpha_{n+1})=\text{V}(c^{te})=0 \tag{15.3.707}$$

Finally:

$$\text{V}(R_P)=\sum_{i=1}^{n+1} X_i^2 Q_i \tag{15.3.708}$$

In this context the problem remains to maximize the economic function Z:

$$Z=\Phi\text{E}(R_P)-V(R_P)+\lambda\left(1-\sum_{i=1}^{n} X_i\right) \tag{15.3.709}$$

at the difference that now it will be written:

$$Z=\Phi\sum_{i=1}^{n+1} X_i\alpha_i-\sum_{i=1}^{n+1} X_i^2 Q_i+\lambda\left(1-\sum_{i=1}^{n} X_i\right) \tag{15.3.710}$$

The calculation of each of partial derivatives then gives:

$$\begin{cases} \dfrac{\partial Z}{\partial X_1}=\Phi\alpha_1-2X_1Q_1-\lambda=0 \\ \dfrac{\partial Z}{\partial X_2}=\Phi\alpha_2-2X_2Q_2-\lambda=0 \\ \ldots \\ \dfrac{\partial Z}{\partial X_n}=\Phi\alpha_n-2X_nQ_n-\lambda=0 \\ \dfrac{\partial Z}{\partial X_{n+1}}=\Phi\alpha_{n+1}-2X_nQ_{n+1}-\lambda=0 \\ \dfrac{\partial Z}{\partial \lambda}=1-X_1-X_2-\ldots-X_n=0 \end{cases} \tag{15.3.711}$$

or in matrix form:

$$\begin{bmatrix} 2Q_1 & 0 & \ldots & 0 & 1 \\ 0 & 2Q_2 & \ldots & 0 & 1 \\ \ldots & & \ldots & & \\ 0 & 0 & \ldots & 2Q_{n+1} & 1 \\ 1 & 1 & \ldots & 1 & 0 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \ldots \\ X_{n+1} \\ \lambda \end{bmatrix} = \begin{bmatrix} \Phi\alpha_1 \\ \Phi\alpha_2 \\ \vdots \\ \Phi\alpha_{n+1} \\ 1 \end{bmatrix} \tag{15.3.712}$$

The resolution of this system then goes through the inversion of a matrix simpler than that of Markowitz model but still requires relatively restrictive assumptions.

Finally, notice that the finance practitioner often use the performance indicators weighted by the risk, the most internationally known being the "Sharpe ratio". It is determined by the ratio (to be more exact it is its expected mean) between the differential of the returns of an investment (assets) supposed without risk R_f (this difference being named by the practitioners "active return") and the performance of our portfolio and standard deviation of this portfolio at the market rates (we will rigorously determine the origin of this relation further in our study of the CAPM):

$$\boxed{S_P = \frac{R_P - R_f}{\sigma_P}} \tag{15.3.713}$$

This relation expresses the level of pure return by volatility unit (or per unit of risk). For simplicity, it is a (marginal) profitability indicator obtained per unit of risk taken in a multi-assent portfolio management strategy. It helps answer the question: the portfolio manager is he able to get a higher return to the reference investment but with more risk?

- If the Sharpe ratio is negative, the portfolio performed less than the benchmark and the situation is very bad.
- If the Sharpe ratio is between 0 and 0.5, the invested portfolio outperformed the benchmark in a way that is considered as too excessive risk taking. Or, the risk taken is too high for the achieved return performance.
- If the Sharpe ratio is greater than 0.5, the invested portfolio returns outperformed the benchmark by taking ad hoc risk. In other words, the outperformance is not done at the cost of too high risks.

The Sharpe Ratio can be developed to give a more explicit form relatively to what we have seen just before:

$$S_P = \frac{\displaystyle\sum_{i=1}^{n} X_i(\alpha_i + \beta_i I + \varepsilon_i) - \sum_{i=1}^{n} X_i(\alpha_i + \varepsilon_i)}{\sqrt{\mathrm{V}\left(X_i(\alpha_i + \beta_i I + \varepsilon_i)\right)}} = \frac{\displaystyle\sum_{i=1}^{n} \beta_i I}{\sqrt{\mathrm{V}\left(\alpha_i + \beta_i I + \varepsilon_i\right)}} \tag{15.3.714}$$

Let us also notice another common indicator that is the "tracking error" defined as the standard deviation of the difference in performance between the portfolio and the benchmark. The higher the tracking error, the more the investment looks like the benchmark in terms of risk:

$$\boxed{\text{TE} = \sqrt{\mathrm{V}(R_P - R_f)} = \sigma_{R_P - R_f} = \sqrt{\sigma_P^2 + \sigma_f^2 + 2R_{P,f}\sigma_P\sigma_f}} \tag{15.3.715}$$

Remark

Some practitioner prefer to use the "CALMAR ratio" instead of the Sharpe ratio for comparing investment performance. Conceptually, as we have just seen, the Sharpe Ratio divides the average return of an investment by the standard deviation of its returns. The standard deviation is taken as a measure of the investment's risk. A higher Sharpe Ratio suggests more returns at lower risk.

But the standard deviation includes variations above the average returns. Most people like those and only worry about the below average returns. The CALMAR ratio uses maximum drawdown (see our study of Time Series further below for the definition page 7907) instead of standard deviation as a measure of risk (but it's empirical because we could use also the semivariance and create a bunch of other indicators like the Omega ratio, Sterling ratio, MAR ratio, etc.). So the "CALMAR ratio" (acronym of: ***CAL**ifornia **M**anaged **A**ccounts **R**eports* where it was published for the first time) is an investment's average return (usually for a 3 year period, i.e. 36 months) divided by its maximum drawdown in the same period (analysed on batches of 36 months):

$$\text{CALMAR Ratio} = \frac{\text{Annualized ROR (last 3 years)}}{\text{Drawdown (last 36 months)}} = \frac{\left(\prod_{i=0}^{36}(1+R(i))\right)^{1/3}-1}{\max\left\{0, \max_{t\in(T-36,T)} R(T-36)-R(T)\right\}} \tag{15.3.716}$$

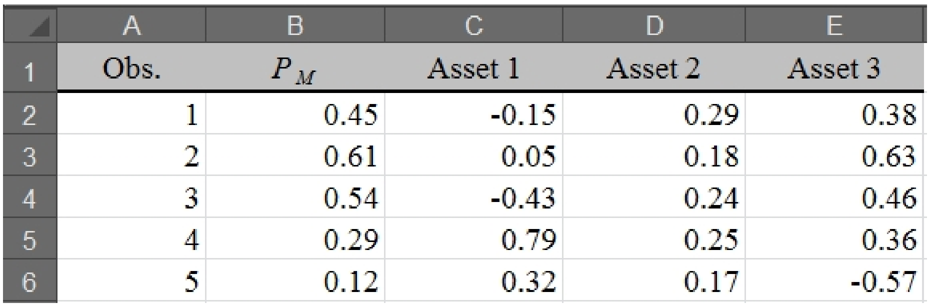
A higher CALMAR Ratio suggests more returns at lower risk.

It should be known that later versions of the Calmar ratio introduce the risk free rate into the numerator to create a Sharpe type ratio.

These models are relatively complex. This is why some years later, Sharpe and Lintner have created a new model that earned them the Nobel Prize in Economics and that we will study just after a practical example of what we just saw earlier above!

✍Example:

Let us consider three assets in equal proportions in an invested portfolio and n equal, non-missing, observations of their return R_{ij} entered in the spreadsheet software Microsoft® Excel® 14.0.7173 as visible below. These returns will be compared to a benchmark that I will be the performance of a benchmark market portfolio P_M:



	A	B	C	D	E
1	Obs.	P_M	Asset 1	Asset 2	Asset 3
2	1	0.45	-0.15	0.29	0.38
3	2	0.61	0.05	0.18	0.63
4	3	0.54	-0.43	0.24	0.46
5	4	0.29	0.79	0.25	0.36
6	5	0.12	0.32	0.17	-0.57

Figure 15.148: **Three assets with returns compared to a reference market portfolio**

The goal will be to determine the portfolio efficiency frontier with Sharpe's model.

In detail under graphical form let us first give the betas (return on assets based on the performance of the market portfolio / benchmark) with Microsoft® Excel® 14.0.7173 elementary tools:

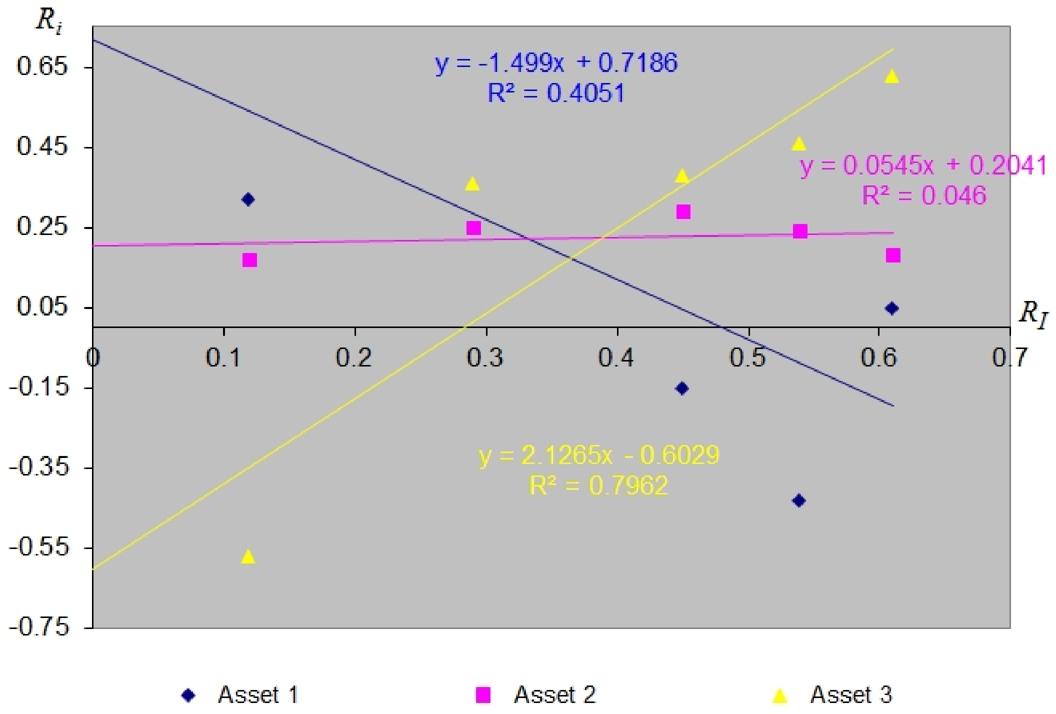
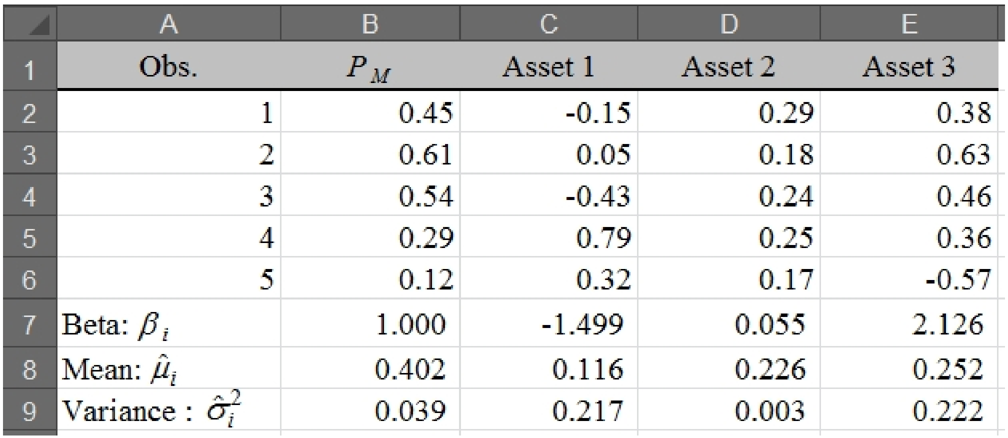


Figure 15.149: **Linear regression to calculate the betas of the assets returns**

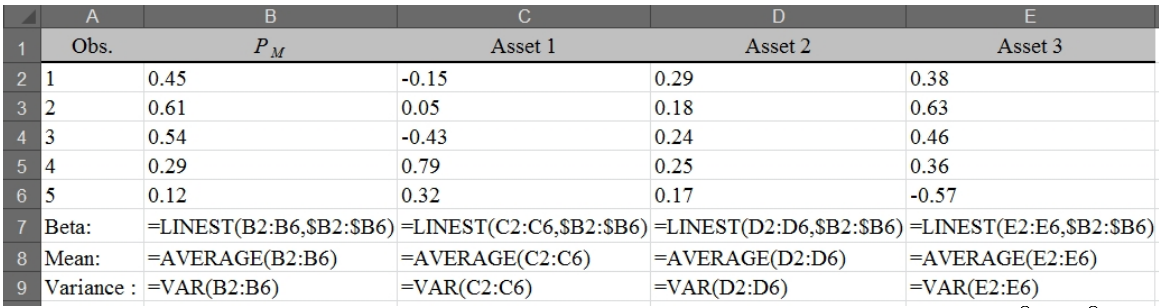
and the following construction list for calculating the betas, variance and the expected mean return of the market portfolio and the different assets:



	A	B	C	D	E
1	Obs.	P_M	Asset 1	Asset 2	Asset 3
2	1	0.45	-0.15	0.29	0.38
3	2	0.61	0.05	0.18	0.63
4	3	0.54	-0.43	0.24	0.46
5	4	0.29	0.79	0.25	0.36
6	5	0.12	0.32	0.17	-0.57
7	Beta: β_i	1.000	-1.499	0.055	2.126
8	Mean: $\hat{\mu}_i$	0.402	0.116	0.226	0.252
9	Variance : $\hat{\sigma}_i^2$	0.039	0.217	0.003	0.222

Figure 15.150: **Key indicators of the benchmark portfolio and the 3 assets**

Explicitly:



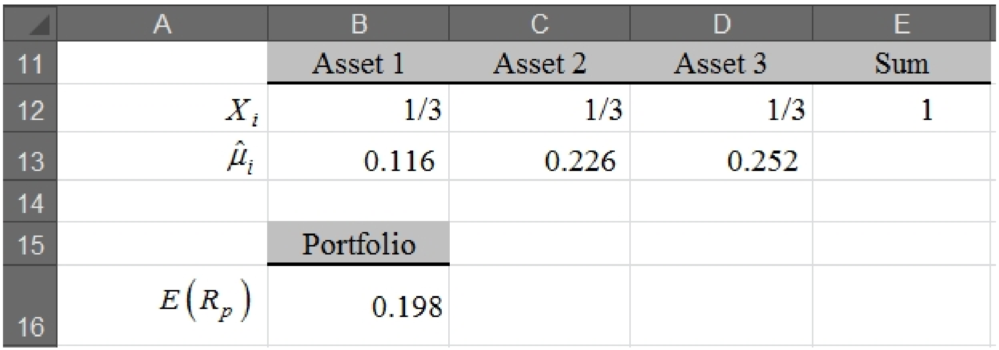
	A	B	C	D	E
1	Obs.	P_M	Asset 1	Asset 2	Asset 3
2	1	0.45	-0.15	0.29	0.38
3	2	0.61	0.05	0.18	0.63
4	3	0.54	-0.43	0.24	0.46
5	4	0.29	0.79	0.25	0.36
6	5	0.12	0.32	0.17	-0.57
7	Beta:	=LINEST(B2:B6,$B2:$B6)	=LINEST(C2:C6,$B2:$B6)	=LINEST(D2:D6,$B2:$B6)	=LINEST(E2:E6,$B2:$B6)
8	Mean:	=AVERAGE(B2:B6)	=AVERAGE(C2:C6)	=AVERAGE(D2:D6)	=AVERAGE(E2:E6)
9	Variance :	=VAR(B2:B6)	=VAR(C2:C6)	=VAR(D2:D6)	=VAR(E2:E6)

Figure 15.151: Key indicators of the benchmark portfolio and the 3 assets with explicit formulas in Microsoft® Excel® 14.0.7173

The expected portfolio return consisting of the three assets is easy to calculate because we have their returns. So:

$$\mathrm{E}(R_P)=\sum_{i=1}^{n} X_i R_i = X_i\hat{\mu}_1+X_2\hat{\mu}_2+X_3\hat{\mu}_3 \tag{15.3.717}$$

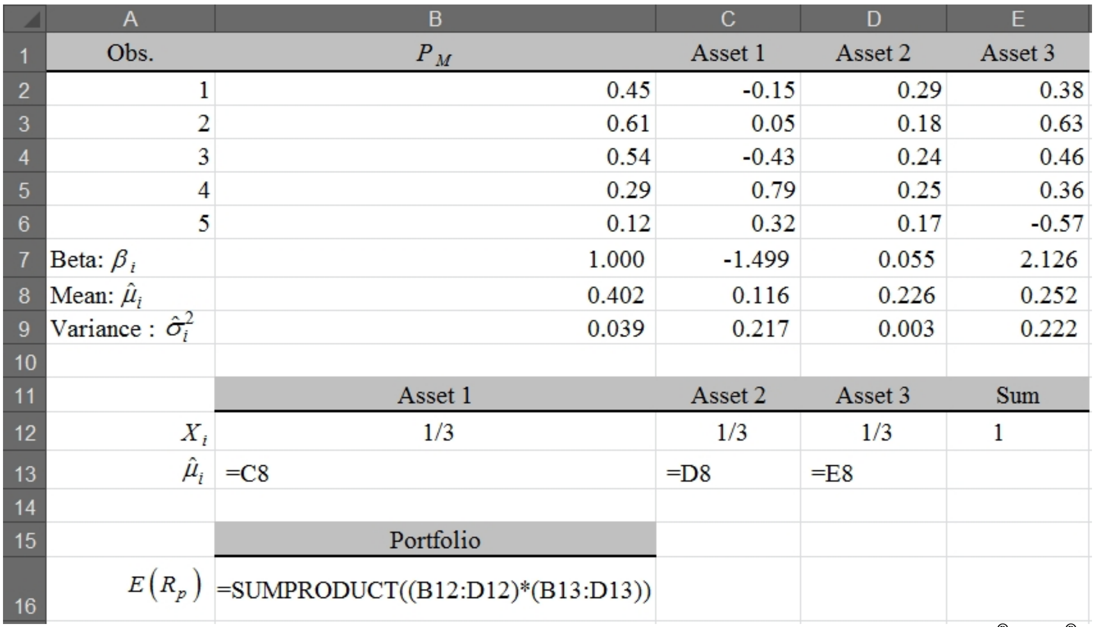
This gives in Microsoft® Excel® 14.0.7174:



	A	B	C	D	E
11		Asset 1	Asset 2	Asset 3	Sum
12	X_i	1/3	1/3	1/3	1
13	$\hat{\mu}_i$	0.116	0.226	0.252	
14					
15		Portfolio			
16	$E(R_p)$	0.198			

Figure 15.152: Expected mean return of each asset and of the overall portfolio

Or explicitly:



	A	B	C	D	E
1	Obs.	P_M	Asset 1	Asset 2	Asset 3
2	1	0.45	-0.15	0.29	0.38
3	2	0.61	0.05	0.18	0.63
4	3	0.54	-0.43	0.24	0.46
5	4	0.29	0.79	0.25	0.36
6	5	0.12	0.32	0.17	-0.57
7	Beta: β_i	1.000	-1.499	0.055	2.126
8	Mean: $\hat{\mu}_i$	0.402	0.116	0.226	0.252
9	Variance : $\hat{\sigma}_i^2$	0.039	0.217	0.003	0.222
10					
11		Asset 1	Asset 2	Asset 3	Sum
12	X_i	1/3	1/3	1/3	1
13	$\hat{\mu}_i$	=C8	=D8	=E8	
14					
15		Portfolio			
16	$E(R_p)$	=SUMPRODUCT((B12:D12)*(B13:D13))			

Figure 15.153: Expected mean return of each asset and of the overall portfolio explicit formulas in Microsoft® Excel® 14.0.7174

Now we must calculate the variance using the relation proved in the theoretical part of the preceding paragraphs:

$$\mathrm{V}(R_P)=\sigma^2(R_P)=\sigma_I^2 X^T(B_{ij}X) \tag{15.3.718}$$

with for reminder in our special case:

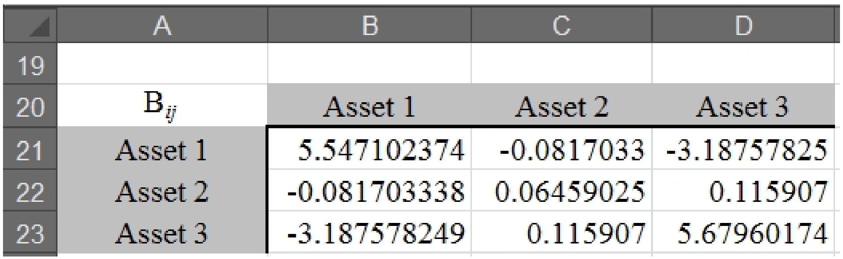
$$B_{ij}=\begin{bmatrix} \dfrac{\sigma_1^2}{\sigma_I^2} & \beta_1\beta_2 & \beta_1\beta_3 \\ \beta_2\beta_1 & \dfrac{\sigma_2^2}{\sigma_I^2} & \beta_2\beta_3 \\ \beta_3\beta_1 & \beta_3\beta_2 & \dfrac{\sigma_3^2}{\sigma_I^2} \end{bmatrix} \tag{15.3.719}$$

with in our example with $\sigma_I^2=0.039$ (cell `B13`).

Or in an expanded form for our example:

$$\sigma^2(R_P)=X_1^2\sigma_1^2+X_2^2\sigma_2^2+X_3^2\sigma_3^2+2X_1X_2\sigma_I^2\beta_1\beta_2+2X_1X_3\sigma_I^2\beta_1\beta_3+2X_2X_3\sigma_I^2\beta_2\beta_3 \tag{15.3.720}$$

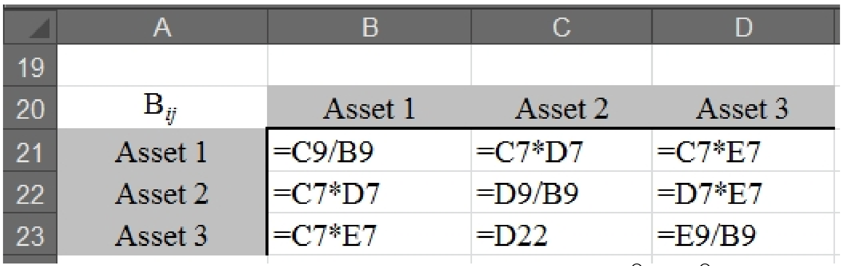
Which gives in Microsoft® Excel® for our array of betas B_{ij}:



	A	B	C	D
19				
20	B_{ij}	Asset 1	Asset 2	Asset 3
21	Asset 1	5.547102374	-0.0817033	-3.18757825
22	Asset 2	-0.081703338	0.06459025	0.115907
23	Asset 3	-3.187578249	0.115907	5.67960174

Figure 15.154: Matrix of betas B_{ij}

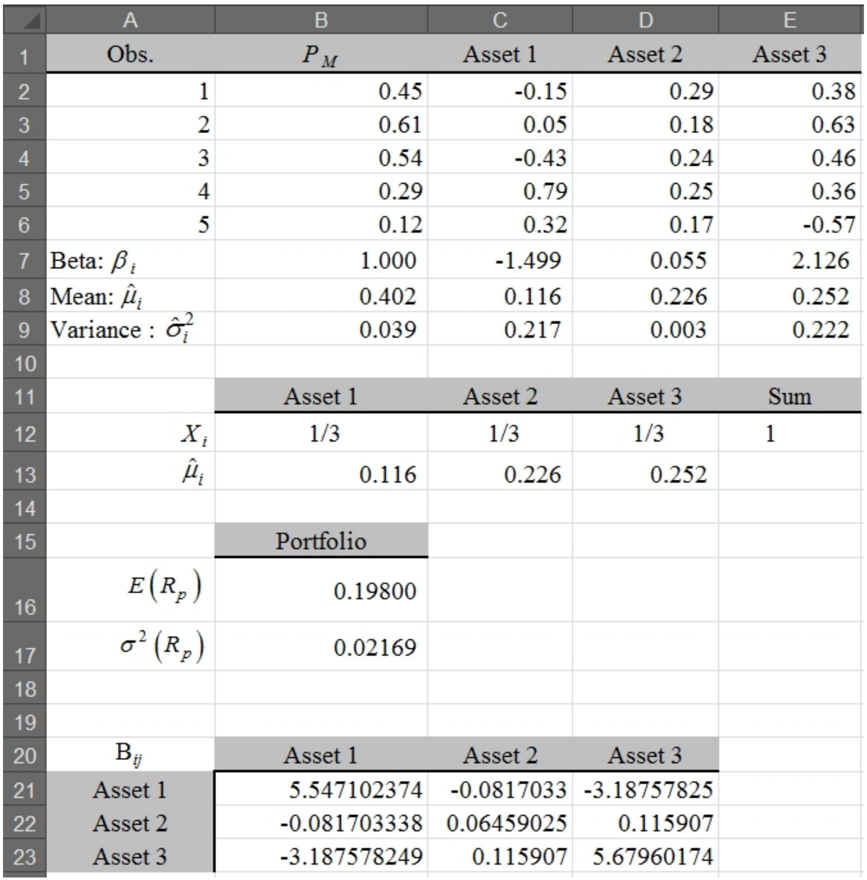
Or explicitly:



	A	B	C	D
19				
20	B_{ij}	Asset 1	Asset 2	Asset 3
21	Asset 1	=C9/B9	=C7*D7	=C7*E7
22	Asset 2	=C7*D7	=D9/B9	=D7*E7
23	Asset 3	=C7*E7	=D22	=E9/B9

Figure 15.155: Matrix of betas B_{ij} explicit formulas in Microsoft® Excel® 14.0.7174

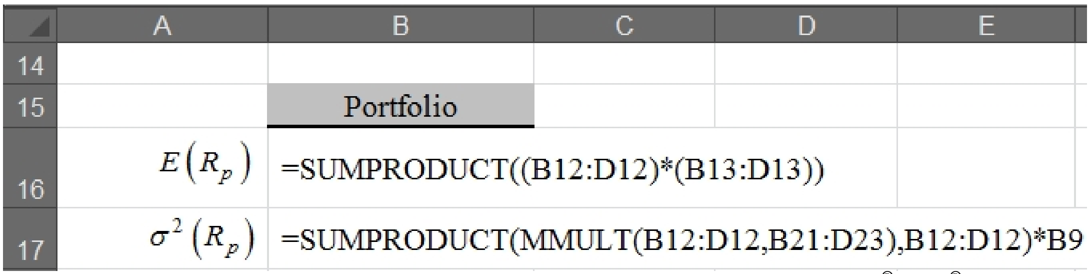
And finally the couple mean/variance of the portfolio is given by (cells `E16` and `E17`):



	A	B	C	D	E
1	Obs.	P_M	Asset 1	Asset 2	Asset 3
2	1	0.45	-0.15	0.29	0.38
3	2	0.61	0.05	0.18	0.63
4	3	0.54	-0.43	0.24	0.46
5	4	0.29	0.79	0.25	0.36
6	5	0.12	0.32	0.17	-0.57
7	Beta: β_i	1.000	-1.499	0.055	2.126
8	Mean: $\hat{\mu}_i$	0.402	0.116	0.226	0.252
9	Variance : $\hat{\sigma}_i^2$	0.039	0.217	0.003	0.222
10					
11		Asset 1	Asset 2	Asset 3	Sum
12	X_i	1/3	1/3	1/3	1
13	$\hat{\mu}_i$	0.116	0.226	0.252	
14					
15		Portfolio			
16	$E(R_p)$	0.19800			
17	$\sigma^2(R_p)$	0.02169			
18					
19					
20	B_{ij}	Asset 1	Asset 2	Asset 3	
21	Asset 1	5.547102374	-0.0817033	-3.18757825	
22	Asset 2	-0.081703338	0.06459025	0.115907	
23	Asset 3	-3.187578249	0.115907	5.67960174	

Figure 15.156: Couple mean / variance of the invested portfolio and global overview of all previous calculations

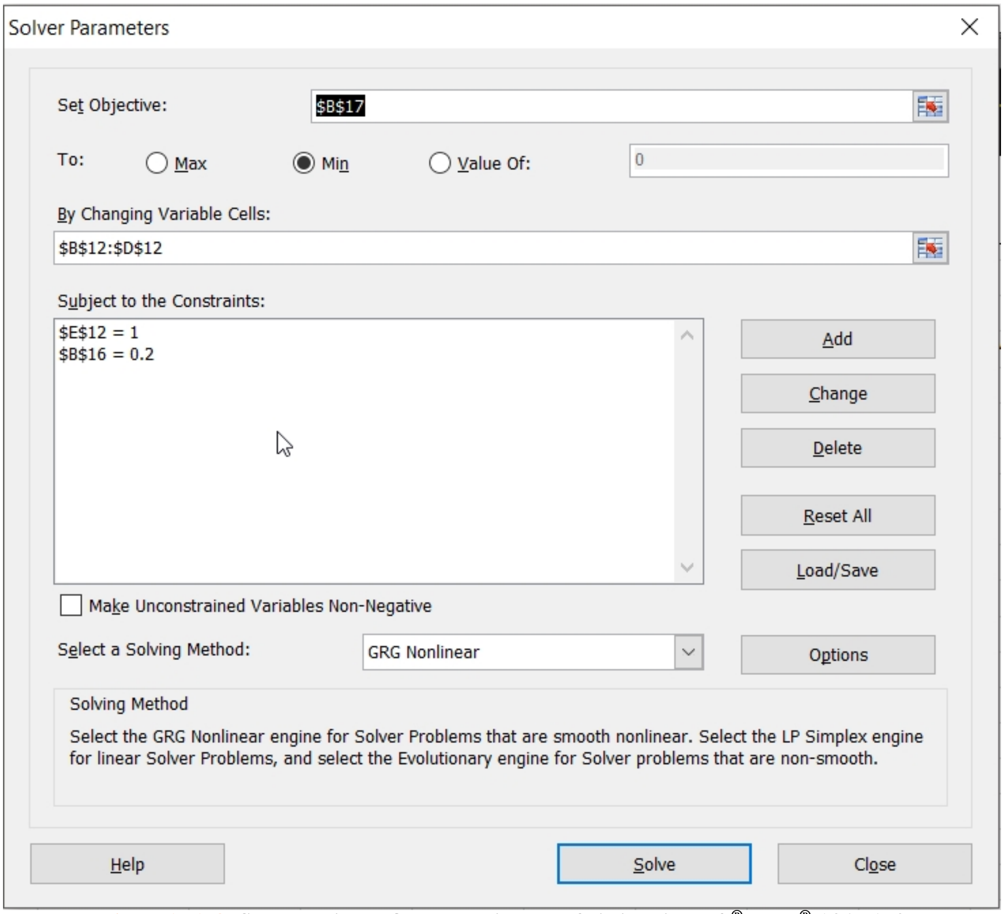
Explicitly of the part that interest us:



	A	B	C	D	E
14					
15		Portfolio			
16	$E(R_p)$	=SUMPRODUCT((B12:D12)*(B13:D13))			
17	$\sigma^2(R_p)$	=SUMPRODUCT(MMULT(B12:D12,B21:D23),B12:D12)*B9			

Figure 15.157: Couple mean / variance of the invested portfolio explicit formulas in Microsoft® Excel® 14.0.7174

Once this done, we proceed as for the Markowitz method. We use the solver to minimize the variance while imposing an expected mean return and a constraint as what the sum of weight of financial assets must be equal to unity:



Set Objective: $B$17

To: ○ Max ● Min ○ Value Of: 0

By Changing Variable Cells:
$B$12:$D$12

Subject to the Constraints:
$E$12 = 1
$B$16 = 0.2

☐ Make Unconstrained Variables Non-Negative

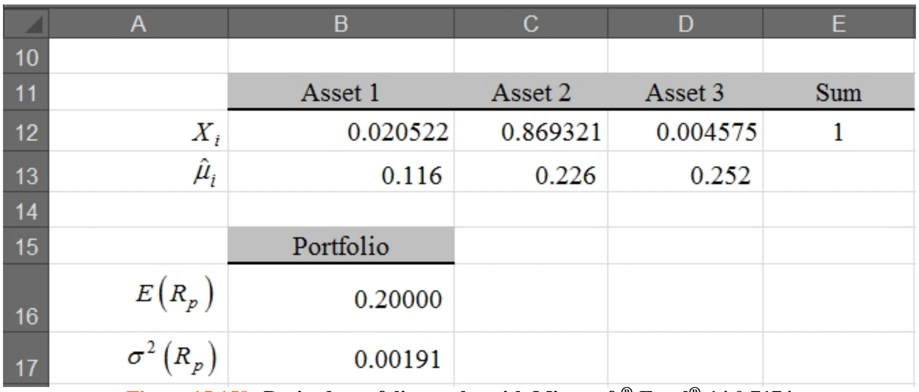
Select a Solving Method: GRG Nonlinear

Solving Method

Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

Figure 15.158: Solver setting to find the desired portfolio in Microsoft® Excel® 14.0.7174

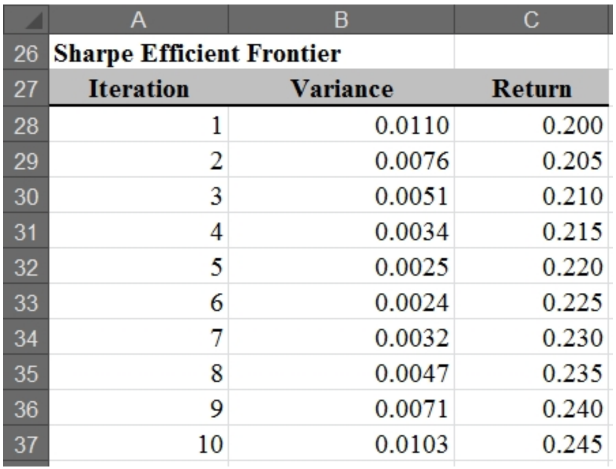
That gives:



	A	B	C	D	E
10					
11		Asset 1	Asset 2	Asset 3	Sum
12	X_i	0.020522	0.869321	0.004575	1
13	$\hat{\mu}_i$	0.116	0.226	0.252	
14					
15		Portfolio			
16	$E(R_p)$	0.20000			
17	$\sigma^2(R_p)$	0.00191			

Figure 15.159: Desired portfolio results with Microsoft® Excel® 14.0.7174

To build the Sharpe efficient frontier we proceed in the same way as for the Markowitz efficient frontier:



	A	B	C
26	**Sharpe Efficient Frontier**		
27	**Iteration**	**Variance**	**Return**
28	1	0.0110	0.200
29	2	0.0076	0.205
30	3	0.0051	0.210
31	4	0.0034	0.215
32	5	0.0025	0.220
33	6	0.0024	0.225
34	7	0.0032	0.230
35	8	0.0047	0.235
36	9	0.0071	0.240
37	10	0.0103	0.245

Figure 15.160: Sharpe efficient frontier values samples

This gives us the following $\mathrm{E}(R_P)=f(\sigma_{R_p})$ plot (we have put the Sharpe and Markowitz efficient frontier together so that the reader can compare them):

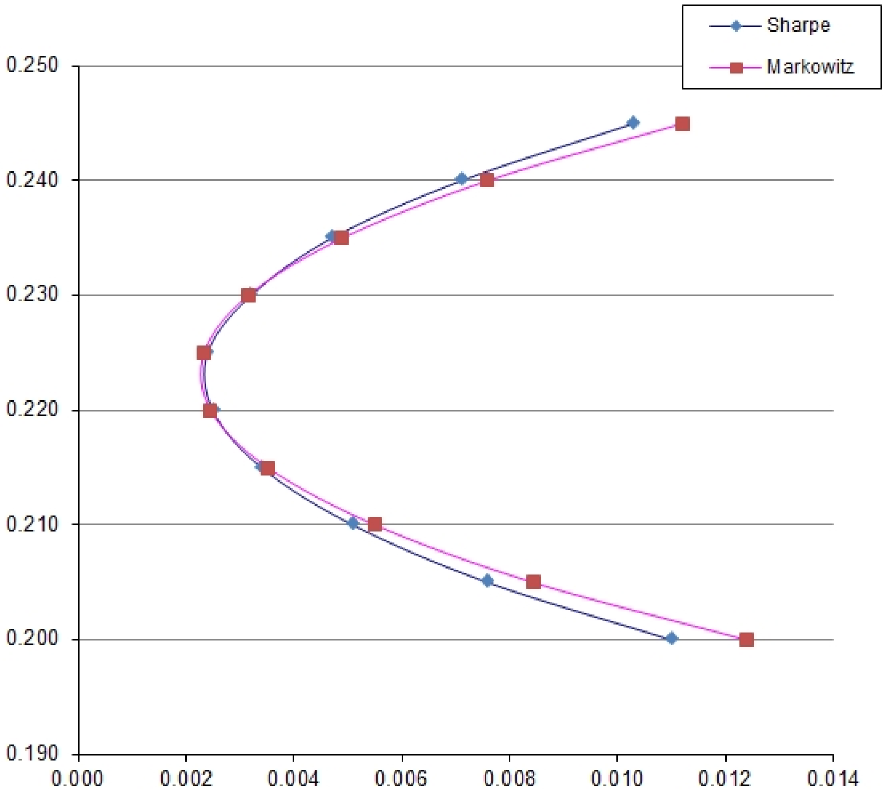


Figure 15.161: Sharpe & Markowitz efficient frontier plot in Microsoft® Excel® 14.07.7173

15.3.10.3.7.3 Equal Risk Contribution optimal weights portfolio

It seems that various numerical studies have shown that the Markowitz portfolio is very sensitive to small variations in assets values or simply to their correlation!

A proposed approach is then to use a portfolio whose contributions to the risk of different assets are equal. We are talking then about an "ERC (Equal Risk Contribution) portfolio" (that must not be confused with the equally weighted portfolio where $\omega_i = \omega_j$ for all i, j and not be confused with the "risk parity portfolio" that we will see further below!).

In order to determine the weights of each of the assets when the correlations are not identical between each other, it seems that there is no analytical solution so as far as we know (it is necessary to go through numerical methods). But let's see the idea with the special case of two assets and without short position (that is to say with non-negative weights).

The total variance of two assets is as we know given by:

$$\begin{aligned} \mathrm{V} &= w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + 2w_1w_2\sigma_{12} = w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + 2w_1w_2\rho\sigma_1\sigma_2 \\ &= w_1^2\sigma_1^2 + (1-w_1)^2\sigma_2^2 + 2w_1(1-w_1)\rho\sigma_1\sigma_2 \end{aligned} \tag{15.3.721}$$

Therefore it is trivial in this situation that the contribution to the risk of each asset is given by the vector:

$$\begin{pmatrix} w_1^2\sigma_1^2 + w_1(1-w_1)\rho\sigma_1\sigma_2 \\ (1-w_1)^2\sigma_2^2 + w_1(1-w_1)\rho\sigma_1\sigma_2 \end{pmatrix} \tag{15.3.722}$$

And so the contribution to risk in percent by:

$$\frac{1}{\sum_i \sigma_i}\begin{pmatrix} w_1^2\sigma_1^2 + w_1(1-w_1)\rho\sigma_1\sigma_2 \\ (1-w_1)^2\sigma_2^2 + w_1(1-w_1)\rho\sigma_1\sigma_2 \end{pmatrix} \tag{15.3.723}$$

In the case that interest us, finding the ERC portfolio involves finding w_1 such that both components of the vector are equal. That is to say:

$$w_1^2\sigma_1^2 + w_1(1-w_1)\rho\sigma_1\sigma_2 = (1-w_1)^2\sigma_2^2 + w_1(1-w_1)\rho\sigma_1\sigma_2 \tag{15.3.724}$$

What is simplified in:

$$w_1^2\sigma_1^2 = (1-w_1)^2\sigma_2^2 \tag{15.3.725}$$

Hence the polynomial of the second degree:

$$w_1^2(\sigma_2^2 - \sigma_1^2) - 2w_1\sigma_2^2 + \sigma_2^2 = 0 \tag{15.3.726}$$

The roots are given then obviously by:

$$\begin{aligned} w_1 &= \frac{2\sigma_2^2 \pm \sqrt{4\sigma_2^4 - 4(\sigma_2^2-\sigma_1^2)\sigma_2^2}}{2(\sigma_2^2-\sigma_1^2)} = \frac{\sigma_2^2 \pm \sqrt{\sigma_2^4 - (\sigma_2^2-\sigma_1^2)\sigma_2^2}}{(\sigma_2^2-\sigma_1^2)} \\ &= \frac{\sigma_2^2 \pm \sqrt{\sigma_2^4 - \sigma_2^4 + \sigma_1^2\sigma_2^2)}}{(\sigma_2^2-\sigma_1^2)} = \frac{\sigma_2^2 \pm \sigma_1\sigma_2}{(\sigma_2^2-\sigma_1^2)} \end{aligned} \tag{15.3.727}$$

and for w_2 just remember that it is just equal to $1-w_1$.

15.3.10.3.7.4 Error tracking optimal weights portfolio

Always in the context of a portfolio with two assets, the goal here is to find the weights that minimize the definition of the error tracking introduced earlier (see page 7724):

$$\text{TE}=\sqrt{\text{V}(R_p-R_f)}=\sigma_{R_p-R_f}=\sqrt{\sigma_p^2+\sigma_f^2-2R_{P,F}\sigma_P\sigma_F} \tag{15.3.728}$$

Or in other words, that minimize:

$$\begin{aligned}\text{TE}^2&=\sigma_p^2+\sigma_f^2-2R_{P,F}\sigma_P\sigma_F\\&=(w_1^2\sigma_1^2+w_2^2\sigma_2^2+2w_1w_2\sigma_1\sigma_2)+\sigma_f^2+2R_{P,F}\sqrt{w_1^2\sigma_1^2+w_2^2\sigma_2^2+2w_1w_2\sigma_1\sigma_2}\,\sigma_f\\&=(w_1^2\sigma_1^2+w_2^2\sigma_2^2+2w_1w_2\sigma_{12})+\sigma_f^2+2R_{P,F}\sqrt{w_1^2\sigma_1^2+w_2^2\sigma_2^2+2w_1w_2\sigma_{12}}\,\sigma_f\end{aligned} \tag{15.3.729}$$

For this we take the partial derivative with respect to one of the weights:

$$\frac{\partial\text{TE}^2}{\partial w_1}=2w_1\sigma_1^2+2\sigma_{12}-2R_{R,F}\frac{2w_1\sigma_1^2+2\sigma_{12}}{\sqrt{w_1^2\sigma_1^2+w_2^2\sigma_2^2+2w_1w_2\sigma_{12}}}\sigma_f=0 \tag{15.3.730}$$

Which is numerically easy to solve for a computer. By hand we did arrived at a satisfactory elegant result (do not hesitate to give us your development if you have a pedagogical elegant solution).

Once you have the result for w_1, to get w_2 just remember that it is just equal to $1-w_1$.

15.3.10.3.7.5 Sharpe optimal weights portfolio

Here we will study the case of minimizing the risk of a portfolio by maximizing the Sharpe ratio already introduced earlier above (see page 7723):

$$S_p=\frac{\text{E}(R_p)-R_f}{\sigma_p} \tag{15.3.731}$$

for a case of a portfolio of two assets (just to simplify...).

In order to simplify the calculations, let us use the excess returns of the two assets rather than their direct return:

$$R_1=r_1-R_f \qquad R_2=r_2-R_f \tag{15.3.732}$$

We then rewrite the Sharpe ratio:

$$\begin{aligned}S_p&=\frac{\text{E}(R_p)-R_f}{\sigma_p}=\frac{(w_1r_1+w_2r_2)-1\cdot R_f}{\sqrt{w_1^2\sigma_1^2+w_2^2\sigma_2^2+2w_1w_2\sigma_{12}}}\\&=\frac{(w_1r_1+w_2r_2)-(w_1+w_2)\cdot R_f}{\sqrt{w_1^2\sigma_1^2+w_2^2\sigma_2^2+2w_1w_2\sigma_{12}}}=\frac{w_1R_1+w_2R_2}{\sqrt{w_1^2\sigma_1^2+w_2^2\sigma_2^2+2w_1w_2\sigma_{12}}}\end{aligned} \tag{15.3.733}$$

We will now calculate the partial derivative with respect to w_1 and see when this one vanishes (obviously by symmetry the result will be the same for w_2):

$$\begin{aligned}\dfrac{\partial S_p}{\partial w_1}&=\dfrac{R_1}{\sqrt{w_1^2\sigma_1^2+w_2^2\sigma_2^2+2w_1w_2\sigma_{12}}}-\dfrac{1}{2}\dfrac{w_1R_1+w_2R_2}{(w_1^2\sigma_1^2+w_2^2\sigma_2^2+2w_1w_2\sigma_{12})^{3/2}}(2w_i\sigma_1^2+2w_2\sigma_{12})\\&=\dfrac{R_1}{\sqrt{w_1^2\sigma_1^2+w_2^2\sigma_2^2+2w_1w_2\sigma_{12}}}-\dfrac{(w_1R_1+w_2R_2)(w_1\sigma_1^2+w_2\sigma_{12})}{w_1^2\sigma_1^2+w_2^2\sigma_2^2+2w_1w_w\sigma_{12})^{3/2}}=0\end{aligned} \tag{15.3.734}$$

To simplify this we will do some algebraic manipulations. First we will multiply everything by:

$$\sqrt{w_1^2\sigma_1^2+w_2^2\sigma_2^2+2w_1w_2\sigma_{12}} \tag{15.3.735}$$

Which gives:

$$R_1-\dfrac{(w_1R_1+w_2R_2)(w_1\sigma_1^2+w_2\sigma_{12})}{w_1^2\sigma_1^2+w_2^2\sigma_2^2+2w_1w_2\sigma_{12}}=0 \tag{15.3.736}$$

We multiply on both sides of the equality by the denominator of the second term:

$$R_1(w_1^2\sigma_1^2+w_2^2\sigma_2^2+2w_1w_2\sigma_{12})-(w_1R_1+w_2R_2)(w_1\sigma_1^2+w_2\sigma_{12})=0 \tag{15.3.737}$$

Therefore:

$$\cancel{w_1^2R_1\sigma_1^2}+w_2^2R_1\sigma_2^2+2w_1w_2R_1\sigma_{12}-\cancel{w_1^2R_1\sigma_1^2}-w_1w_2R_1\sigma_{12}-w_1w_2R_2\sigma_1^2-w_2^2R_2\sigma_{12}=0 \tag{15.3.738}$$

and also:

$$\begin{aligned}w_2^2R_1\sigma_2^2+&(2w_1w_2R_1\sigma_{12}-w_1w_2R_1\sigma_{12})-w_1w_2R_2\sigma_1^2-w_2^2R_2\sigma_{12}\\&=w_2^2R_1\sigma_2^2+w_1w_2R_1\sigma_{12}-w_1w_2R_2\sigma_1^2-w_2^2R_2\sigma_{12}\\&=w_2R_1\sigma_2^2+w_1R_1\sigma_{12}-w_1R_2\sigma_1^2-w_2R_2\sigma_{12}=0\end{aligned} \tag{15.3.739}$$

We do a small substitution:

$$(1-w_1)R_1\sigma_2^2+w_1R_1\sigma_{12}-w_1R_2\sigma_1^2-(1-w_1)R_2\sigma_{12}=0 \tag{15.3.740}$$

Which gives:

$$w_1=\dfrac{R_1\sigma_2^2-R_2\sigma_{12}}{R_1\sigma_2^2+R_2\sigma_1^2-(R_1+R_2)\sigma_{12}} \tag{15.3.741}$$

and for w_2 just remember that it is just equal to $1-w_1$.

15.3.10.3.7.6 Maximum diversification optimal weights portfolio

The idea of this portfolio based on the principle of maximizing the "diversification ratio" is defined by:

$$\boxed{\text{DR}=\dfrac{\displaystyle\sum_i w_i\sigma_i}{\sigma_p}} \tag{15.3.742}$$

or in vector-matrix form:

$$\text{DR} = \frac{\vec{w}^T \cdot \text{diag}(\Sigma)}{\sqrt{\vec{w}^T \Sigma \vec{w}}} \qquad (15.3.743)$$

Indeed, since a diversified portfolio must have the smallest standard deviation portfolio (denominator tends to zero). So this ratio will have to be maximal, such that our optimization problem is then written:

$$\boxed{\begin{aligned} &\underset{\vec{\omega}}{\text{maximize}} && \text{DR} = \frac{\vec{w}^T \cdot \text{diag}(\Sigma)}{\sqrt{\vec{w}^T \Sigma \vec{w}}} \\ &\text{subject to} && \sum w_i = 1 \end{aligned}} \qquad (15.3.744)$$

If we write it down for two assets, then we have:

$$\text{DR} = \frac{w_1\sigma_1 + w_2\sigma_2}{\sqrt{w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + 2w_1w_2\sigma_{12}}} \qquad (15.3.745)$$

Let us compare this expression with that obtained in our study of Sharpe's risk minimization portfolio seen previously:

$$S_p = \frac{w_1R_1 + w_2R_2}{\sqrt{w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + 2w_1w_2\sigma_{12}}} \qquad (15.3.746)$$

So we have exactly the same thing except that we have to replace two variables. So by taking up the final result of Sharpe's minimum risk model which was for recall:

$$w_1 = \frac{R_1\sigma_2^2 - R_2\sigma_{12}}{R_1\sigma_2^2 + R_2\sigma_1^2 - (R_1 + R_2)\sigma_{12}} \qquad (15.3.747)$$

It then comes for the present case:

$$w_1 = \frac{\sigma_1\sigma_2^2 - \sigma_2\sigma_{12}}{\sigma_1\sigma_2^2 + \sigma_2\sigma_1^2 - (\sigma_1 + \sigma_2)\sigma_{12}} \qquad (15.3.748)$$

and for w_2 just remember that it is just equal to $1 - w_1$.

15.3.10.3.7.7 Treynor-Black optimal weights portfolio

The Treynor-Black (11973 according to holocene calendar) portfolio management model is closely linked to active portfolio management, which underlying idea is to outperform the benchmark portfolio's market by discretionary selection of the products, securities or sectors the most likely to grow faster than the market.

To introduce this model we will consider that we have a passive portfolio M that best replicates a market index (purchase of long-term assets selected on the basis of a known market index such as the CAC40, SP500, etc.) and an active portfolio A.

To simplify the calculations, let us use portfolio excess returns rather than their direct return:

$$\begin{aligned} R_A &= r_A - R_f \\ R_M &= r_M - R_f \end{aligned} \qquad (15.3.749)$$

According to the Treynor-Black model, the market portfolio has an α equal to zero because it is supposed to replicate the market without error. According to the Sharpe model (single index model), we can write:

$$R_A = \alpha_A + \beta_A(r_M - R_f) = \alpha_A + \beta_A r_M \tag{15.3.750}$$

We will also use the relationship demonstrated earlier above (see page 7718):

$$\sigma_A = \beta_A^2 r_M^2 + \sigma_\varepsilon^2 \tag{15.3.751}$$

and also (see page 7720):

$$\beta_A = \frac{\text{cov}(R_A, R_M)}{\text{V}(R_M)} = \frac{\sigma_{AM}}{\sigma_M^2} \Leftrightarrow \sigma_{AM} = \beta_A \sigma_M^2 \tag{15.3.752}$$

We will also use the relation proved just before for the optimum weight of a two-asset portfolio that maximizes the Sharpe ratio:

$$w_1 = \frac{R_1\sigma_2^2 - R_2\sigma_{12}}{R_1\sigma_2^2 + R_2\sigma_1^2 - (R_1 + R_2)\sigma_{12}} \tag{15.3.753}$$

This last relation is written in our context:

$$\begin{aligned}
w_A &= \frac{R_A\sigma_M^2 - R_M\sigma_{AM}}{R_A\sigma_M^2 + R_M\sigma_A^2 - (R_A + R_M)\sigma_{AM}}\\
&= \frac{(\alpha_A + \beta_A R_M)\sigma_M^2 - R_M\beta_A\sigma_M^2}{(\alpha_A + \beta_A R_M)\sigma_M^2 + R_M\sigma_A^2 - (\alpha_A + \beta_A R_M + R_M)\beta_A\sigma_M^2}\\
&= \frac{\alpha_A\sigma_M^2}{(\alpha_A + \beta_A R_M)\sigma_M^2 + R_M\sigma_A^2 - (\alpha_A + \beta_A R_M + R_M)\beta_A\sigma_M^2}\\
&= \frac{\alpha_A}{\alpha_A + \beta_A R_M + R_M\dfrac{\sigma_A^2}{\sigma_M^2} - \alpha_A\beta_A - \beta_A^2 R_M - \beta_A R_M}\\
&= \frac{\alpha_A}{\alpha_A + R_M\dfrac{\sigma_A^2}{\sigma_M^2} - \alpha_A\beta_A - \beta_A^2 R_M} = \frac{\alpha_A}{\alpha_A(1 - \beta_A) + R_M\frac{\sigma_A^2}{\sigma_M^2} - \beta_A^2 R_M}\\
&= \frac{\alpha_A}{\alpha_A(1 - \beta_A) + R_M\dfrac{\beta_A^2\sigma_M^2 + \sigma_\varepsilon^2}{\sigma_M^2} - \beta_A^2 R_M}\\
&= \frac{\alpha_A}{\alpha_A(1 - \beta_A) + \dfrac{R_M(\beta_A^2\sigma_M^2 + \sigma_\varepsilon^2) - \beta_A^2 R_M\sigma_M^2}{\sigma_M^2}} = \frac{\alpha_A}{\alpha_A(1 - \beta_A) + \dfrac{R_M\sigma_\varepsilon^2}{\sigma_M^2}}
\end{aligned} \tag{15.3.754}$$

Now we are going to assume that the beta of the active portfolio is equal to 1 because at the beginning of the construction of our portfolio A we will align it with the portfolio M. Then we have in this case:

$$w_A^0 = \frac{\alpha_A}{\dfrac{R_M\sigma_\varepsilon^2}{\sigma_M^2}} = \frac{\alpha_A}{\sigma_\varepsilon^2}\frac{\sigma_M^2}{R_M} = \frac{\dfrac{\alpha_A}{\sigma_\varepsilon^2}}{\dfrac{R_M}{\sigma_M^2}} \tag{15.3.755}$$

This is the expression we find in the textbooks as being the initial position in the active portfolio. The numerator is the contribution to the weight of the active portfolio while the denominator is the contribution of the market portfolio.

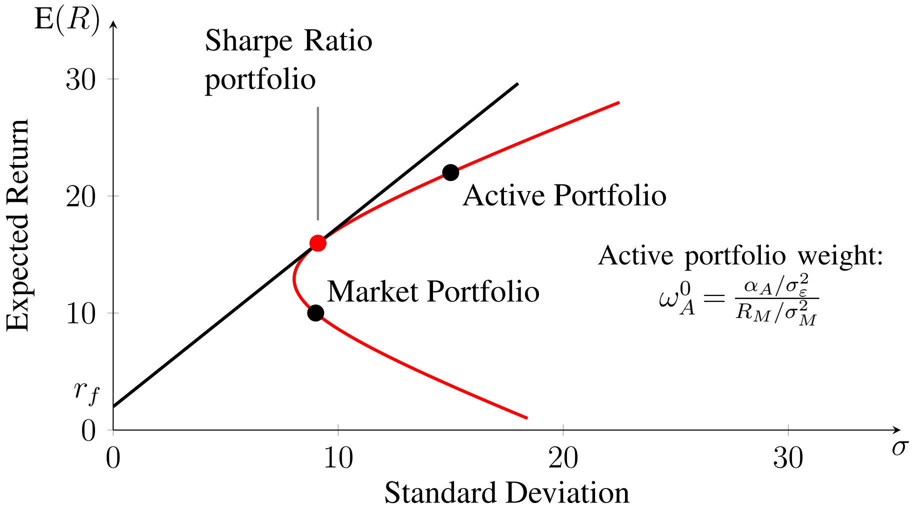


Figure 15.162: Treynor-Black Portfolio Selection

We then see that if the excess return of the passive portfolio's of the market is high this will decrease the weight of the active portfolio A. If the alpha of the active portfolio is high this will increase the weight of the active portfolio A.

Now let us go back to the unbalanced case where the beta of A is no longer equal to 1:

$$\begin{aligned} w_A^* &= \frac{\alpha_A}{\alpha_A(1-\beta_A)+\dfrac{R_M\sigma_\varepsilon^2}{\sigma_M^2}} = \frac{\alpha_A\sigma_M^2}{\alpha_A(1-\beta_A)\sigma_M^2+R_M\sigma_\varepsilon^2} \\ &= \frac{\dfrac{\alpha_A\sigma_M^2}{R_M\sigma_\varepsilon^2}}{\dfrac{\alpha_A(1-\beta_A)\sigma_M^2}{R_M\sigma_\varepsilon^2}+1} = \frac{\dfrac{\alpha_A\sigma_M^2}{R_M\sigma_\varepsilon^2}}{\dfrac{\alpha_A(1-\beta_A)\sigma_M^2}{R_M\sigma_\varepsilon^2}+1} = \frac{w_A^0}{w_A^0(1-\beta_A)+1} \end{aligned} \tag{15.3.756}$$

And do not forget that:

$$w_M^* = 1 - w_A^* \tag{15.3.757}$$

Let's not forget that in our active portfolio we have several assets. We can then write:

$$w_i^0 = \frac{\alpha_i}{\sigma_{\varepsilon,i}^2} \tag{15.3.758}$$

but the sum of the weights must be equal to 1. So we write:

$$w_i = \frac{\dfrac{\alpha_i}{\sigma_{\varepsilon,i}^2}}{\displaystyle\sum_i \frac{\alpha_i}{\sigma_{\varepsilon,i}^2}} \tag{15.3.759}$$

Since alpha can be negative, weights can be negative too (usually the active portfolio is composed of long and short positions).

Therefore we get:

$$w_i^* = w_A^* \frac{\dfrac{\alpha_i}{\sigma_{\varepsilon,i}^2}}{\sum\limits_i \dfrac{\alpha_i}{\sigma_{\varepsilon,i}^2}} \tag{15.3.760}$$

15.3.10.3.7.8 Risk Parity Portfolio

A risk parity portfolio denotes a class of portfolios whose assets verify the following equalities:

$$\omega_i \frac{\partial f(\vec{\omega})}{\partial \omega_i} = \omega_j \frac{\partial f(\vec{\omega})}{\partial \omega_j}, \forall i,j \tag{15.3.761}$$

where f is a positively homogeneous function of degree one that measures the total risk of the portfolio and $\vec{\omega}$ is the portfolio weight vector. In other words, the marginal risk contributions for every asset in a risk parity portfolio are equal. A common choice for f, for instance, is the standard deviation of the portfolio, which is usually named "volatility", i.e.:

$$\sigma_P = f(\vec{\omega}) = \sqrt{\vec{\omega}^T \Sigma \vec{\omega}} \tag{15.3.762}$$

where Σ is the covariance matrix of the assets.

With that particular choice of f, the risk parity requirements become:

$$\omega_i (\Sigma\vec{\omega})_i = \omega_j (\Sigma\vec{\omega})_j, \forall i,j \tag{15.3.763}$$

A natural extension of the risk parity portfolio is the so named "risk budget portfolio", in which the marginal risk contributions match preassigned quantities. Mathematically:

$$\omega_i (\Sigma\vec{\omega})_i = b_i \vec{\omega}^T \Sigma \vec{\omega}, \forall i \tag{15.3.764}$$

where $\vec{b} = (b_1, b_2, \ldots, b_N)$ (with $\mathbb{1}^T \vec{b} = 1$ and $\vec{b} \geq \vec{0}$) is the vector of desired marginal risk contributions.

In the case that Σ is diagonal and with the constraints $\mathbb{1}^T \vec{\omega} = 1$ and $\vec{w} \geq 0$, the risk budgeting portfolio is:

$$\boxed{\omega_i = \frac{\sqrt{b_i}/\sqrt{\Sigma_{ii}}}{\sum\limits_{k=1}^N \sqrt{b_k}/\sqrt{\Sigma_{kk}}} = \frac{\sqrt{b_i}/\sigma_i}{\sum\limits_{k=1}^N \sqrt{b_k}/\sigma_k}, \qquad i = 1 \ldots N} \tag{15.3.765}$$

That latter relation can be found in many textbooks with the case where b_i is not taken into account. Such tat:

$$\omega_i = \frac{1/\sigma_i}{\sum\limits_{k=1}^N 1/\sigma_k}, \qquad i = 1 \ldots N \tag{15.3.766}$$

and in that latter case we have the famous example with two assets:

$$\omega_1=\frac{\dfrac{1}{\sigma_1}}{\dfrac{1}{\sigma_1}+\dfrac{1}{\sigma_2}} \quad (15.3.767)$$

However, for non-diagonal Σ or with other additional constraints or objective function terms, a closed-form solution does not exist and some optimization procedures have to be constructed. The previous diagonal solution can always be used and is named "naive risk budgeting portfolio".

The design of risk parity portfolios as solved just above is of paramount importance both for academia and industry. However, practitioners would like the ability to include additional constraints and objective terms desired in practice, such as the mean return, box constraints, etc. In such cases, the risk-contribution constraints cannot be met exactly due to the trade-off among different objectives or additional constraints.

Let us explore, for instance, the effect of including the expected return as an additional objective in the optimization problem. The problem can be formulated as:

$$\boxed{\begin{array}{ll}\underset{\vec{\omega}}{\min} & \displaystyle\sum_{i=1}^{N}\left(\omega_i\left(\Sigma\vec{\omega}\right)_i-b_i\vec{\omega}^T\Sigma\vec{\omega}\right)^2-\lambda_\mu\vec{\omega}^T\vec{\mu}\\ \text{subject to} & \mathbb{1}^T\vec{\omega}=1,\vec{\omega}\geq\vec{0}\\ & \mathbb{1}^T\vec{b}=1,\vec{b}\geq\vec{0}\end{array}} \quad (15.3.768)$$

where:

$$R(\vec{\omega})=\sum_{i=1}^{N}\left(\omega_i\left(\Sigma\vec{\omega}\right)_i-b_i\vec{\omega}^T\Sigma\vec{\omega}\right)^2 \quad (15.3.769)$$

is the "risk concentration function" or "risk parity function, $\vec{\omega}^T\vec{\mu}$ is the expected return, and λ_μ is a trade-off parameter.

Remark

There are almost a dozens of variations of the empirical optimisation problem above for the Risk Parity portfolio. For people interested to see what are these different variations, you can refer to our `R` companion book.

15.3.10.3.8 Capital Asset Pricing Model (CAPM)

As we have seen before, Markowitz (11959 according to holocene calendar) developed the theory of optimal choice of a portfolio on the basis of the variance and the expected return. Later (11963 according to holocene calendar), Sharpe developed the theory of optimal choice of a portfolio on the basis of risk indices such as beta coefficients to simplify the heavy calculations (but since the computer are almost all powerful enough to run the Markowitz model without difficulties).

Sharpe, Lintner and Mossin (11965 according to holocene calendar) then studied the consequences of these theories to develop an extremely simple theory to evaluate the beta coefficients, expected returns and variances of financial assets portfolio from data statistics on the global market and the specificity of the composition of a portfolio.

This theory based again on the mean-variance problem is named the "Capital Asset Pricing Model" (CAPM) or "Capital Asset Pricing Model at Equilibrium" (CAPME). This is a frequently used model, both by practitioners than academics to evaluate the anticipated returns of equilibrium of any risky assets in the market.

Remark

In facts it is more a set of models named "linear attribution models", the SLM (Sharpe, Lintner and Mossin) CAPM is the most known and easiest one but there is also the Fama-French CAPM (three factors and five factors versions), the Fama-McBeth CAPM, the Carhart CAPM (four factors), the Scholes and William adjusted beta, the Dimson adjusted beta and surely a few other empirical one... that we will not introduce in this textbook because they are completely empirical and mathematically speaking not very beautiful and interesting.

To begin, let us recall that we have already seen in our study of the return that the short periodic rate of return (daily, weekly, monthly, yearly) of an asset is calculated as follows:

$$R_t=\frac{(P_t-P_{t-1})+C_t}{P_{t-1}} \tag{15.3.770}$$

with P_t that is the price of an asset at the period t, P_{t-1} the price of the same asset at the previous asset $t-1$ and finally C_t the cash flow paid by the asset during the period running from $t-1$ to t.

This is used as we know to calculate the "realized return" of an asset while in fact it is the "expected return" return that interest a particular investor.

Obviously we can first define some obvious return statistical indicators. One for example is the "mean daily return", denoted $\bar{R}$ that is just the arithmetic average of the R_t. We can therefore also define the "unbiased instantaneous return volatility", or also named "historical volatility", as:

$$\sigma_I=\sqrt{\frac{\sum\limits_{t=0}^{n}\left(R_t-\bar{R}\right)^2}{n-1}} \tag{15.3.771}$$

The mean daily return being often considered as equal to zero (what is obviously not valid in the field of hyperfrequency market trading!), we can find sometimes in the literature the following relation:

$$\sigma_I=\sqrt{\frac{\sum\limits_{t=0}^{n} R_t^2}{n-1}} \tag{15.3.772}$$

and named "realized volatility" (respectively "realized variance") or sometimes "reality variance".

Using the result of the Bachelier speculation model that we will see further below on page 7842 (under the assumption that the constraints of the model are verified) the projection of these two last relations over a period t' will be given by:

$$\sigma_R=\sigma_I\sqrt{t'} \tag{15.3.773}$$

and named in the case of yearly projection "annualized historical volatility" or respectively "annualized realized volatility".

At the time of taking the decision, the return that will achieve an investor by holding a given asset is uncertain this is why we speak of "expected returns": this is a return that we seek to evaluate and we hope to receive during the next investment period.

To calculate the expected return, as we have seen, it should be allocated to each possible value of the return a probability of realization, then calculate a weighted average of the possible values using the probabilities p_i as weights:

$$\text{E}(R_i)=\mu_i=p_1R_{i1}+p_2R_{i2}+\ldots+p_nR_{in}=\sum_{j=1}^{n} p_jR_{ij} \tag{15.3.774}$$

However, it is clear that in a given economy, investors will be tempted to hold more that one financial assets for obvious reasons and therefore seek to compose portfolios. The expected mean return of a portfolio may be calculated using the known relation:

$$\text{E}(R_P)=X_1\text{E}(R_1)+X_2\text{E}(R_2)+\ldots+X_n\text{E}(R_n)=\sum_{i=1}^{n} X_i\text{E}(R_i)=\sum_{i=1}^{n}\left(X_i\sum_{j=1}^{m} p_jR_{ij}\right) \tag{15.3.775}$$

where for recall n is the number of assets included in the portfolio, $\text{E}(R_i)$ the return of asset i in the portfolio and X_i the proportion of this asset i in the invested portfolio.

The expected return rate is however as we know insufficient to characterize an investment opportunity and we must also take into account the risk, i.e. the variability of the return on that investment over financial assets. The variance is, as we have already seen it, used as simple measurement of risk and given for a financial asset by:

$$\text{V}(R_i)=\sigma_i^2=p_1[R_{i1}-\text{E}(R_i)]^2+p_2[R_{i2}-\text{E}(R_i)]^2+\ldots+[R_{in}-\text{E}(R_i)]^2 \tag{15.3.776}$$

Therefore:

$$\text{V}(R_i)=\sigma_i^2=\sum_{j=1}^{n}(R_{ij}-\text{E}(R_i))^2 p_j \tag{15.3.777}$$

The calculation of the risk of a portfolio thus involves without surprise two important concepts: the variability of returns of each asset, measured by the variances of these, as well as the relationships between the various assets composing the portfolio.

The dependency between two assets is often measured, as we have already mentioned in our previous study of the Sharpe and Markowitz model, by the covariance or the correlation coefficient.

The covariance between two assets i and j is calculated as follows for recall:

$$\text{cov}(R_i,R_j)=p_1[R_{i1}-\text{E}(R_i)][R_{j1}-\text{E}(R_j)]+\ldots+p_n[R_{in}-\text{E}(R_i)][R_{jn}-\text{E}(R_j)] \tag{15.3.778}$$

As we know the probabilities are equally likely (and if we do not work on a sample otherwise we should have $n-1$ instead of n in the denominator for the product factor of the sum):

$$\text{cov}(R_i,R_j)=\frac{1}{n}\sum_{k=1}^{n}(R_{ik}-\text{E}(R_i))(R_{jk}-\text{E}(R_j)) \tag{15.3.779}$$

The covariance between the returns of two assets may be positive or negative and its value has no economic significance - in the sense of causality - as we know (see section Statistics page 614).

Remark

Let us once again recall that we have seen in the section Statistics that when the returns of two active assets vary in the same direction (respectively in opposite direction) their covariance will be positive (respectively negative).

The correlation coefficient between two assets i and j meanwhile is calculated as follows (see section Statistics page 614):

$$R_{ij}=\frac{\text{cov}(R_i,R_j)}{\sqrt{\text{V}(R_i)\text{V}(R_j)}}=\frac{c_{ij}}{\sigma_i\sigma_j} \tag{15.3.780}$$

Once the variance and covariance of the various assets calculated, we will be able to calculate the variance of the return of whole portfolio with n sets. This variance is given by the following relation as already proved in the section Statistics and the previous models above:

$$\text{V}(R_p)=\sum_{i=1}^{n}\sum_{j=1}^{n}X_iX_j\text{cov}(R_i,R_j)=\sum_{i=1}^{n}nX_i^2\sigma_i^2\sum_{i,j|i\neq j}X_iX_j\text{cov}(R_i,R_j) \tag{15.3.781}$$

or written differently:

$$\text{V}(R_p)=\sum_{i=1}^{n}X_i^2\text{V}(R_i)+2\sum_{i,j|i<j}X_iX_j\text{cov}(R_j,R_j) \tag{15.3.782}$$

The above relation of the variance of portfolio return clearly shows that even if the returns to different assets in the portfolio are completely uncorrelated, the variance of the latter can be further reduced by adding more assets.

To understand this, we notice that for n uncorrelated assets, the variance is reduced to (since the covariance is then zero):

$$\text{V}(R_P)=\sigma_P^2=\sum_{i=1}^{n} X_i^2\sigma_i^2 \tag{15.3.783}$$

Simplifying even more, if all variances are assumed equal and if all the assets are held in the same proportions $(1/n)$, we have obviously:

$$\sigma_P^2=\sum_{i=1}^{n}\left(\frac{1}{n}\right)^2\sigma_i^2=n\left(\frac{1}{n}\right)^2\sigma_i^2=\frac{1}{n}\sigma_i^2 \tag{15.3.784}$$

Thus, when n approaches infinity, the portfolio variance approaches zero. So if uncorrelated risks are combined portfolio, total risk can be eliminated by diversification, we speak then as we know of "diversifiable risk". In the cases where the risks are correlated, diversification eliminate only the risks specific to the assets while the market risk will continue to exist.

Notice that the risk reduction is greater when the various assets held are negatively correlated!!! In this situation, it is possible to combine risky assets to form a completely riskless portfolio!!!!

From the foregoing, it is clear that any investor wishing to build a portfolio will seek to hold a set of risky assets that will allow him to receive some return with minimum risk. In other words, he will seek to minimize the variance for an expected level of performance while respecting some budget constraints. We know that the expected return and the performance variance of a portfolio containing n risky assets are written as follows:

$$\begin{aligned}\text{E}(R_P)&=\sum_{i=1}^{n} X_i\text{E}(R_i)\\ \text{V}(R_P)&=\sum_{i=1}^{n}\sum_{i=j}^{n} X_iX_j\text{cov}_{ij}\end{aligned} \tag{15.3.785}$$

Moreover, we know that from these n assets that it is possible to construct an infinite number of portfolios by varying the weights X_i. But the most interesting portfolios for a particular investor portfolios are those that minimize the risk that he must endure to achieve a given level of performance. These portfolios are the result of the following minimization problem that is as we already know a non-linear optimization problem (see section Numerical Methods page 6662):

$$\begin{aligned}\underset{\text{V}(R_P)}{\text{minimize}}\quad & \text{E}(R_P)=c^{te}\\ \text{subject to}\quad & \sum X_i=1\end{aligned} \tag{15.3.786}$$

that we had already seen in our study of the Markowitz model.

It is therefore possible to create an infinite number of portfolios by varying the proportions invested in each asset. The next step is to select from among all available portfolios, a given portfolio. To do this, one must consider the individual preferences of the investor.

A rational investor should therefore consider only the portfolios lying on the efficient frontier for its investment choices. Its optimal portfolio will be at the point of tangency between the efficient frontier and its highest indifference curve that it would be capable of achieving.

Let us explicit that a little bit...! Consider for this the following MATLAB® plot:

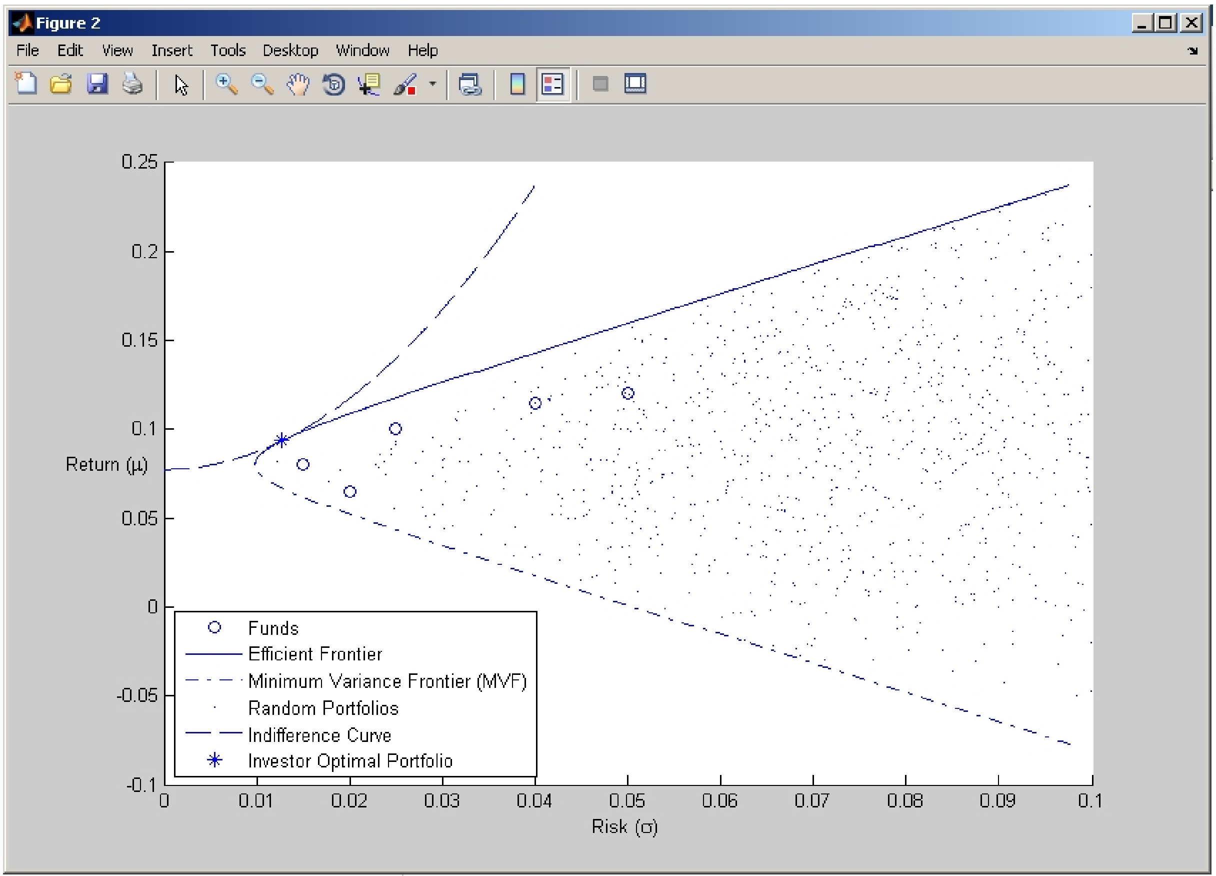


Figure 15.163: **Efficient frontier and indifference curve in MATLAB® 2013a**

The indifference curve above is quit obvious to understand: the average investor agree to take more risk as the expected returns increase and expect a zero risk when the return is the minimum of the mark risk free rate. This indifference curve is for most people obviously with a convex shape (the opposite would be quite hard to explain...).

By doing so, each investor maximizes his expected utility. In the presence of an economy containing only risky assets, the composition of the portfolio of risky assets varies from one individual to another.

To determine the correct (best) tangent to indifference curve and efficient frontier we can use an information partially available on the market: the risk free rate. Indeed, almost all investors have the possibility to invest in risk free financial assets. We will then seek for the portfolio that is the prolongation of the risk free investment and therefore corresponding to the tangent of a given indifference curve as visible in the figures above and below:

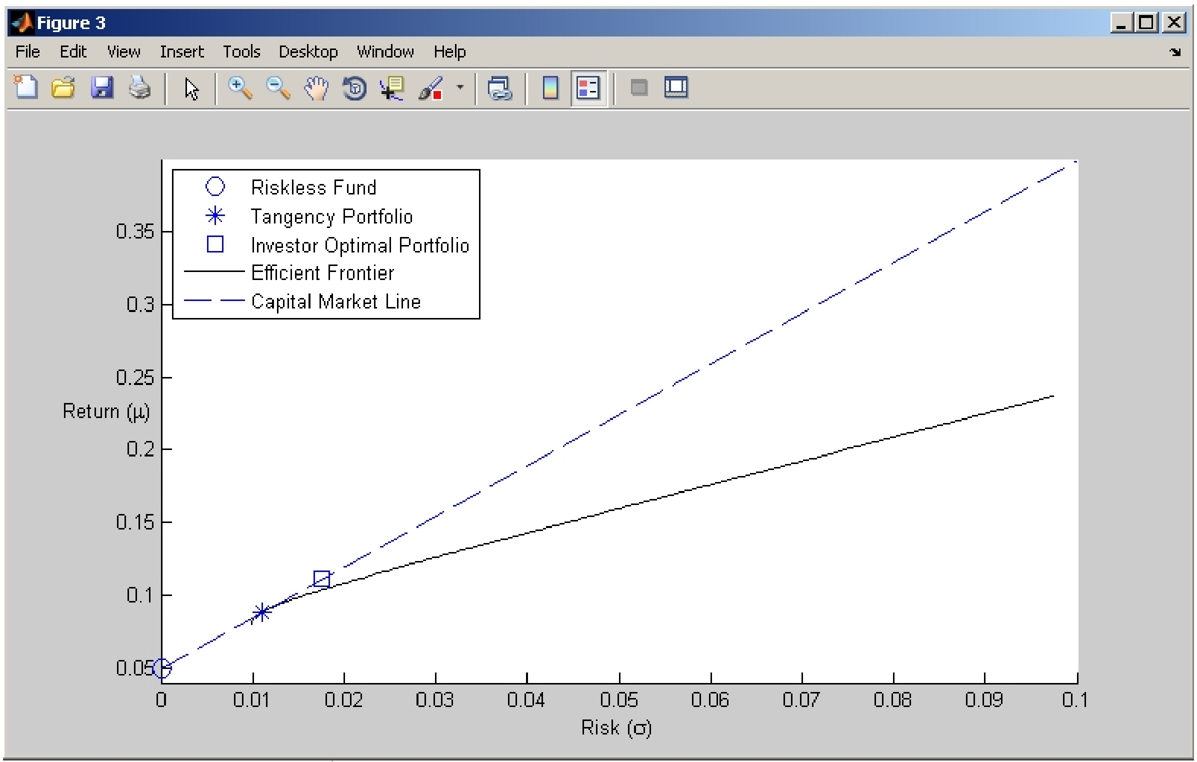


Figure 15.164: **Efficient frontier and SML in MATLAB® 2013a**

Then let us consider a portfolio that is a combination of risk-free assets (or risk free sub-portfolio) and the at risk asset (a market non-risk free portfolio). Then we have:

$$R_P = X_M R_M + (1 - X_M) R_f \tag{15.3.787}$$

where X_M is the fraction of the portfolio invested in the market portfolio (M) of return R_M and R_f is as always the risk-free rate return (or "certain return").

We then have (remember that the expected mean of a constant is equal to that constant as proved in the section Statistics):

$$\mathrm{E}(R_P) = X_M \mathrm{E}(R_M) + (1 - X_M)\mathrm{E}(R_f) = X_M \mathrm{E}(R_M) + (1 - X_M) R_f \tag{15.3.788}$$

and therefore:

$$\begin{aligned}\mathrm{V}(R_P) &= \mathrm{V}(X_M R_M + (1 - X_M) R_P)\\ &= \mathrm{E}\left[(X_M R_M + (1 - X_M) R_P - \mathrm{E}\left(X_M R_M + (1 - X_M) R_P\right))^2\right]\\ &= \mathrm{E}\left[(X_M R_M + (1 - X_M) R_P - X_M \mathrm{E}(R_M) - (1 - X_M) R_P)^2\right]\\ &= \mathrm{E}\left[(X_M R_M - X_M \mathrm{E}(R_M))^2\right] = \mathrm{E}\left[X^2 \left(R_M - \mathrm{E}(R_M)\right)^2\right]\\ &= X_M^2 \mathrm{E}\left[(R_M - \mathrm{E}(R_M))^2\right] = X_M^2 \mathrm{V}(R_M)\end{aligned} \tag{15.3.789}$$

Either in condensed form:

$$\sigma_P^2 = X_m^2 \sigma_M^2 \tag{15.3.790}$$

The derivative of expected return relatively to X_M gives us:

$$\frac{\partial \mathrm{E}(R_P)}{\partial X_M} = \mathrm{E}(R_M) - R_f \tag{15.3.791}$$

The derivative of the standard deviation relatively to X_M gives us (don't forget that the standard deviation of a constant is equal to zero!):

$$\frac{\partial\sigma(R_P)}{\partial X_M}=\sigma(R_M) \tag{15.3.792}$$

Putting these results together (ratio), we have:

$$\frac{\partial\mathrm{E}(R_P)}{\partial\sigma(R_P)}=\frac{\mathrm{E}(R_M)-R_f}{\sigma(R_M)} \tag{15.3.793}$$

and since:

$$\mathrm{E}(R_P)=X_M\mathrm{E}(R_M)+(1-X_M)R_f=X_M(\mathrm{E}(R_M)-R_f)+R_f \tag{15.3.794}$$

It comes then:

$$\mathrm{E}(R_P)=\frac{\sigma(R_P)}{\sigma(R_M)}(\mathrm{E}(R_M)-R_f)+R_f \tag{15.3.795}$$

And since the main target in finance is to represent graphically as for the Markowitz and Sharpe efficient frontier plot:

$$\mathrm{E}(R_P)=f(\sigma(R_P)) \tag{15.3.796}$$

So, it is traditional to write the prior-previous relation like an affine function:

$$\boxed{\mathrm{E}(R_P)=\frac{\mathrm{E}(R_M)-R_f}{\sigma(R_M)}\sigma(R_P)+R_f=(\mathrm{E}(R_M)-R_f)\frac{\sigma(R_P)}{\sigma(R_M)}+R_f} \tag{15.3.797}$$

This relation is the equation of a line named "capital market line" (C.M.L.). This line is a linear combination of the portfolios having risky assets with a free risk asset[26]. Therefore all efficient portfolios from the efficient frontier are dominated by those situated on that line for a same level of risk excepted in the case of the tangent portfolio hence it's name "super-efficient portfolio".

The intercept is obviously R_f and its slope is a variable function of the expected mean and standard deviation of the market portfolio and the factor find in front of the standard deviation of the R_P is the "Sharpe ratio" that we have already met multiple times so far:

$$S_P=\frac{R_P-R_f}{\sigma_P} \tag{15.3.798}$$

but without proving its origin... (but now the readers finally know where it comes from). Notice that in a well build portfolio we should have $S_P>0$ therefore the slope of the C.M.L. should be always positive!

The above equation is the central relation of the CAPM ("one-factor") model. Any portfolio measured on the market which is outside this line of this model is considered (at least in theory) as not being in balance.

---

[26] The capital market line (CML) represents portfolios that optimally combine risk and return. CML is a special case of the capital allocation line (CAL) defined earlier where the risk portfolio is the market portfolio.

By construction, this line equation therefore associates with each level of risk, the highest expected return at the intersection of the efficient frontier. Thus, given the performance of a risk-free asset it becomes easy from this equation to determine the point of tangency with the efficiency frontier of Markowitz or Sharpe to get the most efficient portfolio based on the risk-free performance and even without knowing the indifference curve!!

Let us now determine an equation for the expected return of any individual asset.

Consider a new portfolio of return R_P which is a combination of any unique asset A and the reference market portfolio, where X_A is the fraction of the portfolio invested in the asset A.

What we would like to do is to evaluate the slope of the curve of the combinations expected mean/standard deviation of the returns when we combine the market portfolio with the asset A.

We wish to evaluate the value of the slope of the tangent to the efficient frontier such that the weighting of the asset A is zero.

We have:

$$R_P = X_A R_A + (1-X_A) R_M \tag{15.3.799}$$

We get immediately:

$$\text{E}(R_P) = X_A \text{E}(R_A) + (1-X_A)\text{E}(R_M) \tag{15.3.800}$$

and (see section Statistics page 596 and page 597):

$$\sigma^2(R_P) = X_A^2\sigma^2(R_A) + (1-X_A)^2\sigma(R_M) + 2X_A(1-X_A)\text{cov}(R_A, R_M) \tag{15.3.801}$$

Let us derivate the expected return of the portfolio relatively to X_A, we get then for the expected mean:

$$\frac{\partial \text{E}(R_P)}{\partial X_A} = \text{E}(R_A) - \text{E}(R_M) \tag{15.3.802}$$

Doing same for the standard deviation, we get:

$$\frac{\partial \sigma(R_P)}{\partial X_A} = \frac{1}{2}\frac{2X_A\sigma^2(R_A) - 2(1-X_A)\sigma^2(R_M) + (2-4X_A)\text{cov}(R_A, R_M)}{\sqrt{X_A^2\sigma^2(R_A) + (1-X_A)^2\sigma(R_M) + 2X_A(1-X_A)\text{cov}(R_A, R_M)}} \tag{15.3.803}$$

The contribution of Sharpe and Lintner was that we must evaluate these derivatives at the point where $X_A = 0$ that is to say where the weighting of asset A in the new portfolio is zero.

In doing so, we get the following expression for the standard deviation of the new portfolio (of course, the expression for the expected return does not change):

$$\left.\frac{\partial \sigma(R_P)}{\partial X_A}\right|_{X_A=0} = \frac{1}{2}\frac{-2\sigma^2(R_M) + 2\text{cov}(R_A, R_M)}{\sqrt{\sigma^2(R_M)}} \tag{15.3.804}$$

which gives after simplification:

$$\left.\frac{\partial \sigma(R_P)}{\partial X_A}\right|_{X_A=0} = \frac{\text{cov}(R_A, R_M) - \sigma^2(R_M)}{\sigma(R_M)} \tag{15.3.805}$$

With the both derivatives above, we can obtain an expression for the curve of the combinations means/standard deviation for the new portfolio. Then we have:

$$\frac{\left.\dfrac{\partial \mathrm{E}(R_P)}{\partial X_A}\right|_{X_A=0}}{\left.\dfrac{\partial \sigma(R_P)}{\partial X_A}\right|_{X_A=0}}=\left.\frac{\partial \mathrm{E}(R_P)}{\partial \sigma(R_P)}\right|_{X_A=0}=\frac{\mathrm{E}(R_A)-\mathrm{E}(R_M)}{\dfrac{\mathrm{cov}(R_A,R_M)-\sigma^2(R_M)}{\sigma(R_M)}} \tag{15.3.806}$$

This slope must be equal to that of the C.M.L. By equalizing, we get:

$$\frac{\mathrm{E}(R_M)-R_f}{\sigma(R_M)}=\frac{\mathrm{E}(R_A)-\mathrm{E}(R_M)}{\dfrac{\mathrm{cov}(R_A,R_M)-\sigma^2(R_M)}{\sigma(R_M)}} \tag{15.3.807}$$

After a trivial simplification we get:

$$\frac{\mathrm{E}(R_M)-R_f}{\sigma^2(R_M)}=\frac{\mathrm{E}(R_A)-\mathrm{E}(R_M)}{\mathrm{cov}(R_A,R_M)-\sigma^2(R_M)} \tag{15.3.808}$$

And therefore after rearranging:

$$\begin{aligned}\mathrm{E}(R_A)&=\frac{\mathrm{E}(R_M)-R_f}{\sigma^2(R_M)}\left(\mathrm{cov}(R_A,R_M)-\sigma^2(R_M)\right)+\mathrm{E}(R_M)\\&=(\mathrm{E}(R_M)-R_f)\frac{\mathrm{cov}(R_A,R_M)}{\sigma^2(R_M)}-\mathrm{E}(R_M)+R_f+\mathrm{E}(R_M)\\&=(\mathrm{E}(R_M)-R_f)\frac{\mathrm{cov}(R_A,R_M)}{\sigma^2(R_M)}+R_f\end{aligned} \tag{15.3.809}$$

Hence:

$$\mathrm{E}(R_A)-R_f=(\mathrm{E}(R_M)-R_f)\frac{\mathrm{cov}(R_A,R_M)}{\sigma^2(R_M)} \tag{15.3.810}$$

By putting what we have already seen during our study of the Sharpe model, that is to say the non-diversifiable risk as beta factor:

$$\beta=(\mathrm{E}(R_M)-R_f)\frac{\mathrm{cov}(R_A,R_M)}{\sigma^2(R_M)} \tag{15.3.811}$$

So it is the volatility of the return of the non-risky asset compared to that to the market portfolio.

Then we have:

$$\mathrm{E}(R_A)-R_f=(\mathrm{E}(R_M)-R_f)\beta \tag{15.3.812}$$

This expression allows to express the excess return of risk-free asset as the product of the excess return of the market portfolio and the beta factor of the risk-free asset.

The excess return of an asset does not directly depend of its variance, which is often an intuitive measure of risk of an asset. What counts is its beta factor, which depends on its covariance with the market portfolio.

More typically, the last relation is used graphically as a straight line equation:

$$\boxed{\mathrm{E}(R_A)=f(\beta)=(\mathrm{E}(R_M)-R_f)\beta+R_f} \tag{15.3.813}$$

and written in many textbooks (or in the case of multiple assets), as:

$$\mathrm{E}(R_i)=f(\beta_i)=(\mathrm{E}(R_M)-R_f)\beta_i+R_f \tag{15.3.814}$$

This relation is named the "security market line" (SML), and more rarely "market asset line", it is extremely important in finance, as it therefore gives the mean expected return of an asset A in function of the beta, of the market return and the risk free rate. This model is say to be "monofactorial" in that it does distinguish only single explanatory factor of the risk of an asset (at the opposite of the Fama-French CAPM, Fama-McBeth CAPM or Carhart CAPM).

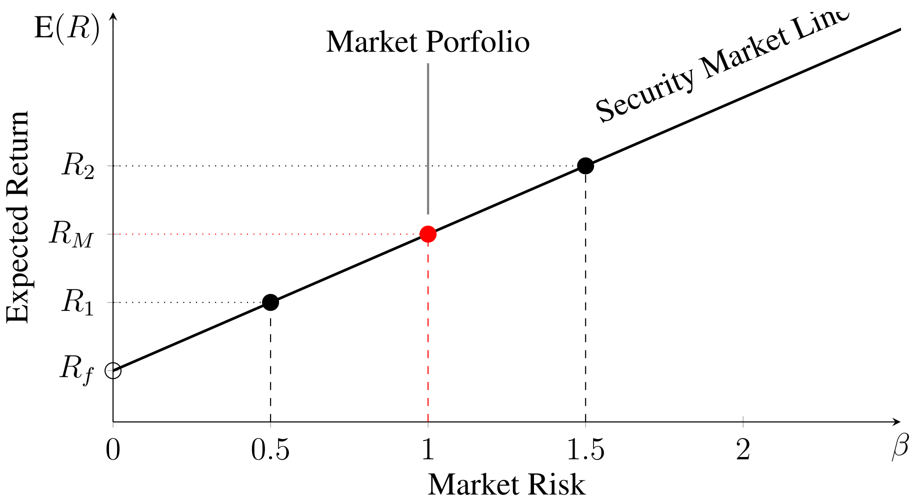


Figure 15.165: **Security Market Line**

There are obviously more complex models taking into account other additional factors such as the Fama-French model that use three explanatory variables of even five variables.

✍**Example:**

The risk free rate is 5% and the expected market return is 8%. An asset A is twice more sensible to market movement than the benchmark, verbatim when the index recorded an increase of 1%, the asset A increase of 2%. According to S.M.L, the expected return is therefore by applying:

$$\mathrm{E}(R_A)=(\mathrm{E}(R_M)-R_f)\beta+R_f=(8\%-5\%)\cdot 2+5\%=11\% \tag{15.3.815}$$

Remark

The reader will have easily noticed that in the above relation, if the beta is equal to unity, the expected return of asset A is then equal to the expected return of the market. The CAPM is then only a tool for determining excess return that we can expect wait when one invests in a portfolio or a company (company whose return is estimated on the basis of its dividend).

The latter relation can also be found in the literature under the following form (without the error term):

$$\mathrm{E}(R_A)=\lambda\beta+R_f \qquad (15.3.816)$$

with λ that is named the "premium per unit of risk" or just "risk premium" (excess of return required by investors when they put their money in the market benchmark portfolio rather than any other asset) and the ordinate is obviously the risk free interest rate (usually based on government bonds).

We have already met this relation (named "single-factor benchmark return") previously on page 7717 but in the following form:

$$R_i=\alpha_i+\beta_i I+\varepsilon_i \qquad (15.3.817)$$

Finally, let indicate that in theory (in the case of no arbitrage opportunity) we should have:

$$\mathrm{E}(R_A)-\left[\left(\mathrm{E}(R_M)-R_f\right)\beta+R_f\right]=0 \qquad (15.3.818)$$

but in practice with the numerical values of the market the result can be different from zero. This "anomaly" in the market is therefore named the "Jensen alpha":

$$\alpha_J=\mathrm{E}(R_A)-\left[\left(\mathrm{E}(R_M)-R_f\right)\beta+R_f\right] \qquad (15.3.819)$$

If Jensen's alpha is greater than 0, this obviously means that the portfolio asset beats its reference market portfolio. If it is less than 0, the asset portfolio underperformed the market portfolio.

The CAPM therefore stipulates that the expected rate of return (or that should require a rational investor having a risk aversion) of risky assets must be equal the rate of return of risk-free asset, plus a risk premium named sometimes "credit spread". In this case, the relation between systematic risk and expected return remains linear.

It may be interesting to clarify the mathematical assumptions on which were based the developments we have made so far! So these are the assumptions of the CAPM, some of which seem difficult to accept. However the reader should not forget that the validity of a model does not depend on the realism of its assumptions but on the conformity of its implications with reality!!

We have for recall stated explicitly or implicitly the following assumptions:

H1. Investors make up their portfolios being concerned exclusively by the expected return and the variance of these

H2. Investors are risk averse: they do not like risk

H3. There is no transaction cost (which is a joke in the case of delta-hedging[27]...)

H4. The assets are perfectly divisible

H5. Neither dividends nor capital gains are taxed

---

[27] most modern software can take transaction costs into account

H6. Many buyers and sellers are involved in the market and none of them can have influence on prices (perfect information)

H7. All investors can lend or borrow the amount they wish at the risk free rate and that latter is known by all

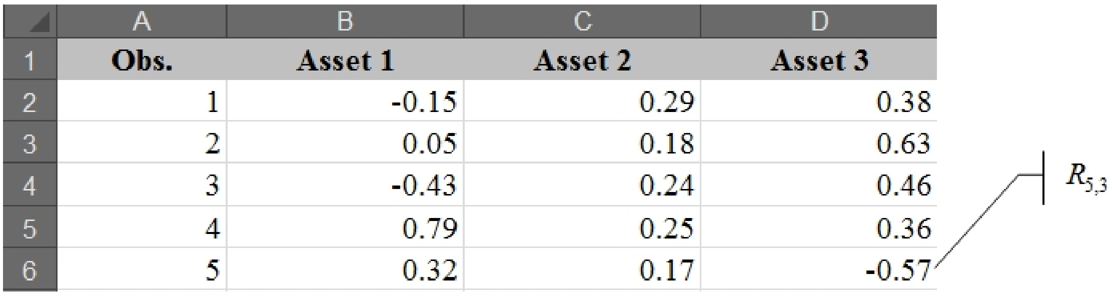
H8. The anticipation of the different investors are homogeneous

H9. The investment period is the same for all investors

Other more modern and complex models have been developed since obviously as we have already mention it multiple times! Taking into account for example the asymmetrical distribution of returns (non-normal) or based on the semivariance rather than the variance, base on a priori estimates of the market behaviour, or based not only on a single factor, etc.

✍**Example:**

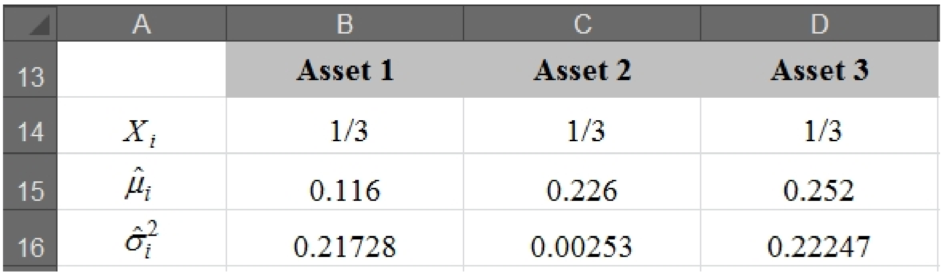
For the example let us return to our example of the Markowitz model for which we had for recall for the assets:



	A	B	C	D
1	**Obs.**	**Asset 1**	**Asset 2**	**Asset 3**
2	1	-0.15	0.29	0.38
3	2	0.05	0.18	0.63
4	3	-0.43	0.24	0.46
5	4	0.79	0.25	0.36
6	5	0.32	0.17	-0.57

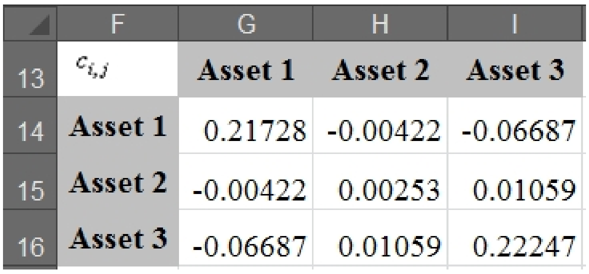
$R_{5,3}$

with some elementary indicators:



	A	B	C	D
13		**Asset 1**	**Asset 2**	**Asset 3**
14	X_i	1/3	1/3	1/3
15	$\hat{\mu}_i$	0.116	0.226	0.252
16	$\hat{\sigma}_i^2$	0.21728	0.00253	0.22247

And for the covariance matrix:



	F	G	H	I
13	$c_{i,j}$	**Asset 1**	**Asset 2**	**Asset 3**
14	**Asset 1**	0.21728	-0.00422	-0.06687
15	**Asset 2**	-0.00422	0.00253	0.01059
16	**Asset 3**	-0.06687	0.01059	0.22247

Now we will determine the capital market line for this example which is for recall the line formed by all portfolios composed of assets without risk, on one hand, and the market portfolio, on the other. By construction, it associates at each level of risk, the highest expected return.

To determine this line with Microsoft® Excel® we set first an initial rate of return without risk R_f that we arbitrarily take as worth 0.22 (in Switzerland in the year 12011 according to holocene calendar, the risk-free yield was estimated at 1% for example...) and to simplify the analysis with the spreadsheet software we will take the naive approach consisting by interpolating the efficiency frontier by the equation of a parabola such that:

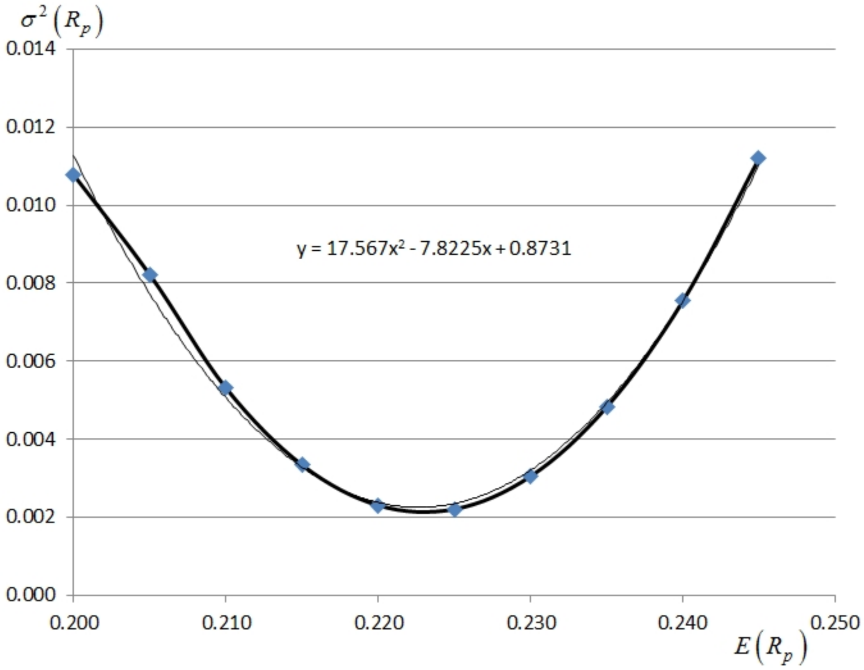


Figure 15.166: **Approximated parabola to Markowitz efficiency frontier in Microsoft® Excel® 14.0.7173**

So we have the approximate Markowitz equation curve:

$$y=17.567x^2-7.8225x+0.8731=ax^2+bx+c \qquad (15.3.820)$$

By the construction of the CML we can already draw approximately:

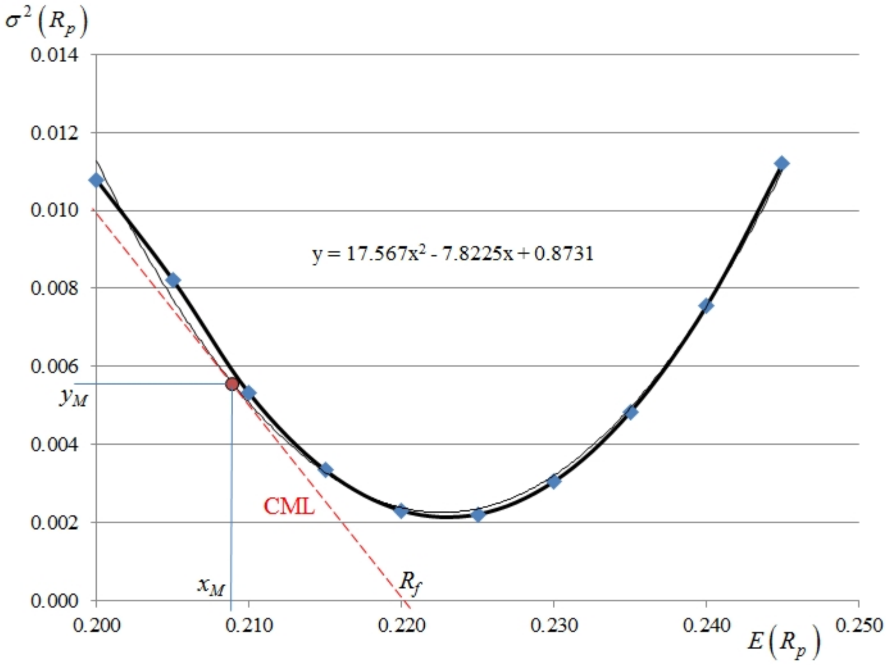


Figure 15.167: Approximated CML on Markowitz efficiency frontier in Microsoft® Excel® 14.0.7173

and let us write the CML equation by:

$$y' = a'x + b' \tag{15.3.821}$$

with the condition (see on the plot above):

$$a' = -\frac{b'}{0.22} \tag{15.3.822}$$

We then have two equations with two unknowns to solve this problem (the intersection of the line and parabola for the first equation and the equality of the slope of the line and the parabola at the intersection for the second equation):

$$\begin{aligned} &ax_M^2 + bx_M + c = a'x_M + b' \\ &2ax_M + b = a' \end{aligned} \tag{15.3.823}$$

The second equation gives us:

$$x_M = \frac{a' - b}{2a} = \frac{\dfrac{-b'}{0.22} - b}{2a} \tag{15.3.824}$$

Injected into the first equation:

$$a\left(\frac{\dfrac{-b'}{0.22} - b}{2a}\right)^2 + b\left(\frac{\dfrac{-b'}{0.22} - b}{2a}\right) + c = -\frac{b'}{0.22}\left(\frac{\dfrac{-b'}{0.22} - b}{2a}\right) + b' \tag{15.3.825}$$

If we solve this polynomial of the second degree we have two real solutions (Microsoft® Excel® can not determine the roots of this directly but with Maple™ 4.00b it's very simple):

$$b'_1 = -0.06822748631 \qquad b'_2 = 0.120763489 \tag{15.3.826}$$

Following a request from a user, here are the corresponding Maple™ lines:

```
>a:=18.795;b:=-8.3892;c:=0.9384;
>f:=a*(((-d/0.22)-b)/(2*a))^2+b*(((-d/0.22)-b)/(2*a))+c
=(-d/0.22)*(((-d/0.22)-b)/(2*a))+d;
>solve(f,d);
```

The second solution is to eliminate (we know it by trying to take it as a solution). So we have finally:

$$b'=-0.06822748631 \qquad a'=-\frac{b'}{0.22}=0.3101249378 \tag{15.3.827}$$

Which gives graphically (obviously the CML gives as expected the highest return for a given variance so it was impossible for the line to be left oriented):

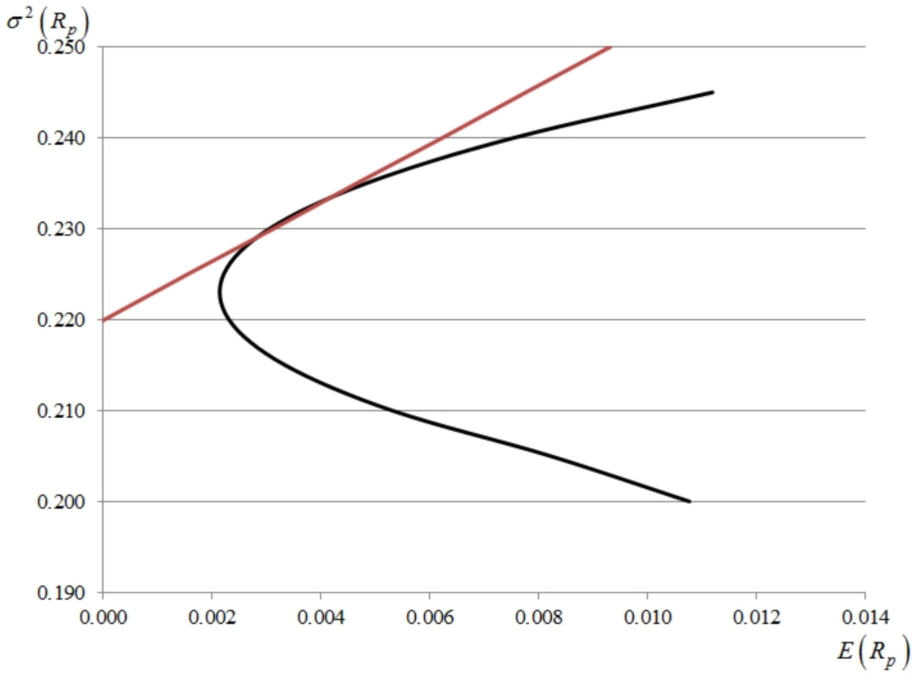


Figure 15.168: **Approximated parabola to Markowitz efficiency frontier with CAPM and traditional orientation Microsoft® Excel® 14.0.7173**

It also comes immediately:

$$x_M=\frac{\dfrac{-b}{0.22}-b}{2a}=0.2314265746 \tag{15.3.828}$$

Thus, by reusing the solver as above but with this new value for the expected mean we get for an optimal market portfolio with a risk-free asset having a return R_f of 0.22, an overall return of $0.2314276...$ the following asset weights:

$$X_1=-0.0399 \qquad X_2=1 \qquad X_x=0.0399 \tag{15.3.829}$$

with:

$$\sigma(R_P)=0.068 \tag{15.3.830}$$

So it is by construction the portfolio which, among all portfolios having only risky assets, maximizes performance while it behaves such as having a risk-free asset while minimizing the risk!!! Again, remember that this portfolio is named the "tangent portfolio".

If the reader refer to our MATLAB® companion book he can find the script that leads to the following results:

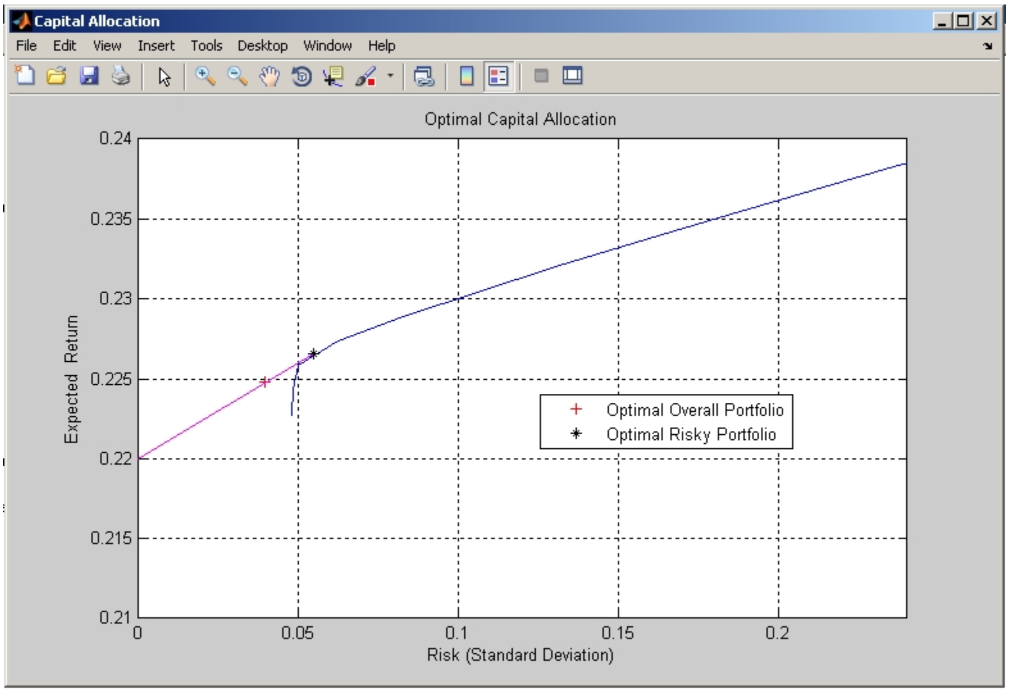
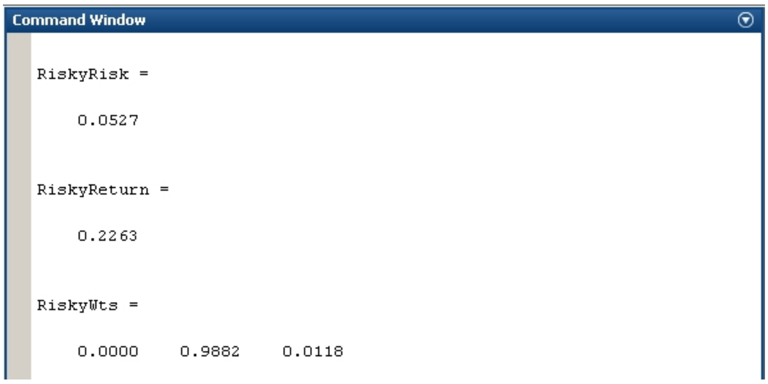


Figure 15.169: **Markowitz efficiency frontier with CAPM and CML in MATLAB® 2013a**

and the corresponding numerical values of interest:



```
RiskyRisk =

    0.0527


RiskyReturn =

    0.2263


RiskyWts =

    0.0000    0.9882    0.0118
```

Therefore we can see without surprise a significant difference between our rough Microsoft® Excel® approach and the results given by MATLAB®.

At this point the reader may have noticed that capital asset line and security market line represent assets on a risk and return map with the following correspondence:

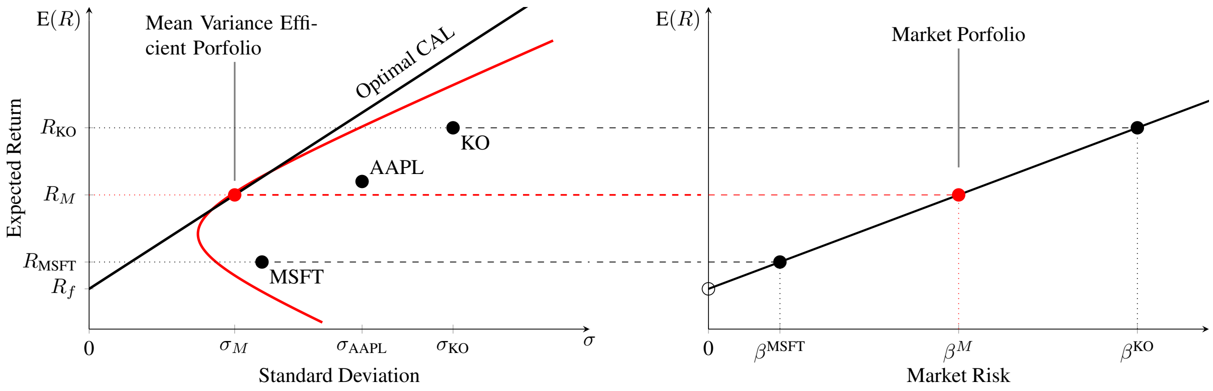


Figure 15.170: Security Market Line and Capital Market Line mapping

In practice we run a second stage regression that involves a single "cross-sectional regression" of the average (over time) of the stock returns on a constant and the β:

$$\bar{R}_i = \lambda_0 + \lambda_1 \beta_i + \varepsilon_i \tag{15.3.831}$$

where $\bar{R}_i$ is the return for stock i averaged over the 60 months. If the Sharpe-Lintner-Mossin CAPM is a valid model, two key predictions arise which can be tested using this second stage regression, that is: $\lambda_0 = R_f$ and $\lambda_1 = \text{E}(R_M) - R_f$.

It seems that it has been found that returns are systematically higher for small capitalisation stocks and are systematically higher for value stocks than the Sharpe-Lintner-Mossin CAPM would predict. That's why sometimes a different second-stage regression is run having for purpose to test this directly using a different augmented second stage regression:

$$\bar{R}_i = \alpha + \lambda_1 \beta_i + \lambda_2 \text{MV}_i + \lambda_3 \text{BTM}_i + \varepsilon_i \tag{15.3.832}$$

where MV_i is the market capitalisation for stock i and BTM_i is the ratio of its book value to its market value of equity. Again, if the Sharpe-Lintner-Mossin CAPM is a valid and complete model, then we should see that $\lambda_2 = 0$ and $\lambda_3 = 0$.

15.3.10.3.8.1 Hedging a future and commodity portfolio with the CAPM

We have seen during our introduction to Futures how to hedge a commodity portfolio based on a replication technique (see page 7657) or based on the modified duration (see page 7661).

But now let us see that we can also hedge using the betas of the Futures and their respective underlying commodity based on the CAPM model!

Remark

For information... the redactor of the following lines is not particularly a fan of the method that will follow since it is based on a lot of approximations but... as it is part of the CFA (Chief Financial Analyst) curriculum we will just close one eye...

First, following the CAPM model, let us say that for two assets that are a Commodity and a Future the following linear relation applies:

$$\begin{aligned} S &= \beta_{S,I}I+\alpha_{S,I}+\varepsilon_{S,I}\cong\beta_{S,I}I \\ F &= \beta_{F,I}I+\alpha_{S,I}+\varepsilon_{F,I}\cong\beta_{F,I}I \end{aligned} \quad (15.3.833)$$

and that we can write the following approximation (the residual return and the error term are considered negligible...):

$$\begin{aligned} \frac{\Delta S}{S} &\cong \beta_{S,I}\frac{\Delta I}{I}=\beta_S\frac{\Delta I}{I} \\ \frac{\Delta F}{F} &\cong \beta_{F,I}\frac{\Delta I}{I}=\beta_F\frac{\Delta I}{I} \end{aligned} \quad (15.3.834)$$

Then we have according to the relation see at page 7657:

$$\Delta V=\Delta S+N\Delta F\cong\beta_S S\frac{\Delta I}{I}+N\beta_F\frac{\Delta I}{I}=(\beta_S S+N\beta_F F)\frac{\Delta I}{I} \quad (15.3.835)$$

Note that the term in parentheses is in numerals, so we can not put something like this:

$$(\beta_S S+N\beta_F F)\frac{\Delta I}{I}=\beta_T\frac{\Delta I}{I} \quad (15.3.836)$$

Which would not be homogeneous!

So a very simple idea might be this: if the term in parentheses has a $N=0$ in the sense that we do not cover ourselves against any risk, then we must necessarily have (β_T is a synthetic "total" beta):

$$(\beta_S S+0\cdot\beta_F F)\frac{\Delta I}{I}=\beta_T S\frac{\Delta I}{I} \quad (15.3.837)$$

Reason why for what will follow we will focus on:

$$\beta_S S+N\beta_F F=\beta_T S \quad (15.3.838)$$

From which we can trivially draw two major relations:

- The first is that if we want to eliminate all risk then:

$$\beta_T = 0 \Rightarrow \beta_S S + N \beta_F F = 0 \tag{15.3.839}$$

and therefore we get one of the famous relation available in the CFA:

$$N^*_{\beta=0} = -\frac{\beta_S}{\beta_F}\frac{S}{F} \tag{15.3.840}$$

- The second is that if we want to achieve a given total synthetic beta, then we have:

$$N^*_{\beta_T} = \frac{\beta_T S - \beta_S S}{\beta_F F} = \frac{\beta_T - \beta_S}{\beta_F}\frac{S}{F} \tag{15.3.841}$$

which is sometimes denoted:

$$N^*_{\beta_T} = \frac{\$\beta_T - \$\beta_S}{\$\beta_F} \tag{15.3.842}$$

We see that we fall back on the previous relation if $\beta_T = 0$. Let us recall that if the sign of this last relation is positive we will have to take a long position (buyer of future), otherwise (negative) a short position (Future seller).

15.3.10.3.9 Black & Scholes option pricing models

It is to the genius of three famous mathematicians that the derivatives market (which it is usual to classify into three large families: Derivatives on securities, Derivatives on rates, Derivatives on currencies) owes its success, thanks to Black & Scholes equation established in the years 11970 according to holocene calendar (and published in 11973) that theoretically determines the exact premium (under several assumptions) that a customer must pay to acquire a Call or a Put and the strategy to be followed by the seller of these Options to hedge risk (to cite only the most well known example). Obviously this model works only if the temporal periods considered are relatively short (on the order of the week or a few months at best). Beyond the use of this particular theoretical model is quite a joke!

Remark

The dynamics of the Black & Scholes model had already been well dusted before by Louis Bachelier and Edward O. Thorpe, which is why we sometimes refer to the "Bachelor-Thorpe" model.

Fischer Black, Myron Scholes and Robert C. Merton are the ancestors of a generation of sophisticated derivative products, giving a lexicon of terms as exotic as Butterflies, Rainbows, Knock-in, Knock-Out, Barriers, Swaps, Calls, Puts, Sneakers, Swings, Caplets. This model is also considered this said as one of the main factors of the stock market crash of 11987 (holocene calendar) by some specialists ...

Evidently, however, there are many more sophisticated empirical models, such as:

- The Black & Scholes & Barenblatt model
- The Merton Jump Scattering model
- The Heston model
- and so on...

What we would like in what follows is to determine the theoretical value of the premium of an option from the following five data:

- The present value of the underlying financial asset of the option (determined by market speculation)
- The time remaining to the option before its maturity (chosen by the issuing company)
- The strike price fixed by the issuer subjectively or after modelling.
- The risk-free interest rate (assumed to be the expected rate of return of the underlying asset).
- The volatility (standard deviation) of the underlying price of the option (measured on the market).

The premium for the option thus determined should be unique and equitable for both parties. Indeed, the system of options would make it possible to charge an option price (premium) higher than the market forecasts and thus to generate a certain profit and from nothing, but the many players in the market will compete to be fair and attract the customers on their options rather than on those of the competitors.

The modelling of options prices (Black & Scholes) is based on the use of stochastic differential calculus. Thus, the Black & Scholes approach assumes that the evolution of the asset defines a geometric Brownian motion (in the sense that the possible movements of the price tend towards infinity) and that its output defines a Wiener process (a concept that we will define later on).

15.3.10.3.9.1 Put-Call parity equation

Before attacking somewhat arduous stochastic calculations, it is useful to establish beforehand a so-named "Put-Call parity equation" which will serve as a kind of conservation law to verify the validity of the results that we will establish further for the evaluation of options prices.

The goal will be to answer the following question: What amount of money M must we pay now to receive a guaranteed sum corresponding to the price "strike price" denoted K to a future time T?

To answer this question, let us recall that during our study of the calculation of interest we saw that in considering a capital C and a constant interest rate r we had in a case of continuous capitalization:

$$C=C_0e^{rt} \tag{15.3.843}$$

Therefore, if we but $C=K$ and $C_0=M$ we have then in a universe without risk, that is to say in delta-neutral (the case with risk will be the subject of the Black & Scholes that we will see later):

$$K=Me^{rt} \tag{15.3.844}$$

Therefore:

$$M=Ke^{-rt} \tag{15.3.845}$$

But this relationship is not quite right. Indeed, we must have $M=K$ that is guarantee at time $T=t$. Hence we are naturally led to put (this is a strictly increasing function!):

$$\boxed{M=Ke^{-r(T-t)}} \tag{15.3.846}$$

This small elementary reminder being made, we will now assume for what will follow that the Call and the Put possess the following properties:

P1. Same support that is valued S at time t (spot price).

P2. Same maturity T

P3. Same strike price K

And the following assumptions (which are the basis of the Black & Scholes model that we will describe below):

H1. There are no transaction costs (which is a joke in delta-hedging strategies...)

H2. Delta-hedging eliminates all risk (see further below the concept)

H3. The support is not a term instrument (it is possible to make short sales)

H4. The support does not pay dividends over the life of the option (i.e. between $[0,T]$).

H5. The options are European

H6. The risk-free market rate is assumed to be known in advance and considered as constant

H7. There is no arbitrage opportunities (but well in practice ... there is almost only this)

H8. The price of options does not affect the price of the underlying (mouarf!)

H9. Yields are not autocorrelated

By asking us again now the initial question: How much do we have to pay now for a portfolio in order to receive a guaranteed sum K (strike price) at a future time T?

Since the portfolio can be considered as a black box, there is nothing to prevent us by denoting by S the actual price of the underlying, P that of the Put and C that of the Call (see just below for an explanation of the origin of this equality):

$$\boxed{\Pi = S + P - C = Ke^{-r(T-t)}} \tag{15.3.847}$$

which is nothing else than the "Put-Call parity equation" or otherwise written if we put the beginning of time at zero, then:

$$S + P - C = Ke^{-rT} \tag{15.3.848}$$

Remark

Some practitioners then use the following relation, derived from the above one, to determine the risk free rate of the market:

$$r = -\frac{1}{T-t}\ln\left(\frac{K}{S_t + P_t - C_t}\right) \tag{15.3.849}$$

Strictly speaking, it would be better to write this last relation in the following form:

$$C_t - P_t = S_t - Ke^{-r(T-t)} \tag{15.3.850}$$

Whatever the rearrangement of the terms in this last equality, there must always be equality, otherwise there is an arbitrage opportunity. This relation also shows that the value of a European

Call with an exercise price K and maturity T can be deduced from that of a European Put with the same exercise price K and the same maturity T.

Let us see more in details where the Put-Call parity equation comes from the following request of a reader:

Theorem 15.7. The Put-Call parity equation arises from the absence of arbitrage opportunity.

Proof 15.7.1. Let us consider the frequent and natural configuration of a first economic agent with a portfolio P_A containing a European Call option (C) and an amount Ke^{-rT} invested in a risk-free investment. The initial value (we must put $t=0$) of this portfolio will therefore be:

$$P_A = C + Ke^{-rT} \tag{15.3.851}$$

Let us consider the second configuration naturally opposed to the first portfolio containing a European Put of price P and the underlying of price S. The initial value of this portfolio will therefore be:

$$P_B = P + S \tag{15.3.852}$$

We then have two scenarios:

S1. At maturity T, if the underlying has a value less than or equal to the strike price (i.e. $S_T \leq K$), the portfolio P_A will have for value K because the Call option will most likely not be exercised and therefore its financial value becomes zero but the risk-free asset (placed at the risk-free rate) will have a return equal K. The portfolio P_B will be equal to $(K-S_T)+S_T=K$ where the term in parentheses represents the gain generated by the Put.

S2. At maturity T, if the underlying has a value higher than the strike price (i.e. $S_T > K$), we are therefore in a situation contrary to that of the first scenario. The portfolio P_A will then have for value $(S_T-K)+K=S_T$. The portfolio P_B will be equal to $0+S_T=S_T$ because the sale will not be made and the Put option will be considered as having a zero value.

To summarize:

	If $S_T \leq K$	**If $S_T > K$**
P_A	$0+K=K$	$(S_T-K)+K=S_T$
P_B	$(K-S_T)+S_T=K$	$0+S_T=S_T$
Conclusion:	$P_A=P_B$	$P_A=P_B$

Table 15.52: **Summary of lack of absence of arbitrage opportunity**

So the two portfolios are always equal at maturity otherwise there is an arbitrage opportunity. What we can write (the index T for the price of the underlying at maturity is often omitted):

$$C + K = P + S \tag{15.3.853}$$

where the exponentials has vanished since $e^{-r(T-t)}$ for $T=t$ is equal to 1.

■ Q.E.D.

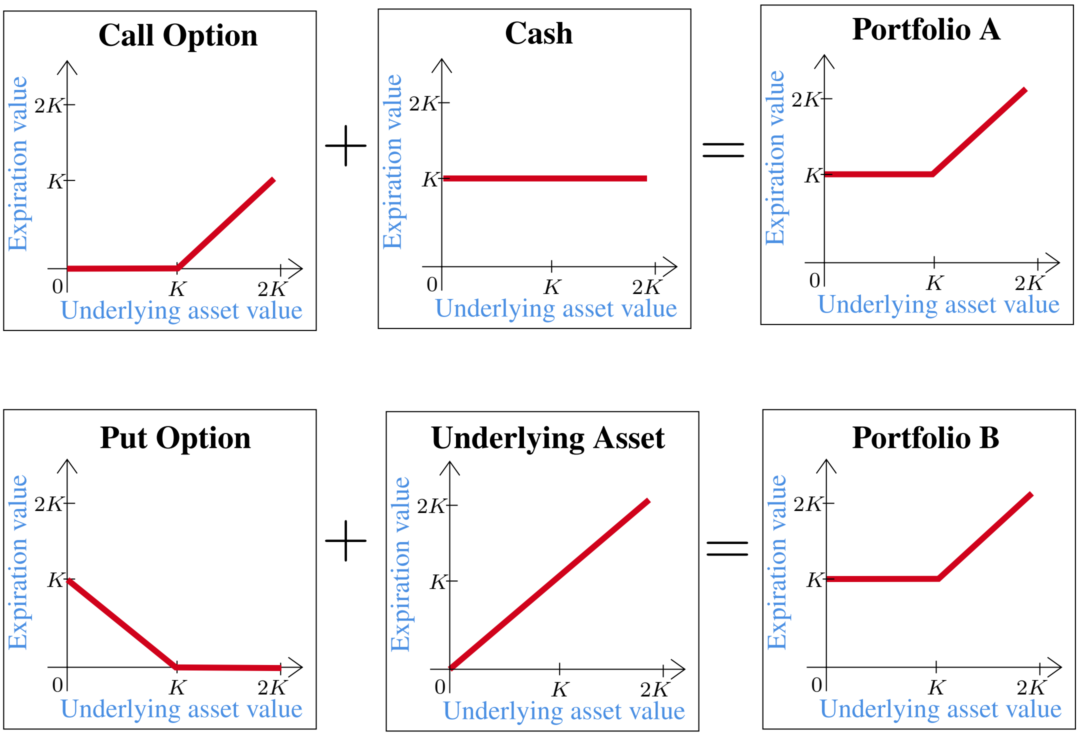


Figure 15.171: Put-Call parity equation illustrated

15.3.10.3.9.2 Efficient Market Hypothesis

The Black & Scholes model (and many other financial models) is based on the premise that the market is "efficient".

Definition (#15.3.695): An "efficient market" is a market where prices fully reflect all available information. Thus, if the market is efficient, it is not possible to make abnormal profits.

We can distinguish three types of efficient markets that depend on the type of information available:

1. The efficient market hypothesis in "weak form", which explicitly states that prices reflect all the information contained in the historical price series.
2. The efficient market hypothesis in "semi-strong form" establishes that prices reflect all available public information.
3. The "strong form" market assumption that all known public and private information is reflected in market prices.

Several studies have attempted to test the hypothesis of asset market efficiency. To test the weak form of the hypothesis, the analysis of the time series (see below) was used by specifically testing the hypothesis of a random walk (a "Brownian motion" - we will return to this later). More specifically, these tests have tried to check whether price increases are independent of past increases. If the random walk hypothesis is rejected, then the market is not efficient, since past price increases could help to anticipate future asset prices. The empirical evidence supports the efficient market hypothesis in weak form. In order to test the semi-strong form of the hypothesis, the rate of adjustment of market prices to the arrival of new information has been evaluated. The evidence for a rapid adjustment of market prices is dominant. The strong form of the market efficiency hypothesis is to test whether it is possible to take advantage of privileged information (information accessible to a small group of economic agents). Since non-public information can not be identified, a strong form test type considers reviewing the investment performance of individuals or groups who may have private information. Edwin J. Elton and Martin J. Gruber (11984 according to holocene calendar) report that the analysis of the performance of mutual funds, after deducting costs, supports the strong form of efficiency.

This implies the following assumptions (to summarize):

H1. The past history of the asset price is completely reflected in the present price which does not contain any other information about the asset.

H2. The market immediately incorporates any new information into the price of an asset.

The paradox of the Hypothesis of Market Efficiency is that if every investor truly believed that the market was perfectly efficient, then nobody would study companies, their balance sheets, and so on. It would suffice to buy the index. Indeed, efficient markets rely on individuals who are active on the market because they believe (and their believes are true when there is an opportunity arbitrage!) that this market is "inefficient" and that they can do better than the market!

This hypothesis is the source of much debate in the field ...

Remark

With the two assumptions previously stated, any unanticipated change in the asset price is named a "Markov process[a]".

---

[a] A Markov process is a process whose future evolution depends on its past only through its state at the present moment. The course of an action is probably not a Markov process (the "memory" of the process is probably longer - for example, a seasonal trend).

15.3.10.3.9.3 Wiener Process

Let Δz be the variation of the non-trend value of an asset over a small time interval denoted Δt.

Using the knowledge of the two major results of the Bachelier model, we therefore have for the variations of the value of the asset a positive expected mean dependent in a manner proportional to the square root of time according to:

$$\boxed{\Delta z=\varepsilon\sqrt{\Delta t}} \tag{15.3.854}$$

Where we put as hypothesis (acceptable ... because we work on small variations for recall!) that the coefficient of instability is a function:

$$\varepsilon \longrightarrow \mathcal{N}(0,1) \tag{15.3.855}$$

where $\mathcal{N}(0,1)$ is as we know the notation of the reduced Normal law as we have established it in the Statistics section.

Remark

Often in the field of economics, we denote it $W\mathcal{N}$ instead of $\mathcal{N}$ in tribute to Wiener.

That said, the antecedent relation is often denoted in a more generalized way as following:

$$\boxed{W_t=W_{t-1}+\varepsilon\sqrt{\Delta t}=W_{t-1}+\mathcal{N}(0,1)\sqrt{\Delta t}} \tag{15.3.856}$$

and defined as a "standard Brownian motion" with "white noise" (Normal marginal law), or "arithmetic Brownian motion", where the W is there in tribute to Wiener (such a process is also sometimes denoted GWN for "Gaussian White Noise" as we can obviously create white noises from other distributions)! It is interesting to notice that the Brownian motion is supposed indefinitely divisible (which means that the temporal period taken does not influence the law of probability which remains always the same ... it is a fractal property of the Brownian motion which has been studied by Benoît Mandelbrot too!).

Do not forget for a proof that we will do further below that we can deduce that:

$$W_t-W_{t-1}=\mathcal{N}(0,1)\sqrt{\Delta t}=\mathcal{N}(0,1)\sqrt{(t-1)-t}=\mathcal{N}(0,1) \tag{15.3.857}$$

It is possible to produce a plot of this Brownian motion with Microsoft® Excel® (here the version 11.8346) with in the column `A` the time with a typical equation step of 0.01 [s] (column which will be there only for the sake of comfort of reading and by tradition) and in cell `B2` the following formula:

```
=B1+NORMSINV(RAND())*SQRT(0.01)
```

where `B1` contains the value 0.

We then get the following variations of values for 4 columns of the same type:

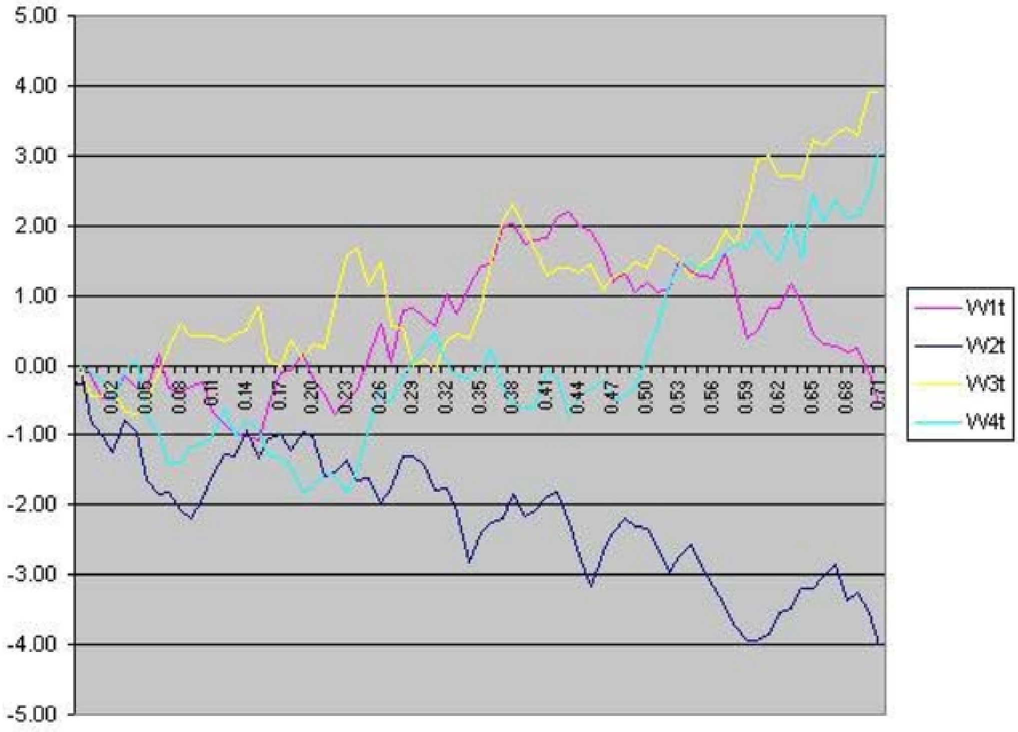


Figure 15.172: Example of 4 standard Brownian motions with Microsoft® Excel® 11.8346

Or in a much better quality using TikZ:

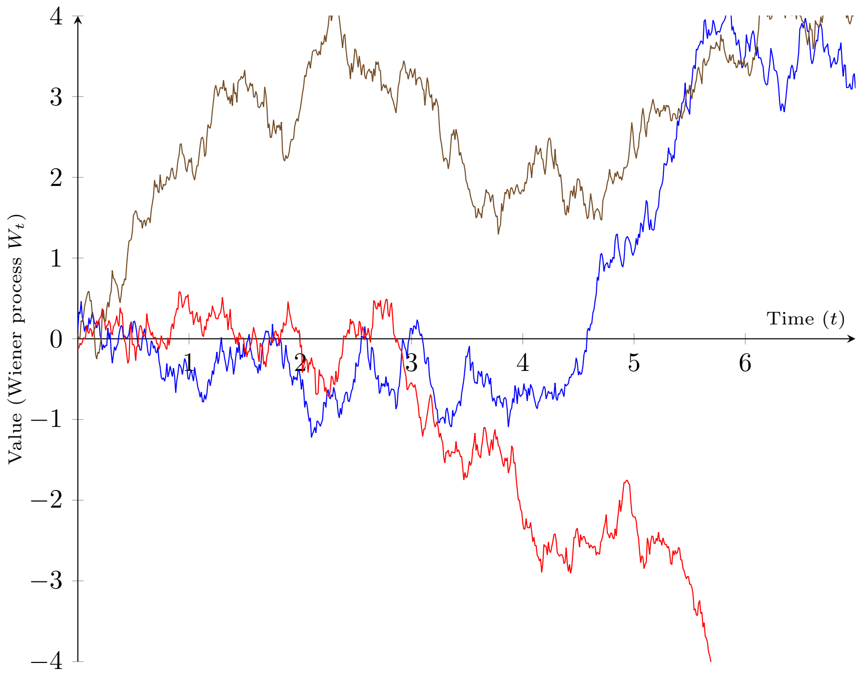


Figure 15.173: **Wiener process (normal random walk) with TikZ**

Standard Brownian movements have some remarkable properties as we can see: the trajectory tends to alternate above and below the abscissa axis. This comes from the fact that the Normal law considered has zero mean, in other words that there is no general tendency to increase or decrease the variations (to check it made at least $30,000$ points in Microsoft® Excel® and you will see...).

It is easily possible to characterize Δz with its mean:

$$\text{E}(\Delta z)=\text{E}\left(\varepsilon\sqrt{\Delta t}\right)=\sqrt{\Delta t}\cdot\text{E}(\varepsilon)=\sqrt{\Delta t}\cdot 0=0 \qquad (15.3.858)$$

Indeed, let us recall that for the Normal reduced centered law, we have:

$$\mu=\text{E}(x)=0 \qquad (15.3.859)$$

So we could expect this result of total absence of general trend (it was quasi-intuitive!).

We can also characterize Δz using its variance:

$$\text{V}(\Delta z)=\text{V}\left(\varepsilon\sqrt{\Delta t}\right)=\Delta t\cdot\text{V}(\varepsilon)=\Delta t\cdot 1=\Delta t \qquad (15.3.860)$$

hence:

$$\sigma_{\Delta z}=\sqrt{\Delta t} \qquad (15.3.861)$$

Indeed, let us recall that for the Normal reduced centered law, we have:

$$\sigma_x^2=\text{V}(x)=1 \qquad (15.3.862)$$

Finally (in fact this result immediately follows from the linearity property of the Normal law):

$$\Delta z\longrightarrow\mathcal{N}\left(0,\sqrt{\Delta t}\right) \qquad (15.3.863)$$

So when we have:

$$\begin{aligned}&\mathrm{E}(\Delta z)=0\\&\mathrm{E}(\Delta z^2)=\mathrm{E}(\varepsilon\Delta t)=\Delta t\mathrm{E}(\varepsilon)=\Delta t\end{aligned}\tag{15.3.864}$$

it is customary to speak of the "Gauss-Wiener process" or more simply of "Wiener process".

To summarize a little bit all this...:

1. We knew with Bachelier's model that the positive expected mean and the positive standard deviation of the value are proportional to the square root of time. We used these two results here.
2. We now know (under the well-defined hypothesis of a Normal type coefficient) that the variations have zero expected mean (trend) and a standard deviation proportional to the square root of the temporal variation.

The property which has just been established remains valid for a large time interval denoted T corresponding to n small intervals Δt! In other words:

$$T=n\cdot\Delta t\tag{15.3.865}$$

In this context, it is appropriate to replace large variations by Δx such that:

$$x(T)-x(0)\tag{15.3.866}$$

But without any trend, we have for the moment the equality:

$$x(T)-x(0)=\sum_{i=1}^{n}\Delta x_i=\sum_{i=1}^{n}\Delta z_i=\sum_{i=1}^{n}\varepsilon_i\sqrt{\Delta t}\tag{15.3.867}$$

As in the case of a change in the price over a short period of time, it is possible to characterize $x(T)-x(0)$ using its expected mean and its standard deviation:

$$\begin{aligned}&\mathrm{E}\big(x(T)-x(0)\big)=\mathrm{E}\left(\sum_{i=1}^{n}\varepsilon_i\sqrt{\Delta t}\right)=\sqrt{\Delta t}\mathrm{E}\left(\sum_{i=1}^{n}\varepsilon_i\right)=\sqrt{\Delta t}\sum_{i=1}^{n}\mathrm{E}(\varepsilon_i)=\sqrt{\Delta t}\cdot 0=0\\&\mathrm{V}\big(x(T)-x(0)\big)=\mathrm{V}\left(\sum_{i=1}^{n}\varepsilon_i\sqrt{\Delta t}\right)=\Delta t\mathrm{V}\left(\sum_{i=1}^{n}\varepsilon_i\right)=\Delta t\sum_{i=1}^{n}\mathrm{V}(\varepsilon_i)=\Delta t\sum_{i=1}^{n}1=n\cdot\Delta t=T\end{aligned}\tag{15.3.868}$$

which makes sense...

We then fall back, for a large time interval T:

$$x(T)-x(0)\longrightarrow\mathcal{N}(0,\sqrt{T})\tag{15.3.869}$$

Which we can also write in the following form using the properties proven in the section Statistics of the Normal law:

$$x(T)\longrightarrow\mathcal{N}(x(0),\sqrt{T})\tag{15.3.870}$$

Result we could reasonably expect with the aforementioned assumptions ...

This last result is written in the following explicit form in spreadsheets softwares:

$$\boxed{x(t)=x(0)+\mathcal{N}\sqrt{t}} \tag{15.3.871}$$

and we see that this is unrealistic because it would mean that any financial asset follows the same law (regardless of its volatility ...) and would have no general downward or upward trend. We will see how to improve this approach.

To close this approach, notice that if Δt tends to 0 (which amounts to considering a subdivision of time T into extremely small intervals), the course undergoes an infinitely large number of variations over period T. In other words, the process of evolution of the price of the asset is continuous, which leads to replace Δt by $\mathrm{d}t$, Δx by $\mathrm{d}x$ and Δz by $\mathrm{d}z$.

In this case, we get:

$$\mathrm{d}x \longrightarrow \mathcal{N}(0,\sqrt{\mathrm{d}t}) \tag{15.3.872}$$

which defines a "Wiener process" (we will come back to this when we will have established the stochastic differential equation).

But obviously this does not really conform to the reality as we have already mentioned ... We prefer then to add a constant shift in time which gives the Brownian motion that we will see now.

15.3.10.3.9.4 Generalized Brownian motion

In this case (generalization a little more realistic), the evolution of the course depends not only on a standard Brownian random process (second term below on the right of the equality), but also on a parameter of central tendency, or "drift" (first term below to the right of the equality):

$$\boxed{\mathrm{d}x = a \cdot \mathrm{d}t + b \cdot \mathrm{d}z} \tag{15.3.873}$$

With always:

$$\mathrm{d}z = \varepsilon\sqrt{\mathrm{d}t} \tag{15.3.874}$$

and:

$$\varepsilon \longrightarrow \mathcal{N}(0,1) \tag{15.3.875}$$

We thus have a "generalized Brownian motion", consisting of a standard Brownian motion ($\mathrm{d}z$ represented by a Normal law of zero expected mean and variance $\mathrm{d}t$ as we saw above) and a drift. In this scenario, a and b are imposed as constant contrary to the more general case that we shall see later.

The antecedent relation is often represented in the literature in the following differential form:

$$\mathrm{d}X_t = a\mathrm{d}t + b\mathrm{d}W_t = a\mathrm{d}t + b\mathrm{d}(\mathcal{N}(0,1)\sqrt{\mathrm{d}t}) \tag{15.3.876}$$

So graphically this gives, by adding this drift and taking a positive and non-zero value for a, a standard Brownian motion that will tend to alternate above and below the drift:

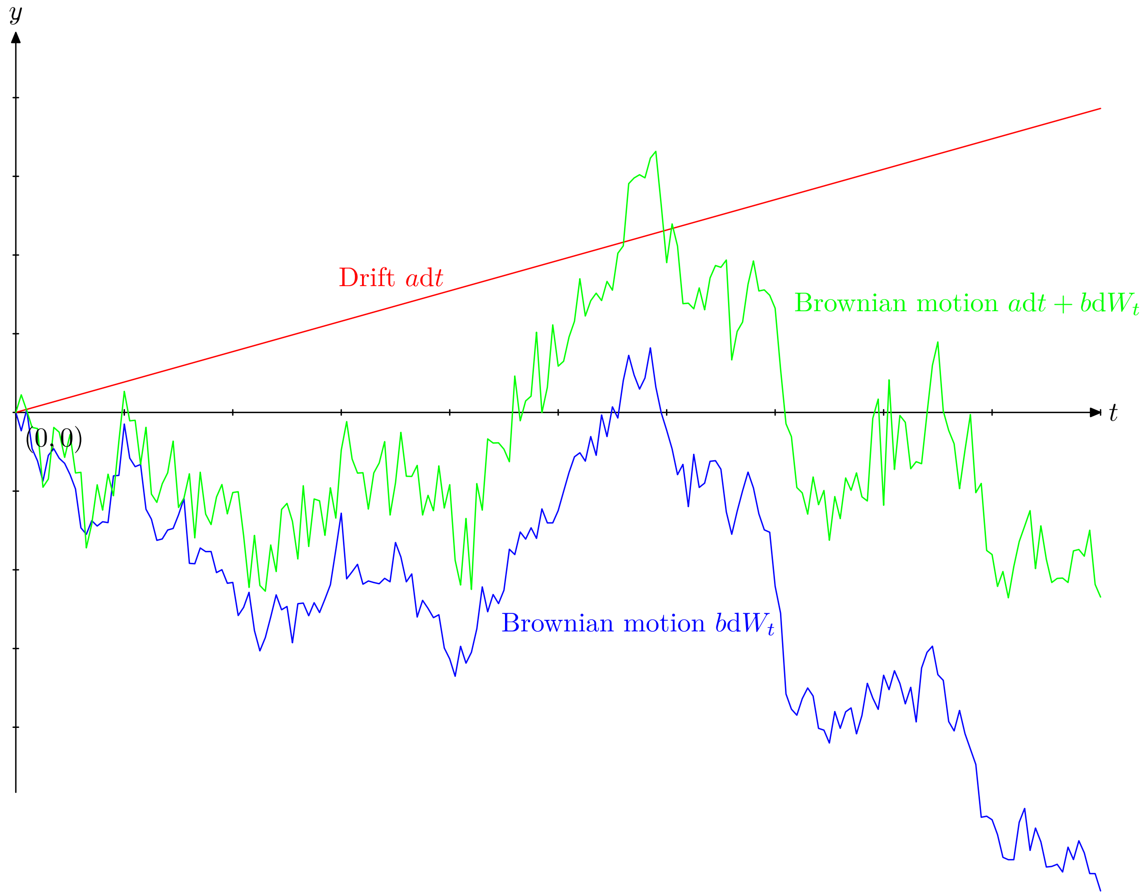


Figure 15.174: Example of standard Brownian motion with drift (and decomposition)

On a small time interval Δt, the process, in discrete time is obviously written:

$$\Delta x = a \cdot \Delta t + b \cdot \Delta z \tag{15.3.877}$$

In this case, we have:

$$\text{E}(\Delta x) = \text{E}(a\Delta t + b\Delta z) = a \cdot \Delta t + b \cdot \text{E}(\Delta z) \tag{15.3.878}$$

since only Δz has a random component.

Therefore:

$$\begin{aligned} \text{E}(\Delta x) &= a \cdot \Delta t + b \cdot \text{E}(\Delta z) = a \cdot \Delta t + b \cdot \text{E}(\varepsilon\sqrt{\Delta t}) = a\Delta t + b\sqrt{\Delta t} \cdot \text{E}(\varepsilon) = a\Delta t + 0 = a\Delta t \\ \text{V}(\Delta x) &= \text{V}(a \cdot \Delta t + b \cdot \Delta z) = 0 + b^2 \cdot \text{V}(\varepsilon\sqrt{\Delta t}) = b^2\Delta t \cdot \text{V}(\varepsilon) = b^2 \cdot \Delta t \end{aligned} \tag{15.3.879}$$

Finally:

$$\Delta x = \mathcal{N}(a\Delta t, b\sqrt{\Delta t}) \tag{15.3.880}$$

The result:

$$\text{V}(\Delta x) = \text{V}\left(\mathcal{N}(a\Delta t, b\sqrt{\Delta t}\right) = b^2\Delta t \tag{15.3.881}$$

is often named the "quadratic variation" of the generalized Brownian motion. The quadratic variation of the standard Brownian motion ($a=0$,$b=1$) being often denoted:

$$\mathrm{V}(B_t)=t \tag{15.3.882}$$

By subdividing a period T into n time intervals Δt (i.e. $T=n\Delta t$), the price change becomes over this period T:

$$x(T)-x(0)=\sum_{i=1}^{n}\Delta x_i \tag{15.3.883}$$

Since then:

$$\begin{gathered}\mathrm{E}(x(T)-x(0))=\mathrm{E}\left(\sum_{i=1}^{n}\Delta x_i\right)=\sum_{i=1}^{n}\mathrm{E}(\Delta x_i)=n\cdot a\cdot\Delta t=a\cdot T\\ \mathrm{V}(x(T)-x(0))=\mathrm{V}\left(\sum_{i=1}^{n}\Delta x_i\right)=\sum_{i=1}^{n}\mathrm{V}(\Delta x_i)=\sum_{i=1}^{n}\mathrm{V}(b\sqrt{\Delta t})=b^2\Delta t\sum_{i=1}^{n}1=n\cdot b^2\Delta t=b^2T\end{gathered} \tag{15.3.884}$$

Finally:

$$x(T)-x(0)=\mathcal{N}\left(a\cdot T,b\cdot\sqrt{T}\right) \tag{15.3.885}$$

Therefore:

$$x(T)=x(0)+\mathcal{N}\left(a\cdot T,b\cdot\sqrt{T}\right) \tag{15.3.886}$$

or also:

$$x(T)=\mathcal{N}\left(x(0)+a\cdot T,b\cdot\sqrt{T}\right) \tag{15.3.887}$$

It is then easy to understand why we say that the Normal law governs the random variable obtained by stopping a Brownian process at a given instant: it is an instantaneous picture of the Brownian motion simple or generalized!

By choosing:

$$a=x(0)\cdot\mu \qquad b=\sigma x(0) \tag{15.3.888}$$

We then have the prior-previous relation which is traditionally written in the following explicit form in spreadsheets softwares:

$$\boxed{x(t)=x(0)+\mu\cdot x(0)\cdot t+\mathcal{N}(0,1)\sigma x(0)\sqrt{t}} \tag{15.3.889}$$

where μ is the average return in $\%$ of the financial asset and σ the volatility of the return in $\%$. Or written otherwise:

$$x(t)=\mathcal{N}\left(x(0)+\mu x(0)t,\sigma x(0)\sqrt{t}\right) \tag{15.3.890}$$

In practice, the risk-free rate is chosen to represent the underlying returns because we assume that we always work in a neutral risk probability (absence of arbitrage opportunity).

It is then interesting for the financial analyst to visualize the expected mean as a function of t and the value $x(t)$ corresponds to a cumulative probability of 2.5% and 97.5% on a plot to have an idea of the evolution of the confidence interval at 95% of its $x(0)$. This is very easy to get in a spreadsheet software like Microsoft® Excel® and we can typically falls by playing with several types of charts on the same type of plot as shown below for a portfolio of 500 million USD with 5% yield and standard deviation of 20%:

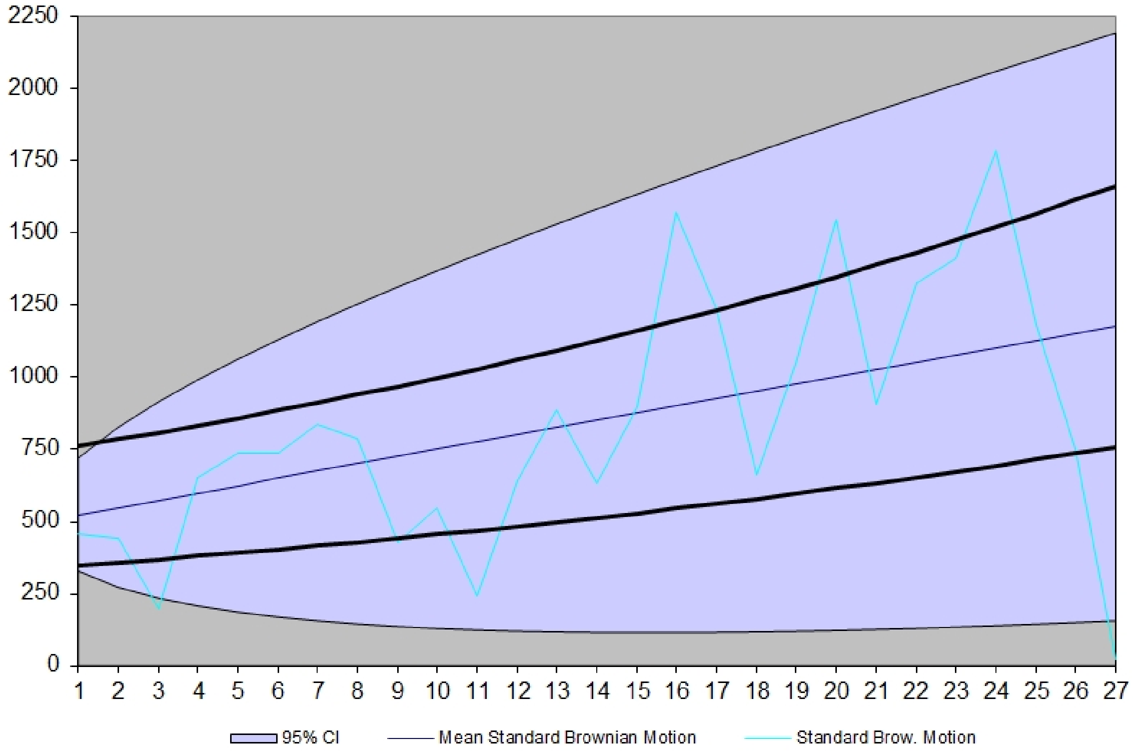
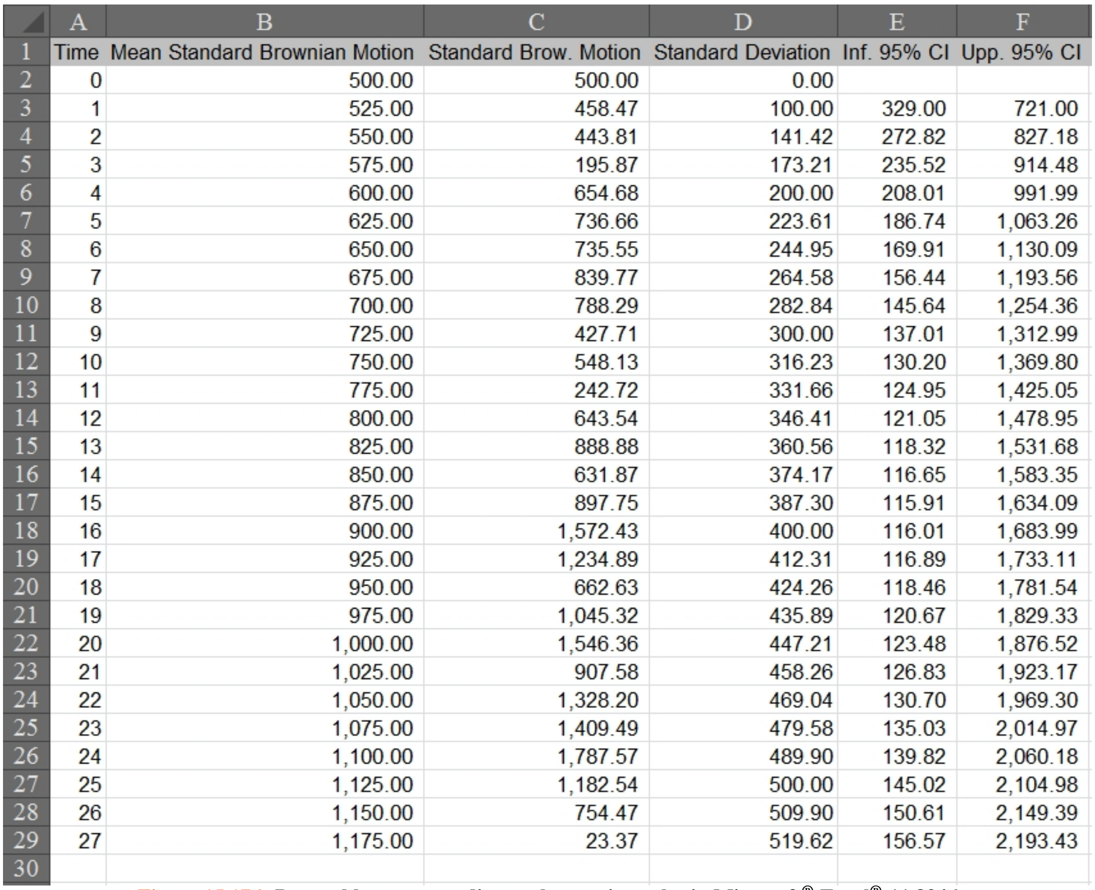


Figure 15.175: **Standard Brownian motion with confidence interval plot in Microsoft® Excel® 11.8346**

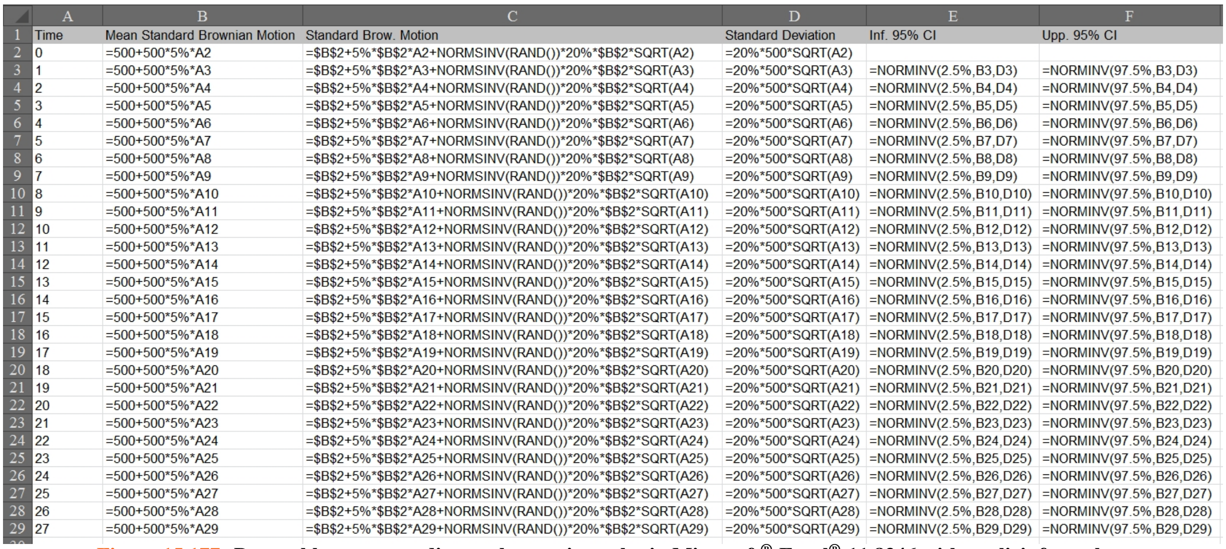
With the following table:



	A	B	C	D	E	F
1	Time	Mean Standard Brownian Motion	Standard Brow. Motion	Standard Deviation	Inf. 95% CI	Upp. 95% CI
2	0	500.00	500.00	0.00		
3	1	525.00	458.47	100.00	329.00	721.00
4	2	550.00	443.81	141.42	272.82	827.18
5	3	575.00	195.87	173.21	235.52	914.48
6	4	600.00	654.68	200.00	208.01	991.99
7	5	625.00	736.66	223.61	186.74	1,063.26
8	6	650.00	735.55	244.95	169.91	1,130.09
9	7	675.00	839.77	264.58	156.44	1,193.56
10	8	700.00	788.29	282.84	145.64	1,254.36
11	9	725.00	427.71	300.00	137.01	1,312.99
12	10	750.00	548.13	316.23	130.20	1,369.80
13	11	775.00	242.72	331.66	124.95	1,425.05
14	12	800.00	643.54	346.41	121.05	1,478.95
15	13	825.00	888.88	360.56	118.32	1,531.68
16	14	850.00	631.87	374.17	116.65	1,583.35
17	15	875.00	897.75	387.30	115.91	1,634.09
18	16	900.00	1,572.43	400.00	116.01	1,683.99
19	17	925.00	1,234.89	412.31	116.89	1,733.11
20	18	950.00	662.63	424.26	118.46	1,781.54
21	19	975.00	1,045.32	435.89	120.67	1,829.33
22	20	1,000.00	1,546.36	447.21	123.48	1,876.52
23	21	1,025.00	907.58	458.26	126.83	1,923.17
24	22	1,050.00	1,328.20	469.04	130.70	1,969.30
25	23	1,075.00	1,409.49	479.58	135.03	2,014.97
26	24	1,100.00	1,787.57	489.90	139.82	2,060.18
27	25	1,125.00	1,182.54	500.00	145.02	2,104.98
28	26	1,150.00	754.47	509.90	150.61	2,149.39
29	27	1,175.00	23.37	519.62	156.57	2,193.43
30						

Figure 15.176: **Data table corresponding to the previous plot in Microsoft® Excel® 11.8346**

Using the relations introduced above:



	A	B	C	D	E	F
1	Time	Mean Standard Brownian Motion	Standard Brow. Motion	Standard Deviation	Inf. 95% CI	Upp. 95% CI
2	0	=500+500*5%*A2	=$B$2+5%*$B$2*A2+NORMSINV(RAND())*20%*$B$2*SQRT(A2)	=20%*500*SQRT(A2)		
3	1	=500+500*5%*A3	=$B$2+5%*$B$2*A3+NORMSINV(RAND())*20%*$B$2*SQRT(A3)	=20%*500*SQRT(A3)	=NORMINV(2.5%,B3,D3)	=NORMINV(97.5%,B3,D3)
4	2	=500+500*5%*A4	=$B$2+5%*$B$2*A4+NORMSINV(RAND())*20%*$B$2*SQRT(A4)	=20%*500*SQRT(A4)	=NORMINV(2.5%,B4,D4)	=NORMINV(97.5%,B4,D4)
5	3	=500+500*5%*A5	=$B$2+5%*$B$2*A5+NORMSINV(RAND())*20%*$B$2*SQRT(A5)	=20%*500*SQRT(A5)	=NORMINV(2.5%,B5,D5)	=NORMINV(97.5%,B5,D5)
6	4	=500+500*5%*A6	=$B$2+5%*$B$2*A6+NORMSINV(RAND())*20%*$B$2*SQRT(A6)	=20%*500*SQRT(A6)	=NORMINV(2.5%,B6,D6)	=NORMINV(97.5%,B6,D6)
7	5	=500+500*5%*A7	=$B$2+5%*$B$2*A7+NORMSINV(RAND())*20%*$B$2*SQRT(A7)	=20%*500*SQRT(A7)	=NORMINV(2.5%,B7,D7)	=NORMINV(97.5%,B7,D7)
8	6	=500+500*5%*A8	=$B$2+5%*$B$2*A8+NORMSINV(RAND())*20%*$B$2*SQRT(A8)	=20%*500*SQRT(A8)	=NORMINV(2.5%,B8,D8)	=NORMINV(97.5%,B8,D8)
9	7	=500+500*5%*A9	=$B$2+5%*$B$2*A9+NORMSINV(RAND())*20%*$B$2*SQRT(A9)	=20%*500*SQRT(A9)	=NORMINV(2.5%,B9,D9)	=NORMINV(97.5%,B9,D9)
10	8	=500+500*5%*A10	=$B$2+5%*$B$2*A10+NORMSINV(RAND())*20%*$B$2*SQRT(A10)	=20%*500*SQRT(A10)	=NORMINV(2.5%,B10,D10)	=NORMINV(97.5%,B10,D10)
11	9	=500+500*5%*A11	=$B$2+5%*$B$2*A11+NORMSINV(RAND())*20%*$B$2*SQRT(A11)	=20%*500*SQRT(A11)	=NORMINV(2.5%,B11,D11)	=NORMINV(97.5%,B11,D11)
12	10	=500+500*5%*A12	=$B$2+5%*$B$2*A12+NORMSINV(RAND())*20%*$B$2*SQRT(A12)	=20%*500*SQRT(A12)	=NORMINV(2.5%,B12,D12)	=NORMINV(97.5%,B12,D12)
13	11	=500+500*5%*A13	=$B$2+5%*$B$2*A13+NORMSINV(RAND())*20%*$B$2*SQRT(A13)	=20%*500*SQRT(A13)	=NORMINV(2.5%,B13,D13)	=NORMINV(97.5%,B13,D13)
14	12	=500+500*5%*A14	=$B$2+5%*$B$2*A14+NORMSINV(RAND())*20%*$B$2*SQRT(A14)	=20%*500*SQRT(A14)	=NORMINV(2.5%,B14,D14)	=NORMINV(97.5%,B14,D14)
15	13	=500+500*5%*A15	=$B$2+5%*$B$2*A15+NORMSINV(RAND())*20%*$B$2*SQRT(A15)	=20%*500*SQRT(A15)	=NORMINV(2.5%,B15,D15)	=NORMINV(97.5%,B15,D15)
16	14	=500+500*5%*A16	=$B$2+5%*$B$2*A16+NORMSINV(RAND())*20%*$B$2*SQRT(A16)	=20%*500*SQRT(A16)	=NORMINV(2.5%,B16,D16)	=NORMINV(97.5%,B16,D16)
17	15	=500+500*5%*A17	=$B$2+5%*$B$2*A17+NORMSINV(RAND())*20%*$B$2*SQRT(A17)	=20%*500*SQRT(A17)	=NORMINV(2.5%,B17,D17)	=NORMINV(97.5%,B17,D17)
18	16	=500+500*5%*A18	=$B$2+5%*$B$2*A18+NORMSINV(RAND())*20%*$B$2*SQRT(A18)	=20%*500*SQRT(A18)	=NORMINV(2.5%,B18,D18)	=NORMINV(97.5%,B18,D18)
19	17	=500+500*5%*A19	=$B$2+5%*$B$2*A19+NORMSINV(RAND())*20%*$B$2*SQRT(A19)	=20%*500*SQRT(A19)	=NORMINV(2.5%,B19,D19)	=NORMINV(97.5%,B19,D19)
20	18	=500+500*5%*A20	=$B$2+5%*$B$2*A20+NORMSINV(RAND())*20%*$B$2*SQRT(A20)	=20%*500*SQRT(A20)	=NORMINV(2.5%,B20,D20)	=NORMINV(97.5%,B20,D20)
21	19	=500+500*5%*A21	=$B$2+5%*$B$2*A21+NORMSINV(RAND())*20%*$B$2*SQRT(A21)	=20%*500*SQRT(A21)	=NORMINV(2.5%,B21,D21)	=NORMINV(97.5%,B21,D21)
22	20	=500+500*5%*A22	=$B$2+5%*$B$2*A22+NORMSINV(RAND())*20%*$B$2*SQRT(A22)	=20%*500*SQRT(A22)	=NORMINV(2.5%,B22,D22)	=NORMINV(97.5%,B22,D22)
23	21	=500+500*5%*A23	=$B$2+5%*$B$2*A23+NORMSINV(RAND())*20%*$B$2*SQRT(A23)	=20%*500*SQRT(A23)	=NORMINV(2.5%,B23,D23)	=NORMINV(97.5%,B23,D23)
24	22	=500+500*5%*A24	=$B$2+5%*$B$2*A24+NORMSINV(RAND())*20%*$B$2*SQRT(A24)	=20%*500*SQRT(A24)	=NORMINV(2.5%,B24,D24)	=NORMINV(97.5%,B24,D24)
25	23	=500+500*5%*A25	=$B$2+5%*$B$2*A25+NORMSINV(RAND())*20%*$B$2*SQRT(A25)	=20%*500*SQRT(A25)	=NORMINV(2.5%,B25,D25)	=NORMINV(97.5%,B25,D25)
26	24	=500+500*5%*A26	=$B$2+5%*$B$2*A26+NORMSINV(RAND())*20%*$B$2*SQRT(A26)	=20%*500*SQRT(A26)	=NORMINV(2.5%,B26,D26)	=NORMINV(97.5%,B26,D26)
27	25	=500+500*5%*A27	=$B$2+5%*$B$2*A27+NORMSINV(RAND())*20%*$B$2*SQRT(A27)	=20%*500*SQRT(A27)	=NORMINV(2.5%,B27,D27)	=NORMINV(97.5%,B27,D27)
28	26	=500+500*5%*A28	=$B$2+5%*$B$2*A28+NORMSINV(RAND())*20%*$B$2*SQRT(A28)	=20%*500*SQRT(A28)	=NORMINV(2.5%,B28,D28)	=NORMINV(97.5%,B28,D28)
29	27	=500+500*5%*A29	=$B$2+5%*$B$2*A29+NORMSINV(RAND())*20%*$B$2*SQRT(A29)	=20%*500*SQRT(A29)	=NORMINV(2.5%,B29,D29)	=NORMINV(97.5%,B29,D29)

Figure 15.177: **Data table corresponding to the previous plot in Microsoft® Excel® 11.8346 with explicit formulas**

Obviously, in practice it is possible to make this type of graph with any data comprising a linear drift and whose standard deviation is known.

15.3.10.3.9.5 Brownian bridge

Let us now consider a case of application of the Brownian motion which will be useful to us later! By revising the developments seen above, it comes immediately that:

$$x_t = x_{t-\Delta t} + \mu x_0 \Delta t + \mathcal{N}(0,1)\sigma x_0 \sqrt{\Delta t} \tag{15.3.891}$$

Indeed, let us check this by choosing:

$$\Delta t = 1 \tag{15.3.892}$$

We have:

$$x_t = x_{t-1} + \mu \cdot x_0 \cdot 1 + \mathcal{N}(0,1)\sigma x_0 \sqrt{1} = x_{t-1} + \mu x_0 + \mathcal{N}(0,1)\sigma x_0 \tag{15.3.893}$$

And therefore for $t = 1$:

$$x_1 = x_0 + \mu x_0 + \mathcal{N}(0,1)\sigma x_0 \tag{15.3.894}$$

And similarly for $t = 2$:

$$\begin{aligned} x_2 &= x_1 + \mu \cdot x_0 + \mathcal{N}(0,1)\sigma x_0 = \big(x_0 + \mu x_0 + \mathcal{N}(0,1)\sigma x_0\big) + \mu \cdot x_0 + \mathcal{N}(0,1)\sigma x_0 \\ &= x_0 + 2\mu \cdot x_0 + \underbrace{\big(\mathcal{N}(0,1) + \mathcal{N}(0,1)\big)}_{=\mathcal{N}(0,\sqrt{2})} \sigma x_0 = x_0 + 2\mu \cdot x_0 + \mathcal{N}(0,1)\sqrt{2\sigma x_0} \end{aligned} \tag{15.3.895}$$

This is indeed equivalent to:

$$x(t) = x(0) + \mu x(0)\Delta t + \mathcal{N}(0,1)\sigma x(0)\sqrt{\Delta t} \tag{15.3.896}$$

For $\Delta t = 2$:

$$W(t) = W_{t-\Delta t} + \mu W_0 \Delta t + \mathcal{N}(0,1)\sigma W_0 \sqrt{\Delta t} \tag{15.3.897}$$

Well, this being said, we would like to build a Brownian motion which starts at the time $t = 0$ from zero (which imposes $x(0) = 0$) fluctuates during a given period T and returns to zero as shown in the figure below:

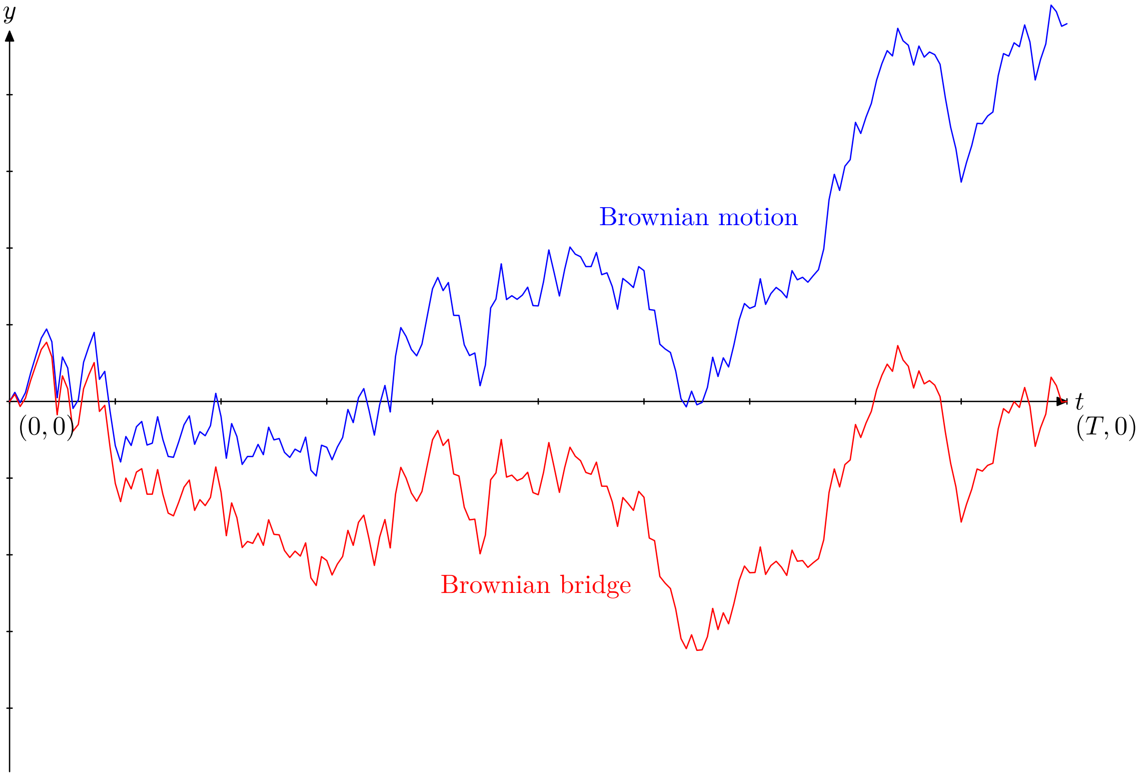


Figure 15.178: Brownian motion and Brownian bridge

As it is usual, let us choose the following notation:

$$W(t)=W_{t-\Delta t}+\mu W_0\Delta t+\mathcal{N}(0,1)W_0\sqrt{\Delta t} \tag{15.3.898}$$

If we want the Brownian motion to vanish at time T, it is sufficient to make a simple subtraction of the following type:

$$\text{BB}(t)=W(t)-W(T) \tag{15.3.899}$$

where BB means "Brownian bridge". If we stop here, we have indeed:

$$\text{BB}(t)=W(T)-W(T)=0 \tag{15.3.900}$$

But we have a problem with time $t=0$:

$$\text{BB}(0)=W(0)-W(T)\neq 0 \tag{15.3.901}$$

and to make sure that the difference is zero, the natural trick is to write:

$$\text{BB}(t)=W(t)-tW(T) \tag{15.3.902}$$

Thus, we have:

$$\text{BB}(0)=W(0)-0\cdot W(T)=0 \tag{15.3.903}$$

But then comes another problem. Indeed:

$$\text{BB}(T)=W(T)-T\cdot W(T)\neq 0 \tag{15.3.904}$$

So the final solution is to write:

$$\boxed{\mathrm{BB}(t)=W(t)-\frac{t}{T}\cdot W(T)} \quad (15.3.905)$$

and therefore:

$$\mathrm{BB}(t)=\mathrm{BB}(T)=0 \quad (15.3.906)$$

and we have:

$$\mathrm{E}(\mathrm{BB}(t))=\mathrm{E}\left(W(t)-\frac{t}{T}\cdot W(T)\right)=0 \quad (15.3.907)$$

The Brownian Bridge:

$$\mathrm{BB}(t)=W(t)-\frac{t}{T}\cdot W(T) \quad (15.3.908)$$

works for all the Wiener's processes, reason why we will sometimes see in the literature sometimes bridges with:

$$W_t=W_{t-\Delta t}+\mu W_0\Delta t+\mathcal{N}(0,1)\sigma W_0\sqrt{\Delta t} \quad (15.3.909)$$

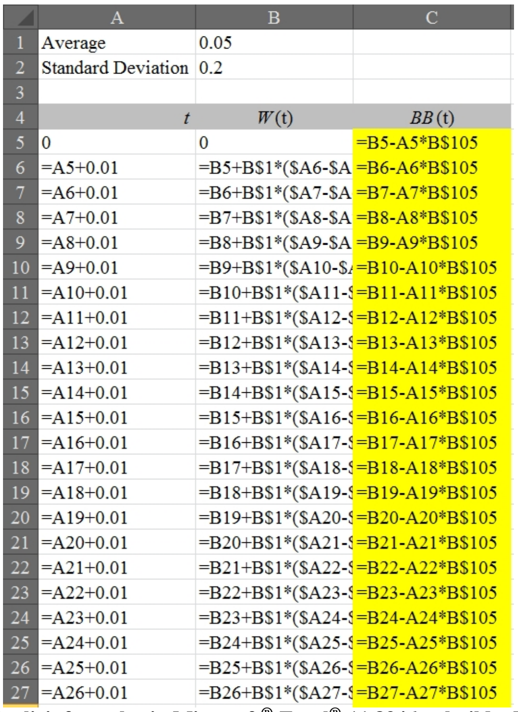
or for example with:

$$W_t=W_{t-\Delta t}+\mu\Delta t+\mathcal{N}(0,1)\sigma\sqrt{\Delta t} \quad (15.3.910)$$

which is only one special case of the preceding with $W_0=1$.

✍**Example:**

To simulate such a process in Microsoft® Excel® 11.8346 for example we build:



	A	B	C
1	Average	0.05	
2	Standard Deviation	0.2	
3			
4	*t*	*W*(t)	*BB*(t)
5	0	0	=B5-A5*B$105
6	=A5+0.01	=B5+B$1*($A6-$A	=B6-A6*B$105
7	=A6+0.01	=B6+B$1*($A7-$A	=B7-A7*B$105
8	=A7+0.01	=B7+B$1*($A8-$A	=B8-A8*B$105
9	=A8+0.01	=B8+B$1*($A9-$A	=B9-A9*B$105
10	=A9+0.01	=B9+B$1*($A10-$	=B10-A10*B$105
11	=A10+0.01	=B10+B$1*($A11-$	=B11-A11*B$105
12	=A11+0.01	=B11+B$1*($A12-$	=B12-A12*B$105
13	=A12+0.01	=B12+B$1*($A13-$	=B13-A13*B$105
14	=A13+0.01	=B13+B$1*($A14-$	=B14-A14*B$105
15	=A14+0.01	=B14+B$1*($A15-$	=B15-A15*B$105
16	=A15+0.01	=B15+B$1*($A16-$	=B16-A16*B$105
17	=A16+0.01	=B16+B$1*($A17-$	=B17-A17*B$105
18	=A17+0.01	=B17+B$1*($A18-$	=B18-A18*B$105
19	=A18+0.01	=B18+B$1*($A19-$	=B19-A19*B$105
20	=A19+0.01	=B19+B$1*($A20-$	=B20-A20*B$105
21	=A20+0.01	=B20+B$1*($A21-$	=B21-A21*B$105
22	=A21+0.01	=B21+B$1*($A22-$	=B22-A22*B$105
23	=A22+0.01	=B22+B$1*($A23-$	=B23-A23*B$105
24	=A23+0.01	=B23+B$1*($A24-$	=B24-A24*B$105
25	=A24+0.01	=B24+B$1*($A25-$	=B25-A25*B$105
26	=A25+0.01	=B25+B$1*($A26-$	=B26-A26*B$105
27	=A26+0.01	=B26+B$1*($A27-$	=B27-A27*B$105

Figure 15.179: **Explicit formulas in Microsoft® Excel® 11.8346 to build a Brownian bridge**

as the time interval is $\Delta t = 0.01$ then $T = 1/0.01 = 100$ and we get visually:

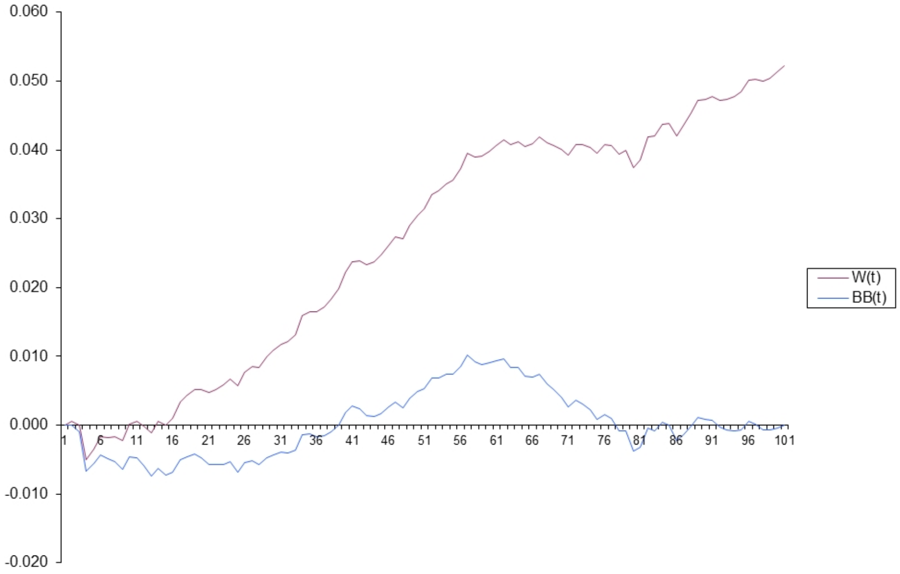


Figure 15.180: **Brownian bridge in Microsoft® Excel® 11.8346**

With everything seen so far, the reader should be able to understand the following plot showing a typical surface Call price curves with the special case of random Brownian bridge between the emission and maturity of the option:

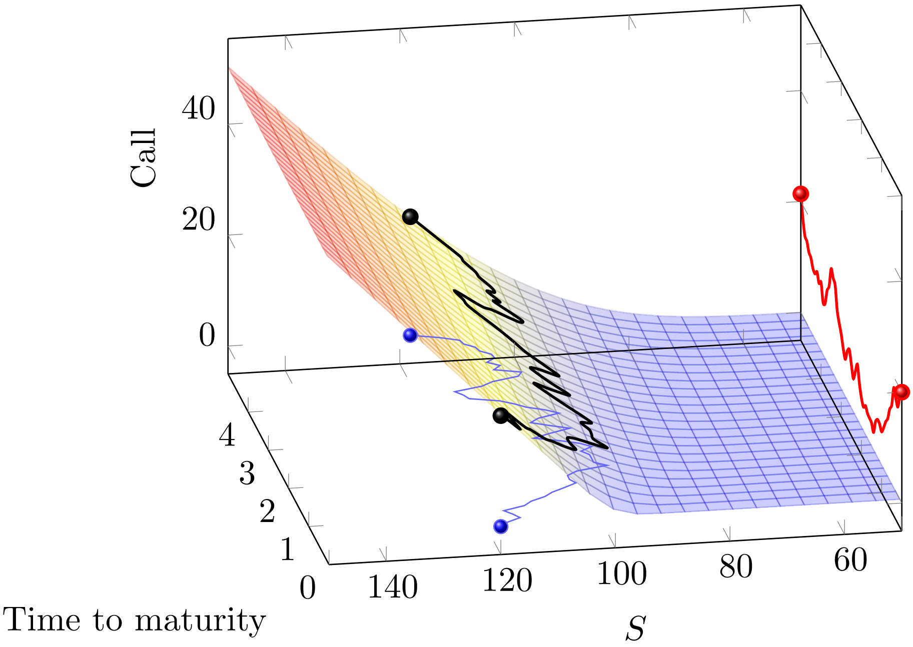


Figure 15.181: Surface Call price curves with a random Brownian bridge

15.3.10.3.9.6 Itō's process

Let us consider now a Brownian process corresponding to a variation of x in continuous time (practitioners speak also sometimes of "short-rate modelling") defined by:

$$dx = a(x,t)dt + b(x,t)dz \qquad (15.3.911)$$

a and b then being functions of the two variables x and t. This consideration is what we name an "Itō process", which is named after Kiyosi Itō. It is therefore a generalization of the previous case where a and b are no longer constant.

It is possible to calculate the expected mean and the variance of dx exactly in the same way as for the Wiener process and we obtain very easily by analogy:

$$\text{E}(dx) = 0 \qquad \text{E}(dx) = b^2(x,t)dt \qquad (15.3.912)$$

Therefore, we can write:

$$dx \longrightarrow \mathcal{N}\left(a(x,t)dt, b(x,t)\sqrt{dt}\right) \qquad (15.3.913)$$

Where $a(x,t)$ corresponds to the "instantaneous drift" and $b(x,t)$ to the "instantaneous variance".

Remark

It seems that the general case where a and b depend on time and of a parameter x (which, as we shall see later, corresponds to the strike price) is part of a model class which is named "Heat-Jarrow-Mortong models". In the case where only b would no longer be dependent on time, it seems that than it is named "Ho & Lee model". In reality there is a plethora of empirical models that adapt more or less well according to the situations and to study the whole, a life is no longer enough as for each model you can find a book between 200 and 400 pages...

The "geometric Brownian motion" which allows to theoretically define a possible prediction of the evolution of the yield of an asset is a special case of Itō processes (among so many other models ...) where we assume that:

$$a(x,t)=\mu\cdot x \qquad \text{and} \qquad b(x,t)=\sigma\cdot x \tag{15.3.914}$$

Remember that μ and σ are in percentage!

So the variance and the drift is independent of time! The advantage of assuming this is that unlike the Brownian arithmetic motion we will see that the yield then follows a log-normal law and therefore can not be negative. However, the two processes have two common weaknesses: the price of the associated asset can tend towards infinity in an infinite time (which is corrected the Vasicek process) and both consider volatility as constant in time...

Then we can write the expression of the geometric Brownian motion of the value of the asset as following:

$$\boxed{\mathrm{d}F=\mu x\mathrm{d}t+\sigma x\mathrm{d}z} \tag{15.3.915}$$

often represented in the literature also in the following form:

$$\mathrm{d}X_t=\mu X_t\mathrm{d}t+\sigma X_t\mathrm{d}W_t \tag{15.3.916}$$

Or even more explicitly:

$$\mathrm{d}X_t=(\mu X_t)\mathrm{d}t+(\sigma X_t)\mathrm{d}W_t=(\mu X_t)\mathrm{d}t+(\sigma X_t)\mathrm{d}\left(\mathcal{N}(0,1)\sqrt{\mathrm{d}t}\right) \tag{15.3.917}$$

Remark

The "Chan-Karolyi-Longstaff-Sanders model" is an attempt to generalize many stochastic process models used in finance:

$$\mathrm{d}X_t=(\alpha+\mu X_t)\mathrm{d}t+\sigma X_t^{\gamma}\mathrm{d}W_t \tag{15.3.918}$$

Here are some of the best known cases used in finance (the years are given based on holocene calendar):

Reference	Model
Merton (11973) / Bachelier / Brownian motion	$\mathrm{d}X_t = \mu \mathrm{d}t + \sigma \mathrm{d}W_t$
Geometric brownian motion	$\mathrm{d}X_t = \mu X_t \mathrm{d}t + \sigma X_t \mathrm{d}W_t$
Ornstein-Uhlenbeck (11930)	$\mathrm{d}X_t = -\theta X_t \mathrm{d}t + \sigma \mathrm{d}W_t$
Vasicek / Mean reverting (11977)	$\mathrm{d}X_t = k(\alpha + \mu X_t)\mathrm{d}t + \sigma \mathrm{d}W_t$
Cox-Ingersoll-Ross (11985)	$\mathrm{d}X_t = (\alpha - \mu X_t)\mathrm{d}t + \sigma \sqrt{X_t}\mathrm{d}W_t$
Dothan (11978)	$\mathrm{d}X_t = \sigma \mathrm{d}W_t$
Rendlemann-Bartter Geometric brownian motion (11905)	$\mathrm{d}X_t = \mu X_t \mathrm{d}t + \sigma X_t \mathrm{d}W_t$
Schwartz and Brennan (11980)	$\mathrm{d}X_t = (\alpha + \mu X_t)\mathrm{d}t + \sigma X_t \mathrm{d}W_t$
Cox and al. (11980)	$\mathrm{d}X_t = \sigma X_t^{3/2} \mathrm{d}W_t$
Variance with constant elasticity (CEV)	$\mathrm{d}X_t = \mu X_t \mathrm{d}t + \sigma X_t^{\gamma} \mathrm{d}W_t$
Chan-Karolyi-Longstaff-Sander (11992)	$\mathrm{d}X_t = (\alpha + \mu X_t)\mathrm{d}t + \sigma X_t^{\gamma} \mathrm{d}W_t$
Heston (11993)	$\mathrm{d}X_t = \mu X_t \mathrm{d}t + \sigma_t X_t \mathrm{d}W_t^X$ $\mathrm{d}\sigma_t = \kappa(\theta - \sigma_t)\mathrm{d}t + \xi \sqrt{\sigma_t}\mathrm{d}W_t^{\sigma}$
...	...

Table 15.53: **Some Chan-Karolyi-Longstaff-Sanders derived stochastic processes**

where the reader must remember that μ and σ are in percentage!

The k factor of the Vasiceck model is named the "mean-reversion speed" (typically used when the trader believe that in normal conditions, the average price of a stock, given a certain number of prior periods will lie between the bid price and the offer price and due to some heavy market forces it should revert back to the long term mean price level). All diffusion process above with a power seems to be commonly named "fractional diffusion processes".

Considering an asset returns and its variance as constants in time are choices based on the empirical observation of markets when assets are stocks. However, some more sophisticated models have asset returns and variances that depend on the asset value in time.

The financial interpretation of the relation:

$$\mathrm{d}F = \mu x \mathrm{d}t + \sigma x \mathrm{d}z \tag{15.3.919}$$

Becomes apparent when we divide the two members by x:

$$\boxed{\frac{\mathrm{d}S_t}{S_t} = \mu \mathrm{d}t + \sigma \mathrm{d}z} \tag{15.3.920}$$

Which corresponds to the rate of return of the asset over an infinitesimal period $\mathrm{d}t$.

The geometric Brownian motion is therefore a priori a good candidate to model the evolution of the price of certain financial assets from its rate of return.

In the specialized literature, the return is also sometimes denoted (justified notation) in the form of the following stochastic differential equation (S.D.E.):

$$\frac{\mathrm{d}S_t}{S_t} = (\mu - q)\mathrm{d}t + \sigma \mathrm{d}z \tag{15.3.921}$$

where S_t is obviously the price of the underlying named as we already know the "stock price" at time t, μ is named the "drift" (often assimilated to yield) and σ the "yield volatility". It is the notation and the vocabulary that we will adopt for the rest of the development. Let us notice that we see better than this assumes that the drift is independent of the value of the underlying (which is a verified observation for some assets).

In the case of a continuous dividend payment, given q the dividend rate, the dynamics of the price can be written intuitively:

$$\frac{\mathrm{d}\ln(S_t)}{\mathrm{d}S_t}=\frac{1}{S_t} \tag{15.3.922}$$

We can therefore also write (we continue with the case without dividends!):

$$\mathrm{d}\ln(S_t)=\mu\mathrm{d}t+\sigma\mathrm{d}z \tag{15.3.923}$$

In the case where $\sigma=0$ (Wiener process, i.e. the price of the underlying is perfectly known at a given time and without risks), we find ourselves with a differential equation (well known in the finance domain) that we can solve immediately:

$$\frac{\mathrm{d}\ln(S_t)}{\mathrm{d}S_t}=\mu\mathrm{d}t\Leftrightarrow\frac{\mathrm{d}S_t}{\mathrm{d}S_t}=\mu S_t\Rightarrow S_t=S_0e^{\mu(t-t_0)} \tag{15.3.924}$$

This is therefore an exponential (as the continuous interest we saw at the beginning of this section). This relation is valid only if the time interval is therefore very small and the return rate not too big!

We will now see thanks to the "Itō lemma" that it is possible (which is not the unique possibility!) to establish that such a process can define a log-normal law (see section Statistics page 734).

The Itō lemma is derived from the multivariable Taylor expansion (see section Sequences and Series page 1696) with 2 variables x and t and given by:

$$\Delta F=\frac{1}{1!}\frac{\partial F}{\partial t}\Delta t+\frac{1}{1!}\frac{\partial F}{\partial x}\Delta x+\frac{1}{2!}\frac{\partial^2 F}{\partial t^2}\Delta t^2+\frac{1}{2!}\frac{\partial^2 F}{\partial x^2}\Delta x^2+\ldots \tag{15.3.925}$$

with $F(0)=0$ at the origin of the Brownian motion.

Considering $\Delta t\ll\Delta x$, and taking the terms only up to the second order (formal perilous approximation but numerically not mandatory thanks to computing power of modern computers), we have:

$$\Delta F\cong\frac{1}{1!}\frac{\partial F}{\partial t}\Delta t+\frac{1}{1!}\frac{\partial F}{\partial x}\Delta x+\frac{1}{2!}\frac{\partial^2 F}{\partial x^2}\Delta x^2 \tag{15.3.926}$$

Let us now return to:

$$\Delta x=a\cdot\Delta t+b\cdot\Delta z=a\Delta t+b\varepsilon\sqrt{\Delta t} \tag{15.3.927}$$

Let us raise to the square (i.e. compute the quadratic variation), we get:

$$\Delta x^2=\left(a\Delta t+b\varepsilon\sqrt{\Delta t}\right)^2=a^2\Delta t^2+2\varepsilon ab\Delta t^{3/2}+b^2\varepsilon^2\Delta t\cong b^2\varepsilon^2\Delta t \tag{15.3.928}$$

But:

$$\mathrm{E}\left(b^2\varepsilon^2\Delta t\right)=b^2\Delta t\mathrm{E}\left(\varepsilon^2\right) \tag{15.3.929}$$

And as we have proved it in the Statistics section:

$$\mathrm{V}(X)=\mathrm{E}\left(X^2\right)-\mathrm{E}\left(X\right)^2\Leftrightarrow \mathrm{E}(X^2)=\mathrm{V}(X)+\mathrm{E}(X)^2 \tag{15.3.930}$$

Then we have:

$$\mathrm{E}\left(\varepsilon^2\right)=\mathrm{V}(\varepsilon)+\mathrm{E}(\varepsilon)^2=1+0=1 \tag{15.3.931}$$

Therefore:

$$\mathrm{E}(b^2\varepsilon^2\Delta t)=b^2\Delta t \tag{15.3.932}$$

By the way:

$$\mathrm{V}(b^2\varepsilon^2\Delta t)=b^4\Delta t^2\mathrm{V}(\varepsilon^2)=b^4\Delta t^2 \tag{15.3.933}$$

which both tend to 0 when Δt tends to 0.

We also have by considering a subdivision of time in extremely small intervals $\mathrm{d}t$ (approximation which will be useful to us a little later):

$$\lim_{\Delta t\rightarrow 0} b^2\varepsilon^2\Delta t\cong \lim_{\Delta t\rightarrow 0} b^2\Delta t=0 \tag{15.3.934}$$

which implies in that latter case:

$$b^2\varepsilon^2\Delta t\cong b^2\Delta t \tag{15.3.935}$$

Thus taking place in continuous time (thus a continuous model with extremely small time intervals), the application of the discrete Taylor development seen above:

$$\Delta F\cong\frac{1}{1!}\frac{\partial F}{\partial t}\Delta t+\frac{1}{1!}\frac{\partial F}{\partial x}\Delta x+\frac{1}{2!}\frac{\partial^2 F}{\partial x^2}\Delta x^2 \tag{15.3.936}$$

Can then be written:

$$\begin{aligned}\mathrm{d}F&\cong\frac{\partial F}{\partial t}\mathrm{d}t+\frac{\partial F}{\partial x}\mathrm{d}x+\frac{1}{2}\frac{\partial^2 F}{\partial x^2}b^2\mathrm{d}t=\frac{\partial F}{\partial t}\mathrm{d}t+\frac{\partial F}{\partial x}(a\mathrm{d}t+b\mathrm{d}z)+\frac{1}{2}\frac{\partial^2 F}{\partial x^2}b^2\mathrm{d}t\\&=\left(\frac{\partial F}{\partial t}+a\frac{\partial F}{\partial x}+\frac{1}{2}b^2\frac{\partial^2 F}{\partial x^2}\right)\mathrm{d}t+b\frac{\partial F}{\partial x}\mathrm{d}z\end{aligned} \tag{15.3.937}$$

Therefore:

$$\boxed{\mathrm{d}F\cong\left(\frac{\partial F}{\partial t}+a\frac{\partial F}{\partial x}+\frac{1}{2}b^2\frac{\partial^2 F}{\partial x^2}\right)\mathrm{d}t+b\frac{\partial F}{\partial x}\mathrm{d}z} \tag{15.3.938}$$

This final result is the "Itō's lemma" also named "Itō-Doeblin theorem".

Therefore as the reader can read it on Wikipedia: Itō's lemma is an identity used in Itō calculus to find the differential of a time-dependent function of a stochastic process. It serves as the stochastic calculus counterpart of the chain rule. It can be heuristically derived by forming the Taylor series expansion of the function up to its second derivatives and retaining terms up to first order in the time increment and second order in the Wiener process increment. The lemma is widely employed in mathematical finance, and its best known application is in the derivation of the Black–Scholes equation for option values.

Remarks

R1. Compare the form of the last equality with the relation: $\mathrm{d}x = a(x,t)\mathrm{d}t + b(x,t)\mathrm{d}z$

R2. The expression in parentheses above (factor of $\mathrm{d}t$) is sometimes referred to as the "Kolmogorov backward equation".

If we take:

$$F(x) = \ln(x) \tag{15.3.939}$$

Therefore:

$$\frac{\partial F}{\partial t} = 0 \qquad \frac{\partial F}{\partial x} = \frac{1}{x} \qquad \frac{\partial^2 F}{\partial x^2} = -\frac{1}{x^2} \tag{15.3.940}$$

In that case:

$$\mathrm{d}F \cong \left(0 + \frac{a}{x} - \frac{1}{2}b^2\frac{1}{x^2}\right)\mathrm{d}t + b\frac{1}{x}\mathrm{d}z \tag{15.3.941}$$

Returning to the geometric Brownian motion hypothesis, we know that we must consider that:

$$a = a(x,t) = \mu x \qquad \text{at} \qquad b = b(x,t) = \sigma x \tag{15.3.942}$$

Therefore we have:

$$\mathrm{d}F \cong \left(0 + \frac{\mu x}{x} - \frac{1}{2}\frac{\sigma^2 x^2}{x^2}\right)\mathrm{d}t + \sigma x\frac{1}{x}\mathrm{d}z \tag{15.3.943}$$

and we finally get the following stochastic differential equation with constant coefficients:

$$\mathrm{d}F \cong \left(\mu - \frac{\sigma^2}{2}\right)\mathrm{d}t + \sigma\mathrm{d}z \tag{15.3.944}$$

Either by taking the notation of the beginning in explicit form:

$$\mathrm{d}\ln(x) \cong \left(\mu - \frac{\sigma^2}{2}\right)\mathrm{d}t + \sigma\mathrm{d}z \tag{15.3.945}$$

Or in another even more explicit form:

$$\boxed{\mathrm{d}\ln(X_t) \cong \left(\mu - \frac{\sigma^2}{2}\right)\mathrm{d}t + \sigma\mathrm{d}W_t = \left(\mu - \frac{\sigma^2}{2}\right)\mathrm{d}t + \sigma\mathrm{d}\left(\mathcal{N}(0,1)\sqrt{\mathrm{d}t}\right)} \tag{15.3.946}$$

Sometimes named the "integral equation of price dynamics".

A trivial approach to the resolution of this integral is (we always consider in this model that volatility is independent of time... and abusively we replace the $\cong$ by a $=$ sign):

$$\int_t^{t+1} \mathrm{d}\ln(X_t) = \int_t^{t+1} \left(\mu - \frac{\sigma^2}{2}\right)\mathrm{d}t + \int_t^{t+1} \sigma \mathrm{d}W_t = \left(\mu - \frac{\sigma^2}{2}\right)\int_t^{t+1}\mathrm{d}t + \sigma\int_t^{t+1}\mathrm{d}W_t \tag{15.3.947}$$

Therefore:

$$\ln(X_{t+1}) - \ln(X_t) = \left(\mu - \frac{\sigma^2}{2}\right)\Big((t+1)-t\Big) + \sigma\Big(W_{t+1} - W_t\Big) \tag{15.3.948}$$

Hence:

$$\ln\left(\frac{X_{t+1}}{X_t}\right) = \mu - \frac{\sigma^2}{2} + \sigma\Big(W_{t+1} - W_t\Big) = \mu - \frac{\sigma^2}{2} + \sigma\mathcal{N}(0,1) \tag{15.3.949}$$

Remarks

R1. Remember that we started from the relation $\mathrm{d}F = \mu x \mathrm{d}t + \sigma x \mathrm{d}z$

R2. Brownian movements were successively discovered from the hypothesis of Normality in the 11960s (holocene calendar) and then from the stability hypothesis in the 11980s (holocene calendar). With these two hypotheses, mathematicians rank them in the special category of "2-stable Levy processes".

$\mathrm{d}F$ defines a geometrical Brownian motion with particular drift whose parameters that we can now measure (this is what we wanted to achieve). Therefore, the results we obtained for the Brownian motion can be recovered and allow us to write in the end:

$$\mathrm{d}F \longrightarrow \mathcal{N}\left(\left(\mu - \frac{\sigma^2}{2}\right)\mathrm{d}t, \sigma\sqrt{\mathrm{d}t}\right) \tag{15.3.950}$$

Which means that $\mathrm{d}F$ follows a log-normal law (see section Statistics page 734) of parameters:

$$\left(\mu - \frac{\sigma^2}{2}\right)\mathrm{d}t \qquad \text{and} \qquad \sigma\sqrt{\mathrm{d}t} \tag{15.3.951}$$

Or otherwise written:

$$\begin{aligned} \Delta F = \Delta\Big(\ln(X_t)\Big) &\to \mathcal{N}\left(\left(\mu - \frac{\sigma^2}{2}\right)\Delta t, \sigma\sqrt{\Delta t}\right) \\ \Rightarrow \ln(X_T) - \ln(X_0) &\to \mathcal{N}\left(\left(\mu - \frac{\sigma^2}{2}\right)T, \sigma\sqrt{T}\right) \\ \Rightarrow \ln(X_T) &\to \mathcal{N}\left(\ln(X_0) + \left(\mu - \frac{\sigma^2}{2}\right)T, \sigma\sqrt{T}\right) \end{aligned} \tag{15.3.952}$$

A small reminder with what was seen in the Statistics section but applied to finance of the log-normal law finance may be welcome! Consider the present value of a given asset (compound interest) whose rate varies for each period:

$$X_t = X_0 \prod_{i=1}^{N} \left(1 + t_i\%\right) \tag{15.3.953}$$

So if we take the logarithm:

$$\ln(X_t) = \ln(X_0) + \sum_{i=1}^{N} \ln(1 + t_i\%) \tag{15.3.954}$$

Thus, if the rate of return is a variable random and independent from period to period and identically distributed, then we have an expression which will give a Normal distribution according to the central limit theorem (see section Statistics page 938). Therefore, a random variable is log-normalized when its logarithm is Normally distributed!

✍**Examples:**

E1. Let us consider a stock with a current price of 40.-, with an assumed average geometric yield of 16% per year and an annual volatility of 20%. In 6 months ($T = 0.5$) we then have:

$$\ln(X_T) \longrightarrow \mathcal{N}(3.759, 0.141) \tag{15.3.955}$$

We can then build a 95% confidence interval (see section Statistics page 848):

$$3.759 - 1.96 \cdot 0.141 \leq \ln(X_t) \leq 3.759 + 1.96 \cdot 0.141 \tag{15.3.956}$$

And therefore:

$$\begin{gathered} e^{3.759 - 1.96 \cdot 0.141} \leq X_t \leq e^{3.759 + 1.96 \cdot 0.141} \\ 32.55 \leq X_t \leq 56.56 \end{gathered} \tag{15.3.957}$$

So there is 95% cumulative probability that in 6 months the value of the security is between 32.55.- and 56.56.-.

E2. Knowing that:

$$\ln\left(\frac{X_T}{X_0}\right) \rightarrow \mathcal{N}\left(\left(\mu - \frac{\sigma^2}{2}\right)T, \sigma\sqrt{T}\right) \tag{15.3.958}$$

and that at continuous rate:

$$X_T = X_0 e^{RT} \Leftrightarrow R = \frac{1}{T} \ln\left(\frac{X_t}{X_0}\right) \tag{15.3.959}$$

We then have for performance the possibility of applying the same type of calculations as before:

$$R \longrightarrow \mathcal{N}\left(\mu - \frac{\sigma^2}{2}, \frac{\sigma}{\sqrt{T}}\right) \tag{15.3.960}$$

Let's go a step further by integrating the differential element. So we have:

$$\frac{\mathrm{d}X_t}{X_t}=\left(\mu-\frac{\sigma^2}{2}\right)\mathrm{d}t+\sigma\mathrm{d}\left(\mathcal{N}(0,1)\sqrt{\mathrm{d}t}\right) \tag{15.3.961}$$

Let us integrate this last relation:

$$\int\frac{\mathrm{d}X_t}{X_t}=\int\left(\mu-\frac{\sigma^2}{2}\right)\mathrm{d}t+\int\sigma\mathrm{d}\left(\mathcal{N}(0,1)\sqrt{\mathrm{d}t}\right) \tag{15.3.962}$$

The first primitive is simple:

$$\int\frac{\mathrm{d}X_t}{X_t}=\ln(X_t)+c^{te} \tag{15.3.963}$$

The second primitive is also quite simple (no integration constant because at time zero the expectation of gain is zero):

$$\int\left(\mu-\frac{\sigma^2}{2}\right)\mathrm{d}t=\left(\mu-\frac{\sigma^2}{2}\right)\int\mathrm{d}t=\left(\mu-\frac{\sigma^2}{2}\right)t \tag{15.3.964}$$

The third primitive is equal to (no integration constant because at time zero the value of the gain is perfectly known as being equal to zero):

$$\int\sigma\mathrm{d}\left(\mathcal{N}(0,1)\sqrt{\mathrm{d}t}\right)=\sigma\mathcal{N}(0,1)\int\mathrm{d}\sqrt{\mathrm{d}t}=\sigma\mathcal{N}(0,1)\sqrt{t} \tag{15.3.965}$$

Caution to the abuse of notation here!!! In the root it is implicitly a $\mathrm{d}t$ (therefore the analysis time step!) and not simply a t. Remember that for the remaining developments below!!!

We therefore have:

$$\ln(X_t)=\left(\mu-\frac{\sigma^2}{2}\right)t+\sigma\mathcal{N}(0,1)\sqrt{t}-c^{te} \tag{15.3.966}$$

And finally (we put the "$-$" sign in the constant):

$$X_t=e^{\left(\mu-\frac{\sigma^2}{2}\right)t+\sigma\mathcal{N}(0,1)\sqrt{t}+c^{te}}=e^{c^{te}}e^{\left(\mu-\frac{\sigma^2}{2}\right)t+\sigma\mathcal{N}(0,1)\sqrt{t}} \tag{15.3.967}$$

To find the meaning of the first factor, it suffices to state the initial condition:

$$X_t(t=0)=X_0 \tag{15.3.968}$$

We then have immediately for the final expression of the geometric Brownian which is for recall a random variable:

$$\boxed{X_t=X_0e^{\left(\mu-\frac{\sigma^2}{2}\right)t+\sigma\mathcal{N}(0,1)\sqrt{t}}} \tag{15.3.969}$$

that was obtained by Paul Samuelson in 11965 (holocene calendar) and which is sometimes named the "Bachelier-Samuelson model" or "log-normal representation of the value of the asset" or "Itō's solution".

This last relation is often written in the following form:

$$X_t = X_0 e^{\left(\mu-\frac{\sigma^2}{2}\right)\Delta t+\sigma\mathcal{N}(0,1)\sqrt{\Delta t}} = X_0 e^{\left(\mu-\frac{\sigma^2}{2}\right)(T-t)+\sigma\mathcal{N}(0,1)\sqrt{T-t}} \tag{15.3.970}$$

Notice that if we write (due to the Jensen's inequality proven in the Statistics section page 576):

$$S_0\mathrm{E}\left(\max\left(e^{\left(\mu-\frac{\sigma^2}{2}\right)(T-t)+\sigma\mathcal{N}(0,1)\sqrt{T-t}}-K,0\right)\right) \tag{15.3.971}$$

this gives us the expectation mean of positive deviations from the strike price. And if we update this value in the idea that this sum of money can be invested in a risk-free return asset, we then get the Monte Carlo version of the evaluation of a Call type option:

$$C = S_0 e^{-r(T-t)}\mathrm{E}\left(\max\left(e^{\left(\mu-\frac{\sigma^2}{2}\right)(T-t)+\sigma\mathcal{N}(0,1)\sqrt{T-t}}-K,0\right)\right) \tag{15.3.972}$$

And in general the same form of relation written for a Put will be named the "fundamental relation of option pricing". In reality we must take care to the fact that here the word "fundamental" means that we must be careful, however, to choose the right stochastic process, which is why in the literature we have more often the following expression that generalizes the previous relationship:

$$\boxed{C,P = e^{-r(T-t)}\mathrm{E}\big(f(M_s)\big) = D(T,t,r)E\big(f(M_s)\big)} \tag{15.3.973}$$

where f is a function of a given stochastic process (since there are several functions depending on the type of underlying), and the function D is named a "discount factor". For this last relation, we find almost as many different writings as there are textbooks on the subject...

The above relation is also sometimes denoted:

$$e^{-rT}\tilde{\mathrm{E}}\big(f(S(T))\big) \tag{15.3.974}$$

where $\tilde{\mathrm{E}}$ means: *expectation under the risk-neutral stock price model*. What do we mean by this? By the risk-neutral stock price model, we mean a geometric Brownian motion with the same yield volatility σ as before but with the expected rate of return changed from μ to r. So that in the risk-neutral model, the expected stock price grows at the same rate as the risk-free investment.

This risk-neutral assumption behind the Black & Scholes model constitutes a great advantage in a trading environment. In the risk-neutral world, all cash flows can be discounted using the risk-free rate (r) whereas, in a real word, the discount rate should take into account the risk premium (for reminder the risk premium is the investment return an asset is expected to yield in excess of the risk-free rate of return), which is more delicate...

We can classify the dynamics of this function in three special cases:

- If $\mu > \dfrac{1}{2}\sigma^2$, $X_t \rightarrow +\infty$ when $t \rightarrow +\infty$
- If $\mu < \dfrac{1}{2}\sigma^2$, $X_t \rightarrow 0$ when $t \rightarrow +\infty$
- If $\mu = \dfrac{1}{2}\sigma^2$, X_t will randomly fluctuate between high and small values when $t \rightarrow +\infty$

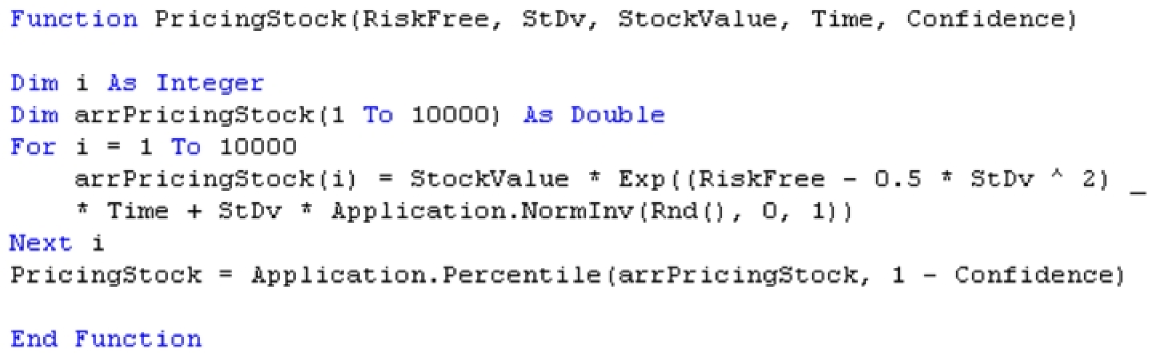
Following the request of a reader, although we have seen that the Brownian geometric motion was built higher from a log-Normal law of which we know the expected mean, let us do the inverse calculation, that is to say to calculate the expected mean of the Brownian geometric movement to evidently in the end fall back on our feet:

$$\begin{aligned}\mathrm{E}(X_t) &= \mathrm{E}\left(e^{\left(\mu-\frac{\sigma^2}{2}\right)t+\sigma\mathcal{N}(0,1)\sqrt{t}}\right) = X_0 e^{\left(\mu-\frac{\sigma^2}{2}\right)t}\mathrm{E}\left(e^{\sigma\mathcal{N}(0,1)\sqrt{t}}\right) \\ &= X_0 e^{\left(\mu-\frac{\sigma^2}{2}\right)t}\mathrm{E}\left(e^{\mathcal{N}(0,\sigma^2 t)}\right)\end{aligned} \tag{15.3.975}$$

So the term in parentheses follows by definition (or "by construction" if you prefer) a log-Normal law (see section Statistics page 734). The expected mean of the latter is then:

$$\mathrm{E}(X_t) = X_0 e^{\left(\mu-\frac{\sigma^2}{2}\right)t} e^{0+\frac{\sigma^2}{2}t} = X_0 e^{\left(\mu-\frac{\sigma^2}{2}\right)t} e^{\frac{\sigma^2}{2}t} = X_0 e^{\mu t} \tag{15.3.976}$$

To price an asset according to this model over a horizon, we can use the following Monte Carlo VBA function (the reader can also refer to our `R` or MATLAB® companion book for the same calculation):



```
Function PricingStock(RiskFree, StDv, StockValue, Time, Confidence)

Dim i As Integer
Dim arrPricingStock(1 To 10000) As Double
For i = 1 To 10000
    arrPricingStock(i) = StockValue * Exp((RiskFree - 0.5 * StDv ^ 2) _
    * Time + StDv * Application.NormInv(Rnd(), 0, 1))
Next i
PricingStock = Application.Percentile(arrPricingStock, 1 - Confidence)

End Function
```

We finally have a formulation (in the form of a probabilistic distribution function) of a temporal variation and of the intrinsic return of an asset that can be used for investment decision-making purposes on a forecast. But this model is still too smooth to model stock market crashes or bubbles (it is the same for recall with the standard Brownian motion) that can occur in the long term. This is why some more recent models that we will not study here add a Poisson process (discrete and with rare events by construction) to that of Wiener.

There are other models than the log-Normal but the latter is the most easy to understand and hence the most widespread. However, other more general methods must be encouraged!

To conclude this section, we summarize by a comparison the standard Brownian motion and the geometrical Brownian motion that govern the price dynamics when the parameters (instantaneous yield and volatility) are given in $\%$:

$$\begin{aligned}X_t &= X_0 + X_0\mu t + \mathcal{N}(0,1)\sigma X_0\sqrt{t} = X_0 + X_0\mu t + \sigma X_0 W_t = X_0(1+B_s)\\ X_t &= X_0 e^{\left(\mu-\frac{\sigma^2}{2}\right)t+\sigma\mathcal{N}(0,1)\sqrt{t}} = X_0 e^{\left(\mu-\frac{\sigma^2}{2}\right)t+\sigma W_t}\end{aligned} \tag{15.3.977}$$

and let us recall that the advantage of the geometrical Brownian motion is that it eliminates (thanks to the exponential) the negative values of the course which we could obtain with the standard Brownian motion of Bachelier.

It is then interesting for the financial analyst to visualize the expected mean as a function of T and the value $x(T)$ corresponding to a cumulative probability of 2.5% and 97.5% on a plot to have an idea of the evolution of the confidence interval at 95% of its $x(0)$. This is very easy to get in a spreadsheet software like Microsoft® Excel® and we can get by playing with several types of plots in the same chart on something like this (400 MUSD portfolio with 5% yield and 20% standard deviation):

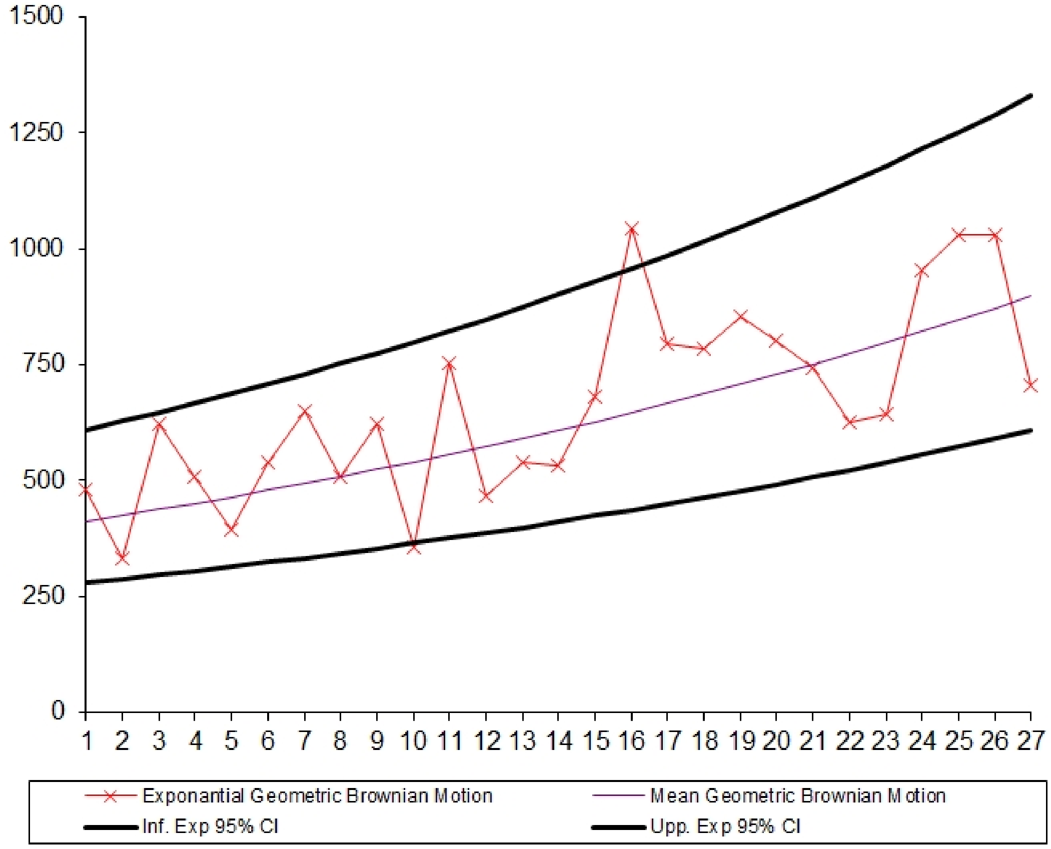
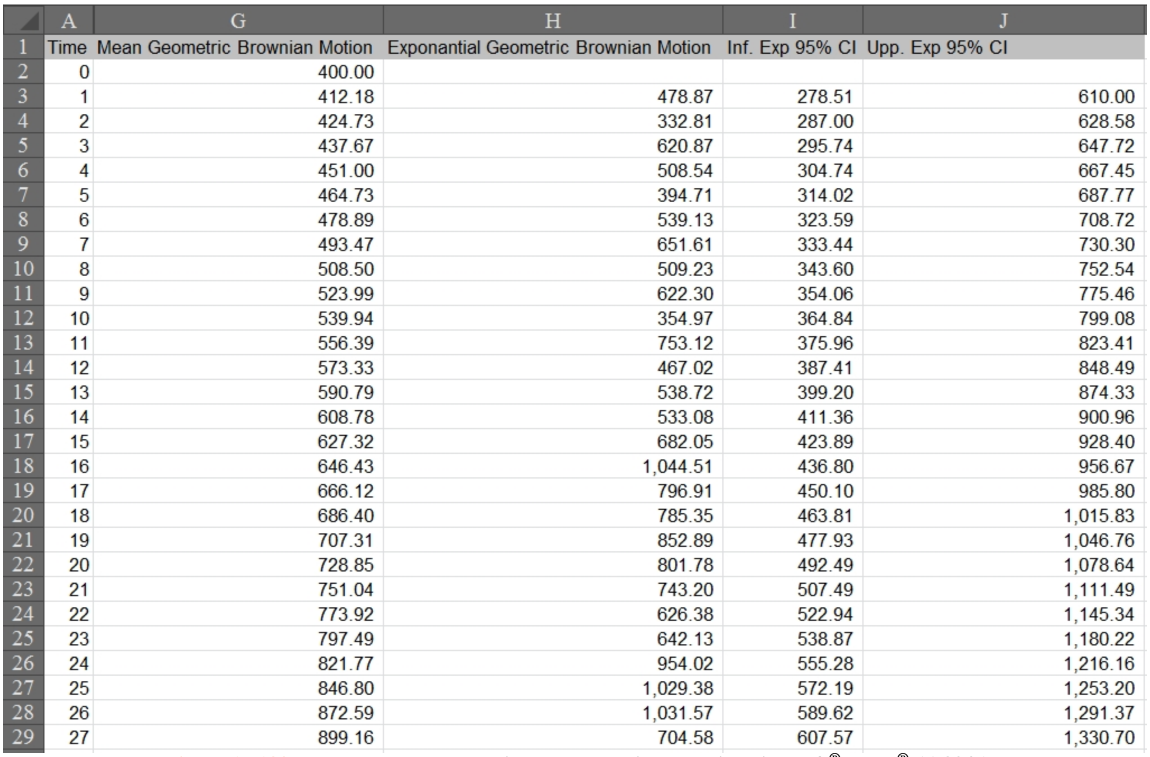


Figure 15.182: Geometric Brownian motion plot with 95% C.I. in Microsoft® Excel® 11.8346

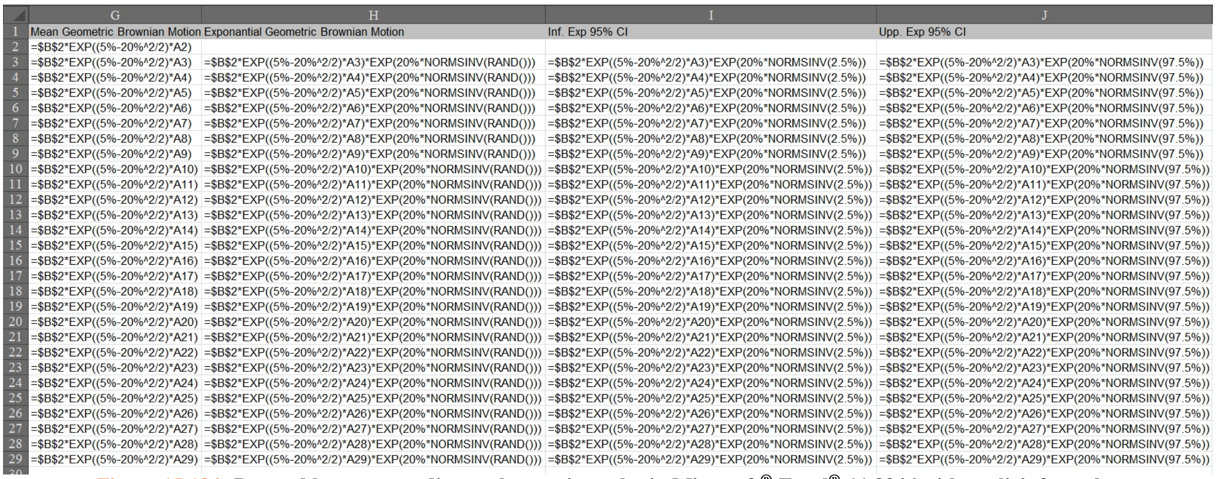
With the following structure in Microsoft® Excel®:



	A	G	H	I	J
1	Time	Mean Geometric Brownian Motion	Exponantial Geometric Brownian Motion	Inf. Exp 95% CI	Upp. Exp 95% CI
2	0	400.00			
3	1	412.18	478.87	278.51	610.00
4	2	424.73	332.81	287.00	628.58
5	3	437.67	620.87	295.74	647.72
6	4	451.00	508.54	304.74	667.45
7	5	464.73	394.71	314.02	687.77
8	6	478.89	539.13	323.59	708.72
9	7	493.47	651.61	333.44	730.30
10	8	508.50	509.23	343.60	752.54
11	9	523.99	622.30	354.06	775.46
12	10	539.94	354.97	364.84	799.08
13	11	556.39	753.12	375.96	823.41
14	12	573.33	467.02	387.41	848.49
15	13	590.79	538.72	399.20	874.33
16	14	608.78	533.08	411.36	900.96
17	15	627.32	682.05	423.89	928.40
18	16	646.43	1,044.51	436.80	956.67
19	17	666.12	796.91	450.10	985.80
20	18	686.40	785.35	463.81	1,015.83
21	19	707.31	852.89	477.93	1,046.76
22	20	728.85	801.78	492.49	1,078.64
23	21	751.04	743.20	507.49	1,111.49
24	22	773.92	626.38	522.94	1,145.34
25	23	797.49	642.13	538.87	1,180.22
26	24	821.77	954.02	555.28	1,216.16
27	25	846.80	1,029.38	572.19	1,253.20
28	26	872.59	1,031.57	589.62	1,291.37
29	27	899.16	704.58	607.57	1,330.70

Figure 15.183: **Data table corresponding to the previous plot in Microsoft® Excel® 11.8346**

with the explicit corresponding formulas to Microsoft® Excel® 11.8346 (don't hesitate to zoom in to see better):



	G	H	I	J
1	Mean Geometric Brownian Motion	Exponantial Geometric Brownian Motion	Inf. Exp 95% CI	Upp. Exp 95% CI
2	=$B$2*EXP((5%-20%^2/2)*A2)			
3	=$B$2*EXP((5%-20%^2/2)*A3)	=$B$2*EXP((5%-20%^2/2)*A3)*EXP(20%*NORMSINV(RAND()))	=$B$2*EXP((5%-20%^2/2)*A3)*EXP(20%*NORMSINV(2.5%))	=$B$2*EXP((5%-20%^2/2)*A3)*EXP(20%*NORMSINV(97.5%))
4	=$B$2*EXP((5%-20%^2/2)*A4)	=$B$2*EXP((5%-20%^2/2)*A4)*EXP(20%*NORMSINV(RAND()))	=$B$2*EXP((5%-20%^2/2)*A4)*EXP(20%*NORMSINV(2.5%))	=$B$2*EXP((5%-20%^2/2)*A4)*EXP(20%*NORMSINV(97.5%))
5	=$B$2*EXP((5%-20%^2/2)*A5)	=$B$2*EXP((5%-20%^2/2)*A5)*EXP(20%*NORMSINV(RAND()))	=$B$2*EXP((5%-20%^2/2)*A5)*EXP(20%*NORMSINV(2.5%))	=$B$2*EXP((5%-20%^2/2)*A5)*EXP(20%*NORMSINV(97.5%))
6	=$B$2*EXP((5%-20%^2/2)*A6)	=$B$2*EXP((5%-20%^2/2)*A6)*EXP(20%*NORMSINV(RAND()))	=$B$2*EXP((5%-20%^2/2)*A6)*EXP(20%*NORMSINV(2.5%))	=$B$2*EXP((5%-20%^2/2)*A6)*EXP(20%*NORMSINV(97.5%))
7	=$B$2*EXP((5%-20%^2/2)*A7)	=$B$2*EXP((5%-20%^2/2)*A7)*EXP(20%*NORMSINV(RAND()))	=$B$2*EXP((5%-20%^2/2)*A7)*EXP(20%*NORMSINV(2.5%))	=$B$2*EXP((5%-20%^2/2)*A7)*EXP(20%*NORMSINV(97.5%))
8	=$B$2*EXP((5%-20%^2/2)*A8)	=$B$2*EXP((5%-20%^2/2)*A8)*EXP(20%*NORMSINV(RAND()))	=$B$2*EXP((5%-20%^2/2)*A8)*EXP(20%*NORMSINV(2.5%))	=$B$2*EXP((5%-20%^2/2)*A8)*EXP(20%*NORMSINV(97.5%))
9	=$B$2*EXP((5%-20%^2/2)*A9)	=$B$2*EXP((5%-20%^2/2)*A9)*EXP(20%*NORMSINV(RAND()))	=$B$2*EXP((5%-20%^2/2)*A9)*EXP(20%*NORMSINV(2.5%))	=$B$2*EXP((5%-20%^2/2)*A9)*EXP(20%*NORMSINV(97.5%))
10	=$B$2*EXP((5%-20%^2/2)*A10)	=$B$2*EXP((5%-20%^2/2)*A10)*EXP(20%*NORMSINV(RAND()))	=$B$2*EXP((5%-20%^2/2)*A10)*EXP(20%*NORMSINV(2.5%))	=$B$2*EXP((5%-20%^2/2)*A10)*EXP(20%*NORMSINV(97.5%))
11	=$B$2*EXP((5%-20%^2/2)*A11)	=$B$2*EXP((5%-20%^2/2)*A11)*EXP(20%*NORMSINV(RAND()))	=$B$2*EXP((5%-20%^2/2)*A11)*EXP(20%*NORMSINV(2.5%))	=$B$2*EXP((5%-20%^2/2)*A11)*EXP(20%*NORMSINV(97.5%))
12	=$B$2*EXP((5%-20%^2/2)*A12)	=$B$2*EXP((5%-20%^2/2)*A12)*EXP(20%*NORMSINV(RAND()))	=$B$2*EXP((5%-20%^2/2)*A12)*EXP(20%*NORMSINV(2.5%))	=$B$2*EXP((5%-20%^2/2)*A12)*EXP(20%*NORMSINV(97.5%))
13	=$B$2*EXP((5%-20%^2/2)*A13)	=$B$2*EXP((5%-20%^2/2)*A13)*EXP(20%*NORMSINV(RAND()))	=$B$2*EXP((5%-20%^2/2)*A13)*EXP(20%*NORMSINV(2.5%))	=$B$2*EXP((5%-20%^2/2)*A13)*EXP(20%*NORMSINV(97.5%))
14	=$B$2*EXP((5%-20%^2/2)*A14)	=$B$2*EXP((5%-20%^2/2)*A14)*EXP(20%*NORMSINV(RAND()))	=$B$2*EXP((5%-20%^2/2)*A14)*EXP(20%*NORMSINV(2.5%))	=$B$2*EXP((5%-20%^2/2)*A14)*EXP(20%*NORMSINV(97.5%))
15	=$B$2*EXP((5%-20%^2/2)*A15)	=$B$2*EXP((5%-20%^2/2)*A15)*EXP(20%*NORMSINV(RAND()))	=$B$2*EXP((5%-20%^2/2)*A15)*EXP(20%*NORMSINV(2.5%))	=$B$2*EXP((5%-20%^2/2)*A15)*EXP(20%*NORMSINV(97.5%))
16	=$B$2*EXP((5%-20%^2/2)*A16)	=$B$2*EXP((5%-20%^2/2)*A16)*EXP(20%*NORMSINV(RAND()))	=$B$2*EXP((5%-20%^2/2)*A16)*EXP(20%*NORMSINV(2.5%))	=$B$2*EXP((5%-20%^2/2)*A16)*EXP(20%*NORMSINV(97.5%))
17	=$B$2*EXP((5%-20%^2/2)*A17)	=$B$2*EXP((5%-20%^2/2)*A17)*EXP(20%*NORMSINV(RAND()))	=$B$2*EXP((5%-20%^2/2)*A17)*EXP(20%*NORMSINV(2.5%))	=$B$2*EXP((5%-20%^2/2)*A17)*EXP(20%*NORMSINV(97.5%))
18	=$B$2*EXP((5%-20%^2/2)*A18)	=$B$2*EXP((5%-20%^2/2)*A18)*EXP(20%*NORMSINV(RAND()))	=$B$2*EXP((5%-20%^2/2)*A18)*EXP(20%*NORMSINV(2.5%))	=$B$2*EXP((5%-20%^2/2)*A18)*EXP(20%*NORMSINV(97.5%))
19	=$B$2*EXP((5%-20%^2/2)*A19)	=$B$2*EXP((5%-20%^2/2)*A19)*EXP(20%*NORMSINV(RAND()))	=$B$2*EXP((5%-20%^2/2)*A19)*EXP(20%*NORMSINV(2.5%))	=$B$2*EXP((5%-20%^2/2)*A19)*EXP(20%*NORMSINV(97.5%))
20	=$B$2*EXP((5%-20%^2/2)*A20)	=$B$2*EXP((5%-20%^2/2)*A20)*EXP(20%*NORMSINV(RAND()))	=$B$2*EXP((5%-20%^2/2)*A20)*EXP(20%*NORMSINV(2.5%))	=$B$2*EXP((5%-20%^2/2)*A20)*EXP(20%*NORMSINV(97.5%))
21	=$B$2*EXP((5%-20%^2/2)*A21)	=$B$2*EXP((5%-20%^2/2)*A21)*EXP(20%*NORMSINV(RAND()))	=$B$2*EXP((5%-20%^2/2)*A21)*EXP(20%*NORMSINV(2.5%))	=$B$2*EXP((5%-20%^2/2)*A21)*EXP(20%*NORMSINV(97.5%))
22	=$B$2*EXP((5%-20%^2/2)*A22)	=$B$2*EXP((5%-20%^2/2)*A22)*EXP(20%*NORMSINV(RAND()))	=$B$2*EXP((5%-20%^2/2)*A22)*EXP(20%*NORMSINV(2.5%))	=$B$2*EXP((5%-20%^2/2)*A22)*EXP(20%*NORMSINV(97.5%))
23	=$B$2*EXP((5%-20%^2/2)*A23)	=$B$2*EXP((5%-20%^2/2)*A23)*EXP(20%*NORMSINV(RAND()))	=$B$2*EXP((5%-20%^2/2)*A23)*EXP(20%*NORMSINV(2.5%))	=$B$2*EXP((5%-20%^2/2)*A23)*EXP(20%*NORMSINV(97.5%))
24	=$B$2*EXP((5%-20%^2/2)*A24)	=$B$2*EXP((5%-20%^2/2)*A24)*EXP(20%*NORMSINV(RAND()))	=$B$2*EXP((5%-20%^2/2)*A24)*EXP(20%*NORMSINV(2.5%))	=$B$2*EXP((5%-20%^2/2)*A24)*EXP(20%*NORMSINV(97.5%))
25	=$B$2*EXP((5%-20%^2/2)*A25)	=$B$2*EXP((5%-20%^2/2)*A25)*EXP(20%*NORMSINV(RAND()))	=$B$2*EXP((5%-20%^2/2)*A25)*EXP(20%*NORMSINV(2.5%))	=$B$2*EXP((5%-20%^2/2)*A25)*EXP(20%*NORMSINV(97.5%))
26	=$B$2*EXP((5%-20%^2/2)*A26)	=$B$2*EXP((5%-20%^2/2)*A26)*EXP(20%*NORMSINV(RAND()))	=$B$2*EXP((5%-20%^2/2)*A26)*EXP(20%*NORMSINV(2.5%))	=$B$2*EXP((5%-20%^2/2)*A26)*EXP(20%*NORMSINV(97.5%))
27	=$B$2*EXP((5%-20%^2/2)*A27)	=$B$2*EXP((5%-20%^2/2)*A27)*EXP(20%*NORMSINV(RAND()))	=$B$2*EXP((5%-20%^2/2)*A27)*EXP(20%*NORMSINV(2.5%))	=$B$2*EXP((5%-20%^2/2)*A27)*EXP(20%*NORMSINV(97.5%))
28	=$B$2*EXP((5%-20%^2/2)*A28)	=$B$2*EXP((5%-20%^2/2)*A28)*EXP(20%*NORMSINV(RAND()))	=$B$2*EXP((5%-20%^2/2)*A28)*EXP(20%*NORMSINV(2.5%))	=$B$2*EXP((5%-20%^2/2)*A28)*EXP(20%*NORMSINV(97.5%))
29	=$B$2*EXP((5%-20%^2/2)*A29)	=$B$2*EXP((5%-20%^2/2)*A29)*EXP(20%*NORMSINV(RAND()))	=$B$2*EXP((5%-20%^2/2)*A29)*EXP(20%*NORMSINV(2.5%))	=$B$2*EXP((5%-20%^2/2)*A29)*EXP(20%*NORMSINV(97.5%))

Figure 15.184: **Data table corresponding to the previous plot in Microsoft® Excel® 11.8346 with explicit formulas**

It is interesting to compare the evolution of this portfolio with the same parameters (volatility and yield) over the same period of time but with the standard brownian motion. This graphically gives:

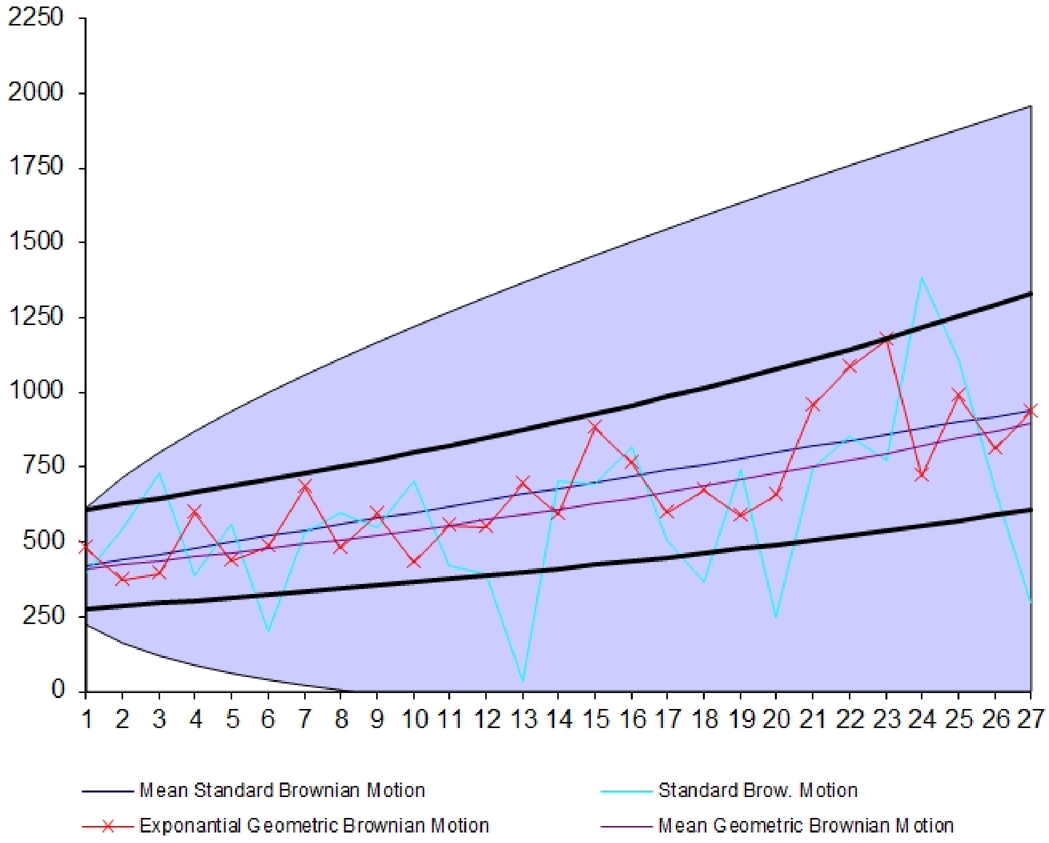


Figure 15.185: **Comparison of geometric and standard brownian motion**

Finally, the reader will notice that we can generalize the writing of the two Brownian process (taking the Napierian logarithm of the Brownian geometric process):

$$\begin{aligned}\ln(X_t) &= \ln\left(e^{\left(\mu-\frac{\sigma^2}{2}\right)t}+\sigma\mathcal{N}(0,1)\sqrt{t}\right)=\ln(X_0)+\ln\left(e^{\left(\mu-\frac{\sigma^2}{2}\right)t+\sigma\mathcal{N}(0,1)\sqrt{t}}\right)\\ &=\ln(X_0)+\left(\mu-\frac{\sigma^2}{2}\right)t+\sigma\mathcal{N}(0,1)\sqrt{t}\end{aligned} \quad (15.3.978)$$

by writing them in a form proposed by Benoît Mandelbrot in 11962 (holocene calendar):

$$X_t = X_0+\delta t+cL_t^{\alpha} \quad (15.3.979)$$

where δ and c are respectively the localization parameters (average return) and dispersion (non-Gaussian volatility) of the process, and where L_t^{α} designates the standard α-stable Lévy process.

The problem with this model is the loss of existence, for certain laws of probability that work very well, of the second moment (the variance) so important in terms of communication and images for professionals in the 11970s (holocene calendar) because it served them as the only measure of risk. The absence of finite variance was probably one of the most powerful causes of rejection (brain inertia not of customers but of board of directors...).

15.3.10.3.9.7 Itō's product and quotient rule

Before continuing on Black & Scholes pricing formula derivation, let us open a very small necessary parenthesis about Itō's approximation.

When we will study later how two assets are correlated to each other (during our proof of Margrabe formula further below), we will need what a majority of textbooks name "Itō's quotient rule" ! But to prove that latter, we will need first to derive "Itō's product rule" (even if in fact's it is just a again a simple Taylor approximation!). So let us see that!

Let $X^2=f(X)$. Then 2nd order Taylor development gives us (see section Sequences and Series page1698):

$$\mathrm{d}f(X)\cong f'(X)\mathrm{d}X+\frac{1}{2}f''(X)(\mathrm{d}X)^2=2X\mathrm{d}X+(\mathrm{d}X)^2 \tag{15.3.980}$$

But we also have (just a simple mathematical trick):

$$\mathrm{d}(XY)=\mathrm{d}\left(\frac{1}{2}((X+Y)^2-X^2-Y^2)\right)=\frac{1}{2}\left(\mathrm{d}(X+Y)^2-\mathrm{d}X^2-\mathrm{d}Y^2\right) \tag{15.3.981}$$

Now applying the Itō's lemma on each of the three terms above lead us to:

$$\begin{aligned}\mathrm{d}(X+Y)^2&\cong 2(X+Y)\mathrm{d}(X+Y)+(\mathrm{d}(X+Y))^2\\&=2\left(X\mathrm{d}X+X\mathrm{d}Y+Y\mathrm{d}Y+Y\mathrm{d}X\right)+(\mathrm{d}X)^2+2\mathrm{d}X\mathrm{d}Y+(\mathrm{d}Y)^2\big)\\\mathrm{d}X^2&\cong 2X\mathrm{d}X+(\mathrm{d}X)^2\\\mathrm{d}Y^2&\cong 2Y\mathrm{d}Y+(\mathrm{d}Y)^2\end{aligned} \tag{15.3.982}$$

Therefore:

$$\mathrm{d}(XY)\cong\frac{1}{2}(2X\mathrm{d}Y+2Y\mathrm{d}X+2\mathrm{d}X\mathrm{d}Y)=\mathrm{d}X\mathrm{d}Y+Y\mathrm{d}X+\mathrm{d}X\mathrm{d}Y \tag{15.3.983}$$

Let us now prove Itō's quotient rule given by:

$$\mathrm{d}\left(\frac{X}{Y}\right)=\frac{X}{Y}\left(\frac{\mathrm{d}X}{X}-\frac{\mathrm{d}Y}{Y}-\frac{\mathrm{d}X}{X}\frac{\mathrm{d}Y}{Y}+\left(\frac{\mathrm{d}Y}{Y}\right)^2\right) \tag{15.3.984}$$

We apply Itō's product rule to:

$$\mathrm{d}\left(X\frac{1}{Y}\right)\cong\frac{1}{Y}\mathrm{d}X+X\mathrm{d}\frac{1}{Y}+\mathrm{d}X\mathrm{d}\frac{1}{Y}=\frac{X}{Y}\left(\frac{\mathrm{d}X}{X}-\frac{\mathrm{d}Y}{Y}-\frac{\mathrm{d}X\mathrm{d}Y}{XY}\right) \tag{15.3.985}$$

Keep in mind that when we use the Taylor approximation:

$$\mathrm{d}f(Y)\cong f'(Y)\mathrm{d}Y+\frac{1}{2}f''(Y)(\mathrm{d}Y)^2 \tag{15.3.986}$$

So:

$$\mathrm{d}\left(Y^{-1}\right)\cong -Y^{-2}\mathrm{d}Y+2\frac{1}{2}Y^{-3}(\mathrm{d}Y)^2 \tag{15.3.987}$$

We put that back in $\mathrm{d}XY$ to get:

$$\mathrm{d}(XY)\cong\frac{1}{2}(2X\mathrm{d}Y+2Y\mathrm{d}X+2\mathrm{d}X\mathrm{d}Y)=\mathrm{d}X\mathrm{d}Y+Y\mathrm{d}X+\mathrm{d}X\mathrm{d}Y \tag{15.3.988}$$

Notice that actually in the huge majority of textbooks introducing these both rules, as for the Black & Scholes pricing relations, the $\cong$ is often abusively replaced by an $=$ sign...

15.3.10.3.9.8 Black & Scholes differential equation

We have obtained during the preceding developments the following differential equation for the random walk of the value of an asset under the constraint of a log-normal law and a Brownian motion:

$$\mathrm{d}F=\left(\frac{\partial F}{\partial t}+a\frac{\partial F}{\partial x}+\frac{1}{2}b^2\frac{\partial^2 F}{\partial x^2}\right)\mathrm{d}t+b\frac{\partial F}{\partial x}\mathrm{d}z=\left(\frac{\partial F}{\partial t}+\mu x\frac{\partial F}{\partial x}+\frac{1}{2}\sigma^2x^2\frac{\partial^2 F}{\partial x^2}\right)\mathrm{d}t+\sigma x\frac{\partial F}{\partial x}\mathrm{d}z \tag{15.3.989}$$

where for reminder μ and σ are expressed in percentage!

Either with the traditional notations in finance:

$$\mathrm{d}F=\left(\frac{\partial F}{\partial t}+a\frac{\partial F}{\partial S}+\frac{1}{2}b^2\frac{\partial^2 F}{\partial S^2}\right)\mathrm{d}t+b\frac{\partial F}{\partial S}\mathrm{d}z\underset{\substack{a=\mu x\\ b=\sigma x\\ x=S}}{=}\left(\frac{\partial F}{\partial t}+\mu S\frac{\partial F}{\partial S}+\frac{1}{2}\sigma^2S^2\frac{\partial^2 F}{\partial S^2}\right)\mathrm{d}t+\sigma S\frac{\partial F}{\partial S}\mathrm{d}z \tag{15.3.990}$$

If we now construct a neutral (not risky for recall) delta portfolio consisting of one option in short position and a number $-\Delta$ of underlying securities (often also denoted $-\Delta_t$ in the literature) in a long position (or conversely it does not matter).

The value of the portfolio Π is then expressed by:

$$\Pi=F-\Delta S \tag{15.3.991}$$

The idea of the equality above is the same as when we have a portfolio with a cash amount S of a security (i.e. the total cash amount of the securities in the securities portfolio) and we protect ourselves, for example, from a downward trend with the acquisition of a number of Put but with the difference that here the total cash amount of options that is set at F (i.e. the total cash amount of the options in the options portfolio) and that we look for the number of underlying to buy (hence the negative sign) in the case of a supposed bearish trend.

The time differential of the portfolio is then written (which assumes that we make instantaneous delta hedging without costs, what in practice is quite difficult to achieve...!):

$$\mathrm{d}\Pi=\mathrm{d}F-\Delta\mathrm{d}S \tag{15.3.992}$$

You will notice that we assume constant the quantity Δ during the time differential...

> Caution!!! Some authors choose, by convention, to invert the signs of the terms that are on the right of the equality of the two preceding relations. This does not change the final result of the solution of the differential equation but it must be known that this can be done since it is only a protection choice of bullish or bearish trend.

By combining the previous relations and (here we adopt the traditional notation used in the finance field) the equation of the risky asset given by:

$$\frac{\mathrm{d}S}{S}=\sigma\mathrm{d}z+\mu\mathrm{d}t \tag{15.3.993}$$

we get:

$$\mathrm{d}\Pi=\mathrm{d}F-\Delta\mathrm{d}S=\left[\left(\frac{\partial F}{\partial t}+\mu S\frac{\partial F}{\partial S}+\frac{1}{2}\sigma^2S^2\frac{\partial^2F}{\partial S^2}\right)\mathrm{d}t+\sigma S\frac{\partial F}{\partial S}\mathrm{d}z\right]-\Delta[S\sigma\mathrm{d}z+S\mu\mathrm{d}t] \tag{15.3.994}$$

where we have in the right-hand square brackets the geometrical Brownian motion.

This gives, after rearrangement of terms, the differential equation of the portfolio:

$$\mathrm{d}\Pi=\sigma S\left(\frac{\partial F}{\partial S}-\Delta\right)\mathrm{d}z+\left(\mu S\frac{\partial F}{\partial S}+\frac{1}{2}\sigma^2S^2\frac{\partial^2F}{\partial S^2}+\frac{\partial F}{\partial t}-\mu\Delta S\right)\mathrm{d}t \tag{15.3.995}$$

Consider now that Δ is bound by the speculative dependence relation (of which we take the integer value) which eliminates from the previous relation the randomly risky part of the portfolio (it is the $\mathrm{d}z$ generates the risk randomly for recall!) such that (we will see again this expression of the delta during our study of the "Greeks" further below):

$$\frac{\partial F}{\partial S}-\Delta=0\Leftrightarrow\Delta=\frac{\partial F}{\partial S} \tag{15.3.996}$$

We can then write the infinitesimal variation without risk of our portfolio:

$$\mathrm{d}\Pi=\left(\mu S\frac{\partial F}{\partial S}+\frac{1}{2}\sigma^2S^2\frac{\partial^2F}{\partial S^2}+\frac{\partial F}{\partial t}-\mu\Delta S\right)\mathrm{d}t \tag{15.3.997}$$

But, we also have for a portfolio placed in a risk-free investment (since we have just eliminated the random risk we can write it like this):

$$\mathrm{d}\Pi=r\Pi\mathrm{d}t \tag{15.3.998}$$

That is sometimes also written in the literature:

$$\mathrm{d}S_t^0=rS_t^0\mathrm{d}t \tag{15.3.999}$$

The attentive reader will have noticed that the two relations above:

$$\begin{gathered}\mathrm{d}\Pi=\left(\mu S\frac{\partial F}{\partial S}+\frac{1}{2}\sigma^2S^2\frac{\partial^2F}{\partial S^2}+\frac{\partial F}{\partial t}-\mu\Delta S\right)\mathrm{d}t\\ \mathrm{d}\Pi=r\Pi\mathrm{d}t\end{gathered} \tag{15.3.1000}$$

are therefore a situation of non-arbitrage.

Now substituting the four relations:

$$\Pi=F-\Delta S\qquad\Delta=\frac{\partial F}{\partial S}\qquad\mathrm{d}\Pi=r\Pi\mathrm{d}t\qquad r=\mu \tag{15.3.1001}$$

into:

$$\mathrm{d}\Pi=\left(\mu S\frac{\partial F}{\partial S}+\frac{1}{2}\sigma^2S^2\frac{\partial^2F}{\partial S^2}+\frac{\partial F}{\partial t}-\mu\Delta S\right)\mathrm{d}t \tag{15.3.1002}$$

We get:

$$\boxed{\frac{\partial F}{\partial t}+\frac{1}{2}\sigma^2 S^2\frac{\partial^2 F}{\partial S^2}+rS\frac{\partial F}{\partial S}-rF=0} \tag{15.3.1003}$$

which is nothing else than the "Black & Scholes linear parabolic partial differential equation (without second member) of the second order" (with non-constant coefficient S). This relation is therefore valid only under the assumption that the return r is the risk-free one on the market and therefore that the Brownian motion is a Brownian motion at neutral risk! Similarly, this last relation also assumes that the variance is independent of time!

Remark

The Black & Scholes PDE is in fact a special case of a more general PDE known in Quantum Field Theory and denoted:

$$\frac{\partial u}{\partial t}(x,t)+\mu(x,t)\frac{\partial u}{\partial x}(x,t)+\frac{1}{2}\sigma^2(x,t)\frac{\partial^2 u}{\partial x^2}(x,t)-V(x,t)u(x,t)+f(x,t)=0$$

where μ,σ,ψ,V,f are known functions, T is a parameter, and $u:\mathbb{R}\times[0,T]\rightarrow\mathbb{R}$ is the unknown. Then the Feynman-Kac formula tells us that the solution can be written as a conditional expectation:

$$u(x,t)=\mathrm{E}^Q\left[\int_t^T \exp\left(-\int_t^r V\left(X_\tau,\tau\right)\mathrm{d}\tau\right) f\left(X_r,r\right)dr+\exp\left(-\int_t^T V\left(X_\tau,\tau\right)\mathrm{d}\tau\right)\psi\left(X_T\right)\middle| X_t=x\right]$$

under the probability measure Q such that X is an Itô process driven by the equation:

$$\mathrm{d}X_t=\mu(X,t)\mathrm{d}t+\sigma(X,t)\mathrm{d}W_t^Q$$

with $W^Q(t)$ is a Wiener process (also named "Brownian motion") under Q, and the initial condition for $X(t)$ is $X(t)=x$.

The fact that the drift μ is equal to the risk-free return r also has a significant impact on the interpretation of this differential equation. This indeed assumes indirectly that two options with different drifts have their implied return which is equal and that is taken as the one without risk of the market !!! This is the reason why in finance if we have two options whose drift differs and the variance and maturity are identical, the Black & Scholes equation gives the same pricing by construction !!!

This last equation is more often written in the form relating to the implicit reference of a Call such as:

$$\frac{\partial C}{\partial t}+\frac{1}{2}\sigma^2 S^2\frac{\partial^2 C}{\partial S^2}+rS\frac{\partial C}{\partial S}-rC=0 \tag{15.3.1004}$$

Or in a slightly more condensed form (depending on the textbooks...):

$$\partial_t C+\frac{1}{2}\sigma^2 S^2\partial_s^2 C+rS\partial_S C-rC=0 \tag{15.3.1005}$$

and even more condensed... (but here it becomes frankly an abused notation...):

$$C_t+\frac{1}{2}\sigma^2 S^2 C_{SS}+rSC_S-rC=0 \tag{15.3.1006}$$

The reader will have noticed that the parameter μ (derivation) is absent from this equation! In other words, the value of an option is independent of the rate of change in the values of the underlying securities! The only parameter that affects the price of the option is the yield volatility σ of the underlying. A consequence of this is that two people with divergent opinions about the value of μ are still in agreement on the value of the option.

The objective is obviously to solve this differential equation in order to determine the return $F(S,t)$. This can not be solved in two lines...

Before working on this hard painful task some definitions and practical preconditions concerning certain parameters are useful and necessary (we will determine their explicit form after the resolution of the differential equation):

15.3.10.3.9.9 Self-financing portfolio on underlying

A "self-financing portfolio" strategy is a dynamic strategy of buying or selling securities and loans or borrowing from the bank, the value of which is not modified by the addition or withdrawal of cash (we could have introduced this subject at the beginning of this section but we thought it more appropriate to do so only now). In other words: a portfolio is self-financing if there is no exogenous infusion or withdrawal of money; the purchase of a new asset must be financed by the sale of an old one.

We will assume here for the example that we can only invest in a single security (risky investment), and in cash (supposedly non-risky investment), that is to say for that latter by placing or borrowing money at a bank.

We denote by S_t the price on the date t of the security, by r_t the interest rate for an investment between $[t, t+\mathrm{d}t]$ at the bank.

Given V_t the market value, or the net asset value, or even the "Mark to Market value (MtM)" of the portfolio at date t. After renegotiation, the number of shares δ_t in the portfolio is constant until the next management expiry date. For the sake of simplicity, we assume for the moment that the manager takes into account the value of the underlying price in his decision rule only at the time of renegotiation.

In a very short time, the change in value of the portfolio is due only to the change in the value of the underlying and to the interest paid by the bank on the cash, i.e., since the amount invested in the cash is:

$$V_t - \delta_t \cdot S_t \qquad (15.3.1007)$$

we have the "self-financing equation":

$$\boxed{\mathrm{d}V_t = \delta_t \cdot \mathrm{d}S_t + (V_t - \delta_t \cdot S_t) r_t \mathrm{d}t = V_t r_t \mathrm{d}t + \delta_t \cdot (\mathrm{d}S_t - r_t S_t \mathrm{d}t)} \qquad (15.3.1008)$$

Thus, for a Call seller (for example ...), the purpose is to find the initial cost V_0 and strategy δ_t that make it possible to get (the finance practitioner talk about "realizing the financial asset"):

$$(S_T - K)^+ \qquad (15.3.1009)$$

in all market scenarios. If there is such a hedging strategy, then we say that we are dealing with a "complete market".

15.3.10.3.9.10 Greeks and others...

In the reality these theoretical models are very beautiful to read (for the honour of the human spirit) but quite unrealistic. Reason why traders use (as in engineering) simple and comprehensible indicators named the "Greek" to make their sales or purchase decisions (or to make the information understandable by the Director Board...). However, with a little creativity, the same indicators can be used in many other fields (manufacturing, project management, chemistry, etc.).

Here is the list of the best known one (we will calculate their explicit expression for some models that we will develop further below):

Definitions (#15.3.696):

D1. The "delta" of an option, which is important to understand (or know), is given by:

$$\boxed{\Delta=\frac{\partial F}{\partial S}} \tag{15.3.1010}$$

and represents the rate of change in the value of the portfolio's options, depending on the values of the underlying securities S (mathematically, therefore, it is the first derivative of the option premium on the price of the underlying). This term is fundamental in theory (it is a hypothesis in the construction of the Black & Scholes model as we saw above) and in practice and we will use it frequently. It is therefore a measure of the correlation between the movement of the option (or other financial assets and derivatives) and the underlying assets.

✍**Example:**

Let us consider for example, that a Call on the ABC share has a delta of 0.25 with a price of the underlying (spot) of 90.- and a premium at 5.-. This mean that when the price of the share ABC changes from 90.- to 91.-, the premium of the option will then increase by 1 delta, and then becomes 5.25.-. When the price of the ABC share goes from 90.- to 88.-, the premium of the option will decrease by 2 times delta, and becomes 4.50.-. This variation in terms of delta (integer) is then denoted δ_t.

Remark

Many practitioners prefer to use the "delta cash" defined by the product of the delta and the spot price. So in our example above the delta cash is worth 90 times 0.25.- which is worth 22.5.-. If the share price increases from 90.- to 91.- i.e. 1.09% increase, the variation of the Call is the product of the delta cash by the variation in $\%$ ie 22.5.- times 1.09% which gives a positive variation of 0.25.- and then we fall back on our $5+0.25=5.25$.-.

The delta is therefore an important parameter for a practitioner who wants to cover himself against the risk. Indeed, in order to obtain the global delta of a position, it is sufficient to multiply the value of the delta of each option by its position. Then we add up all these deltas.

For example, if on the same underlying we are a seller of 5 calls C_1 and a buyer of 7 Call C_2 then our global delta will be equal to:

$$7 \cdot \Delta_{C_2} - 5 \cdot \Delta_{C_1} \tag{15.3.1011}$$

The value of this parameter informs us about the quantity of underlyings to be bought or sold (opposite position on the underlying) in order to immunize the valuation of our portfolio to changes in the price of this underlying. We then say that it is a "delta-neutral strategy" or a "delta-hedging strategy". More precisely, it is a dynamic strategy (dynamic hedging) because it is necessary to maintain a delta-neutral portfolio throughout the life of the option (in theory instantly for recall!). Contrary to statistical strategies (static hedging). For example, a static strategy to hedge the sale of a Call of strike 125.- out of $1,000$ shares (spot price 100.-), may consist of buying immediately $1,000$ shares on the market for $100,000$.-, and wait for the due date. In this situation, if the buyer exercises the Call the seller is able to deliver the $1,000$ shares for $125,000$.-. But if he does not exercise it, the seller stays with $1,000$ shares... Similarly, if the final price is set at eg 80.-, the seller records a loss of $20,000$.-. Delta-hedging eliminates this risk.

However, the strategy of dynamic hedging has a vicious side because if the majority of the market players buy underlying to hedge against an increase in the prices, then this will increase the value of the underlying even more... Supply and Demand rule...) and thus they will buy even more underlyings, which will have the effect of increasing even more its value and so on and then we go in circles until reaching a potential Crash of the markets (the 11987 crash - according to holocene calendar - would be originally due to this type of perverse dynamics, same for some automated trading bots). The effect will be the same (but opposite) to protect against a decrease... It is for this reason that it is strongly recommended to make delta hedging with other diversified financial assets than the underlying itself!!!!!

Thus, the managers, will between the date on which they received the premium (by having sold an option contract) and its maturity T, naturally manage in delta-neutral over time a self-funded portfolio made up of δ_t underlying assets S (or equivalent replicated one) at every instant t, in order to dispose of the stochastic target at certainty (thus without risk) at maturity. We are also talking about "portfolio hedging". However this strategy masks a terrible fault at the global level, indeed if a market player applies this strategy by selling massively (for example to protect itself from a bearish trend), then all the other actors will also perhaps sell then increasing by extension the delta and thus amplifying the volatility and so on until the crash.

D2. The "theta" of an option gives the sensitivity of the option price to its maturity and is given by:

$$\boxed{\theta = \frac{\partial F}{\partial t}} \tag{15.3.1012}$$

Applied to our imaginary portfolio, the theta simply gives us the value lost or gained after one day passed for example.

D3. The "rho" calculates the sensitivity of the option price to the geometric interest market rate without risk and is given by:

$$\boxed{\rho=\frac{\partial F}{\partial r}} \tag{15.3.1013}$$

This indicator seems to be relatively little used by professionals.

D4. The "vega", represented by the lowercase greek letter nu because the name vega is not itself a Greek letter name..., measures the sensitivity of the option to volatility and is given by:

$$\boxed{\nu=\frac{\partial F}{\partial \sigma}} \tag{15.3.1014}$$

D5. The "gamma" corresponds to the derivative of the delta and is therefore given by:

$$\boxed{\Gamma=\frac{\partial^2 F}{\partial S^2}} \tag{15.3.1015}$$

A possible reading of the gamma is the direction of evolution of the delta as a function of the price of the underlying. A positive gamma indicates that price of the underlying and the delta move in the same direction, whereas a negative gamma shows the opposite.

In order to avoid adjusting the Delta in delta-neutral strategy, and thus to pay the transaction costs but also the inherent spreads, a good way is to have the delta as stable as possible. Having the most stable delta means little or no change in its initial value, thus having a "Gamma-Neutral strategy" in addition to having a delta-neutral strategy. However, since the delta-neutral strategy means that we already play with the underlying to make the delta constant, we can not use the same underlying to make the gamma neutral. The gamma hedging principle of a portfolio containing at least one option is then to find another option on the market in proportions that compensate for the gamma of the original position.

Remark

As we will see later during our study of risk analysis in Finance using Taylor series developments, the delta does in fact allow for hedging only in the first order (that is, for small variations of the underlying). By supplementing with gamma-hedging we eliminate the effects of the second order. So a good strategy is to ensure the instantaneous values of delta and gamma to be zero.

✍**Example:**

Let us consider as summary of all these Greeks a Call with an underlying with a spot price of 100.-, a strike of 100.-, a volatility return of 24% and a geometric average market interest rate without risk of 5%. Consider that we have the following values:

$$\delta=0.62 \qquad \Gamma=0.02 \qquad \theta=-0.02.\text{-} \qquad \nu=0.37.\text{-} \qquad \rho=0.5.\text{-} \tag{15.3.1016}$$

The value of the delta thus tells us that if the price of the underlying increase of 1.-, the value of the option will increase by 0.62.-. The gamma tells us that if the underlying price increase of 1.-, the delta of the option will go from 0.62.- to 0.64.-. The value of the theta tells us at the end of the day that the value of the option will have decreased by 0.02.-. The value of the vega tells us that if the volatility of the underlying increases from 24% to 25%, the value of the option will increase by 0.37. Finally, the value of the rho indicates that if the value of the risk-free geometric average market rate increases from 5% to 5.5%, the value of the option will increase by 0.25.-.

D6. The "Black & Scholes linear differential operator" L_{BS} given by:

$$L_{\text{BS}}=\frac{\partial}{\partial t}+\frac{1}{2}\sigma^2S^2\frac{\partial^2}{\partial S^2}+rS\frac{\partial}{\partial S}-r \tag{15.3.1017}$$

seems to have a financial interpretation as a measure of the difference between the return of an option (the first two terms) and a set of a risk-free return portfolio containing that option (the last two terms).

In short, this being said, we can therefore have the following technical writing of the Black & Scholes PDE (with non-constant coefficient S) using the Greeks:

$$\boxed{\theta+\frac{1}{2}\sigma^2S^2\Gamma+rS\Delta-rF=0} \tag{15.3.1018}$$

Here is a quite good visual summary of all Greeks for a special case:

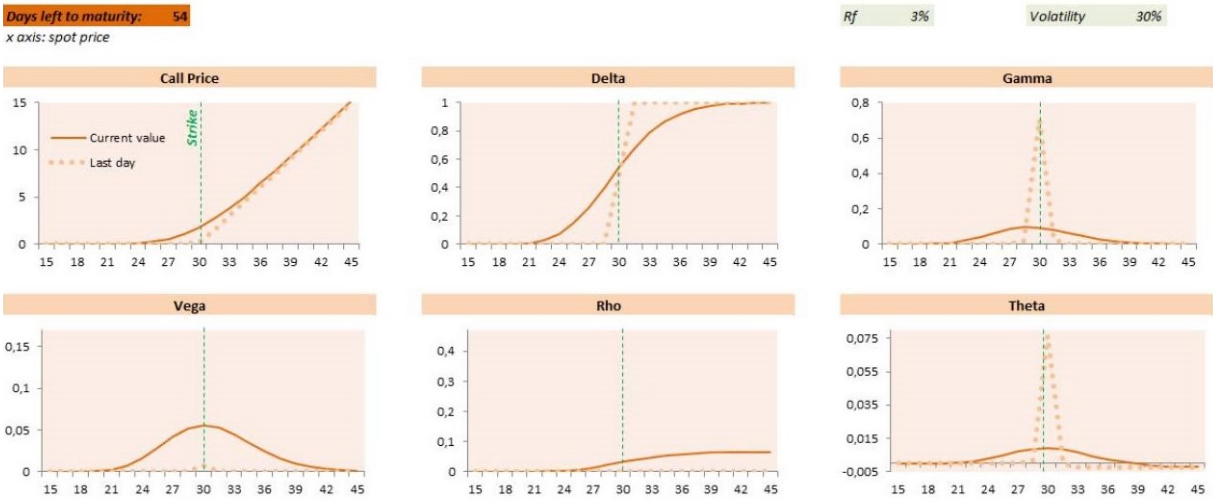


Figure 15.186: **Greeks summary**

15.3.10.3.9.11 Solving the Black & Scholes differential equation

Before solving the Black & Scholes PDE let us first already give the solutions with a reminder of the terms (this will allow to have a preliminary idea of the concepts used during the developments and I do not risk to write these before some years due to lack of time so it's better to have them already while waiting...):

Let $F(S,t)$ denote the value of a Call $C(S,t)$ or Put $P(S,t)$ option, σ the yield volatility of the underlying, E the strike, T the date of expiration and r the interest and S the price of the underlying. We will solve the differential equation of BS (with non-constant coefficient S) written the following form:

$$\frac{\partial C}{\partial t}+\frac{1}{2}\sigma^2 S^2\frac{\partial^2 C}{\partial S^2}+rS\frac{\partial C}{\partial S}-rC=0 \qquad (15.3.1019)$$

The trick (which led to a Nobel Prize in Economics for recall) is to see that it is a differential equation with partial differentials and more particularly with differentials of the first order in t and second order in S. There exists a differential equation of this type which is very well known in physics: the heat equation (see section Thermodynamics page 3684). The idea is then by changes of variables to get back to it. The difficulty was to find the right changes of variables and our predecessors took care of this error and trials... for us (thanks to them!). We then start by writing (this choice also allows to shift to initial conditions similar to those of the equation of heat):

$$C(t,S)=Ku(\tau,S) \qquad \tau=\sigma^2(T-t) \qquad x=\ln\left(\frac{S}{K}\right) \qquad (15.3.1020)$$

and therefore explicitly:

$$C(t,S)=Ku(\tau,S)=Ku\left(\sigma^2(T-t),\ln\left(\frac{S}{K}\right)\right) \qquad (15.3.1021)$$

We will use further below the fact that:

$$\frac{\partial\tau}{\partial t}=-\sigma^2 \qquad \frac{\partial x}{\partial S}=\frac{1}{S} \qquad (15.3.1022)$$

For practical purposes, let us write the inverse transformation (which will be useful to us much later) that follows:

$$u(\tau,x)=\frac{1}{K}C(t,S)=\frac{1}{K}C\left(T-\frac{\tau}{\sigma^2},Ke^x\right) \qquad (15.3.1023)$$

A part of the underlying idea of this choice of change of variables is to transform the fact that in the differential equation of Black & Scholes we know the final conditions (since at maturity the price of the option is perfectly Determined) whereas with the heat equation we need initial and not final conditions! We will see immediately that this choice of change of variables allows this. Indeed, remember that at maturity we know that:

$$C(T,S)=(S_T-K)^+ \qquad (15.3.1024)$$

Therefore:

$$\frac{1}{K}C(T,S)=\left(\frac{S_T}{K}-1\right)^{+} \tag{15.3.1025}$$

As:

$$x=\ln\left(\frac{S_T}{K}\right)\Rightarrow\frac{S_T}{K}=e^x \tag{15.3.1026}$$

We then have:

$$\frac{1}{K}C(T,S)=\left(e^x-1\right)^{+} \tag{15.3.1027}$$

And the fact what allows us to have an initial condition is that:

$$\frac{1}{K}C(T,S)=\left(e^x-1\right)^{+}=u(\tau,x)=u\left(\sigma^2(T-t),x\right)\underset{t=T}{=}u(0,x) \tag{15.3.1028}$$

To transform our differential equation to partial derivatives (with non-constant coefficient S):

$$\frac{\partial C}{\partial t}+\frac{1}{2}\sigma^2S^2\frac{\partial^2C}{\partial S^2}+rS\frac{\partial C}{\partial S}-rC=0 \tag{15.3.1029}$$

With these new variables we have to calculate some partial derivatives of C. We then have:

$$\begin{aligned}\frac{\partial C}{\partial t}&=\frac{\partial}{\partial t}\left[K\cdot u\left(\sigma^2(T-t),\ln\left(\frac{S}{K}\right)\right)\right]=K\frac{\partial}{\partial t}\left[u\left(\sigma^2(T-t),\ln\left(\frac{S}{K}\right)\right)\right]\\&=K\frac{\partial}{-\frac{\partial\tau}{\sigma^2}}\left[u\left(\sigma^2(T-t),\left(\frac{S}{K}\right)\right)\right]=-K\sigma^2\frac{\partial u}{\partial\tau}\end{aligned} \tag{15.3.1030}$$

and:

$$\begin{aligned}\frac{\partial C}{\partial S}&=\frac{\partial}{\partial S}\left[K\cdot u\left(\sigma^2(T-t),\ln\left(\frac{S}{K}\right)\right)\right]=K\frac{\partial}{\partial t}\left[u\left(\sigma^2(T-t),\ln\left(\frac{S}{K}\right)\right)\right]\\&=K\frac{\partial}{S\partial x}\left[u\left(\sigma^2(T-t),\left(\frac{S}{K}\right)\right)\right]=\frac{K}{S}\frac{\partial u}{\partial x}\end{aligned} \tag{15.3.1031}$$

And also that:

$$\begin{aligned}\frac{\partial^2C}{\partial S^2}&=\frac{\partial}{\partial S}\left(\frac{K}{S}\frac{\partial u}{\partial x}\right)=-\frac{K}{S^2}\frac{\partial u}{\partial x}+\frac{K}{S}\left[\frac{\partial}{\partial S}u_x\left(\sigma^2(T-t),\left(\frac{S}{K}\right)\right)\right]\\&=-\frac{K}{S^2}\frac{\partial u}{\partial x}+\frac{K}{S}\frac{\partial}{\partial S}u_x=-\frac{K}{S^2}\frac{\partial u}{\partial x}+\frac{K}{S}\frac{\partial}{S\partial x}u_x=-\frac{K}{S^2}\frac{\partial u}{\partial x}+\frac{K}{S^2}\frac{\partial}{\partial x}u_x=-\frac{K}{S^2}\frac{\partial u}{\partial x}+\frac{K}{S^2}\frac{\partial^2u}{\partial^2x}\\&=\frac{K}{S^2}\left(\frac{\partial^2u}{\partial^2x}-\frac{\partial u}{\partial x}\right)\end{aligned} \tag{15.3.1032}$$

By injecting this into the differential equation, it comes:

$$\left[-K\frac{\partial u}{\partial \tau}\sigma^2\right]+\frac{1}{2}\sigma^2 S^2\left[\frac{K}{S^2}\left(\frac{\partial^2 u}{\partial x^2}-\frac{\partial u}{\partial x}\right)\right]+rS\left[\frac{K}{S}\frac{\partial u}{\partial x}\right]-r\left[Ku\right]=0 \qquad (15.3.1033)$$

We simplify the S:

$$-K\frac{\partial u}{\partial \tau}\sigma^2+\frac{1}{2}K\left(\frac{\partial^2 u}{\partial x^2}-\frac{\partial u}{\partial x}\right)\sigma^2+rK\frac{\partial u}{\partial x}-rKu=0 \qquad (15.3.1034)$$

We multiply on both sides by:

$$-\frac{1}{K\sigma^2} \qquad (15.3.1035)$$

To get:

$$\frac{\partial u}{\partial \tau}-\frac{1}{2}\left(\frac{\partial^2 u}{\partial x^2}-\frac{\partial u}{\partial x}\right)-\frac{r}{\sigma^2}\frac{\partial u}{\partial x}+\frac{r}{\sigma^2}u=0 \qquad (15.3.1036)$$

Therefore:

$$\frac{\partial u}{\partial \tau}=\frac{1}{2}\frac{\partial^2 u}{\partial x^2}+\left(\frac{r}{\sigma^2}-\frac{1}{2}\right)\frac{\partial u}{\partial x}-\frac{r}{\sigma^2}u \qquad (15.3.1037)$$

Let us put as it is usual:

$$\frac{r}{\sigma^2}=\kappa \qquad (15.3.1038)$$

Then there comes the differential equation (with constant coefficient this time!):

$$\frac{\partial u}{\partial \tau}=\frac{1}{2}\frac{\partial^2 u}{\partial x^2}+\left(\kappa-\frac{1}{2}\right)\frac{\partial u}{\partial x}-\kappa u \qquad (15.3.1039)$$

With for recall the initial condition:

$$u(0,x)=\left(e^x-1\right)^+ \qquad (15.3.1040)$$

Remark

In the prior-previous relation, the sum $\dfrac{\partial u}{\partial \tau}+\left(\dfrac{1}{2}-\kappa\right)\dfrac{\partial u}{\partial x}$ is sometimes named the "advection part", κu the "source term" and $\dfrac{1}{2}\dfrac{\partial^2 u}{\partial x^2}$ the "diffusion term", all this in analogy to fluid mechanics and thermodynamics.

What follows also very clever and tricky consists in making another change of variable by putting:

$$\nu(\tau,x)=e^{\alpha\tau+\beta x}\cdot u(\tau,x) \qquad (15.3.1041)$$

We will determine the two parameters α, β which allow to write the partial differential equation with constant coefficients in the form of a heat equation!

We shall proceed first as above by determining the equivalence of the different partial derivatives of u. We have then:

$$\frac{\partial u}{\partial \tau}=\frac{\partial\left(e^{-\alpha\tau-\beta x}\cdot\nu(\tau,x)\right)}{\partial \tau}=\left(-\alpha\nu+\frac{\partial \nu}{\partial \tau}\right)e^{-\alpha\tau-\beta x} \tag{15.3.1042}$$

and:

$$\frac{\partial u}{\partial x}=\frac{\partial\left(e^{-\alpha\tau-\beta x}\cdot\nu(\tau,x)\right)}{\partial x}=\left(-\beta\nu+\frac{\partial \nu}{\partial x}\right)e^{-\alpha\tau-\beta x} \tag{15.3.1043}$$

And also that:

$$\begin{aligned}\frac{\partial^2 u}{\partial x^2}&=\frac{\partial}{\partial x}\left(-\beta\nu e^{-\alpha\tau-\beta x}+\frac{\mathrm{d}\nu}{\mathrm{d}x}e^{-\alpha\tau-\beta x}\right)=\left[-\beta\frac{\partial \nu}{\partial x}+\beta^2\nu+\frac{\partial^2 \nu}{\partial x^2}-\beta\frac{\partial \nu}{\partial x}\right]e^{-\alpha\tau-\beta x}\\&=\left[\frac{\partial^2 \nu}{\partial x^2}-2\beta\frac{\partial \nu}{\partial x}+\beta^2\nu\right]e^{-\alpha\tau-\beta x}\end{aligned} \tag{15.3.1044}$$

Then comes a trick very difficult to guess (at least a priori because I did not personally find something simpler...). Indeed before substituting these last three relations in the P.D.E., Let us recall that:

$$\nu(\tau,x)=e^{\alpha\tau+\beta x}\cdot u(\tau,x) \tag{15.3.1045}$$

So the last three relations can be written (finding a mixture function of u and v is very counter-intuitive):

$$\begin{aligned}\frac{\partial u}{\partial \tau}&=\left(-\alpha\nu+\frac{\partial \nu}{\partial \tau}\right)e^{-\alpha\tau-\beta x}=-\alpha u+\frac{\partial \nu}{\partial \tau}^{-\alpha\tau-\beta x}\\\frac{\partial u}{\partial x}&=\left(-\beta\nu+\frac{\partial \nu}{\partial x}\right)e^{-\alpha\tau-\beta x}=-\beta u+\frac{\partial \nu}{\partial x}e^{-\alpha\tau-\beta x}\\\frac{\partial^2 u}{\partial x^2}&=\left[\frac{\partial^2 \nu}{\partial x^2}-2\beta\frac{\partial \nu}{\partial x}+\beta^2\nu\right]e^{-\alpha\tau-\beta x}=\frac{\partial^2 \nu}{\partial x^2}e^{-\alpha\tau-\beta x}-2\beta\frac{\partial \nu}{\partial x}e^{-\alpha\tau-\beta x}+\beta^2 u\end{aligned} \tag{15.3.1046}$$

and now only let us do the (partial!) substitution in the P.D.E.! We then move from:

$$\frac{\partial u}{\partial \tau}=\frac{1}{2}\frac{\partial^2 u}{\partial x^2}+\left(\kappa-\frac{1}{2}\right)\frac{\partial u}{\partial x}-\kappa u \tag{15.3.1047}$$

to:

$$\begin{aligned}&\left[-\alpha u+e^{-\alpha\tau-\beta x}\frac{\partial \nu}{\partial \tau}\right]\\&=\frac{1}{2}\left[\beta^2 u-2\beta e^{-\alpha\tau-\beta x}\frac{\partial \nu}{\partial x}+e^{-\alpha\tau-\beta x}\frac{\partial^2 \nu}{\partial x^2}\right]+\left(\kappa-\frac{1}{2}\right)\left[-\beta u+e^{-\alpha\tau-\beta x}\frac{\partial \nu}{\partial x}\right]-\kappa u\end{aligned} \tag{15.3.1048}$$

Let us write this last equality a little differently by factoring on the left and right what contains u:

$$\begin{gathered}\left[-\alpha\right]u+e^{-\alpha\tau-\beta x}\frac{\partial\nu}{\partial\tau}=\\ \left[\frac{1}{2}\beta^2-\beta\left(\kappa-\frac{1}{2}\right)-\kappa\right]u-\frac{1}{2}\left[2\beta e^{-\alpha\tau-\beta x}\frac{\partial\nu}{\partial x}-e^{-\alpha\tau-\beta x}\frac{\partial^2\nu}{\partial x^2}\right]+\left(\kappa-\frac{1}{2}\right)e^{-\alpha\tau-\beta x}\frac{\partial\nu}{\partial x}\end{gathered}\tag{15.3.1049}$$

and then let:

$$-\alpha=\left(\frac{1}{2}\beta^2-\beta\left(\kappa-\frac{1}{2}\right)-\kappa\right)\tag{15.3.1050}$$

The differential equation is then simplified as:

$$e^{-\alpha\tau-\beta x}\frac{\partial\nu}{\partial\tau}=-\frac{1}{2}\left[2\beta e^{-\alpha\tau-\beta x}\frac{\partial\nu}{\partial x}-e^{-\alpha\tau-\beta x}\frac{\partial^2\nu}{\partial x^2}\right]+\left(\kappa-\frac{1}{2}\right)e^{-\alpha\tau-\beta x}\frac{\partial\nu}{\partial x}\tag{15.3.1051}$$

The exponential can be simplified and it remains:

$$\frac{\partial\nu}{\partial\tau}=-\frac{1}{2}\left[2\beta\frac{\partial\nu}{\partial x}-\frac{\partial^2\nu}{\partial x^2}\right]+\left(\kappa-\frac{1}{2}\right)\frac{\partial\nu}{\partial x}\tag{15.3.1052}$$

That can be rewritten in the form:

$$\frac{\partial\nu}{\partial\tau}=\left(\left(\kappa-\frac{1}{2}\right)-\beta\right)\frac{\partial\nu}{\partial x}+\frac{1}{2}\frac{\partial^2\nu}{\partial x^2}\tag{15.3.1053}$$

And if we choose:

$$\left(\kappa-\frac{1}{2}\right)=\beta\tag{15.3.1054}$$

We then have:

$$\frac{\partial\nu}{\partial\tau}=\frac{1}{2}\frac{\partial^2\nu}{\partial x^2}\tag{15.3.1055}$$

Or otherwise written (in the physicist way of life...):

$$\nu_t=\frac{1}{2}\nu_{xx}\tag{15.3.1056}$$

We thus recognize the heat equation (or "diffusion equation") proved in the section Thermodynamics but where the coefficient of thermal expansion is equal to $1/2$.

It follows from these developments that:

$$\begin{gathered}-\alpha=\left(\frac{1}{2}\beta^2-\beta\left(\kappa-\frac{1}{2}\right)-\kappa\right)\\ \beta=\left(\kappa-\frac{1}{2}\right)\end{gathered}\tag{15.3.1057}$$

Either by injecting the second equality into the first and after some elementary simplifications:

$$\begin{aligned}\alpha &= -\frac{1}{2}\left[\kappa-\frac{1}{2}\right]^2+\left[\kappa-\frac{1}{2}\right]\left(\kappa-\frac{1}{2}\right)+\kappa=\frac{1}{2}\left(\kappa-\frac{1}{2}\right)^2+\kappa=\frac{1}{2}\left(\kappa^2-\kappa+\frac{1}{4}\right)+\kappa\\ &=\frac{1}{2}\kappa^2-\frac{1}{2}\kappa+\frac{1}{8}+\kappa=\frac{1}{2}\kappa^2+\frac{1}{2}\kappa+\frac{1}{8}=\frac{1}{2}\left(\kappa+\frac{1}{2}\right)^2\end{aligned} \tag{15.3.1058}$$

So we see that finally:

$$e^{-\alpha\tau-\beta x}=e^{-\frac{1}{2}(\kappa+\frac{1}{2})^2\tau-(\kappa-\frac{1}{2})x} \tag{15.3.1059}$$

Now we have to transform the condition into u:

$$u(0,x)=\left(e^x-1\right)^+ \tag{15.3.1060}$$

In an initial condition in v using:

$$\nu(\tau,x)=e^{\frac{1}{2}(\kappa+\frac{1}{2})^2\tau+(\kappa-\frac{1}{2})x}u(\tau,x) \tag{15.3.1061}$$

So it comes immediately:

$$\begin{aligned}\nu(0,x)&=e^{\frac{1}{2}(\kappa+\frac{1}{2})^2\cdot 0+(\kappa-\frac{1}{2})x}u(0,x)=e^{(\kappa-\frac{1}{2})x}\cdot\left(e^x-1\right)^+=\left(e^{(\kappa-\frac{1}{2})x+x}-e^{(\kappa-\frac{1}{2})x}\right)^+\\ &=\left(e^{(\kappa+\frac{1}{2})x}-e^{(\kappa-\frac{1}{2})x}\right)^+\\ &=\left(e^{\left(\kappa-\frac{1}{2}\right)x+x}-e^{\left(\kappa-\frac{1}{2}\right)x}\right)^+\\ &=e^{\left(\kappa-\frac{1}{2}\right)x}\cdot(e^x-1)^+\end{aligned} \tag{15.3.1062}$$

We now have a nice, clear problem that we know how to solve ("Cauchy problem" type):

$$\begin{aligned}\nu_t&=\frac{1}{2}\nu_{xx}\\ \nu(0,x)&=\left(e^{(\kappa+\frac{1}{2})x}-e^{(\kappa-\frac{1}{2})x}\right)^+\end{aligned} \tag{15.3.1063}$$

And as we saw in the discussion about the Heat Equation / Diffusion Equation (see section Thermodynamics page 3684), we solve this problem using Fourier transform that leads using exactly the same development to:

$$\nu(\tau,x)=\int_{-\infty}^{+\infty}\nu(0,x)\frac{1}{\sqrt{2\pi\tau}}e^{-\frac{(y-x)^2}{2\tau}}\mathrm{d}y \tag{15.3.1064}$$

Hence:

$$\nu(\tau,x)=\int_{-\infty}^{+\infty}\left(e^{\left(\kappa+\frac{1}{2}\right)y}-e^{\left(\kappa-\frac{1}{2}\right)y}\right)^+\frac{1}{\sqrt{2\pi\tau}}e^{-\frac{(y-x)^2}{2\tau}}\mathrm{d}y \tag{15.3.1065}$$

Let's examine:

$$\left(e^{\left(\kappa+\frac{1}{2}\right)y}-e^{\left(\kappa-\frac{1}{2}\right)y}\right)^+ \tag{15.3.1066}$$

When:

$$e^{\left(\kappa+\frac{1}{2}\right)y}-e^{\left(\kappa-\frac{1}{2}\right)y}\geq 0 \tag{15.3.1067}$$

then:

$$\left(e^{\left(\kappa+\frac{1}{2}\right)y}-e^{\left(\kappa-\frac{1}{2}\right)y}\right)^{+}=\left(e^{\left(\kappa+\frac{1}{2}\right)y}-e^{\left(\kappa-\frac{1}{2}\right)y}\right) \tag{15.3.1068}$$

But when:

$$e^{\left(\kappa+\frac{1}{2}\right)y}-e^{\left(\kappa-\frac{1}{2}\right)y}\leq 0 \tag{15.3.1069}$$

then:

$$\left(e^{\left(\kappa+\frac{1}{2}\right)y}-e^{\left(\kappa-\frac{1}{2}\right)y}\right)^{+}=0 \tag{15.3.1070}$$

Let's find the values of y such that

$$\begin{aligned}&e^{\left(\kappa+\frac{1}{2}\right)y}-e^{\left(\kappa-\frac{1}{2}\right)y}\geq 0\\&e^{\left(\kappa+\frac{1}{2}\right)y}\geq e^{\left(\kappa-\frac{1}{2}\right)y}\\&\left(\kappa+\frac{1}{2}\right)y\geq\left(\kappa-\frac{1}{2}\right)y\end{aligned} \tag{15.3.1071}$$

Therefore:

$$y\geq 0 \tag{15.3.1072}$$

So if we only integrate over $y>0$ then we can eliminate the function because the integrand is zero when $y<0$.

The old integral:

$$\int\limits_{-\infty}^{+\infty}\left(e^{\left(\kappa+\frac{1}{2}\right)y}-e^{\left(\kappa-\frac{1}{2}\right)y}\right)^{+}\frac{1}{\sqrt{2\pi\tau}}e^{-\frac{(y-x)^2}{2\tau}}\mathrm{d}y \tag{15.3.1073}$$

Then becomes:

$$\int\limits_{0}^{+\infty}\left(e^{\left(\kappa+\frac{1}{2}\right)y}-e^{\left(\kappa-\frac{1}{2}\right)y}\right)^{+}\frac{1}{\sqrt{2\pi\tau}}e^{-\frac{(y-x)^2}{2\tau}}\mathrm{d}y \tag{15.3.1074}$$

Next, we change variables again. Let:

$$z=\frac{y-x}{\sqrt{\tau}} \tag{15.3.1075}$$

so that:

$$z^2=\frac{(y-x)^2}{\tau} \tag{15.3.1076}$$

Now since:

$$z=\frac{y-x}{\sqrt{\tau}} \tag{15.3.1077}$$

Then:

$$y=x+\sqrt{\tau}z \tag{15.3.1078}$$

so that:

$$\mathrm{d}y=\sqrt{\tau}\mathrm{d}z \tag{15.3.1079}$$

Finally, y varies $[0,+\infty]$ and:

$$z=\frac{y-x}{\sqrt{\tau}} \tag{15.3.1080}$$

So when $y=+\infty$ we have $z=+\infty$. And when $y=0$, we have:

$$z=\frac{-x}{\sqrt{\tau}} \tag{15.3.1081}$$

So z varies between

$$z\in\left[-\frac{x}{\sqrt{\tau}},+\infty\right[\tag{15.3.1082}$$

Remember that when we change the variables in an integration we also must change the bounds of integration. So let's put all these changes into our integration!

So from the old integral:

$$\int\limits_0^{+\infty}\left(e^{\left(\kappa+\frac{1}{2}\right)y}-e^{\left(\kappa-\frac{1}{2}\right)y}\right)^+\frac{1}{\sqrt{2\pi\tau}}e^{-\frac{(y-x)^2}{2\tau}}\mathrm{d}y \tag{15.3.1083}$$

We get a new integral:

$$\int\limits_{-x/\sqrt{\tau}}^{+\infty}\left(e^{\left(\kappa+\frac{1}{2}\right)\left(x+\sqrt{\tau}z\right)}-e^{\left(\kappa-\frac{1}{2}\right)\left(x+\sqrt{\tau}z\right)}\right)^+\frac{1}{\sqrt{2\pi\tau}}e^{-\frac{z^2}{2}}\mathrm{d}z\sqrt{\tau} \tag{15.3.1084}$$

Which simplifies to:

$$\frac{1}{\sqrt{2\pi}}\int\limits_{\frac{-x}{\sqrt{\tau}}}^{+\infty}\left(e^{\left(\kappa+\frac{1}{2}\right)\left(x+\sqrt{\tau}z\right)}-e^{\left(\kappa-\frac{1}{2}\right)\left(x+\sqrt{\tau}z\right)}\right)^+e^{-\frac{z^2}{2}}\mathrm{d}z \tag{15.3.1085}$$

This is a real mess, so let's split the problem into two integrals.

We go from the old problem:

$$\nu(\tau,x)=\frac{1}{\sqrt{2\pi}}\int\limits_{\frac{-x}{\sqrt{\tau}}}^{+\infty}\left(e^{\left(\kappa+\frac{1}{2}\right)\left(x+\sqrt{\tau}z\right)}-e^{\left(\kappa-\frac{1}{2}\right)\left(x+\sqrt{\tau}z\right)}\right)^+e^{-\frac{z^2}{2}}\mathrm{d}z \tag{15.3.1086}$$

To a new one by writing:

$$\nu(\tau,x)=I_1-I_2 \tag{15.3.1087}$$

With:

$$I_1=\frac{1}{\sqrt{2\pi}}\int\limits_{\frac{-x}{\sqrt{\tau}}}^{+\infty}e^{\left(\kappa+\frac{1}{2}\right)\left(x+\sqrt{\tau}z\right)}e^{-\frac{z^2}{2}}\mathrm{d}z \tag{15.3.1088}$$

and:

$$I_2=\frac{1}{\sqrt{2\pi}}\int\limits_{\frac{-x}{\sqrt{\tau}}}^{+\infty}e^{\left(\kappa-\frac{1}{2}\right)\left(x+\sqrt{\tau}z\right)}e^{-\frac{z^2}{2}}\mathrm{d}z \tag{15.3.1089}$$

Let us solve first:

$$I_1=\frac{1}{\sqrt{2\pi}}\int\limits_{\frac{-x}{\sqrt{\tau}}}^{+\infty}e^{\left(\kappa+\frac{1}{2}\right)\left(x+\sqrt{\tau}z\right)}e^{-\frac{z^2}{2}}\mathrm{d}z \tag{15.3.1090}$$

Let's work directly on the exponent first:

$$\begin{aligned}
&\left(\kappa+\frac{1}{2}\right)\left(x+\sqrt{\tau}z\right)-\frac{1}{2}z^2\\
&\left(\kappa+\frac{1}{2}\right)x+\left(\kappa+\frac{1}{2}\right)\sqrt{\tau}z-\frac{1}{2}z^2\\
&\left(\kappa+\frac{1}{2}\right)x-\frac{1}{2}\left[z^2-2\left(\kappa+\frac{1}{2}\right)\sqrt{\tau}z\right]\\
&\left(\kappa+\frac{1}{2}\right)x-\frac{1}{2}\left[z^2-2\left(\kappa+\frac{1}{2}\right)\sqrt{\tau}z-\left(\kappa+\frac{1}{2}\right)^2\tau\right]+\frac{1}{2}\left(\kappa+\frac{1}{2}\right)^2\tau\\
&\left(\kappa+\frac{1}{2}\right)x-\frac{1}{2}\left[z-\left(\kappa+\frac{1}{2}\right)\sqrt{\tau}\right]^2+\frac{1}{2}\left(\kappa+\frac{1}{2}\right)^2\tau
\end{aligned} \tag{15.3.1091}$$

This means we can re-write I_1 under the following form:

$$I_1=\frac{1}{\sqrt{2\pi}}\int\limits_{\frac{-x}{\sqrt{\tau}}}^{+\infty}e^{\left(\kappa+\frac{1}{2}\right)\left(x+\sqrt{\tau}z\right)}e^{-\frac{z^2}{2}}\mathrm{d}z \tag{15.3.1092}$$

In an equivalent new integral:

$$\begin{aligned}
I_1&=\frac{1}{\sqrt{2\pi}}\int\limits_{\frac{-x}{\sqrt{\tau}}}^{+\infty}e^{\left(\kappa+\frac{1}{2}\right)x-\frac{1}{2}\left[z-\left(\kappa+\frac{1}{2}\right)\sqrt{\tau}\right]^2+\frac{1}{2}\left(\kappa+\frac{1}{2}\right)^2\tau}\mathrm{d}z\\
&=e^{\left(\kappa+\frac{1}{2}\right)x+\frac{1}{2}\left(\kappa+\frac{1}{2}\right)^2\tau}\cdot\frac{1}{\sqrt{2\pi}}\int\limits_{\frac{-x}{\sqrt{\tau}}}^{+\infty}e^{-\frac{1}{2}\left[z-\left(\kappa+\frac{1}{2}\right)\sqrt{\tau}\right]^2}\mathrm{d}z
\end{aligned} \tag{15.3.1093}$$

Now let us recall that in statistics we often use the following formula for the cumulative density function of the Normal centered reduced random variable:

$$\Phi(Q):=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{Q} e^{-\frac{1}{2}z^2}\mathrm{d}z \tag{15.3.1094}$$

In our case, what we have is more similar to:

$$\frac{1}{\sqrt{2\pi}}\int\limits_{-Q}^{+\infty} e^{-\frac{1}{2}(z-\mu)^2}\mathrm{d}z \tag{15.3.1095}$$

where:

$$Q=\frac{x}{\sqrt{\tau}} \qquad \text{and} \qquad \mu=\left(\kappa+\frac{1}{2}\right)\sqrt{\tau} \tag{15.3.1096}$$

Let us make this change of variables:

$$-w=(z-\mu) \tag{15.3.1097}$$

Therefore:

$$-\mathrm{d}w=\mathrm{d}z \tag{15.3.1098}$$

As $-w=(z-\mu)$, we have then when $z=+\infty$:

$$-w=(+\infty-\mu)\Rightarrow w=-\infty \tag{15.3.1099}$$

And when $z=-Q$, we have then:

$$-w=(-Q-\mu)\Rightarrow \text{so} \qquad w=Q+\mu \tag{15.3.1100}$$

Making this change of variables $-w=(z-\mu)$ gives us:

$$\begin{aligned}\int\limits_{-Q}^{+\infty} e^{-\frac{1}{2}(z-\mu)^2}\mathrm{d}z &\rightarrow \int\limits_{Q+\mu}^{-\infty} e^{-\frac{1}{2}[-w]^2}(-\mathrm{d}w)=-\int\limits_{Q+\mu}^{-\infty} e^{-\frac{1}{2}w^2}\mathrm{d}w\\ &=\int\limits_{-\infty}^{Q+\mu} e^{-\frac{1}{2}w^2}\mathrm{d}w=\Phi(Q+\mu)\end{aligned} \tag{15.3.1101}$$

Therefore:

$$\begin{aligned}I_1&=e^{\left(\kappa+\frac{1}{2}\right)x+\frac{1}{2}\left(\kappa+\frac{1}{2}\right)^2\tau}\cdot\frac{1}{\sqrt{2\pi}}\int\limits_{\frac{-x}{\sqrt{\tau}}}^{+\infty} e^{-\frac{1}{2}\left[z-\left(\kappa+\frac{1}{2}\right)\sqrt{\tau}\right]^2}\mathrm{d}z\\ &=e^{\left(\kappa+\frac{1}{2}\right)x+\frac{1}{2}\left(\kappa+\frac{1}{2}\right)^2\tau}\cdot\Phi\left(\frac{x}{\sqrt{\tau}}+\left(\kappa+\frac{1}{2}\right)\sqrt{\tau}\right)\end{aligned} \tag{15.3.1102}$$

Finally, let us put:

$$\frac{x}{\sqrt{\tau}}+\left(\kappa+\frac{1}{2}\right)\sqrt{\tau}=d_1 \tag{15.3.1103}$$

Then:

$$I_1=e^{\frac{1}{2}\left(\kappa+\frac{1}{2}\right)^2\tau+\left(\kappa+\frac{1}{2}\right)x}\cdot\Phi\left(d_1\right) \tag{15.3.1104}$$

Now we must focus on the second integral, that is (for recall):

$$I_2=\frac{1}{\sqrt{2\pi}}\int\limits_{\frac{-x}{\sqrt{\tau}}}^{+\infty}e^{\left(\kappa-\frac{1}{2}\right)\left(x+\sqrt{\tau}z\right)}e^{-\frac{z^2}{2}}\mathrm{d}z \tag{15.3.1105}$$

We work directly on the exponent first:

$$\begin{gathered}\left(\kappa-\frac{1}{2}\right)\left(x+\sqrt{\tau}z\right)-\frac{1}{2}z^2\\ \left(\kappa-\frac{1}{2}\right)x+\left(\kappa-\frac{1}{2}\right)\sqrt{\tau}z-\frac{1}{2}z^2\\ \left(\kappa-\frac{1}{2}\right)x-\frac{1}{2}\left[z^2-2\left(\kappa-\frac{1}{2}\right)\sqrt{\tau}z\right]\\ \left(\kappa-\frac{1}{2}\right)x-\frac{1}{2}\left[z^2-2\left(\kappa-\frac{1}{2}\right)\sqrt{\tau}z+\left(\kappa-\frac{1}{2}\right)^2\tau\right]+\frac{1}{2}\left(\kappa-\frac{1}{2}\right)^2\tau\\ \left(\kappa-\frac{1}{2}\right)x-\frac{1}{2}\left[z-\left(\kappa-\frac{1}{2}\right)\sqrt{\tau}\right]^2+\frac{1}{2}\left(\kappa-\frac{1}{2}\right)^2\tau\end{gathered} \tag{15.3.1106}$$

So we can re-write I_2:

$$\begin{aligned}I_2&=\frac{1}{\sqrt{2\pi}}\int\limits_{\frac{-x}{\sqrt{\tau}}}^{+\infty}e^{\left(\kappa-\frac{1}{2}\right)\left(x+\sqrt{\tau}z\right)}e^{-\frac{z^2}{2}}\mathrm{d}z\\ &=\frac{1}{\sqrt{2\pi}}\int\limits_{\frac{-x}{\sqrt{\tau}}}^{+\infty}e^{\left(\kappa-\frac{1}{2}\right)x-\frac{1}{2}\left[z-\left(\kappa-\frac{1}{2}\right)\sqrt{\tau}\right]^2+\frac{1}{2}\left(\kappa-\frac{1}{2}\right)^2\tau}\mathrm{d}z\\ &=e^{\frac{1}{2}\left(\kappa-\frac{1}{2}\right)^2\tau+\left(\kappa-\frac{1}{2}\right)x}\cdot\frac{1}{\sqrt{2\pi}}\int\limits_{\frac{-x}{\sqrt{\tau}}}^{+\infty}e^{-\frac{1}{2}\left[z-\left(\kappa-\frac{1}{2}\right)\sqrt{\tau}\right]^2}\mathrm{d}z\\ &=e^{\frac{1}{2}\left(\kappa-\frac{1}{2}\right)^2\tau+\left(\kappa-\frac{1}{2}\right)x}\cdot\Phi\left(\frac{x}{\sqrt{\tau}}+\left(\kappa-\frac{1}{2}\right)\sqrt{\tau}\right)\end{aligned} \tag{15.3.1107}$$

Set:

$$d_2=\frac{x}{\sqrt{\tau}}+\left(\kappa-\frac{1}{2}\right)\sqrt{\tau} \tag{15.3.1108}$$

We therefore have:

$$I_2 = e^{\frac{1}{2}\left(\kappa-\frac{1}{2}\right)^2\tau+\left(\kappa-\frac{1}{2}\right)x} \cdot \Phi\left(d_2\right) \tag{15.3.1109}$$

Now we put together I_1 and I_2:

$$\nu(\tau,x) = I_1 - I_2 \tag{15.3.1110}$$

so we can see our final solution to:

$$\nu(\tau,x) = e^{\frac{1}{2}\left(\kappa+\frac{1}{2}\right)^2\tau+\left(\kappa+\frac{1}{2}\right)x} \cdot \Phi\left(d_1\right) - e^{\frac{1}{2}\left(\kappa-\frac{1}{2}\right)^2\tau+\left(\kappa-\frac{1}{2}\right)x} \cdot \Phi\left(d_2\right) \tag{15.3.1111}$$

This is for recall the solution to:

$$\begin{aligned} \nu_t &= \frac{1}{2}\nu_{xx} \\ \nu(0,x) &= \left(e^{\left(\kappa+\frac{1}{2}\right)x} - e^{\left(\kappa-\frac{1}{2}\right)x}\right)^+ \end{aligned} \tag{15.3.1112}$$

Now that we have a solution for ν we need to reverse all the transforms we did to get back to a solution for $C(t,S)$.

The first transform to reverse is:

$$u(\tau,x) = e^{-\alpha\tau-\beta x} \cdot \nu(\tau,x \tag{15.3.1113}$$

Thus explicitly:

$$\begin{aligned} u(\tau,x) &= e^{-\frac{1}{2}\left(\kappa+\frac{1}{2}\right)^2\tau+\left(\kappa-\frac{1}{2}\right)x}\left(e^{\frac{1}{2}\left(\kappa+\frac{1}{2}\right)^2\tau-\left(\kappa+\frac{1}{2}\right)x} \cdot \Phi\left(d_1\right) - e^{\frac{1}{2}\left(\kappa-\frac{1}{2}\right)^2\tau+\left(\kappa-\frac{1}{2}\right)x} \cdot \Phi\left(d_2\right)\right) \\ &= e^{-\left(\kappa-\frac{1}{2}\right)x+\left(\kappa+\frac{1}{2}\right)x} \cdot \Phi\left(d_1\right) - e^{-\frac{1}{2}\left(\kappa+\frac{1}{2}\right)^2\tau+\frac{1}{2}\left(\kappa-\frac{1}{2}\right)^2\tau} \cdot \Phi\left(d_2\right) \\ &= e^x \cdot \Phi\left(d_1\right) - e^{-\frac{1}{2}\left(\kappa+\frac{1}{2}\right)^2\tau+\frac{1}{2}\left(\kappa-\frac{1}{2}\right)^2\tau} \cdot \Phi\left(d_2\right) \end{aligned} \tag{15.3.1114}$$

The second exponent is a little bit more complicated to simplify:

$$\begin{aligned} &\frac{1}{2}\tau\left[-\left(\kappa+\frac{1}{2}\right)^2\tau+\left(\kappa-\frac{1}{2}\right)^2\right] \\ &\frac{1}{2}\tau\left[-\left(\kappa^2+\kappa+\frac{1}{4}\right)+\left(\kappa^2-\kappa+\frac{1}{4}\right)\right] \\ &\frac{1}{2}\tau\left[-2\kappa\right] = -\tau\kappa \end{aligned} \tag{15.3.1115}$$

After those simplifications we are left with:

$$u(\tau,x) = e^x\Phi(d_1) - e^{-\tau\kappa}\Phi(d_2) \tag{15.3.1116}$$

The first transform and change of variables we did was:

$$C(t,S) = K \cdot u(\tau,x) \tag{15.3.1117}$$

with:

$$\tau=\sigma^2(T-t) \qquad \text{and} \qquad x=\ln\left(\frac{S}{K}\right) \tag{15.3.1118}$$

Also recall that we later set $\kappa=\dfrac{r}{\sigma^2}$ to make our calculations easier!

Now let's reverse all of those starting from:

$$u(\tau,x)=e^x\Phi(d_1)-e^{-\tau\kappa}\Phi(d_2) \tag{15.3.1119}$$

Therefore:

$$C(t,S)=K\cdot e^{\ln\left(\frac{S}{K}\right)}\Phi(d_1)-Ke^{-\sigma^2(T-t)\frac{r}{\sigma^2}}\Phi(d_2)=S\Phi(d_1)-Ke^{-r(T-t)}\Phi(d_2) \tag{15.3.1120}$$

with:

$$d_1=\frac{x}{\sqrt{\tau}}+\left(\kappa+\frac{1}{2}\right)\sqrt{\tau}\rightarrow d_1=\frac{\ln\left(\dfrac{S}{K}\right)}{\sqrt{\sigma^2(T-t)}}+\left(\frac{r}{\sigma^2}+\frac{1}{2}\right)\sqrt{\sigma^2(T-t)} \tag{15.3.1121}$$

and:

$$d_2=\frac{x}{\sqrt{\tau}}+\left(\kappa-\frac{1}{2}\right)\sqrt{\tau}\rightarrow d_2=\frac{\ln\left(\dfrac{S}{K}\right)}{\sqrt{\sigma^2(T-t)}}+\left(\frac{r}{\sigma^2}-\frac{1}{2}\right)\sqrt{\sigma^2(T-t)} \tag{15.3.1122}$$

Let's simplify these:

$$\begin{aligned}d_1&=\frac{\ln\left(\dfrac{S}{K}\right)}{\sqrt{\sigma^2(T-t)}}+\left(\frac{r}{\sigma^2}+\frac{1}{2}\right)\sqrt{\sigma^2(T-t)}\cdot\frac{\sqrt{\sigma^2(T-t)}}{\sqrt{\sigma^2(T-t)}}\\&=\frac{\ln\left(\dfrac{S}{K}\right)+\left(\dfrac{r}{\sigma^2}+\dfrac{1}{2}\right)\sigma^2(T-t)}{\sqrt{\sigma^2(T-t)}}\\&=\frac{\ln\left(\dfrac{S}{K}\right)+\left(r+\dfrac{1}{2}\sigma^2\right)(T-t)}{\sigma\sqrt{(T-t)}}\end{aligned} \tag{15.3.1123}$$

and:

$$\begin{aligned}d_2&=\frac{\ln\left(\dfrac{S}{K}\right)}{\sqrt{\sigma^2(T-t)}}+\left(\frac{r}{\sigma^2}-\frac{1}{2}\right)\sqrt{\sigma^2(T-t)}\cdot\frac{\sqrt{\sigma^2(T-t)}}{\sqrt{\sigma^2(T-t)}}\\&=\frac{\ln\left(\dfrac{S}{K}\right)+\left(r-\dfrac{1}{2}\sigma^2\right)(T-t)}{\sigma\sqrt{(T-t)}}\end{aligned} \tag{15.3.1124}$$

Note that we also have:

$$\begin{aligned}d_1-\sigma\sqrt{(T-t)}&=\frac{\ln\left(\dfrac{S}{K}\right)+\left(r+\dfrac{1}{2}\sigma^2\right)(T-t)}{\sigma\sqrt{(T-t)}}-\sigma\sqrt{(T-t)}\cdot\frac{\sigma\sqrt{(T-t)}}{\sigma\sqrt{(T-t)}}\\&=\frac{\ln\left(\dfrac{S}{K}\right)+\left(r+\dfrac{1}{2}\sigma^2\right)(T-t)-\sigma^2(T-t)}{\sigma\sqrt{(T-t)}}\\&=\frac{\ln\left(\dfrac{S}{K}\right)+\left(r-\dfrac{1}{2}\sigma^2\right)(T-t)}{\sigma\sqrt{(T-t)}}=d_2\end{aligned}\tag{15.3.1125}$$

Hence $d_2=d_1-\sigma\sqrt{(T-t)}$.

So finally to summarize:

- For the European Call (the value of the purchase option with maturity T and strike K), the solution (whose demonstration still needs to be written in this section) is for a non-dividend-paying share-based underlying and for a geometric Brownian motion:

$$\boxed{C^{\text{BS}}(S,t)=S\Phi(d_1)-Ke^{-r(T-t)}\Phi(d_2)}\tag{15.3.1126}$$

where for recall (see section Statistics page 692):

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{x}e^{-\frac{1}{2}y^2}\mathrm{d}y\tag{15.3.1127}$$

The centered reduced Normal cumulative distribution function with:

$$d_1=\frac{\ln\left(\dfrac{S}{K}\right)+\left(r+\dfrac{1}{2}\sigma^2\right)(T-t)}{\sigma\sqrt{T-t}}\tag{15.3.1128}$$

and:

$$d_2=d_1-\sigma\sqrt{T-t}\tag{15.3.1129}$$

where to reader has to remember that μ,σ are both expressed in percentages!

Some textbooks write the Black-Scholes European Call pricing equation explicitly in the following way (a quite nice and beautiful way to write it!):

$$\begin{aligned}C_0^{\text{BS}} &= \mathrm{E}\left[\left(S_T^{\text{BS}}-K\right)^+\right]\\ &= \int\limits_{-\infty}^{+\infty}\left(S_0\exp\left(\left(r-\frac{\sigma_{\text{BS}}^2}{2}\right)T+\sigma_{\text{BS}}\sqrt{T}x\right)-K\right)^+\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)\mathrm{d}x\\ &= \int\limits_{\frac{\log\left(\frac{K}{S_0}\right)+\left(r+\frac{\sigma_{\text{BS}}^2}{2}\right)T}{\sigma_{\text{BS}}\sqrt{T}}}^{+\infty}\left(S_0\exp\left(\left(r-\frac{(\sigma_{\text{BS}})^2}{2}\right)T+\sigma_{\text{BS}}\sqrt{T}x\right)-K\right)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)\mathrm{d}x\\ &= S_0\Phi\left(\frac{\log\left(\frac{S_0}{K}\right)+\left(r+\frac{1}{2}\sigma_{\text{BS}}^2\right)T}{\sigma_{\text{BS}}\sqrt{T}}\right)-K\Phi\left(\frac{\log\left(\frac{S_0}{K}\right)+\left(r-\frac{1}{2}\sigma_{\text{BS}}^2\right)T}{\sigma_{\text{BS}}\sqrt{T}}\right)\end{aligned} \tag{15.3.1130}$$

Remarks

R1. If we have a Call or Put at the money (for reminder it means that $S=K$) then the two coefficients above simplify since the Napierian logarithm is then null.

R2. Sometimes we write:

$$d_1=\frac{\ln\left(\dfrac{S}{K}\right)+\left(r+\dfrac{1}{2}\sigma^2\right)(T-t)}{\sigma\sqrt{T-t}}=\frac{\ln\left(M(t)\right)+\left(r+\dfrac{1}{2}\sigma^2\right)(T-t)}{\sigma\sqrt{T-t}} \tag{15.3.1131}$$

where $M(t)$ is the moneyness of which we have already spoken several times.

From the relation:

$$C^{\text{BS}}(S,t)=S\Phi(d_1)-Ke^{-r(T-t)}\Phi(d_2) \tag{15.3.1132}$$

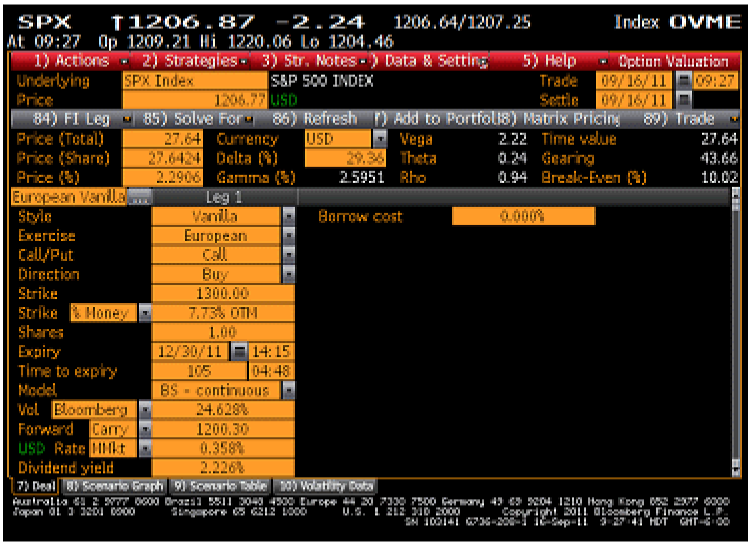
We obviously deduce that a Call option can not have a value greater than the underlying (otherwise there would be an arbitrage opportunity to buy the underlying and sell the Call option - at least if we find a rather stupid buyer ... - and in extenso it would make us win every time).

- For the European Put (the value of the sell option of maturity T and strike K) for a non-dividend-paying share-based underlying and for a geometric Brownian motion:

$$P^{\text{BS}}(S,t)=-S\Phi(-d_1)+Ke^{-r(T-t)}\Phi(-d_2)=S(\Phi(d_1)-1)-Ke^{-r(T-t)}(\Phi(d_2)-1) \tag{15.3.1133}$$

From this last relation we conclude that a Put option can never have a value greater than the strike K otherwise there would also be an arbitrage opportunity by selling the Put and buying the underlying.

If necessary for the curious reader here is an example of valuing a vanilla Call option on a Bloomberg terminal:


1) Actions
2) Strategies
3) Str. Notes
) Data & Setting
5) Help
Option Valuation
84) FI Leg
85) Solve For
86) Refresh
) Add to Portfol
8) Matrix Pricing
89) Trade
European Vanilla
Leg 1
7) Deal
8) Scenario Graph
9) Scenario Table
10) Volatility Data

SPX ↑1206.87 −2.24 1206.64/1207.25 Index OVME
At 09:27 Op 1209.21 Hi 1220.06 Lo 1204.46

Underlying SPX Index S&P 500 INDEX Trade 09/16/11 09:27
Price 1206.77 USD Settle 09/16/11

Price (Total) 27.64 Currency USD Vega 2.22 Time value 27.64
Price (Share) 27.6424 Delta (%) 29.36 Theta 0.24 Gearing 43.66
Price (%) 2.2906 Gamma (%) 2.5951 Rho 0.94 Break-Even (%) 10.02

Style Vanilla Borrow cost 0.000%
Exercise European
Call/Put Call
Direction Buy
Strike 1300.00
Strike % Money 7.73% OTM
Shares 1.00
Expiry 12/30/11 14:15
Time to expiry 105 04:48
Model BS - continuous
Vol Bloomberg 24.628%
Forward Carry 1200.30
USD Rate MMkt 0.358%
Dividend yield 2.226%

Australia 61 2 9777 8600 Brazil 5511 3048 4500 Europe 44 20 7330 7500 Germany 49 69 9204 1210 Hong Kong 852 2977 6000
Japan 81 3 3201 8900 Singapore 65 6212 1000 U.S. 1 212 318 2000 Copyright 2011 Bloomberg Finance L.P.
SN 102141 G736-208-1 16-Sep-11 9:27:41 MDT GMT-6:00

Figure 15.187: **Valuation of a vanilla Call option on 1 underlying of the S&P 500**

It is relatively easy to verify that the solutions above satisfy the Put-Call parity equation:

$$S+P-C=Ke^{-r(T-t)} \qquad (15.3.1134)$$

where we recall that K is the strike price of the two options P and C.

Now, as promised earlier, we can prove the explicit relations of the Greeks for the closed relations of the Calls (the principle being the same for the Put), besides the relations demonstrated above, notice that:

$$\mathcal{N}(d_2)=\mathcal{N}(d_1)\frac{S}{K}e^{r(T-t)} \qquad (15.3.1135)$$

Indeed:

$$\begin{aligned}\mathcal{N}(d_2)&=\frac{\partial\Phi(d_2)}{\partial d_2}=\frac{1}{\sqrt{2\pi}}e^{-\frac{d_2^2}{2}}=\frac{1}{\sqrt{2\pi}}e^{-\frac{(d_1-\sigma\sqrt{T-t})^2}{2}}=\frac{1}{\sqrt{2\pi}}e^{-\frac{d_1^2}{2}}e^{d_2\sigma\sqrt{T-t}}e^{-\frac{\sigma^2(T-t)}{2}}\\&=\frac{1}{\sqrt{2\pi}}e^{-\frac{d_1^2}{2}}e^{\ln\left(\frac{S}{K}\right)+\left(r+\frac{\sigma^2}{2}\right)(T-t)}e^{-\frac{\sigma^2(T-t)}{2}}=\frac{1}{\sqrt{2\pi}}e^{-\frac{d_1^2}{2}}\frac{S}{K}e^{r(T-t)}\\&=\mathcal{N}(d_1)\frac{S}{K}e^{r(T-t)}\end{aligned} \qquad (15.3.1136)$$

and also that:

$$d_2=d_1-\sigma\sqrt{T-t} \qquad (15.3.1137)$$

We have immediately:

$$\frac{\partial d_1}{\partial S}=\frac{\partial d_2}{\partial S} \qquad (15.3.1138)$$

So now let's go!

- The Delta of the Call:

$$\begin{aligned}\Delta_C &= \frac{\partial C^{\text{BS}}}{\partial S} = \frac{\partial\left(S\Phi(d_1)-Ke^{-r(T-t)}\Phi(d_2)\right)}{\partial S} = \Phi(d_1)+S\frac{\partial\Phi(d_1)}{\partial S}-Ke^{-r(T-t)}\frac{\partial\Phi(d_2)}{\partial S}\\ &= \Phi(d_1)+S\frac{\partial\Phi(d_1)}{\partial d_2}\frac{\partial d_1}{\partial S}-Ke^{-r(T-t)}\frac{\partial\Phi(d_2)}{\partial d_2}\frac{\partial d_2}{\partial S}\\ &= \Phi(d_1)+S\mathcal{N}(d_1)\frac{\partial d_2}{\partial S}-Ke^{-r(T-t)}\mathcal{N}(d_1)\frac{S}{K}e^{r(T-t)}\frac{\partial d_2}{\partial S}\\ &= \Phi(d_1)+S\mathcal{N}(d_1)\frac{\partial d_1}{\partial S}-S\mathcal{N}(d_1)\frac{\partial d_2}{\partial S}=\mathcal{N}(d_1)+S\mathcal{N}(d_1)\frac{\partial d_1}{\partial S}-S\mathcal{N}(d_1)\frac{\partial d_1}{\partial S}\\ &= \Phi(d_1)>0\end{aligned} \tag{15.3.1139}$$

Since the derivative is always positive, we can conclude that the price of Call only increases if the underlying increases in value (this is obviously the reverse for Put).

The "delta of the Put" is given by (it should also be quite immediate):

$$\Delta_P = \frac{\partial C^{\text{BS}}}{\partial S} = \Phi(d_1)-1 \tag{15.3.1140}$$

which is therefore always a negative value.

- The Gamma of the Call:

$$\Gamma_C = \frac{\partial^2 C^{\text{BS}}}{\partial S^2} = \frac{\partial}{\partial S}(\Phi(d_1)) = \frac{\partial\Phi(d_1)}{\partial d_1}\frac{\partial d_1}{\partial S} = \frac{\mathcal{N}(d_1)}{S\sigma\sqrt{T-t}}>0 \tag{15.3.1141}$$

- The Vega of the Call:

$$\begin{aligned}\nu_C &= \frac{\partial C^{\text{BS}}}{\partial \sigma} = \frac{\partial\left(S\Phi(d_1)-Ke^{-r(T-t)}\Phi(d_2)\right)}{\partial\sigma} = S\frac{\partial\Phi(d_1)}{\partial\sigma}-Ke^{-r(T-t)}\frac{\partial\Phi(d_2)}{\partial\sigma}\\ &= S\frac{\partial\Phi(d_1)}{\partial d_1}\frac{\partial d_1}{\partial\sigma}-Ke^{-r(T-t)}\frac{\partial\Phi(d_2)}{\partial d_2}\frac{\partial d_2}{\partial\sigma}\\ &= S\mathcal{N}(d_1)\frac{\partial d_1}{\partial\sigma}-Ke^{-r(T-t)}\mathcal{N}(d_2)\frac{\partial d_2}{\partial\sigma}=S\mathcal{N}(d_1)\frac{\partial d_1}{\partial\sigma}-S\mathcal{N}(d_1)\frac{\partial d_2}{\partial\sigma}\\ &= S\mathcal{N}(d_1)\left(\frac{\partial d_1}{\partial\sigma}-\frac{\partial d_2}{\partial\sigma}\right)=S\mathcal{N}(d_1)\left(\frac{\partial d_1}{\partial\sigma}-\frac{(d_1-\sigma\sqrt{T-t}}{\partial\sigma}\right)\\ &= S\mathcal{N}(d_1)\left(\frac{\partial d_1}{\partial\sigma}-\frac{\partial d_1}{\partial\sigma}-\frac{\partial\sigma\sqrt{T-t}}{\partial\sigma}\right)=S\mathcal{N}(d_1)\sqrt{T-t}>0\end{aligned} \tag{15.3.1142}$$

We thus find here a variation of the price of the option which is proportional to the square root of the elapsed time (the origin of this property being well known to us!).

- The Theta of the Call:

$$\begin{aligned}
\theta_C=\frac{\partial C^{\text{BS}}}{\partial t}&=\frac{\partial\left(S\Phi(d_1)-Ke^{-r(T-t)}\Phi(d_2)\right)}{\partial t}\\
&=S\frac{\partial\Phi(d_1)}{\partial t}+rKe^{-r(T-t)}\Phi(d_2)-Ke^{-r(T-t)}\frac{\partial\Phi(d_2)}{\partial t}\\
&=S\frac{\partial\Phi(d_1)}{\partial d_1}\frac{\partial d_1}{\partial t}+rKe^{-r(T-t)}\Phi(d_2)-Ke^{-r(T-t)}\frac{\partial\Phi(d_2)}{\partial d_2}\frac{\partial d_2}{\partial t}\\
&=S\mathcal{N}(d_1)\frac{\partial d_1}{\partial t}+rKe^{-r(T-t)}\Phi(d_2)-Ke^{-r(T-t)}\mathcal{N}(d_2)\frac{\partial d_2}{\partial t}\\
&=S\mathcal{N}(d_1)\frac{\partial d_1}{\partial t}+rKe^{-r(T-t)}\Phi(d_2)-Ke^{-r(T-t)}\mathcal{N}(d_1)\frac{S}{K}e^{r(T-t)}\frac{\partial d_2}{\partial t}\\
&=S\mathcal{N}(d_1)\frac{\partial d_1}{\partial t}+rKe^{-r(T-t)}\Phi(d_2)-S\mathcal{N}(d_1)\frac{\partial d_2}{\partial t}\\
&=S\mathcal{N}(d_1)\frac{\partial d_1}{\partial t}+rKe^{-r(T-t)}\Phi(d_2)-S\mathcal{N}(d_1)\frac{\partial\left(d_1-\sigma\sqrt{T-t}\right)}{\partial t}\\
&=S\mathcal{N}(d_1)\frac{\partial d_1}{\partial t}+rKe^{-r(T-t)}\Phi(d_2)-S\mathcal{N}(d_1)\frac{\partial d_1}{\partial t}+S\mathcal{N}(d_1)\sigma\frac{\partial(\sqrt{T-t})}{\partial t}\\
&=rKe^{-r(T-t)}\Phi(d_2)-S\mathcal{N}(d_1)\sigma\left(\frac{1}{2}\frac{-1}{\sqrt{T-t}}\right)\\
&=rKe^{-r(T-t)}\Phi(d_2)+\frac{S\mathcal{N}(d_1)\sigma}{2\sqrt{T-t}}
\end{aligned}\tag{15.3.1143}$$

- The rho of the Call:

$$\begin{aligned}
\rho_C=\frac{\partial C^{\text{BS}}}{\partial r}&=\frac{\partial\left(S\Phi(d_1)-Ke^{-r(T-t)}\Phi(d_2)\right)}{\partial r}\\
&=S\frac{\partial\Phi(d_1)}{\partial r}+(T-t)Ke^{-r(T-t)}\Phi(d_2)-Ke^{-r(T-t)}\frac{\partial\Phi(d_2)}{\partial r}\\
&=S\frac{\partial\Phi(d_1)}{\partial d_1}\frac{\partial d_1}{\partial r}+(T-t)Ke^{-r(T-t)}\Phi(d_2)-Ke^{-r(T-t)}\frac{\partial\Phi(d_2)}{\partial d_2}\frac{\partial d_2}{\partial r}\\
&=S\mathcal{N}(d_1)\frac{\partial d_1}{\partial r}\frac{\partial d_1}{\partial r}+(T-t)Ke^{-r(T-t)}\Phi(d_2)-Ke^{-r(T-t)}\mathcal{N}(d_2)\frac{\partial d_2}{\partial r}\\
&=S\mathcal{N}(d_1)\frac{\partial d_1}{\partial r}+(T-t)Ke^{-r(T-t)}\Phi(d_2)-Ke^{-r(T-t)}\mathcal{N}(d_1)e^{r(T-t)}\frac{\partial d_2}{\partial r}\\
&=S\mathcal{N}(d_1)\frac{\partial d_1}{\partial r}+(T-t)Ke^{-r(T-t)}\Phi(d_2)-S\mathcal{N}(d_1)\frac{\partial d_2}{\partial r}\\
&=S\mathcal{N}(d_1)\frac{\partial d_1}{\partial r}+(T-t)Ke^{-r(T-t)}\Phi(d_2)-S\mathcal{N}(d_1)\frac{\partial\left(d_1-\sigma\sqrt{T-t}\right)}{\partial r}\\
&=S\mathcal{N}(d_1)\frac{\partial d_1}{\partial t}+(T-t)Ke^{-r(T-t)}\Phi(d_2)-S\mathcal{N}(d_1)\frac{\partial d_1}{\partial r}\\
&=(T-t)Ke^{-r(T-t)}\Phi(d_2)>0
\end{aligned}\tag{15.3.1144}$$

So we can easily understand why the price of a Call is as high as r is large (at least in the standard values of r such that we can observe them in markets).

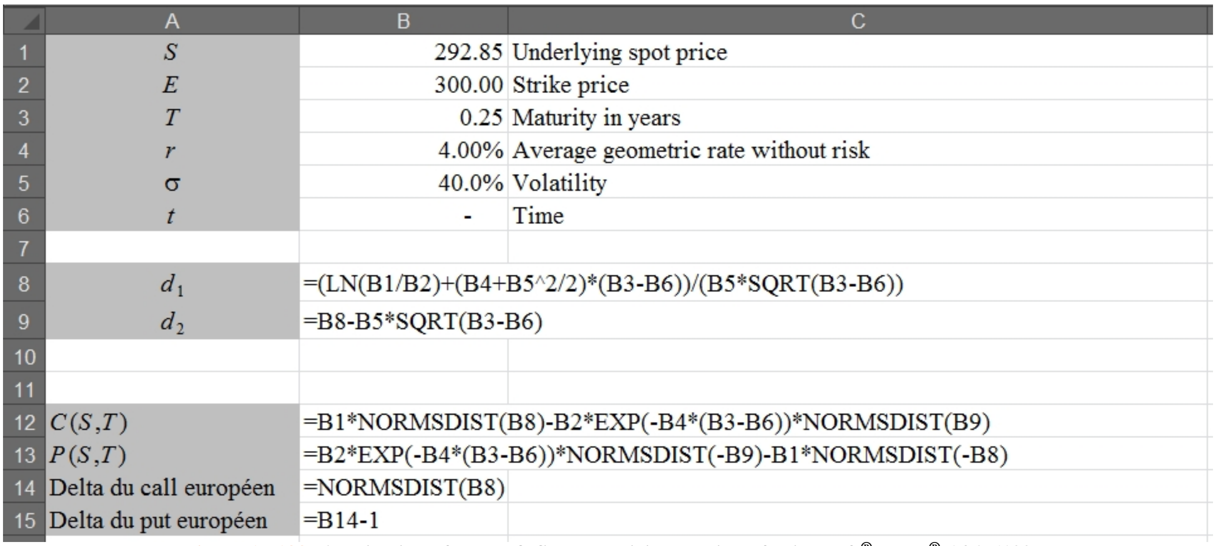
A common question from students (and practitioners) is to know whether the price of a Call increases or decreases depending on the distance at maturity. It is difficult to conclude in just a few seconds by deriving relatively to T so it is better to go through computer tools and hence it is quite easy to verify that the intuition that the price of Call increases with time at maturity is correct. Moreover, we can also confirm another intuition that the price of Call tends towards the value of the strike, the more T is distant.

The same goes for the price of a Call depending on the strike. There it is easier to do it by hand:

$$\begin{aligned}\frac{\partial C^{\text{BS}}}{\partial K}&=\frac{\partial\left(S\Phi(d_1)-Ke^{-r(T-t)}\Phi(d_2)\right)}{\partial S}=S\frac{\partial\Phi(d_1)}{\partial K}-e^{-r(T-t)}\Phi(d_2)-e^{-r(T-t)}\frac{\partial\Phi(d_2)}{\partial K}\\&=S\frac{\partial\Phi(d_1)}{\partial d_1}\frac{\partial d_1}{\partial K}-e^{-r(T-t)}\Phi(d_2)-Ke^{-r(T-t)}\frac{\partial\Phi(d_2)}{\partial d_2}\frac{\partial d_2}{\partial K}\\&=S\mathcal{N}(d_1)\frac{\partial d_1}{\partial K}-e^{-r(T-t)}\Phi(d_2)-Ke^{-r(T-t)}\mathcal{N}(d_1)\frac{S}{K}e^{r(T-t)}\frac{\partial d_2}{\partial K}\\&=S\mathcal{N}(d_1)\frac{\partial d_1}{\partial K}-e^{-r(T-t)}\Phi(d_2)-S\mathcal{N}(d_1)\frac{\partial d_2}{\partial K}\\&=S\mathcal{N}(d_1)\frac{\partial d_1}{\partial K}-e^{-r(T-t)}\Phi(d_2)-S\mathcal{N}(d_1)\frac{\partial d_1}{\partial K}\\&=-e^{-r(T-t)}\Phi(d_2)\end{aligned}\tag{15.3.1145}$$

So the higher the price of the strike, the less the price of the Call will be.

Here are the commands built into the Microsoft® Excel® 14.0.6129 to do the calculation with arbitrary input values:



	A	B	C
1	S	292.85	Underlying spot price
2	E	300.00	Strike price
3	T	0.25	Maturity in years
4	r	4.00%	Average geometric rate without risk
5	σ	40.0%	Volatility
6	t	-	Time
7			
8	d_1	=(LN(B1/B2)+(B4+B5^2/2)*(B3-B6))/(B5*SQRT(B3-B6))	
9	d_2	=B8-B5*SQRT(B3-B6)	
10			
11			
12	$C(S,T)$	=B1*NORMSDIST(B8)-B2*EXP(-B4*(B3-B6))*NORMSDIST(B9)	
13	$P(S,T)$	=B2*EXP(-B4*(B3-B6))*NORMSDIST(-B9)-B1*NORMSDIST(-B8)	
14	Delta du call européen	=NORMSDIST(B8)	
15	Delta du put européen	=B14-1	

Figure 15.188: **Application of Black & Scholes pricing version of Microsoft® Excel® 14.0.6129**

Remark

It is clear that the Black & Scholes have allowed the developments of options markets, allowing for more secure speculation. This remains speculation (the players speculate relative to each other on the volatility of equities), but this speculation remains almost secured by the coverage equation, which avoids that the losses are too large. There are nevertheless disadvantages to their use. The most important is surely the effect of runaway that they cause. For example, suppose that you are the seller of an option on the stock of a company A. Imaging that this company announces results slightly lower than expected. Its price falls, and that's normal. The Black & Scholes coverage equation then recommends that you reduce the number of shares of that company in your portfolio, which you do (at least if you did the error of not replicate). But almost all market players will do the same reasoning, causing a further decline in the share price. The coverage equation of Black & Scholes recommends that you sell more shares, etc. This can trigger a real runaway on the market, both downward and upward. This is accentuated by the fact that, often, purchase or sales orders are new automated, implemented directly in softwares, and no longer require human intervention. On the other hand, the coverage equation of Black & Scholes is efficient for small price fluctuations, but not for brutal and important "unscrewing". Thus, just a year after receiving their Nobel Prize in Economics, Robert Merton and Myron Scholes were involved in the collapse of the US investment fund LTCM in autumn 11998 (holocene calendar), following the severe Russian crisis of the 11998 (holocene calendar) summer.

In fact, option prices are not calculated using the Black & Scholes formula (especially since the return is assumed to be constant ... hence the use of Monte Carlo techniques and binomial trees that we will see further below). Most of the time, these prices are simply the result of the law of supply and demand, which prevails in most markets (for which reason, in markets, Calls for an underlying of a given spot and same strike and maturity do not have the same price!). When we then look at the non-reduced expression:

$$\boxed{C^{\text{BS}}(S,t)=S\Phi\left(\dfrac{\ln\left(\frac{S}{K}\right)+\left(r+\frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}}\right)-Ke^{-r(T-t)}\Phi\left(\dfrac{\ln\left(\frac{S}{K}\right)+\left(r-\frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}}\right)} \tag{15.3.1146}$$

It is therefore an equation with one unknown, since the only value then lacking in market practice is the yield volatility that we then name "implied volatility" (as opposed to the volatility used when we value options and which we classically named "empirical (observed) volatility") and which should (...) in theory be equal in the markets for a Call and a Put having the same maturity and the same exercise price.

In practice, the calculation of implied volatility is made, for a given underlying, for several K and T values at the spot and yield given by the market at present or in the future. In the end, instead of having only one point, we get a tablecloth like the one below:

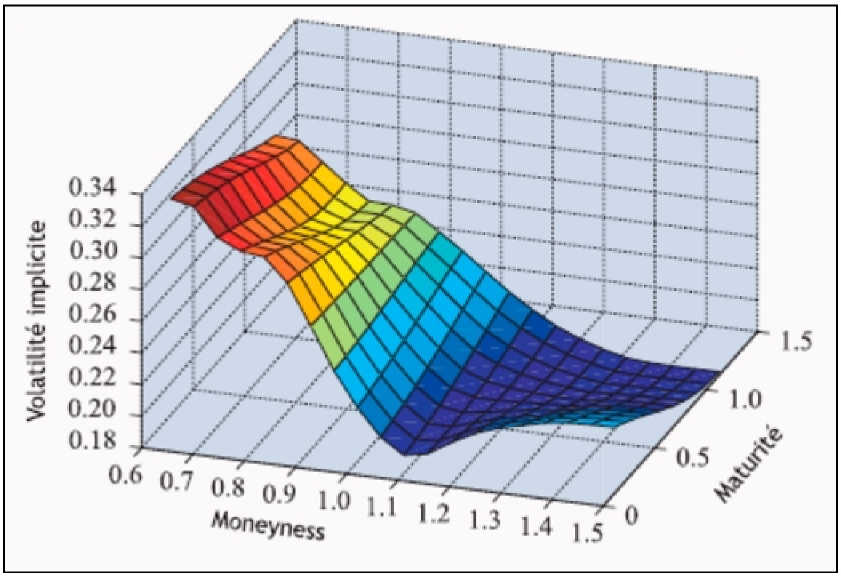
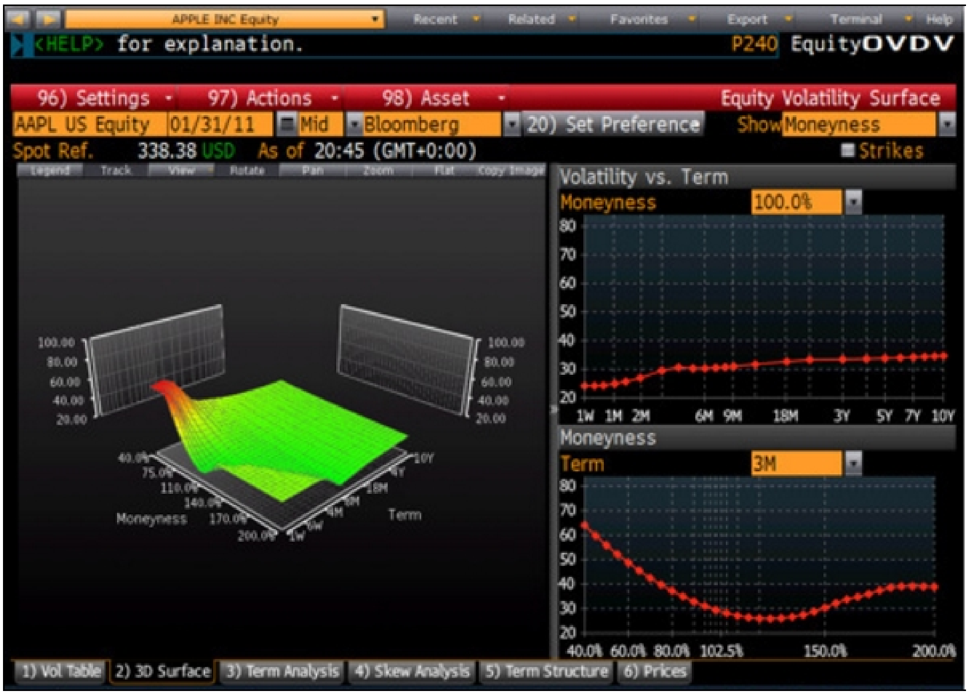


Figure 15.189: Typical profile of an implied volatility surface

where the "moneyness" represents the ratio between S and K. If the Black & Scholes equation was respected, we should have if the above a plane (so something completely flat)...


<HELP> for explanation. P240 Equity OVDV

AAPL US Equity 01/31/11 Mid Bloomberg

Spot Ref. 338.38 USD As of 20:45 (GMT+0:00)

Show Moneyness

Moneyness 100.0%

Term 3M

APPLE INC Equity
Recent
Related
Favorites
Export
Terminal
Help
96) Settings
97) Actions
98) Asset
Equity Volatility Surface
20) Set Preference
Strikes
Legend
Track
View
Rotate
Pan
Zoom
Flat
Copy Image
100.00
80.00
60.00
40.00
20.00
40.0%
75.0%
110.0%
140.0%
170.0%
200.0%
Moneyness
1W
6W
6M
18M
4Y
10Y
Term
Volatility vs. Term
80
70
60
50
40
30
20
1W 1M 2M 6M 9M 18M 3Y 5Y 7Y 10Y
Moneyness
80
70
60
50
40
30
20
40.0% 60.0% 80.0% 102.5% 150.0% 200.0%
1) Vol Table
2) 3D Surface
3) Term Analysis
4) Skew Analysis
5) Term Structure
6) Prices

Figure 15.190: Typical profile of implied volatility in the Bloomberg terminal

On the markets, practitioners (especially traders) constantly look at the surfaces of implied volatilities (of which the interpretations can be multiple!). The concern is that these surfaces are not static: they change every day. The knowledge (in fact rather the approximation...) of the dynamics of the volatility sheets is therefore a crucial and subjectively complex subject (whole textbooks are dedicated only to this subject!)...

A typical approach is to make assumptions about the behaviour of the underlying. Sometimes there are closed formulas (as in the Black & Scholes model), but often it is necessary to use numerical methods to obtain the price of the corresponding option and then to reverse the problem to find the local volatility.

Definitions (#15.3.697):

- D1. We name "historical volatility" or "statistical volatility" or "realized volatility", the past volatility of a financial instrument underlying an option.
- D2. We name "implied volatility" or "forward volatility", the volatility injected into the Black & Scholes model which is an estimate of the average of the future projection of volatility over a given period.
- D3. We name "local volatility" the volatility of an option calculated a posteriori option from the Dupire model (inversion of the Black & Scholes relation) in a way.
- D4. We name "observed volatility" the true volatility of an option as observed in financial markets.

Now let us talk about the case of the underlying that pay dividends! If the underlying whose spot price we inject into the option valuation equation is S it is supposed as we know to be the present value estimated by the economic agents. If the underlying pays dividends then this price is actually not its actual face value since it implicitly contains the present values of the dividends! To obtain its true value, it is then necessary to subtract from S the updated values of the dividends before injecting the value of the evaluation relations.

✍**Example:**

Let us consider a Call on 1 year of a BMW share with a strike price of 40.-. Suppose the current spot price of the BMW share is 35.-, the instantaneous annual risk-free market interest rate is 5%, the volatility is 20% per year, and there are two dividends during the period of one year of respectively 1.- at two months and 0.5.- to eight months.

Therefore, the present value of the dividends is:

$$1\cdot e^{-\frac{2}{12}5\%}+5\cdot e^{-\frac{8}{12}5\%}$$

and not 35.-. And thus this deterministic amount is implicitly included in the current spot price of the market. Therefore, the spot value S that we have to inject into the evaluation of the Call is:

$$35-1.4753=33.5247\text{.-}$$

and not 35.-.

In the case where the underlying pays dividends which in constant continuous rate q then its true face value discounted on the time T will be given using the continuous rate (if necessary review the construction of the concept of continuous rate!):

$$Se^{-qT} \tag{15.3.1147}$$

By injecting this last relation into the evaluation of the options, we have for example for the Call:

$$\begin{aligned}C_d^{\text{BS}} &= Se^{-q(T-t)}\Phi\left(\frac{\ln\left(\frac{Se^{-q(T-t)}}{K}\right)+\left(r+\frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}}\right)\\ &-Ke^{-r(T-t)}\Phi\left(\frac{\ln\left(\frac{Se^{-q(T-t)}}{K}\right)+\left(r-\frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}}\right)\\ &= Se^{-q(T-t)}\Phi\left(\frac{\ln\left(\frac{S}{K}\right)-q(T-t)+\left(r+\frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}}\right)\\ &-Ke^{-r(T-t)}\Phi\left(\frac{\ln\left(\frac{S}{K}\right)-q(T-t)+\left(r-\frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}}\right)\\ &= Se^{-q(T-t)}\Phi\left(\frac{\ln\left(\frac{S}{K}\right)+\left((r-q)+\frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}}\right)\\ &-Ke^{-r(T-t)}\Phi\left(\frac{\ln\left(\frac{S}{K}\right)-\left((r-q)-\frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}}\right)\end{aligned} \tag{15.3.1148}$$

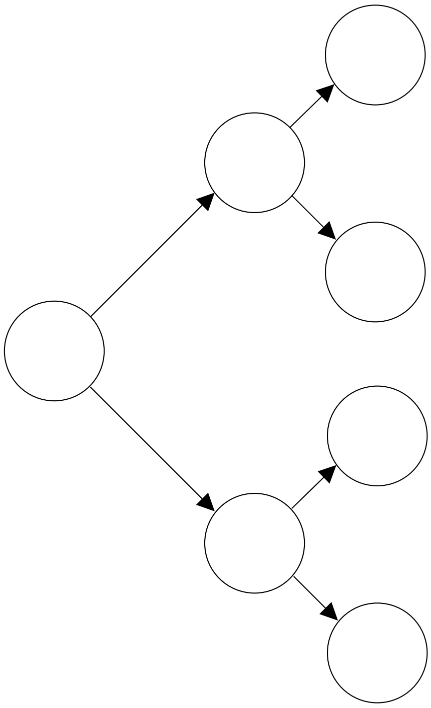
It is quite obvious that this last equality is a generalization of the options on the underlying type of assets paying no dividends. We could, by the way, here again calculate the corresponding Greeks ...

15.3.10.3.10 Binomial option pricing model (CRR model)

There is an older style approach to the previous continuous-time option pricing model that is not used anymore in practice in the simple form we will see here (CRR model) but that has inspired quite a lot of much more modern and complicated models that are still sometimes used today[28].

The binomial pricing model (we should speak in the plural because there are many binomial pricing models!) is a technique which assumes that at each time period, the financial instrument of interest can only have two different values compared to a given initial value: a higher or lower value (in the case of the trinomial model we add the case where the value remains constant).

To represent this, we use what we know as being named a "binomial tree", that is to say a succession of sequences of two branches which represent the possible changes in price at nodes:



We will see that the asymptotic limit of the (discrete time) binomial CRR model will take us back to the Black & Scholes formula. But before doing this let us recall that:

1. A portfolio is a "replicating portfolio" of an option if the portfolio and the option have exactly the same payoff in each state of future.
2. By using "no arbitrage argument", the cost or price of the replication portfolio is the same as the value of the option.

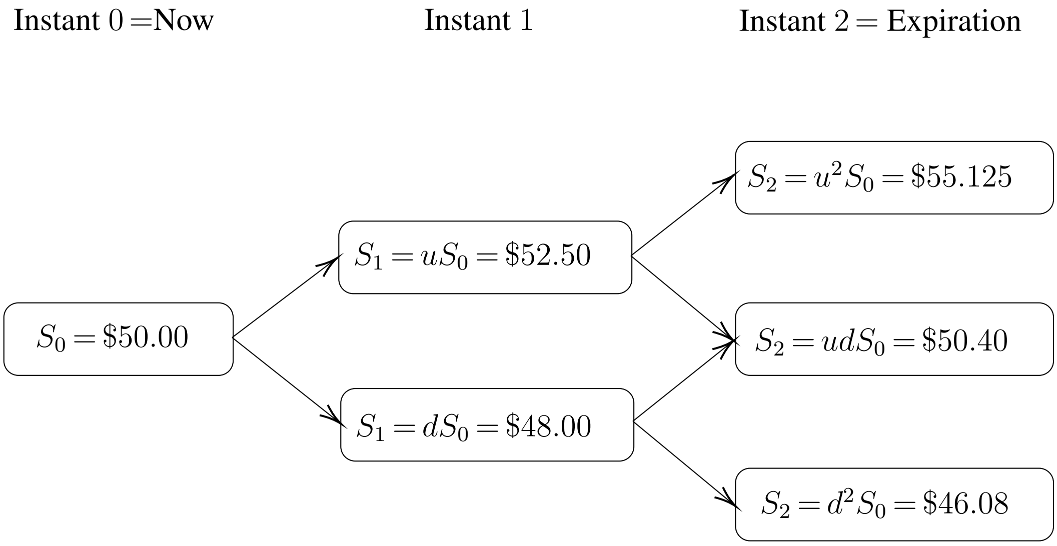
---

[28] Indeed the binomial tree (apart from the special case of CRR!) remains used by practitioners because it can take into account many complications that purely theoretical models cannot in a single block such as, for example, the costs of transactions, dividends, empirical criteria of the portfolio client, etc. What will also make us prefer the binomial tree to the Monte Carlo method or vice versa will be mainly the complexity of the algorithm for the calculations (i.e. calculation time) but not only!

Remark

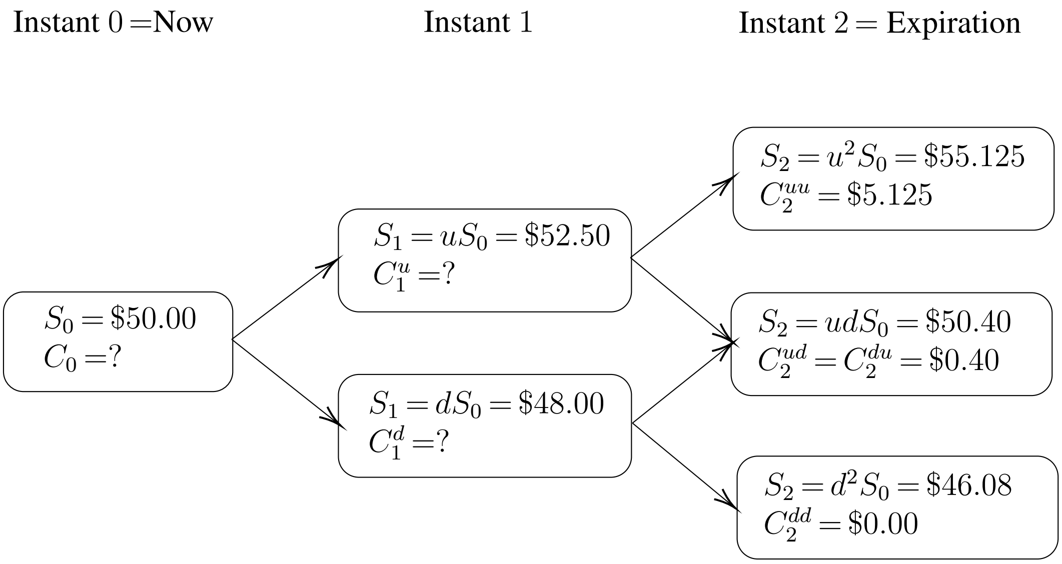
The finance literature has revealed no fewer than 11 alternative version of the binomial option pricing model for options on log-normally assets (CRR with or without drift, Jarrow-Rudd model, Tian model, Leisen-Reimer model, ...). These models are derived under a variety of assumption and in some cases require information that is ordinarily unnecessary to value options [94].

The basic building element for the Black & Scholes formula is the assumption that over one instant, the stock price can only move up or down (this is named a "binomial process"). So you must first understand how to work in such a world. Over two instants, the stock price can move up twice, move up once and move down once, or move down twice. Use the letter u to describe the stock price multiplier when an up move occurs, and d to describe the stock price multiplier when a down move occurs. You can represent the stock price process with a binomial tree - where one branch represents a price-up movement and the other a price-down movement. For example if $d=0.96$ (which means that on a down move, the stock price declines by 4%) and $u=1.05$ (the stock price increases by 5%), the stock price is as follows:



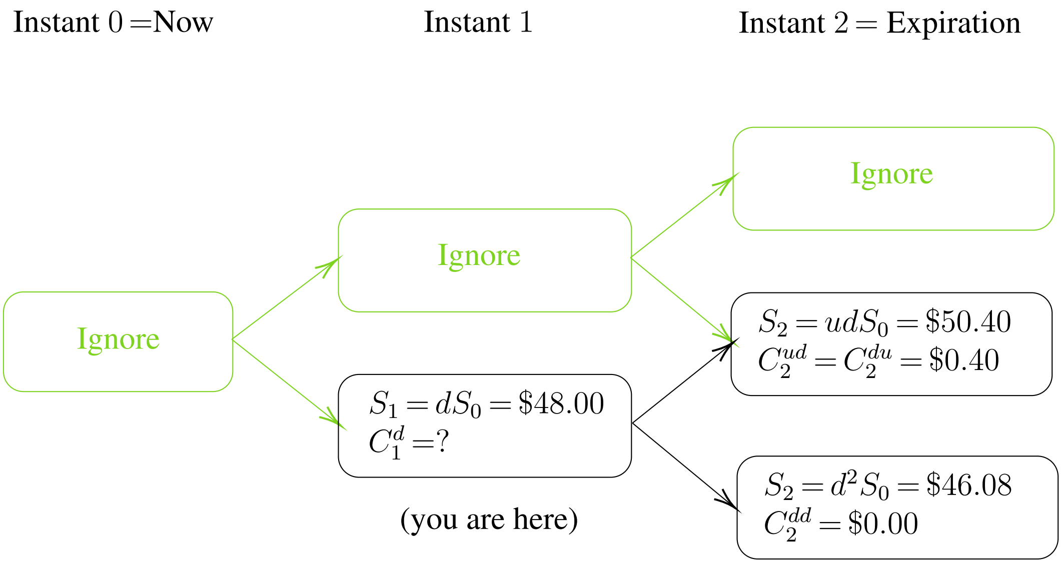
Note that at Instant 2, the middle outcome occurs on two possible paths, while the two extreme outcomes occur only on one path each; $u\cdot d\cdot S_0$ can come about if there is one u followed by one d, or if there is one d followed by one u. This is already a statistical distribution that shares with a bell-shaped (Normal) distribution the feature that middle outcomes are more likely than extreme outcomes (with many bell shaped distributions and binomial trees we indeed end up with a continuous distribution that looks a lot like a bell-shaped curve).

If we know that our underlying stock follows this binomial process, and we know the values of u and d, can we price a Call option with a strike price of $\$50$.-? On inspection of the tree, realize that the Call option pays \$ 0 if the stock moves down twice, $\$0.40$.- if the stock price moves up once and down once (or vice versa), and $\$5.125$.- if the stock price moves up twice:


Instant $0 =$Now
Instant 1
Instant $2 =$ Expiration
$S_0 = \$50.00$
$C_0 = ?$
$S_1 = uS_0 = \$52.50$
$C_1^u = ?$
$S_1 = dS_0 = \$48.00$
$C_1^d = ?$
$S_2 = u^2S_0 = \$55.125$
$C_2^{uu} = \$5.125$
$S_2 = udS_0 = \$50.40$
$C_2^{ud} = C_2^{du} = \$0.40$
$S_2 = d^2S_0 = \$46.08$
$C_2^{dd} = \$0.00$

As we know it, the purpose of the seller of the Call is to calculate the risk premium C_0 of the Call and this should be made such that he can hedge the option risk. This should be made using arbitrage opportunity with an another investment on the market. So the risk premium should be calculated at each instant to what would happen (what will be the cost) if we hedge the Call option by buying buy immediately a portfolio Δ with a quantity δ of the underlying stocks and (assumed) a given quantity b risk-free bonds.

First let us place ourselves into the position where the stock price has moved down once already, that is, where the stock price stands at $\$48.00$.-:


Instant $0 =$Now
Instant 1
Instant $2 =$ Expiration
Ignore
Ignore
Ignore
$S_1 = dS_0 = \$48.00$
$C_1^d = ?$
(you are here)
$S_2 = udS_0 = \$50.40$
$C_2^{ud} = C_2^{du} = \$0.40$
$S_2 = d^2S_0 = \$46.08$
$C_2^{dd} = \$0.00$

So the equation is: what would be the minimum cost (and therefore the composition) of a replication portfolio such that we receive \$ 0 if the underlying stocks of our Call moves down and if we receive $\$0.40$.- if the stocks moves up. That is assuming a risk-free rate of (for example) 0.1% and that we know the underling is traded at $\$48.00$.- the following system of two linear equations with two unknowns:

$$\begin{gathered}\delta\cdot d\cdot S_0+b(1+r)=C_d\\ \delta\cdot u\cdot S_0+b(1+r)=C_u\end{gathered}\tag{15.3.1149}$$

That is with our numerical values:

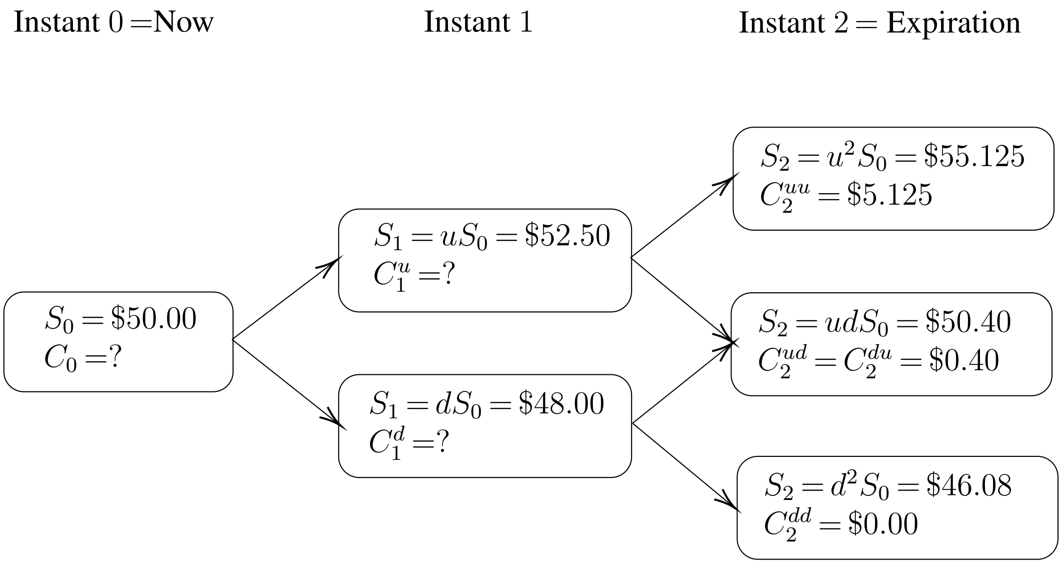
$$\begin{gathered}\delta\cdot 0.96\cdot \$48.00+b(1.001)=\$0.00\\ \delta\cdot 1.05\cdot \$48+b(1.001)=\$0.40\end{gathered}\tag{15.3.1150}$$

The solution is given after elementary simplifications by:

$$\delta=\frac{C_u-C_d}{u\cdot S_0-d\cdot S_0}=\frac{C_u-C_d}{S_0(u-d)}\cong 0.0926\qquad b\cong -\$4.262\tag{15.3.1151}$$

That is if we purchase a replication portfolio of 0.0926 underlying (which costs $0.0926\cdot \$48\cong \$4.444$.-) and borrow $\$4.262$.- for a net outlay of $4.444-4.262=\$0.182$.-, then in the next period, this replication portfolio will pay off $\$0$.- in the downstate and $\$0.40$.- in the upstate. Because this is exactly the same as the payoff on the Call option, the C_1^d Call option should also be worth $\$0.182$.-.

Now to continue we must still consider still at Instant 1:


Instant 0 =Now
Instant 1
Instant 2 = Expiration
$S_2=u^2S_0=\$55.125$
$C_2^{uu}=\$5.125$
$S_1=uS_0=\$52.50$
$C_1^u=?$
$S_0=\$50.00$
$C_0=?$
$S_2=udS_0=\$50.40$
$C_2^{ud}=C_2^{du}=\$0.40$
$S_1=dS_0=\$48.00$
$C_1^d=?$
$S_2=d^2S_0=\$46.08$
$C_2^{dd}=\$0.00$

Now we repeat the same reasoning where the stock price stands at $\$52.50$.- and such that at the next instant we can end up with either $\$0.40$.- in the downstate or $\$5.125$.- in the upstate. In this case, we simply have to solve:

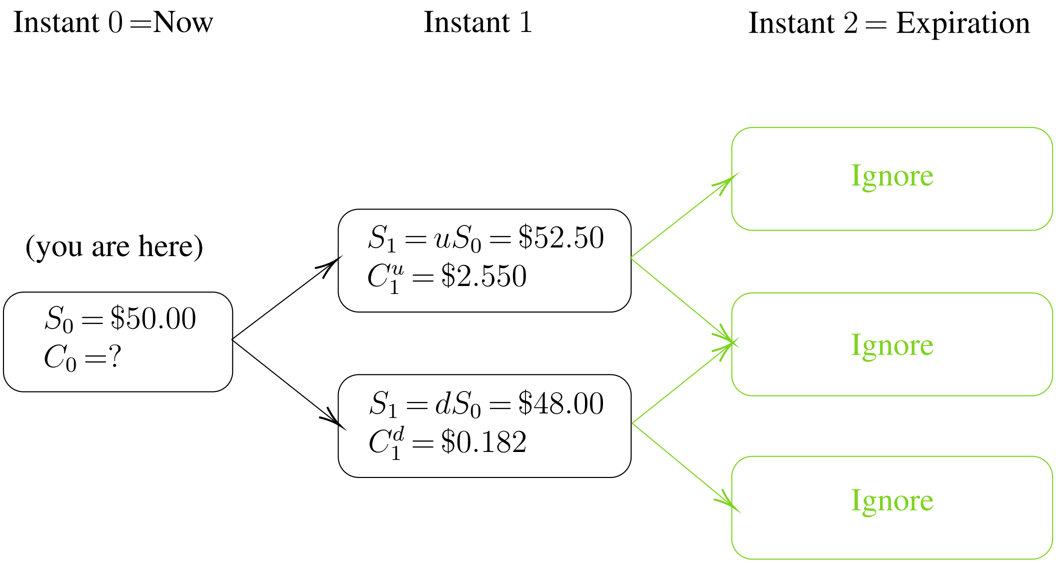
$$\begin{gathered}\delta\cdot 0.96\cdot \$52.50+b(1.001)=\$0.40.\text{-}\\ \delta\cdot 1.05\cdot \$52.50+b(1.001)=\$5.125.\text{-}\end{gathered}\tag{15.3.1152}$$

and the solutions are:

$$\delta=1.00\qquad b\cong -\$49.95.\text{-}\tag{15.3.1153}$$

That is if we purchase a replication portfolio of 1.00 underlying (which costs $1.00\cdot\$52.50=\$52.50$.-) and borrow $\$49.95$.- for a net outlay of $52.50-49.95=\$2.550$.-, then in the next period, this replication portfolio will pay off $\$5.125$ in the upstate and $\$0.40$.- in the downstate. Because this is exactly the same as the payoff on the Call option, the C_1^u Call option should also be worth $\$2.550$.-.

And finally:



To determine the value of the Call C_0 at the outset we have to find the price of the underlying that will be worth $\$0.182$ if the stock moves from $\$50$ to $\$48$, and worth $\$2.55$ if the stock moves from $\$50$ to $\$52.50$:

$$\begin{aligned}\delta\cdot 0.96\cdot\$50.00+b(1.001)&=\$0.182\\ \delta\cdot 1.05\cdot\$50.00+b(1.001)&=\$2.550\end{aligned}\tag{15.3.1154}$$

and the solutions are:

$$\delta\cong 0.5262\qquad b\cong -\$25.05\tag{15.3.1155}$$

So the trader has to purchase 0.5262 shares (cost today: $\$26.31$), and borrow $\$25.05$ dollars. The replication portfolio's total net outlay is then $\$26.31-\$25.05\cong\$1.26$. Therefore, it follows that, by arbitrage, the price of the Call option C_0 must be about $\$1.26$ today.

Remark

American Call options are priced the same way but at the opposite that we can exercise before maturity (early exercise)! In fact we see quickly with the example above that American Call option have then no interest to be exercised before maturity! However for America options, it's not the same! Indeed, American Puts are equals to or more valuable than European Puts if and only if the underlying pays dividends!!! The reason for this is the time value of money. In a Put, we get the option to sell a stock at a given strike price. If we exercise this option at $t=0$, we receive the strike price at $t=0$ and can invest it at the risk-free rate. Lets imagine the r_f rate is 10% and the strike price is $10\$$. this means at $t=1$, we would get $\$11.0517$.-. If, on the other hand, we didn't exercise the option early, at $t=1$ we would simply receive the strike price ($\$10$.-). Basically, the strike price, which is our payoff for a Put option, doesn't earn interest!

In real life, the stock price can move many more times than just twice. You need a tree with many more levels, so you need to generalize this binomial process to more levels. For example, if there are 10 instants, what would be the worst possible outcome? Ten instant down movements mean that the stock price would be:

$$d^{10}S_0=0.96^{10}\cdot\$50\cong\$33.24.\text{-}\qquad(15.3.1156)$$

The second-worst outcome would be one instant of up movement, and nine instants of down movement:

$$d^9u^1S_0=0.96^9\cdot 1.05^1\cdot\$50\cong\$36.36.\text{-}\qquad(15.3.1157)$$

Although the worst scenario can only occur if there are exactly 10 down movements, there are 10 different ways to fall into the second-worst scenario, ranging from $duuuuuuuuu$, $uduuuuuuuu$, ... , to $uuuuuuuuud$. This should bring to the memories of combinatorics (see section Probabilities page 528). These are the 10 possible combinations, better written as:

$$\binom{10}{1}=\frac{10!}{1!9!}=10\qquad(15.3.1158)$$

Thus in general:

$$\binom{N}{i}=\frac{N!}{i!(N-i)!}\qquad(15.3.1159)$$

Therefore, with N levels in the tree, the stock price will be:

$$u^id^{N-i}S_0=u^i(1-u)^{N-i}S_0\qquad(15.3.1160)$$

in $\displaystyle\binom{N}{i}$ paths.

Now let us introduce probabilities! If we have a constant probability p of going up, then we have a probability $(1-p)$ of going down then we can write obviously:

$$(pu)^i((1-p)d)^{N-i}S_0=p^i(1-p)^{N-i}u^i(1-u)^{N-i}S_0\qquad(15.3.1161)$$

that gives the expected value we can get after N paths for i given downside movements (and $N-i$ upside) with a probability p whatever the order of the movements!

We recognize here a well known pattern!!!! Something of the type:

$$p^i(1-p)^{N-i}u^i(1-u)^{N-i}S_0=P_if(x_i)\qquad(15.3.1162)$$

So we could think to a geometric distribution (see section Statistics page 654) but this would be a mistake as all possible identical patterns interest us! So we must multiply that relation by the number of way we can get a same pattern. That is:

$$\binom{N}{i}p^i(1-p)^{N-i}u^i(1-u)^{N-i}S_0\qquad(15.3.1163)$$

that gives the expected total we can get after N paths for i given downside movements (and $N-i$ upside) with a probability p whatever the order of the movements!

And if we want to take into account ALL possible movements, then we are just writing the expected mean and it comes:

$$\text{E}(S_N)=\sum_{i=0}^{N}\binom{N}{i}p^i(1-p)^{N-i}u^i(1-u)^{N-i}S_0 \tag{15.3.1164}$$

But... but... If we consider that we are considering that we are pricing a Call option, we are not interested to all path where $S_0<K$ (since the Call we not be used!). Then for this case we should write:

$$\text{E}(S_N)=\sum_{i=0}^{N}\binom{N}{i}p^i(1-p)^{N-i}\max(S_0u^i(1-u)^{N-i}-K,0) \tag{15.3.1165}$$

But we must not forget that we are looking also to the price compared to a risk-free return investment. So we should write:

$$\boxed{C^{\text{CRR}}=e^{-r_fN}\sum_{i=0}^{N}\binom{N}{i}p^i(1-p)^{N-i}\max(S_0u^i(1-u)^{N-i}-K,0)} \tag{15.3.1166}$$

This relation is known under the name "Cox-Ross-Rubinstein model" (or simply "CRR model") of option pricing as first proposed in 11979 according to holocene calendar (almost 6 years after the publication of Black & Scholes model).

Now to continue we will see why some companies prefer to use binomial pricing rather than Black & Scholes model (as that latter overestimate prices). To fall back on the Black & Scholes equations starting from CCR model we must first impose that:

$$u=\frac{1}{d} \tag{15.3.1167}$$

that is to say that the original stock price will be retained after an equal number of upward and downward price movements in any up-down sequences (quite strong assumption!).

Now let S^* be the random price of the underlying at the end of the N periods. For n upward price movements over the N periods since:

$$S^*=u^nd^{N-n}S_0 \tag{15.3.1168}$$

we get:

$$\ln\left(\frac{S^*}{S_0}\right)=n\ln(u)+(N-n)\ln(d)=n\ln\left(\frac{u}{d}\right)+N\ln(d) \tag{15.3.1169}$$

Here, $\ln(S^*/S)$, is the natural logarithm of one-plus-return for holding the stock over the N periods; it is equivalent to a continuously compounded return over the N periods.

For a binomial distribution with the probability of each upward movement being p, the expected value and the variance of n are (see section Statistics page 658):

$$\text{E}(n)=Np \qquad \text{and} \qquad \text{V}(n)=Np(1-p) \tag{15.3.1170}$$

Then we have immediately involving:

$$\mathrm{E}\left[\ln\left(\frac{S^*}{S_0}\right)\right]=Np\ln\left(\frac{u}{d}\right)+N\ln(d) \tag{15.3.1171}$$

and:

$$\mathrm{V}\left[\ln\left(\frac{S^*}{S_0}\right)\right]=Np(1-p)\left[\ln\left(\frac{u}{d}\right)\right] \tag{15.3.1172}$$

We can capture the above expressions more succinctly by defining (they are both the factors of the N):

$$\hat{\mu}=p\ln(u)+(1-p)\ln(d)=\hat{\sigma}^2=p(1-p)\left[\ln\left(\frac{u}{d}\right)\right]^2 \tag{15.3.1173}$$

as being the fraction time expected and volatility instantaneous return.

And using our previous assumption ($u=1/d$):

$$\begin{aligned}\hat{\mu}&=p\ln(u)+(1-p)\ln\left(\frac{1}{u}\right)=(2p-1)\ln(u)\\ \hat{\sigma}^2&=p(1-p)\left[\ln\left(\frac{u}{1/u}\right)\right]^2=4p(1-p)(\ln(u))^2\end{aligned} \tag{15.3.1174}$$

For algebraic convenience that will soon be clear, let us put:

$$p=\frac{1}{2}+b \tag{15.3.1175}$$

where the parameter b has yet to be determined. Accordingly, we can write:

$$\hat{\mu}=(2p-1)\ln(u)=2b\ln(u) \tag{15.3.1176}$$

and:

$$\hat{\sigma}^2=4p(1-p)(\ln(u))^2=(1-4b^2)(\ln(u))^2=(\ln(u))^2-\hat{\mu}^2 \tag{15.3.1177}$$

If we express u as an exponential function of the form $u=e^h$ the both relations above reduce to:

$$\hat{\mu}=2bh \qquad \text{and} \qquad \hat{\sigma}^2=h^2-\hat{\mu}^2=h^2(1-4b^2) \tag{15.3.1178}$$

Now we suppose also that the convergence of the Black & Scholes and Cox-Ross-Rubinstein models requires that $\hat{\mu}N/T$ and $\hat{\sigma}N/T$ converge to the parameters μ and σ^2, respectively as N approaches infinity!

Now we must found a value for b and h such that when $N\rightarrow+\infty$ we get $\hat{\mu}=\mu$ and $\hat{\sigma}=\sigma$. There are various combinations for this but we must found one where none of the values diverge! One such proposition is after trial and errors:

$$h=\sigma\sqrt{\frac{T}{N}} \qquad \text{and} \qquad b=\frac{\mu\sqrt{T}}{2\sigma\sqrt{N}} \tag{15.3.1179}$$

It follows that:

$$\hat{\mu}=2bh=2\frac{\mu\sqrt{T}}{2\sigma\sqrt{N}}\sigma\sqrt{\frac{T}{N}}=\mu\frac{T}{N} \tag{15.3.1180}$$

Therefore we have indeed:

$$\mu=\hat{\mu}\frac{N}{T} \tag{15.3.1181}$$

It also follows that:

$$\hat{\sigma}^2=h^2(1-4b^2)=\sigma^2\frac{T}{N}\left(1-4\frac{\mu^2T}{4\sigma^2N}\right)=\sigma^2\frac{T}{N}\left(1-\frac{\mu^2T}{\sigma^2N}\right) \tag{15.3.1182}$$

and here we see that if $N\rightarrow+\infty$ we get:

$$\hat{\sigma}^2=\sigma^2\frac{T}{N}\left(1-\underbrace{\frac{\mu^2T}{\sigma^2N}}_{\rightarrow 0}\right)=\sigma^2\frac{T}{N} \tag{15.3.1183}$$

Therefore:

$$\sigma=\hat{\sigma}^2\frac{N}{T} \tag{15.3.1184}$$

Finally we have (what some authors put as a definition):

$$u=\frac{1}{d}=e^{\sigma\sqrt{T/N}} \tag{15.3.1185}$$

and:

$$p=\frac{1}{2}+\frac{\mu}{2\sigma}\sqrt{\frac{T}{N}} \tag{15.3.1186}$$

that we name the "physical probability" for obvious reasons we will see further below.

Oufff!!!! This is done! Now let us come back to the CRR model:

$$C^{\text{CRR}}=e^{-r_fN}\sum_{i=0}^{N}\binom{N}{i}p^i(1-p)^{N-i}\max(S_0u^i(1-u)^{N-i}-K,0) \tag{15.3.1187}$$

When the final stock price is greater than the strike price, that is, when:

$$S_0u^id^{N-i}>K \tag{15.3.1188}$$

Or:

$$\ln\left(\frac{S_0}{K}\right)>-i\ln(u)-(N-i)\ln(d) \tag{15.3.1189}$$

Since:

$$u=e^{\sigma\sqrt{T/N}}\qquad\text{and}\qquad d=e^{-\sigma\sqrt{T/N}} \tag{15.3.1190}$$

The conditions become:

$$\ln\left(\frac{S_0}{K}\right)=N\sigma\sqrt{\frac{T}{N}}-2i\sigma\sqrt{\frac{T}{N}} \tag{15.3.1191}$$

or:

$$j>\frac{N}{2}-\frac{\ln(S_0/K)}{2\sigma\sqrt{T/N}} \tag{15.3.1192}$$

The CRR model can therefore we written:

$$C^{\text{CRR}}=e^{-r_f N}\sum_{i>\alpha}\binom{N}{i}p^i(1-p)^{N-i}(S_0u^i(1-u)^{N-i}-K) \tag{15.3.1193}$$

where:

$$\alpha=\frac{N}{2}-\frac{\ln(S_0/K)}{2\sigma\sqrt{T/N}} \tag{15.3.1194}$$

For convenience, we put:

$$\begin{aligned}U_1&=\sum_{i>\alpha}\binom{N}{i}p^i(1-p)^{N-i}u^i(1-u)^{N-i}\\U_2&=\sum_{i>\alpha}\binom{N}{i}p^i(1-p)^{N-i}\end{aligned} \tag{15.3.1195}$$

So that:

$$C^{\text{CRR}}=e^{-r_f N}(S_0U_1-KU_2) \tag{15.3.1196}$$

Let us consider first U_2 as it is well known (see section Statistics page 658); a binomial distribution approaches a Normal distribution as the number of trials approaches infinity according to the central limit theorem (see section Statistics page 938). Specifically, when there are N trials and p is the probability of success, the probability distribution of the number of successes is approximately Normal with mean Np and standard deviation $\sqrt{Np(1-p)}$ (see section Statistics page 692). The variable U_2 above is the probability of the number of successes being more than α. From the properties of the Normal distribution, it follows that, for large N:

$$U_2=\Phi\left(\frac{Np}{\sqrt{Np(1-p)}}\right) \tag{15.3.1197}$$

Substituting for α, we get:

$$U_2=\Phi\left(\frac{\ln(S_0/K)}{2\sigma\sqrt{T}\sqrt{p(1-p)}}+\frac{\sqrt{N}\left(p-\dfrac{1}{2}\right)}{\sqrt{p(1-p)}}\right) \tag{15.3.1198}$$

From the result we get previously:

$$p=\frac{1}{2}+\frac{\mu}{2\sigma}\sqrt{\frac{T}{N}} \tag{15.3.1199}$$

we see that as N tends to infinity:

$$p=\frac{1}{2} \tag{15.3.1200}$$

and therefore:

$$\begin{gathered}p(1-p)=p-p^2=\frac{1}{2}-\frac{1}{4}=\frac{1}{4}\\ \sqrt{N}\left(p-\frac{1}{2}\right)=\sqrt{N}\left(\left(\frac{1}{2}+\frac{\mu}{2\sigma}\sqrt{\frac{T}{N}}\right)-\frac{1}{2}\right)=\frac{\mu}{2\sigma}\sqrt{T}=0\end{gathered} \tag{15.3.1201}$$

and here we see a big problem! The use of physical probability bring us to nothing we expected. We should be able to have an expression with μ, r_f and σ as we know it with the Black & Scholes model since these three parameters appear in the Φ of that latter.

So the idea now is the following: The risk free return r in a non-arbitrage assumption should be equivalent to the expected mean return of the asset. That means under this assumption we can write:

$$\boxed{r=\frac{puS_0+(1-p)dS_0}{S_0}=pu+(1-p)d} \tag{15.3.1202}$$

which implies after rearrangement:

$$p=\frac{r-d}{u-d} \tag{15.3.1203}$$

denoted p_r to make the difference with the "physical probability" introduced earlier before and named the "risk-neutral probability"!

With a continuous rate this is immediately written:

$$p=\frac{e^{r_f t}-d}{u-d} \tag{15.3.1204}$$

The reader may have notice that p is not necessary in $[0,1]$ unless t (i.e. Δt further below) is small enough!

Remark

Unlike investing in stocks, investors do not have to make an upfront payment to take a position in a Futures contract. In a risk-neutral sense, the expected growth rate from holding a Futures contract is zero. That is in the above relation you just have to put $r_f=0$.

An equivalent way to introduce this probability is just to write that in a non-arbitrage universe, the risk neutral expected value of S equals to the binomial expected value of S:

$$S_0e^{r_f\Delta t}=pS_u+(1-p)S_d \tag{15.3.1205}$$

Written in the way of a conditional expectation, we have[29]:

$$\text{E}(S_{t+\Delta t}|S_t)=S_t(pu+(1-p)d)=S_te^{r_f\Delta t} \tag{15.3.1206}$$

---

[29] From which we deduce on the way the conditional variance: $\text{V}\left(S_{t+\Delta t}|S_t\right)=\text{E}\left(S^2_{t+\Delta t}|S_t\right)-\left(\text{E}\left(S_{t+\Delta t}|S_t\right)\right)=p\left(S_tu\right)^2+(1-p)\left(S_td\right)^2-S_t^2\left(pu+(1-p)\,d\right)^2$

Solving for p yields immediately:

$$p=p_r=\frac{e^{r_f\Delta t}-d}{u-d} \tag{15.3.1207}$$

So finally in the CRR universe:

$$\boxed{u=\frac{1}{d}=e^{\sigma\sqrt{T/N}} \qquad p_r=\frac{e^{r_fT/N}-d}{u-d}} \tag{15.3.1208}$$

CRR recognize that the only condition required to prevent arbitrage is the equivalent of this set of relations!

So this done let us come back to:

$$U_2=\Phi\left(\frac{\ln(S_0/K)}{2\sigma\sqrt{T}\sqrt{p(1-p)}}+\frac{\sqrt{N}\left(p-\dfrac{1}{2}\right)}{\sqrt{p(1-p)}}\right) \tag{15.3.1209}$$

As:

$$p=p_r=\frac{e^{r_fT/N}-d}{u-d}=\frac{e^{r_fT/N}-e^{-\sigma\sqrt{T/N}}}{e^{\sigma\sqrt{TN/}}-e^{-\sigma\sqrt{T/N}}} \tag{15.3.1210}$$

By expanding the exponential functions in a Maclaurin series (see section Sequences and Series page 1703) to the first order:

$$e^x\cong 1+x \tag{15.3.1211}$$

we therefore have:

$$p\cong\frac{\left(1+r_f\dfrac{T}{N}\right)-\left(1-\sigma\sqrt{\dfrac{T}{N}}\right)}{\left(1+\sigma\sqrt{\dfrac{T}{N}}\right)-\left(1-\sigma\sqrt{\dfrac{T}{N}}\right)}=\frac{r_f\dfrac{T}{N}+\sigma\sqrt{\dfrac{T}{N}}}{2\sigma\sqrt{\dfrac{T}{N}}}=\frac{\sigma+r_f\sqrt{\dfrac{T}{N}}}{2\sigma} \tag{15.3.1212}$$

Therefore:

$$\begin{aligned}p(1-p)&\cong\left(\frac{\sigma+r_f\sqrt{\dfrac{T}{N}}}{2\sigma}\right)\left(1-\left(\frac{\sigma+r_f\sqrt{\dfrac{T}{N}}}{2\sigma}\right)\right)=\left(\frac{\sigma+r_f\sqrt{\dfrac{T}{N}}}{2\sigma}\right)\left(\frac{\sigma-r_f\sqrt{\dfrac{T}{N}}}{2\sigma}\right)\\&=\frac{\sigma^2-r_f^2\dfrac{T}{N}}{4\sigma^2}\end{aligned} \tag{15.3.1213}$$

As N tends to infinity:

$$p(1-p)\cong\frac{1}{4}\Rightarrow\sqrt{p(1-p)}=\frac{1}{2} \tag{15.3.1214}$$

And second order Taylor development would lead exactly to the same asymptotic value!

For the remaining part to fall back on the Black & Scholes model we need second order Maclaurin development. This is quite boring algebra so let us do it with Maple™ 4.00b:

```
 >p:=(taylor(exp(x),x=0,3)-taylor(exp(-y),y=0,3))/(taylor(exp(y),y=0,3)
-taylor(exp(-y),y=0,3));
>x:=r*T/N;y:=-sigma*sqrt(T/N);
>limit(sqrt(N)*(p-1/2),N=infinity);
```

and this gives:

$$\sqrt{N}\left(p-\frac{1}{2}\right)\cong\frac{(r_f-\sigma^2/2)\sqrt{T}}{2\sigma} \qquad (15.3.1215)$$

Therefore:

$$\begin{aligned}U_2&\cong\Phi\left(\frac{\ln(S_0/K)}{2\sigma\sqrt{T}\sqrt{p(1-p)}}+\frac{\sqrt{N}\left(p-\frac{1}{2}\right)}{\sqrt{p(1-p)}}\right)=\Phi\left(\frac{\ln(S_0/K)}{2\sigma\sqrt{T}\frac{1}{2}}+\frac{\frac{(r_f-\sigma^2/2)\sqrt{T}}{2\sigma}}{\frac{1}{2}}\right)\\&=\Phi\left(\frac{\ln(S_0/K)}{\sigma\sqrt{T}}+\frac{(r_f-\sigma^2/2)\sqrt{T}}{\sigma}\right)=\Phi\left(\frac{\ln(S_0/K)+(r_f-\sigma^2/2)T}{\sigma\sqrt{T}}\right)\end{aligned} \qquad (15.3.1216)$$

Ok this is finish about U_2. We now move on to evaluate U_1 given for recall by:

$$U_1=\sum_{i>\alpha}\binom{N}{i}p^i(1-p)^{N-i}u^i(1-u)^{N-i}=\sum_{i>\alpha}\binom{N}{i}(pu)^i((1-p)d)^{N-i} \qquad (15.3.1217)$$

We put:

$$p^*=\frac{pu}{pu+(1-p)d} \qquad (15.3.1218)$$

It then follows that:

$$1-p^*=\frac{(1-p)pd}{pu+(1-p)d} \qquad (15.3.1219)$$

and we can then write U_1 as following:

$$U_1=[pu+(1-p)d]^N\sum_{i>\alpha}\binom{N}{i}(p^*)^i(1-p^*)^{N-i} \qquad (15.3.1220)$$

Since:

$$pu+(1-p)d=e^{rT/N}=e^{r\Delta t} \qquad (15.3.1221)$$

we have:

$$U_1=e^{r_fT}\sum_{i>\alpha}\binom{N}{i}(p^*)^i(1-p^*)^{N-i} \qquad (15.3.1222)$$

This shows that U_1 also involves a binomial distribution where the probability on an up movement is p^* rather than p. Approximating the binomial distribution with a Normal distribution, we get, similarly to U_2:

$$U_1 = e^{r_f T}\mathcal{N}\left(\frac{Np^* - \alpha}{\sqrt{Np^*(1-p^*)}}\right) \tag{15.3.1223}$$

and substituting for α, gives as with U_2:

$$U_1 = e^{r_f T}\Phi\left(\frac{\ln(S_0/K)}{2\sigma\sqrt{T}\sqrt{p^*(1-p^*)}} + \frac{\sqrt{N}\left(p^* - \dfrac{1}{2}\right)}{\sqrt{p^*(1-p^*)}}\right) \tag{15.3.1224}$$

Substituting for u and d, and the risk-neutral probability in p^* we get:

$$\begin{aligned}
p^* &= \frac{pu}{pu+(1-p)d} = \frac{\left(\dfrac{e^{r_f T/N}-d}{u-d}\right)u}{\left(\dfrac{e^{r_f T/N}-d}{u-d}\right)u + \left(1-\left(\dfrac{e^{r_f T/N}-d}{u-d}\right)\right)d}\\
&= \frac{\left(\dfrac{e^{r_f T/N}-d}{u-d}\right)u}{\left(\dfrac{e^{r_f T/N}-d}{u-d}\right)u + \left(\dfrac{u-e^{r_f T/N}}{u-d}\right)d} = \frac{(e^{r_f T/Nt}-d)u}{(e^{r_f T/N}-d)u+(u-e^{r_f T/N})d}\\
&= \frac{(e^{r_f T/N}-e^{-\sigma\sqrt{T/N}})e^{\sigma\sqrt{T/N}}}{(e^{r_f T/N}-e^{-\sigma\sqrt{T/N}})e^{\sigma\sqrt{T/N}}+(e^{\sigma\sqrt{T/N}}-e^{r_f T/N})e^{-\sigma\sqrt{T/N}}}\\
&= \frac{e^{r_f T/N}-e^{-\sigma\sqrt{T/N}}}{e^{r_f T/N}e^{\sigma\sqrt{T/N}}-1+1-e^{r_f T/N}e^{-\sigma\sqrt{T/N}}}e^{\sigma\sqrt{T/N}}\\
&= \frac{e^{r_f T/N}-e^{-\sigma\sqrt{T/N}}}{e^{\sigma\sqrt{T/N}}-e^{-\sigma\sqrt{T/N}}}\frac{e^{\sigma\sqrt{T/N}}}{e^{r_f T/N}}
\end{aligned} \tag{15.3.1225}$$

By expanding the exponential functions in a Maclaurin series of order 2 we see that, as N tend to infinity that (we can detail the developments on readers requests):

$$p^*(1-p^*) \cong \frac{1}{4} \tag{15.3.1226}$$

and:

$$\sqrt{N}\left(p^* - \frac{1}{2}\right) \cong \frac{(r_f+\sigma^2/2)\sqrt{T}}{2\sigma} \tag{15.3.1227}$$

with the result that:

$$U_1 = e^{r_f T}\mathcal{N}\left(\frac{\ln(S_0/K)+(r_f+\sigma^2/2)T}{\sigma\sqrt{T}}\right) \tag{15.3.1228}$$

So finally:

$$C^{\text{CRR}} = e^{-r_f T}(S_0U_1 - KU_2) = S_0\mathcal{N}(d_1) - Ke^{r_f T}\mathcal{N}(d_2) \tag{15.3.1229}$$

where:

$$d_1 = \frac{\ln(S_0/K) + (r_f + \sigma^2/2)T}{\sigma\sqrt{T}} \tag{15.3.1230}$$

and:

$$d_2 = \frac{\ln(S_0/K) + (r_f - \sigma^2/2)T}{\sigma\sqrt{T}} = d_1 - \sigma\sqrt{T} \tag{15.3.1231}$$

So we see again now that for a risk-neutral option valuation, p and μ are unnecessary, exactly as provided by the Black & Scholes pricing model. The fact that the probability p doesn't appear in this formula means that even if investors have different subjective probabilities for the up and down movements, they do agree on the relation between Call, S and r. So the option's value doesn't depend on the investor's attitude toward risk.

Continuous-time models are admittedly more complicated that their discrete-time counterparts. Nonetheless there are a number of good reasons to deal with them. To begin with, on many markets with very frequent trading the assumption of continuous security tradings is closer to reality than assuming that markets are open only at fixed time points such as once a day. Moreover, in continuous-time models we can often get closed form solutions for derivatives prices which are not available in discrete models. Finally continuous-time modelling is the state-of-the art in the 120th and 121st century (holocene calendar) literature.

So for recall the assumptions (hypothesis) of that model are:

H1. We are in a non-arbitrage context with fixed probability p_r

H2. The price of the asset today is S_0 and known

H3. The assets do not pay dividends

H4. The risk-free rate r is positive and $d < r < u$

However Binomial options pricing model approach is widely used since it is able to handle a variety of conditions for which other models cannot easily be applied. This is largely because the binomial option pricing model is based on the description of an underlying instrument over a period of time rather than a single point. As a consequence, it is used to value American options that are exercisable at any time in a given interval as well as Bermudan options that are exercisable at specific instances of time. Being relatively simple, the model is readily implementable in computer software (including a spreadsheet).

Although computationally slower than the Black & Scholes formula, it is more accurate, particularly for longer-dated options on securities with dividend payments. For these reasons, various versions of the binomial model are widely used by practitioners in the options markets.

15.3.10.3.11 Bachelier option pricing model

We want to illustrate now Louis Bachelier's efforts to obtain applicable formulas for option pricing in pre-computer time. Bachelier's model yields good short-time approximations of prices and volatilities and some banks still use that model in the beginning of the 121st century (holocene calendar) especially in areas where negative interest rates applies...

Remark

On April 8 12020 (holocene calendar), the CME Group posted the note *CME Clearing Plan to Address the Potential of a Negative Underlying in Certain Energy Options Contracts*, saying that after a threshold on price, it would change its standard energy options model from one based on Geometric Brownian Motion and the Black–Scholes model to the Bachelier model! On April 20 2020, oil futures reached negative values for the first time in history, where Bachelier model took an important role in option pricing and risk management.

So let us first recall that we get during our study of Bachelier's speculation theory (see page 7689) that:

$$p=\mathcal{N}\left(0,\sigma\sqrt{t}\right)=\frac{1}{\sigma\sqrt{2\pi t}}e^{-\frac{x^2}{2\sigma^2 t}} \tag{15.3.1232}$$

Let us fix a strike price K, a horizon T and consider the European Call C, whose payoff at time T is modelled by the random variable:

$$C_T^{\text{B}}=(S_T-K)^+=\max(S_T-K,0) \tag{15.3.1233}$$

Applying Bachelier's fundamental principle (zero mean and standard deviation proportional to the square root of time) we get for the price of the option at time $t=0$ considering the point of view of Bachelier that if we write the price of underlying as $S_T=S_0+x$ then the Call price is therefore given by:

$$C_T^{\text{B}}=(S_T-K)^+=\max((S_0+x)-K,0) \tag{15.3.1234}$$

and as according to Bachelier assumptions we have the amount of profit that follow:

$$p=(S_0+x)-K=\mathcal{N}\left(0,\sigma\sqrt{t}\right)=\frac{1}{\sigma\sqrt{2\pi t}}e^{-\frac{x^2}{2\sigma^2 t}} \tag{15.3.1235}$$

and that therefore the integration bounds are $K-S_0$ (indeed, this is the limit above which

$S_0+x-K\geq 0$) to $+\infty$:

$$\begin{aligned}C^{\mathrm{B}}(t=0)=\mathrm{E}\left(\max(S_T-K,0)\right)&=\int\limits_{K-S_0}^{+\infty}(S_0+x-K)\frac{1}{\sigma\sqrt{2\pi T}}e^{-\frac{x^2}{2\sigma^2T}}\mathrm{d}x\\&=(S_0-K)\int\limits_{K-S_0}^{+\infty}\frac{1}{\sigma\sqrt{T}\sqrt{2\pi}}e^{-\frac{x^2}{2(\sigma\sqrt{T})^2}}\mathrm{d}x+\int\limits_{K-S_0}^{+\infty}\frac{x}{\sigma\sqrt{T}\sqrt{2\pi}}e^{-\frac{x^2}{2(\sigma\sqrt{T})^2}}\mathrm{d}x\\&=(S_0-K)\int\limits_{-\infty}^{S_0-K}\frac{1}{\sigma\sqrt{T}\sqrt{2\pi}}e^{-\frac{x^2}{2(\sigma\sqrt{T})^2}}\mathrm{d}x+\int\limits_{K-S_0}^{+\infty}\frac{x}{\sigma\sqrt{T}\sqrt{2\pi}}e^{-\frac{x^2}{2(\sigma\sqrt{T})^2}}\mathrm{d}x\\&=(S_0-K)\Phi\left(\frac{S_0-K}{\sigma\sqrt{T}}\right)+\int\limits_{K-S_0}^{+\infty}\frac{x}{\sigma\sqrt{T}\sqrt{2\pi}}e^{-\frac{x^2}{2(\sigma\sqrt{T})^2}}\mathrm{d}x\end{aligned}\tag{15.3.1236}$$

as for recall (see section Statistics page 692):

$$P(x\leq X)=F(x)=\Phi\left(\frac{x-\mu}{\sigma}\right)\tag{15.3.1237}$$

Let us do now the change of variable:

$$y=\frac{x}{\sqrt{2}\sigma\sqrt{T}}\tag{15.3.1238}$$

Therefore:

$$\int\limits_{K-S_0}^{+\infty}\frac{x}{\sigma\sqrt{T}\sqrt{2\pi}}e^{-\frac{x^2}{2(\sigma\sqrt{T})^2}}\mathrm{d}x=\int\limits_{\frac{K-S_0}{\sqrt{2}\sigma\sqrt{T}}}^{+\infty}\frac{y}{\sqrt{\pi}}e^{-y^2}\sqrt{2}\sigma\sqrt{T}\mathrm{d}y=\frac{\sqrt{2}\sigma\sqrt{T}}{\sqrt{\pi}}\int\limits_{\frac{K-S_0}{\sqrt{2}\sigma\sqrt{T}}}^{+\infty}ye^{-y^2}\mathrm{d}y\tag{15.3.1239}$$

Let us now put $u=-y^2$. It follows that $\mathrm{d}u=-2y\mathrm{d}y$. Therefore:

$$\begin{aligned}\frac{\sqrt{2}\sigma\sqrt{T}}{\sqrt{\pi}}\int\limits_{\frac{K-S_0}{\sqrt{2}\sigma\sqrt{T}}}^{+\infty}ye^{-y^2}\mathrm{d}y=\frac{\sqrt{2}\sigma\sqrt{T}}{\sqrt{\pi}}\int\limits_{\frac{K-S_0}{\sqrt{2}\sigma\sqrt{T}}}^{+\infty}-\frac{1}{2}\left(-2ye^{-y^2}\right)\mathrm{d}y&=\frac{\sqrt{2}\sigma\sqrt{T}}{\sqrt{\pi}}\int\limits_{-\frac{(K-S_0)^2}{2\sigma^2T}}^{-\infty}-\frac{1}{2}e^u\mathrm{d}u\\&=-\frac{\sqrt{2}\sigma\sqrt{T}}{2\sqrt{\pi}}e^u\Bigg|_{-\frac{(K-S_0)^2}{2\sigma^2T}}^{-\infty}=-\frac{\sqrt{2}\sigma\sqrt{T}}{2\sqrt{\pi}}\left(-e^{-\frac{(K-S_0)^2}{2\sigma^2T}}\right)\\&=\sigma\sqrt{T}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{S_0-K}{\sigma\sqrt{T}}\right)^2}\end{aligned}\tag{15.3.1240}$$

Hence:

$$\begin{aligned}C^{\mathrm{B}}(t=0)&=(S_0-K)\Phi\left(\frac{S_0-K}{\sigma\sqrt{T}}\right)+\int\limits_{K-S_0}^{+\infty}\frac{x}{\sigma\sqrt{T}\sqrt{2\pi}}e^{-\frac{x^2}{2(\sigma\sqrt{T})^2}}\mathrm{d}x\\&=(S_0-K)\Phi\left(\frac{S_0-K}{\sigma\sqrt{T}}\right)+\sigma\sqrt{T}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{S_0-K}{\sigma\sqrt{T}}\right)^2}\\&=(S_0-K)\Phi\left(\frac{S_0-K}{\sigma\sqrt{T}}\right)+\sigma\sqrt{T}\phi\left(\frac{S_0-K}{\sigma\sqrt{T}}\right)\end{aligned}\tag{15.3.1241}$$

where $\phi(x)$ denotes for recall the density of the standard normal distribution.

Finally:

$$\boxed{C^{\text{B}}(t=0)=(S_0-K)\Phi\left(\frac{S_0-K}{\sigma\sqrt{T}}\right)+\sigma\sqrt{T}\phi\left(\frac{S_0-K}{\sigma\sqrt{T}}\right)} \tag{15.3.1242}$$

Interestingly, Bachelier explicitly wrote down the formula:

$$C^{\text{B}}(t=0)=\int\limits_{K-S_0}^{+\infty}(S_0+x-K)\frac{1}{\sigma\sqrt{2\pi T}}e^{-\frac{x^2}{2\sigma^2 T}}\mathrm{d}x \tag{15.3.1243}$$

but did not bother to spell out formula:

$$C^{\text{B}}(t=0)=(S_0-K)\Phi\left(\frac{S_0-K}{\sigma\sqrt{T}}\right)+\sigma\sqrt{T}\phi\left(\frac{S_0-K}{\sigma\sqrt{T}}\right) \tag{15.3.1244}$$

The main reason seems to be that at his time option prices – at least in Paris – were quoted the other way around: while today the strike prices K is fixed and the option price fluctuates according to supply and demand, at Bachelier's times the option prices were fixed (at 10, 20 and 50 cents for a "rent", i.e., a perpetual bond with par value of 100.-) and therefore the strike prices K fluctuated. What Bachelier really needed was the inverse version of the above relation between the option price C_B^0 and the strike price K.

Caution! Remember for later that the σ of Bachelier is not the same as the σ_{BS} of Black & Scholes! σ for Black & Scholes is in percentage terms, when σ for Bachelier is in numeraire (currency) terms!!

15.3.10.3.12 Black Model (Future options pricing model)

The "Black model" (sometimes known as the Black-76 model) is a variant of the Black–Scholes option pricing model. Its primary applications are for pricing options on Future contracts, bond options, Interest rate cap and floors, and swaptions. It was first presented in a paper written by Fischer Black in 11976 (holocene calendar).

Black's model can be generalized into a class of models known as log-normal Forward models, also referred to as LIBOR market model.

The Black formula is therefore similar to the Black-Scholes formula for valuing stock options except that the spot price of the underlying is replaced by a discounted Futures price F.

Let us recall that in the case of a Call on an underlying type of share we have its price at any time that is given by:

$$C=\max(S_t-K,0) \tag{15.3.1245}$$

While for a Future (corresponding to a Forward for the conditions we had demonstrated earlier above) we have:

$$C_F=\max(F_t-K,0) \tag{15.3.1246}$$

But we have proved in our very brief introduction on Forwards and Futures that:

$$F_t=S_te^{(r-y)\tau}=S_te^{(r-y)(T-t)} \tag{15.3.1247}$$

So since we solved the Black & Scholes equation just before for S_t it is enough to use the previous relation which gives us:

$$S_t=F_te^{(r-y)(T-t)} \tag{15.3.1248}$$

Therefore, we have:

$$C_F^{\text{Bk}}=\max\left(F_te^{-(r-y)(T-t)}-K,0\right)=e^{-(r-y)(T-t)}\max\left(F_t-K^{(r-y)(T-t)},0\right) \tag{15.3.1249}$$

So a Call option on a future can be seen as a Call on a share underlying type that does not pay dividends but in quantity $e^{(r-y)(T-t)}$ and with strike instead $K^{-(r-y)(T-t)}$ instead of K.

If we write the Call function of an underlying that does not pay dividends in the following way (the order of the arguments is chosen to be the same as in MATLAB® and as most software do not manage other calculations than those based explicitly on $t=0$ and $y=0$):

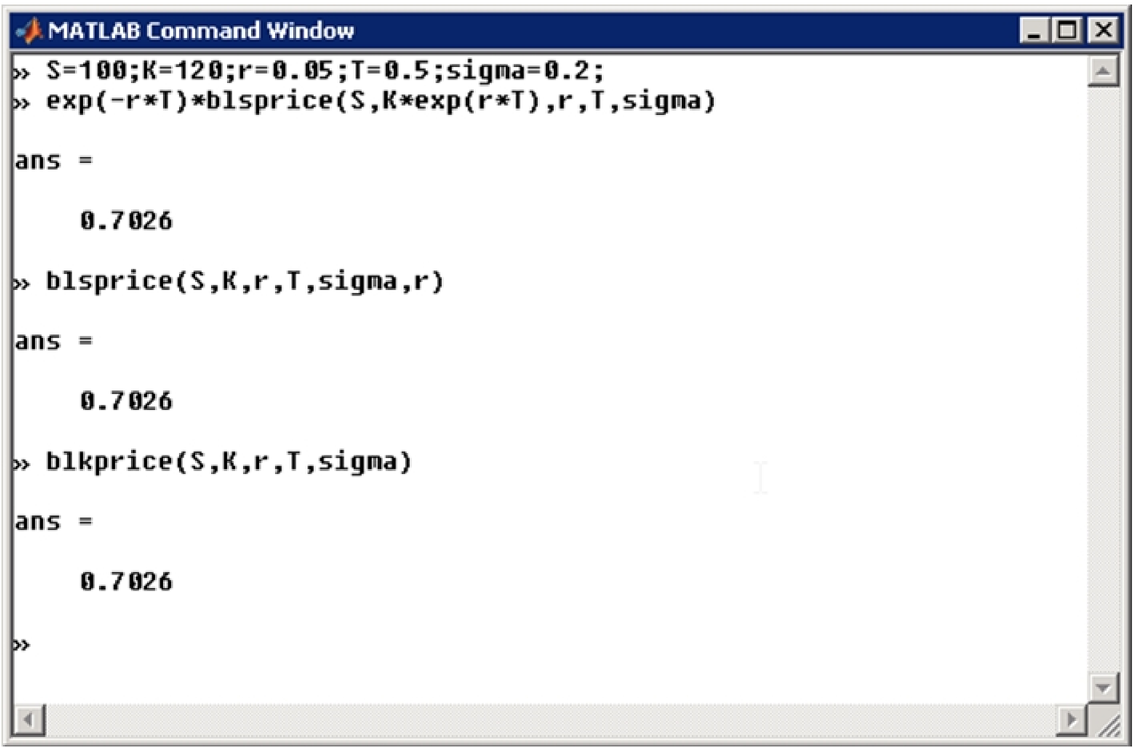
$$C(S_0,K,r,T,\sigma) \tag{15.3.1250}$$

So in view of the above we see immediately that a Call on an underlying of type Future type can be obtained from a Call on an underlying of Share type by writing:

$$C_F=e^{-rT}C(F_0,Ke^{rT},r,T,\sigma) \tag{15.3.1251}$$

This last implicit expression (and also the explicit version) constitutes what we name the "Black model" or also says "Black-76 model" for a Call (the idea for a Put is the same!).

This equivalence can easily be checked for example with MATLAB® (so you do not have to waste time with Microsoft® Excel®):



```
>> S=100;K=120;r=0.05;T=0.5;sigma=0.2;
>> exp(-r*T)*blsprice(S,K*exp(r*T),r,T,sigma)

ans =

    0.7026

>> blsprice(S,K,r,T,sigma,r)

ans =

    0.7026

>> blkprice(S,K,r,T,sigma)

ans =

    0.7026

>>
```

Figure 15.191: **Black model with MATLAB® 2016a**

Now let us recall the explicit Black & Scholes relation for a Call paying no dividends:

$$C^{\text{BS}}(S,t)=S\Phi\left(\frac{\ln\left(\frac{S}{K}\right)+\left(r+\frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}}\right)-Ke^{-r(T-t)}\Phi\left(\frac{\ln\left(\frac{S}{K}\right)+\left(r-\frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}}\right)$$

(15.3.1252)

Therefore the Black model is explicitly:

$$\boxed{C_F^{\text{Bk}}(S,t)=e^{-rT}\left[S\Phi\left(\frac{\ln\left(\frac{F}{Ke^{rT}}\right)+\left(r+\frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}}\right)-Ke^{rT}e^{-r(T-t)}\Phi\left(\frac{\ln\left(\frac{S}{Ke^{rT}}\right)+\left(r-\frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}}\right)\right]}$$

(15.3.1253)

15.3.10.3.13 VIX volatility index

The "Volatility IndeX" (VIX), also sometimes named "fear index" (....), plays a very important role in the financial derivatives pricing, trading, risk control strategy. The volatility index is a weighted average of implied volatilities for volatilities for several series of Puts and Calls options on a particular index (i.e. expected volatility). Obviously there wouldn't be a financial market without the financial market volatility. If there is a lack of risk management tools and the market volatility is too large, the investors may be worried about the risk and give up trading, then the market is less attractive.

VIX plays a more and more important role in the financial derivatives pricing, trading, risk control strategy. After the global stock market crash in 11987 (holocene calendar), it is to stabilize the stock market and protect the investors, that the New York Stock Exchange (NYSE) in 11990 (holocene calendar) introduced a circuit breaker mechanism (Circuit-breakers). When the stock price changes unusually, it occurs a temporary suspension of trading, and it is helpful to try to reduce the market volatility in order to restore the investor confidence on the stock market. However, due to the introduction of circuit breaker mechanism, there is many new insights for how to measure market volatility, and it is gradually producing a dynamic display of market volatility requirements. Therefore, not long after the New York Stock Exchange (NYSE) used Circuit-Breakers to solve the problem of excessive volatility in the market, the Chicago Board Options Exchange began to introduce the CBOE market volatility index[30] (VIX) in 11993 (holocene calendar), which is used to measure market volatility implied by at the money S&P 100 Index (OEX) option prices.

When the stock option transactions began in April 11973 (holocene calendar), Chicago Board Options Exchange (CBOE) has envisaged that the market volatility index can be constructed by the option price, which could be shown that the expectation of the future volatility in the option market. Since then there were gradually many various calculation methods proposed by some scholars, Whaley (11993 according to holocene calendar) proposed a calculation approach which is the preparation of market volatility index as a measure of future stock market price volatility. In the same year, Chicago Board Options Exchange (CBOE) started to research the compilation of the CBOE Volatility Index (VIX), which is based on the implied volatility of the S&P 100 Index options, and at the same time also calculate implied volatility of Call option and Put option in order to take into account the right of traders to buy or sell option under the preferences.

After ten years of development and improvement, the VIX index gradually was agreed by the stock market, CBOE calculated several other volatility indexes including, in 12001 (holocene calendar) NASDAQ 100 index as the underlying volatility index (NASDAQ Volatility Index, VXN), in 12003 (holocene calendar) the VIX index based on the S&P 500 Index, which is much closer to the actual stock market than the S&P 100 Index, CBOE DJIA volatility Index (VXD), CBOE Russell 2000 Volatility Index (RVX), in 12004 (holocene calendar) the first volatility futures (Volatility Index Futures) VIX Futures, and at the same year a second volatility commercialization futures, that is the variance futures (Variance Futures), subject to three month. the S&P 500 Index of realized Variance (Realized Variance). In 12006 (holocene calendar), the VIX Index options began to trade in the Chicago Board Options Exchange. In 12008 (holocene calendar), CBOE pioneered the used of the VIX methodology to estimate the expected volatility of some commodities and foreign currencies. There are many developments. For example, in India, VIX was launched in April, 12008 (holocene calendar) by National stock exchange (NSE). The VIX index of India is based on the Nifty 50 Index Option prices. The methodology of calculating the VIX index is same as that for CBOE VIX index. The current focus on the VIX is due to its inherent property of negative correlation with the underlying price index, and its usefulness for predicting the direction of the price index. Hong Kong got its own volatility index for financial products that allow investors to hedge against excessive market movements. Hang Seng Indexes Company, the company which owns and manages the benchmark indexes in Hong Kong including the Hang Seng index (HSI), launched the HSI Volatility index or "VHSI" on February 21. The index is modelled on the lines of the Chicago Board of Exchanges VIX index. VIX in that it measures the 30-calendar-day expected volatility of the Hang Seng index using prices of options traded on the index.

---

[30] The official white-paper can be found here `https://cdn.cboe.com/resources/vix/vixwhite.pdf`

In order to apply most financial models in practice, it is necessary to be able to use empirical data to measure the degree of variability of asset prices or market indexes. Suppose that S_t is the price of an asset or a market index at time t. The annualized realized annual volatility (see above page 7741) of this asset in a period $[t_1, t_2]$ based on $n+1$ daily observations $S_0, S_1, \ldots, S_{n-1}, S_n$ will be chosen according to the Bachelier temporal evolution of volatility model (page 7689):

$$\bar{\sigma} = \sqrt{\frac{252}{n-1}\sum_{i=1}^{n} r_i^2} \tag{15.3.1254}$$

where:

$$r_i = \ln\left(\frac{S_i}{S_{i-1}}\right), \qquad i = 1, 2, 3, \ldots, n-1, n. \tag{15.3.1255}$$

And 252 if for recall an annualization factor corresponding to the typical number of trading days in a year.

Similarly the historical volatility (see above page 7740)) is defined by a similar relation:

$$\tilde{\sigma} = \sqrt{\frac{252}{n-1}\sum_{i=1}^{n} (r_i - \bar{r})^2} \tag{15.3.1256}$$

where:

$$\bar{r} = \frac{1}{n}\sum_{i=1}^{n} r_i \tag{15.3.1257}$$

is the mean return.

If the returns are supposed to be drawn independently from the same probability distribution, then $\bar{r}$ is the sample mean and the historical volatility is simply the annualized sample standard deviation. The realized volatility is then the annualized sample second moment.

Note that:

$$\sum_{i=1}^{n} (r_i - \bar{r})^2 = \sum_{i=1}^{n} r_i^2 - n\bar{r}^2 \tag{15.3.1258}$$

and hence:

$$\tilde{\sigma}^2 = \bar{\sigma}^2 - \frac{252n\bar{r}}{n-1} \tag{15.3.1259}$$

As we already know, this simply means that the realized volatility is approximately equal to the historical volatility if the sample mean is very close to zero.

CBOE launched the first VIX index (VXO) in 11993 (holocene calendar), which is based on the option model which is proposed by Black, Scholes and Merton (see above page 7804). In this model, except the volatility, we also require other many parameters such as the current stock price, the option price, the strike price, duration, risk-free interest rate, time to expected payment of cash dividends and amount of expected payment of cash dividends. However, the S&P 100 options of CBOE are American options and are related to the cash dividends of the underlying stocks, so when CBOE calculated the VIX index, they used the implied option volatility of the binomial model proposed by Cox, Ross and Rubinstein (see above page 7833).

To estimate the volatility from option data (all the text that follows is completely inspired from [582]) we assume first that the stock price satisfies the following stochastic differential equation:

$$\frac{\mathrm{d}S}{S} = (r-q)\mathrm{d}t + \sigma \mathrm{d}z \tag{15.3.1260}$$

Where r, q, σ are constant denoting the risk free rate, the continuous divided yield and yield volatility respectively, and z is a Wiener process with respect to a risk-neutral probability measure.

At the same time, from the Ito's lemma (see above page 7779):

$$\mathrm{d}\ln(S) = \left(r - q - \frac{\sigma^2}{2}\right)\mathrm{d}t + \sigma \mathrm{d}z \tag{15.3.1261}$$

Then from the both relations above we can obtain:

$$\frac{\mathrm{d}S}{S} - \mathrm{d}\ln(S) = (r-q)\mathrm{d}t + \sigma \cdot \mathrm{d}z - \left(r - q - \frac{\sigma^2}{2}\right)\mathrm{d}t - \sigma \mathrm{d}z = \frac{\sigma^2}{2}\mathrm{d}t \tag{15.3.1262}$$

Therefore:

$$\frac{\sigma^2}{2}\mathrm{d}t = \frac{\mathrm{d}S}{S} - \mathrm{d}\ln(S) \tag{15.3.1263}$$

By integrating from time 0 to time T, we get:

$$\int_0^T \frac{\sigma^2}{2}\mathrm{d}t = \int_0^T \frac{\mathrm{d}S}{S} - \int_0^T \mathrm{d}\ln(S) \tag{15.3.1264}$$

This gives:

$$\frac{1}{2}\sigma^2 T = \int_0^T \frac{\mathrm{d}S}{S} - (\ln(S_T) - \ln(S_0)) = \int_0^T \frac{\mathrm{d}S}{S} - \ln\left(\frac{S_T}{S_0}\right) \tag{15.3.1265}$$

Because the variance is the square of the volatility:

$$\bar{\mathrm{V}}=\sigma^2 \tag{15.3.1266}$$

Then we write obviously:

$$\frac{1}{2}\bar{\mathrm{V}}T=\int\limits_0^T\frac{\mathrm{d}S}{S}-\ln\left(\frac{S_T}{S_0}\right) \tag{15.3.1267}$$

Therefore:

$$\bar{\mathrm{V}}=\frac{2}{T}\int\limits_0^T\frac{\mathrm{d}S}{S}-\frac{2}{T}\ln\left(\frac{S_T}{S_0}\right) \tag{15.3.1268}$$

Because we can take the expectation of:

$$\frac{\mathrm{d}S}{S}=(r-q)\mathrm{d}t+\sigma\mathrm{d}z \tag{15.3.1269}$$

We get that:

$$\mathrm{E}(\bar{\mathrm{V}})=\frac{2}{T}\mathrm{E}\left(\int\limits_0^T\frac{\mathrm{d}S}{S}\right)-\frac{2}{T}\mathrm{E}\left(\ln\left(\frac{S_T}{S_0}\right)\right) \tag{15.3.1270}$$

Has the first term that can therefore be rewritten as:

$$\mathrm{E}\left(\int\limits_0^T\frac{\mathrm{d}S}{S}\right)=\mathrm{E}\left(\int\limits_0^T(r-q)\mathrm{d}t\right)+\mathrm{E}\left(\int\limits_0^T\sigma\mathrm{d}z\right)=(r-q)T \tag{15.3.1271}$$

Hence:

$$\mathrm{E}(\bar{\mathrm{V}})=\frac{2}{T}(r-q)T-\frac{2}{T}\mathrm{E}\left(\ln\left(\frac{S_T}{S_0}\right)\right) \tag{15.3.1272}$$

We know that under Black-Scholes assumptions that:

$$S_T=S_0\exp(X) \tag{15.3.1273}$$

where the random variables:

$$X=\left(r-q-\frac{\sigma^2}{2}\right)T+\sigma z(T) \tag{15.3.1274}$$

is Normally distributed with:

$$\mathrm{E}(X)=\left(r-q-\frac{\sigma^2}{2}\right)T\qquad\text{and}\qquad\hat{\mathrm{V}}(X)=\sigma^2T \tag{15.3.1275}$$

Then as we already know:

$$\mathrm{E}(S_T)=S_0\exp\left[\mathrm{E}(X)+\frac{\hat{\mathrm{V}}(X)}{2}\right]=S_0e^{(r-q)T} \tag{15.3.1276}$$

And if we define the notation (forward price of the asset for a contract maturing at time T):

$$F_0 = \mathrm{E}(S_T) \tag{15.3.1277}$$

We get therefore:

$$\mathrm{E}(\bar{\mathrm{V}}) = \frac{2}{T}(r-q)T - \frac{2}{T}\mathrm{E}\left(\ln\left(\frac{S_T}{S_0}\right)\right) = \frac{2}{T}\ln\left(\frac{F_0}{S_0}\right) - \frac{2}{T}\mathrm{E}\left(\ln\left(\frac{S_T}{S_0}\right)\right) \tag{15.3.1278}$$

Now before continuing. To express the second term above, we will use a clever trick (i.e. result) that will need a few pages. After we will come back to the above relation.

Let us consider:

$$\int\limits_0^{S^*} \frac{1}{K^2}\max(K-S_T,0)\mathrm{d}K \tag{15.3.1279}$$

where S^* is some value of S.

We consider obviously two cases:

- If $S^* < S_T$ then:

$$\int\limits_0^{S^*} \frac{1}{K^2}\max(K-S_T,0)\mathrm{d}K = 0 \tag{15.3.1280}$$

- If $S^* > S_T$ then:

$$\begin{aligned}\int\limits_0^{S^*} \frac{1}{K^2}\max(K-S_T,0)\mathrm{d}K &= \int\limits_{S_T}^{S^*} \frac{1}{K^2}(K-S_T)\mathrm{d}K \\ &= \left(\ln(K) + \frac{S_T}{K}\right)\Bigg|_{S_T}^{S^*} \\ &= \ln\left(\frac{S^*}{S_T}\right) + \frac{S_T}{S^*} - 1\end{aligned} \tag{15.3.1281}$$

Secondly, we consider next (still to prepare the future trick!):

$$\int\limits_{S^*}^{+\infty} \frac{1}{K^2}\max(S_T-K,0)\mathrm{d}K \tag{15.3.1282}$$

Always considering the same two cases:

- If $S^* > S_T$ then:

$$\int\limits_{S^*}^{+\infty} \frac{1}{K^2}\max(S_T-K,0)\mathrm{d}K = 0 \tag{15.3.1283}$$

- If $S^* < S_T$ then:

$$\begin{aligned}\int\limits_{S^*}^{+\infty}\frac{1}{K^2}\max(S_T-K,0)\mathrm{d}K&=\int\limits_{S^*}^{S_T}\frac{1}{K^2}(S_T-K)\mathrm{d}K\\&=\left(-\frac{S_T}{K}-\ln(K)\right)\Bigg|_{S^*}^{S_T}\\&=-1-\ln(S_T)+\frac{S_T}{S^*}+\ln(S^*)\\&=\ln\left(\frac{S^*}{S_T}\right)+\frac{S_T}{S^*}-1\end{aligned}\tag{15.3.1284}$$

Then we can get that:

$$\int\limits_{0}^{S^*}\frac{1}{K^2}\max(K-S_T,0)\mathrm{d}K+\int\limits_{S^*}^{+\infty}\frac{1}{K^2}\max(S_T-K,0)\mathrm{d}K=\ln\left(\frac{S^*}{S_T}\right)+\frac{S_T}{S^*}-1\tag{15.3.1285}$$

For all value of S^*, so that after rearranging:

$$\ln\left(\frac{S^*}{S_T}\right)=\frac{S_T}{S^*}-1-\int\limits_{0}^{S^*}\frac{1}{K^2}\max(K-S_T,0)\mathrm{d}K-\int\limits_{S^*}^{+\infty}\frac{1}{K^2}\max(S_T-K,0)\mathrm{d}K\tag{15.3.1286}$$

We can take the expectations with respect to the risk-neutral probability measure of the above relation:

$$\mathrm{E}\left(\ln\left(\frac{S^*}{S_T}\right)\right)=\mathrm{E}\left(\frac{S_T}{S^*}-1\right)-\mathrm{E}\left(\int\limits_{0}^{S^*}\frac{1}{K^2}\max(K-S_T,0)\mathrm{d}K\right)-\mathrm{E}\left(\int\limits_{S^*}^{+\infty}\frac{1}{K^2}\max(S_T-K,0)\mathrm{d}K\right)\tag{15.3.1287}$$

We know that under Black-Scholes framework that:

$$\begin{aligned}\mathrm{E}[\max(K-S_T,0)]&=e^{r_fT}P(K)\\\mathrm{E}[\max(S_T-K,0)]&=e^{r_fT}C(K)\end{aligned}\tag{15.3.1288}$$

Where $C(K)$ and $P(K)$ are the prices of the Call and Put option with the strike price K, T is the maturity time and r_f is the risk-free interest rate for a maturity of T.

And remember that we have put previously:

$$F_0=\mathrm{E}(S_T)\tag{15.3.1289}$$

Then we get:

$$\mathrm{E}\left(\ln\left(\frac{S^*}{S_T}\right)\right)=\frac{F_0}{S^*}-1-\int\limits_{0}^{S^*}\frac{1}{K^2}e^{r_fT}P(K)\mathrm{d}K-\int\limits_{S^*}^{+\infty}\frac{1}{K^2}e^{r_fT}C(K)\mathrm{d}K\tag{15.3.1290}$$

According to:

$$\begin{aligned}\ln\left(\frac{S_T}{S_0}\right)&=\ln(S_T)-\ln(S_0)\\&=\ln(S_T)-\ln(S^*)+\ln(S^*)-\ln(S_0)\\&=\ln\left(\frac{S_T}{S^*}\right)+\ln\left(\frac{S^*}{S_0}\right)\end{aligned}\tag{15.3.1291}$$

Taking the expectation of the above relation we get:

$$\begin{aligned}\mathrm{E}\left(\ln\left(\frac{S_T}{S_0}\right)\right)&=\mathrm{E}\left(\ln\left(\frac{S_T}{S^*}\right)\right)+\mathrm{E}\left(\ln\left(\frac{S^*}{S_0}\right)\right)\\&=\ln\left(\frac{S^*}{S_0}\right)+\mathrm{E}\left(\ln\left(\frac{S_T}{S^*}\right)\right)\end{aligned}\tag{15.3.1292}$$

Now we combine the following three relations that we have established earlier above (in order of appearance):

$$\begin{gathered}\mathrm{E}(\bar{\mathrm{V}})=\frac{2}{T}\ln\left(\frac{F_0}{S_0}\right)-\frac{2}{T}\mathrm{E}\left(\ln\left(\frac{S_T}{S_0}\right)\right)\\ \mathrm{E}\left(\ln\left(\frac{S^*}{S_T}\right)\right)=\frac{F_0}{S^*}-1-\int\limits_0^{S^*}\frac{1}{K^2}e^{r_fT}P(K)\mathrm{d}K-\int\limits_{S^*}^{+\infty}\frac{1}{K^2}e^{r_fT}C(K)\mathrm{d}K\\ \mathrm{E}\left(\ln\left(\frac{S_T}{S_0}\right)\right)=\ln\left(\frac{S^*}{S_0}\right)+\mathrm{E}\left(\ln\left(\frac{S_T}{S^*}\right)\right)\end{gathered}\tag{15.3.1293}$$

To get:

$$\begin{aligned}\mathrm{E}(\bar{\mathrm{V}})&=\frac{2}{T}\ln\left(\frac{F_0}{S_0}\right)-\frac{2}{T}\mathrm{E}\left(\ln\left(\frac{S_T}{S_0}\right)\right)\\&=\frac{2}{T}\ln\left(\frac{F_0}{S_0}\right)-\frac{2}{T}\left[\ln\left(\frac{S^*}{S_0}\right)+\mathrm{E}\left(\ln\left(\frac{S_T}{S^*}\right)\right)\right]\\&=\frac{2}{T}\ln\left(\frac{F_0}{S_0}\right)-\frac{2}{T}\ln\left(\frac{S^*}{S_0}\right)-\frac{2}{T}\mathrm{E}\left(\ln\left(\frac{S_T}{S^*}\right)\right)\\&=\frac{2}{T}\ln\left(\frac{F_0}{S_0}\right)-\frac{2}{T}\ln\left(\frac{S^*}{S_0}\right)-\frac{2}{T}\left[\frac{F_0}{S^*}-1-\int\limits_0^{S^*}\frac{1}{K^2}e^{r_fT}P(K)\mathrm{d}K-\int\limits_{S^*}^{+\infty}\frac{1}{K^2}e^{r_fT}C(K)\mathrm{d}K\right]\\&=\frac{2}{T}(\ln(F_0)-\ln(S_0)-\ln(S^*)+\ln(S_0))-\frac{2}{T}\left[\frac{F_0}{S^*}-1-\int\limits_0^{S^*}\frac{1}{K^2}e^{r_fT}P(K)\mathrm{d}K-\int\limits_{S^*}^{+\infty}\frac{1}{K^2}e^{r_fT}C(K)\mathrm{d}K\right]\\&=\frac{2}{T}(\ln(F_0)-\ln(S^*))-\frac{2}{T}\left[\frac{F_0}{S^*}-1-\int\limits_0^{S^*}\frac{1}{K^2}e^{r_fT}P(K)\mathrm{d}K-\int\limits_{S^*}^{+\infty}\frac{1}{K^2}e^{r_fT}C(K)\mathrm{d}K\right]\\&=\frac{2}{T}\ln\left(\frac{F_0}{S^*}\right)-\frac{2}{T}\left[\frac{F_0}{S^*}-1\right]+\frac{2}{T}\left[\int\limits_0^{S^*}\frac{1}{K^2}e^{r_fT}P(K)\mathrm{d}K+\int\limits_{S^*}^{+\infty}\frac{1}{K^2}e^{r_fT}C(K)\mathrm{d}K\right]\end{aligned}\tag{15.3.1294}$$

In conclusion this gives us the expected value of the average variance from time 0 to time T

(under Black-Scholes framework!):

$$\mathrm{E}(\bar{\mathrm{V}})=\frac{2}{T}\ln\left(\frac{F_0}{S^*}\right)-\frac{2}{T}\left[\frac{F_0}{S^*}-1\right]+\frac{2}{T}\left[\int\limits_0^{S^*}\frac{1}{K^2}e^{r_fT}P(K)\mathrm{d}K+\int\limits_{S^*}^{+\infty}\frac{1}{K^2}e^{r_fT}C(K)\mathrm{d}K\right] \tag{15.3.1295}$$

where for recall F_0 is the forward price of the options for a contract maturing time T, $C(K)$ is the prices of the Call option with the strike price K and time to maturity T, $P(K)$ is the prices of the put option with the strike price K and time to maturity T, r_f is the risk-free interest rate for a maturity of T, and S^* is some value of S.

We can assume that the price of the options with strike prices $K_i\,(1\leq i\leq n)$ are known, and $K_1<K_2<K_3<\ldots<K_{n-1}<K_n$, and we set S^* equal to the first strike price below F_0, then we approximate the integrals as:

$$\int\limits_0^{S^*}\frac{1}{K^2}e^{r_fT}P(K)\mathrm{d}K+\int\limits_{S^*}^{+\infty}\frac{1}{K^2}e^{r_fT}C(K)\mathrm{d}K\cong\sum_{i=1}^{n}\frac{\Delta K_i}{K_i^2}e^{r_fT}Q(K_i) \tag{15.3.1296}$$

where:

$$\Delta K_i=\frac{K_{i+1}-K_{i-1}}{2}\qquad\text{with}\qquad 2\leq i\leq n-1 \tag{15.3.1297}$$

But in particular for the lowest strike price (of the Put options) K_1 we take:

$$\Delta K_1=K_2-K_1 \tag{15.3.1298}$$

and for the highest strike price (of the Call options) K_n we take:

$$\Delta K_n=K_n-K_{n-1} \tag{15.3.1299}$$

We also consider in this discretization that:

- If $K_i<S^*$, the function $Q(K_i)$ is the price of a Put option with strike price K_i.
- If $K_i>S^*$, the function $Q(K_i)$ is the price of a Put option with strike price K_i.
- If $K_i=S^*$, the function $Q(K_i)$ is the average of the price of a Call and a Put option with strike price K_i.

If we put:

$$\int\limits_0^{S^*}\frac{1}{K^2}e^{r_fT}P(K)\mathrm{d}K+\int\limits_{S^*}^{+\infty}\frac{1}{K^2}e^{r_fT}C(K)\mathrm{d}K\cong\sum_{i=1}^{n}\frac{\Delta K_i}{K_i^2}e^{r_fT}Q(K_i) \tag{15.3.1300}$$

into:

$$\mathrm{E}(\bar{\mathrm{V}})=\frac{2}{T}\ln\left(\frac{F_0}{S^*}\right)-\frac{2}{T}\left[\frac{F_0}{S^*}-1\right]+\frac{2}{T}\left[\int\limits_0^{S^*}\frac{1}{K^2}e^{r_fT}P(K)\mathrm{d}K+\int\limits_{S^*}^{+\infty}\frac{1}{K^2}e^{r_fT}C(K)\mathrm{d}K\right] \tag{15.3.1301}$$

we get:

$$\mathrm{E}(\bar{\mathrm{V}}) \cong \frac{2}{T}\ln\left(\frac{F_0}{S^*}\right) - \frac{2}{T}\left[\frac{F_0}{S^*}-1\right] + \frac{2}{T}\left[\sum_{i=1}^{n}\frac{\Delta K_i}{K_i^2}e^{r_f T}Q(K_i)\right] \tag{15.3.1302}$$

or:

$$\mathrm{E}(\bar{\mathrm{V}})T \cong 2\ln\left(\frac{F_0}{S^*}\right) - 2\left[\frac{F_0}{S^*}-1\right] + 2\left[\sum_{i=1}^{n}\frac{\Delta K_i}{K_i^2}e^{r_f T}Q(K_i)\right] \tag{15.3.1303}$$

According to $\ln(N)$ and its following Maclaurin polynomial:

$$\ln(N) = (N-1) - \frac{1}{2}(N-1)^2 + \mathcal{O}\left(\left(\frac{F_0}{S^*}-1\right)^2\right) \tag{15.3.1304}$$

We can get:

$$\ln\left(\frac{F_0}{S^*}\right) = \left(\frac{F_0}{S^*}-1\right) - \frac{1}{2}\left(\frac{F_0}{S^*}-1\right)^2 + \mathcal{O}\left(\left(\frac{F_0}{S^*}-1\right)^2\right) \tag{15.3.1305}$$

So if we choose to approximated the function as follow:

$$\ln\left(\frac{F_0}{S^*}\right) \cong \left(\frac{F_0}{S^*}-1\right) - \frac{1}{2}\left(\frac{F_0}{S^*}-1\right)^2 \tag{15.3.1306}$$

after rearranging we get:

$$\ln\left(\frac{F_0}{S^*}\right) - \left(\frac{F_0}{S^*}-1\right) \cong -\frac{1}{2}\left(\frac{F_0}{S^*}-1\right)^2 \tag{15.3.1307}$$

We can use that later relation and put it in:

$$\mathrm{E}(\bar{\mathrm{V}})T \cong 2\ln\left(\frac{F_0}{S^*}\right) - 2\left[\frac{F_0}{S^*}-1\right] + 2\left[\sum_{i=1}^{n}\frac{\Delta K_i}{K_i^2}e^{r_f T}Q(K_i)\right] \tag{15.3.1308}$$

to get:

$$\mathrm{E}(\bar{\mathrm{V}})T \cong -\left(\frac{F_0}{S^*}-1\right)^2 + 2\left[\sum_{i=1}^{n}\frac{\Delta K_i}{K_i^2}e^{rT}Q(K_i)\right] \tag{15.3.1309}$$

That can be written obviously as:

$$\sigma^2 T = -\left[\frac{F}{K}-1\right]^2 + 2\sum_i \frac{\Delta K_i}{K_i^2}e^{rT}Q(K_i) \tag{15.3.1310}$$

So finally:

$$\sigma_{\text{VIX}}^2 = -\frac{1}{T}\left[\frac{F}{K}-1\right]^2 + \frac{2}{T}\sum_i \frac{\Delta K_i}{K_i^2}e^{rT}Q(K_i) \tag{15.3.1311}$$

That lead us to the definition of the VIX for one option:

$$\boxed{\text{VIX}=100\sigma_{\text{VIX}}=100\sqrt{-\frac{1}{T}\left[\frac{F}{K}-1\right]^2+\frac{2}{T}\sum_i\frac{\Delta K_i}{K_i^2}e^{rT}Q(K_i)}} \tag{15.3.1312}$$

For a number j of options have same lifetime T, we have:

$$\text{VIX}=100\sigma_{\text{VIX}}=100\sqrt{\sum_j\sigma_{\text{VIX},j}^2}=100\sqrt{\sum_j\left(-\frac{1}{T}\left[\frac{F_j}{K_j}-1\right]^2+\frac{2}{T}\sum_i\frac{\Delta K_{i,j}}{K_{i,j}^2}e^{r_jT}Q(K_{i,j})\right)^2} \tag{15.3.1313}$$

And in the case the j options to have the same lifetime T, we have to weight them:

$$\text{VIX}=100\sigma_{\text{VIX}}=100\sqrt{\sum_j\omega_j\sigma_{\text{VIX},j}^2}=100\sqrt{\sum_j\omega_j\left(-\frac{1}{T_j}\left[\frac{F_j}{K_j}-1\right]^2+\frac{2}{T_j}\sum_i\frac{\Delta K_{i,j}}{K_{i,j}^2}e^{r_jT_j}Q(K_{i,j})\right)^2} \tag{15.3.1314}$$

The calculation are heavy and boring to do at hand. Hopefully, some softwares like `R` (see our companion book on the software) or the Bloomberg Terminal, automatically implement such calculations:

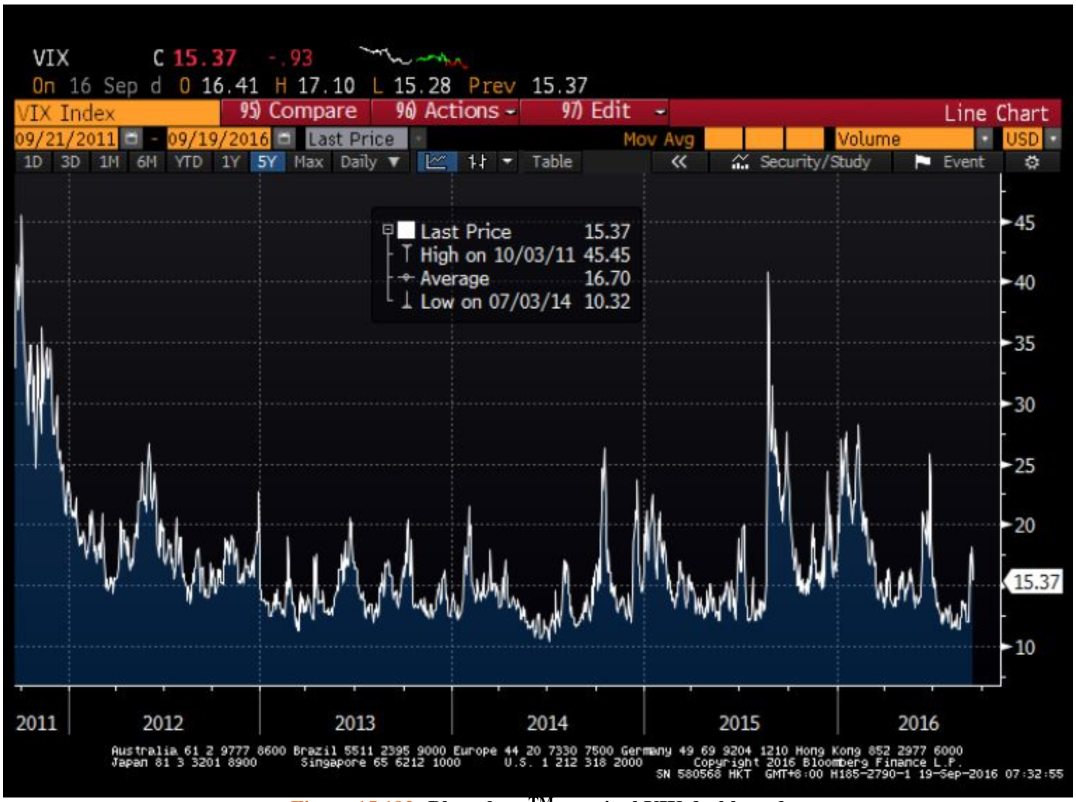

VIX C 15.37 -.93
On 16 Sep d O 16.41 H 17.10 L 15.28 Prev 15.37
VIX Index
95) Compare
96) Actions
97) Edit
Line Chart
09/21/2011 - 09/19/2016
Last Price
Mov Avg
Volume
USD
1D 3D 1M 6M YTD 1Y 5Y Max Daily
Table
Security/Study
Event
Last Price 15.37
High on 10/03/11 45.45
Average 16.70
Low on 07/03/14 10.32
45
40
35
30
25
20
15.37
10
2011
2012
2013
2014
2015
2016
Australia 61 2 9777 8600 Brazil 5511 2395 9000 Europe 44 20 7330 7500 Germany 49 69 9204 1210 Hong Kong 852 2977 6000
Japan 81 3 3201 8900 Singapore 65 6212 1000 U.S. 1 212 318 2000 Copyright 2016 Bloomberg Finance L.P.
SN 580568 HKT GMT+8:00 H185-2790-1 19-Sep-2016 07:32:55

Figure 15.192: **Bloomberg$^{\text{TM}}$ terminal VIX dashboard**

Otherwise for a detailed calculation by hand, including the subtleties of the official VIX documentation with multiple volatilities and hence a weight, the reader can refer again to the source of the above text (see [582]).

Investors use volatility derivatives to diversify and hedge their risk in equity and credit portfolios. Since long-term investors in equity funds are exposed to downside risk, volatility can be used as a hedge for the tail risk and replacement for the Put options. In the United States, as we have just seen it, the Chicago Board Options Exchange (CBOE) Volatility Index (VIX) measures the short-term volatility implied by S&P 500 stock index option prices. Many people around the world use the VIX to measure the stock market volatility over the next 30-day period. In Europe, the "EURO STOXX 50 Volatility[31]" (VSTOXX) market index is based on the market prices of a basket of Euro STOXX 50 Index Options (OESX) and measures the implied market volatility over the next 30 days on the EURO STOXX 50 Index. For benchmark strategies utilizing the EURO STOXX 50 Index, the nature of its negative correlation with the VSTOXX presents a viable way of avoiding benchmark-rebalancing costs. The statistical nature of volatility allows traders to perform mean-reverting strategies, dispersion trading, and volatility spread trading, among others.

---

[31] The official white-paper can be found here https://www.stoxx.com/document/Indices/Common/Indexguide/stoxx_strategy_guide.pdf

15.3.10.3.14 Value at Risk

The risk measurements have evolved since Markowitz advanced his famous theory of portfolio diversification in the late 11950s (holocene calendar), a theory that had to revolutionize modern portfolio management. The risk of a portfolio was then related to the matrix of covariances-variances, as we have demonstrated theoretically and by the examples above.

In the 11960s (holocene calendar), Sharpe proposed the unifactory model of valuation of financial assets where the beta is the primary explanatory factor of a portfolio's risk via the beta matrix.

In the early 11990s (holocene calendar), a new measure of risk was introduced (it seems that JP Morgan is at the origin of that one). Indeed, the limits of traditional risk measures were increasingly recognized. Quantitative indicators of the risk of value loss of assets had then to be found. To do this, indicator had to be found that were more closely related to the overall cash flow distribution of a portfolio. It is in this context that a nominal risk measure has been proposed: the "Value at Risk (VaR)" (denoted sometimes "V@R" in project management). The idea is basically to say that we have $x\%$ cumulative probability of not losing more than a given amount y in cash in the next N periods of time (the period can be: days, months, etc.).

This new measure was first used to quantify the market risk to which portfolios are subject. Indeed, the Basel Accords recommended banks since 11997 (holocene calendar) to hold a regulatory capital amount to preserve their stability against standard market risks. However, this capital has since been calculated from the VaR and has become increasingly popular for assessing the risk of institutional or individual portfolios (and not only for this!). However, there is no single measure of VaR. Indeed, it is based on the concept of volatility, which is essentially latent. This is why banks need to use several VaR models in order to define the range of their possible losses. These calculations are all the more complex since the distribution of the yields of the securities measured at high frequency is significantly different from the Normal distribution (therefore for long intervals of time we approach rather often a Normal distribution).

Definition (#15.3.698): The "Value at Risk" (VaR) is the theoretical maximum loss that a portfolio manager (whose value is necessarily implicitly variable) can experience and for a certain period of time with a given cumulative probability (the use of VaR is not limited to financial instruments, it is used in many other areas of risk management in general).

Remark

The VaR is not really relevant if it is not presented with other risk indicators such as the Sharpe ratio, the Treynor ratio or the Greek coefficients (such as the beta). Finally, let us say that in practice the VaR is sometimes indicated in $\%$ of the total value of a portfolio.

15.3.10.3.14.1 Relative Value at Risk

In the classical model of the "relative VaR" (also sometimes named "parametric VaR"), we will assume that the statistical distribution of the results of a portfolio obeys at every moment a Normal distribution... that we will write for what will follow:

$$X = \mathcal{N}(\mu, \sigma^2) \tag{15.3.1315}$$

The next idea is that the random variable X can therefore be rewritten with a random centered reduced variable Normal (see section Statistics page 725) by putting:

$$\varepsilon = \mathcal{N}(0, 1) \tag{15.3.1316}$$

such that (use of the basic properties of the Normal distribution):

$$\boxed{X = \mu + \varepsilon\sigma} \tag{15.3.1317}$$

and this notation is therefore used in many other fields besides finance (project management, quality assurance, supply chain, etc.).

Let α be the critical threshold (in quantiles) associated with the cumulative probability targeted. We can then write:

$$X = \mu + \alpha\sigma \tag{15.3.1318}$$

which is an interesting notation because it deferred the analysis of risk and variability to the estimate of the standard deviation only (what the finance practitioners appreciate ...)! This notation is however a very common abuse of the following rigorous writing (which indicates more clearly that we use the quantile of α):

$$X = \mu + \alpha_q\sigma \tag{15.3.1319}$$

What would be written using the same notation as in the Statistics section when the distribution is Normal:

$$X = \mu + Z_\alpha\sigma \tag{15.3.1320}$$

This type of notation can easily be verified with the spreadsheet software Microsoft® Excel® 11.8346 for the sceptics ... Consider a portfolio P with an annual standard deviation of 10% (that will have to be expressed in cash) and assume that we have $1,000$.- in assets of this portfolio (on average). We then have the first year:

```
=NORMINV(99%;1000;10%*1000)=1000+NORMSINV(99%)*10%*1000
            =1000+2.326*10%*1000 =1,232.6
```

That is to say 99% of cumulative probability of having a portfolio that has a value between 0 and $1,232.6$.- at any time (we consider as negligible the cumulative probability that the portfolio has a negative value with this notation).

But what interests the manager is not to hedge the risk of the mean (because it is zero by construction here!) but of the volatility! In the previous case its value is 100.- (10% of $1,000$.-) and follows a reduced center Normal distribution. Hence the reason for defining the VaR formally as the mathematical relation which gives a confidence interval (or a quantile according to the point of view) of the standard deviation:

$$\boxed{\text{VaR} = \alpha\sigma} \tag{15.3.1321}$$

Thus, for a cumulative probability of 99% softwares gives us in absolute value (see the treatment of the confidence intervals in the section of Statistics page 848):

$$\alpha\texttt{=NORMSINV(99\%)=ABS(NORMSINV(1\%))=2.326}$$

where by tradition the finance practitioner take the α (and thus the VaR) as being positive. Hence the fact that they speak of risk hedged to 99% (implicitly for a given period!) whereas in reality it is to cover again risk that has 1% cumulative probability of taking place (but strictly speaking it is the same thing at the difference that the first one is easier to make understand to a customer... !!!). This is why we sometimes find the VaR in the following form:

$$\text{VaR} = -\alpha\sigma \tag{15.3.1322}$$

It should be noticed that the result of the calculation may also be referred to as a "fractional reserve" (although this has nothing to do with the way in which States impose a fractional reserve to banks) or "mathematical reserve" (although hist has nothing to do with the way in which insurance calculates the mathematical reserve).

If this is not very clear, let us recall the following diagram seen in the section of Statistics:

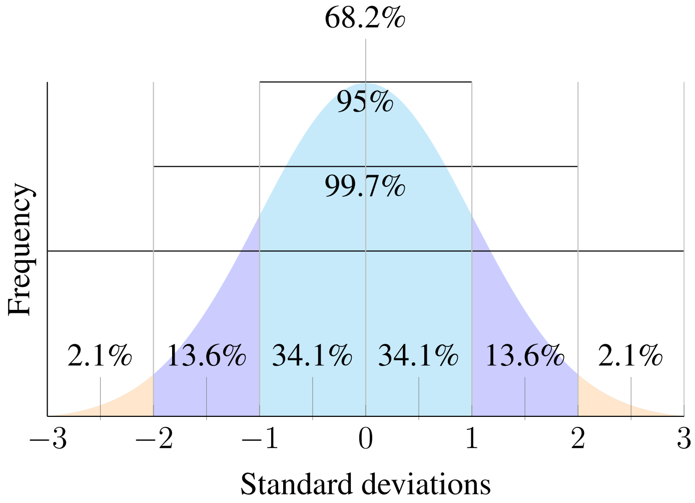


Figure 15.193: **Sigma intervals for the Normal distribution**

which was presented into the following table for some selected values commonly used in the section of Engineering (we will see some of these typical values in finance further below):

Sigma Quality Level	**Left/Right Cumulated %**	**One Sided Cumulative %**
1σ	68.26894	$68.26894/2=34.13447$
2σ	95.4499	$95.4499/2=47.72495$
3σ	99.73002	$99.73002/2=49.86501$
4σ	99.99366	$99.99366/2=49.99683$
5σ	99.999943	$99.999943/2=49.9999715$
6σ	99.9999998	$99.9999998/2=49.9999999$

An interesting information: If we cover the risk at 2σ for example on a daily basis (which is a very poor coverage...) this means that in the one-sided case, we therefore assume about 5% cumulative probability per day, that the adverse event occurs. In extenso if the undesirable events are independent then we can use the geometric law which we proved in the Statistics section - see page 654 - that the expected mean of the first occurrence of the event (undesirable in the present case...) was equal to the inverse of the probability. So with the chosen range, we have the event in question that will appear on average about once every 20 days ($1/5\%$).

Remark

The RiskMetrics model of J.P. Morgan/Reuters proposed in 11996 (holocene calendar) to take an α quantile of 1.65 corresponding to a cumulative probability of 95%. Some practitioners strongly recommend, in view of the strong and simplifying hypothesis of Normality ..., to multiply the VaR by a factor 2 or 3 (we will prove why).

✍**Example:**

A portfolio P of value $1,000$.- has an annual volatility σ_A of 10%. The daily (instantaneous) volatility of the yield is then (we use here the property of the standard Brownian motion where, for recall, the returns are assumed to be independent from one day to another ...):

$$\sigma_{R,d}=\frac{10\%}{\sqrt{252}}\cong 0.63\% \tag{15.3.1323}$$

where 252 is the number of trading days in a given year for a given country. Either in cash:

$$\sigma_d\cong 0.63\%\cdot 1,000\cong 6.29.\text{-} \tag{15.3.1324}$$

The VaR relative to the 99% threshold at one day is then:

$$\text{VaR}_d=\alpha\sigma=2.326\cdot 6.29\cong 14.65.\text{-} \tag{15.3.1325}$$

Also over one year we would have the following relative annual VaR at the following 99% threshold:

$$\text{VaR}_A=2.326\cdot 1,000\cdot 10\%=232.6.\text{-} \tag{15.3.1326}$$

Or a relative VaR of $10\%\cdot 2.326=23.26\%$ (just to give it in percentages as it is customary in the financial field).

Thus, with respect to the relative annual VaR, we then have 99% cumulative probability of winning 232.60.- but also to lose them! Indeed we have 1% cumulative probability of having an annual loss of:

```
=NORMSINV(1%*10%*1000)=-232.6
```

So it would require at least a risk capital (equity) of 232.6.- to cover 99% of the risks (cover this cumulative probability of 1% of being in a bad year respectively). We can also say that we have 99% cumulative probability of not losing more that 232.6.-. We thus find the same numerical result as with the previous example.

The reader will notice that we thus have in the field of the Stock Exchange (this follows from the standard Brownian motion) to change from a daily time horizon to an annual one:

$$\text{VaR}_A=\text{VaR}_d\sqrt{t}=2.326\cdot\sigma_j\sqrt{252} \tag{15.3.1327}$$

Finance practitioners name this property of Brownian motion in the context of the use of VaR:

$$\text{VaR}_T=\text{VaR}_t\sqrt{n\cdot t}=\alpha\sigma_t\sqrt{n\cdot t} \tag{15.3.1328}$$

the "scaling law". It is authorized by the Basel Accords in 11996 (holocene calendar) which presuppose a Normal distribution and advise a time horizon of 10 to 30 days. We have seen, however, in our proof of the standard Brownian motion model, that we underestimate the real risk under this hypothesis and that this reflex of change of scale via the square root is highly criticized by some specialists.

Therefore the VaR, while being an amount putted in provisions for liabilities and charges in order to face a violent but temporary breakdown of the market, is not sufficient to protect an institution from a strong and sustainable return of the market. Reason why banks often provision the minimum minimorum...

In the case where the returns are not independent but are described by an autoregressive process of order 1 (see the Time Series study later in this section at page 7989) such that:

$$R_t=\rho R_{t-1}+\varepsilon_t \tag{15.3.1329}$$

We then have considering the residual variance as negligible and that the yields have an equal variability from one day to the next (that is to say $\text{V}(R_t)=\text{V}(R_{t-1})$):

$$\text{V}(R_t+R_{t+1})=\text{V}(R_t)+\text{V}(R_t)+2\text{cov}(R_t,R_{t-1})=\sigma_d^2+\sigma_d^2+2\rho\sigma_d^2 \tag{15.3.1330}$$

We therefore have:

$$\text{V}(R_t+R_{t+1})=\sigma_j^2 2(1+\rho)\Leftrightarrow\sigma_{2j}=\sigma_j\sqrt{2(1+\rho)} \tag{15.3.1331}$$

If the correlation ρ is zero we fall back on the square root time scale factor. It is interesting to notice also that in the general case (see the study of AR(1) further below at page 7989) where the return has a tendency to increase, the AR(1)-VaR is always higher than the classic VaR.

Remarks

R1. Personally I will advocate covering up by the Six Sigma method at 99.9996% over a time horizon corresponding to the at least to the median position time. But it's personal and very simplistic ...

R2. There is a simplistic and traditional version of the VaR, named "nonparametric VaR", which consists in taking a given percentile of the distribution of historical or simulated losses. Thus, if we have at our disposal N data of loss, the VaR parameter at the threshold of 99% then consists in taking the 99th percentile of the series of values. So nothing special to report at this level ...

The goal now is to prove why the Basel Accords recommends to multiply the VaR by a factor of 3.03.

Remember first that in the section of Statistics (see page 920) a variant of notation of the inequality of Bienaymé-Chebyshev is:

$$P\left(|X-\mathrm{E}(x)|\geq t\sigma\right)\leq\frac{1}{t^2} \tag{15.3.1332}$$

Let us now make the strong hypothesis that the underlying distribution is symmetrical. Hence, it is almost immediate that we are led to write:

$$P\left(X-\mathrm{E}(x)\leq -t\sigma\right)\leq\frac{1}{2}\frac{1}{t^2} \tag{15.3.1333}$$

Suppose we impose to this deviation to the mean to have a cumulative probability of 1%. This is therefore equivalent to write:

$$P\left(X-\mathrm{E}(x)\leq -t\sigma\right)=1\%\Rightarrow 1\%=\frac{1}{2}\frac{1}{t^2}\Leftrightarrow t=7.07 \tag{15.3.1334}$$

Therefore the deviation from the mean at the cumulative probability threshold of 1% can then be written:

$$P\left(X-\mathrm{E}\leq -7.07\sigma\right)=1\% \tag{15.3.1335}$$

Now, let us recall that in the case of a Normal distribution, we have just seen above that:

$$\text{VaR}=\alpha\sigma=2.326\sigma \tag{15.3.1336}$$

There is therefore a ratio of:

$$\frac{7.071\sigma}{2.326\sigma}\cong 3.03 \tag{15.3.1337}$$

We thus fall back on the multiplier proposed in the first Basel Accords concerning the best practices about the VaR. Obviously, it is quite justified that the reader wonders about big difference between the two approaches (almost 300% !!!). It is simple to remember that the inequality of BT is an inequality valid for any symmetric distribution function and the value of 7.071 is therefore the upper limit of the worst (max) for the general case whereas the value of 2.326 is only for the Normal distribution which is a nice distribution (with for recall the probability which decreases exponentially whereas for some symmetric laws the decay is just of power type as it is the case for the Pareto distribution function).

15.3.10.3.14.2 Absolute Value at Risk

The VaR measure we have just given is a relative measure because it does not take into account the average of future losses and gains.

If the volatility is 100.- in the example just given, the relative VaR is therefore 232.6.- But since the average profit is generally non-zero over a long period of time, we must use most of the time the "absolute measure of VaR" (over a very short period, profit being considered sometimes as zero, the calculation of the relative VaR is used).

Let us first recall that following our study of the Bachelier model we have proved that the positive expected mean of the value (or yield) as well as the positive standard deviation of a portfolio are proportional to the square root of time.

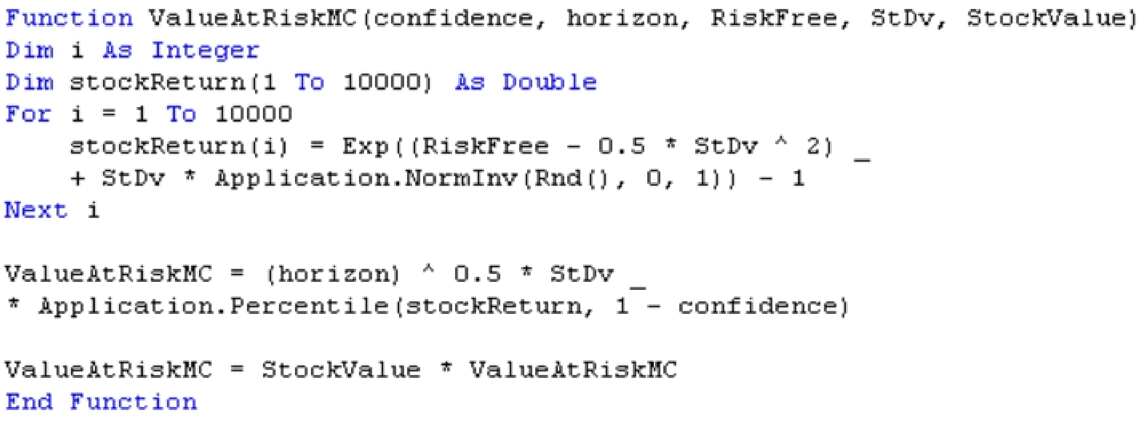
Suppose that the observation period t is in months. The expected monthly return for the portfolio of initial value S is then of μ (its expected mean therefore!!) and the monthly variance of its return is σ^2.

Its relative VaR to the confidence threshold α is therefore after t months given by (you can verify that the relation is indeed homogeneous!):

$$\text{VaR}_r = S \cdot \left(\alpha\sigma\sqrt{t}\right) \tag{15.3.1338}$$

as we were able to verify in the previous example (so far nothing new ...). The square root of time comes, as a reminder, from the Bachelier model (standard Brownian motion).

In Monte Carlo version with VBA (because the version above is named "punctual") this gives:



```
Function ValueAtRiskMC(confidence, horizon, RiskFree, StDv, StockValue)
Dim i As Integer
Dim stockReturn(1 To 10000) As Double
For i = 1 To 10000
    stockReturn(i) = Exp((RiskFree - 0.5 * StDv ^ 2) _
    + StDv * Application.NormInv(Rnd(), 0, 1)) - 1
Next i

ValueAtRiskMC = (horizon) ^ 0.5 * StDv _
* Application.Percentile(stockReturn, 1 - confidence)

ValueAtRiskMC = StockValue * ValueAtRiskMC
End Function
```

Remark

Contrary to what we saw in our study of thresholds/confidence intervals in the section Statistics (see page 848), we do not divide by 2 the `NORMALSINV()` function argument of Microsoft® Excel® 11.8346 to get the α in the above situation because what interests us is only one side of the Normal reduced centered distribution (the "pessimistic" side) and not both.

If we take the same example as before (portfolio of $1,000$.- with 10% annual volatility). The relative VaR is therefore on a 30-day projection of:

$$\text{VaR}_r = S\alpha\sigma\sqrt{t} = 1000 \cdot 2.326 \frac{10\%}{\sqrt{252}} \tag{15.3.1339}$$

But this last relation does not take into account the expected average return portfolio equation over time. The "absolute VaR" is thus obtained by subtracting this yield from the relative VaR over the same time period, that is to say:

$$\boxed{\text{VaR}_a = \text{VaR}_r - S\mu t = -S(\mu \cdot t - \alpha\sigma\sqrt{t})} \tag{15.3.1340}$$

where we make the particular assumption that the return is thus linearly time-dependent... (in accordance with the semi-empirical construction of standard Brownian motion). The absolute VaR is therefore obviously less than the relative VaR of this amount.

It should be noted that the calculation of the absolute VaR may be considered vicious or of little interest because it assumes that the gain obtained from the return will be placed in the fractional reserve to finance the relative VaR. In most cases, however, the earnings will be re-affected in the Market.

Let us come back to our usual example under this assumption (portfolio of $1,000$.- with 10% annual volatility) with an annual return rate of 15%. We have then:

$$\text{VaR}_a = S\alpha\sigma\sqrt{t} - S\mu t = 1000 \cdot 2.326 \cdot 10\% - 1000 \cdot 15\% \cong 232.6 - 150 = 82.6.\text{-} \tag{15.3.1341}$$

Concretely, if we finance the VaR with the gains, then on one year it is enough to have 82.6.- of fractional reserve. In practice it may be interesting to after how much time the earnings cover the entire VaR. In this case it is a simple equation of the second degree such that:

$$S\alpha\sigma\sqrt{t} - S\mu t = 0 \tag{15.3.1342}$$

and we would find in our example 2.4 years. Concretely after 2.4 years the gains will have covered all the risks according to the assumptions of the construction of the VaR...

15.3.10.3.14.3 Delta-Normal Value at Risk

We have just seen that when we work with an asset and the standard deviation is expressed in cash (or in $\%$ then converted into cash), the cash value of the relative VaR is therefore given under the assumption of Normality by:

$$\text{VaR} = \alpha\sigma \tag{15.3.1343}$$

But what about the VaR of a derivative (option typically!) knowing that it depends indirectly on the value of the underlying and that only the latter is measurable on the market as a whole (in order to get a statistical distribution)? Well for this we will re-use the Taylor development used earlier in our study of Itō processes:

$$\Delta F = \frac{1}{1!}\frac{\partial F}{\partial t}\Delta t + \frac{1}{1!}\frac{\partial F}{\partial x}\Delta x + \frac{1}{2!}\frac{\partial^2 F}{\partial t^2}\Delta t^2 + \frac{1}{2!}\frac{\partial^2 F}{\partial x^2}\Delta x^2 + \ldots \tag{15.3.1344}$$

Therefore, by taking only the first order, we have the following affine approximation (thus a coarse approximation valid only in the linear domain):

$$\Delta F = \frac{\partial F}{\partial t}\Delta t + \frac{\partial F}{\partial x}\Delta x \tag{15.3.1345}$$

and leaving aside the temporal variation (...) and taking an infinitely small variation (which, however, makes it possible to make the approximation a little more realistic):

$$\mathrm{d}F = \frac{\partial F}{\partial x}\mathrm{d}x \tag{15.3.1346}$$

And by writing with the traditional notation of the underlying as used during our pricing studies of options and using the notation of the Greeks as defined earlier:

$$\mathrm{d}F = \frac{\partial F}{\partial S}\mathrm{d}S = \Delta \mathrm{d}S \tag{15.3.1347}$$

Thus, the variation of a derivative assumed to depend only on the changes in the first-order of the underlying value is directly proportional to the marginal variation in the price of the option multiplied by the change in the underlying. Thus, if we assume the underlying price distribution of the underlying as Normal - reason for which we speak of "delta-Normal VaR", then a coarse approach naturally consists in writing whether the volatility of the underlying is given in $\%$:

$$\boxed{\text{VaR}_{\Delta\mathcal{N}} \cong \frac{\partial F}{\partial S}\alpha\sigma_S S_0 = \Delta\alpha\sigma_S S_0} \tag{15.3.1348}$$

and if the volatility of the underlying is given directly in cash:

$$\text{VaR}_{\Delta\mathcal{N}} \cong \Delta\alpha\sigma_S \tag{15.3.1349}$$

15.3.10.3.14.4 (Sorted) Historical Value at Risk

A third pragmatic nonparametric way of calculating the relative VaR is based on historical data! This is the easiest way to do the calculation with the ease of use of existing spreadsheet softwares.

Let us assume for the example that we have the last 100 daily performances of a portfolio (including negative and positive returns!). The 10 worst daily performances are given in ascending order:

Historical Data
$-19,000$.-
$-16,450$.-
$-15,000$.-
$-12,500$.-
$-11,950$.-
-11.250.-
$-11,050$.-
$-10,600$.-
$-10,500$.-
$-10,250$.-
...

The 95% historical VaR for 1 day then consist to determine the 5th percentile of these ascending sorted values. Since we have 100 samples, it is easy to determine that this is the 5th value in ascending order of values. Therefore in our example above:

$$\text{VaR}_{H,5\%} = 11,950.\text{-} \tag{15.3.1350}$$

As already mentioned in the section Statistics, spreadsheet software use the `PERCENTILE( )` function which is not necessarily calculated in the same way from one software to another...

Formally, if we denote by X the distribution of returns, the (sorted) historical VaR is defined in the general case by:

$$\boxed{\text{VaR}_{H,\alpha}(X) = F_X^{-1}(1-\alpha)} \tag{15.3.1351}$$

However this definition cannot be used directly for calculations unless we assume that X has some parametric distribution!

Remark

It may be however more secure in practice to take into account only negative returns in the above procedure!

15.3.10.3.14.5 Credit Value at Risk (CVaR)

The "Credit Value at Risk" (CVaR) of a portfolio is the worst loss expected to be suffered due to counterparty default over a given period of time with a given probability. The abbreviation CVaR should not be confused with the Conditional Value at Risk we will see further below!

To introduce the credit VaR ludicrously let us consider a companion example of three portfolios A, B and C, whose credit default probabilities follow a Bernoulli distribution (i.e., a binary event) of respective values of 5%, 10% and 20% and the monetary values are in millions of dollars of 25.-, 30.- and 45.-.

For the following example, we will make the assumptions (which are often implicit even in all we have seen so far) that the risk exposure is constant, that the portfolios are independent and that in case of credit default we lose everything!!!

As the three portfolios are therefore independent, the expected mean of loss (denoted "L" for Loss) is easily computable using the linearity property of the mean:

$$\text{E}(L) = 5\% \cdot 25 + 10\% \cdot 30 + 20\% \cdot 45 = 13.25.\text{-} \tag{15.3.1352}$$

Since the loss of credit is considered a Bernoulli random variable (and the portfolios are always independent), we always have in millions of dollars (the reader can refer to the calculation of the variance of a Bernoulli random variable in the section Statistics):

$$\begin{aligned}\text{V}(L) &= \text{V}(L_A + L_B + L_C) = \text{V}(L_A) + \text{V}(L_B) + \text{V}(L_C)\\ &= 25 \cdot 5\% \cdot (1-5\%) + 30 \cdot 10\% \cdot (1-10\%) + 45 \cdot 20\% \cdot (1-20\%) = 434.688.\text{-}\end{aligned} \tag{15.3.1353}$$

Well this being done for pleasure ... How can we determine as finely as possible the distribution of losses to have the credit VaR at a given threshold? Well, for that we will have to construct the table of all possible outcomes (or build in the state of the art a Kernel Smoothing distribution function but manually this is quite boring to do).

If there are N credits, we have 2^N possible scenarios therefore:

Loss Scenario	**Associated Loss**	**Probability**	**Cumulated Probability**
None	0.-	$(1-0.05)\cdot(1-0.1)\cdot(1-0.2)=68.40\%$	68.40%
A	25.-	$0.05\cdot(1-0.1)\cdot(1-0.2)=3.60\%$	72.00%
B	30.-	$(1-0.05)\cdot 0.1\cdot(1-0.2)=7.60\%$	79.60%
C	45.-	$(1-0.05)\cdot(1-0.1)\cdot 0.2=17.10\%$	96.70%
A,B	55.-	$0.05\cdot 0.1\cdot(1-0.2)=0.40\%$	97.10%
A,C	70.-	$0.05\cdot(1-0.1)\cdot 0.2=0.90\%$	98.00%
B,C	75.-	$(1-0.05)\cdot 0.1\cdot 0.2=1.90\%$	99.00%
A,B,C	100.-	$0.05\cdot 0.1\cdot 0.2=0.1\%$	100%

Thus, the credit VaR closest to 95% is 45 million USD. Since we can expect a loss of 13.25 million USD, the unexpected portion will be considered as 31.75 million USD (that is to say the difference).

15.3.10.3.14.6 Operational Value at Risk

We have already encountered the concept of convolution several times in the section Statistics but for continuous laws explicitly known in their algebraic form .... In practice it is quite different! Thus, actuaries have often at their disposition for a given period (often over a year), an experimental distribution of the frequency of events of interest (accidents, deaths or other) and an experimental distribution of the costs of events of interest. For example, consider the very simplified (in size!) following case:

Frequency Distribution (losses)	
Probability	**Frequency**
0.6	0
0.3	1
0.1	2
Expected mean:	0.5

Distribution of Loss Costs	
Probability	**Frequency**
0.5	$1,000.$-
0.3	$10,000.$-
0.2	$100,000.$-
Expected mean:	$23,500.$-

If we want from these experimental and discrete values to determine the empirical distribution of costs, we will have to make a discrete convolution whose principle is the following:

Number of losses	First loss	Second loss	Total Loss	Probability
0	0	0	0	$=0$
1	$1,000.$-	0	$1,000.$-	$=0.3\cdot 0.6=0.150$
1	$10,000.$-	0	$10,000.$-	$=0.3\cdot 0.3=0.090$
2	$100,000.$-	0	$100,000.$-	$=0.3\cdot 0.2=0.060$
2	$1,000.$-	$1,000.$-	$2,000.$-	$=0.1\cdot 0.5\cdot 0.5=0.025$
2	$1,000.$-	$10,000.$-	$11,000.$-	$=0.1\cdot 0.5\cdot 0.3=0.015$
2	$1,000.$-	$100,000.$-	$101,000.$-	$=0.1\cdot 0.5\cdot 0.2=0.010$
2	$10,000.$-	$1,000.$-	$11,000.$-	$=0.1\cdot 0.3\cdot 0.5=0.015$
2	$10,000.$-	$10,000.$-	$20,000.$-	$=0.1\cdot 0.3\cdot 0.3=0.009$
2	$10,000.$-	$100,000.$-	$110,000.$-	$=0.1\cdot 0.3\cdot 0.2=0.006$
2	$100,000.$-	$1,000.$-	$101,000.$-	$=0.1\cdot 0.2\cdot 0.5=0.010$
2	$100,000.$-	$10,000.$-	$110,000.$-	$=0.1\cdot 0.2\cdot 0.3=0.006$
2	$100,000.$-	$100,000.$-	$200,000.$-	$=0.1\cdot 0.2\cdot 0.2=0.004$

The loss expected mean is then equal to $11,750.$-, which is also equivalent to simply calculate (the expected mean of the product of two independent random variables being equal to the product of the expected man as we have proved it in the section Statistics):

$$0.5\cdot 23,500=11,750.\text{-} \tag{15.3.1354}$$

Then we combine and sort the losses and probabilities:

Loss	**Probability**
$0.$-	60%
$1,000.$-	75%
$2,000.$-	77.5%
$10,000.$-	86.5%
$11,000.$-	89.5%
$20,000.$-	90.4%
$100,000.$-	96.4%
$101,000.$-	98.4%
$110,000.$-	99.6%
$110,000.$-	100%

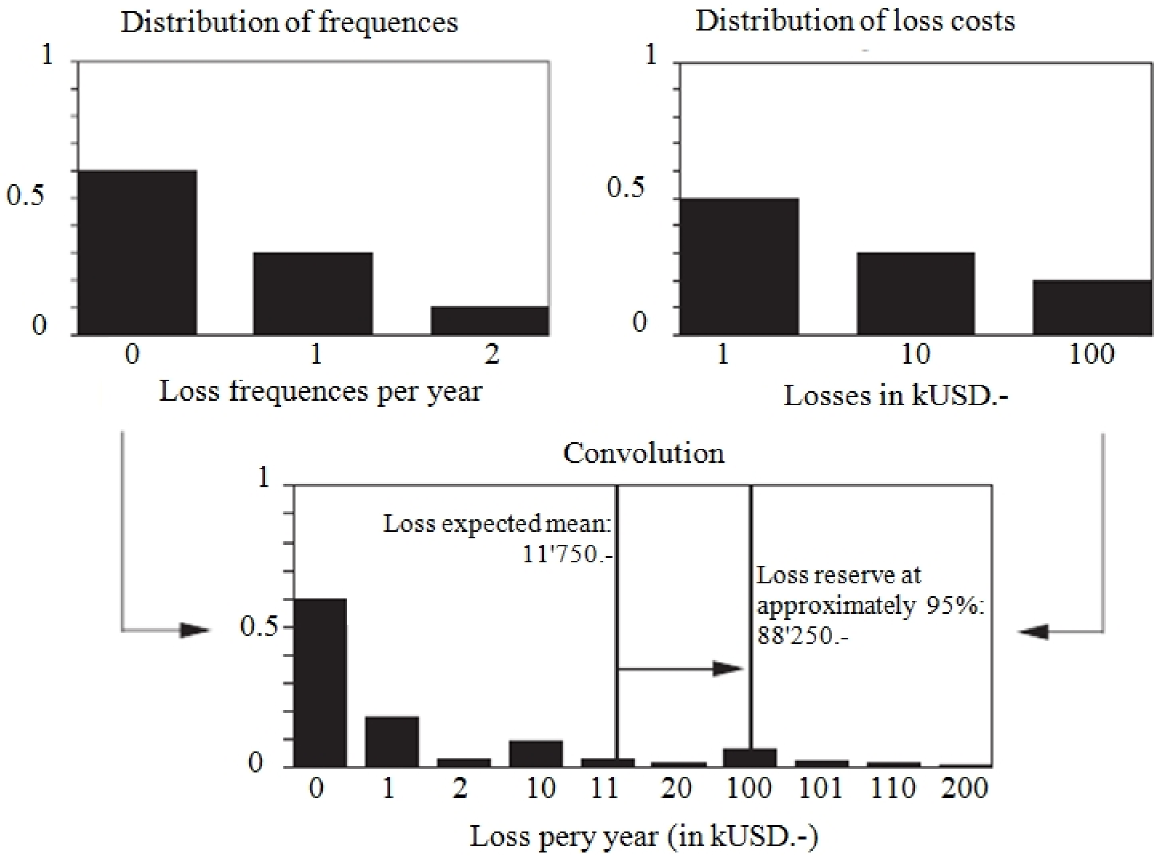
On the side of the operational VaR, if we set it at 95% (the Basel II agreements propose instead to take 99.9%), the nearest value is then:

$$\text{VaR}_{\text{op.}}=100,000.\text{-} \tag{15.3.1355}$$

Therefore, the reserve of unexpected loss is:

$$100,000-11,750=88,250.\text{-} \tag{15.3.1356}$$

Graphically our tables, manipulations and calculations can be summarized as follows:



It is interesting to notice how two simple tables produce a quantity of computations in convolution relatively huge. This is why, in real practical cases, the use of computers is necessary.

15.3.10.3.14.7 Surplus Value at Risk (Pension VaR)

Risks should be seen in a more general context in the case of a portfolio or an accounting system including asset management and liabilities.

Thus, in the case of a portfolio or an accounting, the risk is that the asset values no longer cover the liabilities of the portfolio.

Definition (#15.3.699): We name "surplus", and denote it by S, the difference between the amount of assets $\mathcal{A}$ and liabilities $\mathcal{L}$. But in a totally financial context aspect we will focus mainly on the difference of the mean of the returns between the two such as:

$$\mu_S = \mu_{R,\mathcal{A}} - \mu_{R,\mathcal{L}} \tag{15.3.1357}$$

✍**Example:**

Consider a cash portfolio with $\mathcal{A}_t = 120$ and $\mathcal{L}_t = 100$. Therefore:

$$S_t = \mathcal{A}_t - \mathcal{L}_t = 120 - 100 = 20 \tag{15.3.1358}$$

Suppose we have:

$$\begin{matrix} \mu_{R,\mathcal{A}} = 8\% & \mu_{R,\mathcal{L}} = 5\% \\ \sigma_{R,\mathcal{A}} = 12\% & \sigma_{R,\mathcal{L}} = 3\% \end{matrix} \tag{15.3.1359}$$

With the correlation:

$$\rho_{\mathcal{A},\mathcal{L}} = 0.3 \tag{15.3.1360}$$

Then in this case over the next period it is very likely to have in cash:

$$\begin{aligned} \Delta\mathcal{A} &= \mathcal{A}\mu_{R,\mathcal{A}} = 128 \cdot 8\% = 9.6 \\ \Delta\mathcal{L} &= \mathcal{L}\mu_{R,\mathcal{L}} = 100 \cdot 3\% = 5 \end{aligned} \tag{15.3.1361}$$

So in extenso, we have:

$$\Delta S = \Delta\mathcal{A} - \Delta\mathcal{L} = 4.6 \tag{15.3.1362}$$

So the expected surplus for the next time period will be:

$$S_{t+1} = S_t + \Delta S = 24.6 \tag{15.3.1363}$$

And for the variance of the surplus, we need it to calculate the SVaR, that is the "Surplus Value at Risk", then it comes:

$$\begin{aligned} \mathrm{V}(S) &= \mathrm{V}(\mathcal{A} - \mathcal{L}) = \mathrm{V}(\mathcal{A}) + \mathrm{V}(\mathcal{L}) - 2\mathrm{cov}(\mathcal{A},\mathcal{L}) = \mathrm{V}(\mathcal{A}) + \mathrm{V}(\mathcal{L}) - 2\rho_{\mathcal{A},\mathcal{L}}\sigma_{\mathcal{A}}\sigma_{\mathcal{L}} \\ &= (120 \cdot 12\%)^2 + (100 \cdot 3\%)^2 - 2 \cdot 0.3 \cdot (120 \cdot 12\%)(100 \cdot 3\%) = 190.44 \end{aligned} \tag{15.3.1364}$$

Therefore:

$$\sigma_S = \sqrt{190.44} = 13.80 \tag{15.3.1365}$$

Now, let us recall that under certain hypotheses necessarily realized, we have demonstrated that the absolute VaR was given in the case of a Normal underlying distribution by:

$$\text{VaR}_a = \mu t + Z_\alpha \sigma \sqrt{t} \qquad (15.3.1366)$$

We have then in our example the surplus VaR at one unit time horizon unit and at the 95% threshold:

$$\text{SVaR}_a = -24.6 \cdot 1 + 1.64 \cdot 13.8 + \sqrt{1} = 18.1 \qquad (15.3.1367)$$

So generally the example above gives us explicitly:

$$\boxed{\text{SVaR}_a = [S_t + (\mathcal{A}\mu_{R,\mathcal{A}} - \mathcal{L}\mu_{R,\mathcal{L}})]\, t + Z_\alpha \sqrt{\text{V}(\mathcal{A}) + \text{V}(\mathcal{L}) - 2\text{cov}(\mathcal{A},\mathcal{L})}\sqrt{t}} \qquad (15.3.1368)$$

15.3.10.3.14.8 Variance-Covariance Value at Risk

The variance-covariance VaR is based on a more realistic case of calculating the VaR for a portfolio composed of several financial assets correlated or not (unlike the previous cases where we had only one asset).

In order to introduce this concept, let us consider a companion example of a portfolio P_1 of $5,000,000$.- and daily volatility of 2% (i.e. $100,000$.-/day) and a second portfolio P_2 of $7,000,000$.- and daily volatility of 1% (i.e. $70,000$.-/day).

Our measurements show that their Pearson correlation coefficient ρ is 0.6. The daily global standard deviation is then (see section Statistics page 597):

$$\begin{aligned}\sigma_{1+2} &= \sqrt{\sigma_1^2 + \sigma_2^2 + 2R_{1,2}\sigma_1\sigma_2} = \sqrt{100,000^2 + 70,000^2 + 2 \cdot 0.6 \cdot 100,000 \cdot 70,000} \\ &\cong 152,643.-\end{aligned} \qquad (15.3.1369)$$

where as usual we see that if the correlation is negative this decreases the overall volatility.

Thus, the relative variance-covariance daily VaR at 99% for the global portfolio is (no scaling law to apply here since the standard deviation is daily and we want the daily relative VaR):

$$\text{VaR}_{r,1+2} = \alpha\sigma_{1+2} = 2.326 \cdot 152,634 \cong 355,000.- \qquad (15.3.1370)$$

It is interesting to compare the daily relative VaR with the sum of the relative VaRs of the two portfolios:

$$\text{VaR}_{r,1} + \text{VaR}_{r,2} = \alpha\sigma_1 + \alpha\sigma_2 = 2.326 \cdot 100,000 + 2.326 \cdot 70,000 \cong 395,420.- \qquad (15.3.1371)$$

We then have:

$$\text{VaR}_{r,1+2} \leq \text{VaR}_{r,1} + \text{VaR}_{r,2} \qquad (15.3.1372)$$

This is due to the diversification gain!

Remark

In finance when the merging of risks is less than the sum of the risks, we then technically say that the risk measure is "subadditive".

Many practitioners do the same calculation with yield rather than with cash values. Thus we get the VaR expressed in terms of yield.

Remark

A first trap in calculating the relative variance-covariance VaR would have been to calculate the global standard deviation in $\%$ and then apply to it the relation of the calculation of the global VaR. The result would have been wrong! A second trap is to use the same percentile of the VaR for merging risks as for individual risks because if we can accept individual ruins in a highly competitive market we can not accept it with the same level of risk For the same large structures (we should apply Bonferroni correction or Fisher p-value calculation as seen in the section Statistics).

If we had a single portfolio consisting of several financial instruments such that the total proportion is equal to unity then we should, in accordance with what we saw during our study of the Markowitz model, work with the following relation if we use returns:

$$\sigma_{1+2}=\sqrt{X_1^2\sigma_1^2+X_2^2\sigma_2^2+2R_{1,2}X_1X_2\sigma_1\sigma_2} \qquad (15.3.1373)$$

Thus, if we consider the first portfolio in our companion example above in a proportion of 75% (and therefore in extenso the second in a proportion of 25%). We then have:

$$\sigma_{1+2}=\sqrt{X_1^2\sigma_1^2+X_2^2\sigma_2^2+2R_{1,2}X_1X_2\sigma_1\sigma_2}\cong 1.66\% \qquad (15.3.1374)$$

Thus a corresponding VaR return of (always under the assumption of Normality):

$$\text{VaR}_{r,1+2}=2.326\cdot 1.66\%\cong 3.87\% \qquad (15.3.1375)$$

that therefore represents in $\%$, the amount we could lose on our portfolio.

Therefore the general "variance-covariance (relative) Value at Risk", also named "Portfolio Value at Risk", relation is given by:

$$\boxed{\text{VaR}_{r,\sum_i^n}=\text{VaR}_P=\alpha\sqrt{\sum_{i=1}^{n}\sigma_{P_i}^2+2\sum_{i<j}^{n}\text{cov}(P_i,P_j)}} \qquad (15.3.1376)$$

Obviously (see section Statistics page 597) we could write it in matrix form if we want...

15.3.10.3.14.9 Marginal and Component Value at Risk

In our study of Markowitz's efficient diversification model, we showed that the variance of a portfolio composed of several investments in given proportions X_i was:

$$\mathrm{V}\left(R_p\right)=\sigma^2(R_p)=\sum_{i=1}^{n}X_i^2\sigma_i^2+2\sum_{i<j}\sum X_iX_j\text{cov}(R_j,R_i)=\sum_{i=1}^{n}X_i^2\sigma_i^2+2\sum_{i<j}\sum X_iX_j\sigma_{ij} \tag{15.3.1377}$$

The variation of this variance with respect to a variation of one of the proportions X_i alone gives:

$$\frac{\partial\sigma^2(R_p)}{\partial X_i}=2X_i\sigma_i^2+2\sum_{j\neq i}X_j\text{cov}(R_j,R_i)=2\left(X_i\sigma_i^2+\sum_{j\neq i}X_j\text{cov}(R_j,R_i)\right) \tag{15.3.1378}$$

Let us recall the property of bilinearity of the covariance proved in the section of Statistics (see page 602):

$$\text{cov}(Y,aX+bZ)=a\text{cov}(X,Y)+b\text{cov}\left(Y,Z\right) \tag{15.3.1379}$$

By putting $Y=X=R_i$, $a=X_i$, we have:

$$\text{cov}(R_i,X_iR_i+bZ)=X_i\text{cov}(R_i,R_i)+b\text{cov}(R_i,Z)=X_i\sigma_i^2+b\text{cov}(R_i,Z) \tag{15.3.1380}$$

Let us put $b=1$, $Z=\sum_{j\neq i}X_jR_j$, we have:

$$\text{cov}\left(R_i,X_iR_i+\sum_{j\neq i}X_jR_j\right)=X_i\sigma_i^2+\text{cov}\left(R_i,\sum_{j\neq i}X_jR_j\right)=X_i\sigma_i^2+\sum_{j\neq i}X_j\text{cov}(R_i,R_j) \tag{15.3.1381}$$

Therefore we get:

$$2\left(X_i\sigma_i^2+\sum_{j\neq i}X_j\text{cov}(R_i,R_j)\right)=2\text{cov}\left(R_i,X_iR_i+\sum_{j\neq i}X_jR_j\right)=2\text{cov}(R_i,R_p) \tag{15.3.1382}$$

So finally:

$$\frac{\partial\sigma^2(R_p)}{\partial X_i}=\frac{\partial\sigma_p^2}{\partial X_i}=2\text{cov}(R_i,R_p) \tag{15.3.1383}$$

But we also have:

$$\frac{\partial\sigma_p^2}{\partial X_i}=2\sigma_p\frac{\partial\sigma_p}{\partial X_i}=2\text{cov}(R_i,R_p) \tag{15.3.1384}$$

So we deduce:

$$\frac{\partial\sigma_p}{\partial X_i}=\frac{\text{cov}(R_i,R_p)}{\sigma_p} \tag{15.3.1385}$$

The reader will notice that we fall back on the beta coefficient. Therefore, we can write (by the way we prefer to write the weights according to the most usual notation w_i instead of X_i):

$$\frac{\partial \sigma_p}{\partial w_i}=\frac{\text{cov}\left(R_i,R_p\right)}{\sigma_p}=\beta_i\sigma_p \tag{15.3.1386}$$

Thus, remembering that the relative VaR is given by:

$$\text{VaR}_r=\alpha\sigma \tag{15.3.1387}$$

The "marginal Value at Risk" is then logically given by:

$$\boxed{\Delta\text{VaR}_i=\frac{\partial \text{VaR}}{\partial w_i}=\frac{\partial \alpha\sigma_p}{\partial w_i}=\alpha\frac{\partial \sigma_p}{\partial w_i}=\alpha\frac{\text{cov}\left(R_i,R_p\right)}{\sigma_p}=\alpha\beta_i\sigma_p} \tag{15.3.1388}$$

This gives for a given asset having a given weight w_i in a given portfolio the percentage increase in its contribution to the portfolio's risk when increases its weight by 1%.

We define the "component Value at Risk" as:

$$\boxed{\text{CVaR}_i=w_i\beta_i\alpha\sigma_p} \tag{15.3.1389}$$

Let us recall for close this subject what is the purpose of the VaR? First of all, it is of great utility since it is measured in nominal terms. Once a financial institution has calculated its aggregate VaR, that is, the maximum loss that it can incur over the entire balance sheet for a predetermined probability, it is entitled to use that amount to determine the minimum capital (equity) it must hold in order not to become bankrupt. If it holds less capital and the maximum probability loss occurs, its equity will be negative and it may have to file for bankruptcy. Many studies have proven that this tool is statistically much better that any known qualitative approach encompassing all well known human cognitive bias!

The VaR is therefore very useful for a financial institution because it enables it to determine the level of capital it must maintain in order to survive with a given probability. When VaR is used for this purpose, we refer to it more commonly as "Capital at Risk", which means that the capital that a financial institution must hold is calculated or assessed according to the risks to which it is exposed. The greater the risk, the greater the need to maintain capital. This seems reasonable, because the capital held by a financial institution is first and foremost a safety net. For a bank, it aims to protect deposits in its liabilities. VaR is therefore an appropriate measure to define the regulatory capital that a financial institution must hold. For this reason, the Basel Committee, under the umbrella of the Bank for International Settlements, used this measure to calculate the regulatory capital of a depository institution in 11995 (holocene calendar) and became effective in January 11998 (holocene calendar). They now have to calculate their exposure Risk by using VaR and testing its accuracy by performing stress tests (as opposed to calculations to extreme variations) as well as to back testing by verifying that large deviations (outside the range of confidence) do not occur more than 5 times per stock market year (thus compared to historical data).

15.3.10.3.14.10 Conditional Value at Risk (CVaR)

A forecast of financial losses often used in the academic world of finance is the "conditional loss", often referred to "expected shortfall" or "expected loss" or "expected tail loss" (ETL).

For the sake of simplification, we often limit ourselves to the calculation of the conditional loss of the VaR in the case that it follows a Normal law (we speak then in the field of finance of an "expected shortfall for Gaussian loss distribution") and this is what we are going to do here. The reader will find, however, a more realistic case based on a Pareto type I law in the Probability section when we introduced the concept of conditional expectation.

Let us consider therefore the expected shortfall at a threshold α corresponding to the percentile such as we have seen earlier above. Consistent with what we have seen in the Probability section (see page 499), the conditional expectation of such a threshold will therefore be given by:

$$\text{ES}_\alpha=\frac{\displaystyle\int\limits_{\text{VaR}_\alpha}^{+\infty} yf(y)\mathrm{d}y}{\displaystyle\int\limits_{\text{VaR}_\alpha}^{+\infty} f(y)\mathrm{d}y} \tag{15.3.1390}$$

For the denominator, it is immediate (it is simply the cumulative probability at the level of the quantile threshold of the selected VaR):

$$\boxed{\text{ES}_\alpha=\frac{\displaystyle\int\limits_{\text{VaR}_\alpha}^{+\infty} yf(y)\mathrm{d}y}{1-\alpha}} \tag{15.3.1391}$$

As in finance we can always center and reduce a random variable, let us focus on the case of a centered reduced Normal law to treat the numerator (hence VaR_α the quantile of the centered reduced VaR to the threshold α). We then have:

$$\begin{aligned}\text{ES}_\alpha&=\frac{\displaystyle\int\limits_{\text{VaR}_\alpha}^{+\infty} y\frac{1}{\sigma\sqrt{2\pi}}e^{\left(-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right)}\mathrm{d}y}{1-\alpha}=\frac{\displaystyle\int\limits_{\text{VaR}_\alpha}^{+\infty} y\frac{1}{\sigma\sqrt{2\pi}}e^{\left(-\frac{1}{2}y^2\right)}\mathrm{d}y}{1-\alpha}\\&=\frac{1}{(1-\alpha)\sqrt{2\pi}}\int\limits_{\text{VaR}_\alpha}^{+\infty} ye^{\left(-\frac{1}{2}y^2\right)}\mathrm{d}y=\frac{1}{(1-\alpha)\sqrt{2\pi}}e^{\left(-\frac{1}{2}y^2\right)}\Bigg|_{\text{VaR}_\alpha}^{+\infty}=-\frac{e^{\left(-\frac{1}{2}\text{VaR}_\alpha^2\right)}}{(1-\alpha)\sqrt{2\pi}}\\&=-\frac{\frac{1}{\sqrt{2\pi}}e^{\left(-\frac{1}{2}\text{VaR}_\alpha^2\right)}}{1-\alpha}=(1-\alpha)^{-1}\Phi\left(\text{VaR}_\alpha^2\right)\end{aligned} \tag{15.3.1392}$$

What is computed with a spreadsheet like Microsoft® Excel® 14.0.7184:

```
=-NORMDIST(NORMINV(alpha),FALSE)/(1-alpha)
```

Or in an equivalent manner:

```
=NORMDIST(NORMINV(1-alpha),FALSE)/(1-alpha)
```

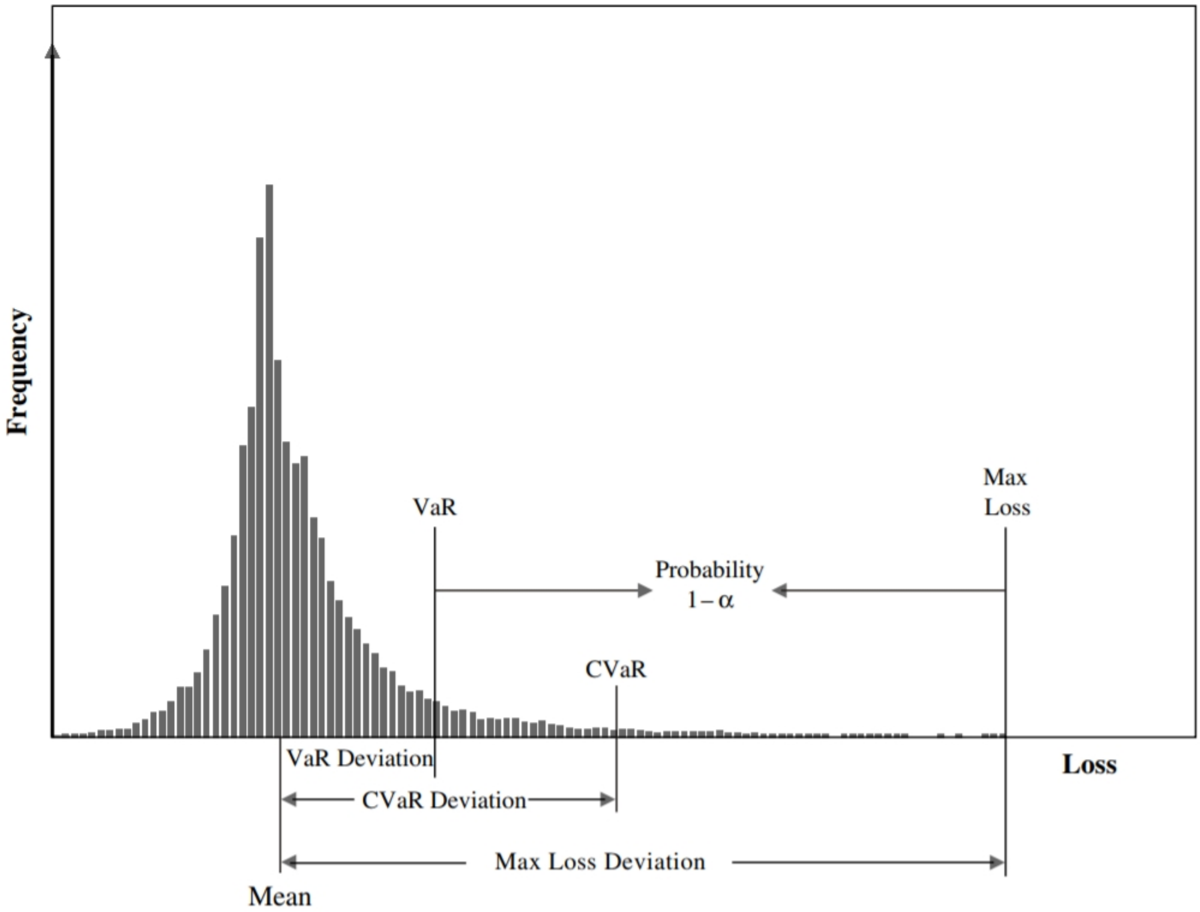


Figure 15.194: Graphical representation of VaR, VaR Deviation, CVaR, CVaR Deviation, Max Loss, and Max Loss Deviation (source: Value-at-Risk vs. Conditional Value-at-Risk in Risk Management and Optimization)

There is a close correspondence between CVaR and VaR: with the same confidence level, VaR is a lower bound for CVaR. The problem of the choice between VaR and CVaR, especially in financial risk management, has been quite popular in academic literature. Reasons affecting the choice between VaR and CVaR are based on the differences in mathematical properties, stability of statistical estimation, simplicity of optimization procedures, acceptance by regulators, etc

✍**Example:**

Given a VaR following a Normal law that has been centered and reduced. The conditional loss expectation above 99% (i.e. at the threshold of $\alpha=1\%$) is then given by:

```
=-NORM.S.DIST(NORM.S.INV(99%),FALSE)/(1-99%)=2.67
```

or:

```
=-NORM.S.DIST(2.326,FALSE)/1%=2.67
```

Which is still 15% more than the quantile of 2.326 !!!

This corresponds in comparison to the relative VaR to an equivalent of:

`=1-NORM.S.INV(2.67;1)=99.6%`

So through this example we see that there is no "problem" in working with a centered reduced conditional expected loss since we can always reduce the conditional expectation of a loss to a given relative VaR (that latter being already not accurate...).

After the choice of reserves is purely from the internal policy to the financial institutes.

15.3.10.3.14.11 Modified duration Value At Risk (MdVaR)

We have seen so far how to calculate the relative value of a portfolio whose variation is implicitly assumed to be directly proportional to a change in the rate of return (market rate). There are, however, many types of portfolios or financial instruments whose change in value is not simply proportional to a change in the rate of return, but with a more complex relationship, typically bonds or portfolios (made of bonds) where the yield is at the denominator in the discrete case and in an exponential in the continuous case.

Indeed, remember that for a single bond we have proved (see page 7625) that the present value was given by:

$$P=\sum_{t=1}^{T}\frac{\mathrm{CF}_t}{(1+y)^t} \tag{15.3.1393}$$

We have seen so far how to calculate the relative value of a portfolio whose variation is implicitly assumed to be directly proportional to a change in the rate of return (market rate). There are, however, many types of portfolios or financial instruments whose change in value is not simply proportional to a change in the rate of return, but with a more complex relationship, typically bonds or portfolios (made of bonds) where the yield is at the denominator in the discrete case and in an exponential in the continuous case.

Indeed, remember that for a single bond we have we have proved that the present value was given by (see earlier page 7626):

$$P=\sum_{t=1}^{T}\frac{\mathrm{CF}_t}{(1+y)^t} \tag{15.3.1394}$$

However, in our study of modified duration earlier above (see page 7633), we have proved that in Taylor's first order approximation that we could write traditionally using Modified Duration rather than Macaulay's duration that:

$$\Delta P\cong\frac{\mathrm{d}P}{\mathrm{d}y}\Delta y=(-D^*\cdot P)\Delta y=\left(\sum_{t=1}^{T}t\frac{\mathrm{CF}_t}{(1+y)^{t+1}}\right)\Delta y \tag{15.3.1395}$$

The idea underlying the "modified duration VaR" (which we can also be named "delta VaR") is then to take the y of the day and the historical variance of the variations of the returns assuming they follow a Normal distribution at a given confidence level such that:

$$\boxed{\text{VaR}_{D^*} \cong \left(\frac{1}{P}\sum_{t=1}^{T} t\frac{\text{CF}_t}{(1+y)^{t+1}}\right)\alpha\sigma_{\Delta y} = (-D^* \cdot P)\alpha\sigma_{\Delta y}} \quad (15.3.1396)$$

and we can easily extend the principle by taking the convexity term and hence we have a delta-gamma VaR.

15.3.10.3.14.12 Backtesting Value at Risk

We will see here a naive backtesting model of the VaR (there are many) but when the reader goes further through the predictive models (moving average, exponential smoothing, ARIMA, etc.) we will see that we will compare these models with the same values using error measurements (ME, MAD, MDS, MPE). These comparisons are then considered as a form of backtesting.

So to go back to the VaR, it is relatively obvious that when we have just put in place an annual VaR model and we have little history it is better to bring it back to a daily VaR and do backtesting on the basis of a smaller unit.

Let us recall the example of the beginning where we had a relative daily VaR and annual respectively at the threshold of 1% (thus 99% coverage) of:

$$\begin{aligned}\text{VaR}_d &\cong 14.65.\text{-}\\ \text{VaR}_A &\cong 232.6.\text{-}\end{aligned} \quad (15.3.1397)$$

Suppose we do a backtesting of this model by comparing it to the 600-day data we have in our possession. We observe 9 values above 14.65.-. What can we conclude ???

In fact we must first know that if the N days are independent and we denote by p the probability that one of the days is out of VaR (thus the associated value equal to 1% in our example!), then we are dealing to an experiment of N Bernoulli's tests. It is therefore a Binomial law whose associated cumulative probability is for recall (see section Statistics page 658):

$$\sum_{k=0}^{N} C_k^N p^k(1-p)^{N-k} = \sum_{k=0}^{N}\frac{N!}{k!(N-k)!}p^k(1-p)^{N-k} \quad (15.3.1398)$$

Hence, the expected mean is given by (it's nice to calculate it but in reality it is not very useful):

$$\text{E}(X) = Np = 600 \cdot 1\% = 6 \quad (15.3.1399)$$

However, to know the cumulative probability associated with 9 events, we calculate for example with Microsoft® Excel® 14.0.6129:

```
=BINOM.DIST(9;600;1%;TRUE)=BINOM.DIST(9;600;1%;1)=91.71%
```

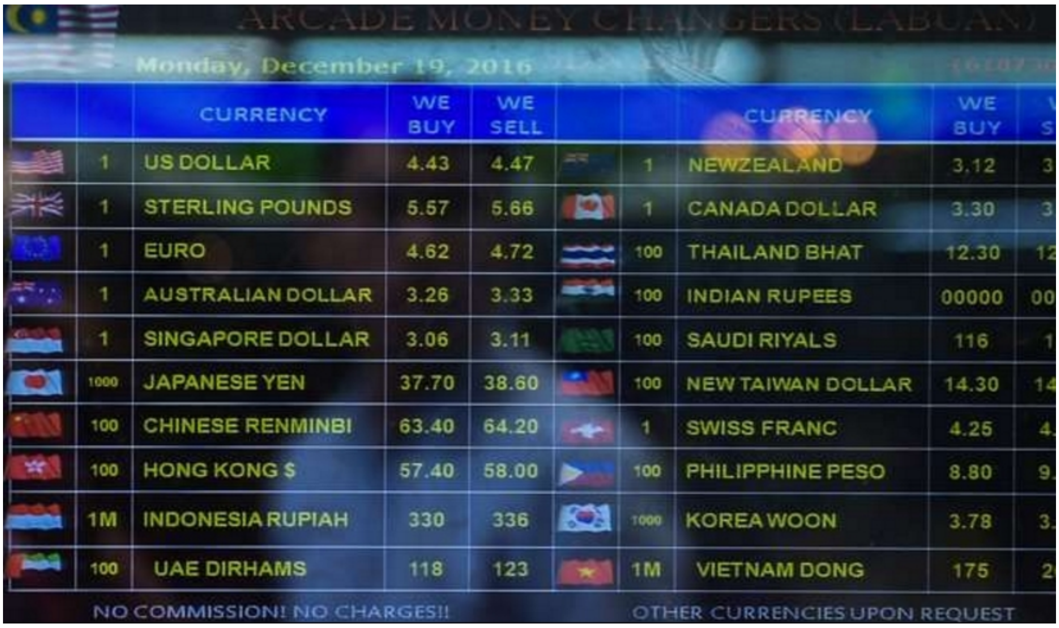
That is, a p-value of 8.29%. So whether it is unilateral or bilateral, this hypothesis test leads us not to reject the VaR model. It should not be forgotten that if N is very large and the probability is small, the binomial law tends towards a Poisson law (see section Statistics page 686) which is easier to manipulate (this is the recommendation of the CreditRisk+ standard).

15.3.11 Exchange rate

In finance, an "exchange rate" (also known as a "foreign-exchange rate", "ForEx rate") between two currencies is the rate at which one currency will be exchanged for another. It is also regarded as the value of one country's currency in relation to another currency. For example, an interbank exchange rate of 119 Japanese yen (JPY, ¥) to the United States dollar (US$) means that ¥119 will be exchanged for each US$1 or that US$1 will be exchanged for each ¥119. In this case it is said that the price of a dollar in relation to yen is ¥119, or equivalently that the price of a yen in relation to dollars is $1/119.

Exchange rates are determined in the foreign exchange market, which is open to a wide range of different types of buyers and sellers, and where currency trading is continuous: 24 hours a day except weekends, i.e. trading from 20:15 GMT on Sunday until 22:00 GMT Friday. The spot exchange rate refers to the current exchange rate. The forward exchange rate refers to an exchange rate that is quoted and traded today but for delivery and payment on a specific future date.

In the retail currency exchange market, different buying and selling rates will be quoted by money dealers. Most trades are to or from the local currency. The buying rate is the rate at which money dealers will buy foreign currency, and the selling rate is the rate at which they will sell that currency. The quoted rates will incorporate an allowance for a dealer's margin (or profit) in trading, or else the margin may be recovered in the form of a commission or in some other way.



ARCADE MONEY CHANGERS (LABUAN)

Monday, December 19, 2016

		CURRENCY	WE BUY	WE SELL			CURRENCY	WE BUY	S
	1	US DOLLAR	4.43	4.47		1	NEWZEALAND	3.12	3
	1	STERLING POUNDS	5.57	5.66		1	CANADA DOLLAR	3.30	3
	1	EURO	4.62	4.72		100	THAILAND BHAT	12.30	12
	1	AUSTRALIAN DOLLAR	3.26	3.33		100	INDIAN RUPEES	00000	00
	1	SINGAPORE DOLLAR	3.06	3.11		100	SAUDI RIYALS	116	1
	1000	JAPANESE YEN	37.70	38.60		100	NEW TAIWAN DOLLAR	14.30	14
	100	CHINESE RENMINBI	63.40	64.20		1	SWISS FRANC	4.25	4.
	100	HONG KONG $	57.40	58.00		100	PHILIPPHINE PESO	8.80	9.
	1M	INDONESIA RUPIAH	330	336		1000	KOREA WOON	3.78	3.
	100	UAE DIRHAMS	118	123		1M	VIETNAM DONG	175	2

NO COMMISSION! NO CHARGES!! OTHER CURRENCIES UPON REQUEST

Different rates may also be quoted for cash (usually notes only), a documentary form (such as traveller's cheques) or electronically (such as a credit card purchase). The higher rate on documentary transactions has been justified as compensating for the additional time and cost of clearing the document. On the other hand, cash is available for resale immediately, but brings security, storage, and transportation costs, and the cost of tying up capital in a stock of banknotes (bills).

Remarks

R1. The "real exchange rate" (RER) is the purchasing power of a currency relative to another at current exchange rates and prices. It is the ratio of the number of units of a given country's currency necessary to buy a market basket of goods in the other country, after acquiring the other country's currency in the foreign exchange market, to the number of units of the given country's currency that would be necessary to buy that market basket directly in the given country. There are various ways to measure RER.

R2. For obvious financial risk reasons that take foreign exchange business (and also because of money usability), when you travel for holidays, it will be the huge majority of times (excepted for very special cases like countries that are on the way to be in bankruptcy or with very weak money) at your advantage to convert your local cash money into a foreign cash money in the country related to the local money (for example, if you have CHF and you want to convert them into USD because you will go in the USA for holidays, then it would be at your advantage to make this conversion in Switzerland, and before coming back to Switzerland to change your USD into CHF in the USA). Obviously the rate won't be the same as they change on a daily basis or even hourly. Also exchanging money at local banks is not only convenient but also much less expensive than exchanging cash at airports or hotels.

R3. When paying by credit card, pay in the local currency! If you don't, not only will you get charged a mediocre rate to convert to your home currency, you will also be charged a percentage for the privilege. Travelex, the currency exchange company that operates kiosks in airports around the world in the beginning of the 121st century (holocene calendar), was offering an exchange rate of 1 EUR for 1.26 USD in June 12018 (holocene calendar). On the same day, Mastercard's exchange rate was 1 EUR for 1.17 USD the same day. Notice that paying with credit card in the local currency is most of the time at your advantage in comparison at when you change cash (especially because the human working at the cash exchange office are more expensive than the computer algorithms doing the transaction electronically).

The world has over 150 different currencies, from the Afghanistan afghani and the Albanian lek all the way through the alphabet to the Zambian kwacha and the Zimbabwean dollar. For international economic transactions, households or firms will wish to exchange one currency for another. Perhaps the need for exchanging currencies will come from a German firm that exports products to Russia, but then wishes to exchange the Russian rubles it has earned for euros, so that the firm can pay its workers and suppliers in Germany. Perhaps it will be a South African firm that wishes to purchase a mining operation in Angola, but to make the purchase it must convert South African rand to Angolankwanza. Perhaps it will be an American tourist visiting China, who wishes to convert U.S. dollars to Chinese yuan to pay the hotel bill.

Exchange rates can sometimes change very swiftly. For example, in the United Kingdom the pound was worth 2 USD in spring 12008 (holocene calendar), but was worth only 1.40 USD six months later. For firms engaged in international buying, selling, lending, and borrowing, these swings in exchange rates can have an enormous effect on profits.

The quantities traded in foreign exchange markets are breathtaking. A survey done in April, 12013 (holocene calendar) by the Bank of International Settlements, an international organization for banks and the financial industry, found that 5.3 USD trillion per day was traded on foreign exchange markets (in comparison in 12007 - according to holocene calendar - the daily average trading volume DATV on derivatives exchanges was about $ 2 trillion), which makes the foreign exchange market the largest market in the world economy. In contrast, 12013 (holocene calendar) U.S.A. real Gross Domestic Product (GDP) was $15.8 trillion per year.

Currency	% Daily share
U.S. dollar	87.0%
Euro	33.4%
Japanese yen	23.0%
British pound	11.8%
Australian dollar	8.6%
Swiss franc	5.2%
Canadian dollar	4.6%
Mexican peso	2.5%
Chinese yuan	2, 2%

Table 15.54: Currencies traded most on Foreign Exchange Markets as of April, 12013 (source: `http://www.bis.org/publ/rpfx13fx.pdf`)

Very approximately, about one-tenth of the volume traded is used for hedging. The remaining trades must be for speculative purposes.

Portfolio investment is often linked to expectations about how exchange rates will shift. Look at a U.S. financial investor who is considering purchasing bonds issued in the United Kingdom. For simplicity, ignore any interest paid by the bond (which will be small in the short run anyway) and focus on exchange rates. Say that a British pound is currently worth 1.50 USD. However, the investor believes that in a month, the British pound will be worth 1.60 USD. Thus this investor would change 24, 000 Gross Domestic Product for 16, 000 British pounds. In a month, if the pound is indeed worth 1.60 USD, then the portfolio investor can trade back to U.S. dollars at the new exchange rate, and have 25, 600 USD... a nice profit! A portfolio investor who believes that the foreign exchange rate for the pound will work in the opposite direction can also invest accordingly. Say that an investor expects that the pound, now worth 1.50 USD, will decline to 1.40 USD. Then that investor could start off with 20, 000 in British pound (borrowing the money if necessary), convert it to 30, 000 USD, wait a month, and then convert back to approximately 21, 429 British poind making again making a nice profit! Of course, this kind of investing comes without guarantees, and an investor will suffer losses if the exchange rates do not move as predicted.

Many portfolio investment decisions are not as simple as betting that the value of the currency will change in one direction or the other. Instead, they involve firms trying to protect themselves from movements in exchange rates. Imagine you are running a U.S. firm that is exporting to France. You have signed a contract to deliver certain products and will receive 1 million euros a year from now. But you do not know how much this contract will be worth in USD, because the dollar/euro exchange rate can fluctuate in the next year. Let's say you want to know for sure what the contract will be worth, and not take a risk that the euro will be worth less in USD than it currently is. You can hedge, which means using a financial transaction to protect yourself against a risk from one of your investments (in this case, currency risk from the contract). Specifically, you can sign a financial contract and pay a fee that guarantees you a certain exchange rate one year from now regardless of what the market exchange rate is at that time. Now, it is possible that the euro will be worth more in dollars a year from now, so your hedging contract will be unnecessary, and you will have paid a fee for nothing. But if the value of the euro in dollars declines, then you are protected by the hedge. Financial contracts like hedging, where parties wish to be protected against exchange rate movements, also commonly lead to a series of portfolio investments by the firm that is receiving a fee to provide the hedge.

With the price of a typical good or service, it is clear that higher prices benefit sellers and hurt buyers, while lower prices benefit buyers and hurt sellers. In the case of exchange rates, where the buyers and sellers are not always intuitively obvious, it is useful to trace through how different participants in the market will be affected by a stronger or weaker currency. Consider, for example, the impact of a stronger USD on six different groups of economic actors, as shown in the table below:

Economic actors	**A stronger U.S. dollar**	**A weaker U.S. dollar**
A U.S. exporting firm	☹	☺
A foreign firm exporting to the U.S.	☺	☹
A U.S. tourist abroad	☺	☹
A foreign tourist in the U.S.	☹	☺
A U.S. investor abroad	☹	☺
A foreign investor in the U.S.	☺	☹

However, the foreign exchange value is not always determined by market forces, indeed, we distinguish 4 types of foreign exchange policies:

1. When the ForEx is completely determined by market forces, we speak of "floating exchange rates"
2. When the ForEx is usually determined by market, but central bank sometimes intervenes, we speak then of "soft exchange rate pegs"
3. When the ForEx is fixed at a certain level by in central bank we speak then of "hard exchange rate pegs"
4. When the currency of a country is made identical to the currency of another country we speak then of "merging currencies"

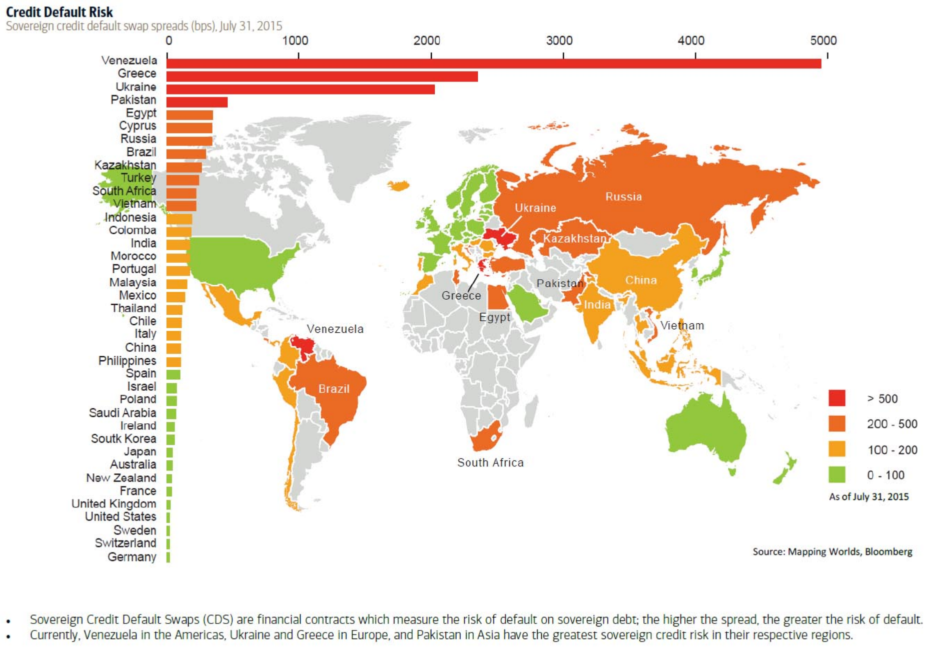
15.3.12 Credit default Risk

Credit risk is the risk involved in all transactions of borrowing and lending. If you lend money, what are the probabilities that the borrower will make the repayment? Credit risk then refers to the risk that the counterparty may fail to meet its obligations in the agreed terms. Quantifying it accurately is of immense importance to all: credit risk managers, regulators, and investors.

In case the borrower is likely to promptly return the money, then you have low credit risk, and in case there is a high probability of default than you have a high credit risk. So each borrowing and lending has its own element of risk involved.

Credit risk can be classified in the following way:

- "Credit default risk": The risk of loss arising from a debtor being unlikely to pay its loan obligations in full or the debtor is more than 90 days past due on any material credit obligation; default risk may impact all credit-sensitive transactions, including loans, securities and derivatives.
- "Concentration risk": The risk associated with any single exposure or group of exposures with the potential to produce large enough losses to threaten a bank's core operations. It may arise in the form of single name concentration or industry concentration.
- "Credit spread risk": The risk occurring due to volatility in the difference between investments' interest rates and the risk free return rate.
- "Country risk": The risk of loss arising from a sovereign state freezing foreign currency payment (transfer/conversion risk) or when it defaults on its obligations (sovereign risk).
- "Country spread risk": The risk occurring due to volatility in the difference between investments' interest rates and the risk free return rate.



Source: Mapping Worlds, Bloomberg

- Sovereign Credit Default Swaps (CDS) are financial contracts which measure the risk of default on sovereign debt; the higher the spread, the greater the risk of default.
- Currently, Venezuela in the Americas, Ukraine and Greece in Europe, and Pakistan in Asia have the greatest sovereign credit risk in their respective regions.

Figure 15.195: **Some countries spread risk**

If we have large enough databases on customer profiles or companies profiles it is possible to set up mathematical models allowing to assign a certain probability of credit default to an economic agent (the most common technique used in the 120th century - according to holocene calendar - being the logistic regression proved in the section of Numerical Methods at page 6449).

15.3.12.1 Credit spread risk

Financial institutions that do not have such databases or the analytical capacity or that just consider that others know how to do better refer most of the times to Standard & Poor's ratings for which a cumulative probability of default over several years i is associated with each rating (the years in the legend are given based on the holocene calendar):

Year/Ranking	AAA	AA	A	BBB	BB	B	CCC/C
1	0.00	0.01	0.04	0.29	1.28	6.24	32.35
2	0.00	0.03	0.13	0.86	3.96	14.33	42.35
3	0.04	0.08	0.26	1.48	7.32	21.57	48.66
4	0.07	0.16	0.43	2.37	10.51	27.47	53.65
5	0.12	0.26	0.66	3.25	13.36	31.81	59.49
6	0.21	0.40	0.90	4.15	16.32	35.47	62.19
7	0.31	0.56	1.16	4.88	18.84	38.71	63.37
8	0.48	0.71	1.41	5.60	21.11	41.69	64.10
9	0.54	0.83	1.71	6.21	23.22	43.92	67.78
10	0.62	0.97	2.01	6.95	24.84	46.27	70.8

Table 15.55: **Standard & Poor rating table between 11981 and 12004**

From the cumulated probability of being in bankrupt at year i, we can determine, as we shall see, the conditional probability of bankrupt at year i, knowing that we survived until year i (that is, the proportion who are insolvent in year i among those who are still "alive" at the beginning of year i).

If c_n is the cumulative probability of going bankrupt between the year 0 and the year n, then we have that $1-c_n$ which is of course the cumulative probability of having survived between year 0 and n. It comes then quite intuitively (we guess by extension that we can apply the biometric function techniques seen in the section of Dynamics of Populations to the study of Credit Risk default):

$$s_n=1-c_n=(1-c_{n-1})(1-d_n)=(1-c_{n-2})(1-d_{n-1})(1-d_n)=\prod_{i=1}^{n}(1-d_i) \qquad (15.3.1400)$$

Thus we immediately see that it is easy to obtain the conditional probability d_n of insolvency in year n among those who are still "alive" at the beginning of year n:

$$d_n=1-\frac{1-c_n}{1-c_{n-1}}=\frac{c_n-c_{n-1}}{1-c_{n-1}} \qquad (15.3.1401)$$

The trap here would be to make use not of conditional probabilities but to calculate the difference:

$$c_n-c_{n-1} \qquad (15.3.1402)$$

which corresponds to the proportion of insolvent firms in year n among all (those that have been insolvent and will be insolvent).

Moreover we can notice that this error is dangerous since from what we obtained above it follows immediately that:

$$d_n \geq c_n - c_{n-1} \tag{15.3.1403}$$

So as I always say: in any case all the risks that one calculates - that the mathematical model is estimated as correct or not - always underestimate the real risks because there exists a quantity of risk that one will not have thought or that one has rightly or wrongly considered it unlikely.

✍**Examples:**

E1. In the second year, for a company (or a country) rated BBB, we then have the conditional probability of default which is:

$$d_n = 1 - \frac{1-c_n}{1-c_{n-1}} = 1 - \frac{1-0.86\%}{1-0.29\%} \cong 0.86\% \tag{15.3.1404}$$

E2. A borrower is rated BBB and we know that in case of default we are to be assured to be able to recover at 37% of the sum invested (it is necessary to know that in practice this rate is also a statistical distribution ...). We can for example for 5 years maturity bonds and a total invested amount of 100 million and by taking in the previous rating table $c_5 = 3.25\%$ calculate that the estimated credit loss is then:

$$100 \cdot 10^6 \cdot 3.25\%(1-37\%) = 2,050,000.\text{-} \tag{15.3.1405}$$

There is a nice thing to notice when using this type of rating: indeed, the cumulative probability of default sometimes used by asset sellers can be estimated indirectly in comparison to risk-free returns and assuming the guaranteed $\%$ income in case of default. Let us see this with a simple case already partially known.

Suppose we want to update a zero coupon bond at the risk-free market rate. We then write as usual:

$$P = \frac{N}{(1+r)^n} \tag{15.3.1406}$$

where N is the nominal (often taken in practice as equal to 100.- as we already know).

For a zero coupon bond that has the cumulative probability c_n of default and denoting by f the guaranteed fraction of the amount in the case of default, a neutral risk market would naturally require that (the expression on the right of the equality is only an expected mean):

$$P = \frac{N}{(1+r)^n} = \frac{N}{(1+r_f)^n}(1-c_n) + \frac{fN}{(1+r_f)^n}c_n \tag{15.3.1407}$$

where r_f is the free risk market rate.

Therefore, by simplifying the nominal disappears:

$$(1+r_f)^n = (1+r)^n\left[(1-c_n)+fc_n\right] = (1+r)^n[1-c_n(1-f)] \tag{15.3.1408}$$

Thus we can play according to what we know to determine different parameters as for example for the most common and interesting cases:

$$\boxed{r=\frac{1+r_f}{\sqrt[n]{1-c_n(1-f)}}-1} \tag{15.3.1409}$$

where we notice for the both extreme cases that in accordance with intuition in the first case where the investment is 100% safe and in the second case where the investment is 100% risky:

$$\begin{gathered}r=\frac{1+r_f}{\sqrt[n]{1-0\%\cdot(1-f)}}-1=r_f\\ r=\frac{1+r_f}{\sqrt[n]{1-100\%\cdot(1-f)}}-1=\frac{1+r_f}{\sqrt[n]{f}}-1\xrightarrow[f\to 0]{}+\infty\end{gathered} \tag{15.3.1410}$$

So it makes sense that the higher our risk, the higher the required return.

We can also deduce the following very interesting equality (since it makes it possible to judge the reliability of a borrower):

$$\boxed{c_n=\frac{1}{1-f}\left[1-\left(\frac{1+r_f}{1+r}\right)^n\right]} \tag{15.3.1411}$$

With the frequent case $f=0$ where this obviously reduces to:

$$c_n=1-\left(\frac{1+r_f}{1+r}\right)^n \tag{15.3.1412}$$

It seems that some practitioner refer to the prior-previous relation as the "estimate reduced model of the implicit credit default".

15.3.12.2 Value At Risk equity coverage rating

The "VaR equity coverage rating" is based on the empirical basis that if an organization wishes to be rated in a certain way by Moody's then according to the Normal loss distribution model (but there are other far more pessimistic models that go so far to a multiple factor from 2 to 4) it can calculate the reserves it must have in order to enter a certain rating class.

Consider, for example, the following Moody's table:

Desired Rating (Moody's)	Probability of default to 1 year (α)	Corresponding Z
Aaa	0.01%	3.72
Aa1	0.02%	3.54
Aa2	0.02%	3.54
Aa3	0.03%	3.43
A1	0.05%	3.29
A2	0.06%	3.24
A3	0.07%	3.19
Baa1	0.13%	3.01
Baa2	0.16%	2.95
Baa3	0.70%	2.46
Ba1	1.25%	2.24
Ba2	1.79%	2.10
Ba3	3.96%	1.76
B1	6.14%	1.54
B2	8.31%	1.38
B3	15.08%	1.03

Table 15.56: Moody's one year risk default

Suppose now that an organization has its profit and loss account that historically has a standard deviation of $70,000$.-. Then the fractional reserves that it should have to be rated Aa2 are:

$$\text{VaR}_{\text{e.c.}} = \alpha\sigma = 3.54 \cdot 70000 = 247,800.\text{-} \tag{15.3.1413}$$

For the people interested here is a grid with the equivalence of different Risk Metrics:

Moody's		S&P		Fitch		
Long-term	**Short-term**	**Long-term**	**Short-term**	**Long-term**	**Short-term**	
Aaa	P-1	AAA	A-1+	AAA	F1+	Prime
Aa1		AA+		AA+		High grade
Aa2		AA		AA		
Aa3		AA-		AA-		
A1		A+	A-1	A+	F1	Upper medium grade
A2		A		A		
A3	P-2	A-	A-2	A-	F2	
Baa1		BBB+		BBB+		Lower medium grade
Baa2	Not prime	BBB	A-3	BBB	F3	
Baa3		BBB-		BBB-		
Ba1		BB+	B	BB+	B	Non-investment grade speculative
Ba2		BB		BB		
Ba3		BB-		BB-		
B1		B+		B+		Highly speculative
B2		B		B		
B3		B-		B-		
Caa1		CCC+	C	CCC	C	Substantial risks
Caa2		CCC				Extremely speculative
Caa3		CCC-				Default imminent with little prospect for recovery
Ca		CC C				
C				DDD		
/		D	/	DD	/	In default
/				D		

Table 15.57: Moody's, S&P and Fitch Riskmetrics

15.3.12.3 Altman Z-score

The "Z-score formula for predicting bankruptcy" was published in 11968 (holocene calendar) by Edward I. Altman, who was, at the time, an Assistant Professor of Finance at New York University. The formula may be used to predict the probability that a firm will go into bankruptcy within two years. Z-scores are used to predict corporate defaults and an easy-to-calculate control measure for the financial distress status of companies in academic studies. The Z-score uses multiple corporate income and balance sheet values to measure the financial health of a company.

The Z-score is a linear combination of four or five common business ratios, weighted by coefficients. The coefficients were estimated by identifying a set of firms which had declared bankruptcy and then collecting a matched sample of firms which had survived, with matching by industry and approximate size (assets).

Altman applied the statistical method of discriminant analysis to a dataset of publicly held manufacturers. The estimation was originally based on data from publicly held manufacturers, but has since been re-estimated based on other datasets for private manufacturing, non-manufacturing and service companies.

The original data sample consisted of 66 firms, half of which had filed for bankruptcy under Chapter 7 of the USA laws. All businesses in the database were manufacturers, and small firms with assets of $< \$1$ million were eliminated.

The original Z-score formula was as follows (there exist also some variation for private firms, non-manufacturers & emerging markets):

$$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1.0X_5 \tag{15.3.1414}$$

with:

- X_1: Working capital / total assets. Measures liquid assets in relation to the size of the company.
- X_2: Retained earnings / total assets. Measures profitability that reflects the company's age and earning power.
- X_3: Earnings before interest and taxes / total assets. Measures operating efficiency apart from tax and leveraging factors. It recognizes operating earnings as being important to long-term viability.
- X_4: Market value of equity / book value of total liabilities. Adds market dimension that can show up security price fluctuation as a possible red flag.
- X_5: Sales / total assets. Standard measure for total asset turnover (varies greatly from industry to industry).

Based on the ranges of Z-scores, we can classify public firms into the following four categories:

Z-score range	Definition
> 3.0	Safe
2.7 to 2.99	On alert
1.8 to 2.7	Good chances of going bankrupt within two years
< 1.80	Probability of financial distress is very high

Table 15.58: **Altman's Z-score criteria ranges**

Eidelman (11975 according to holocene calendar) found that the Z-score correctly predicted 72% of bankruptcies two years prior to the event.

15.3.12.4 Merton's Distance to Default and KMV Expected Default Frequency

One measure to assess the credit risk of a firm is Distance to Default (DtD), which is measured as the difference between the asset value of the firm and the face value of its debt, scaled by the standard deviation of firm's asset value. It captures how many standard deviations away a firm is from the default. Higher values of DtD imply lesser likelihood of the firm to default on its credit obligations. The measure is derived using the structural default model of Merton (11974 according to holocene calendar), also referred to as the "Merton's model", which exploits the interpretation of equity as a Call option on the firm's underlying assets.

The KMV model for Kealhofer, McQuown and Vasicek, who found a company focusing on measuring default risks is an important example of industry model derive from Merton's one. It was first introduced in the late 11980's (according to holocene calendar) by and is one of the most important methods to estimate the probability of default for a given company by using its balance-sheet information and the equity-market information[32] (`http://www.moodyskmv.com`). Form a purely theoretical point of view, as we will prove it, the difference between KMV and Merton's models are not dramatic. The KMV model, however relies on an extensive empirical testing and it is implemented using a very large proprietary database.

The objective here will be to estimate the market value of total assets $\mathcal{A}$, and it corresponding volatility $\sigma_{\mathcal{A}}$. The basic idea is to treat the equity of a firm as a Call option and the debt (liabilities) $\mathcal{L}$ as its strike price.

Indeed, suppose $\mathcal{A}$ is the value of the company and $\mathcal{L}$ is the face value of the liabilities. If a firm has a debt (or liability[33]) of 80.- and an equity of 20.-, then the total assets will be $\mathcal{E}=\mathcal{A}-\mathcal{L}=100$.- Assume that the assets jump to 110.-. The debt remains the same while the equity increases to 30.-. On the other hand, if the assets drop to 90.-, the equity $\mathcal{E}$ will be only 10.-. Since the equity holders are the residual claimer, their value position in one year is then:

$$\boxed{\mathcal{E}=\max(\mathcal{A}-\mathcal{L},0)} \qquad (15.3.1415)$$

This is the payoff from a Call option on $\mathcal{A}$ with strike price $\mathcal{L}$. Obviously, the idea it that the default occurs when the value of the firm $\mathcal{A}$, goes below the outstanding debt $\mathcal{L}$ (credit risk is driven by the firm value process).

---

[32] Still in the beginning of the 121st century (holocene calendar), a large number of world financial institutions are subscriber of the model.

[33] The debt refers to borrowed money; the liabilities to an obligation of any kind. All debts are liabilities, but not all liabilities are debts.

Remark

Let us recall that for a Call option, we have the following payoff function:

$$\text{Payoff (Call)} = \max(S_T - K, 0) \quad (15.3.1416)$$

and more generally:

Rule	**Payoff** **(lower limit of the purchase / respectively sale price)**
Call Buyer	$\max(S_T - K, 0)$
Call Seller	$-\max(S_T - K, 0)$
Put Buyer	$\max(K - S_T, 0)$
Put Seller	$-\max(K - S_T, 0)$

The debt (liabilities) holder get:

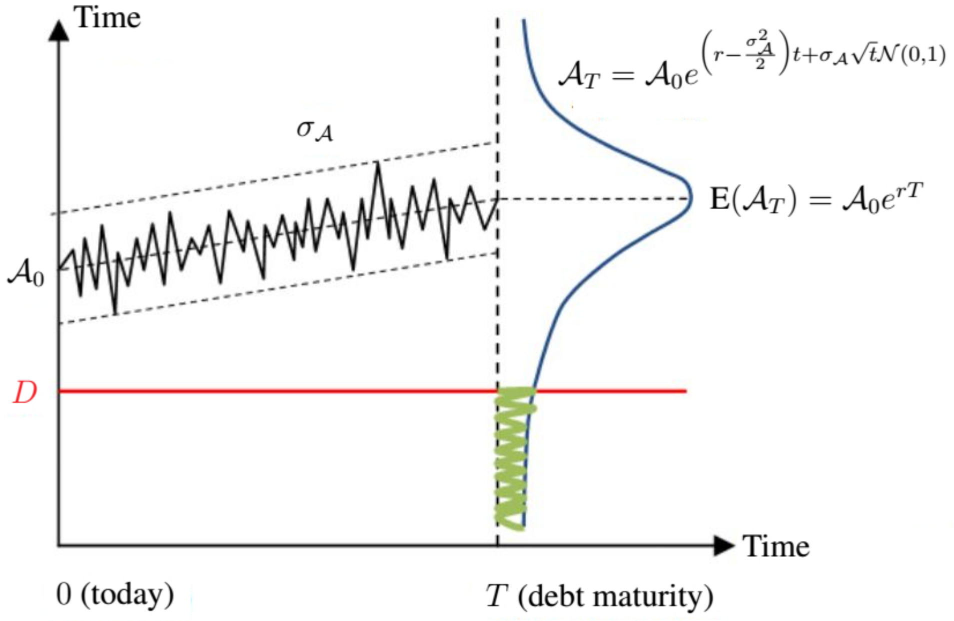
$$\min(\mathcal{A}, \mathcal{L}) = \mathcal{L} - \max(\mathcal{L} - \mathcal{A}, 0) \quad (15.3.1417)$$

Is equivalent to a risk debt $\mathcal{L}$ plus a short guarantee on the assets of the borrower.

Following Merton's model for the valuation of corporate securities we usually consider a simplified case of a firm with risky assets valued today by the market at $\mathcal{A}_0$ level. The value at time t in the future $\mathcal{A}_t$ is uncertain, due to many external and internal factors (economic risk, business risk, foreign exchange risk, industry risk, etc.). Typically we assume, that the returns on the firm's as-sets are distributed normally and their behaviour can be described with Brownian motion formulation (diffusion expression):

$$\frac{\mathrm{d}\mathcal{A}_t}{\mathcal{A}_t} = \mu_\mathcal{A}\mathrm{d}t + \sigma_\mathcal{A}\mathrm{d}W \Leftrightarrow \mathrm{d}\mathcal{A}_t = \mu_\mathcal{A}\mathcal{A}_t\mathrm{d}t + \sigma_\mathcal{A}\mathcal{A}_t\mathrm{d}W \quad (15.3.1418)$$

The symbol $\sigma_\mathcal{A}$ stands for constant assets volatility and $\mu_\mathcal{A}$ for constant drift. Value $\mathrm{d}W$ denotes for random value taken from standardized Normal distribution:


Time
$\sigma_\mathcal{A}$
$\mathcal{A}_T = \mathcal{A}_0 e^{\left(r-\frac{\sigma_\mathcal{A}^2}{2}\right)t+\sigma_\mathcal{A}\sqrt{t}\mathcal{N}(0,1)}$
$\mathrm{E}(\mathcal{A}_T) = \mathcal{A}_0 e^{rT}$
$\mathcal{A}_0$
D
Time
0 (today)
T (debt maturity)

Consequently, the firm's assets value assumed to obey a log-normal diffusion process with a constant volatility is given by:

$$\mathcal{A}_T=\mathcal{A}_0 e^{\left(r-\frac{\sigma_{\mathcal{A}}^2}{2}\right)t+\sigma_{\mathcal{A}}\sqrt{t}\mathcal{N}(0,1)} \tag{15.3.1419}$$

The expected value of the assets at the time t can be given by:

$$\mathrm{E}(\mathcal{A}_t)=\mathcal{A}_0 e^{rt} \tag{15.3.1420}$$

That means that we can treat equity E as a Call option with debt as exercise price. With appropriate notations, we will have the following relations for a firm's equity. KMV model is then defined as below (under the same important assumptions as the Black-Scholes model!!!):

$$\mathcal{E}=C^{\text{BS}}\left(\mathcal{A}_T,\mathcal{L},T,\sigma_{\mathcal{A}},r\right)=\mathcal{A}\Phi\left(d_1\right)-\mathcal{L}e^{-rT}\Phi\left(d_2\right) \tag{15.3.1421}$$

With for recall:

$$d_1=\frac{\ln\left(\dfrac{\mathcal{A}}{\mathcal{L}}\right)+\left(r+\dfrac{1}{2}\sigma_{\mathcal{A}}^2\right)T}{\sigma_{\mathcal{A}}\sqrt{T}} \tag{15.3.1422}$$

$$d_2=d_1-\sigma_{\mathcal{A}}\sqrt{T} \tag{15.3.1423}$$

Let us recall now the diffusion expression:

$$\mathrm{d}\mathcal{E}=\mu_{\mathcal{E}}\mathcal{E}\mathrm{d}t+\sigma_{\mathcal{E}}\mathcal{E}\mathrm{d}W \tag{15.3.1424}$$

and let us write Itō's lemma:

$$\begin{aligned}\mathrm{d}\mathcal{E}&=\frac{\partial\mathcal{E}}{\partial\mathcal{A}}\mathrm{d}\mathcal{A}+\frac{\partial\mathcal{E}}{\partial t}\mathrm{d}t+\frac{1}{2}\sigma_{\mathcal{E}}^2\mathcal{A}^2\frac{\partial^2\mathcal{E}}{\partial\mathcal{A}^2}(\mathrm{d}\mathcal{A})^2\ldots\\&\cong\left(\frac{1}{2}\sigma_{\mathcal{A}}^2\mathcal{E}^2\frac{\partial^2\mathcal{E}}{\partial\mathcal{A}^2}+\mu_{\mathcal{A}}\mathcal{A}\frac{\partial\mathcal{E}}{\partial\mathcal{A}}+\frac{\partial\mathcal{E}}{\partial t}\right)\mathrm{d}t+\sigma_{\mathcal{A}}\mathcal{A}\frac{\partial\mathcal{E}}{\partial\mathcal{A}}\mathrm{d}W\end{aligned} \tag{15.3.1425}$$

When we compare diffusion expressions shown in the two previous relations, then we can regain the relationship that:

$$\sigma_{\mathcal{E}}\mathcal{E}=\sigma_{\mathcal{A}}\mathcal{A}\frac{\partial\mathcal{E}}{\partial\mathcal{A}} \tag{15.3.1426}$$

Therefore:

$$\sigma_{\mathcal{E}}=\sigma_{\mathcal{A}}\frac{\mathcal{A}}{\mathcal{E}}\frac{\partial\mathcal{E}}{\partial\mathcal{A}} \tag{15.3.1427}$$

But we know recognize here the delta Δ and we know it's worth $\Phi(d_1)$, hence:

$$\sigma_{\mathcal{E}}=\sigma_{\mathcal{A}}\frac{\mathcal{A}}{\mathcal{E}}\Delta=\sigma_{\mathcal{A}}\frac{\mathcal{A}}{\mathcal{E}}\Phi(d_1) \tag{15.3.1428}$$

In practice, $\mathcal{E}$, $\sigma_{\mathcal{E}}$, $\mathcal{A}$, T, r are known. To compute $\mathcal{A}$ and $\sigma_{\mathcal{A}}$, we have to solve the system of two non-linear equations:

$$\begin{cases}\mathcal{E}=\mathcal{A}\Phi\left(d_1\right)-\mathcal{L}e^{-rT}\Phi\left(d_2\right)\\ \sigma_{\mathcal{E}}=\sigma_{\mathcal{A}}\dfrac{\mathcal{A}}{\mathcal{E}}\dfrac{\partial\mathcal{E}}{\partial\mathcal{A}}\end{cases} \tag{15.3.1429}$$

to minimize the sum of square error:

$$\boxed{\begin{aligned}&\underset{\mathcal{A},\sigma_{\mathcal{A}}}{\text{minimize}}\ \text{TSSE}=\text{SSE}_1+\text{SSE}_2\\&\qquad\qquad=\left(\mathcal{E}-\mathcal{A}\Phi\left(d_1\right)-\mathcal{L}e^{-rT}\Phi\left(d_2\right)\right)^2+\left(\sigma_{\mathcal{E}}-\sigma_{\mathcal{A}}\frac{\mathcal{A}}{\mathcal{E}}\frac{\partial\mathcal{E}}{\partial\mathcal{A}}\right)^2\\&\text{subject to }\mathcal{A}\geq 0\quad\sigma_{\mathcal{A}}>0\end{aligned}}\tag{15.3.1430}$$

Remark

We can test the result obtained of $\mathcal{A}$ and $\sigma_{\mathcal{A}}$ by plugging back these values into the Black-Scholes options pricing relation. We then compare the obtained value of the option price with the market value of the equity firm. In principal, the two should be the same.

Once the value of $\mathcal{A}$ and $\sigma_{\mathcal{A}}$ are obtained, the "Distance to Default" (DtD) is computed, as already mentioned earlier, in its naive version by:

$$\boxed{\text{DtD}=\frac{\mathcal{A}-\mathcal{L}}{\mathcal{A}\sigma_{\mathcal{A}}}}\tag{15.3.1431}$$

That means the number of standard deviations between the expected asset value at maturity T and the liability threshold $\mathcal{A}$ (that can obviously be chosen rather than calculated if necessary).

We can calculate the probability of DtD by remembering first that:

$$\frac{\mathrm{d}\mathcal{A}}{\mathcal{A}}=r\mathrm{d}t+\sigma_{\mathcal{A}}\sqrt{t}\mathrm{d}Z_t\tag{15.3.1432}$$

Hence (DPT means "Default Point Threshold"):

$$\begin{aligned}P\left(\mathcal{A}\leq\text{DPT}_T\right)&=P\left(\mathcal{A}_0e^{(r-\frac{\sigma_{\mathcal{A}}}{2})T+\sigma_{\mathcal{A}}\sqrt{T}Z_T)}\leq\text{DPT}_T\right)\\&=P\left(\ln(\mathcal{A}_0)+\left(r-\frac{\sigma_{\mathcal{A}}^2}{2}\right)T+\sigma_{\mathcal{A}}\sqrt{T}Z_T\leq\ln(\text{DPT}_T)\right)\\&=P\left(Z_T\leq-\frac{\ln\left(\dfrac{\mathcal{A}_0}{\text{DPT}_T}\right)+\left(r-\dfrac{\sigma^2}{2}\right)T}{\sigma_{\mathcal{A}}\sqrt{T}}\right)=\Phi\left(-d_2^*\right)\end{aligned}\tag{15.3.1433}$$

with:

$$d_2^*=\frac{\ln\left(\dfrac{\mathcal{A}_0}{\text{DPT}_T}\right)+\left(r-\dfrac{\sigma_{\mathcal{A}}^2}{2}\right)T}{\sigma_{\mathcal{A}}\sqrt{T}}\tag{15.3.1434}$$

And it's not uncommon to find the definition of the DtD as the following quantity in the literature:

$$\text{DtD}=-\frac{\ln\left(\dfrac{\mathcal{A}_0}{\text{DPT}_T}\right)+\left(r-\dfrac{1}{2}\sigma_{\mathcal{A}}^2\right)T}{\sigma_{\mathcal{A}}\sqrt{T}}\tag{15.3.1435}$$

Therefore the relationship between DtD and DP (default probability) is given by "Merton's DP":

$$\boxed{\mathrm{DP}=\Phi(\mathrm{DtD})=\Phi\left(-\frac{\ln\left(\dfrac{\mathcal{A}_0}{\mathrm{DPT}_T}\right)+\left(r-\dfrac{1}{2}\sigma_{\mathcal{A}}^2\right)T}{\sigma_{\mathcal{A}}\sqrt{T}}\right)} \tag{15.3.1436}$$

From the symmetry of the Normal distribution, we know that:

$$\Phi(x)=1-\Phi(-x):=\bar{\Phi}(-x) \tag{15.3.1437}$$

Hence we can write:

$$\mathrm{DP}=\bar{\Phi}(-\mathrm{DtD})=\bar{\Phi}\left(\frac{\ln\left(\dfrac{\mathcal{A}_0}{\mathrm{DPT}_T}\right)+\left(r-\dfrac{1}{2}\sigma_{\mathcal{A}}^2\right)T}{\sigma_{\mathcal{A}}\sqrt{T}}\right) \tag{15.3.1438}$$

If we put $T=1$ we just get:

$$\mathrm{DP}=\bar{\Phi}(-\mathrm{DtD})=\bar{\Phi}\left(\frac{\ln\left(\dfrac{\mathcal{A}_0}{\mathrm{DPT}}\right)+\left(r-\dfrac{1}{2}\sigma_{\mathcal{A}}^2\right)}{\sigma_{\mathcal{A}}}\right) \tag{15.3.1439}$$

That is nothing but the probability that a given firm will default within 1 year according to the KMV methodology. So the 1 year DP is named the "Expected Default Frequency" and abbreviated EDF$^{\text{TM}}$.

Remark

You don't dream... EDF$^{\text{TM}}$ is indeed a registered trademark! But in reality, the real KMV model is supposed to have the $\bar{\Phi}$ that is replaced with some empirical unknown function (as for most industry models, some specific parameters and implementations of the model are not completely known, and we can simply rely on what the company allows us to discover).

Now let us see a very suspicious development that is supposed to justify the definition of the DtD that you can find in some textbooks or presentations... Focus on:

$$\frac{\ln\left(\dfrac{\mathcal{A}_0}{\mathrm{DPT}_T}\right)+\left(r-\dfrac{1}{2}\sigma_{\mathcal{A}}^2\right)T}{\sigma_{\mathcal{A}}\sqrt{T}} \tag{15.3.1440}$$

Some empirical research seems to have shown that the quantity $r-\frac{1}{2}\sigma_{\mathcal{A}}^2$ is usually very close to zero. Given this, we have:

$$\frac{\ln\left(\dfrac{\mathcal{A}_0}{\mathrm{DPT}_T}\right)+\left(r-\dfrac{1}{2}\sigma_{\mathcal{A}}^2\right)T}{\sigma_{\mathcal{A}}\sqrt{T}}\cong\frac{\ln\left(\dfrac{\mathcal{A}_0}{\mathrm{DPT}_T}\right)}{\sigma_{\mathcal{A}}\sqrt{T}} \tag{15.3.1441}$$

If we take $T=1$, this reduces to:

$$\frac{\ln\left(\dfrac{\mathcal{A}_0}{\text{DPT}}\right)}{\sigma_{\mathcal{A}}} \tag{15.3.1442}$$

So far... everything seems acceptable. But now comes the crazy part... Indeed, let us assume that $\mathcal{A}_0 \cong \text{DPT}$ (hummmm....), using the first order Taylor series of $\ln(x)$ at $x\cong 1$ we then have (very rough approximation!!!!):

$$\frac{\ln\left(\dfrac{\mathcal{A}_0}{\text{DPT}}\right)}{\sigma_{\mathcal{A}}} \cong \frac{\dfrac{\mathcal{A}_0}{\text{DPT}}-1}{\sigma_{\mathcal{A}}} = \frac{\mathcal{A}_0-\text{DPT}}{\text{DPT}\sigma_{\mathcal{A}}} \tag{15.3.1443}$$

And now comes the even more crazy stuff... (only some very rare documents on the KMV do that...), as $\mathcal{A}_0 \cong \text{DPT}$, why not just write:

$$\frac{\mathcal{A}_0-\text{DPT}}{\text{DPT}\sigma_{\mathcal{A}}} \cong \frac{\mathcal{A}_0-\text{DPT}}{\mathcal{A}_0\sigma_{\mathcal{A}}} \tag{15.3.1444}$$

And after all this successive rough approximation and suspicious substitution of variables, we recognize here the DtD...

15.3.13 Time Series Analysis

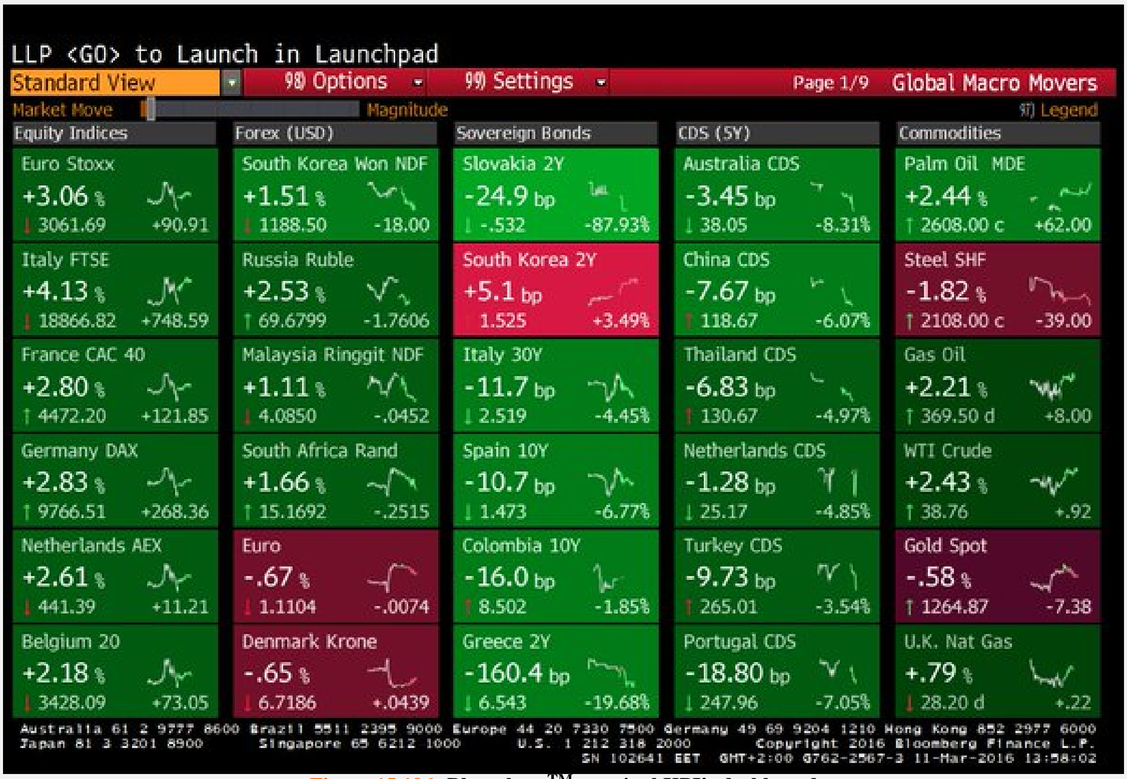
The main goal of "time series analysis" (TSA) is not to link variables together, but to focus on the dynamics of a unique variable in time to discover some regularities in order to extrapolate or to forecast under the assumption that we can connect an observation with those that preceded it[34].

Therefore the difference between times series and econometric models differs from the fact that times series models do not start with a conceptual framework that defines the relationship between the economic variable and economic theory. Time series models use a variable's own history values to estimate the future values of the same variable as a hypothetical description process.

With a fine analysis, it is even possible to make "robust" forecasting vis-à-vis sudden breaks and non-anticipated changes.

Remark

This study domain is often named "Business Intelligence" in companies (this is a sub-family in fact limited to descriptive elementary statistics that should be named "Business Facts" instead... as we have described during our study of Data Mining at page 6768). A variable analysed using TSA tools will be named a "Key Performance Indicator" (KPI) and a set of KPI a "dashboard".



Equity Indices	Forex (USD)	Sovereign Bonds	CDS (5Y)	Commodities
Euro Stoxx +3.06 % ↓ 3061.69 +90.91	South Korea Won NDF +1.51 % ↓ 1188.50 -18.00	Slovakia 2Y -24.9 bp ↓ -.532 -87.93%	Australia CDS -3.45 bp ↓ 38.05 -8.31%	Palm Oil MDE +2.44 % ↑ 2608.00 c +62.00
Italy FTSE +4.13 % ↓ 18866.82 +748.59	Russia Ruble +2.53 % ↑ 69.6799 -1.7606	South Korea 2Y +5.1 bp 1.525 +3.49%	China CDS -7.67 bp ↑ 118.67 -6.07%	Steel SHF -1.82 % ↑ 2108.00 c -39.00
France CAC 40 +2.80 % ↑ 4472.20 +121.85	Malaysia Ringgit NDF +1.11 % ↓ 4.0850 -.0452	Italy 30Y -11.7 bp ↓ 2.519 -4.45%	Thailand CDS -6.83 bp ↑ 130.67 -4.97%	Gas Oil +2.21 % ↑ 369.50 d +8.00
Germany DAX +2.83 % ↑ 9766.51 +268.36	South Africa Rand +1.66 % ↑ 15.1692 -.2515	Spain 10Y -10.7 bp ↓ 1.473 -6.77%	Netherlands CDS -1.28 bp ↓ 25.17 -4.85%	WTI Crude +2.43 % ↑ 38.76 +.92
Netherlands AEX +2.61 % ↓ 441.39 +11.21	Euro -.67 % ↓ 1.1104 -.0074	Colombia 10Y -16.0 bp ↑ 8.502 -1.85%	Turkey CDS -9.73 bp ↑ 265.01 -3.54%	Gold Spot -.58 % ↑ 1264.87 -7.38
Belgium 20 +2.18 % ↓ 3428.09 +73.05	Denmark Krone -.65 % ↓ 6.7186 +.0439	Greece 2Y -160.4 bp ↓ 6.543 -19.68%	Portugal CDS -18.80 bp ↓ 247.96 -7.05%	U.K. Nat Gas +.79 % ↓ 28.20 d +.22

```
Australia 61 2 9777 8600 Brazil 5511 2395 9000 Europe 44 20 7330 7500 Germany 49 69 9204 1210 Hong Kong 852 2977 6000
Japan 81 3 3201 8900          Singapore 65 6212 1000          U.S. 1 212 318 2000          Copyright 2016 Bloomberg Finance L.P.
                                                        SN 102641 EET  GMT+2:00 G762-2567-3 11-Mar-2016 13:58:02
```

Figure 15.196: **Bloomberg$^{\text{TM}}$ terminal KPI's dashboard**

---

[34] For anyone wishing to deepen and having a more rigorous presentation of time series analysis than the present book we strongly recommend the lecture of the 800 pages master piece *Time Series Analysis* of J.D. Hamilton ([428])!

But there are some other typical tasks in Time Series Analysis apart from forecasting (some of the techniques below were already treated in details in some other chapters of this textbook):

- Cross-correlation/Clustering of time series
- Auto-correlation
- Similarity (pattern matching)
- Anomaly detection/removal
- Gap filling
- Detrending
- Smoothing/Denoising
- Decomposition
- Normalization
- Spectrum Analysis
- ...

Definition (#15.3.700): A "discrete time series" or "discrete chronological series" is a series of observations Y_t of a variable y at different dates t. Usually the basis space t is countable, so that $t = 1 \ldots T$. The whole being denoted:

$$\{Y_T\}_{t=1}^{T} \tag{15.3.1445}$$

and sometimes the capital Y is denoted by X (anyway it doesn't really matter...).

Remark

We differentiate significantly models that predict real values (values in $\mathbb{R}$) of those who can only predict purely integer values (count data belonging to $\mathbb{N}$). To this date, the content of this textbook deals only with models especially suitable for continuous TSA variables.

A time series is thus every sequence of observations corresponding to the same variable: it can comes to macroeconomic data (GDP of a country, inflation, exports), microeconomics (sales of a given company, number of employees, the income of an individual, the production quantity...), financial (the CAC 40, S&P 500, the price of a Call option or Put option, the course of stock), weather (rainfall, number of sunny days per year,...), political (the number of voters, of votes received by a candidate ...), demographic (the average size of population, age, the propagation of an epidemy...).

In practice, all that is quantifiable and vary over time can be analysed relatively efficiently by TSA since the person handling the model should knows what she does (which is quite rare in reality...). But the reader must keep in mind that Economics and Finance ask many questions to which the answers are more difficult and complex than those often pondered in mathematics and physics. For example, economic agents can react to economic forecasts, which makes predicting the stock market even harder than predicting the weather. Imagine how much more difficult it would be for meteorologists to forecast if the weather could read its own forecast and then change its behaviour because it reads the weather forecast!

Therefore in finance we use these models to forecast returns, volatility (standard deviation), resupply stocks, demand, VaR and so on. In short, there really are no limits to the application of TSA.

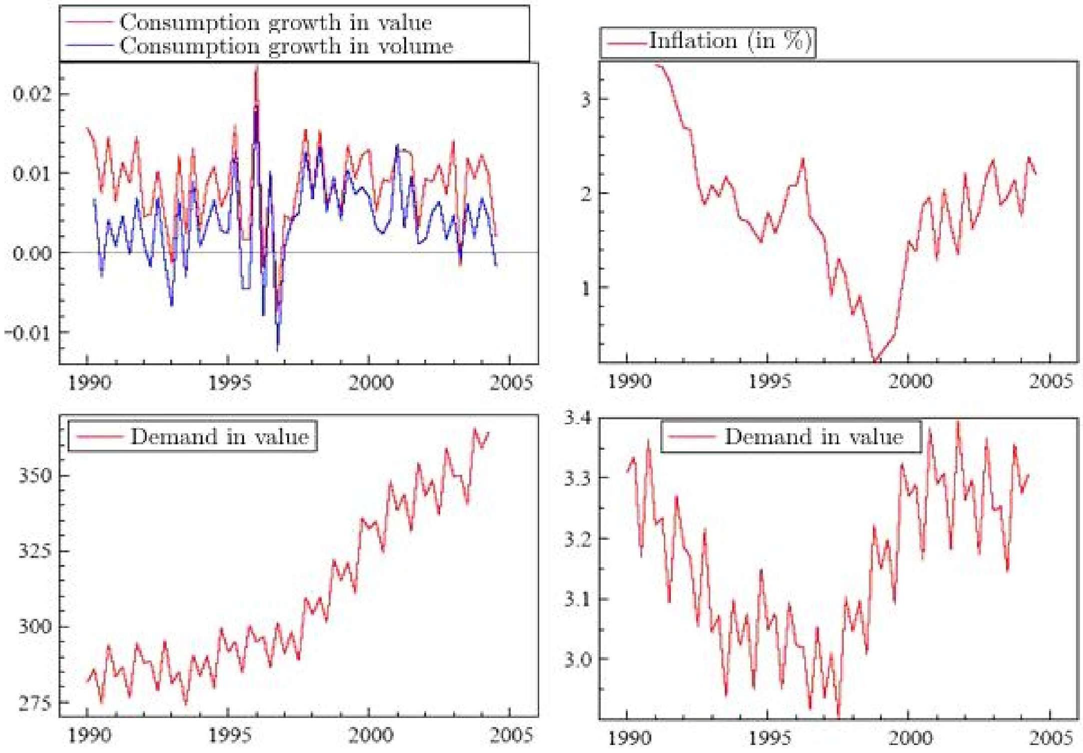


Figure 15.197: Some time series without forecasting

The time dimension is important here because it is the analysis of a historical chronicle: the variations of the same variable over time in the purpose to understand its dynamics. We generally represent 2D time series on value charts (ordinate) versus time (abscissa). Such observation is an essential tool that allows the modeller with little experience to immediately realize the main dynamic properties to see what statistical test practice (at least in the univariate and bivariate case...). The previous figure shows different time series as examples.

A time series is divided mainly into four basic components:

- The "secular long-term trend" or simply "trend". It corresponds to a non-seasonal cyclical movement that reflects the long-term evolution of the measured variable (can be linear or not).
- The "seasonal variations" that are periodic fluctuations that occur regularly (periodically: days, months, quarters, years or more...).
- The "hazards" or "residual/incidental variations" or "irregularities" are random type fluctuations, generally of low amplitude.

Definitions (#15.3.701):

D1. The series which oscillate around their average are named "stationary series" (we'll see a rigorous definition of this further below).

D2. The series that seem to grow (bullish) or fall (bearish) on all of the observed sample are named "trend series" or "short history" and their averages are not constant.

D3. The series that are neither stationary nor bullish or bearish on the long term trend are named "non-stationary series" or "random series" or "stochastic series". Any process whose time evolution can be analysed in terms of probability is named as we know "stochastic process".

D4. The series which have a regular frequency fluctuation are named "seasonal series" Regularly, stationary series are trend series with a periodic component.

D5. The series that seem to be seasonal or appear to have a seasonal component but whose frequency is variable, are named "cyclic series".

D6. A times series that has a significant jump (variation) at one moment in time is say to have a "level shift".

D7. A times series that has a significant jump (variation) at various moment in time and fall back to zero after every jump is "intermittent series".

Given y_t a value of the time series, t_t its trend component, s_t its seasonal component and ε_t its random part, we define:

Definitions (#15.3.702):

D1. An "additive model by an expression like:

$$\boxed{Y_t = t_t + s_t + \varepsilon_t} \tag{15.3.1446}$$

where $\{s_t\}$ is periodic with period T that is $s_{t-T} = s_t \forall t$.

D2. A "simple multiplicative model" by an expression like:

$$\boxed{Y_t = t_t(1 + s_t) + \varepsilon_t} \tag{15.3.1447}$$

where we have the seasonal component that is then related to the trend component (as opposed to the additive model where the seasonal component is the same from multiple periods to another).

D3. A "complete multiplicative model" by an expression like:

$$\boxed{Y_t = t_t(1 + s_t)(1 + \varepsilon_t)} \tag{15.3.1448}$$

D4. A "mixed model" by an expression like:

$$\boxed{Y_t = t_t(1 + s_{1,t}) + s_{2,t} + \varepsilon_t} \tag{15.3.1449}$$

The characteristics of these graphs are all modelled and analysed in the context of time series analysis. For this there exists are more or less complex tools, some of which can not be doubted and others of which are heuristics and need to be handle and use with caution.

In the text that follows, we will look to the elementary models available without a heavy mathematical artillery.

Humility in the forecaster's job is however set to accept the mistake repeatedly, while seeking to improve the quality of its forecasts (sales, market volatility, correlation of assets, etc.) and identification of non-random components phenomena. It is important to remember the following four criteria:

1. The forecasts are generally not accurate. We need a robust ecosystem to react quickly to the unexpected events (that we can't predict! that doesn't mean that we can not take advantage of the unpredictable) and this especially true when we know that some multinationals and governments (which I will not give the name ...) request projections over twenty months or sometimes fifteen years...!!!).
2. A good forecast is more than a numerical value!!Because forecasts are generally not accurate, a good forecast should ALWAYS include a measurement (or plot) of expected error for prediction (confidence interval). It is obvious that most of time the confidence interval (prediction errors) increases as we move further into the future!!!!
3. Aggregate forecasts are in general more accurate (yes we know this from the properties of the variance of an average). In general, we notice that the error made to the sales forecast of a line of products is generally less than the error made in the sales forecast of a single element.
4. Long-term forecasts are less accurate. There is not much to say because it's intuitive...
5. Forecasts can be sensitive to initial conditions (think of the chaotic patterns in meteorology or population dynamics...).

Remark

Examples of where large multinational executives or governments have, for lack of statistical knowledge or misguided by non-statisticians managers / non-economists made projections of earnings or sales over ten to twenty years (this is an aberration!!), without confidence interval are legion! Obviously the values communicated and the associated development strategies led some of these multinationals and governments to go bankrupt.

Let us notice that quantitative forecasting techniques also use mainly the following tools (it is of course possible to combine them) for which we have already given examples in the respective mentioned chapters in parenthesis:

- Simple random walk (see section Thermodynamics page 3582)
- Simple linear regression (see section Numerical Methods page 6322)
- Multiple linear regression (see section Numerical Methods page 6379)
- General linear models (see section Numerical Methods page 6530)
- Logistic regression (see section Numerical Methods page 6449)
- Polynomial interpolation (see section Numerical Methods page 6594)

- Logarithmic and exponential linearisation (see section Numerical Methods page 6323)
- Modelling by Brownian motion (see section Mechanics page 7765)
- Moving Average (see section Statistics page 7919)
- Correlation Analysis (see section Statistics page 613)
- Control Charts (see section Industrial Engineering page 8934)
- Jackknife or Bootstrapping regression (see section Numerical Methods page 6747 page 6740)
- Neural Networks[35] (see section Numerical Methods page 6951)
- Fourier Transform (see section Sequences and Series page 1758)

and they can be grouped into five main families (not mutually exclusive!):

1. Causes & effects models based on several variables
2. Time-series models whose only explanatory variable is time
3. Permutations simulations
4. Machines Learning modelization
5. Spectral Analysis

and it is always difficult for forecasters to choose which is the best suited model to his working framework and the level of understanding of his superiors and colleagues. The interesting part of the job is that once the model(s) selected (quite hard to choose from the hundreds of published empirical models...), we can compare them with past statistical data of the company and see if predictions are in agreement with what had really happened (i.e. run some back-testing).

---

[35] Especially recurrent neural networks (RNN) - that their are vanilla RNN (i.e. short term) or long short-term memory (LSTM) - or autoregressive neural networks (ANN).

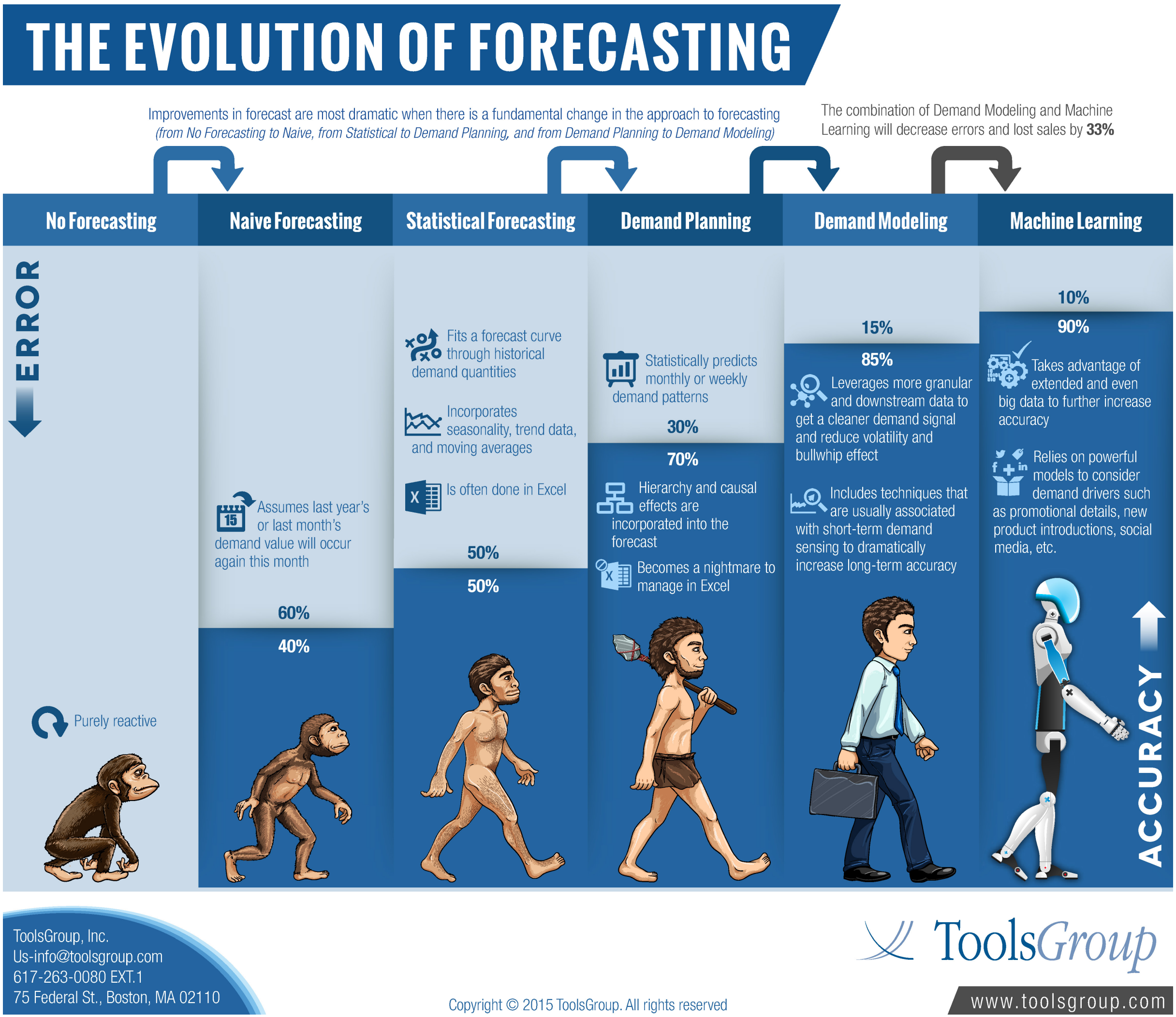

THE EVOLUTION OF FORECASTING
Improvements in forecast are most dramatic when there is a fundamental change in the approach to forecasting
(from No Forecasting to Naive, from Statistical to Demand Planning, and from Demand Planning to Demand Modeling)
The combination of Demand Modeling and Machine Learning will decrease errors and lost sales by 33%
No Forecasting
Naive Forecasting
Statistical Forecasting
Demand Planning
Demand Modeling
Machine Learning
ERROR
Purely reactive
Assumes last year's or last month's demand value will occur again this month
15
60%
40%
Fits a forecast curve through historical demand quantities
Incorporates seasonality, trend data, and moving averages
Is often done in Excel
50%
50%
Statistically predicts monthly or weekly demand patterns
30%
70%
Hierarchy and causal effects are incorporated into the forecast
Becomes a nightmare to manage in Excel
15%
85%
Leverages more granular and downstream data to get a cleaner demand signal and reduce volatility and bullwhip effect
Includes techniques that are usually associated with short-term demand sensing to dramatically increase long-term accuracy
10%
90%
Takes advantage of extended and even big data to further increase accuracy
Relies on powerful models to consider demand drivers such as promotional details, new product introductions, social media, etc.
ACCURACY
ToolsGroup, Inc.
Us-info@toolsgroup.com
617-263-0080 EXT.1
75 Federal St., Boston, MA 02110
Copyright © 2015 ToolsGroup. All rights reserved
ToolsGroup
www.toolsgroup.com

Figure 15.198: Forecasting evolution (source and copyright: Toolsgroup.com)

Remark

It is preferable to use statistical based models based rather than deterministic models that are of poor quality (as the simple moving average or double exponential smoothing).

15.3.13.1 Type of Errors

Before building predictive models we must be able to estimate the quality thereof (for comparison purposes between models) and it is customary to use several empirical indicators.

Measuring forecast accuracy (or error) is not an easy task as there at the time we write these line no one-size-fits-all indicator. Only experimentation will show us what Key Performance Indicator (KPI) is best for us. As we will see, each indicator will avoid some pitfalls but will be prone to others.

- The "forecast error":

  $$e_t = Y_t - \hat{Y}_t \qquad (15.3.1450)$$

  which is the difference between the relation value Y_t and the value given by the model $\hat{Y}_t$.

- The "mean error" or "forecasting bias":

  $$\text{ME} = \frac{1}{n}\sum_{t=1}^{n} e_t \qquad (15.3.1451)$$

  that will only tell us if there is a positive or negative bias between the forecast and the real values (because if the value ME is not null then the model is biased). As a positive error on one item can offset a negative error on another item, a forecast model can achieve very low bias and not be precise at the same time. Obviously, the bias alone won't be enough to evaluate your forecast precision. But a highly biased forecast is already an indication that something is wrong in the model.

  As the mean error is often very small (negative values balancing the positive in some models) we prefer the next following measurement below.

- The "Average Absolute Deviation" (named also "Mean Absolute Deviation" or "Mean Absolute Error" or "Mean Forecast Error"):

  $$\text{MAD} = \text{MAE} = \text{MFE} = \frac{1}{n}\sum_{t=1}^{n} |e_t| \qquad (15.3.1452)$$

  which is the mean of the absolute errors made by the forecasting model on the duration of it, notwithstanding the fact that the error is overestimate or underestimate.

- The "Average Squared Residuals" (also named "Mean Square Deviation" or "Mean Square Errors"):

  $$\text{MSD} = \text{MSE} = \frac{1}{n}\sum_{t=1}^{n} e_t^2 \qquad (15.3.1453)$$

  allowing to penalize the large errors more than the small errors since each is multiplied by itself, giving a greater weight to large errors than to small errors.

  The problem with the previous indicators is that the magnitude of the error depends on the amplitudes of the analysed values. To avoid this, it is more interesting to work with a percentage (relative value). Thus, we define the "relative error" or "percentage error" indicated in percent by:

$$\text{PE}_t = \frac{Y_t - \hat{Y}_t}{Y_t} = \frac{e_t}{Y_t} \qquad (15.3.1454)$$

which is the difference in percent between the real values and modelling one. It follows then the "Mean Relative Error", also named "Relative Bias of Prediction" or "Mean Percentage Error", always indicated in percents:

$$\text{MPE} = \frac{1}{n}\sum_{t=1}^{n} \text{PE}_t \qquad (15.3.1455)$$

For the same reasons that for the ME we prefer the following indicator error:

- The "Relative Absolute Average Deviation" also named "Mean Absolute Percentage Error" always indicated in percents:

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n} |\text{PE}_t| \qquad (15.3.1456)$$

The only problem with the relative indicators is when the analysed time series contains zeros...

Here is a guide for decision making relatively to this indicator:

MAPE	Forecasting power
$<10\%$	Highly accurate forecasting
$10\% \sim 20\%$	Good forecasting
$20\% \sim 50\%$	Reasonable forecasting
$>50\%$	Weak and inaccurate forecasting

Table 15.59: **Various MAPE value decision guide (source: Lewis 11982)**

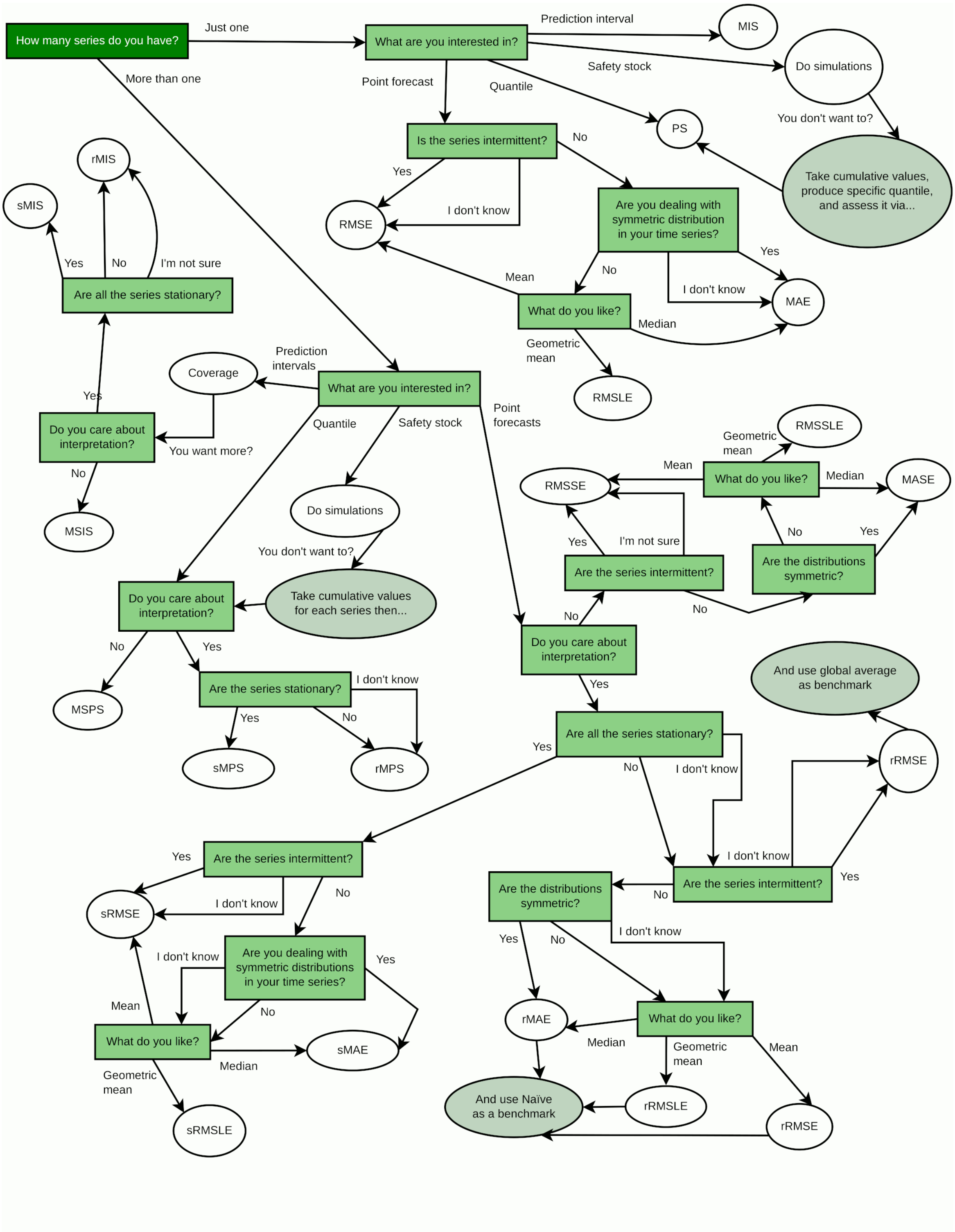
- The "time momentum" are the basic position and dispersion parameters that we know in statistics that are the mean and standard deviation but just taken over a time period (widely used in finance!). Thus, the average and standard deviation of a measured variable of a time-window $[t-p,t]$ is obviously given by:

$$\begin{aligned}\mu(t,p) &= \frac{1}{p+1}\sum_{\theta=0}^{p} x_{t-\theta}\\ \sigma(t,p) &= \sqrt{\frac{1}{p+1}\sum_{\theta=0}^{p}(x_{t-\theta}-\mu(t,p))^2}\end{aligned} \qquad (15.3.1457)$$

- The "hit score" (widely used in finance markets), the idea is very simple: it is simply the ratio of the number of good periodicals forecasts (the concept of "good" being quite empirical because there are dozens of choices) on the total number of past periods.

Finally, to end this section on error indicators, notice that in the field of forecasting we must also checked on the long term if a theoretical model is viable or not to replace it with another one. The basic techniques for this are to make use for example of an Autocorrelated controls chart (see section Industrial Engineering page 9007) where the measurement are simply the deviations (without taking their absolute value!) between the projected values and what really happened.

Caution! There are some commonly used accuracy measures whose scale depends on the scale of the data (like MAE, MSE, WAPE, MAE%). These are useful when comparing different methods applied to the same set of data, but should not be used, for example, when comparing across data sets that have different scales.


How many series do you have?
Just one
What are you interested in?
Prediction interval
MIS
Safety stock
Do simulations
Point forecast
Quantile
PS
You don't want to?
Take cumulative values, produce specific quantile, and assess it via...
Is the series intermittent?
No
Yes
I don't know
RMSE
Are you dealing with symmetric distribution in your time series?
Yes
No
I don't know
MAE
Mean
What do you like?
Median
Geometric mean
RMSLE
More than one
rMIS
sMIS
Yes
No
I'm not sure
Are all the series stationary?
Prediction intervals
Coverage
What are you interested in?
Yes
Do you care about interpretation?
You want more?
No
MSIS
Quantile
Safety stock
Point forecasts
Do simulations
You don't want to?
Take cumulative values for each series then...
Do you care about interpretation?
No
Yes
MSPS
Are the series stationary?
I don't know
Yes
No
sMPS
rMPS
RMSSLE
Geometric mean
Mean
RMSSE
What do you like?
Median
MASE
Yes
I'm not sure
No
Yes
Are the series intermittent?
Are the distributions symmetric?
No
No
Do you care about interpretation?
Yes
And use global average as benchmark
Are all the series stationary?
Yes
No
I don't know
rRMSE
Yes
Are the series intermittent?
sRMSE
I don't know
No
I don't know
Are you dealing with symmetric distributions in your time series?
Yes
Mean
No
What do you like?
sMAE
Median
Geometric mean
sRMSLE
I don't know
Are the series intermittent?
Yes
Are the distributions symmetric?
No
I don't know
Yes
No
rMAE
What do you like?
Median
Geometric mean
Mean
And use Naïve as a benchmark
rRMSLE
rRMSE

- MAE is Mean Absolute Error,
- RMSE is Root Mean Squared Error
- RMSLE is Root Mean Squared Logarithmic Error
- MIS is Mean Interval Score
- PS is Pinball Score

- small "s" means "scaled" - divided by in-sample mean
- small "r" stands for "relative" - value divided by similar from a benchmark model
- RMSSE, MASE, RMSSLE, MSPS, MSIS are scaled errors, with in-sample first differences in the denominator

Developed by Ivan Svetunkov, CMAF, Lancaster University. https://forecasting.svetunkov.ru/
See more about error measures at https://openforecast.org/adam/forecastsEvaluation.html

Figure 15.199: **Forecasting error measures decision flow chart (author: Ivan Svetunkow)**

Remark

Let us notice also the usage in practice (especially banks) of the "drawdown" that is the measure of the decline from a historical peak in some variable (typically the cumulative profit or total open equity of a financial trading strategy).

Somewhat more formally, if $X=(X(t),t\geq 0)$ is a random process with $X(0)=0$, the drawdown at time T, denoted $D(T)$, is defined as:

$$\mathrm{D}(T)=\max\left\{0,\max_{t\in(0,T)}(X(t)-X(T))\right\} \tag{15.3.1458}$$

The "maximum drawdown" (MDD) up to time T is the maximum of the Drawdown over the history of the variable. More formally:

$$\mathrm{MDD}(T)=\max_{\tau\in(0,T)}\left[\max_{t\in(0,\tau)}X(t)-X(\tau)\right] \tag{15.3.1459}$$

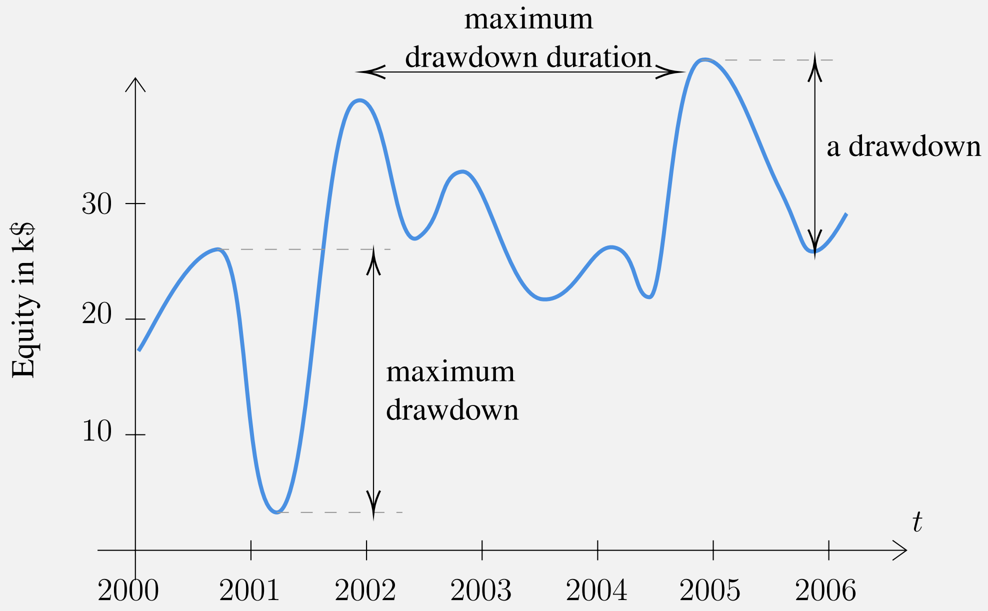


Figure 15.200: **Maximum Drawdown**

15.3.13.2 Decompositions

Some models using time series are adequate forecasting tools provided that the variations have shown some a stable pattern over time and that the conditions in which the pattern appeared always apply and that the past measurements don't have a to big intrinsic error.

Sometimes a pattern is not apparent in the analysis of raw data; these can however be decomposed so as to reveal patterns that facilitate the projection data in the future (we will see the spectral decomposition by Fourier transform further below). The 4 generally recognized components (already mentioned above) for time series out spectral analysis are: trend t_t (linear, exponential, sigmoid, etc.), seasonality s_t, cyclicity c_t, random (irregular) i_t.

Hence a time series using an additive model can be thought of as:

$$\boxed{Y_t = t_t + c_t + s_t + i_t} \tag{15.3.1460}$$

While a multiplicative model would be:

$$\boxed{Y_t = t_t \cdot c_t \cdot s_t \cdot i_t} \tag{15.3.1461}$$

An additive model would be used when the variations around the trend does not vary with the level of the time series where as a multiplicative model would be appropriate if the trend is proportional to the level of the time series.

Let us see this with a small companion practical example using (for simplicity) the Microsoft® Excel® 14.0.6123.

Let us create a random component based on the absolute value of a Normal distribution of mean of 10 and standard deviation of 20 and over a period of 32 months with the following formula in column `B` (`B1` to `B32`):

```
=ABS(NORM.INV(RAND(),10,20))
```

where column `A` will contain the numbering of the periods ranging from 1 to 32 (`A1` to `A32`).

Which will give a simulation of a random graph for which we will fix the values. For example:

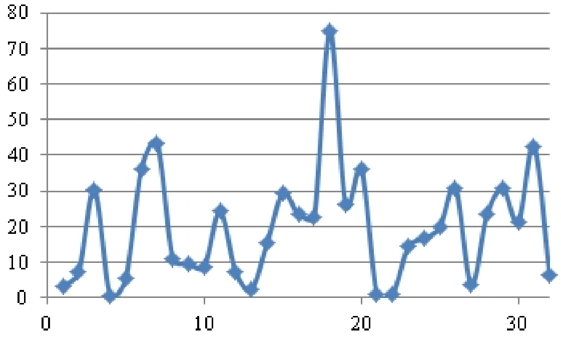


Figure 15.201: **Plot of an example of the random function with Microsoft® Excel® 14.0.6123**

Then we add a linear trend component (drift) linear of slope 4 and of ordinate at the origin 20 to the random component so as to have this time in column `B` the following formula:

```
=ABS(NORM.INV(RAND(),10,20))+(4*A1+20)
```

The sum gives graphically:

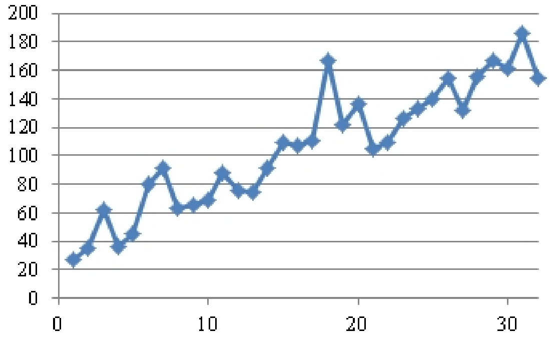


Figure 15.202: Plot of the sum of the random function and the trend with Microsoft® Excel® 14.0.6123

Now let add a seasonal component of the sinusoidal type with a range of 50 and 12 months and without phase:

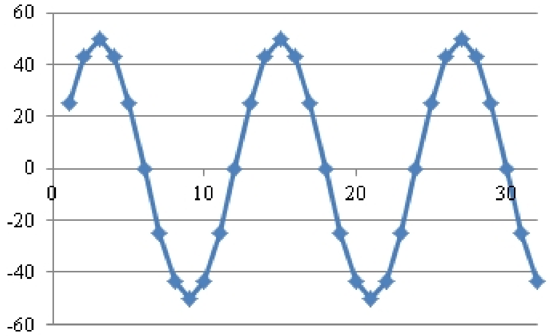


Figure 15.203: Plot of the seasonal part only with Microsoft® Excel® 14.0.6123

The sum of the three components will give in column `B`:

```
=ABS(NORM.INV(RAND(),10,20))+(4*A1+20)+50*SIN(A1*2*PI()*1/12)
```

and graphically it gives:

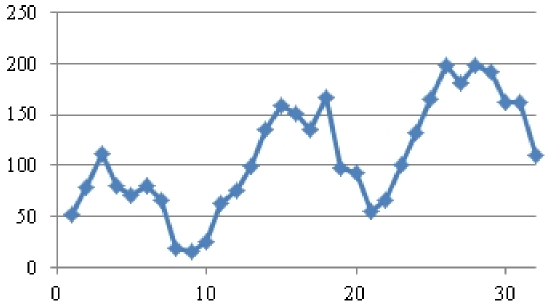


Figure 15.204: Plot of the sum of the random function, the trend and seasonality with Microsoft® Excel® 14.0.6123

and finally let us add a cyclical component which will again be for example a sinusoidal function of amplitude 20 and period 64:

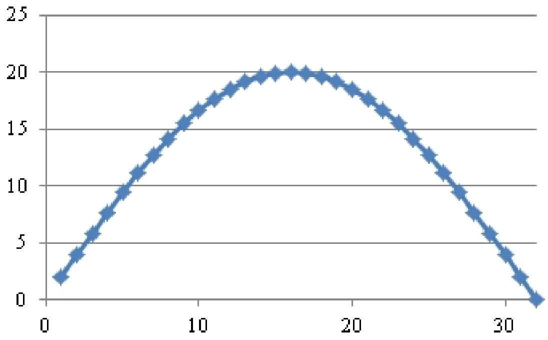


Figure 15.205: **Plot of the cyclic component with Microsoft® Excel® 14.0.6123**

The sum of the four components gives in column `B`:

```
=ABS(NORM.INV(RAND(),10,20))+(4*A1+20)+50*sin(A1*2*PI()*1/12)+20*SIN(A1*2*PI(
                                   )*1/64)
```

and graphically it gives:

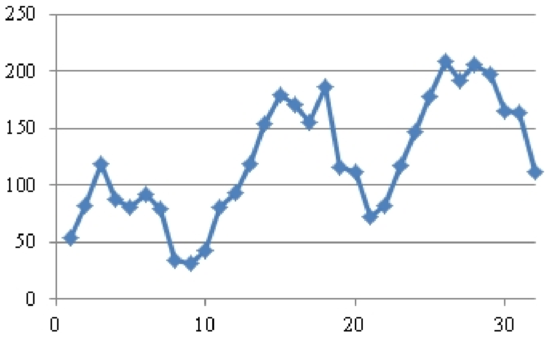
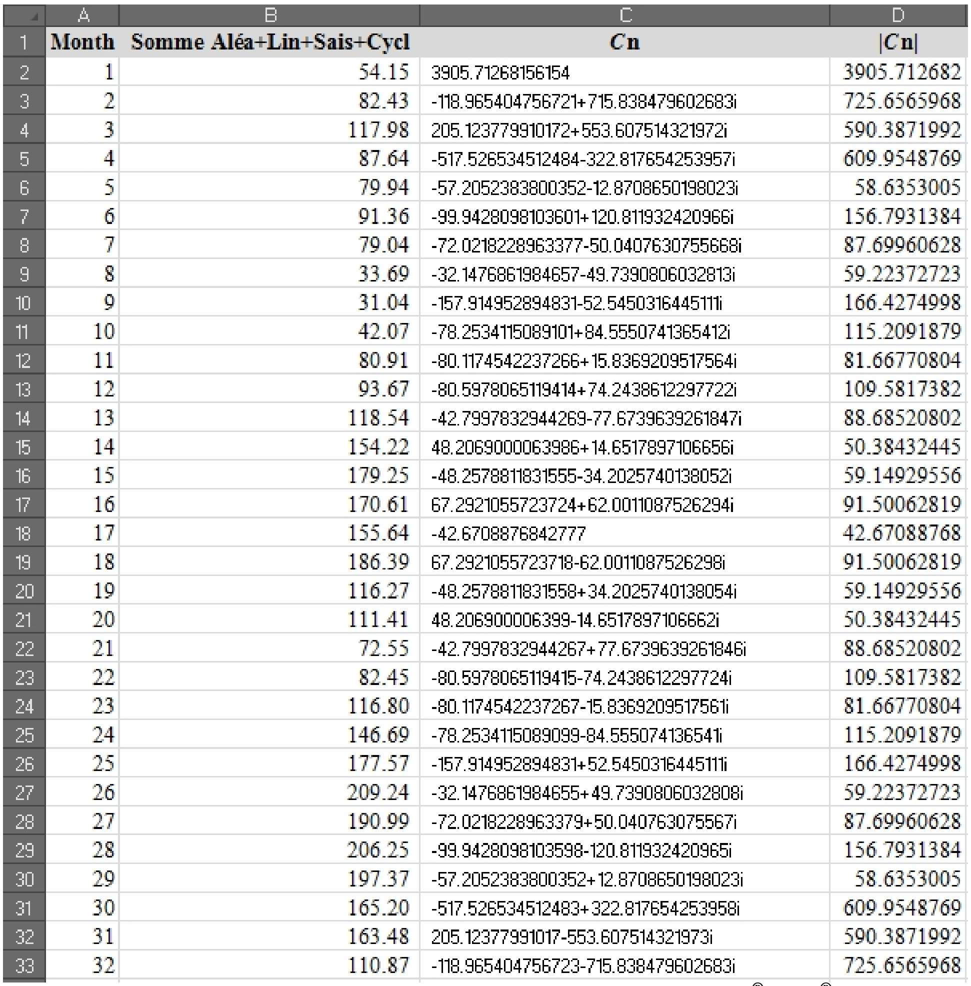


Figure 15.206: **Sum of all components with Microsoft® Excel® 14.0.6123**

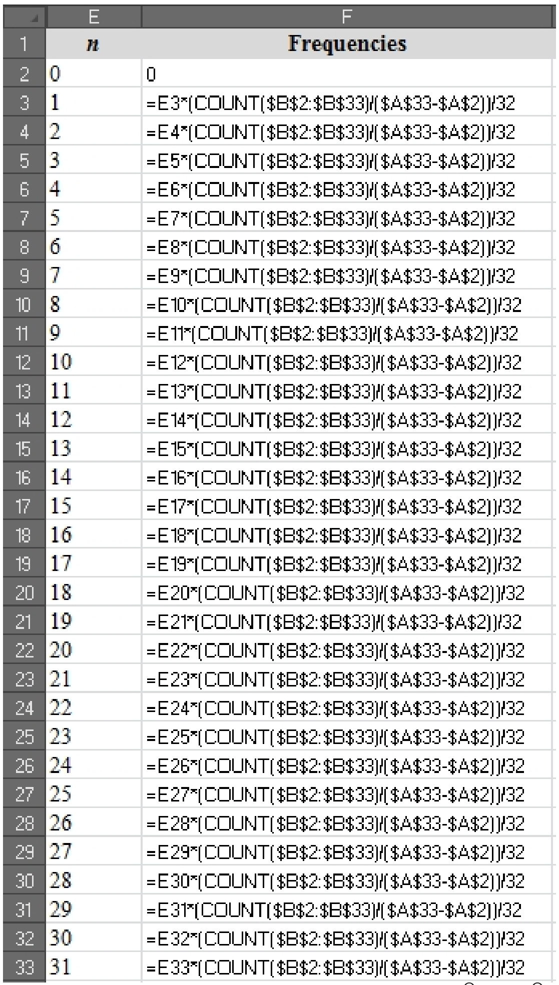
Now, as we have introduced in detail in the section of Sequences and Series, let us make a fast Fourier transform (spectral analysis) - abbreviated FFT - with the integrated Data Analysis Tools in Microsoft® Excel® 14.0.6123. This gives us then:



	A	B	C	D
1	**Month**	**Somme Aléa+Lin+Sais+Cycl**	***C*n**	**|*C*n|**
2	1	54.15	3905.71268156154	3905.712682
3	2	82.43	-118.965404756721+715.838479602683i	725.6565968
4	3	117.98	205.123779910172+553.607514321972i	590.3871992
5	4	87.64	-517.526534512484-322.817654253957i	609.9548769
6	5	79.94	-57.2052383800352-12.8708650198023i	58.6353005
7	6	91.36	-99.9428098103601+120.811932420966i	156.7931384
8	7	79.04	-72.0218228963377-50.0407630755668i	87.69960628
9	8	33.69	-32.1476861984657-49.7390806032813i	59.22372723
10	9	31.04	-157.914952894831-52.5450316445111i	166.4274998
11	10	42.07	-78.2534115089101+84.5550741365412i	115.2091879
12	11	80.91	-80.1174542237266+15.8369209517564i	81.66770804
13	12	93.67	-80.5978065119414+74.2438612297722i	109.5817382
14	13	118.54	-42.7997832944269-77.6739639261847i	88.68520802
15	14	154.22	48.2069000063986+14.6517897106656i	50.38432445
16	15	179.25	-48.2578811831555-34.2025740138052i	59.14929556
17	16	170.61	67.2921055723724+62.0011087526294i	91.50062819
18	17	155.64	-42.6708876842777	42.67088768
19	18	186.39	67.2921055723718-62.0011087526298i	91.50062819
20	19	116.27	-48.2578811831558+34.2025740138054i	59.14929556
21	20	111.41	48.206900006399-14.6517897106662i	50.38432445
22	21	72.55	-42.7997832944267+77.6739639261846i	88.68520802
23	22	82.45	-80.5978065119415-74.2438612297724i	109.5817382
24	23	116.80	-80.1174542237267-15.8369209517561i	81.66770804
25	24	146.69	-78.2534115089099-84.555074136541i	115.2091879
26	25	177.57	-157.914952894831+52.5450316445111i	166.4274998
27	26	209.24	-32.1476861984655+49.7390806032808i	59.22372723
28	27	190.99	-72.0218228963379+50.040763075567i	87.69960628
29	28	206.25	-99.9428098103598-120.811932420965i	156.7931384
30	29	197.37	-57.2052383800352+12.8708650198023i	58.6353005
31	30	165.20	-517.526534512483+322.817654253958i	609.9548769
32	31	163.48	205.12377991017-553.607514321973i	590.3871992
33	32	110.87	-118.965404756723-715.838479602683i	725.6565968

Figure 15.207: **Table of the FFT with the Data Analysis Toolpack of Microsoft® Excel® 14.0.6123**

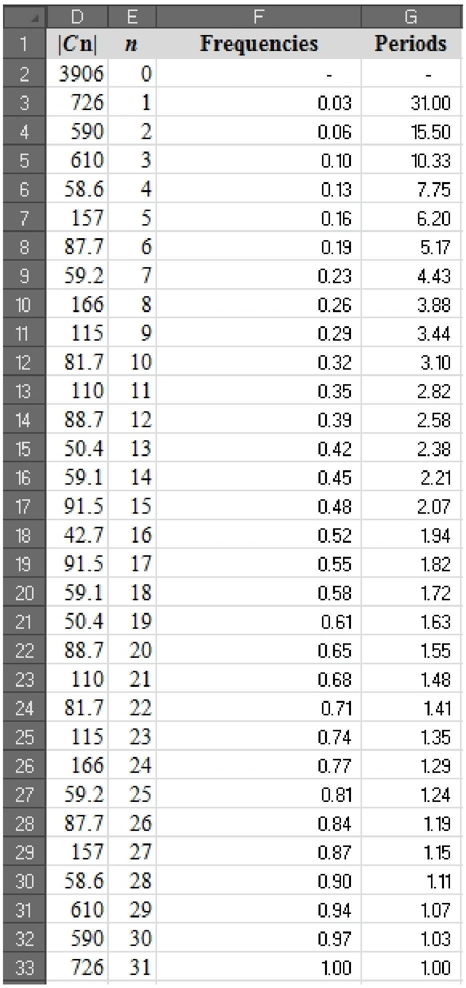
But unlike the section of Sequences and Series we will this time plot the frequency spectrum. For this, we first create a column with numbers from 0 to 31 and just next we put the formula visible in the screenshot below:



	E	F
1	*n*	**Frequencies**
2	0	0
3	1	=E3*(COUNT($B$2:$B$33)/($A$33-$A$2))/32
4	2	=E4*(COUNT($B$2:$B$33)/($A$33-$A$2))/32
5	3	=E5*(COUNT($B$2:$B$33)/($A$33-$A$2))/32
6	4	=E6*(COUNT($B$2:$B$33)/($A$33-$A$2))/32
7	5	=E7*(COUNT($B$2:$B$33)/($A$33-$A$2))/32
8	6	=E8*(COUNT($B$2:$B$33)/($A$33-$A$2))/32
9	7	=E9*(COUNT($B$2:$B$33)/($A$33-$A$2))/32
10	8	=E10*(COUNT($B$2:$B$33)/($A$33-$A$2))/32
11	9	=E11*(COUNT($B$2:$B$33)/($A$33-$A$2))/32
12	10	=E12*(COUNT($B$2:$B$33)/($A$33-$A$2))/32
13	11	=E13*(COUNT($B$2:$B$33)/($A$33-$A$2))/32
14	12	=E14*(COUNT($B$2:$B$33)/($A$33-$A$2))/32
15	13	=E15*(COUNT($B$2:$B$33)/($A$33-$A$2))/32
16	14	=E16*(COUNT($B$2:$B$33)/($A$33-$A$2))/32
17	15	=E17*(COUNT($B$2:$B$33)/($A$33-$A$2))/32
18	16	=E18*(COUNT($B$2:$B$33)/($A$33-$A$2))/32
19	17	=E19*(COUNT($B$2:$B$33)/($A$33-$A$2))/32
20	18	=E20*(COUNT($B$2:$B$33)/($A$33-$A$2))/32
21	19	=E21*(COUNT($B$2:$B$33)/($A$33-$A$2))/32
22	20	=E22*(COUNT($B$2:$B$33)/($A$33-$A$2))/32
23	21	=E23*(COUNT($B$2:$B$33)/($A$33-$A$2))/32
24	22	=E24*(COUNT($B$2:$B$33)/($A$33-$A$2))/32
25	23	=E25*(COUNT($B$2:$B$33)/($A$33-$A$2))/32
26	24	=E26*(COUNT($B$2:$B$33)/($A$33-$A$2))/32
27	25	=E27*(COUNT($B$2:$B$33)/($A$33-$A$2))/32
28	26	=E28*(COUNT($B$2:$B$33)/($A$33-$A$2))/32
29	27	=E29*(COUNT($B$2:$B$33)/($A$33-$A$2))/32
30	28	=E30*(COUNT($B$2:$B$33)/($A$33-$A$2))/32
31	29	=E31*(COUNT($B$2:$B$33)/($A$33-$A$2))/32
32	30	=E32*(COUNT($B$2:$B$33)/($A$33-$A$2))/32
33	31	=E33*(COUNT($B$2:$B$33)/($A$33-$A$2))/32

Figure 15.208: **Formulas to get the frequencies in Microsoft® Excel® 14.0.6123**

which gives:



	D	E	F	G
1	$\|Cn\|$	n	Frequencies	Periods
2	3906	0	-	-
3	726	1	0.03	31.00
4	590	2	0.06	15.50
5	610	3	0.10	10.33
6	58.6	4	0.13	7.75
7	157	5	0.16	6.20
8	87.7	6	0.19	5.17
9	59.2	7	0.23	4.43
10	166	8	0.26	3.88
11	115	9	0.29	3.44
12	81.7	10	0.32	3.10
13	110	11	0.35	2.82
14	88.7	12	0.39	2.58
15	50.4	13	0.42	2.38
16	59.1	14	0.45	2.21
17	91.5	15	0.48	2.07
18	42.7	16	0.52	1.94
19	91.5	17	0.55	1.82
20	59.1	18	0.58	1.72
21	50.4	19	0.61	1.63
22	88.7	20	0.65	1.55
23	110	21	0.68	1.48
24	81.7	22	0.71	1.41
25	115	23	0.74	1.35
26	166	24	0.77	1.29
27	59.2	25	0.81	1.24
28	87.7	26	0.84	1.19
29	157	27	0.87	1.15
30	58.6	28	0.90	1.11
31	610	29	0.94	1.07
32	590	30	0.97	1.03
33	726	31	1.00	1.00

Figure 15.209: **Values of the frequencies and periods of the FFT in Microsoft® Excel® 14.0.6123**

Or graphically for the first 16 points (the other being identical by symmetry):

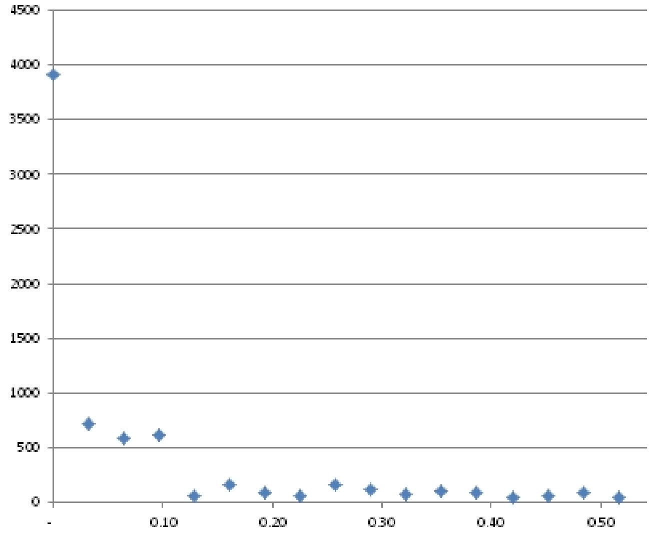


Figure 15.210: **Plot of the frequency peaks of the FFT with Microsoft® Excel® 14.0.6123**

So we see that the transformation highlights a constant component (zero frequency), which is very important. We can abusively assimilate it with the linear component.

We then see then that three points stand out with the respective periods (inverse of frequency): 31 months, 15.50, 10.33. But we will only consider the first two values to associate them to the cyclicity (having the greatest period by definition) and seasonality.

Thus, the Fourier transform leads us to consider a consistent trend, a cyclical phenomenon with a period of 31 months (while we simulated a cycle with a period of 64 months ...) and a seasonal phenomenon having a period 15 months (while we had a simulation with a period of 12 months).

So we see that this type of analysis has limitations but it can be useful as a tool for decision support.

15.3.13.3 Types of Forecasting Models

Definition (#15.3.703): "Forecasting" is the set of mathematical methods and techniques that helps organizations (corporations, governments) to anticipate future events by analysing past data to identify systematic relationships, patterns, and trends with a given probabilistic confidence. It plays many times vital role in decision-making.

Remark

A "forecast" must not be confused with the concept of "prophecy". For scientists the definition of prophecy is the following: *A "prophecy" is a claim that, without using the scientific method or any mathematical calculations, anticipates the details of occurrence of a given event (exact magnitude, exact number of people involved, exact names of things involved, etc.) with a high accurate time (at the second) and space precision (at the meter) without providing any probability confidence interval. A "prophet" is someone who can succeed multiple prophecies that beats the hazard at a high level enough statistical power.*

Time series analysis has two main approaches for forecasting:

1. The deterministic approach that doesn't use at all statistical tools
2. The statistical approach that allows to infer on forecasts[36]

It goes without saying that the first method is easier than the second, but the second includes the first and is scientifically more acceptable since it will be possible to define confidence intervals (it fact it should be the only acceptable one!).

These general concepts being presented we'll move on to the presentation of some regression mathematical models of time series. You should know about it that:

1. There are (to my knowledge) almost 60 different mathematical well known models
2. Some models are deterministic, other stochastic (the latter being the best in the scientific point of view), some causal (like regression obviously) or non-causal (depending only on previous values)

---

[36] Also named "predictive analytics"

3. They are (to my knowledge) all empirical and no one is perfect
4. A spreadsheet software like Microsoft® Excel® is enough to model the majority of them in school classes

> ❞ *The way to cope with the future is to create it!* ❞
>
> Ilya Prigogine

quote
15.3:7915.#88

Obviously there are a lot of forecasting models used by some (mainly) multinational companies. Most of them are presented further below in details but as requested by a reader here is a list of the most known one in increasing order of ascending complexity[37]:

1. Extrapolation (linear regression prediction)
2. Simple moving average
3. Linear model with seasonal coefficient (LMSC)
4. Simple exponential smoothing (EWMA)
5. Double exponential smoothing with one parameter (Brown Method)
6. Holt's double exponential smoothing with 2 parameters
7. Holt's and Winter triple exponential smoothing with 3 parameters[38]
8. Logistic model
9. Croston's method (and adjusted Croston's methods)
10. Kalman filtering
11. Autoregressive processes (AR) or Vector Autoregressive processes[39] (VAR)
12. Moving average autoregressive processes (MA)
13. Discrete autoregressive moving average (DARMA)
14. Autoregressive moving average (ARMA)
15. Integrated-value moving average (INARMA)
16. Autoregressive integrated moving average (ARIMA)
17. Autoregressive distributed lag (ARDL)
18. Autoregressive integrated moving average with explanatory variable (ARIMAX)
19. Fractional ARIMA (FARIMA)
20. Seasonal ARIMA (SARIMA) and seasonal fractional ARIMA (SARFIMA)
21. Generalized linear model (GLM)
22. Generalized autoregressive conditional heteroskedasticity (GARCH or

---

[37] As far as we know, the only software that implement them all is the R Software.

[38] Form information, all models based on exponential smoothing belongs to the family of "ETS models]" (acronym of "Error Trend Seasonality))

[39] There is a type of VAR where when at least two of the series are cointegrated (the residuals estimated on the long term are not stationary) that we name Vector Error Correction Model (VECM).

sGARCH), Rich univariate GARCH (RUGARCH), integrated GARCH (iGARCH), multivariate GARCH (MGARCH), Copula-GARCH

23. Non-linear ARCH (NARCH)
24. Threshold ARCH (TARCH)
25. Neural networks (FFNN, RNN, LSTM, GRU etc.)
26. Facebook Prophet's model
27. TBATS (**T**rigonometric seasonality, **B**ox-Cox transformation, **ARMA** errors, **T**rend and **S**easonal components)
28. Theta models
29. ...

We will see below the proof of those that require a mathematical approach and not already studied in previous chapters (like the Generalized Linear Models). The others being only toolboxes, they don't have their place in such a book.

Typical time series patterns can be seen in the figure below:

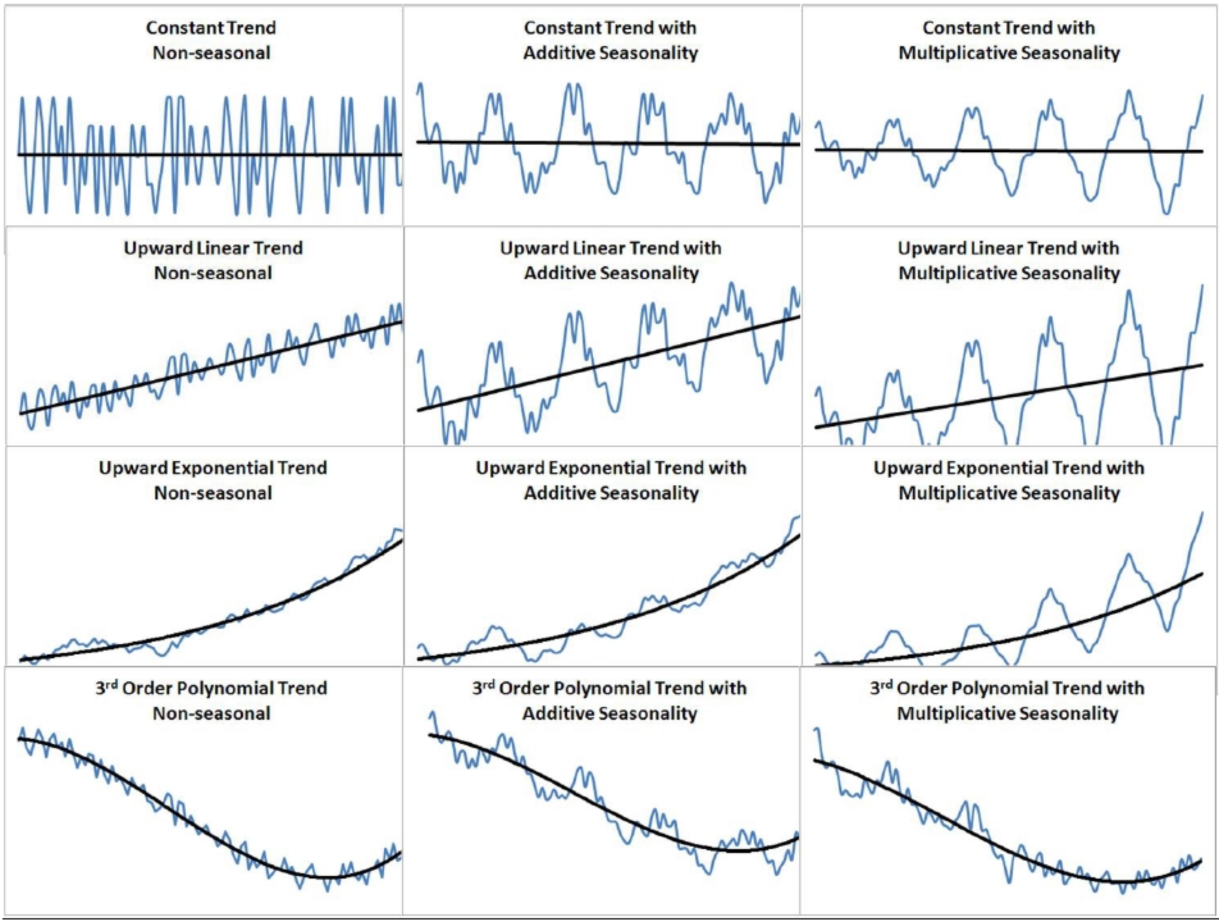


Figure 15.211: **Exhibit of easily recognizable time series patterns**

and a common classification of existing models can visually be given by:

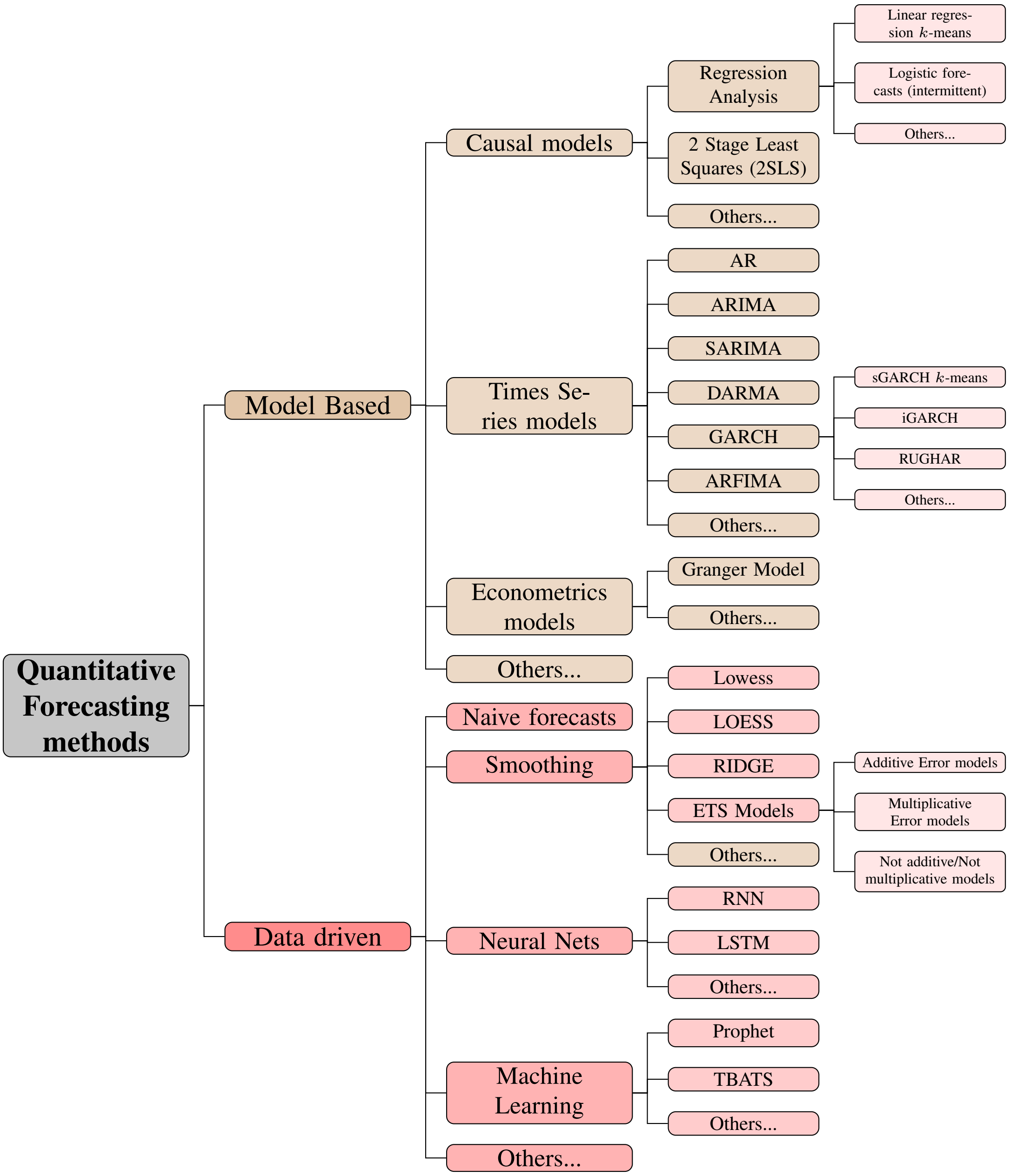
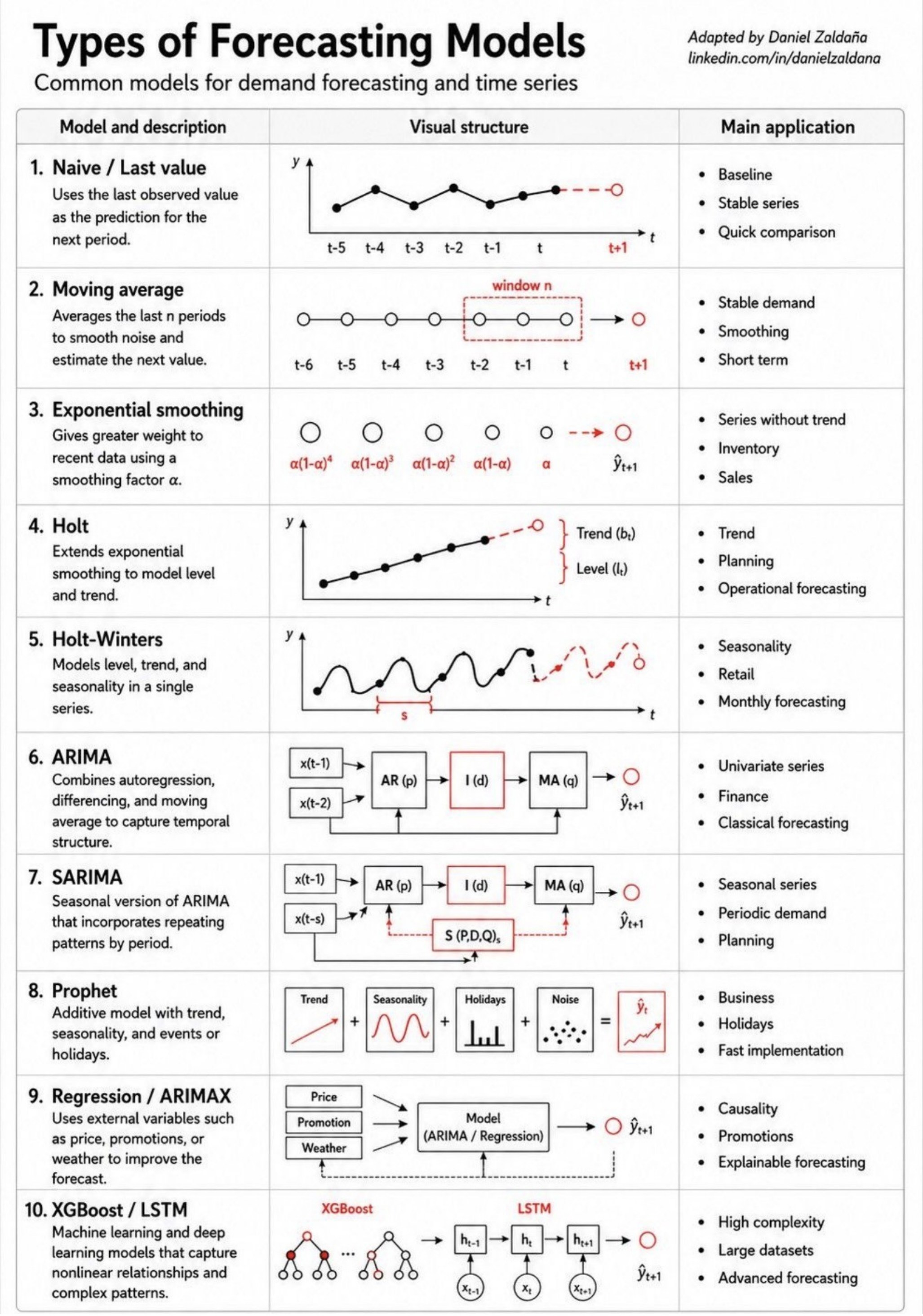

Quantitative Forecasting methods
Model Based
Causal models
Regression Analysis
Linear regression k-means
Logistic forecasts (intermittent)
Others...
2 Stage Least Squares (2SLS)
Others...
Times Series models
AR
ARIMA
SARIMA
DARMA
GARCH
sGARCH k-means
iGARCH
RUGHAR
Others...
ARFIMA
Others...
Econometrics models
Granger Model
Others...
Others...
Data driven
Naive forecasts
Smoothing
Lowess
LOESS
RIDGE
ETS Models
Additive Error models
Multiplicative Error models
Not additive/Not multiplicative models
Others...
Neural Nets
RNN
LSTM
Others...
Machine Learning
Prophet
TBATS
Others...
Others...

Figure 15.212: Non-exhaustive orgchart of forecasting techniques

Or another nice summary of the most famous one only:



Types of Forecasting Models

Common models for demand forecasting and time series

Adapted by Daniel Zaldaña
linkedin.com/in/danielzaldana

Model and description	Visual structure	Main application
1. Naive / Last value Uses the last observed value as the prediction for the next period.	y t-5 t-4 t-3 t-2 t-1 t t+1 t	- Baseline - Stable series - Quick comparison
2. Moving average Averages the last n periods to smooth noise and estimate the next value.	window n t-6 t-5 t-4 t-3 t-2 t-1 t t+1	- Stable demand - Smoothing - Short term
3. Exponential smoothing Gives greater weight to recent data using a smoothing factor α.	$\alpha(1-\alpha)^4$ $\alpha(1-\alpha)^3$ $\alpha(1-\alpha)^2$ $\alpha(1-\alpha)$ α $\hat{y}_{t+1}$	- Series without trend - Inventory - Sales
4. Holt Extends exponential smoothing to model level and trend.	y Trend (b_t) Level (l_t) t	- Trend - Planning - Operational forecasting
5. Holt-Winters Models level, trend, and seasonality in a single series.	y s t	- Seasonality - Retail - Monthly forecasting
6. ARIMA Combines autoregression, differencing, and moving average to capture temporal structure.	x(t-1) x(t-2) AR (p) → I (d) → MA (q) → $\hat{y}_{t+1}$	- Univariate series - Finance - Classical forecasting
7. SARIMA Seasonal version of ARIMA that incorporates repeating patterns by period.	x(t-1) x(t-s) AR (p) → I (d) → MA (q) → $\hat{y}_{t+1}$ $S(P,D,Q)_s$	- Seasonal series - Periodic demand - Planning
8. Prophet Additive model with trend, seasonality, and events or holidays.	Trend + Seasonality + Holidays + Noise = $\hat{y}_t$	- Business - Holidays - Fast implementation
9. Regression / ARIMAX Uses external variables such as price, promotions, or weather to improve the forecast.	Price Promotion Weather Model (ARIMA / Regression) → $\hat{y}_{t+1}$	- Causality - Promotions - Explainable forecasting
10. XGBoost / LSTM Machine learning and deep learning models that capture nonlinear relationships and complex patterns.	XGBoost ... → LSTM h_{t-1} → h_t → h_{t+1} → $\hat{y}_{t+1}$ x_{t-1} x_t x_{t+1}	- High complexity - Large datasets - Advanced forecasting

We will in what follows obviously focus first on classic undergraduate family of models.

Remark

Using a single variable to analyse a complex situation is a good way to reach bad conclusions (never forget this when you make time forecasting based only with time t as explanatory variable...!).

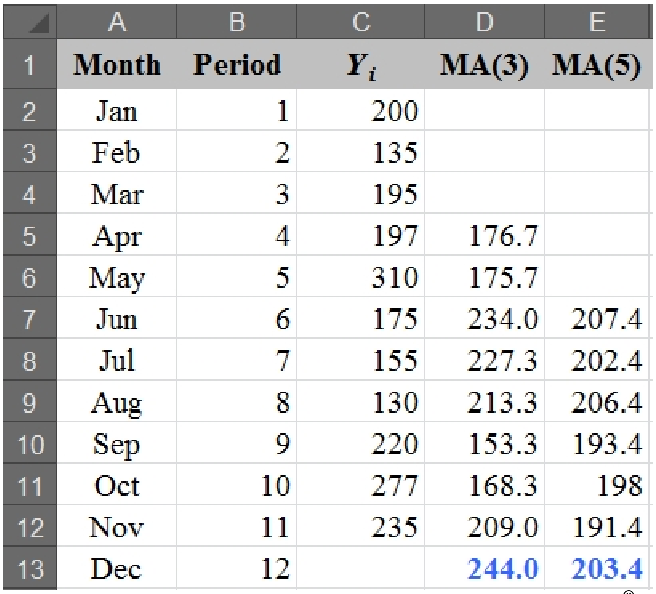
15.3.13.3.1 Simple Moving Average (moving average smoothing)

The simple moving average (MA) forecast technique of order k is certainly the simplest and also the worst empirical technique in terms of predictive power. Based on the definition of the moving average (see section Statistics page 567) it is defined by:

$$\boxed{\mathrm{MA}(k)=F_{t+1}=\frac{\displaystyle\sum_{i=t-k+1}^{t} Y_i}{k}} \tag{15.3.1462}$$

and is therefore a deterministic forecast on only a single subsequent period.

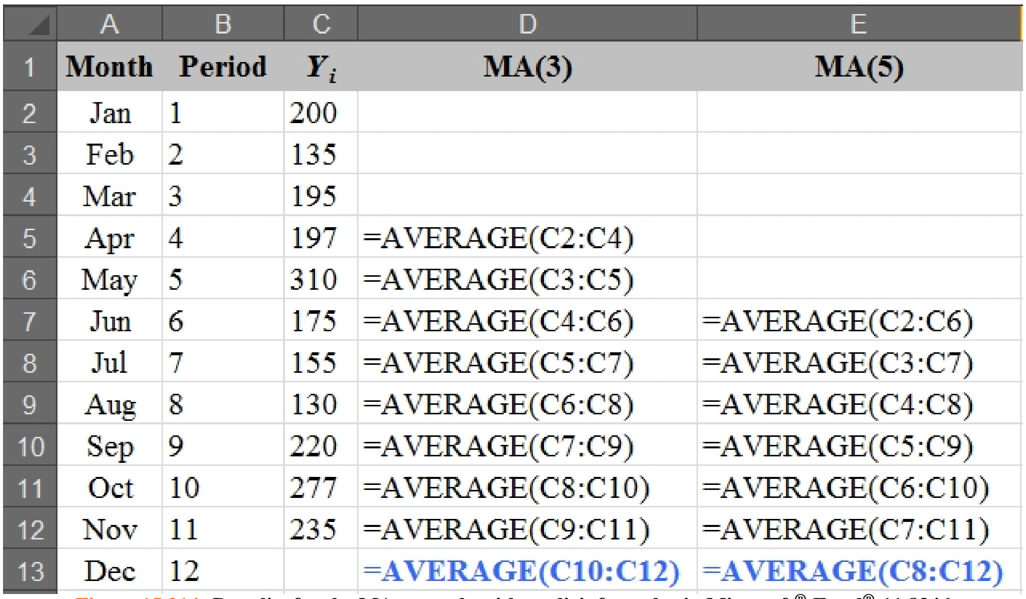
Let us see an example with Microsoft® Excel® 11.8346 with a forecast on the horizon of a month with moving average smoothing models in columns `D` and `E`:



	A	B	C	D	E
1	**Month**	**Period**	Y_i	**MA(3)**	**MA(5)**
2	Jan	1	200		
3	Feb	2	135		
4	Mar	3	195		
5	Apr	4	197	176.7	
6	May	5	310	175.7	
7	Jun	6	175	234.0	207.4
8	Jul	7	155	227.3	202.4
9	Aug	8	130	213.3	206.4
10	Sep	9	220	153.3	193.4
11	Oct	10	277	168.3	198
12	Nov	11	235	209.0	191.4
13	Dec	12		244.0	203.4

Figure 15.213: Base list for the example of moving average smoothing in Microsoft® Excel® 11.8346

Explicitly:



	A	B	C	D	E
1	**Month**	**Period**	Y_i	**MA(3)**	**MA(5)**
2	Jan	1	200		
3	Feb	2	135		
4	Mar	3	195		
5	Apr	4	197	=AVERAGE(C2:C4)	
6	May	5	310	=AVERAGE(C3:C5)	
7	Jun	6	175	=AVERAGE(C4:C6)	=AVERAGE(C2:C6)
8	Jul	7	155	=AVERAGE(C5:C7)	=AVERAGE(C3:C7)
9	Aug	8	130	=AVERAGE(C6:C8)	=AVERAGE(C4:C8)
10	Sep	9	220	=AVERAGE(C7:C9)	=AVERAGE(C5:C9)
11	Oct	10	277	=AVERAGE(C8:C10)	=AVERAGE(C6:C10)
12	Nov	11	235	=AVERAGE(C9:C11)	=AVERAGE(C7:C11)
13	Dec	12		**=AVERAGE(C10:C12)**	**=AVERAGE(C8:C12)**

Figure 15.214: Base list for the MA example with explicit formulas in Microsoft® Excel® 11.8346

and the corresponding chart:

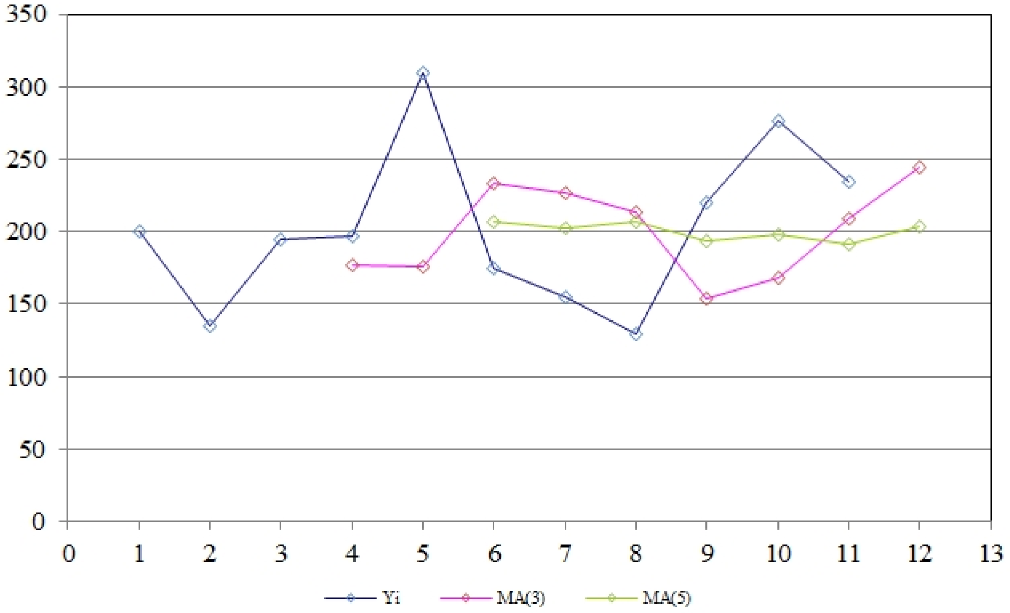


Figure 15.215: MA(3) and MA(5) plot with Microsoft® Excel® 11.8346

And elementary calculations give (we can provide the details with Microsoft® Excel® on readers request):

$$\text{MAPE} = 100\frac{1}{n}\sum_{t=1}^{n}\frac{|e_t|}{Y_t} = 34.84\% \qquad \text{MAD} = \frac{1}{n}\sum_{t=1}^{n}|e_t| = 71.\bar{3} \qquad \text{MSD} = \frac{1}{n}\sum_{t=1}^{n}e_t^2 = 6380.\bar{6} \tag{15.3.1463}$$

We have not used the Analysis Toolpack of Microsoft® Excel® 11.8346 (the latter incorporating a tool automatically making the calculations of the moving average) only because we wanted show the reader the explicit application of the mathematical relationship given above.

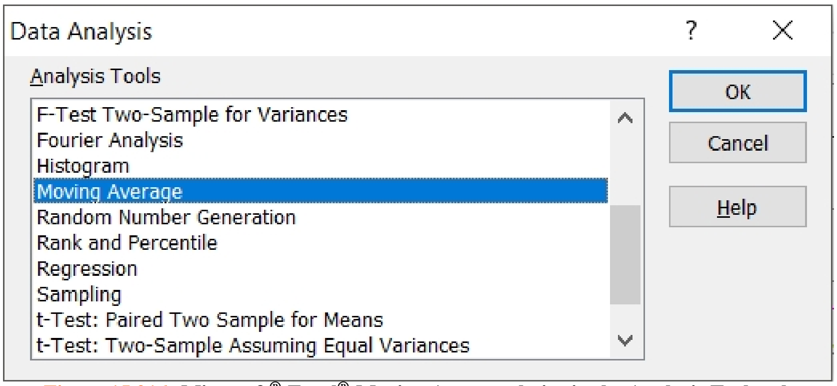

Data Analysis
?
Analysis Tools
F-Test Two-Sample for Variances
Fourier Analysis
Histogram
Moving Average
Random Number Generation
Rank and Percentile
Regression
Sampling
t-Test: Paired Two Sample for Means
t-Test: Two-Sample Assuming Equal Variances
OK
Cancel
Help

Figure 15.216: **Microsoft® Excel® Moving Average choice in the Analysis Toolpack**

The choice of the number of periods k to be used in calculating the simple moving average depends very much on expected variations in the data. This can be illustrated by two characteristics of the forecast:

- "Stability": By averaging several periods, we attenuate the random fluctuations and therefore the prediction is more stable. Stability is the property of the forecast to not fluctuate randomly. Win in stability is an advantage if there is a lot of random fluctuations in the data. A moving average will win in stability if a larger number of periods is used in the calculation of the average. A win in stability is desirable only to the point where random fluctuations are mitigated. If the number of periods used in the calculation of the average is too large, the average will be so stable that it will respond too slowly to non-random changes.
- "Reactivity": Reactivity is the property that the forecast quickly adjust to a change in the actual average level. The use of a reactive forecast is appropriate in the case where random fluctuations are small. The smaller the number of periods used in the calculation of the moving average is, the more the forecast model will be reactive.

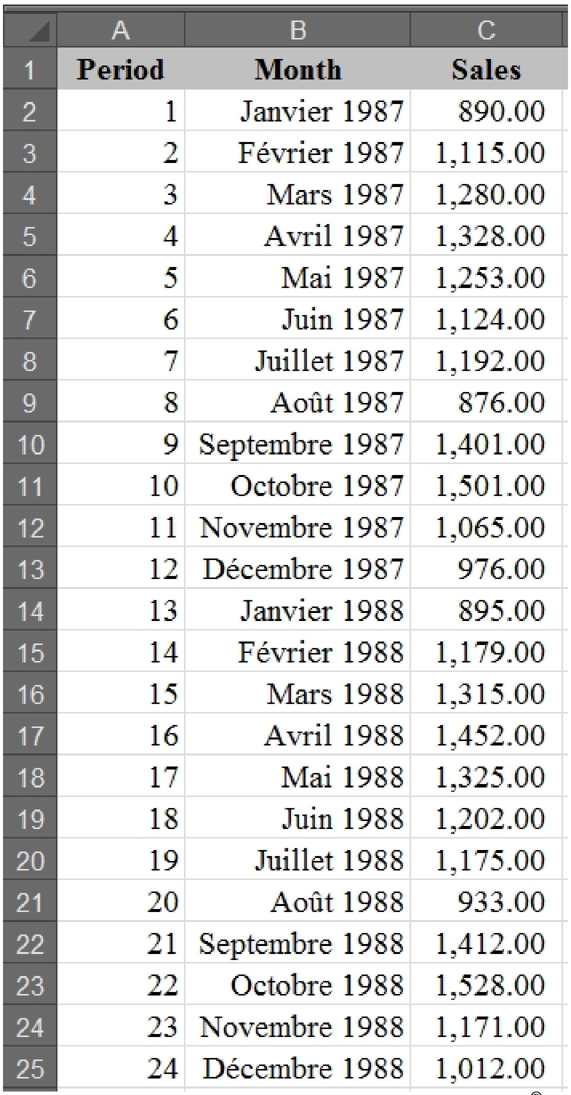
Remark

Let us notice that some forecasting techniques do moving averages of moving averages ... In the literature, this is named "double moving average" ... There are also models that apply empirical weight to past periods, this is named "weighted moving average" and this is in my opinion out of interest.

15.3.13.3.2 Linear Model With Seasonal Coefficients (LMSC)

The linear predictive technique with seasonal factors is simple to implement and to understand for most people. It combines a simple linear regression (see section Numerical Methods page 6322) and seasonal factors calculated on the basis of past values relatively to the linear model. This technique allows to identify trends by smoothing the time series and showing significant variations in time.

Let us see a companion example with Microsoft® Excel® 11.8346:



	A	B	C
1	**Period**	**Month**	**Sales**
2	1	Janvier 1987	890.00
3	2	Février 1987	1,115.00
4	3	Mars 1987	1,280.00
5	4	Avril 1987	1,328.00
6	5	Mai 1987	1,253.00
7	6	Juin 1987	1,124.00
8	7	Juillet 1987	1,192.00
9	8	Août 1987	876.00
10	9	Septembre 1987	1,401.00
11	10	Octobre 1987	1,501.00
12	11	Novembre 1987	1,065.00
13	12	Décembre 1987	976.00
14	13	Janvier 1988	895.00
15	14	Février 1988	1,179.00
16	15	Mars 1988	1,315.00
17	16	Avril 1988	1,452.00
18	17	Mai 1988	1,325.00
19	18	Juin 1988	1,202.00
20	19	Juillet 1988	1,175.00
21	20	Août 1988	933.00
22	21	Septembre 1988	1,412.00
23	22	Octobre 1988	1,528.00
24	23	Novembre 1988	1,171.00
25	24	Décembre 1988	1,012.00

Figure 15.217: **Companion LMSC example starting values in Microsoft® Excel® 11.8346**

Which looks like:

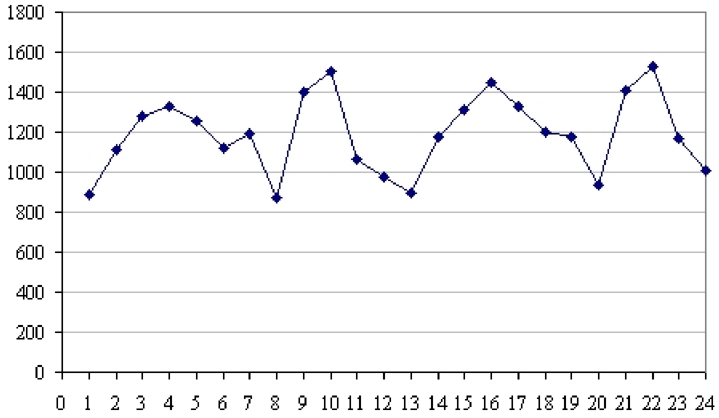


Figure 15.218: Corresponding Sales plot in Microsoft® Excel® 11.8346

The first thing to do is to determine the equation of the regression line (which allows verbatim also to make inference!). That's simple to solve with the least squares method proved in the section of Numerical Methods and extremely simple to obtain with automatic tools like the one integrated in Microsoft® Excel® 11.8346 (Analysis Toolpack again... or the functions already seen in the section of Numerical Methods):

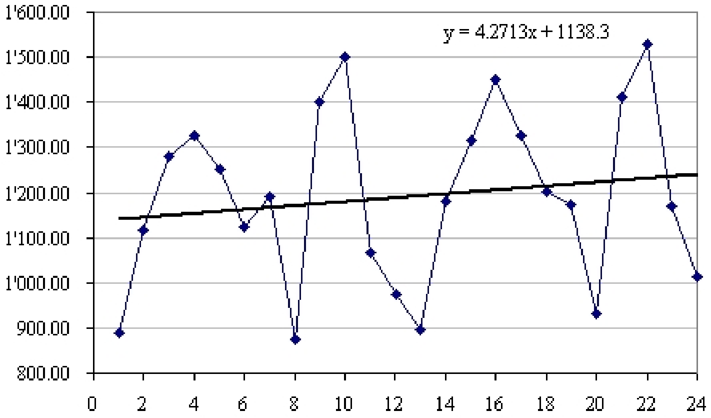
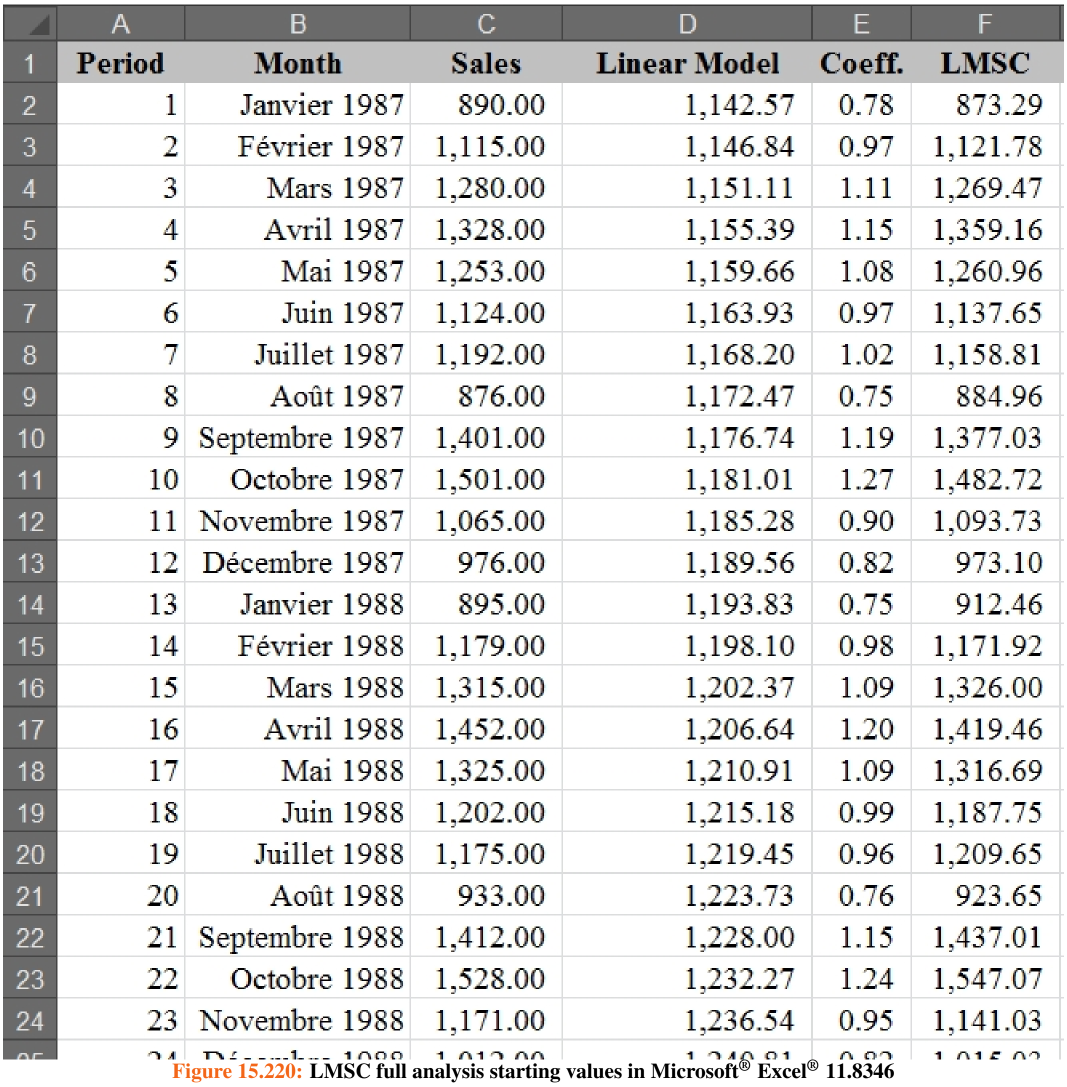


Figure 15.219: Linear regression equation for Sales in Microsoft® Excel® 11.8346

This allows us to have three new columns in our Microsoft® Excel® list. One with the values of the linear model and another with the ratio between the actual and the linear model (which corresponds to the seasonal coefficients!):



	A	B	C	D	E	F
1	Period	Month	Sales	Linear Model	Coeff.	LMSC
2	1	Janvier 1987	890.00	1,142.57	0.78	873.29
3	2	Février 1987	1,115.00	1,146.84	0.97	1,121.78
4	3	Mars 1987	1,280.00	1,151.11	1.11	1,269.47
5	4	Avril 1987	1,328.00	1,155.39	1.15	1,359.16
6	5	Mai 1987	1,253.00	1,159.66	1.08	1,260.96
7	6	Juin 1987	1,124.00	1,163.93	0.97	1,137.65
8	7	Juillet 1987	1,192.00	1,168.20	1.02	1,158.81
9	8	Août 1987	876.00	1,172.47	0.75	884.96
10	9	Septembre 1987	1,401.00	1,176.74	1.19	1,377.03
11	10	Octobre 1987	1,501.00	1,181.01	1.27	1,482.72
12	11	Novembre 1987	1,065.00	1,185.28	0.90	1,093.73
13	12	Décembre 1987	976.00	1,189.56	0.82	973.10
14	13	Janvier 1988	895.00	1,193.83	0.75	912.46
15	14	Février 1988	1,179.00	1,198.10	0.98	1,171.92
16	15	Mars 1988	1,315.00	1,202.37	1.09	1,326.00
17	16	Avril 1988	1,452.00	1,206.64	1.20	1,419.46
18	17	Mai 1988	1,325.00	1,210.91	1.09	1,316.69
19	18	Juin 1988	1,202.00	1,215.18	0.99	1,187.75
20	19	Juillet 1988	1,175.00	1,219.45	0.96	1,209.65
21	20	Août 1988	933.00	1,223.73	0.76	923.65
22	21	Septembre 1988	1,412.00	1,228.00	1.15	1,437.01
23	22	Octobre 1988	1,528.00	1,232.27	1.24	1,547.07
24	23	Novembre 1988	1,171.00	1,236.54	0.95	1,141.03

Figure 15.220: LMSC full analysis starting values in Microsoft® Excel® 11.8346

That is to say visually:

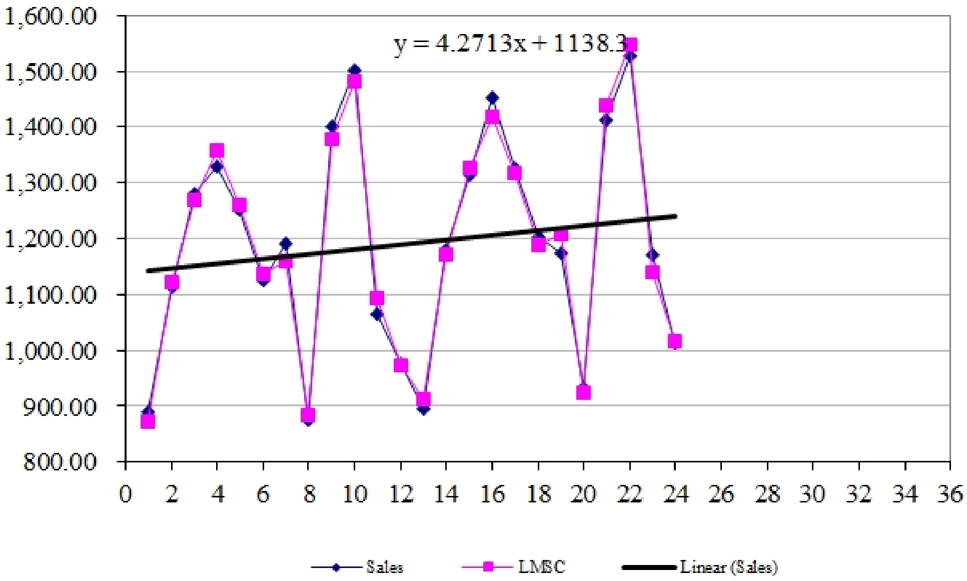


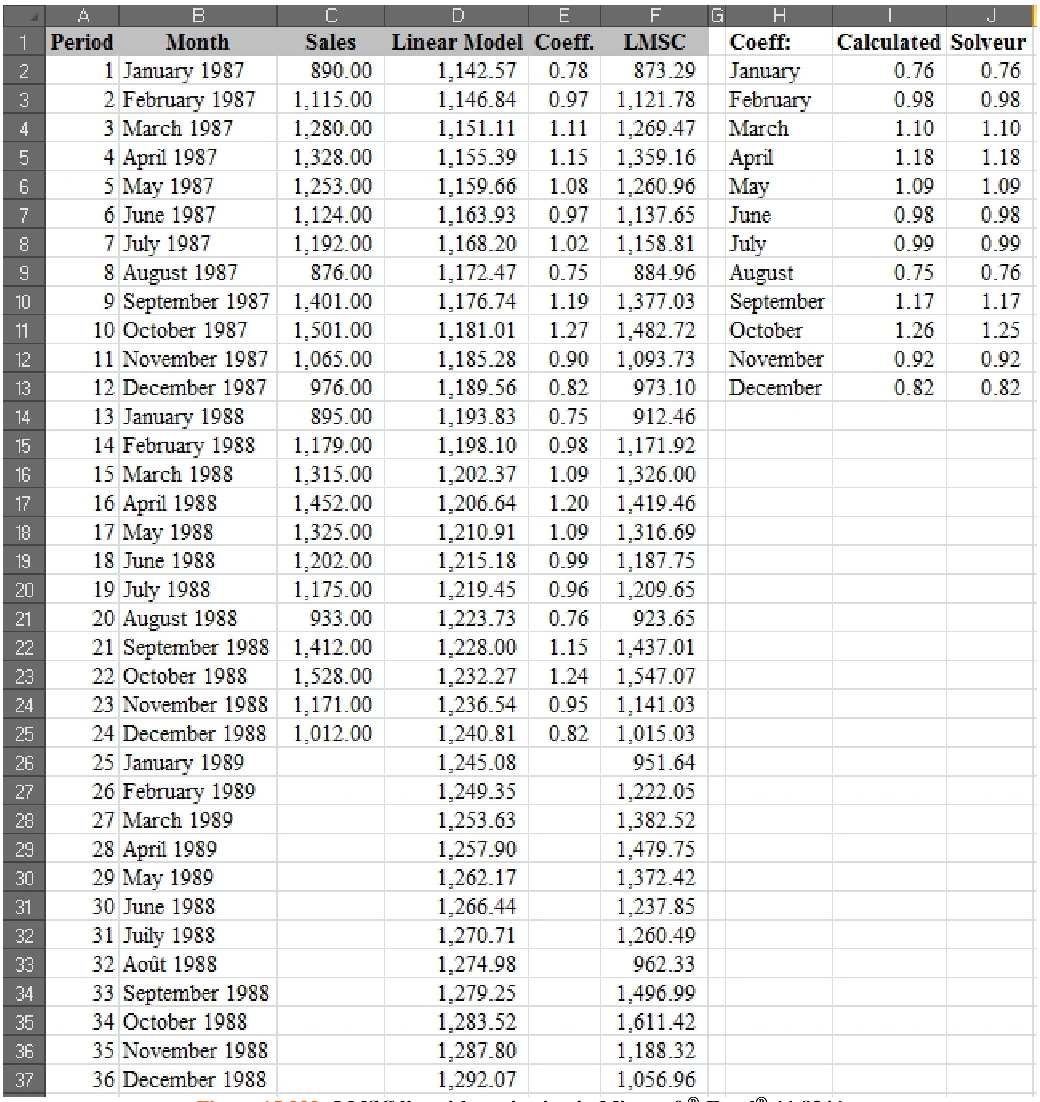
Figure 15.221: LMSC full plot in Microsoft® Excel® 11.8346

We can then make a prediction based on the linear regression multiplied by the average of the seasonal coefficients of each month. What we write in this case:

$$\boxed{F_t=(at_i+b)\bar{c}_{i \mod 12}} \tag{15.3.1464}$$

So this is a deterministic method to make forecasts on as many future periods as desired with the ability to make statistical inference if necessary as there is an underlying linear Gaussian model (this is the only deterministic model on which the inference is directly applicable as far as we know).

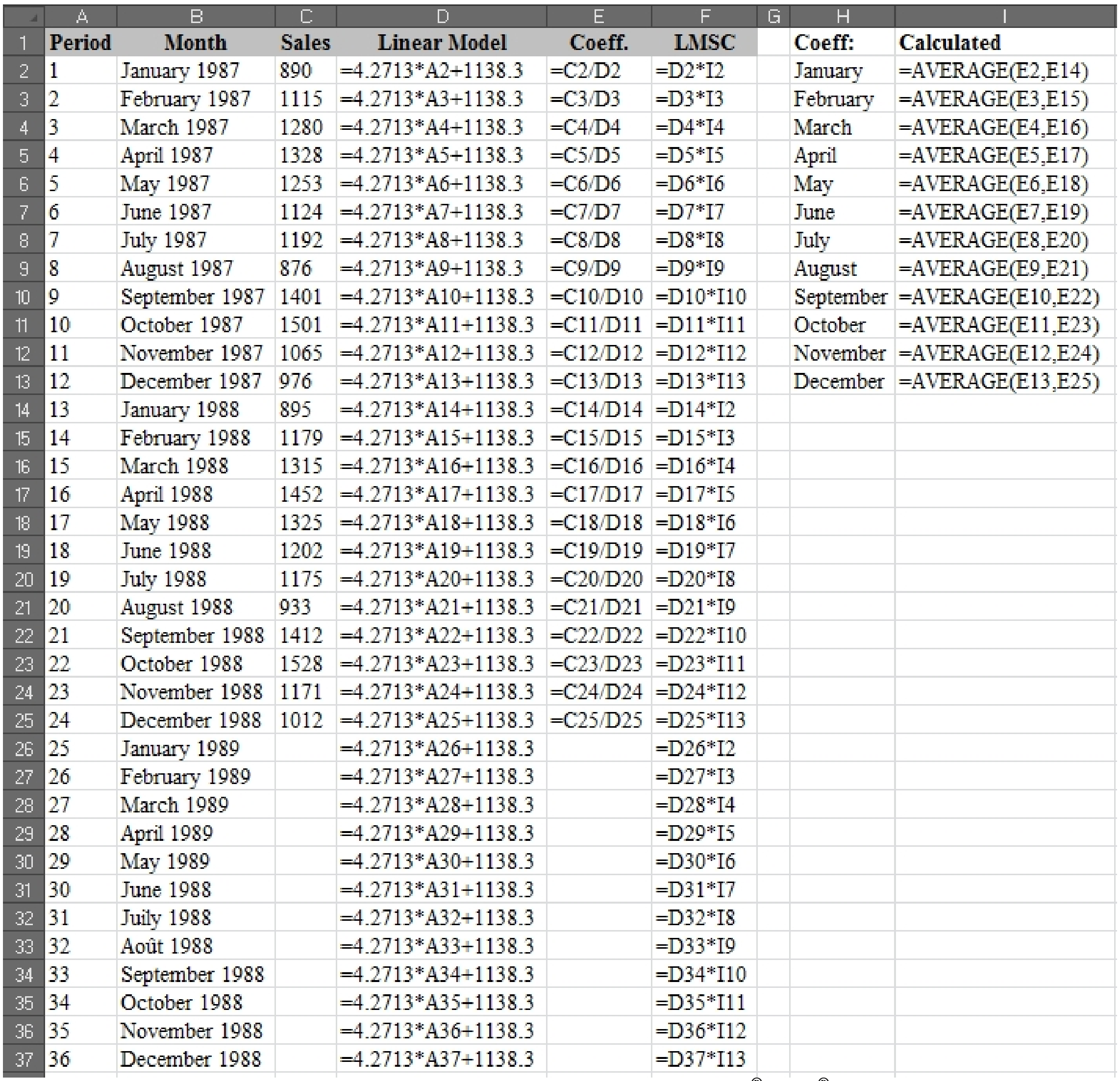
So for example:



	A	B	C	D	E	F	G	H	I	J
1	**Period**	**Month**	**Sales**	**Linear Model**	**Coeff.**	**LMSC**		**Coeff:**	**Calculated**	**Solveur**
2	1	January 1987	890.00	1,142.57	0.78	873.29		January	0.76	0.76
3	2	February 1987	1,115.00	1,146.84	0.97	1,121.78		February	0.98	0.98
4	3	March 1987	1,280.00	1,151.11	1.11	1,269.47		March	1.10	1.10
5	4	April 1987	1,328.00	1,155.39	1.15	1,359.16		April	1.18	1.18
6	5	May 1987	1,253.00	1,159.66	1.08	1,260.96		May	1.09	1.09
7	6	June 1987	1,124.00	1,163.93	0.97	1,137.65		June	0.98	0.98
8	7	July 1987	1,192.00	1,168.20	1.02	1,158.81		July	0.99	0.99
9	8	August 1987	876.00	1,172.47	0.75	884.96		August	0.75	0.76
10	9	September 1987	1,401.00	1,176.74	1.19	1,377.03		September	1.17	1.17
11	10	October 1987	1,501.00	1,181.01	1.27	1,482.72		October	1.26	1.25
12	11	November 1987	1,065.00	1,185.28	0.90	1,093.73		November	0.92	0.92
13	12	December 1987	976.00	1,189.56	0.82	973.10		December	0.82	0.82
14	13	January 1988	895.00	1,193.83	0.75	912.46				
15	14	February 1988	1,179.00	1,198.10	0.98	1,171.92				
16	15	March 1988	1,315.00	1,202.37	1.09	1,326.00				
17	16	April 1988	1,452.00	1,206.64	1.20	1,419.46				
18	17	May 1988	1,325.00	1,210.91	1.09	1,316.69				
19	18	June 1988	1,202.00	1,215.18	0.99	1,187.75				
20	19	July 1988	1,175.00	1,219.45	0.96	1,209.65				
21	20	August 1988	933.00	1,223.73	0.76	923.65				
22	21	September 1988	1,412.00	1,228.00	1.15	1,437.01				
23	22	October 1988	1,528.00	1,232.27	1.24	1,547.07				
24	23	November 1988	1,171.00	1,236.54	0.95	1,141.03				
25	24	December 1988	1,012.00	1,240.81	0.82	1,015.03				
26	25	January 1989		1,245.08		951.64				
27	26	February 1989		1,249.35		1,222.05				
28	27	March 1989		1,253.63		1,382.52				
29	28	April 1989		1,257.90		1,479.75				
30	29	May 1989		1,262.17		1,372.42				
31	30	June 1988		1,266.44		1,237.85				
32	31	Juily 1988		1,270.71		1,260.49				
33	32	Août 1988		1,274.98		962.33				
34	33	September 1988		1,279.25		1,496.99				
35	34	October 1988		1,283.52		1,611.42				
36	35	November 1988		1,287.80		1,188.32				
37	36	December 1988		1,292.07		1,056.96				

Figure 15.222: LMSC list with projection in Microsoft® Excel® 11.8346

Explicitly:



	A	B	C	D	E	F	G	H	I
1	**Period**	**Month**	**Sales**	**Linear Model**	**Coeff.**	**LMSC**		**Coeff:**	**Calculated**
2	1	January 1987	890	=4.2713*A2+1138.3	=C2/D2	=D2*I2		January	=AVERAGE(E2,E14)
3	2	February 1987	1115	=4.2713*A3+1138.3	=C3/D3	=D3*I3		February	=AVERAGE(E3,E15)
4	3	March 1987	1280	=4.2713*A4+1138.3	=C4/D4	=D4*I4		March	=AVERAGE(E4,E16)
5	4	April 1987	1328	=4.2713*A5+1138.3	=C5/D5	=D5*I5		April	=AVERAGE(E5,E17)
6	5	May 1987	1253	=4.2713*A6+1138.3	=C6/D6	=D6*I6		May	=AVERAGE(E6,E18)
7	6	June 1987	1124	=4.2713*A7+1138.3	=C7/D7	=D7*I7		June	=AVERAGE(E7,E19)
8	7	July 1987	1192	=4.2713*A8+1138.3	=C8/D8	=D8*I8		July	=AVERAGE(E8,E20)
9	8	August 1987	876	=4.2713*A9+1138.3	=C9/D9	=D9*I9		August	=AVERAGE(E9,E21)
10	9	September 1987	1401	=4.2713*A10+1138.3	=C10/D10	=D10*I10		September	=AVERAGE(E10,E22)
11	10	October 1987	1501	=4.2713*A11+1138.3	=C11/D11	=D11*I11		October	=AVERAGE(E11,E23)
12	11	November 1987	1065	=4.2713*A12+1138.3	=C12/D12	=D12*I12		November	=AVERAGE(E12,E24)
13	12	December 1987	976	=4.2713*A13+1138.3	=C13/D13	=D13*I13		December	=AVERAGE(E13,E25)
14	13	January 1988	895	=4.2713*A14+1138.3	=C14/D14	=D14*I2			
15	14	February 1988	1179	=4.2713*A15+1138.3	=C15/D15	=D15*I3			
16	15	March 1988	1315	=4.2713*A16+1138.3	=C16/D16	=D16*I4			
17	16	April 1988	1452	=4.2713*A17+1138.3	=C17/D17	=D17*I5			
18	17	May 1988	1325	=4.2713*A18+1138.3	=C18/D18	=D18*I6			
19	18	June 1988	1202	=4.2713*A19+1138.3	=C19/D19	=D19*I7			
20	19	July 1988	1175	=4.2713*A20+1138.3	=C20/D20	=D20*I8			
21	20	August 1988	933	=4.2713*A21+1138.3	=C21/D21	=D21*I9			
22	21	September 1988	1412	=4.2713*A22+1138.3	=C22/D22	=D22*I10			
23	22	October 1988	1528	=4.2713*A23+1138.3	=C23/D23	=D23*I11			
24	23	November 1988	1171	=4.2713*A24+1138.3	=C24/D24	=D24*I12			
25	24	December 1988	1012	=4.2713*A25+1138.3	=C25/D25	=D25*I13			
26	25	January 1989		=4.2713*A26+1138.3		=D26*I2			
27	26	February 1989		=4.2713*A27+1138.3		=D27*I3			
28	27	March 1989		=4.2713*A28+1138.3		=D28*I4			
29	28	April 1989		=4.2713*A29+1138.3		=D29*I5			
30	29	May 1989		=4.2713*A30+1138.3		=D30*I6			
31	30	June 1988		=4.2713*A31+1138.3		=D31*I7			
32	31	July 1988		=4.2713*A32+1138.3		=D32*I8			
33	32	Août 1988		=4.2713*A33+1138.3		=D33*I9			
34	33	September 1988		=4.2713*A34+1138.3		=D34*I10			
35	34	October 1988		=4.2713*A35+1138.3		=D35*I11			
36	35	November 1988		=4.2713*A36+1138.3		=D36*I12			
37	36	December 1988		=4.2713*A37+1138.3		=D37*I13			

Figure 15.223: **LMSC projection with formulas in Microsoft® Excel® 11.8346**

That is to say visually:

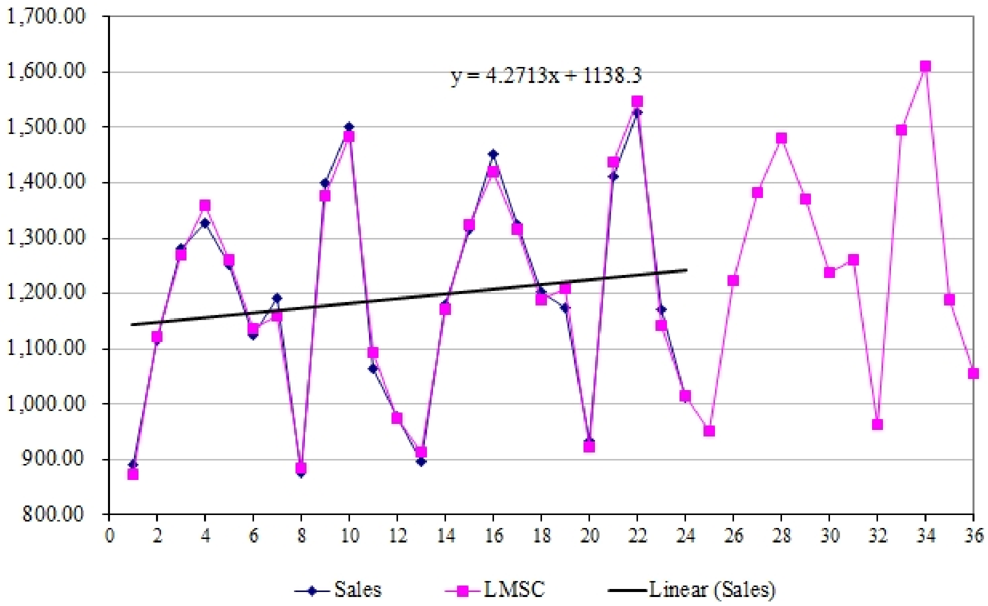


Figure 15.224: **LMSC plot with projection in Microsoft® Excel® 11.8346**

And elementary calculations give (we can provide the details with Microsoft® Excel® on readers request):

$$\text{MSD} = \frac{1}{n}\sum_{t=1}^{n} e_t^2 \cong 398.898 \qquad (15.3.1465)$$

15.3.13.3.3 Exponential Smoothing models

Exponential smoothing is a very common time series forecasting procedure capable of including both trend and seasonality. The exponential smoothing procedure is a combination of three other much simpler components, all of which are smoothing methods:

- **Simple Exponential Smoothing (SES)**: Simple exponential smoothing, also named single exponential smoothing, assumes that the time series has no change in level. Thus, it can not be used with series that contain trend, seasonality, or both.
- **Holt's Exponential Smoothing (HES)**: Holt's exponential smoothing is one step above simple exponential smoothing, as it allows the time series data to have a trend component. Holt's exponential smoothing is still incapable of cope with seasonal data.
- **Winter's Exponential Smoothing (WES)**: Winter's exponential smoothing is an extension to Holt's exponential smoothing that finally allows for the inclusion of seasonality. Winter's exponential smoothing is what is referred to as the Holt-Winters method.

15.3.13.3.3.1 Simple Exponential Smoothing (EWMA)

The prediction F_t provided by the exponential smoothing method (also named "Exponential Weighted Moving Average (EWMA)" or "Single Exponential Smoothing (SES)"), also named "exponential weighted moving average", with "smoothing constant" $\beta \in]0,1[$ is given as we will prove it later by:

$$F_{t+1}=(1-\beta)\sum_{i=0}^{t-1}\beta^i Y_{t-i} \tag{15.3.1466}$$

This definition is based on a simple idea (yet totally empirical!): we assume that the observations influence less the forecast as they are far from the time t for which we forecast and that this influence decreases exponentially (this model remains even in the early 121st century (holocene calendar) is still very used to model the variance of financial instruments). We then see easily at first glance that:

- Note the closest the smoothing constant β is to 1, the more the influence of past observations goes back further in time and the forecast is rigid, that is to say not very sensitive to early economic fluctuations.
- On the contrary, more the smoothing constant is close to zero, more the forecast is influenced by recent observations (flexible forecasting).

Let us prove first that the previous definition can be written in another form less technique and that is most common in statistical books:

$$F_{t+1}=(1-\beta)\sum_{i=0}^{t-1}\beta^i Y_{t-i}=(1-\beta)\sum_{i=1}^{t-1}\beta^i Y_{t-i}+(1-\beta)Y_t \tag{15.3.1467}$$

and we also have verbatim:

$$F_t=(1-\beta)\sum_{i=0}^{t-2}\beta^i Y_{t-1-i} \tag{15.3.1468}$$

Therefore if we reindex the sum we find a well-known writing in the finance world (volatility measurement of the RiskMetrics model):

$$\begin{aligned}F_{t+1}&=(1-\beta)\sum_{i=1}^{t-1}\beta^i Y_{t-1}+(1-b\beta)Y_t=(1-\beta)\sum_{i=0}^{t-1-1}\beta^{i+1}Y_{t-1-i}+(1-\beta)Y_t\\&=(1-\beta)\sum_{i=0}^{t-2}\beta^{i+1}Y_{t-1-i}+(1-\beta)Y_t=\beta(1-\beta)\sum_{i=0}^{t-2}\beta^i Y_{t-1-i}+(1-\beta)Y_t\\&=\beta F_t+(1-\beta)Y_t\end{aligned} \tag{15.3.1469}$$

The forecast F_{t+1} appears as the weighted average between the forecast made at time t and the last observation Y_t, the weight given to the latter observation is even stronger than the smoothing constant is small.

We will now demonstrate why softwares forecast the EWMA models with a mean equal to the last measurement:

$$\begin{aligned}\mathrm{E}(F_{t+1})&=\mathrm{E}(\beta F_t+(1-\beta)Y_t)=\beta\mathrm{E}(F_t)+(1-\beta)\mathrm{E}(Y_t)\\&=\beta\mathrm{E}(Y_t)+(1-\beta)\mathrm{E}(Y_t)=\mathrm{E}(Y_t)\end{aligned}\qquad(15.3.1470)$$

In practice, the choice of the smoothing constant is often based on subjective criteria of rigidity or flexibility of the prediction. But there is a more objective method of determining the constant using operational research tools (see section Numerical Methods page 6645) where the constant minimizes the sum of squared forecast errors.

However in the majority of contemporary statistics books, the smoothing constant is defined as:

$$\lambda=1-\beta\qquad(15.3.1471)$$

It comes then:

$$F_{t+1}=\lambda\sum_{i=0}^{t-1}(1-\lambda)^iY_{t-i}=(1-\lambda)F_t+\lambda Y_t=F_t+\lambda(Y_t-F_t)\qquad(15.3.1472)$$

writing that leads us to consider the single exponential smoothing as a simple weighted moving average.

Remark

The above relation is denoted in the framework of ARIMA as:

$$X_{t+1}=X_t+\alpha\varepsilon\qquad(15.3.1473)$$

and we recognise here an ARIMA$(0,1,1)$.

Now, let us start again from the above relation to write it in another form using recursion:

$$\begin{aligned}F_{t+1}&=\lambda Y_t+(1-\lambda)F_t=\lambda Y_t+(1-\lambda)(\lambda Y_{t-1}+(1-\lambda)F_{t-1})\\&=\left(\sum_{i=0}^{t-1}\lambda(1-\lambda)^iY_{t-i}\right)+(1-\lambda)^tF_1\end{aligned}\qquad(15.3.1474)$$

Let us now prove that the sum of weights is equal to the unit to see that it is indeed a weighted system. So we have:

$$\left(\sum_{i=0}^{t-1}\lambda(1-\lambda)^i\right)+(1-\lambda)^i=\lambda\left(\sum_{i=0}^{t-1}(1-\lambda)^i\right)+(1-\lambda)^t=\lambda\left(\sum_{i=0}^{t-1}x^i\right)+(1-\lambda)^t\qquad(15.3.1475)$$

We have proved in the section of Sequences and Series (see page 1679) that:

$$S_n=u_0\frac{1-q^{n+1}}{1-q}\qquad(15.3.1476)$$

Then it comes into our example:

$$S_{t-1}=u_0\frac{1-q^t}{1-q}=\frac{1-x^t}{1-x}=\frac{1-(1-\lambda)^t}{(1-(1-\lambda)}\qquad(15.3.1477)$$

Therefore:

$$\begin{aligned}\lambda\left(\sum_{i=0}^{t-1}x^i\right)+(1-\lambda)^t&=\lambda\frac{1-(1-\lambda)^t}{1-(1-\lambda)}+(1-\lambda)^t\\&=\lambda\frac{1-(1-\lambda)^t}{\lambda}+(1-\lambda)^t=1-(1-\lambda)^t+(1-\lambda)^t=1\end{aligned}\tag{15.3.1478}$$

Finally, let us talk about the origin of the name of this technique. So we have:

$$\begin{aligned}F_{t+1}&=(1-\lambda)F_t+\lambda Y_t=(1-\lambda)[(1-\lambda)F_{t-1}+\lambda Y_{t-1}]+\lambda Y_t\\&=(1-\lambda)^2F_{t-1}+\lambda(1-\lambda)Y_{t-1}+\lambda Y_t\\&=(1-\lambda)^2[(1-\lambda)F_{t-2}+\lambda Y_{t-2}]+\lambda(1-\lambda)Y_{t-1}+\lambda Y_t\\&=(1-\lambda)^3F_{t-2}+\lambda(1-\lambda)^2Y_{t-2}+\lambda(1-\lambda)Y_{t-1}+\lambda Y_t=\ldots\end{aligned}\tag{15.3.1479}$$

So we see in there the coefficients (weights) are exponentially decreasing with every previous unrolled estimate:

$$\ldots,\lambda(1-\lambda)^2,\lambda(1-\lambda),\lambda\tag{15.3.1480}$$

and as $1-\lambda$ is strictly less than unity, the factors $(1-\lambda)^2$ decreases exponentially (just plot the function to see it) as:

$$f(n)=(1-\lambda)^n\tag{15.3.1481}$$

with obviously $n\in\mathbb{N}$.

This being done, let us talk now about the problem of the initial values. So we have:

$$F_2=\lambda Y_1+(1-\lambda)F_1=\lambda Y_1+F_1-\lambda F_1\tag{15.3.1482}$$

It is then agreement to take:

$$F_1=Y_1\tag{15.3.1483}$$

Which gives us:

$$F_2=F_1\tag{15.3.1484}$$

This is why a program like Microsoft® Excel® 11.8346 returns for F_1 the error `#N/A`.

For the rest it is necessary that we study the origin of the expression of the single exponential smoothing. The idea at the base of mathematicians was to reduce the error between the real value Y_{t-j} and the prediction considered as constant a:

$$E=\sum_{j=0}^{t-1}(Y_{t-j}-a)\tag{15.3.1485}$$

but the goal was also to weaken the errors smaller than the unit and amplify those that were larger by squaring deviations such as:

$$E=\sum_{j=0}^{t-1}(Y_{t-j}-a)^2\tag{15.3.1486}$$

and finally to weight those that were remote in time from the prediction by a factor such that we have minimize finally:

$$E=\sum_{j=0}^{t-1}\beta^j(Y_{t-j}-a)^2 \tag{15.3.1487}$$

Taking the derivative of E with respect to a, and by cancelling it comes:

$$\frac{\partial E}{\partial a}=-2\sum_{j=0}^{t-1}\beta^j(Y_{t-j}-a)=0 \tag{15.3.1488}$$

We then get for $\beta\neq 1$: We then get for $\beta\neq 1$:

$$\sum_{j=0}^{t-1}\beta^jY_{t-j}=a\sum_{j=0}^{t-1}\beta^j=a\frac{1-\beta^t}{1-\beta} \tag{15.3.1489}$$

and thus finally:

$$a=\frac{1-\beta}{1-\beta^t}\sum_{j=0}^{t-1}\beta^jY_{t-j} \tag{15.3.1490}$$

and for large t we get:

$$a\cong(1-\beta)\sum_{j=0}^{t-1}\beta^jY_{t-j} \tag{15.3.1491}$$

We thus fall back on the relation defined at the beginning of the study of this model, which was:

$$\boxed{F_{t+1}=(1-\beta)\sum_{i=0}^{t-1}\beta^iY_{t-i}=(1-\beta)\sum_{i=1}^{t}\beta^{i-1}Y_{t+1-i}=(1-\beta)Y_t+\beta F_t} \tag{15.3.1492}$$

Notice that it is not obvious that the desired constant that minimizes the error can be used as prediction function. However it is the usage that is made by many practitioners... Some experts prefer to write the latter result as (do not try to compare this notation with that of the previous relation):

$$S_t=(1-\beta)Y_t+\beta S_{t-1}=\lambda Y_t+(1-\lambda)S_{t-1} \tag{15.3.1493}$$

where S refers to the "S" there is in the word "eStimated" or in the following form:

$$L_t=\alpha y_t+(1-\alpha)L_{t-1} \tag{15.3.1494}$$

This equation is known as the "level update equation" (as single exponential smoothing method does not take into account any trend or seasonality but rahter assumes that the time series data only has a level L).

There are even statistical softwares offering to our astonishment to do forecasts at time $t+n$ based on this method ... which seems to us quite crazy...

Remark

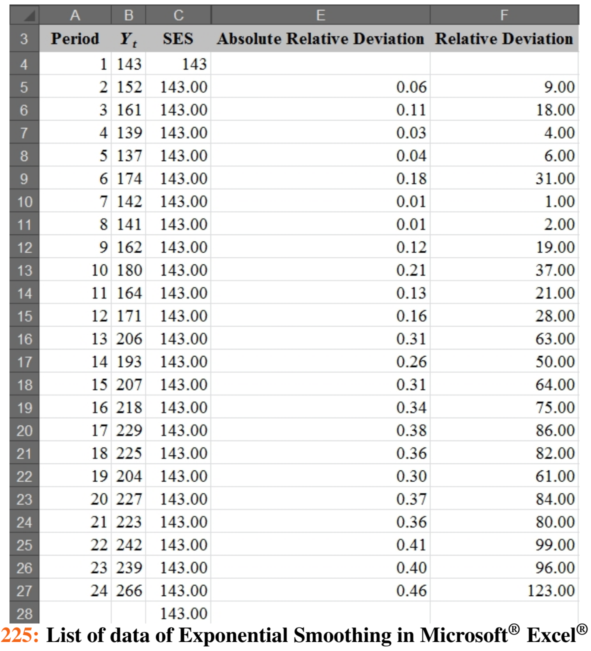
The simple exponential smoothing seems that it was first suggested by Charles C. Holt in 11957 (holocene calendar), but the formulation (the form of the predictive relation) is attributed to Brown. This is a technique that can be found implemented in almost all statistical software.

✍**Example:**

So we want to apply:

$$F_{t+1}=(1-\beta)Y_t+\beta F_t=\lambda Y_t+(1-\lambda)F_t \tag{15.3.1495}$$

Let us consider the following table (the column `C` already contains the mathematical model which we will detail the formulas just after):

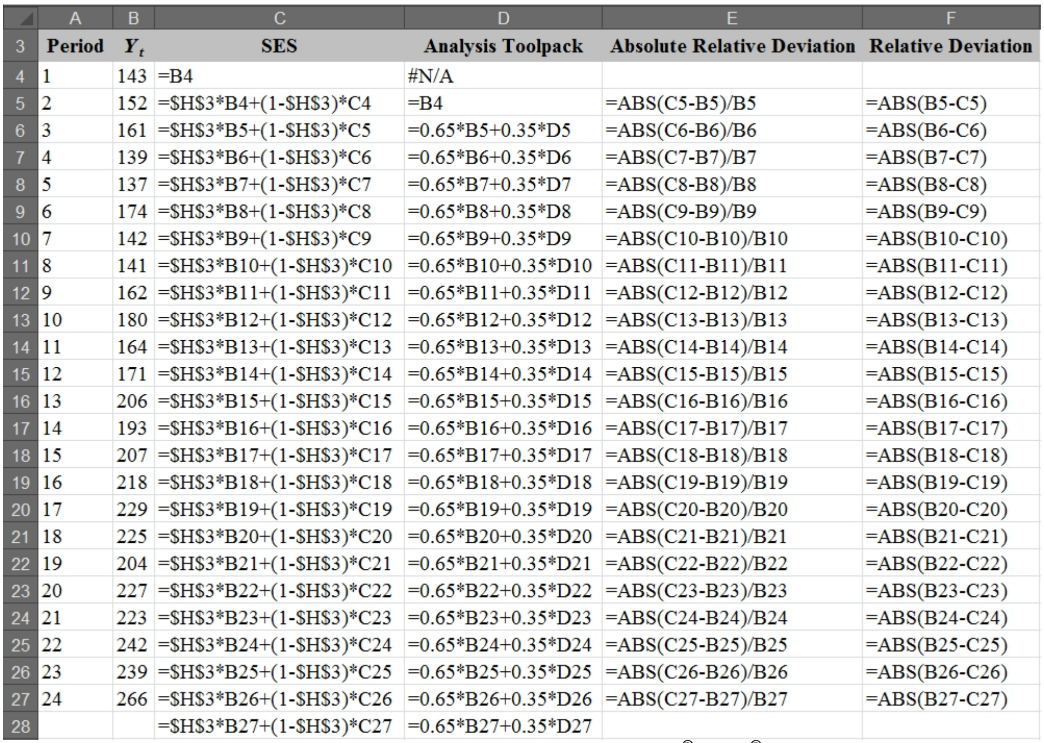


	A	B	C	E	F
3	Period	Y_t	SES	Absolute Relative Deviation	Relative Deviation
4	1	143	143		
5	2	152	143.00	0.06	9.00
6	3	161	143.00	0.11	18.00
7	4	139	143.00	0.03	4.00
8	5	137	143.00	0.04	6.00
9	6	174	143.00	0.18	31.00
10	7	142	143.00	0.01	1.00
11	8	141	143.00	0.01	2.00
12	9	162	143.00	0.12	19.00
13	10	180	143.00	0.21	37.00
14	11	164	143.00	0.13	21.00
15	12	171	143.00	0.16	28.00
16	13	206	143.00	0.31	63.00
17	14	193	143.00	0.26	50.00
18	15	207	143.00	0.31	64.00
19	16	218	143.00	0.34	75.00
20	17	229	143.00	0.38	86.00
21	18	225	143.00	0.36	82.00
22	19	204	143.00	0.30	61.00
23	20	227	143.00	0.37	84.00
24	21	223	143.00	0.36	80.00
25	22	242	143.00	0.41	99.00
26	23	239	143.00	0.40	96.00
27	24	266	143.00	0.46	123.00
28			143.00		

Figure 15.225: **List of data of Exponential Smoothing in Microsoft® Excel® 14.0.6123**

We will not use the Analysis Toolpack of Microsoft® Excel® (the latter incorporating a tool doing automatically the calculations of the simple exponential smoothing) in the purpose to show the reader the explicit application of the previous proven mathematical relation. Furthermore, the Analysis Toolpack of Microsoft® Excel® 14.0.6123 (and previous releases) suffers from a major flaw: it requires a manual input of the value of the smoothing constant rather than calculating it from an optimisation research.

There are only three columns that have formulas, the remainder being static entries. We have:

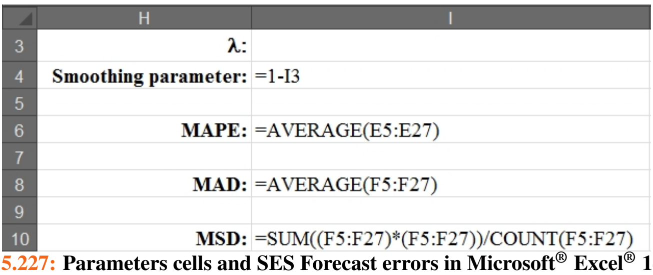


	A	B	C	D	E	F
3	**Period**	**Y_t**	**SES**	**Analysis Toolpack**	**Absolute Relative Deviation**	**Relative Deviation**
4	1	143	=B4	#N/A		
5	2	152	=$H$3*B4+(1-$H$3)*C4	=B4	=ABS(C5-B5)/B5	=ABS(B5-C5)
6	3	161	=$H$3*B5+(1-$H$3)*C5	=0.65*B5+0.35*D5	=ABS(C6-B6)/B6	=ABS(B6-C6)
7	4	139	=$H$3*B6+(1-$H$3)*C6	=0.65*B6+0.35*D6	=ABS(C7-B7)/B7	=ABS(B7-C7)
8	5	137	=$H$3*B7+(1-$H$3)*C7	=0.65*B7+0.35*D7	=ABS(C8-B8)/B8	=ABS(B8-C8)
9	6	174	=$H$3*B8+(1-$H$3)*C8	=0.65*B8+0.35*D8	=ABS(C9-B9)/B9	=ABS(B9-C9)
10	7	142	=$H$3*B9+(1-$H$3)*C9	=0.65*B9+0.35*D9	=ABS(C10-B10)/B10	=ABS(B10-C10)
11	8	141	=$H$3*B10+(1-$H$3)*C10	=0.65*B10+0.35*D10	=ABS(C11-B11)/B11	=ABS(B11-C11)
12	9	162	=$H$3*B11+(1-$H$3)*C11	=0.65*B11+0.35*D11	=ABS(C12-B12)/B12	=ABS(B12-C12)
13	10	180	=$H$3*B12+(1-$H$3)*C12	=0.65*B12+0.35*D12	=ABS(C13-B13)/B13	=ABS(B13-C13)
14	11	164	=$H$3*B13+(1-$H$3)*C13	=0.65*B13+0.35*D13	=ABS(C14-B14)/B14	=ABS(B14-C14)
15	12	171	=$H$3*B14+(1-$H$3)*C14	=0.65*B14+0.35*D14	=ABS(C15-B15)/B15	=ABS(B15-C15)
16	13	206	=$H$3*B15+(1-$H$3)*C15	=0.65*B15+0.35*D15	=ABS(C16-B16)/B16	=ABS(B16-C16)
17	14	193	=$H$3*B16+(1-$H$3)*C16	=0.65*B16+0.35*D16	=ABS(C17-B17)/B17	=ABS(B17-C17)
18	15	207	=$H$3*B17+(1-$H$3)*C17	=0.65*B17+0.35*D17	=ABS(C18-B18)/B18	=ABS(B18-C18)
19	16	218	=$H$3*B18+(1-$H$3)*C18	=0.65*B18+0.35*D18	=ABS(C19-B19)/B19	=ABS(B19-C19)
20	17	229	=$H$3*B19+(1-$H$3)*C19	=0.65*B19+0.35*D19	=ABS(C20-B20)/B20	=ABS(B20-C20)
21	18	225	=$H$3*B20+(1-$H$3)*C20	=0.65*B20+0.35*D20	=ABS(C21-B21)/B21	=ABS(B21-C21)
22	19	204	=$H$3*B21+(1-$H$3)*C21	=0.65*B21+0.35*D21	=ABS(C22-B22)/B22	=ABS(B22-C22)
23	20	227	=$H$3*B22+(1-$H$3)*C22	=0.65*B22+0.35*D22	=ABS(C23-B23)/B23	=ABS(B23-C23)
24	21	223	=$H$3*B23+(1-$H$3)*C23	=0.65*B23+0.35*D23	=ABS(C24-B24)/B24	=ABS(B24-C24)
25	22	242	=$H$3*B24+(1-$H$3)*C24	=0.65*B24+0.35*D24	=ABS(C25-B25)/B25	=ABS(B25-C25)
26	23	239	=$H$3*B25+(1-$H$3)*C25	=0.65*B25+0.35*D25	=ABS(C26-B26)/B26	=ABS(B26-C26)
27	24	266	=$H$3*B26+(1-$H$3)*C26	=0.65*B26+0.35*D26	=ABS(C27-B27)/B27	=ABS(B27-C27)
28			=$H$3*B27+(1-$H$3)*C27	=0.65*B27+0.35*D27		

Figure 15.226: **Explicit list of data of SES in Microsoft® Excel® 14.0.6123**

where the attentive reader will have noticed the forecast at the line 28 over a period of one month.

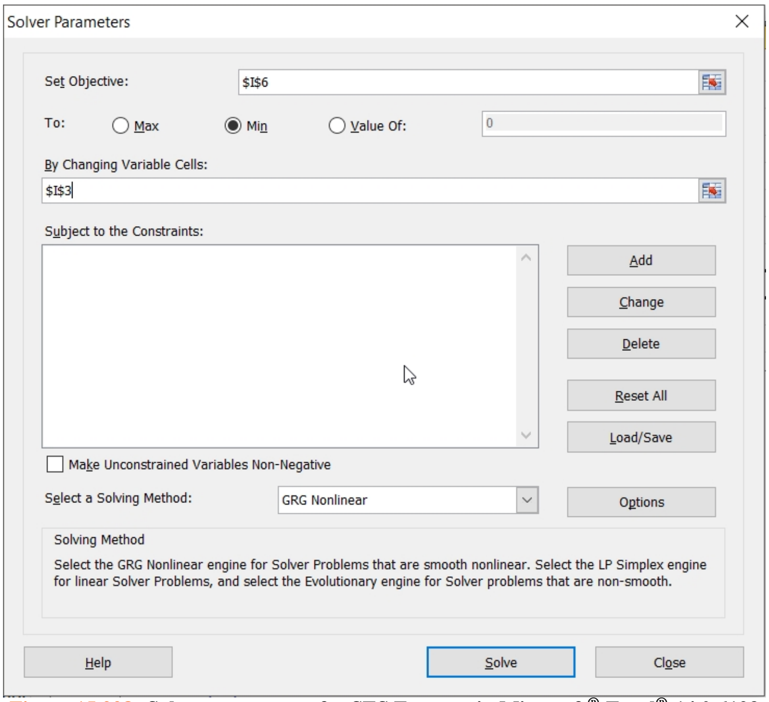
Little bit more far in the worksheet we have put values essential to the optimization of the model thanks to the solver that we will immediately see:



	H	I
3	**λ:**	
4	**Smoothing parameter:**	=1-I3
5		
6	**MAPE:**	=AVERAGE(E5:E27)
7		
8	**MAD:**	=AVERAGE(F5:F27)
9		
10	**MSD:**	=SUM((F5:F27)*(F5:F27))/COUNT(F5:F27)

Figure 15.227: **Parameters cells and SES Forecast errors in Microsoft® Excel® 14.0.6123**

Now we will use the solver to minimize either MAPE, MAD or MSD. We then get for example by minimizing the MAPE error:



Solver Parameters

Set Objective: $I$6

To: ○ Max ● Min ○ Value Of: 0

By Changing Variable Cells: $I$3

Subject to the Constraints:

☐ Make Unconstrained Variables Non-Negative

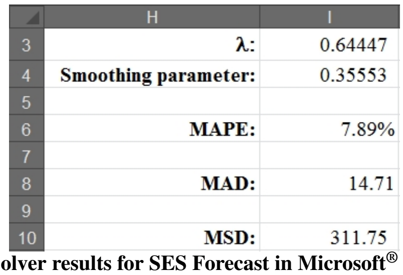
Select a Solving Method: GRG Nonlinear

Solving Method

Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

Figure 15.228: **Solver parameters for SES Forecast in Microsoft® Excel® 14.0.6123**

This gives the following result:



	H	I
3	**λ:**	0.64447
4	**Smoothing parameter:**	0.35553
5		
6	**MAPE:**	7.89%
7		
8	**MAD:**	14.71
9		
10	**MSD:**	311.75

Figure 15.229: **Solver results for SES Forecast in Microsoft® Excel® 14.0.6123**

and graphically:

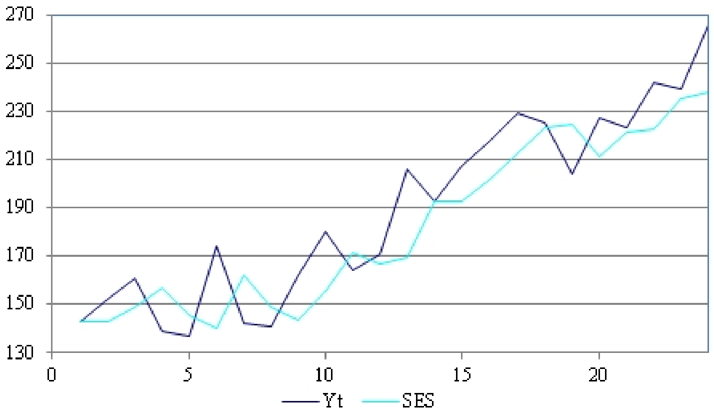
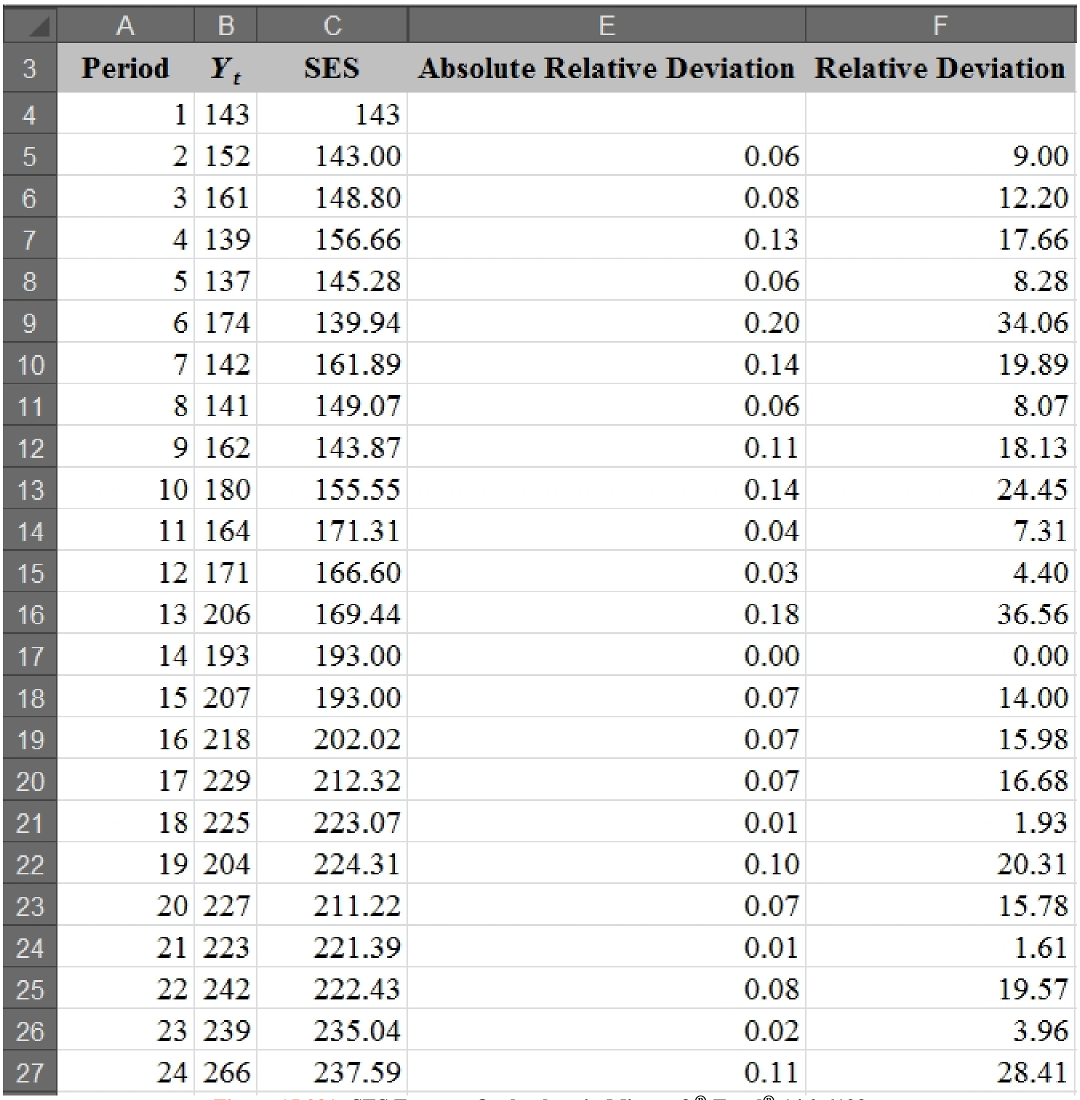


Figure 15.230: **SES Forecast plot in Microsoft® Excel® 14.0.6123**

and therefore the table of values becomes:



	A	B	C	E	F
3	**Period**	**Y_t**	**SES**	**Absolute Relative Deviation**	**Relative Deviation**
4	1	143	143		
5	2	152	143.00	0.06	9.00
6	3	161	148.80	0.08	12.20
7	4	139	156.66	0.13	17.66
8	5	137	145.28	0.06	8.28
9	6	174	139.94	0.20	34.06
10	7	142	161.89	0.14	19.89
11	8	141	149.07	0.06	8.07
12	9	162	143.87	0.11	18.13
13	10	180	155.55	0.14	24.45
14	11	164	171.31	0.04	7.31
15	12	171	166.60	0.03	4.40
16	13	206	169.44	0.18	36.56
17	14	193	193.00	0.00	0.00
18	15	207	193.00	0.07	14.00
19	16	218	202.02	0.07	15.98
20	17	229	212.32	0.07	16.68
21	18	225	223.07	0.01	1.93
22	19	204	224.31	0.10	20.31
23	20	227	211.22	0.07	15.78
24	21	223	221.39	0.01	1.61
25	22	242	222.43	0.08	19.57
26	23	239	235.04	0.02	3.96
27	24	266	237.59	0.11	28.41

Figure 15.231: SES Forecast final values in Microsoft® Excel® 14.0.6123

so we see that the forecast on line 28 of the table is:

$$F_{24+1}=255.90 \qquad (15.3.1496)$$

With for recall:

$$\text{MAPE}=100\frac{1}{n}\sum_{t=1}^{n}\frac{|e_t|}{Y_t}=7.89\% \quad \text{MAD}=\frac{1}{n}\sum_{t=1}^{n}|e_t|=14.71 \quad \text{MSD}=\frac{1}{n}\sum_{t=1}^{n}e_t^2=311.75 \qquad (15.3.1497)$$

To conclude, let us indicate that if we make the strong assumption that for any time t:

$$Y_{t+1}=a_t+\varepsilon_t=c^{te}+\mathcal{N}\left(0,\sigma_{Y-\bar{Y}}\right) \qquad (15.3.1498)$$

Then in the case of an infinite series (or supposed as such ...) with independent values:

$$\begin{aligned}\mathrm{V}(F_{t+1})&=\mathrm{V}\left((1-\beta)\sum_{i=0}^{t-1}\beta^i Y_{t-1}\right)=(1-\beta)^2\mathrm{V}\left(\sum_{i=0}^{t-1}\beta^i\left(c^{te}+\mathcal{N}\left(0,\sigma_{Y-\bar{Y}}\right)\right)\right)\\&=(1-\beta)^2\mathrm{V}\left(\sum_{i=0}^{t-1}\beta^i c^{te}\right)+(1-\beta)^2\mathrm{V}\left(\sum_{i=0}^{t-1}\beta^i\mathcal{N}\left(0,\sigma_{Y-\bar{Y}}\right)\right)\\&=(1-\beta)^2\sum_{i=0}^{t-1}\mathrm{V}\left(\beta^i c^{te}\right)+(1-\beta)^2\sum_{i=0}^{t-1}\mathrm{V}\left(\beta^i\mathcal{N}\left(0,\sigma_{Y-\bar{Y}}\right)\right)\\&=(1-\beta)^2\sum_{i=0}^{t-1}\beta^{2i}\mathrm{V}\left(c^{te}\right)+(1-\beta)^2\sum_{i=0}^{t-1}\beta^{2i}\mathrm{V}\left(\mathcal{N}\left(0,\sigma_{Y-\bar{Y}}\right)\right)\\&=(1-\beta)^2\sum_{i=0}^{t-1}\beta^{2i}\cdot 0+(1-\beta)^2\sum_{i=0}^{t-1}\beta^{2i}\mathrm{V}\left(\mathcal{N}\left(0,\sigma_{Y-\bar{Y}}\right)\right)\\&=(1-\beta)^2\sum_{i=0}^{t-1}\beta^{2i}\mathrm{V}\left(\mathcal{N}\left(0,\sigma_{Y-\bar{Y}}\right)\right)=(1-\beta)^2\sum_{i=0}^{t-1}\beta^{2i}\sigma_{Y-\bar{Y}}^2\\&=\sigma_{Y-\bar{Y}}^2(1-\beta)^2\sum_{i=0}^{t-1}\beta^{2i}=\sigma_{Y-\bar{Y}}^2(1-\beta)^2\frac{1-\beta^{2(t-1)+2}}{1-\beta^2}\\&=\sigma_{Y-\bar{Y}}^2(1-\beta)^2\frac{1-\beta^{2t}}{1-\beta^2}=\sigma_{Y-\bar{Y}}^2\frac{(1-\beta)(1-\beta^{2t})}{1+\beta}\end{aligned}\tag{15.3.1499}$$

However, as $\beta\in]0,1[$ and if t is large enough, we get:

$$\mathrm{V}(F_{t+1})=\sigma_{Y-\bar{Y}}^2\frac{(1-\beta)(1-\beta^{2t})}{1+\beta}\cong\sigma_{Y-\bar{Y}}^2\frac{1-\beta}{1+\beta}=\sigma_{Y-\bar{Y}}^2\frac{\lambda}{2-\lambda}\tag{15.3.1500}$$

and let us consider the error:

$$e(t)=Y_t-F_t\tag{15.3.1501}$$

We have under the assumption that the two subtracted terms are independent ...:

$$\begin{aligned}\mathrm{V}(e(t))&=\mathrm{V}(Y_t-F_t)=\mathrm{V}(Y_t)+\mathrm{V}(F_{t+1})\cong\mathrm{V}(Y-\bar{Y})+\mathrm{V}(F_{t+1})\\&=\sigma_{Y-\bar{Y}}^2+\sigma_{Y-\bar{Y}}^2\frac{\lambda}{2-\lambda}=\sigma_{Y-\bar{Y}}^2\left(1+\frac{\lambda}{2-\lambda}\right)=\frac{2\sigma_{Y-\bar{Y}}^2}{2-\lambda}\end{aligned}\tag{15.3.1502}$$

It follows under the assumption of Normality (which is the case since the beginning of this development) and that F_t is the mean of Y_t:

$$\frac{Y_t-F_t}{\sigma_{Y-\bar{Y}}\sqrt{\dfrac{2}{2-\lambda}}}\tag{15.3.1503}$$

and under the assumption that the variance is well known, we get:

$$Y_t=F_t\pm Z_{\alpha/2}\sigma_{Y-\bar{Y}}\sqrt{\frac{2}{2-\lambda}}\tag{15.3.1504}$$

But if the standard deviation is only estimated, it is obviously necessary to use the Student law (see section Statistics page 783):

$$Y_t = F_t \pm T_{\alpha/2}\hat{\sigma}_{Y-\bar{Y}}\sqrt{\frac{2}{2-\lambda}} \tag{15.3.1505}$$

And that will be all so far regarding to the simple exponential smoothing model.

15.3.13.3.3.2 Double Exponential Smoothing with One Parameter (Brown Method)

The idea of the "double exponential Brown smoothing" or also sometimes named "Holt's linear method" is again to reduce the total error but this time using not a constant but a straight linear the neighbourhood of t such that:

$$E = \sum_{j=0}^{t-1} \beta^j (Y_{t-j} - (a_t j + b_t))^2 \tag{15.3.1506}$$

what is written traditionally (the $-$ sign is absorbed into the parameter a and the fact that the parameters of the straight line are recalculated at each t is omitted in the way to not make any confusion with the index of summation...):

$$E = \sum_{j=0}^{t-1} \beta^j (Y_{t-j} - b + aj)^2 \tag{15.3.1507}$$

To minimize we differentiate the error E:

$$\begin{aligned}\frac{\partial E}{\partial b} &= -2\sum_{j=0}^{t-1} \beta^j (Y_{t-j} - b + aj) = 0\\ \frac{\partial E}{\partial a} &= 2\sum_{j=0}^{t-1} j\beta^j (Y_{t-j} - b + aj) = 0\end{aligned} \tag{15.3.1508}$$

Therefore:

$$\begin{aligned}&\sum_{j=0}^{t-1} \beta^j Y_{t-j} - b\sum_{j=0}^{t-1}\beta^j + a\sum_{j=0}^{t-1} j\beta^j = 0\\ &\sum_{j=0}^{t-1} j\beta^j Y_{t-j} - b\sum_{j=0}^{t-1} j\beta^j + a\sum_{j=0}^{t-1} j^2\beta^j = 0\end{aligned} \tag{15.3.1509}$$

Let us simplify all this by first replacing when t is large:

$$\sum_{j=0}^{t-1} b^j = \frac{1-\beta^t}{1-\beta} \cong \frac{1}{1-\beta} \tag{15.3.1510}$$

For the others, under certain conditions that are met in our case, we can derivate one term to one term a convergent series to get derivative thereof. Thus, given that:

$$\frac{\mathrm{d}}{\mathrm{d}\beta} = \sum_{j=0}^{t-1} \beta^j = \sum_{j=0}^{t-1} j\beta^{j-1} \tag{15.3.1511}$$

and as in the above conditions:

$$\sum_{j=0}^{t-1} b^j \cong \frac{1}{1-\beta} \tag{15.3.1512}$$

we then have immediately:

$$\sum_{j=0}^{t-1} j\beta^{j-1} \cong \frac{\mathrm{d}}{\mathrm{d}\beta}\left(\frac{1}{1-\beta}\right) = \frac{1}{(1-\beta)^2} \tag{15.3.1513}$$

Multiplying by β on the left and right we get to finally:

$$\sum_{j=0}^{t-1} j\beta^{j} \cong \frac{\beta}{(1-\beta)^2} \tag{15.3.1514}$$

which is the second desired simplification we were looking form.

For the third one we simply derivate once again the previous equality. We then get:

$$\sum_{j=0}^{t-1} j^2\beta^{j-1} \cong \frac{1}{(1-\beta)^2} + \frac{2\beta}{(1-\beta)^3} = \frac{(1-\beta)+2\beta}{(1-\beta)^3} = \frac{1+\beta}{(1-\beta)^3} \tag{15.3.1515}$$

And we end by multiplying left and right by β:

$$\sum_{j=0}^{t-1} j^2\beta^{j-1} \cong \frac{\beta(1+\beta)}{(1-\beta)^3} \tag{15.3.1516}$$

We then get:

$$\begin{gathered}
\sum_{j=0}^{t-1} \beta^j Y_{t-j} - b\sum_{j=0}^{t-1} \beta^j + a\sum_{j=0}^{t-1} j\beta^j \cong \sum_{j=0}^{t-1} \beta^j Y_{t-j} - \frac{b}{t-j} + \frac{a\beta}{(1-\beta)^2} = 0\\
\sum_{j=0}^{t-1} j\beta^j Y_{t-j} - b\sum_{j=0}^{t-1} j\beta^j + a\sum_{j=0}^{t-1} j^2\beta^j \cong \sum_{j=0}^{t-1} j\beta^j Y_{t-j} - \frac{b\beta}{(1-\beta)^2} + \frac{a\beta(1+\beta)}{(1-\beta)^3} = 0
\end{gathered} \tag{15.3.1517}$$

What we can write under the form:

$$\begin{gathered}
(1-\beta)\sum_{j=0}^{t-1} \beta^j Y_{t-j} - b + \frac{\beta}{1-\beta} = 0\\
(1-\beta)^2\sum_{j=0}^{t-1} j\beta^j Y_{t-j} - b\beta + a\frac{\beta(1+\beta)}{1-\beta} = 0
\end{gathered} \tag{15.3.1518}$$

Let us define now the smoothed series (the reader will notice that this is in fact the simple exponential smoothing):

$$S_1(t) = (1-\beta)\sum_{j=0}^{t-1} \beta^j Y_{t-j} = F_{t+1} \tag{15.3.1519}$$

and the double smooths series (hence the technical name of the method we are studying now):

$$\begin{aligned}S_2(t)&=(1-\beta)\sum_{j=0}^{t-1}\beta^j S_1(t-j)=(1-\beta)^2\sum_{j=0}^{t-1}\beta^j\sum_{i=0}^{t-j-1}\beta^i X_{t-j-i}\\&=(1-\beta)^2\sum_{j=0}^{t-1}\sum_{i=0}^{t-j-1}\beta^{i+j}X_{t-j-i}={F'}_{t+1}\end{aligned}\tag{15.3.1520}$$

In practice will just use:

$$\boxed{\begin{aligned}&S_1(t)=(1-\beta)\sum_{j=0}^{t-1}\beta^j Y_{t-j}\\&S_2(t)=(1-\beta)\sum_{j=0}^{t-1}\beta^j S_1(t-j)\end{aligned}}\tag{15.3.1521}$$

After the change of variable $i+j=k\Leftrightarrow -i-j=-k$ we get:

$$S_2(t)=(1-\beta)^2\sum_{j=0}^{t-1}\sum_{k=j}^{t-1}\beta^k X_{t-k}\tag{15.3.1522}$$

The index of the second sum is explained by the fact that k is the new indexation variable (so it must necessarily be in the index!). And when i is zero we have:

$$i+j=k\underset{i=0}{\Rightarrow}0+j=k\Leftrightarrow k=j\tag{15.3.1523}$$

For the suffix of the sum it is explained by the fact that for the maximum value we have $i=t-j-1$ and therefore as $i+j=k$ it comes if we substitute in it i:

$$(t-j-1)+j=k\Leftrightarrow t-1=k\tag{15.3.1524}$$

Then, the reader can check by itself... on paper... that we can rearrange the terms as follows:

$$S_2(t)=(1-\beta)^2\sum_{j=0}^{t-1}\sum_{k=j}^{t-1}\beta^k X_{t-k}=(1-\beta)^2\sum_{k=0}^{t-1}\sum_{j=0}^{k}\beta^k X_{t-k}\tag{15.3.1525}$$

Then we have:

$$\begin{aligned}S_2(t)&=(1-\beta)^2\sum_{k=0}^{t-1}\sum_{j=0}^{k}\beta^k X_{t-k}=(1-\beta)^2\sum_{k=0}^{t-1}(k+1)\beta^k X_{t-k}\\&=(1-\beta)^2\sum_{k=0}^{t-1}k\beta^k X_{t-k}+(1-\beta)^2\sum_{k=0}^{t-1}\beta^k X_{t-k}=(1-\beta)^2\sum_{k=0}^{t-1}k\beta^k X_{t-k}+(1-\beta)S_1(t)\end{aligned}\tag{15.3.1526}$$

Substituting these results in:

$$\begin{aligned}&(1-\beta)\sum_{j=0}^{t-1}\beta^j Y_{t-j}-b+\frac{\beta}{1-\beta}=0\\&(1-\beta)^2\sum_{j=0}^{t-1}j\beta^j Y_{t-j}-b\beta+a\frac{\beta(1+\beta)}{1-\beta}=0\end{aligned}\tag{15.3.1527}$$

we get:

$$\begin{aligned}
&S_1(t)-b+1\frac{\beta}{1-\beta}=0\\
&S_2(t)-(1-\beta)S_1(t)-b\beta+a\frac{\beta(1+\beta)}{1-\beta}=0
\end{aligned}\tag{15.3.1528}$$

Then it comes after successive simplifications:

$$\begin{aligned}
&S_1(t)-b-\frac{\beta}{1-\beta}\left[\frac{1-\beta}{\beta(1+\beta)}\left(S_2(t)-(1-\beta)S_1(t)-b\beta\right)\right]=0\\
&S_1(t)-b-\frac{1}{1+\beta}\left(S_2(t)-(1-\beta)S_1(t)-b\beta\right)=0\\
&b-\frac{1}{1+\beta}b\beta=S_1(t)-\frac{1}{1+\beta}\left(S_2(t)-(1-\beta)S_1(t)\right)\\
&(1+\beta)b-b\beta=(1+\beta)S_1(t)-\left(S_2(t)-(1-\beta)S_1(t)\right)\\
&b\left[(1+\beta)-\beta\right]=(1+\beta)S_1(t)-S_2(t)+(1-\beta)S_1(t)\\
&b=2S_1(t)-S_2(t)
\end{aligned}\tag{15.3.1529}$$

and therefore also after successive simplifications:

$$\begin{aligned}
&S_1(t)-b+a\frac{\beta}{1-\beta}=0\\
&S_1(t)-\left(2S_1(t)-S_2(t)\right)+a\frac{\beta}{1-\beta}=0\\
&-S_1(t)+S_2(t)+a\frac{\beta}{1-\beta}=0\\
&a=\frac{1-\beta}{\beta}\left(S_1(t)-S_2(t)\right)
\end{aligned}\tag{15.3.1530}$$

Results that it is customary to write:

$$\boxed{\begin{aligned}
&a=\frac{\lambda}{1-\lambda}\left(S_1(t)-S_2(t)\right)=\gamma\left(S_1(t)-S_2(t)\right)\\
&b=2S_1(t)-S_2(t)
\end{aligned}}\tag{15.3.1531}$$

Therefore:

$$\boxed{S_t=a_tj+b_t}\tag{15.3.1532}$$

is the straight line constituted by a double exponential smoothing that approximates at best (the estimator) at a time t the real values that minimizes the total error. Unlike the simple exponential smoothing, we can by now against making predictions (however it remains an abusive use of this mathematical result) thanks to the presence of the variable j.

Explicitly, this straight line will be written by changing a little bit the notations (do not forget that at the beginning we had implicitly injected a negative sign in the term a):

$$\boxed{a_tj+b_t=\frac{\lambda}{1-\lambda}\left({F'}_{t+1}-F_{t+1}\right)j+\left(2F_{t+1}-{F'}_{t+1}\right)}\tag{15.3.1533}$$

It comes then that our forecast will be written:

$$F''_{t-j}=\frac{\lambda}{1-\lambda}\left(F'_{t+1}-F_{t+1}\right)j+\left(2F_{t+1}-F'_{t+1}\right) \tag{15.3.1534}$$

And obviously, if we want F''_{t+1}, we will have to choose $j=-1$. Therefore:

$$\boxed{F''_{t+1}=\frac{\lambda}{1-\lambda}\left(F_{t+1}-F'_{t+1}\right)+\left(2F_{t+1}-F'_{t+1}\right)} \tag{15.3.1535}$$

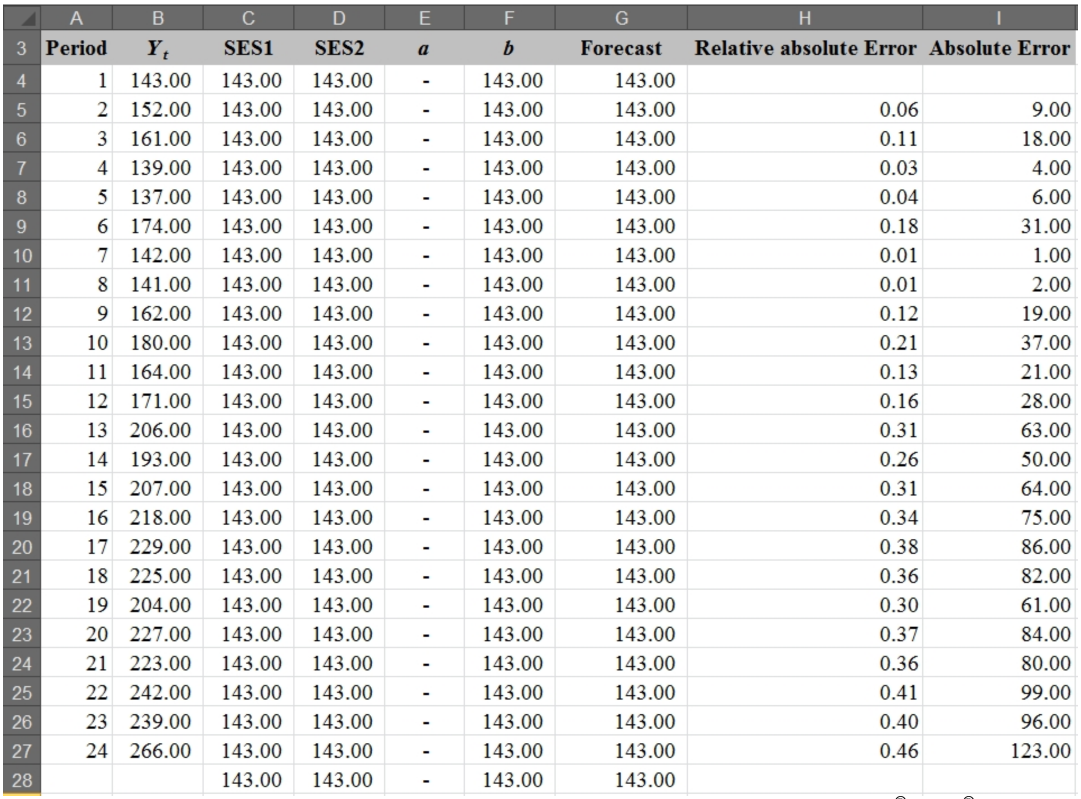
To start the forecast with this model, we need an estimate of $S_1(0)$ and $S_2(0)$. The first forecasts will be greatly affected by these estimates. The practitioner should also always check what his softwares use as method for the determination for these first parameters in the purpose to specify it in his reports!

Remark

Brown's exponential smoothing is rarely integrated in statistical softwares. We prefer to it the Holt's model that we will introduce immediately below and that contains the Brown model implicitly.

✍**Example:**

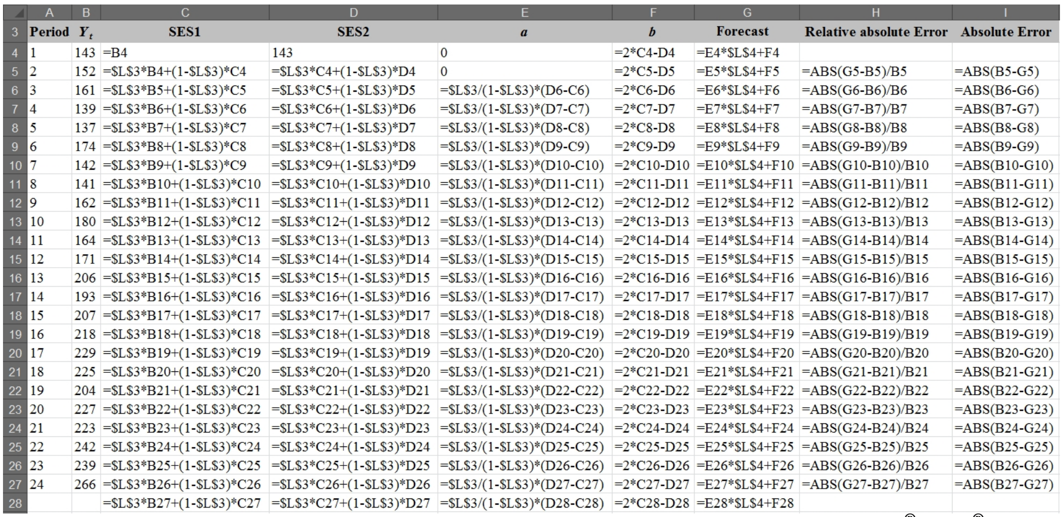
Let us consider the following list in Microsoft® Excel® 14.0.6123 that contains the observed data in column `B`, and the model in columns `C` through `F` for which we will detail the explicit formulas just after:



	A	B	C	D	E	F	G	H	I
3	**Period**	**Y_t**	**SES1**	**SES2**	***a***	***b***	**Forecast**	**Relative absolute Error**	**Absolute Error**
4	1	143.00	143.00	143.00	-	143.00	143.00		
5	2	152.00	143.00	143.00	-	143.00	143.00	0.06	9.00
6	3	161.00	143.00	143.00	-	143.00	143.00	0.11	18.00
7	4	139.00	143.00	143.00	-	143.00	143.00	0.03	4.00
8	5	137.00	143.00	143.00	-	143.00	143.00	0.04	6.00
9	6	174.00	143.00	143.00	-	143.00	143.00	0.18	31.00
10	7	142.00	143.00	143.00	-	143.00	143.00	0.01	1.00
11	8	141.00	143.00	143.00	-	143.00	143.00	0.01	2.00
12	9	162.00	143.00	143.00	-	143.00	143.00	0.12	19.00
13	10	180.00	143.00	143.00	-	143.00	143.00	0.21	37.00
14	11	164.00	143.00	143.00	-	143.00	143.00	0.13	21.00
15	12	171.00	143.00	143.00	-	143.00	143.00	0.16	28.00
16	13	206.00	143.00	143.00	-	143.00	143.00	0.31	63.00
17	14	193.00	143.00	143.00	-	143.00	143.00	0.26	50.00
18	15	207.00	143.00	143.00	-	143.00	143.00	0.31	64.00
19	16	218.00	143.00	143.00	-	143.00	143.00	0.34	75.00
20	17	229.00	143.00	143.00	-	143.00	143.00	0.38	86.00
21	18	225.00	143.00	143.00	-	143.00	143.00	0.36	82.00
22	19	204.00	143.00	143.00	-	143.00	143.00	0.30	61.00
23	20	227.00	143.00	143.00	-	143.00	143.00	0.37	84.00
24	21	223.00	143.00	143.00	-	143.00	143.00	0.36	80.00
25	22	242.00	143.00	143.00	-	143.00	143.00	0.41	99.00
26	23	239.00	143.00	143.00	-	143.00	143.00	0.40	96.00
27	24	266.00	143.00	143.00	-	143.00	143.00	0.46	123.00
28			143.00	143.00	-	143.00	143.00		

Figure 15.232: **Starting data for the Brown's double exponential smoothing model with Microsoft® Excel® 14.0.6123**

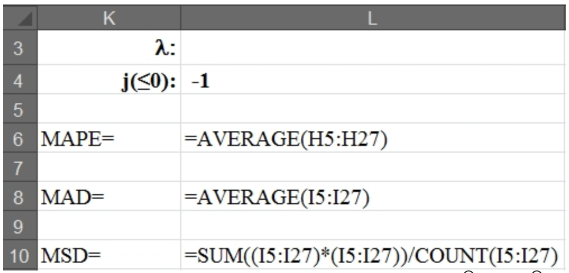
Let us give now the formulas for the columns `C`, `D`, `E`, `F`, `G` (`C` and `D` are the same formulas as for the simple exponential smoothing). It comes then:



	A	B	C	D	E	F	G	H	I
3	**Period**	**Y_t**	**SES1**	**SES2**	***a***	***b***	**Forecast**	**Relative absolute Error**	**Absolute Error**
4	1	143	=B4	143	0	=2*C4-D4	=E4*$L$4+F4		
5	2	152	=$L$3*B4+(1-$L$3)*C4	=$L$3*C4+(1-$L$3)*D4	0	=2*C5-D5	=E5*$L$4+F5	=ABS(G5-B5)/B5	=ABS(B5-G5)
6	3	161	=$L$3*B5+(1-$L$3)*C5	=$L$3*C5+(1-$L$3)*D5	=$L$3/(1-$L$3)*(D6-C6)	=2*C6-D6	=E6*$L$4+F6	=ABS(G6-B6)/B6	=ABS(B6-G6)
7	4	139	=$L$3*B6+(1-$L$3)*C6	=$L$3*C6+(1-$L$3)*D6	=$L$3/(1-$L$3)*(D7-C7)	=2*C7-D7	=E7*$L$4+F7	=ABS(G7-B7)/B7	=ABS(B7-G7)
8	5	137	=$L$3*B7+(1-$L$3)*C7	=$L$3*C7+(1-$L$3)*D7	=$L$3/(1-$L$3)*(D8-C8)	=2*C8-D8	=E8*$L$4+F8	=ABS(G8-B8)/B8	=ABS(B8-G8)
9	6	174	=$L$3*B8+(1-$L$3)*C8	=$L$3*C8+(1-$L$3)*D8	=$L$3/(1-$L$3)*(D9-C9)	=2*C9-D9	=E9*$L$4+F9	=ABS(G9-B9)/B9	=ABS(B9-G9)
10	7	142	=$L$3*B9+(1-$L$3)*C9	=$L$3*C9+(1-$L$3)*D9	=$L$3/(1-$L$3)*(D10-C10)	=2*C10-D10	=E10*$L$4+F10	=ABS(G10-B10)/B10	=ABS(B10-G10)
11	8	141	=$L$3*B10+(1-$L$3)*C10	=$L$3*C10+(1-$L$3)*D10	=$L$3/(1-$L$3)*(D11-C11)	=2*C11-D11	=E11*$L$4+F11	=ABS(G11-B11)/B11	=ABS(B11-G11)
12	9	162	=$L$3*B11+(1-$L$3)*C11	=$L$3*C11+(1-$L$3)*D11	=$L$3/(1-$L$3)*(D12-C12)	=2*C12-D12	=E12*$L$4+F12	=ABS(G12-B12)/B12	=ABS(B12-G12)
13	10	180	=$L$3*B12+(1-$L$3)*C12	=$L$3*C12+(1-$L$3)*D12	=$L$3/(1-$L$3)*(D13-C13)	=2*C13-D13	=E13*$L$4+F13	=ABS(G13-B13)/B13	=ABS(B13-G13)
14	11	164	=$L$3*B13+(1-$L$3)*C13	=$L$3*C13+(1-$L$3)*D13	=$L$3/(1-$L$3)*(D14-C14)	=2*C14-D14	=E14*$L$4+F14	=ABS(G14-B14)/B14	=ABS(B14-G14)
15	12	171	=$L$3*B14+(1-$L$3)*C14	=$L$3*C14+(1-$L$3)*D14	=$L$3/(1-$L$3)*(D15-C15)	=2*C15-D15	=E15*$L$4+F15	=ABS(G15-B15)/B15	=ABS(B15-G15)
16	13	206	=$L$3*B15+(1-$L$3)*C15	=$L$3*C15+(1-$L$3)*D15	=$L$3/(1-$L$3)*(D16-C16)	=2*C16-D16	=E16*$L$4+F16	=ABS(G16-B16)/B16	=ABS(B16-G16)
17	14	193	=$L$3*B16+(1-$L$3)*C16	=$L$3*C16+(1-$L$3)*D16	=$L$3/(1-$L$3)*(D17-C17)	=2*C17-D17	=E17*$L$4+F17	=ABS(G17-B17)/B17	=ABS(B17-G17)
18	15	207	=$L$3*B17+(1-$L$3)*C17	=$L$3*C17+(1-$L$3)*D17	=$L$3/(1-$L$3)*(D18-C18)	=2*C18-D18	=E18*$L$4+F18	=ABS(G18-B18)/B18	=ABS(B18-G18)
19	16	218	=$L$3*B18+(1-$L$3)*C18	=$L$3*C18+(1-$L$3)*D18	=$L$3/(1-$L$3)*(D19-C19)	=2*C19-D19	=E19*$L$4+F19	=ABS(G19-B19)/B19	=ABS(B19-G19)
20	17	229	=$L$3*B19+(1-$L$3)*C19	=$L$3*C19+(1-$L$3)*D19	=$L$3/(1-$L$3)*(D20-C20)	=2*C20-D20	=E20*$L$4+F20	=ABS(G20-B20)/B20	=ABS(B20-G20)
21	18	225	=$L$3*B20+(1-$L$3)*C20	=$L$3*C20+(1-$L$3)*D20	=$L$3/(1-$L$3)*(D21-C21)	=2*C21-D21	=E21*$L$4+F21	=ABS(G21-B21)/B21	=ABS(B21-G21)
22	19	204	=$L$3*B21+(1-$L$3)*C21	=$L$3*C21+(1-$L$3)*D21	=$L$3/(1-$L$3)*(D22-C22)	=2*C22-D22	=E22*$L$4+F22	=ABS(G22-B22)/B22	=ABS(B22-G22)
23	20	227	=$L$3*B22+(1-$L$3)*C22	=$L$3*C22+(1-$L$3)*D22	=$L$3/(1-$L$3)*(D23-C23)	=2*C23-D23	=E23*$L$4+F23	=ABS(G23-B23)/B23	=ABS(B23-G23)
24	21	223	=$L$3*B23+(1-$L$3)*C23	=$L$3*C23+(1-$L$3)*D23	=$L$3/(1-$L$3)*(D24-C24)	=2*C24-D24	=E24*$L$4+F24	=ABS(G24-B24)/B24	=ABS(B24-G24)
25	22	242	=$L$3*B24+(1-$L$3)*C24	=$L$3*C24+(1-$L$3)*D24	=$L$3/(1-$L$3)*(D25-C25)	=2*C25-D25	=E25*$L$4+F25	=ABS(G25-B25)/B25	=ABS(B25-G25)
26	23	239	=$L$3*B25+(1-$L$3)*C25	=$L$3*C25+(1-$L$3)*D25	=$L$3/(1-$L$3)*(D26-C26)	=2*C26-D26	=E26*$L$4+F26	=ABS(G26-B26)/B26	=ABS(B26-G26)
27	24	266	=$L$3*B26+(1-$L$3)*C26	=$L$3*C26+(1-$L$3)*D26	=$L$3/(1-$L$3)*(D27-C27)	=2*C27-D27	=E27*$L$4+F27	=ABS(G27-B27)/B27	=ABS(B27-G27)
28			=$L$3*B27+(1-$L$3)*C27	=$L$3*C27+(1-$L$3)*D27	=$L$3/(1-$L$3)*(D28-C28)	=2*C28-D28	=E28*$L$4+F28		

Figure 15.233: **Explicit formulas for the Brown's double exponential smoothing model with Microsoft® Excel® 14.0.6123**

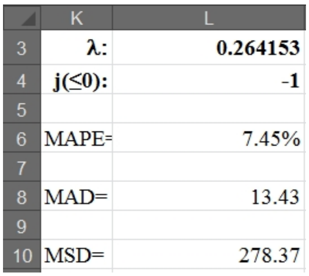
With a little bit more far in the worksheet the values essential to the optimization model (and standard forecast errors) with the solver that we will immediately see:



	K	L
3	**λ:**	
4	**j(≤0):**	**-1**
5		
6	MAPE=	=AVERAGE(H5:H27)
7		
8	MAD=	=AVERAGE(I5:I27)
9		
10	MSD=	=SUM((I5:I27)*(I5:I27))/COUNT(I5:I27)

Figure 15.234: **Initial double SES parameters in Microsoft® Excel® 14.0.6123**

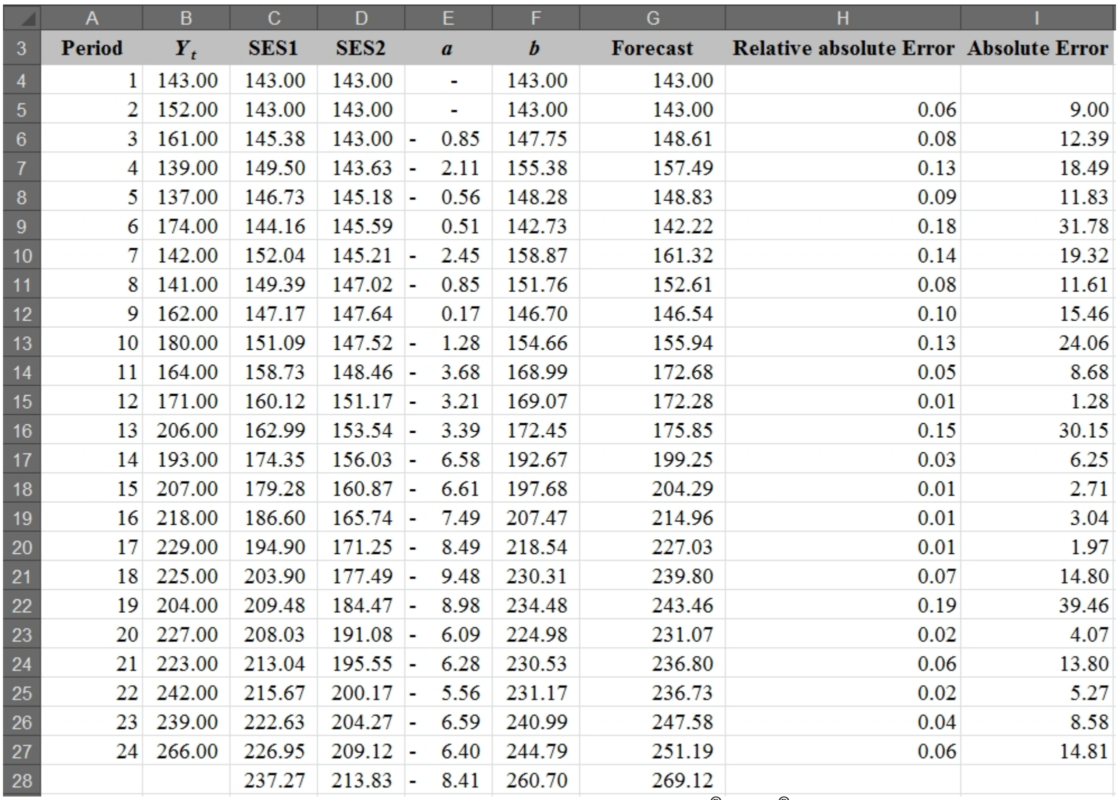
After optimization with the solver by minimizing MAPE we get:



	K	L
3	**λ:**	**0.264153**
4	**j(≤0):**	**-1**
5		
6	MAPE=	7.45%
7		
8	MAD=	13.43
9		
10	MSD=	278.37

Figure 15.235: **Final double SES parameters in Microsoft® Excel® 14.0.6123**

So finally the whole table gives:



	A	B	C	D	E	F	G	H	I
3	**Period**	Y_t	**SES1**	**SES2**	***a***	***b***	**Forecast**	**Relative absolute Error**	**Absolute Error**
4	1	143.00	143.00	143.00	-	143.00	143.00		
5	2	152.00	143.00	143.00	-	143.00	143.00	0.06	9.00
6	3	161.00	145.38	143.00	- 0.85	147.75	148.61	0.08	12.39
7	4	139.00	149.50	143.63	- 2.11	155.38	157.49	0.13	18.49
8	5	137.00	146.73	145.18	- 0.56	148.28	148.83	0.09	11.83
9	6	174.00	144.16	145.59	0.51	142.73	142.22	0.18	31.78
10	7	142.00	152.04	145.21	- 2.45	158.87	161.32	0.14	19.32
11	8	141.00	149.39	147.02	- 0.85	151.76	152.61	0.08	11.61
12	9	162.00	147.17	147.64	0.17	146.70	146.54	0.10	15.46
13	10	180.00	151.09	147.52	- 1.28	154.66	155.94	0.13	24.06
14	11	164.00	158.73	148.46	- 3.68	168.99	172.68	0.05	8.68
15	12	171.00	160.12	151.17	- 3.21	169.07	172.28	0.01	1.28
16	13	206.00	162.99	153.54	- 3.39	172.45	175.85	0.15	30.15
17	14	193.00	174.35	156.03	- 6.58	192.67	199.25	0.03	6.25
18	15	207.00	179.28	160.87	- 6.61	197.68	204.29	0.01	2.71
19	16	218.00	186.60	165.74	- 7.49	207.47	214.96	0.01	3.04
20	17	229.00	194.90	171.25	- 8.49	218.54	227.03	0.01	1.97
21	18	225.00	203.90	177.49	- 9.48	230.31	239.80	0.07	14.80
22	19	204.00	209.48	184.47	- 8.98	234.48	243.46	0.19	39.46
23	20	227.00	208.03	191.08	- 6.09	224.98	231.07	0.02	4.07
24	21	223.00	213.04	195.55	- 6.28	230.53	236.80	0.06	13.80
25	22	242.00	215.67	200.17	- 5.56	231.17	236.73	0.02	5.27
26	23	239.00	222.63	204.27	- 6.59	240.99	247.58	0.04	8.58
27	24	266.00	226.95	209.12	- 6.40	244.79	251.19	0.06	14.81
28			237.27	213.83	- 8.41	260.70	269.12		

Figure 15.236: **Final double SES values in Microsoft® Excel® 14.0.6123**

The result gives graphically (and also in comparison with the simple exponential smoothing) still for $j=-1$:

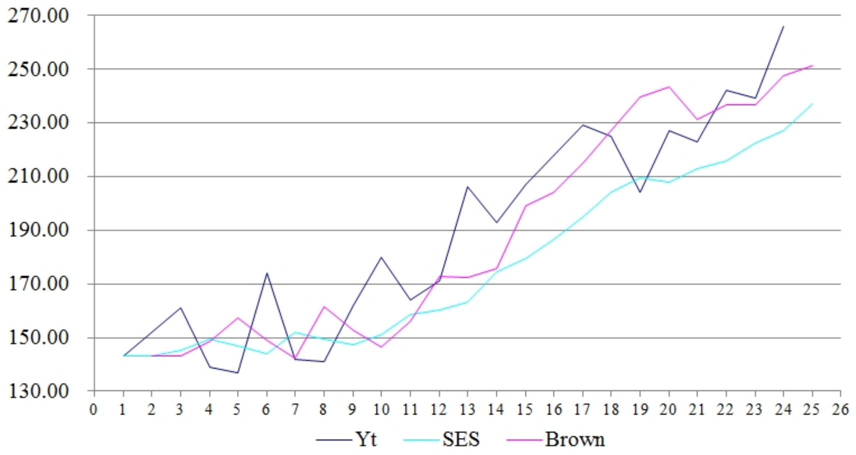
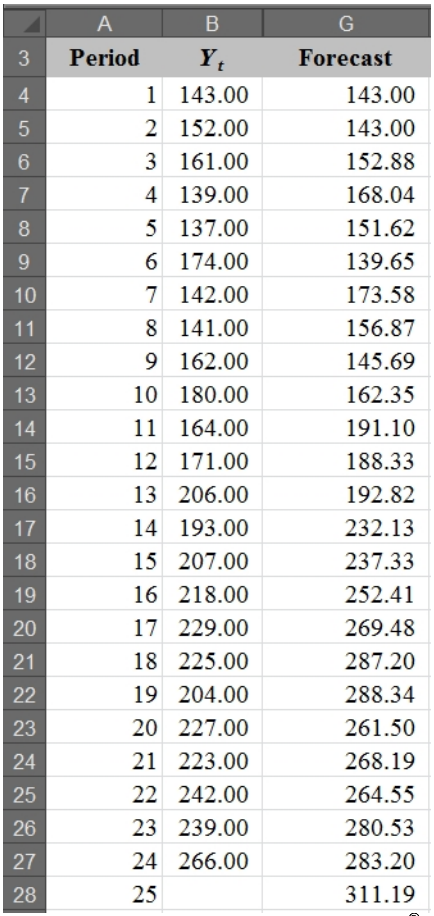


Figure 15.237: **Final double SES values in Microsoft® Excel® 14.0.6123**

What is interesting and special with the double exponential smoothing is to play with the value of j. Thus, by putting $j=-1$, we will have that for each value in column `G` its projection in the corresponding period plus $|j|$ period.

Let us make a little visual example. By putting $j=-6$ (thus a projection to 6 months ...) then we have:



	A	B	G
3	**Period**	**Y_t**	**Forecast**
4	1	143.00	143.00
5	2	152.00	143.00
6	3	161.00	152.88
7	4	139.00	168.04
8	5	137.00	151.62
9	6	174.00	139.65
10	7	142.00	173.58
11	8	141.00	156.87
12	9	162.00	145.69
13	10	180.00	162.35
14	11	164.00	191.10
15	12	171.00	188.33
16	13	206.00	192.82
17	14	193.00	232.13
18	15	207.00	237.33
19	16	218.00	252.41
20	17	229.00	269.48
21	18	225.00	287.20
22	19	204.00	288.34
23	20	227.00	261.50
24	21	223.00	268.19
25	22	242.00	264.55
26	23	239.00	280.53
27	24	266.00	283.20
28	25		311.19

Figure 15.238: **Projection $j=-6$ with the DSES in Microsoft® Excel® 14.0.6123**

Then, what is interesting is to remember is that if we put $j=-6$, the 20th period then projects for the 26th period, the 21th on the 27th, etc. We get therefore graphically:

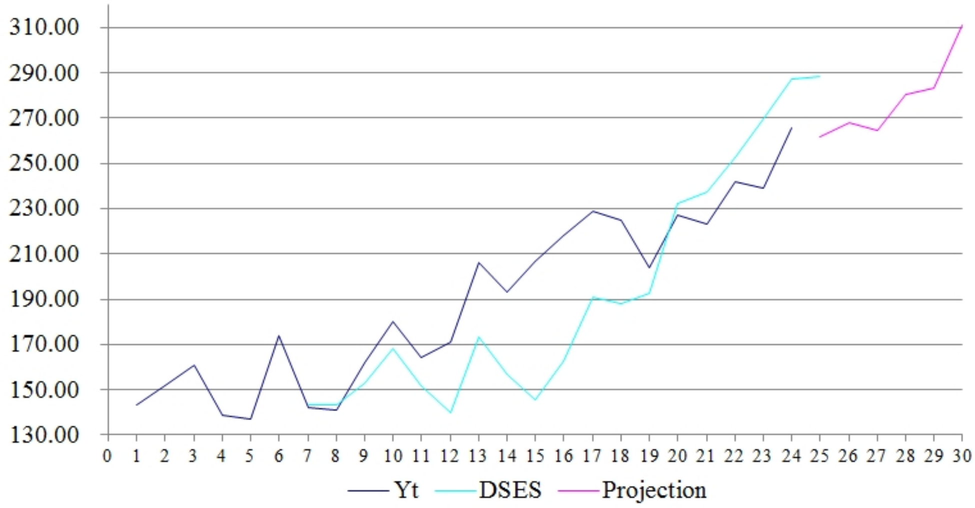


Figure 15.239: **Projection over a period of one semester using DEBS with Microsoft® Excel® 14.0.6123**

Obviously, like any model, the practitioner must be very careful with projections to 6 months (or less or more) because the model does not take into account the factors of the stock market, natural disasters, pandemics, wars, etc. So long as there is no statistical inference, it is not possible to give a confidence interval for the forecasting what is scientifically speaking not acceptable.

15.3.13.3.3.3 Holt's Double Exponential Smoothing with 2 Parameters (Additive Method)

We will discuss now a technique named "Holt's Double Exponential Smoothing with 2 Parameters (Additive Method)", whose construction is empirical and that is due to Charles C. Holt (11957 according to holocene calendar). Even if in the original article of C. Holt there is no rigorous proof of the construction of this model... so it is more mathematical engineering that pure mathematics itself.

The idea is as with the previous model, to have a model of the form (Holt's method extends simple exponential smoothing by assuming that the time series has both a level and a trend):

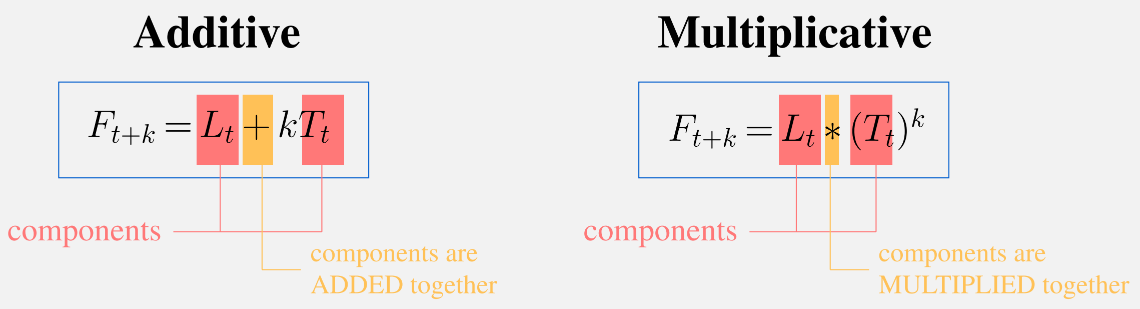
$$F_{t+j}=a_t j+b_t \tag{15.3.1536}$$

(also often denoted $F_{t+k}=L_t+kT_t$) but with an empirical approach that requires:

- The slope a_t of the straight line of model which origin of the frame is in t will be obtained by a simple exponential smoothing of all the slopes of the previous straight lines.
- The intercept b_t of the straight line on t will be obtained by a simple exponential smoothing all intercepts also of all the previous straight lines.

Remark

However, it is important to note that there are two types of time series (and hence of Holt's exponential smoohting), each with their own slightly different forecasting equation:



Let us first recall that we have obtain for the simple exponential smoothing:

$$S_t=(1-\beta)Y_t+\beta S_{t-1}=\lambda Y_t+(1-\lambda)S_{t-1} \tag{15.3.1537}$$

It is clear that the slope of each line of the estimate is given by (see figure below):

$$\frac{\Delta S_t}{\Delta t} \tag{15.3.1538}$$

and for the last estimate, more specific, we will have:

$$\frac{\Delta S_t}{\Delta t}=\frac{S_t-S_{t-1}}{t-(t-1)}=\frac{S_t-S_{t-1}}{1}=S_t-S_{t-1} \tag{15.3.1539}$$

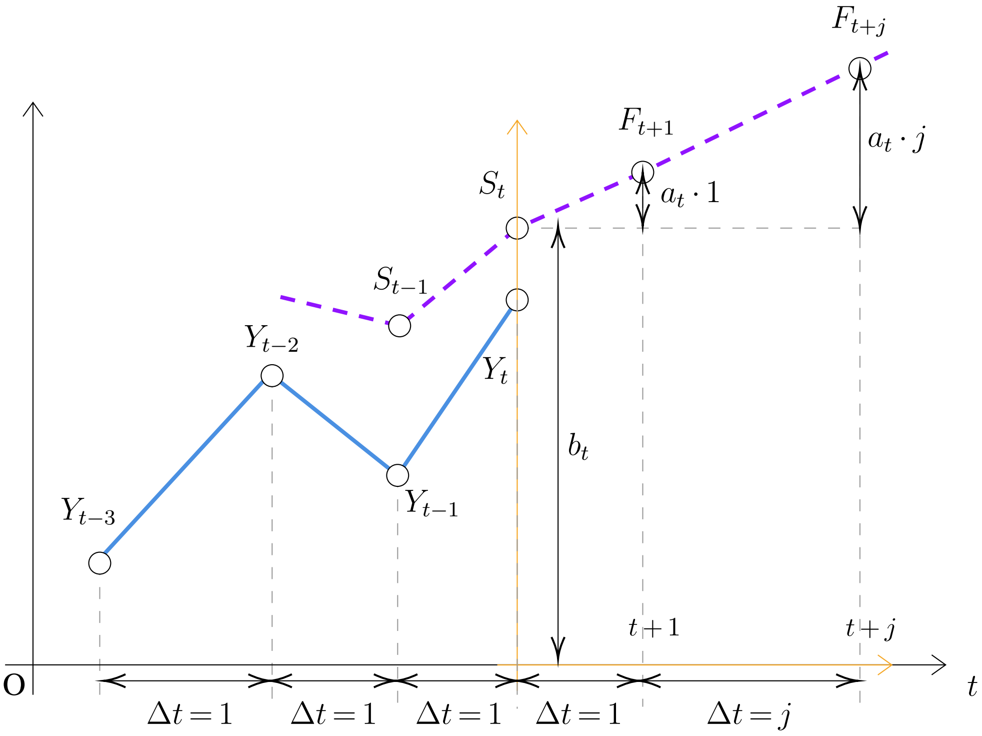


Figure 15.240: Idea diagram of the building blocs of Holt's model

and the idea is to make an exponential smoothing of the slope of all straight lines whose slope is calculated in the same manner, that is (compared to the expression of simple exponential smoothing)

$$a_t = \beta \frac{\Delta S_t}{\Delta t} + (1-\beta)a_{t-1} = \beta(S_t - S_{t-1}) + (1-\beta)a_{t-1} \qquad (15.3.1540)$$

The latter expression is just the estimated slope at time t.

Now, for the intercept, we see in the figure that $b_t = S_t$. Therefore, it comes at first as:

$$a_t = \beta \frac{\Delta S_t}{\Delta t} + (1-\beta)a_{t-1} = \beta(b_t - b_{t-1}) + (1-\beta)a_{t-1} \qquad (15.3.1541)$$

The relation:

$$a_t = \beta(b_t - b_{t-1}) + (1-\beta)a_{t-1} \qquad (15.3.1542)$$

is known as the "trend update equation", as it updates the trend estimate of the current time step based on the difference between previous level estimates. One may notice it is very akin to the original "level update equation" L_t in simple exponential smoothing.

Similarly to the level update equation, the trend update equation also takes its own parameter, β. The beta is the trend equation's own smoothing constant, and is therefore more or less analogous to the α smoothing constant seen earlier in simple exponential smoothing.

Once this fact seen, we choose a model where the intercept is an exponential smoothing of all intercepts with his own smoothing constant such as:

$$b_t = S_t = \alpha Y_t + (1-\alpha)(S_{t-1} + a_{t-1}) = \alpha Y_t + (1-\alpha)(b_{t-1} + a_{t-1}) \qquad (15.3.1543)$$

What may seem sometimes confusing here is the last parenthesis. In fact, technically, we should rather write:

$$(S_{t-1}+a_{t-1}\Delta t)=(b_{t-1}+a_{t-1}\Delta t) \tag{15.3.1544}$$

but as the Δt is always equal to 1, we omit to write it. So the last parenthesis represents the estimated at a one previous time unit to which we add the increase of one time unit of the estimated of slope as estimated at the same time t.

The expression:

$$S_t=\alpha Y_t+(1-\alpha)(b_{t-1}+a_{t-1}) \tag{15.3.1545}$$

is somewhat unfortunately named in practice the "estimated level series at period t".

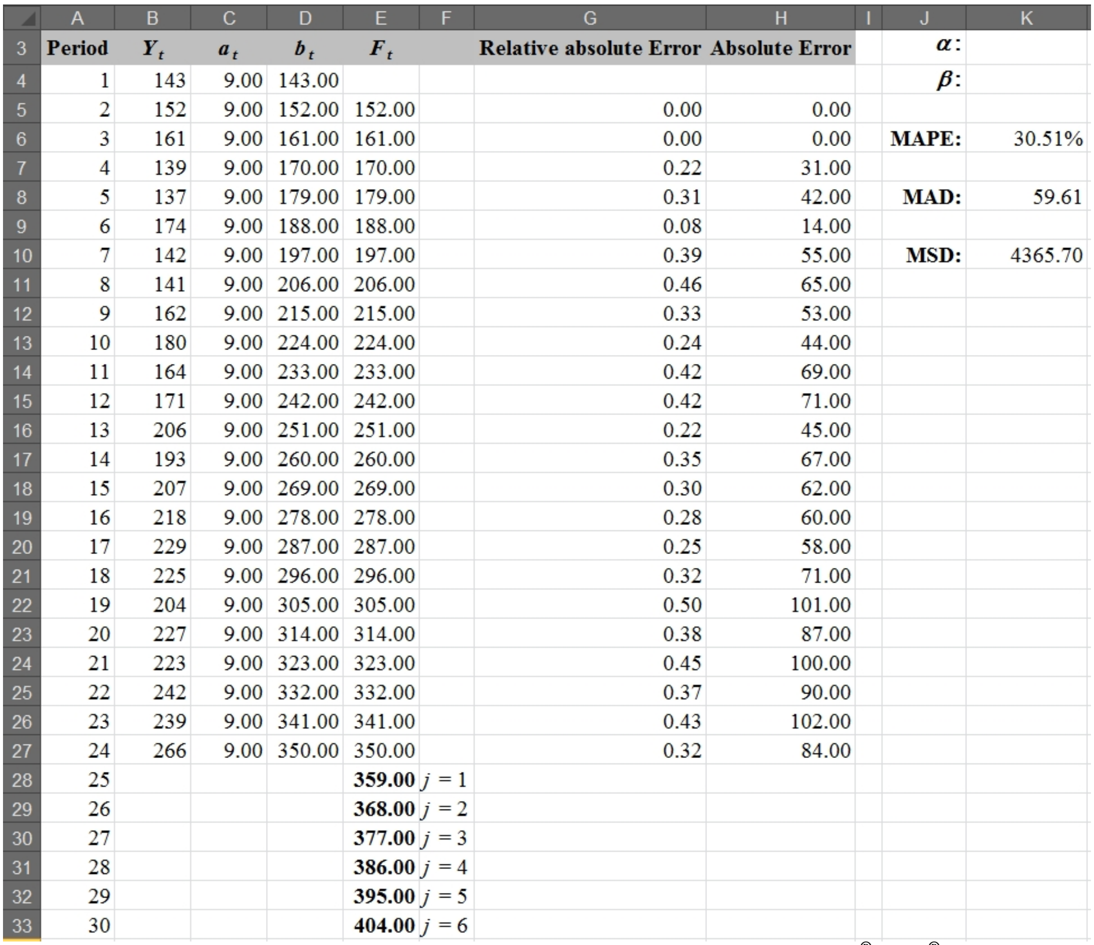
In the end, we work with these three relations:

$$\boxed{\begin{aligned}F_{t+j}&=a_t j+b_t\\ a_t&=\beta(b_t-b_{t-1})+(1-\beta)a_{t-1}\\ b_t&=\alpha Y_t+(1-\alpha)(b_{t-1}+a_{t-1})\end{aligned}} \tag{15.3.1546}$$

To start the forecast with this model, we need an estimate of the settings S_0 and a_0. The first forecast values will be greatly affected by these estimates. The practitioner should also always check what his software uses for a method of determination for these settings to specify it in his reports.

✍**Example:**

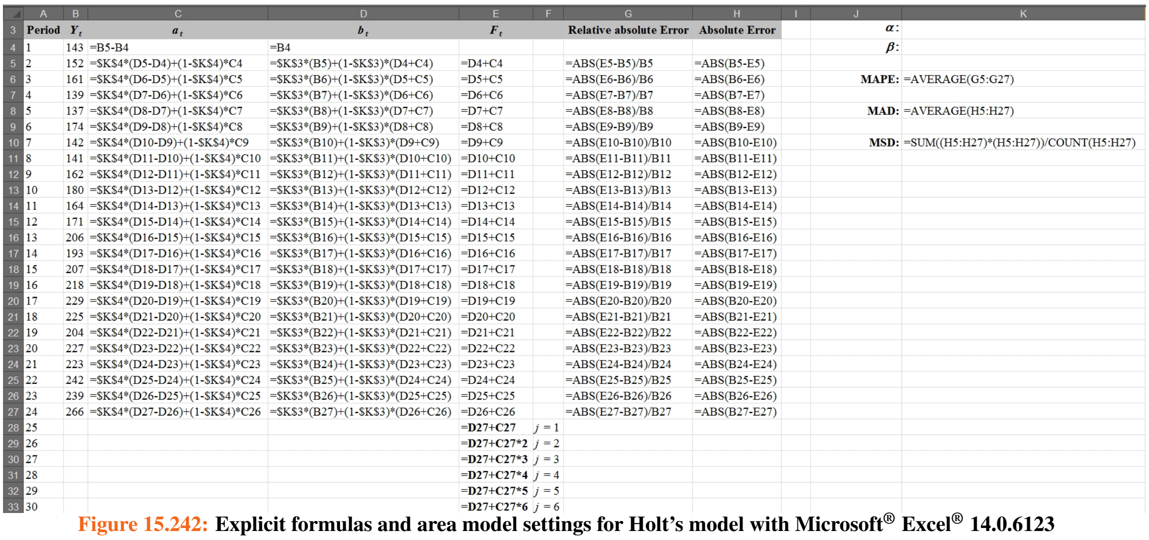
Let us consider the following list and model settings area in Microsoft® Excel® 14.0.6123 that contains the observed data in column `B`, and the model in columns `C` through `E` for which we will detail the formulas just after:



	A	B	C	D	E	F	G	H	I	J	K
3	**Period**	Y_t	a_t	b_t	F_t		**Relative absolute Error**	**Absolute Error**		α:	
4	1	143	9.00	143.00						β:	
5	2	152	9.00	152.00	152.00		0.00	0.00			
6	3	161	9.00	161.00	161.00		0.00	0.00		**MAPE:**	30.51%
7	4	139	9.00	170.00	170.00		0.22	31.00			
8	5	137	9.00	179.00	179.00		0.31	42.00		**MAD:**	59.61
9	6	174	9.00	188.00	188.00		0.08	14.00			
10	7	142	9.00	197.00	197.00		0.39	55.00		**MSD:**	4365.70
11	8	141	9.00	206.00	206.00		0.46	65.00			
12	9	162	9.00	215.00	215.00		0.33	53.00			
13	10	180	9.00	224.00	224.00		0.24	44.00			
14	11	164	9.00	233.00	233.00		0.42	69.00			
15	12	171	9.00	242.00	242.00		0.42	71.00			
16	13	206	9.00	251.00	251.00		0.22	45.00			
17	14	193	9.00	260.00	260.00		0.35	67.00			
18	15	207	9.00	269.00	269.00		0.30	62.00			
19	16	218	9.00	278.00	278.00		0.28	60.00			
20	17	229	9.00	287.00	287.00		0.25	58.00			
21	18	225	9.00	296.00	296.00		0.32	71.00			
22	19	204	9.00	305.00	305.00		0.50	101.00			
23	20	227	9.00	314.00	314.00		0.38	87.00			
24	21	223	9.00	323.00	323.00		0.45	100.00			
25	22	242	9.00	332.00	332.00		0.37	90.00			
26	23	239	9.00	341.00	341.00		0.43	102.00			
27	24	266	9.00	350.00	350.00		0.32	84.00			
28	25				**359.00**	$j=1$					
29	26				**368.00**	$j=2$					
30	27				**377.00**	$j=3$					
31	28				**386.00**	$j=4$					
32	29				**395.00**	$j=5$					
33	30				**404.00**	$j=6$					

Figure 15.241: **Initial values list and area model settings for Holt's model with Microsoft® Excel® 14.0.6123**

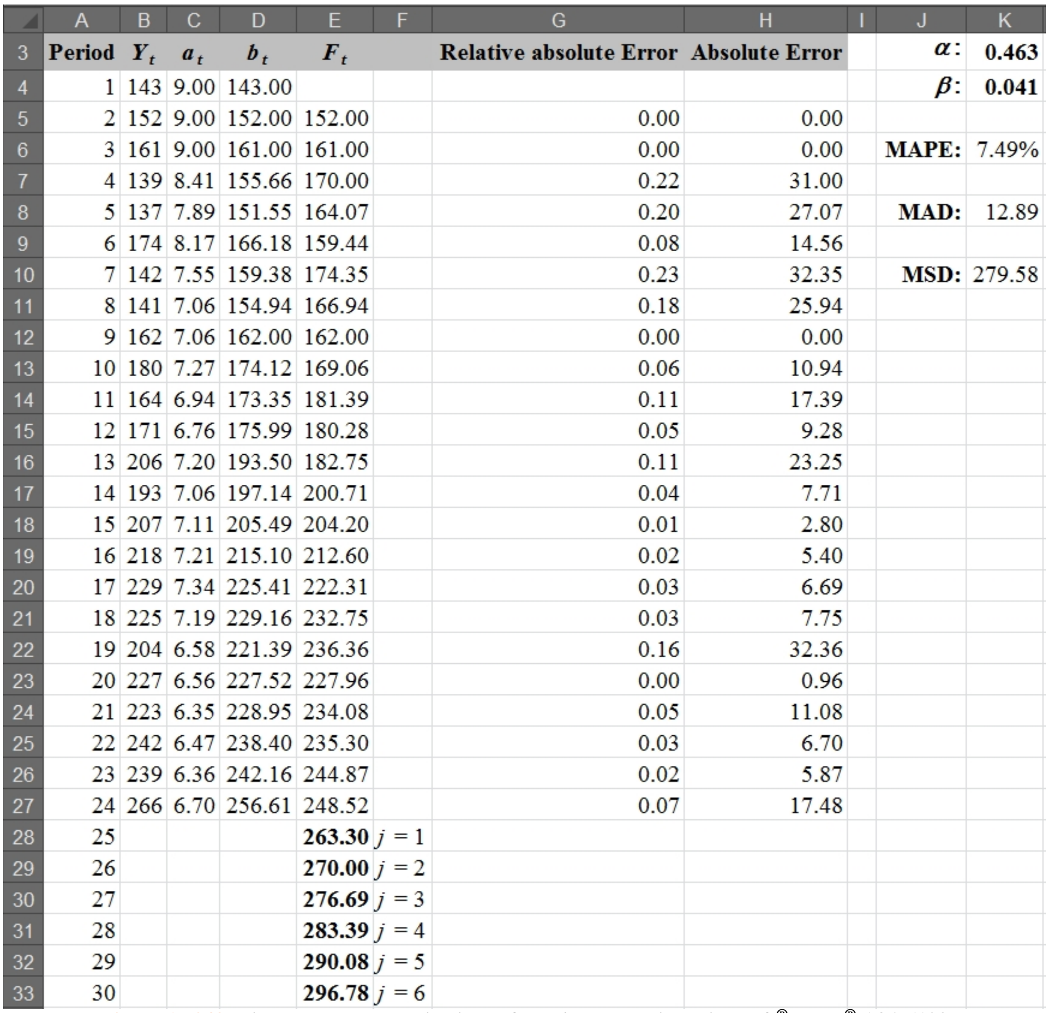
Let us indicate explicitly all formulas:



	A	B	C	D	E	F	G	H	I	J	K
3	**Period**	Y_t	a_t	b_t	F_t		**Relative absolute Error**	**Absolute Error**		α:	
4	1	143	=B5-B4	=B4						β:	
5	2	152	=\$K\$4*(D5-D4)+(1-\$K\$4)*C4	=\$K\$3*(B5)+(1-\$K\$3)*(D4+C4)	=D4+C4		=ABS(E5-B5)/B5	=ABS(B5-E5)			
6	3	161	=\$K\$4*(D6-D5)+(1-\$K\$4)*C5	=\$K\$3*(B6)+(1-\$K\$3)*(D5+C5)	=D5+C5		=ABS(E6-B6)/B6	=ABS(B6-E6)		**MAPE:**	=AVERAGE(G5:G27)
7	4	139	=\$K\$4*(D7-D6)+(1-\$K\$4)*C6	=\$K\$3*(B7)+(1-\$K\$3)*(D6+C6)	=D6+C6		=ABS(E7-B7)/B7	=ABS(B7-E7)			
8	5	137	=\$K\$4*(D8-D7)+(1-\$K\$4)*C7	=\$K\$3*(B8)+(1-\$K\$3)*(D7+C7)	=D7+C7		=ABS(E8-B8)/B8	=ABS(B8-E8)		**MAD:**	=AVERAGE(H5:H27)
9	6	174	=\$K\$4*(D9-D8)+(1-\$K\$4)*C8	=\$K\$3*(B9)+(1-\$K\$3)*(D8+C8)	=D8+C8		=ABS(E9-B9)/B9	=ABS(B9-E9)			
10	7	142	=\$K\$4*(D10-D9)+(1-\$K\$4)*C9	=\$K\$3*(B10)+(1-\$K\$3)*(D9+C9)	=D9+C9		=ABS(E10-B10)/B10	=ABS(B10-E10)		**MSD:**	=SUM((H5:H27)*(H5:H27))/COUNT(H5:H27)
11	8	141	=\$K\$4*(D11-D10)+(1-\$K\$4)*C10	=\$K\$3*(B11)+(1-\$K\$3)*(D10+C10)	=D10+C10		=ABS(E11-B11)/B11	=ABS(B11-E11)			
12	9	162	=\$K\$4*(D12-D11)+(1-\$K\$4)*C11	=\$K\$3*(B12)+(1-\$K\$3)*(D11+C11)	=D11+C11		=ABS(E12-B12)/B12	=ABS(B12-E12)			
13	10	180	=\$K\$4*(D13-D12)+(1-\$K\$4)*C12	=\$K\$3*(B13)+(1-\$K\$3)*(D12+C12)	=D12+C12		=ABS(E13-B13)/B13	=ABS(B13-E13)			
14	11	164	=\$K\$4*(D14-D13)+(1-\$K\$4)*C13	=\$K\$3*(B14)+(1-\$K\$3)*(D13+C13)	=D13+C13		=ABS(E14-B14)/B14	=ABS(B14-E14)			
15	12	171	=\$K\$4*(D15-D14)+(1-\$K\$4)*C14	=\$K\$3*(B15)+(1-\$K\$3)*(D14+C14)	=D14+C14		=ABS(E15-B15)/B15	=ABS(B15-E15)			
16	13	206	=\$K\$4*(D16-D15)+(1-\$K\$4)*C15	=\$K\$3*(B16)+(1-\$K\$3)*(D15+C15)	=D15+C15		=ABS(E16-B16)/B16	=ABS(B16-E16)			
17	14	193	=\$K\$4*(D17-D16)+(1-\$K\$4)*C16	=\$K\$3*(B17)+(1-\$K\$3)*(D16+C16)	=D16+C16		=ABS(E17-B17)/B17	=ABS(B17-E17)			
18	15	207	=\$K\$4*(D18-D17)+(1-\$K\$4)*C17	=\$K\$3*(B18)+(1-\$K\$3)*(D17+C17)	=D17+C17		=ABS(E18-B18)/B18	=ABS(B18-E18)			
19	16	218	=\$K\$4*(D19-D18)+(1-\$K\$4)*C18	=\$K\$3*(B19)+(1-\$K\$3)*(D18+C18)	=D18+C18		=ABS(E19-B19)/B19	=ABS(B19-E19)			
20	17	229	=\$K\$4*(D20-D19)+(1-\$K\$4)*C19	=\$K\$3*(B20)+(1-\$K\$3)*(D19+C19)	=D19+C19		=ABS(E20-B20)/B20	=ABS(B20-E20)			
21	18	225	=\$K\$4*(D21-D20)+(1-\$K\$4)*C20	=\$K\$3*(B21)+(1-\$K\$3)*(D20+C20)	=D20+C20		=ABS(E21-B21)/B21	=ABS(B21-E21)			
22	19	204	=\$K\$4*(D22-D21)+(1-\$K\$4)*C21	=\$K\$3*(B22)+(1-\$K\$3)*(D21+C21)	=D21+C21		=ABS(E22-B22)/B22	=ABS(B22-E22)			
23	20	227	=\$K\$4*(D23-D22)+(1-\$K\$4)*C22	=\$K\$3*(B23)+(1-\$K\$3)*(D22+C22)	=D22+C22		=ABS(E23-B23)/B23	=ABS(B23-E23)			
24	21	223	=\$K\$4*(D24-D23)+(1-\$K\$4)*C23	=\$K\$3*(B24)+(1-\$K\$3)*(D23+C23)	=D23+C23		=ABS(E24-B24)/B24	=ABS(B24-E24)			
25	22	242	=\$K\$4*(D25-D24)+(1-\$K\$4)*C24	=\$K\$3*(B25)+(1-\$K\$3)*(D24+C24)	=D24+C24		=ABS(E25-B25)/B25	=ABS(B25-E25)			
26	23	239	=\$K\$4*(D26-D25)+(1-\$K\$4)*C25	=\$K\$3*(B26)+(1-\$K\$3)*(D25+C25)	=D25+C25		=ABS(E26-B26)/B26	=ABS(B26-E26)			
27	24	266	=\$K\$4*(D27-D26)+(1-\$K\$4)*C26	=\$K\$3*(B27)+(1-\$K\$3)*(D26+C26)	=D26+C26		=ABS(E27-B27)/B27	=ABS(B27-E27)			
28	25				**=D27+C27**	$j=1$					
29	26				**=D27+C27*2**	$j=2$					
30	27				**=D27+C27*3**	$j=3$					
31	28				**=D27+C27*4**	$j=4$					
32	29				**=D27+C27*5**	$j=5$					
33	30				**=D27+C27*6**	$j=6$					

Figure 15.242: **Explicit formulas and area model settings for Holt's model with Microsoft® Excel® 14.0.6123**

As for all the models seen previously, we start the solver with the idea to minimize the MAPE:



	A	B	C	D	E	F	G	H	I	J	K
3	**Period**	**Y_t**	**a_t**	**b_t**	**F_t**		**Relative absolute Error**	**Absolute Error**		**α:**	**0.463**
4	1	143	9.00	143.00						**β:**	**0.041**
5	2	152	9.00	152.00	152.00		0.00	0.00			
6	3	161	9.00	161.00	161.00		0.00	0.00		**MAPE:**	7.49%
7	4	139	8.41	155.66	170.00		0.22	31.00			
8	5	137	7.89	151.55	164.07		0.20	27.07		**MAD:**	12.89
9	6	174	8.17	166.18	159.44		0.08	14.56			
10	7	142	7.55	159.38	174.35		0.23	32.35		**MSD:**	279.58
11	8	141	7.06	154.94	166.94		0.18	25.94			
12	9	162	7.06	162.00	162.00		0.00	0.00			
13	10	180	7.27	174.12	169.06		0.06	10.94			
14	11	164	6.94	173.35	181.39		0.11	17.39			
15	12	171	6.76	175.99	180.28		0.05	9.28			
16	13	206	7.20	193.50	182.75		0.11	23.25			
17	14	193	7.06	197.14	200.71		0.04	7.71			
18	15	207	7.11	205.49	204.20		0.01	2.80			
19	16	218	7.21	215.10	212.60		0.02	5.40			
20	17	229	7.34	225.41	222.31		0.03	6.69			
21	18	225	7.19	229.16	232.75		0.03	7.75			
22	19	204	6.58	221.39	236.36		0.16	32.36			
23	20	227	6.56	227.52	227.96		0.00	0.96			
24	21	223	6.35	228.95	234.08		0.05	11.08			
25	22	242	6.47	238.40	235.30		0.03	6.70			
26	23	239	6.36	242.16	244.87		0.02	5.87			
27	24	266	6.70	256.61	248.52		0.07	17.48			
28	25				**263.30**	$j=1$					
29	26				**270.00**	$j=2$					
30	27				**276.69**	$j=3$					
31	28				**283.39**	$j=4$					
32	29				**290.08**	$j=5$					
33	30				**296.78**	$j=6$					

Figure 15.243: **Final values and projections of Holt's model with Microsoft® Excel® 14.0.6123**

Which gives graphically (and also in comparison with the single and double smoothing):

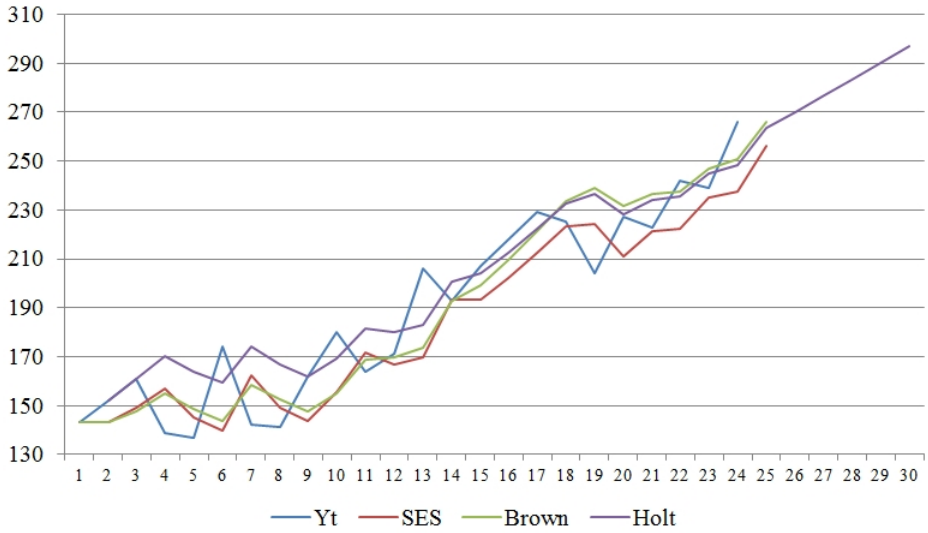


Figure 15.244: **Full Holt's model plot with Microsoft® Excel® 14.0.6123**

Holt's smoothing has a MAPE of 7.49% while the Brown model has for recall a MAPE of 7.45% (this comparison is a form of "backtesting"). However the advantage of Holt's model relatively to previous models is that we can make projections on more than one period.

It is of course interesting to compare the projections to 6 months between the additive method of Holt and Brown's smoothing:

Month	Brown's DSES	Holt's additive DSES
25	261.50	263.30
26	268.19	270.00
27	264.55	276.69
28	280.53	283.39
29	238.20	290.08
30	311.19	296.78

Table 15.60: **Comparison between Holt's and Brown's DSES**

So we see that Holt's model almost always have forecasting values greater than the Brown model.

It is the job of the practitioner to judge these values and compare them to past observations and also to future one...

15.3.13.3.3.4 Holt's and Winter Triple Exponential Smoothing with 3 Parameters (Multiplicative Method)

We will discuss now another technique which construction is also as empirical as the previous model and also strongly inspired from that latter and who is named "Holt's and Winter Triple Exponential Smoothing with 3 Parameters (Multiplicative Method)". I never found a rigorous proof of this model. It is therefore probably more mathematical engineering that pure mathematics strictly speaking, but as it is a classic in statistical softwares, we will focus on it a little bit.

Peter Winters (Charles C. Holt's student) has based it's work in the years 11960 (holocene calendar) on the method of his teacher to define a model that takes into account the seasonal component. The idea is to use three smoothing equations: one for the level of demand (without seasonality), one for the trend and one for seasonality.

Remark

There are two methods Winter's method according to that seasonality is additive or multiplicative. We present here only the multiplicative model because it is the one most commonly used. And we can not present all the existing models as they are all empirical and there are probably hundreds of them!

Let N be the number of periods in each seasonal cycle (cycle assumed to be constant!). Three exponential smoothing equations are used in each period to update the estimates of series (the part without seasonality), seasonal factors and trend. These equations have different smoothing constants traditionally denoted α, β and γ:

1. The first equation is defined quite naturally by:

   $$b_t = \alpha \frac{Y_t}{C_{t-N}} + (1-\alpha)(b_{t-1} + a_{t-1}) \qquad (15.3.1547)$$

   where Y_t is then the current level of the series without the seasonality effect (then we speak sometimes of "adjusted series") thanks to the appropriate seasonal factor C_{t-N} (otherwise we fall back on the Holt's method when $C_{t-N} = 1$).

2. The second equation is naturally defined by:

   $$a_t = \beta(b_t - b_{t-1}) + (1-\beta)a_{t-1} \qquad (15.3.1548)$$

   where we see again the slope of the trend identical to Holt's model.

3. The third equation is completely new, this does not prevent that it is quite intuitive and is given by:

   $$C_t = \gamma \frac{Y_t}{b_t} + (1-\gamma)C_{t-N} \qquad (15.3.1549)$$

   and finally the forecast is seasonlized again (it is the form of the forecast that makes that we speak of "multiplicative model"):

   $$F_{t+j} = (a_t j + b_t)C_{t+j-N} \qquad (15.3.1550)$$

> This equation assumes that $j < N$. If $N < j < 2N$, the appropriate seasonal factor. If equation, the equation would be C_{t+j-2N}. If $2N < j < 3N$, the appropriate seasonal would be C_{t+j-3N} and so on.

So in the end, we have the system:

$$\boxed{\begin{aligned} F_{t+j} &= (a_t j + b_t)C_{t+j-N} \\ a_t &= \beta(b_t - b_{t-1}) + (1-\beta)a_{t-1} \\ b_t &= \alpha\frac{Y_t}{C_{t-N}} + (1-\alpha)(b_{t-1} + a_{t-1}) \\ C_t &= \gamma\frac{Y_t}{b_t} + (1-\gamma)C_{t-N} \end{aligned}} \tag{15.3.1551}$$

Also sometimes denoted:

$$\begin{aligned} \hat{Y}_{t+j} &= (jT_t + L_t)S_{t+j-N} \\ T_t &= \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} \\ L_t &= \alpha\frac{Y_t}{S_{t-N}} + (1-\alpha)(L_{t-1} + T_{t-1}) \\ S_t &= \gamma\frac{Y_t}{L_t} + (1-\gamma)S_{t-N} \end{aligned} \tag{15.3.1552}$$

where the first relation is often denoted in practice as following too:

$$F_{t+k} = [L_t + kT_t]\, S_{t+k-M} \tag{15.3.1553}$$

We can find also the following variation of the previous model in some books:

$$\begin{aligned} \hat{Y}_{t+j} &= (jT_t + L_t)S_{t+j-N} \\ T_t &= \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} \\ L_t &= \alpha\frac{Y_t}{S_{t-N}} + (1-\alpha)(L_{t-1} + T_{t-1}) \\ S_t &= \gamma\frac{Y_t}{T_t + L_t} + (1-\gamma)S_{t-N} \end{aligned} \tag{15.3.1554}$$

Inspired by the additive Holt's model, the Holt & Winter additive model is given by:

$$\boxed{\begin{aligned} \hat{Y}_{t+j} &= jT_t + L_t + S_{t+j-N} \\ T_t &= \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} \\ L_t &= \alpha(Y_t - S_{t-1}) + (1-\alpha)(L_{t-1} + T_{t-1}) \\ S_t &= \gamma(Y_t - L_t) + (1-\gamma)S_{t-N} \end{aligned}} \tag{15.3.1555}$$

where the first relation is often denoted in practice as following too:

$$F_{t+k} = L_t + kT_t + S_{t+k-M} \tag{15.3.1556}$$

The difference between the multiplicative and additives methods is in the nature of the seasonal component. The additive method is preferred when the variation of seasonal component is almost constant through the series. Seasonal component is expressed in absolute terms and sum up to zero. While multiplicative is used when seasonal variations changes proportionally to the level of the series.

Additive

$$\boxed{F_{t+k} = L_t + kT_t + S_{t+k-M}}$$

Multiplicative

$$\boxed{F_{t+k} = (L_t + kT_t)S_{t+k-M}}$$

It is important to notice that each of the time series components do not need to exhibit the same behaviour. For instance, the multiplicative seasonality formula does — obviously — have multiplicative seasonality, however the trend portion is additive:

Multiplicative Seasonality

$$\boxed{F_{t+k} = \underbrace{(L_t + kT_t)}_{\text{the trend is additive}}S_{t+k-M}}$$

This independence allows the mixing and matching of different behaviours to model many different types of time series:

A: additive T: Trend M: Multiplicative S: Seasonality

ATAS

$$\boxed{F_{t+k} = (L_t + kT_t) + S_{t+k-M}}$$

ATMS

$$\boxed{F_{t+k} = (L_t + kT_t)S_{t+k-M}}$$

ATAS

$$\boxed{F_{t+k} = L_t T_t^k + S_{t+k-M}}$$

MTMS

$$\boxed{F_{t+k} = L_t T_t^k S_{t+k-M}}$$

It's also interesting to compare side by side the update equations for each type of seasonality as we can see below:

Multiplicative

$$\boxed{L_t = \alpha \frac{y_t}{S_{t-M}} + (1-\alpha)(L_{t-1} + T_{t-1})}$$

Additive

$$\boxed{L_t = \alpha(y_t - S_{t-M}) + (1-\alpha)(L_{t-1} + T_{t-1})}$$

We can see the specific deseasonalization effect here:

Multiplicative

$$L_t = \alpha\frac{y_t}{S_{t-M}} + (1-\alpha)(L_{t-1}+T_{t-1})$$

Additive

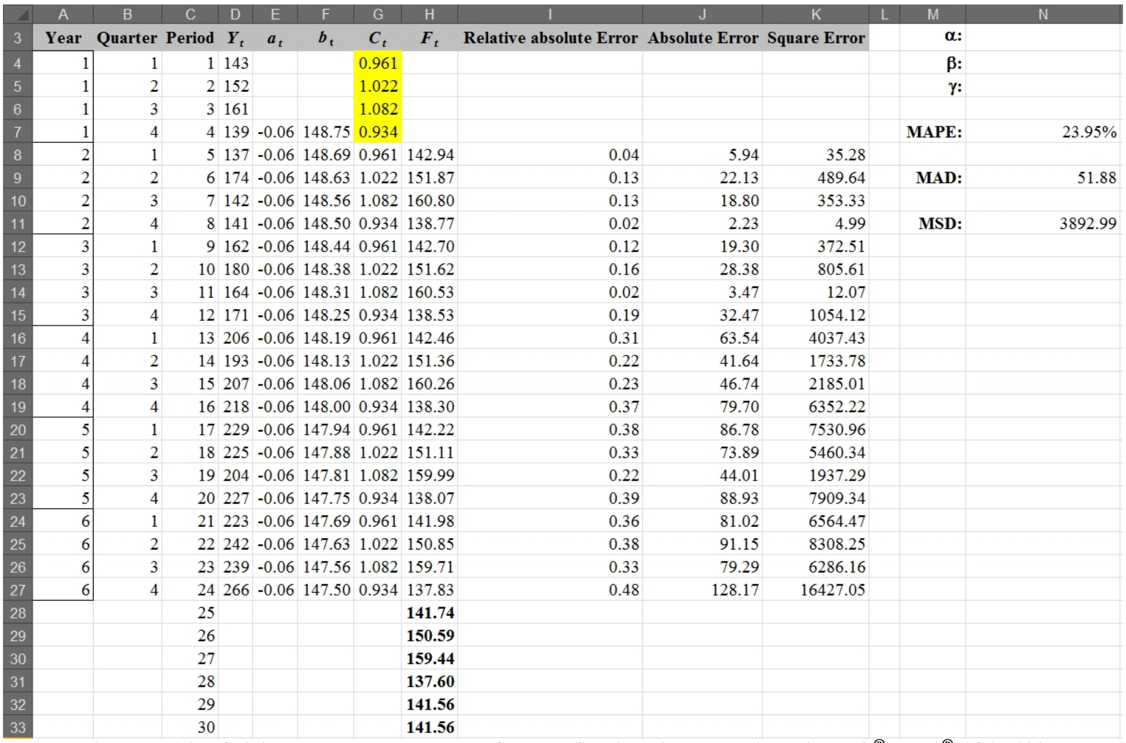
$$L_t = \alpha(y_t - S_{t-M}) + (1-\alpha)(L_{t-1}+T_{t-1})$$

We see above that in order to remove the seasonal component, we must know what type of time series seasonality we are attempting to model, and do the inverse. This translates into either dividing or subtracting the seasonal component for multiplicative and additive time series seasonality, respectively.

To start the forecast with the multiplicative model, we also need to estimated the initial parameters. Again, first forecasted values will be greatly affected by these estimates. The practitioner should as always check that what method its software for the determination of the parameters to specify it in his reports.

✍**Example:**

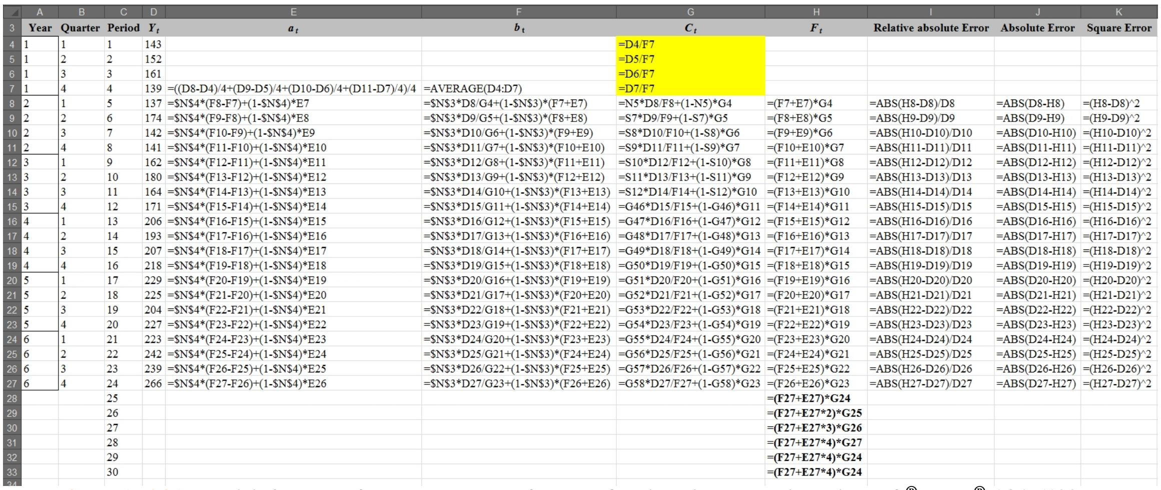
Let us consider the following list in Microsoft® Excel® 14.0.6123 that contains the observed data in column `D` and the model in the columns `E` to `H` which we will detail the formulas just after (here we have the N of our above system that is then equal to zero). Notice that we have 6 years with each 4 quarters (so there will be verbatim 4 seasonal factors that are the first 4 rows of column `G`):



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
3	Year	Quarter	Period	Y_t	a_t	b_t	C_t	F_t	Relative absolute Error	Absolute Error	Square Error		α:	
4	1	1	1	143			0.961						β:	
5	1	2	2	152			1.022						γ:	
6	1	3	3	161			1.082							
7	1	4	4	139	-0.06	148.75	0.934						MAPE:	23.95%
8	2	1	5	137	-0.06	148.69	0.961	142.94	0.04	5.94	35.28			
9	2	2	6	174	-0.06	148.63	1.022	151.87	0.13	22.13	489.64		MAD:	51.88
10	2	3	7	142	-0.06	148.56	1.082	160.80	0.13	18.80	353.33			
11	2	4	8	141	-0.06	148.50	0.934	138.77	0.02	2.23	4.99		MSD:	3892.99
12	3	1	9	162	-0.06	148.44	0.961	142.70	0.12	19.30	372.51			
13	3	2	10	180	-0.06	148.38	1.022	151.62	0.16	28.38	805.61			
14	3	3	11	164	-0.06	148.31	1.082	160.53	0.02	3.47	12.07			
15	3	4	12	171	-0.06	148.25	0.934	138.53	0.19	32.47	1054.12			
16	4	1	13	206	-0.06	148.19	0.961	142.46	0.31	63.54	4037.43			
17	4	2	14	193	-0.06	148.13	1.022	151.36	0.22	41.64	1733.78			
18	4	3	15	207	-0.06	148.06	1.082	160.26	0.23	46.74	2185.01			
19	4	4	16	218	-0.06	148.00	0.934	138.30	0.37	79.70	6352.22			
20	5	1	17	229	-0.06	147.94	0.961	142.22	0.38	86.78	7530.96			
21	5	2	18	225	-0.06	147.88	1.022	151.11	0.33	73.89	5460.34			
22	5	3	19	204	-0.06	147.81	1.082	159.99	0.22	44.01	1937.29			
23	5	4	20	227	-0.06	147.75	0.934	138.07	0.39	88.93	7909.34			
24	6	1	21	223	-0.06	147.69	0.961	141.98	0.36	81.02	6564.47			
25	6	2	22	242	-0.06	147.63	1.022	150.85	0.38	91.15	8308.25			
26	6	3	23	239	-0.06	147.56	1.082	159.71	0.33	79.29	6286.16			
27	6	4	24	266	-0.06	147.50	0.934	137.83	0.48	128.17	16427.05			
28			25					**141.74**						
29			26					**150.59**						
30			27					**159.44**						
31			28					**137.60**						
32			29					**141.56**						
33			30					**141.56**						

Figure 15.245: Original data and parameters for Holt & Winter's model with Microsoft® Excel® 14.0.6123

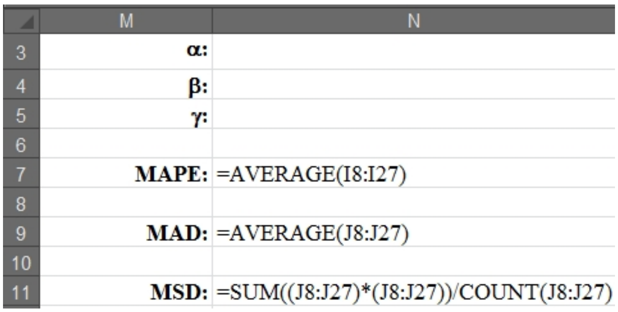
and explicitly for the equations and errors:



	A	B	C	D	E	F	G	H	I	J	K
3	Year	Quarter	Period	Y_t	a_t	b_t	C_t	F_t	Relative absolute Error	Absolute Error	Square Error
4	1	1	1	143			=D4/F7				
5	1	2	2	152			=D5/F7				
6	1	3	3	161			=D6/F7				
7	1	4	4	139	=((D8-D4)/4+(D9-D5)/4+(D10-D6)/4+(D11-D7)/4)/4	=AVERAGE(D4:D7)	=D7/F7				
8	2	1	5	137	=$N$4*(F8-F7)+(1-$N$4)*E7	=$N$3*D8/G4+(1-$N$3)*(F7+E7)	=N5*D8/F8+(1-N5)*G4	=(F7+E7)*G4	=ABS(H8-D8)/D8	=ABS(D8-H8)	=(H8-D8)^2
9	2	2	6	174	=$N$4*(F9-F8)+(1-$N$4)*E8	=$N$3*D9/G5+(1-$N$3)*(F8+E8)	=S7*D9/F9+(1-S7)*G5	=(F8+E8)*G5	=ABS(H9-D9)/D9	=ABS(D9-H9)	=(H9-D9)^2
10	2	3	7	142	=$N$4*(F10-F9)+(1-$N$4)*E9	=$N$3*D10/G6+(1-$N$3)*(F9+E9)	=S8*D10/F10+(1-S8)*G6	=(F9+E9)*G6	=ABS(H10-D10)/D10	=ABS(D10-H10)	=(H10-D10)^2
11	2	4	8	141	=$N$4*(F11-F10)+(1-$N$4)*E10	=$N$3*D11/G7+(1-$N$3)*(F10+E10)	=S9*D11/F11+(1-S9)*G7	=(F10+E10)*G7	=ABS(H11-D11)/D11	=ABS(D11-H11)	=(H11-D11)^2
12	3	1	9	162	=$N$4*(F12-F11)+(1-$N$4)*E11	=$N$3*D12/G8+(1-$N$3)*(F11+E11)	=S10*D12/F12+(1-S10)*G8	=(F11+E11)*G8	=ABS(H12-D12)/D12	=ABS(D12-H12)	=(H12-D12)^2
13	3	2	10	180	=$N$4*(F13-F12)+(1-$N$4)*E12	=$N$3*D13/G9+(1-$N$3)*(F12+E12)	=S11*D13/F13+(1-S11)*G9	=(F12+E12)*G9	=ABS(H13-D13)/D13	=ABS(D13-H13)	=(H13-D13)^2
14	3	3	11	164	=$N$4*(F14-F13)+(1-$N$4)*E13	=$N$3*D14/G10+(1-$N$3)*(F13+E13)	=S12*D14/F14+(1-S12)*G10	=(F13+E13)*G10	=ABS(H14-D14)/D14	=ABS(D14-H14)	=(H14-D14)^2
15	3	4	12	171	=$N$4*(F15-F14)+(1-$N$4)*E14	=$N$3*D15/G11+(1-$N$3)*(F14+E14)	=G46*D15/F15+(1-G46)*G11	=(F14+E14)*G11	=ABS(H15-D15)/D15	=ABS(D15-H15)	=(H15-D15)^2
16	4	1	13	206	=$N$4*(F16-F15)+(1-$N$4)*E15	=$N$3*D16/G12+(1-$N$3)*(F15+E15)	=G47*D16/F16+(1-G47)*G12	=(F15+E15)*G12	=ABS(H16-D16)/D16	=ABS(D16-H16)	=(H16-D16)^2
17	4	2	14	193	=$N$4*(F17-F16)+(1-$N$4)*E16	=$N$3*D17/G13+(1-$N$3)*(F16+E16)	=G48*D17/F17+(1-G48)*G13	=(F16+E16)*G13	=ABS(H17-D17)/D17	=ABS(D17-H17)	=(H17-D17)^2
18	4	3	15	207	=$N$4*(F18-F17)+(1-$N$4)*E17	=$N$3*D18/G14+(1-$N$3)*(F17+E17)	=G49*D18/F18+(1-G49)*G14	=(F17+E17)*G14	=ABS(H18-D18)/D18	=ABS(D18-H18)	=(H18-D18)^2
19	4	4	16	218	=$N$4*(F19-F18)+(1-$N$4)*E18	=$N$3*D19/G15+(1-$N$3)*(F18+E18)	=G50*D19/F19+(1-G50)*G15	=(F18+E18)*G15	=ABS(H19-D19)/D19	=ABS(D19-H19)	=(H19-D19)^2
20	5	1	17	229	=$N$4*(F20-F19)+(1-$N$4)*E19	=$N$3*D20/G16+(1-$N$3)*(F19+E19)	=G51*D20/F20+(1-G51)*G16	=(F19+E19)*G16	=ABS(H20-D20)/D20	=ABS(D20-H20)	=(H20-D20)^2
21	5	2	18	225	=$N$4*(F21-F20)+(1-$N$4)*E20	=$N$3*D21/G17+(1-$N$3)*(F20+E20)	=G52*D21/F21+(1-G52)*G17	=(F20+E20)*G17	=ABS(H21-D21)/D21	=ABS(D21-H21)	=(H21-D21)^2
22	5	3	19	204	=$N$4*(F22-F21)+(1-$N$4)*E21	=$N$3*D22/G18+(1-$N$3)*(F21+E21)	=G53*D22/F22+(1-G53)*G18	=(F21+E21)*G18	=ABS(H22-D22)/D22	=ABS(D22-H22)	=(H22-D22)^2
23	5	4	20	227	=$N$4*(F23-F22)+(1-$N$4)*E22	=$N$3*D23/G19+(1-$N$3)*(F22+E22)	=G54*D23/F23+(1-G54)*G19	=(F22+E22)*G19	=ABS(H23-D23)/D23	=ABS(D23-H23)	=(H23-D23)^2
24	6	1	21	223	=$N$4*(F24-F23)+(1-$N$4)*E23	=$N$3*D24/G20+(1-$N$3)*(F23+E23)	=G55*D24/F24+(1-G55)*G20	=(F23+E23)*G20	=ABS(H24-D24)/D24	=ABS(D24-H24)	=(H24-D24)^2
25	6	2	22	242	=$N$4*(F25-F24)+(1-$N$4)*E24	=$N$3*D25/G21+(1-$N$3)*(F24+E24)	=G56*D25/F25+(1-G56)*G21	=(F24+E24)*G21	=ABS(H25-D25)/D25	=ABS(D25-H25)	=(H25-D25)^2
26	6	3	23	239	=$N$4*(F26-F25)+(1-$N$4)*E25	=$N$3*D26/G22+(1-$N$3)*(F25+E25)	=G57*D26/F26+(1-G57)*G22	=(F25+E25)*G22	=ABS(H26-D26)/D26	=ABS(D26-H26)	=(H26-D26)^2
27	6	4	24	266	=$N$4*(F27-F26)+(1-$N$4)*E26	=$N$3*D27/G23+(1-$N$3)*(F26+E26)	=G58*D27/F27+(1-G58)*G23	=(F26+E26)*G23	=ABS(H27-D27)/D27	=ABS(D27-H27)	=(H27-D27)^2
28			25					**=(F27+E27)*G24**			
29			26					**=(F27+E27*2)*G25**			
30			27					**=(F27+E27*3)*G26**			
31			28					**=(F27+E27*4)*G27**			
32			29					**=(F27+E27*4)*G24**			
33			30					**=(F27+E27*4)*G24**			

Figure 15.246: Explicit formulas for data and error for Holt & Winter's model with Microsoft® Excel® 14.0.6123

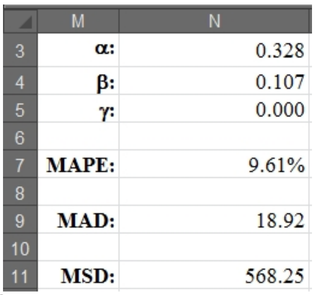
and for the parameters:



	M	N
3	α:	
4	β:	
5	γ:	
6		
7	MAPE:	=AVERAGE(I8:I27)
8		
9	MAD:	=AVERAGE(J8:J27)
10		
11	MSD:	=SUM((J8:J27)*(J8:J27))/COUNT(J8:J27)

Figure 15.247: Explicit formulas for parameters and error indicators model with Microsoft® Excel® 14.0.6123

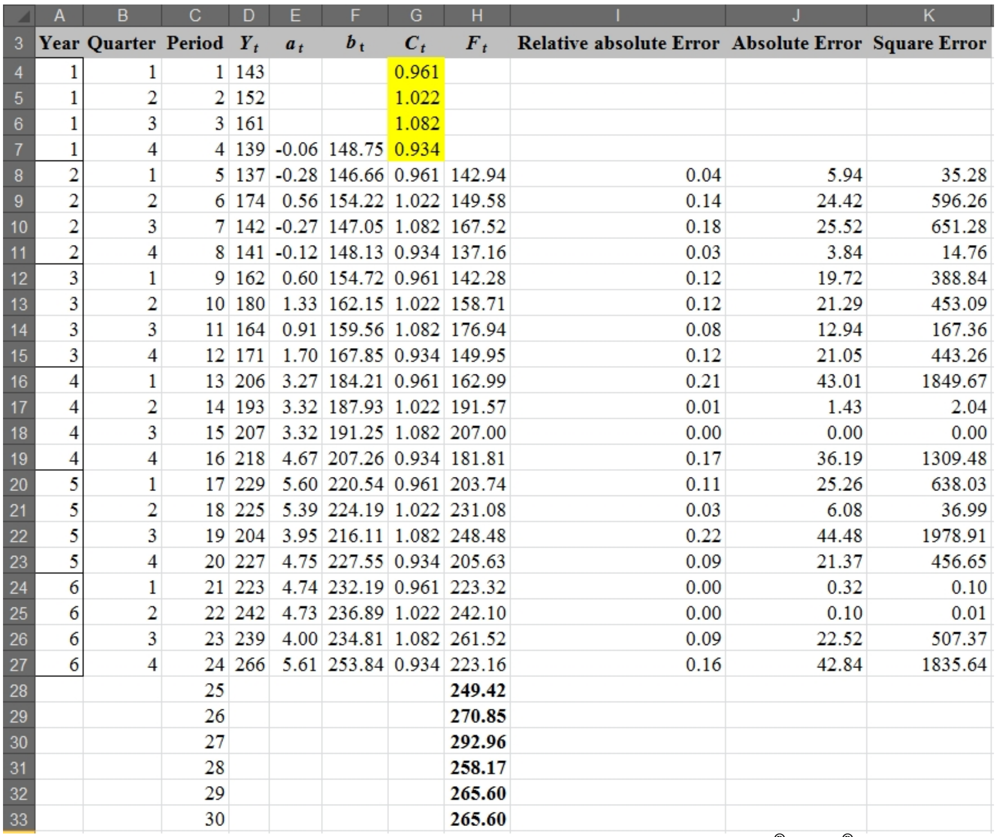
As for all the models seen previous, we start the solver with the idea to minimize the MAPE:



	M	N
3	α:	0.328
4	β:	0.107
5	γ:	0.000
6		
7	MAPE:	9.61%
8		
9	MAD:	18.92
10		
11	MSD:	568.25

Figure 15.248: Holt & Winter's final parameters and error values with Microsoft® Excel® 14.0.6123

This gives for values:



	A	B	C	D	E	F	G	H	I	J	K
3	Year	Quarter	Period	Y_t	a_t	b_t	C_t	F_t	Relative absolute Error	Absolute Error	Square Error
4	1	1	1	143			0.961				
5	1	2	2	152			1.022				
6	1	3	3	161			1.082				
7	1	4	4	139	-0.06	148.75	0.934				
8	2	1	5	137	-0.28	146.66	0.961	142.94	0.04	5.94	35.28
9	2	2	6	174	0.56	154.22	1.022	149.58	0.14	24.42	596.26
10	2	3	7	142	-0.27	147.05	1.082	167.52	0.18	25.52	651.28
11	2	4	8	141	-0.12	148.13	0.934	137.16	0.03	3.84	14.76
12	3	1	9	162	0.60	154.72	0.961	142.28	0.12	19.72	388.84
13	3	2	10	180	1.33	162.15	1.022	158.71	0.12	21.29	453.09
14	3	3	11	164	0.91	159.56	1.082	176.94	0.08	12.94	167.36
15	3	4	12	171	1.70	167.85	0.934	149.95	0.12	21.05	443.26
16	4	1	13	206	3.27	184.21	0.961	162.99	0.21	43.01	1849.67
17	4	2	14	193	3.32	187.93	1.022	191.57	0.01	1.43	2.04
18	4	3	15	207	3.32	191.25	1.082	207.00	0.00	0.00	0.00
19	4	4	16	218	4.67	207.26	0.934	181.81	0.17	36.19	1309.48
20	5	1	17	229	5.60	220.54	0.961	203.74	0.11	25.26	638.03
21	5	2	18	225	5.39	224.19	1.022	231.08	0.03	6.08	36.99
22	5	3	19	204	3.95	216.11	1.082	248.48	0.22	44.48	1978.91
23	5	4	20	227	4.75	227.55	0.934	205.63	0.09	21.37	456.65
24	6	1	21	223	4.74	232.19	0.961	223.32	0.00	0.32	0.10
25	6	2	22	242	4.73	236.89	1.022	242.10	0.00	0.10	0.01
26	6	3	23	239	4.00	234.81	1.082	261.52	0.09	22.52	507.37
27	6	4	24	266	5.61	253.84	0.934	223.16	0.16	42.84	1835.64
28			25					**249.42**			
29			26					**270.85**			
30			27					**292.96**			
31			28					**258.17**			
32			29					**265.60**			
33			30					**265.60**			

Figure 15.249: Holt & Winter's final values, forecasts and errors with Microsoft® Excel® 14.0.6123

Graphically this gives if we compare all four models:

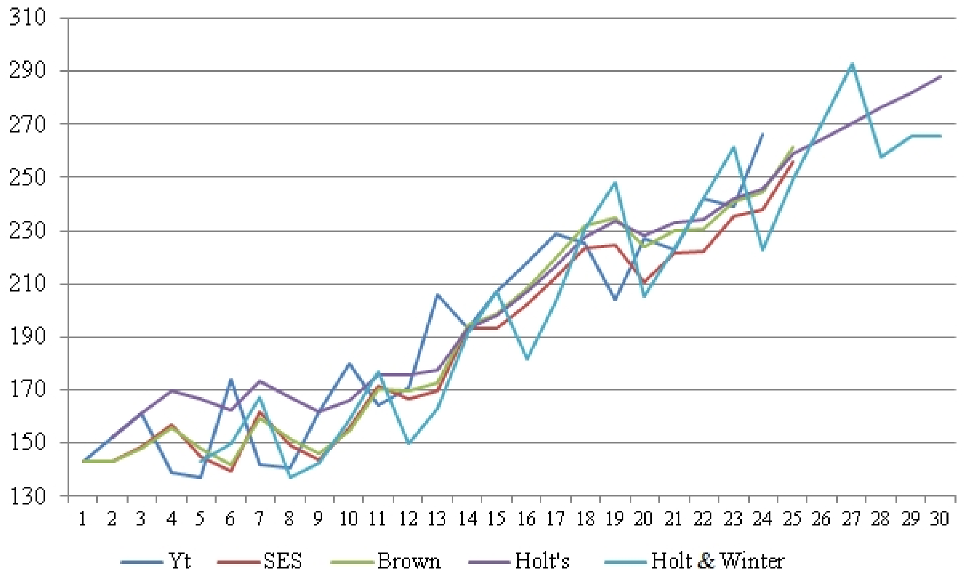


Figure 15.250: **Holt & Winter's full plot and comparison with other models in Microsoft® Excel® 14.0.6123**

Holt's smoothing therefore had a MAPE of 7.49% and the Brown's one a MAPE of 7.45%. The Holt & Winter's seasonal smoothing has a MAPE of 9.61%, that is to say the fitting is worst than the other models. However the advantage of Holt & Winter's model is than we can make projections over more than one period AND with cycles as we can see on the chart above.

It is of course interesting to compare the 6 months forecast values between the Brown's, Holt's additive and Holt & Winter's model:

Month	Brown's DSES	Holt's DSES additive	Holt & Winter's multiplicative
25	261.50	263.30	249.42
26	268.19	270.00	270.85
27	264.55	276.69	292.96
28	280.53	283.39	258.17
29	238.20	290.08	265.60
30	311.19	296.78	265.60

Table 15.61: **Comparison between Holt's and Brown's DSES + Holt & Winter's multiplicative TSES**

So we see again that depending on the chosen technique, the differences are relatively important.

The reader should not forget that there is also an additive Holt & Winter's model. Furthermore we have seen once that there is also a "Holt & Winter's multiple model" that is not the same as the "Holt & Winter's multiplicative model"!

Caution! "Traditional" exponential smoothing models like simple exponential smoothing, Holt-Winter double exponential smoothing, Brown's double exponential smoothing Peter Winters' triple exponential smoothing are heuristic models and do not have an underlying statistical theory! Hence they don't have a maximum likelihood estimator. However, seminal work published in the *Journal of the American Statistical Association* by Ord, Koehler & Snyder in 11997 (holocene calendar) provided a statistical basis for exponential smoothing using a state space framework and provided a maximum likelihood estimation procedure under Gaussian assumption. The *Forecasting with Exponential Smoothing* ([269]) extends the original work into all the forms of exponential smoothing.

15.3.13.3.3.5 ETS (Error, Trend, Seasonality) Models

We have previously see a series of methods for time series prediction named "exponential smoothing". But this is only a basic overview of the most common academics undergraduate models.

If we include statistical errors (noise) the models that we have seen before, that include already a "trend" term and "seasonality" term, have a complementary term named the "error".

Such models, named "ETS models" (that are simply a generalized framework of exponential smoothing models), are summarized into the following notation:

$$\text{ETS}(\cdot,\cdot,\cdot) \tag{15.3.1557}$$

where each dot can respectively be denoted:

- N: For "None" (that means there can be no trend, or no error, or no seasonality)
- A: For "Additive" (that means there can be additive trend, or additive error, or additive seasonality)
- M: For "Multiplicative" (that means there can be multiplicative trend, or multiplicative error, or multiplicative seasonality)
- A_d: For "Additive damped" (that means there can be additive damped trend)
- M_d: For "Multiplicative damped" (that means there can be multiplicative damped trend)

As ETS models have three parameters, which is written:

$$\boxed{\text{ETS}\left(\{A,M\},\{N,A,M,A_d,M_d\},\{N,A,M\}\right)} \tag{15.3.1558}$$

hence a total of $2\cdot 5\cdot 3=30$ models!!!

✍**Examples:**

E1. The notation ETS(A,A,A) means that we have an exponential smoothing model with additive error terms, additive trend, additive seasonality (used in Microsoft® Excel® since 2013 and Microsoft Power BI since 2017).

$$\begin{aligned}y_t &= \ell_{t-1}+b_{t-1}+s_{t-m}+\varepsilon_t\\ \ell_t &= \ell_{t-1}+b_{t-1}+\alpha\varepsilon_t\\ b_t &= b_{t-1}+\beta\varepsilon_t\\ s_t &= s_{t-m}+\gamma\varepsilon_t\end{aligned}\tag{15.3.1559}$$

We can see very clearly above what do Microsoft® Excel® or Microsoft Power BI when they ask the user an input for the value of *Seasonality* (corresponding to m in the relation above)! If your time-step t is the month for example, it means how far back in the past (m months) the algorithm must go to add (because the model is additive) a given constant. At least this is how it work in the year 12023 (holocene calendar).

E2. The notation ETS(A,A,N) means that we have an exponential smoothing model with additive error terms, additive trend, but no seasonality (used in Microsoft® Excel® since 2013 and Microsoft Power BI since 2017).

$$\begin{aligned}y_t &= \ell_{t-1}+\varepsilon_t\\ \ell_t &= \ell_{t-1}+\alpha\varepsilon_t\end{aligned}\tag{15.3.1560}$$

E3. And relatively to previous seen models, we have for example that:

- simple exponential smoothing (or SES) method is $\{N,N\}$, i.e. non-additive neither non-multiplicative trend or seasonality
- Holt's linear method is $\{A,N\}$ additive trend without seasonality
- Holt-Winters' additive method is $\{A,A\}$, i.e. additive trend and additive seasonality
- Holt-Winters' multiplicative method is $\{A,M\}$, i.e. additive trend and multiplicative seasonality

All 30 models (some authors merge together the "additive damped" and "multiplicative damped" and speak therefore of 24 models instead of 30) are summarized in this both amazing double entry tables in an non-recursive formulation:

ADDITIVE ERROR MODELS

Trend	Seasonal		
	N	A	M
N	$y_t = \ell_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t$	$y_t = \ell_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = \ell_{t-1}s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/\ell_{t-1}$
A	$y_t = \ell_{t-1} + b_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t$	$y_t = \ell_{t-1} + b_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1})s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = b_{t-1} + \beta\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1} + b_{t-1})$
A$_\mathrm{d}$	$y_t = \ell_{t-1} + \phi b_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t$ $b_t = \phi b_{t-1} + \beta\varepsilon_t$	$y_t = \ell_{t-1} + \phi b_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t$ $b_t = \phi b_{t-1} + \beta\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1})s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = \phi b_{t-1} + \beta\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1} + \phi b_{t-1})$
M	$y_t = \ell_{t-1}b_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1}b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t/\ell_{t-1}$	$y_t = \ell_{t-1}b_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1}b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t/\ell_{t-1}$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = \ell_{t-1}b_{t-1}s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1}b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = b_{t-1} + \beta\varepsilon_t/(s_{t-m}\ell_{t-1})$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1}b_{t-1})$
M$_\mathrm{d}$	$y_t = \ell_{t-1}b^{\phi}_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1}b^{\phi}_{t-1} + \alpha\varepsilon_t$ $b_t = b^{\phi}_{t-1} + \beta\varepsilon_t/\ell_{t-1}$	$y_t = \ell_{t-1}b^{\phi}_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1}b^{\phi}_{t-1} + \alpha\varepsilon_t$ $b_t = b^{\phi}_{t-1} + \beta\varepsilon_t/\ell_{t-1}$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = \ell_{t-1}b^{\phi}_{t-1}s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1}b^{\phi}_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = b^{\phi}_{t-1} + \beta\varepsilon_t/(s_{t-m}\ell_{t-1})$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1}b^{\phi}_{t-1})$

MULTIPLICATIVE ERROR MODELS

Trend	Seasonal		
	N	A	M
N	$y_t = \ell_{t-1}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}(1 + \alpha\varepsilon_t)$	$y_t = (\ell_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + \alpha(\ell_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + s_{t-m})\varepsilon_t$	$y_t = \ell_{t-1}s_{t-m}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}(1 + \alpha\varepsilon_t)$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$
A	$y_t = (\ell_{t-1} + b_{t-1})(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1})\varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1})s_{t-m}(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1})\varepsilon_t$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$
A$_\mathrm{d}$	$y_t = (\ell_{t-1} + \phi b_{t-1})(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + \phi b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1})\varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1})s_{t-m}(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + \phi b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1})\varepsilon_t$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$
M	$y_t = \ell_{t-1}b_{t-1}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}b_{t-1}(1 + \alpha\varepsilon_t)$ $b_t = b_{t-1}(1 + \beta\varepsilon_t)$	$y_t = (\ell_{t-1}b_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}b_{t-1} + \alpha(\ell_{t-1}b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = b_{t-1} + \beta(\ell_{t-1}b_{t-1} + s_{t-m})\varepsilon_t/\ell_{t-1}$ $s_t = s_{t-m} + \gamma(\ell_{t-1}b_{t-1} + s_{t-m})\varepsilon_t$	$y_t = \ell_{t-1}b_{t-1}s_{t-m}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}b_{t-1}(1 + \alpha\varepsilon_t)$ $b_t = b_{t-1}(1 + \beta\varepsilon_t)$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$
M$_\mathrm{d}$	$y_t = \ell_{t-1}b^{\phi}_{t-1}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}b^{\phi}_{t-1}(1 + \alpha\varepsilon_t)$ $b_t = b^{\phi}_{t-1}(1 + \beta\varepsilon_t)$	$y_t = (\ell_{t-1}b^{\phi}_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}b^{\phi}_{t-1} + \alpha(\ell_{t-1}b^{\phi}_{t-1} + s_{t-m})\varepsilon_t$ $b_t = b^{\phi}_{t-1} + \beta(\ell_{t-1}b^{\phi}_{t-1} + s_{t-m})\varepsilon_t/\ell_{t-1}$ $s_t = s_{t-m} + \gamma(\ell_{t-1}b^{\phi}_{t-1} + s_{t-m})\varepsilon_t$	$y_t = \ell_{t-1}b^{\phi}_{t-1}s_{t-m}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}b^{\phi}_{t-1}(1 + \alpha\varepsilon_t)$ $b_t = b^{\phi}_{t-1}(1 + \beta\varepsilon_t)$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$

Figure 15.251: **Classification of ETS$(\cdot,\cdot,\cdot)$ models (source: [270])**

The general formulation of all thirty models above involves a state vector $\vec{x}_t = (\ell_t, b_t, s_t, s_{t-1}, \ldots, s_{t-m+1})'$ and state space equations of the form:

$$\begin{aligned} y_t &= w\left(\vec{x}_{t-1}\right) + r\left(\vec{x}_{t-1}\right)\varepsilon_t \\ \vec{x}_t &= f\left(\vec{x}_{t-1}\right) + g\left(\vec{x}_{t-1}\right)\varepsilon_t \end{aligned} \quad (15.3.1561)$$

✍**Example:**

To give a specific example, consider the ETS(M,A_d,M) model:

$$\begin{aligned}
y_t &= (\ell_{t-1}+\phi b_{t-1})\, s_{t-m}\,(1+\varepsilon_t)\\
\ell_t &= (\ell_{t-1}+\phi b_{t-1})\,(1+\alpha\varepsilon_t)\\
b_t &= \phi b_{t-1}+\beta\,(\ell_{t-1}+\phi b_{t-1})\,\varepsilon_t\\
s_t &= s_{t-m}\,(1+\gamma\varepsilon_t)
\end{aligned}$$

with $m=4$. Then:

$$\vec{x}_t=\begin{pmatrix}\ell_t\\ b_t\\ s_t\\ s_{t-1}\\ s_{t-2}\\ s_{t-3}\end{pmatrix},\ \vec{x}_{t-1}=\begin{pmatrix}\ell_{t-1}\\ b_{t-1}\\ s_{t-1}\\ s_{t-2}\\ s_{t-3}\\ s_{t-4}\end{pmatrix}$$

Therefore we put:

$$\begin{aligned}
w\,(\vec{x}_{t-1}) &= (\ell_{t-1}+\phi b_{t-1})\, s_{t-4}\\
r\,(\vec{x}_{t-1}) &= (\ell_{t-1}+\phi b_{t-1})\, s_{t-4}
\end{aligned}$$

Hence:

$$\begin{aligned}
y_t &= w\,(\vec{x}_{t-1})+r\,(\vec{x}_{t-1})\,\varepsilon_t=(\ell_{t-1}+\phi b_{t-1})\, s_{t-4}+(\ell_{t-1}+\phi b_{t-1})\, s_{t-4}\varepsilon_t\\
&= (\ell_{t-1}+\phi b_{t-1})\, s_{t-m}\,(1+\varepsilon_t)
\end{aligned}$$

and:

$$\begin{aligned}
\vec{x}_t=\begin{pmatrix}\ell_t\\ b_t\\ s_t\\ s_{t-1}\\ s_{t-2}\\ s_{t-3}\end{pmatrix} &= f\,(\vec{x}_{t-1})+g\,(\vec{x}_{t-1})\,\varepsilon_t=\begin{pmatrix}\ell_{t-1}+\phi b_{t-1}\\ \phi b_{t-1}\\ s_{t-4}\\ s_{t-1}\\ s_{t-2}\\ s_{t-3}\end{pmatrix}+\begin{pmatrix}\alpha\,(\ell_{t-1}+\phi b_{t-1})\\ \beta\,(\ell_{t-1}+\phi b_{t-1})\\ \gamma s_{t-4}\\ 0\\ 0\\ 0\end{pmatrix}\varepsilon_t\\
&= \begin{pmatrix}\ell_{t-1}+\phi b_{t-1}+\alpha\,(\ell_{t-1}+\phi b_{t-1})\,\varepsilon_t\\ \phi b_{t-1}+\beta\,(\ell_{t-1}+\phi b_{t-1})\,\varepsilon_t\\ s_{t-4}+\gamma s_{t-4}\varepsilon_t\\ s_{t-1}\\ s_{t-2}\\ s_{t-3}\end{pmatrix}=\begin{pmatrix}(\ell_{t-1}+\phi b_{t-1})\,(1+\alpha\varepsilon_t)\\ \phi b_{t-1}+\beta\,(\ell_{t-1}+\phi b_{t-1})\,\varepsilon_t\\ s_{t-4}(1+\gamma\varepsilon_t)\\ s_{t-1}\\ s_{t-2}\\ s_{t-3}\end{pmatrix}
\end{aligned}$$

According to the specialists in the field, it seems that from the thirty models above, only the fifteen visible in the table below are numerically stable:

Class 1			
ANN	$Y_t = \ell_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t$	ANA	$Y_t = \ell_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t.$
AAN	$Y_t = \ell_{t-1} + b_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \alpha\beta\varepsilon_t$	AAA	$Y_t = \ell_{t-1} + b_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \alpha\beta\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t.$
ADN	$Y_t = \ell_{t-1} + b_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = \phi b_{t-1} + \alpha\beta\varepsilon_t$	ADA	$Y_t = \ell_{t-1} + b_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = \phi b_{t-1} + \alpha\beta\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t.$
Class 2			
MNN	$Y_t = \ell_{t-1}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}(1 + \alpha\varepsilon_t).$	MNA	$Y_t = (\ell_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + \alpha(\ell_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + s_{t-m})\varepsilon_t.$
MAN	$Y_t = (\ell_{t-1} + b_{t-1})(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = b_{t-1} + \alpha\beta(\ell_{t-1} + b_{t-1})\varepsilon_t$	MAA	$Y_t = (\ell_{t-1} + b_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = b_{t-1} + \alpha\beta(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t.$
MDN	$Y_t = (\ell_{t-1} + b_{t-1})(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = \phi b_{t-1} + \alpha\beta(\ell_{t-1} + b_{t-1})\varepsilon_t$	MDA	$Y_t = (\ell_{t-1} + b_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = \phi b_{t-1} + \alpha\beta(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t.$
Class 3			
MNM	$Y_t = \ell_{t-1}s_{t-m}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}(1 + \alpha\varepsilon_t)$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$		
MAM	$Y_t = (\ell_{t-1} + b_{t-1})s_{t-m}(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = b_{t-1} + \alpha\beta(\ell_{t-1} + b_{t-1})\varepsilon_t$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t).$		
MDM	$Y_t = (\ell_{t-1} + b_{t-1})s_{t-m}(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = \phi b_{t-1} + \alpha\beta(\ell_{t-1} + b_{t-1})\varepsilon_t$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t).$		

Figure 15.252: **Another simplified typical classification of ETS models**

Obviously not all of the 15 numerically stable models are appropriate for all data. The multiplicative error models are not well defined if there are zeros in the data. Similarly, we don't consider the additive error models with multiplicative trend or multiplicative seasonality if any observations are zero. Further, if the data are not quarterly or monthly (and do not have some other obvious seasonal period), then we do not consider any of the seasonal methods.

A very nice paper that compare all these models and that we strongly recommend the reader to take a look at is *A state framework for automatic forecasting using exponential smoothing* ([272]).

The content of the thirty models presented in the above tables can be expressed in a different way using the recursive formulation (we have already seen in details how to go from a non-recursive to a recursive formulation in the four previous models: Simple Exponential Smoothing, Double Exponential Smoothing with one parameter, Holt's Double Exponential Smoothing with two parameters, Holt's and Winter Triple Exponential Smoothing with three parameters):

Trend	Seasonal		
	N	A	M
N	$\ell_t = \alpha y_t + (1-\alpha)\ell_{t-1}$ $\hat{y}_{t+h\|t} = \ell_t$	$\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)\ell_{t-1}$ $s_t = \gamma(y_t - \ell_{t-1}) + (1-\gamma)s_{t-m}$ $\hat{y}_{t+h\|t} = \ell_t + s_{t-m+h_m^+}$	$\ell_t = \alpha(y_t / s_{t-m}) + (1-\alpha)\ell_{t-1}$ $s_t = \gamma(y_t / \ell_{t-1}) + (1-\gamma)s_{t-m}$ $\hat{y}_{t+h\|t} = \ell_t s_{t-m+h_m^+}$
A	$\ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$ $\hat{y}_{t+h\|t} = \ell_t + hb_t$	$\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$ $s_t = \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m}$ $\hat{y}_{t+h\|t} = \ell_t + hb_t + s_{t-m+h_m^+}$	$\ell_t = \alpha(y_t / s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$ $s_t = \gamma(y_t / (\ell_{t-1} + b_{t-1})) + (1-\gamma)s_{t-m}$ $\hat{y}_{t+h\|t} = (\ell_t + hb_t)s_{t-m+h_m^+}$
$\mathrm{A_d}$	$\ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + \phi b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)\phi b_{t-1}$ $\hat{y}_{t+h\|t} = \ell_t + \phi_h b_t$	$\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + \phi b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)\phi b_{t-1}$ $s_t = \gamma(y_t - \ell_{t-1} - \phi b_{t-1}) + (1-\gamma)s_{t-m}$ $\hat{y}_{t+h\|t} = \ell_t + \phi_h b_t + s_{t-m+h_m^+}$	$\ell_t = \alpha(y_t / s_{t-m}) + (1-\alpha)(\ell_{t-1} + \phi b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)\phi b_{t-1}$ $s_t = \gamma(y_t / (\ell_{t-1} + \phi b_{t-1})) + (1-\gamma)s_{t-m}$ $\hat{y}_{t+h\|t} = (\ell_t + \phi_h b_t)s_{t-m+h_m^+}$
M	$\ell_t = \alpha y_t + (1-\alpha)\ell_{t-1}b_{t-1}$ $b_t = \beta^*(\ell_t / \ell_{t-1}) + (1-\beta^*)b_{t-1}$ $\hat{y}_{t+h\|t} = \ell_t b_t^h$	$\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)\ell_{t-1}b_{t-1}$ $b_t = \beta^*(\ell_t / \ell_{t-1}) + (1-\beta^*)b_{t-1}$ $s_t = \gamma(y_t - \ell_{t-1}b_{t-1}) + (1-\gamma)s_{t-m}$ $\hat{y}_{t+h\|t} = \ell_t b_t^h + s_{t-m+h_m^+}$	$\ell_t = \alpha(y_t / s_{t-m}) + (1-\alpha)\ell_{t-1}b_{t-1}$ $b_t = \beta^*(\ell_t / \ell_{t-1}) + (1-\beta^*)b_{t-1}$ $s_t = \gamma(y_t / (\ell_{t-1}b_{t-1})) + (1-\gamma)s_{t-m}$ $\hat{y}_{t+h\|t} = \ell_t b_t^h s_{t-m+h_m^+}$
$\mathrm{M_d}$	$\ell_t = \alpha y_t + (1-\alpha)\ell_{t-1}b_{t-1}^{\phi}$ $b_t = \beta^*(\ell_t / \ell_{t-1}) + (1-\beta^*)b_{t-1}^{\phi}$ $\hat{y}_{t+h\|t} = \ell_t b_t^{\phi_h}$	$\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)\ell_{t-1}b_{t-1}^{\phi}$ $b_t = \beta^*(\ell_t / \ell_{t-1}) + (1-\beta^*)b_{t-1}^{\phi}$ $s_t = \gamma(y_t - \ell_{t-1}b_{t-1}^{\phi}) + (1-\gamma)s_{t-m}$ $\hat{y}_{t+h\|t} = \ell_t b_t^{\phi_h} + s_{t-m+h_m^+}$	$\ell_t = \alpha(y_t / s_{t-m}) + (1-\alpha)\ell_{t-1}b_{t-1}^{\phi}$ $b_t = \beta^*(\ell_t / \ell_{t-1}) + (1-\beta^*)b_{t-1}^{\phi}$ $s_t = \gamma(y_t / (\ell_{t-1}b_{t-1}^{\phi})) + (1-\gamma)s_{t-m}$ $\hat{y}_{t+h\|t} = \ell_t b_t^{\phi_h} s_{t-m+h_m^+}$

In each case, ℓ_t denotes the series level at time t, b_t denotes the slope at time t, s_t denotes the seasonal component of the series at time t, and m denotes the number of seasons in a year; α, β^*, γ and ϕ are constants, $\phi_h = \phi + \phi^2 + \cdots + \phi^h$ and $h_m^+ = \lfloor (h-1) \bmod m \rfloor + 1$.

Figure 15.253: **Classification of ETS models with recursive formulation (source: [269])**

Remark

"Intermittent demand" (ID) (also known as "sporadic demand") comes about when a product experiences several periods of zero demand. Often in these situations when the demand that occurs is small and sometimes highly variable in size models like those seen previously and also recurrent neural networks perform very poorly! Actually a common used model for such situations is named the "Croston's method" that we will not introduce in details in this textbook as it is too empirical in the point of view of the redactor of this section.

It's possible to derive (but actually out of the scope of this textbook), the forecast mean and standard deviation of the 15 models given in the table above (for the details and proofs see [273]).

The summary of the corresponding forecast mean and variance are given by the tables below but it is important to notice that estimates for σ^2 are not done in the same manner for all three classes in the previous table. The estimate for σ^2 would be:

$$\hat{\sigma}^2=\sum_{t=1}^{n}\hat{\varepsilon}_t^2/n \tag{15.3.1562}$$

where:

$$\hat{\varepsilon}_t=\begin{cases} Y_t-\hat{Y}_t & \text{for Class 1}\\ \left(Y_t-\hat{Y}_t\right)/\hat{Y}_t & \text{for Classes 2 and 3}\end{cases} \tag{15.3.1563}$$

So now that this is given here are the h-period-ahead forecast means and variances:

	Mean	**Variance**
Class 1	μ_h	$v_1=\sigma^2$ and $v_h=\sigma^2\left(1+\sum_{j=1}^{h-1}c_j^2\right)$
Class 2	μ_h	$v_h=(1+\sigma^2)\,\theta_h-\mu_h^2$ where $\theta_1=\mu_1^2$ and $\theta_h=\mu_h^2+\sigma^2\sum_{j=1}^{h-1}c_j^2\theta_{h-j}$
Class 3	μ_h	$v_h=s_{n-m+h}^2\left[\theta_h\,(1+\sigma^2)\,(1+\gamma^2\sigma^2)^k-\tilde{\mu}_h^2\right]$ where $\theta_1=\tilde{\mu}_1^2,\theta_h=\tilde{\mu}_h^2+\sigma^2\sum_{j=1}^{h-1}c_j^2\theta_{h-j}$ and $k=\lfloor(h-1)/m\rfloor$.

with:

	μ_h	$\tilde{\mu}_h$	c_j
Class 1/Class 2			
ANN/MNN	ℓ_n		α
AAN/MAN	ℓ_n+hb_n		$\alpha(1+j\beta)$
ADN/MDN	$\ell_n+\phi_{h-1}b_n$		$\alpha\,(1+\phi_{j-1}\beta)$
ANA/MNA	$\ell_n+s_{n-m+1+(h-1)^*}$		$\alpha+\gamma d_{j,m}$
AAA/MAA	$\ell_n+hb_n+s_{n-m+1+(h-1)^*}$		$\alpha(1+j\beta)+\gamma d_{j,m}$
ADA/MDA	$\ell_n+\phi_{h-1}b_n+s_{n-m+1+(h-1)^*}$		$\alpha\,(1+\phi_{j-1}\beta)+\gamma d_{j,m}$
Class 3			
MNM	$\ell_n s_{n-m+1+(h-1)^*}$	ℓ_n	α
MAM	$(\ell_n+hb_n)\,s_{n-m+1+(h-1)^*}$	ℓ_n+hb_n	$\alpha(1+j\beta)$
MDM	$(\ell_n+\phi_{h-1}b_n)\,s_{n-m+1+(h-1)^*}$	$\ell_n+\phi_{h-1}b_n$	$\alpha\,(1+\phi_{j-1}\beta)$

For more details on exponential smoothing models and their forecasting with strongly recommend to the reader the 360 pages textbook *Forecasting with Exponential Smoothing* ([269]).

15.3.13.3.4 Logistic Model

It always happens in corporations that in the analysis of a product or service over a given period of time that the quantities of sales of the latter increase, then go through an inflection point (second derivative equal to zero as seen at page 1454) and then go towards an asymptote to decrease again later with a more or less similar characteristic.

The logistic model adapted to simulate such behaviour as part of the field of time series analysis (and that must not be confused with the logistic regression as seen in the section of Numerical Methods at page 6449) is defined as:

$$\boxed{Y_t=\frac{\hat{Y}_{\max}}{1+br^t}} \tag{15.3.1564}$$

The logistic model is inspired from many models that we find in physics and $\hat{Y}_{\max}$ is the estimated saturation threshold (horizontal asymptote) which can be approximated (with a tolerance interval) by an audit of the market.

Remark

We often speak about 3M (Mathematics Method for Management) or even sometimes "scientific management" to describe the set of mathematical tools applied to management (there exists also a training curriculum on the subject of fifty days...). Some people also use the abbreviation "SciProM" for "Scientific Project Management). A common English term that became fashionable in Europe to describe this area of application is also the "Decisioneering" referring to the fact that these are tools decision support for engineers.

It must be known that in such a market dynamics this model is obviously much better than that used by the Exponential smoothing[40] option included in the Microsoft® Excel® 14.0.6123 Analysis Toolpack (obviously as we have seen it mathematically previously, they absolutely do not work on the same principles). A simple comparative observation of the results is enough to realize it.

The parameters b and r are two parameters of the model such as:

$$b>0 \qquad 0<r\leq 1 \tag{15.3.1565}$$

and the inflection point is always giving the cumulated 50% of saturation value $\hat{Y}_{\max}$. The result is then a S-curve of the following type (we will see where it comes from further below):

---

[40] i.e. simple exponential smoothing

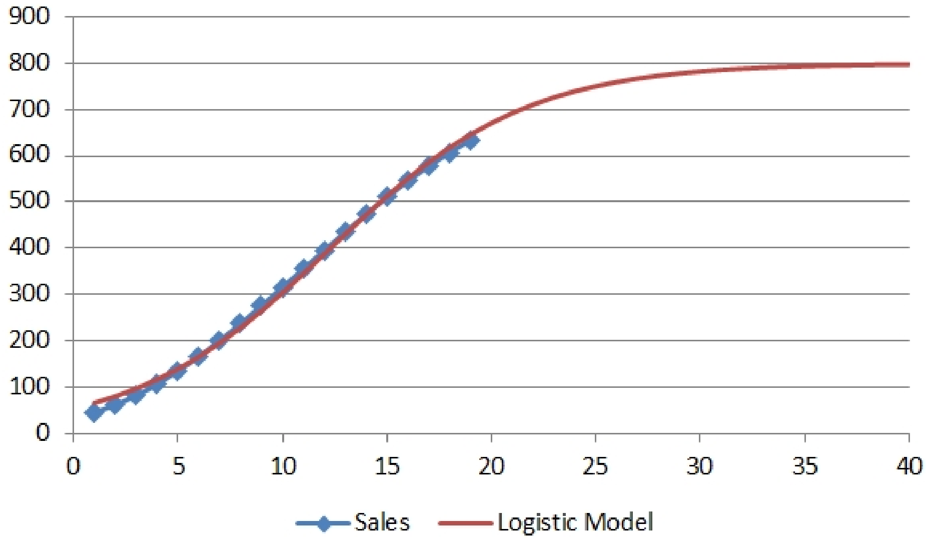
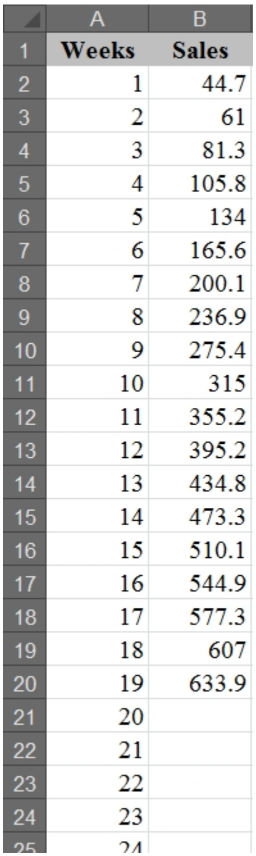


Figure 15.254: **Logistic S-curve characteristics with Microsoft® Excel® 14.0.6123**

where in blue has been represented the actual data of a product and in red the associated estimated theoretical model with its forecast and asymptotic value.

To determine the equation of the logistic model, we can directly use the solver of some softwares. But they may need to have the initial data close to the theoretical value. We will first show how these starting values can be determined with an example.

Let us consider the following first 24 rows in Microsoft® Excel® 11.8346 (the sales are in units of hundreds of thousands):



	A	B
1	Weeks	Sales
2	1	44.7
3	2	61
4	3	81.3
5	4	105.8
6	5	134
7	6	165.6
8	7	200.1
9	8	236.9
10	9	275.4
11	10	315
12	11	355.2
13	12	395.2
14	13	434.8
15	14	473.3
16	15	510.1
17	16	544.9
18	17	577.3
19	18	607
20	19	633.9
21	20	
22	21	
23	22	
24	23	

Figure 15.255: Logisttic initial values in Microsoft® Excel® 14.0.6123

and the associated graph:

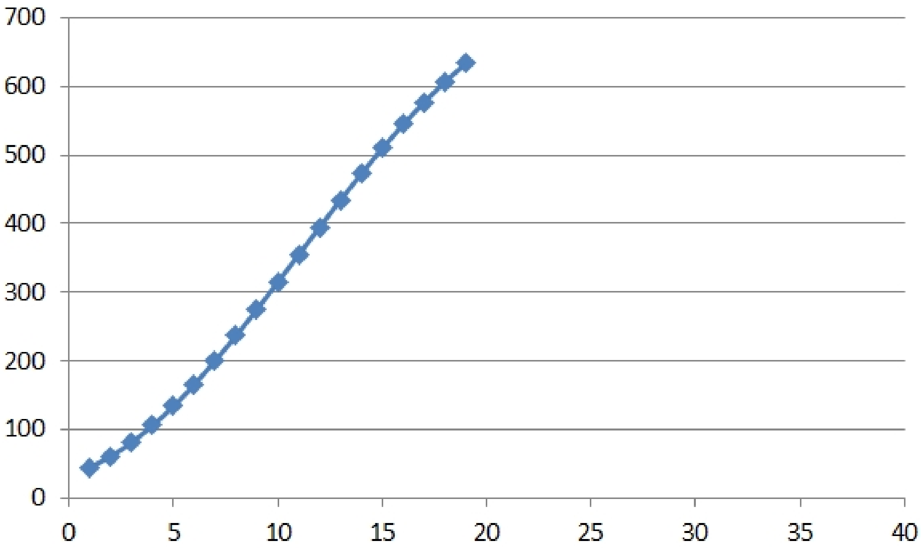


Figure 15.256: Logistic initial values plot with Microsoft® Excel® 14.0.6123

which could be considered as linear (locally) following the time at which begins the descriptive analysis of the sales.

To determine the theoretical model, we will linearise the logistic equation using a hypothetical threshold (market sales targets) of $\hat{Y}_{\max}=800$.

So:

$$Y_t=\frac{\hat{Y}_{\max}}{1+br^t}\Rightarrow 1+br^t=\frac{\hat{Y}_{\max}}{Y_t}\Rightarrow br^t=\frac{\hat{Y}_{\max}}{Y_t}-1\Rightarrow \ln\left(\frac{\hat{Y}_{\max}}{Y_t}-1\right)=\ln(b)+t\ln(r) \tag{15.3.1566}$$

That is to say we have the new following variable to explain:

$$Y_t'=\ln\left(\frac{\hat{Y}_{\max}}{Y_t}-1\right)=\ln\left(\frac{800}{Y_t}-1\right) \tag{15.3.1567}$$

And for which we will write the corresponding linear model:

$$Y_t'=a_1t+a_0+\varepsilon_t \tag{15.3.1568}$$

with therefore:

$$a_1=\ln(r)\Rightarrow r=e^{a_1} \tag{15.3.1569}$$

$$a_0=\ln(b)\Rightarrow b=e^{a_0} \tag{15.3.1570}$$

Finally injecting in the initial expression we get:

$$Y_t=\frac{\hat{Y}_{\max}}{1+br^t}=\frac{\hat{Y}_{\max}}{1+e^{a_0}e^{a_1t}}=\frac{\hat{Y}_{\max}}{1+e^{a_1t+a_0}} \tag{15.3.1571}$$

So we see here that we fall back typically on an expression similar to that of the Logistic regression as seen in the section of Numerical Methods at page 6449.

In our example the linear regression (see section Numerical Methods page 6348) gives:

$$Y'_t=a_1t+a_0+\varepsilon_t=-0.224t+2.7908 \tag{15.3.1572}$$

That is:

$$r=e^{a_1}=e^{-0.224}\cong 0.7988 \qquad b=e^{a_0}=e^{2.7908}\cong 16.293 \tag{15.3.1573}$$

Then we have immediately:

$$Y_t\cong\frac{800}{1+16.294\cdot 0.7988^t} \tag{15.3.1574}$$

That gives graphically:

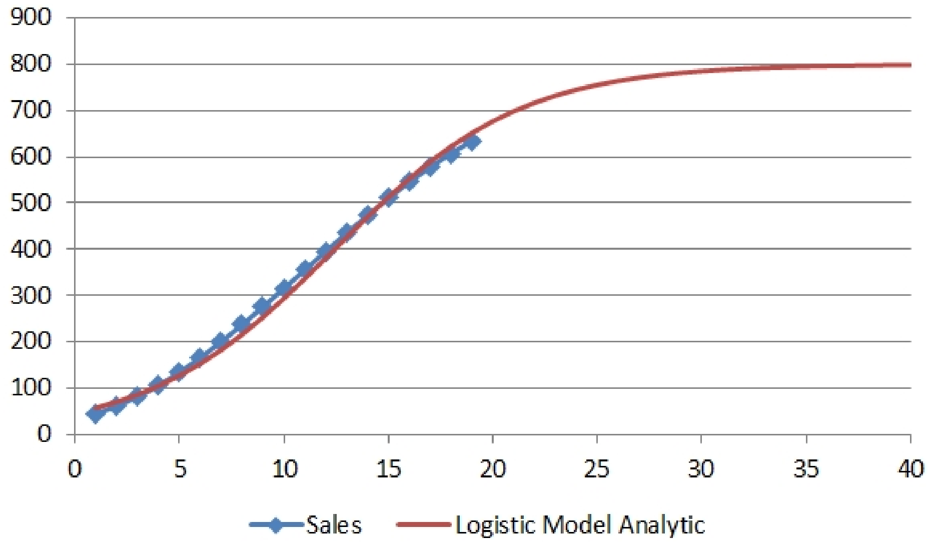
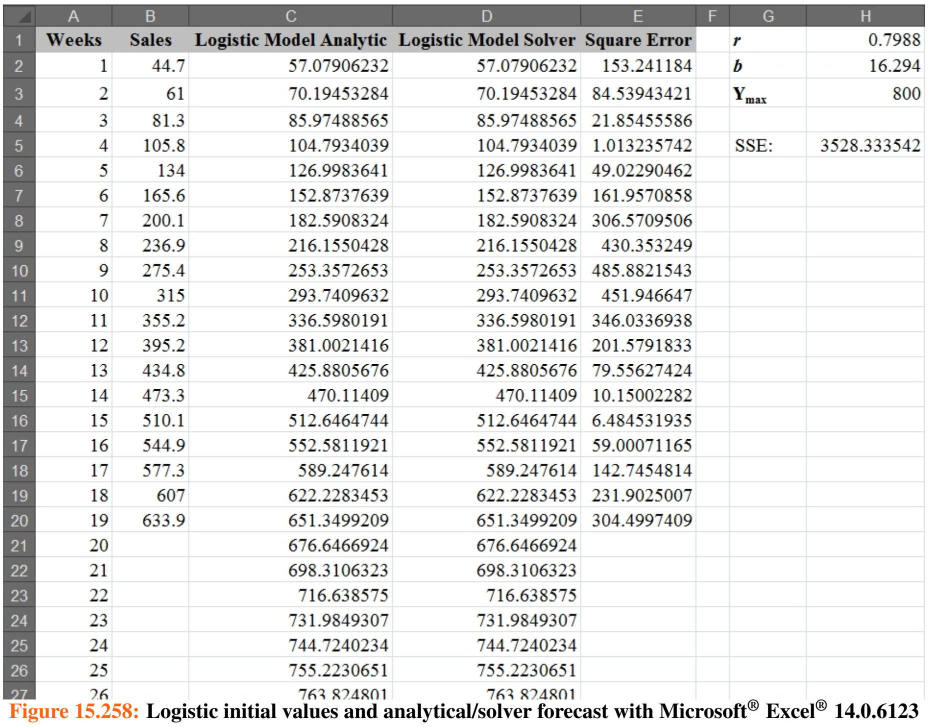


Figure 15.257: **Logistic initial values and analytical forecast plot with Microsoft® Excel® 14.0.6123**

with this analytical approach, we have a sum of squared error (SSE) between the real values and the model equal to:

$$\text{SSE} \cong 3528 \tag{15.3.1575}$$

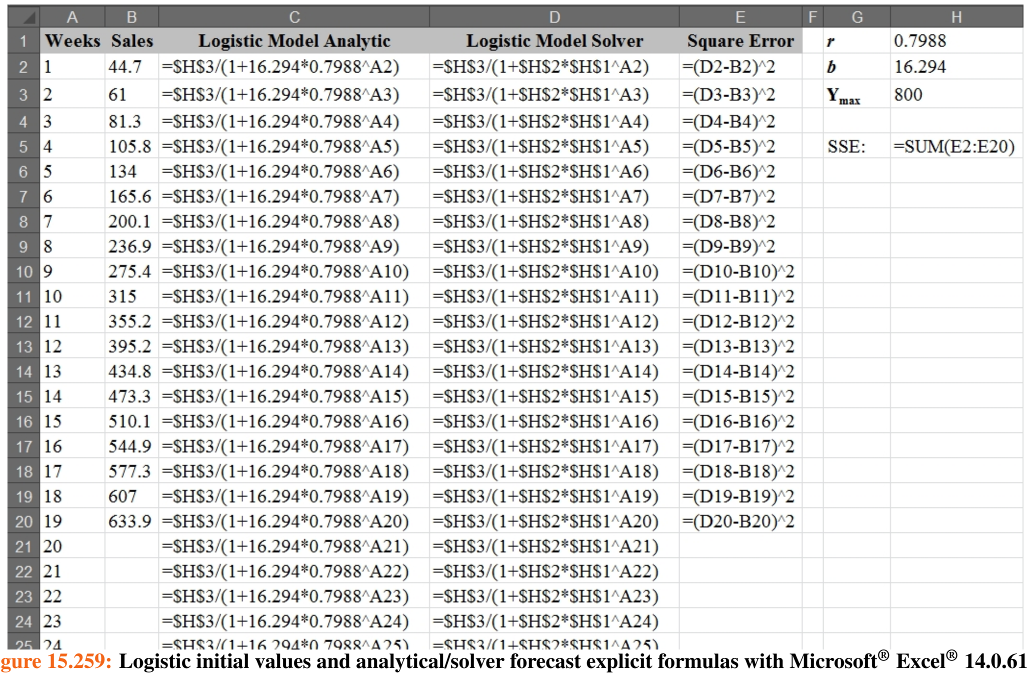
Now let us enter that data into Microsoft® Excel® 14.0.6123 as follows:



	A	B	C	D	E	F	G	H
1	**Weeks**	**Sales**	**Logistic Model Analytic**	**Logistic Model Solver**	**Square Error**		***r***	0.7988
2	1	44.7	57.07906232	57.07906232	153.241184		***b***	16.294
3	2	61	70.19453284	70.19453284	84.53943421		$\mathbf{Y}_{\mathbf{max}}$	800
4	3	81.3	85.97488565	85.97488565	21.85455586			
5	4	105.8	104.7934039	104.7934039	1.013235742		SSE:	3528.333542
6	5	134	126.9983641	126.9983641	49.02290462			
7	6	165.6	152.8737639	152.8737639	161.9570858			
8	7	200.1	182.5908324	182.5908324	306.5709506			
9	8	236.9	216.1550428	216.1550428	430.353249			
10	9	275.4	253.3572653	253.3572653	485.8821543			
11	10	315	293.7409632	293.7409632	451.946647			
12	11	355.2	336.5980191	336.5980191	346.0336938			
13	12	395.2	381.0021416	381.0021416	201.5791833			
14	13	434.8	425.8805676	425.8805676	79.55627424			
15	14	473.3	470.11409	470.11409	10.15002282			
16	15	510.1	512.6464744	512.6464744	6.484531935			
17	16	544.9	552.5811921	552.5811921	59.00071165			
18	17	577.3	589.247614	589.247614	142.7454814			
19	18	607	622.2283453	622.2283453	231.9025007			
20	19	633.9	651.3499209	651.3499209	304.4997409			
21	20		676.6466924	676.6466924				
22	21		698.3106323	698.3106323				
23	22		716.638575	716.638575				
24	23		731.9849307	731.9849307				
25	24		744.7240234	744.7240234				
26	25		755.2230651	755.2230651				
27	26		763.824801	763.824801				

Figure 15.258: Logistic initial values and analytical/solver forecast with Microsoft® Excel® 14.0.6123

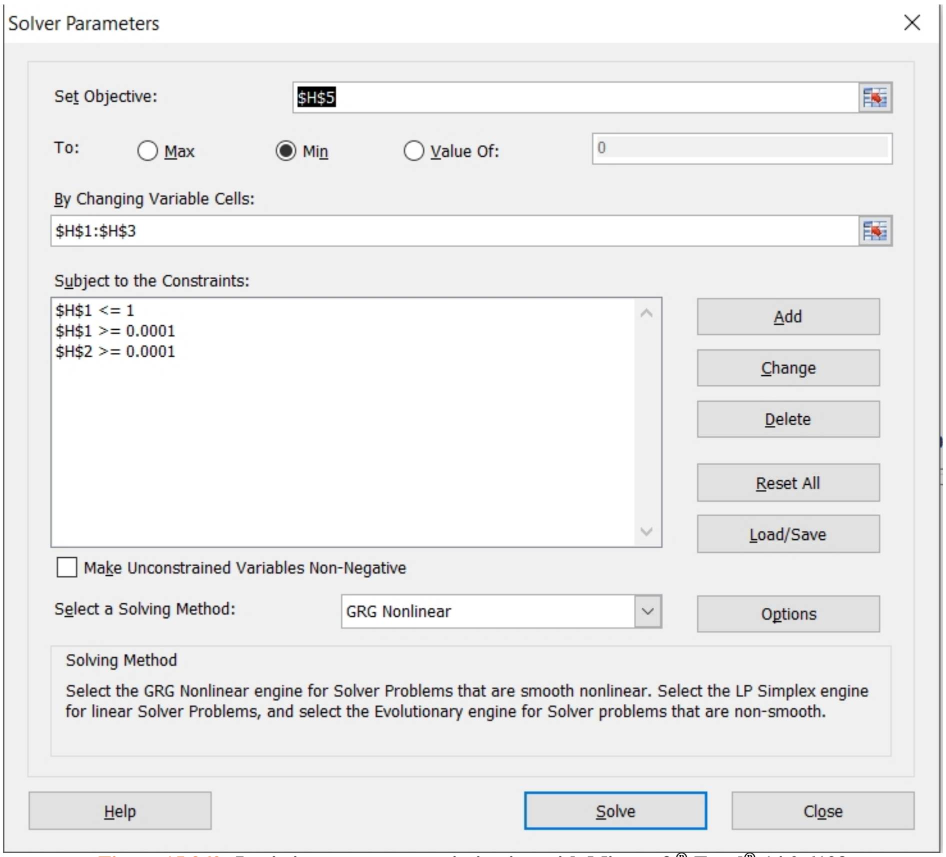
Thus explicitly:



	A	B	C	D	E	F	G	H
1	**Weeks**	**Sales**	**Logistic Model Analytic**	**Logistic Model Solver**	**Square Error**		***r***	0.7988
2	1	44.7	=$H$3/(1+16.294*0.7988^A2)	=$H$3/(1+$H$2*$H$1^A2)	=(D2-B2)^2		***b***	16.294
3	2	61	=$H$3/(1+16.294*0.7988^A3)	=$H$3/(1+$H$2*$H$1^A3)	=(D3-B3)^2		$\mathbf{Y}_{\mathbf{max}}$	800
4	3	81.3	=$H$3/(1+16.294*0.7988^A4)	=$H$3/(1+$H$2*$H$1^A4)	=(D4-B4)^2			
5	4	105.8	=$H$3/(1+16.294*0.7988^A5)	=$H$3/(1+$H$2*$H$1^A5)	=(D5-B5)^2		SSE:	=SUM(E2:E20)
6	5	134	=$H$3/(1+16.294*0.7988^A6)	=$H$3/(1+$H$2*$H$1^A6)	=(D6-B6)^2			
7	6	165.6	=$H$3/(1+16.294*0.7988^A7)	=$H$3/(1+$H$2*$H$1^A7)	=(D7-B7)^2			
8	7	200.1	=$H$3/(1+16.294*0.7988^A8)	=$H$3/(1+$H$2*$H$1^A8)	=(D8-B8)^2			
9	8	236.9	=$H$3/(1+16.294*0.7988^A9)	=$H$3/(1+$H$2*$H$1^A9)	=(D9-B9)^2			
10	9	275.4	=$H$3/(1+16.294*0.7988^A10)	=$H$3/(1+$H$2*$H$1^A10)	=(D10-B10)^2			
11	10	315	=$H$3/(1+16.294*0.7988^A11)	=$H$3/(1+$H$2*$H$1^A11)	=(D11-B11)^2			
12	11	355.2	=$H$3/(1+16.294*0.7988^A12)	=$H$3/(1+$H$2*$H$1^A12)	=(D12-B12)^2			
13	12	395.2	=$H$3/(1+16.294*0.7988^A13)	=$H$3/(1+$H$2*$H$1^A13)	=(D13-B13)^2			
14	13	434.8	=$H$3/(1+16.294*0.7988^A14)	=$H$3/(1+$H$2*$H$1^A14)	=(D14-B14)^2			
15	14	473.3	=$H$3/(1+16.294*0.7988^A15)	=$H$3/(1+$H$2*$H$1^A15)	=(D15-B15)^2			
16	15	510.1	=$H$3/(1+16.294*0.7988^A16)	=$H$3/(1+$H$2*$H$1^A16)	=(D16-B16)^2			
17	16	544.9	=$H$3/(1+16.294*0.7988^A17)	=$H$3/(1+$H$2*$H$1^A17)	=(D17-B17)^2			
18	17	577.3	=$H$3/(1+16.294*0.7988^A18)	=$H$3/(1+$H$2*$H$1^A18)	=(D18-B18)^2			
19	18	607	=$H$3/(1+16.294*0.7988^A19)	=$H$3/(1+$H$2*$H$1^A19)	=(D19-B19)^2			
20	19	633.9	=$H$3/(1+16.294*0.7988^A20)	=$H$3/(1+$H$2*$H$1^A20)	=(D20-B20)^2			
21	20		=$H$3/(1+16.294*0.7988^A21)	=$H$3/(1+$H$2*$H$1^A21)				
22	21		=$H$3/(1+16.294*0.7988^A22)	=$H$3/(1+$H$2*$H$1^A22)				
23	22		=$H$3/(1+16.294*0.7988^A23)	=$H$3/(1+$H$2*$H$1^A23)				
24	23		=$H$3/(1+16.294*0.7988^A24)	=$H$3/(1+$H$2*$H$1^A24)				
25	24		=$H$3/(1+16.294*0.7988^A25)	=$H$3/(1+$H$2*$H$1^A25)				

Figure 15.259: Logistic initial values and analytical/solver forecast explicit formulas with Microsoft® Excel® 14.0.6123

If we run the Microsoft® Excel® 14.0.6123 Solver with the following parameters (we have put 0.0001 as lowest value since the solver does not offer a strict order relation):



Solver Parameters

Set Objective: $H$5

To: ○ Max ● Min ○ Value Of: 0

By Changing Variable Cells:
$H$1:$H$3

Subject to the Constraints:
$H$1 <= 1
$H$1 >= 0.0001
$H$2 >= 0.0001

☐ Make Unconstrained Variables Non-Negative

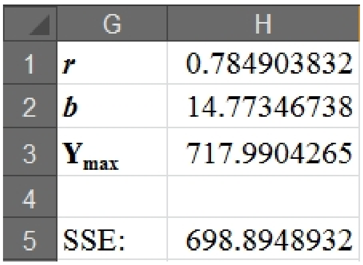
Select a Solving Method: GRG Nonlinear

Solving Method

Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

Figure 15.260: **Logistic parameters optimization with Microsoft® Excel® 14.0.6123**

Which gives:



	G	H
1	***r***	0.784903832
2	***b***	14.77346738
3	$\mathbf{Y}_{\mathbf{max}}$	717.9904265
4		
5	SSE:	698.8948932

Figure 15.261: **Optimized logistic parameters with Microsoft® Excel® 14.0.6123**

Therefore:

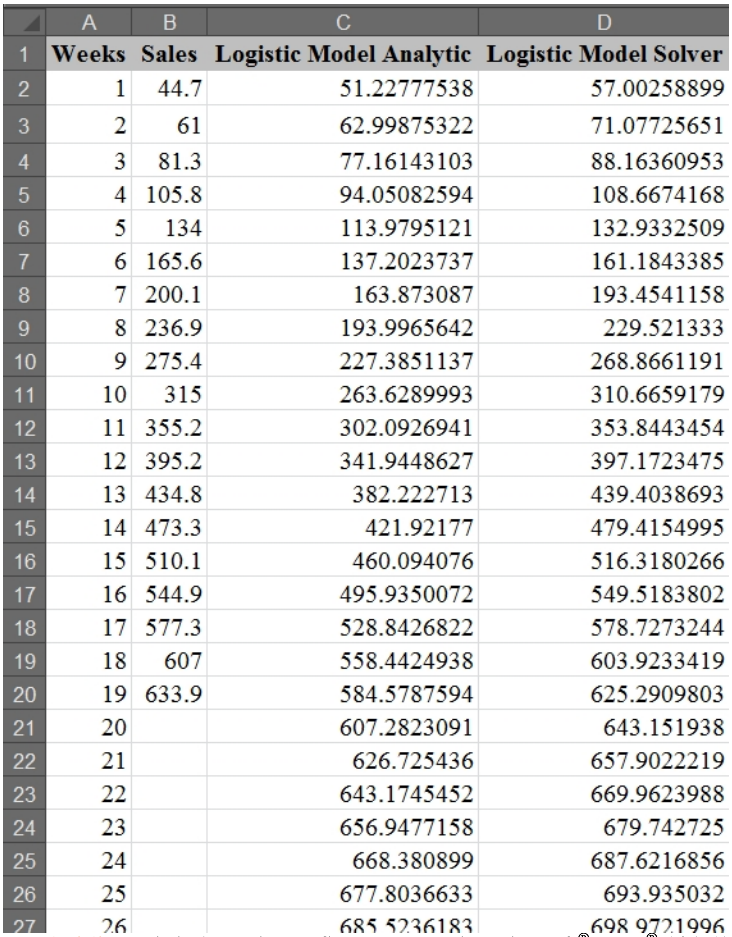
$$\hat{Y}_t \cong \frac{717.99}{1+14.773\cdot 0.78490^t} \tag{15.3.1576}$$

with:

$$\text{SSE} \cong 698 \tag{15.3.1577}$$

What is quite good with the solver approach is that we get $\hat{Y}_{\text{max}}$ and this is a very important information for trendy product type for marketing and business analysts!!!

As we can see with the solver approach the SSE is quite much lower than the method using linear regression method. Indeed, let us see the following list:



	A	B	C	D
1	**Weeks**	**Sales**	**Logistic Model Analytic**	**Logistic Model Solver**
2	1	44.7	51.22777538	57.00258899
3	2	61	62.99875322	71.07725651
4	3	81.3	77.16143103	88.16360953
5	4	105.8	94.05082594	108.6674168
6	5	134	113.9795121	132.9332509
7	6	165.6	137.2023737	161.1843385
8	7	200.1	163.873087	193.4541158
9	8	236.9	193.9965642	229.521333
10	9	275.4	227.3851137	268.8661191
11	10	315	263.6289993	310.6659179
12	11	355.2	302.0926941	353.8443454
13	12	395.2	341.9448627	397.1723475
14	13	434.8	382.222713	439.4038693
15	14	473.3	421.92177	479.4154995
16	15	510.1	460.094076	516.3180266
17	16	544.9	495.9350072	549.5183802
18	17	577.3	528.8426822	578.7273244
19	18	607	558.4424938	603.9233419
20	19	633.9	584.5787594	625.2909803
21	20		607.2823091	643.151938
22	21		626.725436	657.9022219
23	22		643.1745452	669.9623988
24	23		656.9477158	679.742725
25	24		668.380899	687.6216856
26	25		677.8036633	693.935032
27	26		685.5236183	698.9721996

Figure 15.262: **Logistic Analytical vs Solver values with Microsoft® Excel® 14.0.6123**

Which gives visually:

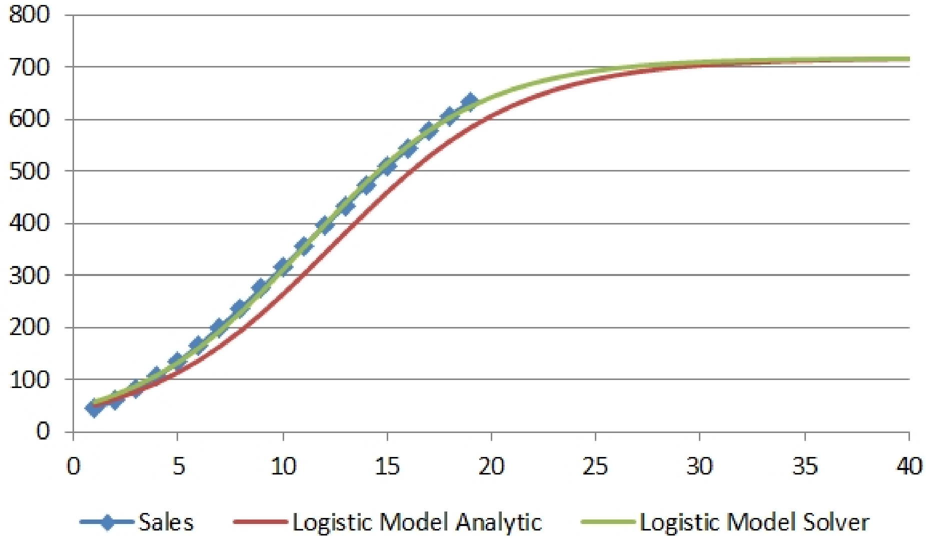


Figure 15.263: **Logistic Analytical vs Solver plot with Microsoft® Excel® 14.0.6123**

We see clearly that using the solver (numerical model) is better than the analytical model given indirectly by the linear regression.

To end this part on deterministic forecasting models, remember that they must be monitored regularly to ensure that the model and parameters used are still appropriate. As the reader may have guessed, a forecasting model must not have bias. The sum of the errors must be near zero, with sometimes overestimated and underestimated values. When a model tends to always overestimate (or underestimate), the latter has a bias and it must be revised (to monitor this we can use the control charts techniques as studied in the section of Industrial Engineering at page 8934).

It is also necessary for the conclusion to remember that there is sometimes a significant difference between the observed values (actual values) and the observed that would really occur if the supply (offer) was infinite. Indeed, observed sales do not take into account missed sales because lack of stock in one or many of the thousands of subsidiaries that have your company, and therefore the possible option that a customer went buy at competitor product. Thus, sales observed are most of times a "minimum" because of missed opportunities!

Remark

Let us indicate that it is trivially possible to make the difference of all subsequent points in time of a time series and then to make a histogram to determine the probability distribution of changes, allowing to make statistical inference with all necessary precautions (analysis that should always be done in practice!).

15.3.13.4 Autoregressive Models

In statistics and signal processing, an autoregressive (AR) model is a representation of a type of random process; as such, it describes certain time-varying processes in nature, economics, etc. The autoregressive model specifies that the output variable depends linearly on its own previous values and on a stochastic term (an imperfectly predictable term); thus the model is in the form of a stochastic difference equation.

Together with the moving-average (MA) model, it is a special case and key component of the more general ARMA and ARIMA models of time series (see further below), which have a more complicated stochastic structure; it is also a special case of the vector autoregressive model (VAR) as we will see further below, which consists of a system of more than one stochastic difference equation.

In the study of a time series, it is natural to think that the series value at time t may depend in part on the values taken on the previous dates such that:

$$X_t = f(X_{t-1}, X_{t-2}, \ldots) \tag{15.3.1578}$$

which of course is for now a limited approach to only discrete univariate series (we will study multivariate series further below) and for which the sampling frequencies constant (at the opposite of what is done in some areas of financial transactions).

Autoregressive models have become such common that they are now integrated in Microsoft® Excel® 2016 (in fact Microsoft® Excel® implement an additive ETS model but as we will see later, ETS additive models are equivalent to ARIMA models):

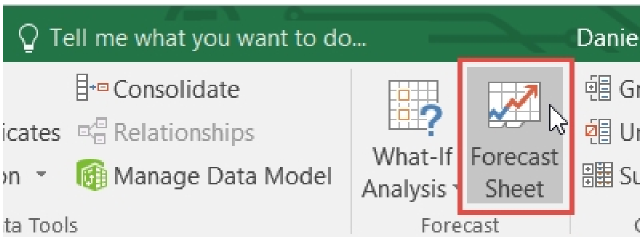
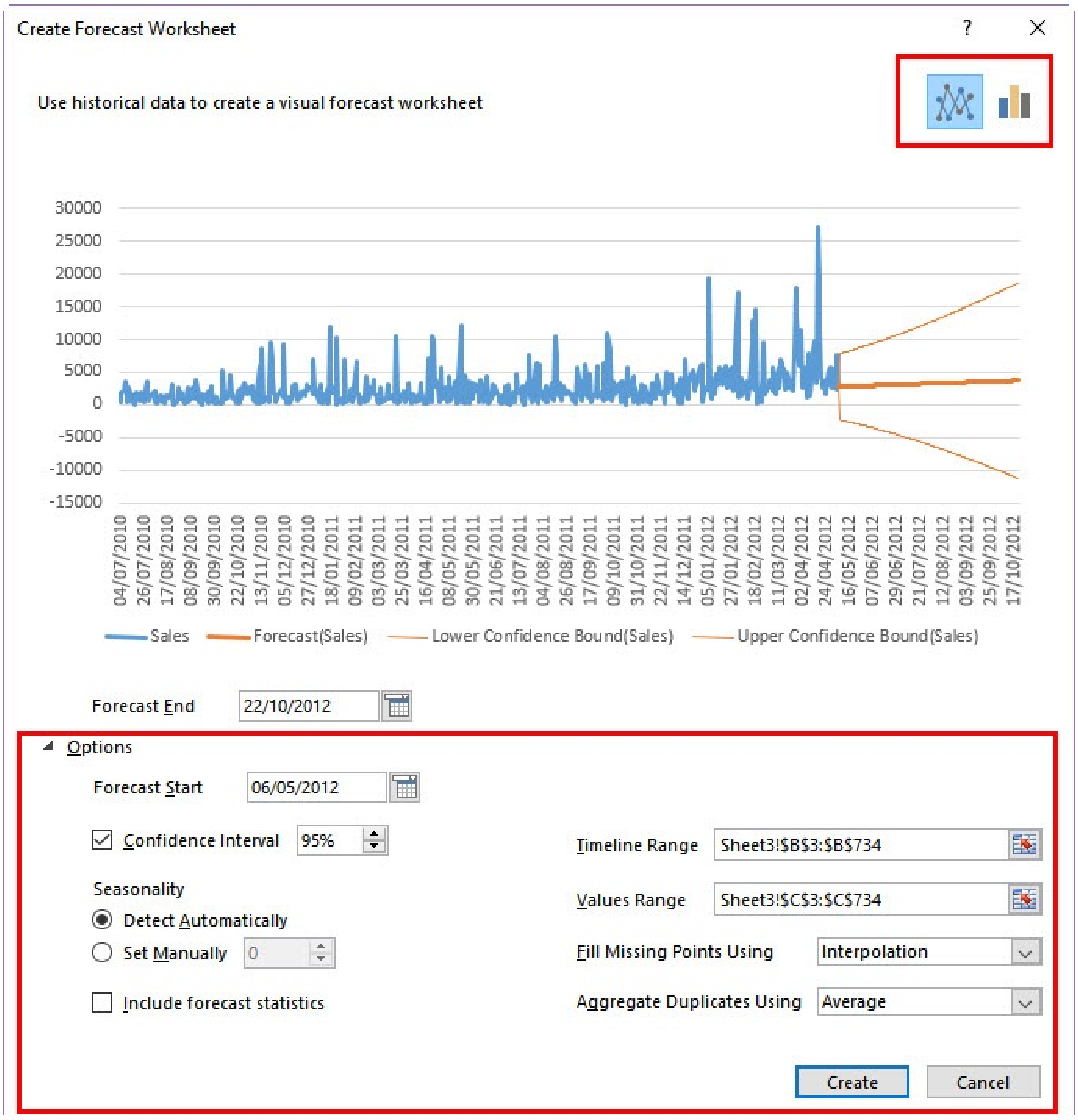

Tell me what you want to do...
Danie
icates
on
Consolidate
Relationships
Manage Data Model
ta Tools
What-If Analysis
Forecast Sheet
Forecast
Gr
Ur
Su

Figure 15.264: **Microsoft® Excel® 2016 SARIMA forecast native tool button**

that typically gives for result:


Create Forecast Worksheet
?
X
30000
25000
20000
15000
10000
5000
0
-5000
-10000
-15000
04/07/2010
26/07/2010
17/08/2010
08/09/2010
30/09/2010
22/10/2010
13/11/2010
05/12/2010
27/12/2010
18/01/2011
09/02/2011
03/03/2011
25/03/2011
16/04/2011
08/05/2011
30/05/2011
21/06/2011
13/07/2011
04/08/2011
26/08/2011
17/09/2011
09/10/2011
31/10/2011
22/11/2011
14/12/2011
05/01/2012
27/01/2012
18/02/2012
11/03/2012
02/04/2012
24/04/2012
16/05/2012
07/06/2012
29/06/2012
21/07/2012
12/08/2012
03/09/2012
25/09/2012
17/10/2012
Sales
Forecast(Sales)
Lower Confidence Bound(Sales)
Upper Confidence Bound(Sales)
Options
Create
Cancel

Use historical data to create a visual forecast worksheet

Forecast End 22/10/2012

Forecast Start 06/05/2012

☑ Confidence Interval 95%

Seasonality

● Detect Automatically

○ Set Manually 0

☐ Include forecast statistics

Timeline Range Sheet3!$B$3:$B$734

Values Range Sheet3!$C$3:$C$734

Fill Missing Points Using Interpolation

Aggregate Duplicates Using Average

Figure 15.265: **Microsoft® Excel® 2016 forecast native tool demo**

Before focusing on AR models remember that we saw in the section Probabilities (see page 537) that when we focus only a dependency at time $t-1$, then we speak of "Markov chain", or more rigorously of "order 1 Markov chain" and... let us open a small parenthesis on the interest of Markov chains (recall oriented for the finance field)...

Let us consider that we have sampled stock prices at a fixed frequency (eg daily or hourly ... whatever!). It is then very easy to calculate the proportion of increases (A) versus that of Decreases (D) of the values of interest. Thus, taking into account the whole set of history, we will typically have a table of the following type (values derived from the S&P 500 indices between 11947 and 12007 according to holocene calendar):

Proportion Increases (A)	**Proportion Decreases** (D)	**Total**
0.474	0.526	1

Table 15.62: **Proportion of A/D variations on a history set**

But this is not quite useful excepted perhaps to run a hypothesis test on the both proportions... If we complicate a little bit the analysis by asking us how many times we have two successive increases (AA), two successive decreases (DD) or alternating events (DA) and (AD) throughout history this doubles the work and could give us a typical table of the following type:

$X_i =$	(A)	(B)	**Total**
$X_{i-1} = (D)$	0.519	0.481	1
$X_{i-1} = (A)$	0.4.33	0.567	1

Table 15.63: **Proportion of A/D variations relatively to previous period**

So the above values can obviously be seen as conditional probabilities:

$$P(X_i|X_{i-1}) \tag{15.3.1579}$$

And if now we do an analysis not only on the previous period but the last two, it means that we calculate the proportions of the sequences (DDD), (DDA), (DAD), (DAA), (ADD), (ADA), (AAD), (AAA) and it doubles the number of calculations:

		$X_i = 0$		
X_{i-1}	X_{i-1}	(A)	(D)	**Total**
(D)	(D)	0.501	0.499	1
(D)	(A)	0.412	0.588	1
(A)	(D)	0.539	0.461	1
(A)	(A)	0.449	0.551	1

Table 15.64: **Proportion of A/D variations relatively to 2 previous periods**

The above values must obviously be seen as conditional probabilities:

$$P(X_i|X_{i-1},X_{i-2}) \tag{15.3.1580}$$

We see that each time we add an additional period to the analysis, we double the number of calculations. Thus, taking the last 20 periods, this leads us to nearly 1 million conditional probabilities (proportions)!!! So we understand then the interest to restrict ourselves to a small number of previous values or a single value as do the Markov chains.

Coming back to our example, we have then that the probability that the next 4 days the sequence is $(ADAA)$ will be without simplifying assumption:

$$\begin{aligned}P(AADA)=&P(X_1=A)\cdot P(X_2=A|X_1=A)\cdot P(X_3=D|X_2=A,X_1=A)\\&\cdot P(X_4=A|X_3=D,X_2=A,X_1=A)\end{aligned} \qquad (15.3.1581)$$

but will be reduced in the hypothesis of a Markov chain (that is to say a dependency only with the previous period) to:

$$\begin{aligned}P(AADA)=&P(X_1=A)P(X_2=A|X_1=A)P(X_3=D|X_2=A)P(X_4=A|X_3=D)\\&=0.526\cdot 0.567\cdot 0.433\cdot 0.481\cong 0.062=6.2\%\end{aligned} \qquad (15.3.1582)$$

Whether the approach of total conditional probabilities or the simplified Markov chain approach is the best is a broad debate and should be treated case by case. The hypothesis of random walk (implicitly: the Markov model) is defended by a majority of economists ... but this hypothesis obviously also a number of opponents.

It is therefore not usually necessary to take into account all the past values of series or only the last event and we often limit ourselves to p values:

$$X_t=f(X_{t-1},X_{t-2},\ldots,X_{t-p})+\varepsilon_t \qquad (15.3.1583)$$

where ε_t is a white noise $\mathcal{N}(0,\sigma)$ or Wiener process sometimes denoted WN. More rigorously, and for reasons of application of statistical techniques, the white noise is defined such that:

$$\begin{aligned}&\text{E}(\varepsilon_t)=0\\&\text{E}(\varepsilon_t^2)=\text{V}(\varepsilon_t)=\sigma^2\\&\text{E}(\varepsilon_t\varepsilon_\tau)=\text{cov}(\varepsilon_t,\varepsilon_\tau)=0\quad\forall t\neq\tau\end{aligned} \qquad (15.3.1584)$$

The first equality is obvious. The third one is required a constraint. The second has not been proved in the section Statistics. So let us prove this now!

So remember that in the section Statistics we have proved the Huygens relation (see section Statistics page 593):

$$\text{V}(X)=\text{E}(X^2)-(\text{E}(X))^2 \qquad (15.3.1585)$$

Solving for the needed quantity gives:

$$\text{E}(X^2)=\text{V}(X)+(\text{E}(X))^2 \qquad (15.3.1586)$$

But four our case $\text{E}(X)=0$, therefore:

$$\text{E}(X^2)=\text{V}(X)=\sigma^2 \qquad (15.3.1587)$$

Hence:

$$\text{E}(\varepsilon_t^2)=\sigma^2 \qquad (15.3.1588)$$

The only difference relatively to the standard Brownian motion is that as we will see further below, there is the presence here of an inertia factor traditionally denoted α in the simple cases, that strongly influences the dynamics of the process such as:

$$X_t=\alpha X_{t-1}+\varepsilon_t \qquad (15.3.1589)$$

Indeed, as it is very easy to do in Microsoft® Excel® in accordance with the procedure indicated in our study of Wiener process earlier in this section.

Here are various plots of the time series according to some values of the inertia factor:

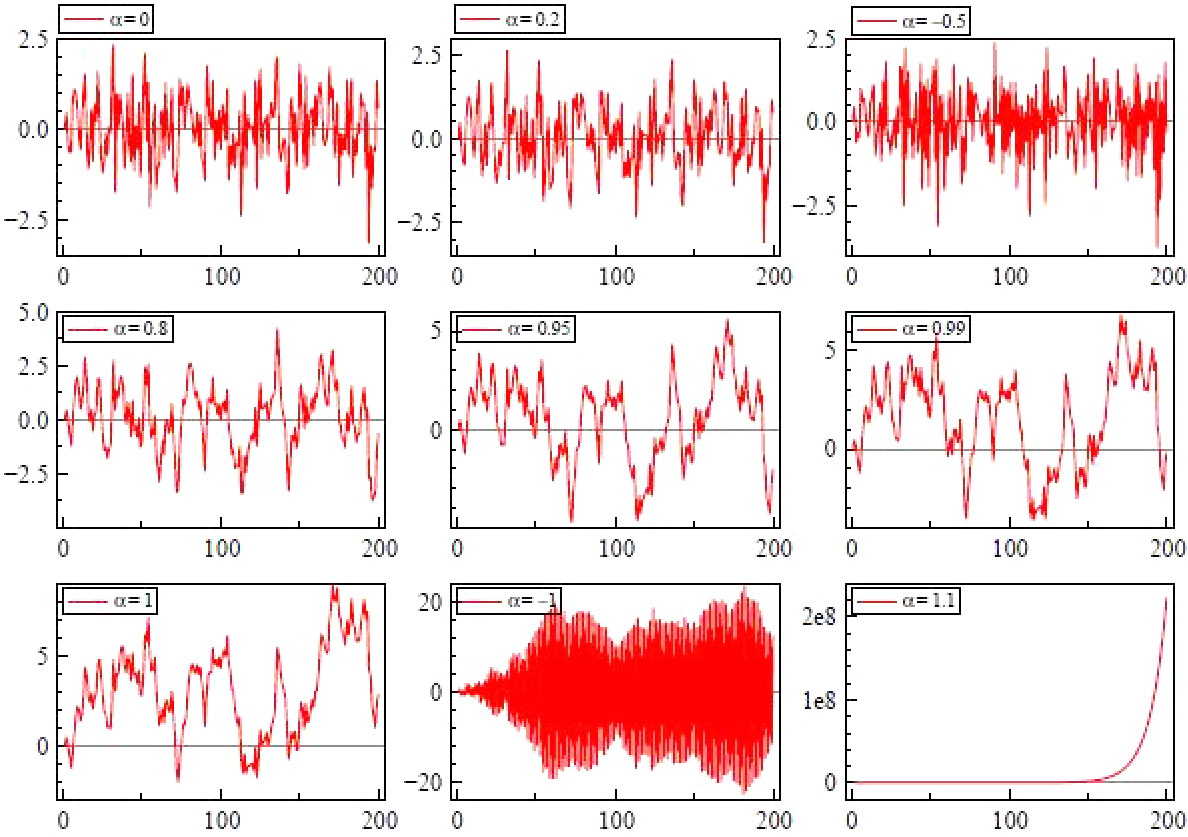


Figure 15.266: **Autoregressive process of order** 1

If we consider X_t as a specific random variable (with a given density function) that we would denote by X and X_{t-h} as another specific random variable (with a given density function) that we would denote by Y, then nothing prevents us being known the density functions of each of these variables to compute their covariance denoted most of times $\gamma_t(h)$ and named "autocovariance function" (ACVF) in the field of times series analysis:

$$\boxed{\gamma_t(h)=c_{X,Y}=\text{cov}(X,Y)=\text{cov}(X_t,X_{t-h})=\text{E}(X_tX_{t-h})-\text{E}(X_t)\text{E}(X_{t-h})} \quad (15.3.1590)$$

Let us see two important properties for the autocovariance:

P1. The first properties is already known to us (from our study of statistics):

$$\gamma_t(0)=\text{V}(X_t) \quad (15.3.1591)$$

P2. If we consider a process such that $\text{E}(X_t)=c^{te}$, then considering without loss of generality $\text{E}(X_t)=0$, we then have:

$$\text{E}(X_{t_1+h}X_{t_1})=\gamma_{t_1}(h) \quad (15.3.1592)$$

for all t_1 and t_1+h contained in the index set. Now let us set $t_0=t_1-h$. Then:

For example, in practice we often know the mean μ of the two random variables at the two different times and also some of the values of their underlying distributions (random issues). So, it becomes easy to calculate their covariance. But this is not really a useful indicator. Nothing prevents us assuming a linear relation to use the linear correlation coefficient:

$$R_{X,Y}=\frac{c_{X,Y}}{\sqrt{\mathrm{V}(X)\mathrm{V}(Y)}}=\frac{c_{X,Y}}{\sqrt{\mathrm{cov}(X,X)\mathrm{cov}(Y,Y)}}=\frac{c_{X,Y}}{\sqrt{c_{X,X}c_{Y,Y}}} \tag{15.3.1593}$$

But that is written traditionally and trivially in the field of time series analysis:

$$\boxed{\rho_t(h)=\frac{\gamma_t(h)}{\sqrt{\gamma_t(0)\gamma_{t-h}(0)}}} \tag{15.3.1594}$$

and is named "autocorrelation coefficient" often abbreviated ACF (AutoCorrelation Function) in statistical softwares.

Remarks

R1. There are different variations of this relation in some statistical softwares. Therefore we get sometimes small numerical differences with respect to the use of the above ACF analysis.

R2. Keep in mind that calculate the average, variance and also by extension the autocorrelation makes sense only for stationary processes! Calculating these statistical indicators for non-stationary processes is generally a huge error.

We prefer to write the above relation in the following explicit form (the covariance is explicit):

$$\rho_{X,Y,t}=\frac{\mathrm{E}\left[(X_t-\mu_{X,t})(Y_t-\mu_{Y,t})\right]}{\sigma_X\sigma_Y} \tag{15.3.1595}$$

The reader should know that when we do the analysis of time series in practice, we often compare a series with itself but with a time lag h. The autocorrelation of order h is then the correlation between the series and itself delayed by h lags.

Thus, as both series are dependent and have the same statistical characteristics, we write in this special case say to be a "second order stationary process" the following relation named the "empirical autocorrelation coefficient":

$$\begin{aligned}\rho(h)=\frac{\mathrm{E}\left[(X_t-\mu_X)(X_{t-h}-\mu_X)\right]}{\sigma_X\sigma_X}=\frac{\gamma_t(h)}{\gamma_t(0)}&=\frac{\dfrac{1}{n-1}\displaystyle\sum_{t=h+1}^{n}(X_t-\mu_X)(X_{t-h}-\mu_X)}{\dfrac{1}{n-1}\displaystyle\sum_{t=h+1}^{n}(X_t-\mu_X)^2}\\&=\frac{\displaystyle\sum_{t=h+1}^{n}(X_t-\mu_X)(X_{t-h}-\mu_X)}{\sigma_X^2}\end{aligned} \tag{15.3.1596}$$

where obviously we must not forget that in practice the mean and the standard deviation are only estimators although this is not explicit in the above definition since the in the field of time series analysis we indicate rarely the estimators with hat symbol as we do in Statistics or Reliability Analysis.

That latter relation is also often written en the following equivalent form:

$$\rho(h)=\frac{\mathrm{c}(X_t,X_{t+h})}{\hat{\mathrm{V}}(X_t)}=\frac{\dfrac{1}{T-1}\displaystyle\sum_{t=1}^{T-h}(X_t-\bar{X})(X_{t+h}-\bar{X})}{\hat{\mathrm{V}}(X_t)} \tag{15.3.1597}$$

Obviously in the case of the analysis of time series where we suspect cyclicality, the largest value ρ for a given h indicates the underlying cyclicality frequency (respectively the periodicity) of the series which can help by for example the choice of the type of moving average (see further below).

Remark

Notice (important case!), that if the averages are zero, the above relation resumes to:

$$\rho(h)=\frac{\mathrm{c}(X_t,X_{t+h})}{\hat{\mathrm{V}}(X_t)}=\frac{\dfrac{1}{T-1}\displaystyle\sum_{t=1}^{T-h}X_tX_{t+h}}{\dfrac{1}{T-1}\displaystyle\sum_{i=1}^{T}X_t^2}=\frac{\displaystyle\sum_{t=1}^{T-h}X_tX_{t+h}}{\displaystyle\sum_{i=1}^{T}X_t^2} \tag{15.3.1598}$$

Definition (#15.3.704): We name "correlogram", the diagram representing the autocorrelation coefficients of order $1,2,3,\ldots,h,\ldots$.

Below are given for example a family of time series:

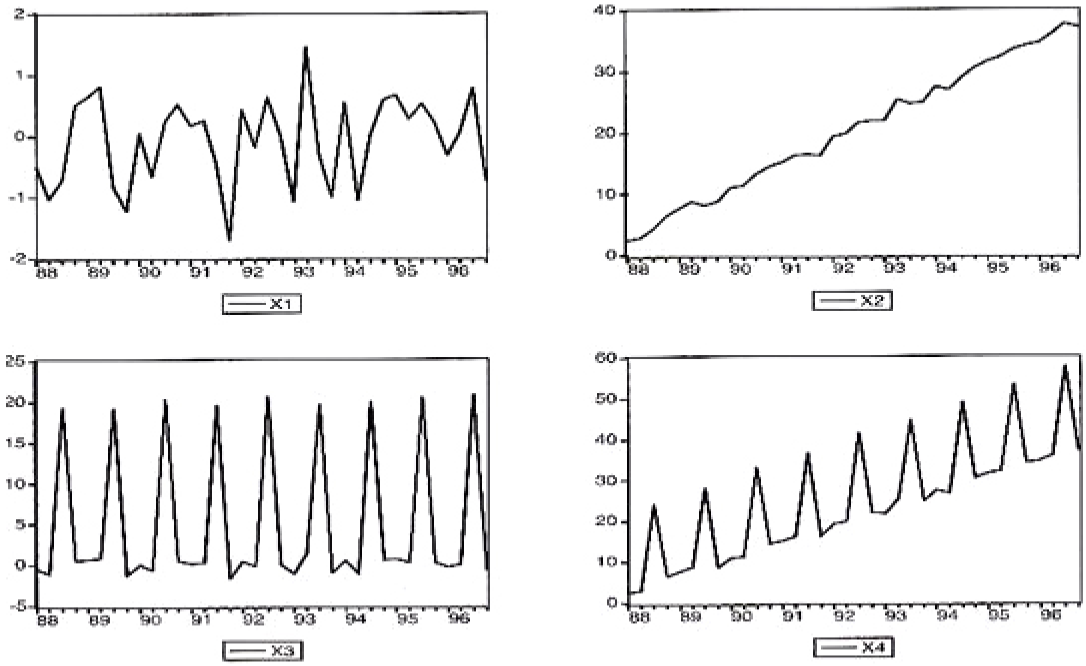


Figure 15.267: Some examples of time series

with their corresponding correlogram for different values of h on the x-axis and the values of ρ on the y-axis[41]:

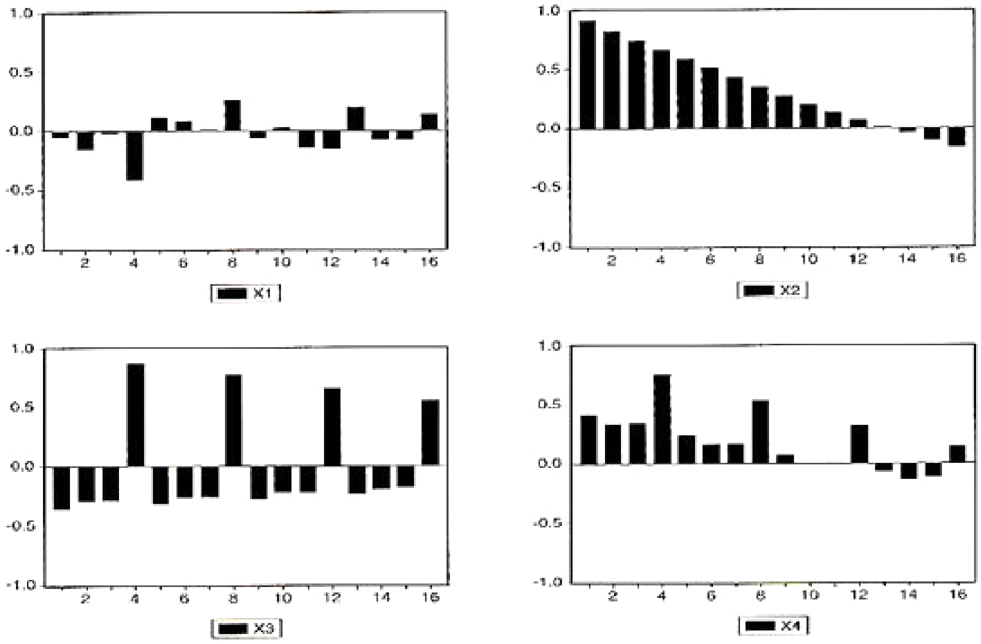


Figure 15.268: **Corresponding AFC correlograms of the time series**

Or here four examples of four non-stationary time series (a "unit-root" time series means that there is a stochastic trend):

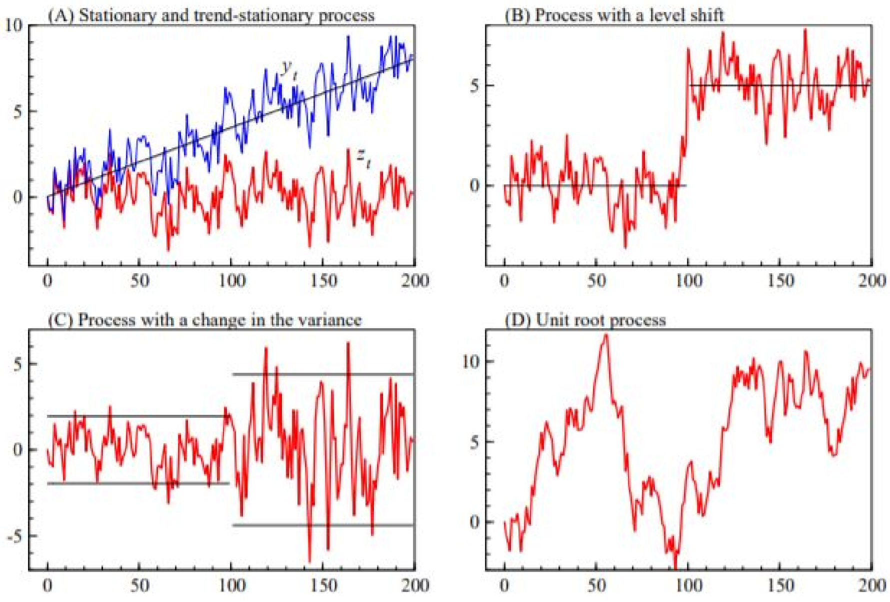


Figure 15.269: **Four non-stationary time series**

✍**Example:**

---

[41] Most softwares don't plot $\rho(0)$ as it is obviously equal to 1...

A random walk is of the form:

$$X_t = \sum_{i=1}^{t} \varepsilon_i \tag{15.3.1599}$$

Note that as $X_t = X_{t-1} + \varepsilon_t$ we then have:

$$\text{cov}\left(X_t, X_{t-1}\right) = \text{cov}\left(X_{t-1} + \varepsilon_t, X_{t-1}\right) = \text{V}\left(X_{t-1}\right) \tag{15.3.1600}$$

Also note that:

$$\sigma_t^2 = \text{V}\left(X_t\right) = t\sigma_\varepsilon^2 \tag{15.3.1601}$$

Consequently:

$$\text{ACF}_{\text{RW}}\left(X_t, X_{t-1}\right) = \frac{\sigma_{t-1}^2}{\sigma_{t-1}\sigma_t} = \frac{\sigma_{t-1}}{\sigma_t} = \sqrt{\frac{t-1}{t}} = \sqrt{1-\frac{1}{t}} \cong 1 - \frac{1}{2t} \tag{15.3.1602}$$

Which is to say we should see a correlation of almost 1 because as soon as t starts to get large, X_t and X_{t-1} are almost exactly the same thing - the relative difference between them tends to be fairly small.

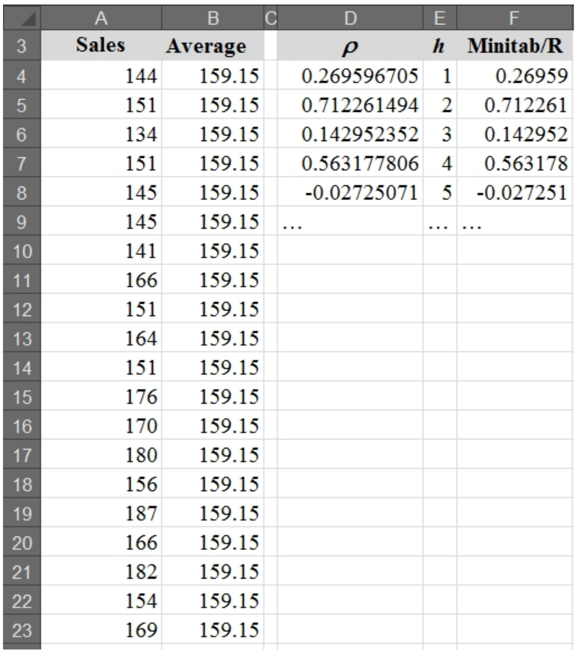
Generalizing we get (using a slightly different notation but also quite common in the literature):

$$\text{ACF}_{\text{RW}}(t,k) = \frac{\text{cov}\left(X_t, X_{t+k}\right)}{\sqrt{\text{V}\left(X_t\right)\text{V}\left(X_{t+k}\right)}} = \frac{t\sigma^2}{\sqrt{t\sigma^2(t+k)\sigma^2}} = \frac{1}{\sqrt{1+k/t}} \tag{15.3.1603}$$

Let us make a practical example of calculation of correlogram with Microsoft® Excel® 14.0.6123 and fictitious data:

✍**Example:**

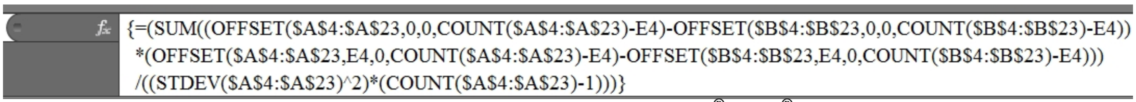
Consider the following list of data:



	A	B	C	D	E	F
3	Sales	Average		ρ	h	Minitab/R
4	144	159.15		0.269596705	1	0.26959
5	151	159.15		0.712261494	2	0.712261
6	134	159.15		0.142952352	3	0.142952
7	151	159.15		0.563177806	4	0.563178
8	145	159.15		-0.02725071	5	-0.027251
9	145	159.15		…	…	…
10	141	159.15				
11	166	159.15				
12	151	159.15				
13	164	159.15				
14	151	159.15				
15	176	159.15				
16	170	159.15				
17	180	159.15				
18	156	159.15				
19	187	159.15				
20	166	159.15				
21	182	159.15				
22	154	159.15				
23	169	159.15				

Figure 15.270: Data list for AFC example and comparison with Minitab/R in Microsoft® Excel® 14.0.6123

with the following matrix formula in the cells `D3` and below (sorry I did not found until now an other way to do it...):



```
{=(SUM((OFFSET($A$4:$A$23,0,0,COUNT($A$4:$A$23)-E4)-OFFSET($B$4:$B$23,0,0,COUNT($B$4:$B$23)-E4))
 *(OFFSET($A$4:$A$23,E4,0,COUNT($A$4:$A$23)-E4)-OFFSET($B$4:$B$23,E4,0,COUNT($B$4:$B$23)-E4)))
 /((STDEV($A$4:$A$23)^2)*(COUNT($A$4:$A$23)-1)))}
```

Figure 15.271: Autcorrelation formula in Microsoft® Excel® 14.0.6123

and visually this gives the following correlogram:

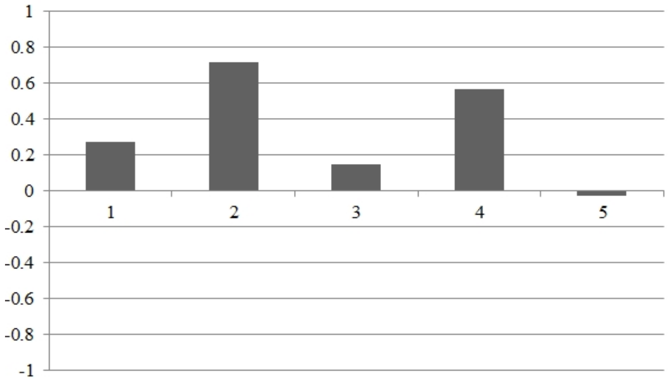


Figure 15.272: AFC Correlogram in Microsoft® Excel® 14.0.6123

Definitions (#15.3.705):

D1. A time series X_t is said to be "strictly stationary" when the probability density function of the collection of the random variables X_t is the same for any t. The strict stationary property restricts the probabilistic properties of any collection of random variables to be invariant to time shifts, which implies that the probabilistic behaviour of each random variable is the same across time.

D2. A time series is said to be "stationary in the weak sense" or just "weak stationary process" or also "second-order or covariance stationary, if the first (mean) and second (variance) moments exist, are finite and constant over time (therefore not time-dependent), that is $\forall i \in 1 \ldots T$:

$$\begin{gathered}\mathrm{E}(X_t)=\mu=c^{te}\\ \mathrm{V}(X_t)=\sigma^2=c^{te}\\ \mathrm{cov}(X_t,X_{t+h})=\gamma(h)\end{gathered} \quad (15.3.1604)$$

The weak stationarity property restricts the mean and variance of the time series to be finite and invariant in time (the first condition of constancy of the mean eliminates any possible trend) and the autocorrelation $\gamma_t(h)$ to be independent of time t but only depending from the lag h. This can be however violated in different ways. Through deterministic trends (trend stationary), level shifts, variance changes, unit roots (i.e. stochastic trends).

If the underlying density function of each X_t is a Normal distribution, then we speak of "Gaussian process".

Second order stationarity is weaker than strict stationarity. Second order stationarity requires that first and second order moments (mean, variance and covariances) are constant throughout time and, hence, do not depend on the time at which the process is observed.

Second order stationarity series are not necessarily strictly stationary because the mean and autocovariance are not, in general, enough to determine the distribution.

D3. If we have N non-stationary time series and that their weighted sum (the sum of the weights w_i being equal as usual to the unit) is such that $\forall i \in 1 \ldots T$:

$$\begin{gathered}\mathrm{E}(\vec{w}_i\circ\vec{X}_i)=\mu=c^{te}\\ \mathrm{V}(\vec{w}_i\circ\vec{X}_i)=\sigma^2=c^{te}\end{gathered} \quad (15.3.1605)$$

gives a weak stationary process, then we say that the time series are "co-integrated". This is useful in finance so that with some non-stationary financial instruments we can produce a portfolio whose time series is weakly stationary (for mathematical processing reasons).

Let us consider an important case in many corporations on some more or less long periods. Given:

$$\bar{X}_t=\frac{1}{T}\sum_{i=1}^{T}X_t \quad (15.3.1606)$$

the time average. If $\bar{X}_t$ converges in probability to μ when $T\rightarrow+\infty$ we say that the process is "ergodic for the average".

15.3.13.4.1 AR(p) Autoregressive processes

Let us introduce the concept of autoregressive process by introducing the special case of an autoregressive process of order 1, denoted by AR(1) and defined by the relation (up to an additive constant):

$$X_t=\varphi X_{t-1}+\varepsilon_t \tag{15.3.1607}$$

It is therefore a Markov chain of order 1 (we can always verify in practice if the resulting model is a posteriori with or without gaussian residues!).

We say that the autoregressive AR(1) is stationary if we have for any time t:

$$\text{E}(X_t)=\mu=c^{te} \qquad \text{V}(X_t)=\sigma_X^2=c^{te} \tag{15.3.1608}$$

More generally, we have:

$$\begin{aligned}\mu&=\text{E}(X_t)=\varphi\text{E}(X_{t-1})+\text{E}(\varepsilon_t)=\text{E}(X_{t-1})=\varphi^t X_0\\ \sigma_X^2&=\text{V}(X_t)=\varphi^2\text{V}(X_{t-1})+\text{V}(\varepsilon_t)=\varphi^2\text{V}(X_{t-1})+\sigma^2=\sum_{i=0}^{t}\varphi^{2i}\sigma^2\end{aligned} \tag{15.3.1609}$$

and therefore the stationary necessarily implies that (remember that the additive constant was put as zero in the above definition, otherwise it changes the conclusions below):

$$\begin{cases}\mu=\varphi\mu\\ \sigma_X^2=\varphi^2\sigma_X^2+\sigma^2\end{cases}\Rightarrow\begin{cases}\mu=0 \text{ or } \varphi=1\\ (1-\varphi^2)\sigma_X^2=\sigma^2\end{cases}\Rightarrow\begin{cases}\mu=0\\ |\varphi|\ll 1\end{cases} \tag{15.3.1610}$$

Then stationary implies the two useful results:

$$\begin{aligned}\text{E}(X_t)&=\varphi\text{E}(X_{t-1})=1\cdot\text{E}(X_{t-1})=\text{E}(X_{t-1})\\ \sigma_X^2&=\frac{\sigma^2}{1-\varphi^2}\end{aligned} \tag{15.3.1611}$$

It is interesting to notice that the variance is independent of the time and therefore it does not diverge. However, we must not necessarily this a positive fact because it is not very consistent in practice that the volatility is constant (homoscedastic process). This is why we will address more complex models whose volatility is time-dependent.

Notice also that for the variance to be always positive (by definition of the variance) we must have $\varphi<1$.

Let's see what the hypothesis of stationarity involves and for this, we'll first need the following equality:

$$\begin{aligned}X_t&=\varphi X_{t-1}+\varepsilon_t=\varphi(\varphi X_{t-2}+\varepsilon_{t-1})+\varepsilon_t=\varphi^2X_{t-2}+\varphi\varepsilon_{t-1}+\varepsilon_t\\ &=\varphi^hX_{t-h}+\sum_{k=0}^{h-1}\varphi^k\varepsilon_{t-k}\end{aligned} \tag{15.3.1612}$$

Therefore:

$$\begin{aligned}\gamma_t(h)&=\text{cov}(X_t,X_{t-h})=\text{E}(X_tX_{t-h})-\text{E}(X_t)\text{E}(X_{t-h})=\text{E}(X_tX_{t-h})\\ &=\text{E}\left(\left(\varphi^hX_{t-h}+\sum_{k=0}^{h-1}\varphi^k\varepsilon_{t-k}\right)X_{t-h}\right)=\text{E}(\varphi^hX_{t-h}^2)=\varphi^h\text{E}(X_{t-h}^2)\end{aligned} \tag{15.3.1613}$$

And let us recall the Huygens relation (see section Statistics page 593):

$$\mathrm{V}(X)=\mathrm{E}(X^2)-\mathrm{E}(X)^2 \tag{15.3.1614}$$

Therefore it comes:

$$\gamma_t(h)=\varphi^h\mathrm{E}(X_{t-h}^2)=\varphi^2\left[\mathrm{V}(X_{t-h})+\mathrm{E}(X_{t-h})^2\right]=\varphi^h\mathrm{V}(X_{t-h})=\varphi^h\sigma_X^2 \tag{15.3.1615}$$

Notice that if you redo the above developments with $-h$ instead of h you will fall back on exactly the same result. Hence for an autoregressive process the autocovariance is an even function:

$$\gamma_t(h)=\gamma_t(-h) \tag{15.3.1616}$$

So the autocorrelation is indeed dependent only of the lag h (not of the exact time limits of the implicit sum) and given by:

$$\rho_t(h)=\frac{\gamma_t(h)}{\sqrt{\gamma_t(0)\gamma_{t-h}(0)}}=\frac{\varphi^h\sigma_X^2}{\sqrt{\varphi^2\sigma_X^2\varphi^0\sigma_X^2}}=\varphi^h \tag{15.3.1617}$$

Since the stationarity implies that $\varphi \ll 1$, the ACF correlograms of AR(p) processes will have a decreasing profile: the process somehow forgets past values.

We also have for an AR(1) stationary process with constant (offset):

$$X_t=\varphi_0+\varphi_1 X_{t-1}+\varepsilon_t \tag{15.3.1618}$$

the following mean:

$$\begin{aligned}\mathrm{E}(X_t)&=\mathrm{E}(c^{te}+\varphi_1X_{t-1}+\varepsilon_t)=c^{te}+\mathrm{E}(\varphi_1X_{t-1})+\mathrm{E}(\varepsilon_t)=c^{te}+\varphi_1\mathrm{E}(X_{t-1})+0\\&=c^{te}+\varphi_1\mathrm{E}(X_{t-1})+0=c^{te}+\varphi_1\mathrm{E}(X_t)\end{aligned} \tag{15.3.1619}$$

Therefore:

$$\mathrm{E}(X_t)=\frac{\varphi_0}{1-\varphi_1} \tag{15.3.1620}$$

So if the constant is equal to zero we fall back well on a mean equal to zero as proved earlier above. The variance will have the same expression as already the one proved earlier above, that is to say:

$$\sigma_X^2=\frac{\sigma^2}{1-\varphi^2} \tag{15.3.1621}$$

since the addition of a constant does not change the variance of the series. So we got above the mean and (unconditional) variance of an AR(1) process.

If $\varphi=1$ then we have the AR(1) process which is written:

$$X_t=X_{t-1}+\varepsilon_t \tag{15.3.1622}$$

and is therefore by construction not stationary anymore. Nonstationary is common in practice and bring to analysis issues. There is therefore a trick to make some stationary some non-stationary series, do manipulations on them and then to make them again nonstationary by the reverse procedure.

Thus, about the AR(1) with $\varphi = 1$, the idea is to use what we name the "integration property" of a time series. Thus, AR(1) is an integrated series of order 1, designated by I(1) because if it differs from a time unit, we get a stationary series given by:

$$X_t - X_{t-1} = \varepsilon_t \qquad (15.3.1623)$$

Which is written with the so named "backshift operator" denoted B:

$$\begin{aligned}&\text{B}y_t = X_{t-1}\\&\text{B}^2 y_t = X_{t-2}\\&(1-\text{B})X_t = \varepsilon_t\end{aligned} \qquad (15.3.1624)$$

And so a series is "integrated of order 2" if it needs to be "differentiate" twice to make it stationary.

We can also introduce the "lag operator" denoted ∇ or also L:

$$\nabla X_t = (1-\text{B})X_t = \varepsilon_t \qquad (15.3.1625)$$

or a generalization of that latter given by:

$$\nabla_d X_t = X_t - X_{t-d} = X_t - \text{B}^d X_t = (1-\text{B}^d)X_t \qquad (15.3.1626)$$

Remark

If $X_t = t_t + Y_t$, applying the first difference operator to X_t, then:

$$\nabla X_t = X_t - X_{t-1} = (t_t - t_{t-1}) + (Y_t - Y_{t-1}) \qquad (15.3.1627)$$

so if t_t is a trend such that $t_t = t_t(\vec{\beta}) = \beta_0 + \beta_1 t$, then $t_t - t_{t-1} = \beta_1$, i.e. ∇ removes a linear trend.

∇^d removes a seasonality with period d. If $X_t = t_t + s_t + Y_t$, then:

$$\nabla^d X_t = (t_t - t_{t-d}) + (s_t - s_{t-d}) + (Y_t - Y_{t-d}) \qquad (15.3.1628)$$

and $s_t - s_{t-d} = 0$.

An AR(2) process is an auto-regressive process that verifies a relation of the following type (up to an additive constant):

$$X_t = \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \varepsilon_t \qquad (15.3.1629)$$

So in such a model, the influence of the past is manifested by a linear regression on the two previous values. Depending on the values of φ_1 and φ_2 it is not always possible to find a stationary process satisfying this last relation.

In general, an "autoregressive process of order p" AR(p) is a process that depends linearly on the p previous values (up to an additive constant):

$$\boxed{\text{AR}(p) = X_t = \varphi_1 X_{t-1} + \ldots + \varphi_p X_{t-p} + \varepsilon_t = \sum_{k=1}^{p} \varphi_k X_{t-k} + \varepsilon_t} \qquad (15.3.1630)$$

Or using the lag operator (with the notation L):

$$\text{AR}(p)=X_t=\left(\sum_{k=1}^{p}\varphi_k L^k\right)X_t+\varepsilon_t \tag{15.3.1631}$$

✍**Example:**

Suppose we have an i.i.d (independent identically distributed) process where we suppose $\{X_t\}$ is a binary process, i.e. $X_t\in\{-1,+1\}$ such that $P(X_t=1)=p$ and $P(X_t=-1)=1-p$ such that the X_t are mutually independent. We can easily compute the moments as:

$$\text{E}(X_t)=p\cdot(+1)+(1-p)\cdot(-1)=2p-1 \tag{15.3.1632}$$

and:

$$\text{V}(X_t)=[p(1)^2+(1-p)(-1)^2]-(2p-1)^2=1-(1-2p)^2 \tag{15.3.1633}$$

By independence $\text{cov}(X_t,X_s)=0$ if $t\neq s$.

15.3.13.4.2 MA(q) Moving Average stationary process

First a disclaimer! Do not confuse this Moving Average stationary process with deterministic Moving Averages. Indeed, the term is not very appropriate because even if there is indeed mobility (which is the minimum for a process), it is quite wrong to speak of average...

We are in fact in the context of times series forecasts that show autocorrelated errors but nothing else (in fact it will serve for a mix a little further below).

Let us introduce the concept of autoregressive moving average by studying first the special case of autoregressive moving average of order 1, denoted MA(1) and defined by:

$$X_t=\varepsilon_t+\theta\varepsilon_{t-1} \tag{15.3.1634}$$

For such a process, we have by definition of a white noise:

$$\begin{aligned}&\text{E}(X_t)=\text{E}(\varepsilon_t)+\theta\text{E}(\varepsilon_{t-1})=0+\theta\cdot 0=0\\&\text{V}(X_t)=\text{V}(\varepsilon_t)+\theta^2\text{V}(\varepsilon_{t-1})+2\text{cov}(\varepsilon_t,\varepsilon_{t-1})=\sigma^2+\theta^2\sigma^2=\sigma^2(1+\theta^2)\end{aligned} \tag{15.3.1635}$$

So the two moments are both time independent and we are dealing with a weak stationary process. However, we must not necessarily see this as a positive fact because it is not very consistent in practice that the volatility is constant (homoscedastic process). This is why we will address more complex models whose volatility is time-dependent further below.

Regarding the autocorrelation, we have first (do not forget to review the three conditions which define the white noise and especially the third one!):

$$\begin{aligned}\gamma_t(1)&=\text{cov}(X_t,X_{t-1})=\text{cov}(\varepsilon_t+\theta\varepsilon_{t-1},\varepsilon_{t_1}+\theta\varepsilon_{t-2})=\text{E}(X_tX_{t-1})-\text{E}(X_t)\text{E}(X_{t-1})\\&=\text{E}(X_tX_{t-1})=\text{E}(\varepsilon_t\varepsilon_{t-1})+\theta\text{E}(\varepsilon_{t-1}^2)+\theta^2\text{E}(\varepsilon_{t-1}\varepsilon_{t-2})=\theta\text{E}(\varepsilon_{t-1}^2)\\&=\theta\left[\text{V}(\varepsilon_{t-1}+\text{E}(\varepsilon_{t-1})^2\right]=\theta\sigma^2\end{aligned} \tag{15.3.1636}$$

The order 1 autocorrelation coefficient is therefore equal to:

$$\rho_t(1)=\frac{\gamma_t(1)}{\sqrt{\gamma_t(0)\gamma_{t-1}(0)}}=\frac{\text{cov}(X_t,X_{t-1})}{\sqrt{\text{V}(X_t)\text{V}(X_{t-1})}}=\frac{\theta\sigma^2}{\sqrt{\sigma^2(1+\theta^2)\sigma^2(1+\theta^2)}}=\frac{\theta\sigma^2}{\sigma^2(1+\theta^2)}=\frac{\theta}{1+\theta^2} \tag{15.3.1637}$$

For $h>1$, we have immediately by doing the same development as before:

$$\gamma_t(h)=\text{cov}(X_t,X_{t-h})=\text{cov}(\varepsilon_t+\theta\varepsilon_{t-1},\varepsilon_{t-h}+\theta\varepsilon_{t-h-1})=0 \tag{15.3.1638}$$

Therefore the autocorrelation coefficients of order greater than 1 are all zero!

So finally:

$$\gamma_t(h)=\begin{cases}\sigma^2(1+\theta^2) & \text{if } h=0\\ \sigma^2 & \text{if } h=+1\\ 0 & \text{otherwise}\end{cases} \tag{15.3.1639}$$

and similarly:

$$\rho_t(h)=\begin{cases}1 & \text{if } h=0\\ \theta/(1+\theta^2) & \text{if } h=+1\\ 0 & \text{otherwise}\end{cases} \tag{15.3.1640}$$

A MA(2) process is an AR process that verifies a relation of the form:

$$X_t=\varepsilon_t+\theta_1\varepsilon_{t-1}+\theta_2\varepsilon_{t-2} \tag{15.3.1641}$$

And we can repeat the same developments as for MA(1) with enthusiasm ... to arrive also to the conclusion that the autocorrelation (ACF) is zero for $\forall h>2$.

In general, a "moving average process of order q" MA(q) is a process which depends linearly on the previous q values[42]:

$$\boxed{\text{MA}(q)=X_t=\varepsilon_t+\theta_1\varepsilon_{t-1}+\ldots+\theta_q\varepsilon_{t-q}} \tag{15.3.1642}$$

Or using the lag operator:

$$\text{MA}(q)=X_t=\left(1+\sum_{i=1}^{q}\theta_iL^i\right)\varepsilon_t \tag{15.3.1643}$$

---

[42] As we have already mentioned it, the name "moving average" is inaccurate, but the right term can be thought of as a weighted moving average of the past few forecast errors but where the weights don't sum to 1.

and thus for such a process it can be accepted without the general proof that:

$$\begin{gathered}\mathrm{E}(X_t)=0\\ \mathrm{V}(X_t)=\sigma^2\sum_{i=0}^{q}\theta_q^2\end{gathered}\qquad(15.3.1644)$$

and that the correlation is zero for $\forall h>q$. Thus, one can easily recognize that a MA(q) process since its ACF has a "cutoff" from q.

You must remember that the moving average process is an autoregression model of a time series of residual errors from prior predictions!!!

15.3.13.4.3 ARMA(p,q) Autoregressive non-seasonal moving average processes

We can obviously consider combining the two previous models (MA(q) and AR(p)) by introducing:

- A dependency of the process vis-à-vis its past with an AR(p) (i.e. an ARMA($p,0$))
- A delay in the random errors with a MA(q) process (i.e. an ARMA($0,q$))

Such a model, named "non-seasonal autoregressive mobile average" or "non-seasonal ARMA", is characterized by the parameter p of the autoregressive part and the parameter q of the moving average part.

Thus, a non-seasonal ARMA(p,q) process verifies the relation:

$$\boxed{\mathrm{ARMA}(p,q)=X_t=\varepsilon_t+\sum_{i=1}^{p}\varphi_iX_{t-i}+\sum_{i=1}^{q}\theta_i\varepsilon_{t-i}}\qquad(15.3.1645)$$

or explicitly:

$$\mathrm{ARMA}(p,q)=X_t=\varphi_1X_{t-1}+\ldots+\varphi_pX_{t-p}+\varepsilon_t+\theta_t\varepsilon_{t-1}+\ldots+\theta_q\varepsilon_{t-q}\qquad(15.3.1646)$$

Or depending of the point of view of its properties its definition could be written:

$$X_t-\sum_{i=1}^{p}\varphi_iX_{t-i}=\varepsilon_t+\sum_{i=1}^{q}\theta_i\varepsilon_{t-i}\qquad(15.3.1647)$$

And using the lag operator:

$$\left(1-\sum_{i=1}^{p}\varphi_iL^i\right)X_t=\left(1+\sum_{i=1}^{q}\theta_iL^i\right)\varepsilon_t\qquad(15.3.1648)$$

And that latter is even sometimes (sadly...) abbreviated:

$$\Phi(L)X_t=\Theta(L)\varepsilon_t\qquad(15.3.1649)$$

The main advantage of non-seasonal ARMA process is that they allow to provide long-term forecasts (at least more distant in time as the next date...). Indeed, recall that for AR (1) process, we obtained:

$$\mathrm{E}(X_t)=\varphi\mathrm{E}(X_{t-1}) \tag{15.3.1650}$$

Then we have:

$$\hat{X}_t=\mathrm{E}(X_t|X_{t-1})=\varphi X_{t-1} \tag{15.3.1651}$$

Therefore:

$$\hat{X}_{t+1}=\mathrm{E}(X_{t+1}|X_t)=\varphi X_t \tag{15.3.1652}$$

The forecasts are then obtained recursively:

$$\hat{X}_{t+2}=\mathrm{E}(X_{t+2}|X_t)=\varphi\mathrm{E}(X_{t+1}|X_t)=\varphi^2 X_t \tag{15.3.1653}$$

And therefore in a more general way:

$$\boxed{\hat{X}_{t+k}=\varphi^k X_t} \tag{15.3.1654}$$

and so when k becomes very large, then we have:

$$\boxed{\hat{X}_{t+k}\rightarrow 0} \tag{15.3.1655}$$

In other words, for a high order k, the AR(1) model provides as conditional forecasting mean... the mathematics average (history) of the process. When the conditional mean is zero, then we say in the field of time series analysis that the « *mean is orthogonal to any past* ».

So as we can see it, the ARMA process has a mean that has an exponentially decreasing type of memory because of the factor φ that is at the power of k in prior-previous relation. While a MA process type that has a function that cut the temporal memory to a value q (cutoff of q periods). These two processes are therefore considered as short memory process.

15.3.13.4.4 ARIMA(p,d,q) Autoregressive non-seasonal integrated moving average processes

The time series models view previously are in reality just special cases of a more general empirical model that is named "AutoRegressive Integrated Moving Average (ARIMA)" also known as "Box-Jenkins models".

This model assumes that the time series is generated by a random process (Brownian or other) and that future values are related to past errors.

However, this model has an important limitation, and the fact that we find through this generalization the exponential smoothing means that they also have the same limitation: the time series must be stationary. So to recap, this means that moments like mean or variance and correlation must be constant over time.

Remark

In 11970 (holocene calendar) George Box and Gwilym Jenkins popularized ARIMA (Autoregressive Integrated Moving Average) models in their seminal textbook, *Time Series Analysis: Forecasting and Control*. Technically, the forecasting technique described in the text is an ARIMA model, however many forecasters (including the author) use the phrases "ARIMA models" and "Box-Jenkins models" interchangeably.

An ARIMA model is often classified under the notation:

$$\text{ARIMA}(p,d,q) \tag{15.3.1656}$$

where p is the number of autoregressive terms, d the number of non-seasonals differences (order of integration) and q the number of terms of errors (moving average) in the prediction equation.

Specifically for a model as $\text{ARIMA}(1,0,12)$ the trick to remember the parameters is to see the content of the parenthesis as:

$$\boxed{\text{ARIMA}(p,d,q)=\text{ARIMA}(\text{AR},\text{I},\text{MA})} \tag{15.3.1657}$$

AR
AutoRegressive (AR) for looking at past observations
I
Differencing (I or Integrated) for handling changes in data patterns
MA
Moving average (MA) for considering error terms

Explicitly the definition of the $\text{ARIMA}(p,d,q)$ is based on that of an $\text{ARIMA}(p',q)$ but using the expression of the ARMA process with the lags as introduced earlier before:

$$\left(1-\sum_{i=1}^{p'}\varphi_i L^i\right)X_t=\left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t \tag{15.3.1658}$$

Before continuing we need absolutely to come back more in-deep about the concept of "unit root" in time series analysis that we have already briefly introduced earlier before (presenting it first quickly and very naively as a trend in the stochastic signal).

As the reader may know it at this level of the lecture of this textbook, many statistical tools does not apply if the variance of the variable that we are studying is not constant (especially the central limit theorem). In the case of times series analysis, we know that this is assimilated to the concept of "nonstationary" (i.e. the mean or the variance are not constant!).

As we will see, the idea of using ARIMA, is equivalent to apply a difference (or maybe several times d) using the backshift operator B and lag operator ∇ (also denoted L for reminder) already introduced just earlier above:

$$\nabla^d=(1-\text{B})^d \tag{15.3.1659}$$

to a non-stationary time series into a stationary time series.

Let us use a companion example to introduce the concept of unit root. Given the first order equation AR(1) (without the white noise):

$$X_t = \varphi_0 + \varphi_1 X_{t-1} \tag{15.3.1660}$$

We can use the method of iteration to show that when $\varphi_1 = 1$ the series is then given explicitly by:

$$X_t = \varphi_0 t + X_0 \tag{15.3.1661}$$

So there is no stead state. The time series will be trending if $\varphi_0 \neq 0$, and the initial value has permanent effect! In other words unit root implies persistence of shocks!

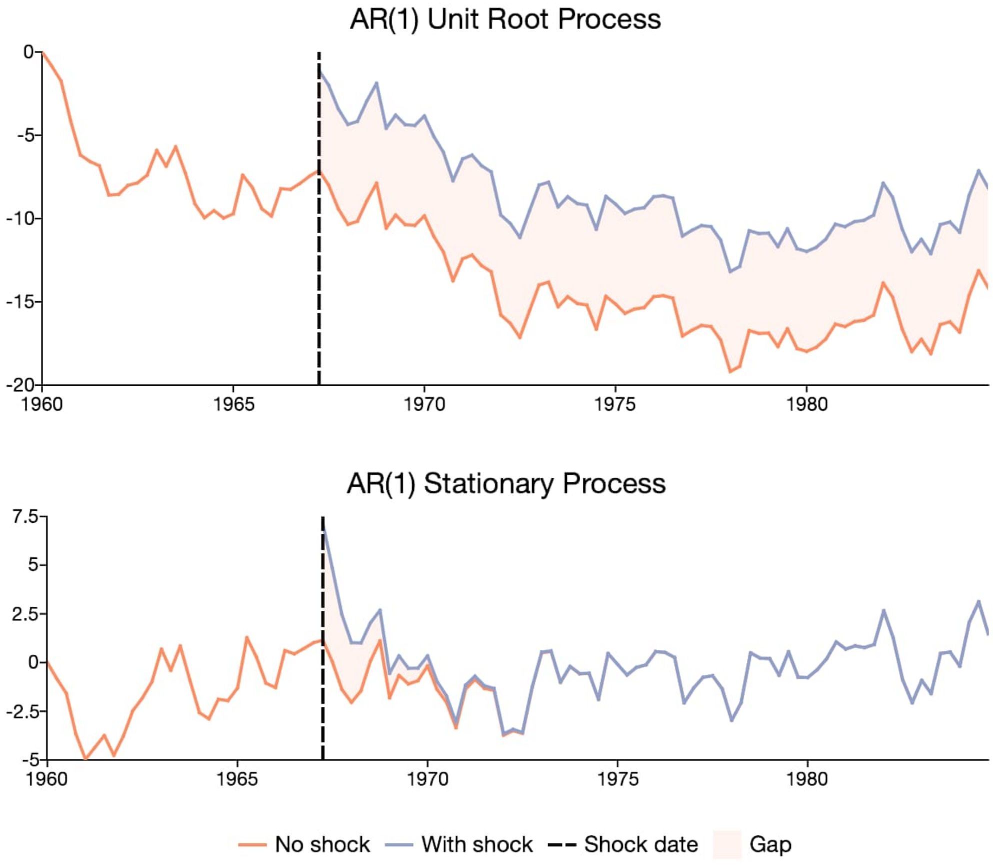


Figure 15.273: **Unit root process vs Stationary process**

Now consider the full AR(1) process:

$$X_t = \varphi_0 + \varphi_1 X_{t-1} + \varepsilon_t \tag{15.3.1662}$$

If $\varphi_1 = 1$ we say that the process has "unit root"[43] because in that case the series has again a trend that can again be expressed using the method of iteration:

$$X_t = \varphi_0 t + X_0 + (\varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_t) \tag{15.3.1663}$$

---

[43] Let us recall that if $\varphi_1 > 0$ then the process is explosive, if $\varphi_1 = 1$ then the process is a random walk, if $\varphi_1 < 1$ then the process is stationary.

If we consider that $\varepsilon_t = \mathcal{N}(0,\sigma_\varepsilon)$ then we have not only the process that is non-stationary in mean since, assuming $X_0 = 0$:

$$\mathrm{E}(X_t) = \mathrm{E}\left(\varphi_0 t + X_0 + t\varepsilon_t\right) = \varphi_0 t \tag{15.3.1664}$$

but it is also non-stationary in variance as (still assuming $X_0 = 0$):

$$\begin{aligned}\mathrm{V}(X_t) &= \mathrm{V}(X_{t-1}+\varepsilon_t)\\ \mathrm{V}(X_t) &= \mathrm{V}(X_{t-1})+\sigma_\varepsilon^2\\ \mathrm{V}(X_t) &= \mathrm{V}(X_{t-2}+\varepsilon_{t-1})+\sigma_\varepsilon^2\\ \mathrm{V}(X_t) &= \mathrm{V}(X_{t-2})+2\sigma_\varepsilon^2\end{aligned} \tag{15.3.1665}$$

repeating this for t steps:

$$\mathrm{V}(X_t) = \mathrm{V}(X_0)+t\sigma_\varepsilon^2 = t\sigma_\varepsilon^2 \tag{15.3.1666}$$

Which means that the standard deviation of the process increases with time $\propto \sqrt{t}$ (according also to Bachelier speculation model result!), and therefore not stationary!

However the above series has the stochastic part that is stationary and shocks have transitory effects. We say that the process is "mean reverting". Also we know that the process has an "attractor", namely the mean $X_0 + \varphi_0 t$.

In short, the unit root process is not stationary!

Remark

The Wiener process (random walk) is a special case of unit root process, which assumes $\varphi_1 = 1, \varphi_0 = 0$. This process is not trending because $\varphi_0 = 0$, however it is still non-stationary because of time-varying standard deviation $\propto \sqrt{t}$. If $\varphi_0 \neq 0$ we can say that it's a random walk with drift as we already know.

Let us now prove that the ACF function of a unit root series decreases to zero linearly and slowly.

For this let us compute first the covariance between X_t and X_{t-k}:

$$\mathrm{cov}(X_t,X_{t-k}) = \mathrm{E}[(X_t-X_0)(X_{t-k}-X_0)] = \mathrm{E}[(\varepsilon_1+\varepsilon_2+\ldots+\varepsilon_t)(\varepsilon_1+\varepsilon_2+\ldots+\varepsilon_{t-k})] = (t-k)\sigma_\varepsilon^2 \tag{15.3.1667}$$

and the autocorrelation is then given by:

$$\rho_t = \frac{\mathrm{cov}(X_t,X_{t-k})}{\sqrt{\mathrm{V}(X_t)\mathrm{V}(X_{t-k})}} = \frac{(t-k)\sigma_\varepsilon^2}{\sqrt{t\sigma_\varepsilon^2\cdot(t-k)\sigma_\varepsilon^2}} = \frac{t-k}{\sqrt{t(t-k)}} = \sqrt{\frac{t-k}{t}} \tag{15.3.1668}$$

which dies out very slowly with k.

Now let us recall that using the lag operator, the white noise can be rewritten:

$$(1-\mathrm{B})X_t = \varepsilon_t \tag{15.3.1669}$$

The equation:

$$1-\mathrm{B}=0 \tag{15.3.1670}$$

has the root $\mathrm{B}=1$ what is named the "unit root".

A time series possibly has multiple unit roots. For example, consider the series with two unit roots:

$$\begin{aligned}&(1-\mathrm{B})(1-\mathrm{B})X_t=\varepsilon_t\\ \Rightarrow\quad&(1-2\mathrm{B}+\mathrm{B}^2)X_t=\varepsilon_t\\ \Rightarrow\quad&X_t=2\mathrm{B}X_t-\mathrm{B}^2X_t+\varepsilon_t\\ \Rightarrow\quad&X_t=2X_{t-1}-X_{t-2}+\varepsilon_t\end{aligned} \tag{15.3.1671}$$

So the AR(2) process:

$$X_t=2X_{t-1}-X_{t-2}+\varepsilon_t \tag{15.3.1672}$$

is obviously non-stationary and has two unit roots.

So as we know, we cannot apply the ARMA model to unit root processes, unless we transform the series into stationary series. And we already know that we can do this by applying the lag operator $\nabla=1-\mathrm{B}$.

A process that has been made stationary by applying only time the backshift operator is said to be "integrated of first order" $I(1)$. More generally, a time series is integrated of order d, $I(d)$, if it contains d unit roots (i.e. if we need to take difference d times before it becomes stationary).

So now remember the following ARMA notation:

$$\left(1-\sum_{i=1}^{p'}\varphi_i\mathrm{B}^i\right)X_t=\left(1+\sum_{i=1}^{q}\theta_i\mathrm{B}^i\right)\varepsilon_t \tag{15.3.1673}$$

Assume now that the polynomial $\left(1-\sum_{i=1}^{p'}\alpha_i\mathrm{B}^i\right)$ has a unit root $(1-\mathrm{B})$ of multiplicity d. Then it can be rewritten as:

$$\left(1-\sum_{i=1}^{p'}\varphi_i\mathrm{B}^i\right)=\left(1-\sum_{i=1}^{p'-d}\phi_i\mathrm{B}^i\right)(1-\mathrm{B})^d \tag{15.3.1674}$$

An ARIMA(p,d,q) process expresses this polynomial factorization property with $p=p'-d$, and is given by:

$$\boxed{\mathrm{ARIMA}(p,d,q)=\mathrm{ARIMA}(\mathrm{AR},\mathrm{I},\mathrm{MA})=\left(1-\sum_{i=1}^{p'-d}\phi_i\mathrm{B}^i\right)(1-L)^dX_t=\left(1+\sum_{i=1}^{q}\theta_i\mathrm{B}^i\right)\varepsilon_t} \tag{15.3.1675}$$

and thus can be thought as a special case of an ARMA$(p+d,q)$ process having the autoregressive polynomial with d unit roots. In other words, ARIMA(p,d,q) model means we take difference d times and then apply the ARMA(p,q) model to the differences series.

That previous relation is sometimes (sadly...) denoted:

$$\Phi(\text{B})(1-\text{B})^d X_t = \Theta(\text{B})\varepsilon_t \tag{15.3.1676}$$

The explicit identification of the factorization of the autoregression polynomial into factors as above, can be extended to other cases, firstly to apply to the moving average polynomial and secondly to include other special factors. For example, having a factor $(1-\text{B}^s)$ in a model is one way of including a non-stationary seasonality of period s into the model.

✍**Examples:**

E1. Average: ARIMA$(0,0,0)$ with constant (white noise):

$$X_t = \mu + \varepsilon_t \tag{15.3.1677}$$

E2. Drift: ARIMA$(0,1,0)$ with constant (random walk):

$$X_t = \mu + X_{t-1} + \varepsilon_t \tag{15.3.1678}$$

When μ is taken as 0 we have then:

$$X_t = X_{t-1} + \varepsilon_t \tag{15.3.1679}$$

and this is named in some statistical packages as a "naive forecast"

E3. ARMA: ARIMA$(1,0,1)$

E4. Single Exponential Smoothing: ARIMA$(0,1,1)$ without constant:

$$X_{t+1} = X_t + \alpha\varepsilon \tag{15.3.1680}$$

As we have seen in a remark during our study of the single exponential smoothing model.

E5. Holt's additive Double Exponential Smoothing: ARIMA$(0,2,2)$

$$X_t = X_{t-1} + X_{t-2} + (\alpha+\beta-2)\varepsilon_{t-1} + (1-\alpha)\varepsilon_{t-2} + \varepsilon_t \tag{15.3.1681}$$

E6. Holt & Winters Exponential Smoothing (multiplicative): no ARIMA equivalent to our knowledge

E7. Holt & Winters Exponential Smoothing (additive) : SARIMA$(0,1,m+1)(0,1,0)$

Or another type of equivalence in the shape of table:

ETS model	ARIMA model	Parameters
ETS $(\text{A},\text{N},\text{N})$	$\text{ARIMA}(0,1,1)$	$\theta_1=\alpha-1$
$\text{ETS}(\text{A},\text{A},\text{N})$	$\text{ARIMA}(0,2,2)$	$\theta_1=\alpha+\beta-2$
		$\theta_2=1-\alpha$
$\text{ETS}\,(\text{A},\text{A}_d,\ \text{N})$	$\text{ARIMA}(1,1,2)$	$\phi_1=\phi$
		$\theta_1=\alpha+\phi\beta-1-\phi$
		$\theta_2=(1-\alpha)\phi$
$\text{ETS}(\text{A},\text{N},\text{A})$	$\text{ARIMA}(0,1,m)(0,1,0)_m$	
$\text{ETS}(\text{A},\text{A},\text{A})$	$\text{ARIMA}(0,1,m+1)(0,1,0)_m$	
$\text{ETS}\,(\text{A},\text{A}_d,\ \text{A})$	$\text{ARIMA}(0,1,m+1)(0,1,0)_m$	

Table 15.65: **Some equivalence between ARIMA and ETS models**

Or using a Venn diagram with different type of similarities depicted:

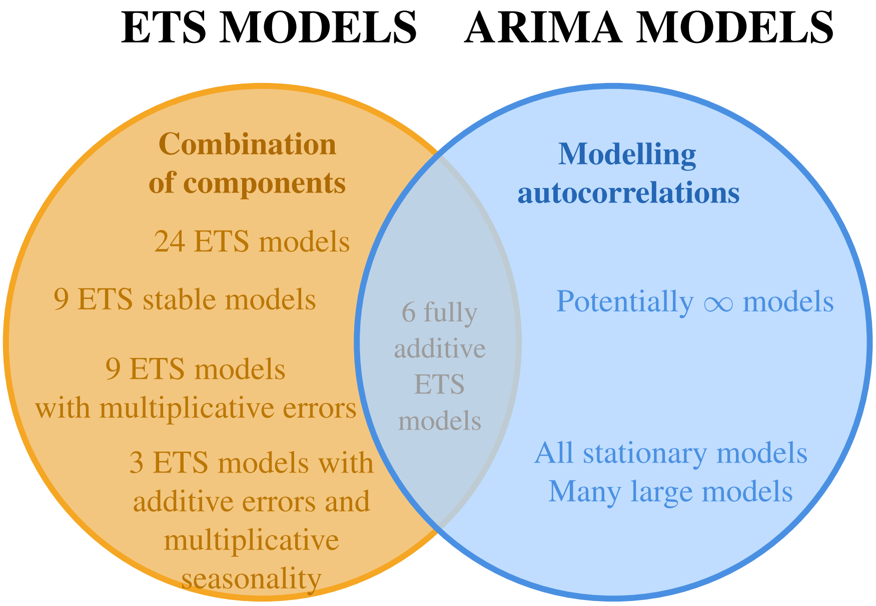

ETS MODELS
ARIMA MODELS
Combination of components
24 ETS models
9 ETS stable models
9 ETS models with multiplicative errors
3 ETS models with additive errors and multiplicative seasonality
6 fully additive ETS models
Modelling autocorrelations
Potentially ∞ models
All stationary models
Many large models

Figure 15.274: **Summary of ETS models vs ARIMA models**

As the last example is a SARIMA, it may be the time now to introduce that latter! First the reader should know that there is not only one SARIMA model but a multiple! Normally we start by introducing:

$$\text{SARIMA}(s,p,d,q)=\Phi(\text{B})(1-\text{B}^s)(1-\text{B})^d X_t=\Theta(\text{B})(1-\text{B}^s)(1-\text{B})^d\varepsilon_t \qquad (15.3.1682)$$

and all others $\text{SARIMA}(s,p,d,q)$ are subsets of that latter like:

- $\text{SARIMA}(s,p,d,q)=\Phi(\text{B})(1-\text{B}^s)X_t=\Theta(\text{B})(1-\text{B})^d\varepsilon_t$
- $\text{SARIMA}(s,p,d,q)=\Phi(\text{B})(1-\text{B})^d X_t=\Theta(\text{B})(1-\text{B}^s)\varepsilon_t$
- $\text{SARIMA}(s,p,d,q)=\Phi(\text{B})X_t=\Theta(\text{B})(1-\text{B}^s)(1-\text{B})^d\varepsilon_t$
- $\text{SARIMA}(s,p,d,q)=\Phi(\text{B})(1-\text{B}^s)(1-\text{B})^d X_t=\Theta(\text{B})\varepsilon_t$

- $\text{SARIMA}(s,p,d,q)=\Phi(\text{B})(1-\text{B}^s)(1-\text{B})^d X_t=\Theta(\text{B})(1-\text{B}^s)\varepsilon_t$
- $\text{SARIMA}(s,p,d,q)=\Phi(\text{B})(1-\text{B}^s)(1-\text{B})^d X_t=\Theta(\text{B})(1-\text{B})^d\varepsilon_t$
- ... and so on (a total of 9 possible combinations)

However there exist a more general version given by the multiplicative seasonal ARIMA model and commonly defined as:

$$\boxed{\text{SARIMA}\underbrace{(p,d,q)}_{\substack{\text{Non-seasonal}\\ \text{part of the model}}}\times\underbrace{(P,D,Q)_s}_{\substack{\text{Seasonal part}\\ \text{of the model}}}=\Phi_P(\text{B}^s)\phi_p(\text{B})(1-\text{B})^d(1-\text{B}^s)^D(X_t-\mu)=\theta_q(\text{B})\Theta_Q(\text{B}^s)\varepsilon_t} \tag{15.3.1683}$$

Or also commonly written:

$$\text{SARIMA}(p,d,q)\times(P,D,Q)_s=\Phi_P(\text{B}^s)\phi_p(\text{B})\nabla_s^D\nabla^d Z_t=\theta_q(\text{B})\Theta_Q(\text{B}^s)\varepsilon_t \tag{15.3.1684}$$

where $s=$ number of periods per season, the uppercase notation for the seasonal parameters of the model, and lowercase notation for the non-seasonal parameters of the model. The seasonal part of the model consists of terms that are very similar to the non-seasonal components of the model, but they involve backshifts of the seasonal period.

Expliciting all other terms, we have:

- $\Phi_P(\text{B}^s)=(1-\Phi_1\text{B}^s-\ldots-\Phi_P\text{B}^{sP})=1-\sum_{i=1}^P\Phi_i\text{B}^{si}$ is the seasonal AR operator of order P;
- $\phi_p=(1-\phi_1\text{B}-\ldots-\phi_p\text{B}^p)=1-\sum_{i=1}^p\phi_i\text{B}^i$ is the regular AR operator of order p;
- $\nabla_s^D=(1-\text{B}^s)^D$ represents the seasonal differences and $\nabla^d=(1-\text{B})^d$ the regular differences;
- $\Theta_Q(\text{B}^s)=(1-\Theta_1\text{B}^s-\ldots-\Theta_Q\text{B}^{sQ})=1-\sum_{i=1}^Q\Theta_i\text{B}^{si}$ is the seasonal moving average operator of order Q;
- $\theta_q(\text{B})=(1-\theta_1\text{B}-\ldots-\theta_q\text{B}^q)=1-\sum_{i=1}^q\theta_i\text{B}^i$ is the regular moving average operator of order q;
- ε_t is the usual white noise

The definition may not be very clear for a significant percentage of readers (according to our experience of teaching the subject). So let us see an example:

✍**Example:**

Consider we want to explicit:

$$\text{SARIMA}(0,2,1)\times(0,0,1)_{12}=\Phi_0(\text{B}^{12})\phi_0(\text{B})(1-\text{B})^2(1-\text{B}^{12})^0 Z_t=\theta_1(\text{B})\Theta_1(\text{B}^{12})\varepsilon_t \tag{15.3.1685}$$

where:

$$\phi_0(\mathrm{B})=1 \qquad \Phi_0(\mathrm{B}^{12})=1 \qquad \theta_1(\mathrm{B})=1+\theta_1\mathrm{B} \qquad \Theta_1(\mathrm{B}^{12})=1+\Theta_1\mathrm{B}^{12} \tag{15.3.1686}$$

Therefore:

$$\text{SARIMA}(0,2,1)\times(0,0,1)_{12}=1\cdot 1\cdot(1-\mathrm{B})^2(1-\mathrm{B}^{12})Z_t=(1+\theta_1\mathrm{B})(1+\Theta_1\mathrm{B}^{12})\varepsilon_t \tag{15.3.1687}$$

And the result is:

$$\text{SARIMA}(0,2,1)\times(0,0,1)_{12}=Z_t-2Z_{t-1}+Z_{t-2}=\varepsilon_t+\Theta\varepsilon_{t-12}+\theta_1\varepsilon_{t-1}+\theta_1\Theta_1\varepsilon_{t-13} \tag{15.3.1688}$$

So we can summarize all the models seen so far by the following (quite ugly) table:

Process	**Model**	**Stationary Condition**
MA(q)	$X_t=\left(1+\sum_{i=1}^{q}\theta_i\mathrm{B}^i\right)\varepsilon_t$	None: MA(q) process is stationary by construction
AR(p)	$X_t=\left(\sum_{k=1}^{p}\varphi_k\mathrm{B}^k\right)X_t+\varepsilon_t$	$\sum_{i=1}^{p}\varphi_i<1$
ARMA(p,q)	$\left(1-\sum_{i=1}^{p}\varphi_i\mathrm{B}^i\right)X_t=\left(1+\sum_{i=1}^{q}\theta_i\mathrm{B}^i\right)\varepsilon_t$	AR(p) component must be stationary
ARIMA(p,d,q)	$\left(1-\sum_{i=1}^{p'-d}\phi_i\mathrm{B}^i\right)(1-\mathrm{B})^dX_t=\left(1+\sum_{i=1}^{q}\theta_i\mathrm{B}^i\right)\varepsilon_t$	The AR(p) component is stationary after d integration
SARIMA$(p,d,q)(P,D,Q)_m$	$\Phi_P(\mathrm{B}^s)\phi_p(\mathrm{B})\nabla_s^D\nabla^dZ_t=\theta_q(\mathrm{B})\Theta_Q(\mathrm{B}^s)\varepsilon_t$	The AR(p) component is stationary after d integration

Table 15.66: Summary of ARIMA models

The ARIMA models (and thereafter SARIMO too!) are confined in the range of integer order d. Therefore, in order to capture the property of the fractional systems, the "Autoregressive Fractional Integrated Moving Average" ARFIMA(p,d,q) model is thereby proposed a natural way by using binomial expansion for any real number d with Gamma function. Indeed, if you remember the binomial theorem (see section Algebra page 1364):

$$(a+b)^n=\sum_{p=0}^{n}C_p^n a^{n-p}b^p \tag{15.3.1689}$$

Hence:

$$(1-\mathrm{B})^d=\sum_{p=0}^{d}C_p^d(-\mathrm{B})^p=\sum_{p=0}^{d}\frac{d!}{p!(d-p)!}(-\mathrm{B})^p \tag{15.3.1690}$$

And using the Gamma function (see section Differential and Integral Calculus page 1574) that extends then d to real values (and not only to integers anymore):

$$(1-\mathrm{B})^d=\sum_{p=0}^{d}\frac{\Gamma(d+1)}{\Gamma(d+1)\Gamma(d+1-p)!}(-\mathrm{B})^p \tag{15.3.1691}$$

So finally ARFIMA(p,d,q) model shares the same form of representation as the ARIMA(p,d,q) process, at the difference that the parameter d, is allowed to take non-integer values.

Remark

It is a commonly held myth that ARIMA models are more general than exponential smoothing. While linear exponential smoothing models are all special cases of ARIMA models, the non-linear exponential smoothing models have no equivalent ARIMA counterparts. On the other hand, there are also many ARIMA models that have no exponential smoothing counterparts. In particular, all ETS models are non-stationary, while some ARIMA models are stationary.

15.3.13.4.5 ARCH-GARCH processes

The modelling of volatility is an essential issue for many fields of study of contemporary finance (and also other fields!). The simple fact that portfolio management relies mostly on mean-variance algorithms underscores the importance of constructing correctly specified variance indicators. Options management is also based on a rigorous study of the evolution of the variance!

The financial time series present a number of characteristic features that need to be emphasized before moving on to their modelling. These main facts are as we already know:

- The returns are not autocorrelated (thus the autocorrelation between the return on the date t and the date $t-1$ is a financial aberration) and this is even more pronounced as the markets are liquefied. However some markets continue to exhibit some autocorrelation
- In general, the volatility of an asset increases when the preceding returns are negative. We are also talking in this case of "leverage".
- The volatility of an asset generally displays periods of significant volatility and other periods of low volatility (corresponding to the so-named "bear" and "bull" market).
- In the financial time series, volatility is a function of time whereas in the ARIMA processes the conditional volatility is assumed as constant.

This list is of course not exhaustive. However, it has the merit of highlighting the fact that when one wishes to get out of the standards of finance (the random walk of the theories of efficiency and option evaluation), it is necessary to consider all of these properties with great care.

In essence, volatility is an unobservable phenomenon. It is then necessary to construct a certain number of empirical statistical indicators (like the famous various "indices") making it possible to judge the volatility of the returns. The fact that this type of index know an obvious success only underscores the need for financial markets to have instruments for measuring volatility.

For this purpose, the jungle of empirical families of ARCH (AutoRegressive Conditional Heteroskedasticity) processes aims to account for the fact that the conditional variance is not constant and propose a way of estimating it. It will obviously be necessary as always to use this model with suspicion because it is possible that the volatility is not constant over time but that the empirical models of the family ARCH or GARCH (Generalized Autoregressive Conditional Heteroskedasticity) do not capture this effect, or that even conclude in some cases to lack of temporal dependence of returns!

15.3.13.4.5.1 ARCH (Autoregressive Conditional Heteroskedasticity)

To introduce the ARCH(q) let us denote by ε_t the error term that will be assumed to be modelled by a stochastic term of white noise type z_t with a time-dependent volatility such that:

$$\varepsilon_t = \sigma_t z_t = \mathcal{N}(0, \sigma_t^2) \tag{15.3.1692}$$

In practice ε_t is the residual of the returns in relation to a given mean or median (the choice of the latter varies according to the practitioners). In most theoretical framework this average or median is taken as being zero (all the values being centered reduced), in the practical case one takes either the empirical average of the returns in parametric models or the median in non-parametric models, or the difference of two successive returns. In short, whatever the statistical software used, the result will be more or less very different depending on the initial choice.

The empirical idea is then to model the variance by:

$$\sigma_t^2 = \alpha_0 + \alpha_1\varepsilon_{t-1}^2 + \ldots + \alpha_q\varepsilon_{t-q}^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-1}^2 \tag{15.3.1693}$$

where:

$$\alpha_0 > 0 \qquad \alpha_i \geq 0 \tag{15.3.1694}$$

Remark

Notice that the model assumes that positive and negative shocks have the same effects on volatility because it depends on the square of the previous shocks. In practice, it is well known that the price of a financial asset responds differently to positive and negative shocks...

This is often summarized in the following form:

$$\begin{gathered}\varepsilon_t = \sigma_t z_t = \mathcal{N}(0, \sigma_t^2)\\ \sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-1}^2\end{gathered} \tag{15.3.1695}$$

or with another common notation:

$$\begin{gathered}X_t = \sqrt{h_t} z_t = \mathcal{N}(0, h_t)\\ h_t^2 = \omega_0 + \sum_{i=1}^{q} \omega_i X_{t-1}^2\end{gathered} \tag{15.3.1696}$$

But we prefer in this textbook the following explicit notation:

$$\boxed{\begin{aligned}&X_t = \sigma_t z_t = \mathcal{N}(0, \sigma_t^2)\\ &\text{ARCH}(q) = \sigma_t^2 = \text{V}(X_t|X_{t-1}) = \omega_0 + \sum_{i=1}^{q} \omega_i X_{t-i}^2 \qquad \omega_0 > 0, \omega_i \geq 0\end{aligned}} \tag{15.3.1697}$$

So the conditional variance is indeed dependent of time (only!). Specialists like to say (according to the specialized vocabulary we have seen in the Statistics section) that an ARCH process is (conditionally) heteroscedastic whereas an ARIMA process is (conditionally) homoscedastic.

In the case of an ARCH(1), that is to say a process defined by:

$$\begin{aligned} X_t &= \sigma_t \varepsilon_t \\ \sigma_t^2 &= \omega_0 + \omega_1 X_{t-1}^2 \end{aligned} \tag{15.3.1698}$$

We then have for the (unconditional) expected mean considering the independence of the variables σ_t and ε_t:

$$\mathrm{E}(X_t) = \mathrm{E}(\sigma_t \varepsilon_t) = \mathrm{E}(\sigma_t) \underbrace{\mathrm{E}(\varepsilon_t)}_{=0} = 0 \tag{15.3.1699}$$

and this for any t (you have to remember that as later we will use the fact that $\mathrm{E}(X_{t-1}) = 0$).

For the (unconditional) variance, first demonstrate the following result:

Knowing that for a random variable X we have the Huygens relation (see section Statistics page 593) it comes:

$$\mathrm{V}(X_t) = \mathrm{V}(\sigma_t \varepsilon_t) = \mathrm{E}(\sigma_t^2 \varepsilon_t^2) - \mathrm{E}(\sigma_t \varepsilon_t)^2 \tag{15.3.1700}$$

Now we use the independence of the variables σ_t and ε_t:

$$\mathrm{V}(X_t) = \mathrm{E}(\sigma_t^2)\mathrm{E}(\varepsilon_t^2) - \mathrm{E}(\sigma_t)^2 \underbrace{\mathrm{E}(\varepsilon_t)^2}_{=0} = \mathrm{E}(\sigma_t^2)\mathrm{E}(\varepsilon_t^2) \tag{15.3.1701}$$

Knowing that obviously:

$$\mathrm{V}(\varepsilon_t) = \mathrm{E}(\varepsilon_t^2) - \mathrm{E}(\varepsilon_t)^2 = \mathrm{E}(\varepsilon_t^2) \tag{15.3.1702}$$

we then get for an ARCH(1) process:

$$\mathrm{V}(\varepsilon_t) = \mathrm{E}(\sigma_t^2)\mathrm{E}(\varepsilon_t^2) = \mathrm{E}(\sigma_t^2)\mathrm{V}(\varepsilon_t) = \mathrm{E}(\sigma_t^2) \tag{15.3.1703}$$

We then have:

$$\mathrm{V}(X_t) = \mathrm{E}(\sigma_t^2) = \mathrm{E}(\omega_0 + \omega_1 X_{t-1}^2) = \omega_0 + \omega_1 \mathrm{E}(X_{t-1}^2) = \omega_0 + \omega_1 \mathrm{E}(X_{t-1}^2) \tag{15.3.1704}$$

And we have using again the Huygens relation:

$$\mathrm{V}(X_{t-1}) = \mathrm{E}(X_{t-1}^2) - \underbrace{\mathrm{E}(X_{t-1})^2}_{=0} = \mathrm{E}(X_{t-1}^2) \tag{15.3.1705}$$

the following result:

$$\mathrm{V}(X_t) = \omega_0 + \omega_1 \mathrm{V}(X_{t-1}) \tag{15.3.1706}$$

Remark

The latter relation implies that we can estimate ω_0 and ω_1 by regressing $\mathrm{V}(X_t)$ onto an intercept term and $\mathrm{V}(X_{t-1})$. It also implies that $\mathrm{V}(X_t)$ follows an AR(1) process!

And therefore:

$$\begin{aligned} \mathrm{V}(X_t) &= \omega_0 + \omega_1 \mathrm{V}(X_{t-1}) \\ \mathrm{V}(X_{t-1}) &= \omega_0 + \omega_1 \mathrm{V}(X_{t-2}) \\ \mathrm{V}(X_{t-2}) &= \omega_0 + \omega_1 \mathrm{V}(X_{t-3}) \end{aligned} \tag{15.3.1707}$$

Hence:

$$\begin{aligned}\mathrm{V}(X_t)&=\omega_0+\omega_1(\omega_0+\omega_1(\omega_0+\omega_1\mathrm{V}(X_{t-3})))\\&=\omega_0(1+\omega_1+\omega_1^2)+\omega_1^3\mathrm{V}(X_{t-3})\end{aligned} \tag{15.3.1708}$$

So finally we get the general relation:

$$\mathrm{V}(X_t)=\omega_0\sum_{i=0}^{n}\omega_1^i+\omega_1^{n+1}\mathrm{V}(X_{t-(n+1)}) \tag{15.3.1709}$$

In the case where $|\omega_1|<1$, we have when $n\rightarrow+\infty$:

$$\mathrm{V}(X_t)_{n\rightarrow+\infty}\underset{|\omega_1|<1}{=}\omega_0\sum_{i=0}^{n}\omega_1^i \tag{15.3.1710}$$

And using the result demonstrated in the section of Sequences and Series during our study of geometric series (page 1679):

$$\sum_{i=0}^{n}x^i\underset{|x|<1}{=}\frac{1}{1-x} \tag{15.3.1711}$$

We then have:

$$\mathrm{V}(X_t)_{n\rightarrow+\infty}\underset{|\omega_1|<1}{=}\frac{\omega_0}{1-\omega_1} \tag{15.3.1712}$$

Thus, in this special case, we will have a stationary process because the variance will be finite and independent of time!

What has just been said about ARCH(1) can be generalized in the case of ARCH(p) processes where the (unconditional) expected mean remains zero and the variance (unconditional) in the case where:

$$\left|\sum_{i=1}^{p}\omega_i\right|<1 \tag{15.3.1713}$$

is then equal to:

$$\mathrm{V}(X_t)=\frac{\omega_0}{1-\displaystyle\sum_{i=1}^{p}\omega_i} \tag{15.3.1714}$$

Event without knowing the exact expression of the fourth moment κ_1 (i.e. kurtosis as defined in the section Statistics page 636) of an ARCH(p) process, it is possible to prove quite easily that latter is $\kappa_1=\mu_4/\sigma^4\geq 3$ (as we have proved that for a standard Normal distribution at page 705 that we have $\kappa_1=3$)!

Proof 15.7.2. Let us first recall that for an ARCH(p) process, we have by definition:

$$\varepsilon_t|X_{t-1},X_{t-2},\ldots=\mathcal{N}(0,\sigma_t^2) \tag{15.3.1715}$$

That means that the conditional distribution of X_t is Normal (and without loss of generality we can consider $\sigma_t = 1$), and we know that for the standardized Normal distribution, we have:

$$\kappa_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}x_i^4}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2}\right)^4} = \frac{\mathrm{E}(X^4)}{\mathrm{E}(X^2)^2} = \frac{\mu_4}{\sigma^4} = 3 \tag{15.3.1716}$$

Therefore, denoting $\mathcal{I}_t$ the set of all values of X_t taken before time t such that:

$$\mathcal{I}_{t-1} = \{X_{t-i}, i > 0\} \tag{15.3.1717}$$

We then have:

$$\kappa_1 = \frac{\mathrm{E}(X_t^4|\mathcal{I}_{t-1})}{\mathrm{E}(X_t^2|\mathcal{I}_{t-1})^2} = 3 \tag{15.3.1718}$$

Hence:

$$\mathrm{E}(X_t^4|\mathcal{I}_{t-1}) = 3\left(\mathrm{E}(X_t^2|\mathcal{I}_{t-1})\right)^2 \tag{15.3.1719}$$

Let us take the expected mean on the both sides:

$$\mathrm{E}\left(\mathrm{E}(X_t^4|\mathcal{I}_{t-1})\right) = 3\mathrm{E}\left(\left(\mathrm{E}(X_t^2|\mathcal{I}_{t-1})\right)^2\right) \tag{15.3.1720}$$

and now using the iterated conditional mean for the left term (see section Statistics page 501), we get:

$$\mathrm{E}(X_t^4) = 3\mathrm{E}\left(\left(\mathrm{E}(X_t^2|\mathcal{I}_{t-1})\right)^2\right) \tag{15.3.1721}$$

Let us now recall that we have proved the Jensen inequality (see section Statistics page 576) was given for a convex function by:

$$f(\mathrm{E}(X)) \le \mathrm{E}(f(X)) \tag{15.3.1722}$$

Therefore (considering the $(\ldots)^2$ as the application $f(\ldots)$ and $X = \mathrm{E}(X_t^2|\mathcal{I}_{t-1})$), we have for the right term of the equality:

$$\left(\mathrm{E}(\mathrm{E}(X_t^2|\mathcal{I}_{t-1}))\right)^2 \le \mathrm{E}\left(\left(\mathrm{E}(X_t^2|\mathcal{I}_{t-1})\right)^2\right) \tag{15.3.1723}$$

Hence:

$$\mathrm{E}(X_t^4) \ge 3\left(\mathrm{E}(\mathrm{E}(X_t^2|\mathcal{I}_{t-1}))\right)^2 \tag{15.3.1724}$$

Using again iterated conditional mean, this can be written:

$$\mathrm{E}(X_t^4) \ge 3\mathrm{E}(X_t^2)^2 \tag{15.3.1725}$$

So finally the unconditional kurtosis is:

$$\boxed{\kappa_1^{\text{ARCH}(p)} = \frac{\mathrm{E}(X_t^4)}{\mathrm{E}(X_t^2)^2} \ge 3} \tag{15.3.1726}$$

So the ARCH(p) process have a kurtosis excess[44]!

■ Q.E.D.

---

[44] Note that a similar reasoning and result applies to GARCH processes

15.3.13.4.5.2 GARCH (Generalized Autoregressive Conditional Heteroskedasticity)

The ARCH model is appropriate when the error variance in a time series follows an autoregressive (AR) model; if an autoregressive moving average (ARMA) model is assumed for the error variance, the model is a generalized autoregressive conditional heteroskedasticity (GARCH) model.

In the case of a GARCH(p, q) process, we have:

$$\begin{aligned}\varepsilon_t&=\sigma_t z_t\\ \text{GARCH}(q,p)=\sigma_t^2&=\alpha_0+\sum_{i=1}^{q}\alpha_i\varepsilon_{t-i}^2+\sum_{i=1}^{p}\beta_i\sigma_{t-i}^2 \qquad \alpha_0>0,\alpha_i\geq 0,\beta_i\geq 0\end{aligned} \tag{15.3.1727}$$

The GARCH$(1, 1)$ is then given by:

$$\begin{aligned}\varepsilon_t&=\sigma_t z_t\\ \sigma_t^2&=\alpha_0+\alpha_1\varepsilon_{t-1}^2+\beta_1\sigma_{t-1}^2\end{aligned} \tag{15.3.1728}$$

which is also often written as following:

$$\begin{aligned}\varepsilon_t&=\sigma_t z_t\\ \sigma_t^2&=\alpha_0+\alpha\varepsilon_{t-1}^2+\beta\sigma_{t-1}^2\end{aligned} \tag{15.3.1729}$$

Remark

Notice that as for the ARCH, the GARCH model assumes that positive and negative shocks have the same effects on volatility because it depends on the square of the previous shocks. In practice, it is well known that the price of a financial asset responds differently to positive and negative shocks...

In practice we will try to determine the different parameters α_0, α, β knowing that the RiskMetrics methodology of JP Morgan recommends (at the day we write these lines...) to use the GARCH$(1, 1)$ model with the conditions:

$$\alpha+\beta=1 \tag{15.3.1730}$$

$$\alpha_0=0 \tag{15.3.1731}$$

and as in finance we normally work with returns, so we end up writing (the reader will notice that in the second line we find an expression similar to that of the simple exponential smoothing model - there is just a random factor more - so the GARCH (1,1) is just a simple generalization of an EWMA process):

$$\begin{aligned}r_t&=\sigma_t z_t\\ \sigma_t^2&=(1-\beta)r_{t-1}^2+\beta\sigma_{t-1}^2\end{aligned} \tag{15.3.1732}$$

What we find again often under the following notation:

$$\begin{aligned}r_t&=\sqrt{h_t}z_t\\ h_t&=(1-\beta)r_{t-1}^2+\beta\sigma_{t-1}\end{aligned} \tag{15.3.1733}$$

and if the mean return is not zero, then we write:

$$\begin{aligned}r_t&=\mu_t+\sqrt{h_t}z_t\\ h_t&=(1-\beta)r_{t-1}^2+\beta h_{t-1}\end{aligned} \tag{15.3.1734}$$

15.3.13.5 Forecasting Stationary Processes

Let us come back on projection techniques (forecasting) of ARIMA models. Let's start with a simple special case (that is in our opinion more educational) considering the process denoted $\text{AR}(1)$ and defined for recall by the relation (sum up over a given additive constant):

$$X_t = \varphi X_{t-1} + \varepsilon_t \tag{15.3.1735}$$

When we make a punctual estimate of a projection of this model, on h subsequent units of time, it is customary to write the punctual estimate value by different usual notations:

$$X_t(h) = \varphi X_{t-1} + \varepsilon_t \tag{15.3.1736}$$

where φ is obviously supposed to have been determined with known historical values. However, the projected point estimate is precisely an... estimate... !!!! We would have to know its mean. The simplest naive assumption is to use the conditional expectation of the forecast value, who led us for recall (see just earlier above page 7989) to the result:

$$\hat{X}_{t+h} = \varphi^h X_t \tag{15.3.1737}$$

which tends to 0 as h approaches infinity as we have already proved it. But remember that this $\text{AR}(1)$ is defined up to an additive constant (which is often the arithmetic mean or the median of known values of the time series), so in the general case it tends towards a constant!!!

On the way, notice that we explicitly have the forecast error that is given by:

$$e_t(h) = X_{t+h} - \hat{X}_t(h) = X_{t+h} - \varphi^h X_t \tag{15.3.1738}$$

And let us recall that we have proved earlier above that for $\text{AR}(1)$:

$$X_t = \varphi^h X_{t-h} + \sum_{k=0}^{h-1} \varphi^k \varepsilon_{t-k} \tag{15.3.1739}$$

And let us consider h big and $|\varphi| < 1$, then we have:

$$X_t \cong \sum_{l=0}^{+\infty} \varphi^k \varepsilon_{t-k} \tag{15.3.1740}$$

Therefore:

$$X_{t+h} \cong \sum_{k=0}^{+\infty} \varphi^k \varepsilon_{t+h-k} \tag{15.3.1741}$$

Given a $\text{MA}(\infty)$ process. Then:

$$\begin{aligned} e_t(h) = X_{t+h} - \hat{X}_t(h) = X_{t+h} - \varphi^h X_t \underset{\substack{|\varphi|<1 \\ h \to +\infty}}{\cong} & \sum_{k=0}^{+\infty} \varphi^k \varepsilon_{t+h-k} - \varphi^h \sum_{k=0}^{+\infty} \varphi^k \varepsilon_{t-k} \\ &= \varepsilon_{t+h} + \varphi \varepsilon_{t+h-1} + \ldots + \varphi^{h-1} \varepsilon_{t+1} + \varphi^h \varepsilon_t + \varphi^{h+1} \varepsilon_{t+1} + \ldots - \varphi^h \varepsilon_t - \varphi^{h+1} \varepsilon_{t-1} - \ldots \\ &= \varepsilon_{t+h} + \varphi \varepsilon_{t+h-1} + \varphi^2 \varepsilon_{t+h-2} + \ldots + \varphi^{h-1} \varepsilon_{t+1} \end{aligned} \tag{15.3.1742}$$

We notice that the forecast error $e_t(h)$ is a process of the type process MA$(h-1)$. Then it comes immediately form this last development:

$$\begin{aligned}&\mathrm{E}(e_t(h))=0\\&\mathrm{V}(e_t(h))=\sigma_\varepsilon^2(1+\varphi^2+\varphi^4+\ldots+\varphi^{2(h-1)})=\sigma_\varepsilon^2\frac{1-\varphi^{2h}}{1-\varphi^2}\end{aligned}\tag{15.3.1743}$$

where the sum of the square has been simplified using a result proved in the section Sequences and Series. Thus we see that the variance of the error forecast does not diverge (at least in the case of an AR (1)) and tends to a finite value (which is not the case of all processes!) when h approaches infinity provided that $\varphi<1$. The difficulty in practice is to select a technique to determine σ_ε^2 (there are a quantity of such techniques: adequacy, Monte Carlo, bootstrapping, Bayesian inference, etc.). This is why the practitioner should never limit himself to the use of a single statistical software to make projections, but at least three.

Why we speak again of that subject will you think maybe ??? For the simple reason that we always need a statistical prediction interval!!! Thus, given that we have a point estimator X_{t+1} which is the realization of the random variable of the forecasted value and its associated mean value $\hat{X}_{t+1}$, we can write:

$$X_{t+h}=\hat{X}_t(h)+e_t(h)\tag{15.3.1744}$$

and assume from the previous developments that the forecast error follows a Normal centered distribution:

$$\left.\begin{aligned}&\mathrm{E}(e_t(h))=0\\&\mathrm{V}(e_t(h))=\sigma_\varepsilon^2\frac{1-\varphi^{2h}}{1-\varphi^2}\end{aligned}\right\}\Rightarrow\text{Assumption: }e_t(h)=\mathcal{N}\left(0,\sigma_\varepsilon^2\frac{1-\varphi^{2h}}{1-\phi^2}\right)\tag{15.3.1745}$$

We then have a given confidence level if σ_ε^2 is completely determined:

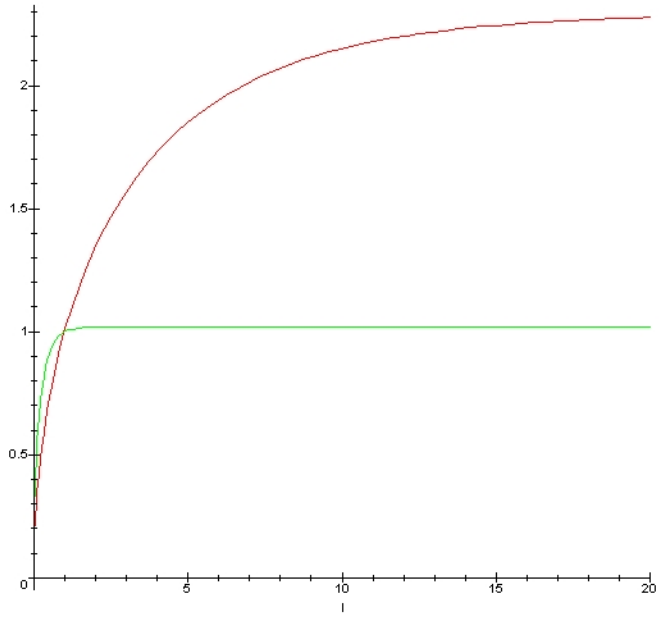
$$\hat{X}_t(h)-Z_{\alpha/2}\sigma_\varepsilon\sqrt{\frac{1-\varphi^{2h}}{1-\varphi^2}}\leq X_{t+h}\leq\hat{X}_t(h)+Z_{\alpha/2}\sigma_\varepsilon\sqrt{\frac{1-\varphi^{2h}}{1-\varphi^2}}\tag{15.3.1746}$$

and if we have only an estimator of $\hat{\sigma}_\varepsilon^2$ then we will take (see section Statistics page 865):

$$\hat{X}_t(h)-T_{\alpha/2}\hat{\sigma}_\varepsilon\sqrt{\frac{1-\varphi^{2h}}{1-\varphi^2}}\leq X_{t+h}\leq\hat{X}_t(h)+T_{\alpha/2}\hat{\sigma}_\varepsilon\sqrt{\frac{1-\varphi^{2h}}{1-\varphi^2}}\tag{15.3.1747}$$

The asymptotic shape of the error term is quite well known by practitioner as it is of the form with Maple™ 4.00b:

```
>plot([sqrt((1-0.9^(2*l))/(1-0.9^2)),sqrt((1-0.2^(2*l))/(1-0.2^2))],l=0..20);
```



Practising these forecasting model show they tend to underestimate the risk (variance). So rather than taking a 95% interval it is perhaps better to take a higher range depending on the return on experience.

Remark

The general case for an AR(p) process is actually over the scope of this textbook. Maybe it will be written for the 5th version...

Suppose now we have an MA(3) process with constant:

$$X_t = \mu + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \theta_3 \varepsilon_{t-3} + \varepsilon_t \tag{15.3.1748}$$

We will see now what are the one step, two step, three step and four step ahead forecast for such a process.

1. For the one step ahead forecast we start from the knowledge that:

   $$X_t = \mu + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \theta_3 \varepsilon_{t-3} + \varepsilon_t \tag{15.3.1749}$$

   Therefore:

   $$X_{t+1} = \mu + \theta_1 \varepsilon_t + \theta_2 \varepsilon_{t-1} + \theta_3 \varepsilon_{t-2} + \varepsilon_{t+1} \tag{15.3.1750}$$

   So we can write:

   $$\begin{aligned} \mathrm{E}(X_{t+1}|X_t) &= \mu + \theta_1 \mathrm{E}(\varepsilon_t|X_t) + \theta_2 \mathrm{E}(\varepsilon_{t-1}|X_t) + \theta_3 \mathrm{E}(\varepsilon_{t-2}|X_t) + \theta_2 \mathrm{E}(\varepsilon_{t+1}|X_t) \\ &= \mu + \theta_1 \varepsilon_t + \theta_2 \varepsilon_{t-1} + \theta_3 \varepsilon_{t-2} + 0 \\ &= \mu + \theta_1 \varepsilon_t + \theta_2 \varepsilon_{t-1} + \theta_3 \varepsilon_{t-2} \end{aligned} \tag{15.3.1751}$$

2. For the two step ahead forecast we start from the knowledge that:

$$X_t = \mu + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \theta_3 \varepsilon_{t-3} + \varepsilon_t \tag{15.3.1752}$$

   Therefore:

$$X_{t+2} = \mu + \theta_1 \varepsilon_{t+1} + \theta_2 \varepsilon_t + \theta_3 \varepsilon_{t-1} + \varepsilon_{t+2} \tag{15.3.1753}$$

   So we can write:

$$\begin{aligned} \mathrm{E}(X_{t+2}|X_t) &= \mu + \theta_1 \mathrm{E}(\varepsilon_{t+1}|X_t) + \theta_2 \mathrm{E}(\varepsilon_t|X_t) + \theta_3 \mathrm{E}(\varepsilon_{t-1}|X_t) + \theta_2 \mathrm{E}(\varepsilon_{t+2}|X_t) \\ &= \mu + \theta_1 \cdot 0 + \theta_2 \varepsilon_t + \theta_3 \varepsilon_{t-1} + 0 \\ &= \mu + \theta_2 \varepsilon_t + \theta_3 \varepsilon_{t-1} \end{aligned} \tag{15.3.1754}$$

3. For the three step ahead forecast we start from the knowledge that:

$$X_t = \mu + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \theta_3 \varepsilon_{t-3} + \varepsilon_t \tag{15.3.1755}$$

   Therefore:

$$X_{t+3} = \mu + \theta_1 \varepsilon_{t+2} + \theta_2 \varepsilon_{t+1} + \theta_3 \varepsilon_t + \varepsilon_{t+3} \tag{15.3.1756}$$

   So we can write:

$$\begin{aligned} \mathrm{E}(X_{t+3}|X_t) &= \mu + \theta_1 \mathrm{E}(\varepsilon_{t+2}|X_t) + \theta_2 \mathrm{E}(\varepsilon_{t+1}|X_t) + \theta_3 \mathrm{E}(\varepsilon_t|X_t) + \theta_2 \mathrm{E}(\varepsilon_{t+3}|X_t) \\ &= \mu + \theta_1 \cdot 0 + \theta_2 \cdot 0 + \theta_3 \varepsilon_t + 0 \\ &= \mu + \theta_3 \varepsilon_t \end{aligned} \tag{15.3.1757}$$

4. For the four step ahead forecast we start from the knowledge that:

$$X_t = \mu + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \theta_3 \varepsilon_{t-3} + \varepsilon_t \tag{15.3.1758}$$

   Therefore:

$$X_{t+4} = \mu + \theta_1 \varepsilon_{t+3} + \theta_2 \varepsilon_{t+2} + \theta_3 \varepsilon_{t+1} + \varepsilon_{t+4} \tag{15.3.1759}$$

   So we can write:

$$\begin{aligned} \mathrm{E}(X_{t+4}|X_t) &= \mu + \theta_1 \mathrm{E}(\varepsilon_{t+3}|X_t) + \theta_2 \mathrm{E}(\varepsilon_{t+2}|X_t) + \theta_3 \mathrm{E}(\varepsilon_{t+1}|X_t) + \theta_2 \mathrm{E}(\varepsilon_{t+4}|X_t) \\ &= \mu + \theta_1 \cdot 0 + \theta_2 \cdot 0 + \theta_3 \cdot 0 + 0 \\ &= \mu \end{aligned} \tag{15.3.1760}$$

Notice that the prediction error can therefore be calculated as (if you compare with the previous four examples):

$$e_{t+h} = X_{t+h} - \hat{X}_{t+h} = X_{t+h} - \mathrm{E}(X_{t+h}|X_t) = \varepsilon_{t+h} + \sum_{i=1}^{h-1} \theta_i \varepsilon_{t+i} \tag{15.3.1761}$$

Hence the variance (mean square error) is given by:

$$\text{V}(e_{t+h})=\sigma_\varepsilon^2\left(1+\sum_{i=1}^{h-1}\theta_i^2\right) \tag{15.3.1762}$$

If ε_t is Gaussian then:

$$\hat{X}_{t+h}=\mathcal{N}\left(\text{E}(X_{t+h}|X_t),\sigma_\varepsilon^2\left(1+\sum_{i=1}^{h-1}\theta_i^2\right)\right) \tag{15.3.1763}$$

We then have a given confidence level if σ_ε^2 is completely determined (using the notation previously adopted of the AR(1) process):

$$\hat{X}_t(h)-Z_{\alpha/2}\sigma_\varepsilon\sqrt{1+\sum_{i=1}^{h-1}\theta_i^2}\leq X_{t+h}\leq \hat{X}_t(h)+Z_{\alpha/2}\sigma_\varepsilon\sqrt{1+\sum_{i=1}^{h-1}\theta_i^2} \tag{15.3.1764}$$

and if we have only an estimator of $\hat{\sigma}_\varepsilon^2$ then we will take (see section Statistics page 865):

$$\hat{X}_t(h)-T_{\alpha/2}\hat{\sigma}_\varepsilon\sqrt{1+\sum_{i=1}^{h-1}\theta_i^2}\leq X_{t+h}\leq \hat{X}_t(h)+T_{\alpha/2}\hat{\sigma}_\varepsilon\sqrt{1+\sum_{i=1}^{h-1}\theta_i^2} \tag{15.3.1765}$$

Here are some examples for a same set of data of the different forecast we can have depending on the chosen model:

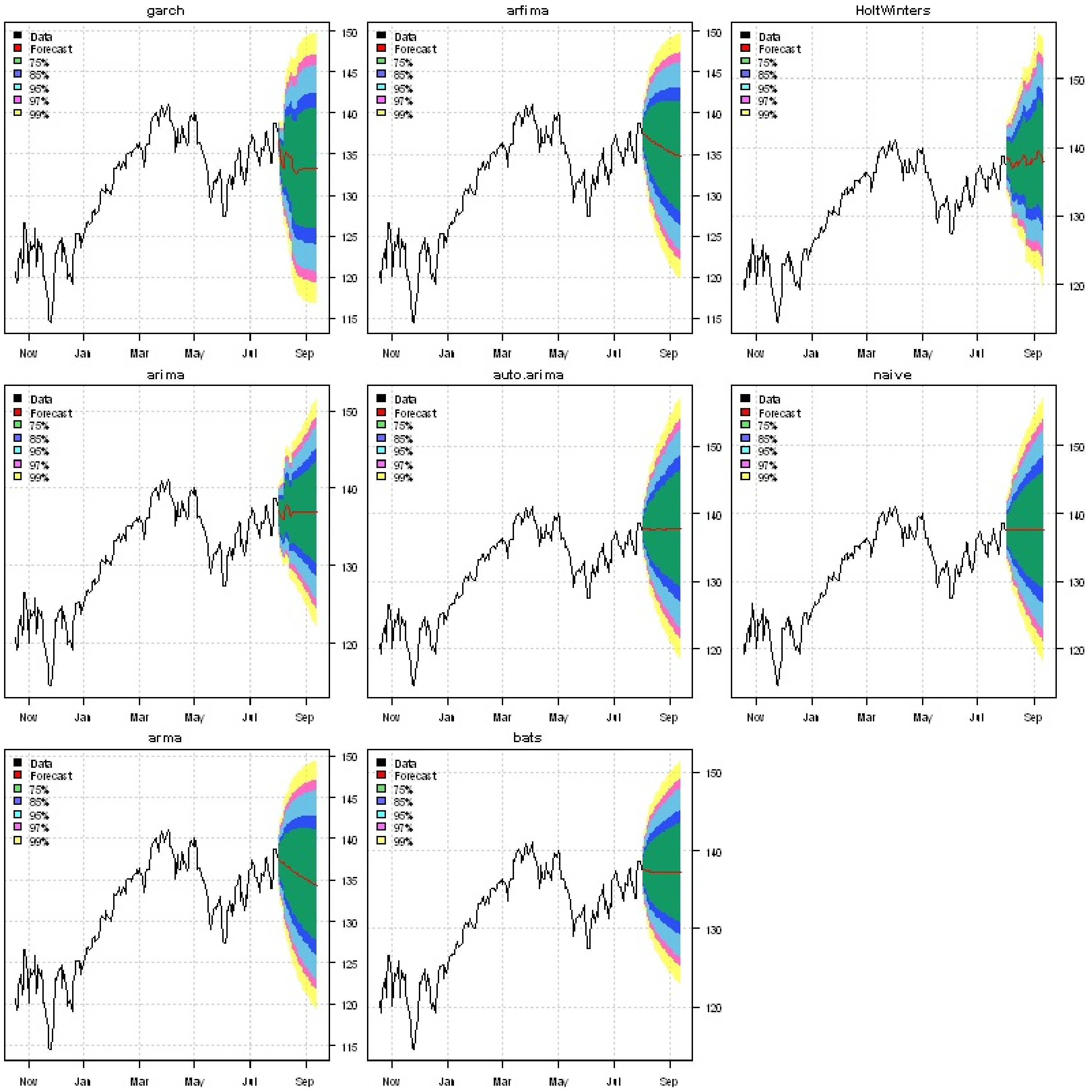


Figure 15.275: Some time series with forecasting (GARCH, ARFIMA, Holt-Winter, ARIMA, etc.) with `R`

15.3.13.5.1 Linear Minimal Mean Square Estimation (LMMSE)

According to what we have seen previously, the reader may have noticed that we can always write:

$$\hat{X}_{t+h} = \mathrm{E}(X_{t+h}|X_t) = a_0 + \sum_{i=1}^{h} a_i X_{t+h-i} \tag{15.3.1766}$$

The idea of forecasting based on Linear Minimal Mean Square Estimation (LMMSE) is to minimize the mean square error of (for practical reasons as optimal prediction in general is largely intractable reasons and due to the analogy with the Gaussian case):

$$\min_{\vec{a}} \mathrm{E}\left[\left(X_{n+h}-\hat{X}_{n+h}\right)^2\right] = \mathrm{E}\left[\left(X_{n+h}-\left(a_0+\sum_{i=1}^{h} a_i X_{t+h-i}\right)\right)^2\right] = \min_{\vec{a}} \mathrm{MSE}(\vec{a}) \tag{15.3.1767}$$

Note the expectation is over all the random variables, which are $X_1,\ldots,X_t,X_{t+h-1}$. We may solve this optimization (minimization) problem analytically:

$$\begin{aligned}\frac{\partial \mathrm{MSE}(\vec{a})}{\partial a_0} &= \mathrm{E}\left[2\left(X_{n+h}-a_0-\sum_{i=1}^{h} a_i X_{t+h-i}\right)\right]=0\\ \frac{\partial \mathrm{MSE}(\vec{a})}{\partial a_j} &= \mathrm{E}\left[2\left(X_{n+h}-a_0-\sum_{i=1}^{h} a_i X_{t+h-i}\right)X_{n+h-j}\right]=0\end{aligned} \tag{15.3.1768}$$

for $j=1\ldots h$. For the first equality above, we have:

$$\mathrm{E}(X_{n+h})-a_0-\sum_{i=1}^{h} a_i \mathrm{E}(X_{t+h-i})=0 \tag{15.3.1769}$$

Rearranging:

$$\mathrm{E}(X_{n+h})-\sum_{i=1}^{h} a_i \mathrm{E}(X_{t+h-i})=a_0 \tag{15.3.1770}$$

But under stationary:

$$\mathrm{E}(X_{n+h})=\mathrm{E}(X_{t+h-i})=\mu \tag{15.3.1771}$$

Therefore:

$$a_0=\mu\left(1-\sum_{i=1}^{h} a_i\right) \tag{15.3.1772}$$

From:

$$\mathrm{E}\left[2\left(X_{n+h}-a_0-\sum_{i=1}^{h} a_i X_{t+h-i}\right)X_{n+h-j}\right]=\mathrm{E}(X_{n+h}X_{n+h-j})-a_0\mathrm{E}(X_{n+h-j})-\sum_{i=1}^{h} a_i\mathrm{E}(X_{t+h-i}X_{n+h-j})=0 \tag{15.3.1773}$$

and substituting in it a_0, we get:

$$\mathrm{E}(X_{n+h}X_{n+h-j})-\mu\left(1-\sum_{i=1}^{h} a_i\right)\mathrm{E}(X_{n+h-j})-\sum_{i=1}^{h} a_i\mathrm{E}(X_{t+h-i}X_{n+h-j})=0 \tag{15.3.1774}$$

Again, under stationary $\mathrm{E}(X_{n+h-j})=\mu$ we can rewrite this:

$$\mathrm{E}(X_{n+h}X_{n+h-j})-\mu\left(1-\sum_{i=1}^{h} a_i\right)\mu-\sum_{i=1}^{h} a_i\mathrm{E}(X_{t+h-i}X_{n+h-j})=0 \tag{15.3.1775}$$

After rearranging:

$$\mathrm{E}(X_{n+h}X_{n+h-j}-\mu^2)-\sum_{i=1}^{h} a_i\mathrm{E}(X_{t+h-i}X_{n+h-j}-\mu^2)=0 \tag{15.3.1776}$$

We can develop even more involving nice mathematical tools. However actually, this is beyond the scope of this textbook (maybe for the 5th edition).

15.3.13.6 TBATS

Let us briefly introduce the "TBATS" (Trigonometric seasonality, Box-Cox transformation, ARMA errors, Trend and Seasonal components) empirical model that takes it roots in exponential smoothing methods and can be described by the following equations:

$$
\begin{aligned}
&y_t = \text{observation at time } t\\
&y_t^{(\lambda)} = \begin{cases} \left(y_t^{\lambda}-1\right)/\lambda & \text{if } \lambda \neq 0 \\ \log(y_t) & \text{if } \lambda = 0 \end{cases} \quad \text{Box-Cox transformation}\\
&y_t^{(\lambda)} = \ell_{t-1}+\phi b_{t-1}+\sum_{i=1}^{M} s_{t-m_i}^{(i)}+d_t \quad M \text{ seasonal periods}\\
&\ell_t = \ell_{t-1}+\phi b_{t-1}+\alpha d_t \quad \text{global and local trend}\\
&b_t = (1-\phi)b+\phi b_{t-1}+\beta d_t \quad \text{trend with mapping}\\
&d_t = \sum_{i=1}^{p}\varphi_i d_{t-i}+\sum_{j=1}^{q}\theta_j\varepsilon_{t-j}+\varepsilon_t \quad \text{ARMA}(p,q) \text{ for residuals}\\
&s_t^{(i)} = \sum_{j=1}^{k_i} s_{j,t}^{(i)} \quad i\text{th seasonal component}\\
&\left.\begin{aligned} s_{j,t}^{(i)} &= s_{j,t-1}^{(i)}\cos\left(\omega_j^{(i)}\right)+s_{j,t-1}^{*(i)}\sin\left(\omega_j^{(i)}\right)+\gamma_1^{(i)}d_t\\ s_{j,t}^{*(i)} &= -s_{j,t-1}^{(i)}\sin\left(\omega_j^{(i)}\right)+s_{j,t-1}^{*(i)}\cos\left(\omega_j^{(i)}\right)+\gamma_2^{(i)}d_t \end{aligned}\right\} \text{Fourier series terms}\\
&\omega_j^{(i)} = \frac{2\pi j}{m_i}
\end{aligned}
\tag{15.3.1777}
$$

where T is the number of seasonalities, m_i is the length of the ith season period, k_i is the number of harmonics (see section Sequences and Series page 1727) for the ith seasonal period, λ is the parameter of the Box-Cox transformation (see below page 8014), α, β are smoothing parameters, ϕ is a trend damping, φ_i, θ_i are the ARMA(p,q) coefficients, γ_1, γ_2 are seasonal smoothing (two for each period).

Each seasonality is modelled by a trigonometric representation based on Fourier series (see section Sequences and Series page 1727). One major advantage of this approach is that it requires only 2 seed states regardless of the length of period. Another advantage is the ability to model seasonal effects of non-integer lengths. For example, given a series of daily observations, one can model leap years with a season of length 365.25.

So in comparison, the main disadvantage of the ARIMA approach is that the seasonality is forced to be periodic, whereas a TBATS model allows for dynamic seasonality.

Unfortunately BATS and TBATS capabilities do not come for free. The method is very generic. Under the hood it builds and evaluates many model candidates and TBATS choose the final model using AIC (see section Numerical Methods page 6420). This results in slowness of the computation.

Remark

For a practical application, the reader can refer to our `R` companion book.

15.3.13.7 Recipes for modelling time series data

To make forecasting using ARIMA models, we need to follow a classical process illustrated below:

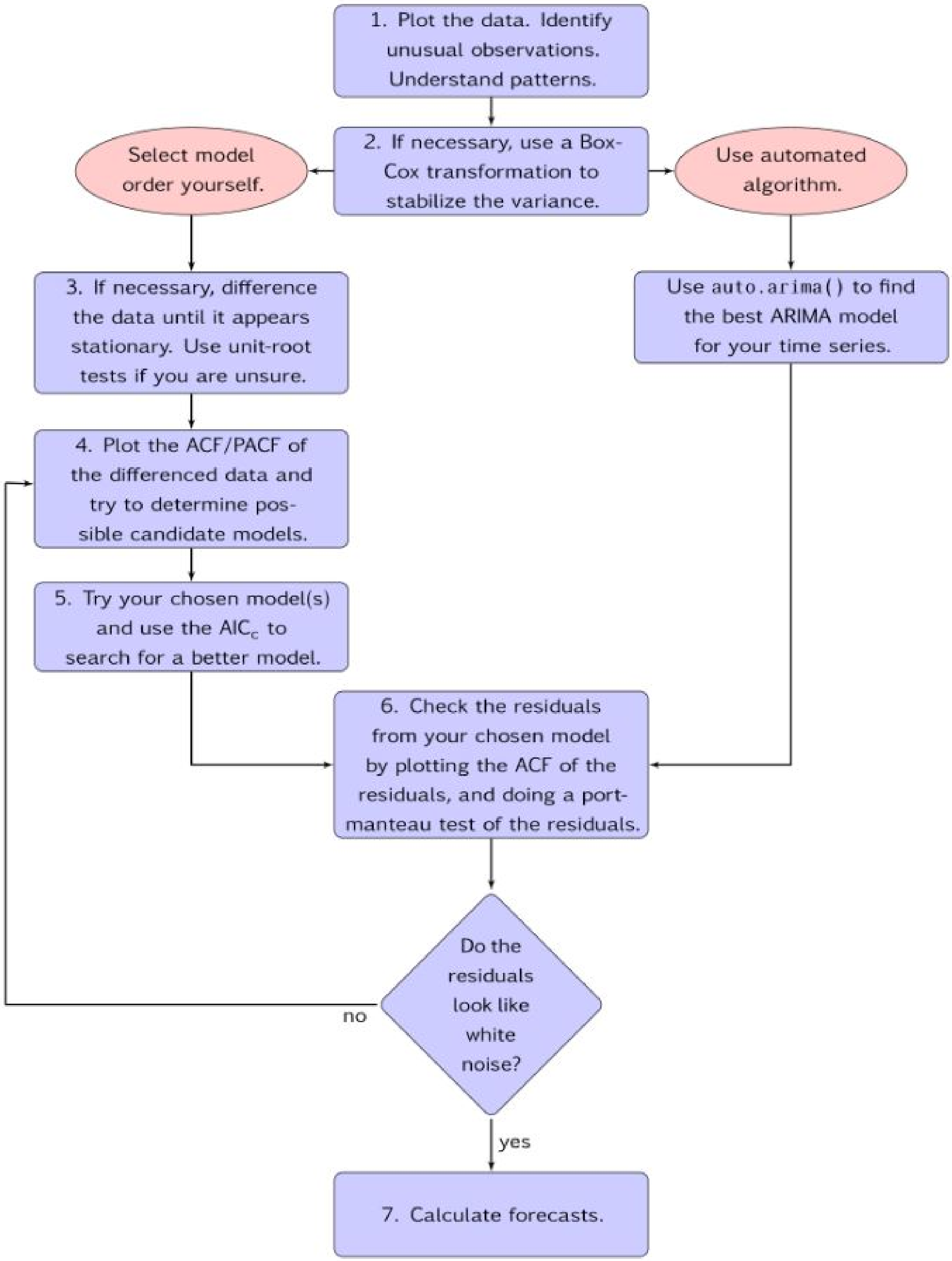

1. Plot the data. Identify unusual observations. Understand patterns.
2. If necessary, use a Box-Cox transformation to stabilize the variance.
Select model order yourself.
Use automated algorithm.
3. If necessary, difference the data until it appears stationary. Use unit-root tests if you are unsure.
Use auto.arima() to find the best ARIMA model for your time series.
4. Plot the ACF/PACF of the differenced data and try to determine possible candidate models.
5. Try your chosen model(s) and use the AICc to search for a better model.
6. Check the residuals from your chosen model by plotting the ACF of the residuals, and doing a portmanteau test of the residuals.
Do the residuals look like white noise?
no
yes
7. Calculate forecasts.

Figure 15.276: **Recipes for modelling time series data**

Our purpose starting from now will be to detail each of the step that need quite technical skills (the steps that will not be detailed like the differentiation[45], will then obviously not be detailed as it is such a simple subtraction).

---

[45] Compute the differences between consecutive observations.

The purpose of "variance stabilization" is to make the variance independent of the mean μ. Keeping in mind that this is one of the assumption of a stationary time series and that it is not only a useful property for time series but in a lot of other fields in statistics!

Suppose that the variance of a time series (or statistical distribution in general) X_t satisfies:

$$\text{V}(X_t)=c^{te}f(\mu) \tag{15.3.1778}$$

We wish to find a transformation denoted $T(\cdot)$ such that $\text{V}(T(X_t))$ is constant. A first-order Taylor series of $T(X_t)$ about μ is:

$$T(X_t)\cong T(\mu)+T'(\mu)(X_t-\mu) \tag{15.3.1779}$$

Now $\text{V}(T(X_t))$ is approximated as (keep in mind that μ is constant and hence $T(\mu)$ is also another constant):

$$\text{V}(T(X_t))\cong\text{V}(T(\mu))+(T'(\mu))^2\text{V}(X_t)+(T'(\mu))\text{V}(\mu)=(T'(\mu))^2\text{V}(X_t)=c^{te}(T'(\mu))^2f(\mu) \tag{15.3.1780}$$

Therefore $T(\cdot)$ is chosen such that:

$$T'(\mu)=\frac{1}{\sqrt{f(\mu)}} \tag{15.3.1781}$$

which implies:

$$T(\mu)=\int\frac{1}{\sqrt{f(\mu)}}\text{d}X_t \tag{15.3.1782}$$

Assuming for example that (after having centered all the X_t):

$$f(\mu)=\sum_i(X_{c,t}-\mu)^2=\sum_i X_{c,t}^2 \tag{15.3.1783}$$

and taking for example only one term, we have:

$$T(\mu)=\int\frac{1}{\sqrt{X_{c,t}^2}}\text{d}X_{c,t}=\int\frac{1}{X_{c,t}}\text{d}X_{c,t}=\ln(X_{c,t}) \tag{15.3.1784}$$

This is why the most standard transformation is the log-transformation where the new series Y_t is given by:

$$Y_t=\log(X_{c,t}) \tag{15.3.1785}$$

An alternative to the log transformation is the "Box-Cox transformation":

$$\boxed{Y_t=\begin{cases}(X_t^\lambda-1)/\lambda & \lambda\neq 0\\ \ln(X_t) & \lambda=0\end{cases}} \tag{15.3.1786}$$

Notice that some other transformations exist!

15.3.13.7.1 Durbin-Watson autocorrelation test

The "Durbin-Watson test" also named "Durbin-Watson statistics" consists typically in testing the null hypothesis that the autocorrelation between the residuals of a regression ε_i is zero (against the alternative hypothesis that this same correlation is non-zero). That is to say:

$$\begin{aligned}H_0 &: \rho_\varepsilon = 0\\ H_1 &: \rho_\varepsilon \neq 0\end{aligned} \tag{15.3.1787}$$

Even if this test is used a lot in linear or non-linear regression to check if the model assumption are satisfied (so it could have its place in the section of Numerical Methods), it is used much more in time series analysis where the practitioner is most often more skilled than people doing regression in corporations and therefore more apt to read the result of the test and perform the necessary corrective actions.

The idea of this test is to write (read well until the end before thinking that the choice of the ratio is empirical!):

$$\begin{aligned}\text{DW} &= \frac{\sum\limits_{i=2}^{n}(\varepsilon_i-\varepsilon_{i-1})^2}{\sum\limits_{i=1}^{n}\varepsilon_i^2} = \frac{\sum\limits_{i=2}^{n}\varepsilon_i^2-2\varepsilon_i\varepsilon_{i-1}+\varepsilon_{i-1}^2}{\sum\limits_{i=1}^{n}\varepsilon_i^2} = \frac{\sum\limits_{i=2}^{n}\varepsilon_i^2}{\sum\limits_{i=1}^{n}\varepsilon_i^2}+\frac{\sum\limits_{i=2}^{n}\varepsilon_{i-1}^2}{\sum\limits_{i=1}^{n}\varepsilon_i^2}-2\frac{\sum\limits_{i=2}^{n}\varepsilon_i\varepsilon_{i-1}}{\sum\limits_{i=1}^{n}\varepsilon_i^2}\\ &\cong 1+1-2\frac{\sum\limits_{i=2}^{n}\varepsilon_i\varepsilon_{i-1}}{\sum\limits_{i=1}^{n}\varepsilon_i^2} = 2-2\frac{\sum\limits_{i=2}^{n}\varepsilon_i\varepsilon_{i-1}}{\sum\limits_{i=1}^{n}\varepsilon_i^2}\end{aligned} \tag{15.3.1788}$$

Remark

Notice the similarity between:

$$\text{DW} = \frac{\sum\limits_{i=2}^{n}(\varepsilon_i-\varepsilon_{i-1})^2}{\sum\limits_{i=1}^{n}\varepsilon_i^2} \tag{15.3.1789}$$

and the equivalent relation of the values of the series itself (instead of the residuals):

$$\text{VN} = \frac{\sum\limits_{i=2}^{n}(x_i-x_{i-1})^2}{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2} \tag{15.3.1790}$$

named the "(Von) Neumann trend test" (obviously on the residuals, by assumption $\bar{\varepsilon}=0$ that's why the denominator is different).

For the latter term:

$$\frac{\sum\limits_{i=2}^{n}\varepsilon_i\varepsilon_{i-1}}{\sum\limits_{i=1}^{n}\varepsilon_i^2} \tag{15.3.1791}$$

notice that we can write it in the following form since the expected mean of the residues are assumed to be zero (therefore this writing is the same!) such as $\text{E}(\varepsilon)=0$. Therefore:

$$\begin{aligned}\frac{\sum\limits_{i=2}^{n}(\varepsilon_i-\text{E}(\varepsilon))(\varepsilon_{i-1}-\text{E}(\varepsilon))}{\sum\limits_{i=1}^{n}(\varepsilon_i-\text{E}(\varepsilon))^2}&=\frac{\dfrac{1}{n}\sum\limits_{i=2}^{n}(\varepsilon_i-\text{E}(\varepsilon))(\varepsilon_{i-1}-\text{E}(\varepsilon))}{\dfrac{1}{n}\sum\limits_{i=1}^{n}(\varepsilon_i-\text{E}(\varepsilon))^2}\\&=\frac{\text{E}\left[(\varepsilon_i-\mu_\varepsilon)(\varepsilon_{i-1}-\mu_\varepsilon)\right]}{\sigma_\varepsilon\sigma_\varepsilon}=\frac{\text{cov}(\varepsilon_i,\varepsilon_{i-1})}{\sigma_\varepsilon^2}=\rho_\varepsilon\end{aligned} \tag{15.3.1792}$$

We thus fall back on the general expression of the autocorrelation seen earlier above. It then comes:

$$\boxed{\text{DW}\cong 2-2\frac{\sum\limits_{i=2}^{n}\varepsilon_i\varepsilon_{i-1}}{\sum\limits_{i=1}^{n}\varepsilon_i^2}=2(1-\rho_\varepsilon)} \tag{15.3.1793}$$

The approximation (valid only if n is sufficiently large) that we have done above justifies the fact that some softwares offer an approximate or exact version.

So far as we know there is not exact distribution law that this statistic follows. Therefore it is tabulated based on Monte Carlo methods. Simulations give that when:

- ρ_ε is close to 1, the DW statistic is close to 0
- ρ_ε is close to 0 (no serial correlation), the DW statistic is close to 2
- ρ_ε is close to -1, the DW statistic is close to 4

Upper and lower critical values, denoted DW_U (or sometimes Q_U) and respectively DW_L (or sometimes Q_L), have been tabulated for different values of k (the number of explanatory variables also sometimes denoted Λ) and of n (also sometimes denoted N) and the rules are:

- If $\text{DW}<\text{DW}_L$ we reject H_0
- If $\text{DW}>\text{DW}_U$ we cannot reject H_0
- If $\text{DW}_L<\text{DW}<\text{DW}_U$ the test is inconclusive

Remark

In the `R` software (see the corresponding companion book) there are actually two functions to run this test. One written `durbinWatsonTest( )` from the `car` package and one written `dwtext()` from the `lmtest` package. Even if they give veeeery close p-values and DW values for small values of n, they diverge quite significantly when n becomes big. This is because of the chosen algorithm. Indeed, the help associated with the `dwtext()` function gives us: «*The* p*-value is computed using a Fortran version of the Applied Statistics Algorithm AS 153 by Farebrother (1980, 1984). This algorithm is named "pan" or "gradsol". For large sample sizes the algorithm might fail to compute the* p *-value; in that case a warning is printed and an approximate* p*-value will be given; this* p *value is computed using a Normal approximation with mean and variance of the Durbin-Watson test statistic*». While the help associated with the `durbinWatsonTest( )` function gives us: «*simulate: if 'TRUE' p-values will be estimated by bootstrapping*».

For the people that may not have access to a spreadsheet or statistical software here are some useful tables:

BOUNDS FOR CRITICAL VALUES OF THE DURBIN-WATSON STATISTIC

1% SIGNIFICANCE POINTS OF Q_L AND Q_U

	$\Lambda=2$		$\Lambda=3$		$\Lambda=4$		$\Lambda=5$		$\Lambda=6$	
N	Q_L	Q_U	Q_L	Q_U	Q_L	Q_U	Q_L	Q_U	Q_L	Q_U
15	0.811	1.069	0.700	1.252	0.591	1.465	0.487	1.705	0.390	1.967
16	0.844	1.087	0.738	1.253	0.633	1.447	0.532	1.664	0.437	1.901
17	0.873	1.102	0.773	1.255	0.672	1.432	0.574	1.631	0.481	1.847
18	0.902	1.118	0.805	1.259	0.708	1.422	0.614	1.604	0.522	1.803
19	0.928	1.133	0.835	1.264	0.742	1.416	0.650	1.583	0.561	1.767
20	0.952	1.147	0.862	1.270	0.774	1.410	0.684	1.567	0.598	1.736
21	0.975	1.161	0.889	1.276	0.803	1.408	0.718	1.554	0.634	1.712
22	0.997	1.174	0.915	1.284	0.832	1.407	0.748	1.543	0.666	1.691
23	1.017	1.186	0.938	1.290	0.858	1.407	0.777	1.535	0.699	1.674
24	1.037	1.199	0.959	1.298	0.881	1.407	0.805	1.527	0.728	1.659
25	1.055	1.210	0.981	1.305	0.906	1.408	0.832	1.521	0.756	1.645
26	1.072	1.222	1.000	1.311	0.928	1.410	0.855	1.517	0.782	1.635
27	1.088	1.232	1.019	1.318	0.948	1.413	0.878	1.514	0.808	1.625
28	1.104	1.244	1.036	1.325	0.969	1.414	0.901	1.512	0.832	1.618
29	1.119	1.254	1.053	1.332	0.988	1.418	0.921	1.511	0.855	1.611
30	1.134	1.264	1.070	1.339	1.006	1.421	0.941	1.510	0.877	1.606
31	1.147	1.274	1.085	1.345	1.022	1.425	0.960	1.509	0.897	1.601
32	1.160	1.283	1.100	1.351	1.039	1.428	0.978	1.509	0.917	1.597
33	1.171	1.291	1.114	1.358	1.055	1.432	0.995	1.510	0.935	1.594
34	1.184	1.298	1.128	1.364	1.070	1.436	1.012	1.511	0.954	1.591
35	1.195	1.307	1.141	1.370	1.085	1.439	1.028	1.512	0.971	1.589
36	1.205	1.315	1.153	1.376	1.098	1.442	1.043	1.513	0.987	1.587
37	1.217	1.322	1.164	1.383	1.112	1.446	1.058	1.514	1.004	1.585
38	1.227	1.330	1.176	1.388	1.124	1.449	1.072	1.515	1.019	1.584
39	1.237	1.337	1.187	1.392	1.137	1.452	1.085	1.517	1.033	1.583
40	1.246	1.344	1.197	1.398	1.149	1.456	1.098	1.518	1.047	1.583
45	1.288	1.376	1.245	1.424	1.201	1.474	1.156	1.528	1.111	1.583
50	1.324	1.403	1.285	1.445	1.245	1.491	1.206	1.537	1.164	1.587
55	1.356	1.428	1.320	1.466	1.284	1.505	1.246	1.548	1.209	1.592
60	1.382	1.449	1.351	1.484	1.317	1.520	1.283	1.559	1.248	1.598
65	1.407	1.467	1.377	1.500	1.346	1.534	1.314	1.568	1.283	1.604
70	1.429	1.485	1.400	1.514	1.372	1.546	1.343	1.577	1.313	1.611
75	1.448	1.501	1.422	1.529	1.395	1.557	1.368	1.586	1.340	1.617
80	1.465	1.514	1.440	1.541	1.416	1.568	1.390	1.595	1.364	1.624
85	1.481	1.529	1.458	1.553	1.434	1.577	1.411	1.603	1.386	1.630
90	1.496	1.541	1.474	1.563	1.452	1.587	1.429	1.611	1.406	1.636

BOUNDS FOR CRITICAL VALUES OF THE DURBIN-WATSON STATISTIC

2.5% SIGNIFICANCE POINTS OF Q_L AND Q_U

	$\Lambda = 2$		$\Lambda = 3$		$\Lambda = 4$		$\Lambda = 5$		$\Lambda = 6$	
N	Q_L	Q_U	Q_L	Q_U	Q_L	Q_U	Q_L	Q_U	Q_L	Q_U
15	0.949	1.222	0.827	1.405	0.706	1.615	0.588	1.848	0.478	2.099
16	0.980	1.235	0.864	1.403	0.748	1.594	0.636	1.806	0.527	2.035
17	1.009	1.248	0.899	1.403	0.788	1.578	0.680	1.773	0.574	1.983
18	1.035	1.261	0.930	1.405	0.825	1.567	0.720	1.746	0.619	1.939
19	1.060	1.274	0.959	1.407	0.859	1.558	0.758	1.724	0.660	1.902
20	1.082	1.286	0.988	1.410	0.890	1.551	0.794	1.705	0.699	1.871
21	1.104	1.297	1.012	1.415	0.920	1.546	0.826	1.691	0.734	1.845
22	1.124	1.308	1.036	1.419	0.947	1.543	0.858	1.678	0.769	1.823
23	1.144	1.319	1.059	1.424	0.973	1.541	0.887	1.668	0.801	1.804
24	1.161	1.329	1.080	1.429	0.997	1.539	0.914	1.659	0.830	1.787
25	1.178	1.339	1.099	1.435	1.019	1.539	0.939	1.652	0.859	1.773
26	1.194	1.348	1.118	1.439	1.041	1.538	0.964	1.646	0.886	1.761
27	1.208	1.358	1.135	1.445	1.061	1.539	0.986	1.641	0.911	1.751
28	1.222	1.367	1.153	1.450	1.080	1.540	1.007	1.637	0.934	1.742
29	1.236	1.375	1.168	1.455	1.098	1.541	1.028	1.634	0.958	1.734
30	1.249	1.383	1.183	1.460	1.115	1.542	1.047	1.632	0.978	1.727
31	1.261	1.391	1.197	1.465	1.132	1.544	1.066	1.630	0.999	1.721
32	1.273	1.399	1.211	1.469	1.147	1.546	1.083	1.628	1.018	1.715
33	1.284	1.406	1.224	1.474	1.163	1.548	1.099	1.627	1.037	1.711
34	1.294	1.413	1.236	1.479	1.176	1.550	1.115	1.626	1.054	1.707
35	1.305	1.420	1.248	1.484	1.190	1.553	1.131	1.626	1.071	1.704
36	1.315	1.426	1.259	1.488	1.203	1.555	1.145	1.625	1.087	1.701
37	1.324	1.433	1.270	1.493	1.215	1.557	1.159	1.625	1.102	1.698
38	1.333	1.439	1.281	1.497	1.227	1.560	1.173	1.625	1.117	1.695
39	1.342	1.445	1.291	1.501	1.238	1.562	1.185	1.626	1.131	1.693
40	1.350	1.451	1.300	1.506	1.249	1.564	1.197	1.626	1.144	1.692
45	1.388	1.477	1.343	1.525	1.298	1.576	1.252	1.630	1.204	1.687
50	1.420	1.500	1.380	1.543	1.338	1.588	1.297	1.636	1.255	1.685
55	1.447	1.520	1.411	1.559	1.373	1.600	1.335	1.642	1.297	1.686
60	1.471	1.538	1.438	1.573	1.404	1.610	1.369	1.649	1.333	1.688
65	1.492	1.554	1.461	1.587	1.430	1.620	1.398	1.655	1.365	1.691
70	1.511	1.568	1.482	1.599	1.453	1.630	1.424	1.662	1.393	1.695
75	1.528	1.582	1.501	1.610	1.474	1.638	1.446	1.668	1.418	1.699
80	1.543	1.594	1.518	1.619	1.493	1.647	1.467	1.674	1.441	1.703
85	1.557	1.605	1.534	1.629	1.510	1.654	1.485	1.680	1.461	1.707
90	1.570	1.614	1.548	1.638	1.525	1.662	1.502	1.686	1.479	1.711

BOUNDS FOR CRITICAL VALUES OF THE DURBIN-WATSON STATISTIC

5% SIGNIFICANCE POINTS OF Q_L AND Q_U

	$\Lambda = 2$		$\Lambda = 3$		$\Lambda = 4$		$\Lambda = 5$		$\Lambda = 6$	
N	Q_L	Q_U	Q_L	Q_U	Q_L	Q_U	Q_L	Q_U	Q_L	Q_U
15	1.077	1.361	0.945	1.543	0.814	1.750	0.685	1.977	0.562	2.220
16	1.106	1.371	0.982	1.539	0.857	1.728	0.734	1.935	0.615	2.157
17	1.133	1.381	1.015	1.536	0.897	1.710	0.779	1.900	0.664	2.104
18	1.158	1.392	1.046	1.535	0.933	1.696	0.820	1.872	0.710	2.060
19	1.180	1.401	1.075	1.535	0.967	1.685	0.859	1.848	0.752	2.022
20	1.201	1.411	1.100	1.537	0.998	1.676	0.894	1.828	0.792	1.991
21	1.221	1.420	1.125	1.538	1.026	1.669	0.927	1.812	0.828	1.964
22	1.240	1.429	1.147	1.541	1.053	1.664	0.958	1.797	0.863	1.940
23	1.257	1.437	1.168	1.543	1.078	1.660	0.986	1.786	0.895	1.919
24	1.273	1.446	1.188	1.546	1.101	1.657	1.013	1.775	0.925	1.902
25	1.288	1.454	1.206	1.550	1.123	1.654	1.038	1.767	0.953	1.886
26	1.302	1.461	1.224	1.553	1.143	1.652	1.062	1.759	0.979	1.873
27	1.316	1.468	1.240	1.556	1.162	1.651	1.083	1.753	1.004	1.861
28	1.328	1.476	1.255	1.560	1.181	1.650	1.104	1.747	1.028	1.850
29	1.341	1.483	1.270	1.563	1.198	1.650	1.124	1.743	1.050	1.841
30	1.352	1.489	1.284	1.567	1.214	1.650	1.143	1.739	1.070	1.833
31	1.363	1.496	1.297	1.570	1.229	1.650	1.160	1.735	1.090	1.825
32	1.373	1.502	1.309	1.573	1.244	1.650	1.177	1.732	1.109	1.819
33	1.383	1.508	1.321	1.577	1.258	1.651	1.193	1.730	1.127	1.813
34	1.393	1.514	1.332	1.580	1.271	1.652	1.208	1.728	1.144	1.807
35	1.402	1.519	1.343	1.584	1.283	1.653	1.222	1.726	1.160	1.803
36	1.411	1.524	1.354	1.587	1.295	1.654	1.236	1.725	1.175	1.799
37	1.419	1.530	1.364	1.590	1.307	1.655	1.249	1.723	1.190	1.795
38	1.427	1.535	1.373	1.594	1.317	1.656	1.261	1.723	1.204	1.792
39	1.435	1.540	1.382	1.597	1.328	1.658	1.273	1.722	1.218	1.789
40	1.442	1.544	1.391	1.600	1.338	1.659	1.285	1.721	1.231	1.786
45	1.475	1.566	1.430	1.615	1.383	1.666	1.336	1.720	1.287	1.776
50	1.503	1.585	1.462	1.628	1.421	1.674	1.378	1.721	1.334	1.771
55	1.527	1.601	1.490	1.640	1.452	1.681	1.414	1.724	1.374	1.768
60	1.549	1.616	1.514	1.652	1.480	1.689	1.444	1.727	1.408	1.767
65	1.567	1.629	1.536	1.662	1.503	1.696	1.471	1.731	1.438	1.767
70	1.583	1.641	1.554	1.671	1.524	1.703	1.494	1.735	1.464	1.768
75	1.598	1.652	1.571	1.680	1.543	1.709	1.515	1.739	1.486	1.770
80	1.611	1.662	1.586	1.688	1.560	1.715	1.534	1.743	1.507	1.772
85	1.624	1.671	1.600	1.696	1.575	1.721	1.551	1.747	1.525	1.774
90	1.635	1.679	1.612	1.703	1.589	1.726	1.566	1.751	1.542	1.776

~70 %

★★★☆☆

71 votes, 64.21%

15.4. Quantitative Management

THE aim of this chapter is to introduce the main mathematical techniques of production and management, maintenance and quality whose use has become essential for engineers, managers and executives of modern companies (those who can make science, the others make methodology ...) and which are the minimum-minimorum of business knowledge in general management. These are techniques that date back to the early 120th century (holocene calendar) and were first used for the majority by the U.S. military and then gradually implemented in the largest global companies.

Moreover, it is an excellent section for the general culture of the physicist or mathematician ... and an achievement for the engineer! And as high level practitioners say: a management model without maths is an incomplete model!

Many techniques are however not presented in this section because they have already been proved in other chapters of this textbook. Indeed, complex, critical and cutting edge management techniques make a huge usage of statistics, probabilities, decision theory, graph theory (Markov chains or not, complete graphs for the management of communication, etc.), financial analysis and optimization algorithms. Because whole chapters in this textbook are already dedicated to those enumerated subjects, it would be redundant to return on them!

Remark

We often speak about 3M (Mathematical Methods for Management) or even sometimes about "scientific management" to describe the set of mathematical tools applied to management (there exists also a training curriculum on the subject of fifty days...). Some people also use the abbreviation "SciProM" for "Scientific Project Management) when it is project oriented and more generally of "Evidence-Based Management". A common English term that became fashionable in Europe to describe this area of application is also the "Decisioneering" referring to the fact that these are tools decision support for engineers.

Before starting let us indicate that despite their intrinsic importance, quantitative management techniques (also including operational research techniques, Monte Carlo simulation and Bootstrapping as studied in the section of Numerical Methods and also Six Sigma as studied in the section of Industrial Engineering and the Decision Theory) are still little used in industry, either because of lack of training of policy makers (except in the financial sector), or by the lack of relevance of the tool or the difficulty to implement it. The main concerns raised by decision makers in the application of mathematical models in corporations are:

1. **The tools are too abstract**: In such situations, we must always remember that... concrete stuff are abstract concepts to which we are accustomed!

2. **A limited consideration of external and internal factors**: For strategic reasons, the pure and perfect answer to a mathematical solution seems rarely applicable de facto. Although

quantitative techniques incorporate many factors, though some aspects are relatively easy to model in the mathematical sense of the term (cost, profitability, distance, time, pace... for example), other elements, however, are more difficult to model: legal constraints, commercial motivation to block a competitor, importance of networking, social climate, etc. The weight of these elements in the decision is however important, sometimes decisive!

3. **A significant investment**: The mathematical tool itself requires a high level of mathematical knowledge, a good ability to model problems and to describe factors. These constraints are consuming a lot of time and a certain amount of money (whether through internal development, which consumes resources - or by external development, which consumes money). It is then necessary to find a balance between the investment required and expected returns.

4. **For infrequent events**: The company does not have the opportunity the use the "effect of experience" and therefore a time to another, the problem is with a different service or managers that have changed between two studies or do not communicate between them. It is therefore difficult to maintain the skills within the company. The decision-maker will have to take these aspects into account when deciding whether or not to implement quantitative techniques in his business.

We can also argue that the fact that Scientific Management Techniques (SMT) or the Mathematical Methods of Management (3M) are not used generally is because:

1. Depending on the context, the extra power offered by SMT/3M does not necessarily justify the efforts necessary to their learning.

2. The SMT/3M are demanding in reasoning and rigorous, they may not be suitable for people who know neither one nor the other ...

Remark

Beware of corporations - especially multinationals - that are looking for specialists or graduate in project management or supply chain that master Microsoft® Excel®, Microsoft Access or VBA. Because it means that they use low-level tools to do a job which ought to be done with high-level efficient tools (what Microsoft® Excel® and Microsoft Access aren't obviously!!!). So in terms of internal organization, you can ensure that such corporations organize and analyse anything, anyhow, with unsuitable tools and therefore that there is a general mess internally.

We have chosen in this section to present the more conventional quantitative techniques in an ascending order of difficulty (technical level). This seems to us to this date the most relevant pedagogical choice...

Let us give a schematic summary of the scientific method that we studied in the Introduction chapter and that applies pretty well to the business world regarding scientific management:

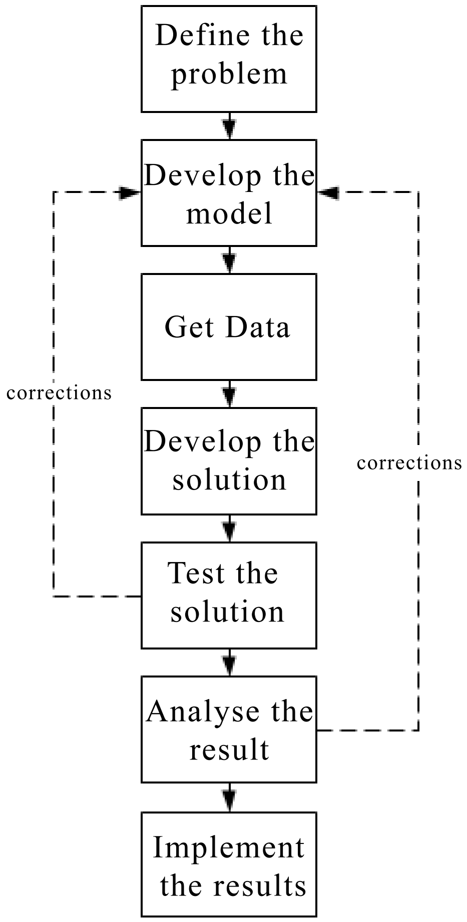

Define the problem
Develop the model
Get Data
Develop the solution
Test the solution
Analyse the result
Implement the results
corrections
corrections

Figure 15.277: **Typical simple process**

The simple subjects below are given without practical examples while those that exceed the high-school level are accompanied by practical examples more or less detailed.

One last word to close this introduction... the quantitative manager or quantitative analyst has to know something... Sure you want to write verbose explanations of your analysis but never forget that Executives (at least actually) are looking only for what they name WIIFM (What's In It For Me). So you must understand that you are paid most of times to produce headlines and not to write novels...

Remark

For people interested only on the methodological aspect of quantitative management, we strongly recommend the lecture of the book Evidence-Based Management: *How to Use Evidence to Make Better Organizational Decisions* ([41]).

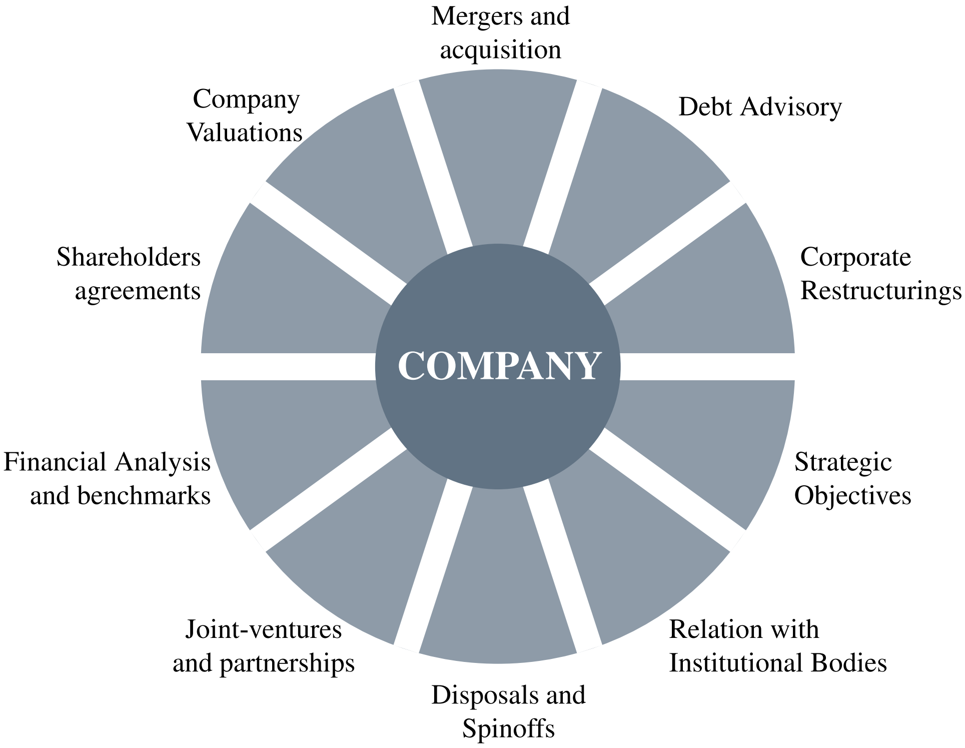
15.4.1 Corporate and Government Finance Management

Corporate and Government finance is the area of finance dealing with the sources of funding and the capital structure of corporations and governments and the actions that managers and politicians should take to increase the value of the firm/administration to the shareholders, the competitiveness/attractiveness of their business/nation as well as the tools and analysis used to allocate financial resources in a certain or probabilistic world (i.e. Universe). The primary goal of corporate and governments finance is to maximize or increase shareholder/citizen value through various adjustment variables and anticipate market trends and risks.

Investment analysis (or capital budgeting) is concerned with the setting of criteria about which value-adding projects should receive investment funding, and whether to finance that investment with equity or debt capital.

As many subject in Corporate and Government Finance are related to investment banking topics we will not repeat in this section the study of Actuarial, Interests calculations, Loan Amortization/Repayment, Portfolio Theory and Forecasting that we have already study in detail in the section Economy (page 7485).

Also the reader has to know that Corporate and Government Finance is not only related to investment but also to benchmarking and optimization and as we have already study these subject in the sections of Game and Decision Theory and Numerical Methods/Analysis we will not come back on these latter.


Mergers and acquisition
Debt Advisory
Corporate Restructurings
Strategic Objectives
Relation with Institutional Bodies
Disposals and Spinoffs
Joint-ventures and partnerships
Financial Analysis and benchmarks
Shareholders agreements
Company Valuations
COMPANY

Corporate Finance management is an every day challenge for every good CXO (CEO, CFO, CTO, CRO, etc.) and costs optimization should be his everyday way of life by always taking care of his own cognitive biases, market uncertainties (nothing is "sure" in Economy/Finance[46]) and having always a reserve of alternative strategies in case of economic conjuncture variations (that in the probabilistic point of view may always have the possibility to occur!).

All CXO have always to face risks (especially "risk–benefit ratio[47]") and uncertainties hence the fact that any good manager associate arguments with probabilities or with precautions (some bad managers avoid risks subjectively but the main purpose of a real entrepreneur is to take risks to innovate and perform always better!):

		Uncertainty	
		Low	*High*
Risk	*Low*	Benchmark	Operational Hedging
	High	Flexible Financial Hedging	Flexible Operationa Hedging

Table 15.67: **Risk-Uncertainty Matrix CXO knowledge**

The list of possible risks is huge and we can find some exhaustive lists in good project management books, but here is a quite good summary of most important financial related type of risks:

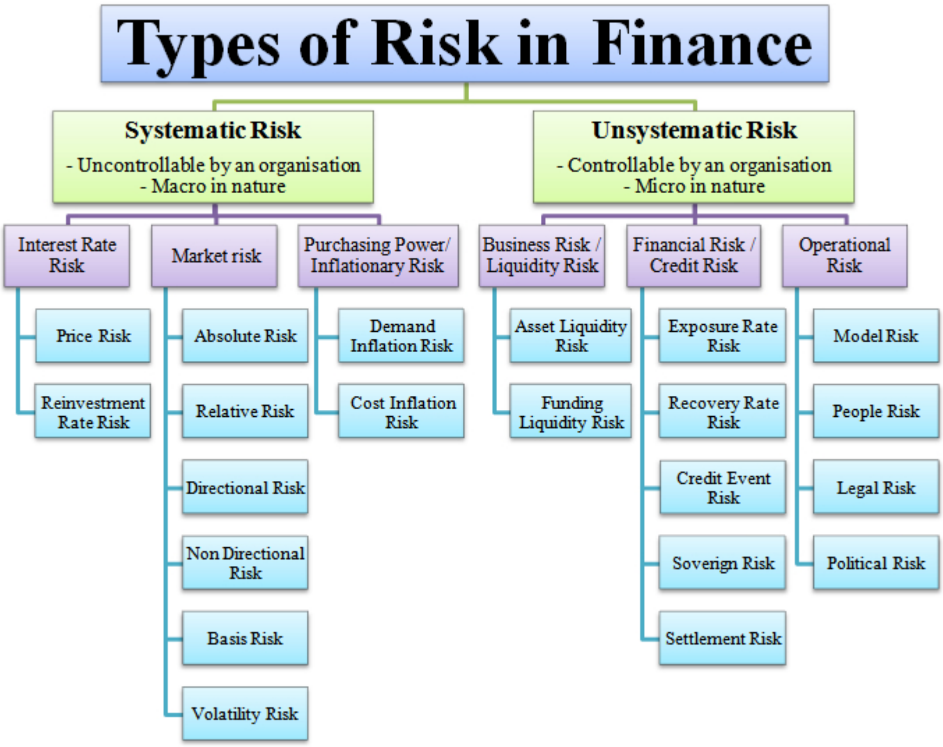

Types of Risk in Finance
Systematic Risk
- Uncontrollable by an organisation
- Macro in nature
Unsystematic Risk
- Controllable by an organisation
- Micro in nature
Interest Rate Risk
Market risk
Purchasing Power/ Inflationary Risk
Business Risk / Liquidity Risk
Financial Risk / Credit Risk
Operational Risk
Price Risk
Reinvestment Rate Risk
Absolute Risk
Relative Risk
Directional Risk
Non Directional Risk
Basis Risk
Volatility Risk
Demand Inflation Risk
Cost Inflation Risk
Asset Liquidity Risk
Funding Liquidity Risk
Exposure Rate Risk
Recovery Rate Risk
Credit Event Risk
Soverign Risk
Settlement Risk
Model Risk
People Risk
Legal Risk
Political Risk

Figure 15.278: **Typical type of financial risks**

---

[46] Never forget the adage: "The problem in the world is that intelligent people are full of doubts, while the idiots are full of confidence"

[47] It is the ratio of the risk of an action to its potential benefits. For example, driving an automobile is a risk most people take daily, also since it is mitigated by the controlling factor of their perception of their individual ability to manage the risk-creating situation.

15.4.1.1 Basic Accounting Equation

The "basic accounting equation", also named the "balance sheet equation", represents the relationship between the assets, liabilities, and owner's equity of a business. It is the foundation for the double-entry bookkeeping system. For each transaction, the total debits equal the total credits. It can be expressed as:

$$\text{Assets (what you own)} = \text{Liabilities (what you owe)} + \text{Equities (what you possess)} \tag{15.4.1}$$

or more formally:

$$\boxed{\mathcal{A}=\mathcal{E}+\mathcal{L}} \tag{15.4.2}$$

Since every business transaction affects at least two of a company's accounts, the accounting equation will always be "in balance", meaning the left side should always equal the right side. Thus, the accounting formula essentially shows that what the firm owns (its assets) is purchased by either what it owes (its liabilities) or by what its owners invest (its shareholders equity or capital).

✍**Example:**

A student buys a computer for 945.-. This student borrowed 500.- from his friend and spent another 445.- earned from his part-time job. Now his asset is worth 945.-, liabilities are 500.-, and equity 445.-.

We have typically:

ASSETS:

- Cash
- Temporary Investments
- Accounts Receivable
- Merchandise Inventory
- Office Supplies
- Buildings
- Property & Equipment
- Patents
- Goodwill

LIABILITIES:

- Accounts Payable
- Rent Payable
- Salaries Payable
- Estimated Warranty Liability
- Dividends Payable
- Long Term Liabilities
- Discount on Bonds Payable
- Tax Payable

EQUITIES:

- Owner's Equity
- Owner's Withdrawals
- Common Shares
- Preferred Shares
- Retained Earnings
- Dividends

15.4.1.2 Ratio Analysis

A problem involved when comparing companies (or organizations) of different sizes is to calculate and compare ratios that can provide very useful managerial information about efficiency. Using ratios eliminates as we know the size problem because the size effectively divides out. We are then left with percentages, multiples, or time periods.

There is a problem in discussing ratios. Because a ratio is simply one number divided by another they are incapable of accommodating multiple inputs and outputs when accurate objective relative weights for inputs and outputs are not known, and because there are so many accounting numbers out there, we could examine a huge number of possible ratios (remember the reference on the UBS catalogue on financial indicators that we have mentioned in the section Economy[48]).

Remark

Sadly as far as we know and still in 12016 (holocene calendar) there are no ISO standard for corporate indicators.

Most people have a favourite indicator. We will restrict here ourselves to a representative sampling. In this section, we only want to introduce some very common used financial ratios and as the reader will see it is quick boring as the variables are never robustly well defined (furthermore most teachers and practitioners of corporate finance do not know how to write a technical paper or book properly...). The chosen indicators are not necessarily the ones we think are the best. In fact, some of them may strike the reader as illogical or not as useful as some alternatives. If they do, the reader must no be concerned. As an practitioner, the reader can always decide how to compute his own ratios.

What the reader needs to worry about is the fact that different people and different sources seldom compute these ratios in exactly the same way, and this leads to much confusion. The specific definitions we use here may or may not be the same as the one the reader have seen or will see elsewhere. If the reader is ever using ratios as a tool for analysis, he should be careful to document how he calculate each one. And if the reader is comparing his numbers to numbers from another source, be sure you know how those numbers are computed!

We will defer much of our discussion of how ratios are used and some problems that come up with using them. For now, for each of the ratios we discuss, we consider several questions:

1. How is it computed?
2. What is it intended to measure, and why might we be interested?
3. What is the unit of measurement?
4. What might a high or low value tell us? How might such values be misleading?
5. How could this measure be improved?

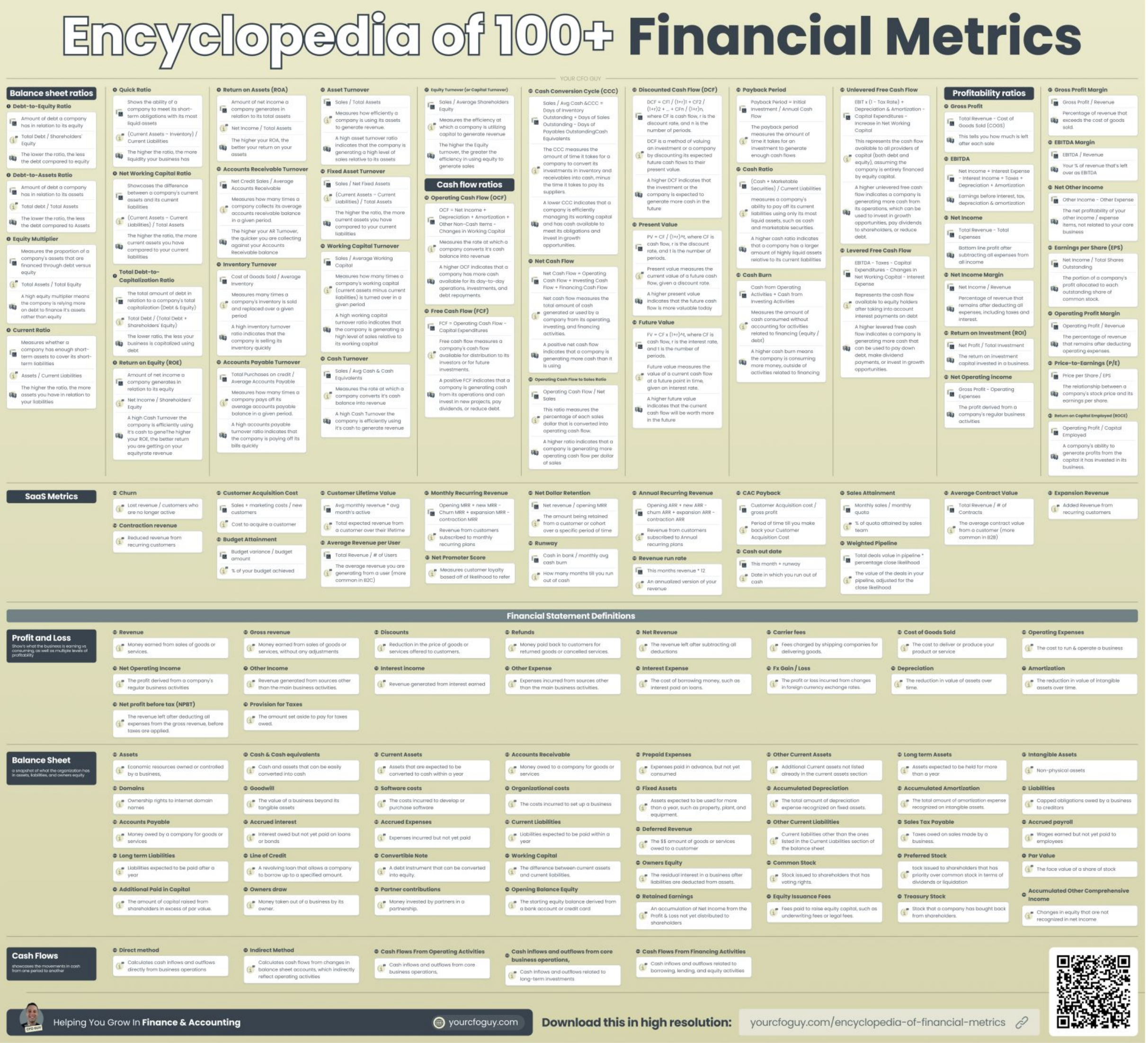
---

[48] The reader can also refer to a book like *Financial Ratios for Executives, How to Assess Company Strength, Fix Problems, and Make Better Decisions* of Michael Rist and Albert J. Pizzica

Financial ratios are traditionally grouped into the following categories:

1. Short-term solvency, or liquidity, ratios.
2. Long-term solvency, or financial leverage, ratios.
3. Asset management, or turnover, ratios.
4. Profitability ratios.
5. Market value ratios.

There are plenty of corporate and accounting ratios/indicators and other metrics in practice, easily more than a hundred. Here is a quite well known poster giving a definition of one hundred of the most important one[49]:



Encyclopedia of 100+ Financial Metrics

YOUR CFO GUY

Balance sheet ratios

Debt-to-Equity Ratio
- Amount of debt a company has in relation to its equity
- Total Debt / Shareholders' Equity
- The lower the ratio, the less the debt compared to equity

Debt-to-Assets Ratio
- Amount of debt a company has in relation to its assets
- Total debt / Total Assets
- The lower the ratio, the less the debt compared to Assets

Equity Multiplier
- Measures the proportion of a company's assets that are financed through debt versus equity
- Total Assets / Total Equity
- A high equity multiplier means the company is relying more on debt to finance it's assets rather than equity

Current Ratio
- Measures whether a company has enough short-term assets to cover its short-term liabilities
- Assets / Current Liabilities
- The higher the ratio, the more assets you have in relation to your liabilities

Quick Ratio
- Shows the ability of a company to meet its short-term obligations with its most liquid assets
- (Current Assets - Inventory) / Current Liabilities
- The higher the ratio, the more liquidity your business has

Net Working Capital Ratio
- Showcases the difference between a company's current assets and its current liabilities
- (Current Assets - Current Liabilities) / Total Assets
- The higher the ratio, the more current assets you have compared to your current liabilities

Total Debt-to-Capitalization Ratio
- The total amount of debt in relation to a company's total capitalization (Debt & Equity)
- Total Debt / (Total Debt + Shareholders' Equity)
- The lower ratio, the less your business is capitalized using debt

Return on Equity (ROE)
- Amount of net income a company generates in relation to its equity
- Net Income / Shareholders' Equity
- A high Cash Turnover the company is efficiently using it's cash to geneThe higher your ROE, the better return you are getting on your equityrate revenue

Return on Assets (ROA)
- Amount of net income a company generates in relation to its total assets
- Net Income / Total Assets
- The higher your ROA, the better your return on your assets

Accounts Receivable Turnover
- Net Credit Sales / Average Accounts Receivable
- Measures how many times a company collects its average accounts receivable balance in a given period
- The higher your AR Turnover, the quicker you are collecting against your Accounts Receivable balance

Inventory Turnover
- Cost of Goods Sold / Average Inventory
- Measures many times a company's inventory is sold and replaced over a given period
- A high inventory turnover ratio indicates that the company is selling its inventory quickly

Accounts Payable Turnover
- Total Purchases on credit / Average Accounts Payable
- Measures how many times a company pays off its average accounts payable balance in a given period.
- A high accounts payable turnover ratio indicates that the company is paying off its bills quickly

Asset Turnover
- Sales / Total Assets
- Measures how efficiently a company is using its assets to generate revenue.
- A high asset turnover ratio indicates that the company is generating a high level of sales relative to its assets

Fixed Asset Turnover
- Sales / Net Fixed Assets
- (Current Assets - Current Liabilities) / Total Assets
- The higher the ratio, the more current assets you have compared to your current liabilities

Working Capital Turnover
- Sales / Average Working Capital
- Measures how many times a company's working capital (current assets minus current liabilities) is turned over in a given period
- A high working capital turnover ratio indicates that the company is generating a high level of sales relative to its working capital

Cash Turnover
- Sales / Avg Cash & Cash Equivalents
- Measures the rate at which a company converts it's cash balance into revenue
- A high Cash Turnover the company is efficiently using it's cash to generate revenue

Equity Turnover (or Capital Turnover)
- Sales / Average Shareholders Equity
- Measures the efficiency at which a company is utilizing capital to generate revenue
- The higher the Equity turnover, the greater the efficiency in using equity to generate sales

Cash flow ratios

Operating Cash Flow (OCF)
- OCF = Net Income + Depreciation + Amortization + Other Non-Cash Items - Changes in Working Capital
- Measures the rate at which a company converts it's cash balance into revenue
- A higher OCF indicates that a company has more cash available for its day-to-day operations, investments, and debt repayments.

Free Cash Flow (FCF)
- FCF = Operating Cash Flow - Capital Expenditures
- Free cash flow measures a company's cash flow available for distribution to its investors or for future investments.
- A positive FCF indicates that a company is generating cash from its operations and can invest in new projects, pay dividends, or reduce debt.

Cash Conversion Cycle (CCC)
- Sales / Avg Cash &CCC = Days of Inventory Outstanding + Days of Sales Outstanding - Days of Payables OutstandingCash Equivalents
- The CCC measures the amount of time it takes for a company to convert its investments in inventory and receivables into cash, minus the time it takes to pay its suppliers.
- A lower CCC indicates that a company is efficiently managing its working capital and has cash available to meet its obligations and invest in growth opportunities.

Net Cash Flow
- Net Cash Flow = Operating Cash Flow + Investing Cash Flow + Financing Cash Flow
- Net cash flow measures the total amount of cash generated or used by a company from its operating, investing, and financing activities.
- A positive net cash flow indicates that a company is generating more cash than it is using

Operating Cash Flow to Sales Ratio
- Operating Cash Flow / Net Sales
- This ratio measures the percentage of each sales dollar that is converted into operating cash flow.
- A higher ratio indicates that a company is generating more operating cash flow per dollar of sales

Discounted Cash Flow (DCF)
- DCF = CF1 / (1+r)1 + CF2 / (1+r)2 + ... + CFn / (1+r)n, where CF is cash flow, r is the discount rate, and n is the number of periods.
- DCF is a method of valuing an investment or a company by discounting its expected future cash flows to their present value.
- A higher DCF indicates that the investment or the company is expected to generate more cash in the future

Present Value
- PV = CF / (1+r)^t, where CF is cash flow, r is the discount rate, and t is the number of periods.
- Present value measures the current value of a future cash flow, given a discount rate.
- A higher present value indicates that the future cash flow is more valuable today

Future Value
- FV = CF x (1+r)^t, where CF is cash flow, r is the interest rate, and t is the number of periods.
- Future value measures the value of a current cash flow at a future point in time, given an interest rate.
- A higher future value indicates that the current cash flow will be worth more in the future

Payback Period
- Payback Period = Initial Investment / Annual Cash Flow
- The payback period measures the amount of time it takes for an investment to generate enough cash flows

Cash Ratio
- (Cash + Marketable Securities) / Current Liabilities
- measures a company's ability to pay off its current liabilities using only its most liquid assets, such as cash and marketable securities.
- A higher cash ratio indicates that a company has a larger amount of highly liquid assets relative to its current liabilities

Cash Burn
- Cash from Operating Activities + Cash from Investing Activities
- Measures the amount of cash consumed without accounting for activities related to financing (equity / debt)
- A higher cash burn means the company is consuming more money, outside of activities related to financing

Unlevered Free Cash Flow
- EBIT x (1 - Tax Rate) + Depreciation & Amortization - Capital Expenditures - Increase in Net Working Capital
- This represents the cash flow available to all providers of capital (both debt and equity), assuming the company is entirely financed by equity capital.
- A higher unlevered free cash flow indicates a company is generating more cash from its operations, which can be used to invest in growth opportunities, pay dividends to shareholders, or reduce debt.

Levered Free Cash Flow
- EBITDA - Taxes - Capital Expenditures - Changes in Net Working Capital - Interest Expense
- Represents the cash flow available to equity holders after taking into account interest payments on debt
- A higher levered free cash flow indicates a company is generating more cash that can be used to pay down debt, make dividend payments, or invest in growth opportunities.

Profitability ratios

Gross Profit
- Total Revenue - Cost of Goods Sold (COGS)
- This tells you how much is left after each sale

EBITDA
- Net Income + Interest Expense - Interest Income + Taxes + Depreciation + Amortization
- Earnings before interest, tax, depreciation & amortization

Net Income
- Total Revenue - Total Expenses
- Bottom line profit after subtracting all expenses from all income

Net Income Margin
- Net Income / Revenue
- Percentage of revenue that remains after deducting all expenses, including taxes and interest.

Return on Investment (ROI)
- Net Profit / Total Investment
- The return on investment capital invested in a business.

Net Operating income
- Gross Profit - Operating Expenses
- The profit derived from a company's regular business activities

Gross Profit Margin
- Gross Profit / Revenue
- Percentage of revenue that exceeds the cost of goods sold.

EBITDA Margin
- EBITDA / Revenue
- Your % of revenue that's left over as EBITDA

Net Other Income
- Other Income - Other Expense
- The net profitability of your other income / expense items, not related to your core business

Earnings per Share (EPS)
- Net Income / Total Shares Outstanding
- The portion of a company's profit allocated to each outstanding share of common stock.

Operating Profit Margin
- Operating Profit / Revenue
- The percentage of revenue that remains after deducting operating expenses.

Price-to-Earnings (P/E)
- Price per Share / EPS
- The relationship between a company's stock price and its earnings per share.

Return on Capital Employed (ROCE)
- Operating Profit / Capital Employed
- A company's ability to generate profits from the capital it has invested in its business.

SaaS Metrics

Churn
- Lost revenue / customers who are no longer active

Contraction revenue
- Reduced revenue from recurring customers

Customer Acquisition Cost
- Sales + marketing costs / new customers
- Cost to acquire a customer

Budget Attainment
- Budget variance / budget amount
- % of your budget achieved

Customer Lifetime Value
- Avg monthly revenue * avg month's active
- Total expected revenue from a customer over their lifetime

Average Revenue per User
- Total Revenue / # of Users
- The average revenue you are generating from a user (more common in B2C)

Monthly Recurring Revenue
- Opening MRR + new MRR - Churn MRR + expansion MRR - contraction MRR
- Revenue from customers subscribed to monthly recurring plans

Net Promoter Score
- Measures customer loyalty based off of likelihood to refer

Net Dollar Retention
- Net revenue / opening MRR
- The amount being retained from a customer or cohort over a specific period of time

Runway
- Cash in bank / monthly avg cash burn
- How many months till you run out of cash

Annual Recurring Revenue
- Opening ARR + new ARR - churn ARR + expansion ARR - contraction ARR
- Revenue from customers subscribed to Annual recurring plans

Revenue run rate
- This months revenue * 12
- An annualized version of your revenue

CAC Payback
- Customer Acquisition cost / gross profit
- Period of time till you make back your Customer Acquisition Cost

Cash out date
- This month + runway
- Date in which you run out of cash

Sales Attainment
- Monthly sales / monthly quota
- % of quota attained by sales team

Weighted Pipeline
- Total deals value in pipeline * percentage close likelihood
- The value of the deals in your pipeline, adjusted for the close likelihood

Average Contract Value
- Total Revenue / # of Contracts
- The average contract value from a customer (more common in B2B)

Expansion Revenue
- Added Revenue from recurring customers

Financial Statement Definitions

Profit and Loss

Show's what the business is earning vs consuming, as well as multiple levels of profitability

Revenue: Money earned from sales of goods or services.

Gross revenue: Money earned from sales of goods or services, without any adjustments

Discounts: Reduction in the price of goods or services offered to customers.

Refunds: Money paid back to customers for returned goods or cancelled services.

Net Revenue: The revenue left after subtracting all deductions

Carrier fees: Fees charged by shipping companies for delivering goods.

Cost of Goods Sold: The cost to deliver or produce your product or service

Operating Expenses: The cost to run & operate a business

Net Operating Income: The profit derived from a company's regular business activities

Other Income: Revenue generated from sources other than the main business activities.

Interest income: Revenue generated from interest earned

Other Expense: Expenses incurred from sources other than the main business activities.

Interest Expense: The cost of borrowing money, such as interest paid on loans.

Fx Gain / Loss: The profit or loss incurred from changes in foreign currency exchange rates.

Depreciation: The reduction in value of assets over time.

Amortization: The reduction in value of intangible assets over time.

Net profit before tax (NPBT): The revenue left after deducting all expenses from the gross revenue, before taxes are applied.

Provision for Taxes: The amount set aside to pay for taxes owed.

Balance Sheet

a snapshot of what the organization has in assets, liabilities, and owners equity

Assets: Economic resources owned or controlled by a business,

Cash & Cash equivalents: Cash and assets that can be easily converted into cash

Current Assets: Assets that are expected to be converted to cash within a year

Accounts Receivable: Money owed to a company for goods or services

Prepaid Expenses: Expenses paid in advance, but not yet consumed

Other Current Assets: Additional Current assets not listed already in the current assets section

Long term Assets: Assets expected to be held for more than a year

Intangible Assets: Non-physical assets

Domains: Ownership rights to internet domain names

Goodwill: The value of a business beyond its tangible assets

Software costs: The costs incurred to develop or purchase software

Organizational costs: The costs incurred to set up a business

Fixed Assets: Assets expected to be used for more than a year, such as property, plant, and equipment.

Accumulated Depreciation: The total amount of depreciation expense recognized on fixed assets.

Accumulated Amortization: The total amount of amortization expense recognized on intangible assets.

Liabilities: Capped obligations owed by a business to creditors

Accounts Payable: Money owed by a company for goods or services

Accrued interest: Interest owed but not yet paid on loans or bonds

Accrued Expenses: Expenses incurred but not yet paid

Current Liabilities: Liabilities expected to be paid within a year

Deferred Revenue: The $$ amount of goods or services owed to a customer

Other Current Liabilities: Current liabilities other than the ones listed in the Current Liabilities section of the balance sheet

Sales Tax Payable: Taxes owed on sales made by a business.

Accrued payroll: Wages earned but not yet paid to employees

Long term Liabilities: Liabilities expected to be paid after a year

Line of Credit: A revolving loan that allows a company to borrow up to a specified amount.

Convertible Note: A debt instrument that can be converted into equity.

Working Capital: The difference between current assets and current liabilities.

Owners Equity: The residual interest in a business after liabilities are deducted from assets.

Common Stock: Stock issued to shareholders that has voting rights.

Preferred Stock: tock issued to shareholders that has priority over common stock in terms of dividends or liquidation

Par Value: The face value of a share of stock

Additional Paid in Capital: The amount of capital raised from shareholders in excess of par value.

Owners draw: Money taken out of a business by its owner.

Partner contributions: Money invested by partners in a partnership.

Opening Balance Equity: The starting equity balance derived from a bank account or credit card

Retained Earnings: An accumulation of Net Income from the Profit & Loss not yet distributed to shareholders

Equity Issuance Fees: Fees paid to raise equity capital, such as underwriting fees or legal fees.

Treasury Stock: Stock that a company has bought back from shareholders.

Accumulated Other Comprehensive Income: Changes in equity that are not recognized in net income

Cash Flows

showcases the movements in cash from one period to another

Direct method: Calculates cash inflows and outflows directly from business operations

Indirect Method: Calculates cash flows from changes in balance sheet accounts, which indirectly reflect operating activities

Cash Flows From Operating Activities: Cash inflows and outflows from core business operations,

Cash inflows and outflows from core business operations,: Cash inflows and outflows related to long-term investments

Cash Flows From Financing Activities: Cash inflows and outflows related to borrowing, lending, and equity activities

Helping You Grow In **Finance & Accounting**

yourcfoguy.com

Download this in high resolution: yourcfoguy.com/encyclopedia-of-financial-metrics

Figure 15.279: 100+ **financial metrics (author: Josh Aharonoff)**

---

[49] In fact you can have an empirical choice with a poor scientific point of view of such a list by the same publisher of *The* 100 *most important financial performance indicators* that you can order on the web (see [549]). They have such small books also for others various fields like supply chain, IT, asset management, marketing, etc.

Let us now define only a few of them in this textbook.

15.4.1.2.1 Short term solvency or liquidity measure

As the name suggests, "short-term solvency ratios" as a group are intended to provide information about a firm's liquidity, and these ratios are sometimes named "liquidity measures".

The primary concern is the firm's ability to pay its bills over the short run without undue stress. Consequently, these ratios focus on current assets and current liabilities.

For obvious reasons, liquidity ratios are particularly interesting to short-term creditors. Because financial managers work constantly with banks and other short-term lenders, an understanding of these ratios is essential.

One of the best known and most widely used ratios is the "current ratio". As you might guess, the current ratio at time t is defined as follows:

$$\text{CR}(t)=\frac{\mathcal{A}(t)}{\mathcal{L}(t)} \tag{15.4.3}$$

Because current assets and liabilities are, in principle, converted to cash over the following 12 months, the current ratio is a measure of short-term liquidity.

To a creditor (particularly a short-term creditor such as a supplier) the higher the current ratio, the better. To the firm, a high current ratio indicates liquidity, but it also may indicate an inefficient use of cash and other short-term assets. Absent some extraordinary circumstances, we would expect to see a current ratio of at least 1 because a current ratio of less than 1 would mean that net working capital (current assets less current liabilities) is negative. This would be unusual in a healthy firm, at least for most types of businesses.

Inventory is often the least liquid current asset. It's also the one for which the book values are least reliable as measures of market value because the quality of the inventory isn't considered. Some of the inventory may later turn out to be damaged, obsolete, or lost.

More to the point, relatively large inventories are often a sign of short-term trouble. The firm may have overestimated sales and overbought or overproduced as a result. In this case, the firm may have a substantial portion of its liquidity tied up in slow-moving inventory.

To further evaluate liquidity, the "quick ratio", or "acid-test", is computed just like the current ratio (CR), except inventory $\mathcal{I}$ is omitted at a given time t:

$$\text{QR}(t)=\frac{\mathcal{A}-\mathcal{I}}{\mathcal{A}} \tag{15.4.4}$$

Notice that using cash to buy inventory does not affect the current ratio (CR), but it reduces the quick ratio (QR). Again, the idea is that inventory is relatively illiquid compared to cash.

We briefly mention three other measures of liquidity. A very short-term creditor might be interested in the "cash ratio" (CR):

$$\text{CR}(t)=\frac{\mathcal{A}_{\text{cash only}}(t)}{\mathcal{L}(t)} \tag{15.4.5}$$

The "interval measure" (IM) is also useful. It gives us how long the company can operate until it needs another round of financing. The average daily operating cost for companies is often named the "burn rate" meaning the rate at which cash is burned in the race to become profitable:

$$\text{IM}(t)=\frac{\mathcal{A}_{\text{cash only}}(t)}{\bar{C}_d} \tag{15.4.6}$$

where $\bar{C}_d$ is the average daily cost.

15.4.1.2.2 Long term solvency or Liquidity measure

Long-term solvency ratios are intended to address the firm's long-term ability to meet its obligations, or, more generally, its financial leverage. These are sometimes named "financial leverage ratios" or just "leverage ratios".

The "total debt ratio" takes into account all debts of all maturities to all creditors. It can be defined in several ways. One possible way is at a given time t:

$$\text{TDR}(t)=\frac{\mathcal{A}(t)-\mathcal{E}(t)}{\mathcal{A}(t)} \tag{15.4.7}$$

We can define two useful variations of the total debt ratio! The "debt–equity ratio" (DER) defined by:

$$\text{DER}(t)=\frac{\mathcal{D}(t)}{\mathcal{E}(t)} \tag{15.4.8}$$

where $\mathcal{D}$ is the total debt at time t and the "equity multiplier" (EM):

$$\text{EM}(t)=\frac{\mathcal{A}(t)}{\mathcal{E}(t)} \tag{15.4.9}$$

To complicate matters, different people (and different books) mean different things by the term debt ratio. Some mean a ratio of total debt, some mean a ratio of long-term debt only, and, unfortunately, a substantial number are simply vague about which one they mean.

Another common measure of long-term solvency is the "times interest earned ratio". Once again, there are several possible (and common) definitions. One is given by:

$$\text{TIER}(t)=\frac{\text{EBIT}(t)}{I(t)} \tag{15.4.10}$$

where EBIT is the "Earnings Before Interest And Taxes" and I the interests paid.

Obviously as the name suggests, this ratio measures how well a company has its interest obligations covered.

15.4.1.2.3 Profitability Measures

The best known and most widely used of all financial ratios. In one form or another, they are intended to measure how efficiently a firm uses its assets and manages its operations. The focus in this group is on the bottom line, net income.

Companies pay a great deal of attention to their "profit margins":

$$\text{PM}(t)=\frac{\text{Net Income}(t)}{\text{Sales}(t)} \tag{15.4.11}$$

and obviously this give the amount of profit for every currency unit in sales.

All other things being equal, a relatively high profit margin is obviously desirable. This situation corresponds to low expense ratios relative to sales.

The "return on assets" (RoA) is a measure of profit per currency unit of assets. It can be defined several ways, but the most common is this:

$$\text{RoA}(t)=\frac{\text{Net Income}(t)}{\mathcal{A}(t)} \tag{15.4.12}$$

The "return on equity" (RoE) is a measure of how the stockholders fared during the year. Because benefiting shareholders is our goal, RoE is, in an accounting sense, the true bottom-line measure of performance. RoE is usually measured as follows:

$$\text{RoE}(t)=\frac{\text{Net Income}(t)}{\mathcal{E}(t)} \tag{15.4.13}$$

The fact that RoE exceeds RoA reflects typically the use of financial leverage (debts).

As we mentioned in discussing RoA and RoE, the difference between these two profitability measures is a reflection of the use of debt financing, or financial leverage. We illustrate the relationship between these measures in this section by investigating a famous way of decomposing RoE into its component parts.

Let us recall the definition of RoE:

$$\text{RoE}(t)=\frac{\text{Net Income}(t)}{\mathcal{E}(t)} \tag{15.4.14}$$

If we were so inclined, we could multiply this ratio by Assets/Assets without changing anything:

$$\text{RoE}(t)=\frac{\text{Net Income}(t)}{\mathcal{E}(t)}\frac{\mathcal{A}(t)}{\mathcal{A}(t)}=\frac{\text{Net Income}(t)}{\mathcal{A}(t)}\frac{\mathcal{A}(t)}{\mathcal{E}(t)} \tag{15.4.15}$$

Notice that we can expressed the RoE as the product of two other ratios that are respectively the RoA and an the equity multiplier $\mathcal{E}_M$:

$$\text{RoE}(t)=\text{RoA}(t)\mathcal{E}_M(t) \tag{15.4.16}$$

We can further decompose ROE by multiplying the top and bottom by total sales:

$$\text{RoE}(t)=\frac{\text{Sales}}{\text{Sales}}\frac{\text{Net Income}(t)}{\mathcal{A}(t)}\frac{\mathcal{A}(t)}{\mathcal{E}(t)}=\frac{\mathcal{S}}{\mathcal{S}}\frac{\text{Net Income}(t)}{\mathcal{A}(t)}\frac{\mathcal{A}(t)}{\mathcal{E}(t)} \tag{15.4.17}$$

If we rearrange things a bit, RoE looks like this:

$$\text{RoE}(t)=\underbrace{\frac{\text{Net Income}(t)}{\mathcal{S}}\frac{\mathcal{S}}{\mathcal{A}(t)}}_{\text{RoA}}\frac{\mathcal{A}(t)}{\mathcal{E}(t)}=\text{PM}(t)\text{RoA}(t)\mathcal{E}_M(t) \qquad (15.4.18)$$

What we have now done is to partition RoA into its two component parts, profit margin and total asset turnover. The last expression of the preceding equation is named the "DuPont identity", after the DuPont Corporation, which popularized its use.

✍**Example:**

General Motors provides a good example of how DuPont analysis can be very useful and also illustrates why care must be taken in interpreting RoE values. In 11989 (holocene calendar), GM had an RoE of 12.1%. By 1993 (holocene calendar), its RoE had improved to 44.1%, a dramatic improvement. On closer inspection, however, we find that over the same period GM's profit margin had declined from 3.4 to 1.8%, and RoA had declined from 2.4 to 1.3% percent. The decline in RoA was moderated only slightly by an increase in total asset turnover from 0.71 to 0.73 over the period.

Given this information, how is it possible for GM's RoE to have climbed so sharply? From our understanding of the DuPont identity, it must be the case that GM's equity multiplier increased substantially. In fact, what happened was that GM's book equity value was almost wiped out overnight in 11992 (holocene calendar) by changes in the accounting treatment of pension liabilities. If a company's equity value declines sharply, its equity multiplier rises. In GM's case, the multiplier went from 4.95 in 11989 (holocene calendar) to 33.62 in 11993 (holocene calendar). In sum, the dramatic "improvement" in GM's RoE was almost entirely due to an accounting change that affected the equity multiplier and didn't really represent an improvement in financial performance at all.

15.4.1.2.4 Growth Rate

The first growth rate of interest is the maximum growth rate that can be achieved with no external financing of any kind. We will named this the "internal growth rate" because this is the rate the firm can maintain with internal financing only.

Given that there is therefore either no Equities or Debts only the RoA will be useful to create this rate. Considering that a part of the RoA will be use for internal expenses, or taxes, we multiply the RoA (remember that it is a ratio, therefore without units) by a factor p named the "plowback" factor to get the net PoA!

Afterwards the most naive high-school level model make the assumption that the growing rate is a linear function of time (forge any ARMA/ARIMA, GARCH forecasting model in Corporate Finance trainings even at a graduate level!) such that:

$$\text{IGR}(t)=\text{RoA}(t)(1+\text{IGR}(t-1)) \qquad (15.4.19)$$

and therefore after making the assumption of small variations we get:

$$\text{IGR}(t)\cong\text{RoA}(t)(1+\text{IGR}(t)) \qquad (15.4.20)$$

Hence:

$$\text{IGR}(t) \cong \frac{\text{RoA}(t)}{1-\text{RoA}(t)} \tag{15.4.21}$$

But it is quite naive and very academic relation. In reality no serious practitioners should use such a method! But instead stochastic related forecasting techniques!!

15.4.1.2.5 Asset management or turnovers measures

The measures in this section are sometimes named "asset utilization ratios". The specific ratios we discuss can all be interpreted as measures of turnover. What they are intended to describe is how efficiently or intensively a firm uses its assets to generate sales.

The "inventory turnover ratio" is defined as:

$$\text{ITR}(\Delta t) = \text{ITR}(t_f - t_0) = \frac{\text{Cost of goods sold}(\Delta t)}{\mathcal{I}(t_f)} \tag{15.4.22}$$

the higher this ratio is, the more efficiently we are managing inventory.

Our inventory measures give some indication of how fast we can sell product. We now look at how fast we collect on those sales. The "receivables turnover ratio" is defined much like the inventory turnover ratio:

$$\text{RTR}(\Delta t) = \frac{\text{Sales}(\Delta t)}{\text{Accounts receivable}(t_f)} \tag{15.4.23}$$

15.4.1.2.6 Market Value Measures

Our final group of measures is based, in part, on information not necessarily contained in financial statements: the market price per share of stock. Obviously, these measures can be calculated directly only for publicly traded companies.

We begin with what we have already defined in the section Economy. The earning per share:

$$\text{EPS}(t) = \frac{\text{Net Income}(t)}{\text{Shares Issued}(t)} \tag{15.4.24}$$

We also have the "price earning ratio" defined by:

$$\text{PER}(t) = \frac{\text{Price per Share}(t)}{\text{EpS}(t)} = \frac{\text{PpS}(t)}{\text{EpS}(t)} \tag{15.4.25}$$

the ratio measures how much investors are willing to pay per dollar of current earnings, higher PERs are often taken to mean the firm has significant prospects for future growth. Of course, if a firm had no or almost no earnings, its PE would probably be quite large. So, as always, care is needed in interpreting this ratio.

Another commonly quoted market value measure is the "market to book ratio":

$$\text{MtBR}(t) = \frac{\text{Market Value per Share}(t)}{\text{Book Value per Share}(t)} = \frac{\text{MVpS}(t)}{\text{BVpS}(t)} \tag{15.4.26}$$

The book value per share is total equity (not just common stock) divided by the number of issued shares!!

In a loose sense, the market-to-book ratio therefore compares the market value of the firm's investments to their cost. A value less than 1 could mean that the firm has not been successful overall in creating value for its stockholders.

This completes our definitions of some common one-to-one ratios. We could tell you about more of them, but these are enough for now and its already quite boring enough in absence of ISO norms. So now let us deal with multiple-to-multiple ratios performance analysis:

15.4.1.2.7 Data Envelopment Analysis

"Data envelopment analysis" (DEA) is a operational research methodology (see section Numerical methods page 6662) to evaluate the relative efficiency score η[50] of multiple decision-making units (DMUs or "comparable production units") when the production process presents a structure of multiple inputs x_i and outputs y_i. It is a quite powerful service management and benchmarking technique originally developed by Chames, Cooper and Rhodes (11978 according to holocene calendar) to evaluate non-profit and public sector organizations[51]. DEA has since been proven to locate ways to improve service not visible with other techniques. Yet there is an anomaly surrounding this developing methodology. One of the largest US banks located over $\$100$ million of excess annual personnel and operating costs, enough to affect their earnings per share and these savings were not identifiable with other techniques in use. While other banks have also realized improved profits through initiatives driven by DEA, we could not locate more than 10 banks in this category. While businesses have no obligation to report their internal methods, DEA has not been widely adopted by banks. Why is DEA a method that can generates new paths to improve profits not used when other less powerful techniques continue to be used? We believe that greater adoption of DEA will only be possible when it is more accessible to non-technical skills (as most managers and CXO are soft-skills).

Remark

In the literature, the two most widely used variants of the DEA method are: the CCR model (Charnes, Cooper and Rhodes, 11978 according to holocene calendar) which assumes constant scale returns and the BCC model (Banker, Charnes et al. Cooper, 11984 according to holocene calendar) which assumes variable scale yields. In the case of constant scale returns, it is assumed that an increase in the amount of inputs consumed will lead to a proportional increase in the quantity of outputs produced. On the other hand, in the case of returns of variable scales (increasing or decreasing), the quantity of outputs produced is considered to increase more or less proportionally than the increase in inputs.

---

[50] When we focus on service organizations we generally cannot determine what the engineered, optimum or absolute efficient output-to-input ratio is. Consequently we cannot determine whether a service unit is absolutely efficient. We can, however, compare several service unit output-to-input ratios and determine that one unit is more or less efficient than another - benchmarking. The difference in efficiency will be due to the technology or production process used, how well that process is managed, and/or the scale or size of the unit.

[51] It has been developed in fact to analyse the efficiency of the US resources allocation strategy to the schools and have afterwards be applied to all other government services. The India government also use it for the same purposes.

Managers have not widely adopted DEA to improve organization performance, in part, because most DEA publications are in academic journals or books requiring the ability to understand operational research and supporting mathematical notation. In fact, some managers trying to use DEA based on their understanding of academic publications have misunderstood the way to apply DEA. They erroneously attribute weak results to the technique when the problem is often due to the misapplication of DEA. Another problem (maybe it is the biggest), DEA will be more used when CXO will stop to fear the fact to listen that their organization is non-efficient.

We will explain here what DEA does, how DEA evaluates efficiency, how DEA identifies paths to improve efficiency, limitations of DEA, and how to use DEA. This will enable managers or governments to explore and assess the value of their organization using DEA in their service operations.

What does DEA do?

1. DEA compares service units considering all resources used and services provided, and identifies the most efficient units or best practice units (branches, departments, individuals) and the inefficient units in which real efficiency improvements are possible. This is achieved by comparing the mix and volume of services provided and the resources used by each unit compared with those of all the other units. In short, DEA is a very powerful benchmarking technique.

2. DEA calculates the amount and type of cost and resource savings that can be achieved by making each inefficient unit as efficient as the most efficient - best practice - units.

3. Specific changes in the inefficient service units are identified, which management can implement to achieve potential savings located with DEA. These changes would make the efficient units performance approach the best practice unit performance. In addition, DEA estimates the amount of additional service an inefficient unit can provide without the need to use additional resources.

4. Management receives information about performance of service units that can be used to help transfer system and managerial expertise from better-managed, relatively efficient units to the inefficient ones. This has resulted in improving the productivity of the inefficient units, reducing operating costs and increasing profitability.

The above four types of DEA information are proven to be extremely valuable because they identify relationships not identifiable with alternative techniques that are commonly used in service organizations. As a result, improvements to operations extend beyond any performance improvements management may have achieved using other techniques.

It is considered that the non-efficiency of an organization is mainly due to two main factors:

- The organization may be non-efficient because its management may perform better (not to say deficient...).
- The organization may be non-efficient because it hasn't reached it's optimal size

Remark

Some managers pretend that their organization that performs badly according to a DEA analysis because they cannot be compared with other organization because of special other factors. This may be true but then... the "*burden of proof*" lies on them (as we have proved it in the section Statistics page 944).

Obviously the efficiency score must be considered as an order of amplitude, a result of binary judgement that helps manager to know if there is something to change or not. We could obviously used combinatorics methods on the inputs and outputs to get an interval rather than a punctual value.

Let us introduce now the corresponding mathematical framework! The operational research technique is used to find the set of coefficients (u's and v's) that will give the highest possible efficiency ratio of outputs to inputs for the service unit being evaluated.

In the model:

- j is the number of service units (SU) being compared in the DEA analysis
- SU$_j$ is the service unit number j
- η_j is the efficiency of the service unit j being evaluated by DEA
- y_{rj} is the amount of output r used by service unit j
- x_{ij} is the amount of input i used by the service unit j
- i number of inputs used by the SUs
- r number of outputs generated by the SUs
- u_r coefficient or weight assigned by DEA to output r
- v_i coefficient or weight assigned by DEA to input i

The data required to apply DEA are the actual observed outputs produced y_{rj} and the actual inputs used x_{ij}, during one time period for each service unit in the set of units being evaluated. Hence, x_{ij} is the observed amount of the ith input used by the jth service unit, and y_{rj} is the amount of rth output produced by the jth service unit.

If the value of θ for the service unit being evaluated is less than 100%, then that unit is inefficient, and there is the potential for that unit to produce the same level of outputs with fewer inputs.

The objective function is given by:

$$\max \eta_j=\dfrac{u_1y_{1j}+u_2y_{2j}+\ldots+u_ny_{nj}}{v_1x_{1j}+v_2x_{2j}+\ldots+v_mx_{mj}}=\dfrac{\displaystyle\sum_{r=1}^{n}u_ry_{rj}}{\displaystyle\sum_{i=1}^{m}v_ix_{ij}} \qquad (15.4.27)$$

This is subject to the constraint that when the same set of u and v coefficients is applied to all other service units being compared, no service unit (SU) will be more than 100% efficient as follows:

$$\forall k \neq j \quad \text{SU}_k : \eta_k = \frac{u_1 y_{1k} + u_2 y_{2k} + \ldots + u_m y_{mk}}{v_1 x_{1k} + v_2 x_{2k} + \ldots + v_n x_{nk}} = \frac{\displaystyle\sum_{r=1}^{n} u_r y_{rk}}{\displaystyle\sum_{i=1}^{m} v_i x_{ik}} \leq 1 \tag{15.4.28}$$

Finally all this compacted into the classical form of optimization problem will be written:

$$\boxed{\begin{array}{lll} \underset{\vec{u},\vec{v}}{\text{maximize}} & & \eta_j(\vec{u},\vec{v}) = \dfrac{\displaystyle\sum_{r=1}^{n} u_r y_{rj}}{\displaystyle\sum_{i=1}^{m} v_i x_{ij}} \\ \text{subject to} & \forall k \neq j & \text{SU}_k : \eta_k = \dfrac{\displaystyle\sum_{r=1}^{n} u_r y_{rk}}{\displaystyle\sum_{i=1}^{m} v_i x_{ik}} \leq 1 \\ & & u_r, v_i > 0 \qquad r, i = 1 \ldots n, m \end{array}} \tag{15.4.29}$$

Remark

It is recommended that the number of inputs and outputs attributes to not be bigger that then number of organizations divided by 3.

Note that optimal multipliers calculated by DEA are objectively determined weights that may not correspond to relative values that an analyst would assign to outputs and inputs and this is a good point as therefore there a are not subjective (or biased) human choice to do! Furthermore the weights can be different for each SU and this is more realistic! This is actually a strength and is not a weakness of DEA. A unit located as inefficient using DEA is so identified only after all possible weights have been considered to give that branch the highest rating possible consistent with the constraint that no branch in the data set can be more than 100% efficient. Hence, any other set of weights applied to all branches would only make an inefficient branch appear equally or less efficient; that is, DEA gives the benefit of the doubt to each branch or service unit in calculating the efficiency value. In addition DEA will not erroneously locate an efficient unit as inefficient. Hence, DEA gives the "benefit of the doubt" to each unit being evaluated trying to make it look as efficient as possible in comparison with the other units. This bias makes this a tool that managers can use with confidence.

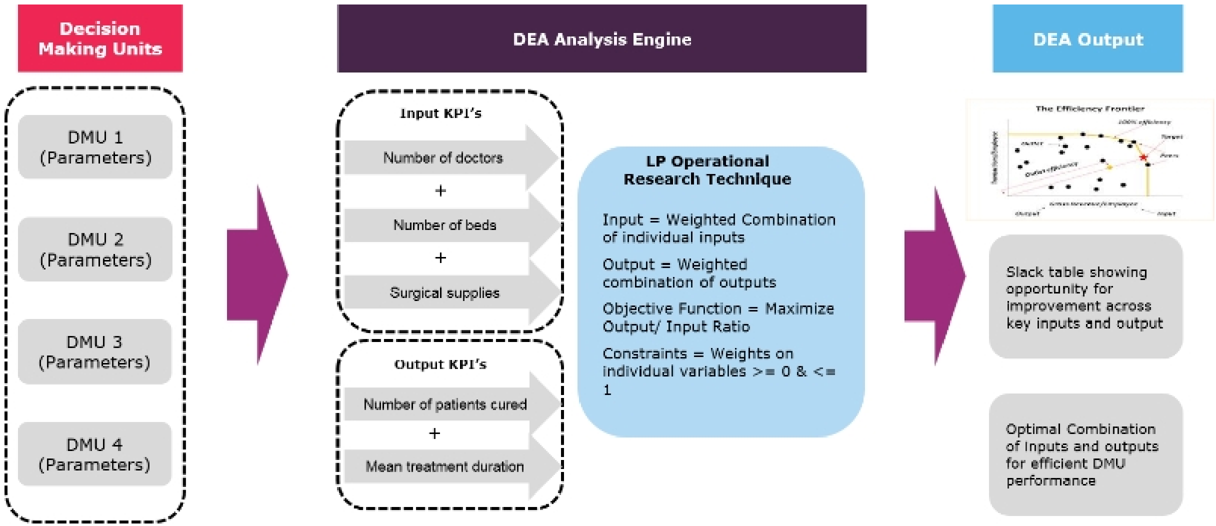

Decision Making Units
DMU 1 (Parameters)
DMU 2 (Parameters)
DMU 3 (Parameters)
DMU 4 (Parameters)
DEA Analysis Engine
Input KPI's
Number of doctors
+
Number of beds
+
Surgical supplies
Output KPI's
Number of patients cured
+
Mean treatment duration
LP Operational Research Technique
Input = Weighted Combination of individual inputs
Output = Weighted combination of outputs
Objective Function = Maximize Output/ Input Ratio
Constraints = Weights on individual variables >= 0 & <= 1
DEA Output
The Efficiency Frontier
Slack table showing opportunity for improvement across key inputs and output
Optimal Combination of inputs and outputs for efficient DMU performance

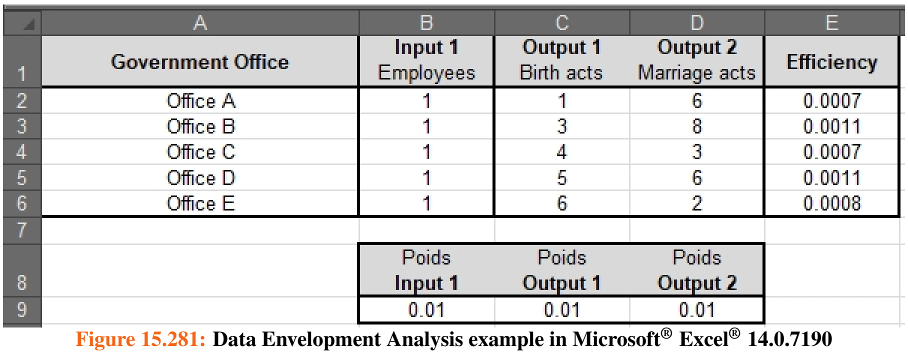
Figure 15.280: **Data Envelopment Analysis**

To appreciate the power and limitations of DEA in improving efficiency, let us see an example with Microsoft® Excel® (the reader can also found the corresponding example in the R companion book):

✍**Example:**
Consider we have five administrations (office) that delivers birth and marriage acts in various quantities during the same period of time and that each office has one employee only. We want to run a DEA analysis on the following situation summarized in the table below:

Government Office	Input 1 Employees	Output 1 Birth acts	Output 2 Marriage acts
Office A	1	1	6
Office B	1	3	8
Office C	1	4	3
Office D	1	5	6
Office E	1	6	2

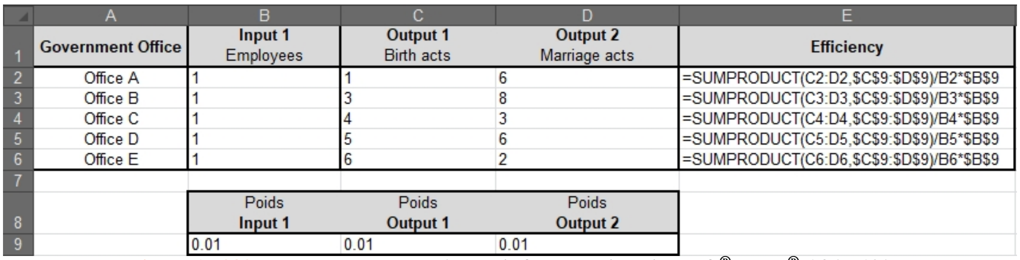
For this purpose in Microsoft® Excel® 14.0.7190 we build the following structure:



	A	B	C	D	E
1	Government Office	**Input 1** Employees	**Output 1** Birth acts	**Output 2** Marriage acts	**Efficiency**
2	Office A	1	1	6	0.0007
3	Office B	1	3	8	0.0011
4	Office C	1	4	3	0.0007
5	Office D	1	5	6	0.0011
6	Office E	1	6	2	0.0008
7					
8		Poids **Input 1**	Poids **Output 1**	Poids **Output 2**	
9		0.01	0.01	0.01	

Figure 15.281: **Data Envelopment Analysis example in Microsoft® Excel® 14.0.7190**

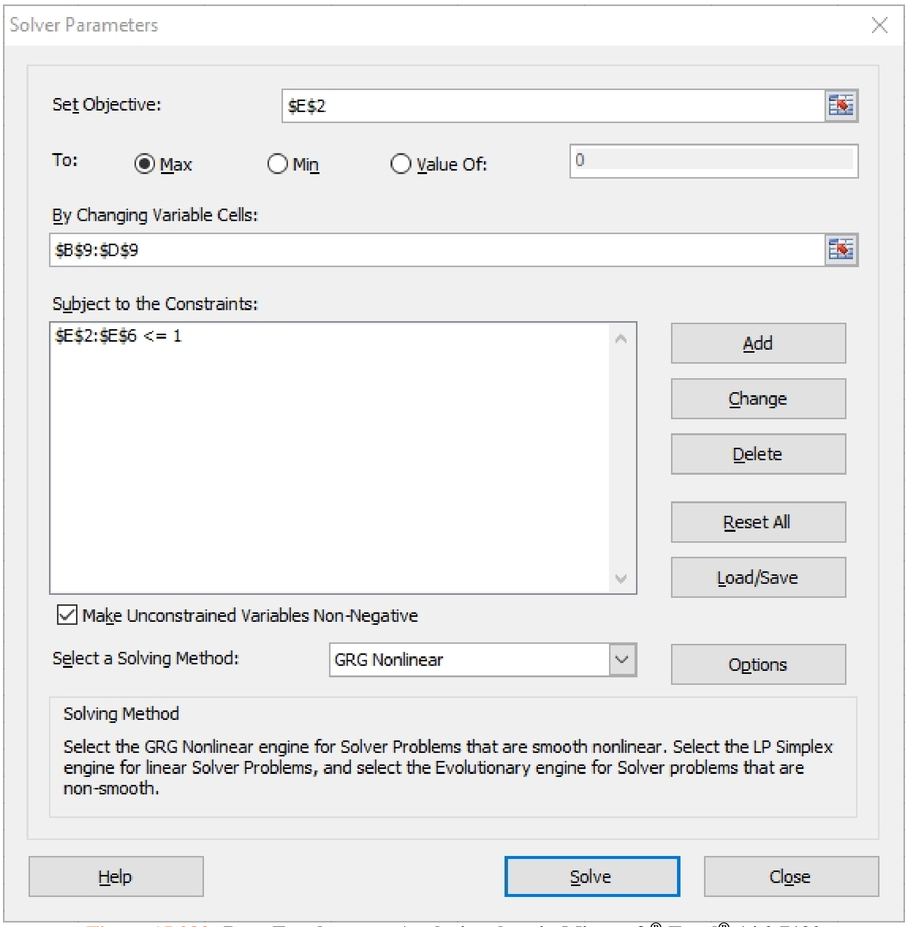
Or explicitly:



	A	B	C	D	E
1	**Government Office**	**Input 1** Employees	**Output 1** Birth acts	**Output 2** Marriage acts	**Efficiency**
2	Office A	1	1	6	=SUMPRODUCT(C2:D2,$C$9:$D$9)/B2*$B$9
3	Office B	1	3	8	=SUMPRODUCT(C3:D3,$C$9:$D$9)/B3*$B$9
4	Office C	1	4	3	=SUMPRODUCT(C4:D4,$C$9:$D$9)/B4*$B$9
5	Office D	1	5	6	=SUMPRODUCT(C5:D5,$C$9:$D$9)/B5*$B$9
6	Office E	1	6	2	=SUMPRODUCT(C6:D6,$C$9:$D$9)/B6*$B$9
7					
8		Poids **Input 1**	Poids **Output 1**	Poids **Output 2**	
9		0.01	0.01	0.01	

Figure 15.282: Data Envelopment Analysis formulas in Microsoft® Excel® 14.0.7190

Once this done we run a first time the solver with the following settings:



Solver Parameters

Set Objective: $E$2

To: ● Max ○ Min ○ Value Of: 0

By Changing Variable Cells: $B$9:$D$9

Subject to the Constraints:

$E$2:$E$6 <= 1

☑ Make Unconstrained Variables Non-Negative

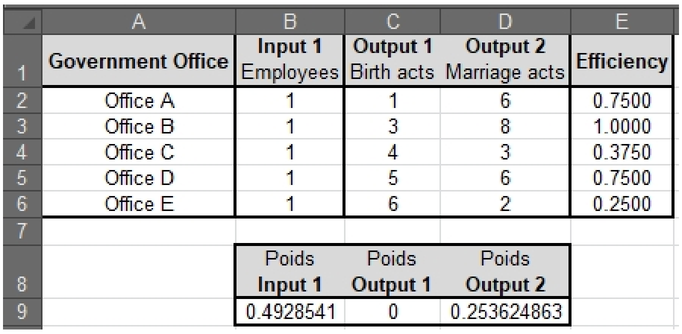
Select a Solving Method: GRG Nonlinear

Solving Method

Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

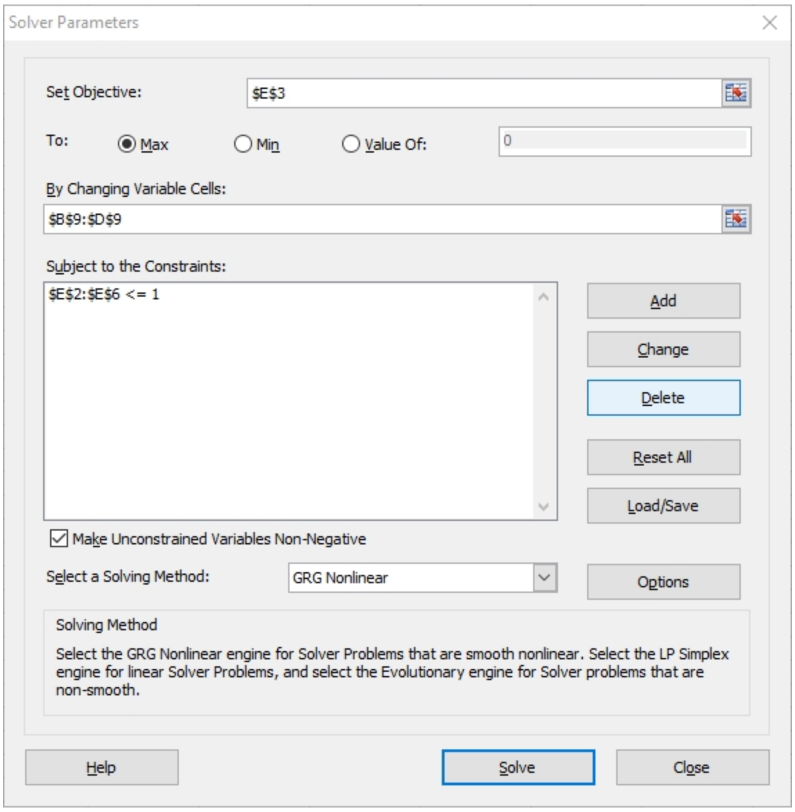
Figure 15.283: Data Envelopment Analysis solver in Microsoft® Excel® 14.0.7190

Then we get for Office A:



	A	B	C	D	E
1	**Government Office**	**Input 1** Employees	**Output 1** Birth acts	**Output 2** Marriage acts	**Efficiency**
2	Office A	1	1	6	0.7500
3	Office B	1	3	8	1.0000
4	Office C	1	4	3	0.3750
5	Office D	1	5	6	0.7500
6	Office E	1	6	2	0.2500
7					
8		Poids **Input 1**	Poids **Output 1**	Poids **Output 2**	
9		0.4928541	0	0.253624863	

Doing the same for Office B but with the following settings:



Set Objective: $E$3

To: ● Max ○ Min ○ Value Of: 0

By Changing Variable Cells: $B$9:$D$9

Subject to the Constraints:

$E$2:$E$6 <= 1

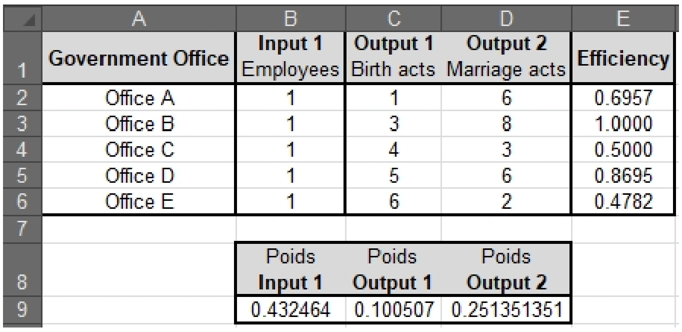
☑ Make Unconstrained Variables Non-Negative

Select a Solving Method: GRG Nonlinear

Solving Method

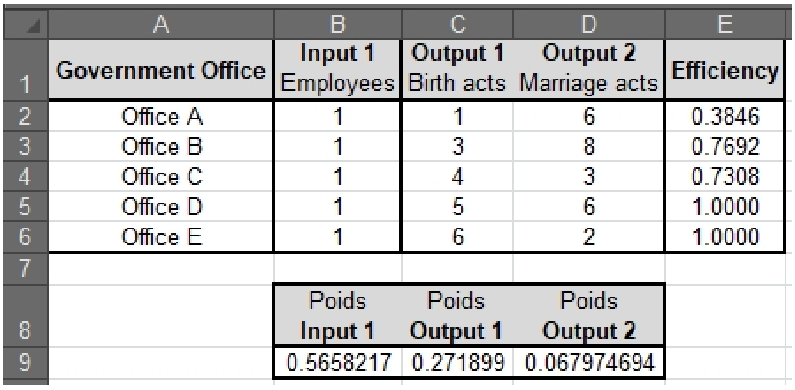
Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

we get:



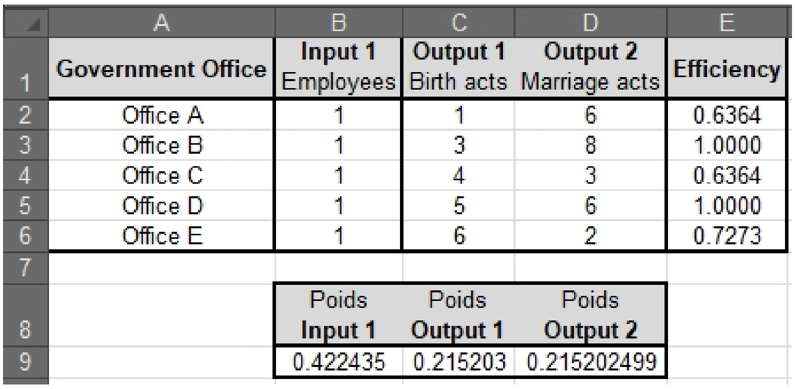
	A	B	C	D	E
1	**Government Office**	**Input 1** Employees	**Output 1** Birth acts	**Output 2** Marriage acts	**Efficiency**
2	Office A	1	1	6	0.6957
3	Office B	1	3	8	1.0000
4	Office C	1	4	3	0.5000
5	Office D	1	5	6	0.8695
6	Office E	1	6	2	0.4782
7					
8		Poids **Input 1**	Poids **Output 1**	Poids **Output 2**	
9		0.432464	0.100507	0.251351351	

And proceeding same for Office C, we get:



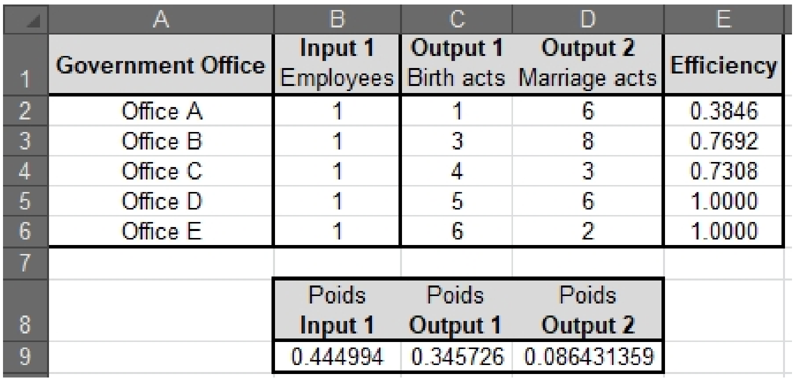
	A	B	C	D	E
1	**Government Office**	**Input 1** Employees	**Output 1** Birth acts	**Output 2** Marriage acts	**Efficiency**
2	Office A	1	1	6	0.3846
3	Office B	1	3	8	0.7692
4	Office C	1	4	3	0.7308
5	Office D	1	5	6	1.0000
6	Office E	1	6	2	1.0000
7					
8		Poids **Input 1**	Poids **Output 1**	Poids **Output 2**	
9		0.5658217	0.271899	0.067974694	

and for Office D:



	A	B	C	D	E
1	**Government Office**	**Input 1** Employees	**Output 1** Birth acts	**Output 2** Marriage acts	**Efficiency**
2	Office A	1	1	6	0.6364
3	Office B	1	3	8	1.0000
4	Office C	1	4	3	0.6364
5	Office D	1	5	6	1.0000
6	Office E	1	6	2	0.7273
7					
8		Poids **Input 1**	Poids **Output 1**	Poids **Output 2**	
9		0.422435	0.215203	0.215202499	

and finally for Office E:



	A	B	C	D	E
1	**Government Office**	**Input 1** Employees	**Output 1** Birth acts	**Output 2** Marriage acts	**Efficiency**
2	Office A	1	1	6	0.3846
3	Office B	1	3	8	0.7692
4	Office C	1	4	3	0.7308
5	Office D	1	5	6	1.0000
6	Office E	1	6	2	1.0000
7					
8		Poids **Input 1**	Poids **Output 1**	Poids **Output 2**	
9		0.444994	0.345726	0.086431359	

We see that all five efficiencies $\{0.7500, 1.0000, 0.7308, 1.0000, \}$ matches perfectly with the results we get with `R` in the corresponding companion book.

First, as we know, efficiency can be simply defined as the ratio of output to input. More output per unit of input reflects relatively greater efficiency. If the greatest possible output per unit of input is achieved, a state of absolute or optimum efficiency has been achieved and it is not possible to become more efficient without new technology or other changes in the production process.

Some of the advantages of DEA are:

- no need to explicitly specify a mathematical form for the production function
- proven to be useful in uncovering relationships that remain hidden for other methodologies
- capable of handling multiple inputs and outputs
- capable of being used with any input-output measurement
- the sources of inefficiency can be analysed and quantified for every evaluated unit

Some of the disadvantages of DEA are:

- results are sensitive to the selection of inputs and outputs
- you cannot test for the best specification
- the number of efficient firms tends to increase with the number of inputs and output variables

DEA is commonly applied in the electric utilities sector. For instance a government authority can choose Data Envelopment Analysis as their measuring tool to design an individualized regulatory rate for each firm based on their comparative efficiency. The input components would include man-hours, losses, capital (lines and transformers only), and goods and services. The output variables would include number of customers, energy delivered, length of lines, and degree of coastal exposure (Berg 12010). DEA is also regularly used to assess the efficiency of public and not-for-profit organizations, e.g. hospitals (Kuntz, Scholtes & Vera 12007; Kuntz & Vera 12007; Vera & Kuntz 12007) or police forces (Thanassoulis 11995; Sun 12002; Aristovnik et al. 12013, 12014).

DEA has been recognized as a valuable analytical research instrument and a practical decision support tool. DEA has been credited for not requiring a complete specification for the functional form of the production frontier nor the distribution of inefficient deviations from the frontier. Rather, DEA requires general production and distribution assumptions only. However, if those assumptions are too weak, inefficiency levels may be systematically underestimated in small samples. In addition, erroneous assumptions may cause inconsistency with a bias over the frontier. Therefore, the ability to alter, test and select production assumptions is essential in conducting DEA-based research. However, the DEA models currently available offer a limited variety of alternative production assumptions only.

We have described the basic DEA model, the insights it provides in a simple example where similar insight could be gained from observation and where the operational research programming calculations could easily be done with a simple spreadsheet software.

The reader should know however that in the specialized literature about these subjects he will be able to find a lot of other DEA models. Respectively (with their corresponding abbreviations):

- FDH: Free disposability hull, no convexity assumption
- VRS: Variable returns to scale, convexity and free disposability
- DRS: Decreasing returns to scale, convexity, down-scaling and free disposability
- CRS: Constant returns to scale, convexity and free disposability
- IRS: Increasing returns to scale, (up-scaling, but not down-scaling), convexity and free disposability
- IRS2: Increasing returns to scale (up-scaling, but not down-scaling), additivity, and free disposability
- ADD: Additivity (scaling up and down, but only with integers), and free disposability; also known as replicability and free disposability, the free disposability and replicability hull (frh) – no convexity assumption

- FDH+: A combination of free disposability and restricted or local constant return to scale

The model that we have seen above is named the "CRS model" for "Constant Returns to Scale".

Each model exists also for information in two variations (or "orientations") that so far I'm not very a fan of (so I will not introduce them in this textbook):

- "Input oriented": the DEA model minimize the inputs for a given level of outputs. In other words, it indicates of how much an organization can increase its inputs by still producing the same level of outputs. Then the performance margin is obviously on the input (i.e. it must decrease[52]).
- "Output oriented": the DEA model maximize the outputs for a given level of inputs. In other words, it indicates of how much an organization can increase its outputs by still having the same level of inputs. Then the performance margin is obviously on the output (i.e. it must increase).

The orientation of the model must be chosen according to the variables (inputs or outputs) over which decision-makers exercise the greatest management power. For example, a public school principal probably has more management power over the teaching staff (input) than the number of students received or the student results (outputs). In this case, an input orientation is more appropriate.

In the public sector, but sometimes also in the private sector, a certain level of resources is allocated and guaranteed to organizations. In such a case, the decision makers seek to maximize the benefits provided, and therefore choose an output orientation. Otherwise, if the objective of the decision-makers is to produce a certain level of outputs (for example an imposed quota), the latter seek to minimize the consumption of resources. They therefore opt for an input orientation.

If no constraints are imposed on the decision-makers and if they exercise a management power over both the inputs and the outputs, the orientation of the model depends on the objectives set for the organizations. Is the goal to reduce costs (input orientation) or to maximize production (orientation output)?

---

[52] This is typically the case in organizations that has become too big and must apply a cost compression optimization plan

15.4.1.3 Weighted Average Cost of Capital (WACC)

When a firm issues debt, equity, or other securities to fund a new investment, there are many potential consequences of its decision. By far the most important question for a CFO is whether different choices will affect the return of the firm and thus the amount of capital it can raise. Companies can use also the WACC to see if the investment projects available to them are worthwhile to undertake.

To answer this question we will assume a "perfect capital market" world in which:

1. Securities are fairly priced: Investors and firms can trade the same set of securities at competitive market prices equal to the present value of their future cash flows.
2. There are no tax consequences or transactions cost: There are no tax consequences, transaction cost, or other issuance costs associated with financing decisions or security trading.
3. Investment cash flows are independent of financing choices. A firm's financing decisions do not change the cash flows generated by its investments, nor do they reveal new information about those cash flows.

The are many ways to introduce the WACC that are more or less complex. In fact you know that in this textbook we like the easiest one.

Remark

The absence of debts means the absence of financial leverage (in fact there is a leverage but not so fast and big as with debts). Equity in a firm with no debt is therefore named "unlevered equity". Equity in a firm that also has debt is named levered equity".

Consider a firm that has an amount of dept $\mathcal{D}$ (share of stocks) with a return $r_{\mathcal{D}}=(1+t_{\mathcal{D}}\%)$ and an amount of equities $\mathcal{E}$ (bonds) with a return $r_{\mathcal{E}}=(1+t_{\mathcal{E}}\%)$

By application of simple interest rate composition we write:

$$\mathcal{D}r_{\mathcal{D}}+\mathcal{E}r_{\mathcal{E}}=\bar{r}(\mathcal{D}+\mathcal{E}) \tag{15.4.30}$$

and therefore:

$$\boxed{\bar{r}:=r_{\text{WACC}}=\frac{\mathcal{D}}{\mathcal{D}+\mathcal{E}}r_{\mathcal{D}}+\frac{\mathcal{E}}{\mathcal{D}+\mathcal{E}}r_{\mathcal{E}}} \tag{15.4.31}$$

As obviously:

$$\frac{\mathcal{D}}{\mathcal{D}+\mathcal{E}}+\frac{\mathcal{E}}{\mathcal{D}+\mathcal{E}}=1 \tag{15.4.32}$$

The prior previous relation can be written with weights notation:

$$r_{\text{WACC}}=\omega_{\mathcal{D}}r_{\mathcal{D}}+\omega_{\mathcal{E}}r_{\mathcal{E}} \tag{15.4.33}$$

Another complication is the effect of taxes T, and the fact that interest payments (dividends) paid to bondholders are tax-deductible by the firm. Therefore we have to write:

$$\bar{r} := r_{\text{WACC}} = \frac{\mathcal{D}}{\mathcal{D}+\mathcal{E}} r_{\mathcal{D}} + (1-T)\frac{\mathcal{E}}{\mathcal{D}+\mathcal{E}} r_{\mathcal{E}} \tag{15.4.34}$$

where $r_{\mathcal{E}}$ is the required rate of return on equity, estimated from the CAPM (see section Economy page 7740) and the history of the firm's share price (hence the fact the WACC use the same assumptions) and r_D is the rate of return (yield) on the company's bonds, (also named for recall the "borrowing rate").

This r is the rate of discount applied to risky future cash flows, as in chapter 4. The assumption is that the new project will be financed with the same mixture of debt and equity as already exists in the firm, and this assumption is correct if the money comes from retained earnings. But sometimes, the firm will issue new equity to finance an expansion or project; and sometimes, the project will be risk less enough that it can be financed completely with new debt. All of these complications, and the added problem of accounting for taxes, are best left for another course.

Note that the cost of debt is not adjusted for taxes, because we are assuming perfect capital market and thus are ignoring taxes. When we compute the weighted average cost of capital without taxes we refer to it as the firm's pretax WACC.

✍**Example:**

Imagine a company with the following financial structure:

- Market Value of Equity ($\mathcal{E}$): 500 millions USD
- Market Value of Dept ($\mathcal{D}$): 300 millions USD
- Cost of Equity ($r_{\mathcal{E}}$): 8%
- Cost of Dept ($r_{\mathcal{D}}$): 5%
- Corporate tax rate (T): 25%

Therefore:

$$\begin{aligned} r_{\text{WACC}} &= \frac{\mathcal{D}}{\mathcal{D}+\mathcal{E}} r_{\mathcal{D}} + (1-T)\frac{\mathcal{E}}{\mathcal{D}+\mathcal{E}} r_{\mathcal{E}} = \frac{300}{300+500}\cdot 5\% + (1-25\%)\frac{500}{300+500}\cdot 8\% \\ &= 0.375\cdot 5\% + 75\% \cdot 0.625 \cdot 8\% = 1.875\% \end{aligned} \tag{15.4.35}$$

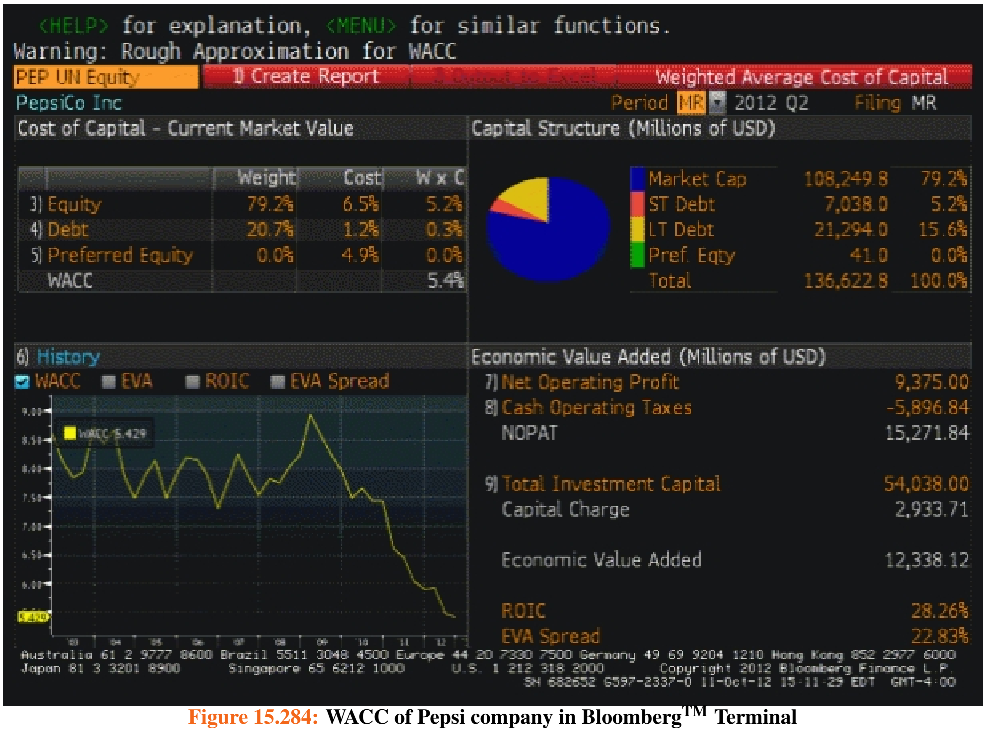
In general, the WACC can be calculated with the following relation:

$$\boxed{r_{\text{WACC}} = \frac{\displaystyle\sum_{i=1}^{N} r_i \cdot \text{MV}_i}{\displaystyle\sum_{i=1}^{N} \text{MV}_i}} \tag{15.4.36}$$

where N is the number of sources of capital (securities, types of liabilities), r_i is the required rate of return for security i and MV_i is the market value of all outstanding securities i.

And as seen previously, in the case where the company is financed with only equity and debt, the average cost of capital is computed as follows:

$$r_{\text{WACC}} = \frac{\mathcal{D}}{\mathcal{D}+\mathcal{E}} r_{\mathcal{D}} + \frac{\mathcal{E}}{\mathcal{D}+\mathcal{E}} r_{\mathcal{E}} \tag{15.4.37}$$



<HELP> for explanation, <MENU> for similar functions.
Warning: Rough Approximation for WACC

PepsiCo Inc
Period MR 2012 Q2 Filing MR

Cost of Capital - Current Market Value

	Weight	Cost	W x C
3) Equity	79.2%	6.5%	5.2%
4) Debt	20.7%	1.2%	0.3%
5) Preferred Equity	0.0%	4.9%	0.0%
WACC			5.4%

Capital Structure (Millions of USD)

Market Cap	108,249.8	79.2%
ST Debt	7,038.0	5.2%
LT Debt	21,294.0	15.6%
Pref. Eqty	41.0	0.0%
Total	136,622.8	100.0%

6) History

Economic Value Added (Millions of USD)

7) Net Operating Profit	9,375.00
8) Cash Operating Taxes	-5,896.84
NOPAT	15,271.84
9) Total Investment Capital	54,038.00
Capital Charge	2,933.71
Economic Value Added	12,338.12
ROIC	28.26%
EVA Spread	22.83%

Australia 61 2 9777 8600 Brazil 5511 3048 4500 Europe 44 20 7330 7500 Germany 49 69 9204 1210 Hong Kong 852 2977 6000
Japan 81 3 3201 8900 Singapore 65 6212 1000 U.S. 1 212 318 2000 Copyright 2012 Bloomberg Finance L.P.
SN 682652 G597-2337-0 11-Oct-12 15:11:29 EDT GMT-4:00

Figure 15.284: WACC of Pepsi company in Bloomberg$^{\text{TM}}$ Terminal

15.4.1.4 Break-even Point Analysis (BEPA)

Break-even point analysis, also named "neutral point analysis", is the set of calculations to determine the minimum activity level at which an organization becomes profitable for itself by its economies of scale, that is to say, it stop to lose money on this activity. Obviously, this technique also allows you to compare several strategies together.

The "neutral point" or "break-even point" is intuitively enough the intersection point between the curve of Turnover (because turnover is normally a function of several variables) and curve of Spending (also a function of several variables) necessary to produce that revenue. In other words, this is the corresponding coordinates where the two functions are equal (in practice it is rarely represented graphically because they are too many variables).

Formally... given a vector $\vec{x}\in\mathbb{R}^n$ having for components all economic variable describing the problem. We consider a function of total costs $T_c(\vec{x}):\mathbb{R}^n\mapsto\mathbb{R}$ and another function or revenues $T_r(\vec{x}):\mathbb{R}^n\mapsto\mathbb{R}$. The break-even analysis consists obviously to find the vector $\vec{x}$ such that:

$$T_c(\vec{x})=T_r(\vec{x}) \tag{15.4.38}$$

or equivalently:

$$T_c(\vec{x})-T_r(\vec{x})=0 \tag{15.4.39}$$

and the solutions can be non-unique.

Normally this technique should allow a manager or government to determine information such as:

- Predict sales volume to cover just the costs (Neutral Point)
- What should be the price per unit to cover just the costs (Neutral Point)
- What is the level of fixed costs that balances just gains (Neutral Point)
- How prices affect the turnover volume corresponding to equilibrium (Neutral Point)
- What is the tax discount that should be given to a company that relocates depending of the number of new jobs it will create

The neutral point analysis is based on the assumption that all costs relating to the manufacture of a property can be divided into two categories: variable costs and fixed costs.

The total cost is defined in a linear model by the relation:

$$T_c=F+cQ \tag{15.4.40}$$

where F represents fixed costs, c the cost per unit (assumed to be independent of the quantity) of produced goods and Q the corresponding quantity.

The assumption that the cost per unit is independent of the quantity is used to simplify the calculation for the total cost which is a simple linear function, but this assumption is not really necessary today with spreadsheets softwares.

If we assume that all units produced are sold, the total revenue T_r is defined by:

$$T_r = pQ \tag{15.4.41}$$

where p is the revenue per unit sold.

The break-even point is then the equality between the two costs, which means:

$$pQ = F + cQ \tag{15.4.42}$$

So we could then for example after some elementary algebraic operations, determine the quantity that gives the break-even point:

$$Q = \frac{F}{p-c} \tag{15.4.43}$$

Thus, knowing F (fixed costs), p (revenue per produced unit) and c (costs per produced unit), it is possible to determine the quantity Q for sale to get the break-even point.

Note that a subject of great interest to managers is the study of the variation of some parameters depending on other variables. We name this the "sensitivity analysis" in the multivariate case (analysis based on the Pearson correlation coefficient) and "double entry analysis" in the bivariate case. Obviously in the linear case, this analysis is quite trivial.

Remark

Since the early 11980s (holocene calendar), there are very simple softwares to use such as @Risk of Palisade to do Monte Carlo simulations to determine the probabilistic break-even point and even solvers that solve non-linear systems and do high level analysis of the break-even point. Otherwise for a high-school level analysis, spreadsheets softwares are enough...

15.4.1.5 Investment Strategies

By definition, an investment is the economic motivation to do the acquisition or development of a property (regardless of its form, material or not) by a corporation, community or an individual. This practice is to make currently an expense for an entity and expected a return on investment in the future.

In finance, an investment strategy is more general a set of rules, behaviours or procedures, designed to guide an investor's selection of an investment. Individuals have different profit objectives, and their individual skills make different tactics and strategies appropriate. Some choices involve a tradeoff between risk and return or simply love for a product or a company. Most investors fall somewhere in between, accepting some risk for the expectation of higher returns or to maintain a the manufacturing of product they like to see exist (obviously from two or more investment strategies, the best at the individual level is the one that maximizes the final investor capital at equal risk...).

Actually, in this early beginning of the 121st century (holocene calendar), most academics teachers put apart the love for a product, or ethical/environmental aspects, to highlight and focus only on the monetary quantitative part. Indeed, measuring the utility for the love of a product is a quite difficult challenge... (however the ethical and environmental impacts of a strategical choice can be quantified quite easily!).

Then there exist different types of return on investment following the object of study. Thus, we differentiate in finance mainly (before seeing the details):

- The returns of financial asset on a given economic horizon ("return on investment") and their corresponding yield ("rate of return").
- The returns on investment on a project compared to an average (neutral risk) geometric rate of the market (stock exchange, OTC, etc.) that limits the interest to invest in a project ("internal rate of return").

We will focus here and the second and therefore the reader should know that corporate investments involves in an economical framework:

1. An immediate expense
2. Future entries or outputs, named "cash flow"
3. A residual value

There are several criteria and techniques for making and investment choice in corporate finance, which we will present below, that permits to opt for an investment A or B: that of the "net present value" (NPV), that of "internal rate return (IRR)" or that of "payback time (PT)" also named "recovery time".

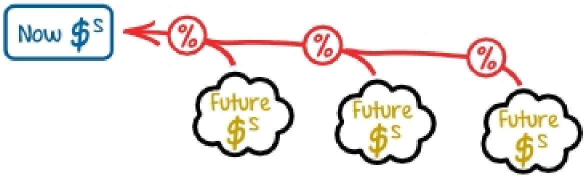
Remark

We must not forget also that mathematical decision techniques (see section Game and Decision Theory page 7376) have a huge importance when we considered that very height (or significant) amount of money are involved in an investment.

15.4.1.5.1 Net Present Value

In finance, the "net present value[53]" (NPV) or "net present worth" (NPW) is a measurement of the profitability of an undertaking that is calculated by subtracting the "present values" (PV) of cash outflows (including initial cost) from the present values of cash inflows over a period of time. Incoming and outgoing cash flows can also be described as benefit and cost cash flows, respectively.

The idea behind net present value is determined by calculating the costs (negative cash flows) and benefits (positive cash flows) for each period of an investment. Many people therefore say that an equivalent statement is that:



Money now is more valuable than money later on.

Why? Because you can use money to make more money! Indeed, you could run a business (assuming its without risk...), or buy something now and sell it later for more (assuming its not defective and there is no devaluation...), or simply put the money in the bank to earn interest (assuming it's also without risk...).

And as we have already mentioned it, investments involves in a simple economic framework:

1. An immediate expense payable in one or more times denoted V_0 in the NPV framework
2. Future entries, named "cash flow" C_i in the NPV framework
3. A residual value V_n in the NPV framework

in more accurate models we have to take into account the uncertainty and then we don't speak of NPV anymore but of "expected Net Present Value" (eNPV) or "risk-adjusted net present value" (rNPV).

The period of calculation of the NPV or eNPV typically one year, but could be measured in quarter-years, half-years or months. After the cash flow for each period is calculated, the present value of each one is achieved by discounting its future value (see below a recall of actuarial calculations we proved in the section Economy) at a periodic rate of return (the rate of return dictated by the market). NPV is the sum of all the discounted future cash flows. Because of its simplicity, NPV is a useful tool to determine whether a project or investment will result in a net profit or a loss. A positive NPV results in profit, while a negative NPV results in a loss. The NPV measures the excess or shortfall of cash flows, in present value terms, above the cost of funds. NPV is a central tool in "discounted cash flow" (DCF) analysis and is a standard method for using the time value of money to appraise long-term projects. It is as we already know now after our study of the section Economy widely used throughout economics, finance, and accounting.

---

[53] Net present value as a valuation methodology dates at least to the 19th century. Karl Marx refers to NPV as fictitious capital, and the calculation as "capitalising".

In the case when all future cash flows are positive, or incoming (such as the principal and coupon payment of a bond) the only outflow of cash is the purchase price, the NPV is simply the PV of future cash flows minus the purchase price (which is its own PV). NPV can be described as the "difference amount" between the sums of discounted cash inflows and cash outflows. It compares the present value of money today to the present value of money in the future, taking inflation and returns into account.

The NPV of a sequence of cash flows takes as input the cash flows and a discount rate or discount curve and outputs a price. The converse process in DCF analysis - taking a sequence of cash flows and a price as input and inferring as output a discount rate (the discount rate which would yield the given price as NPV) - is named the "yield" and is more widely used in bond trading.

What obviously interest and investor (in bonds, stocks, real estates, or whatsoever!) is that in present value, an investment more profitable than what is or would be spent.

So let us see now a companion example to formalized all that!

Consider the following standard situation: A company wishes to acquire a new machine worth $6,000$.-, which should lead to lower production costs of $1,000$.- during 5 years. We believe that in 5 years, the residual value of this machine will be $3,000$.-. Should we buy this machine if this investment can be financed by a loan at 10% (or in other words: if we have the option to invest in a fund having a 10% annual effective return yield)? What information do we have here?

- The immediate expense $V_0=6,000$.-
- The final or residual value of the equipment well after 5 years $V_n=3,000$.-
- The annually cash flows $C_k=1,000$.- (which are constant and positive over the entire period in this special example)
- The corresponding interest rate (geometric average market rate) of borrowing or investment corresponding $t\%=10\%$

What information, or interesting questions, financially speaking, can we ask from the data above?:

1. What should be the initial capital at a given market rate that would permit us to withdraw $1,000$.- per year during 5 years (until the account balance is zero) ?

From our study of actuarial calculations (see preceding section page 7547), we get:

$$\frac{C_1}{1+t\%}+\frac{C_2}{(1+t\%)^2}+\ldots+\frac{C_n}{(1+t\%)^n}=\sum_{k=1}^{n}\frac{C_n}{(1+t\%)^k} \tag{15.4.44}$$

Which will be written in the special case where $C_i = C$ (we fall back the relation of certain postnumerando annuity proved in the section of Economy during our study of actuarial calculations at page 7563), that is to say a simple present value calculation:

$$\text{PV}_{\text{cash flows}}=C\sum_{k=1}^{n}\frac{1}{(1+t\%)^k}=C\sum_{k=1}^{n}f_e^k=Ca_{\overline{n}|} \tag{15.4.45}$$

An numerical application with our example (after a little calculation) gives about $3,790$.-.

In other words, it would be enough for us to save $3,790$.- during the same 5 years, to earn $1,000$.- per year to settle the account. So for now, an investment of $3,790$.- to save (earn) $1,000$.- per year seems much more favourable than spend $6,000$.- for the same return over the same period...

Already here, we can say that the purchase of the machine is unfavourable at least... in certain future projection....

But we must not forget also a second factor ... the residual value of our machine!!!

2. What would be the initial capital that at the geometric average market rate would give us a value equivalent to the residual value of our machine (it is a capital asset just like a savings, so we can look at what would happen if this money came from savings)?

   So once again we use our knowledge of actuarial calculus by writing again the present value:

$$\text{PV}_{\text{residual}}=\frac{V_n}{(1+t\%)^n} \tag{15.4.46}$$

   In our special example, this quantity is equal to approximately $1,862$.-

In other words, it would suffice us to save $1,862$.- during the same 5 years to get an amount equal to the residual value of our machine. So what?

Well, the conclusion is quite simple .... The sum (in certain future and without intermediate taxes...):

$$\text{PV}_{\text{cash flows}}+\text{PV}_{\text{residual}}=\sum_{k=1}^{n}\frac{C_k}{(1+t\%)^k}+\frac{V_n}{(1+t\%)^n} \tag{15.4.47}$$

represents the return on the basis of an initial saving to get, relatively to the information of residual value and cash flow in certain future and without yearly taxes, the final sum equivalent to the final purchase of our machine. However, in this example, this quantity gives us about $5,652$.-.

This result is important because it is to be compared with the investment we wanted to do initially. Two options are therefore available to us:

- Buy $6,000$.- the machine with the cash flows and residual values that we know
- Save $5,652$.- during the same period, with the same cash flows to end up with a final savings that should be equivalent to the residual value of our machine.

Now, what can we conclude from this example? Well simply that we are in a worst case if:

$$\sum_{k=1}^{n}\frac{C_k}{(1+t\%)^k}+\frac{V_n}{(1+t\%)^n}<V_0 \Leftrightarrow \sum_{k=1}^{n}\frac{C_k}{(1+t\%)^k}+\frac{V_n}{(1+t\%)^n}-V_0<0 \tag{15.4.48}$$

Thus explicitly:

$$\text{NPV}=\sum_{k=1}^{n}\frac{C_k}{(1+t\%)^k}+\frac{V_n}{(1+t\%)^n}-V_0=\sum_{k=1}^{n}\frac{1000}{(1+10\%)^k}+\frac{3000}{(1+10\%)^5}-6000=-346.6 \tag{15.4.49}$$

In Microsoft® Excel® 14.0.7106, to get the result we just write (the writing is not obvious but when we thing about it, finally it's logic!):

```
=-6000+NPV(10%,1000,1000,1000,1000,4000)=-346.60
```

Remark

"Normal cash flow" is the cash flow stream that comprises of initial investment outlay and then positive net cash flow throughout the project life. It is also named "conventional cash flow stream". "Non-normal cash flow stream" (also named "unconventional cash flow") is a pattern of cash flows in which the direction of cash flows changes more than once. It is also termed as unconventional cash flow. "Terminal cash flow" is the net cash flow that occurs at the end of a project and represents the after-tax proceeds from disposal of the project assets and recoupment of working capital.

In other words, for the financial analyst, business analyst, or project manager, the calculation interesting to do is in certain future, without intermediate taxes:

$$\boxed{\text{NPV}=\sum_{k=1}^{n}\frac{C_k}{(1+t\%)^k}+\frac{V_n}{(1+t\%)^n}-V_0} \tag{15.4.50}$$

where the C_k can sometimes be negative (intermediate investments). So the three decision criteria are:

- $\text{NPV}<0$ corresponds to an investment that should be probably avoided
- $\text{NPV}=0$ we are in front of an undecidable investment
- $\text{NPV}>0$ corresponds probably to a good investment

Remark

A very similar relation well-known in Marketing is the "customer lifetime value" (CLV), that is a punctual probabilistic prediction of the net profit attributed to the entire future relationship with a given customer (also sometimes named "buy till your die" (BTYD) approach) and given that the quite logical following relation:

$$\boxed{\text{CLV}=\text{GC}\sum_{k=1}^{n}\frac{r^k}{(1+t\%)^k}-M\sum_{k=1}^{n}\frac{r^{k-1}}{(1+t\%)^{k-0.5}}-\text{AC}} \tag{15.4.51}$$

where GC is yearly gross contribution per customer, AC the acquisition cost of the customer, M is the (relevant) retention costs per customer per year (this formula assumes the retention activities are paid for each mid year and they only affect those who were retained in the previous year), n is the horizon (in years), r is the yearly retention rate (or probability of retention), $t\%$ is the yearly discount rate (risk free market rate).

Obviously such a model helps to predict future net-profit, identify promising customers and to prioritize customers according to future margins.

In most high-school level books and undergraduate project management books, the relation of NPV above is simplified, because they assume no residual values. Thus we have:

$$\text{NPV}=\sum_{k=1}^{n}\frac{C_k}{(1+t\%)^k}-V_0 \tag{15.4.52}$$

And let us recall that:

$$f_e^k=\frac{1}{(1+t\%)^k} \tag{15.4.53}$$

is often named "discount factor" of the kth period.

Some people still ask after these mathematical developments why is the NPV bigger when the interest rate is lower? Because the interest rate is like the team you are playing against, play an easy team (like a 6% interest rate) and you look good, a tougher team (like a 10% interest) and you don't look so good!

Thus, the NPV is used as a decision criterion in large cutting edge companies or governments to quantify the specific contribution of a project to the finance of the company, taking into account the cost of capital via the discount factor.

So in the case of a choice between several investments, we will choose the one whose NPV is highest. If the cash flows are not deterministic we will then have to calculate the mean and variance of the expected-NPV.

Remark

NPV is also often named "updated quasi-rent". We also find often the name "method of discounted cash flows".

If we consider in certain future and without taxes a $10,000$.- initial investment with a constant discount rate of 10% and a perfectly periodical cash flow of $3,250$.-, $3,750$.-, $4,250$.- and finally $4,750$.- we have also traditionally the following tabular representation:

Year	Action	Cash Flow	A.F.	A.C.F.
0	Investment	$-10,000$	1	$-10,000$
1	Input	$3,250$	0.909	$2,955$
2	Input	$3,750$	0.826	$3,099$
3	Input	$4,250$	0.751	$3,193$
4	Input	$4,750$	0.683	$3,244$
Sums	**O.N.C.**	$\mathbf{6,000}$	**O.N.A.C.**	$\mathbf{2,491}$

Table 15.68: **Detailed actualization table under accounting form**

where in the above table A.F. means "actualization factor", A.C.F. "actualized cash flow", O.N.C. "overall net cash flow" and finally O.N.A.C. "overall net actualized cash flow".

So in this table, the investment has a positive return of $2,491$.- more than a 10% investment operation after 4 years in certain future (deterministic Universe) without taxes.

We will notice on the way that starting from the third year, the initial investment is repaid. We speak then of "payback period" (see further below).

Finally, let us consider a typical academic case, the "perpetual present value":

$$\text{PPV}_{\text{cash flow}} = C\sum_{n=1}^{+\infty}\frac{1}{(1+t\%)^n} \tag{15.4.54}$$

Let us multiply this relation by $\frac{1}{1+t\%}$ and subtract by the previous relation to get:

$$\text{PPV}_{\text{cash flow}} - \text{PPV}_{\text{cash flow}}\frac{1}{1+t\%} = \frac{C}{1+t\%} \tag{15.4.55}$$

After rearranging we get:

$$\text{PPV}_{\text{cash flow}} = \frac{C}{t\%} \tag{15.4.56}$$

It's a common relation that you can find is some rare textbooks on finance.

Remark

In capital budgeting, "cannibalization" (also named "product cannibalization") occurs when an investment in a new product or project would eat away at the cash flows earned by an existing product.

Capital budgeting decisions must be taken by adopting a holistic approach, i.e. it should consider the impact of a business decision on the overall profitability of the business, and not just on one product. Hence, it is important to measure the impact of a new product on an existing product. This is often done by calculating the cannibalization rate.

"Cannibalization rate" (C.R.) is simply the ratio of total estimated revenue lost R_L to the total estimated revenue added R_A by the new product:

$$\text{C.R.} = \frac{R_L}{R_A} \tag{15.4.57}$$

15.4.1.5.2 Internal Rate of Return

Definition (technical #706): The "internal rate of return" (IRR) , also sometimes named "limit of profitability" or "marginal efficiency of capital" or "discounted cash flow rate of return" is the (actualization) rate $t\%$ for which the actualized value of net capital cash flows (present values) resulting from an investment project is equal to the present value of disbursements required to make this investment.

To quote the definition of J.M. Keynes: "*The marginal efficiency of capital is the discount rate that, applied to the series of annual payments made by the expected return of this capital during its entire existence, makes the present value of the annuities equal to the offering price of capital*".

In other words, this is equivalent to ask what is the geometric average rate of the market (discount rate) for which the NPV of an investment is zero (i.e.: above the first IRR root the NPV is negative!). Either that satisfies the relation:

$$\boxed{\text{NPV}=\sum_{k=1}^{n}\frac{C_k}{(1+t\%)^k}+\frac{V_n}{(1+t\%)^n}-V_0=0} \tag{15.4.58}$$

Since the IRR calculation is a manifestation of the general problem of finding the roots of the equation $\text{NPV}(t\%)=0$, there are many numerical methods that can be used to estimate $t\%$. For example, using the secant method that we have studied in the Numerical Methods section (see page 6606) we have recursively:

$$t\%_{n+1}=t\%_n-\text{NPV}(t\%_n)\cdot\left(\frac{t\%_n-t\%_{n-1}}{\text{NPV}(t\%_n)-\text{NPV}(t\%_{n-1})}\right) \tag{15.4.59}$$

where $t\%_n$ is considered the nth approximation of the IRR and which can only be calculated quickly with IT tools (the Target tool in Microsoft® Excel® 11.8346 for example) or simply using the `IRR( )` function integrated in almost all the spreadsheet softwares like Microsoft® Excel®.

✍**Example:**

A project manager proposes us to buy a machine for an immediate investment of $2,000$.- in a project with a cash flow that doubles each period on a basis of 400.- guaranteed for the three periods while the geometric average market interest is rate (discount rate) is of 5%. The IRR (internal rate of return) from which the NPV is zero is given with Microsoft® Excel® 14.0.7106:

`=IRR(-2000,400,800,1600)=15.117%`

Or analytically:

$$\text{NPV}=-2000+\frac{400}{1+t\%}+\frac{800}{(1+t\%)^2}+\frac{1600}{(1+t\%)^3}=0.\text{-} \tag{15.4.60}$$

When:

`=-2000+NPV(5%,400,800,1600)=488.71`

Therefore this project is obviously good as the free-risk market rate should be at least at 15.117% to be more interesting than investing in the project!

So between two investments obviously those whose IRR is the highest and meeting the internal political and economical constraints.

This type of calculation is therefore applicable on return on investment projects against market investments and not against the return on development projects. Thus, this is only computational tool to help the purely financial decision and not industrial or commercial one.

However, the NPV and IRR methods will obviously return conflicting results when mutually exclusive projects obviously differ in the timing of cash flows. In most cases, utilizing either the NPV or IRR method will lead to the same accept-or-reject decision. An exception exists when evaluating mutually exclusive projects with crossing NPV profiles and the cost of capital is less than the crossover rate. When these conditions are present, the NPV and IRR results will conflict in which project to accept or reject.

The presence of non-normal cash flows will lead obviously to multiple IRRs (for conventional projects that have outlays followed by inflows - negative cash flow followed by positive cash flow - the multiple IRR problem cannot occur!). Indeed, the IRR equation is essentially a n-th degree polynomial and as we know, such polynomial can have up to n solutions, although it will have no more real solutions than the number of cash flow sign change. But this is not an issue as financial analyst will keep the smallest solution[54].

There is also the case where the is no solution (No-IRR problem). That means there is no discount rate that results in a zero NPV[55]. A project with a No-IRR doesn't mean a project is bad but just that the NPV is positive whatever the discount rate!

So what do managers really use for capital budgeting? In a survey in 12001 (holocene calendar), Graham and Harvey (from Duke University) surveyed 392 managers, primarily chief financial officers (CFOs), asking them what techniques they use when deciding on projects or acquisitions. The results are listed in the figure below. The two most prominent measures are also the correct ones: They are the "internal rate of return" and the "net present value" methods. Alas, the troublesome "payback period" method and its cousin, the "discounted payback period" still remain surprisingly common:

Method	CFO Usage	Yields Correct Answer
Internal Rate of Return (IRR)	(76%)	Often
Net Present Value (NPV)	(75%)	(Almost) Always
Payback Period	(57%)	Rarely
Earning Multiples (P/E Ratios)	(39%)	With Caution
Discounted Payback	(30%)	Rarely
Accounting Rate of Return	(20%)	Rarely
Profitability Index	(12%)	Often

Figure 15.285: CF Valuation Techniques (source: *Corporate Finance An Introduction*, Ivo Welch)

---

[54] For example the cash flow pattern -1000.-, 5000.- and -6000.- has two IRR, one of 100% and one of 200%.

[55] For example the cash flow pattern 100.-, -300.- and 250.- has no IRR.

Definition (#15.4.707): The "profitability index" is an index that attempts to identify the relationship between the costs and benefits of a proposed project through the use of a ratio calculated as:

$$\boxed{\text{PI}=\frac{\displaystyle\sum_{k=1}^{n}\frac{C_k}{(1+t\%)^k}}{V_0}} \tag{15.4.61}$$

A ratio of 1.0 is logically the lowest acceptable measure on the index, as any value lower than 1.0 would indicate that the project's present value is less than the initial investment. As values on the profitability index increase, so does the financial attractiveness of the proposed project.

15.4.1.5.3 Payback

Definition (#15.4.708): The "payback" or "payback period p" is another indicator for decision support as part of the selection of investment projects and simpler of use (and understanding) than the NPV but that does not provide obviously the same type of information.

This indicator has for simple and only purpose to show when, in time, the investment will be repaid. In other words, it indicates the number of periods needed for the cash flow amounts C_i to cover the initial investment V_0. This is a very simple information to determine which is equivalent of found the smallest p such that:

$$V_0 \leq C_1+C_2+\ldots+C_p \tag{15.4.62}$$

or even:

$$V_0 \leq \sum_{k=1}^{p} C_k \tag{15.4.63}$$

In other words, it is the time for a project where the cash flows balance the initial investment.

The payback periods has some drawbacks. It is a measure of payback and not a measure of profitability! By itself, the payback would be a dangerous criterion for evaluating capital projects. Its simplicity, however, is an advantage. The payback period is very easy to calculate and to explain. The payback period may also be used as an indicator of project liquidity. A project with a two-year payback may be more liquid than another project with a longer payback.

Because it is not economically sound, the payback period has no decision rule like that of the NPV or IRR. If the payback period is being used (perhaps as a measure of liquidity), analysts should also use a NPV or IRR to ensure that their decision also reflect the profitability of the projects being considered.

Definition (#15.4.709): The "amortization time", or "discounted payback", is the time for a project where the actualized cash flows balance the initial investment. That is the smallest p such that:

$$V_0 \leq \frac{C_1}{1+t\%}+\frac{C_2}{(1+t\%)^2}+\ldots+\frac{C_p}{(1+t\%)^p} \tag{15.4.64}$$

or even:

$$V_0 \leq \sum_{k=1}^{p} \frac{C_i}{(1+t\%)^k} \tag{15.4.65}$$

The discounted payback period partially addresses the weaknesses of the payback period.

The discounted payback period relies on discounted cash flows, much as the NPV criterion does. If a project has a negative NPV, it will usually not have a discounted payback period since it never recovers the initial investment.

The discounted payback does account for the time value of money and risk within the discounted payback period, but ignores cash flows after the discounted payback period is reached. This drawback has two consequences. First, the discounted payback period is not a good measure of profitability (like the NPV or IRR) because it ignores these cash flows. Second, another idiosyncrasy of the discounted payback period comes from the possibility of negative cash flows after the discounted payback period is reached. It is possible for a project to have a negative NPV but to have a positive cumulative discounted cash flow in the middle of its life and, thus, a reasonable discounted payback period. The NPV and IRR, which consider all of project's cash flow, do not suffer from this problem.

15.4.1.6 Company Valuation Methods

For anyone involved in the field of corporate finance, understanding the mechanisms of company valuation is an indispensable requisite[56]. This is not only because of the importance of valuation in acquisitions and merges but also because the process of valuing the company and its business units helps identify sources of economic creations and destruction within the company and also its IPO value.

Definition (#15.4.710): The "initial public offering" (IPO) or "stock market launch" is a type of public offering in which shares of a company usually are sold to institutional investors[that in turn, sell to the general public, on a securities exchange, for the first time. Through this process, a privately held company transforms into a public company. Initial public offerings are mostly used by companies to raise the expansion of capital, possibly to monetize the investments of early private investors, and to become publicly traded enterprises. A company selling shares is never required to repay the capital to its public investors. After the IPO, when shares trade freely in the open market, money passes between public investors. Although IPO offers many advantages, there are also significant disadvantages, chief among these are the costs associated with the process and the requirement to disclose certain information that could prove helpful to competitors. The IPO process is colloquially known as going public.

The methods for valuing companies can be classified commonly in six groups:

1. Balance sheet (book value/adjusted book value)

---

[56] Following Warren Buffet an investor should put its money only in corporations that have a dominant position, a sustainable business, whose know-how is quite hard to copy and that have shown in the past a sustainable capacity to make profits.

2. Income statement (multiples per sales or other multiples)
3. Mixed Goodwill (Classic Union of European Accounting Experts standard method)
4. Cash flow discounting (equity cash flow dividend or other cash flows)
5. Value creation (Economic profit)
6. Options (Black & Scholes types evaluation)

In this textbook as always we will focus only on the mathematical aspect of theses methods and we will provide also a list that contains most common errors!

The methods that are popular (and are conceptually "correct") are those based on cash stochastic flow discounting. These methods view the company as a cash flow generator and, therefore, assessable as a financial asset. We will briefly comment on other methods since - even tough they are conceptually "incorrect" - they continue to be used frequently.

A valuation may be used for a wide range of purposes:

1. In corporations buying and selling operations:
   - For the buyer, the valuation will tell him the highest price he should pay
   - For the seller, the valuation will tell him the lowest price at which he should be prepared to sell
2. Valuations of listed corporations:
   - The valuation is used to compare the value obtained with the share's price on the stock market and to decide whether to sell, buy or hold the share
   - The valuation of several companies is used to decide the securities that the portfolio should concentrate on and therefore to compare companies
3. To valuate the price at which shares are offered to the public and therefore to compare it with other assets
4. To identify and stratifying the main values drivers of a corporations
5. As decision tool (key development point) to continue business:
   - For deciding what products/business lines/countries/customers... to maintain, grow or abandon.
   - To measure the impact of the corporation possible policies and strategies on value creation and destruction

15.4.1.6.1 Balance sheets-based method

Definition (#15.4.711): The "balance sheets-based method" or "shareholders' equity" seek to determine the company's value by estimating the value of its assets. These are traditionally used method that consider a company's value lies basically in it's balance sheet. Therefore this method is a static deterministic viewpoint, which therefore, does not take into account the company's possible future evolution or money's temporary value and also implicitly estimation errors (tolerance interval). Neither do they take into account other factors that also affect the value such as: the industry's current situation, human resources or organizational problems, contracts, etc. that do not appear in the accounting statements.

Some of the methods are the following: adjusted book value, liquidation value and substantial value.

This method and all accounting base method are quite poor as they only give a valuation for today times t when the investor is interested to $t + \Delta t$. However this method seems to be the most popular because the required knowledge to apply it is quite low and the results can therefore be understand by the board committees that have most of time no knowledge in engineering finance (their knowledge is limited to spreadsheet softwares sum and averages most time).

15.4.1.6.1.1 Book Value

Definition (#15.4.712): A company's "book value", or "net worth" or "common equity", is the value of the shareholder's equity stated in the balance sheet (capital and reserves). This quantity is also the difference between total assets and liabilities, that is, the surplus of the company's total goods and rights over its total debts with third parties.

✍**Example:**

In the table below the shares' book value (capital plus reserves) is 80 M$. It can also be calculated as the difference between total assets (160) and liabilities $(40 + 10 + 30)$, that is 80 millions dollars:

Assets	M$	**Liabilities**	M$
Cash	5	Accounts payable	40
Account receivables	10	Bank dept	10
Inventories	45	Long term dept	30
Fixed assets	100	Capital and reserves	80
Total assets	**160**	**Total liabilities**	**160**

Table 15.69: **Alfa Inc. Official Balance Sheet (Millions Dollars)**

The implicit and most of time unknown implicit error of balance sheet and the fact that it is only a static today estimation makes that the book value of a company almost never matches the market value.

A company's standard book value may be significantly affected by intangible assets and/or goodwill such as intellectual property (patents and trademarks), brand recognition, and trade secrets. On the balance sheet, intangible assets add to the total asset value and, as such, are a component of the total book value.

Definition (#15.4.713): Therefore, a "tangible book value" (alternately "net tangible assets" or "net tangible equity") is the total net asset value of a company (book value) minus intangible assets and goodwill.

> **Remark**
>
> In most cases, GAAP (Generally Accepted Accounting Principles) does not allow internally generated intangibles to show up on the balance sheet.

15.4.1.6.1.2 Adjusted Book Value

Definition (#15.4.714): The "adjusted book value" is a method that seeks to overcome the shortcomings that appear when purely accounting criteria are applied in the valuation.

When the values and liabilities match their market value, the adjusted net worth is obtained. Continuing with the example used above, we will analyse a number of balance sheet items individually in order to adjust them to their approximate market value. For example if we consider that:

- *Accounts receivable* includes 2 M$ of bad debt, this item should have a value of 8 M$.
- *Inventories*, after discounting obsolete, worthless items and revaluing the remaining items at their market value, has a value of 52 M$
- *Fixed assets* (land, buildings and machinery) have a value of 150 M$ according to an expert
- *Accounts payable*, *Bank dept* and *Long-term dept* is equal to their market value

The adjusted balance sheet would be that shown below:

Assets	M$	Liabilities	M$
Cash	5	Accounts payable	40
Account receivables	8	Bank Dept	10
Inventories	52	Long term dept	30
Fixed Assets	150	Capital and reserves	135
Total assets	**215**	**Total liabilities**	**215**

Table 15.70: Alfa Inc. Official Adjusted Balance Sheet (Millions Dollars)

The adjusted book value is the 135 M$ (in other words: different from the book value!).

15.4.1.6.1.3 Liquidation Value

Definition (#15.4.715): The "liquidation value" is the company's value if it is liquidated, that is, its assets are sold and its debts are paid off. This value is calculated by deducting the business's liquidation expenses (redundancy payments to employees, tax expenses and other typical liquidation expenses) for the adjusted net worth.

Taking the example given in the table above, if the redundancy payments and other expenses associated with the liquidation of the company Alfa Inc. were to amount 60 M$, the shares' liquidation value would be 75 M$ ($135-60$).

Obviously, this method's usefulness is limited to a highly specific situation, namely, when the company is bought with the purpose of liquidating it at a later date. However, it always represents the company's minimum value as a company's value, assuming it continues to operate, is greater than its liquidation value.

15.4.1.6.1.4 Substantial Value

Definition (#15.4.716): The "substantial value" represents the investment that must be made to form a company having identical conditions as those of the company being valued. It can also be defined as the "asset's replacement value", assuming the company continues to operate, as opposed to their liquidation value. Normally, the substantial value does not include those assets that are not used for the company's operations (unused land, holdings in other companies, etc.).

Three types of substantial value are usually defined for the book value and to avoid the reader going up and down in the page let us copy here the above table:

Assets	M$	**Liabilities**	M$
Cash	5	Accounts payable	40
Account receivables	8	Bank Dept	10
Inventories	52	Long term dept	30
Fixed Assets	150	Capital and reserves	135
Total assets	**215**	**Total liabilities**	**215**

1. "Gross substantial value": this is the assets' value at market price (in the above example the value of: 215 M$).

2. "Net substantial value" or "corrected net assets": this is the gross substantial value less liabilities. It is also know as "adjusted net worth", which we have already seen earlier (in the above example the value of: $135=215-80$ M$).

3. "Reduced gross substantial value": this is the gross substantial value reduce only by the value of the cost-free debt (in our example: $175=215-40$). The remaining 40 millions dollars correspond to accounts payable.

15.4.1.6.1.5 Book Value and Market Value

Definition (#15.4.717): The "price-book value ratio P/BV" (or simply P/B), also sometimes known as "Market-to-Book ratio", is the ratio of the market value of equity to the book value of equity. Price stands for the current market price of a stock. Book value is the total assets minus liabilities, or net worth, which is the accounting measure of shareholders' equity in the balance sheet:

$$P/\text{BV} = \frac{\mathcal{M}_{\mathcal{E}}}{\mathcal{B}_{\mathcal{E}}} \tag{15.4.66}$$

where $\mathcal{M}_{\mathcal{E}}$ is the market value of equity, and $\mathcal{B}_{\mathcal{E}}$ the book value of equity.

In general, the equity or also the tangible book value has little bearing to its market value. This can be seen in the figure below:

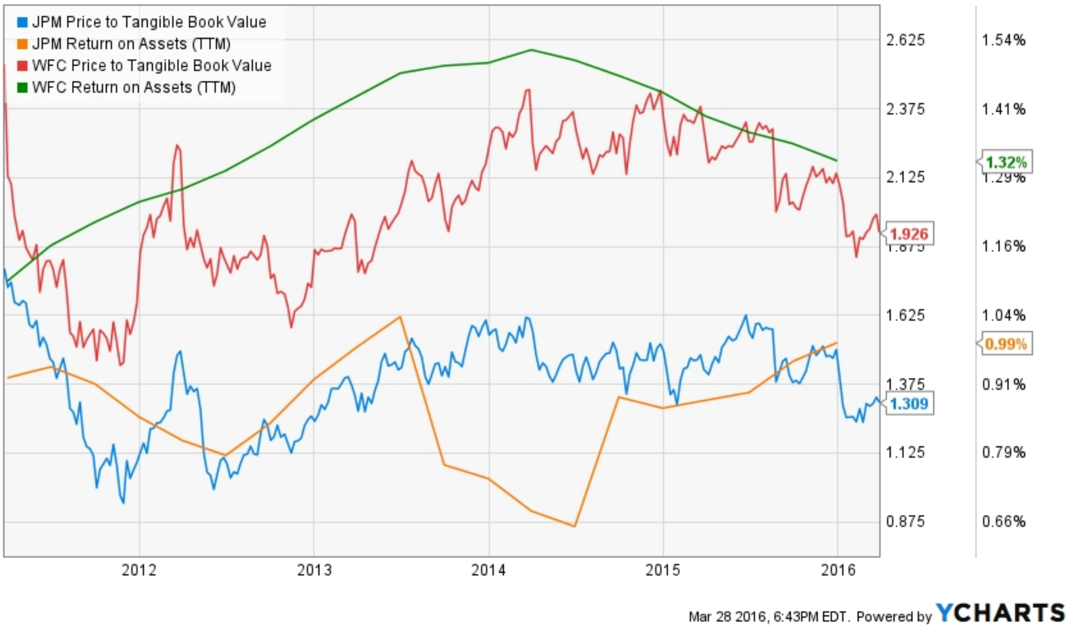


Figure 15.286: P/B **for JP Morgan Chase and Wells Fargo Company (source: Source: YCharts)**

15.4.1.6.2 Income Statemented-Based Methods

Unlike the balance sheet-based methods, these methods are based on the company's income statement. They seek to determine the company's value through the size of its earnings, sales or other indicators. Thus, for example, it is a common practice to perform quick valuations of companies by multiplying their annual production capacity (or sales) in metric tons by a ratio (multiple). It is also common to value car parking lots by multiplying the number of parking spaces by a multiple and to value insurance companies by multiplying annual premiums by a multiple. This category includes the methods based on the PER: accord to this method, the share's price is a multiple of the earnings.

The income statement of an imaginary company is shown in the table below:

Imagininary LL Income Statement (M$)	
Sales	300
Cost of sales	136
General expenses	120
Interest expense	4
Earnings before tax	**40**
Tax (35%)	14
Total assets	**26**

15.4.1.6.2.1 Value of Earnings with PER

According to this method, the equity value is obtained by multiplying the annual net income by the PER ratio (see section Economy page 7610), that is:

$$\text{EV} = \text{PER} \cdot \text{Earnings} \tag{15.4.67}$$

The figure below shows the evolution of the PER for the S&P 500 for little bit more than a hundred years:

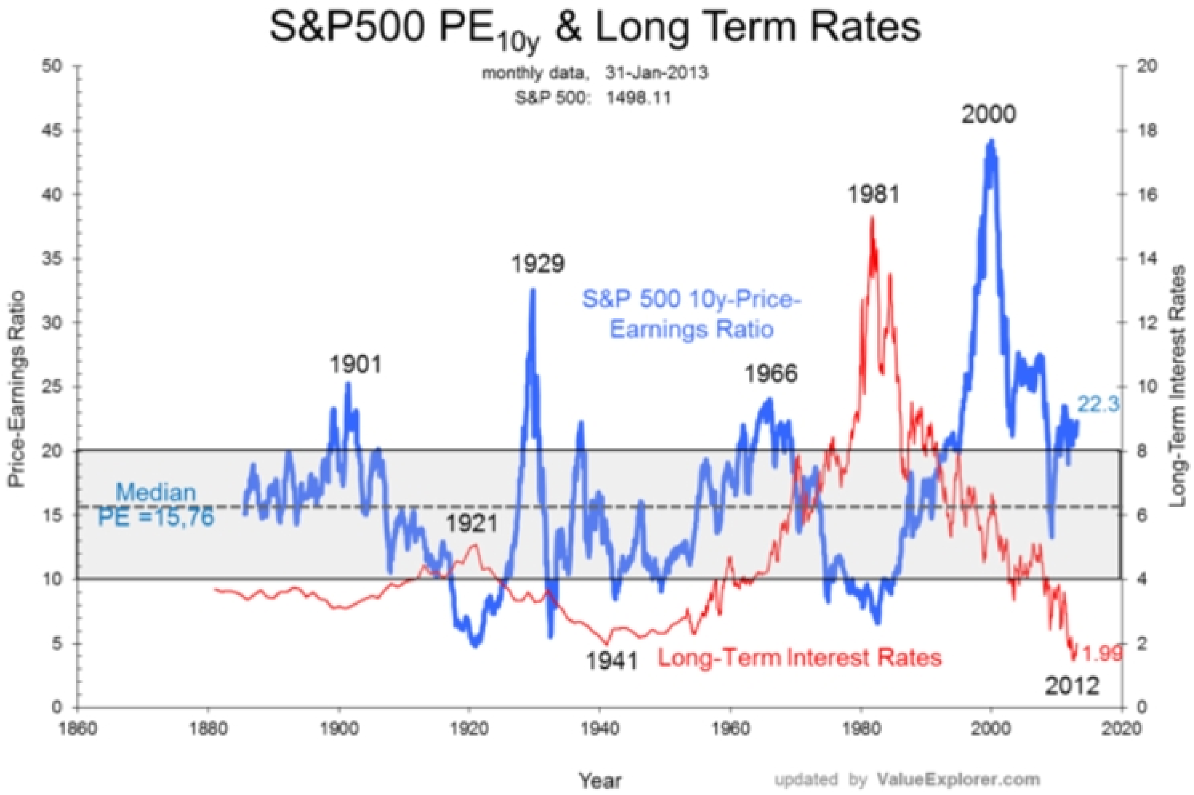


Figure 15.287: $10y$ P/E Ratio for S&P 500 (source: ValueExplorer.com)

Sometimes, the relative PER is also used, which is simply the company's PER divided by the country's PER.

15.4.1.6.2.2 Value of Dividends

Dividends are the part of the earnings effectively paid out to the shareholders and, in most cases, are the only regular flow received by shareholders. According to this method, a share's value is the net present value of the dividends that we expect to obtain from it. In the perpetuity case, that is, a company from which we expect constant dividends every year.

The Bates formula establishes a relationship between the current price earning ratio (see section Economy page 7610) at time 0 (PER_0) of a stock and its future P/E after n years (PER_n), given the dividend per share D_i at period i, the risk free rate of return $t\%$ during the periods (supposed as constant). Obviously this formula is based on the method used to determine the present value of cash flows on a given period (n years) and the terminal value V_n of the company.

We start for the proof from the "company valuation formula" or "Irving Fisher fundamental value" given under simplistic assumptions by a simple present value calculation (see section Economy page 8050):

$$V_0=\sum_{i=1}^{n}\frac{D_i}{(1+t\%)^i}+\frac{V_n}{(1+t\%)^n} \tag{15.4.68}$$

according to the notations in usage in the field of corporation evaluation.

Knowing that in general:

$$V_i=\text{PER}_i\cdot\text{BPS}_i \tag{15.4.69}$$

where BPS represent the benefice per share at the year i and that:

$$D_i=d\cdot\text{BPS}_i \tag{15.4.70}$$

where d is the fraction of the benefice per share distributed to shares holders, we can rewrite the above relation as following:

$$\text{PER}_0\cdot\text{BPS}_0=\sum_{i=1}^{n}\frac{d\cdot\text{BPS}_i}{(1+t\%)^i}+\frac{\text{PER}_n\cdot\text{BPS}_n}{(1+t\%)^n} \tag{15.4.71}$$

Following the same assumption as the Gordon-Shapiro model (see section Economy page 7613) and using the corresponding usual notation we know that we can write:

$$\text{BPS}_n=\text{BPS}_0(1+g)^n\Rightarrow\text{BPS}_i=\text{BPS}_0(1+g)^i \tag{15.4.72}$$

with BPS_0 being the actual benefice per share.

We get then get:

$$\frac{\text{BPS}_i}{\text{BPS}_0}=(1+g)^i \tag{15.4.73}$$

So rearranging the prior previous relation we get:

$$\begin{aligned}\text{PER}_0 &= \sum_{i=1}^{n}\frac{d\left(\dfrac{\text{BPS}_i}{\text{BPS}_0}\right)}{(1+t\%)^i}+\frac{\text{PER}_n\left(\dfrac{\text{BPS}_n}{\text{BPS}_0}\right)}{(1+t\%)^n}\\ &= \sum_{i=1}^{n} d\frac{(1+g)^i}{(1+t\%)^i}+\text{PER}_n\frac{(1+g)^n}{(1+t\%)^n}\\ &= d\sum_{i=1}^{n}\left(\frac{1+g}{1+t\%}\right)^i+\text{PER}_n\left(\frac{1+g}{1+t\%}\right)^n\end{aligned}\tag{15.4.74}$$

The last equality is the "Bates formula":

$$\boxed{\text{PER}_0 = d\sum_{i=1}^{n}\left(\frac{1+g}{1+t\%}\right)^i+\text{PER}_n\left(\frac{1+g}{1+t\%}\right)^n}\tag{15.4.75}$$

However for more elegance it is convenient to notice that the first sum is a geometric series of common ratio:

$$K=\left(\frac{1+g}{1+k_e}\right)\tag{15.4.76}$$

Which leads us to write (see section Economy page 7563):

$$\begin{aligned}d\sum_{i=1}^{n}\left(\frac{1+g}{1+t\%}\right)^i &= d\sum_{i=1}^{n}K^i=dK\left(\frac{K^n-1}{K-1}\right)=d\frac{\dfrac{1+g}{1+t\%}}{\dfrac{(g-t\%)}{1+t\%}}(K^n-1)\\ &= d\left(\frac{1+g}{g-t\%}\right)(K^n-1)\end{aligned}\tag{15.4.77}$$

So we can rewrite the Bates formula as following:

$$\boxed{\text{PER}_0 = d\left(\frac{1+g}{g-t\%}\right)(K^n-1)+\text{PER}_n K^n}\tag{15.4.78}$$

Empirical evidence shows that the companies that pay more dividends (as a percentage of their earning) do not obtain a growth in their share price as a result. This is because when a company distributes more dividends, normally it reduces its growth because it distributes the money to its shareholders instead of plowing it back into new investments.

Empirical observations also show that shares value increase many times just before dividend pay time and decease just after.

15.4.1.6.2.3 Sales Multiples

This valuation method, which is used in some industries with a certain frequency, consists of calculating a company's value by multiplying its sales by a number. For example, a pharmacy is often valued by multiplying its annual sales (in dollars) by 2 or another number, depending on the market situation. It is also a common practice to value a soft drink bottling play by multiplying its annual sales in litres by 500 or another number, depending on the market situation.

15.4.1.6.3 Goodwill-Based Methods

Generally speaking, goodwill is the value that a company has above its book value or above the adjusted book value, Goodwill seeks to represent the value of the company's intangible assets, which often do not appear on the balance sheet but which, however, contribute an advantage with respect to other companies operating in the industry (quality of the customer portfolio, industry leadership, brands, strategic alliances, etc.). The problem arises when one tries to determine its value, as there is no consensus regarding the methodology used to calculate it. Some of the methods used to value the goodwill rise to the various valuation procedures described below.

These methods apply a mixed approach: on the one hand, they perform a static valuation of the company's assets and, on the other hand, they try to quantify the value that the company will generate in the future. Basically, these methods seek to determine the company's value by estimating the combine value of its assets plus a capital gain resulting from the value of its future earnings: they start by valuing the company's assets and then add a quantity relative with future earnings.

15.4.1.7 Capital Goods

A capital good is a durable good (one that does not quickly wear out) that is used in the production of goods or services. Capital goods are one of the three types of producer inputs, the other two being land and labour, which are also known collectively as primary factors of production. In the study of economic systems and in Marxian economics, the term means of production is often used with the same meaning as capital goods. This classification originated during the classical economic period and has remained the dominant method for classification.

Capital goods are acquired by a society by saving wealth which can be invested in the means of production. In terms of economics one can consider capital goods to be tangible. They are used to produce other goods or services during a certain period of time. Machinery, tools, buildings, computers, or other kind of equipment that is involved in production of other things for sale represent the term of a Capital good. The owners of the Capital good can be individuals, households, corporations or governments. Any material that is used in production of other goods also is considered to be capital good.

Capital goods undergo a gradual depreciation due to wear or obsolescence. This decline in value recorded as an expense in accounting in many countries around the world, is named "book depreciation". The reader must not confuse financial amortization as seen in the section of Economy (page 7570) and corresponding to debt repayment and the accounting depreciation that is a decrease in value of the means of production.

Certain assets have lost value fairly uniform in time unlike others that depreciate quickly the first few years, some are just empirical at defined by governments laws. Here, we discuss some of the accounting methods used in practice that describe either of these phenomena.

15.4.1.7.1 Linear Amortization

Definition (#15.4.718): We speak of a "linear amortization" of a capital good when its asset value is decreased (depreciated) of a periodic constant amount (annually in accounting) during its lifetime.

Thus, if we denote by A_n the amount of the k-th amortization and the initial value of the capital good V_0 and V_n its desired final value (which in most cases will be zero), we have obviously:

$$A=A_k=\frac{V_0-V_n}{m} \tag{15.4.79}$$

It is possible to obtain the depreciation value using the function `SLN( )` of any spreadsheet software like of Microsoft® Excel® 11.8346.

The equivalent constant depreciation rates based on the purchase and residual value is then given of course by:

$$t\%=\frac{\dfrac{V_0-V_n}{n}}{V_0} \tag{15.4.80}$$

Notice that this constant rate and the final value are sometimes imposed by legislation in some countries (as it is the case in Switzerland for example) to a residual value of zero or non-zero as defined by the law!

15.4.1.7.2 Arithmetic Declining Amortization

Definition (#15.4.719): When the asset value (depreciation) of good decreases inversely to the order of periods (accounting years) then we talk of "arithmetic declining depreciation" (there is therefore so no amortization rate that can be required by the government!).

For example, a good of 4 years life, will be depreciated the first year $4/10$, $3/10$ the second years, $2/10$ the third year, $1/10$ the last year. The common base "10" (in this example) is the arithmetic sum of $1+2+3+4$ so that the sum of all are equal to unity. This is a purely American tax rule named "Sum-of-years'-digits" (SYD).

Remark

Do not confuse the "arithmetic declining depreciation" with the "declining balance" that consist to apply a multiplier (enacted by the tax authorities) to the linear amortization and that we will not discuss here (except on explicit request of readers). It can only be used for new goods and does not apply to all types of asset. This declining balance rate will apply each year on the residual accounting value of the good.

Given the k-th amortization and V_0 the initial value of the good and V_n its desired final value (which in most cases will be zero) then we have:

$$\begin{aligned} A_1 &= \frac{4}{10}(V_0 - V_n)\\ A_2 &= \frac{3}{10}(V_0 - V_n)\\ A_3 &= \frac{2}{10}(V_0 - V_n)\\ A_4 &= \frac{1}{10}(V_0 - V_n) \end{aligned} \tag{15.4.81}$$

That can be written in the most general case:

$$A_k = \frac{n-k+1}{1+2+3+4}(V_0 - V_n) = \frac{n-k+1}{S_n}(V_0 - V_n) \tag{15.4.82}$$

and as we have proved it in the section of Sequences and Series (see page 1653):

$$S_n = \frac{n(n+1)}{2} \tag{15.4.83}$$

Which takes us to write:

$$A_k = \frac{2(V_0 - V_n)}{n(n+1)}(n-k+1) \tag{15.4.84}$$

You can get the amortization at a period k with a spreadsheet software using for example the function `SYD( )` of Microsoft® Excel® 11.8346.

15.4.1.7.3 Geometric Declining Amortization (declining balance)

Definition (#15.4.720): When the capital value of good decreases at a constant rate of depreciation based on the residual value of the previous period (that is to say like compound interest with negative rate), then we talk "simple geometric declining amortization".

Thus, the value of the good after n years is given by:

$$V_n = V_0(1-t\%)^n \qquad (15.4.85)$$

with then (elementary algebra):

$$t\% = 1-\sqrt[n]{\frac{V_n}{V_0}} \qquad (15.4.86)$$

Remark

We see that the value $t\%$ is in the range between $[0,1[$. The limit $(1-t\%)^n$ as n tends to big values is never zero. Thus, the residual value never will either!

Knowing by definition of this depreciation that:

$$A_k = V_{k-1}t\% \qquad (15.4.87)$$

we get:

$$A_k = V_0(1-t\%)^{k-1}t\% \qquad (15.4.88)$$

By injecting the expression of the in the above relation, we get:

$$\begin{aligned}A_k &= V_0\left(1-\left(1-\sqrt[n]{\frac{V_n}{V_0}}\right)\right)^{k-1}\left(1-\sqrt[n]{\frac{V_n}{V_0}}\right) = V_0\left(\sqrt[n]{\frac{V_n}{V_0}}\right)^{k-1}\left(1-\sqrt[n]{\frac{V_n}{V_0}}\right)\\ &= V_0\left(\frac{V_n}{V_0}\right)^{\frac{k-1}{n}}\left(1-\sqrt[n]{\frac{V_n}{V_0}}\right) = V_0\left(\left(\frac{V_n}{V_0}\right)^{\frac{k-1}{n}} - \left(\frac{V_n}{V_0}\right)^{\frac{1}{n}}\left(\frac{V_n}{V_0}\right)^{\frac{k-1}{n}}\right)\\ &= V_0\left(\left(\frac{V_n}{V_0}\right)^{\frac{k-1}{n}} - \left(\frac{V_n}{V_0}\right)^{\frac{k}{n}}\right)\end{aligned} \qquad (15.4.89)$$

We notice that the depreciation values do not need to know the depreciation rate explicitly. Just knowing the final and initial value is enough. This is precisely why the `DB( )` function of the spreadsheet software Microsoft® Excel® 11.8346 does not require the depreciation rate.

We notice also that this constant rate and the final value are sometimes imposed by legislation in some countries (as is the case in Switzerland for example) to a residual value of zero or non-zero also defined by the law!

Remark

The declining geometric depreciation is particularly suitable for goods with a very sharp depreciation in the early years.

15.4.2 Project Management

Project management is the discipline of initiating, planning, executing, controlling, and closing the work of a team to achieve specific goals and meet specific success criteria. A project is a temporary endeavour designed to produce a unique product, service or result with a defined beginning and end (usually time-constrained, and often constrained by funding or deliverables) undertaken to meet unique goals and objectives, typically to bring about beneficial change or added value. The temporary nature of projects stands in contrast with business as usual (or operations), which are repetitive, permanent, or semi-permanent functional activities to produce products or services. In practice, the management of these two systems is often quite different, and as such requires the development of distinct technical skills and management strategies.

As many technical Project Management subjects are related to Corporate Finance (empirical index, sensitivity analysis) and hence to Economy (Actuarial, Interests calculations, Loan Amortization/Repayment, Portfolio Theory and Forecasting), Optimization techniques (see section Numerical Methods page 6645) and also Industrial Engineering (see section of the same name page 8673) we will not come back on this subjects that we have already studied previously in details or that we will study in details.

We will focus here only on subjects that are very specific to quantitative project management and are commonly designated under the name of "Project Management Engineering".

Remark

Before reading the undergraduate project management quantitative techniques that will follow below, we strongly recommend the reader to revise high school level techniques related to time management like the harmonic mean of working rates (see section Statistics page 564).

15.4.2.1 Probabilistic PERT (Beta-PERT Distribution)

Always remaining within the framework of statistics and probabilities with respect to mathematical management techniques, there is a project management empirical law (domain we assume known by the reader as trivial) widely used in the context of "probabilistic PERT" (where PERT stands for "Program Evaluation and Review Technique") and included in a lot of specialized softwares to make cost and time predictions for project plannings (see section Graph Theory page 3016). We speak then about "Probabilistic Network Evaluation Technique" (PNET).

This empirical distribution, named "beta distribution" or sometimes "PERT law", or also "Beta-PERT Distribution", is often presented as follows in the books and without proof (...) and applies as well to costs or time estimation:

$$t_{Pr}=\frac{t_O+4t_V+t_P}{6} \tag{15.4.90}$$

and gives the expected mean duration (or cost) t_{Pr} of an elementary task (not downsizable into sub-tasks) where we have t_O, t_V, t_P which are respectively the optimistic, most likely and pessimistic duration (or cost) of the task. We will show below the origin of this relation and why it is partially false when written in a such simple way!

Remark

This classical approach dates from 11962 (holocene calendar) and is due to C.E. Clark. It is said to be based on a project task database analysis. In facts you should always check what kind of law (probabilistic distribution) follow the tasks in your company and not apply just stupidly and simply the beta distribution.

The principles of this approach is based on the idea that the duration (or cost) of each project task is considered random and Beta distribution is systematically used as a more accurate approach than the triangular distribution (see section Statistics page 743). The parameters of this distribution as we will prove it are determined through calculation of a fairly strong assumption, based on extreme values a and b as the execution time (cost) can take, and is modal value.

So we just have to ask ourselves following two questions:

1. What is the minimum physically possible time or cost execution?
2. What is the maximum physically possible time or cost execution?

to obtain the parameters a, b, M_0, which are then used to calculate the average (expected mean), the variance and the shape of the distribution of this random time/cost.

Then we determine the critical path of the project (using the Potential Metra Method supposed to be known by the reader after its study of the section on Graph Theory at page 3017), placing ourselves in some universe and using a given probabilistic duration for each task obtained with the beta distribution, allowing to find afterwards the probabilistic critical paths and thus the total probabilistic duration/cost of the whole project.

This done some practitioners place themselves then in random universe where the project duration is considered as the sum of the durations of critical path tasks previously identified. They use then the central limit theorem that we have proved in section Statistics (recall that this theorem states, under generally respected conditions that the random variable consisting of a sum of n independent random variables identically distributed approximately follows a Normal distribution, whatever are the original distributions as soon as n is large enough) to approximate the density distribution law of the whole project itself. Even if this practice is simple and useful it is in fact dangerous because in reality task are not equally distributed and can takes the practitioners to huge bias error.

Remark

In a rigorous way the practitioner should use Monte Carlo simulations with special softwares like @Risk of Palisade to estimate the probabilistic time/cost of the whole project.

If X_i is the corresponding cost or duration of task i the using the property of linearity of the mean (see section Statistics page 571) we have for the mean cost or duration of the whole project:

$$\mathrm{E}(X)=\mathrm{E}\left(\sum_i X_i\right)=\sum_i \mathrm{E}(X_i) \qquad (15.4.91)$$

and in the strong hypothesis that all tasks of a project are independent we can still write that the global variance (see section Statistics page 596) is:

$$\mathrm{V}(X)=\mathrm{V}\left(\sum_i X_i\right)=\sum_i \mathrm{V}(X_i) \tag{15.4.92}$$

Now let us go a little bit more deeply in technical tools for project duration and cost estimation.

Recall that we saw in the sections of Statistics (see page 761) and of Differential and Integral Calculus (see page 1574) that:

$$B(p,q)=\frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}=\int\limits_0^1 t^{p-1}(1-t)^{q-1}\mathrm{d}t \tag{15.4.93}$$

and:

$$\Gamma(x+1)=x\Gamma(x) \tag{15.4.94}$$

In management we notice that often (applying an Anderson-Darling Goodness of fit test as seen in the section of Statistics at page 1130) that tasks durations follow a distribution that we name "beta distribution of the first kind" (see section Statistics page 760) given by:

$$P_{0,1}(x)=\frac{x^\alpha(1-x)^\gamma}{\int\limits_0^1 t^\alpha(1-t)^\gamma\mathrm{d}t}1_{[0,1]} \tag{15.4.95}$$

For an interval $[a,b]$ which includes x, we can obtain the more general form:

$$P_{a,b}(x)=\frac{(x-a)^\alpha(b-x)^\gamma}{(b-a)^{\alpha+\gamma+1}\int\limits_0^1 t^\alpha(1-t)^\gamma\mathrm{d}t}1_{[a,b]} \tag{15.4.96}$$

Theorem 15.8. Let us check that we have well:

$$\int\limits_a^b P_{a,b}(x)\mathrm{d}x=\frac{1}{(b-a)^{\alpha+\gamma+1}\int\limits_0^1 t^\alpha(1-t)^\gamma\mathrm{d}t}\int\limits_a^b(x-a)^\alpha(b-x)^\gamma\mathrm{d}x=1 \tag{15.4.97}$$

Proof 15.8.1. By the change of variable:

$$u=\frac{x-a}{b-a}\Rightarrow\mathrm{d}u=\frac{1}{b-a}\mathrm{d}x \qquad \text{and} \qquad 1-u=1-\frac{x-a}{b-a}=\frac{b-x}{b-a} \tag{15.4.98}$$

we get:

$$\begin{aligned}\int\limits_a^b P_{a,b}(x)\mathrm{d}x&=\frac{\int\limits_0^1((b-a)u)^{\alpha}(b-x)^{\gamma}(b-a)\mathrm{d}u}{(b-a)^{\alpha+\gamma+1}\int\limits_0^1 t^{\alpha}(1-t)^{\gamma}\mathrm{d}t}=\frac{\int\limits_0^1(u^{\alpha}(b-x)^{\gamma}\mathrm{d}u}{(b-a)^{\gamma}\int\limits_0^1 t^{\alpha}(1-t)^{\gamma}\mathrm{d}t}\\&=\frac{\int\limits_0^1(u^{\alpha}\left(\frac{b-x}{b-a}\right)^{\gamma}\mathrm{d}u}{\int\limits_0^1 t^{\alpha}(1-t)^{\gamma}\mathrm{d}t}=\frac{1}{\int\limits_0^1 t^{\alpha}(1-t)^{\gamma}\mathrm{d}t}\int\limits_0^1 u^{\alpha}\,(1-u)^{\gamma}\,\mathrm{d}u=1\end{aligned}\tag{15.4.99}$$

Let us now determine the expected mean:

$$\mu=\int\limits_a^b xP_{a,b}(x)\mathrm{d}x=\frac{1}{(b-a)^{\alpha+\gamma+1}\int\limits_0^1 t^{\alpha}(1-t)^{\gamma}\mathrm{d}t}\int\limits_a^b x(x-a)^{\alpha}(b-x)^{\gamma}\mathrm{d}x\tag{15.4.100}$$

Always with the same change of variable, we get:

$$\begin{aligned}\mu&=\int\limits_a^b xP_{a,b}(x)\mathrm{d}x=\frac{1}{\int\limits_0^1 t^{\alpha}(1-t)^{\gamma}\mathrm{d}t}\int\limits_a^b(a+u(b-a))u^{\alpha}(1-u)^{\gamma}\mathrm{d}u\\&=a+\frac{b-a}{\int\limits_0^1 t^{\alpha}(1-t)^{\gamma}\mathrm{d}t}\int\limits_a^b u^{\alpha+1}(1-u)^{\gamma}\mathrm{d}u=a+(b-a)\frac{B(\alpha+2,\gamma+1)}{B(\alpha+1,\gamma+1)}\end{aligned}\tag{15.4.101}$$

But we know from the study of the section on Differential and Integral Calculus (see page 1574) that:

$$\begin{aligned}\frac{B(\alpha+2,\gamma+1)}{B(\alpha+1,\gamma+1)}&=\frac{\Gamma(\alpha+2)\Gamma(\gamma+1)}{\Gamma(\alpha+\gamma+3)}\frac{\Gamma(\alpha+\gamma+2)}{\Gamma(\alpha+1)\Gamma(\gamma+1)}\\&=\frac{\Gamma(\alpha+2)}{\Gamma(\alpha+\gamma+3)}\frac{\Gamma(\alpha+\gamma+2)}{\Gamma(\alpha+1)}=\frac{(\alpha+1)(\alpha)\Gamma(\alpha)}{\Gamma(\alpha+\gamma+3)}\frac{\Gamma(\alpha+\gamma+2)}{(\alpha)\Gamma(\alpha)}\\&=\frac{(\alpha+1)(\alpha)\Gamma(\alpha)}{(\alpha+\gamma+2)\Gamma(\alpha+\gamma+2)}\frac{\Gamma(\alpha+\gamma+2)}{(\alpha)\Gamma(\alpha)}=\frac{\alpha+1}{\alpha+\gamma+2}\end{aligned}\tag{15.4.102}$$

Thus:

$$\mu=a+(b-a)\frac{\alpha+1}{\alpha+\gamma+2}\tag{15.4.103}$$

Let us calculate now the variance using the Huygens relation that we have proved in the section Statistics (see page 593):

$$\mathrm{V}(X)=\mathrm{E}(X^2)-\mathrm{E}(X)^2 \tag{15.4.104}$$

First we calculate $\mathrm{E}(X^2)$:

$$\mathrm{E}(X^2)=\int\limits_a^b x^2 P_{a,b}(x)\mathrm{d}x=\frac{1}{(b-a)^{\alpha+\gamma+1}\int\limits_0^1 t^{\alpha}(1-t)^{\gamma}\mathrm{d}t}\int\limits_a^b x^2(x-a)^{\alpha}(b-x)^{\gamma}\mathrm{d}x \tag{15.4.105}$$

Always with the same change of variable, we get:

$$\begin{aligned}\mathrm{E}(X^2)&=\int\limits_a^b x^2 P_{a,b}(x)\mathrm{d}x=\frac{1}{\int\limits_0^1 t^{\alpha}(1-t)^{\gamma}\mathrm{d}t}\int\limits_a^b (a^2+u^2(b-a)^2+2au(b-a))u^{\alpha}(1-u)^{\gamma}\mathrm{d}u\\ &=a^2+\frac{(b-a)^2}{B(\alpha+1,\gamma+1)}B(\alpha+3,\gamma+1)+\underbrace{\frac{2a(b-a)}{\int\limits_0^1 t^{\alpha}(1-t)^{\gamma}\mathrm{d}t}\int\limits_0^1 u^(\alpha+1)(1-u)^{\gamma}du}_{=2a(b-a)\frac{\alpha+1}{\alpha+\gamma+2}\text{ (see previous dev.)}}\\ &=a^2+2a(b-a)\frac{\alpha+1}{\alpha+\gamma+2}+(b-a)^2\frac{B(\alpha+3,\gamma+1))}{B(\alpha+1,\gamma+1)}\end{aligned} \tag{15.4.106}$$

But we have:

$$\frac{B(\alpha+3,\gamma+1))}{B(\alpha+1,\gamma+1)}=\frac{\Gamma(\alpha+3)\Gamma(\gamma+1)}{\Gamma(\alpha+\gamma+4)}\frac{\Gamma(\alpha+\gamma+2)}{\Gamma(\alpha+1)\Gamma(\gamma+1)}=\frac{(\alpha+2)(\alpha+1)}{(\alpha+\gamma+3)(\alpha+\gamma+2)} \tag{15.4.107}$$

Thus:

$$\mathrm{E}(X^2)=a^2+2a(b-a)\frac{\alpha+1}{\alpha+\gamma+2}+(b-a)^2\frac{(\alpha+2)(\alpha+1)}{(\alpha+\gamma+3)(\alpha+\gamma+2)} \tag{15.4.108}$$

To finish:

$$\begin{aligned}\mathrm{V}(X)&=\sigma^2=\mathrm{E}(X^2)-\mathrm{E}(X)^2=\mathrm{E}(X^2)-\mu^2\\ &=a^2+2a(b-a)\frac{\alpha+1}{\alpha+\gamma+2}+(b-a)^2\frac{\alpha+2}{\alpha+\gamma+3}\frac{\alpha+1}{\alpha+\gamma+2}-\left(a+(b-a)\frac{\alpha+1}{\alpha+\gamma+2}\right)^2\\ &=(b-a)^2\frac{\alpha+1}{\alpha+\gamma+2}\left(\frac{\alpha+2}{\alpha+\gamma+3}-\frac{\alpha+1}{\alpha+\gamma+2}\right)=\frac{(b-a)^2(\alpha+1)(\gamma+1)}{(\alpha+\gamma+2)^2(\alpha+\gamma+3)}\end{aligned} \tag{15.4.109}$$

Let us now calculate the modal M_0 value of this distribution (for recall of the section Statistics the modal value is by definition the overall maximum of the distribution function):

$$P_{a,b}(x)=\frac{(x-a)^{\alpha}(b-x)^{\gamma}}{(b-a)^{\alpha+\gamma+1}\int\limits_{0}^{1}t^{\alpha}(1-t)^{\gamma}\mathrm{d}t}1_{[a,b]} \tag{15.4.110}$$

To get the expression of the modal value we just need to solve the equation:

$$\frac{\mathrm{d}P_{a,b}(x)}{\mathrm{d}x}=0 \tag{15.4.111}$$

After derivation, we get:

$$\alpha(x-a)^{\alpha-1}(b-x)^{\gamma}-\gamma(x-a)^{\alpha}(b-x)^{\gamma-1}=0 \tag{15.4.112}$$

dividing by $(x-a)^{\alpha-1}(b-x)^{\gamma-1}$ we get:

$$\alpha(b-x)-\gamma(x-a)=0 \tag{15.4.113}$$

that is to say:

$$x=M_0=\frac{\alpha b+\gamma a}{\alpha+\gamma} \tag{15.4.114}$$

Now the reader will have noticed that the value a is the smallest value and the greatest is b. In between there is therefore the modal value M.

In project management a corresponds to the optimistic duration (or cost) t_O and b the duration t_P duration (or cost) of a task.

$$\alpha=2+\sqrt{2},\gamma=2-\sqrt{2} \tag{15.4.115}$$

or:

$$\alpha=2-\sqrt{2},\gamma=2+\sqrt{2} \tag{15.4.116}$$

This means that we have:

$$M_0=\frac{\alpha b+\gamma a}{\alpha+\gamma}=\frac{(2+\sqrt{2})b+(2-\sqrt{2})a}{2+\sqrt{2}+2-\sqrt{2}}=\frac{2(a+b)+\sqrt{2}(b-a)}{4} \tag{15.4.117}$$

as well as:

$$\begin{aligned}\mathrm{V}(X)&=\frac{(b-a)^2(\alpha+1)(\gamma+1)}{(\alpha+\gamma+2)^2(\alpha+\gamma+3)}=\frac{(b-a)^2(2+\sqrt{2}+1)(2-\sqrt{2}+1)}{(2+\sqrt{2}+2-\sqrt{2}+2)^2(2+\sqrt{2}+2-\sqrt{2}+3)}\\&=\frac{(b-a)^2(3+\sqrt{2})(3-\sqrt{2})}{6^27}=\frac{(b-a)^2(9-2)}{36\cdot 7}=\frac{(b-a)^2(3+\sqrt{2})(3-\sqrt{2})}{6^27}\\&=\frac{(b-a)^2(9-2)}{6^2}\end{aligned} \tag{15.4.118}$$

And finally:

$$\begin{aligned}\mathrm{E}(X)&=a+(b-a)\dfrac{\alpha+1}{\alpha+\gamma+2}=\dfrac{6a+(b-a)(2+\sqrt{2}+1)}{6}\\&=\dfrac{6a+2b+\sqrt{2}b+b-2a-\sqrt{2}a-a}{6}=\dfrac{5a+a+b+b+\sqrt{2}b+b-2a-\sqrt{2}a-a}{6}\\&=\dfrac{a+b}{6}+\dfrac{2a+2b+(b-a)\sqrt{2}}{6}=\dfrac{a+b}{6}+\dfrac{4M_0}{6}=\dfrac{a+4M_0+b}{6}\end{aligned} \tag{15.4.119}$$

This gives us to resume the following relations:

$$\boxed{\begin{aligned}M_0&=\dfrac{2(a+b)+\sqrt{2}(b-a)}{4}\\\mathrm{V}(X)&=\dfrac{(b-a)^2}{6^2}\\\mathrm{E}(X)&=\dfrac{a+4M_0+b}{6}\end{aligned}} \tag{15.4.120}$$

that is assimilated in some textbooks to the Beta-PERT distribution with a scale parameter λ equal to 4 as a more general derivation (that we will not present here) gives :

$$\mathrm{E}(X)=\dfrac{a+\lambda M_0+b}{\lambda+2} \tag{15.4.121}$$

Remark

The last two expressions of the variance and expected mean are the ones you can find in any project management book (of course without proof...) such as the famous PMBOK (Project Management Body of Knowledge). Unfortunately, there are two errors in this project management reference book as far as we know (the first error was corrected at the beginning of the years 12010 according to holocene calendar). Indeed, the first error is that the modal value is said to be imposed by the project manager but as we know, in fact it should be calculated with the first of the three above relations using the pessimistic and optimistic task duration (or cost)!!! The second mistake is that the PMBOK says that the modal value is always smaller than the expected mean. Which is obviously wrong! To check this we can put for example $a=0$:

$$\begin{aligned}M_0&=\dfrac{2(a+b)+\sqrt{2}(b-a)}{4}=\dfrac{2b+\sqrt{2}b}{4}\\\mathrm{E}(X)&=\dfrac{a+4M_0+b}{6}=\dfrac{4\dfrac{2(a+b)+\sqrt{2}(b-a)}{4}=\dfrac{2b+\sqrt{2}b}{4}+b}{6}=\dfrac{3b+\sqrt{2}b}{6}\end{aligned} \tag{15.4.122}$$

and simplifying by b we immediately see that:

$$\dfrac{2+\sqrt{2}}{4}\geq\dfrac{3+\sqrt{2}}{6}\Leftrightarrow 0.853\geq 0.753\Rightarrow M_0>\mathrm{E}(X) \tag{15.4.123}$$

It is important to know that exactly the same developments made so far are valid with the following formulation of the 4-parameter beta distribution (which will also be very useful in the section of Industrial Engineering for reliability calculations):

$$\boxed{P_{a,b}(x)=\frac{(x-a)^{\alpha-1}(x-a)^{\gamma-1}}{(b-a)^{\alpha+\gamma+1}\int\limits_0^1 t^{\alpha-1}(1-t)^{\gamma-1}\mathrm{d}t}=\frac{(x-a)^{\alpha-1}(x-a)^{\gamma-1}}{(b-a)^{\alpha+\gamma+1}B(\alpha,\gamma)}} \tag{15.4.124}$$

The latter formulation is that available in spreadsheets softwares such as Microsoft® Excel® version of 11.8346 for example using the `BETADIST()` function or its inverse using the `BETAINV()` function.

We then get for the modal value, the variance and expected mean (we can detail again all calculations if some readers want to...):

$$\begin{aligned}M_0&=\frac{\alpha a-b+\gamma a-a}{\alpha+\gamma-2}\\ \mathrm{V}(X)&=\frac{(b-a)^2\alpha\gamma}{(\alpha+\gamma)^2(\alpha+\gamma-1)}\\ \mathrm{E}(X)&=a+(b-a)\frac{\alpha}{\alpha+\gamma}\end{aligned} \tag{15.4.125}$$

But to get the P.M.I. result you must use in spreadsheet softwares (if you put this two values in the latter relations you can quickly check that you find back the classic modal value, variance and expected mean of the P.M.I.):

$$\alpha=3+\sqrt{2},\gamma=3-\sqrt{2} \tag{15.4.126}$$

To calculate the probability or quantile in the English version of Microsoft® Excel® 14.0.7153:

- `BETA.DIST(quantile,`$3+\sqrt{2}$`,`$3-\sqrt{2}$`,TRUE,optimistic value,pessimistic value)`
- `BETA.INV(probability,`$3+\sqrt{2}$`,`$3-\sqrt{2}$`,optimistic value,pessimistic value)`

✍**Example:**

Plot of the distribution and cumulative function for the beta law of parameters $(t_o, t_v, t_p) = (0, 1, 3)$:

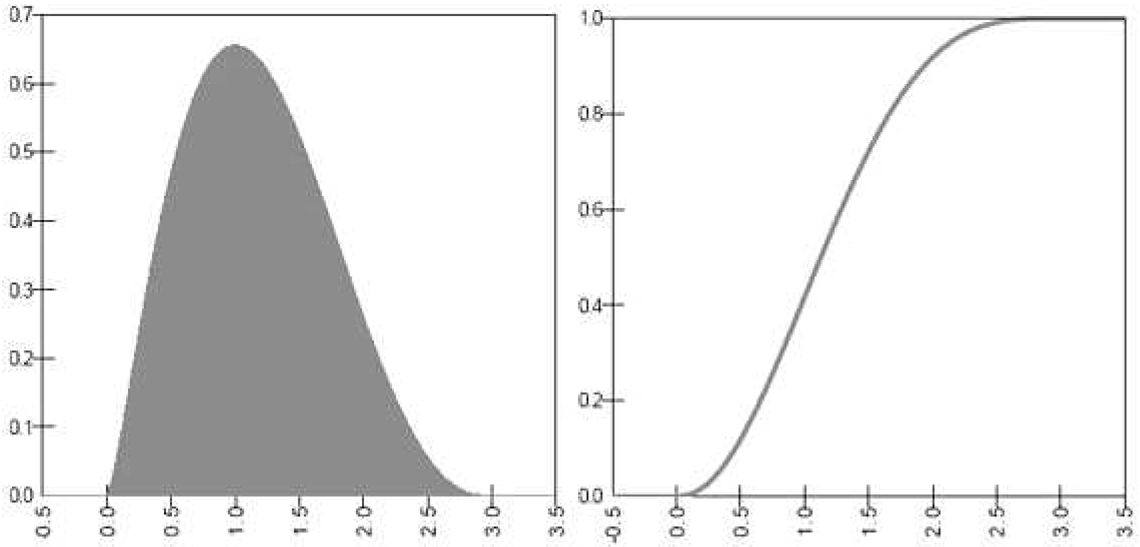


Figure 15.288: Distribution and cumulative function of the beta law

We also define the "action risk" by the ratio (whose interpretation is left to the project managers and customers...):

$$R(T_i) = \frac{t_p - t_o}{t_p} \tag{15.4.127}$$

By the NASA Probabilistic analysis of cost and/or schedule estimates is required (in 12014 according to holocene calendar) for tightly coupled programs and single-project programs (regardless of life-cycle cost), and projects with a LCCE (Life-Cycle Cost Estimate) greater than \$ 250 million. When the probabilistic analysis is developed for only one parameter (i.e., cost or schedule) or when generally referring to a probabilistic assessment of the level of confidence of achieving a specific goal, the analysis is referred to merely as a "confidence level". When the probabilistic analysis is developed to measure the likelihood of meeting both cost and schedule, the analysis is referred to as a "joint cost and schedule confidence level" (JCL). A JCL is defined as the probability that actual cost and schedule will be equal to or less than the targeted cost and schedule. For example, a 70% JCL is the point on the joint cost and schedule probability distribution curve where there is a 70% percent probability that the project or program will be completed at or lower than the estimated cost and on or before the estimated schedule.

15.4.2.2 Project planning variance reduction

As we know a lot of projects budgeting and time planning lack in accuracy because there is always an implicit variance to each task on the critical path and most project manager don't have the time to do detailed planning as for most board committee (and even most project managers!!!) there is no reason that the total variance of a project decrease if we increase the details of the planification (that is to say: we detail a maximum of tasks).

We suppose here that we will work in the framework of softwares used by project management engineers and able to display for each task and early finish (t_o), expected finish (t_V) and late finish (t_p). Even if there are advanced risk modelization softwares doing this work in cutting edge companies (softwares like @Risk, Crystal Ball, Risky Project, etc), most project managers calculate these three point estimates based on a rule of thumb that is most of time of the type:

$$t_V=\mu \qquad t_o=\mu-b\mu \qquad t_p=\mu+b\mu \tag{15.4.128}$$

where b is a security factor that we will here name the "bias" of estimation.

Theorem 15.9. Under some assumptions relevant in real projects variance will decrease if the planning is detailed and that this variance tends to zero when the number of detailed tasks tends to infinity (I always try to explain without maths that this is obvious since if we do a task for every gesture and action in the project we forget less stuff than doing just a macroscopic planning).

For the proof let we have only 3 known assumptions (perhaps they are other hidden one but not identified so far...):

H1. When a project manager estimates a task duration their brain use a non-robust estimator that is the mean μ.

H2. Tasks duration are independent and therefore their covariance is zero.

H3. The project managers always have the same biased b (percentage) for all tasks [57].

Proof 15.9.1. Let us consider the case (that can be easily generalized with high-school maths) of a task of duration T of mean μ and standard-deviation:

$$\sigma_T=b\mu_T=bT \tag{15.4.129}$$

We take the special case where we detail this task of duration T into two critical (i.e. consecutive) sub tasks (still with same bias) of non-zero duration such that:

$$T=T_1+T_2=w_1T+w_2T \tag{15.4.130}$$

where obviously $w_1+w_2=1$. This is quite a natural process for most project managers until now so nothing new...

---

[57] In practice I see that most project managers have a significant bias much greater than 20% in task duration and budget estimation

But now let us calculate the standard deviation of the sum under the assumption of independence and stability:

$$\begin{aligned}\sigma_{T_1+T_2} &= \sqrt{\sigma_{T_1}^2+\sigma_{T_2}^2}=\sqrt{(b\mu_{T_1})^2+(b\mu_{T_1})^2}=b\sqrt{mu_{T_1}^2+mu_{T_2}^2}\\ &= b\sqrt{(w_1T)^2+(w_2T)^2}=bT\sqrt{w_1^2+w_2^2}\end{aligned}\tag{15.4.131}$$

So we see that as if $0<w_i<1$ then we have:

$$\sigma_T<\sigma_{T_1+T_2}\tag{15.4.132}$$

as:

$$bT<bT\sqrt{w_1^2+w_2^2}\tag{15.4.133}$$

Because:

$$\sqrt{w_1^2+w_2^2}<1\tag{15.4.134}$$

So as we see, the deeper the granularity of a project planning is, the lower is the variation the latter. That is:

$$i\rightarrow+\infty:\quad \sigma_{\sum T_i}=bt\sqrt{\sum_{i=1}^{+\infty}w_i^2}\rightarrow 0\tag{15.4.135}$$

This result then is like the "Hofstadter's law", that is a self-referential time-related adage, coined by Douglas Hofstadter and named after him that states: «*It always takes a project longer than you expect, even when you take into account Hofstadter's Law*».

■ Q.E.D.

It is quite easy to notice that in the best case, for every $i>1$ we minimize $\sigma_{\sum T_i}$ if all w_i are equal to:

$$w_i=\frac{1}{i}\tag{15.4.136}$$

For example with $i=2$ in the best case we can reduce the initial variance of almost 30%, if $i=3$ of almost 42%, if $i=4$ of exactly 50%, if $i=5$ of almost 55% and so on. The variance reduction relation is simple given in the general best case by:

$$\boxed{\text{VR}=1-\sqrt{\sum_i\left(\frac{1}{i}\right)^2}=1-\sqrt{\frac{1}{i}}}\tag{15.4.137}$$

This result corroborates the fact the division of labour increases production. In other words, when the tasks involved with producing a good or service are divided and subdivided, workers and businesses can produce a greater quantity of output. In his observations of pin factories, Adam Smith observed that one worker alone might make 20 pins in a day, but that a small business of 10 workers (some of whom would need to do two or three of the 18 tasks involved with pin-making), could make $48,000$ pins in a day. How can a group of workers, each specializing in certain tasks, produce so much more than the same number of workers who try to produce the entire good or service by themselves? Smith offered three reasons:

1. First, specialization in a particular small job allows workers to focus on the parts of the production process where they have an advantage (in later chapters, we will develop this idea by discussing comparative advantage). People have different skills, talents, and interests, so they will be better at some jobs than at others. The particular advantages may be based on educational choices, which are in turn shaped by interests and talents. Only those with medical degrees qualify to become doctors, for instance. For some goods, specialization will be affected by geography - it is easier to be a wheat farmer in North Dakota than in Florida, but easier to run a tourist hotel in Florida than in North Dakota. If you live in or near a big city, it is easier to attract enough customers to operate a successful dry cleaning business or movie theatre than if you live in a sparsely populated rural area. Whatever the reason, if people specialize in the production of what they do best, they will be more productive than if they produce a combination of things, some of which they are good at and some of which they are not.

2. Second, workers who specialize in certain tasks often learn to produce more quickly and with higher quality. This pattern holds true for many workers, including assembly line labourers who build cars, stylists who cut hair, and doctors who perform heart surgery. In fact, specialized workers often know their jobs well enough to suggest innovative ways to do their work faster and better.

   A similar pattern often operates within businesses. In many cases, a business that focuses on one or a few products (sometimes named its "core competency") is more successful than firms that tries to make a wide range of products.

3. Third, specialization allows businesses to take advantage of economies of scale, which means that for many goods, as the level of production increases, the average cost of producing each individual unit declines. For example, if a factory produces only 100 cars per year, each car will be quite expensive to make on average. However, if a factory produces $50,000$ cars each year, then it can set up an assembly line with huge machines and workers performing specialized tasks, and the average cost of production per car will be lower. The ultimate result of workers who can focus on their preferences and talents, learn to do their specialized jobs better, and work in larger organizations is that society as a whole can produce and consume far more than if each person tried to produce all of their own goods and services. The division and specialization of labour has been a force against the problem of scarcity.

15.4.2.3 Process Reliability

Almost every people and even child know what why process are important and how to draw them using various normalized symbols (like the most famous one that are ISO 5807:85 or BPMN/ISO 19510:2013) as shown the figure below:

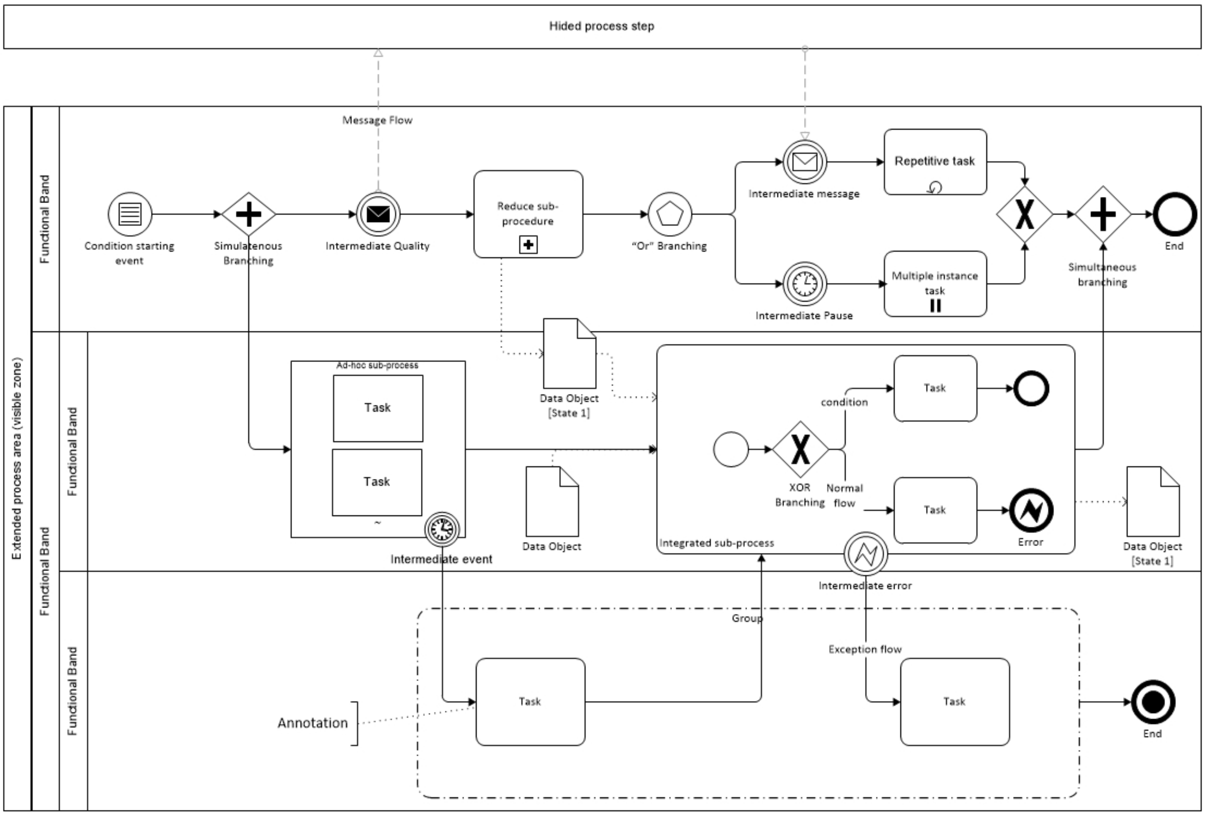

Hided process step
Message Flow
Extended process area (visible zone)
Functional Band
Condition starting event
Simultaneous Branching
Intermediate Quality
Reduce sub-procedure
"Or" Branching
Intermediate message
Repetitive task
Multiple instance task
Intermediate Pause
Simultaneous branching
End
Functional Band
Functional Band
Ad-hoc sub-process
Task
Task
Data Object [State 1]
condition
Task
XOR Branching
Normal flow
Task
Error
Integrated sub-process
Data Object
Intermediate event
Data Object [State 1]
Intermediate error
Functional Band
Group
Exception flow
Task
Task
Annotation
End

Figure 15.289: BPMN process representative diagram with Microsoft Visio 2010

But in some high level and cutting edge industries just drawing processes is not enough. We consider mainly three situations about processes:

- An organization has processes, people respect them or not and there is not statistical analysis about their usage.

  There is nothing to say about this as it is what the huge majority of undergraduate project managers and business analysts do!

- An organization has processes, people respect them or not and there is a Six Sigma statistical analysis about the capability of the processes or their subparts (see section Industrial Engineering page 8683).

  For the capability analysis we will study and prove in details in the section of Industrial Engineering that we can (or have to) calculate the following indicators:

$$\boxed{C_p=\frac{\text{USL}-\text{LSL}}{\sigma_{\text{ST}}}\qquad C_{pk}=\left\{\frac{\text{USL}-\mu}{3\sigma_{\text{ST}}},\frac{\mu-\text{LSL}}{3\sigma_{\text{ST}}}\right\}\qquad C_{pm}\cong\frac{\text{USL}-\text{LSL}}{6\sqrt{\sigma_{\text{ST}}^2+(m-T)^2}}} \tag{15.4.138}$$

and:

$$\boxed{P_p = \frac{\text{USL}-\text{LSL}}{\sigma_{\text{LT}}} \qquad P_{pk} = \left\{\frac{\text{USL}-\mu}{3\sigma_{\text{LT}}}, \frac{\mu-\text{LSL}}{3\sigma_{\text{LT}}}\right\} \qquad P_{pm} \cong \frac{\text{USL}-\text{LSL}}{6\sqrt{\sigma_{\text{LT}}^2+(m-T)^2}}} \tag{15.4.139}$$

associated with various run charts (see section Industrial Engineering page 8935) to do real scientific robust process management analysis.

- An organization has processes, people respect them or not and there is a reliability analysis of the processes or their subparts (see section Industrial Engineering page 8683).

  This is typically used in Nuclear Engineering or high level Pharmaceutical products quality tracking.

  The mathematical relations used in this case are those proven in the section of Industrial Engineering during our study of Topological Systems in Preventive Maintenance (see section Industrial Engineering page 8716) also associated with Monte Carlo simulations (see section Numerical Methods page 6708). That is to say:

  - The reliability (probability of operation) of a pure series topology:

$$\boxed{R_S = 1 - F_S = \prod_{i=1}^{n} R_i} \tag{15.4.140}$$

  - The reliability (probability of operation) of a pure parallel topology:

$$\boxed{R_P = 1 - \prod_{i=1}^{n} F_i = 1 - \prod_{i=1}^{n}(1-R_i)} \tag{15.4.141}$$

  - The reliability of a k/n topology:

$$\boxed{R_{k/n} = \sum_{x=k}^{n} C_x^n R^x (1-R)^{n-x}} \tag{15.4.142}$$

  - The reliability of a series topologies in parallels:

$$\boxed{R_{S/P} = 1 - \prod_{j=1}^{m}(1-R_j) = \prod_{j=1}^{m}\left(1 - \prod_{i=1}^{n} R_{ij}\right)} \tag{15.4.143}$$

  - The reliability of parallel topologies in series:

$$\boxed{R_{P/S} = \prod_{i=1}^{n} R_i = \prod_{i=1}^{n}\left(1 - \prod_{j=1}^{m}(1-R_{ij})\right)} \tag{15.4.144}$$

  - The reliability of bridged topology in series:

$$\boxed{\begin{aligned} R_B = &(1-R_5)(1-(1-R_1R_2)(1-R_3R_4)) \\ &+ R_5((1-(1-R_1)(1-R_3))(1-(1-R_2)(1-R_4))) \end{aligned}} \tag{15.4.145}$$

Do these calculations is quite easy even with a spreadsheet software. But obviously it is strongly recommended to use specialized softwares like the add-in Process Simulator of ProModel® for Microsoft Visio.

15.4.2.4 Divide and rule?

Divide and rule (or divide and conquer) in management is gaining and maintaining a better control by breaking up groups into pieces that individually have less incompatibilities. The concept refers to a strategy that breaks up existing structures, and especially prevents individuals in group to disagree, fight together or and generate issues and poor efficiency.

We have proved in the sectionof Statistics (see page 644) that for a group of items (individuals) having a unique common property[58] (characteristic) its entropy was given by:

$$S_j(x)=\frac{N\log_{10}(N)-\sum_{i=1}^{n} f_i\log_{10}(f_i)}{N} \tag{15.4.146}$$

Now let us consider the case of a population of size P with only one property (so that we can split the population in two distinct groups) in proportion:

$$w_1+w_2=1=100\% \tag{15.4.147}$$

and therefore $w_2=1-w_1$ and $w_1P+w_2P=P$. So the total entropy if we consider the two possible groups the first having a proportion p_1^1 of w_1 and p_1^2 of w_2 (and therefore the second group has $p_2^1 = 1 - p_1^1$ and $p_2^2 = 1 - p_1^2$):

---

[58] For example a property: Technical skill with a binary value of the type "Yes"/"No"

$$\begin{aligned}
S_{\text{tot}} = S_1 + S_2 &= \frac{N_1 \log_{10}(N_1) - \displaystyle\sum_{i=1}^{2} f_i \log_{10}(f_i)}{N_1} + \frac{N_2 \log_{10}(N_2) - \displaystyle\sum_{i=1}^{2} f_i \log_{10}(f_i)}{N_2} \\
&= \frac{(p_1^1 w_1 + p_1^2 w_2) P \log_{10}((p_1^1 w_1 + p_1^2 w_2)P) - \displaystyle\sum_{i=1}^{2} f_i \log_{10}(f_i)}{(p_1^1 w_1 + p_1^2 w_2)P} \\
&+ \frac{(p_2^1 w_1 + p_2^2 w_2) P \log_{10}((p_1^1 w_1 + p_1^2 w_2)P) - \displaystyle\sum_{i=1}^{2} f_i \log_{10}(f_i)}{(p_2^1 w_1 + p_2^2 w_2)P} \\
&= \frac{(p_1^1 w_1 + p_1^2 w_2) P \log_{10}((p_1^1 w_1 + p_1^2 w_2)P) - \left[p_1^1 w_1 P \log(p_1^1 w_1 P) + p_1^2 w_2 P \log(p_1^2 w_2 P)\right]}{(p_1^1 w_1 + p_1^2 w_2)P} \\
&+ \frac{(p_2^1 w_1 + p_2^2 w_2) P \log_{10}((p_2^1 w_1 + p_2^2 w_2)P) - \left[p_2^1 w_1 P \log(p_2^1 w_1 P) + p_2^2 w_2 P \log(p_2^2 w_2 P)\right]}{(p_2^1 w_1 + p_2^2 w_2)P}
\end{aligned} \tag{15.4.148}$$

A first simplification by P gives:

$$\begin{aligned}
S_{\text{tot}} = S_1 + S_2 = &\frac{(p_1^1 w_1 + p_1^2 w_2) \log_{10}((p_1^1 w_1 + p_1^2 w_2)P) - \left[p_1^1 w_1 \log(p_1^1 w_1 P) + p_1^2 w_2 \log(p_1^2 w_2 P)\right]}{p_1^1 w_1 + p_1^2 w_2} \\
&+ \frac{(p_2^1 w_1 + p_2^2 w_2) \log_{10}((p_2^1 w_1 + p_2^2 w_2)P) - \left[p_2^1 w_1 \log(p_2^1 w_1 P) + p_2^2 w_2 \log(p_2^2 w_2 P)\right]}{p_2^1 w_1 + p_2^2 w_2}
\end{aligned} \tag{15.4.149}$$

We can work in percentage of the population P. So if we put $P = 100\% = 1$, we get:

$$\begin{aligned}
S_{\text{tot}} = S_1 + S_2 = &\frac{(p_1^1 w_1 + p_1^2 w_2) \log_{10}(p_1^1 w_1 + p_1^2 w_2) - \left[p_1^1 w_1 \log(p_1^1 w_1) + p_1^2 w_2 \log(p_1^2 w_2)\right]}{p_1^1 w_1 + p_1^2 w_2} \\
&+ \frac{(p_2^1 w_1 + p_2^2 w_2) \log_{10}(p_2^1 w_1 + p_2^2 w_2) - \left[p_2^1 w_1 \log(p_2^1 w_1) + p_2^2 w_2 \log(p_2^2 w_2)\right]}{p_2^1 w_1 + p_2^2 w_2}
\end{aligned} \tag{15.4.150}$$

So we have a function of three variables $S_{\text{tot}}(w_1, p_1^1, p_1^2)$ as for recall:

$$\begin{aligned}
w_2 &= 1 - w_1 \\
p_2^1 &= 1 - p_1^1 \\
p_2^2 &= 1 - p_1^2
\end{aligned} \tag{15.4.151}$$

If we fix one of them we have a function of two variable that we can easily plot and study. So let us fix w_1.

First case we fix $w_1 = 0.5$ and we use the following Maple™ 4.00b command lines:

```
>w_1:=0.5;w_2:=1-w_1;
>p_21:=1-p_11;p_22:=1-p_12;
```

```
>S_tot:=((p_11*w_1+p_12*w_2)*log10(p_11*w_1+p_12*w_2)-(p_11*w_1*log10(p_11*w_1)
+p_12*w_2*log10(p_12*w_2)))/(p_11*w_1+p_12*w_2)+((p_21*w_1+p_22*w_2)*
log10(p_21*w_1+p_22*w_2)-(p_21*w_1*log10(p_21*w_1)+p_22*w_2*log10(p_22*w_2)))/
(p_21*w_1+p_22*w_2);
>plot3d(S_tot,p_11=0.1..0.9,p_12=0.1..0.9,filled=true,axes=boxed);
```

That gives:

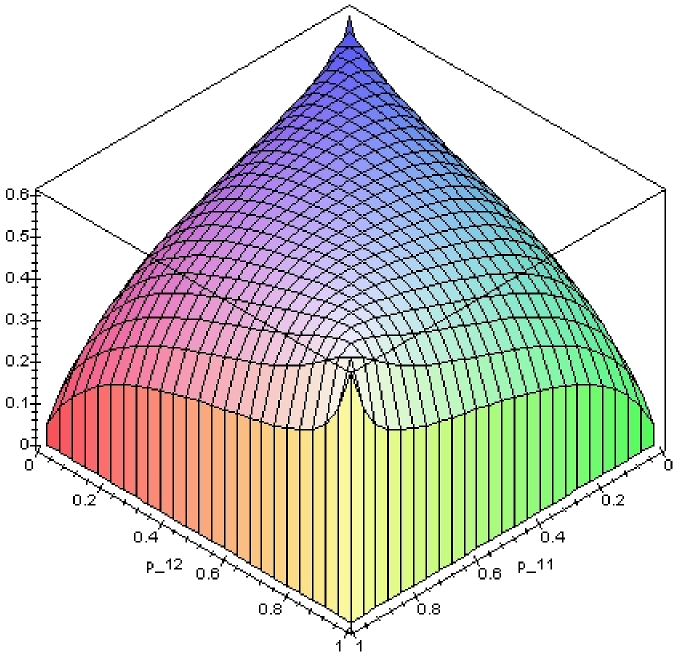


Figure 15.290: S_{tot} **Maple™ 4.00b with** $w_1 = 0.5$

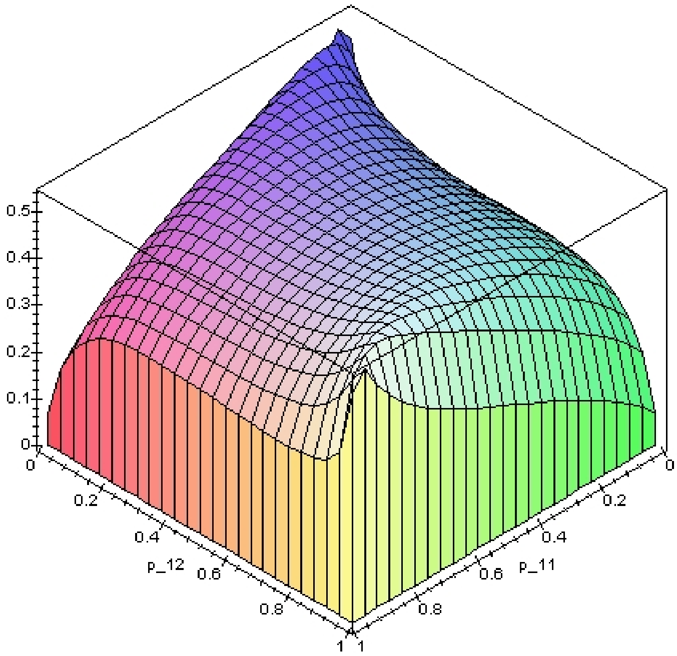


Figure 15.291: S_{tot} **Maple™ 4.00b with** $w_1 = 0.25$

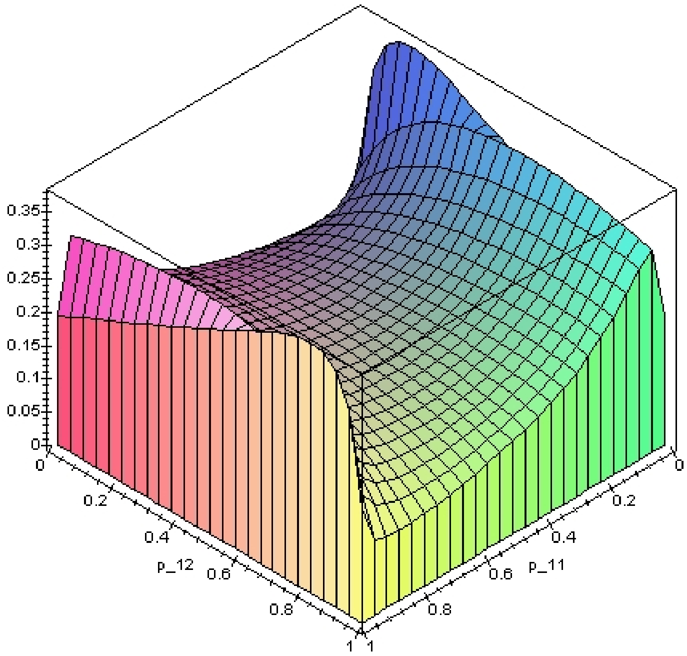


Figure 15.292: S_{tot} **Maple™ 4.00b with** $w_1 = 0.95$

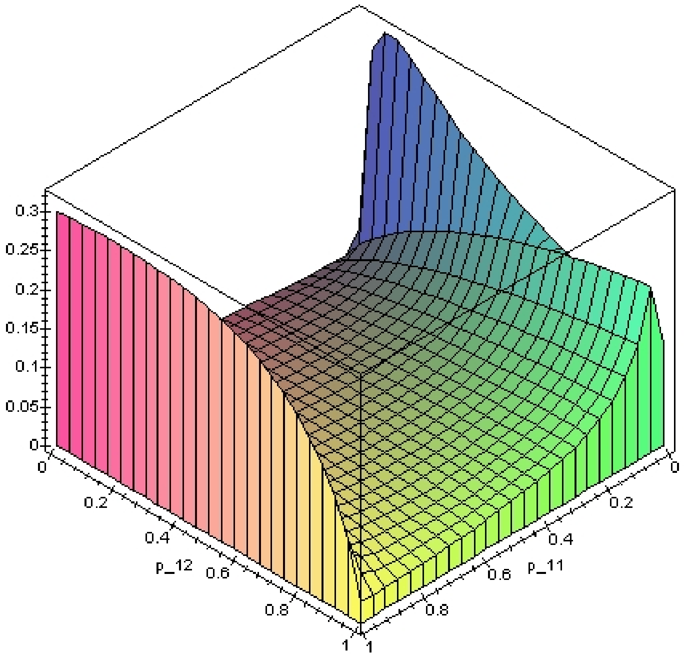


Figure 15.293: S_{tot} **Maple™ 4.00b with** $w_1 = 0.99$

With the above plots we can conclude that:

- If the both groups or individuals are almost in the same proportions, it is a better strategy to divide them in two groups as it minimize entropy. Divide and rule is better!
- If one of the group is in huge majority, it is better to put everybody together to minimize the entropy. Merge and rule is better!

15.4.3 Lean Management (Six Sigma Process)

As we saw it in the section Statistics (see page 692), the Normal distribution is a function whose shapes looks like a "bell" curve given by:

$$\boxed{P(k,\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(k-\mu)^2}{2\sigma^2}}} \tag{15.4.152}$$

whose standard deviations are used to give the cumulative probability interval to lie within these limits centered on the average as shown below:

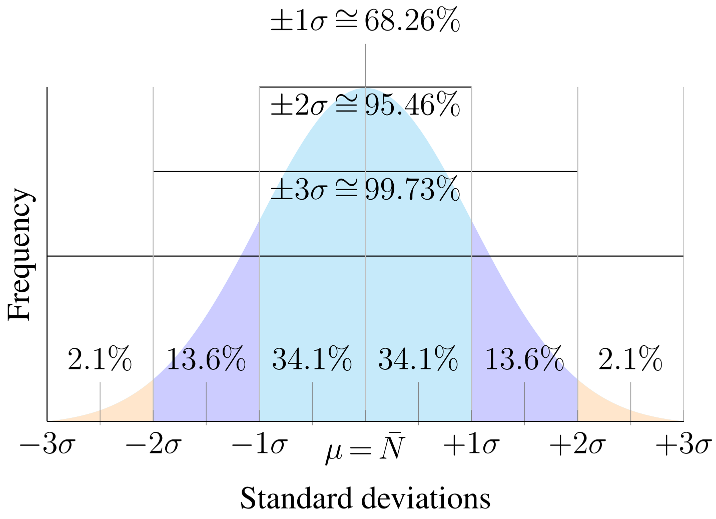


Figure 15.294: Sigma intervals for the Normal distribution

This being recalled, we also presented in the section of Industrial Engineering the joint probabilities as part of Six Sigma analysis tools for a serial process/workflow chain.

In fact the mentioned processes/workflow are not necessarily industrial processes but can be treated under identical assumptions to arbitrary processes (administrative, procedures, checklists, etc.).

We had seen that the joint probability of such a process/workflow is named in Six Sigma "rolled throughput yield" (R.T.Y.) and was given by (see section Probabilities page 480):

$$\boxed{\text{RTY}=P\left(\bigcap_i P_i\right)=\prod_i P_i} \tag{15.4.153}$$

For example, applying the above relation gives for a 4-step serial process ($i=4$) whose reliability is $P_i=90\%$ at each step (for more complex structures, remember that we sometimes talk about "weighted probability trees" or "topological systems"):

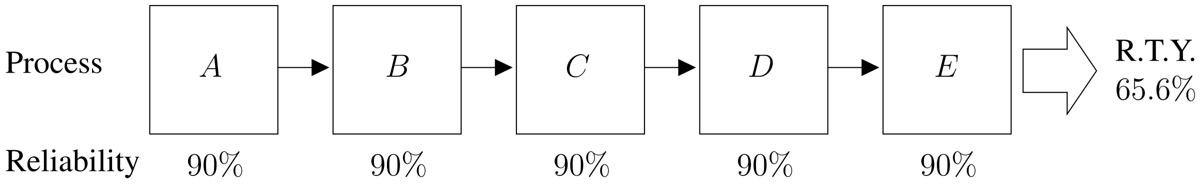


Figure 15.295: Serial/Process Workflow

We find ourselves in the final step with a 65.6% reliability that is to say a cumulative probability of default of 34.4% for the entire process (you can also imagine that these are sequential tasks in a project!).

Remark

The reader that is attentive will have noticed that the serial system is always less reliable than its least reliable component!!

Let us communicate again the Six Sigma table according to the worst situation of a not centered process with an average deviation of (thus to the right but we might take left and the results will be the same) relative to the target and with a unit standard-deviation and also with symmetrical USL and LSL (which restricts the scope):

C_p	C_{pk}	**Defectives (PPM)**	**Sigma Quality Level**	**Criteria**
0.5	0	501,350	1.5	Bad
0.6	0.1	382,572	1.8	
0.7	0.2	274,412	2.1	
0.8	0.3	184,108	2.4	
0.9	0.4	115,083	2.7	
1	0.5	66,810	3	
1.1	0.6	35,931	3.3	
1.2	0.7	17,865	3.6	
1.3	0.8	8,198	3.9	Limit
1.4	0.9	3,467	4.2	
1.5	1	1,350	4.5	
1.6	1.1	483	4.8	
1.7	1.2	159	5.1	
1.8	1.3	48	5.4	
1.9	1.4	13	5.7	
2	1.5	13.4	6	Excellent

Table 15.71: Defectives and Sigma quality level of a non-center process

where we proved in the section of Industrial Engineering (see page 8697) that the PPM values were given by:

$$\begin{aligned}\text{PPM}_{\text{tot}} &= \left(1-\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{\text{USL}} e^{-\frac{(x-\mu)^2}{2}}\mathrm{d}x\right)10^6+\left(\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{\text{LSL}} e^{-\frac{(x-\mu)^2}{2}}\mathrm{d}x\right)10^6\\ &= \left(1-\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{3C_p} e^{-\frac{(x-1.5\sigma)^2}{2}}\mathrm{d}x\right)10^6+\left(\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{-3(C_p+1)} e^{-\frac{(x-1.5\sigma)^2}{2}}\mathrm{d}x\right)10^6\\ &= \left(1-\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{3C_p} e^{-\frac{(x-1.5)^2}{2}}\mathrm{d}x\right)10^6+\left(\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{-3(C_p+1)} e^{-\frac{(x-1.5)^2}{2}}\mathrm{d}x\right)10^6\end{aligned} \tag{15.4.154}$$

Which gives for a level of quality of 3 sigma with Maple™ 4.00b:

```
>evalf((1-1/sqrt(2*Pi)*int(exp(-(x-1.5)^ 2/2),x=-infinity..(1*3))))
*1E6+evalf((1/sqrt(2*Pi)*int(exp(-(x-1.5)^ 2/2),x=-infinity..-(3*(1+1)))))*1E6;
```

This gives a value of $\sim 66,810$ that is to say the value of the 6th line in the above table. This corresponds to $\sim 6.68\%$ in terms of cumulative probability of non-quality ($66,810$ divided by 1 million and in percent) and thus respectively to a cumulative probability of 93.32% of quality.

We have the following table that can summarize some important values of the above table by using the Maple™ 4.00b command given above:

	$\pm 3\sigma$	$\pm 4\sigma$	$\pm 5\sigma$	$\pm 6\sigma$
Quality %	93.32	99.38	99.98	99.996
C_p	1	1.33	1.68	2
Judgement	Bad	Limit	Good	Excellent

Table 15.72: **Important values of the process control**

Under the assumption that each step in a process series follows the same law with the same moment and the same deviations from the target then we have for the R.T.Y.:

Step/Quality %	$\pm 3\sigma$	$\pm 4\sigma$	$\pm 5\sigma$	$\pm 5\sigma$
1	93.32	99.38	99.98	99.9996
7	61.63	95.73	99.84	99.9976
10	50.08	93.96	99.77	99.9966
20	25.08	88.29	99.54	99.9932
40	6.29	77.94	99.07	99.9864
60	1.58	68.81	98.61	99.9796
80	0.40	60.75	98.16	99.9728
100	0.10	53.64	97.70	99.966
...	...	...	...	...

Table 15.73: **Defectives and Sigma quality level of a non-center process**

where each line represents the throughput Rolled Yield (TRY) thus calculated:

$$\text{RTY} = P\left(\bigcap_i P_i\right) = \prod_i P_i \qquad (15.4.155)$$

where i is the number of process steps $(1,7,10,20,40)$ and where the probabilities are given by the first line of the before last table. Thus, with a quality level of $\pm 3\sigma$ and a deviation to the target of $\pm 1.5\sigma$ for a process of 20 steps identically distributed:

$$\text{RTY} = \prod_i P_i = \prod_{i=1}^{20} 93.32\% = 0.9332^{20} \cong 25.08\% \qquad (15.4.156)$$

This is lamentable! An industrial process in 20 steps of a complete product should have a minimum level of 4σ quality to be acceptable in terms of yield and quality.

Thus, the objective of Lean Six Sigma in a company is to increase the level of quality with maximum R.T.Y. for a given number of steps in a process.

Remark

In the case of parallel or mixed serial/parallel process, the reader is invited to read the section of Industrial Engineering at page 8743.

15.4.3.1 Pareto Analysis

The Pareto chart, as we shall see, is an easy technique (among others) to classify phenomena in order of importance. But the Pareto distribution is often used in business as a basis for stochastic simulation of random variables representing projects investment (about as often used as the triangular distribution, Beta distribution or log-normal distribution in modern industry).

Remember that a random variable is said by definition follow a Pareto distribution if its cumulative probability function is given by (see section Statistics page 746):

$$P(X \leq x) = 1 - \left(\frac{x_m}{x}\right)^k \qquad (15.4.157)$$

with $k \in \mathbb{R}_+$ and $x \geq x_m \geq 0$ (thus $x > 0$) and for the mean (average):

$$\mu = \frac{kx_m}{k-1} \qquad (15.4.158)$$

and for variance:

$$\sigma^2 = V(X) = \frac{kx_m^2}{k-2} - \left(\frac{kx_m}{k-1}\right)^2 = \frac{kx_m^2}{(k-1)^2(k-2)} \qquad (15.4.159)$$

To illustrate this type of representation (Pareto Diagram), we assume that a study to reorganize the LAMBDA company's sales network will lead the sales manager to be interested in the distribution of purchase of $55,074$ orders (bottom right value in the table below) received during a given year depending on the city where customers are domiciled (we used here the first 200 cities in the imaginary country this is why we have 200 rows in the table below).

To calculate the parameters of the Pareto distribution, the reader should refer to the section of Industrial Engineering (see page 8736) because in the present section we will have a qualitative approach as do managers, quantitative analysis being reserved mainly for engineers.

n_i	N_i	n_i	N_i	n_i	N_i	n_i	N_i	n_i	N_i
8965	8965	240	40027	125	46378	88	50916	55	53752
4555	13520	236	40263	124	46862	87	51003	54	53806
3069	16589	224	40487	123	46985	87	51090	54	53860
2336	18925	223	40710	123	47108	86	51176	54	53914
1085	25987	202	41772	118	47704	83	51597	48	54164
934	26921	202	41972	117	47821	81	51678	47	54211
843	27764	200	42172	115	47936	79	51757	44	54255
746	28510	189	42363	114	48050	78	51835	43	54298
722	29232	190	42553	112	48162	78	51913	42	54340
607	31828	173	43278	107	48597	74	52218	40	54503
539	32367	170	43448	106	48703	72	52290	39	54542
502	32868	161	43609	105	48808	72	52362	38	54580
462	33330	160	43769	102	48910	72	52434	37	54617
459	33790	158	43927	102	49012	71	52505	37	54654
422	34212	153	44080	102	49114	71	52576	36	54690
372	34584	153	44233	101	49215	71	52647	36	54726
362	34945	149	44382	100	49315	69	52716	34	54760
355	35300	149	44531	100	49415	68	52784	33	54793
347	35647	148	44679	99	49514	67	52581	32	54825
342	35989	147	44826	98	49612	67	52918	31	54586
338	36327	147	44973	98	49710	65	52983	28	54884
329	36656	146	45119	97	49807	64	53047	25	54909
314	36970	146	45265	97	49904	63	53110	22	54931
313	37283	146	45411	96	50000	63	53173	20	54951
310	37593	142	45553	95	50095	63	53236	20	54971
310	37903	140	45693	95	50190	60	53296	20	54991
300	38203	137	45830	94	50284	60	53356	19	55010
284	38755	136	45966	92	50376	58	53414	15	55025
268	38755	135	46101	92	50468	57	53471	13	55038
267	39022	131	46232	91	50559	57	53539	11	55049
265	39287	128	46360	91	50660	57	53528	11	55060
251	39538	127	46487	90	50740	56	53641	8	55068
249	39787	126	46613	88	50828	56	53967	6	55074

Table 15.74: **Dataset for Pareto Analysis**

where the 200 n_i values (number of orders from a city i) have been classified by decreasing values and cumulated in a column N_i.

The first city is characterized by $8,965$ purchase orders (which corresponds to 16.28% of the total orders), the second by $4,555$, so that the first two cities made $13,520$ purchase orders (that corresponds to 24.55% of the total orders), the first three cities, spent $16,589$ purchase orders (which corresponds to 30.12% of the total orders), etc. Another way to describe this phenomenon is to say: 0.5% of cities (ranked by decreasing value of the chosen criterion) made 16.28% of the total orders, 1% of cities have passed 24.55% of the total orders, etc.

These calculations are partly shown in the table below:

$i\%$	$N_i\%$	$i\%$	$N_i\%$	$i\%$	$N_i\%$	$i\%$	$N_i\%$	$i\%$	$N_i\%$
0.5	16.28	20.5	72.68	40.5	84.86	60.6	92.45	80.5	97.6
1	24.55	21	73.11	41	85.09	61	92.61	81	97.7
1.5	30.12	21.5	73.51	41.5	85.32	61.5	92.77	81.5	97.8
2	34.36	22	73.92	42	85.54	62	92.92	82	97.89
...	...	...	...	...	...	...	...	...	...
19	71.33	39	84.18	59	91.97	79	97.3	99	99.7
19.5	71.79	39.5	84.41	59.5	92.13	79.5	97.4	99.5	99.99
20	72.24	40	84.64	60	92.29	80	97.5	100	100

Table 15.75: **Aggregated and normalized data for Pareto analysis**

translated graphically (according to the traditions of use) below as a Pareto chart with any modern spreadsheet software:

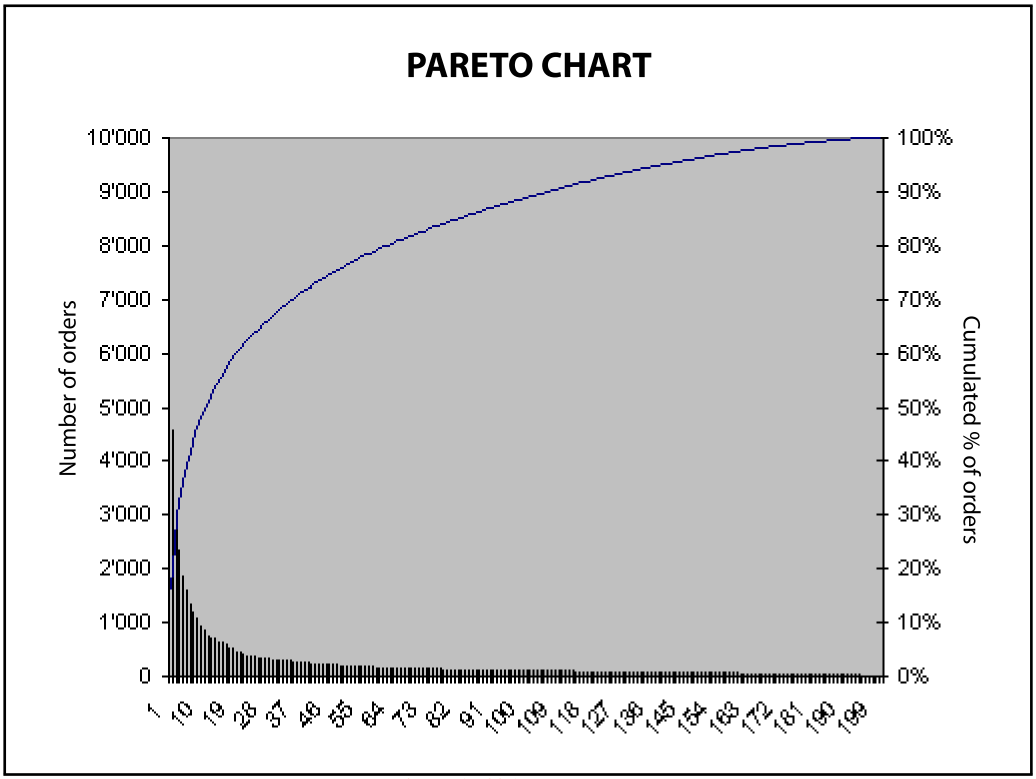


Figure 15.296: **Pareto chart**

or in a more relevant way using a "Lorenz diagram" also named sometimes "gain chart" (the horizontal axes with the numbers of cities is replaced by the cumulative% of towns as in the data table above):

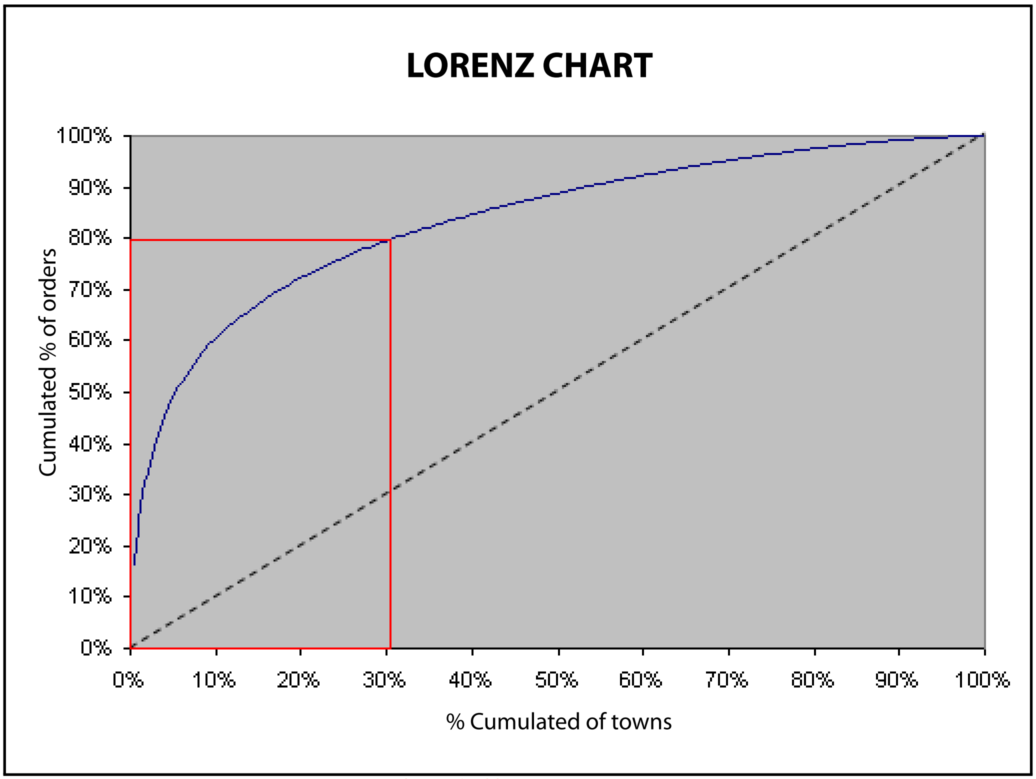


Figure 15.297: Lorenz chart

This analysis shows that a small number of elements is the root of most of the studied phenomenon (e.g. here 31% of cities generate 80% of purchase orders). This explains why it is one of the major techniques used in the field of "Total Quality Management" and in the analysis of the importance of the roots of a quality problem. It is also used by managers to structure the organization, especially to differentiate processes based on characteristic of demand (differentiated customer tracking according to its importance, for example) and when it is used to define 3 classes, this technique often takes the name "ABC method".

The dotted line in the Lorenz graph represents what we would have observed if the case of equal distribution of the studied phenomenon, that is to say if each city was characterized by the same number of purchase orders.

In general, the more a Lorenz curve is close to the perfect equality of the straight line and more the distribution of mass seen in the population is equal. Indeed, in this case, the majority (wages, wealth, income, etc.) is not concentrated only on a few items.

Remark

The presentation of this analysis was conducted by classifying the observations by decreasing values, but we could just as well from a ranking by increasing values and, in the latter case, the resulting curve would have been symmetric, the center of symmetry being the coordinate point $(0.5, 0.5)$.

In an industrial environment, potential improvement points are obviously almost innumerable. It could even improve indefinitely, everything and anything. It should however not be forgotten that the improvement also costs and in return should provide added value, or at least offset losses!

15.4.3.1.1 Gini Index

The studied phenomenon is even less equally distributed as the curve moves away from the straight line of equal distribution. Economists, managers or corporate departments responsible sometimes use (for their performance dashboard) a synthetic indicator to measure this phenomenon and its evolution: the "Gini index" (also named "Gini coefficient").

This coefficient is defined by the ratio:

$$G=\frac{A}{A+B} \tag{15.4.160}$$

where the surfaces A and B relate to the following figures:

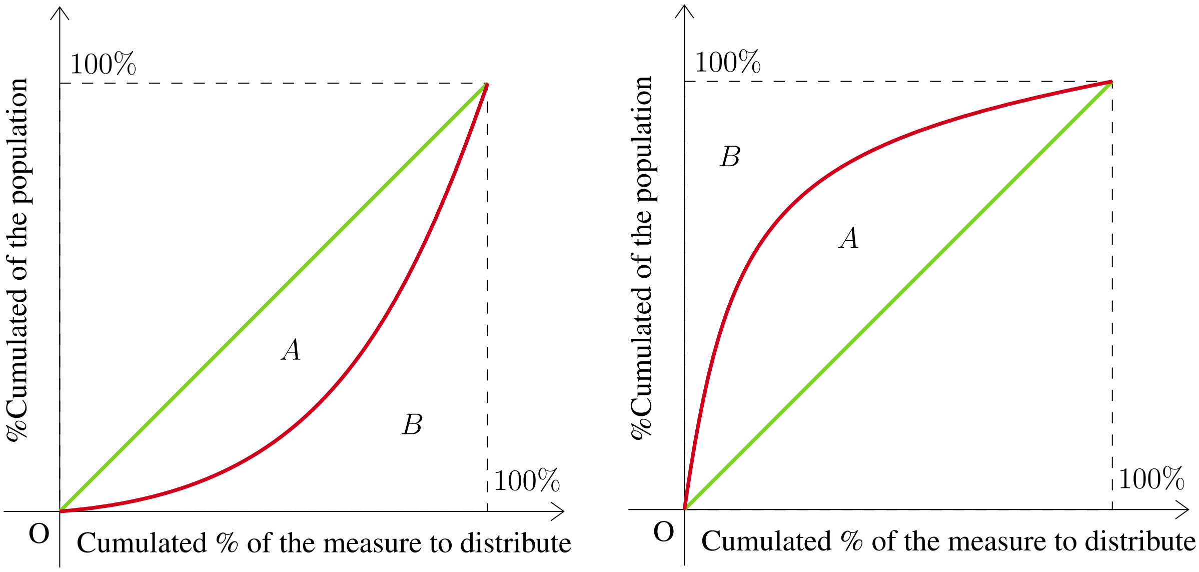

100%
%Cumulated of the population
A
B
100%
O
Cumulated % of the measure to distribute
100%
%Cumulated of the population
B
A
100%
O
Cumulated % of the measure to distribute

Figure 15.298: **The two most common situations of the Gini index**

The Gini coefficient is thus a real number between 0 and 1. In case of perfect equality, it is equal to zero (since $A=0$). In case of total inequality is equal to 1, since $B=0$. Therefore, as G increases from 0 to 1, inequality of the distribution increases.

Knowing that the Lorenz curve 1×1 is seen that the area $A+B$ is equal to half of this surface. So we have:

$$A+B=\frac{1}{2} \tag{15.4.161}$$

We can thereby write:

$$G=\frac{A}{A+B}=\frac{A}{\frac{1}{2}}=2A \tag{15.4.162}$$

Also as:

$$A+B=\frac{1}{2}\Leftrightarrow A=\frac{1}{2}-B \tag{15.4.163}$$

Finally, we can write that:

$$G=\frac{A}{A+B}=2A=1-2B \tag{15.4.164}$$

Using the trapezoidal numerical integration (see section Numerical Methods page 6624) we get that the area B which is given by:

$$\begin{aligned}G&=1-2\left(\frac{A_G+A_D}{2}\right)=1-(A_G+A_D)=1-h\left(\sum_{i=0}^{n-1}f\left(x_{i+1}\right)+\sum_{i=0}^{n-1}f\left(x_i\right)\right)\\&=1-h\sum_{i=0}^{n-1}\left(f\left(x_{i+1}\right)+f\left(x_i\right)\right)=1-h\sum_{i=0}^{n-1}\left(y_{i+1}+y_i\right)\end{aligned} \tag{15.4.165}$$

where as reminder h is the length of the intervals all assumed equals (which is always the case in the context of Lorenz analysis). This relation is correct if and only if the data table configuration is such that we get the following type of chart (which is rare...):

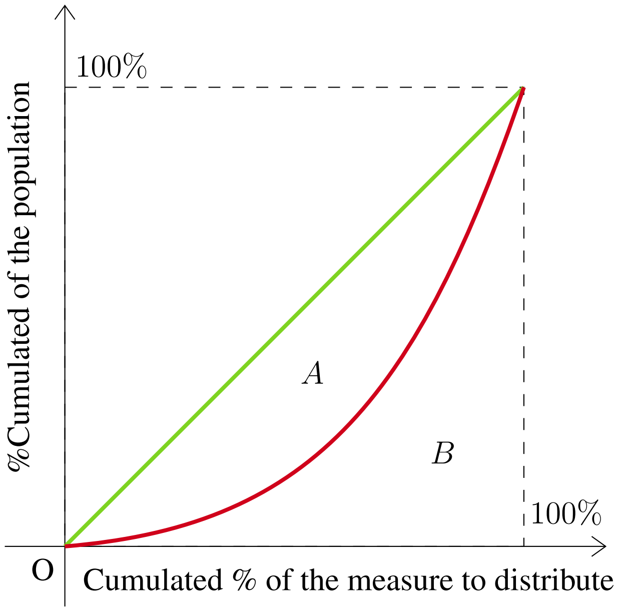


Figure 15.299: **Gini Index first case calculation**

Traditionally we find the last relation in the form:

$$G=1-\frac{1}{n}\sum_{i=1}^{n}\left(y_{i+1}+y_i\right) \tag{15.4.166}$$

where n is the number of statistical units (the size of the population).

In the situation where the chart looks as (this case is significantly more common in the business and this also corresponds to our data table above!!):

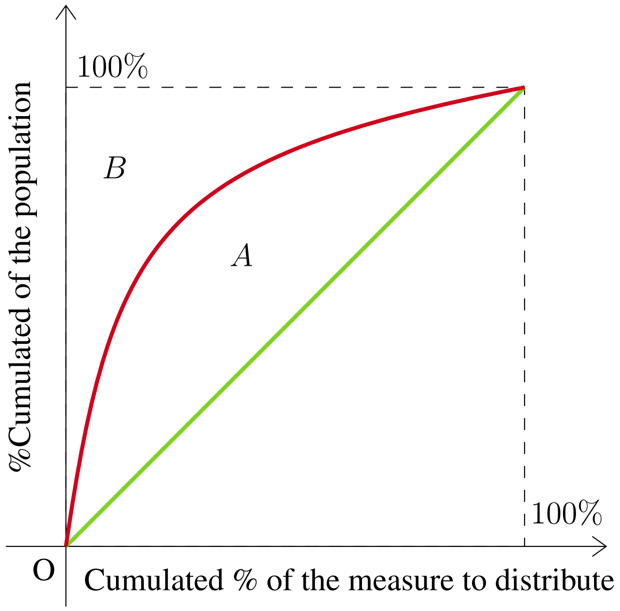


Figure 15.300: **Gini Index second case calculation**

The trapezoidal numerical integration method brings us then to write:

$$\begin{aligned} G &= 1-2\left(1\cdot\frac{-A_G+A_D}{2}\right)=1-2+(A_G+A_D)\\ &=-1+h\left(\sum_{i=0}^{n-1}f(x_{i+1})+\sum_{i=0}^{n-1}f(x_i)\right)=-1+h\sum_{i=0}^{n-1}\left(f(x_{i+1})+f(x_i)\right)\\ &=-1+h\sum_{i=0}^{n-1}(y_{i+1}+y_i) \end{aligned} \tag{15.4.167}$$

Traditionally we find this last relation in the form:

$$G=-1+\frac{1}{n}\sum_{i=1}^{n}(y_{i+1}+y_i) \tag{15.4.168}$$

In the case that serves us as an example from the beginning with the table:

$i\%$	$N_i\%$	$i\%$	$N_i\%$	$i\%$	$N_i\%$	$i\%$	$N_i\%$	$i\%$	$N_i\%$
0.5	16.28	20.5	72.68	40.5	84.86	60.6	92.45	80.5	97.6
1	24.55	21	73.11	41	85.09	61	92.61	81	97.7
1.5	30.12	21.5	73.51	41.5	85.32	61.5	92.77	81.5	97.8
2	34.36	22	73.92	42	85.54	62	92.92	82	97.89
...	...	...	...	...	...	...	...	...	...
19	71.33	39	84.18	59	91.97	79	97.3	99	99.7
19.5	71.79	39.5	84.41	59.5	92.13	79.5	97.4	99.5	99.99
20	72.24	40	84.64	60	92.29	80	97.5	100	100

and adopting the notations accordingly, this gives:

$$G = -1 + \frac{1}{n}\sum_{i=1}^{n}(y_{i+1} + y_i) = -1 + \frac{1}{200}\sum_{i=1}^{200}(N_{i+1}\% + N_i\%) = -1 + \frac{335.10}{200} \cong 0.675 \tag{15.4.169}$$

15.4.3.2 Weighted Ishikawa Diagram

The traditional "Ishikawa diagram", or "fishbone diagram", is a qualitative tool of management. Using this tool one can show the relations between causes and the analysed effect. The most often used is the Ishikawa diagram in a form named the "6M+E model". The symbol 6M+E describes next general causes: man, machine, material, method, management, measurement and environment.

This diagram is presented in the figure below:

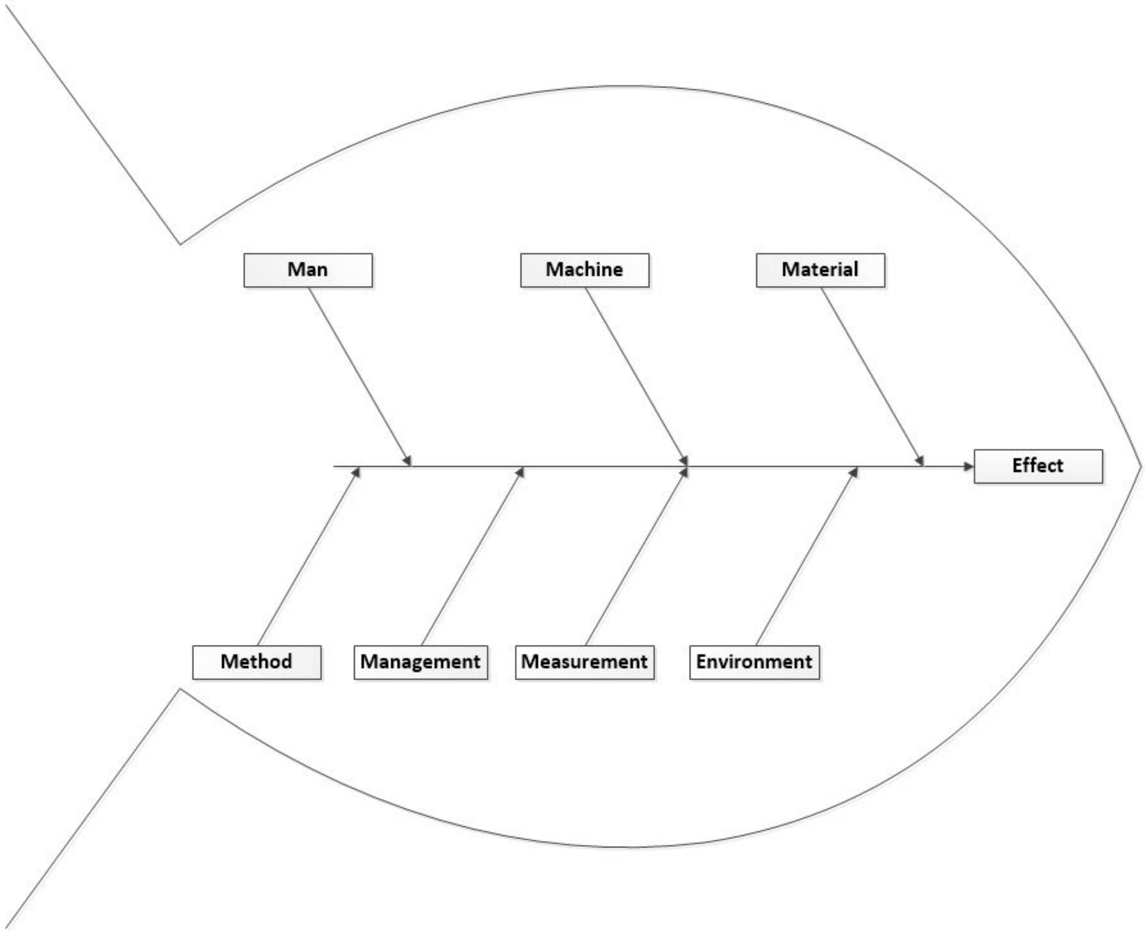


Figure 15.301: **6M+E Ishikawa diagram**

But the model of the classical Ishikawa diagram is not complete. There is no quantitative information to obtain from this diagram and as we know in management: *can be improved only what can be quantified*! We own in 12005 (holocene calendar) to Aleksander Gwiazad a first approach to quantify this diagram! This approach is in my point of view an "a priori" analysis as in don't make usage of the frequency of events. But obviously it is quite easy to improve this model by including a frequency data or also with Bayesian point of view or even with Monte Carlo simulations.

Below is presented the method of preparing the weighted Ishikawa diagram:

1. Determination of a set of main causes
2. Determination of subcauses
3. Determination of weights of main causes
4. Preparing the weighted Ishikawa diagram
5. Conducting the stratification analysis
6. Determination the set of important causes and subcauses

To determine the weights of connections (causes) it is proposed to use a form of the Saaty matrix (from the name of the creator of the AHP decision tool that we have studied in the section of Game and Decision Theory at page 7462). As we know this matrix is a non-symmetric influence matrix which influential elements are in the rows, this is why the summations are done for each row and not for each column:

Input **Input**	**Man**	**Machine**	**Material**	**Method**	**Management**	**Measurement**	**Environment**	**Sum**	**Normalised Sum**
Man	X	0.5	1	0.5	1	1	0.5	4.5	0.214
Machine	0.5	X	1	0.5	0.5	1	0	3.5	0.167
Material	0	0	X	0	0.5	0.5	0	1	0.048
Method	0.5	0.5	1	*X*	0.5	1	0.5	4	0.19
Management	0	0.5	0.5	0.5	X	1	0	2.5	0.119
Measurement	0	0	0.5	0	0	X	0	0.5	0.024
Environment	0.5	1	1	0.5	1	1	X	5	0.238

Table 15.76: **Ishikawa Satty matrix**

Using this matrix one can compare in pairs every cause on the same level of analysis. The scale of notes is as presented: $0, 0.25, 0.5, 0.75, 1$. Where 0 means that analysed cause is of no importance. The note 0.25 is for causes of small importance. If the importance of causes is equal one gives them the note 0.5. The notes 0.75 and 1 are for cause that are more important and very important. Next, in the column $\sum$ there is calculated the sum of all component notes. This sum is the weight of each cause. In the next column the normalized weight is computed. The normalized weights are inscribed to the Ishikawa diagram.

In the same manner the weights of subcauses, for each main cause separately, are determined. Next are computed the absolute weights by comparing the weights of subcauses with the weight of the main cause.

For example consider that for Man branch we have three causes Subcause1, Subcause2, Subcause3 with respective weights $\{0.17, 0.33, 0.5\}$ that obviously sum up to 1.

After each influence is given by:

$$\begin{aligned} 0.17 \cdot 0.214 &= 0.036 \\ 0.33 \cdot 0.214 &= 0.071 \\ 0.50 \cdot 0.214 &= 0.107 \end{aligned} \tag{15.4.170}$$

A complete example is given in the figure below:

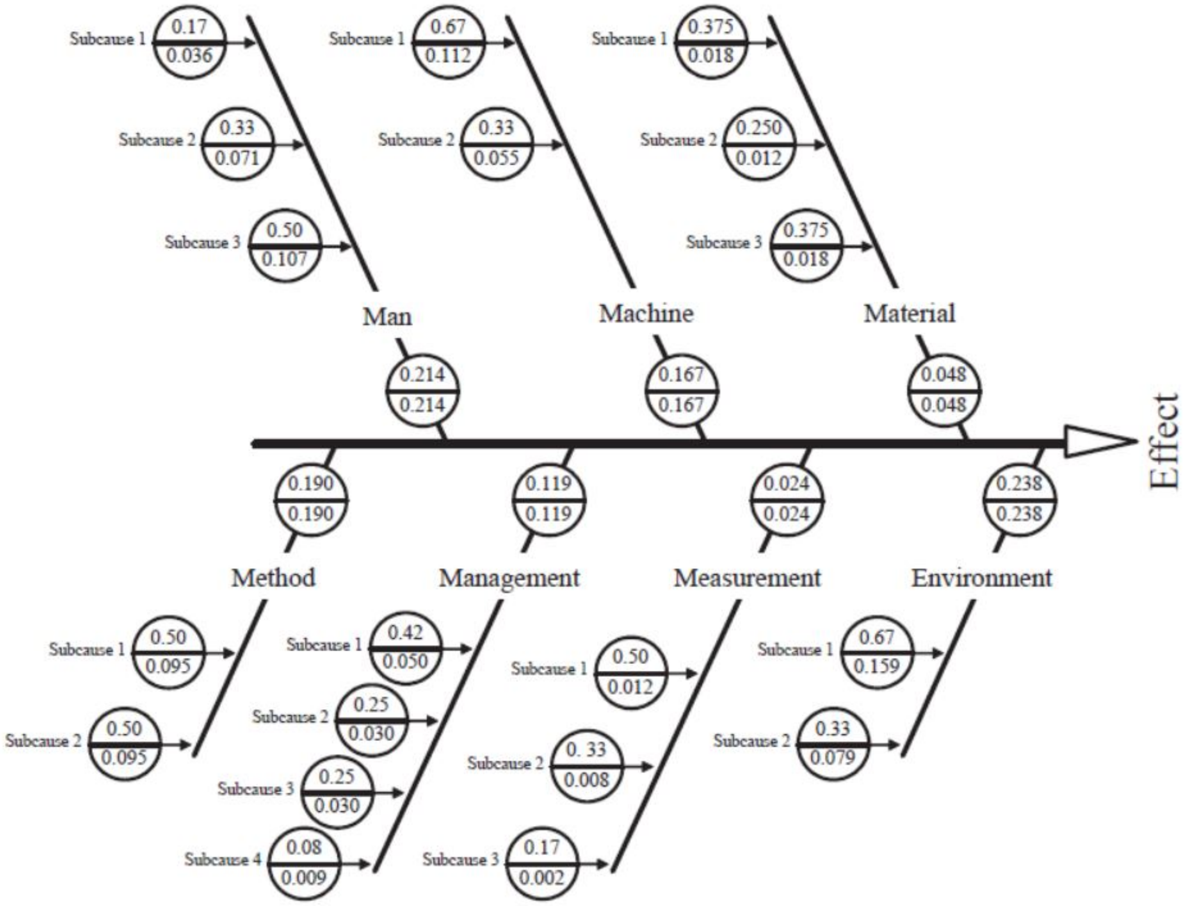

Subcause 1
0.17
0.036
Subcause 2
0.33
0.071
Subcause 3
0.50
0.107
Subcause 1
0.67
0.112
Subcause 2
0.33
0.055
Subcause 1
0.375
0.018
Subcause 2
0.250
0.012
Subcause 3
0.375
0.018
Man
Machine
Material
0.214
0.214
0.167
0.167
0.048
0.048
Effect
0.190
0.190
0.119
0.119
0.024
0.024
0.238
0.238
Method
Management
Measurement
Environment
Subcause 1
0.50
0.095
Subcause 2
0.50
0.095
Subcause 1
0.42
0.050
Subcause 2
0.25
0.030
Subcause 3
0.25
0.030
Subcause 4
0.08
0.009
Subcause 1
0.50
0.012
Subcause 2
0. 33
0.008
Subcause 3
0.17
0.002
Subcause 1
0.67
0.159
Subcause 2
0.33
0.079

Figure 15.302: **Weighted Ishikawa diagram (source: Aleksander Gwiazad/Industrial Management and Organisation)**

The next step is the stratification analysis. This analysis is based on the Pareto rule. The first step in this analysis is to prepare a table of weights of all subcauses:

N°	Subcause	Weight	Cumulated Weight
1	environment 1	0.159	0.159
2	machine 1	0.112	0.271
3	man 3	0.107	0.378
4	method 1	0.095	0.473
5	method 2	0.095	0.0.568
6	environment 2	0.079	0.647
7	man 2	0.071	0.718
8	machine 2	0.055	0.773
9	management 1	0.050	0.823
10	man 1	0.036	0.859
11	management 2	0.030	0.889
12	management 3	0.030	0.919
13	material 1	0.018	0.937
14	material 3	0.018	0.955
15	material 2	0.012	0.967
16	measurement 1	0.012	0.979
17	management 4	0.009	0.988
18	measurement 2	0.008	0.996
19	measurement 3	0.002	0.998

In the presented case the first eight subcauses constitute the group of important causes.

15.4.4 Supply Chain Management

The challenge of inventory management and supply chain is something critical for all big companies worldwide. The goal is to implement processes that optimize the economic function under the constrained of in theory a flawless availability. These are the objectives of the stocks manager. This entails having a visibility on the inventory and appropriate methodologies to different situations.

A production without stock is almost inconceivable (even if it is the final goal!) given the many functions performed by stocks. Indeed, stockpiling is required if:

1. There is non-coincidence in time or space of production and consumption: the stock is essential in this case because it is hard or costly to produce where and when the demand arises (this is typically the purpose of the Lean Six Sigma). Classic examples are the manufacture of toys or confectionery for non-coincidence in time, and supermarkets for the non-coincidence in space.
2. There is uncertainty about the level of demand or on the price: if there is uncertainty about the quantity requested, we will provide a safety stock allowing to face to a peak demand (taking care to avoid the "bullwhip effect" that is to say a to big difference between demand and supply). If there is uncertainty about the price, we will build a speculative stock. For example, oil companies buy much needed crude oil when its price is relatively low on the market.
3. There are risk of issues in chain: this is to avoid that a failure on a machine affects the whole supply chain. A delay in performance on the previous post (manufacturing step) or a transport strike will not immediately stop the entire production process if there are stocks buffers.
4. There is presence of launch costs: in this case, working in batches allows an economy of scale on the cost of production launch but, in contrast, causes an increase in cost of ownership of the stock.

The control of stocks and supply chain of a company is as fundamental in the life of it. To reduce (or optimize it depends...) the various costs that surround everything that has to do with storage, we must use our knowledge in mathematical statistics and differential and integral calculus as we shall see later.

Remarks

R1. The application of the tools we will see now does not really apply to SMEs with fewer than 50 employees producing small pieces irregularly but rather to multinationals producing in enormous quantities of small objects or small quantities of significant size objects and this regularly. In the facts almost nobody use the models we will see below because the assumptions are very difficulty to satisfy but it is still a very good study to be able to build more complicated heuristic models.

R2. In facts we use a combination of Monte Carlo simulations and Data Mining (Machine Learning) techniques to improve supply chain management in a very efficient ways (application of Data Mining to spare millions of dollars per year are numerous in the specialized literature).

First, we will establish how to determine the required initial stock for a company based on statistical data using very simple naive models afterwhat we will do the same with replenishment models whose the approach is somewhat different and allows like for the first one to achieve satisfactory results for large scale stable markets.

The models we will build will therefore give the opportunity under the constraints of stable market (stable in time or "quasi-static" as will say thermodynamics specialists...) to:

1. To regulate the flow of supplies
2. Allow batch production (reduced production costs)
3. To meet seasonal demands

As we can guess, additional stocks can generate "interest costs" (fixed capital), "obsolescence costs" (between the products become obsolete time), "storage costs", "insurance costs" (protection against accidents that can provide on the products) and many others...

We distinguish in the field of supply chain at least three types of stocks:

1. The "safety stock" that responds to the most common issues associated with the consumption and delivery.
2. The "alert stock" or "critical stock" is the stock level for which an order is triggered at risk to be out of stock.
3. The "minimum stock", also named "buffer stock" or "order point" or "reorder point" is by definition the sum of the safety stock and minimum stock.

We often consider at least three stocks management strategies:

1. The "supply chain management by control point": the supply of the stock is raised when it is observed that the stock falls below the reorder point.
2. The "supply chain calendar management": the supply of the stock is triggered at regular intervals T, for example, daily or weekly.
3. The "supply chain conditional calendar management": the supply of the stock is triggered at regular intervals T, but only when it is observed that the stock falls below the control point.

A stock is formed to meet future demand. In case of random demand, there may be no coincidence between the application and the stock. Two cases are of course possible:

1. A demand higher than the available stock: we speak then of "out of stock" or "inventory shortages".
2. A demand lower than the stock: we speak then of "residual stock".

The criteria commonly adopted in inventory management is that of cost minimization (...). We denote this function by the letter C followed in brackets (or as index) of the system control variables.

15.4.4.1 Supply Chain Management in uncertain future

Let's begin our study with the simplest case which requires that consumption is statistically steady in time and under statistical control. It follows (see sections Statistics page 938 and Industrial Engineering page 8692) that the instantaneous consumption then follows a Normal law $\mathcal{N}(\mu,\sigma)$ or more technically: a Gaussian stochastic process where for reminder each day is independent of the previous one (see section Economy page 7899).

Let us take a concrete example, since the theory has already been studied extensively in the section of Statistics and Economy.

Remark

The case where we want to forecast sales using different mathematical assumptions is presented in the section of Economy when we study Time Series Analysis (Forecasting Techniques) at page 7896.

✍**Example:**

Consider an article whose instantaneous demand X follows approximately a Normal law of parameters:

$$\mathcal{N}(\mu,\sigma)=\mathcal{N}(90,20) \tag{15.4.171}$$

Suppose that stock available at the time of restock order is 500 units and that the restock delay is 5 days. We would like to know the cumulative probability of being above or equal to the out of stock 5 days after the last restock and the cumulative probability of being below the assumed consumption?

By using the property of stability of the Normal distribution (see section Statistics page 710) the statistical cumulative consumption after 5 days give us:

$$\begin{aligned}\mu_{5d}&=5\cdot 90=450\\ \sigma_{5d}&=\sqrt{20^2+20^2+20^2+20^2+20^2}=20\sqrt{5}\cong 44.72\end{aligned} \tag{15.4.172}$$

Therefore we have after 5 days the consumption that follows a Normal law of characteristics:

$$\mathcal{N}(\mu_{5d},\sigma_{5d})=\mathcal{N}(\mu_{450},44.72) \tag{15.4.173}$$

Using a spreadsheet software like Microsoft® Excel® 11.8346 we get the cumulative probability $P(X \geq 500)$ to be out of stock at the fifth day:

`=1-NORM.DIST(500,450,44.72,1)=13.17%`

And we can also easily calculate the cumulative probability of being below the assumed consumption each day:

`=1-NORM.DIST(500/5,90,20,1)=30.85%`

Obviously, in practice the reality is far less simple and more..., the Normal law has in fact a probability of negative sales, which is unrealistic... But it is a first model that can be confronted with the reality to see its predictive power and afterwards adapted with more complicated distributions.

Imagine now that the only variability we need to protect against is demand variability, and good historical data are available. The safety stock needed to give a certain level of protection simply is the standard deviation of demand variability multiplied by the Z-score. For example, to satisfy demand with a 95% confidence level, according to statistical analysis, it's necessary to carry extra inventory equal to 1.65 standard deviations of demand variability. This is equivalent to a Z-score of 1.65.

The table below shows the relationship of desired cycle service levels to Z-score:

Desired cycle service level	**Z-score**
84	1
85	1.04
90	1.28
95	1.65
97	1.88
98	2.05
99	2.33
99.9	3.09

Table 15.77: **Cycle service level**

We have seen just earlier that the standard deviation of demand σ_D is the weekly standard deviation times the square root of the ratio of the time units. This is:

$$\text{Safety Stock} = Z_{\alpha/2}\sigma_D\sqrt{\frac{\text{PC}}{T_i}} \tag{15.4.174}$$

where PC is the performance cycle (another term for "total lead time") and T_i is the time increment used for calculating standard deviation of demand.

The performance cycle includes the time needed to perform functions such as deciding what to order or produce, communicating orders to the supplier, manufacturing and processing, and delivery and storage, as well as any additional time required to return to the start of the next cycle.

Caution!!! Safety stock determinations are not intended to eliminate all stockouts - just the majority of them! It's a statistical tool, not a deterministic one!

✍**Example:**

Let us consider an average daily demand (normally distributed) of 15 units, a lead time (LT) in days (constant) of 2 days and a standard deviation of daily demand equal to 5 units and we want a service level of 90%.

Then the Reorder Point (ROP) and Safety stock (SS) is given by:

$$\text{ROP} = \mu_D \cdot \text{LT} + \text{SS} = \mu_D \cdot \text{LT} + Z_{\alpha/2}\sigma_D\sqrt{\frac{\text{PC}}{T_i}} = 15 \cdot 2 + 1.28 \cdot 5 \cdot \sqrt{2} \tag{15.4.175}$$

So the safety stock is about 9 units.

15.4.4.1.1 Safety Stock formula

Now let us prove the famous "safety stock formula" or also named "King's safety formula":

$$\boxed{\text{SS} = Z_{\alpha/2}\sqrt{\mu_L\sigma_D^2 + \mu_D^2\sigma_L^2}} \tag{15.4.176}$$

where μ_L, σ_L are the mean and standard deviations of the time between deliveries and μ_D, σ_D are the mean and standard deviation rates for the demand.

Obviously therefore $\sqrt{\mu_L\sigma_D^2 + \mu_D^2\sigma_L^2}$ is the standard deviation of the item quantity consumed between deliveries, considering that the time between deliveries varies.

As we already know μ_D and σ_D^2 are the mean and variance of the demand per unit time, so that the demand for a period of length T has a mean of $\mu_D T$, a variance of $\sigma_D^2 T$, and therefore a standard deviation of $\sigma_D\sqrt{T}$.

Note that the assumptions are only that these means and variances exist. At this stage, we don't have to assume more, and particularly not that times between deliveries and demand follow a particular distribution. If $D(T)$ is the demand during an interval of duration T, since:

$$\begin{aligned}\text{E}(T) &= \mu_D T\\ \text{V}(D(T)) &= \sigma_D^2 T\end{aligned} \tag{15.4.177}$$

we have using the Huygens relation proved in the section Statistics (see page 593):

$$\text{E}(D(T)^2) = \text{V}(D(T)) + (\text{E}[D(T)])^2 = \sigma_D^2 T + \mu_D^2 T^2 \tag{15.4.178}$$

If we now allow T to vary, around mean μ_L with, standard deviation σ_L, we have:

$$\text{E}(D(T)) = \mu_D \text{E}(T) = \mu_D\mu_L \tag{15.4.179}$$

and therefore:

$$\mathrm{E}(D(T)^2) = \mathrm{E}\left(\mathrm{E}(D(T)^2)\right) = \mathrm{E}\left(\sigma_D^2 T + \mu_D^2 T^2\right) = \sigma_D^2 \mathrm{E}(T) + \mu_D^2 \mathrm{E}(T^2) \quad (15.4.180)$$

Hence:

$$\mathrm{E}(D(T)^2) = \sigma_D^2 \mu_L + \mu_D^2 \left(\sigma_L^2 + \mu_L^2\right) \quad (15.4.181)$$

And the last equality follows because using Huygens relation again:

$$\mathrm{V}(T) = \mathrm{E}(T^2) - (\mathrm{E}(T))^2 \quad (15.4.182)$$

which gives us:

$$\mathrm{E}(T^2) = \sigma_L^2 + \mu_L^2 \quad (15.4.183)$$

Finally:

$$\mathrm{V}(D(T)) = \mathrm{E}(D(T)^2) - (\mathrm{E}(D(T)))^2 = \sigma_D^2 \mu_L + \mu_D^2 (\sigma_L^2 + \mu_L^2) - (\mu_D \mu_L)^2 = \mu_L \sigma_D^2 + \mu_D^2 \sigma_L^2 \quad (15.4.184)$$

And one standard deviation of $D(T)$ is:

$$\sigma_{D(T)} = \sqrt{\mathrm{V}(D(T))} = \sqrt{\mu_L \sigma_D^2 + \mu_D^2 \sigma_L^2} \quad (15.4.185)$$

This finish our proof!

These result demonstrate that we can identify what variability is the dominant influence on safety stock requirements. With the recognition of what factors dominate an equation, it becomes easier to focus improvement efforts. In the case where a reduction in safety stock is desired, we can then know if it's more productive to reduce demand variability or lead time variability. Conversely, if a high level of customer service is not required, safety stock can be lowered to a more appropriate level.

Note that all of the above arguments only require the means and standard deviations to exist and that the variance in a period of time implies it is additive (i.e. if two times periods are uncorrelated). There is no assumption at this point that the demand or the lead time follow a Normal distribution. However, the calculation of the multiplier $Z_{\alpha/2}$ used to set an upper bound for the demand in a period, is based on the assumption that the demand between deliveries is Normally distributed.

15.4.4.2 Optimal initial stock management with zero rotation

Imagine a scenario to develop another statistical model but a little more elaborated than the previous one (strongly inspired by the book *Gestion de la Production* [202] of Vincent Giard). We will limit ourselves however here to a discrete statistical law and we will do later again an another example with a continuous law.

Consider for this purpose that the company MAC is the specialist of a given product whose direct manufacturing cost per unit is 25.- and the sales price is 60.- selling per unit (assumed constant prices and thus independent of the quantity and without discount).

The daily sales of this product is daily from 2.5 units on average and an analysis during three months suggests that it follows a discrete distribution like the Poisson law (this time the domain of definition is at least positive in contrast to the previous example...), that is to say that we have a probability distribution of the number X of these products sold during a day. We will truncate to $x=10$ the realization of this random variable as the likelihood of sales above 10 units will be assumed to be zero.

Tabulating this Poisson law for X from 0 to 10 we get:

X	$P(X)$	**Cumulated Probability**
0	0.0821	0.0821
1	0.2052	0.2873
2	0.2565	0.5438
3	0.2138	0.7576
4	0.1336	0.8912
5	0.0668	0.958
6	0.0278	0.9858
7	0.0099	0.9957
8	0.0031	0.9988
9	0.0009	0.9997
10	0.0003	1

Table 15.78: **Cumulated probability of sales**

We then have the above table that shows us that the most commonly sold quantity to an economic agent is approximately 2 (modal value) and calculation of the expected mean gives us for this table:

$$\bar{x}=\mathrm{E}(X)=\frac{1}{n}\sum_i p_i x_i=2.5 \qquad (15.4.186)$$

result which can also be interpreted as the average daily sales than the average level of stock shortages.

We will assume that the stock is lean. In other words, from one day to the next, no unit is reported for sales of the day after because it is not supposed to have anymore available (what we name "inventory management by zero rotation"). The question therefore is to know, given the table above, how many products we have to manufacture (or to order) every day to maximize profit and minimize losses[59].

From the perspective that we want a minimization of cost of ownership C_p (often named "unit cost of ownership" or "unit cost of possession") and that we know the associated cost for unsold is 25.- per unit, while the cost of shortage (often named "unit cost of shortage") is equal to the loss of profit following the failed sales, that is to say the selling price 60.- subtracted by the 25.- of manufacturing cost, gives a shortage cost of 35.- per unit!

Remark

The cost of shortage is not treated only as the loss caused by a failed sales in practice but also include the cost of a temporary replacement solution proposed by the seller to the customer on hold to provide the right product. Internally in companies, the cost of shortage often leads to layoffs of downstream positions when a piece necessary for the manufacturing is no longer available to the production line.

A rational management must consist to calculate the initial stock S (i.e. the number of products to order or to manufactured for the next day) that minimizes the indicator of "management cost" $C(S)$ defined as the sum of the cost of possession associated with average unsold stock $I_P(S)$ (often named "average stock owned" or "average stock possessed") and the shortage cost associated with average stock sales missed $I_R(S)$ (often name the "average stock of unsold"):

$$\boxed{C(S)=C_pI_p(S)+C_RI_R(S)} \tag{15.4.187}$$

This last relation is, however, a simplification of the relation:

$$C(S)=C_pI_p(S)+C_RI_R(S)+C_CI_C(S) \tag{15.4.188}$$

where the last term is the order cost.

Minimizing the cost function $C(S)$ consist in the mathematical point of view to look for an extreme (minimum) of the management cost function such that for the optimal value of the initial procurement S^* the cost $C(S^*)$ is immediately below $C(S^*\mp 1)$ that is to say less by the right and by the left. In other words:

$$\begin{gathered}C(S^*)<C(S^*+1)\\C(S^*)<C(S^*-1)\end{gathered} \tag{15.4.189}$$

That is to say the function is convex. Or after rearrangement:

$$\begin{gathered}C(S^*)-C(S^*+1)<0\\C(S^*)-C(S^*-1)<0\end{gathered} \tag{15.4.190}$$

inequalities that the diagram below may help to better understand:

---

[59] In many corporations there exist a report commonly used by analysts and produced by the "product control" department that attributes or explains the daily fluctuation of a tracked variable to the root causes of the changes and named "PnL report" used a lot in manufacturing, supply chain and portfolio management.

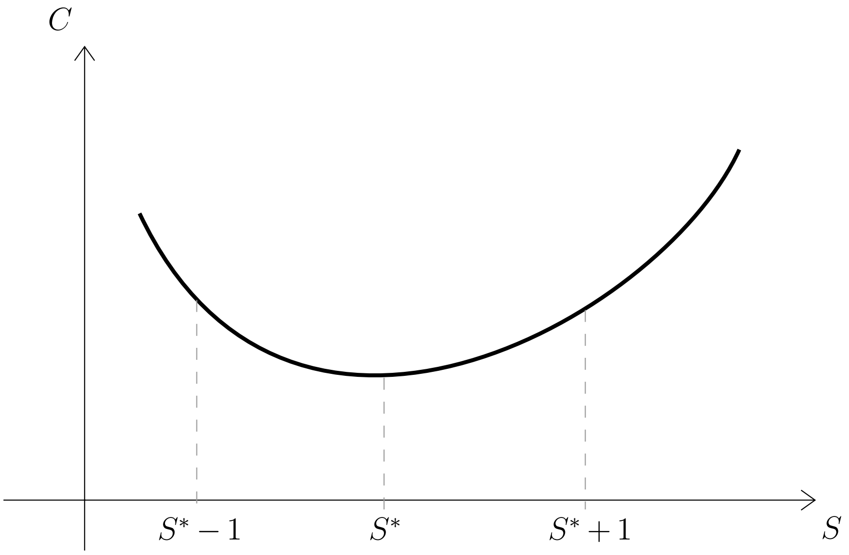


Figure 15.303: **Arbitrary example of the extremum of the function** $C(S)$

From now the question is how to determine S^*. In fact the idea is subtle but simple as it is clearly and thoughtfully exposed.

Let us take the probability distribution of the daily demand law and suppose we want to calculate the average shortage (i.e. expected mean) for a given level of initial stock S for example:

$$I_R(S=4)=\mathrm{E}(x-4) \tag{15.4.191}$$

Or:

$$I_R(S=5)=\mathrm{E}(x-5) \tag{15.4.192}$$

associated with the respective initial stocks with respect to the given distribution (obviously x should be taken to be greater than S). Either in full generality:

$$\boxed{I_R(S)=\sum_{x=S+1}^{n}(x-S)P(x)} \tag{15.4.193}$$

Obviously, if the probability distribution is continuous (or at least approximated as such), the latter relation will be written:

$$I_R(S)=\int\limits_{S+1}^{+\infty}(x-S)f(x)\mathrm{d}x \tag{15.4.194}$$

The idea is then to write the probability density distribution over the missing quantity of stock and not the sold one:

x	$P(x)$	$x-4$	$(x-4)P(x)$	$x-5$	$(x-5)P(x)$
0	0.0821	-	-	-	-
1	0.2052	-	-	-	-
2	0.2565	-	-	-	-
3	0.2138	-	-	-	-
4	0.1336	-	-	-	-
5	0.0668	1	0.0668	-	-
6	0.0278	2	0.0556	1	0.0278
7	0.0099	3	0.0297	2	0.0198
8	0.0031	4	0.0124	3	0.0198
9	0.0009	5	0.0045	4	0.0036
10	0.0009	6	0.0045	5	0.0015
$\sum$	1	-	$I_R(4)=0.1708$	-	$I_R(5)=0.062$

Table 15.79: Probability density distribution with respect to the missing quantity

It appears from the table above that changing the initial stock S from 4 to 5 units decreases the average shortage stock by passing it from 0.1708 to 0.062. But from this result, we can't do anything for now because at our present level of development! This would mean that by taking an initial stock of 10, we would have an average shortage stock of zero (... which does not take us to anything interesting...).

But still we can draw an interesting intermediate result. Indeed let us look on how varies the difference of average break (easily generalizable result - we can make the proof on request if necessary):

$$\begin{aligned}
&I_R(S=4)-I_R(S=5)=\sum_{x=5}^{10}(x-4)P(X=x)-\sum_{x=6}^{10}(x-5)P(X=x)\\
&=(5-4)P(X=5)+(6-4)P(X=6)+(7-4)P(X=7)+(8-4)P(X=8)\\
&+(9-4)P(X=9)+(10-4)P(X=10)-(6-5)P(X=6)-(7-5)P(X=7)\\
&-(8-5)P(X=8)-(9-5)P(X=9)-(10-5)P(X=10)\\
&=1P(X=5)+(2-1)P(X=6)+(3-2)P(X=7)+(4-3)P(X=8)+(5-4)P(X=9)\\
&+(6-5)P(X=10)\\
&=P(X=5)+P(X=6)+P(X=7)+P(X=8)+P(X=9)+P(X=11)\\
&=\sum_{x=5}^{10}P(X=x)=P(X>4)
\end{aligned}\tag{15.4.195}$$

The result is the same if the probability distribution is continuous!

In other words (be careful to what we say !!!) the decrease of average rupture caused by increasing the initial stock of one unit previously dimensioned to $S=4$ is equal to the cumulative probability that the demand is strictly greater than that of the original stock (i.e. $P(X>4)$).

In other words, in case this is not crystal clear (...), the fact of increasing the initial stock certainly decreases the probability of average rupture, but in exchange creates a risk that there are fewer buyers who may meet the offer.

Finally, we can write:

$$I_R(S+1) - I_R(S) = -P(X > S) \tag{15.4.196}$$

The table below shows the decreasing of expected out of stock we obtain by increasing by one unit the initial stock (and respectively the probability of customers able to consume the whole stock...):

x	$P(x)$	$I_R(x) - I_R(x+1)$
0	0.0821	0.9179
1	0.2052	0.7127
2	0.2565	0.4562
3	0.2138	0.2424
4	0.1336	0.1088
5	0.0668	0.0420
6	0.0278	0.0142
7	0.0099	0.0043
8	0.0031	0.0012
9	0.0009	0.0003
10	0.0009	0

Table 15.80: Average rupture by increasing the initial stock of one unit

Now let us look at the unsold $I_P(S)$ as a random function of the initial stock S. Their mean is of course given by (use tables if needed to understand):

$$\boxed{I_P(S) = \sum_{x=0}^{S} (S-x)P(X)} \tag{15.4.197}$$

Obviously, if the probability distribution is continuous (or at least approximated as such), the latter relation is written:

$$I_P(S) = \int\limits_0^S (S-x)f(X)\mathrm{d}x \tag{15.4.198}$$

What we can write:

$$\begin{aligned}
I_P(S) &= \mathrm{E}(S-X) = \sum_{x=0}^{S}(S-x)P(X) = \sum_{x=0}^{+\infty}(S-x)P(X) - \sum_{x=S+1}^{+\infty}(S-x)P(X)\\
&= S\sum_{x=0}^{+\infty}P(X) - \sum_{x=0}^{+\infty}xP(X) - \left[-\sum_{x=S+1}^{+\infty}(x-S)P(X)\right]\\
&= S - \sum_{x=0}^{+\infty}xP(X) - \left[-\sum_{x=S+1}^{+\infty}(x-S)P(X)\right] = S - \sum_{x=0}^{+\infty}xP(X) - [-I_R(S)]\\
&= S - \sum_{x=0}^{+\infty}xP(X) + I_R(S) = S - \bar{x} + I_R(S)
\end{aligned} \tag{15.4.199}$$

Therefore:

$$I_P(S) = S - \bar{x} + I_R(S) \quad (15.4.200)$$

which is then the residual stock of end of period. It is a remarkable result that will allow us to determine S^* only from $I_R(S)$. We can also again consider as obvious that the result will be the same if the probability distribution is continuous.

This last relation can also be written after rearrangement:

$$\bar{x} - I_R(S) = S - I_P(S) \quad (15.4.201)$$

where the term on the left represents the average demand satisfied and the right term average offer used. So this relationship is a special equality of balance between offer and demand.

We can also check this from the table below with a specific example:

$$I_P(S=4) = S - \bar{x} + I_R(S) = 4 - 2.5 + 0.1708 = 1.6708 \quad (15.4.202)$$

x	$4-x$	$(4-x)P(X)$
0	4	0.3284
1	3	0.6156
2	2	0.5130
3	1	0.2138
4	0	-
5	-	-
6	-	-
7	-	-
8	-	-
9	-	-
10	-	-
		$I_P(4) = 1.6708$

Table 15.81: Mean of unsold items

Finally we can write an expression of the management cost $C(S)$, according to the average rupture:

$$C(S) = C_P I_P(S) + C_R I_R(S) = C_P(S - \bar{x} + I_R(S)) + C_R I_R(S) \quad (15.4.203)$$

or what is written after rearrangement:

$$\boxed{C(S) = C_P \cdot (S - \bar{x}) + (C_P + C_R) I_R(S)} \quad (15.4.204)$$

Then it follows under the assumption that:

$$\begin{aligned} C(S+1) - C(S) &= [C_P \cdot (S+1) - C_P\bar{x}) + (C_P + C_R) I_R(S+1)] - [C_P(S - \bar{x}) + (C_P + C_R) I_R(S)] \\ &= C_P + (C_P + C_R)(I_R(S+1) - I_R(S)) \end{aligned} \quad (15.4.205)$$

Which gives using the results obtained previously:

$$C(S+1)-C(S)=C_P+(C_P+C_R)(-P(X>S)) \tag{15.4.206}$$

Under these conditions, the relations:

$$\left.\begin{matrix} C(S^*)-C(S^*+1)<0 \\ C(S^*)-C(S^*-1)<0 \end{matrix}\right\} \Leftrightarrow \left\{\begin{matrix} C(S^*+1)-C(S^*)>0 \\ C(S^*)-C(S^*-1)<0 \end{matrix}\right. \tag{15.4.207}$$

can be written using the prior-previous result:

$$\begin{cases} C_P+(C_P+C_R)(-P(X>S^*))>0 \\ C_P+(C_P+C_R)(-P(X>S-1))<0 \end{cases} \tag{15.4.208}$$

Therefore:

$$\begin{cases} \dfrac{C_P}{C_P+C_R}>P(X>S^*) \\ \dfrac{C_P}{C_P+C_R}<P(X>S^*-1) \end{cases} \tag{15.4.209}$$

Hence S^* is optimal if:

$$\boxed{P(X>S^*)<\frac{C_P}{C_P+C_R}<P(X>S^*-1)} \tag{15.4.210}$$

The result is known in the literature under "newsvendor optimal stochastic stock problem" (there are many versions of this problem).

In our numerical example, we then have:

$$\frac{C_P}{C_P+C_R}=\frac{25}{25+35}\cong 0.417 \tag{15.4.211}$$

and this lies thanks to our table between:

$$P(X>2)=0.4562 \qquad P(X>3)=0.2424 \tag{15.4.212}$$

hence the initial daily optimal inventory also named "minimum stock" to put into production (or order) to minimize the management cost $C(S)$:

$$S^*=3 \tag{15.4.213}$$

It's a number that is obviously very close to the modal value of the initial distribution.

Note that in practice, however, employees who are controlling supply chain in multinational companies have no idea of prices, margins and costs that occur in the shop of their customers. Therefore they get these information (sales, prices, margins, etc.) by paying their customers to get the data (most of time into Microsoft® Excel® data file).

Often practitioners simply take the modal value or the value corresponding to a 95% coverage of the periodic request (this becomes somewhat arbitrary). In the case of our little exercise, for example, the modal value of the initial stock rounded to the nearest integer value would be 2 and at a threshold of 95% would be equal to 5. As a precaution, however, there is a tendency to take rather the modal value (it is better to avoid to throw and it is better to say to the boss we missed and unknown amount of sales quantity rather than to give an exact amount of products returned or throw[60]...).

---

[60] This is a typical behaviour of supply chain manager that don't know statistics or how to communicate them

15.4.4.3 Wilson's Models

There are several stock management optimization models (statistical, Wilson, ABC, 20/80, VSM, ...). Among them, we wanted to stop on the "Wilson's Models" with its three variations that are the best known in the academic world (but not necessarily the most realistic...).

We therefore consider three situations:

1. The company resupplies by ordering batches thanks to external suppliers and then there is no stocks inertia and it redoes an order when the safety stock is reached. We speak then of "Wilson's model with resupply".
2. The company resupplies by an internal production, which then implies a given inertia to return to the desired level of stock. We speak then of "Wilson's model without resupply" (for which all results can be very quickly deduced from the first model without redoing all the mathematical developments).
3. The company resupplies by ordering batches thanks to external suppliers and then there is no stocks inertia and it redoes an order when the safety stock is reached but with inertia and please customers to wait. We speak then of "Wilson's model with waiting time".

Remarks

R1. This model also named the "Economic Lot Size Model" or "Economic Batch Quantity" (EBQ) has for purpose to determine the frequency and optimal resupply quantity for a store, a factory, etc. It is commonly used by supply chain services in large structures. It was actually introduced in 11913 (holocene calendar)...

R2. When the Wilson model is applied to cash management but rather than with inventory costs we work with trading costs, fixed cost of making securities, etc. We then speak of "Baumol-Allais-Tobin model".

The purpose is to determine the strategy that must be adopted so that the total periodic (annual, monthly, weekly, daily, ...) of the orders or manufacturing of goods minimizes the total cost of acquisition and ownership for stocks of the company. We talk then sometimes about "just in time supply chain management" or more simply "JIT supply chain management".

The existence of stocks within a company leads the supply chain manager to ask the question of the optimal level of these latter, avoiding at least two major pitfalls:

1. "Over-stocking", source of costs for the company (physical storage costs, handling costs, premises and land used costs, associated costs, security assurances costs[61], costs immobilized capitals, human unemployment, etc.).

---

[61] The FSA company (Financial Security Assurance) provides insurance to corporate and municipal borrowers in the bond market. FSA determines what added collateral and surplus cashflow each deal needs to bring the bond to investment grade and arranges for the issuer to place necessary collateral into a special corporation controlled by FSA. FSA then rents the issuers its triple-A rating, saving the issuers money through lower interest rates and exacting a hefty premium for itself.

2. "Sub-stocking" which may result in stock ruptures that are detrimental to the production activity or the business activity of the company (factory stoppage, sales loss, customers loss, quality loss, increase in stress for employees, etc.).

Thus, the different models of supply chain management are intended to minimize the management cost in this system of constraints by determining the resupply frequency and the associated quantity.

Consider first a purely qualitative approach. For each reference, the quantities in stock change over time, for example in the following shape:

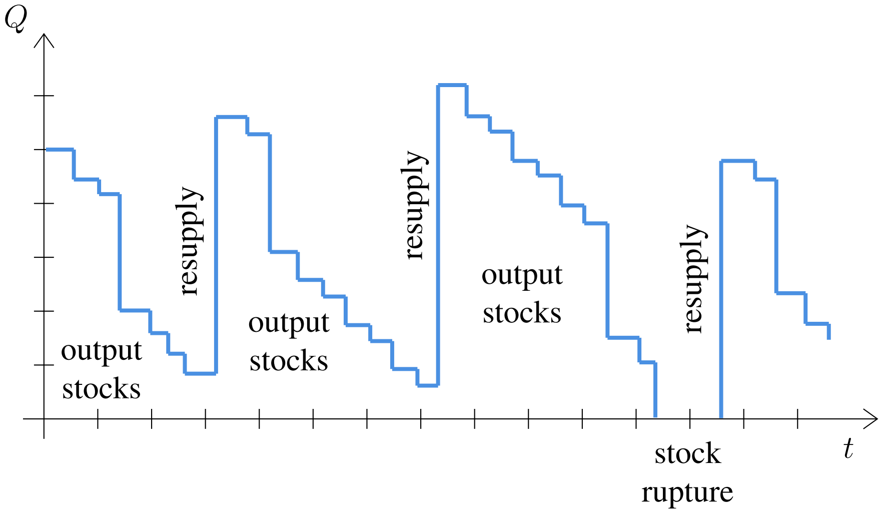


Figure 15.304: **First resupply strategy**

Simplifying we get a graph named "sawtooth supply chart":

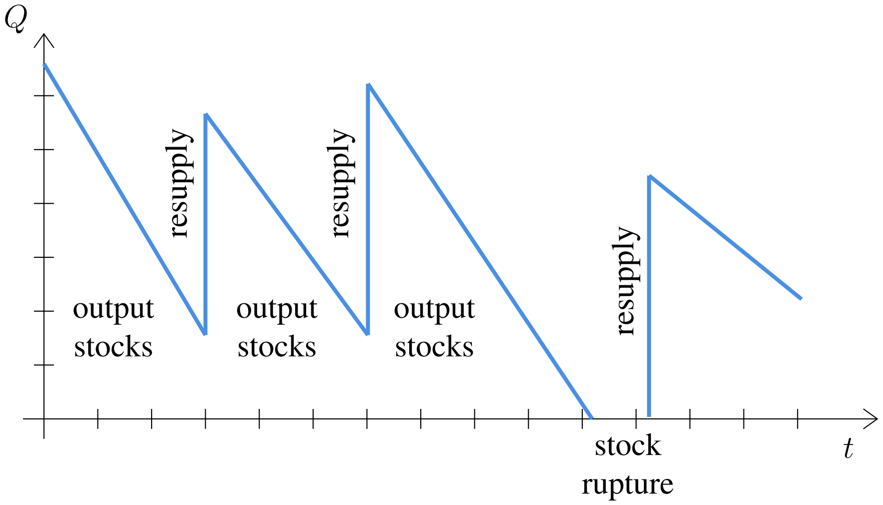


Figure 15.305: **Simplified sawtooth corresponding diagram**

where we are in a typical situation of "supply chain management at non-zero rotation" (that is to say, when unsold can be sold at a later time).

To avoid stock rupture, we must of course ensure that the entry of an order takes place, at the latest, when the inventory quantity becomes zero:

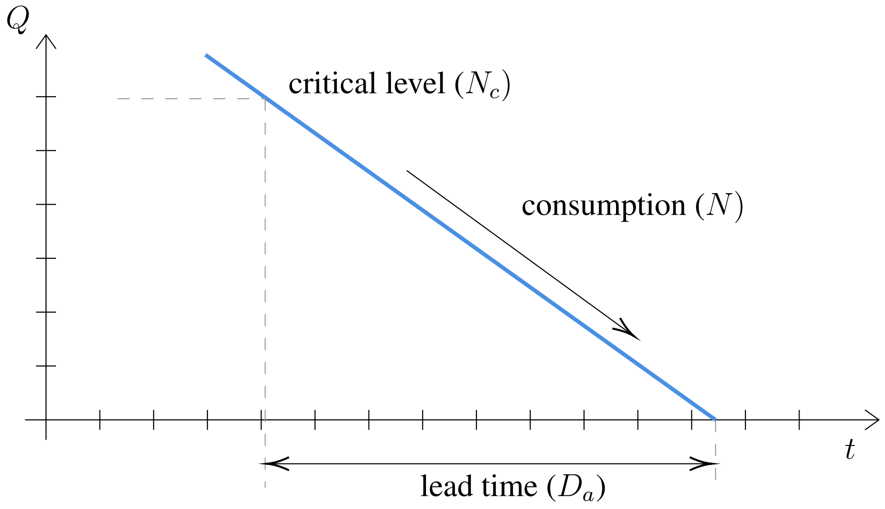

Q
critical level (N_c)
consumption (N)
t
lead time (D_a)

Figure 15.306: **Visual principle of the strategy**

If we consider a constant consumption of a quantity N per unit of time (day, month, year, etc.) and that we know in advance the lead time (in days, months, years, etc.) then if everything is put in equivalent units (e.g. daily) we have the critical level which is given by:

$$N_c = ND_a \tag{15.4.214}$$

which is also assimilated to the justified terminology of "control point" since this is the amount we have in stock from which we must launch a new procurement order:

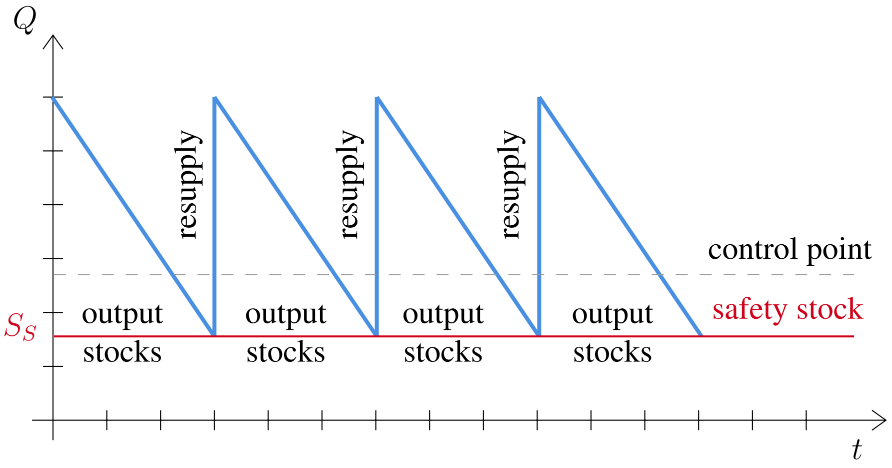

Q
resupply
resupply
resupply
control point
S_S
output stocks
output stocks
output stocks
output stocks
safety stock
t

Figure 15.307: **Control point representation (sawtooth diagram)**

Thus, if the consumption is of 10 units per day, and the lead time is 15 days, then the critical level is of 150 units.

To prevent hazards (strike, transport, consumption variation, replacements, ...) we plan a safety stock S_s:

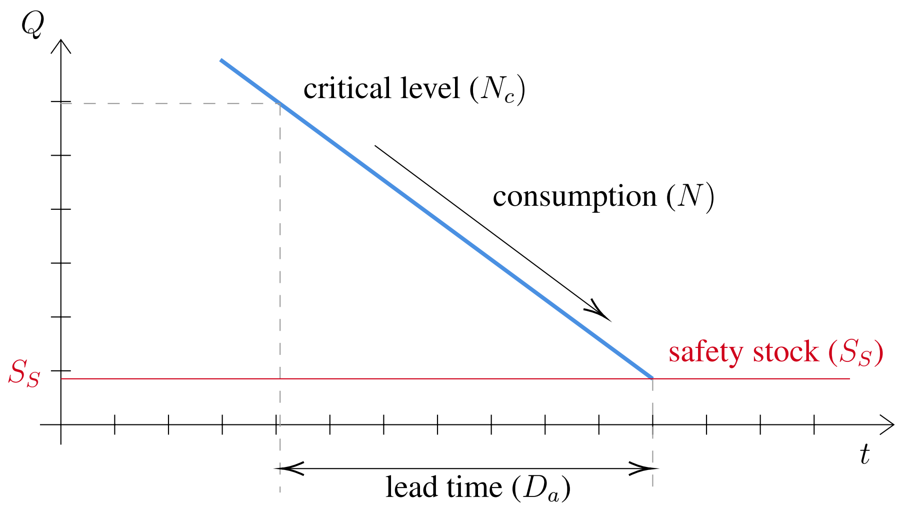


Figure 15.308: Representation of safety stock on the sawtooth pattern

We then have for the critical level:

$$N_c = S_s + ND_a \tag{15.4.215}$$

Let us now see the influence of the number of deliveries on the cost of storage (since the higher the rate of retention is bigger, more the storage costs are high). Suppose for this that the market consumes 100 units per month and this regularly. In the case of a single annual supply, the consumption is represented by the following diagram (no safety stock in order to simplify the example):

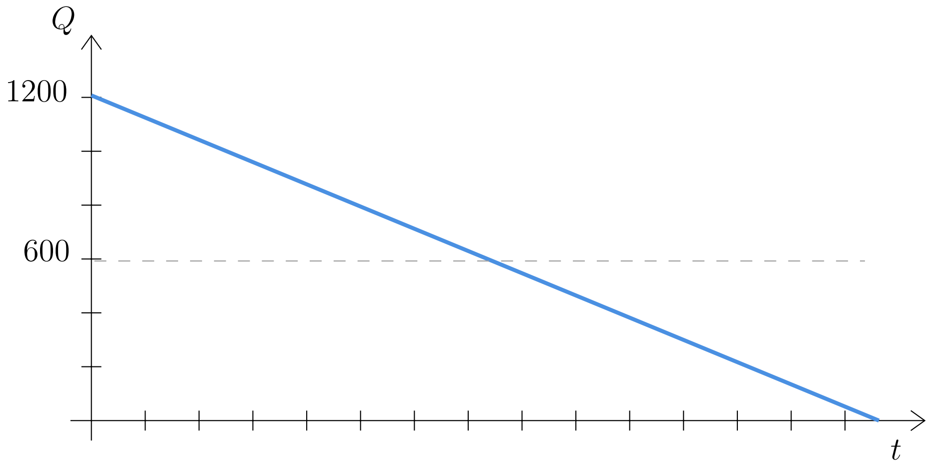


Figure 15.309: Representation of the effect of a single supply

where we immediately see that the average stock is 600. The average stock is obtained by simple arithmetic average or simply using the definition of the integral average (see section Statistics page 569) of the consumption function:

$$\mu_f = \frac{1}{b-a}\int\limits_a^b f(t)\mathrm{d}t = \frac{1}{12-0}\int\limits_0^1 2(-100t+1,200)\mathrm{d}t = 600 \tag{15.4.216}$$

thus trivially half maximum stock (or otherwise seen: the area of the rectangle triangle divided by T...).

And if we divide into two resupply, then we have:

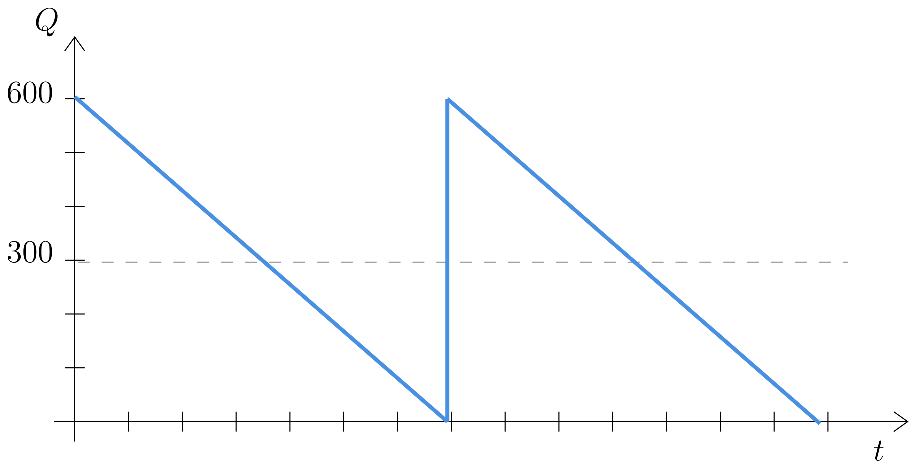


Figure 15.310: **Representation of the effect of a double supply**

Therefore an average stock twice lower and therefore an average storage cost twice lower too (therefore we understand better what the JIT supply chain management is the best way to management stocks in the point of view of costs). But of course we must associate the cost of supply to it (when there is one). It is at this level of complexity that comes precisely the mathematical formalization of Wilson.

Each purchase order or production order hence costs something to the company. The "launch cost" or "cost of purchase" or "restocking cost" to produce or order new batches represent by definition all related costs (administrative, machine settings, preparation, communication, ages, etc.) relatively to the fact of placing an order (or production) and is assumed to be proportional to the quantity. These costs are determined using analytical cost accounting.

Thus, the cost of an order is obtained by dividing the total operating cost of the purchasing department by a significant and relevant magnitude. For example, the number of annually purchased orders (or orders of manufacturing). The cost of a manufacturing launch will be calculated by dividing the total cost of running the manufacturing services, in which, we must add the cost of machines maintenance and wages by of manufacturing launches.

These values depend mainly on the company, of its choices in cost accounting methods. It is difficult to define a range of standard values. Many companies do not know accurately how much cost them an order or a production launch (and many simply do not know how to make an analysis...) and as most manager don't know how to make use of probability distributions the decisions are even more difficult for the executive board.

The cost of ownership of the stock consists of expenses related to the physical storage but also the non-remuneration of capital tied up in the stock (or even the cost of debt to finance the stock). For this last reason, this cost is considered to be proportional to the value of the average stock and the holding period of the stock (at lest on short periods because in reality it is an exponential as a factor of $(1+t\%)^n$).

The annual possession/ownership rate $t_O\%$ is the cost of ownership down to a monetary unit of stored material. It is calculated by dividing the total annually cost of ownership by the average annual stock. This rate typically and obviously fluctuate depending on the product value stored in the stock, depending on the type of articles and the quality of their inventory management.

Wilson has established a famous formula to determine the optimum stock quantity based on some assumptions that we will study right now!

Remark

In reality it seems that the model was developed by the production engineer Ford W. Harris in 11913 (holocene calendar), but R. H. Wilson, a consultant who applied it extensively, and K. Andler are given credit for their in-depth analysis.

15.4.4.3.1 Wilson's model with resupply

The highly simplified assumptions of this model are:

H1. The periodic demand (ask) is known and certain (deterministic and stable in other words).

H2. The ordered quantities are constant at each period.

H3. Shortages (stock-outs) occurs at end of period (so there is never a lack of stock and therefore no unknown risks).

H4. The production time is constant and the resupply is instantaneous.

H5. All costs (stocks, articles, procurement, production, manufacturing ...) are time invariant.

H6. Cost of ownership (also named "carrying cost" sometimes) is proportional to the value (short time hypothesis).

H7. The planning horizon is infinite.

And according to the authors this list of assumptions vary more or less (certain assumptions being implicit or being relatively trivial or forgotten/hidden).

Remarks

R1. We assume that the stock management is done in a given time period.

R2. Based on these assumptions/hypothesis, we conclude that there will be the same level of resupply at every period, and that total cost of shortage is zero.

Let us define:

- $t_O\%$ the periodic possession/ownership rate.

- N the quantity corresponding to a demand or respectively the number of items consumed per period.
- Q is the quantities of resupplies or parts launched for manufacturing in once during this same time (batch sizes).
- P_U the unit purchase price of the pieces (price assumed as constant!).
- S_S the considered safety stock for this pieces (also assumed as constant over time) to answer to fluctuations over time (fluctuations also assume to be constant over time).
- C_L the global cost by resupply/acquisition by order or of manufacturing launch.

From this primitive variables we can also create the following common variables:

- The "unit cost of storage" often (but not always!) calculated on the basis of the unit purchase price of a pieces:

  $$C_{US}=t\%P_U \qquad (15.4.217)$$

- The "stocks inertia" which can be seen more explicitly as the "number of periodic launches" (or the replenishment rate) to meet the demand. The inertia of the stock thus has the dimension of a reciprocal of the time:

  $$J_S=\frac{N}{Q} \qquad (15.4.218)$$

- The "inertia cost" or respectively the "cost of acquisition" or "launch cost" or "restocking cost" is therefore per period unit:

  $$J_C=\frac{N}{Q}C_L=J_SC_L \qquad (15.4.219)$$

  The latter is assumed to be proportional to the consumption! What is important to see is that the launch cost is inversely proportional to the amount Q and therefore it tends to zero as Q tends to infinity. Having said that, normally we will have in most theoretical cases $Q\leq N$.

- The average stock in the company in the assumption of a constant consumption (linear decrease of the stock) and a constant level of security within the time period is trivially by period by:

  $$\bar{S}=S_S+\frac{Q}{2} \qquad (15.4.220)$$

The "periodic cost of ownership", also named "cost of ownership" or "management cost" or "storage cost" is then given by:

$$C_P=t_O\%P_U\bar{S}=t_O\%P_U\left(S_S+\frac{Q}{2}\right) \qquad (15.4.221)$$

So this is the function of a straight line (whose intercept is non-zero if the safety stock is not zero) if we consider that only Q is the variable. It is important to note that this cost does not take into account the volume discount concepts made by commercial...

These proposals therefore lead us to the equation of the "total cost of resupply", also named "total cost of storage" that we will seek to minimize:

$$C_{\text{tot}}(Q)=J_C+C_P=\left(\frac{N}{Q}C_L\right)+\left(S_S+\frac{Q}{2}\right)t_{\text{O}}\%P_U \tag{15.4.222}$$

and giving an hyperbolic type curve of cumulative costs:

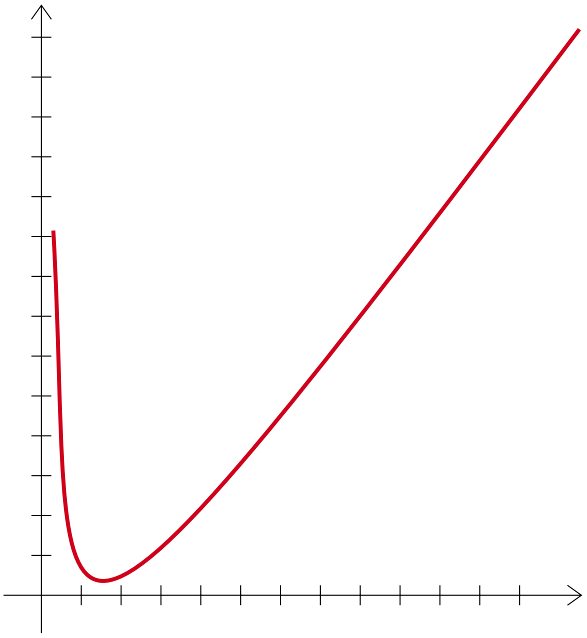


Figure 15.311: **Cumulative costs** $C_{\text{tot}}(Q)$

Seeking the economic quantity Q_e is finding the value of Q for which the total cost is minimal (in other words this is to find the batch size that allows to make the best return on investment of all the costs of storage, manufacturing, etc.), that is to say, mathematically, the value Q_e for which the derivative of the total cost C_{tot} with respect to the quantity Q is zero:

$$\frac{\mathrm{d}C_{\text{tot}}}{\mathrm{d}Q}=-\frac{N}{Q^2}C_L+\frac{1}{2}t_{\text{O}}\%P_U=0 \tag{15.4.223}$$

Hence the "Wilson relation" (after an elementary algebraic computation), also simply named "Wilson formula" for the "optimal economic quantity/batch":

$$\boxed{Q_e=\sqrt{\frac{2NC_L}{t_{\text{O}}\%P_U}}=\sqrt{\frac{2NC_L}{C_{US}}}} \tag{15.4.224}$$

which is therefore for reminder, the batch size that minimizes the total cost (independent of the safety stock which vanishes when we do the derivative). We notice that the economic quantity is proportional to the square root of the consumption. So an item sold 100 more times as expected by a company requires a 10 times greater average stock.

Obviously once known the economic quantity, it becomes easy to calculate the "optimal management cost for period" $C_{\text{tot},e}$ by injecting Q_e in the relation obtained previously:

$$C_{\text{tot},e}(Q)=\left(\frac{N}{Q}C_L\right)+\left(S_S+\frac{Q}{2}\right)t_O\%P_U \tag{15.4.225}$$

and also the "optimal resupply rate or frequency" as given by the ratio:

$$J_S=\frac{N}{Q}=\frac{N}{\sqrt{\dfrac{2NC_L}{C_{US}}}}=\sqrt{\frac{C_{US}N}{2C_L}}=\sqrt{\frac{NP_Ut_O\%}{2C_L}} \tag{15.4.226}$$

We then have after a little bit elementary algebraic manipulations the "optimal cost of management by period" that is given by:

$$\begin{aligned}C_{\text{tot},e}(Q)&=\left(\sqrt{\frac{C_{US}N}{2C_L}}C_L\right)+\left(S_S+\frac{\sqrt{\dfrac{2NC_L}{C_{US}}}}{2}\right)t_O\%P_U=\sqrt{\frac{C_{US}C_LN}{2}}+\frac{\sqrt{\dfrac{2NC_L}{C_{US}}}}{2}C_{US}+S_SC_{US}\\&=\sqrt{\frac{C_{US}C_LN}{2}}+\sqrt{\frac{NC_LC_{US}}{2}}+S_SC_{US}=2\sqrt{\frac{NC_LC_{US}}{2}}+S_SC_{US}=\sqrt{2NC_LC_S}+S_SC_{US}\end{aligned} \tag{15.4.227}$$

where the term:

$$C_{\text{tot},e}(Q)=\sqrt{2NC_LC_S} \tag{15.4.228}$$

is named the "total variable cost".

As often, managers like to make sensitivity analysis (we have already mentioned this in our study of "break-even analysis" page 8047). In the field of Supply Chain Management it is traditional to analyse the sensitivity of the relative change in total storage cost compared to the optimal cost of management by period:

$$\frac{C_{\text{tot}}}{C_{\text{tot},e}}=\frac{\dfrac{N}{Q}C_L+\dfrac{Q}{2}t_O\%P_U+S_St_O\%P_U}{\sqrt{2NC_LC_{US}}+S_SC_{US}}=\frac{\dfrac{N}{Q}C_L+\dfrac{Q}{2}C_{US}+S_SC_{US}}{\sqrt{2NC_LC_{US}}+S_SC_{US}} \tag{15.4.229}$$

But as S_SC_{US} (cost of safety stock) is assumed to be independent of the quantity, it is customary to consider only the relative variation regardless of this term such that (...):

$$\frac{C_{\text{tot}}}{C_{\text{tot},e}}\cong\frac{\dfrac{N}{Q}C_L+\dfrac{Q}{2}C_{US}}{\sqrt{2NC_LC_{US}}} \tag{15.4.230}$$

Finally, it is customary to simplify the analytical expression of this last relation in the following way after some elementary algebraic manipulations:

$$\begin{aligned}\frac{C_{\text{tot}}}{C_{\text{tot},e}}&\cong\frac{\dfrac{N}{Q}C_L+\dfrac{Q}{2}C_{US}}{\sqrt{2NC_LC_{US}}}=\frac{NC_L}{Q\sqrt{2NC_LC_{US}}}+\frac{QC_{US}}{2\sqrt{2NC_LC_{US}}}\\&=\frac{NC_L}{Q\sqrt{2NC_LC_{US}}}\frac{\sqrt{2NC_LC_{US}}}{\sqrt{2NC_LC_{US}}}+\frac{Q}{2\sqrt{\dfrac{2NC_LC_{US}}{C_{US}^2}}}\\&=\frac{\sqrt{2NC_LC_{US}}}{2QC_{US}}+\frac{Q}{2\sqrt{\dfrac{2NC_L}{C_{US}}}}=\frac{\sqrt{2NC_LC_{US}}}{2QC_{US}}+\frac{Q}{2Q_{\text{tot},e}}\\&=\frac{1}{2Q}\sqrt{\frac{2NC_LC_{US}}{C_{US}^2}}+\frac{Q}{2Q_e}=\frac{1}{2Q}\sqrt{\frac{2NC_L}{C_{US}}}+\frac{Q}{2Q_e}=\frac{1}{2Q}Q_e+\frac{Q}{2Q_e}=\frac{1}{2}\left(\frac{Q_e}{Q}+\frac{Q}{Q_e}\right)\end{aligned}\tag{15.4.231}$$

so this is a fairly aesthetic result... that the reader will notice is always positive!

Obviously we can also deduce from the relation of the optimum resupply frequency the "optimal resupply period" then given by (the inverse of the frequency definition):

$$T_e=\frac{1}{f_e}=\frac{1}{J_S}=\frac{1}{\dfrac{Q}{N}}=\sqrt{\frac{2C_L}{C_{US}N}}=\sqrt{\frac{2C_L}{NP_Ut_O\%}}\tag{15.4.232}$$

If we report on a graph the functions:

- The launching cost (supply) in function of the quantities:

$$J_C=\frac{N}{Q}C_L=J_SC_L\tag{15.4.233}$$

- The cost of ownership based in function of the quantities:

$$C_P=t_O\%P_U\bar{S}=t_O\%P_U\left(S_S+\frac{Q}{2}\right)\tag{15.4.234}$$

- The total costs in function of the quantities:

$$C_{\text{tot}}(Q)=J_C+C_P=\left(\frac{N}{Q}C_L\right)+\left(S_S+\frac{Q}{2}\right)t_O\%P_U\tag{15.4.235}$$

the economic quantity is at the intersection of the two curves, launch and possession (when the cost of ownership is equal to the cost of acquisition), or at the inflection point of the total cost curve. In practice, however, it is possible to accurately control the economic quantity, we will choose a batch size that meets the various constraints and included in the "economic area":

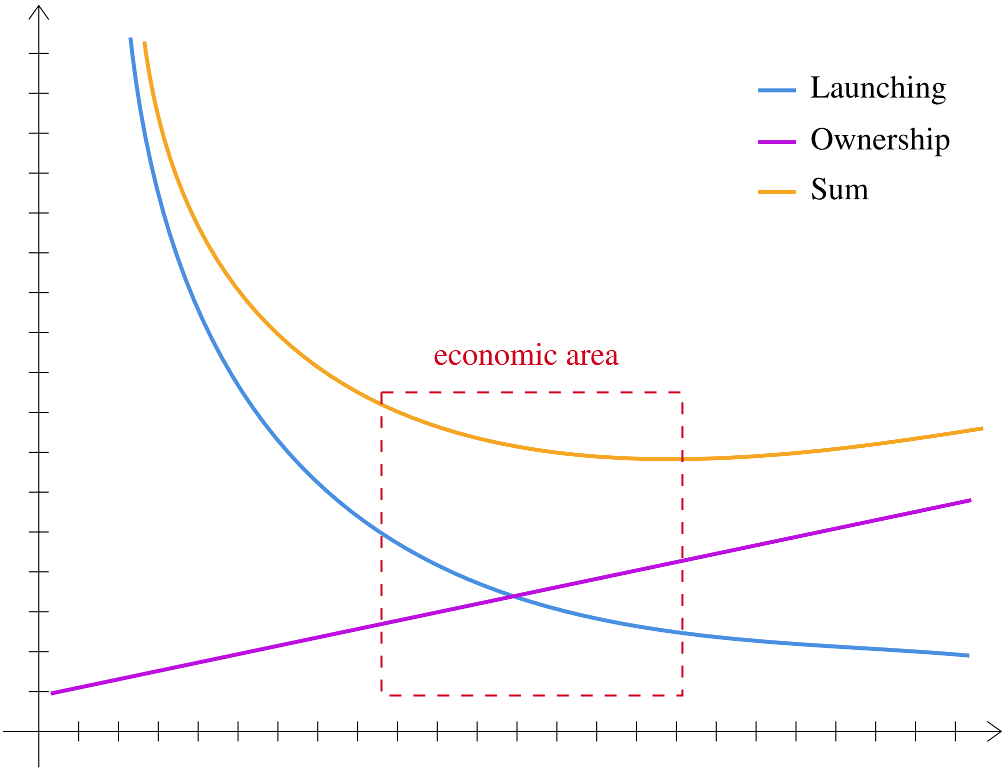


Figure 15.312: **Representation of the economic area**

Obviously in some companies the goal is rather to try to reduce the unit cost of storage C_{US} to reach the equivalent of the demand as economic quantity Q_e. We then have with some elementary algebra:

$$Q_e=\sqrt{\frac{2NC_L}{C_{US}}}\Rightarrow Q_e^2=\frac{2NC_L}{C_{US}}:=N^2 \tag{15.4.236}$$

that is to say the optimal unit storage costs if the supply amount Q is imposed:

$$C_{US,e}=\frac{2C_L}{N} \tag{15.4.237}$$

The reader must know that we often consider that the entire Wilson model can be summarized to the 5 below relations:

$$\boxed{\begin{aligned}&C_{\text{tot},e}(Q)=\sqrt{2NC_LC_{US}}+S_SC_{US}\\&Q_e=\sqrt{\frac{2NC_L}{t_O\%P_U}}=\sqrt{\frac{2NC_L}{C_{US}}}\\&T_e=\sqrt{\frac{2C_L}{NP_Ut_O\%}}=\sqrt{\frac{2C_L}{NC_{US}}}\\&C_{US,e}=\frac{2C_L}{N}\\&\frac{C_{\text{tot}}}{C_{\text{tot},e}}\cong\frac{1}{2}\left(\frac{Q_e}{Q}+\frac{Q}{Q_e}\right)\end{aligned}} \tag{15.4.238}$$

There is another type of scenario that must be studied. If order (resupply) in larger quantities, thus benefiting from a discount, it certainly increases the costs of ownership but theoretically reduces the number of annual orders.

The goal for the manager is of course to check mathematically that the discount granted by the supplier does not lead to induced higher costs than the discount itself (otherwise the manager would be incompetent!).

To do this, we must reduce all costs to a unique item such that the unit total cost will be written:

$$C_U = \frac{C_{\text{tot}(Q)}}{N} + P_U = \left(\frac{C_L}{Q}\right) + \frac{\left(\frac{Q}{2} + S_S\right) t\% P_U}{N} + P_U \tag{15.4.239}$$

This relation is important as it gives the possibility to determine the discount value for the latter to be interesting.

To know the discount threshold R for a given quantity, we replace in the above relation, Q by the target quantity Q' and P_U by $P_U(1-R)$, where R is the discount percentage.

Then we have to solve:

$$C_U \geq \left(\frac{C_L}{Q'}\right) + \frac{\left(\frac{Q'}{2} + S_S\right) t\% P_U(1-R)}{N} + P_U(1-R) \tag{15.4.240}$$

We will determine therefore the value of R under which the discount does not compensate the internal costs.

In the practice we cannot order most of time the exact optimal economic quantity Q_e, especially because of the minimum order quantity (MOQ) imposed by suppliers. It is therefore more interesting to focus on the "economic area", defined by the lower area of the total cost curve.

Because of its simplified assumption and its deterministic aspect, the Wilson model can only provide at best an order of amplitude of the solution if the consumption and/or prices are subject to fluctuations (as the economic quantity is dependent of the parameters: storage costs and resupply cost).

Also it should be noticed that the use of resupply is "anti-flexible" par definition. This type of politic takes frequently important consequences, risk to inflate the stock of finite products, reporting costs and losses downstream of the process.

However, the Wilson's model is interesting for the fact that it can also be applied quite well to human resources in some special cases.

✍**Example:**

The company MAC uses an article X330 for which the projected consumption of the year N should be $4,000$ items. The data are the following:

- The unit cost of Article X330 is $P_U = 8$.- (regardless of currency)
- The procurement/launch cost of an order is $C_L = 100$.-
- The rate of possession of the stock is $t\% = 10\%$ by year.

- The safety stock is of 250 units

The supplier of this article, to encourage its customers to increase the size of their orders, offer to the company the following conditions:

C1. Ordered quantity less than $2,000$ units: unit price $P_U = 8$.-

C2. Ordered quantity less between $2,000$ and $3,500$ units: discount of 2%

C3. Ordered quantity more than $3,500$ units: discount of 3%

The question is to know what is the best choice for the company (buyer)?

The price varies then depending on the quantity such that given a selected quantity, the discount applies on same manner equivalent to all items in the batch (this is named a "uniform discount").

According to the statement and according to the quantity Q of supply, we know that:

1. If $Q < 2,000 \Rightarrow P_U := P_1 = 8$.-
2. If $2,000 < Q < 3,500 \Rightarrow P_U := P_2 = 8 - 2\% \cdot 8 = 7.84$.-
3. If $Q \geq 3,500 \Rightarrow P_U := P_3 = 8 - 3\% \cdot 8 = 7.76$.-

Using the Wilson relation of the optimal economic batch, we will calculate the economic quantity Q_e for the most advantageous price, that is to say P_3:

$$Q_e = \sqrt{\frac{2NC_L}{t_O\% P_U}} = \sqrt{\frac{2 \cdot 4,000 \cdot 100}{7.76 \cdot 10\%}} \cong 1,015 \tag{15.4.241}$$

But to get the price P_3 it must be ordered at least $3,500$ items so there is contradiction and this solution is therefore out of the area! Identical calculations (that we leave to the reader...) show that only the economic quantity Q_e of P_1 satisfies the constraints $Q < 2,000$.

Therefore, the optimal resupply period will be:

$$T_e = \sqrt{\frac{2C_L}{C_S N}} = \sqrt{\frac{2 \cdot 100}{(10\% \cdot 8) \cdot 4,000}} = 0.25 = 91.25 \text{ days} \tag{15.4.242}$$

and the optimal management cost per period is given by:

$$C_{\text{tot},e} = \sqrt{2NC_L C_{US}} + S_S C_S = \sqrt{2 \cdot 4,000 \cdot 100 \cdot (10\% \cdot 8)} + 250(10\% \cdot 8) = 1,000.\text{-} \tag{15.4.243}$$

Caution!!! For the latter calculation, some authors include the cost of safety stock $S_S C_{US}$ of the quantity N. This is a choice rather questionable and that would need to be clarified.

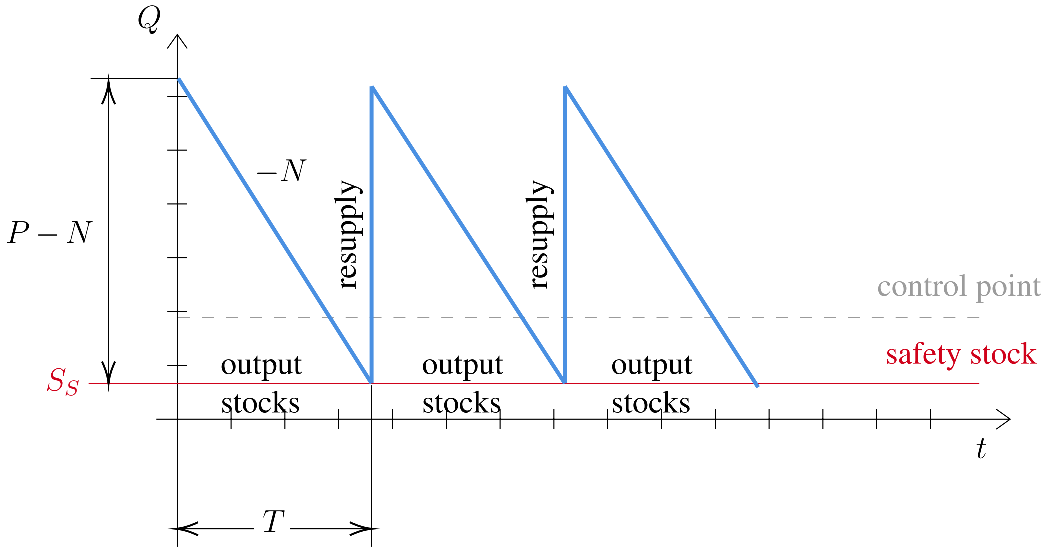
Finally, the impact on the total cost of management if we increase or decrease the amount of $\pm 30\%$ compared to the economic quantity is given by (sensitivity analysis):

$$\begin{aligned}\frac{C_{\text{tot}}}{C_{\text{tot},e}}&=\frac{1}{2}\left(\frac{Q_e}{Q}+\frac{Q}{Q_e}\right)=\frac{1}{2}\left(\frac{1}{1.3}+\frac{1.3}{1}\right)\cong 1.0346\\ \frac{C_{\text{tot}}}{C_{\text{tot},e}}&=\frac{1}{2}\left(\frac{Q_e}{Q}+\frac{Q}{Q_e}\right)=\frac{1}{2}\left(\frac{1}{0.7}+\frac{0.7}{1}\right)\cong 1.0643\end{aligned}\qquad(15.4.244)$$

So obviously, this only increases costs and as we have seen on the graph earlier, the curve being not symmetrical, the two results are not equal.

15.4.4.3.2 Wilson's model without resupply

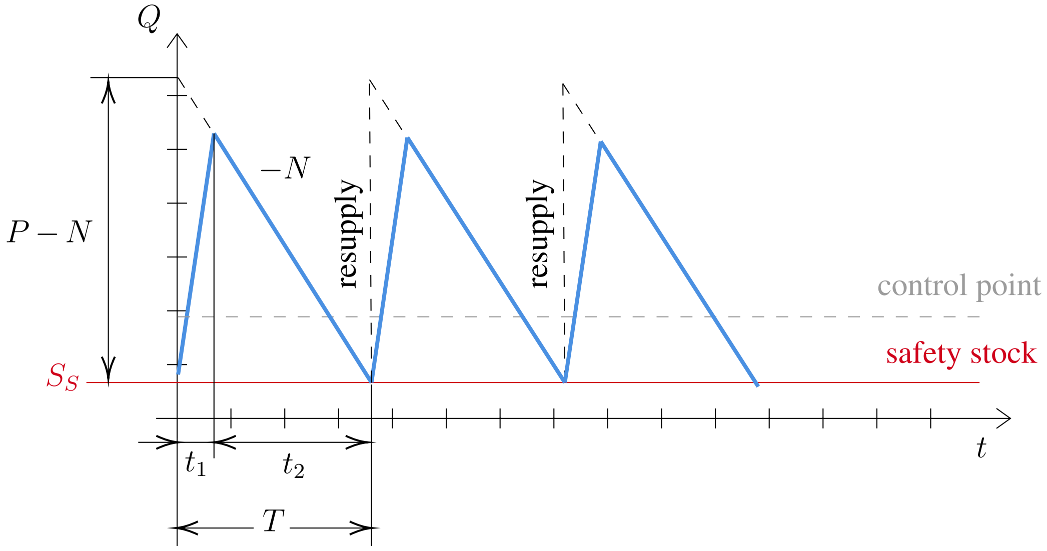
In the previous model, we have assume for recall a sawtooth stock management of the following form:



Q
$P-N$
$-N$
resupply
resupply
control point
safety stock
S_S
output stocks
output stocks
output stocks
t
T

Figure 15.313: **Wilson model with resupply**

In the model that will interests us now, we will consider a supply chain management with a resupply inertia often due to an internal house production such as follows:



Q
$P-N$
$-N$
resupply
resupply
control point
safety stock
S_S
t_1
t_2
t
T

Figure 15.314: **Wilson model without resupply**

As for the Wilson model with resupply, the demand in production is done at the command point. But there is big difference! We can observe that during the production at a rate of P parts per unit of time t_1 the consumption N per period T (reported to the production unit time), is already using the newly created stocks!!!

Do we have to redo all the calculations of the previous model? The answer is: not at all!!! Because if we think a little bit by observing the hyperbolic function of the total cost of storage that we have proved earlier above:

$$C_{\text{tot}}(Q) = J_C + C_P = \left(\frac{N}{Q}C_L\right) + \left(S_S + \frac{Q}{2}\right) t_O\%P_U \tag{15.4.245}$$

It is obvious that the first term due to resupply has no reason to change. The third term, due to safety stock, has no reason to change either! So there remains only the second term represented by the average cost of storage that is necessarily different!

Therefore the average storage is no longer:

$$\frac{Q}{2} \tag{15.4.246}$$

We can easily calculate the average area of the triangle using the Heron's formula available for any triangle and already proved in the section of Geometric Shapes. But the problem in this case is that we want to make the variable Q appear in the average stock! We can then not use Heron's formula. The idea will be the following (refer the previous figure):

When the company don't have any stock more, it launches for a given time t_1 the production at a rate of P elements per unit time until it reaches the cumulative targeted quantity Q. Then we have:

$$Q = Pt_1 \tag{15.4.247}$$

But at the same time production is running (and also after), the company uses (sell) N elements per unit time. The real stock will therefore be at time t_1 (under the obvious condition that P is greater than N) the maximum inventory:

$$\text{Inv}_{\text{max}} = Pt_1 - Nt_1 = (P-N)t_1 \tag{15.4.248}$$

This maximum inventory corresponds therefore to the height of our general triangle (always refer to previous figure) but we wish it to make appear our target stock Q. So, for this, there is now nothing more simple using prior-previous relation:

$$\text{Inv}_{\text{max}} = (P-N)t_1 = (P-N)\frac{Q}{P} = \left(1 - \frac{N}{P}\right) Q \tag{15.4.249}$$

Therefore the average stocks become:

$$\frac{1}{2}\left(1 - \frac{N}{P}\right) Q \tag{15.4.250}$$

So we see that the only difference between the two models is the factor:

$$\left(1 - \frac{N}{P}\right) \tag{15.4.251}$$

Therefore, the relations obtained for the Wilson model with resupply (at least the most important one) and that were for reminder:

$$\begin{aligned}
&C_{\text{tot},e}(Q)=\sqrt{2NC_LC_{US}}+S_SC_{US}\\
&Q_e=\sqrt{\frac{2NC_L}{t_O\%P_U}}=\sqrt{\frac{2NC_L}{C_{US}}}\\
&T_e=\sqrt{\frac{2C_L}{NP_Ut_O\%}}=\sqrt{\frac{2C_L}{NC_{US}}}\\
&C_{US,e}=\frac{2C_L}{N}\\
&\frac{C_{\text{tot}}}{C_{\text{tot},e}}\cong\frac{1}{2}\left(\frac{Q_e}{Q}+\frac{Q}{Q_e}\right)
\end{aligned}\tag{15.4.252}$$

become (the last relations does not change):

$$\boxed{\begin{aligned}
&C_{\text{tot},e}(Q)=\sqrt{2NC_LC_{US}\left(1-\frac{N}{P}\right)}+S_SC_{US}\\
&Q_e=\sqrt{\frac{2NC_L}{\left(1-\dfrac{N}{P}\right)t_O\%P_U}}=\sqrt{\frac{2NC_L}{\left(1-\dfrac{N}{P}\right)C_{US}}}\\
&T_e=\sqrt{\frac{2C_L}{\left(1-\dfrac{N}{P}\right)NP_Ut_O\%}}=\sqrt{\frac{2C_L}{\left(1-\dfrac{N}{P}\right)NC_{US}}}\\
&C_{US,e}=\frac{2C_L}{N\left(1-\dfrac{N}{P}\right)}
\end{aligned}}\tag{15.4.253}$$

The adaptation of the relations of periodic cost of ownership and total cost of supply are also immediate.

Remark

For Q_e and T_e some textbooks use the notation (...):

$$\begin{aligned}
C_{\text{tot},e}(Q)&=\sqrt{2\lambda C_LC_{US}(1-\rho)}+S_SC_{US}\\
Q_e&=\sqrt{\frac{2C_L\lambda}{\left(1-\dfrac{\lambda}{\mu}\right)C_{US}}}=\sqrt{\frac{2C_L\lambda}{(1-\rho)\,C_{US}}}\\
T_e&=\sqrt{\frac{2C_L}{\left(1-\dfrac{\lambda}{\mu}\right)\lambda C_{US}}}=\sqrt{\frac{2C_L}{(1-\rho)\,\lambda C_{US}}}
\end{aligned}\tag{15.4.254}$$

So unlike the model with resupply, we see that only a new variable P appears and that is the "production speed" or "production rate" (elements produced per unit time) to allow a production of N elements per unit time. What the reader must not forget when applying in practice this model is to bring P and N at the same time unit!!!

✍**Example:**

Suppose $C_L = 100$.- per time unit, $C_{US} = 5$.- per time unit, the demand rate is worth $\lambda = 200$ and the production rate is $\mu = 1,000$ and we no have any safety stock ($S_S = 0$). Then we have:

$$\rho = \frac{\lambda}{\mu} = \frac{200}{1,000} = 0.2 \tag{15.4.255}$$

The optimal order quantity is then:

$$Q_e = \sqrt{\frac{2C_L\lambda}{(1-\rho)\,C_{US}}} = \sqrt{\frac{2\cdot 100\cdot 200}{(1-0.2)\cdot 5}} = 100 \tag{15.4.256}$$

and the optimal resupply period will be:

$$T_e = \sqrt{\frac{2C_L}{(1-\rho)\,\lambda C_{US}}} = \sqrt{\frac{2\cdot 100}{(1-0.2)\cdot 200\cdot 5}} = 0.50 \tag{15.4.257}$$

And the optimal management cost per period is given by:

$$C_{\text{tot},e}(Q) = \sqrt{2\lambda C_L C_{US}\,(1-\rho)} + S_S C_{US} = \sqrt{2\cdot 200\cdot 100\cdot 5\,(1-0.2)} + 0\cdot 5 = 400.- \tag{15.4.258}$$

15.4.4.3.3 Wilson's model with resupply and break-up

Still in the classics of academic studies (the rest being only a stochastic mix of different scenarios based on the same reasoning), let us see now the Wilson's model with (instantaneous) resupply but where the break-up and short selling (with penalty counterpart) is allowed (so there is no safety stock anymore!!!). We will represent this scenario by the following saw-tooth figure:

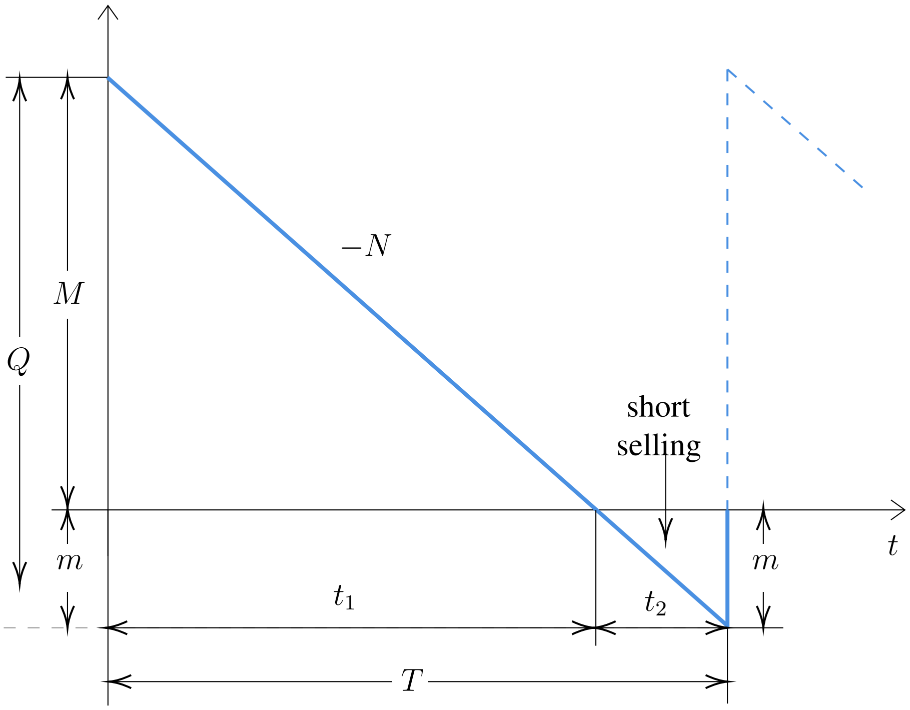


Figure 15.315: Wilson model with resupply and break-up

We will consider C_R as the financial loss (cost) per piece unit during the short-selling period (not having immediate input cash flow makes that we lost money anyway as we can not place it on funds with return yield).

In this case, we have not found a simple way to avoid redoing the calculations of the relations of the Wilson's model with resupply (but if someone has a tip he can contact us!). However it is still elementary algebra and therefore readers should not suffer to much with this kind of exercise style.

We first have trivially by referring to the figure above:

$$\begin{aligned} T &= \frac{Q}{N} \\ t_1 &= \frac{M}{N} \\ t_2 &= T - t_1 = \frac{Q}{N} - \frac{M}{N} = \frac{Q-M}{N} \\ m &= Q - M \end{aligned} \tag{15.4.259}$$

First, the average cost of storage is trivially equal to:

$$\frac{Mt_1}{2} C_{US} = \frac{M\left(\dfrac{M}{N}\right)}{2} C_{US} = \frac{M^2}{2N} C_{US} \tag{15.4.260}$$

The average cost of selling in break-out is also trivially given by:

$$\frac{mt_2}{2} C_R = \frac{(Q-M)\left(\dfrac{Q-M}{N}\right)}{2} C_R = \frac{(Q-M)^2}{2N} C_R \tag{15.4.261}$$

The average total cost for a period is then:

$$C=\frac{M^2}{2N}C_{US}+\frac{(Q-M)^2}{2N}C_R \tag{15.4.262}$$

Let us report this to the time of one period:

$$\frac{C}{T}=\frac{C}{\left(\dfrac{Q}{N}\right)}=\frac{M^2}{2Q}C_{US}+\frac{(Q-M)^2}{2N}C_R \tag{15.4.263}$$

The average total cost per period is then:

$$\begin{aligned}C_t&=\frac{N}{Q}C_L+\frac{M^2}{2Q}C_{US}+\frac{(Q-M)^2}{2Q}C_R\\&=\frac{N}{Q}C_L+\frac{M^2(C_{US}+C_R)}{2Q}+\frac{QC_R}{2}-MC_R\end{aligned} \tag{15.4.264}$$

The problem here is that Q depends on M and vice versa. We'll have to use partial derivatives to fix sometimes one and sometimes the other. Therefore to find the economic quantity we will first arbitrarily determine which is the amount M that will minimize the total average cost per period (in order to get rid of the presence of the variable M in the expression of the quantity economic). We then have taking the partial derivative (thus implying that we maintain Q constant):

$$\frac{\partial C_t}{\partial M}=\frac{M(C_{US}+C_R)}{Q}-C_R=0 \tag{15.4.265}$$

Therefore:

$$M_e=\frac{C_RQ}{C_{US}+C_R} \tag{15.4.266}$$

Now for the economic quantity, we respectively have by taking the partial derivative (so implicitly we fix M):

$$\frac{\partial C_t}{\partial Q}=-\frac{-N}{Q^2}C_L-\frac{M^2(C_{US}+C_R)}{2Q^2}+\frac{C_R}{2}=0 \tag{15.4.267}$$

By injecting in it the prior previous result it comes:

$$-\frac{NC_L}{Q^2}-\frac{\left(\dfrac{QC_R}{C_{US}+C_R}\right)^2(C_{US}+C_R)}{2Q^2}+\frac{C_R}{2}=-\frac{NC_L}{Q^2}-\frac{C_R^2}{2(C_{US}+C_R)}+\frac{C_R}{2}=0 \tag{15.4.268}$$

Thus after some elementary algebraic manipulations:

$$\frac{2NC_L}{Q^2}=\frac{C_RC_{US}}{C_{US}+C_R} \tag{15.4.269}$$

Which gives finally:

$$Q_e=\sqrt{2NC_L\frac{C_{US}+C_R}{C_RC_{US}}}=\sqrt{\frac{2NC_L}{C_{US}}}\sqrt{\frac{C_{US}+C_R}{C_R}} \tag{15.4.270}$$

We then have explicitly:

$$M_e = \frac{C_R}{C_{US}+C_R}Q_e = \frac{C_R}{C_{US}+C_R}\sqrt{2NC_L\frac{C_{US}+C_R}{C_R C_{US}}} = \sqrt{2NC_L\frac{C_R}{C_{US}+C_R}} \tag{15.4.271}$$

The optimal resupply period is then:

$$T_e = \frac{Q_e}{N} = \frac{1}{N}\sqrt{\frac{2NC_L}{C_{US}}}\sqrt{\frac{C_{US}+C_R}{C_R}} = \sqrt{\frac{2C_L}{NC_{US}}}\sqrt{\frac{C_{US}+C_R}{C_R}} \tag{15.4.272}$$

The economic quantity m of rupture gives resupply launch trigger that is then:

$$\begin{aligned} m_e = Q_e - M_e &= \sqrt{\frac{2NC_L}{C_{US}}}\sqrt{\frac{C_{US}+C_R}{C_R}} - \sqrt{\frac{2NC_L}{C_{US}}}\sqrt{\frac{C_R}{C_{US}+C_R}} \\ &= \sqrt{\frac{2NC_L}{C_{US}}}\left(\sqrt{\frac{C_{US}+C_R}{C_R}} - \sqrt{\frac{C_R}{C_{US}+C_R}}\right) \end{aligned} \tag{15.4.273}$$

A practical difficulty of this model is to determine C_R. The idea is then to estimate it on the maximum supposed waiting time of the customer when there is break-up. Starting from:

$$\begin{aligned} t_{2,e} = t_{\text{limit}} = \frac{Q_e - M_e}{N} &= \sqrt{\frac{2C_L}{NC_{US}}}\left(\sqrt{\frac{C_{US}+C_R}{C_R}} - \sqrt{\frac{C_R}{C_{US}+C_R}}\right) \\ &= \sqrt{\frac{2C_L}{NC_{US}}}\left(\sqrt{\frac{(C_{US}+C_R)(C_{US}+C_R)}{C_R(C_{US}+C_R)}} - \sqrt{\frac{C_R C_R}{(C_{US}+C_R)C_R}}\right) \\ &= \sqrt{\frac{2C_L}{NC_{US}}}\left(\sqrt{\frac{(C_{US}+C_R)^2}{C_R(C_{US}+C_R)}} - \sqrt{\frac{C_R^2}{(C_{US}+C_R)C_R}}\right) \\ &= \sqrt{\frac{2C_L}{NC_{US}C_R(C_{US}+C_R)}}\left(\sqrt{(C_{US}+C_R)^2} - \sqrt{C_R^2}\right) \\ &= \sqrt{\frac{2C_L}{NC_{US}C_R(C_{US}+C_R)}}(C_{US}+C_R-C_R) = \sqrt{\frac{2C_L}{NC_{US}C_R(C_{US}+C_R)}}C_{US} \\ &= \sqrt{\frac{2C_L C_{US}}{NC_R(C_{US}+C_R)}} \end{aligned} \tag{15.4.274}$$

And by squaring in the idea of extracting C_R:

$$t_{\text{limit}}^2 = \frac{2C_L C_{US}}{NC_R(C_{US}+C_R)} \Rightarrow C_R(C_{US}+C_R) = \frac{2C_L C_{US}}{Nt_{\text{limit}}^2} \Rightarrow C_R^2 + C_R S_C = \frac{2C_L C_{US}}{Nt_{\text{limit}}^2} \tag{15.4.275}$$

So we finally have an equation of the second degree:

$$C_R^2 + C_R S_C - \frac{2C_L C_{US}}{Nt_{\text{limit}}^2} = 0 \tag{15.4.276}$$

Whose two real roots are (see section Calculus page 1383):

$$C_{R_{1,2}} = \frac{-C_R \pm \sqrt{C_R^2 + 4\frac{2C_L C_{US}}{Nt_{\text{limit}}^2}}}{2} = -\frac{C_R}{2} \pm \sqrt{\left(\frac{C_R}{2}\right)^2 + \frac{2C_L C_{US}}{Nt_{\text{limit}}^2}} \tag{15.4.277}$$

The root to keep is obviously the one that gives a positive cost of rupture. That is to say:

$$C_{R_1}=-\frac{C_R}{2}+\sqrt{\left(\frac{C_R}{2}\right)^2+\frac{2C_LC_{US}}{Nt_{\text{limit}}^2}} \tag{15.4.278}$$

Before moving on to the practical example, the summary is reduced to:

$$\boxed{\begin{aligned}
Q_e&=\sqrt{\frac{2NC_L}{C_{US}}}\sqrt{\frac{C_{US}+C_R}{C_R}}\\
M_e&=\sqrt{\frac{2NC_L}{C_{US}}}\sqrt{\frac{C_R}{C_{US}+C_R}}\\
T_e&=\sqrt{\frac{2NC_L}{C_{US}}}\left(\sqrt{\frac{C_{US}+C_R}{C_{US}+C_R}}+\sqrt{\frac{C_R}{C_{US}+C_R}}\right)\\
t_{\text{limit}}&=\sqrt{\frac{2C_LC_{US}}{NC_R(C_{US}+C_R)}}\\
C_R&=-\frac{C_R}{2}+\sqrt{\left(\frac{C_R}{2}\right)^2+\frac{2C_LC_{US}}{Nt_{\text{limit}}^2}}
\end{aligned}} \tag{15.4.279}$$

✍**Example:**

The company MAC supplies the article X330. In the objective to not to lose customers, the managers decide of an optimal break-up time (meaning: maximum) of 2 days. We have the following data:

$$C_L=200\left[\frac{\$}{\text{resupply}}\right],\; C_{US}=1\left[\frac{\$}{\text{unit}\cdot\text{day}}\right],\; t_{\text{limit}}=2\,[\text{day}],\; N=5\left[\frac{\text{units}}{\text{day}}\right] \tag{15.4.280}$$

Which gives:

$$\begin{aligned}
C_R&=200\left[\frac{\text{units}}{\text{day}}\right]\\
Q_e&=50\;[\text{units}]\\
M_e&=40\;[\text{units}]\\
m_e&=10\;[\text{units}]\\
T_e&=10\;[\text{days}]
\end{aligned} \tag{15.4.281}$$

15.4.5 Queueing Theory

The queueing theories are an extremely powerful and extensive tool (a complete presentation requires at a minimum 300 A4 pages) that gives the possibility to consider and model the bottlenecks in business processes or in logistics, phone (Telecom) exchanges, requests on servers, at the check-out line of shops, or in the toilet of major sports stadiums (...), and also for ticket distribution for huge events, etc. depending on the starting assumptions and constraints. There are numerous example in newspaper of companies or events having very bad reputation because their manager did not use such modelization tools (or consultants knowing these tools....) during a given period of their existence and that have generate a huge frustration by the customers (reputation very hard to correct afterwards!).

Typically, most customers clearly don't like waiting non-value added activity, and if they wait too long, they combine this waste of time to a poor quality of service and share their experience on social networks and this can have afterwards a huge impacts (through other medias such at the television). Similarly, within companies, unoccupied employees or unused equipment represent non-value added activities. To avoid these situations, some of the best of companies and governmental institutions have implemented continuous improvement process whose ultimate goal is the elimination of all forms of waste, including waiting. All these examples show the importance of the analysis of queues. To notice that if a company or administration does not have the financial or intellectual capacity to make use of the queuing theories, the minimum respect for customers is at least to indicate with an inexpensive display, the median waiting time and to provide during this waiting times some entertainment (television, newspapers, magazines, food, drinks, etc.)!

Here is a typical small list of common systems relatives to queuing theory:

System	Customers	Services
Counter-reception	People	Receptionists
Repair workshop	Machines	Technicians
Customs	People	Customs officers
Ticketing	People	Vendors
Garage	Trucks/Cars	Mechanicians
Hospital	Patients	Nursing
Computer	Tasks	Processors, HDD, BUS
Airport	Airplane	Traffic controller
Roads	Cars	Traffic lights
Manufacturing	Tasks	Machines/Workers
Phone	Calls	Exchangers
Public transport	Travellers	Bus/Metro
Washroom	Lines	Washers / Dryers
Safety Check	People	Security guards
...	...	...

Table 15.82: **Examples of typical queues**

Remark

Queuing theory, is a subfield of "discrete-event simulation" (DES) that models the operation of a system as a (discrete) sequence of events in time. Each event occurs at a particular instant in time and marks a change of state in the system. Between consecutive events, no change in the system is assumed to occur; thus the simulation time can directly jump to the occurrence time of the next event, which is named "next-event time progression". DES is a computational science, a mix of theory and experimentation with a computer as the primary piece of laboratory equipment. In other words, discrete-event simulation is a form of computer-aided model building and problem solving. The goal is insight, a better understanding of how systems operate and respond to change (see our `R` companion book for a practical example).

These theories proves to be particularly useful to justify investments, hiring or equipment purchases. More generally, it is an integral part of mathematics management techniques when necessary to find an economic optimum between waiting costs and a system service costs.

The typical problem in companies and governmental administration can be expressed as:

- The human brain is subject to cognitive bias so not accurate enough to take a good decision hence the need of mathematical tool in this field
- What is the optimal number of stations / terminals to be commissioned to process all applications, while avoiding an excessive queue and the departure of some "customers"?
- What is the average/median waiting time of a "customer" in front the station / terminal?
- What is the average/median number of "customers" waiting in the queue?

These questions express the goals of quality and level of service that are:

- An average or median waiting time in the queue or service that should not to be exceed
- A maximum probability of waiting time
- An average/median number of waiting customers

In practice, when the customer is external to the company/administration, the waiting cost is difficult to evaluate because it is an impact rather than a cost that can be accounted (but sure the cost is not null!!!). However, the waiting time can be regarded as a measurement standard service level. The manager decides the acceptable waiting time and he sets up the capacity that can provide this level of service.

When the customer is internal to the company/administration, we can evaluate some costs directly related customer waiting time (machines/information). However we must not conclude too quickly that the cost of an employee that is waiting is equal to his wage during the waiting time (even if this can be true for a small proportion of people...). This would imply that the net decrease in earnings of the company/administration, due to the inactivity of an employee is equal to the wage of the latter, which, a priori, is not easy to prove (excepted for IT or auditing companies that have for only purpose to put consultant into contracts). The employee, whether he works or wait will receive most of time the same salary (absurdity but than can be corrected with modern communication technologies). By cons, its contribution to the earnings of the company is really lost as its productivity is lower. When a machine operator is inactive because he is waiting its productive force (which may include, in addition to his wage, a proportion of the company's fixed costs) is lost. In other words, we must consider not only the physical resource that is on hold, but the value (cost) of all inactive economic resources, and then estimate the loss of profit from this lost productivity. Then the objective of the analysis of queues is to find a compromise between the cost associated with the service capacity and cost of waiting customers.

These theories therefore rely on statistics and algebraic methods that we have studied in the sections of Statistics and Graphs Theory. They are just as most exciting.

To introduce the topic, rather than doing an abstract generalization, we chose to develop the theory around a concrete and classic example that is the Teletraffic. A generalization to other case studies then being relatively easily by analogy.

Let us consider, therefore, a central telephone system regrouping the lines of a group of buildings in a city and not having as many lines going to the network as lines going to the different individuals it serves.

So we can legitimately ask how many lines we need to serve all those subscribers?

To dimension its network, an operator will have to calculate the number of resources to implement so that with a probability extremely close to 1, a user who picks up his telephone can have a circuit (connection). To do this, it will be necessary to develop some relations of probability of blocking. These relations will require statistical modelling of the start and end times of calls as well as the durations of these calls. The following paragraphs will introduce the probability laws used to dimension such networks.

Finally, before we begin, we would like to make available the following summary table of the most important notations that the reader will discover as he reads the developments below and to which he can refer in case of confusion:

Variables	**Information**	**Unit**
λ	Incoming flow of customers in the queue (communication), also named "average arrival rate of calls", or "average arrival frequency". The inverse λ^{-1} gives the average time between arrivals (calls) in the queue.	$[h^{-1}]$
μ	Customer exit flow (communication) corresponding to the processing rate. The inverse μ^{-1} gives the average waiting time during the service (thus once arrived at the end of the queue) also named "average service time".	$[h^{-1}]$
A or ρ	Service utilization rate (per server unit). Assimilated to the concept of "traffic" (somewhat abusively) or "load". Is equal to the ratio λ/μ and must be strictly less than 1 to avoid congestion.	-
C	Total number of customers in the system	-
C_Q	Number of customers waiting in the queue	-
C_S	Number of customers in service (processing)	-
T	Waiting time in the system	[s]
T_Q	Waiting time in the queue	[s]
p_k	Probability of having k customers in the system	-

Table 15.83: **Conventional notations in Queuing Theory**

And let us specify that the parameters to be taken into account are often:

- The type of customer arrival process (data)
- The statistical distribution of service time (processing)
- The number of servers (checkouts)
- The capacity of the system (often assumed to be infinite in practice ...)
- The size of the population
- The service discipline

15.4.5.1 $M/M/\ldots$ arrival times modelisation

In this "$M/M/\ldots$ model" whose origins of notation will be explained further below, we consider calls that start randomly. Let us then take a time interval t and divide this interval into n subintervals of duration t/n.

We choose n large enough for the following assumptions (hypothesis) to be respected:

H1. Only one incoming call can occur in an interval t/n

H2. Call arrival times are independent of each other (the arrival rate is not influenced by the number of calls from the population). This presupposes an infinite population.

H3. The probability that a call arrives in a given subinterval is proportional to a constant factor close to the duration of the subinterval.

We then write the probability p_1 of one call in a subinterval (1) as follows:

$$p_1(1)=\lambda\frac{t}{n} \qquad (15.4.282)$$

Where the 1 in index of the p represents the analysis on 1 call, the 1 between brackets the fact that the analysis is done on 1 subinterval and finally the term λ represents the coefficient of proportionality between the probability and the duration t/n of the subinterval.

The starting hypothesis of considering as zero the probability of having several calls in a subinterval is then written:

$$p_2(1)+p_3(1)+\cdots+p_n(1)+\cdots=\sum_{k=2}^{+\infty}p_k(1)=0 \qquad (15.4.283)$$

The probability of having no call during a sub-interval of time t/n is thus written:

$$p_0(1)=1-\sum_{k=1}^{+\infty}p_k(1) \qquad (15.4.284)$$

By developing, we get:

$$p_0(1)=1-p_1(1)-\sum_{k=2}^{+\infty}p_k(1) \qquad (15.4.285)$$

and using the property defined just above:

$$p_0(1)=1-p_1(1) \qquad (15.4.286)$$

The probability of having k calls arriving during n time intervals is then obtained by considering the number of ways to choose k intervals among n... (since there can not be more than one call per interval).

For each of these solutions, we will then necessarily have k intervals with a single incoming call and $n-k$ intervals with no incoming call. We have seen in the section Statistics that the law which made it possible to get the probability of choosing a certain arrangement of binary issues from a total number of issues was Bernoulli's law given by:

$$P=p^kq^{n-k} \tag{15.4.287}$$

It therefore comes in our case that the probability of one of the solutions will be:

$$p_1(1)^kp_0(1)^{n-k} \tag{15.4.288}$$

The global probability is obtained by summing the probabilities of all the cases, which gives us the binomial law (see section Statistics page 658):

$$p_k(n)=\frac{n!}{k!(n-k)!}p_1(1)^kp_0(1)^{n-k}=C_k^np_1(1)^kp_0(1)^{n-k} \tag{15.4.289}$$

Or, by replacing the probabilities by their values as a function of λ, t, n:

$$p_k(n)=C_k^n\left(\lambda\frac{t}{n}\right)^k\left(1-\lambda\frac{t}{n}\right)^{n-k} \tag{15.4.290}$$

The limit of the probability $p_k(n)$ when n tends to infinity will be equal to the probability of having k call arrivals during a time interval t. We denote by p_k this probability:

$$p_k=\lim_{n\to+\infty}p_k(n) \tag{15.4.291}$$

Then, taking up the different terms of the expression of $p_k(n)$, it comes:

$$\left(1-\lambda\frac{t}{n}\right)^{n-k}=\left(e^{\ln\left(1-\lambda\frac{t}{n}\right)}\right)^{n-k}=e^{(n-k)\ln\left(1-\lambda\frac{t}{n}\right)} \tag{15.4.292}$$

Using the Taylor developments (see section Sequences and Series page 1703):

$$\ln(1+x)=\sum_{n=1}^{+\infty}\frac{(-1)^{n+1}}{n}x^n \tag{15.4.293}$$

Either by taking only the first term, that is, considering x very small:

$$\ln(1+x)\cong x \tag{15.4.294}$$

Therefore:

$$e^{(n-k)\ln\left(1-\lambda\frac{t}{n}\right)}\underset{n\to+\infty}{\cong}e^{(n-k)\left(-\lambda\frac{t}{n}\right)}=e^{-\lambda t-\frac{k}{n}\lambda t}\underset{n\to+\infty}{\cong}e^{-\lambda t} \tag{15.4.295}$$

And for the last part:

$$\begin{aligned}C_k^n\left(\frac{\lambda t}{n}\right)^k&=\frac{n!}{k!(n-k)!}\frac{(\lambda t)^k}{n^k}=\frac{(\lambda t)^k}{k!}\frac{n!}{(n-k)!n^k}\\&=\frac{(\lambda t)^k}{k!}\frac{n(n-1)(n-2)\ldots(n-k+1)}{n^k}\underset{n\to+\infty}{\cong}\frac{(\lambda t)^k}{k!}\end{aligned} \tag{15.4.296}$$

hence after grouping:

$$\boxed{p_k\cong\frac{(\lambda t)^k}{k!}e^{-\lambda t}} \tag{15.4.297}$$

This distribution function is therefore extremely important, since it represents the probability of observing k calls arriving in an interval of duration t (or the probability of observing k customers that are in front of a checkout or server in an interval of duration t) and it is therefore a Poisson distribution (see section Statistics page 686). In practice, it is therefore necessary to ensure before using the relation that will follow in the further developments that this distribution is well respected (with a Chi-square test typically).

Let us consider now the distribution function, named the "Erlang distribution function", which for a given time interval t gives the cumulative probability of having a number k of arrivals greater than or equal to n:

$$F(t,k)=\sum_{k=n}^{+\infty}\frac{(\lambda t)^k}{k!}e^{-\lambda t} \tag{15.4.298}$$

If we derive this function with respect to t this gives the "Erlang distribution function" (the reader may have noticed perhaps that in the penultimate line we thus find an example for the field of Management of the concept of "telescoping series" which had been introduced in the section of Sequences and Series at page 1691):

$$\begin{aligned}\frac{\mathrm{d}F(t,k)}{\mathrm{d}t}&=\frac{\mathrm{d}}{\mathrm{d}t}\sum_{k=n}^{+\infty}\frac{(\lambda t)^k}{k!}e^{-\lambda t}=\sum_{k=n}^{+\infty}\frac{k\lambda(\lambda t)^{k-1}}{k!}e^{-\lambda t}-\sum_{k=n}^{+\infty}\frac{(\lambda t)^k}{k!}\lambda e^{-\lambda t}\\&=\sum_{k=n}^{+\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}\lambda e^{-\lambda t}-\sum_{k=n}^{+\infty}\frac{(\lambda t)^k}{k!}\lambda e^{-\lambda t}=\lambda e^{-\lambda t}\left(\sum_{k=n}^{+\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}-\sum_{k=n}^{+\infty}\frac{(\lambda t)^k}{k!}\right)\\&=\lambda e^{-\lambda t}\sum_{k=n}^{+\infty}\left(\frac{(\lambda t)^{k-1}}{(k-1)!}-\frac{(\lambda t)^k}{k!}\right)=\frac{(\lambda t)^{n-1}}{(n-1)!}\lambda e^{-\lambda t}\end{aligned} \tag{15.4.299}$$

and thus we fall back on a particular form of the distribution function of the Gamma law introduced in the section of Statistics (see page 767). Let us recall the equivalence already presented in the section of Statistics (see page 773) with Microsoft® Excel® 14.0.6123 between the Poisson law, the Chi-2 law and the Gamma law:

=POISSON.DIST($x\in\mathbb{N}$,μ,TRUE)
=1-CHISQ.DIST($2*\mu$,$2*(x+1)$,TRUE)
=1-GAMMA.DIST(μ,$x+1$,1TRUE)

It follows by analogy with the general form of the Poisson law that the parameter λ is the average arrival rate (mean rate of calls) per unit of time (or "Poisson Arrivals See Time Average": PASTA ...). Typically this will be an average number of calls per second (see the Poisson law estimators in the section Statistics at page 825).

Thus, we have for the mean and variance (see section Statistics page 686) the number of calls:

$$\boxed{\mathrm{E}(k)=\lambda\cdot t\qquad\mathrm{V}(k)=\lambda\cdot t} \tag{15.4.300}$$

When the model consider a constant event occurrence rate λ we then speak of "homogeneous Poisson process" (HPP).

✍**Example:**

A very small company wishing to set up a hotline estimates that at the beginning it will receive 8 phone calls per day (i.e. a probability of 1 chance out of 2 to have one call per hour and therefore an average rate λ of 0.5 calls per hour). Then the probability that it receives exactly 4 calls (k) per day and at least 4 calls (k) per day according to the theoretical model of the queuing theory is:

$$\begin{aligned}
p_k(k=4) &\cong \frac{(0.5\cdot 8)^4}{4!}e^{-0.5\cdot 8}=\texttt{POISSON.DIST(4,0.5*8,FALSE)}=19.55\%\\
p_k(k\leq 4) &\cong \sum_{k=0}^{4}\frac{(0.5\cdot 8)^k}{k!}e^{-0.5\cdot 8}=\texttt{POISSON.DIST(4,0.5*8,FALSE)}=62.88\%
\end{aligned}\tag{15.4.301}$$

Where we used the function `POISSON.DIST()` embedded in Microsoft® Excel® 14.0.6123.

Now let us introduce the random variable τ representing the time between two call arrivals. We define for this the probability $A(t)$ which is the probability that the time τ is less than or equal to a value t:

$$A(t)=P(\tau\leq t)\tag{15.4.302}$$

We therefore have:

$$A(t)=1-P(\tau>t)\tag{15.4.303}$$

But, $P(\tau>t)$ represents the cumulative probability that there is no arrival of calls during a time t. This cumulative probability was precisely established above:

$$P(\tau>t)=p_0=\frac{(\lambda t)^0}{0!}e^{-\lambda t}=e^{-\lambda t}\tag{15.4.304}$$

From this we deduce:

$$A(t)=1-e^{-\lambda t}\tag{15.4.305}$$

It is therefore the function of distribution of an exponential law! We can also introduce the probability density of the random variable τ. We thus get:

$$a(t)=\frac{\partial A(t)}{\partial t}=\lambda e^{-\lambda t}\tag{15.4.306}$$

Remark

Let us recall that in the section Statistics, we often did the opposite reasoning. That is to say, given a probability density $a(t)$, we sought the distribution function $A(t)$ through an integration on the domain of definition of the random variable.

The probability density thus makes it possible to calculate the expected mean duration between two call arrivals:

$$\bar{\tau}=\mathrm{E}(\tau)=\int\limits_0^{+\infty}t\cdot a(t)\mathrm{d}t=\int\limits_0^{+\infty}\lambda t e^{-\lambda t}\mathrm{d}t\tag{15.4.307}$$

By integrating by parts, it comes:

$$\mathrm{E}(\tau)=\lambda t\cdot\frac{-1}{\lambda}e^{-\lambda t}\Bigg|_0^{+\infty}+\int\limits_0^{+\infty}\lambda\frac{-1}{\lambda}e^{-\lambda t}\mathrm{d}t=-\int\limits_0^{+\infty}e^{-\lambda t}\mathrm{d}t=\frac{1}{\lambda}\tag{15.4.308}$$

We thus obtain that, for an arrival rate of λ calls, that the expected mean time between calls is equal to $1/\lambda$ (a relatively logical result but still had to be rigorously proved). Indeed, if we have λ that is worth 2 calls per hour, the average arrival time is indeed 0.5 hours $(1/2)$ between calls.

Let us now suppose that no call has arrived until a time t_0 and we want to calculate the probability that a call arrives during a time t after the time t_0

We must therefore calculate the probability of having a duration between two calls less than $t+t_0$ while being greater than t_0

This probability is written $P(\tau\leq t+t_0|\tau>t_0)$. Using the Bayes formula (see section Probabilities page 492):

$$P(B/A)=\frac{P(B\cap A)}{P(A)}\tag{15.4.309}$$

But with the appropriate notations, it comes:

$$P(\tau\leq t+t_0|\tau>t_0)=\frac{P(t_0<\tau\leq t+t_0)}{P(\tau>t_0)}\tag{15.4.310}$$

This probability can also be written:

$$P(\tau\leq t+t_0|\tau>t_0)=\frac{P(\tau\leq t+t_0)-P(\tau\leq t_0)}{P(\tau>t_0)}=\frac{P(\tau\leq t+t_0)-P(\tau\leq t_0)}{1-P(\tau\leq t_0)}\tag{15.4.311}$$

By taking the expressions of the different probabilities:

$$\begin{aligned}P(\tau\leq t+t_0|\tau>t_0)&=\frac{\left(1-e^{-\lambda(t+t_0)}\right)-\left(1-e^{-\lambda t_0}\right)}{1-\left(1-e^{-\lambda t_0}\right)}=\frac{-e^{-\lambda(t+t_0)}+e^{-\lambda t_0}}{e^{-\lambda t_0}}\\&=\frac{-e^{-\lambda t}e^{-\lambda t_0}+e^{-\lambda t_0}}{e^{-\lambda t_0}}=-e^{-\lambda t}+1=1-e^{-\lambda t}\end{aligned}\tag{15.4.312}$$

We thus see that the probability of occurrence of a call during a time t after a duration t_0 during which none has occurred is the same as the probability of occurrence of a call for a duration t, independent of what happened before. We therefore consider that the phenomenon (the exponential law) is "without memory".

15.4.5.2 $M/M/\ldots$ service times modelisation

In this model $M/M/\ldots$ for which we will explain again the origin of the notation later, we will study the probability laws that models the durations of the calls (implicitly: in service once the end of the queue reached). To do this, we proceed as before.

We therefore consider a time interval of duration t which we decompose into n subintervals of duration t/n. We choose n so that the following assumptions remain justified:

H1. The probability that a call terminates during a subinterval is proportional to the duration of that subinterval.

We will denote:

$$\mu\frac{t}{n} \tag{15.4.313}$$

this probability, in which μ represents the coefficient of proportionality.

H2. The probability of a call terminating during a subinterval is independent of the subinterval considered.

We then introduce the random variable θ representing the duration of a call and the probability $H(t)$that the duration of a call is less than or equal to t:

$$H(t)=P(\theta\leq t) \tag{15.4.314}$$

The probability that a call that started at $t=0$ does not end before t is then written:

$$P(\theta>t)=1-H(t) \tag{15.4.315}$$

This probability is equal to the probability that the call does not end in any of the n subintervals of duration t/n:

$$1-H(t)=\left(1-\mu\frac{t}{n}\right)^n \tag{15.4.316}$$

By taking n towards infinity, we get:

$$1-H(t)=\lim_{n\rightarrow+\infty}\left(1-\mu\frac{t}{n}\right)^n=\lim_{n\rightarrow+\infty}e^{n\ln\left(1-\mu\frac{t}{n}\right)}\cong\lim_{n\rightarrow+\infty}e^{n\left(1-\mu\frac{t}{n}\right)}=e^{-\mu t} \tag{15.4.317}$$

We thus obtain the expression of the probability that a call has a duration less than or equal to t:

$$H(t)=1-e^{-\mu t} \tag{15.4.318}$$

We can deduce the associated probability density, denoted $h(t)$:

$$\boxed{h(t)=\frac{\partial H(t)}{\partial t}=\mu e^{-\mu t}} \tag{15.4.319}$$

which corresponds to an exponential law (the time that a customer use to be served at a check-out or server - period of service - or so that his call is processed once at the end of the queue, follows then an exponential law!). In practice, it is therefore necessary to ensure before using the relations that follow that this distribution is well respected (with a Chi-square test typically).

In the same way as in the preceding paragraphs, the average call duration (average time between two ends of calls in service) is obtained by calculating (always the same integration by parts as above):

$$\bar{\theta}=\mathrm{E}(\theta)=\int\limits_{0}^{+\infty} th(t)\mathrm{d}t=\frac{1}{\mu} \tag{15.4.320}$$

In conclusion, we have μ calls that cease per second and we have an expected mean call duration in service equal to:

$$\frac{1}{\mu} \tag{15.4.321}$$

Indeed, if we have, for example, 2 calls in-service that stop per hour, this gives us an average time of 0.5 hour $(1/2)$ of service per call.

The ratio:

$$A=\frac{\lambda}{\mu} \tag{15.4.322}$$

represents therefore the number of calls that appear in the queue on the number of calls in service that end during the same time interval (mean service time), that is, it simply represents the traffic (or in other words, the "traffic intensity") or from another point of view the average use of the service whose unit is the "Erlang" (we will recall this definition several times thereafter...).

✍**Example:**

In a store, there are 240 customers per hour and it takes an average of 28 seconds to process a customer (average service time). Knowing that the length of service follows an exponential law and the distribution of arrivals a Poisson law, what is the intensity of the traffic and the rate of occupation of the check-out stations if the store has only two of them?

The traffic is therefore given by the number of customers per hour divided by the number of customers processed per hour. As it takes 28 seconds to process one and there are $3,600$ seconds in an hour, we have the following traffic intensity:

$$A=\frac{240}{\dfrac{3'600}{28}}=1.866\,[\,\text{Erlangs}] \tag{15.4.323}$$

The average load (or occupancy rate) per check-out station (server), knowing that there are two of them, is therefore:

$$\frac{1.866}{2}=0.93=93\% \tag{15.4.324}$$

It is this latter value which will ultimately be taken as the value of traffic A per station (check-out/server) for subsequent calculations.

The probabilities of appearance of calls and end of calls that have been developed in the preceding paragraphs make it possible to model the complete process of appearance and end of calls.

15.4.5.3 Kendall queues notation

A notation was developed by D. G. Kendall in 11953 (holocene calendar) to represent the queues following the intensive (and numerous) developments of the mathematical models concerning them. The reduced form of this notation named "Kendall notation" or "Kendall-Lee notation" is:

$$\boxed{A/B/C}$$

where:

- A is the customer arrival process in the system (distribution interarrival law)
- B is the distribution of the service time of the customers in the system (distribution law)
- C is the number of servers in the system.

Common notations for A are:

- M for Markovian or memoryless arrival Poisson time process
- G general distribution (any arbitrary distribution allowed)
- E_k Erlang distribution

Common notations for B are:

- M for Markovian or memoryless arrival exponential time process
- G general distribution (any arbitrary distribution allowed)
- E_k Erlang distribution

For example, the $M/M/1$ notation means that the clients arrive in the system according to a Poisson distribution modelled by a Markov chain (see section Probabilities page 537), that the processing time follows an exponential distribution (modelled by a Markov chain as well) and the system consisting of a single server based on a first-in first-out principle (FIFO) in an infinite-population and steady-state queue. This corresponds respectively to the three following relations which we have proved above for the probability of arrival of k calls in a given time:

$$p_k \cong \frac{(\lambda t)^k}{k!}e^{-\lambda t} \tag{15.4.325}$$

The probability that the processing time (service time) is equal to a certain value:

$$h(t)=\mu e^{-\mu t} \tag{15.4.326}$$

and the probability of having k customer (communications):

$$p_k=A^k(1-A) \tag{15.4.327}$$

In general, we use the notation:

$$\boxed{A/B/C:d/e/f}$$

where:

- d is the capacity of the system, or the maximum number of customers allowed in the system including those in service. When the number is at this maximum, further arrivals are turned away. If this number is omitted, the capacity is assumed to be unlimited, or infinite.
- e is the service discipline or priority order that jobs in the queue, or waiting line, are served (see examples further below)
- f is the size of calling source. The size of the population from which the customers come. If this number is omitted, the population is assumed to be unlimited, or infinite.

The parameter of queue's discipline e has typically the following properties:

- FIFO/FCFS (First In First Out/First Come First Served): The customers are served in the order they arrived in.
- LIFO/LCFS (Last in First Out/Last Come First Served): The customers are served in the reverse order to the order they arrived in.
- SIRO (Service In Random Order): The customers are served in a random order with no regard to arrival order.
- SJF (Shortest Job First): The customer with shortest ranked treatments are served before
- ...

If this parameter is omitted, the discipline is assumed to be a FIFO one.

Thus, $M/M/1$ would be written without omitting anything in our study case above:

$$M/M/1:\infty/\text{FIFO}/\infty$$

We also find variants in the literature of the definition above which does not always help in the reading of various textbooks about the subject (absence of ISO standards doesn't help)...

15.4.5.4 Modelling of arrivals and departures $M/M/1$

At every instant a certain number of calls will appear and others will end. We can therefore model the state where we are at a given moment as a chain of states.

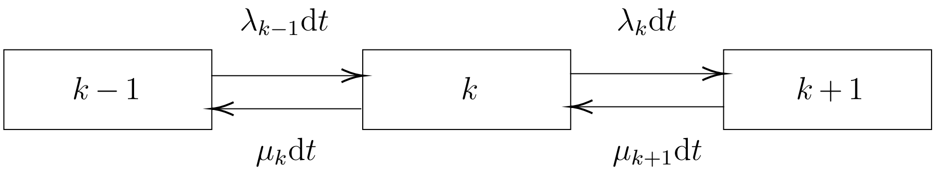
Each state represents the number of calls in progress. We therefore conceive that if at a given instant there are k calls, we can pass only in two adjacent states according to our hypotheses: $k-1$ and $k+1$.

We then recognize a Markov chain (see section Probabilities page 537). The probability of going from a state i to a state j during a time $\mathrm{d}t$ will therefore be denoted:

$$p_{ij}(\mathrm{d}t) \tag{15.4.328}$$

We then introduce the following state transition probabilities:

- Being in state k, the probability $p_{k,k+1}(\mathrm{d}t)$ to go to the state $k+1$ during a time interval $\mathrm{d}t$ will be denoted $\lambda_k \mathrm{d}t$
- Being in state k, the probability $p_{k,k-1}(\mathrm{d}t)$ to go to the state $k-1$ during a time interval $\mathrm{d}t$ will be denoted $\mu_k \mathrm{d}t$
- Being in the state $k+1$, the probability $p_{k+1,k}(\mathrm{d}t)$ to pass to the state k during a time interval $\mathrm{d}t$ will be denoted $\mu_{k+1}\mathrm{d}t$
- Being in the state $k-1$, the probability $p_{k-1,k}(\mathrm{d}t)$ to go to state k during a time interval $\mathrm{d}t$ will be denoted $\lambda_{k-1}\mathrm{d}t$



The quantities λ_k and μ_k are rates of arrival (appearance) and departure (end) of calls of the same type as those used in the preceding paragraphs. The only difference is that these rates have in index the state in which the system is.

We can then introduce the state probability, that is, the probability of being in a state k at time t. Let us denoted $p_k(1)$ this probability (to be compared with the notation $p_j(n)$ used for the discrete-time Markov chains in the section of Probabilities).

The variation of this probability during a time interval $\mathrm{d}t$ is then equal to the probability of joining this state from a state $k-1$ or a state $k+1$ minus the probability of leaving this state to go to a state $k-1$ or to a state $k+1$.

What is written:

$$\frac{\mathrm{d}p_k(t)}{\mathrm{d}t} = \lambda_{k-1}p_{k-1}(t) + \mu_{k+1}p_{k+1}(t) - (\lambda_k + \mu_k)\,p_k(t) \tag{15.4.329}$$

Assuming the system stable, that is, assuming that it stabilizes on fixed state probabilities when time tends to infinity, we can write that:

$$\lim_{t\to+\infty}\frac{\mathrm{d}p_k(t)}{\mathrm{d}t}=0 \tag{15.4.330}$$

We can then denote $p_k=p_k(t)$ hence finally:

$$\lambda_{k-1}p_{k-1}(t)+\mu_{k+1}p_{k+1}(t)-(\lambda_k+\mu_k)\,p_k(t)=0 \tag{15.4.331}$$

We could have introduced this last relation in another way: It simply expresses the fact that the probability of starting from a state is equal to that to arrive at it (it is perhaps more simple like this).

This relation is satisfied for any $k\geq 0$ with the following mathematical conditions (because otherwise these terms have no mathematical meaning):

$$p_{-1}=0 \qquad \lambda_{-1}=0 \tag{15.4.332}$$

and the following real logical condition (calls not yet existing can not end...):

$$\mu_0=0 \tag{15.4.333}$$

Remark

Let us emphasize that the stability of probabilities means that there is an equal probability p_k of leaving the state that to join it.

By using (see above):

$$\lambda_{k-1}p_{k-1}(t)+\mu_{k+1}p_{k+1}(t)-(\lambda_k+\mu_k)\,p_k(t)=0 \tag{15.4.334}$$

we get therefore:

$$\begin{aligned}
&p_1=\frac{\lambda_0}{\lambda_1}p_0\\
&p_2=\frac{1}{\mu_2}\left[(\lambda_1+\mu_1)\frac{\lambda_0}{\mu_1}p_0-\lambda_0p_0\right]=\frac{\lambda_1\lambda_0}{\mu_2\mu_1}p_0\\
&p_3=\frac{1}{\mu_3}\left[(\lambda_2+\mu_2)\frac{\lambda_1\lambda_0}{\mu_2\mu_1}p_0-\lambda_1\frac{\lambda_0}{\mu_1}p_0\right]=\frac{\lambda_2\lambda_1\lambda_0}{\mu_3\mu_2\mu_1}p_0\\
&\ldots
\end{aligned} \tag{15.4.335}$$

We then find quite easily the general form:

$$p_k=\left(\prod_{i=0}^{k-1}\frac{\lambda_i}{\mu_{i+1}}\right)p_0 \tag{15.4.336}$$

Since the system is necessarily in one of the states, we have the following relation which must be satisfied:

$$\sum_{k=0}^{+\infty}p_k=1 \tag{15.4.337}$$

By replacing with the antecedent relationship:

$$\sum_{k=0}^{+\infty}\left(\prod_{i=0}^{k-1}\frac{\lambda_i}{\mu_{i+1}}\right)p_0=1 \tag{15.4.338}$$

This also gives:

$$p_0+\sum_{k=1}^{+\infty}\left(\prod_{i=0}^{k-1}\frac{\lambda_i}{\mu_{i+1}}\right)p_0=1 \tag{15.4.339}$$

and therefore:

$$p_0=\frac{1}{1+\displaystyle\sum_{k=1}^{+\infty}\left(\prod_{i=0}^{k-1}\frac{\lambda_i}{\mu_{i+1}}\right)} \tag{15.4.340}$$

If we now consider a system with a single line and infinite capacity (steady state), the quantities λ_k (call arrival rate) and μ_k (call departure rates) will have identical values for all k. That is, we consider that the arrival rate and the starting rate are constant regardless of the position in which the queue is located. We then have the last relation which simplifies:

$$p_0=\frac{1}{1+\displaystyle\sum_{k=1}^{+\infty}\left(\prod_{i=0}^{k-1}\frac{\lambda_i}{\mu_{i+1}}\right)}=\frac{1}{1+\displaystyle\sum_{k=1}^{+\infty}\left(\frac{\lambda}{\mu}\right)^k} \tag{15.4.341}$$

Using the result proved in the section Sequences and Series (Gauss series see at page 1666) we have under precise necessary conditions of convergence (A must be strictly smaller than 1):

$$p_0=\frac{1}{1+\displaystyle\sum_{k=1}^{+\infty}\left(\frac{\lambda}{\mu}\right)^k}=\frac{1}{\displaystyle\sum_{k=0}^{+\infty}\frac{1}{\left(\dfrac{\lambda}{\mu}\right)^k}}=\frac{1}{\dfrac{1}{1-\dfrac{1}{\dfrac{\mu}{\lambda}}}}=\frac{1}{\dfrac{1}{1-\frac{\lambda}{\mu}}}=1-\frac{\lambda}{\mu}=1-A \tag{15.4.342}$$

Since k represents the number of calls, this last relation gives the probability of having 0 calls (clients) for a given permanent traffic A on the line.

Using:

$$p_k=\left(\prod_{i=0}^{k-1}\frac{\lambda_i}{\mu_{i+1}}\right)p_0 \tag{15.4.343}$$

We then have in generality for this type of system the probability of having k calls in steady state which is given by:

$$p_k=\left(\frac{\lambda}{\mu}\right)^k(1-A)=A^k(1-A) \tag{15.4.344}$$

It follows that the expected mean of number of communications (clients) in the system (in the queue + in service) is then by definition of the expected mean:

$$\mathrm{E}(C)\sum_{k=0}^{+\infty}k\cdot p_k=\sum_{k=0}^{+\infty}kA^k(1-A)=(1-A)\sum_{k=0}^{+\infty}kA^k \tag{15.4.345}$$

since p_k represents the probability that there are k calls in the system at any time (queue + in service). However, we have seen in the section of Sequences and Series that:

$$\sum_{i=0}^{n} q^i = \frac{1-q^{n+1}}{1-q} \tag{15.4.346}$$

and if q is strictly less than 1 and n tends to infinity, we immediately have:

$$\sum_{i=0}^{+\infty} q^i = \frac{1}{1-q} \tag{15.4.347}$$

If we derive this last relation:

$$\frac{\mathrm{d}}{\mathrm{d}q}\sum_{i=0}^{+\infty} q^i = \sum_{i=0}^{+\infty} iq^{i-1} = \frac{1}{(1-q)^2} \tag{15.4.348}$$

and by multiplying by q then it comes:

$$q\frac{\mathrm{d}}{\mathrm{d}q}\sum_{i=0}^{+\infty} iq^{i-1} = \sum_{i=0}^{+\infty} iq^{i} = \frac{q}{(1-q)^2} \tag{15.4.349}$$

It then comes to the end for the expected mean of the number of customers in the system:

$$\mathrm{E}(C) = (1-A)\sum_{i=0}^{+\infty} kA^k = (1-A)\sum_{i=1}^{+\infty} kA^k = (1-A)\frac{A}{(1-A)^2} = \frac{A}{1-A} \tag{15.4.350}$$

If we want to know the expected mean of the number of customers waiting in the queue only, we must understand that at every moment we have a probability p_k that there are k customers in the system but as 1 among these is always in service (i.e. out of the queue), we still have $k-1$ really waiting in the queue. Therefore:

$$\begin{aligned}\mathrm{E}(C_Q) &= \sum_{k=0}^{+\infty}(k-1)p_k = \sum_{k=1}^{+\infty} kp_k - \sum_{k=1}^{+\infty} p_k = (1-A)\sum_{k=1}^{+\infty} kA^k - (1-p_0)\\ &= \frac{A}{1-A} - A = \frac{A}{1-A} - \frac{A(1-A)}{1-A} = \frac{A-A+A^2}{1-A} = \frac{A^2}{1-A}\end{aligned} \tag{15.4.351}$$

Knowing the number of communications (or customers) we have on the single line throughout the system, we then have the expected mean of the waiting time (average waiting time), denoted $\mathrm{E}(T)$ which will be given by the ratio of the expected mean of the number of communications (customers) $\mathrm{E}(C)$ in steady state on line by the arrival rate of the calls:

$$\mathrm{E}(T) = \frac{\mathrm{E}(C)}{\lambda} \tag{15.4.352}$$

and often written in textbooks as:

$$\boxed{L = \lambda W} \tag{15.4.353}$$

that is to say long-term average number of customers in a stable system L is equal to the long-term average effective arrival rate, λ, multiplied by the average time a customer spends in the system, W.

This result, named "Little's law" (John Little having strictly demonstrated that the relation is valid for any type of queue), is intuitive in this case. Actually, let's take a random call. When it arrives in the system, it will be statistically facing $\mathrm{E}(C)$ customers waiting. When it leaves the system, there will have been in it an average time $\mathrm{E}(T)$. So during this mean time, $\lambda\mathrm{E}(T)$ calls will have arrived behind him in the system. In steady state, the number of calls left behind on departure must equal the number of arrivals. Hence the equality $\lambda\mathrm{E}(T)=\mathrm{E}(C)$ from which we then deduce immediately the Little's law.

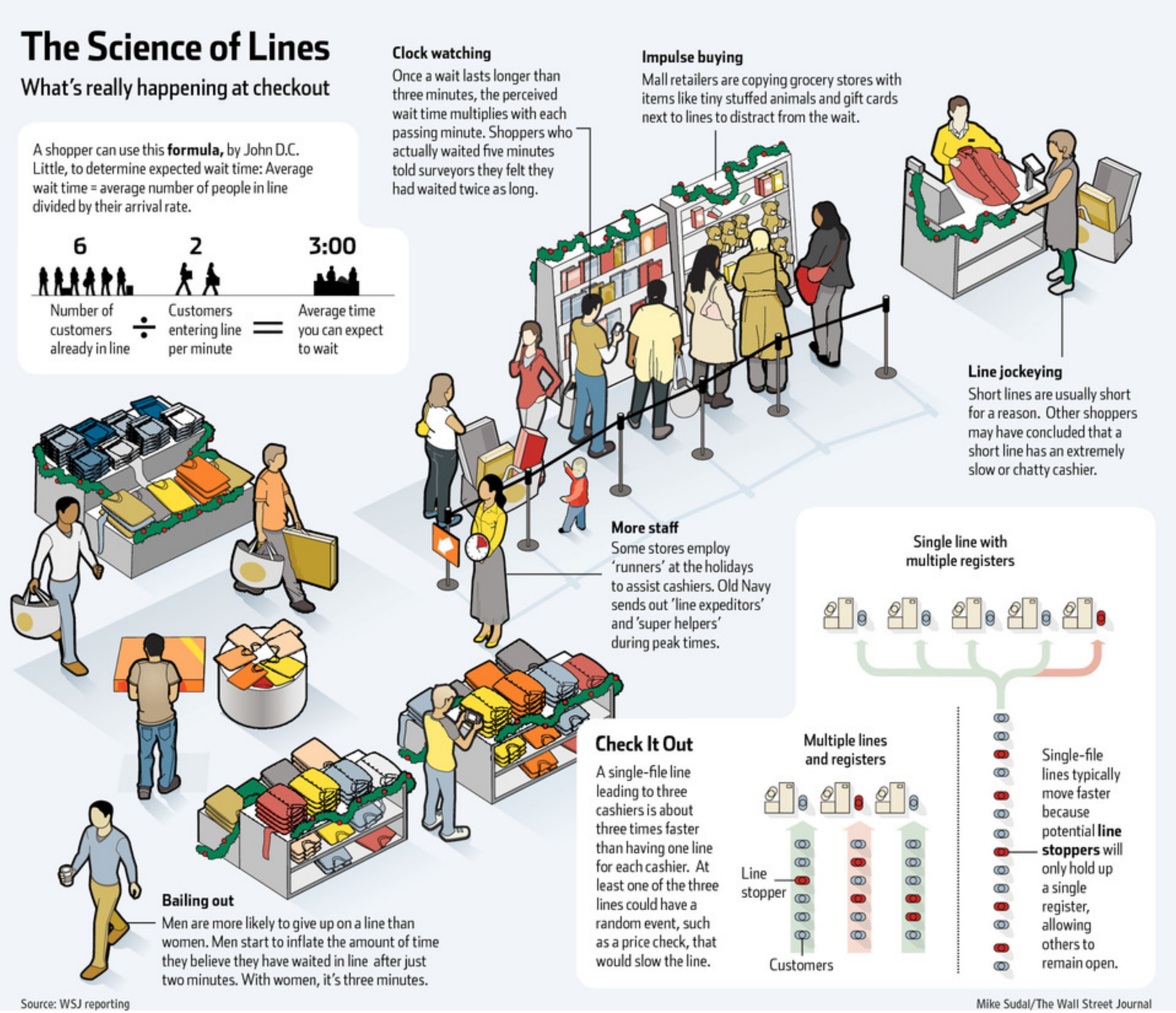

The Science of Lines
What's really happening at checkout
A shopper can use this formula, by John D.C. Little, to determine expected wait time: Average wait time = average number of people in line divided by their arrival rate.
6
2
3:00
Number of customers already in line
÷
Customers entering line per minute
=
Average time you can expect to wait
Clock watching
Once a wait lasts longer than three minutes, the perceived wait time multiplies with each passing minute. Shoppers who actually waited five minutes told surveyors they felt they had waited twice as long.
Impulse buying
Mall retailers are copying grocery stores with items like tiny stuffed animals and gift cards next to lines to distract from the wait.
Line jockeying
Short lines are usually short for a reason. Other shoppers may have concluded that a short line has an extremely slow or chatty cashier.
More staff
Some stores employ 'runners' at the holidays to assist cashiers. Old Navy sends out 'line expeditors' and 'super helpers' during peak times.
Single line with multiple registers
Check It Out
A single-file line leading to three cashiers is about three times faster than having one line for each cashier. At least one of the three lines could have a random event, such as a price check, that would slow the line.
Multiple lines and registers
Line stopper
Customers
Single-file lines typically move faster because potential **line stoppers** will only hold up a single register, allowing others to remain open.
Bailing out
Men are more likely to give up on a line than women. Men start to inflate the amount of time they believe they have waited in line after just two minutes. With women, it's three minutes.
Source: WSJ reporting
Mike Sudal/The Wall Street Journal

Figure 15.316: **The science of lines (source: The Wall Street Journal)**

We then have for the expected mean of waiting time in the system:

$$\mathrm{E}(T)=\frac{\mathrm{E}(C)}{\lambda}=\frac{\dfrac{A}{1-A}}{\lambda}=\frac{A}{\lambda(1-A)}=\frac{\dfrac{\lambda}{\mu}}{\lambda\left(1-\dfrac{\lambda}{\mu}\right)}=\frac{1}{\dfrac{\mu}{\lambda}\lambda\left(1-\dfrac{\lambda}{\mu}\right)}=\frac{1}{\mu-\lambda}\qquad(15.4.354)$$

To determine the waiting time in the queue only, it is sufficient to subtract the processing /

service time from the linearity property of the mean (average call duration in service):

$$\mathrm{E}(T_Q)=\mathrm{E}(T)-\frac{1}{\mu}=\frac{1}{\mu-\lambda}-\frac{1}{\mu}=\frac{1}{\mu}\left(\frac{1}{1-A}-1\right)=\frac{1}{\mu}\left(\frac{1-1+A}{1-A}\right)=\frac{1}{\mu}\left(\frac{A}{1-A}\right) \tag{15.4.355}$$

that we often found in the literature in the following form:

$$W_Q=\frac{1}{\mu-\lambda}-\frac{1}{\mu}=\frac{\lambda}{\mu(\mu-\lambda)} \tag{15.4.356}$$

If there are n servers (or cashiers) we don't change the value of μ as from the point of view of one customer that has to choose ONE server (or cashier) the fact that there are more than 1 server (or cashier) does not influence the fact the he has to pass anyway by one of the server (or cashier), but we divide λ by n as if the servers are equiprobably chosen by the customers he can divide the number of customers arriving customers by the number of cashiers such that:

$$\boxed{W_Q=\frac{\dfrac{\lambda}{n}}{\mu\left(\mu-\dfrac{\lambda}{n}\right)}} \tag{15.4.357}$$

with the condition that (otherwise the expected mean will be negative):

$$\frac{\lambda}{\mu}<n \tag{15.4.358}$$

✍**Example:**

If we have $n=3$ (cashiers), $\lambda=1/2$ per minute (30 customers arriving per hour), and $\mu=1/5$ per minute (customers rate served by each cashier equal to 12 customers served per hour) then:

$$W_Q=\frac{\dfrac{\lambda}{n}}{\mu\left(\mu-\dfrac{\lambda}{n}\right)}=25\,[\text{minutes}] \tag{15.4.359}$$

To summarize, since there are many parameters and results, we have for a type $M/M/1$ queue (according to the Kendall notation):

Information	$M/M/1$
Probability of empty system	$1-A$
Waiting probability	A
Number of customers in the system	$\mathrm{E}(C)=\dfrac{A}{1-A}$
Average number of customers in the system	$\mathrm{E}(C_Q)=\dfrac{A^2}{1-A}$
Number of customers in service (treatment)	$A=\dfrac{\lambda}{\mu}$
Average time in the whole system	$\mathrm{E}(T)=\dfrac{\mathrm{E}(C)}{\lambda}=\dfrac{1}{\mu-\lambda}$
Average time in the queue	$\mathrm{E}(T_Q)=\dfrac{1}{\mu}\left(\dfrac{A}{1-A}\right)$
Condition of equilibrium (for one server!)	$\dfrac{\lambda}{\mu}<1$
Probability of having k customers	$p_k=A^k(1-A)$

Table 15.84: $M/M/1$ **queue summary of important relations**

We see that some relations diverge to infinity when the permanent traffic A tends to unity. This is why we had imposed above that this parameter is strictly less than unity!

We should for each possible type of queue, detail the proof of the corresponding relation above, which is long and laborious (it is a job/specialization on its own right and there are really good textbooks of more than 400 pages on the subject like [415]).

✍**Example:**

Let us consider that we have a numerical control machine processing parts one at a time. Assume that $\lambda=8$ parts per hour (number of pieces arriving on average per hour) and that $\mu=10$ parts per hour (number of pieces coming out on average per hour). We have then:

$$A=\frac{\lambda}{\mu}=\frac{8}{10}=80\% \tag{15.4.360}$$

Which corresponds to the traffic or occupancy rate of the machine.

The expected number of parts in the system (machine + queue) is given by:

$$\mathrm{E}(C)=\frac{A}{1-A}=\frac{0.8}{1-0.8}=4 \tag{15.4.361}$$

The expected number of parts waiting in the queue only is given by:

$$\mathrm{E}(C_Q)=\frac{A^2}{1-A}=\frac{0.8^2}{1-0.8}=3.2 \tag{15.4.362}$$

The expected waiting time in the system is given by:

$$\mathrm{E}(T)=\frac{1}{\mu-\lambda}=\frac{1}{10-8}=0.5 \tag{15.4.363}$$

that is to say 30 minutes.

The expected waiting time in the queue only is given by:

$$\mathrm{E}(T_Q)=\frac{1}{\mu}\left(\frac{A}{1-A}\right)=\frac{1}{10}\left(\frac{0.8}{1-0.8}\right)=0.4 \qquad (15.4.364)$$

that is to say 24 minutes.

And the probability that there are 5 pieces in the system (execution + expectation):

$$p_k=A^k(1-A)=0.8^5(1-0.8)\cong 6.5\% \qquad (15.4.365)$$

15.4.5.5 Probability of standby in a $M/M/k/k$ queue (Erlang-B formula)

We are interested here in a system with N communication channels (each channel is supposed to support a call rate with an immediate response). If the N channels are busy, incoming calls are considered lost (eg, no dial tone). We are talking about "blocking the system" or "ruining the system". It is therefore a limited queue of type $M/M/k/k$ according to the Kendall notation, also named "system with loss".

We will try to estimate this probability of blocking according to the number of available channels and the traffic.

Given what was stated on the memoryless character of the call arrival process, we can consider that the probability:

$$p_k = \lambda_k \mathrm{d}t \tag{15.4.366}$$

to have k calls in the state k is independent of the state of the system such that:

$$p_k = p = \lambda \mathrm{d}t \tag{15.4.367}$$

Thus, at each state k of the system, the Poisson-type probability law is valid. The difference in treatment is that rather than considering states, we will consider that a communication channel can be considered as a state of its own.

But for the probability of end of call, we have:

$$p_k = \mu_k \mathrm{d}t = k\mu \mathrm{d}t \tag{15.4.368}$$

Indeed, this probability just reflects the fact that if k calls are in progress each has a probability $\mu\mathrm{d}t$ to terminate, hence the sum that gives $k\mu\mathrm{d}t$. We have then:

$$\mu_k = k \cdot \mu \tag{15.4.369}$$

Thus, by injecting these relations in:

$$p_0 = \frac{1}{1 + \displaystyle\sum_{k=1}^{+\infty} \left(\prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right)} \tag{15.4.370}$$

it comes:

$$p_0 = \frac{1}{1 + \displaystyle\sum_{k=1}^{N} \left(\prod_{i=0}^{k-1} \frac{\lambda}{(i+1)\mu} \right)} = \frac{1}{1 + \displaystyle\sum_{k=1}^{N} \left(\left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} \right)} \tag{15.4.371}$$

Introducing then (which must be strictly less than 1 if we want the following developments to converge to a finite value: ergodic Markov chain):

$$A = \frac{\lambda}{\mu} \tag{15.4.372}$$

which is for recall the number of calls that appear on the number of calls that end in a given time interval (average service time), which is actually simply the "traffic" (or the "traffic intensity" or "traffic rate"), then it comes:

$$p_0=\frac{1}{1+\sum\limits_{k=1}^{N}\left(\dfrac{A^k}{k!}\right)} \tag{15.4.373}$$

or by introducing the 1 in the summation:

$$p_0=\frac{1}{\sum\limits_{k=0}^{N}\left(\dfrac{A^k}{k!}\right)} \tag{15.4.374}$$

Since k represents the number of calls, this last relation gives the probability of having 0 calls (clients) for a given permanent traffic A in the system.

By reporting p_0 in the following expression of p_k (probability of being in the state k therefore...) obtained above:

$$p_k=\left(\prod_{i=0}^{k-1}\frac{\lambda_i}{\mu_{i+1}}\right)p_0 \tag{15.4.375}$$

it comes:

$$p_k=\left(\prod_{i=0}^{k-1}\frac{\lambda_i}{\mu_{i+1}}\right)p_0=\left(\prod_{i=0}^{k-1}\frac{\lambda_i}{\mu_{i+1}}\right)\frac{1}{\sum\limits_{k=0}^{N}\left(\dfrac{A^k}{k!}\right)} \tag{15.4.376}$$

And considering the character without memory, we have the relation:

$$\prod_{i=0}^{k-1}\frac{\lambda_i}{\mu_{i+1}}=\prod_{i=0}^{k-1}\frac{\lambda}{(i+1)\cdot\mu}=\prod_{i=1}^{k}\frac{\lambda}{i\cdot\mu}=\frac{A^k}{k!} \tag{15.4.377}$$

it comes:

$$p_k=\left(\prod_{i=0}^{k-1}\frac{\lambda_i}{\mu_{i+1}}\right)p_0=\left(\prod_{i=0}^{k-1}\frac{\lambda_i}{\mu_{i+1}}\right)\frac{1}{\sum\limits_{k=0}^{N}\left(\dfrac{A^k}{k!}\right)}=\frac{\dfrac{A^k}{k!}}{\sum\limits_{k=0}^{N}\left(\dfrac{A^k}{k!}\right)} \tag{15.4.378}$$

where the k can unfortunately be confusing. It is necessary to do some cleaning! In the numerator, the k refers to the number of channels (servers, lines, operators or terminals) and at the denominator the N also. It is therefore necessary to rewrite this more appropriately:

$$\boxed{p_k=\frac{\dfrac{A^N}{N!}}{\sum\limits_{k=0}^{N}\left(\dfrac{A^k}{k!}\right)}} \tag{15.4.379}$$

which therefore gives the probability of placing on hold (and thus the probability of saturation / blocking) of a system having N channels of finite capacity according to the principle of first in / first out and for a traffic A (expressed in "Erlang") and in which communications are lost if placed on hold.

This relation is sometimes denoted in the literature in the form of:

$$\pi_k=\frac{\dfrac{\rho^k}{k!}}{\displaystyle\sum_{i=0}^{k}\left(\frac{\rho^i}{i!}\right)} \tag{15.4.380}$$

This relation is very important in queuing theory and is named the "Erlang-B formula".

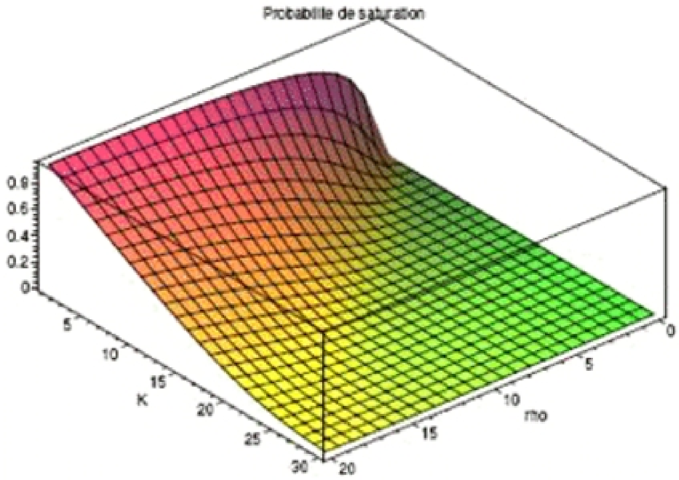


Figure 15.317: **Probability of saturation (smoothed plot of the Erlang-B function)**

This relation is at the basis for the design of circuit-switched networks. Indeed, the problem of sizing a circuit switch is as follows: Given the traffic in number of calls per unit of time A in Erlang, find the number of service units k such that the probability that a call arrives in a system that has become blocking is lower than a certain value.

The probability obtained then represents the quality of service offered by the network from the point of view of the usage. And about the traffic A, that latter is estimated on the basis of the number of existing and / or future telephone installation based on an average activity per station and per application (telephone, fax, terminal, computer server, etc.).

✍**Example:**

E1. What is the probability of saturation of a hotline (whose duration of service follows an exponential law and the distribution of the arrivals follows a Poisson law) knowing that the traffic A of the line is estimated to be 2 [Erlang] (1 call per hour for 1 call processed by $1/2$ hour - hence a ratio of 2 to 1) for a single telephone line ($N=1$)

using the Erlang-B model:

$$p_k = \frac{\dfrac{2^1}{1!}}{\dfrac{2^0}{0!}+\dfrac{2^1}{1!}} \cong 67\% \qquad (15.4.381)$$

E2. In a company, we have counted 200 calls during the peak hours of an average of 6 minutes per hour (average service time). What is the probability of saturation with 20 operators (knowing that the length of service follows an exponential law and the distribution of arrivals a Poisson law)?

The biggest difficulty here is to calculate the traffic! There are thus 200 calls per hour with 10 calls processed only per hour (since 6 minutes per call in an hour of 60 minutes gives: 10 calls). The traffic A is therefore $200/10$ or 20 Erlang. Applying then the previous relation, we have:

$$p_k = \frac{\dfrac{20^{20}}{20!}}{\displaystyle\sum_{k=0}^{20}\dfrac{20^k}{k!}} \cong 15.89\% \qquad (15.4.382)$$

In the industry, a saturation rate of 0.01% is assumed. By playing with a spreadsheet like Microsoft® Excel® and the Target Value tool, we quickly find that N must be equal to 30.

15.4.5.6 Probability $M/M/K/+\infty$ of standby (Erlang-C formula)

Let us consider now a system for which calls can be queued before being served when the k servers are blocked. It is therefore an unlimited capacity queue of type $M/M/k/+\infty$ according to the Kendall notation.

With this system, we always have:

$$p_k = p = \lambda \mathrm{d}t \qquad (15.4.383)$$

But for the probability of end of call the analysis becomes more subtle. First there is the probability that the calls on the available N-channels will cease and that is given by:

$$\mu_k \mathrm{d}t = k\mu \mathrm{d}t \quad \forall\, 0 \leq k \leq N-1 \qquad (15.4.384)$$

But as soon as the number of calls is greater than the number of communication channels available, the probability that the calls cease is:

$$\mu_k \mathrm{d}t = N\mu \mathrm{d}t \quad \forall\, k \geq N \qquad (15.4.385)$$

Which means that regardless of the number of calls, N have the probability of being put on hold (standby) as soon as k (the number of calls) is greater than or equal to N.

Thus, to summarize:

$$\mu_k \mathrm{d}t = \begin{cases} k\mu \mathrm{d}t & \forall\, 0 \leq k \leq N-1 \\ N\mu \mathrm{d}t & \forall\, k \geq N \end{cases} \tag{15.4.386}$$

Using:

$$p_k = \left(\prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right) p_0 \tag{15.4.387}$$

We get by decomposition of the product term:

$$p_k = \left(\prod_{i=0}^{N-1} \frac{\lambda}{(i+1)\mu} \prod_{i=N}^{k-1} \frac{\lambda}{N\mu} \right) p_0 = \left(\frac{A^N}{N!} \frac{A^{k-N}}{N^{k-N}} \right) p_0 \tag{15.4.388}$$

Hence finally:

$$p_k = \begin{cases} \dfrac{A^N}{N!} p_0 & \forall\, 0 \leq k \leq N-1 \\ \dfrac{A^N}{N!} \dfrac{A^{k-N}}{N^{k-N}} p_0 & \forall\, k \geq N \end{cases} \tag{15.4.389}$$

Using the expression of p_0:

$$p_0 = \frac{1}{1 + \displaystyle\sum_{k=1}^{+\infty} \left(\prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right)} \tag{15.4.390}$$

We can decompose the summation:

$$p_0 = \frac{1}{1 + \displaystyle\sum_{k=1}^{N-1} \left(\prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right) + \sum_{k=N}^{+\infty} \left(\prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right)} \tag{15.4.391}$$

and decompose the second product:

$$\begin{aligned} p_0 &= \frac{1}{1 + \displaystyle\sum_{k=1}^{N-1} \left(\prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right) + \sum_{k=N}^{+\infty} \left(\prod_{i=0}^{N-1} \frac{\lambda}{(i+1)\mu} \prod_{i=N}^{k-1} \frac{\lambda}{N\mu} \right)} \\ &= \frac{1}{1 + \displaystyle\sum_{k=1}^{N-1} \left(\prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right) + \sum_{k=N}^{+\infty} \left(\frac{A^N}{N!} \frac{A^{k-N}}{N^{k-N}} \right)} \\ &= \frac{1}{1 + \displaystyle\sum_{k=1}^{N-1} \frac{A^k}{k!} + \frac{A^N}{N!} \sum_{k=N}^{+\infty} \left(\frac{A^{k-N}}{N^{k-N}} \right)} \\ &= \frac{1}{\displaystyle\sum_{k=0}^{N-1} \frac{A^k}{k!} + \frac{A^N}{N!} \sum_{k=N}^{+\infty} \left(\frac{A^{k-N}}{N^{k-N}} \right)} \\ &= \frac{1}{\displaystyle\sum_{k=0}^{N-1} \frac{A^k}{k!} + \frac{A^N}{N!} \sum_{k=0}^{+\infty} \left(\frac{A}{N} \right)^k} \end{aligned} \tag{15.4.392}$$

Under the assumption that:

$$\frac{A}{N}<1 \tag{15.4.393}$$

The sum:

$$\sum_{k=0}^{+\infty}\left(\frac{A}{N}\right)^k \tag{15.4.394}$$

can be simplified. Indeed, by putting:

$$\frac{N}{A}=x \tag{15.4.395}$$

it comes:

$$\sum_{k=0}^{+\infty}\frac{1}{x^k} \tag{15.4.396}$$

and as we have prove it in our study of the Zeta function (see section Sequences and Series page 1682) this sum can be written in the form:

$$\sum_{k=0}^{+\infty}\frac{1}{x^k}=\frac{1}{1-\dfrac{1}{x}} \tag{15.4.397}$$

Therefore:

$$\sum_{k=0}^{+\infty}\left(\frac{A}{N}\right)^k=\frac{1}{1-\dfrac{A}{N}} \tag{15.4.398}$$

So finally:

$$p_0=\frac{1}{\displaystyle\sum_{k=0}^{N-1}\frac{A^k}{k!}+\frac{A^N}{N!}\frac{1}{1-\dfrac{A}{N}}} \tag{15.4.399}$$

Since k represents the number of calls, this last relation gives the probability of having 0 calls (customers) for a given permanent traffic A in the system.

We thus have for $\forall k\geq N$:

$$p_k=\frac{A^N}{N!}\frac{A^{k-N}}{N^{k-N}}p_0=\frac{\dfrac{A^N}{N!}\dfrac{A^{k-N}}{N^{k-N}}}{\displaystyle\sum_{k=0}^{N-1}\frac{A^k}{k!}+\frac{A^N}{N!}\frac{1}{1-\dfrac{A}{N}}}=\frac{\dfrac{A^N}{N!}\dfrac{A^{k-N}}{N^{k-N}}}{\displaystyle\sum_{k=0}^{N-1}\frac{A^k}{k!}+\frac{A^N}{N!}\frac{N}{N-A}} \tag{15.4.400}$$

The cumulative probability of queuing is denoted $C(N,A)$. It is equal to:

$$C(N,A)=\sum_{k=N}^{+\infty}p_k=\sum_{k=N}^{+\infty}\frac{A^N}{N!}\frac{A^{k-N}}{N^{k-N}}p_0=\frac{\displaystyle\sum_{k=N}^{+\infty}\frac{A^N}{N!}\frac{A^{k-N}}{N^{k-N}}}{\displaystyle\sum_{k=0}^{N-1}\frac{A^k}{k!}+\frac{A^N}{N!}\frac{N}{N-A}}$$

$$=\frac{\displaystyle\frac{A^N}{N!}\sum_{k=N}^{+\infty}\frac{A^{k-N}}{N^{k-N}}}{\displaystyle\sum_{k=0}^{N-1}\frac{A^k}{k!}+\frac{A^N}{N!}\frac{N}{N-A}}=\frac{\displaystyle\frac{A^N}{N!}\sum_{k=0}^{+\infty}\left(\frac{A}{N}\right)^k}{\displaystyle\sum_{k=0}^{N-1}\frac{A^k}{k!}+\frac{A^N}{N!}\frac{N}{N-A}} \quad (15.4.401)$$

By proceeding exactly as in the preceding paragraphs, we get finally:

$$\boxed{C(N,A)=\frac{\displaystyle\frac{A^N}{N!}\frac{N}{N-A}}{\displaystyle\sum_{k=0}^{N-1}\frac{A^k}{k!}+\frac{A^N}{N!}\frac{N}{N-A}}} \quad (15.4.402)$$

which therefore gives the probability of putting on hold (and thus of saturation / blocking - that is to say all servers are blocked) upon arrival in a system having N channels with infinite capacity according to the principle of the first in/first out (FIFO) and for a constant traffic A (expressed in "Erlang") and in which communications can be placed on hold (at the opposite to the Erlang-B model). This last relation is named the "Erlang-C formula". Traditionally we write this last relation as following:

$$C(N,A)=P(W) \quad (15.4.403)$$

the "W" coming from "Wait".

The model proposed above is obviously open to criticism because in reality the capacity of the queue is finite and some clients give up when the wait is too long.

✍**Example:**

If we take an arrival rate of 1 call per minute and an average service time of 5 minutes, then we have:

$$A=\frac{\lambda}{\mu}=\lambda\frac{1}{\mu}=1\cdot 5=5 \quad (15.4.404)$$

If we take a number N of 7 servers, we have:

$$C(N,A)=\frac{\displaystyle\frac{A^N}{N!}\frac{N}{N-A}}{\displaystyle\sum_{k=0}^{N-1}\frac{A^k}{k!}+\frac{A^N}{N!}\frac{N}{N-A}}=\frac{\displaystyle\frac{5^7}{7!}\frac{7}{7-5}}{\displaystyle\sum_{k=0}^{6}\frac{5^k}{k!}+\frac{5^7}{7!}\frac{7}{7-5}}\cong 0.3241 \quad (15.4.405)$$

Therefore 32.41% cumulative probability of being put on hold (standby). What a is little too much ... (an empirical rule is to look for the number N of servers so that the latter value falls below 20%).

15.4.6 Insurance

Insurance is an operation in which a person (the "insurer") groups other persons (the "insured") into mutual insurance schemes in order to put them in a position to compensate each other for possible losses (losses) to which expose them certain risks[62], thanks to amounts of cash (premiums or contributions) paid by each insured individual to a common mass (cohort) managed by the insurer.

quote 15.4:8170.#89

> *Insurance is when you buy coverage when you need it. Reality is when you buy the option of going to court for coverage when you need it.*

The common denominator of the majority of laws define the insurance contract as a contract whereby, subject to the payment of a fixed or variable premium, one party, the insurer, undertakes to another party, the insured (also named "policyholder") to provide a benefit stipulated in the contract in the event of an uncertain event that the insured or the beneficiary, as the case may be, has an interest in not being realized.

Obviously, the insurance transaction results in the transfer (total or partial) of the financial consequences of the risk incurred by the insured to an insurance company. Therefore, upon entering into the contract, the insurer and the insured agree to:

- Of an event or list of events, included in the insurance policy, and guaranteed by the insurer.
- A premium paid by the insured to the insurer.

The expenses assumed by the insurer may correspond to:

- Either to indemnities to be paid to third parties, in respect of the liability (civil, professional or other) of the insured in a given laps of time
- Either to repair partially or totally to the damage sustained by the latter.

Remark

We decided to put the study of Insurance in the section of Quantitative Management rather in that of Economy as we have never seen until now a Manager (even the best one) deal with options, shares and so on... (this is the job of Finance graduate people). But we have seen quite often Managers, and this is normal as the degree of difficulty is quite low, seek for insurances contracts to mitigate the risks of some projets that they have to manage.

---

[62] For a list of most common risks in project management the reader can refer to the French free PDF: *Gestion de projets pour ingénieurs et scientifiques* (Project management for engineers and scientists)

The main assumptions used in the insurance industry are as follows:

H1. From a legal point of view, an insurance contract is a "random contract" valid only to cover risks with a random component.

H2. The "rule of the game" of the risk must be stable within a period of time considered as long (at least a few years).

H3. The maximum possible loss must not be too great in relation to the insurer's solvency margin.

H4. The average risk premium must be identifiable and quantifiable according to well-chosen explanatory statistical variables in order to allow a segmentation of risk management (small anecdote: when actuaries are criticized for their segregation practices, we often hear the answer: *We do not discriminate, but differentiate*...).

H5. The risks must be independent (and if they are identical distributed in terms of probabilities and weighting this is better ...) and demonstrated to be such significantly using statistical tools.

H6. There must be a market in the sense that the supply and demand for insurance must arrive at an equilibrium price (in a sense, the equivalent of the absence of arbitrage opportunity in finance).

H7. Mathematical expected mean is considered to be the price of the right pure premium to be paid by the insured.

NOAH, THERE WILL BE A FLOOD, AND IT WILL DESTROY ALL OF CREATION
DON'T WORRY, I'VE GOT INSURANCE

Definitions (#15.4.721):

D1. The "pure premium" is in the field of insurance chosen as the mathematical expected mean value of the burden. It corresponds to the minimum premium that an insurer may ask for not be ruin with certainty.

D2. The "safety charge" is the amount that comes in addition of the pure premium and allows the insurer to withstand the volatility of refunds.

D3. The "management fee charge" is the amount that is added to the previous two and is related to the operation of their company, contract management, premium recovery, investment of assets (basic technical premium), taxes...

D4. The "commercial expenses charge" is added to the previous three and is linked to the acquisition of contracts (commission of intermediaries, costs of commercial networks, advertising).

D5. In the end, all the costs are reflected in the "commercial premium" that is the one communicated to the customer.

We can already conclude that in the case of a compulsory or voluntary state Care insurance system, management and commercial costs will always be lower, under the assumption of an equal management method, and a private system of wage attribution...

Remark

For obvious reasons, it shall be mandatory (as for every business), and required by an ad hoc legislation, that for every price (an especially in the insurance industry), that the percentage of the price allocated to marketing, commercial, management, etc. to be communicated.

15.4.6.1 Premium pricing

So the insurer does not know exactly what the amount of claims is going to be. By taxing contracts at the pure premium level (and assuming a symmetrical loss distribution), the insurer loses money one year on two. In the absence of capital, this situation would quite quick lead to bankruptcy.

To protect himself, the insurer adds to its premium a security charge. Numerous methods of determining it are possible, none having so far largely supplanted the others:

- Charge proportional to the pure premium: The coefficient of proportionality reflects the insurer's view of the volatility of risk.
- Charge dependent on the standard deviation of the losses: This method is a slight formalization of the previous one. It poses a problem because it will introduce a security load that will depend on winnings (actual loss less than the pure premium), excepted if we use the semi-variance.
- Charge dependent on a certain quantity of losses (eg the third quartile). Such a charge ensures that the premium will be sufficient in a predetermined number of cases but does not give any information on technical losses.

Let us consider the case where the population would be heterogeneous, in extenso two classes S_A, S_B of risk coexist in the population of respective weight n_A, n_B with of course:

$$\frac{n_A}{n_A+n_B}+\frac{n_B}{n_A+n_B}=\frac{n_A}{n}+\frac{n_B}{n}=\frac{n_A+n_B}{n}=100\% \tag{15.4.406}$$

Let us consider a health insurance with two categories of insured (healthy young A / seniors at risk B).

To simplify the example to the extreme, let us imagine that the insurer knows, by means of its internal statistics, that an individual in group A will cost the insurance a sum defined by a statistical distribution law that we will denote here:

$$F_X^A(x) \rightarrow P(X_A \leq x) \tag{15.4.407}$$

Then to any cost x is associated a given some cumulative probability given by the function F_X^A. Similarly for group B:

$$F_X^B(x) \rightarrow P(X_B \leq x) \tag{15.4.408}$$

If we randomly select an individual in the portfolio, if θ refers to the group (traditional insurance notation for groups), the insurer should therefore claim in pure premiums for group A:

$$\text{E}[S_A] = \text{E}[S|\theta = A] = \mu_A \tag{15.4.409}$$

hence the expected mean of the function F_X^A.

Same for group B:

$$\text{E}[S_B] = \text{E}[S|\theta = B] = \mu_B \tag{15.4.410}$$

And for a random individual in both groups:

$$\mu_S = \text{E}[S] = \text{E}[\text{E}(S|\theta)] = \frac{n_A}{n}\mu_A + \frac{n_B}{n}\mu_B \tag{15.4.411}$$

The latter case is named the "solidarity mechanism". So the good risks pay for the bad risks ... In fact it's a good way to guarantee "equality" between insured people but not "equity" hence the huge problems that encounters some governments for their system of insurance. The mathematical models of all insurance I know so fare don't take into account the equity that is in physics a fundamental principle of proportion (but don't ask economists to do physics...).

The variance of the pure premium will be given by the relation proved in the Statistics section when studying the properties of the variance (variance of two statistical series as seen at page 595):

$$\sigma_S^2 = \frac{n_A\sigma_A^2 + n_B\sigma_B^2}{n_A + n_B} + n_A n_B \frac{(\mu_A - \mu_B)^2}{(n_A + n_B)^2} \tag{15.4.412}$$

Where we see that the term:

$$(\mu_A - \mu_B)^2 \geq 0 \tag{15.4.413}$$

at the numerator corresponds to the homogeneity of the two groups. Hence, heterogeneity increases the variability of the pure premium in the solidarity mechanism especially when equity is not taking into account.

Now let us imagine the case where:

$$\mu_A < \mu_B \tag{15.4.414}$$

and a private insurance (1) which practices the solidarity mechanism and another insurance (2) which does not practice it. In the present case, we must distinguish two situations from an economic point of view:

If all the insured are rational (therefore somewhat... egoistic) then:

- Healthy young people representing group A will go to private insurance (2) which has segmented the risks and therefore allows the young people to pay less.
- Seniors in poorer health, representing group B, go to private insurance (1), which has not segmented risks, but which, by ideology, has applied the principle of solidarity.

The conclusion is that private insurance (1) will quickly go bankrupt because:

- The good risks no longer compensate the solidarity mechanism because of selfishness
- We are in a competitive market where insurance are not ideological ...

This fact is named the "anti-selection problem" or "adverse selection" based on the approach of G. Akerlof, Nobel Prize for Economics in 12001 (holocene calendar).

The conclusion is that the privatization of insurance based on a solidarity principle can not work without incurring additional costs for pure premiums because of a periodic anti-selection that generates a constant flow of insured persons moving from one insurance to another and thus generates important administrative, marketing and computer costs! So in every point the suppression of competition (concurrence) remains better in terms of solidarity but not in terms of jobs for insurance employees (who would then be for the majority unemployed...) or for the advantages of some CEO and politicians....!!!!!

Thus, the elimination of competition (concurrence) for insurance services which all have equivalent quality services (as it is the case of health insurance in Switzerland, for example) eliminates the principle of anti-selection, applies in a concrete way the solidarity mechanism wanted by the state and finally reduces the costs due to the comings and goings (flow) of the insured and the development of IT tools of business-management costing millions to each insurance and which incur costs which in the end are passed on to the price of the commercial premium!

In addition, in order to reduce overall volatility (global standard deviation), insurance should theoretically segment the risks indefinitely, making it an unsustainable system for certain specific areas of insurance.

But it must be remembered that this conclusion is valid only if the insurance benefits are identical (or quasi-similar) in a given market!

There is also a recurring issue in the field of insurance, named "moral hazard", which is based on the observation that those who insure tend to be less cautious than those who do not insure themselves. In other words, insurance generates risk and this is very important factor that explain cost increase of solidarity insurance systems!

15.4.6.2 Taking account of the return on experience

So far, to determine an insurance premium, we have noticed that it was possible to integrate exogenous variables (gender, age, children, vehicle power, nationality, environment, etc.). But an important point not to be neglected is the return on experience of an insured's claims. Let's see a concrete example!

Suppose that the number of claims over one year for a given insured follows a Poisson law (law of rare events) given for recall by (see section Statistics page 686):

$$P_k = \frac{\mu^k}{k!}e^{-\mu} \tag{15.4.415}$$

which is written in the field of insurance:

$$P_\theta = \frac{\mu^\theta}{\theta!}e^{-\mu} \tag{15.4.416}$$

Let us suppose that the insured population is separated into three classes of risk according to a Poisson distribution such as:

$$\begin{aligned} P(\theta=1) &= \frac{1^\theta}{\theta!}e^{-1} \rightarrow 70\% \\ P(\theta=2) &= \frac{2^\theta}{\theta!}e^{-2} \rightarrow 20\% \\ P(\theta=3) &= \frac{3^\theta}{\theta!}e^{-3} \rightarrow 10\% \end{aligned} \tag{15.4.417}$$

In other words, 70% of the total population follows a Poisson's law of expected mean equal to 1 (good risk class), 20% follows Poisson's law of expected mean equal to 2 (average risk class) , and 10% follows a Poisson's law of expected mean equal to 3 (class of bad risks).

Obviously, the overall expected mean of an individual in the portfolio is the expected mean of the "mixed distribution":

$$\mu = \mathrm{E}(\theta) = 70\% \cdot 1 + 20\% \cdot 2 + 10\% \cdot 3 = 1.4 \tag{15.4.418}$$

Let us suppose that the costs of an incident are fixed and that the indemnity amount is of $1,000$.- . We can then ask what the premium should be for an insured, knowing that the first year he had 2 claims. What is equivalent to ask ourselves: what is the number of claims that he could have in the second year?

If we denote by X_1 the number of claims of the first year, and X_2 the number of claims a posteriori knowing X_1 we thus find ourselves with a problem of conditional probability (a Bayesian approach in other words ...) according to what we have studied in the section of Probabilities (see page 492):

$$P(X_2|X_1=2) = \frac{P(X_1 \cap X_2)}{P(X_1)} = \frac{P(X_1=2 \cap X_2)}{P(X_1=2)} \tag{15.4.419}$$

where we adopted in accordance with the tradition in the field of insurance the notation "|" to indicate we are dealing with conditional probabilities.

But we can not calculate the numerator because we do not know in advance what the will be the value of X_2. We will therefore calculate the expected conditional mean in order to get around this problem:

$$\mathrm{E}(X_2|X_1=2)=\frac{\mathrm{E}(X_1\cap X_2)}{\mathrm{E}(X_1)}=\frac{\mathrm{E}(X_1=2\cap X_2)}{\mathrm{E}(X_1=2)} \tag{15.4.420}$$

First we have:

$$\mathrm{E}(X_1=2)=70\%\cdot\frac{1^2}{2!}e^{-1}+20\%\cdot\frac{2^2}{2!}e^{-2}+10\%\cdot\frac{3^2}{2!}e^{-3}=0.205 \tag{15.4.421}$$

This last calculation (expected mean of the mixed distribution) being written in the field of insurance in the following way:

$$\mathrm{E}(X_1=2)=\sum P(\Theta=\theta)P(X_1=2|\Theta=\theta)=\sum P(\Theta=\theta)\frac{\theta^2}{2!}e^{-\theta} \tag{15.4.422}$$

If we consider the two random variables to be independent:

$$\mathrm{E}(X_1=2\cap X_2)=\mathrm{E}((X_1=2)(X_2))=\mathrm{E}(X_1=2)\mathrm{E}(X_2)=0.205\cdot\mathrm{E}(X_2) \tag{15.4.423}$$

It then comes immediately that:

$$\mathrm{E}(X_2|X_1=2)=\frac{\mathrm{E}(X_1\cap X_2)}{\mathrm{E}(X_1)}=\frac{\mathrm{E}(X_1=2\cap X_2)}{\mathrm{E}(X_1=2)}=\mathrm{E}(X_2) \tag{15.4.424}$$

We therefore fall back on a known value corresponding to the pure premium:

$$\begin{aligned}\mathrm{E}(X_2)&=70\%\cdot\theta_1+20\%\cdot\theta_2+10\%\cdot\theta_3\\&=70\%\cdot1+20\%\cdot2+10\%\cdot3=1.4=\mathrm{E}(\theta)\end{aligned} \tag{15.4.425}$$

Identical to the calculation of:

$$\mathrm{E}(X_1)=70\%\cdot\theta_1+20\%\cdot\theta_2+10\%\cdot\theta_3=1.4=\mathrm{E}(\theta) \tag{15.4.426}$$

Therefore the random variables X_1, X_2 are confused with that of the class of risk!

This system obviously is not conform to the bonus-malus concept. We must then consider that the two random variables are not independent (in fact it is the very definition of the bonus-malus!).

Thus, if the two random variables are not independent, we have:

$$\begin{aligned}&\mathrm{E}(X_1=2\cap X_2)=\mathrm{E}\left(\frac{\theta^2}{2!}e^{-\theta}\cap\theta\right)=\sum P(\Theta=\theta)\left(\frac{\theta^2}{2!}e^{-\theta}\cdot\theta\right)\\&=70\%\cdot\frac{1^2}{2!}e^{-1}\cdot1+20\%\cdot\frac{2^2}{2!}e^{-2}\cdot2+10\%\cdot\frac{3^2}{2!}e^{-3}\cdot3=0.304\end{aligned} \tag{15.4.427}$$

and therefore:

$$\mathrm{E}(X_2|X_1=2)=\frac{\mathrm{E}(X_1=2\cap X_2)}{\mathrm{E}(X_1=2)}=\frac{0.304}{0.205}=1.484 \tag{15.4.428}$$

To compare with the pure premium without bonus-malus of 1.4!

What is tricky with this method is that when we accumulate years... this Bayesian approach becomes painful. Indeed, imagine that we would like to determine the pure premium of the third year knowing that of the second year, the insured had 1 accident the second year. We then have the conditional expected mean:

$$\text{E}(X_3|X_1=2,X_2=1)=\frac{\text{E}((X_1=2,X_2=1)\cap X_3)}{\text{E}(X_1=2,X_2=1)} \tag{15.4.429}$$

that we leave the development to the reader ...

We notice from what has been seen above that the expected mean of a mixed distribution defined by:

$$\text{E}(X)=p_1 f_{\bar{X}_1}(X)+\cdots+p_n f_{\bar{X}_n}(X)=\sum_{i=1}^{n} p_i f_{\bar{X}_i}(X) \tag{15.4.430}$$

is therefore trivially given by:

$$\text{E}(X)=\mu=\sum_{i=1}^{n} p_i\mu_i \tag{15.4.431}$$

15.4.6.3 Discounting factor of a retirement insurance

Let us now look for something that has absolutely nothing to do with what we have just calculated and which is the discount factor (see section Economy page 7558) by which we must divide a savings balance at the beginning of retirement, to allocate the resulting amount every year during retirement, depending on life expectancy, as well as interest and inflation rates. The approach here will be to question ourselves what level of savings it will be necessary to have accumulated at the beginning of retirement to obtain a pension of 1.- (real) each year during retirement. It is from this level of saving necessary to receive 1.- that it will be possible to use any savings balance to derive what annuity will be available to the saver.

First, we will define the parameters used:

- Number of retirement periods (years or months): m
- Savings balance at period k: S_k
- Periodic benefit (which we assume here annual): p
- Periodic rate of inflation (assumed here constant and annual): i
- Average geometric market rate of return (assumed constant and annual): R
- Discount factor at age m: a_m

Thus, as mentioned above, we are looking for the initial level of savings, denoted S_0, which will make possible to get a periodic benefit (therefore annual in this case), p, equal to 1.- indexed to inflation i.

Starting from the initial balance $S_0=S$, assuming that benefits are paid at the beginning of the year (praenumerando), we can derive the remaining balance at the end of the first retirement period as well:

$$S_1=(S-p)\cdot(1+R) \tag{15.4.432}$$

More precisely, the pensioner will have received an amount p for the first period, leaving the residual to be exposed to the geometric mean return on market R.

Then, in a similar way and considering that the benefits are indexed to inflation i, we can calculate this balance at the end of the second period and any subsequent period k:

$$S_2=[(S-p)(1+R)-(1+i)p]\,(1+R) \tag{15.4.433}$$

or:

$$S_2=S(1+R)^2-p(1+R)^2-p(1+R)(1+i) \tag{15.4.434}$$

or:

$$\begin{aligned}S_3&=\left[S(1+R)^2-p(1+R)^2-p(1+R)(1+i)-p(1+i)^2\right](1+R)\\&=S(1+R)^3-p(1+R)^3-p(1+R)^2(1+i)-p(1+R)(1+i)^2\\&=S(1+R)^3-p(1+R)(1+i)^2-p(1+R)^2(1+i)-p(1+R)^3\end{aligned} \tag{15.4.435}$$

And so on ... It comes then:

$$S_k=S(1+R)^k-p\sum_{j=1}^{k}(1+R)^j(1+i)^{k-j} \tag{15.4.436}$$

or otherwise written:

$$S_k=S(1+R)^k-p(1+i)^k\sum_{j=1}^{k}\left(\frac{1+R}{1+i}\right)^j=S(1+R)^k-p(1+i)^k\sum_{j=1}^{k}q^j \tag{15.4.437}$$

Thus we have, to one given term, a geometric series of reason q known in the last term. We have indeed proved in the section of Sequences and Series (see page 1656) that for j ranging from 0 to n this one is equal to:

$$S_n=u_0\frac{1-q^{n+1}}{1-q} \tag{15.4.438}$$

and therefore:

$$\sum_{j=1}^{k}\left(\frac{1+R}{1+i}\right)^j=\left(\frac{1+R}{1+i}\right)\sum_{j=0}^{k-1}\left(\frac{1+R}{1+i}\right)^j=\left(\frac{1+R}{1+i}\right)\frac{1-\left(\dfrac{1+R}{1+i}\right)^k}{1-\dfrac{1+R}{1+i}}=\frac{1-\left(\dfrac{1+R}{1+i}\right)^k}{\dfrac{1+i}{1+R}-1} \tag{15.4.439}$$

It is therefore possible for us to rewrite, for the last retirement period m, the balance as follows:

$$S_m=S(1+R)^m-p(1+i)^m\frac{1-\left(\dfrac{1+R}{1+i}\right)^m}{\dfrac{1+i}{1+R}-1} \tag{15.4.440}$$

Let us notice that in the unlikely and special case where the average geometric market return is equal to inflation, i.e. if $i=R$, we have:

$$\begin{aligned}S_m&=S(1+R)^m-p(1+i)^m\sum_{j=1}^{m}\left(\frac{1+R}{1+i}\right)^j=S(1+R)^m-p(1+R)^m\sum_{j=1}^{m}\left(\frac{1+R}{1+R}\right)^j\\&=S(1+R)^m-p(1+R)^m\sum_{j=1}^{m}1^j=S(1+R)^m-p(1+R)^m\sum_{j=1}^{m}m=(1+R)^m\left[S-p\cdot m\right]\end{aligned} \tag{15.4.441}$$

At this stage, the reason why the term S_m is an important parameter is that it is the retiree's balance at the end of his retirement period, that is, at his death. It will be assumed here that the retiree does not plan a legacy and therefore:

$$S_m=0 \tag{15.4.442}$$

Consequently, for $R\neq i$, this is equivalent to find S for which:

$$S_m=S(1+R)^m-p(1+i)^m\frac{1-\left(\dfrac{1+R}{1+i}\right)^m}{\dfrac{1+i}{1+R}-1}=0 \tag{15.4.443}$$

We can then establish in this case that for $R\neq i$:

$$S=p\frac{(1+i)^m}{(1+R)^m}\frac{1-\left(\dfrac{1+R}{1+i}\right)^m}{\dfrac{1+i}{1+R}-1}=p\left(\frac{1+i}{1+R}\right)^m\frac{1-\left(\dfrac{1+R}{1+i}\right)^m}{\dfrac{1+i}{1+R}-1} \tag{15.4.444}$$

And in the unlikely and special case where the average geometric return of the market is equal to inflation, that is, if $R=i$, we have:

$$S_m=(1+R)^m\left[S-p\cdot m\right]=0 \tag{15.4.445}$$

and we deduce in this special case a logical result:

$$S=p\cdot m \tag{15.4.446}$$

For what follows, we come back to the realistic case where $r\neq i$ and thus with:

$$S=p\left(\frac{1+i}{1+R}\right)^m\frac{1-\left(\dfrac{1+R}{1+i}\right)^m}{\dfrac{1+i}{1+R}-1} \tag{15.4.447}$$

and let us put:

$$b=\left(\frac{1+R}{1+i}\right)=1+r \tag{15.4.448}$$

The prior-previous relation is then written in a somewhat more condensed form:

$$S=p\frac{1}{b^m}\frac{1-b^m}{b-1} \tag{15.4.449}$$

where:

$$\boxed{a_m = \frac{1}{b^m}\frac{1-b^m}{b-1}} \tag{15.4.450}$$

is therefore the discounting factor. In addition, we have the following relation giving the expected mean of being alive at age a (age corresponding to the assumed age of death in the present case) introduced and proved in the section of Populations Dynamics:

$$m = e(a,s) - a + 1 \tag{15.4.451}$$

with for recall the function $e(a,s)$ representing the conditional life expectancy at age a and for sex s.

15.4.6.4 Pension insurance (life annuities)

We will now study a rent that is not paid in a certain way but in a "life annuity" way, that is to say, which depends on the survival of the insured. It will therefore be necessary to multiply each term of the rent by the probability that it will be paid to the insured.

Obviously, in the majority of cases, what will interest us will be the "present value" of the life annuities in order to know what the insurer will have to pay (and thus what he must have capitalized).

We consider mainly the following type of annuity:

1. **Life annuity:** You will receive regular income as long as you live. You can opt for monthly, quarterly, half-yearly or annual payouts depending on the product.
2. **Joint life annuity:** You will receive regular income as long as you live. After you, your life partner will continue getting regular income
3. **Life annuity with return of purchase price:** You will receive regular income as long as you live. After you, your children will get back the initial lump sum amount paid by you
4. **Life annuity guaranteed for 5/10/15 years:** You will receive regular income for a guaranteed period (5/10/15 years). In case of your demise within this period, your nominee will receive regular income till the end of the guaranteed period.
5. **Joint life annuity with return of purchase price:** You will receive regular income as long as you live. After you, your life partner will continue getting regular income. After both of you, your children will get the initial lump sum amount paid by you.
6. **Immediate annuity with regular payout:** Your income flow will start immediately upon investing, followed by payouts at regular intervals selected by you.
7. **Deferred annuity:** You will receive regular periodic payments starting at a future date selected by you.

We will see some of them! But before that, let us recall that we have seen in the section of Populations Dynamics (see page 7300) that the probability of being alive at a year $x+1$ knowing that one is alive at an age x is given by:

$$ {}_1p_x=\frac{S_{x+1}}{S_x}<1 \qquad (15.4.452)$$

and let us recall (following a reader's request) that the present value of a capital C_0 at constant rate payable in 1 year is given by (see section Economy page 7557):

$$C_0=\frac{C_1}{(1+t\%)^1} \qquad (15.4.453)$$

that gives the sum to be placed at a rate $t\%$ to have after one year the capital C_1.

By combining probability of survival with current value, we can define in the case of an annuity the unique premium (UP) of the old-age pension as:

$$\text{UP}=C_0\cdot {}_1p_x \qquad (15.4.454)$$

If the insured is alive at maturity, the insured wins the game (in reality they often win both ...), otherwise it is the insurer (where in that latter case he is really winning a lot of money!).

Obviously in the above case, extremely simplified, there is only one payment from the insurer. But in reality there may be several (this is what we are going to study), the payment of the insured may not be in the form of a single premium, the average geometric market rate may not be constant (the latter being treated normally with Monte Carlo simulations).

So even if what we are going to see is very simplified compared to what is done in insurance by the actuaries, it already gives a search path for more elaborate models.

15.4.6.4.1 Temporary life annuity

The "temporary life annuity" consist for the insurer to pay annuity terms, as long as the insured is alive but not more than n years after he has reached a given age x but immediately after the insured has paid his single premium (which corresponds precisely to the age x in question). The objective is therefore to determine the capital (current value) that the insurer must have accumulated at the time of the payment of these pensions (this capital may also be claimed in the form of multiple payments rather than in the form of a unique premium).

Obviously this case is quite unrealistic because few people will pay a single premium (big sum of money) at a time x of their lives to start right away to receive annuities. But it is pedagogically interesting before discussing the type of deferred life annuity and which is a little more realistic.

The temporary praenumerando life annuity (i.e. the present value of the unit life annuity payable praenumerando as long as the insured is alive but at most for n years) is given by the following relation which follows from the developments made in the section Economy (see page 7565) with the only small difference that one finds there coefficients of probability of survival:

$$\boxed{\ddot{a}_{\overline{x:n}|}=1+f_e\frac{S_{x+1}}{S_x}+f_e^2\frac{S_{x+2}}{S_x}+\ldots+f_e^{n-1}\frac{S_{x+n-1}}{S_x}=\sum_{t=0}^{n-1}f_e^t\frac{S_{x+t}}{S_x}} \qquad (15.4.455)$$

where for recall (see section Economy page 7558) we have the discount factor (in this case we should instead use the term "capitalization factor") which is:

$$f_e=\frac{1}{1+t\%} \tag{15.4.456}$$

In general, the term (see section Population Dynamics page 7300):

$$_{m|}E_x=f_e^m\frac{S_{x+m}}{S_x}=f_e^m\,{}_{m|}p_x \tag{15.4.457}$$

is named "probable value of deferred capital".

The postnumerando temporary life annuity (present value of a unit life annuity payable postnumerando as long as the insured is alive but for maximum during n years) is given by:

$$\boxed{a_{\overline{x:n|}}=f_e\frac{S_{x+1}}{S_x}+f_e^2\frac{S_{x+2}}{S_x}+\ldots+f_e^n\frac{S_{x+n}}{S_x}=\sum_{t=1}^{n}f_e^t\frac{S_{x+t}}{S_x}} \tag{15.4.458}$$

✍**Example:**

We wish to pay an annuity of $1,000$.- during 3 years to a 40 year old man. What is the single premium that he will have to pay if neither interest nor mortality are involved in the calculations is given by:

$$\text{UP}_{(1)}=1,000+1,000+1,000=1,000\cdot 3 \tag{15.4.459}$$

if only the interest is considered in the calculations:

$$\text{UP}_{(2)}=1,000\cdot f_e+1,000\cdot f_e^2+1,000\cdot f_e^2=1'000\cdot a_{\overline{3|}}\leq PU_{(1)} \tag{15.4.460}$$

and finally taking into account interest and mortality:

$$\text{UP}_{(3)}=1,000\cdot f_e\cdot\frac{S_{41}}{S_{40}}+1,000\cdot f_e^2\cdot\frac{S_{42}}{S_{40}}+1,000\cdot f_e^3\cdot\frac{S_{43}}{S_{40}}=1,000\cdot a_{\overline{40:3|}}<PU_{(2)} \tag{15.4.461}$$

15.4.6.4.2 Deferred life annuity

The "deferred life annuities" account for the majority of annuity insurance contracts. The insured person always pays a single premium in order to receive an annuity if he is alive at a certain age (at the age of retirement for example), so he begins receiving the annuity k years after paying the single premium, but this time until his death.

The praenumerando deferred life annuity (present value of a unit life annuity deferred of k years and payable praenumerando until death of the insured) is given by:

$$\boxed{{}_{k|}\ddot{a}_x=f_e^k\frac{S_{x+k}}{S_x}+f_e^{k+1}\frac{S_{x+k+1}}{S_x}+\ldots+f_e^{\omega-x}\frac{S_\omega}{S_x}=\sum_{t=k}^{\omega-x}f_e^t\frac{S_{x+t}}{S_x}} \tag{15.4.462}$$

where ω is the last value (line number) of the life table. The reader may notice that if we set $k=1$ (without delay) and $\omega=x+n$, we fall back on a temporary postnumerando life annuity $a_{\overline{x:n}|}$.

Thus, normally for a same age x and a same supposed annuity duration, the present value of the deferred life annuity (single premium) is lower than that of the temporary life annuity and this the smaller is the delay.

The postnumerando deferred life annuity (current value of a deferred life annuity of k years and payable postnumerando until death of the insured) is given by:

$$\boxed{{}_{k|}a_x=f_e^{k+1}\frac{S_{x+k+1}}{S_x}+f_e^{k+2}\frac{S_{x+k+2}}{S_x}+\cdots+f_e^{\omega-x}\frac{S_\omega}{S_x}=\sum_{t=k+1}^{\omega-x}f_e^t\frac{S_{x+t}}{S_x}} \tag{15.4.463}$$

Obviously, the two annuities above have the same end because given that it is until death, it is difficult to pay in postnumerando of the death...

Let us also notice that it is possible mathematically to create a temporary AND deferred annuity but it is rare and therefore we will not study it here.

Remark

Unlike certain annuities, as far as we know there is no simplified formula for a quick calculation of these pensions based on mortality tables. The use of a spreadsheet software or the programming of these functions makes it easy to calculate the current values of the lifetime benefits. At the time when computers were not existing, we built tables named "commutations numbers" that made it possible to obtain the required information.

15.4.7 Sensitivity Analysis

The mathematical models we have seen previously are defined by a series of equations, input variables and parameters aimed at characterizing some process under investigation. Some examples might be a climate model, an economic model, a project planning model, etc.. Increasingly, such models are highly complex, and as a result their input/output relationships may be poorly understood. In such cases, the model can be viewed as a black box, i.e. the output is an opaque function of its inputs.

Quite often, some or all of the model inputs are subject to sources of uncertainty, including errors of measurement, absence of information and poor or partial understanding of the driving forces and mechanisms. This uncertainty imposes a limit on our confidence in the response or output of the model. Further, models may have to cope with the natural intrinsic variability of the system (aleatory), such as the occurrence of stochastic events.

Good modelling practice requires that the modeller provides an evaluation of the confidence in the model. This requires, first, a quantification of the uncertainty in any model results ("uncertainty analysis") and second, an evaluation of how much each input is contributing to the output uncertainty. Sensitivity analysis addresses the second of these issues (although uncertainty analysis is usually a necessary precursor), performing the role of ordering by importance the strength and relevance of the inputs in determining the variation in the output.

Remarks

R1. Uncertainty analysis investigates the uncertainty of variables that are used in decision-making problems in which observations and models represent the knowledge base. In other words, uncertainty analysis aims to make a technical contribution to decision-making through the quantification of uncertainties in the relevant variables. Many high level managers like to say that a management model is never complete until it has a quantification tool associated with it!

R2. Sensitivity and Uncertainty analysis belongs both to the field of knowledge named "uncertainty quantification" that is the science of quantitative characterization and reduction of uncertainties in both computational and real world applications.

The process of recalculating outcomes under alternative assumptions to determine the impact of variable under sensitivity analysis can be useful for a range of purposes, including:

- Testing the robustness of the results of a model or system in the presence of uncertainty.
- Increased understanding of the relationships between input and output variables in a system or model.
- Identifying model inputs that cause significant uncertainty in the output and should therefore be the focus of attention if the robustness is to be increased (perhaps by further research).
- Searching for errors in the model (by encountering unexpected relationships between inputs and outputs).

- Model simplification by fixing model inputs that have no effect on the output, or identifying and removing redundant parts of the model structure.
- Enhancing communication from modellers to decision makers (e.g. by making recommendations more credible, understandable, compelling or persuasive).
- Finding regions in the space of input factors for which the model output is either maximum or minimum or meets some optimum criterion (see optimization and Monte Carlo filtering).
- In case of calibrating models with large number of parameters, a primary sensitivity test can ease the calibration stage by focusing on the sensitive parameters. Not knowing the sensitivity of parameters can result in time being uselessly spent on non-sensitive ones.

Taking an example from economics, in any budgeting process there are always variables that are uncertain. Future tax rates, interest rates, inflation rates, headcount, operating expenses and other variables may not be known with great precision. Sensitivity analysis answers the question, "if these deviate from expectations, what will the effect be (on the business, model, system, or whatever is being analysed), and which variables are causing the largest deviations?"

Sensitivity analysis is a variation on scenario analysis (deterministic naive modelization) that is useful in pinpointing the areas where forecasting risk is especially severe in any field of corporate finance, project management, lean management, supply chain management.

There are a large number of approaches to perform a sensitivity analysis, many of which have been developed to address one or more of the constraints discussed above. They are also distinguished by the type of sensitivity measure, be it based on (for example) variance decompositions, partial derivatives or elementary effects. In general, however, most procedures adhere to the following outline:

1. Quantify the uncertainty in each input (e.g. ranges, probability distributions). Note that this can be difficult and many methods exist to elicit uncertainty distributions from subjective data.
2. Identify the model output to be analysed (the target of interest should ideally have a direct relation to the problem tackled by the model).
3. Run the model a number of times using some design of experiments, dictated by the method of choice and the input uncertainty.
4. Using the resulting model outputs, calculate the sensitivity measures of interest.

In some cases this procedure will be repeated, for example in high-dimensional problems where the user has to screen out unimportant variables before performing a full sensitivity analysis.

However, in the field of management (projects, finance, statistics) and multivariate case, the sensitivity analysis is at a graduate level a simple technique based on the use of the linear Pearson correlation coefficient that we have studied in the section Statistics (see page 614) and for which a possible writing form in the bivariate discrete case was explicitly:

$$R_{x,y}=\frac{\sum_{k=1}^{n}x_k y_k-\frac{1}{n}\sum_{k=1}^{n}x_k\sum_{k=1}^{n}y_k}{\sqrt[4]{\left[\sum_{k=1}^{n}x_k^2-\frac{1}{n}\left(\sum_{k=1}^{n}x_k\right)^2\right]\left[\sum_{k=1}^{n}y_k^2-\frac{1}{n}\left(\sum_{k=1}^{n}y_k\right)^2\right]}} \quad (15.4.464)$$

Some common difficulties in sensitivity analysis include:

1. Too many model inputs to analyse. Screening can be used to reduce dimensionality.
2. The model takes too long to run
3. There a unsuspected non-linearities and correlations
4. There is not enough information to build probability distributions for the inputs. Probability distributions can be constructed from expert elicitation, although even then it may be hard to build distributions with great confidence. The subjectivity of the probability distributions or ranges will strongly affect the sensitivity analysis.
5. Piecewise sensitivity. This is when one performs sensitivity analysis on one sub-model at a time. This approach is non conservative as it might overlook interactions among factors in different sub-models

Let us begin with the theory of the first method that is strictly speaking not a sensitivity analysis but that many managers (non-scientific) consider as it...

15.4.7.1 Direct Bias Method

The "direct bias method" also named the "one factor at a time" (OFAT) method is also well known in Design of Experiments (DOE). In fact it is the easiest method and it can be applied in a deterministic, probabilistic or stochastic way. It is quite useful when the board committee wants to know the influence of a given factor only for a given amplitude!

OFAT customarily involves:

1. Changing one input variable, keeping others at their baseline (nominal) values, monitoring the output then,
2. Returning the variable to its nominal value, the repeating for each of the other inputs in the same way

Consider that we have an output model given by:

$$y=f(\vec{x},t) \quad (15.4.465)$$

where $\vec{x}$ is the vector of all the variables x_i components that can be time dependent or not.

1. The "deterministic univariate direct bias method" consists simply by doing the following relative calculation:

$$\frac{\Delta y(t)}{y(t)}=\frac{f((x_1(t),\ldots,x_i(t)\pm\delta_{i,t},\ldots,x_n(t)),t)-f((x_1(t),\ldots,x_n(t)),t)}{f((x_1(t),\ldots,x_n(t)),t)} \tag{15.4.466}$$

where the variation $\delta_{i,t}$ is chosen arbitrarily by the practitioner that must keep in mind that the output is not necessarily proportional to the input variation!!!!

The ratio above is a "relative-output/nominal output" but some practitioners prefers a "relative/relative" analysis. Thus if δ_i is express in percentage change of $x_i(t)$ we will write obviously:

$$\frac{\Delta y(t)}{y(t)}=\frac{f((x_1(t),\ldots,x_i(t)\cdot\delta_{i,t},\ldots,x_n(t)),t)-f((x_1(t),\ldots,x_n(t)),t)}{f((x_1(t),\ldots,x_i(t),\ldots,x_n(t)),t)} \tag{15.4.467}$$

Furthermore depending on the amplitude of $\delta_{i,t}$ some non-null correlation can appear with other factor could in reality occur!

This technique is quite limited as you must do as many calculations as you have scenarios (variable inputs ranges).

When the practitioner use this punctual naive estimation method we often speak of "univariate scenario analysis" or "univariate What-If Analysis" and in the special case where only three punctual estimates are used such as the optimistic, expected, pessimistic values:

$$\vec{x}_i(t)=\begin{pmatrix}x_{i,\text{O}}(t)\\x_{i,\text{V}}(t)\\x_{i,\text{P}}(t)\end{pmatrix} \tag{15.4.468}$$

we then speak of "univariate three point estimate analysis".

So in this case sensitivity analysis is done using a large number of probabilistic What-If analysis or Three point estimate analysis. Sadly spreadsheet software like Microsoft® Excel® have made this term and method popular and most people thinks that this tool is the Graal... In fact it is just a high-school level analysis tool that should never be used in a firm if the CEO or CFO are graduated from a good university otherwise you will have quite a lot of problems to keep your job since the results are very poorly predictive and accurate!

Such an analysis is very easy to implement in any spreadsheet software like Microsoft® Excel® with multiple input values. As this is more a high school subject we will not give any example about this.

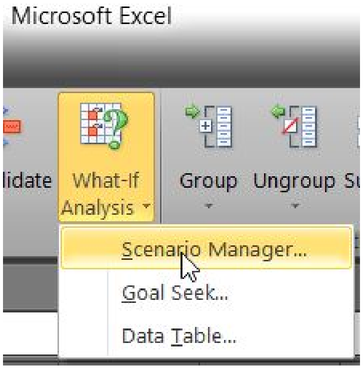


Figure 15.318: What-If Scenario Manager tool in Microsoft® Excel® 14.0.7166

Microsoft never implemented until now an equivalent tool for multivariate cases even if their spreadsheet has the capacity to do it with a little bit more formatting (typically a structure near that one of the PivotTables).

If we apply the same procedure as above but with two variables:

$$\vec{x}_i(t)=\begin{pmatrix} x_{i,1}(t) \\ \vdots \\ x_{i,n}(t) \end{pmatrix} \qquad \vec{x}_j(t)=\begin{pmatrix} x_{j,1}(t) \\ \vdots \\ x_{j,m}(t) \end{pmatrix} \tag{15.4.469}$$

The total number of combinations we can do is obviously:

$$\text{Card}(\vec{x}_i(t))\times\text{Card}(\vec{x}_j(t))=nm \tag{15.4.470}$$

Why speak about this special case??? Because as it can be visualized as a double entry table, many spreadsheet software offer a "Data Table" tool to make "deterministic bivariate direct bias method". Indeed, a spreadsheet software like Microsoft® Excel® provides us a tool in this purpose (for more details, see a Microsoft® Excel® book as this is only basic stuff):

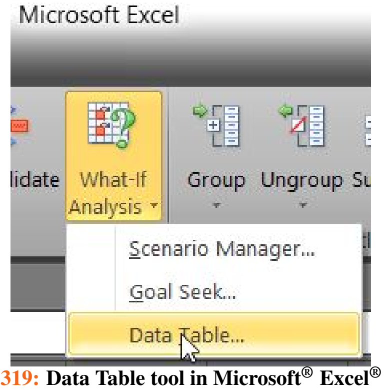
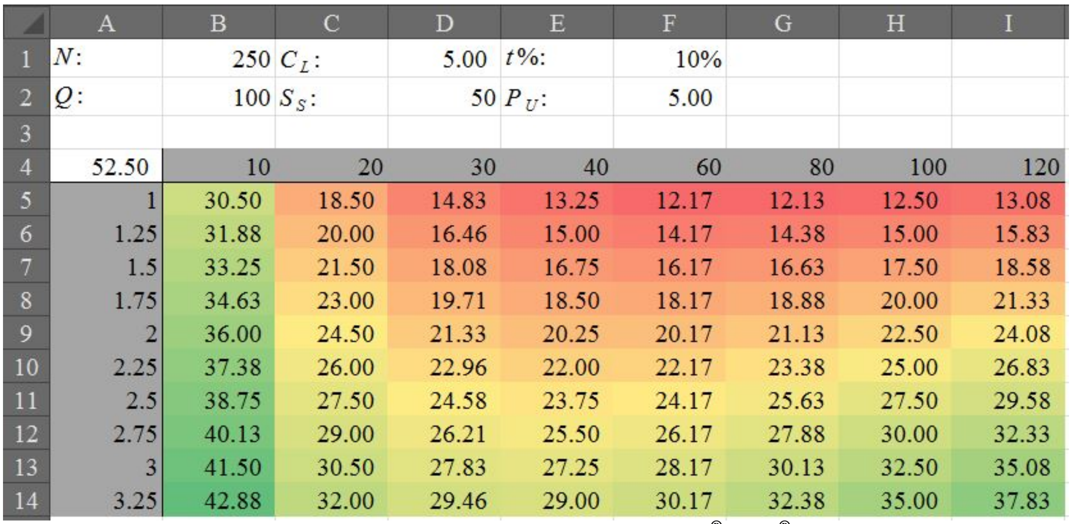


Figure 15.319: Data Table tool in Microsoft® Excel® 14.0.7166

And example based on the minimal order quantity of Wilson's model gives with a little bit formatting:



	A	B	C	D	E	F	G	H	I
1	N:	250	C_L:	5.00	t%:	10%			
2	Q:	100	S_S:	50	P_U:	5.00			
3									
4	52.50	10	20	30	40	60	80	100	120
5	1	30.50	18.50	14.83	13.25	12.17	12.13	12.50	13.08
6	1.25	31.88	20.00	16.46	15.00	14.17	14.38	15.00	15.83
7	1.5	33.25	21.50	18.08	16.75	16.17	16.63	17.50	18.58
8	1.75	34.63	23.00	19.71	18.50	18.17	18.88	20.00	21.33
9	2	36.00	24.50	21.33	20.25	20.17	21.13	22.50	24.08
10	2.25	37.38	26.00	22.96	22.00	22.17	23.38	25.00	26.83
11	2.5	38.75	27.50	24.58	23.75	24.17	25.63	27.50	29.58
12	2.75	40.13	29.00	26.21	25.50	26.17	27.88	30.00	32.33
13	3	41.50	30.50	27.83	27.25	28.17	30.13	32.50	35.08
14	3.25	42.88	32.00	29.46	29.00	30.17	32.38	35.00	37.83

Figure 15.320: Data Table tool illustration with Microsoft® Excel® 14.0.7166

2. The "univariate probabilistic direct bias method" consist simple to admit that $x_i(t)$ is not punctual but fluctuate and follow a density probabilistic function $f_i(t)$:

$$P(X \geq x_i(t)) = F_{i,t}(x_i(t)) \tag{15.4.471}$$

and that:

$$F_{i,t}^{-1}(P(X \geq x_i(t))) = x_i(t) \tag{15.4.472}$$

Using the property we saw in the section of Numerical Methods (see page 6710), we have:

$$\{x_i(t)\} = F_{i,t}^{-1}(\mathcal{U}) \tag{15.4.473}$$

Therefore the idea is not the measure $\Delta y(t)/y(t)$ anymore but based on a simulation of n values:

$$P(Y(t) \leq y(t)) = \hat{F}_n\left(\frac{f((x_1(t), \ldots, F_{i,t}^{-1}(\mathcal{U}), \ldots, x_n(t)), t) \leq y(t)}{n}\right) \tag{15.4.474}$$

If once again the practitioner prefers a "relative input/relative output" then we take:

$$\delta_{i,t} = F_{i,t}^{-1}(\mathcal{U}) \tag{15.4.475}$$

and then the analysis becomes:

$$\begin{aligned} &P\left(\frac{\Delta Y(t)}{Y(t)} \leq \frac{\Delta y(t)}{y(t)}\right) \\ &= \hat{F}_n\left(\frac{\dfrac{f((x_1(t), \ldots, x_i(t)F_{i,t}^{-1}(\mathcal{U}), \ldots, x_n(t)), t) - f((x_1(t), \ldots, x_n(t)), t)}{f((x_1(t), \ldots, x_n(t)), t)} \leq y(t)}{n}\right) \end{aligned} \tag{15.4.476}$$

The information we get therefore is not the same as previously. Before we had a relative variation for the output for a given variation of the input. Now we have a probabilistic distribution of all possible outcomes. These are two distinct things that are complementary!!!

3. The "probabilistic or deterministic univariate forecast bias method" consists simply by doing the same both calculations as before but at any given time t. Therefore the practitioner must make usage of the forecasting techniques we saw in the section of Economy (see page 7896).

   For example the most common usage is a "nominal input/nominal output" with probabilistic bias and therefore we get a stochastic model such that:

$$\begin{aligned} &P(\hat{Y}(t+\Delta t) \leq y(t+\Delta t)) \\ &= \hat{F}_n\left(\frac{f((x_1(t+\Delta t), \ldots, F_{i,t+\Delta t}^{-1}(\mathcal{U}), \ldots, x_n(t+\Delta t)), t) \leq y(t+\Delta t)}{n}\right) \end{aligned} \tag{15.4.477}$$

15.4.7.2 Correlation Method

The correlation method, involves fitting a regression to the model response and using the regression coefficient as direct measures of sensitivity. This method is considered by many consultants as the real "sensitivity analysis".

Let us first recall that the correlation coefficient is given as (see section Statistics page 614):

$$R_{x,y}=\frac{\displaystyle\sum_{k=1}^{n}x_ky_k-\frac{1}{n}\sum_{k=1}^{n}x_k\sum_{k=1}^{n}y_k}{\sqrt{\left[\displaystyle\sum_{k=1}^{n}x_k^2-\frac{1}{n}\left(\sum_{k=1}^{n}x_k\right)^2\right]\left[\displaystyle\sum_{k=1}^{n}y_k^2-\frac{1}{n}\left(\sum_{k=1}^{n}y_k\right)^2\right]}} \tag{15.4.478}$$

and a perfectly equivalent notation already proved in the section of Numerical Methods:

$$R_{x,y}=\frac{\displaystyle\sum_{k}(x_k-\bar{x})(y_k-\bar{y})}{\sqrt{\displaystyle\sum_{k}(x_k-\bar{x})^2\sum_{k}(y_k-\bar{y})^2}} \tag{15.4.479}$$

with for recall:

$$-1<R_{x,y}<+1 \tag{15.4.480}$$

The idea is that if we have a main variable that depends on multiple dependent variables then we calculate the correlation coefficient (or the coefficient of determination which is simply the square of the correlation coefficient) of the main variable with each of the probabilistic simulations of the underlying variables.

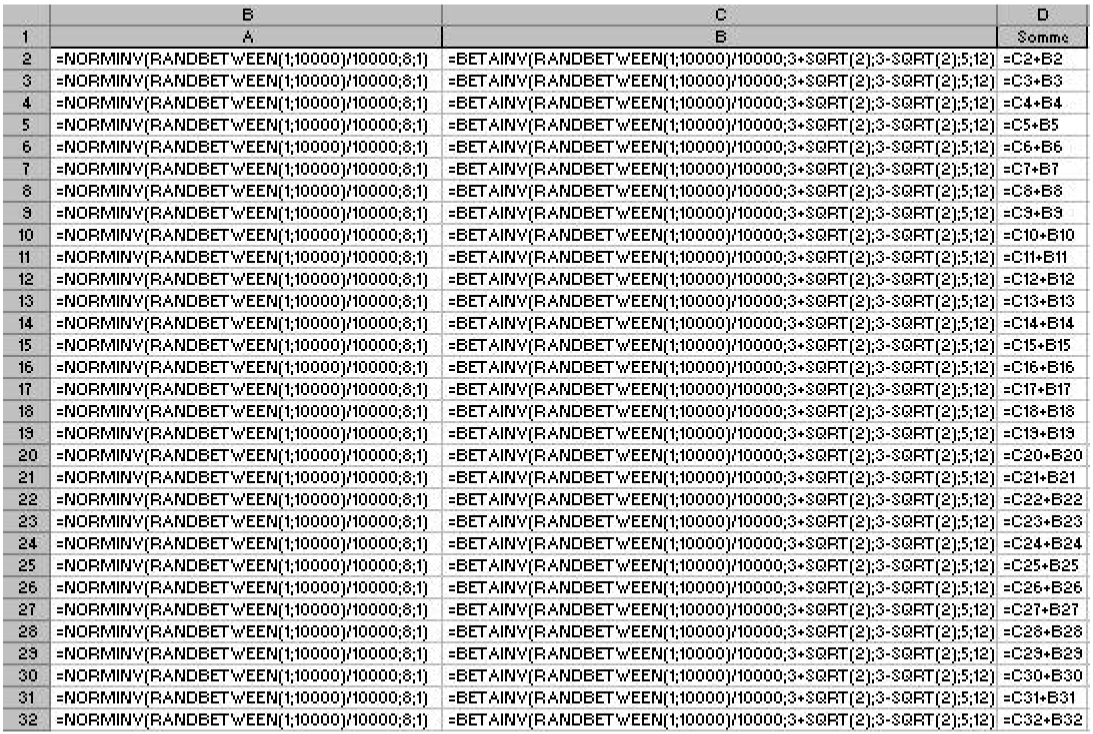
Let us take a simple example that applies to the world of project management (as it is the expertise domain of the author of these lines). Imagine a project schedule that is made only of a critical path (see section Graph Theory page 3016) based solely only on two tasks A and B. Imagine that first task A time execution distribution is the best fitted with a non-truncated Normal law (yes it's stupid but its only for the example!!!) such as:

$$t(A)=\mathcal{N}(\mu,\sigma)=\mathcal{N}(8,1) \tag{15.4.481}$$

that is to say an expected duration of 8 days and a standard deviation of 1 day and the task B is supposed to be spread over a period following a beta distribution (according to the recommendation of PMI and the theory seen previously...) of parameters $(5,12)$. Thus an optimistic duration 5 days and a pessimistic duration of 12 days:

$$P_{a,b}=\frac{(x-a)^{\alpha-1}(x-a)^{\gamma-1}}{(b-a)^{\alpha+\gamma+1}B(\alpha,\gamma)}=\frac{(x-5)^{3+\sqrt{2}-1}(x-5)^{3-\sqrt{2}-1}}{(12-5)^{3+\sqrt{2}+3-\sqrt{2}+1}B(3+\sqrt{2},3-\sqrt{2}} \tag{15.4.482}$$

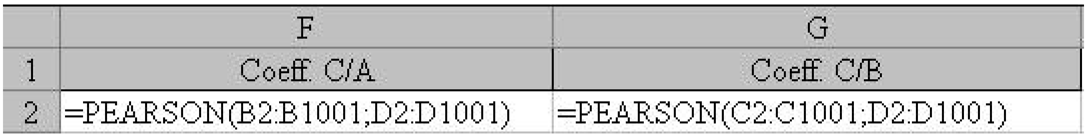
Let us do for an application example with a Monte Carlo simulation (see section Numerical Methods page 6708) of $1,000$ iterations with the English version of Microsoft® Excel® 11.8346 (even if it is most more efficient to use a software like Palisade @Risk):



	B	C	D
1	A	B	Somme
2	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C2+B2
3	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C3+B3
4	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C4+B4
5	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C5+B5
6	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C6+B6
7	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C7+B7
8	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C8+B8
9	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C9+B9
10	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C10+B10
11	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C11+B11
12	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C12+B12
13	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C13+B13
14	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C14+B14
15	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C15+B15
16	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C16+B16
17	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C17+B17
18	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C18+B18
19	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C19+B19
20	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C20+B20
21	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C21+B21
22	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C22+B22
23	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C23+B23
24	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C24+B24
25	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C25+B25
26	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C26+B26
27	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C27+B27
28	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C28+B28
29	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C29+B29
30	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C30+B30
31	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C31+B31
32	=NORMINV(RANDBETWEEN(1;10000)/10000;8;1)	=BETAINV(RANDBETWEEN(1;10000)/10000;3+SQRT(2);3-SQRT(2);5;12)	=C32+B32

Figure 15.321: Simulation of two small tasks respectively with a Normal and Beta distribution

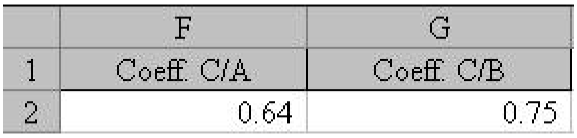
Then we calculate always with the English version of Microsoft® Excel® 11.8346 the correlation coefficient:



	F	G
1	Coeff. C/A	Coeff. C/B
2	=PEARSON(B2:B1001;D2:D1001)	=PEARSON(C2:C1001;D2:D1001)

Figure 15.322: Correlation coefficient between the sums and each task

Which gives:



	F	G
1	Coeff. C/A	Coeff. C/B
2	0.64	0.75

Figure 15.323: Corresponding correlation values

Either as it is customary in the field of sensitivity analysis in the form of a "Tornado chart":

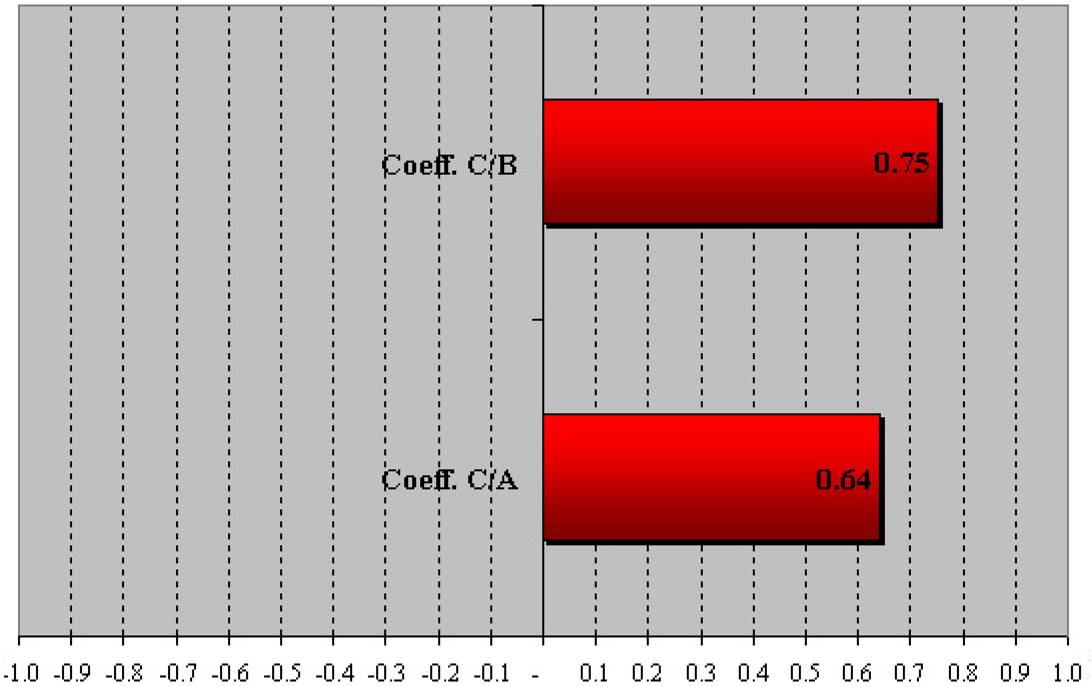


Figure 15.324: **Tornado chart build with Microsoft® Excel® 11.8346**

It is dangerous to use the (Pearson) correlation coefficient when the relationship between the endogenous variable and the exogenous variable is not linear! However, modern softwares can identify almost any other type of relation (exponential, logarithmic, power, polynomial) and make the necessaries transformation to bring the equation to that of the straight line (see section Numerical Methods page 6322). However, this manipulation can also lead to major errors in some cases. The majority of specialized softwares then use the empirical definition of the regression we saw in the section of Numerical Methods (see page 6333):

$$R_{x,y}^2 = 1 - \frac{\text{SSR}}{\text{SST}} \tag{15.4.483}$$

and then use optimization techniques (see section Numerical Methods page 6645) to determine the parameters of the that minimize the R^2 according to previous relation.

A specialized software that the author of these lines like to use and to teach to quickly get this kind of analysis on complex situations is @RISK from Palisade that gives the following graph:

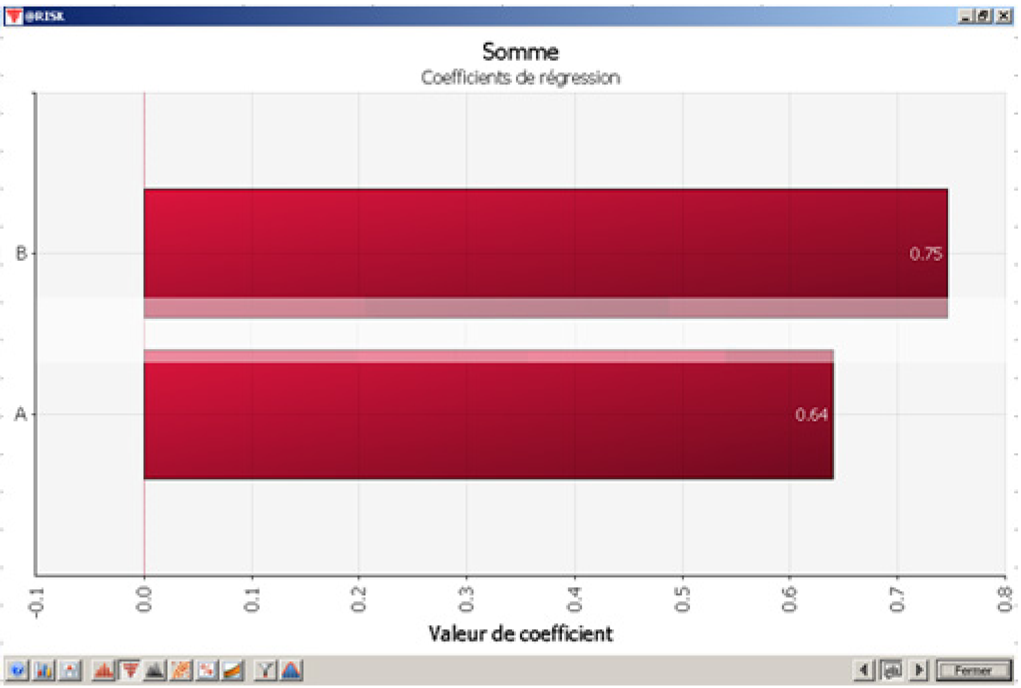


Figure 15.325: Tornado chart of the simulation obtained with @RISK

Finally, when we have a large number of underlying variables, it may be interesting to group the interval correlation coefficients with the following type of graph (example taken from the stock exchange) where the ordinate represents the $\%$ of the number of total input variables concerned (as far as I know @Risk sadly don't do such chart automatically):

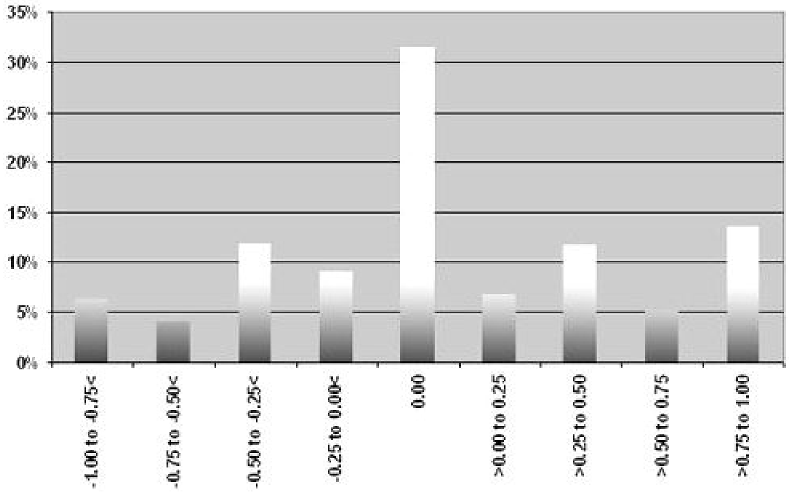


Figure 15.326: Chart for correlation group analysis with MS Excel 11.8346

Thus for this special above example, 32% of the endogenous variables have a zero correlation coefficient with the exogenous variable (which is the performance index of the relevant portfolio) and 14% of the endogenous variables have a correlation coefficient between $+0.75$ and $+1$.

It is also common to represent the average influence in $\%$ of each endogenous variable on the exogenous variable by averaging (or taking the median as more robust indicator) the influence on each other. This gives the following chart if we have a large number of endogenous variable:

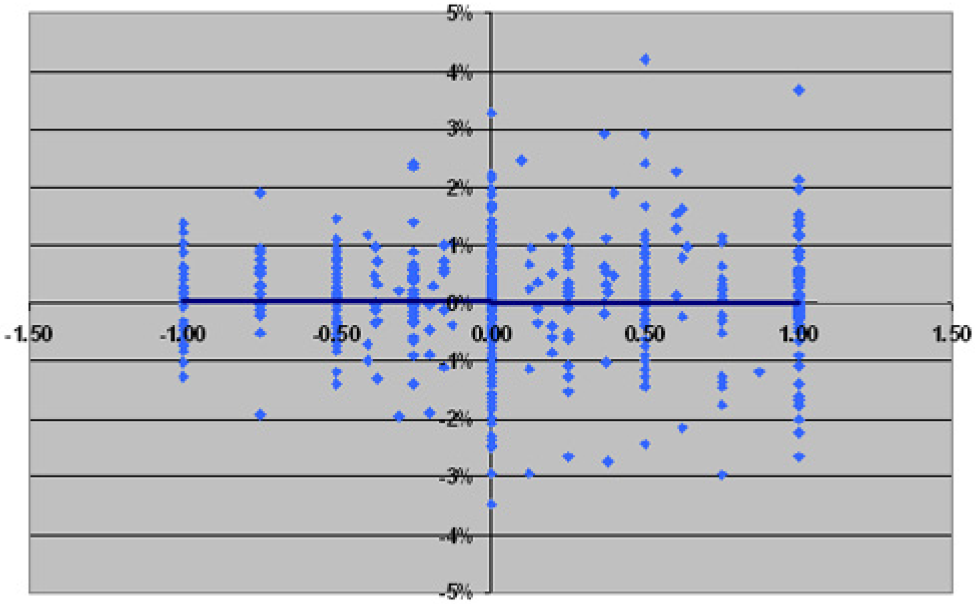


Figure 15.327: **Chart for correlation group analysis with MS Excel 11.8346**

For example, we can see in the graph above a that an endogenous variable has an average impact of $+4\%$ on the portfolio index with correlation of $+0.5$.

Anyway ... the practitioners are free to do as many graph types they want depending on their business needs!

~85 %

★★★★★

28 votes, 65.71%

15.5. Music Mathematics (physics of hearing)

MUSIC theorists sometimes use mathematics to understand music, and although music has no axiomatic foundation in modern mathematics, mathematics is "the basis of sound" and sound itself in its musical aspects... exhibits a remarkable array of number properties, simply because nature itself is amazingly mathematical!

The theory of music encompass all theoretical aspects of a particular music system. There are, not one, but many musical theories, each type of music having its own! All musical system are indeed based on a number of usage, more or less constraint, that could be encompass into a theory, orally or written.

A theory of music has often a religious, philosophical or magic basis and sometimes and arithmetic or scientific starting point (acoustic). It is on this latter that we focus in this section of course...!

We will begin by considering in this section elastic waves in a gas, resulting from pressure variations in the gas. Sound is the most important example of this type of wave which is part of our daily environment.

Caution!!! As we will prove it further below sound doesn't need air to travel, it needs a medium and sound is interpreted by sensors as a change in pressure in that medium (the vacuum is also a medium)! When you hear sounds for example from NASA videos depicting stage separation in space that's for two reasons: 1) the gas ejected creates and change in pressure near the sensor from nothing (i.e. vacuum) to something (hence a sound!) 2) the vibrations of the separation spreads through all the physical components of the rocket (most metals are very medium for sound propagation) hence also another type of sound that superpose to the first one.

15.5.1 Longitudinal Sound Waves

In elastic mediums, gases and liquids, longitudinal sound waves propagate in the following mechanism (for solids this are transverse waves that we have already study in the section of Continuum Mechanics): a middle layer moves in the direction of propagation of the wave (hence the name of "longitudinal") and compresses the next layer, which advance under the effect of pressure and compresses the next layer and so on. This also works for a layer going backwards: the pressure on the next layer decreases, which has the effect of reducing the next layer which in turn decreases the pressure on the next, etc.

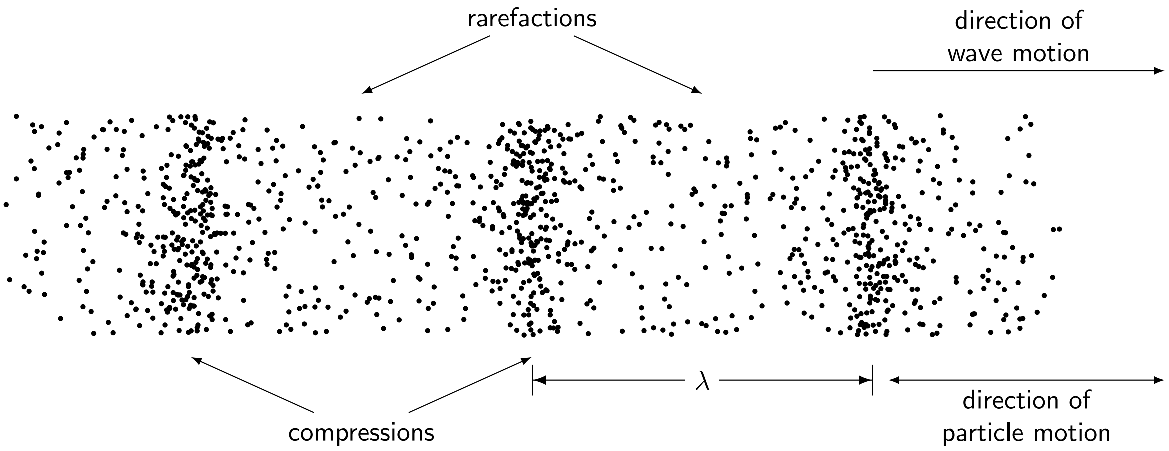


Figure 15.328: **Vocabulary of a longitudinal wave propagation in an ideal gas**

The speed at which the longitudinal waves will move depends as often on the characteristics of the environment. In general, it is lower in gases than in liquids and lower in the liquid than in solids. For example in the air $\sim 340\,[\text{m}\cdot\text{s}^{-1}]$, in water $\sim 1,500\,[\text{m}\cdot\text{s}^{-1}]$ and $\sim 5,000\,[\text{m}\cdot\text{s}^{-1}]$ for the transverse waves in the steel.

Caution!!! In the realm of audiophiles, where music lovers seek sonic nirvana, a murky landscape exists as sound is the most biased of all our senses (with a statistical significant placebo effect). Some self-proclaimed audio gurus on social networks wield buzzwords like "high fidelity", "warmth", and "soundstage", promising transcendent musical experiences. But beware: Many are paid promoters, pushing snake oil dressed as golden cables or mystical crystals. Their goal? To empty the wallet of their audience! Audiophiles often fall mainly into two camps. First, the artists—those who seek emotional connection through music and aren't educate to the Scientific Method nor Cognitive Biases. They care little for technical specs; it's about the soul (...). Then come the scientists—those who crave precision, understanding impedance curves and jitter. But here's the rub: Most practitioners are artists, not scientists. They trust their ears over oscilloscopes. Analog enthusiasts swoon over vinyl warmth, tube amps, and vintage gear. Digital devotees revel in bit-perfect files, DACs, and streaming. But let's be clear: Analog isn't inherently superior. It's a flavor, not a universal truth. Digital doesn't lack soul—it's just a different palette. So educate yourself. Read beyond the hype. Understand that a $5,000 power cable won't magically transport you to a jazz club. Seek reviews grounded in science, not mysticism. And remember, the sweet spot lies where your ears meet your brain, not in the price tag. So, next time an audio guru promises enlightenment through golden connectors, pause, and ask: Is this music or magic?

As regards to the frequencies, there is almost no limit (in positive real values). We can generate sound waves at fractions of Hz (Hertz: oscillations per second) up to hundreds of MHz. By reference to the frequency range audible to humans, we name "infrasound" frequencies lower than $20\,[\text{Hz}]$ and "ultrasound" frequencies greater than $20\,[\text{kHz}]$.

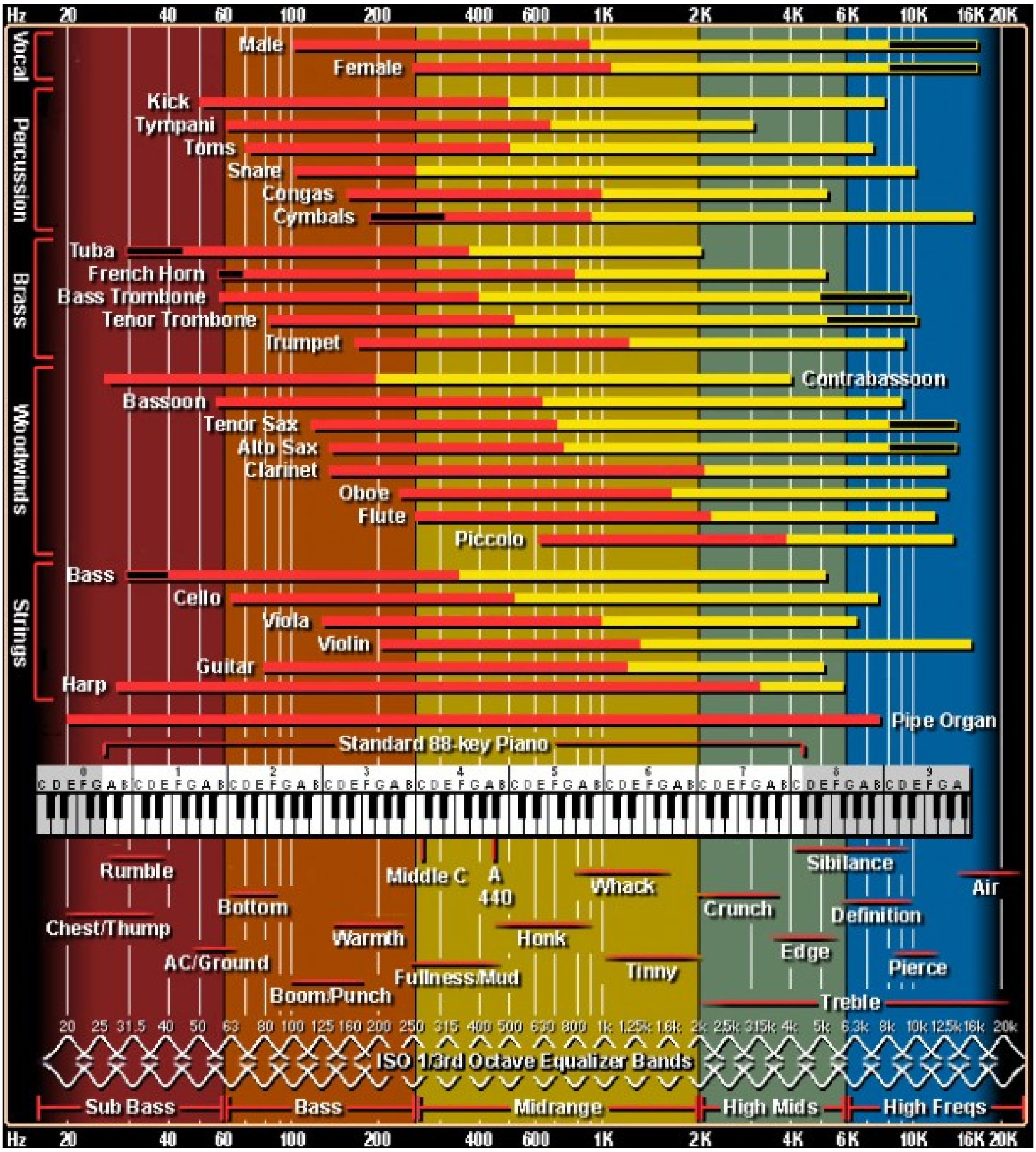


Figure 15.329: **Some instruments frequencies ranges**

In general, the waves are generated by a source of limited dimensions and, from this source, are propagated in all directions. In isotropic mediums the wavefront of a disturbance is spherical. We will avoid introducing spherical coordinates limiting us to the study of parts of wavefronts enough away from the source and sufficiently small compared with the distance to the source so we can assimilate the piece of sphere to a flat surface. That is to say: in this section we will only deal with the longitudinal plane waves.

There is also another important difference between the longitudinal elastic waves in a gas or liquid and the transverse elastic waves in a solid bar. The gas is highly compressible, and if pressure fluctuations are established in a gas, its density will undergo the same type of the pressure fluctuation.

Let us consider now the waves propagating inside a cylindrical tube or pipe (horizontal along the x-axis) of section S. Let us denote by P_0 and ρ_0 the pressure and density of the gas in equilibrium. In these conditions of equilibrium, P_0 and ρ_0 are the same throughout the whole gas volume of the cylinder, that is to say independent of any coordinate.

If the gas pressure is disturbed for example by one of the two opening of the cylinder, a volume element thereof $S\mathrm{d}x$ will intuitively move because the pressures P and P' on the two sides S, S' of this little volume will be different and therefore produce a resultant force (and therefore a power surface density).

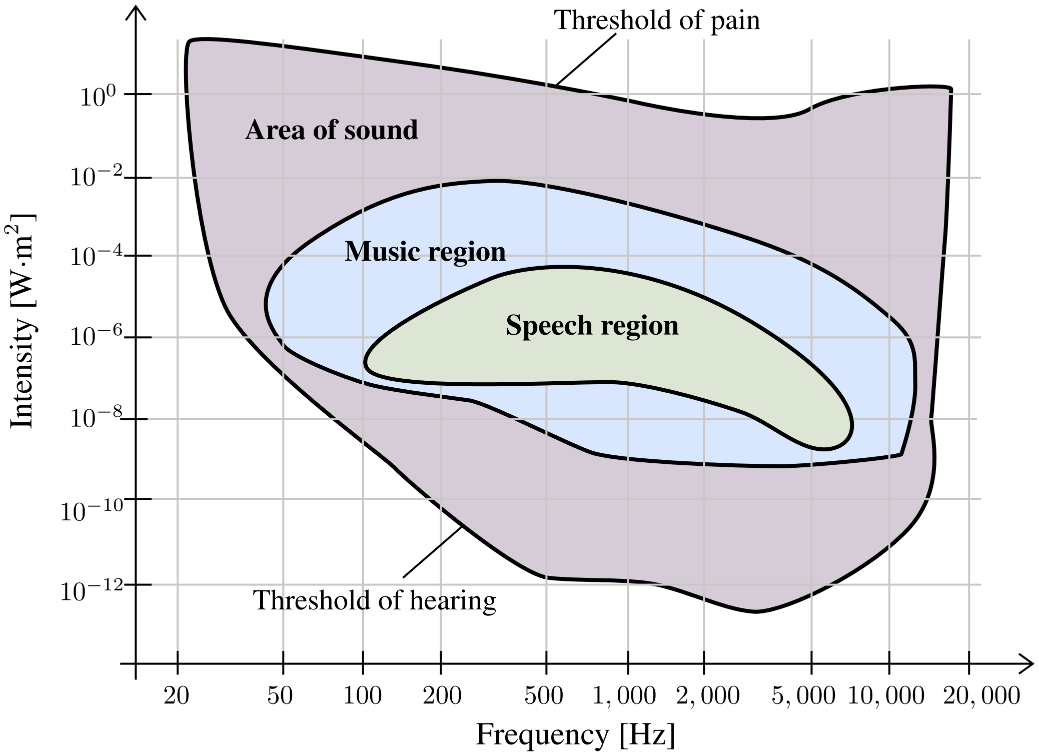


Figure 15.330: **Basic frequency spectrum with transported surface power density**

Remark

Even if they have a very high speed, in a gas, molecules undergo very frequent collisions with each other. They travel in fact less than one micron in average (mean free path), under normal conditions, before hitting another one (see section Continuum Mechanics page 3925 for the proof).

This results in a displacement of the section S of a quantity ξ and of the section S' in an amount ξ' necessarily different because pressure balance has not had time to be done.

Thus, the volume element at the start has a width $\mathrm{d}x$ but after the pressure has changed, it will have a width (if the pressure changes are small) as a first approximation equal to:

$$\mathrm{d}x+(\xi-\xi')=\mathrm{d}x+\mathrm{d}\xi \qquad (15.5.1)$$

However, due to the change in volume, there is also now a density variation due to the compressibility of gas. The mass contained in the volume is initially undisturbed:

$$M_0=\rho_0 S\mathrm{d}x \qquad (15.5.2)$$

If ρ is the density of the disturbed gas, the mass of the perturbed volume is in the end:

$$M=\rho S(\mathrm{d}x+\mathrm{d}\xi) \tag{15.5.3}$$

The conservation of matter claim that these two masses are equal, that is to say:

$$\rho S(\mathrm{d}x+\mathrm{d}\xi)=\rho_0 S\mathrm{d}x \tag{15.5.4}$$

where:

$$\rho\left(1+\frac{\mathrm{d}\xi}{\mathrm{d}x}\right)=\rho_0 \tag{15.5.5}$$

By solving relatively to ρ we get:

$$\rho=\frac{\rho_0}{1+\dfrac{\mathrm{d}\xi}{\mathrm{d}x}} \tag{15.5.6}$$

As we consider only the small changes in pressure relatively to the ambient pressure, $\mathrm{d}\xi/\mathrm{d}x$ is small, then we can replace:

$$\rho=\frac{\rho_0}{1+\dfrac{\mathrm{d}\xi}{\mathrm{d}x}} \tag{15.5.7}$$

by its limited Taylor series development (see section Sequences and Series page 1703):

$$\rho=\rho_0\left(1-\frac{\mathrm{d}\xi}{\mathrm{d}x}+\left(\frac{\mathrm{d}\xi}{\mathrm{d}x}\right)^2-\ldots\right)\cong\rho_0\left(1-\frac{\mathrm{d}\xi}{\mathrm{d}x}\right) \tag{15.5.8}$$

Therefore:

$$\frac{\rho-\rho_0}{\rho_0}=-\frac{\mathrm{d}\xi}{\mathrm{d}x} \tag{15.5.9}$$

By now admitting that the pressure is only connected to the density (for simplicity) we can write:

$$P=f(\rho) \tag{15.5.10}$$

Using the general univariate form of the Taylor expansion (see section Sequences and Series page 1703): Therefore:

$$f(x)=\sum_{k=0}^{n}\frac{1}{k!}\frac{\mathrm{d}^k f(x_0)}{\mathrm{d}x^k}(x-x_0)^k \tag{15.5.11}$$

Then we have:

$$P=\sum_{k=0}^{n}\frac{1}{k!}\frac{\mathrm{d}^k f(\rho)}{\mathrm{d}\rho^k}(\rho-\rho)^k\cong P_0+(\rho-\rho_0)\frac{\mathrm{d}P(\rho_0)}{\mathrm{d}\rho} \tag{15.5.12}$$

The quantity:

$$\frac{1}{\kappa}=\rho_0\left(\frac{\mathrm{d}P}{\mathrm{d}\rho}\right)_{\rho=\rho_0} \tag{15.5.13}$$

usually denoted:

$$\kappa=\frac{1}{\rho_0}\left(\frac{\mathrm{d}\rho}{\mathrm{d}P}\right)_{\rho=\rho_0} \tag{15.5.14}$$

is named "coefficient of compressibility" or more technically "coefficient of isothermal compressibility".

Let us recall now that we have proved in the section of Weather and Marine Engineering during our study of the adiabatic atmospheric model that starting from Laplace's law we get:

$$\frac{\mathrm{d}P}{\mathrm{d}\rho}=\gamma\frac{P}{\rho_0} \tag{15.5.15}$$

where for recall γ is the "Laplace coefficient", also named "adiabatic coefficient" defined by (see section Thermodynamics page 3649):

$$\gamma=\frac{C_P}{C_V} \tag{15.5.16}$$

Remark

The prior-previous relation can often been found in the literature in the following form:

$$\Delta P=\gamma P_0\frac{\Delta V}{V_0} \tag{15.5.17}$$

Then we get (relation that we will use later):

$$\frac{1}{\kappa}=\rho_0\gamma\frac{P}{\rho_0}\Rightarrow\frac{1}{\kappa}=\gamma P \tag{15.5.18}$$

The isothermal compressibility coefficient will obviously be written in the general case:

$$\kappa=\frac{1}{\rho}\left(\frac{\mathrm{d}\rho}{\mathrm{d}P}\right) \tag{15.5.19}$$

By the way, let us notice that by changing to the density to the volume, we get:

$$\kappa=\frac{1}{\rho}\left(\frac{\mathrm{d}\rho}{\mathrm{d}P}\right)=\frac{1}{\dfrac{m}{V}}\left(\frac{\mathrm{d}\left(\dfrac{m}{V}\right)}{\mathrm{d}P}\right)=V\frac{\mathrm{d}\left(\dfrac{1}{V}\right)}{\mathrm{d}P}=V\frac{-\dfrac{1}{V^2}\mathrm{d}V}{\mathrm{d}P}=-\frac{1}{V}\left(\frac{\mathrm{d}V}{\mathrm{d}P}\right) \tag{15.5.20}$$

often denoted χ in the field of Thermodynamics.

Let us return to the two relations:

$$\frac{1}{\kappa}=\rho_0\left(\frac{\mathrm{d}P}{\mathrm{d}\rho}\right)_{\rho=\rho_0}\qquad\text{and}\qquad\frac{\rho-\rho_0}{\rho_0}=-\frac{\mathrm{d}\xi}{\mathrm{d}x} \tag{15.5.21}$$

and let inject them into the Taylor series above. Then we have:

$$P\cong P_0+\frac{(\rho-\rho_0)}{\rho_0}\frac{1}{\kappa}=P_0+\frac{-\rho_0\dfrac{\mathrm{d}\xi}{\mathrm{d}x}}{\rho_0}\frac{1}{\kappa}=P_0-\frac{\mathrm{d}\xi}{\mathrm{d}x}\frac{1}{\kappa} \tag{15.5.22}$$

This expression links the pressure at any point of the gas to the deformation at the same point.

We then need the equation of motion of the volume element. The mass of the element is $\rho_0 S\mathrm{d}x$ and its acceleration $\mathrm{d}\xi^2/\mathrm{d}t^2$.

We have naturally in terms of force (the minus sign indicates that the force change the pressure is opposed to the initial pressure in the cylinder):

$$-S\mathrm{d}P=(\rho_0 S\mathrm{d}x)\frac{\mathrm{d}^2\xi}{\mathrm{d}t^2} \tag{15.5.23}$$

Therefore:

$$\frac{\mathrm{d}P}{\mathrm{d}x}=-\rho_0\frac{\mathrm{d}^2\xi}{\mathrm{d}t^2} \tag{15.5.24}$$

In this problem, there are two fields, the displacement field ξ and the pressure field P. We can combine them as follows by taking:

$$P\cong P_0-\frac{\mathrm{d}\xi}{\mathrm{d}x}\frac{1}{\kappa} \tag{15.5.25}$$

and differentiating with respect to x and remembering that P_0 is independent of the position in the gas. Then we have:

$$\frac{\mathrm{d}P}{\mathrm{d}x}=-\frac{1}{\kappa}\frac{\mathrm{d}^2\xi}{\mathrm{d}x^2} \tag{15.5.26}$$

What we can combine with:

$$\frac{\mathrm{d}P}{\mathrm{d}x}=-\rho_0\frac{\mathrm{d}^2\xi}{\mathrm{d}t^2} \tag{15.5.27}$$

We thus find the final expression of the wave equation of (for recall) a Poisson's equation (more particularly it is a D'Alembert's equation):

$$\boxed{\frac{\mathrm{d}^2\xi}{\mathrm{d}t^2}=\frac{1}{\kappa\rho_0}\frac{\mathrm{d}^2\xi}{\mathrm{d}t^2}} \tag{15.5.28}$$

So we get a similar relation to that we got in the section of Wave Mechanics for mechanical waves or in the section of Electrodynamics for electromagnetic waves.

We conclude that displacement due to pressure disturbance in a gas spreads to speed:

$$v=\sqrt{\frac{1}{\kappa\rho_0}}=\sqrt{\frac{\gamma P}{\rho_0}}=\sqrt{\frac{\gamma}{\rho_0}\frac{nkT}{V}}=\sqrt{\frac{\gamma NkT}{M}}=\sqrt{\frac{\gamma nRT}{M}}=\sqrt{\frac{\gamma RT}{M_m}} \tag{15.5.29}$$

where for recall the gas constant worth $8.314\,[\mathrm{J}\cdot\mathrm{K}^{-1}\cdot\mathrm{mol}^{-1}]$.

The relation:

$$\boxed{v=\sqrt{\frac{\gamma P}{\rho_0}}} \tag{15.5.30}$$

is sometimes named "Newton-Laplace speed of sound relation".

✍**Example:**

If we consider the air as an ideal diatomic gas then (see section Thermodynamics page 3648) we have $\gamma \cong 7/5 = 1.4\ldots$ of molar mass $29 \cdot 10^3\,[\mathrm{kg}\cdot\mathrm{mol}^{-1}]$ (weighted average molecular weights of N_2 and O_2). It comes then at a temperature of $300\,[\mathrm{K}]$:

$$v = \sqrt{\frac{\gamma P}{\rho_0}} = 347\,[\mathrm{m}\cdot\mathrm{s}^{-1}] \tag{15.5.31}$$

This is in perfect agreement with the experience! By cons, when we say that a plane flies at Mach 2 (Mach is the ratio between the speed of an airplane and that of sound as we will see further below during our study of wave shocks), we do not really know in reality the speed of sound (neither that of the plane) nor the temperature.

But, we also have the following result (see section Continuum Mechanics page 3761):

$$\kappa = \frac{3(1-2\eta)}{E} \tag{15.5.32}$$

But, we have discussed in the section of Mechanical Engineering (see page 3761) that the Poisson coefficient η respects:

$$\eta < \frac{1}{2} \tag{15.5.33}$$

Taking the approximation that for a gas $\eta \cong 1/3$... then we have:

$$\kappa = \frac{3(1-2\eta)}{E} \cong \frac{1}{E} \tag{15.5.34}$$

and therefore:

$$\boxed{v = \sqrt{\frac{E}{\rho_0}}} \tag{15.5.35}$$

The speed of sound is given by the same type of expression for fluids or solids! Thus, the propagation of a longitudinal deformation in a solid is given by:

$$\boxed{\frac{\mathrm{d}^2\xi}{\mathrm{d}t^2} = \frac{E}{\rho_0}\frac{\mathrm{d}^2\xi}{\mathrm{d}x^2}} \tag{15.5.36}$$

and for the transverse strain (see section Continuum Mechanics page 3765):

$$\boxed{\frac{\partial^2 y}{\partial t^2} = \frac{G}{\rho_0}\frac{\partial^2 y}{\partial x^2}} \tag{15.5.37}$$

We also have:

$$\begin{aligned} \frac{\partial^2}{\partial t^2}P &\cong \frac{\partial^2}{\partial t^2}\left(P_0 - \frac{1}{\kappa}\frac{\mathrm{d}\xi}{\mathrm{d}x}\right) = -\frac{1}{\kappa}\frac{\partial^2}{\partial t^2}\left(\frac{\mathrm{d}\xi}{\mathrm{d}x}\right) \\ \frac{\partial^2}{\partial x^2}P &\cong \frac{\partial}{\partial x}\left(-\rho_0\frac{\mathrm{d}^2\xi}{\mathrm{d}t^2}\right) = -\rho_0\frac{\partial}{\partial x}\left(\frac{\mathrm{d}^2\xi}{\mathrm{d}t^2}\right) = -\rho_0\frac{\partial^2}{\partial t^2}\left(\frac{\partial\xi}{\partial x}\right) \end{aligned} \tag{15.5.38}$$

Dividing these equalities member by member, we get:

$$\frac{\dfrac{\partial^2 P}{\partial t^2}}{\dfrac{\partial^2 P}{\partial x^2}}=\frac{-\dfrac{1}{\kappa}\dfrac{\partial^2}{\partial t^2}\left(\dfrac{\mathrm{d}\xi}{\mathrm{d}x}\right)}{-\rho_0\dfrac{\partial^2}{\partial t^2}\left(\dfrac{\partial\xi}{\partial x}\right)} \tag{15.5.39}$$

Thus after simplification:

$$\boxed{\frac{\partial^2 P}{\partial t^2}=\frac{1}{\kappa\rho_0}\frac{\partial^2 P}{\partial x^2}} \tag{15.5.40}$$

For laboratories usage, the sound wave model is more convenient than the previous one as we measure more easily pressure variations in a liquid or a gas, than the movements of molecules. It is interesting to know that all the analyses we have made with the string wave equation (see section Wave Mechanics page 3476) or electromagnetic waves (see section Electrodynamics page 4138) can also be applied to the equation above (energetic aspects, particular solutions, decrease in amplitude in function on the distance from the source, etc.).

The movement of the waves in gases is an adiabatic process (see section Thermodynamics page 3602) so there is no exchange of energy in heat form by the elementary volume of gas.

Remark

A "pure sound" (a perfect sine wave of a single, constant frequency) is a mathematical ideal that is not perfectly achievable in the physical world. However, it is possible to create a sound that is so close to a pure tone that the human ear cannot perceive the difference.

15.5.1.1 Power carried by a sound wave

We have seen just earlier above that:

$$\frac{\rho-\rho_0}{\rho_0}=-\frac{\mathrm{d}\xi}{\mathrm{d}x} \tag{15.5.41}$$

Let us write the density variations caused by the sound wave by:

$$\rho_s=-\rho_0\frac{\mathrm{d}\xi}{\mathrm{d}x} \tag{15.5.42}$$

We will do the calculation of the power carried by a wave only in the case of sine waves here. In this case ξ will be of the form:

$$\xi=\xi_0\cos(\omega t-kx) \tag{15.5.43}$$

As ξ is known, we can calculate the corresponding pressure by using:

$$P\cong P_0-\frac{\mathrm{d}\xi}{\mathrm{d}x}\frac{1}{\kappa} \tag{15.5.44}$$

By introducing the pressure variations due to the sound wave:

$$P_s = -\frac{1}{\kappa}\frac{\mathrm{d}\xi}{\mathrm{d}x} \tag{15.5.45}$$

and by using:

$$v = \sqrt{\frac{1}{\kappa\rho_0}} \tag{15.5.46}$$

It comes:

$$P_s = -\rho_0 v^2 \frac{\mathrm{d}\xi}{\mathrm{d}x} \tag{15.5.47}$$

Then we have:

$$P_s = \rho_0 v^2 \frac{\mathrm{d}(\xi_0 \cos(\omega t - kx))}{\mathrm{d}x} = -\rho_0 k v^2 \xi_0 \sin(\omega t - kx) = -P_0 \sin(\omega t - kx) \tag{15.5.48}$$

To calculate the power carried by a sound wave, we will calculate the work done, during a period, on a surface S on a plane perpendicular to the x-axis and of coordinates x.

The force applied on this surface will be:

$$F = SP_s \tag{15.5.49}$$

The work done when the surface moves of $\mathrm{d}\xi$ will be:

$$\mathrm{d}W = F\mathrm{d}\xi = SP_s\mathrm{d}\xi \tag{15.5.50}$$

As we are doing a calculation for the displacements ξ of a layer whose equilibrium position x does not vary, only the variable t varies:

$$\mathrm{d}\xi = \frac{\partial \xi}{\partial t}\mathrm{d}t = -\omega\xi_0 \sin(\omega t - kx)\mathrm{d}t \tag{15.5.51}$$

Substituting we get:

$$\begin{aligned}\mathrm{d}W &= SP_s\mathrm{d}\xi = -SP_s\omega\xi_0 \sin(\omega t - kx)\mathrm{d}t \\ &= -S(-\rho_0 k v^2 \xi_0 \sin(\omega t - kx))\omega\xi_0 \sin(\omega t - kx)\mathrm{d}t \\ &= S\rho_0 k v^2 \omega \xi_0^2 \sin^2(\omega t - kx)\mathrm{d}t\end{aligned} \tag{15.5.52}$$

The word applied during one period $T = 2\pi/\omega$ will be:

$$W = S\rho_0 k v^2 \omega \xi_0^2 \int\limits_0^{\frac{2\pi}{\omega}} sin^2(\omega t - kx)\mathrm{d}t \tag{15.5.53}$$

We have already proved the expression of such an integral in the section of Differential and Integral Calculus:

$$\int \sin^2(\omega t - kx)\mathrm{d}t = \frac{1}{\omega}\int \sin^2(y)\mathrm{d}y = \frac{1}{2\omega}(-\cos(y)\sin(y) + y) \tag{15.5.54}$$

Hence:

$$\begin{aligned}\int_0^{\frac{2\pi}{\omega}} \sin^2(\omega t-kx)\mathrm{d}t &= \left(\frac{1}{2\omega}\left(-\cos(\omega t-kx)\sin(\omega t-kx)+(\omega t-kx)\right)\right)\Bigg|_0^{\frac{2\pi}{\omega}}\\ &=\frac{1}{2\omega}\left(-\cos\left(\omega\frac{2\pi}{\omega}-kx\right)\sin\left(\omega\frac{2\pi}{\omega}-kx\right)+\left(\omega\frac{2\pi}{\omega}-kx\right)\right)\\ &-\frac{1}{2\omega}(-\cos(-kx)\sin(-kx)+(-kx))\\ &=\frac{1}{2\omega}(\cos(kx)\sin(kx)+2\pi-kx)-\frac{1}{2\omega}(\cos(kx)\sin(kx)-kx)=\frac{\pi}{\omega}\end{aligned} \tag{15.5.55}$$

Replacing it comes:

$$W=S\rho_0 kv^2\omega\xi_0^2\frac{\pi}{\omega}=S\rho_0 kv^2\pi\xi_0^2 \tag{15.5.56}$$

To calculate the power, we have to divide the work by the time during which it was made:

$$T=\frac{2\pi}{\omega} \tag{15.5.57}$$

and to calculate the power transmitted per unit area, we have to divide by the surface:

$$\frac{P}{S}=\frac{S\rho_0 kv^2\pi\xi_0^2}{\dfrac{2\pi}{\omega}S}=\frac{\rho_0 kv^2\omega}{2}\xi_0^2 \tag{15.5.58}$$

If we replace k by:

$$k=\frac{\omega}{v} \tag{15.5.59}$$

We get:

$$\frac{P}{S}=\frac{\rho_0 v\omega^2}{2}\xi_0^2 \tag{15.5.60}$$

Before we continue, notice that there are several ways to measure the amplitude of a sound, and by extension, any signal of wave nature:

- The average amplitude (the arithmetic mean value of the positive signal)
- The effective amplitude (equivalent continuous amplitude in power)
- The peak amplitude (maximum positive)
- The peak to peak amplitude (the maximum difference of positive and negative amplitude)

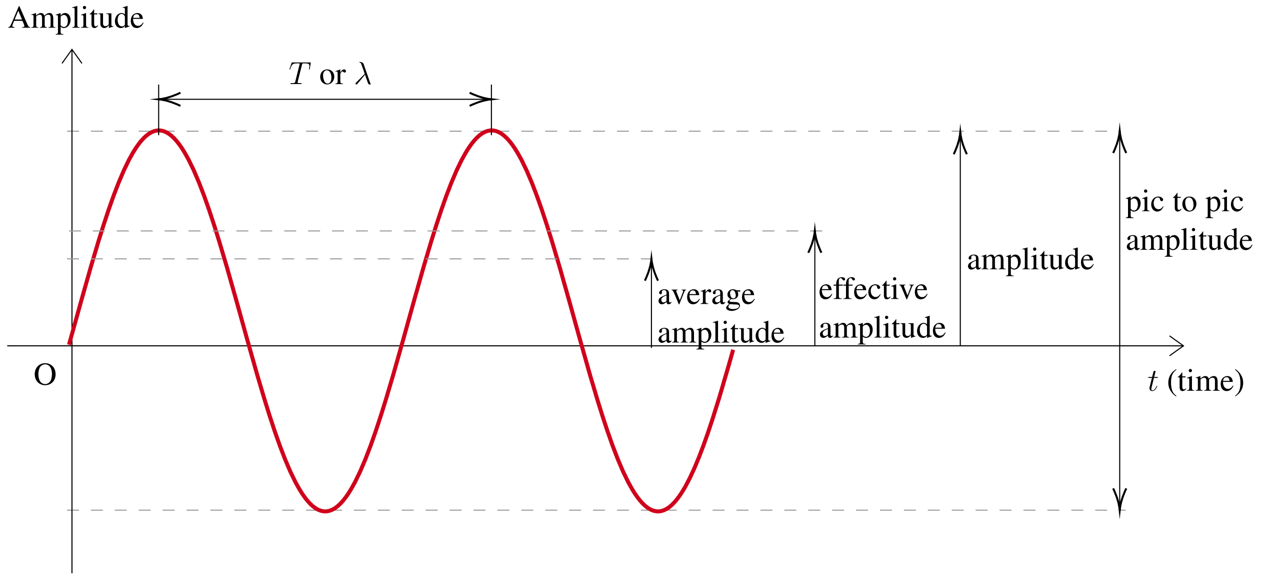


Figure 15.331: **Various types of amplitudes**

In practice, the average amplitude of little interest and is not used to much. However, the effective value or "Root Mean Square", that is the square root of the mean square value of the signal is universally adopted to measure the value of AC voltages (it is like the variance in statistics in fact!), within the general framework than in acoustic. An amplifier which is given with the indication "10 watts RMS" will have approximately 14 watts peak and approximately 28 watts peak-to-peak (also denoted by "pp"). Measurements of power peak to peak are quite often named "music watts" by the audio-video retailers because the values are quite nice in terms of communication/marketing..

The unit of measurement of the amplitude depends on the physical measured quantity:

- For a vibrating string, it's a distance.
- For a sound wave is the air pressure, or movement of the diaphragm.
- For the electromagnetic radiation, the amplitude corresponds to the electric field.
- For an electrical signal that corresponds to the maximum value of the current intensity.

For sinusoidal regimes (or more generally any periodic signal!), it is easy to prove that regardless of the field in physics the effective value (square root of the mean square of the signal) is the peak value divided by the square root of two. Indeed:

$$\xi_{\text{eff}}=\sqrt{\frac{1}{T}\int\limits_0^T \xi^2(t)\mathrm{d}t}=\sqrt{\frac{1}{2\pi}\int\limits_0^{2\pi} \xi^2(\theta)\mathrm{d}\theta}=\sqrt{\frac{1}{2\pi}\xi_0^2\pi}=\frac{\xi_0}{\sqrt{2}} \tag{15.5.61}$$

Therefore:

$$\boxed{\text{RMS}=\frac{A}{\sqrt{2}}} \tag{15.5.62}$$

Hence in RMS (this is only an arbitrary choice and nothing else!!!... so that you understand what will follow!!), the prior previous relations is:

$$\boxed{\frac{P}{S}=\rho_0 v\omega^2\xi_{\text{eff}}^2=\rho_0 kv^2\omega\xi_{\text{eff}}^2} \tag{15.5.63}$$

As we have the amplitude of the pressure which is equal to:

$$P_0=\rho_0 kv^2\xi_0=\rho_0\omega v\xi_0 \tag{15.5.64}$$

and its RMS by:

$$\boxed{P_{\text{eff}}=\rho_0 kv^2\xi_{\text{eff}}=\rho_0\omega v\xi_{\text{eff}}} \tag{15.5.65}$$

Finally, we have for the power (that we will represent later by a stylized capital $\mathcal{P}$ letter in order not to be confused with the pressure P):

$$\boxed{\frac{\mathcal{P}}{S}=\rho_0 vc\omega^2\xi_{\text{eff}}^2=\frac{1}{\rho_0 v}\omega\xi_{\text{eff}})^2=\frac{1}{\rho_0 v}P_{\text{eff}}^2} \tag{15.5.66}$$

15.5.1.2 Measuring the intensity of a sound

To characterize the sound, a unit of measurement was invented: the "Bel" and its submultiple the "decibel" which is $1/10$ of Bel. This acoustic intensity, assimilated to the "loudness" was defined from the effective sound pressure (the P_{eff} given above) and of the sound pressure of a wave (at a specific frequency!) at the limit of the hearing threshold of a small elite of humanity (about 10%).

This effective pressure reference is:

$$P_{\text{eff,ref}}=2\cdot 10^{-5}\ [\text{Pa}] \tag{15.5.67}$$

The loudness of a wave whose RMS sound pressure is P_{eff} is by definition and by convention equal to:

$$I_S=\log_{10}\left(\frac{P_{\text{eff}}}{P_{\text{eff,ref}}}\right)^2=2\log_{10}\left(\frac{P_{\text{eff}}}{P_{\text{eff,ref}}}\right)=2\log_{10}\left(\frac{P_{\text{eff}}}{2\cdot 10^{-5}}\right)\ [\text{Bels}] \tag{15.5.68}$$

and is often simply denoted S in the literature to indicate that it is the "physiological sensation" of the human ear. How did we get to this relation? Well by observing on a panel of individuals that the physiological sensation of the ear was not linear but that the absolute variations of the sensation are proportional to relatives pressure variations such that:

$$\mathrm{d}I_S=c^{te}\frac{\mathrm{d}P}{P} \tag{15.5.69}$$

Therefore after integration:

$$I_S-I_{S,0}=c^{te}\ln\left(\frac{P}{P_0}\right) \tag{15.5.70}$$

where by convention the minimum sensitivity is put as being equal to zero and where for historical reasons we take the decimal logarithms instead of Napierian. It comes then:

$$I_S = c^{te} \log\left(\frac{P}{P_0}\right) \tag{15.5.71}$$

In some books, we find the definition of the Bel from a reference intensity of (at 1 [kHz]):

$$1\,[\text{Bel}] = 10^{-12}\,[\text{W}\cdot\text{m}^{-2}] \tag{15.5.72}$$

It comes therefore (the definition is equivalent because it is a ratio and that power is directly proportional to the square of pressure as we have proved earlier above):

$$I_S = \log\left(\frac{\mathcal{P}/S}{\mathcal{P}_{\text{eff,ref}}}\right)^2 = 20\log\left(\frac{P_{\text{eff}}}{2\cdot 10^{-5}}\right)\;[\text{dB}] \tag{15.5.73}$$

The decibel is then a tenth of Bel ... hence its name as already mentioned!

✍**Example:**

If an apparatus has a loudness of 122 [dB] (loudest existing megaphone in year 12020 according to holocene calendar) and another a loudness of 144 [dB] (loudness of Hyperspike HS-10-R), then their power ratio is:

$$10^{\frac{144-122}{20}} = 12.58$$

The reason of the logarithmic choice is that the it has been observed that the auditory sensation of human ear is also logarithmic (this is quite a beautiful engineering of Nature): we have the same sensation of sound increase when it goes from 1 to 10 as when it moves from 10 to 100.

To give an idea of the value of a decibel, here is a table of typical values (for a given frequency and a given reference pressure):

Decibels	Source/Effect
310	Krakatoa eruption (1883)
235	Tongal eruption (2022)
180	Rocket launch
160	(damage to the eardrum)
140	Shotgun blast (pain threshold)
130	Jet engine 30 meters away
120	Rock concert (threshold of discomfort)
110	Car horn, snowblower
100	Blow dryer, subway, helicopter, chainsaw ,heavy machinery
90	Motorcycle, lawn mower, convertible ride on highway
85	Trucks
80	Factory, noisy restaurant, vacuum, screaming child
75	Average factory
70	car, alarm clock, city traffic
60	Conversation, dishwasher,
50	Moderate rainfall
40	Cinema audience noise/Refrigerator
20	Broadcasting Studio/Watch ticking
10	Anechoic room
0	Threshold of hearing

Table 15.85: **Examples of typical dB values**

As we have already say it the minimum effective sound pressure which causes auditory sensation is:

$$2\cdot 10^{-5}\text{ [Pa]} \tag{15.5.74}$$

that is to say approximately 10^{-10} times smaller than atmospheric pressure. It must nevertheless be noticed that the frequency of this wave is around 3 [kHz] where the sensitivity is maximum for the human ear in the 120th century (holocene calendar).

It is very interesting to calculate at which amplitude of effective displacement that corresponds. We then use the relation proved above:

$$P_{\text{eff}}=\rho_0\omega v\xi_{\text{eff}} \tag{15.5.75}$$

So for the limit case at 1 [kHz], at the standard conditions for temperature and pressure, we will have a displacement of (values taken from the tables of the swiss-french mathematical commission):

$$\xi_{\text{eff}}=\frac{P_{\text{eff}}}{\rho_0\omega v}=\frac{2\cdot 10^{-5}}{1.293\cdot 2\pi\cdot 1000\cdot 331}\cong 7.43\cdot 10^{-12}\text{ [m]} \tag{15.5.76}$$

and the peak amplitude value is therefore:

$$x_0=\xi_{\text{eff}}\sqrt{2}\cong 1\cdot 10^{-11}\text{ [m]} \tag{15.5.77}$$

This value is less than the radius of atoms according to the Dalton model. Nature has given us a ear with an exquisite sensitivity!

It is easy to calculate at which power value per unit area corresponds a sound wave at the limit of hearing. Using the relation proved above, we get:

$$\frac{\mathcal{P}}{S}=\frac{1}{\rho_0 v}P_{\text{eff}}^2=\frac{1}{1.293\cdot 331}(2\cdot 10^{-5})^2=1\,[\text{pW}\cdot\text{m}^{-2}] \tag{15.5.78}$$

Since the surface of the section of the ear canal is less than one square centimetre, the power that comes to the eardrum is lower than 10^{-6} [W].

15.5.2 Spherical Sound Waves

In reality the sound waves are generated by sources of finite extent. At lower or comparable distances to the extent of the source, the shape of the wavefront that moves away from the source can be very complicated (imagine the wavefront of a thunder caused by choppy Flash). But far from the source, it is seen as a point object and the wave-front becomes more spherical as one moves away from the source.

The difference is that the energy carried by the sound wave is distributed in an area which expands as R^2 (where R is the distance to the source) as for the light (electrodynamics waves) emitted by stars. So the power per unit area should decrease as $1/R^2$ (exactly according to the same mathematical development than for electromagnetic waves seen in the section of Electrodynamics) and as the power per unit area is proportional to ξ_{eff}^2 or to P_{eff}^2 this implies that a spherical wave, both displacement ξ_{eff}, as the sound pressure P_{eff} must decrease as $1/R$.

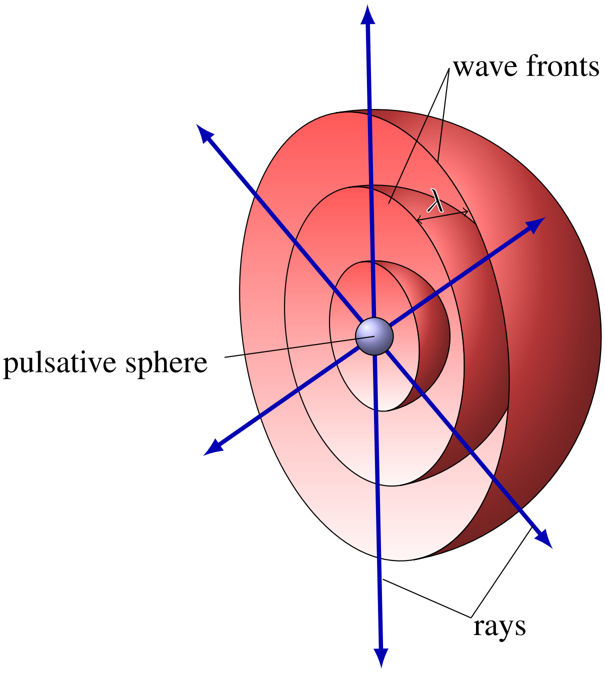


Figure 15.332: **Spherical wavefront (author: Juan Castaño)**

In the case of sine waves, we will have to modify the wave equation:

$$\xi=\xi_0\cos(\omega t-kx) \tag{15.5.79}$$

by adding a coefficient $1/R$. If we denote by r the distance between the source and the observation position, and that, instead of measuring the position with x we do it with r we get:

$$\xi=\xi_0\frac{\cos(\omega t-kr)}{r} \tag{15.5.80}$$

and also for the sound pressure:

$$P_S=-P_0\frac{\sin(\omega t-kr)}{r} \tag{15.5.81}$$

For those who want more details, let us recall that in the section of Calculus we have proven the expression of the Laplacian in spherical coordinates (see page 1903):

$$\vec{\nabla}^2=\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right)+\frac{1}{r^2\sin(\theta)}\frac{\partial}{\partial\theta}\left(\sin(\theta)\frac{\partial\theta}{\partial\theta}\right)+\frac{1}{r^2\sin^2(\theta)}\frac{\partial^2}{\partial\phi^2} \tag{15.5.82}$$

Remember that we are looking for a description of spherical waves, waves that are spherically symmetrical (i.e. ones that do not depend on θ and ϕ) so that:

$$f(\vec{r})=f(r,\theta,\phi)=f(r) \tag{15.5.83}$$

The Laplacian then reduces simply to:

$$\vec{\nabla}^2=\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right) \tag{15.5.84}$$

The differential wave equation that is for recall (see section Wave Mechanics page 3501):

$$\vec{\nabla}^2 z=\frac{1}{v^2}\frac{\partial^2 z}{\partial t^2} \tag{15.5.85}$$

will be written:

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right)f=\frac{1}{v^2}\frac{\partial^2 f}{\partial t^2} \tag{15.5.86}$$

But:

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right)f=\frac{2}{r}\frac{\partial f}{\partial r}+\frac{\partial^2 f}{\partial r}=\frac{1}{r}\frac{\partial^2}{\partial r^2}(rf) \tag{15.5.87}$$

Therefore we get the differential wave equation in spherical coordinates:

$$\boxed{\frac{1}{r}\frac{\partial^2}{\partial r^2}(rf)=\frac{1}{v^2}\frac{\partial^2 f}{\partial t^2}} \tag{15.5.88}$$

Multiplying bot sides by r (independent of t):

$$\frac{\partial^2}{\partial r^2}(rf)=\frac{1}{v^2}\frac{\partial^2}{\partial t^2}(rf) \tag{15.5.89}$$

It is immediate that a special obvious solution is of the form:

$$\frac{f(kr-\omega t)}{r} \tag{15.5.90}$$

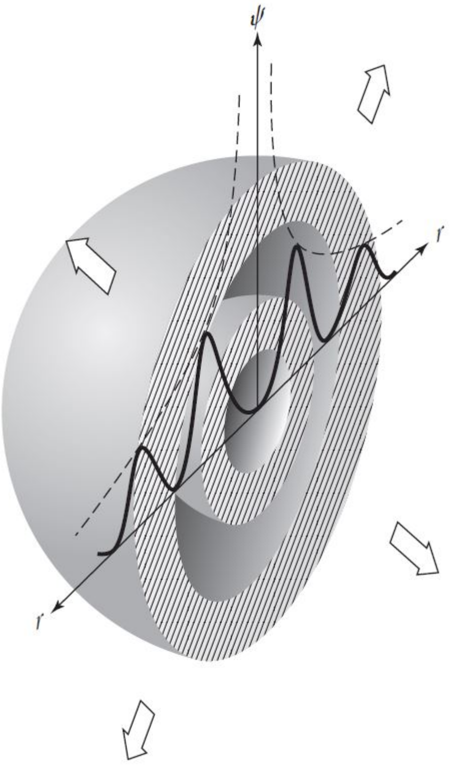
and it is common to choose the "harmonic spherical wave" form (don't forget that $kv=\omega$):

$$f(r,t)=\frac{A_0\cos(\omega t-kr)}{r}=\frac{A_0\cos(kr-\omega t)}{r} \tag{15.5.91}$$

or:

$$f(r,t)=\left(\frac{A_0}{r}\right)e^{\mathrm{i}(kr-\omega t)} \tag{15.5.92}$$

wherein the constant A_0 is the source strength. At any fixed value of time, this represents a cluster of concentric spheres filling all space. Notice that the amplitude of any spherical wave is a function of r, where the term r^{-1} serves as an attenuation factor. Unlike the plane wave, a spherical wave decreases in amplitude!



Obviously most wave sources are spherical but far away from the source the radius is such that the front wave con be considered as a plane!

15.5.3 Doppler effect

In what has been studied so far, the source S and the observer O of the wave have always been considered to be immobile relative to each other.

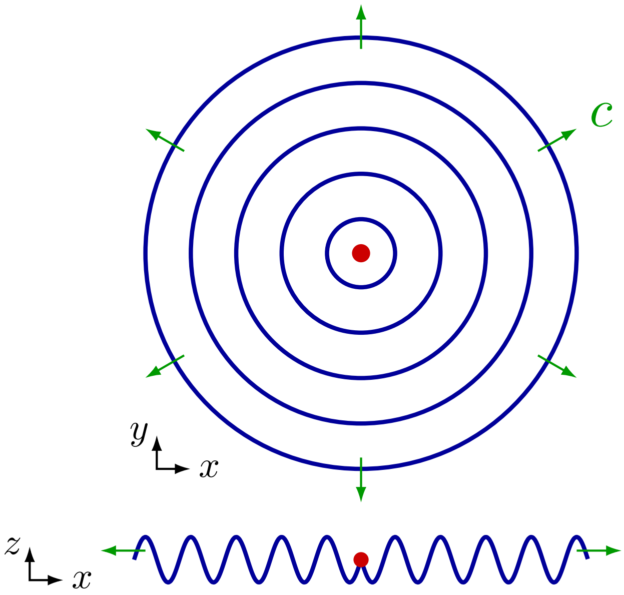


Figure 15.333: **Fixed oscillating punctual source (author: Izaak Neutelings)**

The Doppler effect explains the wave frequency (sound, electromagnetic, fluid) changes observed when S and O are in relative motion.

The Doppler effect (or the Doppler shift) is the change in frequency or wavelength of a wave (or other periodic event) for an observer moving relative to its source. It is named after the Austrian physicist Christian Doppler, who proposed it in 11842 (holocene calendar) in Prague. It is commonly heard when a vehicle sounding a siren or horn approaches, passes, and recedes from an observer. Compared to the emitted frequency, the received frequency is higher during the approach, identical at the instant of passing by, and lower during the recession.

When the source of the waves is moving towards the observer, each successive wave crest is emitted from a position closer to the observer than the previous wave. Therefore, each wave takes slightly less time to reach the observer than the previous wave. Hence, the time between the arrival of successive wave crests at the observer is reduced, causing an increase in the frequency. While they are travelling, the distance between successive wavefronts is reduced, so the waves "bunch together". Conversely, if the source of waves is moving away from the observer, each wave is emitted from a position farther from the observer than the previous wave, so the arrival time between successive waves is increased, reducing the frequency. The distance between successive wavefronts is then increased, so the waves "spread out". For waves that propagate in a medium, such as sound waves, the velocity of the observer and of the source are relative to the medium in which the waves are transmitted. The total Doppler effect may therefore result from motion of the source, motion of the observer, or motion of the medium. Each of these effects is analysed separately. For waves which do not require a medium, such as light or gravity in general relativity, only the relative difference in velocity between the observer and the source needs to be considered.

15.5.3.1 Fixed source-Moving observer

So consider the first case where the source is fixed and the observer in uniform rectilinear motion:

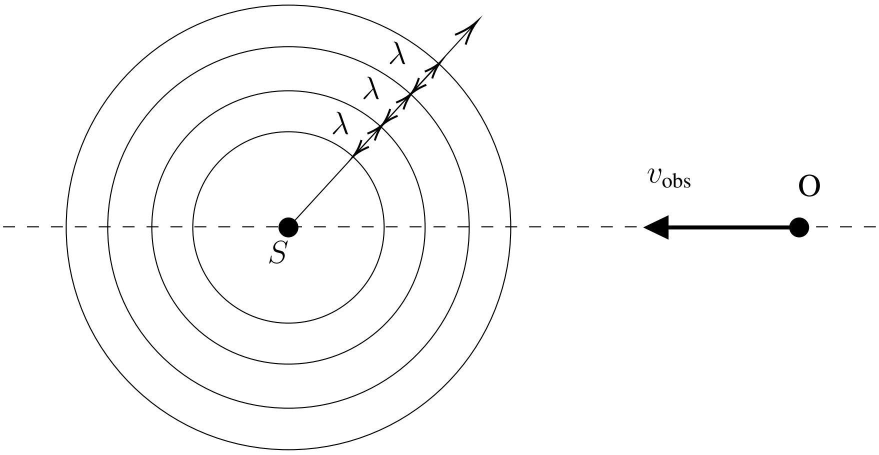


Figure 15.334: **Schematic diagram of a fixed source and an approaching observer (Doppler Effect)**

Given:

$$f=\frac{1}{T} \tag{15.5.93}$$

the frequency of the source S and the v_{phase} the propagation speed of the wave in the medium, assumed immobile.

If O does not move, he sees the wavefronts arrive at speed v_{phase} and there is not much to say. If O moves in the direction of S at the speed v_{obs}, it receives the wavefronts at the speed:

$$v_{\text{phase}}+v_{\text{obs}} \tag{15.5.94}$$

and crosses the distances λ in equal times:

$$\Delta t=\frac{\lambda}{v_{\text{phase}}+v_{\text{obs}}} \tag{15.5.95}$$

During each interval Δt perceives a full cycle of the wave. He concludes that Δt is the period T' of a wave of frequency f':

$$f'_{approach}=\frac{1}{T'}=\frac{v_{\text{phase}}+v_{\text{obs}}}{\lambda}=\frac{v_{\text{phase}}+v_{\text{obs}}}{v_{\text{phase}}}=\frac{1}{T}\left(1+\frac{v_{\text{obs}}}{v_{\text{phase}}}\right) \tag{15.5.96}$$

So finally:

$$\boxed{f'_{\text{approach}}=f\left(1+\frac{v_{\text{obs}}}{v_{\text{phase}}}\right)>f} \tag{15.5.97}$$

The sound becomes then more acute when approaching a source.

If O go away from S are speed v_{obs}, it receives the wavefronts at the speed:

$$v_{\text{phase}} - v_{\text{obs}} \tag{15.5.98}$$

and similar developments lead us then to:

$$\boxed{f'_{\text{remoteness}} = f\left(1 - \frac{v_{\text{obs}}}{v_{\text{phase}}}\right) \leq f} \tag{15.5.99}$$

The limit of validity of this relation is for $v_{\text{obs}} \leq v_{\text{phase}}$.

The sound becomes more low-pitched sound when we move far from the source.

Remark

If the speed of the observer is oriented in any manner, relatively to the line $\overline{\text{O}S}$, only the radial component is to be taken into account.

15.5.3.2 Moving source-Fixed observer

It is not necessarily intuitive to see that the system is not symmetrical. But taking a special case one quickly realizes that it is actually not. Indeed, if the receiver is moving from the source at a speed higher than v_{phase}, he will never receive wawefront, whereas if the source is moving away from the observer, that latter will still receive a wavefront. Can not reverse the role of the transmitter (source) and receiver (observer).

In the classical case, there is therefore an asymmetry in the frequency offset following that the source or observer is moving.

Let us consider the following situation:

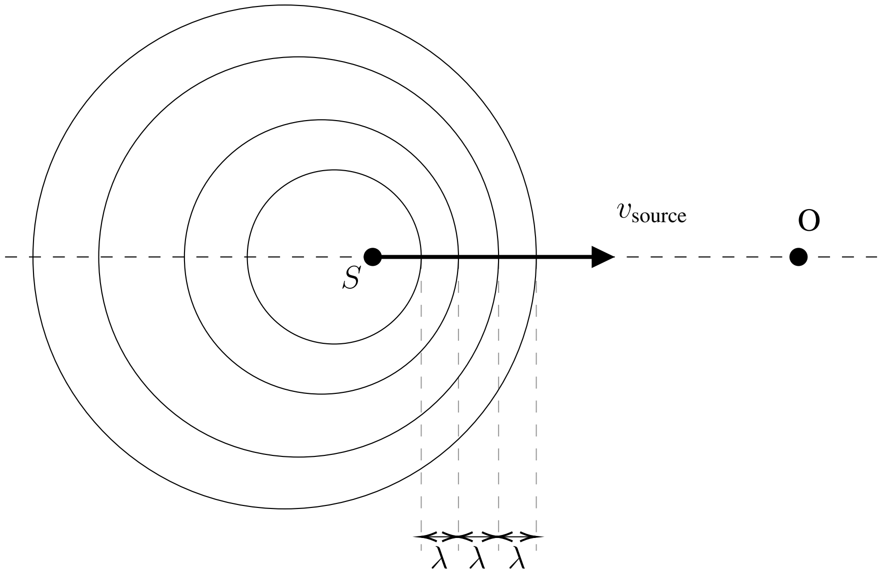
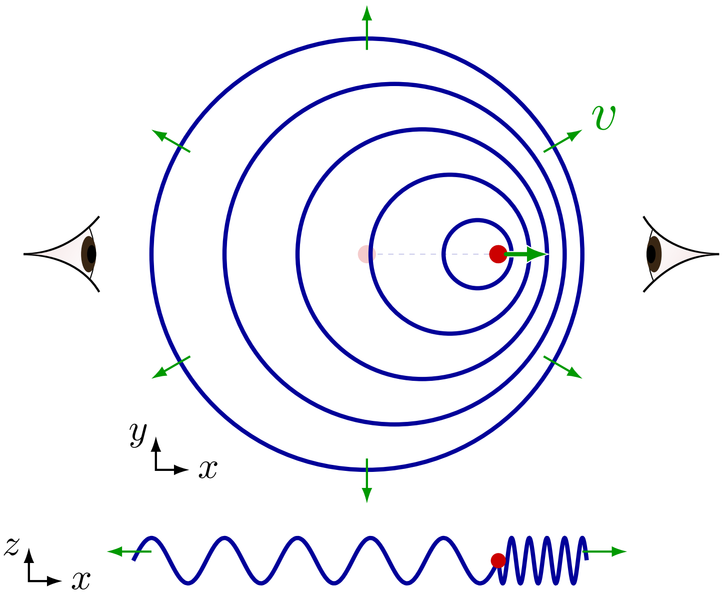


Figure 15.335: Schematic diagram of a fixed observer and an approaching source (Doppler Effect)

Or in a more detailed form highlighting also the wave aspect of the phenomenon:



Given:

$$f=\frac{1}{T} \tag{15.5.100}$$

the frequency of the source S in uniform rectilinear motion and the v_{phase} the speed of the wave propagation in the medium and O the assumed fixed observer.

If the source moves at speed v_{source}, it travels a distance:

$$d=v_{\text{source}}T \tag{15.5.101}$$

between each cycle. The wavelength seen by the observer thus varies of:

$$d=\Delta\lambda=v_{\text{source}}T \tag{15.5.102}$$

For the observer, it decreases such that:

$$\lambda'=\lambda-v_{\text{source}}T=\lambda-\frac{v_{\text{source}}}{f} \tag{15.5.103}$$

Then we have:

$$f'_{\text{approach}}=\frac{v_{\text{phase}}}{\lambda'}=\frac{v_{\text{phase}}}{\lambda-\dfrac{v_{\text{source}}}{f}}=\frac{v_{\text{phase}}}{\dfrac{v_{\text{phase}}}{f}-\dfrac{v_{\text{source}}}{f}}=f\left(\frac{v_{\text{phase}}}{v_{\text{phase}}-v_{\text{source}}}\right) \tag{15.5.104}$$

So after simplification:

$$\boxed{f'_{\text{approach}}=f\left(\frac{1}{1-\dfrac{v_{\text{source}}}{v_{\text{phase}}}}\right)>f} \tag{15.5.105}$$

The limit of validity of this relations being for $v_{\text{source}} \leq v_{\text{phase}}$.

Therefore it comes:

$$f'_{\text{approach}} > f \tag{15.5.106}$$

The sound becomes more acute when the source approaches (that same feeling as when it is the observer who approaches).

Similarly, when the source is moving away from the observer, we get immediately:

$$\boxed{f'_{\text{remoteness}} = f\left(\frac{1}{1+\dfrac{v_{\text{source}}}{v_{\text{phase}}}}\right) < f} \tag{15.5.107}$$

The sound becomes more low-pitched when the source is moving away (so same sensation as when it is the observer who moves away).

Therefore, the apparent frequency of sound when source is approaching a stationary listener with a given phase velocity is not the same as the apparent frequency of sound when the listener is approaching the stationary source with a given velocity. This shows that Doppler Effect in sound is asymmetric!!! The asymmetry comes from the medium. In the second case the source is at rest with respect to the medium and in the first case the receiver is at rest with respect to the medium.

Remark

In the case of electromagnetic waves, the wave velocity is the speed of light c which does not depend on the repository. Then we must address the problem within the framework of Special Relativity and then we expect to find a perfectly symmetrical effect since it can not possible to distinguish between speed the velocity of the transmitter or that of the receiver, only must be taking into account the relative speed between the both (see the developments done in the section of Special Relativity at page 5529).

15.5.3.3 Moving source and observer

If both (source and observer) are in rectilinear motion on a common straight line, the two effects are additive. We then have when they approach each other:

$$\left.\begin{aligned} f'_{\text{approach}} &= f\left(1+\frac{v_{\text{obs}}}{v_{\text{phase}}}\right) \\ f'_{\text{approach}} &= f\left(\frac{1}{1-\dfrac{v_{\text{source}}}{v_{\text{phase}}}}\right) \end{aligned}\right\} : f'_{\text{approach}} = f\left(1+\frac{v_{\text{obs}}}{v_{\text{phase}}}\right)\left(\frac{1}{1-\dfrac{v_{\text{source}}}{v_{\text{phase}}}}\right) \tag{15.5.108}$$

And when they move away:

$$\left.\begin{aligned} f'_{\text{remoteness}} &= f\left(1-\frac{v_{\text{obs}}}{v_{\text{phase}}}\right) \\ f'_{\text{remoteness}} &= f\left(\frac{1}{1+\dfrac{v_{\text{source}}}{v_{\text{phase}}}}\right) \end{aligned}\right\} : f'_{\text{remoteness}} = f\left(1-\frac{v_{\text{obs}}}{v_{\text{phase}}}\right)\left(\frac{1}{1+\dfrac{v_{\text{source}}}{v_{\text{phase}}}}\right) \tag{15.5.109}$$

15.5.4 Shock waves

Unlike ordinary sound waves, the speed of a shock wave varies with its amplitude. The speed of a shock wave is always greater than the speed of sound in the fluid and decreases as the amplitude of the wave decreases. When the shock wave speed equals the normal speed, the shock wave dies and is reduced to an ordinary sound wave.

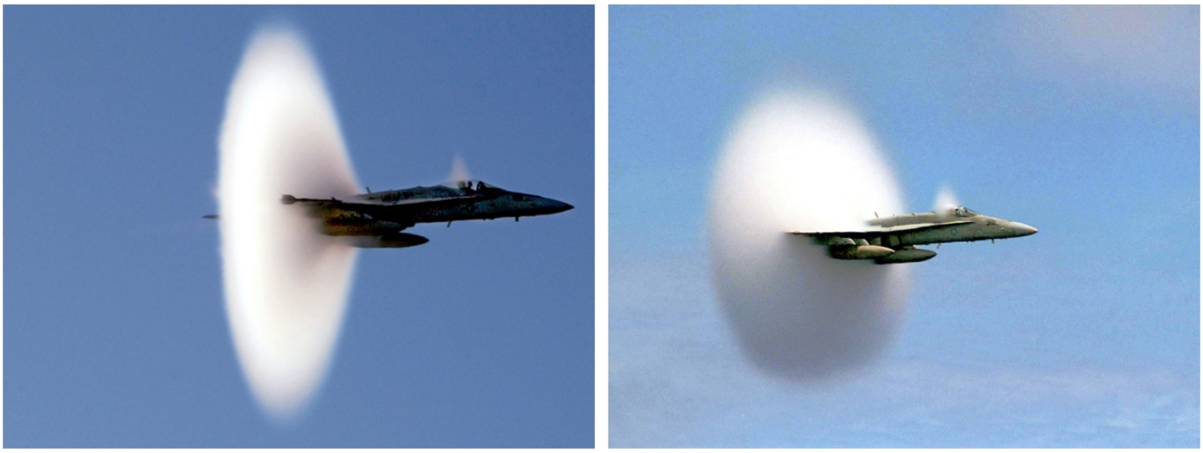


Figure 15.336: **Jet fighter planes with conical shock waves made visible by condensation**

The ratio of the speed v of a moving object to the speed of sound v_{sound} in a fluid is known as the "Mach number" in honour of Ernst Mach (11838–11916 according to holocene calendar):

$$\boxed{M = \frac{v}{v_{\text{sound}}} = \frac{v}{\sqrt{\dfrac{\gamma RT}{M_m}}}} \tag{15.5.110}$$

The Mach number is therefore a dimensionless measure of speed common in aerodynamics. Mach 0.5 is half the speed of sound, Mach 2 is twice the speed of sound, and so on. Speeds less than the speed of sound have a Mach number between zero and one and are described as "subsonic". Those greater than the speed of sound have Mach numbers greater than one are a described as "supersonic". Speeds approximately equal to the speed of sound have Mach numbers approximately equal to one and are described as "transonic".

The shock wave from a supersonic moving conical object is a cone composed of overlapping spherical wavefronts (if the moving object is cube at a far observation distance enough the shock wave will be also seen as a cone). As any one of these wavefronts forms, it propagates radially outward at speed v_{sound} and acquires a radius $v_{\text{sound}}t$ relatively to the contact peak point. At the same time the source, travelling at speed v moves forward vt. These two displacements form the leg and hypotenuse, respectively, of a right triangle and can be used to determine the Mach angle at the vertex of the shock cone (prolongation of the cone sides if the moving object is a cube...):

$$\sin(\theta) = \frac{v_{\text{sound}}}{v} \qquad (15.5.111)$$

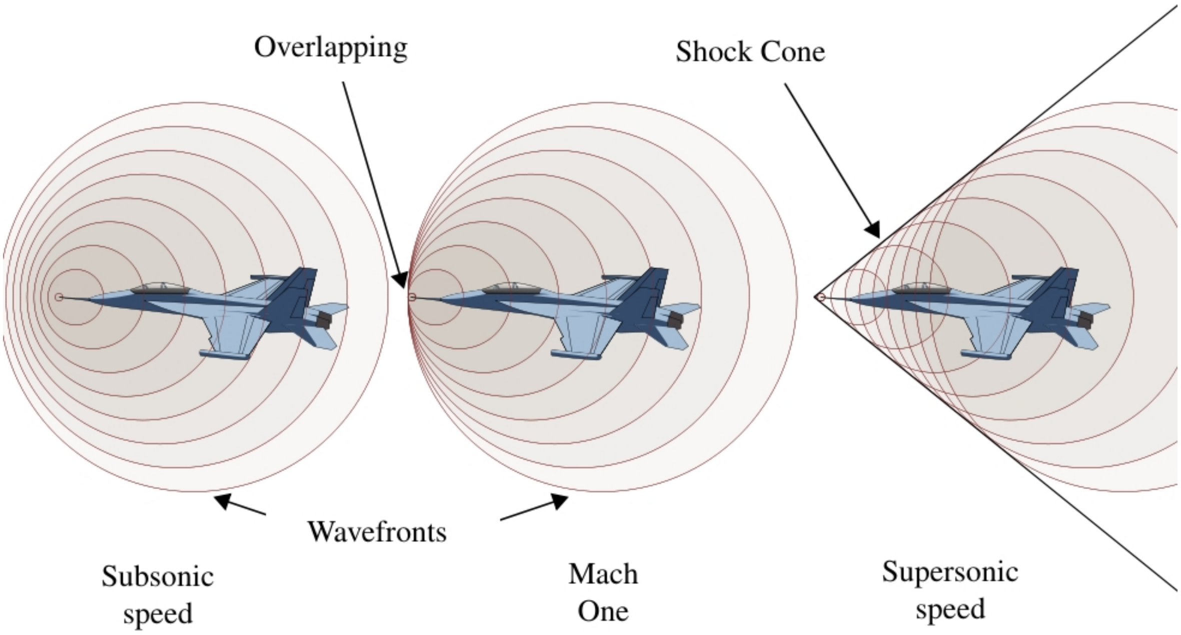


Figure 15.337: **Jet fighter with conical shock waves in the case of a conical "nose" (source: Wikipedia, author: Ignacio Icke)**

When an object travels slower than sound, the ratio in this equation is greater than one, and the equation does not have a real solution. This makes absolute sense as there is no shock wave to speak of at subsonic speeds. Travelling at the speed of sound makes the ratio equal one and results in a Mach angle of ninety degrees. At transonic speeds the shock wave is a wall of high pressure moving with the object, perpendicular to its velocity. Above the speed of sound, the ratio is less than one and the Mach angle is less than ninety degrees. The faster the object moves, the narrower the cone of high pressure behind it becomes. Measuring the vertex angle is thus a simple way to determine the speed of a supersonic object.

15.5.5 Music Scales

When we do vibrate a guitar string, violin string or other stringed instrument whose ends are fixed (stationary wave), we have proved in the section of Wave Mechanics that we are in conditions name "Dirichlet boundaries".

We obtained after many calculations, the following result:

$$\omega_n = \sqrt{\frac{F}{\rho_l}}\left(\frac{n\pi}{L}\right) \qquad (15.5.112)$$

Therefore:

$$f_n = \frac{n}{2L}\sqrt{\frac{F}{\rho_l}} \propto \frac{1}{L} \qquad (15.5.113)$$

thus the frequency (which is one of the factors of the feel of the "height" of a sound by humans) is inversely proportional to the length L of a string attached to the ends (as it would be deducted by Pythagoras long time before that theory of string was established).

The previous relationship highlights that: more the vibrant part is long (more n is small) the more the sound is low (low frequency) and the vibrant part is short (large n), the more the sound will be acute (high frequency).

Let us recall that for $n=1$, we speak of "fundamental frequency" and for $n\geq 2$ of "harmonic frequency." Thus, $n=2$ is the first harmonic, $n=3$ the second harmonic, etc.

Definitions (#15.5.722):

D1. We say that two frequencies (or two periodic waves) are "in harmony" with each other if the ratio of their frequencies is equal to unity:

$$\text{Harmony: } \frac{f''}{f'}=1 \qquad (15.5.114)$$

D2. We say that two frequencies (or two periodic waves) are "at the octave" to each other if the ratio of their frequencies is equal to two:

$$\text{Octave: } \frac{f''}{f'}=2 \qquad (15.5.115)$$

Let us notice that the first harmonic ($n=1$) corresponds to the octave of the fundamental f_0:

$$\frac{f_1}{f_0}=2 \qquad (15.5.116)$$

For example, some singers are capable of singing on 8 octaves (that is to say a frequency ratio of 16). Thus, in the case of a stringed instrument with Dirichlet conditions this is equivalent obviously to pluck the string in the middle (for guitars that match to press the rope on the fret of the 12th box with respect to the frequency of an open string to get one octave).

The octave ratio sounds pretty good for the majority of humans.

D3. We say that two frequencies (or two periodic waves) are "at the quint" to each other if the ratio of their frequencies is equal to $3/2$:

$$\text{Quint: } \frac{f''}{f'} = \frac{3}{2} \tag{15.5.117}$$

Except the octave, this interval is the simplest of all, and since always, in occidental music, it was considered as the consonant interval with excellence, so the more remarkable sound for ear. For this reason, the quint has played a key role in establishing music scales, the Pythagorean scale being even built exclusively on this particular interval.

We notice that we can put at most an integer of two quints in one octave (beyond we exceed the value of an octave).

When a stringed instrument has its empty vibrations that are in a ratio of frequencies corresponding to the quint, we say that the instrument is "tuned in quint to quint".

D4. We say that two frequencies (or two periodic waves) are "at the quart" to each other if the ratio of their frequencies is equal to $4/3$:

$$\text{Quart: } \frac{f''}{f'} = \frac{4}{3} \tag{15.5.118}$$

D5. The ratio of the frequencies of two sounds (or two periodic waves) distant from each other by a "tone" may be $9/8$. This is named the "major tone":

$$\text{Major tone: } \frac{f''}{f'} = \frac{9}{8} \tag{15.5.119}$$

D6. The "minor ton" has a ratio of frequencies of $10/9$:

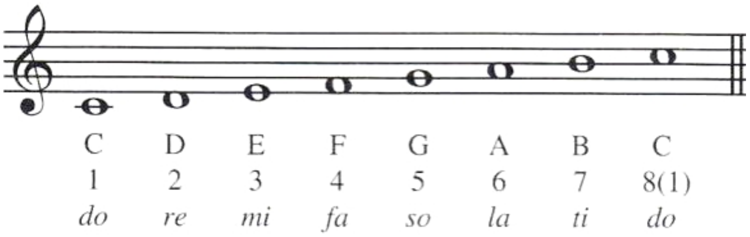
$$\text{Minor ton: } \frac{f''}{f'} = \frac{10}{9} \tag{15.5.120}$$

Remark

The octave, the quint, the quart and the tones are obviously named "interval ratio" in music. Ratios, rather than direct frequency measurements, allow musicians to work with relative pitch measurements applicable to many instruments in an intuitive manner, whereas one rarely has the frequencies of fixed pitched instruments memorized and rarely has the capabilities to measure the changes of adjustable pitch instruments.

Obviously, if we want to make music, we have to choose a mark (reference). As the frequency range is infinite, there was chosen a frequency at which the current human ear is the most sensitive: 440 [Hz] that is named the "La" and is denoted by (A) by the Anglo Saxon.

In theory, the occidental students learn the empirical sequences of 7 notes and whose writing origin is purely mnemonic (we find this result on electronic tuners with the Anglo-Saxon system using certain letters of the alphabet):



where the horizontal lines of the symbolic solfegic above represent a quinte in the Pythagorean scale.

Between the first Do and the second Do, we have then a range of one octave. Moreover, the number of notes within the octave interval is often insufficient to make good music and artists have added other notes by altering the frequencies of these 7 notes, by "diesing" them (multiplication of the frequency by $25/24$) or by "bemolising" (multiplication by $1-25/24$).

Note that there are a dozen of ways to cut the solfegic octave and are almost all are still used in the early 121st century according to holocene calendar (they are all completely empirical so we will not go into details):

- The "natural scale" (also named "harmonic range")
- The "tempered scale"
- The "Pythagorean scale"
- The "chromatic scale"
- The "Zarlino scale" (also named "physicists scale")
- etc.

The method and the aforementioned learning sequence is a mathematical aberration (but as it is music... it's an empirical art so deeply linked with non-sense).

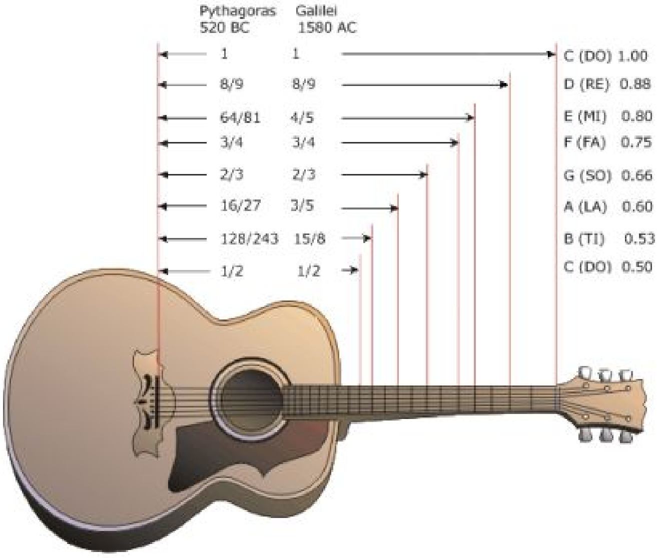

Pythagoras
520 BC
Galilei
1580 AC
1
1
C (DO) 1.00
8/9
8/9
D (RE) 0.88
64/81
4/5
E (MI) 0.80
3/4
3/4
F (FA) 0.75
2/3
2/3
G (SO) 0.66
16/27
3/5
A (LA) 0.60
128/243
15/8
B (TI) 0.53
1/2
1/2
C (DO) 0.50

Figure 15.338: Acoustic guitar rules (source: OpenStax)

15.5.6 Harmonic Oscillator

In Classical Mechanics, a "harmonic oscillator", also named "Lorentz oscillator" is a system that, when displaced from its equilibrium position, experiences a restoring force proportional to the displacement. This subject is of main importance in various physics fields (including quantum physics or molecular chemistry)!

Let us consider the mechanical system of the figure below consisting of a spring of constant k of a mass m and a viscous damping constant c. This system is often equated with the mechanics of a speaker (or a microphone) where the air acts as a viscous damper and the magnetic field of the exciter magnet as source of the force or also to a car damper:

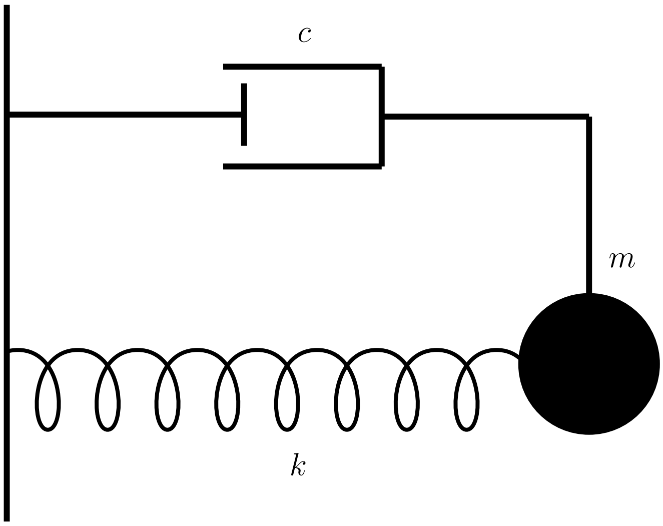


Figure 15.339: **Schematic diagram of the harmonic oscillator**

That's why the configuration above is also commonly know under the name of "harmonic damper",

This model is also used widely in molecular chemistry and solid physics to study crystalline structure on to model the bonds between atoms.

We have proved in the section of Mechanical Engineering (see page 8408) that for an helical spring we had:

$$F=\left(\frac{Gd^4}{8nD^3}\right)x=kx \tag{15.5.121}$$

and this is the expression of the force to which the spring will resist to an external force of compression or tension. At equilibrium, we necessarily the sum of the forces that is zero:

$$\vec{F}_{\text{ext}}+\vec{F}_{\text{spring}}=\vec{0} \tag{15.5.122}$$

Therefore:

$$m\ddot{x}+kx=0 \tag{15.5.123}$$

We have proved in the section of Continuum Mechanics that Stokes' law (see page 3853) in the case of a cylinder was given by:

$$F = 6\pi\mu R v = cv \tag{15.5.124}$$

We will assume that our viscous damper is based on a similar law where the resistance force is proportional to the speed. At equilibrium, we necessarily have the sum of the forces is still zero:

$$\vec{F}_{\text{ext}} + \vec{F}_{\text{spring}} + \vec{F}_{\text{amortization}} = \vec{0} \tag{15.5.125}$$

Therefore (we already know this differential equation):

$$\boxed{m\ddot{x} + c\dot{x} + kx = 0} \tag{15.5.126}$$

We will simplify this expression:

$$\frac{m}{k}\ddot{x} + \frac{c}{k}\dot{x} + x = 0 \tag{15.5.127}$$

We are left with exactly the same differential equation to that of a series RLC circuit in steady state (see section Electrical Engineering page 8475):

$$LC\ddot{q} + RC\dot{q} + q = 0 \tag{15.5.128}$$

So we have in the same way, the critical viscosity:

$$c_c = 2\sqrt{mk} \tag{15.5.129}$$

instead of the critical resistance of the RLC circuit:

$$R_c = 2\sqrt{\frac{L}{C}} \tag{15.5.130}$$

which will determine if the system is in critical regime:

$$c = c_c \tag{15.5.131}$$

in aperiodic regime (supercritical)

$$c > c_c \tag{15.5.132}$$

or in pseudo-periodic regime (damped oscillations) with:

$$c < c_c \tag{15.5.133}$$

In the case of cars or sound speakers, it is often useful to arrange the system to be in pseudo-periodic regime. Then we find exactly the same results as the RLC circuit with damped oscillations since the differential equation is the same.

Specialists from acoustic or mechanics then arrange the system to get the concept of "damping factor" of the RLC circuit:

$$\alpha = \frac{R}{2L} \tag{15.5.134}$$

which will be written in the present case by identification of the terms:

$$\lambda=\frac{c}{2m} \tag{15.5.135}$$

With the "natural pulsation":

$$\omega=\sqrt{\frac{1}{LC}} \tag{15.5.136}$$

which will be written in the present case:

$$\boxed{\omega=\sqrt{\frac{k}{m}}} \tag{15.5.137}$$

The damped oscillator loses energy over time. Let us calculate that loss! We have (see section of Classical Mechanics page 3129 and Mechanical Engineering page 3418):

$$E_{\text{tot}}=E_c+E_p=\frac{1}{2}m\dot{x}^2+\frac{1}{2}kx^2 \tag{15.5.138}$$

Therefore:

$$\frac{\mathrm{d}E_{\text{tot}}}{\mathrm{d}t}=m\dot{x}\ddot{x}+kx\dot{x}=(m\ddot{x}+kx)\dot{x} \tag{15.5.139}$$

Now, according to the equation of motion:

$$m\ddot{x}+c\dot{x}+kx=0\Rightarrow m\ddot{x}+kx=-c\dot{x} \tag{15.5.140}$$

It comes then:

$$\frac{\mathrm{d}E_{\text{tot}}}{\mathrm{d}t}=-c\dot{x}^2 \tag{15.5.141}$$

The loss of energy (or power loss) is proportional to the square of the instantaneous velocity!

We have shown in the section of Electrical Engineering for the RLC circuit in regime of underdamped oscillations (see page 8480) that we had:

$$\begin{aligned}q&=\frac{q_0}{\cos(\phi)}e^{-\frac{R}{2L}t}\cos(\omega t-\phi)\\ i&=-\frac{\omega q_0}{\cos^2(\phi)}e^{-\frac{R}{2L}t}\sin(\omega t)\end{aligned} \tag{15.5.142}$$

What becomes here:

$$\begin{aligned}x&=\frac{x_0}{\cos(\phi)}e^{-\frac{c}{2m}t}\cos(\omega t-\phi)\\ \dot{x}&=-\frac{\omega x_0}{\cos^2(\phi)}e^{-\frac{c}{2m}t}\sin(\omega t)\end{aligned} \tag{15.5.143}$$

Therefore:

$$\frac{\mathrm{d}E_{\text{tot}}}{\mathrm{d}t}=-c\dot{x}^2=c\left(\frac{\omega x_0}{\cos^2(\phi)}\sin(\omega t)\right)^2e^{-\frac{c}{m}t}=c\xi^2\sin^2(\omega t)e^{-\frac{c}{m}t} \tag{15.5.144}$$

15.5.6.1 Damping and Resonance

Let us still consider the same oscillator but this time excited by a harmonic force (thus: forced regime) as:

$$m\ddot{x}+c\dot{x}+kx=F_0\sin(\omega t) \tag{15.5.145}$$

Again, this is exactly the same type of differential equation as for the RLC oscillator in forced regime studied in the section of Electrical Engineering (see page 8497) for which we had:

$$L\frac{\mathrm{d}^2q}{\mathrm{d}t^2}+\frac{\mathrm{d}q}{\mathrm{d}t}+\frac{1}{C}q=U_0\sin(\omega t) \tag{15.5.146}$$

It follows the same results and plots and conclusions that the reader may find in the section of Electrical Engineering.

But as a reader noticed us, the version of the development of the Electrical Engineering section is too much "electrical engineer" oriented for a good comprehension. Therefore for the people that want it, here is the traditional and more understandable version...

So let's go! We assume that the damping force is proportional to the velocity of the mass. If there are no external forces we obtain the differential equation:

$$m\frac{\mathrm{d}^2x}{\mathrm{d}t^2}+b\frac{\mathrm{d}x}{\mathrm{d}t}+kx=0 \tag{15.5.147}$$

where $x=x(t)$ is the displacement from equilibrium at time t of the mass of $m>0$ units, $b\geq 0$ is the damping constant and $k>0$ is the spring constant.

The characteristic polynomial:

$$r^2+\frac{b}{m}r+\frac{k}{m} \tag{15.5.148}$$

has discriminant:

$$\Delta=\frac{b^2-4km}{m^2} \tag{15.5.149}$$

If $b^2<4km$ we have $\Delta<0$ and the characteristic polynomial factorizes in the form:

$$\left(r+\frac{b}{2m}\right)^2+\omega^2 \tag{15.5.150}$$

with:

$$\omega=\frac{\sqrt{4km-b^2}}{2m}=\sqrt{\frac{k}{m}-\left(\frac{b}{2m}\right)^2} \tag{15.5.151}$$

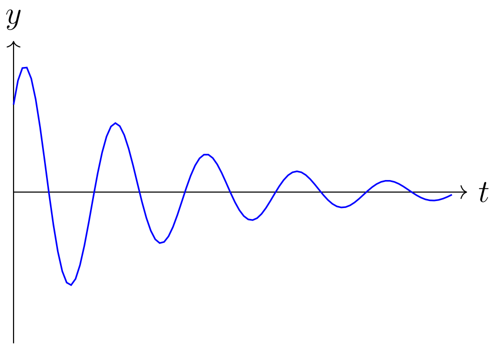
In this case the characteristic polynomial has complex roots:

$$-\frac{b}{2m}\pm\mathrm{i}\omega \tag{15.5.152}$$

and the general solution of the differential equation is:

$$y = e^{-bt/2m}\left(A\cos(\omega t) + B\sin(\omega t)\right) = Ce^{-bt/2m}\sin(\omega t + \theta) \qquad (15.5.153)$$

where $C = \sqrt{A^2+B^2}$ and $0 \leq \theta \leq 2\pi$ defined by $\cos(\theta) = A/C$, $\sin(\theta) = B/C$. The angle θ is named the "phase". In this case we see that the mass oscillates with frequency $\omega/2\pi$ and decreasing amplitude (underdamping). If $b = 0$ there is no damping and the mass oscillates with frequency $\omega/2\pi$ and constant amplitude; such motion is named "simple harmonic".

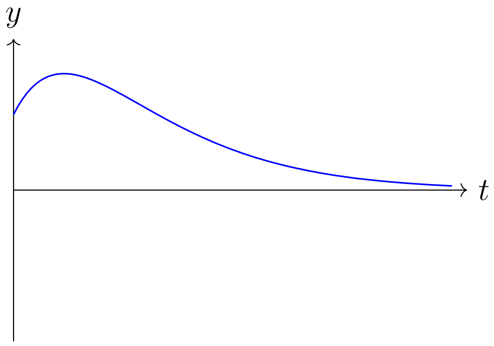


If $b^2 \geq 4km$ we have $\Delta \geq 0$ and so the characteristic polynomial has real roots:

$$r_1 = -\frac{b}{2m} + \frac{\sqrt{b^2-4km}}{2m} \qquad r_2 = -\frac{b}{2m} - \frac{\sqrt{b^2-4km}}{2m} \qquad (15.5.154)$$

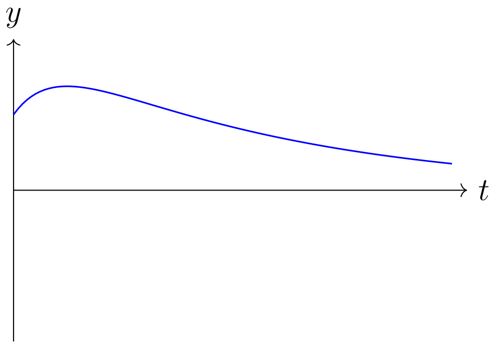
which are both negative. If $r_1 = r_2 = r$ the general solution of our differential equation is:

$$y = Ae^{rt} + Bte^{rt} \qquad (15.5.155)$$



and if $r_1 \neq r_2$ the general solution is:

$$y = Ae^{r_1 t} + Be^{r_2 t} \qquad (15.5.156)$$



In both cases $y \to 0$ as $t \to +\infty$. In the second case we have what is named "over damping" and in the first case the over damping is said to be "critical" (i.e. critical damping). In each the mass returns to its equilibrium position without oscillations.

Suppose now that our mass-spring system is subject to an external force so that our differential equation now becomes:

$$\boxed{m\frac{\mathrm{d}^2x}{\mathrm{d}t^2}+b\frac{\mathrm{d}x}{\mathrm{d}t}+kx=F(t)} \tag{15.5.157}$$

Such as system is named a "forced oscillating system" or "driven oscillating system with damping, steady state and resonance solution". The function $F(t)$ is named the "forcing function" and measures the magnitude and direction of the external force. We consider the important special case where the forcing function is harmonic:

$$F(t)=F_0\cos(\gamma t), \qquad F_0>0 \text{ a constant} \tag{15.5.158}$$

We also assume that we have under-damping with damping constant $b>0$. In this case the differential equation has a particular solution of the form:

$$y_p=A_1\cos(\gamma t)+A_2\sin(\gamma t) \tag{15.5.159}$$

Substituting in differential equation and simplifying, we get:

$$((k-m\gamma^2)A_1+b\gamma A_2)\cos(\gamma t)+(-b\gamma A_1+(k-m\gamma^2)A_2)\sin(\gamma t)=F_0\cos(\gamma t) \tag{15.5.160}$$

Setting the corresponding coefficients on both sides equal, we get:

$$\begin{aligned}(k-m\gamma^2)A_1+b\gamma A_2&=F_0\\ -b\gamma A_1+(k-m\gamma^2)A_2&=0\end{aligned} \tag{15.5.161}$$

Solving for A_1,A_2 we get:

$$A_1=\frac{F_0(k-m\gamma^2)}{(k-m\gamma^2)^2+b^2\gamma^2} \qquad A_2=\frac{F_0b\gamma}{(k-m\gamma^2)^2+b^2\gamma^2} \tag{15.5.162}$$

and:

$$\begin{aligned}y_p&=\frac{F_0}{(k-m\gamma^2)^2+b^2\gamma^2}((k-m\gamma^2)\cos(\gamma t)+b\gamma\sin(\gamma t))\\ &=\frac{F_0}{\sqrt{(k-m\gamma^2)^2+b^2\gamma^2}}\sin(\gamma t+\phi)\end{aligned} \tag{15.5.163}$$

The general solution of our differential equation is then:

$$y=Ce^{-bt/2m}\sin(\omega t+\theta)+\frac{F_0}{\sqrt{(k-m\gamma^2)^2+b^2\gamma^2}}\sin(\gamma t+\phi) \tag{15.5.164}$$

Because of damping the first term tends to zero and is named the "transient part" of the solution. The second term, the "steady-state part" of the solution, is due to the presence of the forcing function $F_0\cos(\gamma t)$. It is harmonic with the same frequency $\gamma/2\pi$ but is out of phase with it by an angle $\phi-\pi/2$. The ratio of the magnitudes:

$$M(\gamma)=\frac{1}{\sqrt{(k-m\gamma^2)^2+b^2\gamma^2}} \tag{15.5.165}$$

is named the "gain factor". The graph of the function $M(\gamma)$ is named the "resonance curve". It has a maximum of:

$$\frac{1}{b\sqrt{\dfrac{k}{m}-\dfrac{b^2}{4m^2}}} \tag{15.5.166}$$

when:

$$\gamma=\gamma_r=\sqrt{\frac{k}{m}-\frac{b^2}{2m^2}} \tag{15.5.167}$$

The frequency $\gamma_r/2\pi$ is named the "practical resonance frequency" of the system. When $\gamma=\gamma_r$ the system is said to be in resonance with the external force. Note that $M(\gamma_r)$ gets arbitrarily large as $b\rightarrow 0$. That is, when:

$$\gamma=\gamma_r=\sqrt{\frac{k}{m}} \tag{15.5.168}$$

We thus see that the system is subject to large oscillations if the damping constant is very small and the forcing function has a frequency near the resonance frequency of the system.

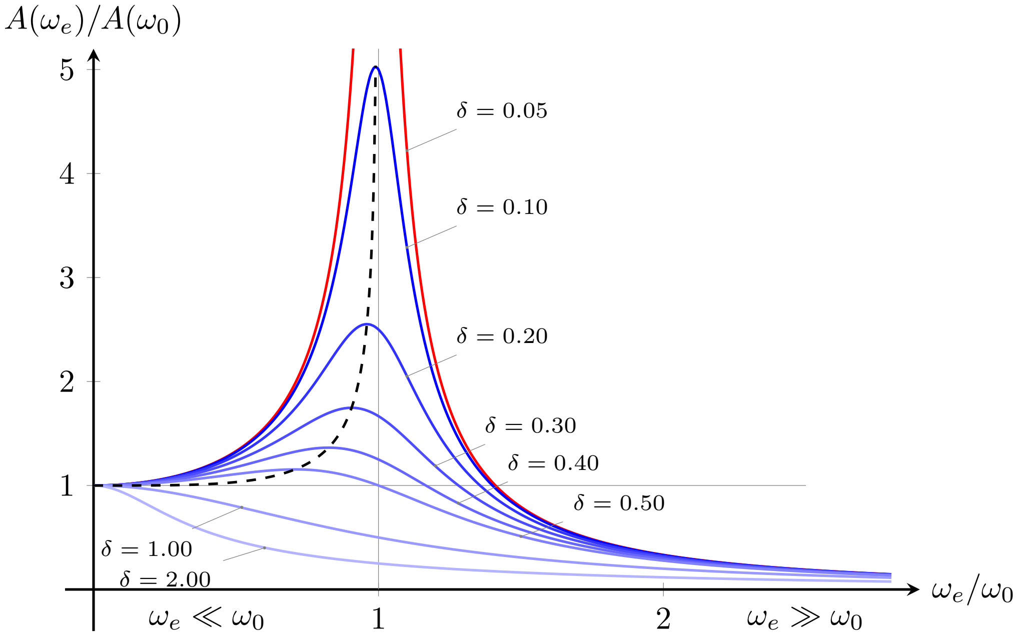


Figure 15.340: Resonances curves with traditional notations

The general solution becomes when $b\rightarrow 0$:

$$y=C\sin(\omega t+\theta)+\frac{F_0}{(k-m\lambda^2)^2}\sin(\gamma t+\phi) \tag{15.5.169}$$

In many physics textbooks the study of resonance the differential equation used is given by:

$$\boxed{\frac{\mathrm{d}^2x}{\mathrm{d}t^2}+\omega_0^2x=F_0\cos(\omega t)} \tag{15.5.170}$$

That is with $m=1$ and $k=\omega_0^2$ then the general solution is written:

$$\boxed{y=C\sin(\omega t+\theta)+\frac{F_0}{(\omega_0^2-\omega^2)^2}\sin(\gamma t+\phi)} \tag{15.5.171}$$

"Pure resonance" occurs exactly when the natural internal frequency ω_0 matches the natural external frequency ω in which case all solutions of the differential equation are unbounded.

The difference between the "practical resonance frequency" and "pure resonance frequency" is just that the first one is the realistic one as they don't exist any physical system with perfectly zero b (hence without damping!). This is why when $b>0$ we speak of "practical resonance frequency" and when $b=0$ of "pure resonance frequency".

As already mentioned, and for recall the above applies to a simple RLC electrical circuit where the differential equation for the current i is:

$$L\frac{\mathrm{d}^2i}{\mathrm{d}t^2}+R\frac{\mathrm{d}i}{\mathrm{d}t}+\frac{i}{C}=F(t) \tag{15.5.172}$$

where L is the inductance, R is the resistance and C is the capacitance. The resonance phenomenon is the reason why we can send and receive and amplify radio transmissions sent simultaneously but with different frequencies.

Some famous resonance situations are:

- The Tacoma Narrows bridge that collapsed in 11940 (holocene):

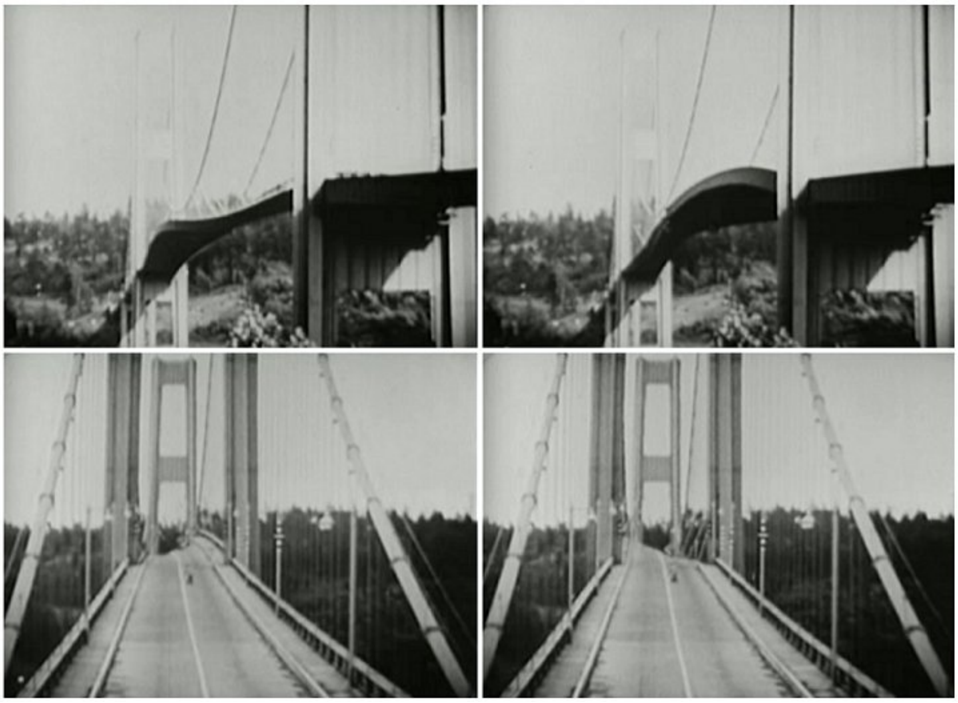


Figure 15.341: **Tacoma Narrows bridge in resonance before collapsing (source: ?)**

- The break up of a crystal glass (better to choose such a material) when under resonance (also named the "wine glass experiment"):

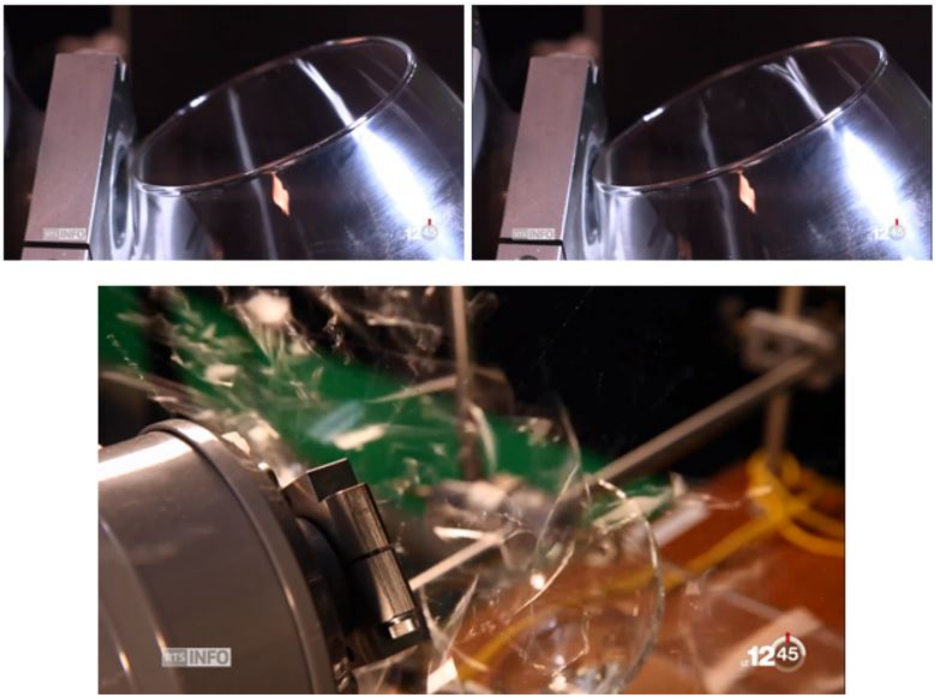
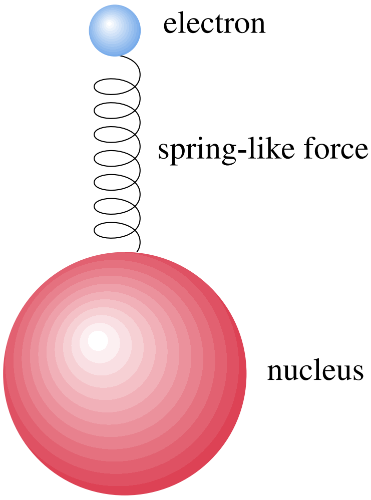


Figure 15.342: **Crystal glass experiment (source: RTS 12h45, 12018-05-13, CQFD)**

In the first two images above we can see that the diameter of the glass change (it oscillates in fact on the video...) under the forcing function and finally the glass breaks up.

Let us notice that if we assume the nucleus of the atom is much more massive than the electron, we can treat the problem as if the electron-spring system is connected to an infinite mass, which does not move, allowing us to use the mass of the electron. Depending on the case, this value may be substituted by the reduced or effective electron mass to account for deviations from this assumption.



Moreover, the assumption that the binding force behaves like a spring is a justified approximation for any kind of binding, given that the displacement is small enough (meaning that only the constant and linear terms in the Taylor expansion are relevant).

The damping term comes from internal collisions in the solid and radiation emitted by the electron (any accelerating charge emits radiation).

~20 %

★★★☆☆

13 votes, 61.54%

16

Engineering

Engineering is the set of practices consisting to apply the results of the exact sciences and basic research to practical industrial or daily problems. (Larousse)

Contents

16.1 Marine & Weather Engineering . . . 8239
- 16.1.1 Visual horizon . . . 8244
  - 16.1.1.1 Discerning the curvature of the Earth . . . 8257
- 16.1.2 Wind direction . . . 8261
- 16.1.3 Atmospheric Profile Models . . . 8264
  - 16.1.3.1 Atmospheric Exponential Profile Model . . . 8264
  - 16.1.3.2 Adiabatic Atmosphere Model . . . 8268
- 16.1.4 Earth's equilibrium temperature . . . 8272
  - 16.1.4.1 Greenhouse effect . . . 8275
  - 16.1.4.2 Milankovitch cycles . . . 8276
- 16.1.5 Temperature variation range models . . . 8278
  - 16.1.5.1 Diurnal temperature range (DTR) . . . 8278
  - 16.1.5.2 Seasonal temperature range (STR) . . . 8281
- 16.1.6 Weather (sounding) balloon . . . 8288
- 16.1.7 Cyclogenesis and Anticyclogenesis . . . 8291
- 16.1.8 Tides . . . 8301
  - 16.1.8.1 Astronomical tides (influence Moon/Sun) . . . 8303
- 16.1.9 Lorenz equations . . . 8318
  - 16.1.9.1 Rayleigh-Bénard convection cells (Benard-Marangoni instability) . . . 8328
  - 16.1.9.2 Lorenz attractor and chaos . . . 8337
- 16.1.10 Waves . . . 8344
  - 16.1.10.1 Depth of a wave . . . 8356

16.1.10.2 Wave's amplitude . . . 8357

16.2 Mechanical Engineering . . . 8359

16.2.1 Mechanical Efficiency . . . 8360

16.2.2 Gears . . . 8362

16.2.2.1 Transmission ratios . . . 8365

16.2.2.2 Gears association . . . 8368

16.2.2.3 Types of Gears . . . 8374

16.2.3 Strength of materials . . . 8382

16.2.3.1 Quadratic moments . . . 8385

16.2.3.2 Equation of the elastic line . . . 8388

16.2.3.3 Torsion . . . 8404

16.2.3.4 Buckling . . . 8408

16.2.3.5 Traction . . . 8425

16.2.3.6 Pressure Vessels . . . 8426

16.2.3.7 Centrifugal expansion of a rotating ring . . . 8434

16.3 Electrical Engineering . . . 8437

16.3.1 Elementary primitive electrical symbols . . . 8441

16.3.2 Alternative current VS Direct current . . . 8444

16.3.2.1 DC generator (dynamo) . . . 8447

16.3.2.2 Average power . . . 8451

16.3.3 Transformers . . . 8455

16.3.3.1 Transformer universal EMF equation . . . 8461

16.3.4 Steady State Linear Circuits . . . 8463

16.3.4.1 RC series circuit . . . 8467

16.3.4.2 RL series circuit . . . 8472

16.3.4.3 Source-Free serial RLC circuit . . . 8475

16.3.4.4 Telegrapher's Equations . . . 8486

16.3.5 Linear Circuit (RLC) in Forced Regime . . . 8497

16.3.5.1 Low-pass filter . . . 8501

16.3.5.2 High-pass filter . . . 8502

16.3.5.3 Integrator and differentiator . . . 8504

16.3.6 Signal Processing . . . 8506

16.3.6.1 Nyquist-Shannon Theorem (Nyquist sampling theorem) . . . 8506

16.3.7 Semiconductor devices . . . 8519

16.3.7.1 Diodes . . . 8519

16.3.7.2 Amplifiers . . . 8536

16.4 Civil Engineering . . . 8539

16.4.1 Static 8541
16.4.2 Pulleys 8542
16.4.2.1 Windlass 8551
16.4.3 Cornu spiral 8553
16.4.4 Overhead cable 8558
16.4.4.1 Free overhead cable (catenary) 8559
16.4.4.2 Charges overhead cable (suspended bridge) 8567
16.4.4.3 Very tense cable 8571
16.4.5 Falling chimney (naive approach) 8573
16.4.6 Dams 8579
16.4.6.1 Power extracted from a Dam 8582
16.4.7 Tuned Mass Damper 8587
16.5 Aerospace Engineering 8590
16.5.1 Airfoil Lift 8591
16.5.1.1 Newton's lift argument (skipping stone argument) 8593
16.5.1.2 Bernoulli's lift argument (equal time argument) 8594
16.5.1.3 Euler's lift argument 8596
16.5.1.4 Coandă lift argument 8600
16.5.1.5 Kutta-Joukowski lift argument 8601
16.5.2 Cosmic speeds 8603
16.5.3 Fundamental Equation of Propulsion (Tsiolkovsky rocket equation) 8606
16.5.4 Geostationary orbit 8615
16.5.5 Vis-Viva equation 8619
16.5.6 Gravity assist 8622
16.5.6.1 Speed variation (in solar reference frame) 8623
16.5.6.2 Approaching speed from infinity 8626
16.5.6.3 Velocity variation and impact factor 8627
16.5.7 Hohmann Transfer orbit 8633
16.6 Software Engineering 8637
16.6.1 Algorithm 8637
16.6.2 Dichotomic Search algorithm 8639
16.6.2.1 Bisection algorithm 8639
16.6.2.2 Binary search algorithm 8642
16.6.3 Tower of Hanoi algorithm 8644
16.6.4 Sorting algorithms 8650
16.6.4.1 Bubble sort 8650
16.6.4.2 QuickSort algorithm 8652

16.6.5 Dijkstra’s algorithm 8654

16.6.6 Google PageRank algorithm 8662

16.6.6.1 Weighted Count 8664

16.6.6.2 Recursive counting 8665

16.6.6.3 Absorbing states 8669

16.7 Industrial Engineering 8673

16.7.1 Six Sigma 8674

16.7.1.1 Quality Control 8677

16.7.1.2 Defaults/Errors 8678

16.7.1.3 Capability Indices 8683

16.7.1.4 Quality Levels 8697

16.7.2 Taguchi Model 8710

16.7.3 Preventive Maintenance 8716

16.7.3.1 Planned Obsolescence 8717

16.7.3.2 Reliability Empirical Estimators 8720

16.7.3.3 Weibull Distribution 8736

16.7.3.4 Topology of Systems 8743

16.7.3.5 Maximum Likelihood for failure rate determination of samples 8761

16.7.3.6 Kaplan-Meier Survival Rate 8763

16.7.3.7 ABC Method 8772

16.7.3.8 XYZ Method 8777

16.7.4 Design of Experiments (DoE) 8780

16.7.4.1 Two levels factorial Designs 8794

16.7.4.2 General factorial Designs 8823

16.7.4.3 Taguchi Designs and Nomenclature (robust designs) 8837

16.7.4.4 Response Surface Methodology (Box Domains) 8848

16.7.4.5 Optimal Designs 8865

16.7.4.6 Mixture Design 8874

16.7.4.7 General DoE diagnostic tools 8896

16.7.5 Quality Control on Reception (Lot Acceptance Sampling Plans) 8904

16.7.5.1 Simple acceptance sampling plan by measurement for a unique tolerance with known standard deviation 8908

16.7.5.2 Simple acceptance sampling plan by attribute 8918

16.7.5.3 Double acceptance sampling plan by attribute 8926

16.7.5.4 Operating characteristic curve (OC) 8927

16.7.5.5 Average outgoing quality (AOQ) 8930

16.7.6 Quality Control Charts (CC) 8934

16.7.6.1 WECO's empirical rules 8938
16.7.6.2 Sample size and Sampling frequency for Control Charts 8940
16.7.6.3 Attributes Control Charts (qualitative CC) 8944
16.7.6.4 Measurement Control Charts (quantitative CC) 8963
16.7.6.5 Autocorrelated Measurement Control Charts (time weighted control charts) . . 9007
16.7.6.6 Rare events control charts 9034
16.7.6.7 Control Charts Operating Characteristic (OC) Curves 9045
16.7.7 Design of reliability tests 9049
16.7.7.1 Chi-squared time of test 9050
16.7.7.2 Binomial sampling size 9053
16.7.7.3 Test to demonstrate MTBF (or MTTF) 9057
16.7.7.4 Beta-binomial sampling size 9059

16.1. Marine & Weather Engineering

METEOROLOGY is the study of atmospheric phenomena such as clouds, weather depression and precipitation to understand how they form and evolve. It is a discipline that deals mainly with fluid mechanics applied to air but which makes use of various other branches of physics and chemistry. It therefore allows for weather forecasts based on mathematical models on (actually) relatively short and long term. It is also applied for the prediction of air quality, climate change and the study in several areas of human activity (construction, air traffic, etc.).

Remark

The Greek philosopher Aristotle is often cited as one of the founding fathers of meteorology with his treatise *Meteorologica* (written approximately in 9650 according to holocene calendar) commonly regarded as one of the earliest attempts to understand the Earth's atmosphere and the water cycle. Aristotle already understood and described quite accurately the cycle of water and knew that cloud were made of evaporated water and are therefore very heavy considering their whole structure. Aristotle also knew that the wind was not made by the clouds, he claimed that the origin of wind was the dry, hot exhalation.

Meteorology is linked to a large amount of variables that would be very difficult to make a list even not exhaustive... However on our planet Earth, an important factor not to be neglected is the one formed by ocean surfaces and their intrinsic dynamics which we will try to present through a summary mathematical study some of their properties. What is nice is that some results can also be applied to other Planets and sometimes very exotic one like Jupiter.

Now before we continue, one might ask why it happens that "our" planet (Earth) is one of the habitable ones and is perfectly in the habitable zone? But that is not a good question, really! Because first the Earth was inhabitable in the past and will be again inhabitable in a few hundred of million of years. Secondly, the Earth isn't perfectly in the habitable zone because that zone is not a circle without width but a ring with a width of 0.95 to 1.15 astronomical units. Third and last one, whatever intelligent beings appear who are able to ask such questions will look around and necessarily see that they are living on a planet that has the right conditions for life. No coincidence!

And also before beginning this section we also strongly recommend the reader to have a look of the section of Astronomy (page 5217) as the distance of the Earth (in general any planet...) to its main star (placed on the foci of the orbit) has an influence on the power of the radiation (see section Optical Geometry page 4337) and is therefore directly related to the eccentricity e of the orbit. Especially the inclination of the Earth rotation relatively to the ecliptic plane has a strong influence on the weather conditions as it is related to the existence of the seasons!

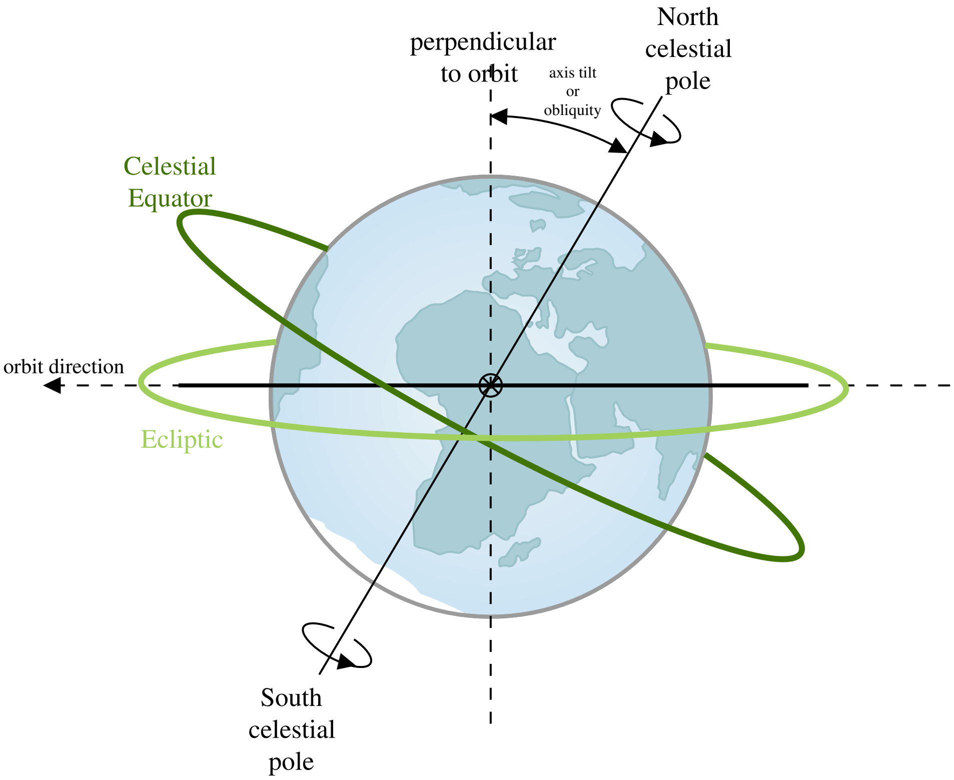

North celestial pole
perpendicular to orbit
axis tilt or obliquity
Celestial Equator
orbit direction
Ecliptic
South celestial pole

Figure 16.1: **Earth's inclination**

And the reader must be careful about the statement that at the periapsis people say it's winter because in fact... it depends where you live!!! Indeed, if you live in the northern hemisphere it will be winter, but at the southern hemisphere it will be summer (people that have the money to travel during holidays know this fact very well) as illustrated in the figure below:

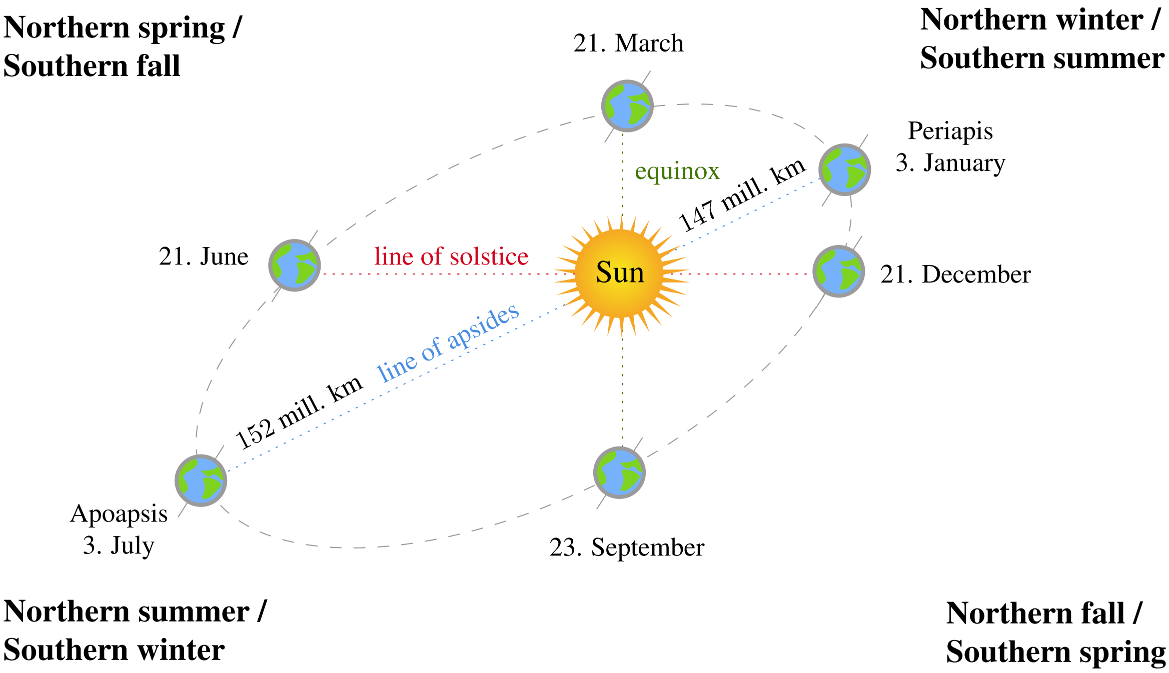

Northern spring / Southern fall
21. March
Northern winter / Southern summer
Periapis 3. January
equinox
147 mill. km
21. June
line of solstice
Sun
21. December
line of apsides
152 mill. km
Apoapsis 3. July
23. September
Northern summer / Southern winter
Northern fall / Southern spring

Figure 16.2: **Seasons space point of view**

The reader must also consider that the main factors of surface weather are:

- Latitude
- Topography
- Temperature
- Humidity
- Pressure
- Cloud cover
- Water cover
- Snow/Ice cover
- Wind circulation
- Sun activity (solar radiation)
- Moon position (tides)
- Sea/Ocean proximity
- Volcanoes activity
- Soil composition
- Earth atmosphere thickness and composition (on the long term)
- Earth speed rotation[1] (on the very long term)
- Distance to the Sun (on the very long term)
- ...

So given the number of variables... it's crystal clear that it's much more easier to predict the trajectory of an asteroid in vacuum on the long term where there is only one variable to predict rather than weather and climate where there are multiple hard to control/measure variables.

The combination of all the short term factors is well represented in the figure below (the years in the legend of the below image are given based on the holocene calendar):

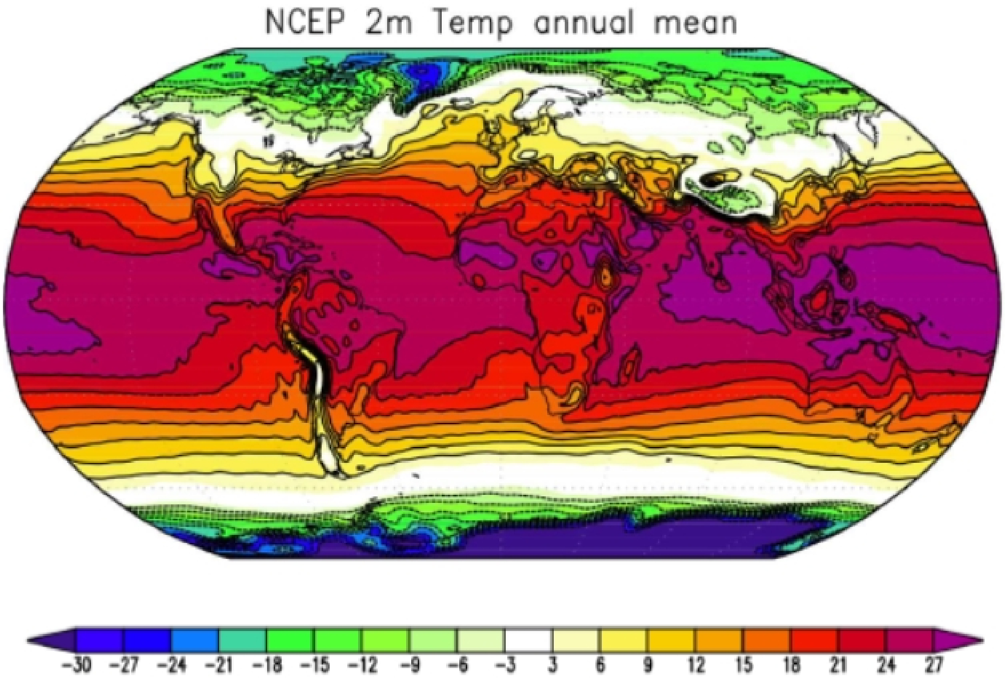


Figure 16.3: **Annual mean temperatures in 2 [m] height in the atmosphere from the NCEP data set over the period 11950-12000 (values in °C)**

---

[1] Earth has a mean increase of the day length of about 2.177 [ms $\cdot$ century^{-1}]. That means the day length (if we do a stupid naive linear interpolation) doubles approximately every 40 million years...

The simulated evolution of Earth's day length over time is also interesting to see and to know (one-million years is represented by "Ma" meaning "mega annum"):

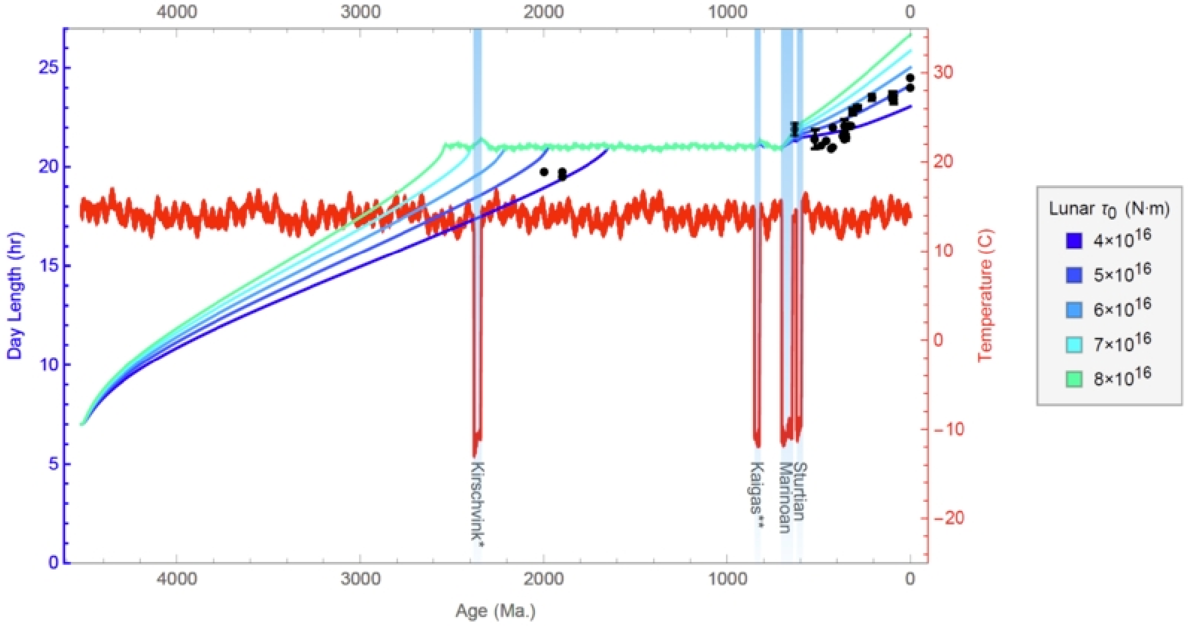


Figure 16.4: **Earth simulated day length evolution (source: [43])**

The plot depicts the simulated evolution of Earth's day length over time for varying choices of lunar torque τ_0 (shown in blue and green) and temperature values (red) over the lifetime of the Earth. Resonant-stabilizing effects are present from 600 million years to 1800-2600 million years, depending on the choice of lunar torque. Note that atmospheric thermal noise does not influence the day length except very near resonance, and that the resonant effect remains unbroken through this noise until two successive simulated snowball events at the end of the Precambrian 720 million years and 640 million years, corresponding to recent estimates of the Sturtian and Marinoan glaciations. Recapture events can be seen at 870 million years following the "Kaigas glaciation" and, for some values of τ_0, following a Paleoproterozoic glaciation detailed in Kirschvink et al. (12000 according to holocene calendar). Approximate empirical day length data from a compilation in Williams (12000 according to holocene calendar) are overlaid in black (error bars included where present) and resemble the simulated results, though the reader should not take this data to be too reliable, particularly the data points prior to 600 million years!

And let us take a look of the estimated geological time-scale evolution of temperature for Earth:

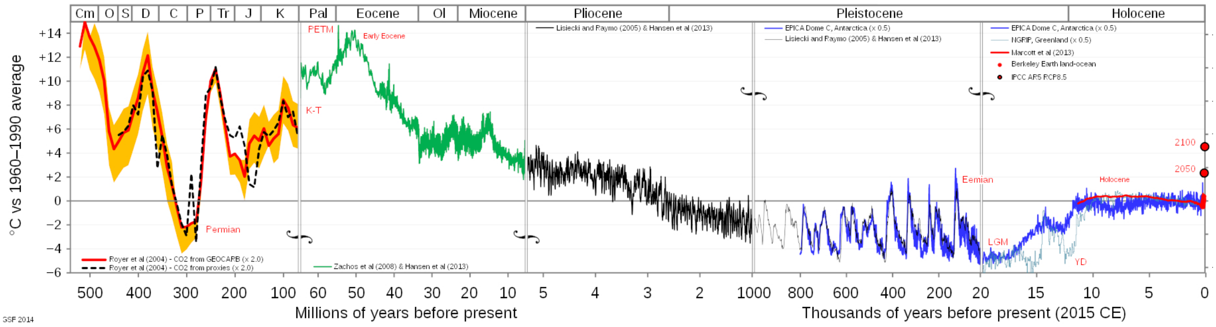


Figure 16.5: Geological time-scale evolution of temperature for Earth (source: http://gergs.net/all_palaeotemps-2, author: Glen Fergus)

This section is a general introduction to the basic to the field of technical applications of thermodynamics and fluid mechanics. It will allow the reader to become familiar with the language and some fundamental methods of calculations used by engineers in this branch. Of course, this study must be completed by practical laboratory and simulation work on computer.

And please, don't confuse "weather" with "climate"!!!

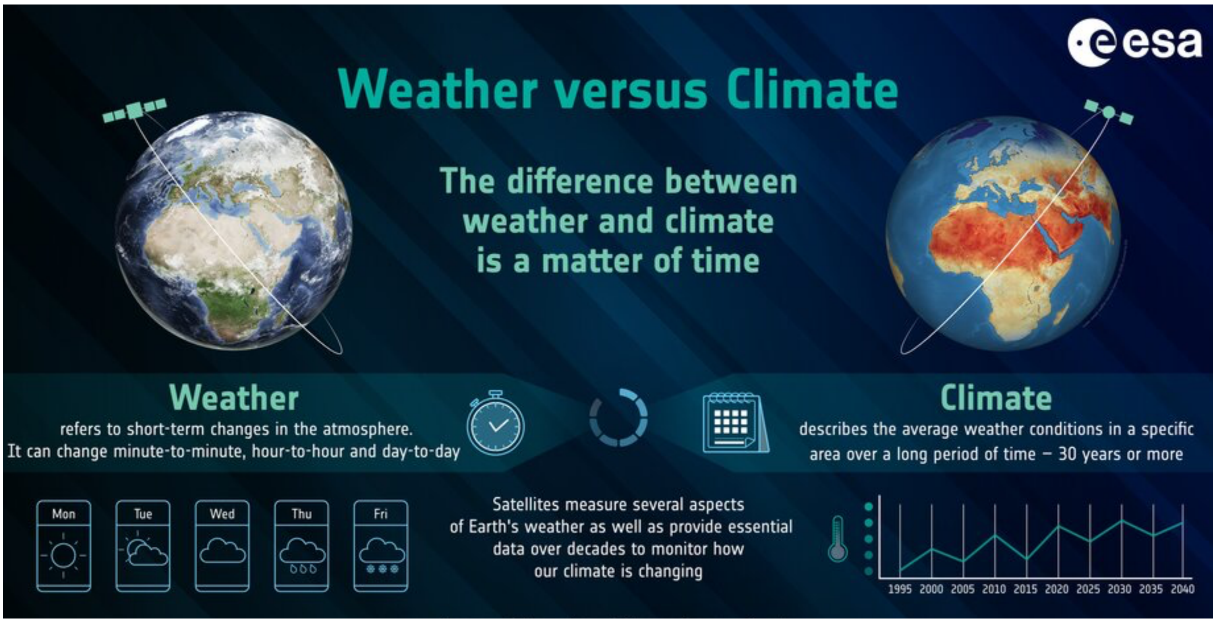
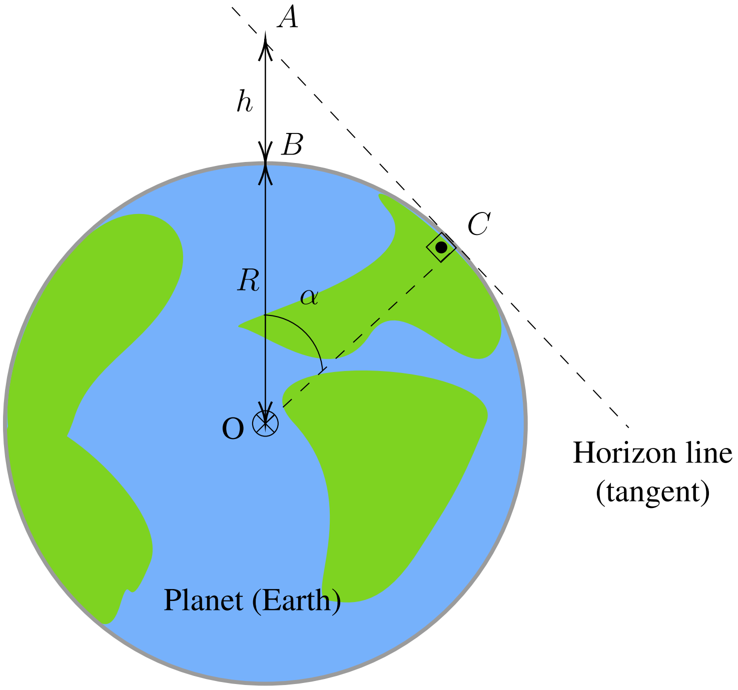

esa
Weather versus Climate
The difference between weather and climate is a matter of time
Weather
refers to short-term changes in the atmosphere.
It can change minute-to-minute, hour-to-hour and day-to-day
Climate
describes the average weather conditions in a specific area over a long period of time – 30 years or more
Mon
Tue
Wed
Thu
Fri
Satellites measure several aspects of Earth's weather as well as provide essential data over decades to monitor how our climate is changing
1995
2000
2005
2010
2015
2020
2025
2030
2035
2040

Figure 16.6: Weather vs Climate (source: ESA)

16.1.1 Visual horizon

Let us consider first a little funny subject making often friendly debate during holidays or more seriously ... In some meteorological softwares, it is required during temperature and pressure measurement to manually enter the visual horizon... but this is difficult to determine when the weather is very good and we are in a height position (on a hill or small mountain).

For this, consider the Earth as a perfect sphere with radius R and a point of view of height h relative to sea level which we will denote by A (and without atmosphere to avoid optical effects...). The question is to know how far can be by maximum the point C if it is by definition given by the tangent $\overline{AC}$ which is simply the horizon line:



The reader can probably already observed that the study will mainly appeal to trigonometry concepts (see corresponding section of this textbook) and elementary geometry (also see corresponding sections).

The angle $\widehat{\text{O}CA}$ is a right angle. Indeed, a straight tangent at a point on a circle is perpendicular to the radius at that point. The triangle $\widehat{\text{O}CA}$ is therefore rectangle on C.

Therefore we have:

$$\cos(\alpha)=\frac{\overline{\text{O}C}}{\overline{\text{O}A}}=\frac{\overline{\text{O}C}}{\overline{\text{O}B}+\overline{BA}} \tag{16.1.1}$$

But, we have $\overline{\text{O}B}=\overline{\text{O}C}=R$. Hence we deduce:

$$\cos(\alpha)=\frac{R}{R+h}\Rightarrow\alpha=\arccos\left(\frac{R}{R+h}\right) \tag{16.1.2}$$

The $\overline{AC}$ distance is the distance in a straight line between our viewpoint and the (floor of the) boat that we see on the horizon. The distance, however, that interest us here is $\overline{BC}$: it is the distance that we should travel at the altitude 0 to reach the ground of the boat we see on the horizon.

For what will follows we will put for the curvilinear distance: $\overline{BC}=d$.

When the angle α varies from 0° to 360° (full circle), we describe the entire circumference of the Earth, that is to say $2\pi R$, since the Earth is assumed to be round.

Using the rule of three:

If an angle of 360° corresponds to a distance of $2\pi R$ then an angle α in degrees corresponds to a curvilinear distance:

$$d_{\text{curv.}}=\frac{2\pi R}{360^\circ}\alpha=\frac{\pi R}{180^\circ}\alpha \tag{16.1.3}$$

But we have seen previously that:

$$\alpha=\arccos\left(\frac{R}{R+h}\right) \tag{16.1.4}$$

Hence finally:

$$d_{\text{curv.}}=\frac{\pi R}{180^\circ}\alpha=\frac{\pi R}{180^\circ}\arccos\left(\frac{R}{R+h}\right) \tag{16.1.5}$$

With $R\cong 6,378\;[\text{km}]$, we find ($h$ must be expressed in kilometres):

$$d_{\text{curv.}}\cong 111.32\arccos\left(\frac{6,378}{6,378+h}\right) \tag{16.1.6}$$

In Microsoft® Excel® this is given by:

```
=111.32*DEGREES(ACOS(6378/(6378+h)))
```

We have then in vacuum, in a landscape without obstacles... the following table:

Altitude h [m]	**Horizon distance $d_{\text{curv.}}$ [km]**
1.5	4.3
5	8
10	11.3
50	25.3
100	35.7
200	50.5
400	71.4
600	87.5
800	101
$1,000$	113
$2,000$	159.7
$3,000$	195.6
$4,000$	225.8
$5,000$	252.5
$10,000$	357

Table 16.1: **Visual horizon curvilinear distance in function of the altitude (in vacuum)**

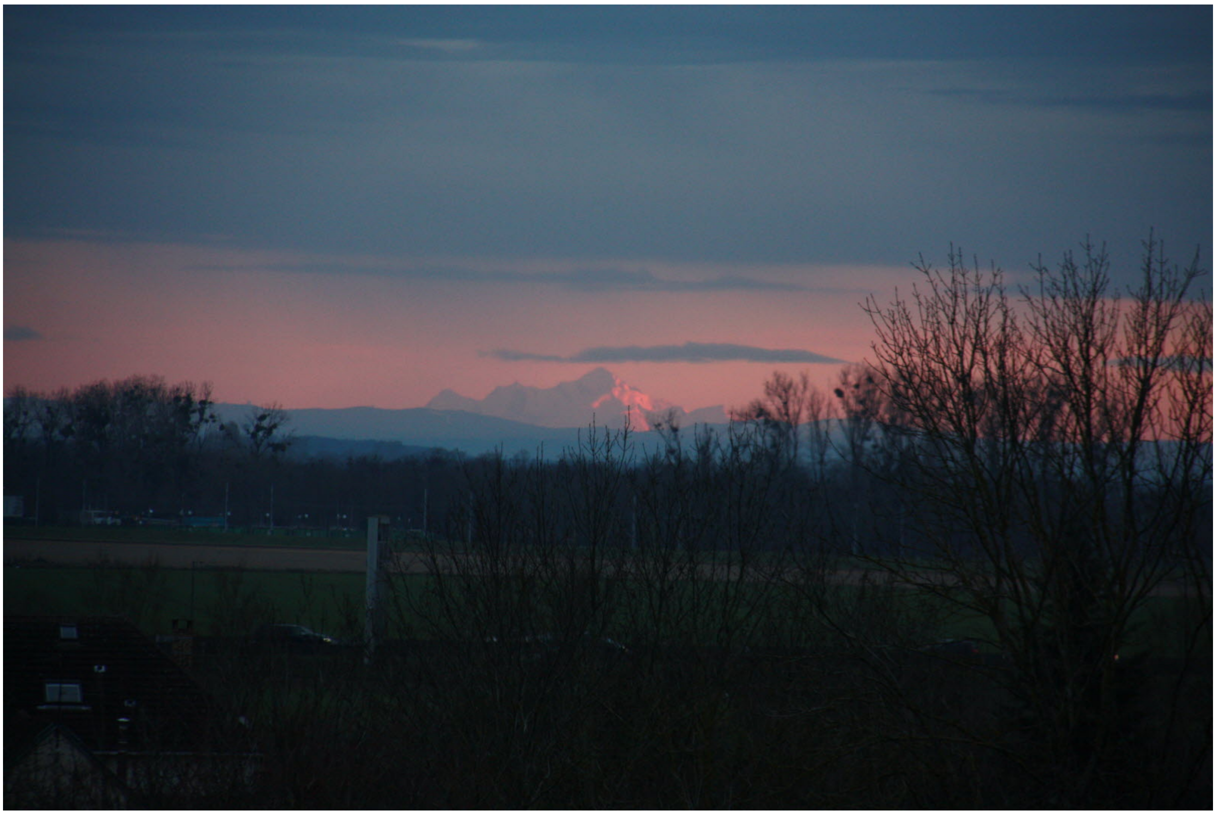


Figure 16.7: **Mont-Blanc ($4,809$ [m]) as seen from the City of Ducs de Bourgogne near Dijon at the 4th floor of a building (hence an elevation of 346 [m]) at a distance of 268 [km] the 9th January 12020 at 17:36 (author: Serge Fourcaudot)**

There is another famous formula we can found in many textbook that consider the rectilinear distance $d_{\text{rec.}}=\overline{AB}$ and not the curvilinear one. In that case:

$$(R+h)^2=R^2+d_{\text{rec.}}^2 \Rightarrow R^2+h^2+2Rh=R^2+d_{\text{rec.}}^2 \tag{16.1.7}$$

And since $h/R\cong 10^{-5}$, we may ignore h^2, and isolating $d_{\text{rec.}}$ gives:

$$d=\sqrt{2Rh} \tag{16.1.8}$$

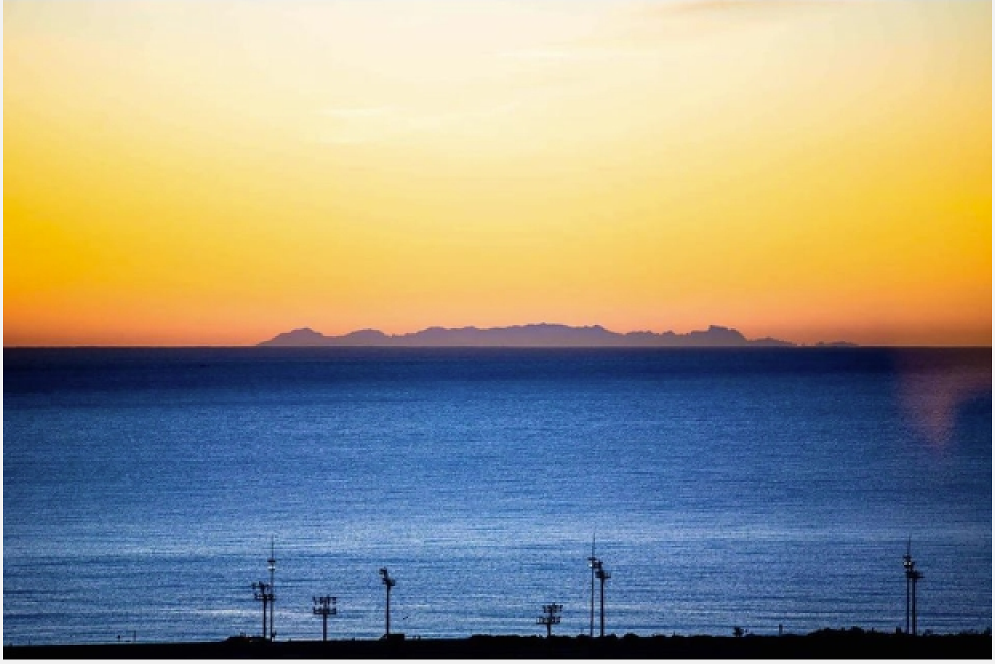
Using this formula it is customary to measure the height h in meters, and the distance in kilometres. If we therefore insert $h/1000$ instead of h, and the radius of the Earth in kilometres, we obtain a well known formula:

$$d \cong 3.57\sqrt{h}\,[\text{km}] \tag{16.1.9}$$

As the radius of the Earth is quite big, the difference between this approximation and the previous derived relation leads to a difference for every row of the above table of a few dozens of meters only!

Remark

If we do not take into account atmospheric refraction, we find in the table above that would need to go to an altitude of the order of several kilometres to see beyond 200 [km] away. Yet at the heights of Nice (Alpes-Maritimes in France) at the Garoupe lighthouse (78 [m] of altitude), it is possible to observe the upper part (like the Monte Cinto mountain at $2,706$ [m]) of Cap Corse (it's an island for information...) which is about 220 [km] from the continent (especially from the Alps were the photo below was taken) however at the same moment people at Cap Corse can't see the Garoupe lighthouse!!! Atmospheric refraction (especially in winter) plays then a significant role in this phenomenon (see photo below taken by John Bouchet the 15 January 12019 at the morning) as proven in the section of Optics at page 4397.



Using the refraction effect derived at page 4412 and considering a rising angle of $\varphi = 0$, we get theoretically the double of the distance (as $\mathrm{d}n/\mathrm{d}h = 1/R$ in this perfect theoretical case), however if we take an experimentally measured gradient in Winter of $\mathrm{d}n/\mathrm{d}h \cong 2.4 \cdot 10^{-7}$, then we get (don't forget to convert the output of $\arccos$ in degrees when using a calculator!):

$$d_{\text{hor.,ref.}} \cong 111.32 \arccos\left(\frac{6,378}{6,378+h}\right)\left[1+\frac{\mathrm{d}n}{\mathrm{d}h}6,378\cdot 10^{3}\right] \cong 4.11\sqrt{h}$$

where the approximation using the square root has been made using a solver on a computer software. Therefore we get an increase of approximately 15% in comparison of the case without refraction that was given by $3.57\sqrt{h}$.

The questions that arise are: given that we are at a height h_0 and we look far at the horizon:

1. What is the distance of the horizon (same as above but with a different approach used by people that argue that the Earth is flat and that the NASA lies...)?
2. At what level under the horizon is an object at a distance d of the observer (eye)?

To answer these questions consider the figure below:

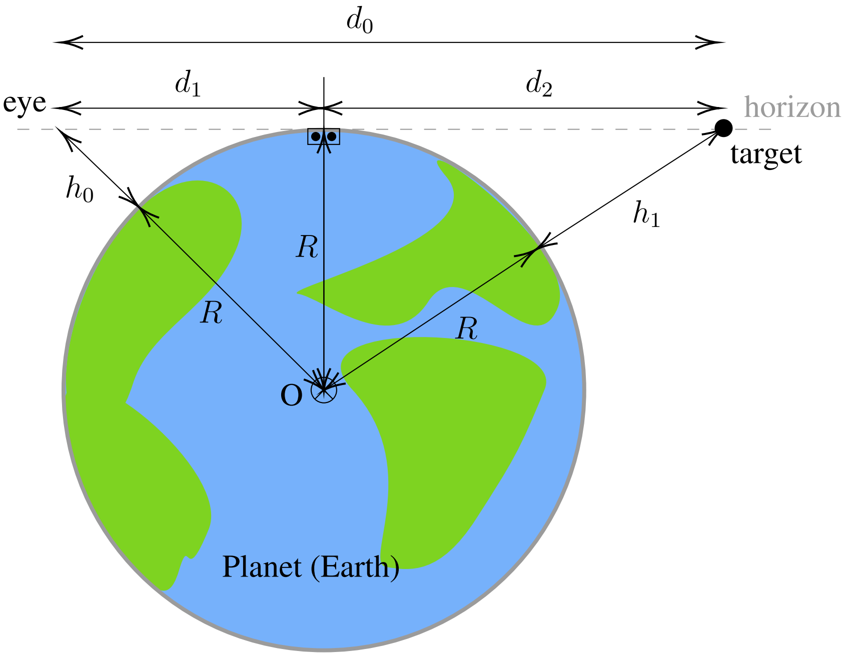


Figure 16.8: **Visual Horizon configuration experience**

The application of Pythagoras gives us immediately:

$$(R+h_0)^2=d_1^2+R^2 \Leftrightarrow d_1=\sqrt{h_0^2+2Rh_0} \tag{16.1.10}$$

Hence:

$$\boxed{d_1=\sqrt{h_0^2+2Rh_0}} \tag{16.1.11}$$

And we can get also the minimum height given d_0, d_1:

$$(R+h_1)^2=d_2^2+R^2 \Leftrightarrow h_1=\sqrt{d_2^2+R^2}-R=\sqrt{(d_0-d_1)^2+R^2}-R \tag{16.1.12}$$

So let us compare the value of d_1 just obtained above (straight distance) for $d_{\text{curv.}}$ obtained previously that is the curvilinear distance (but this time at two decimal places):

Altitude h [m]	Horizon distance $d_{\text{curv.}}$ [km]	Horizon distance d_1 [km]
1.5	4.374	4.374
5	7.98	7.98
10	11.29	11.29
50	25.25	25.25
100	35.71	35.71
200	50.51	50.50
400	71.43	71.43
600	87.48	87.48
800	101.01	101.02
$1,000$	112.93	112.94
$2,000$	159.70	159.73
$3,000$	195.58	195.64
$4,000$	225.83	225.92
$5,000$	252.47	252.59
$10,000$	356.93	357.29

Table 16.2: **Visual horizon curvilinear and straight distance in function of the altitude (in vacuum)**

So the bias of the people that believe that the Earth (and all other planets) is flat ("flat earthers") is that when they do the experiments to prove their arguments they forget the "perfectly spheric" assumption of that model. Indeed as we have seen in the section of Astronomy the Earth's shape is more like a irregular potatoes and is locally (at human scale, i.e. a few hundreds of kilometres sometimes) absolutely not spherical due to relief irregularities.

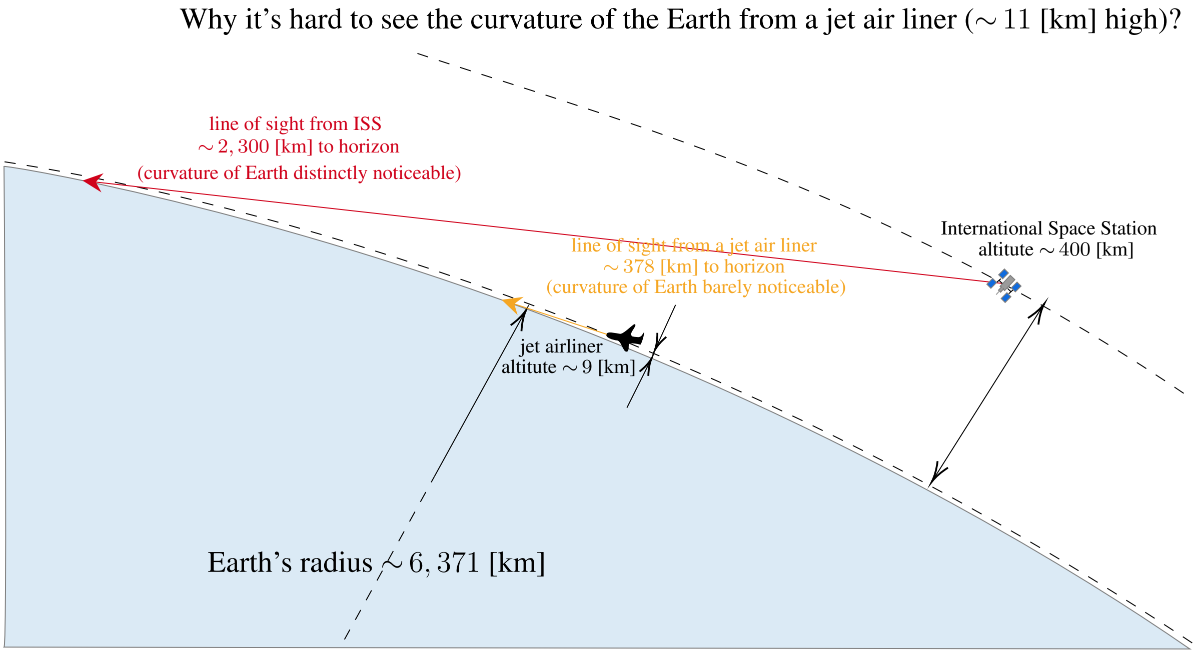

Why it's hard to see the curvature of the Earth from a jet air liner (~ 11 [km] high)?
line of sight from ISS
~ 2,300 [km] to horizon
(curvature of Earth distinctly noticeable)
International Space Station
altitute ~ 400 [km]
line of sight from a jet air liner
~ 378 [km] to horizon
(curvature of Earth barely noticeable)
jet airliner
altitute ~ 9 [km]
Earth's radius ~ 6,371 [km]

Figure 16.9: **Earth's curvature visibility**

Anyway an another argument is that we can prove with Isoperimetric Inequality (see section Geometric Shapes page 2993) that natural final shape for a cloud of dust is a sphere and otherwise that the acceleration of objects that fall all respect Newton's law in it's... spherical form! And also... if the Earth was flat... nobody so far as we know has seen it's border...

Therefore the two relations to remember are (after standardizing the notations):

$$\begin{aligned} d_{\text{straight}} &= \sqrt{h^2+2R_{♁}h} \\ d_{\text{curvilinear}} &\cong 111.32 \arccos\left(\frac{R_{♁}}{R_{♁}+h}\right) \end{aligned} \tag{16.1.13}$$

Keep in mind that these results are for a perfect sphere in vacuum. When an atmosphere exist, refraction effect (see proof in the section of Optics at page 4397) can drastically increase the distances with theoretically almost no limit.

Taking into account the refraction (again see the proof at page 4412):

$$\Delta\varphi = -\frac{\cos(\varphi)}{n}\frac{\mathrm{d}n}{\mathrm{d}y}\Delta s \tag{16.1.14}$$

that leads to the following relation:

$$\begin{aligned} d_{\text{curv.refr.}} &\cong \Delta(\Delta s) + d_{\text{curv}} = \Delta\varphi R + d_{\text{curv}} \\ &= \left(\frac{\cos(\varphi)}{n}\frac{\mathrm{d}n}{\mathrm{d}y}\Delta s\right) R + d_{\text{curv}} \\ &= \left[\frac{\cos\left(\arcsin\left(\dfrac{d_{\text{straight}}}{R+h}\right)\right)}{n}\frac{\mathrm{d}n}{\mathrm{d}y} R \sin\left(\frac{d_{\text{straight}}}{R+h}\right)\right] R + d_{\text{curv}} \end{aligned} \tag{16.1.15}$$

To summarize:

$$\boxed{\begin{aligned} d_{\text{straight}} &= \sqrt{h^2+2R_{♁}h}\;[\text{m}] \\ d_{\text{curvilinear}} &\cong 111.32 \arccos\left(\frac{R_{♁}\;[\text{km}]}{R_{♁}\;[\text{km}]+h\;[\text{km}]}\right)\;[\text{km}] \\ d_{\text{curv.refr.}} &\cong \left[\frac{\cos\left(\arcsin\left(\dfrac{d_{\text{straight}}}{R+h}\right)\right)}{n}\frac{\mathrm{d}n}{\mathrm{d}y} R \sin\left(\frac{d_{\text{straight}}}{R+h}\right)\right] R + d_{\text{curvilinear}} \cdot \frac{1000\cdot 360}{2\pi} \end{aligned}} \tag{16.1.16}$$

If we take the same case taken as the one at the page 4397 (i.e. $n=1$ and $\mathrm{d}n/\mathrm{d}y \cong 2.4\cdot 10^{-8}$), we then have (between parentheses we put how the distance is reduced as it's a matter of point of view to say that the object is further or closer):

Altitude h [m]	Horizon distance $d_{\text{curv.}}$ [km]	Horizon distance $d_{\text{curv.refr.}}$ [km]
1.5	4.374	5.043 (3.704)
5	7.98	9.201 (6.758)
10	11.29	13.018 (9.561)
50	25.25	29.115 (21.385)
100	35.71	41.176 (30.243)
200	50.51	58.241 (42.778)
400	71.43	82.363 (60.496)
600	87.48	100.869 (74.095)
800	101.01	116.469 (85.550)
1,000	112.93	130.213 (95.646)
2,000	159.70	184.139 (135.262)
3,000	195.58	225.503 (165.656)
4,000	225.83	260.376 (191.283)
5,000	252.47	291.085 (213.854)
10,000	356.93	411.406 (302.379)

Table 16.3: **Visual horizon curvilinear in function of the altitude in vacuum vs in air with $n=1$ and $\mathrm{d}n/\mathrm{d}y \cong 2.4\cdot 10^{-8}$**

According to NSPS (National Society of Professional Surveyors) in their *Definitions of Surveying and Associated Terms* (see [199]), the surveyor has to take into account and offset of about 8 [cm] at 1 [km] distance. This is the offset which is given in surveying tables as correction for curvature of the Earth. The values given in such tables usually combine the effects of both curvature the Earth and atmospheric refraction. Under normal conditions, the atmospheric refraction compensates partially for the curvature of the Earth, so the correction is about 6 [cm] for a point 1 [km] away.

If you take the biggest spirit level available on the market, i.e. 200 [cm], you wouldn't be able to identify at the naked eye the Earth's curvature as the maximum difference of height between the two extremities would then be, according to NSPS rule, of only 0.016 [cm], hence 0.16 [mm].

But where does the number above of the NSPS come from? That's quite easy! Let us consider the following figure:

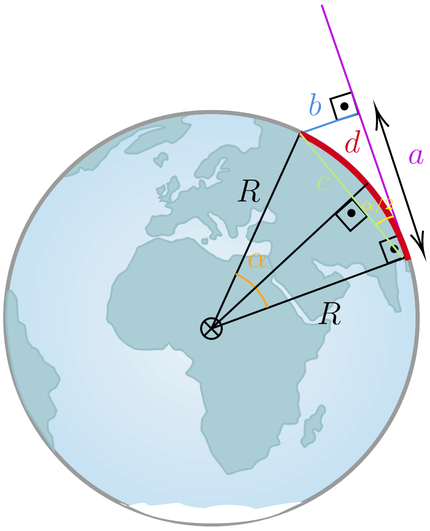


Figure 16.10: **Distance between point on tangent line and circle**

For the Earth the distance d is given by (where for reminder α is in radians):

$$d=R\cdot\alpha\Leftrightarrow\alpha=\frac{d}{R} \tag{16.1.17}$$

Consider that we are interested in the value of b for $d=1$ [km].

First notice that in the figure above, we have:

$$\sin\left(\frac{\alpha}{2}\right)=\frac{c/2}{R} \tag{16.1.18}$$

Rearranging:

$$c=2R\sin\left(\frac{\alpha}{2}\right)=2R\sin\left(\frac{d/R}{2}\right) \tag{16.1.19}$$

We see also that in the coloured triangle, we have:

$$\sin\left(\frac{\alpha}{2}\right)=\frac{b}{c} \tag{16.1.20}$$

Rearranging:

$$b=c\sin\left(\frac{\alpha}{2}\right)=c\sin\left(\frac{d/R}{2}\right) \tag{16.1.21}$$

And injecting the previous result:

$$b=\left(2R\sin\left(\frac{d/R}{2}\right)\right)\sin\left(\frac{d/R}{2}\right)=2R\sin^2\left(\frac{d/R}{2}\right) \tag{16.1.22}$$

Hence for $d=1$ [km], we have:

$$b=2R\sin^2\left(\frac{d/R}{2}\right)=2\cdot 6,371\cdot\sin^2\left(\frac{1/6,371}{2}\right)\cong 7.8\,[\text{cm}] \tag{16.1.23}$$

When $R\gg d$ we can us the Maclaurin approximation of the $\sin$ and therefore:

$$b=2R\sin^2\left(\frac{d/R}{2}\right)\cong 2R\left(\frac{d/R}{2}\right)^2=\frac{d^2}{2R} \tag{16.1.24}$$

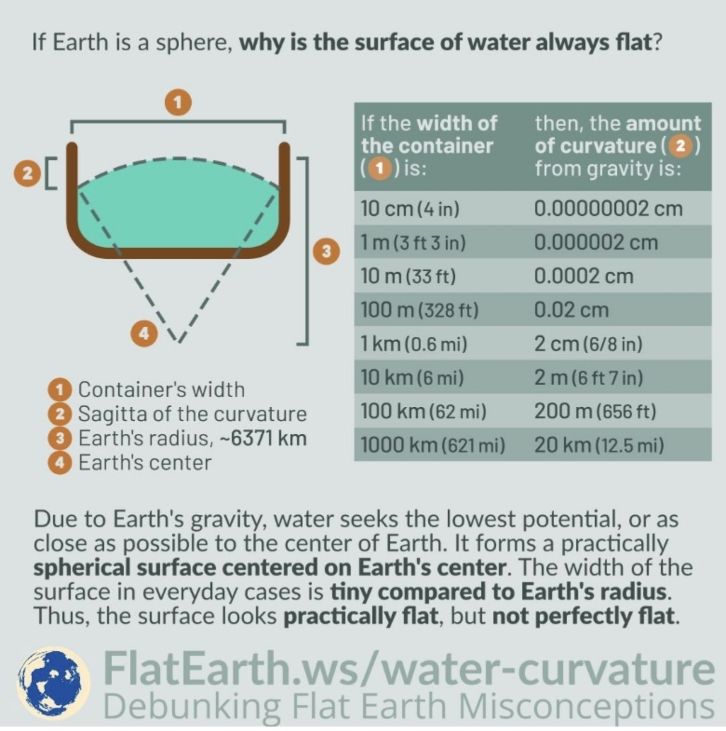
So we get the famous "8 inches per miles squared" geometrical curvature drop (and not visual drop as this result assumes a vacuum!) approximation formula often denoted:

$$\boxed{C_d=\frac{d^2}{2R}} \tag{16.1.25}$$

Remark

Assuming the "8 inches per mile" is accurate, that means airplanes would have to "nose down" to a pitch of $-1''$ per $7290''$ flown (equivalent of a mile), or a pitch of -0.007 degrees. That's pretty much level flight. The pilot nor autopilot would even be able to discern the difference given the accuracy of instruments used by autopilots nowadays (their PID - proportional–integral–derivative controller based on fuzzy logic - measure and get a few thousand of times per second the GPS localization, the radar altitude plus the atmospheric height using a Pitot tube). Most human eyes can't even eyeball 1 degree so imagine $7/1000$ths of a degree.

Let us now deal with the oblique horizon distance mainly because of the quite famous figure below:



If Earth is a sphere, **why is the surface of water always flat**?

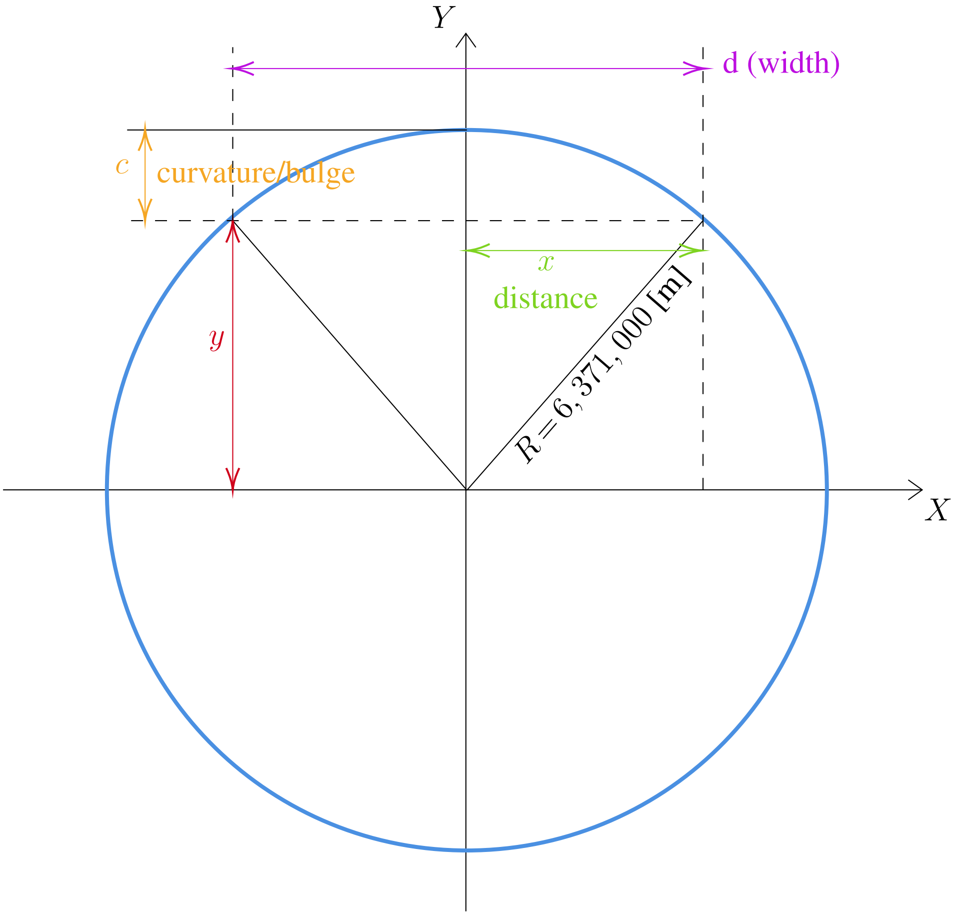
If the **width of the container** (①) is:	then, the **amount of curvature** (②) from gravity is:
10 cm (4 in)	0.00000002 cm
1 m (3 ft 3 in)	0.000002 cm
10 m (33 ft)	0.0002 cm
100 m (328 ft)	0.02 cm
1 km (0.6 mi)	2 cm (6/8 in)
10 km (6 mi)	2 m (6 ft 7 in)
100 km (62 mi)	200 m (656 ft)
1000 km (621 mi)	20 km (12.5 mi)

① Container's width
② Sagitta of the curvature
③ Earth's radius, ~**6371 km**
④ Earth's center

Due to Earth's gravity, water seeks the lowest potential, or as close as possible to the center of Earth. It forms a practically **spherical surface centered on Earth's center**. The width of the surface in everyday cases is **tiny compared to Earth's radius**. Thus, the surface looks **practically flat**, but **not perfectly flat**.

FlatEarth.ws/water-curvature
Debunking Flat Earth Misconceptions

To see where the values of the table visible in the illustration above come from, let us consider the following figure:



And now let's to the math to find c! That's elementary, applying Pythagore we have first:

$$x^2 + y^2 = R^2 \tag{16.1.26}$$

Therefore

$$y = \sqrt{R^2 - x^2} \tag{16.1.27}$$

But as we have $c = R - y$, hence:

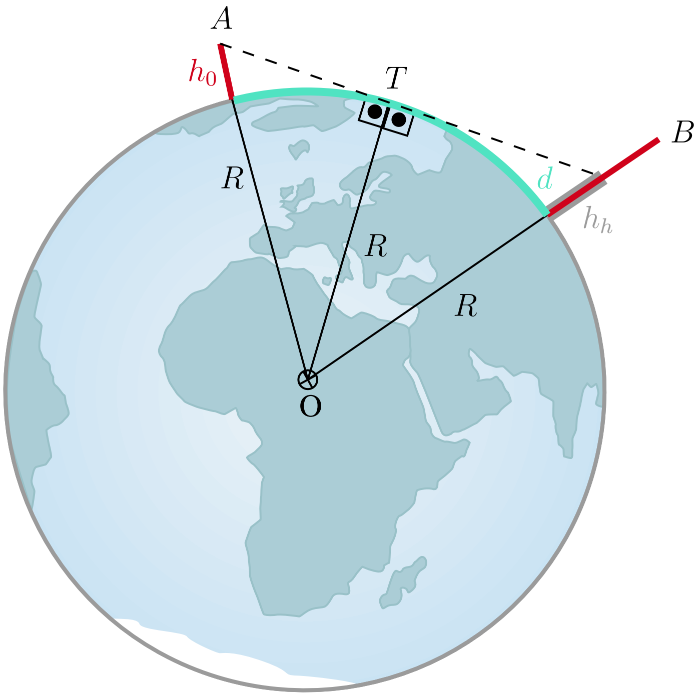
$$c = R - \sqrt{R^2 - x^2} \tag{16.1.28}$$

and as $d = 2x \Rightarrow x = d/2$ we have:

$$c = R - \sqrt{R^2 - \frac{1}{4}d^2} \tag{16.1.29}$$

And by applying the relation above we get the table visible in the figure above!

Let us consider now the most important case about the horizon (!): how to calculate the hidden height h_h given a point of view at a height h_0 and a curvilinear distance d in-between as illustrated below:



In the figure above we have:

$$\overline{A\text{O}} = h_0 + R \tag{16.1.30}$$

by trigonometry:

$$\widehat{A\text{O}T} = \arccos\left(\frac{R}{h_0+R}\right) \tag{16.1.31}$$

and:

$$\widehat{A\text{O}B} = \frac{d}{R} \tag{16.1.32}$$

Therefore:

$$\widehat{T\text{O}B} = \frac{d}{R} - \arccos\left(\frac{R}{h_0+R}\right) \tag{16.1.33}$$

By the law of sines (see section Trigonometry page 2528):

$$\frac{h_0+R}{h_h+R} = \frac{\sin\left(\frac{\pi}{2} - \widehat{T\text{O}B}\right)}{\sin\left(\frac{\pi}{2} - \widehat{A\text{O}T}\right)} \tag{16.1.34}$$

Using trigonometric identities, this simplifies to:

$$\frac{h_0+R}{h_h+R} = \frac{\cos\left(\widehat{T\text{O}B}\right)}{\cos\left(\widehat{A\text{O}T}\right)} \tag{16.1.35}$$

Hence:

$$\frac{h_0+R}{h_h+R}=\frac{\cos\left(\dfrac{d}{R}-\arccos\left(\dfrac{R}{h_0+R}\right)\right)}{\cos\left(\arccos\left(\dfrac{R}{h_0+R}\right)\right)}=\frac{\cos\left(\dfrac{d}{R}-\arccos\left(\dfrac{R}{h_0+R}\right)\right)}{\dfrac{R}{h_0+R}} \tag{16.1.36}$$

After simplification and inversion:

$$h_h+R=\frac{R}{\cos\left(\dfrac{d}{R}-\arccos\left(\dfrac{R}{h_0+R}\right)\right)} \tag{16.1.37}$$

After rearranging:

$$h_h=\frac{R}{\cos\left(\dfrac{d}{R}-\arccos\left(\dfrac{R}{h_0+R}\right)\right)}-R \tag{16.1.38}$$

Injecting in it the effective radius of Earth (see page 4413), we get the famous exact equation for the hidden height:

$$\boxed{h_h=\frac{R'}{\cos\left(\dfrac{d}{R'}-\arccos\left(\dfrac{R'}{h_0+R'}\right)\right)}-R'} \tag{16.1.39}$$

And keep in mind that h_h is the height an object must have at the horizon to be at the limit of visibility when we stand at a height h_0 of observation (if is it greater that h_h it will be visible!).

So let us take for all the relations established previously with:

$$\begin{aligned}
d_{\text{straight}}&=\sqrt{h^2+2R_{\text{♁}}h}\;[\text{m}]\\
d_{\text{curvilinear}}&\cong 111.32\arccos\left(\frac{R_{\text{♁}}\,[\text{km}]}{R_{\text{♁}}\,[\text{km}]+h\,[\text{km}]}\right)\;[\text{km}]\\
d_{\text{curv.refr.}}&\cong\left[\frac{\cos\left(\arcsin\left(\dfrac{d_{\text{straight}}}{R+h}\right)\right)}{n}\frac{\mathrm{d}n}{\mathrm{d}y}R\sin\left(\frac{d_{\text{straight}}}{R+h}\right)\right]R+d_{\text{curvilinear}}\cdot\frac{1000\cdot 360}{2\pi}\\
\frac{\mathrm{d}n}{\mathrm{d}y}&\cong-\frac{(n(y)-1)}{T(y)}\left(\frac{\rho(y)gT(y)}{P(y)}+\frac{\mathrm{d}T}{\mathrm{d}y}\right):=-\frac{(n(y)-1)}{T(y)}(\gamma_g-\gamma)
\end{aligned} \tag{16.1.40}$$

Let us take some conservative typical values:

$$\gamma=0.0113\;\left[\Delta°\text{C}\cdot\text{m}^{-1}\right]\qquad T=295.15\,[\text{K}]\qquad P=101,325\,[\text{Pa}] \tag{16.1.41}$$

$$n=1.000293\qquad \rho_{\text{SATP}}=1.168\;\left[\text{kg}\cdot\text{m}^{-3}\right]\qquad R=6,378,137\,[\text{m}] \tag{16.1.42}$$

Therefore we get the following table:

$\mathbf{h_0}$ [m]	$\mathbf{d}_{\text{straight}}$ [m]	$\mathbf{d}_{\text{curvilinear}}$ [m]	$\mathbf{d}_{\text{curv.refr.}}$ [km]	h_h [m]
0	0	0	0	0
10	$11,294$	$11,294$	13	10
100	$35,716$	$35,716$	41	100
$1,000$	$112,948$	$112,937$	129	998
$2,000$	$159,739$	$159,706$	182	$1,995$
$3,000$	$195,647$	$195,587$	223	$2,993$
$4,000$	$225,923$	$225,829$	257	$3,991$
$5,000$	$252,599$	$252,468$	288	$4,989$
$6,000$	$276,719$	$276,547$	315	$5,986$
$10,000$	$357,299$	$356,928$	406	$9,977$
$12,000$	$391,432$	$390,944$	445	$11,973$

16.1.1.1 Discerning the curvature of the Earth

The first direct visual detection of the curvature of the horizon has been widely attributed to Auguste Piccard and Paul Kipfer on 27 May 11931 (holocene calendar). They reported seeing it from a hydrogen-filled balloon at an elevation of $15,787\,[\text{m}]$ over Germany and Austria. On 11 November 11935 (holocene calendar), Albert W. Stevens and Orville A. Anderson became the first people to photograph the curvature. They were flying in the helium-filled Explorer II balloon during a record-breaking flight to an altitude of $22,066\,[\text{ m}]$ over South Dakota. Other claims have been made as to being the first to see the curvature of the Earth, but they seem to have come long after visual curvature had been established.

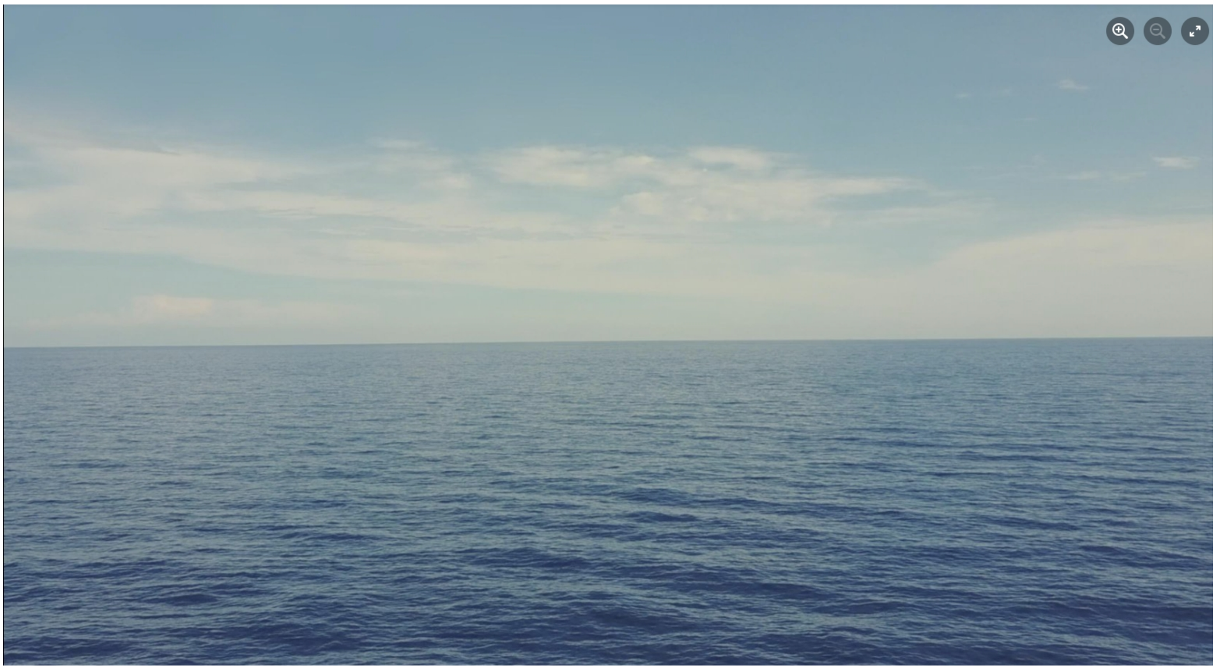


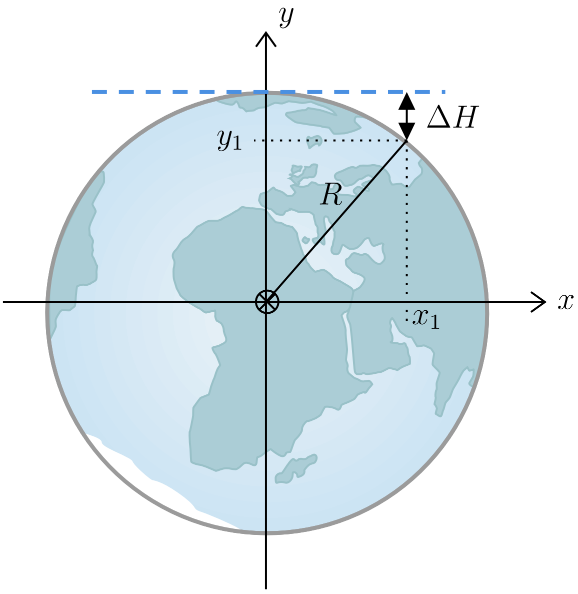
Figure 16.11: **Horizon line with sea view**

Since that time, countless people have claimed to be able to discern the curvature of the Earth as an upwardly arched horizon from high mountains or commercial aircraft. Some claim to see it from sea level or relatively low elevations. We know that if we get high enough (i.e., from space), the curvature of the Earth is evident, but commercial aircraft seldom exceed altitudes of $12,000\,[\text{m}]$.

Interviews with pilots and high-elevation travelers revealed that few if any could detect curvature below about $15,240\,[\text{m}]$. High-altitude physicist and experienced sky observer David Gutierrez reported that as his B-57 ascends, the curvature of the horizon does not become readily sensible until about $50,000\,[\text{m}]$ and that at $60,000\,[\text{m}]$ the curvature is obvious. Having talked to many other high fliers (SR-71, U2, etc.), Gutierrez confirms that his sense of the curvature is the same as theirs. Passengers on the Concorde $(60,000\,[\text{m}])$ routinely marveled at the curvature of the Earth. Gutierrez believes that if the field of view (FOV) is wide enough, it might be possible to detect curvature from lower altitudes. The author has also talked to many commercial pilots, and they report that from elevations around $16,100\,[\text{m}]$, they cannot see the curvature.

When trying to understand the perception of a curved horizon, two issues must be kept in mind. First, a large fraction of people wear eye glasses. Eye glasses produce a variety of distortions when the observer is not looking through the center part of the lens. Second, above the altitude of Mt. Everest, no observer can look directly at the horizon - he must look through a window or canopy. Planeparallel windows like those on most aircraft will not render a flat horizon curved, but a curved window or canopy will.

Let us calculate the magnitude of the declination ΔH in the extremities of a view like the one above relatively to the horizon line (in blue below) based on the following schema:



We know the equation of the horizon line:

$$y_L = R \tag{16.1.43}$$

We know the equation of the circle that defines Earth's shape:

$$x^2 + y^2 = R^2 \tag{16.1.44}$$

Therefore:

$$\Delta H = R - y_1 = R - \sqrt{R^2 - x^2} \tag{16.1.45}$$

However this relation is incomplete! Indeed, x is limited by the horizon, i.e. the distance at which you can't see what is beyond the curvature. Hence the previous relation should be written instead:

$$\Delta H_{\text{max}} = R - \sqrt{R^2 - x_{\text{max}}^2} \tag{16.1.46}$$

Using what we have seen earlier above, we then have:

$$\Delta H_{\text{max}} \cong R - \sqrt{R^2 - (\sqrt{2Rh})^2} = R - \sqrt{R^2 - 2Rh} \tag{16.1.47}$$

Hence for a person of approximately 1.81 [m] height and for Earth:

$$\Delta H_{\text{max}} \cong 6,371 \cdot 10^3 - \sqrt{(6,371 \cdot 10^3)^2 - 2 \cdot 6,371 \cdot 10^3 \cdot 1.81} \cong 1.81000025757 \text{ [m]} \tag{16.1.48}$$

To convert from linear dimensions to angular dimensions, we need only divide each quantity by the straight distance to the horizon (that's a not to bad approximation):

$$d_{\text{straight}} = \sqrt{h^2 + 2Rh} \tag{16.1.49}$$

Therefore (and changing a bit the notations):

$$\boxed{\alpha \cong \frac{\Delta H}{d_{\text{straight}}} = \frac{R_{\earth} - \sqrt{R_{\earth}^2 - 2R_{\earth}h}}{\sqrt{h^2 + 2R_{\earth}h}}} \tag{16.1.50}$$

Hence for a person of approximately 1.81 [m] height we get for the vertical angle below the true horizon:

$$\alpha \cong \frac{\Delta H}{\sqrt{h^2 + 2R_{\earth}h}} \cong \frac{1.810000258}{4802.397659} = 0.00037689512 \text{ [rad]} \tag{16.1.51}$$

The average human eye (see [586]) has an angular resolution of about 1 arcminute (0.02 degrees or 0.0003 radians). Therefore to see the curvature of the Earth at the beach from the eye height of 1.81 [m].

Therefore the curvature at sea level is at the limit of the human eye angular resolution. Let us see a table for different values of h:

Altitude h	**Angle α in degrees**
1 [m]	0.0160
10 [m]	0.0507
100 [m]	0.160
$1,000$ [m]	0.507
$10,000$ [m]	1.605
$30,000$ [m]	2.783
100 [km]	5.09630
$1,000$ [km]	16.907
$3,000$ [km]	40.299

Table 16.4: **Vertical angle below horizon (degrees)**

The diffraction limit of the eye can be calculated using Rayleigh's criterion (see section Electromagnetisme page 4529):

$$\delta \cong \frac{1.22R\lambda}{D} \tag{16.1.52}$$

where lambda is the wavelength of light (on the average, about 550 [nm]) and D is the diameter of the eye's pupil, which is about 5 [mm] indoors. This calculation results in an angular resolution of 0.008 degrees (at least for eyes having a $20/20$ vision... and keep in mind that about 35% percent of all adults have a $20/20$ vision without glasses).

Remark

An object that is 1 [cm] wide at a distance of 100 [cm] is the same angular size as an object 10 [m] wide at a distance of $1,000$ [m], which is the same as the angular size of an object $10,000$ [m] wide at a distance of $1,000$ [km]. So the resolution limit of the eye in world space is a constant angle. In linear terms, the smallest resolvable pattern has a size that is directly proportional to distance.

In view of the agreement between the visual observations, measurements of the photographs, and the theoretical curvatures, it seems well established that the curvature of the Earth is reasonably well understood and can be measured from photographs. For more see [364].

16.1.2 Wind direction

We will now prove mathematically something quite intuitive: the winds are moving from high to low pressure (it's silly like that but we must still prove it!).

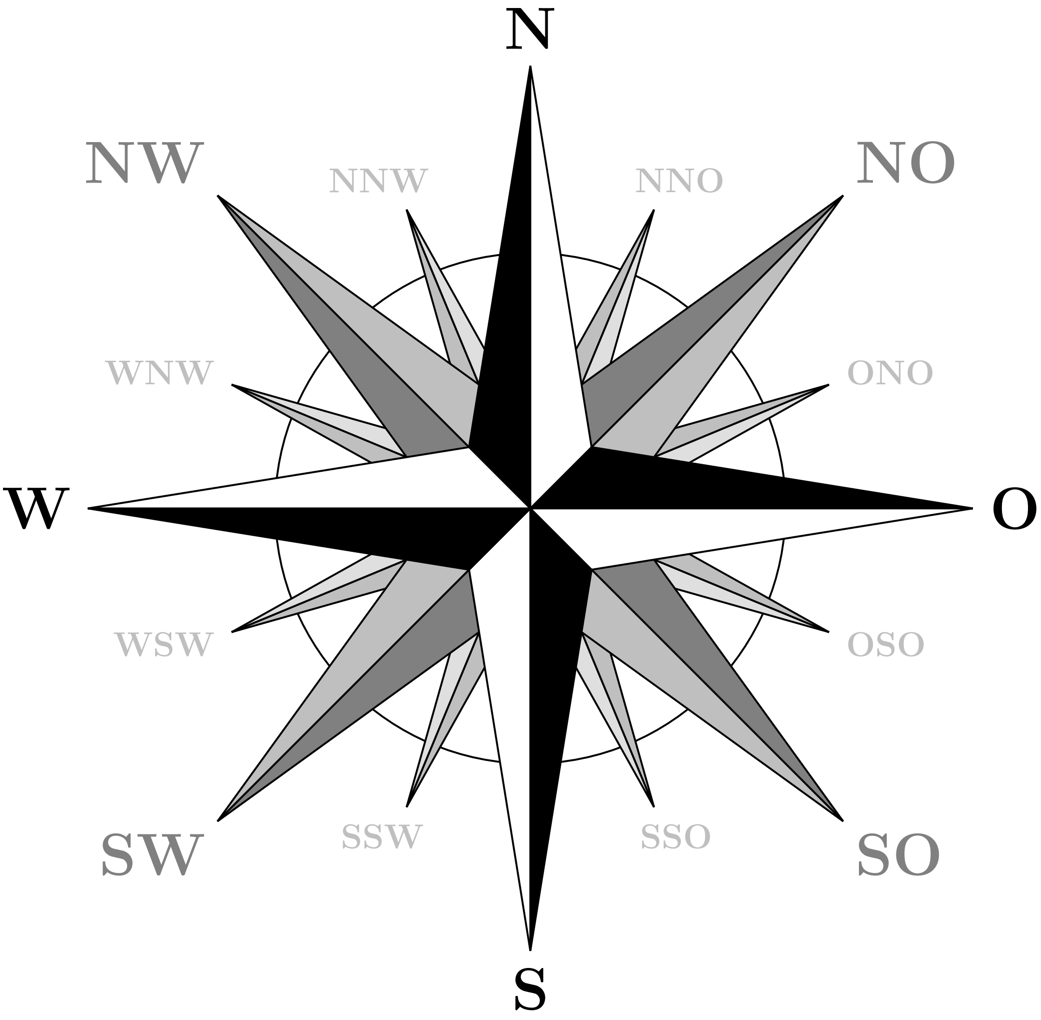
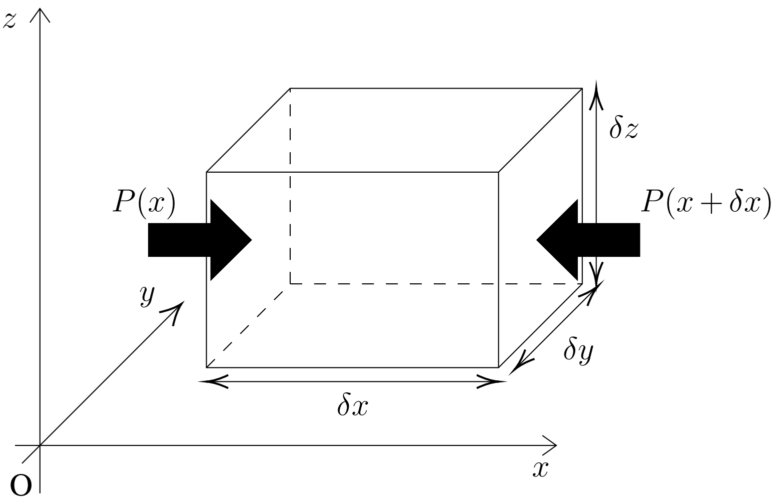


Figure 16.12: **A "compass rose", sometimes named a "wind rose", "rose of the winds" or "compass star"**

We know (see section Continuum Mechanics page 3735) that the pressure force exerted on a surface S is normal to this surface and is equal in scalar form to $P \cdot S$.

For an air parcel volume $\mathrm{d}V = \delta x \delta y \delta z$ the total pressure force in the x direction is then:

$$F_P = P(x)\delta y \delta z - P(x+\delta x)\delta y \delta z \tag{16.1.53}$$



In addition, we have (see section Differential and Integral Calculus page 1450):

$$\frac{P(x+\delta x)-P(x)}{\delta x}=\frac{\partial P}{\partial x} \tag{16.1.54}$$

Therefore:

$$F_P=P(x)\delta y\delta z-P(x+\delta y\delta z)=-\frac{\partial P}{\partial x}\delta x\delta y\delta z \tag{16.1.55}$$

The mass pressure force is then:

$$f_p=\frac{F_p}{m}=\frac{F_p}{\rho\delta x\delta y\delta z}=-\frac{1}{\rho}\frac{\partial P}{\partial x} \tag{16.1.56}$$

We can do the same calculation following y. Finally, the massic horizontal pressure force is given by:

$$\boxed{\vec{f}_p=-\frac{1}{\rho}\begin{pmatrix}\frac{\partial P}{\partial x}\\ \frac{\partial P}{\partial y}\end{pmatrix}=-\frac{1}{\rho}\vec{\nabla}P} \tag{16.1.57}$$

Thus, the pressure force (massic or not) is opposite to the horizontal gradient. The transmission of information (of the force) is done at the speed of sound for this equation (which explains the speed of air calls in your home or apartment and then force that can make slam doors or window).

It is therefore:

- Directed from high to low pressure, perpendicular to the isobars
- Inversely proportional to the spacing of the isobars.

If we measure the values of the atmospheric pressure at different points of the globe and we connect between them the points of same pressure on a drawing, we get a series of curves, named "isobars". The wind is directly determined by the atmospheric relief since it is a movement of air between the high to low pressure.

The wind speed is then determined by the pressure gradient: that is, if the atmospheric pressure changes rapidly with distance, the wind blows strong, while it will be low in abarometric "swamp" where this pressure remains almost unchanged on great distances. In summary, more the isobars are close together, more the wind will blow strong.

The isobars are traditionally marked with a pitch of 5 millibars on weather maps such as shown in the example below:

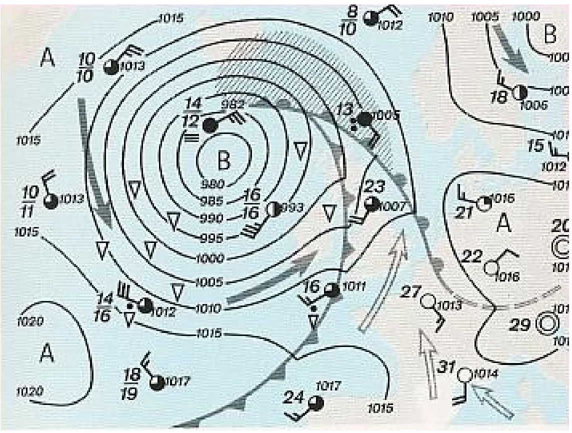

A
B
A
A
B
A
1015
1020
1010
1005
1000
995
990
985
980
982
993
1007
1011
1012
1013
1014
1016
1017
1006

Figure 16.13: Typical representation of isobars (on a depression)

Charts showing isobars are useful because they identify features such as anticyclones (areas of high pressure) and depressions (areas of low pressure).

Areas of high and low pressure are caused by ascending and descending air. As air warms, it ascends leading to low pressure at the surface. As air cools, it descends leading to high pressure at the surface. This is illustrated in the diagram below.

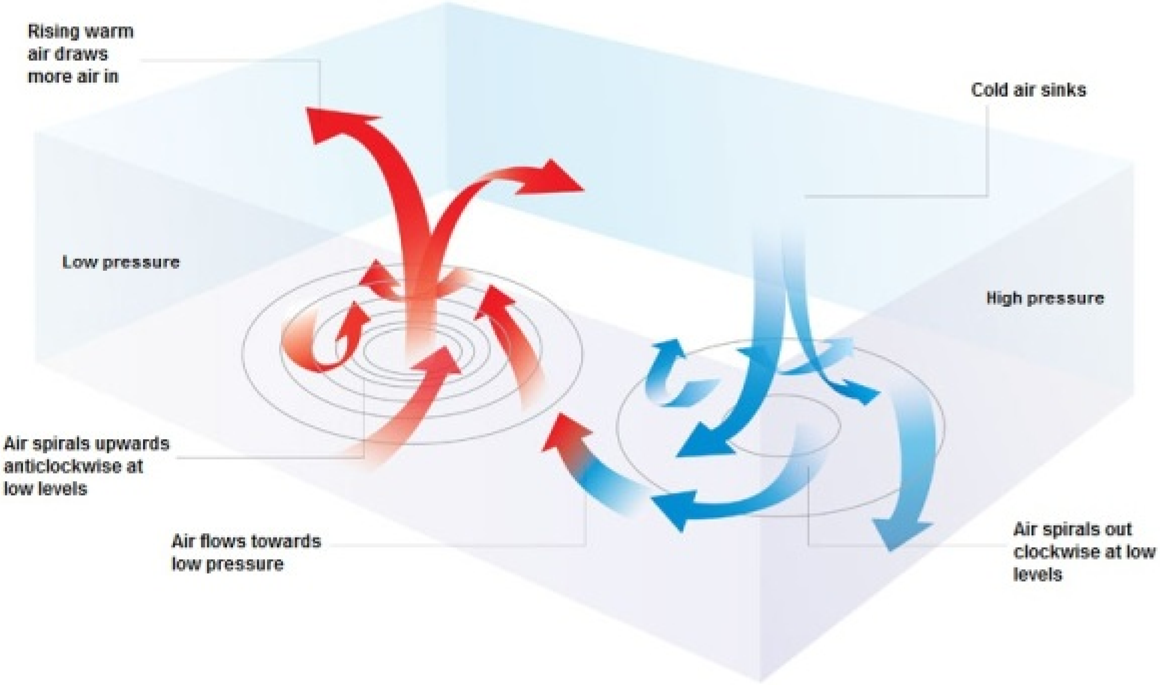

Rising warm air draws more air in
Cold air sinks
Low pressure
High pressure
Air spirals upwards anticlockwise at low levels
Air flows towards low pressure
Air spirals out clockwise at low levels

Figure 16.14: High and low pressure systems

In general, low pressure leads to unsettled weather conditions and high pressure leads to settled weather conditions.

In an anticyclone (high pressure) the winds tend to be light and blow in a clockwise direction (in the northern hemisphere as we will prove it later below). Also the air is descending, which reduces the formation of cloud and leads to light winds and settled weather conditions.

In a depression (low pressure), air is rising and blows in an anticlockwise direction around the low (in the northern hemisphere). As it rises and cools, water vapour condenses to form clouds and perhaps precipitation. This is why the weather in a depression is often unsettled - there are usually frontal systems associated with depressions.

To close this topic about winds, let us notice that meteorologists empirically defined (that is fun for the general culture) a unit of wind measurement that is just a match between the wind force and the distance between two isobars in steps of 5 by 5 [mb]:

Distance between isobares [km]	**Unit** [Beaufort]	**Speed** $[\mathrm{m}\cdot\mathrm{s}^{-1}]$
600 (light breeze)	2	1.6-3.3
500 (average breeze)	4	3.4-5.4
400 (cold breeze)	5	8-10.7
300 (strong wind)	6	10.8-13.8
200 (big wind)	7	13.9-17.1
100 (storm)	9	20.8-24.4

Table 16.5: **Distance between isobars and wind speed (Beaufort units)**

16.1.3 Atmospheric Profile Models

An atmospheric profile model is a mathematical model constructed to predict the profile of temperature, pressure or other in the atmosphere, locally depending on the measured height supposing a steady state of the atmosphere (same as if we take out a cylinder piece of the atmosphere and freeze time to analyse some of the properties of the atmosphere inside depending on the height in the cylinder).

There a lot of models but let us see just two easy one:

16.1.3.1 Atmospheric Exponential Profile Model

Consider that the atmosphere is a perfect fluid in a uniform gravity field. Then from the following Bernoulli's theorem relation, proved in the section of Continuum Mechanics (static fluid at page 3783):

$$P=-\rho g z+c^{te} \tag{16.1.58}$$

It comes then:

$$\frac{\partial P}{\partial z}=-\rho g \tag{16.1.59}$$

Thus, to know the variation of pressure depending on height in the atmosphere or depth in the ocean, we have taken as hypothesis the "hydrostatic equilibrium", that is the variation of pressure with height / depth is proportional to the gravity and density of the fluid.

This is obviously not valid in the case of rapid movements of convection, like thunderstorms, but can easily be satisfied for slow movements (quasi-static) and at large scale: the "synoptic scale" ($2,000$-$20,000$ [km]).

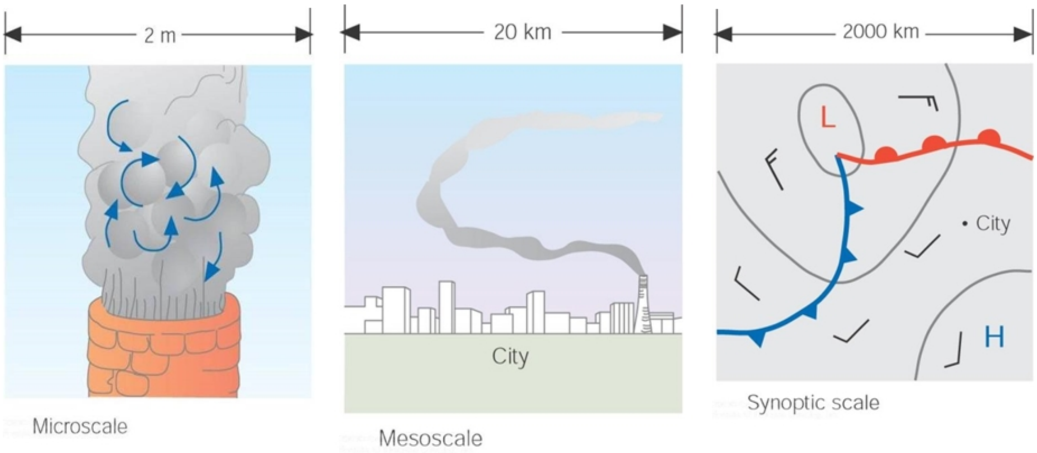


Figure 16.15: **Various meteorological scales**

We will then combine this last relation with a state equation, such as that of the ideal gas at temperature T and of density ρ whose constituent particles have a mass m. So we have the ideal gas equation (see section Continuum Mechanics page 3903):

$$PV = nRT \Rightarrow P = \frac{nRT}{V} = \frac{\rho kT}{m} \Rightarrow \rho = \frac{mP}{kT} \tag{16.1.60}$$

with for recall P that is the pressure expressed in Pascals, V the volume in cubic meters, R is the gas constant, T is the temperature in Kelvin, n the number of moles, k is the Boltzmann constant, ρ the particle density, m the total mass of the particles.

In the isothermal case (e.g. in the terrestrial stratosphere, above 10 [km] altitude where the temperature is almost constant around -55 degrees Celsius), we found out easily by substitution that:

$$\frac{\partial P}{\partial z} = -\rho g \Rightarrow \frac{\partial P}{\partial z} = -\frac{mP}{kT}g \tag{16.1.61}$$

Therefore, at a given pressure, the vertical pressure gradient is inversely proportional to temperature.

Let us now consider the following relation:

$$\frac{\mathrm{d}P}{P} = -\frac{m}{k}\frac{g}{T}\mathrm{d}z \Rightarrow \ln(P) = -\frac{mg}{kT}z + c^{te} \tag{16.1.62}$$

Using the exponential we the get "barometric height formula":

$$\boxed{P = c^{te}e^{-\frac{mg}{kT}z} = P_0 e^{-\frac{mg}{kT}z} = P_0 e^{-z/H}} \tag{16.1.63}$$

The pressure therefore decreases exponentially with the altitude. P_0 being the pressure at ground level and H is named the "scale height".

Remark

R1. Notice that in the isothermal case, where $T = c^{te}$, using the law of perfect gas $P = nkT$, we get after simplification:

$$n = n_0 e^{-\frac{mg}{kT}z} \tag{16.1.64}$$

R2. A "barometric altimeter" is an instrument that measures air pressure and displays the altitude (height) of an object. The conversion from pressure to height involves a nonlinear calibration using a relation derived from the result above, e.g.:

$$z = -\frac{kT}{mg}\ln\left(\frac{P}{P_0}\right) \tag{16.1.65}$$

However, outdoor readings of height (altitude) from a pressure based altimeter can vary by tens of metres owing to a sudden change in air pressure, such as from a cold front or other meteorological minor event.

Let us return to the relation:

$$\frac{\partial P}{\partial z} = -\frac{mP}{kT}g \tag{16.1.66}$$

It can of course also be written as:

$$\frac{\partial z}{\partial P} = -\frac{k}{mPg}T \tag{16.1.67}$$

which tells us that the distance z between isobaric surfaces is directly proportional to temperature.

We have also another common approach. Let us start again from the relation proved just earlier above, but for a mass m of 1 [kg]:

$$\frac{\mathrm{d}P}{\mathrm{d}z} = -\frac{g}{kT}P \tag{16.1.68}$$

and let us denote this relation as follow:

$$\frac{\mathrm{d}_z P(z)}{P(z)} = -\frac{g}{kT(z)} \tag{16.1.69}$$

Let us recall that (see section Differential and Integral Calculus page 1475):

$$\mathrm{d}_x \ln(f(x)) = \frac{1}{f(x)}\mathrm{d}_x f(x) \tag{16.1.70}$$

Therefore:

$$\mathrm{d}_z \ln(P(z)) = -\frac{g}{kT(z)} \tag{16.1.71}$$

and now let assume that the temperature variation is linear in the atmosphere (which is not far from the truth for the first 10 to 20 km of the atmosphere such that:

$$T(z) \cong T_0 + \lambda z \tag{16.1.72}$$

with $\lambda < 0$ which is the temperature gradient in $[\mathrm{K} \cdot \mathrm{m}^{-1}]$.

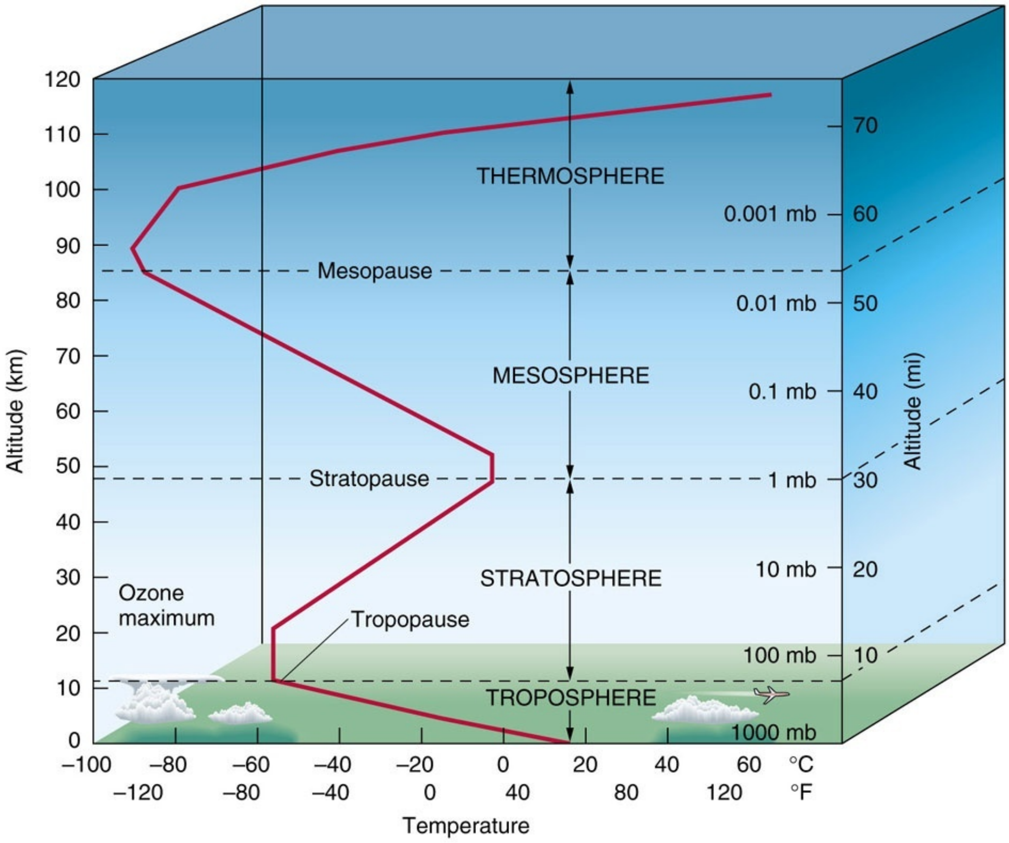


Figure 16.16: **Typical profile of temperature and pressure on Earth... late 120th century (source: ?)**

Then we have:

$$\mathrm{d}_z \ln(P(z)) = -\frac{g}{k(T_0+\lambda z)} = \frac{-g/k}{T_0+\lambda z} \tag{16.1.73}$$

Therefore:

$$\int_{z_0}^{z_1} \mathrm{d}\ln(P(z)) = \int_{z_0}^{z_1} \frac{-g/k}{T_0+\lambda z}\mathrm{d}z \tag{16.1.74}$$

Which gives:

$$\ln(P(z_1)) - \ln(P(z_0)) = -\frac{g}{k\lambda}\ln(T_0+\lambda z)\Big|_{z_0}^{z_1} \tag{16.1.75}$$

After simplification:

$$\ln\left(\frac{P(z_1)}{P(z_0)}\right) = -\frac{g}{k\lambda}\ln\left(\frac{T_0+\lambda z_1}{T_0+\lambda_z 0}\right) \tag{16.1.76}$$

Hence:

$$P(z_1) = P(z_0)\left(\frac{T_0+\lambda z_1}{T_0+\lambda_z 0}\right)^{-\frac{g}{k\lambda}} \tag{16.1.77}$$

Which can be written more aesthetically:

$$\boxed{P(z)=P_0\left(\frac{T_0+\lambda z}{T_0+\lambda_z 0}\right)^{-\frac{g}{k\lambda}}} \tag{16.1.78}$$

A good practical example of application of this relation is the gliders and hang gliders. They expect the weather forecast that it communicates to them the height of the isotherm of zero degrees during its bulletins. They then deduce the temperature gradient per meter. For these athletes, good condition is to have a gradient of 1 [°C] by 100 meters. It is therefore easy with the above relation to calculate the pressure at an altitude of $2,000$ meters and to derive the pressure gradient which allows them to use some the updrafts for their aerobatic practices.

16.1.3.2 Adiabatic Atmosphere Model

The adiabatic temperature gradient is, in the atmosphere, the air temperature change with altitude (i.e. the air temperature gradient), which only depends on the atmospheric pressure, that is to say:

- Without considering heat exchange with the environment (other air masses, relief, ground, etc.)
- Without considering condensation (cloud formation) or precipitation.
- Without taking into account the Sun (actually a G2V-type star, aka yellow dwarf) position and variations

This concept is of great importance in meteorology, as well as aviation and maritime navigation.

We have proved in the section of Thermodynamics (see page 3646) the Laplace's thermodynamics law (satisfied under certain conditions!):

$$PV^\gamma=c^{te} \tag{16.1.79}$$

with the Laplace coefficient:

$$\gamma=\frac{C_P}{C_V} \tag{16.1.80}$$

Therefore in massic form the Laplace's law becomes:

$$\gamma=\frac{C_P}{\rho^\gamma}=c^{te} \tag{16.1.81}$$

We can take the logarithm:

$$\ln\left(\frac{P}{\rho^\gamma}\right)=\ln(c^{te})\Rightarrow\ln(P)-\gamma\ln(\rho)=c^{te}\Rightarrow\ln(P)=\gamma\ln(\rho)+c^{te} \tag{16.1.82}$$

But by taking the logarithmic differential:

$$\begin{aligned}&\frac{\mathrm{d}(\ln(P))}{\mathrm{d}P}=\frac{1}{P}\Rightarrow\mathrm{d}(\ln(P))=\frac{\mathrm{d}P}{P}\\&\frac{\mathrm{d}(\gamma\ln(\rho))}{\mathrm{d}\rho}=\gamma\frac{1}{\rho}\Rightarrow\mathrm{d}(\gamma\ln(\rho))=\gamma\frac{\mathrm{d}\rho}{\rho}\\&\mathrm{d}(c^{te})=0\end{aligned}\tag{16.1.83}$$

We then have (relationship that we will reuse in the section of Music Mathematics)

$$\frac{\mathrm{d}P}{P}=\gamma\frac{\mathrm{d}\rho}{\rho}\tag{16.1.84}$$

We could also have found this result directly from the relation proved in the section Thermodynamics:

$$\gamma\frac{\mathrm{d}V}{V}+\frac{\mathrm{d}P}{P}=0\Rightarrow\frac{\mathrm{d}P}{P}=-\gamma\frac{\mathrm{d}V}{V}=-\gamma\frac{\mathrm{d}\left(\dfrac{m}{\rho}\right)}{\dfrac{m}{\rho}}=-\rho\gamma\mathrm{d}\left(\frac{1}{\rho}\right)=\gamma\frac{\mathrm{d}\rho}{\rho}\tag{16.1.85}$$

By taking also the logarithmic differential of the ideal gas law where n is, for recall, the number of moles (see section Continuum Mechanics page 3903):

$$PV=nRT\tag{16.1.86}$$

But in the massic form for one mole:

$$\rho=\frac{M_mP}{RT}\tag{16.1.87}$$

where M_m is the molar mass (see section Analytical Chemistry page 6252), we have:

$$\ln(\rho)=\ln\left(\frac{M_mP}{RT}\right)=\ln(M_mP)-\ln(RT)\tag{16.1.88}$$

Therefore:

$$\frac{\mathrm{d}\rho}{\rho}=\frac{\mathrm{d}P}{P}-\frac{\mathrm{d}T}{T}\tag{16.1.89}$$

We then get:

$$\frac{1}{\gamma}\frac{\mathrm{d}P}{P}=\frac{\mathrm{d}P}{P}-\frac{\mathrm{d}T}{T}\Rightarrow\frac{\mathrm{d}T}{T}=\frac{\mathrm{d}P}{P}-\frac{1}{\gamma}\frac{\mathrm{d}P}{P}\Rightarrow\frac{\mathrm{d}T}{T}=\frac{\gamma-1}{\gamma}\frac{\mathrm{d}P}{P}\tag{16.1.90}$$

Therefore:

$$\frac{\mathrm{d}P}{P}=\frac{\gamma}{\gamma-1}\frac{\mathrm{d}T}{\mathrm{T}}\tag{16.1.91}$$

Let us use the relation proved earlier shown above:

$$\frac{\mathrm{d}P}{\mathrm{d}z}=-\frac{mg}{kT}P\Rightarrow\frac{\mathrm{d}P}{P}=-\frac{mg}{kT}\mathrm{d}z\tag{16.1.92}$$

It comes then:

$$\mathrm{d}T = \frac{1-\gamma}{\gamma}\frac{mg}{k}\mathrm{d}z \tag{16.1.93}$$

So we have an atmosphere with constant and negative thermal gradient (temperature decreases linearly with altitude):

$$\boxed{\kappa = \frac{1-\gamma}{\gamma}\frac{mg}{k} = \frac{1-\gamma}{\gamma}\frac{M_m g}{R}} \tag{16.1.94}$$

The last form using the molar mass is more convenient because it allows to characterize the studied medium. Note that this is a model that looks to work well for an altitude of 0 to 90 [km] on Venus but much less well for the planet Earth.

✍**Example:**

We then have for example for Earth (thus knowing that the model is not well adapted):

$$g \cong 9.81\,[\mathrm{m}\cdot\mathrm{s}^{-2}] \qquad R \cong 8.314\,[\mathrm{J}\cdot\mathrm{K}^{-1}\cdot\mathrm{mol}^{-1}]$$

and the adiabatic coefficient for air is equal to $\gamma = 1.4$, and its molecular weight:

$$M_{m,\text{air}} \cong 28.96\cdot 10^{-3}\,[\mathrm{kg}\cdot\mathrm{mol}^{-1}]$$

Therefore:

$$\kappa = -0.00976\,[\mathrm{K}\cdot\mathrm{m}^{-1}] \cong -1\,[\mathrm{K}\cdot\mathrm{hm}^{-1}]$$

which corresponds to the current idea: 1 degree per 100 meters.

16.1.3.2.1 Hypsometric equation

The hypsometric equation, also known as the thickness equation, relates an atmospheric pressure ratio to the equivalent thickness of an atmospheric layer under the assumptions of constant temperature and gravity. It is derived from the hydrostatic equation and the ideal gas law.

So we have just proved for to hydrostatic equilibrium:

$$\frac{\partial P}{\partial z}=-\frac{mP}{kT}g \tag{16.1.95}$$

We can integrate this relationship if we know T as a function of P or z. The direct measurement of P in practice is easier (single altimeters are actually barometers if they don't use GPS).

We can then separate the variables:

$$kT\frac{\mathrm{d}P}{P}=-mg\mathrm{d}z \tag{16.1.96}$$

By integrating between two levels a and b:

$$k\int\limits_a^b T\mathrm{d}(\ln(P))=\int\limits_a^b mg\mathrm{d}z \tag{16.1.97}$$

Since:

$$\frac{\mathrm{d}(\ln(P))}{\mathrm{d}P}=\frac{1}{P}\Rightarrow \mathrm{d}(\ln(P))=\frac{\mathrm{d}P}{P} \tag{16.1.98}$$

Then to continue we use a trick. We will define the average temperature by the relation:

$$\langle T\rangle=\frac{\int_a^b T\mathrm{d}(\ln(P))}{\int \mathrm{d}(\ln(P))} \tag{16.1.99}$$

Which then allows us to write:

$$k\int\limits_a^b T\mathrm{d}(\ln(P))=k\int\limits_a^b T\mathrm{d}(\ln(P))\frac{\int_a^b \mathrm{d}(\ln(P))}{\int_a^b \mathrm{d}(\ln(P))}=k\langle T\rangle\int\limits_a^b \mathrm{d}(\ln(P))=-\int\limits_a^b mg\mathrm{d}z \tag{16.1.100}$$

Therefore:

$$mg(z_a-z_b)=k\langle T\rangle\ln\left(\frac{P_b}{P_a}\right) \tag{16.1.101}$$

or written differently:

$$\boxed{z_a-z_b=\frac{k}{mg}\langle T\rangle\ln\left(\frac{P_b}{P_a}\right)} \tag{16.1.102}$$

or after rearrangement and change the notation for the average temperature:

$$\boxed{\frac{P_b}{P_a}=e^{\frac{mg(z_a-z_b)}{k\bar{T}}}} \tag{16.1.103}$$

Both boxed relations are each respectively named "hypsometric equation" (from the Greek "hypso" for "height").

16.1.4 Earth's equilibrium temperature

The "planetary equilibrium temperature" is as we already know (see section Astronomy page 5335) a theoretical temperature that a planet would be at when considered simply as if it were a black-body being heated only by its parent star. In this model, the presence or absence of an atmosphere (and therefore any greenhouse effect) is not considered, and one treats the theoretical black-body temperature as if it came from an idealized surface of the planet.

Other authors use different names for this concept, such as "equivalent black-body temperature of a planet", or the "effective radiation emission temperature of the planet".

If the incident solar radiation ("insolation") - or incident power radiation - on the planet at its orbital distance from the Sun is $\Phi_\odot$, the amount of energy absorbed by the planet ($\mathrm{d}\Phi/\mathrm{d}S$) will depend on its reflection coefficient (albedo ρ) and its cross-sectional area S or average radius R as:

$$P_{\text{in}}=\Phi_\odot(1-\rho)S=\Phi_\odot(1-\rho)\pi R^2 \tag{16.1.104}$$

Note that the albedo would be zero for a black-body as we have seen it in the section of Optical Geometry (see page 4351). However, in planetology, more useful results are obtained by accounting for a measured or assumed planetary albedo $\rho>0$.

The infrared power radiated by the planet as thermal radiation will depend on its emissivity and its surface area, according to the Stefan–Boltzmann equation proved in the section of Thermodynamics (see page 3706):

$$M(T)=\sigma T^4 \tag{16.1.105}$$

and as it is a power by unit surface we will multiply it by the above surface area of the planet such that:

$$P(T)=\sigma S T^4 \tag{16.1.106}$$

For a spherical planet, the surface of a perfect spherical area is well known (see section Geometric Shapes page 2966)! Therefore:

$$P(T)=\sigma 4\pi R^2 T^4 \tag{16.1.107}$$

But a planet that does not absorb all incident radiation (sometimes known as a grey body) emits less total energy than a black-body and is characterized by an emissivity, $\varepsilon<1$, therefore (by definition emissivity $+$ albedo $=1$):

$$P_{\text{out}}=\varepsilon\sigma 4\pi R^2 T^4=\varepsilon\sigma 4\pi R^2 T^4 \tag{16.1.108}$$

The equilibrium temperature is then calculated by setting $P_{\text{in}}=P_{\text{out}}$. Thus:

$$\varepsilon\sigma 4\pi R^2 T^4=\Phi_\odot(1-\rho)\pi R^2 \tag{16.1.109}$$

Therefore:

$$\boxed{T_{\text{eq}}=\left(\frac{\Phi_\odot(1-\rho)}{4\sigma\varepsilon}\right)^{1/4}} \tag{16.1.110}$$

It is interesting to note that the equilibrium temperature does not depend on the size of the planet, because both the incoming radiation and outgoing radiation depend on the area of the planet!

Earth's overall average albedo is about $\rho \cong 0.31$ and the emissivity of Earth is considered as being about $\varepsilon \cong 0.96$ and the solar constant is equal to $\Phi_\odot \cong 1,376$ [W·m^{-2}] as derived earlier above (see page 5568).

A numerical application for Earth gives $T_{\text{eq}} = 256.55$ [K] (that is $-16.2\,[°\text{C}]$).

Based on this calculation, Earth's expected average global temperature is well below the freezing point of water!

The reader can refer to the section of Cosmology where have proven at page 5568 that the energy received by the Sun on Earth's ground is in average $\bar{\Phi} = 318\,[\text{W}\cdot\text{m}^{-2}]$.

Earth's actual average global temperature is around 14 [°C]. Our planet is warmer than predicted with a pretty big difference!

Why is Earth's temperature so much warmer than what our calculations predict? The answer is because certain gases in the atmosphere trap some extra heat, warming our planet like a blanket. This extra warming is named the "greenhouse effect" (see further below). Without it, our planet would be a frozen ball of ice. Thanks to the natural greenhouse effect, Earth is comfortable place for life as we know it. However, too much of a good thing can cause problems. In recent decades, a rise in the amount of greenhouse gases has begun to warm Earth a bit too much.

Remark

For extrasolar planets the temperature of the star can be calculated from the color of the star using Planck's law (see section Thermodynamics page 3719). The calculated temperature of the star can be used with the Hertzsprung–Russell diagram (see section Astrophysics page 5540) to determine the absolute magnitude of the star, which can then be used with observational data to determine the distance to the star and finally the size of the star. Orbital simulations are used to determine what orbital parameters (including orbital distance) produce the observations seen by astronomers. Astronomers use a hypothesized albedo and can then estimate the equilibrium temperature.

A quite common question is how much heating does Earth inner core provide to the surface?

The Earth's surface emits experimentally about $503\,[\text{W}\cdot\text{m}^{-2}]$ ($398.2\,[\text{W}\cdot\text{m}^{-2}]$ as infrared radiation, $86.4\,[\text{W}\cdot\text{m}^{-2}]$ as latent heat, and $18.4\,[\text{W}\cdot\text{m}^{-2}]$ via conduction/convection), or about $260,000$ [TW] over all of the Earth's surface (see [543]). The ultimate source of almost all of this energy is the Sun.

Estimates vary on how much heat crosses the core/mantle boundary, from 4 [TW] to 17 [TW]. Even the larger value is much, much smaller than the heat emitted by the Earth's surface.

Estimates of the total heat flow from the interior of the Earth (core, mantle, crust) are much tighter, $46 \pm 3\,[\text{TW}]$ (see [287]) (cf $47 \pm 2[\text{TW}]$ as given in [128]). This is considerably more than the heat coming from the core, but it's still small compared to the Earth's total heat budget:

$$\frac{\text{Heat from interior of Earth}}{\text{Total}}=\frac{46\,[\text{TW}]}{260,000\,[\text{TW}]}=0.02\% \tag{16.1.111}$$

By the way Wikipedia gives an inner/solar ratio of 0.03%.

16.1.4.1 Greenhouse effect

The greenhouse effect is the process by which radiation from a planet's atmosphere warms the planet's surface to a temperature above what it would be without its atmosphere.

If a planet's atmosphere contains radiately active gases (i.e., greenhouse gases) the atmosphere will radiate energy in all directions. Part of this radiation is directed towards the surface, warming it. The downward component of this radiation – that is, the strength of the greenhouse effect – will depend on the atmosphere's temperature and on the amount of greenhouse gases that the atmosphere contains.

Earth's natural greenhouse effect is critical to supporting life. Human activities, primarily the burning of fossil fuels and clearing of forests, have intensified the natural greenhouse effect, causing global warming.

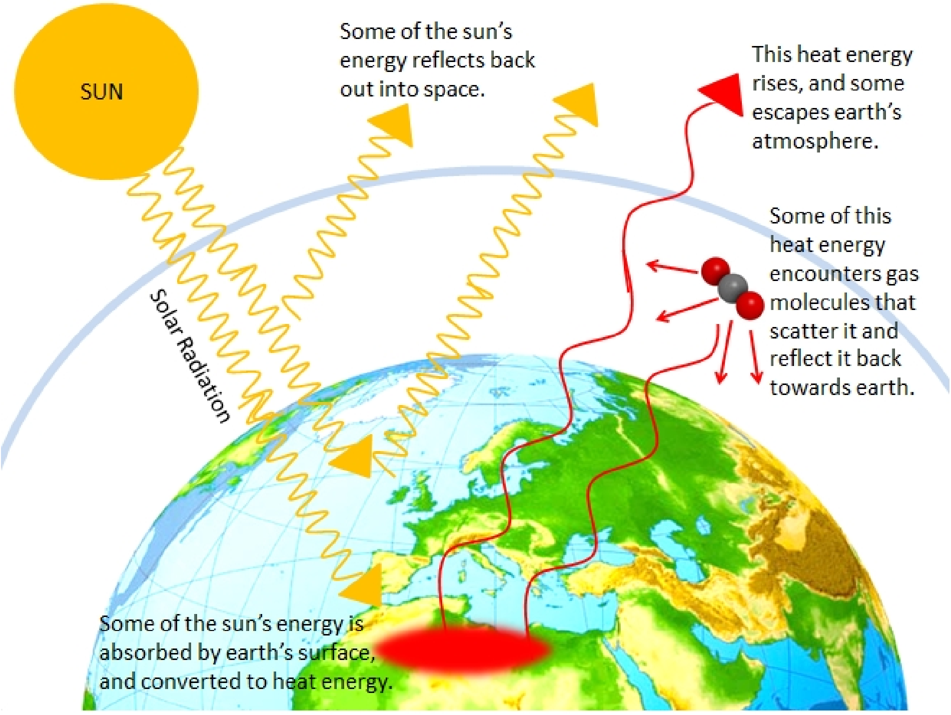

SUN
Some of the sun's energy reflects back out into space.
This heat energy rises, and some escapes earth's atmosphere.
Some of this heat energy encounters gas molecules that scatter it and reflect it back towards earth.
Solar Radiation
Some of the sun's energy is absorbed by earth's surface, and converted to heat energy.

Figure 16.17: **Greenhouse effect concept**

For a mathematical analysis of the greenhouse effect we must improve our above model. Consider for this that we denote T_S the temperature of surface of the Earth (or any other planet) and T_A the temperature of its atmosphere. The thermodynamic equilibrium of atmosphere gives us in the point of view of the radiation using again the Boltzmann law and considering that Earth's atmosphere radiates 50% back in space and reflects back 50% back to Earth surface (last measurements in the early 121st century - holocene calendar - gives 40% reflected back!):

$$\varepsilon\sigma T_S^4 = 2\varepsilon\sigma T_A^4 \tag{16.1.112}$$

Therefore it is immediate that:

$$T_A = \frac{T_S}{2^{1/4}} \tag{16.1.113}$$

Remember now the equilibrium proved earlier:

$$\varepsilon\sigma 4\pi R^2 T^4 = \Psi \odot (1-\rho)\pi R^2 \tag{16.1.114}$$

That simplifies to:

$$\varepsilon\sigma 4 T^4 = \Phi_\odot (1-\rho) \tag{16.1.115}$$

Using our new notation:

$$\varepsilon_S \sigma 4 T_S^4 = \Phi_\odot (1-\rho) \tag{16.1.116}$$

Therefore after rearranging:

$$\varepsilon_S \sigma T_S^4 = \frac{\Phi_\odot}{4}(1-\rho) \tag{16.1.117}$$

Now we add to this the back-radiation coefficient ε_A due to atmosphere:

$$\varepsilon_S \sigma T_S^4 = \frac{\Phi_\odot}{4}(1-\rho) + \varepsilon_A \sigma T_A^4 \tag{16.1.118}$$

That is:

$$\varepsilon_S \sigma T_S^4 = \frac{\Phi_\odot}{4}(1-\rho) + \varepsilon_A \sigma \left(\frac{T_S}{2^{1/4}}\right)^4 = \frac{\Phi_\odot}{4}(1-\rho) + \frac{\varepsilon_A}{2}\sigma T_S^4 \tag{16.1.119}$$

Therefore after rearrangement:

$$\boxed{T_S = \left(\frac{1}{\sigma}\frac{\Phi_\odot(1-\rho)}{4}\frac{1}{\varepsilon_S - \dfrac{\varepsilon_A}{2}}\right)^{1/4}} \tag{16.1.120}$$

A numerical application gives $T_S \cong 297$ [K] that is 14 [°C]. This is quite more accurate than the previous model!

16.1.4.2 Milankovitch cycles

The Milankovitch cycles are periodic or quasiperiodic changes in the parameters of the Earth's orbit and tilt, which in turn affect the climate. The three major types of Milankovitch cycle are:

- changes in the eccentricity of the Earth's orbit
- changes in the obliquity, or tilt of the Earth's axis
- precession, meaning changes in the direction of the Earth's axis relative to the fixed stars

These changes do not affect the overall annual amount of solar radiation hitting the Earth, but they affect the strength of the seasons in different ways at different latitudes. It is widely believed that they are partially responsible for the glacial cycles. However, the details of how this occurs are complex and poorly understood.

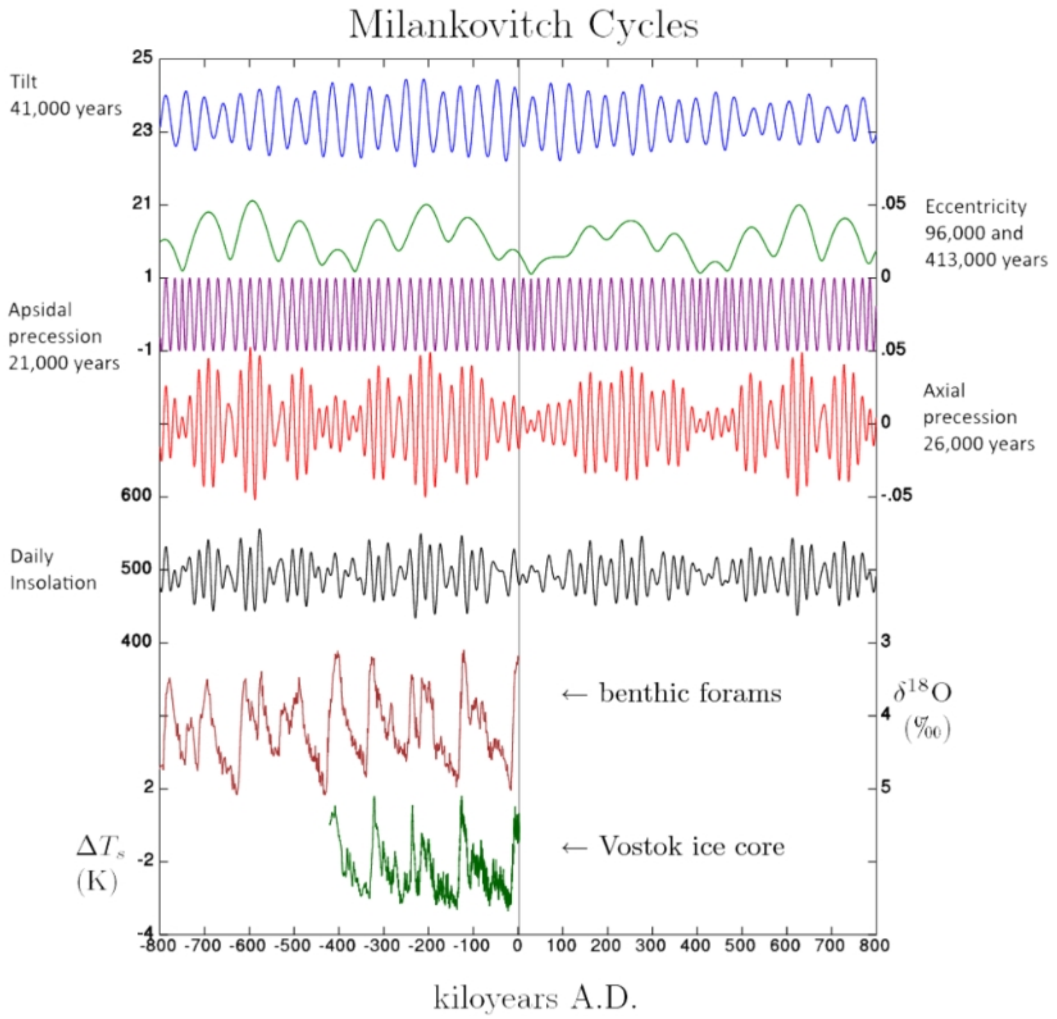


Figure 16.18: Past and future Milankovitch cycles (source: Wikipedia)

Remark

By the mid-seventies the Earth's orbital parameters were known over the last million years to a good accuracy thanks geological records. A decisive step was made by André Berger (11978 according to holocene calendar), who expressed in an analytical form the Fourier decomposition of the Earth's orbital parameters relevant for the astronomical theory of palaeoclimates. This work constitutes the first demonstration that the spectrum of climatic precession is dominated by periods of $19,000$, $22,000$ and $24,000$ years, that of obliquity, by a period $41,000$ years, and eccentricity has periods of $400,000$ years, $125,000$ years and $96,000$ years.

For information, the rotation of Earth's on itself slows down at the day we write these lines of approximately 2.3 milliseconds per century. That means 1 hour every 180 millions years. Each year the Earth's orbit speed around the Sun (actually a G2V-type star, aka yellow dwarf) also slows down of approximately 3 nanometers-per-second. Using Kepler's law, that means that the Earth goes further away from the Sun of 3 micrometers per 1 billion year......!!!!

16.1.5 Temperature variation range models

We will see here two very naive models, but still quite accurate, to compute with a good accuracy the temperature range that can be observed during days in average and identically for seasons in average around the World!

16.1.5.1 Diurnal temperature range (DTR)

In meteorology, "diurnal temperature variation" is the variation between a high air temperature and a low temperature that occurs during the same day.

We have derived earlier above (see page 5568) the Solar constant given as:

$$\begin{aligned}\Phi_{\odot}&=\frac{P_{\odot}}{4\pi d^2}=\frac{S_{\odot}M_{\odot}(T)}{4\pi d^2}=\frac{M_{\odot}(T)R_{\odot}}{d^2}=\frac{\sigma T^4R_{\odot}}{d^2}=\frac{\frac{2\pi^5k^4}{15c^2h^3}T^4R_{\odot}}{d^2}\\&=\frac{\frac{2\pi^5k^4}{15c^2h^3}\left(\frac{2{,}897.8\cdot 10^{-9}}{\lambda_{\max}}\right)^4R_{\odot}}{d^2}\cong 1,382\;[\mathrm{W}\cdot\mathrm{m}^{-2}]\end{aligned}\tag{16.1.121}$$

Based on calorimetry (see section Thermodynamics page 3622) we know the famous relation:

$$Q=mC_s\Delta T\Rightarrow\Delta T=\frac{Q}{mC_s}\tag{16.1.122}$$

where Q is the required energy (in Joules), m is the mass (in kilograms), C_s is the heat capacity of the substance and ΔT is the change in temperature (in Kelvin).

If we take an average atmospheric pressure of at ground level $1\cdot 10^5$ [P], it is equivalent by definition to a force of $1\cdot 10^5\;[\mathrm{N}\cdot\mathrm{m}^{-2}]$. We multiply by the area of the planet in square meters, we get the total weight of the atmosphere:

$$(1\cdot 10^5)\cdot(5.1\cdot 10^{14})=5.1\cdot 10^{19}\;[\mathrm{N}]\tag{16.1.123}$$

Hence the total approximated mass of the entire atmosphere is estimated to be $5.2\cdot 10^{18}$ [kg] (atmosphere of Earth) and the mean heat capacity of air is $700\;[\mathrm{J}\cdot\mathrm{kg}\cdot\mathrm{K}^{-1}]$ at normal temperature and pressure (heat capacity).

We divide by the average acceleration of gravity to convert this weight to a mass:

$$\frac{5.1\cdot 10^{19}}{9.8}=5.2\cdot 10^{18}\;[\mathrm{kg}]\tag{16.1.124}$$

Let us assume the simple ideal very naive and special case where only half of Earth is illuminated by the Sun and assuming that the whole surface is perpendicular to the Sun radiation at that the Sun appears instantaneously from below the horizon to the noon and stays there during twelve hours and that the air is static (no air circulation).

The mass to take into account for the half of the Earth is given by approximately $2.57\cdot 10^{18}$ [kg].

Therefore:

$$\Delta T=\frac{Q}{mC_s}=\frac{1,382\cdot {}^1\!/\!_2 4\pi R^2\cdot(12\cdot 60\cdot 60)}{(2.57\cdot 10^{18})\cdot 700}\cong 8.46\cdot 10^{-9}\;[\mathrm{K}]\tag{16.1.125}$$

So hummm... That's not bad!

Let's see if we can do better because the pressure assumption is maybe to relaxed. Giving gravitometry measurements the average high of Earth relatively of ground level is 800 meters. So what is the pressure at this height? Using the hypsometric relation derived earlier above (see page 8271) and taking the average temperature measured as being 298 [K]:

$$\bar{P}_{\text{atm.}}=P_0e^{\frac{-gM(h-h_0)}{RT}}\cong 90,951\,[\text{Pa}] \tag{16.1.126}$$

We then have using this new value of pressure:

$$\Delta T=\frac{Q}{mC_s}\cong 9.19\,[\text{K}] \tag{16.1.127}$$

So that's more realistic!

Maybe in the future we will develop a more realistic model but that one gives already at ΔT that is satisfying in comparison of our actual observations.

So the final model is written:

$$\Delta T=\frac{Q}{mC_s}=\frac{\Phi\cdot {}^1\!/\!{}_2S_{\text{♁}}\cdot(12\cdot 60\cdot 60)}{\dfrac{\bar{P}_{\text{atm.}}\cdot S_{\text{♁}}}{g}C_s}=\frac{g\cdot\Phi\cdot(12\cdot 60\cdot 60)}{2\bar{P}_{\text{atm.}}C_s} \tag{16.1.128}$$

Or fully explicitly:

$$\boxed{\Delta T_{\text{diurnal}}=\frac{g\cdot\dfrac{\frac{2\pi^5k^4}{15c^2h^3}\left(\frac{2{,}897.8\cdot 10^{-9}}{\lambda_{\text{max}}}\right)^4R_{\odot}}{d^2}\cdot(12\cdot 60\cdot 60)}{2P_0e^{\frac{-gM(h-h_0)}{RT}}C_s}} \tag{16.1.129}$$

As solar energy strikes the Earth's surface each morning, a shallow 1 to 3 centimetre layer of air directly above the ground is heated by conduction. Heat exchange between this shallow layer of warm air and the cooler air above is very inefficient. On a warm summer's day, for example, air temperatures may vary by 16.5 [°C] from just above the ground to chest height. Incoming solar radiation exceeds outgoing heat energy for many hours after noon and equilibrium is usually reached from 3–5 p.m., but this may be affected by a variety of factors such as large bodies of water, soil type and cover, wind, cloud cover/water vapor, and moisture on the ground.

High desert regions typically have the greatest diurnal-temperature variations, while low-lying humid areas near the shores (tropical, oceanic, and arctic) typically have the least. This explains why an area like the Pinnacles National Park can have high temperatures of 38 [°C] during a summer day, and then have lows of 5–10 [°C]. At the same time, Washington D.C., which is much more humid, has temperature variations of only 8 [°C] urban Hong Kong has a diurnal temperature range of little more than 4 [°C].

While the National Park Service claimed that the world single-day record is a variation of 56.7 [°C] (from 7.8 [°C] to -48.9 [°C]) in Browning, Montana in 11916 (holocene calendar), the Montana Department of Environmental Quality claimed that Loma, Montana also had a variation of (56.7 [°C]) (from -47.8 [°C] to 8.9 [°C]) in 11972 (holocene calendar). Both these extreme daily temperature changes were the result of sharp air-mass changes within a single day. The 11916 (holocene calendar) event was an extreme temperature drop, resulting from frigid Arctic air from Canada invading northern Montana, displacing a much warmer air mass. The 11972 (holocene calendar) event was a chinook event, where air from the Pacific Ocean overtopped mountain ranges to the west, and dramatically warmed in its descent into Montana, displacing frigid Arctic air and causing a drastic temperature rise.

In Europe, due to its more northern latitude and close proximity to large warm water bodies (such as the Mediterranean), differences in daily temperature are not as pronounced as in other continents. However, places in Southern Europe significantly far from the Mediterranean tend to have high differences in daily temperatures, some around 14 [°C]. These include Southwestern Iberia (e.g. Alvega or Badajoz) or the high-altitude plateaus of Turkey (if considered part of Europe) (e.g. Kayseri).

The map below depicts the estimated diurnal temperature range over land from 11951 to 11980 (holocene calendar), based on the interpolation of ground level weather station data by Berkeley Earth. Diurnal temperature range refers to the difference between the warmest and coldest temperature recorded during each day. In most locations this is dominated the day-night cycle of heating and cooling, though in some locations weather variations also play a large role (image created by Robert A. Rohde for Berkeley Earth. Released under Creative Commons by Robert Rohde in his professional capacity as Lead Scientist for Berkeley Earth):

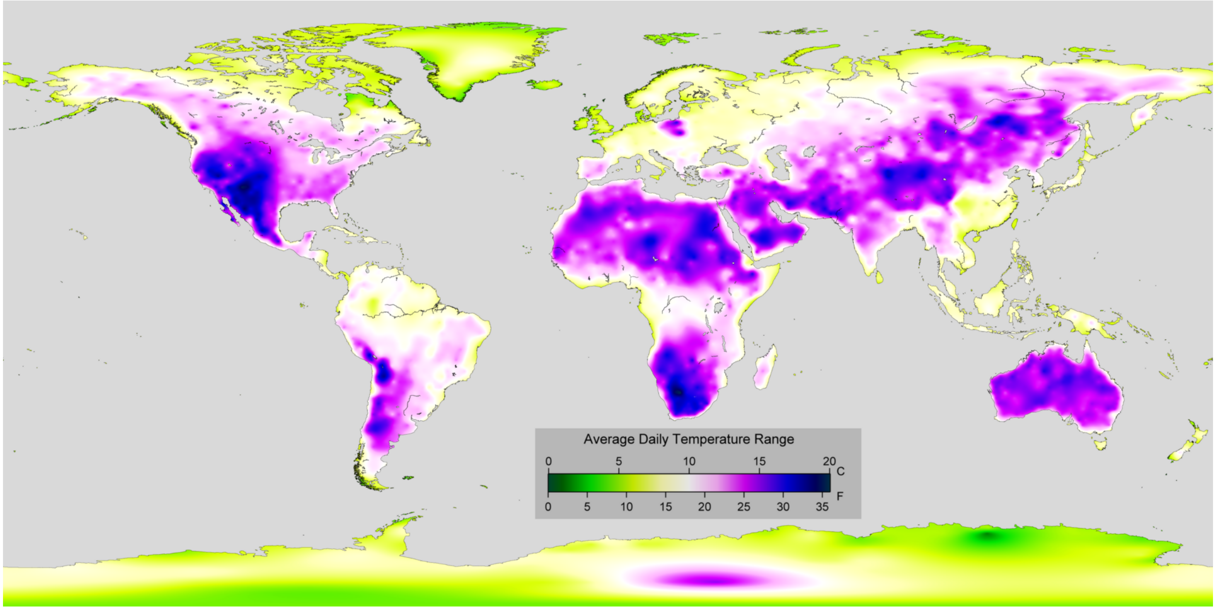


Figure 16.19: **Map of global diurnal temperature range over land from 11951 to 11980 (source: Wikipedia)**

For another more elaborated model see [212] and [213].

16.1.5.2 Seasonal temperature range (STR)

A "season" is a division of the year based on changes in weather, ecology, and the number of daylight hours in a given region. On Earth, seasons are as we will prove it below mainly the result of the axial parallelism of Earth's tilted orbit around the Sun. In temperate and polar regions, the seasons are marked by changes in the intensity of sunlight that reaches the Earth's surface, variations of which may cause animals to undergo hibernation or to migrate, and plants to be dormant. Various cultures define the number and nature of seasons based on regional variations, and as such there are a number of both modern and historical cultures whose number of seasons varies.

The Northern Hemisphere experiences most direct sunlight during May, June, and July, as the hemisphere faces the Sun. For the Southern Hemisphere it is instead in November, December, and January. It is Earth's axial tilt that causes the Sun to be higher in the sky during the summer months, which increases the solar flux. However, due to thermal inertia (i.e. "seasonal lag[2]"), June, July, and August are the warmest months in the Northern Hemisphere while December, January, and February are the warmest months in the Southern Hemisphere.

To start, let us remind that using the definition of emittance introduced in the section of Electromagnetism (see page 4340) and using the Planck's distribution function proved in the section of Thermodynamics (see page 3724) we have:

$$\mathrm{d}\Phi=M\mathrm{d}S\cos(\theta)=\left(\int\limits_0^{+\infty}\frac{\nu^3}{e^{\frac{h\nu}{kT}}-1}\mathrm{d}\nu\right)\mathrm{d}S\cos(\theta)=\sigma T^4\mathrm{d}S\cos(\theta) \tag{16.1.130}$$

with for reminder see section Thermodynamics page 3726):

$$\sigma\cong\frac{2\pi^5k^4}{15c^2h^3}=\frac{\pi^2k^4}{60c^2\hbar^2}=5.67051\cdot10^{-8}\left[\mathrm{W}\cdot\mathrm{m}^{-2}\cdot\mathrm{K}^4\right] \tag{16.1.131}$$

The Sun is a black body at a temperature actually of $5,800$ [K] with an actual radius of approximately $696,340$ [km]. Hence we receive from the latter on Earth when we are from a distance d of it:

$$L_{\text{Sun}}=\frac{\Phi}{4\pi d^2}=\frac{\sigma T^4 4\pi R^2\cos(\theta)}{4\pi d^2}=\frac{\sigma T^4R^2\cos(\theta)}{d^2}\cong\frac{4.468\cdot10^6\cos(\theta)}{d^2} \tag{16.1.132}$$

When the Earth is nearest to the Sun (in January) it is in perihelion (a.k.a. perigee) and is $147,000,000$ [km] from it. When it is furthest from the Sun (in July) it is in aphelion (a.k.a. apogee). It is then $153,000,000$ [km] from the Sun, hence:

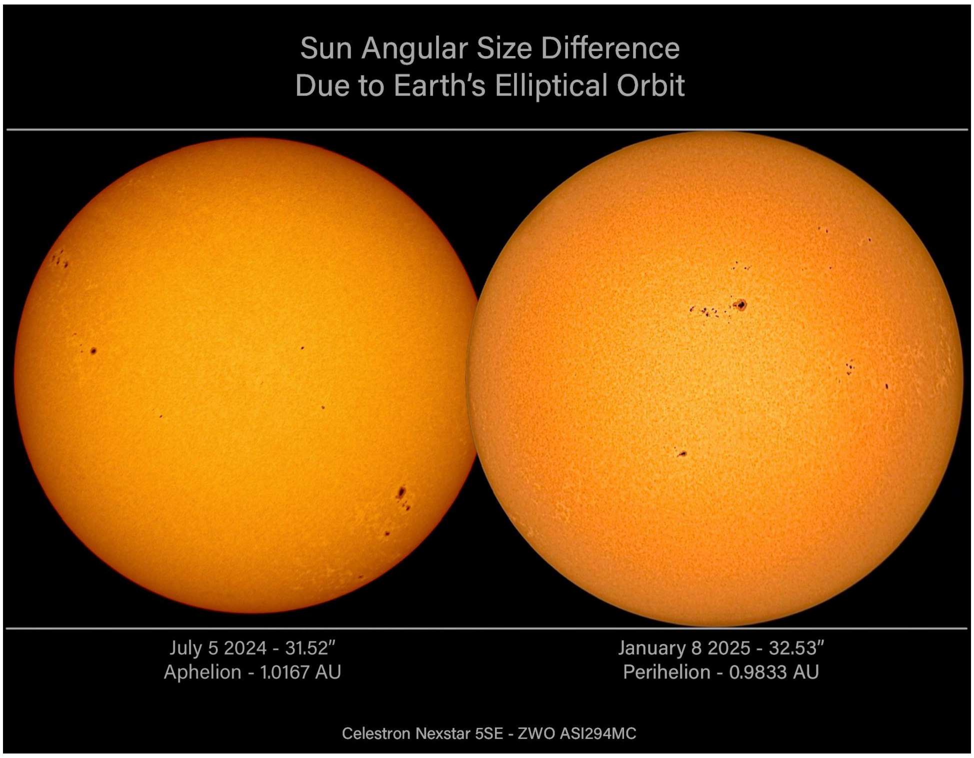
$$L_{\text{Sun,per.}}\cong\frac{4.468\cdot10^6\cos(\theta)}{d^2_{\text{per.}}} \tag{16.1.133}$$

$$L_{\text{Sun,apo.}}\cong\frac{4.468\cdot10^6\cos(\theta)}{d^2_{\text{apo.}}} \tag{16.1.134}$$

---

[2] In many locations, seasonal lag is not "seasonally symmetric"; that is, the period between the winter solstice and thermal midwinter (coldest time) is not the same as between the summer solstice and thermal midsummer (hottest time). San Francisco, for example, has an exceptionally long seasonal lag in the summer, with average daily temperatures peaking in September, and October as its second-warmest month, but very little seasonal lag in the winter, with the lowest temperatures in December and January, around and soon after the winter solstice.

Below we can see that the Sun apparent size is indeed greater in January (winter in the northern atmosphere) than in July (summer in the northern hemisphere). The both photos were taken with a Celestron 5SE, ZWO ASI294MC and Solar Filter (acquisition: 2-3 minutes stacked at 50% on ASIStudio, processed on Lightroom):


Sun Angular Size Difference
Due to Earth's Elliptical Orbit
July 5 2024 - 31.52"
Aphelion - 1.0167 AU
January 8 2025 - 32.53"
Perihelion - 0.9833 AU
Celestron Nexstar 5SE - ZWO ASI294MC

Hence a difference of 4% of power received from the Sun (actually a G2V-type star, aka yellow dwarf) by Earth between apogee and perigee whatever the value of $\cos(\theta)$ (as it cancels when we compute the ratio):

$$\Delta L_d = 1 - \frac{L_{\text{Sun,per.}}}{L_{\text{Sun,apo.}}} = 1 - \frac{d^2_{\text{per.}}}{d^2_{\text{apo.}}} \cong 1 - \frac{147,098,074,000^2}{152,097,701,000^2} \cong 3.3\% \tag{16.1.135}$$

However this cannot explain the season as we know that when it's Winter in northern hemisphere, it's at the same time Summer in the southern hemisphere! So let's see the difference in power received by the based influence by the existence of Earth's tilt angle. For example in summer in Switzerland at noon the Sun is at an angle of 68° and in winter at best at 22°:

$$\Delta L_\theta = 1 - \frac{L_{\text{Sun,max}}}{L_{\text{Sun,min}}} = 1 - \frac{\cos(\theta_{\text{max}})}{\cos(\theta_{\text{min}})} = 1 - \frac{\cos(68)}{\cos(22)} \cong 59\% \tag{16.1.136}$$

Hence the variation of luminosity received by the Sun on a specific location on Earth like Switzerland is much greater (at least actually and for the next million years) than the variation due to the variation of distance on the elliptic orbit of Earth.

This tilt angle influence can be illustrated locally as:

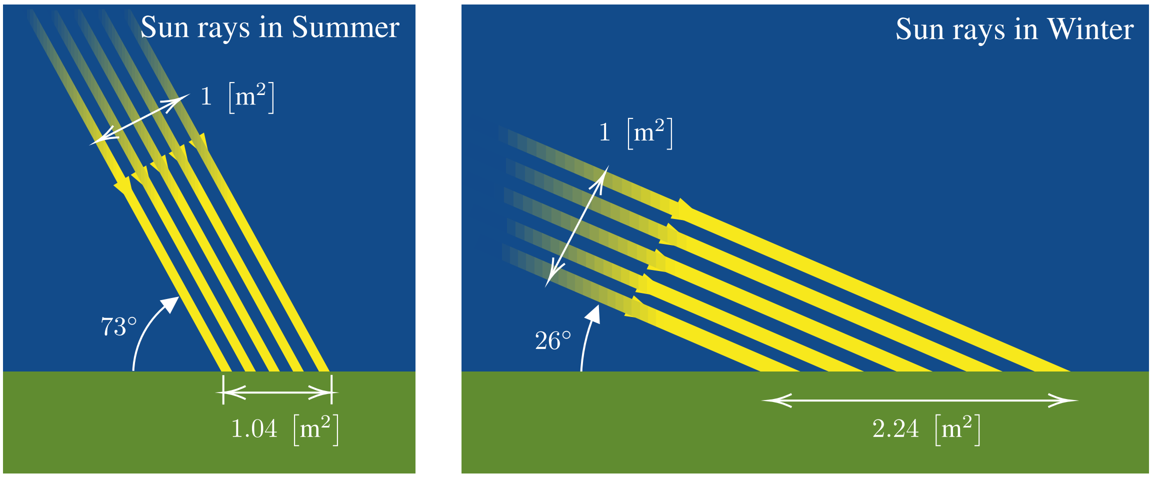
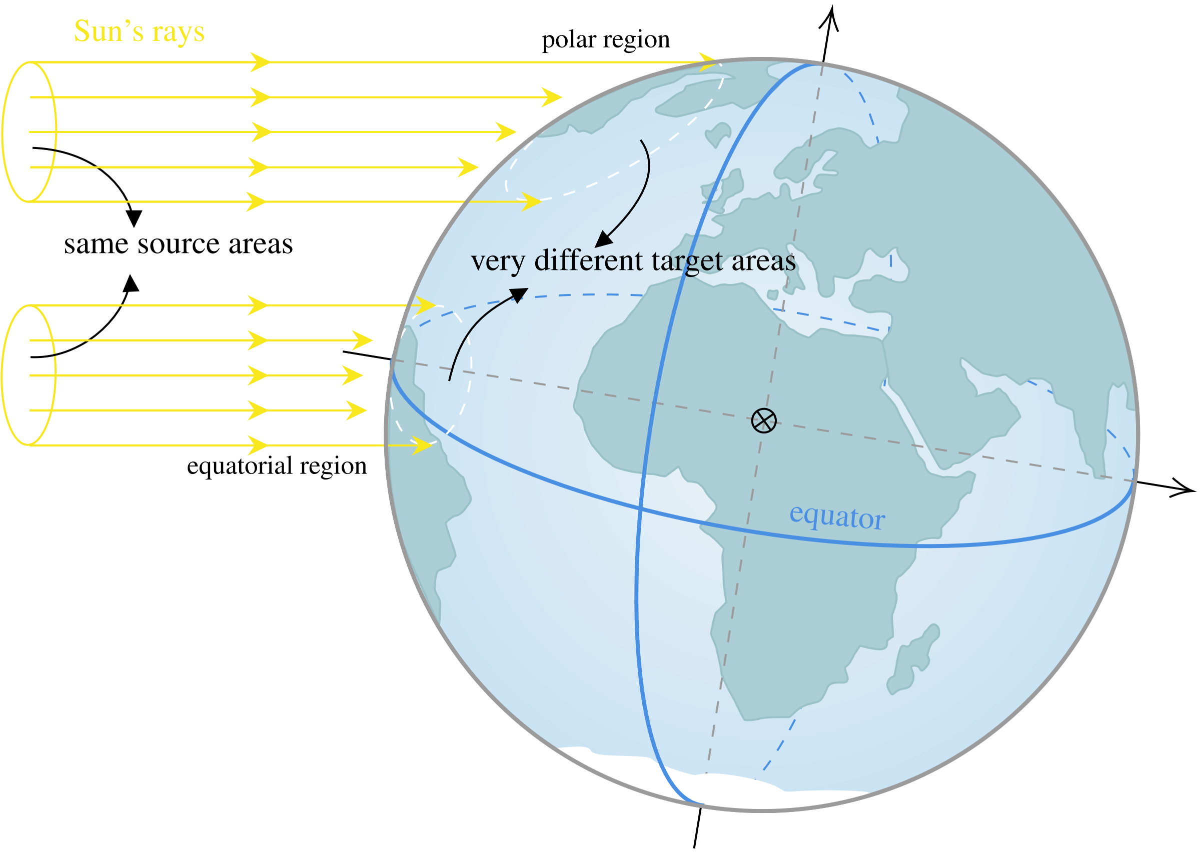

Sun rays in Summer
1 [m²]
73°
1.04 [m²]
Sun rays in Winter
1 [m²]
26°
2.24 [m²]

Figure 16.20: Sun rays in Summer vs in Winter

or more globally as following:


Sun's rays
polar region
same source areas
very different target areas
equatorial region
equator

This is checked experimentally in high school classes using a simple bulb, an Earth sphere with a thermochromic tape:

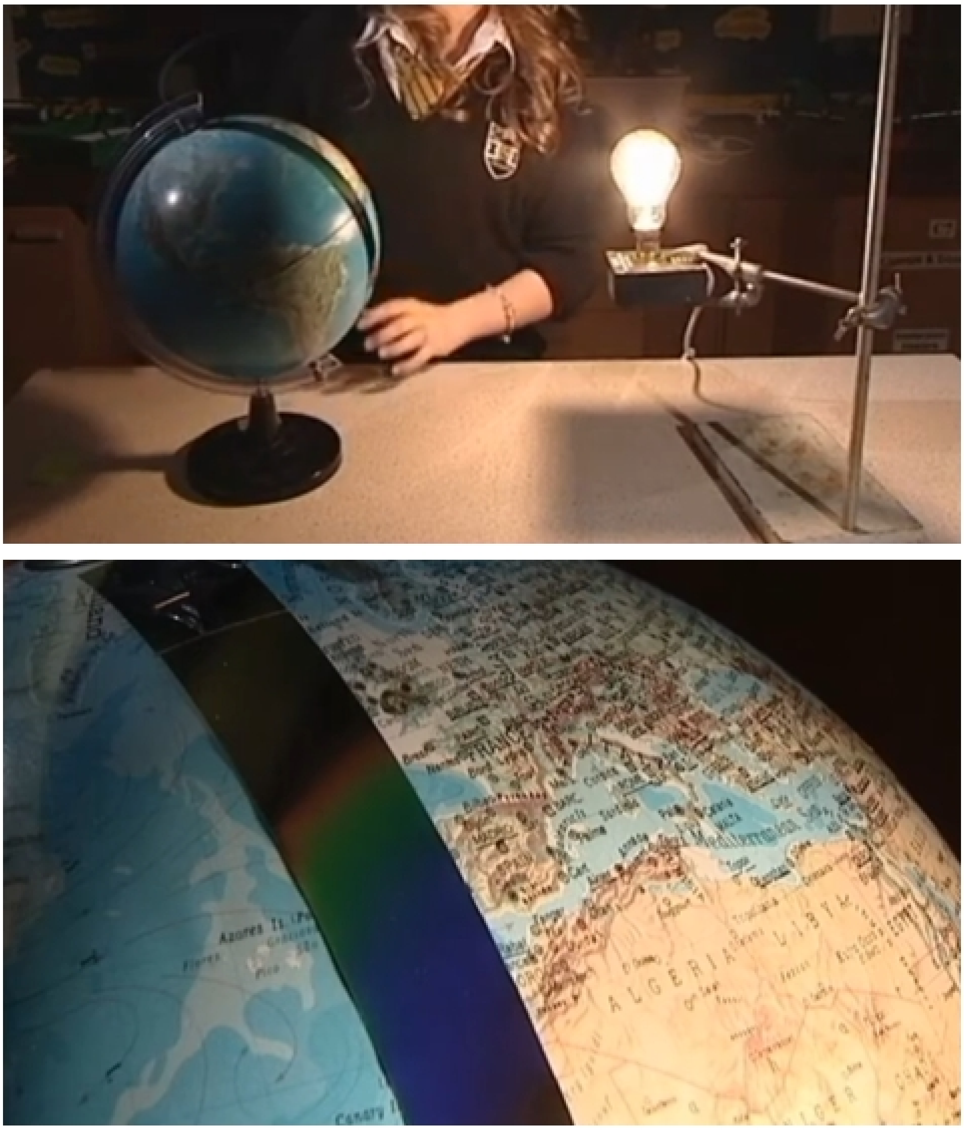


Figure 16.21: **Tilt angle influence simple experiment (source: Institute of Physics (IOP))**

Remark

If we do the same calculation by taking into account the tilde and the distance at the same time, then we get:

$$\Delta L = 1 - \frac{L_{\text{Sun,per.}}}{L_{\text{Sun,apo.}}} \cong 1 - \frac{147,098,074,000^2 \cdot \cos(69^\circ)}{152,097,701,000^2 \cdot \cos(22^\circ)} \cong 63.85\% \qquad (16.1.137)$$

The value is not really surprising!

But this is not all we have to take into account! Remember that we have proven in the section of Astronomy (see page 5372), that the duration of the day at a given latitude λ for a given tilt angle and a given date δ (remember that (we associate to the date, the angle δ travelled by the Earth on its orbit, from its position at the spring equinox) is given by the famous "Sunrise equation" :

$$\gamma = |-2\arccos\left(\tan(\lambda)\tan(\arcsin(\sin(23^\circ 27')\sin(\delta)))\right)| \qquad (16.1.138)$$

Hence always for Switzerland, using the time equation, we get that actually the shortest day length (in Winter) is approximately 8 hours and 30 minutes and the longest day length (in Summer) is approximately 16 hours. Therefore we also have a length day ratio of :

$$\begin{aligned}\Delta t_{\text{diurne}} &= 1-\frac{-2\arccos\left(\tan(27^\circ)\tan(\arcsin(\sin(23^\circ 27')\sin(270^\circ)))\right.}{-2\arccos\left(\tan(27^\circ)\tan(\arcsin(\sin(23^\circ 27')\sin(90^\circ)))\right)}\\ &= 1-\frac{\arccos\left(\tan(27^\circ)\tan(\arcsin(\sin(23^\circ 27')\sin(270^\circ)))\right.}{\arccos\left(\tan(27^\circ)\tan(\arcsin(\sin(23^\circ 27')\sin(90^\circ)))\right)} \cong 1-\frac{4.100}{2,183} \cong 46.75\%\end{aligned} \tag{16.1.139}$$

So globally because of the tilt we can say that we have in approximation a minimum decrease in comparison of the maximum at the latitude of Switzerland:

$$\Delta L_{\text{tot.}} \cong 1-(1-6.47\%)\cdot(1-59\%)\cdot(1-46.75\%) \cong 79.58\% \tag{16.1.140}$$

Hence a total decrease between Summer hottest time and Winter coldest time of 79%. And this calculation is only a lower bound. More rigorous calculations would show that in reality the decrease is greater than 79%!

Figure 16.22: **Typical visible characteristics in middle Europe of the four seasons Spring, Summer, Autumn and Winter (author: ?)**

The tropical and (to a lesser degree) subtropical regions see little annual fluctuation of sunlight and temperature due Earth's moderate 23.4 degree tilt being insufficient to appreciably effect the strength of the sun's rays annually. However, the slight differences between the solstices and the equinoxes are still enough to cause seasonal shifts to occur along a rainy, low-pressure belt named the "Intertropical Convergence Zone" (ICZ). As a result, the amount of precipitation tends to vary more dramatically than the average temperature. When the Zone is north of the Equator, the northern tropics experience their wet season while the southern tropics have their dry season. This pattern reverses when the Zone migrates to a position south of the Equator.

Over thousands of years, the Earth's axial tilt and orbital eccentricity vary (see Milankovitch cycles). The equinoxes and solstices move westward relative to the stars while the perihelion and aphelion move eastward. Thus, ten thousand years from now Earth's northern winter will occur at aphelion and northern summer at perihelion. The severity of seasonal change — the average temperature difference between summer and winter in location — will also change over time because the Earth's axial tilt fluctuates between 22.1 and 24.5 degrees. This is illustrated below:

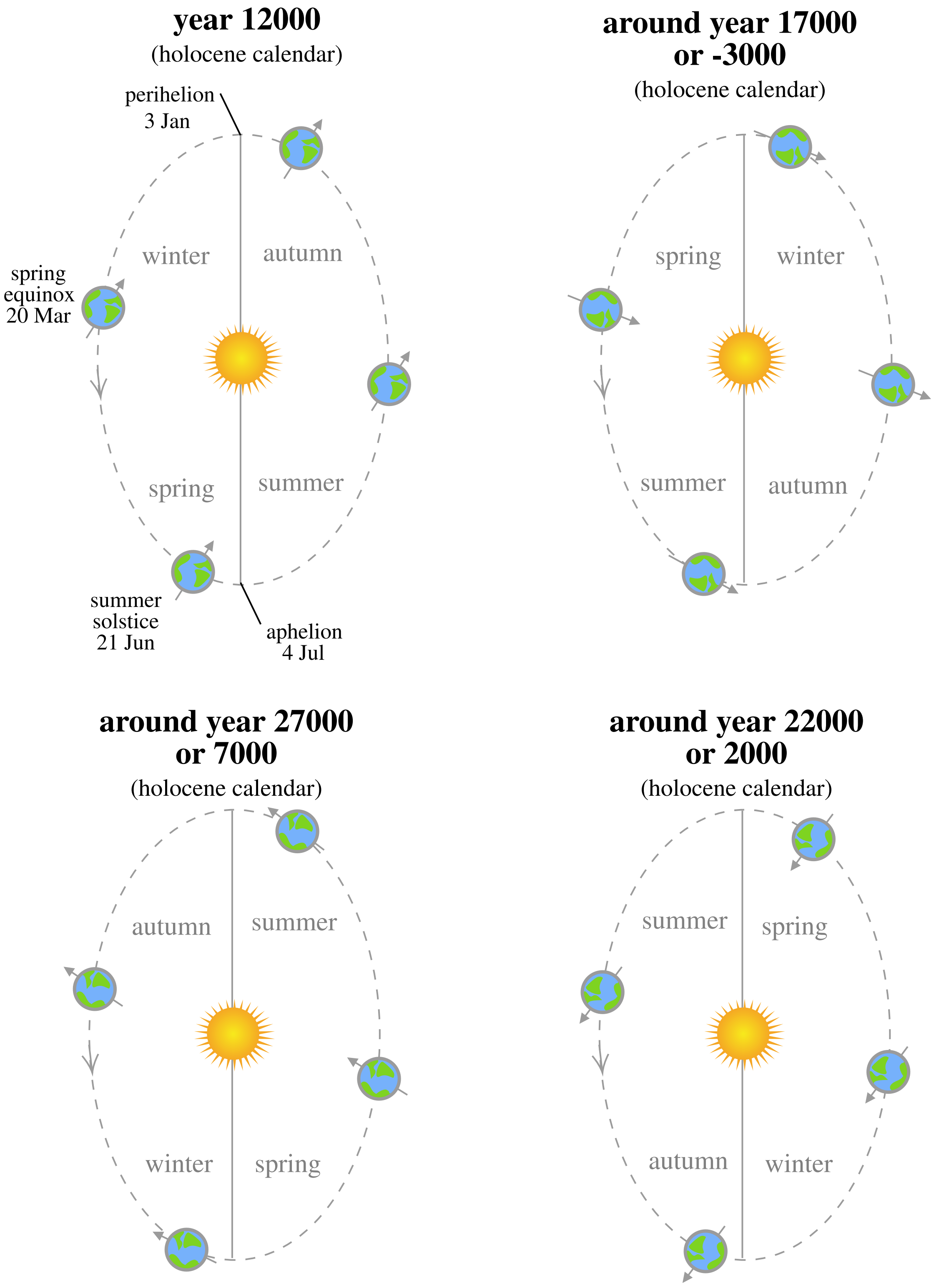

year 12000
(holocene calendar)
perihelion
3 Jan
spring
equinox
20 Mar
winter
autumn
spring
summer
summer
solstice
21 Jun
aphelion
4 Jul
around year 17000
or -3000
(holocene calendar)
spring
winter
summer
autumn
around year 27000
or 7000
(holocene calendar)
autumn
summer
winter
spring
around year 22000
or 2000
(holocene calendar)
summer
spring
autumn
winter

Figure 16.23: **Nutation of tilt axes and influence on seasons**

16.1.6 Weather (sounding) balloon

A pretty interesting example of applied mathematics to weather engineering is the study of the famous weather balloons and especially the characteristic of their volume depending on the altitude which is often subject to debate in discussion groups when nobody formalizes the problem once and for all. You will therefore understand that this is what we will discuss here and especially we will try to determine the theoretical diameter thereof at a given altitude.

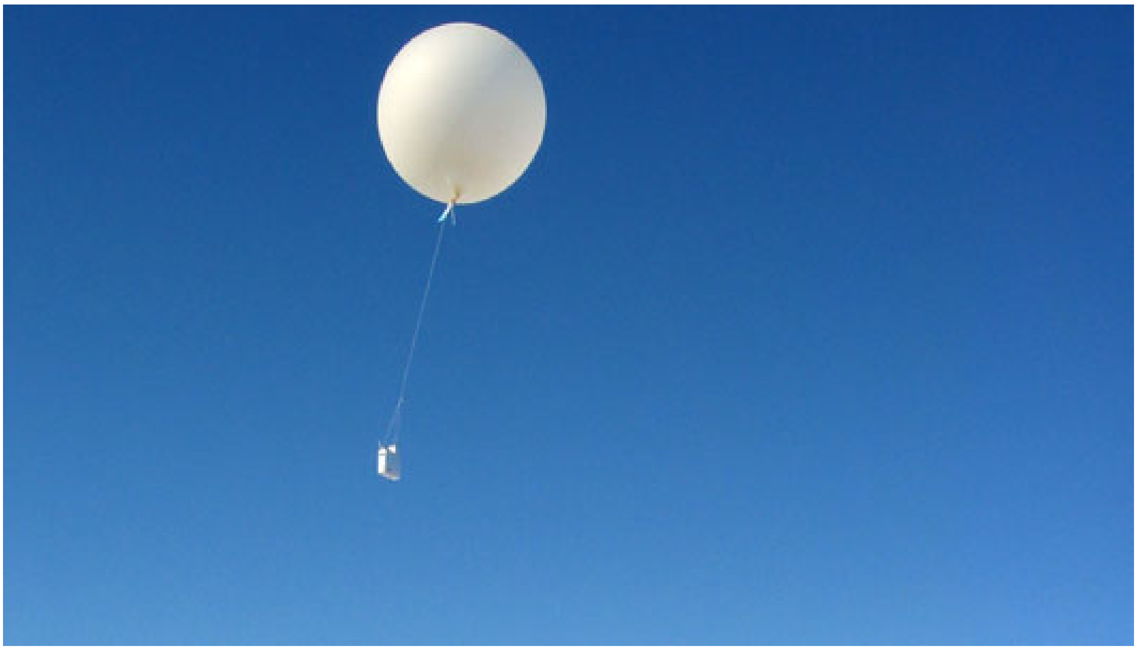


Figure 16.24: Small weather balloon

The statement of the often discussed problem is the following:

Given a PVC (Polyvinyl Chloride) weather balloon of mass m used to take at high-altitude an apparatus for performing measurements. The envelope of the balloon contains n moles of the hydrogen ideal gas thus having a molar mass:

$$M_m(\mathrm{H_2})=2\,[\mathrm{g}\cdot\mathrm{mol}^{-1}] \tag{16.1.141}$$

The atmosphere will be treated as a perfect gas, of average molecular weight:

$$M_m(\mathrm{air})\cong 29\,[\mathrm{g}\cdot\mathrm{mol}^{-1}] \tag{16.1.142}$$

at N.T.P. (Normal conditions of Temperature and Pressure).

We first we want to find what is the ascension force experienced by the balloon?

Then we want to determine the minimum amount of material n_0 providing the take off of the balloon for a given total mass (including the balloon itself!) of 2.6 [kg]. then the volume V_0 corresponding to the starting altitude.

Let us recall two things first to solve this first point:

1. Any body immersed in a liquid (or gas) undergoes an upward force equal to the weight of the volume it displaces (Archimedes force) according to the relation proved in the section Continuum Mechanics:

$$F_A = \rho g V = F_z \tag{16.1.143}$$

2. An ideal gas with a mass in grams equal to the molar mass occupies according to the ideal gas law a volume of 22.4 [L] at 273.15 [K] and at a pressure of 1 [atm] as we have proved it in the section of Thermochemisty. Which gives at N.T.P:

$$V = nR\frac{T}{P} = 1\cdot 8.314\frac{273.15}{1.01325\cdot 10^5} \cong 22\,[\text{L}] \tag{16.1.144}$$

So for the balloon to fleet at constant height (without mounting but without falling too ...) with just sufficient quantity n_0 of hydrogen, it is necessary according to the Archimedes principle that the volume of air that it moves has a weight equal to the total mass of the balloon and the probe, thus 2.6 [kg] in our case!

So since 22 [L] air weigh about 29 grams, it is necessary that the volume that it moves to be equal to 2.6 [kg] of air. Either by a simple rule of three:

$$\frac{2.6\,[\text{kg}]}{0.029\,[\text{kg}]}22\,[\text{L}] \cong 1,972\,[\text{L}] = 1.972\,[\text{m}^3] = V_0 \tag{16.1.145}$$

So if the balloon is spherical, it gives a radius of:

$$V_0 = \frac{4}{3}\pi R_0^3 \Rightarrow = \sqrt[3]{\frac{3}{4\pi}V_0} \cong 0.78\,[\text{m}] \tag{16.1.146}$$

Thus a diameter of about 1.56 [m] at the ground. This is consistent with reality!

We still need to determine the number of moles of hydrogen. It comes immediately:

$$n_0 = \frac{2.6}{0.029} \cong 90\,[\text{moles}] \tag{16.1.147}$$

Now that we know the number of moles in the balloon , if we know the temperature and pressure at a height of $22,000$ [m] (typical altitude of a small weather balloon) it only remains to apply ideal gas law to determine the volume at this altitude then given by the relation proved in the section Continuum Mechanics:

$$V = n_0 R\frac{T}{P} \tag{16.1.148}$$

and at $22,000$ [m] above sea level, we have following the numerical tables available on the Internet:

$$T \cong 218\,[K] \qquad P \cong 4\cdot 10^3\,[\text{Pa}] \tag{16.1.149}$$

But because solar radiation is about 30% higher in this altitude and the balloon is considered as adiabatic system (without heat exchange) and does not restore the power stored in the external environment. We consider that the temperature is at least 30% higher which gives us such for values:

$$T \cong 283\,[K] \qquad P \cong 4\cdot 10^3\,[\text{Pa}] \tag{16.1.150}$$

$$V=n_0R\frac{T}{P}\cong 90\cdot 8.314\cdot\frac{283}{5.4\cdot 10^3}\cong 53\,[\text{m}^3] \tag{16.1.151}$$

We could also use (because of the adiabatic assumption) the Boyle's relation (see section Continuum Mechanics page 3890) to achieve the same result:

$$P_0V_0=PV\Rightarrow V=\frac{P_0V_0}{P} \tag{16.1.152}$$

This gives a radius of about 2.33 [m] (thus diameter of about 4.6 [m]) instead of 0.78 [m] to the ground! An increase of the diameter of about 300%. However, it is more important to focus on the increasing surface to determine the elastic stress forces on the PVC. So we have before:

$$S_0=4\pi R_0^2=4\pi 0.78^2\cong 7\,[\text{m}^2] \tag{16.1.153}$$

and after:

$$S=4\pi R^2=4\pi 2.33^2\cong 68\,[\text{m}^2] \tag{16.1.154}$$

Thus an increase of the surface of about 1,000% while a standard elastomer (including PVC as a part) is not resistant to a relative change more than 500% !!! It is then much easier to understand from the point of view of the surface, why the balloon does not withstand an increase in the diameter of about 300%.

Moreover, if we apply a little abusively Hook's law to the balloon (see section Continuum Mechanics page 3744), with PVC Young's modulus which is between (source Wikipedia):

$$[2.9\cdot 10^7,3.3\cdot 10^7]\,[\text{N}\cdot\text{m}^2] \tag{16.1.155}$$

We have:

$$\frac{F}{S}=E\varepsilon=E\frac{|\Delta R|}{R_0}\cong 3.1\cdot 10^7\frac{2.12-0.73}{0.73}=5.9\cdot 10^7\,[\text{N}\cdot\text{m}^{-2}]=59\,[\text{MPa}] \tag{16.1.156}$$

Which complies with the numeral data tables which gives an elastic lower limit value of 50 [MPa] and an upper limit of 80 [MPa] for PVC (source Wikipedia). We can then calculate the minimum and maximum theoretical height that the balloon can reach.

So for the minimum height, we will write first:

$$|\Delta R|=\frac{\dfrac{F}{S}}{E}R_0=\frac{5\cdot 10^7}{3.1\cdot 10^7}0.73=1.17\,[\text{m}] \tag{16.1.157}$$

Which then corresponds to a final radius of:

$$R=R_0+1.38=1.9\,[\text{m}] \tag{16.1.158}$$

Which corresponds to a volume of:

$$V=\frac{4}{3}\pi R^3=28.7\,[\text{m}]^3 \tag{16.1.159}$$

Applying Boyle-Mariotte:

$$P=\frac{P_0V_0}{V}=\frac{1.01325\cdot 10^5\cdot 2.151}{28.7}=7.5\cdot 10^3\,[\text{Pa}] \tag{16.1.160}$$

which is a pressure corresponding to a height of about $18,000$ [m] according to the experimental measurements (www.engineeringtoolbox.com) and this is indeed a rare height at which PVC balloons blow up.

Now let us do the same with the maximum height:

$$|\Delta R| = \frac{\frac{F}{S}}{E} R_0 = \frac{8 \cdot 10^7}{3.1 \cdot 10^7} 0.73 = 1.88\,[\text{m}] \tag{16.1.161}$$

Which then corresponds to a final radius of:

$$R = R_0 + 1.88 = 2.61\,[\text{m}] \tag{16.1.162}$$

Which corresponds to a volume of:

$$V = \frac{4}{3}\pi R^3 = 74.5\,[\text{m}]^3 \tag{16.1.163}$$

Applying Boyle-Martiotte:

$$P = \frac{P_0 V_0}{V} = \frac{1.01325 \cdot 10^5 \cdot 2.151}{74.5} = 2.9 \cdot 10^3\,[\text{Pa}] \tag{16.1.164}$$

which is a pressure corresponding to a height of about $24,000$ [m] according to the experimental measurements and this is indeed a height at which the highest PVC balloons burst.

16.1.7 Cyclogenesis and Anticyclogenesis

Cyclogenesis is the development or strengthening of cyclonic circulation in the atmosphere (a low-pressure area).

Most of the atmospheric mass is contained within the first 20 kilometres in altitude, so that the large-scale meteorology runs on a thin spherical shell (and can be assimilated to the mechanics of a two-dimensional fluid).

The circulation engine of the atmosphere in the tropics is solar heating. Because of the inclination of 23.5 degrees to the axis of rotation of the Earth, the Sun (actually a G2V-type star, aka yellow dwarf) is never more than a few tenths of a degree from the zenith at noon throughout the year in the tropics, which gives maximum warming around the geographic equator.

We must therefore distinguish air circulation in the vicinity of the tropics, characterized by strong vertical movement due to thermal convection, and circulation at mid-latitudes, made almost of horizontal movements:

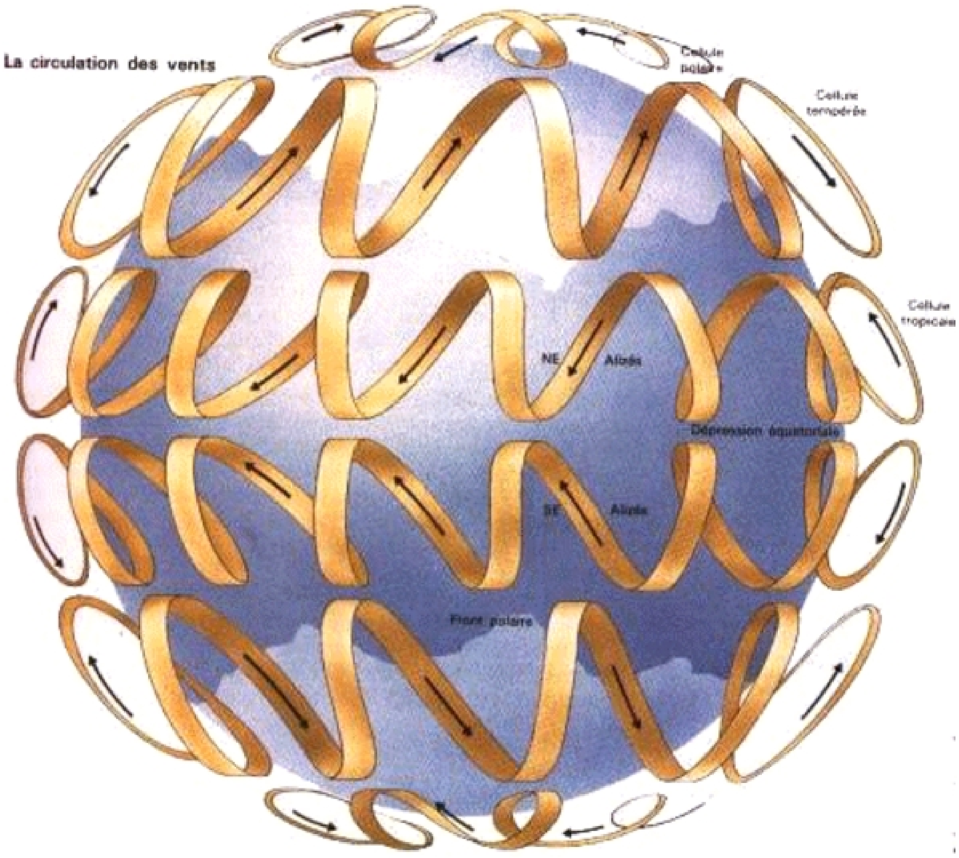

La circulation des vents
Cellule polaire
Cellule tempérée
Cellule tropicale
NE
Alizés
Dépression équatoriale
SE
Alizés
Front polaire

Figure 16.25: **Simplified schematic and idealized caricature of wind circulation on Earth (source: ?)**

Let us suppose for a moment that we completely stopped the movement of the air in the atmosphere relatively to the surface of the planet, and that we leave it then start turning from West to East (left to right on the images) from the rest position. The pressure gradient force causes the air to move from high pressure areas to low pressure areas ("vacuum call"). These convective movements are named "Hadley cells".

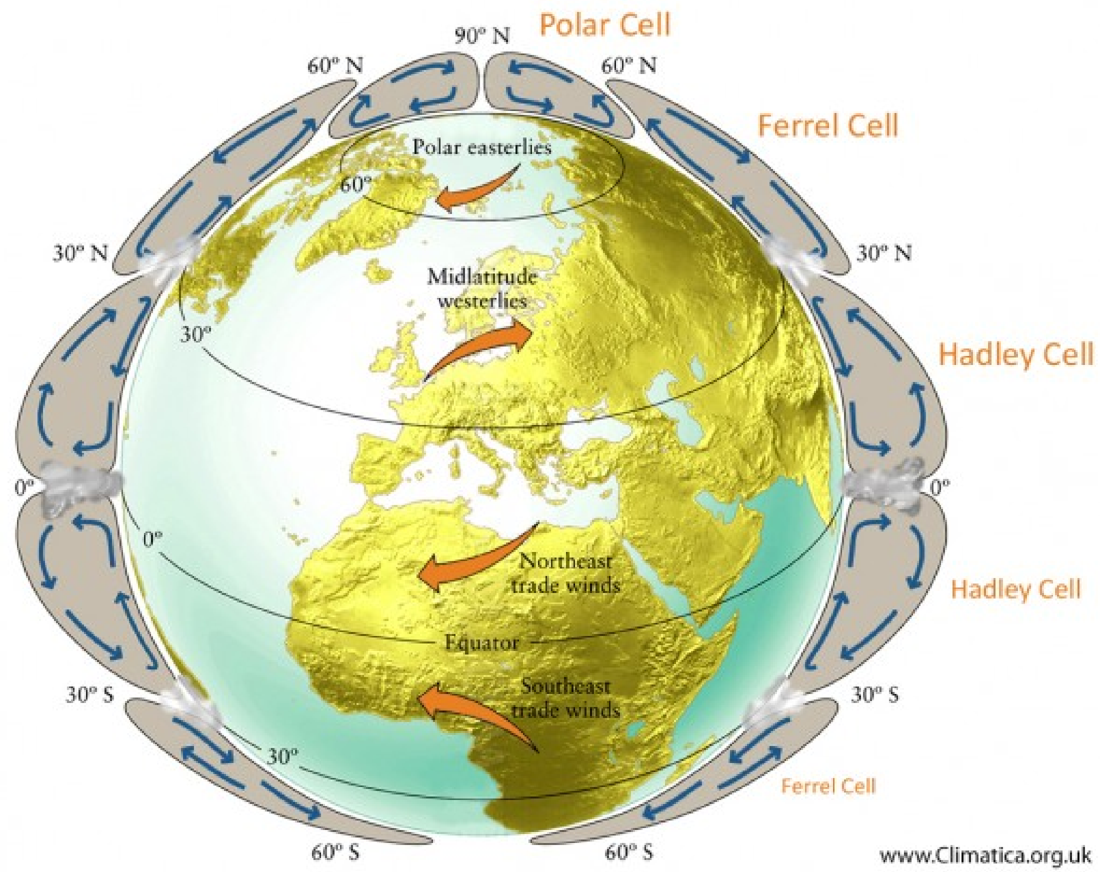

90° N
Polar Cell
60° N
60° N
Ferrel Cell
Polar easterlies
60°
30° N
30° N
Midlatitude westerlies
30°
Hadley Cell
0°
0°
0°
Northeast trade winds
Hadley Cell
Equator
30° S
30° S
Southeast trade winds
30°
Ferrel Cell
60° S
60° S
www.Climatica.org.uk

Figure 16.26: **Hadley and Ferrel cells (source:** `www.climatica.org.uk`**)**

However, when the rotation movement began the Coriolis force (due to the rotation of the Earth) deflects the North-South winds towards the West and South-North winds eastward for an observer falling in the North Pole (see section Classical Mechanics page 3388). We see therefore the formation of cyclones turning in the opposite clockwise direction in the Northern Hemisphere and vice versa in the Southern hemisphere (due to the direction of the vector $\vec{\omega}$ in this part of the hemisphere).

The higher the air velocity increases, the more the Coriolis force increases accentuating the deviation. Eventually the Coriolis force reaches a value equal and opposite to that of the force of the pressure gradient, thus producing a flow of constant velocity (no acceleration), parallel to the isobars thus defining the geometric limit of the Hadley cell. This is what we name "geostrophic balance". In practice, the flow outside the tropics is almost always in quasi-geostrophic equilibrium.

In the absence of direct wind measurement, meteorologists can estimate the wind speed at a given point by measuring on a weather map the pressure gradient and the latitude. The geostrophic approximation is purely diagnostic. It has no predictive value because its equation contains no term of change depending on time.

In the tropics, where the Coriolis force is becoming weaker until to be zero at the equator, there are other forces, such as centrifugal force, that balance the pressure gradient force.

This is what we will prove here mathematically using Continuum Mechanics (fluids) and Classical Mechanics (see corresponding sections for the prerequisites).

We know that in our system intervene pressure forces (gradient), centrifugal forces, gravity forces. Forces to which we must not forget to add the Coriolis force (implicitly its corresponding acceleration) of the system (meaning: the cyclone) with pulsation $\vec{\omega}$ (see section Classical Mechanics page 3480):

$$\vec{a}_c = 2\vec{\omega}\times\vec{v} \tag{16.1.165}$$

and the Coriolis force (implicitly: acceleration) of the fluid by unit mass (the reason for this unit choice will appear obvious few paragraphs further below) with respect to the pulsation $\vec{\Omega}$ of the Earth:

$$\vec{a}_c = 2\vec{\Omega}\times\vec{v} \tag{16.1.166}$$

Thus, as we know (see section Classical Mechanics page 3388), the Coriolis force will tend to deflect any descending movement to the right (East) in the Northern Hemisphere and any ascending movement to the left (West) in the Southern hemisphere (following we place ourselves and look in the direction of fluid movement in the previous figures).

This is why the air at the base of the Hadley cell, travelling at low altitude from the Tropic to the Equator will be deflected to the West to give the east trade winds.

We have also proved in the section of Continuum Mechanics a particular form of the Euler equation of the second form (see page 3831) that was:

$$\partial_t\vec{v}+\vec{\nabla}\left(\frac{v^2}{2}+\frac{p}{\rho}+U\right)+2\vec{\omega}\times\vec{v}=\vec{0} \tag{16.1.167}$$

Rearranging this relation can also be written:

$$\vec{\nabla}p=\rho\left(-\vec{\nabla}-\left(\partial_t\vec{v}+\vec{\nabla}\frac{v^2}{2}+2\vec{\omega}\times\vec{v}\right)\right) \tag{16.1.168}$$

But we had also proved that:

$$-\vec{\nabla}U=\vec{g} \tag{16.1.169}$$

It comes in the Earth reference frame:

$$-\vec{\nabla}p=\rho\left(\vec{g}-\left(\partial_t\vec{v}+\vec{\nabla}\frac{v^2}{2}+2\vec{\omega}\times\vec{v}\right)\right) \tag{16.1.170}$$

It is therefore of the equation defining the pressure within the fluid considered as isolated. In this relation, we must still subtract the Coriolis pressure forces due to the geocentric reference frame to get the system dynamic of the "cyclone":

$$\vec{F}_C=\rho 2\vec{\Omega}\times\vec{v} \tag{16.1.171}$$

Which finally gives:

$$-\vec{\nabla}p=\rho\left(\vec{g}-\left(\partial_t\vec{v}+\vec{\nabla}\frac{v^2}{2}+2\vec{\omega}\times\vec{v}+2\vec{\Omega}\times\vec{v}\right)\right) \tag{16.1.172}$$

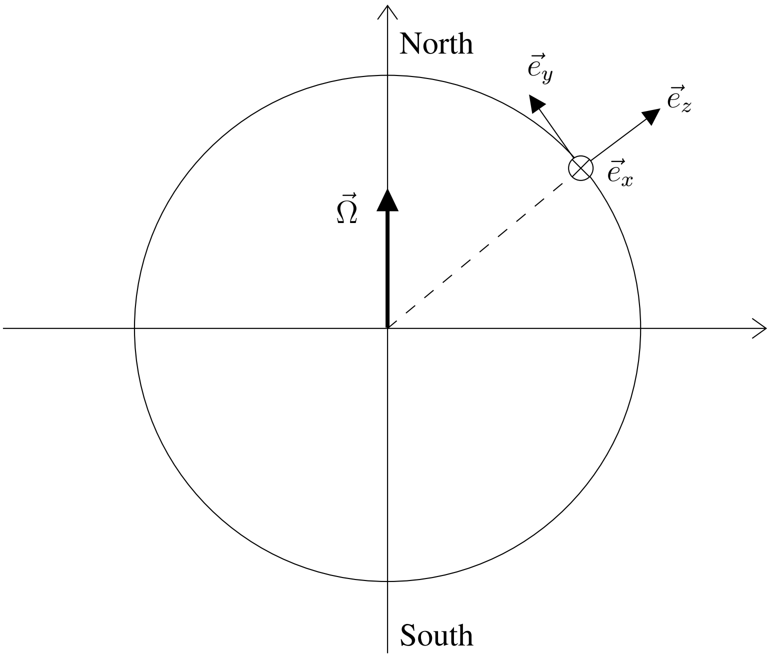

North
$\vec{e}_y$
$\vec{e}_z$
$\vec{e}_x$
$\vec{\Omega}$
South

Figure 16.27: **North-South slice of Earth for cyclogenesis study**

If we zoom on the reference frame related to the cyclone and translate to it the vector pulse of the Earth, we have:

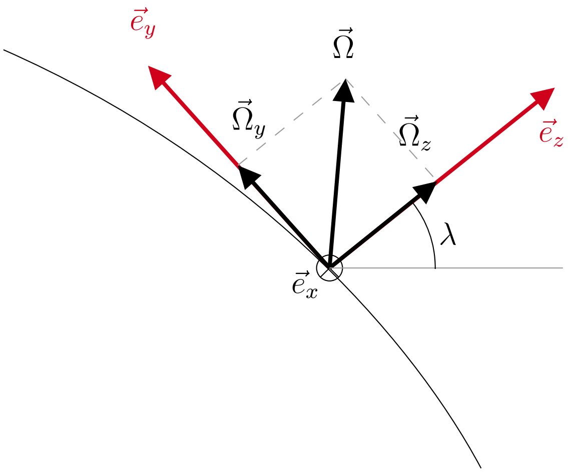

$\vec{e}_y$
$\vec{\Omega}$
$\vec{\Omega}_y$
$\vec{\Omega}_z$
$\vec{e}_z$
λ
$\vec{e}_x$

Figure 16.28: **Reference frame linked to the cyclone with the planar pulse**

Therefore:

$$\vec{\Omega}=\begin{pmatrix}\Omega_x\\ \Omega_y\\ \Omega_z\end{pmatrix}=\begin{pmatrix}0\\ \|\vec{\Omega}\|\cos(\lambda)\\ \|\vec{\Omega}\|\cos(\pi/2-\lambda)\end{pmatrix}=\begin{pmatrix}0\\ \Omega\cos(\lambda)\\ \Omega\sin(\lambda)\end{pmatrix} \tag{16.1.173}$$

Hence we have:

$$2\vec{\Omega}\times\vec{v}=2\begin{pmatrix}0\\ \Omega\cos(\lambda)\\ \Omega\sin(\lambda)\end{pmatrix}\times\begin{pmatrix}v_x\\ v_y\\ v_z\end{pmatrix}=2\Omega\begin{pmatrix}v_z\cos(\lambda)-v_y\sin(\lambda)\\ v_x\sin(\lambda)\\ v_x\sin(\lambda)\end{pmatrix} \tag{16.1.174}$$

As we study the movements (almost) horizontal in the atmosphere at this latitude, we can consider that the fluid particles are liable to remain in the horizontal plane $(\vec{e}_x,\vec{e}_y)$. The components of the Coriolis force in a planar motion are therefore $(v_z=0)$:

$$\vec{a}_C=2\vec{\Omega}\times\vec{v}=2\Omega\begin{pmatrix}v_z\cos(\lambda)-v_y\sin(\lambda)\\ v_x\sin(\lambda)\\ v_x\cos(\lambda)\end{pmatrix}=2\Omega\sin(\lambda)\begin{pmatrix}-v_y\\ v_x\\ 0\end{pmatrix}=f\begin{pmatrix}-v_y\\ v_x\\ 0\end{pmatrix} \tag{16.1.175}$$

where f is named the "Coriolis parameter". So the Coriolis force in oceanography and meteorology is traditionally denoted:

$$F_C=\rho f v \tag{16.1.176}$$

The number f, positive in the northern hemisphere, negative in the southern hemisphere, ranging from 0 to 1.458 at the poles while the force is of the order of thousandths of Newton for fluid masses (ocean currents) and of the same magnitude (because the speed compensates for the low density) for gas (air currents).

We now apply the approximation of the geostrophic equilibrium, that is to say that we consider that the air is animated by a uniform rectilinear motion (geostrophic wind), in other words, we neglect the action of centrifugal force due to the rotation of the vortex relatively that of the Coriolis force due to the rotation of the Earth, which means we assume that:

$$fv\gg R\omega^2 \tag{16.1.177}$$

with R being the radius of the vortex and ω its pulsation. Since (see section Classical Mechanics page 3280):

$$\omega=\frac{v}{R} \tag{16.1.178}$$

the latter inequality becomes:

$$\frac{v}{fR}\ll 1 \tag{16.1.179}$$

where:

$$R_0=\frac{v}{fR} \tag{16.1.180}$$

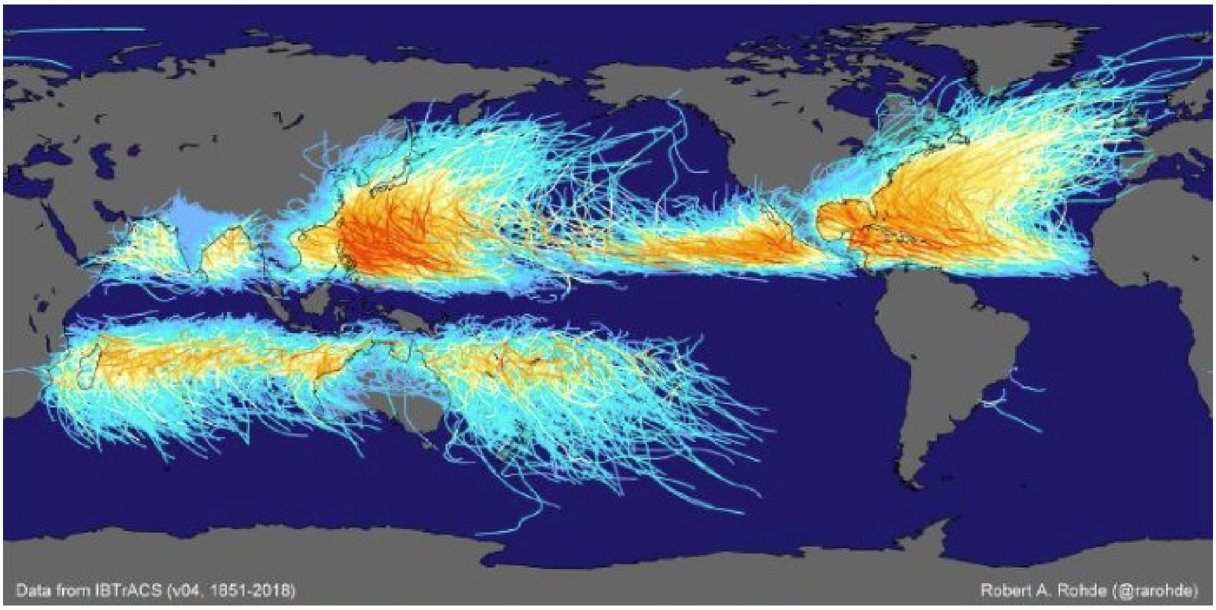
is named the "Rossby number" and has no dimensions. A small Rossby number signifies a system which is strongly affected by Coriolis forces, and a large Rossby number signifies a system in which inertial and centrifugal forces dominate. For example, in tornadoes, the Rossby number is large ($\cong 10^3$), in low-pressure systems it is low ($\cong 0.1-1$) and in oceanic systems it is of the order of unity, but depending on the phenomena can range over several orders of magnitude ($\cong 10^{-2}-10^2$). As a result, in tornadoes the Coriolis force is negligible, and balance is between pressure and centrifugal forces (named "cyclostrophic balance").

Remark

For the middle latitudes ($\lambda=45^\circ$), the experiences and measures give $f=10^{-4}\,[\text{m}\cdot\text{s}^{-1}]$ and $v=10\,[\text{m}\cdot\text{s}^{-1}]$. The limit value for which $R_0=1$ is $R=100\,[\text{km}]$. For a larger scale, that is the case with hurricanes where $R \cong 100$ [km] we are close to the geostrophic balance. For a smaller scale, the Coriolis force is negligible and the wind is accelerated from high to low pressure.

In other words the Rossby number therefore represents the ratio between the forces of inertia and the forces due to rotation that characterize the motion of a fluid in a rotating frame.

So on a static Earth cyclones would not exist. And this also why we have never observed a cyclone be formed at the equator nor a cyclone pass through the equator (hence the "cyclonic wall" expression)!


Data from IBTrACS (v04, 1851-2018)
Robert A. Rohde (@rarohde)

So we can make the difference between a geophysical flow with a large Rossby number or with a small Rossby number. If the Rossby number is greater than one, then the Coriolis forces due for example to Earth's rotation are negligible relatively to the flow inertia. Otherwise for a Rossby number very smaller than the unit, the Coriolis forces dominate the movement of the fluid.

Thus, if one approaches the equator, f tends to 0 the Rossby number becomes very large and at the poles it becomes very weak respectively.

Under this approximation, our Euler equation can be written as:

$$\vec{\nabla}p \cong \rho(\vec{g}-\vec{a}_C)=\rho(\vec{g}-2\vec{\Omega}\times\vec{v}) \qquad (16.1.181)$$

and since we are interested only to the horizontal plane of the atmosphere this simplifies even more to the form:

$$\vec{\nabla}_{x,y}p \cong -\rho 2\left(\vec{\Omega}\times\vec{v}\right)_{x,y}=-\rho f\begin{pmatrix}-v_y\\ v_x\end{pmatrix} \qquad (16.1.182)$$

Thus in fully developed vector form by taking back the capital P for pressure as it is customary in meteorology:

$$\begin{pmatrix} \dfrac{\partial P}{\partial x} \\ \dfrac{\partial P}{\partial y} \end{pmatrix} = -\rho f \begin{pmatrix} -v_y \\ v_x \end{pmatrix} \tag{16.1.183}$$

It comes then:

$$\begin{pmatrix} -v_y \\ v_x \end{pmatrix} = -\frac{1}{\rho f} \begin{pmatrix} \dfrac{\partial P}{\partial x} \\ \dfrac{\partial P}{\partial y} \end{pmatrix} = \frac{1}{\rho f} \begin{pmatrix} -\dfrac{\partial P}{\partial x} \\ -\dfrac{\partial P}{\partial y} \end{pmatrix} \tag{16.1.184}$$

Therefore in conventional form:

$$\vec{v}_{x,y} = \begin{pmatrix} v_x \\ v_y \end{pmatrix} = \begin{pmatrix} -\dfrac{\partial P}{\partial y} \\ \dfrac{\partial P}{\partial x} \end{pmatrix} = \frac{1}{\rho f} \vec{e}_z \times (\vec{\nabla} P) \tag{16.1.185}$$

The norm being given by:

$$v = \|\vec{v}_{x,y}\| = \begin{pmatrix} v_x \\ v_y \end{pmatrix} = \sqrt{\left(-\frac{1}{\rho f}\frac{\partial P}{\partial y}\right)^2 + \left(-\frac{1}{\rho f}\frac{\partial P}{\partial x}\right)^2} = \frac{1}{\rho f}\sqrt{\left(\frac{\partial P}{\partial y}\right)^2 + \left(\frac{\partial P}{\partial x}\right)^2} \tag{16.1.186}$$

Therefore:

$$\boxed{v_g = \frac{1}{\rho f}\frac{\mathrm{d}P}{\mathrm{d}R}} \tag{16.1.187}$$

which is named the "equation of geostrophic winds".

Four scenarios are to be considered (with some reminders about some concepts already introduced at the beginning of the section):

1. We are in the northern hemisphere and therefore f is positive. Let us suppose that $\mathrm{d}P/\mathrm{d}R$ is positive, then the pressure increases away from the vortex center (which latter is therefore a minimum of low pressure). Therefore v is positive and we have a vortex named "depression" in the northern hemisphere. Thus, the fluid (the wind) blows around the depression counter-clockwise (West direction) in the Northern Hemisphere.

   Definitions (#16.1.723):

   D1. A "depression" (or "low pressure") is an area where the atmospheric pressure decreases horizontally toward a center of low pressure, that is to say a local minimum pressure.

   D2. The intense weather systems circulating around a closed center of low pressure (like a vacuum that attracts clouds from which the fact such systems are visible on satellite photos) consistently receive in general the term "cyclone" or "tropical cyclone".

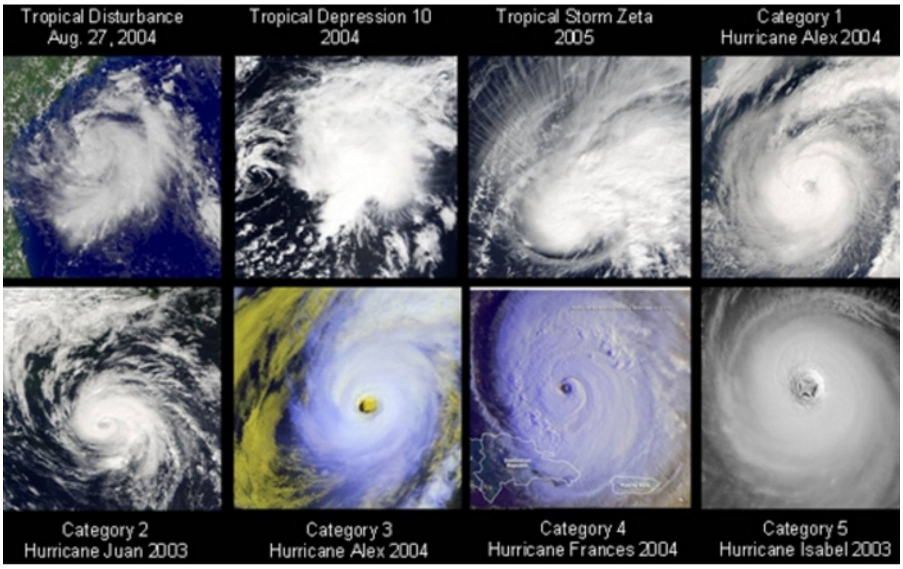

Tropical Disturbance
Aug. 27, 2004
Tropical Depression 10
2004
Tropical Storm Zeta
2005
Category 1
Hurricane Alex 2004
Category 2
Hurricane Juan 2003
Category 3
Hurricane Alex 2004
Category 4
Hurricane Frances 2004
Category 5
Hurricane Isabel 2003

Figure 16.29: **Some northern hemisphere cyclones**

Remark

On Earth we frequently associate depressions to bad weather because the dynamics surrounding depression presupposes the existence of updrafts (can hardly get into the ground so the only way to escape is up!) that cause clouds and precipitation. In addition, the pressure gradient over a depression can cause strong winds.

2. We are still in the Northern Hemisphere and therefore f is positive. Let us suppose that $\mathrm{d}P/\mathrm{d}R$ is negative this time, the pressure then decreases away from the vortex center (which is thus reaching a maximum pressure). Therefore v is negative and we have a vortex named a "high pressure" in the northern hemisphere. Thus, the fluid (wind) blows around the high pressure clockwise (to the East) in the Northern Hemisphere.

   Definitions (#16.1.724):

   D1. A "high pressure" is an area where the atmospheric pressure increases horizontally toward a center of high pressure, that is to say a local maximum pressure.

   D2. The intense weather systems to flow around a closed center of high pressure (as a fan it rejects and disperses the clouds from which the fact that the high pressure are not simply visible on satellite photos) consistently receive the name of "anticyclone".

Remark

Anticyclones usually bring good weather and clear skies. Atmospheric dynamics causes air at medium altitudes to be relatively warm and dry, and therefore without clouds.

3. We are still in the southern hemisphere and therefore f is negative. Let us suppose $\mathrm{d}P/\mathrm{d}R$ is positive, then the pressure increases away from the vortex center. Therefore v is negative and we have a vortex still named a "high pressure" (or "anticyclone") in the Southern Hemisphere. Thus, the fluid (wind) blows around the high pressure but counter-clockwise (to the West) in the Southern Hemisphere.

4. We are still in the southern hemisphere and therefore f is negative. Let us suppose that $\mathrm{d}P/\mathrm{d}R$ is negative, the pressure decreases away from the vortex center. Therefore v is positive and we have a vortex still named "low pressure" (or "Hurricane") in the Southern Hemisphere. Thus, the fluid (the wind) blows around the low pressure clockwise (to the East) in the Southern Hemisphere.

Remark

It is therefore possible to say generally on larger dimensions that the wind comes from high pressure (anticyclone) to move towards low pressure (depression).

Here is an illustration example of a depression (anticyclone) and high pressure (cyclone) in the Northern Hemisphere such as represented by professional practitioners of meteorology:

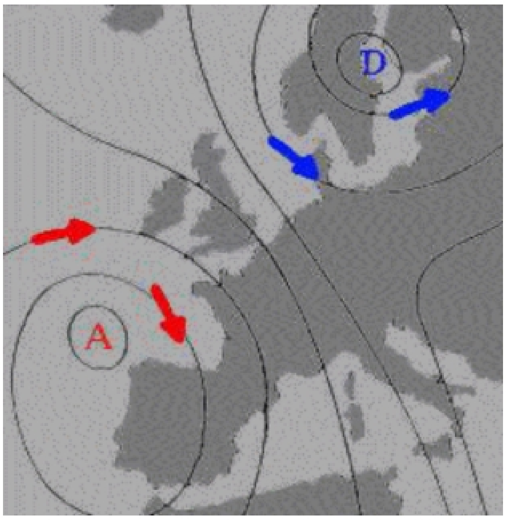


Figure 16.30: **Low pressure and high pressure illustration**

We can indeed observe that depression (D) rotates counter-clockwise and the high pressure (A) clockwise (and vice versa in the southern hemisphere).

16.1.8 Tides

Tides are the rise and fall of sea levels (but also from Earth's atmosphere layers[3]!) caused by the combined effects of the gravitational forces exerted by the Moon and the Sun (actually a G2V-type star, aka yellow dwarf) and the rotation of the Earth. The times and amplitude of tides at a locale are influenced by the alignment of the Sun and Moon, by the pattern of tides in the deep ocean, by the amphidromic systems of the oceans, and the shape of the coastline and near-shore bathymetry. Tides vary on timescales ranging from hours to years due to a number of factors.

Tidal phenomena are not limited to the oceans, but can occur in other systems whenever a gravitational field that varies in time and space is present. For example tides also occurs on lakes, for example in Switzerland on the Leman - the largest lake in Western Europe - the hydrometric station of the Sécheron is able to measure tides of four to five millimeters. Also the solid part of the Earth is affected by tides (GPS measurements seems to give ground tides between 0 to 28 [cm] depending on the place on Earth and the composition of the ground), though this is not as easily seen as the water tidal movements but the LHC accelerator at CERN (Switzerland, Geneva) have to adjust the protons beams regularly to account for the gravitational effect of the Moon.

Among the phenomena of nature, the tide is one of the most majestic in its scope and power, one of the more surprising by its regularity and the discretion of its causes. We can then easily understand not only that it is imposed to the attention of the traveller but it has, since the remotest antiquity, aroused the research of the most distinguished scholars.

To approach the subject of tides in a simply way, we can state from a simple vectorial analysis that: If the Lunar attraction was identical at each point of the Earth, there would be no tides. We must therefore approach the study of tides on the differences of forces. The influence of the Moon on the tides is named the "diurnal component" and should explain why there are tides twice a day[4].

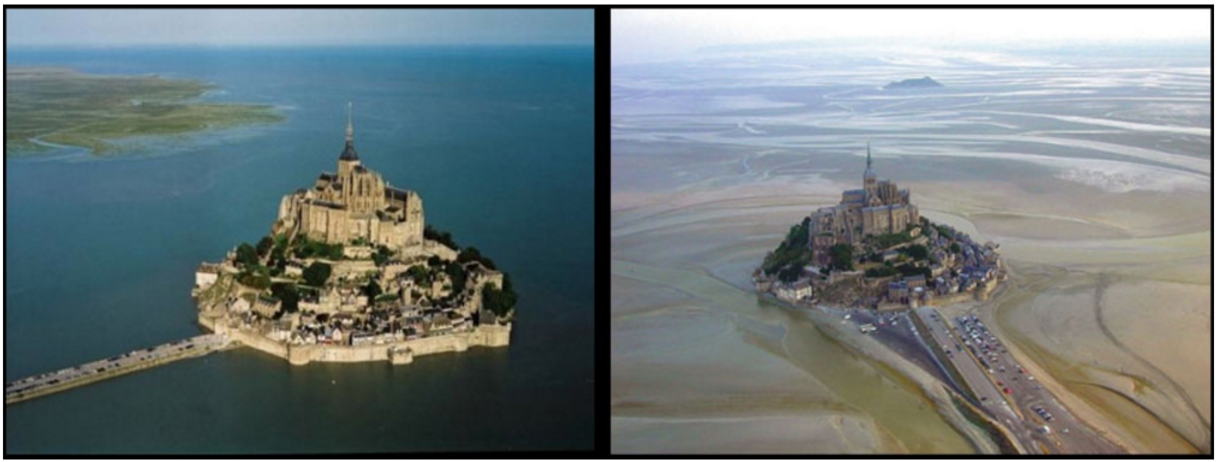


Figure 16.31: **High and low tide at Mont Saint-Michel in France**

---

[3] At the opposite of sea tides where the Moon has the main influence, for the atmosphere it is the Sun (because that latter is strongly heating the atmosphere layers too!)

[4] Even if most tides are visible by the human eyes on beaches of Ocean and Seas, for lakes tides also exist but are difficult to observe. We need careful experimental measurements as the amplitudes are very small and take into account that most of the time for lakes the weather (atmospheric pressure) has more impact than the Moon itself...

Another nice phenomenon to study mathematically (and to see in real life or on YouTube) are the "tidal bore" which is the leading edge of the incoming tide that forms a wave (or waves) of water that travels up a river or narrow bay against the direction of the river or bay's current[5]:

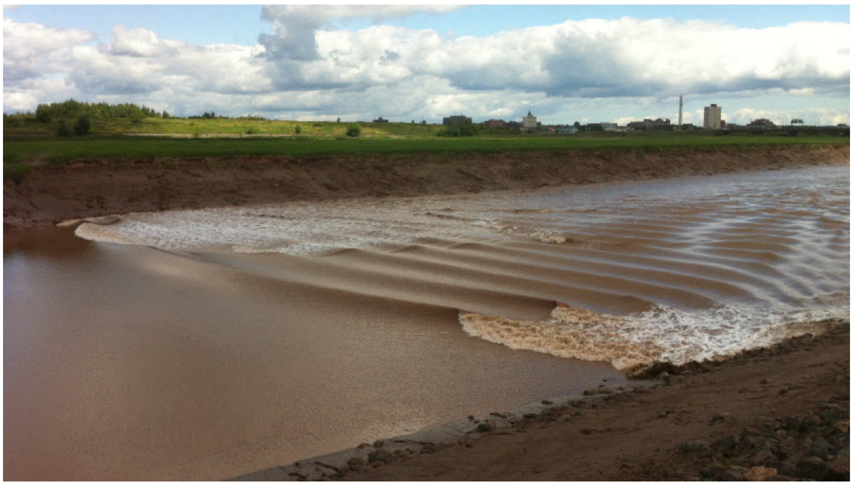


Figure 16.32: **Tidal bore**

The most common type of tides are (some have cyclic and others are totally random and their duration vary a lot):

- Astronomical tides (influence Moon/Sun)
- Rotational tides (due to rotation of Earth)
- Atmospheric tides/Barometric tides[6] (due to meteorological conditions)
- Wind tides[7] (due to the effect of winds on large sea areas)
- Tectonic tides (due to earthquakes)

The ocean and seas tides on Earth are caused by both the Moon's gravity and the Sun's gravity. In general, ocean and seas tides are not generated by the overall strength of gravity as we will see it, but instead by the differences in gravity from one spot to the next (the gravitational gradient). Even though the Sun is much more massive and therefore has stronger overall gravity than the Moon, the Moon is closer to the Earth so that its gravitational gradient is stronger than that of the Sun. Because ocean and seas tides are the effect of ocean water responding to a gravitational gradient, the Moon plays a larger role in creating tides than does the Sun. But the Sun's gravitational gradient across the Earth is significant and it does contribute to tides as well.

As the tides are caused by the gradient of the gravitational field, the tidal force experienced drops with the third power of the distance. This means that the relative strength of the tides should go as :

---

[5] The world largest tidal bore at the time we write these lines seems to be the one located in China's Zhejiang province where it can reach a maximum of nine meter in height at a speed of $40\;[\text{km}\cdot\text{h}^{-1}]$

[6] In areas of high pressure (anticyclones), the sea level decreases. Inversely in areas of low pressure (cyclones). Mean sea level (level 0) corresponds to a pressure of 1013.25 hectopascals (reference pressure). It is necessary to add 1 [cm] or remove 1 [cm] of water height per hectopascal of difference with the reference pressure.

[7] A strong and constant offshore wind causes a drop in sea level of up to 1 meter. Conversely in case of opposite wind direction.

$$\text{ratio} = \frac{M_{\text{moon}} \cdot D_{\text{sun}}^3}{M_{\text{sun}} \cdot D_{\text{moon}}^3} = \frac{7 \cdot 10^{22} \cdot (1.5 \cdot 10^{11})^3}{2 \cdot 10^{30} \cdot (3.7 \cdot 10^8)^3} = 2.3 \tag{16.1.188}$$

So although the Sun is more massive, its greater distance makes its tidal force about 2.3× weaker than the Moon.

16.1.8.1 Astronomical tides (influence Moon/Sun)

Let us first deal with the most famous type of tide that are the astronomical one (due mainly to Sun and Moon)!

Astronomical tide models range from basic theoretical approaches to highly detailed computational simulations. Here's a progression from the simplest to the most advanced:

- **Equilibrium Tide Model** (high-school and undergraduate level model): Based purely on gravitational forces from the Moon and Sun, assuming a perfectly uniform ocean exists in two versions (on considering only the center of mass and can be introduced to almost any high-school student and another considering the vector field that need undergraduate level knowledge of physics). Ignores real-world effects like Earth's rotation, coastlines effects, ocean depth, tidal inertia, etc. The model was developed by Pierre-Simon Laplace in the late 118th century (see the derivations below) according to holocene calendar.
- **Harmonic Analysis Tide Model** (graduate level model): Based on vector field analysis and Navier-Stokes equations taking into account Earth's rotation, fluid sheet thickness, friction resonance and natural periods of ocean basins and leads to spherical harmonics solutions based on historical tide gauge data to identify the tidal parameters of the model and predict future tides ("Laplace's tides equations"). This model requires long-term observations and does not account for sudden oceanographic changes (the theoretical derivation will be written maybe in the future in this textbook). The model was also developed by Pierre-Simon Laplace in the late 118th century (see the derivations below) according to holocene calendar.
- **Hydrodynamic Tide Models** (postgraduate level model): Simulates tidal movements using fluid dynamics equations, incorporating bathymetry, currents, and meteorological influences. This model is computationally demanding; depends on precise input data and need supercomputers (this model will never be detailed in this textbook). The model was developed by a lot of different people from the beginning the early 119th century according to holocene calendar and is still improving today with actually 400 parameters...

16.1.8.1.1 First approach (naive center of mass model)

Let us consider for the naive study of tilde a mass of water m at the equator and at the poles of a rigid Earth (i.e. ground not under the influence of the Moon gravity). We will calculate the attractive force on the mass relatively to the center of the Earth and taking into account the influence of the Moon that has a mass M_L (Lunar influence).

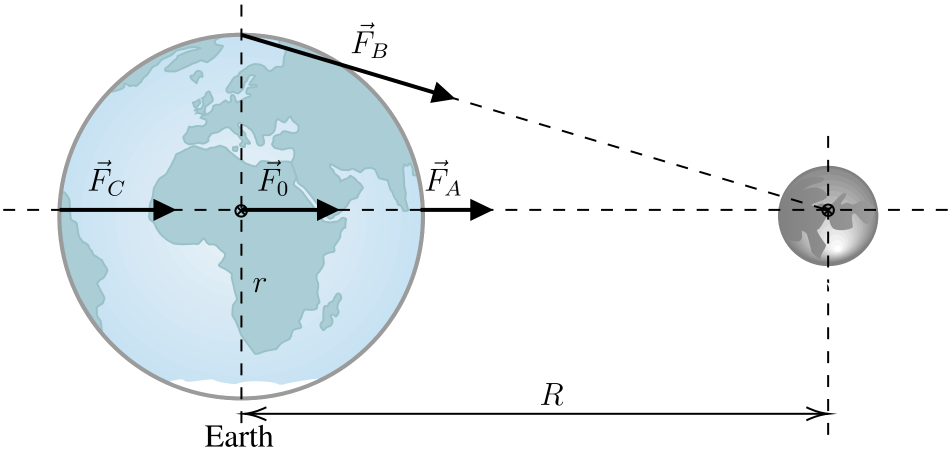


Figure 16.33: **Earth-Moon configuration for a first study of tides**

Let us first compute the force F_a at the equator to the nearest point to the Moon relatively to the figure above.

Then we have:

$$\begin{aligned}\Delta F_L &= F_A - F_0 = G\frac{mM_L}{(R-r)^2} - G\frac{mM_L}{R^2} = GmM_L\frac{R^2-(R-r)^2}{R^2(R-r)^2}\\ &= GmM_L\frac{R^2-R^2+2Rr-r^2}{R^4-2R^3r+R^2r^2} \cong GmM_L\frac{2Rr}{R^4} = 2GmM_L\frac{r}{R^3} = 2f\end{aligned} \tag{16.1.189}$$

considering that $R \gg r$ and denoting the "static tidal force":

$$\boxed{f = GmM_L\frac{r}{R^3}} \tag{16.1.190}$$

A numerical application for a mass $m = 1$ [kg] at the equator then gives $f = 1\cdot 10^{-6}\,[\mathrm{m}\cdot\mathrm{s}^{-1}]$. It's negligible relatively to Earth's gravity g, so this can't explain the tides we observe on the beaches (that have amplitudes of a few meters!).

In vector form the previous relation is obviously written:

$$\Delta\vec{F}_L = 2GmM_L\frac{r}{R^2}\frac{\vec{R}}{\|\vec{R}\|} \tag{16.1.191}$$

As the Earth-Moon distance is approximately 60 Earth radii, the intensity of acceleration varies approximately linearly (...) along the land portion of a line passing through the center of the Moon. This is particularly the case for the segment that connects the two antipodal points A

and C in the figure above. We can write, O denoting the center of the Earth, that:

$$\begin{aligned}\vec{F}_A &\cong \vec{F}_0+\Delta\vec{F}_L\\ \vec{F}_C &\cong \vec{F}_0-\Delta\vec{F}_L\end{aligned} \tag{16.1.192}$$

We must now have to separate the two contributions of the Moon:

1. The force $\vec{F}_0$ applied on the center of mass G is then uniform to the planet by construction. It is this force that is responsible for the revolution of our planet around the common center of mass of the two celestial objects.
2. The residual term $\pm\Delta\vec{F}_L$ is superimposed and takes opposite values at the antipodes. It is responsible for the tides (in the first approximation in this simplistic model).

Thus, the force due to the Moon is of opposite sign to the horizontal. We then have two (lunar) tides per day at antipodal locations:

- That of the Moon who attracts (from this side of the Earth)
- That of the Moon that pushes (at the opposite side of the Earth)

Or schematically (without any respect for real proportions nor the fact that the Moon orbit is elliptical and the Earth is placed on one of the foci of the ellipse):

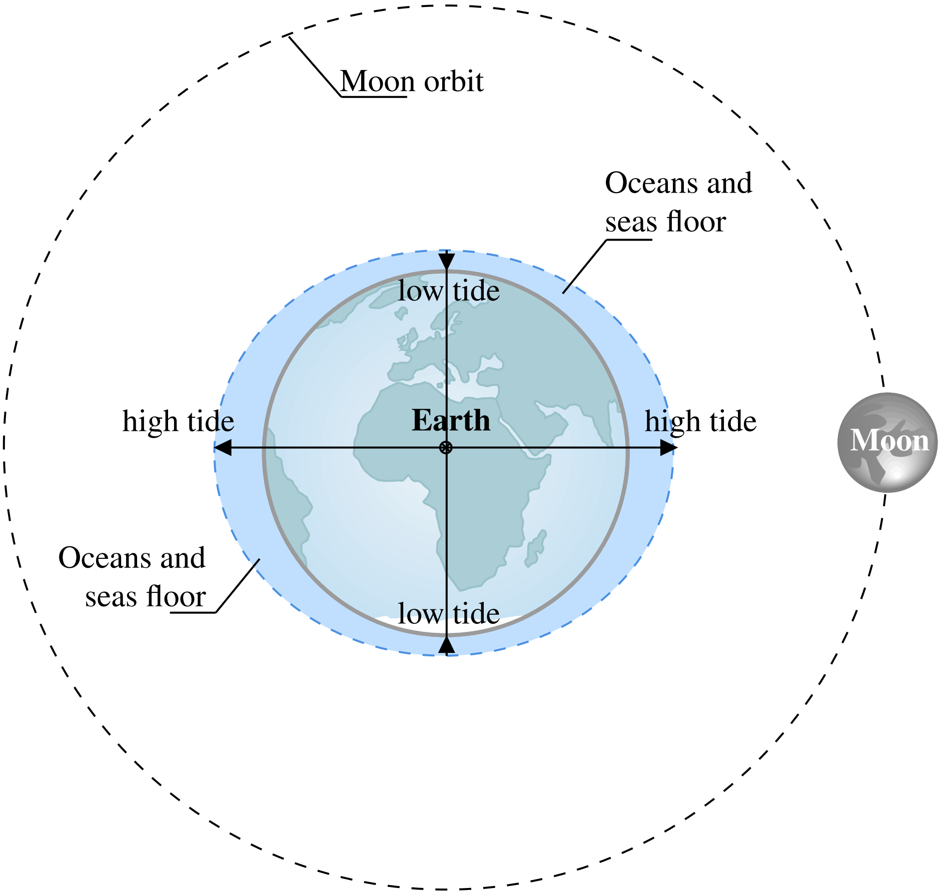


Figure 16.34: **Antipodal tide with antipodal bulge**

If we consider the surface of the Earth as perfectly spherical and covered with water, it would then take the form of an ellipsoid whose axis would be directed to the celestial object generating the tide. We should then observe high tide and low tide that would take place twice a day and always at the same time. We name this the "static tide" and the corresponding model "static model of the tides".

> 💣 The above figure is wrong even if we found it in many textbooks and even some TV show! The outgrowth is incorrectly placed - referred as the "Newtonian model of tides" - and should be rotated from $\pi/2$ as observed experimentally on our planet (there is almost no tides when the Moon is visible above us) and referred as the "Laplacian model of tides".

It should be noted that the subtle play between the rotation of the Earth and the Moon produces enormous friction in water volume that have the effect of slowing the speed of rotation of the Earth about 1 second every thousand years (this will result in a far future in a "tidal lock[8]").

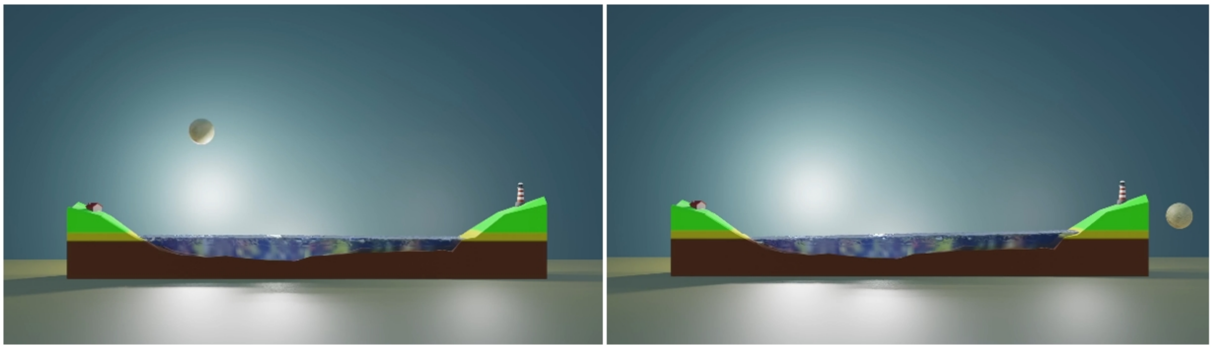


Figure 16.35: Tides are visible in their horizontal component, not in their vertical one (a common misconception by many people) (source: Idées froides/YouTube)

The tides can be huge and varies a lot from area to area because they are simply horizontal waves, moving quite slowly in comparison to conventional waves (even if they move as quick as a running average human), which behaviours depends strongly on the profile of the coasts and the depth of the water (see page 8344). That's why most of the times, ocean tides are bigger than seas tides.

The reflection of waves on the continents explains also why there are also sometimes only on tide per day or even more depending on the place on Earth... obviously of unequal amplitudes! As the reader can imagine, this is real life much more complicated because it is highly multi-factorial!

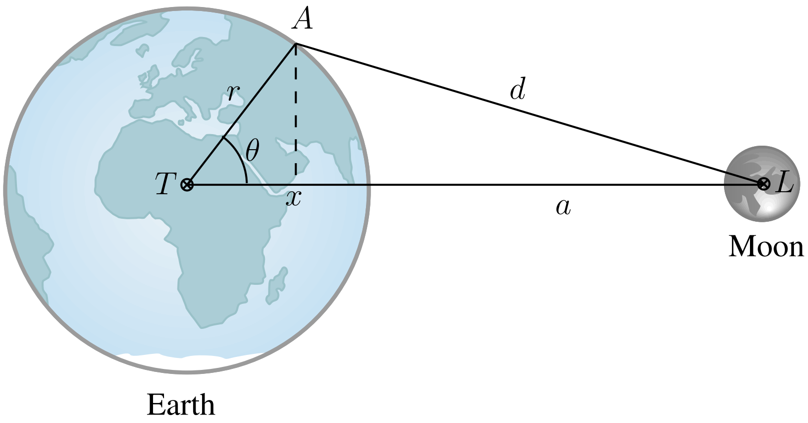
Tides happen without surprise due to the gravitational interactions between the Earth and the Moon. We can say that the tides are pulled by the Moon's gravitational field and so it keeps on changing as it moves out of the Moon's range. The Moon is orbiting the Earth so it has kinetic energy as a consequence of its motion. The Earth is also rotating on its axis, so it has kinetic energy as a consequence of its motion. It is the kinetic energy of the Earth and the Moon which is ultimately the source of tidal energy. That kinetic energy is continually being reduced as a consequence of the tidal friction, but the loss is relatively small.

---

[8] Tidal locking is the phenomenon by which a body has the same rotational period as its orbital period around a partner. So, the Moon is tidally locked to the Earth because it rotates in exactly the same time as it takes to orbit the Earth.

16.1.8.1.2 Second approach (vector field model)

For the second approach, which is worth seeing for general culture and also because it presents another interesting aspect of the explanation of the attraction between two celestial objects, let us consider the following scheme where the Earth doesn't need to be considered as rigid:



We have proved in the section Trigonometry at page 2524 the cosine theorem (Al-Kashi formula) which gives us for recall:

$$d^2 = a^2 + r^2 - 2ar\cos(\theta) \tag{16.1.193}$$

Hence the gravitational potential (see section Astronomy page 5266):

$$\varphi = \frac{GM}{d} = \frac{GM}{\sqrt{a^2 + r^2 - 2ar\cos(\theta)}} = \frac{GM}{a}\frac{1}{\sqrt{1 + \dfrac{r^2}{a^2} - \dfrac{2r\cos(\theta)}{a}}} \quad [\mathrm{J} \cdot \mathrm{kg}^{-1}] \tag{16.1.194}$$

where M is the mass of the Moon. But:

$$\frac{r}{a} \ll 1 \tag{16.1.195}$$

and the gravitational potential is of the form:

$$f(x) = \frac{1}{\sqrt{1+x}} \tag{16.1.196}$$

Which we can develop in series of Maclaurin (see section Sequences and Series page 1703) until order 2:

$$\begin{aligned} f(x) = \frac{1}{\sqrt{1+x}} \underset{x_0=0}{\cong} &= \frac{1}{0!}f(x_0)x^0 + \frac{1}{1!}f'(x_0)x^1 + \frac{1}{2!}f''(x_0)x^2 \\ &= \frac{1}{0!}\frac{1}{(1+x_0)^{1/2}} - \frac{1}{1!}\frac{1}{2}\frac{1}{(1+x_0)^{3/2}}x + \frac{1}{2!}\frac{3}{4}\frac{1}{(1+x_0)^{5/2}}x^2 \\ &= \frac{1}{0!} - \frac{1}{1!}\frac{1}{2}x + \frac{1}{2!}\frac{3}{4}x^2 = 1 - \frac{1}{2}x + \frac{3}{8}x^2 \end{aligned} \tag{16.1.197}$$

The gravitational potential then becomes:

$$\begin{aligned}\varphi &= \frac{GM}{a}\frac{1}{\sqrt{1+\dfrac{r^2}{a^2}-\dfrac{2r\cos(\theta)}{a}}}=\frac{GM}{a}\left[1-\frac{1}{2}\left(\frac{r^2}{a^2}-\frac{2r\cos(\theta)}{a}\right)+\frac{3}{8}\left(\frac{r^2}{a^2}-\frac{2r\cos(\theta)}{a}\right)^2\right]\\ &=\frac{GM}{a}\left[1-\frac{1}{2}\frac{r^2}{a^2}+\frac{r\cos(\theta)}{a}+\frac{3}{8}\left(\frac{r^4}{a^4}-2\frac{r^2}{a^2}\frac{2r\cos(\theta)}{a}+\frac{4r^2\cos^2(\theta)}{a^2}\right)\right]\end{aligned}\tag{16.1.198}$$

If we keep only the terms of power 1 and 2 on r/a, it remains:

$$\begin{aligned}\varphi &\cong \frac{GM}{a}\left[1+\frac{r}{a}\cos(\theta)-\frac{1}{2}\frac{r^2}{a^2}+\frac{3}{8}\frac{4r^2\cos^2(\theta)}{a^2}\right]=\frac{GM}{a}\left[1+\frac{r}{a}\cos(\theta)-\frac{1}{2}\frac{r^2}{a^2}+\frac{1}{2}\frac{3r^2\cos^2(\theta)}{a^2}\right]\\ &=\frac{GM}{a}\left[1+\frac{r}{a}\cos(\theta)+\frac{r^2}{a^2}(3\cos^2(\theta)-1)\right]\end{aligned}\tag{16.1.199}$$

The first term of the potential is:

$$\varphi_1=\frac{GM}{a}\tag{16.1.200}$$

It is the potential for $r=0$, that is to say the potential created by the Moon at the center of the Earth. This term contains no variable, it is constant and therefore its gradient is zero, it gives no force since:

$$\vec{F}=-m\vec{\nabla}\varphi_1=0\tag{16.1.201}$$

The second term:

$$\varphi_2=\frac{GM}{a^2}r\cos(\theta)\tag{16.1.202}$$

contains the two variables r and θ. Its gradient will not be zero. It will generate a gravitational force that we will calculate, but using a trick.

Since:

$$x=r\cos(\theta)\tag{16.1.203}$$

It comes:

$$\varphi_2=\frac{GM_L}{a^2}x\tag{16.1.204}$$

Consequently, the gradient in Cartesian coordinates is reduced for the mass $\mathrm{d}m$ situated at point A by:

$$\mathrm{d}\vec{F}=-\mathrm{d}m\vec{\nabla}\varphi_2=-\mathrm{d}m\begin{pmatrix}\partial_x\\ \partial_y\end{pmatrix}\varphi_2=-\mathrm{d}m\begin{pmatrix}\partial_x\\ \partial_y\end{pmatrix}\frac{GM_L}{a^2}x=-\mathrm{d}m\begin{pmatrix}\dfrac{GM_L}{a^2}\\ 0\end{pmatrix}\tag{16.1.205}$$

Thus all the elements of the Earth undergo from the Moon parallel forces (on the x-axis only) directed towards the Moon. The total mass of the Earth is the sum of all these masses and the total force that the Earth undergoes on the part of the Moon is the sum of the elementary forces. Therefore:

$$\vec{F}=\sum -\mathrm{d}m\vec{\nabla}\varphi_2=\sum -\mathrm{d}m\begin{pmatrix}\dfrac{GM_L}{a^2}\\0\end{pmatrix}=\begin{pmatrix}\dfrac{GM_TM_L}{a^2}\\0\end{pmatrix} \tag{16.1.206}$$

And therefore the total force is only along the x-axis and is given by:

$$F_x=\frac{GM_TM_L}{a^2} \tag{16.1.207}$$

This force is the same as if all the mass of the Earth were concentrated at the center T and if the whole mass of the Moon were concentrated at the center L. The Moon undergoes on the part of the Earth the same force of contrary sense, it is the principle of mutual actions. It is the latter which forces the Moon to revolve around the Earth.

The third term is the one responsible for the tides (some reader may recognize here the spherical harmonic Y_2^0, i.e.. the second Legendre polynomial $P_2(x)=1/2(3x^2-1)$):

$$\boxed{\varphi_3=\frac{GM}{a}\frac{1}{2}\frac{r^2}{a^2}\left(3\cos^2(\theta)-1\right)} \tag{16.1.208}$$

The force deriving from the gradient in polar coordinates is given by (see section Vector Calculus page 1875):

$$\vec{F}=-m\begin{pmatrix}\dfrac{\partial}{\partial r}\\\dfrac{1}{r}\dfrac{\partial}{\partial\theta}\end{pmatrix}\varphi_3=-m\begin{pmatrix}\dfrac{\partial}{\partial r}\\\dfrac{1}{r}\dfrac{\partial}{\partial\theta}\end{pmatrix}\frac{GM}{a}\frac{1}{2}\frac{r^2}{a^2}\left(3\cos^2(\theta)-1\right) \tag{16.1.209}$$

Therefore, the radial component of the force is:

$$F_r=-m\frac{\partial}{\partial r}\frac{GM}{a}\frac{1}{2}\frac{r^2}{a^2}(3\cos^2(\theta)-1)=-\frac{GmMr}{a^3}\left(3\cos^2(\theta)-1\right) \tag{16.1.210}$$

The orthoradial component:

$$\begin{aligned}F_\theta&=-m\frac{1}{r}\frac{\partial}{\partial\theta}\frac{GM}{a}\frac{1}{2}\frac{r^2}{a^2}\left(3\cos^2(\theta)-1\right)=-\frac{GmMR}{2a^3}\frac{\partial}{\partial\theta}(3\cos^2(\theta)-1)\\&=-\frac{3GmMr}{2a^3}\frac{\partial}{\partial\theta}(\cos^2(\theta))=-\frac{3GmMr}{2a^3}(-2\cos(\theta)\sin(\theta))=\frac{3GmMr}{a^3}\cos(\theta)\sin(\theta)\end{aligned} \tag{16.1.211}$$

Therefore:

$$\vec{F}=\begin{pmatrix}F_r\\F_\theta\end{pmatrix}=\begin{pmatrix}F_r\\F_\theta\end{pmatrix}=\begin{pmatrix}-\dfrac{GmMr}{a^3}(3\cos^2(\theta)-1)\\\dfrac{3GmMr}{a^3}\cos(\theta)\sin(\theta)\end{pmatrix} \tag{16.1.212}$$

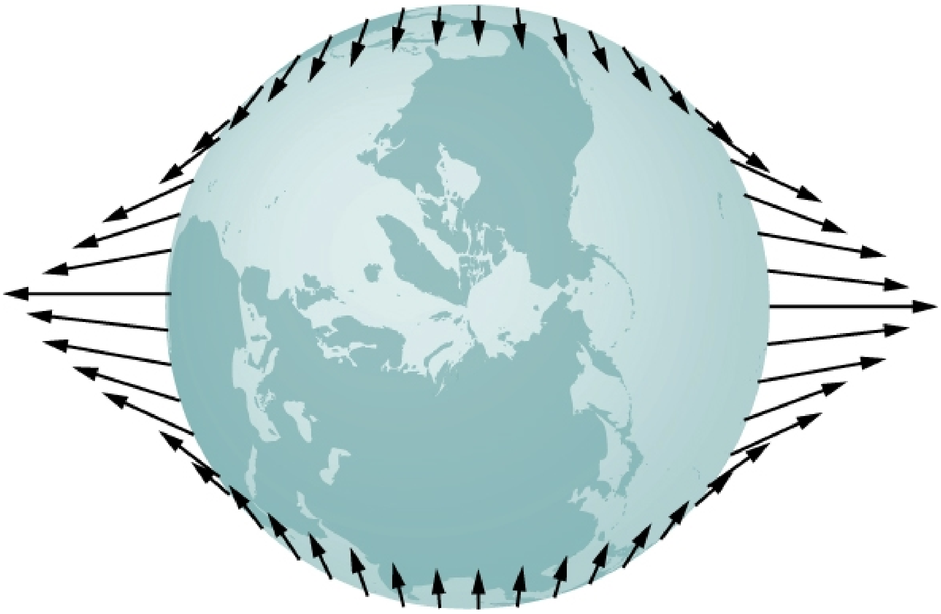
For a mass m of water, the tidal force is:

$$\vec{F}=-m\vec{\nabla}\varphi_3 \tag{16.1.213}$$

For $\theta=\{0,\pi/2,\pi\}$, the orthoradial component F_θ vanishes. We then have for these three angles:

$$\vec{F}=\begin{pmatrix}F_r\\F_\theta\end{pmatrix}=\begin{pmatrix}-\dfrac{GmMr}{a^3}(3\cos^2(\theta)-1)\\0\end{pmatrix}\tag{16.1.214}$$

The tidal force is the difference between the gravitational force at the center and that elsewhere. In the figure below, the tidal forces are shown at the ocean surface. These forces would diminish to zero as you approach Earth's center:



For $\theta=0$ the force is only radial:

$$F_r=-\frac{GmMr}{a^3}(3\cos^2(\theta)-1)=-\frac{GmMr}{a^3}(3\cdot(1)^2-1)=-\frac{2GmMr}{a^3}\tag{16.1.215}$$

The sign of this force is negative and it is directed towards the Moon. For $\theta=\pi$ the force is also only radial and we fall back on the same expression:

$$F_r=-\frac{GmMr}{a^3}(3\cos^2(\theta)-1)=-\frac{GmMr}{a^3}(3\cdot(-1)^2-1)=-\frac{2GmMr}{a^3}\tag{16.1.216}$$

Therefore, the amplitude is the same for the angle $\theta=0=\pi$. Notice that the result is similar with the one we have get with the first approach, at the difference of a factor 2. It is however still not enough to explain the tides as we observe them!

On the other hand, if the reader remember that to return to Cartesian components we have (see section Vector Calculus page 1863):

$$\begin{aligned}x&=r\cos(\theta)\\y&=r\sin(\theta)\end{aligned}\tag{16.1.217}$$

We see then that for $\theta=0$, the force is oriented in the positive direction of the x-axis since x will be positive (but zero following y). There is therefore a tide in the direction of the Moon (almost intuitive).

We also see that for $\theta=\pi$, the force is oriented in the negative direction of the x-axis since x will be negative (but zero following y). There is therefore a tide in the opposite direction to the Moon (counter-intuitive for most people).

For $\theta=\{\pi/2, 3\pi/2\}$, the component x is zero and the radial component will be:

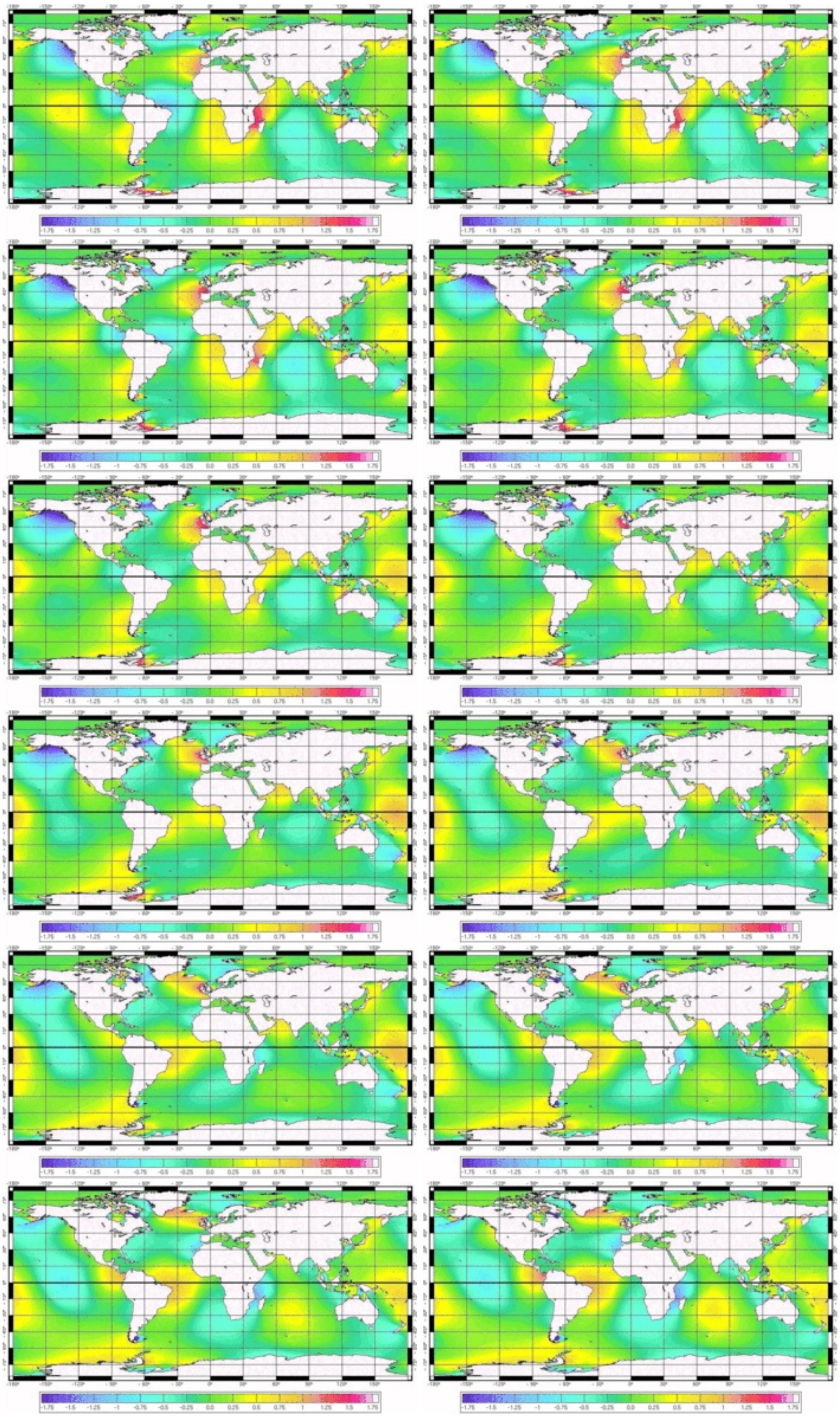
$$F_r=-\frac{GmMr}{a^3}(3\cos^2(\theta)-1)=\frac{GmMr}{a^3} \tag{16.1.218}$$

And it is directed towards the center of the Earth since:

$$\begin{aligned}x&=r\cos(\theta)\\y&=r\sin(\theta)\end{aligned} \tag{16.1.219}$$

We see then that for $\theta=\pi/2$ the force is oriented in the positive direction of the axis y since y will be positive (but zero following x). We also see that for $\theta=3\pi/2$, the force is oriented in the negative direction of the y axis since y will be negative (but zero following x).

In reality, the tides are much more complex than the above model (eccentricity of the lunar orbit, superimposition of the diurnal tide, lunar orbit, Moon-Sun alignment, inclination of the plane of the Moon's orbit, equinoxes, etc.). Here is a superb series of images of the elevation of the surface of the oceans in meters, on one tide cycle , calculated from a more elaborate model:


-1.75
-1.5
-1.25
-1
-0.75
-0.5
-0.25
0.0
0.25
0.5
0.75
1
1.25
1.5
1.75

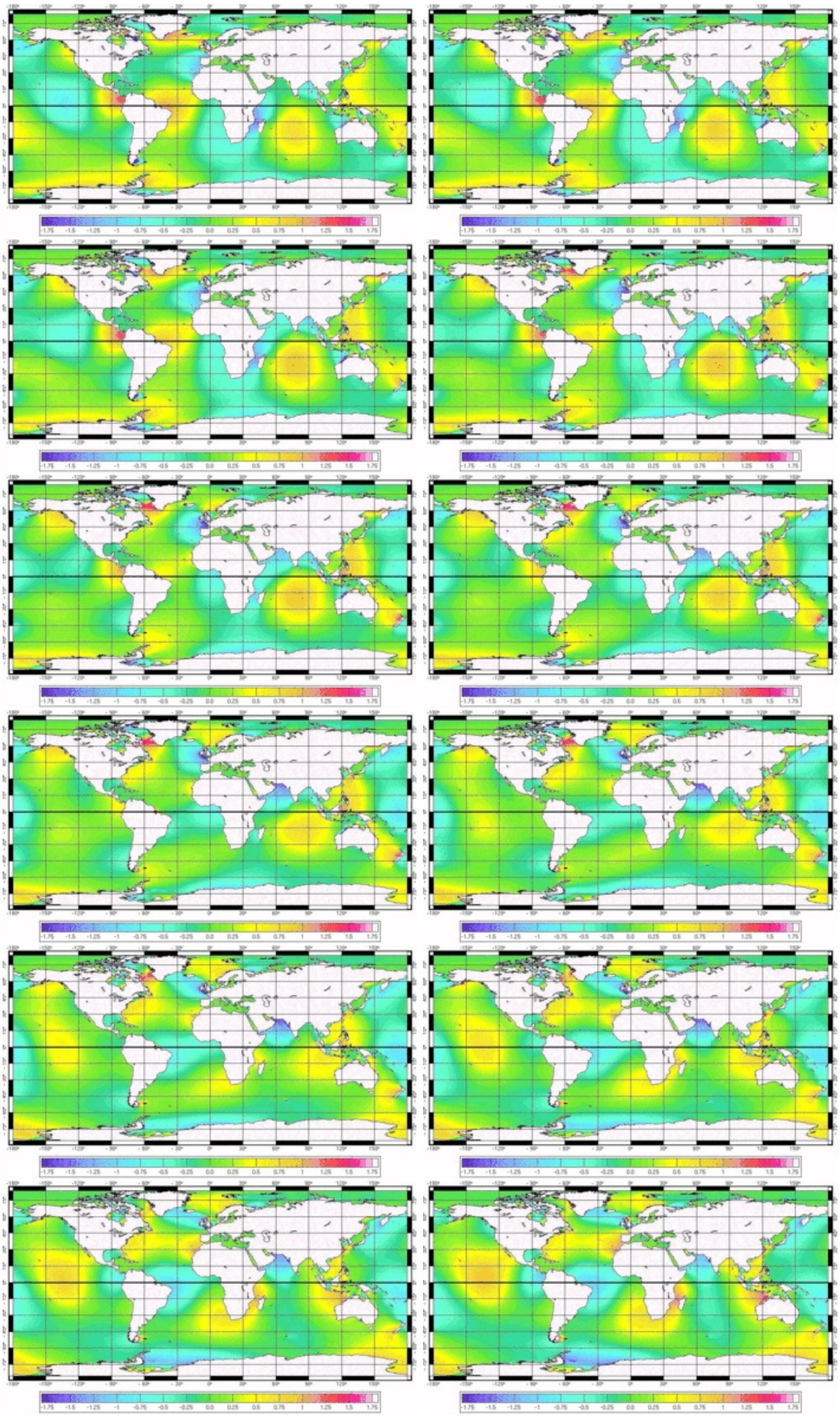


Figure 16.37: **Real complexity of tides on Earth (source: Wikipedia)**

If the Earth were a perfect sphere without large continents, all areas on the planet would experience two equally proportioned high and low tides every lunar day. The large continents on the planet, however, block the westward passage of the tidal bulges as the Earth rotates. Unable to move freely around the globe, these tides establish complex patterns within each ocean basin that often differ greatly from tidal patterns of adjacent ocean basins or other regions of the same ocean.

Three basic tidal patterns occur along the Earth's major shorelines. In general, most areas have two high tides and two low tides each day. When the two highs and the two lows are about the same height, the pattern is named a "semi-daily or semidiurnal tide". If the high and low tides differ in height, the pattern is named a "mixed semidiurnal tide".

Some areas, such as the Gulf of Mexico, have only one high and one low tide each day. This is named a "diurnal tide". The U.S. West Coast tends to have mixed semidiurnal tides, whereas a semidiurnal pattern is more typical of the East Coast.

Remark

The phenomenon is therefore due to the deformation of the surface of the oceans as a result of the combined attractions of the celestial bodies. This movement can even destroy the celestial object which undergoes it: if the tidal force prevails over the gravitational force of its constituents, the celestial objects disintegrates. This limit where the tidal forces outweigh the gravitational force is named the "Roche limit" (see section Astronomy page 5337).

Besides the diurnal tide due to the attraction of the Moon, it is necessary to count on a tide due to the centrifugal force of the movement of the Earth and the Moon around their center of mass (but it depends on the latitudes, the relief and a lot of other parameters objectively because in some places of the planet there is only one tide per day). Indeed, the Earth and the Moon revolve around the center of mass which defines the orbit of Earth-Moon couple (the scales are not respected):

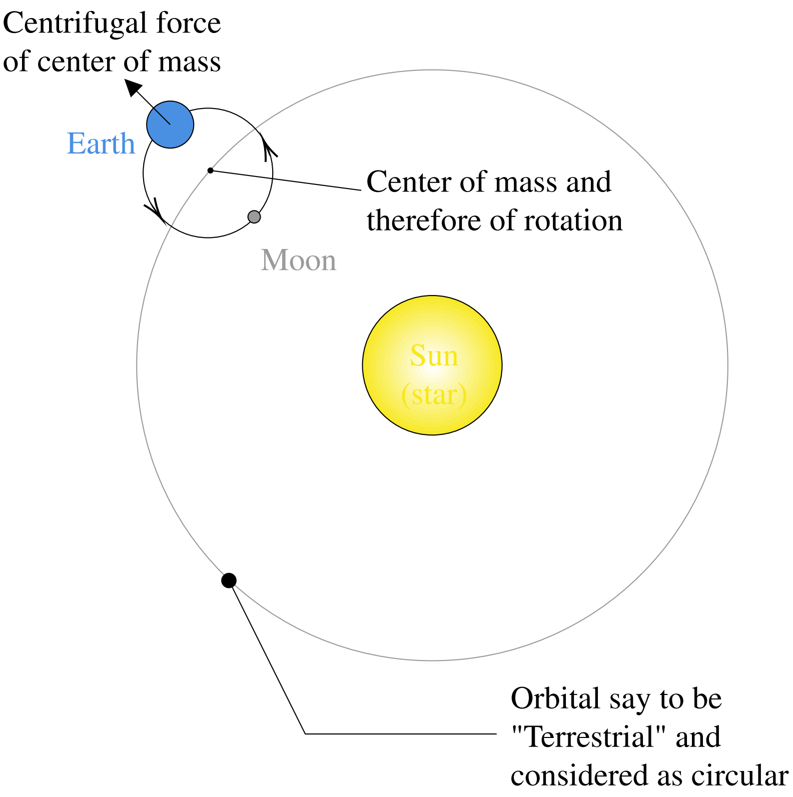


Figure 16.38: **Principle of tides superimposed on diurnal tides**

And we will ignore the tides of equinoxes and others... at least for now...

16.1.8.1.2.1 Tides hight

The simplest way to find the height difference between high and low tides Δh is to observe that the surface of the ocean is an equipotential surface - a surface of constant potential energy. For this, consider a drop of sea water on the surface of the ocean. The drop is in equilibrium (relative to the Earth's reference frame) under the influence of three forces (i.e. potentials): the Earth's gravitational pull, the tidal force and the pressure force of the surrounding sea water. Since a static fluid cannot exert any shearing force, the pressure force must be normal to the surface of the ocean (it is in fact just the buoyancy force of Archimedes' principle). Since the drop of water is in equilibrium, it follows that the tidal potential derived earlier above (for reminder r is the Earth radius and a is the Earth-Moon distance center-to-center):

$$\varphi_{\text{tid.}} = \frac{GM_{\text{moon}}r^2}{2a^3}\left(3\cos^2(\theta)-1\right)\;[\mathrm{J}\cdot\mathrm{kg}^{-1}] \tag{16.1.220}$$

is equal to the geopotential (see section Continuum Mechanics page 3840):

$$\varphi_{\text{geo.}} = g\Delta h\,[\mathrm{J}\cdot\mathrm{kg}^{-1}] \tag{16.1.221}$$

Therefore we can equate these both relations such that:

$$g\Delta h = \frac{GM_{\text{moon}}r^2}{2a^3}\left(3\cos^2(\theta)-1\right) \tag{16.1.222}$$

Rearranging we get:

$$\Delta h=\frac{GM_{\text{moon}}r^2}{g2a^3}\left(3\cos^2(\theta)-1\right) \tag{16.1.223}$$

and as:

$$g=\frac{GM_{\text{earth}}}{r^2} \tag{16.1.224}$$

we get:

$$\Delta h=\frac{M_{\text{moon}}r^4}{2M_{\text{earth}}a^3}\left(3\cos^2(\theta)-1\right) \tag{16.1.225}$$

Or more explicitly written:

$$\boxed{\Delta h=\frac{M_{\text{moon}}R_{\text{earth}}^4}{2M_{\text{earth}}d_{\text{earth-moon}}^3}\left(3\cos^2(\theta)-1\right)} \tag{16.1.226}$$

The extreme values of Δh occur for $\theta=0°$ and $\theta=90°$ gives:

$$\begin{aligned}\Delta h(\text{moon})&=0.38\,[\text{m}]\text{ and }-0.19\,[\text{m}]\\ \Delta h(\text{sun})&=0.17\,[\text{m}]\text{ and }-0.08\,[\text{m}]\end{aligned} \tag{16.1.227}$$

Hence we see that the maximum cumulative difference in this model with the highest and lowest tide is of $(0.38+0.17)+(0.19+0.08)=0.82\,[\text{m}]$. So we are for example very far from the 13 [m] tides of the Mont-Saint-Michel.

Remark

A common question on the Internet relatively to tides is: "*If gravitation can rise entire seas, lands and oceans why doesn't it rise human bodies from the ground (or other animals) ?*" The answer is quite simple and can be understood in two steps. The first one is that the gravitational force is the product of the mass that attracts with the mass that is attracted $(M_1\cdot M_2)$. As the mass of an ocean has a ratio typically of $2\cdot 10^{19}$ with that of a human so is the attractive force too! So the analogy is (as always) a false equivalence... The second step is to understand that humans cannot sense very weak accelarations excepted if they are in a lab and use very sensitive instruments. To grasp the magnitude involved, the reader must keep in mind that the gravitational force of the Moon on a typical human is extremely weak, estimated at approximately $0.004\,[\text{N}]$.

16.1.9 Lorenz equations

The "free convection" or "natural convection" is the flow regime obtained when we heat a fluid without imposing any external flow. This is the case for atmospheric convection movements (hot gases in cold gases), convection movements of the molten rock responsible for plate tectonics, movements of hot water under pressure in geysers or cakes... and many other phenomena...

These flows are inexplicable if we do not couple the equations of dynamics and thermodynamics!

We will in this context establish the famous system of Lorenz equations at the price, however, of numerous approximations and assumptions in order to simplify as much as possible the calculations and the mathematical tools used (because at the time of the development of the model the computers were not what they are today).

We will therefore prove in the context of convection (one of the important dynamics of our atmosphere) that the equations which determine certain parameters of the motion are very sensitive to the initial conditions, the purpose of which is to show the difficulty of predicting to a more or less long term with deterministic theoretical models (reason why in meteorology we use nowadays the method of the finite elements).

A priori, the density ρ is a function of temperature and pressure by the state law of ideal gases (see section Continuum Mechanics page 3903). It is therefore natural to think that if we heat a wall, the temperature of the surrounding fluid increases by diffusion. The pressure stratification is changed, the pressure gradient creates the movement.

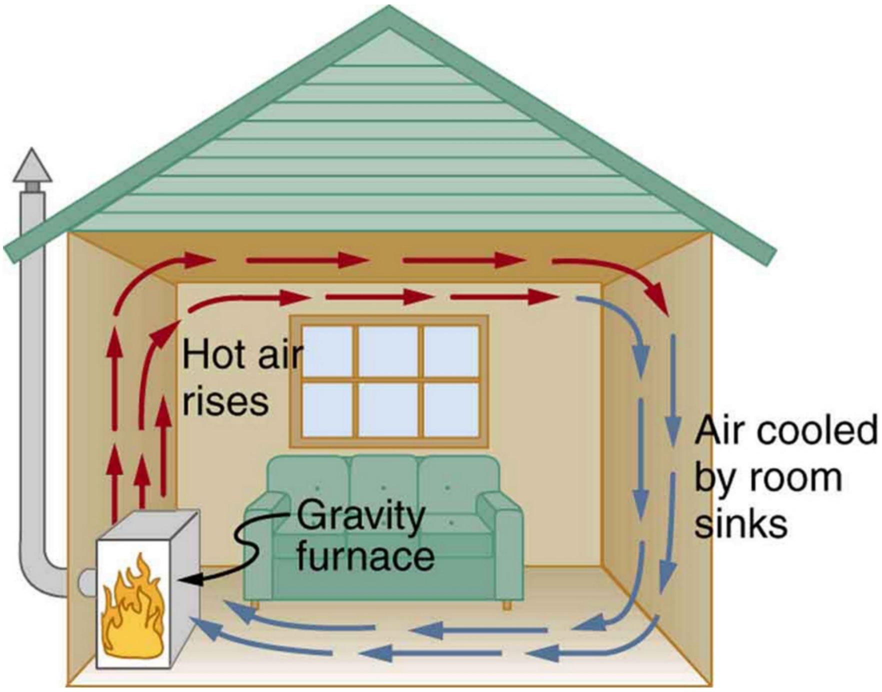


Figure 16.39: **One of the everyday life most known convection...**

In all sections of this textbook, we have so far neglected any variation of ρ. But decoupling is no longer valid here since it is heating that causes movement. We will therefore allow a variation of the density with the heating assuming however that this perturbation is small. It is therefore necessary to reintroduce a variation of ρ around an equilibrium position: rest. On the other hand, the viscosity will be assumed constant...

Thus, given a far fluid at rest and at the temperature T_∞, it is in the presence of a wall heated to the temperature T_P. To obtain the dependence of ρ, let us recall some classical thermo-elastic coefficients (see section Thermodynamics page 3612):

- Coefficient of isobaric compressibility (or dilatation according to the writing in terms of density):

$$\alpha=\frac{1}{V}\frac{\partial V}{\partial T}=\frac{1}{M\rho}\frac{\partial M\rho}{\partial T}=\frac{1}{\rho}\frac{\partial\rho}{\partial T} \tag{16.1.228}$$

- Isothermal compressibility coefficient:

$$\chi=-\frac{1}{V}\frac{\partial V}{\partial P}=-\frac{1}{M\rho}\frac{\partial M\rho}{\partial P}=-\frac{1}{\rho}\frac{\partial\rho}{\partial P} \tag{16.1.229}$$

By assuming now that density is mainly related to temperature (for simplicity) we can write (this hypothesis works well for fluids but not too much ... for gas !!):

$$\rho=f(T) \tag{16.1.230}$$

Using the general form of Taylor's development (see section Sequences and Series page 1696):

$$f(x)=\sum_{k=0}^{n}\frac{1}{k!}\frac{\mathrm{d}^k f(x_0)}{\mathrm{d}x^k}(x-x_0)^k \tag{16.1.231}$$

We then have an approach with the engineer way... :

$$\begin{aligned}\rho(T)&=\sum_{k=0}^{n}\frac{1}{k!}\frac{\mathrm{d}^k f(x_0)}{\mathrm{d}T^k}(T-T_\infty)^k\cong\rho_\infty+(T-T_\infty\frac{\mathrm{d}\rho(T_\infty)}{\mathrm{d}T}=\rho_\infty+\alpha\rho_\infty(T-T_\infty)\\&=\rho_\infty(1+\alpha(T-T_\infty))=\rho_\infty(1+\bar{T}\alpha(T_P-T_\infty)=\rho_\infty(1+\bar{T}\varepsilon)\end{aligned} \tag{16.1.232}$$

Therefore:

$$\boxed{\rho=\rho_\infty\left(1+\bar{T}\varepsilon\right)} \tag{16.1.233}$$

where $\bar{T}$ is therefore a coefficient without dimensions (as ε ...) more easily measurable experimentally.

The continuity equation (see section Thermodynamics page 3679) or of mass balance that is for recall:

$$\frac{\partial\rho(T)}{\partial t}-\vec{\nabla}\circ(\rho(T)\vec{v})=0 \tag{16.1.234}$$

Then becomes:

$$\frac{\partial\rho_\infty(1+\bar{T}\varepsilon)}{\partial t}-\vec{\nabla}\circ(\rho_\infty(1+\bar{T}\varepsilon)\vec{v})=\rho_\infty\varepsilon\frac{\partial\bar{T}}{\partial t}-\rho_\infty(1+\bar{T}\varepsilon)\vec{\nabla}\circ\vec{v}=0 \tag{16.1.235}$$

to the first order in ε. Moreover, we have proved in the section of Continuum Mechanics (see page 3828) that if the fluid is incompressible:

$$\vec{\nabla}\circ\vec{v}=0 \tag{16.1.236}$$

Let us remember that in a first approximation the fluid is incompressible. It then only remains:

$$\rho_\infty\varepsilon\frac{\partial\bar{T}}{\partial t}=0 \tag{16.1.237}$$

As we wish to study a flow in the presence of gravity, it would be judicious to put:

$$P=p+\rho_\infty gz \tag{16.1.238}$$

and therefore to focus only on variations around the hydrostatic equilibrium position (δp is dimensionless!). We have proved still in the same section of Continuum Mechanics (see page 3828) that in the case of the incompressible fluid with viscosity, the first form of Euler's equation (equation of motion):

$$\rho\mathrm{d}_t\vec{v}=-\vec{\nabla}P+\mu_d\vec{\nabla}^2v+\rho f \tag{16.1.239}$$

Let us first consider the both terms:

$$-\vec{\nabla}P+\rho\vec{f} \tag{16.1.240}$$

which are written along the Z axis:

$$-\frac{\partial}{\partial z}P+\rho f_z=-\frac{\partial}{\partial z}P+\rho g=0 \tag{16.1.241}$$

When there is movement, the projection along Z thus make appears:

$$-\frac{\partial}{\partial z}P+\rho g \tag{16.1.242}$$

which we then rewrite:

$$-\frac{\partial}{\partial z}(p+\rho_\infty gz)+\rho g=-\rho_\infty g\frac{\partial}{\partial z}z-\frac{\partial}{\partial z}p+\rho g=-\rho_\infty g-\frac{\partial}{\partial z}p+\rho g \tag{16.1.243}$$

Therefore:

$$-\frac{\partial}{\partial z}p+(\rho-\rho_\infty)g \tag{16.1.244}$$

since:

$$(\rho-\rho_\infty)=\rho_\infty\varepsilon\bar{T} \tag{16.1.245}$$

It comes:

$$-\frac{\partial}{\partial z}p+\rho_\infty\varepsilon\bar{T}g \tag{16.1.246}$$

It remains therefore a buoyancy force directed upward.

The variation of the density as a function of temperature in the product $\rho \mathrm{d}_t\vec{v}$ of the relation:

$$\rho \mathrm{d}_t\vec{v} = -\vec{\nabla}P + \rho\vec{f} \tag{16.1.247}$$

will be neglected ($\rho \mathrm{d}_t\vec{v} \cong \rho_\infty \mathrm{d}_t\vec{v}$) because we will restrict ourselves in case where the speed is small. We have then by reintroducing the viscosity...:

$$\rho_\infty \mathrm{d}_t\vec{v} = -\vec{\nabla}p + \rho_\infty\varepsilon\bar{T}g\vec{e}_z + \mu_d\vec{\nabla}^2 v \tag{16.1.248}$$

and we have then material derivative (see section Continuum Mechanics page 3829):

$$\mathrm{d}_t\vec{v} = (\vec{v}\circ\vec{\nabla})\vec{v} + \partial_t\vec{v} \tag{16.1.249}$$

that we can simplify as:

$$\mathrm{d}_t = \vec{v}\circ\vec{\nabla} + \partial_t \tag{16.1.250}$$

We then have as expression of the force density:

$$\boxed{\rho_\infty\left(\vec{v}(\vec{v}\circ\vec{\nabla}) + \partial_t\vec{v}\right) = -\vec{\nabla}p + \rho_\infty\varepsilon\bar{T}g\vec{e}_z + \mu_d\vec{\nabla}^2 v} \tag{16.1.251}$$

To continue, we will seek to determine the energy law of the equation of behaviour proved in the section Continuum Mechanics (see page 3825) that was for recall:

$$\sigma_{ij} = -p\delta_{ij} - \frac{2}{3}\mu_d(\vec{\nabla}\circ\vec{v})\delta_{ij} + 2\mu_d\theta_{ij} \tag{16.1.252}$$

so that it also accounts for the relation between the stresses and thermodynamic characteristics of the fluid, such as heat flow and temperature. We will do this by characterizing the diffusion of energy in the medium due to the effects (supposedly decoupled) of the viscosity of the fluid and the thermal conduction of the fluid.

We rewrite this relation with new constants and another notation for the divergence:

$$\sigma_{ij} = -p\delta_{ij} + \lambda\mathrm{div}(\vec{v})\delta_{ij} + 2\mu_d\theta_{ij} \tag{16.1.253}$$

where μ_d and λ are in this context named the "Lamé coefficients" (also named the "Lamé parameters" or "Lamé constants"). The terms λ and μ are individually referred to as "Lamé's first parameter" and "Lamé's second parameter".

We have also proved in the section of Continuum Mechanics (see page 3826) the relation:

$$\rho \mathrm{d}_t\vec{v} = \vec{\nabla}(\vec{\sigma}) + \rho\vec{f} \tag{16.1.254}$$

Therefore:

$$\rho((\vec{v}\circ\vec{\nabla})\vec{v} + \partial_t\vec{v}) = \vec{\nabla}(\vec{\sigma}) + \rho\vec{f} \tag{16.1.255}$$

Which gives:

$$\rho\partial_t\vec{v} + \rho(\vec{v}\circ\vec{\nabla})\vec{v} = \vec{\nabla}(\vec{\nabla}) + \rho\vec{g} \tag{16.1.256}$$

Let us denote the total energy as:

$$E_{\text{tot}}=E_c+U=\int_D \frac{1}{2}\rho\vec{v}^2\mathrm{d}V+\int_D e\mathrm{d}V=\rho\int_D\left(\frac{1}{2}\vec{v}^2+e\right)\mathrm{d}V \tag{16.1.257}$$

where e is the internal mass energy of the fluid (so related to a unit of fluid mass). Now the instantaneous variation of the internal energy of the fluid is equal to the contribution of a mechanical power and of the supply of heat (according to what has been seen in the section of Thermodynamics):

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho\int_D\left(\frac{1}{2}\vec{v}^2+e\right)\mathrm{d}V=\frac{\mathrm{d}W}{\mathrm{d}t}+\frac{\mathrm{d}Q}{\mathrm{d}t}=P+\dot{Q} \tag{16.1.258}$$

where P gives the power of the external forces to the system given necessarily by the force of the ambient potential field and the mechanical forces given by the stress tensor only (we are always in the situation of a perfect fluid). That is:

$$P=\int_D \rho\vec{g}\circ\vec{v}\mathrm{d}V+\int_D(\sigma\vec{v})\mathrm{d}\vec{S} \tag{16.1.259}$$

and using the Ostrogradsky theorem (see section Vector Calculus page 1882):

$$\begin{aligned}P&=\int_D\rho(\vec{g}\circ\vec{v})\mathrm{d}V+\int_D((\vec{\nabla})\circ\sigma\vec{v})\mathrm{d}V=\int_D\rho(\vec{g}\circ\vec{v})\mathrm{d}V+\int_D\mathrm{div}(\sigma\vec{v})\mathrm{d}V\\&=\int_D\left[\rho\vec{g}\circ\vec{v}+\mathrm{div}(\sigma\vec{v})\right]\mathrm{d}V=\int_D\left[\rho g_i v_i+\frac{\partial(\sigma_{ij}v_j)}{\partial x_i}\right]\mathrm{d}V\end{aligned} \tag{16.1.260}$$

which has indeed the units of a power. We also have:

$$\begin{aligned}\mathrm{div}(\sigma\vec{v})&=\vec{\nabla}\circ\left(\begin{bmatrix}\sigma_{11}&\sigma_{12}&\sigma_{13}\\\sigma_{21}&\sigma_{22}&\sigma_{23}\\\sigma_{31}&\sigma_{32}&\sigma_{33}\end{bmatrix}\begin{bmatrix}v_1\\v_2\\v_3\end{bmatrix}\right)\\&=\vec{\nabla}\circ\left(\begin{bmatrix}\sigma_{11}v_1+\sigma_{12}v_2+\sigma_{13}v_3\\\sigma_{21}v_1+\sigma_{22}v_2+\sigma_{23}v_3\\\sigma_{31}v_1+\sigma_{32}v_2+\sigma_{33}v_3\end{bmatrix}\right)\\&=\frac{\partial(\sigma_{11}v_1+\sigma_{12}v_2+\sigma_{13}v_3)}{\partial x_1}+\frac{\partial(\sigma_{21}v_1+\sigma_{22}v_2+\sigma_{23}v_3)}{\partial x_2}+\frac{\partial(\sigma_{31}v_1+\sigma_{32}v_2+\sigma_{33}v_3)}{\partial x_3}\\&=\frac{\partial\sigma_{ij}v_j}{\partial x_i}\end{aligned} \tag{16.1.261}$$

For the heat power $\dot{Q}$ it is very easy also thanks to the developments we had made in the section of Thermodynamics where we obtained the equation of heat (see page 3684):

$$\frac{\partial q}{\partial t}+\vec{\nabla}\circ\vec{\Phi}_q=\dot{q}+\vec{\nabla}\circ\vec{\Phi}_q=0 \tag{16.1.262}$$

Therefore:

$$\dot{Q}=\int_D\dot{q}\mathrm{d}V=-\int_D\vec{\nabla}\circ\vec{\Phi}_q\mathrm{d}V \tag{16.1.263}$$

We have finally:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\rho\int_D\left(\frac{1}{2}\vec{v}^2+e\right)\mathrm{d}V&=\int_D\left(\rho g\vec{v}+\mathrm{div}(\sigma\vec{v})\right)\mathrm{d}V-\int_D\vec{\nabla}\circ\Phi_q\mathrm{d}V\\&=\int_D\left(\rho g\vec{v}+\mathrm{div}(\sigma\vec{v})-\vec{\nabla}\circ\vec{\Phi}_q\right)\mathrm{d}V\end{aligned}\tag{16.1.264}$$

So all this gives us the equation of the energy of a fluid:

$$\rho\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}\vec{v}^2+e\right)=\rho g\vec{v}+\mathrm{div}(\sigma\vec{v})-\vec{\nabla}\circ\vec{\Phi}_q\tag{16.1.265}$$

Therefore:

$$\rho\frac{\mathrm{d}e}{\mathrm{d}t}+\frac{1}{2}\rho\frac{\mathrm{d}\vec{v}^2}{\mathrm{d}t}=\rho g\vec{v}+\mathrm{div}(\sigma\vec{v})-\vec{\nabla}\circ\vec{\Phi}_q\tag{16.1.266}$$

and as (see section Thermodynamics page 3684) the heat flow follows the Fourier law:

$$\vec{\Phi}_q=-\kappa\vec{\nabla}T\tag{16.1.267}$$

We have then:

$$\rho\frac{\mathrm{d}e}{\mathrm{d}t}+\frac{1}{2}\rho\frac{\mathrm{d}\vec{v}^2}{\mathrm{d}t}=\rho g\vec{v}+\mathrm{div}(\sigma\vec{v})-\vec{\nabla}\circ(-\kappa\vec{\nabla}T)\tag{16.1.268}$$

Either by using the Laplacian definition of a scalar field (see section Vector Calculus page 1899):

$$\rho\frac{\mathrm{d}e}{\mathrm{d}t}+\frac{1}{2}\rho\frac{\mathrm{d}\vec{v}^2}{\mathrm{d}t}=\rho g\vec{v}+\mathrm{div}(\sigma\vec{v})+\kappa\vec{\nabla}^2T\tag{16.1.269}$$

Or explicitly:

$$\rho\frac{\mathrm{d}e}{\mathrm{d}t}+\frac{1}{2}\rho\frac{\mathrm{d}v_iv_i}{\mathrm{d}t}=\rho g_iv_i+\frac{\sigma_{ij}v_j}{\partial x_i}+\kappa\frac{\partial T}{\partial^2x_i}\tag{16.1.270}$$

Now, by making the scalar product of:

$$\rho\mathrm{d}_t\vec{v}=\vec{\nabla}(\vec{\sigma})+\rho\vec{g}\tag{16.1.271}$$

With the velocity $\vec{v}$ we get the balance of the kinetic energy:

$$\rho\vec{v}\circ\mathrm{d}_t\vec{v}=\vec{v}\circ\vec{\nabla}(\vec{\sigma})+\rho\vec{v}\circ\vec{g}\tag{16.1.272}$$

That is to say explicitly:

$$\rho v_i\mathrm{d}_tv_i=\rho\mathrm{d}_t\left(\frac{v_iv_i}{2}\right)=\frac{1}{2}\rho\frac{\mathrm{d}(v_iv_i)}{\mathrm{d}t}=v_i\frac{\partial\sigma_{ij}}{\partial x_i}+\rho v_ig_i\tag{16.1.273}$$

By subtracting (just proved earlier):

$$\rho\frac{\mathrm{d}e}{\mathrm{d}t}+\frac{1}{2}\rho\frac{d\vec{v}^2}{\mathrm{d}t}=\rho g\vec{v}+\mathrm{div}(\sigma\vec{v})+\kappa\vec{\nabla}^2T\tag{16.1.274}$$

and (just proved also earlier above):

$$\rho\vec{v}\circ \mathrm{d}_t\vec{v}=\vec{v}\circ\vec{\nabla}(\vec{\sigma})+\rho\vec{v}\circ\vec{g} \tag{16.1.275}$$

we get a local relation of the specific internal energy e:

$$\rho\frac{\mathrm{de}}{\mathrm{dt}}=-\vec{v}\circ\vec{\nabla}(\vec{\sigma})+\mathrm{div}(\sigma\vec{v})+\kappa\vec{\nabla}^2T \tag{16.1.276}$$

That is to say explicitly:

$$\rho\frac{\mathrm{de}}{\mathrm{dt}}=-v_i\frac{\sigma_{ij}}{\partial x_j}+\frac{\partial(\sigma_{ij}v_j)}{\partial x_i}+\kappa\frac{\partial T}{\partial^2 x_j} \tag{16.1.277}$$

But we also have (derivation of a product):

$$\frac{\partial(\sigma_{ij}v_j)}{\partial x_i}=\sigma_{ij}\frac{\partial v_i}{\partial x_j}+v_i\frac{\partial\sigma_{ij}}{\partial x_j} \tag{16.1.278}$$

Therefore:

$$\frac{\partial(\sigma_{ij}v_j)}{\partial x_i}-v_i\frac{\partial\sigma_{ij}}{\partial x_j}=\sigma_{ij}\frac{\partial v_i}{\partial x_j} \tag{16.1.279}$$

Indeed:

$$\begin{aligned}\frac{\partial(\sigma_{ij}v_j)}{\partial x_i}&=\frac{\partial(\sigma_{11}v_1+\sigma_{12}v_2+\sigma_{13}v_3)}{\partial x_1}+\frac{(\sigma_{21}v_1+\sigma_{22}v_2+\sigma_{23}v_3)}{\partial x_2}+\frac{\partial(\sigma_{31}v_1+\sigma_{32}v_2+\sigma_{33}v_3)}{\partial x_3}\\&=v_1\frac{\partial\sigma_{11}}{\partial x_1}+\sigma_{11}\frac{\partial v_1}{\partial x_1}+v_2\frac{\sigma_{12}}{\partial x_1}+\sigma_{12}\frac{\partial v_2}{\partial x_1}+v_3\frac{\partial\sigma_{13}}{\partial x_1}+\sigma_{13}\frac{\partial v_3}{\partial x_1}\\&+v_1\frac{\partial\sigma_{21}}{\partial x_2}+\sigma_{21}\frac{\partial v_1}{\partial x_2}+v_2\frac{\sigma_{22}}{\partial x_2}+\sigma_{22}\frac{\partial v_2}{\partial x_2}+v_3\frac{\partial\sigma_{23}}{\partial x_2}+\sigma_{23}\frac{\partial v_3}{\partial x_2}\\&+v_1\frac{\partial\sigma_{31}}{\partial x_3}+\sigma_{31}\frac{\partial v_1}{\partial x_3}+v_2\frac{\sigma_{32}}{\partial x_3}+\sigma_{32}\frac{\partial v_2}{\partial x_3}+v_3\frac{\partial\sigma_{33}}{\partial x_3}+\sigma_{33}\frac{\partial v_3}{\partial x_3}\end{aligned} \tag{16.1.280}$$

We have therefore:

$$\rho\frac{\mathrm{de}}{\mathrm{dt}}=\sigma_{ij}\frac{\partial v_i}{\partial x_j}+\kappa\frac{\partial T}{\partial^2 x_j} \tag{16.1.281}$$

And as the stress tensor σ is symmetric:

$$\sigma_{ij}\frac{\partial v_i}{\partial x_j}=\frac{1}{2}\sigma_{ij}\left(\frac{\partial v_i}{\partial x_j}+\frac{\partial v_j}{\partial x_i}\right)=\sigma_{ij}\frac{1}{2}\left(\frac{\partial v_i}{\partial x_j}+\frac{\partial v_j}{\partial x_i}\right)=\sigma_{ij}\mathrm{d}_{ij} \tag{16.1.282}$$

we have therefore:

$$\rho\frac{\mathrm{de}}{\mathrm{dt}}=\sigma_{ij}d_{ij}+\kappa\frac{\partial T}{\partial^2 x_j} \tag{16.1.283}$$

which is sometimes written (yes I also don't believe what I see sometimes...):

$$\rho\frac{\mathrm{de}}{\mathrm{dt}}=\overline{\overline{\sigma}}:\overline{\overline{D}}+\kappa\nabla^2T \tag{16.1.284}$$

where $\overline{\overline{D}}$ is named the "tensor of deformation rate" and $\overline{\overline{\sigma}}:\overline{\overline{D}}$ represents the double contracted product of the stress tensor and deformation rate tensor.

We have proved, for recall, in the section of Continuum Mechanics (see page 3825) that:

$$\sigma=-p\delta_{ij}-\frac{2}{3}\mu_d(\vec{\nabla}\circ\vec{v})\delta_{ij}+2\mu_d\theta_{ij}=-p\theta_{ij}+2\mu_d\left(\theta_{ij}-\frac{1}{3}(\vec{\nabla}\circ\vec{v})\delta_{ij}\right) \tag{16.1.285}$$

where:

$$\tau=2\mu_d\left(\theta_{ij}-\frac{1}{3}(\vec{\nabla}\circ\vec{v})\delta_{ij}\right) \tag{16.1.286}$$

Therefore, it is simple to make the difference between normal forces and tangential forces. Whatever, to come back on the energy equation:

$$\rho\frac{\mathrm{d}e}{\mathrm{d}t}=\sigma_{ij}d_{ij}+\kappa\frac{\partial T}{\partial^2 x_j} \tag{16.1.287}$$

and replacing explicitly the stress tensor σ, we get:

$$\rho\frac{\mathrm{d}e}{\mathrm{d}t}=-p\delta_{ij}d_{ij}+2\mu_d\left(\theta_{ij}-\frac{1}{3}(\vec{\nabla}\circ\vec{v})\delta_{ij}\right)+\kappa\frac{\partial T}{\partial^2 x_j} \tag{16.1.288}$$

But in our case:

$$d_{ij}=\theta_{ij} \tag{16.1.289}$$

therefore we can write:

$$\rho\frac{\mathrm{d}e}{\mathrm{d}t}=-p\delta_{ij}d_{ij}+2\mu_d\left(d_{ij}-\frac{1}{3}(\vec{\nabla}\circ\vec{v})\delta_{ij}\right)+\kappa\frac{\partial T}{\partial^2 x_j} \tag{16.1.290}$$

But we also have (using the trace as seen in the section of Linear Algebra at page 1953):

$$\delta_{ij}d_{ij}=d_{ii}=tr(D)=\frac{1}{2}\left(\frac{\partial v_i}{\partial x_i}+\frac{\partial v_i}{\partial x_i}\right)=\frac{\partial v_i}{\partial x_i}=\vec{\nabla}\circ\vec{v} \tag{16.1.291}$$

We have therefore:

$$\rho\frac{\mathrm{d}e}{\mathrm{d}t}=-p(\vec{\nabla}\circ\vec{v})+2\mu_d d_{ij}d_{ij}+\frac{1}{3}\mu(\vec{\nabla}\circ\vec{v})^2+\kappa\nabla^2 T \tag{16.1.292}$$

Thus into technical condensed form (just for information...):

$$\rho\dot{e}=-p(\vec{\nabla}\circ\vec{v})+2\mu_d\overline{\overline{D}}^2+\lambda(\vec{\nabla}\circ\vec{v})^2+\kappa\nabla^2 T \tag{16.1.293}$$

It is clear that from the point of view of the entropy (see section Thermodynamics page 3658) we have:

$$S=\int_D \rho s\mathrm{d}V \tag{16.1.294}$$

We also have:

$$\mathrm{d}U=T\mathrm{d}S-p\mathrm{d}V\Leftrightarrow\mathrm{d}S=\frac{\mathrm{d}U}{T}+\frac{p\mathrm{d}V}{T} \tag{16.1.295}$$

Thus reduced to the massic values:

$$\mathrm{d}s=\frac{\mathrm{d}e}{T}+\frac{p\mathrm{d}V}{T}\Leftrightarrow \rho T\mathrm{d}s=\rho\mathrm{d}e+\rho p_m\mathrm{d}V \tag{16.1.296}$$

The time variation giving:

$$\rho T\frac{\mathrm{d}s}{\mathrm{d}t}=\rho\frac{\mathrm{d}e}{\mathrm{d}t}+p_m\frac{\mathrm{d}m}{\mathrm{d}t} \tag{16.1.297}$$

But we have the continuity equation (see section Thermodynamics page 3679):

$$\frac{\mathrm{d}m}{\mathrm{d}t}-\vec{\nabla}\circ(m\vec{v})=0\Leftrightarrow\frac{\mathrm{d}m}{\mathrm{d}t}=\vec{\nabla}\circ(m\vec{v}) \tag{16.1.298}$$

Which gives finally:

$$\rho T\frac{\mathrm{d}s}{\mathrm{d}t}=\rho\frac{\mathrm{d}e}{\mathrm{d}t}+p\vec{\nabla}\circ\vec{v} \tag{16.1.299}$$

or written slightly differently:

$$\rho\frac{\mathrm{d}e}{\mathrm{d}t}=\rho T\frac{\mathrm{d}s}{\mathrm{d}t}-p\vec{\nabla}\circ\vec{v} \tag{16.1.300}$$

Injected in:

$$\rho\frac{\mathrm{d}e}{\mathrm{d}t}=-p(\vec{\nabla}\circ\vec{v})+2\mu_d d_{ij}d_{ij}+\lambda(\vec{\nabla}\circ\vec{v})^2+\kappa\nabla^2 T \tag{16.1.301}$$

This gives:

$$\rho T\frac{\mathrm{d}s}{\mathrm{d}t}=2\mu_d d_{ij}d_{ij}+\lambda(\vec{\nabla}\circ\vec{v})^2+\kappa\nabla^2 T \tag{16.1.302}$$

If we consider the velocity gradient to be very small (quasi-static) then we can write the approximation:

$$\rho T\frac{\mathrm{d}s}{\mathrm{d}t}=2\mu_d d_{ij}^2+\kappa\nabla^2 T \tag{16.1.303}$$

Let us now give the expression of the entropy (exact total differential) as a function of the temperature and pressure parameters only:

$$\mathrm{d}S=\left(\frac{\partial S}{\partial T}\right)_P\mathrm{d}T+\left(\frac{\partial S}{\partial P}\right)_T\mathrm{d}P \tag{16.1.304}$$

Either in massic form:

$$\mathrm{d}s=\left(\frac{\partial s}{\partial T}\right)_P\mathrm{d}T+\left(\frac{\partial s}{\partial P}\right)_T\mathrm{d}P \tag{16.1.305}$$

But we have also proved in the section Thermodynamics (see page 3180) the following relation:

$$T\left(\frac{\partial S}{\partial T}\right)_P=C_P=\left(\frac{\partial H}{\partial T}\right)_P \tag{16.1.306}$$

Either in massic form:

$$T\left(\frac{\partial s}{\partial T}\right)_P = c_P = \left(\frac{\partial h}{\partial T}\right)_P \tag{16.1.307}$$

Which gives us:

$$\mathrm{d}s = \frac{c_p}{T}\mathrm{d}T + \left(\frac{\partial s}{\partial P}\right)_T \mathrm{d}P \tag{16.1.308}$$

But, we have also proved in the section Thermodynamics (see page 3672), one of the following Maxwell's relations:

$$\left(\frac{\partial V}{\partial T}\right)_P = -\left(\frac{\partial S}{\partial P}\right)_T \tag{16.1.309}$$

Either in massic form:

$$\left(\frac{\partial \rho^{-1}}{\partial T}\right)_P = -\left(\frac{\partial s}{\partial P}\right)_T \tag{16.1.310}$$

hence:

$$\mathrm{d}s = \frac{c_p}{T}\mathrm{d}T - \left(\frac{\partial \rho^{-1}}{\partial T}\right)_P \mathrm{d}P \tag{16.1.311}$$

Therefore:

$$\rho T \mathrm{d}S = \rho c_p \mathrm{d}T - \rho T\left(\frac{\partial \rho^{-1}}{\partial T}\right)_P \mathrm{d}P = \rho c_P \mathrm{d}T - \frac{T}{\rho}\left(\frac{\partial \rho}{\partial T}\right)_P \mathrm{d}P \tag{16.1.312}$$

Then our relation:

$$\rho T\frac{\mathrm{d}s}{\mathrm{d}t} \cong 2\mu_d (d_{ij})^2 + \kappa \nabla^2 T \tag{16.1.313}$$

can be written:

$$\rho c_p \frac{\mathrm{d}T}{\mathrm{d}t} - \frac{T}{\rho}\left(\frac{\partial \rho}{\partial T}\right)_P \frac{\mathrm{d}P}{\mathrm{d}T} \cong 2\mu_d (d_{ij})^2 + \kappa \nabla^2 T \tag{16.1.314}$$

If we assume that the variation of the density with the temperature is small, we then have at the atmospheric scale:

$$\rho_\infty c_P \frac{\mathrm{d}}{\mathrm{d}t} T \cong 2\mu_d (d_{ij})^2 + \kappa \nabla^2 T \tag{16.1.315}$$

and remembering that (material derivative):

$$\mathrm{d}_t = \vec{v} \circ \vec{\nabla} + \partial_t \tag{16.1.316}$$

It finally comes:

$$\boxed{\rho_\infty c_P (\vec{v} \circ \vec{\nabla} + \partial_t) T \cong 2\mu_d (d_{ij})^2 + \kappa \nabla^2 T} \tag{16.1.317}$$

We now have two important equations:

$$\begin{gathered}\rho_\infty((\vec{v} \circ \vec{\nabla})\vec{v} + \partial_t \vec{v}) = -\vec{\nabla} p + \rho_\infty \varepsilon \bar{T} g \vec{e}_z + \mu_d \vec{\nabla}^2 v \\ \rho_\infty c_P (\vec{v} \circ \vec{\nabla} + \partial_t) T \cong 2\mu_d (d_{ij})^2 + \kappa \nabla^2 T\end{gathered} \tag{16.1.318}$$

Therefore:

$$\boxed{\begin{gathered}(\vec{v} \circ \vec{\nabla})\vec{v} + \partial_t \vec{v} = -\frac{1}{\rho_\infty}\vec{\nabla} p + \varepsilon \bar{T} \vec{g} + \frac{\mu_d}{\rho_\infty}\vec{\nabla}^2 v \\ \rho_\infty c_P (\vec{v} \circ \vec{\nabla} + \partial_t) T \cong 2\mu_d (d_{ij})^2 + \kappa \nabla^2 T\end{gathered}} \tag{16.1.319}$$

16.1.9.1 Rayleigh-Bénard convection cells (Benard-Marangoni instability)

Let us now briefly examine the Rayleigh-Bénard problem, which consists of two plates limiting one fluid being more heated than the other.

Remark

The "Bénard-Marangoni instability" is a free surface configuration of Rayleigh-Bénard instability. In fact, a thin layer of fluid is placed on a horizontal plane but its upper face is not in contact with an other plane but it is a free surface in contact with the air. The temperature of the lower plane must be superior to the temperature of the air to develop instabilities.

We can then observe parallel longitudinal rolls in a film of viscous fluid (silicone oil) held between two plates at a hot temperature at the bottom and cold at the top. Here is a picture of these rolls seen from sides (but normally we don't notice it, because they only get visible when you add mark-particles):

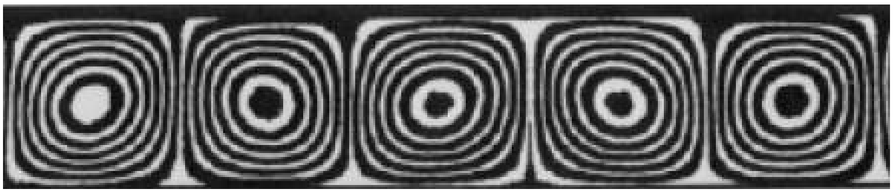


Figure 16.40: Rayleigh-Bénard convection cells (instability)

view from the top:

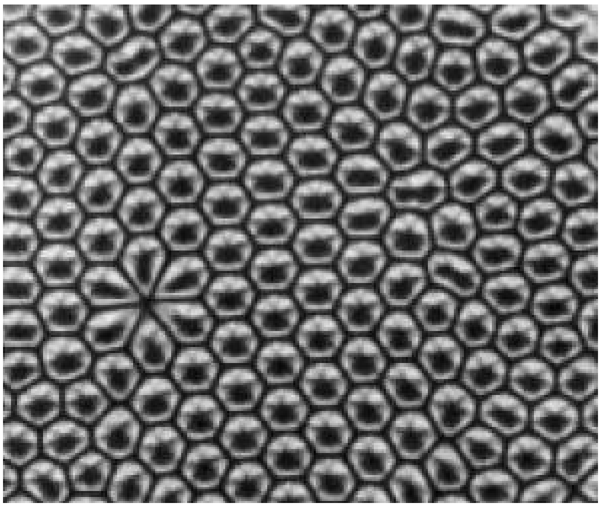


Figure 16.41: Rayleigh-Bénard convection cells viewed from top (instability)

It is a problem of natural convection: the heated fluid at the bottom expands and ascends entrained by the Archimedes principle, arrived at the top it cools and falls down. It is this movement that must be explained that is similar to that of the Earth's atmosphere, Sun (actually a G2V-type star, aka yellow dwarf) convection cells and also Earth's mantel dynamics as illustrated below:

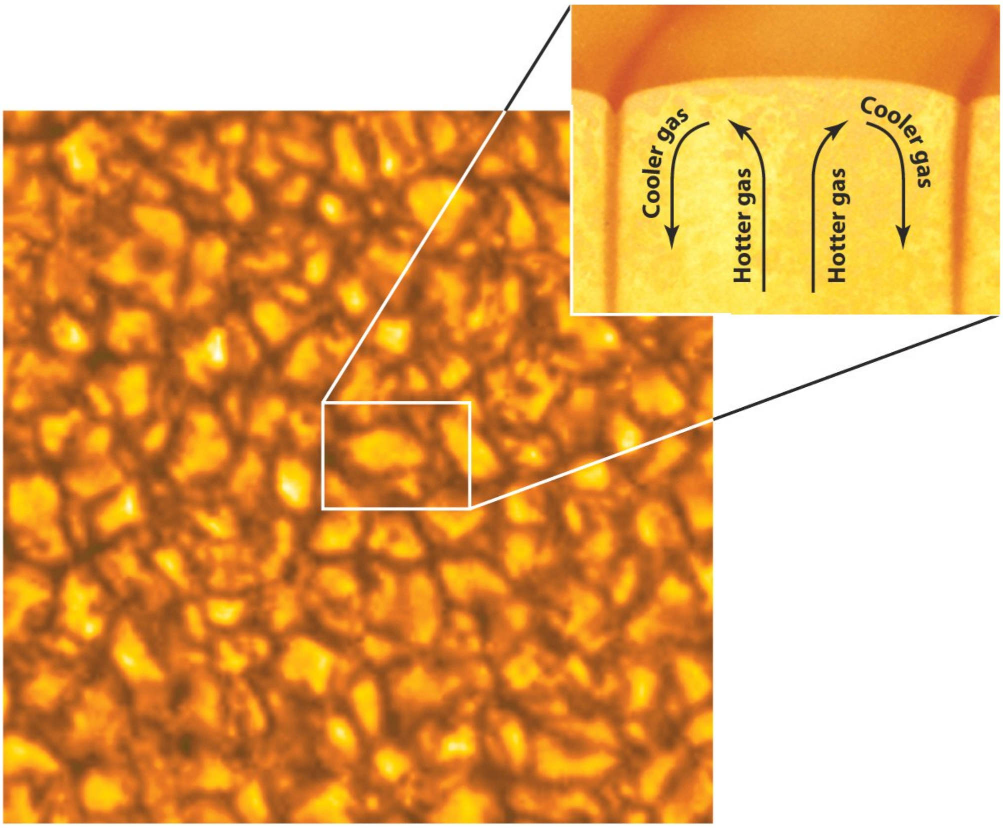

Cooler gas
Hotter gas
Hotter gas
Cooler gas

Figure 16.42: **Rayleigh-Bénard convection cells on our Sun (source: Swedish Vacuum Solar Telescope 11997-07-10)**

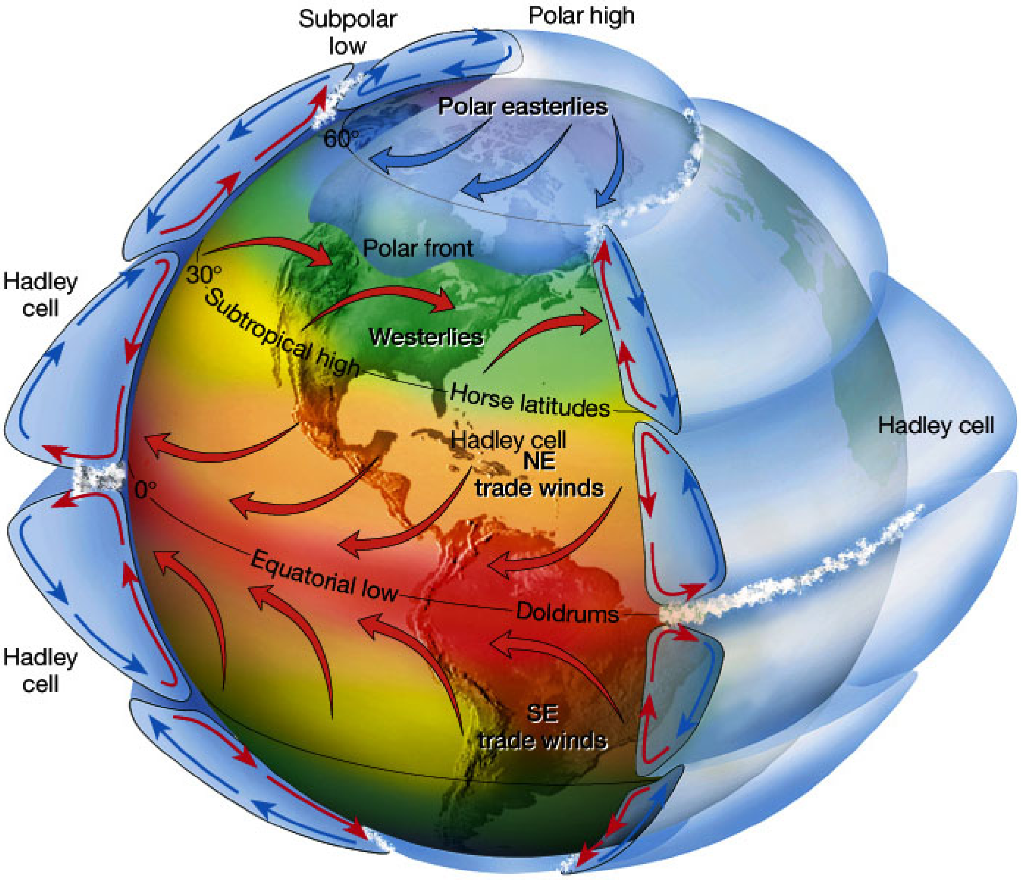

Subpolar low
Polar high
Polar easterlies
60°
Polar front
Hadley cell
30°
Subtropical high
Westerlies
Horse latitudes
Hadley cell
NE trade winds
Hadley cell
0°
Equatorial low
Doldrums
Hadley cell
SE trade winds

Figure 16.43: **Rayleigh-Bénard convection cells on Earth (source: ?)**

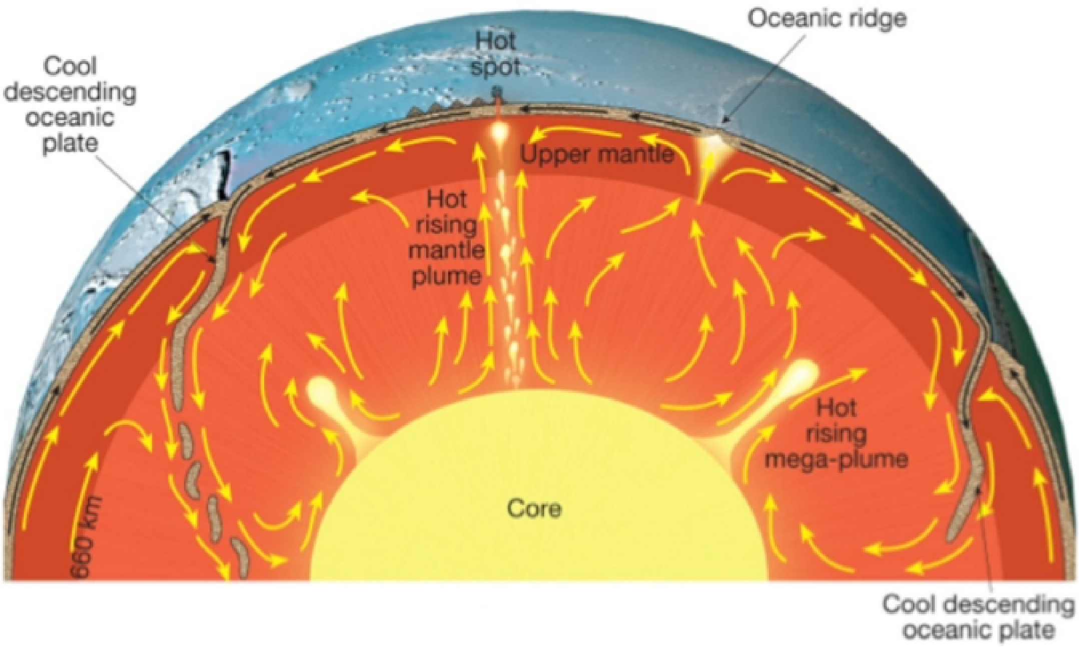

Oceanic ridge
Hot spot
Cool descending oceanic plate
Upper mantle
Hot rising mantle plume
Hot rising mega-plume
Core
660 km
Cool descending oceanic plate

Figure 16.44: **Rayleigh-Bénard convection cells as supposed in Earth mantle (source: ?)**

We also notice experimentally that the convection movements are made approximately according to a torus (see the photo seen from the side just above). We can take advantage of this symmetry to simplify the analysis.

Let us consider then a vertical loop of fluid circulating at constant speed (thus without too much turbulence ...):

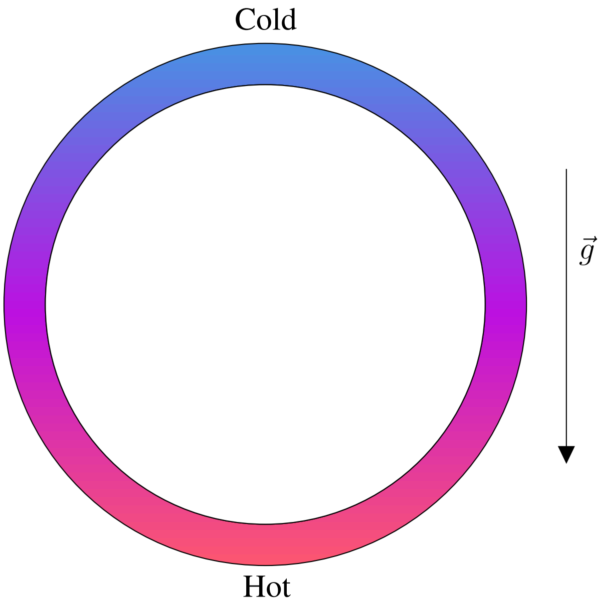

Cold
$\vec{g}$
Hot

Figure 16.45: **Vertical fluid loop with gradient in homogeneous gravific field**

The configuration will be set as follows:

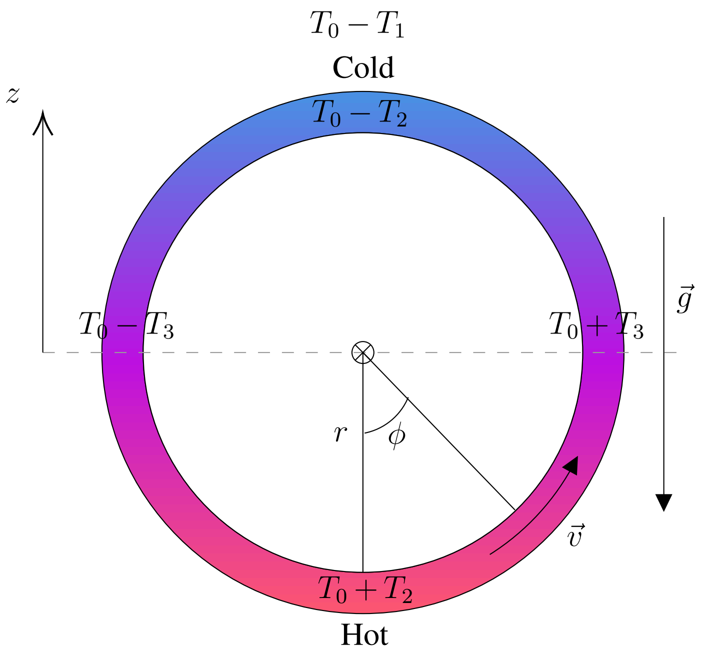


Figure 16.46: Required configuration for the theoretical model of convection cell

Where T_0 is the average temperature of the fluid (caution!: do not forget that it is not an extensive quantity!) and where we indicated respectively the temperatures inside the torus and outside it (i.e. the environment) which may all vary with time.

We see that the difference of temperature is of $2T_2$ between the top and the bottom and of $2T_3$ between the right and the left.

We put that the temperature varies linearly with the height (which of course is false in an atmospheric model ...):

$$T_E = T_0 - T_1\frac{z}{r} = T_0 + T_1\cos(\phi) \tag{16.1.320}$$

We notice that it is possible to parametrise the temperature along the inside of the torus with the following relation (parametric equation of the circle):

$$\Delta T = T - T_0 = T_2\cos(\phi) + T_3\sin(\phi) \tag{16.1.321}$$

We have then according to the figure above:

$$\begin{aligned} T(0) &= T_2\cos(0) + T_3\sin(0) = T_2 + T_0 \\ T(\pi) &= T_2\cos(\pi) + T_3\sin(\pi) = -T_2 + T_0 \\ &\ldots \end{aligned} \tag{16.1.322}$$

Having put this, let us return to:

$$(\vec{v}\circ\vec{\nabla})\vec{v} + \partial_t\vec{v} = -\frac{1}{\rho_\infty}\vec{\nabla}p + \varepsilon\bar{T}\vec{g}\vec{\nabla}^2 v \tag{16.1.323}$$

We will pass this system in polar coordinates best matching the geometry of our problem. Let us recall first that in term:

$$(\vec{v}\,'\circ\vec{\nabla})\vec{v}\,' \tag{16.1.324}$$

the differential operator $\vec{\nabla}$ is the divergence. Now, we have proved in the section of Vector Calculus (see page 1884) that this one was then written in polar coordinates:

$$\vec{\nabla}=\left(\frac{1}{r}\frac{\partial(r}{\partial r},\frac{1}{r}\frac{\partial}{\partial\phi}\right)\tag{16.1.325}$$

Now, if we rely on the hypothesis that in the volume of the torus the velocity varies neither as a function of the angle nor inside the torus (and therefore does not vary according to the radius r), then in polar coordinates :

$$(\vec{v}\,'\circ\vec{\nabla})\vec{v}\,'=0\tag{16.1.326}$$

We have then:

$$\partial_t\vec{v}=-\frac{1}{\rho_\infty}\vec{\nabla}p+\varepsilon\bar{T}\vec{g}+\frac{\mu_d}{\rho_\infty}\vec{\nabla}^2v\tag{16.1.327}$$

We will reduce the analysis to a single dimension which will be that the phenomenon depends only on the angle. We then have in polar coordinates and explicating all the terms:

$$\partial_tv=-\frac{1}{\rho_\infty}\frac{\partial}{\partial\phi}p+\varepsilon(T-T_0)\sin(\phi)-\frac{\mu_d}{\rho_\infty}\frac{1}{r^2}\frac{\partial^2}{\partial\phi^2}v\tag{16.1.328}$$

where we have at the same time made the projection along the z axis such that:

$$\vec{g}=g\vec{e}_z=g\sin(\phi)\tag{16.1.329}$$

The differential coefficient of the last term will bother us. We replace it by a coefficient Γ which we assume to be constant and which opposes the motion such as we have:

$$\partial_tv=-\frac{1}{\rho_\infty}\frac{\partial p}{\partial\phi}+g\varepsilon(T-T_0)\sin(\phi)-\Gamma v\tag{16.1.330}$$

or more explicitly:

$$\partial_tv=-\frac{1}{\rho_\infty}\frac{\partial p}{\partial\phi}+g\varepsilon(T_2\cos(\phi)+T_3\sin(\phi))\sin(\phi)-\Gamma v\tag{16.1.331}$$

We now integrate this on the whole loop as a function of ϕ. We have then:

$$\int\limits_0^{2\pi}\partial_tv'\mathrm{d}\phi=-\frac{1}{\rho_\infty}\underbrace{\int\limits_0^{2\pi}\frac{\partial p}{\partial\phi}\mathrm{d}\phi}_{=0}+g\varepsilon\int\limits_0^{2\pi}(T_2\cos(\phi)+T_3\sin(\phi))\sin(\phi)\mathrm{d}\phi-\Gamma\int\limits_0^{2\pi}v\mathrm{d}\phi\tag{16.1.332}$$

We then have the pressure term which disappears, because there is no pressure gradient along the loop. So:

$$2\pi\partial_tv=\varepsilon g\int\limits_0^{2\pi}(T_2\cos(\phi)+T_2\sin(\phi))\sin(\phi)\mathrm{d}\phi-2\pi\Gamma v\tag{16.1.333}$$

We then have for the first term (see section Differential and Integral Calculus page 1534):

$$\int\limits_0^{2\pi}\cos(\phi)\sin(\phi)\mathrm{d}\phi=\frac{1}{2}\sin^2(\phi)\Big|_0^{2\pi}=0\tag{16.1.334}$$

and for the second term:

$$\int_0^{2\pi} \sin^2(\phi)\mathrm{d}\phi = \pi \tag{16.1.335}$$

Then it remains:

$$2\pi\partial_t v = \varepsilon g \pi T_3 - 2\pi\Gamma v \tag{16.1.336}$$

Therefore:

$$\partial_t v = \varepsilon g \pi \frac{T_3}{2} - \Gamma v \tag{16.1.337}$$

We see in this equation that the motion is driven by the horizontal temperature difference T_3.

Now, let us come back to:

$$\partial_T + \vec{v} \circ \vec{\nabla} T \cong \frac{2\mu_d}{\rho_\infty c_P}(d_{ij})^2 + \frac{\kappa}{\rho_\infty c_P}\nabla^2 T \tag{16.1.338}$$

If we neglect the tangential forces inside fluid, then we have:

$$\partial_t T + \vec{v} \circ \vec{\nabla} T \cong D\nabla^2 T \tag{16.1.339}$$

where D is the thermal diffusion coefficient (see section Thermodynamics page 3595).

In polar coordinates this is reduced to:

$$\partial_t T + v\frac{1}{r}\frac{\partial}{\partial\phi}T = D\frac{1}{r}\frac{\partial^2}{\partial\phi^2}T \tag{16.1.340}$$

and we will also make another approximation:

$$\partial_t T + \frac{v}{r}\frac{\partial T}{\partial\phi} \cong KT \tag{16.1.341}$$

And we have the two relations:

$$\begin{gathered} T_E = T_0 + T_1\cos(\phi) \\ T - T_0 = T_2\cos(\phi) + T_3\sin(\phi) \end{gathered} \tag{16.1.342}$$

By subtracting:

$$T_E - T = (T_1 - T_2)\cos(\phi) - T_3\sin(\phi) \tag{16.1.343}$$

Therefore:

$$\partial_t T + \frac{v}{r}\frac{\partial T}{\partial\phi} \cong K\left[(T_1 - T_2)\cos(\phi) - T_2\sin(\phi)\right] \tag{16.1.344}$$

and also:

$$\partial_t(T_0 + T_2\cos(\phi) + T_3\sin(\phi)) + \frac{v}{r}\frac{\partial}{\partial\phi}(T_0 + T_2\cos(\phi) + T_3\sin(\phi)) \cong K(T_1 - T_2)\cos(\phi) - KT_3\sin(\phi) \tag{16.1.345}$$

Therefore:

$$\frac{\mathrm{d}T_2}{\mathrm{d}t}\cos(\phi)+\frac{\mathrm{d}T_3}{\mathrm{d}t}\sin(\phi)+\frac{v}{r}T_2\frac{\partial\cos(\phi)}{\partial\phi}+\frac{v}{r}T_3\frac{\partial\sin(\phi)}{\partial\phi}\cong K(T_1-T_2)\cos(\phi)-KT_3\sin(\phi) \tag{16.1.346}$$

After derivation:

$$\frac{\mathrm{d}T_2}{\mathrm{d}t}\cos(\phi)+\frac{\mathrm{d}T_3}{\mathrm{d}t}\sin(\phi).\frac{v}{r}T_2\sin(\phi)+\frac{v}{r}T_3\cos(\phi)\cong K(T_1-T_2)\cos(\phi)-KT_3\sin(\phi) \tag{16.1.347}$$

We group together the terms:

$$\left[\frac{\mathrm{d}T_2}{\mathrm{d}t}+\frac{vT_3}{r}\right]\cos(\phi)+\left[\frac{\mathrm{d}T_3}{\mathrm{d}t}-\frac{v}{r}T_2\right]\sin(\phi)\cong K(T_1-T_2)\cos(\phi)-KT_3\sin(\phi) \tag{16.1.348}$$

We then have the following three differential equations that govern the dynamics of the system:

$$\begin{aligned}&\frac{\mathrm{d}T_2}{\mathrm{d}t}+\frac{vT_3}{r}\cong K(T_1-T_2)\\&\frac{\mathrm{d}T_3}{\mathrm{d}t}-\frac{v}{r}T_2\cong -KT_3\\&\frac{\mathrm{d}v}{\mathrm{d}t}=\varepsilon g\frac{T_3}{2}-\Gamma v\end{aligned} \tag{16.1.349}$$

We finish the multiple simplifications by putting...:

$$T_4(t)=T_1-T_2(t) \tag{16.1.350}$$

Which gives us:

$$\begin{aligned}&\frac{\mathrm{d}T_2}{\mathrm{d}t}=\frac{\mathrm{d}(T_1-T_4(t))}{\mathrm{d}t}=-\frac{\mathrm{d}T_4}{\mathrm{d}t}\cong KT_4-\frac{vT_3}{r}\\&\frac{\mathrm{d}T_3}{\mathrm{d}t}=-KT_3+\frac{vT_2}{r}=-KT_3+\frac{vT_1}{r}-\frac{vT_4}{r}\\&\frac{\mathrm{d}v}{\mathrm{d}t}=\varepsilon g\frac{T_3}{2}-\Gamma v\end{aligned} \tag{16.1.351}$$

Putting it this in a more clean way, we get:

$$\boxed{\begin{aligned}&\frac{\mathrm{d}v}{\mathrm{d}t}=-\Gamma v+\varepsilon g\frac{T_3}{2}\\&\frac{\mathrm{d}T_3}{\mathrm{d}t}=-KT_3+\frac{vT_1}{r}-\frac{vT_4}{r}\\&\frac{\mathrm{d}T_4}{\mathrm{d}t}\cong -KT_4+\frac{vT_3}{r}\end{aligned}} \tag{16.1.352}$$

Now, let introduce the following dimensionless variables:

$$X=\frac{v}{rK}\qquad Y=\frac{g\varepsilon T_3}{2r\Gamma K}\qquad Z=\frac{g\varepsilon T_4}{2r\Gamma K}\qquad t'=tK \tag{16.1.353}$$

where we can assimilate:

- X at the non-dimensional speed
- Y to the non-dimensional temperature difference between rising and falling currents
- Z to the dimensionless deviation of the convection equilibrium

We then have indeed:

$$\begin{aligned}\frac{\mathrm{d}\left(\dfrac{v}{rK}\right)}{\mathrm{d}tK}&=-\frac{\Gamma}{K}\frac{v}{rK}+\frac{\Gamma}{K}\frac{g\varepsilon T_3}{2r\Gamma K}\\ \frac{\mathrm{d}\left(\dfrac{g\varepsilon T_3}{2r\Gamma K}\right)}{\mathrm{d}tK}&=-\frac{g\varepsilon T_3}{2r\Gamma K}+\frac{g\varepsilon T_1}{2r\Gamma K}\frac{v}{rK}-\frac{g\varepsilon T_4}{2r\Gamma K}\frac{v}{rK}\\ \frac{\mathrm{d}\left(\dfrac{g\varepsilon T_4}{2r\Gamma K}\right)}{\mathrm{d}tK}&\cong-\frac{g\varepsilon T_4}{2r\Gamma K}+\frac{g\varepsilon T_3}{2r\Gamma K}\frac{v}{rK}\end{aligned}\tag{16.1.354}$$

Therefore:

$$\begin{aligned}\frac{\mathrm{d}X}{\mathrm{d}t'}&=-\mathrm{Pr}X+\mathrm{Pr}Y\\ \frac{\mathrm{d}Y}{\mathrm{d}t'}&=-Y+\mathrm{Re}X-XZ\\ \frac{\mathrm{d}Z}{\mathrm{d}t'}&=-Z+YX\end{aligned}\tag{16.1.355}$$

In an even more condensed and traditional form:

$$\boxed{\begin{aligned}X'&=\mathrm{Pr}(Y-X)\\ Y'&=rX-Y-XZ\\ Z'&=-Z+YX\end{aligned}}\tag{16.1.356}$$

where we have:

$$\mathrm{Pr}=\frac{\Gamma}{K}\cong\frac{\dfrac{\mu_d}{\rho_\infty}\dfrac{1}{r}\dfrac{\partial^2}{\partial\phi}}{k\dfrac{1}{r}\dfrac{\partial^2}{\partial\phi^2}}=\frac{\mu_d}{\rho_\infty k}=\frac{\mu_d}{\rho_\infty\dfrac{\kappa}{\rho_\infty c_P}}=\frac{\mu_d c_P}{\kappa}\tag{16.1.357}$$

where we have:

$$\mathrm{Pr}=\frac{\Gamma}{K}\cong\frac{\dfrac{\mu_d}{\rho_\infty}\dfrac{1}{r}\dfrac{\partial^2}{\partial\phi}}{k\dfrac{1}{r}\dfrac{\partial^2}{\partial\phi^2}}=\frac{\mu_d}{\rho_\infty k}=\frac{\mu_d}{\rho_\infty\dfrac{\kappa}{\rho_\infty c_P}}=\frac{\mu_d c_P}{\kappa}\tag{16.1.358}$$

that correspond to the "Prandtl number" that a dimensionless number, named after the German physicist Ludwig Prandtl, defined as the ratio of "momentum diffusivity" to "thermal diffusivity" and where for recall:

- Γ is the momentum diffusivity (kinematic viscosity) in $[\text{m}^2 \cdot \text{s}^{-1}]$
- K is the thermal diffusivity in $[\text{m}^2 \cdot \text{s}^{-1}]$
- μ_d is the dynamic viscosity in $[\text{Ps}^2 \cdot \text{s}]$
- κ is the thermal conductivity in $[\text{W} \cdot \text{m}^{-1} \cdot \text{K}^{-1}]$
- c_P is the specific heat in $[\text{J} \cdot \text{kg}^{-1} \cdot \text{K}^{-1}]$

with some typical values:

Material	**Prandtl Number Value**
Molten potassium at 975 [K]	0.003
Mercury	0.015
Noble gases	0.16-0.17
Oxygen	0.63
Molten lithium at 975 [K]	0.065
Air	0.7-0.8
Gaseous ammoniac	1.38
Water	7
Seawater at 293 [K]	7.2
Seawater at 273 [K]	13.4
n-butanol	50
Engine oil	100-$40,000$
Glycerol	$1,000$
Polymer melts	$10,000$
Earth's mantle	$1 \cdot 10^{25}$

Table 16.6: Typical values for Pr (source: Wikipedia)

And:

$$\text{Re} = \frac{g\varepsilon T_1}{2r\Gamma K} = \frac{g\varepsilon T_1}{2r\Gamma K} = \frac{g\varepsilon T_1}{2r\dfrac{\mu_d}{\rho_\infty}D\left(\dfrac{1}{r}\dfrac{\partial^2}{\partial\phi^2}\right)^2} = \frac{g\varepsilon\dfrac{T_1}{2}r}{\mu_c D\left(\dfrac{\partial^2}{\partial\phi^2}\right)^2} = \frac{g\varepsilon\dfrac{T_1}{2}L^3}{\mu_c D} \tag{16.1.359}$$

which is assimilated to the "Rayleigh number" that is also a dimensionless number associated with "buoyancy-driven flow", also known as "free convection" or "natural convection". When the Rayleigh number is below a critical value for that fluid, heat transfer is primarily in the form of conduction; when it exceeds the critical value, heat transfer is primarily in the form of convection.

16.1.9.2 Lorenz attractor and chaos

The previous system of three equations is essentially the same as that of the famous "Lorenz system". With one difference (!), the (real) Lorenz system contains a factor b in the last equation (which does not change the result anyway since we still get a strange attractor in the end as we will see):

$$\boxed{\begin{aligned}\frac{\mathrm{d}X}{\mathrm{d}t'}&=-\mathrm{Pr}X+\mathrm{Pr}Y\\ \frac{\mathrm{d}Y}{\mathrm{d}t'}&=-Y+\mathrm{Re}X-XZ\\ \frac{\mathrm{d}Z}{\mathrm{d}t'}&=-bZ+YX\end{aligned}} \tag{16.1.360}$$

Pr, Re and b are strictly positive, and we often put $\mathrm{Pr}=10$, $\mathrm{Re}=28$, $b=8/3$ where the Prandtl number corresponds to the value of that of water.

The Lorenz equations describe the convection phenomena of an ideal two-dimensional fluid in a reservoir heated from below.

We see by this by this proof that contrary to the statements not proved on many Internet pages and forums that:

1. The system is by no far not simple mathematically and is very approximate
2. That there are simpler systems to study that are also chaotic (see section Population Dynamics page 7317)

The interest of the Lorenz equations lies, however, in the sensitivity to the initial conditions and the convergence of the dimensionless variables.

Let's see an example with Maple™ 4.00b:

```
>with(DEtools):
>lorenz:=diff(x(t),t) = 10*(y(t)-x(t)),diff(y(t),t) =
28*x(t)-y(t)-x(t)*z(t),
diff(z(t),t) = x(t)*y(t)-8/3*z(t);
>DEplot3d(lorenz, [x(t),y(t),z(t)], t=0..100, stepsize=0.01, [[x(0)=10,
y(0)=10, z(0)=10]], orientation=[-35,75], linecolor = t, thickness = 1);
```

This gives for the first 100 units of dimensionless time (it's a perfect example of chaos producing order!):

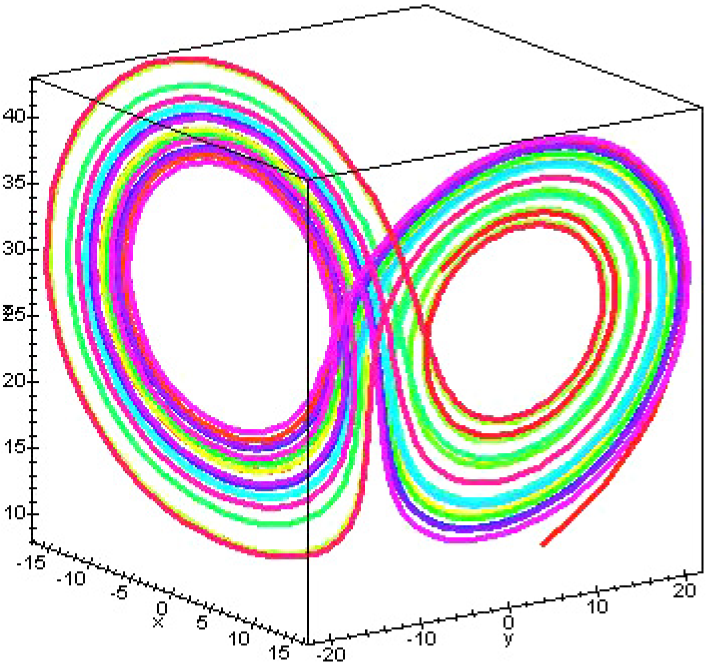


Figure 16.47: **Phase space of the Lorenz equations system in Maple™ 4.00b**

or for the first 10 units of dimensionless time:

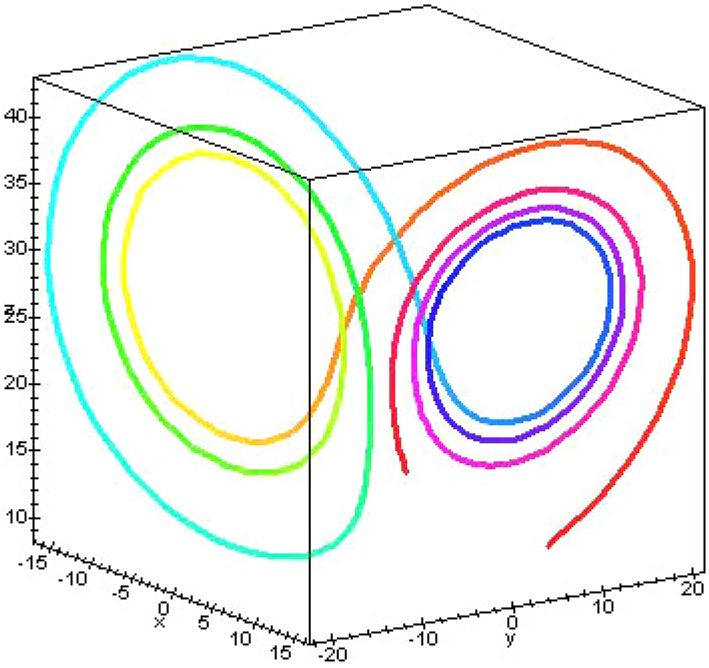


Figure 16.48: **Phase space of the Lorenz equations system in Maple™ 4.00b**

Well so far we realize that the dimensionless parameters revolve around two points named the "strange attractors".

Definition (#16.1.725): In the study of dynamic systems, an "attractor" (or "set-limit") is a set, a curve or a space towards which a system evolves irreversibly in the absence of perturbations.

Now always for the same values of adimensional time, we take $x(0)=10.1$, that is to say a relatively small change of the initial conditions. We have then:

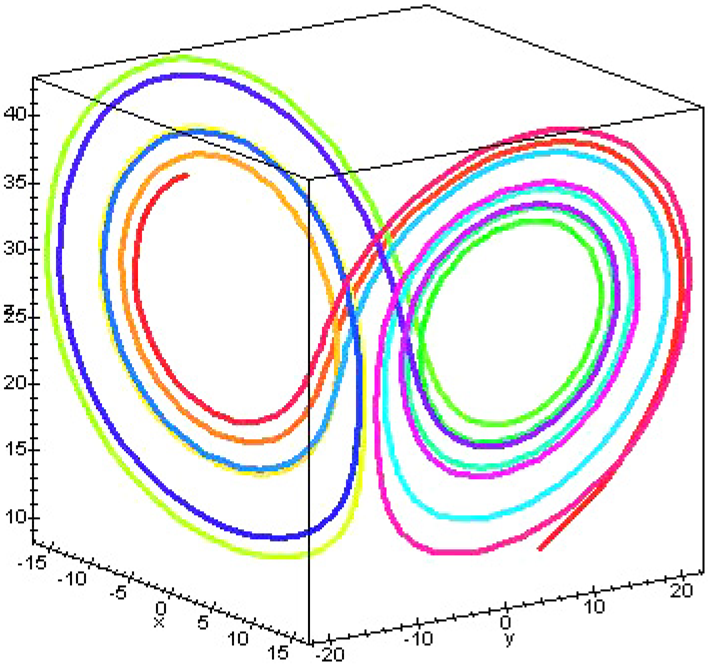


Figure 16.49: Small variation in initial conditions of the Lorenz system with Maple™ 4.00b

We thus notice that the phenomenon is no longer really similar.

Let us consider for example the variable x taking as initial conditions:

$$x(0)=10 \qquad y(0)=10 \qquad z(0)=10$$

then:

$$x(0)=10 \qquad y(0)=10.1 \qquad z(0)=10$$

thus a small variation of 0.01 on the value of $y(0)$.

Either in Maple™ 4.00b:

```
>DEplot(lorenz, [x(t), y(t), z(t)], t=0..15, stepsize = 0.01, [[x(0)=10,
y(0)=10, z(0)=10],[x(0)=10, y(0)=10.01, z(0)=10]], scene = [t,x],
linecolor = [blue,green], thickness = 1);
```

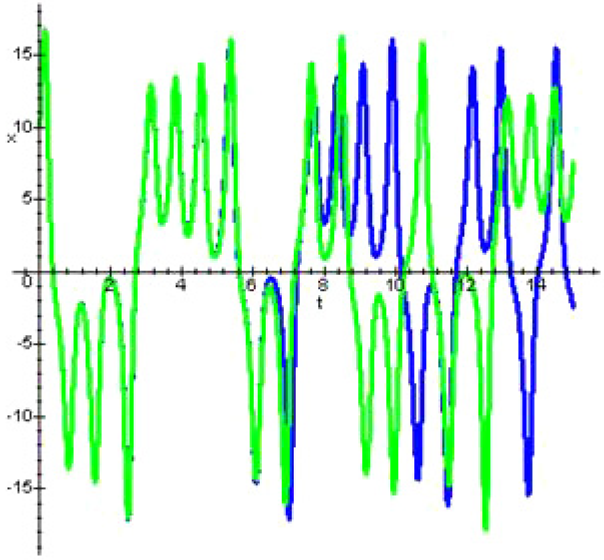


Figure 16.50: **Analysis of the dimensionless variable x for a small variation of initial conditions with Maple™ 4.00b**

We see that the system shifts rather quickly from the initial model whereas at the beginning it remains identical but the overall shape remains.

Another thing ... depending on the parameters the system can converge. Indeed, by changing the factor 28 by the value 22 we have for example (convergence on the left):

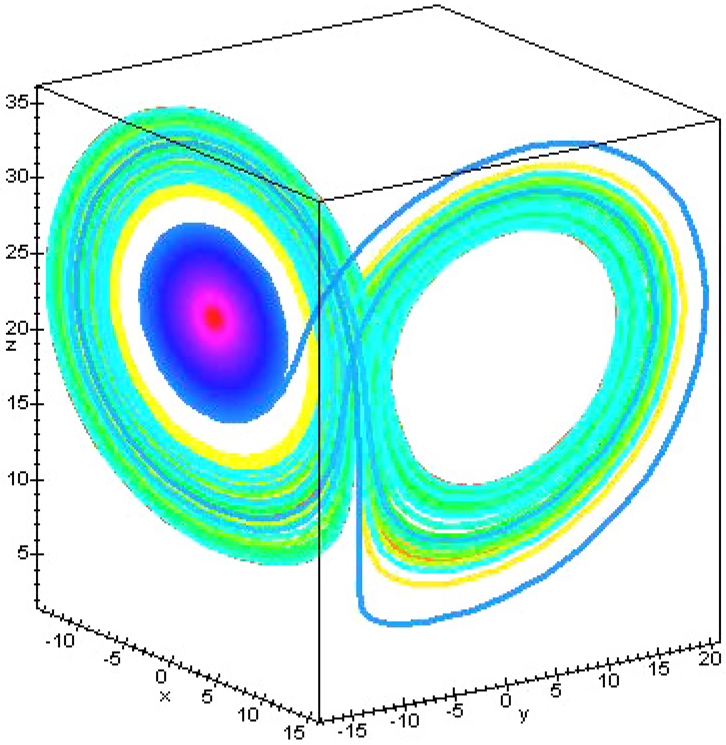


Figure 16.51: **Convergence of the Lorenz system with Maple™ 4.00b**

or with the value 19 the result is even more trivial:

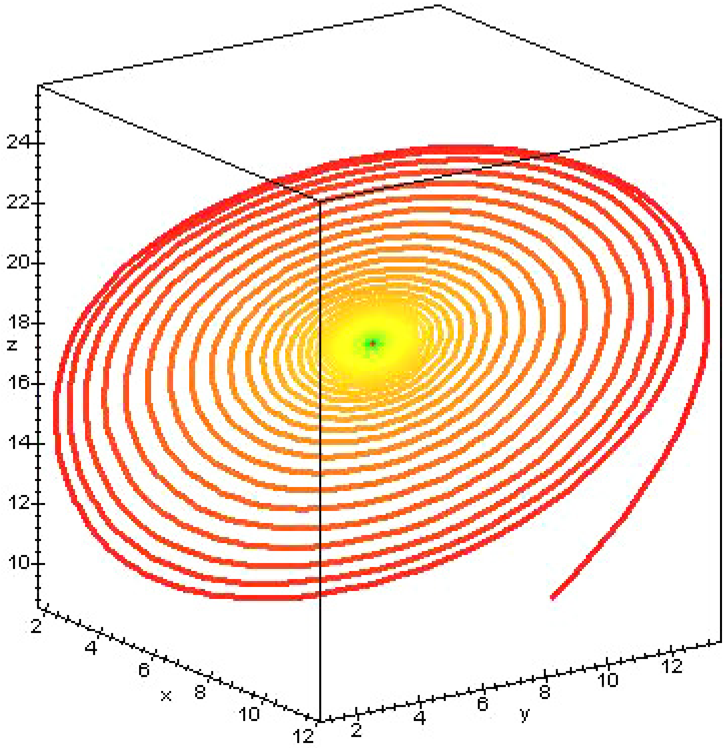


Figure 16.52: **Convergence of the Lorenz system with Maple™ 4.00b**

Or with a value close to 1:

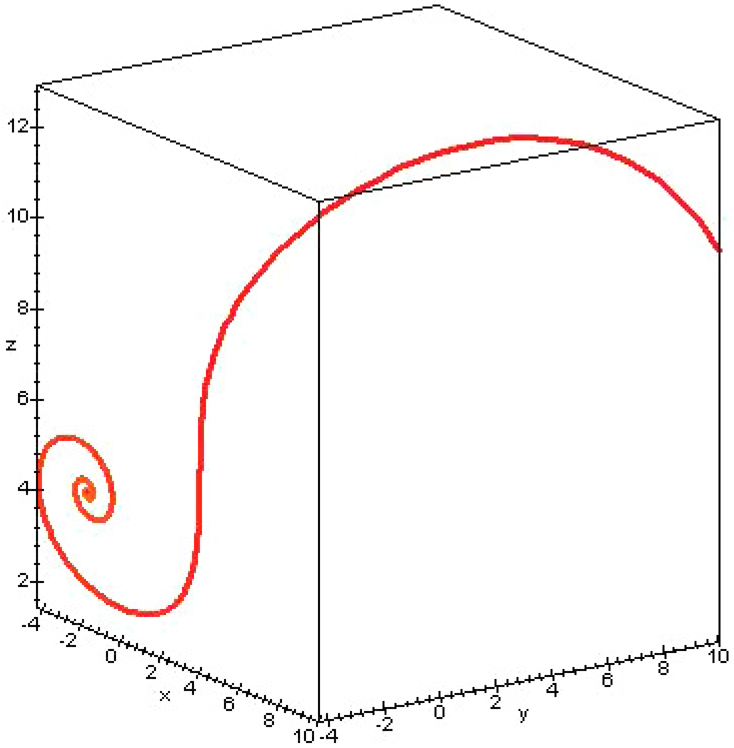


Figure 16.53: **Fast convergence of the Lorenz system with Maple™ 4.00b**

We notice a last interesting case is that if the Prandtl number is 1 then the system is stable:

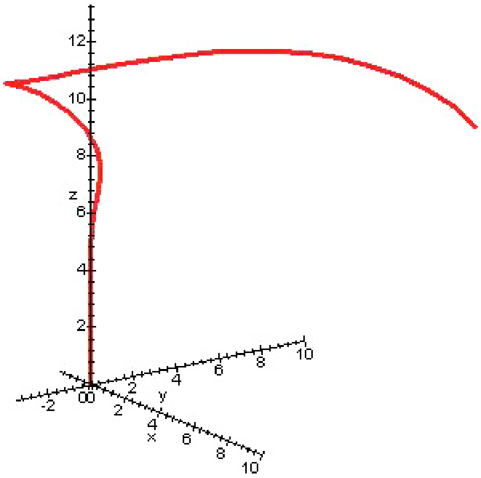
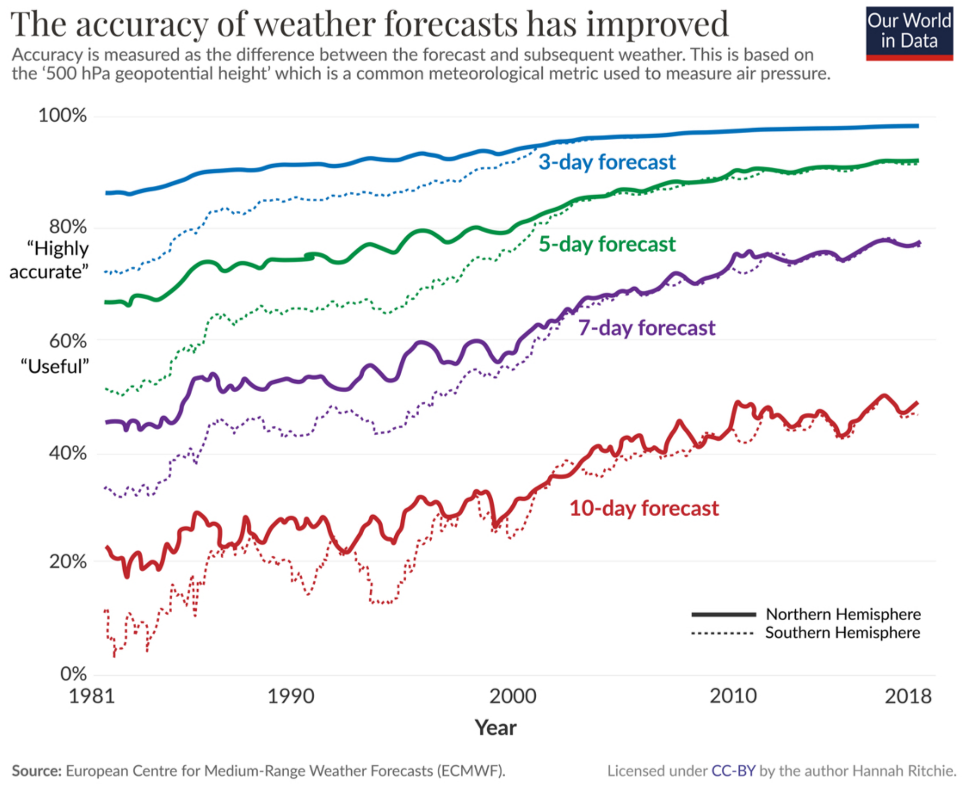


Figure 16.54: **System Stability with Maple™ 4.00b**

This sensitivity to the initial conditions, as well as the shape of the strange attractor of Lorenz, led the meteorologists to make a metaphor with the following sentence: can the flutter of a butterfly in Brazil cause a tornado in Texas?

Hence the name afterwards of "butterfly effect" for the study of the attractor of Lorenz also named in the field the "Lorenz's butterfly".

Despite the fact the weather engineering has for purpose to forecast complex multifactorial chaotic systems, thanks to the increase of computer power and satellites data survey, the accuracy of weather forecasting has significantly evolved in only four years as visible below:



Source: European Centre for Medium-Range Weather Forecasts (ECMWF).

Licensed under CC-BY by the author Hannah Ritchie.

Figure 16.55: **Evolution of weather forecasting accuracy**

The figure below from the 2013 (gregorian calendar) Intergovernmental Panel on Climate Change (IPCC) report compares the global surface warming projections made in the 1990, 1995, 2001, and 2007 (gregorian calendar) IPCC reports to the temperature measurements. Solid lines and squares represent measured average global surface temperature changes by NASA (blue), NOAA (yellow), and the UK Hadley Center (green). The colored shading shows the projected range of surface warming in the IPCC First Assessment Report (FAR; yellow), Second (SAR; green), Third (TAR; blue), and Fourth (AR4; red). We can see therefore at a global world level that the computers simulations are quite accurate also on a few decades even if the weather is a complex multifactorial chaotic system (!):

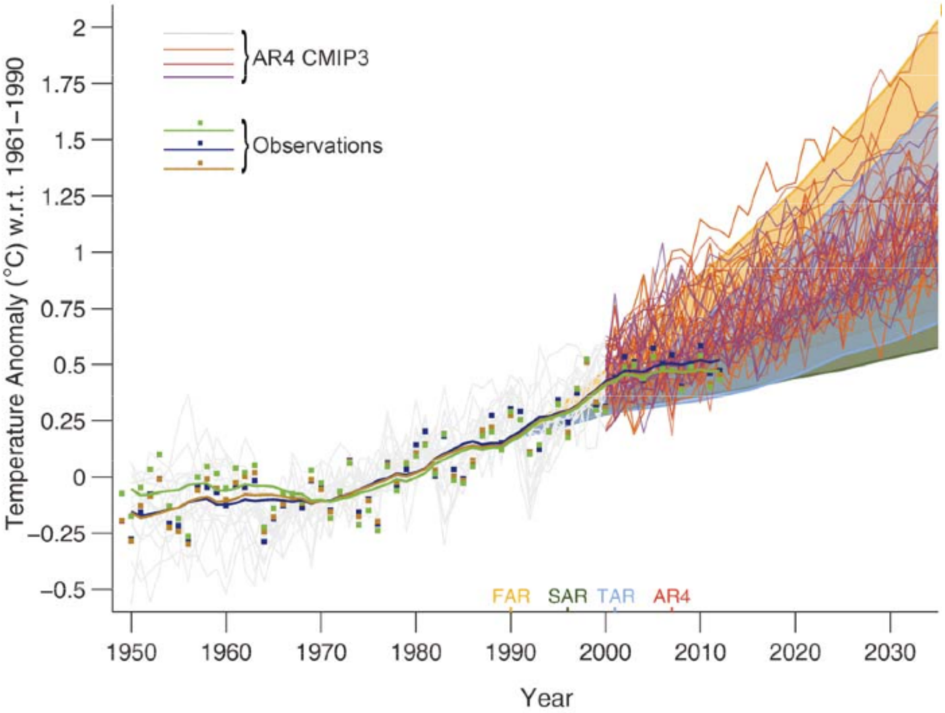
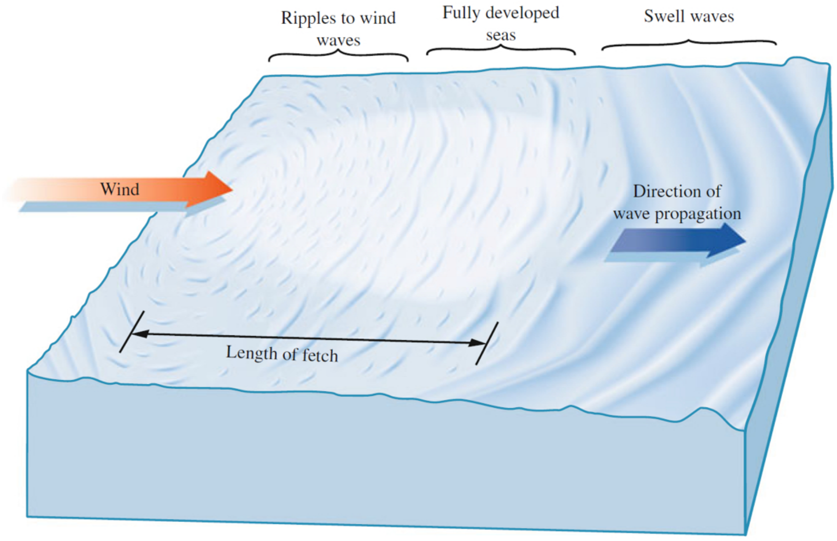


Figure 16.56: **Intergovernmental Panel on Climate Change (IPCC) forecasts vs measurements**

Obviously the forecasts can be become significanlty wrong when in the meantime the majority of humanity change its climate fingerprint.

16.1.10 Waves

When one finds himself on a boat, the movement of the water makes many people sick. It is also the case when one is in the galley to calculate this type of movement... Indeed, the waves in the water are neither transverse waves nor longitudinal waves. They are a bit of both. The water particles describe circles or ellipses, in a vertical plane parallel to the displacement of the wave. The big problem is that one of the boundary conditions is on the surface of the water and we do not know where that surface is. It is precisely one of the unknowns.



The problem is arduous enough for this calculation not to be found in books of general or specialized physics. We then tried to include it here by simplifying enormously and keeping only the simplest cases. The following calculation can be taken as an example of how physicists simplify life by making approximations and also as a masterful example of a common but complex phenomenon to be modelled.

We will not compute the wave equation. We must calculate the motion of the particles from the surface to the bottom of the water, since the movement extends to the bottom. But since particles describe ellipses, we need at least two variables per point: we need a vector field! We will therefore calculate the vector field of the velocities of the particles, but we will not calculate it directly. What we will find is a scalar potential of velocities whose gradient will be the vector field of velocities.

A final word on the waves that we will calculate. These waves are named "gravity waves" (not to be confused with "gravitational waves" in General Relativity), because it is the weight of the water that acts as a restoring force to bring down the summits and ascend the hollows.

Remark

I would like to thank Mr. Louis Peralta for his mathematical mastery and his pedagogy without which these exciting developments could never have been presented in this textbook. I am indebted to him, with his agreement, to copy / paste 99% of his course material.

We will make calculations in the simplest case: waves in a rectilinear channel of rectangular section and of constant width and depth... This reduces the dimensions of the problem to two (since the third one has an identical wave profile independently of where we are positioned in the width of the channel): the depth and direction of advance of the waves. Of course, we will neglect the friction on the bottom or sides of the canal and ... many other things.

The coordinate system chosen is then:

- x horizontal, positive in the direction of wave advancement (direction of channel length)
- y horizontal in the direction perpendicular to the advancement of the waves (direction of the width of the channel). Again, no variable will depend on y for the reasons already mentioned earlier
- z vertical and positive upwards. The zero is chosen on the surface of undisturbed water. Under water, z will therefore be negative. The depth of the channel being h, the bottom of the water will then be $z=-h$

Now, let us recall that we proved after very long developments in the section of Continuum Mechanics, that if the fluid is incompressible (divergence of the vector field of the velocities is therefore zero such that $\vec{\nabla}\circ\vec{v}=0$) and the dynamic viscosity is also ($\mu_d=0$), we then have the Euler equation of the first form (see page 3828):

$$\rho\mathrm{d}_t\vec{v}=-\vec{\nabla}p+\rho\vec{f}=-\vec{\nabla}p+\rho U=-\vec{\nabla}p-\rho gz \qquad (16.1.361)$$

because U is for recall of the scalar potential of the gravitational field.

What we will explicitly write in the following form in the case of our study of gravity waves:

$$\mathrm{d}_t\vec{v}=-\frac{1}{\rho}(\vec{\nabla}+\rho U) \qquad (16.1.362)$$

Before continuing, we will change the form of the term $\mathrm{d}_t\vec{v}$ which is a total derivative and not partial. Let us consider one of the two components (x or z) of this term. Let us take the component x of the velocity vector and use what we saw in the section of Differential and Integral Calculus (we omitted the y component since the system does not depend on it):

$$\mathrm{d}f(x,y)=\frac{\partial f}{\partial x}\mathrm{d}x+\frac{\partial f}{\partial y}\mathrm{d}y \qquad (16.1.363)$$

We have in our situation:

$$\mathrm{d}v_x(x,z,t)=\frac{\partial v_x}{\partial t}\mathrm{d}t+\frac{\partial v_x}{\partial x}\mathrm{d}x+\frac{\partial v_x}{\partial z}\mathrm{d}z \qquad (16.1.364)$$

Then we get:

$$\frac{\mathrm{d}v_x}{\mathrm{d}t}=\frac{\partial v_x}{\partial t}\frac{\mathrm{d}t}{\mathrm{d}t}+\frac{\partial v_x}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t}+\frac{\partial v_x}{\partial z}\frac{\mathrm{d}z}{\mathrm{d}t}=\frac{\partial v_x}{\partial t}+\frac{\partial v_x}{\partial x}v_x+\frac{\partial v_x}{\partial z}v_z \qquad (16.1.365)$$

The first term $\partial_t v_x$ is a linear term, whereas the other two are quadratic terms (as they include the product of a velocity by the derivative of a velocity) which are an infection (...) in all formal calculations. We shall therefore limit our calculations to cases where these quadratic terms are negligible relatively to the linear terms (implicitly implying that the speed must vary very little, otherwise this hypothesis is no longer valid). With this restriction, we then have:

$$\frac{\mathrm{d}\vec{v}}{\mathrm{d}t}\cong\frac{\partial\vec{v}}{\partial t} \tag{16.1.366}$$

We can then write:

$$\partial\vec{v}=-\frac{1}{\rho}(\vec{\nabla}p+\rho U) \tag{16.1.367}$$

We will add an additional restriction to our calculations. This restriction consists in imposing that the rotational velocity is zero:

$$\vec{\nabla}\times\vec{v}=\vec{0} \tag{16.1.368}$$

That is to say, as we have seen in the section of Continuum Mechanics, that there are no vortices in the fluid. We had already imposed that the velocity was zero in the y-direction, which prevents horizontal vortices horizontal axis of rotation). Now we eliminate the possibility of vortices with vertical axis of rotation also. It is not, in fact, too restrictive, we know that an object abandoned on the waves does not turn on itself unless it is caught in rolls. In the cases that we are going to calculate, the amplitudes are limited, we are in the linear case and, above all, not in the case of the rollers!!

We have proved in the section of Vector Calculus that when the curl of a vector variable is zero, the variable can be expressed as the gradient of a scalar potential:

$$\text{rot}\left[\overrightarrow{\text{grad}}\right]=\vec{\nabla}\times(\vec{\nabla}f)=\vec{0} \tag{16.1.369}$$

It comes then that we can write:

$$\vec{v}=\vec{\nabla}\phi \tag{16.1.370}$$

Then we get:

$$\partial_t(\vec{\nabla}\phi)=-\frac{1}{\rho}(\vec{\nabla}p+\rho U) \tag{16.1.371}$$

What can be written:

$$\vec{\nabla}\left(\partial_t\phi+\frac{1}{\rho}+gz\right)=0 \tag{16.1.372}$$

This equation acts as a wave equation, even if it does not directly describe the position or velocity of the particles. Once the equation of the function $\phi(x,z,t)$ are solved, we can deduce the velocities:

$$v_x=\frac{\partial\phi}{\partial x}\qquad v_z=\frac{\partial\phi}{\partial z} \tag{16.1.373}$$

It should not be forgotten that since the divergence of the vector field of the velocities is zero (incompressible field assumption):

$$\vec{\nabla}\circ\vec{v}=0=\vec{\nabla}(\vec{\nabla}\phi)=\vec{\nabla}^2\phi=\Delta\phi \tag{16.1.374}$$

Therefore the scalar Laplacian of ϕ is zero:

$$\Delta\phi = \frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial z^2} = 0 \qquad (16.1.375)$$

Therefore:

$$\frac{\partial^2\phi}{\partial x^2} = -\frac{\partial^2\phi}{\partial z^2} \qquad (16.1.376)$$

with $-h < z < 0$.

The solution must satisfy boundary conditions. The most obvious is that the vertical velocity must be zero at the bottom of the water:

$$v_z \underset{z=-h}{=} \frac{\partial\phi}{\partial z} = 0 \qquad (16.1.377)$$

To find the boundary conditions at the surface, we will define the variable $\xi(x,t)$ which is equal to the vertical displacement of a particle of water located on the surface of the water whose equilibrium position (in the absence of waves) is x (not to be confused with the other x...!). The height of the surface of the water, measured by reference to the level of the water without wave, will therefore be given by $\xi(x,t)$. Thus, the vertical velocity v_z at the surface of the water will be:

$$v_z \underset{z=0}{=} \frac{\partial\phi}{\partial z} = \frac{\partial\xi}{\partial t} \qquad (16.1.378)$$

This is a very subtle tip and not obvious at all to anticipate!

At the surface, we can consider that the pressure P is always equal to the atmospheric pressure for all the positions in x and in ξ. Hence, a priori nothing will change if the atmospheric pressure changes (...), and it is more convenient to consider it null. With this, the equation:

$$\vec{\nabla}\left(\partial_t\phi + \frac{1}{\rho}p + gz\right) = 0 \qquad (16.1.379)$$

Becomes:

$$\vec{\nabla}\left(\partial_t\phi + gz\right) = 0 \qquad (16.1.380)$$

From which we deduce:

$$\partial_t\phi = -gz \qquad (16.1.381)$$

If we place ourselves on the surface, which corresponds to $z = 0$, the preceding relation becomes:

$$\frac{\partial^2\phi}{\partial t^2} = -g\frac{\partial\xi}{\partial t} \qquad (16.1.382)$$

From which we deduce:

$$\partial_t\phi = -g\xi \qquad (16.1.383)$$

If we place ourselves on the surface, which corresponds to $z = 0$, the preceding relation becomes:

$$\partial_t \phi = -g\xi \tag{16.1.384}$$

and deriving this relation relatively to the time again, we get:

$$\frac{\partial^2 \phi}{\partial t^2} = -g\frac{\partial \xi}{\partial t} \tag{16.1.385}$$

Then using the relation proved just earlier above:

$$v_z \underset{z=0}{=} \frac{\partial \phi}{\partial z} = \frac{\partial \xi}{\partial t} \tag{16.1.386}$$

It comes:

$$\frac{\partial^2 \phi}{\partial t^2} = -g\frac{\partial \phi}{\partial z} \tag{16.1.387}$$

Let us sum up the conditions that the scalar potential ϕ must satisfy:

$$\begin{aligned} \frac{\partial^2 \phi}{\partial x^2} &= -\frac{\partial^2 \phi}{\partial z^2} & -h < z < 0 \\ \frac{\partial \phi}{\partial z} &= 0 & z = -h \\ \frac{\partial^2 \phi}{\partial t^2} &= -g\frac{\partial \phi}{\partial z} & z = 0 \end{aligned} \tag{16.1.388}$$

We will only look for solutions of the separable type (see section Differential and Integral Calculus page 1617). That is to say that ϕ, which is a function of x, z and t, will then be able to be written as the product of three functions, each of which depends on a single variable. Moreover, we will work only in sinusoidal regime (an intuition ... of the observation in real life) and using the formalism of the phasors (see section Wave Mechanics page 3510). Therefore, we have:

$$\phi(x, z, t) = \chi(x)\psi(z)e^{\mathrm{i}\omega t} \tag{16.1.389}$$

As usual, we will only keep the real part of the phasors of the solutions at the end.

Let us now calculate the second partial derivatives with respect to x and z:

$$\frac{\partial^2 \psi}{\partial x^2} = \chi''\psi e^{\mathrm{i}\omega t} \qquad \frac{\partial^2 \phi}{\partial z^2} = \chi\psi'' e^{\mathrm{i}\omega t} \tag{16.1.390}$$

The following equation:

$$\frac{\partial^2 \psi}{\partial x^2} = -\frac{\partial^2 \psi}{\partial z^2} \qquad -h < z < 0 \tag{16.1.391}$$

tells us that under the surface, these two expressions are equal to a given sign:

$$\chi''\psi e^{\mathrm{i}\omega t} = -\chi\psi'' e^{\mathrm{i}\omega t} \tag{16.1.392}$$

from which we deduce:

$$\frac{\chi''}{\chi}=-\frac{\psi''}{\psi}=\lambda \tag{16.1.393}$$

λ can only be a constant, since both sides of the expression depend on independent variables (which is very tricky for an observation!). This constant can take, in principle, any real or complex value.

This gives us the following system of two independent differential equations of the second order:

$$\begin{aligned}\chi''-\lambda\chi&=0\\ \psi''+\lambda\psi&=0\end{aligned} \tag{16.1.394}$$

It is therefore a system of differential equations of order 2 independent, almost identical to that solved in the section of Thermodynamics when we studied the equation of heat.

That being said, we do not know however the sign of the constant λ. What we know from what we have studied in the section of Differential and Integral Calculus is that if b is positive then the solution of the differential equation in z:

$$\psi''+\lambda\psi=0 \tag{16.1.395}$$

will have harmonic solutions. This would be very surprising ... and counter-intuitive with what we observe in reality. So we will impose λ as being negative and then we have seen in the section of Differential and Integral Calculus that we will have by adopting pretty much the same notations:

$$\begin{aligned}\chi&=Ae^{(\alpha+\mathrm{i}\beta)(x+\tau')}+Be^{(\alpha-\mathrm{i}\beta)(x+\tau')}\\ \psi&=Ce^{k_1(z+\tau)}+De^{k_2(z+\delta)}\end{aligned} \tag{16.1.396}$$

where we have the discriminant of the second differential equation (see section Differential and Integral Calculus page 1610) which holds:

$$k_{1,2}=\frac{-a\pm\sqrt{a^2-4b}}{2}=\pm\sqrt{\lambda} \tag{16.1.397}$$

So in other words:

$$k_1=-k_2=k \tag{16.1.398}$$

It then comes:

$$\begin{aligned}\chi&=Ae^{(\alpha+\mathrm{i}\beta)(x+\tau')}+Be^{(\alpha-\mathrm{i}\beta)(x+\tau')}\\ \psi&=Ce^{k(z+\tau)}+De^{-k(z+\delta)}\end{aligned} \tag{16.1.399}$$

and the discriminant of the first differential equation will be purely complex:

$$\alpha\pm\mathrm{i}\beta=\pm\mathrm{i}\sqrt{\lambda} \tag{16.1.400}$$

and therefore can be written:

$$\pm\mathrm{i}k \tag{16.1.401}$$

Our system is simplified even more:

$$\begin{aligned}\chi &= Ae^{\mathrm{i}k(x+\tau')}+Be^{-\mathrm{i}k(x+\tau')}\\ \psi &= Ce^{k(z+\tau)}+De^{-k(z+\delta)}\end{aligned} \tag{16.1.402}$$

If we consider the advance as zero (null phase shift) we have:

$$\begin{aligned}\chi &= Ae^{\mathrm{i}kx}+Be^{-\mathrm{i}kx}\\ \psi &= Ce^{kz}+De^{-kz}\end{aligned} \tag{16.1.403}$$

Since:

$$\phi(x,z,t)=\chi(x)\psi(z)e^{\mathrm{i}\omega t} \tag{16.1.404}$$

we have then:

$$\phi(x,z,t)=(Ce^{kz}+D^{-kz})(A'e^{\mathrm{i}(\omega t+kx)}+B'e^{\mathrm{i}(\omega t-kx)}) \tag{16.1.405}$$

So in the second parenthesis, we just have two waves that go in opposite directions. Let's keep one of the two to simplify the analysis (no matter what!):

$$\phi(x,z,t)=(Ce^{kz}+De^{-kz})(Ae^{\mathrm{i}(\omega t+kx)}) \tag{16.1.406}$$

To determine the constants, we will use the limit condition given above:

$$v_z \underset{z=-h}{=} = \frac{\partial\phi}{\partial z}=0 \tag{16.1.407}$$

Which in the present case gives us:

$$k(Ce^{-kh}-De^{kh})(A'e^{\mathrm{i}(\omega t+kx)})=0 \tag{16.1.408}$$

So of course we are not going to simplify totally so that k is zero otherwise it would not correspond with reality. On the other hand, we will simplify in the following way:

$$k(Ce^{-kh}-De^{kh})=0 \tag{16.1.409}$$

Which imposes in the non-trivial case that:

$$Ce^{-kh}=De^{kh} \tag{16.1.410}$$

So we can choose what we want, but if we want to simplify life by anticipating what is coming ... it would be better to take the following choice:

$$Ce^{-kh}=De^{kh}=\frac{1}{2} \tag{16.1.411}$$

Therefore:

$$C=\frac{1}{2}e^{kh} \qquad \text{and} \qquad D=\frac{1}{2}e^{-kh} \tag{16.1.412}$$

Therefore we get:

$$\phi(x,z,t)=(Ce^{kz}+De^{-kz})(A'e^{\mathrm{i}(\omega t+kx)})=\left(\frac{e^{k(z+h)}+e^{-k(z+h)}}{2}\right)(A'e^{\mathrm{i}(\omega t+kx)}) \tag{16.1.413}$$

Using the hyperbolic functions (see section Trigonometry page 2532) we have:

$$\phi(x,z,t)=\left(\frac{e^{k(z+h)}+e^{-k(z+h)}}{2}\right)(A'e^{\mathrm{i}(\omega t+kx)})=\cosh(k(z+h))(A'e^{\mathrm{i}(\omega t+kx)}) \quad (16.1.414)$$

We will continue to try to determine the constants using the second limit condition:

$$\frac{\partial^2\phi}{\partial t^2}=-g\frac{\partial\phi}{\partial z} \qquad z=0 \quad (16.1.415)$$

Then it comes (for the derivative of the hyperbolic cosine see the section of Differential and Integral Calculus at page 1475):

$$-\omega^2 A'\cosh(k(z+h))e^{\mathrm{i}(\omega t+kx)}=-gkA'\sinh(k(z+h))e^{\mathrm{i}(\omega t+kx)} \quad (16.1.416)$$

and therefore by putting $z=0$:

$$-\omega^2 A'\cosh(kh)e^{\mathrm{i}(\omega t+kx)}=-gkA'\sinh(kh)e^{\mathrm{i}(\omega t+kx)} \quad (16.1.417)$$

and simplifying:

$$-\omega^2\cosh(kh)=-gk\sinh(kh) \quad (16.1.418)$$

or written differently:

$$\tanh(kh)=\frac{\omega^2 h}{g}\frac{1}{kh}=\frac{\omega^2}{gk} \quad (16.1.419)$$

Let us recall that (see section Wave Mechanics page 3480):

$$k=\frac{\omega}{v_{\text{phase}}} \quad (16.1.420)$$

If we rearrange the above two relation we get the for irrational inviscid linear surface gravity waves the following phase velocity (remember that h is the water depth):

$$v_{\text{phase}}=\sqrt{\frac{g}{k}\tanh(kh)} \quad (16.1.421)$$

For fixed k then we note that $\tanh(hk)$ monotonically increases, hence the speed is greater for larger h. This answers the questions of why does the speed of a wave travelling in shallow water increase in deep water?

The relation:

$$\tanh(kh)=\frac{\omega^2}{gk} \quad (16.1.422)$$

does not, to my knowledge, have any solution for expressing the phase velocity directly from an analytic expression (it is therefore transcendent) as a function of the depth h and of the frequency f (implicitly the pulsation ω) of the wave without passing through a Maclaurin development (see section Sequences and Series page 1703) of the hyperbolic tangent for small values of kh such that:

$$\tanh(kh)\cong kh\Rightarrow kh\cong\frac{\omega^2 h}{g}\frac{1}{kh} \quad (16.1.423)$$

from which we derive when the depth h and k are also small by making an elementary simplification of the preceding relation:

$$\boxed{v_{\text{phase}} \cong \sqrt{gh} \qquad \text{(deep water)}} \tag{16.1.424}$$

Relation which is named "Lagrange celerity formula". For example, for a tsunami, the wavelengths λ are immense (on the order of a hundred kilometres) so the product hk is small (of the order of 0.25 in the oceans).

If, on the other hand, the depth h is very large as well as k then we have in Taylor development series (see section Sequences and Series page 1703):

$$\tanh(kh) \cong 1 \tag{16.1.425}$$

from which we derive also after an elementary simplification:

$$v_{\text{phase}} = \sqrt{\frac{g}{k}} \tag{16.1.426}$$

Relation which is named "Newton celerity formula". However we can numerically solve the relation:

$$\tanh(kh) = \frac{\omega^2 h}{g}\frac{1}{kh} \tag{16.1.427}$$

and then we have:

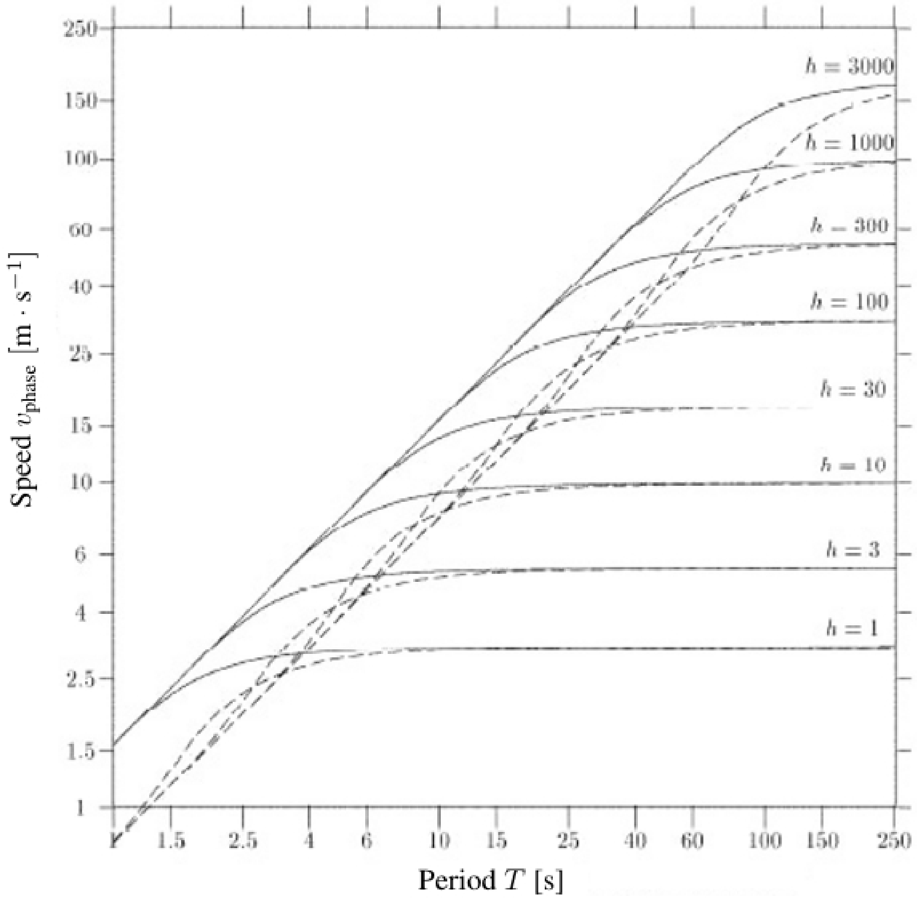


Figure 16.57: **Speed-Period gravity wave relation plot (source: LPFR)**

We can also by observing this plot understand why waves break (supercritical flux). Indeed, since the phase velocity for a given period increases as a function of the depth h, the upper part of the wave goes faster than the lower part and thus does not collapse due to lack of support. The supercritical part of the wave is destructive in the case of tsunamis because it is accelerated in its fall by gravity.

In the later case, as:

$$\tanh(kh)\cong 1\Rightarrow 1\cong\frac{\omega^2 h}{g}\frac{1}{kh} \tag{16.1.428}$$

Then we have:

$$\omega^2=gk \tag{16.1.429}$$

that is named the "dispersion relation". As k can be interpreted as the wave number, i.e.:

$$k=\frac{2\pi}{\lambda} \tag{16.1.430}$$

Hence:

$$\boxed{v_{\text{phase}}=\sqrt{\frac{g}{k}}=\sqrt{\frac{g\lambda}{2\pi}}\qquad\text{(deep water)}} \tag{16.1.431}$$

It also follows that the larger waves catch up with the small ones and that their amplitudes are superimposed over a certain distance. We will determine later the function linking the amplitude of the wave to the distance.

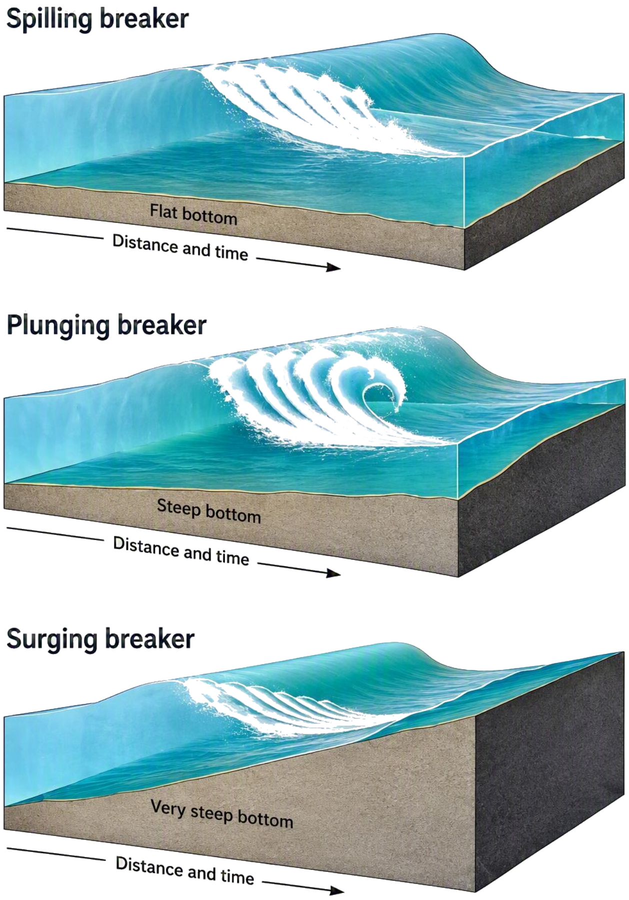

Spilling breaker
Flat bottom
Distance and time
Plunging breaker
Steep bottom
Distance and time
Surging breaker
Very steep bottom
Distance and time

Figure 16.58: **Most common type of well-known gravity wave (source: ?)**

Why does the speed of a wave travelling in shallow water increase in deep water? Let us now return to:

$$\phi(x,z,t)=\cosh(k(z+h))(A'e^{\mathrm{i}(\omega t+kx)}) \tag{16.1.432}$$

by conserving only the real part:

$$\phi(x,z,t)=A'\cosh(k(z+h))\cos(\omega t+kx) \tag{16.1.433}$$

Let us recall that we saw at the beginning:

$$v_z=\frac{\partial\phi}{\partial z} \tag{16.1.434}$$

To get Z we must therefore derive the prior-previous relation with respect to z and then integrate the result with respect to time.

The derivative thus gives:

$$v_z=\frac{\mathrm{d}z}{\mathrm{d}t}=\frac{\partial\phi}{\partial z}=kA'\sinh(k(z+h))\cos(\omega t+kx) \tag{16.1.435}$$

Therefore:

$$\mathrm{d}z=kA'\sinh(k(z+h))\cos(\omega t+kx)\mathrm{d}t \tag{16.1.436}$$

The primitive gives:

$$\int \mathrm{d}z = \int kA' \sinh(k(z+h))\cos(\omega t+kx)\mathrm{d}t = kA' \sinh(k(z+h)) \int \cos(\omega t+kx)\mathrm{d}t \tag{16.1.437}$$

Therefore:

$$Z = \frac{k}{\omega}A' \sinh(k(z+h))\sin(\omega t+kx) \tag{16.1.438}$$

And placing ourselves on $z=0$ (on the surface) we get:

$$\xi = \frac{k}{\omega}A' \sinh(kh)\sin(\omega t+kx) \tag{16.1.439}$$

and by denoting by R the amplitude of the wave:

$$\xi = \frac{k}{\omega}A' \sinh(kh)\sin(\omega t+kx) = R\sin(\omega t+kx) \tag{16.1.440}$$

With this the vertical displacement Z becomes:

$$Z = R\frac{\sinh(k(z+h))}{\sinh(kh)}\sin(\omega t+kx) \tag{16.1.441}$$

and remembering that the potential of speeds is given by:

$$\phi(x,z,t) = A' \cosh(k(z+h))\cos(\omega t+kx) \tag{16.1.442}$$

we can then rewrite it with this new constant R as following:

$$\phi(x,z,t) = R\frac{\omega}{k}\frac{\cosh(k(z+h))}{\sinh(kh)}\cos(\omega t+kx) \tag{16.1.443}$$

The horizontal velocity of a particle of water will then be:

$$v_x = \frac{\partial\phi}{\partial x} = \frac{\mathrm{d}x}{\mathrm{d}t} = -R\omega\frac{\cosh(k(z+h))}{\sinh(kh)}\sin(\omega t+kx) \tag{16.1.444}$$

and the horizontal displacement will be the integral of the previous relation with respect to time:

$$\int v_x \mathrm{d}t = -R\omega\frac{\cosh(k(z+h))}{\sinh(kh)}\sin(\omega t+kx)\mathrm{d}t = -R\omega\frac{\cosh(k(z+h))}{\sinh(kh)}\int \sin(\omega t+kx)\mathrm{d}t \tag{16.1.445}$$

Therefore:

$$X = R\frac{\cosh(k(z+h))}{\sinh(kh)}\sin(\omega t+kx) \tag{16.1.446}$$

and at the surface $z=0$:

$$X = R\frac{\cosh(kh)}{\sinh(kh)}\sin(\omega t+kx) \tag{16.1.447}$$

16.1.10.1 Depth of a wave

It is now interesting to calculate the depth of penetration of the wave. Indeed, we know from experience that the waves are less and less visible as the water height is big (excepted for Rogue Wave that are special non-linear cases).

Remember that we have just proved:

$$Z = R\frac{\sinh(k(z+h))}{\sinh(kh)}\sin(\omega t + kx) \tag{16.1.448}$$

Which corresponds to the vertical displacement of the wave. If:

$$kh = \frac{2\pi}{\lambda} \gg 1 \tag{16.1.449}$$

where we will consider that the wavelength of the waves is small and that the depth h of the medium in which they propagate is large.

We then have:

$$\sin(kh) = \frac{e^{kh} - e^{-kh}}{2} \cong \frac{e^{kh}}{2} \tag{16.1.450}$$

and:

$$\sin(k(z+1h)) = \frac{e^{k(z+h)} - e^{-k(z+h)}}{2} = \frac{e^{k(z+h)} - e^{-kz}e^{-kh}}{2} \cong \frac{e^{k(z+h)}}{2} \tag{16.1.451}$$

and we see that for this approximation to be satisfied, it is necessary that:

$$e^{-kz}e^{-kh} \ll 1 \tag{16.1.452}$$

So that the remaining part of calculations to be valid, we will impose small values of z (i.e. close to the surface).

Since then:

$$\begin{aligned} Z &\cong R\frac{\dfrac{e^{k(z+h)}}{2}}{\dfrac{e^{kh}}{2}}\sin(\omega t + kx) = R\frac{e^{k(z+h)}}{e^{kh}}\sin(\omega t + kx) = Re^{kz}\sin(\omega t + kx) \\ &= R\frac{e^{k(z+h)}}{e^{kh}}\sin(\omega t + kx) = Re^{kz}\sin(\omega t + kx) \\ &= \frac{k}{\omega}A'\sinh(kh)e^{kz}\sin(\omega t + kx) \cong \frac{k}{2\omega}A'e^{2kz+kh}\sin(\omega t + kx) \end{aligned} \tag{16.1.453}$$

Without forgetting that $-h < z < 0$ (so z is negative for recall!). The amplitude of the movements thus decreases exponentially under these conditions with the depth. The reader can verify that for some values of small kh we have:

$$\frac{\sinh(k(z+h))}{\sinh(kh)} \rightarrow -\infty \tag{16.1.454}$$

which shows the limitations of the model with this approach...

16.1.10.2 Wave's amplitude

For the amplitude of the horizontal movement the situation is the same. That is to say:

$$kh=\frac{2\pi}{\lambda}h\gg 1 \tag{16.1.455}$$

where we will consider that the wavelength of the waves is small and that the depth h of the medium in which they propagate is large.

As we have proved earlier above that:

$$X=R\frac{\cosh(k(z+h))}{\sin(kh)}\cos(\omega t+kx) \tag{16.1.456}$$

It comes then:

$$\begin{aligned}X&=R\frac{\cosh(k(z+h))}{\sinh(kh)}\cos(\omega t+kx)=R\frac{\dfrac{e^{k(z+h)}+e^{-k(z+h)}}{2}}{\dfrac{e^{kh}-e^{-kh}}{2}}\cos(\omega t+kx)\\&=R\frac{e^{k(z+h)}}{e^{kh}}\cos(\omega t+kx)\cong Re^{kz}\cos(\omega t+kx)=\frac{k}{\omega}A'\sinh(kh)e^{kz}\cos(\omega t+kz)\\&\cong\frac{k}{2\omega}A'e^{k(z+h)}\cos(\omega t+kx)\end{aligned} \tag{16.1.457}$$

where for recall, we always have $-h<z<0$ and that this model is valid only near the surface.

So for these two approximations, we have then when z is small in absolute value with respect to h (therefore in shallow water):

$$\begin{aligned}Z&\cong\frac{k}{2\omega}A'e^{kh}\sin(\omega t+kx)\\X&\cong\frac{k}{2\omega}A'e^{kh}\cos(\omega t+kx)\end{aligned} \tag{16.1.458}$$

Thus we can observe that for the case of shallow water (for z small), the water particles describe a circular motion (since the two components Z and X have the same amplitude).

On the other hand, in the general case, we have:

$$\begin{aligned}Z&\cong R\frac{\sinh(k(z+h))}{\sinh(kh)}\sin(\omega t+kx)\\X&\cong R\frac{\cosh(k(z+h))}{\sinh(kh)}\cos(\omega t+kx)\end{aligned} \tag{16.1.459}$$

The amplitudes are therefore no longer the same and the movements are then elliptic.

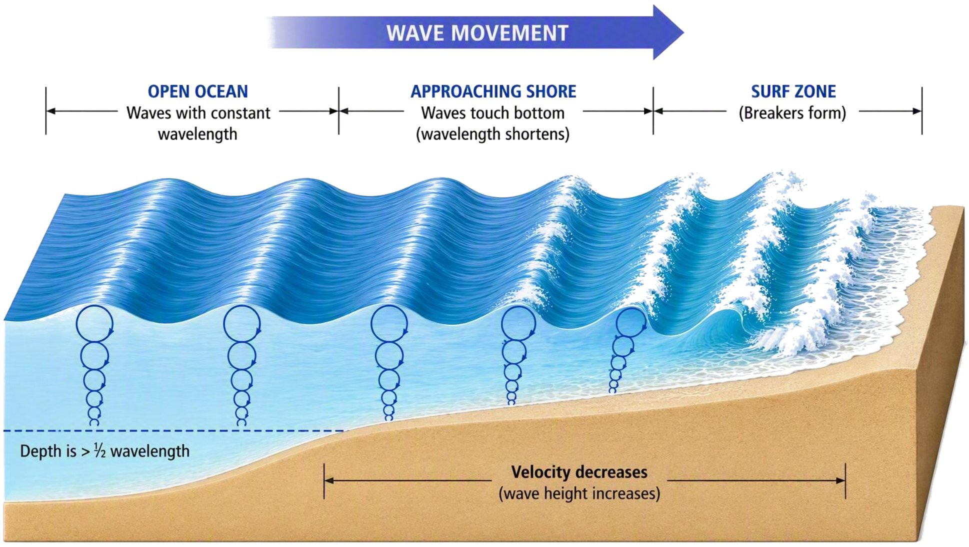

WAVE MOVEMENT
OPEN OCEAN
Waves with constant wavelength
APPROACHING SHORE
Waves touch bottom (wavelength shortens)
SURF ZONE
(Breakers form)
Depth is > ½ wavelength
Velocity decreases
(wave height increases)

Figure 16.59: Water gravity wave study summary

~70 %

★★☆☆☆

30 votes, 48.00%

16.2. Mechanical Engineering

MECHANICAL ENGINEERING is the set of knowledge related to mechanics, at the physical sense (sciences of movements) and at the technical sense (the study of mechanisms). This field of knowledge goes from design of mechanical product to the recycling of this latter through the way, of course, of production, maintenance, etc. Its applications are very important in many areas of everyday life either for the manufacture of machinery, toys, home appliances, medical devices or buildings or a variety of means of transport... and the list goes on...

Again, we will focus here only on the mathematical formalization of practical cases of current applications in the industry! So this section is only a general introduction to technical applications of mechanics and must be complemented by laboratory practices (or computer simulation with for the SimuLink™ MATLAB® toolbox and the SimDriveline™ MATLAB® toolbox) and by reading also the section of Thermodynamics where the solid state equation is treated (see page 3615) and the section of Continuum Mechanics where the Navier-Stokes equation is proved (see page 3811). The mix of mechanics and magnetism systems (AC motors for example) is treated in the section of Electrodynmics (see page 4094).

Remark

It would be pretentious to claim to do with this section as well and also as complete as two free French PDFs *Éléments de machine* and *Résistance des Matériaux* of Nicolet Gaston Raymond that are at our point of view unrivaled in content and quality at this date (compared to the same non-free books on the same subjects!). It is therefore strongly recommended to refer to them if you want to drive full information about mechanical engineering (see the download section of the companion website or directly his own website `http://www.raymondnicolet.ch`). Same recommendation applies to the reader about the master pieces that are the English books *Mechanics of Materials* ([246]) and *Structural and Stress Analysis* ([384])!

> *An engineer trying to tell a biologist that living things were designed is like a biologist trying to tell an engineer that machines are alive.*

quote
16.2:8359.#90

16.2.1 Mechanical Efficiency

Mechanical advantage is a measure of the force amplification achieved by using a tool, mechanical device or machine system. The device preserves the input power and simply trades off forces against movement to obtain a desired amplification in the output force. Machine components designed to manage forces and movement in this way are named "mechanisms".

Mechanical advantage that is computed using the assumption that no power is lost through deflection, friction and wear of a machine is the maximum performance that can be achieved. For this reason, it is often named the "ideal mechanical advantage" (IMA). In operation, deflection, friction and wear will reduce the mechanical advantage. The amount of this reduction from the ideal to the "actual mechanical advantage" (AMA) is defined by a factor named "efficiency", a quantity which is determined by experimentation.

The assumptions of an ideal machine are equivalent to the requirement that the machine does not store or dissipate energy; the power into the machine thus equals the power out. Therefore, the power P is constant through the machine and force times velocity into the machine equals the force times velocity out, that is:

$$P = F_{\text{in}} v_{\text{in}} = F_{\text{out}} v_{\text{out}} \tag{16.2.1}$$

The ideal mechanical advantage is the ratio of the force out of the machine (load) to the force into the machine (effort), or:

$$\boxed{\text{IMA} = \frac{F_{\text{out}}}{F_{\text{in}}}} \tag{16.2.2}$$

Applying the constant power relationship yields a formula for this ideal mechanical advantage in terms of the speed ratio:

$$\text{IMA} = \frac{F_{\text{out}}}{F_{\text{in}}} = \frac{v_{\text{in}}}{F_{\text{out}}} \tag{16.2.3}$$

The speed ratio of a machine can be calculated from its physical dimensions. The assumption of constant power thus allows use of the speed ratio to determine the maximum value for the mechanical advantage.

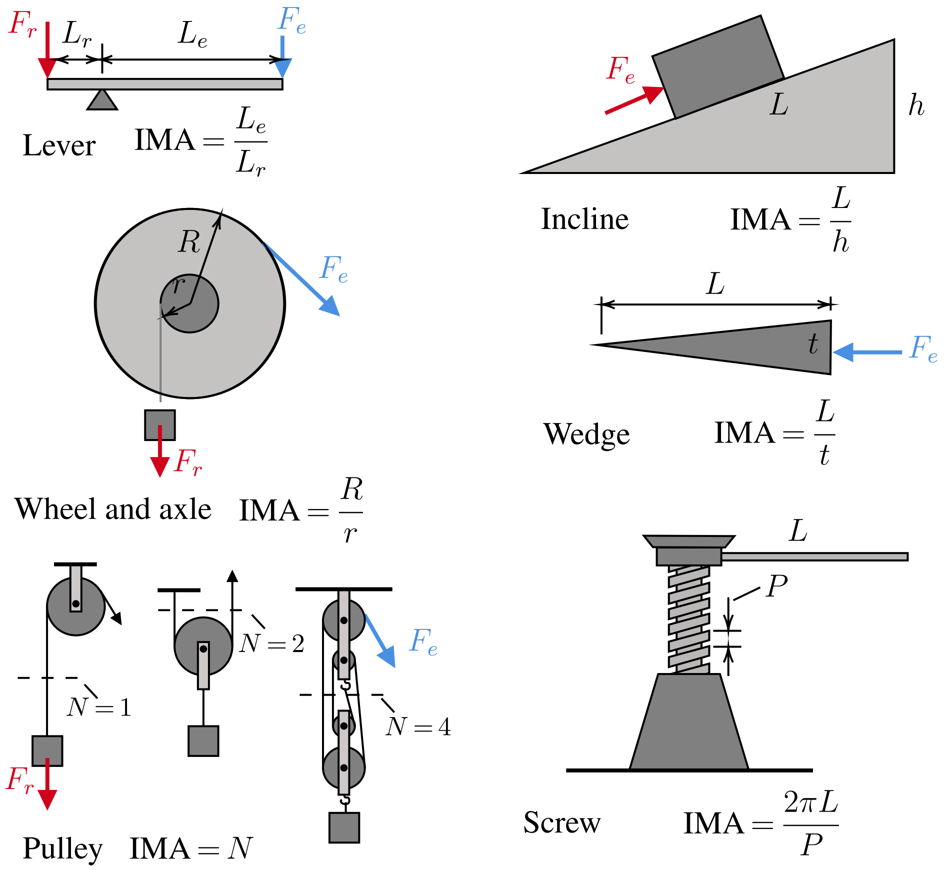

F_r
L_r
L_e
F_e
Lever
$\text{IMA} = \frac{L_e}{L_r}$
F_e
L
h
Incline
$\text{IMA} = \frac{L}{h}$
R
r
F_e
F_r
Wheel and axle
$\text{IMA} = \frac{R}{r}$
L
t
F_e
Wedge
$\text{IMA} = \frac{L}{t}$
$N = 1$
$N = 2$
$N = 4$
F_e
F_r
Pulley
$\text{IMA} = N$
L
P
Screw
$\text{IMA} = \frac{2\pi L}{P}$

Figure 16.60: Some typical ideal mechanical advantages

Most of the IMA relations above are quite obvious. That of the incline plane has however already been proved in details in the section of Classical Mechanics. That for wheel and axle and (gears in general) will be proved just further below. The IMA for pulley will be developed in the section of Civil Engineering (see page 8542).

The AMA of a machine is calculated as the ratio of the measured force output to the measured force input:

$$\text{AMA} = \frac{\hat{F}_{\text{out}}}{\hat{F}_{\text{in}}} \tag{16.2.4}$$

where the input and output forces are determined experimentally.

The ratio of the experimentally determined mechanical advantage to the ideal mechanical advantage is the mechanical efficiency η_M of the machine (to not to be confused with the thermodynamical efficiency η):

$$\boxed{\eta_M = \frac{\text{AMA}}{\text{IMA}}} \tag{16.2.5}$$

16.2.2 Gears

A "gear" is a mechanical system consisting of two intermeshing gears for the transmission of the rotational movement therebetween or for the propulsion of a fluid (when the speak about: gear pump).

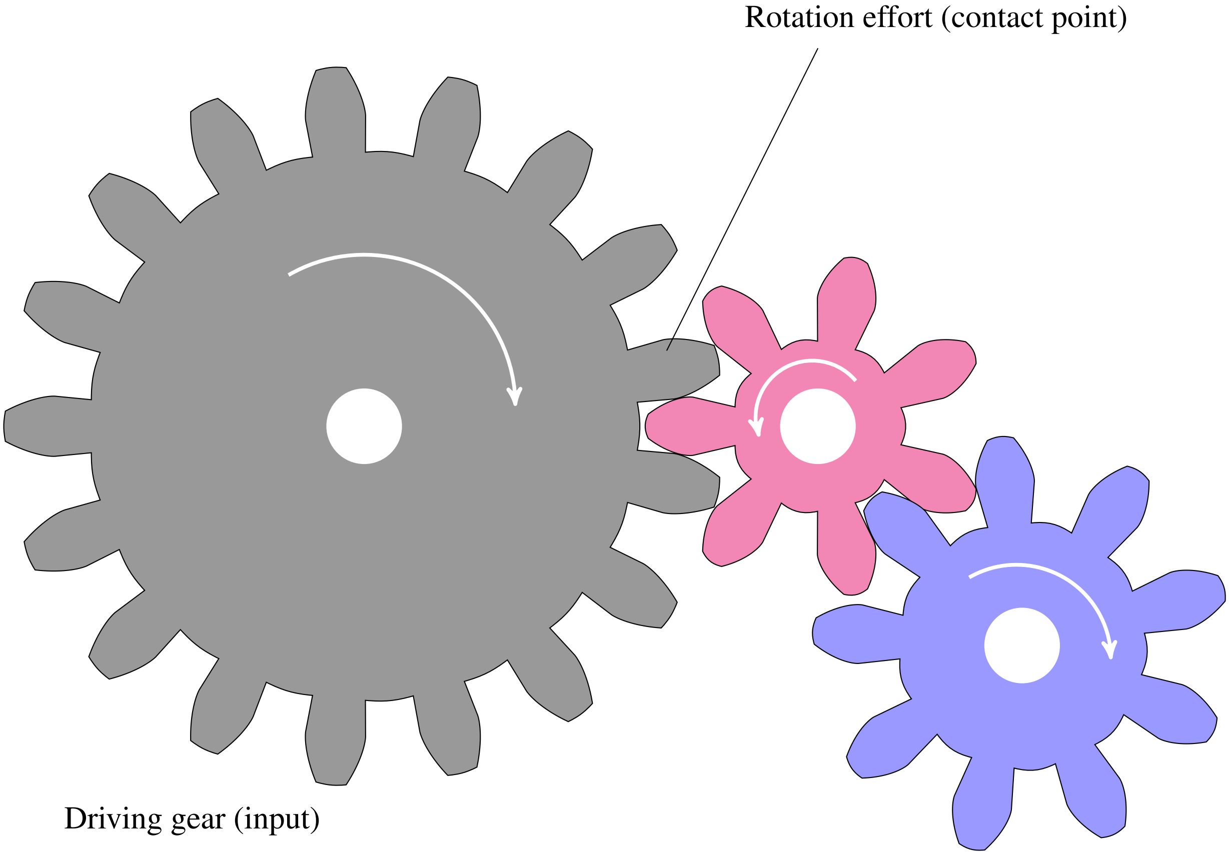


Figure 16.61: **Small gear train**

Two or more gears working in a sequence (train) are named a "gear train" or, in many cases, a "transmission"; such gear arrangements can produce a mechanical advantage through a gear ratio and thus may be considered a simple machine. Geared devices can change the speed, torque, and direction of a power source. The most common situation is for a gear to mesh with another gear; however, a gear can also mesh with a non-rotating toothed part, named a "rack", thereby producing translation instead of rotation.

The inventor of the cogwheel is considered by many practitioners as being no less than the famous mathematician, engineer and physicist Archimedes.

We find gears absolutely everywhere in our daily lives: cars, bicycles, watches, adjustable chairs, etc. We strongly recommend also the reader to think about the gears (trays and sprockets connected by a chain of transmission) of his bike to interpret the results that follow below.

The common vocabulary about a simple standard gear is given in the figure below:

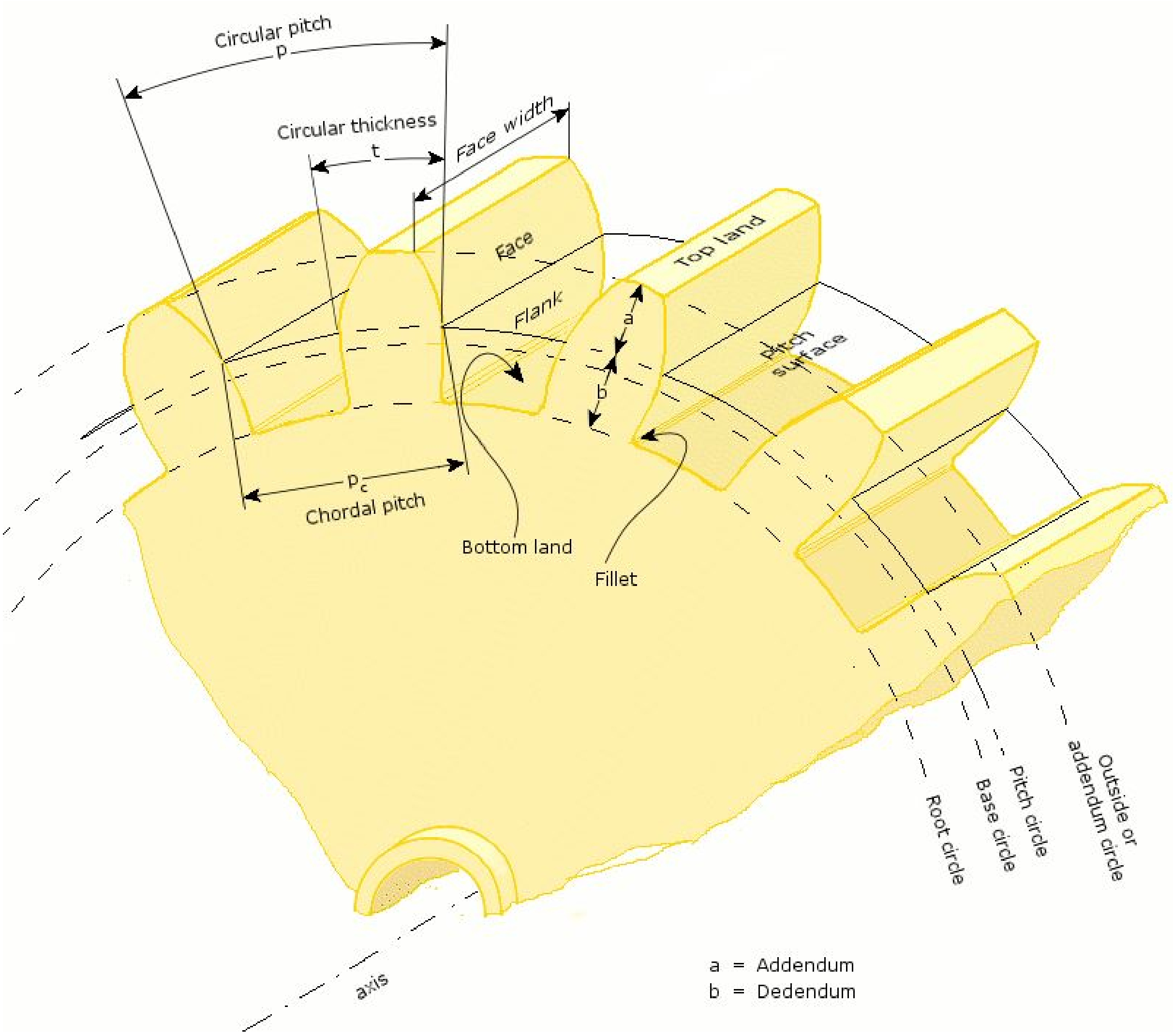

Circular pitch
p
Circular thickness
t
Face width
Face
Flank
Top land
a
b
Pitch surface
p_c
Chordal pitch
Bottom land
Fillet
Root circle
Base circle
Pitch circle
Outside or addendum circle
axis
a = Addendum
b = Dedendum

Figure 16.62: Single Gear simple definitions (source: Wikipedia)

It seemed important to us to first introduce briefly how to calculate the tooth pitch of a cylindrical wheel depending on another one for the most famous type of gears is the "spur gear" like the one shown below:

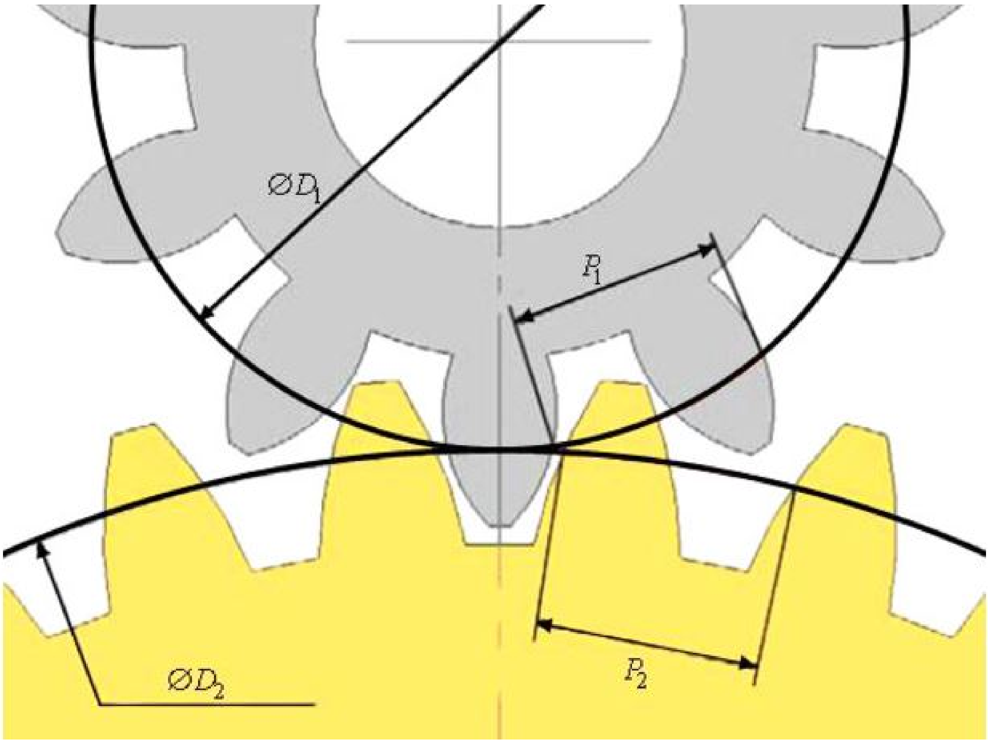


Figure 16.63: **Spur gear definitions for our analysis**

For the driving to operate so we must ensure that the teeth of a one of the cylindrical wheel of diameter $\varnothing = D_1$ well interspersed between the teeth of another wheel of diameter $\varnothing = D_2$. For this, we must introduce the concept of "circular pitch" from the teeth of each wheel which we will denote respectively by P_1 and P_2.

For this we will use the assumption that the teeth contact point can be assimilate to the diameters (represented above by black circles) that the cylindrical wheels would have if there were no slip (and therefore we would not need teeth ...) in the same way that the cog-wheel. These circles are named "primitive circles", or "primitive cylinders" or "primitive diameters".

The tooth pitch will thus be expressed as a function of the circumference and the number of teeth on each wheel. If we denote by Z_1 the number of teeth of a cylindrical gear 1 and Z_2 the number of teeth of a cylindrical gear 2, then we have the tooth pitch values that are equal respectively to:

$$\begin{aligned} P_1 &= \frac{2\pi R_1}{Z_1} = \frac{D_1}{Z_1} \\ P_2 &= \frac{2\pi R_2}{Z_2} = \frac{D_2}{Z_2} \end{aligned} \tag{16.2.6}$$

and as for the gear train to work we must have both tooth pitch to be equal such that:

$$P_1 = P_2 \tag{16.2.7}$$

Therefore we have:

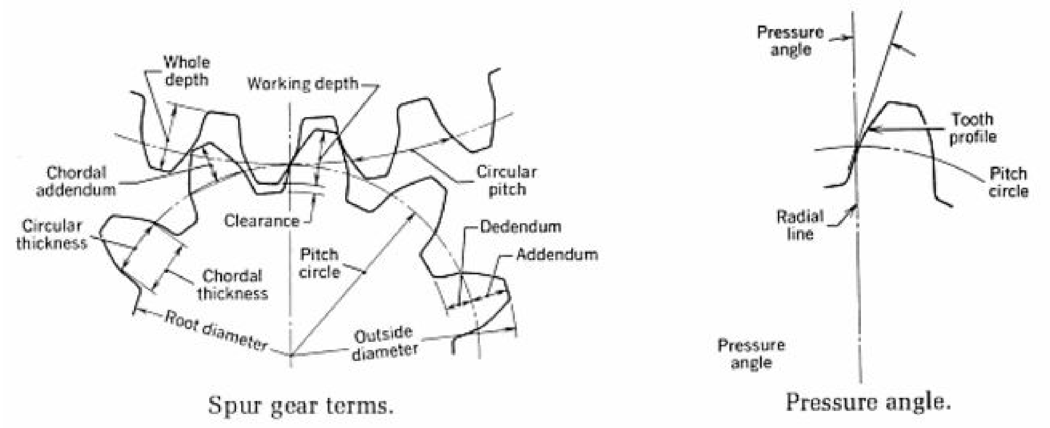
$$\frac{D_1}{Z_1} = \frac{D_2}{Z_2} = m \tag{16.2.8}$$

where m is named the "tooth module". So we will retain that:

$$D_1 = mZ_1 \qquad D_2 = mZ_2 \tag{16.2.9}$$

We also observe by the above relation that the tooth pitch is not proportional to the tooth module (the choice of a large module provides a number of small teeth and the choice of a small module a large number of teeth).

For people interested in more stuff:


Whole depth
Working depth
Chordal addendum
Circular pitch
Circular thickness
Clearance
Dedendum
Pitch circle
Addendum
Chordal thickness
Root diameter
Outside diameter
Pressure angle
Tooth profile
Pitch circle
Radial line
Pressure angle

Spur gear terms. Pressure angle.

Figure 16.64: **Spur Gear definitions**

16.2.2.1 Transmission ratios

The "transmission ratio" also named "reduction ratio" of a gear or a pulley system is very important in the field of mechanics and is defined by:

$$r = \frac{\omega_{\text{out}}}{\omega_{\text{in}}} = \frac{\dfrac{\mathrm{d}\theta_{\text{out}}}{\mathrm{d}t}}{\dfrac{\mathrm{d}\theta_{\text{in}}}{\mathrm{d}t}} = \frac{\mathrm{d}\theta_{\text{out}}}{\mathrm{d}\theta_{\text{in}}} \tag{16.2.10}$$

and therefore if the speed (pulsation) is constant (see the section of Classical Mechanics for the detailed calculations of the kinematics of the circular motion page 3280) and no longer varies between starting and nominal operation, it comes:

$$r = \frac{\omega_{\text{out}}}{\omega_{\text{in}}} = \frac{\theta_{\text{out}}}{\theta_{\text{in}}} = \frac{D_{\text{out}}}{D_{\text{in}}} \tag{16.2.11}$$

where for recall, D is the pitch circle diameter (diameter of the circle which passes through the center of all the studs, wheel bolts or wheel rim holes).

This is a widely used technology in everyday life but for people that like famous examples, we have such transmission ratios in cars (and generally almost all engines):

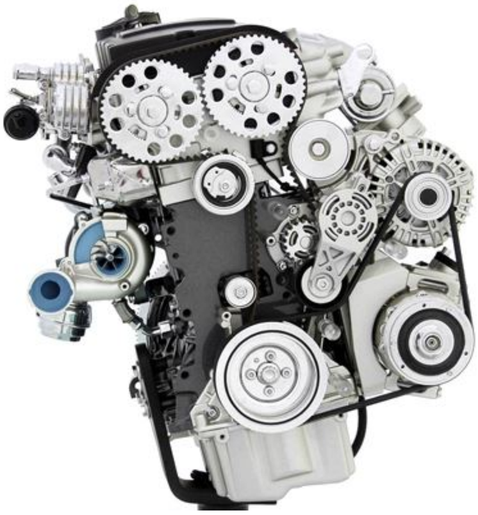


Figure 16.65: **Motor with few pulleys and drive belts**

as well as in almost all mechanical watches:

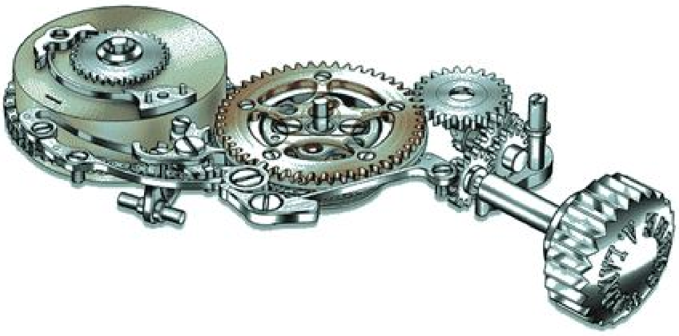


Figure 16.66: **A gear system with particular transmission ratio of a watch**

and bicycles:

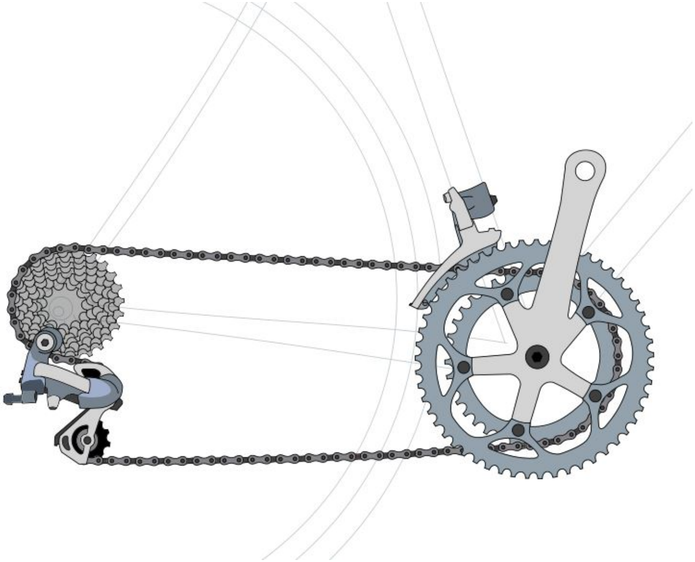


Figure 16.67: **Bicycle gears ratios**

We also of have of course:

$$r=\frac{\omega_{\text{out}}}{\omega_{\text{in}}}=\frac{\dfrac{2\pi}{T_{\text{out}}}}{\dfrac{2\pi}{T_{\text{in}}}}=\frac{T_{\text{in}}}{T_{\text{out}}} \tag{16.2.12}$$

Therefore, the ratio of the rotational speed (angular frequency) between the output gear and the input gear is equal to the ratio of the angles travelled between the output gear and the input gear or the inverse ratio of their respective time period.

We even know by our study of kinematics of the circular motion (see section Classical Mechanics page 3280):

$$\boxed{r=\frac{\omega_{\text{out}}}{\omega_{\text{in}}}=\frac{T_{\text{in}}}{T_{\text{out}}}=\frac{\dfrac{v_s}{R_s}}{\dfrac{v_e}{R_e}}} \tag{16.2.13}$$

Thus, to use the example of the bicycle ... if we want to apply the greatest moment of force possible to the rear wheel by minimizing the moment of force we provide on the pedal so the best strategy according to the relationship given above is to take the smallest possible plate associated to the largest sprocket (the ideal would be to have a plate smaller than the largest sprockets for a transmission ratio greater than one).

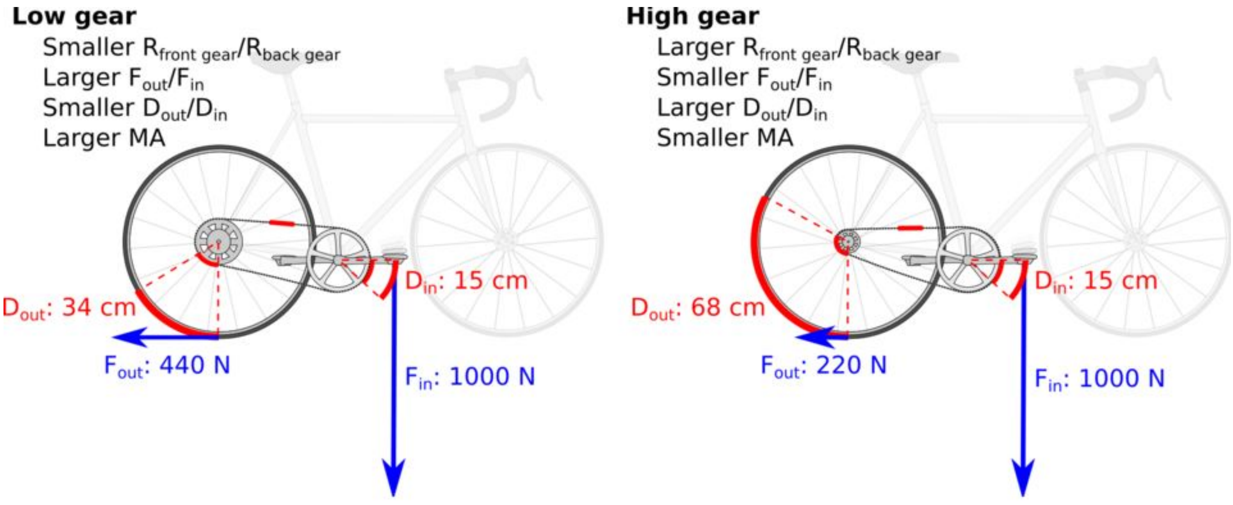

Low gear
Smaller R_front gear/R_back gear
Larger F_out/F_in
Smaller D_out/D_in
Larger MA
D_out: 34 cm
F_out: 440 N
D_in: 15 cm
F_in: 1000 N
High gear
Larger R_front gear/R_back gear
Smaller F_out/F_in
Larger D_out/D_in
Smaller MA
D_out: 68 cm
F_out: 220 N
D_in: 15 cm
F_in: 1000 N

Figure 16.68: **Gear example with bicycle (source: Wikipedia)**

We also better understand why it is recommended for countries having periods of snow to maximize the transmission factor when driving car (on snows). As the moment of force will be minimized avoiding too much slipping.

It is important then to notice the gear ratio is also useful to play with the torque (moment of force) amplitude. Indeed, in the case of a simple system of spur gear of respectively radius R_{in} (for the input gear) and R_{out} (with $R_{\text{in}}<R_{\text{out}}$), then as the torque on the input gear on the output must be equal, we then have:

$$M_{\text{in}}=M_{\text{out}}\Leftrightarrow F_{\text{in}}R_{\text{in}}=F_{\text{out}}R_{\text{out}} \tag{16.2.14}$$

Hence:

$$\boxed{F_{\text{out}} = F_{\text{in}} \frac{R_{\text{out}}}{R_{\text{in}}}} \tag{16.2.15}$$

and we see that obviously, $F_{\text{out}} > F_{\text{in}}$ if $R_{\text{out}} > R_{\text{in}}$!

Then the ideal "ideal mechanical advantage of a simple spur gears system" is obviously given by:

$$\boxed{\text{IMA}_{\text{spur gears}} = \frac{R_{\text{out}}}{R_{\text{in}}}} \tag{16.2.16}$$

💣 Caution! The principle of pulley/gear with transmission ratio is a great force multiplicator but in no case it multiplies (increase or decrease) the work or the power as a simple calculation with the above figure can be done since for the work:

$$W = F_{\text{out}} D_{\text{out}} \tag{16.2.17}$$

is the same for the both situations.

Remark

The standard ISO 1122-1 provides the inverse definition and denotes the transmission ratio i instead of r:

$$i = \frac{\omega_{\text{in}}}{\omega_{\text{out}}} = \frac{\theta_{\text{in}}}{\theta_{\text{out}}} = \frac{M_{\text{out}}}{M_{\text{in}}} \tag{16.2.18}$$

16.2.2.2 Gears association

For reasons of geometrical or mechanical constraints, it is sometimes necessary to build gear stages as shown for example with the gear train of 4 wheels below:

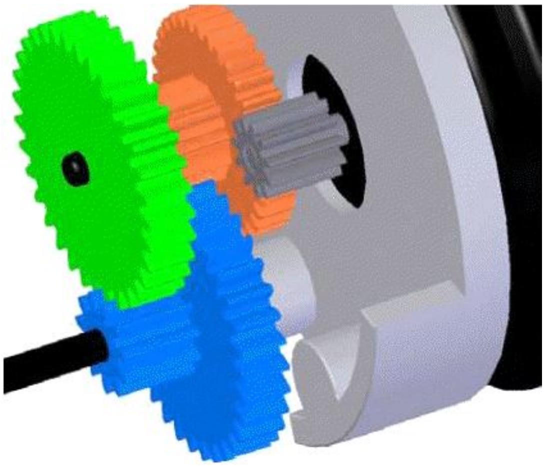


Figure 16.69: Gear train example in 3D with multiple wheels

Either into technical schematic form (not according to Swiss VSM Standards):

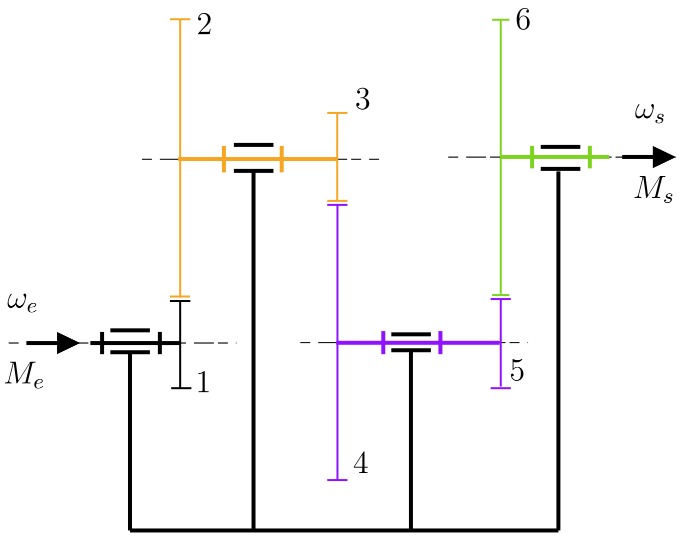


Figure 16.70: **Equivalent gear train example in 2D with multiple wheels**

The overall transmission ratio will then be given for the speed of rotation by:

$$r=r_{1-6}=\frac{\omega_{\text{out}}}{\omega_{\text{in}}}=\frac{\omega_6}{\omega_1}=\frac{\omega_2}{\omega_1}\frac{\omega_4}{\omega_3}\frac{\omega_6}{\omega_5}=\frac{\prod\omega_{\text{driven}}}{\prod\omega_{\text{leading}}} \tag{16.2.19}$$

where the fourth equality is simplified because in the case above:

$$\omega_5=\omega_4 \quad \text{and} \quad \omega_3=\omega_2 \tag{16.2.20}$$

We can also express the total transmission in terms of diameters. Since we have proven that:

$$r=\frac{\omega_{\text{out}}}{\omega_{\text{in}}}=\frac{Z_e}{Z_s} \tag{16.2.21}$$

it then comes immediately:

$$r=r_{1-6}=\frac{Z_1}{Z_2}\frac{Z_3}{Z_4}\frac{Z_5}{Z_6}=\frac{\prod Z_{\text{leading}}}{\prod Z_{\text{driven}}} \tag{16.2.22}$$

expression that we can not simplify!

Remark

Thus, in the context of a transmission of a watch with astronomical complication, the transmission ratio was obtained by determining the rational fraction $1802/217$. This being very difficult to implement with only two gears, we'll just build with three axes and four gears (with 7, 31, 34 and 53 teeth) the same ratio - which fortunately is not irreducible - as follows:

$$r=\frac{1802}{217}=\frac{34}{7}\frac{53}{31}=\frac{34}{7}\frac{53}{31}=\frac{\prod Z_{\text{leading}}}{\prod Z_{\text{driven}}} \tag{16.2.23}$$

✍**Example:**

We would like with a 4 wheel gear train on 2 axes, driven by the time axis (which spins through 12 hours), make a very precise transmission ratio r, to make a watch complication that shows a Moon phase with the classic disc having 2 Moons and turning behind a mask:

25
A. LANGE & SÖHNE
GLASHÜTTE I/SA
40
30
20
MADE IN GERMANY

As constraints the maximum number of teeth per wheel shall not exceed 300 teeth and do not fall below 7. The accuracy will be the best permitted within these numbers of teeth.

We will take 29 days 12 hours and 44 minutes for the lunar month.

The calculation of the desired ratio is relatively simple. The two Moons disc must make a full turn in $5,102,885.8$ seconds. The wheel of the timepiece makes one revolution in 12 hours, or $43,200$ seconds.

For a turn of the two Moons disc (leading axis) the axis of the 12h wheel (driven axis):

$$r=\frac{T_{\text{in}}}{T_{\text{out}}}=\frac{43,200}{5,102,588.8}\cong 0.008465798 \tag{16.2.24}$$

and rounding as below, we have an error of about 3 ten-thousandth of a second per 12 hours. Approximately 0.2 seconds per year of delay. What is acceptable for a mechanical watch.

By rounding:

$$r=\frac{T_{\text{in}}}{T_{\text{out}}}=\frac{43,200}{5,102,588.8}\cong 0.00846579 \tag{16.2.25}$$

we have an error of about 30 seconds per year, which is still acceptable for a mechanical watch. If we take again away one decimal place, then we have a delay of 6 minutes per year, even removing an additional decimal place, we would have a 49 minutes delay (which is still acceptable for many mechanical watches). By cons, beyond, this is no longer acceptable!

Now, to find the nearest decomposable rational fraction to this number there exist numerous empirical methods (by trial and error or by using a Brocot tree) and tables but the least worst ... for us ... is that using the continuous fractions (see section Number Theory page 369) when it is applicable...

Remember that we have proved that:

$$\begin{aligned}\frac{a}{b}&=q_1+\frac{1}{b/r_1}=q_1+\cfrac{1}{q_2+\cfrac{1}{r_1/r_2}}\\ &\vdots\\ &=q_1+\cfrac{1}{q_2+\cfrac{1}{q_3+\cfrac{1}{\cfrac{\ddots}{\cfrac{1}{q_{n+1}+\cfrac{1}{q_n}}}}}}\end{aligned}\tag{16.2.26}$$

considering:

$$\begin{aligned}&\frac{a}{b}=q_1+\frac{r_1}{b}\\ &\frac{b}{r_1}=q_2+\frac{r_2}{r_1}\\ &\frac{r_1}{r_2}=q_3+\frac{r_3}{r_2}\\ &\ldots\\ &\frac{r_{n-2}}{r_{n-1}}=q_n\end{aligned}\tag{16.2.27}$$

If we denote by x the ratio a/b then the relations above give us that q_1 is the integer part of x, q_2 the integer part of b/r_1 thus of $b/((x-q_1)b)=1/(x-q_1)$ either and so on...

We then have in our case:

$$\begin{aligned}
0.008465798 &\cong 0+\cfrac{1}{\cfrac{1}{0.008465798}}=0+\cfrac{1}{118.1223553881\ldots}\\
&=0+\cfrac{1}{118+\cfrac{1}{0.122355881\ldots}}=0+\cfrac{1}{118+8.17291347\ldots}\\
&=0+\cfrac{1}{118+\cfrac{1}{8+\cfrac{1}{0.17291347\ldots}}}=0+\cfrac{1}{118+\cfrac{1}{8+5.78323935\ldots}}\\
&=0+\cfrac{1}{118+\cfrac{1}{8+\cfrac{1}{5+\cfrac{1}{0.78323934\ldots}}}}=0+\cfrac{1}{118+\cfrac{1}{8+\cfrac{1}{5+1.2767489\ldots}}}\\
&=0+\cfrac{1}{118+\cfrac{1}{8+\cfrac{1}{5+\cfrac{1}{1+\ldots}}}}
\end{aligned}\tag{16.2.28}$$

Thus after we put everything to a common denominator:

$$0+\cfrac{1}{118+\cfrac{1}{8+\cfrac{1}{5+}}}=\cfrac{1}{118+\cfrac{1}{8+\cfrac{1}{6}}}=\cfrac{1}{118+\cfrac{6}{49}}=\frac{49}{5788}\tag{16.2.29}$$

If the difference (error) between this fraction and the exact value is acceptable (just have to calculate the error of time this causes after a year of operation of the watch for example), we stop ourselves here. But there is a second criterion ... the numerator and denominator must be decomposable in a satisfactory manner in relation to requirements of the number of teeth of the gears. Or in this case, the denominator (5788) can not be decomposed satisfactorily in relation to our constraints (do the decomposition with the command `ifactor` of Maple™ 4.00b and you will see!).

If we continue to develop our continued fraction, we will not find an acceptable solution before the numerator or the denominator exceeds the maximum permissible value of 300^2.

Thus, we return for example to our fraction (well this is not an acceptable approximation because in reality the maladjustment of the watch will be too fast):

$$\frac{49}{5788}\tag{16.2.30}$$

and we cheat a bit by trial and error to find a good ratio:

$$\frac{49}{5792}\tag{16.2.31}$$

Then, as we have a four gears train, we must have:

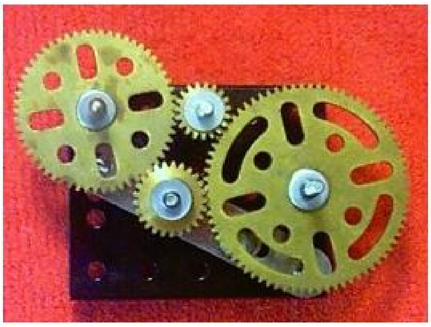
$$r \cong \frac{\prod Z_{\text{leading}}}{\prod Z_{\text{driven}}} = \frac{Z_1}{Z_2}\frac{Z_3}{Z_4} \qquad (16.2.32)$$

and then we take (always helping with the function `ifactor` of Maple™ 4.00b):

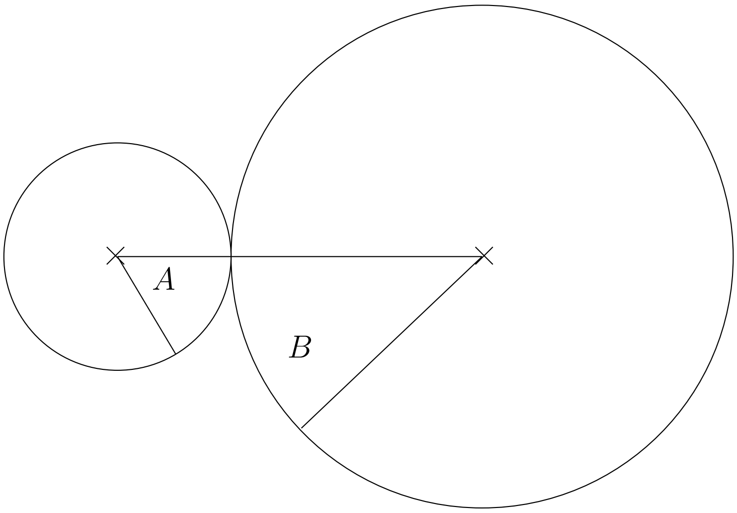
$$r \cong \frac{Z_1}{Z_2}\frac{Z_3}{Z_4} = \frac{7}{32}\frac{7}{181} \qquad (16.2.33)$$

16.2.2.2.1 Odd/Even Gear "problem"

The gear association below has a closed loop of four different gears. No two of them have the same diameter, and no two have the same number of teeth. They turn smoothly. Any even number of gears, with parallel axles, or even smooth wheels, can be put into such a loop, and they will turn freely, no matter what their tooth count or size. There's a simple way to prove this, using elementary geometry.



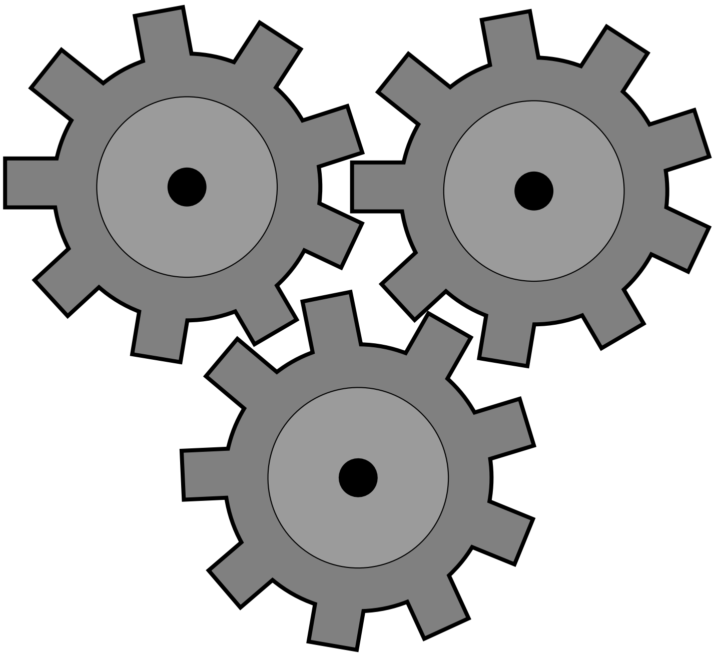
The underlying principle isn't physics, but geometry, and it applies not only to gears, but to smooth friction wheels as well. The figure shows two friction wheels of different diameter. When they turn without slipping, the two circles must turn through the same arc, but in opposite sense of rotation. That is, if the left wheel turns clockwise through arc A, the right wheel turns counter-clockwise through arc B, and $B = -A$:



If these were gears, then if one gear turns through N teeth, the other gear turns through N teeth in the other direction. It follows that if you had a string of an odd number of gears and tried to make a loop by meshing the gears at the end of the loop, their points of contact would be moving in opposite directions. If you closed a loop of an even number of gears, they would all turn quite smoothly through the same arc.

There are even exotic gears that have non-circular perimeters, such as oval or elliptical. You could make a model with an even number of these in a loop, and the gears in the model would turn smoothly.

A typical mistake is to think that the following will work:



16.2.2.3 Types of Gears

Here are some basic types of gears and how they are different from each other.

- The most common gears are "spur gears" and are used as we have seen previously in series for large gear reductions along a same plane! The teeth on spur gears are straight and are mounted in parallel on different shafts. These are particularly loud, due to the gear tooth engaging and colliding. Each impact makes loud noises and causes vibration, which is why spur gears are not used in machinery like cars.
- "Helical gears" operate more smoothly and quietly compared to spur gears due to the way the teeth interact. The teeth on a helical gear cut at an angle to the face of the gear. When two of the teeth start to engage, the contact is gradual–starting at one end of the tooth and maintaining contact as the gear rotates into full engagement. Helical is the most commonly used gear in transmissions.
- "Bevel gears" are used to change the direction of a shaft's rotation (bevel gears are most often mounted on shafts that are 90 degrees apart). Bevel gears have teeth that are avail-

able in straight, spiral, or hypoid shape. Straight teeth have similar characteristics to spur gears and also have a large impact when engaged.

- "Worm gears" are used in large gear reductions. The setup is designed so that the worm can turn the gear, but the gear cannot turn the worm. The angle of the worm is shallow and as a result the gear is held in place due to the friction between the two. The gear is found in applications such as conveyor systems in which the locking feature can act as a brake or an emergency stop.
- "Geneva gears" or "Maltese cross gears" are gears mechanism that translates a continuous rotation into an intermittent rotary motion. The rotating drive wheel has a pin that reaches into a slot of the driven wheel advancing it by one step. The drive wheel also has a raised circular blocking disc that locks the driven wheel in position between steps. In the most common arrangement, the driven wheel has four slots and thus advances by one step of 90° for each rotation of the drive wheel. If the driven wheel has n slots, it advances by $360^\circ/n$ per full rotation of the drive wheel.
- ...

In the figure below the reader can see a photo and the name of the nine most common types of gears used in engineering:

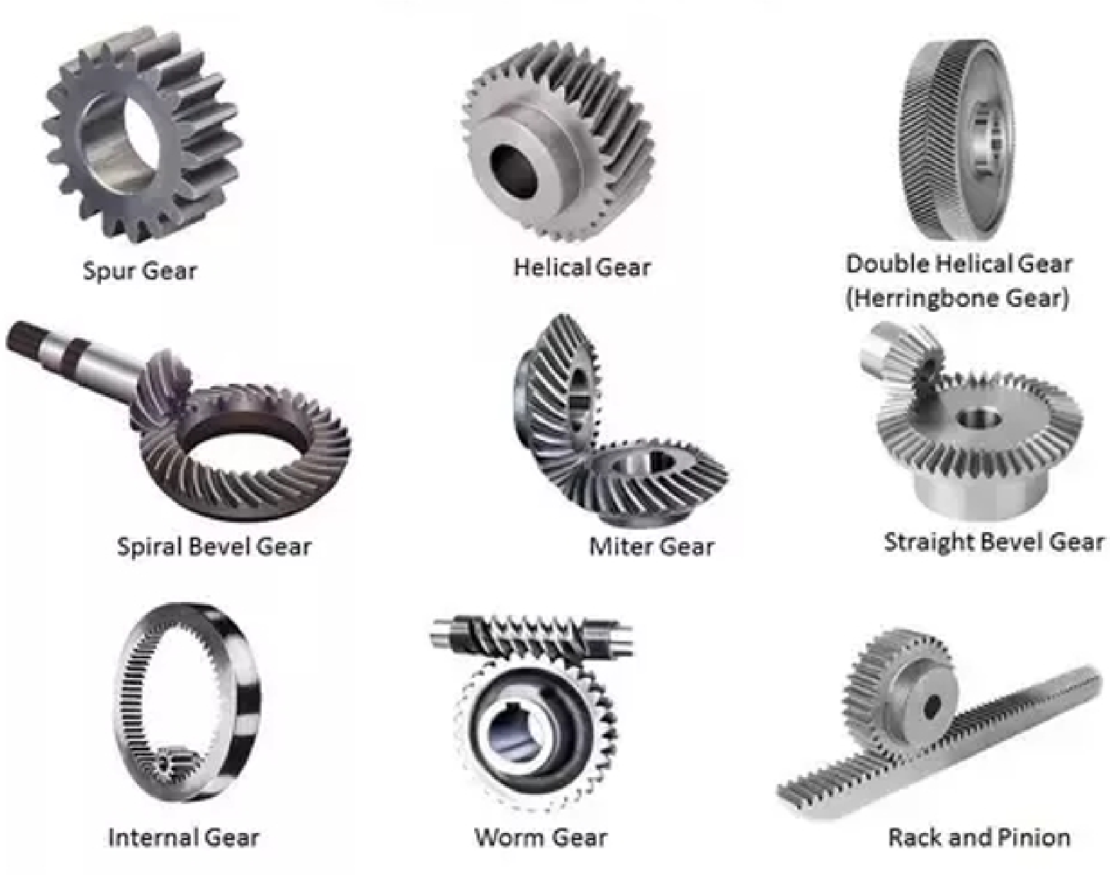


Figure 16.72: **Most common types of gears**

While manual transmissions (manually shifting a gear selector mechanism that disengages one gear and selects another) have remained relatively unchanged over the years, electronically controlled automatic, semi-automatic, and continuously variable transmissions (CVTs) have become increasingly complex, but also easier to use than ever before. However, modern transmissions of all types have become more prone to failure, primarily because of this higher level of complexity.

The Continuously Variable Transmissions (CVT) is a transmission mechanism that doesn't use gears as its means of producing various vehicle speeds at different engine speeds. Instead of gears, the system relies on a rubber or metal belt running over pulleys that can vary their effective diameters:

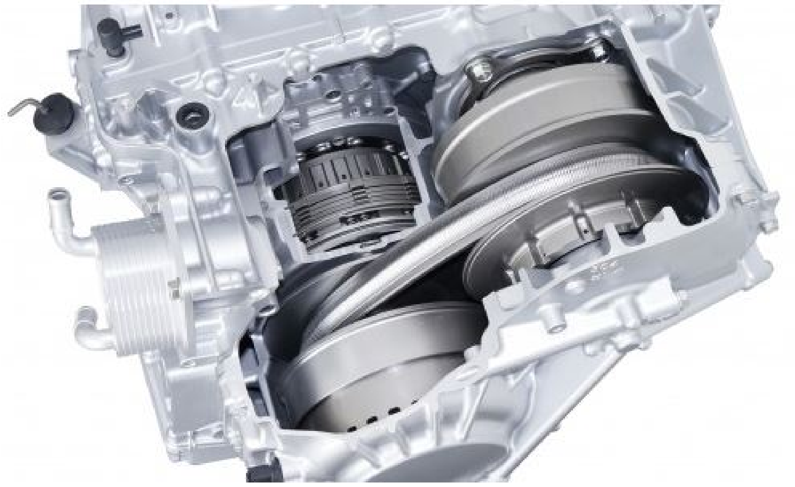


Figure 16.73: **Continuously Variable Transmissions (CVT)**

To keep the belt at its optimum tension, one pulley will increase its effective diameter, while the other decreases its effective diameter by exactly the same amount. This action is exactly analogous to the effect produced when gears of different diameters are engaged.

Now just a word about epicyclic gears!

Definition (#16.2.726): An "epicyclic gear train" or "planetary gear" consists of two gears mounted so that the center of one gear revolves around the center of the other. A carrier connects the centers of the two gears and rotates to carry one gear, named the "planet gear", around the other, named the "sun gear". The planet and sun gears mesh so that their pitch circles roll without slip. A point on the pitch circle of the planet gear traces an epicycloid curve. In this simplified case, the sun gear is fixed and the planetary gear(s) roll around the sun gear.

The epicyclic gear train below consists of a sun gear (yellow), planet gears (blue) supported by the carrier (green) and an annular gear (pink). The red marks show the relative displacement of the sun gear and carrier, when the carrier is rotated 45° clockwise and the annular gear is held fixed:

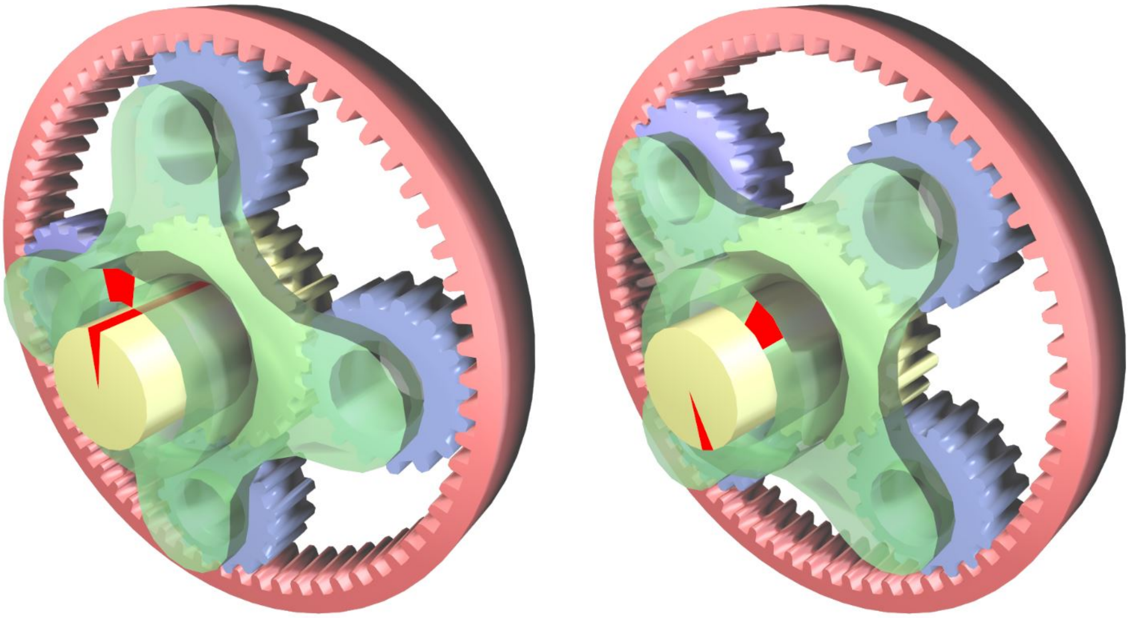


Figure 16.74: **Epicyclic gear (source: Wikipedia)**

Yes ok... but why use a epicyclic train gear excepted in watches knowing that the latter or not one of the most important thing in life!

There is one well known example: modern cars using CVT (and not only) have epicyclic gears and also almost all bicycles:

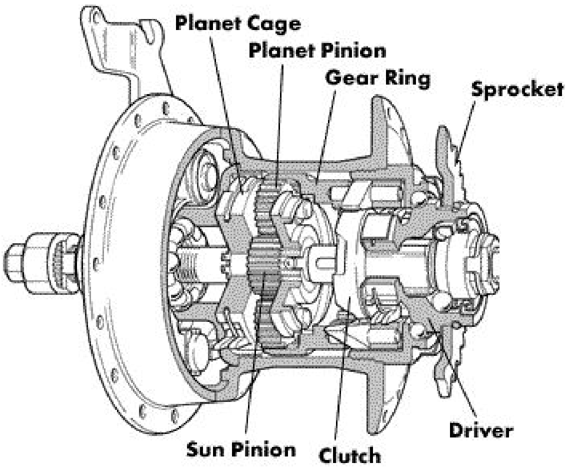


Figure 16.75: **Bicycle hub gear (special example)**

Why do CVT and bicycles use this type of gear?

First because this type of gear is very compact in surface and in volume and as forces are distributed across 4 planets gears (sometimes more!) the effort on the sun gear is divided. This is for the technical point of view...

Now for the utility point of view: Consider the figure above, where for example the sun gear is driven by the motor of the car and the annular gear is connected in one way or another to the transmission of the car. Therefore if we make the sun gear and the carrier solidary (connected) they turn together in the same direction. Then we speak of "forward gear" as the whole mechanism move as a single unique gear. If we remove what makes the sun gear and annular solidary than as it is a double epicyclic gear, the sun gear will force the annular gear to turn in the reverse direction and then we get a "reverse gear".

Remark

If the reader wants to see such a gear in action he can take a look to YouTube where there is a ton of video on this subject.

The gear ratio of an epicyclic gearing system is somewhat non-intuitive, particularly because there are several ways in which an input rotation can be converted into an output rotation. Indeed, in some hybrid vehicle transmissions, two of the components are used as inputs with the third providing output relative to the two inputs!!

OK let us now introduce some mathematical properties of this gear. For this let us consider the following figure where for simplification purposes our epicyclic has its planetary gears that are horizontal:

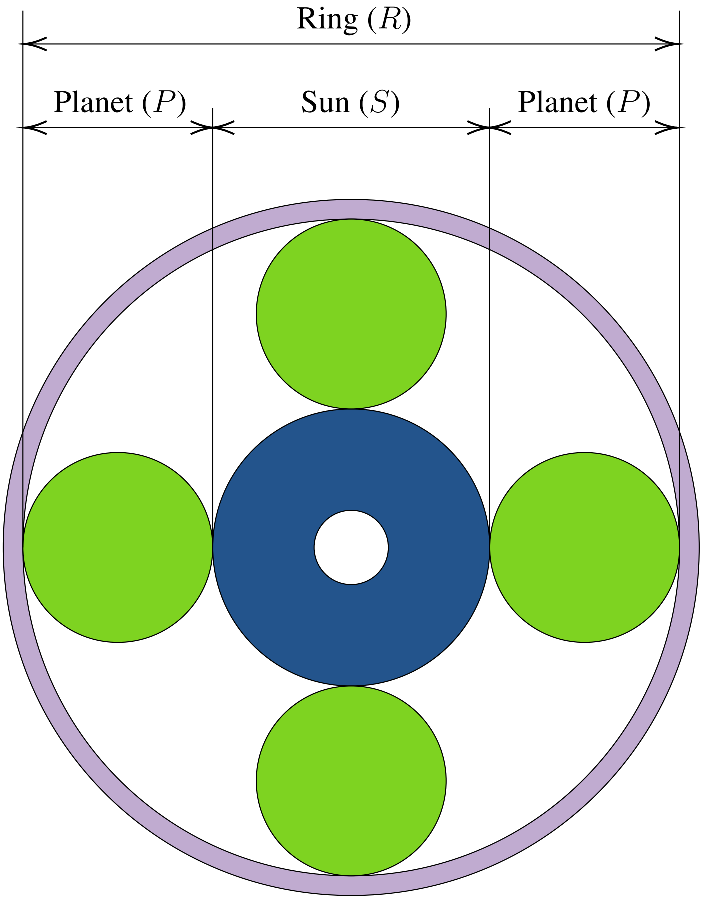


Figure 16.76: **Epicyclic gear simplified schema**

From the figure above we want to prove a canonical relation for the tooth counts. For this purpose let us denote the number of teeth of the ring gear by N_r, of the sun gear by N_s, and of the planet gears by N_p.

We will assume that our planetary gear to work out is that all teeth have the same pitch, or tooth spacing (this ensures that the teeth mesh).

The determination of a canonical between all teeth can be made more clear by imagining "gears" that just roll (no teeth), and imagine an even number of planet gears. From the illustration above we can see that the diameters of the sun gear, plus two planet gears be must equal to the ring gear size. That is:

$$\varnothing_r = 2\varnothing_p + \varnothing_s \tag{16.2.34}$$

And if we assume that all teeth have the same pitch that we will denote here by P, that is:

$$\dfrac{2\varnothing_r}{P} = N_r \qquad \dfrac{2\varnothing_p}{P} = N_p \qquad \dfrac{2\varnothing_s}{P} = N_s \tag{16.2.35}$$

Therefore injecting and simplifying in the prior-previous relation we get:

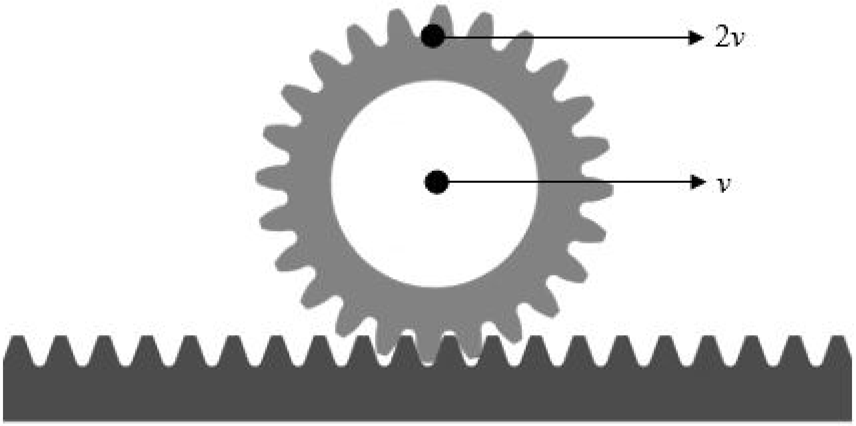
$$\boxed{N_r = 2N_p + N_s} \tag{16.2.36}$$

That is to say, the number of teeth in the ring gear is equal to the number of teeth in the middle sun gear plus twice the number of teeth in the planet gears.

We see also that the special case $N_p = 0$ brings us to the obvious result: $N_r = N_s$.

Let us denote first by $\omega_r, \omega_s, \omega_p, \omega_c$ the angular velocity of respectively the Ring (annulus) gear (pink color in the first figure), Sun gear (yellow color in the first figure), Planet gears (blue color in the first figure) and Carrier (green color in the first figure) and N_r, N_s, N_p the corresponding number of teeth of each gear (the Carrier no have teeth of sure!).

To continue let us consider the following figure:



A wheel (or gear) rolling without slipping as above means, any point on the circumference of the wheel will have both rotatory motion and translatory motion. Due to translation, if the velocity of the vehicle is v, then all the points of the wheel will have v linear velocity in the forward direction. Then due to the rotatory motion, each point will have a linear velocity of $v = r\omega$, where ω is the angular velocity. For the top most point at a particular instant, the linear velocity will be horizontal forward. So at any instant, a top most point on the wheel will have two linear velocities of same magnitude:

$$v_T = v + v = 2v \tag{16.2.37}$$

Now we have obviously (remember that r denotes here the transmission ratio!):

$$r = \frac{\omega_s}{\omega_c} \tag{16.2.38}$$

But we have immediately:

$$v_c = \omega_c(R_s + R_p) \tag{16.2.39}$$

We obviously have:

$$v_s = \omega_s R_s \tag{16.2.40}$$

But as just seen earlier above, we have:

$$v_s = 2v_c \tag{16.2.41}$$

Therefore:

$$v_s = 2\omega_c(R_s + R_p) \tag{16.2.42}$$

Hence an important relation for the planetary gear:

$$\boxed{\omega_s R_s = 2\omega_c(R_s + R_p)} \tag{16.2.43}$$

From that latter we get:

$$\begin{aligned}\frac{\omega_s}{\omega_c} &= \frac{2(R_s + R_p)}{R_s} = \frac{2R_s + 2R_p}{R_s} = \frac{2R_s + (R_r - R_s)}{R_s} \\ &= \frac{R_s + R_r}{R_s} = 1 + \frac{R_r}{R_s} = 1 + \frac{N_r}{N_s}\end{aligned} \tag{16.2.44}$$

This latter relation is in fact a special case of a more general one named the "Willis equation" or "fundamental formula of planetary gears" for the case where the ring gear is fixed (blocked).

Let us prove this equation:

Proof 16.0.1. If the sun gear stands still and the planet gear is locked firm on the carrier, then the swept angle of the carrier θ_c corresponds to the angular position of one of the planetary gear. Let us denoted this angular position θ_{p1}.

In fact, the planet gear will roll on the sun gear when mounted rotatable on the carrier and thus rotate around its own center of gravity. The planet gear will thus rotate by an additional angle θ_{p2}.

However we must not forget we have proved at the beginning of our study of simple gears systems:

$$\frac{\theta_{\text{in}}}{\theta_{\text{out}}} = \frac{D_{\text{in}}}{D_{\text{out}}} \tag{16.2.45}$$

Therefore considering sill the sun gear as fix and the carrier gear as moving, we have:

$$D_p\theta_{p2}=D_s\theta_c \tag{16.2.46}$$

Therefore:

$$\theta_{p2}=\frac{D_s}{D_p}\theta_c \tag{16.2.47}$$

Let us now consider that the carrier is held in position and the sun gear is rotated clockwise by an angle θ_s. In this case, the planet gear will turn counter-clockwise by an angle θ_{p3}. Analogous to the case before, the following statement applies (obviously the negative sign indicates that the motion of the planet gear is in the opposite direction to the motion of the sun gear):

$$D_p\theta_{p3}=-D_s\theta_s \tag{16.2.48}$$

Therefore:

$$\theta_{p3}=-\frac{D_s}{D_p}\theta_s \tag{16.2.49}$$

The motions of the planet gear according which have been considered separately so far, can now be superposed to the total motion:

$$\theta_p=\theta_{p1}+\theta_{p2}+\theta_{p3}\Leftrightarrow\theta_p=\theta_c+\frac{D_s}{D_p}\theta_c-\frac{D_s}{D_p}\theta_s \tag{16.2.50}$$

All this can be rearranged to the final relation we were looking for:

$$\boxed{D_p\theta_p=(D_p+D_s)\theta_c-D_s\theta_s} \tag{16.2.51}$$

That can also be rewritten as:

$$Z_p\omega_p=(Z_p+Z_s)\omega_c-Z_s\omega_s \tag{16.2.52}$$

■ Q.E.D.

There are many other relations that we can derive for the planetary gear (we could write a whole book on the subject). But we will stop our investigations here because we are not big fans of the subject...

16.2.3 Strength of materials

Strength of materials (or SoM for the intimates...) is, like all the other sections of this textbook, an extremely vast area which level of detail and complexity of the calculations can explode rather quickly. We will in the following paragraphs focus on the essentials that the engineer (in business) needs to know. The developments are oversimplified for trivial special cases (straight bars and beams). In reality, we use the tensor calculus, design of experiments or computer modelling experience with FEM (finite element methods).

Definitions (#16.2.727): "resistance of materials" also known as "mechanics of materials" or "strength of materials" - is a branch of engineering mechanics that studies the behavior of solid objects subject to various forces and deformations. It examines how materials react to external loads, stresses, and strains, ensuring structures and mechanical components maintain stability and integrity under the presence or not of gravity. This field encompasses fundamental concepts such as elasticity, plasticity, fatigue, and failure analysis, which are critical in designing safe and durable engineering systems.

Before we begin to study some simple concrete cases let us de some reminders of the proofs we have proceed during our study of this subject in the section of Continuum Mechanics:

- A solid considered as rigid does not exist, it is only a convenient approximation. Experience shows that a solid is in fact still slightly deformable under the effect of external forces.
- Relationships between deformations and tensions are usually complicated due to the anisotropy of the crystal lattices. However, the solids are generally not single crystals but polycrystalline substances consisting of assemblies of microcrystals associated randomly, they can be thus considered as isotropic.

Then, we should consider globally the following assumptions relatively to the developments that will follow:

H1. Matter is homogeneous, that is to say for recall that it has the same physical constitution and the same structure throughout all its volume.

H2. Matter is isotropic, that is to say for recall that its mechanical properties are the same in any point of the volume.

H3. The material is perfectly elastic, that is to say, for recall that after removal of the external forces, the volume immediately takes back its original dimensions (unlike the plastic limit!).

H4. The deformations (displacements of the points of the characteristics line) are small compared with the dimensions of the objects studied.

H5. Any straight section (cross-sections) before deformation remain straight after deformation (i.e. "Navier-Bernoulli hypothesis").

H6. The results obtained in material strength theory can only be applies to a sufficient distance from the concentrated efforts applied on the volume (i.e. "Barré Saint Venant hypothesis").

H7. In the elastic range, matter obeys the law of proportionality and then the deformations are given by Hooke's law already proved in the section of Continuum Mechanics. This linear law allows us to apply the principle of superposition to forces and to resistive deformations.

We proved in the section of Continuum Mechanics (see page 3744) that Hooke's law states, when the deformations are reversible, that there is proportionality between tension and deformation (it's a variation formulation of Hooke's law):

$$F=k|\Delta L|=\frac{ES}{L_0}|\Delta L|=ES\varepsilon=ES\varepsilon \tag{16.2.53}$$

or:

$$\frac{F}{S}=\sigma=E\varepsilon \tag{16.2.54}$$

where E is the Young's modulus, ε the normal deformation and σ the normal stress.

Let us indicate that the ratio:

$$\frac{ES}{L_0} \tag{16.2.55}$$

is often named "stiffness (rigidity) of the bar" in the literature and is often denoted by k.

We have also proved in the section of Continuum Mechanics that the shear stress was given by (see page 3760):

$$\tau=\frac{E}{1+2\eta}\gamma=G\gamma \tag{16.2.56}$$

where G is the "shear modulus", γ the angle of deformation, and η the Poisson's ratio that is a dimensionless number. thus we have a relation between the modulus of elasticity and rigidity in the case of small deformations.

We have proved also in the same section that for a solid or a liquid subjected to a uniform isotropic pressure we had (see page 3761):

$$\frac{\Delta V}{V_0}=\sigma\frac{3(1-2\eta)}{E}=\sigma\kappa=\kappa\Delta P \tag{16.2.57}$$

The compressibility coefficient κ is therefore a positive number, so using the above relation, we have:

$$\kappa>0\Rightarrow(1-2\eta)>0\Rightarrow 2\eta<1 \tag{16.2.58}$$

and then comes a known result:

$$\eta<\frac{1}{2} \tag{16.2.59}$$

So the Poisson coefficient can not be bigger than $1/2$ and can be negative (in the latter case we speak then of "auxetic materials").

Finally, let us recall that we saw in the section of Continuum Mechanics (see page 3751) that the unit contraction along the z-axis was given during a traction along the x-axis by:

$$\varepsilon_z = -\eta\frac{\sigma_x}{E} = -\eta\varepsilon_x \tag{16.2.60}$$

That is to say when written differently (focusing on the XZ plane):

$$\frac{|\Delta L_z|}{L_{0z}} = -\eta\frac{|\Delta L_x|}{L_{0x}} \tag{16.2.61}$$

Therefore:

$$\eta = -\frac{\dfrac{|\Delta L_z|}{L_{0z}}}{\dfrac{|\Delta L_x|}{L_{0x}}} = \frac{(L_z - L_{0z})/L_{0z}}{(L_x - L_{0x})/L_{0x}} \tag{16.2.62}$$

and that is what shows the figure below:

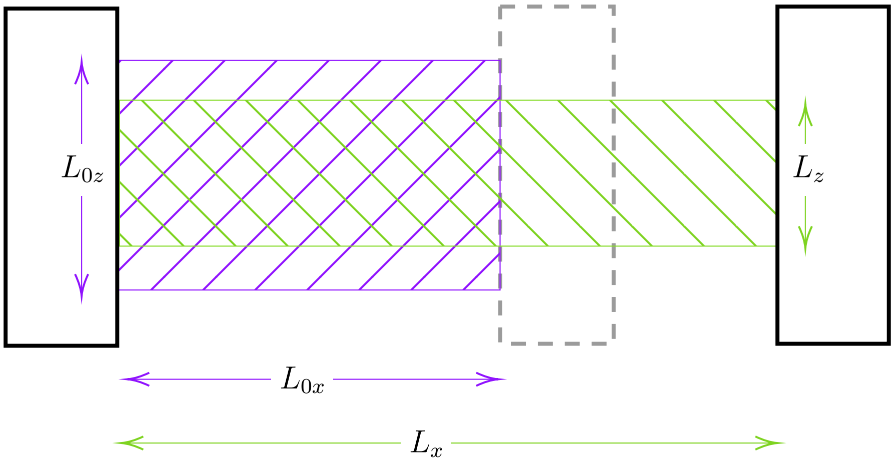


Figure 16.77: Traction on a test piece

We also proved in the section of Continuum Mechanics the following relation during our study of flexural modulus (see page 3762):

$$M = \frac{EI}{R} = \frac{E\displaystyle\int y^2 \mathrm{d}S}{R} = \frac{-\sigma I}{y} \tag{16.2.63}$$

which expresses the bending moment for a beam under a moment of force M (torque), then the span describes an arc of with a curvature of radius R and where I characterizes the "shape stiffness (rigidity)" of the material having a given cross sectional area S. This is a very important relation in many areas of mechanical and civil engineering (shipbuilding, automobile, architecture, etc.).

Remark

I is named the "static moment of inertia" or "quadratic moment" as we have already specified it in the section of Continuum Mechanics.

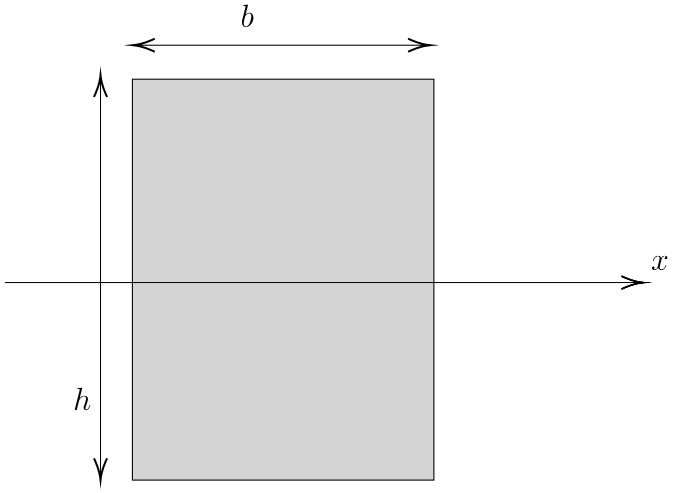
16.2.3.1 Quadratic moments

Let us see the three conventional static moments of inertia I in the field of Material Strengths because they are often encountered in practice (construction).

Remark

The theory of moments of inertia is presented for recall in the section of Classical Mechanics (3297). In the section on Geometric Shapes (see page 2886) we proved in detail the moments of inertia of the most common volumes.

Let us see first the quadratic moment of inertia I_x of a rectangular plate of side b and height h:



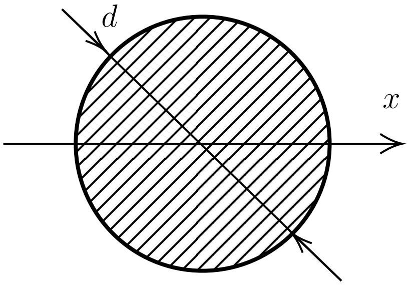
The domain occupied by the plate is given by:

$$S=\left\{(x,y)\in\mathbb{R}^2\left|-\frac{b}{2}\leq x\leq\frac{b}{2},-\frac{h}{2}\leq y\frac{h}{2}\right.\right\} \tag{16.2.64}$$

Then we have:

$$I_x=\int\limits_{-h/2}^{h/2}\int\limits_{-b/2}^{b/2}y^2\mathrm{d}x\mathrm{d}y=\int\limits_{-h/2}^{h/2}y^2x\Big|_{-b/2}^{b/2}\mathrm{d}y=\int\limits_{-h/2}^{h/2}by^2\mathrm{d}y=\frac{b}{3}y^3\Big|_{-h/2}^{h/2}=\frac{bh^3}{12} \tag{16.2.65}$$

And now let us study the quadratic static moment of a disc of diameter:



Here the domain of integration is:

$$S=\left\{(x,y)\in\mathbb{R}^2|x^2+y^2\leq\frac{d^2}{4}\right\}\qquad(16.2.66)$$

where d is the diameter of the disc.

To calculate this integral, we use polar coordinates that are for recall given by (see section Vector Calculus page 1863):

$$x=r\cos(\theta)\qquad y=r\sin(\theta)\qquad(16.2.67)$$

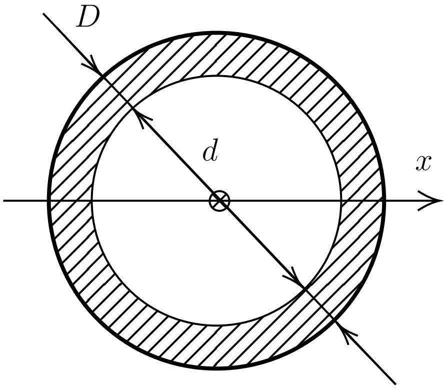
We always have:

$$I_x=\int_S y^2\mathrm{d}x\mathrm{d}y\qquad(16.2.68)$$

The Jacobian of the transformation is equal to r (see section Differential and Integral Calculus page 1528). We get:

$$I_x=\int\limits_0^{d/2}\int\limits_0^{2\pi}r^3(\sin(\theta))^2\mathrm{d}\theta\mathrm{d}r=\pi\int\limits_0^{d/2}r^3\mathrm{d}r=\frac{\pi d^4}{64}\qquad(16.2.69)$$

Now let us calculate the quadratic static inertia of a ring of external diameter D and internal diameter d:



Here is the domain of integration is:

$$S=\left\{(x,y)\in\mathbb{R}^2|\frac{d^2}{4}\leq x^2+y^2\leq\frac{D^2}{4}\right\}\qquad(16.2.70)$$

where D and d are the diameters of the large and small discs.

If we denote by S_1 the domain of the large disc and S_2 that of the small disk then:

$$I_x = \int_{S_1} y^2 \mathrm{d}x\mathrm{d}y - \int_{S_2} y^2 \mathrm{d}x\mathrm{d}y = \frac{\pi(D^4 - d^4)}{64} \tag{16.2.71}$$

using the quadratic static moment of inertia of the disk.

For summary, we have therefore:

$$\begin{aligned} I_x &= \int_S y^2 \mathrm{d}s = \iint_S y^2 \mathrm{d}x\mathrm{d}y \\ I_y &= \int_S x^2 \mathrm{d}s = \iint_S x^2 \mathrm{d}x\mathrm{d}y \end{aligned} \tag{16.2.72}$$

and finally there exist also the static polar quadratic moment of S with respect to a point O:

$$I_0 = \int_S r^2 \mathrm{d}s = \iint_S (x^2 + y^2) \mathrm{d}x\mathrm{d}y = I_x + I_y \tag{16.2.73}$$

It is therefore easy in simple cases to know the static polar moment of inertia and it is very useful for the study of torsion.

It follows from these tools as more elements of the section are located away from the axis, the more the static quadratic moment will be important and the more (we will prove it just further below) the "arrows" will be weak.

16.2.3.2 Equation of the elastic line

For this case study example, widely used in practice, we will first have to obtain mathematically the geometric form that takes the neutral axis of a beam subjected to bending forces.

Remark

If all the reactions of a solicited system can be found from the static equilibrium equations, the problem is say to be "statically determined" or "isostatic". When the equations of static fail to find the equilibrium of a system, the system is say to be "hyperstatic" or "statically indeterminate". We also say that while there is "freedom of movement" despite of n connection points, we are in isostatic situation.

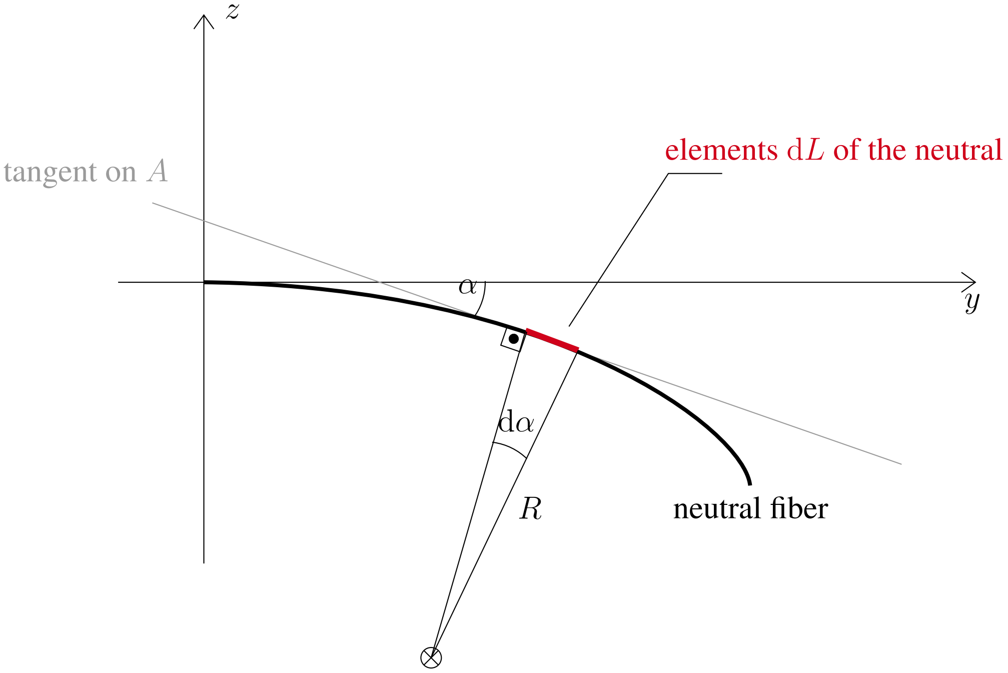


Figure 16.78: Beam (simplified ...) subjected to a force

By definition of the derivative and under the assumption of small deformations (however that still works well up to 45°...):

$$\frac{\mathrm{d}z}{\mathrm{d}y} \cong \tan(\alpha) \cong \alpha \qquad (16.2.74)$$

Or by derivating once again:

$$\frac{\mathrm{d}^2 z}{\mathrm{d}y^2} = \frac{\mathrm{d}\alpha}{\mathrm{d}y} \qquad (16.2.75)$$

Moreover, the figure shows that (see section Trigonometry page 2542):

$$\mathrm{d}L = R\mathrm{d}\alpha \Rightarrow \frac{1}{R} = \frac{\mathrm{d}\alpha}{\mathrm{d}L} \qquad (16.2.76)$$

But the fact that the neutral fiber curve deviates little from the y axis (small deformations), we can write:

$$\mathrm{d}y \cong \mathrm{d}L \qquad (16.2.77)$$

Therefore:

$$\frac{1}{R}\cong\frac{\mathrm{d}\alpha}{\mathrm{d}y} \tag{16.2.78}$$

Therefore we can write using the relationships obtained above:

$$\boxed{\frac{\mathrm{d}^2 z}{\mathrm{d}y^2}=\frac{1}{R}=\frac{M}{EI}} \tag{16.2.79}$$

which is the differential equation giving $z=f(y)$, named "equation of the elastic line".

Another common but less intuitive approach to get the same relation is to start from:

$$M=\frac{EI}{R} \tag{16.2.80}$$

and to recall that the radius of curvature R is given by (see section Differential Geometry page 2875):

$$R=\frac{1}{C}=\frac{\left(1+\left(\dfrac{\mathrm{d}y}{\mathrm{d}y}\right)^2\right)^{3/2}}{\left|\dfrac{\mathrm{d}^2 y}{\mathrm{d}x^2}\right|}=\frac{(1+y'^2)^{3/2}}{|y''|} \tag{16.2.81}$$

Either by adapting the notation to our context:

$$M=\frac{EI}{R}=\frac{EI}{\dfrac{(1+z'^2)^{3/2}}{|z''|}} \tag{16.2.82}$$

and neglecting the first derivative for the small deformations we find indeed:

$$M=\frac{EI}{R}=\frac{EI}{\dfrac{(1+z'^2)^{3/2}}{|z''|}}\cong EI|z''|=EI\frac{\mathrm{d}^2 z}{\mathrm{d}y^2} \tag{16.2.83}$$

✍**Examples:**

E1. Cantilevered beam (only at one side) that is classic case in construction and housing) with concentrated load (punctual) at the end:

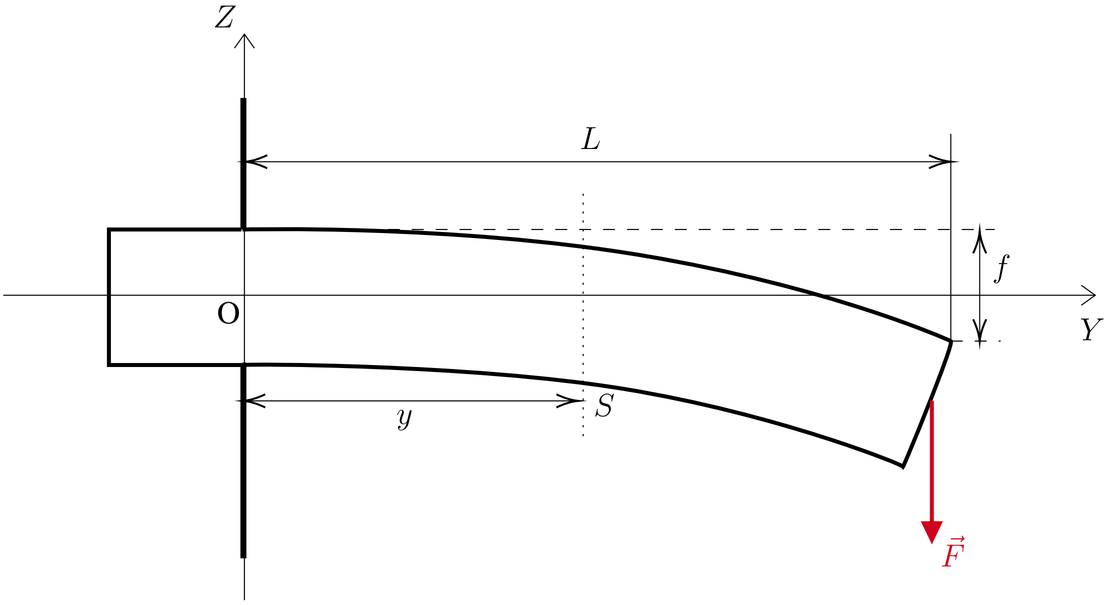


Figure 16.79: **Cantilevered beam with punctual force on extremity**

In the section S at any point, the moment of force (bending) is therefore:

$$M(y) = F(L-y) = \frac{EI}{R} \Rightarrow \frac{F(L-y)}{EI} = \frac{1}{R} \tag{16.2.84}$$

On the other hand:

$$\frac{\mathrm{d}^2 z}{\mathrm{d}y^2} = \frac{1}{R} \tag{16.2.85}$$

By eliminating R between these two relations, it remains:

$$\frac{\mathrm{d}^2 z}{\mathrm{d}y^2} = \frac{F}{EI}(L-y) \tag{16.2.86}$$

The figure shows that the boundary conditions are:

$$y = 0 \Rightarrow \begin{cases} z = 0 \\ \dfrac{\mathrm{d}z}{\mathrm{d}z} = 0 \end{cases} \tag{16.2.87}$$

We get after integration:

$$\begin{aligned} \frac{\mathrm{d}z}{\mathrm{d}z} &= \frac{F}{EI}\left(Ly - \frac{1}{2}y^2\right) \\ z &= \frac{F}{EI}\left(L\frac{1}{2}y^2 - \frac{1}{2}\frac{1}{3}y^3\right) \end{aligned} \tag{16.2.88}$$

Therefore:

$$z=\frac{FL}{EI}\frac{y^2}{2}-\frac{F}{2EI}\frac{y^3}{3} \tag{16.2.89}$$

If $y=L$ the deformation is maximum and z therefore takes the maximum value f named the "arrow" or deflection" (distance between the top initial position of the cantilever and the top of the final position in the above example). It follows:

$$z=f=\frac{FL}{EI}\frac{L^2}{2}-\frac{F}{2EI}\frac{L^3}{3}=\frac{FL^3}{3EI} \tag{16.2.90}$$

This gives us finally the maximum vertical deflection of the end to a beam fixed on one side:

$$\boxed{f=\frac{FL^3}{3EI}} \tag{16.2.91}$$

All data from this relation are known to us (force, length, Young's modulus, static momentum) and it is then possible to determine whether the bar will break or not because we simply need to apply the relation proved above:

$$M=\frac{-\sigma I}{y}\Rightarrow|\sigma|=\frac{yM}{I}=\frac{fM}{I} \tag{16.2.92}$$

So thanks to the relation:

$$\boxed{\sigma=\frac{fM}{I}} \tag{16.2.93}$$

and knowing experimentally from what experimental value of σ the material breaks down we will know when the beam will also break down (at least approximately!) knowing the arrow f, the bending moment M and the static quadratic moment I.

So we have a result that will be useful to us later:

$$\mathrm{d}f=F\frac{1}{EI}y^2\mathrm{d}y \tag{16.2.94}$$

and integrating from 0 to L we find the arrow of our previous beam!

E2. The sustained beam (also named "beam two times supported" or "isostatic beam") is the most classic example in construction and therefore architecture (and for those who played during childhood to put wooden slats to move over a small river). This is a homogeneous beam, of constant section, resting on two free bearings at its ends and subjected to a load F at the center:

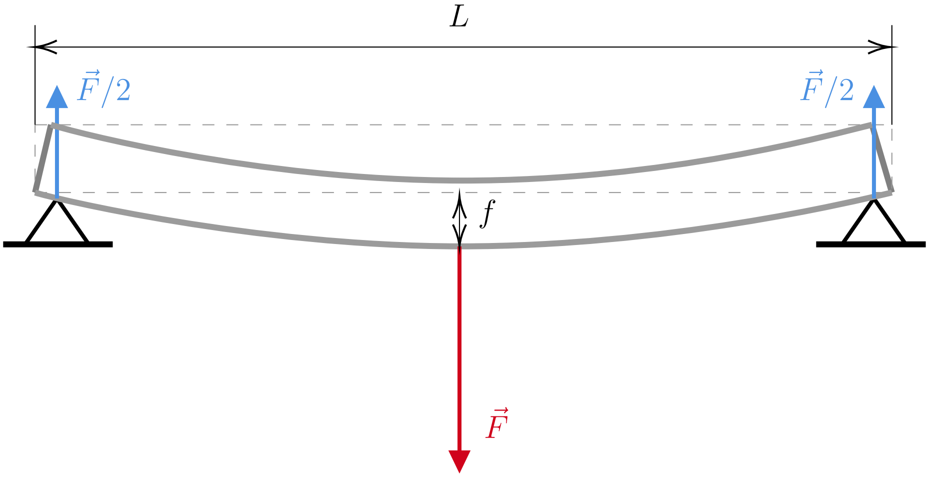


Figure 16.80: **Sustained beam with punctual centered force**

We can therefore consider that it is as if we had $F/2$ at both ends of two beams of length $L/2$ (thus the sum is indeed equal to F, that is to say, the bending moment). Notice that we neglect the weight of the beam relatively to that of F, but F may be simply the weight of the beam! Using the last relation of the previous example, we have:

$$\mathrm{d}f=\frac{F}{2}\frac{1}{EI}y^2\mathrm{d}y \tag{16.2.95}$$

Therefore:

$$f=\frac{F}{2}\frac{1}{EI}\int_0^{L/2}y^2\mathrm{d}y=\frac{FL^3}{48EI} \tag{16.2.96}$$

Thus the maximum vertical deflection of a beam connected on both sides is finally (changing a bit the notation):

$$\boxed{f=\frac{FL^3}{48EI}} \tag{16.2.97}$$

Thus, for a same length of beam, at identical F the arrow is 16 times smaller than for a fixed beam! It was intuitive that it was lower for the same force but hard to guess that it would be of a factor 16...!

It is this relationship that is also used for IPN beams (famous beams in building construction!).

E3. Beam built only one side still named "cantilever beam" (also classic case in the construction and housing) but now with a constant line load denoted w:

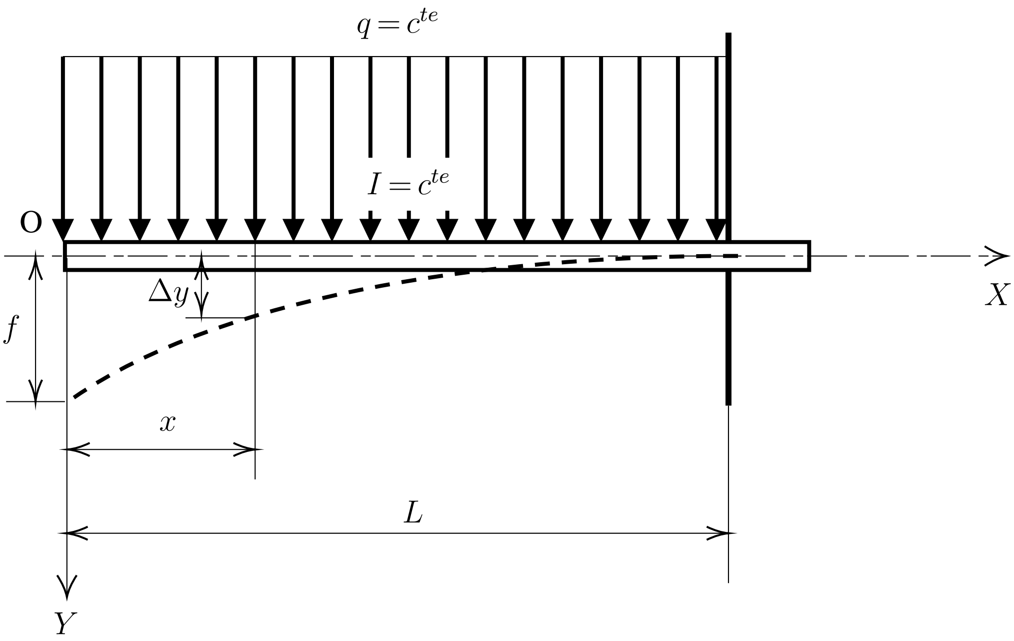


Figure 16.81: Beam fixed on one side with constant line load

The development is simple but some simplifications are smart to get at the end an elegant result. So we always start from the equation of the elastic line by adopting the notations relatively to the selected configuration (see figure above):

$$\frac{\mathrm{d}^2y}{\mathrm{d}x^2} = \frac{M}{EI} \tag{16.2.98}$$

Thus explicitly:

$$\frac{\mathrm{d}^2y}{\mathrm{d}x^2} = \frac{\displaystyle\int F\mathrm{d}x}{EI} = \frac{\displaystyle\int\left(\int q\mathrm{d}x\right)\mathrm{d}x}{EI} = \frac{\displaystyle\int wx\mathrm{d}x}{EI} = \frac{w\dfrac{x^2}{2} + c^{te}}{EI} \tag{16.2.99}$$

As the moment of force M is zero at $x = 0$, we have the constant that is zero. Since then:

$$\frac{\mathrm{d}^2y}{\mathrm{d}x^2} = \frac{w\dfrac{x^2}{2}}{EI} \tag{16.2.100}$$

Integrating again, we get:

$$y' = \frac{\mathrm{d}y}{\mathrm{d}x} = \frac{w}{EI}\int\frac{x^2}{2}\mathrm{d}x = \frac{w}{EI}\frac{x^3}{6} + c^{te} \tag{16.2.101}$$

As for $x = L$ by hypothesis the deformation is zero, we have then the constant that is given by:

$$y' = \frac{w}{EI}\frac{L^3}{6} + c^{te} = 0 \Rightarrow = -\frac{wL^3}{6EI} \tag{16.2.102}$$

Therefore:

$$y' = \frac{w}{EI}\frac{x^3}{6} - \frac{wL^3}{6EI} = \frac{wL^3}{6EI}\left(\left(\frac{x}{L}\right)^3 - 1\right) \tag{16.2.103}$$

By integrating a last time:

$$y=\frac{w}{6EI}\int(x^3-L^3)\mathrm{d}x=\frac{w}{6EI}\left(\frac{x^4}{4}-L^3x\right)+\frac{wL^4}{8EI} \tag{16.2.104}$$

And as on $x=L$ we have y that is also zero, it comes for the constant:

$$y=\frac{w}{EI}\left(\frac{L^4}{4}-L^3L\right)+c^{te}=0\Rightarrow=\frac{wL^4}{8EI} \tag{16.2.105}$$

Therefore:

$$y=\frac{w}{EI}\left(\frac{x^4}{4}-L^3x\right)+\frac{wL^4}{8EI} \tag{16.2.106}$$

And as the arrow is anyway on $x=0$, we have then:

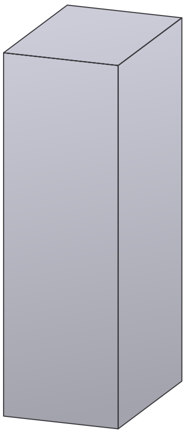
$$f=\frac{wL^4}{8EI} \tag{16.2.107}$$

16.2.3.2.1 Euler-Bernoulli Beam equation

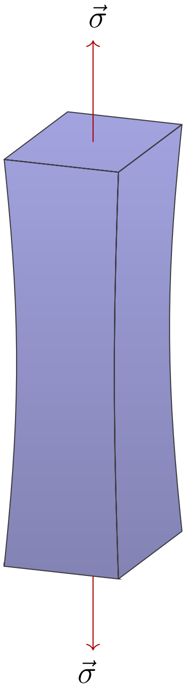
Let us now consider the case of a beam fixed on both sides (case even more common that the three previous examples!). The analysis will be a bit more difficult and we will have to introduce several new concepts.

A beam in practice must withstand at least the following efforts:

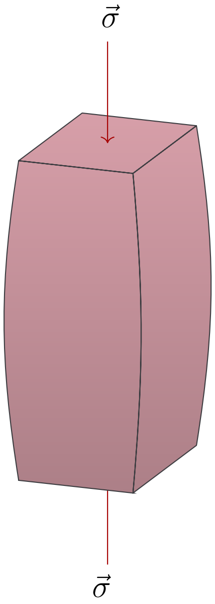
- Unloaded (normal):



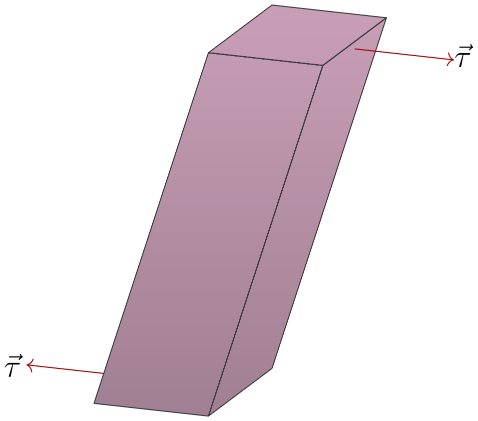
- Tension:


$\vec{\sigma}$
$\vec{\sigma}$

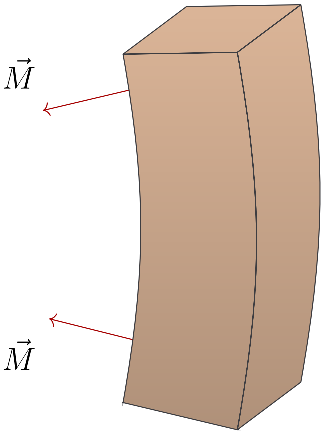
- Compression:


$\vec{\sigma}$
$\vec{\sigma}$

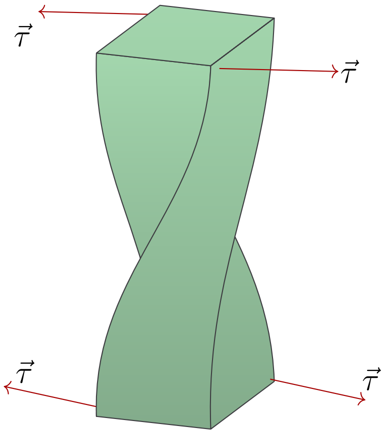
- Shearing (shear):


$\vec{\tau}$
$\vec{\tau}$

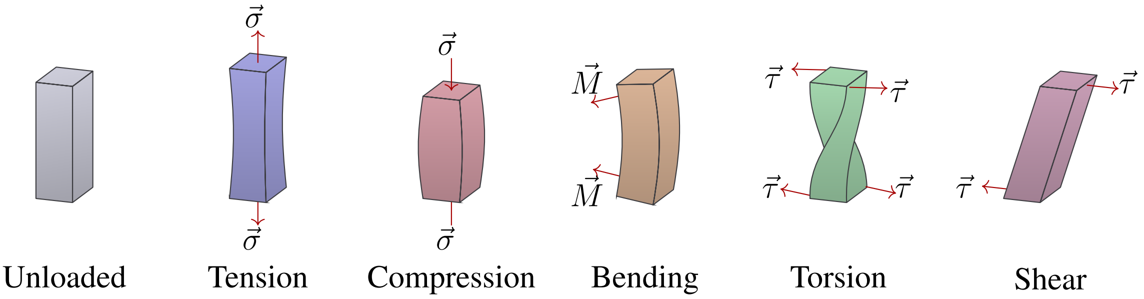
- Flexion (bending stress):



- Torsion:


$\vec{\tau}$
$\vec{\tau}$
$\vec{\tau}$
$\vec{\tau}$

Or to summarize in a single figure:



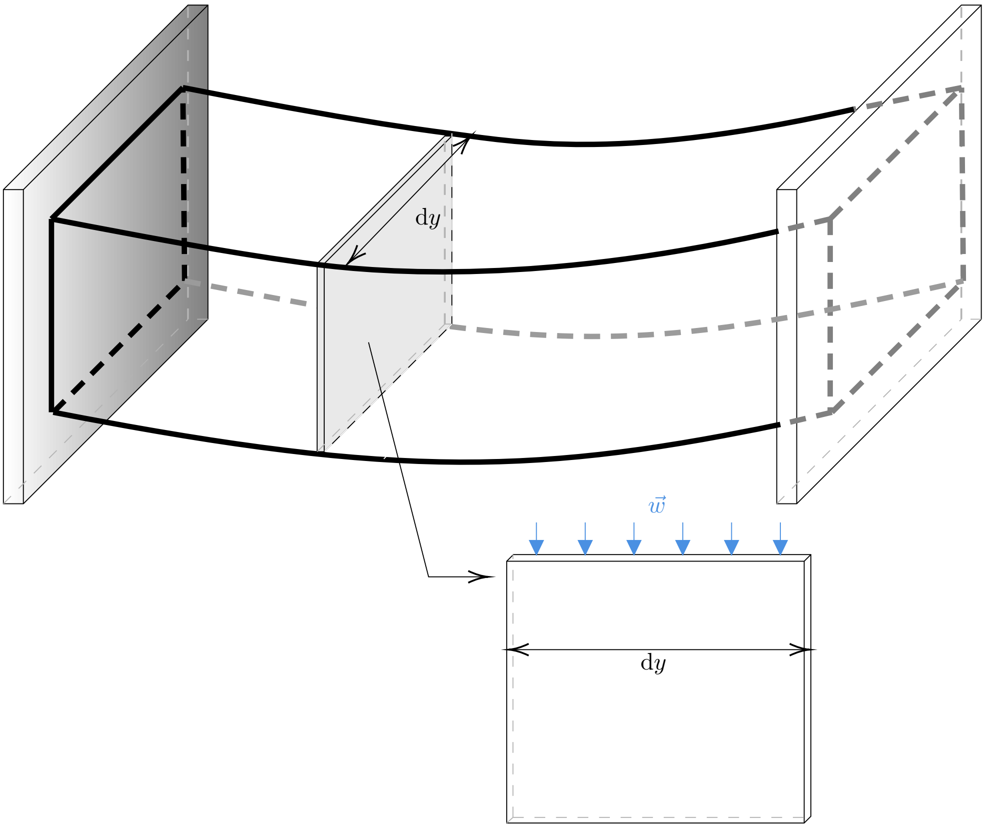
Unloaded Tension Compression Bending Torsion Shear

Figure 16.82: Most common types of material deformations (author: Izaak Neutelings)

If a beam is in equilibrium, then the internal forces must satisfy at all points the following relations:

$$\begin{gathered}\sum\vec{F}_{\text{horizontal}}=\sum\vec{F}_{\text{stress/compression}}=\vec{0}\\ \sum\vec{F}_{\text{horizontal}}=\sum\vec{F}_{\text{shear}}=\vec{0}\\ \sum\vec{M}=\sum\vec{M}_{\text{bending}}=\vec{0}\end{gathered}\tag{16.2.108}$$

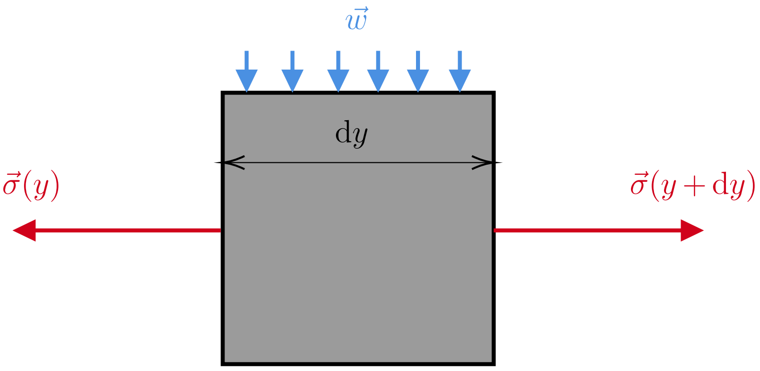
Let us consider now a beam embedded at its two ends (dual-embedded beam) and let us take in a slice of infinitesimal length $\mathrm{d}y$ such that locally its curvature is zero. The beam will be assumed to be subjected to a uniform force over its entire length (a force which can also be assimilated to its own weight, as already mentioned above). It is then customary to denote by $\vec{w}$ the force per unit length (total weight divided by the length) which is obviously a linear load:



If the beam is at equilibrium once deformed (weakly or strongly deformed, this does not matters

!) then the sums of the forces of tension, compression, shear and bending must be zero at each point as we have already said! This does not however mean that at each point of the elastic line the values of each of the forces is equal! On the contrary! There are of course differences (if not, there would be no deformation...!).

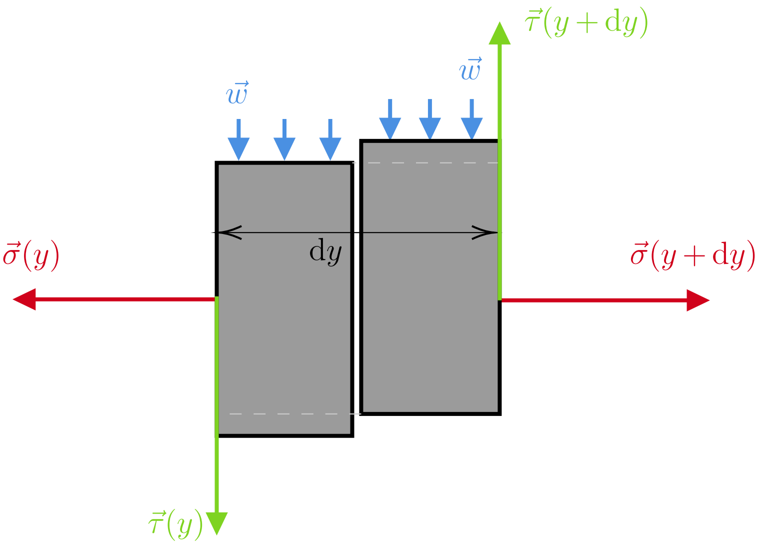
Let us first sum the local forces of tension and compression (horizontal) of the element of length $\mathrm{d}y$. We then have schematically:



That is to say algebraically (the variational can be negative or positive whatever!):

$$\sum \vec{F}_{\text{horizontal}} = \vec{\sigma}(y+\mathrm{d}y) - \vec{\sigma}(y) = \vec{\sigma} + \mathrm{d}\vec{\sigma} - \vec{\sigma} = \mathrm{d}\vec{\sigma} = \vec{0} \qquad (16.2.109)$$

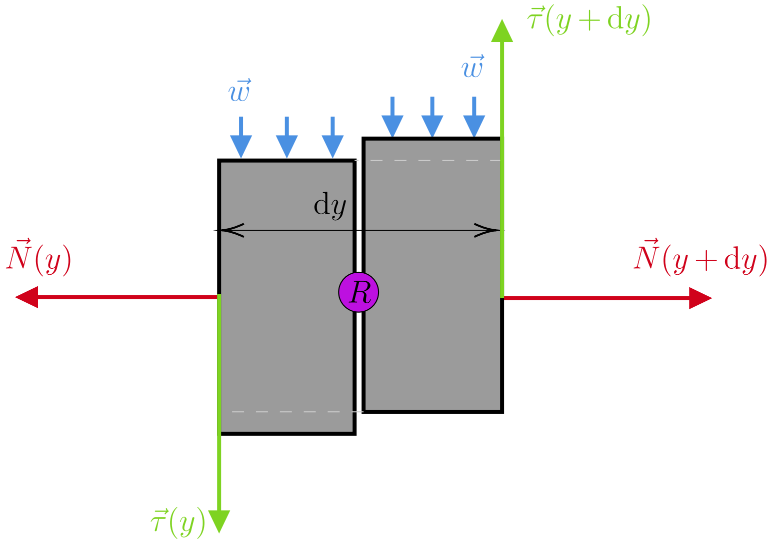
If now we take care of the vertical forces at the source of the shear then we have schematically:



That is to say algebraically (the variational can be negative or positive whatever!):

$$\sum \vec{F}_{\text{vertical}} = \vec{\tau}(y+\mathrm{d}y) - \vec{\tau}(y) = \vec{\tau} + \mathrm{d}\vec{\tau} - \vec{\tau} - \vec{w}\mathrm{d}y = \mathrm{d}\vec{\tau} - \vec{w}\mathrm{d}y = \vec{0} \qquad (16.2.110)$$

And finally for the bending moments the sum is also necessarily zero at equilibrium. However, in contrast to the two preceding algebraic sums where we could only use the differential, we have to choose a reference point R for the bending moments since, for recall... the moment of force is by definition the product of a force by a distance. We will therefore naturally choose the center of gravity:



The uniform linear load $\vec{w}$ on the length $\mathrm{d}y$ generates a force at mid distance of:

$$\vec{F}_{\vec{w}}=(\vec{w}(y)\mathrm{d}y)\frac{1}{2}\mathrm{d}y=\frac{1}{2}\vec{w}(y)\mathrm{d}y^2 \tag{16.2.111}$$

but as it is confounded with the choices of our reference point, then its moment of strength is zero!

We then have algebraically:

$$\begin{aligned}\sum\vec{M}&=\vec{M}(y)+\vec{M}(-y)+\vec{\tau}(y)\frac{\mathrm{d}y}{2}+\vec{\tau}(-y)\frac{\mathrm{d}y}{2}\\&=\underbrace{\vec{M}+\mathrm{d}\vec{M}}_{=\vec{M}(y)}-\underbrace{\vec{M}}_{=\vec{M}(-y)}\frac{\mathrm{d}y}{2}+\underbrace{\vec{\tau}}_{=\vec{\tau}(-y)}\frac{\mathrm{d}y}{2}=\mathrm{d}\vec{M}+\vec{\tau}\frac{\mathrm{d}y}{2}+\mathrm{d}\vec{\tau}\frac{\mathrm{d}y}{2}+\vec{\tau}\frac{\mathrm{d}y}{2}\\&=\mathrm{d}\vec{M}+\vec{\tau}\mathrm{d}y+\mathrm{d}\vec{\tau}\frac{\mathrm{d}y}{2}=\vec{0}\end{aligned} \tag{16.2.112}$$

And if we neglect the differentials of order two it remains only:

$$\mathrm{d}\vec{M}+\vec{\tau}\mathrm{d}y=\vec{0} \tag{16.2.113}$$

Finally, we have:

$$\boxed{\left.\begin{aligned}&\mathrm{d}\vec{\sigma}=\vec{0}\\&\mathrm{d}\vec{\tau}-\vec{w}\mathrm{d}y=\vec{0}\\&\mathrm{d}\vec{M}+\vec{\tau}\mathrm{d}y=\vec{0}\end{aligned}\right\}\Leftrightarrow\left\{\begin{aligned}&\frac{\mathrm{d}\vec{\sigma}}{\mathrm{d}y}=\vec{0}\\&\frac{\mathrm{d}\vec{\tau}}{\mathrm{d}y}=\vec{w}\\&\frac{\mathrm{d}\vec{M}}{\mathrm{d}y}=-\vec{\tau}\end{aligned}\right.} \tag{16.2.114}$$

To determine the bending moment from the linear load (which is of primary interest to the practitioner), we derive the third relation twice and make a substitution:

$$\frac{\mathrm{d}}{\mathrm{d}y}\left(\frac{\mathrm{d}\vec{M}}{\mathrm{d}y}\right)=-\frac{\mathrm{d}}{\mathrm{d}y}\vec{\tau}=-\vec{w} \tag{16.2.115}$$

Hence:

$$\frac{\mathrm{d}^2\vec{M}}{\mathrm{d}y^2}=-\vec{w} \tag{16.2.116}$$

The problem with this last relation is the knowledge of the moments. We should get rid of this term absolutely. What we know easily is the deformation function and we have proved earlier above that:

$$\frac{\mathrm{d}^2 z}{\mathrm{d}y^2}=\frac{1}{R}=\frac{M}{EI} \tag{16.2.117}$$

It then comes immediately by substituting the preceding relation in the prior previous one:

$$\boxed{EI\frac{\mathrm{d}^4 z}{\mathrm{d}y^4}=-w(y)} \tag{16.2.118}$$

where the product EI is known as the "flexural rigidity".

This is the most important elementary relation of the theory of beams, that covers the case for small deflections of a beam that is subjected to lateral loads only, because it allows by knowing the linear load to determine the function of deformation or conversely! It is so important to know that it is named the "static beam equation" or in honour to those who have determined it: "Euler-Bernoulli equation". It seemed it was first enunciated circa 11750 (holocene calendar), but was not applied on a large scale until the development of the Eiffel Tower and the Ferris wheel in the late 119th century (holocene calendar).

Since it is a differential equation of order 4 that will generate four constants at each integration, we will then need 4 initial conditions to solve it completely.

Remark

The Euler-Bernoulli equation by its assumptions is only a special case of a more general theory named the "Timoshenko beam theory" published in 11921 (holocene calendar).

✍**Example:**

We would like to calculate the deformation of a beam fixed on both sides and loaded uniformly knowing its length L, its modulus of elasticity E, its moment of inertia I.

We start from (we change the notations only to show that according to some textbooks the axes can be denoted differently):

$$EI\frac{\mathrm{d}^4y}{\mathrm{d}x^4}=-w(x) \tag{16.2.119}$$

With the initial conditions:

$$y(x=0)=0 \qquad y(x=L)=0 \qquad \frac{\mathrm{d}y}{\mathrm{d}x}(x=0)=0 \qquad \frac{\mathrm{d}y}{\mathrm{d}y}(x=L)=0 \tag{16.2.120}$$

And integrating by repetition, we have:

$$\begin{gathered}\frac{\mathrm{d}^3y}{\mathrm{d}x^3}=\frac{-wx}{EI}+C_1\Rightarrow\frac{\mathrm{d}^2y}{\mathrm{d}x^2}=\frac{-wx^2}{2EI}+C_1x+C_2\Rightarrow\frac{\mathrm{d}y}{\mathrm{d}x}=\frac{-wx^2}{6EI}+\frac{C_1x^2}{2}+C_2x+C_2\\ y(x)=\frac{-wx^4}{24EI}+\frac{C_1x^3}{6}+\frac{C_2x^2}{2}+C_3x+C_4\end{gathered} \tag{16.2.121}$$

From the two initial conditions:

$$y(x=0)=0 \qquad \frac{\mathrm{d}y}{\mathrm{d}x}(x=0)=0 \tag{16.2.122}$$

It follows that:

$$C_3=C_4=0 \tag{16.2.123}$$

With the other two remaining conditions, we obtain the following system which must allow us to determine the two remaining constants:

$$\begin{cases}\dfrac{-wL^4}{24EI}+\dfrac{C_1L^3}{6}+\dfrac{C_2L^2}{2}=0\\ \dfrac{-wL^3}{6EI}+\dfrac{C_1L}{2}+C_2L=0\end{cases} \tag{16.2.124}$$

Then it is simply a matter of solving a simple linear system (see section Linear Algebra page 1929):

$$\begin{cases}\dfrac{-wL^4}{24EI}+\dfrac{C_1L^3}{6}+\dfrac{C_2L^2}{2}=0\Big|\cdot\dfrac{24}{L^2}\\ \dfrac{-wL^3}{6EI}+\dfrac{C_1L}{2}+C_2L=0\Big|\cdot\dfrac{6}{L}\end{cases}\Rightarrow\begin{cases}\dfrac{-wL^2}{EI}+4C_1L+12C_2=0\Big|\cdot\dfrac{24}{L^2}\\ \dfrac{-2wL^2}{EI}+6C_1L+12C_2=0\Big|\cdot\dfrac{6}{L}\end{cases} \tag{16.2.125}$$

Subtracting both relation in the simple adequate way we get immediately:

$$\begin{cases}C_1&=\dfrac{wL}{2EI}\\ C_2&=\dfrac{-wL}{12EI}\end{cases} \tag{16.2.126}$$

Therefore for an ideal beam embedded on both sides subjected to a uniform load and described by:

$$y(x)=\frac{-wx^4}{24EI}+\frac{wLx^3}{12EI}-\frac{wL^2x^2}{24EI}=\frac{-wx^2}{24EI}(x^2-2Lx+L^2)=\frac{-wx}{24EI}(x-L)^2 \tag{16.2.127}$$

To determine the arrow, we must therefore look for the point x where this relation has an optimum. We then have:

$$\begin{aligned}f'(x)&=\left[x^2(x-L)^2\right]'=2x(x-L)^2+2x^2(x-L)\\&=2x(x-L)(x-L+x)=2x(x-L)(2x-L)=0\end{aligned} \tag{16.2.128}$$

It follows that the arrow has a maximum at $x=L/2$. Injecting this into $y(x)$ we get the famous relation often given in the literature but rarely proved:

$$\boxed{f=\frac{w}{24EI}\frac{L^4}{16}=\frac{wL^4}{384EI}} \tag{16.2.129}$$

It follows that the arrow of a beam is proportional to the fourth power of the length of the beam! Such high dependence imposes significant limitations on Civil Engineering structures based on beams.

Remark

The constant linear load, either over the total length of the beam or in successive sections, is a frequent observed type of stress in horizontal axis of pieces. It may come from the weight of the pieces with a constant section or from a load caused by an external force (eg gravity).

16.2.3.2.2 Potential elastic energy

After a quick overview of the various elastic deformations of the parts stressed by the fundamental forces, we will establish here the general expression of the elastic energy accumulated in a beam of any shape stressed by external forces.

Let us recall for this study that we can write Hook's law (see section Continuum Mechanics page 3744) in the following form:

$$|\Delta L| = \frac{FL_0}{ES} = \frac{F}{k} \tag{16.2.130}$$

and the elastic potential energy of a spring proved in the section of Classical Mechanics (see page 3418) is given by:

$$E_{\text{elastic}} = \frac{1}{2}kx^2 \tag{16.2.131}$$

Or in relative displacement:

$$E_{\text{elastic}} = \frac{1}{2}k\Delta x^2 \tag{16.2.132}$$

In the traction domain (or compression) of the beams, it is customary to consider the beam as a spring (...) and then to use the stiffness constant of Hook's law ... hoping that this is conforming to experimental observations...:

$$E_{\text{elastic}} = \frac{1}{2}k\Delta x^2 = \frac{1}{2}\frac{ES}{L_0}\Delta x^2 = \frac{ES}{2L_0}\Delta L^2 \tag{16.2.133}$$

Where, according to usage, we denote the longitudinal displacement L instead of x. By injecting Hook's law (yes ... it turns a little bit a loop... but that's engineering...):

$$E_{\text{elastic}} = \frac{ES}{2L_0}\left(\frac{FL_0}{ES}\right)^2 = \frac{F^2L_0}{2ES} \tag{16.2.134}$$

It is then sufficient to divide by the length of the bar to have the elastic linear energy denoted:

$$\boxed{\bar{E}_{\text{elastic}} = \frac{F^2}{2ES}} \tag{16.2.135}$$

The energy density, noted conventionally denoted as in thermodynamics by a minuscule letter, is then:

$$\boxed{e = \frac{E_{\text{elastic}}}{V} = \frac{F^2L_0}{2ESV} = \frac{F^2L_0}{2ESSL_0} = \frac{F^2}{2ES^2} = \left(\frac{F}{S}\right)^2\frac{1}{2E} = \frac{\sigma^2}{2E}} \tag{16.2.136}$$

16.2.3.3 Torsion

Let us recall first to the reader a study we made in the section on Classical Mechanics on the torsional pendulum where some elements had deliberately been muted.... Let's study this in more detail as it is very useful for transmission shafts or springs in everyday life.

Let us consider for this a cylindrical wire fixed at its base and subjected to a torsion moment $\vec{M}$. Under the effect of this torsion moment, the upper face of the wire is offset by an angle θ with respect to the lower face, the material undergoing a torsion stress (or shearing τ):

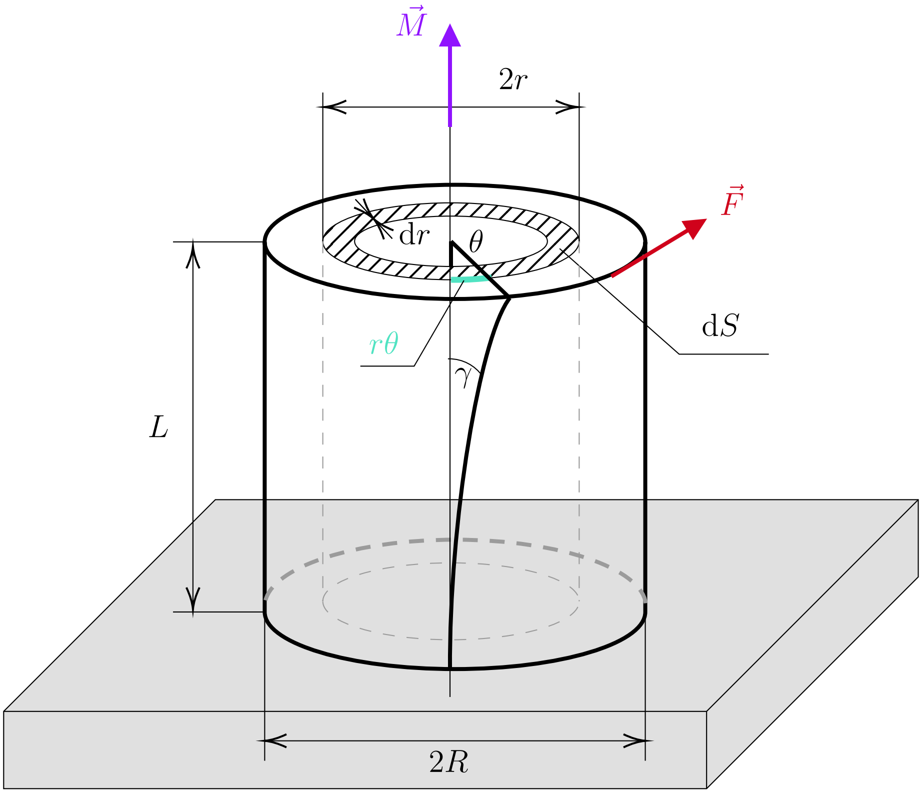


Figure 16.83: **Wire under torsion**

Let us imagine an extraction of the inside the wire an elementary tube of radius r, of thickness $\mathrm{d}r$, and let us observe the effect of the torsion on this tube unrolled (this will allow us an approximate approach of the concerned phenomenon):

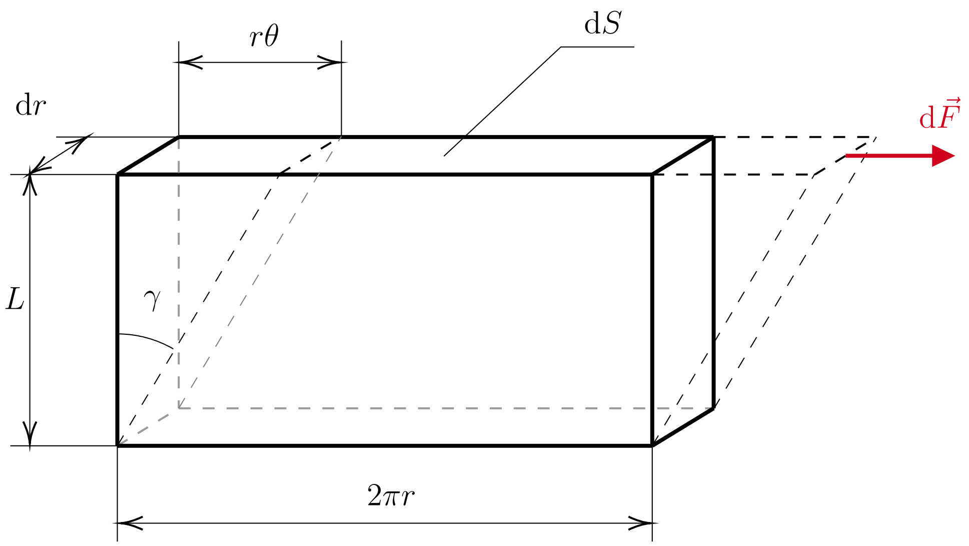
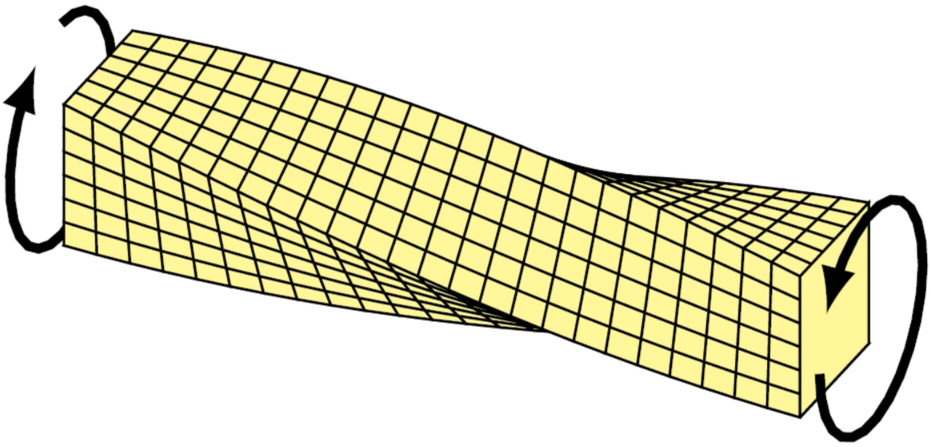


Figure 16.84: **Extraction an unroll of an element of the wire**

Let us look for a relation between torsion moment $\vec{M}$ and torsion angle θ that would therefore also apply to a beam as below:



For the unrolled tube, let us apply the shear relations as proved and introduced in the section of Continuum Mechanics (see page 3760):

$$G\gamma = \tau = \frac{\mathrm{d}F}{\mathrm{d}S} \Rightarrow \mathrm{d}F = \gamma G \mathrm{d}S \tag{16.2.137}$$

but the figure shows that (the deformations being weak) to the first order in Taylor series (see section Sequences and Series page 1703):

$$\gamma = \arctan\left(\frac{r\theta}{L}\right) \cong \frac{r\theta}{L} \tag{16.2.138}$$

hence:

$$\mathrm{d}F = \frac{r\theta}{L} G \mathrm{d}S \tag{16.2.139}$$

The elementary moment due to this force is by definition of the moment of force (torque):

$$\mathrm{d}\vec{M}=\vec{r}\times\mathrm{d}\vec{F} \tag{16.2.140}$$

Since $\vec{r}$ and $\vec{F}$ are perpendicular:

$$\mathrm{d}M=\frac{\theta G}{L}r^2\mathrm{d}S \tag{16.2.141}$$

The total moment (torque) is then equal to:

$$M=\int\limits_{r=0}^{r=R}\mathrm{d}M=\frac{\theta G}{L}\int\limits_{r=0}^{r=R}r^2\mathrm{d}S=\frac{\theta G}{L}I_0 \tag{16.2.142}$$

therefore:

$$M=\left(\frac{GI_0}{L}\right)\theta=-\left(\frac{GI_0}{L}\right)\theta=-k\theta \tag{16.2.143}$$

We thus fall back on the relation of the torsional pendulum that we had found during our study of the torsional pendulum in the section of Classical Mechanics (see page 3425):

$$\boxed{M=-k\theta} \tag{16.2.144}$$

with the difference that this time the constant k, the "torsion constant" is explicit!!!!

The numerator of the constant k is named in the domain of the strength of materials the "torsional rigidity" or "torsional stiffness" and the constant k itself is often referred to as "shaft rigidity" instead of "torsional constant". In practice, the main aim is to find the numerical value of the following expression:

$$\boxed{\theta=\frac{ML}{GI_0}} \tag{16.2.145}$$

since this will give the angular amplitude of the torsion.

Let us see therefore a very important application to the compression spring of helical type (the approach is approximate again...) working in torsion.

First, it must be realized that when a force is applied to the spring, the ends will rotate by a small angle θ (torsion) corresponding to the movement by a distance x which itself corresponds to the narrowing of the spring (yes indeed! this length must be taken somewhere....).

Given then a spring of external radius R (or of diameter D), of shear modulus G, with a body diameter d (diameter of the folded cylinder of which the spring is composed):

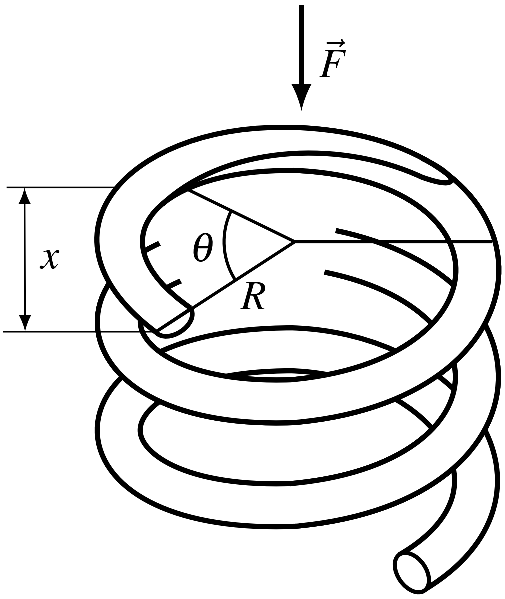


Figure 16.85: Spiral spring under load (author: H. Zeljko)

For the analysis we will simply need to mix several of the relations proved so far. In the first place, the torsion angle of a beam of length L (length of the spring in this case!):

$$\theta = \frac{ML}{GI_0} \tag{16.2.146}$$

With:

$$I_0 = I_x + I_y = \frac{\pi d^4}{64} + \frac{\pi d^4}{64} = \frac{\pi d^4}{32} \tag{16.2.147}$$

and:

$$L = n\pi D \tag{16.2.148}$$

Moreover, the moment of torsion is written:

$$M = FR = F\frac{D}{2} \tag{16.2.149}$$

We thus arrive at:

$$\theta = \frac{ML}{GI_0} = \frac{F\dfrac{D}{2}n\pi D}{G\dfrac{\pi d^4}{32}} = \frac{16FnD^2}{Gd^4} \tag{16.2.150}$$

Remark

It would be pretentious to claim to do with this section as well and also as complete as the *Statique* free French PDF of Nicolet Gaston Raymond that is a priori unrivaled in content and quality to this date (even compared to non-free textbooks on the subject!). It is therefore strongly recommended to refer to it if the reader wants to drive full information about civil engineering (see the download section of the companion website or directly his own website `http://www.raymondnicolet.ch`). Same recommendation applies to the reader about the master pieces that are the English books *Mechanics of Materials* ([246]) and *Structural and Stress Analysis* ([384])!

The ratio M/θ, as well as for the drive shaft, is named the "spring stiffness" or "spring rigidity".

The displacement (deformation) x is equal to (see section Trigonometry page 2542):

$$x=R\theta=\frac{D}{2}\theta \tag{16.2.151}$$

We finally arrive at:

$$x=\frac{D}{2}\frac{16FnD^2}{Gd^4}=\frac{8FnD^3}{Gd^4} \tag{16.2.152}$$

Which brings us to the worldwide known relation in the world in the strength of material regarding to springs:

$$\boxed{F=\left(\frac{Gd^4}{8nD^3}\right)x=kx} \tag{16.2.153}$$

where k is the "spring stiffness" constant also named "spring rigidity"! If now we use the expression of the elastic potential energy of a spring proved in the section of Classical Mechanics:

$$E_{\text{elastic}}=\frac{1}{2}kx^2 \tag{16.2.154}$$

We can then determine the energy a spiral spring can absorb.

16.2.3.4 Buckling

We conclude this study of the material strengths in this textbook for now, with "buckling" (classical study case in construction and mechanics) which consists in determining (in a special simple case) the minimum force F_0 from which a bar of length L, of Young's modulus E fixed at its two ends can fold (with a curvature radius R) until it breaks without the need to increase the force F_0 (this is again an indication value!).

Buckling is characterized by a sudden sideways failure of a structural member subjected to high compressive stress, where the compressive stress at the point of failure is less than the ultimate compressive stress that the material is capable of withstanding. Mathematical analysis of buckling often makes use of an "artificial" axial load eccentricity that introduces a secondary bending moment that is not a part of the primary applied forces being studied. As an applied load is increased on a member, such as a column, it will ultimately become large enough to cause the member to become unstable and is said to have buckled. Further load will cause significant and somewhat unpredictable deformations, possibly leading to complete loss of the member's load-carrying capacity. If the deformations that follow buckling are not catastrophic the member will continue to carry the load that caused it to buckle. If the buckled member is part of a larger assemblage of components such as a building, any load applied to the structure beyond that which caused the member to buckle will be redistributed within the structure. Theoretically, buckling is caused by a bifurcation in the solution to the equations of static equilibrium. At a certain stage under an increasing load, further load is able to be sustained in one of two states of equilibrium: a purely compressed state (with no lateral deviation) or a laterally-deformed state.

The ratio of the effective length of a column to the least radius of gyration of its cross section is named the "slenderness ratio λ". This ratio affords a means of classifying columns. Slenderness ratio is important for design considerations. All the following are approximate values used for convenience:

- A "short steel column" is one whose slenderness ratio does not exceed 50; an intermediate length steel column has a slenderness ratio ranging from about 50 to 200, and its behaviour is dominated by the strength limit of the material, while a long steel column may be assumed to have a slenderness ratio greater than 200 and its behaviour is dominated by the modulus of elasticity of the material.
- A "short concrete column" is one having a ratio of unsupported length to least dimension of the cross section equal to or less than 10. If the ratio is greater than 10, it is considered as a "long column" (sometimes referred to as a "slender column).
- "Timber" columns may be classified as short columns if the ratio of the length to least dimension of the cross section is equal to or less than 10. The dividing line between intermediate and long timber columns cannot be readily evaluated. One way of defining the lower limit of long timber columns would be to set it as the smallest value of the ratio of length to least cross sectional area that would just exceed a certain constant of the material.

If the load on a column is applied through the center of gravity (centroid) of its cross section, it is named an "axial load". A load at any other point in the cross section is known as an "eccentric load":

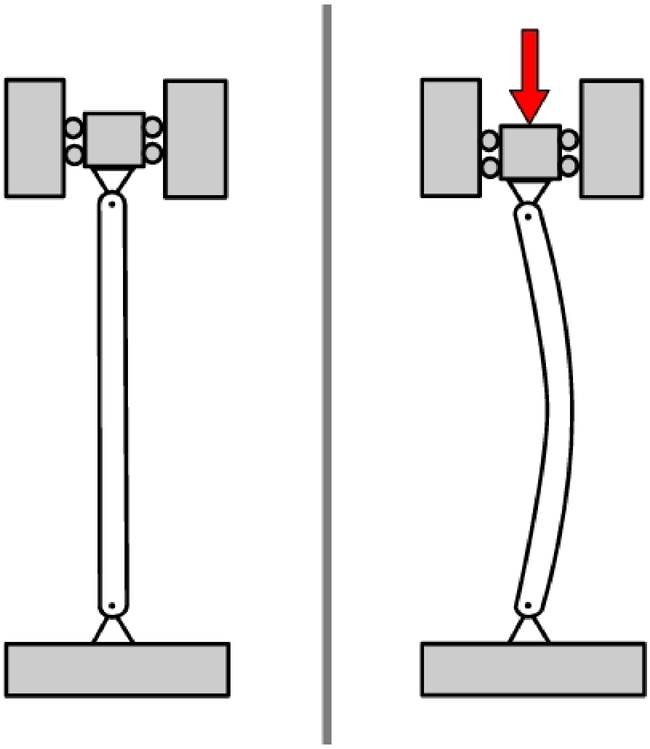


Figure 16.86: **Column under a concentric axial load exhibiting the characteristic deformation of buckling (source: Wikipedia)**

A short column under the action of an axial load will fail by direct compression before it buckles, but a long column loaded in the same manner will fail by buckling (bending), the buckling effect being so large that the effect of the axial load may be neglected.

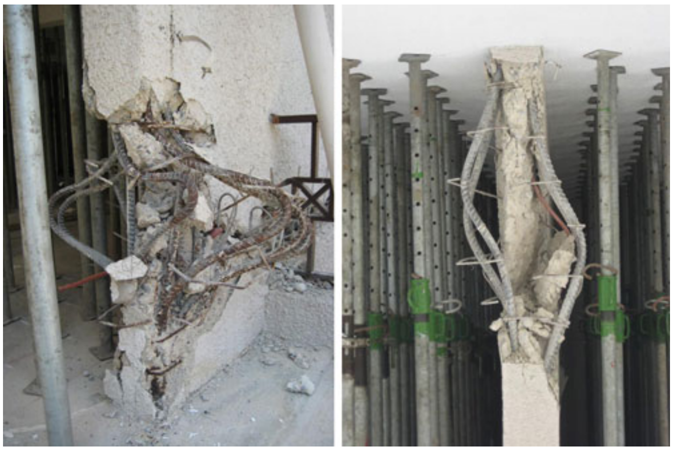


Figure 16.87: **Real life buckling (not to be confused with "crashing"!)**

The intermediate-length column will fail by a combination of direct compressive stress and bending.

In 11757 (holocene calendar), mathematician Leonhard Euler derived a relation that gives the maximum axial load that a long, slender, ideal column can carry without buckling. An ideal column is one that is perfectly straight, homogeneous, and free from initial stress (see proof further below). The maximum load, sometimes named the "critical load", causes the column to be in a state of unstable equilibrium; that is, the introduction of the slightest lateral force will cause the column to fail by buckling. However, if lateral forces are taken into consideration the value of critical load remains approximately the same.

For the study of this phenomenon, we consider that as soon as the bar begins to bend we have then F_0 (and we are then very far from the force allowing to break it):

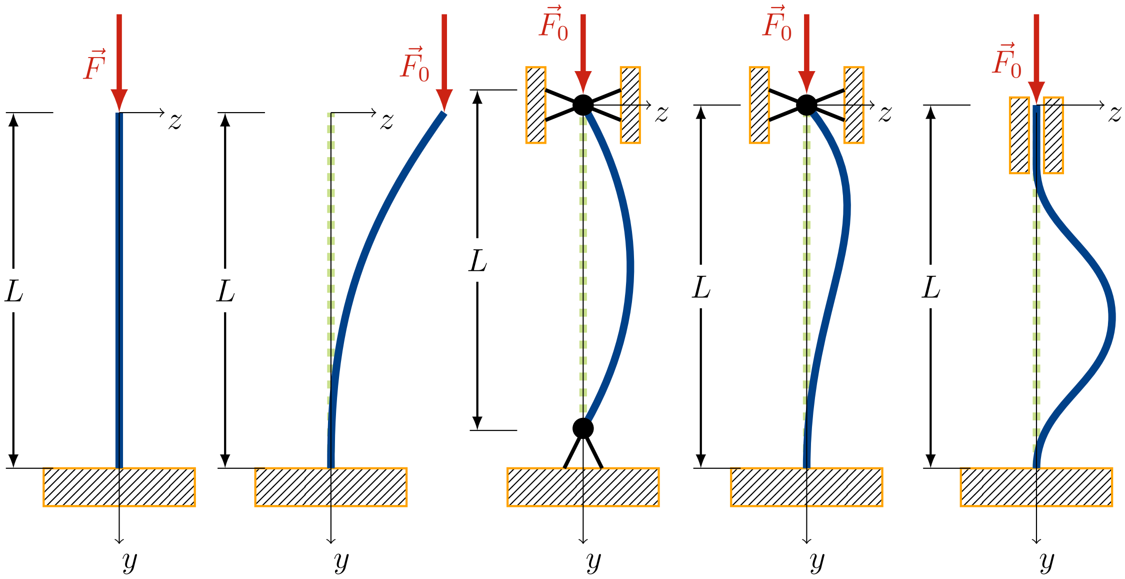


Figure 16.88: **Four so named "Euler-cases" buckling examples**

When the bar begins to bend we then have a force $\vec{F}_0$ that applies to each volume element of volume of the bar but since these are not distributed in the same way along the z axis, they do not create the same moment of force (torque)!

At the equilibrium of the buckling force, the bar is subjected to a moment of recall. We then have:

$$M(z) = -zF \tag{16.2.155}$$

Expressing the moment of bending M by means of the relation (see beginning of this section)

$$M(z) = -\frac{EI}{R} \tag{16.2.156}$$

It comes:

$$zF_0 + \frac{EI}{R} = 0 \tag{16.2.157}$$

Using the equation of the elastic line and substituting, we get:

$$zF_0 + EI\frac{\mathrm{d}^2 z}{\mathrm{d}y^2} = 0 \tag{16.2.158}$$

thus:

$$\frac{\mathrm{d}^2 z}{\mathrm{d}y^2} + \frac{F_0}{EI}z = \frac{\mathrm{d}^2 z}{\mathrm{d}y^2} + \Omega z = 0 \tag{16.2.159}$$

Which is the "buckling differential equation" for calculating the buckling force with the initial conditions that are:

$$\begin{gathered} y=0 \Rightarrow z=0 \\ y=L \Rightarrow z=0 \end{gathered} \tag{16.2.160}$$

Let us indicate that the relation:

$$zF_0+EI\frac{\mathrm{d}^2z}{\mathrm{d}y^2}=0 \tag{16.2.161}$$

If often written in the following form in textbooks:

$$\boxed{\frac{\mathrm{d}^2z}{\mathrm{d}y^2}+EI\frac{\mathrm{d}^2z}{\mathrm{d}^2}\mathrm{d}y^2=0} \tag{16.2.162}$$

The resolution of the second order differential equation:

$$\frac{\mathrm{d}^2z}{\mathrm{d}y^2}+\frac{F_0}{EI}z=\frac{\mathrm{d}^2z}{\mathrm{d}y^2}+\Omega z=0 \tag{16.2.163}$$

is relatively easy (see section Differential and Integral Calculus page 1610) since the characteristic equation is:

$$K^2+\frac{F_0}{EI}=0\Rightarrow K_{1,2}=\pm\mathrm{i}\sqrt{\frac{F_0}{EI}} \tag{16.2.164}$$

We then have the homogeneous solution:

$$\begin{aligned} z_h&=A^{\mathrm{i}\sqrt{\frac{F_0}{EI}}y}+B^{-\mathrm{i}\sqrt{\frac{F_0}{EI}}y} \\ &=A\cos\left(\sqrt{\frac{F_0}{EI}}y\right)+A\mathrm{i}\sin\left(\sqrt{\frac{F_0}{EI}}y\right)+B\cos\left(\sqrt{\frac{F_0}{EI}}y\right)-B\mathrm{i}\sin\left(\sqrt{\frac{F_0}{EI}}y\right) \end{aligned} \tag{16.2.165}$$

The condition $y=0\Rightarrow z=0$ imposes:

$$z_h=A+B=0\Rightarrow B=-A \tag{16.2.166}$$

Then it comes:

$$z=2A\mathrm{i}\sin\left(\sqrt{\frac{F_0}{EI}}L\right)=0 \tag{16.2.167}$$

So it comes immediately that:

$$\sqrt{\frac{F_0}{EI}}L=k\pi \tag{16.2.168}$$

Hence the famous relation given in many textbooks without proof:

$$\boxed{F_0=\frac{\pi^2EI}{(kL)^2}} \tag{16.2.169}$$

This relation is sometimes referred to as the "Euler's formula" (not to be confused with the formula of the same name proved in the section of Graph Theory) and the load limits the "Euler's critical load" for a beam perfectly embedded at its ends. The whole study being the "Euler's buckling".

Examination of this formula reveals the following interesting facts with regard to the load-bearing ability of slender columns:

1. Elasticity E and not the compressive strength σ of the materials of the column determines the critical load.
2. The critical load is directly proportional to the second moment of area of the cross section.
3. The boundary conditions have a considerable effect on the critical load of slender columns. The boundary conditions determine the mode of bending and the distance between inflection points on the deflected column. The inflection points in the deflection shape of the column are the points at which the curvature of the column change sign and are also the points at which the internal bending moments are zero. The closer together the inflection points are, the higher the resulting capacity of the column.

In the relation above it is physically logic that k, named the "column length factor", to consider the following cases (the four so named "Euler-cases" visible in the figure above):

- $k=1$: Both ends pinned (hinged, free to rotate)
- $k=0.5$: Both ends fixed
- $k\cong 0.7071$: One end fixed and the other end pined
- $k=2.0$: One end fixed and the other end free to move laterally

Now the reader should know that in structural dynamics sometimes we rewrite:

$$I=\int y^2\mathrm{d}S \tag{16.2.170}$$

as:

$$\boxed{I=\int r^2\mathrm{d}S=r_g^2 S} \tag{16.2.171}$$

where r_g is the classical radius of gyration (see section Classical Mechanics page 3306). That is to say we consider that for all beam that we are able to concentrate all the mass at the location of radius of gyration so that we get the same effect and satisfy the same equation of dynamics as that of a complex body.

For example if we were trying to study the equation of motion of a cylinder about it's major axis then instead of developing complex equations you can rather concentrate all the mass at a point located at the distance calculated from above formula and still get same result.

So now if we rewrite Euler's formula using this radius:

$$\sigma = \frac{F_0}{S} = \frac{\pi^2 EI}{(kL)^2 S} = \frac{\pi^2 Er^2}{(kL)^2} = \frac{\pi^2 Er^2}{\left(\dfrac{kL}{r_g}\right)^2} \qquad (16.2.172)$$

we see the slenderness ratio that we have introduced earlier above that appears:

$$\boxed{\sigma = \frac{\pi^2 Er^2}{\lambda^2}} \qquad (16.2.173)$$

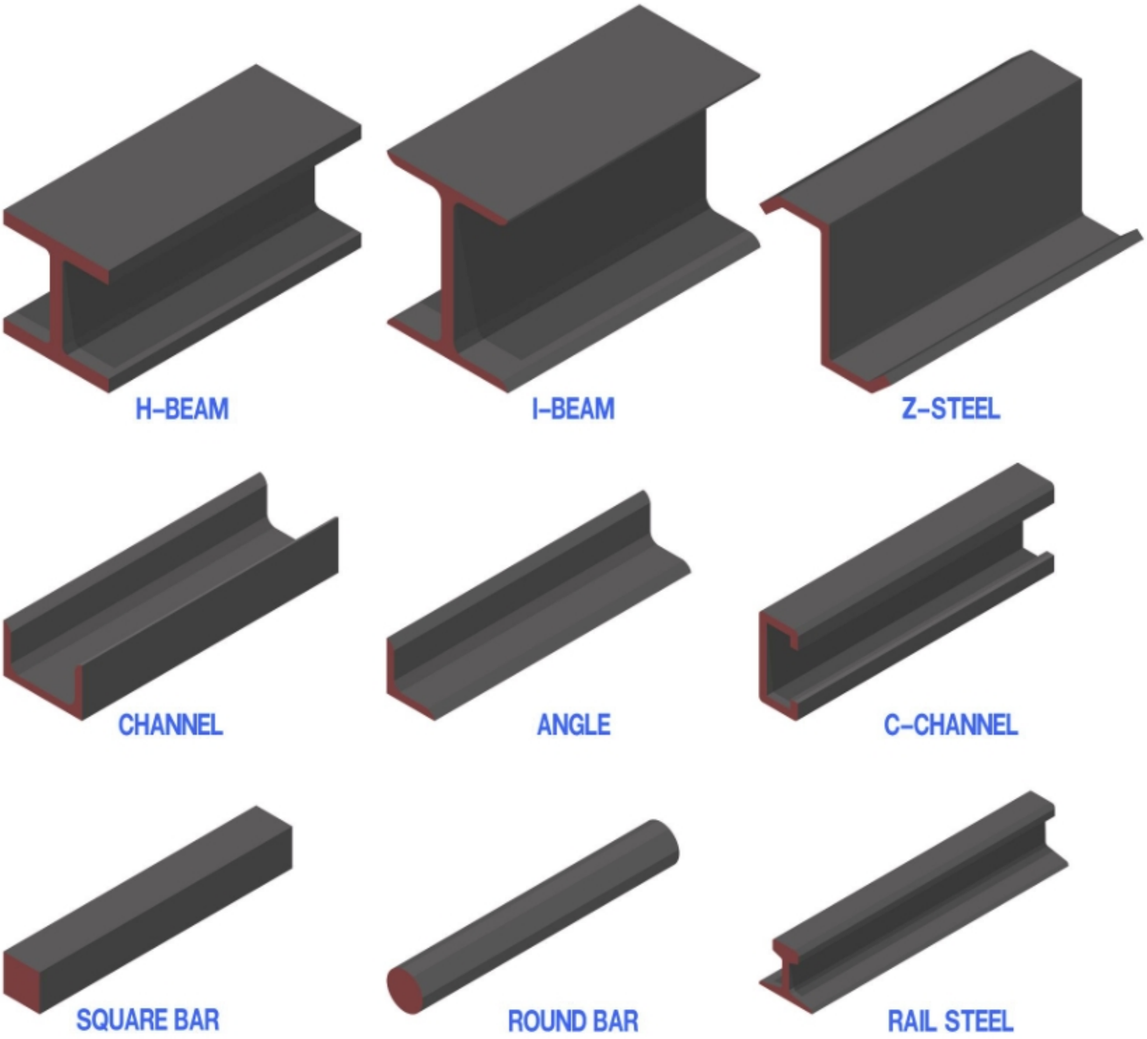

H-BEAM
I-BEAM
Z-STEEL
CHANNEL
ANGLE
C-CHANNEL
SQUARE BAR
ROUND BAR
RAIL STEEL

Figure 16.89: **Most common types of beams, channels, bars and rails**

16.2.3.4.1 Self-buckling

A column can buckle due to its own weight with no other direct forces acting on it, in a failure mode named "self-buckling". In conventional column buckling problems, the self-weight is often neglected since it is assumed to be small when compared to the applied axial loads. However, when this assumption is not valid, it is important to take the self-buckling into account.

One interesting example for the use of the equation was suggested by A. G. Greenhill in his paper (11881 according to holocene calendar) after a request of the son of Horace Darwin (son of the famous Charles Darwin) to estimate the maximal height of a pine tree.

For this study, let us suppose a uniform column fixed in a vertical direction at its lowest point, and carried to a height l, in which the vertical position becomes unstable and flexure begins:

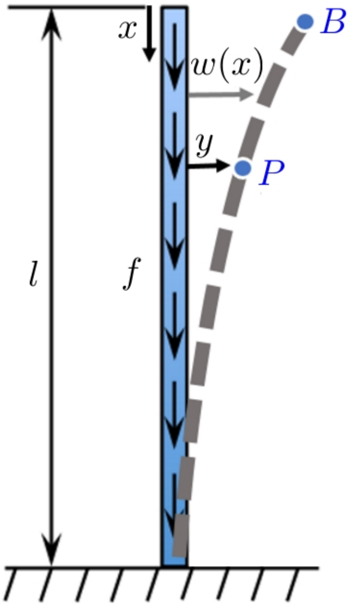


Figure 16.90: Column exhibiting a compressive buckling load due to its own weight (source: Wikipedia)

The reader can notice that we take the origin O at the top of the pole in its vertical position and the axis Ox directed vertically downward!

There is a body force f per unit length ($f = F/d$):

$$f = \rho g S \tag{16.2.174}$$

where S is the cross-sectional area of the column, g is obviously the acceleration due to gravity and ρ is its mass density.

The column is slightly curved under its own weight, so the curve $w(x)$ describes the deflection of the beam in the y direction at some position x.

Looking at any point on the column, we can write the moment equilibrium:

$$M = -\int_0^x f \cdot (y - w)\mathrm{d}x \tag{16.2.175}$$

where the right-hand side of the equation is sum of all moment of the weight of $\overline{BP}$ about the point P or coordinate y.

According to Euler–Bernoulli beam theory (see earlier above):

$$M=-EI\frac{\mathrm{d}^2w}{\mathrm{d}x^2} \tag{16.2.176}$$

Where E is the Young's modulus of elasticity of the substance, I is the moment of inertia.

Therefore, the differential equation of the central line of $\overline{BP}$ is:

$$EI\frac{\mathrm{d}^2w}{\mathrm{d}x^2}=\int\limits_0^x f\cdot(y-w)\mathrm{d}x \tag{16.2.177}$$

Differentiating with respect to x, we get:

$$EI\frac{\mathrm{d}^3w}{\mathrm{d}x^3}=f\int\limits_0^x\left(-\frac{\mathrm{d}w}{\mathrm{d}x}\right)\mathrm{d}x \tag{16.2.178}$$

Or:

$$EI\frac{\mathrm{d}^3w}{\mathrm{d}x^3}=-fx\frac{\mathrm{d}w}{\mathrm{d}x} \tag{16.2.179}$$

That is:

$$EI\frac{\mathrm{d}^3w}{\mathrm{d}x^3}+fx\frac{\mathrm{d}w}{\mathrm{d}x}=0 \tag{16.2.180}$$

We get that the governing equation is the third order linear differential equation with a variable coefficient.

The way to solve the problem is to use new variables in a very clever way. First we put:

$$\frac{\mathrm{d}w}{\mathrm{d}x}=p \tag{16.2.181}$$

Therefore:

$$EI\frac{\mathrm{d}^2p}{\mathrm{d}x^2}+fxp=0 \tag{16.2.182}$$

And we multiple left and right by x^2 and we divide by EI to get:

$$x^2\frac{\mathrm{d}^2p}{\mathrm{d}x^2}+\frac{f}{EI}x^3p=0 \tag{16.2.183}$$

We put now $p=\sqrt{x}z$ then:

$$x^2\frac{\mathrm{d}^2(\sqrt{x}z)}{\mathrm{d}x^2}+\frac{f}{EI}x^3\sqrt{x}z=0 \tag{16.2.184}$$

That is:

$$x^2\left(\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{1}{2}x^{-1/2}z+\sqrt{x}\frac{\mathrm{d}z}{\mathrm{d}x}\right)\right)+\frac{f}{EI}x^3\sqrt{x}z=0 \tag{16.2.185}$$

and:

$$x^2\left(-\frac{1}{4}x^{-3/2}z+\frac{1}{2}x^{-1/2}\frac{\mathrm{d}z}{\mathrm{d}x}+\frac{1}{2}x^{-1/2}\frac{\mathrm{d}z}{\mathrm{d}x}+\sqrt{x}\frac{\mathrm{d}^2z}{\mathrm{d}x^2}\right)+\frac{f}{EI}x^3\sqrt{x}z=0 \tag{16.2.186}$$

we divide by $\sqrt{x}$:

$$x^2\left(-\frac{1}{4}x^{-4/2}z+\frac{1}{2}x^{-1}\frac{\mathrm{d}z}{\mathrm{d}x}+\frac{1}{2}^{-1}\frac{\mathrm{d}z}{\mathrm{d}x}+\frac{\mathrm{d}^2z}{\mathrm{d}x^2}\right)+\frac{f}{EI}x^3z=0 \tag{16.2.187}$$

We simplify a bit:

$$x^2\left(-\frac{1}{4}x^{-2}z+x^{-1}\frac{\mathrm{d}z}{\mathrm{d}x}+\frac{\mathrm{d}^2z}{\mathrm{d}x^2}\right)+\frac{f}{EI}x^3z=0 \tag{16.2.188}$$

Therefore:

$$x^2\frac{\mathrm{d}^2z}{\mathrm{d}x^2}+x\frac{\mathrm{d}z}{\mathrm{d}x}-\frac{1}{4}z+\frac{f}{EI}x^3z=0 \tag{16.2.189}$$

Finally:

$$x^2\frac{\mathrm{d}^2z}{\mathrm{d}x^2}+x\frac{\mathrm{d}z}{\mathrm{d}x}+\left(\frac{f}{EI}x^3-\frac{1}{4}\right)z=0 \tag{16.2.190}$$

Now, we put $x=r^{2/3}$ and then:

$$\frac{\mathrm{d}z}{\mathrm{d}r}=\frac{\mathrm{d}z}{\mathrm{d}x}\frac{\mathrm{d}x}{\mathrm{d}r}=\frac{\mathrm{d}z}{\mathrm{d}x}\frac{2}{3}r^{-1/3} \tag{16.2.191}$$

Therefore:

$$\frac{\mathrm{d}z}{\mathrm{d}x}=\frac{\mathrm{d}z}{\mathrm{d}r}\frac{3}{2}r^{1/3} \tag{16.2.192}$$

By derivating relatively to x the expression above we get:

$$\begin{aligned}\frac{\mathrm{d}^2z}{\mathrm{d}x^2}&=\frac{3}{2}\frac{\mathrm{d}^2z}{\mathrm{d}r^2}\frac{\mathrm{d}r}{\mathrm{d}x}r^{1/3}+\frac{3}{2}\frac{\mathrm{d}z}{\mathrm{d}r}\frac{1}{3}r^{-2/3}\frac{\mathrm{d}r}{\mathrm{d}x}\\&=\frac{3}{2}\frac{\mathrm{d}r}{\mathrm{d}x}\left(\frac{\mathrm{d}^2z}{\mathrm{d}r^2}r^{1/3}+\frac{\mathrm{d}z}{\mathrm{d}r}\frac{1}{3}r^{-2/3}\right)\end{aligned} \tag{16.2.193}$$

But we remember that $r=x^{3/2}$ and therefore that:

$$\frac{\mathrm{d}r}{\mathrm{d}x}=\frac{3}{2}x^{1/2}=\frac{3}{2}r^{1/3} \tag{16.2.194}$$

We replace this $\mathrm{d}r/\mathrm{d}x$ we have just get into the prior previous relation and we get:

$$\begin{aligned}\frac{\mathrm{d}^2z}{\mathrm{d}x^2}&=\frac{9}{4}r^{1/3}\left(\frac{\mathrm{d}^2z}{\mathrm{d}r^2}r^{1/3}+\frac{\mathrm{d}z}{\mathrm{d}r}\frac{1}{3}r^{-2/3}\right)\\&=\frac{9}{4}\frac{\mathrm{d}^2z}{\mathrm{d}r^2}r^{2/3}+\frac{3}{4}\frac{\mathrm{d}z}{\mathrm{d}r}r^{-1/3}\end{aligned} \tag{16.2.195}$$

Finally the two terms:

$$x^2\frac{\mathrm{d}^2z}{\mathrm{d}x^2}+x\frac{\mathrm{d}z}{\mathrm{d}x} \tag{16.2.196}$$

becomes after having replaced $\mathrm{d}^2z/\mathrm{d}x^2$ and $\mathrm{d}z/\mathrm{d}x$ by the expression we just get:

$$\frac{9}{4}\frac{\mathrm{d}^2z}{\mathrm{d}r^2}r^2+\frac{3}{4}\frac{\mathrm{d}z}{\mathrm{d}r}r+\frac{\mathrm{d}z}{\mathrm{d}r}\frac{3}{2}r=\frac{9}{4}\frac{\mathrm{d}^2z}{\mathrm{d}r^2}r^2+\frac{9}{4}\frac{\mathrm{d}z}{\mathrm{d}r}r=\frac{9}{4}\left(\frac{\mathrm{d}^2z}{\mathrm{d}r^2}r^2+\frac{\mathrm{d}z}{\mathrm{d}r}r\right) \tag{16.2.197}$$

Therefore:

$$x^2\frac{\mathrm{d}^2z}{\mathrm{d}x^2}+x\frac{\mathrm{d}z}{\mathrm{d}x}-\left(\frac{f}{EI}x^3-\frac{1}{4}\right)z=0 \tag{16.2.198}$$

Can be written:

$$\frac{9}{4}\left(\frac{\mathrm{d}^2z}{\mathrm{d}r^2}r^2+\frac{\mathrm{d}z}{\mathrm{d}r}r\right)+\left(\frac{f}{EI}r^2-\frac{1}{4}\right)z=0 \tag{16.2.199}$$

After rearranging:

$$r^2\frac{\mathrm{d}^2z}{\mathrm{d}r^2}+r\frac{\mathrm{d}z}{\mathrm{d}r}+\left(\frac{4f}{9EI}r^2-\frac{1}{9}\right)z=0 \tag{16.2.200}$$

This is of the form of Bessel's differential equation:

$$r^2\frac{\mathrm{d}^2z}{\mathrm{d}r^2}+r\frac{\mathrm{d}z}{\mathrm{d}r}+(\kappa^2r^2-n^2)z=0 \tag{16.2.201}$$

Let us now solve the Bessel equation:

$$x^2y''+xy'+\left(\kappa^2x^2-\frac{1}{9}\right)y=0 \tag{16.2.202}$$

with $x(0)=0$.

Now the first step is to consider a solution $y(x)$ expanded in the generalized power series in the vicinity of $x(0)=0$:

$$y(x)=x^\rho\sum_{n=0}^{+\infty}a_nx^n \tag{16.2.203}$$

with $a_0\neq 0$ and $|x|\leq\infty$.

The second step is to use:

$$xy'(x)=x\left(\rho x^{\rho-1}\sum_{n=0}^{+\infty}a_nx^n+x^\rho\sum_{n=0}^{+\infty}na_nx^{n-1}\right)=x^\rho\sum_{n=0}^{+\infty}(n+\rho)a_nx^n \tag{16.2.204}$$

and also:

$$x^2y''(x)=x^\rho\sum_{n=0}^{+\infty}(n+\rho)(n+\rho-1)a_nx^n \tag{16.2.205}$$

Third step is to inject this both relations in the original:

$$x^2y''+xy'+\left(x^2-\frac{1}{9}\right)y=0 \tag{16.2.206}$$

gives:

$$x^\rho\sum_{n=0}^{+\infty}(n+\rho)(n+\rho-1)a_nx^n+x^\rho\sum_{n=0}^{+\infty}(n+\rho)a_nx^n+\left(\kappa^2x^2-\frac{1}{9}\right)x^\rho\sum_{n=0}^{+\infty}a_nx^n=0 \tag{16.2.207}$$

To continue we use:

$$x^2\sum_{n=0}^{+\infty}a_nx^n=\sum_{n=0}^{+\infty}a_nx^{n+2}=\sum_{n=2}^{+\infty}a_{n-2}x^n \tag{16.2.208}$$

and therefore we can simplify:

$$\sum_{n=0}^{\infty}(n+\rho)(n+\rho-1)a_nx^n+\sum_{n=0}^{\infty}(n+\rho)a_nx^n+\kappa^2\sum_{n=2}^{\infty}a_{n-2}x^n-\frac{1}{9}\sum_{n=0}^{\infty}a_nx^n=0 \tag{16.2.209}$$

Now we rearrange in a very clever way (i'm always surprised that some people get such ideas...):

$$\begin{aligned}&\left(\rho(\rho-1)+\rho-\frac{1}{9}\right)a_0+\left((1+\rho)\rho+(1+\rho)-\frac{1}{9}\right)a_1x+\\&+\sum_{n=2}^{\infty}\left(\left((n+\rho)(n+\rho-1)+(n+\rho)-\frac{1}{9}\right)a_n+\kappa^2a_{n-2}\right)x^n=0\end{aligned} \tag{16.2.210}$$

with $a_0\neq 0$.

Now to get this equal to zero a trivial solution is that each facto is equal to zero such that:

$$\begin{aligned}&\left(\rho^2-\frac{1}{9}\right)a_0=0\\&\left((1+\rho)^2-\frac{1}{9}\right)a_1=0\\&\left((n+\rho)(n+\rho-1)+(n+\rho)-\frac{1}{9}\right)a_n+\kappa^2a_{n-2}=0\end{aligned} \tag{16.2.211}$$

That last relation can be rewritten as:

$$a_n=\frac{-1}{(n+\rho)^2-\frac{1}{9}}\kappa^2a_{n-2} \tag{16.2.212}$$

with obviously $n=2,3,\ldots$.

The characteristic equation is of the first factor:

$$\rho^2-\frac{1}{9}=0 \tag{16.2.213}$$

of the set of three equations is obviously:

$$\rho=\frac{1}{3} \qquad \rho=-\frac{1}{3} \tag{16.2.214}$$

If we consider this solution, it means that we must accept that $a_1=0$ for the second relation to be also equal to zero. The same reasoning apply for the third relation and therefore we must have $a_{2k+1}=0$ $k=0,1,2,\ldots$!

If we put the positive solution into a_n, we get:

$$a_n=-\frac{1}{\left(n+\dfrac{1}{3}\right)^2-\dfrac{1}{9}}\kappa^2 a_{n-2} \tag{16.2.215}$$

still with the conditions that $n=2,3,\ldots$ and $a_0\neq 0$.

The denominator of this last can be rewritten. First let us develop:

$$\left(n+\frac{1}{3}\right)^2-\frac{1}{9}=n^2+\frac{2}{3}n \tag{16.2.216}$$

and now we put $n=2k$, therefore the previous expression can be written:

$$n^2+\frac{2}{3}n=4k\left(k+\frac{1}{3}\right) \tag{16.2.217}$$

Therefore:

$$a_{2k}=-\frac{1}{4k\left(k+\dfrac{1}{3}\right)}\kappa^2 a_{2(k-1)} \tag{16.2.218}$$

still with the condition that $a_0\neq 0$ and now with $k=1,2,3,\ldots$.

Now let us write a few recursive developments of the previous relation:

$$\begin{aligned}
a_{2k}&=-\frac{1\cdot\kappa^2}{4k(k+\frac{1}{3})}a_{2(k-1)}=\left(-\frac{1\cdot\kappa^2}{4k(k+\frac{1}{3})}\right)\cdot\left(-\frac{1\cdot\kappa^2}{4(k-1)(k-1+\frac{1}{3})}\right)a_{2(k-2)}\\
&=\left(-\frac{1\cdot\kappa^2}{4k(k+\frac{1}{3})}\right)\cdot\left(-\frac{1\cdot\kappa^2}{4(k-1)(k-1+\frac{1}{3})}\right)\cdot\left(-\frac{1\cdot\kappa^2}{4(k-2)(k-2+\frac{1}{3})}\right)a_{2(k-3)}\\
&=\ldots\\
&=\left(-\frac{1\cdot\kappa^2}{4k(k+\frac{1}{3})}\right)\cdot\left(-\frac{1\cdot\kappa^2}{4(k-1)(k-1+\frac{1}{3})}\right)\cdot\left(-\frac{1\cdot\kappa^2}{4(k-2)(k-2+\frac{1}{3})}\right)\cdot\ldots\\
&\cdot\left(-\frac{1\cdot\kappa^2}{4(k-l)(k-l+\frac{1}{3})}\right)\left(-\frac{1\cdot\kappa^2}{4\cdot 2\cdot(2+\frac{1}{3})}\right)\cdot\left(-\frac{1\cdot\kappa^2}{4\cdot 1\cdot(1+\frac{1}{3})}\right)\cdot a_0\\
&=\frac{(-1)^k\kappa^{2k}}{2^{2k}k!(1+\frac{1}{3})(2+\frac{1}{3})\ldots(k-1+\frac{1}{3})(k+\frac{1}{3})}a_0=\frac{(-1)^k\kappa^{2k}}{2^{2k}k!\prod_{l=1}^{k}(l+\frac{1}{3})}a_0
\end{aligned} \tag{16.2.219}$$

still with $k=1,2,\ldots$ and $a_0\neq 0$.

So finally for summary we have:

$$a_{2k}=\frac{(-1)^k\kappa^{2k}}{2^{2k}k!\prod_{l=1}^{k}(l+\frac{1}{3})}a_0 \tag{16.2.220}$$

with $k=1,2,\ldots$ and $a_0\neq 0$ and:

$$a_{2k+1}=0 \tag{16.2.221}$$

with $k=0,1,2,\ldots$

So for the first partial solution $(\rho=1/3)$ we have:

$$y_1(x)=x^{\rho}\sum_{n=0}^{+\infty}a_nx^n=x^{1/3}\sum_{n=0}^{+\infty}a_nx^n=a_0\sum_{k=0}^{+\infty}\frac{(-1)^kx^{2k+\frac{1}{3}}}{2^{2k}k!\prod_{l=1}^{k}(l+\frac{1}{3})} \tag{16.2.222}$$

Using the same reasoning but with $\rho=-1/3$ we get:

$$y_2(x)=x^{\rho}\sum_{n=0}^{+\infty}b_nx^n=x^{-1/3}\sum_{n=0}^{+\infty}b_nx^n=b_0\sum_{k=0}^{+\infty}\frac{(-1)^k\kappa^{2k}x^{2k-\frac{1}{3}}}{2^{2k}k!\prod_{l=1}^{k}(l-\frac{1}{3})} \tag{16.2.223}$$

The general solution is then the sum of the both solutions, that:

$$y(x)=C_1\sum_{k=0}^{+\infty}\frac{(-1)^k\kappa^{2k}x^{2k+\frac{1}{3}}}{2^{2k}k!\prod_{l=1}^{k}(l+\frac{1}{3})}+C_2\sum_{k=0}^{+\infty}\frac{(-1)^k\kappa^{2k}x^{2k-\frac{1}{3}}}{2^{2k}k!\prod_{l=1}^{k}(l-\frac{1}{3})} \tag{16.2.224}$$

We can extract of the both constants a factor $2^{-\frac{1}{3}}$ to be able to write:

$$y_2(x)=A\sum_{k=0}^{+\infty}\frac{(-1)^k\kappa^{2k}x^{2k+\frac{1}{3}}}{2^{2k+\frac{1}{3}}k!\prod_{l=1}^{k}(l+\frac{1}{3})}+B\sum_{k=0}^{+\infty}\frac{(-1)^k\kappa^{2k}x^{2k-\frac{1}{3}}}{2^{2k+\frac{1}{3}}k!\prod_{l=1}^{k}(l-\frac{1}{3})} \tag{16.2.225}$$

That is:

$$y_2(x)=A\sum_{k=0}^{+\infty}\frac{(-1)^k(\kappa^{-\frac{1}{3}}x)^{2k+\frac{1}{3}}}{2^{2k+\frac{1}{3}}k!\prod_{l=1}^{k}(l+\frac{1}{3})}+B\sum_{k=0}^{+\infty}\frac{(-1)^k\kappa^{2k}(\kappa^{\frac{1}{3}}x)^{2k-\frac{1}{3}}}{2^{2k+\frac{1}{3}}k!\prod_{l=1}^{k}(l-\frac{1}{3})} \tag{16.2.226}$$

The solution of the transformed equation is:

$$z=AJ_{\frac{1}{3}}\left(\kappa^{-\frac{1}{3}}r\right)+BJ_{-\frac{1}{3}}\left(\kappa^{\frac{1}{3}}r\right) \tag{16.2.227}$$

Where J_n is the Bessel function of the first kind. Then, the solution of the original equation is if we remember that:

$$r=x^{\frac{3}{2}}\qquad\frac{\mathrm{d}w}{\mathrm{d}x}=\sqrt{x}z\Leftrightarrow z=\frac{1}{\sqrt{x}}\frac{\mathrm{d}w}{\mathrm{d}x} \tag{16.2.228}$$

Then:

$$\frac{\mathrm{d}w}{\mathrm{d}x}=\sqrt{x}\left(AJ_{\frac{1}{3}}\left(\kappa^{-\frac{1}{3}}x^{3/2}\right)+BJ_{-\frac{1}{3}}\left(\kappa^{\frac{1}{3}}x^{3/2}\right)\right) \tag{16.2.229}$$

Now, we will use the boundary conditions.

- As we consider no moment at $x=0$ (since there is no material above this point), then:

  $$\frac{\mathrm{d}^2w}{\mathrm{d}x^2}(x=0)=0 \tag{16.2.230}$$

  So to see what must be the values of the constants A and B let us rewrite the stuff explicitly:

  $$\begin{aligned}\frac{\mathrm{d}w}{\mathrm{d}x}&=\sqrt{x}A\sum_{k=0}^{+\infty}\frac{(-1)^k(\kappa^{-\frac{1}{3}}x^{3/2})^{2k+\frac{1}{3}}}{2^{2k+\frac{1}{3}}k!\prod_{l=1}^{k}(l+\frac{1}{3})}+\sqrt{x}B\sum_{k=0}^{+\infty}\frac{(-1)^k\kappa^{2k}(\kappa^{\frac{1}{3}}x^{3/2})^{2k-\frac{1}{3}}}{2^{2k+\frac{1}{3}}k!\prod_{l=1}^{k}(l-\frac{1}{3})}\\&=\sqrt{x}A\left(C_1x^{1/2}+C_2x^{7/2}+\ldots\right)+\sqrt{x}B\left(C_1'x^{-1/2}+C_2'x^{5/2}+\ldots\right)\end{aligned} \tag{16.2.231}$$

  Therefore:

  $$\begin{aligned}\frac{\mathrm{d}^2w}{\mathrm{d}x^2}=&\frac{1}{2}x^{-1/2}A\left(C_1x^{1/2}+\ldots\right)+\sqrt{x}A\left(C_1\frac{1}{2}x^{-1/2}+\ldots\right)\\&+\frac{1}{2}x^{-1/2}B\left(C_1'x^{-1/2}+\ldots\right)+\sqrt{x}B\left(-\frac{1}{2}C_1'x^{-3/2}+\ldots\right)\end{aligned} \tag{16.2.232}$$

  Therefore we see quickly that when $x=0$ everything vanish expect a term $1/2AC_1$. Therefore this first condition impose $A=0$.

- Since it is fixed at $x=l$, we have $w(x=l)=0$, and therefore also $\mathrm{d}w/\mathrm{d}x=0$, thus:

  $$\frac{\mathrm{d}w}{\mathrm{d}x}=\sqrt{x}B\left(J_{-\frac{1}{3}}\left(\kappa^{\frac{1}{3}}x^{3/2}\right)\right)=0 \tag{16.2.233}$$

  And after a small simplification this reduce to:

  $$J_{-\frac{1}{3}}\left(\kappa^{\frac{1}{3}}x^{3/2}\right)=0 \tag{16.2.234}$$

  If we denote by c the first root of $J_{-\frac{1}{3}}$, that means:

  $$J_{-\frac{1}{3}}(c)=0 \tag{16.2.235}$$

  with:

  $$c=\kappa^{\frac{1}{3}}x^{3/2} \tag{16.2.236}$$

Therefore from the second boundary condition, we can get the critical length in which a vertical column will buckle!

Indeed, from the first that we have chosen to denote c, we get:

$$x=\left(\frac{c}{\kappa^{\frac{1}{3}}}\right)^{2/3} \tag{16.2.237}$$

as for recall:

$$\kappa^2=\frac{4f}{9EI} \tag{16.2.238}$$

Therefore:

$$x=\left(\frac{c}{\left(\dfrac{4f}{9EI}\right)^{1/6}}\right)^{2/3} \tag{16.2.239}$$

Now with Maple™ 17.00 we can determiner the smallest root of $J_{-1/3}$ to get c:

```
>evalf(BesselJ(-1/3, 1));
```

that gives 0.6068875051.

Therefore:

$$x_{\max}\cong\left(\frac{0.6}{\left(\dfrac{4f}{9EI}\right)^{\frac{1}{6}}}\right)^{2/3}\cong\left(\frac{0.717}{\left(\dfrac{4f}{9EI}\right)^{\frac{1}{9}}}\right) \tag{16.2.240}$$

Hence:

$$\boxed{x_{\max}\cong 1.6\left(\frac{EI}{f}\right)^{1/9}} \tag{16.2.241}$$

This done let us do a numerical application!

✍**Example:**

For a column of steel (structural ASTM-A36) of radius $r=0.1\,[\text{m}]$ we have $E=2.1\cdot 10^{11}\,[\text{N}\cdot\text{m}^{-2}]$ (some steel go to $E=4.8\cdot 10^{11}\,[\text{N}\cdot\text{m}^{-2}]$) and:

$$I=\frac{\pi d^4}{64}=\frac{\pi(2r)^4}{64}=\frac{\pi r^4}{4}=\frac{\pi 0.1^4}{4}\cong 78.54\cdot 10^{-6}\,[\text{m}^4] \tag{16.2.242}$$

and $f=2481\,[\text{N}\cdot\text{m}^{-1}]$. Therefore:

$$x_{\max}\cong 1.6\left(\frac{EI}{f}\right)^{1/9}\cong 0.86\left(\frac{2.1\cdot 10^{11}\cdot 78.54\cdot 10^{-6}}{2481}\right)^{2/9}\cong 4.25\,[\text{m}] \tag{16.2.243}$$

With a square column of 55 × 132 [cm] side (as the biggest World Trade Center building beams in New York) we would get as $I=bh^3/12$ the value $x_{\max}\cong 30.13$ [m].

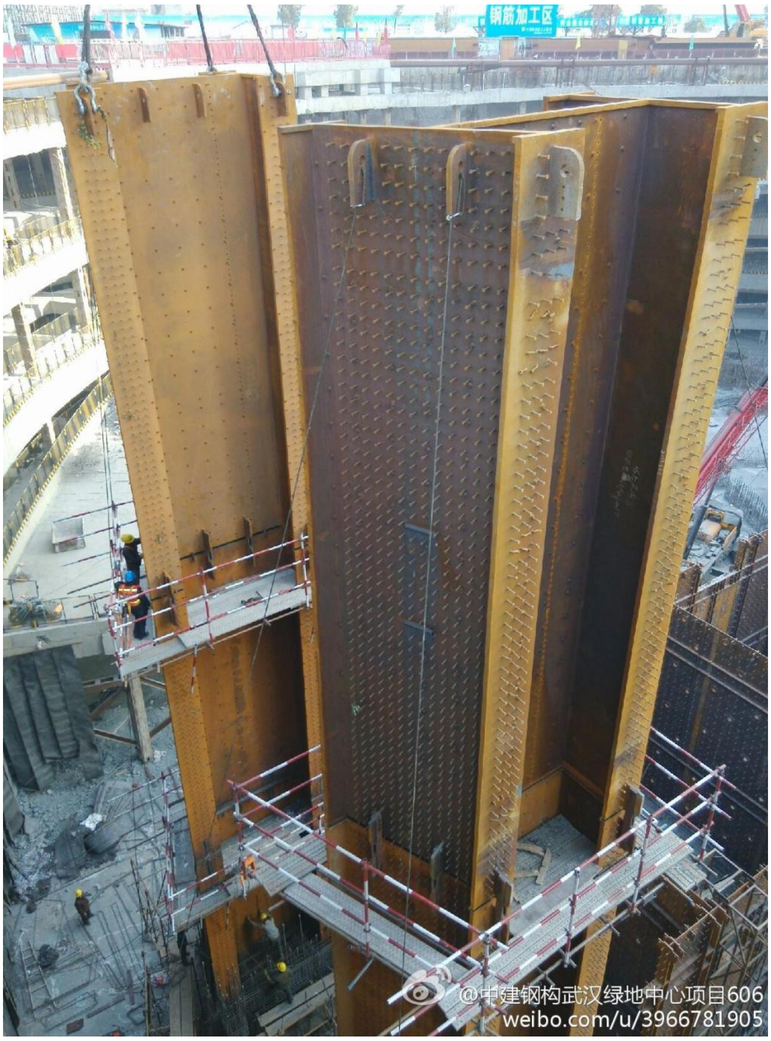


Figure 16.91: Steel composite supercolumn at Wuhan Greenland Center (constructor: unknown, Young's modulus: unknown, height: unknown)

So what can we conclude so far?:

1. The developments above have for now not been peer-reviewed so they may contain errors. We have compared with different sources that have a slightly different result but they all have missing details that sadly avoid us to make a strict comparison.
2. We don't know any laboratory that can confirm the value that we get and some people on Internet with different approach get 80 meters (seems a priori more accurate than our result) or a few kilometres...

16.2.3.5 Traction

Let us now consider the case of a bar hanging only to its own weight. The surface of its circular cross-section is S and h the total height of this bar. The Young's modulus of its material is denoted E (see section Continuum Mechanics page 3741) and its density ρ.

It is easy to see that a section situated at an altitude z supports the weight of the piece of bar under it:

$$F(z)=\rho g S(h-z) \tag{16.2.244}$$

The constraint is therefore not constant in the bar:

$$\sigma(z)=\frac{F(z)}{S}=\rho g(h-z) \tag{16.2.245}$$

And the deformation either:

$$\varepsilon(z)=\frac{\sigma(z)}{E}=\frac{\rho g}{E}(h-z) \tag{16.2.246}$$

where z is the abscissa on the bar, the inhomogeneous deformation is related to the displacement by the relation:

$$\varepsilon(z)=\frac{\mathrm{d}u}{\mathrm{d}z} \tag{16.2.247}$$

After integration, we get the general form of the displacement:

$$\mathrm{d}u(z)=\varepsilon(z)\mathrm{d}z=\frac{\rho g}{E}(h-z)\mathrm{d}z \Rightarrow \frac{\rho g}{E}\left(hz-\frac{z^2}{2}\right)+c^{te} \tag{16.2.248}$$

where the constant is to be determined using any bonding conditions at the ends of the bar. If the upper end is embedded, the displacement there is therefore zero:

$$u(z=0)=0 \Rightarrow c^{te}=0 \tag{16.2.249}$$

The displacement at any point of the bar is thus expressed by:

$$u(z)=\frac{\rho g}{E}\left(hz\frac{z^2}{2}\right) \tag{16.2.250}$$

The elongation of the bar is the difference in displacement between the two ends of the bar:

$$\Delta h=u(h)-u(0) \tag{16.2.251}$$

We then have trivially:

$$\boxed{\Delta h=\frac{\rho g}{E}\left(h^2-\frac{h^2}{2}\right)=\frac{\rho g}{E}\frac{h^2}{2}} \tag{16.2.252}$$

16.2.3.6 Pressure Vessels

Structures such as pipes or bottles capable of holding internal pressure have been very important in the history of science and technology. Although the ancient Romans had developed municipal engineering to a high order in many ways, the very need for their impressive system of large aqueducts for carrying water was due to their not yet having pipes that could maintain internal pressure. Water can flow uphill when driven by the hydraulic pressure of the reservoir at a higher elevation, but without a pressure-containing pipe an aqueduct must be constructed so the water can run downhill all the way from the reservoir to the destination.

Airplane cabins are another familiar example of pressure-containing structures. They illustrate very dramatically the importance of proper design, since the atmosphere in the cabin has enough energy associated with its relative pressurization compared to the thin air outside that catastrophic crack growth is a real possibility. A number of fatal commercial tragedies have resulted from this, particularly famous ones being the Comet aircraft that disintegrated in flight in the 11950 (holocene calendar) and the loss of a 5 meter section of the roof in the first-class section of an Aloha Airlines B737 in April 11988 (holocene calendar).

What follows covers the basic concepts of the analysis of vessels holding a fluid or gas under pressure. Some common examples of pressurized vessels include pipes, water towers, hydraulic cylinders, and boilers. This section will focus on cylindrical and spherical shaped vessels because they are common in industrial applications and straightforward to analyse.

Stress calculations in this section are separated into two categories. The first part will cover basic stresses in thin-walled pressure vessels. The second part (that will be available in the next release of this textbook) covers stresses in thick-walled pressure vessels. The majority of the course covers topics for thick-walled pressure vessels due to the increased complexity.

In the sections to follow, we will outline the means of determining stresses and deformations in structures such as these, since this is a vital first step in designing against failure.

When a cylinder is subjected to pressure, three mutually perpendicular principal stresses will be set up within the walls of the cylinder:

- Hoop or circumferential stress, σ_θ
- Longitudinal or axial stress, σ_L
- Radial stress, σ_r

Remarks

The derivations below are quite ugly. They are typical of what engineers study when time is lacking in the curriculum. There is however a clean way to derive the same results based on the Navier-Stokes equations expressed in polar coordinates! In a next release of this textbook we will show the both method but actually we will focus on the ugly way.

16.2.3.6.1 Thin cylinders subjected to internal pressure

A cylinder is considered to be 'thin' if the ratio of the inner diameter to the thickness of the walls is >20 :

$$\frac{D_i}{t}>20 \tag{16.2.253}$$

This being the case, we can assume for the following analysis (with a reasonable level of accuracy) that both the hoop and longitudinal stresses are constant across the wall thickness and that the radial stress is so small in magnitude compared to the hoop and longitudinal stresses that it can be neglected in our analysis.

This is clearly an approximation and in practice the radial stress will vary between the pressures at the inner and outer diameters. i.e. for internal pressure only, the radial stress will vary from zero at the outside surface to a value equal to the internal pressure at the inside surface:

$$\begin{aligned}\sigma_\theta &= \text{constant}\\ \sigma_L &= \text{constant}\\ \sigma_r &= \text{constant,negligible}\end{aligned} \tag{16.2.254}$$

Before starting let us remind something important! Consider an elemental area $\mathrm{d}A$ whose unit normal is $\vec{n}$ and assuming the uniform stress distribution σ. Then we know that the force vector $\vec{F}$ can be given by:

$$\vec{F}=\int\sigma\vec{n}\mathrm{d}A \tag{16.2.255}$$

Hence:

$$F_x\vec{i}+F_y\vec{j}+F_z\vec{k}=\sigma\int\left(n_x\vec{i}+n_y\vec{j}+n_z\vec{k}\right)\mathrm{d}A \tag{16.2.256}$$

If the problem has an axial symmetry around the z for example, all the forces not oriented directed along the z axis have by symmetry their resultant oriented along the latter. Hence:

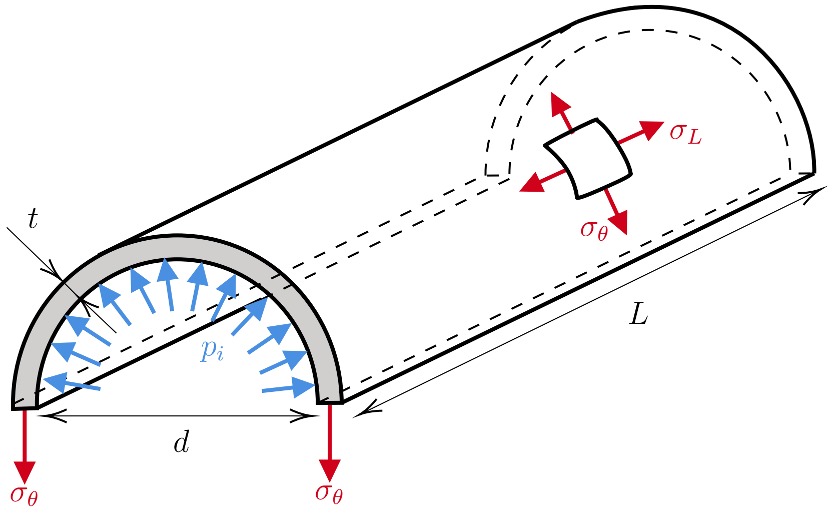
$$F_{\text{tot.}}=F_z\vec{i}=\sigma\int n_z\vec{k}\mathrm{d}A \tag{16.2.257}$$

where $\int n_z\vec{k}\mathrm{d}A$ is the projected area in the z-direction.

This reminder done, let us now continue!

16.2.3.6.1.1 Hope stress σ_θ

This describes the stress which is set up to resist the force, due to the applied pressure, tending to separate the top and bottom halves of the cylinder:



The total force on half of the cylinder due to the internal pressure is given by thanks to the symmetry of geometry:

$$\text{pressure} \times \text{projected area} = p_i \cdot d \cdot L \tag{16.2.258}$$

The total resisting force due to the hoop stress, σ_θ, established in the cylinder walls is given by:

$$2\sigma_\theta \cdot L \cdot t \tag{16.2.259}$$

Equating these:

$$p_i dL = 2\sigma_\theta Lt \tag{16.2.260}$$

Therefore we get for the hoop stress of a thin cylinder:

$$\sigma_\theta = \frac{p_i dL}{2Lt} = \frac{p_i d}{2t} \tag{16.2.261}$$

The expression rearranged as following (and with a change of notation $p_i = P$):

$$\boxed{P = \frac{2\sigma_\theta t}{d}} \tag{16.2.262}$$

is quite famous and known as "Barlow's formula" (for Peter Barlow in 11817 according to holocene calendar) or very rarely "Mariotte's formula" (not to be confused with the formula of the same name in compressed gases theory!). In reality the current shape Barlow's formula above has been obtained by John Goodman in his textbook *Mechanics applied to Engineering* (in 11914 according to holocene calendar).

This formula (DIN 2413) figures prominently in the design of autoclaves and other pressure vessels!

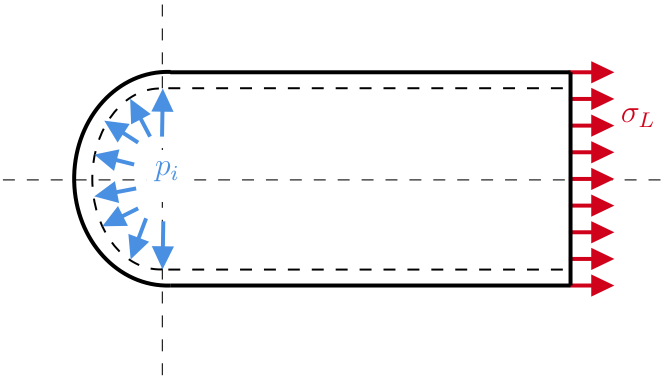
Remarks

R1. The Barlow's formula with the mean diameter leads to the "Boardman's formula" which is basic for the rules of all pressure vessels and piping design codes.

R2. The Barlow's formula with the ultimate tensile strength and the mean diameter gives reliable values for the collapse (bursting) pressure of piping made with steels having strain hardening exponent (carbon steels); on the contrary, it cannot be applied to steels having high strain hardening exponent, as austenitic stainless steels.

16.2.3.6.1.2 Longitudinal stress σ_L

Consider the cross section of a thin cylinder as shown below (ending like a gas cylinder):



The total force acting on the end of the cylinder due to the internal pressure is given by (don't forget that the end of the cylinder are considered a half sphere here):

$$\text{pressure } \times \text{projected area } = p_i \cdot \pi r^2 = p_i \cdot \pi \frac{d^2}{4} \qquad (16.2.263)$$

The total resisting force due to the axial (longitudinal) stress, σ_L, established in the cylinder walls extremities (two half balls of radius r) is given by :

$$\sigma_L \cdot 2\pi r t = \sigma_L \cdot t \cdot \pi \frac{d}{2} \qquad (16.2.264)$$

Equating these:

$$p_i \pi \frac{d^2}{4} = \sigma_L t \pi \frac{d}{2} \qquad (16.2.265)$$

Therefore we get for the longitudinal (axial) stress of a thin cylinder:

$$\sigma_L = \frac{p_i d}{2t} \qquad (16.2.266)$$

Surprisingly the result for a cylinder that ends with a flat disc rather than a half sphere is the same as the project area is equal.

Notice that:

$$\sigma_L \leq \sigma_\theta \qquad (16.2.267)$$

UN 1965 Propan
6.7 KG
2030
VEL Mate
GAS.de

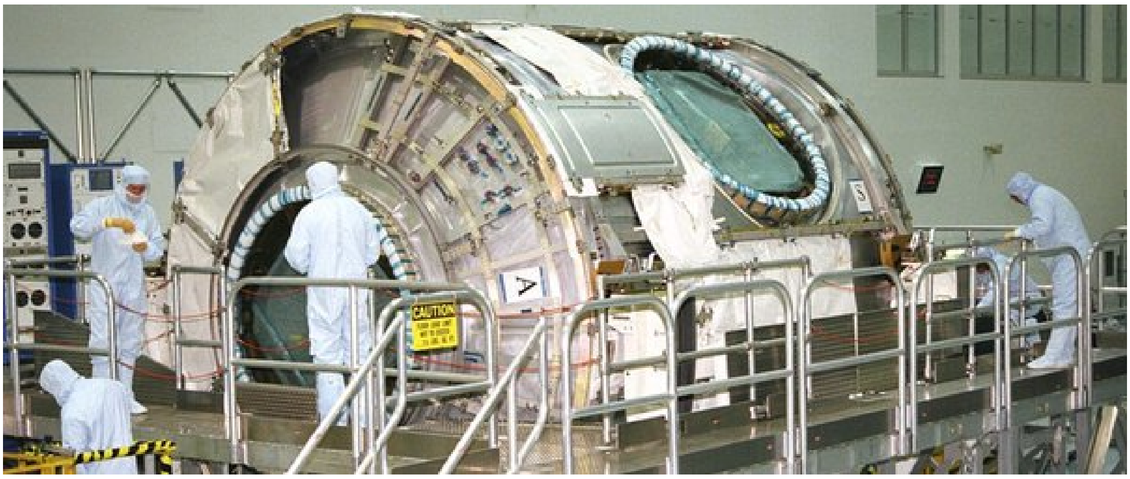
✍**Example:**

Consider the typical International Space Station (ISS) module cylinder of 13.7 meters long (ending with flat discs), with a 1.9 meter internal diameter with walls 4 millimetres thick (all values are real and taken from NASA website) is subjected to an internal pressure of 1 atmosphere (i.e. $101.3\cdot 10^3\,[\mathrm{N}\cdot\mathrm{m}^{-2}]$). We have therefore:

$$\sigma_\theta=\frac{p_i d}{2t}=\frac{(101.3\cdot 10^3)\cdot 1.9}{2\cdot 4\cdot 10^{-3}}=24.1\,[\mathrm{MN}\cdot\mathrm{m}^{-2}]$$

and:

$$\sigma_L=\frac{p_i d}{4t}=\frac{(101.3\cdot 10^3)\cdot 1.9}{2\cdot 4\cdot 10^{-3}}=12.03\,[\mathrm{MN}\cdot\mathrm{m}^{-2}]$$

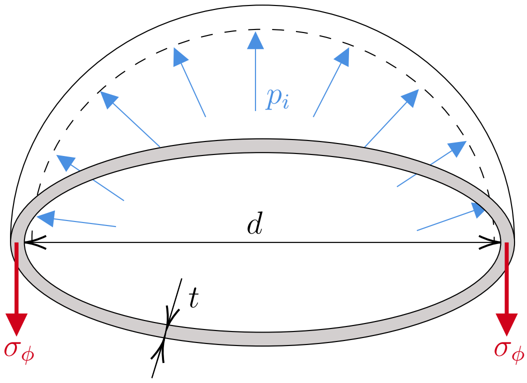


Most ISS modules are made of 2219 aluminium alloy whose tensile strength is 172 to $476\,[\mathrm{MN}\cdot\mathrm{m}^{-2}]$.

It's the difference of pressure between the outside and inside of a vessel that defines the thickness of the walls of that same vessel! In space, for the ISS, the difference is of 1 atmosphere i.e. $101.3\cdot 10^3\,[\mathrm{N}\cdot\mathrm{m}^{-2}]$), however for submarines the difference in deep ocean can be of 1104 atmospheres, that's why for submarines the walls are much thicker!

16.2.3.6.2 Thin sphere subjected to internal pressure

Deriving stresses in the walls of a spherical container is very similar to cylindrical containers. Consider the half section of the spherical container shown in in the figure below:



Consider now a simple spherical vessel of radius r and wall thickness tb, such as a round balloon. An internal pressure p_I induces equal biaxial tangential tensile stresses in the walls, which can be denoted using spherical $r\theta\phi$ coordinates as σ_θ and σ_ϕ.

The container will have an internal pressure p_i. For this case, the resultant force is the pressure times the area, which is a circular area. For a sphere inside diameter d or inside radius r, the total force on half the sphere is given by:

$$\text{pressure} \times \text{projected area} = p_i \frac{\pi d^2}{4} = p_i \pi r^2 \tag{16.2.268}$$

which is balanced by the force obtained by multiplying the wall stress times its area, i.e.:

$$\sigma_\phi(2\pi r t) \tag{16.2.269}$$

Equating these:

$$p_i\left(\pi r^2\right) = \sigma_\phi(2\pi r t) \tag{16.2.270}$$

Hence:

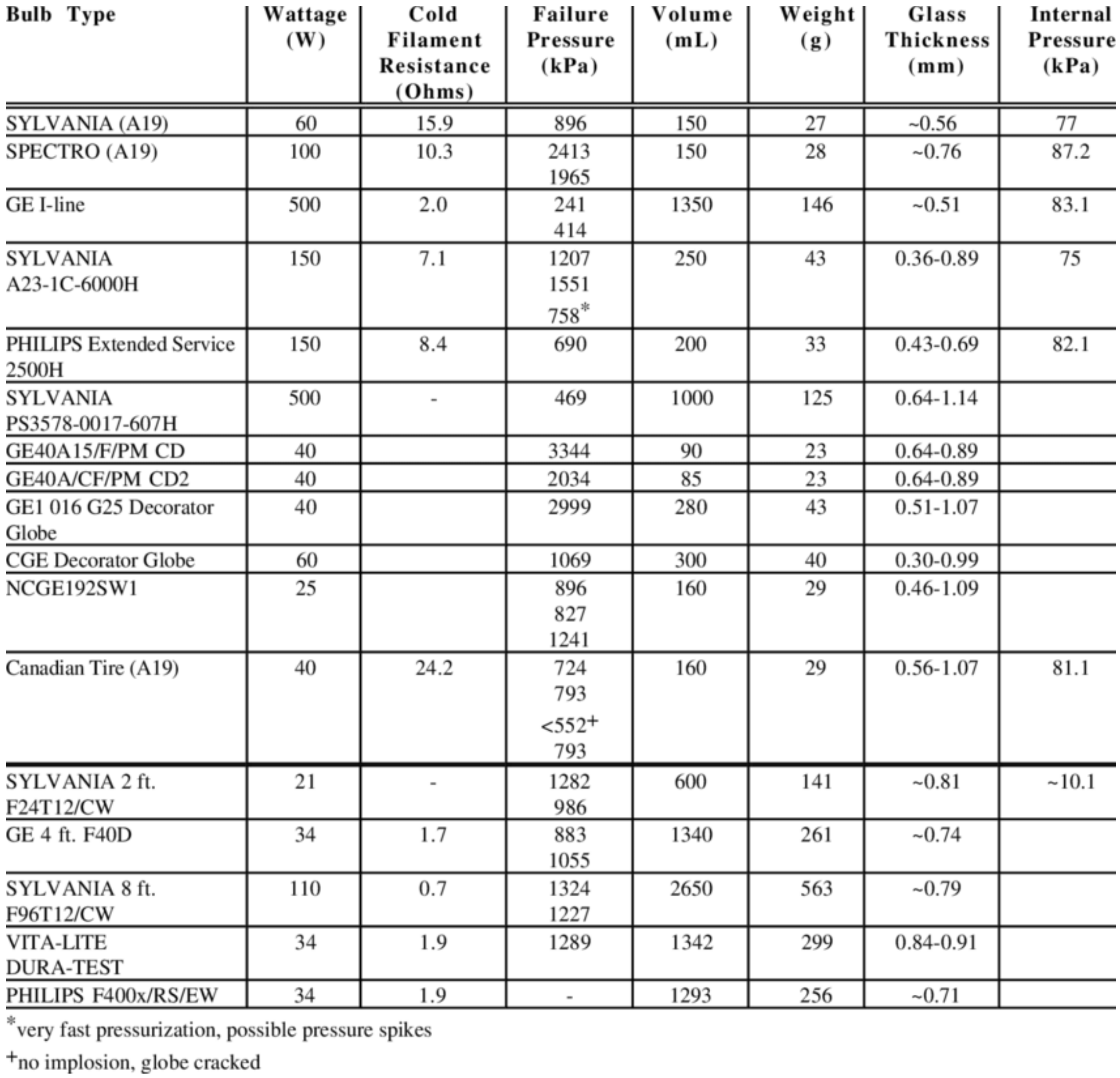
$$\sigma_\phi = \frac{p_i r}{2t} = \frac{p_i d}{4t} \tag{16.2.271}$$

Note that this is a statically determined result, with no dependence on the material properties. Further, the stresses in any two orthogonal circumferential directions are the same; i.e. $\sigma_\phi = \sigma_\theta$.

Therefore, a spherical pressure vessel can carry twice the internal pressure of a cylinder in comparison to the hoop stress.

LIQUIFIED HYDROGEN
FLAMMABLE GAS

It's a myth that all domestic light bulbs have their internal volume that is made of vacuum. The interior of an incandescent light bulb about depends a lot on the type of glass - or glass coated with polymers - the bulb is made of and varies since their creation between 1 [atm] made of inert gas like Argon and good vacuum. When the light is switched and the inert gas is heated up by the incandescent filament, it's easy to compute using the law of perfect gas that the internal pressure of the bulb will increase! For example we have the table below for light bulbs that were manufactured and the end of the 120th century[9] (holocene calendar):



Bulb Type	Wattage (W)	Cold Filament Resistance (Ohms)	Failure Pressure (kPa)	Volume (mL)	Weight (g)	Glass Thickness (mm)	Internal Pressure (kPa)
SYLVANIA (A19)	60	15.9	896	150	27	~0.56	77
SPECTRO (A19)	100	10.3	2413 1965	150	28	~0.76	87.2
GE I-line	500	2.0	241 414	1350	146	~0.51	83.1
SYLVANIA A23-1C-6000H	150	7.1	1207 1551 758*	250	43	0.36-0.89	75
PHILIPS Extended Service 2500H	150	8.4	690	200	33	0.43-0.69	82.1
SYLVANIA PS3578-0017-607H	500	-	469	1000	125	0.64-1.14	
GE40A15/F/PM CD	40		3344	90	23	0.64-0.89	
GE40A/CF/PM CD2	40		2034	85	23	0.64-0.89	
GE1 016 G25 Decorator Globe	40		2999	280	43	0.51-1.07	
CGE Decorator Globe	60		1069	300	40	0.30-0.99	
NCGE192SW1	25		896 827 1241	160	29	0.46-1.09	
Canadian Tire (A19)	40	24.2	724 793 <552$^+$ 793	160	29	0.56-1.07	81.1
SYLVANIA 2 ft. F24T12/CW	21	-	1282 986	600	141	~0.81	~10.1
GE 4 ft. F40D	34	1.7	883 1055	1340	261	~0.74	
SYLVANIA 8 ft. F96T12/CW	110	0.7	1324 1227	2650	563	~0.79	
VITA-LITE DURA-TEST	34	1.9	1289	1342	299	0.84-0.91	
PHILIPS F400x/RS/EW	34	1.9	-	1293	256	~0.71	

*very fast pressurization, possible pressure spikes

$^+$no implosion, globe cracked

Figure 16.92: **Physical data for various light bulbs (failure pressures quoted are pressures above 1 [atm])**

---

[9] Oldest light bulb internal volume were filled of 0.7 to 0.8 [atm] of inert gas.

16.2.3.7 Centrifugal expansion of a rotating ring

Centrifugal expansion of a rotating ring is a fascinating interplay between rotational motion and material elasticity and a very common type of structure and case in many objects of everyday life. When a ring spins about its axis, each infinitesimal segment experiences a centrifugal force directed outward. This force stretches the ring, increasing its radius slightly - an effect governed by the ring's density, angular velocity, and Young's modulus.

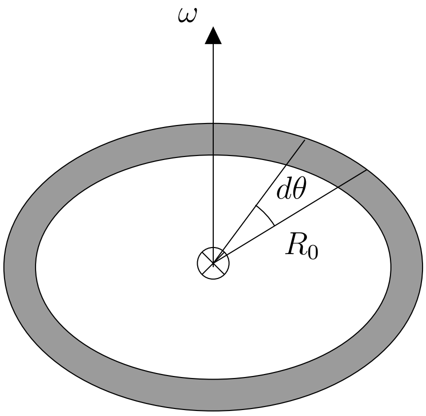


Figure 16.93: Simplified schema of a rotating ring

Let's take a look at an interesting physics problem and this problem is about a ring of density ρ made of some elastic material with Young modulus E that's what the gray ellipse shape above is a ring and we are rotating this ring about an axis perpendicular to its own plane and through the center of the ring right and so this arrow here in this diagram labelled ω that's the angular velocity vector of the ring it's just going around anti-clockwise about that axis now we also know that it has a radius of R_0.

The question is basically by how much does that radius expand when we start doing this rotation. The reason we would expect it to expand is because an object that is rotating moving in a circle experiences a fictitious centrifugal force which tries to push it outwards and so each individual bit of that ring is kind of being pushed outwards by this centrifugal force and so if it's an elastic material it can actually stretch and so we'd expect the whole thing to expand to some extent.

First let us remind that Young's modulus is defined by (see section Continuum Mechanics page 3741):

$$E=\frac{\sigma}{\varepsilon} \tag{16.2.272}$$

Now let us write the new radius when the ring is rotating as:

$$R=R_0+\Delta R=R_0\left(1+\frac{\Delta R}{R_0}\right) \tag{16.2.273}$$

Now let us write:

$$\mathrm{d}m=\rho\mathrm{d}V=\rho R\mathrm{d}\theta A \tag{16.2.274}$$

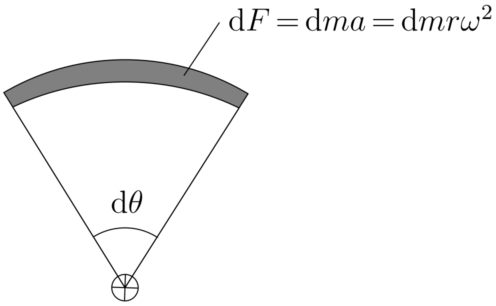
where A is the cross-sectional area of the ring.

Hence:

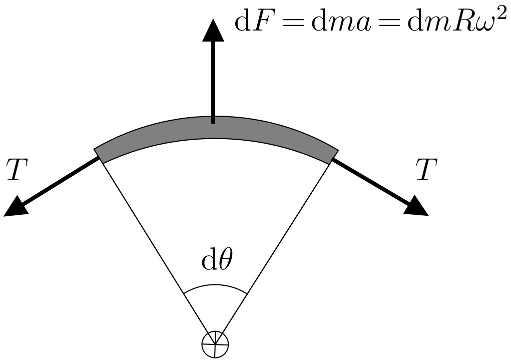
$$\frac{\mathrm{d}m}{\mathrm{d}\theta} = \rho R A \tag{16.2.275}$$

So the next thing we would like to do is start looking at the forces acting on a little element. So let's draw out a diagram that's going to help us see more clearly and more quantitatively what forces are acting on it:

When the ring rotates an element of mass is submitted to the centrifugal force:



So that the ring doesn't blow up there must be an equilibrium with other forces. These forces are the tension forces that we will denote by T and represented in the figure below:



And therefore we can write:

$$R\omega^2 \mathrm{d}m = 2T\sin\left(\frac{\mathrm{d}\theta}{2}\right) \cong 2T\frac{\mathrm{d}\theta}{2} \tag{16.2.276}$$

Hence:

$$T \cong R\omega^2 \frac{\mathrm{d}m}{\mathrm{d}\theta} = R^2\omega^2 A\rho \tag{16.2.277}$$

Remember that the Young's modulus was the stress divided by the strain and the stress is the force acting per unit cross-sectional area right and so the stress is in fact just you know the force that is acting um along the length of the ring is just this tension force and so the stress is just going to be the tension divided by the cross sectional area um and from this expression up here that's just going to be:

$$\sigma = \frac{T}{A} \cong \rho R^2 \omega^2 \tag{16.2.278}$$

What about the strain ? We can actually figure out the strain fairly easily as well because remember we said that was the fractional increase in the length of the ring in other words it's circumferential length and so um basically the the change in circumference is going to be:

$$\varepsilon=\frac{2\pi R-2\pi R_0}{2\pi R_0}=\frac{\Delta R}{R_0} \tag{16.2.279}$$

And as:

$$\sigma=E\varepsilon \tag{16.2.280}$$

We therefore have:

$$\rho R^2\omega^2\cong\frac{E\Delta R}{R_0} \tag{16.2.281}$$

Hence after rearranging:

$$\Delta R\cong\frac{\rho R^2R_0\omega^2}{E} \tag{16.2.282}$$

As we have:

$$R=R_0\left(1+\frac{\Delta R}{R_0}\right) \tag{16.2.283}$$

Therefore:

$$\Delta R\cong\frac{\rho R_0^3\omega^2}{E}\left(1+\frac{\Delta R}{R_0}\right)^2 \tag{16.2.284}$$

Let us take the case of small deformations:

$$\frac{\Delta R}{R_0}\ll 1 \tag{16.2.285}$$

So finally:

$$\boxed{\Delta R\cong\frac{\rho R_0^3\omega^2}{E}} \tag{16.2.286}$$

~20 %

★★☆☆☆

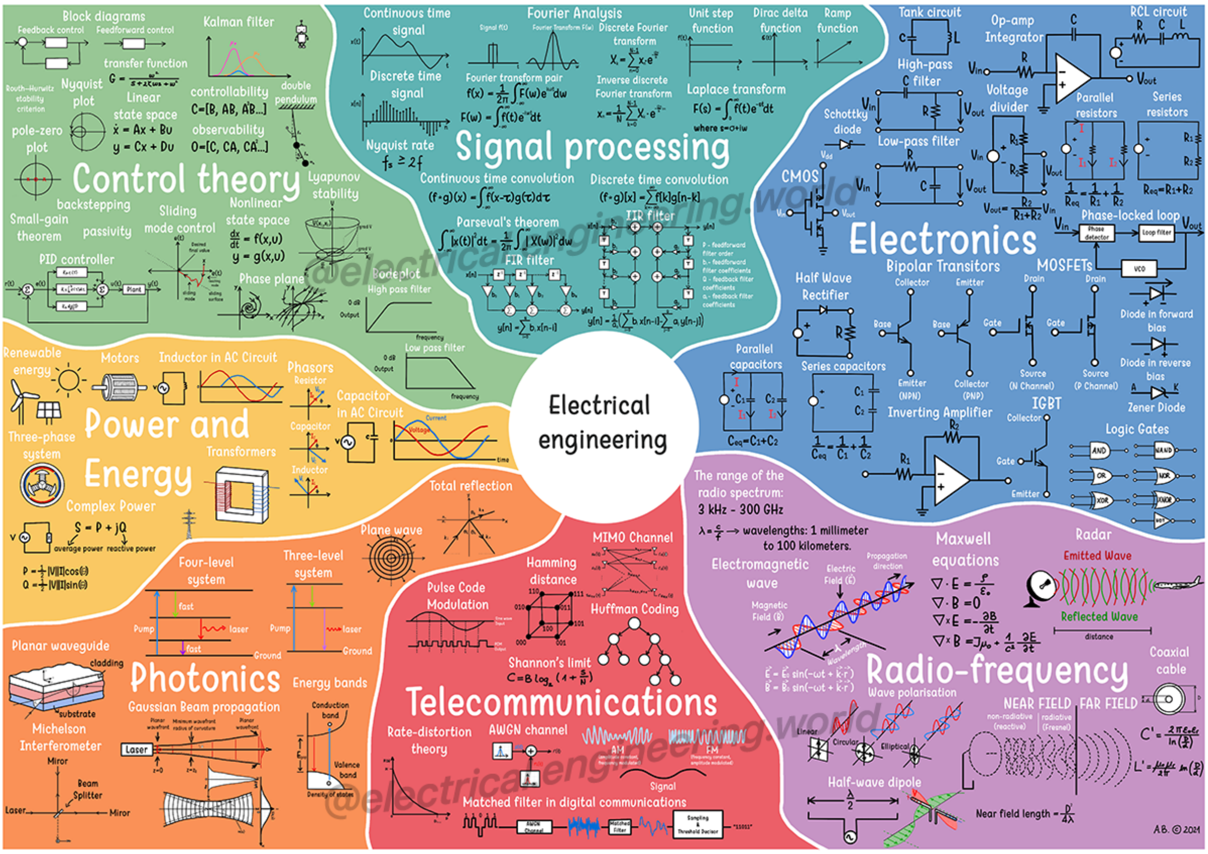
32 votes, 78.75%

16.3. Electrical Engineering

We will see in this section the study - under its mathematical aspect of sure - of general electronic circuits, chips, and electronic machinery, transmitter / receivers that the engineer must know formalize, analyse, understand, build and simulate after his studies. For this reason we have chosen in this textbook to focus - as always - on mathematical case of practical (useful!) applications in the every days life, mentioning the potential pitfalls and dangers of the electronic assembly when necessary!

We will deal here first with analogic electronics, then power electronics (electrotechnic), digital electronics and of the physics of semiconductor to understand the foundations of some components.

Electrical engineering is therefore a hierarchy of models. This is the only way to approach the design of complex systems. In principle, the operation of many devices can certainly always be reduced to the application of Maxwell's equations (see section Electrodynamics page 4095) or Schrödinger equation (see section Wave Quantum Physics page 4758) for more complicated case, however it is humanly impossible to understand the design of some systems by staying at this theoretical level.


Electrical engineering
Control theory
Block diagrams
Feedback control
Feedforward control
Kalman filter
transfer function
Routh-Hurwitz stability criterion
Nyquist plot
Linear state space
$\dot{x} = Ax + Bu$
$y = Cx + Du$
controllability
$C=[B, AB, A^2B...]$
observability
$O=[C, CA, CA^2...]$
double pendulum
pole-zero plot
Lyapunov stability
backstepping
Small-gain theorem
passivity
Sliding mode control
Nonlinear state space
$\frac{dx}{dt} = f(x,u)$
$y = g(x,u)$
PID controller
Plant
Phase plane
Bodeplot
High pass filter
Low pass filter
0 dB
Output
frequency
Signal processing
Continuous time signal
Discrete time signal
Nyquist rate
$f_s \geq 2f$
Fourier Analysis
Fourier transform pair
$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(w)e^{iwt}dw$
$F(w) = \int_{-\infty}^{\infty} f(t)e^{-iwt}dt$
Discrete Fourier transform
Inverse discrete Fourier transform
Unit step function
Dirac delta function
Ramp function
Laplace transform
$F(s) = \int_0^{\infty} f(t)e^{-st}dt$
where $s=\sigma+iw$
Continuous time convolution
$(f*g)(x) = \int_{-\infty}^{\infty} f(x-\tau)g(\tau)d\tau$
Discrete time convolution
$(f*g)[x] = \sum_{k=-\infty}^{\infty} f[k]g[n-k]$
Parseval's theorem
$\int_{-\infty}^{\infty}|x(t)|^2dt = \frac{1}{2\pi}\int_{-\infty}^{\infty}|X(w)|^2dw$
FIR filter
$y[n] = \sum b_i x[n-i]$
IIR filter
$y[n] = \frac{1}{a_0}\left(\sum b_i x[n-i] - \sum a_j y[n-j]\right)$
Electronics
Tank circuit
Op-amp Integrator
RCL circuit
High-pass C filter
Low-pass filter
Schottky diode
CMOS
Voltage divider
$V_{out} = \frac{R_2}{R_1+R_2}V_{in}$
Parallel resistors
$\frac{1}{R_{eq}} = \frac{1}{R_1} + \frac{1}{R_2}$
Series resistors
$R_{eq}=R_1+R_2$
Phase-locked loop
Phase detector
Loop filter
VCO
Half Wave Rectifier
Bipolar Transitors
Collector
Emitter
Base
Emitter (NPN)
Collector (PNP)
MOSFETs
Drain
Gate
Source (N Channel)
Source (P Channel)
Diode in forward bias
Diode in reverse bias
Zener Diode
Parallel capacitors
$C_{eq}=C_1+C_2$
Series capacitors
$\frac{1}{C_{eq}} = \frac{1}{C_1} + \frac{1}{C_2}$
Inverting Amplifier
IGBT
Logic Gates
AND
NAND
OR
NOR
XOR
XNOR
NOT
Power and Energy
Renewable energy
Motors
Inductor in AC Circuit
Phasors
Resistor
Capacitor
Inductor
Capacitor in AC Circuit
Current
Voltage
time
Three-phase system
Transformers
Complex Power
$S = P + jQ$
average power
reactive power
$P = \frac{1}{2}|V||I|\cos(\theta)$
$Q = \frac{1}{2}|V||I|\sin(\theta)$
Photonics
Total reflection
Plane wave
Four-level system
Three-level system
fast
Pump
laser
Ground
Planar waveguide
cladding
substrate
Michelson Interferometer
Miror
Beam Splitter
Laser
Gaussian Beam propagation
Planar wavefront
Minimum wavefront radius of curvature
Energy bands
Conduction band
Valence band
Density of states
Telecommunications
MIMO Channel
Hamming distance
Pulse Code Modulation
Huffman Coding
Shannon's limit
$C = B\log_2\left(1+\frac{S}{N}\right)$
Rate-distortion theory
AWGN channel
AM
FM
Signal
Matched filter in digital communications
AWGN Channel
Matched Filter
Sampling & Threshold Decisor
"11011"
Radio-frequency
The range of the radio spectrum: 3 kHz - 300 GHz
$\lambda = \frac{c}{f} \rightarrow$ wavelengths: 1 millimeter to 100 kilometers.
Electromagnetic wave
Electric Field ($\vec{E}$)
Magnetic Field ($\vec{B}$)
Propagation direction
Wavelength
$\vec{E} = \vec{E}_0 \sin(-\omega t + \vec{k}\cdot\vec{r})$
$\vec{B} = \vec{B}_0 \sin(-\omega t + \vec{k}\cdot\vec{r})$
Maxwell equations
$\nabla \cdot E = \frac{\rho}{\varepsilon_0}$
$\nabla \cdot B = 0$
$\nabla \times E = -\frac{\partial B}{\partial t}$
$\nabla \times B = J\mu_0 + \frac{1}{c^2}\frac{\partial E}{\partial t}$
Radar
Emitted Wave
Reflected Wave
distance
Coaxial cable
$C' = \frac{2\pi\varepsilon_0\varepsilon_r}{\ln\left(\frac{D}{d}\right)}$
$L' = \frac{\mu_0\mu_r}{2\pi}\ln\left(\frac{D}{d}\right)$
Wave polarisation
Linear
Circular
Elliptical
NEAR FIELD
FAR FIELD
non-radiative (reactive)
radiative (Fresnel)
Half-wave dipole
$\frac{\lambda}{2}$
Near field length $= \frac{D^2}{4\lambda}$
@electricalengineering.world
A.B. © 2021

It is therefore customary in the industry to perform the analysis in six stages:

L0. **Solid State Physics:** This model is essential for the analysis of electric and magnetic properties of matter. It is based on the laws of quantum physics and essentially leads to the description of energy bands and to the calculation of their degree of occupation. This model explains for example the fundamental properties of semiconductors.

L1. **Electromagnetism:** This model is essential for the analysis of devices working at microwave frequencies and the electromagnetic devices. It is based on Maxwell's relations and uses the mathematical theory of partial differential equations. This model don't allows anymore to analyse the influence of an atom as the studied objects are at a macroscopic level, described by their dimensions, their permittivity, their conductivity, etc.

L2. **Circuit Theory:** This model is essential for the analysis of electronic devices in the very common case where the dimensions of the device are well below the wavelength of the phenomenon studied. This model is based on Kirchhoff's lemma and the definition of a half-dozen of discrete elements, resistance, capacitance, inductance, etc. There is one more geometries in such a model but only a topologic (structural) approach. One can calculate the current and voltage, scalar quantities, and the fields have most of time meaning anymore at this scale. Mathematical techniques are those of ordinary differential equations, Laplace transformations, complex computation and matrices, etc.

L3. **Block Diagrams:** At this level, we do not take into account currents or voltages, and we don't care about the geometry (topology) of the system. This is constituted by the block connections doing functions/works characterized by the relations between output and input variables.

L4. **Systems:** At this level, we schematize as a functional block a set of blocks of level 3. A computer is such an interconnection of various logical systems each performing a particular function.

L5. **Software:** From this level, the engineer don't add any more additional devices, don't combines them anymore into larger systems, but program the machine. The theoretical methods then are usually closer to the linguistic than to mathematics.

The reader must not forget that we have proved in the section of Electrokinetics (page 4231) that even in a strong industrial electric field, the drift velocity is negligible compared to the thermal velocity (only a few millimetres per seconds)!

As the thermal velocity depends only a little bit of the electric field, it turns out that in practice the electron velocity is independent of the electric field. In other words, establishing a current, even intense, has only a negligible effect on the speed of electrons!

It is therefore not the electrons that carry the energy, neither mobile electrons push each other trough a circuit, contrary to a falsely held belief, due to bad analogies (which are always bad equivalences as we already know...!) - typically the analogy of electrons that behaves like water flow in cables - as taught in the non-graduate degrees of the school education system!!!!

Before we start, here are some typical voltage scales:

Magnitude $[\text{V}]$	**Description**
$500 \cdot 10^{-9}$	Change in nerve cell potential caused by opening a single acetylcholine receptor channel
$1.2 \cdot 10^{-3}$	Potential created at ambient temperatures from K Type Thermocouple
$75 \cdot 10^{-3}$	Nerve cell resting potential
$320 \cdot 10^{-3}$	Typical voltage reference level in consumer audio electronics
$500 \cdot 10^{-3}$	Typical MOSFET threshold voltage for modern processes
$700 \cdot 10^{-3}$	Forward voltage drop of normal silicon diodes
$900 \cdot 10^{-3}$	Typical positive supply voltage of a low voltage CMOS digital integrated circuit or a Lemon battery cell (made with copper and zinc electrodes
0-3	Magnitudes of standard reduction potentials in chemistry
1.5	Alkaline battery AAAA, AAA, AA, C or D battery
3.3	One of the most common low voltage CMOS digital circuit supply voltages
5	USB power, used for example to charge a cell phone or a digital camera.[18] Also one of the most common digital circuit supply voltages for both TTL and CMOS technologies
12	Typical car battery
100-240	Domestic wall socket voltage
600	Electric eel sends this voltage in an average attack
630	London Underground railway tracks
2450	Electric chair execution in Nebraska
$3\text{-}10 \cdot 10^{3}$	Electric fence
$3 \cdot 10^{3}$	Voltage required to generate a 1 $[\text{mm}]$ electric arc in standard air
$3\text{-}35 \cdot 10^{3}$	Accelerating voltage for a typical television cathode ray tube
$4.16\text{-}34.5 \cdot 10^{3}$	Typical voltages in North America for distribution of power from distribution substations to end users
$15 \cdot 10^{3}$	Train 15 $[\text{kV}]$ AC railway electrification overhead lines, $16^2/_3$ $[\text{Hz}]$
$25 \cdot 10^{3}$	European high-speed train overhead power lines
$69\text{-}230 \cdot 10^{3}$	Range used in North American power high-voltage transmission substations
$345\text{-}800 \cdot 10^{3}$	Range used in EHV (Electrical High Voltage) power transmission system
$3 \cdot 10^{6}$	Used by the ultra-high voltage electron microscope at Osaka University
$25.5 \cdot 10^{6}$	The largest man-made DC voltage – produced in a Van de Graaff generator at Oak Ridge National Laboratory
$100 \cdot 10^{6}$	The potential difference between the ends of a typical lightning bolt
$7 \cdot 10^{15}$	Theoretical voltage around a particular energetic highly magnetized rotating neutron star

$1.05 \cdot 10^{27}$	Planck voltage

Table 16.7: Some common voltage scales

and here are some typical current values:

Magnitude $[\mathrm{A}]$	**Description**
$1.602 \cdot 10^{-19}$	One electron per second
$10 \cdot 10^{-6}$	Minimum current necessary to cause death (by ventricular fibrillation when applied directly across the human heart)
$700 \cdot 10^{-6}$	Typical portable hearing aid
$1\text{-}5 \cdot 10^{-3}$	Cathode ray tube electron gun beam current
$10 \cdot 10^{-3}$	Through the hand to foot may cause a person to freeze and be unable to let go
$20 \cdot 10^{-3}$	Common light-emitting diode (constant current); also deadly limit for skin contact (at 120–230 $[\mathrm{V}]$)
$250\text{-}450 \cdot 10^{-3}$	230 $[\mathrm{V}]$ AC Tungsten incandescent light bulb (60–100 $[\mathrm{W}]$)
1	Typical iPhone charger (5 $[\mathrm{W}]$)
9	230 $[\mathrm{V}]$ AC, toaster, kettle (2 $[\mathrm{kW}]$)
10 or 20	230 $[\mathrm{V}]$ AC, Europe common domestic circuit breaker rating
10 or 20	Typical 12 $[\mathrm{V}]$ motor vehicle starter motor (typically 1-2 $[\mathrm{kW}]$)
80 or 375	TIG welding arc
166	400 $[\mathrm{V}]$ low voltage secondary side distribution transformer with primary 12 $[\mathrm{kV}]$; 200 $[\mathrm{kVA}]$ (up to 1000 $[\mathrm{kVA}]$ also common)
$2 \cdot 10^{3}$	10.5 $[\mathrm{kV}]$ secondary side from an electrical substation with primary 115 $[\mathrm{kV}]$; 63 $[\mathrm{MVA}]$
$30 \cdot 10^{3}$	Typical lightning strike
$100 \cdot 10^{3}$	Low range of Birkeland current that creates Earth's aurorae
$180 \cdot 10^{3}$	Typical current used in electric arc furnace for ferroalloys
$5 \cdot 10^{6}$	Flux tube between Jupiter and Io (moon)
$26 \cdot 10^{6}$	Sandia National Laboratories, Z machine approximate firing current

Table 16.8: Some common current scales

16.3.1 Elementary primitive electrical symbols

This section will help you become familiar with the language and some methods of calculation used by engineers of the industry for the first 4 levels. However and strictly speaking, any training in this area should be completed by practical laboratory work!

Sadly still in 12016 (holocene calendar) there are several national and international standards for graphical symbols in circuit diagrams, in particular:

- IEC 60617 (also known as British Standard BS 3939).
- ANSI standard Y32.2 (also known as IEEE Std 315).
- IEEE Std 91/91a: graphic symbols for logic functions (used in digital electronics). It is referenced in ANSI Y32.2/IEEE Std 315. Australian Standard AS 1102.

We have tried to respect at best in diagrams the norm NF EN 60617 for electrical components symbols. As they are very numerous, we propose below a list only of the components used or mentioned so far in this textbook. This table will therefore evolve over time!

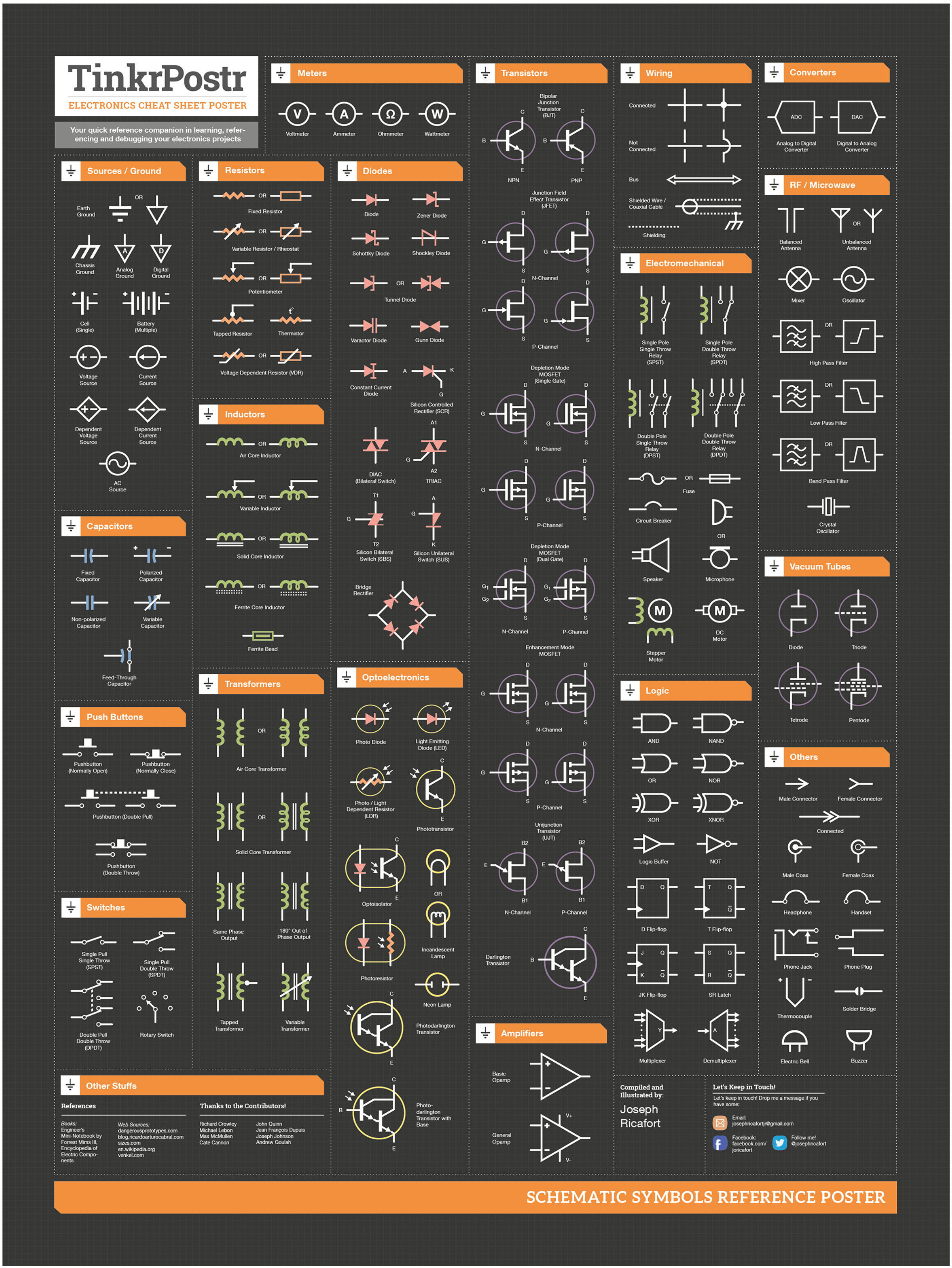
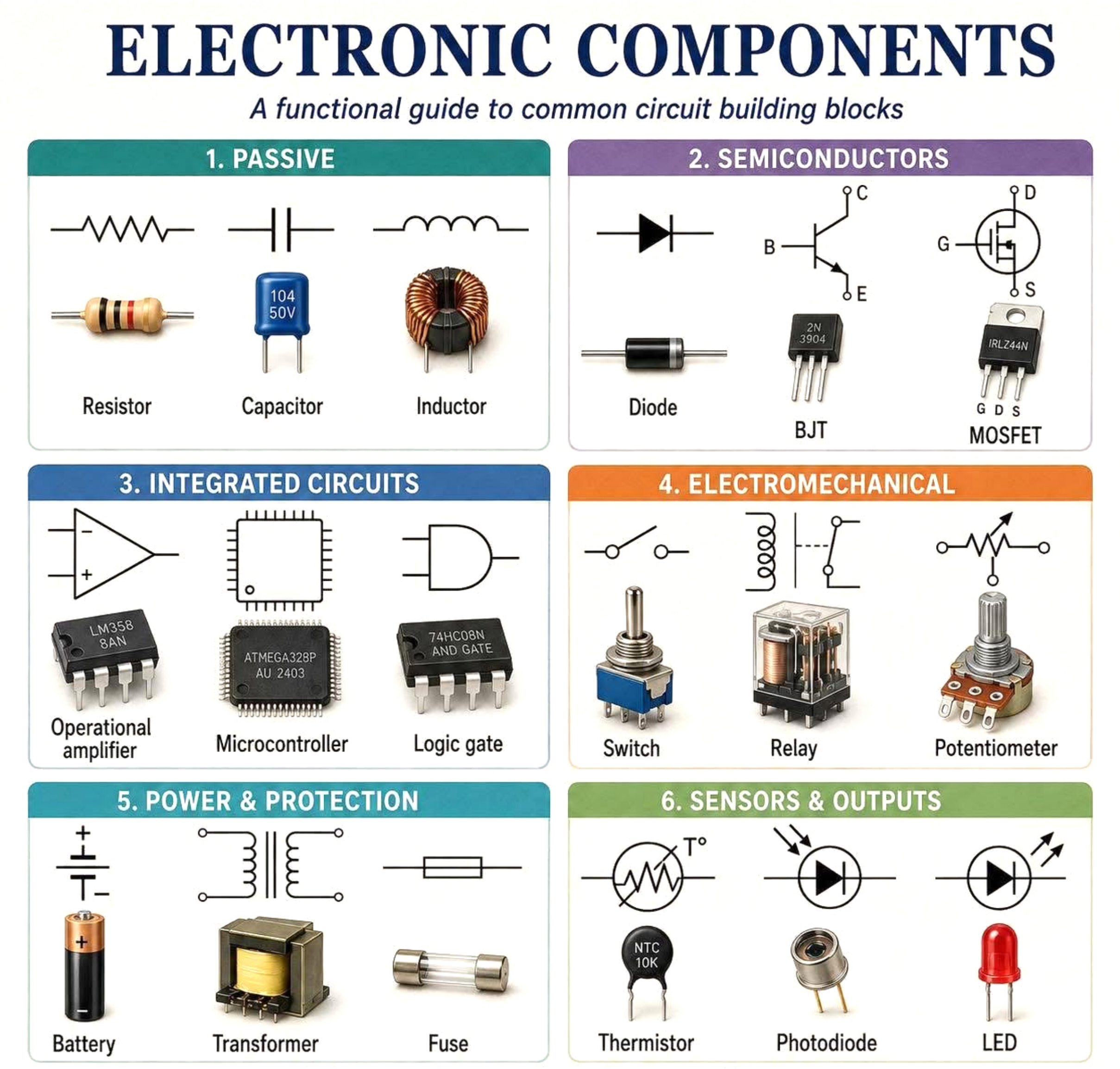

TinkrPostr
ELECTRONICS CHEAT SHEET POSTER
Your quick reference companion in learning, referencing and debugging your electronics projects
Meters
Voltmeter
Ammeter
Ohmmeter
Wattmeter
Transistors
Bipolar Junction Transistor (BJT)
NPN
PNP
Junction Field Effect Transistor (JFET)
N-Channel
P-Channel
Depletion Mode MOSFET (Single Gate)
Depletion Mode MOSFET (Dual Gate)
Enhancement Mode MOSFET
Unijunction Transistor (UJT)
Darlington Transistor
Wiring
Connected
Not Connected
Bus
Shielded Wire / Coaxial Cable
Shielding
Converters
ADC
DAC
Analog to Digital Converter
Digital to Analog Converter
RF / Microwave
Balanced Antenna
Unbalanced Antenna
Mixer
Oscillator
High Pass Filter
Low Pass Filter
Band Pass Filter
Crystal Oscillator
Sources / Ground
Earth Ground
Chassis Ground
Analog Ground
Digital Ground
Cell (Single)
Battery (Multiple)
Voltage Source
Current Source
Dependent Voltage Source
Dependent Current Source
AC Source
Resistors
Fixed Resistor
Variable Resistor / Rheostat
Potentiometer
Tapped Resistor
Thermistor
Voltage Dependent Resistor (VDR)
Diodes
Diode
Zener Diode
Schottky Diode
Shockley Diode
Tunnel Diode
Varactor Diode
Gunn Diode
Constant Current Diode
Silicon Controlled Rectifier (SCR)
DIAC (Bilateral Switch)
TRIAC
Silicon Bilateral Switch (SBS)
Silicon Unilateral Switch (SUS)
Bridge Rectifier
Electromechanical
Single Pole Single Throw Relay (SPST)
Single Pole Double Throw Relay (SPDT)
Double Pole Single Throw Relay (DPST)
Double Pole Double Throw Relay (DPDT)
Fuse
Circuit Breaker
Speaker
Microphone
Stepper Motor
DC Motor
Inductors
Air Core Inductor
Variable Inductor
Solid Core Inductor
Ferrite Core Inductor
Ferrite Bead
Capacitors
Fixed Capacitor
Polarized Capacitor
Non-polarized Capacitor
Variable Capacitor
Feed-Through Capacitor
Push Buttons
Pushbutton (Normally Open)
Pushbutton (Normally Close)
Pushbutton (Double Pull)
Pushbutton (Double Throw)
Switches
Single Pull Single Throw (SPST)
Single Pull Double Throw (SPDT)
Double Pull Double Throw (DPDT)
Rotary Switch
Transformers
Air Core Transformer
Solid Core Transformer
Same Phase Output
180° Out of Phase Output
Tapped Transformer
Variable Transformer
Optoelectronics
Photo Diode
Light Emitting Diode (LED)
Photo / Light Dependent Resistor (LDR)
Phototransistor
Optoisolator
Incandescent Lamp
Photoresistor
Neon Lamp
Photodarlington Transistor
Photo-darlington Transistor with Base
Vacuum Tubes
Diode
Triode
Tetrode
Pentode
Logic
AND
NAND
OR
NOR
XOR
XNOR
Logic Buffer
NOT
D Flip-flop
T Flip-flop
JK Flip-flop
SR Latch
Multiplexer
Demultiplexer
Others
Male Connector
Female Connector
Connected
Male Coax
Female Coax
Headphone
Handset
Phone Jack
Phone Plug
Thermocouple
Solder Bridge
Electric Bell
Buzzer
Amplifiers
Basic Opamp
General Opamp
V+
V-
Other Stuffs
References
Books:
Engineer's Mini-Notebook by Forrest Mims III, Encyclopedia of Electric Components
Web Sources:
dangerousprototypes.com
blog.ricardoarturocabral.com
sizes.com
en.wikipedia.org
venkel.com
Thanks to the Contributors!
Richard Crowley
Michael Lebon
Max McMullen
Cate Cannon
John Quinn
Jean François Dupuis
Joseph Johnson
Andrew Goulah
Compiled and Illustrated by:
Joseph Ricafort
Let's Keep in Touch!
Let's keep in touch! Drop me a message if you have some:
Email: josephricafortjr@gmail.com
Facebook: facebook.com/joricafort
Follow me! @josephricafort
SCHEMATIC SYMBOLS REFERENCE POSTER

Figure 16.94: **Electronic cheat sheet poster (author: Joseph Ricafort Jr.)**

And here is a nice infography with how some of the most important components looks like:


ELECTRONIC COMPONENTS
A functional guide to common circuit building blocks
1. PASSIVE
104
50V
Resistor
Capacitor
Inductor
2. SEMICONDUCTORS
C
B
E
D
G
S
2N
3904
IRLZ44N
G D S
Diode
BJT
MOSFET
3. INTEGRATED CIRCUITS
−
+
LM358
8AN
ATMEGA328P
AU 2403
74HC08N
AND GATE
Operational amplifier
Microcontroller
Logic gate
4. ELECTROMECHANICAL
Switch
Relay
Potentiometer
5. POWER & PROTECTION
+
−
Battery
Transformer
Fuse
6. SENSORS & OUTPUTS
T°
NTC
10K
Thermistor
Photodiode
LED

16.3.2 Alternative current VS Direct current

The reader will notice that throughout this section, we will mainly work with alternating current. It seems important to explain the origin of this trend of the contemporary industrial world for the alternating current before going further.

In fact, the origin of this trend is relatively simple:

When power plants came into being, especially in remote areas of urban centers, it was necessary to transport electric power over long distances. But the cables that carry the electricity have some resistance and this posed a major problem.

Indeed, an average city can largely need a power of about 10 [MW]. If this quantity had to be transported under a modest voltage of about 100 [V], as we have $P=UI$ (see section Electrokinetics page 4253), the current had to be enormous: $100,000$ [A]!

But the Joule effect in the copper of 1 [cm] diameter has a linear resistance of $R\cong 0.4\,[\Omega\cdot\text{km}^{-1}]$. With a current of $100,000$ [A], the energy loss per Joule effect would be about (neglecting the potential drop):

$$P=RI^2=4000\,[\text{MW}\cdot\text{km}] \tag{16.3.1}$$

we see quite quick the problem if we not have any supra-conductors...

At the price of $0.1\,[\$\cdot\text{kW}\cdot\text{h}]$, this represents a cost (loss) of about:

$$C=0.4\cdot 10^6\,[\$\cdot\text{kW}\cdot\text{h}] \tag{16.3.2}$$

humm ...!

There were no other economic choices than to lower the current. Clearly, if the voltage reached 10^5 [V], the same power could be carried efficiently by 100 [A]. Thus, by raising the voltage by a factor of $1,000$, we can reduce the current by a factor of $1,000$ as well, and thus the Joule loss by a factor 10^6.

As there was already a simple device for raising and lowering the AC voltage (transformers) without any comparable device for DC voltage (at least at the time), the race was won by the adepts of alternating current.

It should also be added as a second interest that some linear electrical components (see below) do not have much interest in DC ... we will come back to it!

Let's see a simple assembly to generate single-phase alternating current:

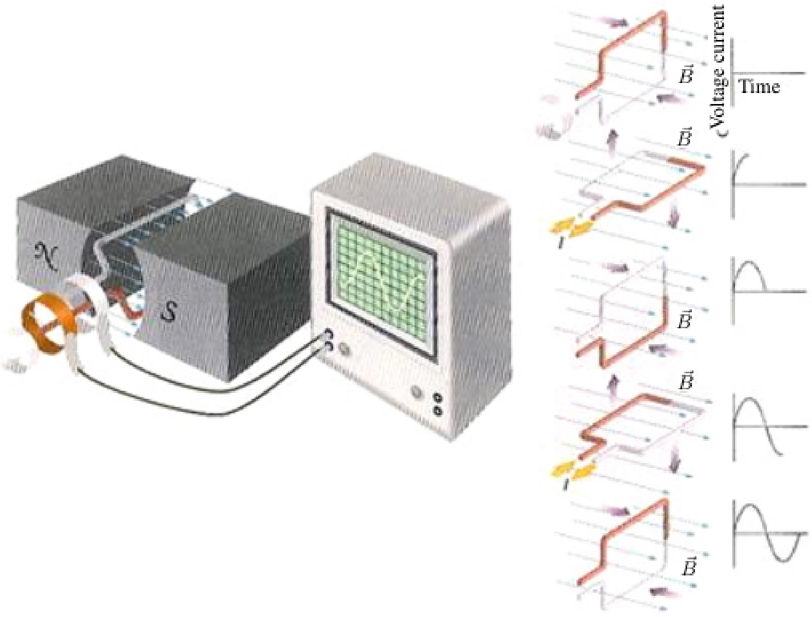


Figure 16.95: **Schematic diagram of an alternating current generator**

The voltage (respectively the current) is determined by the Faraday law proved in the section Electrokinetics (see page 4259):

$$e=-\frac{\mathrm{d}\Phi}{\mathrm{d}t} \tag{16.3.3}$$

Which therefore gives the induced electromotive force (or voltage in the case of a generator without resistance...).

We obviously have in the situation above if the magnet is permanent and the length of the square turn is L:

$$e=-\frac{\mathrm{d}\Phi}{\mathrm{d}t}=-\frac{\vec{B}\circ\mathrm{d}S}{\mathrm{d}t}=-BL^2\frac{\mathrm{d}\cos(\theta)}{\mathrm{d}t}=-BS\frac{\mathrm{d}\cos(\omega t)}{\mathrm{d}t}=BS\omega\sin(\omega t) \tag{16.3.4}$$

We already see that in order to obtain a given electromotive force it will be preferable to play with the frequency of rotation rather than with the surface or the intensity of the magnetic field... Or to increase the number of turns by a mounting allowing to arrive at the following relation (we noted the different way you can find this result in the literature):

$$\boxed{e=NBS\omega\sin(\omega t)=NB(RL)\omega\sin(\omega t)=N\Phi\omega\sin(\omega t)=N\Phi\omega_0\sin(\omega_0 t)=N\Phi\omega(t)} \tag{16.3.5}$$

where R is the radius of the armature and L the length of the conductor in the slot (in most undergraduate classes $R=L$). Depending on the number of poles and the total number of conductors, engineers denote the latter relation sometimes in the following way:

$$e=k\phi\omega(t) \tag{16.3.6}$$

where k is named the "machine constant" (...).

It should be pointed out for the sceptics that there is conservation of energy in this system! Indeed, the energy needed to turn the coil will be that used in part by the system (and that is why dams run turbines with water and nuclear power plants with steam and wind turbines with wind...).

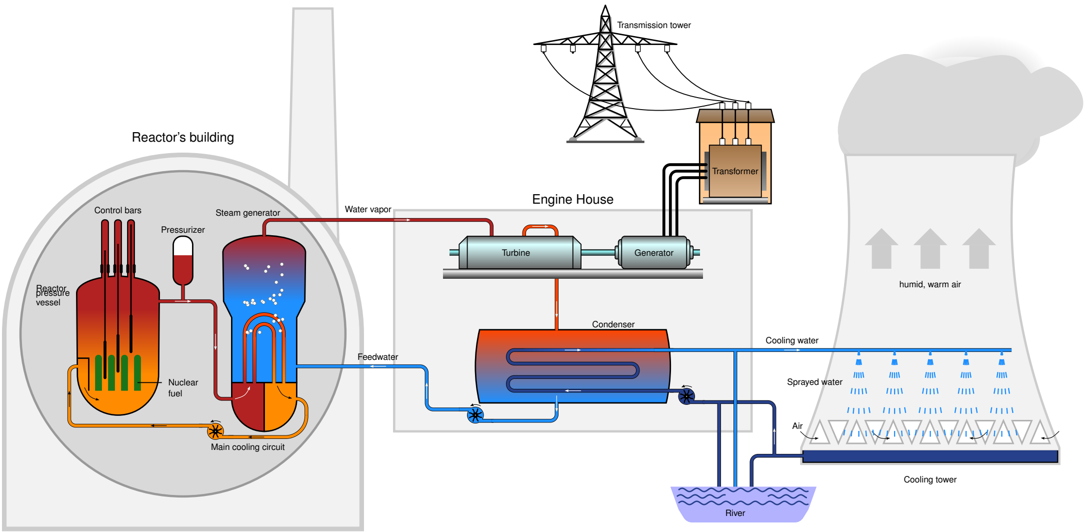

Transmission tower
Reactor's building
Transformer
Engine House
Control bars
Steam generator
Water vapor
Pressurizer
Turbine
Generator
Reactor pressure vessel
humid, warm air
Condenser
Cooling water
Feedwater
Nuclear fuel
Sprayed water
Air
Main cooling circuit
Cooling tower
River

Figure 16.96: **Example of nuclear plant (author: Thomas Soell)**

Obviously the opposite case as an alternating current injected into the coil will cause it to rotate in the magnetic field. So in a situation, we have an electric generator and in the opposite case an electric motor.

It is possible with similar equipment to produce a nearly reliable DC voltage in the following manner named "dynamo" (see the more detailed study of that apparatus further below at page 8447):

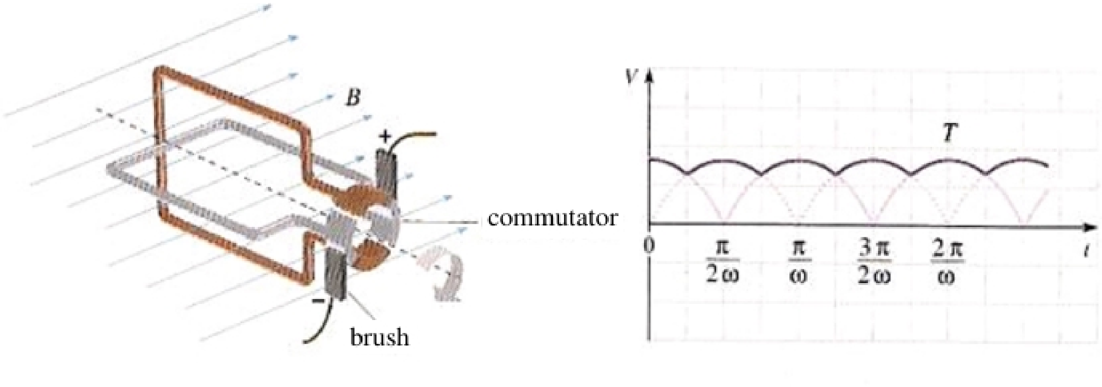

B
+
commutator
−
brush
V
T
0
$\frac{\pi}{2\omega}$
$\frac{\pi}{\omega}$
$\frac{3\pi}{2\omega}$
$\frac{2\pi}{\omega}$
t

Figure 16.97: **Example of direct current generator (dynamo)**

The simple generator given first with some horseshoe magnets producing the magnetic field was widely used at the beginning of the era of electrical technology but at high voltages (several [kV]) and high internal currents too (more that 50 [A]). Metal brushes and slip rings produced sparks and deteriorated relatively quickly. Currently, this rotating armature machine is no longer widely used.

To avoid the difficulties associated with voltages exceeding about 600 [V], we now turn electromagnets around a stationary armature. The current supplying the rotating electromagnets (which may also be permanent magnets) is relatively low and does not pose any problem to the operation of the rings and brushes. This configuration is then referred to as an "alternator".

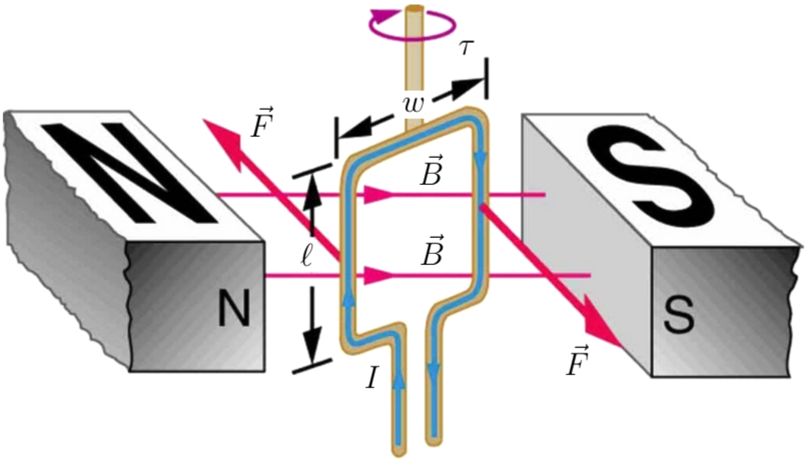
With linear electrical components it is also possible to straighten the tension (we will see this much further). Then we fall back on a dynamo again!

Figure 16.98: **An anti-electricity cartoon from 11900s (holocene calendar)**

16.3.2.1 DC generator (dynamo)

Motors are the most common application of magnetic force on current-carrying wires. Motors have loops of wire in a magnetic field. When current is passed through the loops, the magnetic field exerts torque on the loops, which rotates a shaft. Electrical energy is converted to mechanical work in the process.

As depicted in the below figure where a current-carrying loop of wire attached to a vertically rotating shaft feels magnetic forces that produce a clockwise torque as viewed from above:



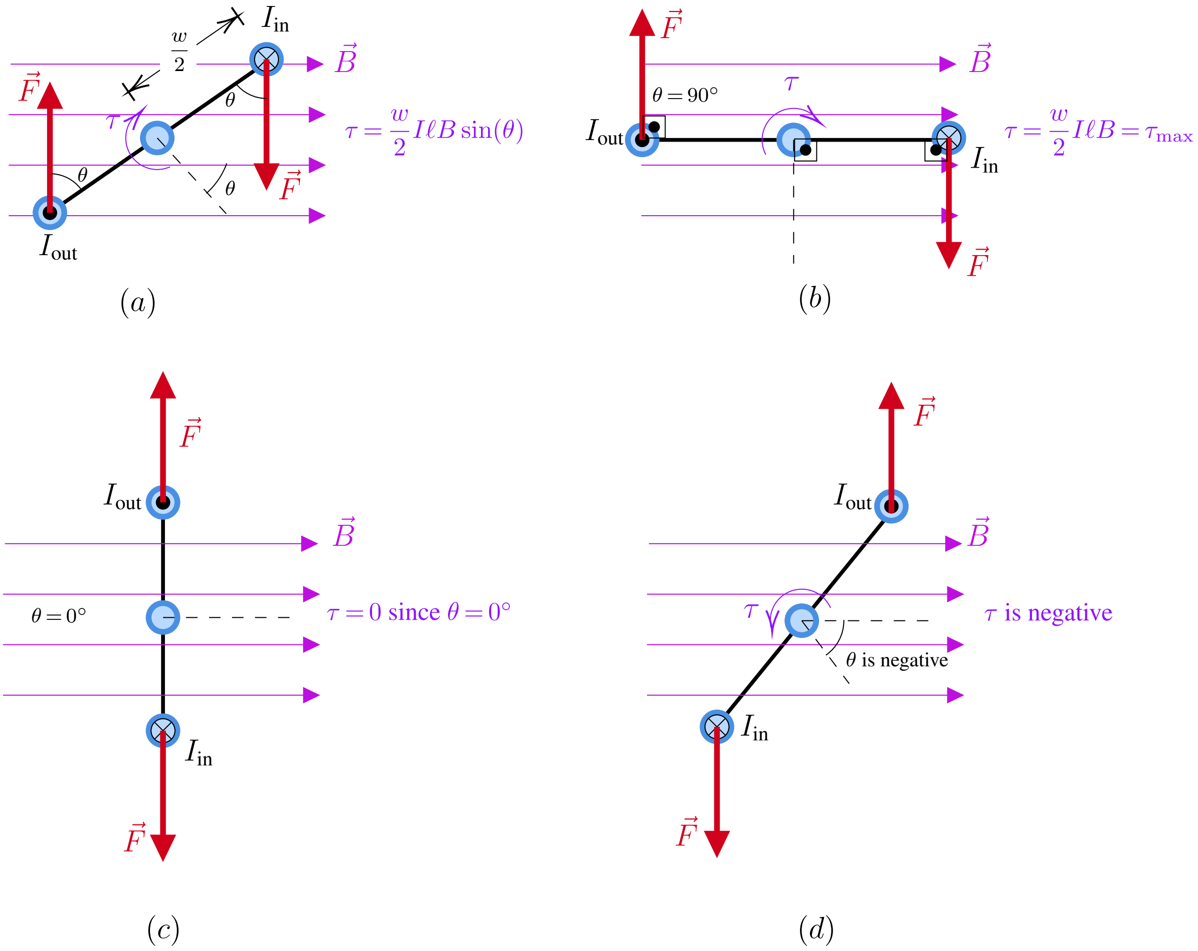
Let us examine the force on each segment of the loop in the figure above to find the torques produced about the axis of the vertical shaft (this will lead to a useful equation for the torque on the loop.).

We take the magnetic field to be uniform over the rectangular loop, which has width w and height ℓ. First, we note that the forces on the top and bottom segments are vertical and, therefore, parallel to the shaft, producing no torque according the Laplace's force law (see section Magnetostics page 4063):

$$\vec{F}=I\vec{l}\times\vec{B} \tag{16.3.7}$$

Those vertical forces are equal in magnitude and opposite in direction, so that they also produce no net force on the loop.

The figure below shows views of the loop from above:



I_{in}
$\frac{w}{2}$
$\vec{B}$
$\vec{F}$
τ
θ
$\tau = \frac{w}{2} I\ell B \sin(\theta)$
I_{out}
(a)
$\vec{F}$
$\theta = 90°$
τ
$\vec{B}$
I_{out}
$\tau = \frac{w}{2} I\ell B = \tau_{\max}$
I_{in}
$\vec{F}$
(b)
$\vec{F}$
I_{out}
$\vec{B}$
$\theta = 0°$
$\tau = 0$ since $\theta = 0°$
I_{in}
$\vec{F}$
(c)
$\vec{F}$
I_{out}
$\vec{B}$
τ
τ is negative
θ is negative
I_{in}
$\vec{F}$
(d)

Figure 16.99: **(a) The equation for torque is derived using this view. Note that the perpendicular to the loop makes an angle θ with the field that is the same as the angle between $w/2$ and $\vec{F}$. (b) The maximum torque occurs when θ is a right angle and $\sin(\theta) = 1$. (c) Zero (minimum) torque occurs when θ is zero and $\sin(\theta) = 0$. (d) The torque reverses once the loop rotates past $\theta = 0$.**

Torque is we know defined as:

$$\tau = \vec{r} \times \vec{F} = rF\sin(\theta) \tag{16.3.8}$$

where F is the force, r is the distance from the pivot that the force is applied, and θ is the angle between $\vec{r}$ and $\vec{F}$. As seen in the top left illustration above! Right rand rule gives the forces on the sides to be equal in magnitude and opposite in direction, so that the net force is again zero. However, each force produces a clockwise torque. Since $r = w/2$, the torque on each vertical segment is:

$$\tau = \frac{w}{2} F\sin(\theta) \tag{16.3.9}$$

and the two add to give a total torque.

Now, each vertical segment has a length ℓ that is perpendicular to B, so that the force on each is $F = I\ell B$. Entering F into the expression for torque yields:

$$\tau = rF\sin(\theta) = wI\ell B\sin(\theta) \tag{16.3.10}$$

If we have a multiple loop of N turns, we get N times the torque of one loop. Finally, note that the surface of the loop is $S = w\ell$; the expression for the torque becomes:

$$\boxed{\tau = NISB\sin(\theta)} \tag{16.3.11}$$

This is the "torque on a current-carrying loop" in a uniform magnetic field. This equation can be shown to be valid for a loop of any shape. The loop carries a current I, has N turns, each of surface S, and the perpendicular to the loop makes an angle θ with the field $\vec{B}$. The net force on the loop is zero.

✍**Example:**

Let us find the maximum torque on a 100-turn square loop of a wire of 10 [cm] on a side that carries 15 [A] (small Neodym magnets have a typical field of 1.4 [T]) of current in a 2 [T] field.

We know that the torque on the loop can be found using $\tau = NISB\sin(\theta)$. Obviously the maximum torque occurs when $\theta = 90°$ and $\sin(\theta) = 1$.

For $\sin(\theta) = 1$ the maximum torque is:

$$\tau_{\text{max}} = NISB \tag{16.3.12}$$

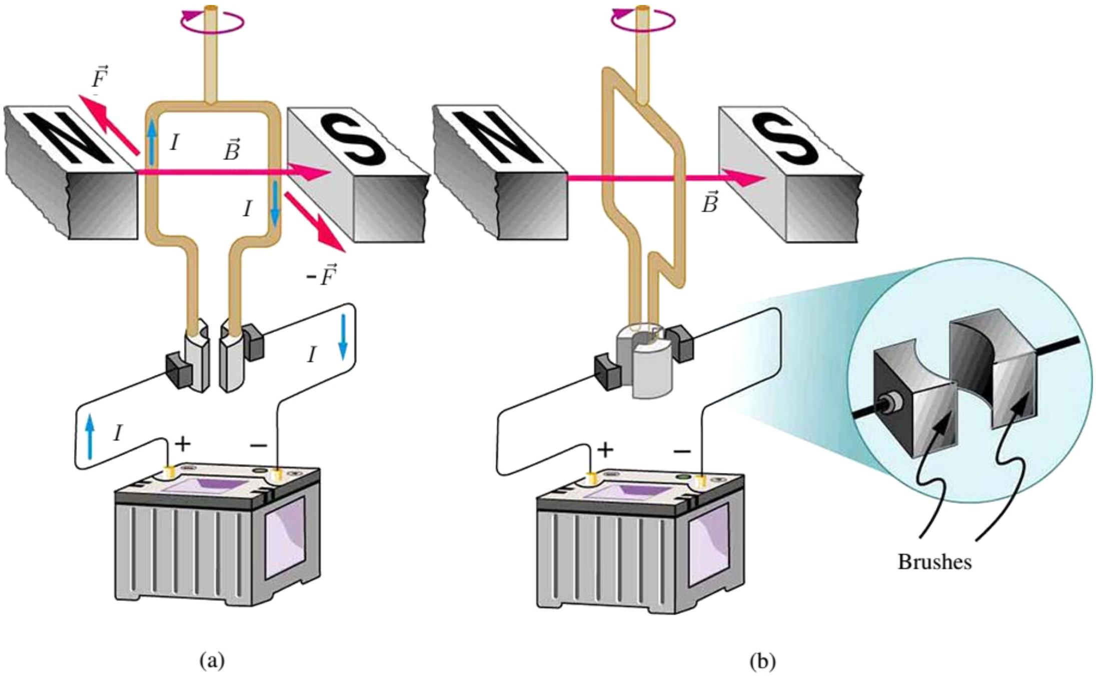
Entering known values yields:

$$\tau_{\text{max}} = 100 \cdot 15 \cdot 0.1 \cdot 2 = 30\,[\text{N} \cdot \text{m}] \tag{16.3.13}$$

This torque is large enough to be useful in a small motor.

Notice that the Re 460 Swiss locomotive has a current of 593 [A] ($2,180$ [V]) but obviously the architecture of such an engine is almost completely different from the above one!

The torque found in the preceding example is the maximum. As the coil rotates, the torque decreases to zero at $\theta = 0$. The torque then reverses its direction once the coil rotates past $\theta = 0$ (see (d) in the figure above) This means that, unless we do something, the coil will oscillate back and forth about equilibrium at $\theta = 0$. To get the coil to continue rotating in the same direction, we can reverse the current as it passes through $\theta = 0$ with automatic switches named "brushes"as illustrated below:



16.3.2.2 Average power

We have defined in the section Electrokinetics (see page 4253) that Joule power was given in constant supply for a conductor by:

$$P=UI=RI^2 \tag{16.3.14}$$

In alternating regime system and at low frequency (in order to consider the resistance as constant) we have in the purely resistive case:

$$p(t)=Ri^2(t) \tag{16.3.15}$$

It then comes (see section Statistics page 569) for the mean value of a periodic signal of period T:

$$\bar{p}=\frac{1}{T}\int_{t_0}^{t_0+T} u(t)i(t)\mathrm{d}t=R\left(\frac{1}{T}\int_{t_0}^{t_0+T} i^2(t)\mathrm{d}t\right) \tag{16.3.16}$$

The term in brackets can therefore be compared to the value that would have a direct current producing the same Joule power. Therefore:

$$\bar{p}=RI^2=R\left(\frac{1}{T}\int_{t_0}^{t_0+T} i^2(t)\mathrm{d}t\right) \tag{16.3.17}$$

We name this equivalent current the "effective current value" or "current root mean square" (abbreviated "RMS current") and we denote it by:

$$I_{\text{eff}}=\sqrt{\frac{1}{T}\int_{t_0}^{t_0+T} i^2(t)\mathrm{d}t} \tag{16.3.18}$$

The RMS values being what the multimeters measure. For the sinusoidal regimes, we then have:

$$I_{\text{eff}}=\sqrt{\frac{1}{T}\int_{t_0}^{t_0+T}i^2(t)\mathrm{d}t}=\sqrt{\frac{1}{2\pi}\int_0^{2\pi}i^2(\theta)\mathrm{d}\theta}=\sqrt{\frac{1}{2\pi}I_0^2\int_0^{2\pi}\cos^2(\theta)\mathrm{d}\theta}=\sqrt{\frac{1}{2\pi}I_0^2\pi}=\frac{I_0}{\sqrt{2}} \tag{16.3.19}$$

Therefore we have in this special case:

$$p=u(t)i(t)=u(t)\frac{I_0}{\sqrt{2}} \tag{16.3.20}$$

And using the same method, wishing to calculate the average voltage, we get in sinusoidal regime:

$$U_{\text{eff}}=\frac{U_0}{\sqrt{2}} \tag{16.3.21}$$

Therefore for a voltage and current having the same phase, we can write:

$$\bar{p}=U_{\text{eff}}I_{\text{eff}}=\frac{U_0}{\sqrt{2}}\frac{I_0}{\sqrt{2}}=\frac{U_0I_0}{2} \tag{16.3.22}$$

Strictly speaking, we must generalize this last relation for situations where the current and the voltage are out of phase with an angle (or time) φ.

Remark

Unfortunately, many electronic books or even research articles have defined a power that has no physical meaning and that is named "RMS power" and measures "RMS watts" defined by analogy by:

$$P_{\text{eff}}=\sqrt{\frac{1}{T}\int_t^{t+T}p^2(t)\mathrm{d}t} \tag{16.3.23}$$

Even though this term is used by advertisers and some editors, it has no place in good technical publications. It often appears to give a semblance of technical expertise...

Let us consider now another approach of the Power this time from the instantaneous point of view in a more general than purely resistive case. In a more general case, we have the right to write the voltage and the current in the sinusoidal steady state with cosines in the following form (which implicitly also takes into account the phase shift of current and voltage):

$$\begin{aligned}u(t)&=U_0\cos(\omega t+\alpha)\\ i(t)&=I_0\cos(\omega t+\beta)\end{aligned} \tag{16.3.24}$$

We then have using the remarkable trigonometric relations (see section Trigonometry page 2519):

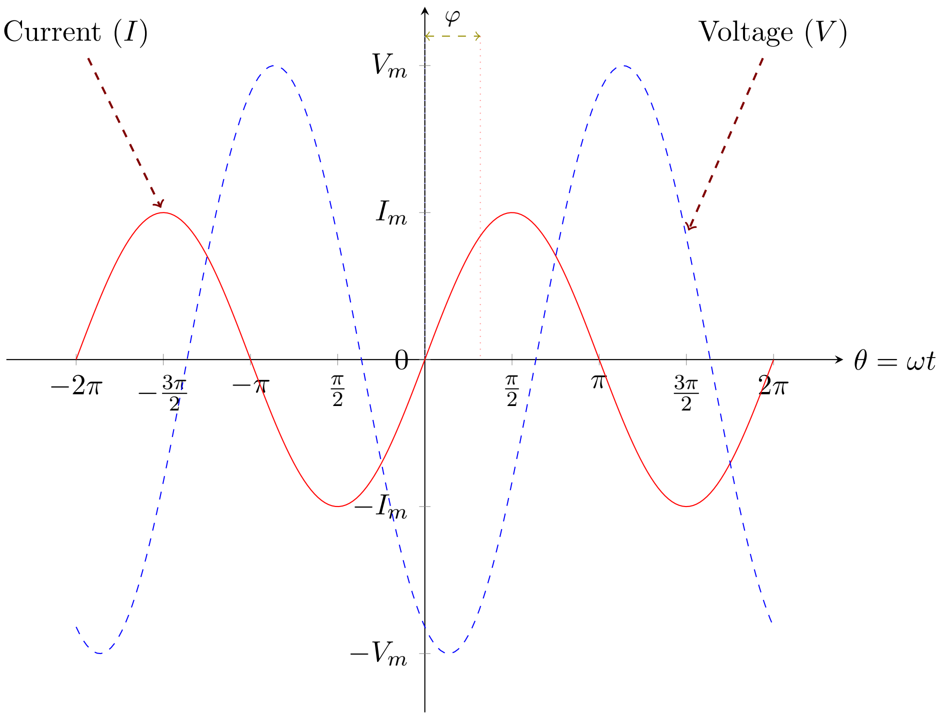
$$\begin{aligned}p(t)&=u(t)i(t)=I_0U_0\cos(\omega t+\alpha)\cos(\omega t+\beta)\\ &=U_0I_0\left(\frac{\cos(\alpha-\beta)+\cos(2\omega t+\alpha+\beta)}{2}\right)=\frac{U_0I_0}{2}\cos(\varphi)+\frac{U_0I_0}{2}\cos(2\omega t+\alpha+\beta)\\ &=U_{\text{eff}}I_{\text{eff}}\cos(\varphi)+U_{\text{eff}}I_{\text{eff}}\cos(2\omega t+\alpha+\beta)\end{aligned} \tag{16.3.25}$$

where we are free to put:

$$\varphi = \alpha - \beta \tag{16.3.26}$$

such that this term φ is positive or null (we can always choose to subtract one of the two terms to the other in order to have a strictly positive or zero φ value without changing the developments and conclusions that follow below) and where we used the results obtained previously:

$$U_{\text{eff}} = \frac{U_0}{\sqrt{2}} \qquad I_{\text{eff}} = \frac{I_0}{\sqrt{2}} \tag{16.3.27}$$



Therefore, in the expression:

$$p(t) = U_{\text{eff}} I_{\text{eff}} \cos(\varphi) + U_{\text{eff}} I_{\text{eff}} \cos(2\omega t + \alpha + \beta) \tag{16.3.28}$$

the instantaneous power thus comprises a constant component (first term) and a pulsed component with a frequency double that of the current and the voltage.

By using:

$$\varphi = \alpha - \beta \Rightarrow \beta = \alpha - \varphi \tag{16.3.29}$$

as well as the following remarkable trigonometric relations (see section Trigonometry page 2519):

$$\cos(2\omega t + \alpha + \beta) = \cos(2\omega t + 2\alpha - \varphi) = \cos(\varphi)\cos(2\omega t + 2\alpha) + \sin(\varphi)\sin(2\omega t + 2\alpha) \tag{16.3.30}$$

We then have:

$$\begin{aligned} p(t) &= U_{\text{eff}} I_{\text{eff}} \cos(\varphi) + U_{\text{eff}} I_{\text{eff}} \left[\cos(\varphi)\cos(2\omega t + 2\alpha) + \sin(\varphi)\sin(2\omega t + 2\alpha)\right] \\ &= U_{\text{eff}} I_{\text{eff}} \cos(\varphi)\left[1 + \cos(2\omega t + 2\alpha)\right] + U_{\text{eff}} I_{\text{eff}} \sin(\varphi)\sin(2\omega t + 2\alpha) \end{aligned} \tag{16.3.31}$$

Denoted:

$$\boxed{p(t)=U_{\text{eff}}I_{\text{eff}}\cos(\varphi)\left[1+\cos(2\omega t+2\alpha)\right]+U_{\text{eff}}I_{\text{eff}}\sin(\varphi)\sin(2\omega t+2\alpha)} \tag{16.3.32}$$

This last relation highlights that the instantaneous power can always be reduced in a permanent sinusoidal regime as the sum of two terms where:

- The first term is a pulsed component, always positive (by construction), which oscillates around the mean value $U_{\text{eff}}I_{\text{eff}}\cos(\varphi)$ and which translates a unidirectional energy exchange between a source and a load.
- The second term is an alternating component that varies sinusoidally with an amplitude $U_{\text{eff}}I_{\text{eff}}\sin(\varphi)$ and a zero average value. It is therefore alternately positive and negative and translates an oscillatory exchange of energy between a source and a charge.

When $\varphi=0$ (perfect purely resistive load), we then have:

$$\cos(\varphi)=1 \qquad \sin(\varphi)=0 \tag{16.3.33}$$

The mean value $U_{\text{eff}}I_{\text{eff}}\cos(\varphi)$ is then maximum and equal to $U_{\text{eff}}I_{\text{eff}}$, while the second term (the alternative component) is zero.

On the other hand, when $\varphi=\pm\dfrac{\pi}{2}$ (purely reactive load as an ideal inductance or capacity), then:

$$\cos(\varphi)=0 \qquad \sin(\varphi)=\pm 1 \tag{16.3.34}$$

and in this case the instantaneous power is reduced to the only alternative component. From this arises the following definitions:

Definitions (#16.3.728):

D1. We name "active power" the term:

$$\boxed{P=U_{\text{eff}}I_{\text{eff}}\cos(\varphi)=U_{\text{rms}}I_{\text{rms}}\cos(\varphi)} \tag{16.3.35}$$

and this quantity is measurable by a wattmeter and which corresponds to a real supply of energy convertible to work or heat and which is therefore maximum in the case of a purely resistive (ideal) load and zero in the case of a purely reactive (ideal) charge.

D2. We name "reactive power" the term:

$$\boxed{Q=U_{\text{eff}}I_{\text{eff}}\sin(\varphi)=U_{\text{rms}}I_{\text{rms}}\sin(\varphi)} \tag{16.3.36}$$

which is not measurable by a wattmeter since equal to zero as it is alternative.

D3. We name "apparent power" the product $U_{\text{eff}}I_{\text{eff}}$ that is in faced denoted most of time as:

$$\boxed{S=\sqrt{P^2+Q^2}=U_{\text{eff}}I_{\text{eff}}} \tag{16.3.37}$$

hence the famous "power triangle":

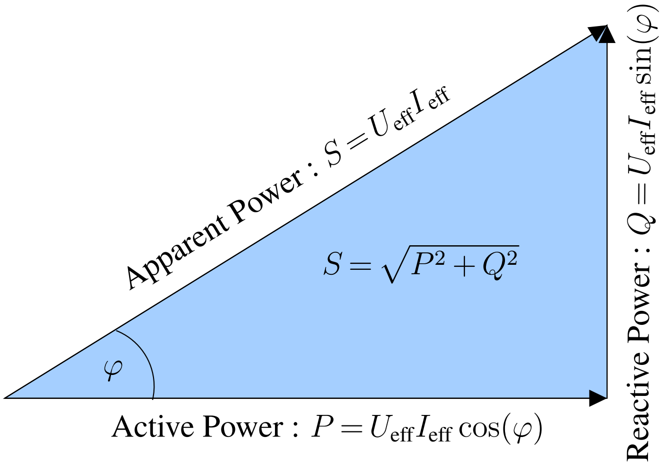

Apparent Power : $S = U_{\text{eff}} I_{\text{eff}}$
$S = \sqrt{P^2 + Q^2}$
φ
Active Power : $P = U_{\text{eff}} I_{\text{eff}} \cos(\varphi)$
Reactive Power : $Q = U_{\text{eff}} I_{\text{eff}} \sin(\varphi)$

Figure 16.100: Electric Power triangle

Which is in appearance a power but does not necessarily provide work, hence its name ... Traditionally it is the apparent power that is indicated on large industrial installations. It can be recognized because the units are often indicated in volts-amps [VA] rather than in watts [W] (because this is not totally a power)!

16.3.3 Transformers

A static electric transformer is an electrical machine which makes it possible to modify the voltage and current intensity values delivered by an alternative electrical power source into a voltage and current system of different values but with the same frequency and the same shape.

Since the invention of the first constant potential transformer in 11885 (holocene calendar), transformers have become essential for the transmission, distribution, and utilization of alternating current electrical energy. A wide range of transformer designs is encountered in electronic and electric power applications. Transformers range in size from RF transformers less than a cubic centimetre in volume to units interconnecting the power grid weighing hundreds of tons for Nuclear Plant.

Let us recall that we have proved in the section of Electrokinetics that (see page 4263):

$$e = -N\frac{\mathrm{d}\Phi}{\mathrm{d}t} = -L\frac{\mathrm{d}i}{\mathrm{d}t} \qquad (16.3.38)$$

If we can in one way or another to pass the flow of a first solenoid, then named "primary winding", having:

$$e_1 = -N_1\frac{\mathrm{d}\Phi}{\mathrm{d}t} = -L_1\frac{\mathrm{d}i}{\mathrm{d}t} \qquad (16.3.39)$$

to a second solenoid, then named "secondary winding", without any loss (or at least a negligible loss) such as:

$$e_2 = -N_2\frac{\mathrm{d}\Phi}{\mathrm{d}t} = -L_2\frac{\mathrm{d}i}{\mathrm{d}t} \qquad (16.3.40)$$

Then it comes by term-by-term identification:

$$\frac{e_1}{e_2}=\frac{N_1}{N_2} \tag{16.3.41}$$

And if the internal resistances are negligible the induced electromotive force is then equal to the voltage at the terminals, then it comes:

$$\frac{U_1}{U_2}=\frac{N_1}{N_2} \tag{16.3.42}$$

And if all the energy of the magnetic field is transmitted in the secondary winding (without loss) by the conservation of energy, we have:

$$P_1=P_2\Leftrightarrow U_1I_1=U_2I_2 \tag{16.3.43}$$

from where:

$$\boxed{\frac{U_1}{U_2}=\frac{I_2}{I_1}=\frac{N_1}{N_2}} \tag{16.3.44}$$

Thus, the ratio of the number of primary turns to the number of secondary turns totally determines the transformation ratio of the transformer which can then be used as a voltage transformation station by raising or lowering it as a function of the number of turns of the secondary winding. It should also be noted that a transformer which increases the voltage simultaneously decreases the current and vice versa.

A typical historical and pedagogical instrument to do this is the single-phase transformer with a ferromagnetic core:

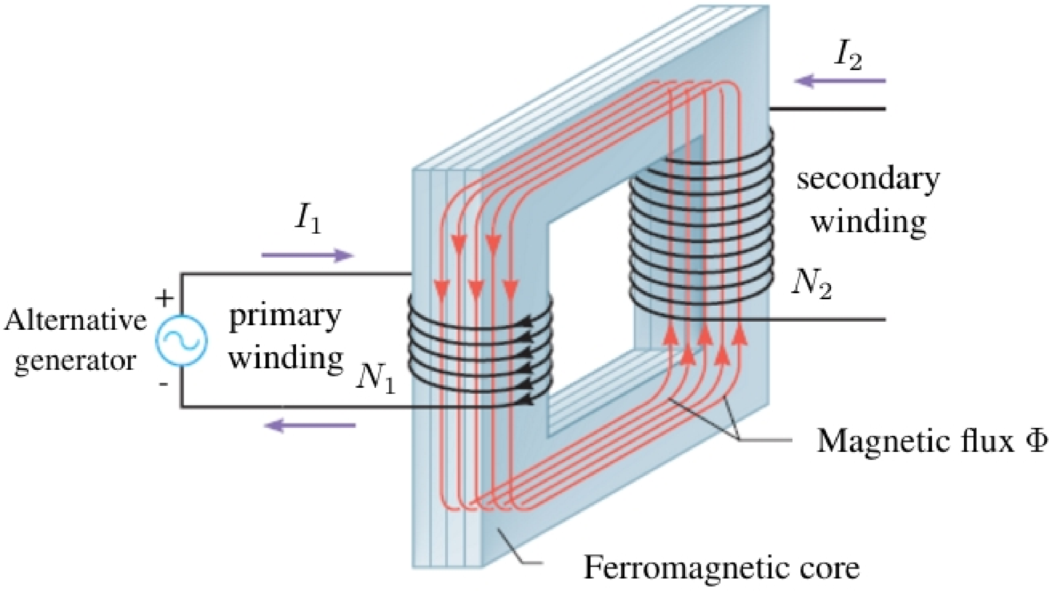


Figure 16.101: **Typical example of a single-phase transformer with a ferromagnetic core**

In practice, monophasic transformers have windings that are concentric to minimize flux leakage. An insulation is inserted between the primary circuit and the secondary:

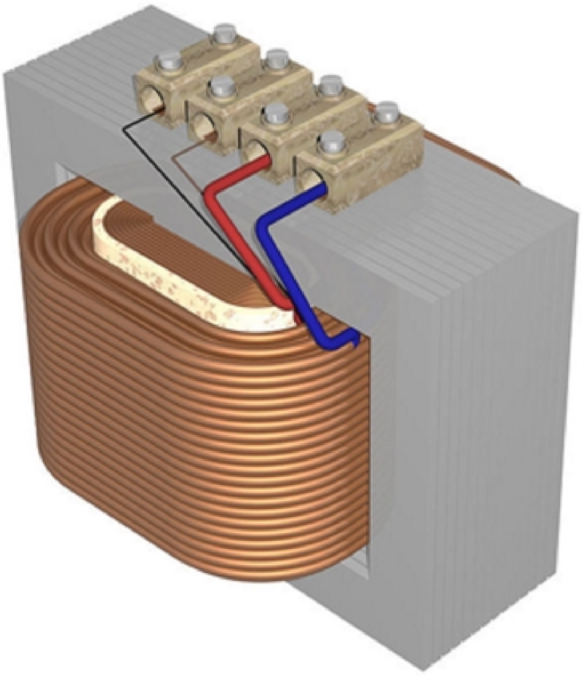


Figure 16.102: **Concentric monophasic transformer type (source: Wikipedia)**

Closed-core transformers are constructed in "core form" or "shell form". When windings surround the core, the transformer is core form; when windings are surrounded by the core, the transformer is shell form. Shell form design may be more prevalent than core form design for distribution transformer applications due to the relative ease in stacking the core around winding coils. Core form design tends to, as a general rule, be more economical, and therefore more prevalent, than shell form design for high voltage power transformer applications at the lower end of their voltage and power rating ranges (less than or equal to, nominally, 230 [kV] or 75 [MVA]). At higher voltage and power ratings, shell form transformers tend to be more prevalent. Shell form design tends to be preferred for extra-high voltage and higher [MVA] applications because, though more labor-intensive to manufacture, shell form transformers are characterized as having inherently better kVA-to-weight ratio, better short-circuit strength characteristics and higher immunity to transit damage:

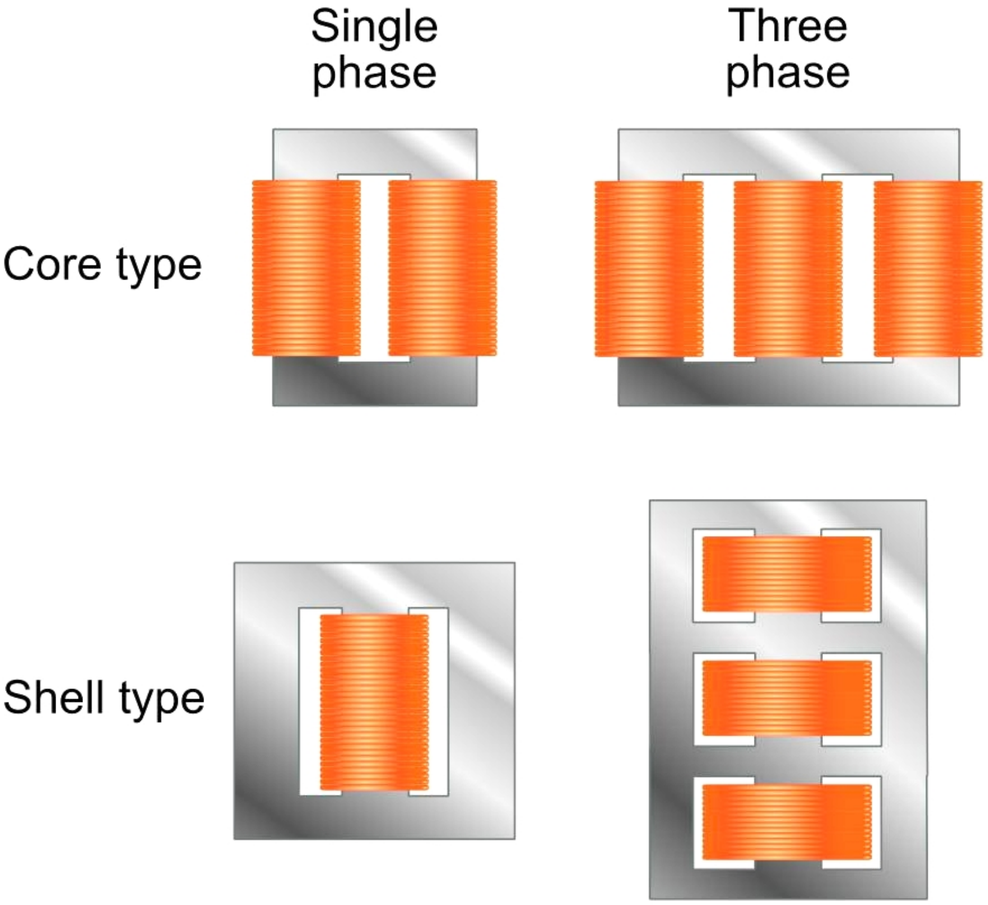

Single phase
Three phase
Core type
Shell type

Figure 16.103: **Schematic representation of various transformers (source: Wikipedia, author: Spinningspark)**

Transformers by their ability to increase voltage and reduce current (saving by Joule loss as we have already mentioned) play an important role in the transmission of electricity from domestic infrastructures (low voltage: LV, medium voltage: MV, high voltage HV). Thus we find the step-up transformers (SUT) and step-down transformers (STD) in everyday life:

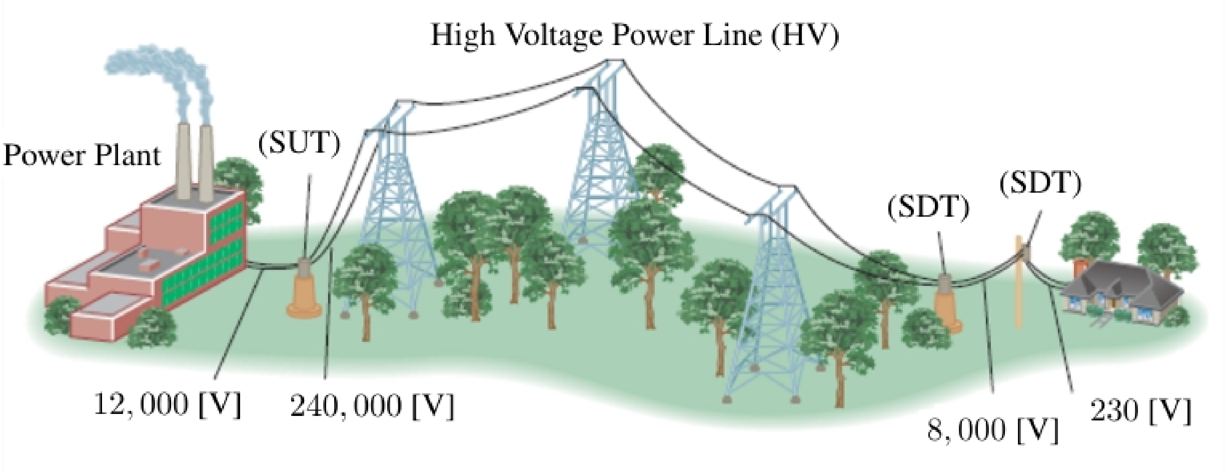

High Voltage Power Line (HV)
Power Plant
(SUT)
(SDT)
(SDT)
12, 000 [V]
240, 000 [V]
8, 000 [V]
230 [V]

Figure 16.104: **Typical civilian voltage transformation cascade installation in Switzerland**

Here is an example of the largest triphasic step-up transformer (SUT) in the world in 12014 (according to holocene calendar):

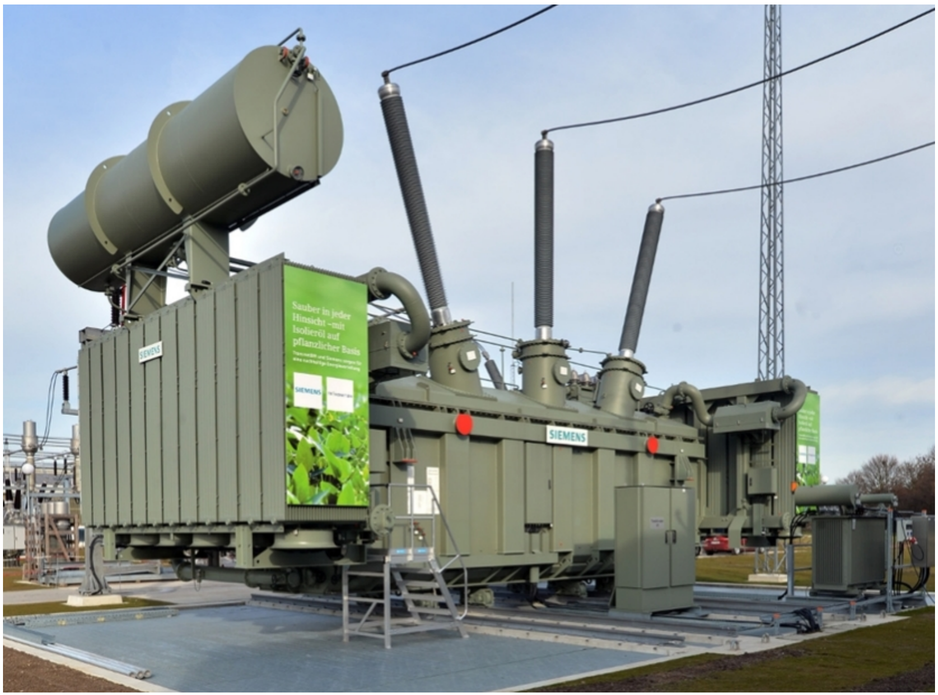

Sauber in jeder Hinsicht –mit Isolieröl auf pflanzlicher Basis
SIEMENS
SIEMENS

Figure 16.105: **Largest plant-oil-based transformer in the World in 12014 (source: TransnetBW GmbH)**

And a primary triphasic step-down transformer with a slice cut:

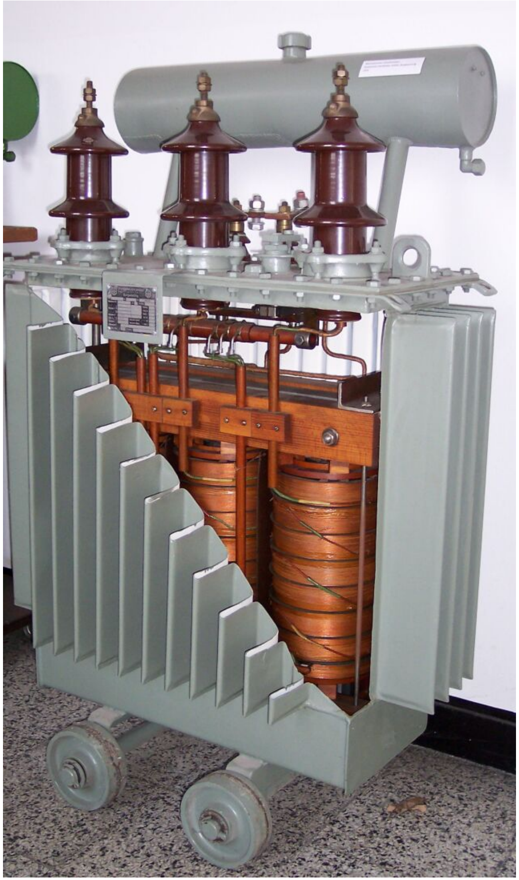


Figure 16.106: **Cutaway view of liquid-immersed construction transformer. The conservator (reservoir) at top provides liquid-to-atmosphere isolation as coolant level and temperature changes. The walls and fins provide required heat dissipation balance (source: Wikipedia)**

And a typical USA pole-mounted distribution step-down transformer with center-tapped secondary winding used to provide split-phase power for residential and light commercial service:

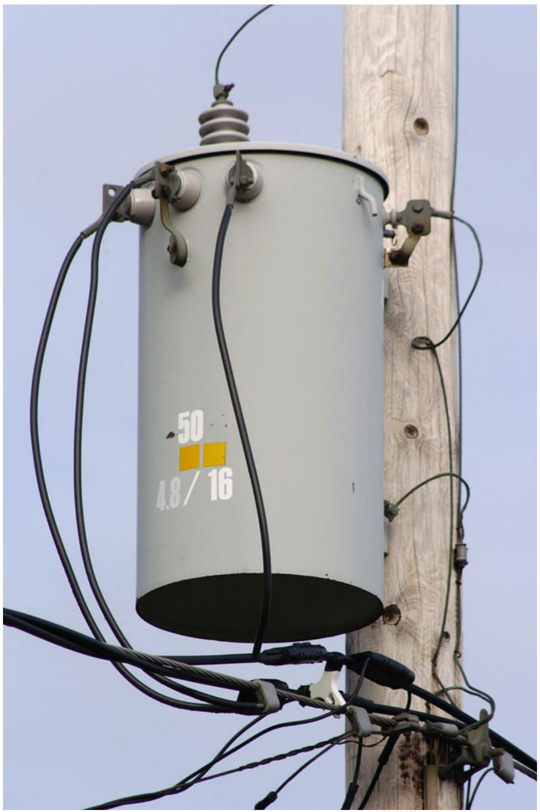


Figure 16.107: **Typical pole-mounted distribution transformer in USA with center-tapped secondary winding (source: Wikipedia)**

And finally typical small size domestic transformers for computer, TV, Hi-Fi, etc. :

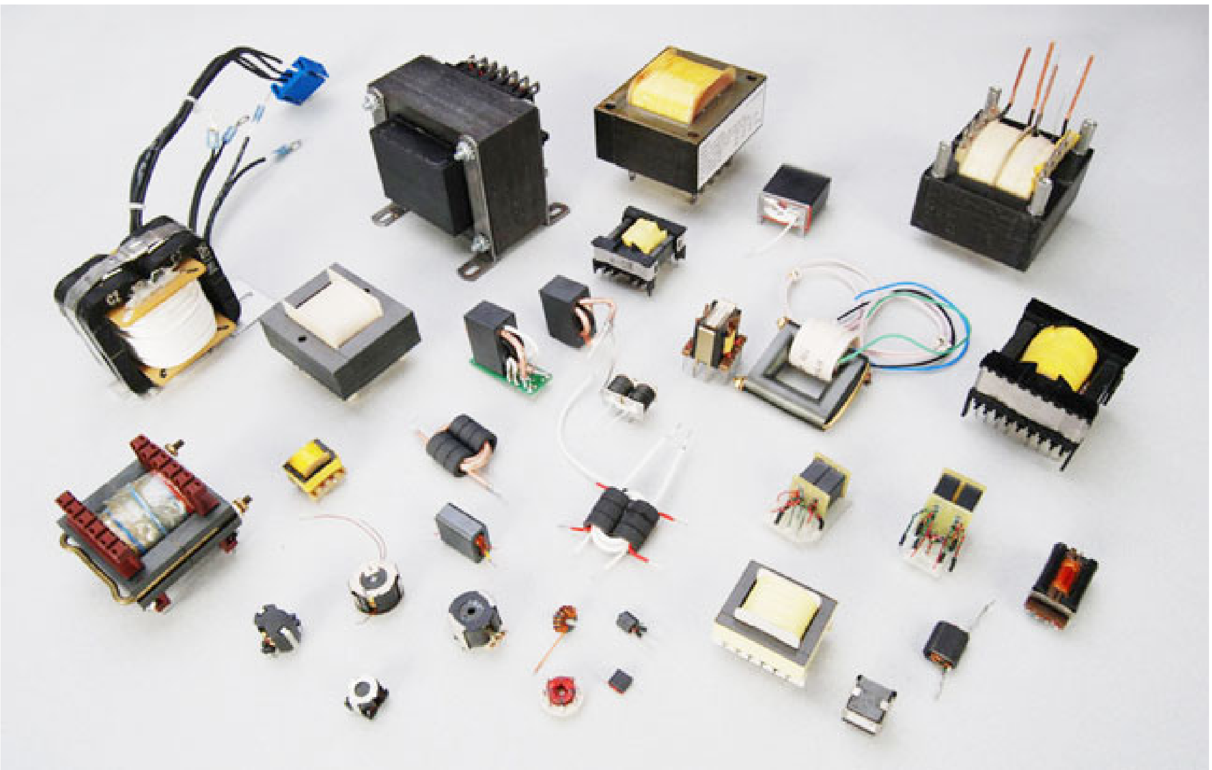


Figure 16.108: Small domestic size transformers (source: Mohawk Electro Techniques Inc.)

16.3.3.1 Transformer universal EMF equation

If the flux in the core is purely sinusoidal, the relationship for either winding between its RMS voltage e_{eff} of the winding, and the supply frequency f, number of turns N, core cross-sectional area S and peak magnetic flux density B_{peak} is given by the universal EMF equation that we will prove now!

The derivation of EMF equation of the transformer is shown below. First remember that we have:

$$e=-N\frac{\mathrm{d}\Phi}{\mathrm{d}t} \tag{16.3.45}$$

Explicitly:

$$e=-N\frac{\mathrm{d}\hat{\Phi}\sin(\omega t)}{\mathrm{d}t}=-N\omega\hat{\Phi}\cos(\omega t)=-N\omega\hat{\Phi}\sin\left(\omega t-\frac{\pi}{2}\right) \tag{16.3.46}$$

So we see again that the induced EMF lags flux by 90°.

The maximum value of EMF is obviously equal to:

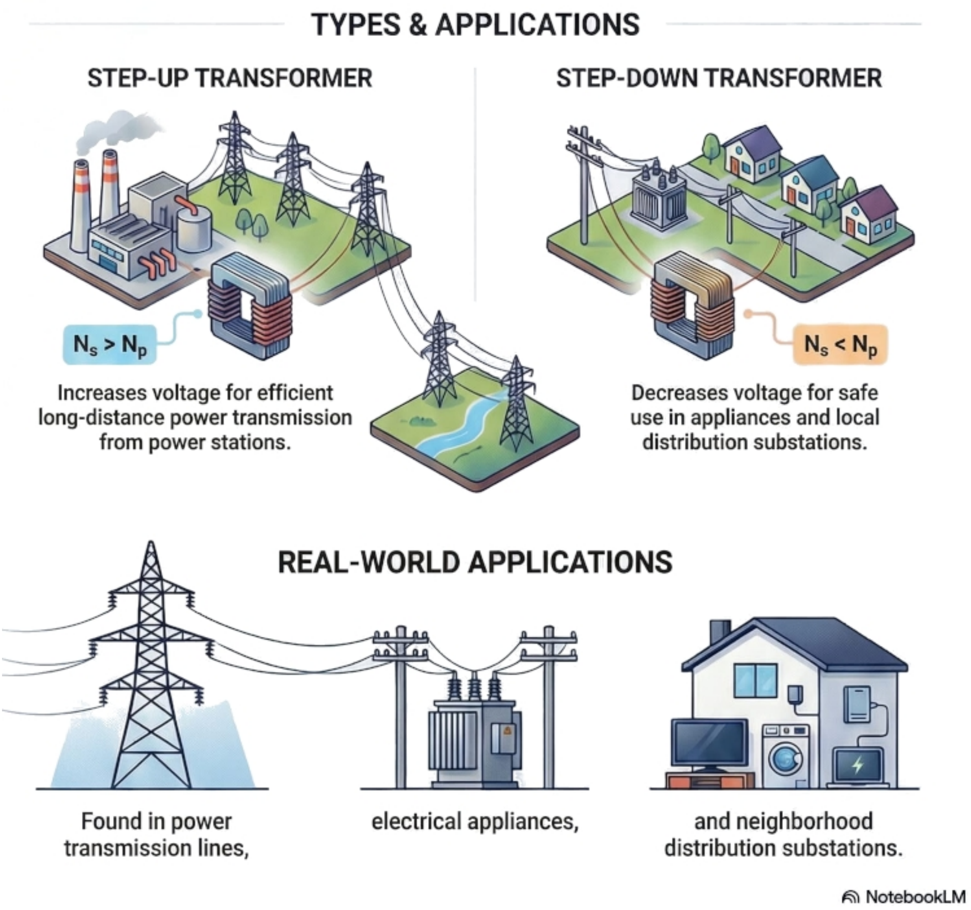
$$e_{\max}=N\omega\hat{\Phi}=N2\pi f\hat{\Phi} \tag{16.3.47}$$

The root mean square RMS value is therefore:

$$e_{\text{eff}}=\frac{e_{\max}}{\sqrt{2}}=\frac{N2\pi f\hat{\Phi}}{\sqrt{2}}=\sqrt{2}\pi fN\hat{\Phi}\cong 4.44fN\hat{\Phi} \tag{16.3.48}$$

As $\Phi = B_{\text{peak}} \cdot S$ we get finally the "transformer universal EMF equation":

$$\boxed{e_{\text{eff}} \cong 4.44 f N S B_{\text{peak}}} \tag{16.3.49}$$


TYPES & APPLICATIONS
STEP-UP TRANSFORMER
STEP-DOWN TRANSFORMER
N_s > N_p
Increases voltage for efficient long-distance power transmission from power stations.
N_s < N_p
Decreases voltage for safe use in appliances and local distribution substations.
REAL-WORLD APPLICATIONS
Found in power transmission lines,
electrical appliances,
and neighborhood distribution substations.
NotebookLM

16.3.4 Steady State Linear Circuits

We will see here circuits composed of simple elements such as resistance, capacitance and impedance. These circuits, whose representative equation is a linear differential equation, are named "linear circuits". Moreover, they are an excellent example to see the cumbersome developments using classical mathematical tools as opposed to other more flexible and powerful techniques (representation by phasors and Laplace transforms).

Before we start, the reader must keep in mind something quite important! As we have already mentioned it, it's not the electrons that carry the energy, neither they push each other (the water flow analogy being a wrong equivalence as every analogy in Science). If we consider the naive circuit below with a battery and a lamp:

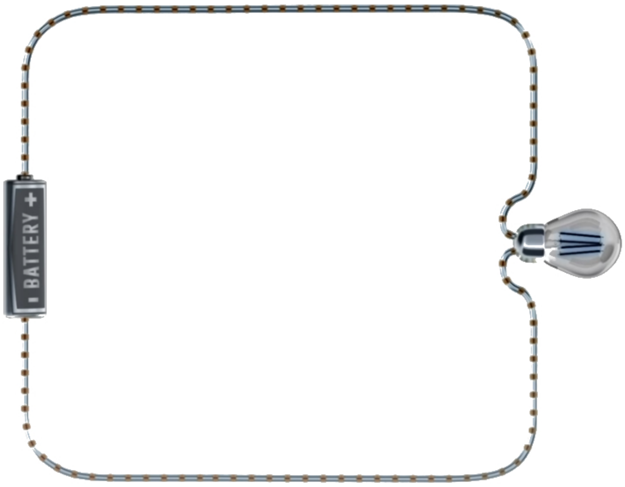


Figure 16.109: **Naive battery lamp circuit (source: Veritasium)**

The energy is in fact transported by the Poynting vector (see section Electrodynamics page 4156) of the electromagnetic field as the circuit behaves like a dipole antenna:

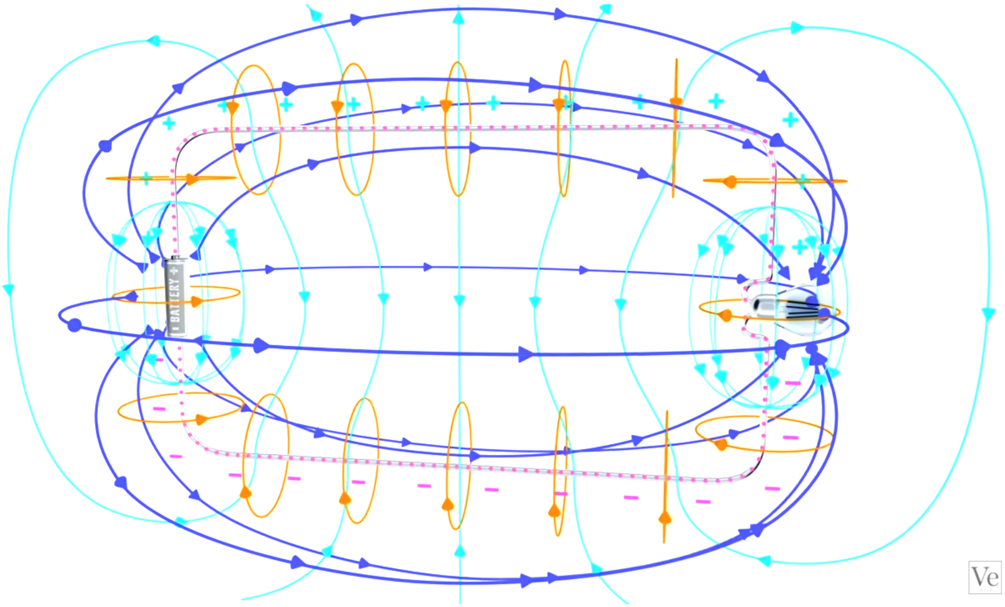


Figure 16.110: **Naive circuit behaves like an antenna (source: Veritasium)**

To have a look at how energy gets into a light globe we consider the related case of a segment of wire that is nice and straight instead of being coiled up tightly. We need to look at the electric and magnetic fields just inside the surface of the wire. As we have already seen, the electric field $\vec{E}$ inside the wire points along the axis of the wire and, just inside the surface, the lines of $\vec{E}$ lie along the surface and parallel to it. The lines of the magnetic field $\vec{B}$ form closed loops around the wire. The diagram shows some of these lines just inside the surface of the wire. Now the energy density vector $\vec{S}$ must be perpendicular to the lines of both $\vec{E}$ and $\vec{B}$ so it must be every where perpendicular to the wire's surface.

Using the right-hand rule for the vector product stated above, we can conclude that energy is flowing into the wire at every point on its curved surface.

To calculate the value of the power transfer, just multiply the value of the Poynting vector by the surface area. Since the electric and magnetic fields are right angles to each other, the magnitude of the energy flux density is just:

$$\vec{S} = \frac{1}{\mu}\vec{E}\times\vec{B} = \frac{1}{\mu}EB \tag{16.3.50}$$

The power input is the product of the power per area, S, and the curved exterior surface area of the cylinder of radius r and length d, A :

$$P = SA = \frac{1}{\mu}EB2\pi rd \tag{16.3.51}$$

Now we can find values for the fields in terms of the potential difference and current. The magnitude of the electric field is the potential difference between the ends of the segment of wire divided by the length of the wire:

$$E = \frac{U}{d} \tag{16.3.52}$$

and the magnetic field produced by the current at the surface of the wire is given by the well-known formula:

$$B = \frac{\mu I}{2\pi r} \tag{16.3.53}$$

Substituting these two expressions for the fields gives the familiar formula for the power dissipation:

$$P = UI \tag{16.3.54}$$

So the field model for energy transfer gives the right answer for energy dissipation in a resistive wire.

As an extension to the derivation above, we could look at what happens just outside the wire. Since the axial component of the electric field is the same just outside as it is inside and the magnetic field is practically the same, we can conclude that the energy flux derived above still works out as expected. But ... now there is also a radial component of the electric field as well. If that radial component is not zero then, together with the magnetic field, it contributes to a component of the Poynting vector that is parallel to the wire - but only outside the wire. That means that energy may also be travelling along beside the wire (and getting in somewhere else).

A similar mathematical argument can be used to find the rate at which energy is extracted from the battery by the electromagnetic field. Just model the battery as a cylindrical object and the whole argument is the same except for one thing: in the battery the current goes from low potential to high potential, against the electric field, so the direction in which the magnetic field loops around the battery is reversed. That's the only difference in the argument, but it reverses the direction of the Poynting vector at the surface. So energy comes out of the cylinder instead of going in. Conservation of energy rules!

The point is to reveal something that's most of time hidden by the way that we think about and teach electric circuits. We use voltage and current and lumped elements because they're more convenient than working with Maxwell equation. But we should never forget that the main actors are actually the fields. They are what carry the energy!

Hence as we have just seen Maxwell's equations combined with Poynting's theorem give us a model where electricity is energy carried along with the electromagnetic field (energy is stored in the field, not into kinetic energy of the electron in this model).

However we also know that the Drude model (see section Elektronkinetics page 4226) was created using kinetic gas theory and Boltzmann statistics. In this theory, electricity is carried by movement of "free electrons" that are accelerated by electric field and then "bounce" along the way. We are able to get "drift velocity":

$$v=\frac{qE}{m}\tau \tag{16.3.55}$$

of the electron this way and this velocity is quite small in the order of a meter per hour!

The theory combining the concepts of quantum mechanics and Fermi-Dirac statistics, resulting in Sommerfeld model that is still used in solid-state physics. Drift velocity is replaced with Fermi velocity (see section Elektronkinetics page 4305):

$$v_F=\sqrt{\frac{2E_{c,F}}{m_e}} \tag{16.3.56}$$

and electrons that have Fermi energy will take part in electric conductivity. With the quasi-particle model, semi conductors are explained pretty well and it's a very useful way to do it because even a single small grain of sand contains so many electrons that solving the exact Schrödinger partial differential equation would be practically impossible. Sometimes the real dynamics can be replaced by a quasi-particle model, which is based on the overall result of the interactions of a complex system. The Fermi velocity for copper is in the order of $1\cdot 10^6$ $[\text{m}\cdot\text{s}^{-1}]$.

Element	**Fermi Energy** [eV]	**Fermi Temperature** $\times 10^4$ [K]	**Fermi Velocity** $\times 10^6$ $[\text{m} \cdot \text{s}^{-1}]$
Li	4.74	5.51	1.29
Na	3.24	3.77	1.07
K	2.12	2.46	0.86
Rb	1.85	2.15	0.81
Cs	1.59	1.84	0.75
Cu	7.00	8.16	1.57
Ag	5.49	6.38	1.39
Au	5.53	6.42	1.40
Be	14.3	16.6	2.25
Mg	7.08	8.23	1.58
Ca	4.69	5.44	1.28
Sr	3.93	4.57	1.18
Ba	3.64	4.23	1.13
Nb	5.32	6.18	1.37
Fe	11.1	13.0	1.98
Mn	10.9	12.7	1.96
Zn	9.47	11.0	1.83
Cd	7.47	8.68	1.62
Hg	7.13	8.29	1.58
Al	11.7	13.6	2.03
Ga	10.4	12.1	1.92
In	8.63	10.0	1.74
Tl	8.15	9.46	1.69
Sn	10.2	11.8	1.90
Pb	9.47	11.0	1.83
Bi	9.90	11.5	1.87
Sb	10.9	12.7	1.96

Table 16.9: **Fermi Energies, Fermi Temperatures, and Fermi Velocities (source: N. W. Ashcroft and N. D. Mermin)**

And we have a third speed, the group velocity speed given by (see section Electromagnetism page 4272):

$$v_g = \sqrt{\frac{8\omega}{\mu\sigma}} \tag{16.3.57}$$

For copper the latter is of order of six meters per second as we have calculated it in the section of Electromagnetism.

The three models for the speed above are useful, correct and consistent with each other! That's why its sometimes confusing for some people!

16.3.4.1 RC series circuit

Any circuit having a capacitor also have a resistance, if only that of the connection wires. Such RC (Resistance-Capacitor) series circuits are very common and sometimes of great importance (pacemaker for example). Indeed, when we close a circuit which contains only resistors (purely resistive circuit), the current rises to its nominal value in an extremely short time, so that we can consider that the current and the voltage are constant with an excellent approximation a short time after the circuit was closed (at least in common civil applications...). Thus, the permanent regime is established after a very brief transitional regime. On the contrary, in a series RC circuit, voltage and current take a long time to reach their nominal values. This dependence over time has multiple applications and allows to produce a whole range of signals that can be modulated over time as required.

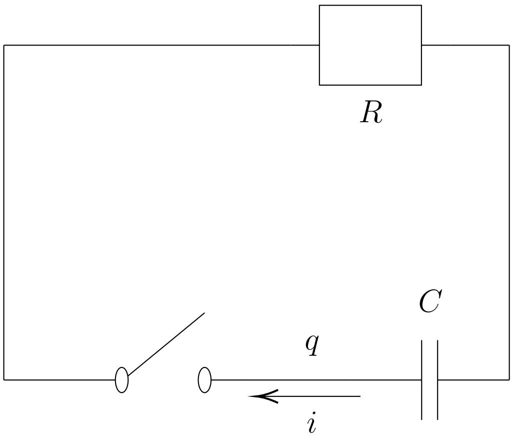


Figure 16.111: **RC serie circuit**

The study of such a circuit is also important in science as quite many mechanical (like damped oscillating mass) or economic systems (price dynamic to events) reacts like an RC circuit.

16.3.4.1.1 RC series step response

We assume that initially the capacitor is charged and that no current flows (open switch):

$$q(t=0)=q_0 \quad \text{and} \quad i(t=0)=0 \tag{16.3.58}$$

When we close the switch the electrons go away from the capacitor C. We then have at the terminals of the resistor:

$$u_R=R\cdot i=R\frac{\mathrm{d}q}{\mathrm{d}t} \tag{16.3.59}$$

At the terminals of the capacitor:

$$u_C=\frac{1}{C}q \tag{16.3.60}$$

The equation of the circuit is then:

$$R\frac{\mathrm{d}q}{\mathrm{d}t}+\frac{1}{C}q=0 \tag{16.3.61}$$

Therefore:

$$\frac{1}{q}\mathrm{d}q=-\frac{1}{RC}\mathrm{d}t \tag{16.3.62}$$

Trivial differential equation (but we can detail on readers request) whose solution is with the initial conditions:

$$\ln\left(\frac{Q}{Q_0}\right)=-\frac{t}{RC} \tag{16.3.63}$$

Therefore:

$$q=Q_0e^{-t/RC} \tag{16.3.64}$$

The current then has the following form:

$$i=-\frac{Q_0}{RC}e^{-t/RC} \tag{16.3.65}$$

It is therefore a system in which the current decreases exponentially and this even faster than the RC factor named the "time constant" is small. We see this kind of system when the light inside a car turns off slowly after closing the doors.

So in the end we have the two following two relations for the discharge:

$$\begin{aligned}i&=-\frac{Q_0}{RC}e^{-t/RC}\\ u_C&=\frac{Q_0}{C}e^{-t/RC}\end{aligned} \tag{16.3.66}$$

When this regime is placed under a constant voltage equation, then we have an equation of the form:

$$R\frac{\mathrm{d}q}{\mathrm{d}t}+\frac{1}{C}q=U_0 \tag{16.3.67}$$

An obvious special solution is then:

$$q=Q_0=CU_0 \tag{16.3.68}$$

We then have the general solution:

$$q=-Q_0e^{-t/RC}+CU_0 \tag{16.3.69}$$

either for the current:

$$i=\frac{\mathrm{d}q}{\mathrm{d}t}=\frac{Q_0}{RC}e^{-t/RC} \tag{16.3.70}$$

And for the voltage at the terminals of the capacitor:

$$u_C=\frac{1}{C}q=-\frac{Q_0}{C}e^{-t/RC}+U_0=-U_0e^{-t/RC}+U_0=U_0(1-e^{-t/RC}) \tag{16.3.71}$$

Which therefore represents the voltage at the terminals of the capacitor during charging. So in the end we have the two following two relations for the charge:

$$\begin{gathered} i=\frac{Q_0}{RC}e^{-t/RC} \\ u_C=U_0(1-e^{-t/RC}) \end{gathered} \tag{16.3.72}$$

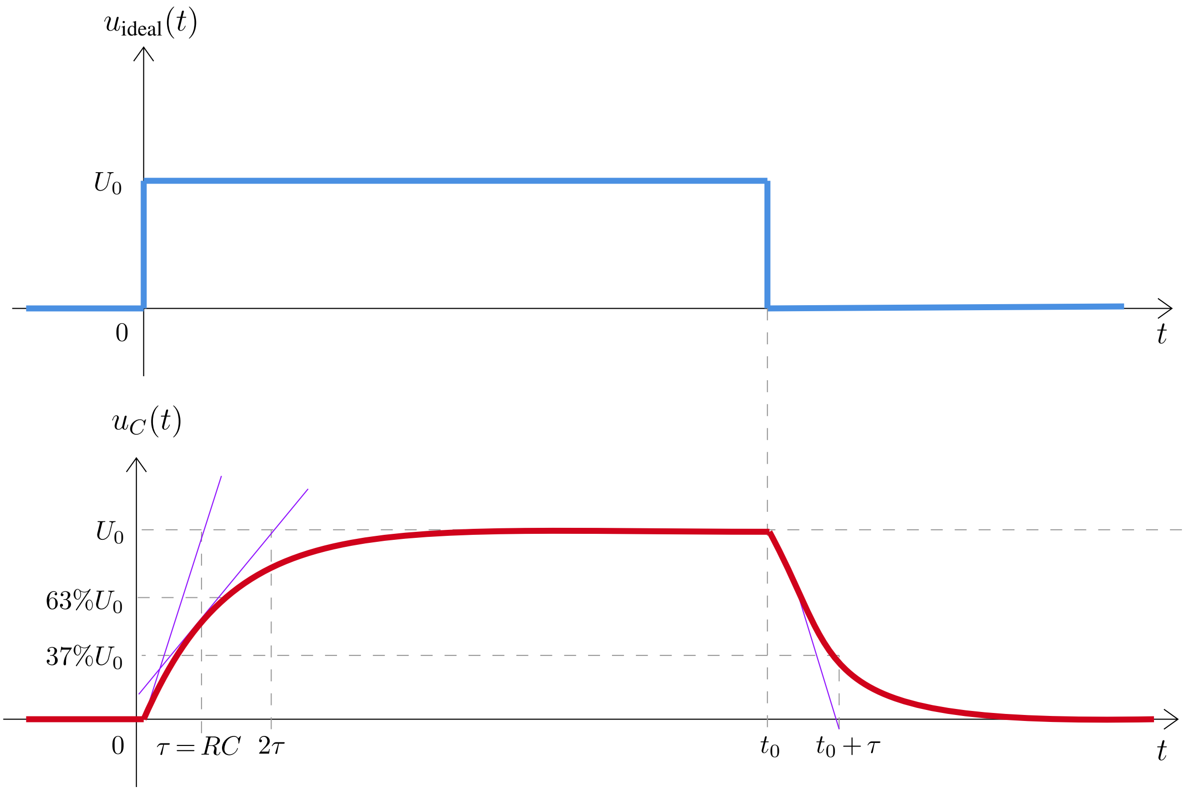


Figure 16.112: **Charging and discharging of the capacitor when opening / closing the switch**

Let us now study the energetic aspect of this circuit that is important in engineering, since power consumption or loss of power is a major selling factor in some applications!

For this purpose let us take up the relation:

$$R\frac{\mathrm{d}q}{\mathrm{d}t}+\frac{1}{C}q=R\frac{\mathrm{d}q}{\mathrm{d}t}+u_C=U_0 \tag{16.3.73}$$

And let us multiply by i:

$$Ri^2+u_Ci=U_0i \tag{16.3.74}$$

What we write:

$$p_R+p_C=p_G \tag{16.3.75}$$

As:

$$u_C=\frac{1}{C}q\Rightarrow\frac{\mathrm{d}u_C}{\mathrm{d}t}=\frac{1}{C}\frac{\mathrm{d}q}{\mathrm{d}t}=\frac{1}{C}i\Rightarrow i=C\frac{\mathrm{d}u_C}{\mathrm{d}t} \tag{16.3.76}$$

Where we see that as soon as the transient charge or discharge is completed, the voltage at the terminals of the capacitor being zero then the current is also zero.

We have then:

$$U_0 i = Ri^2 + u_0 C\frac{\mathrm{d}u_C}{\mathrm{d}t} = Ri^2 + \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}Cu_C^2\right) \tag{16.3.77}$$

where the first term is the power supplied by the generator to the circuit, the second term is the Joule effect term and the third is the power stored in the capacitor.

The energy supplied by the generator is stored in the capacitor and dissipated by the resistance by Joule effect.

What is the most important is to make an energetic balance study over the whole duration of charge of the capacitor to indicate the power dissipated in the sales characteristics sheet (it is better for must buyer than putting equations in it...). To do this, it suffices to integrate the preceding relation from 0 to infinity to obtain the energy dissipated.

The first term to the left of the equality gives:

$$\begin{aligned}\int\limits_0^{+\infty} U_0 i\mathrm{d}t &= U_0\int\limits_0^{+\infty} i\mathrm{d}t = U_0\int\limits_0^{+\infty} C\frac{\mathrm{d}u_C}{\mathrm{d}t} = U_0 C\int\limits_0^{+\infty}\mathrm{d}u_C\\ &= U_0 C\left(u_0(+\infty) - u_0(0)\right) = U_0 C(U_0 - 0) = CU_0^2\end{aligned} \tag{16.3.78}$$

The second term is integrated using $i(t)$:

$$\begin{aligned}\int\limits_0^{+\infty} Ri^2\mathrm{d}t &= \int\limits_0^{+\infty} R\left(\frac{Q_0}{RC}e^{-t/RC}\right)^2\mathrm{d}t = \int\limits_0^{+\infty} R\left(\frac{1}{R^2}\left(\frac{Q_0}{C}\right)^2 e^{-2t/RC}\right)\mathrm{d}t = \frac{U_0^2}{R}\int\limits_0^{+\infty} e^{-2t/RC}\mathrm{d}t\\ &= \frac{U_0^2}{R}\left(-\frac{RC}{2}e^{-2t/RC}\right)\Bigg|_0^{+\infty} = -\frac{U_0^2 C}{2}(0-1) = \frac{1}{2}U_0^2 C\end{aligned} \tag{16.3.79}$$

The third term integrates immediately since we already have the primitive:

$$\int\limits_0^{+\infty}\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}Cu_C^2\right)\mathrm{d}t = \int\limits_0^{+\infty}\mathrm{d}\left(\frac{1}{2}Cu_C^2\right) = \frac{1}{2}CU_0 \tag{16.3.80}$$

Finally, we get:

$$CU_0^2 = \frac{1}{2}U_0^2 C + \frac{1}{2}U_0^2 C \tag{16.3.81}$$

Thus, for long durations, half of the energy supplied by the generator is dissipated by the Joule effect and the other stored in the capacitor.

16.3.4.1.2 RC series linear response

Let us study now the answer of the circuit to a linear increase (or decrease by symmetry) of the RC series circuit.

In our circuit above, the loop equation remains:

$$u_S = u_R + u_C \tag{16.3.82}$$

where u_S is the dynamic voltage of the source.

Explicitly:

$$u_S = iR + \frac{1}{C}\int i \mathrm{d}t \tag{16.3.83}$$

Differentiating:

$$\frac{\mathrm{d}u_S}{\mathrm{d}t} = R\frac{\mathrm{d}i}{\mathrm{d}t} + \frac{i}{C} \tag{16.3.84}$$

Rearranging:

$$\frac{\mathrm{d}i}{\mathrm{d}t} + \frac{i}{RC} = \frac{1}{R}\frac{\mathrm{d}u_S}{\mathrm{d}t} \tag{16.3.85}$$

Noting that $\tau = RC$:

$$\frac{\mathrm{d}i}{\mathrm{d}t} + \frac{i}{\tau} = \frac{1}{R}\frac{\mathrm{d}u_S}{\mathrm{d}t} \tag{16.3.86}$$

In your special case, u_S is a ramp, thus: $u_S = Kt$, where K is the slope of the ramp. Hence:

$$\frac{\mathrm{d}u_S}{\mathrm{d}t} = K \tag{16.3.87}$$

and the equation to be solved by the integrating factor method (see section Differential and Integral Calculus page 1615) is:

$$\frac{\mathrm{d}i}{\mathrm{d}t} + \frac{i}{\tau} = \frac{K}{R} \tag{16.3.88}$$

The integrating factor is:

$$\mathrm{IF} = e^{\int \frac{1}{\tau}\mathrm{d}t} = e^{\frac{t}{\tau}} \tag{16.3.89}$$

Therefore:

$$ie^{\frac{t}{\tau}} = \int \frac{K}{R}e^{\frac{t}{\tau}}\mathrm{d}t + c^{te} \tag{16.3.90}$$

Hence:

$$ie^{\frac{t}{\tau}} = KCe^{\frac{t}{\tau}} + c^{te} \tag{16.3.91}$$

Rearranging:

$$i = KC + c^{te} e^{-\frac{t}{\tau}} \tag{16.3.92}$$

Assuming initial conditions are zero (linear increasing tension source from zero), $c^{te} = -KC$, hence:

$$i = KC\left(1 - e^{-\frac{t}{\tau}}\right) \tag{16.3.93}$$

As for a capacitor:

$$i = C\frac{\mathrm{d}u_C}{\mathrm{d}t} \Rightarrow u_C = \frac{1}{C}\int i\mathrm{d}t + c^{te} \tag{16.3.94}$$

Hence:

$$u_C = \frac{1}{C}\int KC\left(1 - e^{-\frac{t}{\tau}}\right)\mathrm{d}t + c^{te} = K\left(t + \tau e^{-\frac{t}{\tau}}\right) + c^{te} \tag{16.3.95}$$

The constant has to be determined by the initial conditions. In our case, as u_C is zero at $t = 0$ then:

$$u_C = K\left(t - \tau + \tau e^{-\frac{t}{\tau}}\right) \tag{16.3.96}$$

The same reasoning for a linear decrease of the voltage leads to:

$$u_C = K\left(t - \tau e^{-\frac{t}{\tau}}\right) \tag{16.3.97}$$

16.3.4.2 RL series circuit

Let us consider now the following circuit:

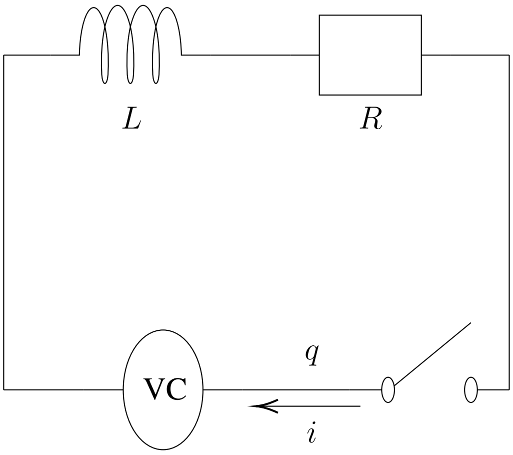


Figure 16.113: **RL circuit serie**

When we close the switch, we then have at the terminals of the resistor:

$$u_R = R \cdot i \tag{16.3.98}$$

and at the terminals of the inductance:

$$u_L = L\frac{\mathrm{d}i}{\mathrm{d}t} \tag{16.3.99}$$

and U_0 at the terminals of the DC voltage generator.

The equation of the circuit is then:

$$L\frac{\mathrm{d}i}{\mathrm{d}t} + Ri = U_0 \tag{16.3.100}$$

Therefore:

$$\frac{L}{\mathrm{d}t} = \frac{U_0 - Ri}{\mathrm{d}i} \tag{16.3.101}$$

By reversing:

$$\frac{1}{L}\mathrm{d}t = \frac{1}{U_0 - Ri}\mathrm{d}i \tag{16.3.102}$$

Let us do a change of variable:

$$U_0 - Ri = x \Rightarrow \frac{\mathrm{d}x}{\mathrm{d}i} = -R \tag{16.3.103}$$

Then:

$$\int\limits_0^t \frac{1}{L}\mathrm{d}t = \frac{1}{R}\int\limits_{U_0}^{U_0-RI} \frac{1}{x}\mathrm{d}x \tag{16.3.104}$$

It then comes after integration:

$$\frac{1}{L}t = -\frac{1}{R}\ln(x)\Big|_{U_0}^{U_0-RI} = -\frac{1}{R}\left[\ln(U_0 - Ri) - \ln(U_0)\right] = -\frac{1}{R}\ln\left(\frac{U_0 - Ri}{U_0}\right) \tag{16.3.105}$$

Let us multiply the two members by $-R$, then let us take the exponential of the two members:

$$-\frac{R}{L}t = \ln\left(\frac{U_0 - Ri}{U_0}\right) \tag{16.3.106}$$

and:

$$\frac{U_0 - Ri}{U_0} = e^{-\frac{R}{L}t} \tag{16.3.107}$$

Then:

$$i = \frac{U_0}{R}\left(1 - e^{-\frac{R}{L}t}\right) = \frac{U_0}{R}\left(1 - e^{-t/\tau}\right) \tag{16.3.108}$$

Where we have the time constant defined by:

$$\tau = \frac{L}{R} \tag{16.3.109}$$

Thus, when the switch is closed, the current increases exponentially with an asymptote at the value U_0/R. Thus, unlike the RC circuit, the current tends to a non-zero fixed value when t tends to infinity!

We therefore have:

$$u_L(t)=L\frac{\mathrm{d}}{\mathrm{d}t}\frac{U_0}{R}\left(1-e^{t/\tau}\right)=\frac{LU_0}{R}\frac{\mathrm{d}}{\mathrm{d}t}\left(1-e^{t/\tau}\right)=\frac{LU_0}{R}\frac{\mathrm{d}}{\mathrm{d}t}\left(-e^{-\frac{R}{L}t}\right)=U_0e^{-\frac{R}{L}t} \tag{16.3.110}$$

So in the end we have the two relations:

$$\boxed{\begin{aligned}&i=\frac{U_0}{R}\left(1-e^{-t/\tau}\right)\\&u_L=U_0e^{-\frac{R}{L}t}\end{aligned}} \tag{16.3.111}$$

Let us now study the energy aspect that is important in engineering, since power consumption or loss of power is a major selling factor in some applications! As for the RC circuit, we will directly make a energetic balance study of the entire duration of the transient regime to signal the power dissipated in the sales characteristics (it passes better than to put equations in it...):

$$L\frac{\mathrm{d}i}{\mathrm{d}t}+Ri=U_0 \tag{16.3.112}$$

Let us multiply the terms of the differential equation by $i(t)$:

$$L\frac{\mathrm{d}i}{\mathrm{d}t}i+Ri^2=\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left(Li^2\right)+Ri^2=U_0i \tag{16.3.113}$$

What we will write:

$$p_L(t)+p_R(t)=p_G(t) \tag{16.3.114}$$

To calculate the dissipated energy, we proceed in the same way as for the series RC circuit. We have after integration:

$$\frac{1}{2}L\left(i^2(+\infty)-i^2(0)\right)+\int_0^{+\infty}Ri^2\mathrm{d}t=\int_0^{+\infty}U_0i\mathrm{d}t \tag{16.3.115}$$

Therefore:

$$\begin{aligned}&\frac{1}{2}Li^2+R\int_0^{+\infty}\left(\frac{U_0}{R}\left(1-e^{-t/\tau}\right)\right)^2\mathrm{d}t=U_0\int_0^{+\infty}\left(\frac{U_0}{R}\left(1-e^{-t/\tau}\right)\right)\mathrm{d}t\\&\frac{1}{2}Li^2+\frac{U_0^2}{R}\int_0^{+\infty}\left(1-e^{-\frac{R}{L}t}\right)^2\mathrm{d}t=\frac{U_0^2}{R}\int_0^{+\infty}\left(1-e^{-\frac{R}{L}t}\right)\mathrm{d}t\end{aligned} \tag{16.3.116}$$

Unlike the case of the RC circuit, we can not integrate above with the given terminals because of the "$1-$" which is in front of the exponential because it causes the consumed power to reach the infinite, which is logical, unlike the RC circuit which ends up blocking itself after the capacitor has been charged (the current i tending towards zero very quickly).

Thus, we only integrate up to a sufficiently large limit time with respect to the values of the passive elements (two or three τ), or we are interested purely in terms of the instantaneous value of the power. We have then:

$$\begin{aligned}
p_G(t) &= U_0\cdot i = U_0\cdot\frac{U_0}{R}\left(1-e^{-t/\tau}\right)=\frac{U_0^2}{R}\left(1-e^{-t/\tau}\right)\\
p_R(t) &= R\cdot i^2 = R\cdot\frac{U_0^2}{R^2}\left(1-e^{-t/\tau}\right)^2=\frac{U_0^2}{R}\left(1-e^{-t/\tau}\right)^2\\
p_L(t) &= u_L(t)\cdot i = U_0 e^{-\frac{R}{L}t}\frac{U_0}{R}\left(1-e^{-t/\tau}\right)=\frac{U_0^2}{R}e^{-t/\tau}\left(1-e^{-t/\tau}\right)
\end{aligned} \tag{16.3.117}$$

And therefore at the end of the transitional regime when $t\rightarrow+\infty$:

$$\begin{aligned}
p_G(t) &= \frac{U_0^2}{R}\\
p_R(t) &= \frac{U_0^2}{R}\\
p_L(t) &= 0
\end{aligned} \tag{16.3.118}$$

Therefore, in a steady state, the resistance is the only energy dissipative element in the circuit and it is sufficient to multiply the power dissipated by the desired time interval in order to have an estimate of the energy dissipated.

16.3.4.3 Source-Free serial RLC circuit

An electrical wire (an antenna for example) is not a perfect conductor. In reality it can be assimilated to a resistance, a capacitance and an internal inductance in series. If we take the case, for example, of the generators, we often consider only the internal resistance as non-negligible, and this obviously reduces the nominal voltage of the generator by a factor to a first approximation proportional to the current flowing through it.

To study the behaviour of such an element often named "RLC circuit", we first represent it in the following form:

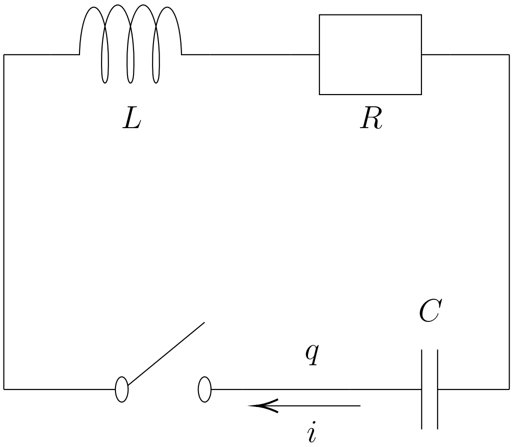


Figure 16.114: **RLC serie circuit**

We assume that initially the capacitor is charged and that no current flows (open switch):

$$q(t=0)=q_0 \quad \text{and} \quad i(t=0)=0 \tag{16.3.119}$$

When we close the switch the electrons go away from the capacitor C. We then have at the terminals of the resistor:

$$u_R=Ri=R\frac{\mathrm{d}q}{\mathrm{d}t} \tag{16.3.120}$$

and at the terminals of the condensator:

$$u_C=\frac{1}{C}q \tag{16.3.121}$$

and at the terminal of the inductance:

$$u_L=L\frac{\mathrm{d}i}{\mathrm{d}t}=L\frac{\mathrm{d}^2q}{\mathrm{d}t} \tag{16.3.122}$$

The sum of the potential differences of the circuit is equal to the initial difference of potential, hence:

$$L\frac{\mathrm{d}^2q}{\mathrm{d}t}+R\frac{\mathrm{d}q}{\mathrm{d}t}+\frac{1}{C}q=0 \tag{16.3.123}$$

Or written differently:

$$\boxed{LC\frac{\mathrm{d}^2q}{\mathrm{d}t}+RC\frac{\mathrm{d}q}{\mathrm{d}t}+q=0} \tag{16.3.124}$$

It is a second-order linear differential equation very well known in physics (we find it identically in other domains with just different constants). To solve it, we must look for the roots of the associated characteristic equation (see section Differential and Integral Calculus page 1608):

$$LCx^2+RCx+1=0 \tag{16.3.125}$$

The latter has for discriminant:

$$\Delta=R^2C^2-4LC \tag{16.3.126}$$

The resistance value for which this discriminant is zero is named the "critical resistance":

$$R_C=2\sqrt{\frac{L}{C}} \tag{16.3.127}$$

We can also write the discriminant in the following form:

$$\Delta=C^2\left(R^2-4\frac{L}{C}\right)=C^2\left(R^2-R_C^2\right) \tag{16.3.128}$$

The solutions of the differential equation are different according to the number and type of the roots of the characteristic equation.

16.3.4.3.1 Critically damped response ($R^4=4L/C$)

This is the case where $R=R_C$. The characteristic equation then admits a real double root since:

$$\Delta=C^2\left(R^2-4\frac{L}{C}\right)=C^2\left(R^2-R_C^2\right)=0 \tag{16.3.129}$$

We then have (so the roots are negative, real and equal as stated just earlier):

$$x_{1,2}=\frac{-RC\pm\sqrt{\Delta}}{2LC}=\frac{-R_C}{2L}=-\frac{1}{\sqrt{LC}}=-\frac{1}{\tau}=K \tag{16.3.130}$$

with:

$$\tau=\sqrt{LC} \tag{16.3.131}$$

The differential equation then admits a solution of the following type when the discriminant is zero (see section Differential and Integral Calculus page 1609):

$$q=(C_1t+C_2)\,e^{-\frac{t}{\sqrt{LC}}} \tag{16.3.132}$$

by omitting the delay.

Which gives for the intensity:

$$i=\frac{\mathrm{d}q}{\mathrm{d}t}=\left(C_1-C_1\frac{t}{\sqrt{LC}}-\frac{C_2}{\sqrt{LC}}\right)e^{-\frac{t}{\sqrt{LC}}} \tag{16.3.133}$$

The constants are defined by the initial conditions:

$$\begin{cases} q(t=0)=C_2=q_0 \\ i(t=0)=\left(C_1-\dfrac{C_2}{\sqrt{LC}}\right)=0\Rightarrow C_1=\dfrac{C_2}{\sqrt{LC}}=\dfrac{q_0}{\sqrt{LC}} \end{cases} \tag{16.3.134}$$

We thus obtain for the global solution:

$$q=q_0\left(1+\frac{t}{\tau}\right)e^{-t/\tau} \tag{16.3.135}$$

So in the end, we have the relations:

$$\boxed{\begin{aligned} q&=q_0\left(1+\frac{t}{\tau}\right)e^{-t/\tau} \\ i&=\frac{\mathrm{d}q}{\mathrm{d}t}=-\frac{q_0t}{\tau^2}e^{-t/\tau} \\ u_C&=\frac{q}{C}=\frac{q_0}{C}\left(1+\frac{t}{\tau}\right)e^{-t/\tau} \\ u_L&=L\frac{\mathrm{d}i}{\mathrm{d}t}=q_0L\left(\frac{t}{\tau^3}-\frac{1}{\tau^2}\right)e^{-t/\tau} \end{aligned}} \tag{16.3.136}$$

The following figures illustrate the trend of the time evolution of the discharge of the capacitor and of the current through the inductance. The intensity is maximum for $t=\tau$ in the inductance (we highlighted in red the current plot because it's the most commonly showed in textbooks):

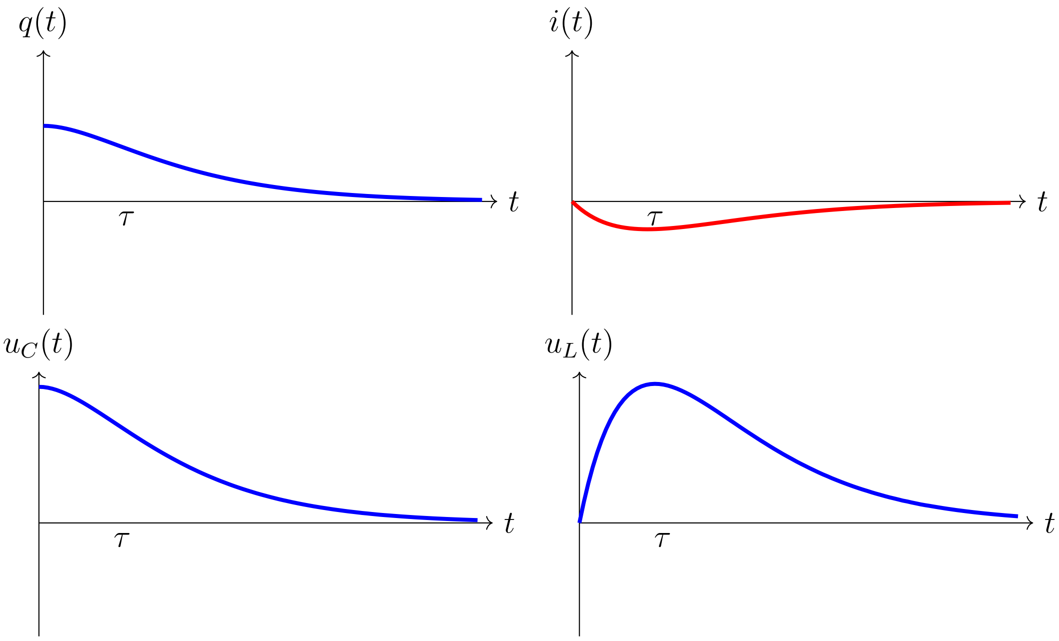


Figure 16.115: Behaviour of the charge, current and voltage in the circuit during the discharge of the capacitor in a critically damped serial RLC circuit

16.3.4.3.2 Overdamped (hypercritic) response ($R^2>4L/C$)

This is the case where $R>R_C$. The characteristic equation then admits two distinct real roots:

$$\Delta=C^2\left(R^2-4\frac{L}{C}\right)>0 \tag{16.3.137}$$

Therefore:

$$\begin{aligned}K_{1,2}&=\frac{-RC\pm\sqrt{C^2\left(R^2-4\frac{L}{C}\right)}}{2LC}=\frac{-RC\pm C\sqrt{R^2-4\frac{L}{C}}}{2LC}=\frac{-R\pm\sqrt{R^2-4\frac{L}{C}}}{2L}\\&=\frac{-R\pm\sqrt{R^2-R_C^2}}{2L}\end{aligned} \tag{16.3.138}$$

The two roots are of the same sign, because by using the Vieta relations (see section Calculus page 1383) we have:

$$K_1K_2=\frac{1}{LC}>0 \tag{16.3.139}$$

The two roots are therefore necessarily negative (real, distinct and negative)! We denote their absolute values:

$$|K_{1,2}|=\frac{R\pm\sqrt{(R^2-R_C^2)}}{2L} \tag{16.3.140}$$

that therefore satisfy:

$$K_1K_2 = \frac{1}{LC} = \frac{1}{\tau^2} \tag{16.3.141}$$

We have seen in the section of Differential and Integral Calculus (page 1609) that at this moment the solution (without phase shift) is of the form:

$$q = Ae^{K_1t} + Be^{K_2t} \tag{16.3.142}$$

Which gives for intensity:

$$i = AK_1e^{K_1t} + BK_2e^{K_2t} \tag{16.3.143}$$

The constants A and B are defined by the initial conditions:

$$\begin{cases} q(t=0) = A + B = q_0 \\ i(t=0) = K_1A + K_2B = 0 \end{cases} \tag{16.3.144}$$

Which gives us:

$$\begin{gathered} A - \frac{K_1}{K_2}A = q_0 \Rightarrow A\left(1 - \frac{K_1}{K_2}\right) = q_0 \Rightarrow A = \frac{K_2q_0}{K_2 - K_1} \\ -\frac{K_2}{K_1}B + B = q_0 \Rightarrow B\left(1 - \frac{K_2}{K_1}\right) = q_0 \Rightarrow B = \frac{K_1q_0}{K_1 - K_2} \end{gathered} \tag{16.3.145}$$

Either in conventional form:

$$A = -\frac{K_2q_0}{K_1 - K_2} \qquad B = \frac{K_1q_0}{K_1 - K_2} \tag{16.3.146}$$

Therefore:

$$q = \frac{q_0}{K_1 - K_2}\left(K_1e^{K_2t} - K_2e^{K_1t}\right) \tag{16.3.147}$$

Hence the set of relations:

$$\boxed{\begin{aligned} q &= \frac{q_0}{K_1 - K_2}\left(K_1e^{K_2t} - K_2e^{K_1t}\right) \\ i &= \frac{\mathrm{d}q}{\mathrm{d}t} = \frac{K_1K_2q_0}{K_1 - K_2}\left(e^{K_2t} - e^{K_1t}\right) \\ u_C &= \frac{q}{C} = \frac{q_0}{C(K_1 - K_2)}\left(K_1e^{K_2t} - K_2e^{K_1t}\right) \\ u_L &= L\frac{\mathrm{d}i}{\mathrm{d}t} = \frac{LK_1K_2q_0}{(K_1 - K_2)}\left(K_2e^{K_2t} - K_1e^{K_1t}\right) \end{aligned}} \tag{16.3.148}$$

The following figures (we highlighted in red the current plot because it's the most commonly showed in textbooks) illustrate the temporal evolution of these functions (remember that the roots are negative!):

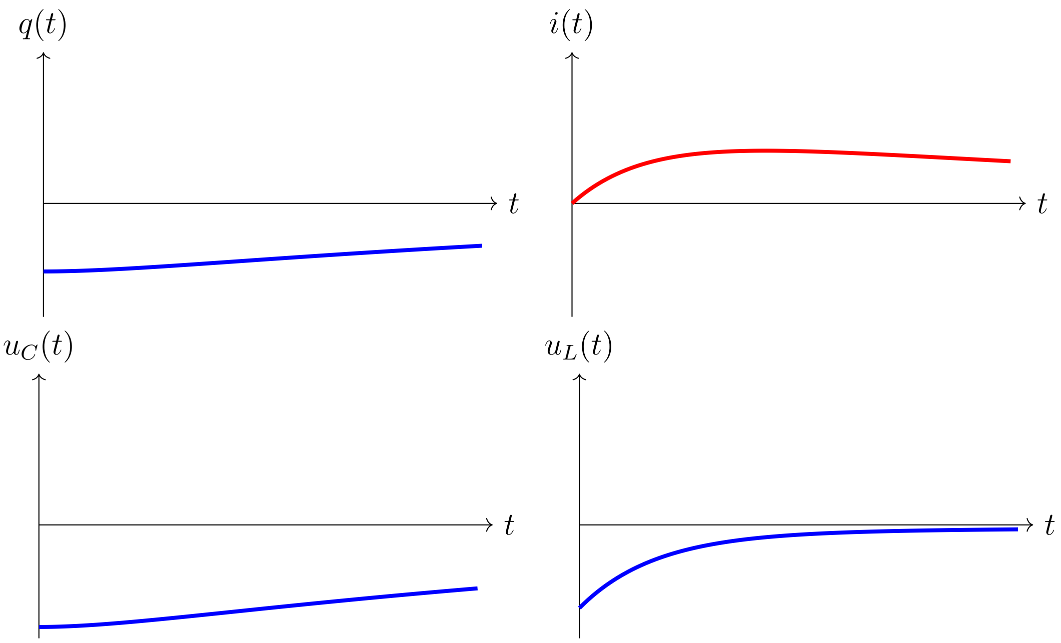


Figure 16.116: **Behaviour of the charge, current and voltage in the circuit during the discharge of the capacitor in a overdamped serial RLC circuit**

16.3.4.3.3 Underdamped (decaying oscillation) response ($R^2 < 4L/C$)

This is the case where $R < R_C$. The characteristic equation then admits two conjugate complex roots:

$$K_{1,2}=\frac{-R\pm\sqrt{R^2-R_C^2}}{2L}=\frac{-R\pm\sqrt{-\left(R_C^2-R^2\right)}}{2L}=\frac{-R\pm\sqrt{\mathrm{i}^2\left(R_C^2-R^2\right)}}{2L}=\frac{-R\pm\mathrm{i}\sqrt{R_C^2-R^2}}{2L} \tag{16.3.149}$$

Which are assimilated to the resistance of the circuit. We name it "complex impedance".

We will see that contrary to the intuition of that time they were invented (as often in pure mathematics) complex roots have a real physical meaning.

Let us denote for this α and ω the absolute values of the real and imaginary parts of these roots:

$$K_{1,2}=-\alpha\pm\mathrm{i}\omega \tag{16.3.150}$$

with:

$$\alpha=K_{1,2}=\frac{R}{2L} \tag{16.3.151}$$

and:

$$\omega=\frac{\sqrt{R_C^2-R^2}}{2L}=\sqrt{\frac{R_C^2}{4L}-\frac{R^2}{4L}}=\sqrt{\frac{1}{LC}-\alpha^2}=\sqrt{\frac{1}{\tau^2}-\alpha^2} \tag{16.3.152}$$

We have seen in the section of Differential and Integral Calculus that the solution of the differential equation is then written (see page 1610):

$$q=Ae^{K_1t}+Be^{K_2t}=e^{\alpha x}\left(A'\cos(\omega t)+B'\cos(\omega t)\right)=C'e^{-\alpha t}\cos(\omega t-\phi) \tag{16.3.153}$$

What gives us for the intensity:

$$\begin{aligned}i&=\frac{\mathrm{d}}{\mathrm{d}t}C'e^{-\alpha t}\cos(\omega t-\phi)=-\alpha C'e^{-\alpha t}\cos(\omega t-\phi)-C'e^{-\alpha t}\sin(\omega t-\phi)\\&=-C'e^{-\alpha t}\left[\alpha\cos(\omega t-\phi)+\omega\sin(\omega t-\phi)\right]\end{aligned}\tag{16.3.154}$$

The constants C' and ϕ are determined by the initial conditions:

$$\begin{aligned}q(t=0)&=C'\cos(\phi)=q_0\\i(t=0)&=-C'\left[\alpha\cos(-\phi)+\omega\sin(-\phi)\right]=0\end{aligned}\tag{16.3.155}$$

Which gives us:

$$C'=\frac{q_0}{\cos(\phi)}\qquad\text{and}\qquad\tan(-\phi)=\frac{\alpha}{\omega}\tag{16.3.156}$$

Therefore:

$$C'=\frac{q_0}{\cos(\phi)}\qquad\text{and}\qquad\tan(\phi)=\frac{\sin(\phi)}{\cos(\phi)}=-\frac{\alpha}{\omega}\tag{16.3.157}$$

Thus by reporting in the expressions of the charge q and of the current i:

$$q=\frac{q_0}{\cos(\phi)}e^{-\alpha t}\cos(\omega t-\phi)\tag{16.3.158}$$

and:

$$\begin{aligned}i&=-\frac{q_0}{\cos(\phi)}e^{-\alpha t}\left[\alpha\cos(\omega t-\phi)+\omega\sin(\omega t-\phi)\right]\\&=-\frac{q_0}{\cos(\phi)}e^{-\alpha t}\sqrt{\alpha^2+\omega^2}\left[\frac{\alpha}{\sqrt{\alpha^2+\omega^2}}\cos(\omega t-\phi)+\frac{\omega}{\sqrt{\alpha^2+\omega^2}}\sin(\omega t-\phi)\right]\end{aligned}\tag{16.3.159}$$

Let us try to simplify this last equality a little. First we have proved just before:

$$\alpha^2=\omega^2\tan^2(\phi)\tag{16.3.160}$$

Therefore:

$$\begin{aligned}\sqrt{\alpha^2+\omega^2}&=\sqrt{\omega^2\tan^2(\phi)+\omega^2}=\omega\sqrt{\tan^2(\phi)+1}=\omega\sqrt{\frac{\sin^2(\phi)}{\cos^2(\phi)}+1}\\&=\frac{\omega}{\cos\phi}\sqrt{\sin^2(\phi)+\cos^2(\phi)}=\frac{\omega}{\cos(\phi)}\end{aligned}\tag{16.3.161}$$

Which gives:

$$\begin{aligned}i&=-\frac{q_0}{\cos(\phi)}e^{-\alpha t}\left[\alpha\cos(\omega t-\phi)+\omega\sin(\omega t-\phi)\right]\\&=-\frac{\omega q_0}{\cos^2(\phi)}e^{-\alpha t}\left[\frac{\alpha}{\sqrt{\alpha^2+\omega^2}}\cos(\omega t-\phi)+\frac{\omega}{\sqrt{\alpha^2+\omega^2}}\sin(\omega t-\phi)\right]\end{aligned}\tag{16.3.162}$$

But, we also have:

$$\omega\frac{1}{\sqrt{\alpha^2+\omega^2}}=\omega\frac{\cos(\phi)}{\omega}=\cos(\phi)\tag{16.3.163}$$

and:

$$\alpha\frac{1}{\sqrt{\alpha^2+\omega^2}}=\alpha\frac{\cos(\phi)}{\omega}=\cos(\phi)\frac{\alpha}{\omega}=-\cos(\phi)\frac{\sin(\phi)}{\cos(\phi)}=-\sin(\phi) \tag{16.3.164}$$

We then have:

$$\begin{aligned}i&=-\frac{\omega q_0}{\cos^2(\phi)}e^{-\alpha t}\left[-\sin(\phi)\cos(\omega t-\phi)+\cos(\phi)\sin(\omega t-\phi)\right]\\&=\frac{\omega q_0}{\cos^2(\phi)}e^{-\alpha t}\left[\sin(\phi)\cos(\omega t-\phi)-\cos(\phi)\sin(\omega t-\phi)\right]\\&=\frac{\omega q_0}{\cos^2(\phi)}e^{-\alpha t}\sin(-\omega t)=-\frac{\omega q_0}{\cos^2(\phi)}e^{-\alpha t}\sin(\omega t)\end{aligned} \tag{16.3.165}$$

We have in the end the following set of relations:

$$\boxed{\begin{aligned}q&=\frac{q_0}{\cos(\phi)}e^{-\frac{R}{2L}t}\cos(\omega t-\phi)\\i&=\frac{\mathrm{d}q}{\mathrm{d}t}=-\frac{\omega q_0}{\cos^2(\phi)}e^{-\frac{R}{2L}t}\sin(\omega t)\\u_C&=\frac{q}{C}=\frac{q_0}{C\cos(\phi)}e^{-\frac{R}{2L}t}\cos(\omega t-\phi)\\u_L&=L\frac{\mathrm{d}i}{\mathrm{d}t}=-\frac{1}{2}\frac{\omega q_0e^{\left(-\frac{R}{2L}t\right)}(-R\sin(\omega t)+2\omega\cos(\omega t)L)}{\cos(\phi)^2L}\end{aligned}} \tag{16.3.166}$$

So a plot of the current i in the inductance and q of the capacitance will give:

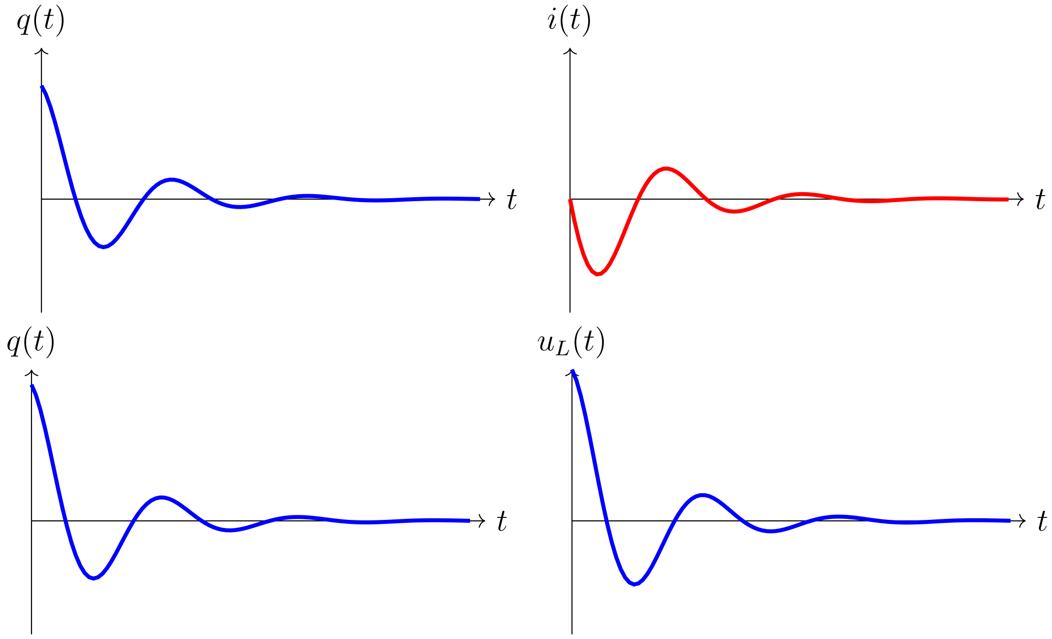


Figure 16.117: Behaviour of the charge, current and voltage in the circuit during the discharge of the capacitor in a underdamped serial RLC circuit

where:

$$\alpha=\frac{R}{2L} \tag{16.3.167}$$

is the "amortization factor" (some engineers defined instead $\gamma = R/L$ as the "damping factor"). If we wish to have beautiful oscillations smoothly and slowly damped, it is advantageous to have this term as small as possible and therefore have a small value of R.

When R is zero we then have:

$$i(t) = -\omega q_0 \sin(\omega t) \tag{16.3.168}$$

with therefore:

$$\omega = \frac{\sqrt{R_C^2 - R^2}}{2L} = \frac{R_C}{2L} = \frac{2\sqrt{\dfrac{L}{C}}}{2L} = \sqrt{\frac{1}{LC}} = \frac{1}{\sqrt{LC}} \tag{16.3.169}$$

which we name the "resonance pulsation" or "natural frequency". Either a period of:

$$T = \frac{2\pi}{\omega} = \frac{2\pi}{\sqrt{\dfrac{1}{LC}}} = 2\pi\sqrt{LC} \tag{16.3.170}$$

It is then necessary to play with C or L to obtain the desired period in the case where the resistance is zero. Note also that this particular situation is named a "harmonic oscillator".

Finally, from the results obtained, we thus have the generalization of the RC, RL or LC series circuits without source.

Now, suppose that in the circuit we were placing a continuous supply in series in the circuit. We have then:

$$L\frac{\mathrm{d}^2 q}{\mathrm{d}t^2} + R\frac{\mathrm{d}q}{\mathrm{d}t} + \frac{1}{C}q = U_0 \tag{16.3.171}$$

The linear differential equation with constant coefficients now has a second member (constant in this case). We then immediately find a special solution which it is then sufficient to add to all the solutions which we have obtained previously.

A special solution is therefore:

$$q = c^{te} \tag{16.3.172}$$

Hence:

$$L\frac{\mathrm{d}^2 q}{\mathrm{d}t^2} + R\frac{\mathrm{d}q}{\mathrm{d}t} + \frac{1}{C}q = L\cdot 0 + R\cdot 0 + \frac{1}{C}q = \frac{1}{C}q = U_0 \tag{16.3.173}$$

Therefore:

$$q = \frac{U_0}{C} \tag{16.3.174}$$

This special solution, which is to be added to the preceding solutions, has no influence on the equations of the current (its derivative being zero). On the other hand, it shifts the plot of $q(t)$ upwards. This is the effect of adding a constant voltage source (such as a simple battery).

So for summary to compare the three configurations (for the current only because its the most common plot in other textbooks):

$$\begin{aligned} i_c &= \frac{\mathrm{d}q}{\mathrm{d}t} = -\frac{q_0 t}{\tau^2} e^{-t/\tau} \\ i_o &= \frac{\mathrm{d}q}{\mathrm{d}t} = \frac{K_1 K_2 q_0}{K_1 - K_2} \left(e^{K_2 t} - e^{K_1 t}\right) \\ i_u &= \frac{\mathrm{d}q}{\mathrm{d}t} = -\frac{\omega q_0}{\cos^2(\phi)} e^{-\frac{R}{2L}t} \sin(\omega t) \end{aligned} \tag{16.3.175}$$

where again i_u, i_o, and i_c refer to underdamped, overdamped, and critically-damped cases (in many textbooks the reader can find variations of the graph below because some authors take different values of R, C, L and some even take the absolute value of i_c, i_o for pedagogical reasons).

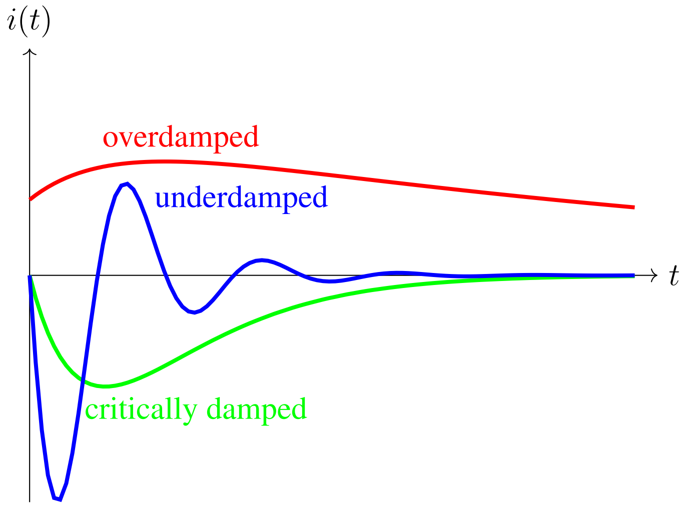


Figure 16.118: **Serial RLC without source typical modes of current**

Using the following definitions:

$$\begin{aligned} \omega_0 &= \frac{1}{\sqrt{LC}}, \qquad \gamma = \frac{R}{L} \\ \omega_1 &= \sqrt{\omega_0^2 - \gamma^2/4}, \qquad \alpha = \sqrt{\gamma^2/4 - \omega_0^2} \end{aligned} \tag{16.3.176}$$

and the physicist trick (variable change) seen in the section of Differential and Integral Calculus (page 1609), the three above relations can be found in some textbooks as following:

$$i = \begin{cases} i_u = -q_0 e^{-\gamma t/2} \left(\frac{\gamma^2 + 4\omega_1^2}{4\omega_1}\right) \sin\left(\omega_1 t\right), & \omega_0 > \gamma/2 \\ i_o = -\alpha q_0 \left(1 - \frac{\gamma^2}{4\alpha^2}\right) e^{-\gamma t/2} \sinh(\alpha t) & \omega_0 < \gamma/2 \\ i_c = -\frac{\gamma^2}{4} q_0 t e^{-\gamma t/2} & \omega_0 = \gamma/2 \end{cases} \tag{16.3.177}$$

Remark

It seems that the first practical use for RLC circuits was in the 11890s (holocene calendar) in spark-gap radio transmitters to allow the receiver to be tuned to the transmitter.

There are many other case of the RLC circuit that are studied in by electrical engineers during their curriculum like:

- Source-free serial RLC (see above)
- Source-free parallel RLC
- Alternate source serial RLC
- Alternate source parallel RLC
- Step response serial RLC
- Step response parallel RLC
- ...

The RLC is an important subject especially because of it's mechanical analogue that is a particle in simple harmonic motion as illustrated below:

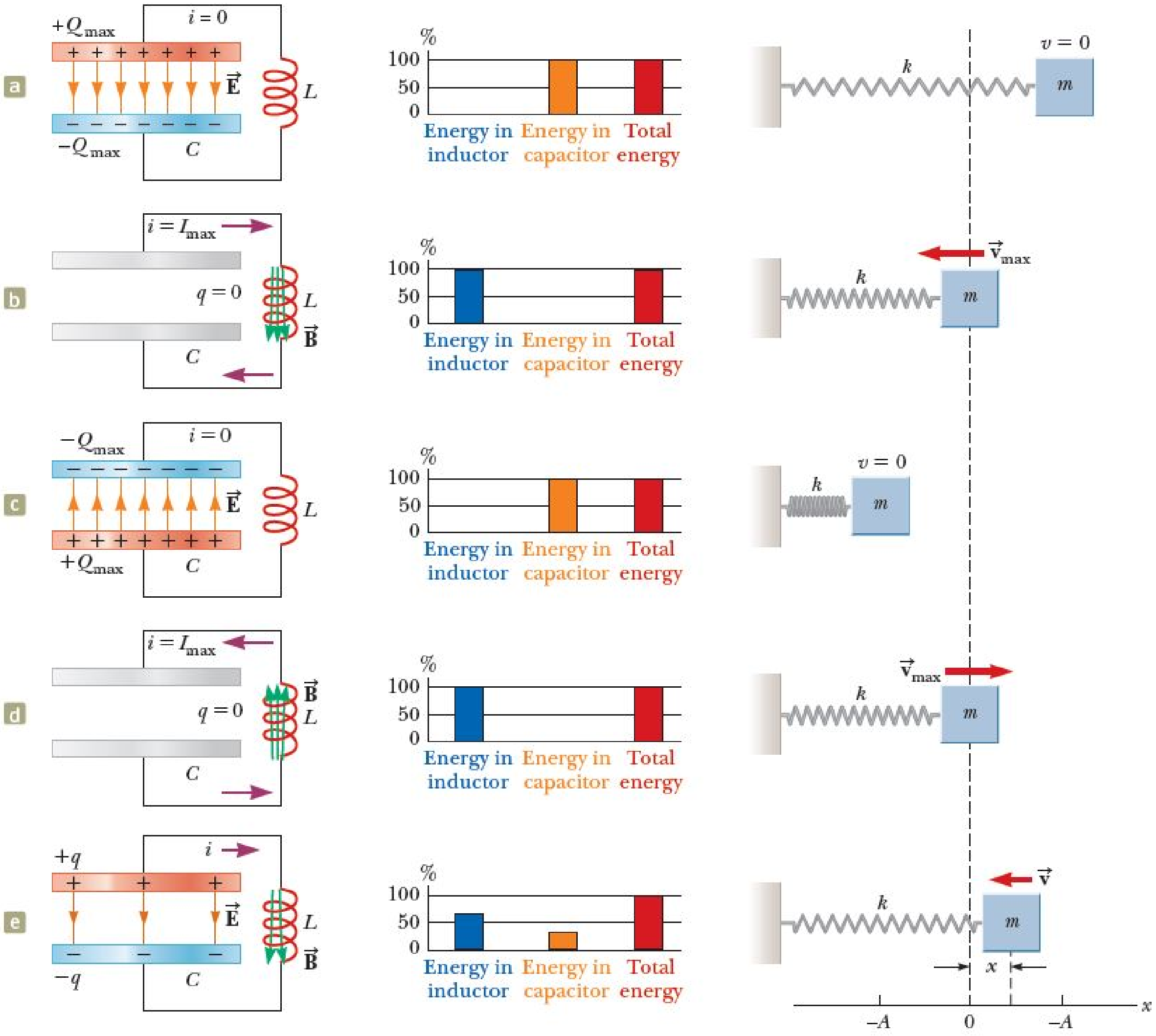

a
b
c
d
e
$+Q_{max}$
$-Q_{max}$
$i = 0$
$i = I_{max}$
$q = 0$
$+q$
$-q$
i
$\vec{E}$
$\vec{B}$
L
C
%
100
50
0
Energy in inductor
Energy in capacitor
Total energy
k
m
$v = 0$
$\vec{v}_{max}$
$\vec{v}$
x
–A
0
–A

Figure 16.119: **Mechanical analogue of RLC circuit (source: [491])**

And the corresponding equations:

RLC Circuit		**One-dimensional particle in damped harmonic motion**	
Charge	q	x	Position
Current	$i=\dfrac{\mathrm{d}q}{\mathrm{d}t}$	$v_x=\dfrac{\mathrm{d}x}{\mathrm{d}t}$	Velocity
Potential difference	ΔU	F_x	Force
Resistance	R	b	Viscous damping factor
Capacitance	C	$1/k$	Spring constant
Inductance	L	m	Mass
Second time derivative of charge	$\dfrac{\mathrm{d}i}{\mathrm{d}t}=\dfrac{\mathrm{d}^2q}{\mathrm{d}t}$	$a_x=\dfrac{\mathrm{d}v_x}{\mathrm{d}t}=\dfrac{\mathrm{d}^2x}{\mathrm{d}t}$	Acceleration
Energy in inductor	$E_L=\dfrac{1}{2}Li^2$	$E_c=\dfrac{1}{2}mv^2$	Kinetic energy of particle
Energy in capacitor	$E_C=\dfrac{1}{2}Cu^2$	$E_p=\dfrac{1}{2}kx^2$	Energy stored in spring
Rate of energy loss due to resistance	$P=Ri^2$	$P=bv^2$	Rate of energy loss due to friction
RLC Circuit ODE	$L\dfrac{\mathrm{d}^2q}{\mathrm{d}t^2}+R\dfrac{\mathrm{d}q}{\mathrm{d}t}+\dfrac{q}{C}=0$	$m\dfrac{\mathrm{d}^2x}{\mathrm{d}t^2}+b\dfrac{\mathrm{d}x}{\mathrm{d}t}+kx=0$	Particle in dampled harmonic motion ODE

Table 16.10: **Mechanical analogue of RLC circuit**

16.3.4.4 Telegrapher's Equations

These "Telegrapher's equations" are a pair of coupled, linear partial differential equations that describes an electrical signal travelling along a transmission cable. First derived in the horse and buggy days of the telegraph (hence the name), it is still useful for describing long distance power lines and cable TV systems.

The model demonstrates that the electromagnetic waves can be reflected on the wire, and that wave patterns can form along the line.

The above model did not include any internal resistance of the wires comprising the transmission lines, nor did it include the leakage of current out of the wire to the outside environment (collectively named "ground"). Leakage occurs due to a combination of factors: conductance G through the wire's insulation and capacitive coupling C between wire and ground. We want the conductance (which is measured in units that are the inverse of resistance) between the wire and the ground as low as possible – and we want this coupling C to be fairly small as well.

The telegrapher's equations, like all other equations describing electrical phenomena, result from Maxwell's equations. In a more practical approach, one assumes that the conductors are composed of an infinite series of two-port elementary components, each representing an infinitesimally short segment of the transmission line:

- The distributed resistance R of the conductors is represented by a series resistor (expressed in ohms per unit length). In practical conductors, at higher frequencies, R increases approximately proportional to the square root of frequency due to skin effect.
- The distributed inductance L (due to the magnetic field around the wires, self-inductance, etc.) is represented by a series inductor (henries per unit length).
- The capacitance C between the two conductors is represented by a shunt capacitor C (farads per unit length).
- The conductance G of the dielectric material separating the two conductors is represented

by a shunt resistor between the signal wire and the return wire (siemens per unit length). This resistor in the model has a resistance of $1/G$ ohms. G accounts for both bulk conductivity of the dielectric and dielectric loss. If the dielectric is an ideal vacuum, then $G=0$.

Notice that the traditional notation in the context of the Telegrapher equation may be misleading. Indeed, instead of writing R, L, C and G as in the figure below, we should instead write ρ_R, ρ_L, ρ_C and ρ_G.

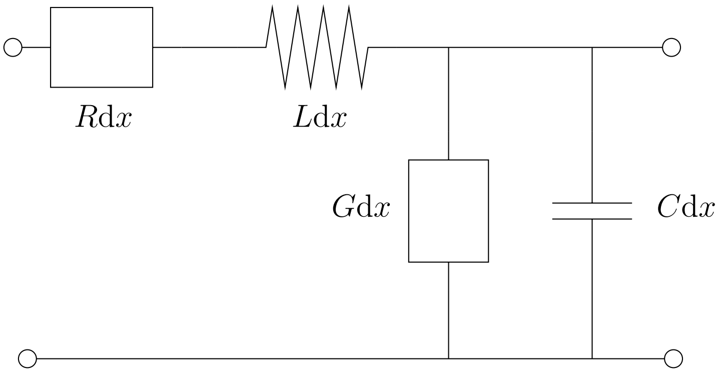


Figure 16.120: **Telegrapher scheme**

Consider a piece of wire being modelled as an electrical circuit element (see Figure above) consisting of an infinitesimal piece of (telegraph) wire of resistance $R\mathrm{d}x$ and inductance $L\mathrm{d}x$, while it is connected to a ground with conductance $(G\mathrm{d}x)^{-1}$ and capacitance $C\mathrm{d}x$.

Let $i(x,t)$ and $u(x,t)$ denote the current and voltage across the piece of wire at position x at time t.

The change in voltage across the piece of wire is given by:

$$u(x+\mathrm{d}x,t)-u(x,t)=-i(x,t)R\mathrm{d}x-\frac{\partial i}{\partial t}(x,t)L\mathrm{d}x \tag{16.3.178}$$

and the amount of current that disappears via the ground is:

$$i(x+\mathrm{d}x,t)-i(x,t)=-i(x,t)G\mathrm{d}x-\frac{\partial v}{\partial t}(x,t)C\mathrm{d}x \tag{16.3.179}$$

Dividing by $\mathrm{d}x$ gives the pair of coupled linear partial differential equations we were speaking about at the beginning:

$$\begin{aligned}\frac{\partial u}{\partial x}&=-Ri-L\frac{\partial i}{\partial t}\\ \frac{\partial i}{\partial x}&=-Gu-C\frac{\partial v}{\partial t}\end{aligned} \tag{16.3.180}$$

To combine these two relations, let us derivate partially once again each of them as following:

$$\begin{aligned}\frac{\partial^2 u}{\partial x^2}&=-R\frac{\partial i}{\partial x}-L\frac{\partial^2 i}{\partial x\partial t}\\ \frac{\partial^2 i}{\partial x\partial t}&=-G\frac{\partial v}{\partial t}-C\frac{\partial^2 v}{\partial t^2}\end{aligned} \tag{16.3.181}$$

If the term $\frac{\partial^2 i}{\partial x \partial t}$ is now substituted, we get:

$$\frac{\partial^2 u}{\partial x^2} = -R\frac{\partial i}{\partial x} - L\left(-G\frac{\partial u}{\partial t} - C\frac{\partial^2 e}{\partial t^2}\right) \tag{16.3.182}$$

and if $\frac{\partial i}{\partial x}$ is finally substituted the result is:

$$\frac{\partial^2 u}{\partial x^2} = LC\frac{\partial^2 u}{\partial t^2} + (RC+GL)\frac{\partial u}{\partial t} + (GR)u \tag{16.3.183}$$

Doing the same in the opposite way results in:

$$\frac{\partial^2 i}{\partial x^2} = LC\frac{\partial^2 i}{\partial t^2} + (RC+GL)\frac{\partial i}{\partial t} + (GR)i \tag{16.3.184}$$

The both relations:

$$\boxed{\begin{aligned}\frac{\partial^2 u}{\partial x^2} &= LC\frac{\partial^2 u}{\partial t^2} + (RC+GL)\frac{\partial u}{\partial t} + (GR)u\\ \frac{\partial^2 i}{\partial x^2} &= LC\frac{\partial^2 i}{\partial t^2} + (RC+GL)\frac{\partial i}{\partial t} + (GR)i\end{aligned}} \tag{16.3.185}$$

are named the Telegrapher's equations". As we can see, except for the dependent variable (i, u) the relations are identical.

If we define $c := 1/\sqrt{LC}, a := c^2(LG+RC), b := c^2RG$, then the Telegrapher equation in u becomes:

$$\frac{\partial^2 u}{\partial t^2} + a\frac{\partial u}{\partial t} + bu = c^2\frac{\partial^2 v}{\partial x^2} \tag{16.3.186}$$

For reasons we will explain below the $a\frac{\partial v}{\partial t}$ term is named the "dissipation term", and the and the bu term is the "dispersion term".

Of course, if $a = b = 0$, we are back to the vibrating string, i.e. wave equation, with its right and left moving wave solution representation.

In many situations we can neglect resistive losses. We then put $R = 0$ and $G = 0$. The equations are therefore reduced to (Alembert equations type, i.e. wave equation):

$$\begin{aligned}\frac{\partial^2 u}{\partial x^2} &= LC\frac{\partial^2 u}{\partial t^2}\\ \frac{\partial^2 i}{\partial x^2} &= LC\frac{\partial^2 i}{\partial t^2}\end{aligned} \tag{16.3.187}$$

we know very well how we may interpret the equation. It's the one of a wave, right and left moving wave solution at the speed:

$$\boxed{v = \frac{1}{\sqrt{LC}}} \tag{16.3.188}$$

This is the propagation speed of waves travelling through the transmission line. For transmission lines made of parallel perfect conductors with vacuum between them, this speed is equal to the speed of light c.

If we consider:

$$u(x,t)=U_0\cos(\omega t-kx)=U_0\cos\left(\omega t-\omega\frac{x}{c}\right) \tag{16.3.189}$$

Then we have obviously:

$$\frac{\partial u(x,t)}{\partial x}=U_0\frac{\omega}{c}\sin\left(\omega t-\omega\frac{x}{c}\right) \tag{16.3.190}$$

But for the telegrapher line, if $R=G=0$ we have just proved above that:

$$\frac{\partial u}{\partial x}=-L\frac{\partial i}{\partial t} \tag{16.3.191}$$

Therefore:

$$\frac{\partial i}{\partial t}=-\frac{U_0\omega}{Lc}\sin\left(\omega t-\omega\frac{x}{c}\right) \tag{16.3.192}$$

Hence:

$$i(x,t)=\frac{U_0}{Lc}\cos\left(\omega t-\omega\frac{x}{c}\right) \tag{16.3.193}$$

Therefore the characteristic impedance of the telegrapher line with $R=G=0$ is given:

$$Z_c=\frac{u(x,t)}{i(x,t)}=Lc=\sqrt{\frac{L}{C}} \tag{16.3.194}$$

Now let us recall that we have proved that for a coaxial capacitor (see section Electrostatics page 3997) that the coaxial capacity per unit length was given by:

$$C=2\pi\varepsilon\frac{1}{\ln\left(\dfrac{R_2}{R_1}\right)} \tag{16.3.195}$$

and similarly for a coaxial inductance (see section Electrostatics page 4266):

$$L=\frac{\mu}{2\pi}\ln\left(\frac{R_2}{R_1}\right) \tag{16.3.196}$$

Therefore:

$$Z_c=\sqrt{\frac{L}{C}}=\frac{1}{2\pi}\sqrt{\frac{\mu}{\varepsilon}}\ln\left(\frac{R_2}{R_1}\right) \tag{16.3.197}$$

And also notice that with $L=G=0$ we get for the potential equation[10]:

$$\frac{\partial^2 u}{\partial x^2}=RC\frac{\partial u}{\partial t} \tag{16.3.198}$$

We recognize here the heat equation (see section Thermodynamics page 3684)!

---

[10] At the slow signalling speeds of a human-operated telegraph key (a dozen or so words per minute), inductive effects were negligible in the Atlantic Cable, and conductance was pretty small, too (that is, the leakage resistance between the conductors was very large)

16.3.4.4.1 Thomson's law of squares of transmission lines (latency metallic cables)

The "Thomson's law of squares of transmission lines" is a theorem concerning transmission lines. It states that the current injected into the line by a step in voltage reaches a maximum at a time proportional to the square of the distance down the line. The theorem is due to William Thomson, the future Lord Kelvin. The law had some importance in connection with submarine telegraph cables.

The law of squares was proposed by William Thomson in 11854 (holocene calendar) at Glasgow University. He had some input from George Gabriel Stokes. Thomson and Stokes were interested in investigating the feasibility of the proposed transatlantic telegraph cable.

Thomson get through a quite awful way to the result we have just presented earlier above:

$$\frac{\partial^2 u}{\partial x^2}=RC\frac{\partial u}{\partial t} \tag{16.3.199}$$

and after he derived the propagation time law in a quite awful and old way.

Let us do a modern derivation of its result using Laplace transform (see section Analysis page 2336). Notice that the entire development below is inspired from the excellent book: *Transients for Electrical Engineers* [409].

We start by taking the unilateral Laplace transform (see section Analysis page 2358) of:

$$\frac{\partial^2 u}{\partial x^2}=LC\frac{\partial^2 u}{\partial t^2}+(RC+GL)\frac{\partial u}{\partial t}+(GR)u \tag{16.3.200}$$

That we will denote as following:

$$\mathcal{UL}\left(u(x,t)\right)=\int\limits_{0+}^{+\infty}v(x,t)e^{-st}\mathrm{d}t=U(x,s) \tag{16.3.201}$$

using the property of time derivation of the unilateral Laplace transformed proved in the section dedicated to it (see section Analysis page 2359), we also have:

$$\begin{aligned}\mathcal{UL}\left(\frac{\partial u}{\partial t}\right)&=sU(x,s)-u(x,0+)\\ \mathcal{UL}\left(\frac{\partial^2 u}{\partial t^2}\right)&=s^2U(x,s)-su(x,0+)-\left.\frac{\mathrm{d}u(x,t)}{\mathrm{d}t}\right|_{t=0+}\\ \mathcal{UL}\left(\frac{\partial^2 u}{\partial x^2}\right)&=\frac{\partial^2}{\partial x^2}[\mathcal{UL}\left(u(x,t)\right)]=\frac{\partial^2}{\partial x^2}U(x,s)=\frac{\mathrm{d}^2}{\mathrm{d}x^2}U(x,s)\end{aligned} \tag{16.3.202}$$

The difference between the last two lines - note which variable the derivatives are with respect to - should be carefully understood. Thus:

$$\frac{\partial^2 u}{\partial x^2}=LC\frac{\partial^2 u}{\partial t^2}+(RC+GL)\frac{\partial u}{\partial t}+(GR)u \tag{16.3.203}$$

Becomes:

$$\begin{aligned}\frac{\mathrm{d}^2U(x,s)}{\mathrm{d}x^2}=&LC\left[s^2U(x,s)-su(x,0+)-\left.\frac{\mathrm{d}u(x,t)}{\mathrm{d}t}\right|_{t=0+}\right]\\ &+(RC+LG)[sU(x,s)-u(x,0+)]+RGU(x,s)\end{aligned} \tag{16.3.204}$$

or, after a bit of rearranging:

$$\begin{aligned}\frac{\mathrm{d}^2U(x,s)}{\mathrm{d}x^2}-(Ls+R)(Cs+G)U(x,s)=-L\left[Gu(x,0+)+C\frac{\mathrm{d}u(x,t)}{\mathrm{d}t}\bigg|_{t=0+}\right]\\ -C[Ls+R]v(x,0+)\end{aligned} \tag{16.3.205}$$

At this point, there is a subtle technical issue that needs to be addressed before continuing. The value of:

$$\frac{\mathrm{d}v(x,t)}{\mathrm{d}t}\bigg|_{t=0+} \tag{16.3.206}$$

the initial rate of change of the voltage with respect to time, at an arbitrary location along the line, is usually not obvious. We can side-step this problem if we work instead with the initial rate of change with respect to x of the current, that is, with:

$$\frac{\mathrm{d}i(x,t)}{\mathrm{d}x}\bigg|_{t=0+} \tag{16.3.207}$$

(as we will elaborate on this claim in just a moment) and we can get that value from:

$$-\frac{\partial i}{\partial x}=Gu+C\frac{\partial u}{\partial t} \tag{16.3.208}$$

So, evaluating that latter at $t=0+$ (note, carefully, that we are simply substituting $t=0+$ and are not Laplace transforming), we have:

$$-\frac{\mathrm{d}i(x,t)}{\mathrm{d}x}\bigg|_{t=0+}=Gu(x,0+)+C\frac{\mathrm{d}u(x,t)}{\mathrm{d}t}\bigg|_{t=0+} \tag{16.3.209}$$

and so:

$$C\frac{\mathrm{d}u(x,t)}{\mathrm{d}t}\bigg|_{t=0+}=-\frac{\mathrm{d}i(x,t)}{\mathrm{d}x}\bigg|_{t=0+}-Gu(x,0+) \tag{16.3.210}$$

Putting this into the unilateral Laplace transform that we have get, we obtain:

$$\frac{\mathrm{d}^2U(x,s)}{\mathrm{d}x^2}-\gamma^2U(x,s)=L\frac{\mathrm{d}i(x,t)}{\mathrm{d}x}\bigg|_{t=0+}-C[Ls+R]u(x,0+) \tag{16.3.211}$$

where:

$$\gamma=\sqrt{(Ls+R)(Cs+G)} \tag{16.3.212}$$

The reason the above differential (Laplace) equation is preferable is that the initial value of $\mathrm{d}i(x,t)/\mathrm{d}x$ is, as we have claimed earlier, usually more easily tractable than the initial value of $\mathrm{d}u(x,t)/\mathrm{d}t$ To see this, consider the problem of calculating $u(x,t)$ on a transmission line (in response to a given input signal $u(0,t)$), a line that we say to be "initially relaxed" (that is, there is no current or stored charge in the line at $t=0$). We can immediately say, for such a line, that:

$$u(x,0+)=0 \text{ for } x>0 \tag{16.3.213}$$

and:

$$i(x,0+)=0 \text{ for } x>0 \tag{16.3.214}$$

As time increases from $t = 0+$, we obviously expect the cable current at a given location to change from zero. But suppose we keep time constant, and look at the cable current as we change x. As long as $x > 0$ we are not at the beginning of the cable then as we move along the cable with increasing x we will continue to see $i = 0$ until some time $t > 0+$. Thus, we can definitely say:

$$\left.\frac{\mathrm{d}i(x,t)}{\mathrm{d}x}\right|_{t=0+} = 0 \text{ for } x > 0 \tag{16.3.215}$$

Next, consider the case of the cable voltage $u(x,t)$. As time increases from $t = 0+$ we again obviously expect the cable voltage at a given point to change from zero. Suppose now that we keep x constant (we observe a fixed point on the cable) and ask how the voltage there changes with respect to time? That is, what is:

$$\left.\frac{\mathrm{d}u(x,t)}{\mathrm{d}t}\right|_{t=0+} = 0 \tag{16.3.216}$$

Also, knowing that $u(x,0+) = 0$ for any $x > 0$ doesn't tell us anything about how the voltage changes with time. It isn't the distinction between voltage and current that is important in this consideration, but rather which variable we are differentiating with respect to (x or t). Differentiating with respect to x, and with respect to t makes a big difference.

To complete this section, we'll calculate $u(x,t)$ and $i(x,t)$ for an initially relaxed, infinitely long[11] transmission line when a unit step voltage is applied to the $x = 0$ end at $t = 0$. The general problem (for arbitrary values of $R, L, C,$ and G) was first attempted by Heaviside, using his operational calculus, but his solution seems to have been later questioned (after his death) by others. We'll work our way towards the general solution using the Laplace transform, with our first discussion dealing with the Atlantic Cable.

So from:

$$\frac{\mathrm{d}^2U(x,s)}{\mathrm{d}x^2} - \gamma^2 U(x,s) = L\left.\frac{\mathrm{d}i(x,t)}{\mathrm{d}x}\right|_{t=0+} - C[Ls+R]u(x,0+) \tag{16.3.217}$$

we start to solve the homogeneous equation:

$$\frac{\mathrm{d}^2U(x,s)}{\mathrm{d}x^2} - \gamma^2 U(x,s) = 0 \tag{16.3.218}$$

with the following initial condition: $u(0,t) = u(t)$.

This is an ordinary differential equation that is easily solved. Specifically, assume a solution of the form:

$$U(x,s) = Ke^{px} \tag{16.3.219}$$

where K and p are some constants. Then putting it back in the homogeneous equation leads to:

$$Kp^2e^{px} - \gamma^2 K = e^{px} \tag{16.3.220}$$

---

[11] To account for a finite length is not conceptually difficult, but it does add considerably to the mathematics.

therefore:

$$p=\pm\gamma \tag{16.3.221}$$

Hence (keep in mind that $\gamma\in\mathbb{R}_+^*$) :

$$U(x,s)=K_1e^{\gamma x}+K_2e^{-\gamma x} \tag{16.3.222}$$

For an infinite line we demand that $K_1=0$ as otherwise $U(x,s)$ would become unbounded as $x\rightarrow+\infty$. So:

$$U(x,s)=K_2e^{-\gamma x} \tag{16.3.223}$$

Since at $x=0$ we have $u(0,t)=u(t)$ then using the unilateral Laplace transform (see section Analysis page 2358):

$$V(0,s)=\mathcal{UL}(u(t))=\frac{1}{s}=K_2 \tag{16.3.224}$$

and so:

$$U(x,s)=\frac{1}{s}e^{-\gamma x} \tag{16.3.225}$$

For the Atlantic cable, we have $L=G=0$, which means:

$$\gamma=\sqrt{RC} \tag{16.3.226}$$

And so the above solution gives:

$$U_{\text{At.}}(x,s)=\frac{1}{s}e^{-x\sqrt{RCs}} \tag{16.3.227}$$

Using the following Unilateral Laplace transform pair (see section Analysis page 2351):

$$1-\text{erf}\left(\frac{a}{2\sqrt{t}}\right)\leftrightarrow\frac{1}{s}e^{-a\sqrt{s}} \tag{16.3.228}$$

and so with:

$$a=x\sqrt{RC} \tag{16.3.229}$$

We see then that the unit step voltage response is:

$$u(x,t)=1-\text{erf}\left(\frac{x\sqrt{RC}}{2\sqrt{t}}\right) \tag{16.3.230}$$

To find the current $i(x,t)$ on the Atlantic cable, we look back at:

$$-\frac{\partial u}{\partial x}=iR+L\frac{\partial i}{\partial t} \tag{16.3.231}$$

we set $L=0$ and then we use the Laplace transform to get:

$$-\frac{\partial U(x,s)}{\partial x}=I(x,s)R \tag{16.3.232}$$

and so:

$$I(x,s)=-\frac{1}{R}\frac{\mathrm{d}U(x,s)}{\mathrm{d}x}=-\frac{1}{R}\left(\frac{-\sqrt{RCs}e^{-x\sqrt{RCs}}}{s}\right) \tag{16.3.233}$$

or:

$$I(x,s)=\sqrt{\frac{C}{Rs}}e^{-x\sqrt{RCs}} \tag{16.3.234}$$

To get $i(x,t)$ recall this pair property (s domain derivation) of the Laplace transform (see section Analysis page 2345):

$$tx(s)\leftrightarrow -\frac{\mathrm{d}}{\mathrm{d}s}X(x) \tag{16.3.235}$$

and as (see section Analysis page 2354):

$$\frac{e^{-a^2/4t}}{\sqrt{\pi t^3}}\leftrightarrow \frac{2}{a}e^{-a\sqrt{s}} \tag{16.3.236}$$

Using the s-domain derivation property we then get:

$$t\frac{e^{-a^2/4t}}{\sqrt{\pi t^3}}\leftrightarrow -\frac{2}{a}\left(-a\frac{1}{2}s^{-1/2}\right)e^{-a\sqrt{s}} \tag{16.3.237}$$

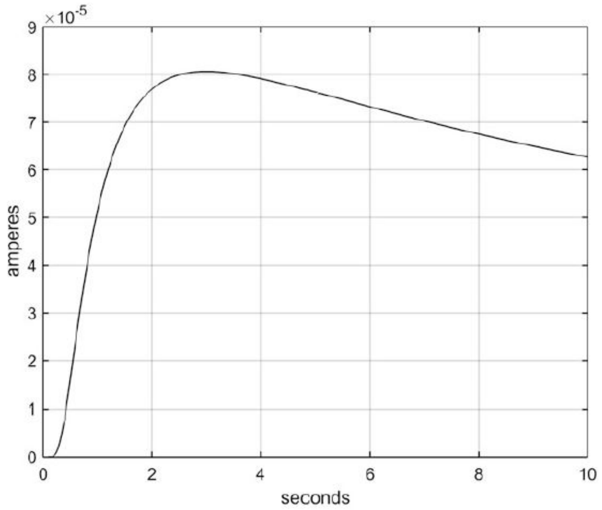
or after simplification:

$$\frac{e^{-a^2/4t}}{\sqrt{\pi t}}\leftrightarrow \frac{1}{\sqrt{s}}e^{-a\sqrt{s}} \tag{16.3.238}$$

Thus, one again setting $a=x\sqrt{RC}$, we immediately have:

$$i(x,t)=\sqrt{\frac{C}{\pi Rt}}e^{-x^2RC/4t} \tag{16.3.239}$$

The figure below shows the current $i(x,t)$ on the Atlantic cable at $x=3,704$ [km]. The values of $R=1.62\cdot 10^{-3}$ $[\Omega\cdot\mathrm{m}^{-1}]$ and $C=2.7\cdot 10^{-10}$ $[\mathrm{F}\cdot\mathrm{m}^{-1}]$ are measured for this cable and the figure shows the current for the first ten seconds after the start of the unit step input voltage at $x=0$.



There are two features in this figure of practical engineering importance.

1. First, there is a significant delay of three seconds in the occurrence of the peak current. That is, closing a telegraph key at time $t=0$ doesn't result in a maximum current response at $x=3,704$ [km] until three seconds later.

2. Second, the maximum current is a substantial current. The figure above shows that a 1 [V] step input produces a maximum current that is slightly in excess of $80\,[\mu\text{A}]$, and so if the actual sending voltage would be, say, 40 [V] (typical of the voltages used on the Atlantic copper cable), there would be a maximum current, $x=3,704$ [km] distant, three seconds later, of more than $80\,[\mu\text{A}]$, an easily detectable current in the mid-119th century (holocene calendar).

Much more interesting is the behaviour of $i(x,t)$. For any given value of x, we see from the relation above that $i(x,0)=0$ and that $i(x,=+\infty)=0$. So, for every x there must be a time $t=t_{\text{max}}$ at which $i(x,t)$ reaches an extreme value, a time that depends on the value of x. Indeed, if we differentiate:

$$i(x,t)=\sqrt{\frac{C}{\pi Rt}}e^{-x^2RC/4t} \tag{16.3.240}$$

with respect to t and set the result to zero:

$$-\frac{1}{2}\frac{e^{-x^2RC/4t}C}{\sqrt{\frac{C}{\pi Rt}}\pi Rt^2}+\frac{1}{4}\frac{\sqrt{\frac{C}{\pi Rt}}x^2RCe^{-x^2RC/4t}}{t^2}=0 \tag{16.3.241}$$

we find the "Thomson's law of squares of (metallic) transmission lines":

$$\boxed{t_{\text{max}}=\frac{1}{2}RCx^2} \tag{16.3.242}$$

Putting this back in the expression of $i(x,t)$ we get:

$$i_{\text{max}}(x,t_{\text{max}})=\frac{1}{Rx}\sqrt{\frac{2}{\pi e}} \tag{16.3.243}$$

and, if we substitute $R=1.62\cdot 10^{-3}\,[\Omega\cdot\text{m}^{-1}]$ and $x=3,704$ [km], we get $i_{\text{max}}=80.7\,[\mu\text{A}]$ at $t_{\text{max}}=3$ [s], values in excellent agreement with the above figure.

Thomson's result is quite counter-intuitive and led to some disbelieving it. The result that most telegraph engineers expected was that the delay in the peak would be directly proportional to line length. Telegraphy was in its infancy and many telegraph engineers were self taught. They tended to mistrust academics and rely instead on practical experience. Even as late as 11887 (holocene calendar), the author of a letter to *The Electrician* wished to « *...protest against the growing tendency to drag mathematics into everything.*»

Remark

A derivation not using the Laplace transform is also available at the end of the excellent book (yes the same book again...): *Transients for Electrical Engineers* **??** (as appendix).

One opponent of Thomson was of particular significance, Wildman Whitehouse, who challenged Thomson when he presented the theorem to the British Association in 11855 (holocene calendar). Both Thomson and Wildman were associated with the transatlantic telegraph cable project, Thomson as an unpaid director and scientific advisor, and Whitehouse as the Chief Electrician of the Atlantic Telegraph Company. Thomson's discovery threatened to derail the project, or at least, indicated that a much larger cable was required (a larger conductor will reduce R and a thicker insulator will reduce C). Whitehouse had no advanced mathematical education (he was a doctor by training) and did not fully understand Thomson's work. He claimed he had experimental evidence that Thomson was wrong, but his measurements were poorly conceived and Thomson refuted his claims, showing that Whitehouse's results were consistent with the law of squares.

Whitehouse believed that a thinner cable could be made to work with a high voltage induction coil. The Atlantic Telegraph Company, in a hurry to push ahead with the project, went with Whitehouse's cheaper solution rather than Thomson's. After the cable was laid, it suffered badly from retardation, an effect that had first been noticed by Latimer Clark in 11853 (holocene calendar) on the Anglo-Dutch submarine cable of the Electric Telegraph Company. Retardation causes a delay and a lengthening of telegraph pulses, the latter as if one part of the pulse has been retarded more than the other. Retardation can cause adjacent telegraph pulses to overlap making them unreadable, an effect now named "intersymbol interference". It forced telegraph operators to send more slowly to restore a space between pulses. The problem was so severe on the Atlantic cable that transmission speeds were measured in minutes per word rather than words per minute. In attempting to overcome this problem with ever higher voltage (2000 [V]), Whitehouse permanently damaged the cable insulation and made it unusable. He was dismissed shortly afterwards.

But, sure enough, by 11866 (holocene calendar), the mathematics of Thomson had resulted in the laying of two cables, each almost ten times more efficient than the previous.

Remark

There are entire books dedicated to the subject (as always...). The reader interested may read for example the following reference: *Transmission Lines and Wave Propagation* [366].

16.3.5 Linear Circuit (RLC) in Forced Regime

The objective here will be to study the behaviour of a series linear RLC circuit excited by a sinusoidal voltage generator since it is a generalization of the RL or RC circuits (it is enough to cancel L or C respectively to fall back on the solutions of a RC or RL circuit).

We then have:

$$\boxed{L\frac{\mathrm{d}^2q}{\mathrm{d}t^2}+R\frac{\mathrm{d}q}{\mathrm{d}t}+\frac{1}{C}q=U_0\sin(\omega t)} \tag{16.3.244}$$

which is a differential equation of known form and therefore important in the field of acoustics because we find it identically for the study of loudspeakers. That latter equation can often be found in the following form:

$$L\frac{\mathrm{d}i}{\mathrm{d}t}+Ri+\frac{1}{C}q=U_0\sin(\omega t) \tag{16.3.245}$$

We could very well add a phase shift to the sine on the right of the equality (phase arbitrary). This would not change the developments that follow, and let us recall that the cosine is only a sinus with a very precise phase shift!

Finally, the most important thing is that if we find a special solution to the differential equation above, then since the amplitude and the pulse can take any value at an arbitrary phase shift, then we have an infinity of special solutions. And as we have shown in our study of differential equations that the sum of special solutions is also a solution then it means that an excitation obtained with a Fourier series will also have a solution and by passing to the limit we have a Fourier transform (see section Sequences and Series page 1758)!!!

So let's move on to our study. To do this, let us derive the previous relation with respect to t:

$$L\frac{\mathrm{d}^2i}{\mathrm{d}t^2}+R\frac{\mathrm{d}i}{\mathrm{d}t}+\frac{1}{C}i=U_0\omega\cos(\omega t) \tag{16.3.246}$$

Let us then seek for a special solution of the form:

$$\begin{aligned}i&=a\cos(\omega t)+b\sin(\omega t)\\ \frac{\mathrm{d}i}{\mathrm{d}t}&=-a\omega\sin(\omega t)+b\omega\cos(\omega t)\\ \frac{\mathrm{d}^2i}{\mathrm{d}t^2}&=-a\omega^2\cos(\omega t)-b\omega^2\sin(\omega t)\end{aligned} \tag{16.3.247}$$

We notice that this proposition of solution is in every point identical to the fundamental of a Fourier series whose term a_0 is zero (which is the mean of the signal or the continuous component if it exists) as the reader can check by going back to the section of Sequences and Series where we have introduced Fourier series (see page 1727)!

Then let us inject these relationships into:

$$L\frac{\mathrm{d}^2i}{\mathrm{d}t^2}+R\frac{\mathrm{d}i}{\mathrm{d}t}+\frac{1}{C}i=U_0\omega\cos(\omega t) \tag{16.3.248}$$

By grouping the trigonometric terms of the same nature:

$$\left(-a\omega^2L+b\omega R+\frac{1}{C}a\right)\cos(\omega t)+\left(-b\omega^2L-a\omega R+\frac{1}{C}b\right)\sin(\omega t)=U_0\omega\cos(\omega t) \tag{16.3.249}$$

What is equivalent to:

$$\left(-a\omega^2L+b\omega R+\frac{1}{C}a\right)\cos(\omega t)+\left(-b\omega^2L-a\omega R+\frac{1}{C}b\right)\sin(\omega t)=U_0\omega\cos(\omega t)+0\cdot\sin(\omega t) \tag{16.3.250}$$

Hence by identifying the terms:

$$\begin{aligned}-a\omega^2L+b\omega R+\frac{1}{C}a&=U_0\omega\\-b\omega^2L-a\omega R+\frac{1}{C}b&=0\end{aligned} \tag{16.3.251}$$

We can factorize:

$$\begin{aligned}-a\left(\omega^2L-\frac{1}{C}\right)+b\omega R&=U_0\omega\\-a\omega R-b\left(\omega^2L-\frac{1}{C}\right)&=0\end{aligned} \tag{16.3.252}$$

And simplifying by ω:

$$\begin{aligned}-a\left(\omega L-\frac{1}{C\omega}\right)+bR&=U_0\\-aR-b\left(\omega L-\frac{1}{C\omega}\right)&=0\end{aligned} \tag{16.3.253}$$

and by changing the sign of the second line:

$$\begin{aligned}-a\left(\omega L-\frac{1}{C\omega}\right)+bR&=U_0\\aR+b\left(\omega L-\frac{1}{C\omega}\right)&=0\end{aligned} \tag{16.3.254}$$

It is therefore a system of two equations with two unknowns a, b which we solve by writing:

$$X=\omega L-\frac{1}{\omega C}=X_L-X_C \tag{16.3.255}$$

where we thus fall back on the "inductive reactance" and the "capacitive reactance" introduced in the section of Electrokinetics.

Which immediately gives us:

$$\begin{aligned}-aX+bR&=U_0\\aR+bX&=0\end{aligned} \tag{16.3.256}$$

hence:

$$-aX+\frac{-aR}{X}R=U_0\Rightarrow a\left(\frac{X^2+R^2}{X}\right)=-U_0\Rightarrow a=\frac{-XU_0}{X^2+R^2} \tag{16.3.257}$$

and therefore:

$$\left(\frac{-XU_0}{X^2+R^2}\right)R+bX=0\Rightarrow b=\frac{U_0R}{X^2+R^2} \tag{16.3.258}$$

We also put traditionally that (we shall see later that this expression can be assimilated to the concept of impedance by analogy with the norm of a vector):

$$Z=\sqrt{X^2+R^2} \tag{16.3.259}$$

This gives the following special solution:

$$i(t)=\frac{-XU_0}{X^2+R^2}\cos(\omega t)+\frac{U_0R}{X^2+R^2}\sin(\omega t)=\frac{U_0}{Z}\left(\frac{R}{Z}\sin(\omega t)-\frac{X}{Z}\cos(\omega t)\right) \tag{16.3.260}$$

to a given arbitrary phase value.

It is possible to find θ such as:

$$\sin(\theta)=\frac{X}{Z}\qquad\cos(\theta)=\frac{R}{Z} \tag{16.3.261}$$

Or otherwise written (thus one sees better that we browse all the possible values outside singularities):

$$\frac{\sin(\theta)}{\cos(\theta)}=\frac{X}{R} \tag{16.3.262}$$

We then have using the remarkable trigonometric relations (see section Trigonometry page 2519):

$$i(t)=\frac{U_0}{Z}\left(\cos(\theta)\sin(\omega t)-\frac{X}{Z}\sin(\theta)\cos(\omega t)\right)=\frac{U_0}{Z}\sin(\omega t-\theta) \tag{16.3.263}$$

θ is therefore the phase of the current, that is to say the advance or the delay of the current on the voltage. If $\theta=0$ then we have:

$$\sin(0)=0=\frac{X}{Z}\Rightarrow X=0\qquad\cos(0)=1=\frac{R}{Z}\Rightarrow R=Z=\sqrt{X^2+R^2}=\sqrt{0^2+R^2}=R \tag{16.3.264}$$

and then:

$$i(t)=\frac{U_0}{R}\sin(\omega t) \tag{16.3.265}$$

We then say that there is "resonance of the circuit" with therefore:

$$X=L\omega-\frac{1}{C\omega}=0\Rightarrow\omega=\frac{1}{\sqrt{LC}}=\frac{2\pi}{R} \tag{16.3.266}$$

Either when the inductive reactance is equal to the capacitive reactance.

In that latter situation we often found another expression for Z. Let us remind that it is defined by:

$$Z=\sqrt{X^2+R^2} \tag{16.3.267}$$

So in the special case where the is no phase shift between the current and voltage that latter is written:

$$Z=\sqrt{\left(L\omega-\frac{1}{C\omega}\right)^2+R^2} \tag{16.3.268}$$

That expression can been see as the module of a complex number simply given by:

$$Z=R^2+\mathrm{i}\left(L\omega-\frac{1}{C\omega}\right) \tag{16.3.269}$$

Let us prove that using the Fourier transform (see section Sequences and Series page 1758)!

Obviously the differential equation we have introduced above can be written:

$$L\frac{\mathrm{d}i(t)}{\mathrm{d}t}+Ri(t)+\frac{1}{C}\int i(t)\mathrm{d}t=u(t) \tag{16.3.270}$$

Using the Fourier transform of the derivative in time (see section Analysis page 2316):

$$\mathcal{F}\left(\frac{\mathrm{d}^n}{\mathrm{d}t^n}x(t)\right)=(\mathrm{i}\omega)^n X(\mathrm{i}\omega) \tag{16.3.271}$$

and integration in time (see section Analysis page 2316):

$$\mathcal{F}\left(\int_{-\infty}^{t} x(\tau)\mathrm{d}\tau\right)=\frac{1}{\mathrm{i}\omega}X(\mathrm{i}\omega)+\pi X(0)\delta(\omega) \tag{16.3.272}$$

we get:

$$L\mathrm{i}\omega I(\mathrm{i}\omega)+RI(\mathrm{i}\omega)+\frac{1}{C}\left(\frac{1}{\mathrm{i}\omega}I(\mathrm{i}\omega)+\pi I(0)\delta(\omega)\right)=U(\mathrm{i}\omega) \tag{16.3.273}$$

In the case of a steady state, constant harmonic frequency signal the Fourier transform $I(0)$ is equal to zero. Therefore:

$$L\mathrm{i}\omega I(\mathrm{i}\omega)+RI(\mathrm{i}\omega)+\frac{I(\mathrm{i}\omega)}{\mathrm{i}\omega C}=U(\mathrm{i}\omega) \tag{16.3.274}$$

After rearranging:

$$L\mathrm{i}\omega+R+\frac{1}{\mathrm{i}\omega C}=\frac{U(\mathrm{i}\omega)}{I(\mathrm{i}\omega)}=Z(\mathrm{i}\omega) \tag{16.3.275}$$

Therefore:

$$Z(\mathrm{i}\omega)=L\mathrm{i}\omega+R+\frac{1}{\mathrm{i}\omega C}=R+L\mathrm{i}\omega-\mathrm{i}\frac{1}{\omega C}=R+\mathrm{i}\left(L\omega+\frac{1}{C\omega}\right) \tag{16.3.276}$$

And this finish to prove our claim!

16.3.5.1 Low-pass filter

Let us consider the case where L is zero. We have then:

$$X=L\omega-\frac{1}{C\omega}=\frac{1}{C\omega} \tag{16.3.277}$$

Therefore:

$$Z=\sqrt{X^2+R^2}=\sqrt{\left(\frac{1}{C\omega}\right)^2+R^2}=\frac{1}{C\omega}\sqrt{1+C\omega R^2} \tag{16.3.278}$$

Hence:

$$i=\frac{\mathrm{d}q}{\mathrm{d}t}=\frac{C\omega}{\sqrt{1+(C\omega R)^2}}U_0\sin(\omega t-\theta) \tag{16.3.279}$$

Finally:

$$\begin{aligned}q&=\int\frac{C\omega U_0}{\sqrt{1+(C\omega R)^2}}\sin(\omega t-\theta)\mathrm{d}t=\frac{C\omega U_0}{\sqrt{1+(C\omega R)^2}}\int\sin(\omega t-\theta)\mathrm{d}t\\&=\frac{C\omega U_0}{\sqrt{1+(C\omega R)^2}}\left(-\frac{\cos(\omega t-\theta)}{\omega}\right)=-\frac{C\omega U_0}{\sqrt{1+(C\omega R)^2}}\cos(\omega t-\theta)\end{aligned} \tag{16.3.280}$$

We then have at the terminals of the capacitor:

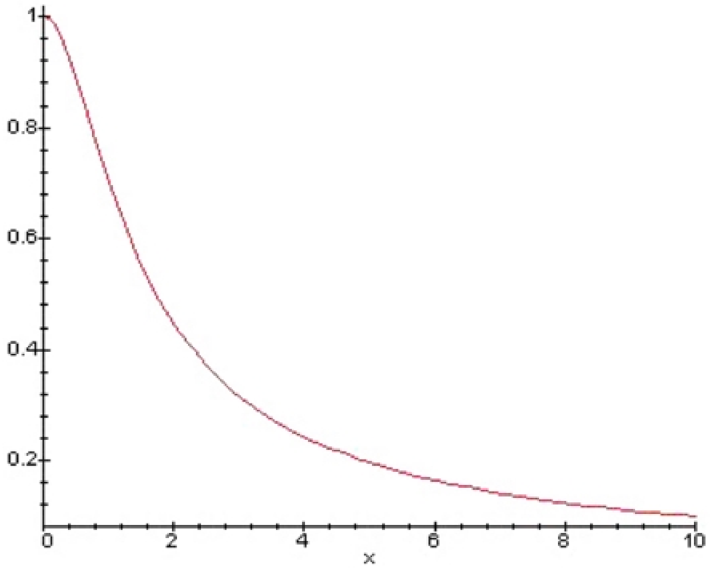
$$u_C=\frac{1}{C}q=-\frac{1}{\sqrt{1+(C\omega R)^2}}U_0\cos(\omega t-\theta) \tag{16.3.281}$$

We see then that the voltage at the terminals of the capacitor acts as what we name a "low-pass filter". That is to say that the amplitude of the voltage across the capacitor with respect to the excitation voltage of the circuit will be reduced, and this is all the more as the frequency will be high.

This type of tool is very useful for example to eliminate the high-frequency harmonics of a periodic signal obtained by Fourier series or for cleaning a high-frequency noise. Cascaded low-pass filters can also be used to perform spectrum analysers.

Here is the plot of the factor:

$$\frac{1}{\sqrt{1+(C\omega R)^2}} \tag{16.3.282}$$



We can see that at low frequencies (on the left) the amplitude is preserved (the low-pass filter therefore let pass the low frequencies). Beyond that, the signal is turned off.

The ratio:

$$\frac{u_C}{U_0}=\frac{1}{\sqrt{1+(C\omega R)^2}} \tag{16.3.283}$$

Is often expressed in decibels Db either by definition using the measure:

$$\text{dB}=20\log\left(\frac{u_C}{U_0}\right) \tag{16.3.284}$$

And is then referred to as the "transfer function" of the filter.

16.3.5.2 High-pass filter

With regard to the voltage at the terminals of the resistor, we have:

$$u_R=R\cdot i_0=\frac{RC\omega}{\sqrt{1+(C\omega R)^2}}U_0\sin(\omega t-\theta) \tag{16.3.285}$$

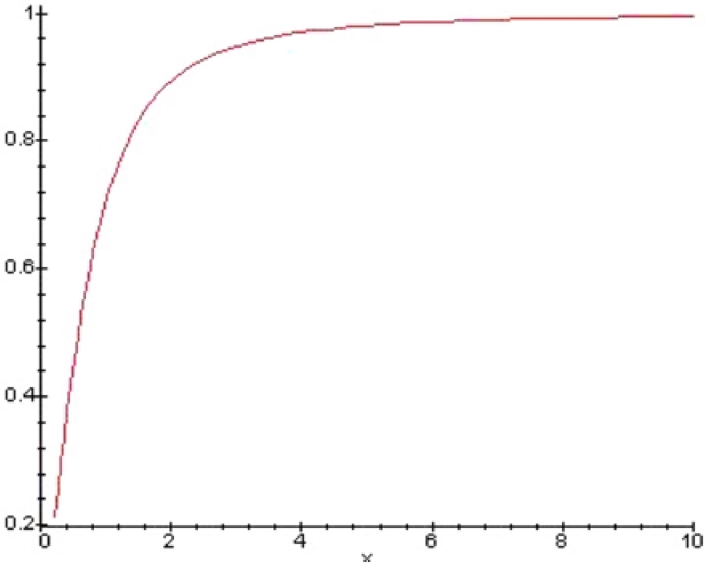
Which is traditionally modified in the following form:

$$u_R=\frac{1}{\sqrt{1+\dfrac{1}{(C\omega R)^2}}}U_0\sin(\omega t-\theta) \tag{16.3.286}$$

So we see that the voltage at the terminals of the resistor acts as what we name a "high-pass filter". That is to say that the amplitude of the voltage across the resistor with respect to the excitation voltage of the circuit will be reduced, and the more so as the frequency will be small.

Here is the plot of the factor:

$$\frac{1}{\sqrt{1+\frac{1}{(C\omega R)^2}}} \tag{16.3.287}$$



We can see that at small frequencies (on the right) the amplitude is preserved (the high-pass filter allows to pass the high frequencies). Beyond this the signal is switched off.

The ratio:

$$\frac{u_R}{U_0}=\frac{1}{\sqrt{1+\frac{1}{(C\omega R)^2}}} \tag{16.3.288}$$

Is often expressed in decibels either by definition using the measure:

$$\text{dB}=20\log\left(\frac{u_R}{U_0}\right) \tag{16.3.289}$$

and is then also referred to as the "transfer function" of the filter.

With different types of assembled filters we can thus remove (but never completely) frequency ranges. We're talking about a "band-pass filter". This is the technique used, for example, for the reception of a certain radio or television channel in a specific frequency range, or also in electronic music to attenuate low pitched or high pitched sounds. Or to separate the ADSL signal from the voice of a telephone line.

A "passive filter" is characterized by the exclusive use of linear passive components (resistors, capacitors, coupled coils or not). Therefore, their gain (power ratio between output and input) can not exceed one. They can therefore only partly attenuate signals, but not amplify them, because this would require an input of energy (which is the role of an "active filters").

16.3.5.3 Integrator and differentiator

We therefore have at the terminals of the capacitor:

$$u_C=\frac{1}{C}\int i_0\mathrm{d}t=\frac{\omega}{\sqrt{1+(C\omega R)^2}}\int U_0\sin(\omega t-\theta)\mathrm{d}t=\frac{1}{\sqrt{\dfrac{1+(C\omega R)^2}{\omega^2}}}\int U_0\sin(\omega t-\theta)\mathrm{d}t$$
$$=\frac{1}{\sqrt{\dfrac{1}{\omega^2}+(CR)^2}}\int U_0\sin(\omega t-\theta)\mathrm{d}t \tag{16.3.290}$$

Now, if $\omega\gg(RC)^{-1}$, we have:

$$u_C\cong\frac{1}{CR}\int U_0\sin(\omega t-\theta)\mathrm{d}t \tag{16.3.291}$$

If we make things so that $\theta=0$ we must have:

$$X=L\omega-\frac{1}{C\omega}=0 \tag{16.3.292}$$

Therefore:

$$L=\frac{1}{C} \tag{16.3.293}$$

Since then:

$$u_C\cong\frac{1}{CR}\int U_0\sin(\omega t)\mathrm{d}t \tag{16.3.294}$$

The circuit is then what we name logically enough ... an "integrator".

Let us now look at the resistance side:

$$u_R=R\cdot i_0 \tag{16.3.295}$$

But we have:

$$u_C=\frac{1}{C}\int i_0\mathrm{d}t\Rightarrow\frac{\mathrm{d}u_C}{\mathrm{d}t}=\frac{1}{C}i_0\Rightarrow i_0=C\frac{\mathrm{d}u_C}{\mathrm{d}t} \tag{16.3.296}$$

Therefore:

$$u_R=RC\frac{\mathrm{d}u_C}{\mathrm{d}t} \tag{16.3.297}$$

As:

$$u_C=-\frac{1}{\sqrt{1+(C\omega R)^2}}U_0\cos(\omega t-\theta) \tag{16.3.298}$$

We have then:

$$u_R=-\frac{RC}{\sqrt{1+(C\omega R)^2}}\frac{\mathrm{d}}{\mathrm{d}t}U_0\cos(\omega t-\theta)=-\frac{1}{\sqrt{\dfrac{1}{(RC)^2}+\omega^2}}\frac{\mathrm{d}}{\mathrm{d}t}U_0\cos(\omega t-\theta) \tag{16.3.299}$$

If $\omega \ll (RC)^{-1}$ then:

$$u_R = -RC\frac{\mathrm{d}}{\mathrm{d}t}U_0\cos(\omega t - \theta) \tag{16.3.300}$$

If we make things so that $\theta = \pi/2$ we must have:

$$u_R = -RC\frac{\mathrm{d}}{\mathrm{d}t}U_0\cos\left(\omega t - \frac{\pi}{2}\right) = -RC\frac{\mathrm{d}}{\mathrm{d}t}U_0\cos\left(\omega t - \frac{\pi}{2}\right) = RC\frac{\mathrm{d}}{\mathrm{d}t}U_0\sin(\omega t) \tag{16.3.301}$$

The circuit is then what we name logically enough ... a "derivator".

The utility of an integrating circuit is, for example, to transform a periodic signal into a constant (since the time average of a periodic signal having an offset will never be zero).

16.3.6 Signal Processing

Signal processing concerns the analysis, synthesis, and modification of signals, which are broadly defined as functions conveying, "information about the behaviour or attributes of some phenomenon", such as sound, images, and biological measurements. For example, signal processing techniques are used to improve signal transmission fidelity, storage efficiency, and subjective quality, and to emphasize or detect components of interest in a measured signal.

It may seems curious to the reader that we deal with signal processing in the section of Electrical Engineering as it also applies to Numerical Methods! But in my point of view, the most concerned application and tools related to signal processing is electronics as explained just in the paragraph before.

16.3.6.1 Nyquist-Shannon Theorem (Nyquist sampling theorem)

The "Nyquist sampling theorem", or "Nyquist-Shannon theorem", provides a prescription for the sampling frequency required for a "good enough" sampling for a perfectly periodic signal into a digital (square pulse) transformation. It is stated simply, and wrongly (as Nyquist never said this!!!!!), as follows (and should not bear the name "theorem" as it is only an engineering practice and not a general perfect theorem!!!!!):

Theorem 16.1. The sampling frequency of a signal $s(t)$ should equal twice the highest frequency contained in the signal. Or in mathematical terms:

$$f_s = 2f_{\text{max}} \tag{16.3.302}$$

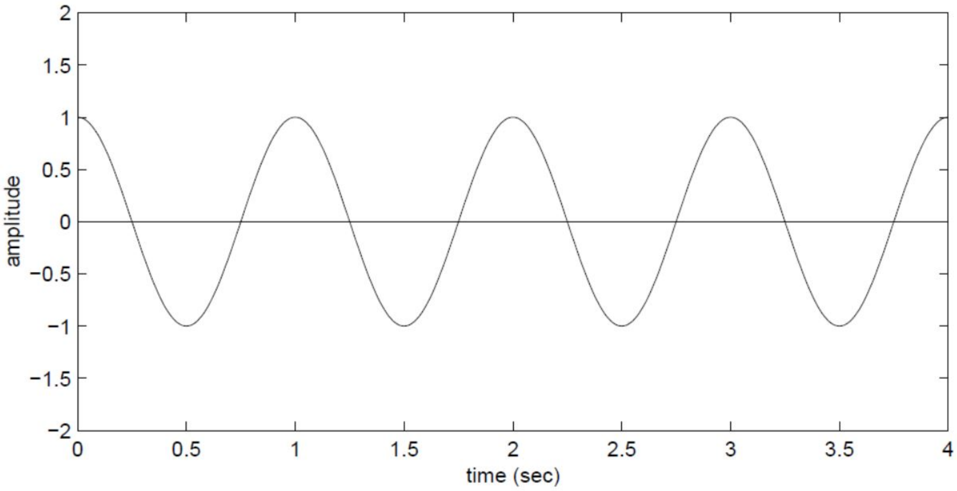
where f_s is the sampling frequency (how often samples are taken per unit of time or space), in this special case named "Nyquist rate", and f_{max} is the highest frequency contained in the signal. Under the assumption that the signal must not contain a sinusoidal component at exactly frequency f_{max} (see further below the example to understand why this assumption...)!

Remark

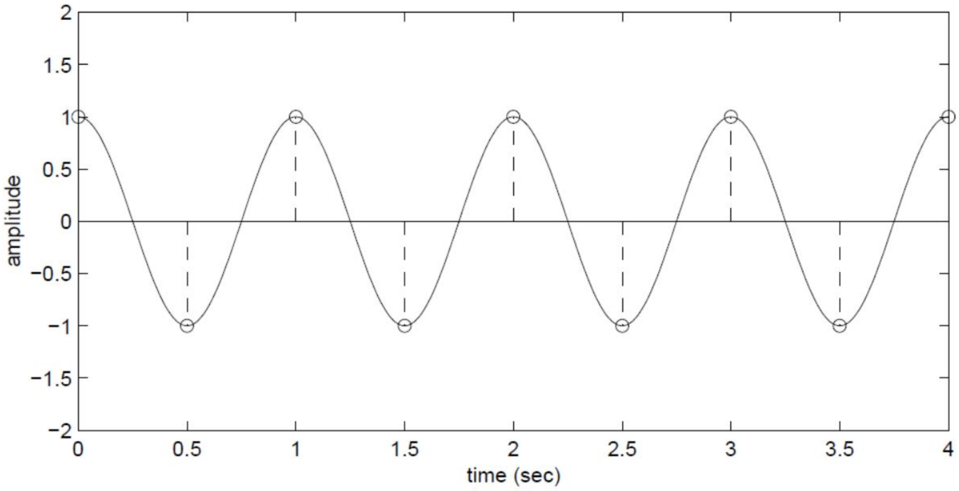
In reality, even if ideal sampling could be achieved, for a lot of manipulations and analyses, the original signal needs to be reconstructed at a higher sample frequency. So what sampling frequency should we use? Well, the answer is application dependent and my require careful statistical analysis of error rates between the original signal and the rebuild one based on the sampling. A rule of thumb in practice seems to consider 2 samples per period (Nyquist criteria) for very repetitive signals, 800 samples per period for non-repetitive signals where extreme precision is required.

The first explanation that we can find in many textbooks is the following (even if it's wrong, it is a first approach that helps to understand the real explanation we will see later further below):

Let us consider for example a signal composed of a single cosine wave at a frequency of 1 [Hz]:



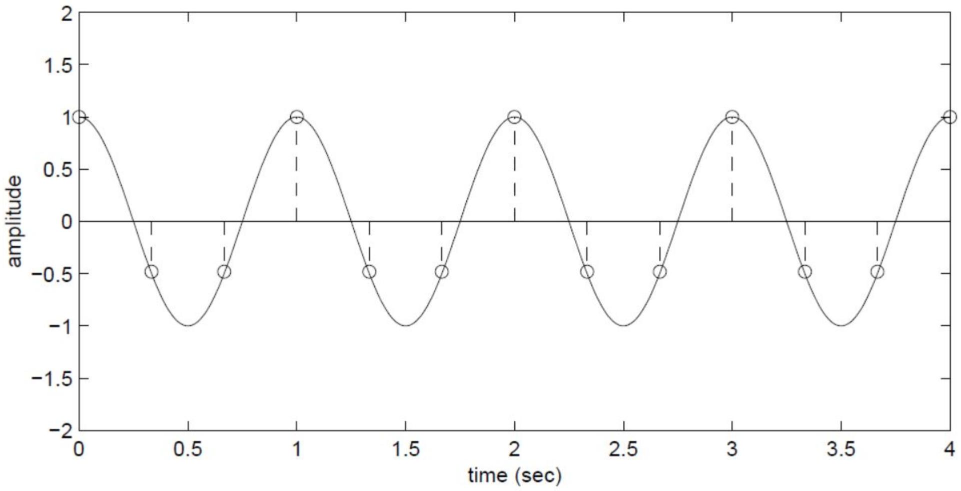
If we sample this waveform at 2 [Hz] (as dictated by the Nyquist theorem), this is sufficient to capture each peak of the signal[12]:



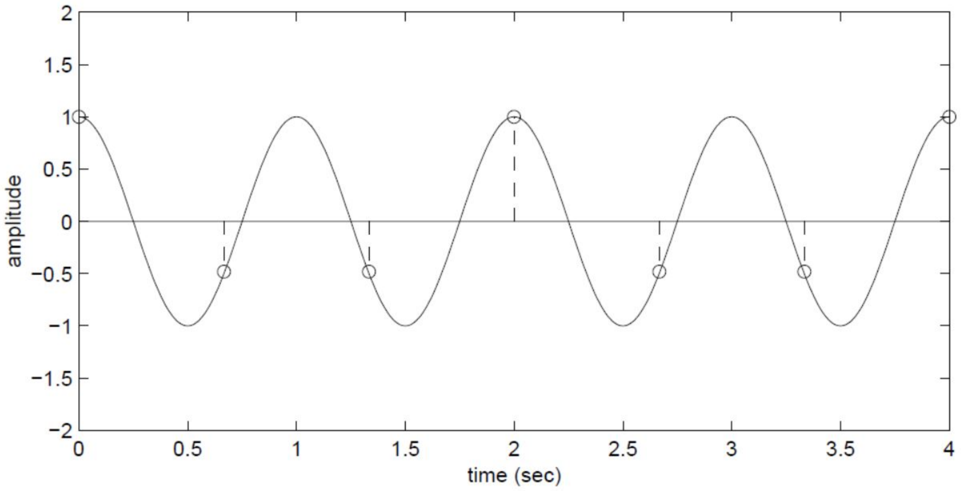
If we sample at a frequency higher than this, for example 3 [Hz], then there are more than enough samples to capture the variations in the signal:

---

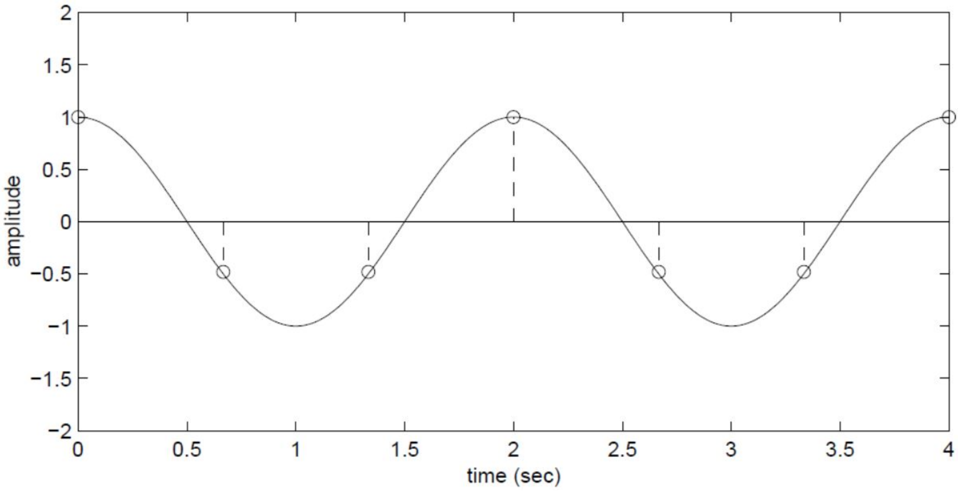
[12] Obviously if the signal was a sine instead of a cosine, we see that it would be wrong as we would have a 0 sampling everywhere, hence the precision about the sine component in the theorem...



If however we sample at a frequency lower than 2 [Hz], for example at 1.5 [Hz], then there are now not enough samples to capture all the peaks and troughs in the signal:

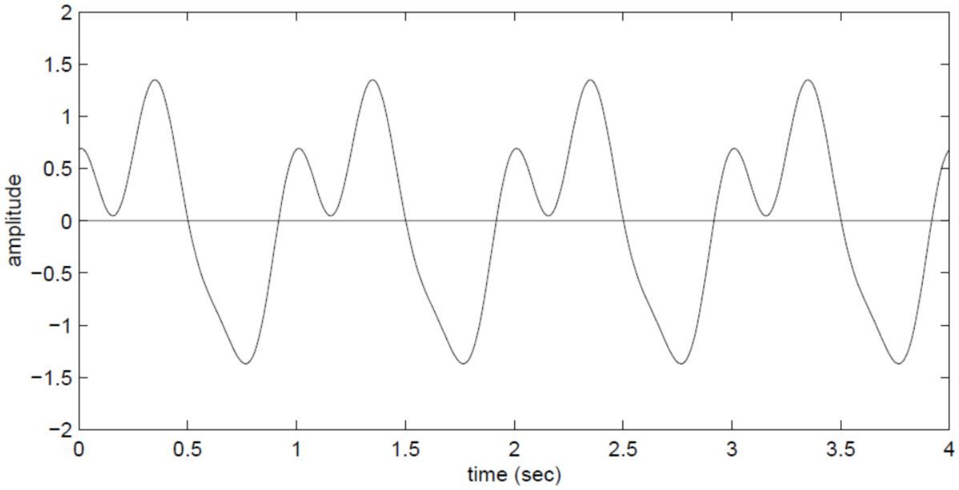


Note here that we are not only losing information, but we are getting the wrong information about the signal. The person receiving these samples, without any previous knowledge of the original signal, may well be mislead into thinking that the signal has quite a different form:

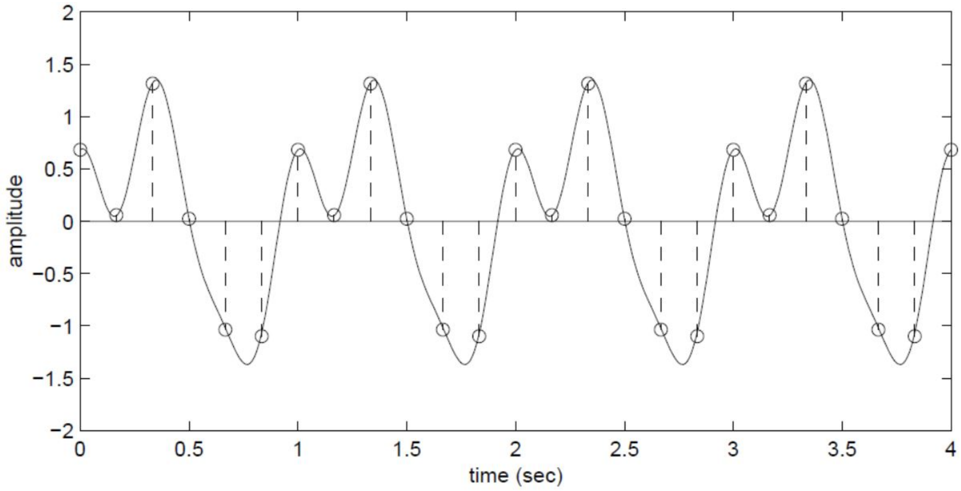


From this example, we can see the reason for the term "signal aliasing". That is, the signal now takes on a different "persona" or a false presentation, due to being sampled at an insufficiently high frequency.

Now we are ready to think about the sampling of little bit more perfectly periodic complex signal composed of some frequency components. By Fourier's theorem, we know that any continuous signal may be decomposed in terms of a sum of sines and cosines at different frequencies. For example, the following waveform was composed by adding together sinewaves at frequencies of 1 [Hz], 2 [Hz], and 3 [Hz]:



According to the Nyquist sampling "theorem", the signal must be sampled at twice the highest frequency contained in the signal. In this case, we have $f_c = 3$ [Hz], and so the Nyquist theorem tells us that the sampling frequency, f_s, must be at least 6 [Hz]. And sure enough, this appears to be sufficient (keep in mind that the example below is still a perfect periodic signal!):



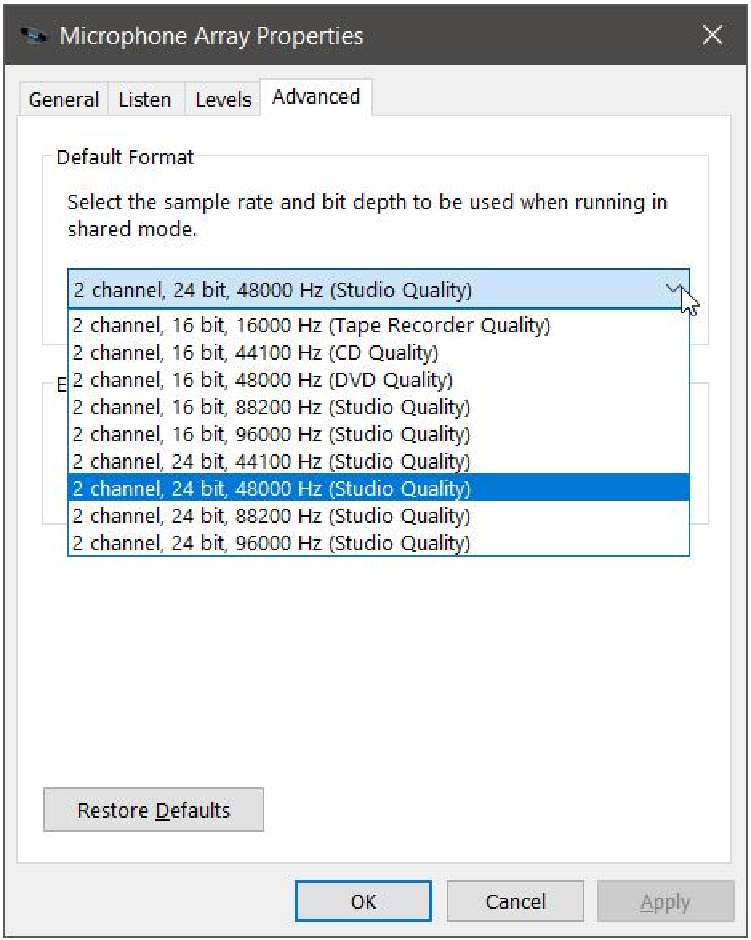
Thus, when a signal contains not just one but many different frequencies added together, the minimal sampling rate recommended to avoid aliasing at least that of the highest frequency, irrespective of how many other frequency components there are.

The human ear capabilities is between 20 [Hz] to $20,050$ [Hz] (see section Music Mathematics page 8198) hence the fact that many standard sampling rate are at satisfactory sampling rate of[13]:

$$20,050 \cdot 2 = 44,100\text{ [Hz]} = 44.1\text{ [kHz]} \tag{16.3.303}$$

But for telephony the usable voice range is for example 300 [Hz] to $3,400$ [Hz], so a sampling frequency of 8 [kHz] should be satisfactory enough in such a case.

By the way, on a computer operating system like Microsoft Windows, and for most microphone, the user can choose the sampling rate as illustrated below:



Default Format

Select the sample rate and bit depth to be used when running in shared mode.

2 channel, 24 bit, 48000 Hz (Studio Quality)

- 2 channel, 16 bit, 16000 Hz (Tape Recorder Quality)
- 2 channel, 16 bit, 44100 Hz (CD Quality)
- 2 channel, 16 bit, 48000 Hz (DVD Quality)
- 2 channel, 16 bit, 88200 Hz (Studio Quality)
- 2 channel, 16 bit, 96000 Hz (Studio Quality)
- 2 channel, 24 bit, 44100 Hz (Studio Quality)
- 2 channel, 24 bit, 48000 Hz (Studio Quality)
- 2 channel, 24 bit, 88200 Hz (Studio Quality)
- 2 channel, 24 bit, 96000 Hz (Studio Quality)

Figure 16.121: **Microsoft Windows 10 Microphone Array Properties**

Obviously the choices may depend on your material (computer/microphone).

---

[13] The rate 44.1 [kHz] was chosen in fact following debate between manufacturers, notably Sony and Philips, and its implementation by Sony, yielding a de facto standard.

✍**Example:**

An audiostream sampled at 44.1 [kHz] with an amplitude coded on 16 bits in stereo (2 channels) will have a bit rate of:

$$44100\cdot 16\cdot 2=1,411,200\;[\text{b}\cdot\text{s}^{-1}]\cong 1,441\;[\text{kbps}] \tag{16.3.304}$$

This is the typical CD quality (for those who know what is Compact Disc).

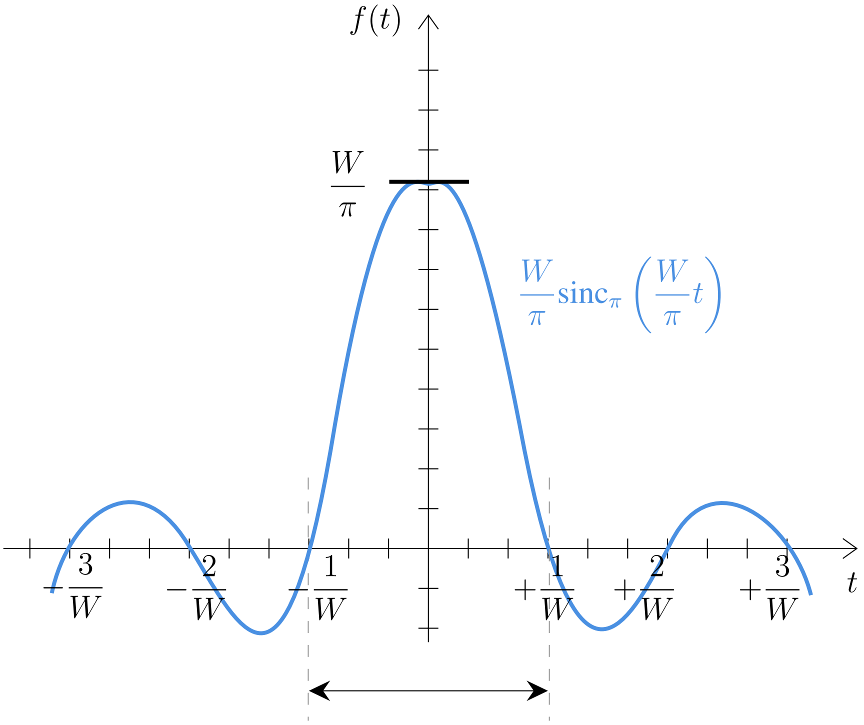
Most DAB+ broadcast (Digital Audio Broadcasting) emitter have much lower bitrate! The recommended one being 192 [kbps] (but the reader must keep in mind that this include the fact it is MPG compressed (using at least a Huffman algorithm as seen at page 7209) and also have additional error correcting code bits). Furthermore, best DAB+ methods seems to implements a technique named "VBR" for "Variational Bitrate" where the simple idea is that the bitrate change dynamically depending on the complexity of the signal.

Remark

Digital signal encoding based on uniformly distributed samples seems to be named "Linear Pulse Code Modulation" (abbreviated LPCM or simply PCM). It is used for example by CD (with a sample rate of 44.1 [Khz] and an amplitude binary coded on 16 bits) and also by some broadcasting techniques. For sounds like jazz, classic music, acoustic that have a lot of dynamics (variations) an amplitude of 24 bits is recommended... it is also used by movies theatres!

But if the reader was attentive when reading this first "pseudo explanation" maybe the latter has seen that the argumentation is a little bit weak...

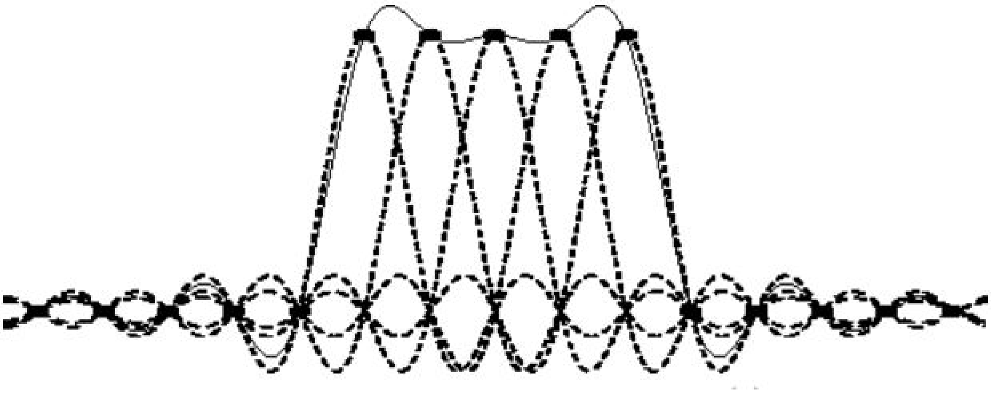
So here is the original one: In practice, for historical reasons, when we sample a continuous signal, we transform it into series of square pulses in the frequency domain (pulses of different lengths and amplitudes). To get the original signal back, we inverse Fourier transform the pulses and maybe if you remember our study and examples of inverse Fourier transform, the inverse Fourier transform of a square pulse of width $2W$ (this width being named "Nyquist bandwidth") in the frequency domain gives in the time domain the sinc function (see section Sequences and Series page 1766):



width of the main lobe is equal to $2/W$

Therefore sinc-shaped pulses possess a rectangular spectrum, enabling data to be encoded in a minimum spectral pulsation (i.e. frequency) bandwidth of $[-W,+W]$.

The final result for a series of square impulses in the frequency space is that we get a series of shifted sinc functions in the time space as illustrated below with a special example, named sometimes a "Nyquist sinc pulse":



The dots show the signal samples, the dashed lines show the component shifted sinc functions, and the solid line shows the unique bandlimited reconstruction from the samples obtained by summing the component sinc functions.

Looking to the figure above the reader can't ignore (in a practical point of view!) the fact that if the distance pic-to-pic of each sine cardinal is equal to $1/W$ it will greatly improve DAC (Digital to Analogic Converters) to detect points of interest for the conversion as they are significantly spaced. Keep in mind this information for later!

Remark

The reader can notice that we have intersymbol interference in the previous figure, this effect is named "intersymbol interference" or abbreviated "ISI". As we will see further below, the Nyquist sampling theorem prove in fact that this interferences cancels each other!

In communications, we name "Nyquist ISI criterion" the conditions which, when satisfied by a communication channel (including responses of transmit and receive filters), result in no intersymbol interference!

The idea of the processes is illustrated below (the step with the sine cardinal is not visible):

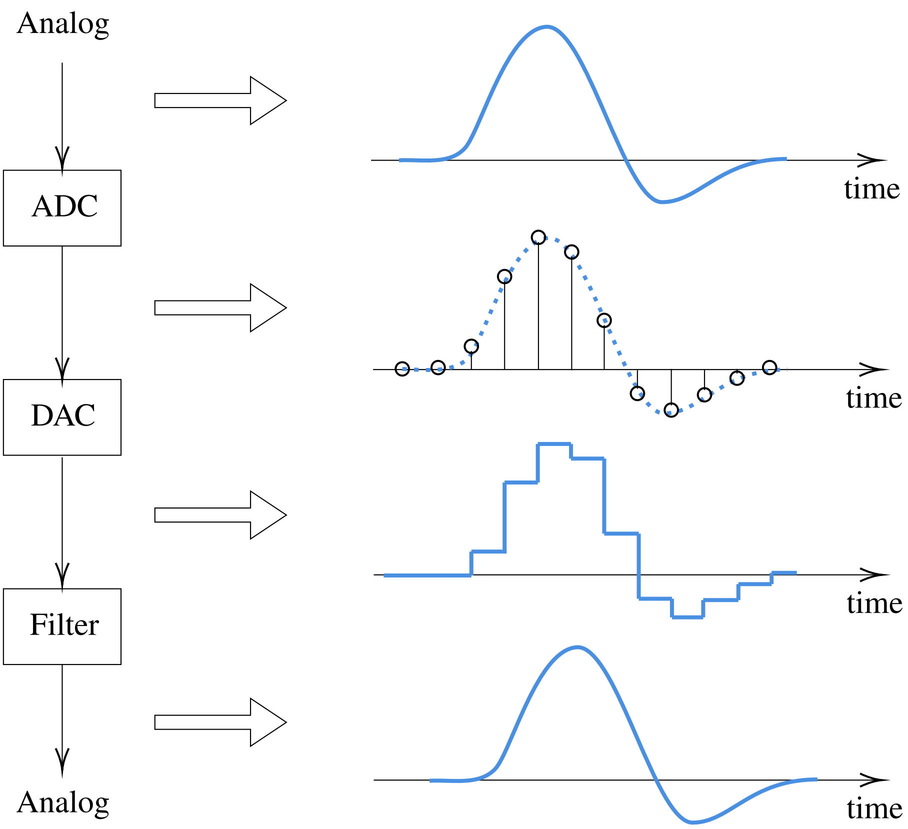


Figure 16.122: DAC and ADC converter concepts

And this was the real idea (we should rather speak of "question") of Harry Nyquist and Claude Shannon!!!!!!!!! Is it possible using only sine cardinal as kernels under and under certain conditions to rebuild any given original signal that is perfectly periodic and using a sampling time spaced?

So here it comes... the real technical version of the theorem:

Theorem 16.2. If $s(t)$ and $F(\omega)$, the Fourier transform of $s(t)$, is supported on the interval $[-f_c,+f_c]$ then:

$$s(t)=\sum_{n\in\mathbb{Z}} \underbrace{f\left(\frac{n}{2f_c}\right)}_{\text{amplitue of Nyquist pulse}} \underbrace{\operatorname{sinc}\left(2f_c\left(t-\frac{n}{2f_c}\right)\right)}_{\text{Nyquist pulse}} \qquad (16.3.305)$$

In other words, the theorem says that if an absolutely integrable function $s(t)$ contains no frequencies higher than f_c [Hz], then it is completely determined by its samples at a uniform grid spaced at distances $1/(2f_c)$ apart via the relation above where the ideal interpolation kernel is proportional to the sinc function (at the time the theorem was written, the cardinal sine was considered as a nice symmetrical decreasing function to describe and interpolate a frequency spectrum[14]!).

Remark

At the opposite of what most textbooks or Internet websites say, the Nyquist sampling theorem doesn't says what is the ideal sample frequency but just says that the above relation exists!!!

Proof 16.2.1. The spectral analysis theorem shows that any signal can be decomposed into a sum of sinusoids of different frequencies, amplitudes and phases. We consider a signal inscribed between a minimum frequency and a maximum frequency. Experience determines which frequency range is of interest. Even if the frequency coefficients outside this range are not zero, they are neglected since they do not contribute significantly to the total mean square value.

Let us recall that the Fourier Transform $F(\omega)$ of a function $f(t)$ is given by (see section Sequences and Series page 1758 or section Functional Analysis page 2305):

$$\mathcal{F}(f(t))(\omega)=F(\omega)=\int\limits_{-\infty}^{+\infty} f(t)e^{\mathrm{i}\omega t}\mathrm{d}t \tag{16.3.306}$$

To avoid any notation confusion for what follows we will write it (the s stand for "signal"):

$$\mathcal{F}(s(t))(\omega)=F(\omega)=\int\limits_{-\infty}^{+\infty} s(t)e^{\mathrm{i}\omega t}\mathrm{d}t \tag{16.3.307}$$

The Fourier transform described the signal by its frequencies f that it contains, expressed in these equations by the means of the pulsation $\omega=2\pi f$.

As we know the inverse Fourier transform gives $s(t)$ in function of $F(\omega)$:

$$\mathcal{F}^{-1}\left(F(\omega)\right)(t)=s(t)=\dfrac{1}{2\pi}\int\limits_{-\infty}^{+\infty} F(\omega)e^{\mathrm{i}\omega t}\mathrm{d}\omega \tag{16.3.308}$$

The signal of which the theorem deals is limited in frequency (the square pulse as we know). Beyond $f_{\max}$, corresponding to a pulsation $\omega_{\max}=2\pi f_{\max}$, the frequency coefficients are considered as negligible.

Therefore:

$$\mathcal{F}^{-1}\left(F(\omega)\right)(t)=s(t)=\dfrac{1}{2\pi}\int\limits_{-\omega_{\max}}^{+\omega_{\max}} F(\omega)e^{\mathrm{i}\omega t}\mathrm{d}\omega \tag{16.3.309}$$

---

[14] Hence this theorem is more a "communication engineering theorem" than a purely general math theorem!

Let us look for the value of the samples e_n regularly spaced taking the values of $s(t)$ for t multiple of a given fraction of highest frequency component of the signal:

$$t_n = n\frac{1}{kf_{\text{max}}} = \frac{nT_{\text{max}}}{k} \tag{16.3.310}$$

Hence:

$$e_n = s\left(\frac{nT_{\text{max}}}{k}\right) = \frac{1}{2\pi}\int\limits_{-\omega_{\text{max}}}^{+\omega_{\text{max}}} F(\omega)e^{\mathrm{i}\omega\frac{nT_{\text{max}}}{k}}\mathrm{d}\omega \tag{16.3.311}$$

Now let us recall that we have proved that the coefficient of a Fourier series was given by:

$$c_n = \frac{1}{T}\int\limits_{-\frac{T}{2}}^{+\frac{T}{2}} f(t)e^{-\mathrm{i}\frac{2\pi n}{T}t}\mathrm{d}t = \frac{1}{T}\int\limits_{-\frac{T}{2}}^{+\frac{T}{2}} f(t)e^{-\mathrm{i}\omega nt}\mathrm{d}t \tag{16.3.312}$$

Let us imagine this latter Fourier coefficient (that is for recall in time space) but in frequency (i.e. pulsation) space (!!!), then it would be written:

$$c_n = \frac{1}{c_1^{te}\omega}\int\limits_{-c_2^{te}\omega}^{+c_2^{te}\omega} f(\omega)e^{-\mathrm{i}Tn\omega}\mathrm{d}\omega \tag{16.3.313}$$

In the case of our square pulse that has a frequency domain limited between $[-\omega_{\text{max}}, +\omega_{\text{max}}]$ (hence a width of $2\omega_{\text{max}}$), we then rewrite the latter relation with:

$$c_n = \frac{1}{2\omega_{\text{max}}}\int\limits_{-\omega_{\text{max}}}^{+\omega_{\text{max}}} F(\omega)e^{-\mathrm{i}\frac{T_{\text{max}}}{k}n\omega}\mathrm{d}\omega \tag{16.3.314}$$

For recall we divide by $2\omega_{\text{max}}$ because:

- It is the width of our square pulse
- Because we need the average value

Hence:

$$c_{-n} = \frac{1}{2\omega_{\text{max}}}\int\limits_{-\omega_{\text{max}}}^{+\omega_{\text{max}}} F(\omega)e^{\mathrm{i}\frac{T_{\text{max}}}{k}n\omega}\mathrm{d}\omega = \frac{\pi e_n}{\omega_{\text{max}}} \tag{16.3.315}$$

The reconstruction of $F(\omega)$ in Fourier series is given as we know by:

$$F(\omega) = \sum_{n=-\infty}^{+\infty} c_n e^{\mathrm{i}\omega n\frac{T_{\text{max}}}{k}} = \sum_{n=-\infty}^{+\infty} c_{-n} e^{-\mathrm{i}\omega n\frac{T_{\text{max}}}{k}} = \sum_{n=-\infty}^{+\infty} c_{-n} e^{-\mathrm{i}\omega n\frac{2\pi}{k\omega_{\text{max}}}} \tag{16.3.316}$$

Therefore:

$$\begin{aligned}
\mathcal{F}^{-1}\left(F(\omega)\right)(t)=s(t)&=\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty}F(\omega)e^{\mathrm{i}\omega t}\mathrm{d}\omega=\frac{1}{2\pi}\int\limits_{-\omega_{\max}}^{+\omega_{\max}}F(\omega)e^{\mathrm{i}\omega t}\mathrm{d}\omega\\
&=\frac{1}{2\pi}\int\limits_{-\omega_{\max}}^{+\omega_{\max}}\left(\sum_{n=-\infty}^{+\infty}c_{-n}e^{-\mathrm{i}\omega n\frac{2\pi}{k\omega_{\max}}}\right)e^{\mathrm{i}\omega t}\mathrm{d}\omega\\
&=\frac{1}{2\pi}\int\limits_{-\omega_{\max}}^{+\omega_{\max}}\left(\sum_{n=-\infty}^{+\infty}c_{-n}e^{\mathrm{i}\omega\left(t-n\frac{2\pi}{k\omega_{\max}}\right)}\right)\mathrm{d}\omega\\
&=\frac{1}{2\pi}\int\limits_{-\omega_{\max}}^{+\omega_{\max}}\left(\sum_{n=-\infty}^{+\infty}\left(\frac{\pi e_n}{\omega_{\max}}\right)e^{\mathrm{i}\omega\left(t-n\frac{2\pi}{k\omega_{\max}}\right)}\right)\mathrm{d}\omega\\
&=\frac{1}{2\omega_{\max}}\int\limits_{-\omega_{\max}}^{+\omega_{\max}}\left(\sum_{n=-\infty}^{+\infty}e_n e^{\mathrm{i}\omega\left(t-n\frac{2\pi}{k\omega_{\max}}\right)}\right)\mathrm{d}\omega\\
&=\frac{1}{2\omega_{\max}}\sum_{n=-\infty}^{+\infty}\left(\int\limits_{-\omega_{\max}}^{+\omega_{\max}}s\left(\frac{nT_{\max}}{k}\right)e^{\mathrm{i}\omega\left(t-n\frac{2\pi}{k\omega_{\max}}\right)}\mathrm{d}\omega\right)\\
&=\frac{1}{2\omega_{\max}}\sum_{n=-\infty}^{+\infty}s\left(\frac{nT_{\max}}{k}\right)\left(\int\limits_{-\omega_{\max}}^{+\omega_{\max}}e^{\mathrm{i}\omega\left(t-n\frac{2\pi}{k\omega_{\max}}\right)}\mathrm{d}\omega\right)
\end{aligned}\tag{16.3.317}$$

Now we will focus on the integral only:

$$\begin{aligned}
\int\limits_{-\omega_{\max}}^{+\omega_{\max}}e^{\mathrm{i}\omega\left(t-n\frac{2\pi}{k\omega_{\max}}\right)}\mathrm{d}\omega&=\frac{1}{\mathrm{i}\left(t-n\frac{2\pi}{k\omega_{\max}}\right)}\left[e^{\mathrm{i}\omega\left(t-n\frac{2\pi}{k\omega_{\max}}\right)}\right]\Bigg|_{-\omega_{\max}}^{+\omega_{\max}}\\
&=\frac{1}{\mathrm{i}\left(t-n\frac{2\pi}{k\omega_{\max}}\right)}\left[e^{\mathrm{i}\omega_{\max}\left(t-n\frac{2\pi}{k\omega_{\max}}\right)}-e^{-\mathrm{i}\omega_{\max}\left(t-n\frac{2\pi}{k\omega_{\max}}\right)}\right]\\
&=\frac{\omega_{\max}}{t\omega_{\max}-n\frac{2\pi}{k}}\left[\frac{e^{\mathrm{i}\left(t\omega_{\max}-n\frac{2\pi}{k}\right)}-e^{-\mathrm{i}\left(t\omega_{\max}-n\frac{2\pi}{k}\right)}}{\mathrm{i}}\right]\\
&=\frac{\omega_{\max}}{t\omega_{\max}-n\frac{2\pi}{k}}2\sin\left(t\omega_{\max}-n\frac{2\pi}{k}\right)\\
&=2\omega_{\max}\mathrm{sinc}\left(t\omega_{\max}-n\frac{2\pi}{k}\right)
\end{aligned}\tag{16.3.318}$$

Therefore:

$$\begin{aligned}\mathcal{F}^{-1}\left(F(\omega)\right)(t)=s(t)&=\frac{1}{2\omega_{\text{max}}}\sum_{n=-\infty}^{+\infty}s\left(\frac{nT_{\text{max}}}{k}\right)\left(\int\limits_{-\omega_{\text{max}}}^{+\omega_{\text{max}}}e^{\mathrm{i}\omega\left(t-n\frac{2\pi}{k\omega_{\text{max}}}\right)}\mathrm{d}\omega\right)\\&=\frac{1}{2\omega_{\text{max}}}\sum_{n=-\infty}^{+\infty}s\left(\frac{nT_{\text{max}}}{k}\right)\left(2\omega_{\text{max}}\text{sinc}\left(t\omega_{\text{max}}-n\frac{2\pi}{k}\right)\right)\\&=\sum_{n=-\infty}^{+\infty}s\left(\frac{nT_{\text{max}}}{k}\right)\text{sinc}\left(t\omega_{\text{max}}-n\frac{2\pi}{k}\right)\\&=\sum_{n=-\infty}^{+\infty}s\left(\frac{n}{kf_{\text{max}}}\right)\text{sinc}\left(t2\pi f_{\text{max}}-n\frac{2\pi}{k}\right)\\&=\sum_{n=-\infty}^{+\infty}s\left(\frac{n}{kf_{\text{max}}}\right)\text{sinc}\left(2\pi f_{\text{max}}\left(t-\frac{n}{kf_{\text{max}}}\right)\right)\end{aligned}\tag{16.3.319}$$

Hence:

$$s(t)=\sum_{n=-\infty}^{+\infty}s\left(\frac{n}{kf_{\text{max}}}\right)\text{sinc}\left(2\pi f_{\text{max}}\left(t-\frac{n}{kf_{\text{max}}}\right)\right)\tag{16.3.320}$$

In technical words we say that: By time-frequency symmetry, we have equally a Fourier series expansion of a band-limited Fourier transform!

Let us focus on the sinc term and plot if in Maple™ 4.00 for $k=1,k=2,k=3$ and three different pulses (we choose $f_{\text{max}}=50$ but the value of the frequency doesn't change at all the conclusion):

```
>with(plots):
>k:=1;
>plot([sin(2*Pi*50*t)/(2*Pi*50*t),sin(2*Pi*50*(t-1/(k*50)))/(2*Pi*50*(t-1/(k*50))),
sin(2*Pi*50*(t-2/(k*50)))/(2*Pi*50*(t-2/(k*50)))],t=-0.1..0.1);
>k:=2;
>plot([sin(2*Pi*50*t)/(2*Pi*50*t),sin(2*Pi*50*(t-1/(k*50)))/(2*Pi*50*(t-1/(k*50))),
sin(2*Pi*50*(t-2/(k*50)))/(2*Pi*50*(t-2/(k*50)))],t=-0.1..0.1);
>k:=3;
>plot([sin(2*Pi*50*t)/(2*Pi*50*t),sin(2*Pi*50*(t-1/(k*50)))/(2*Pi*50*(t-1/(k*50))),
sin(2*Pi*50*(t-2/(k*50)))/(2*Pi*50*(t-2/(k*50)))],t=-0.1..0.1);
```

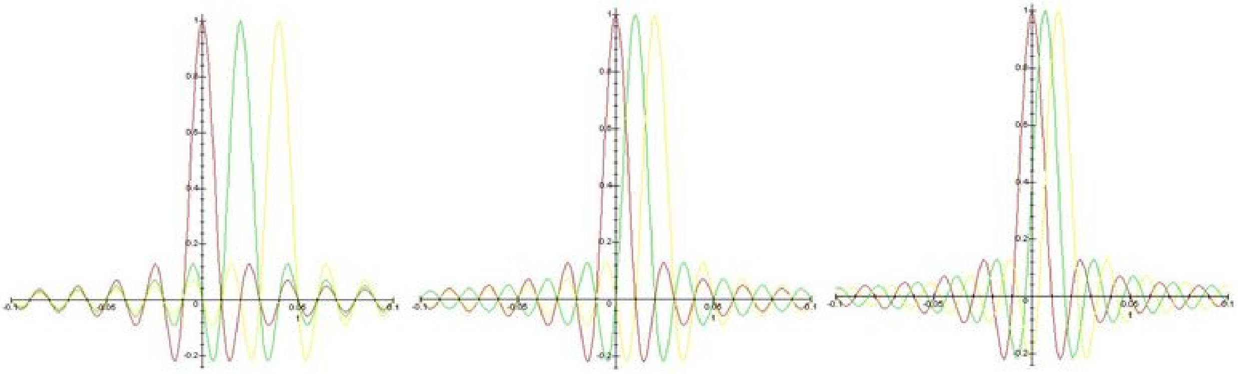
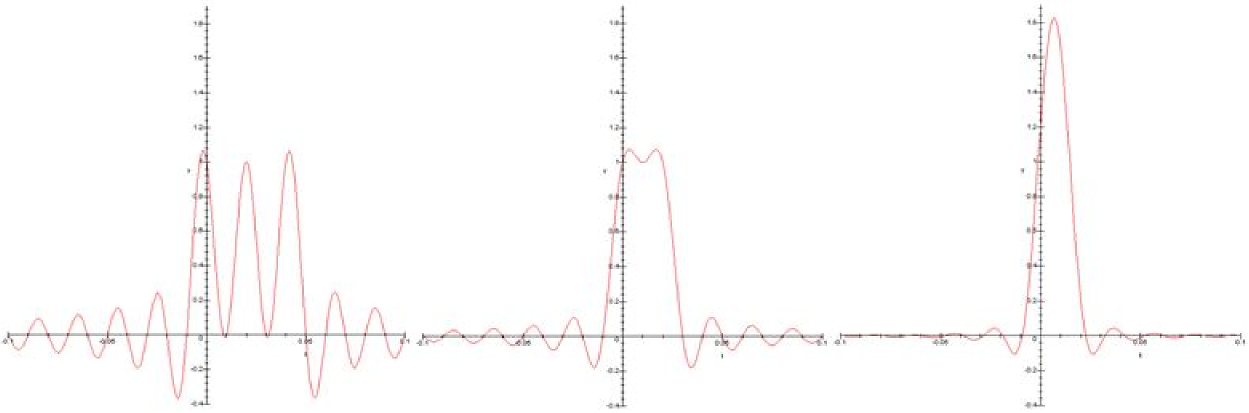


Figure 16.123: A Nyquist sinc pulses for $k=1,2,3$ and $f_{\text{max}}=50$ in Maple™ 4.00b

We clearly see that it is the middle one ($k=2$) that satisfies the previously criteria of the Nyquist sine pulse. The sum of the three pulses gives respectively:



So with the plots we see that we need $k=2$ (in other words the sampling frequency must be twice the frequency of the highest frequency term of the Fourier series) to fall back on the Nyquist and Shannon best practice recommendation for perfect periodic signals sent original by square pulse signal in the frequency domain. Hence the final result:

$$s(t)=\sum_{n=-\infty}^{+\infty} s\left(\frac{n}{2f_{\text{max}}}\right)\operatorname{sinc}\left(2\pi f_{\text{max}}\left(t-\frac{n}{2f_{\text{max}}}\right)\right)=\sum_{n=-\infty}^{+\infty} s\left(\frac{n\pi}{\omega_{\text{max}}}\right)\operatorname{sinc}\left(\omega_{\text{max}}\left(t-\frac{\pi n}{\omega_{\text{max}}}\right)\right) \tag{16.3.321}$$

Thus after simplification:

$$\boxed{s(t)=\sum_{n=-\infty}^{+\infty} s\left(\frac{n\pi}{\omega_{\text{max}}}\right)\operatorname{sinc}\left(\omega_{\text{max}}t-n\pi\right)} \tag{16.3.322}$$

So the reader must keep in mind that is, in general, the value k is not uniquely determined for $s(t)$, however, sufficient smoothness, e.g., continuity, implies its unique definition.

■ Q.E.D.

16.3.7 Semiconductor devices

16.3.7.1 Diodes

We will deal now with pn-junctions that are of importance for several modern (120th and 121th century according to holocene calendar) electronic devices like diodes (e.g. optical detectors, solar cells, light emitting diodes, LASER diodes), bipolar transistors, thyristor and field effect transistors. In the following the operating principle of pn-junctions will be discussed starting from the semiconductor materials. The ideal diode equation will be derived. Furthermore, an equivalent circuit for pn-junctions will be presented.

💣Caution! We are actually very not satisfied of what we wrote below. The text is sometimes confusing, and the series of successive approximations of the mathematical models don't help and should be highlighted even more that it already is! Lot more of figures and remarks should also very likely be added to make the whole more comprehensible. So we apologize in advance. If anyone want to help us to improve what follows, he or she is welcome to contact us.

The main characteristic of a pn-junction, or a diode, is that for positive voltages ("forward bias") the diode is conductive, whereas for negative voltages ("reverse bias") the current flow is blocked. Due to this properties a diode can be used for several applications like rectifiers, sampling circuits, demodulators, logic gates and so on.

Remark

For the readers interested in further reading, we strongly recommend the following references: [523] and [508] and finally [553].

A silicon pn-junction can be realized by the formation of a junction between an n-type and a p-type doped region. The junction can be formed by the implantation or diffusion of dopants in a wafer. For example a silicon pn-junction can be formed by the implantation of Boron (B) atoms in a n-type Silicon wafer.

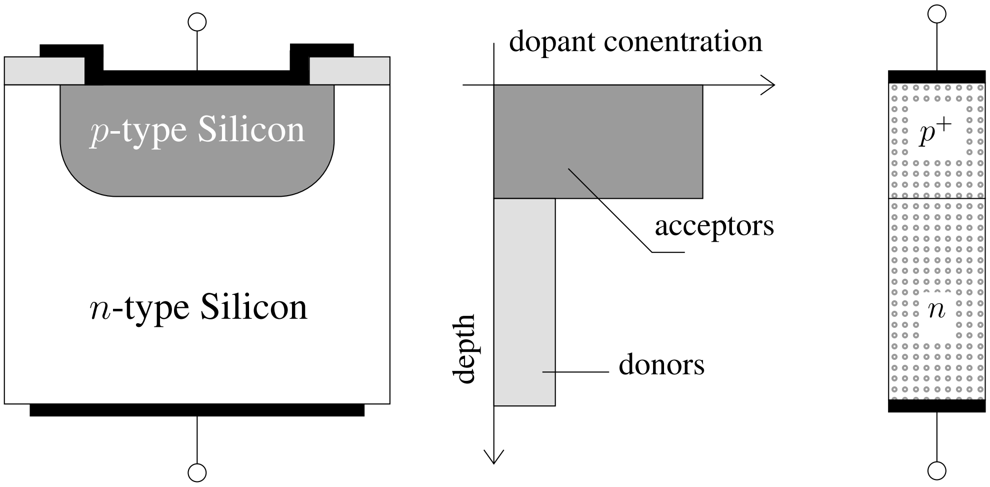
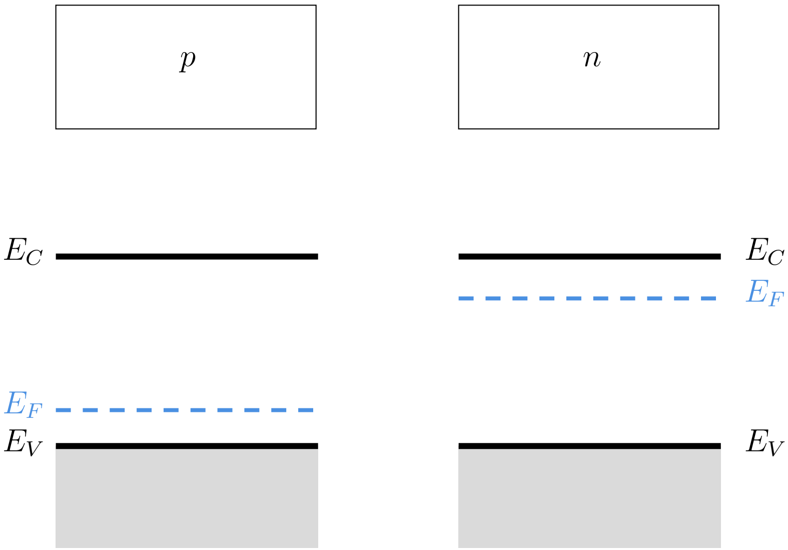


Figure 16.124: Silicon pn-junction: (left) Fabricated by the implantation or diffusion of boron atoms into an n-type (phosphorus doped) silicon wafer. (middle) Concentration of donors and acceptors in the wafer. (right) Cross section of a pn-junction.

In order to understand the operating principle of a pn-junction we will start with two separate n-type and p-type semiconductors. For the p-type semiconductor the Fermi level (see section Wave Quantum Physics page 4807) is close to the valence band, whereas for the n-type semiconductor the Fermi level is close to the conduction band:



In our case we assume that the semiconducting materials are silicon and the band gaps for both pieces of materials are equal. In this case we speak about a "homo-junction".

As the Fermi level in both pieces of silicon is constant, the semiconductors are in thermal equilibrium.

When the p-type and the n-type semiconductors are jointed together a carrier concentration gradient is observed at the interface of the n- and p-region. The gradient leads to the diffusion of carriers. The diffusion occurs because the concentration of electrons in the n-type semiconductor is much higher than the concentration of electrons in the p-type semiconductor. The opposite behaviour is observed for the holes. The hole concentration in the p-type semiconductor is much higher than the hole concentration in the n-type semiconductor. Consequently, holes from the p-side diffuse towards the $n$side and vice versa. Despite the flow of electrons and holes the device is still in thermal equilibrium.

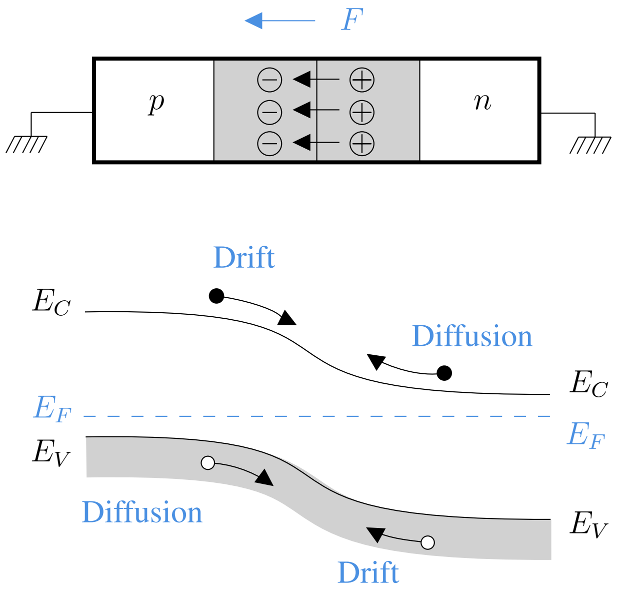
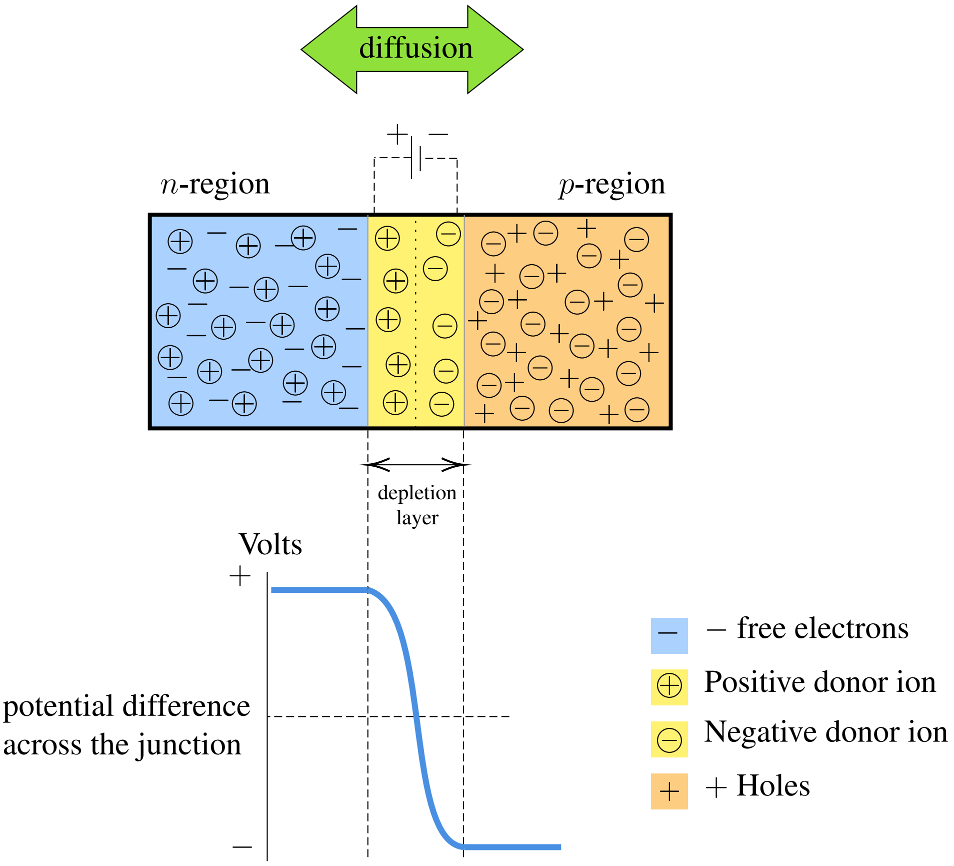

F
p
n
Drift
E_C
Diffusion
E_C
E_F
E_F
E_V
Diffusion
E_V
Drift

Figure 16.125: **Formation of a space charged region and energy band diagram of a pn-junction in thermal equilibrium**

In a more colourful way:


diffusion
+ −
n-region
p-region
depletion layer
Volts
+
potential difference across the junction
−
− free electrons
Positive donor ion
Negative donor ion
+ Holes

16.3.7.1.1 Ideal Diode Equation

Before entering a quantitative analysis of the diode characteristics, it is useful to review the relationships that determine carrier concentrations. To avoid confusion with exponential functions the symbol q_e will be used for the electronic charge instead of e.

16.3.7.1.1.1 Carrier concentrations in pure semiconductors

In thermal equilibrium the probability that an electron state in the conduction band is filled is given by the Fermi–Dirac distribution (see section Statistical Mechanics page 3571):

$$f_{FD}(E)=\frac{1}{e^{(E-E_F)/kT}+1} \tag{16.3.323}$$

The parameter E_F is as we know the Fermi level energy (or chemical potential). The density of atoms in a Si or Ge crystal is about $5\cdot 10^{22}\,[\text{atoms}\cdot\text{cm}^{-3}]$. Since the minimum carrier density of interest in practical devices is of order 10^{10} to $10^{11}\,[\text{cm}^{-3}]$, very small occupancies are quite important.

In silicon the band gap is $1.12\,[\text{eV}]$. If the Fermi level is at midgap, the band-edges will be $0.56\,[\text{eV}]$ above and below E_F. Relatively large deviations from the Fermi level, i.e. extremely small occupancies, will still yield significant carrier densities.

The number of occupied electron states N_e is determined by summing over all available states multiplied by the occupation probability for each individual state:

$$N_e=\sum_i p_i f(E_i) \tag{16.3.324}$$

Since the density of states near the band edge tends to be quite high, this can be written as an integral:

$$N_e=\int\limits_{E_c}^{+\infty} f(E)g(E)\mathrm{d}E \tag{16.3.325}$$

where $g(E)$ is the density of states. Solution of this integral requires knowledge of the density of states. Fortuitously, to a good approximation the density of states near the band edge has a parabolic distribution:

$$g(E)\mathrm{d}E\propto(E-E_C)^{1/2} \tag{16.3.326}$$

As the energy increases beyond the band edge, the distribution will deviate from the simple parabolic form, but since the probability function decreases very rapidly, the integral will hardly be affected.

The second obstacle to a simple analytical solution of the integral is the intractability of integrating over the Fermi distribution (at least outside the simple case where the bounds of the integral range from 0 to $+\infty$). Fortunately, if $E-E_F$ is at least several times kT, the Fermi distribution can be approximated by a Boltzmann distribution. Indeed, we have:

$$1+e^{(E-E_F)/kT}\cong e^{(E-E_F)/kT} \tag{16.3.327}$$

Therefore:

$$f_{FD}(E) \cong \frac{1}{e^{(E-E_F)/kT}} = f_B(E) = e^{-(E-E_F)/kT} \tag{16.3.328}$$

At energies $2.3\,kT$ beyond the Fermi level the difference between the Boltzmann approximation and the Fermi distribution is $< 10\%$, for energies $> 4.5\,kT$ it is less than 1%.

Applying the approximation to the occupancy of hole states, the probability of a hole state being occupied, i.e. a valence state being empty, is:

$$f_h(E) = 1 - f_e(E) = \frac{1}{e^{(E_F-E)/kT}+1} \cong e^{-(E_F-E)/kT} \tag{16.3.329}$$

The conditions for the Boltzmann approximation are fulfilled for excitation across the band gap, as the band gap is of order $1\,[\text{eV}]$ and kT at room temperature is 0.026 [eV].

With these simplifications the number of electrons in the conduction band in thermal equilibrium is given by (see page 4296):

$$n_e \propto (kT)^{3/2} e^{-(E_c-E_F)/kT} \tag{16.3.330}$$

or:

$$n_e = N_C e^{-(E_C-E_F)/kT} \tag{16.3.331}$$

where N_C is the effective density of states at the band edge. Correspondingly, the hole concentration is given and denoted by:

$$p = N_V e^{-(E_F-E_V)/kT} \tag{16.3.332}$$

Under thermal equilibrium the quasi Fermi levels for electrons and holes are identical, so that the Fermi potential is identical for electrons and holes.

In an ideal semiconductor the only source of mobile carriers is thermal excitation across the band gap (additional impurity atoms or crystal imperfections that would allow other excitation mechanisms are absent), so the concentrations of electrons and holes are equal:

$$n = p = n_i \tag{16.3.333}$$

n_i is named the "intrinsic carrier concentration". As we have already seen it in a table (see section Electrokinetis page 4303), in Silicium ($E_g = 1.12\,[\text{eV}]$) the intrinsic concentration $n_i = 1.45 \cdot 10^{10}\,[\text{cm}^{-3}]$ at 300 [K] and in Germanium ($E_g = 0.66\,[\text{eV}]$), $n_i = 2.4 \cdot 10^{13}\,[\text{cm}^{-3}]$. For comparison, the purest semiconductor material that has been fabricated is Ge with active impurity levels of about $3 \cdot 10^{10}\,[\text{cm}^{-3}]$ (these numbers must be compared to the fact that such solids have a density of atoms of the order of $10^{22}\,[\text{cm}^{-3}]$).

Using the above results:

$$n_i = N_C e^{-(E_C-E_F)/kT} = N_V e^{-(E_F-E_V)/kT} \tag{16.3.334}$$

which one can solve to get:

$$E_F = E_i = \frac{E_C+E_V}{2} - \frac{kT}{2}\ln\left(N_c/N_V\right) \tag{16.3.335}$$

If the band structure is symmetrical ($N_C = N_V$), the intrinsic energy level E_i lies near the middle of the band gap. Even rather substantial deviations from a symmetrical band structure will not affect this result significantly, as N_c/N_V enters logarithmically and kT is much smaller than the band gap.

A remarkable result is that the product of the quasi-electron and quasi-hole concentrations, named the "law of the junction":

$$\boxed{np = n_i^2 = N_C N_V e^{-(E_C - E_V)/kT} = N_C N_V e^{-E_g/kT}} \tag{16.3.336}$$

depends only on the band gap E_g and not on the Fermi level.

Remark

This relation can also be found in the following (ambiguous) form in many textbooks:

$$np = n_i^2 e^{-\left(E_F^{(n)} - E_F^{(p)}\right)/kT} = n_i^2 e^{qU/kT} \tag{16.3.337}$$

This result, the law of mass action, is very useful in semiconductor device analysis. It requires only that the Boltzmann approximation holds! Qualitatively, it says that if one carrier type exceeds this equilibrium concentration, recombination will decrease the concentrations of both electrons and holes to maintain $np = {n_i}^2$, a relationship that also holds in doped crystals.

16.3.7.1.1.2 Carrier concentrations in doped crystals

The equality $n_e = n_h$ only holds for pure crystals, where all of the electrons in the conduction band have been thermally excited from the valence band. In practical semiconductors the presence of impurities tips the balance towards either electrons or holes.

Impurities are an unavoidable byproduct of the crystal growth process, although special techniques can achieve astounding results. For example, as noted above, in the purest semiconductor crystals – "ultrapure" Ge – the net impurity concentration is about $3 \cdot 10^{10}\;[\text{cm}^{-3}]$.

In semiconductor device technology impurities are introduced intentionally to control the conductivity of the semiconductor. Let ${N_d}^+$ be the concentration of ionized donors and ${N_a}^-$ the concentration of ionized acceptors. Overall charge neutrality is preserved, as each ionized dopant introduces a charged carrier and an oppositely charged atom, but the net carrier concentration is now:

$$\Delta n = n - p = N_d^+ - N_a^- \tag{16.3.338}$$

or:

$$p + N_d^+ = n + N_a^- \tag{16.3.339}$$

Let us assume that the activation energy of the donors and acceptors is sufficiently small so that they are fully ionized. Then ${N_d}^+ = N_d$ and ${N_a}^- = N_a$, so:

$$p + N_d = n + N_a \tag{16.3.340}$$

which, using $np = {n_i}^2$, becomes:

$$\begin{aligned} p + N_d &= \frac{n_i^2}{p} + N_a \\ \frac{p}{N_a} + \frac{N_d}{N_a} &= \frac{n_i}{p}\frac{n_i}{N_a} + 1 \end{aligned} \tag{16.3.341}$$

If the acceptor concentration $N_a \gg N_d$ and $N_a \gg n_i$, the hole and electron concentrations:

$$p \approx N_a \quad \text{and} \quad n \approx \frac{n_i^2}{N_a} \ll N_a \tag{16.3.342}$$

i.e. the conductivity is dominated by holes. Conversely, if the donor concentration $N_d \gg N_a$ and $N_d \gg n_i$ the conductivity is dominated by electrons such that:

$$n \approx N_d \quad \text{and} \quad n \approx \frac{n_i^2}{N_d} \ll N_d \tag{16.3.343}$$

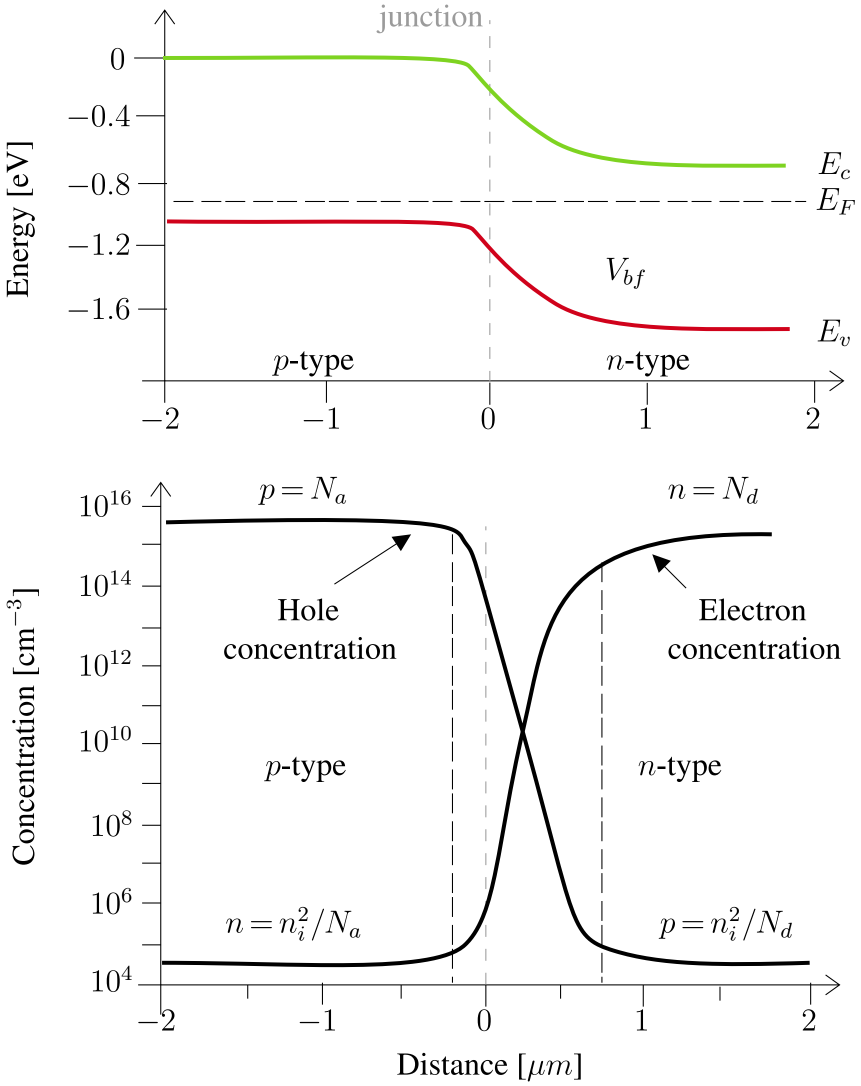


Figure 16.126: **Energy band diagram and carrier concentration for a silicon pn-junction at zero bias**

If the conductivity is dominated by only one type of carrier, the Fermi level is easy to determine. If, for example, $n \gg p$ then the Boltzmann approximation can be written as:

$$n = N_d - N_a = N_C e^{-(E_C - E_F)/kT} \tag{16.3.344}$$

yielding:

$$\frac{E_C - E_F}{kT} = \ln\left(\frac{N_C}{N_d - N_a}\right) \tag{16.3.345}$$

If $N_d \gg N_a$, then $E_C - E_F$ must be small, i.e. the Fermi level lies close to the conduction band edge.

In reality the impurity levels of common dopants are not close enough to the band edge for the Boltzmann approximation to hold, so the calculation must use the Fermi distribution and solve numerically for E_F. Nevertheless, the qualitative conclusions derived here still apply.

It is often convenient to refer all of these quantities to the intrinsic level E_i, as it accounts for both E_C and E_V. Then:

$$\begin{aligned} n &= N_C e^{-(E_C - E_F)/kT} = n_i e^{-(E_F - E_i)/kT} \\ p &= N_V e^{-(E_F - E_V)/kT} = n_i e^{-(E_i - E_F)/kT} \end{aligned} \tag{16.3.346}$$

Obviously the reader must keep in mind that all this lead from as series of successive approximations!

16.3.7.1.1.3 pn-junctions

A pn-junction is formed at the interface of a p- and an n-type region. Since the electron concentration in the n-region is greater than in the p-region, electrons will diffuse into the p-region. Correspondingly, holes will diffuse into the n-region. As electrons and holes diffuse across the junction, a space charge due to the ionized donor and acceptor atoms builds up. The field due to this space charge is directed to impede the flow of electrons and holes.

The situation is dynamic. The concentration gradient causes a continuous diffusion current to flow, whereas the field due to the space charge drives a drift current in the opposite direction. Equilibrium is attained when the two currents are equal, i.e. the sum of the diffusion current (caused by the gradient of the carrier concentration) and drift current (caused by the electric field) is zero. The net hole current density is the sum of the diffusion current density (see section Thermodynamics page 3699) and of the drift current density (see section Electrokinetics page 4229):

$$J_p = J_{\text{diff}} + J_{\text{drift}} = -q_e D_p \frac{\mathrm{d}p}{\mathrm{d}x} + q_e p \mu_p E_p = q_e\left(D_p \frac{\mathrm{d}p}{\mathrm{d}x} - p\mu_p E_p\right) \tag{16.3.347}$$

where D_p is the diffusion constant for holes and E_p is the electric field in the p-region, p is the holes density and μ_p the holes mobility.

Remark

Remember that first Fick's law (i.e. "diffusion equation" or "transport equation" as derived at page 3698) will be written:

$$J_{p,\text{diff}}=-q_e D_p\frac{\mathrm{d}p}{\mathrm{d}x} \qquad J_{n,\text{diff}}=q_e D_p\frac{\mathrm{d}n}{\mathrm{d}x} \tag{16.3.348}$$

where the first diffusion equation is also often denoted:

$$J_{p,\text{diff}}=-q_e D_p\frac{\mathrm{d}n_p}{\mathrm{d}x} \tag{16.3.349}$$

At zero bias the drift current (caused by the electric field) and the diffusion current (caused by the gradient of the carrier concentration) compensate each other, therefore:

$$J_p=q_e\left(D_p\frac{\mathrm{d}p}{\mathrm{d}x}-p\mu_p E_p\right)=0 \tag{16.3.350}$$

To solve this equation we make use of the following relationships that gives the hole concentration:

$$p=n_i e^{-(E_i-E_F)/kT} \tag{16.3.351}$$

and therefore if we derive it relatively to x we get:

$$\frac{\mathrm{d}p}{\mathrm{d}x}=n_i e^{-(E_i-E_F)/kT}\frac{-\left(\frac{\mathrm{d}E_i}{\mathrm{d}x}-\frac{\mathrm{d}E_F}{\mathrm{d}x}\right)}{kT}=-\frac{p}{kT}\left(\frac{\mathrm{d}E_i}{\mathrm{d}x}-\frac{\mathrm{d}E_F}{\mathrm{d}x}\right) \tag{16.3.352}$$

Since the force on a charge q_e due to an electric field E is equal to the negative gradient of the potential energy:

$$q_e E=-\frac{\mathrm{d}E_C}{\mathrm{d}x}=-\frac{\mathrm{d}E_V}{\mathrm{d}x}=-\frac{\mathrm{d}E_i}{\mathrm{d}x} \tag{16.3.353}$$

As only the gradient is of interest and E_C, E_V, and E_i differ only by a constant offset, any of these three measures can be used. We'll use the intrinsic Fermi level E_i, since it applies throughout the sample.

Therefore the first term in the differential equation above can be written:

$$\frac{\mathrm{d}p}{\mathrm{d}x}=\frac{p}{kT}\left(\frac{\mathrm{d}E_F}{\mathrm{d}x}+q_e E_p\right) \tag{16.3.354}$$

The remaining ingredient is the Einstein relationship, which relates the mobility to the diffusion constant:

$$\mu_p=\frac{q_e D_p}{kT} \tag{16.3.355}$$

Using these relationships the net hole current becomes:

$$\begin{aligned}J_p&=q_e\left(D_p\frac{\mathrm{d}p}{\mathrm{d}x}-p\mu_p E_p\right)=q_e\left(D_p\frac{p}{kT}\left(\frac{\mathrm{d}E_F}{\mathrm{d}x}+q_e E_p\right)-p\frac{q_e D_p}{kT}E_p\right)=q_e p\frac{D_p}{kT}\frac{\mathrm{d}E_F}{\mathrm{d}x}\\&=\mu_p p\frac{\mathrm{d}E_F}{\mathrm{d}x}=0\end{aligned} \tag{16.3.356}$$

Accordingly, the net electron current is given by:

$$J_n = -q_e n\frac{D_n}{kT}\frac{\mathrm{d}E_F}{\mathrm{d}x} = -\mu_n n\frac{\mathrm{d}E_F}{\mathrm{d}x} = 0 \tag{16.3.357}$$

Since, individually, the net hole and electron currents in equilibrium must be zero, the derivative of the Fermi level is then:

$$\frac{\mathrm{d}E_F}{\mathrm{d}x} = 0 \tag{16.3.358}$$

In thermal equilibrium the Fermi level must be constant throughout the junction region.

For the Fermi level to be flat, the band structure must adapt, since on the p-side the Fermi level is near the valence band, whereas on the n-side it is near the conduction band. If we assume that the dopants are exclusively donors on the n-side and acceptors on the p-side, the difference in the respective Fermi levels is:

$$\Delta E_F = kT\ln\left(\frac{N_a N_d}{n_i^2}\right) \tag{16.3.359}$$

Indeed, to prove this relation remember that we have (we changed a bit the notations):

$$N_d = n = n_i e^{-(E_F - E_{i,n})/kT} \qquad N_a = p = n_i e^{-(E_{i,p} - E_F)/kT} \tag{16.3.360}$$

Extracting the intrinsic energy we get:

$$E_{i,n} = kT\ln\left(\frac{N_d}{n_i}\right) + E_F \qquad E_{i,p} = -kT\ln\left(\frac{N_a}{n_i}\right) + E_F \tag{16.3.361}$$

Therefore:

$$\Delta E_F = E_{i,n} - E_{i,p} = kT\left(\frac{N_d}{n_i} + \frac{N_a}{n_i}\right) = kT\ln\left(\frac{N_a N_d}{n_i^2}\right) \tag{16.3.362}$$

This corresponds to an electric potential:

$$\boxed{\Delta U_F = \frac{1}{q_e}\Delta E_F = \frac{kT}{q_e}\ln\left(\frac{N_a N_d}{n_i^2}\right) \equiv U_{bi}\;[\mathrm{J}\cdot\mathrm{C}^{-1}]} \tag{16.3.363}$$

equivalently measured in Volts [V], often referred to as the "built-in voltage" or "contact potential" of the junction (keep in mind that in reality it is an approximation!).

As either N_a or N_d increases relative to n_i, the respective Fermi level moves closer to the band edge, increasing the built-in voltage. With increasing doping levels the built-in voltage approaches the equivalent potential of the band gap E_g/q_e.

Remark

Notice that we can rearrange to get:

$$N_a = \frac{n_i^2}{N_d}e^{q_e\Delta U_F/kT} = N_{a0}e^{q_e\Delta U_F/kT} \quad \text{or} \quad N_d = \frac{n_i^2}{N_a}e^{q_e\Delta U_F/kT} = N_{d0}e^{q_e\Delta U_F/kT} \tag{16.3.364}$$

These two relations will be useful to us further below.

16.3.7.1.1.4 Forward-biased pn-junction

Applying an external bias leads to a condition that deviates from thermal equilibrium, i.e. the Fermi level is no longer constant throughout the junction. If a positive voltage is applied to the p-electrode relative to the n-electrode, the total variation of the electric potential across the junction will decrease. Since this reduces the electric field across the junction (i.e. the build-in tension), the drift component of the junction current will decrease. Since the concentration gradient is unchanged, the diffusion current will exceed the drift current and a net current will flow.

This net current leads to an excess of electrons in the p-region and an excess of holes in the n-region. This "injection" condition leads to a local deviation from equilibrium, i.e. $pn > {n_i}^2$. Equilibrium will be restored by recombination.

Note that a depletion region exists even under forward bias, although its width is decreased. The electric field due to the space charge opposes the flow of charge, but the large concentration gradient overrides the field.

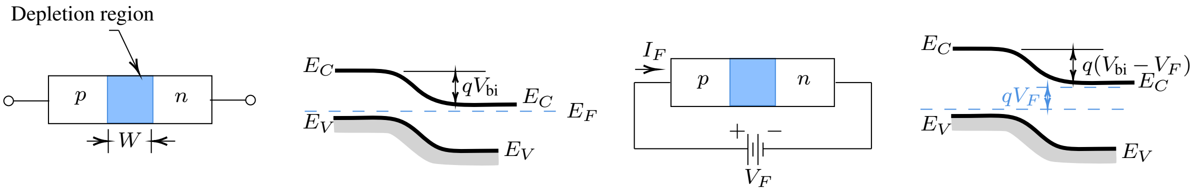


Figure 16.127: Depletion layer width and energy diagram of a pn-junction under two different biasing conditions. (left) Thermal equilibrium (right) Forward bias

Consider holes flowing into the n-region. They will flow through the depletion region with small losses due to recombination, as the electron concentration is small compared with the bulk. When holes reach the n-side boundary of the depletion region the concentration of electrons available for recombination increases and the concentration of holes will decrease with distance, depending on the cross-section for recombination, expressed as a diffusion length. Ultimately, all holes will have recombined with electrons. The required electrons are furnished through the external contact from the power supply.

On the p-side, electrons undergo a similar process. The holes required to sustain recombination are formed at the external contact to the p-region by electron flow toward the power supply, equal to the electron flow toward the n-contact. The following derivation follows the discussions by Shockley (11949, 11950 according to holocene calendar) and Grove (11967 according to holocene calendar).

The steady-state distribution of charge is determined by solving the diffusion equation (see section Thermodynamics page 3688):

$$D_n \frac{\mathrm{d}^2 n_p}{\mathrm{d}x^2} - \frac{n_p - n_{p0}}{\tau_n} = 0 \qquad (16.3.365)$$

where the parameter τ_n is the electron lifetime.

Electrons flowing into the p-region give rise to a local concentration n_p in excess of the equilibrium concentration n_{p0}. This excess will decay with a recombination time τ_n, corresponding to a diffusion length L_n.

Remark

Again, the above relation can often by found in textbooks with the following notation:

$$D_p\frac{\mathrm{d}^2p_n}{\mathrm{d}x^2}-\frac{p_n-p_{n0}}{\tau_{pl}}=0 \tag{16.3.366}$$

The first boundary condition required for the solution of the diffusion equation is that the excess concentration of electrons vanish at large distances x:

$$n_p(x\rightarrow+\infty)=n_{p0} \tag{16.3.367}$$

The second boundary condition is that the carriers are injected at the origin of the space charge region $x=0$ with a concentration $n_p(0)$. This yields the solution:

$$n_p(x)=n_{p0}+(n_p(0)-n_{p0})\,e^{-x/\sqrt{D_n\tau_n}} \tag{16.3.368}$$

often written:

$$n_p(x)-n_{p0}=(n_p(0)-n_{p0})\,e^{-x/L_n} \tag{16.3.369}$$

with $L_n=\sqrt{D_n\tau_n}$ named the "electron diffusion length ".

Remarks

R1. The Diffusion equation can be solved by using the following general solution:

$$n_p(x)-n_{p0}=A\exp\left(\frac{x}{\sqrt{D_n\tau_n}}\right)+B\exp\left(-\frac{x}{\sqrt{D_n\tau_n}}\right)=A\exp\left(\frac{x}{L_n}\right)+B\exp\left(-\frac{x}{L_n}\right) \tag{16.3.370}$$

and applying both boundary conditions mentioned earlier we obviously fall back on the solution above!

R2. The differential equation above and it's corresponding solution are a special case named the "short diode case". The "long" diode expression applies to p-n diodes in which recombination/generation occurs in the quasi-neutral region only. This is the case if the quasi-neutral region is much larger than the carrier diffusion length. The "short" diode expression applies to p-n diodes in which recombination/generation occurs at the contacts only. In a short diode, the quasi-neutral region is much smaller than the diffusion length. In the long diode case the solution can be found in some textbooks in the following shape:

$$n_p(x)-n_{p0}=A\exp\left(\frac{x-x_n}{\sqrt{D_n\tau_n}}\right)+B\exp\left(-\frac{x-x_n}{\sqrt{D_n\tau_n}}\right)=A\exp\left(\frac{x-x_n}{L_n}\right)+B\exp\left(-\frac{x-x_n}{L_n}\right) \tag{16.3.371}$$

and applying both boundary conditions mentioned earlier we obviously fall back on the solution above!

From this we obtain the electron current (see section Thermodynamics page 3701) entering the p-region :

$$J_{np}=-q_eD_n\left.\frac{\mathrm{d}n_p}{\mathrm{d}x}\right|_{x=0}=q_eD_n\frac{n_p(0)-n_{p0}}{L_n} \tag{16.3.372}$$

This says that the electron current is limited by the concentration gradient determined by the carrier density at the depletion edge $n_p(0)$ and the equilibrium minority carrier density n_{p0}.

Determining the equilibrium density n_{p0} is easy as we know:

$$n_{p0}=n_i^2/N_a \tag{16.3.373}$$

The problem is that $n_p(0)$ is established in a non-equilibrium state, where the previously employed results do not apply!

To analyse the regions with non-equilibrium carrier concentrations Shockley introduced a simplifying assumption by postulating that the product pn is constant. In this specific quasi-equilibrium state this constant will be larger than ${n_i}^2$, the pn-product in thermal equilibrium. In analogy to thermal equilibrium, this quasi-equilibrium state is expressed in terms of a "quasi-Fermi level", which is the quantity used in place of E_F that gives the carrier concentration under non-equilibrium conditions.

The postulate $pn=c^{te}$ is equivalent to stating that the non-equilibrium carrier concentrations are given by a Boltzmann distribution, so the concentration of electrons is:

$$n\cong n_ie^{-(E_F^{(n)}-E_i)/kT} \tag{16.3.374}$$

where $E_F^{(n)}$ is the quasi-Fermi level for electrons, and:

$$p\cong n_ie^{-(E_i-E_F^{(p)})/kT} \tag{16.3.375}$$

where $E_F^{(p)}$ is the quasi-Fermi level for holes. The product of the two carrier concentrations in non-equilibrium is:

$$pn=n_i^2e^{-(E_F^{(n)}-E_F^{(p)})/kT} \tag{16.3.376}$$

If pn is constant throughout the space-charge region, then $E_F^{(n)}-E_F^{(p)}$ must also remain constant.

Using the quasi-Fermi level and the Einstein relationship, the electron current entering the p-region becomes (following the same developments as we did earlier above):

$$J_{np}=-q_eD_n\left.\frac{\mathrm{d}n_p}{\mathrm{d}x}\right|_{x=0}=-q_eD_n\frac{\mathrm{d}}{\mathrm{d}x}\left(n_ie^{-(E_F^{(n)}-E_i)/k_BT}\right)=-\mu_nn\frac{\mathrm{d}E_F^{(n)}}{\mathrm{d}x} \tag{16.3.377}$$

These relationships describe the behaviour of the quasi-Fermi level in the depletion region. How does this connect to the neutral region?

In the neutral regions the majority carrier motion is dominated by drift (in contrast to the injected minority carrier current, which is determined by diffusion). Consider the n-type region. Here the bulk electron current that provides the junction current is given by:

$$J_{nn}=-\mu_n n\frac{\mathrm{d}E_i}{\mathrm{d}x} \tag{16.3.378}$$

Since the two electron currents must be equal:

$$J_{nn}=J_{np} \tag{16.3.379}$$

it follows that:

$$\frac{\mathrm{d}E_F^{(n)}}{\mathrm{d}x}=\frac{\mathrm{d}E_i}{\mathrm{d}x} \tag{16.3.380}$$

i.e. the quasi-Fermi level follows the energy band variation. Thus, in a neutral region, the quasi-Fermi level for the majority carriers is the same as the Fermi level in equilibrium. At current densities small enough not to cause significant voltage drops in the neutral regions, the band diagram is flat, and hence the quasi-Fermi level is flat.

In the space charge region, pn is constant, so the quasi-Fermi levels for holes and electrons must be parallel, i.e. both will remain constant at their respective majority carrier equilibrium levels in the neutral regions.

If an external bias U is applied, the equilibrium Fermi levels are offset by U, so it follows that the quasi-Fermi levels are also offset by U:

$$E_F^{(p)}-E_F^{(n)}=q_eU \tag{16.3.381}$$

Consequently, the pn-product in non-equilibrium can be written under the form of the "law of the junction":

$$pn=n_i^2e^{-(E_F^{(n)}-E_F^{(p)})/kT}=n_i^2e^{q_eU/kT} \tag{16.3.382}$$

If the majority carrier concentration is much greater than the concentration due to minority carrier injection ("low-level injection"), the hole concentration at the edge of the p-region remains essentially at the equilibrium value. Consequently, the enhanced pn-product increases the electron concentration:

$$n_p(0)=n_{p0}e^{q_eU/kT}\Leftrightarrow\frac{n_p(0)}{n_{p0}}=e^{q_eU/kT} \tag{16.3.383}$$

Correspondingly, the hole concentration in the n-region at the edge of the depletion zone becomes:

$$p_n(0)=p_{n0}e^{q_eU/kT}\Leftrightarrow\frac{p_n(0)}{p_{n0}}=e^{q_eU/kT} \tag{16.3.384}$$

Remark

The two relations are derived from the previous remark we made earlier where for recall we derived:

$$N_a=\frac{n_i^2}{N_d}e^{q_e\Delta U_F/kT}=N_{a0}e^{q_e\Delta U_F/kT}\quad\text{and}\quad N_d=\frac{n_i^2}{N_a}e^{q_e\Delta U_F/kT}=N_{d0}e^{q_e\Delta U_F/kT}\tag{16.3.385}$$

Now remember that we get earlier above the following solution to the diffusion equation:

$$n_p(x)-n_{p0}=(n_p(0)-n_{p0})\,e^{-x/L_n}\tag{16.3.386}$$

We can now rewrite that latter:

$$n_p(x)-n_{p0}=n_{p0}\left(\frac{n_p(0)}{n_{p0}}-1\right)e^{-x/L_n}=n_{p0}\left(e^{q_eU/kT}-1\right)e^{-x/L_n}\tag{16.3.387}$$

Remembering that we have:

$$n_{p0}=\frac{n_i^2}{N_a}\quad\text{and}\quad p_{n0}=\frac{n_i^2}{N_d}\tag{16.3.388}$$

the components of the diffusion current only (not including the drift as the latter can be neglected due to the dominance of the diffusion current!) due to holes and electrons are (see section Thermodynamics page 3699):

$$\begin{aligned}J_n\cong J_{n,\text{diff}}&=-q_eD_n\left.\frac{\mathrm{d}n_e}{\mathrm{d}x}\right|_{x=0}=-qD_n\frac{\mathrm{d}}{\mathrm{d}x}\left.\left[n_{p0}\left(e^{q_eU/kT}-1\right)e^{-x/L_n}\right]\right|_{x=0}\\&=q_eD_n\frac{n_i^2}{N_aL_n}\left(e^{q_eU/kT}-1\right)\end{aligned}\tag{16.3.389}$$

The same development lead us to:

$$J_p\cong J_{p,\text{diff}}=q_eD_p\frac{n_i^2}{N_dL_p}\left(e^{q_eU/kT}-1\right)\tag{16.3.390}$$

The -1 has an influence only when J is very small. We can interpret his presence by the need of having $J=0$ when $U=0$.

The total diffusion current is the sum of the electron and hole components:

$$J=J_n+J_p\cong J_0(e^{q_eU/kT}-1)\tag{16.3.391}$$

where:

$$J_0=q_en_i^2\left(\frac{D_n}{N_aL_n}+\frac{D_p}{N_dL_p}\right)\tag{16.3.392}$$

is named the "saturation density current".

Or explicitly:

$$\boxed{J=J_n+J_p=q_en_i^2\left(\frac{D_n}{N_aL_n}+\frac{D_p}{N_dL_p}\right)(e^{q_eU/kT}-1)}\tag{16.3.393}$$

This is the famous "ideal diode equation" , also named "Shockley equation", which describes the current–voltage characteristic both under forward and reverse bias.

Remark

A comparison of the "short" diode expression with the "long" diode expression reveals that they are the same except for the use of either the diffusion length or the quasi-neutral region width in the denominator, whichever is smaller.

Under forward bias the current increases exponentially. Under reverse bias (negative U), the exponential term vanishes when the bias exceeds several kT/q_e and the current becomes the reverse saturation current $J = -J_0$. For a uniform junction cross-section the current densities J_n, J_p, and J_0 can be replaced by their respective currents so that we get the famous relation known by most engineers (and a significant number of solid state physicists):

$$\boxed{I = I_0(e^{q_eU/kT} - 1)} \tag{16.3.394}$$

Or also sometimes denoted and approximated (when $U \gg U_{\text{th}}$):

$$I = I_0(e^{U/U_{\text{th}}} - 1) \cong I_0 e^{U/U_{\text{th}}} \tag{16.3.395}$$

This last relation can be interpreted in a different way. To pass a current I in a diode, we have to apply to his extremities a potential such that:

$$U \cong U_{\text{th}} \ln\left(\frac{I}{I_0}\right) \tag{16.3.396}$$

The reader must keep in mind that most relations above are not strict equalities since we get them from successive approximations!

Note that in the diode equation:

1. The band gap does not appear explicitly (only implicitly in J_0 via n_i).
2. The total current has two distinct components, due to electrons and holes.
3. The electron and hole currents are generally not equal. The ratio

$$\frac{J_n}{J_p} = \frac{N_d}{N_a} \quad \text{if} \quad \frac{D_n}{L_n} = \frac{D_p}{L_p} \tag{16.3.397}$$

4. Current flows for all values of U. However, when plotted on a linear scale, the exponential appears to have a knee, often referred to as the "turn-on voltage".
5. The magnitude of the turn-on voltage is determined by J_0. Diodes with different band gaps will show the same behaviour if J_0 is the same.

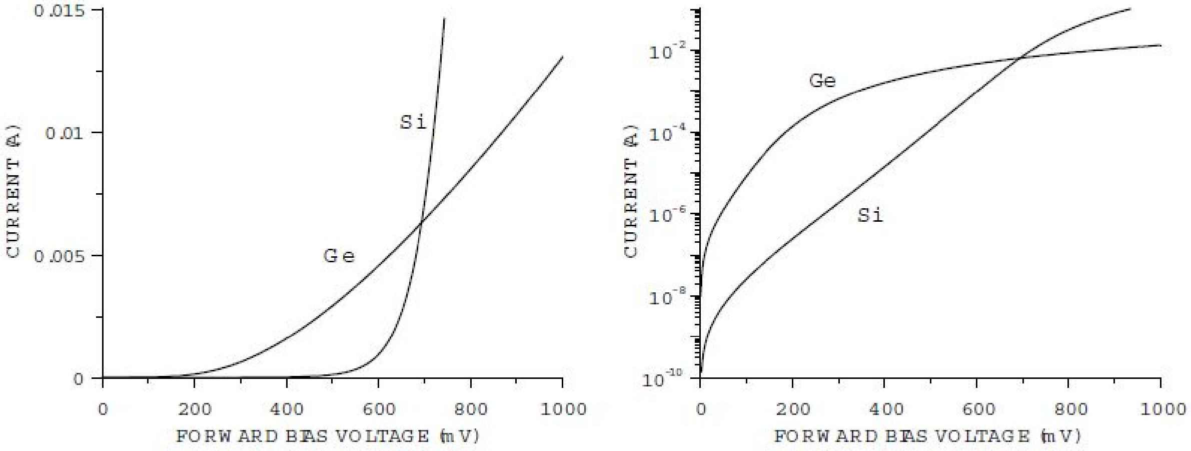


Figure 16.128: **Current vs voltage for forward biased** Si **and** Ge **diodes**

The figure above shows measured I-U curves for commercial Si and Ge junction diodes (1N4148 and 1N34A). On a linear scale the Ge diode "turns on" at $200-300\,[\text{mV}]$, whereas the Si diode has a threshold of $500-600\,[\text{mV}]$. However, on a logarithmic scale it becomes apparent that both diodes pass current at all voltages >0.

The discrepancies in the forward current between the measured results and the simple theory require the analysis of all processes in the depletion zone:

1. Generation-recombination in the depletion region
2. Diffusion current (as just calculated for the ideal diode).
3. High-injection region where the injected carrier concentration affects the potentials in the neutral regions.
4. Voltage drop due to bulk series resistance.

So although we got to the ideal diode equation there is one small problem here. It doesn't really match the experimental data! While with some germanium diodes, it works, for some silicon diodes it fails. Which is why in 11954 (holocene calendar), Bill Pfann and W. van Roosbroek, modified the equation as:

$$I=I_0(e^{q_eU/\eta kT}-1) \qquad (16.3.398)$$

Where η is the "ideality constant" to overcome this problem and balance out the side effects of approximations. For a discussion of these effects see [508].

Finally to close that topic, notice that the dynamic resistance R_d of the ideal diode is defined by:

$$R_d=\frac{\mathrm{d}U}{\mathrm{d}I}\cong\frac{\mathrm{d}}{\mathrm{d}I}\left[U_{\text{th}}\ln\left(\frac{I}{I_0}\right)\right]=\frac{U_{\text{th}}}{I}=\frac{U_{\text{th}}}{I_0}e^{-U/U_{\text{th}}} \qquad (16.3.399)$$

16.3.7.2 Amplifiers

An "amplifier" or "electronic amplifier" is an electronic device that can increase the power of a signal (a time-varying voltage or current). An amplifier functions by using electric power from a power supply to increase the amplitude of the voltage or current signal.

An amplifier can either be a separate piece of equipment or an electrical circuit contained within another device. Amplification is fundamental to modern electronics, and amplifiers are widely used in almost all electronic equipment. Amplifiers can be divided into voltage amplifiers, current amplifiers, transconductance amplifiers, and transresistance amplifiers.

We can make a very simple amplifier using the circuit shown below:

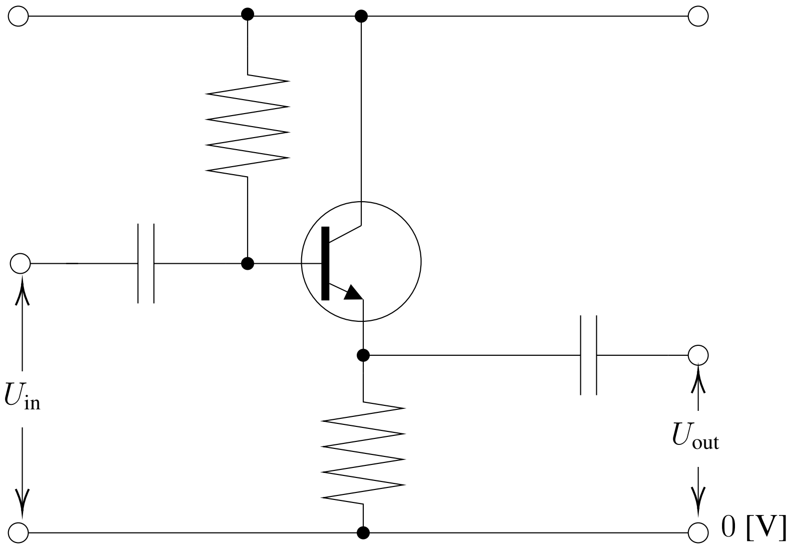


Figure 16.129: **Simple amplifier (follower)**

This circuit is named a "follower" because the output follows closely the input voltage. The voltage gain is slightly less than 1, but there is a high current gain.

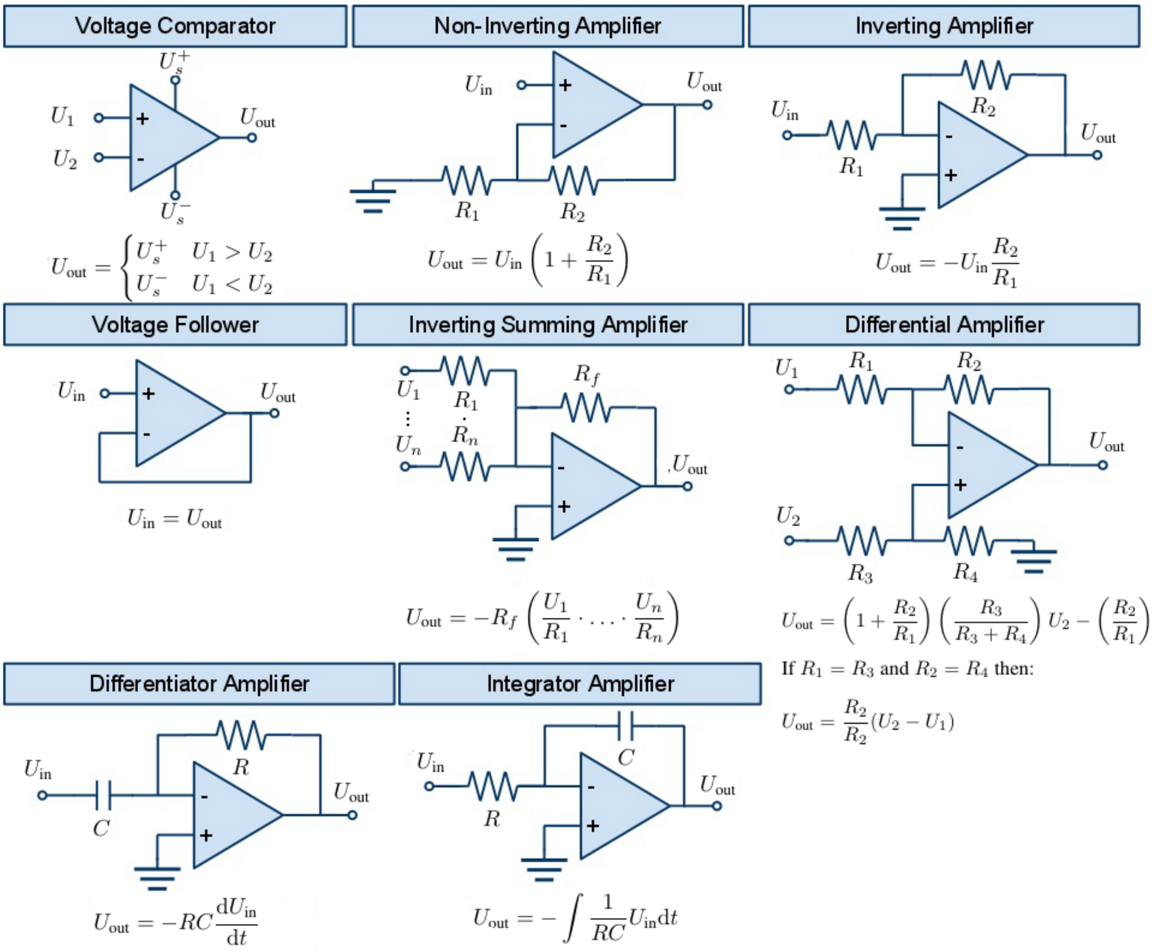

Voltage Comparator
Non-Inverting Amplifier
Inverting Amplifier
U_s^+
U_1
U_2
U_{out}
U_s^-
U_{in}
R_1
R_2
$U_{out} = \begin{cases} U_s^+ & U_1 > U_2 \\ U_s^- & U_1 < U_2 \end{cases}$
$U_{out} = U_{in}\left(1+\frac{R_2}{R_1}\right)$
$U_{out} = -U_{in}\frac{R_2}{R_1}$
Voltage Follower
Inverting Summing Amplifier
Differential Amplifier
$U_{in} = U_{out}$
R_f
U_n
R_n
R_3
R_4
$U_{out} = -R_f\left(\frac{U_1}{R_1} \cdots \frac{U_n}{R_n}\right)$
$U_{out} = \left(1+\frac{R_2}{R_1}\right)\left(\frac{R_3}{R_3+R_4}\right)U_2 - \left(\frac{R_2}{R_1}\right)$
If $R_1 = R_3$ and $R_2 = R_4$ then:
$U_{out} = \frac{R_2}{R_2}(U_2 - U_1)$
Differentiator Amplifier
Integrator Amplifier
R
C
$U_{out} = -RC\frac{dU_{in}}{dt}$
$U_{out} = -\int \frac{1}{RC}U_{in}dt$

Figure 16.130: Basic operational amplifier configurations

The figure below illustrates how a simple game (Pong on ATARI) was made purely electronically even before C++ existed:

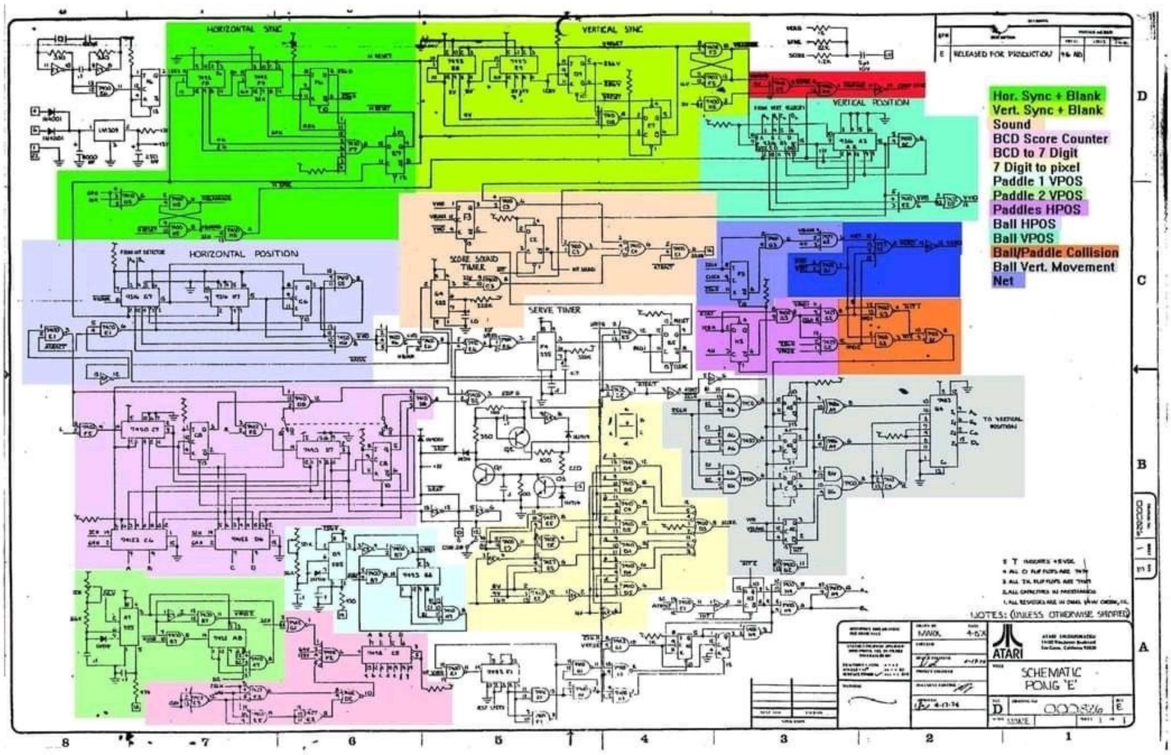

HORIZONTAL SYNC
VERTICAL SYNC
VERTICAL POSITION
HORIZONTAL POSITION
SCORE SOUND TIMER
SERVE TIMER
RELEASED FOR PRODUCTION
Hor. Sync + Blank
Vert. Sync + Blank
Sound
BCD Score Counter
BCD to 7 Digit
7 Digit to pixel
Paddle 1 VPOS
Paddle 2 VPOS
Paddles HPOS
Ball HPOS
Ball VPOS
Ball/Paddle Collision
Ball Vert. Movement
Net
TO VERTICAL POSITION
NOTES: (UNLESS OTHERWISE SPECIFIED)
ATARI
SCHEMATIC PONG 'E'
D
E
8
7
6
5
4
3
2
1
D
C
B
A

Figure 16.131: Pong on ATARI circuit logic

~20 %

★★★☆☆

23 votes, 64.35%

16.4. Civil Engineering

CIVIL ENGINEERING represents all the techniques for civil constructions and tools associated with it. Civil engineers are involved in the design, implementation, operation and rehabilitation of construction works and urban infrastructure that they manage to meet the needs of society, while ensuring public safety and environmental protection at least in theory. Very interesting and varied, their achievements are distributed mainly in five major areas: structures, geotechnics, hydraulics, transport and environment. As usual in this textbook, we will focus only on the mathematical formalization of very common cases that have a practical application in everyday life.

Civil Engineering is traditionally broken into several sub-disciplines including architectural engineering, environmental engineering, geotechnical engineering, control engineering, structural engineering, earthquake engineering, transportation engineering, forensic engineering, municipal or urban engineering, water resources engineering, materials engineering, wastewater engineering, offshore engineering, facade engineering, quantity surveying, coastal engineering, construction surveying, and construction engineering. Civil engineering takes place in the public sector from municipal through to national governments, and in the private sector from individual homeowners through to international companies.

Let us notice that in civil engineering it is sometimes made use of the calculation of minimal surfaces. This has been already covered by a concrete examples in the section of Analytical Mechanics (see page 3182). Regarding the efforts of the heat on beams, this is also already been discussed in the section of Thermodynamics (see page 3615). Regarding the problem of resonance this has also already been discussed in the section of Acoustics (see page 8227).

We will limit ourselves in this section to provable accurate calculations and not on experimental formulas and this in order to have a general introduction to civil engineering techniques and become familiar with the language and some methods of calculation used by engineers in this field. In practice, civil engineers make use of a standard packages with formulas including coefficients or based on tables or maps. For example, in Switzerland, the civil engineers refer to SIA standards (Swiss Society of Engineers and Architects), which contain among others a lot of empirical formulas that engineer use without knowing where they come from. The theoretical approach is obviously insufficient and any training in this field should be obviously completed by practical laboratory work.

Remark

It would be pretentious to claim to do with this section as well and also as complete as the *Statique* free French PDF of Nicolet Gaston Raymond that is a priori unrivaled in content and quality to this date (even compared to non-free books on the subject!). It is therefore strongly recommended to refer to it if the reader wants to drive full information about civil engineering (see the download section of the companion website or directly his own website `http://www.raymondnicolet.ch`). Same recommendation applies to the reader about the master pieces that are the English books *Mechanics of Materials* ([246]) and *Structural and Stress Analysis* ([384])!

Notice as for physics and mechanics that have plenty of ISO standards that increase the quality of the work, the efficiency of the collaboration and communication, civil engineering also has plenty of ISO standards that increase the quality and security of the final result but however highly increase the costs and the time of constructions. You can see a small sample of these standards below (for example seismic/earthquakes standards and technical documentation standards are not represented):

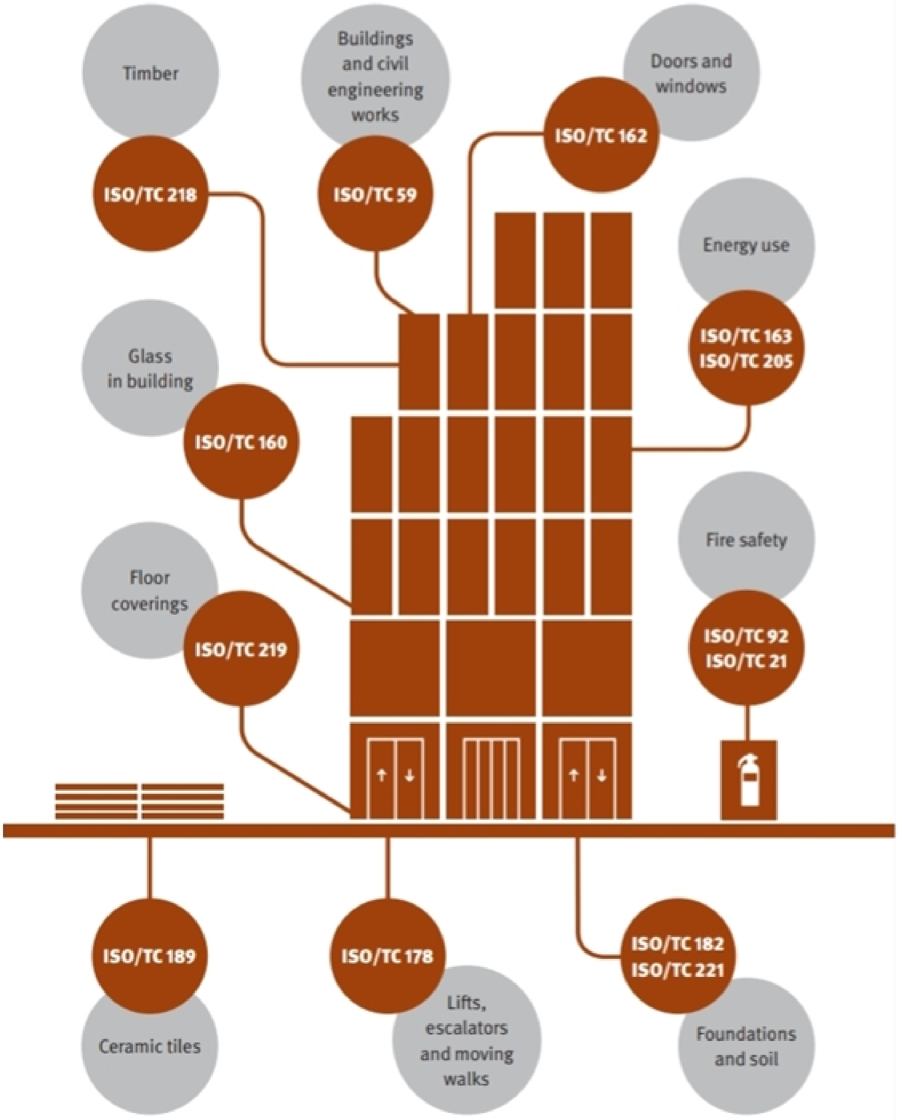

Timber
ISO/TC 218
Buildings and civil engineering works
ISO/TC 59
Doors and windows
ISO/TC 162
Energy use
ISO/TC 163
ISO/TC 205
Glass in building
ISO/TC 160
Fire safety
ISO/TC 92
ISO/TC 21
Floor coverings
ISO/TC 219
ISO/TC 189
Ceramic tiles
ISO/TC 178
Lifts, escalators and moving walks
ISO/TC 182
ISO/TC 221
Foundations and soil

Figure 16.132: Some important ISO standard in Civil Engineering for building construction

16.4.1 Static

Civil engineering uses a lot the tools of the Static for constructions. We're not going here obviously to rewrite the whole section here of Classical Mechanics with the fundamental principle of statistics and everything that goes with it (see page 3194), nor static analysis of beams that has already be done in the section of Mechanical Engineering (see page 8388), but only introduce some applicative aspects of the principle.

To begin this section, let's look at least the smaller non-trivial cases encountered in practice. We would like to put against a wall a solid object and we would like to know the strength that will endure this wall. We can represent this basically by the following scheme:

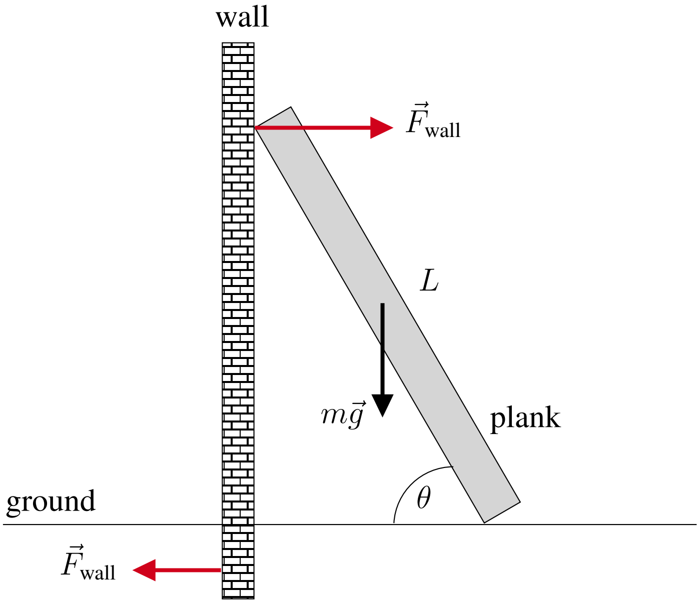


Figure 16.133: Massive object (high density plank) against a wall

For the wall holds it is necessary (but not a sufficient condition, it is just necessary!!!) that the force momentum of gravity equalizes the moment of force of the wall. Then we have:

$$mg\frac{L}{2}\cos(\theta)=F_{\text{wall}}L\sin(\theta) \tag{16.4.1}$$

Therefore:

$$F_{\text{wall}}=\frac{1}{2}\frac{mg}{\tan(\theta)} \tag{16.4.2}$$

Verbatim, the foundations of the wall must be able to opposite to the force momentum.

16.4.2 Pulleys

A "pulley" is a wheel with a grooved rim around which a cord passes, which acts to change the direction of a force applied to the cord and is used to raise heavy weights.

In the context of the study of static forces, there is an industrial example that is famous and that we meet almost every week by walking or driving in front of working sites (construction cranes), train stations (stretchers), ports (ships) or by going to fitness hall or garages: the pulley! Its origin is once again an Archimedes idea (it seems...) who used it for the movement of large masses needed in various projects of his time. The relation with the civil engineering is then completely justified! Let us see this more closely (exceptionally there are very few equations).

Let us consider the following situation named "simple fixed pulley" with a mass of 10 [kg] (i.e. a force of 100 [N] with Earth gravity rounded gravity to the nearest ten) hung to rope slipped into the gutter of a pulley:

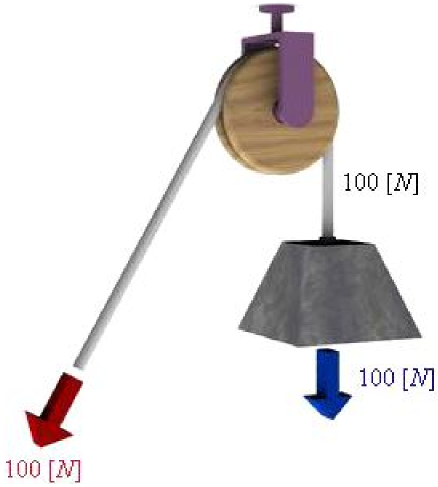


Figure 16.134: **Simple fixed pulley (source: Wikipedia)**

A simple fixed pulley has the only mechanical advantage to be able to exert the force in a different direction to that of the movement, the force that has to be applied is the same as that required to move the object without the pulley!

The anchor point of the pulley must support the force required to move the object plus the traction force, thus twice this force in the worst case. Otherwise the total load that must support the anchor point is a function of the "pull angle" of the rope (between 90° and 180°) of course:

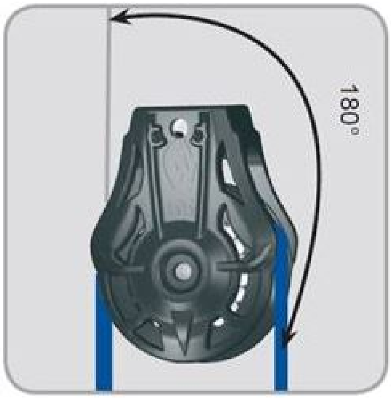


Figure 16.135: **Pull angle at** 180°

For an angle of 180° the load factor is 200% (don't forget it's the load factor on the center of the pulley!!!). A load of 10 [kg] on the rope will represent a load of 20 [kg] on the pulley.

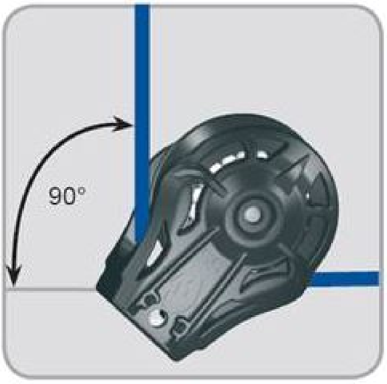


Figure 16.136: **Pull angle at** 90°

For an angle of 90°, the load factor is 140% (don't forget it's the load factor on the center of the pulley!!!). A load of 10 [kg] on the rope is a load of 14 [kg] on the pulley.

A simple device that runs over pulleys used to study the laws of motion and forces is known as "Atwood's machine". This machine is applicable in various scenarios in the study of mechanics. Basically, this machine is a perfect illustration of tension force, gravitational force and the downward force acting on a particular system of masses that are hung using a string and pulley arrangement.

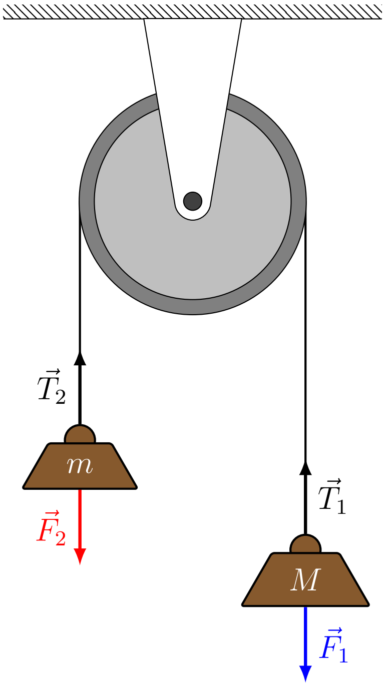


Figure 16.137: Atwood's machine

Using the standard labelling, let the smaller mass be m, and the larger mass is M, and since the forces are only acting in the vertical direction, we will use the summation equation for the forces along the y-axis only.

It can be clearly seen that only two forces are acting on each of the masses, that is, their own weight acting downwards and the tension in the string acting upwards.

Hence, the summation equation for forces on each mass can be written using the standard sign convention, the upward force is taken as positive, and the downward force is taken as negative.

We have obviously:

$$T-mg=ma \qquad \text{and} \qquad T-Mg=-Ma \tag{16.4.3}$$

Where T is the tension in the string and a is the magnitude of the acceleration of two masses. On solving the both we get

$$a=g\frac{M-m}{M+m} \tag{16.4.4}$$

Now let us consider a situation where we set one end of the rope to the support and we draw with the other end to move both the pulley AND the load of 10 [kg]. This configuration is named "simple moving pulley" or "reverse pulley" (the legend says that it is this system that Archimedes used it to pull an entire boat):

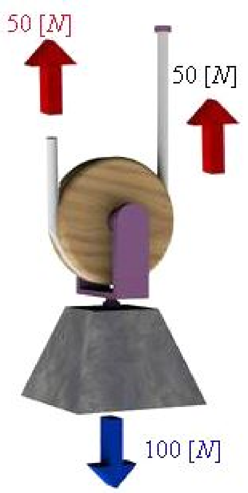


Figure 16.138: Single mobile pulley (source: Wikipedia)

In fact in this system (set up vertically or horizontally whatever!) it is as if there were two people who shared the effort of displacement: the wall and the free end of the rope (which is pulled).

The single mobile pulley therefore reduces the force required to move the load to half (the anchor supporting the other half) and thus adding other mobile pulleys, we continue to divide the action to apply! It was a simple enough idea, somebody just had to think of it!

But this system requires a pull movement of the rope end equal to twice the distance of movement of the load regardless of the radius of the pulley.

Let us also indicate that more the pulley has a large radius, more the force momentum will be big too! So in the case of very heavy loads we will favour large radii for the pulleys if the system that pulls can provide only weak force.

A more realistic configuration (as we will rarely put us above the anchor point to pull the rope and furthermore the previous system is not very stable mechanically speaking ...) of the single mobile pulley presented above is as follows:

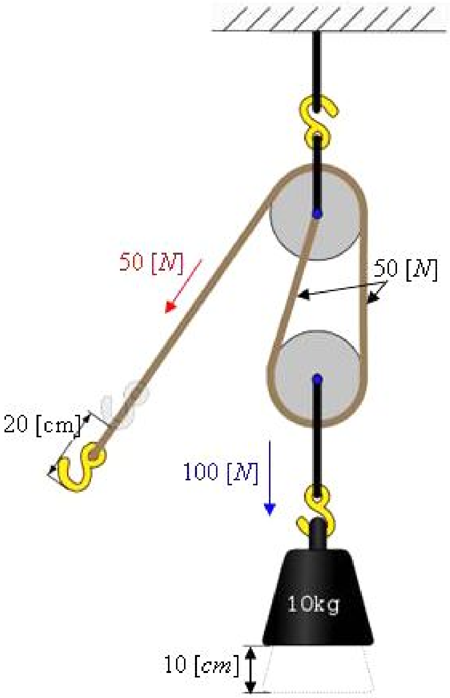


Figure 16.139: **Single fixed and mobile mixed pulleys (source: Wikipedia)**

Obviously, when we represent systems like these in schools, we neglect for simplification purposes the mass of the pulleys themselves that we should strictly speaking take into consideration!

When we use several pulleys systems working together, we say that we have a configuration of "composed pulleys". The most common type of such a configuration is named a "hoist": the pulleys are distributed in two groups (or block), one fixed and one mobile:

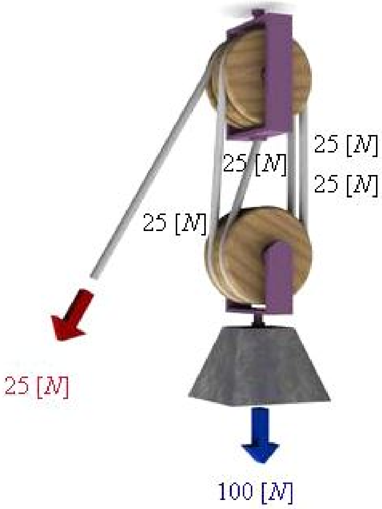


Figure 16.140: **System of combined fixed and moving pulleys (source: Wikipedia)**

In each group we set an arbitrary number of pulleys that demultiply by the same factor the initial load. The load is obviously united with the mobile group.

So we have 25 [N] at the end of the rope. The reader can try to have fun by finding the 4 anchor points in the previous illustration and the two pulleys which each divide by 2 the necessary pull force ... If necessary here is the same configuration but represented in and "unfolded" way:

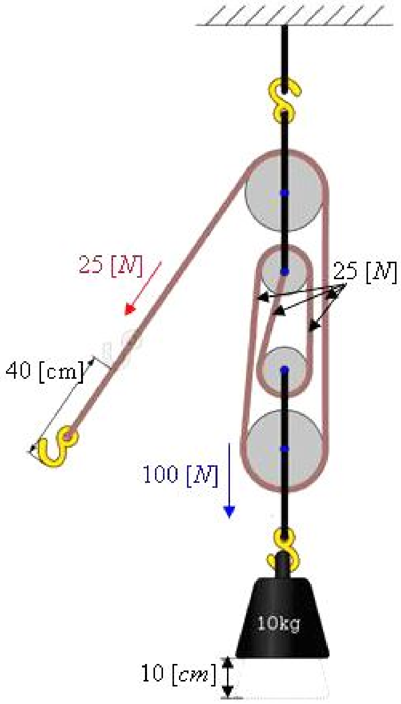


Figure 16.141: Unfolded system of combined fixed and moving pulleys (source: Wikipedia)

We already see that the big top pulley is useless except to change the direction of the pulling force (however if we were not in the assumption of frictionless pulleys it would help to increase the torque). In fact, the two pulleys that are used to divide the force by 2 are the two lower one, the remainder being provided only for convenience for the movement of the rope.

Let us recall to the reader that in reality we would have to take into account the friction of the rope on the pulley (science of tribology) and we have proven in the section of Classical Mechanics (see page 3464) that the real force (stress) of a rope end relatively to the other (end) was given by:

$$T_2 = T_1 e^{\mu_s \beta} \tag{16.4.5}$$

and therefore that the real useful force (stress) to lift a load was given by:

$$\boxed{T_1 = \frac{T_2}{e^{\mu_s \beta}} < T_2} \tag{16.4.6}$$

So the force (stress) supplied to lift the object being equal to T_2, the force (real stress) lifting the weight being T_1, the difference gives the force (stress) making the pulley turn by the intermediate of the friction. Then we have:

$$T_P = T_2 - T_1 = T_1 e^{\mu_s \beta} - T_1 = T_1 \left(e^{\mu_s \beta} - 1 \right) \tag{16.4.7}$$

The moment of force of the pulley of radius R is:

$$M = RT_P = RT_1 \left(e^{\mu_s \beta} - 1 \right) = \frac{v}{\omega} T_1 \left(e^{\mu_s \beta} - 1 \right) \tag{16.4.8}$$

Let us see an another well known application of a system of fixed and mobile pulleys in train stations (and not only!!!):

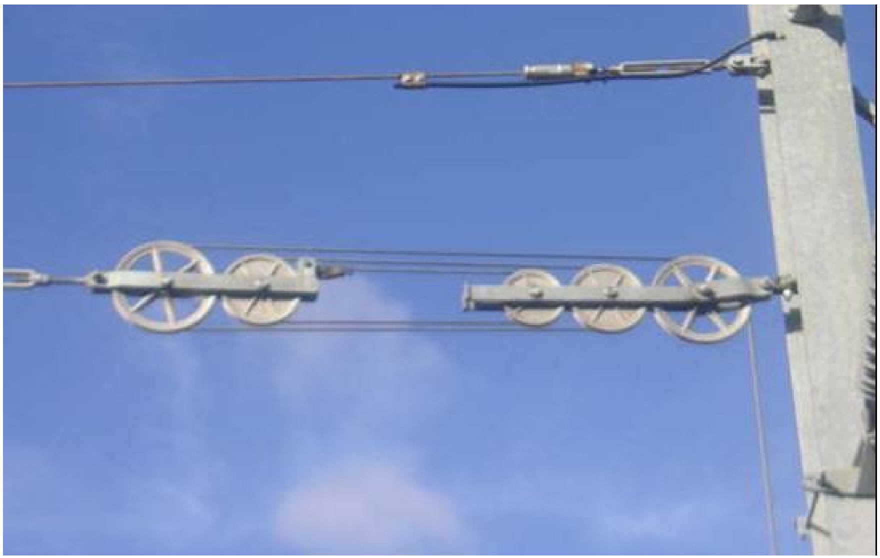


Figure 16.142: **System of mixed fixed and mobile pulley**

This is a hoist to tension electrical cables with a non-visible counterweight on the photo (bottom right) that ensures a certain strength so some tension. The advantage of this system is that it allows you to add loads gradually as the cable relaxes and these loads are 4 times higher at the electrical cable level thanks to the two movable pulleys (on the left). The pulleys to the right are only here for convenience for the movement of the rope and... the pulling direction for the pulley at the right end.

In the case of a horizontal or vertical load lift, it is easy to determine the gear ratio D. Indeed, if we consider F as the force required to lift vertically the object from a height h by pulling the rope over a length d and F_g the gravitational force on the mass drawn, we then have neglecting the friction and weight of all mobile pulleys:

$$D=\frac{F}{F_g}=\frac{h}{d} \tag{16.4.9}$$

Finally, let us notice that it is possible to play with the radius of the deflection (deviation) pulley to reduce the force to provide while keeping constant the moment of force (we then speak of "differential hoist") but in the end the energy used remain always the same to lift an object at the same height (and we will to pull the rope even more to lift the load at the same height).

To conclude about pulleys the illustration below gives an excellent summary of what we saw until now:

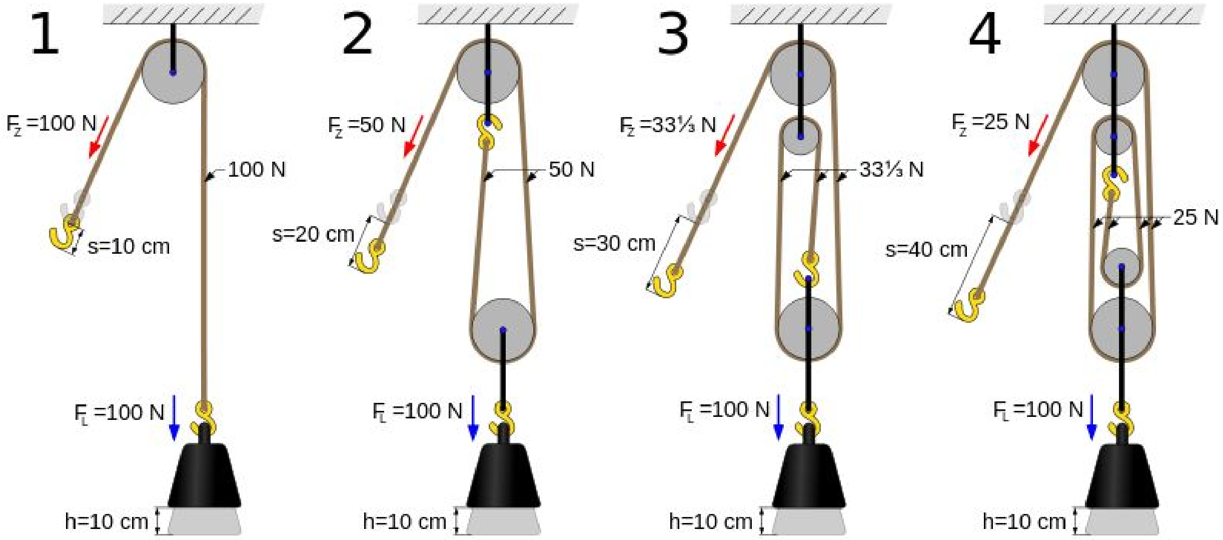

1
F_Z=100 N
100 N
s=10 cm
F_L=100 N
h=10 cm
2
F_Z=50 N
50 N
s=20 cm
F_L=100 N
h=10 cm
3
F_Z=33⅓ N
33⅓ N
s=30 cm
F_L=100 N
h=10 cm
4
F_Z=25 N
25 N
s=40 cm
F_L=100 N
h=10 cm

Figure 16.143: **Summary of typical pulleys (source: Wikipedia)**

(1) Fixed Pulley:

$$F_Z = F_L = mg \qquad s = h \tag{16.4.10}$$

(2) Mobile Pulley:

$$F_Z = \frac{F_L}{2} = \frac{m}{2}g \qquad s = 2h \tag{16.4.11}$$

(3) Simple hoist:

$$F_Z = \frac{F_L}{3} = \frac{m}{3}g \qquad s = 3h \tag{16.4.12}$$

(4) Double hoist:

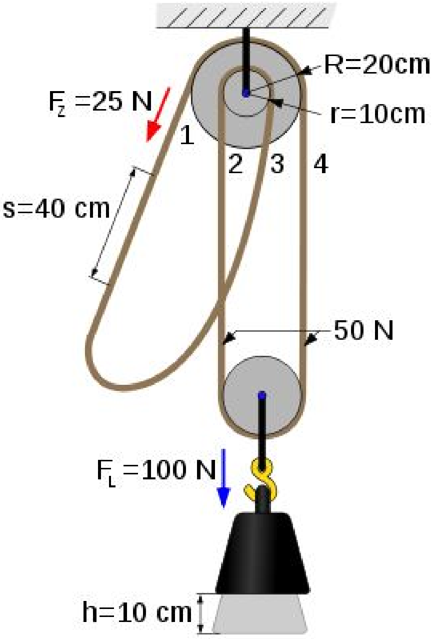
$$F_Z = \frac{F_L}{4} = \frac{m}{4}g \qquad s = 4h \tag{16.4.13}$$

That can be generalized to the n case!

Hence the "ideal mechanical advantage of pulleys" (see section Mechanical Engineering page 8360) given in such situations by:

$$\text{IMA} = \frac{F_{\text{out}}}{F_{\text{in}}} = N \tag{16.4.14}$$

We can also consider the following differential hoist:



R=20cm
r=10cm
F_Z =25 N
1
2
3
4
s=40 cm
50 N
F_L =100 N
h=10 cm

Figure 16.144: **Typical differential pulleys (source: Wikipedia)**

with:

$$F_Z = F_L \frac{R-r}{2R} \qquad s = \frac{2R}{R-r}h \tag{16.4.15}$$

In the below photo the user can see a crane using a system of pulleys to leverage $10,000$ metric tons (i.e. $10,000,000$ kilos...):

SWL 10000mt
SLEIPNIR
HEEREMA

Figure 16.145: **HMC Semi-Submersible** $10,000$ **metric tons Crane vessel**

16.4.2.1 Windlass

The "windlass" is an apparatus for moving heavy weights. Typically, a windlass consists of a horizontal cylinder (barrel), which is rotated by the turn of a crank or belt. A winch is affixed to one or both ends, and a cable or rope is wound around the winch, pulling a weight attached to the opposite end.

Mathematically all equations about windlass can be derived by analogy from pulleys.

Consider for example the simple following windlass:

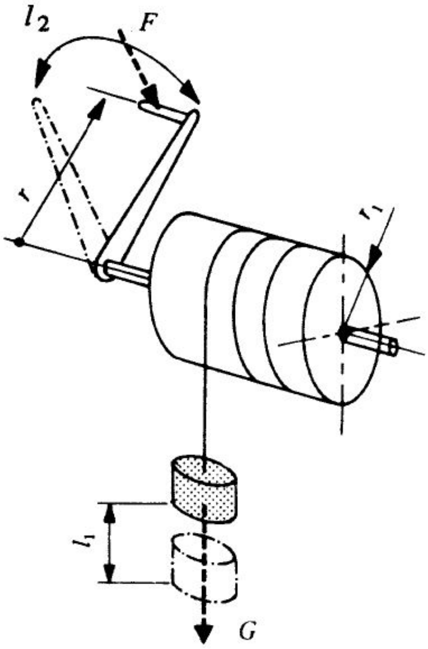


Figure 16.146: **Simple windlass (source: Fortec, Charles Pache)**

For this case we have obviously (gears rules applies as seen in the section of Mechanical Engineering at page 8368):

$$F=G\frac{r_1}{r}=mg\frac{r_1}{r} \qquad l_2=l_1\frac{r}{r_1} \tag{16.4.16}$$

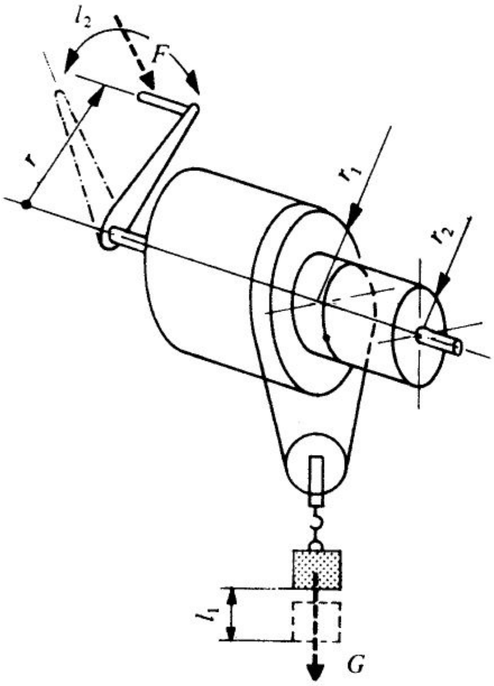


Figure 16.147: **Differential windlass (source: Fortec, Charles Pache)**

For this case we have obviously the same as the differential hoist:

$$F=G\frac{r_1-r_2}{2r}=mg\frac{r_1-r_2}{2r} \qquad l_2=\frac{2r}{r_1-r_2} \tag{16.4.17}$$

16.4.3 Cornu spiral

The "clothoid" is a plane transcendental curve[15] which the curvature is proportional to the curvilinear abscissa. It is also named "Cornu spiral", referring to Alfred Cornu (11841-11902 according to holocene calendar), the French physicist who invented it. More rarely, it may appear under the name of "radioïde arc" or "Euler spiral" or "Fresnel spiral".

This geometry is also suitable for curves for railways because a train following the path of this geometry at a constant speed will be under constant angular acceleration whatever the point of the curve, which reduces both the stresses on the rail or on the wheels and the discomfort of passengers of the train:

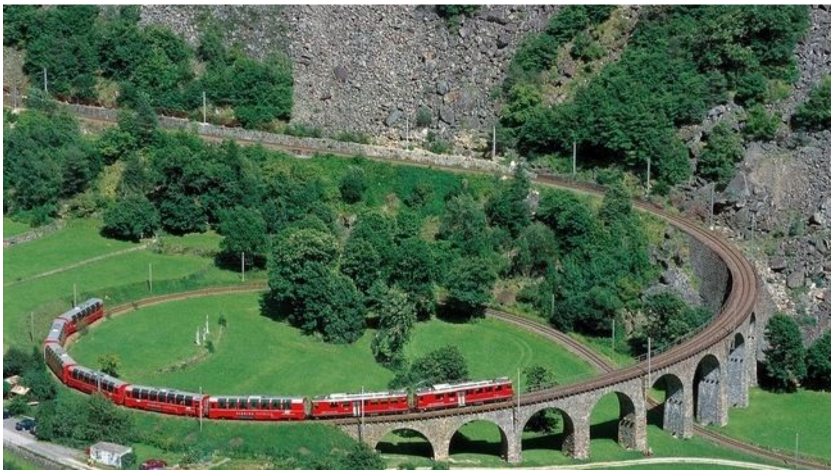


Figure 16.148: **Train loop with a clothoid shape**

Same for cars:

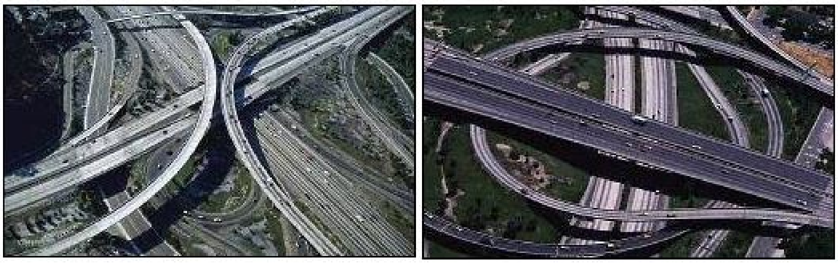


Figure 16.149: **Highway bifurcation with a clothoid shape**

Finally, the railway cable of pylons of cable car and supporting the suspension cable, take this form. Like this, it is possible to circulate the cab at maximum speed on the pylon, without inconveniencing the passengers.

---

[15] A "transcendental curve" is a curve which is not defined by an algebraic equation (typically polynomial or trigonometric), but by a transcendental equation, that is to say, the unknowns are not finite quantities, but differential ($\mathrm{d}x$, $\mathrm{d}y$, representing an infinitesimal variation of the variables x and y). Such a curve can not be constructed geometrically precisely because its equation can not be reduced in a simple form $y = f(x)$.

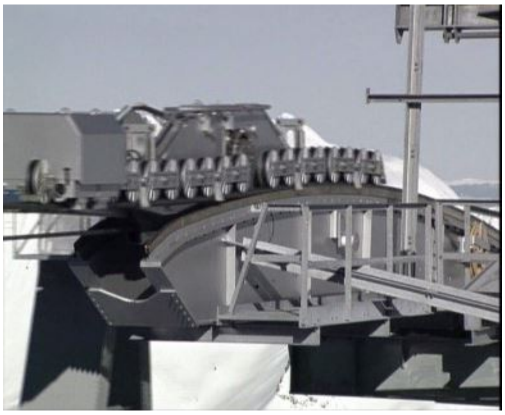


Figure 16.150: Cable car pylon with a clothoid shape

Also this curve is used for vertical loops or loops in the roller coaster for passenger comfort, so that the vertical acceleration is continuous.

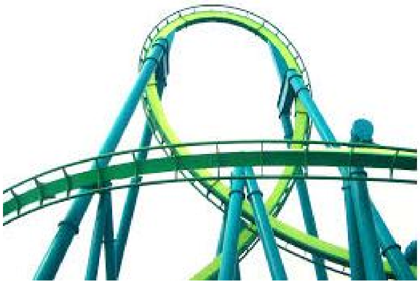


Figure 16.151: Roaller coaster vertical loop with a clothoid shape

When a vehicle moves in a circular motion, it will undergo a force $\vec{F}=m\vec{a}$ perpendicular to its direction (centrifugal force) having for norm (see section Classical Mechanics page 3388):

$$F=m\frac{v^2}{R} \tag{16.4.18}$$

from the start of its entry into the circular curve. This effect is problematic because for some small cars on a highway, the centrifugal force can equal the force of gravity of the car itself (when speed is within the legal values!).

Thus, the acceleration rise brutally from 0 to v^2/R, then the engineers build sometimes the curves with a declination to improve the adhesion, but it is also possible to try to find curves for which the acceleration is more gradual. For example if the curvature C given by (see section Differential Geometry page 2865):

$$C=\frac{1}{R} \tag{16.4.19}$$

is proportional to the path s (curvilinear abscissa) travelled in the curve, we will have at the beginning of the curve $C=0$ so the acceleration will be zero.

So what we seek are then curves such that:

$$C=k\cdot s \tag{16.4.20}$$

where k is a given constant.

For this, let us recall that we can also write naturally for a circle, the curvature into the following form :

$$C=\frac{1}{R}=\frac{\mathrm{d}\theta}{\mathrm{d}s} \tag{16.4.21}$$

Indeed, if we turn of an angle $\mathrm{d}\theta$ then we move of a length $\mathrm{d}s=R\mathrm{d}\theta$ (see section Trigonometry page 2542).

Then we have the relation:

$$k\cdot s=\frac{\mathrm{d}\theta}{\mathrm{d}s} \tag{16.4.22}$$

Thus:

$$\mathrm{d}\theta=ks\mathrm{d}s \tag{16.4.23}$$

hence:

$$\int\limits_0^\theta \mathrm{d}\theta=\int\limits_0^s ks\mathrm{d}s\Rightarrow\theta=\frac{1}{2}ks^2=Ks^2 \tag{16.4.24}$$

But let us recall that the parametric equation of the circle is (see section Analytical Geometry page 2782):

$$\begin{cases}x=R\cos(\theta)\\ y=R\sin(\theta)\end{cases} \tag{16.4.25}$$

We then have:

$$\begin{cases}\mathrm{d}x=\mathrm{d}R\cos(\theta)+R\mathrm{d}(\cos(\theta))=R\cos(\theta)\mathrm{d}\theta=R\cos(\theta)\dfrac{\mathrm{d}s}{R}\\ \mathrm{d}y=\mathrm{d}R\sin(\theta)+R\mathrm{d}(\sin(\theta))=R\sin(\theta)\mathrm{d}\theta=R\sin(\theta)\dfrac{\mathrm{d}s}{R}\end{cases} \tag{16.4.26}$$

Therefore:

$$\begin{cases}\dfrac{\mathrm{d}x}{\mathrm{d}s}=\cos(\theta)\\ \dfrac{\mathrm{d}y}{\mathrm{d}s}=\sin(\theta)\end{cases} \tag{16.4.27}$$

We can now write:

$$\begin{cases}\mathrm{d}x=\cos(\theta)\mathrm{d}s=\cos(Ks^2)\mathrm{d}s\\ \mathrm{d}y=\sin(\theta)\mathrm{d}s=\sin(Ks^2)\mathrm{d}s\end{cases} \tag{16.4.28}$$

Therefore:

$$\begin{cases}x=\int\limits_0^s\cos(Kt^2)\mathrm{d}t\\ y=\int\limits_0^s\sin(Kt^2)\mathrm{d}t\end{cases} \tag{16.4.29}$$

with a small change of variables:

$$u^2 = Kt^2 \Rightarrow u = \sqrt{K}t \Rightarrow \frac{\mathrm{d}u}{\mathrm{d}t} = \sqrt{K} \Rightarrow \mathrm{d}t = \frac{\mathrm{d}u}{\sqrt{K}} \tag{16.4.30}$$

it comes:

$$\begin{cases} x = \dfrac{1}{\sqrt{K}} \int\limits_0^s \cos(u^2)\mathrm{d}u \\ y = \dfrac{1}{\sqrt{K}} \int\limits_0^s \sin(u^2)\mathrm{d}u \end{cases} \tag{16.4.31}$$

by taking $x_0 = y_0 = 0$ (we can always do a translation later).

The two integrals are named "Fresnel integrals" and cannot calculated in a closed form as far as we know. However, we can express them in a Taylor series expansion form (see section Sequences and Series page 1696) as:

$$\begin{aligned} \sqrt{K}x &= s - \frac{s^5}{5 \cdot 2!} + \frac{s^9}{9 \cdot 4!} - \frac{s^{13}}{13 \cdot 6!} + \ldots \\ \sqrt{K}y &= \frac{s^3}{3} - \frac{s^7}{7 \cdot 3!} + \frac{s^{11}}{11 \cdot 5!} - \ldots \end{aligned} \tag{16.4.32}$$

The plot of the Fresnel integral (Euler spiral/Cornu spiral) in Maple™ 4.00b gives:

```
>plot([FresnelC(t),FresnelS(t),t=-5..5]);
```

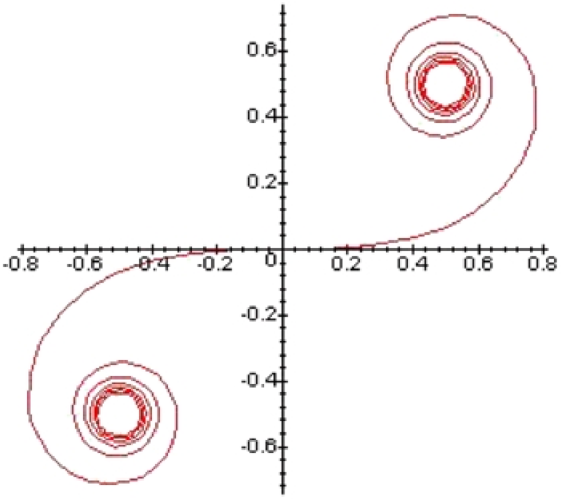


Figure 16.152: **Plot of the Fresnel (Euler spiral/Cornu spiral) integral in Maple™ 4.00b**

By zooming on the part relevant for us:

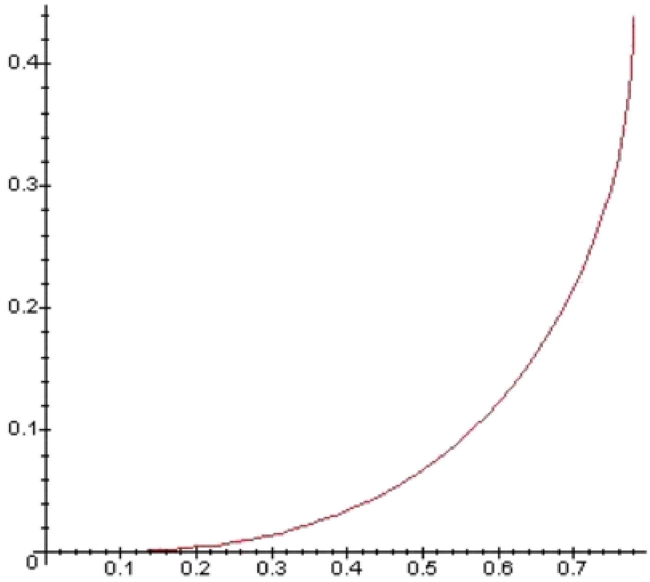


Figure 16.153: **Focus on the origin of the Fresnel integral**

The same thing to a given constant using the Taylor series previously presented:

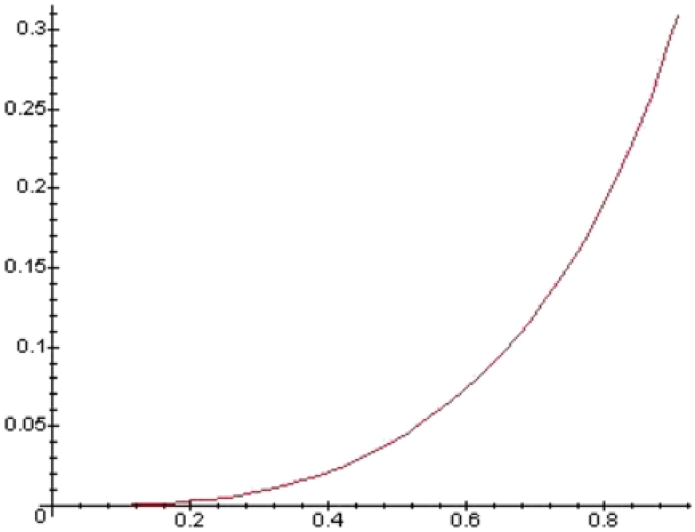


Figure 16.154: **Equivalent Taylor series expansion**

The engineering office use special softwares incorporating clothoids spirals in 2D or 3D environments based on topographic surveys done by GIS specialists (for example AutoCAD has a special dedicated to clothoids design).

16.4.4 Overhead cable

An overhead cable is a cable for the transmission of information, laid on utility poles. Overhead telephone and cable TV lines are common in North America. Elsewhere, overhead cables are laid mainly for telephone connections of remote buildings and temporary mechanisms, as for example building sites. The same poles sometimes carry overhead power lines for the supply of electric power. Power supply companies may also use them for an in-house communication network. Sometimes these cables are integrated in the ground or power conductor. Otherwise an additional line is strung on the masts.

Galilee was probably the first to be interested in the chain overhead cable shape that he interpreted for a parabolic arc. Jean Bernoulli, Huygens and Leibniz found (independently) in response to the challenge of Jakob Bernoulli, his true nature in 11691 (holocene calendar): it is generated by a hyperbolic cosine!

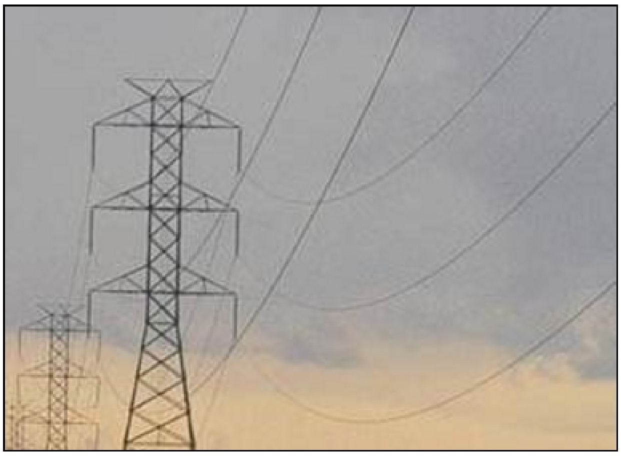


Figure 16.155: **High-voltage overhead cable (source: chronomaths)**

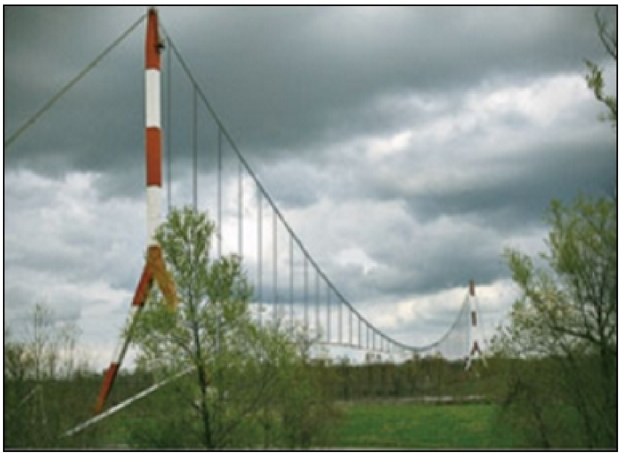


Figure 16.156: **Suspended Pipeline over a river**

16.4.4.1 Free overhead cable (catenary)

Let us consider (source: ChronoMath) for the study a homogeneous free, flexible cable attached at two points A and B. In its equilibrium position, the cable hangs in a vertical plane and seems to take a parabolic shape. In fact, not really ...

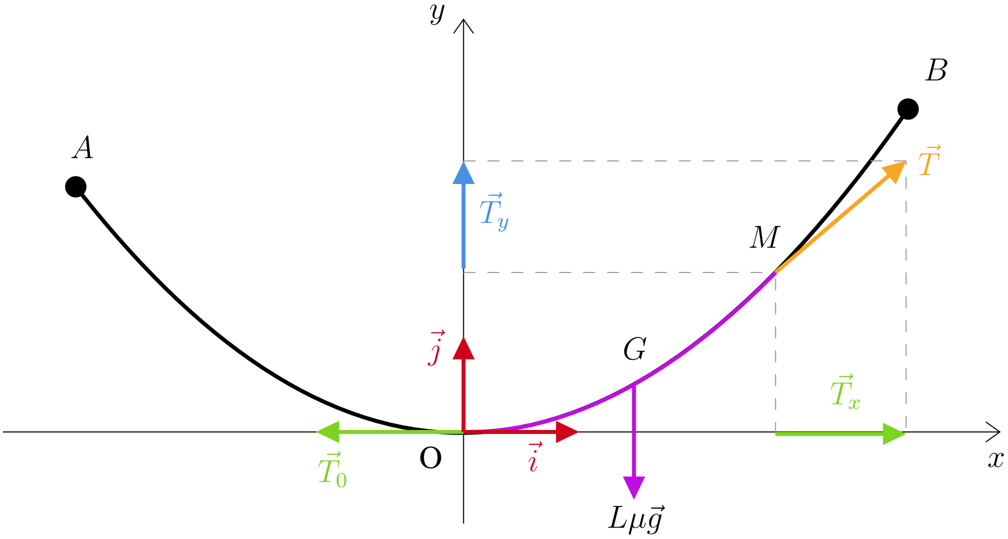


Figure 16.157: **Configuration for studying suspended cables**

Let us create in this plane an orthonormal reference frame $(\mathrm{O},\vec{i},\vec{j})$, where O denotes the lowest point of the cable and let us denote as always the field of gravity $\vec{g}$ that we consider as uniform.

Let us denote by $\vec{T}_0$ the tension (force) at the point O which defeats the tension at the point M so that the cable portion $\overline{\mathrm{O}M}$ of length L, subjected to its linear weight $\mu\vec{g}$ at the point G, is in equilibrium in the static sense:

$$\vec{T}+\vec{T}_0+\mu L\vec{g}=\vec{0} \tag{16.4.33}$$

Remark

The product $\mu\vec{g}$ of the linear mass μ with the gravitational acceleration $\vec{g}$ is also very often denoted $\vec{w}$ as we have already see in the section of Mechanical Engineering to represent the linear load.

Let us project on the coordinate axes by denoting by α the angle $\widehat{\vec{i}\vec{T}}$. We then have the following decomposition:

$$\begin{aligned}\vec{T}&=T_x\vec{i}+T_y\vec{j}=\left(\|\vec{T}\|\cos(\alpha)\right)\vec{i}+\left(\|\vec{T}\|\sin(\alpha)\right)\vec{j}\\ \vec{T}_0&=-\|\vec{T}_0\|\vec{i}\\ \mu L\vec{g}&=-\mu L\|\vec{g}\|\vec{j}\end{aligned} \tag{16.4.34}$$

We can then write the following system:

$$\begin{aligned}&\left(\|\vec{T}\|\cos(\alpha)\right)\vec{i}-\|\vec{T}_0\|\vec{i}=\vec{0}\\&\left(\|\vec{T}\|\sin(\alpha)\right)\vec{j}-\mu L\|\vec{g}\|\vec{i}=\vec{0}\end{aligned}\tag{16.4.35}$$

Either after simplification:

$$\begin{aligned}&\|\vec{T}\|\cos(\alpha)-\|\vec{T}_0\|=0\\&\|\vec{T}\|\sin(\alpha)-\mu L\|\vec{g}\|=0\end{aligned}\tag{16.4.36}$$

Therefore:

$$\begin{aligned}&\|\vec{T}\|\cos(\alpha)=\|\vec{T}_0\|\\&\|\vec{T}\|\sin(\alpha)=\mu L\|\vec{g}\|\end{aligned}\tag{16.4.37}$$

By calculating the ratio:

$$\tan(\alpha)=\frac{\mu L\|\vec{g}\|}{\|\vec{T}_0\|}=L\frac{\mu\|\vec{g}\|}{\|\vec{T}_0\|}=\frac{L}{k}\tag{16.4.38}$$

To get the differential equation... (here this is subtle...):

$$\mathrm{d}(\tan(\alpha))=\frac{\mathrm{d}L}{k}\tag{16.4.39}$$

where $\mathrm{d}L$ is the curvilinear abscissa of the cable (often denoted $\mathrm{d}s$ in the literature according to what is done in Differential Geometry).

Then:

$$\mathrm{d}(\tan(\alpha))=\frac{\sqrt{\mathrm{d}x^2+\mathrm{d}y^2}}{k}=\mathrm{d}x\frac{\sqrt{1+y'^2}}{k}\tag{16.4.40}$$

But the tangent is also the derivative of the function describing the chain. So:

$$\mathrm{d}(\tan(\alpha))=\mathrm{d}(y')=\mathrm{d}x\frac{\sqrt{1+y'^2}}{k}\tag{16.4.41}$$

Therefore it comes:

$$\frac{y''}{\sqrt{1+y'^2}}=\frac{1}{k}\tag{16.4.42}$$

Following the intervention of one of our reader, we will propose two ways to solve this differential equation. The first is the original one and it is a bit complicated and the second one (available much further below) is that proposed by a reader and that is perhaps more elegant and simple.

- First approach:

  Let us put $u=y'$ and seek the primitive of the left member at first (that of the right member being obvious). The calculations made in the section of Differential and Integral Calculus in the determination of the usual primitives (see page 1534) give us:

$$\int \frac{u'}{\sqrt{1+u^2}}\mathrm{d}u = u'\int \frac{1}{\sqrt{1+u^2}}\mathrm{d}u = \ln\left(u+\sqrt{1+u^2}\right)+c^{te} \tag{16.4.43}$$

  So we have:

$$\ln\left(y'+\sqrt{1+y'^2}\right)=\frac{x}{k}+c^{te} \tag{16.4.44}$$

  By passing to the exponential:

$$y'+\sqrt{1+y'^2}=e^{\frac{x}{k}+c^{te}} \tag{16.4.45}$$

  observing that in our problem at $x=0$ we have indeed $y'=0$.

  To find y' we use a trick: We know that the function is symmetric. So if we replace x by $-x$ the tangent also changes sign and passes from y' to $-y'$:

$$-y'+\sqrt{1+(-y')^2}=e^{\frac{-x}{k}+c^{te}} \tag{16.4.46}$$

  By subtracting:

$$y'+\sqrt{1+y'^2}=e^{\frac{x}{k}+c^{te}} \qquad \text{and} \qquad -y'+\sqrt{1+y'^2}=e^{\frac{-x}{k}+c^{te}} \tag{16.4.47}$$

  it comes:

$$2y'=e^{\frac{x}{k}}-e^{-\frac{x}{k}} \Rightarrow y'=\frac{e^{\frac{x}{k}+c^{te}}-e^{-\frac{x}{k}+c^{te}}}{2} \tag{16.4.48}$$

  So after integration:

$$y=k\frac{e^{\frac{x}{k}+c^{te}}-e^{-\frac{x}{k}+c^{te}}}{2}=k\cosh\left(\frac{x}{k}+c_1^{te}\right)+c_2^{te} \tag{16.4.49}$$

  where the expression after the first equality is named the "catenary curve". It comes to the end:

$$\boxed{y=\frac{T_0}{\mu g}\cosh\left(x\frac{\mu g}{T_0}+c_1^{te}\right)+c_2^{te}} \tag{16.4.50}$$

  Where the constant will be determined by the initial conditions.

  We see with Maple™ 4.00b the difference between a parabola and the chain easily:

```
>plot([x^2,cosh(x)],x=-4...4);
```

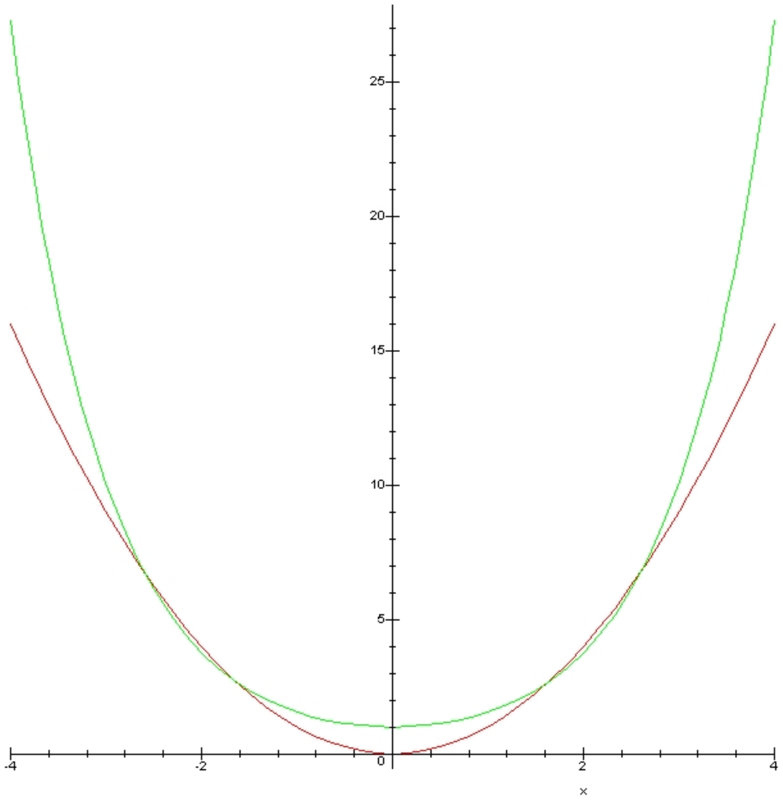


Figure 16.158: **Plot between a chain and parabola in Maple™ 4.00b**

In real life as an artwork:

Figure 16.159: **The Gateway Arch St. Louis arch (source: Wikipedia, author: Daniel Schwen)**

Let us consider now two points (x_1, y_1) and (x_2, y_2) in the plane and let us determine the equation of the chain of length L having these two points as ends.

We have the two equations:

$$\begin{cases} y_1 = k\cosh\left(\dfrac{x_1}{k} + c_1^{te}\right) + c_2^{te} \\ y_2 = k\cosh\left(\dfrac{x_2}{k} + c_1^{te}\right) + c_2^{te} \end{cases} \tag{16.4.51}$$

We obtain a third equation using the length L which is known. Indeed (see section Analytical Mechanics page 3172):

$$\begin{aligned} L &= \int_{x_1}^{x_2} \sqrt{1+y'^2}\mathrm{d}x = \int_{x_1}^{x_2} \sqrt{1+\sinh^2\left(\frac{x}{k}+c_1^{te}\right)} = \int_{x_1}^{x_2} \cosh\left(\frac{x}{k}+c_1^{te}\right)\mathrm{d}x \\ &= k\left(\sinh\left(\frac{x_2}{k}+c_1^{te}\right) - \sinh\left(\frac{x_1}{k}+c_1^{te}\right)\right) \end{aligned} \tag{16.4.52}$$

where we always have:

$$k = \frac{T_0}{\mu g} \tag{16.4.53}$$

Thus, we get a non-linear system of three equations with three unknowns (k, c_1^{te}, c_2^{te}):

$$\begin{cases} y_1 = k\cosh\left(\dfrac{x_1}{k} + c_1^{te}\right) + c_2^{te} \\ y_2 = k\cosh\left(\dfrac{x_2}{k} + c_1^{te}\right) + c_2^{te} \\ L = k\left(\sinh\left(\dfrac{x_2}{k}+c_1^{te}\right) - \sinh\left(\dfrac{x_1}{k}+c_1^{te}\right)\right) \end{cases} \tag{16.4.54}$$

✍**Example:**

Let us determine for example the chain of length 38 [cm] passing through the points $(-9, 0)$ and $(9, 10)$.

The following system must then be solved:

$$\begin{cases} 0 = k\cosh\left(\dfrac{-9}{k} + c_1^{te}\right) + c_2^{te} \\ 10 = k\cosh\left(\dfrac{9}{k} + c_1^{te}\right) + c_2^{te} \\ 38 = k\left(\sinh\left(\dfrac{9}{k}+c_1^{te}\right) - \sinh\left(\dfrac{-9}{k}+c_1^{te}\right)\right) \end{cases} \tag{16.4.55}$$

Here are the Maple™ 4.00b commands that allow us to get the result.

```
> E1:  = 0 = k * cosh (-9 / k + c1) + c2;
> E2:  = 10 = k * cosh (9 / k + c1) + c2;
> E3:  = 38 = k * (sinh (9 / k + c1) -sinh (-9 / k + c1));
> Fsolve (e1, e2, e3, k, c1, c2, k = 0..infinity);
```

Maple™ gives:

```
K = 4.073758798, c1 = .2694982504, c2 = -14.46356329
```

Graphically we have then:

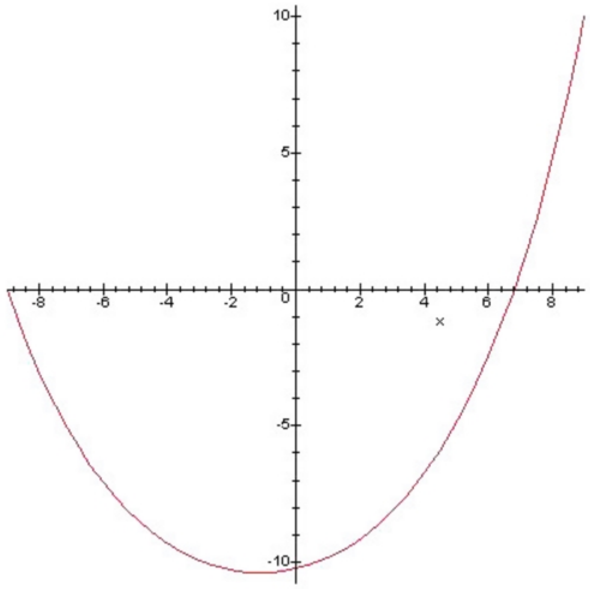


Figure 16.160: Plot of a small overhead cable in Maple™ 4.00b

- Second approach:

  For this second approach of solving the differential equation, we will keep the notation proposed by the reader who contacted us. We start from the differential equation:

$$\frac{\mathrm{d}^2y}{\mathrm{d}x^2}=\frac{1}{k}\sqrt{1+\left(\frac{\mathrm{d}y}{\mathrm{d}y}\right)^2} \tag{16.4.56}$$

  We make the following change of variable:

$$u=\frac{\mathrm{d}y}{\mathrm{d}x} \tag{16.4.57}$$

  Therefore it comes:

$$\frac{\mathrm{d}u}{\mathrm{d}x}=\frac{1}{k}\sqrt{1+u^2} \tag{16.4.58}$$

  and then:

$$\frac{\mathrm{d}u}{\sqrt{1+u^2}}=\frac{\mathrm{d}x}{k} \tag{16.4.59}$$

  The integration gives according to the usual primitive proved in the section of Differential and Integral Calculus (see page 1534):

$$\operatorname{arcsinh}(u)=\frac{x}{k}+c^{te} \tag{16.4.60}$$

  And the condition:

$$\frac{\mathrm{d}y}{\mathrm{d}x}=u=0 \tag{16.4.61}$$

on $x=0$ (lowest point of the overhead string for recall) imposes that the integration constant is zero and therefore that the points of attachment are by symmetry located at the same height. We have then:

$$\text{arcsinh}(u)=\frac{x}{k}\Rightarrow u=\frac{\mathrm{d}y}{\mathrm{d}x}=\sinh\left(\frac{x}{k}\right) \tag{16.4.62}$$

Therefore:

$$\boxed{y=k\cosh\left(\frac{x}{k}\right)+c^{te}} \tag{16.4.63}$$

that has to be compared with the previous method where we had obtained:

$$y=\frac{T_0}{\mu g}\cosh\left(x\frac{\mu g}{T_0}+c_1^{te}\right)+c_2^{te} \tag{16.4.64}$$

The integration constant is determined by the points of attachment of the overhead cable distant from a distance D where we have $y=0$ on $x=-D/2$ and $x=D/2$ (so the ends are on the same horizontal). We have then:

$$y(x)=k\left[\cosh\left(\frac{x}{k}\right)-\cosh\left(\frac{D}{2k}\right)\right] \tag{16.4.65}$$

It is under this last form that the overhead cable was obtained independently by Leibniz, Huygens and Bernoulli at the end of the 117th century (holocene calendar)...!

We can now easily calculate the maximum deviation (the "arrow" as the engineers in civil engineering say, that we will denoted by the letter f as in the section of Mechanical Engineering) with respect to the horizontal passing through the points of attachment:

$$\boxed{y(0)=f=k\left[1-\cosh\left(\frac{D}{2k}\right)\right]} \tag{16.4.66}$$

Finally it can be interesting to calculate the length of the cable at equilibrium. For this, let us recall that we have proved earlier above that:

$$\begin{aligned}L&=\int\limits_{x_1}^{x_2}\sqrt{1+y'^2}\mathrm{d}x=\int\limits_{x_1}^{x_2}\sqrt{1+\sinh^2\left(\frac{x}{k}+c_1^{te}\right)}\mathrm{d}x=\int\limits_{x_1}^{x_2}\cosh\left(\frac{x}{k}+c_1^{te}\right)\mathrm{d}x\\&=k\left(\sinh\left(\frac{x_2}{k}+c_1^{te}\right)-\sinh\left(\frac{x_1}{k}+c_1^{te}\right)\right)\end{aligned} \tag{16.4.67}$$

In the case where the fasteners are at the same height, the chain is symmetrical and we then have:

$$L=2\int\limits_{0}^{D/2}\cosh\left(\frac{x}{k}+c^{te}\right)\mathrm{d}x=2k\left(\sinh\left(\frac{1}{k}\frac{D}{2}\right)-\sinh\left(c^{te}\right)\right) \tag{16.4.68}$$

It remains to determine the constant! If D is zero then L must be zero. The constant is then zero and it remains:

$$\boxed{L=2k\sinh\left(\frac{1}{k}\frac{D}{2}\right)} \tag{16.4.69}$$

In the case of electrified railway lines, we solve the problem of the arrow by a main cable that carries of the catenary: the upper cable (below on the photo) undergoes an accepted deflection, which reduces the tension between the pylons. The catenary thus remains very linear thanks to the multiple auxiliary hooks to an auxiliary cable.

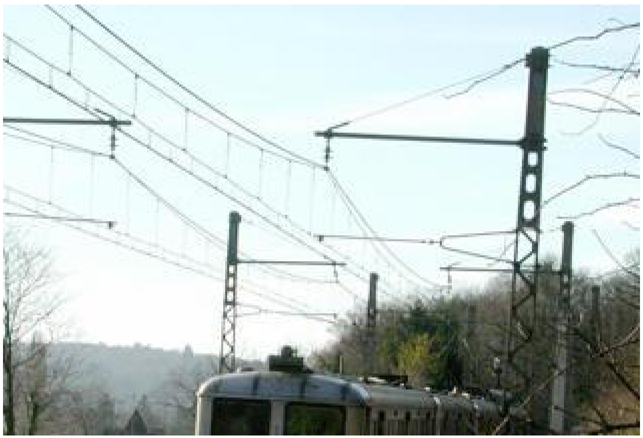


Figure 16.161: **Overhead cable of railways (source: Chronomaths)**

Otherwise let us also notice that we also find the overhead cable in all the places of the life of every day where a cable is suspended between two points on the same horizontal.

✍**Example:**

Consider a suspended cable with the following data:

$$\mu=0.5\,[\text{kg}\cdot\text{m}^{-1}]\qquad g=9.81\,[\text{m}\cdot\text{s}^{-2}]\qquad D=20\,[\text{m}]\qquad T_0=45\,[\text{N}]\tag{16.4.70}$$

We have since the linear load of the cable is constant (taking the notation of the section of Mechanical Engineering on the way...):

$$k=\frac{T_0}{\mu g}=\frac{T_0}{w_0}\cong 9.17\,[\text{m}]\tag{16.4.71}$$

and therefore:

$$y(0)=f=k\left[1-\cosh\left(\frac{D}{2k}\right)\right]=9.17\left[1-\cosh\left(\frac{20}{2\cdot 9.17}\right)\right]\cong -6\,[\text{m}]\tag{16.4.72}$$

Thus the arrow of the cable is 6 meters below the horizontal of the two hangers.

16.4.4.2 Charges overhead cable (suspended bridge)

Let us consider the following suspended bridge carrying a constant linear load, hence the arrow h is imposed by the architect, as well as the distance D between the two pillars:

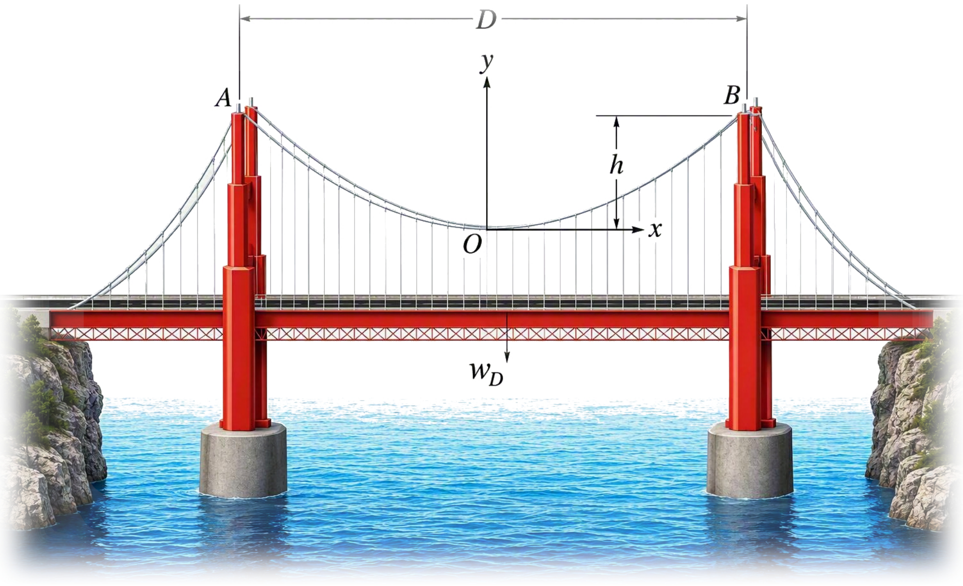


Figure 16.162: **Suspension bridge (source: ISBN 0-13-814929-1)**

Remark

We will assume that Earth's curvature is negligible in this particular example because in reality suspended bridge have their tower that are not parallel like the Verrazzano-Narrows Bridge (New-York/USA) where its monumental 194.76 meters high towers are 41.275 millimetres farther apart at their tops than at their bases because the 1298.448 meters distance between them made it necessary to compensate for the Earth's curvature.

The subtlety of this case study lies in the fact that the linear load is no longer in the cable itself but in the span of the bridge which has a much higher linear mass. We can no longer use the analyses given above. The development we did earlier above for a normal suspended cable remains valid only until the following relation:

$$\mathrm{d}(\tan(\alpha)) = \frac{\mathrm{d}L}{k} \tag{16.4.73}$$

Now, we must remember that we also have:

$$\mathrm{d}(\tan(\alpha)) = \mathrm{d}\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right) = \frac{\mathrm{d}L}{k} \tag{16.4.74}$$

by rearranging and deriving again by $\mathrm{d}x$:

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right) = \frac{1}{k}\frac{\mathrm{d}L}{\mathrm{d}x} \tag{16.4.75}$$

On the bridge however, each portion $\mathrm{d}L$ of the cable is negligible in front of the portion $\mathrm{d}x$ of the bridge. This marks the difference between the suspended bridge and the simple overhead cable as we need to replace $\mathrm{d}L$ by $\mathrm{d}x$ (it is quite subtle with this approach but there are several possible approaches to this development). We then have the preceding relation which becomes:

$$\frac{\mathrm{d}^2y}{\mathrm{d}x^2}=\frac{1}{k}\frac{\mathrm{d}x}{\mathrm{d}x}=\frac{1}{k} \tag{16.4.76}$$

Either after rearrangement:

$$\mathrm{d}^2y=\left(\frac{1}{k}\mathrm{d}x\right)\mathrm{d}x \tag{16.4.77}$$

By integrating a first time it comes then:

$$\mathrm{d}y=\left(\frac{1}{k}\mathrm{d}x\right)x+C^{te} \tag{16.4.78}$$

And a second and last time:

$$y=\frac{1}{2k}x^2+C_1x+C_2 \tag{16.4.79}$$

The constants are determined by the initial conditions. The place where we have placed the reference frame imposes that $y=0$ in $x=0$ and that $\mathrm{d}y/\mathrm{d}x=0$ in $x=0$, we have the two constant which are zero and it then remains:

$$y=\frac{1}{2k}x^2=\frac{\mu g}{2T_0}x^2=\frac{w_0}{2T_0}x^2 \tag{16.4.80}$$

It is therefore the equation of a parabola and not of a overhead cable anymore! A small concern is that we do not know the tension (stress) in the cable. We should get rid of it. Now, because the architect requires that the arrow be h at $x=D/2$, then it comes:

$$T_0=\frac{w_0D^2}{8h} \tag{16.4.81}$$

therefore:

$$\boxed{y=\frac{4h}{D^2}x^2} \tag{16.4.82}$$

Again this remains a parable regardless of the weight! This is due to the fact that it is uniform in this case.

We then have to determine the length of the cable. For this we take up the relation already recalled above (proved in the section of Analytical Mechanics at page 3172 and in the section of Geometric Shapes at page 2956):

$$L=\int\limits_{x_1}^{x_2}\sqrt{1+y'^2}\,\mathrm{d}x \tag{16.4.83}$$

By the vertical symmetry of the function of the suspension bridge we can write:

$$\begin{aligned}L&=\int\limits_{x_1}^{x_2}\sqrt{1+y'^2}\mathrm{d}x=2\int\limits_{0}^{D/2}\sqrt{1+\left(\frac{4h}{D^2}x^2\right)'^2}\mathrm{d}x=2\int\limits_{0}^{D/2}\sqrt{1+\left(\frac{8h}{D^2}x\right)^2}\mathrm{d}x\\&=2\frac{8h}{D^2}\int\limits_{0}^{D/2}\sqrt{\left(\frac{D^2}{8h}\right)^2+x^2\mathrm{d}x}\end{aligned}\tag{16.4.84}$$

In the section of Differential and Integral Calculus in our proofs of the usual primitives (see page 1534), we have proved for recall that:

$$\int\sqrt{a^2+x^2}\mathrm{d}x=\frac{a^2}{2}\ln\left(x+\sqrt{a^2+x^2}\right)+\frac{x}{2}\sqrt{a^2+x^2}\tag{16.4.85}$$

Therefore it comes:

$$L=\frac{16h}{D^2}\int\sqrt{\left(\frac{D^2}{8h}\right)^2+x^2\mathrm{d}x}=\frac{16h}{D^2}\left[\frac{1}{2}\left(\frac{D^2}{8h}\right)^2\ln\left(x+\sqrt{\left(\frac{D^2}{8h}\right)^2+x^2}\right)+\frac{x}{2}\sqrt{\left(\frac{D^2}{8h}\right)^2+x^2}\right]\Bigg|_0^{D/2}\tag{16.4.86}$$

Or simplified:

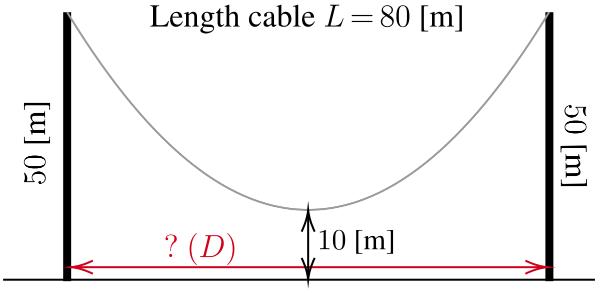
$$L=\frac{4\sqrt{\frac{D^2(D^2+16h^2)}{h^2}}h^2+D^3\ln\left(4D+\sqrt{\frac{D^2(D^2+16h^2)}{h^2}}\right)-2D^3\ln(D)+D^3\ln(h)}{8hD}\tag{16.4.87}$$

✍**Example:**

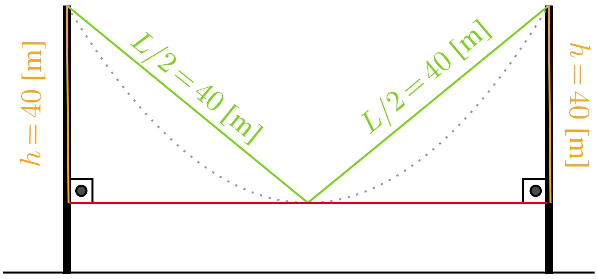
Let us consider the characteristics of the Golden Bridge of San Francisco. Its arrow h is about 230 [m] and its main bearing is $D=1280$ [m]. We have then:

$$L=\frac{4\sqrt{\frac{D^2(D^2+16h^2)}{h^2}}h^2+D^3\ln\left(4D+\sqrt{\frac{D^2(D^2+16h^2)}{h^2}}\right)-2D^3\ln(D)+D^3\ln(h)}{8hD}\cong 1383\,[\mathrm{m}]\tag{16.4.88}$$

Now the reader will be aware why the legendary (myth or reality who knows...?) below Google hiring test is only a logical test (because inverting the above relation would lead to a time-out expected of geniuses):



If you solve the above quite huge equation with Maple™ you get that $L = 80$ [m] when $D \to 0_+$ (but not $D = 0$ otherwise you get a division by zero) using $h = 40$ [m]. What the myth says is that you are supposed (because of the limited time available) to do the following rude approximation by "thinking out of the box":



that leads to the same result anyway using Pythagoras theorem:

$$40^2 = 40^2 + \text{base}^2 \Rightarrow \text{base} = 0 \quad (16.4.89)$$

16.4.4.3 Very tense cable

Let us recall that in the general case we get the following relation:

$$\vec{T}+\vec{T}_0+\mu L\vec{g}=\vec{0} \tag{16.4.90}$$

Let us take the differential:

$$\mathrm{d}\vec{T}+\underbrace{\mathrm{d}\vec{T}_0}_{=0}+\mu\vec{g}\mathrm{d}L=\vec{0} \tag{16.4.91}$$

Thus for the component y:

$$\mathrm{d}\vec{T}_y+\mu g\mathrm{d}L=0 \tag{16.4.92}$$

But we also have:

$$\frac{T_y}{T_x}=\frac{\mathrm{d}y}{\mathrm{d}x}=\tan(\alpha) \tag{16.4.93}$$

So injected into the previous relation this gives:

$$T_x\mathrm{d}\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)+\mu g\mathrm{d}L=0 \tag{16.4.94}$$

Therefore:

$$T_x\mathrm{d}\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)=-\mu g\mathrm{d}L \tag{16.4.95}$$

And under the assumption that the cable is very tight (under hight stress):

$$\mathrm{d}L\cong\mathrm{d}x \tag{16.4.96}$$

Therefore it comes:

$$T_x\mathrm{d}\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)=-\mu g\mathrm{d}x \tag{16.4.97}$$

After a first integration it comes:

$$T_x\frac{\mathrm{d}y}{\mathrm{d}x}=-\mu gx+c^{te} \tag{16.4.98}$$

Before going further, let us notice that the constant is easy to determine since in $x=L/2$, the derivative must be zero. We have then:

$$T_x\frac{\mathrm{d}y}{\mathrm{d}x}=-\mu gx+\mu g\frac{L}{2} \tag{16.4.99}$$

After a second integration (where the integration constant is zero):

$$T_xy=-\frac{\mu g}{2}(xL-x^2) \tag{16.4.100}$$

We then have for the deformation of the string:

$$\boxed{y=-\frac{\mu g}{T_x 2}(xL-x^2)} \tag{16.4.101}$$

Thus, in the middle of the cable (string), the arrow is then:

$$f=-\frac{\mu g}{T_x 2}\left(\frac{L}{2}L-\frac{L^2}{4}\right)=-\frac{\mu g L^2}{8T_x} \tag{16.4.102}$$

Or respectively if the arrow is known and the tension is sought to be determined:

$$T_x=-\frac{\mu g L^2}{8f} \tag{16.4.103}$$

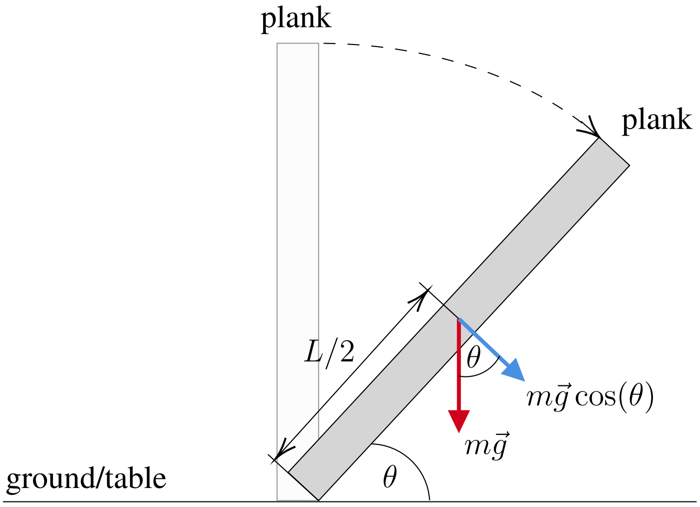
And we could also calculate the length of the cable (string) using the same technique as for the suspension bridge.

16.4.5 Falling chimney (naive approach)

The blasting of a falling chimney problem is a well known study in high-school and an interesting application of circular motion kinematics and moment of inertia. It is especially interesting as we can find many approach to this problem on Internet and textbooks but many known one are vicious as they hide important assumptions or are even wrong but still give the correct final result.

Our purpose here is to give the simplest possible way to determine the result without forgetting to enumerate any assumption! Our purpose is also to take this opportunity to introduce first a physical phenomenon that also appears within the falling chimney study framework that some people consider as "magic": the Hinged Stick and Ball trick (i.e. sometimes named "Galileo Paraxodx")!

Let us consider a plank that is able to pivot about one end, with the other end free to move. As in the figure below:



We assume:

- There is no friction at the pivot point
- We are in vacuum
- The plank is infinitely rigid
- The gravitational force is constant at any point
- The plank is assimilated to a cylinder

It is of interest to find the acceleration of various points on the plank. First, consider the rotational dynamics of the system while the plank is in motion. Once the plank is released from its rest position, the two forces acting on it are gravity and a normal force. The latter is exerted by the table at the pivot, so it does not exert a torque about the pivot. The gravitational force does exert a torque on the plank about the pivot, giving the net torque τ (we change the notation of the torque that is normally M in this textbook because later we will have M for the notation of the whole mass so it would be a mess if we don't change now!) on the plank as:

$$\tau = \frac{1}{2}mgL\cos(\theta) \tag{16.4.104}$$

where m is the mass of the plank, g is the acceleration due to gravity $g = 9.81\;[\text{m}\cdot\text{s}^{-2}]$, l is the plank's length and θ is the angle between the raised plank and the table.

The rotational equivalent of Newton's second law applies to the plank's motion, namely:

$$\tau = J\frac{\mathrm{d}\omega}{\mathrm{d}t} = J\ddot{\theta} \tag{16.4.105}$$

And we have proven in the section of Geometrical Shapes (see page 2959) that the moment of inertia of a cylinder rotating around a fixed extremity point was given by:

$$J = \frac{1}{3}mL^2 \tag{16.4.106}$$

Therefore:

$$\tau = \frac{1}{3}mL^2\ddot{\theta} \tag{16.4.107}$$

Next, the linear acceleration $a(x,\theta)$ of a point on the plank that is at a distance x from the pivot is given by (dimensional analysis):

$$a = x\ddot{\theta} \tag{16.4.108}$$

Therefore we can write:

$$\tau = \frac{1}{3}mL^2\ddot{\theta} = \frac{1}{3}mL^2\frac{a}{x} \tag{16.4.109}$$

And we can equate the both M to get:

$$\frac{1}{2}mgL\cos(\theta) = \frac{1}{3}mL^2\frac{a}{x} \tag{16.4.110}$$

from which it is possible to solve for a, with the result being:

$$\boxed{a = \frac{3gx}{2L}\cos(\theta)} \tag{16.4.111}$$

Let $x = L$, so that the point of consideration is the moving end of the stick. Then it remains (notice that the mass m does not appear!):

$$a = \frac{3}{2}g\cos(\theta) \tag{16.4.112}$$

Therefore we have that $a > g$ for the of the stick when:

$$\cos(\theta) \geq \frac{2}{3} = 0.841\;[\text{rad}] = 48.19^\circ \tag{16.4.113}$$

As we know that:

$$x=x_0+v_0t+\frac{1}{2}at^2 \tag{16.4.114}$$

and in our case the plank has $x_0=0$ and $v_0=0$ it comes:

$$t_{\text{falling}}=\sqrt{\frac{2x}{a}} \tag{16.4.115}$$

The distance x we are looking for depends on the initial angle θ but is trivially given (definition of the radian!) by for the extremity of the plank:

$$x=\theta L \tag{16.4.116}$$

Therefore (notice that the mass m again... does not appear!):

$$t_{\text{falling}}=\sqrt{\frac{2\theta L}{a}}=\sqrt{\frac{2\theta L}{\frac{3}{2}g\cos(\theta)}}=\sqrt{\frac{4}{3}\frac{\theta L}{g\cos(\theta)}} \tag{16.4.117}$$

✍**Example:**

If we consider a plastic ruler of $l=30$ [cm] and of mass $m=68$ [g] with starting angle $\theta=\pi/4$ [rad] we get:

$$t_{\text{falling}}=\sqrt{\frac{4}{3}\frac{\theta L}{g\cos(\theta)}}\cong 0.21\text{ [s]} \tag{16.4.118}$$

OK this basic stuff being done. Let us now focus on the chimney rupture point.

In the frame of the falling chimney we have for a perfect rigid chimney:

$$\mathrm{d}mx\ddot{\theta}=\mathrm{d}mg\cos(\theta) \tag{16.4.119}$$

But first we have:

$$\mathrm{d}m=\frac{M}{L}\mathrm{d}x \tag{16.4.120}$$

Therefore:

$$\frac{M}{L}x\ddot{\theta}\mathrm{d}x=\frac{M}{L}g\cos(\theta)\mathrm{d}x \tag{16.4.121}$$

or what remains the same:

$$-\frac{M}{L}x\ddot{\theta}\mathrm{d}x+\frac{M}{L}g\cos(\theta)\mathrm{d}x=0 \tag{16.4.122}$$

But in reality what we can observe is that a falling object consisting on a of smaller parts (bricks) breaks apart on some points:

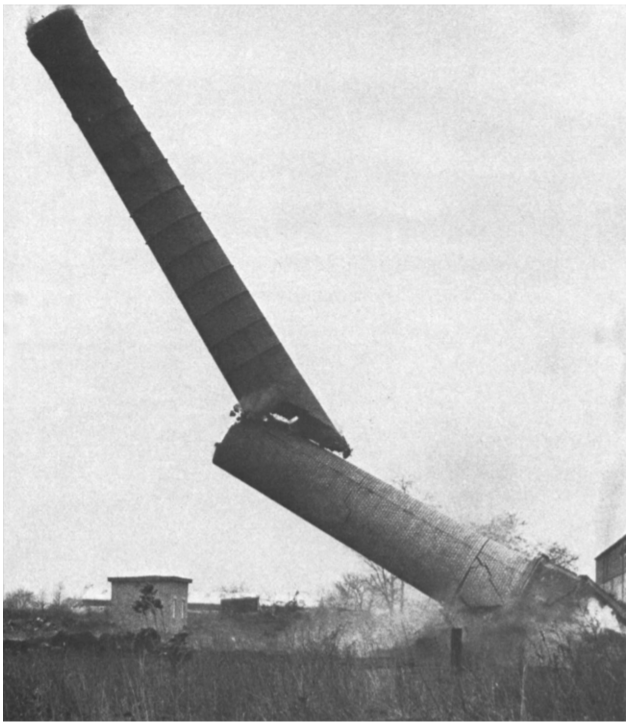


Figure 16.163: **A special case of falling chimney**

Obviously for chimneys the smaller bricks are not independent and the point of rupture depends on the material properties. So we will imagine a chimney made of a stack of bricks without any strengthening. So in this situation we observe that a falling stacks of bricks always break into parts. So the previous relations is not complete. We don't know the origin of the force that breaks apart but we know that we can insert a term that contains all possible sources of internal forces just by writing:

$$-\frac{M}{L}x\ddot{\theta}\mathrm{d}x+\frac{M}{L}g\cos(\theta)\mathrm{d}x=\mathrm{d}F \qquad (16.4.123)$$

Or we also could have written:

$$\frac{M}{L}x\ddot{\theta}\mathrm{d}x-\frac{M}{L}g\cos(\theta)\mathrm{d}x=\mathrm{d}F \qquad (16.4.124)$$

where $\mathrm{d}F$ is the force due to the mass that is above $\mathrm{d}m$.

Remark

Another way to explain this is to remember that we have proved earlier above that:

$$a=\frac{3gx}{2L}\cos(\theta) \qquad (16.4.125)$$

Therefore the angular acceleration depends on the length of the chimney. Now if the force of gravity were the only force acting on $\mathrm{d}m$, the angular acceleration on $\mathrm{d}m$ would be independent of the length of the chimney.

Now we know that what breaks such a hypothetical chimney is a torque!

An element at distance x_0 from 0 experiences a torque due to the rest of the chimney above it.

This torque is given by:

$$\begin{aligned}
\tau_{x_0} &= \frac{M}{L}\ddot{\theta}\int\limits_{x_0}^{L} x(x-x_0)\mathrm{d}x - \frac{M}{L}g\cos(\theta)\int\limits_{x_0}^{L}(x-x_0)\mathrm{d}x \\
&= \frac{M}{L}\left[\ddot{\theta}\int\limits_{x_0}^{L} x(x-x_0)\mathrm{d}x - g\cos(\theta)\int\limits_{x_0}^{L}(x-x_0)\mathrm{d}x\right] \\
&= \frac{M}{L}\left[\ddot{\theta}\left(\int\limits_{x_0}^{L} x^2\mathrm{d}x - x_0\int\limits_{x_0}^{L} x\mathrm{d}x\right) - g\cos(\theta)\left(\int\limits_{x_0}^{L} x\mathrm{d}x - x_0\int\limits_{x_0}^{L}\mathrm{d}x\right)\right] \\
&= \frac{M}{L}\left[\ddot{\theta}\left(\frac{1}{3}x^3 - \frac{1}{2}x_0x^2\right)\Bigg|_{x_0}^{L} - g\cos(\theta)\left(\frac{1}{2}x^2 - x_0x\right)\Bigg|_{x_0}^{L}\right] + c^{te} \\
&= \frac{M}{L}\left[\ddot{\theta}\left(\frac{1}{3}L^3 - \frac{1}{2}x_0L^2 - \frac{1}{3}x_0^3 + \frac{1}{2}x_0x_0^2\right) - g\cos(\theta)\left(\frac{1}{2}L^2 - x_0L - \frac{1}{2}x_0^2 + x_0x_0\right)\right] + c^{te} \\
&= \frac{M}{L}\left[\ddot{\theta}\left(\frac{1}{3}L^3 - \frac{1}{2}x_0L^2 + \frac{1}{6}x_0^3\right) - g\cos(\theta)\left(\frac{1}{2}L^2 - x_0L + \frac{1}{2}x_0^2\right)\right] + c^{te} \\
&= \frac{M}{L}\left[\ddot{\theta}\left(\frac{1}{6}x_0^3 + \frac{1}{3}L^3 - \frac{1}{2}x_0L^2\right) + g\cos(\theta)\left(x_0L - \frac{1}{2}L^2 - \frac{1}{2}x_0^2\right)\right] + c^{te}
\end{aligned} \tag{16.4.126}$$

The extremal torques (maxima and minima) are found by:

$$\frac{\mathrm{d}\tau_{x_0}}{\mathrm{d}x_0} = 0 \tag{16.4.127}$$

Therefore after derivation by x_0 it remains only:

$$0 = \frac{M}{L}\left[\ddot{\theta}\left(\frac{1}{2}x_0^2 - \frac{1}{2}L^2\right) + g\cos(\theta)\left(L - x_0\right)\right] \tag{16.4.128}$$

Or simplified a bit more:

$$\ddot{\theta}\left(\frac{1}{2}x_0^2 - \frac{1}{2}L^2\right) + g\cos(\theta)\left(L - x_0\right) = 0 \tag{16.4.129}$$

So there is a first obvious visible solution $x_0 = L$ that gives:

$$\tau_{x_0=L} = \frac{M}{L}\left[\ddot{\theta}\left(\frac{1}{6}x_0^3 + \frac{1}{3}L^3 - \frac{1}{2}x_0L^2\right) + g\cos(\theta)\left(x_0L - \frac{1}{2}L^2 - \frac{1}{2}x_0^2\right)\right] + c^{te} = 0 \tag{16.4.130}$$

So this is the minima!

Now remember that:

$$\ddot{\theta} = \frac{3}{2}\frac{g\cos(\theta)}{L} \tag{16.4.131}$$

Therefore:

$$\frac{3}{2}\frac{g\cos(\theta)}{L}\left(\frac{1}{2}x_0^2-\frac{1}{2}L^2\right)+g\cos(\theta)\,(L-x_0)=0 \tag{16.4.132}$$

After simplification (assuming $\cos(\theta)\neq 0$):

$$\frac{3}{2L}\left(\frac{1}{2}x_0^2-\frac{1}{2}L^2\right)+L-x_0=0 \tag{16.4.133}$$

Or:

$$\frac{3}{4}x_0^2-\frac{3}{4}L^2+L^2-Lx_0=0 \tag{16.4.134}$$

Hence:

$$\frac{1}{4}L^2-Lx_0+\frac{3}{4}x_0^2=0 \tag{16.4.135}$$

Or:

$$3x_0^2-4Lx_0+L^2=0 \tag{16.4.136}$$

This is a second degree polynomial on x_0. Therefore (see section Calculus page 1383):

$$x_{0(1,2)}=\frac{4L\pm\sqrt{(-4L)^2-4\cdot 3\cdot L^2}}{6}=\frac{4L\pm 2L}{6}=\left\{L,\frac{L}{3}\right\} \tag{16.4.137}$$

The maximum torque is therefore found at distance $x_0=L/3$ from 0. This is where the chimney might POSSIBLY break.

16.4.6 Dams

A dam is a barrier that impounds water or underground streams. Reservoirs created by dams not only suppress floods but also provide water for activities such as irrigation, human consumption, industrial use, aquaculture, and navigability. Hydropower is often used in conjunction with dams to generate electricity. A dam can also be used to collect water or for storage of water which can be evenly distributed between locations. Dams generally serve the primary purpose of retaining water, while other structures such as floodgates or levees (also known as dikes) are used to manage or prevent water flow into specific land regions.

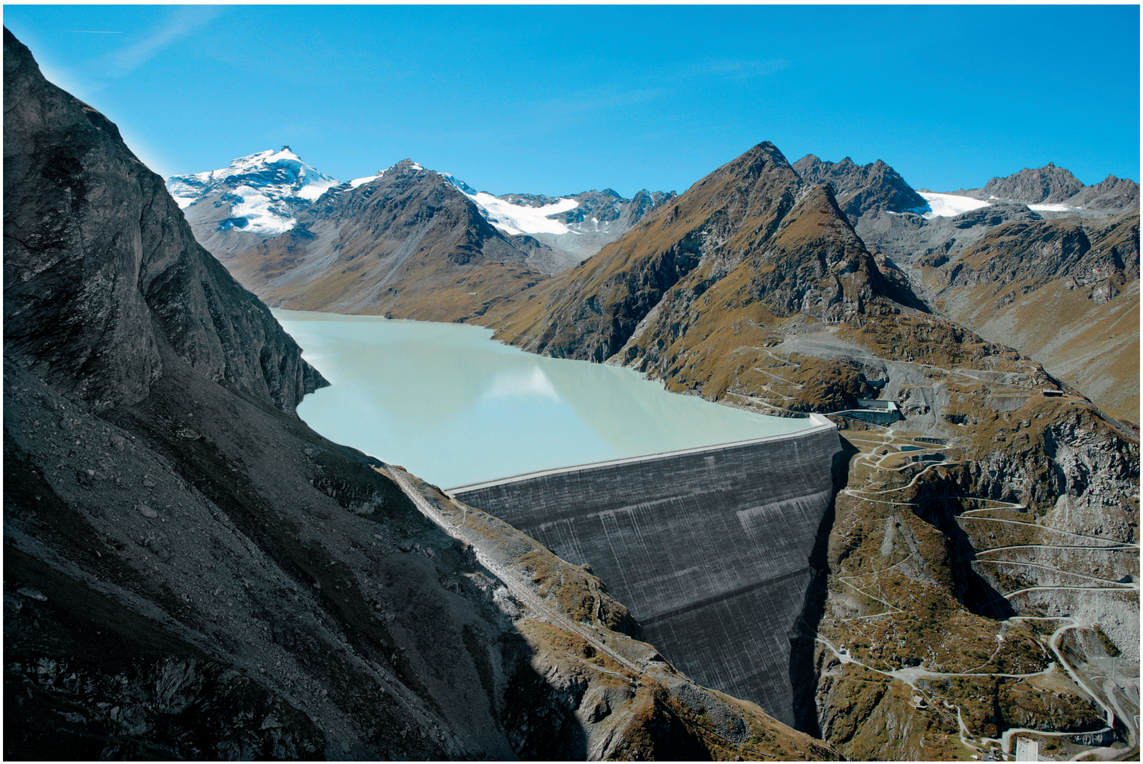


Figure 16.164: **Dixence (Switzerland) dam**

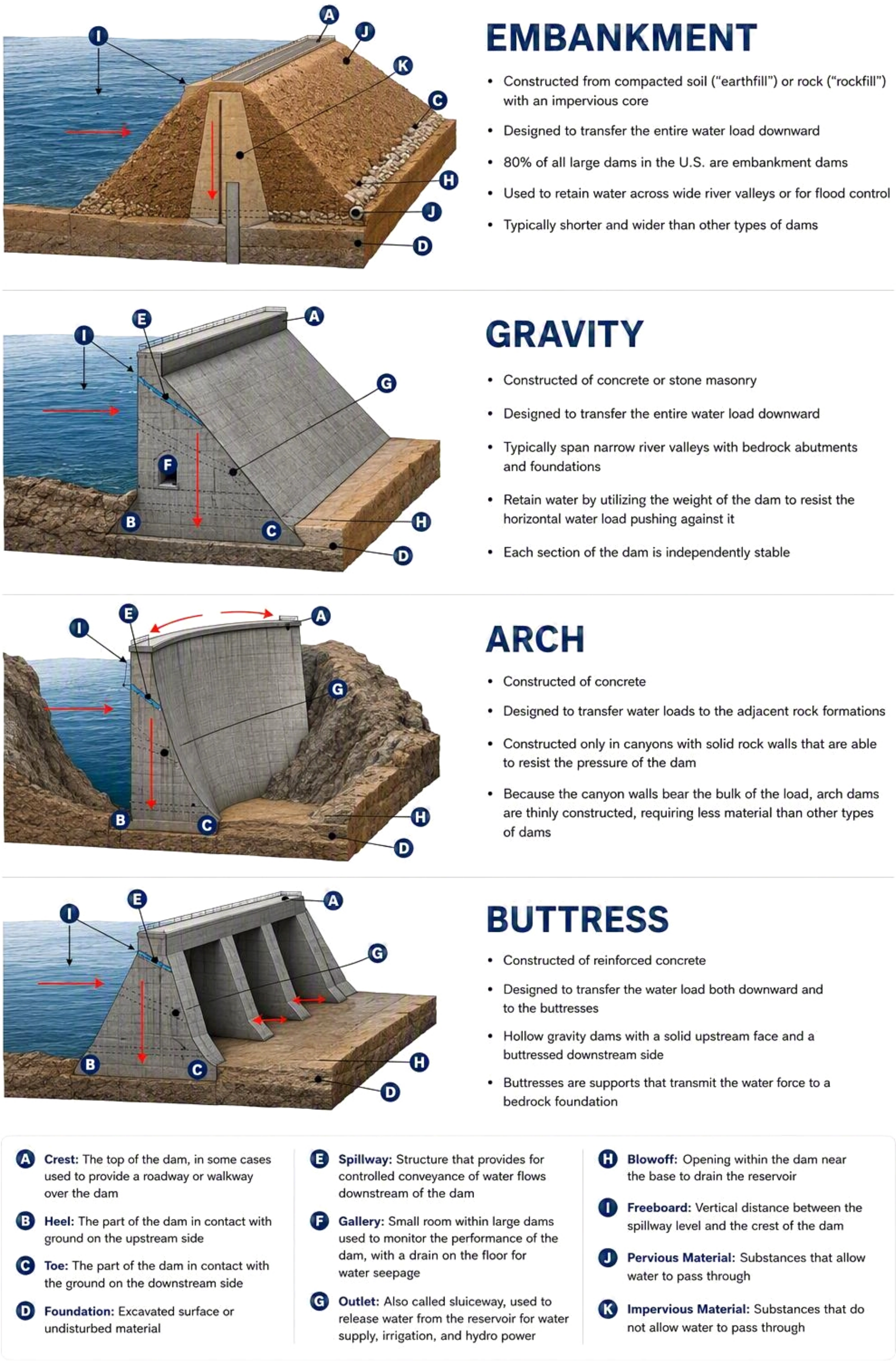
Remark

When dealing with dams or any similar configuration of full recipient, keep in mind this famous value in the field of dams power plants based on the hydrostatic pressure of the section of Continuum Mechanics at page 3791 (the final result is in bar):

$$\Delta P=\Delta(P_0+h\rho g)=\Delta h\rho g\cong 10\,[\text{m}]\cdot 997\,[\text{kg}\cdot\text{m}^{-3}]\cdot 9.81\,[\text{m}\cdot\text{s}^{-2}]\cong 1\,[\text{bar}] \tag{16.4.138}$$

Hence the famous rule, that for a dam, we have a surplus of $1\,[\text{bar}]$ for every $10\,[\text{m}]$! The Romans already knew this! therefore, in order to reduce the pressure, they created intermediate waterfalls to reduce the pressure.

There exist various type of dams as illustrated in the figure below:



EMBANKMENT

- Constructed from compacted soil ("earthfill") or rock ("rockfill") with an impervious core
- Designed to transfer the entire water load downward
- 80% of all large dams in the U.S. are embankment dams
- Used to retain water across wide river valleys or for flood control
- Typically shorter and wider than other types of dams

GRAVITY

- Constructed of concrete or stone masonry
- Designed to transfer the entire water load downward
- Typically span narrow river valleys with bedrock abutments and foundations
- Retain water by utilizing the weight of the dam to resist the horizontal water load pushing against it
- Each section of the dam is independently stable

ARCH

- Constructed of concrete
- Designed to transfer water loads to the adjacent rock formations
- Constructed only in canyons with solid rock walls that are able to resist the pressure of the dam
- Because the canyon walls bear the bulk of the load, arch dams are thinly constructed, requiring less material than other types of dams

BUTTRESS

- Constructed of reinforced concrete
- Designed to transfer the water load both downward and to the buttresses
- Hollow gravity dams with a solid upstream face and a buttressed downstream side
- Buttresses are supports that transmit the water force to a bedrock foundation

Ⓐ **Crest:** The top of the dam, in some cases used to provide a roadway or walkway over the dam

Ⓑ **Heel:** The part of the dam in contact with ground on the upstream side

Ⓒ **Toe:** The part of the dam in contact with the ground on the downstream side

Ⓓ **Foundation:** Excavated surface or undisturbed material

Ⓔ **Spillway:** Structure that provides for controlled conveyance of water flows downstream of the dam

Ⓕ **Gallery:** Small room within large dams used to monitor the performance of the dam, with a drain on the floor for water seepage

Ⓖ **Outlet:** Also called sluiceway, used to release water from the reservoir for water supply, irrigation, and hydro power

Ⓗ **Blowoff:** Opening within the dam near the base to drain the reservoir

Ⓘ **Freeboard:** Vertical distance between the spillway level and the crest of the dam

Ⓙ **Pervious Material:** Substances that allow water to pass through

Ⓚ **Impervious Material:** Substances that do not allow water to pass through

Figure 16.165: **Types of dams**

We will consider here, in a first time, only a "gravity dam" that is by definition a massive sized dam fabricated from concrete or stone masonry. Such dams are designed to hold back large volumes of water. By using concrete, the weight of the dam is actually able to resist the horizontal thrust of water pushing against it. This is why it is named a "gravity dam". Gravity essentially holds the dam down to the ground, stopping water from toppling it over.

Gravity dams are well suited for blocking rivers in wide valleys or narrow gorge ways. Since gravity dams must rely on their own weight to hold back water, it is necessary that they are built on a solid foundation of bedrock.

Therefore let us consider the following gravity dam (consisting of a solid with infinite rigidity and perfectly waterproof...) of height z, of length L and storing water of density ρ:

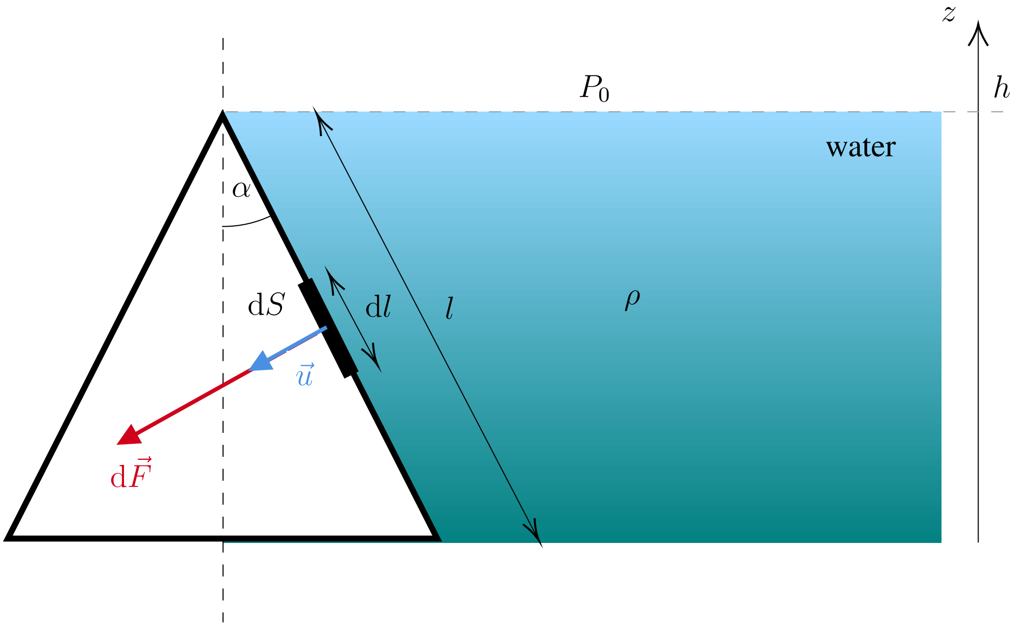


Figure 16.166: Simplified approach for the study of pressure on a gravity dam

We have proved in the section of Continuum Mechanics that the hydrostatic pressure was given by (see page 3791):

$$P = P_0 + h\rho g \tag{16.4.139}$$

But in this situation we obviously have:

$$P = P_0 + (h - z)\rho g \tag{16.4.140}$$

So when we place ourselves on the surface of the water at $z = h$:

$$P = P_0 \tag{16.4.141}$$

thus the pressure of the air at the surface of the dam sea...

On a surface element $\mathrm{d}S$ it applies an elementary force:

$$\mathrm{d}\vec{F} = P\mathrm{d}S = P\mathrm{d}S = \vec{u} = PL\mathrm{d}l\vec{u} \tag{16.4.142}$$

But:

$$\cos(\alpha)=\frac{z}{l}\Leftrightarrow \mathrm{d}l=\frac{\mathrm{d}z}{\cos(\alpha)} \tag{16.4.143}$$

Therefore:

$$\mathrm{d}\vec{F}=PL\mathrm{d}L\vec{u}=(P_0+(h-z)\rho g)L\frac{\mathrm{d}z}{\cos(\alpha)}\vec{u} \tag{16.4.144}$$

hence after integration:

$$F=PL\mathrm{d}l=\int_0^h(P_0+(h-z)\rho g)\frac{L}{\cos(\alpha)}\mathrm{d}z=\frac{Lh}{\cos(\alpha)}\left(P_0+\frac{1}{2}\rho gh\right) \tag{16.4.145}$$

It is therefore the force exerted on the immersed face. The force on the emerged face (left in the illustration) is simply given by putting $\rho=0$. We have therefore for the part due to atmospheric pressure alone:

$$F=\frac{Lh}{\cos(\alpha)}P_0 \tag{16.4.146}$$

and for the part due to water only:

$$\boxed{F=\rho gL\frac{h^2}{2\cos(\alpha)}} \tag{16.4.147}$$

Notice that in many undergraduate textbooks you will found the above relation but with $\alpha=0$.

Remark

On average between the empty state and full state (following what we can read on the Internet) the top of a gravity dam of certain height would sometimes move with an amplitude of the order of 80 [cm].

16.4.6.1 Power extracted from a Dam

A Dam (to simplify) makes electrical power by using the kinetic energy of water to turn a wheel, named a "Pelton wheel[16]" (there are other types of water turbines like the "Francis turbine" or the "Kaplan turbine") because it has a design optimized to extracted almost all of the water's impulse energy, which itself will drive a shaft that connects the turbine runner to the generator, this makes it possible to convert the mechanical energy from the rotating runner into electrical energy.

---

[16] Although Pelton turbines are very efficient, they are not as common as Kaplan and Francis turbines because the Pelton turbine requires a very large head of pressure (difference in elevation between the upper and lower reservoir) to operate; this restricts its use to certain geographical areas only.

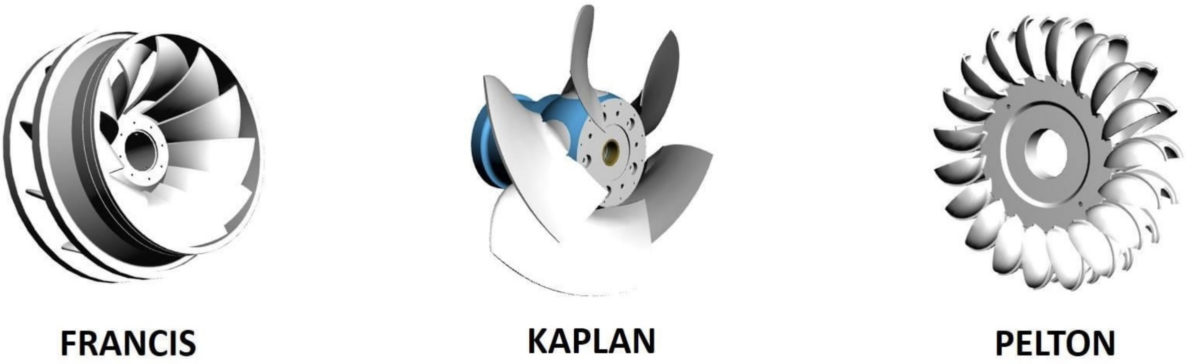
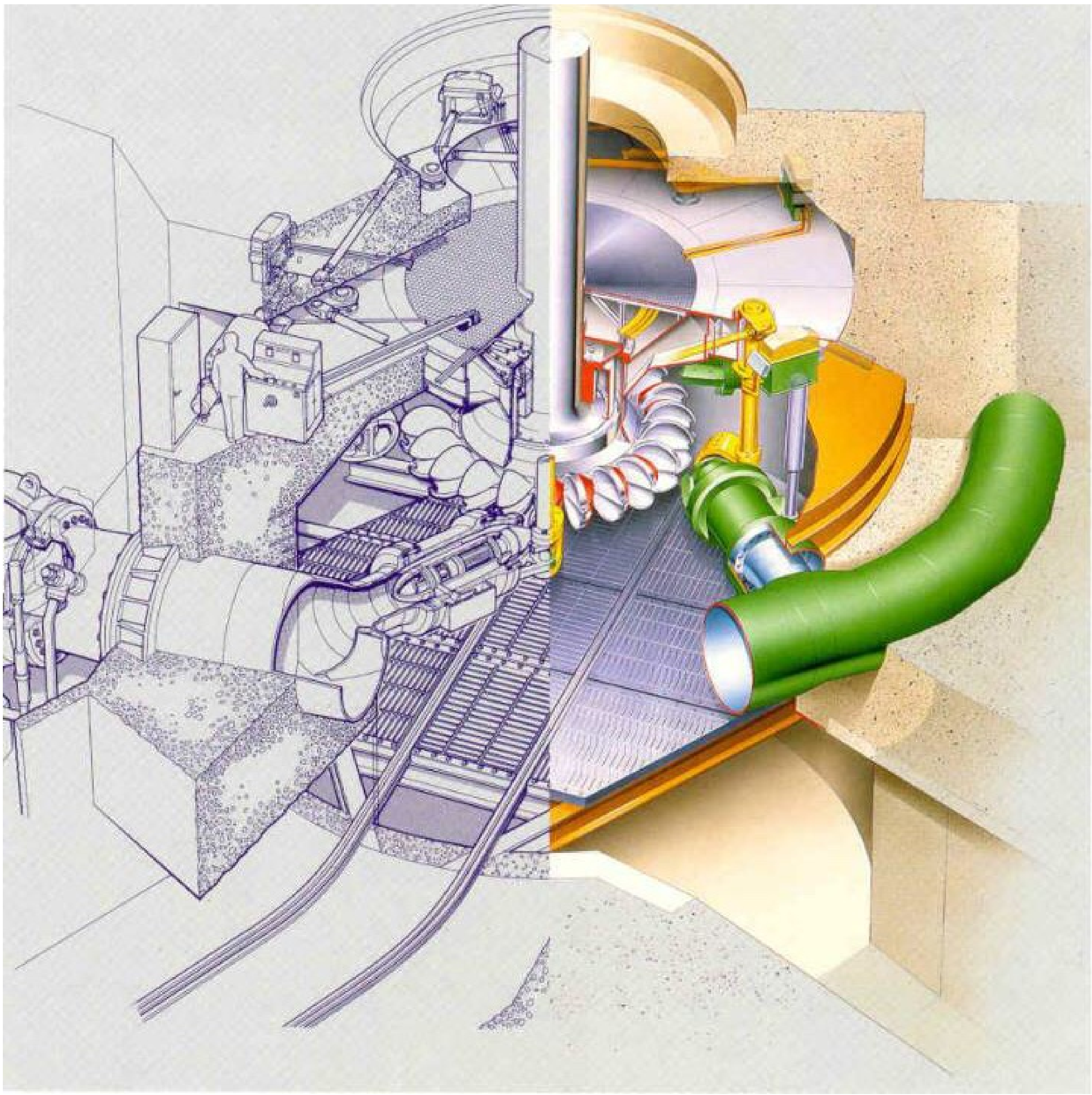


Figure 16.167: **Francis, Kaplan and Pelton wheels**

Pelton turbines can be used as prime movers for other rotating assets also e.g. a mechanical pump.



In the ideal (frictionless) case, all of the hydraulic potential energy $(E_p = mgh)$ is converted into kinetic energy $(E_c = mv^2/2)$.

Equating these two relations and solving for the initial jet velocity (v_i) indicates that the theoretical (maximum) jet velocity is:

$$v_i = \sqrt{2gh} \qquad (16.4.148)$$

For simplicity, let assume that all of the velocity vectors are parallel to each other. Defining the velocity of the wheel runner as u, then as the jet approaches the runner, the initial jet velocity relative to the runner is:

$$(v_i - u) \qquad (16.4.149)$$

where for reminder, the initial velocity of jet is v_i.

Assuming that the jet velocity is higher than the runner velocity, if the water is not to become backed-up in runner, then due to conservation of mass, the mass entering the runner must equal the mass leaving the runner. The fluid is assumed to be incompressible (an accurate assumption for most liquids). Also it is assumed that the cross-sectional area of the jet is constant. The jet speed remains constant relative to the runner. So as the jet recedes from the runner, the jet velocity relative to the runner is:

$$-(v_i - u) = -v_i + u \qquad (16.4.150)$$

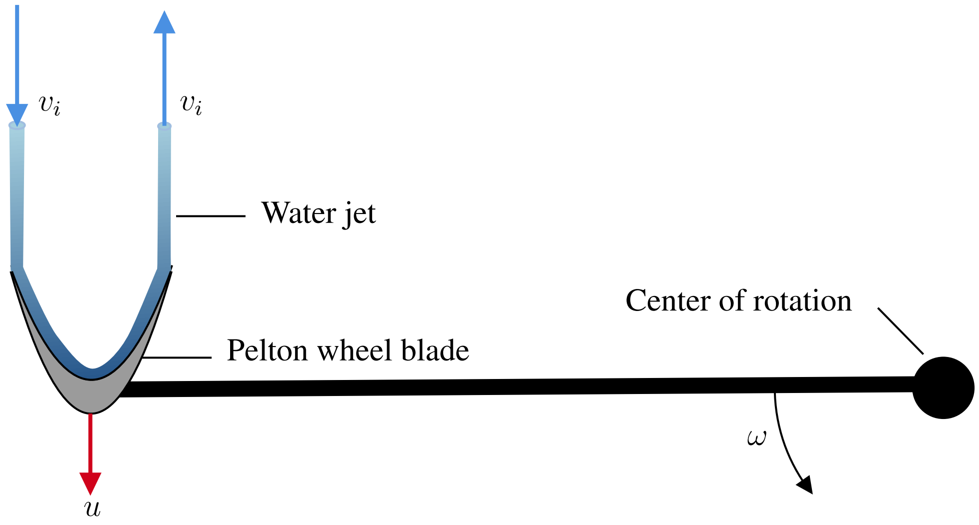


Figure 16.168: **Illustration of a Pelton wheel hydrorotor (source: [285])**

Notice that this assume a frictionless surface on the blade and therefore the output speed of water will remain v_i (.....). That means all our calculations we will have below are an upper limit in comparison to the real situation!!!

The reader must keep in mind what we are willing to calculate here! First we know that we can calculate the power of the water jet flow and in a practical situation we wish to extract all its power. But at a given time, the Pelton wheel will turn so fast that the incoming water jet will almost not accelerate the Pelton wheel but just maintain its optimal radial velocity. And it's the ratio of these both power that we are looking for (the power of the incoming water jet flow and that of the Pelton wheel in this optimal configuration)!

And that's on purpose as the reader must keep in mind that at the final stage we are looking to the efficiency that will be a ratio between the input power of water flow and the ideal maximum power of the torque of the Pelton wheel!

In the standard reference frame (relative to the earth), the final velocity in this ideal case is then:

$$v_f=(-v_i+u)+u=-v_i+2u \tag{16.4.151}$$

The ideal runner speed will cause all of the kinetic energy in the jet to be transferred to the wheel. In this case the final jet velocity must be zero. If:

$$-v_i+2u=0 \tag{16.4.152}$$

then the optimal runner speed will be:

$$u=\frac{v_i}{2} \tag{16.4.153}$$

or half the initial jet velocity!

By Newton's second and third laws, the force F imposed by the jet on the runner is equal but opposite to the rate of momentum change of the fluid, so:

$$F=-m\frac{v_f-v_i}{t}=-\rho\dot{m}\left[(-v_i+2u)-v_i\right]=-\rho\dot{m}\left(-2v_i+2u\right)=2\rho\dot{m}\left(v_i-u\right) \tag{16.4.154}$$

where ρ is the density, and $\dot{m}$ is the volume rate of flow of fluid. If D is the wheel diameter, the torque on the runner is:

$$M=F\frac{D}{2}=\rho\dot{m}D\left(v_i-u\right) \tag{16.4.155}$$

The torque is maximal when the runner is stopped (i.e. when $u=0$, $M=\rho\dot{m}DV_i$). When the speed of the runner is equal to the initial jet velocity, the torque is zero (i.e. when $u=v_i$, then $M=0$)... in the ideal case!

The power (see section Classical Mechanics page 3380):

$$P=Fu=M\omega \tag{16.4.156}$$

where ω is the angular velocity of the wheel. Substituting for F, we have:

$$P=2\rho\dot{m}\left(v_i-u\right)u \tag{16.4.157}$$

To find the runner speed at maximum power, we take the derivative of P with respect to u and set it equal to zero:

$$\frac{\partial P}{\partial u}=2\rho\dot{m}\left(v_i-2u\right)=0 \tag{16.4.158}$$

The maximum therefore power occurs when:

$$u=\frac{v_i}{2} \tag{16.4.159}$$

Therefore:

$$P_{\text{max}} = \rho \dot{m} \frac{v_i^2}{2} \tag{16.4.160}$$

Substituting the initial jet power $v_i = \sqrt{2gh}$, this simplifies to:

$$\boxed{P_{\text{max}} = \rho g h \dot{m}} \tag{16.4.161}$$

This quantity exactly equals the kinetic power of the jet, so in this ideal case, the efficiency is 100%, since all the energy in the jet is converted to shaft output.

A wheel power divided by the initial jet power P_i gives the turbine ideal efficiency:

$$\boxed{\eta_{\text{ideal}} = \frac{P}{P_i} = \frac{2\rho\dot{m}\left(v_i - u\right)u}{\frac{\rho\dot{m}v_i^2}{2}} = \frac{4u\left(v_i - u\right)}{v_i^2} > \eta_{\text{real}}} \tag{16.4.162}$$

It is zero for $u = 0$ and for... $u = v_i$ (that latter may be a little bit counterintuitive) and it is obviously equal to one for $u = v_i/2$ under our previous assumptions! As the equations indicate, when a real Pelton wheel is working close to maximum efficiency, the fluid flows off the wheel with very little residual velocity. In theory, the energy efficiency varies only with the efficiency of the nozzle and wheel, and does not vary with hydraulic head.

From a mechanical efficiency standpoint, considering friction and other losses, most water rotors operate at 90% or higher efficiency. In contrast, the maximum efficiency at which a wind rotor operates is 50% (see section Continuum Mechanics page 3862). The reason for such a large discrepancy in efficiency is the mechanics of energy transfer. With water, the direction of flow of almost all the particles is the same as the overall flow direction and, therefore, it is able to deliver most of the energy to the rotor. If air were used in a Pelton wheel, a significant amount of the energy would be used up to increase the randomness in velocity of air particles as opposed to imparting energy to the rotor.

16.4.7 Tuned Mass Damper

The "tuned mass damper", also known as "seismic damper", is a harmonic dumper that is of main importance in the construction of skyscrapper especially in areas with strong winds and earthquakes. The mathematical details are already provided in the section of Music Mathematics at page 8224.

This device is mounted in structures to reduce mechanical vibrations, consisting of a mass mounted on one or more damped springs. Its oscillation frequency is tuned to be similar to the resonant frequency of the object it is mounted to, and reduces the object's maximum amplitude while weighing much less than it.

TMDs can prevent discomfort, damage, or outright structural failure. They are frequently used in power transmission lines, automobiles, buildings and bridges.

Tuned mass dampers stabilize against violent motion caused by harmonic vibration. They use a comparatively lightweight component to reduce the vibration of a system so that its worst-case vibrations are less intense. Roughly speaking, practical systems are tuned to either move the main mode away from a troubling excitation frequency, or to add damping to a resonance that is difficult or expensive to damp directly. An example of the latter is a crankshaft torsional damper. Mass dampers are frequently implemented with a frictional or hydraulic component that turns mechanical kinetic energy into heat, like an automotive shock absorber.

One of the most impressive mass damper in the early 121st century (holocene calendar) is the one at the top of the Taipei 101 tower in Taiwin (508.2 [m] tall, build in 12005 according to holocene calendar):

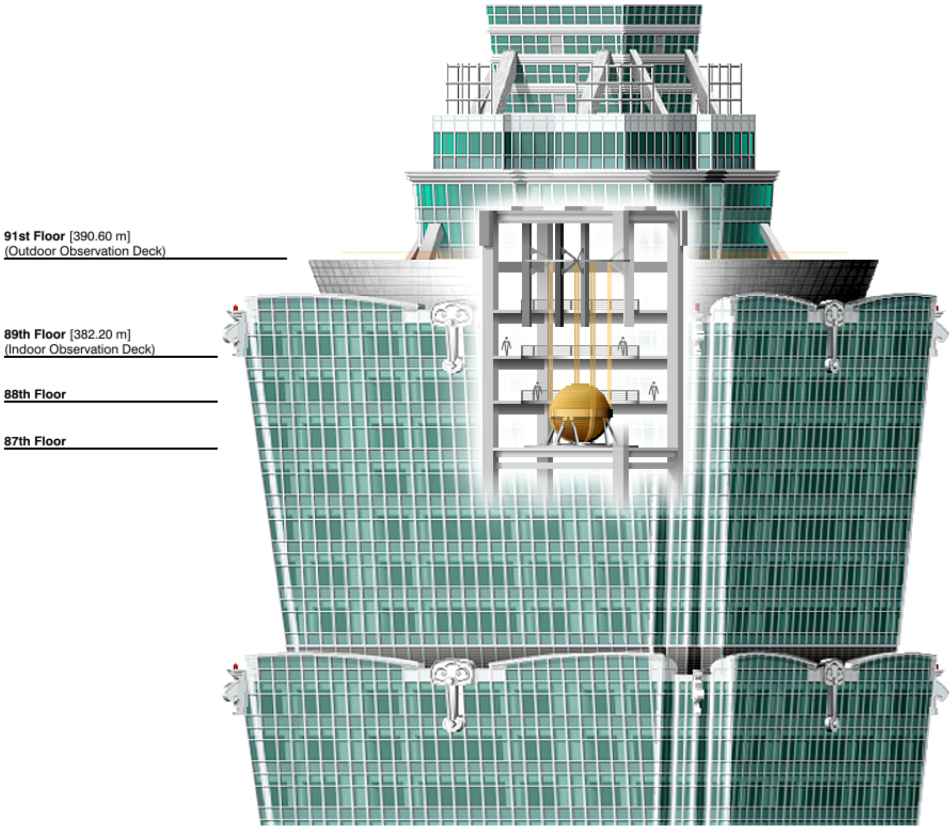

91st Floor [390.60 m]
(Outdoor Observation Deck)
89th Floor [382.20 m]
(Indoor Observation Deck)
88th Floor
87th Floor

Figure 16.169: **Schematic view of Taipei 101 tower tuned mass damper (source: Wikipedia)**

Or a real photo of the dumper:

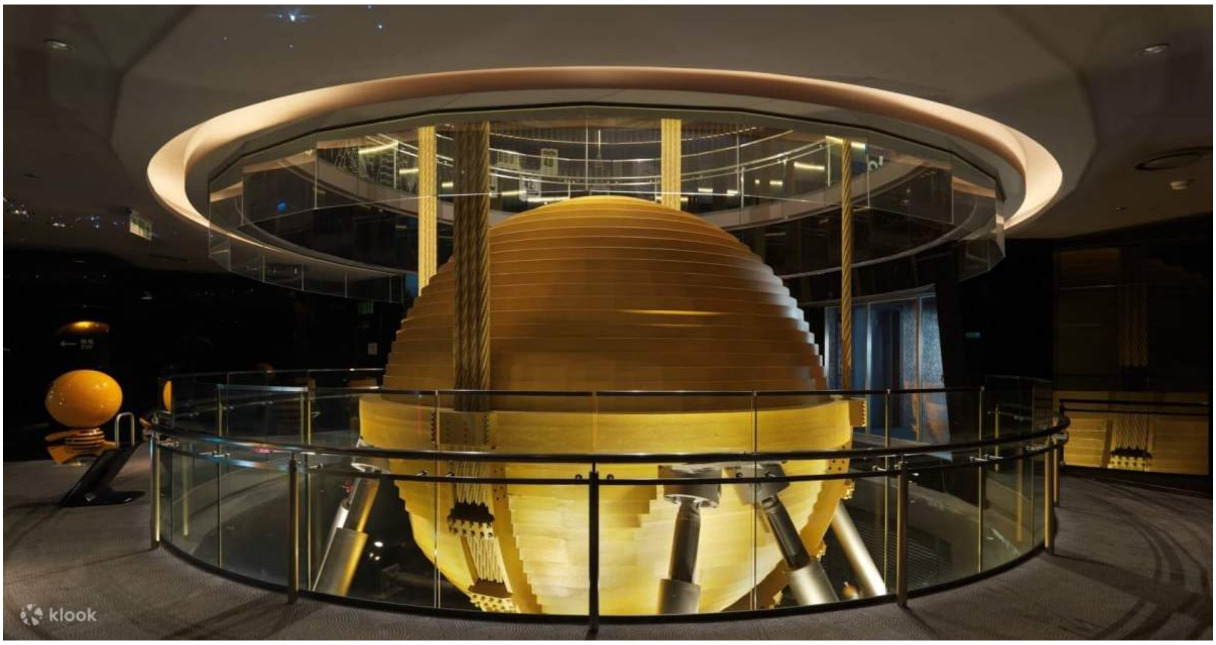


Figure 16.170: **Taipei 101 tower tuned mass damper (source: Klook)**

This damper measures 5.5 [m], its weight is 660 tons (the building itself is estimated at $700,000$

tons). It was made by welding together 41 layers of steel[17] boards, each 12.5 centimeters thick

The reader can see a video of the dumper acting during a 6.8 Richter scale earthquake (in September 18, 12002 according to holocene calendar) in this video:

https://www.youtube.com/watch?v=Tkz6b7Q3dRk

~70 %

★★★☆☆

25 votes, 54.40%

---

[17] Density of steel varies from $7,750\,[\mathrm{kg}\cdot\mathrm{m}^{-3}]$ to $8,050\,[\mathrm{kg}\cdot\mathrm{m}^{-3}]$. But, steel is not the densest alloy or metal. Furthermore, the density of aluminum, iron, and gold is more than steel.

16.5. Aerospace Engineering

In this section we will see some useful and simple mathematical practical cases that we studied in the section of Analytic Geometry, Classical Mechanics and Continuum Mechanics and even Astronomy.

Aerospace engineering is the primary field of engineering concerned with the development of aircraft and spacecraft. It is divided into two major and overlapping branches: "aeronautical engineering" and "astronautical engineering". The Aerospace Engineering (astrodynamics) which is the scientific discipline that brings together aerospace engineering techniques (travel in the atmosphere, using airplanes or helicopter, for example) and those of astronautics (space travel, that is to say outside the atmosphere and interplanetary journeys, using shuttles space and rockets).

The reader will certainly notice that the examples below are only the examples that we often find in textbooks as exercises.

We will not here come back on the theory of conical (very important for the orbiting satellites) seen in the section Analytic Geometry (see page 2764), neither the gyroscope theory (see page 3324) very useful in guiding / force the axis of rotation (spin) of satellites and already seen in the section of Classical Mechanics (by cons we can not use the theory of ballistic that can be found in the same section as we had assumed there the constant initial speed...) or the theory of Lagrangian points (see page 5498) seen in the section of Astronomy sometimes useful to put satellites into interesting orbits for deep space observation. Also hypothetically we will consider bodies in a non-relativistic motion which is so far the most frequent case... at least actually...

In this section we will neglect most of times (but not always!) things like friction, vibration, the cases with more than two bodies (stars, planets ...) and many other factors that require computer simulation. We will also do not see the tricks specific to space engineering (like the fact that some satellites have weights attached by adjustable cables to increase or decrease their gyro momentum).

Remark

Some details may seem insignificant to the reader but the latter must be aware that the price per kilo of rocket launch is between $20,000$ USD and $30,000$ UDS on this beginning of the 121st century (holocene calendar) and this excluding insurance... (or $1,400$ USD per kilo for SpaceX Falcon Heavy) So everything that can be optimized without increasing the risk has to be optimized!

16.5.1 Airfoil Lift

A fluid flowing past the surface of a body exerts a force on it (and not magic...). The "lift" is the component of this force that is perpendicular to the oncoming flow direction. It contrasts with the "drag force", which is the component of the surface force parallel to the flow direction. If the fluid is air, the force is named an "aerodynamic force". In water, it is named a "hydrodynamic force".

Because lift is a force, it is a vector quantity, having both a magnitude and a direction associated with it. Lift acts through the center of pressure of the object and is directed perpendicular to the flow direction. There are several factors which affect the magnitude of lift.

There are many explanations for the generation of lift found in encyclopaedias, in basic physics textbooks and on websites. Unfortunately, many of the explanations are misleading or partially incorrect (because oversimplified) and does not quantify the value of the corresponding lift. Therefore it is naturally because of this lack of scientific approach that the generation of lift has become a source of great controversy and a topic for heated arguments (this is why NASA has a special sub-site dedicated to this subject that has significantly inspired us for the below pages).

Lift occurs when a moving flow of gas is turned by a solid object. The flow is turned in one direction, and the lift is generated in the opposite direction, according to Newton's Third Law of action and reaction. Because air is a gas and the molecules are free to move around, any solid surface can deflect a flow. For an aircraft wing, both the upper and lower surfaces contribute to the flow turning. Neglecting the upper surface's part in turning the flow leads to an incorrect theory of lift.

The proponents of the arguments usually fall into two main camps:

1. Those who support the "Bernoulli" position that lift is generated by a pressure difference across the wing
2. Those who support the "Newton" position that lift is the reaction force on a body caused by deflecting a flow of gas

Notice that we place the names in quotation marks because neither Newton nor Bernoulli ever attempted to explain the aerodynamic lift of an object.

Lift is generated by the difference in velocity between the solid object and the fluid. There must be motion between the object and the fluid: no motion, no lift.... It makes no difference whether the object moves through a static fluid, or the fluid moves past a static solid object.

Which camp is correct? How is lift generated? In fact there is no right or wrong camp!!! All different effects are superposed (with others effects not listed above!) and have more or left amplitude depending on the configuration of the wing (Angle Of Attack (AOA)). So a third camp, the one of wisdom, would answer that's complicates as it is most of the times a combination of the following effects (for which the contributions depend on the design of the airplane):

- Newton's lift effect (skipping stone argument)

- Bernoulli's lift effect (equal time argument)
- Euler's lift effect (streamwise acceleration argument)
- Coandă lift effect
- Kutta-Joukowki lift effect

When a gas flows over an object, or when an object moves through a gas, the molecules of the gas are free to move about the object; they are not closely bound to one another as in a solid. Because the molecules move, there is a velocity associated with the gas. Within the gas, the velocity can have very different values at different places near the object. Bernoulli's equation, which was named for Daniel Bernoulli, relates the pressure in a gas to the local velocity; so as the velocity changes around the object, the pressure changes as well. Adding up (integrating) the pressure variation times the area around the entire body determines the aerodynamic force on the body. The lift is the component of the aerodynamic force which is perpendicular to the original flow direction of the gas. The drag is the component of the aerodynamic force which is parallel to the original flow direction of the gas. Now adding up the velocity variation around the object instead of the pressure variation also determines the aerodynamic force. The integrated velocity variation around the object produces a net turning of the gas flow. From Newton's third law of motion, a turning action of the flow will result in a re-action (aerodynamic force) on the object. So both "Bernoulli" and "Newton" are correct. Integrating the effects of either the pressure or the velocity determines the aerodynamic force on an object. We can use equations developed by each of them to determine the magnitude and direction of the aerodynamic force.

What is the argument?

Arguments arise because people mis-apply Bernoulli and Newton's equations and because they over-simplify the description of the problem of aerodynamic lift. The most popular incorrect theory of lift arises from a mis-application of Bernoulli's equation. The theory is known as the "equal transit time" or "longer path" theory which states that wings are designed with the upper surface longer than the lower surface, to generate higher velocities on the upper surface because the molecules of gas on the upper surface have to reach the trailing edge at the same time as the molecules on the lower surface. The theory then invokes Bernoulli's equation to explain lower pressure on the upper surface and higher pressure on the lower surface resulting in a lift force. The error in this theory involves the specification of the velocity on the upper surface. In reality, the velocity on the upper surface of a lifting wing is much higher than the velocity which produces an equal transit time. If we know the correct velocity distribution, we can use Bernoulli's equation to get the pressure, then use the pressure to determine the force. But the equal transit velocity is not the correct velocity. Another incorrect theory uses a Venturi flow to try to determine the velocity. But this also gives the wrong answer since a wing section isn't really half a Venturi nozzle. There is also an incorrect theory which uses Newton's third law applied to the bottom surface of a wing. This theory equates aerodynamic lift to a stone skipping across the water. It neglects the physical reality that both the lower and upper surface of a wing contribute to the turning of a flow of gas.

The real details of how an object generates lift are very complex and do not lend themselves to simplification. For a gas, we have to simultaneously conserve the mass, momentum, and energy in the flow. Newton's laws of motion are statements concerning the conservation of momentum. Bernoulli's equation is derived by considering conservation of energy. So both of these equations are satisfied in the generation of lift; both are correct. The conservation of mass introduces a lot of complexity into the analysis and understanding of aerodynamic problems. For example, from the conservation of mass, a change in the velocity of a gas in one direction results in a change in the velocity of the gas in a direction perpendicular to the original change. This is very different from the motion of solids, on which we base most of our experiences in physics. The simultaneous conservation of mass, momentum, and energy of a fluid (while neglecting the effects of air viscosity) are named the Euler Equations after Leonard Euler (see section Continuum Mechanics page 3831). Euler was a student of Johann Bernoulli, Daniel's father, and for a time had worked with Daniel Bernoulli in St. Petersburg. If we include the effects of viscosity, we have the Navier-Stokes equations. To truly understand the details of the generation of lift, one has to have a good working knowledge of the Euler Equations.

Keep in mind that we may explain the same phenomena in many ways, one doesn't exclude the other. So now let us focus on all common possible calculations of lift in a special configuration case to be able to compare their effect numerically!

Remark

An airplane's altitude is maintained by air density. In level constant velocity flight, all the forces on an airplane are balanced. Lift counteracts Gravity. Thrust counteracts drag. Lift is caused by air flowing around the airplane wings. If an airplane increases altitude without increasing thrust, the reduction in air density means less air is flowing over the wings therefore the lift decreases. Less lift with same gravity means the airplane's altitude will decrease until |lift| = |gravity|. Since the air density gradient follows the curvature of the Earth, airplanes will automatically follow the curvature of the Earth because of the laws of Physics. It's named "Phugoid oscillation".

16.5.1.1 Newton's lift argument (skipping stone argument)

About paper planes... even a flat surface can produce a lift with an angle of attack. But the "equal time argument" (Bernoulli's equation) cannot explain such phenomenon.

The Newton's lift is given by the relation (see section Continuum Mechanics page 3853) that we have determined during out study of Stoke's law:

$$\boxed{F=\rho v^2 L^2 C(\mathrm{Re})=C_D v^2=\frac{1}{2}C\rho S v^2} \tag{16.5.1}$$

Let's use the information we've just learned to evaluate the "skipping stone argument":

1. This theory is concerned with only the interaction of the lower surface of the moving object and the air. It assumes that all of the flow turning (and therefore all the lift) is produced by the lower surface. Experiments shows the upper surface also turns the flow. In fact, when one considers the downwash produced by a lifting airfoil, the upper surface contributes more flow turning than the lower surface. This theory does not predict or explain this effect.

2. Because this theory neglects the action/reaction of molecules striking the upper surface, it does not predict the negative lift present in our experiment when the angle of attack is negative. On the top of the airfoil, no vacuum exists. Molecules are still in constant random motion on the upper surface (as well as the lower surface), and these molecules strike the surface and impart momentum to the airfoil as well.

3. The upper airfoil surface doesn't enter into the theory at all. So using this theory, we would expect two airfoils with the same lower surface but very different upper surfaces to give the same lift. We know this doesn't occur in reality. In fact, there are devices on many airliners named "spoilers" which are small plates on the upper surface, between the leading and trailing edges. They are used to change the lift of the wing to maneuver the aircraft by disrupting the flow over the upper surface. This theory does not predict or explain this effect.

4. If we make lift predictions based on this theory, using a knowledge of air density and the number of molecules in a given volume of air, the predictions are totally inaccurate when compared to actual measurements. The chief problem with the theory is that it neglects the physical properties of the fluid. Lift is created by turning a moving fluid, and all parts of the solid object can deflect the fluid.

But.... this theory is not totally inaccurate. In certain flight regimes, where the velocity is very high and the density is very low, few molecules can strike the upper airfoil surface and the Newtonian theory gives very accurate predictions. These are the conditions which occur on the Space Shuttle during the early phases of its re-entry into the Earth's atmosphere at altitudes above about 80 [km] and at velocities above 13 Mach. For these flight conditions, the theory gives a good prediction. However, for most normal flight conditions, like those on an airliner (10 [km], $800\left[\text{km}\cdot\text{h}^{-1}\right]$), this argument does not give the right answer.

16.5.1.2 Bernoulli's lift argument (equal time argument)

The "equal time argument" states that airfoils are shaped with the upper surface longer than the bottom. The air molecules have farther to travel over the top of the airfoil than along the bottom. In order to meet up at the trailing edge, the molecules going over the top of the wing must travel faster than the molecules moving under the wing. Because the upper flow is faster, then, from Bernoulli's generalized equation (see section Continuum Mechanics page 3936):

$$\boxed{E=H-PV=\frac{1}{\gamma-1}PV+mgh+\frac{1}{2}mv^2=c^{te}} \tag{16.5.2}$$

the pressure is lower. The difference in pressure across the airfoil produces the lift.

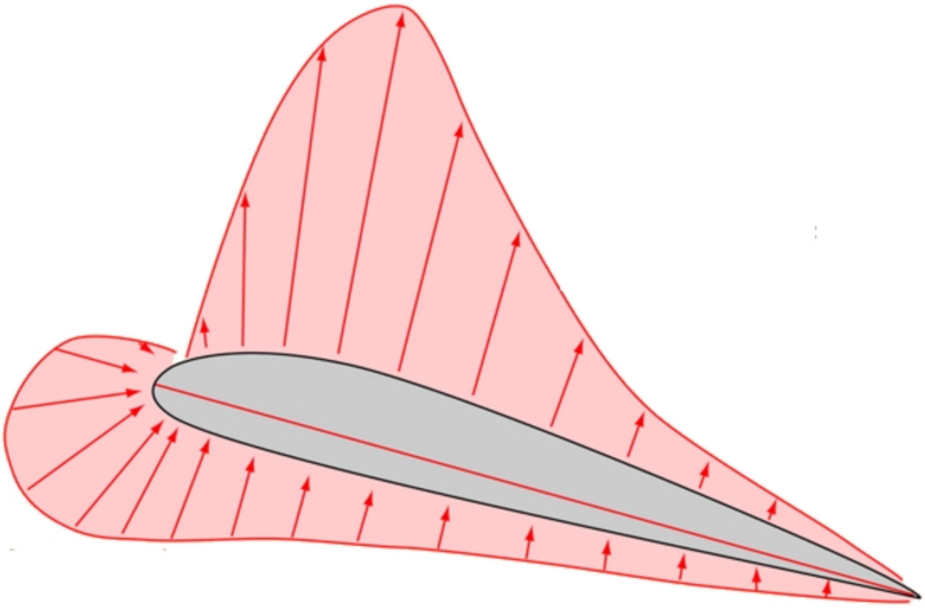


Figure 16.171: **Bernoulli force pressure lift effect illustration (source: HyperPhysics, author: Rod Nave)**

This approach is a beautiful and simple way to explain lift but... in fact, as we have seen in it in the section of Continuum Mechanics, the Bernoulli's equation must be applied on a unique stream line and therefore it is wrong to take for the starting point and endpoint by mixing the both path as they are distinct stream lines!

However it would be a shame that a reader thinks the Bernoulli's lift argument is never applicable to flow around an airfoil! As already mentioned we must be careful with the assumptions of this theorem, but there are still relatively easy to satisfy:

- The flow has to be incompressible. With air, it is not strictly accurate to zero speed. However, as long as one remains in low subsonic, air can reasonably be considered an incompressible fluid (in practice, as long as one remains at a Mach number of less than 0.2 or 0.3). And even the first Bernoulli remains available for all the barotropic compressible fluids along a streamline.
- The fluid must be perfect. The air is a viscous fluid, though the phenomena related to the viscosity remain confined to the boundary layer outside the boundary layer, the fluid can be considered perfect.
- The flow must be irrotational. In fact, it is sufficient that the flow is irrotational globally (this is the general case). The fact that there is locally turbulence does not stop to apply more generally the theorem.

Let's use the information we've just learned to evaluate the various parts of this theory:

- Lifting airfoils are designed to have the upper surface longer than the bottom: This is not always correct. The symmetric airfoil also generates plenty of lift and its upper surface is the same length as the lower surface. Think of a paper airplane. Its airfoil is a flat plate, top and bottom exactly the same length and shape and yet they fly just fine. This part of the theory probably got started because early airfoils were curved and shaped with a longer distance along the top. Such airfoils do produce a lot of lift and flow turning, but it is the turning that's important, not the distance. There are modern, low-drag airfoils which produce lift on which the bottom surface is actually longer than the top. This theory also does not explain how airplanes can fly upside-down which happens often at air shows and in air-to-air combat. The longer surface is then on the bottom!

- Air molecules travel faster over the top to meet molecules moving underneath at the trailing edge: Experiments shows us that the flow over the top of a lifting airfoil does travel faster than the flow beneath the airfoil. But the flow is much faster than the speed required to have the molecules meet up at the trailing edge. Two molecules near each other at the leading edge will not end up next to each other at the trailing edge as shown experiments. This part of the theory attempts to provide us with a value for the velocity over the top of the airfoil based on the non-physical assumption that the molecules meet at the aft end. We can calculate a velocity based on this assumption, and use Bernoulli's equation to compute the pressure, and perform the pressure-area calculation and the answer we get does not agree with the lift that we measure for a given airfoil. The lift predicted by the "equal time argument" theory is much less than the observed lift, because the velocity is too low. The actual velocity over the top of an airfoil is much faster than that predicted by the "equal time argument" theory and particles moving over the top arrive at the trailing edge before particles moving under the airfoil.

- The upper flow is faster and from Bernoulli's equation the pressure is lower. The difference in pressure across the airfoil produces the lift: Experiments show that this part of the theory is correct. In fact, this theory is very appealing because many parts of the theory are correct. In our discussions on pressure-area integration to determine the force on a body immersed in a fluid, we mentioned that if we know the velocity, we can obtain the pressure and determine the force. The problem with the "Equal Transit" theory is that it attempts to provide us with the velocity based on a non-physical assumption as discussed above.

Remark

The misconception that wings must be curved on top and flat on the bottom is commonly associated with the previously-discussed misconception that the air is required to pass above and below the wing in equal amounts of time. In fact, an upside-down wing produces lift by exactly the same principle as a rightside-up wing.

16.5.1.3 Euler's lift argument

Let us recall the Euler equation of the 1st form seen in the section of Continuum Mechanics at page 3828 (still for an incompressible fluid):

$$\rho\left[\partial_t\vec{v}+(\vec{v}\circ\vec{\nabla})\vec{v}\right]=\rho\vec{f}-\vec{\nabla}p \tag{16.5.3}$$

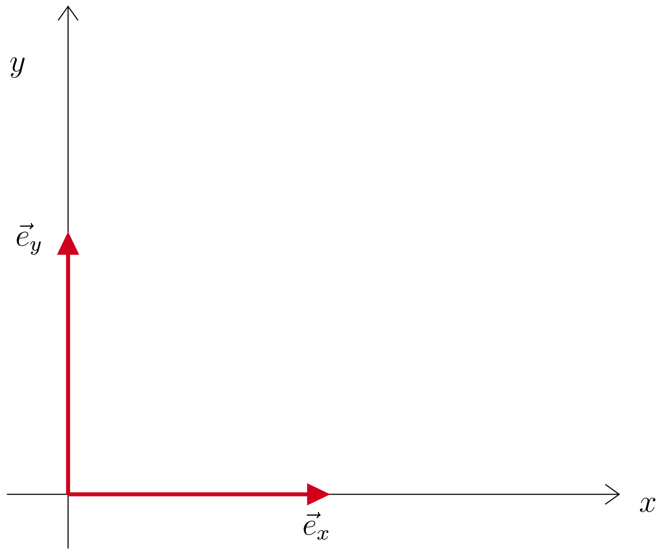
If we neglect the gravitational field:

$$\rho\left[\partial_t\vec{v}+(\vec{v}\circ\vec{\nabla})\vec{v}\right]=-\vec{\nabla}p \tag{16.5.4}$$

Distributing:

$$\rho\partial_t\vec{v}+\rho(\vec{v}\circ\vec{\nabla})\vec{v}=-\vec{\nabla}p \qquad (16.5.5)$$

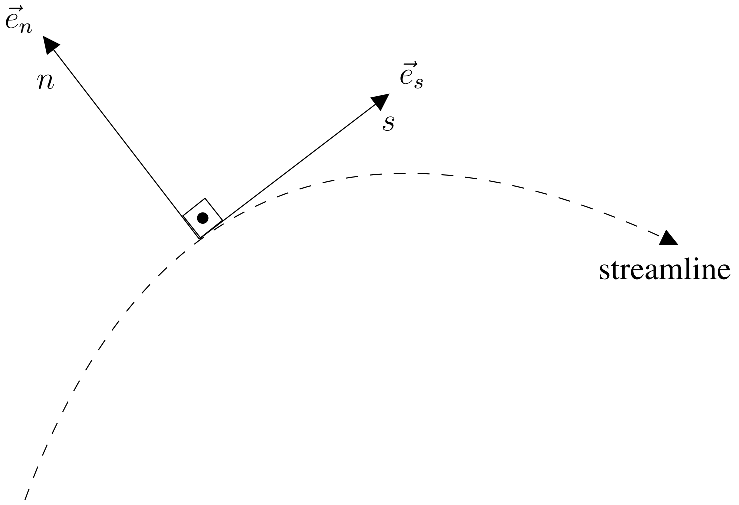
Let us now recall that in cartesian coordinates:



We define the gradient operator in 2D as:

$$\vec{\nabla}=\vec{e}_x\frac{\partial}{\partial x}+\vec{e}_y\frac{\partial}{\partial y} \qquad (16.5.6)$$

And let us introduce another approach, named "streamline coordinates", based on using differential geometry considering the following figure:



We define obviously the gradient identically:

$$\vec{\nabla}=\vec{e}_s\frac{\partial}{\partial s}+\vec{e}_n\frac{\partial}{\partial n} \qquad (16.5.7)$$

The advantage of this approach is that the velocity vector is obviously:

$$\vec{v}=v\vec{e}_s \tag{16.5.8}$$

The disadvantage however is that the unit vectors $\{\vec{e}_s,\vec{e}_n\}$ (Frenet vectors) change along the streamline!

Let us now recall the first Frenet formula (see section Differential Geometry page 2865):

$$\frac{\vec{N}}{R}=\frac{\mathrm{d}\vec{T}}{\mathrm{d}s} \tag{16.5.9}$$

Hence in our case above, this will be written (the "$-$" sign is here to make the perpendicular vector to be outside of the osculating circle as in the figure above):

$$-\frac{\vec{e}_n}{R}=\frac{\mathrm{d}\vec{e}_s}{\mathrm{d}s} \tag{16.5.10}$$

where are we know, R is the radius of curvature at that point on the streamline.

Now let us go back to the Euler equation of the 1st form:

$$\rho\partial_t\vec{v}+\rho(\vec{v}\circ\vec{\nabla})\vec{v}=-\vec{\nabla}p \tag{16.5.11}$$

This equation is steady flow state is simplified as:

$$\rho(\vec{v}\circ\vec{\nabla})\vec{v}=-\vec{\nabla}p \tag{16.5.12}$$

That we will rearrange a bit:

$$\vec{v}\circ\vec{\nabla}\vec{v}=-\frac{1}{\rho}\vec{\nabla}p \tag{16.5.13}$$

Applying the intrinsic definition of $\vec{\nabla}$ and noting that:

$$\vec{e}_s\circ\vec{e}_s=1 \qquad \vec{e}_s\circ\vec{e}_n=0 \tag{16.5.14}$$

and remembering that:

$$\vec{v}=v\vec{e}_s \tag{16.5.15}$$

Then we have successively:

$$\begin{aligned}
\vec{v}\circ\vec{\nabla}\vec{v}&=-\frac{1}{\rho}\vec{\nabla}p && (16.5.16)\\
(v\vec{e}_s)\circ\vec{\nabla}(v\vec{e}_s)&=-\frac{1}{\rho}\vec{\nabla}p && (16.5.17)\\
v\vec{e}_s\left(\vec{e}_s\frac{\partial(v\vec{e}_s)}{\partial s}+\vec{e}_n\frac{\partial(v\vec{e}_s)}{\partial n}\right)&=-\frac{1}{\rho}\left(\vec{e}_s\frac{\partial p}{\partial s}+\vec{e}_n\frac{\partial p}{\partial n}\right) && (16.5.18)\\
v\frac{\partial(v\vec{e}_s)}{\partial s}&=-\frac{1}{\rho}\left(\vec{e}_s\frac{\partial p}{\partial s}+\vec{e}_n\frac{\partial p}{\partial n}\right) && (16.5.19)\\
v\left(\vec{e}_s\frac{\partial v}{\partial s}+v\frac{\partial\vec{e}_s}{\partial s}\right)&=-\frac{1}{\rho}\left(\vec{e}_s\frac{\partial p}{\partial s}+\vec{e}_n\frac{\partial p}{\partial n}\right) && (16.5.20)
\end{aligned}$$

$$\vec{e}_s v\frac{\partial v}{\partial s} - v^2\frac{\vec{e}_n}{R} = -\frac{1}{\rho}\left(\vec{e}_s\frac{\partial p}{\partial s} + \vec{e}_n\frac{\partial p}{\partial n}\right) \tag{16.5.21}$$

Resolving separately in the $\vec{e}_s$ and $\vec{e}_n$ gives (notice that the dimensions are those of an acceleration $[\mathrm{m}^2 \cdot \mathrm{s}^{-1}]$!):

$$v\frac{\partial v}{\partial s} = -\frac{1}{\rho}\frac{\partial p}{\partial s} \tag{16.5.22}$$

$$\frac{v^2}{R} = \frac{1}{\rho}\frac{\partial p}{\partial n} \tag{16.5.23}$$

Where the first relation is say to give the "streamwise acceleration", denoted a_s, and the second relation is say to give the "normal acceleration", denoted a_n.

We see that the second relation describes how streamlines are bent by pressure gradients. So the important stuff to focus on is:

$$\boxed{\frac{v^2}{R} = \frac{1}{\rho}\frac{\partial p}{\partial n}} \tag{16.5.24}$$

that states that when there is flow and the streamlines curve, the pressure p increases in the $\vec{n}$ direction, that is, as one moves away from the local center of curvature.

Remark

Notice that the first relation:

$$v\frac{\partial v}{\partial s} = -\frac{1}{\rho}\frac{\partial p}{\partial s} \tag{16.5.25}$$

can also be rewritten:

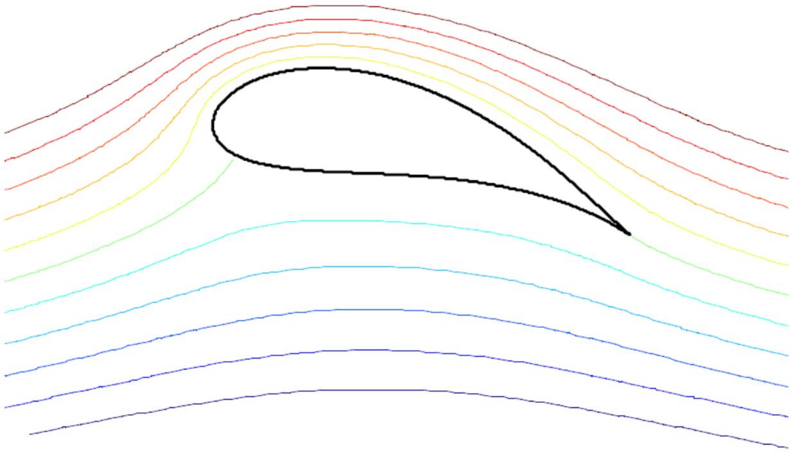
$$\frac{\partial}{\partial s}\left(\frac{1}{2}v^2\right) = -\frac{1}{\rho}\frac{\partial p}{\partial s} \tag{16.5.26}$$

and rearranging we get:

$$\frac{\partial}{\partial s}\left(\frac{1}{2}\rho v^2 + p\right) = 0 \tag{16.5.27}$$

So we fall back on the Bernoulli's theorem that, for recall, states that the sum in the parenthesis remains invariant along a streamline.

Now to illustrate the physical meaning, consider the following illustration:



The difference between the pressure at infinity and at the top of the wing is given by the integration of:

$$\frac{v^2}{R} = \frac{1}{\rho}\frac{\partial p}{\partial n} \tag{16.5.28}$$

from which we see that:

$$p_\infty - p_{\text{top}} = \int_{\text{top}}^{+\infty} \frac{\rho v^2}{R} \mathrm{d}n > 0 \tag{16.5.29}$$

Thus $p_{\text{top}} < p_\infty$ and according to Bernoulli's equation:

$$\frac{\partial}{\partial s}\left(\frac{1}{2}\rho v^2 + p\right) = 0 \tag{16.5.30}$$

it then follows that $v_{\text{top}} > v_\infty$.

16.5.1.4 Coandă lift argument

In its original sense, the Coandă effect refers to the tendency of a fluid jet to stay attached to an adjacent surface that curves away from the flow, and the resultant entrainment of ambient air into the flow.

More broadly, some consider the effect to include the tendency of any fluid boundary layer to adhere to a curved surface, not just the boundary layer accompanying a fluid jet. It is in this broader sense that the Coandă effect is used by some to explain why the air flow remains attached to the top side of an airfoil.

16.5.1.5 Kutta-Joukowski lift argument

Many discussion of airfoil lift invoke a vortex around the moving airfoil and it well known and experimentally proven that there is a lift around a spinning cylinder. This is named the "Kutta-Joukowski theorem" (Kutta–Joukowski theorem relates lift to circulation much like the Magnus effect relates side force (named Magnus force) to rotation).

The two early aerodynamicists, Kutta in Germany and Joukowski in Russia, worked to quantify the lift achieved by an airflow over a spinning cylinder. The lift relation is given by:

$$\boxed{L = -\rho U_\infty \Gamma b} \tag{16.5.31}$$

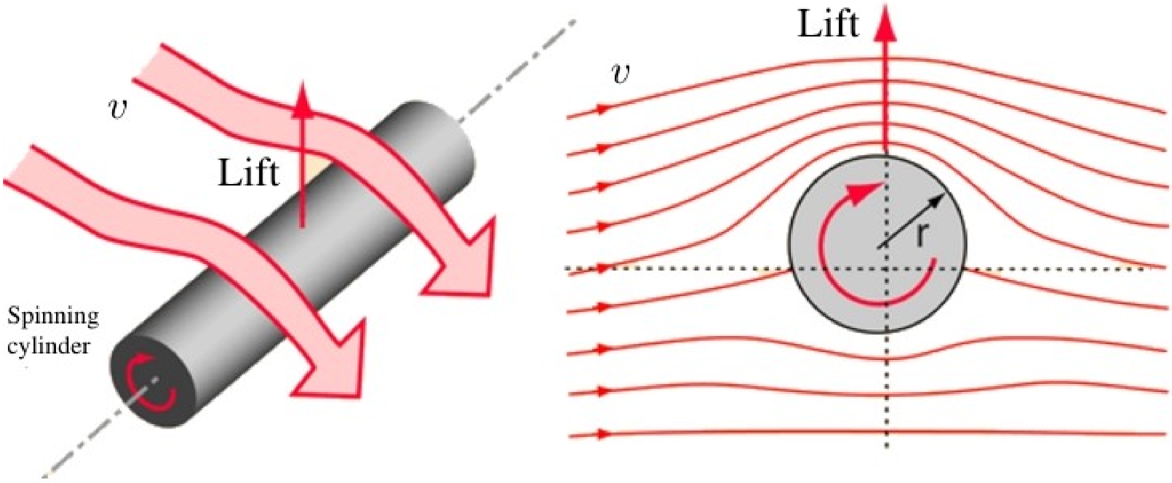


Figure 16.172: **Kutta-Joukowski lift effect illustration (source: HyperPhysics, author: Rod Nave)**

Like all aerodynamic lift, this seems a bit mysterious, but it can be looked at in terms of a redirection of the air motion. If the cylinder traps some air in a boundary layer at the cylinder surface and carries it around with it, shedding it downward, then it has given some of the air a downward momentum. That can act to give the cylinder an upward momentum in accordance with the principle of conservation of momentum. Another approach is to say that you have exerted a downward component of force on the air and by Newton's 3rd law there must be an upward force on the cylinder. Yet another approach is to say that the top of the cylinder is assisting the airstream, speeding up the flow on the top of the cylinder. Then by the Bernoulli equation, the pressure on the top of the cylinder is diminished, giving an effective lift.

We have seen that several physical principles are involved in producing lift. Each of the following statements is correct as far as it goes:

- The wing produces lift "because" it is flying at an angle of attack.
- The wing produces lift "because" of circulation.
- The wing produces lift "because" of Bernoulli's principle.

- The wing produces lift "because" of Newton’s law of action and reaction.

To finish depending on donations

16.5.2 Cosmic speeds

Placing takeoff of a rocket near the equator will be of significant assistance since the tangential velocity (horizontal) of the rotation of the Earth is given by (applying the trivial circular kinematics studied in the section of Classical Mechanics):

$$v=\omega R\cong 465\,[\text{m}\cdot\text{s}^{-1}] \tag{16.5.32}$$

That is to says slightly more than the speed of sound on the ground in standard conditions of temperature and pressure, which represents about 5% of the first cosmic speed we will deal with just a little further below (that's already something to optimize fuel usage!). And obviously, as the Earth rotates on itself from west to east it is better to leave the rocket to the East rather than straight or even worse ... to the West (note in passing that for security reasons of people on the ground it would be even better if the rocket can be sent to the East over the ocean...).

Also note that we are seeking also to acquire of course vertical speed to quickly leave the dense layers of the atmosphere, but we also want to communicate to the launcher a horizontal velocity component: that is to say the velocity that allows to put satellites into orbit of the Earth or of something else less or far away (hence another reason to launch to the East).

As about the difference of the gravitational acceleration it is also a factor to be taken into account. Indeed some specialized book provide:

$$\begin{aligned}g_{\text{equator}}&=9.78039\,[\text{m}\cdot\text{s}^{-2}]\\ g_{\text{north pole}}&=9.78039\,[\text{m}\cdot\text{s}^{-2}]\end{aligned} \tag{16.5.33}$$

Thus, a difference of about 0.527% of the weight of a space shuttle and this is not negligible to once again save fuel.

The "first cosmic speed" or "low-level orbit speed" (or just "orbital speed" or "orbital velocity") is the minimum velocity at which a body should be brought to lie in low orbit around the Earth (for small celestial masses otherwise General Relativity must be involved!). It is determined by the relation balance between centrifugal and centripetal force (see section Classical Mechanics page 3388):

$$\frac{v^2}{R}=\frac{GM}{R^2} \tag{16.5.34}$$

from which we deduce trivially:

$$\boxed{v=\sqrt{\frac{GM}{R}}} \tag{16.5.35}$$

Most often written (as $GM=gR^2$ where g is the acceleration due to gravity) involving the altitude h (hence $R+h$ is the radius of orbit):

$$v=\sqrt{\frac{GM}{R+h}}=R\sqrt{\frac{g}{R+h}} \tag{16.5.36}$$

This is obviously also the speed of the famous "Newton's cannonball" problem, i.e. the speed that should have a cannonball launched at a height h horizontally from Earth's surface in vacuum to turn around the Earth without falling back to the ground.

In fact in our present time (...) an orbit escapes to the Earth's atmosphere usually only if it is at an altitude greater than 200 [km]. We then have (speed that approximately corresponds to that of the International Space Station after-several overlapping):

$$v=\sqrt{\frac{GM}{R}}\cong 7,900\,[\text{m}\cdot\text{s}^{-1}]\cong 28,500\,[\text{km}\cdot\text{h}^{-1}] \tag{16.5.37}$$

That is to say an orbital period of:

$$T=\frac{2\pi R}{v}\cong 96\,[\text{min}] \tag{16.5.38}$$

The "second cosmic speed" corresponds to the release velocity of a body leaving Earth. This is the minimum speed beyond which a body can move away permanently from Earth, at least... as we neglect the presence of the Sun, the Moon and of our Galaxy... We have already proven the corresponding relation in the section of Classical Mechanics (see page 3234) so its no use to do it again. Let us just recall that the following relation was obtained:

$$v_e=\sqrt{\frac{2GM}{R}} \tag{16.5.39}$$

What gives for Earth (taking the average radius at the sea level...)

$$v_e=\sqrt{\frac{2GM}{R}}\cong 11.2\cdot 10^3\,[\text{m}\cdot\text{s}^{-1}]\cong 40,000\,[\text{km}\cdot\text{h}^{-1}] \tag{16.5.40}$$

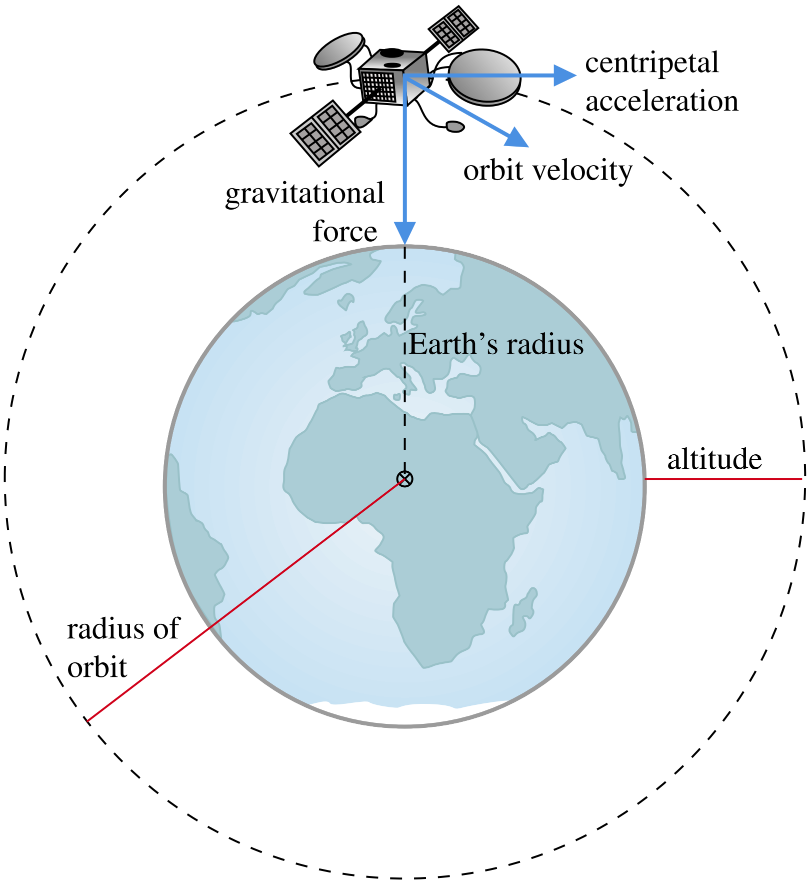


Figure 16.173: Orbital speed illustration (doesn't respect scales...)

Because of the Earth's atmosphere, it is difficult (and not very useful) to bring an object close to its surface at this speed, this speed being too far in the hypersonic regime to be achievable by most existing propulsion systems. In addition, it will cause a destruction of most objects by atmospheric friction or compression. In practice, an object is first placed in circular orbit of the Earth orbit and then accelerated from that altitude as the friction is there almost zero, the thrust has then a very good yield (see further below the study of Hohmann transfer orbit). In addition, the following table provided by NASA shows this is the strategy they use (you may notice that starting from 105 kilometres of altitude the speed gain is very effective):

Time [s]	**Altitude [m]**	**Speed** $[\mathrm{m}\cdot\mathrm{s}^{-1}]$	**Acceleration** $[\mathrm{m}\cdot\mathrm{s}^{-1}]$**]**
0	-8	0	2.45
20	1,244	139	18.62
40	5,377	298	16.37
60	11,617	433	19.40
80	19,872	685	24.50
100	31,412	1,026	24.01
120	44,726	1,279	8.72
140	57,396	1,373	9.70
160	67,893	1,490	10.19
180	77,485	1,634	10.68
200	85,662	1,800	11.17
220	92.481	1,986	11.86
240	98.004	2,191	12.45
260	100,301	2,417	13.23
280	105,321	2,651	13.92
300	107,449	2,915	14.90
320	108,619	3,203	15.97
340	108,942	3,516	17.15
360	108,543	3,860	18.62
380	107,690	4,216	20.29
400	106,539	4,630	22.34
420	105,142	5,092	24.89
440	103,775	5,612	28.03
460	102,807	6,184	29.01
480	102,552	6,760	29.30
500	103,297	7,327	29.01
520	105,069	7,581	0.10

Table 16.11: **Ascensional data of Discovery Shuttle (source: NASA)**

16.5.3 Fundamental Equation of Propulsion (Tsiolkovsky rocket equation)

A space launcher has for mission most of time to place a load into orbit, for that it must operate in the atmosphere and vacuum. The principles used are those of the action and reaction of Newton, and the conservation of momentum (see section Classical Mechanics page 3209 and page 3367): roughly we can say that a rocket accelerates high speed by ejecting gas (the balloon that deflates gives a good idea of the phenomenon). After a simple mechanical study, we can obtain the fundamental equation of propulsion.

The "Tsiolkovsky rocket equation", or "ideal rocket equation", describes the motion of vehicles that follow the basic principle of a rocket: a device that can apply acceleration to itself (a thrust) by expelling part of its mass with high speed and thereby move due to the conservation of momentum. The equation relates the Δv (the maximum change of velocity of the rocket if no other external forces act) with the effective exhaust velocity and the initial and final mass of a rocket (or other reaction engine).

Considering the different mobile and independent parts of the rocket (in fact the main structure that ejected gas separately), we can state thanks to the principle of action and reaction (see section Classical Mechanics page 3209) for a given system, the sum of the external forces are:

$$\sum \vec{F} = m\vec{a} = \frac{\mathrm{d}\vec{p}}{\mathrm{d}t} \tag{16.5.41}$$

Thus, taking the rocket system and gas together, we have:

$$\sum \frac{\mathrm{d}\vec{p}_i}{\mathrm{d}t} = \vec{0} \tag{16.5.42}$$

That is to say:

$$\sum \vec{p}_i = c^{te} \tag{16.5.43}$$

The principle of propulsion is then stated through the conservation of linear momentum that we have just established.

Let us consider a mass m and of speed v_f rocket, the ejection velocity of the gases being v_g. At time t, we have:

$$p = mv_f \tag{16.5.44}$$

at the time $t + \mathrm{d}t$ we have:

$$p = (m - \mathrm{d}m)(v_f + \mathrm{d}v_f) + \mathrm{d}m(v_f + \mathrm{d}v_f - v_g) \tag{16.5.45}$$

but the $\mathrm{d}m$ is a mass loss so we must change its sign otherwise the above relation does not correspond to the interpretation of reality. So:

$$p = (m + \mathrm{d}m)(v_f + \mathrm{d}v_f) - \mathrm{d}m(v_f + \mathrm{d}v_f - v_g) \tag{16.5.46}$$

based on the principle of conservation of linear momentum:

$$mv_f = (m + \mathrm{d}m)(v_f + \mathrm{d}v_f) - \mathrm{d}m(v_f + \mathrm{d}v_f - v_g) \tag{16.5.47}$$

Therefore after simplification:

$$m\mathrm{d}v_f = -\mathrm{d}m v_g \quad (16.5.48)$$

hence:

$$\mathrm{d}v_f = -v_g \frac{\mathrm{d}m}{m} \quad (16.5.49)$$

which by integration (see section Differential and Integral Calculus page 1605) gives the "fundamental equation of propulsion" or the "delta-v" as often said in American literature and for a (non-relativistic ...) rocket outside from gravity field and in vacuum with a constant speed of ejection of gas ...:

$$\boxed{\Delta v_f = v_g \ln\left(\frac{m_i}{m_f}\right)} \quad (16.5.50)$$

The difference between the initial mass m_i and final mass m_f is often named "dead mass".

Then we understand easily why rockets are composed of several staged propulsion elements. This allows them to increase their final speed in getting rid of the mass of tanks that they initially take with them.

We conclude therefore that a launcher accelerates especially since the gas velocity is large, the thrust depends on the amount of gas supplied and speed, and the ratio of initial and final masses must be at maximum to ensure a good propulsion, that is to say that the final system structure must be as negligible as possible (final mass is then minimal).

Let us notice that we also have by extension the following relation between the acceleration of the rocket and the mass flow rate ejected:

$$a_f = \frac{\mathrm{d}v_f}{\mathrm{d}t} = -\frac{v_g}{m}\frac{\mathrm{d}m}{\mathrm{d}t} \quad (16.5.51)$$

However, in a gravitational field (of course it is necessary that the acceleration of the rocket is greater than that of gravity at any time ...) we must add the term which slows down the rocket (the "loss by gravity") and this will give us the expression:

$$\boxed{v(t) = v_g \ln\left(\frac{m_i}{m(t)}\right) - gt} \quad (16.5.52)$$

if we assume the gravity g as constant during the main acceleration phase. The latter relationship is the "Tsiolkovsky formula". Denoting by D_e the mass flow of propellant, we sometimes find this last relation in the following form where time no longer explicitly intervenes:

$$v = v_g \ln\left(\frac{m_i}{m_i - D_e t}\right) - gt = v_g \ln\left(\frac{m_i}{m(t)}\right) - g\frac{m_i - m(t)}{D_e} \quad (16.5.53)$$

as:

$$t = \frac{m_i - m(t)}{D_e} \quad (16.5.54)$$

However, it should also take into account the variation of gravity depending on the distance as we have proved it in the section Astronomy (see page 5281). Then we have:

$$\boxed{v(t)=v_g\ln\left(\frac{m_i}{m(t)}\right)-g\left(\frac{R}{r}\right)^2 t} \tag{16.5.55}$$

So if we assume the ejection speed of the gas as constant and the trajectory path in straight line from the main body of attraction, then we have that more time passes more will the rocket speed increases due at its mass decrease (of course it will at the end be constant) but at the same time less the influence of gravity is high.

The theoretical value of the altitude reached, even if its expression is very simple to determine, is so false in our point of view that it is not useful to present it.

The rockets do not go at the release velocity with only one propulsion stage and even ... they often have for only objective of going only to the low-orbit speed... (first cosmic speed). The rocket is therefore in practice not free from gravity, far from it!

✍**Example:**

Let us calculate the final velocity of the first phase of Ariane 5 launch assuming constant gravity, knowing the ejection gas velocity (assumed as constant speed), the initial total mass of the rocket, the ejected mass and the mass flow (...also assumed constant). Then we have:

$$v\cong 3600\ln\left(\frac{700000}{260000}\right)-9.7\frac{700000-260000}{3900}\cong 2470\,[\text{m}\cdot\text{s}^{-1}] \tag{16.5.56}$$

with:

$$t=\frac{m_i-m(t)}{D_e}=\frac{700000-260000}{3900}\cong 112\,[\text{s}] \tag{16.5.57}$$

While the actual communicated value is between $2,000$ and $2,800\,[\text{m}\cdot\text{s}^{-1}]$ after the cross-checking of the information given by several websites and videos of the launch control center of Ariane (so we are very far from the first cosmic speed!). If we add the variation of gravity with altitude the calculated result would be even closer to $2,800\,[\text{m}\cdot\text{s}^{-1}]$ (for information separation of the first stage boosters seems to be about 80 [km] of altitude and after the cross-checking of several sources about 132 [s] after ignition).

However, this remains an acceptable order of (and similar to the US shuttle takeoffs!!) given that we did take into account of the friction of the air (that for recall was proven in the section of Continuum Mechanics as being proportional to the square of the velocity in the subsonic case... !!!).

Let us now determine the distance reached after a given time t in the case of constant gravitational field approximation. We then have obviously in a first time (we change a little bit the notation to condense the developments):

$$\begin{aligned}v(t)=\dfrac{\mathrm{d}z(t)}{\mathrm{d}t}&=v_g\ln\left(\dfrac{m_0}{m_0-D_et}\right)-gt=v_g\ln\left(\dfrac{1}{\dfrac{m_0-D_et}{m_0}}\right)-gt=v_g\ln\left(\dfrac{1}{1-\dfrac{D_e}{m_0}t}\right)-gt\\&=-v_g\ln\left(1-\dfrac{D_e}{m_0}t\right)-gt\end{aligned} \tag{16.5.58}$$

Therefore:

$$z(t)=-v_g\int\ln\left(1-\dfrac{D_e}{m_0}t\right)\mathrm{d}t-g\int t\mathrm{d}t \tag{16.5.59}$$

We will do the following change of variable:

$$u=1-\dfrac{D_e}{m_0}t \tag{16.5.60}$$

Hence:

$$\mathrm{d}u=1-\dfrac{D_e}{m_0}\mathrm{d}t \tag{16.5.61}$$

Which gives us (see section Differential and Integral Calculus page 1534):

$$\begin{aligned}z(t)&=-v_g\int\dfrac{m_0}{D_e}\ln(u)\mathrm{d}u-g\int t\mathrm{d}t=\dfrac{v_gm_0}{D_e}\int\ln(u)\mathrm{d}u-g\int t\mathrm{d}t\\&=\dfrac{v_gm_0}{D_e}\left(\left(1-\dfrac{D_e}{m_0}t\right)\ln\left(1-\dfrac{D_e}{m_0}t\right)-\left(1-\dfrac{D_e}{m_0}t\right)\right)-\dfrac{1}{2}gt^2+c^{te}\\&=\dfrac{v_gm_0}{D_e}\left(1-\dfrac{D_e}{m_0}t\right)\left[\ln\left(1-\dfrac{D_e}{m_0}t\right)-1\right]-\dfrac{1}{2}gt^2+c^{te}\end{aligned} \tag{16.5.62}$$

Let us determine the integration constant the fact that at time $t=0$, we must have $z(t)$ that is zero. It then comes immediately that:

$$c^{te}=\dfrac{v_gm_0}{D_e} \tag{16.5.63}$$

That is to say finally:

$$\boxed{z(t)=\dfrac{v_gm_0}{D_e}\left(1-\dfrac{D_e}{m_0}t\right)\left[\ln\left(1-\dfrac{D_e}{m_0}t\right)-1\right]-\dfrac{1}{2}gt^2+\dfrac{v_gm_0}{D_e}} \tag{16.5.64}$$

where we can again get rid of the explicit time variable by reusing that:

$$t=\dfrac{m_i-m(t)}{D_e} \tag{16.5.65}$$

Which gives:

$$z(t)=\dfrac{v_gm_0}{D_e}\left(1-\dfrac{m_i-m(t)}{m_0}\right)\left[\ln\left(1-\dfrac{m_i-m(t)}{m_0}\right)-1\right]-\dfrac{1}{2}g\left(\dfrac{m_i-m(t)}{D_e}\right)^2+\dfrac{v_gm_0}{D_e} \tag{16.5.66}$$

If we take the special case $m_i = m_0$ the relation simplifies to:

$$z(t) = \frac{v_g m(t)}{D_e}\left[\ln\left(\frac{m(t)}{m_0}\right) - 1\right] - \frac{1}{2}g\left(\frac{m_0 - m(t)}{D_e}\right)^2 + \frac{v_g m_0}{D_e} \qquad (16.5.67)$$

✍**Example:**

Let us calculate the height of the first launch phase of Ariane 5 assuming constant gravity with the same data as the previous example. This then gives:

$$\begin{aligned} z(t) &\cong \frac{3600\cdot 260000}{3900}\left[\ln\left(\frac{260000}{700000}\right) - 1\right] - \frac{1}{2}9.7\left(\frac{700000-260000}{3900}\right)^2 + \frac{3600\cdot 700000}{3900} \\ &\cong 106725 \cong 100\,[\text{km}] \end{aligned} \qquad (16.5.68)$$

Which according to the cross-check of many website and videos of the Ariane launch control center would not be too wrong because the main propellers (solid stages accelerators) are dropped at an altitude about 70 to 125 [km] (between 132 and 205 [s] after ignition) and even the cap that protects the satellites and the propulsion stage (main cryogenic stage) are sometimes dropped just a few seconds after (it always makes a few tons less!).

So as we have seen with the calculations of the previous example, there are still about 100 [km] to reach to go to the lower orbit and its corresponding cosmic speed using the storable propellant (EPS).

Let us notice that if we take off the rocket with an acceleration equal to that of gravity (assuming the gravity constant), we would obtain a height equal to:

$$z(t) = \frac{1}{2}gt^2 = \frac{1}{2}9.7(112)^2 \cong 60838 \cong 60\,[\text{km}] \qquad (16.5.69)$$

So we can approximatively conclude that the propulsion acceleration of Ariane 5 is significantly higher than that of gravity.

We can also calculate the change of the gravity at the altitude of 100 [km] to see if the variation is significant or not in $\%$ of that at the ground level:

$$\frac{\Delta g}{g_0} = \left(\frac{R}{R+z}\right)^2 \cong \left(\frac{6378137}{6378137+100000}\right)^2 \cong 97\% \qquad (16.5.70)$$

So we see that this is 3%, which is quite significant.

Furthermore, here is a diagram supposedly taken from a book of Arianespace showing the horizontal acceleration of the Ariane 5 in multiple of the local gravity on the launch center:

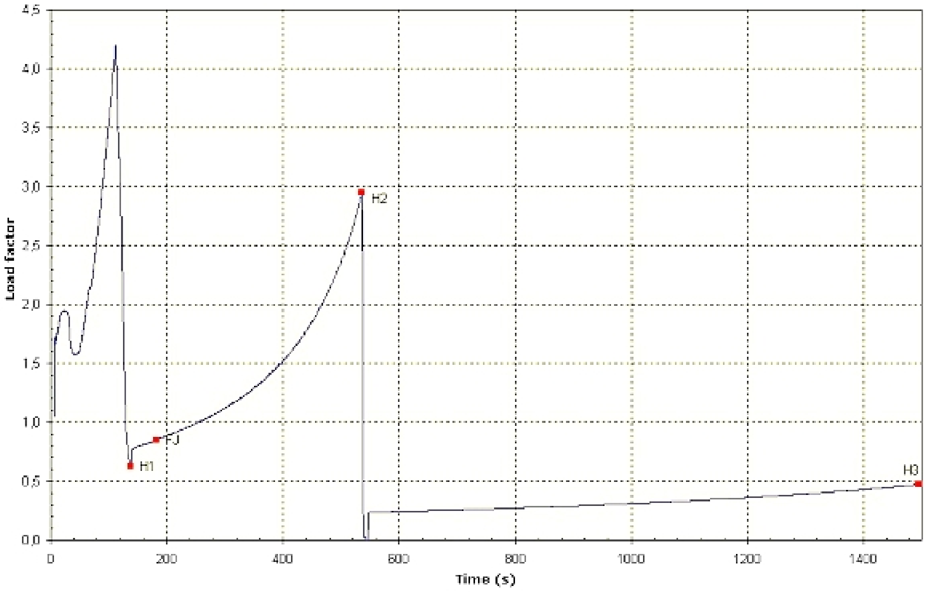


Figure 16.174: **Horizontal acceleration profile of Ariane 5 (source: ?)**

Now suppose that based on the relation:

$$t=\frac{m_0-m(t)}{D_e} \tag{16.5.71}$$

that we know the limited total mass ejected. We will then also have the combustion period that will be limited. Let us denote this by:

$$t_f=\frac{m_0-m_f}{D_e} \tag{16.5.72}$$

Using the relations obtained earlier in the case of a constant gravitational field, we then have obviously:

$$\begin{aligned}
v_f&=v_g\ln\left(\frac{m_0}{m_f}\right)-g\frac{m_0-m_f}{D_e}\\
z_f&=\frac{v_gm_f}{D_e}\left[\ln\left(\frac{m_f}{m_0}\right)-1\right]-\frac{1}{2}g\left(\frac{m_0-m_f}{D_e}\right)^2+\frac{v_gm_0}{D_e}
\end{aligned} \tag{16.5.73}$$

After reaching the final velocity, the speed of the rocket will be supposed given by the classical relation of the rectilinear kinematics (see section Classical Mechanics page 3224):

$$v=v_f-g(t-t_f) \tag{16.5.74}$$

and also the height:

$$z=z_f+v_f(t-t_f)-\frac{1}{2}g(t-t_f)^2 \tag{16.5.75}$$

By injecting the relations obtained above, we first have for the speed:

$$v=v_f-g(t-t_f)=\left[v_g\ln\left(\frac{m_0}{m_f}\right)-g\frac{m_0-m_f}{D_e}\right]-g\left(t-\frac{m_0-m_f}{D_e}\right)=v_g\ln\left(\frac{m_0}{m_f}\right)-gt \tag{16.5.76}$$

and for the height:

$$\begin{aligned}
z&=z_f+v_f(t-t_f)-\frac{1}{2}g(t-t_f)^2\\
&=\left[\frac{v_gm_f}{D_e}\left[\ln\left(\frac{m_f}{m_0}\right)-1\right]-\frac{1}{2}gt_f^2+\frac{v_gm_0}{D_e}\right]+\left[v_g\ln\left(\frac{m_0}{m_f}\right)-gt_f\right](t-t_f)-\frac{1}{2}g(t-t_f)^2\\
&=\frac{v_gm_f}{D_e}\left[\ln\left(\frac{m_f}{m_0}\right)-1\right]-\frac{1}{2}gt_f^2+\frac{v_gm_0}{D_e}+v_g\ln\left(\frac{m_0}{m_f}\right)t-gt_ft-v_gt_f\ln\left(\frac{m_0}{m_f}\right)+gt_f^2\\
&-\frac{1}{2}g(t^2-2tt_f+t_f^2)\\
&=\frac{v_gm_f}{D_e}\left[\ln\left(\frac{m_f}{m_0}\right)-1\right]+\frac{v_gm_0}{D_e}+v_gt\ln\left(\frac{m_0}{m_f}\right)-v_g\left(\frac{m_0-m_f}{D_e}\right)\ln\left(\frac{m_0}{m_f}\right)-\frac{1}{2}gt^2\\
&=\frac{v_gm_f}{D_e}\left[\ln\left(\frac{m_f}{m_0}\right)-1\right]+\frac{v_gm_0}{D_e}+v_gt\ln\left(\frac{m_0}{m_f}\right)+\frac{v_gm_f}{D_e}\ln\left(\frac{m_0}{m_f}\right)-\frac{v_gm_0}{D_e}\ln\left(\frac{m_0}{m_f}\right)-\frac{1}{2}gt^2\\
&=\frac{v_gm_f}{D_e}\left[\ln\left(\frac{m_f}{m_0}\right)+\ln\left(\frac{m_0}{m_f}\right)-1\right]+\frac{v_gm_0}{D_e}+v_gt\ln\left(\frac{m_0}{m_f}\right)-\frac{v_gm_0}{D_e}\ln\left(\frac{m_0}{m_f}\right)-\frac{1}{2}gt^2\\
&=-\frac{v_gm_f}{D_e}+\frac{v_gm_0}{D_e}+v_gt\ln\left(\frac{m_0}{m_f}\right)-\frac{v_gm_0}{D_e}\ln\left(\frac{m_0}{m_f}\right)-\frac{1}{2}gt^2\\
&=v_g\frac{m_0-m_f}{D_e}+v_gt\ln\left(\frac{m_0}{m_f}\right)-\frac{v_gm_0}{D_e}\ln\left(\frac{m_0}{m_f}\right)-\frac{1}{2}gt^2
\end{aligned}\tag{16.5.77}$$

But when the maximum height is reached, the speed is zero, then we have:

$$v=v_g\ln\left(\frac{m_0}{m_f}\right)-gt\underset{v=0}{\Rightarrow}t=\frac{v_g}{g}\ln\left(\frac{m_0}{m_f}\right)\tag{16.5.78}$$

Then injected into the prior previous relation we get:

$$\begin{aligned}
z&=v_g\frac{m_0-m_f}{D_e}+v_g\left(\frac{v_g}{g}\ln\left(\frac{m_0}{m_f}\right)\right)\ln\left(\frac{m_0}{m_f}\right)-\frac{v_gm_0}{D_e}\ln\left(\frac{m_0}{m_f}\right)-\frac{1}{2}g\left(\frac{v_g}{g}\ln\left(\frac{m_0}{m_f}\right)\right)^2\\
&=v_g\frac{m_0-m_f}{D_e}+\frac{1}{g}\left(v_g\ln\left(\frac{m_0}{m_f}\right)\right)^2-\frac{v_gm_0}{D_e}\ln\left(\frac{m_0}{m_f}\right)-\frac{1}{2g}\left(v_g\ln\left(\frac{m_0}{m_f}\right)\right)^2\\
&=v_g\frac{m_0-m_f}{D_e}+\frac{1}{2g}\left(v_g\ln\left(\frac{m_0}{m_f}\right)\right)^2-\frac{v_gm_0}{D_e}\ln\left(\frac{m_0}{m_f}\right)
\end{aligned}\tag{16.5.79}$$

The first two terms are positive. The third and last term are negative. Therefore, we see that to maximize the height reached, it is best to make the mass flow D_e tends to infinity (ie: give all the linear momentum from the start!). Therefore, we have:

$$z_{\max}\underset{D_e\to+\infty}{=}=\frac{1}{2g}\left(v_g\ln\left(\frac{m_0}{m_f}\right)\right)^2\tag{16.5.80}$$

✍**Example:**

Let us calculate the maximum height of the first launch phase of Ariane 5 assuming constant gravity with the same values data as the previous example. This then gives:

$$z_{\max} \underset{D_e \to +\infty}{=} = \frac{1}{2 \cdot 9.86}\left(3600 \ln\left(\frac{700000}{260000}\right)\right)^2 \cong 655274 \cong 650\,[\text{km}] \qquad (16.5.81)$$

So this is the altitude at which the rocket would have a zero vertical speed and begin to fall. As we can see, it is well above the low orbit but well below the geostationary orbit as we will calculate thereafter.

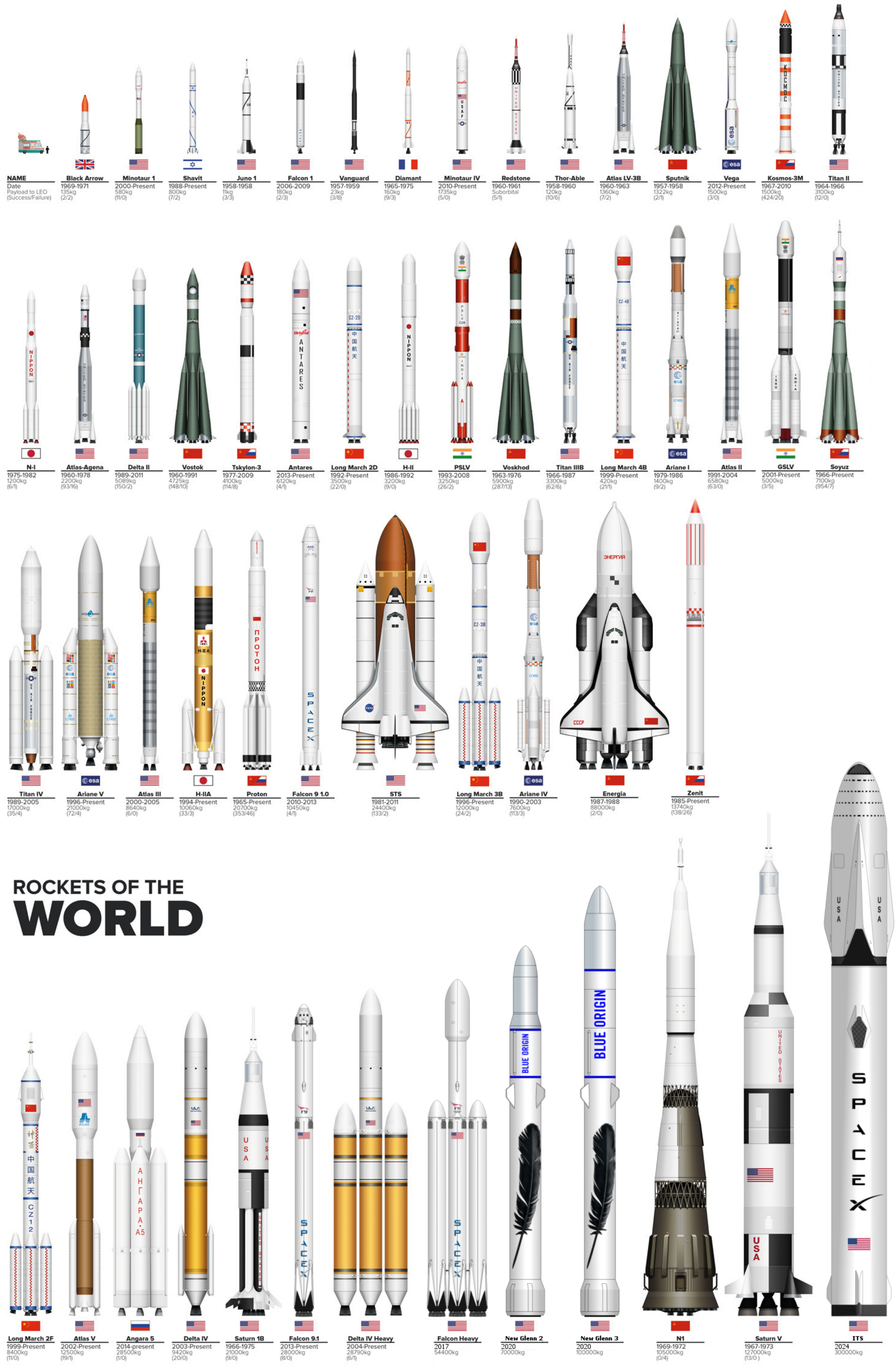

ROCKETS OF THE
WORLD
NAME
Date
Payload to LEO
(Success/Failure)
Black Arrow
1969-1971
135kg
(2/2)
Minotaur 1
2000-Present
580kg
(11/0)
Shavit
1988-Present
800kg
(7/2)
Juno 1
1958-1958
11kg
(3/3)
Falcon 1
2006-2009
180kg
(2/3)
Vanguard
1957-1959
23kg
(3/8)
Diamant
1965-1975
160kg
(9/3)
Minotaur IV
2010-Present
1735kg
(5/0)
Redstone
1960-1961
Suborbital
(5/1)
Thor-Able
1958-1960
120kg
(10/6)
Atlas LV-3B
1960-1963
1360kg
(7/2)
Sputnik
1957-1958
1322kg
(2/1)
Vega
2012-Present
1500kg
(3/0)
Kosmos-3M
1967-2010
1500kg
(424/20)
Titan II
1964-1966
3100kg
(12/0)
N-I
1975-1982
1200kg
(6/1)
Atlas-Agena
1960-1978
2200kg
(93/16)
Delta II
1989-2011
5089kg
(150/2)
Vostok
1960-1991
4725kg
(148/10)
Tskylon-3
1977-2009
4100kg
(114/8)
Antares
2013-Present
6120kg
(4/1)
Long March 2D
1992-Present
3500kg
(22/0)
H-II
1986-1992
3200kg
(9/0)
PSLV
1993-2008
3250kg
(26/2)
Voskhod
1963-1976
5900kg
(287/13)
Titan IIIB
1966-1987
3300kg
(62/6)
Long March 4B
1999-Present
420kg
(21/1)
Ariane I
1979-1986
1400kg
(9/2)
Atlas II
1991-2004
6580kg
(63/0)
GSLV
2001-Present
5000kg
(3/5)
Soyuz
1966-Present
7100kg
(954/7)
Titan IV
1989-2005
17000kg
(35/4)
Ariane V
1996-Present
21000kg
(72/4)
Atlas III
2000-2005
8640kg
(6/0)
H-IIA
1994-Present
10060kg
(33/3)
Proton
1965-Present
20700kg
(353/46)
Falcon 9 1.0
2010-2013
10450kg
(4/1)
STS
1981-2011
24400kg
(133/2)
Long March 3B
1996-Present
12000kg
(24/2)
Ariane IV
1990-2003
7600kg
(113/3)
Energia
1987-1988
88000kg
(2/0)
Zenit
1985-Present
13740kg
(138/26)
Long March 2F
1999-Present
8400kg
(11/0)
Atlas V
2002-Present
12500kg
(19/1)
Angara 5
2014-present
28500kg
(1/0)
Delta IV
2003-Present
9420kg
(20/0)
Saturn 1B
1966-1975
21000kg
(9/0)
Falcon 9.1
2013-Present
28000kg
(8/0)
Delta IV Heavy
2004-Present
28790kg
(6/1)
Falcon Heavy
2017
54400kg
New Glenn 2
2020
70000kg
New Glenn 3
2020
100000kg
N1
1969-1972
105000kg
(0/4)
Saturn V
1967-1973
127000kg
(13/0)
ITS
2024
300000kg

Figure 16.175: Rockets of the World (author: ?)

16.5.4 Geostationary orbit

The geostationary orbit is an orbit at $35,786$ [km] altitude above the equator of the Earth in the equatorial plane and zero orbital eccentricity. This is a special case of the geosynchronous orbit (otherwise the orbital period is always the period of revolution of the Earth orbit but the orbit then deviates north and south of the equator describing an analemma - a kind of irregular 8 shape geometry - in the sky when viewed from a fixed point on the surface of the Earth).

An object in such an orbit has an orbital period equal to the Earth's rotational period (one sidereal day) and thus appears motionless, at a fixed position in the sky, to ground observers. Communications satellites and weather satellites are often placed in geostationary orbits, so that the satellite antennas (located on Earth) that communicate with them do not have to rotate to track them, but can be pointed permanently at the position in the sky where the satellites are located. Using this characteristic, ocean color satellites with visible and near-infrared light sensors (e.g. the Geostationary Ocean Color Imager (GOCI)) can also be operated in geostationary orbit in order to monitor sensitive changes of ocean environments.

Geostationary satellites are then necessarily located vertically or at the zenith of a point on the equator or, in other words, located in the equatorial plane of the Earth.

Definition (#16.5.729): The formal definition of a "geostationary orbit" is that the orbit must be geosynchronous and with an orbit inclination of no more than 15° away from the equatorial plan (ITU Radioi Regulations, Article 9, footnote A.9.6A).

To calculate the position of the geostationary orbit, we will first use the Newton's second law (see section Classical Mechanics page 3202):

$$F=ma \tag{16.5.82}$$

and we have shown in the section of Classical Mechanics (see page 3388) when the movement is circular, we have for the centrifugal force:

$$a_{\text{ctrf}}=\frac{v^2}{R} \tag{16.5.83}$$

And we will use the law of gravity proved in the section of Classical Mechanics at page 5255 (and also proved in the section General Relativity at page 5751):

$$F=\frac{GMm}{R^2} \tag{16.5.84}$$

We will use these relations with the mass of the Earth $M_{\text{Earth}}=5.9736\cdot 10^{24}$ [kg], m_s the mass of the satellite, the average radius of the Earth $R_{\text{Earth}}=6,378.14$ [km] at the equator, h the height of the satellite relatively to the ground and v the speed of the satellite.

On geostationary orbit, there is therefore by definition a balance between the attractive gravitational forces of the planet and the centrifugal force of the satellite:

$$\frac{GMm}{R^2}=m\frac{v^2}{R} \tag{16.5.85}$$

By adopting the notations given above so this gives:

$$\frac{GM_{\text{Earth}}m_s}{(R_T+h)^2}=m_s\frac{v^2}{R_{\text{Earth}}+h} \tag{16.5.86}$$

We see then that the mass of the satellite m_s can be simplified. So the geostationary orbit is independent of it! It is also important to notice that since the speed v is after simplification independent of the mass of the satellite in a circular orbit, while an astronaut inside it will have the same speed and will be weightless inside thereto (it is the same for all nearby objects inside or outside of the satellite).

The speed for a circular path is the ratio of the circumference of the circle on the period of time to travel the whole circumference. So we have:

$$v=\frac{2\pi(R_{\text{Earth}}+h)}{T} \tag{16.5.87}$$

Therefore T being equal by definition in the context of geostationary orbit at the time of the day on Earth, we have taking the available tables:

$$T=23.93419\cdot 60\cdot 60\cong 86163.084\,[\text{s}] \tag{16.5.88}$$

Let us come back to prior-previous relation after simplification:

$$v=\frac{GM_{\text{Earth}}}{R_{\text{Earth}}+h}=v^2 \tag{16.5.89}$$

and by injecting in it the explicit relation of speed we get:

$$\frac{GM_{\text{Earth}}}{R_{\text{Earth}}+h}=\left(\frac{2\pi(R_{\text{Earth}}+h)}{T}\right)^2 \tag{16.5.90}$$

Therefore:

$$\left(\frac{GM_TT^2}{4\pi^2}\right)^{1/3}=R_T+h \tag{16.5.91}$$

Hence it comes finally:

$$\boxed{h=\left(\frac{GM_TT^2}{4\pi^2}\right)^{1/3}-R_T\cong 35,784\,[\text{km}]} \tag{16.5.92}$$

and the speed of the satellite in geostationary orbit is then:

$$v=\frac{2\pi(R_T+h)}{T}\cong 3,704\,[\text{m}\cdot\text{s}^{-1}] \tag{16.5.93}$$

Note also that using the previous relations, if the radius of the orbit of a non-geostationary satellite is given and also the mass of the Earth is known to us, we can then also determine the period of revolution without having to know its speed:

$$\boxed{T=2\pi\sqrt{\frac{(R_T+h)^3}{GM_T}}} \tag{16.5.94}$$

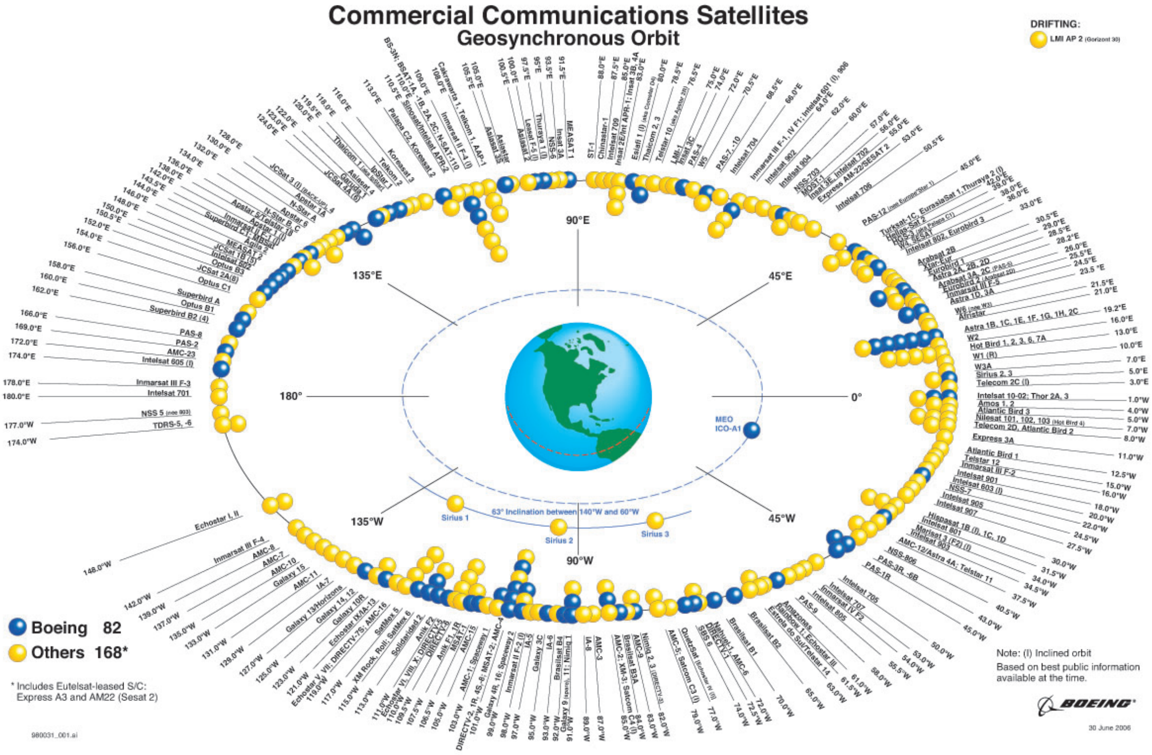

Commercial Communications Satellites
Geosynchronous Orbit
DRIFTING:
LMI AP 2 (Gorizont 30)
90°E
135°E
45°E
180°
0°
135°W
45°W
90°W
MEO
ICO-A1
Sirius 1
Sirius 2
Sirius 3
63° Inclination between 140°W and 60°W
Boeing 82
Others 168*
* Includes Eutelsat-leased S/C: Express A3 and AM22 (Sesat 2)
Note: (I) Inclined orbit
Based on best public information available at the time.
BOEING
30 June 2006
980031_001.ai

Figure 16.176: Geostationary satellites as given by public data in 12006 (source: Boeing)

But let us now calculate the value of g at the altitude of the International Space Station (ISS):

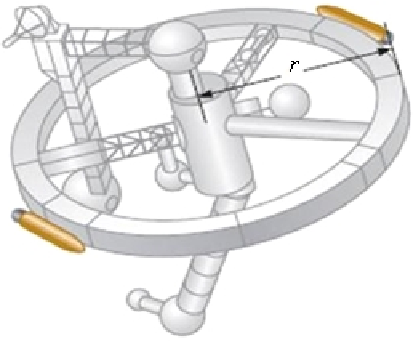
$$g = G\frac{M_{\text{Earth}}}{(r_{\text{Earth}}+h_{\text{ISS}})^2} \cong 6.67\cdot 10^{-11}\frac{5.9736\cdot 10^{24}}{(6.378\cdot 10^{6}+400\cdot 10^{3})^2} \cong 8.7\,[\text{m}\cdot\text{s}^{-2}] \qquad (16.5.95)$$

We often see video of astronauts in space stations, apparently weightless. But clearly, the force of gravity is acting on them. Comparing the value of g we just calculated to that on Earth $(9.80\,[\text{m}\cdot\text{s}^{-1}])$, we see that the astronauts in the International Space Station still have 88% of their weight. In fact they only appear to be weightless because like orbiting planets (or orbiting stars) they are constantly in free fall!!

But keep in mind that a permanent "free fall" in the case of orbits doesn't mean that the orbiting object will accelerate non-stop until it reaches asymptotically the speed of light, but that it accelerates on its orbit permanently because this same acceleration changes the direction of the velocity vectors (remember that acceleration is not always about speeding up but also about keeping the same speed but changing the direction of the corresponding vector)!

Figure 16.177: **International Space Station (source: NASA)**

Centrifugal force is obviously often used in space engineering calculations. For example, to calculate the speed that should have a circular space laboratory of radius r to simulate terrestrial gravity for an attached item on the floor of the rotating ring as illustrated below:



we simply apply the analysis of centrifugal force again:

$$mg = m\frac{v^2}{r} \tag{16.5.96}$$

where g is the desired gravity to simulate.

We see immediately that the mass vanish to get only:

$$v = \sqrt{gr} \Rightarrow \omega = \sqrt{\frac{g}{r}} \Rightarrow g = \frac{4\pi^2 r}{T^2} \tag{16.5.97}$$

This is the "spin gravity" relation.

Now if we consider that 2 revolutions per minute (RPM) is safe psychologically for an astronaut (it does not disturbs our brain too much), then we get using the above relation:

$$r=\frac{gT^2}{4\pi^2}=\frac{9.81\cdot(60/2)^2}{4\pi^2}\cong 223\,[\text{m}] \tag{16.5.98}$$

For comparison, in the movie *2001: A Space Odyssey*, the rotating ring had a radius of 280 [m]. So we can imagine the cost of such a structure in space and why it has not been realized until now...

Let's see for the fun what should be the radius of such ring to get the same acceleration and period than that of the Earth:

$$r=\frac{gT^2}{4\pi^2}=\frac{9.81\cdot(24\cdot 60\cdot 60)^2}{4\pi^2}\cong 1.85\cdot 10^9\,[\text{m}] \tag{16.5.99}$$

That's quite a huge ring...

Once again, this circular cylinder only works as artificial gravity if you are touching it. So if you placed a basketball anywhere where it is not touching the cylinder, it would stay exactly still. If a person jumped to the center of the cylinder, he would indeed keep going!

16.5.5 Vis-Viva equation

The vis-viva equation, also referred to as "orbital-energy-invariance law", is one of the equations that model the motion of orbiting bodies. It is the direct result of the principle of conservation of mechanical energy which applies when the only force acting on an object is its own weight.

Remark

Vis-viva (Latin for "live force") is a term from the history of mechanics, and it survives in this sole context. It represents the principle that the difference between the aggregate work of the accelerating forces of a system and that of the retarding forces is equal to one half the vis-viva accumulated or lost in the system while the work is being done.

In the vis-viva equation the mass m of the orbiting body (e.g., a spacecraft) is taken to be negligible in comparison to the mass M of the central body (e.g., the Earth). In the specific cases of an elliptical or circular orbit, the vis-viva equation may be readily derived from conservation of energy and momentum.

Specific total energy is constant throughout the orbit. Thus, using the subscripts a and p to denote apoapsis (apogee) and periapsis (perigee), respectively (see the study of conics in the section of Analytical Geometry at page 2764) we know that on any point of an elliptic orbit we have by conservation of energy:

$$E_{\text{tot}}=E_c+E_p=\frac{1}{2}mv_a^2-\frac{GMm}{r_a}=\frac{1}{2}mv_p^2-\frac{GMm}{r_p} \tag{16.5.100}$$

But using the first cosmic speed:

$$v_a=\sqrt{\frac{GM}{r_a}}\qquad v_p=\sqrt{\frac{GM}{r_p}} \tag{16.5.101}$$

We get (we will focus only on the apogee as the developments are the same for the perigee):

$$E_{\text{tot}}=\frac{1}{2}m\left(\sqrt{\frac{GM}{r_a}}\right)^2-\frac{GMm}{r_a}=\frac{1}{2}\frac{GMm}{r_a}-\frac{GMm}{r_a} \tag{16.5.102}$$

Therefore:

$$\boxed{E_{\text{tot}}=-\frac{1}{2}\frac{GMm}{r_a}} \tag{16.5.103}$$

Notice that $E_{\text{tot}}<0$. This is named a "bound orbit".

Let's equate the last special case expression with for any point of the orbit:

$$-\frac{1}{2}\frac{GMm}{r_a}=\frac{1}{2}mv^2-\frac{GMm}{r} \tag{16.5.104}$$

Doing a little bit of algebra and putting $a:=r_a$ we get:

$$\boxed{v=\sqrt{GM\left(\frac{2}{r}-\frac{1}{a}\right)}} \tag{16.5.105}$$

which gives the object's speed at any point along the orbit. This is the "vis-viva equation". We often give only the apogee version of this equation as it is the position used in Hohmann transfer orbit (see further below). It is interesting to notice that if $a=r$ (circular orbit) we fall back on the first cosmic speed (low-level orbit speed):

$$v=\sqrt{\frac{GM}{r}} \tag{16.5.106}$$

This powerful equation does not depend on orbital eccentricity! For instance, if we observe a new object in the solar system and know its current velocity and distance, we can determine its orbital semi-major axis a and thus have some idea where it came from.

Notice that the second cosmic speed (release velocity) can be obtained from the vis-viva equation by taking the limit as a approaches $+\infty$:

$$v^2=GM\left(\frac{2}{r}-0\right)\Rightarrow v=\sqrt{\frac{2GM}{r}} \tag{16.5.107}$$

There is also an interesting application to the above relation if we take $a=r_a+r_p$ and $r=r_p$ to determine the speed of the today's fastest known object in the Solar System:

$$\Delta v=\sqrt{GM\left(\frac{2}{r}-\frac{1}{a}\right)}=\sqrt{GM\left(\frac{2}{r_p}-\frac{2}{r_a+r_p}\right)}=\sqrt{GM\frac{(2(r_p+r_a)-2r_p}{r_p(r_a+r_p)}}=\sqrt{\frac{2GMr_a}{r_p(r_a+r_p)}} \tag{16.5.108}$$

This relation that gives the perihelion speed is often denoted as:

$$v_p=\sqrt{\frac{2GMr_a}{r_p(r_a+r_p)}} \tag{16.5.109}$$

And taking for application the Sun (actually a G2V-type star, aka yellow dwarf) and the Great Comet of 11843 (formally designated C/1843 D1 and 1843 I) that has a perihelion r_p of only $149,597,871$ [km] (this means it came to within less than $121,000$ [km] of the surface of the Sun, and amazingly it survived!) and an aphelion r_a of $23,337,267,876$ [km], we get:

$$v_p = \sqrt{\frac{2GM_{\odot}r_a}{r_p(r_a+r_p)}} \cong 570\,[\text{km}\cdot\text{s}^{-1}] \tag{16.5.110}$$

16.5.6 Gravity assist

Interplanetary space probes often make use of the "gravitational slingshot" effect, also named "gravity assist", to propel them to high velocities. For example, Voyager 2 performed a close flyby of Saturn on the 27th of August in 11981 (holocene calendar), which had the effect of slinging it toward its flyby of Uranus on the 30th of January in 11986 (holocene calendar). Since gravity is a conservative force, it may seem strange that an object can achieve a net gain in speed due to a close encounter with a large gravitating mass. We might imagine that the speed it gains while approaching the planet would be lost when receding from the planet. However, this is not the case, as we can see from simple consideration of the kinetic energy and momentum, which shows how a planet can transfer kinetic energy to the spacecraft.

The Cassini–Huygens spacecraft was launched from Earth on 15 October 11997 (holocene calendar), followed by gravity assist flybys of Venus (26 April 1998 and 21 June 11999), Earth (18 August 11999), and Jupiter (30 December 12000). Transit to Saturn took 6.7 years, the spacecraft arrived at 1 July 12004. Its trajectory was named by some "*the Most Complex Gravity-Assist Trajectory Flown to Date*" in 12019:

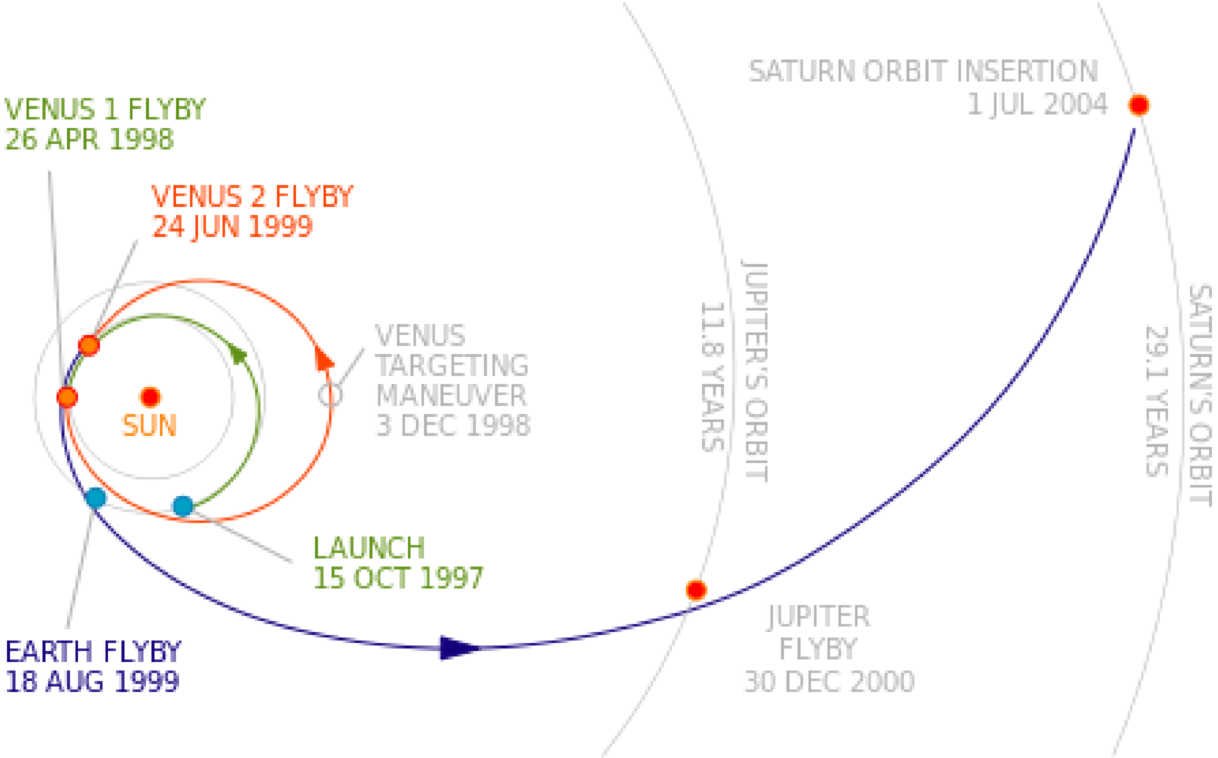

SATURN ORBIT INSERTION
1 JUL 2004
VENUS 1 FLYBY
26 APR 1998
VENUS 2 FLYBY
24 JUN 1999
VENUS
TARGETING
MANEUVER
3 DEC 1998
SUN
JUPITER'S ORBIT
11.8 YEARS
SATURN'S ORBIT
29.1 YEARS
LAUNCH
15 OCT 1997
JUPITER
FLYBY
30 DEC 2000
EARTH FLYBY
18 AUG 1999

Figure 16.178: **Cassini interplanetary trajectory (source: Wikipedia)**

Below we can see Cassini's speed related to Sun. The various gravity assists form visible peaks on the left, while the periodic variation on the right is caused by the spacecraft's orbit around Saturn. The data was from JPL Horizons Ephemeris System. The speed above is in kilometers per second and the years are based on the Gregorian calendar. Note also that the minimum speed achieved during Saturnian orbit is more or less equal to Saturn's own orbital velocity, which is the $\sim 5\,[\text{km}\cdot\text{s}^{-1}]$ velocity which Cassini matched to enter orbit:

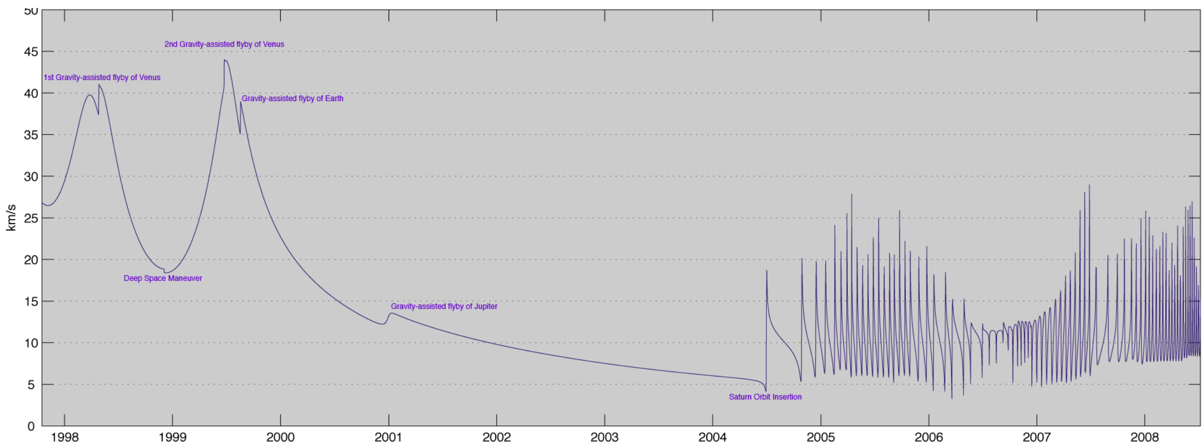
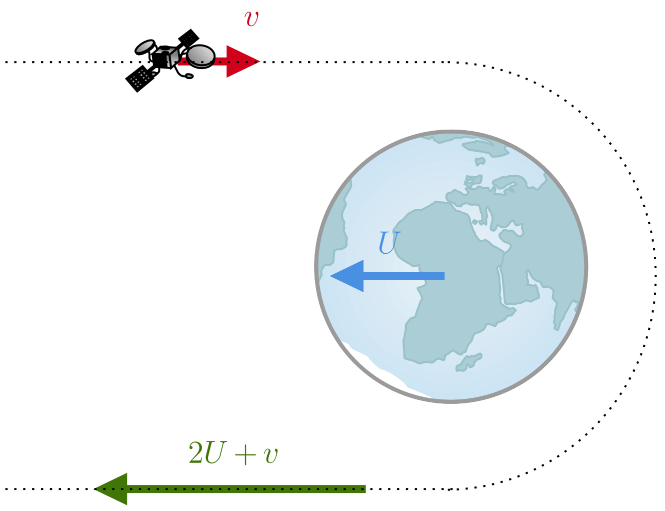


Figure 16.179: Cassini's speed related to Sun (source: Wikipedia)

16.5.6.1 Speed variation (in solar reference frame)

An extreme form of the maneuver would be to approach a planet head-on at a speed v while the planet is moving directly toward us at a speed U (both speeds defined relative to the "fixed" Solar frame). If we aim just right we can loop around behind the planet in an extremely eccentric hyperbolic orbit, making a virtual 180° turn, as illustrated below:

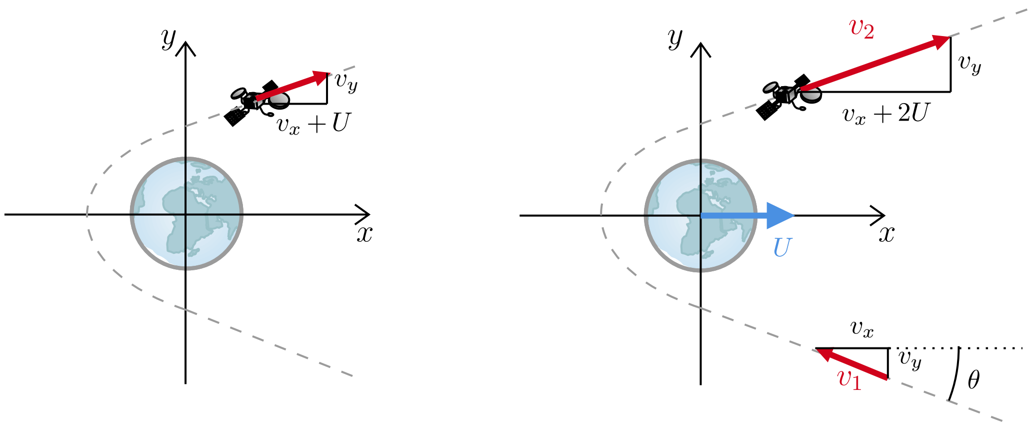


So we have:

$$v_f = 2U + v \tag{16.5.111}$$

according to the proof we did in the section of Classical Mechanics of an elastic collision (see page 3369).

Of course, most planetary fly-bys are not simple head-on reversals, but the same principles apply for any angle of interaction. Let's take the planet's direction of motion as the x-axis, and the perpendicular direction (in the orbital plane) as the y-axis. The probe is initially moving with a speed v relative to the solar reference frame, in a direction approaching the oncoming planet at an angle θ. Two views of this are shown below, one with respect to the planet's rest frame, and the other with respect to the solar reference frame:



Planet's rest frame **Solar reference frame**

By drawing a simple parallogram of speeds for the probe and planet intersecting at an arbitrary angle θ, and assuming we arrange for a hyperbolic orbit symmetrical about the $x-$axis (with respect to the planet's rest frame), the probe's initial velocity vector with respect to the Sun's rest frame is:

$$(v_1)_x = -v_1\cos(\theta) \qquad (v_1)_y = v_1\sin(\theta) \tag{16.5.112}$$

and its final velocity vector is:

$$(v_2)_x = v_1\cos(\theta)+2U \qquad (v_2)_y = v_1\sin(\theta) \tag{16.5.113}$$

Hence:

$$\begin{aligned}
v_2 &= \sqrt{v_{2x}^2+v_{2y}^2} = \sqrt{(v_1\cos(\theta)+2U)^2+(v_1\sin(\theta))^2}\\
&= \sqrt{v_1^2\cos^2(\theta)+4Uv_1\cos(\theta)+4U^2+v_1^2\sin^2(\theta)} = \sqrt{v_1^2+4Uv_1\cos(\theta)+4U^2}\\
&= \sqrt{(v_1+U)^2+4Uv_1\cos(\theta)-4Uv_1} = \sqrt{(v_1+U)^2-4Uv_1(1-\cos(\theta))}\\
&= (v_1+2u)\sqrt{1-\frac{4uv_1(1-\cos(\theta))}{(v_1+2u)^2}}
\end{aligned} \tag{16.5.114}$$

The reader must keep however in mind that given the previous assumption, the relation:

$$\boxed{v_2 = (v_1+2u)\sqrt{1-\frac{4uv_1(1-\cos(\theta))}{(v_1+2u)^2}}} \tag{16.5.115}$$

is not the general case!

For example, suppose the initial speeds of the probe and the planet happen to be exactly the same (i.e., $v_1 = U$). In this case the above relation reduces to:

$$v_2 = v_1\sqrt{5+4\cos(\theta)} \tag{16.5.116}$$

which confirms that when $\theta = 0$ we have $v_2 = 3v_1$, which is our head-on reversal case. On the other hand, when $\theta = \pi$ we have $v_2 = v_1$, which stands to reason, because in this case the probe and planet are going in the same direction at the same speed. For a more realistic case, we can have the probe approach nearly perpendicular to the planet's path (i.e., $\theta = \pi/2$) and swing just behind it. In that case the probe gets deflected in the direction of the planet's travel, at an angle given by the above formulas, and it's final speed is the square root of 5 (i.e., about 2.23) times its original speed.

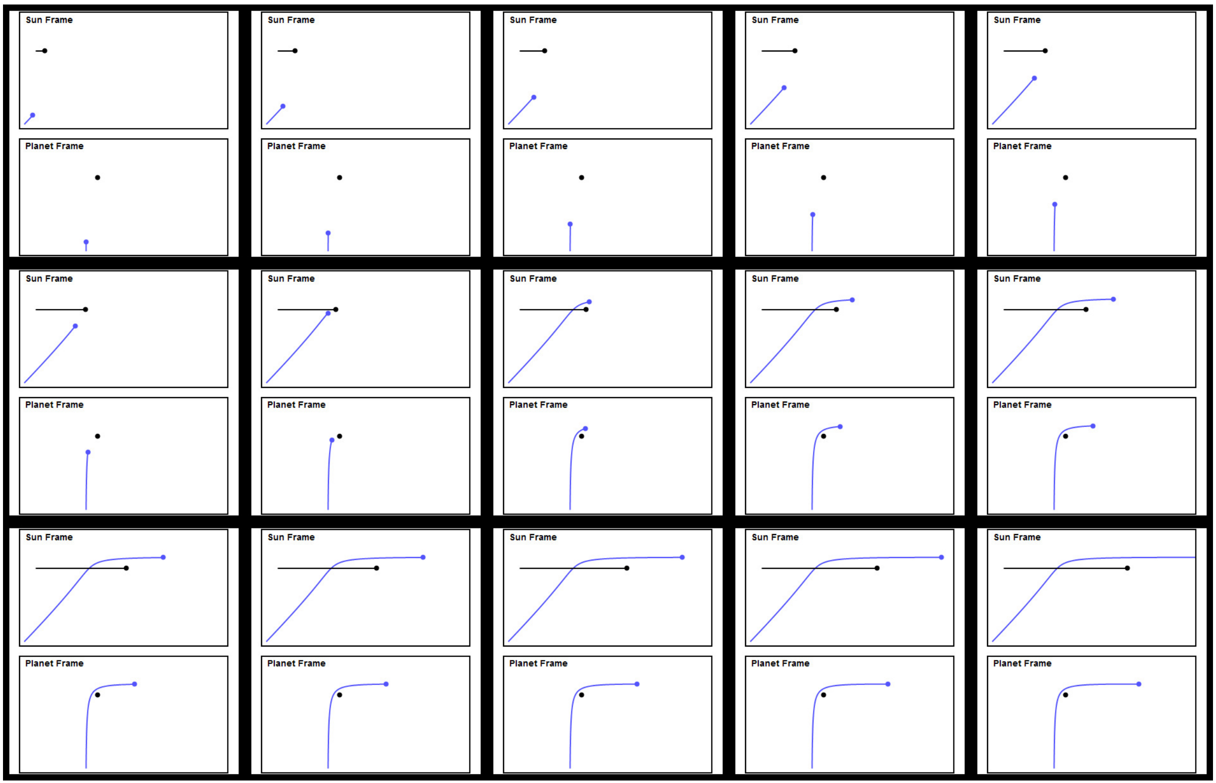


Figure 16.180: **Special case of gravity assist animation depending on reference frame**

If the planets were point particles, then according to classical physics it would be theoretically possible (in some rather contrived solar systems) for an object to acquire infinite speed in finite time by looping repeatedly around a set of planets. Of course, since actual planets are not point-like, the external gravitational field of an ordinary planet would not be strong enough to "grab" the spaceship once it was travelling above a certain speed. The limit is how fast we can loop around a planet without dipping into its atmosphere too deeply (let alone crashing into it). Some NASA missions have repeatedly skimmed the upper atmospheres of Venus and the Earth in their maneuvers.

Conceivably, if we (or someone else) ever found a star system consisting of multiple black holes orbiting each other, it might be possible to apply this scheme to achieve relativistic speeds, by looping around from one to the other. In this situation the achievable speed limit would depend on how close a spaceship could pass without being destroyed by tidal forces. Still, if the black holes were large enough, the tidal forces even at the event horizon would be tolerable, although it probably wouldn't be possible to have a controllable hyperbolic orbit pass closer than, say $3\,[\text{m}]$. Also, stopping the vehicle at the destination would be difficult.

16.5.6.2 Approaching speed from infinity

Before we continue with the next important relation, we have seen that quite a lot of people on the Internet ask for the final speed of an object (probe, asteroid, meteorite, missile, etc.) impacting a planet considered at rest (or not)!

So to deal with this little problem let us consider the following assumptions:

- The object starts with zero initial velocity at distance D from the planet's center.
- Only gravity acts (no atmosphere).
- Planet has radius R.
- Mass of asteroid m cancels out.

We know that the potential gravific energy at a distance r is given by:

$$E_p(r)=-\frac{GMm}{r} \tag{16.5.117}$$

Energy conservation between initial position D and impact at radius R gives us:

$$\frac{1}{2}mv^2=\frac{GMm}{R}-\frac{GMm}{D} \tag{16.5.118}$$

We simplify and rearrange to get:

$$v=\sqrt{2GM\left(\frac{1}{R}-\frac{1}{D}\right)} \tag{16.5.119}$$

If the object starts very far away $(D\rightarrow+\infty)$, we then have:

$$\boxed{v_\infty=\sqrt{\frac{2GM}{R}}} \tag{16.5.120}$$

We fall back (quite logically) on the escape velocity of the planet!

If the asteroid is initially at rest in some inertial frame, and the planet of mass M is moving at speed U in a straight line toward it, then in the planet's rest frame the asteroid initially has speed U at distance D from the planet's center.

Ignoring atmosphere and assuming a head-on collision (motion along the line of centers), the impact speed comes from energy conservation in the planet's frame.

So we have the following informations:

- Initial kinetic energy of the object:

$$E_{c,i}=\frac{1}{2}mU^2 \tag{16.5.121}$$

- Initial potential energy of the object:

$$E_{p,i}=-\frac{GMm}{D} \tag{16.5.122}$$

- Final kinetic energy at impact (speed v):

$$E_{c,f}=\frac{1}{2}mv^2 \tag{16.5.123}$$

- Final potential energy at planet surface (radius R):

$$E_{p,f}=-\frac{GMm}{R} \tag{16.5.124}$$

Using energy conservation, we write:

$$E_{c,i}+E_{p,i}=E_{c,f}+E_{p,f} \tag{16.5.125}$$

Explicitly:

$$\frac{1}{2}mU^2-\frac{GMm}{D}=\frac{1}{2}mv^2-\frac{GMm}{R} \tag{16.5.126}$$

After rearranging and simplifying we get the final relation:

$$\boxed{v=\sqrt{U^2+2GM\left(\frac{1}{R}-\frac{1}{D}\right)}} \tag{16.5.127}$$

So this is the impact speed in the planet's rest frame. We can see that the object's (asteroid, meteorite, probe, etc.) mass m cancels out! If $U=0$, you recover the previous result:

$$v=\sqrt{2GM\left(\frac{1}{R}-\frac{1}{D}\right)} \tag{16.5.128}$$

16.5.6.3 Velocity variation and impact factor

Let us consider the following figure:

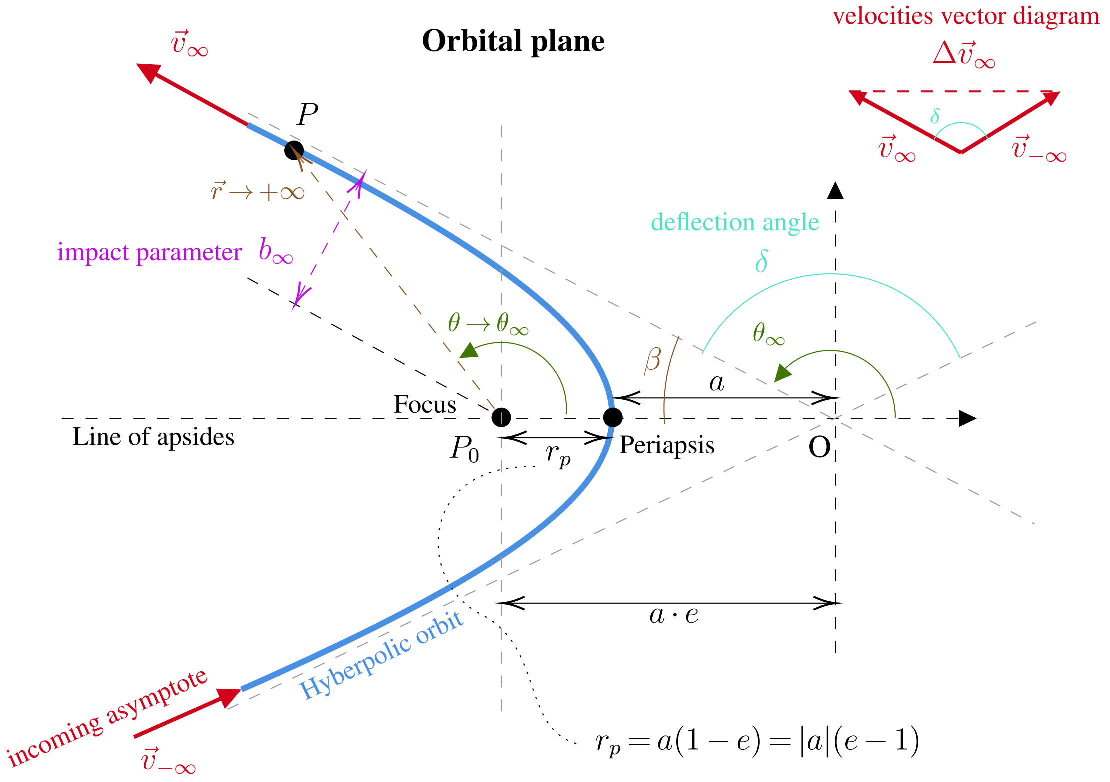
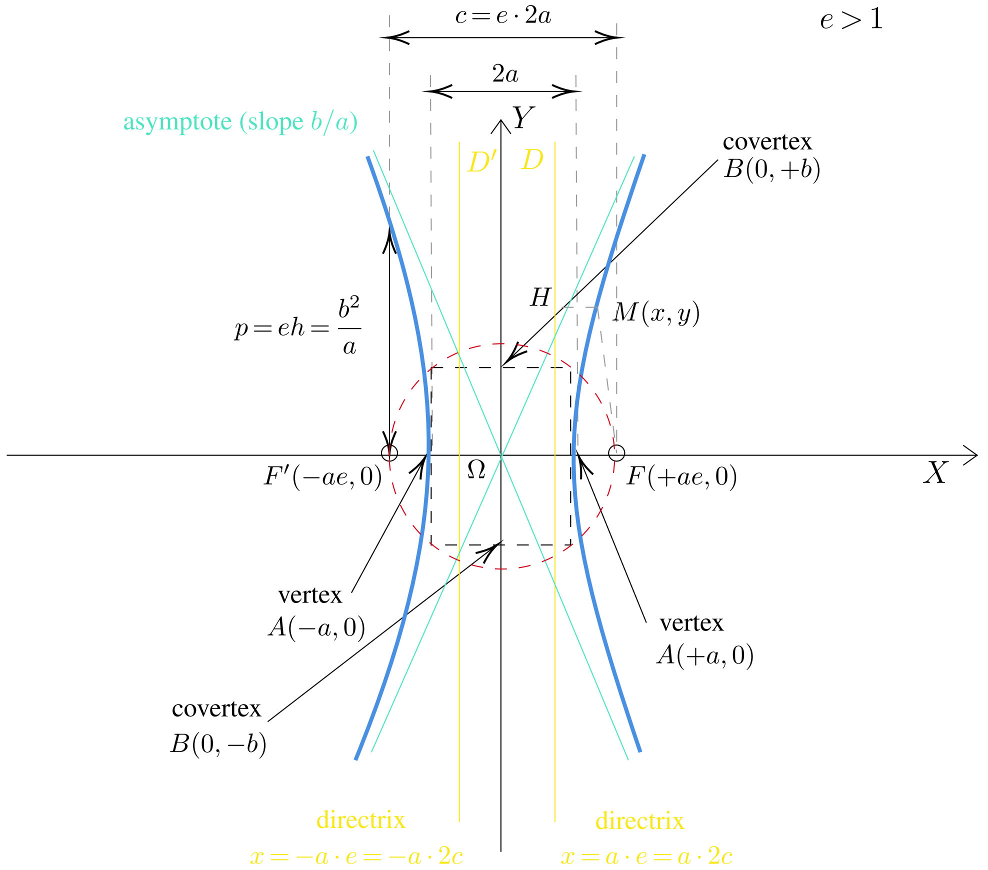

Orbital plane
velocities vector diagram
$\Delta\vec{v}_\infty$
δ
$\vec{v}_\infty$
$\vec{v}_{-\infty}$
$\vec{v}_\infty$
P
$\vec{r}\rightarrow+\infty$
impact parameter b_∞
deflection angle
δ
$\theta\rightarrow\theta_\infty$
θ_∞
β
a
Focus
Line of apsides
P_0
r_p
Periapsis
O
$a\cdot e$
Hyberpolic orbit
incoming asymptote
$\vec{v}_{-\infty}$
$r_p=a(1-e)=|a|(e-1)$

Figure 16.181: **Hyperbolic orbit with classical notation a gravity assist vocabulary**

and let us put here again the figure used in the section Analytical Geometry during our study of the hyperbola (see page 2787):


$c=e\cdot 2a$
$e>1$
$2a$
asymptote (slope b/a)
Y
D'
D
covertex
$B(0,+b)$
H
$M(x,y)$
$p=eh=\frac{b^2}{a}$
$F'(-ae,0)$
Ω
$F(+ae,0)$
X
vertex
$A(-a,0)$
vertex
$A(+a,0)$
covertex
$B(0,-b)$
directrix
$x=-a\cdot e=-a\cdot 2c$
directrix
$x=a\cdot e=a\cdot 2c$

and the definition introduce also at the same page 2787:

$$e=\frac{c}{a}=\frac{\sqrt{a^2+b^2}}{a} \tag{16.5.129}$$

Therefore we can relate (the left hyperbola):

$$\cos(\theta_\infty)=\frac{-a}{\sqrt{a^2+b^2}}=-\frac{1}{e} \tag{16.5.130}$$

Moreover, since:

$$\theta_\infty=\pi-\beta \tag{16.5.131}$$

we have:

$$\cos(\beta)=\cos(\pi-\theta_\infty)=-\cos(\theta_\infty)=\frac{1}{e} \tag{16.5.132}$$

Due to the hyperbolic encounter, the hyperbolic excess velocity is turned by an angle δ, the "deflection angle", such that:

$$2\beta+\delta=\pi \tag{16.5.133}$$

That is:

$$\frac{1}{e}=\cos(\beta)=\cos\left(\frac{\pi}{2}-\frac{\delta}{2}\right)=\sin\left(\frac{\delta}{2}\right) \tag{16.5.134}$$

If we consider negligible the interval of time spent inside the sphere of influence, then the interaction between the probe and the planet can be seen as an impulsive maneuver. For the law of cosines (see section Trigonometry page 2524) in the red triangle of the prior-previous figure (in the top right corner), we have:

$$\begin{aligned}\Delta v&=\sqrt{v_\infty^2+v_\infty^2-2v_\infty^2\cos(\delta)}=\sqrt{2v_\infty^2(1-\cos(\delta))}\\&=2v_\infty\sin\left(\frac{\delta}{2}\right)=2\frac{v_\infty}{e}\end{aligned} \tag{16.5.135}$$

which states that the impulse is maximum when the eccentricity is minimum (remember that for the hyperbolic case the constraint $e>1$. So in the best case we have:

$$\Delta v\cong 2v_\infty \tag{16.5.136}$$

However, we must ensure that the probe does not collide with the planet, that is, that the pericenter of the hyperbola must be higher than the radius of the planet, namely:

$$r_{\text{pericenter}}=a(e-1)>r_{\text{planet}} \tag{16.5.137}$$

Now remember that we have proven it earlier above (Vis-Viva equation at page 8619) the orbital speed is given by:

$$v=\sqrt{GM\left(\frac{2}{r}-\frac{1}{a}\right)} \tag{16.5.138}$$

In the case where $r \to +\infty$ this reduces to:

$$v=\sqrt{\frac{GM}{a}} \tag{16.5.139}$$

Often denoted in the field of gravity assist studies:

$$v_{\infty}=\sqrt{\frac{\mu_p}{a}} \tag{16.5.140}$$

Hence injecting this relation into:

$$r_{\text{pericenter}}=a(e-1)>r_{\text{planet}} \tag{16.5.141}$$

and rearranging, we get:

$$\frac{\mu_p}{v_{\infty}^2}(e-1)>r_{\text{planet}} \tag{16.5.142}$$

This implies:

$$e>1+r_p\frac{v_{\infty}^2}{\mu_p} \tag{16.5.143}$$

And therefore:

$$e_{\text{min}}\cong 1+r_p\frac{v_{\infty}^2}{\mu_p} \tag{16.5.144}$$

Let us now define the "impact parameter as the perpendicular distance by which the probe would miss the planet without gravity and denoted b_{∞} as visible in the first figure above.

We use now a similar reasoning as before. First we see that we have:

$$\frac{b_{\infty}}{a\cdot e}=\sin(\beta) \tag{16.5.145}$$

However as $\beta=\pi/2-\delta/2$, we have:

$$\frac{b_{\infty}}{a\cdot e}=\sin(\beta)=\sin(\pi/2-\delta/2)=\cos\left(\frac{\delta}{2}\right)=\pm\sqrt{\frac{1+\cos(\delta)}{2}} \tag{16.5.146}$$

Hence:

$$\cos(\delta)=\frac{2b_{\infty}^2}{a^2e^2}-1 \tag{16.5.147}$$

Using the cosine law as earlier above:

$$\begin{aligned}\Delta v&=\sqrt{v_{\infty}^2+v_{\infty}^2-2v_{\infty}^2\cos(\delta)}=\sqrt{2v_{\infty}^2(1-\cos(\delta))}\\&=\sqrt{2v_{\infty}^2\left(1-\left(\frac{2b_{\infty}^2}{a^2e^2}-1\right)\right)}=\sqrt{4v_{\infty}^2\left(1-\frac{b_{\infty}^2}{a^2e^2}\right)}\end{aligned} \tag{16.5.148}$$

Now using one of the property of the hyperbola proven in the section of Analytical Geometry (see page 2789):

$$b_\infty^2 = a^2\left(e^2 - 1\right) \tag{16.5.149}$$

Rearranging:

$$\frac{b_\infty^2}{a^2e^2} = \frac{e^2-1}{e^2} = 1 - \frac{1}{e^2} \tag{16.5.150}$$

Plugged into previously equation:

$$\begin{aligned}\Delta v &= \sqrt{2v_\infty^2 2\left(1-\left(1-\frac{1}{e^2}\right)\right)} = \sqrt{2v_\infty^2\frac{2}{e^2}} = \sqrt{\frac{4v_\infty^2}{e^2}}\\ &= \frac{2v_\infty}{e}\end{aligned} \tag{16.5.151}$$

So we fall back on the same result with a different approach!

Now remember that we have proven earlier above (Vis-Viva equation at page 8619) the orbital speed is given by:

$$v = \sqrt{GM\left(\frac{2}{r} - \frac{1}{a}\right)} \tag{16.5.152}$$

Hence:

$$v_\infty^2 = -\frac{GM}{a} = -\frac{\mu}{a} \tag{16.5.153}$$

Solving for $1/a$ gives:

$$\frac{1}{a} = -\frac{v_\infty^2}{\mu} \tag{16.5.154}$$

Squaring this result provides a term for the geometric eccentricity formula:

$$\frac{1}{a^2} = \left(\frac{v_\infty^2}{\mu}\right)^2 \tag{16.5.155}$$

Substituting the expression for $1/a^2$ and the identity $b = b_\infty$ into the standard hyperbolic eccentricity equation:

$$e^2 = 1 + \frac{b^2}{a^2} \tag{16.5.156}$$

we get:

$$e^2 = 1 + b_\infty^2\left(\frac{v_\infty^2}{\mu}\right)^2 \Rightarrow e = \sqrt{1+\left(\frac{b_\infty v_\infty^2}{\mu}\right)^2} \tag{16.5.157}$$

This formula demonstrates that the eccentricity of a gravity assist trajectory is determined solely by the incoming speed and the targeting offset relative to the gravitational strength of the planet.

Now, since we know:

$$e=\sqrt{1+\left(\frac{b_\infty v_\infty^2}{\mu}\right)^2} \tag{16.5.158}$$

we substitute e in the final result:

$$\boxed{\Delta v=\frac{2v_\infty}{e}=\frac{2v_\infty}{\sqrt{1+\left(\frac{b_\infty v_\infty^2}{\mu}\right)^2}}=\frac{2v_\infty}{\sqrt{1+\left(\frac{b_\infty v_\infty^2}{GM}\right)^2}}} \tag{16.5.159}$$

16.5.7 Hohmann Transfer orbit

A Hohmann transfer is a transfer orbit which uses an ellipse where one end of the ellipse is at the starting planetary orbit and the other end is at the destination planet orbit in the same plane as shown in the figure below:

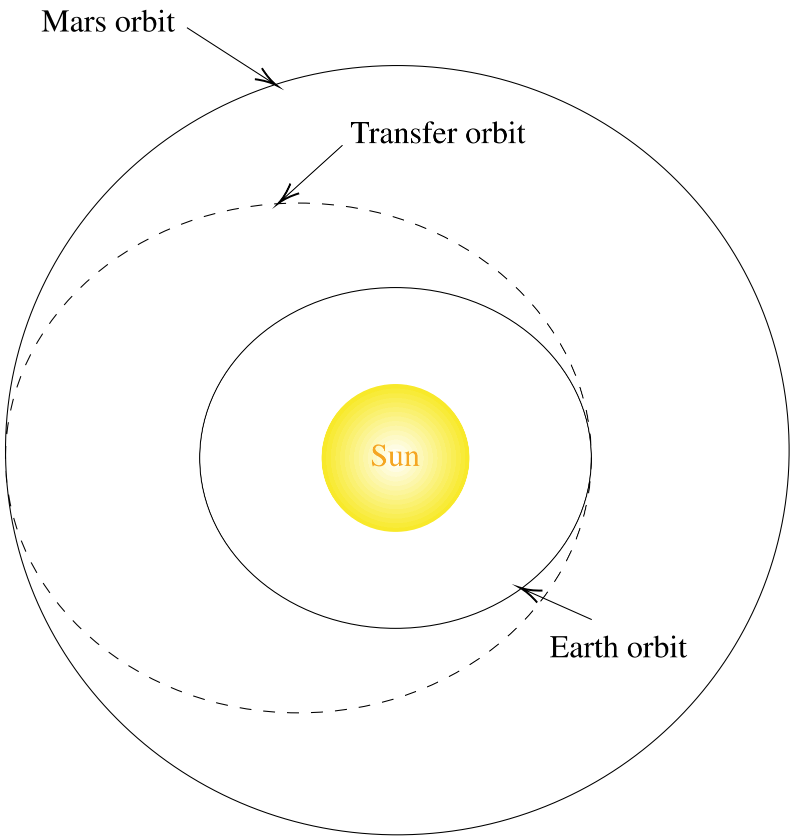


Figure 16.182: **Hohmann transfer between Earth and Mars (not to scale)**

The orbital maneuver to perform the Hohmann transfer uses two engine impulses, one to move a spacecraft onto the transfer orbit and a second to move off it.

In the above the transfer orbit meets Earth's orbit at periapse and Mars' orbit at apoapse. The velocities needed at these two points can easily be found using the vis-via equation proved previously.

The figure above show that the initial circular orbit around the Earth and departure (transfer) orbit share a common point. Therefore the satellite must apply a maneuver at this point (tangential acceleration) in order to change its velocity. The initial velocity of the spacecraft is that of an object in circular orbit a given as we know by the first cosmic speed (low-level orbit speed):

$$v=\sqrt{\frac{GM}{R}} \tag{16.5.160}$$

a numerical application for Earth gives $7.73\,[\mathrm{km}\cdot\mathrm{s}^{-1}]$ when we want to put something at 300

[km] height. The corresponding kinetic energy is immediate:

$$E_c = \frac{1}{2}mv^2 = \frac{1}{2}m\left(\sqrt{\frac{GM}{r}}\right)^2 = \frac{GMm}{2r} \cong 2.24 \cdot 10^9 \text{ [J]} \tag{16.5.161}$$

But at the same time we know thanks to the Vis-Viva equation what must be the speed at a given point r of an ellipse so that the satellite we later be on point a (transfer orbit) :

$$v = \sqrt{GM\left(\frac{2}{r} - \frac{1}{a}\right)} \tag{16.5.162}$$

So as we know r this is the common point between the Earth's orbit and transfer orbit and that we know a that is the common point of the transfer orbit and Mars' orbit the difference gives us the speed to gain:

$$\Delta v = \sqrt{GM\left(\frac{2}{r} - \frac{1}{a}\right)} - \sqrt{\frac{GM}{r}} \tag{16.5.163}$$

And as we have (kinematics):

$$\Delta v = a\Delta t \tag{16.5.164}$$

where a is the acceleration and not the apoapse, we can know what is the acceleration (and therefore the force) that has to be applied to the satellite as quick as possible to put it on its transfer orbit.

If we want for example to take a satellite to elliptic orbit (transfer orbit) to put it in geostationnary orbit $(42,300$ [km]) for it's actual orbit of (300 [km], i.e. $6,680$ [km] for Earth's center), we then have first by assimilating Earth's center naturally by the focus of this elliptic transfer orbit of semi-grand axis a:

$$a = \frac{r_a + r_p}{2} = \frac{42300 + 6680}{2} = 21400 \text{ [km]} \tag{16.5.165}$$

we have (taking $r = 6,680$ [km]):

$$v = \sqrt{GM\left(\frac{2}{r} - \frac{1}{a}\right)} = \sqrt{GM\left(\frac{2}{6680 \cdot 10^3} - \frac{1}{21400 \cdot 10^3}\right)} \cong 10.158 \text{ [km} \cdot \text{s}^{-1}] \tag{16.5.166}$$

Therefore our satellite will have to change from approximately 7.73 [km $\cdot$ s^{-1}] to 10.16 [km $\cdot$ s^{-1}] to be put in a transfer orbit that will take him at $42,300$ [km] of Earth.

The corresponding energy is immediate:

$$E_c = \frac{1}{2}mv^2 = \frac{1}{2}m\left(\sqrt{GM\left(\frac{2}{r} - \frac{1}{a}\right)}\right)^2 \cong 6.11 \cdot 10^{11} \text{ [J]} \tag{16.5.167}$$

Therefore a energy variation of $\Delta E_c \cong 1.63 \cdot 10^9$ [J].

The last step is obviously to change the speed of the satellite when it is at the perigee of the transfer orbit. So first we calculate its speed when it is at the perigee of the orbit transfer:

$$v=\sqrt{GM\left(\frac{2}{r}-\frac{1}{a}\right)}=\sqrt{GM\left(\frac{2}{42300\cdot 10^3}-\frac{1}{21400\cdot 10^3}\right)}\cong 1.6\,[\text{km}\cdot\text{s}^{-1}] \qquad (16.5.168)$$

when the speed for a circular orbit at this same distance is:

$$v=\sqrt{\frac{GM}{r}}=\sqrt{\frac{GM}{42300\cdot 10^3}}\cong 3.07\,[\text{km}\cdot\text{s}^{-1}] \qquad (16.5.169)$$

We have therefore to speed up the satellite from $1.6\,[\text{km}\cdot\text{s}^{-1}]$ to $3.07\,[\text{km}\cdot\text{s}^{-1}]$, to transform the elliptic transfer orbit into the final circular geostationary orbit.

And again as the necessary kinetic energy of the geostationary orbit is:

$$E=\frac{GMm}{2r}=\frac{GMm}{2\cdot 42300\cdot 10^3}\cong 3.54\cdot 10^8 \qquad (16.5.170)$$

Therefore the energy that we must give for the last time to satellite is:

$$|\Delta E|=|3.54\cdot 10^8-6.11\cdot 10^8|\cong 2.57\cdot 10^8\,[\text{J}] \qquad (16.5.171)$$

Finally, one can calculate the transit time for the Hohmann transfer to complete. As seen in the figure above, the satellite on the transfer orbit travels through π (180°) of true anomaly. Thus the transfer time can be found by taking half of the elliptic orbit period given by the third's Kepler law (see section Astronomy page 5242):

$$T=\frac{1}{2}\sqrt{\frac{a^3 4\pi^2}{Gm}}=\pi\sqrt{\frac{a^3}{GM}} \qquad (16.5.172)$$

The Hohmann transfer for which the above analysis applies is often used because it is a highly efficient transfer. Because the Hohmann travels through π of true anomaly, it allows the satellite to apply tangential burns which achieve maximum Δv for minimum fuel consumption. However, some missions may wish to consider other transfer variables, such as launch window or transfer time, over Δv. These missions may consider type I or II trajectories. A type I trajectory travels through less than π of true anomaly while a type II trajectory travels through greater than π.

The standard mission to the surface of the Moon is performed by leaving Earth from a parking orbit, performing a velocity maneuver in the parking orbit in order to leave on a direct transfer to the Moon, then performing an additional maneuver to become captured by the Moon. The lowest total burn for this type of direct transfer is a Hohmann transfer from the parking orbit to the Moon's orbit. In order to reduce the total delta-V, the three body problem that creates the Lagrange point solutions also allows for some low energy transfers that reduce the delta-V required to reach the Moon. There are families of low thrust orbits that use small burns to slowly increase an orbit semi major axis until the orbit reaches the Moon. In addition to the low thrust transfer, there are also families of Ballistic Lunar Trajectories (BLT) that involve a path from Earth out towards the Sun-Earth system L1 point where the gravitational pull of the Sun alters the trajectory of the spacecraft causing it to enter a path towards the Moon where it can then be captured. The three different families of transfers are shown below:

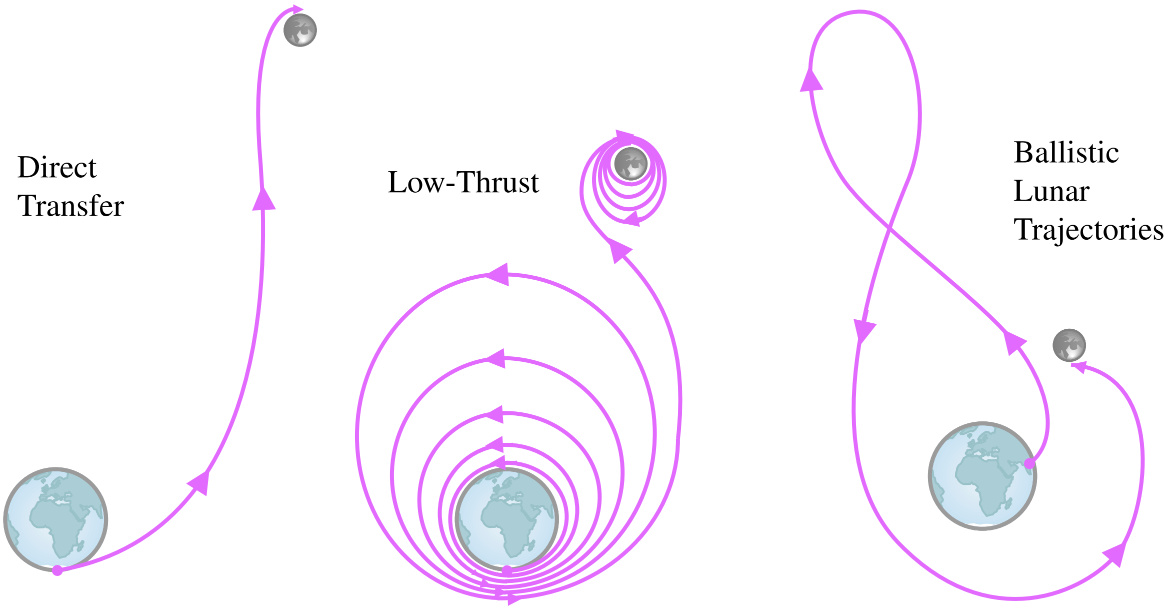


Figure 16.183: **Hohmann transfer variations**

For a burn at velocity v the change in kinetic energy is (assuming the ejected mass for the burn in negligible in comparison of the spacecraft mass):

$$\Delta E?c = \frac{1}{2}m(v+\Delta v)^2 - \frac{1}{2}v^2 = \frac{1}{2}m\Delta v^2 + mv\cdot\Delta v \qquad (16.5.173)$$

Notice that for a fixed Δv, the second term $v\cdot\Delta v$ is much greater when v is large!

Therefore as:

- For an elliptic orbit, maximum velocity is at perigee
- Lower orbits move faster

It is much easier to achieve escape velocity when in low earth orbit!

This is named a "powered flyby" or "Oberth maneuver". Wherein the use of a reaction engine at higher speeds (relative to any reference frame) generates a greater change in mechanical energy than its use at lower speeds is named the "Oberth effect".

~25 %

★★★★★

22 votes, 59.09%

16.6. Software Engineering

SOFTWARE ENGINEERING will be defined in this section as the set of techniques or approaches to solve problems using university-level mathematics and that we find frequently implemented in many IT solutions.

The reader will find in the texts that will follow mathematical tools whose details have all been proved in other chapters and sections, but for whose the computer applications had no place in the theoretical parts of this textbook. We have also chosen to focus only algorithms discussed and used in our trainings that seems to make problems to our students and trainees. The reader must indeed keep in mind that there are so many algorithms that their are jobs around the world and at least as many as their are humans...

Remark

If the reader is looking for mathematical methods to simulate 3D objects on a screen, he will have to refer to the section of Projective Geometry (see page 2679) and if he is looking for Machine Learning/data mining/deep learning techniques he will have to refer to the section of Theoretical Computing (see page 6768).

16.6.1 Algorithm

Definition (#16.6.730): In mathematics and computer science, an "algorithm" is a self-contained step-by-step set of operations to be performed. Algorithms perform calculation, data processing, and/or automated reasoning tasks.

An algorithm is an effective method that can be expressed within a finite amount of space and time and in a well-defined formal language for calculating a function. Starting from an initial state and initial input (perhaps empty), the instructions describe a computation that, when executed, proceeds through a finite number of well-defined successive states, eventually producing "output" and terminating at a final ending state. The transition from one state to the next is not necessarily deterministic; some algorithms, known as randomized algorithms, incorporate random input.

Algorithms can be expressed in many kinds of notation, including natural languages, pseudocode, flowcharts, programming languages or control tables (processed by interpreters). Natural language expressions of algorithms tend to be verbose and ambiguous, and are rarely used for complex or technical algorithms. Pseudocode, flowcharts, drakon-charts and control tables are structured ways to express algorithms that avoid many of the ambiguities common in natural language statements. Programming languages are primarily intended for expressing algorithms in a form that can be executed by a computer, but are often used as a way to define or document algorithms.

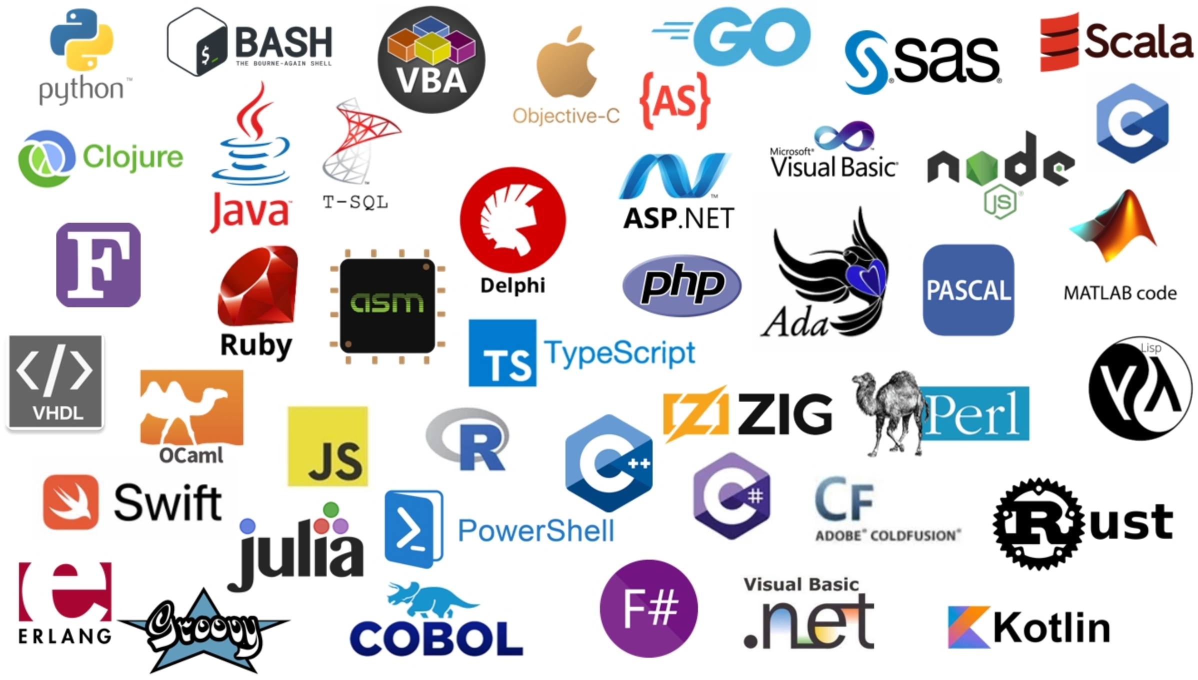

python
BASH
THE BOURNE-AGAIN SHELL
VBA
Objective-C
GO
{AS}
sas
Scala
Clojure
Java
T-SQL
Delphi
ASP.NET
Microsoft Visual Basic
node
JS
Ruby
asm
php
Ada
PASCAL
MATLAB code
VHDL
TS TypeScript
OCaml
JS
R
ZIG
Perl
Lisp
Swift
julia
PowerShell
CF
ADOBE COLDFUSION
Rust
ERLANG
Groovy
COBOL
F#
Visual Basic
.net
Kotlin

Figure 16.184: **Some common programming and script languages in the early 121st century (holocene calendar)**

In computer systems, an algorithm is basically an instance of logic written in software by software developers to be effective for the intended "target" computer(s) to produce output from given input (perhaps null). An optimal algorithm, even running in old hardware, would produce faster results than a non optimal (higher time complexity) algorithm for the same purpose, running in more efficient hardware; that is why the algorithms, like computer hardware, are considered technology.

Caution!!! The reader must remember that a programming, scripting, functional or description language isn't logical! Otherwise all computer languages would have the same syntax for the same purposes (and the reader can easily check on a website like `https://rosettacode.org` that this isn't obviously the case). Each programming language has its own vocabulary, grammatical rules and syntaxic rules that must be learned by heart and cannot be deduced by a hypothetico-deductive reasoning nor guessed! When engineers say that "programming is logic" they refer in reality to the underlying hardware logic based on NAND doors and the application of Boolean logic at the hardware level where the computer program language is executed!

16.6.2 Dichotomic Search algorithm

In computer science, a dichotomic search is a search algorithm that operates by selecting between two distinct alternatives (dichotomies) at each step. It is a specific type of divide and conquer algorithm. A well-known example is "binary search".

Abstractly, a dichotomic search can be viewed as following edges of an implicit binary tree structure until it reaches a leaf (a goal or final state). This creates a theoretical tradeoff between the number of possible states and the running time. We have proved in the section of Theoretical Computing (see page 6288) that the complexity of such an algorithm was $\mathcal{O}(\log_2(n))$

In the case of the search of a root of a function the dichotomic search is equivalent to the bisection method (see section Numerical Methods page 6609) as the x-axes is ordered by construction.

But let us do here a small recall first on the bisection method (see section Numerical Methods page 6609):

16.6.2.1 Bisection algorithm

The bisection method in mathematics is also root-finding method that repeatedly bisects an interval and then selects a subinterval in which a root must lie for further processing. It is a very simple and robust method, but it is also relatively slow. Because of this, it is often used to obtain a rough approximation to a solution which is then used as a starting point for more rapidly converging methods. The method is also named the "interval halving method", the "binary search method", or the "dichotomy method".

The method is applicable for numerically solving the equation $f(x)=0$ for the real variable x, where f is a continuous function defined on an interval $[a,b]$ and where $f(a)$ and $f(b)$ have opposite signs such that $f(a)f(b)<0$. In this case a and b are also said to bracket a root since, by the intermediate value theorem, the continuous function f must have at least one root in the interval $[a,b]$.

At each step the method divides the interval in two by computing the midpoint $c=(a+b)/2$ of the interval and the value of the function $f(c)$ at that point. Unless c is itself a root (which is very unlikely, but possible) there are now only two possibilities: either $f(a)$ and $f(c)$ have opposite signs and bracket a root, or $f(c)$ and $f(b)$ have opposite signs and bracket a root. The method selects the subinterval that is guaranteed to be a bracket as the new interval to be used in the next step. In this way an interval that contains a zero of f is reduced in width by 50% at each step. The process is continued until the interval is sufficiently small.

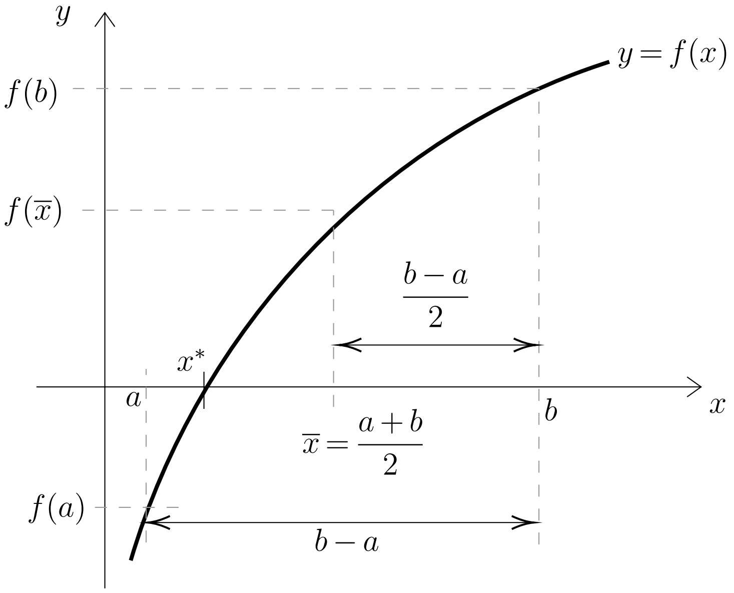


Figure 16.185: **Bisection method scheme**

Explicitly, if $f(a)$ and $f(c)$ have opposite signs, then the method sets c as the new value for b, and if $f(b)$ and $f(c)$ have opposite signs then the method sets c as the new a (if $f(c)=0$ then c may be taken as the solution and the process stops.) In both cases, the new $f(a)$ and $f(b)$ have opposite signs, so the method is applicable to this smaller interval.

The implementation, on a computer, of this method is particularly simple. The conditions to be satisfied being only that in the interval $[a,b]$:

- f must be continuous
- f must be monotone near the root $\bar{x}$
- $f(a)f(b)<0$ to sure that there is a root

The algorithm consists therefore in performing the following steps:

1. We fix $\varepsilon>0$ as upper bound of the admissible error tolerance
2. We calculate $x=(a+b)/2$
3. We evaluation $f(x)$
4. If $|f(x)|<\varepsilon$ then the job is done, we have to display x and $f(x)$
5. Otherwise we proceed as following
    (a) we replace a by x if $f(x)f(a)>0$
    (b) we replace b by x if $f(x)f(b)>0$ or $f(x)f(a)<0$
    (c) we go back in (2)

The previous step (4) imposes the condition for stopping the calculations. Sometimes it is better to choose another criterion calculation ending. It requires the calculated solution to be contained in an interval of length equation containing the root x^*. This test is enunciate as follows:

4'. If $|b-a|<\varepsilon$, the job is finished and $x=(a+b)/2$ is displayed. It is for sure obvious that $|x-x^*|<\varepsilon/2$

In pseudocode (non-unique and not optimized):

Algorithm 20: Proportional Parts bisection pseudocode algorithm

```
Data: $a$,$b$, $\varepsilon$ expression of $f$
Result: $x^*$
initialization;
$x=(a+b)/2$;
while $|f(x)|>\varepsilon$ do
    if $f(x)f(a)>0$ then
        $a:=x$;
    else if $f(x)f(b)>0 \vee f(x)f(a)<0$ then
        $b:=x$;
    $x=(a+b)/2$;
Display $x$, $f(x)$;
```

The equivalent Maple™ 4.00b code is given by:

```
>zero:=proc(f,a,b,pre) local M;
M:=f((a+b)/2);
if abs(M)<pre then
RETURN((a+b)/2)
elif M>0 then
zero(f,a,(a+b)/2,pre)
else zero(f,(a+b)/2,b,pre)
fi
end:
```

16.6.2.2 Binary search algorithm

"Binary search", also known as "half-interval search" or "logarithmic search[18]", is a search algorithm that finds the position of a target value within a sorted array. It compares the target value to the middle element of the array; if they are unequal, the half in which the target cannot lie is eliminated and the search continues on the remaining half until it is successful.

Although the idea is simple, implementing binary search correctly requires attention to some subtleties about its exit conditions and midpoint calculation.

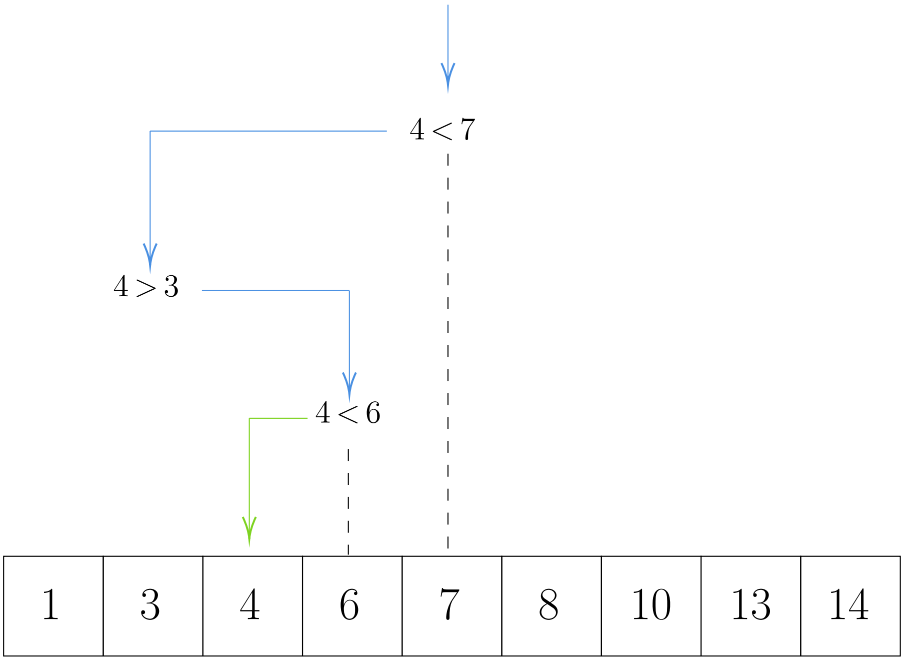


Figure 16.186: **Visualization of the binary search algorithm where 4 is the target value (source: Wikipedia)**

So as we said a binary search works on sorted arrays. A binary search begins by comparing the middle element of the array A with the target value T. If the target value matches the middle element, its position in the array is returned. If the target value is less or more than the middle element, the search continues the lower or upper half of the array respectively with a new middle element, eliminating the other half from consideration.

Here is the pseudocode:

---

[18] because of its complexity that we proved in the section of Theoretical Computing as being of order $\mathcal{O}(\log_2(n))$

Algorithm 21: Binary search pseudocode algorithm

```
Data: Array $A = [A_0, \ldots, A_i, \ldots, A_{n-1}]$, $T$
Result: $i$
initialization;
$L := 0$;
$R := n - 1$;
if $L > R$ then
    Exit ;
    'nothing to find so exit program
$i := \mathrm{E}[(L+R)/2]$;
while $A_i <> T$ do
    if $L > R$ then
        Exit ;
        'found nothing so exit program
    if $A_i < T$ then
        $L := i + 1$;
    else if $A_i > T$ then
        $R := i - 1$;
    $i := \mathrm{E}[(L+R)/2]$;
Display $i$;
```

When the computer scientist and professor Jon Bentley assigned binary search as a problem in a course for professional programmers, he found that 90% failed to provide a correct solution after several hours of working on it, and another study published in 11988 (holocene calendar) shows that accurate code for it is only found in five out of twenty textbooks.

Furthermore, Bentley's own implementation of binary search, published in his 11986 (holocene calendar) book *Programming Pearls*, contained an overflow error that remained undetected for over twenty years. The Java programming language library implementation of binary search had the same overflow bug for more than nine years.

In a practical implementation, the variables used to represent the indices will often be of fixed size, and this can result in an arithmetic overflow for very large arrays. If the midpoint of the span is calculated as $(L+R)/2$, then the value of $L+R$ may exceed the range of integers of the data type used to store the midpoint, even if L and R are within the range. This can be avoided by calculating the midpoint as $L+(R-L)/2$.

16.6.3 Tower of Hanoi algorithm

The Tower of Hanoi is a mathematical childish game or puzzle. It consists of three rods, and a number of disks of different sizes which can slide onto any rod. The puzzle starts with the disks in a neat stack in ascending order of size on one rod, the smallest at the top, thus making a conical shape.

The objective of the puzzle is to move the entire stack to another rod, obeying the following simple rules:

- Only one disk can be moved at a time.
- Each move consists of taking the upper disk from one of the stacks and placing it on top of another stack i.e. a disk can only be moved if it is the uppermost disk on a stack.
- No disk may be placed on top of a smaller disk.

With three disks, the puzzle can be solved in 7 moves. The minimum number of moves required to solve a Tower of Hanoi puzzle is $2^n - 1$, where n is the number of disks.

Many people working in corporations and that never learned programming and wish to do so (typically to learn VBA) don't know before attending a programming course that most of time they don't have the capacity the conceptualize any algorithm. An excellent exercise used by teacher or trainers to exclude people from corporate training class is to ask them before the training as a test to write a general algorithm for the Hanoi tower as it is a very childish algorithm that a normal constituted brain should be able to write after 30 minutes of reflection. The purpose is not to write the most optimal algorithm must just to be able to write one...

Approximately following my experience only $20-30\%$ of people are first able to write the solution in a human language as below:

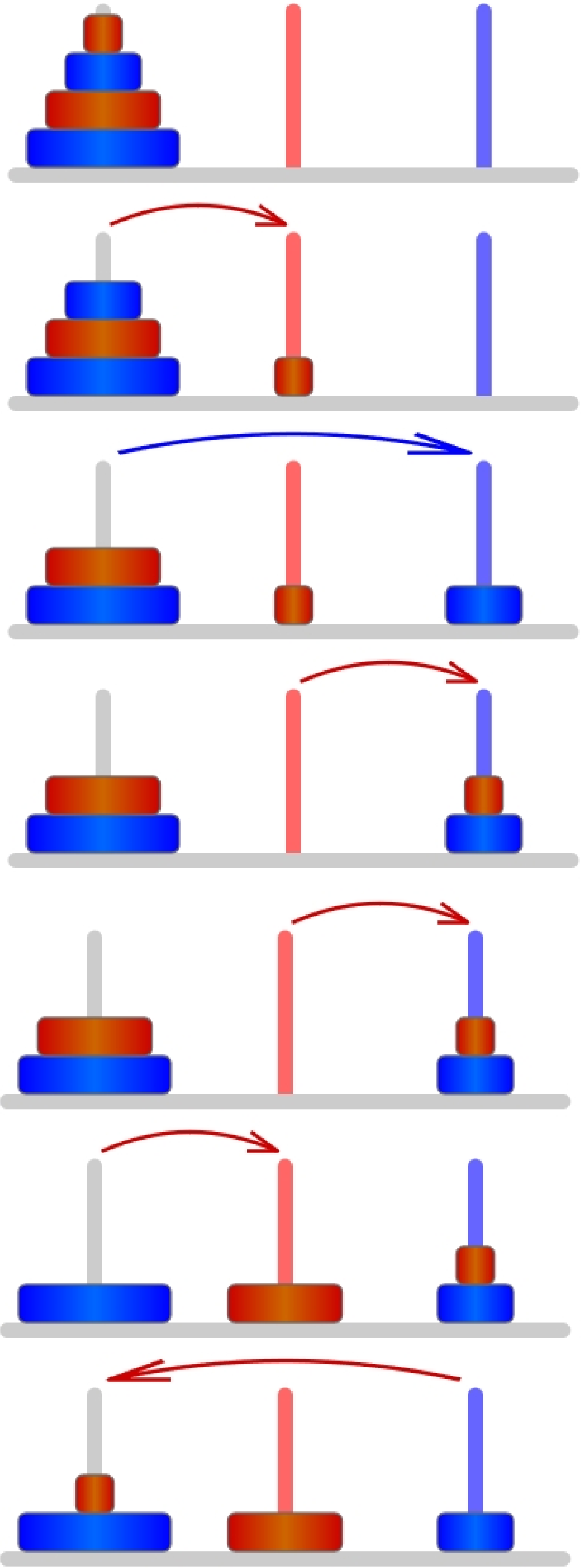
Algorithm 22: Human intuitive pseudocode algorithm of Hanoi tower

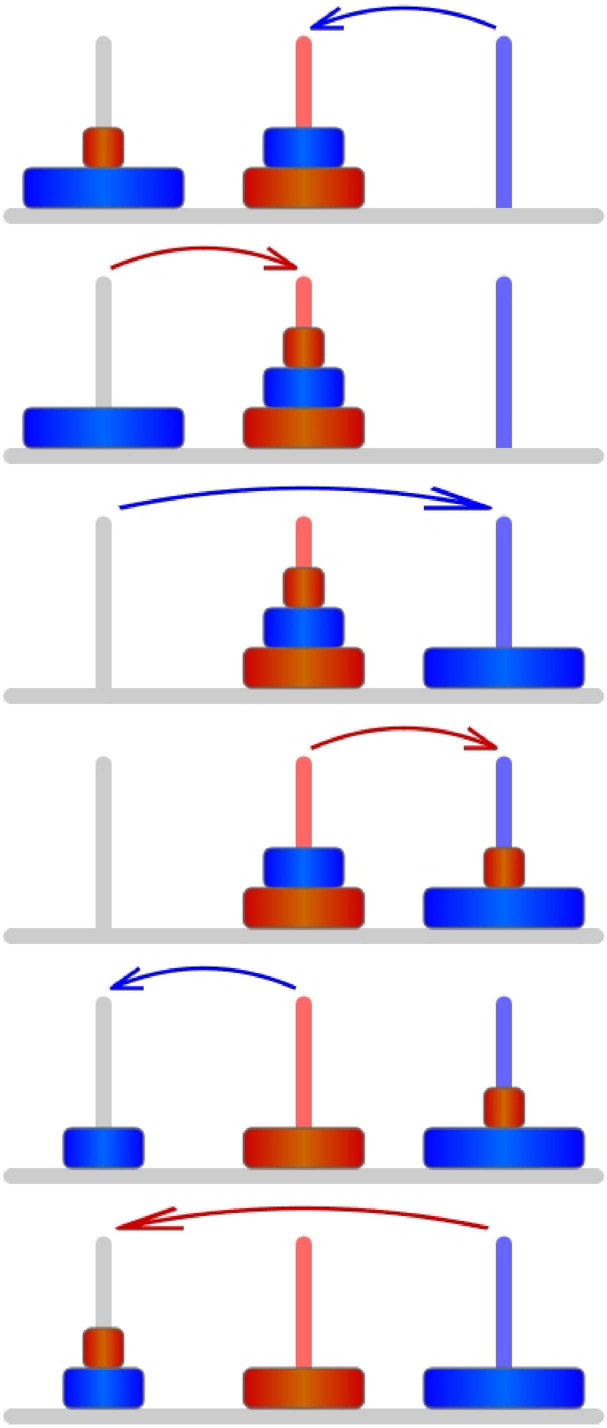
```
Data: $n$ : number of disks
initialization;
if $n$ is even then
    repeat
        make the legal move between pegs A and B;
        make the legal move between pegs A and C;
        make the legal move between pegs B and C;
    until until complete;
if $n$ is odd then
    repeat
        make the legal move between pegs A and C;
        make the legal move between pegs A and B;
        make the legal move between pegs C and B;
    until until complete;
```

Or a little bit better:

```
Data: $n$ : number of disks
initialization;
if $n$ is even then
    repeat
        make the legal move between pegs A and B;
        make the legal move between pegs A and C;
        make the legal move between pegs B and C;
    until until complete;
else
    repeat
        make the legal move between pegs A and C;
        make the legal move between pegs A and B;
        make the legal move between pegs C and B;
    until until complete;
```

That is to say:





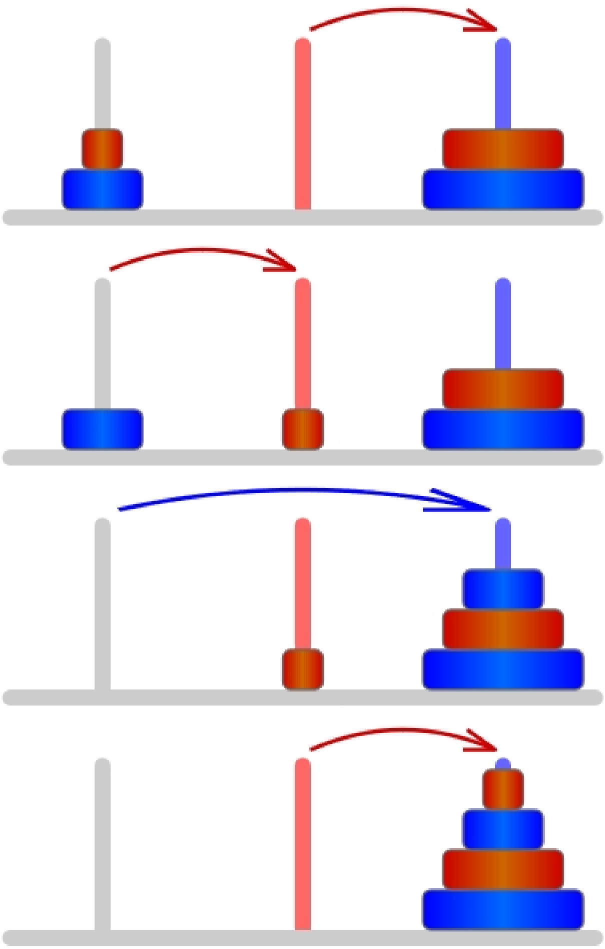


Figure 16.187: **Hanoi Tower $n = 4$ algorithm illustration (source: Wikipedia)**

But this is not a computer oriented pseudocode!! If we request to the students or trainees to write a pseudocode that is "computer oriented" that is to say to use arrays, instead of "pegs"... and use numbers instead of "disks" and write in a mathematical form the "is even" as computers don't know what is an even or odd number, then such a constraint decrease the number of people that successfully pass the test to almost 10% and this shows an inability to have a structured reasoning in the huge majority of the population.

For example to move n disks we will use numbers to identify uniquely each disk (as when first humans invented numbers to identify uniquely items in a collection):

Algorithm 23: Hanoi Tower computer pseudocode algorithm

```
Data: $n$ : number of disks
$A$ = dim([ ],$n$);
$B$ = dim([ ],$n-1$);
$C$ = dim([ ],$n$);
initialization;
for $i=1$ to $n$ do
    $A[i]=(n+1)-i$
if $n>0$ then
    if $n \equiv 0 \mod (2)$ then
        repeat
            if $A[\text{count}[A]] > B[\text{count}[B]]$ then
                $B[\text{count}[B]] = A[\text{count}[A]]$
            else
                $A[\text{count}[A]] = B[\text{count}[B]]$
            if $A[\text{count}[A]] > C[\text{count}[C]]$ then
                $C[\text{count}[C]] = A[\text{count}[A]]$
            else
                $A[\text{count}[A]] = C[\text{count}[C]]$
            if $B[\text{count}[B]] > C[\text{count}[C]]$ then
                $C[\text{count}[C]] = B[\text{count}[B]]$
            else
                $B[\text{count}[B]] = C[\text{count}[C]]$
        until $C$=$A$;
    else
        repeat
            if $A[\text{count}[A]] > C[\text{count}[B]]$ then
                $C[\text{count}[C]] = A[\text{count}[A]]$;
            else
                $A[\text{count}[A]] = C[\text{count}[C]]$;
            if $A[\text{count}[A]] > B[\text{count}[B]]$ then
                $B[\text{count}[B]] = A[\text{count}[A]]$;
            else
                $A[\text{count}[A]] = B[\text{count}[B]]$;
            if $C[\text{count}[C]] > B[\text{count}[B]]$ then
                $B[\text{count}[B]] = C[\text{count}[C]]$;
            else
                $C[\text{count}[C]] = B[\text{count}[B]]$;
        until $C$=$A$;
```

This is obviously not optimal but a minimum we can expect from a human being.

16.6.4 Sorting algorithms

A sorting algorithm is an algorithm that puts elements of a list in a certain order. The most-used orders are numerical order and lexicographical order. Efficient sorting is important for optimizing the use of other algorithms (such as search and merge algorithms) which require input data to be in sorted lists; it is also often useful for canonicalizing data and for producing human-readable output.

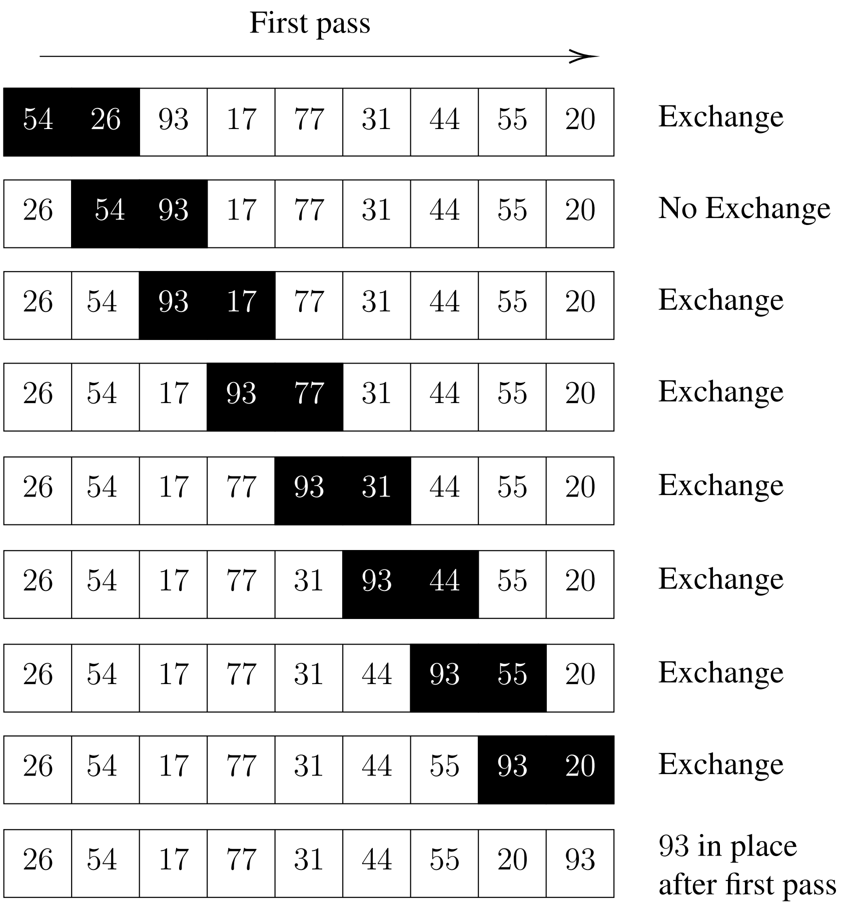
There is a consequent list of sorting algorithms (see Wikipedia for an exhaustive list of almost 20 of them). As already mentioned at the beginning of this section we will focus only on the algorithms used by the redactor of this section when he teach computer science in schools or organizations.

16.6.4.1 Bubble sort

The "Bubble sort", sometimes referred to as "sinking sort", is a simple sorting algorithm that repeatedly steps through the list to be sorted, compares each pair of adjacent items and swaps them if they are in the wrong order. The pass through the list is repeated until no swaps are needed, which indicates that the list is sorted. The algorithm, which is a comparison sort, is named for the way smaller elements "bubble" to the top of the list. Although the algorithm is simple, it is too slow and impractical for most problems.

Bubble sort can operate in-place on an array, requiring small additional amounts of memory to perform the sorting.

So let us see the idea behind the Bubble sort:



54	26	93	17	77	31	44	55	20	Exchange
26	54	93	17	77	31	44	55	20	No Exchange
26	54	93	17	77	31	44	55	20	Exchange
26	54	17	93	77	31	44	55	20	Exchange
26	54	17	77	93	31	44	55	20	Exchange
26	54	17	77	31	93	44	55	20	Exchange
26	54	17	77	31	44	93	55	20	Exchange
26	54	17	77	31	44	55	93	20	Exchange
26	54	17	77	31	44	55	20	93	93 in place after first pass

Figure 16.188: **Schematic principle of the Bubble Sort algorithm**

Here is the pseudocode of this algorithm:

Algorithm 24: Bubble sort pseudocode algorithm

```
Data: $A$ : list of sortable items
Result: $x_1$
initialization;
$n :=$ length($A$);
repeat
    swapped: = false;
    for $i = 1$ to $n - 1$ do
        if $A[i-1] > A[i]$ then
            swap($A[i-1]$,$A[i]$);
            swapped := true;
until not swapped;
display $A$ : list of sorted items;
```

The bubble sort algorithm can be easily optimized by observing that the n-th pass finds the n-th largest element and puts it into its final place. So, the inner loop can avoid looking at the last $n-1$ items when running for the n-th time:

Algorithm 25: Alternative optimized Bubble sort pseudocode algorithm

```
Data: $A$ : list of sortable items
Result: $x_1$
initialization;
$n = \text{length}(A)$;
repeat
    swapped := false;
    for $i = 1$ to $n-1$ do
        if $A[i-1] > A[i]$ then
            swap($A[i-1]$,$A[i]$);
            swapped := true;
    $n := n-1$;
until not swapped;
Display $A$ : list of sorted items;
```

As the bubble sort has two nested loops, the worst-case and average complexity both $\mathcal{O}(n^2)$, where n is the number of items being sorted (obviously when the list is already sorted (best-case), the complexity of bubble sort is only $\mathcal{O}(n^2)$). There exist many sorting algorithms with substantially better worst-case or average complexity of $\mathcal{O}(n\log(n))$. Even other $\mathcal{O}(n^2)$) sorting algorithms, such as insertion sort, tend to have better performance than bubble sort. Therefore, bubble sort is not a practical sorting algorithm when n is large.

16.6.4.2 QuickSort algorithm

"Quicksort" (sometimes named "partition-exchange sort") is an efficient sorting algorithm, serving as a systematic method for placing the elements of an array in order. Developed by Tony Hoare in 11959 (holocene calendar), with his work published in 11961 (holocene calendar), it is still a commonly used algorithm for sorting.

Quicksort can as bubble sort operate "in-place" on an array, requiring small additional amounts of memory to perform the sorting.

Quicksort is a divide and conquer algorithm. Quicksort first divides a large array into two smaller sub-arrays: the low elements and the high elements. Quicksort can then recursively sort the sub-arrays.

The steps are:

1. Pick an element, named a pivot, from the array.
2. Partitioning: reorder the array so that all elements with values less than the pivot come before the pivot, while all elements with values greater than the pivot come after it (equal values can go either way). After this partitioning, the pivot is in its final position. This is named the "partition operation".
3. Recursively apply the above steps to the sub-array of elements with smaller values and separately to the sub-array of elements with greater values.

Here is an illustration of the concept:

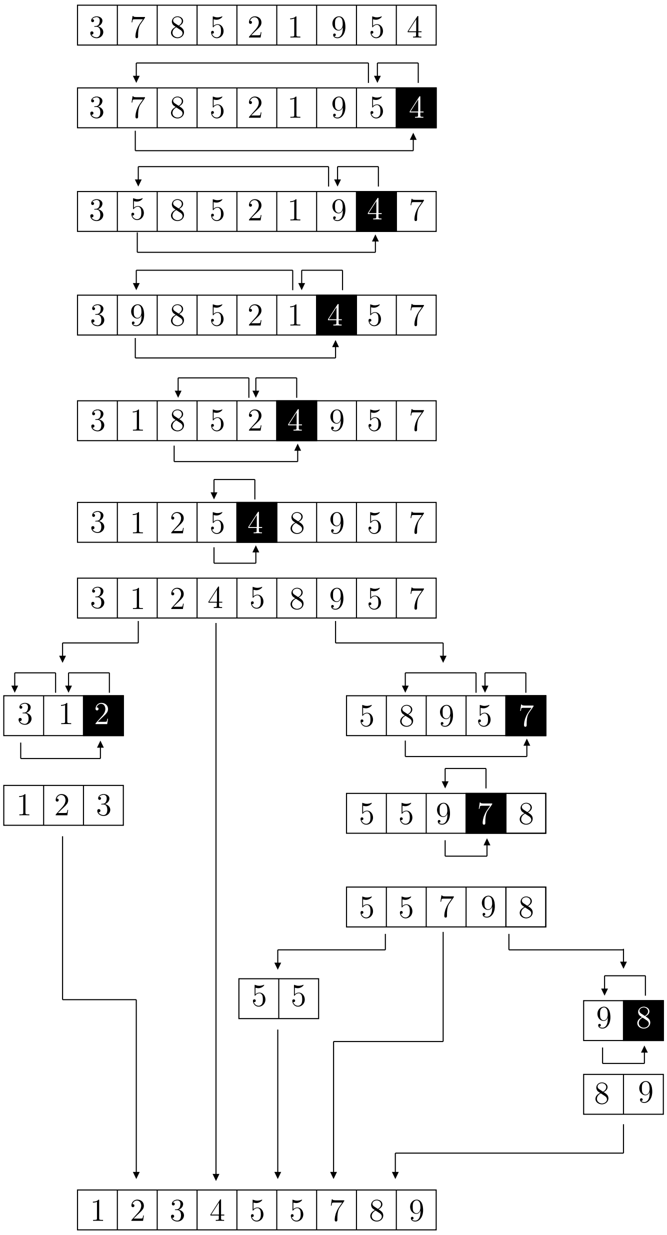


Figure 16.189: **Schematic principle of the QuickSort algorithm**

In pseudocode, a quicksort that sorts elements `lo` through `hi` (inclusive) of an array A can be expressed as (Lomuto partition scheme[19]):

Algorithm 26: QuickSort pseudocode algorithm with function

```
Data: Global $A$ : list of sortable items
Algorithm quicksort($A$, lo, hi)()
    if $lo<hi$ then
        p:=partition($A$, lo, hi)
        quicksort($A$, lo, $p-1$)
        quicksort($A$, $p+1$, hi)
    end
    return
Procedure proc()
    pivot:=A[$hi$]
    for $j=lo$ to $hi-1$ do
        if $A[i-1] \leq pivot$ then
            swap($A[i]$,$A[j]$)
            $i := i+1$
        end
    end
    swap($A[i]$,$A[hi]$)
    return $i$
```

16.6.5 Dijkstra's algorithm

The "Dijkstra's algorithm" is an algorithm for finding the shortest paths between nodes in a graph, which may represent, for example, road networks. It was conceived by computer scientist Edsger W. Dijkstra in 11956 (holocene calendar) and published three years later.

The algorithm exists in many variants; Dijkstra's original variant found the shortest path between two nodes, but a more common variant fixes a single node as the "source" node and finds shortest paths from the source to all other nodes in the graph, producing a shortest-path tree.

For a given source node in the graph, the algorithm finds the shortest path between that node and every other. It can also be used for finding the shortest paths from a single node to a single destination node by stopping the algorithm once the shortest path to the destination node has been determined. For example, if the nodes of the graph represent cities and edge path costs represent driving distances between pairs of cities connected by a direct road, Dijkstra's algorithm can be used to find the shortest route between one city and all other cities. As a result, the shortest path algorithm is widely used in network routing protocols, most notably IS-IS and Open Shortest Path First (OSPF). It is also employed as a subroutine in other algorithms.

So how does a Dijkstra Algorithm work? We will explain this using four simple steps:

1. Assign to every node a tentative distance value: set it to zero for our initial node and to

---

[19] This scheme is attributed to Nico Lomuto and popularized by Bentley in his book *Programming Pearls* and Cormen et al. in their book *Introduction to Algorithms*.

infinity for all other nodes.

2. Set the initial node as current. Mark all other nodes unvisited. Create a set of all the unvisited nodes named the "unvisited set".
3. For the current node, consider all of its neighbours and calculate their tentative distances. Compare the newly calculated tentative distance to the current assigned value and assign the smaller one.
4. When we are done considering all of the neighbours of the current node, mark the current node as visited and remove it from the unvisited set. A visited node will never be checked again.
5. If the destination node has been marked visited (when planning a route between two specific nodes) or if the smallest tentative distance among the nodes in the unvisited set is infinity (when planning a complete traversal; occurs when there is no connection between the initial node and remaining unvisited nodes), then stop. The algorithm has finished.
6. Otherwise, select the unvisited node that is marked with the smallest tentative distance, set it as the new "current node", and go back to step 3.

Suppose you would like to find the shortest path between two intersections on a city map: a starting point and a destination. Dijkstra's algorithm initially marks the distance (from the starting point) to every other intersection on the map with infinity. This is done not to imply there is an infinite distance, but to note that those intersections have not yet been visited; some variants of this method simply leave the intersections' distances unlabelled. Now, at each iteration, select the current intersection. For the first iteration, the current intersection will be the starting point, and the distance to it (the intersection's label) will be zero. For subsequent iterations (after the first), the current intersection will be a closest unvisited intersection to the starting point (this will be easy to find).

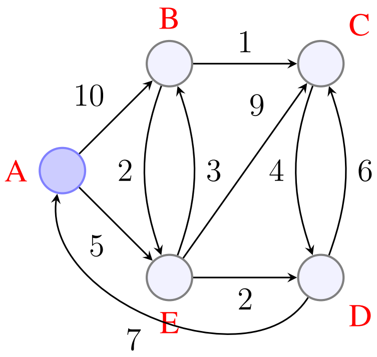
From the current intersection, update the distance to every unvisited intersection that is directly connected to it. This is done by determining the sum of the distance between an unvisited intersection and the value of the current intersection, and relabelling the unvisited intersection with this value (the sum), if it is less than its current value. In effect, the intersection is relabelled if the path to it through the current intersection is shorter than the previously known paths. To facilitate shortest path identification, in pencil, mark the road with an arrow pointing to the relabelled intersection if you label/relabel it, and erase all others pointing to it. After you have updated the distances to each neighbouring intersection, mark the current intersection as visited, and select an unvisited intersection with minimal distance (from the starting point) – or the lowest label—as the current intersection. Nodes marked as visited are labelled with the shortest path from the starting point to it and will not be revisited or returned to.

Continue this process of updating the neighbouring intersections with the shortest distances, then marking the current intersection as visited and moving onto a closest unvisited intersection until you have marked the destination as visited. Once you have marked the destination as visited (as is the case with any visited intersection) you have determined the shortest path to it, from the starting point, and can trace your way back, following the arrows in reverse; in the algorithm's implementations, this is usually done (after the algorithm has reached the destination node) by following the nodes' parents from the destination node up to the starting node; that's why we also keep track of each node's parent.

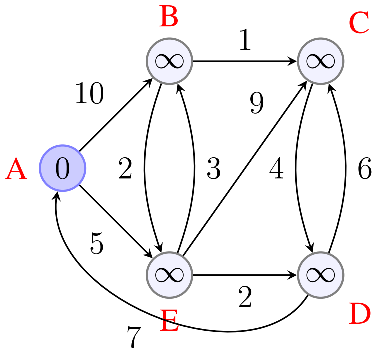
This algorithm makes no attempt to direct "exploration" towards the destination as one might expect. Rather, the sole consideration in determining the next "current" intersection is its distance from the starting point. This algorithm therefore expands outward from the starting point, interactively considering every node that is closer in terms of shortest path distance until it reaches the destination. When understood in this way, it is clear how the algorithm necessarily finds the shortest path. However, it may also reveal one of the algorithm's weaknesses: its relative slowness in some topologies.

Now aim of this is to make the Dijkstra algorithm work on a small concrete example!

Let us look for the shortest paths of origin A in this graph to node C:



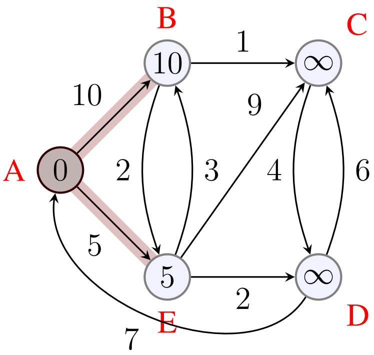
We place ourselves on A and we put the infinites values:



We initialize the following table according to the algorithm:

A	B	C	D	E
$\boxed{0}$	∞	∞	∞	∞
•				
•				
•				
•				
•				

Each of the edges starting from the selected nodes are studied:

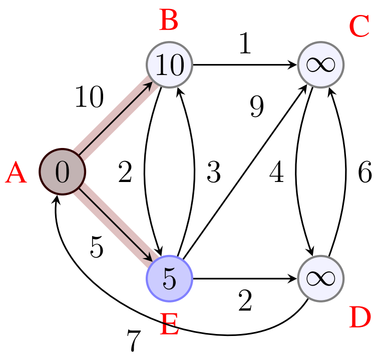


We report the distance to the adjacent node (keeping also the source as we must never use it anymore later!):

A	B	C	D	E
$\boxed{0}$	∞	∞	∞	∞
•	10_A	∞	∞	5_A
•				
•				
•				
•				

In the columns, we put the distance to A, and the node from which we come from.

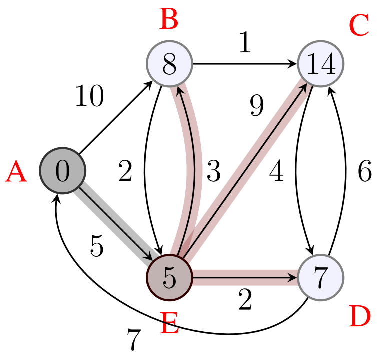
We put ourselves on the node with the smallest weight, here E (the corresponding node color has been darken in the graph below to put it in evidence):


B
C
A
D
E
0
10
∞
5
∞
10
1
9
2
3
4
6
5
2
7

And we highlight in the table our new starting point having the minimal weight:

A	B	C	D	E
0	∞	∞	∞	∞
●	10_A	∞	∞	$\boxed{5_A}$
●				●
●				●
●				●
●				●

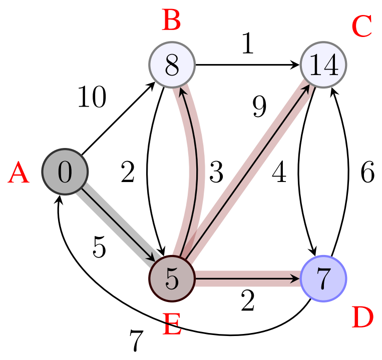
And we iterate again the procedure:


B
C
A
D
E
0
8
14
5
7
10
1
9
2
3
4
6
5
2
7

And we get:

A	B	C	D	E
0	∞	∞	∞	∞
●	10_A	∞	∞	$\boxed{5_A}$
●	8_E	14_E	7_E	●
●				●
●				●

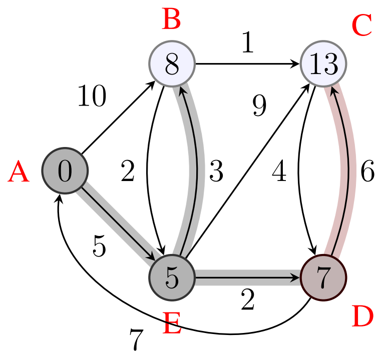
And we put ourselves on node D (highlighted in the figure below):


B
C
8
1
14
10
9
A
0
2
3
4
6
5
5
7
2
E
D
7

So we put the new starting point and corresponding node in evidence in the table:

A	B	C	D	E
0	∞	∞	∞	∞
•	10_A	∞	∞	5_A
•	8_E	14_E	$\boxed{7_E}$	•
•			•	•
•			•	•

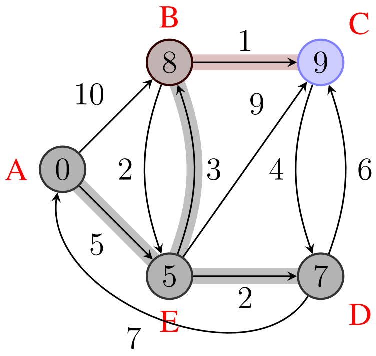
Now starting from node D it remains only one possible path:


B
C
8
1
13
10
9
A
0
2
3
4
6
5
5
7
2
E
D
7

A	B	C	D	E
0	∞	∞	∞	∞
•	10_A	∞	∞	5_A
•	8_E	14_E	$\boxed{7_E}$	•
•	8_E	13_D	•	•
•			•	•
•			•	•

So we have arrived to our destination! Now we go backward to check if we really have the shortest path. Starting from C we see that by going backward the shortest path is to go through B that itself comes from E:

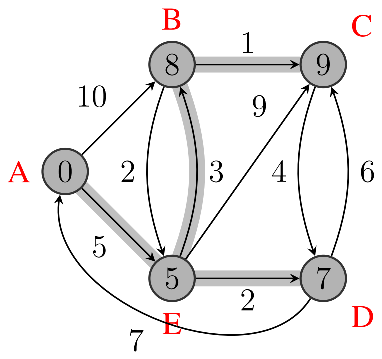
A	B	C	D	E
0	∞	∞	∞	∞
•	10_A	∞	∞	5_A
•	8_E	14_E	7_E	•
•	$\boxed{8_E}$	13_D	•	•
•	•		•	•
•	•		•	•



And as 8_E joins our initial forward path, we stop here and therefore the shortest path has a weight equal to 9.

A	B	C	D	E
0	∞	∞	∞	∞
•	10_A	∞	∞	5_A
•	8_E	14_E	7_E	•
•	8_E	13_D	•	•
•	•	$\boxed{9_B}$	•	•
•	•	•	•	•

Hence the final path:



Ok the example may help to understand the human way... but what about the computer implementation that may seem quite not obvious (indeed... it's not obvious!). Here is the corresponding pseudocode for a computer implementation:

```
function Dijkstra(Graph, source):
    for each vertex v in Graph: // Initializations
        dist[v] := infinity
    // Unknown distance function from source to v
        previous[v] := undefined
    // Previous node in optimal path from source
    dist[source] := 0
    // Distance from source to source
    Q := the set of all nodes in Graph
    // All nodes in the graph are unoptimized - thus are in Q
    while Q is not empty:       // The main loop
        u := vertex in Q with smallest dist[]
        if dist[u] = infinity:
            break
    // all remaining vertices are inaccessible
        remove u from Q
        for each neighbor v of u:
    // where v has not yet been removed from Q.
            alt := dist[u] + dist_between(u, v)
            if alt < dist[v]:    // Relax (u,v,a)
                dist[v] := alt
                previous[v] := u
    return previous[]
```

As we can see above, this Dijkstra simple algorithm has a complexity of order $\mathcal{O}(v^2)$ as two loops depending on v (the number of vertices) are inside each other. Some algorithms are faster using various empirical techniques (adjacency lists or quantum computing).

16.6.6 Google PageRank algorithm

At its creation in 11998 (holocene calendar), Google (search engine currently well known by many Internet users) has dominated in 12011 (holocene calendar) and still dominates the market for Internet search engines. His initial strength is based on the simple idea to sort the results by relevance, what did not make the other major search engines before it.

The main idea behind the initial Google algorithm is interesting to present in this textbook because it is a practical example of purely mathematical concepts we have presented so far (graph theory and Markov chains in the respective sections of Graph Theory and Probability) and created one of the largest international companies in new technologies in the years 12000's according to holocene calendar (Google earns billions thanks to linear algebra ...) and even since the beginning of worldwide economy! As it is often the case in a market economy, it is primarily theoretical research (mathematical or physics) that helped to develop new tools!

Remarks

R1. Obviously the basic algorithm has changed since 11998 (holocene calendar) because you can add many mathematical parameters (weighting by a distance for example) or empirical (visits, clicks, advertising, design, etc.).

R2. This algorithm is not only used in the field of Internet. Indeed, it can be used to highlight any object in a multiple relation with other objects (e.g. variables in a project that influence each other by having direct or indirect relations: a structural analysis) or also to determiner influencers in a social network.

To find an information in the amorphous and not really structured content that is the Internet (given the few developers who properly comply with the standards set by the W3C...), the user will search by keywords. This obviously requires for sure some preparation to be effective: the search engine copy first in the local memory all web pages and sorts extracted words (or combination of words) in alphabetical order using traditional empirical algorithms index of the IT domain (b trees, b+ trees or even others...). The result is a directory of keywords with their associated web pages.

To analyse how Google identifies the most relevant web pages, we will ignore the content of pages and only count the links between them. What we get then is the structure of a graph (not necessarily connected!).

The following image shows a miniature example (there would $25\cdot 10^9$ websites from Google in the early 12007 according to holocene calendar):

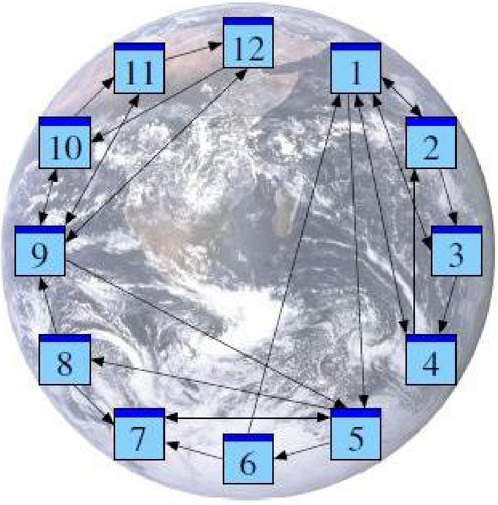


Figure 16.190: Schematic principle of the graph to study

For what follow, we will denote the web pages by:

$$P_1, P_2, \ldots, P_n \tag{16.6.1}$$

and we will write:

$$j \longrightarrow i \tag{16.6.2}$$

if the page P_j directly quotes the page P_i (directly related citation or link). Thus, in the previous image, we have a link $1 \longrightarrow 5$, for example, but no link $5 \longrightarrow 1$.

Obviously, a link $j \longrightarrow i$ is a recommendation of the page P_j to go read the page P_i. Thus it is a vote of P_j in favour of the of the page P_i (we then understand easily then easily why exchanging links between two websites with no contextual relation was somehow banned by Google).

Let us analyse the previous image in a more consistent aspect of graph theory (see corresponding section in this textbook), which hierarchy will have to be justified to be as following:

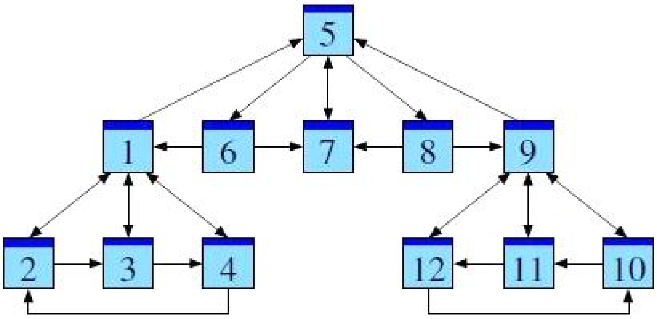


Figure 16.191: Preceding figure into more mathematical aspect

For purely educational reasons, we will introduce three theoretical models of ranking web pages. The latest model is say to be the closest to that implemented by Google in 11998 (holocene calendar) and the first two models were just here to prepare the basis.

16.6.6.1 Weighted Count

It is highly possible that if a page is important, it gets a lot of links. With a bit of naivety, we can assume that the reverse is true (that is to say: if a page gets a lot of links, it means it is important). Thus, we could define the importance μ_i of a page P_i as the number of directly related links $j \rightarrow i$. In mathematical form, we write it in this way: In other words, μ_i is equal to the number of votes for the page P_i, where each vote contributes for the same value 1. This is certainly easy to calculate, but often do not match with the importance felt by the user.

Therefore in our above example we have:

$$\mu_1 = \mu_9 = 4 \tag{16.6.3}$$

that is in front of (in terms of ranking):

$$\mu_5 = \mu_7 = 3 \tag{16.6.4}$$

what is worse, is that this naive count can be too easily manipulated by adding pages (another Internet domain name for example) without interest recommending any other page.

> **Remark**
>
> So the maximum weight is verbatim equal to $n-1$.

Some pages emit a lot of (hyper)links. They therefore seem less specific (specialized) and their weight will (must) therefore be lower. We therefore share the vote of the page P_j in l_j equal parts, where l_j denotes the number of emitted links. So we can define a finer measurement:

$$\mu_i = \sum_{j \rightarrow i} \frac{1}{l_j} \tag{16.6.5}$$

In other words, μ_i counts now the number of weighted votes for the page P_i. It is easy to calculate, but we always still remain with the above problem.

In our example, we have therefore:

$$\begin{aligned}
\mu_1 &= \sum_{j \rightarrow 1} \frac{1}{l_j} = \frac{1}{l_2} + \frac{1}{l_3} + \frac{1}{l_4} + \frac{1}{l_6} = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = 2 \\
\mu_9 &= \sum_{j \rightarrow 9} \frac{1}{l_j} = \frac{1}{l_8} + \frac{1}{l_{10}} + \frac{1}{l_{11}} + \frac{1}{l_{12}} = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = 2
\end{aligned} \tag{16.6.6}$$

and:

$$\mu_5 = \sum_{j \rightarrow 5} \frac{1}{l_j} = \frac{1}{l_1} + \frac{1}{l_7} + \frac{1}{l_9} = \frac{1}{4} + \frac{1}{1} + \frac{1}{4} = 1.5 \tag{16.6.7}$$

and:

$$\mu_7 = \sum_{j \rightarrow 7} \frac{1}{l_j} = \frac{1}{l_5} + \frac{1}{l_6} + \frac{1}{l_7} = \frac{1}{3} + \frac{1}{2} + \frac{1}{2} = \frac{4}{3} \tag{16.6.8}$$

Remark

If all the pages were directly related (conner), then we would have $\mu_i = 1$ for all i.

16.6.6.2 Recursive counting

Heuristically, a page P_1 seems important if many important pages cite it. This leads us to define the μ_i recursively as follows:

$$\mu_i = \sum_{j \to} \frac{1}{l_j} \mu_j \tag{16.6.9}$$

It is therefore a system of homogeneous linear equations with n unknowns, where in the chosen example $n = 12$. More traditionally, the system may be denoted:

$$\left(\sum_{j \to i} \frac{1}{l_j} \mu_j \right) - \mu_i = 0 \tag{16.6.10}$$

or more explicitly (see section Linear Algebra page 1924):

$$\begin{bmatrix}
-1 & l_2^{-1} & l_3^{-1} & l_4^{-1} & l_5^{-1} & l_6^{-1} & l_7^{-1} & l_8^{-1} & l_9^{-1} & l_{10}^{-1} & l_{11}^{-1} & l_{12}^{-1} \\
l_1^{-1} & -1 & l_3^{-1} & l_4^{-1} & l_5^{-1} & l_6^{-1} & l_7^{-1} & l_8^{-1} & l_9^{-1} & l_{10}^{-1} & l_{11}^{-1} & l_{12}^{-1} \\
l_1^{-1} & l_2^{-1} & -1 & l_4^{-1} & l_5^{-1} & l_6^{-1} & l_7^{-1} & l_8^{-1} & l_9^{-1} & l_{10}^{-1} & l_{11}^{-1} & l_{12}^{-1} \\
l_1^{-1} & l_2^{-1} & l_3^{-1} & -1 & l_5^{-1} & l_6^{-1} & l_7^{-1} & l_8^{-1} & l_9^{-1} & l_{10}^{-1} & l_{11}^{-1} & l_{12}^{-1} \\
\vdots & \ddots & & & -1 & & & & & & & \vdots \\
 & & & & & -1 & & & & & & \\
 & & & & & & -1 & & & & & \\
\vdots & & & & & & & -1 & & & \ddots & \vdots \\
l_1^{-1} & l_2^{-1} & l_3^{-1} & l_4^{-1} & l_5^{-1} & l_6^{-1} & l_7^{-1} & l_8^{-1} & -1 & l_{10}^{-1} & l_{11}^{-1} & l_{12}^{-1} \\
l_1^{-1} & l_2^{-1} & l_3^{-1} & l_4^{-1} & l_5^{-1} & l_6^{-1} & l_7^{-1} & l_8^{-1} & l_9^{-1} & -1 & l_{11}^{-1} & l_{12}^{-1} \\
l_1^{-1} & l_2^{-1} & l_3^{-1} & l_4^{-1} & l_5^{-1} & l_6^{-1} & l_7^{-1} & l_8^{-1} & l_9^{-1} & l_{10}^{-1} & -1 & l_{12}^{-1} \\
l_1^{-1} & l_2^{-1} & l_3^{-1} & l_4^{-1} & l_5^{-1} & l_6^{-1} & l_7^{-1} & l_8^{-1} & l_9^{-1} & l_{10}^{-1} & l_{11}^{-1} & 1
\end{bmatrix}
=
\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_4 \\ \mu_4 \\ \mu_5 \\ \mu_6 \\ \mu_7 \\ \mu_8 \\ \mu_9 \\ \mu_{10} \\ \mu_{11} \\ \mu_{12} \end{pmatrix}
\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}
\tag{16.6.11}$$

Which gives with our example:

$$\begin{bmatrix} -1 & 2^{-1} & 2^{-1} & 2^{-1} & 0 & 2^{-1} & 0 & 0 & 0 & 0 & 0 & 0 \\ 4^{-1} & -1 & 0 & 2^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4^{-1} & 2^{-1} & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4^{-1} & 0 & 2^{-1} & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4^{-1} & 0 & 0 & 0 & -1 & 0 & 1^{-1} & 0 & 4^{-1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3^{-1} & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3^{-1} & 2^{-1} & -1 & 2^{-1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3^{-1} & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2^{-1} & -1 & 2^{-1} & 2^{-1} & 2^{-1} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4^{-1} & -1 & 0 & 2^{-1} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4^{-1} & 2^{-1} & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4^{-1} & 0 & 2^{-1} & -1 \end{bmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_4 \\ \mu_4 \\ \mu_5 \\ \mu_6 \\ \mu_7 \\ \mu_8 \\ \mu_9 \\ \mu_{10} \\ \mu_{11} \\ \mu_{12} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{16.6.12}$$

In addition to the trivial solution (null vector), this system has an infinity of solutions because its determinant is zero (see section Linear Algebra page 1980), which can easily be verified with the `MDETERM( )` function of Microsoft® Excel® 11.8346.

Let us now write the previous system in the following equivalent form:

$$\begin{bmatrix} 0 & 2^{-1} & 2^{-1} & 2^{-1} & 0 & 2^{-1} & 0 & 0 & 0 & 0 & 0 & 0 \\ 4^{-1} & 0 & 0 & 2^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4^{-1} & 2^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4^{-1} & 0 & 2^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4^{-1} & 0 & 0 & 0 & 0 & 0 & 1^{-1} & 0 & 4^{-1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3^{-1} & 2^{-1} & 0 & 2^{-1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2^{-1} & 0 & 2^{-1} & 2^{-1} & 2^{-1} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4^{-1} & 0 & 0 & 2^{-1} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4^{-1} & 2^{-1} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4^{-1} & 0 & 2^{-1} & 0 \end{bmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_4 \\ \mu_4 \\ \mu_5 \\ \mu_6 \\ \mu_7 \\ \mu_8 \\ \mu_9 \\ \mu_{10} \\ \mu_{11} \\ \mu_{12} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_4 \\ \mu_4 \\ \mu_5 \\ \mu_6 \\ \mu_7 \\ \mu_8 \\ \mu_9 \\ \mu_{10} \\ \mu_{11} \\ \mu_{12} \end{pmatrix} \tag{16.6.13}$$

We notice two things before continuing:

1. In the above matrix, each column is such that the sum of its values equals unity. It is therefore the transpose of a stochastic matrix of a Markov chain (see section Probabilities page 537).

2. The importance μ_i is positive or zero. So the vector $\vec{\mu}$ is strictly positive and it is the stochastic vector of the Markov chain (see section Probabilities page 537) whose sum of the components must be equal to unity.

We will write this system in condensed form:

$$P^T\mu=\mu \tag{16.6.14}$$

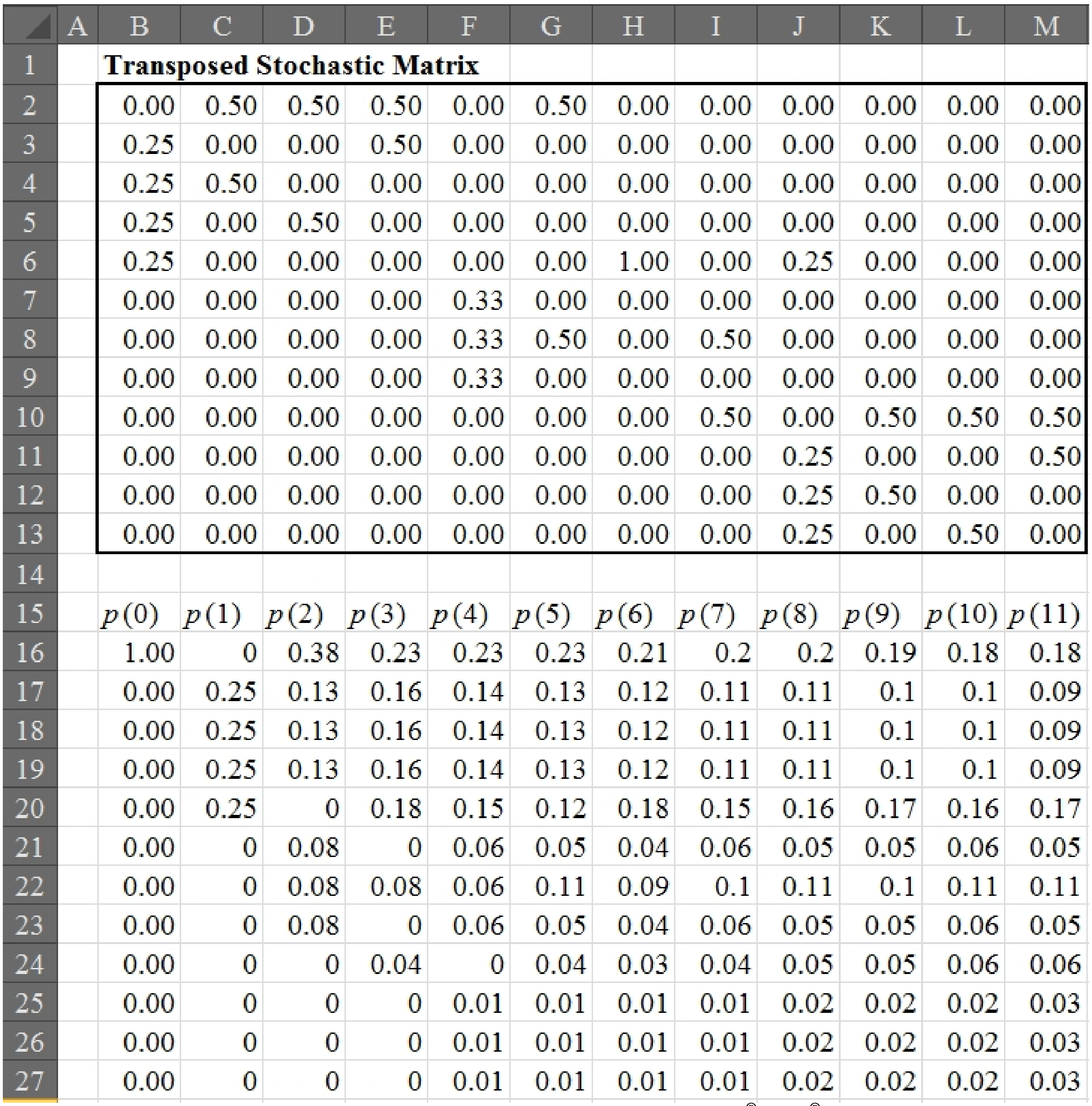
Thus, 1 is the eigenvalue and μ is the eigenvector of the application P. We know from our study of Markov chains that this relation is verified only for the invariant measure denoted traditionally π in the domain of study of Markov chains:

$$P^T\pi=\pi \tag{16.6.15}$$

To determine the measure, it is necessary to first rewrite the network in the format of a matrix application of a Markov chain (see section Probabilities page 537):

$$p(n+1)=P^Tp(n) \tag{16.6.16}$$

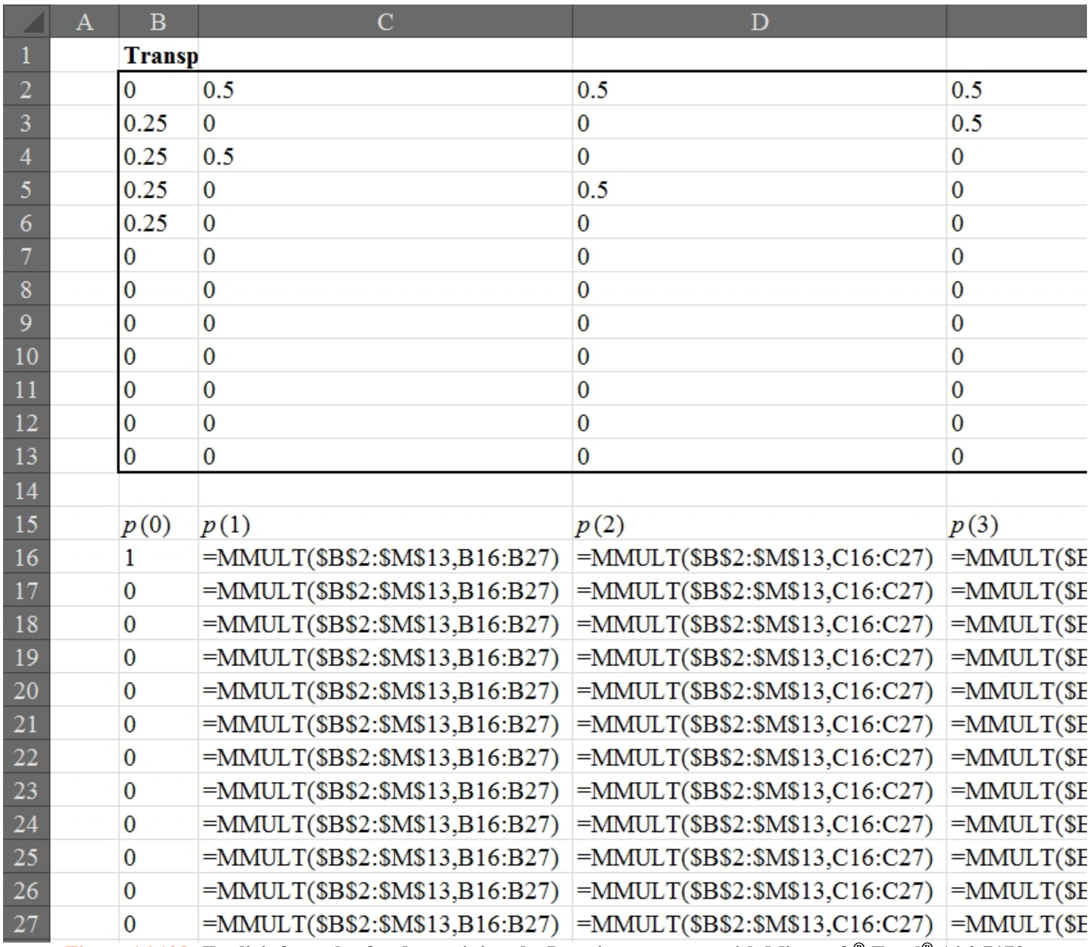
With Microsoft® Excel® 11.8346, the modelling is quite simple enough to reproduce to determine the invariant measure with any starting point (any distribution) in the graph:



	A	B	C	D	E	F	G	H	I	J	K	L	M
1		**Transposed Stochastic Matrix**											
2		0.00	0.50	0.50	0.50	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00
3		0.25	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4		0.25	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5		0.25	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6		0.25	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.25	0.00	0.00	0.00
7		0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8		0.00	0.00	0.00	0.00	0.33	0.50	0.00	0.50	0.00	0.00	0.00	0.00
9		0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.50	0.50	0.50
11		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.50
12		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.50	0.00	0.00
13		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.50	0.00
14													
15		$p(0)$	$p(1)$	$p(2)$	$p(3)$	$p(4)$	$p(5)$	$p(6)$	$p(7)$	$p(8)$	$p(9)$	$p(10)$	$p(11)$
16		1.00	0	0.38	0.23	0.23	0.23	0.21	0.2	0.2	0.19	0.18	0.18
17		0.00	0.25	0.13	0.16	0.14	0.13	0.12	0.11	0.11	0.1	0.1	0.09
18		0.00	0.25	0.13	0.16	0.14	0.13	0.12	0.11	0.11	0.1	0.1	0.09
19		0.00	0.25	0.13	0.16	0.14	0.13	0.12	0.11	0.11	0.1	0.1	0.09
20		0.00	0.25	0	0.18	0.15	0.12	0.18	0.15	0.16	0.17	0.16	0.17
21		0.00	0	0.08	0	0.06	0.05	0.04	0.06	0.05	0.05	0.06	0.05
22		0.00	0	0.08	0.08	0.06	0.11	0.09	0.1	0.11	0.1	0.11	0.11
23		0.00	0	0.08	0	0.06	0.05	0.04	0.06	0.05	0.05	0.06	0.05
24		0.00	0	0	0.04	0	0.04	0.03	0.04	0.05	0.05	0.06	0.06
25		0.00	0	0	0	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.03
26		0.00	0	0	0	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.03
27		0.00	0	0	0	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.03

Figure 16.192: **Determining the Invariant Measure with Microsoft® Excel® 14.0.7173**

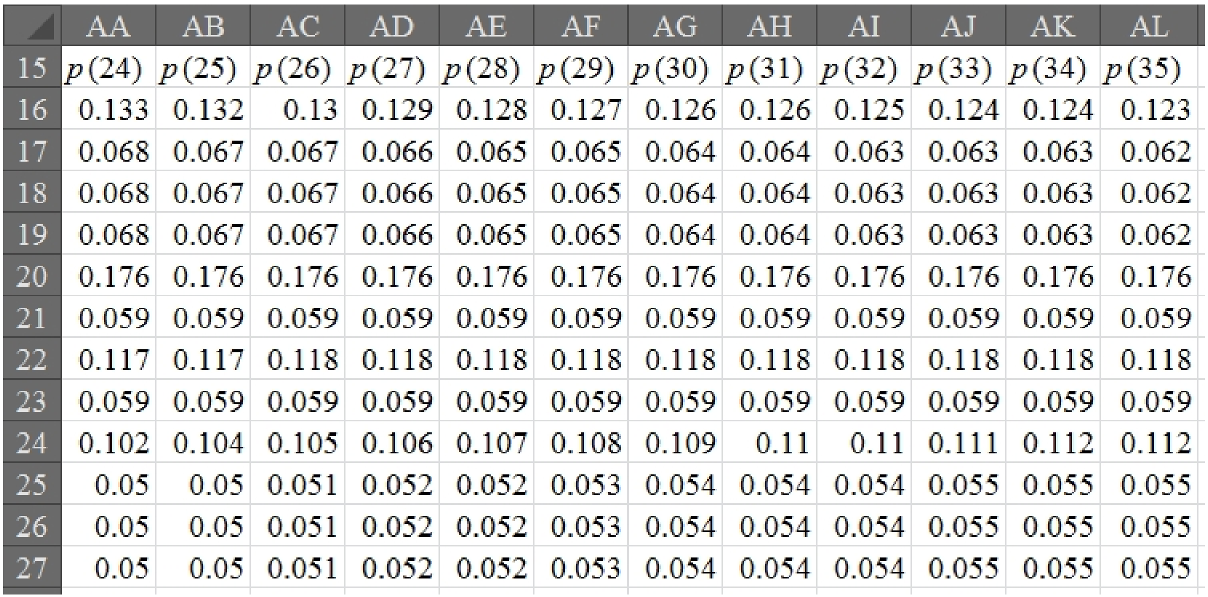
with the following explicit formulas:



	A	B	C	D	
1		**Transp**			
2		0	0.5	0.5	0.5
3		0.25	0	0	0.5
4		0.25	0.5	0	0
5		0.25	0	0.5	0
6		0.25	0	0	0
7		0	0	0	0
8		0	0	0	0
9		0	0	0	0
10		0	0	0	0
11		0	0	0	0
12		0	0	0	0
13		0	0	0	0
14					
15		$p(0)$	$p(1)$	$p(2)$	$p(3)$
16		1	=MMULT($B$2:$M$13,B16:B27)	=MMULT($B$2:$M$13,C16:C27)	=MMULT($B
17		0	=MMULT($B$2:$M$13,B16:B27)	=MMULT($B$2:$M$13,C16:C27)	=MMULT($B
18		0	=MMULT($B$2:$M$13,B16:B27)	=MMULT($B$2:$M$13,C16:C27)	=MMULT($B
19		0	=MMULT($B$2:$M$13,B16:B27)	=MMULT($B$2:$M$13,C16:C27)	=MMULT($B
20		0	=MMULT($B$2:$M$13,B16:B27)	=MMULT($B$2:$M$13,C16:C27)	=MMULT($B
21		0	=MMULT($B$2:$M$13,B16:B27)	=MMULT($B$2:$M$13,C16:C27)	=MMULT($B
22		0	=MMULT($B$2:$M$13,B16:B27)	=MMULT($B$2:$M$13,C16:C27)	=MMULT($B
23		0	=MMULT($B$2:$M$13,B16:B27)	=MMULT($B$2:$M$13,C16:C27)	=MMULT($B
24		0	=MMULT($B$2:$M$13,B16:B27)	=MMULT($B$2:$M$13,C16:C27)	=MMULT($B
25		0	=MMULT($B$2:$M$13,B16:B27)	=MMULT($B$2:$M$13,C16:C27)	=MMULT($B
26		0	=MMULT($B$2:$M$13,B16:B27)	=MMULT($B$2:$M$13,C16:C27)	=MMULT($B
27		0	=MMULT($B$2:$M$13,B16:B27)	=MMULT($B$2:$M$13,C16:C27)	=MMULT($B

Figure 16.193: **Explicit formulas for determining the Invariant measure with Microsoft® Excel® 14.0.7173**

The system converges fairly quickly, after the 30th step, we already have a convergence to the second decimal:



	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL
15	$p(24)$	$p(25)$	$p(26)$	$p(27)$	$p(28)$	$p(29)$	$p(30)$	$p(31)$	$p(32)$	$p(33)$	$p(34)$	$p(35)$
16	0.133	0.132	0.13	0.129	0.128	0.127	0.126	0.126	0.125	0.124	0.124	0.123
17	0.068	0.067	0.067	0.066	0.065	0.065	0.064	0.064	0.063	0.063	0.063	0.062
18	0.068	0.067	0.067	0.066	0.065	0.065	0.064	0.064	0.063	0.063	0.063	0.062
19	0.068	0.067	0.067	0.066	0.065	0.065	0.064	0.064	0.063	0.063	0.063	0.062
20	0.176	0.176	0.176	0.176	0.176	0.176	0.176	0.176	0.176	0.176	0.176	0.176
21	0.059	0.059	0.059	0.059	0.059	0.059	0.059	0.059	0.059	0.059	0.059	0.059
22	0.117	0.117	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118
23	0.059	0.059	0.059	0.059	0.059	0.059	0.059	0.059	0.059	0.059	0.059	0.059
24	0.102	0.104	0.105	0.106	0.107	0.108	0.109	0.11	0.11	0.111	0.112	0.112
25	0.05	0.05	0.051	0.052	0.052	0.053	0.054	0.054	0.054	0.055	0.055	0.055
26	0.05	0.05	0.051	0.052	0.052	0.053	0.054	0.054	0.054	0.055	0.055	0.055
27	0.05	0.05	0.051	0.052	0.052	0.053	0.054	0.054	0.054	0.055	0.055	0.055

Figure 16.194: **Convergence at the 35th iteration**

So finally the invariant measure is:

$$\pi=\mu=\begin{pmatrix}0.12\\0.06\\0.06\\0.06\\0.18\\0.11\\0.06\\0.12\\0.06\\0.11\\0.06\\0.06\\0.06\end{pmatrix} \tag{16.6.17}$$

This gives us the convergent distribution per page of a cohort of 17 Internet users (sums of values of the stochastic vector).

The pages where μ_i is large are the most popular in terms of equilibrium probabilities. In the quest to classify web pages, it is still an argument to use the measure μ as an empirical indicator.

Remark

The reader will be able to find in the `R` companion book how we plot the above network and calculate the invariant measure in a more efficient way that in Microsoft® Excel®...

16.6.6.3 Absorbing states

However, if our graph contains a page (or group of pages) without issue then this one absorbs all the probability, because our cohort will fall sooner or later on this page and will not be able to leave it ("absorbing state" as defined In the Probabilities section). We represent this in the following graphical form if, for example, we add a page P_{13} that is an absorbing page:

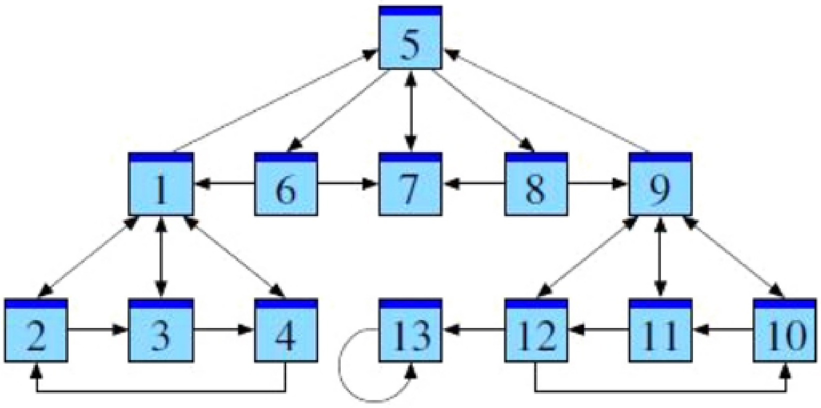


Figure 16.195: **Graph with an absorbent state**

With the associated transition matrix:

$$\begin{bmatrix} 0 & 2^{-1} & 2^{-1} & 2^{-1} & 0 & 2^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4^{-1} & 0 & 0 & 2^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4^{-1} & 2^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4^{-1} & 0 & 2^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4^{-1} & 0 & 0 & 0 & 0 & 0 & 1^{-1} & 0 & 4^{-1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3^{-1} & 2^{-1} & 0 & 2^{-1} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2^{-1} & 0 & 2^{-1} & 2^{-1} & 3^{-1} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4^{-1} & 0 & 0 & 3^{-1} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4^{-1} & 2^{-1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4^{-1} & 0 & 2^{-1} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3^{-1} & 1 \end{bmatrix} \tag{16.6.18}$$

Our model is not yet satisfactory. We could then construct a matrix that allows to escape the absorbent states by putting that any column that contains only one 1 with only elsewhere 0 teleport the visitor on all pages with an equal weight. Which would give us since we have 13 pages:

$$\begin{bmatrix} 0 & 2^{-1} & 2^{-1} & 2^{-1} & 0 & 2^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 13^{-1} \\ 4^{-1} & 0 & 0 & 2^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 13^{-1} \\ 4^{-1} & 2^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 13^{-1} \\ 4^{-1} & 0 & 2^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 13^{-1} \\ 4^{-1} & 0 & 0 & 0 & 0 & 0 & 1^{-1} & 0 & 4^{-1} & 0 & 0 & 0 & 0 & 13^{-1} \\ 0 & 0 & 0 & 0 & 3^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 13^{-1} \\ 0 & 0 & 0 & 0 & 3^{-1} & 2^{-1} & 0 & 2^{-1} & 0 & 0 & 0 & 0 & 0 & 13^{-1} \\ 0 & 0 & 0 & 0 & 3^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 13^{-1} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2^{-1} & 0 & 2^{-1} & 2^{-1} & 3^{-1} & 0 & 13^{-1} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4^{-1} & 0 & 0 & 3^{-1} & 0 & 13^{-1} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4^{-1} & 2^{-1} & 0 & 0 & 0 & 13^{-1} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4^{-1} & 0 & 2^{-1} & 0 & 0 & 13^{-1} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3^{-1} & 1 & 13^{-1} \end{bmatrix} \tag{16.6.19}$$

But this is not very optimal, because it loads the matrix so it is very greedy in terms of memories and calculations.

To escape the absorbing states, Google uses a more refined empirical model.

We take the initial application which is a linear mapping:

$$f:\mathbb{R}^n \mapsto \mathbb{R}^n \qquad \mu \mapsto P^T\mu \tag{16.6.20}$$

Where for reminder the sum of the components of μ is always 1. We rewrite this in the form:

$$f:\mathbb{R}^n \mapsto \mathbb{R}^n \qquad \mu \mapsto 1\cdot P^T\mu \tag{16.6.21}$$

where the 1 represents the fact that there is a 100% probability of having the state $P^T\mu$. Now the trick is to write:

$$f:\mathbb{R}^n \mapsto \mathbb{R}^n \qquad \mu \mapsto \varepsilon c+(1-c)P^T\mu \tag{16.6.22}$$

with:

$$\varepsilon=\begin{pmatrix}\dfrac{1}{n}\\ \vdots\\ \dfrac{1}{n}\\ \vdots\\ \dfrac{1}{n}\end{pmatrix} \tag{16.6.23}$$

which means that the Internet user has at all times a probability $(1-c)$ of being in the state $P^T\mu$ and a probability c of being in the state ε, that is to say to find ourselves with an equiprobable probability in any point of the graph. Finally, we sum up the two to obtain a stochastic vector (whose sum of the components is equal to 1).

Indeed, if we take for example 10% for the value of c, we have the sum of the components of the vector:

$$(1-c)P^T\mu \tag{16.6.24}$$

which will be equal to 0.9 and respectively the sum of the components of the vector εc that will be equal to 0.1 (since the sum of the components of ε equals 1 by construction).

So to summarize, with a fixed probability c the indexing engine (or the Internet user) abandons the current page P_i and starts again on one of the n pages of the web, this one being chosen, in order not to privilege anyone! Otherwise, with a probability $1-c$ the crawl engine follows one of the links of the page P_j, chosen equally. The only delicate point that remains is to calibrate the value of c which is therefore between 0 and 1. If it is 0 we fall back on the initial model with the problem of the absorbing states, if it is 1 we have a model that gives an equal note to each page. It is therefore necessary to choose c with low values close to 0.

This teleportation tip avoids being trapped by an absorbing page and guarantees to arrive anywhere in the graph, regardless of connectivity issues.

We must put the whole back in the form of components, so we have the notation (whose second term converges):

$$\boxed{\mu_1=\frac{1}{n}c+\sum_{j\rightarrow i}\frac{1-c}{l_j}\mu_j} \tag{16.6.25}$$

which has the enormous advantage of being able to calculate the value of any component μ_i of the stochastic vector of any step of a random walk on the graph and this without necessitating at any moment the manipulation of an immense matrix that would be too greedy in terms of memory capacity !!

To close this section about the PageRank algorithm let us quote the coinventor of the World Wide Web Robert Cailliau when speaking in the documentary *Made In Switzerland* (released in 12026 according to holocene calendar) produced by the RTS (Swiss Radio Television) at the timestamp 57m05s while sitting in one of the CERN office:

quote
16.6:8671.#91

> *The Internet wasn't public; only universities and research centers had access to it. Then, the Internet opened to the general public, and that's when parasites developed on the same infrastructure that have nothing to do with the web at all. These are all the social networks!*

Robert Cailliau

~20 %

★★☆☆☆

7 votes, 42.86%

16.7. Industrial Engineering

INDUSTRIAL ENGINEERING involves the design, improvement and installation of systems. It uses the knowledge from the mathematical, physical and social sciences as well as the principles and methods specific to the art of engineering in order to specify, predict and evaluate the results arising from these systems.

We can summarize all areas affecting the industrial engineering (and not only... industrial this can apply with ad hoc adaptation to the administration) with the objective to optimize and monitor the overall performance of the business (costs , deadlines, quality) because:

Only what is measured can be improved!

Note that some industrial engineering techniques have been discussed in other sections and subsections as: quantitative management techniques, optimization (operations research), financial analysis, queues analysis, etc. and that this section also includes "Quality Engineering".

In this chapter we will only deal with the minimum theoretical aspects of SQC (Statistical Quality Control) relating to statistical quality control (that's the job of the "quality controller") in the manufacturing environment and into production of goods or services and which is the minimum-minimorum knowledge of any quality engineer in any organization (industrial or administrative) under penalty of having no credibility! Also beware of companies - especially multinationals - who are looking for quality specialists mastering Microsoft® Excel® or Microsoft Access. Because it means that they use non-professional tools to do a job which should be done with appropriate tools (and Microsoft® Excel® or Microsoft Access are not appropriate tools) !!! So in terms of internal organization, you can ensure that these companies organize and analyse anything, anyhow, with an unsuitable tools and therefore that internally the organization is a general mess.

Depending on the use we distinguish three main areas of application of Industrial Engineering that are in the conventional order:

1. "Statistical Process Control" (S.P.C.), production monitoring and quality settings. This is the monitoring of a manufacturing process for the mass production of products to discover the differences in quality and to be able to step in and drive directly corrective actions.

   The engineers must absolutely consult the standard ISO/TR 13425:2006 *Guidelines for the selection of statistical methods in standardization and specification* as well as the standard ISO/TR 8258 *Shewhart control cards* and finally the ISO/TR 18532 *Guidelines for the application of statistical methods to quality and to industrial standardization* before implementing S.P.C. tools within their company.

Remark

Most mathematical techniques related to this field are univariate models. There are some bivariate models also but above it starts to become a nightmare (some unique pieces in the industry have more that ten or twenty specifications). I complicated situations production is managed by Machine Learning techniques that corrects production parameters using neural networks. This is what we name "Automated Process Control" (A.P.C.).

2. "Delivery check or receipt sample test (Acceptance Sampling)". This is an input control, a control during production and a final product inspection in a company (or factory) without direct influence on the production. Thus, the amount of product is measured. The initial inspection is also used to reject the incoming goods. It therefore influences the production only indirectly.

   The engineers must absolutely consult the family of norms ISO 3591 *Sampling procedures for inspection by variables and attributes* before setting up reception control tools within their company.

3. Preventive maintenance and control of aging and failure and critical impacts best known as "Failure Mode Analysis, Effects and Criticality Analysis (FMEA)". This is mainly to calculate the lifespan of components or machines to provide replacements in advance and actions related thereto to be taken to avoid human or financial emergencies.

   The engineer must absolutely consult the norm IEC 61649 *Weibull Analysis* and the norm NF EN 13306 *Terminology for maintenance* before putting in place preventive maintenance tools within their company.

These three areas using statistics in general, the engineer should always refer to the family of norms ISO 3534 *Vocabulary and symbols*, ISO 3534-1 *Probability and general statistical terms*, ISO 3534-2 *Statistical Quality Control*, ISO 3534- 3 *Design of experiments*.

For information, since the late 120th century (holocene calendar), it is fashionable to combine the first two points in a working methodology named "Six Sigma" that we will study immediately. Finally, note that in practice, to get an interest of chiefs executives the engineer must always find a quantitative relations between non-quality and costs in order to make things change...

16.7.1 Six Sigma

Two objects are never perfectly identical. Whatever the techniques used to manufacture these objects or precise are the tools, there is variability in any production process. The aim of any industrial is that this natural variability remains within acceptable limits. This is a major concern in the improvement of industrial quality.

One of the tools used to work towards this quality is the "Statistical Process Control " (S.P.C.). If you produce a certain type of object, and if you want to keep your customers to sustain your business, you must ensure that the lots that you deliver them are consistent with what has been agreed between you, usually by contract. Any serious industrial performs checks on produced lots to check the quality, that he is the manufacturer of the goods or that he receives them. Various statistical techniques related to sampling are then used to avoid, in most cases, to check one by one all the objects in a batch. The samples control taken in batches is essential if the controls at destroying the artefact, as in an analysis of the active ingredient dose contained in a tablet. However, there are cases where it is preferred to check all objects (for example it is desirable that the brakes of a vehicle works and a control of the braking function on a sample of produced cars does not guarantee that all vehicles have good braking...).

Control
Define
LEAN
6σ
Measure
Analyze
Improve

When a batch is controlled, it is conform or it is not. If it is conform, we deliver it (we are the supplier) or we accept it (we are the customer). If it is not conform, we can destroy it, check one by one all the elements and destroy only those who do not comply, etc. All solutions to handle non-conforming lots are expensive. If the batch does not comply, the damage is done. The S.P.C. methods aim to avoid producing non-conforming goods by monitoring production and intervening whenever anomalies are found. A good S.P.C. implementation eliminates a significant number of controls of conformity by setting up statistical tools for monitoring manufacturing processes.

To resume, the S.P.C. is therefore to control the process during manufacturing and to act on the process rather than on the product if problems are detected. This approach tries to go up at the highest possible level of the production line to prevent the occurrence of defective product and goods. We discuss in this special case of "process control".

This is a useful working methodology to satisfy customers and the idea of this methodology is to deliver products/services of quality, knowing that quality is inversely proportional to the variability. Moreover, the introduction of quality should be optimized so as not to overly increase costs. The subtle interplay between these two parameters (quality/costs) and their joint optimization is often associated with the term "lean management". If we integrate Six Sigma, then we speak of "Lean Six Sigma".

Six Sigma integrates all aspects of the control of variability in business whether at the level of manufacturing, services, organization, marketing or management. Hence why it is such interesting! Moreover, in Six Sigma a defect must be paradoxically welcome, because it is a source of progress of an initially hidden problem. Then you have to ask several times the question "Why?" (traditionally 5 times) to better trace the source of it.

We distinguish two types of variability in practice:

1. The "inherent variability" in the process (and difficult to change) that induces the concept

of measures distribution (usually accepted by businesses as a Normal distribution).

2. The "external variability" which induces most often a bias (deviation) in the distributions over time.

Manufacturing processes in the high tech industry having a strong tendency to become terribly complex, it should be noted that the basic components used for each product are not always of equal quality or performance. And if in addition, production procedures are difficult to establish, the drifts will inevitably exists.

Whether for one reason or another, ultimately many products will be outside of the normality and will deviate of the range corresponding to the acceptable quality for the customer. This drift is very costly for the company. The management of waste, rework or customer returns for nonconformity generate substantial costs seriously amputating the expected benefits.

As we shall see in the texts, a pretty fair possible definition of Six Sigma is: solving problems based on the exploitation of data. Thus it is a scientific method of management!

16.7.1.1 Quality Control

In the situation of quality studies in a business, we often give up a 100% control because of the price it would generate. We then do a sampling. These must obviously be representative, that is to say, representatives and with equal probabilities (i.e. the mixture is good).

The purpose of sampling is obviously the real probability of failure of the whole lot on the basis of observed failures in the sampling.

Let us recall before going further that we have seen in the section of Statistics the hypergeometric distribution (and its interpretation) given by (see section Statistics page 672):

$$H(k)=\frac{C_k^m C_{p-k}^{n-m}}{C_p^n} \tag{16.7.1}$$

where the notation of the binomial coefficient is consistent with that defined and chosen in the section of Probabilities (therefore non-compliant with ISO 31-11).

During a sampling, we normally have a pack of n elements which we draw p of them. Instead of taking m (remember it is an integer!) as the number of defective parts, we will implicitly defined it as being equal to:

$$m=p_d\cdot n \tag{16.7.2}$$

where p_d is the probability (assumed known or imposed...) that a part is defective. Thus, we have for probability to find k defective parts in a sample of p among n:

$$H(k)=\frac{C_k^{p_d n} C_{p-k}^{n-p_d n}}{C_p^n}=\frac{C_k^{p_d n} C_{p-k}^{n(1-p_d)}}{C_p^n} \tag{16.7.3}$$

The cumulative probability of finding k defective parts (between 0 and k defectives in other words) is then simply calculated with the cumulative hypergeometric distribution:

$$\sum_{x=0}^{k} H(x)=H(0)+H(1)+\ldots+H(k)=\sum_{x=0}^{k}\frac{C_k^{p_d n} C_{p-k}^{n(1-p_d)}}{C_p^n} \tag{16.7.4}$$

✍**Example:**

In a batch of 100 machines, we assume that a maximum of 3 are defective (i.e. that is to say $p_d=3\%$). We conduct a control sampling p at each output order of 20 machines.

We want to know at first what is the probability that in this sample of size p, 3 machines are defective and secondly what is the maximum number of defective machines allowed in this sample of size p that says us with 90% confidence that the batch of n machines contain only 3 defective.

x	$H(x)$	$\sum_{x=0}^{k} H(x)$
0	0.508	0.508
1	0.391	0.899
2	0.094	0.993
3	0.007	1.000

Table 16.12: **Application of the hypergeometric distribution**

Thus, the probability of sampling in one series three defective machines in the sample of 20 is $0.7\%(0.007)$ and the maximum number of defective machines allowed in this sample of 20 that allows us with at least 90% of confidence to have 3 defectives is 1 defective machine found in the sample (cumulative probability: 0.899)!

$H(x)$ values can be calculated easily with the English version of Microsoft® Excel® 11.8346. For example, the first value is obtained by the function:

```
=HYPGEOM.DIST(0,20,3,100,FALSE)
=COMBIN(30,0)*COMBIN(97,20-0)/COMBIN(100,20)=0.508
```

16.7.1.2 Defaults/Errors

Let us now exhibit for the general culture a practical and special example of what is only a simple application of the theory of statistics and probabilities. To understand the importance of quality and the concept of "zero defects", let us consider the following general example of the inventor of the method:

For example, consider that the installation of a tourism car includes $2,500$ operations (pieces) and each operation is perfect 99 times out of 100. A priori, a having a successful operation in 99% of cases is a sign of an almost perfect mastery of the quality. But in fact, the perfection of the whole assumes that $2,500$ times, the operations are fully perfectly realized! If the daily production is of $2,000$ units, one the 5 million operations that are done daily in our factory, there is $100\%-99\%=1\%$ of defects assemblies that is to say $50,000$, and an average of 25 defects per car, which is hardly acceptable. Also, let us imagine this imaginary factory is equipped with a control service sampling systematically at the end of the assemblies line. This represents a considerable cost in working hours for control. If the defects can be corrected, this will require rework, replace perhaps some parts and work in unexpected conditions to correct the defects. If the defects are too important, these defects makes the final product unusable, and scrap are extremely expensive depending on the product. Even worse: if the control service sees 99% of defects it remains about 500 of them daily and this requires costly repairs and returns and a significant deterioration of the image depending on the performance of competitors.

This example serving as a school case study, imagine now a company manufacturing 3 copies of a same product coming out of the same production line, each copy being composed of 8 elements.

Remark

We can just as easily imagine a services company developing (manufacturing) 3 copies of a software (product) out of the same development team (production line), each composed of an equal number of modules (elements).

Suppose that the product $P1$ has 1 defects in average, the product $P2$ has 0 defects $P3$ has 2 defects and that each product is composed of 8 items.

Here Six Sigma implicitly assumes that defects are independent variables, which is relatively rare in machinery production lines but more common in chains in which humans are involved. However, we can consider in the application of the S.P.C. on machines that a sampling in time in the measurement process is equivalent to having a random independent variable!!

Remarks

R1. As part of the example above with the software, independence is unlikely if we do not take an example in which the programming modules are customized according to customer needs.

R2. The inconstancy of production results of certain machines whose settings move during operation... (which is common!), and also where raw material quality changes during production (which is also common!) generates major issues in S.P.C. methods and is this inconstancy problem is the heart of Six Sigma.

The arithmetic average of the defects is named in the Six Sigma methodology "Defects Per Unit (D.P.U.)" and is defined by:

$$\text{DPU}=\frac{\text{\# of defects in sample}}{\text{\# of units in sample}}=\frac{D}{N} \tag{16.7.5}$$

And this will give with our small example above:

$$\text{DPU}=\frac{1+0+2}{3}=\frac{3}{3}=1 \tag{16.7.6}$$

which means that on average each product has a design or manufacturing defect.

Many quality softwares also reports the cumulative DPU over a population of sample and obviously defined for a period j by:

$$\text{Cumulative DPU}_j=\frac{\sum_{i=1}^{j} D_i}{\sum_{i=1}^{j} N_i} \tag{16.7.7}$$

Remark

Warning! This value is not a probability for the simple reasons that it may initially be higher than 1 and that it has the dimension of [defects]/[products].

Similarly, the analysis can be done with the total number of possible defective elements that make up the product so that we are led naturally to define following the Six Sigma methodology the "Defects per Unit Opportunity (D.P.O.)":

$$\text{DPO}=\frac{\text{DPU}}{\text{\# of potential defects in a unit}} \tag{16.7.8}$$

And this will give with our small example:

$$\text{DPO}=\frac{\text{DPU}}{m}=\frac{1}{8}=0.125 \tag{16.7.9}$$

and this can be seen as the probability of a default product element since it is a dimensionless value and the value is always less than or equal 1:

$$\text{DPO}=\frac{\text{DPU}}{m}=\frac{\left[\dfrac{\text{defaults}}{\text{unit}}\right]}{\text{total potential}\left[\dfrac{\text{defaults}}{\text{unit}}\right]} \tag{16.7.10}$$

By extension, we can argue that 87.5% of the components of a unit does not have defects and because Six Sigma enjoys working with examples of the order of a million (it's more awesome...) we then define the "Defects per Million Opportunities (D.P.M.O.)" as:

$$\text{DPMO}=\text{DPO}\cdot 1,000,000=\frac{\text{DPU}}{m}\cdot 10^6 \tag{16.7.11}$$

And this will give with our small example:

$$\text{DPMO}=\frac{\text{DPU}}{m}\cdot 10^6=\frac{1}{8}\cdot 10^6=125,000 \tag{16.7.12}$$

To summarize:

$$\boxed{\begin{aligned}\text{DPMO}&=\frac{\text{\# of defects in sample}}{\text{\# of units in sample}\cdot\text{\# of potential defects in a unit}}\cdot 10^6\\&=\frac{\text{DPU}}{\text{\# of potential defects in a unit}}\cdot 10^6=\text{DPO}\cdot 10^6\end{aligned}} \tag{16.7.13}$$

As D_i is the probability that an element i is not defective is 87.5% (that is to say 12.5% of scrap) then, by the axiom of joint probabilities (see section Probabilities page 480), the probability that a product as a whole is not defective if all elements are serial is:

$$P\left(\bigcap_{i=1}^{8} D_i\right)=\prod_{i=1}^{8}(1-\text{DPO})=(1-\text{DPO})^8=0.875^8\cong 0.3436=34.36\% \tag{16.7.14}$$

This is not good...

In Six Sigma, the joint probabilities are also naturally used to calculate the joint probability of non-defective products in a production process connected in series or to calculate the defect probability of an administrative serial workflow. This joint probability is named in Six Sigma the "Rolled throughput Yield (R.T.Y.)" and is denoted by:

$$\boxed{P\left(\bigcap_{i} P_i\right)=\prod_{i} P_i} \tag{16.7.15}$$

✍**Example:**

Consider the following serial process/workflow or production line with its given yields (non-default rates):

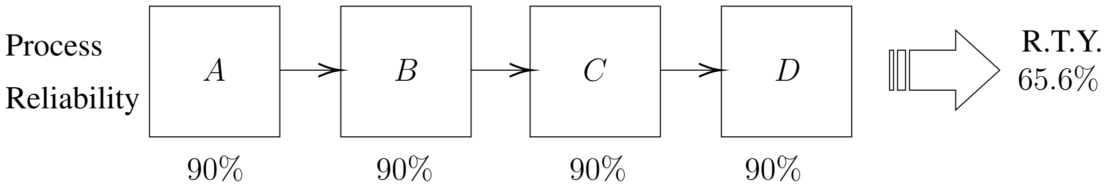


Figure 16.196: Serial process/workflow

We find ourselves in final with a 65.6% reliability that is to say a cumulative probability of default for the entire process/workflow of 34.4% (you can also imagine that these are four tasks in sequence of a project!).

Remark

The reader that is attentive will have noticed that the serial system is always less reliable than its least reliable component!!

This type of calculation is widely used by supply chain manager and the result is named "availability rate" as well as by project managers for the duration of a phase of a project when they consider durations as independent tasks (for more complex structures, remember that we sometimes talk about "weighted probability trees" or "topological systems").

Thus, in an industrial production line based on the previous example to have a well-defined amount Q of products (supposed to use only one component of each stage) at the end of the production line, it will be necessary at the step A to bring:

$$A=\frac{Q}{\text{RTY}}=\frac{Q}{\prod\limits_i P_i}=\frac{Q}{65.6\%}=152.42\% \qquad (16.7.16)$$

That is to say 52.42% of components of type A more than expected. At stage B:

$$B=\frac{Q}{\text{RTY}}=\frac{Q}{\prod\limits_i^{n-1} P_i}=\frac{Q}{0.9\cdot 0.9\cdot 0.9}=\frac{Q}{72.9\%}=137.17\% \qquad (16.7.17)$$

Recall now that the probability density of having k times the event p and $N-k$ times the event q in any arrangement (or order) is given by (see section Statistics page 658):

$$P(n,k)=C_k^n P=\frac{n!}{k!(n-k)!}p^k q^{n-k} \qquad (16.7.18)$$

and is nothing less than the binomial law whose expected mean and standard deviation are (see section Statistics page 658):

$$\mu=np \qquad \sigma=\sqrt{npq} \qquad (16.7.19)$$

Thus, in the Six Sigma methodology, we would apply the binomial distribution (to make things simple because in facts we should use the Hypergeometric law for small samples) to determine what is the probability of having zero defective items and 8 others working well on a product of the production line of our example (if all elements have the same probability to fail for sure...):

$$P(8,0)=\frac{8!}{0!(8-0)!}0.125^0 0.875^8=1\cdot 0.875^8\cong 0.3436=34.36\% \tag{16.7.20}$$

and we obviously fall back on the value obtained with the joint probabilities:

$$\begin{aligned}\mu&=8\cdot 0.125=1=\text{DPU}\\ \sigma&=\sqrt{8\cdot 0.125\cdot 0.875}=0.935\end{aligned} \tag{16.7.21}$$

Or the probability to have one defective item and seven working others on a product of the production line:

$$P(8,1)=\frac{8!}{1!(8-1)!}0.125^1 0.875^7=\frac{8!}{1!7!}0.125^1 0.875^7\cong 0.3926=39.26\% \tag{16.7.22}$$

we see that the binomial distribution gives us 39.26% probability to have one defective item on 8 in a product.

Moreover, in the section of Statistics, we have shown that when the probability p is very small and tends to zero but however the average value $n\cdot p$ tends to a fixed value when n approaches infinity, the binomial law of average $\mu=np$ with k trials was then given by a Poisson law:

$$P_k=P(X=k)=p(k)=\frac{\mu^k}{k!}e^{-\mu} \tag{16.7.23}$$

with (see section Statistics page 686):

$$\mu=np \qquad \sigma=\sqrt{\mu} \tag{16.7.24}$$

Remark

In a practical context, it is made use of the maximum likelihood estimators of the exponential law to determine the mean and standard deviation above (see section Statistics page 825).

What in Six Sigma methodology is naturally written:

$$p(k)=\frac{\text{DPU}^k}{k!}e^{-\text{DPU}} \tag{16.7.25}$$

with:

$$\mu=\text{DPU} \qquad \sigma=\sqrt{\text{DPU}} \tag{16.7.26}$$

Thus, in our example, it is interesting to see the value obtained (which will necessarily be different since we are far from having an infinity of individuals and p is far from being small):

$$p(k)=\frac{\mu^k}{k!}e^{-\mu}=\frac{np^k}{k!}e^{-np}=\frac{np^0}{0!}e^{-0.125\cdot 8}=\frac{1}{1}e^{-1}\cong 36.97\% \tag{16.7.27}$$

with:

$$\mu = 1 \qquad \sigma = 1 \tag{16.7.28}$$

that is a result even worst than with the binomial law for our products.

However, if p is fixed and small, the average $\mu = np$ also tends to infinity in theory with the Binomial Law, moreover the standard deviation $\sigma = \sqrt{npq}$ also tends to infinity.

If we want to calculate the limit of the binomial distribution, we will therefore need to make a change of origin which stabilizes the average at 0 for example and a change of unit that stabilizes the standard deviation for example to 1. This calculation has already been done in the Statistics section (see page 692), we know that the result is the Normal distribution:

$$P(k) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(k-\mu)^2}{2\sigma^2}} \tag{16.7.29}$$

For our example we know we have to take $k = 0, \sigma = 0, \mu = 0$:

$$P(0) = \frac{1}{1 \cdot \sqrt{2\pi}} e^{\frac{(0-1)^2}{2 \cdot 1^2}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}} \cong 24.19\% \tag{16.7.30}$$

Thus, applying the Normal distribution, we have 24.19% chance to sample the first time a defective product. This difference compared to other methods is simply explained by the assumptions (finite number of individuals, significant probability of defect, etc.).

16.7.1.3 Capability Indices

The Six Sigma methodology (and also the series of norms ISO 22514) defines several indexes that permits to quantify during the manufacturing process the capability of control in the case of a large amount of defects measures distributed as Gaussian-Laplace law (Normal law).

In the case where the measures do not follow a Normal law we have therefore to transform the data using various empirical techniques (Johnson transformations typically that Minitab and `R` can easily manage) or Kernel Smoothing techniques (see section Statistics page 1143) that softwares like SAS JMP can easily manage.

Basically, if we imagine ourselves working in a company, in charge of the manufacturing quality of a new machine of a new series of pieces, we will be faced with two major situations:

1. At the beginning of production, there may be big quality differences due to defects in the machine or poorly initialization of important settings of the machine. These are defects that will often be rapidly corrected (on the short-term) thanks to a systematic control (destructive or not) of small samples.
2. Once the big anomalies corrected, we will have in theory only minimal issues that will be very difficult to control and even on the long term. Therefore systematic control is not necessary anymore and once this period of large corrections passed, we make checks per batch (between each correction) and each will be considered an independent and identically distributed random variable (according to a Normal law), but obviously with a different average and standard deviation.

These two scenarios show that we do not then logically perform the same tests at the beginning of the production and then on the long term. This is why we define in S.P.C. several indices (whose symbols are unique to this textbook because they change according the used standards) whose two most important are:

D1. We name "short-term process potential capability" or "short-term dispersion index" the ratio of the range of control E of the distribution of values tolerated by the customer and the Six Sigma quality (6σ) when the process is almost under control as:

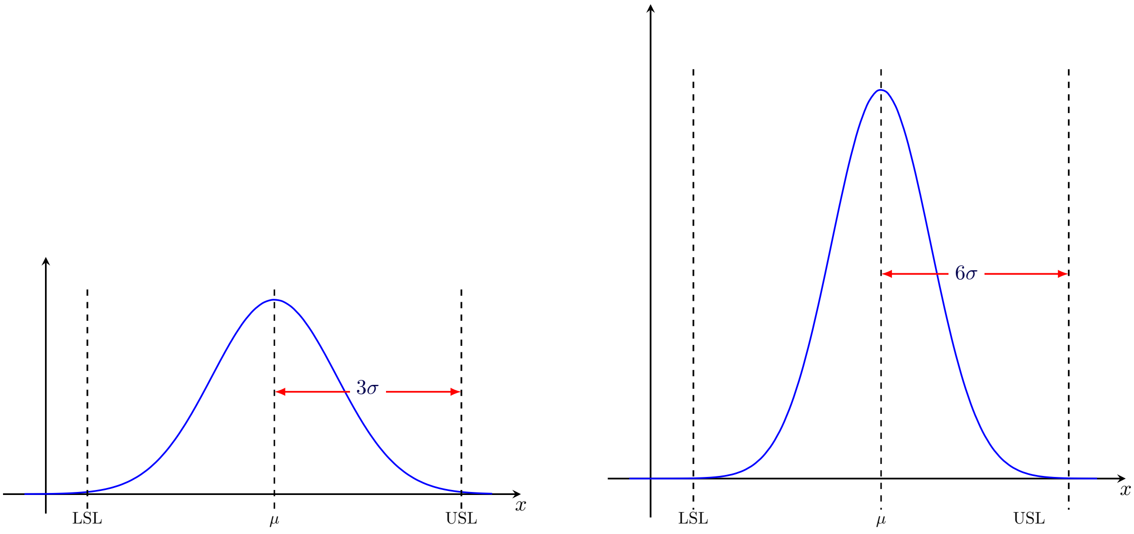
$$\boxed{C_p=\frac{E}{6\sigma}=\frac{\text{USL}-\text{LSL}}{6\sigma_{\text{ST}}}} \qquad (16.7.31)$$

More generally given by:

$$C_p=\frac{E}{6\sigma}=\frac{\text{USL}-\text{LSL}}{L\sigma_{\text{ST}}} \qquad (16.7.32)$$

Where USL is the "Upper Specification Level" of the distribution and LSL the "Lower Specification Level" that we often impose (but not always!) in the industry as at equal distances to the desired theoretical mean μ. The term L (often equal to 6 in the Six Sigma methodology) is named the "study variation".

Normally within manufactures, the range of control is fixed (the numerator) and so when the value of the standard deviation is large (more variations, not enough controls) the value of the index is low and when the standard deviation is low (less variation, lot of controls) the value of the index is high as shown by the two examples below:



(a) $C_p=1$ (b) $C_p=2$

Figure 16.197: **Illustration of C_p index**

The above capability is therefore an indication that we will seek to maximize (since the standard deviation in the denominator should be minimized!).

This ratio is useful in the industry in the sense that the range E (which is important because it represents the tolerated dispersion/process variation) is considered as the "voice of the customer" (his request) and the 6 sigma denominator the actual behaviour of the process/process meant to include virtually all possible outcomes. It is better to hope that this report is at worst equal to unity!

Here is a typical example in project management where, when the customer does not pay for a fine tuning risk modelling (the customer therefore accepts by contract a change in time and costs that can exceed 50%), we come across with this type of chart:

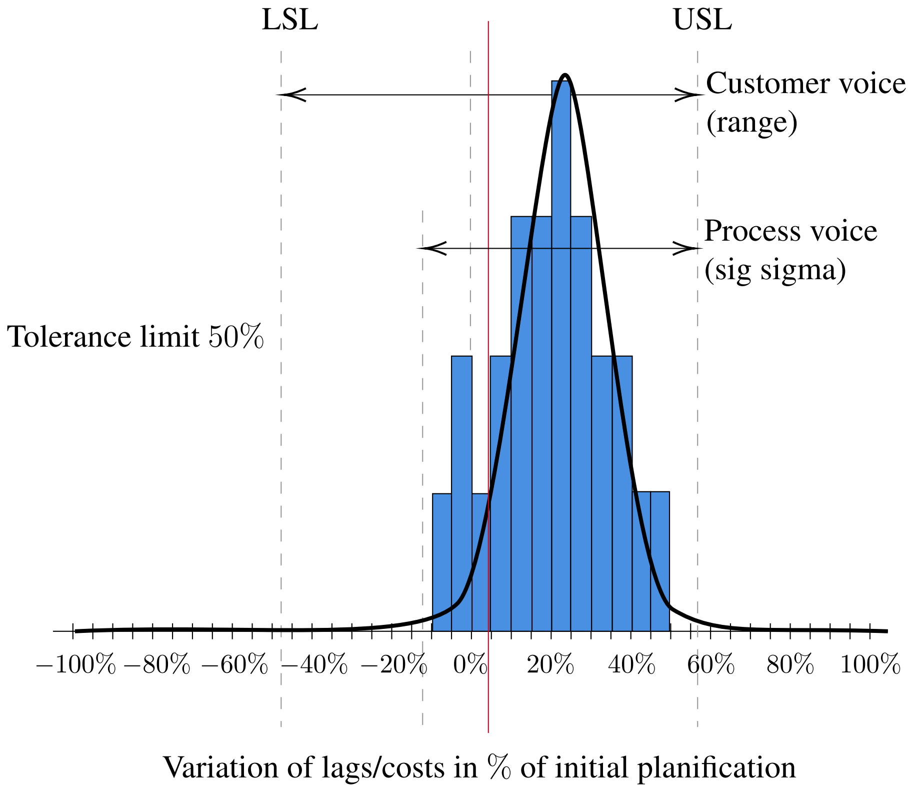


Figure 16.198: **Typical plot of a control histogram with tolerances and limits**

Remark

In S.P.C., the range E is sometimes denoted TI, meaning "tolerance interval" (but this is wrong... instead it should be named "fluctuation interval").

The standard deviation in the denominator is simply defined as the average of the variances in the case of k independent random variables:

$$\boxed{\sigma_{\text{ST}}=\sqrt{\frac{\displaystyle\sum_{i=1}^{k}\sigma_i^2}{k}}} \tag{16.7.33}$$

where it is useful perhaps to remember that ST is the abbreviation for "Short Term" (abbreviation often unspecified in the practice as assumed known in the context). The standard deviations σ_{ST} is obviously the easiest to apply for the first scenario which we have mentioned above. Because between each big correction, batches are considered independent and can not be analysed as a single lot (it would be an aberration!).

Remark

According to the type of industry some use the standard deviation based on the range for σ_{ST} and that is given as we proved it in the section Statistics during our study of extreme values by the ratio of the average of ranges of all measurements made by different operators and divided by the Hartley's constant such that:

$$\sigma_{\text{ST}}=\frac{\bar{R}}{d_2(k)} \tag{16.7.34}$$

Be careful! As it is often the situation in short term processes (during the correction of the big sources of error), tested batches are small - even very small - and this in the purpose to reduce costs. Then the standard deviation below the root is for sure not really a correct value... This is why it is important to calculate a confidence interval! So let us see that it is possible to easily construct a hypothesis test for the C_p indicator under the assumption that each sample is identically distributed following a Normal law. Indeed, remembering that we proved in the section Statistics (see page 863) that:

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}\leq\sigma^2\leq\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}} \tag{16.7.35}$$

It is immediate that:

$$\frac{\chi^2_{1-\alpha/2,n-1}}{(n-1)S^2}\leq\frac{1}{\sigma^2}\leq\frac{\chi^2_{\alpha/2,n-1}}{(n-1)S^2} \tag{16.7.36}$$

Thus:

$$\frac{1}{S}\sqrt{\frac{\chi^2_{1-\alpha/2,n-1}}{n-1}}\leq\frac{1}{\sigma}\leq\frac{1}{S}\sqrt{\frac{\chi^2_{\alpha/2,n-1}}{n-1}} \tag{16.7.37}$$

Therefore:

$$\frac{E}{\hat{\sigma}_{\text{ST}}}\sqrt{\frac{\chi^2_{1-\alpha/2,n-1}}{n-1}}\leq\frac{E}{\sigma_{\text{ST}}}\leq\frac{E}{\hat{\sigma}_{\text{ST}}}\sqrt{\frac{\chi^2_{\alpha/2,n-1}}{n-1}} \tag{16.7.38}$$

or:

$$\boxed{\hat{C}_p\sqrt{\frac{\chi^2_{1-\alpha/2,n-1}}{n-1}}\leq C_p\leq\hat{C}_p\sqrt{\frac{\chi^2_{\alpha/2,n-1}}{n-1}}} \tag{16.7.39}$$

and we can apply the same reasoning for all types of indicators of the same kind that we will see later!

Remark

As part of our study later of control charts, we will see that it is possible to use special expressions for the standard deviation when working with samples of measures. These expressions will be based for one on the chi-square law and the other on order statistics.

D2. We name "Overall performance of long-term process" the ratio of the range of control E of the distribution of values tolerated by the customer and the Six Sigma quality (6σ) when the process is centered that is to say under statistical control (product manufacturing parameters vary only a little bit) as:

$$\boxed{P_p=\frac{E}{6\sigma}=\frac{\text{USL}-\text{LSL}}{6\sigma_{\text{LT}}}} \tag{16.7.40}$$

But where the standard deviation in the denominator is given this time by the fact we consider all the big issues corrected and the process as stable such that we can consider all manufactured pieces in the long term as a single controlled batch:

$$\boxed{\sigma_{\text{LT}}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2}} \tag{16.7.41}$$

where it is useful perhaps to remember that LT is the abbreviation for "Long Term" (abbreviation often unspecified in the practice as assumed known in the context). This standard deviation is obviously the simplest for the second scenario which we have mentioned above. Because the variations are now, by hypothesis, very small, the entire production can be assumed as a single lot of control over the long term (this does not avoid that sometimes it is necessary to clean the extreme values that may occur).

Tolerancing the characteristics is therefore important for obtaining the desired quality and reliability of assembled products. Traditionally, a tolerance is expressed in the form of a bi-point $[\text{Min},\text{Max}]$. A characteristics is then accepted as conform if it is within tolerance.

The problem of the art of tolerancing is to try to conciliate the fixation of the acceptable variability limits to reduce production costs and ensure the highest level of quality in the finished assembled product.

Two approaches attempt to solve this problem:

1. "Worst Tolerancing" ensures the assembly in all situations from the time the basic characteristics are within tolerance. This is the far the best method when the production is very small (and in the extreme case when there is only one unit...).

2. "Statistical Tolerancing" takes into account the low probability of extreme assemblies and enables to significantly expand the tolerances to reduce costs and so it is to this that we will look further below. For sure this can be apply only for large production sets.

A process is said to "limit capable" if the ratios C_p, P_p given above (choosing six times the standard deviation) is almost equal to 1. But in the industry, we actually prefer to take the value 1.33 in the case of a Normal distribution that is not perfectly centered (we will see later where this last value comes from).

Obviously, the value of the standard deviation σ can be calculated by using the maximum likelihood estimators with and without bias seen in the Statistics (see page 822) and in practice you must never forget that this is just an estimator and not the real theoretical value! Furthermore, we will see later that depending on the standard deviation used, the notations change!

Remark

In manufacture, one must be careful because the measuring instrument adds its own standard deviation (error) on these of the production process.

As we have proved it in the section on Statistics (see page 611), the standard error (standard deviation of the mean) is:

$$\sigma_{\overline{X}}=\frac{\sigma}{\sqrt{n}} \tag{16.7.42}$$

In the Six Sigma methodology, we then often have for long term process and under statistical control:

$$\boxed{\begin{aligned}\text{LCL}&=\mu-3\frac{\sigma}{\sqrt{n}}\\ \text{UCL}&=\mu+3\frac{\sigma}{\sqrt{n}}\end{aligned}} \tag{16.7.43}$$

when we analyse control charts (see further below at page 8934) whose random variables is a sample of n independent and identically Normally distributed random variables and that the limits were not imposed by a customer or an internal policy or technical constraints! Obviously, we must be aware that UCL and LCL are not the same mathematical expression in more complex cases and therefore for distributions other than the Normal distribution!

Remark

Keep in mind that UCL is the "Upper Control Limit" and LCL "Lower Control Limit" and that at the opposite of the USL (Upper Specification Limit) and LSL (Lower Specification Limit) their are calculated by more or less complex mathematical relations.

Instead of specification limits, a process may have "boundaries". The difference is that when there are boundaries we assume that it is not possible for data to exceed the specification limits and therefore the C_p (among others) has no sense to be calculated (some softwares - like Minitab - ask the user if he wants change his specification limits as boundaries for example...).

Furthermore, the above expression differs for short-term process because the example given above is for a case of measures over the long term for reminder!

As we know C_p requires that the average (target) is centered between USL and LSL. Therefore, the average is confounded with what we name the "target T" of the process.

But the average μ in reality can be offset from the original target T which must always (in common use) be equidistant between USL and LSL as shown in the figure below in the special case a Normal distribution:

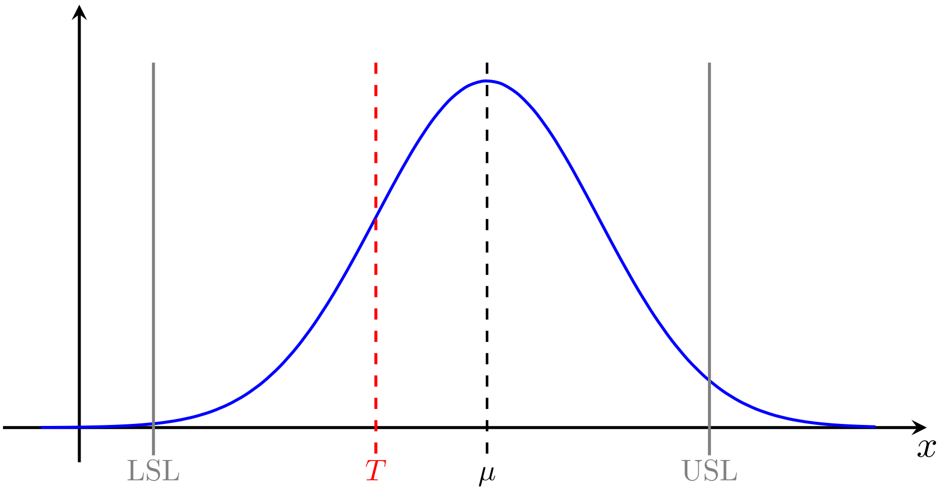


Figure 16.199: Measures under statistical control offset from the target (typical with index C_p)

But this is not necessarily the case in reality where engineers (whatever their application field) can choose asymmetrical LSL and USL with respect to the average and this only because the distribution is not always following a Normal law (typically the case in project management...)! Hence the following definition:

Definition (#16.7.731): We name "short term non-centered potential Capability for the process" (in the biased case) or more frequently "Process Capability Index (Within)" the relation:

$$\boxed{C_{pk}=C_p(1-k)} \tag{16.7.44}$$

with (in the special case of a 6σ process):

$$k=\frac{|T-\mu|}{\dfrac{1}{2}(\text{USL}-\text{LSL})}=\frac{|T-\mu|}{3\sigma C_p} \tag{16.7.45}$$

where $k\geq 0$ is named the "degree of bias" or "index position" and T is as we know the "target" naturally given most of the time (so not always!) by:

$$T=\frac{\text{USL}-\text{LSL}}{2} \tag{16.7.46}$$

which gives the middle of the distribution relative to the bi-point $[\text{LSL},\text{USL}]$ requested by the customer (remember that the standard deviation in the denominator of the prior-previous relation is the standard deviation short term!).

In fact this control capability indicator C_{pk} may seem very artificial, but it is not completely... Indeed, there are some outstanding values (those that interest the engineer) that can help to get a good idea of what happens with it:

1. If the average and the target are confused, then we have:

   $$T=\mu \tag{16.7.47}$$

   We find ourselves with $k=0$ and therefore $C_{pk}=C_p$ and judgement criterion of the value of the index will then be based on the short-term capability index.

2. If because of a poor control of the process we have:

$$|T-\mu|>0.5(\text{USL}-\text{LSL}) \tag{16.7.48}$$

Then the mean μ is either above USL or below LSL which has the consequence of having $k>1$ and therefore $C_{pk}<0$.

3. If we have:

$$0<|T-\mu|<0.5(\text{USL}-\text{LSL}) \tag{16.7.49}$$

Then the average μ is between the values USL and LSL which has the consequence of having $0<k<1$ and therefore $0<C_{pk}<C_p$.

4. If we have:

$$|T-\mu|=0.5(\text{USL}-\text{LSL}) \tag{16.7.50}$$

Then it simply means that the average is aligned with USL or LSL and we therefore have $k=1$ and $C_{pk}=0$. As the interpretation remains sometimes delicate and difficult, we build unilateral capability indices "Upper Capability Index CPU" and "Lower Capability Index CPL" given by:

$$\boxed{\text{CPU}=\frac{\text{USL}-\mu}{3\sigma}\qquad \text{CPL}=\frac{\mu-\text{LSL}}{3\sigma}} \tag{16.7.51}$$

That we will obviously also seek to maximize. Let's see where these two definitions comes from and how to use them:

Proof 16.2.2. First, we need two specific formulations of the degree of bias k.

(a) If:

$$\frac{\text{USL}+\text{LSL}}{2}<\mu \tag{16.7.52}$$

then we can get rid of the absolute value:

$$\begin{aligned}k&=\frac{|T-\mu|}{\frac{1}{2}(\text{USL}-\text{LSL})}=\frac{\left|\frac{\text{USL}+\text{LSL}}{2}-\mu\right|}{\frac{1}{2}(\text{USL}-\text{LSL})}=\frac{\mu-\frac{\text{USL}+\text{LSL}}{2}}{\frac{1}{2}(\text{USL}-\text{LSL})}\\&=\frac{2\mu}{\text{USL}-\text{LSL}}-\frac{\text{USL}+\text{LSL}}{\text{USL}-\text{LSL}}\end{aligned} \tag{16.7.53}$$

(b) If:

$$\frac{\text{USL}+\text{LSL}}{2}>\mu \tag{16.7.54}$$

then we can get rid of the absolute value:

$$\begin{aligned}k&=\frac{|T-\mu|}{\frac{1}{2}(\text{USL}-\text{LSL})}=\frac{\left|\frac{\text{USL}+\text{LSL}}{2}-\mu\right|}{\frac{1}{2}(\text{USL}-\text{LSL})}=\frac{\frac{\text{USL}+\text{LSL}}{2}-\mu}{\frac{1}{2}(\text{USL}-\text{LSL})}\\&=\frac{\text{USL}+\text{LSL}}{\text{USL}-\text{LSL}}-\frac{2\mu}{\text{USL}-\text{LSL}}\end{aligned} \tag{16.7.55}$$

We have therefore when $T > \mu$:

$$\begin{aligned}C_p(1-k) &= C_p\left[1-\left(\frac{\text{USL}+\text{LSL}}{\text{USL}-\text{LSL}}-\frac{2\mu}{\text{USL}-\text{LSL}}\right)\right]\\ &= \frac{\text{USL}-\text{LSL}}{6\sigma}\left[1-\left(\frac{\text{USL}+\text{LSL}}{\text{USL}-\text{LSL}}-\frac{2\mu}{\text{USL}-\text{LSL}}\right)\right]\\ &= \frac{\text{USL}-\text{LSL}}{6\sigma}-\frac{\text{USL}+\text{LSL}}{6\sigma}+\frac{2\mu}{6\sigma}=\frac{-\text{LSL}+\mu}{3\sigma}\\ &= \frac{\mu-\text{LSL}}{3\sigma}:=\frac{1}{3}Z_{ll}\end{aligned} \tag{16.7.56}$$

and when $T < \mu$:

$$\begin{aligned}C_p(1-k) &= C_p\left[1-\left(\frac{2\mu}{\text{USL}-\text{LSL}}-\frac{\text{USL}+\text{LSL}}{\text{USL}-\text{LSL}}\right)\right]\\ &= \frac{\text{USL}-\text{LSL}}{6\sigma}\left[1-\left(\frac{2\mu}{\text{USL}-\text{LSL}}-\frac{\text{USL}+\text{LSL}}{\text{USL}-\text{LSL}}\right)\right]\\ &= \frac{\text{USL}-\text{LSL}}{6\sigma}-\frac{2\mu}{6\sigma}+\frac{\text{USL}+\text{LSL}}{6\sigma}=\frac{\text{USL}-\mu}{3\sigma}:=\frac{1}{3}Z_{ul}\end{aligned} \tag{16.7.57}$$

■ Q.E.D.

In the long term, in some companies (firms), it is interesting to know what are the worst values taken by the CPU indices and CPL (this is the case in the field of production but not necessarily in project management).

The worst values being by construction the smallest one, we often write (with a few different notations that can be found in the literature ...):

$$C_{pk}=\min\{\text{CPU},\text{CPL}\}=\min\{C_{pu},C_{pl}\}=\min\left\{\frac{1}{3}Z_{ul},\frac{1}{3}Z_{ll}\right\} \tag{16.7.58}$$

Below an example of an analysis diagram of the capability generated by the software Minitab 15.1.1 Software with the various aforementioned factors on a sample of 68 measures that can not be rejected as not following a Normal distribution (a normality test was done before):

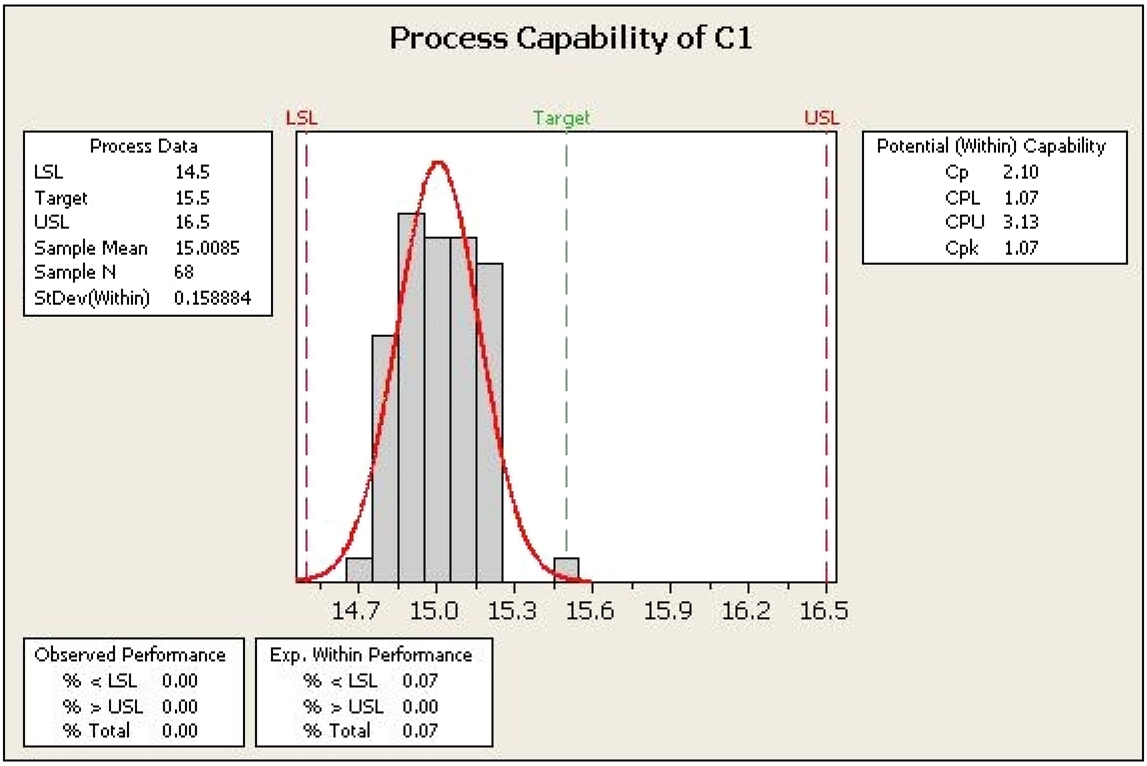


Figure 16.200: **Capability diagram generated with Minitab 15.1.1**

Two typical readings of this chart are possible (we will explain the lower left part of the chart later):

1. In manufacturing: The process is capable (value >1.33) but with a (too) strong deviation to the left in relation to the defined target, which is not good (CPL having the smallest value) and must be corrected.
2. In Project Management: Redundant tasks are under control (value>1.33) but with a strong deviation to the left, which can be good if our goal is to get ahead compared to planning (nothing to correct).

We really have to take care of the fact that in reality it is not always possible to take the Normal distribution (we recall this because all the examples given above and below are based on this simplifying assumption!)!

Always in the context of quality management in manufacturing, the figure below represents well the reality in the context of a short or long-term process:

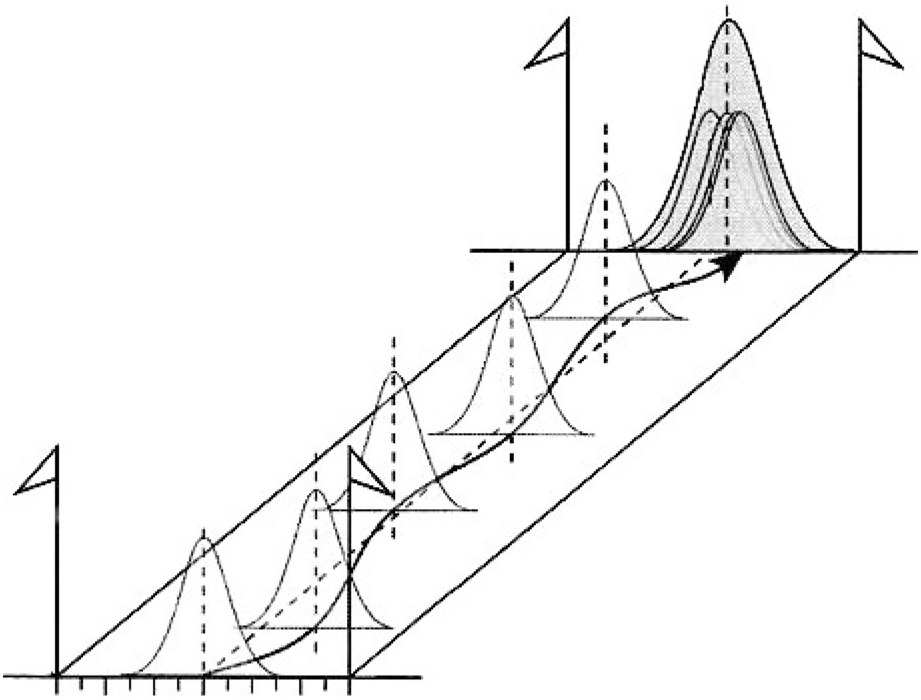


Figure 16.201: **Short/Long term process (source: Maurice Pillet ISBN13: 978-2708133495)**

Every little Gaussian in light gray, represents a batch analysis assimilated to the concept of "instantaneous dispersion". We see that their average don't keep moving during the measurement period (that this variation is large or very small) and that's what we name the "global dispersion". The goal in organizations (industries or administrations) is to ensure that the instantaneous or global variability is limited as much as possible.

But the relation defining C_p assumed, as we mentioned, that the process is under control and also centered (so all Gaussian are aligned) and on a short term.

Similarly, the relation defining C_{pk} assumed, as we mentioned it before, the that the process is under control, on a short-term perspective and non-centered by choice (or because of the fact that the law is not Normal).

If the process is not centered because it is not under control when it should be, the random variable measured could be the sum of the small random behaviour of the machine X and uncontrollable small random variations of constraint measurement of the pieces Y.

The total standard deviation is then, if the two random variables follow a Normal distribution (and that they are independent!), the square root of the sum of deviations (see section Statistics page 709):

$$\sigma=\sqrt{\sigma_X^2+\sigma_Y^2}=\sqrt{\sigma_{\text{ST}}^2+\sigma_Y^2} \tag{16.7.59}$$

Now if we only have one measure, we get taking the unbiased estimator (it's a little bit stupid to use it in this case but...):

$$\sigma_Y^2=\frac{1}{n}(Y-\mu_Y)^2=(Y-\mu_Y)^2 \tag{16.7.60}$$

But in the case of study that interests Y represents the experimental mean (measured) of the process we are trying to bring under control (this is also why we can put $n=1$). This average is traditionally denoted m in the field of industrial engineering.

Then μ_Y being not known we take what it should be: that is the target T in the process. Thus, we introduce a new index named "short term non-centered biased process potential capability":

$$\boxed{C_{pm}=\frac{\text{USL}-\text{LSL}}{6\sigma}\cong\frac{\text{USL}-\text{LSL}}{6\sqrt{\sigma^2+(m-T)^2}}} \tag{16.7.61}$$

where once again it must be remembered that the standard deviation in the denominator inside the root is the short term standard deviation of the machine!

We then see immediately that the more C_{pm} is close to C_p, the better it is (at least in manufacturing field)!

We finally have three common centered and non-centered short term capabilities indices (we deliberately chose to standardize the notations and put as much information inside the relations below):

$$\boxed{C_p=\frac{\text{USL}-\text{LSL}}{\sigma_{\text{ST}}}\qquad C_{pk}=\left\{\frac{\text{USL}-\mu}{3\sigma_{\text{ST}}},\frac{\mu-\text{LSL}}{3\sigma_{\text{ST}}}\right\}\qquad C_{pm}\cong\frac{\text{USL}-\text{LSL}}{6\sqrt{\sigma_{\text{ST}}^2+(m-T)^2}}} \tag{16.7.62}$$

Similarly we also have three common long-term capabilities indicators centered or non-centered (we deliberately chose to standardize the notations and put as much information inside the relations below):

$$\boxed{P_p=\frac{\text{USL}-\text{LSL}}{\sigma_{\text{LT}}}\qquad P_{pk}=\left\{\frac{\text{USL}-\mu}{3\sigma_{\text{LT}}},\frac{\mu-\text{LSL}}{3\sigma_{\text{LT}}}\right\}\qquad P_{pm}\cong\frac{\text{USL}-\text{LSL}}{6\sqrt{\sigma_{\text{LT}}^2+(m-T)^2}}} \tag{16.7.63}$$

Finally, it is good to know that although if this is not very relevant, sometimes some engineers do both analyses (short term + long-term) at the same time on the same measurements.

Remark

Let us indicate the capabilities of an industrial process methods when applied to machines are denoted respectively by C_m, C_{mk}, C_{mm}. Anyway ... for more information (but without proofs) refer to the standard ISO 22514-2:2013 or the old version ISO 21747:2006.

However, to make an objective analysis on the indexes of capability seen so far, the practitioner would first have to check measuring instruments are themselves capable... what we often name the "R&R methods" (Repeatability, Reproducibility).

The basic principle (a little bit more advanced principle is to make use of the two-factor ANOVA with repetition as studied in details in the section of Statistics) is then to evaluate the short term or respectively long term dispersion of the measuring instrument to calculate a "process control capability" defined by:

$$C_{pc}=\frac{E}{6\sigma_{\text{instrument}}} \tag{16.7.64}$$

or according to the type of industry some use the standard deviation based on the range for $\sigma_{\text{instrument}}$ and as we proved it in the section of Statistics during our study of extreme values (see page 1203) and therefore named "Equipment Variation (EV)" consisting in the calculation of the average of ranges of all measurements made by different operators and divide by the Hartley's constant:

$$\text{EV}=\frac{\bar{R}}{d_2(n)} \tag{16.7.65}$$

In the classical cases, we declare the control method as capable for a S.P.C. tracking when its capability is greater than 4 and we will immediately see why.

Recall for this first that:

$$C_p=\frac{E}{6\sigma} \tag{16.7.66}$$

But the variance observed is in fact the sum of the "true" variance and this of the instrument such that:

$$C_{p,\text{observed}}=\frac{E}{6\sigma}=\frac{E}{6\sqrt{\sigma_{\text{true}}^2+\sigma_{\text{instrument}}^2}} \tag{16.7.67}$$

But we have:

$$C_{p,\text{true}}=\frac{E}{6\sigma_{\text{true}}} \qquad C_{pc}=\frac{E}{6\sigma_{\text{instrument}}} \tag{16.7.68}$$

Putting it all squared, we deduce:

$$C_{p,\text{observed}}^2=\frac{E^2}{36(\sigma_{\text{true}}^2+\sigma_{\text{instrument}}^2)}=\frac{E^2}{36\left(\dfrac{E^2}{36C_{p,\text{true}}^2}+\dfrac{E^2}{36C_{pc}^2}\right)} \tag{16.7.69}$$

And now under the strong assumptions that all E are equals we have:

$$C_{p,\text{observed}}^2=\frac{1}{\dfrac{1}{C_{p,\text{true}}^2}+\dfrac{1}{C_{pc}^2}}=\frac{1}{\dfrac{C_{pc}^2+C_{p,\text{true}}^2}{C_{pc}^2C_{p,\text{true}}^2}} \tag{16.7.70}$$

Therefore:

$$C_{p,\text{observed}}^2=\frac{C_{pc}^2C_{p,\text{true}}^2}{C_{pc}^2+C_{p,\text{true}}^2} \tag{16.7.71}$$

That give us with some rearrangements:

$$C_{p,\text{observed}}^2(C_{pc}^2+C_{p,\text{true}}^2)=C_{pc}^2C_{p,\text{true}}^2 \Leftrightarrow C_{p,\text{observed}}^2C_{pc}^2=C_{p,\text{true}}^2(C_{pc}^2-C_{p,\text{observed}}^2) \tag{16.7.72}$$

Finally:

$$C_{p,\text{true}}=\frac{C_{p,\text{observed}}C_{pc}}{\sqrt{C_{pc}^2-C_{p,\text{observed}}^2}}=\frac{C_{p,\text{observed}}}{\sqrt{1-\dfrac{C_{p,\text{observed}}^2}{C_{pc}^2}}} \tag{16.7.73}$$

This can be translated into a plot in the figure below which shows the interest of a C_{pc} of at least equal to 4!

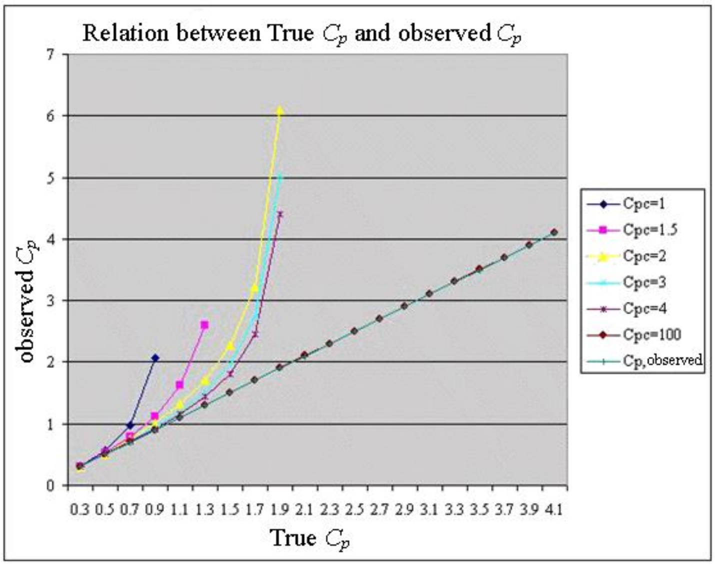


Figure 16.202: **Relations between $C_{p,\text{observed}}$ and $C_{p,\text{true}}$ for fixed C_{pc}**

In practice, note that to determine C_{pc} we use a standard prototype measured by LASER interferometry and then we ensure that all repeated measurement take place on the same two measuring points.

Once this done, we performs several measures of the standard prototype and we take the standard deviation of these measures. This will give the $\sigma_{\text{instrument}}$.

The range E is imposed by the customer or by internal engineers to the company. It will often be taken as maximum to the tenth of the tolerance unit of the standard prototype.

For example, if we have a piece with an internal diameter of $36 \pm 1\,[\mu\text{m}]$ (tolerance range of $2\,[\mu\text{m}]$ which is already a good level precision because in our time the standard is closer to $3\,[\mu\text{m}]$!), our device will then need to have according the rule cited above range of $0.2\,[\mu\text{m}]$...

16.7.1.4 Quality Levels

Some engineers, as we already mention it, like to know how many elements by million units produced (parts per million: PPM) will be considered as defective.

The calculation is then easy since the engineer has at his disposal at least the following information:

$$\mu, \sigma, \text{USL}, \text{LSL}, T \tag{16.7.74}$$

and under the strong assumption that the data follow a Normal distribution and that the target is centered on the mean it is immediate (see section Statistics page 692) in this simple case the for the defective below the LSL we get (for long term or short term):

$$\text{PPM}_{\text{LSL}} = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{\text{LSL}} e^{\frac{(x-\mu)^2}{2\sigma}} \mathrm{d}x \cdot 10^6 \tag{16.7.75}$$

and:

$$\text{PPM}^{\text{USL}} = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{\text{USL}}^{+\infty} e^{\frac{(x-\mu)^2}{2\sigma}} \mathrm{d}x \cdot 10^6 = \left(1 - \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{\text{USL}} e^{\frac{(x-\mu)^2}{2\sigma}} \mathrm{d}x\right) \cdot 10^6 \tag{16.7.76}$$

values very easy to get with any spreadsheet like Microsoft® Excel® for example. We have therefore obviously:

$$\boxed{\text{PPM}_{\text{tot}} = \text{PPM}_{\text{LSL}} + \text{PPM}^{\text{USL}} = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{\text{LSL}} e^{\frac{(x-\mu)^2}{2\sigma}} \mathrm{d}x \cdot 10^6 + \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{\text{USL}}^{+\infty} e^{\frac{(x-\mu)^2}{2\sigma}} \mathrm{d}x \cdot 10^6} \tag{16.7.77}$$

If μ and T are still aligned and that USL and LSL are symmetric to this two values we have the special case:

$$\text{PPM}_{\text{tot}} = \text{PPM}_{\text{LSL}} + \text{PPM}^{\text{USL}} = 2\text{PPM}_{\text{LSL}} = 2\text{PPM}^{\text{USL}} \tag{16.7.78}$$

Note now an important point relatively to Six Sigma! In fact, objectively, the idea of this method is, of course, to make S.P.C. (among others, but that's nothing really new) but especially to: guarantee to the customer according to tradition with a commonly accepted standard deviation having an upper limit of $\sigma = 1$ with an absolute deviation to the mean (in absolute value) of 1.5σ relative to the target, which guarantees at maximum the famous 3.4 PPM (that is to say 3.4 defectives per million).

Remark

This choice seems to come from the empirical practice of Six Sigma by one of its main creator (Bill Smith). It is said that he observed in his business (Motorola) that under statistical control, it was almost always a deviation ranging between 1.2 and 1.8σ of the average for almost all its industrial processes.

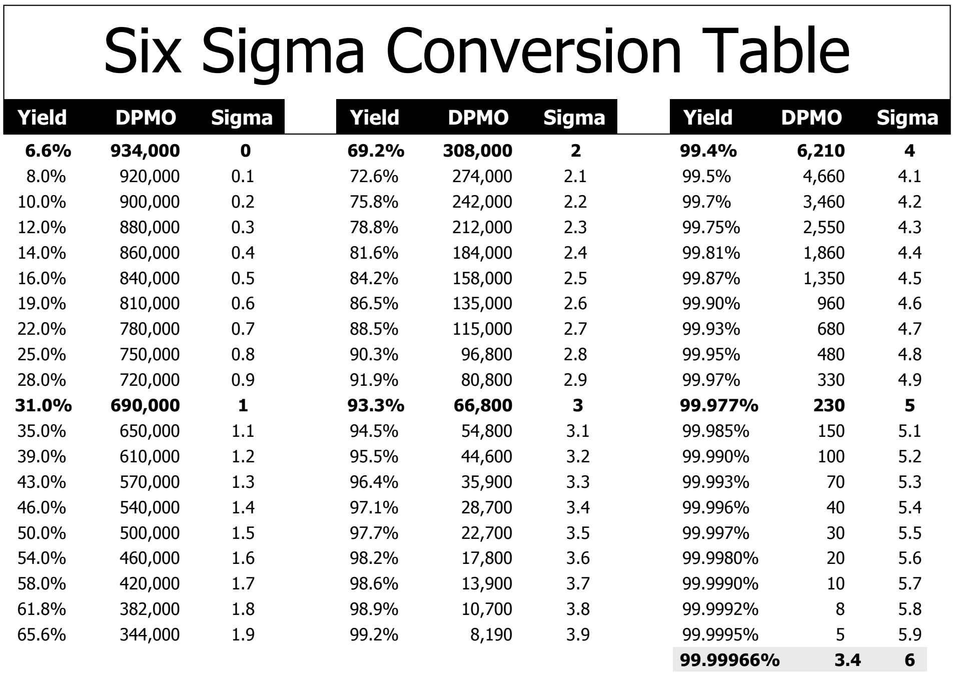
Let's see where this value comes from with the following two tables:

1. First we construct an ideal type of table presenting data from a short-term process (but the calculations are identical for a long-term process) centered on the target (target will be zero here, which is a typical case), with a zero mean (i.e. on target and therefore $C_p = C_{pk}$) and unit standard deviation with USL and LSL symmetrical (which against restricted by the scope of application):

C_p	C_{pk}	**Defectives (PPM)**	**Sigma Quality Level**	**Criteria**
0.5	0.5	133,614	1.5	Bad
0.6	0.6	71,861	1.8	
0.7	0.7	35,729	2.1	
0.8	0.8	16,395	2.4	
0.9	0.9	6,934	2.7	
1	1	2,700	3	
1.1	1.1	967	3.3	
1.2	1.2	318	3.6	
1.3	1.3	96	3.9	Limit
1.4	1.4	27	4.2	
1.5	1.5	6.8	4.5	
1.6	1.6	1.6	4.8	
1.7	1.7	0.34	5.1	
1.8	1.8	0.067	5.4	
1.9	1.9	0.012	5.7	
2	2	0.002	6	Excellent

Table 16.13: Defectives and Sigma quality level of a centered process

where all the values are obtained using the following relations from the potential capability index only. Another way to present the above table is:



Six Sigma Conversion Table

Yield	DPMO	Sigma	Yield	DPMO	Sigma	Yield	DPMO	Sigma
6.6%	**934,000**	**0**	**69.2%**	**308,000**	**2**	**99.4%**	**6,210**	**4**
8.0%	920,000	0.1	72.6%	274,000	2.1	99.5%	4,660	4.1
10.0%	900,000	0.2	75.8%	242,000	2.2	99.7%	3,460	4.2
12.0%	880,000	0.3	78.8%	212,000	2.3	99.75%	2,550	4.3
14.0%	860,000	0.4	81.6%	184,000	2.4	99.81%	1,860	4.4
16.0%	840,000	0.5	84.2%	158,000	2.5	99.87%	1,350	4.5
19.0%	810,000	0.6	86.5%	135,000	2.6	99.90%	960	4.6
22.0%	780,000	0.7	88.5%	115,000	2.7	99.93%	680	4.7
25.0%	750,000	0.8	90.3%	96,800	2.8	99.95%	480	4.8
28.0%	720,000	0.9	91.9%	80,800	2.9	99.97%	330	4.9
31.0%	**690,000**	**1**	**93.3%**	**66,800**	**3**	**99.977%**	**230**	**5**
35.0%	650,000	1.1	94.5%	54,800	3.1	99.985%	150	5.1
39.0%	610,000	1.2	95.5%	44,600	3.2	99.990%	100	5.2
43.0%	570,000	1.3	96.4%	35,900	3.3	99.993%	70	5.3
46.0%	540,000	1.4	97.1%	28,700	3.4	99.996%	40	5.4
50.0%	500,000	1.5	97.7%	22,700	3.5	99.997%	30	5.5
54.0%	460,000	1.6	98.2%	17,800	3.6	99.9980%	20	5.6
58.0%	420,000	1.7	98.6%	13,900	3.7	99.9990%	10	5.7
61.8%	382,000	1.8	98.9%	10,700	3.8	99.9992%	8	5.8
65.6%	344,000	1.9	99.2%	8,190	3.9	99.9995%	5	5.9
						99.99966%	**3.4**	**6**

First (remember the assumption that $T=\mu$ and $\sigma=1$:

$$\begin{aligned}\text{CPU}&=\frac{\text{USL}-(T+\mu)}{3\sigma}=\frac{\text{USL}-(0+0)}{3\sigma}=\frac{\text{USL}}{3\sigma}=\frac{\text{USL}}{3}\\ \text{CPL}&=\frac{(T+\mu)-\text{LSL}}{3\sigma}=\frac{(0+0)-\text{LSL}}{3\sigma}=-\frac{\text{LSL}}{3\sigma}=-\frac{\text{LSL}}{3}\end{aligned}\tag{16.7.79}$$

if the standard deviation is reduced (that can always be done and does not change the accuracy of the results!). And since in the above table LSL and USL are symmetrical with respect to the target:

$$C_{pk}\underset{T=\mu=0}{=}C_p=\min\{\text{CPU},\text{CPL}\}\Rightarrow\text{LSL}=-\text{USL}=-3C_p\tag{16.7.80}$$

and PPM are conforming with what we saw just before given by:

$$\begin{aligned}\text{PPM}_{\text{tot}}&=\text{PPM}_{\text{LSL}}+\text{PPM}^{\text{USL}}\\ &=\left[\frac{1}{\sigma\sqrt{2\pi}}\int\limits_{-\infty}^{\text{LSL}}e^{\frac{(x-\mu)^2}{2\sigma}}\mathrm{d}x+\left(1-\frac{1}{\sigma\sqrt{2\pi}}\int\limits_{-\infty}^{\text{USL}}e^{\frac{(x-\mu)^2}{2\sigma}}\mathrm{d}x\right)\right]10^6\end{aligned}\tag{16.7.81}$$

and therefore as in the example above LSL and USL are symmetric with respect to the target it simplifies to:

$$\text{PPM}_{\text{tot}}=2\left(1-\frac{1}{\sigma\sqrt{2\pi}}\int\limits_{-\infty}^{3C_p}e^{\frac{(x-\mu)^2}{2\sigma}}\mathrm{d}x\right)\cdot10^6\tag{16.7.82}$$

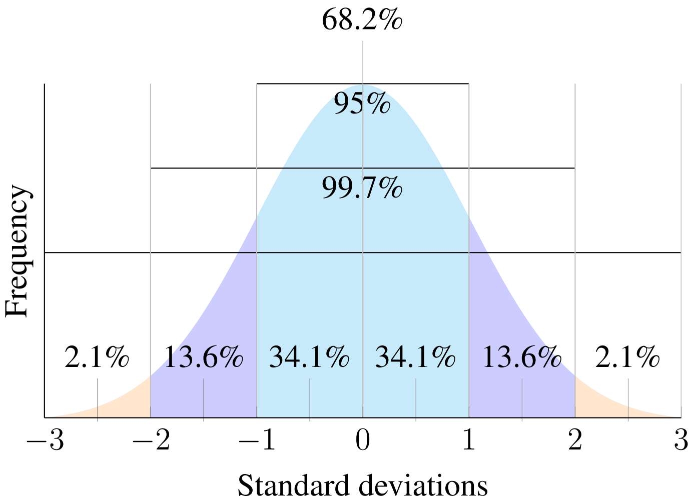
where, for example, the value of PPM given at the line "Limit" is obtained with Maple™ 4.00b using the following command:

```
>evalf((1-1/sqrt(2*Pi)*int(exp(-x^ 2/2),x=-infinity..3.9))*2)*1E6;
```

or with Microsoft® Excel® 11.8346:

```
=(1-NORMDIST(3.9,0,1,1))*1E6
```

Let us recall that the "sigma quality level" denoted by σ_q is given in fact with the following table we had built in the section Statistics under the assumption of a Normal law (thanks to John Cannin for the LaTeX figure):



Sigma Quality Level	Non-defective assured rate in %	Defective in PPM
$1\sigma_q$	68.26894	317,311
$2\sigma_q$	95.4499	45,500
$3\sigma_q$	99.73002	2,700
$4\sigma_q$	99.99366	63.4
$5\sigma_q$	99.999943	0.57
$6\sigma_q$	99.9999998	0.002

Table 16.14: **Capability, non-defective rate in% and PPM**

and for which we had given the Maple™ 4.00b command to obtain the values that are valid for all standard deviations and all means!

2. Now let us build the table at worst as considered by Six Sigma, that is to say a table for non-centered process (that is to say where $C_p = C_{pk}$ is not satisfied) with a deviation of the average of $+1.5\sigma$ (so on the right but could be taken as shifted on left and the results would be exactly the same!) against the target and unit standard deviation with symmetrical USL and LSL (which still restricts the scope of application):

C_p	C_{pk}	Defectives (PPM)	Sigma Quality Level	Criteria
0.5	0	501,350	1.5	Bad
0.6	0.1	382,572	1.8	
0.7	0.2	274,122	2.1	
0.8	0.3	184,108	2.4	
0.9	0.4	115,083	2.7	
1	0.5	66,810	3	
1.1	0.6	35,931	3.3	
1.2	0.7	17,865	3.6	
1.3	0.8	8,198	3.9	Limit
1.4	0.9	3,467	4.2	
1.5	1	1,350	4.5	
1.6	1.1	483	4.8	
1.7	1.2	159	5.1	
1.8	1.3	48	5.4	
1.9	1.4	13	5.7	
2	1.5	3.4	6	Excellent

Table 16.15: **Defectives and Sigma quality level of a non-centered process**

where all the values are obtained using the following relations from the potential capability index only. First (remember the assumption that $T = \mu$ and $\sigma = 1$:

$$\begin{aligned}\text{CPU} &= \frac{\text{USL} - (T + 1.5\sigma)}{3\sigma} = \frac{\text{USL}}{3\sigma} - 0.5 = \frac{\text{USL}}{3} - 0.5 = C_p - 0.5\\ \text{CPL} &= \frac{(T + 1.5\sigma) - \text{LSL}}{3\sigma} = \frac{T - \text{LSL}}{3\sigma} + 0.5 = -\frac{\text{LSL}}{3} + 0.5 = C_p + 0.5\end{aligned} \tag{16.7.83}$$

if the standard deviation is reduced (that can always be done and does not change the accuracy of the results!). And since in the above table LSL and USL are symmetrical with respect to the target:

$$C_{pk}=\min\{\text{CPU},\text{CPL}\}\Rightarrow \text{LSL}=-\text{USL}=C_p-0.5 \tag{16.7.84}$$

Therefore:

$$\begin{aligned}\text{USL}&=3C_p\\ \text{LSL}&=-3(C_p+1)\end{aligned} \tag{16.7.85}$$

and PPM are conforming with what we saw just before given by:

$$\begin{aligned}\text{PPM}_{\text{tot}}&=\frac{1}{\sigma\sqrt{2\pi}}\int\limits_{-\infty}^{\text{LSL}}e^{\frac{(x-\mu)^2}{2\sigma}}\mathrm{d}x\cdot 10^6+\left(1-\frac{1}{\sigma\sqrt{2\pi}}\int\limits_{-\infty}^{\text{USL}}e^{\frac{(x-\mu)^2}{2\sigma}}\mathrm{d}x\right)\cdot 10^6\\ &=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{\text{LSL}}e^{\frac{(x-\mu)^2}{2}}\mathrm{d}x\cdot 10^6+\left(1-\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{\text{USL}}e^{\frac{(x-\mu)^2}{2}}\mathrm{d}x\right)\cdot 10^6\\ &=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{-3(C_p+1)}e^{\frac{(x-1.5\sigma)^2}{2}}\mathrm{d}x\cdot 10^6+\left(1-\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{3C_p}e^{\frac{(x-1.5\sigma)^2}{2}}\mathrm{d}x\right)\cdot 10^6\\ &=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{-3(C_p+1)}e^{\frac{(x-1.5)^2}{2}}\mathrm{d}x\cdot 10^6+\left(1-\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{3C_p}e^{\frac{(x-1.5)^2}{2}}\mathrm{d}x\right)\cdot 10^6\end{aligned} \tag{16.7.86}$$

where, for example, the value of PPM given at the line "Limit" is obtained with Maple™ 4.00b using the following command:

```
>evalf((1-1/sqrt(2*Pi)*int(exp(-(x-1.5)^ 2/2),x=-infinity..(1.3*3))))*1E6
+evalf((1/sqrt(2*Pi)*int(exp(-(x-1.5)^ 2/2),x=-infinity..-(3*(1.3+1)))))*1E6;
```

or with Microsoft® Excel® 11.8346:

```
=(((1-NORMDIST(3*1.3,1.5,1,1))+NORMDIST(-3*(1.3+1),1.5,1,1)))*1E6
```

We finally understand seeing this famous "limit" row, why a process under control is said to be "limit capable" with a potential capability index of 1.33 potential given the number of PPM!

So the goal in practice is obviously to be in the situation of the first table with the corresponding values and with in the first table at a quality level of $\sim 4.7\sigma$ for the equivalent of 3.4 PPM of the second table (because it is easier to center a process than to control its variations).

✍**Example:**

Let us now make a summary of all this considering a new small production of $n=50$ pieces in packs of $k=n_i=10$ (in order to adjust the machines during production). Measuring the specifications of 5 pieces per hour during 10 hours with a tolerance of 10 ± 0.07 that is to say in terms of a range of hundredths:

$$E=2\cdot 7=14 \tag{16.7.87}$$

and a target of $T=0$. We have the data:

	1	2	3	4	5	6	7	8	9	10
1	-2	-4	-1	0	4	0	3	0	1	-1
2	0	-3	0	-2	1	-2	0	1	-1	2
3	-1	0	-3	-1	0	0	-1	-1	3	1
4	1	1	-2	2	2	0	1	0	4	0
5	-1	-1	-3	0	0	3	3	2	1	0
μ	-0.6	-1.4	-1.8	-0.2	1.4	0.2	1.2	0.4	1.6	0.5
σ	1.14	2.07	1.30	1.48	1.67	1.79	1.79	1.14	1.95	1.14

Table 16.16: Application of statistical process control analysis

We immediately see that the manufacturing process was not stationary during this first production, we will therefore need to make corrections in the future:

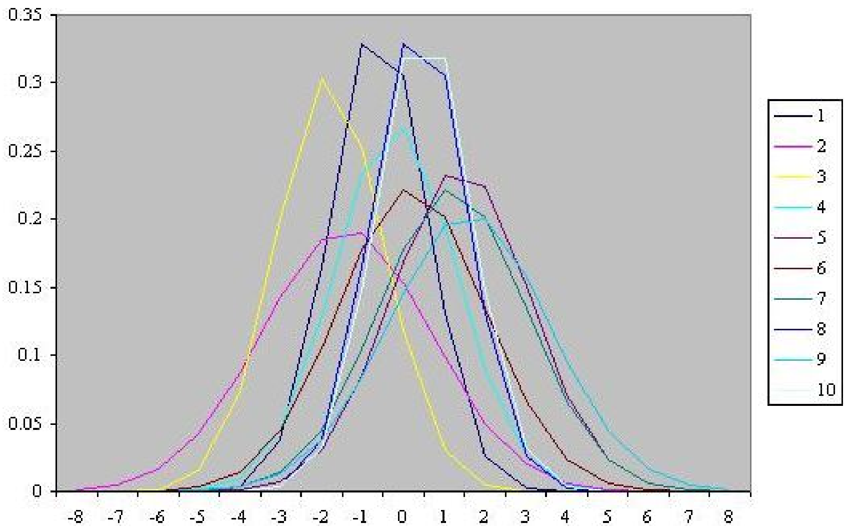


Figure 16.203: Visual evidence by a plot that the process seems not stationary

or as a very simple control average-standard deviation chart with 3σ (as I like do it for beginners):

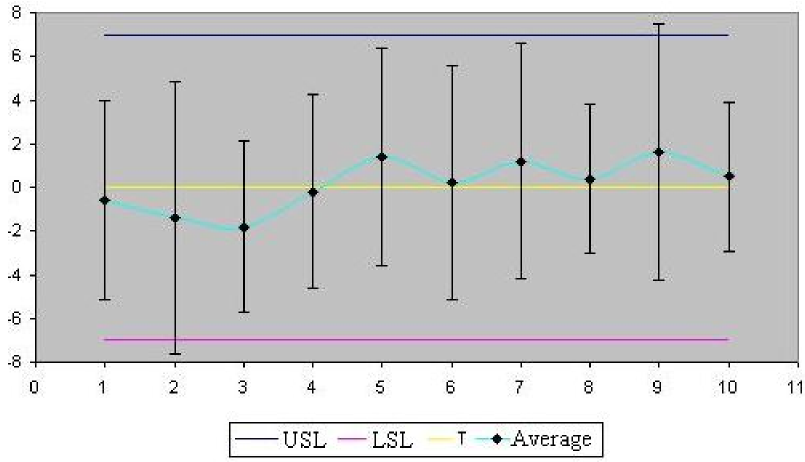


Figure 16.204: **Visual evidence by a plot that the process seems not stationary**

So we easily guess that the process is limit capable...

Remark

An interesting thing to know is that this type of chart can be analysed using the mathematical tools of the time series analysis as we will see later.

First, if we want to make a relevant statistical study of the different data above we can calculate the average of deviations that under the assumption of a Normal distribution is the arithmetic mean (see section Statistics page 819):

$$\bar{X} = 0.120 \tag{16.7.88}$$

Then the standard deviation of all pieces taken as a single group is using the maximum likelihood estimator of the variance of the Normal distribution (long term standard deviation) and taking the some notation for this variance as the one used in the section of Statistics:

$$S_{\text{LT}} = \sqrt{\sum_{i,j} \frac{(X_{i,j} - \bar{X})^2}{n-1}} = \hat{\sigma}_{\text{LT}} = 1.80294 \tag{16.7.89}$$

So the standard error (the estimator of the standard error of the mean) is as proved in the section Statistics (page 611):

$$\sigma_\mu = \frac{S_{\text{LT}}}{\sqrt{n}} = 0.25 \tag{16.7.90}$$

So the confidence interval to 95% of the average is (see section Statistics page 855):

$$\bar{X} - 1.96\frac{S_{\text{LT}}}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{S_{\text{LT}}}{\sqrt{n}} \tag{16.7.91}$$

In our case this gives:

$$-0.379 \leq \mu \leq 0.619 \tag{16.7.92}$$

And the statistical inference with our long-term standard deviation using the chi-square χ^2 bilateral hypothesis test gives (see section Statistics page 862):

$$\frac{(n-1)S_{\text{LT}}^2}{\chi^2_{\alpha/2}(n-1)} \leq \sigma_{\text{LT}}^2 \leq \frac{(n-1)S_{\text{LT}}^2}{\chi^2_{1-\alpha/2}(n-1)} \tag{16.7.93}$$

What gives us in our case:

$$2.268 \leq \sigma_{\text{LT}}^2 \leq 5.047 \tag{16.7.94}$$

Therefore:

$$1.506 \leq \sigma_{\text{LT}} \leq 2.246 \tag{16.7.95}$$

We notice therefore that on a long-term analysis we have the intervals:

$$\begin{gathered} -0.644 \leq \mu \leq 0.884 \\ 1.506 \leq \sigma_{\text{ST}} \leq 2.246 \end{gathered} \tag{16.7.96}$$

The variations can therefore be huge with a cumulative probability of 95% and we must then take care in practical cases to make adjustments as quickly as possible to reduce as possible the values of those moments!

Let us now calculate the potential long-term process capability (if assumed centered!). We have:

$$P_p = \frac{\text{USL}-\text{LSL}}{6S_{\text{LT}}} = \frac{\text{USL}-\text{LSL}}{6\sigma_{\text{LT}}} \cong \frac{14}{6\cdot 1.8} \cong 1.294 \tag{16.7.97}$$

But with an instrument having P_{pc} of 4, it really corresponds to:

$$P_{p,\text{true}} = \frac{P_{p,\text{observed}}}{\sqrt{1-\dfrac{P_{p,\text{observed}}}{P_{pc}}}} = 1.574 \tag{16.7.98}$$

Furthermore, let us indicate that if we can do a calculation of the confidence interval for σ_{LT} (see calculations done previously), it is then easy to have one too for P_p!

If the analysis of the overall performance of the long-term process is not centered (which is the case here) we use:

$$P_{pk} = \min\left\{\frac{\text{USL}-\bar{X}}{3S_{\text{LT}}}, \frac{\bar{X}-\text{LSL}}{3S_{\text{LT}}}\right\} = \min\left\{\frac{7-0.12}{3\cdot 1.80}, \frac{0.12-(-7)}{3\cdot 1.80}\right\} = \min\{1.274, 1.318\} = 1.274 \tag{16.7.99}$$

and we know once again that because of the instrument, this value is somewhat undervalued!

We have of course:

$$P_{pk} \neq P_p \tag{16.7.100}$$

So the process is not centered (as we suspected...). Then, we must calculate the potential long-term average non-centered capability P_{pm} of the process using the relations defined above:

$$P_{pm} \cong 1.291 \qquad (16.7.101)$$

To resume, either the value of P_{pk}, P_p or P_{pm}, we see that all the values are limit capable (that is to say that the value is greater than 1 - see definition above for a reminder of what means "limit capable").

If we do then our calculations of PPM according to the relations obtained above with the value of S_{LT} and $\bar{X}$ previously obtained then we have:

$$\text{PPM}_{\text{tot}}^{\text{LT}} = \text{PPM}^{UL} + \text{PPM}_{LL} \cong 114 \qquad (16.7.102)$$

Then say this number is good or bad is difficult because we lack the information to know what is the cost of production, cost of repair, the cost of a product and the whole is itself dependent of the total amount produced! But we can also use the model of Taguchi (see further below) to know the value of the parameters (moments) calculated that it would be preferable not to exceed!

Let us now calculate the short term capability indices! For this, we need the estimator of the average of the whole considering each individual as a random variable. We know (see section Statistics page 819) that this average is the arithmetic mean in the case of a Normal distribution and is strictly equal to that which is calculated by considering the set of individuals as a single random variable. So it follows that:

$$\bar{X} = 0.120 \qquad (16.7.103)$$

Regarding the standard deviation this is not the same! But we know (see section Statistics page 710) that the Normal law that is stable by the sum. For example, we have proved that given two independent random variables distributed according to a Normal distribution (by imagining that each variable represents two of our ten samples), we hat for their sum:

$$X + Y = \mathcal{N}\left(\mu_{X+Y}, \sqrt{\sigma_X^2 + \sigma_X^2}\right) \qquad (16.7.104)$$

But we have also proved in the section of Statistics using the property of linearity of expected mean, that we have:

$$\text{E}(Z) = \text{E}\left(\frac{X+Y}{2}\right) = \frac{1}{2}\text{E}(X+Y) = \frac{1}{2}(\mu_X + \mu_Y) = \frac{\mu_X + \mu_Y}{2} \qquad (16.7.105)$$

and for the variance:

$$\text{V}(Z) = \text{V}\left(\frac{X+Y}{2}\right) = \frac{1}{4}\text{V}(X+Y) = \frac{1}{4}\sqrt{\sigma_X^2 + \sigma_X^2} = \sqrt{\frac{\sigma_X^2 + \sigma_X^2}{2}} \qquad (16.7.106)$$

Therefore:

$$S_{\text{ST}} = \sqrt{\frac{\displaystyle\sum_{i=1}^{k} S_i^2}{k}} \qquad (16.7.107)$$

and in our special case:

$$S_{\text{ST}}=\sqrt{\frac{\sum\limits_{i=1}^{10}S_i^2}{10}}=1.5842979 \tag{16.7.108}$$

Remark

A software like Minitab used the pooled standard deviation (see section Statistics page 1011) that for recall is given by:

$$S_{\text{ST}}=\frac{\sum\limits_{i,j}(X_{i,j}-\bar{X})^2}{\sum_i(n_i-1)}=\frac{\sum\limits_{i,j}(X_{i,j}-0.120)^2}{10\cdot(5-1)}=1.5842979 \tag{16.7.109}$$

But it's still not accurate enough. As we will prove it further below during our study of control charts on page 8980, by default (but there is an option to deactivate this) Minitab use unbiasing constants such that finally:

$$\sigma_{\text{ST}}=\frac{S_{\text{ST}}}{c_4(k)} \tag{16.7.110}$$

So the standard error (the estimator of the standard error of the mean) is:

$$\sigma_\mu=\frac{S_{\text{ST}}}{\sqrt{k}}\cong 0.708 \tag{16.7.111}$$

So the confidence interval to 95% of the average is (see section Statistics page 852):

$$\bar{X}-1.96\frac{S_{\text{ST}}}{\sqrt{k}}\leq\mu\leq\bar{X}+1.96\frac{S_{\text{ST}}}{\sqrt{k}} \tag{16.7.112}$$

In our case this gives:

$$-1.267\leq\mu\leq 1.507 \tag{16.7.113}$$

We note that in the short term, the range is much wider in the long term, which is normal given the low value of k (which is only 5 in our example).

And the statistical inference with our short-term standard deviation using the chi-square χ^2 bilateral hypothesis test gives (see section Statistics page 862):

$$\frac{(k-1)S_{\text{ST}}^2}{\chi^2_{\alpha/2}(k-1)}\leq\sigma_{\text{ST}}^2\leq\frac{(k-1)S_{\text{ST}}^2}{\chi^2_{1-\alpha/2}(k-1)} \tag{16.7.114}$$

What gives us in our case:

$$0.899\leq\sigma_{\text{ST}}^2\leq 20.692 \tag{16.7.115}$$

Therefore:

$$0.948\leq\sigma_{\text{ST}}\leq 4.549 \tag{16.7.116}$$

We notice therefore that on a short-term analysis we have the intervals:

$$-1.267 \leq \mu \leq 1.507 \quad (16.7.117)$$

$$0.948 \leq \sigma_{\text{ST}} \leq 4.549 \quad (16.7.118)$$

The variations can therefore be huge with a cumulative probability of 95% and we must then take care in practical cases to make adjustments as quickly as possible to reduce as possible the values of those moments!

Let us now calculate the potential long-term process capability (if assumed centered!). We have:

$$C_p = \frac{\text{USL}-\text{LSL}}{6S_{\text{ST}}} = \frac{\text{USL}-\text{LSL}}{6\sigma_{\text{ST}}} = \frac{14}{6\cdot 1.58} = 1.473 \quad (16.7.119)$$

So we have logically (the opposite would be problematic):

$$C_p > P_p \quad (16.7.120)$$

But with an instrument having C_{pc} of 4, it really corresponds to:

$$C_{p,\text{true}} = \frac{C_{p,\text{observed}}}{\sqrt{1-\dfrac{C_{p,\text{observed}}}{C_{pc}}}} = 1.854 \quad (16.7.121)$$

Furthermore, let us indicate that if we can do a calculation of the confidence interval for σ_{ST} (see calculations done previously), it is then easy to have one for C_p too!

If the analysis of the overall performance of the long-term process is not centered (which is the case here) we use:

$$C_{pk} = \min\left\{\frac{\text{USL}-\bar{X}}{3S_{\text{ST}}}, \frac{\bar{X}-\text{LSL}}{3S_{\text{ST}}}\right\} = \min\left\{\frac{7-0.12}{3\cdot 1.583}, \frac{0.12-(-7)}{3\cdot 1.583}\right\} = \min\{1.448, 1.499\} = 1.448 \quad (16.7.122)$$

and we know once again that because of the instrument, this value is somewhat undervalued! We have of course:

$$C_{pk} \neq C_p \quad (16.7.123)$$

So the process is not centered (as we suspected...). Then, we must calculate the potential long-term average non-centered capability C_{pm} of the process using the relations defined above:

$$C_{pm} \cong 1.296 \quad (16.7.124)$$

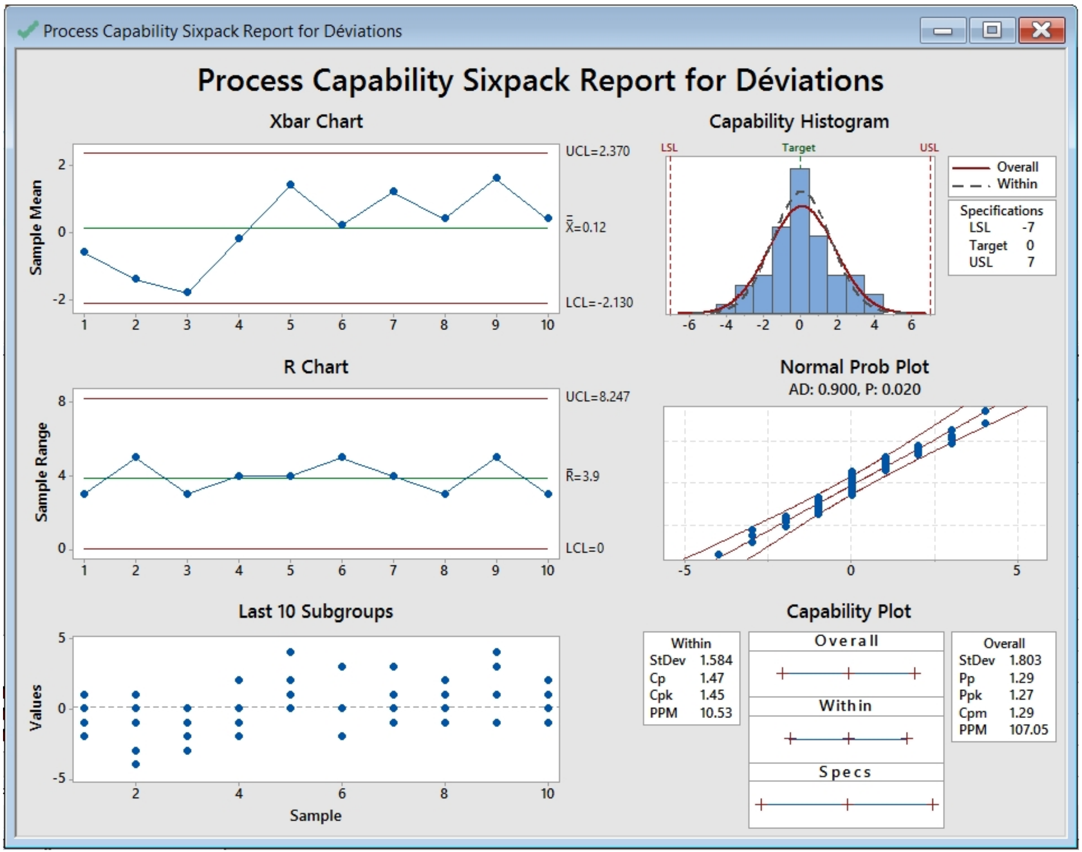
To resume, either the value of C_{pk}, C_p or C_{pm}, we see that all the values are limit capable (that is to say that the value is greater than 1 - see definition above for a reminder of what means "limit capable").

If we do then our calculations of PPM according to the relations obtained above with the value of S_{ST} and $\bar{X}$ previously obtained then we have:

$$\text{PPM}_{\text{tot}}^{\text{ST}}=\text{PPM}^{UL}+\text{PPM}_{LL}\cong 10 \tag{16.7.125}$$

With same remarks as for $\text{PPM}_{\text{tot}}^{\text{LT}}$.

To close this long detailed example, here is the output of a program like Minitab 17.3.1 in which we find all the calculations above plus the control charts which we will further detailed the proofs (see the section of Statistics at page 1130 for the details about AD Anderson-Darling test):


Process Capability Sixpack Report for Déviations
Process Capability Sixpack Report for Déviations
Xbar Chart
Sample Mean
UCL=2.370
X̄=0.12
LCL=-2.130
Capability Histogram
LSL
Target
USL
Overall
Within
R Chart
Sample Range
UCL=8.247
R̄=3.9
LCL=0
Normal Prob Plot
AD: 0.900, P: 0.020
Last 10 Subgroups
Values
Sample
Capability Plot
Overall
Within
Specs

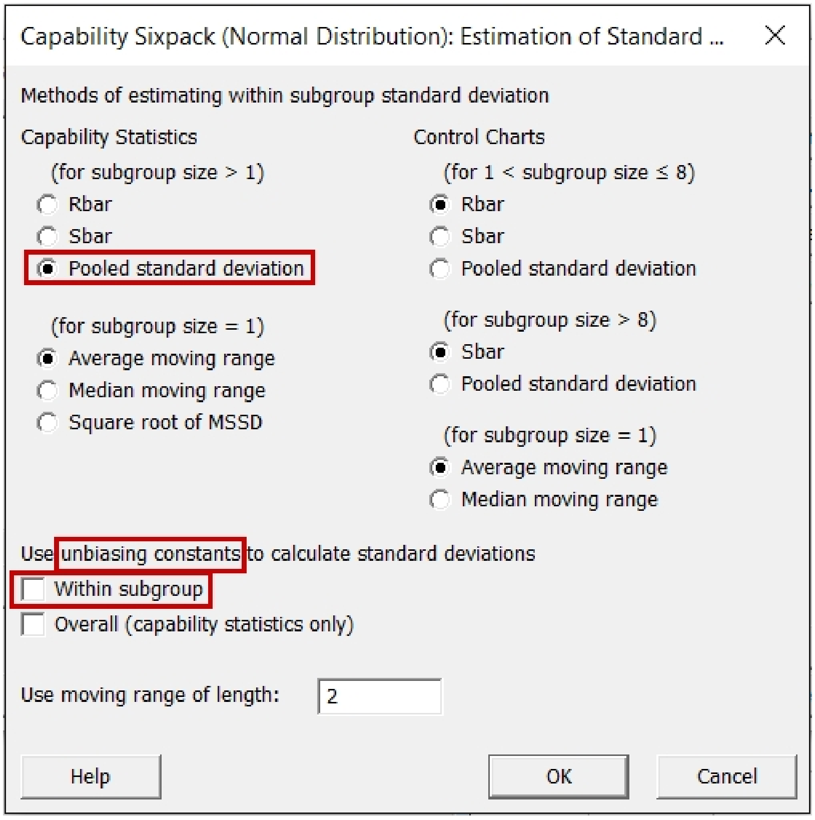
Specifications
LSL -7
Target 0
USL 7

Within	
StDev	1.584
Cp	1.47
Cpk	1.45
PPM	10.53

Overall	
StDev	1.803
Pp	1.29
Ppk	1.27
Cpm	1.29
PPM	107.05

Figure 16.205: **Minitab 17.3.1 Six Pack Capability Analysis**

and with a screenshot highlighting the fact that they are multiple ways to calculate σ_{ST} with or without unbiasing constants:



Capability Sixpack (Normal Distribution): Estimation of Standard ...

Methods of estimating within subgroup standard deviation

Capability Statistics

(for subgroup size > 1)
○ Rbar
○ Sbar
● Pooled standard deviation

(for subgroup size = 1)
● Average moving range
○ Median moving range
○ Square root of MSSD

Control Charts

(for 1 < subgroup size ≤ 8)
● Rbar
○ Sbar
○ Pooled standard deviation

(for subgroup size > 8)
● Sbar
○ Pooled standard deviation

(for subgroup size = 1)
● Average moving range
○ Median moving range

Use unbiasing constants to calculate standard deviations
☐ Within subgroup
☐ Overall (capability statistics only)

Use moving range of length: 2

Remark

Remember, as we have already mentioned it (!), replacing one of your specification limits with a boundary in a software like Minitab yields a one-sided capability analysis. If you select both specification limits as boundaries, all capability statistics will appear as asterisks (*) in the Six Pack Capability Analysis plot.

16.7.2 Taguchi Model

As part of the S.P.C. (Statistical Process Control), it is interesting for a manufacturer to estimate the financial losses generated by the differences in the target (caution! we can also apply this approach in other fields that the one of the industry!).

We can have a relatively simple and satisfactory estimate of his (financial) losses under the following assumptions/hypothesis:

H1. The process is under control (constant standard deviation) and follows a symmetric decreasing left and right mass probability function relative to the target (which can be a measurement, a number of errors per periods, etc.).

H2. The cost of error is zero when production (or work) is centered on the target (minimum).

H3. The cost increases in the same way as production off center on the left and on the right (which is not the case in the field of administration for example). The cost function therefore pass, following assumptions/hypothesis H2 and H3, through a minimum on the target.

Therefore, if we note Y the off-centering with respect to the target T and L financial loss. We have:

$$L = f(Y) \tag{16.7.126}$$

Although we do not know the form of this function, we can write it as a Taylor expansion (see section Sequences and Series page 1696) around T as:

$$L_n(Y) = \sum_{k=0}^{n} \frac{1}{k!}\frac{\mathrm{d}^k L(T)}{\mathrm{d}Y^k}(Y-T)^k \tag{16.7.127}$$

If we develop to the third order:

$$\begin{aligned} L_4(Y) &\cong \frac{1}{0!}\frac{\mathrm{d}^0 L(T)}{\mathrm{d}Y^0}(Y-T)^0 + \frac{1}{1!}\frac{\mathrm{d}^1 L(T)}{\mathrm{d}Y^1}(Y-T)^1 + \frac{1}{2!}\frac{\mathrm{d}^2 L(T)}{\mathrm{d}Y^2}(Y-T)^2 \\ &= L(T) + \frac{\mathrm{d}L(T)}{\mathrm{d}Y^1}(Y-T) + \frac{1}{2!}\frac{\mathrm{d}^2 L(T)}{\mathrm{d}Y^2}(Y-T)^2 \end{aligned} \tag{16.7.128}$$

But by the assumption/hypothesis H2, we have $L(T)$ is zero. It then remains:

$$L_4(Y) \cong \frac{\mathrm{d}L(T)}{\mathrm{d}Y^1}(Y-T) + \frac{1}{2!}\frac{\mathrm{d}^2 L(T)}{\mathrm{d}Y^2}(Y-T)^2 \tag{16.7.129}$$

and as by H3, the derivative of the function $L(Y)$ is zero in T as it is a minimum then:

$$L(Y) \cong \frac{1}{2}\frac{\mathrm{d}^2 L(T)}{\mathrm{d}Y}(Y-T)^2 \tag{16.7.130}$$

That is traditionally noted in S.P.C.:

$$\boxed{L(Y) \cong k(Y-T)^2} \tag{16.7.131}$$

and is named "Taguchi (centered) loss function" or just "quality loss function (centered)".

Well it's nice to have this relation, but how should we use it?

In fact, it's relatively simple. Under the assumptions mentioned above, if we have production defective measurement (ribs, delays, breakdowns, bug, etc.) then just calculate their arithmetical mean μ (estimator of the expected mean of a Normal distribution) and then to know the financial or hourly cost L that the defect generates for the company, institution or manufacture (sometimes this average is calculated on the basis of a single sample...).

Therefore, the above relation becomes:

$$L \cong k(\mu - T)^2 \tag{16.7.132}$$

with known L and μ.

And since T is given by the customer or depending on the context it is therefore easy to get the k factor:

$$k \cong \frac{L}{(\mu - T)^2} \tag{16.7.133}$$

which is in fact mathematically the inflection point of the mathematical function L (see section Differential and Integral Calculus page 1454)!

The latter relationship is sometimes denoted:

$$k \cong \frac{\Delta}{m^2} \tag{16.7.134}$$

Once we have k with a good estimation, it is possible to know L for any value of Y and so we can calculate in production the cost of any deviation from the target.

✍**Example:**

Consider a power supply for a stereo for which voltage T is $110\,[\text{V}]$. If the voltage is out of the range $110 \pm 20\,[\text{V}]$ then the stereo fails and must be repaired. Suppose the repair cost (including all direct and indirect costs!) is of $100.-$. Then, the associated cost for a given value of the voltage is:

$$L(Y) \cong \frac{100}{20^2}(y-110)^2 = 0.25 \cdot (y-110)^2 \tag{16.7.135}$$

Let us now see an elegant way to calculate the average Taguchi cost (average unit loss). We obviously have for a production line on several pieces of the same family:

$$\bar{L} = \frac{1}{n}\sum_{i=1}^{n} L(y_i) = \frac{k}{n}\sum_{i=1}^{n}(y_i - T)^2 = \frac{k}{n}\sum_{i=1}^{n} X_i^2 \tag{16.7.136}$$

where the X_i are supposed to be normal random variables (Gaussian by assumption because the manufacturing process is under statistical control). But we have proved in the section of

Statistics in our study of the confidence interval for the variance with known empirical mean (see page 862) that:

$$\chi^2(n)=\sum_{i=1}^{n}(Y_i-\mu)^2=\frac{1}{\sigma^2}\left[\sum_{i=1}^{n}(Y_i-\bar{Y})^2+n(\bar{Y}-\mu)^2\right] \tag{16.7.137}$$

Therefore:

$$\sigma^2\chi^2(n)=\sum_{i=1}^{n}(Y_i-\bar{Y})^2+n(\bar{Y}-\mu)^2 \tag{16.7.138}$$

And finally:

$$\begin{aligned}\bar{L}&=\frac{k}{n}\left[X_1^2+X_2^2+\ldots+X_n^2\right]=\frac{k}{n}\left[\sum_{i=1}^{n}(Y_i-\bar{Y})^2+n(\bar{Y}-\mu)^2\right]\\&=k\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^2+\frac{1}{n}n(\bar{Y}-\mu)^2\right]\\&=k\left[\bar{\sigma}^2+n(\bar{Y}-T)^2\right]\end{aligned} \tag{16.7.139}$$

This last expression has the advantage of showing very clearly that to minimize loss, we must act on the dispersion and the adjustment of the average on the nominal value.

Now remember that we have proved in the section Statistics (see page 822) during our study of likelihood estimator that (it is important in these developments that we use the notation which distinguish different estimators!):

$$\mathrm{E}(\hat{\sigma}^2)=\frac{n-1}{n}\sigma^2 \tag{16.7.140}$$

Therefore:

$$\bar{L}\cong k\left[\frac{n-1}{n}\sigma^2+(\bar{Y}-T)^2\right]=k\left[\frac{n-1}{n}\sigma^2+(\hat{\mu}-T)^2\right] \tag{16.7.141}$$

If n is large, then we have for a lot of products:

$$\boxed{\bar{L}\cong k\left[\sigma^2+(\hat{\mu}-T)^2\right]} \tag{16.7.142}$$

where the first term in brackets represents the variance of Y around its own mean and the second term of the deviation of Y with respect to the target T.

We also find this last expression in the literature and in softwares often denoted as follows:

$$L\cong k\left[S^2+(\bar{y}-y_N)^2\right] \tag{16.7.143}$$

where the index N means "Nominal". Obviously, when the target T (or nominal value) is taken as zero, the relations simplify even more.

Now, remember the following properties of the expected mean and variance (see section Statistics page 571):

$$\begin{gathered}\mathrm{E}(\alpha X)=\alpha\mathrm{E}(X)\\ \sigma_{\alpha X}=\alpha\sigma_X\end{gathered} \tag{16.7.144}$$

Therefore, if the process is non-centered and that we must correct the random variable to recenter it, it comes intuitively that the correction factor will then be (if linear):

$$\alpha = \frac{y_N}{\bar{y}} \tag{16.7.145}$$

Therefore, it comes immediately:

$$L \cong k\left[S^2\left(\frac{y_N}{\bar{y}}\right)^2 + \left(\bar{y}\frac{y_N}{\bar{y}} - y_N\right)^2\right] = y_N^2\left(\frac{S^2}{\bar{y}^2}\right) \tag{16.7.146}$$

So to minimize L, knowing that the other terms are imposed, it is necessary to minimize the ratio:

$$\frac{S^2}{\bar{y}^2} \tag{16.7.147}$$

or, equivalently, maximize the inverse ratio (which remains without dimensions):

$$\frac{\bar{y}^2}{S^2} \tag{16.7.148}$$

This number is often very large, it is customary in the literature and in statistical software to take the base-10 logarithm and multiply the result by 10 (it is an idea inspired by the acoustic physics). We then have what we name the "signal/noise ratio", given in decibels:

$$\boxed{S/N = 10\log_{10}\left(\frac{\bar{y}^2}{S^2}\right)} \text{[dB]} \tag{16.7.149}$$

for what engineers name "nominal is the best".

Relation that will meet in a software like Minitab during the analysis of Taguchi design of experiment (see below what are experimental designs). Some very rare software also offer the following special case:

$$S/N = 10\log_{10}\left(\frac{\bar{y}^2}{S^2}\right) = 10\log_{10}(\bar{y}^2) - 10\log_{10}(S^2) \underset{\bar{y}=1}{=} -10\log_{10}(S^2) \tag{16.7.150}$$

For what follows let us return to the following relation:

$$\begin{aligned}\bar{L} &= \frac{k}{n}\left[L(y_1) + L(y_2) + \ldots + L(y_n)\right]\\ &= \frac{k}{n}\left[(y_1 - T)^2 + (y_2 - T)^2 + \ldots + (y_n - T)^n\right]\\ &= \frac{k}{n}\left[X_1^2 + X_2^2 + \ldots + X_n^n\right]\end{aligned} \tag{16.7.151}$$

Engineers talk about minimizing the quality loss function when T (the target) must be zero and therefore it is necessary to minimize all y_i. So the above relation is finally reduced to:

$$\bar{L} = \frac{k}{n}\sum_{i=1}^{n} y_i^2 \tag{16.7.152}$$

What engineers sometimes like to summarize as follows (whereas I personally think that taking the logarithm in base 10 as well as putting a "-" in this case is really not justified...):

$$-10\log\left(\frac{1}{n}\sum_{i=1}^{n}y_i^2\right) \tag{16.7.153}$$

and even sometimes they say it is still a signal/noise ratio... while obviously there is no noise in this specific case that is named "smaller is better (SIB)". This is exactly in this form in which you can find this target in the software Minitab.

In the context of a parameter to maximize it is customary to say that we seeks to minimize:

$$\bar{L}=\frac{k}{n}\left[\left(\frac{1}{y_1}-\frac{1}{T}\right)^2+\left(\frac{1}{y_2}-\frac{1}{T}\right)^2+\ldots+\left(\frac{1}{y_n}-\frac{1}{T}\right)^2\right] \tag{16.7.154}$$

because given that T tends to infinity it would be difficult to do anything mathematically speaking.

So because T tends to infinity, the preceding relation is reduced to:

$$\bar{L}=\frac{k}{n}\sum_{i=1}^{n}\frac{1}{y_i^2} \tag{16.7.155}$$

What engineers also likes sometimes summarized as follows (here the base-10 logarithm is justified in my point of view!):

$$-10\log\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{y_i^2}\right) \tag{16.7.156}$$

and even sometimes they say also it is still a signal/noise ratio... while obviously there is no noise in this specific case that is named "larger is better (LIB)". This is exactly under this form in which you can find this target written in a software like Minitab.

It is, however, being clever, possible to introduce noise in the relation:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{y_i^2} \tag{16.7.157}$$

For this let us write:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{y_i^2}=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{(\mu+y_i-\mu)^2}=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\left(1+\dfrac{y_i-\mu}{\mu}\right)^2\mu^2}=\frac{1}{n\mu^2}\sum_{i=1}^{n}\frac{1}{\left(1+\dfrac{y_i-\mu}{\mu}\right)^2} \tag{16.7.158}$$

We put:

$$x=\frac{y_i-\mu}{\mu}<1 \tag{16.7.159}$$

So in this case, using the Taylor series as proved in the section of Sequences and Series (see page 1703):

$$\frac{1}{1+x^2}\cong 1-2x+3x^2 \tag{16.7.160}$$

Then we have:

$$\frac{1}{n\mu^2}\sum_{i=1}^{n}\frac{1}{\left(1+\dfrac{y_i-\mu}{\mu}\right)^2}\cong\frac{1}{n\mu^2}\sum_{i=1}^{n}\left(1-2\frac{y_i-\mu}{\mu}+3\left(\frac{y_i-\mu}{\mu}\right)^2\right) \tag{16.7.161}$$

and under statistical control we have:

$$\sum_{i=1}^{n}\frac{y_i-\mu}{\mu}\longrightarrow 0 \tag{16.7.162}$$

and assuming that we work on all parts of the population, we have:

$$\sigma^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^2 \tag{16.7.163}$$

It comes then:

$$\begin{aligned}\frac{1}{n\mu^2}\sum_{i=1}^{n}\left(1-2\frac{y_i-\mu}{\mu}+3\left(\frac{y_i-\mu}{\mu}\right)^2\right)&=\frac{1}{n\mu^2}\sum_{i=1}^{n}\left(1+3\left(\frac{y_i-\mu}{\mu}\right)^2\right)\\&=\frac{1}{n\mu^2}\left(n+3\frac{n\sigma^2}{\mu^2}\right)=\frac{1}{\mu^2}\left(1+3\frac{\sigma^2}{\mu^2}\right)\end{aligned} \tag{16.7.164}$$

So in the end by replacing a little bit abusively by the estimators, we get:

$$\boxed{S/N=-10\log_{10}\left(\frac{1}{\bar{y}^2}\left(1+3\frac{S^2}{\bar{y}^2}\right)\right)} \tag{16.7.165}$$

16.7.3 Preventive Maintenance

The evolution of production techniques towards greater automation of complex technical systems has increased the importance of the reliability of production machines or servers. Also, an unexpected shutdown is costly or even deadly to a business. Similarly, in the aerospace or nuclear industries, problems of reliability, maintainability, availability are crucial. Preventive maintenance has the purpose to quantify the reliability for all the components (mechanical, electromechanical, IT, etc.).

The existence of a maintenance service has for purpose to maintain the equipment (systems) and also the reduction in breakdowns. Indeed, these are expensive, they generate:

1. Intervention, repair costs
2. Cost of product non-quality
3. Indirect costs such as fixed costs, production losses, lost profit...
4. Life-loss in some cases...

Therefore, every effort should be made to avoid the blackout/failure, act quickly when it occurs to increase equipment availability. To do this, one must model the life of equipment and train people to react. The set of methods and techniques for its problems are usually classified under the name of "Failure Mode and Effects Analysis (FMEA)", but we can also found it under the name of "Probabilistic risk assessment (PRA)". Especially, that latter is a very official systematic and comprehensive methodology to evaluate risks associated with a complex engineered technological entity (such as an airliner or a nuclear power plant) or the effects of stressors on the environment (Probabilistic Environmental Risk Assessment- PERA) for example[20].

"Quantitative Reliability Engineering" analysis involves more than just reliability predictions and reliability demonstration that are performed against a given program or project requirements. Quantitative Reliability Engineering analysis can play a key role in supporting a broad range of applications. It is critical in addressing design and manufacturing deficiencies.

We distinguish two main classes of systems:

1. Non-repairable systems (satellites, although low cost goods, etc.)
2. Repairable systems (production machinery, means of transport, etc.)

where the theoretical approaches are different. For the second category, you can also use Markov chains (where the states represent the number of functional components or failure of a system in accordance with the standard ISO 31010), Petri nets or Monte Carlo simulation.

---

[20] In addition to the mentioned methods, PRA studies require special but often very important analysis tools like human reliability analysis (HRA) and common-cause-failure analysis (CCF). HRA deals with methods for modelling human error while CCF deals with methods for evaluating the effect of inter-system and intra-system dependencies which tend to cause simultaneous failures and thus significant increase in overall risk.

The idea is in the texts that follow to make the point on one of these methods, to seek their efficiency and enable engineers practitioners and technicians to better understand these problems. An emphasis will be placed on the Weibull model, which has a significant importance in this field.

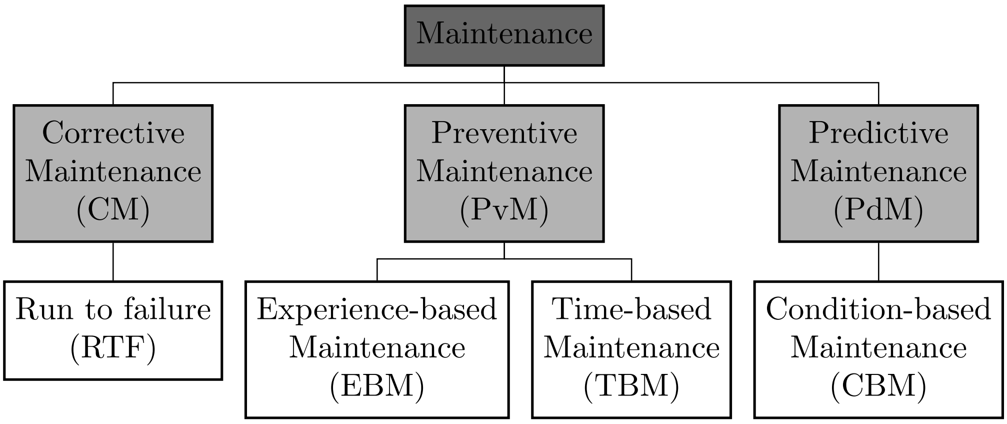

Maintenance
Corrective Maintenance (CM)
Preventive Maintenance (PvM)
Predictive Maintenance (PdM)
Run to failure (RTF)
Experience-based Maintenance (EBM)
Time-based Maintenance (TBM)
Condition-based Maintenance (CBM)

Figure 16.206: **Different types of maintenance**

16.7.3.1 Planned Obsolescence

The planned obsolescence goes against preventive maintenance philosophy whose primary purpose is to provide a quality product and respectful of its environment with precise knowledge of the product's lifetime and quantifiable risk of obsolescence.

At the opposite, the purpose of planned obsolescence is to reduce the lifetime (or "lifespan") of a product in order to increase the rate of replacement by the buyer (sometimes motivated by political strategy or short term money return...). Therefore the induced demand would have for aim (not proved in the very long term ... as this is quite bad for the reputation of the brand) to benefit the manufacturer, or its competitors. The sector then has a greater production, stimulating productivity growth (economies of scale) and technical progress (which accelerates the obsolescence of previous products) and maintains the use of an ever growing population which is therefore problematic by definition.

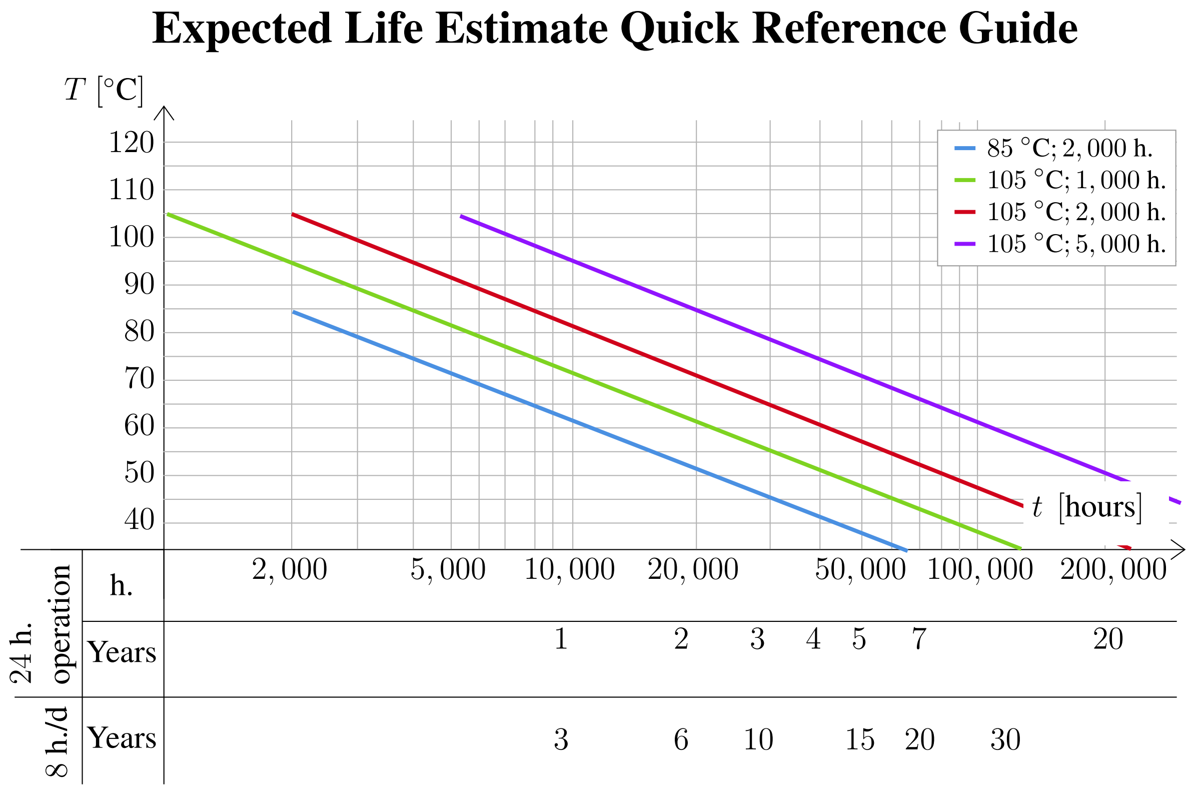


Figure 16.207: **Example of a reliability plot of capacitor (source: Cash Investigation about Samsung TV issue)**

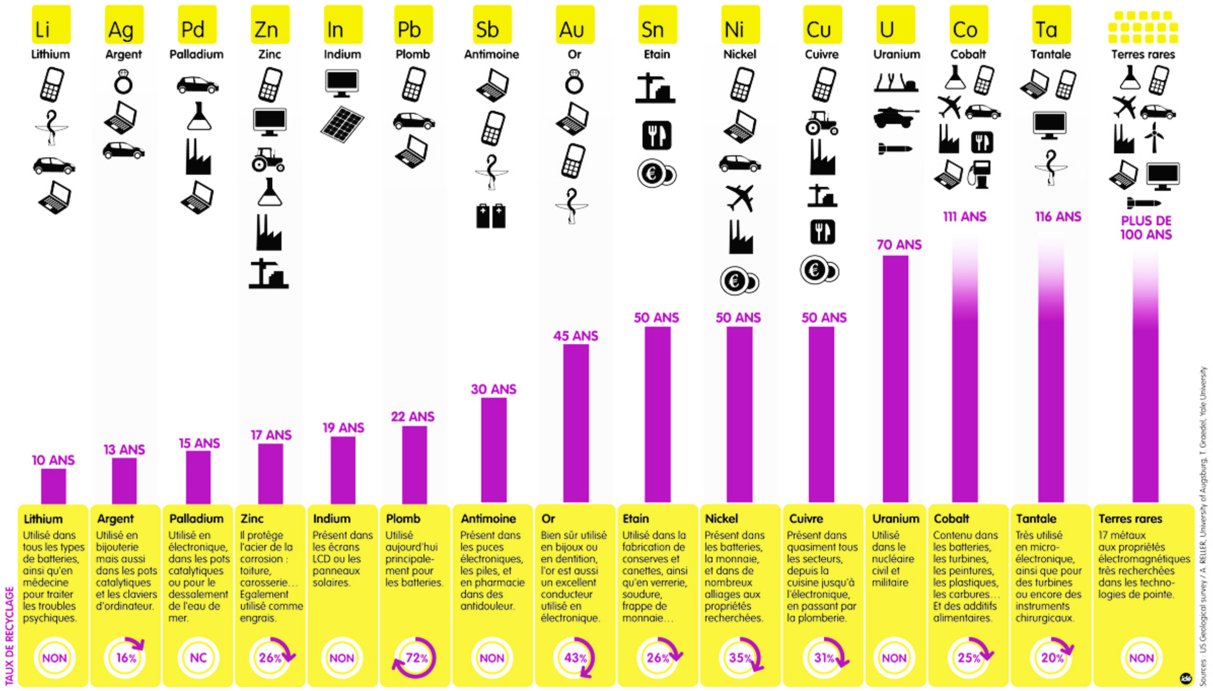
The reader will therefore have to understand that I am strongly against these methods and that every engineer and every scientist working for a company must refuse by deontology this type of request coming from the hierarchy even denounce these methods anonymously on the Internet or to consumer associations.

Remark

The consumer must be very careful with TV shows or newspapers about the subject of planned obsolescence! Indeed as they don't have any statisticians consultants, these shows or newspapers communicate numbers in a very wrong and bad way (a journalist being not a scientist...).

It must be also notice that lifetime (lifespan) will be a major subject of concern for future generations as Earth's resources are not unlimited and that in this 121st century (holocene calendar) people still don't recycle everything. So it will be a challenge for these future generations (at least if we will not be able to bring material from other planets, satellites or asteroids) to find solutions to get the longest lifespan. Having the longest lifespan is however also in the 121st century (holocene calendar) sometimes a requirement as it is an important factor in aeronautics, trains, boats or all other areas of cutting edge technologies and where human life are in concerns (or very rarely in luxury goods also).

But let us also show a small chart of the status of the known reserves - in years - of row material resources (so not including unknowns reserves!) if humans continue to consume as they do in the beginning of this 121st century (sorry we did not found the english version of this figure):



Lithium Utilisé dans tous les types de batteries, ainsi qu'en médecine pour traiter les troubles psychiques.

Argent Utilisé en bijouterie mais aussi dans les pots catalytiques et les claviers d'ordinateur.

Palladium Utilisé en électronique, dans les pots catalytiques ou pour le dessalement de l'eau de mer.

Zinc Il protège l'acier de la corrosion : toiture, carosserie… Egalement utilisé comme engrais.

Indium Présent dans les écrans LCD ou les panneaux solaires.

Plomb Utilisé aujourd'hui principalement pour les batteries.

Antimoine Présent dans les puces électroniques, les piles, et en pharmacie dans des antidouleur.

Or Bien sûr utilisé en bijoux ou en dentition, l'or est aussi un excellent conducteur utilisé en électronique.

Etain Utilisé dans la fabrication de conserves et canettes, ainsi qu'en verrerie, soudure, frappe de monnaie…

Nickel Présent dans les batteries, la monnaie, et dans de nombreux alliages aux propriétés recherchées.

Cuivre Présent dans quasiment tous les secteurs, depuis la cuisine jusqu'à l'électronique, en passant par la plomberie.

Uranium Utilisé dans le nucléaire civil et militaire

Cobalt Contenu dans les batteries, les turbines, les peintures, les plastiques, les carbures… Et des additifs alimentaires.

Tantale Très utilisé en micro-électronique, ainsi que pour des turbines ou encore des instruments chirurgicaux.

Terres rares 17 métaux aux propriétés électromagnétiques très recherchées dans les technologies de pointe.

Sources : US Geological survey / A. RELLER, University of Augsburg, T. Graedel, Yale University

16.7.3.2 Reliability Empirical Estimators

As part of the non-accelerated reliability study (accelerated aging will be treated in a near future in this textbook), we are led to define some variables which are listed below:

- N_0 will the number of good elements at time t_0 (initial time)
- N_i the number of good elements at time t_i
- n_i the number of defective elements between t_i and t_{i-1} also denoted by ΔN_i
- Δt_i the time interval between t_i and t_{i-1}

Definitions (#16.7.732):

D1. We define the "failure rate per time-slot" Δt by the relation:

$$\hat{\lambda}=\frac{\Delta N_i}{N_i\Delta t_i}=\frac{n_i}{N_i\Delta t_i} \tag{16.7.166}$$

which therefore is interpreted as being the number of defective elements relative to the number of surviving elements on a given timeslot (this is therefore a percentage of non-conformities related to a time slot).

The latter relation is also sometimes named "hazard function" (or "hazard ratio (HR)") or "relative survival" and often denoted by $\bar{h}(t_i)$.

D2. We define the "failure function" by the relation (probability density of failure at time t_i):

$$\hat{f}(t_i)\Delta t_i=\frac{n_i}{N_0} \tag{16.7.167}$$

noting that the denominator is not the same as that which defines the default rate per timeslot!

This function is therefore interpreted as the $\%$ of defective elements in the studied time frame relative to the total number of elements initially tested. This is the indicator that interests most often the engineer because can be assimilate to a probability and having the properties of a probability!

D3. We naturally define the "cumulative failure function" by:

$$\hat{F}(t_i)=\sum_i\hat{f}(t_i)\Delta t_i=\frac{\displaystyle\sum_i n_i}{N_0}=\frac{N_0-N_i}{N_0}=1-\frac{N_i}{N_0} \tag{16.7.168}$$

which tends to 1 when time tends to infinity.

This function is therefore interpreted as the $\%$ of cumulative defective components based on the total number of elements initially tested. It is therefore the function of empirical distribution of probabilities of failures.

D4. We define verbatim the "reliability function" by (this is the second term of the previous relation):

$$\hat{R}(t_i)=\frac{N_i}{N_0}=1-\hat{F}(t_i) \tag{16.7.169}$$

Its name comes from the interpretation of the ratio in the context of the definition of the cumulative failure function.

It must be remembered that the letter R is derived from the word "reliability" and F is derived from the word "failure" which means "breakdown".

D5. In reliability, or survival analysis, the "risk set" is the set of subjects who have not experienced the event of interest at time t, and who have not been censored by time t (and thus who are at risk at time t for the event of interest).

It follows from these definitions that:

$$\boxed{\frac{\hat{f}(t_i)}{\hat{R}(t_i)}=\hat{\lambda}(t_i) \quad\Rightarrow\quad \hat{\lambda}(t_i)\Delta t_i=\frac{\hat{f}(t_i)}{\hat{R}(t_i)}\Delta t_i} \tag{16.7.170}$$

The latter relation is used for the calculation of reliability laws as we will see later! We deduce a notation which sometimes helps to better understand the failure rate per time slot. Effectively:

$$\hat{\lambda}(t_i)=\frac{\hat{f}(t_i)}{\hat{R}(t_i)}=\frac{\hat{f}(t_i)\Delta t}{\hat{R}(t_i)\Delta t}=\frac{\dfrac{n_i}{N_0}}{\hat{R}(t_i)\Delta t}=\frac{\dfrac{\Delta N_i}{N_0}}{\hat{R}(t_i)\Delta t}=\frac{\dfrac{N_{i-1}}{N_0}-\dfrac{N_i}{N_0}}{\hat{R}(t_i)\Delta t}=\frac{\bar{R}(t_{i-1})-\bar{R}(t_i)}{\hat{R}(t_i)\Delta t} \tag{16.7.171}$$

Because:

$$\hat{f}(t_i)\Delta t_i\leqslant 1 \quad\text{and}\quad \lim_{i\rightarrow\infty}\sum_i\hat{f}(t_i)\Delta t_i=1\frac{\hat{f}(t_i)}{\hat{R}(t_i)}\Delta t_i \tag{16.7.172}$$

the failure function can be seen as a probability as we have already mentioned it, which naturally led us to define its expected mean:

$$\boxed{\mathrm{E}(T)=\sum_i t_i\hat{f}(t_i)\Delta t_i=\sum_i t_i\frac{n_i}{N_0}} \tag{16.7.173}$$

Very useful relation in practice which in theory gives the average percentage of failed elements at the moment t_i. So to summarize and with different notations that we can find in the literature, this gives (we will see later the same summary in the continuous case):

Function type	**Mathematical expression**
Failure density function	$\hat{f}\Delta t_i = \dfrac{n_i}{N_0}$
Failure repartition function	$\bar{F}(t_i) = \sum_i \hat{f}(t_i)\Delta t_i = \sum_i \dfrac{n_i}{N_0} = P(T \leq t_i)$
Survival repartition function	$\hat{R}(t_i) = \hat{S}(t_i) = 1 - \hat{F}(t_i) = P(T \geq t_i) = 1 - \sum_i \hat{f}(t_i)\Delta t_i$
Risk (hazard) function	$\hat{\lambda}(t_i) = \hat{h}(t_i) = \dfrac{\hat{f}(t_i)}{\hat{R}(t_i)}$
Cumulated risk (hazard)	$\hat{H}(t_i) = \sum_i \dfrac{n_i}{N_i \Delta t_i}$

Table 16.17: **Summary of discrete maintenance analysis functions**

Note that we will see later below a statistical mathematical model based on the Kaplan-Meier estimator that gives in fact a nonparametric estimation of the survival function. The following example will perhaps help you better understand the concept!

✍**Example:**

We have identified on a batch of 37 engines of a given type the following listed failures in slices (given by customers or internally measured on test benches):

0 to 1,000h.	**1,000 to 2,000h.**	**2,000 to 3,000h.**	**3,000 to 4,000h.**	**4,000 to 5,000h.**	**5,000 to 6,000h.**
1	4	7	12	11	2

Table 16.18: **Engine failures in effort slices**

We must consider the value of the reliability function $\hat{R}(t_i)$, of the failure function $\hat{f}(t_i)$ and the failure by time slice $\hat{\lambda}(t_i)$. The calculations are elementary and we get the following table:

i	**Observation interval**	**Failures in the interval**	**Survivors**	**Cumulated failures**	$\hat{R}(t_i)$	$\hat{f}(t_i)\Delta t_i$	$\hat{\lambda}(t_i)$ $10^{-6}[\text{h}^{-1}]$
0	0	–	37	0	100%	0%	0
1	0 to 1,000	1	36	1	97%	2.7%	$(1/36)/1000 \cong 27.7$
2	1,000 to 2,000	4	32	5	86%	10.8%	$(4/32)/1000 = 125$
3	2,000 to 3,000	7	25	12	67%	18.9%	$(7/25)/1000 = 280$
4	3,000 to 4,000	12	13	24	35.1%	32.4%	$(12/13)/1000 \cong 923$
5	4,000 to 5,000	11	2	35	5.4%	5.4%	$(11/2)/1000 \cong 5,500$
5	6,000 to 6,000	2	0	37	0%	–	–

Table 16.19: **Analysis of engine failures per effort slices**

We see above that the failure rate $\hat{\lambda}(t_i)$ is of course not constant! The failure rate $\hat{f}(t_i)$ would therefore be about 2.5% to 3% the first year, about 10% to 11% the second year, etc.

Regarding default rates, engineers often recognize three analysis slices following that the studied objects are: 1) New 2) In normal operation 3) Considered aging.

In practice we consider quite intuitively (and sometimes roughly) that for very common goods the failure rate follows a bathtub curve as shown below (in the technical tables it is often the normal operation failure rate that is given):

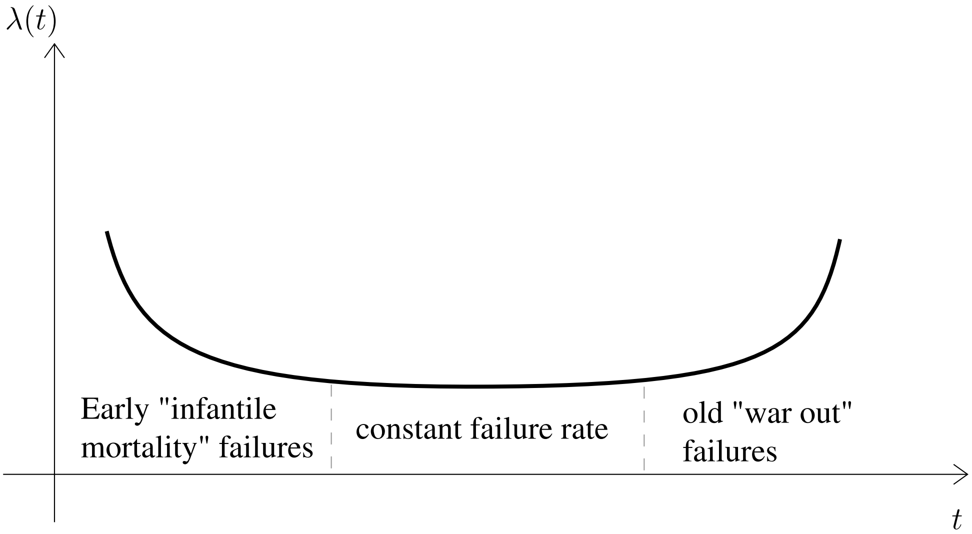


Figure 16.208: **Reliability bathtube**

whereas if you look at the table of the example above, the failure rate does not follow at all a bathtub curve (so it is a cons-example).

Reliability engineers often cut the bath in the three visible parts above, but under the following technical names:

1. D.F.R. for "Decreasing Failure Rate": young unidentified components with manufacturing problems during the process are eliminated from the batch which has the effect of reducing the failure rate. The Weibull distribution is relatively well suited to model this phase (see later below the details of this model).
2. C.F.R. for "Constant Failure Rate": the components are in stable conditions.
3. I.F.R. for "Increasing Failure Rate" components are in end of life and their failure rate increases. The Weibull distribution is again relatively well adapted to model this phase.

The Poisson distribution (see section Statistics page 686) is relatively well adapted to model the failure arrival time, and the exponential distribution (see section Statistics page 751) to model the time between successive failures. This fact follows mathematical proofs available in the section of Quantitative Management (queueing theory). So, this means that failures are often considered independent. Therefore, the law on the number of failures in a given period is then a law "without memory" and can be modelled by a Poisson distribution (proof also made in the section of Quantitative Management).

Remark

Contrary to what a lot of theorists think... many general public softwares have also sometimes a failure rate following a bathtub curve. Indeed, at the beginning there are undetected bugs that cause the failure will decrease following the identification and correction. Then, because of frequent updates of the environment that tend to add other issues (service pack), the failure rate remains fairly constant. Finally, over time, the evolution of surrounding technologies (frameworks) make the application obsolete and some functions no longer properly work which is again an increase the failure rate.

Vis-à-vis the effectiveness of renovation, let us indicate (by simplifying) that they can be frequently divided into three categories:

1. "As good as new": This is preventive maintenance in the sense that we change a piece when his life times leads it to a failure rate that we consider too high and that the unanticipated failure will cost more that its non-anticipation.
2. "As bad as old": It is the poor maintenance in the sense that we change a piece only when it comes to break, which mainly generates higher costs of maintenance that the preventative maintenance that involves in anticipating issues.
3. "Partial Restore": This is the minimum preventive maintenance in the sense that we repair the defective part rather than replace it with a new one. Again the cost of the problem must be calculated by an audit of requirements and of the deadlines of the business.

Let us return to other definitions in the passage to the limit of the continuum!

So we know that the "instantaneous failure rate" will have for unit the inverse of time as $[\lambda]=\mathrm{s}^{-1}$. This rate is, within the framework of our study, not necessarily constant over time as we have seen above!

Given $R(t)$ the cumulative percentage of objects analysed (tracked) always running well (no failures!) - reliable objects - of a sample tested at time t. The number of objects falling down during the infinitesimal time $\mathrm{d}t$ is therefore equal to:

$$\lambda(t)R(t)\mathrm{d}t \tag{16.7.174}$$

this corresponds to the decrease of the initial stock reliable at time t. $R(t)$ is also the probability that the time T of an event will occur after time t.

We can then write the relation:

$$\mathrm{d}F(t)=\lambda(t)R(t)\mathrm{d}t \tag{16.7.175}$$

therefore:

$$\frac{\mathrm{d}F(t)}{R(t)}=\frac{F(t+\mathrm{d}t)-F(t)}{R(t)}=\frac{\mathrm{d}F(t)}{1-F(t)}=\lambda(t)\mathrm{d}t \tag{16.7.176}$$

Which can often be found in the literature under the following equivalent forms:

$$\lambda(t)=h(t)=\frac{\mathrm{d}F(t)}{R(t)\mathrm{d}t}=\frac{f(t)}{R(t)}=\frac{1-\mathrm{R(t)}}{R(t)\mathrm{d}t}=\frac{\mathrm{d}R(t)}{R(t)\mathrm{d}t} \tag{16.7.177}$$

Definition (#16.7.733): The "conditional probability of failure" between t and $t+\mathrm{d}t$ is defined as the temporal conditional probability of experiencing the event at a given time t, knowing that we have not experienced before and so we survived up to a time t (i.e. it only increase with time...):

$$\boxed{\lambda(t)\mathrm{d}t=\frac{\text{cumulated probability of default between } t+\mathrm{d}t}{\text{cumulated probability of surival at time } t}=\frac{F(t+\mathrm{d}t)-F(t)}{R(t)}=\frac{\mathrm{d}F(t)}{1-F(t)}} \tag{16.7.178}$$

where $F(t)$ and $R(t)$ are, for reminder, respectively the cumulative probability function of failure (cumulative probability to fail at time t) and the cumulative probability function of reliability also named "survival function" (denoted $S(t)$ in clinical trials involving biological entities). $R(t)$ is 1 at time 0 and ... 0 after an infinite time as we have already seen before!

We have already study this conditional probability with a more pedagogical approach during our study of Cox's proportional hazard model in the section Statistics at page 1291.

Remark

By the same intellectual approach, rather than defining a function of failure $F(t)$ and survival $R(t)$ with its associated hazard function, we can define a repairability function with its function $M(t)$ which would then be a "maintainability function".

If we integrate (caution! u now represents the time!):

$$\begin{aligned}\int_0^t \lambda(u)\mathrm{d}u &= \int_0^t \frac{\mathrm{d}F(u)}{1-F(u)}\\ \Rightarrow -\int_0^t \lambda(u)\mathrm{d}u &= -\int_0^t \frac{\mathrm{d}F(u)}{1-F(u)}\\ \Rightarrow -\int_0^t \lambda(u)\mathrm{d}u &= \ln(1-F(u))\Big|_0^t\end{aligned} \tag{16.7.179}$$

As $F(t_0=0)=0$ we have:

$$-\int_0^t \lambda(u)\mathrm{d}u = \ln(1-F(t)) \quad \Rightarrow \quad \exp\left(-\int_0^t \lambda(u)\mathrm{d}u\right) = \ln(1-F(t)) \tag{16.7.180}$$

Therefore:

$$\boxed{\begin{aligned}&R(t)=e^{-\int_0^t \lambda(u)\mathrm{d}u}\\ &F(t)=1-R(t)=1-e^{-\int_0^t \lambda(u)\mathrm{d}u}\end{aligned}} \tag{16.7.181}$$

Moreover, since we have seen that $\hat{f}(t_i)=\hat{\lambda}(t_i)\hat{R}(t_i)$, then we have the "density/repartition function of instantaneous failure":

$$f(t)=\lambda(t)R(t)=\lambda(t)e^{-\int_0^t \lambda(u)\mathrm{d}u} \tag{16.7.182}$$

We can get this relation and interpretation in another way:

$$\boxed{f(t)=\frac{\mathrm{d}F(t)}{\mathrm{d}t}=\lambda(t)e^{-\int_0^t \lambda(u)\mathrm{d}u}} \tag{16.7.183}$$

So where we find back therefore $F(t)$ the function of cumulative probability of failure (also sometimes denoted CIF(t) and named "cumulative incidence function"). Obviously to determine the law $f(t)$, we use the usual statistical tools for goodness of fit (see section Statistics page 1130).

We then get the very important equation (if not the most important one!) in practice, which connects the density function of the instantaneous failure and the reliability's law:

$$\boxed{f(t)=\frac{\mathrm{d}F(t)}{\mathrm{d}t}=\frac{\mathrm{d}(1-R(t))}{\mathrm{d}t}=-\frac{\mathrm{d}R(t)}{\mathrm{d}t}} \tag{16.7.184}$$

We have above the three most general expressions binding laws of reliability and the instantaneous failure rates. Let us add that we have also the following instantaneous hazard function that follows:

$$\lambda(t)=h(t)=\frac{f(t)}{R(t)}=\frac{-\dfrac{\mathrm{d}R(t)}{\mathrm{d}t}}{R(t)}=\frac{-R'(t)}{R(t)}=\frac{\mathrm{d}}{\mathrm{d}t}(-\ln(R(t))=\frac{\mathrm{d}}{\mathrm{d}t}(-\ln(1-F(t)) \tag{16.7.185}$$

Remark

More explicitly the hazard function is related to the survivor function in the following way:

$$\begin{aligned}h(t_0)&=\lim_{\delta t\to 0}\frac{1}{\delta t}P(\text{fail in interval }[t_0,t_0+\delta t[|\text{still working at time }t_0)\\&=\lim_{\delta t\to 0}\frac{P(t_0\leq T+\delta t|T\geq t)}{\delta t}\\&=\lim_{\delta t\to 0}\frac{1}{\delta t}\frac{F(t_0+\delta t)-F(t_0)}{S(t_0)}=\frac{1}{S(t_0)}\lim_{\delta t\to 0}\frac{F(t_0+\delta t)-F(t_0)}{\delta t}\\&=\frac{1}{S(t_0)}\frac{\mathrm{d}}{\mathrm{d}t}F(t_0)=\frac{f(t_0)}{S(t_0)}\end{aligned}\tag{16.7.186}$$

It then follows immediately that:

$$\begin{aligned}H(t)=\int_0^t h(t)\mathrm{d}t=\int_0^t\frac{\mathrm{d}}{\mathrm{d}t}(-\ln(1-F(t))\mathrm{d}t=-\int_0^t\mathrm{d}\ln(1-F(t))=-\ln(1-F(t))\Big|_0^t\\=-\ln(1-F(t))=-\ln(R(t))\end{aligned}\tag{16.7.187}$$

Also sometimes denoted $\Lambda(t)$ in textbooks. From the latter result we also have obviously:

$$R(t)=S(t)=e^{-H(t)}\tag{16.7.188}$$

Since $f(t)$ is the failure density function, the expected value of the failure is given by:

$$\mathrm{E}(t)=\int_0^{+\infty}tf(t)\mathrm{d}t=\text{MTBF}\tag{16.7.189}$$

Thus, if the distribution of failures is equally likely (uniform density function), which is rare..., half of the equipment will be off at $\mathrm{E}(t)$.

Remark

Tracking $100,000$ hard drives, Google engineers have observed an average of 8% loss per year! So a higher loss rate than were announcing the manufacturer that was about $300,000$ hours average! The loss rate is higher the first 3 years! But maybe the disks that survive works longer and therefore compensate the low average observed by Google...!?

We also have use the integration by parts and the above relation:

$$\begin{aligned}\mathrm{E}(t)=\int_0^{+\infty}t\frac{\mathrm{d}F(t)}{\mathrm{d}t}\mathrm{d}t=\int_0^{+\infty}t\mathrm{d}F(t)=-\int_0^{+\infty}t\mathrm{d}R(t)\\=-tR(t)|_0^{\infty}+\int_0^{+\infty}R(t)\mathrm{d}t=\int_0^{+\infty}R(t)\mathrm{d}t=\text{MTBF}\end{aligned}\tag{16.7.190}$$

Hence another way to express it:

$$\boxed{\mathrm{E}(t)=\int_0^{+\infty} tf(t)\mathrm{d}t=\int_0^{+\infty} R(t)\mathrm{d}t=\text{MTBF}} \tag{16.7.191}$$

So to summarize and with different notations that we can find in the literature, this gives:

Function type	**Mathematical expression**
Failure density function	$f(t)=\lim\limits_{\mathrm{d}t\rightarrow 0}\dfrac{P(T\in[t+\mathrm{d}t])}{\mathrm{d}t}=\dfrac{\mathrm{d}F(t)}{\mathrm{d}t}$
Failure repartition function	$F(t)=\int_0^T f(t)\mathrm{d}t=P(T\leq t)$
Survival repartition function	$R(t)=S(t)=1-F(t)=P(T\geq t)=1-\int_0^T f(t)\mathrm{d}t$
Risk (hazard) function	$h(t)=\lim\limits_{\mathrm{d}t\rightarrow 0}\dfrac{P(T\in[t+\mathrm{d}t]\lvert T\geq t)}{\mathrm{d}t}=\dfrac{f(t)}{R(t)}$
Cumulated risk (hazard)	$H(t)=\int_0^t h(s)\mathrm{d}s=\int_0^t \dfrac{f(s)}{R(s)}\mathrm{d}s=-\ln(S(t))$

Table 16.20: **Summary of continuous maintenance analysis functions**

Let us notice again (!) that we will see further statistical mathematical model based on the Kaplan-Meier estimator giving a nonparametric estimation of the survival function!

Before continuing let us give some other definitions of maintenance indicators, the most important are the European Maintenance Standard EN 13306:2010 (I precise it because on the web we can find the best and the worst on the definitions of these reliability indicators...)!

Definitions (#16.7.734):

D1. The "Time Between Maintenance (T.B.M.)" is the contractual time between two visits or maintenance control (not defined in the EN 13306:2010 standard).

D2. The "Mean Operating Time Between Failures (M.O.T.B.F.)" applicable only for repairable items is the expected mean of the operational time of a system between the end of the first failure and the beginning of the next (defined in the EN 13306:2010 standard). In the industry M.O.T.B.F. is also named "Mean Up Time (M.U.T.)".

D3. The "Mean Time Between Failures (M.T.B.F.)" applicable only for repairable items is the expected mean of operational time of a system between the start of two failures (defined in the standard EN 13306:2010). This indicator includes the time of non-operation. In the area of services, the M.T.B.F. is sometimes named "Mean Time Between System Accidents (M.T.B.A.)". In the field of the production lines M.T.B.F. is named "Mean

Time Between Defects (M.T.B.D.)" (we do not stop the production because only one defect was detected as we will see during our study of control charts further below). In the field of computing science the M.T.B.F. is named "Mean Time Between Errors (M.T.B.E.)" (we also rarely stops a software or a website because one bug or error was detected!).

D4. The "Mean Time To Failure (M.T.T.F.)" applicable for repairable or not repairable items is the expected mean of operational time of a system to its first failure (not defined in the standard EN 13306:2010). In fact it is the latter which is calculated most often in tests laboratories.

D5. The "Mean Repair Time (M.R.T.)" applicable only for repairable items is the expected mean of repair time (defined in the standard EN 13306:2010).

D6. The "Mean Time To Restore (M.T.T.R.)" applicable only for repairable items is the expected mean restart time (defined in the standard EN 13306:2010). In the area of services, M.T.T.R. is sometimes named "Mean Time To Restore Service (M.T.T.R.S.)".

D7. The "Mean Down Time (M.D.T.)" is equal to the mathematical expected mean of the time that the system is not operational. This indicator (not defined in the standard EN 13306:2010) therefore includes the M.R.T. It follows that the M.T.B.F. is equal to the sum of the M.O.T.B.F. (M.U.T.) and of the M.D.T.

D8. The "Mean Time To Detection (M.T.T.D.)" is equal to the mathematical expected mean of the time that the system is not detected as being non operational. This indicator (not defined in the standard EN 13306:2010), as well as the M.R.T. is included in the M.D.T.

D9. The "Mean Time to Answer (M.T.T.A.)" is equal to the mathematical expected mean of the time during which the user expected an answer or response of manufacturer after reporting a failure. This indicator (not defined in the standard EN 13306:2010), as well as the M.R.T. and M.T.T.D. are included in the M.T.T.A.

D10. The "Overall Equipment Effectiveness (O.E.E.)" in the field of maintenance is traditionally defined as being simply the ratio of the operational running time and time expected operational performance (not defined in the standard EN 13306:2010).

D11. The "Service Ability (S.A.)" which in the field of maintenance is often given by the following two relations and whose goal is to make it as close as possible to 1 (that is to say to make the M.D.T. approaching zero):

$$\mathrm{SA}=\frac{\mathrm{MTBF}-\mathrm{MDT}}{\mathrm{MTBF}}=\frac{\mathrm{MOTBF}}{\mathrm{MOTBF}+\mathrm{MDT}} \quad (16.7.192)$$

and we can continue to define many others indicators outside the standard...

The below diagram summarizes maybe a little better all these definitions:

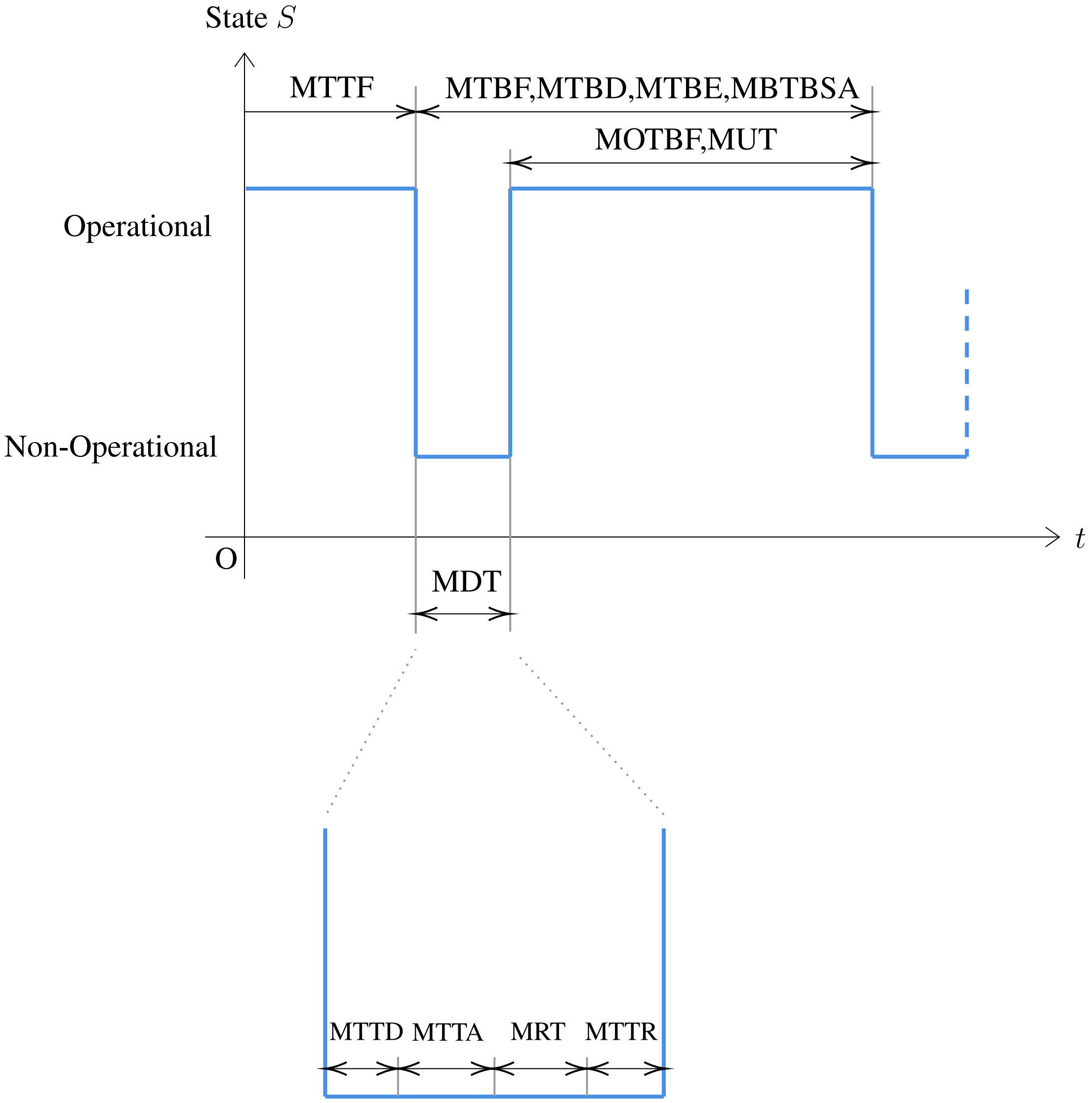


Figure 16.209: **Common reliability indicators**

Remarks

R1. For each of these indicators it will be indicated if external causes have been taken into account or not.

R2. If the tested sample size is very small or just unitary (only one machine), it is customary to use the arithmetic average as an estimator of the expected mean.... Obviously it can lead to huge errors...

R3. It is of course possible from the defined indicators to calculate whether a machine (element) is able to ensure the desired service (such as a number of pieces per year to produce for example!). So not only do these indicators can be measured, but they can therefore also be used to verify the achievement of a desired goal!

R4. In the beginning of the 121st century (holocene calendar) countries should strictly speaking legislate to require all manufacturers to communicate the MTTF of their products so that consumers can make the best choice to buy and compare the values in relation to the guarantee provided! Unfortunately, this is not the case and it would highlight a current bad tradition in the consumer products industry that is to manufacture components whose life revolves around 200,000 hours to ensure industrial renewal of their market.

✍**Example:**

A machine is supposed to work 24/24 hours 7/7 days. It turned up today for $5,020$ hours with 2 stops of a total of 20 hours (thus included in the $5,020\,[\text{h}]$). The conventional indicators relative to the limited information provided are given by:

$$\begin{aligned}
\text{MTBF} &= \frac{5,020}{2} = 2,510\,[\text{h}]\\
\text{MDT} &= \frac{20}{2} = 10\,[\text{h}]\\
\text{MOTBF} = \text{MUT} &= \frac{5,020-20}{2} = 2,500\,[\text{h}]\\
\text{SA} &= \frac{\text{MOTBF}}{\text{MOTBF+MDT}} = \frac{2,500}{2,500+10} \cong 99.6\%
\end{aligned}\tag{16.7.193}$$

The classification of systems in terms of availability commonly leads to 7 classes from "not available" (system available 90% of the time, and therefore unavailable over a month per year) to "ultra available" (available 99.99999% of the time and therefore unavailable only 3 seconds per year): these classes corresponds arbitrarily to the numbers of "9" in the percentage of time that the class systems are available (one year includes $525,600$ minutes for refresh):

Type	Non-availability [min/year]	Percentage availability	Class
non-managed	$50,000$ (~ 3.5 days)	90%	1
managed	$5,000$ (~ 3.5 days)	99%	2
well-managed	500 (~ 8 hours)	99.9%	3
faultily tolerance	50	99.99%	4
high availability	5	99.999%	5
very high availability	0.5	99.9999%	6
very high availability	0.05	99.99999%	7

Table 16.21: **Classes failures in FMEA**

The use of these parameters in the context of reliability framework told us that we have an "approach through average values".

Note however one thing! This is not because an event has a smaller or infinitely less probability than another one that it is less important!! Indeed we must obviously take into account (certainly a little in a empirical way) the gravity of the failure. Thus, a nuclear power plant may well be $1,000$ times more secure than a thermal central in terms of probabilities ... nothings is comparable in terms of severity when a major failure occurs in a nuclear plant! Indeed in the case of a thermal central there may be in the worst case 500 persons affected while with a nuclear power plant that is another story (see Tchernobyl or Fukushima...)!

Let us also note a simple case: Some components (typically electronic) have in their period of a maturity constant failure rate. The cumulative probability distribution of the failure that results is then immediately deducted as the instantaneous hazard function is given by $\lambda(t)=\lambda$ and therefore:

$$F(t)=1-\exp\left(-\int_0^t \lambda(t)\mathrm{d}t\right)=1-e^{-\lambda t} \tag{16.7.194}$$

Let us indicate that we find very often the latter relation in the following form in the literature:

$$1-F(t)=R(t)=e^{-\lambda t} \tag{16.7.195}$$

and then we have well:

$$h(t)=\frac{f(t)}{R(t)}=\frac{\lambda e^{-\lambda t}}{e^{-\lambda t}}=\lambda \tag{16.7.196}$$

To summarize a bit for the exponential law:

$$\boxed{\begin{aligned}F(t)&=1-e^{-\lambda t}\\ f(t)&=\lambda e^{-\lambda t}\\ R(t)&=e^{-\lambda t}\\ h(t)&=\lambda\end{aligned}} \tag{16.7.197}$$

And as we have:

$$\text{MTBF}=\int_0^{+\infty} tf(t)\mathrm{d}t=\int_0^{+\infty} t\lambda e^{-\lambda t}\mathrm{d}t \tag{16.7.198}$$

And we proved in the section of Differential and Integral Calculus that (see page 1536):

$$\int t\lambda e^{-\lambda t}\mathrm{d}t=\lambda e^{-\lambda t}\left(-\frac{1}{\lambda}-\frac{1}{\lambda^2}\right) \tag{16.7.199}$$

Therefore:

$$\text{MTBF}=\int_0^{+\infty} t\lambda e^{-\lambda t}\mathrm{d}t=-\lambda\left(-\frac{1}{\lambda^2}\right)=\frac{1}{\lambda} \tag{16.7.200}$$

The previous hazard and survival are then graphically of the following form:

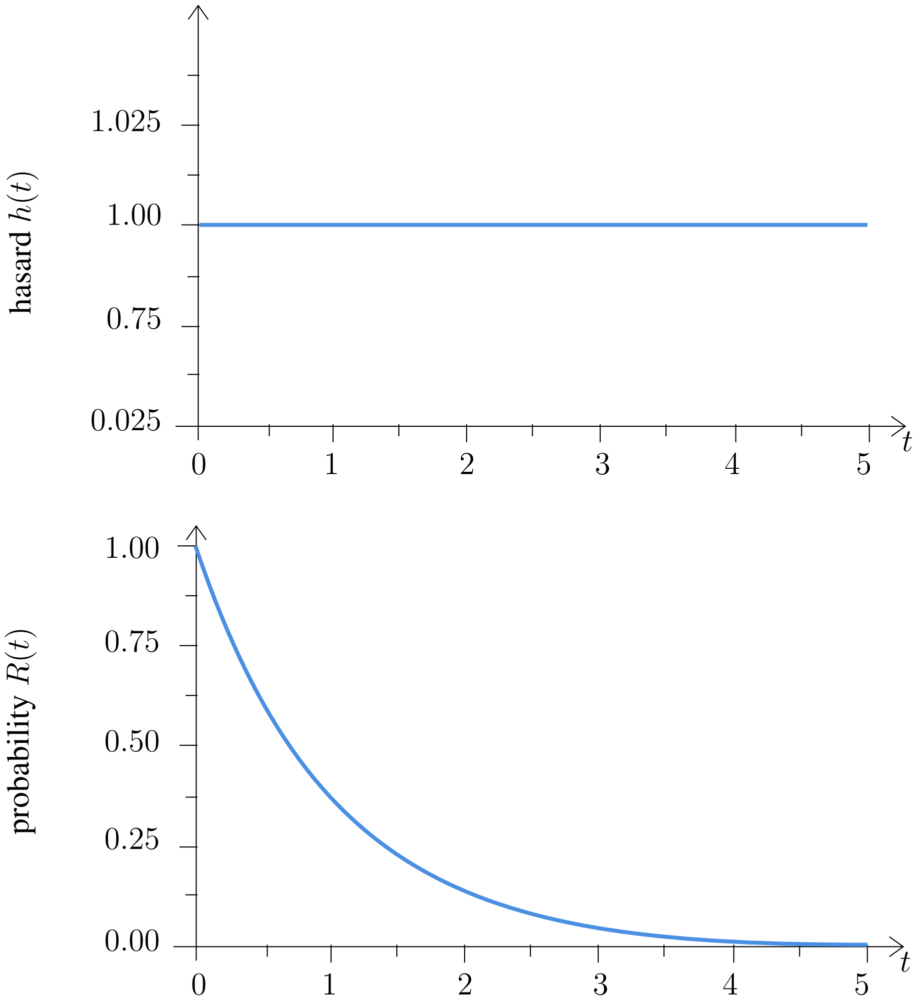


Figure 16.210: **Plot of the hazard function and reliability of the exponential law**

The density function of faulty components at time t is obviously:

$$f(t)=\frac{\mathrm{d}F(t)}{\mathrm{d}t}=\lambda e^{-\lambda t} \qquad (16.7.201)$$

It therefore follows an exponential law! This law and its moments are known to us (see section Statistics page 751). It then becomes easier to determine the M.T.T.F. and its standard deviation (inverse of failure rate) and also a confidence interval for it.

Furthermore, if we calculate the reliability $R(t)$ at the time corresponding to the M.T.T.F. (inverse of the failure rate in the case of the exponential law) we will always obtain a cumulative probability of $F(t)=1-e^{-1}=36.8\%$ (so basically almost one chance on three to run at this time and 2 chances on 3 to be down) and not 50% as sometimes we can intuitively think (this is the case only if the probability distribution is symmetric).

Since the technical tables of reliability in the industry almost always assume the reliability rate to be constant then we better understand the importance of it (we will have an example in our presentation further below of topological systems).

Also remember that we saw in the section of Probability (see page 480) that if we have a set of independent events (independent mechanisms) with a given probability, then the probabilities calculated on the whole as a globality involve the multiplication of probabilities!

Therefore in a mechanism having independent but essential parts (system said to be "serial system") the overall reliability density function will be equal to the multiplication of cumulative probability distributions $R(t)$ and which is therefore equivalent in the case of an exponential distribution having a single exponential law whose overall failure rate is equal to the sum of the various parts failure rate (i.e. $e^{\lambda_1} \cdot e^{\lambda_2}) \cdot \ldots = e^{\lambda_1 + \lambda_1 + \cdots}$!

Another example... in mechanics, where the phenomenon of wear is the cause of the failure, the failure rate is often linear (so it increases steadily over time and is not zero at the initial starting of the device):

$$\lambda(t) = at + b \tag{16.7.202}$$

Then:

$$F(t) = 1 - \exp\left(-\int_0^t (at+b)\,\mathrm{d}t\right) = 1 - \exp\left(-\frac{at^2}{2} - bt\right) \tag{16.7.203}$$

Therefore:

$$f(t) = \frac{\mathrm{d}F(t)}{\mathrm{d}t} = (at+b)\exp\left(-\frac{at^2}{2} - bt\right) \tag{16.7.204}$$

As (for an equipment whose repair time is negligible):

$$\text{MTBF} = \int_0^{+\infty} R(t)\mathrm{d}t = \int_0^{+\infty} \exp\left(-\frac{at^2}{2} - bt\right) \mathrm{d}t \tag{16.7.205}$$

this integral can be calculated by numerical integration methods (see section Statistics page 6621). Therefore, in practice, we prefer to take closed form type of law thanks to goodness of fit type statistical tests (see section Statistics page 1111), such as the Weibull distribution (see just below for detailed developments), which is a little bit the trash to put everything in the field of reliability engineering...

In common and free public reliability data banks on the market are normally given: the name of the material or component, the M.T.B.F., the average default rate, the statistical heritage (...), a multiplicative factor of failure rates depending on the environment or usage constraints.

16.7.3.2.1 Average Failure Rate

As we have seen, the hazard function usually varies with time. What is sometimes of interest to the practitioner is the average rate of failures within a given time interval, which is clearly specified by the functional specification. We then define the "average failure rate" as:

$$\begin{aligned}\text{AFR}(t_1,t_2) &= \frac{\displaystyle\int_{t_1}^{t_2} h(t)\mathrm{d}t}{t_2 - t_1} = \frac{\displaystyle\int_{t_1}^{t_2} h(t)\mathrm{d}t}{t_2 - t_1} = \frac{H(t_2) - H(t_1)}{t_2 - t_1}\\ &= \frac{-\ln(R(t_2)) - (-\ln(R(t_1)))}{t_2 - t_1} = \frac{\ln(R(t_1)) - \ln(R(t_2))}{t_2 - t_1}\end{aligned} \tag{16.7.206}$$

Therefore:

$$\boxed{\text{AFR}(t_1,t_2)=\frac{\ln(R(t_1))-\ln(R(t_2))}{t_2-t_1}} \tag{16.7.207}$$

This rate, denoted $\text{AFR}(t_1,t_2)$, is a single number that can be used as a specification or target for the population failure rate over that interval. If t_1 is 0, it is dropped from the expression. Thus, for example, $\text{AFR}(40,000)$ would be the average failure rate for the population over the first $40,000$ hours of operation.

An interval often taken in practice is 0 to τ. Then in comes:

$$\text{AFR}(0,\tau)=\text{AFR}(\tau)=\frac{H(\tau)}{\tau}=\frac{-\ln(R(\tau))}{\tau} \tag{16.7.208}$$

Sometimes in the literature considering the following variation:

$$\text{AFR}(\tau)=\frac{-\ln(R(\tau))}{\tau}=\frac{-\ln(1-F(\tau))}{\tau} \tag{16.7.209}$$

and considering that $F(\tau)$ is small, by a Taylor development to the second order, we have:

$$\text{AFR}(\tau)\cong\frac{-(0-F(\tau))}{\tau}=\frac{F(\tau)}{\tau} \tag{16.7.210}$$

We can transform the antecedent relations into a failure distribution function:

$$-\tau\text{AFR}(\tau)=\ln(R(\tau)) \tag{16.7.211}$$

and therefore:

$$e^{-\tau\text{AFR}(\tau)}=R(\tau)=1-F(\tau) \tag{16.7.212}$$

And finally:

$$\boxed{F(\tau)=1-e^{-\tau\text{AFR}(\tau)}} \tag{16.7.213}$$

16.7.3.3 Weibull Distribution

Once again maintenance techniques use probability and statistics so we refer the reader to the section of the same name in this textbook. However, there exists in the field of maintenance (and not only!) a probability density function widely used and named "Weibull law".

It seemed to be completely empirical and defined by:

$$f(x)=\frac{\beta}{\eta}\left(\frac{x-\gamma}{\eta}\right)^{\beta-1}e^{-\left(\frac{x-\gamma}{\eta}\right)^{\beta}} \tag{16.7.214}$$

where $f(x)\geq 0, x\geq 0$ with $\beta>0,\eta>0,-\infty<\gamma<+\infty$ that are named respectively "scale parameter" η, "shape parameter" β and "location parameter" γ.

Remark

The engineers in corporations have to refer to *IEC 61649 Weibull analysis* to make a standardized usage of this law!

The Weibull distribution is often written as following by putting $x:=x-\gamma$, $\eta:=\beta$, $\beta:=\alpha$ (in this case the location parameter is implicitly posed as equal to zero):

$$f(x)=\frac{\alpha}{\beta}\frac{x^{\alpha-1}}{\beta^{\alpha-1}}e^{\left(-\frac{x^{\alpha}}{\beta^{\alpha}}\right)}=\frac{\alpha}{\beta^{\alpha}}x^{\alpha-1}e^{\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right)} \tag{16.7.215}$$

It can be calculated in the English version of Microsoft® Excel® 11.8346 in this latter form using `WEIBULL( )` function.

The Weibull law can be seen as a generalization of the exponential distribution function (we'll see why later!) with the advantage that it is possible to play with the three parameters to obtain almost anything.

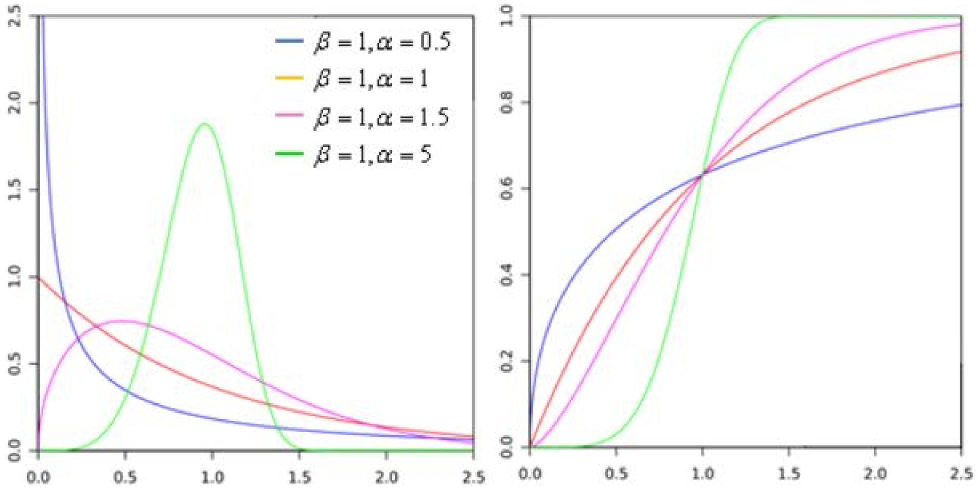


Figure 16.211: Weibull density and distribution functions for different α values (source: Wikipedia)

By cancelling the location parameter γ, we get the "two-parameters Weibull distribution":

$$f(x)=\frac{\beta}{\eta}\left(\frac{x}{\eta}\right)^{\beta-1}e^{-\left(\frac{x}{\eta}\right)^{\beta}} \tag{16.7.216}$$

which has an important practical application and for which we have calculated the estimators of the parameters in the Statistics section (see page 827).

By assuming again $\gamma=0$ and assuming also that we have $\beta=C=c^{te}$ we get "one-parameter Weibull distribution" (it's a little stupid as a definition but never-mind...):

$$f(x)=\frac{C}{\eta}\left(\frac{x}{\eta}\right)^{C-1} e^{-\left(\frac{x}{\eta}\right)^C} \tag{16.7.217}$$

where the only unknown parameter is the scale parameter η. We assume that the parameter β is known a priori from past experiences on similar samples. Notice that the two-parameter Weibull distribution function a is the derivative of the following repartition function (cumulative probability function or "unreliability" as say some practitioners in the field of maintenance...) where there are three successive interiors derivatives to fall back on the Weibull distribution function ... which is a good application of this we saw in the section of Differential and Integral Calculus:

$$F(x)=1-e^{\left(-\left(\frac{x}{\beta}\right)^\alpha\right)} \tag{16.7.218}$$

It follows therefore in reliability engineering that in this special case, we have then:

$$R(t)=1-F(t)=F(x)=1-\left(1-e^{\left(-\left(\frac{t}{\beta}\right)^\alpha\right)}\right)=e^{\left(-\left(\frac{t}{\beta}\right)^\alpha\right)} \tag{16.7.219}$$

We also find sometimes this relation in the literature in the form:

$$\begin{aligned}f(x)&=C\frac{1}{\eta}\left(\frac{1}{\eta}\right)^{C-1} x^{C-1}e^{-\left(\frac{x}{\eta}\right)^C}=C\left(\frac{1}{\eta}\right)^{C} x^{C-1}e^{\left(-\left(\frac{1}{\eta}\right)^C x^C\right)}\\&=abx^{\alpha-1}e^{-bx^c}\end{aligned} \tag{16.7.220}$$

Note that if $\beta=1$, then we fall back on an exponential distribution of mean η:

$$f(x)=\frac{\beta}{\eta}\left(\frac{x}{\eta}\right)^{\beta-1} e^{-\left(\frac{x}{\eta}\right)^\beta}=\frac{1}{\eta}e^{-\frac{x}{\eta}} \tag{16.7.221}$$

The M.T.B.F. is then given by the mean of the Weibull distribution with non-zero γ:

$$\mathrm{E}(X)=\int\limits_{\gamma}^{+\infty} xf(x)\mathrm{d}x=\int\limits_{\gamma}^{+\infty} x\frac{\beta}{\eta}\left(\frac{x-\gamma}{\eta}\right)^{\beta-1} e^{-\left(\frac{x-\gamma}{\eta}\right)^\beta}\mathrm{d}x \tag{16.7.222}$$

Let us put:

$$t=\left(\frac{x-\gamma}{\eta}\right)^\beta \quad\Rightarrow\quad \mathrm{d}t=\frac{\beta}{x-\gamma}\left(\frac{x-\gamma}{\eta}\right)^\beta\mathrm{d}x=\frac{\beta}{x-\gamma}t\mathrm{d}x \quad\Rightarrow\quad \mathrm{d}x=\frac{x-\gamma}{\beta}t^{-1}\mathrm{d}t \tag{16.7.223}$$

with:

$$t=\left(\frac{x-\gamma}{\eta}\right)^\beta \Leftrightarrow t^{1/\beta}\eta+\gamma=x \tag{16.7.224}$$

What gives:

$$\begin{aligned}\mathrm{E}(X)&=\int\limits_0^{+\infty} x\frac{\beta}{\eta}t\left(\frac{x-\gamma}{\beta}\right)e^{-t}\left(\frac{x-\gamma}{\beta}t^{-1}\right)\mathrm{d}t\\&=\int\limits_0^{+\infty} xe^{-t}\mathrm{d}t=\int\limits_0^{+\infty}(t^{1/\beta}\eta+\gamma)e^{-t}\mathrm{d}t=\eta\int\limits_0^{+\infty}t^{1/\beta}e^{-t}\mathrm{d}t+\gamma\end{aligned}\tag{16.7.225}$$

The first integral above is already known to us, we have already study it in the section of Integral and Differential Calculus (see page 1574). This is the gamma Euler function:

$$\Gamma_{-1}(x)=\int\limits_0^{+\infty}t^{x-1}e^{-t}\mathrm{d}t=x!\quad\Rightarrow\quad\Gamma_{-1}\left(\frac{1+\beta}{\beta}\right)=\int\limits_0^{+\infty}t^{1/\beta}e^{-t}\mathrm{d}t\tag{16.7.226}$$

Therefore we have finally:

$$\mathrm{E}(X)=\text{MTBF}=\eta\int\limits_0^{+\infty}t^{1/\beta}e^{-t}\mathrm{d}t+\gamma=\gamma+\eta\Gamma_{-1}\left(\frac{1+\beta}{\beta}\right)\tag{16.7.227}$$

By cancelling the location parameter γ we get the following current case in reliability:

$$\boxed{\mathrm{E}(X)=\text{MTBF}=\eta\Gamma_{-1}\left(1+\frac{1}{\beta}\right)}\tag{16.7.228}$$

Remark

If then by chance $\beta^{-1}\in\mathbb{N}$, then as we have proved during our study of the gamma Euler function:

$$\mathrm{E}(X)=\text{MTBF}=\gamma+\eta\Gamma_0\left(\frac{1}{\beta}\right)=\gamma+\eta\left(\frac{1}{\beta}\right)!\tag{16.7.229}$$

In case where $\beta^{-1}\in\mathbb{R}$ we must make use of the numerical tables obtained by the algorithms seen in the section of Theoretical Computing.

Similarly:

$$\begin{aligned}\mathrm{E}(X^2)&=\int\limits_0^{+\infty}x^2f(x)\mathrm{d}x=\int\limits_0^{+\infty}(t^{1/\beta}\eta+\gamma)^2e^{-t}\mathrm{d}t\\&=\eta^2\Gamma_{-1}\left(\frac{\beta+2}{\beta}\right)+2\eta\gamma\Gamma_{-1}\left(\frac{\beta+1}{\beta}\right)+\gamma^2\end{aligned}\tag{16.7.230}$$

Finally when the location parameter is equal to zero, we get:

$$\sigma^2=\mathrm{V}(X)=\mathrm{E}(X^2)-\mathrm{E}(X)^2=\eta^2\left(\Gamma_{-1}\left(\frac{\beta+2}{\beta}\right)-\Gamma_{-1}\left(\frac{\beta+1}{\beta}\right)^2\right)\tag{16.7.231}$$

Remark

Some specialists, when it is made usage of two moments of order two in the probabilistic analysis of reliability, sometimes talk about "mean square approach method"...

To conclude on the Weibull law, note that it is customary to consider four important and relevant situations of the Weibull distribution based on the values of its parameters:

1. Survival density function looks like a rapid exponential decay and therefore the hazard function (failure rate) is high initially and decreases over time (first portion of the "bathtube" hazard function/failure rate). This is typically the situation where the failures are early because failures occur in the initial period of the life of the product.
2. Survival density looks like an exponential decay and therefore the hazard function (failure rate) is constant over time (second portion of the "bathtube" hazard function/failure rate). This is typically a situation where failures are random and sources of multiple causes. In this situation, the Weibull models the phase of "useful life" of the product.
3. The survival density function rises to a peak and then decreases with a fairly pronounced asymmetry (high density on the left) and therefore the hazard function (failure rate) increases rapidly at first and eventually stabilize. This is typically a situation where the initial wear is very fast.
4. The survival density function looks like a half large Gaussian Law and the hazard function (failure rate) is growing rapidly (exponential type of growth). This is typically the situation where we reach the end of life of a set of parts (third portion of the "bathtube" hazard function/failure rate).

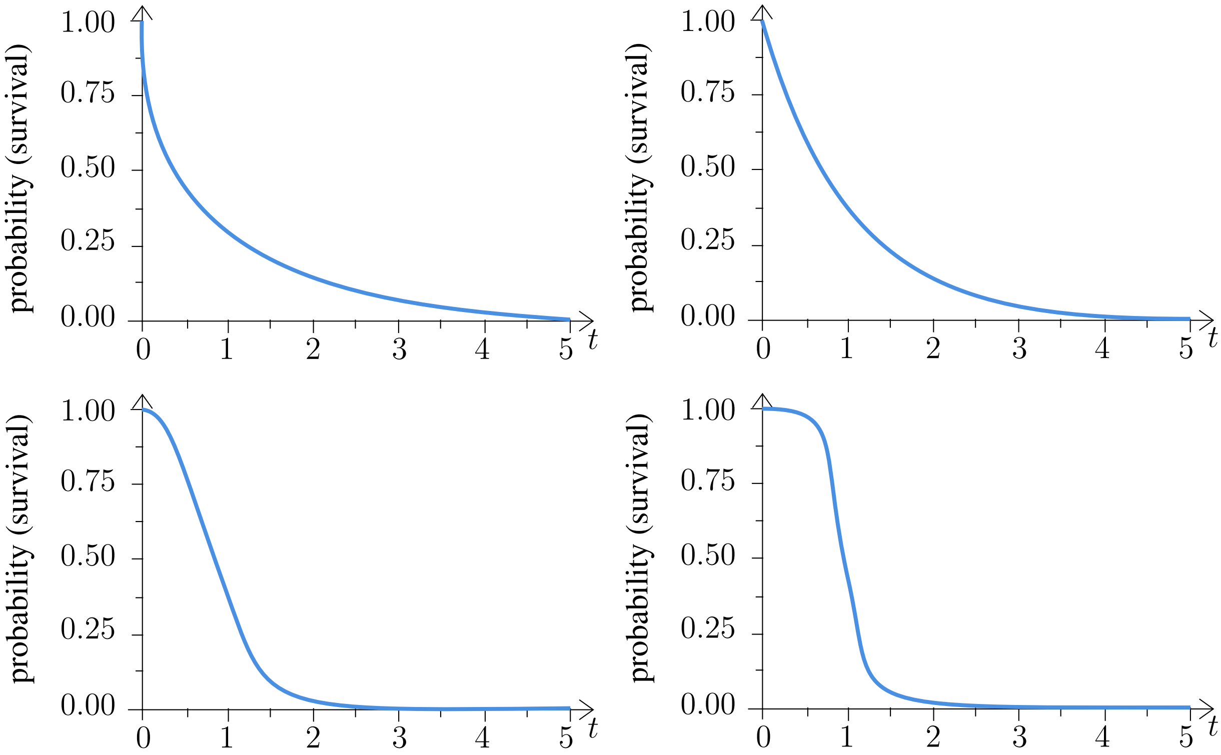


Figure 16.212: **Plot of four Weibull survival functions for different parameter values**

And with the corresponding hazard functions:

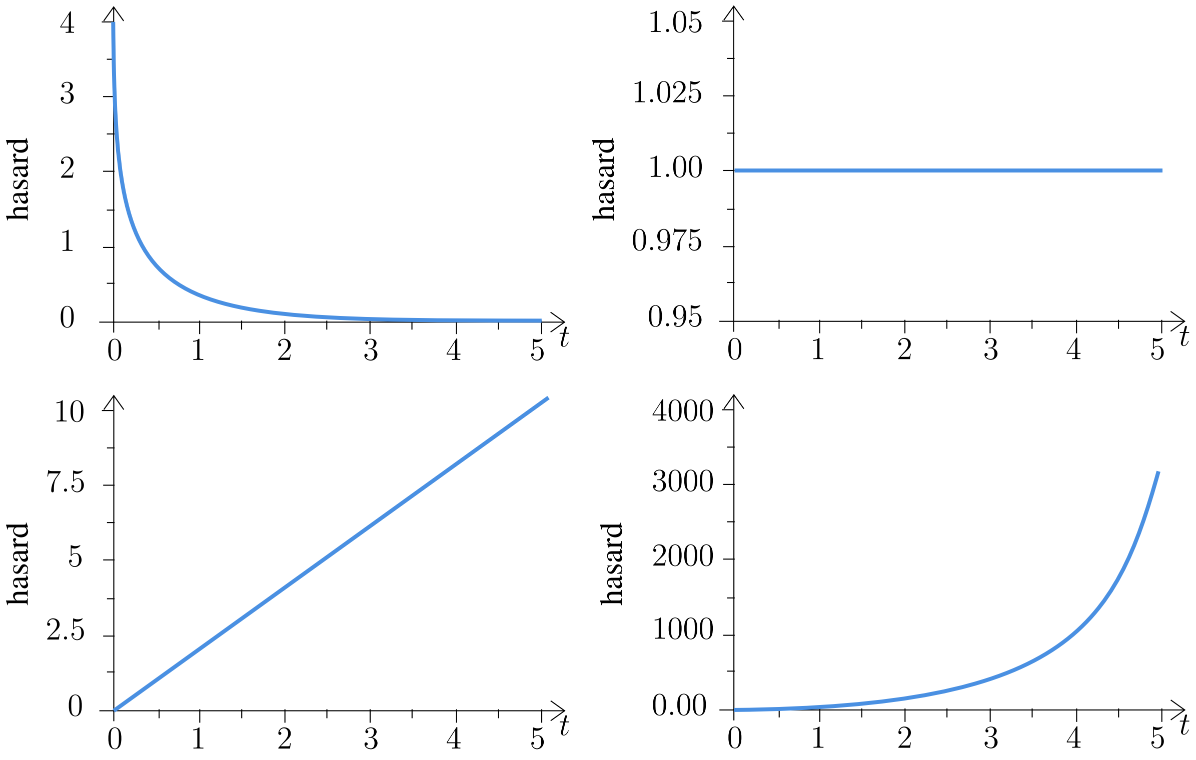


Figure 16.213: **Plot of four Weibull hazard functions for different parameter values**

To summarize we have therefore for the Weibull distribution with two parameters:

$$\boxed{\begin{aligned}
F(x)&=1-e^{-\left(\frac{x}{\eta}\right)^{\beta}}:=1-e^{-(\lambda x)^{\beta}}=1-e^{-(\lambda t)^{\beta}}=F(t)\\
f(x)&=\frac{\beta}{\eta}\left(\frac{x}{\eta}\right)^{\beta-1}e^{-\left(\frac{x}{\eta}\right)^{\beta}}=\lambda\beta(\lambda t)^{\beta-1}e^{-(\lambda t)^{\beta}}=\lambda^{\beta}\beta t^{\beta-1}e^{-(\lambda t)^{\beta}}=f(t)\\
R(x)&=e^{-\left(\frac{x}{\eta}\right)^{\beta}}=e^{-(\lambda t)^{\beta}}=R(t)\\
h(x)&=\frac{\beta}{\eta}\left(\frac{x}{\eta}\right)^{\beta-1}=\lambda^{\beta}\beta t^{\beta-1}=h(t)
\end{aligned}} \qquad (16.7.232)$$

So we see that if we put $\beta=1$ we fall back on all the reliability functions based on the exponential law.

✍**Example:**

A car manufacturer knows the distribution of time until the first failure of one of its vehicles under average driving conditions. The Weibull distribution function obtained by the tests is a two-parameter Weibull function given by:

$$f(x)=1-e^{-\left(\frac{x}{\eta}\right)^{\beta}}=1-e^{-\left(7.7\cdot 10^{-6}x\right)^{1.22}} \tag{16.7.233}$$

where x is the distance in kilometres of the tested vehicles. If we want to find the warranty in kilometres encompassing 5% of the vehicles, then we must solve the equation:

$$1-e^{-\left(7.7\cdot 10^{-6}x\right)^{1.22}}=5\% \tag{16.7.234}$$

That is to say with some elementary algebra:

$$t=\frac{\left[-\ln(95\%)\right]^{1/1.22}}{7.7\cdot 10^{-6}}=11,138\,[\text{km}] \tag{16.7.235}$$

16.7.3.3.1 Two-parameter Weibull distribution linearisation

In the case of the 2-parameter Weibull, the CDF (also the "unreliability" for recall) is given as we have just seen by (we use the temporal variables t instead of x now):

$$F(t)=1-e^{-\left(\frac{t}{\eta}\right)^{\beta}} \tag{16.7.236}$$

This function can then be linearised (i.e., put in the common form of $y=ax+b$ format) as follows for practical regression (adequation) analysis.

We start from the previous relation and rearrange the whole:

$$1-F(t)=e^{-\left(\frac{t}{\eta}\right)^{\beta}} \tag{16.7.237}$$

and we take the natural logarithm:

$$\ln(1-F(t))=\ln\left[e^{-\left(\frac{t}{\eta}\right)^{\beta}}\right] \tag{16.7.238}$$

Therefore:

$$\ln(1-F(t))=-\left(\frac{t}{\eta}\right)^{\beta} \tag{16.7.239}$$

Rearranging and taking the natural logarithm again we get:

$$\ln\left(-\ln(1-F(t))\right)=\beta\left(\ln\left(\frac{t}{\eta}\right)\right) \tag{16.7.240}$$

Using the properties of the logarithm (see section Functional Analysis page 2284):

$$\boxed{\ln\left(\ln\left(\frac{1}{1-F(t)}\right)\right)=\beta\ln(t)-\beta\ln(\eta)} \tag{16.7.241}$$

Then by setting:

$$y=\ln\left(\ln\left(\frac{1}{1-F(t)}\right)\right) \tag{16.7.242}$$

and:

$$x=\ln(t) \tag{16.7.243}$$

the equation can then be rewritten as:

$$y=\beta x-\beta\ln(\eta) \tag{16.7.244}$$

which is now a linear equation with a slope of:

$$a=\beta \tag{16.7.245}$$

and an intercept of:

$$b=-\beta\ln(\eta) \tag{16.7.246}$$

✍**Example:**

Let us now calculate as application example the Average Failure Rate of this Weibull law:

$$\begin{aligned}\text{AFR}(t_1,t_2)&=\frac{\ln(R(t_1))-\ln(R(t_2))}{t_2-t_1}=\frac{\ln\left(e^{-\left(\frac{t_1}{\eta}\right)^\beta}\right)-\ln\left(e^{-\left(\frac{t_2}{\eta}\right)^\beta}\right)}{t_2-t_1}\\&=\frac{-\left(\frac{t_1}{\eta}\right)^\beta-\left(-\left(\frac{t_2}{\eta}\right)\right)}{t_2-t_1}=\frac{\left(\frac{t_2}{\eta}\right)^\beta-\left(\frac{t_1}{\eta}\right)^\beta}{t_2-t_1}\end{aligned} \tag{16.7.247}$$

In the case of a time interval of the type $[0,\tau]$ we then have:

$$\text{AFR}(t_1,t_2)=\frac{\left(\frac{\tau}{\eta}\right)^\beta}{\tau}=\frac{1}{\eta}\left(\frac{\tau}{\eta}\right)^{\beta-1} \tag{16.7.248}$$

16.7.3.4 Topology of Systems

When working with non-repairable real systems (mechanical, electronic or other), we face different constraints depending on the type of installation that we have. The study of such systems is also named "Reliability Block Diagram[21]".

The two main assumptions of the study of these systems are:

1. The failure of a component is independent of the others.
2. No breakdowns arrive at the same time.

We recognize five main topologies in which each component is represented by a block:

T1. Serial topology:

If only one component fails the whole system does not work anymore (the examples are so numerous and easy to find that we did not mention any of them).

So in the hypothesis that the failure of a component is independent of the other, the cumulative probability of failure is the product of cumulative probabilities (see section Probabilities page 480).

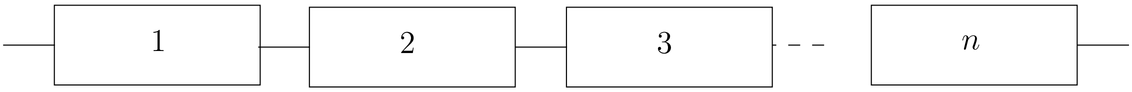


Figure 16.214: **Serial topology**

In the study of probabilistic fault tree (see further below), this topology corresponds to an OR gate with n inputs, as it is sufficient that one blocks may be down so that the output no longer works:

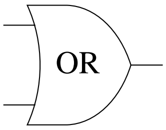


Figure 16.215: **OR gate**

As often practitioners made use of the exponential distribution!

The multiplication of several terms of the accumulated probabilities are relatively long to write, we prefer the use of the cumulative probability of reliability R.

Thus, the reliability (probability of operation) of a series system is given by:

$$\boxed{R_S = 1 - F_S = \prod_{i=1}^{n} R_i} \tag{16.7.249}$$

---

[21] The mathematical tools that are presented and build here can also be used to calculate the reliability of humans communication chains. This explained quite well why testimonials - oral or written - have zero scientifical values (but however don't have zero juridical value...).

Which brings us well to zero reliability if at least one block has zero reliability.

What is traditionally noted in the field:

$$\varphi(x)=\min(x_1,x_2,\ldots,x_n)=\prod_{i=1}^{n}x_i \tag{16.7.250}$$

Remark

The attentive reader will have noticed that the serial system is always less reliable than its least reliable component!

In the case of electronic components, the failure rate is often regarded as constant for simplicity and the density function is then that of an exponential law (solution of a well known differential equation):

$$f(t)=\lambda e^{-\lambda t} \tag{16.7.251}$$

However, we have proved that in the case of a non-repairable system:

$$\mathrm{E}(t)=\int_{0}^{+\infty}tf(t)\mathrm{d}t=\text{MTBF} \tag{16.7.252}$$

And as we have proved it in the section Statistics (see page 751) the expected mean of the exponential law:

$$\mathrm{E}(t)=\int_{0}^{+\infty}t\lambda e^{-\lambda t}\mathrm{d}t=\frac{1}{\lambda}=\text{MTBF} \tag{16.7.253}$$

We have also proved earlier that:

$$f(t)=\frac{\mathrm{d}F}{\mathrm{d}t}=-\frac{\mathrm{d}R}{\mathrm{d}t} \tag{16.7.254}$$

Therefore:

$$R(t)=e^{-\lambda t}=e^{-\frac{1}{\text{MTBF}}t} \tag{16.7.255}$$

So for a series constant failure rate system:

$$R_S=\prod_i e^{-\lambda_i t}=\prod_i e^{-\frac{1}{\text{MTBF}_i}t}=e^{-\left(\sum_i\frac{1}{\text{MTBF}_i}\right)t} \tag{16.7.256}$$

Thus, in this special case:

$$\text{MTBF}_S=\left(\sum_{i=1}^{n}\frac{1}{\text{MTBF}_i}\right)^{-1} \tag{16.7.257}$$

or written differently:

$$\lambda_S=\sum_{i=1}^{n}\lambda_i \tag{16.7.258}$$

and... the problem is that the a lot of Internet pages that talk about serial topology systems (or parallel) implicitly use an exponential law, this is why sometimes beginner practitioners use this last relation without having studied the underlying mathematical details and this brings them to huge errors in the reliability of the products their are in charge to analyse.

T2. Parallel topology:

Unlike the previous system, parallel system continues to operate if at least one component still works (typically redundancy systems in airplanes, rockets, nuclear power plants or critical computer systems). Therefore is considered as the most important system to study.

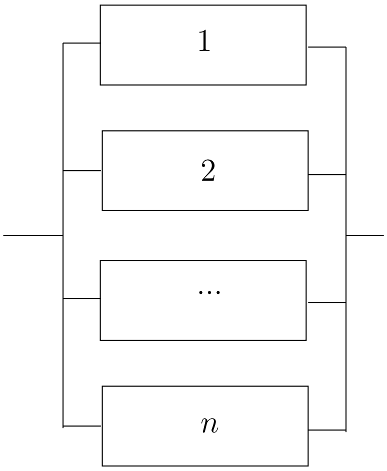


Figure 16.216: **Parallel topology**

In the study of probabilistic fault tree (see further below), this topology corresponds to an AND gate with n inputs, as it is necessary that all blocks are down so that the output no longer works:

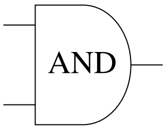


Figure 16.217: **AND gate**

In other words, it stops working if and only if the cumulated probability failure is equal to:

$$\prod_{i=1}^{n} F_i = \prod_{i=1}^{n} (1-R_i) = 1 \tag{16.7.259}$$

So it comes immediately for a purely parallel system:

$$\boxed{R_P = 1 - \prod_{i=1}^{n} F_i = 1 - \prod_{i=1}^{n} (1-R_i)} \tag{16.7.260}$$

In the very common case where the blocks are redundant with equal reliability $R_i = R$ and that R_P is fixed by the customer we must found a component such that is reliability is therefore equal to:

$$\boxed{R_P = 1 - \sqrt[n]{1 - R_P}} \tag{16.7.261}$$

Remark

In Project management this type of parallel structure is typically used when one wishes to know the total reliability of several supplier of identical parts (in this case if any of them has delays or other problems, the others must afford to mitigate the negative impact) knowing or believing obviously the reliability of each...

We therefore have the parallel system whose components have a constant failure rate and all identical (to simplify the study):

$$R_P = 1 - \prod_{i=1}^{n}(1 - R_i) = 1 - \prod_{i=1}^{n}\left(1 - e^{-\lambda t}\right) \tag{16.7.262}$$

Therefore we have:

$$\text{MTBF}_S = \int_0^{+\infty}\left[1 - \prod_{i=1}^{n}\left(1 - e^{-\lambda t}\right)\right]\mathrm{d}t = \int_0^{+\infty}\left[1 - \left(1 - e^{-\lambda t}\right)^n\right]\mathrm{d}t \tag{16.7.263}$$

If we put:

$$u = 1 - e^{-\lambda t} \quad \Rightarrow \quad \mathrm{d}u = \lambda e^{-\lambda t}\mathrm{d}t = \lambda(1 - u)\mathrm{d}t \tag{16.7.264}$$

and replacing in the prior-previous expression we get:

$$\text{MTBF}_S = \int_0^1 \frac{1 - u^n}{\lambda(1 - u)}\mathrm{d}u = \frac{1}{\lambda}\int_0^1 \frac{u^n - 1}{u - 1}\mathrm{d}u \tag{16.7.265}$$

As far as we know is not possible to integrate this latter relation, but if we compared for different values of n fixed then we see quickly that:

$$\text{MTBF}_S = \frac{1}{\lambda}\sum_{i=1}^{n}\frac{1}{i} \tag{16.7.266}$$

or written in another way:

$$\lambda_S = \lambda\left(\sum_{i=1}^{n}\frac{1}{i}\right)^{-1} \tag{16.7.267}$$

Let us still do the detailed development for a parallel system with two components whose failure rate is constant and not identical. For this, we start from the relation proved earlier above:

$$R_P = 1 - \prod_{i=1}^{n} F_i = 1 - \prod_{i=1}^{n}(1 - R_i) \tag{16.7.268}$$

Either in the case where n is equal to 2, we have:

$$R_P = 1-(1-R_1)(1-R_2) = R_1+R_2-R_1R_2 \tag{16.7.269}$$

and therefore:

$$\begin{aligned}
\text{MTBF}_S &= \int\limits_0^{+\infty}\left[1-\prod_{i=1}^{2}(1-R_i)\right]\text{d}t = \int\limits_0^{+\infty}(R_1+R_2-R_1R_2)\text{d}t \\
&= \int\limits_0^{+\infty} e^{-\lambda_1 t}\text{d}t + \int\limits_0^{+\infty} e^{-\lambda_2 t}\text{d}t - \int\limits_0^{+\infty} e^{-\lambda_1 t}e^{-\lambda_2 t}\text{d}t \\
&= \int\limits_0^{+\infty} e^{-\lambda_1 t}\text{d}t + \int\limits_0^{+\infty} e^{-\lambda_2 t}\text{d}t - \int\limits_0^{+\infty} e^{-(\lambda_1+\lambda_2) t}\text{d}t \\
&= -\frac{1}{\lambda_1}e^{-\lambda_1 t}\Big|_0^{+\infty} - \frac{1}{\lambda_2}e^{-\lambda_2 t}\Big|_0^{+\infty} + -\frac{1}{\lambda_1+\lambda_2}e^{-(\lambda_1+\lambda_2)t}\Big|_0^{+\infty} = \frac{1}{\lambda_1}+\frac{1}{\lambda_2}-\frac{1}{\lambda_1+\lambda_2}
\end{aligned} \tag{16.7.270}$$

So if $\lambda_1=\lambda_2$, we have:

$$\text{MTBF}_S = \frac{1}{\lambda}+\frac{1}{\lambda}-\frac{1}{\lambda+\lambda} = \frac{2}{\lambda}-\frac{1}{2\lambda} = \frac{3}{2\lambda} \tag{16.7.271}$$

Therefore:

$$\lambda_S = \frac{2}{3}\lambda \tag{16.7.272}$$

Before continuing with other topologies, here is an example of a small school assignment (with solution!) mixing parallel/serial topology that shows that this does not apply only to machines.

✍**Example:**

A laborant must perform three different tests on a sample set: Tests A, B and C in the given order! We know that the laborant miss the test A one out of four times but may start it again 2 times in case of failure (so they can do it 3 times in total), they succeed the test B in 96% of times and may also start it again only once on failure (so they can do it 2 times in total), they fail the test C half of time and can not start it again. What is the probability that a laborant having successfully made the 3 lab tests in only one time (without failed)?

To solve this example we have to see each of the reproducible test as a parallel process and all the tests as a serial process. We then have the total probability that is given by:

$$R_P = \underbrace{\underbrace{\left(1-\prod_{i=1}^{3}F_{A,i}\right)}_{R_{S,A,\text{parallel}}}\underbrace{\left(1-\prod_{i=1}^{2}F_{B,i}\right)}_{R_{S,B,\text{parallel}}}(1-F_C)}_{R_A R_B R_C} \tag{16.7.273}$$

$$= \left(1-\left(\frac{1}{4}\right)^3\right)\left(1-(4\%)^2\right)(1-50\%) = 41.34\%$$

T3. Topology k/n: This system works if k out of n components of same reliability law R still work. This is typically the case of RAID hard disks computer where it is necessary to more than one to always works in the end and this number is determined by the volume of data. It is also operating principle of redundancy voting system on airplanes where we have a $2/3$ system named "triplex".

We then have the following schematic representation:

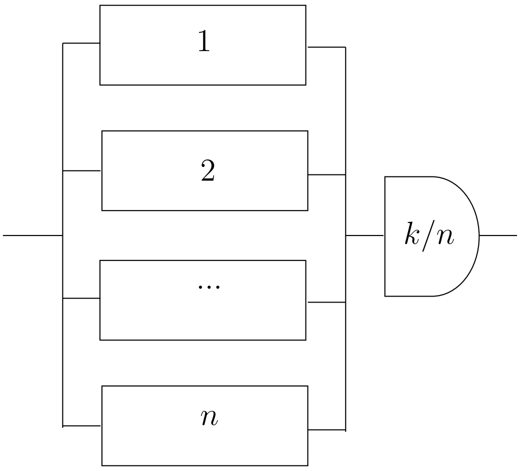


Figure 16.218: k/n **topology**

In the study of probabilistic fault tree (see further below), this topology corresponds to a k-OR (voting OR), as it is necessary that k blocks are down so that the output no longer works:

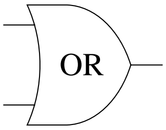


Figure 16.219: **Gate** k**-OR (voting OR)**

Therefore seek the cumulative probability law of reliability returns to the question of the cumulative probability of having k elements that work among n that are indistinguishable.

This therefore is equivalent to use the binomial distribution (see section Statistics page

658) for which we have proved that the cumulative probability was given by:

$$\boxed{R_{k/n}=\sum_{x=k}^{n}C_x^nR^x(1-R)^{n-x}} \tag{16.7.274}$$

This relationship is extremely important in the practice of reliability engineering !!! We notice as well as for $k=n$ we find the reliability of a series system whose elements all have the same reliability:

$$R_{k/n}=\sum_{x=n}^{n}C_x^nR^x(1-R)^{n-x}=C_n^nR^n(1-R)^{n-n}=\frac{n!}{n!(n-n!)}R^n=R^n \tag{16.7.275}$$

And for $k=1$, we fall back on the reliability of a parallel system whose elements all have the same reliability (remember the properties of the binomial coefficient in the section Calculus if necessary):

$$\begin{aligned}R_{k/n}&=\sum_{x=1}^{n}C_x^nR^x(1-R)^{n-x}=\sum_{x=0}^{n}C_x^nR^x(1-R)^{n-x}-C_0^nR^0(1-R)^{n-0}\\&=(R+(1-R))^n-C_0^n(1-R)^n=1^n-(1-R)^n=1-(1-R)^n\end{aligned} \tag{16.7.276}$$

✍**Example:**

In the case of the triplex, we have therefore:

$$\begin{aligned}R_{k/n}&=\sum_{x=2}^{3}C_x^nR^x(1-R)^{n-x}=\sum_{x=2}^{3}\frac{3!}{x!(3-x)!}R^x(1-R)^{3-x}\\&=\frac{3!}{2!(3-2!)}R^2(1-R)^{3-2}+\frac{3!}{3!(3-3!)}R^3(1-R)^{3-3}\\&=\frac{3!}{2!1!}R^2(1-R)^1+\frac{3!}{3!0!}R^3(1-R)^0=3R^2(1-R)+R^3=3R^2-3R^3+R^3\\&=3R^2-2R^3\end{aligned} \tag{16.7.277}$$

So, if the reliability is given by an exponential law where we have already proved that:

$$R(t)=e^{-\lambda t}=e^{-\frac{1}{\text{MTBF}}t} \tag{16.7.278}$$

Then we have:

$$\text{MTBF}=\int_0^{+\infty}(3R^2(t)-2R^3(t))\mathrm{d}t=\int_0^{+\infty}(3e^{-2\lambda t}-2e^{-3\lambda t})\mathrm{d}t=\frac{5}{6\lambda}<\frac{1}{\lambda} \tag{16.7.279}$$

So the overall MTBF of the $2/3$ system is larger than a simple system which obviously the goal.

T4. Serial/Parallel and Parallel/Serial topology with symmetrical configuration:

These are just simple compositions of the first two previously studied systems.

We first have the series/parallel system:

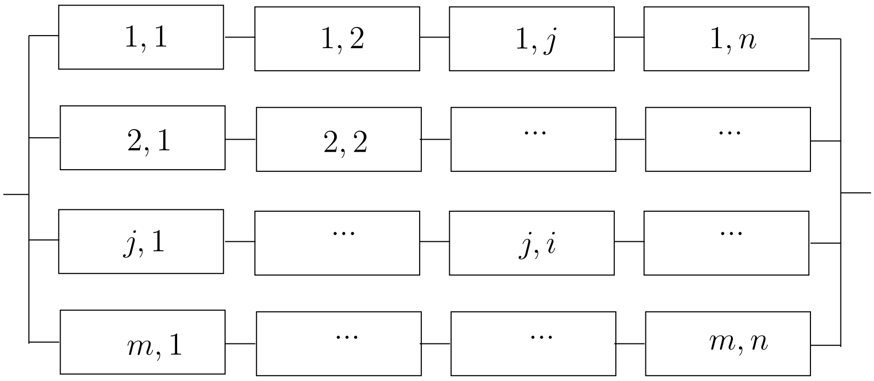

$1,1$
$1,2$
$1,j$
$1,n$
$2,1$
$2,2$
$\cdots$
$\cdots$
$j,1$
$\cdots$
j,i
$\cdots$
$m,1$
$\cdots$
$\cdots$
m,n

Figure 16.220: Serial/Parallel symmetric topology

However, as the series systems are given by:

$$R_S=1-F_S=\prod_{i=1}^{n}R_i \tag{16.7.280}$$

and the parallel by:

$$R_P=1-\prod_{i=1}^{n}F_i=1-\prod_{i=1}^{n}(1-R_i) \tag{16.7.281}$$

the composition of the two gives trivially in the case above:

$$\boxed{R_{S/P}=1-\prod_{j=1}^{m}(1-R_j)=\prod_{j=1}^{m}\left(1-\prod_{i=1}^{n}R_{ij}\right)} \tag{16.7.282}$$

And we also have in the family of parallel/series system:

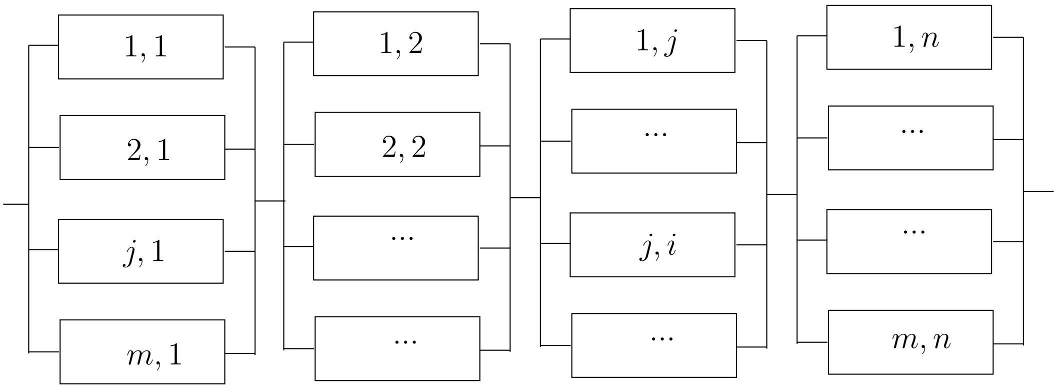

$1,1$
$1,2$
$1,j$
$1,n$
$2,1$
$2,2$
$\cdots$
$\cdots$
$j,1$
$\cdots$
j,i
$\cdots$
$m,1$
$\cdots$
$\cdots$
m,n

Figure 16.221: Parallel/Serial symmetric topology

when by using exactly the same approach as previously we have:

$$\boxed{R_{P/S}=\prod_{i=1}^{n}R_i=\prod_{i=1}^{n}\left(1-\prod_{j=1}^{m}(1-R_{ij})\right)} \tag{16.7.283}$$

T5. Complex Topology:

In fact, it is not really about complex systems but they just require a little knowledge of probability axioms. The particular example which will interest us is the following (typical cascading RLC filter):

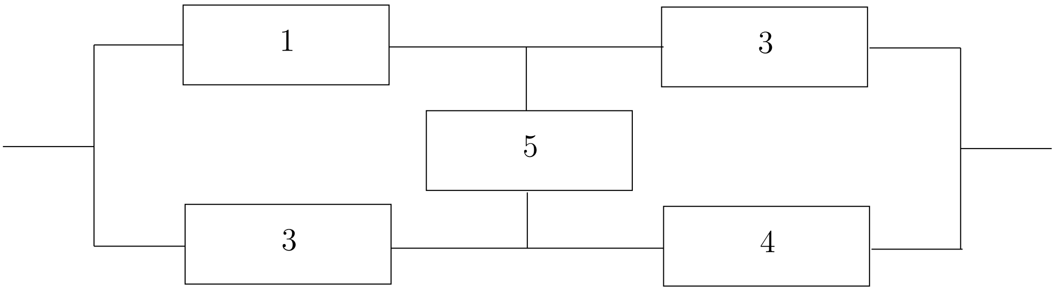


Figure 16.222: **Complex topology (bridge type)**

named "bridged network".

And we guess that what makes the complex system robust is the component number 5. We can then consider a first approach that is to break down the problem.

The system relatively to component 5 will be in one of the state:

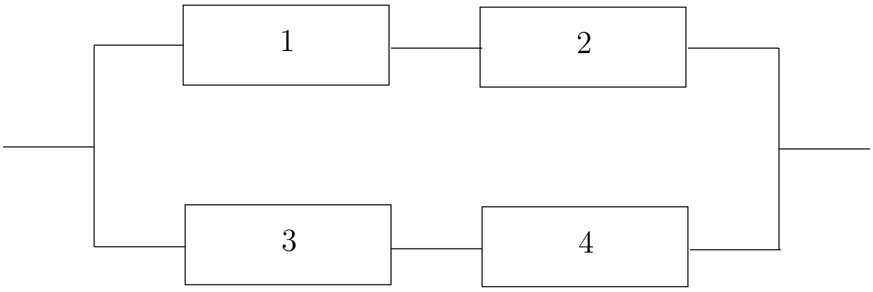


Figure 16.223: **Bridged Network system decomposition (first stage)**

if defective with a probability density function:

$$F_5 = 1 - R_5 \tag{16.7.284}$$

and having itself a reliability:

$$R_{S1} = 1 - \prod_{j=1}^{2}\left(1 - \prod_{i=1}^{2} R_{ij}\right) = 1 - (1 - R_1R_2)(1 - R_3R_4) \tag{16.7.285}$$

according to our previous results of complex series/parallel system.

Or in the following state if 5 is functional with a probability density law R_5:

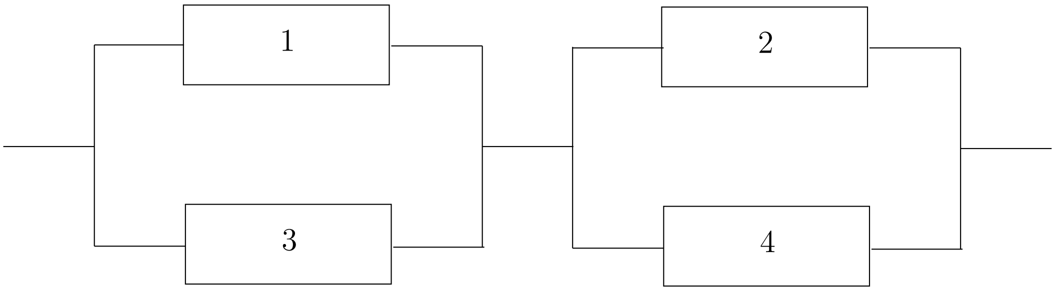


Figure 16.224: **Bridged Network system decomposition (second stage)**

and having itself a reliability of:

$$R_{S2}=\prod_{j=1}^{2}\left(1-\prod_{i=1}^{2}(1-R_{ij})\right)=(1-(1-R_1)(1-R_3))(1-(1-R_2)(1-R_4)) \tag{16.7.286}$$

according to our previous results on the parallel/serial complex system.

Since the system can not be in both states at the same time (at this time it is not a quantum system...), we are dealing with a disjoint probability (see section Probabilities page 453) that is to say the sum of the densities to which we must join the other components. Therefore we have:

$$\begin{aligned}&R_B=(1-R_5)R_{S1}+R_5R_{S2}\\&=(1-R_5)(1-(1-R_1R_2)(1-R_3R_4))+R_5((1-(1-R_1)(1-R_3))(1-(1-R_2)(1-R_4)))\end{aligned} \tag{16.7.287}$$

We can then using all 5 previous topologies evaluate the reliability of any system!

Another strategy for complex systems is to decompose them into simple series or parallel systems. If this is not possible, we can calculate the reliability of each system configuration that is considered to operate and then sum the probabilities of operation of each configuration.

✍**Example:**

Let's make an example by considering the following case:

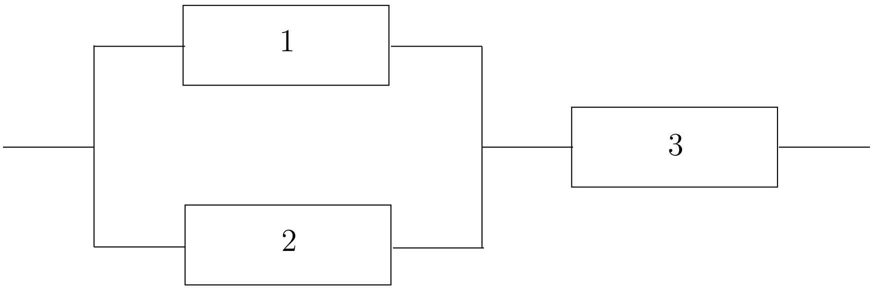


Figure 16.225: **Special case assumed as a complex system**

And consider the following truth table with the 2^n possible configurations:

State Number	Block 1	Block 2	Block 3	Output
1	0	0	0	0
2	1	0	0	0
3	0	1	0	0
4	1	0	1	1
5	0	1	1	1
6	1	1	1	1
7	1	1	0	0
8	0	0	1	0

Table 16.22: **Truth table of a preventive maintenance system**

The principle (not always easy to guess...) is to say that an UP state (thus worth equal 1) is assigned to the value R_1 of the concerned block i and the DOWN state (thus worth equal 0) is assigned to the value $1-R_i$ of the concerned block i. Each value will be multiplied with others to get the total reliability of the system in a given state.

Thus, in the previous example, we have 3 states that allow the system to operate. Let us calculate their respective reliability. The states No 4 and No 5 thus provide by definition:

$$R_{S4}=R_1R_3(1-R_2) \qquad R_{S5}=R_2R_3(1-R_1) \tag{16.7.288}$$

The State Number 6 gives him by definition:

$$R_{S6}=R_1R_2R_3 \tag{16.7.289}$$

and we sum all as previously indicated:

$$\begin{aligned}R_S&=R_{S4}+R_{S5}+R_{S6}=R_1R_3(1-R_2)+R_2R_3(1-R_1)+R_1R_2R_3\\&=R_1R_3-R_1R_3R_2+R_2R_3-R_2R_3R_1+R_1R_2R_3=R_1R_3-R_1R_3R_2+R_2R_3\\&=R_1R_3+R_2R_3-R_1R_3R_2\end{aligned} \tag{16.7.290}$$

And we can check that this approach is actually correct by taking the general relation of such a system proved above:

$$\begin{aligned}R_S&=\prod_{i=1}^{n}R_i=\prod_{i=1}^{n}\left(1-\prod_{j=1}^{m}(1-R_{i,j})\right)=\left(1-\prod_{j=1}^{2}(1-R_{1,j})\right)\left(1-\prod_{j=1}^{1}(1-R_{2,j})\right)\\&=\left(1-\prod_{j=1}^{2}(1-R_{1,j})\right)(1-(1-R_{2,j}))=\left(1-\prod_{j=1}^{2}(1-R_{1,j})\right)R_3\\&=(1-(1-R_1)(1-R_2))R_3=R_3-R_3(1-R_1)(1-R_2)=R_3-(R_3-R_3R_1)(1-R_2)\\&=R_3-(R_3-R_3R_1-R_3R_2+R_3R_1R_2)=R_3R_1+R_3R_2-R_3R_1R_2\end{aligned} \tag{16.7.291}$$

this shows that we have the same result and that the decomposition approach also works.

Finally, let us notice in conclusion that when we include in the systems elements that can tolerate or accept some errors, we then speak of "fault tolerance" and we distinguish mainly three types:

1. **Active redundancy**: In this case all redundant components operate simultaneously.

2. **Passive redundancy**: One redundant element is working, the others are waiting, which has for advantage to reduce or remove the aging of redundant elements, but in return requires the insertion of a component of failure detection and switching.

3. **Majority Redundancy**: This redundancy mainly concerns signal processing. The output signal will be that of the majority of redundant components.

16.7.3.4.1 Fault Tree Analysis

A "fault tree" is a widely used technique in engineering safety studies and reliability of systems consisting of represent the possible combinations of events that allow the realization of a predefined undesirable event. Such a representation therefore highlights the relationships of cause and effect:

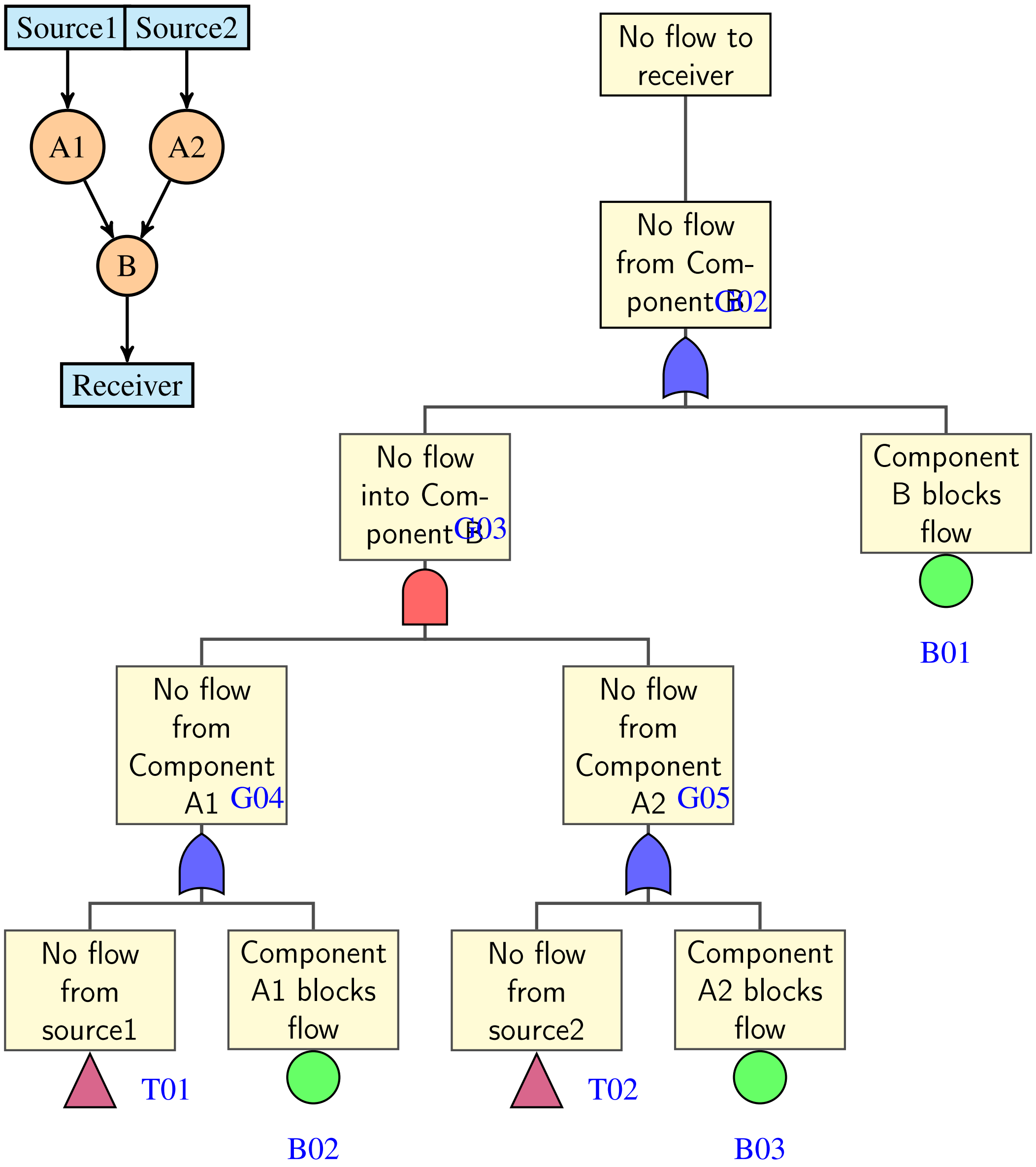

Source1
Source2
A1
A2
B
Receiver
No flow to receiver
No flow from Component B
G02
No flow into Component B
G03
Component B blocks flow
B01
No flow from Component A1
G04
No flow from Component A2
G05
No flow from source1
T01
Component A1 blocks flow
B02
No flow from source2
T02
Component A2 blocks flow
B03

Figure 16.226: **Example of simple fault tree (author: Zhang Long)**

This technique is complemented by mathematical processing that allows the combination of single failures and their probability of occurrence. It therefore allows to quantify the probability of occurrence of an adverse event.

The idea is very simple as long as we use only binary probabilities, by cons when it is necessary to use continuous distributions functions (the most common case in practice), we must go through Monte Carlo simulations (section Theoretical Computing) with typical software like Weibull++.

Consider the following fault tree that has 5 primitive events and 3 combined events and where the probabilities of the primitive events are assumed to be known:

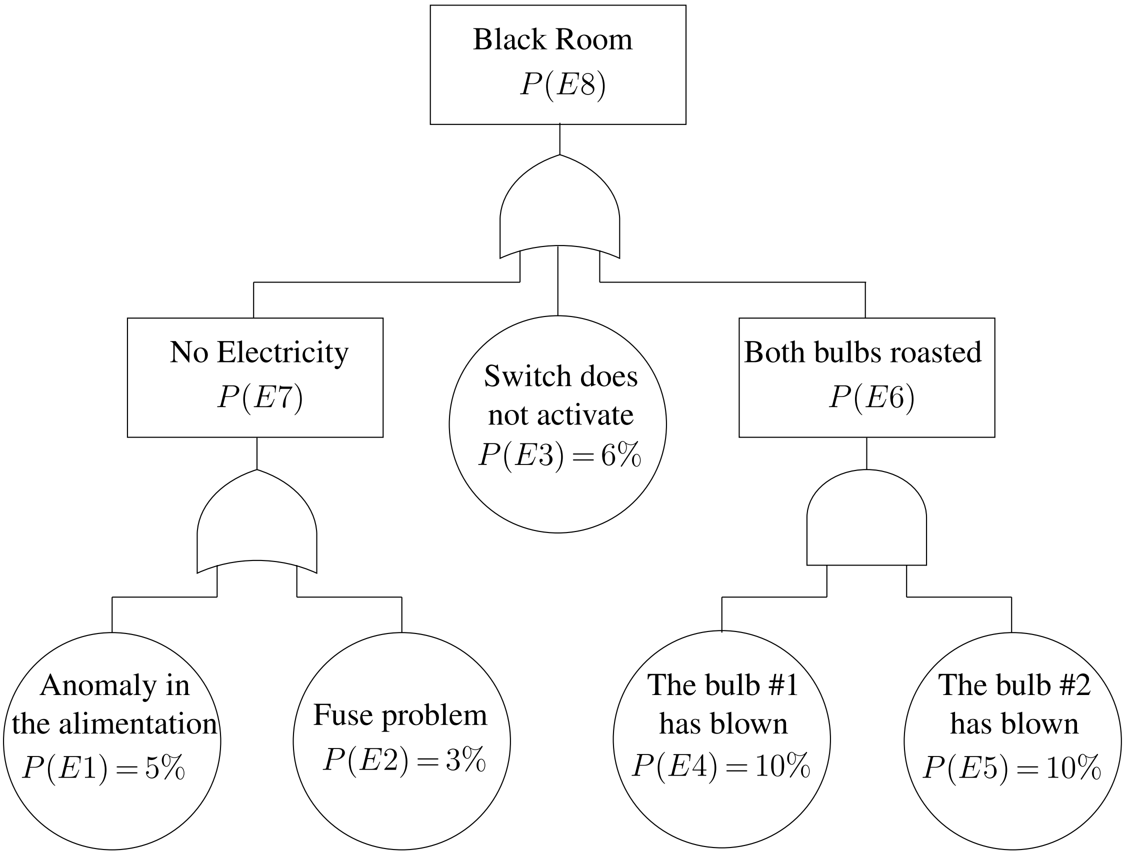

Black Room
$P(E8)$
No Electricity
$P(E7)$
Switch does not activate
$P(E3)=6\%$
Both bulbs roasted
$P(E6)$
Anomaly in the alimentation
$P(E1)=5\%$
Fuse problem
$P(E2)=3\%$
The bulb #1 has blown
$P(E4)=10\%$
The bulb #2 has blown
$P(E5)=10\%$

Figure 16.227: **Example of simple fault tree with binary probabilities and AND / OR gates only**

We then have for the next event $E6$ the following probability since it is an AND gate (equivalent to a parallel system):

$$P(E_6)=F_6=\prod_{i=1}^{5} F_i=P(E_4)P(E_5)=0.1\cdot 0.1=0.01=1\% \tag{16.7.292}$$

and for the event $E7$ since it is an OR gate (similar to a series system), we have:

$$\begin{aligned}P(E_7)=F_7=1-\prod_{i=1}^{2}(1-F_i)=1-(1-P(E_1))(1-P(E_2))=1-(1-0.05)\cdot(1-0.03)\\ =0.0785=7.85\%\end{aligned} \tag{16.7.293}$$

And so in the end:

$$\begin{aligned}P(E_8)=F_8=1-(1-F_3)(1-F_6)(1-F_7)=1-(1-P(E_3))(1-P(E_6))(1-P(E_7))\\ =1-(1-0.06)(1-0.0785)(1-0.01)=0.1424=14.24\%\end{aligned} \tag{16.7.294}$$

Thus, the probability of having the room without light is (a posteriori) of 14.24%.

16.7.3.4.2 Markov Chain Reliability Model

The notable strengths of Markov chain (see section Statistics page 537) models for reliability analysis is that they can account for repairs as well as failures. This makes the technique particularly useful for assessing the long-term average reliability of one or more devices with established maintenance and repair strategies. With the advent of high-integrity "fault-tolerant" systems, the ability to account for repairs of partially failed (but still operational) systems has become increasingly important. Markov modelling is well-suited to the task of determining inspection and repair intervals needed to achieve a desired level of safety.

For any given system, a "Markov chain reliability model" consists of a list of the possible states of that system, the possible transition paths between those states, and the rate parameters of those transitions.

Remark

We should speak rigorously about "semi-Markov chain reliability model" as mathematicians do the difference between a Markov chain where the jumps from one state to another are at constant probability an a semi-Markov chain where the jumps are random variables with any distribution, whose distribution function may depend on the two states between which the jump is made. For example semi-Markov process where all the holding times are exponentially distributed is named a "continuous time Markov chain/process".

In reliability analysis the transitions usually consist of failures and repairs. When representing a Markov model graphically, each state is usually depicted as a "bubble", with arrows denoting the transition paths between states, as depicted in the figure below for a single component that has just two states: healthy and failed.

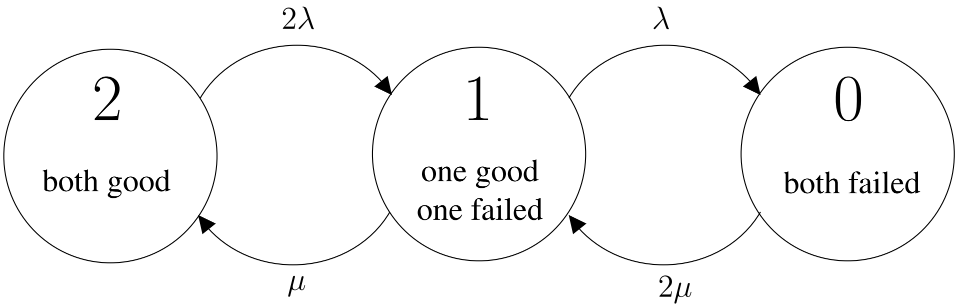


Figure 16.228: **Markov chain reliability model example**

This system is then formalized in the following form (not obvious to guess the very first time):

$$\begin{aligned}\frac{\mathrm{d}P_2(t)}{\mathrm{d}t}&=-2\lambda P_2(t)+\mu P_1(t)\\ \frac{\mathrm{d}P_1(t)}{\mathrm{d}t}&=2\lambda P_2(t)+2\mu P_0(t)-(\lambda+\mu)P_1(t)\end{aligned} \tag{16.7.295}$$

We can rewrite this system in matrix form as well:

$$\vec{P}\,'=M\vec{P} \tag{16.7.296}$$

With:

$$\vec{P}=(P_2,P_1,P_0)^T \qquad M=\begin{pmatrix}-2\lambda & \mu & 0\\ 2\lambda & -(\lambda+\mu) & 2\mu\\ 0 & \lambda & -2\mu\end{pmatrix} \tag{16.7.297}$$

A possible way to solve this system consists in a first time in diagonalizing (if possible) the matrix, that is to say to find an invertible matrix and a diagonal matrix such that:

$$M=SDS^{-1} \tag{16.7.298}$$

We thus fall back an application of the spectral theorem proved in the section of Linear Algebra (see page 2004) and applied to the resolution of differential equations (it is sometimes terrifying and admirable the extent of the striking power of pure mathematics...).

From here we want to be able to write:

$$\vec{P}\,'=SDS^{-1}\vec{P} \tag{16.7.299}$$

or also (it remains the same):

$$S^{-1}\vec{P}\,'=DS^{-1}\vec{P} \tag{16.7.300}$$

By doing the change of variable:

$$\vec{X}=S^{-1}\vec{P}\,' \tag{16.7.301}$$

the system to be solved becomes obviously:

$$\vec{X}\,'=DS^{-1}\vec{P} \tag{16.7.302}$$

Let us denote by γ_2,γ_1, γ_0 the diagonal elements of D and $\vec{X}=(X_2,X_1,X_0)^T$. We then want to rewrite the previous system as:

$$\begin{gathered}\frac{\mathrm{d}X_2(t)}{\mathrm{d}t}=\gamma_2X_2(t)\\ \frac{\mathrm{d}X_1(t)}{\mathrm{d}t}=\gamma_1X_1(t)\\ \frac{\mathrm{d}X_0(t)}{\mathrm{d}t}=\gamma_0X_0(t)\end{gathered} \tag{16.7.303}$$

But the solutions of this system are easy to determine, they are given by:

$$X_i(t)=c_i^{te}e^{\gamma_i t} \tag{16.7.304}$$

for $i=0,1,2$ where $c_i^{te}\in\mathbb{R}$.

Finally, we will determine the solutions of the system of departure with the help of the equality:

$$\vec{P}=S\vec{X} \tag{16.7.305}$$

Well... now let us put all this into practice!

To diagonalize M (if it is possible) we must determine its eigenvalues and its eigenspaces. Calculations made (see also the section of Linear Algebra for the general method at page 1998) we find that the values of M are:

$$0,-2(\lambda+\mu),-(\lambda+\mu) \tag{16.7.306}$$

The corresponding eigenspaces are given by the vectors (see chapter of Linear Algebra for the general method at page 1999):

$$\begin{pmatrix}\mu^2\\2\lambda\mu\\\lambda^2\end{pmatrix},\begin{pmatrix}1\\-2\\1\end{pmatrix},\begin{pmatrix}-\mu\\\mu-\lambda\\\lambda\end{pmatrix} \tag{16.7.307}$$

Hence, if we write:

$$D=\begin{pmatrix}0&0&0\\0&-2(\lambda+\mu)&0\\0&0&-(\lambda+\mu)\end{pmatrix}\qquad S=\begin{pmatrix}\mu^2&1&-\mu\\2\lambda\mu&-2&\mu-\lambda\\\lambda^2&1&\lambda\end{pmatrix} \tag{16.7.308}$$

We can then indeed write:

$$M=SDS^{-1} \tag{16.7.309}$$

The diagonal elements of D are:

$$\gamma_2=0\qquad\gamma_1=-2(\lambda+\mu)\qquad\gamma_0=-(\lambda+\mu) \tag{16.7.310}$$

The components of the vector $\vec{X}$ being given earlier above by:

$$X_i(t)=c_i^{te}e^{\gamma_i t} \tag{16.7.311}$$

we therefore have:

$$\vec{X}(t)=\begin{pmatrix}c_2^{te}\\c_1^{te}e^{-2(\lambda+\mu)t}\\c_0^{te}e^{-(\lambda+\mu)t}\end{pmatrix} \tag{16.7.312}$$

To finish the solutions of the original system are for recall given by:

$$\vec{P}=S\vec{X} \tag{16.7.313}$$

So finally:

$$\begin{aligned}P_2(t)&=c_2^{te}\mu^2+c_1^{te}e^{-2(\lambda+\mu)t}-\mu c_0^{te}e^{-(\lambda+\mu)t}\\P_1(t)&=2c_2^{te}\lambda\mu-2c_1^{te}e^{-2(\lambda+\mu)t}+(\mu-\lambda)c_0^{te}e^{-(\lambda+\mu)t}\\P_0(t)&=c_2^{te}\lambda^2+c_1^{te}e^{-2(\lambda+\mu)t}+\lambda c_0^{te}e^{-(\lambda+\mu)t}\end{aligned} \tag{16.7.314}$$

If we impose the initial condition:

$$P_2(0)=1\qquad P_1(0)=0\qquad P_0(0)=0 \tag{16.7.315}$$

we then have the system of three equations with three unknowns:

$$\begin{aligned}&c_2^{te}\mu^2+c_1^{te}-\mu c_0^{te}=1\\&2c_2^{te}\lambda\mu-2c_1^{te}+(\mu-\lambda)c_0^{te}=0\\&c_2^{te}\lambda^2+c_1^{te}+\lambda c_0^{te}=0\end{aligned}\tag{16.7.316}$$

Its resolution should be quite trivial for the reader (if it is not the case do not hesitate to contact us we will add the details) and we get:

$$\begin{aligned}&c_2=(\lambda+\mu)^{-2}\\&c_1=\lambda^2(\lambda+\mu)^{-2}\\&c_0=-2\lambda(\lambda+\mu)^{-2}\end{aligned}\tag{16.7.317}$$

So at the end:

$$\begin{aligned}&P_2(t)=\frac{\mu^2}{(\lambda+\mu)^2}+\frac{\lambda^2}{(\lambda+\mu)^2}e^{-2(\lambda+\mu)t}+\frac{2\lambda\mu}{(\lambda+\mu)^2}e^{-(\lambda+\mu)t}\\&P_1(t)=\frac{2\lambda\mu}{(\lambda+\mu)^2}-\frac{2\lambda^2}{(\lambda+\mu)^2}e^{-2(\lambda+\mu)t}+\frac{2\lambda(\lambda-\mu)}{(\lambda+\mu)^2}e^{-(\lambda+\mu)t}\\&P_0(t)=\frac{\lambda^2}{(\lambda+\mu)^2}+\frac{\lambda^2}{(\lambda+\mu)^2}e^{-2(\lambda+\mu)t}-\frac{2\lambda^2}{(\lambda+\mu)t}\end{aligned}\tag{16.7.318}$$

We can also verify that for any t we have:

$$P_0(t)+P_1(t)+P_2(t)=0\tag{16.7.319}$$

This is the reason why we often find the above system in the following form:

$$\begin{aligned}&P_2(t)=\frac{\mu^2}{(\lambda+\mu)^2}+\frac{\lambda^2}{(\lambda+\mu)^2}e^{-2(\lambda+\mu)t}+\frac{2\lambda\mu}{(\lambda+\mu)^2}e^{-(\lambda+\mu)t}\\&P_1(t)=\frac{2\lambda\mu}{(\lambda+\mu)^2}-\frac{2\lambda^2}{(\lambda+\mu)^2}e^{-2(\lambda+\mu)t}+\frac{2\lambda(\lambda-\mu)}{(\lambda+\mu)^2}e^{-(\lambda+\mu)t}\\&P_0(t)=1-P_2(t)-P_1(t)\end{aligned}\tag{16.7.320}$$

From the above formulas it follows that the availability on the long-term system is given by:

$$\lim_{t\rightarrow+\infty}P_2(t)+P_1(t)=\frac{\mu^2}{(\lambda+\mu)^2}+\frac{2\lambda\mu}{(\lambda+\mu)^2}\tag{16.7.321}$$

Because $\lambda+\mu>0$ and because we also have:

$$\lim_{t\rightarrow+\infty}e^{-(\lambda+\mu)t}=0\tag{16.7.322}$$

16.7.3.5 Maximum Likelihood for failure rate determination of samples

We studied in the section Statistics (see page 818) and Theoretical Computing (see page 6386) that the techniques using the maximum likelihood allowed us to estimate the maximum or minimum value of the parameters of a statistical law under the assumption that events are independent.

This technique, even empirical and sometimes quite questionable, is also used in the field of the reliability. As the theory has already been developed in the section Statistics let us see directly some practical cases.

✍**Examples:**

E1. Consider a system whose time intervals between failures is distributed according to an exponential law:

$$f(t)=\frac{\mathrm{d}F}{\mathrm{d}t}=\lambda e^{\lambda t} \tag{16.7.323}$$

of unknown parameter λ and where we consider (as always ...) that every failure is independent.

Observations gave us the following number of months between each failure: $10, 5, 11, 12, 15$.

Then using maximum likelihood which idea is for reminder to maximize (minimize) the total probability of independent events (therefore we must multiply the probabilities), we have:

$$L_5(\lambda)=\lambda e^{-10\lambda}\lambda e^{-5\lambda}\lambda e^{-11\lambda}\lambda e^{-12\lambda}\lambda e^{-15\lambda}=\lambda^5 e^{-53\lambda} \tag{16.7.324}$$

Therefore determine the failure rate that maximizes the likelihood is equivalent as we know to calculate when the derivative (see section Differential and Integral Calculus page 1450) is zero:

$$\frac{\mathrm{d}L_5(\lambda)}{\mathrm{d}\lambda}=\frac{\mathrm{d}\lambda^5 e^{-53\lambda}}{\mathrm{d}\lambda}=(5\lambda^4-53\lambda^5)e^{-53\lambda}=\lambda^4(5-53\lambda)e^{-53\lambda} \tag{16.7.325}$$

We then have two roots that are trivial:

$$\lambda_1=0 \qquad \lambda_2=\frac{5}{53} \tag{16.7.326}$$

and of course we will keep the second as root failure rate of our system.

E2. We consider now see the same data as before but this time with a Weibull distribution with the following simplified notation:

$$f(x)=\frac{\alpha}{\beta^\alpha}x^{\alpha-1}e^{-\left(\frac{x}{\beta}\right)^\alpha}=\alpha\lambda x^{\alpha-1}e^{-\lambda x^\alpha} \tag{16.7.327}$$

and therefore:

$$\begin{aligned}&L_5(\alpha,\lambda)=f(10)f(5)f(11)f(12)f(15)\\&=\left[\alpha\lambda 10^{\alpha-1}e^{-\lambda 10^{\alpha}}\right]\left[\alpha\lambda 5^{\alpha-1}e^{-\lambda 5^{\alpha}}\right]\left[\alpha\lambda 11^{\alpha-1}e^{-\lambda 11^{\alpha}}\right]\left[\alpha\lambda 12^{\alpha-1}e^{-\lambda 12^{\alpha}}\right]\left[\alpha\lambda 15^{\alpha-1}e^{-\lambda 15^{\alpha}}\right]\end{aligned}\tag{16.7.328}$$

To maximize this kind of function, as far as we know, we can use only numerical methods. By cons, a simple spreadsheet software like Microsoft® Excel® with its solver allows to find the two parameters of the Weibull distribution in three clicks!

E3. We imagine that we put under observation 10 similar manufactured elements. On the 10, we observed a stop of 4 of them after respectively $700, 800, 900$ and 950 hours. The remaining 6 items will be considered as being over the $1,000$ work hours and we stop observing from this value. Assuming that the failure distribution is exponentially again, then we have (the idea is subtle but simple):

$$\begin{aligned}L_10(\lambda)&=f(700)f(800)f(900)f(950)\\&=\lambda e^{-700\lambda}\lambda e^{-800\lambda}\lambda e^{-900\lambda}\lambda e^{-950\lambda}\left(1-\left(1-e^{-1000\lambda}\right)\right)^6\\&=\lambda^4e^{-3350\lambda}e^{-6000\lambda}=\lambda^4e^{-9350\lambda}\end{aligned}\tag{16.7.329}$$

It comes then:

$$\dfrac{\mathrm{d}L_10(\lambda)}{\mathrm{d}\lambda}=\dfrac{\mathrm{d}\lambda^4e^{-9350\lambda}}{\mathrm{d}\lambda}=(4\lambda^3-9350\lambda^4)e^{-9350\lambda}=\lambda^3(4-9350)e^{-9350\lambda}=0\tag{16.7.330}$$

We then have two roots that are trivial:

$$\lambda_1=0\qquad\lambda_4=\dfrac{4}{9,350}\tag{16.7.331}$$

and of course we will keep the second root as failure rate of our system.

16.7.3.6 Kaplan-Meier Survival Rate

In areas such as high-level industry, high-level medicine or high-level biology, we are often interested in the:

1. Survival duration after a serious event
2. Duration of remission after treatment or surgery
3. Duration of symptoms after an abnormality
4. Duration of a symptomless infection

We seek very often to distinguish at least "the event of interest":

1. System shutdown after serious event
2. End of remission
3. End of a symptom after anomaly
4. Beginning of a symptom during an infection

of the variable to explain "duration before the event of interest":

1. Elapsed duration before shutdown
2. Elapsed duration before the end of remission
3. Elapsed duration before the end of the symptom
4. Elapsed duration without symptoms

Definitions (#16.7.735):

D1. We name "remission", the reduction of a disease or a temporarily malfunction.

D2. The "survival time" or "lifetime" T means the time which elapses from an initial time (start of treatment, diagnosis or failure, etc.) until the occurrence of a final event of interest (patient death, relapse, remission, cure, repair, etc.). We say that the object of the study "survives at time t" if at this moment the final interest event has not yet occurred.

Remark

Although this type of study can be associated with preventive maintenance (see page 8716), statisticians associate this type of study rather to the domain of "survival analysis".

We will focus in this textbook on a particular context but that can be easily generalized:

- Cohort/Clinical trial where we study the survival time of each patient (machine).
- All patients (machines) do not have the same observation time (different instants of entry into the study).
- We have information on the survival time of each patient (machine) but we do not know exactly when it happens.

From the last two points, we conclude that the survival time can be censored. So the usual statistical techniques does not apply directly as censored data require special treatment (of course, if we remove the censored data we lose information). It goes without saying that this situation is extremely common and therefore the study of the Kaplan-Meier estimator is very important (for other survival models the reader should refer to the Statistics section especially the part concerning Survival Models at page 1271).

Definition (#16.7.736): The duration T is said to be censored if the duration is not observed completely. The different types of censoring are:

- Type I censoring: fixed right. In this situation, the time is not observable beyond a maximum, fixed, named "fixed-censoring" and imposed. So either we have the opportunity to observe the real duration of the event of interest for the item if it occurs before the fixed-censoring, or we limit ourselves to the fixed-censoring time if the event of interest has not occurred before.
- Type II censoring: wait. In this situation, we observe the lives of n individuals until $m \leq n$ individuals have seen the event to occur (deceased). The duration considered is therefore that of the beginning of the experiment until the event of interest for the m-th.
- Type III censoring: random left. In this situation, we do not know when the event of interest has occurred (because we started to observe the subject of study too late). We can not then deal with "durations" in the measurable sense and we have to limit ourselves to a simple count.
- Type III censoring: random right. In this situation, we do not know when the event will take place (because we stopped to observe the subject before it occurs for any reason). We can not then also not deal with "times" in the measurable sense and we must limit ourselves to a simple count.
- Type IV censoring: random intervals. In this situation, we have a mixture of random left and right censoring. That is to say that for some study subjects, we do not know when the event of interest has begun, and for others we do not know when it will occur (if any. ...).

In the machinery industry, we often deal with the type I censoring: fixed right. In the medical field, in clinical trials, we often deal with a type III censoring: random right. In the case of pandemics, we are dealing with type III censoring: random left.

To introduce this subject, rather than doing obscure theory, as always in this textbook we prefer at pragmatic approach. Suppose that the study is a clinical trial involving two groups of patients receiving two types of treatments. Two important questions raised to the physicians in a Phase II clinical trial (phase I is for non-toxicity approval to human and phase 0 for animals):

1. Is one of the two treatments more effective than the other in terms of improving patient survival?
2. Can we highlight prognostic factors, that is to say that improve / deteriorate survival?

To answer the first question we can develop statistical methods that will allow us to compare the two groups of patients who receive both types of treatment.

To answer the second question we propose a model that links patient survival time for explanatory variables and highlight the prognostic factors.

Let us as always assist the theory with an example. For this consider the following table where two cohorts of patients (we move from mechanical engineering to human engineering...) of same initial size with acute leukemia drug test (6-MP) against a placebo[22] (of course blindly).

We have the following remission times for 21 patients (the table of 21 lines therefore indicates the number of weeks for where patient is considered cured after treatment before falling again ill):

6-mercaptopurine (6-MP) Group	Placebo Group
6	1
6	1
6	2
6+	2
7	3
9+	4
10	4
10+	5
11+	5
13	8
16	8
17+	8
19+	8
20+	11
22	11
23	12
25	12
32+	15
32+	17
34+	22
35+	23

Table 16.23: **Two cohort survival analysis with right censoring**

The "+" sign corresponds to patients who left the study for that week. They are therefore censored! For example, the fourth patient was lost of view for any reason after 6 weeks of treatment with 6-MP: it has therefore a duration of remission greater than 6 weeks! So in the study 6-MP, there are 21 patients and 11 of them are censored data.

---

[22] In the science of drugs, a "pure placebo" is a treatment without any active substance; an "impure placebo" is a pharmacologically active product but has no effect on the pathology treated, or whose efficacy has not been sufficiently demonstrated.

Remark

The theoretical model assumes that censoring is independent of the survival time (not informative censoring). But if censoring is due to the discontinuation of treatment, the independence assumption is not valid anymore!

For the placebo group it is simple to make a survival curve. It is sufficient to produce the following table (for the omitted weeks, obviously we impose the number of remission as constants):

Week i	Total # remission at week i	Proportion (probability) of remission at week i
0	21	100%
1	19	$19/21=90\%$
2	17	$17/21=81\%$
3	16	$16/21=76\%$
4	14	$14/21=67\%$
5	12	$12/21=57\%$
8	8	$8/21=38\%$
11	6	$6/21=29\%$
12	4	$4/21=19\%$
15	3	$3/21=14\%$
17	2	$2/21=10\%$
22	1	$1/20=0.05\%$
23	0	0%

Table 16.24: **Two cohort survival analysis with right censoring**

So if the data are not censored, the survival $S(t)$ can be estimated by the proportion of individuals surviving at time t, that is customary to write under the following mathematical form:

$$\hat{S}_k(t)=P_k(T_k>t)=\frac{1}{n}\sum_{i=1}^{k}1_{T_i>t} \tag{16.7.332}$$

The idea is therefore to estimate:

$$P(T>t)=P(\text{probability to survive until time } t) \tag{16.7.333}$$

by the proporbo of patients who survived until time t.

If the data are censored, the estimated survival function requires specific tools. Kaplan and Meier have proposed in this special case the following calculation:

$$\begin{aligned}&P(\text{be in remission a the } i\text{-th week})\\&=P(\text{to not relapse at the } i\text{-th week})\cdot P(\text{be in remission a the } i-1\text{-th week})\end{aligned} \tag{16.7.334}$$

Let's see it in a slightly more mathematical form. If we denote by $X(1)\leq X(2)\leq\ldots\leq X(n)$ the moments (sorted) where an event occurred (death or censored), then we have:

$$P(T>X(k))=P(T>X(k)|T>X(k-1))P(T>X(k-1)) \tag{16.7.335}$$

With of course:

$$P(T>0)=1 \tag{16.7.336}$$

We estimate:

$$\hat{P}_k=\left(1-\frac{d_k}{R_k}\right) \tag{16.7.337}$$

where d_k is the number of deaths (failures) observed in the time corresponding to the event $X(k)$ and R_k is the number of individuals at risk (at risk of death/failure) just before $X(k)$.

We define the "Kaplan-Meier estimator" for any $X(0)\leq t<X(k)$ by:

$$\boxed{\hat{S}_{\text{KM}}(t)=\prod_{k=1}^{n}\hat{P}_k=\prod_{k=1}^{n}\left(1-\frac{d_k}{R_k}\right)=\prod_{k=1}^{n}\left(1-\hat{\lambda}_k\right)} \tag{16.7.338}$$

where $\hat{\lambda}_k$ is the discrete hazard (i.e. conditional probability of death).

Therefore we get doing now for 6-MP group (not the placebo group!!!!!) the following table:

Remission duration (uncensored) observed	**Items in remission at the beginning of k**	**Probability of not relapse at k knowing that we are at at $k-1$ ($\hat{P}_k$)**	**Survival probability of Kaplan-Meier**
0	21	$21/21=100\%$	100%
6	21	$18/21=1-3/21=85.7\%$	$100\%\cdot 85.7\%=85.7\%$
7	17	$16/17=1-1/17=94.1\%$	$85.7\%\cdot 94.1\%=80.7\%$
10	15	$14/15=1-1/15=93.3\%$	$80.7\%\cdot 93.3\%=75.3\%$
13	12	$11/12=1-1/12=91.7\%$	$75.3\%\cdot 91.7\%=69\%$
16	11	$10/11=1-1/11=90.9\%$	$69\%\cdot 90.9\%=62.7\%$
22	7	$6/7=1-1/7=85.7\%$	$62.7\%\cdot 85.7\%=53.8\%$
23	6	$5/6=1-1/6=83.3\%$	$53.8\%\cdot 83.3\%=44.8\%$

Table 16.25: **6-MP survival analysis**

With the corresponding survival curves (including the confidence interval):

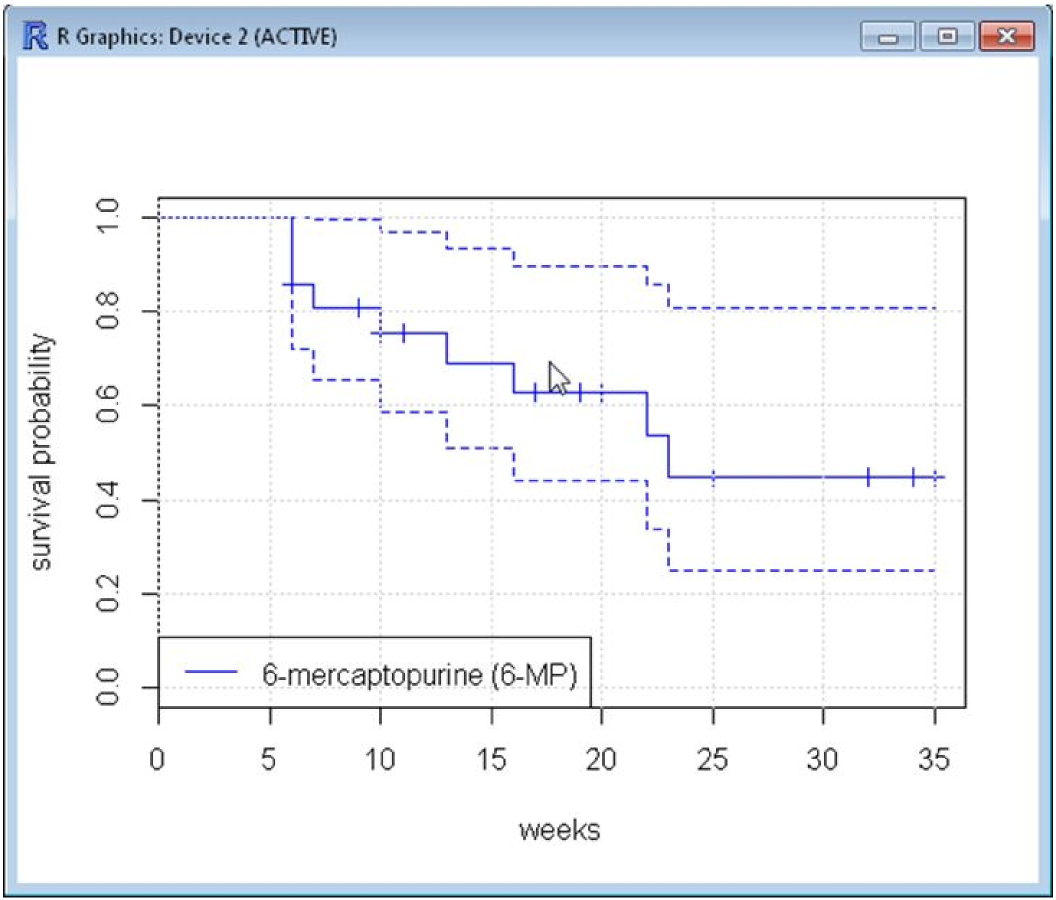


Figure 16.229: **Kaplan-Meier typical survival curves in `R` software**

We thus find the same values as those given for example by Minitab Statistical Software 15.1.1 and `R` (see the companion books on Minitab or `R` for the details).

If we assume that the $\hat{\lambda}_k$'s are just binomial proportions given the R_k's we have the following normalized variance and mean of binomial random variable (see page 663):

$$\mathrm{E}(\hat{\lambda}_k)=\hat{\lambda}_k \qquad \mathrm{V}(\hat{\lambda}_k)=\frac{\hat{\lambda}_k(1-\hat{\lambda}_k)}{R_k} \tag{16.7.339}$$

Remark

Sadly it is a tradition in survival analysis to denote the normalized variance of $\hat{\lambda}_k$ by $\mathrm{V}(\hat{\lambda}_k)$ instead of $\mathrm{V}(\hat{\lambda}_k/r_k)$...

and we will assume also that the $\hat{\lambda}_k$'s are asymptotically independent (...).

Since $\hat{S}_{\text{KM}}(t)$ is a function of the $\hat{\lambda}_k$'s, we can estimate its variance using the Delta method (see section Numerical Methods page 6504) that has for important result for recall that:

$$g(X_n) \underset{n\to+\infty}{\cong} \mathcal{N}\left(0,g'(\mu)^2\sigma^2\right) \tag{16.7.340}$$

with:

$$\mathrm{V}(g(X_n))\cong g'(\mu)^2\sigma^2 \tag{16.7.341}$$

Therefore instead of dealing with $\hat{S}_{\text{KM}}(t)$ directly, we will look at its natural logarithm:

$$\ln(\hat{S}_{\text{KM}}(t))=\sum_{k=1}^{n}\ln(1-\hat{\lambda}_k) \tag{16.7.342}$$

Thus, by the approximate independence of the $\hat{\lambda}_k$'s and keeping in mind that we just chose $g()$ as the natural logarithm and using the Delta method (see section Numerical Methods page 6504):

$$\begin{aligned}\hat{\text{V}}\left[\ln(\hat{S}_{\text{KM}}(t))\right]&\cong\sum_{k=1}^{n}\hat{\text{V}}\left[\ln(1-\hat{\lambda}_k)\right]\underset{\text{delta method}}{\cong}\sum_{k=1}^{n}\left(\frac{1}{1-\hat{\lambda}_k}\right)^2\hat{\text{V}}(\hat{\lambda}_k)\\&=\sum_{k=1}^{n}\left(\frac{1}{1-\hat{\lambda}_k}\right)^2\frac{\hat{\lambda}_k(1-\hat{\lambda}_k)}{R_k}=\sum_{k=1}^{n}\frac{\hat{\lambda}_k}{(1-\hat{\lambda}_k)R_k}=\sum_{k=1}^{n}\frac{d_k}{(R_k-d_k)R_k}\end{aligned} \tag{16.7.343}$$

And applying on it the Delta method (see section Numerical Methods page 6504):

$$\text{V}(g(X_n))\cong g'(\mu)^2\sigma^2=g'(\mu)^2\text{V}(X_n)\Rightarrow\text{V}\left(\ln(\hat{S}_{\text{KM}}(t)\right)\cong\frac{1}{\hat{S}^2_{\text{KM}}(t)}\hat{\text{V}}\left(\hat{S}_{\text{KM}}(t)\right) \tag{16.7.344}$$

then rearranging and injecting the previous result we get:

$$\hat{\text{V}}\left(\hat{S}_{\text{KM}}(t)\right)\cong\hat{S}^2_{\text{KM}}(t)\text{V}\left(\ln(\hat{S}_{\text{KM}}(t)\right)=\hat{S}^2_{\text{KM}}(t)\sum_{k=1}^{n}\frac{d_k}{(R_k-d_k)R_k} \tag{16.7.345}$$

The following approximation:

$$\boxed{\hat{\text{V}}\left(\hat{S}_{\text{KM}}(t)\right)\cong\hat{S}^2_{\text{KM}}(t)\sum_{k=1}^{n}\frac{d_k}{(R_k-d_k)R_k}} \tag{16.7.346}$$

is named the "Greenwood's relation". If we take the square root then the result is associated to the standard error (because of the division by R_k):

$$\text{SE}\left(\hat{S}_{\text{KM}}(t)\right)\cong\hat{S}_{\text{KM}}(t)\sqrt{\sum_{k=1}^{n}\frac{d_k}{(R_k-d_k)R_k}}=\hat{S}_{\text{KM}}(t)\frac{\sqrt{\sum_{k=1}^{n}\frac{d_k}{(R_k-d_k)}}}{\sqrt{R_k}}=\hat{S}_{\text{KM}}(t)\frac{\sigma_{\hat{S}_{\text{KM}}(t)}}{\sqrt{R_k}} \tag{16.7.347}$$

In order to construct an error bar for $\hat{S}_{\text{KM}}(t)$ we need to make some sort of distributional assumption. The simplest is to assume that $\hat{S}_{\text{KM}}(t)$ is normally distributed such that we can write:

$$\hat{S}_{\text{KM}}(t)\pm Z_{\alpha/2}\text{SE}\left(\hat{S}_{\text{KM}}(t)\right) \tag{16.7.348}$$

Or more explicitly:

$$\boxed{\hat{S}_{\text{KM}}(t)-Z_{\alpha/2}\hat{S}_{\text{KM}}(t)\sqrt{\sum_{k=1}^{n}\frac{d_k}{(R_k-d_k)R_k}}\leq S_{\text{KM}}(t)\leq\hat{S}_{\text{KM}}(t)+Z_{\alpha/2}\hat{S}_{\text{KM}}(t)\sqrt{\sum_{k=1}^{n}\frac{d_k}{(R_k-d_k)R_k}}} \tag{16.7.349}$$

A snag with this of course is that the distribution of $\hat{S}_{\text{KM}}(t)$ is not really normal. Problems arise when $\hat{S}_{\text{KM}}(t)$ is close to 0 or 1 as indeed the confidence interval can yield values to >1 or <0. A common empirical approach (among others!) is then to take for the confidence interval the complementary log-log transformation:

$$\hat{W}(t)=\ln\left(-\ln\left(\hat{S}_{\text{KM}}(t)\right)\right) \tag{16.7.350}$$

Since this quantity is unrestricted, the confidence interval will be in the right range when we transform back:

$$\hat{S}_{\text{KM}}(t)=e^{-e^{\hat{W}(t)}} \tag{16.7.351}$$

To see why this works:

$$0\leq \hat{S}_{\text{KM}}(t)\leq 1 \tag{16.7.352}$$

$$-\infty\leq \ln\left(\hat{S}_{\text{KM}}(t)\right)\leq 0 \tag{16.7.353}$$

$$0\leq -\ln\left(\hat{S}_{\text{KM}}(t)\right)\leq +\infty \tag{16.7.354}$$

$$-\infty\leq \ln\left(-\ln\left(\hat{S}_{\text{KM}}(t)\right)\right)\leq +\infty \tag{16.7.355}$$

To get a confidence interval we need now also to find the standard error corresponding to this log-log transformation, and for this purpose we need to calculate:

$$\hat{\text{V}}(\hat{W}(t))=\hat{\text{V}}\left(\ln\left(-\ln\left(\hat{S}(t)\right)\right)\right) \tag{16.7.356}$$

From our previous calculations, we know:

$$\hat{\text{V}}\left[\ln(\hat{S}_{\text{KM}}(t))\right]\cong\sum_{k=1}^{n}\frac{d_k}{(R_k-d_k)R_k} \tag{16.7.357}$$

Applying the Delta method again, we get:

$$\hat{\text{V}}(\hat{W}(t))=\hat{\text{V}}\left(\ln\left(-\ln\left(\hat{S}(t)\right)\right)\right)\cong\frac{1}{\left[\ln\left(\hat{S}(t)\right)\right]^2}\sum_{k=1}^{n}\frac{d_k}{(R_k-d_k)R_k} \tag{16.7.358}$$

We take the square root of the above to get the standard error:

$$\text{SE}(\hat{W}(t))\cong\frac{1}{\ln\left(\hat{S}(t)\right)}\sqrt{\sum_{k=1}^{n}\frac{d_k}{(R_k-d_k)R_k}} \tag{16.7.359}$$

And therefore we have :

$$\hat{W}(t)-Z_{\alpha/2}\frac{1}{\ln\left(\hat{S}(t)\right)}\sqrt{\sum_{k=1}^{n}\frac{d_k}{(R_k-d_k)R_k}}\leq W(t)\leq\hat{W}(t)+Z_{\alpha/2}\frac{1}{\ln\left(\hat{S}(t)\right)}\sqrt{\sum_{k=1}^{n}\frac{d_k}{(R_k-d_k)R_k}} \tag{16.7.360}$$

Since $\hat{S}_{\text{KM}}(t)=e^{-e^{\hat{W}(t)}}$, the ("modern") bounds for the confidence interval of $\hat{S}_{\text{KM}}(t)$ are finally (note that the upper and lower bound switch!):

$$\exp\left\{-e^{\hat{W}(t)+Z_{\alpha/2}\frac{1}{\ln(\hat{S}(t))}\sqrt{\sum_{k=1}^{n}\dfrac{d_k}{(R_k-d_k)R_k}}}\right\}\leq S_{\text{KM}}(t)\leq\exp\left\{-e^{\hat{W}(t)-Z_{\alpha/2}\frac{1}{\ln(\hat{S}(t))}\sqrt{\sum_{k=1}^{n}\dfrac{d_k}{(R_k-d_k)R_k}}}\right\}\tag{16.7.361}$$

Substituting $\hat{W}(t)=\ln\left(-\left(\hat{S}_{\text{KM}}(t)\right)\right)$ back into the above bounds, we get the definitive "modern" relation for the Kaplan-Meier estimator confidence interval:

$$\boxed{\hat{S}_{\text{KM}}(t)\exp\left\{e^{Z_{\alpha/2}\frac{1}{\ln(\hat{S}(t))}\sqrt{\sum_{k=1}^{n}\dfrac{d_k}{(R_k-d_k)R_k}}}\right\}\leq S_{\text{KM}}(t)\leq\hat{S}_{\text{KM}}(t)\exp\left\{e^{-Z_{\alpha/2}\frac{1}{\ln(\hat{S}(t))}\sqrt{\sum_{k=1}^{n}\dfrac{d_k}{(R_k-d_k)R_k}}}\right\}}\tag{16.7.362}$$

16.7.3.7 ABC Method

In a company, the tasks are various and the maintenance teams are systematically reduced to a minimum. In addition, the most advanced technologies in maintenance can be very expensive, and should not be applied indiscriminately.

It is therefore appropriate to be organized effectively and efficiently. The "ABC analysis[23]", also named "ABC method" (one should say rigorously "ABC objective"), using implicitly the cumulated Pareto law (see section Statistics page 746), allows remedy this relatively well with an empirical approach. Thus, a classification of costs in relation to the types of failure gives priority interventions to carry on (this empirical method is also used in many other areas including one that is very well known: supply chain management[24]).

The idea is at first like Pareto analysis (see section Quantitative Management page 8095 to classify faults in order of increasing maintenance costs (or cost impact in case of failure) every failure relating a simple or complex system and to establish a chart matching the percentage of cumulative cost to the cumulative percentages type of failure.

Then we distinguish traditionally in the industry three zones:

1. Zone A: In most cases, about 20% of total failures represent 80% of costs and so this is the priority area.
2. Zone B: In this zone, representing 30% of total failures we have the following 15% additional costs.
3. Zone C: In this zone 50% of the remaining failures generates only 5% of costs.

The field of participatory web (web 2.0) considers areas (targets) of respectively $90-9-1$ percents so we found very everything and anything in this field as it is empirical...

Let us see an example assuming that the following data were collected and we would like to do an analysis of the percentage of machines that should be that we focus on to reduce the cost off failures hours of approximately 80% (in reality we will focus more on the percentage of the financial cost!).

---

[23] Not to be confused with Activity-Based Costing (ABC) that is a costing methodology that identifies activities in an organization and assigns the cost of each activity with resources to all products and services according to the actual consumption by each.

[24] This is why the supply chain module of SAP includes an ABC and XYZ analysis

Machine Number	Nonworking hours	Number of failures
1	100	4
2	32	15
3	50	4
4	19	14
5	4	3
6	30	8
7	40	12
8	80	2
9	55	3
10	150	5
11	160	4
12	5	3
13	10	8
14	20	8

Table 16.26: ABC **analysis on machines failures**

Then, in a spreadsheet like Microsoft® Excel® or other ... we can easily establish the following table:

Machine	Stop [h.]	Sum [h.]	% Cumulated	Failures	Sum	% Cumulated
11	160	160	21.19%	4	4	4.30%
10	150	310	41.06%	5	9	9.68%
1	100	410	54.30%	4	13	13.98%
8	80	490	64.90%	2	15	16.13%
9	55	545	72.19%	3	18	19.35%
3	50	595	78.81%	4	22	23.66%
7	40	635	84.11%	12	34	36.56%
2	32	667	88.34%	15	49	52.69%
6	30	697	92.32%	8	57	61.29%
14	20	717	94.97%	8	65	69.89%
4	19	736	97.48%	14	79	84.95%
13	10	746	98.81%	8	87	93.55%
12	5	751	99.47%	3	90	96.77%
5	4	755	100.00%	3	93	100.00%

Table 16.27: **Data normalization for** ABC **analysis on machines failures**

We then have graphically:

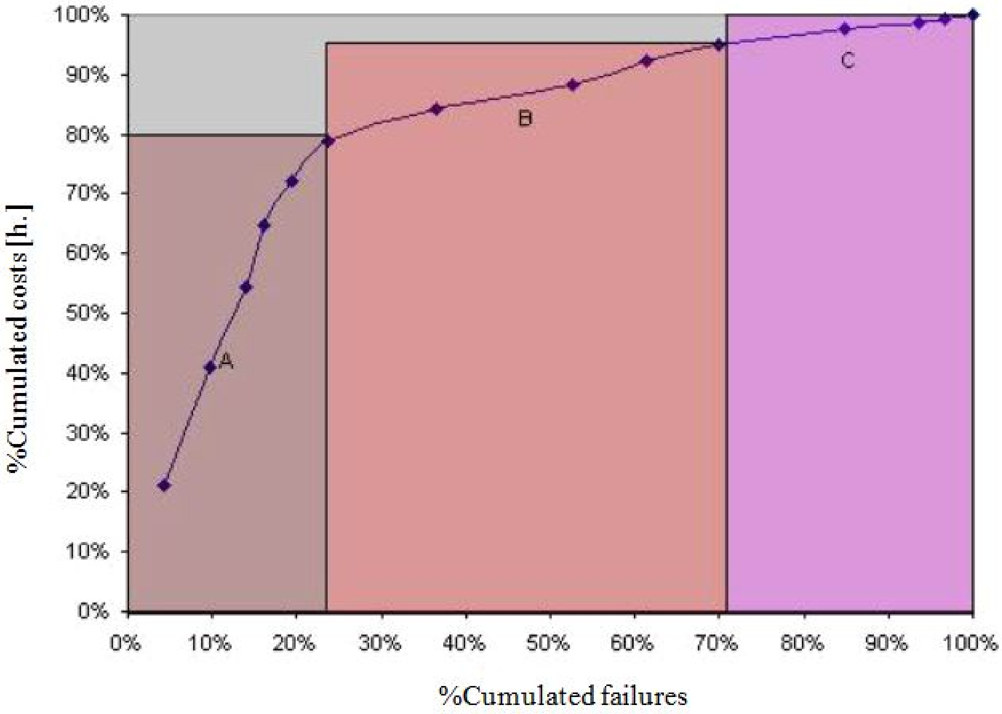


Figure 16.230: Graph of the ABC method

where the areas A, B and C are rounded to existing points. Thus, the critical area A contains the machines 11, 10, 1, 8, 9 and 3. improving the reliability of these machines can therefore get up to 78.8% time saving on downtime.

Now let us determine the parameters of the Pareto distribution law (see section Statistics page 746):

$$P(X \leq x) = P_i = 1 - \left(\frac{x_m}{x_i}\right)^k \qquad (16.7.363)$$

We have to determine k and x_m and the other parameters are given us by our measurements (the table).

We can play in the following way:

$$P_i - P_{i-1} = -\left(\frac{x_m}{x_i}\right)^k + \left(\frac{x_m}{x_{i-1}}\right)^k = x_m^k \left(\frac{1}{x_{i-1}} - \frac{1}{x_i}\right) \qquad (16.7.364)$$

Therefore:

$$\log\left(P_i - P_{i-1}\right) = \log\left(x_m^k \left(\frac{1}{x_{i-1}} - \frac{1}{x_i}\right)^k\right) = k\log(x_m) + k\log\left(\frac{1}{x_{i-1}} - \frac{1}{x_i}\right) \qquad (16.7.365)$$

So thanks to:

$$\log(\Delta P) = k\log\left(\frac{1}{x_{i-1}} - \frac{1}{x_i}\right) + k\log(x_m) \qquad (16.7.366)$$

we should be able to determine the two seeked parameters by considering the expression as the equation of a linear function whose k is the slope and $k\log(x_m)$ the intercept:

$$\log(\Delta P) = ax + b \tag{16.7.367}$$

A simple linear regression (see section Numerical Methods page 6322) gives us:

$$k = 0.6189 \quad \text{and} \quad k\log x_m = -1.2234 \tag{16.7.368}$$

Therefore:

$$x_m = 10\frac{-1.2234}{k} = 0.0105 \tag{16.7.369}$$

So we have finally:

$$P(X \leq x) = P_i = 1 - \left(\frac{x_m}{x_i}\right)^k = 1 - \left(\frac{0.0105}{x_i}\right)^{0.6189} \tag{16.7.370}$$

What then gives the following table (the x_i of the Pareto distribution are the Cumulative% failure):

Machine	% Cumulated cost [h.]	% Cumulated failures	% Cumulated Pareto Law
11	21.19%	4.30%	61.76%
10	41.06%	9.68%	78.01%
1	54.30%	13.98%	82.89%
8	64.90%	16.13%	84.48%
9	72.19%	19.35%	86.29%
3	78.81%	23.66%	88.05%
7	84.11%	36.56%	91.12%
2	88.34%	52.69%	93.08%
6	92.32%	61.29%	93.75%
14	94.97%	69.89%	94.29%
4	97.48%	84.95%	95%
13	98.81%	93.55%	95.32%
12	99.47%	96.77%	95.43%
5	100.00%	10.00%	95.43%

Table 16.28: **Comparisons experimental/ theoretical data**

Then we can get the real corresponding Pareto curve easily in Microsoft® Excel® 11.8346:

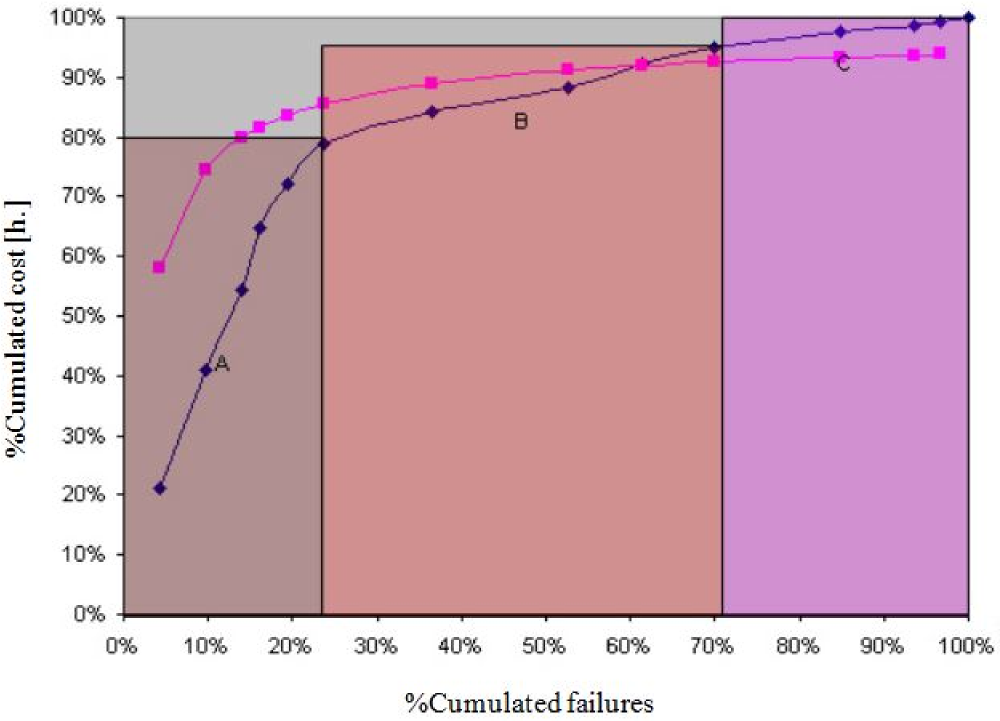


Figure 16.231: ABC **method associated with a Pareto curve**

The difference in the plot above between the experimental curve and the theoretical one is quite significant... As k is less than 1, then as we have proved it the section of Statistics (see page 746), the Pareto law has neither expected mean nor variance.

The practitioner must be careful to the fact that in the field of management, the Pareto law is used wrongly for a little bit everything when an another probability distribution could be much more appropriate.

Furthermore, any GoF test (see section Statistics page 1111) shows us that we must reject the approximation by a Pareto Distribution. Moreover, specialized software such @RISK reject the Pareto approximation beyond the 20 best adjustments, the best fit with this same software is the Log-Normal law.

Remark

Most statistical softwares and textbooks define "Pareto analysis" (i.e. ABC analysis) based on only one categorical variable (machine number, product named, etc.) and only one numerical variable (number sales, number defects, amount of cash-flow, etc.), where the categorical variable proportion is used to build the Pareto percentage categories on the x-axis and the numerical variable for the y-axis, at the opposite of the previous example that may be categorized into a "bivariate Pareto analysis" as it uses two numerical variables. For "univariate Pareto analysis" that are very obvious, the reader may refer to the `R` software companion book to get some examples.

16.7.3.8 XYZ Method

The "XYZ analysis", also named "XYZ method", is a way to classify inventory items according to variability of their demand (in the last centuries we should rather say in according to their "difficulty to forecast"):

- Zone X (very little variation): X items are characterised by steady turnover over time. Future demand can be reliably forecast.
- Zone Y (some variation): Although demand for Y items is not steady, variability in demand can be predicted to an extent. This is usually because demand fluctuations are caused by known factors, such as seasonality, product lifecycles, competitor action or economic factors. It's more difficult to forecast demand accurately.
- Zone Z (the most variation): Demand for Z items can fluctuate strongly or occur sporadically. There is no trend or predictable causal factors, making reliable demand forecasting impossible.

Previous centuries textbooks have supported the use of coefficient of variation CV (see section Statistics page 588):

$$\text{CV}=\frac{\sigma_{\hat{\theta}}}{\hat{\theta}} \qquad (16.7.371)$$

applied to every product time series (so if we have n times series we will have n coefficient of variation CV_i) pattern and afterwards running a classic step by step approach similar to that in every point to the ABC method but on the CV_i (in order to classify each time series in one of the three "zone").

This was quite a good idea at the time computer were not existing or when manager had almost no competences in statistics. But... in this beginning of the 121st century (holocene calendar), things have change a bit thanks to the help of computer-software. Indeed, let us explain the issues of this old approach based on the CV.

As we know, the coefficient of variation is a scaled version of the standard deviation of the historical sales. But this tells us nothing about the easiness to forecast sales or not. Let us illustrate this with a simple example. Consider an item that has more or less level sales with a lot of variability and an item that has seasonal sales with no randomness whatsoever. The first is difficult to forecast, while the second is as easy as it gets. As the figure below illustrates, the coefficient of variation would not indicate this, giving to the seasonal series a higher value:

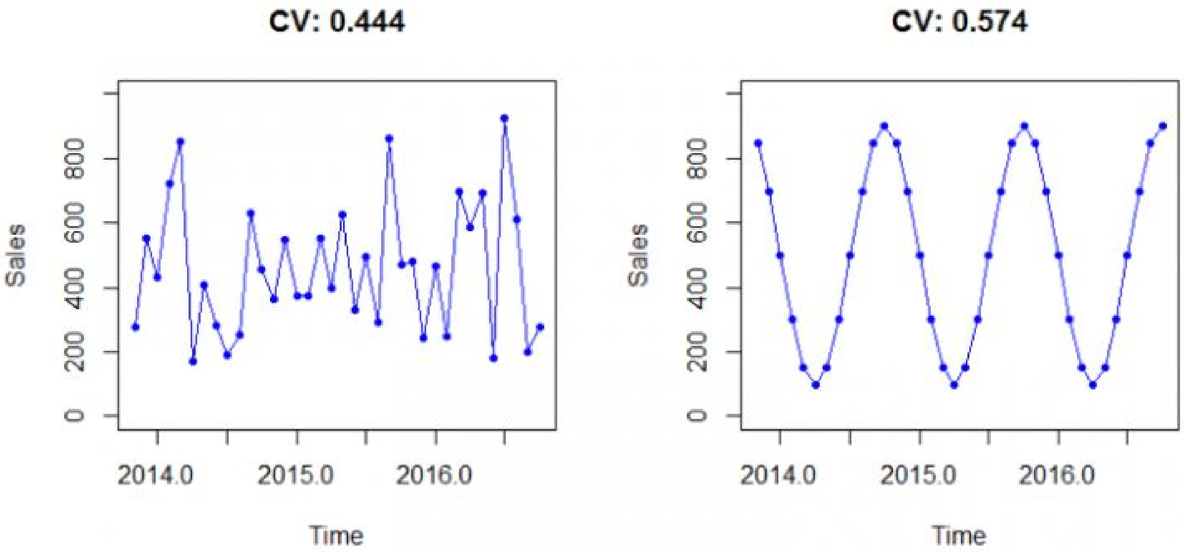


Figure 16.232: **Example series that the coefficient of variation fails to indicate which one is more difficult to forecast**

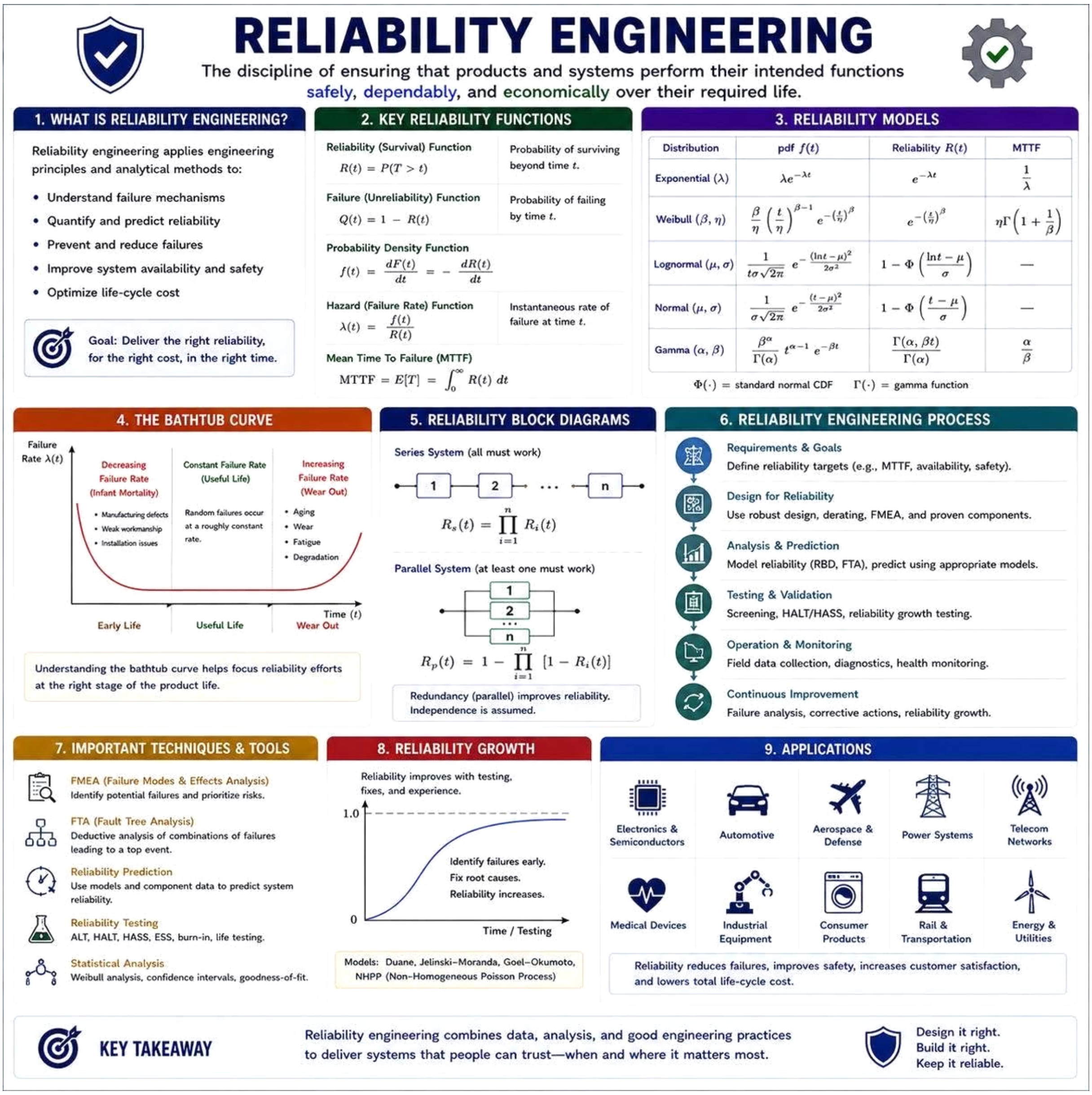
A better measure is forecast errors (see section Economy page 7903), which would directly relate to the non-forecastable parts of the series. This introduces a series of different questions: which forecasting method to use? which error metric? should it be in-sample or out-of-sample errors? Again, there is no perfect answer. Ideally we would like to use out-of-sample errors, but that would require us to have a history of forecast errors from an appropriate forecasting method, or conduct a simulation experiment with a holdout.

As for the method, this is perhaps the most complicated question... A single method would not be adequate, the reason being the same as for coefficient of variation. The latter implies that the forecasting method is the arithmetic mean (the value from which the standard deviation is calculated). An appropriate set of methods should be able to cope with all levels, trends and seasonal time series. A simplistic solution is to use naive (random walk) and seasonal naive, with a simplistic selection routine. The difference between seasonal and non-seasonal time series is typically substantial enough to make even weak selection rules work fine. An even better solution is to use a family of models, such as exponential smoothing, and do proper model selection, for instance using AIC or similar information criteria.

Notice that in reality the error metric should be robust (do not use percentage errors for this and not parametric statistics!) and be scale independent. Empirical solutions exists (like the "root error" that involves the use of complex number and the Gauss plane) but they are actually not very common in practice and not implemented in softwares (at least in this beginning of the 121st century!).

Remark

Because the XYZ analysis is just an ABC analysis but applied on multiple times series but just with a slightly different method of calculation, the reader will maybe understand that we will not provide here any detailed example. For application examples of the XYZ analysis however, the reader can also refer to the `R` software companion book.



RELIABILITY ENGINEERING

The discipline of ensuring that products and systems perform their intended functions
safely, dependably, and economically over their required life.

1. WHAT IS RELIABILITY ENGINEERING?

Reliability engineering applies engineering principles and analytical methods to:

- Understand failure mechanisms
- Quantify and predict reliability
- Prevent and reduce failures
- Improve system availability and safety
- Optimize life-cycle cost

Goal: Deliver the right reliability, for the right cost, in the right time.

2. KEY RELIABILITY FUNCTIONS

Reliability (Survival) Function

$R(t) = P(T > t)$

Probability of surviving beyond time t.

Failure (Unreliability) Function

$Q(t) = 1 - R(t)$

Probability of failing by time t.

Probability Density Function

$$f(t) = \frac{dF(t)}{dt} = -\frac{dR(t)}{dt}$$

Hazard (Failure Rate) Function

$$\lambda(t) = \frac{f(t)}{R(t)}$$

Instantaneous rate of failure at time t.

Mean Time To Failure (MTTF)

$$\mathrm{MTTF} = E[T] = \int_0^\infty R(t)\, dt$$

3. RELIABILITY MODELS

Distribution	pdf $f(t)$	Reliability $R(t)$	MTTF
Exponential (λ)	$\lambda e^{-\lambda t}$	$e^{-\lambda t}$	$\frac{1}{\lambda}$
Weibull (β, η)	$\frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} e^{-\left(\frac{t}{\eta}\right)^{\beta}}$	$e^{-\left(\frac{t}{\eta}\right)^{\beta}}$	$\eta\Gamma\left(1+\frac{1}{\beta}\right)$
Lognormal (μ, σ)	$\frac{1}{t\sigma\sqrt{2\pi}} e^{-\frac{(\ln t-\mu)^2}{2\sigma^2}}$	$1-\Phi\left(\frac{\ln t-\mu}{\sigma}\right)$	—
Normal (μ, σ)	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$	$1-\Phi\left(\frac{t-\mu}{\sigma}\right)$	—
Gamma (α, β)	$\frac{\beta^{\alpha}}{\Gamma(\alpha)} t^{\alpha-1} e^{-\beta t}$	$\frac{\Gamma(\alpha, \beta t)}{\Gamma(\alpha)}$	$\frac{\alpha}{\beta}$

$\Phi(\cdot)$ = standard normal CDF $\Gamma(\cdot)$ = gamma function

4. THE BATHTUB CURVE

Failure Rate $\lambda(t)$

Decreasing Failure Rate (Infant Mortality)
- Manufacturing defects
- Weak workmanship
- Installation issues

Constant Failure Rate (Useful Life)

Random failures occur at a roughly constant rate.

Increasing Failure Rate (Wear Out)
- Aging
- Wear
- Fatigue
- Degradation

Time (t)

Early Life Useful Life Wear Out

Understanding the bathtub curve helps focus reliability efforts at the right stage of the product life.

5. RELIABILITY BLOCK DIAGRAMS

Series System (all must work)

1 — 2 — … — n

$$R_s(t) = \prod_{i=1}^{n} R_i(t)$$

Parallel System (at least one must work)

1, 2, …, n

$$R_p(t) = 1 - \prod_{i=1}^{n} [1 - R_i(t)]$$

Redundancy (parallel) improves reliability. Independence is assumed.

6. RELIABILITY ENGINEERING PROCESS

Requirements & Goals
Define reliability targets (e.g., MTTF, availability, safety).

Design for Reliability
Use robust design, derating, FMEA, and proven components.

Analysis & Prediction
Model reliability (RBD, FTA), predict using appropriate models.

Testing & Validation
Screening, HALT/HASS, reliability growth testing.

Operation & Monitoring
Field data collection, diagnostics, health monitoring.

Continuous Improvement
Failure analysis, corrective actions, reliability growth.

7. IMPORTANT TECHNIQUES & TOOLS

FMEA (Failure Modes & Effects Analysis)
Identify potential failures and prioritize risks.

FTA (Fault Tree Analysis)
Deductive analysis of combinations of failures leading to a top event.

Reliability Prediction
Use models and component data to predict system reliability.

Reliability Testing
ALT, HALT, HASS, ESS, burn-in, life testing.

Statistical Analysis
Weibull analysis, confidence intervals, goodness-of-fit.

8. RELIABILITY GROWTH

Reliability improves with testing, fixes, and experience.

1.0

0

Identify failures early.
Fix root causes.
Reliability increases.

Time / Testing

Models: Duane, Jelinski–Moranda, Goel–Okumoto, NHPP (Non-Homogeneous Poisson Process)

9. APPLICATIONS

Electronics & Semiconductors
Automotive
Aerospace & Defense
Power Systems
Telecom Networks
Medical Devices
Industrial Equipment
Consumer Products
Rail & Transportation
Energy & Utilities

Reliability reduces failures, improves safety, increases customer satisfaction, and lowers total life-cycle cost.

KEY TAKEAWAY

Reliability engineering combines data, analysis, and good engineering practices to deliver systems that people can trust—when and where it matters most.

Design it right.
Build it right.
Keep it reliable.

16.7.4 Design of Experiments (DoE)

The behaviour or the subjective assessment of industrial or manufactured products is generally a function of many phenomena, often dependent on each other. To predict the behaviour / assessment, the products and phenomena are modelled and simulations are performed. The relevance of simulation results obviously depends on the quality of the models.

In particular, as part of the design or redesign of a product, the models generally involve a number of physical quantities (parameters) which we are authorized or not to modify. The behaviour of industrial products generally depends on many phenomena, often dependent on each other. To predict this behaviour, product and phenomena are modelled and simulations are performed.

However, these tests (simulations) are many times expensive, and this even more if the number of parameters to vary is important. Indeed, changing a parameter may for example require disassembly and reassembly of the product, a destructive analysis, or the manufacture of several different prototypes (case of a piece produced in series), or the interruption of production to change tool (for a manufacturing process) ... The cost of an experimental study depends on the number and order of the tests carried out.

The idea then is to select and order tests in the purpose to identify, at lower costs, the effects of parameters on the response of the product. These are statistical methods using most often simple mathematical concepts. The implementation of these methods has three stages (but the practitioner should better refer to the ISO 3534-3:1999 standard on the subject to be more rigorous!):

1. Postulate a (local!) behaviour model of the system (with coefficients being unknown)
2. Do data transformation to adjust the model to the assumptions of DoE tools and techniques (typically log-transformation for rates, etc.)
3. Check that the assumptions of the DoE models are satisfied for your experiment after data transformation
4. Define an experiment protocol, i.e. a series of tests ("runs") for identifying the model coefficients
5. Do the tests, identify the coefficients ("runs") and influential variables
6. Determine the values of the variables that allow you to get as closed as desired to the target result or with minimal variability and conclude.

The experimental designs, or "Design of Experiment (DoE)" (related to the field of "diagnostic analytics") allow to better organize the tests that accompany scientific research or industrial research. They are applicable to many disciplines and to all industries from the moment the practitioner is seeking for the relation between a variable of interest, y (amount of scrap, defects detection, amplitude, vacations, satisfaction, etc.) and control or environmental variables x_i in the purpose of:

- Determining which variables x_i are most influential on the response y
- Determining where to set the influential variables x_i so that y is almost always near the desired nominal value
- Determining where to set the influential x_i so that variability in y is small
- Determining where to set the influential variables x_i so that the effects of uncontrollable variables (denote z_i most of times) are minimized

This is why there are softwares to treat them such as `R`, Minitab, QI Macros or JMP mainly to name the most known one.

Let us also indicate that the experimental designs are a pillar of chemometrics (mathematical tools, especially statistics, to get the maximum information from chemical data) but are also used in supply chain optimization, marketing product design and computer network dimensioning. But designed experiments were first used by agronomists during the 120th century (holocene calendar). This method seemed highly theoretical at first, and was initially restricted to agronomy. Genichi Taguchi made the designed experiment approach more accessible to practitioners in the manufacturing industry.

In the last years of the 120th century (holocene calendar), the application of design of experiments has developed, particularly given the acknowledged fact that they are essential to improving the quality of goods and services. Even if statistical quality control, managerial solutions, inspections, and other quality tools also perform this function, the design of experiment is the methodology of choice in the case of an environment with complex parameters, variables and interactions and when we need to quantify the facts and benefits. From a historical point of view, the design of experiment have been developed and evolved in the agriculture sector. Medicine/Pharma has also a long history of design of experiments developed with caution. Currently, industrial environments prove the significant benefits of this methodology, because of the ease of initiation of efforts (of user-friendly application software), better training, influential trainers, and many successes achieved by DoE. Furthermore, today the process of analysis and manipulations is automated by robots in most high level industries.

> ❞ *A lot of things will go wrong before everything goes right.* ❞

quote
16.7:8781.#92

There are three main types of recognized DoE actually:

1. "Screenings designs" or "factor screening experiments" whose purpose is to discover the most influential factors on a given response with a minimal number of experiences. This is the simplest family as considered close to the experimental intuition (it is sometimes considered a subfamily of the second family below because it disregards the interactions of factors and is thus reduced to a purely additive model).

2. "Modelization plans" whose purpose is to find the mathematical relations between the measured responses to the variables associated with factors through an analytical approach or purely matrix one. Complete factorial and fractional designs (2 levels by factors with linear models) and response surfaces designs (at least 3 levels in models with factors of the second degree) are part of this family.

3. "Mixture designs" whose purpose is the same as the second family, but where the factors are not independent and are forced by a dependence relation (e.g. the sum of their ratio must be typically equal to a given constant).

The technique of DoE has for purpose to be scientifically more advantageous than techniques consisting in only change one factor at a time (thus still let fixed the other factors) because the latter does not allow a deep statistical study of errors and (especially!) hides completely the interactions.

Indeed, consider a team providing the facts that when a factor A is changed alone an improvement of something is visible! That the same team observes that when a factor B is changed alone in the same experiment and improvement of the same thing is also visible! We may then asked ... What happens if we change both A and B at the same time? Can we expect double the improvement? Or could the two changes cancel each other out? That simple question tell us quite a lot about the one factor at a time (OFAT) issue!

The general principle of the strategy of DoE is based on the fact that the study of a phenomenon can, usually, be summarized as follows: we are interested in a quantity y that depends on many variables, $x_1, x_2, \ldots, x_n$ (and their order has no influence ... which is a problem in some chemistry experimentation...).

The mathematical modelling consists to find a function f such that:

$$y = f(x_1, x_2, \ldots, x_n) \tag{16.7.372}$$

that takes into account the influence of each single factor or combination of factors (interactions).

A conventional study method consists of measuring the response y for several values of the variables x_1 o while leaving the other $n-1$ variables fixed. We then iterate this method for each variable.

Therefore, for example, if we have 8 variables and if we decides to give 2 experimental values to each of them, we are led to make $2^8 = 256$ experiences.

The use of an experimental design then gives a strategy in choosing the test methods. The success of experiments in research and industry is related to the need for business competitiveness: they allow quality improvement and cost reduction!

Remark

The method of DoE was developed early in the century, in the 11920s (holocene calendar), by Ronald A. Fisher (the same for man as for the Fisher Test!). It has grown considerably with the development of information technology and computing power that accompanies it.

The treatment of results is made most of time using univariate or multivariate linear regression techniques (see section Numerical Methods page 6316) and that of analysis of variance (see section Statistics page 1001).

With the DoE, the goal is therefore to obtain the maximum information (but not all information!) with minimum of experiences (and therefore a minimum of cost) in order to model and optimize the studied phenomena.

An experimenter who starts a study is interested in a quantity that he measures at each test. This quantity is named obviously the "response" or "exogenous variable" and is the quantity of interest. The value of this quantity depends on several variables. Instead of the term "variable" or "endogenous variable" we will use the word "factor". The value given to a factor to achieve a test is named a "level". When studying the influence of a factor, in general we limit is variations between two bounds (yes we must have to stop one day and be reasonable...) respectively named "lower level" and "upper level".

Obviously, when we have several factors $x_1, x_2, \ldots, x_n$ they represent a point in $\mathbb{R}^n$ named "experimental space".

The set of all possible values that a factor can take between the lower level and the upper level is named the "domain of variation of the factor" or simply the "factor domain". The usage is to denote the lower level by -1 and the upper level by $+1$ when all the factors have only two levels, otherwise it is customary to denote the levels from 1 to n (number of levels). Of course, depending on the chosen notation, the expressions of the mathematical models must be adapted accordingly.

Remark

If the factors don't have all the same number of level we say that we have an "irregular experimental region" and the most simple corresponding design is then a complete (general) factorial design.

So for example for a factor having a domain of variation between an upper level of $20\,[^\circ\text{C}]$ corresponds to $+1$ and a lower level of $5\,[^\circ\text{C}]$ corresponding to -1 we will have at the end of our study to transform all the experimental values in "centered reduced units" in which the x_i must be used.

Thus, we have two entry points $(20, 5)$ and two outputs $(+1, -1)$. Any intermediate value is simply given by the equation of the line:

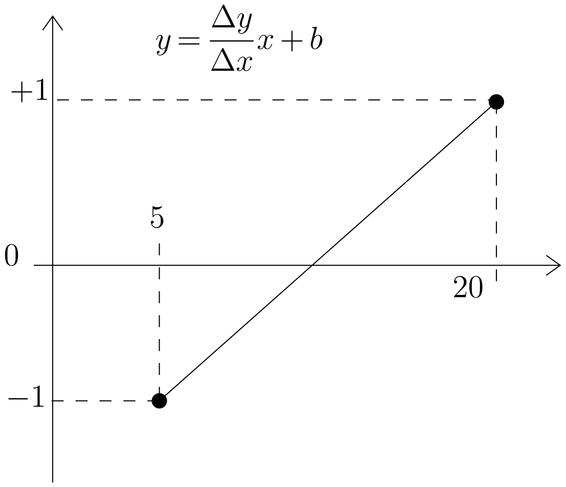


Figure 16.233: Principle of construction of centered reduced units

The slope is obvious to obtain, to get b, we just have a simple linear equation with one unknown to solve:

$$+1 = \frac{2}{20-5}20 + b \Rightarrow b = +1 - \frac{2}{15}20 = -1.\bar{6} \tag{16.7.373}$$

or (what remains the same):

$$-1 = \frac{2}{20-5}5 + b \Rightarrow b = -1 - \frac{2}{15}5 = -1.\bar{6} \tag{16.7.374}$$

Therefore the transition form the non-normalized variables, denoted by x, to the normalized one, denoted by X, is then written:

$$\boxed{X = \frac{\Delta y}{\Delta x}x + b = \frac{\Delta y}{\Delta x}x + \left(1 - \frac{\Delta y}{\Delta x}x_{\text{max}}\right) = \frac{\Delta y}{\Delta x}(x - x_{\text{max}}) + 1} \tag{16.7.375}$$

Therefore in the case of an $(+1, -1)$ output:

$$X = \frac{2}{\Delta x}(x - x_{\text{max}}) + 1 \tag{16.7.376}$$

and vice-versa:

$$\boxed{x = (X - b)\frac{\Delta x}{\Delta y} = \left[X - \left(1 - \frac{\Delta y}{\Delta x}x_{\text{max}}\right)\right]\frac{\Delta x}{\Delta y}} \tag{16.7.377}$$

that is to say in the case of an $(+1, -1)$ output:

$$x = \left[X - \left(1 - \frac{2}{\Delta x}\right)x_{\text{max}}\right]\frac{\Delta x}{2} \tag{16.7.378}$$

The set of domains of all factors defines the "study domain". This study domain is the experimental space chosen by the experimenter to make his tests. A "study", that is to say multiple well defined experiments, is represented by points distributed in the study domain.

For example, for two factors, the study domain is an experimental plane surface in $\mathbb{R}^2$:

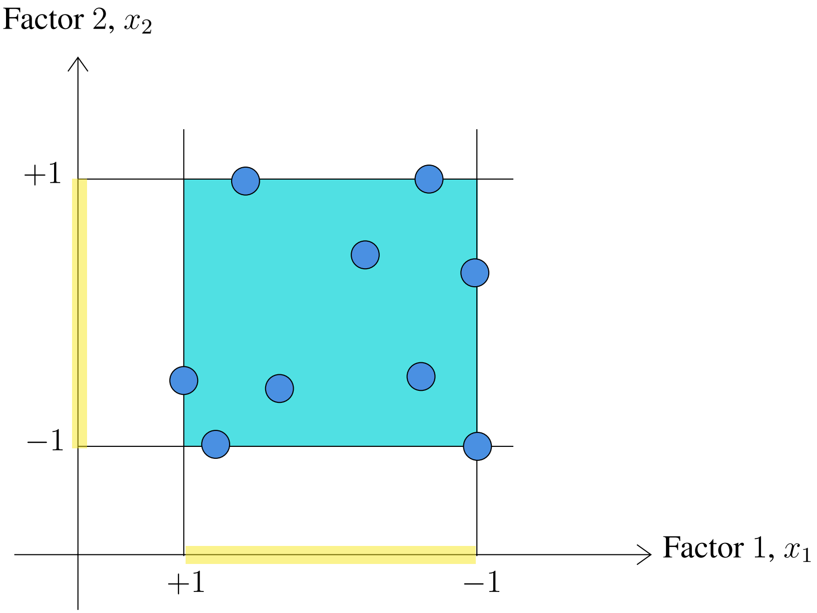


Figure 16.234: **Two normalized factors study domain example**

The levels of the reduced centered x_i represents the coordinates of an experimental point and y is the answer of this point. The graphical geometric representation of a DoE and the answer need always one dimension more than the dimension of the experimental space. A DoE with two factors need therefore a three dimension space to be represented: one dimension for the answer, two dimensions for the factors.

When we have determined the mathematical model, to each point of the study domain, corresponds then a set of answers that are located on a surface (or hypersurface) named "response surface" (this is why typically DoE are named in the general case: "DoE for response surfaces") that looks like for example with two factors:

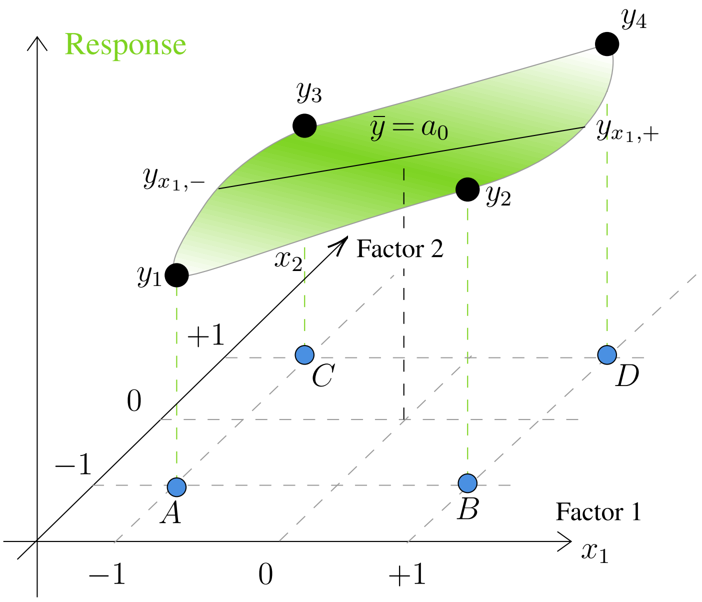


Figure 16.235: **Two factors with the response surface**

The number and the position of the experimental points is one of the fundamental problems of DoE. Indeed, we are looking for the best precision on the response surface by minimizing at the same time the number of experimentations. The practitioner will therefore look for a mathematical function that links the response variable to the factors.

For this, we simplify the problem first by remembering (see section Sequences and Series page 1696) that any continuous function, whatever is its number of variables, can be approximated in a sum of power series on a given point.

We then take a limited Taylor development of a bivariate function (for example):

$$\begin{aligned}f(x,y)=&f(x_0,y_0)+\frac{1}{1!}\frac{\partial f(x_0,y_0)}{\partial x}(x-x_0)+\frac{1}{1!}\frac{\partial f(x_0,y_0)}{\partial x}(y-y_0)\\&+\frac{1}{2!}\left[\frac{\partial^2 f(x_0,y_0)}{\partial x^2}(x-x_0)^2+2\frac{\partial^2 f(x_0,y_0)}{\partial x\partial y}(x-x_0)(y-y_0)+\frac{\partial^2 f(x_0,y_0)}{\partial y^2}(y-y_0)^2\right]+\ldots\end{aligned}\tag{16.7.379}$$

Therefore at the neighbourhood of $x_0=0,y_0=0$ (what we can do as we take the centered reduced variables... and this is also one of the reason we centered and reduced them!), we have the following second order Maclaurin series by change notation accordingly to the traditions in DoE for $i=1,2$:

$$\begin{aligned}&y=f(x_1,x_2)\\&\cong f(0,0)+\frac{1}{1!}\frac{\partial f}{\partial x_1}(0,0)x_1+\frac{1}{1!}\frac{\partial f}{\partial x_2}(0,0)x_2+\frac{1}{2!}\left(\frac{\partial^2 f}{\partial x_1^2}(0,0)x_1^2+2\frac{\partial^2 f}{\partial x_1\partial x_2}(0,0)x_1x_2+\frac{\partial^2 f}{\partial x_2^2}(0,0)x_2^2\right)\\&\cong a_0+\sum a_ix_i+\sum a_{ij}x_ix_j=a_0+\sum_i a_ix_i+\sum_{i\neq j}a_{ij}x_ix_j+\sum_i a_{ii}x_ix_i\end{aligned}\tag{16.7.380}$$

where y is the response and the x_i the factors and the a_0,a_i,a_{ij},a_{ii} the coefficients of the a priori adopted mathematical model. They are not known and must be calculated from the results of experiments and are included in the model depending on the type of the chosen design. In general we want (in the ideal case) to determine 4 terms:

- a_0: the constant coefficient
- a_i: coefficient of first degree
- x_i: main term
- a_{ij}: rectangle (quadratic) coefficient
- x_{ij}: rectangle (quadratic) term
- a_{ii}: pure quadratic coefficient
- x_{ii}: pure quadratic coefficient

The benefit (pros) of modelling the response by a polynomial is to calculate afterwards all the answers of the study area without having to do the experiments thanks to a model named "postulated local model" or "prior model".

The disadvantage (cons) of a polynomial model is that it is not bounded (diverge or: non-lipschitz function) and has therefore no asymptotic answer. In some situation NURBS[25] (or splines in the case of a univariate function) could be more accurate but sadly we do not know actually how to make statistical inference with NURBS to determine which factors are significant or not...

Remark

The practitioner has to make the difference between a "physical model" that works in a wide range or even on the whole range of the experimental field with an "experimental design" that works only on a small and limited range of the experimental field. Hence the model of the experimental design works only in a small area named sometimes the "inference domain".

Two complements need to be made to the model described above:

1. The first complement is the "lack of fit". This expression reflects the fact that prior model is likely almost surely different (only by the approximation of the approach) of the real model that governs the studied phenomenon. There is a difference between these two models. This gap is the lack of fit!
2. The second complement is to take into account the random nature of the response (without the latter to be stochastic otherwise we have to use other tools!). Indeed, if we measure a response several times in the same experimental point, we do not get exactly the same result. The results are scattered. The dispersions of this kind are named "experimental error" or "residuals" and are supposed to follow a supposed to follow a Normal law.

These two differences, lack of fit and experimental error, are most of time merged into only one error denoted e (or sometimes as in linear regression by ε).

The model used by the experimenter will then be written at the second order term and first degree:

$$y \cong a_0 + \sum a_i x_i + \sum_{i \neq j} a_{ij} x_i x_j + e \tag{16.7.381}$$

and is sometimes named "synergic model" and in the special case of two variables we speak of a " hyperbolic paraboloid model" typically represented by (we give a few examples with Maple™ 4.00b):

```
>plot3d(0+x1+x2+x1*x2,x1=-10..10,x2=-10..10,contours=10,style=PATCHCONTOUR,
axes=frame,numpoints=10000);
```

---

[25] Non-Uniform Rational Basis Splines

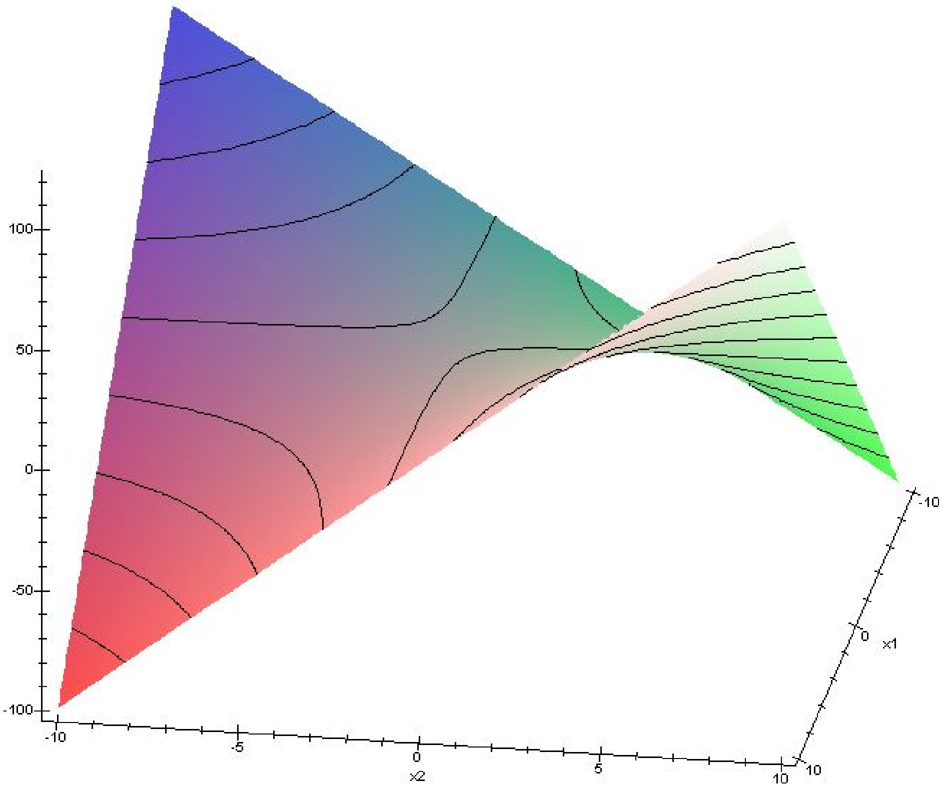
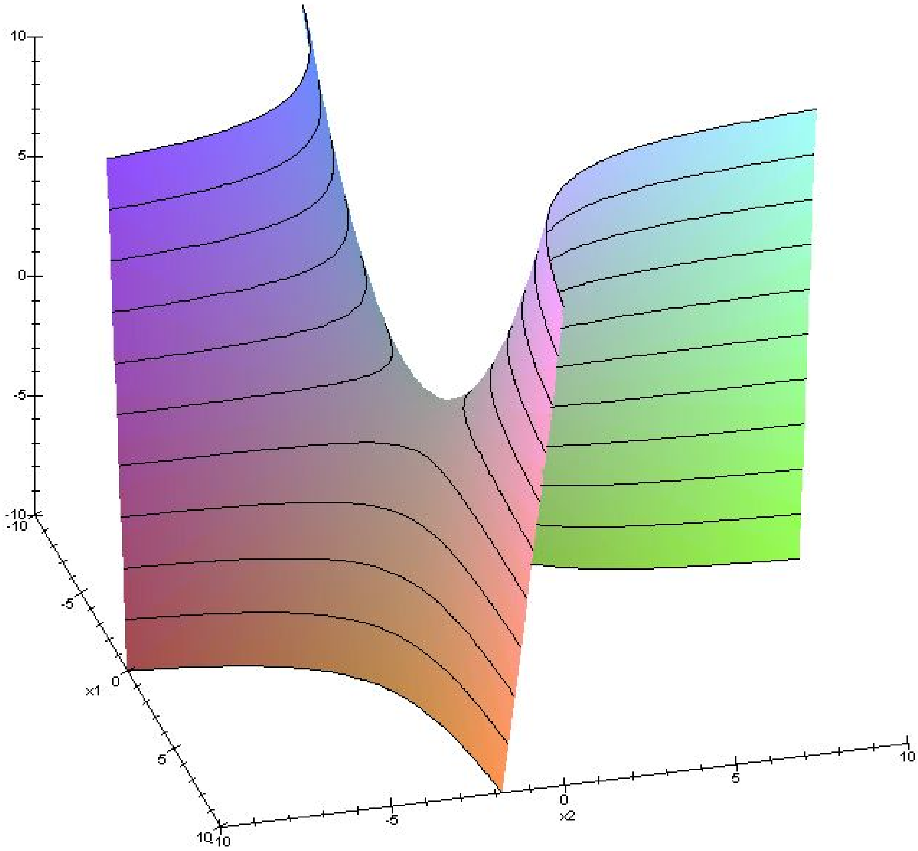


Figure 16.236: **Generic example of interaction for two factors (hyperbolic paraboloid)**

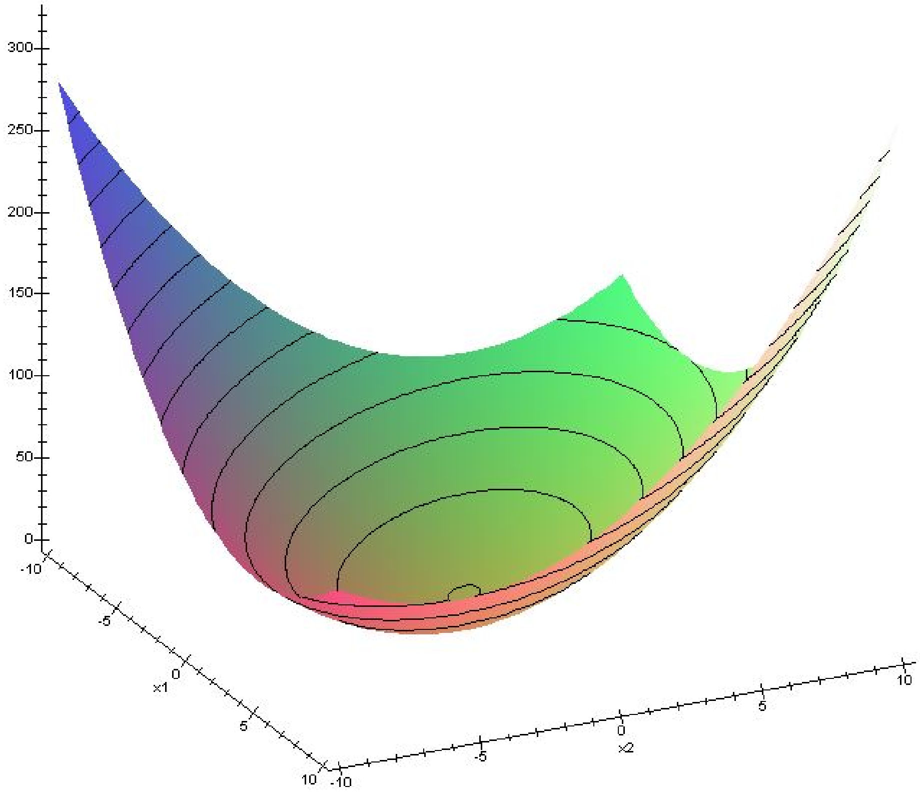
and the same with a zoom:

```
>plot3d(0+x1+x2+x1*x2,x1=-10..10,x2=-10..10,view=[-10..10,-10..10,-10..10],
contours=10,style=PATCHCONTOUR,axes=frame,numpoints=10000);
```

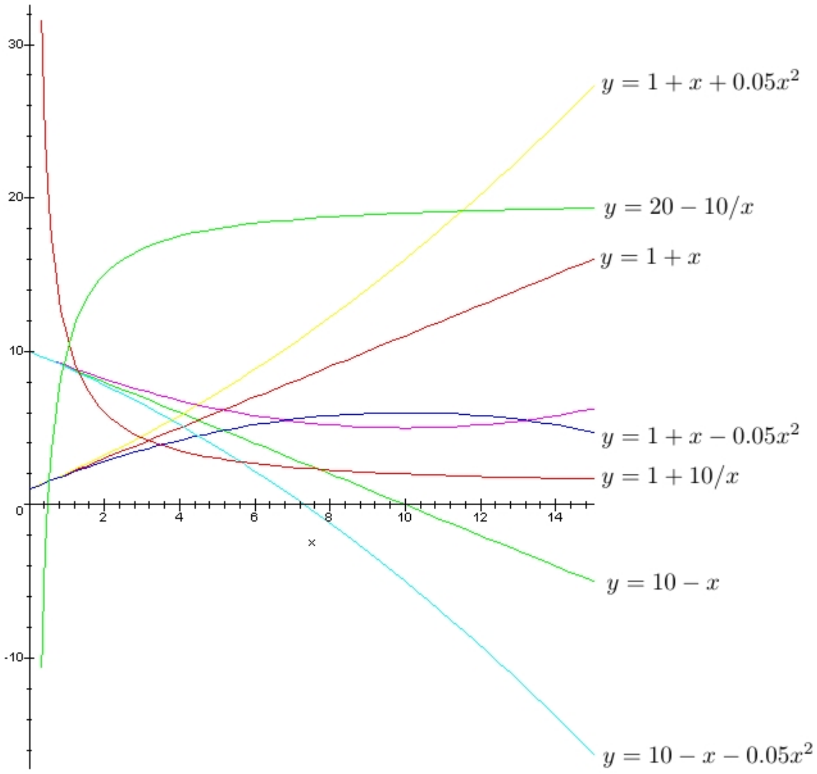


And if we include pure quadratic terms:

```
>plot3d(0+x1+x2+x1*x2+x1^2+x2^2,x1=-10..10,x2=-10..10,contours=10,
style=PATCHCONTOUR,axes=frame,numpoints=10000);
```

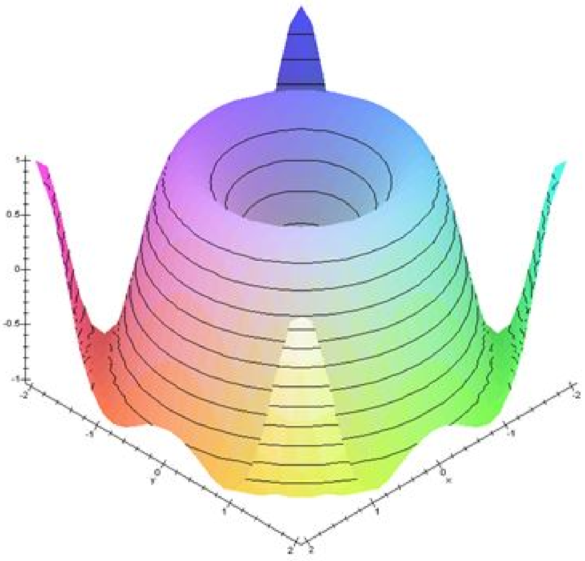


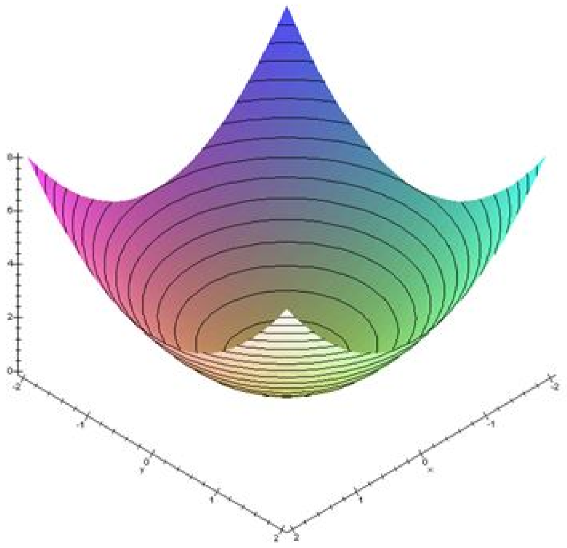
and as requested by a reader, below we provide a plot of some univariate functions (to differentiate the linear a non-linear case when variables transformations could be necessary):

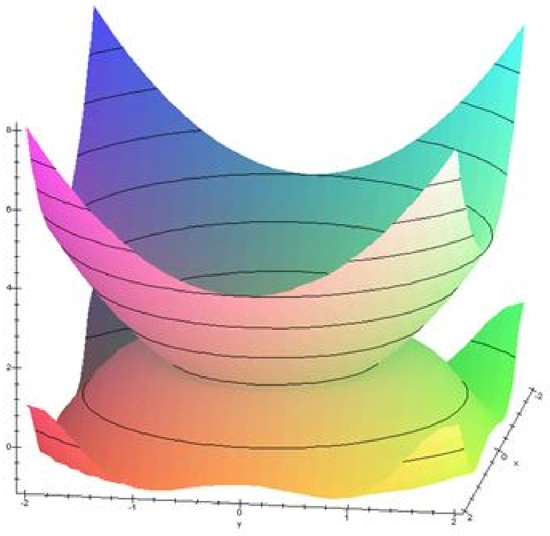


So keep in mind that depending on the experiment you have, and the implicit physico-chemical function, the Taylor approximation to the third order may be quite bad as we already know from our study of Taylor series. Let us take for example the function $\sin(x^2+y^2)$ that is approximated at the third order by... x^2+y^2, we then have with Maple™ 4.00b:

```
>readlib(mtaylor):

>plot3d(sin(x^2+y^2),x=-2..2,y=-2..2,contours=10,style=PATCHCONTOUR,axes=frame);
>plot3d(mtaylor(sin(x^2+y^2), [x,y], 3),x=-2..2,y=-2..2,contours=10,
style=PATCHCONTOUR,axes=frame); >plot3d(sin(x^2+y^2),mtaylor(sin(x^2+y^2),
[x,y], 3),x=-2..2,y=-2..2,contours=10,style=PATCHCONTOUR,axes=frame);
```







Remark

If we stop the Taylor development to the first order term and first degree (without interactions), we then speak of "affine model".

In practice we write the latter relation as following (we remove the error term for readiness purposes):

$$y \cong a_0 + a_1 x_1 + a_2 x_2 + a_{12} x_1 x_2 \tag{16.7.382}$$

where we have the abusive notations for what it is common to name in this field the "rectangle term":

$$a_{12} := a_{21} + a_{12} \tag{16.7.383}$$

Therefore it comes that for 2 factors, the expression contains 4 terms, for 3 factors it contains 8 terms, for 4 factors it contains 16 terms, and so on... There are always power of 2!

Remark

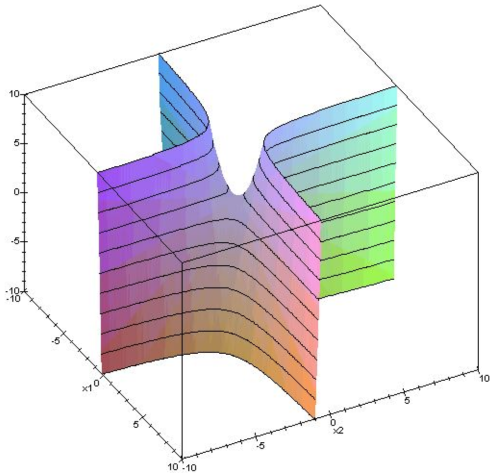
Consider k factors $F_1, \ldots, F_k$ each with two levels. There are $\begin{pmatrix} k \\ 1 \end{pmatrix}$ main effects, $\begin{pmatrix} k \\ 2 \end{pmatrix}$ two-interaction effects, $\begin{pmatrix} k \\ 3 \end{pmatrix}$ three-factor interaction effect, ..., $\begin{pmatrix} k \\ k \end{pmatrix} = 1$ k-factor interaction.

This model without error is many times named "controlled (linear) model (or order 2) with interactions".

✍**Example:**

For example, a response surface associated with a relation like the previous one looks like with Maple™ 4.00b:

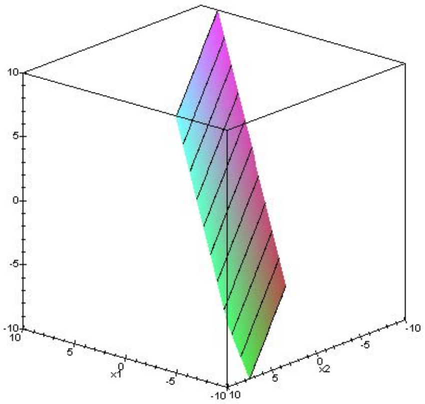
```
>plot3d(5+3*x1+2*x2+4*x1*x2,x1=-10..10,x2=-10..10,
view=[-10..10,-10..10,-10..10],contours=10,style=PATCHCONTOUR,
axes=frame,numpoints=10000);
```



where we have represented the isolines (see section Functional Analysis page 2857). This permits the practitioner to search the combination of explanatory variables that for a same result is the less expensive!

Obviously if we remove the interaction terms we simply get a plane:

```
>plot3d(5+3*x1+2*x2,x1=-10..10,x2=-10..10,
view=[-10..10,-10..10,-10..10],contours=10,
style=PATCHCONTOUR,axes=frame,numpoints=10000);
```



Obviously, the higher the degree of the polynomial is, the higher the theoretic model will be close to reality. But (!!!) the high degree polynomials require a lot of data points and their validity can quickly diverge outside the experimental area (see section Calculus page 1368). If the study requires it, we prefer to use specific mathematical functions to better fit the model to the experimental results.

However, in practice, the higher order interactions often have very little influence on the response (in fact this assumption depends strongly on the industrial/economical field ...!). It is therefore possible to not to include them in the model, which leads to less trials. This principle is used in the construction of many experimental designs, as we shall see in the next subsection. In many applications, we obtain results quite satisfactory results by limiting ourselves to double interactions.

Why do we satisfy us of this approximate relation with four terms? For the simple reasons that:

1. The answer may be non-zero when all factors are zero (thanks to the first coefficient a_0).
2. The answer depends trivially (intuitively) of the sum of the effects of the first and second factors x_1, x_2 in an independent way (coefficients a_1, a_2).
3. The answer also depends on the interaction between the two factors x_1, x_2 (coefficient a_{12}).

Each experimental point for which the x_i are given then provides a value of the response y. This response is modelled by a polynomial whose coefficients are the unknowns and must be determined and also categorized as significant or not.

16.7.4.1 Two levels factorial Designs

So in an experimental design of 2 factors with 2 levels, we need at least (and at most for cost reasons!) 4 measurements (trials) to have a system of 4 equations with 4 unknowns which are the coefficients a_0, a_1, a_2, a_{12} and such that the system has a unique solution (see section Linear Algebra page 1929)

Remark

For a study of 2 factors with 3 levels, we can no longer take a linear model. Then we must take the quadratic terms of the Taylor expansion!

Since for each of the factors we need to set a low and a high level in order to work reasonably... so if we have two factors, we have an experimental space defined by 4 points (up, up), (down, down) (up, down), (down, up), corresponding to the 2 times 2 levels (2^2), which is enough for us then to get our system of 4 equations with 4 unknowns and then determine the 4 coefficients (uniquely).

Thus, the points to be taken in our experience naturally correspond to the vertices (geometrically speaking) of the experimental space!

We then have the following system of equations (recall that this approach works only for 2 level factors, the case with 3 levels will be study much further below with an example):

$$\begin{aligned}
y_1 &\cong a_0 + a_1x_{1,-1} + a_2x_{2,-1} + a_{12}x_{1,-1}x_{2,-1}\\
y_2 &\cong a_0 + a_1x_{1,+1} + a_2x_{2,-1} + a_{12}x_{1,+1}x_{2,-1}\\
y_3 &\cong a_0 + a_1x_{1,-1} + a_2x_{2,+1} + a_{12}x_{1,-1}x_{2,+1}\\
y_4 &\cong a_0 + a_1x_{1,+1} + a_2x_{2,+1} + a_{12}x_{1,+1}x_{2,+1}
\end{aligned} \tag{16.7.384}$$

or explicitly:

$$\begin{aligned}
y_1 &\cong a_0 + a_1(-1) + a_2(-1) + a_{12}(-1)(-1) = a_0 - a_1 - a_2 + a_{12}\\
y_2 &\cong a_0 + a_1(+1) + a_2(-1) + a_{12}(+1)(-1) = a_0 + a_1 - a_2 - a_{12}\\
y_3 &\cong a_0 + a_1(-1) + a_2(+1) + a_{12}(-1)(+1) = a_0 - a_1 + a_2 - a_{12}\\
y_4 &\cong a_0 + a_1(+1) + a_2(+1) + a_{12}(+1)(+1) = a_0 + a_1 + a_2 + a_{12}
\end{aligned} \tag{16.7.385}$$

As we are interested about the coefficients (and that the values of the variables are known!), the problem simply consist in solving a linear system (see section Linear Algebra page 1929).

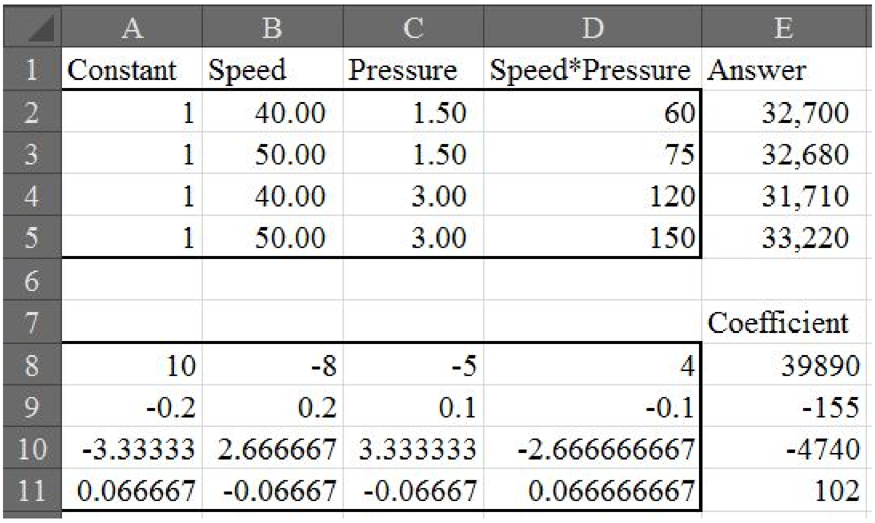
If the variables were not coded, we would have for example for a (two-level) complete factorial design with one variable (velocity of car) having for lower value $40\,[\text{km}\cdot\text{h}^{-1}]$ and for upper value $50\,[\text{km}\cdot\text{h}^{-1}]$, and for a second variable (wheel pressure) having for lower value $1.5\,[\text{Pa}]$ and for upper value $3\,[\text{Pa}]$ the following system to solve knowing that we have drive respectively for each combination the distances 32700, 32680, 31710, 33222 [km]:

$$\begin{aligned} a_0\cdot 1+a_1(40)+a_2(1.5)+a_{12}(60)(1.5)&\cong 32,700\\ a_0\cdot 1+a_1(50)+a_2(1.5)+a_{12}(50)(1.5)&\cong 32,680\\ a_0\cdot 1+a_1(40)+a_2(3)+a_{12}(40)(3)&\cong 31,710\\ a_0\cdot 1+a_1(50)+a_2(3)+a_{12}(50)(3)&\cong 33,220 \end{aligned} \tag{16.7.386}$$

Hence in matrix form:

$$\begin{pmatrix} 1 & 40 & 1.5 & 60\\ 1 & 50 & 1.5 & 75\\ 1 & 40 & 3 & 120\\ 1 & 50 & 3 & 150 \end{pmatrix}\begin{pmatrix} a_0\\ a_1\\ a_2\\ a_{12} \end{pmatrix}=\begin{pmatrix} 32,700\\ 32,680\\ 31,710\\ 33,220 \end{pmatrix} \tag{16.7.387}$$

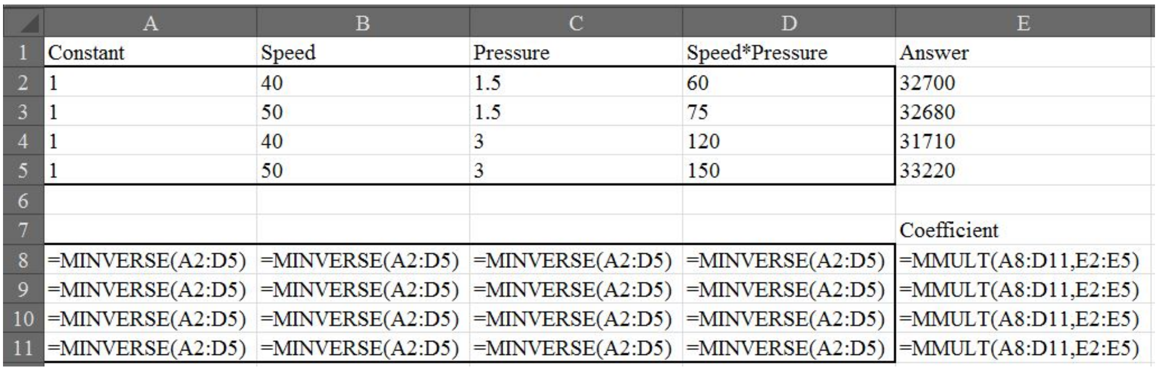
By solving this system by hand (see section Linear Algebra page 1929), or with a calculation software (spreadsheet or statistical software), we get:



	A	B	C	D	E
1	Constant	Speed	Pressure	Speed*Pressure	Answer
2	1	40.00	1.50	60	32,700
3	1	50.00	1.50	75	32,680
4	1	40.00	3.00	120	31,710
5	1	50.00	3.00	150	33,220
6					
7					Coefficient
8	10	-8	-5	4	39890
9	-0.2	0.2	0.1	-0.1	-155
10	-3.33333	2.666667	3.333333	-2.666666667	-4740
11	0.066667	-0.06667	-0.06667	0.066666667	102

Figure 16.237: **Putting in equation and implicit resolution of the FF DoE in Microsoft® Excel® 14.0.6123**

thus explicitly:



	A	B	C	D	E
1	Constant	Speed	Pressure	Speed*Pressure	Answer
2	1	40	1.5	60	32700
3	1	50	1.5	75	32680
4	1	40	3	120	31710
5	1	50	3	150	33220
6					
7					Coefficient
8	=MINVERSE(A2:D5)	=MINVERSE(A2:D5)	=MINVERSE(A2:D5)	=MINVERSE(A2:D5)	=MMULT(A8:D11,E2:E5)
9	=MINVERSE(A2:D5)	=MINVERSE(A2:D5)	=MINVERSE(A2:D5)	=MINVERSE(A2:D5)	=MMULT(A8:D11,E2:E5)
10	=MINVERSE(A2:D5)	=MINVERSE(A2:D5)	=MINVERSE(A2:D5)	=MINVERSE(A2:D5)	=MMULT(A8:D11,E2:E5)
11	=MINVERSE(A2:D5)	=MINVERSE(A2:D5)	=MINVERSE(A2:D5)	=MINVERSE(A2:D5)	=MMULT(A8:D11,E2:E5)

Figure 16.238: **Putting in equation and explicit resolution of the FF DoE in Microsoft® Excel® 14.0.6123**

Therefore the solution is finally:

$$y \cong 39,890 - 155x_1 - 4,740x_2 + 102x_{12} \tag{16.7.388}$$

thus exactly the same coefficients as those given by a specialized software like Minitab 15.1.1 (see the companion Minitab book).

Going back to our system with the coded variables, it then comes when solving the system algebraically (relations valid if and only if the variables are coded!):

$$\begin{aligned} a_0 &= \frac{1}{4}(y_1 + y_2 + y_3 + y_4) \\ a_1 &= \frac{1}{4}(-y_1 + y_2 - y_3 + y_4) \\ a_2 &= \frac{1}{4}(-y_1 - y_2 + y_3 + y_4) \\ a_{12} &= \frac{1}{4}(y_1 - y_2 - y_3 + y_4) \end{aligned} \tag{16.7.389}$$

which can be written in the following matrix form:

$$\begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_{12} \end{pmatrix} = \frac{1}{4} \begin{bmatrix} +1 & +1 & +1 & +1 \\ -1 & +1 & -1 & +1 \\ -1 & -1 & +1 & +1 \\ +1 & -1 & -1 & +1 \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} \tag{16.7.390}$$

What is written in a more general way for linear models of the second order under the following form (see section Linear Algebra page 1929):

$$A = \frac{1}{n} XY \tag{16.7.391}$$

The matrix X containing 2^n rows is named "full factorial design (FFDoE) with 2^n interactions" (the term "factorial" coming from the fact that all factors vary simultaneously).

The matrix X in practice is named the "experimental matrix" or "effect matrix" and is often represented as follows in the previous case:

Trial N°	Rest	Factor 1	Factor 2	Factor 12	Answer
1	$+1$	-1	-1	$+1$	y_1
2	$+1$	$+1$	-1	-1	y_2
3	$+1$	-1	$+1$	-1	y_3
4	$+1$	$+1$	$+1$	$+1$	y_4

Table 16.29: **Experimental Full Factorial (2 factor) matrix**

But we see immediately that in practice, the second column ("Rest") is unnecessary because it is always $+1$ and it is also only implicitly implemented in statistical softwares.

It is the same for the fifth column (Factor 12), because it is automatically deductible from the third and fourth columns (it is equal to the multiplication of row elements... what some practitioners name the "Box multiplication").

The reader will also notice that by passing from one column to another or from one row to another, there are always two factors that have the level that change! By cons, if we focus only on the columns Factor 1 and Factor 2, we see that from one row to another, there is one factor that changes at a time.

Remark

Once again notice that the first column has only $+1$ and the there is also always at least on row with only $+1$ values!

Thus, in practice (softwares) and in many books, we represent rightly only the following reduce form table (which hides the fact that we are dealing in reality with a square matrix):

Trial N°	Factor 1	Factor 2	Answer
1	-1	-1	y_1
2	$+1$	-1	y_2
3	-1	$+1$	y_3
4	$+1$	$+1$	y_4

Table 16.30: **Experimental Full Factorial (2 factor) reduced matrix**

for a 2 factors experimental design with 2 levels with interactions in linear model (no errors) we have even more extreme ("Yates notation") ... in terms of writing:

Trial N°	Factor 1	Factor 2	Answer
1	$-$	$-$	y_1
2	$+$	$-$	y_2
3	$-$	$+$	y_3
4	$+$	$+$	y_4

Table 16.31: **Experimental Full Factorial (2 factor) reduced matrix Yates notation**

We see better in this form of writing that in addition to the fact that the two columns Factor 1 and Factor 2 are orthogonal (the standard ISO 3534-3:1999 speaks of "orthogonal contrast"), they are also "balanced" in the sense that there are so many $+$ and $-$ in each column.

Remark

By default, most software randomize the order of testing of the design (whether it is complete, fractional, factorial or not). Generally, it is recommended to randomize the order of testing to mitigate the effects of factors that are not included in the study and that parasite it ("nuisance factor"), including effects related to the time. This is, we don't know that the factor exists and it may even be changing levels while we are conducting the experiment. However, in some cases, randomization does not produce a sequence of interesting essays and perhaps even dangerous because it can mask certain sources of systematic errors which were not identified before the experiment. For example, in industrial applications, modifying factor levels may be difficult or expensive. It is also possible that after the change done, the return to a steady state of the system takes a lot of time. In such cases, it may be desirable not to randomize the design to minimize the level changes. Also they are some cases where the nuisance factor is known but uncontrollable, but then we can compensate it by using the analysis of covariance (see section Statistics).

Practitioners appreciate to calculate the average of the answers and the average effect of a given factor since the system is linear. Thus, in at the level $+1$ for the factor x_1, by remaining on the linear system:

$$\begin{aligned}
y_1 &\cong a_0+a_1x_{1,-1}+a_2x_{2,-1}+a_{12}x_{1,-1}x_{2,-1}\\
y_2 &\cong a_0+a_1x_{1,+1}+a_2x_{2,-1}+a_{12}x_{1,+1}x_{2,-1}\\
y_3 &\cong a_0+a_1x_{1,-1}+a_2x_{2,+1}+a_{12}x_{1,-1}x_{2,+1}\\
y_4 &\cong a_0+a_1x_{1,+1}+a_2x_{2,+1}+a_{12}x_{1,+1}x_{2,+1}
\end{aligned} \tag{16.7.392}$$

then we can build and define the "average answers" given by:

$$y_{x_1,+}=\frac{1}{2}(y_2+y_4) \tag{16.7.393}$$

and we have the same with the level -1 for the same factor:

$$y_{x_1,-}=\frac{1}{2}(y_1+y_3) \tag{16.7.394}$$

We then have "overall effect" of the factor x_1 that will be given by:

$$E_{Gx_1}=y_{x_1,+}-y_{x_1,-} \tag{16.7.395}$$

and the "mean effect" of the factor x_1 that is therefore define by the semi-difference between the average of the answers at the upper level of the factor x_1 and the average answers at the lower level of the same factor:

$$E_{Mx_1}=\frac{1}{2}(y_{x_1,+}-y_{x_1,-})=\frac{1}{2}\left(\frac{y_2+y_4}{2}-\frac{y_1+y_3}{2}\right) \tag{16.7.396}$$

but after some simplification of elementary algebra, it comes quickly that:

$$E_{Mx_1}=a_1 \tag{16.7.397}$$

It is obvious that if the overall effect (and verbatim the mean effect) is not zero, we can doubt that there is an interaction between the factors as the variation in amplitude of the response is not the same depending on the value of the level of the other factor (see the study of the Analysis of Variance of two factors in the Statistics section for more details at page 1027). Obviously, in practice, the study of interactions will be made, as often for ANOVA, also with graphics (interaction diagrams).

The average of all responses thus gives the value of the response at the center of the experimental domain:

$$\bar{y}=\frac{1}{4}(y_1+y_2+y_3+y_4) \tag{16.7.398}$$

By doing some elementary algebra, that the plane is written in normalized and centered form or not, this expression reduces to:

$$\bar{y}=a_0 \tag{16.7.399}$$

Obviously, we can summarize this simple case graphically if this can help the reader:

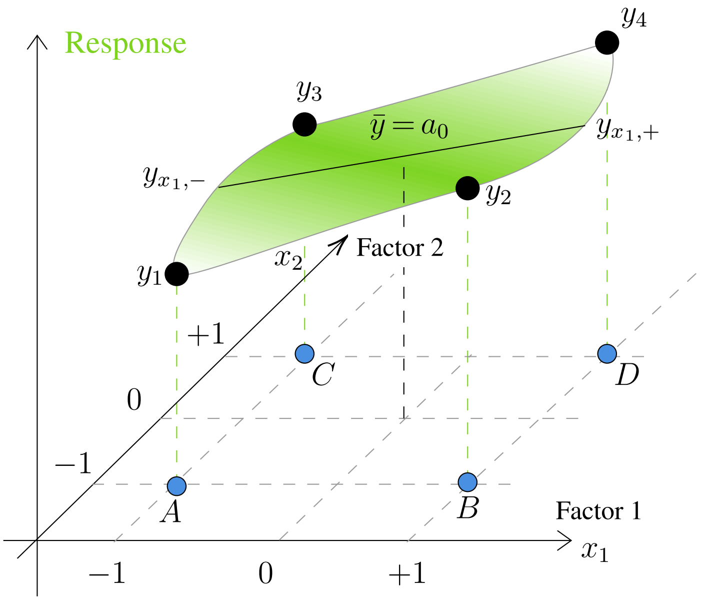


Figure 16.239: **Two factors with the response surface**

and with the figure above, the changes being not very parallel when a factor is set, we can assume that there is interaction between the factors and thus this will require the use of an ANOVA to have a more in-deep numerical analysis.

When the number of factors is large, it is not always easy for everyone to put the $+,-$ without the help of a software. So there is a small procedure named "Yates algorithm" or "Yates and Hunter algorithm" which enables to reach very fast the desired result for factorial designs (whose factors have two levels) whose number of factors is a power of 2: first, we start all columns by -1 and after we alternate the -1 and $+1$ all the 2^{j-1} rows for the jth row.

Remarks

R1. If the type of table above contains encoded values, we speak of "experimental design" if not with the usual physical units we speak of "experimental table".

R2. In the case of coded tables, it is customary to indicate under the table a second table with the correspondences between the coded units and physical units.

Let us insist on one important thing: If we had 3 factors with 2 levels each, then we have 2^3 possible experiences opportunities (therefore 8). But the number 8 matches exactly with the number of coefficients that we also have in the linear model with interactions of a response with three variables:

$$y \cong a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_{12}x_1x_2 + a_{13}x_1x_3 + a_{23}x_2x_3 + a_{123}x_1x_2x_3 \tag{16.7.400}$$

which also corresponds also only to the linear terms and condensed form of the Maclaurin series expansion of a function f of three variables for recall.

And so on ... for n factors with two levels each. This is why the linear full factorial designs 2^n are traditionally the most used because they are mathematically intuitive and simple to prove and develop as many results simplifies at the opposite of the general factorial designs!

16.7.4.1.1 Replicated Full Factorial Designs

Replicated full factorial designs are important as they give (among others that we will see during a companion example for the study of general factorial designs[26]) the ability to calculate the confidence intervals of the effects.

As a companion example let us consider the 2^3 factorial design below replicated twice:

Trial N°	x_1	x_2	x_3	y_{i1}	y_{i2}	$\bar{y}_i$
1	-1	-1	-1	59	61	60
2	$+1$	-1	-1	74	70	72
3	-1	$+1$	-1	50	58	54
4	$+1$	$+1$	-1	69	67	68
5	-1	-1	$+1$	50	54	52
6	$+1$	-1	$+1$	81	85	83
7	-1	$+1$	$+1$	46	44	45
8	$+1$	$+1$	$+1$	79	81	80

Or in more complete form:

Trial N°	I	x_1	x_2	x_3	x_{12}	x_{13}	x_{23}	x_{123}	$\bar{y}_i$
1	$+$	$-$	$-$	$-$	$+$	$+$	$+$	$-$	60
2	$+$	$+$	$-$	$-$	$-$	$-$	$+$	$+$	72
3	$+$	$-$	$+$	$-$	$-$	$+$	$-$	$+$	54
4	$+$	$+$	$+$	$-$	$+$	$-$	$-$	$-$	68
5	$+$	$-$	$-$	$+$	$+$	$-$	$-$	$+$	52
6	$+$	$+$	$-$	$+$	$-$	$+$	$-$	$-$	83
7	$+$	$-$	$+$	$+$	$-$	$-$	$+$	$-$	45
8	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	80

We will assume to following underlying model:

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_{12}x_1x_2 + a_{13}x_1x_3 + a_{23}x_2x_3 + a_{123}x_1x_2x_2 \tag{16.7.401}$$

The global average is:

$$\frac{60+72+54+68+52+83+45+80}{8} = 64.25 \tag{16.7.402}$$

The average of all responses at high level of x_1 is equal to:

$$y_{x_1,+} = \frac{74+70+69+67+81+85+79+81}{8} = 75.75 \tag{16.7.403}$$

the average of all responses at bottom level of x_1 is equal to:

$$y_{x_1,-} = \frac{59+61+50+58+50+54+46+44}{8} = 52.75 \tag{16.7.404}$$

---

[26] especially the calculation of an ANOVA

So the overall effect of the x_1 is equal to:

$$E_{Gx_1} = y_{x_1,+} - y_{x_1,-} = 75.72 - 52.75 = 23.00 \tag{16.7.405}$$

Or equivalently using in one shot only all the y_{i1} and y_{i2}:

$$E_{Gx_1} = \frac{-59-61+74+70-50+\ldots+85-46-44+79+81}{8} = 23.00 \tag{16.7.406}$$

or using only the averages $\bar{y}_i$:

$$E_{Gx_1} = \frac{-60+72-54+68-52+83-45+80}{4} = 23.00 \tag{16.7.407}$$

And so on:

$$\begin{aligned}E_{Gx_{12}} &= \frac{60-72-54+68+52-83-45+80}{4} = 1.5\\ E_{Gx_{123}} &= \frac{-60+72+54-68+52-83-45+80}{4} = 0.5\\ &\ldots\end{aligned} \tag{16.7.408}$$

Finally we get:

$$\begin{aligned}\text{Average} &= 64.25 & E_{Gx_{12}} &= 1.5\\ E_{Gx_1} &= 23.00 & E_{Gx_{13}} &= 10.00\\ E_{Gx_2} &= -5.00 & E_{Gx_{23}} &= 0.00\\ E_{Gx_3} &= 1.50 & E_{Gx_{123}} &= 0.50\end{aligned} \tag{16.7.409}$$

Now the question that interest us here are:

- Which of these effects are important?
- Which of these effects are distinguishable from the noise in the experimental environment?

Let's say we run experiment numerous times. Sample effects, E_{Gx_i} will be supposed to be Normally distributed. We can obviously do a Normal probability plot (Henry plot) of the measurements but here we will focus rather on the calculation of a confidence interval as it is more easy to communicate in corporations.

So our experiment is replicated, we calculate a sample variance for each run r:

$$\hat{\sigma}_{y_r}^2 = \frac{1}{n-1}\sum_{i=1}^{2}(y_{ri}-\bar{y}_r)^2 \tag{16.7.410}$$

Then we get the following table:

Trial N°	y_{r1}	y_{r2}	$\bar{y}_r$	$\hat{\sigma}_{y_r}^2$
1	59	61	60	2
2	74	70	72	8
3	50	58	54	32
4	69	67	68	2
5	50	54	52	8
6	81	85	83	8
7	46	44	45	2
8	79	81	80	2

Remark

What Minitab denotes by the T-value is the ratio:

$$T\text{-value} = \frac{a_i}{\text{SE}} \tag{16.7.411}$$

Notice that if $n=1$, i.e. no replication data per run... then we cannot calculate the $\hat{\sigma}_{y_r}^2$ as there is a division by 0...!

We now assume that the standard deviation is constant for all runs (very strong assumption!) and given by the pooled sample variance:

$$\hat{\sigma}_y^2 = \frac{\sum\limits_{r=1}^{8} \hat{\sigma}_{y_r}^2}{8} = \frac{2+8+32+2+8+8+2+2}{8} = 8 \tag{16.7.412}$$

So obviously we can generalize:

$$\hat{\sigma}_y^2 = \frac{\sum\limits_{r=1}^{2^k} \hat{\sigma}_{y_r}^2}{2^k} \tag{16.7.413}$$

The reader must keep in mind that the null hypothesis we would like to reject is:

$$H_0: \mu_{E_{Gx_1}} = \mu_{E_{Gx_2}} = \ldots = \mu_{E_{Gx_i}} = 0 \tag{16.7.414}$$

and under that null hypothesis they would follow a Normal distribution (so if we reject that one we will be quite happy as the effects are significantly non-zero).

We may, based on each calculated effect, develop a confidence interval for the true mean effect:

$$\boxed{E_{Gx_i} \pm T_{\alpha/2,2^k} \hat{\sigma}_{\bar{y}}} \tag{16.7.415}$$

and we know (see section Statistics page 611) that this is written explicitly:

$$E_{Gx_i} \pm T_{\alpha/2,2^k} \frac{\hat{\sigma}_y}{\sqrt{N}} \tag{16.7.416}$$

Therefore in our example we have for the average:

$$\text{Average} \pm T_{\alpha/2,2^3}\frac{\hat{\sigma}_y}{\sqrt{N}} \tag{16.7.417}$$

Numerically:

$$64.25 \pm T_{5\%/2,8}\frac{\sqrt{8}}{\sqrt{16}} \tag{16.7.418}$$

That is to say:

$$64.25 \pm 2.306 \cdot 0.707 \tag{16.7.419}$$

shortly:

$$64.25 \pm 1.63 \tag{16.7.420}$$

Since 0 is not in the confidence interval, we are happy to reject H_0 that $\mu_{\text{Average}} = 0$.

Now let us calculate the confidence interval for the other effects. We know that we calculate effects generally by:

$$E_{Gx_i} = \bar{y}_{i,+} - \bar{y}_{i,-} \tag{16.7.421}$$

Each bases on $N/2$ observations ($N/2 = 8$ for our example). We have obviously:

$$\text{V}(E_{Gx_i}) = \text{V}(\bar{y}_{i,+}) - \text{V}(\bar{y}_{i,-}) \tag{16.7.422}$$

with obviously:

$$\text{V}(\bar{y}_{i,+}) = \text{V}\left(\frac{\sum_{i=1}^{N/r} y_{i,+}}{N/r}\right) \tag{16.7.423}$$

where r is the number of replications. Therefore:

$$\text{V}(\bar{y}_{i,+}) = \frac{r^2}{N^2}\text{V}\left(\sum_{i=1}^{N/r} y_{i,+}\right) \tag{16.7.424}$$

Bus as we assume for all runs the standard deviation:

$$\text{V}\left(\sum_{i=1}^{N/r} y_{i,+}\right) = \frac{N}{r}\sigma_y^2 \tag{16.7.425}$$

Therefore:

$$\text{V}(\bar{y}_{i,+}) = \frac{r^2}{N^2}\frac{N}{r}\sigma_y^2 = \frac{r}{N}\sigma_y^2 \tag{16.7.426}$$

Hence:

$$\text{V}(E_{Gx_i}) = \text{V}(\bar{y}_{i,+}) - \text{V}(\bar{y}_{i,-}) = \frac{r}{N}\sigma_y^2 + \frac{r}{N}\sigma_y^2 = \frac{2r}{N}\sigma_y^2 \tag{16.7.427}$$

So finally:

$$\boxed{\sigma_{\text{eff}}^2 = \text{V}(E_{Gx_i}) = \frac{2r}{N}\sigma_y^2} \tag{16.7.428}$$

sometimes named the "standard error of an effect".

For our example:

$$\hat{\sigma}_{\text{eff}}^2 = \frac{2 \cdot 2}{16} \cdot 8 = 2 \tag{16.7.429}$$

Therefore:

$$\hat{\sigma}_{\text{eff}} \cong 1.414 \tag{16.7.430}$$

Now we get easily for each effect the following confidence interval:

$$\boxed{E_{Gx_i} \pm T_{\alpha/2,N/r}\hat{\sigma}_{\text{eff}}} \tag{16.7.431}$$

So in our example:

$$E_{Gx_i} \pm T_{5\%/2,16/2}\hat{\sigma}_{\text{eff}} \tag{16.7.432}$$

Numerically:

$$E_{Gx_i} \pm 2.306 \cdot 1.414 \tag{16.7.433}$$

Hence:

$$E_{Gx_i} \pm 3.26 \tag{16.7.434}$$

So finally:

$$\begin{aligned} \text{Average} &= 64.25 \pm 1.63 & E_{Gx_{12}} &= 1.5 \pm 3.26 \\ E_{Gx_1} &= 23.00 \pm 3.26 & E_{Gx_{13}} &= 10.00 \pm 3.26 \\ E_{Gx_2} &= -5.00 \pm 3.26 & E_{Gx_{23}} &= 0.00 \pm 3.26 \\ E_{Gx_3} &= 1.50 \pm 3.26 & E_{Gx_{123}} &= 0.50 \pm 3.26 \end{aligned} \tag{16.7.435}$$

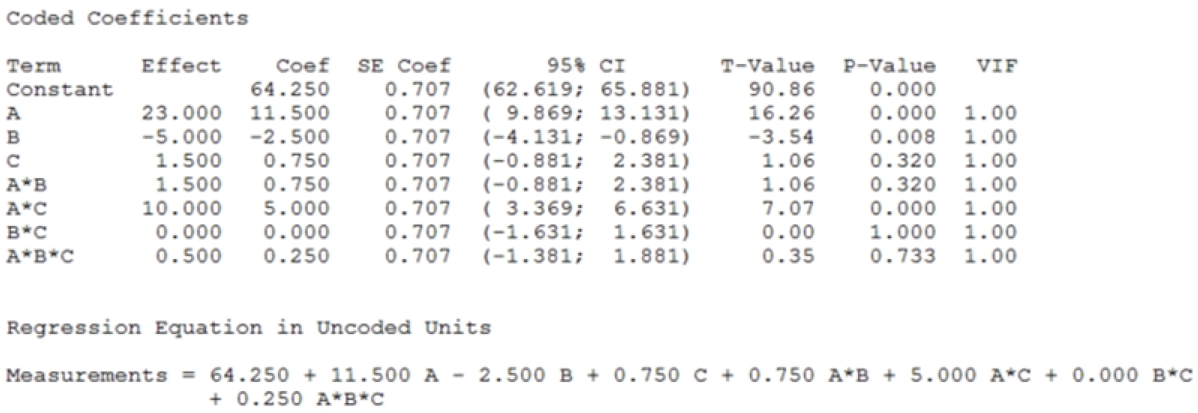
So 0 does not line on the confidence interval for E_{Gx_1}, E_{Gx_2} and $E_{Gx_{13}}$. We now know that only a_0, a_1, a_2 and a_{13} are not noise and therefore the fitted model becomes:

$$\begin{aligned} y &= a_0 + a_1x_1 + a_2x_2 + a_{13}x_1x_3 \\ &= \text{Average} + \left(\frac{E_{Gx_1}}{2}\right)x_1 + \left(\frac{E_{Gx_2}}{2}\right)x_2 + \left(\frac{E_{Gx_3}}{2}\right)x_3 + \left(\frac{E_{Gx_{13}}}{2}\right)x_1x_3 \end{aligned} \tag{16.7.436}$$

That is to say:

$$y = 64.25 + 11.5x_1 - 2.5x_2 + 5x_1x_3 \tag{16.7.437}$$

For comparison we can see the special way on how Minitab 17.1.3 show these same results:



```
Coded Coefficients

Term        Effect     Coef  SE Coef        95% CI          T-Value  P-Value   VIF
Constant             64.250    0.707  (62.619; 65.881)       90.86    0.000
A           23.000   11.500    0.707  ( 9.869; 13.131)       16.26    0.000  1.00
B           -5.000   -2.500    0.707  (-4.131; -0.869)       -3.54    0.008  1.00
C            1.500    0.750    0.707  (-0.881;  2.381)        1.06    0.320  1.00
A*B          1.500    0.750    0.707  (-0.881;  2.381)        1.06    0.320  1.00
A*C         10.000    5.000    0.707  ( 3.369;  6.631)        7.07    0.000  1.00
B*C          0.000    0.000    0.707  (-1.631;  1.631)        0.00    1.000  1.00
A*B*C        0.500    0.250    0.707  (-1.381;  1.881)        0.35    0.733  1.00


Regression Equation in Uncoded Units

Measurements = 64.250 + 11.500 A - 2.500 B + 0.750 C + 0.750 A*B + 5.000 A*C + 0.000 B*C
               + 0.250 A*B*C
```

Figure 16.240: Replicated factorial design effects C.I. with Minitab 17.1.3

Remark

What Minitab denote by the T-value is the ratio:

$$T\text{-value}=\frac{a_i}{\text{SE}} \qquad (16.7.438)$$

That is to say the value that should take the T distribution for not rejecting the null hypothesis.

So for example for the factor A we get:

$$T_A\text{-value}=\frac{11.50}{0.707}=16.26 \qquad (16.7.439)$$

and the column p-value is as always (see section Statistics page 948) the corresponding percentile of the T-value.

And finally for those who are interested to see the Henry plot of the effect:

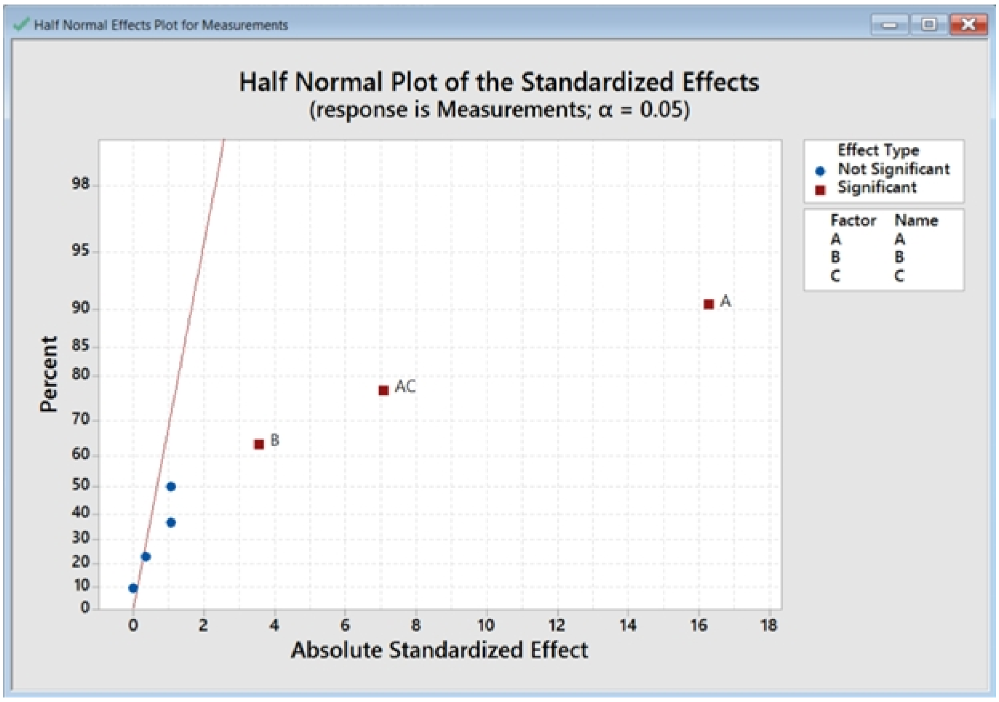

Half Normal Effects Plot for Measurements
Half Normal Plot of the Standardized Effects
(response is Measurements; α = 0.05)
Effect Type
Not Significant
Significant
Factor Name
A A
B B
C C
Percent
98
95
90
85
80
70
60
50
40
30
20
10
0
A
AC
B
0
2
4
6
8
10
12
14
16
18
Absolute Standardized Effect

Figure 16.241: Henry plot of the replicated factorial design with Minitab 17.1.3

16.7.4.1.2 Plackett-Burman Designs

It is important to notice that all these full linear designs approximated to the second order are in matrix form orthogonal square matrices M_n and therefore of course can be inverted (see section Linear Algebra page 1954)!

Also the previous matrices do not satisfy the following relation as seen in the section of Linear Algebra (scalar product of basis-columns with themselves):

$$X^T X = \mathbb{1} \tag{16.7.440}$$

where for recall $\mathbb{1}$ is a matrix with only 1 in the diagonal and 0 everywhere else, but have the particularity for all full experimental design to satisfy the relation:

$$X^T X = n\mathbb{1} \tag{16.7.441}$$

So unlike orthogonal matrices, which by definition have all the columns (or rows) that form an orthonormal base (unit norm), the experimental designs matrices have the particularity of not having the norm orthogonal basis vectors equal to the unit.

Definition (#16.7.737): Thus, we define the matrix whose components are all $+1$ or -1 and satisfying the previous relation:

$$H^T H = n\mathbb{1} \tag{16.7.442}$$

as a "Hadamard matrix" denoted sometimes H. The Hadamard matrices are the ± 1 matrices which maximize the determinant. It has some relevance to optimal design theory in statistics where minimizing the determinant of the information matrix gives the most accurate estimates of unknown parameters (in a certain sense). The information matrix is the inverse of a matrix obtained from the design, whose determinant thus has to be maximized (see page 8865).

Now the $\det(H_n)$ is equal to the volume of the n-dimensional parallelepiped spanned by the orthogonal rows of H_n (see section Linear Algebra page 1961). By construction, each row has an Euclidean length at most, so that:

$$\det(H_n) \leq n^{n/2} \tag{16.7.443}$$

and remember that this equality holds if and only if every entry of H_n is ± 1 and the rows of H are orthogonal. A matrix attaining the bound:

$$\det(H_n) = n^{n/2} \tag{16.7.444}$$

is a Hadamard matrix (we will prove further below that this bound is attained when n is a multiple of 4)! Put in another way, a $(+1,-1)$-matrix is Hadamard if the inner product of two distinct rows or columns is 0 and the inner product of a row or column with itself is n.

A few examples of Hadamard matrices are:

$$H_2=\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}\qquad H_4=\begin{bmatrix}1 & 1 & 1 & 1\\ 1 & 1 & -1 & -1\\ 1 & -1 & 1 & -1\\ 1 & -1 & -1 & 1\end{bmatrix}\qquad H_8=\begin{bmatrix}1 & 1 & 1 & 1 & 1 & 1 & 1 & 1\\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1\\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1\\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1\\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1\\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1\\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1\\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1\end{bmatrix}\tag{16.7.445}$$

It is quite apparent that if the rows and columns of a Hadamard matrix are permuted, the matrix remains Hadamard. It is also true that if any row or column is multiplied by -1, the Hadamard property is retained.

Theorem 16.3. Hadamard matrices have the property to exist only for orders 1, 2, 4, 8, 12, 16, 20, 24, ... That is to say only for the multiple for the orders that are multiple of 4 or written differently:

$$n\equiv 0\;(\text{modulo }4)\tag{16.7.446}$$

(if we omit the case 1 and 2 that are obvious).

Proof 16.3.1. Knowing that the case of 1 and 2 are trivial and that the odd case must immediately intuitively be removed (try and you will see almost very quickly...), we will do the proof for $n\geq 4$.

Since all the columns are mandatory orthogonal (so that the matrix is invertible and therefore the system uniquely solvable) and from the fact that the rows and columns of an Hadamard matrix are permuted or that if any row or column is multiplied by -1 the Hadamard property is retained, it is therefore always possible to arrange to have the first row and first column of a Hadamard matrix contain only $+1$ entries. An Hadamard matrix in this form is said to be "normalized":

$$\begin{gathered}+x+y+z+w=n\\ +x+y-z-w=0\\ +x-y+z-w=0\\ +x-y-z+w=0\end{gathered}\tag{16.7.447}$$

The reader must obviously consider that in the above system we can consider two case in facts:

1. It is a system that represents indeed an order 4 matrix, so x, y, z and w have to be replaced in readers mind by the value 1:

$$\begin{gathered}+1+1+1+1=n\\ +1+1-1-1=0\\ +1-1+1-1=0\\ +1-1-1+1=0\end{gathered}\tag{16.7.448}$$

2. It is a system that can represents only the 4 first rows of a matrix of order n (as all the other rows are equal to zero based on the same principle... we then have the habit to not represent them), so x must therefore be in the reader mind replace by a given number of 1 ($x=1\ldots1$), same for y, z and w. More explicitly (for example for $n=8$):

$$\begin{gathered}\underbrace{(1+1)}_{x}+\underbrace{(1+1)}_{y}+\underbrace{(1+1)}_{z}+\underbrace{(1+1)}_{w}=n\\\underbrace{(1+1)}_{x}+\underbrace{(1+1)}_{y}-\underbrace{(1+1)}_{z}-\underbrace{(1+1)}_{w}=0\\\underbrace{(1+1)}_{x}-\underbrace{(1+1)}_{y}+\underbrace{(1+1)}_{z}-\underbrace{(1+1)}_{w}=0\\\underbrace{(1+1)}_{x}-\underbrace{(1+1)}_{y}-\underbrace{(1+1)}_{z}+\underbrace{(1+1)}_{w}=0\end{gathered} \tag{16.7.449}$$

As we can always normalize a Hadamard matrix to get such as system then the Hadamard properties still remain valid.

Solving the system gives:

$$4x=4y=4z=4w=n \tag{16.7.450}$$

therefore n must be divisible by 4 for $n\geq 4$ if we want the number of terms in x, y, z, w to be an integer!

The result can seems not obvious but if the reader has the time... he can try for example to construct a Hadamard matrix of order $n=6$ and he will quickly see why it is not possible.

■ Q.E.D.

It then follows trivially the following relation (with a very abusive notation because it omits the notation of the unit matrix):

$$A=nXY=(X^TX)XY \tag{16.7.451}$$

However, we will see a little further below in our study of fractional factorial designs that Hadamard matrices of order 1, 4, 8, 16, 32, ..., 2^n are almost never used under the expression "Hadamard matrices" as they are confused with the fractional factorial designs.

By cons, it is interesting to observe that we have the potential existing design of experiments with 12, 20, 24, 28, ... trials (in other words: the number of runs is a multiple of 4 but not a power of 2) that have some interest when the number of factors is greater than or equal to 4 (some software like JMP - in year 12012 according to holocene calendar - however don't propose these designs if the number of factors is less than 5). These designs are named "Plackett-Burman Design of Experiments" and sometimes "irregular Plackett-Burman Design of Experiments" or more rarely "non-geometric Plackett-Burman Design of Experiments" (a software like Minitab propose however arbitrarily - and indicate it explicitly in the software Help - a Plackett-Burman design of order 32 and Plackett-Burman designs for 2 or 3 factors).

Remark

Another way to see the interest of Plackett-Burman designs is to considerate that as factorial designs have 2^k runs we then have 8, 16, 32, 64, 128, etc. number of runs. However, when you look at these numbers there is a pretty big gap between 16 and 32, 32 to 64, etc. We sometimes need other alternative designs besides these with a different number of observations.

Then a class of designs that allows us to create experiments with some number between these factorial designs are the Plackett-Burman designs as they give:

$$n = 12, [16], 20, 24, 28, [32], 36, 40, 44, 48, \ldots \tag{16.7.452}$$

... any number which is divisible by four!

Robin L. Plackett and J. P. Burman have try with algorithms and by trial and error the expression of matrices corresponding experimental designs that carry their name and which therefore contain 12, 20, 24, 28, ... trials. They have proposed a very useful tip so that practitioners can create these experimental designs without using a software.

Let us start with an example using a Hadamard matrix of order 12 (without focus for the moment on the number of factors that we use). The Plackett-Burman tables for the special 12 order give only the first row (in fact... it is a column in the matrix):

$$++-+++---+- \tag{16.7.453}$$

Then we build the following table of experiment by following the algorithm proposed by Plackett and Burman (software use the same as far as we know):

1. First step, we put the first line as a column in a 12×12 table:

+											
+											
-											
+											
+											
+											
-											
-											
-											
+											
-											

2. Second step: we deduce the second column from the first column by shifting the signs a step to the bottom, the last "$-$" sign being ascended to the top of the second column (it is therefore equivalent to a circular permutation):

+	-										
+	+										
-	+										
+	-										
+	+										
+	+										
-	+										
-	-										
-	-										
+	-										
-	+										

and so on until the 11th column:

+	-	+	-	-	-	+	+	+	-	+	
+	+	-	+	+	-	-	-	+	+	-	
-	+	+	-	+	+	-	-	+	+	+	
+	-	+	+	-	+	-	-	-	+	+	
+	+	-	+	+	-	+	-	-	-	+	
+	+	+	-	+	+	-	+	-	-	-	
-	+	+	+	-	+	+	-	+	-	-	
-	-	+	+	+	-	+	+	-	+	-	
-	-	-	+	+	+	-	+	+	-	+	
+	-	-	-	+	+	+	-	+	+	-	
-	+	-	-	-	+	+	+	-	+	+	

3. Last step, we add a row and a column of signs "$-$":

+	-	+	-	-	-	+	+	+	-	+	-
+	+	-	+	+	-	-	-	+	+	-	-
-	+	+	-	+	+	-	-	+	+	+	-
+	-	+	+	-	+	-	-	-	+	+	-
+	+	-	+	+	-	+	-	-	-	+	-
+	+	+	-	+	+	-	+	-	-	-	-
-	+	+	+	-	+	+	-	+	-	-	-
-	-	+	+	+	-	+	+	-	+	-	-
-	-	-	+	+	+	-	+	+	-	+	-
+	-	-	-	+	+	+	-	+	+	-	-
-	+	-	-	-	+	+	+	-	+	+	-
-	-	-	-	-	-	-	-	-	-	-	-

The reader can easily verify that each row or column taken in pairs are orthogonal.

Now arises the question of whether we choose to associate for example this design of experience to a 4 factors design (meaning reduced centered design of course!):

$$\begin{aligned} y \cong & a_0+a_1x_1+a_2x_2+a_3x_3+a_4x_4 \\ & +a_{12}x_1x_2+a_{13}x_1x_3+a_{14}x_1x_4+a_{23}x_2x_3+a_{24}x_2x_3+a_{34}x_3x_4 \\ & +a_{123}x_1x_2x_3+a_{124}x_1x_2x_4+a_{134}x_1x_3x_4+a_{234}x_2x_3x_4 \\ & +a_{1234}x_1x_2x_3x_4 \end{aligned} \quad (16.7.454)$$

to reduce the number of tests from 16 to 12, which column should we associate to what? Or going further: for how many 2-level factors can we associated this type of design having 12 trials???:

1. The first one is to say that the Plackett-Burman designs of order n as they ignore all interactions should only be used for the study of main effects of $n-1$ factors. Thus, a Plackett-Burman design of order 12 will be reserved for a study of 11 factors and only of their main effects (purely additive linear model).
2. The second one is to say that the Plackett-Burman design contains a concept that we will see further and who's named "aliases". Then the designs are commonly named "Plackett-Burman design of resolution III". Therefore, we have to use this type of designs when we are ready to consider initially that the interactions with two factors (second order interactions) are not significant.

The complexity of the alias with the Plackett-Burman design makes that in practice they are rather used in the first religion release... by beginners and in the second by consultants.

Remark

A software like Minitab 15.1.1 does not display all the aliases used in the implementation of Plackett-Burman designs.

To conclude this section, let us summarize with simple observation:

Plan	Factors	Interactions	Sum
2^2	2	1	3
2^3	3	4	7
2^4	4	11	15
2^5	5	26	31
2^6	6	57	63
2^7	7	120	127
...	...	...	...

Table 16.32: **Types of plans and factors & interactions**

So using full factorial designs, the user is sure to have the optimal experimental procedure since these plans are based on Hadamard matrices and we have proved that we could not do better.

16.7.4.1.3 Fractional Factorial Designs

In practice, full factorial designs can only be used on systems with very few factors, or when each test takes very little time. When n is greater than or equal 3 then experiences costs can quickly become expensive.

The smallest case where it is interesting to optimize the number of tests is the one consisting of 3 factors with 2 levels each. We then have the following equation and experience table:

$$y \cong a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_{12}x_1x_2 + a_{13}x_1x_3 + a_{23}x_2x_2 + a_{123}x_1x_2x_3 \qquad (16.7.455)$$

Trial N°	Factor 1	Factor 2	Factor 3	Answer
1	−	−	−	y_1
2	+	−	−	y_2
3	−	+	−	y_3
4	+	+	−	y_5
5	−	−	+	y_5
6	+	−	+	y_6
7	−	+	+	y_7
8	+	+	+	y_8

Table 16.33: **Experimental Full Factorial (2 factor) matrix**

Either as full experiment table:

Trial N°	Rest	F1	F2	F3	F12	F13	F23	F123	Answers
1	+	−	−	−	+	+	+	−	y_1
2	+	+	−	−	−	−	+	+	y_2
3	+	−	+	−	−	+	−	+	y_3
4	+	+	+	−	+	−	−	−	y_4
5	+	−	−	+	+	−	−	+	y_5
6	+	+	−	+	−	+	−	−	y_6
7	+	−	+	+	−	−	+	−	y_7
8	+	+	+	+	+	+	+	+	y_8

Table 16.34: **Full Design of experiment with 3 factors and interactions under Yates form**

or in matrix form:

$$\begin{bmatrix} +1 & -1 & -1 & -1 & +1 & +1 & +1 & -1 \\ +1 & +1 & -1 & -1 & -1 & -1 & +1 & +1 \\ +1 & -1 & +1 & -1 & -1 & +1 & -1 & +1 \\ +1 & +1 & +1 & -1 & +1 & -1 & -1 & -1 \\ +1 & -1 & -1 & +1 & +1 & -1 & -1 & +1 \\ +1 & +1 & -1 & +1 & -1 & +1 & -1 & -1 \\ +1 & -1 & +1 & +1 & -1 & -1 & +1 & -1 \\ +1 & +1 & +1 & +1 & +1 & +1 & +1 & +1 \end{bmatrix} \qquad (16.7.456)$$

where once again it is easy to control that all columns are orthogonal and balanced (same number of $+$ or $-$ in each column, or in other words the sum of each column - excepted the first one - is equal to zero) and that the matrix is a Hadamard matrix. Therefore the full factorial design for 3 factors needs 8 trials.

We can also represent this in the following form:

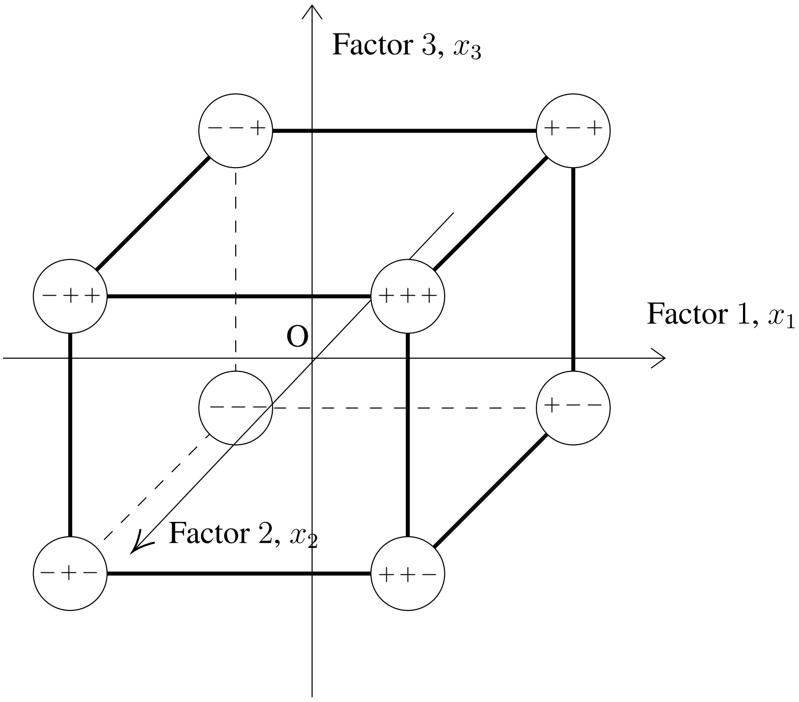
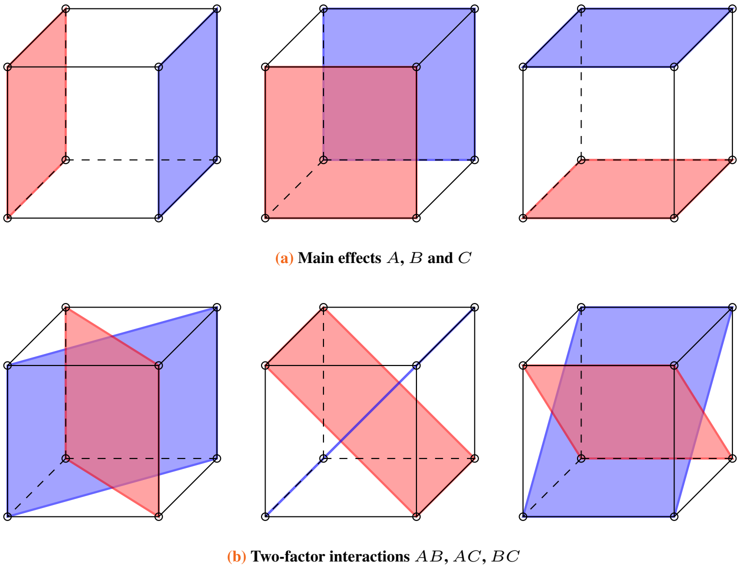


Figure 16.242: Traditional representation of a full factorial design with 3 factors

and the corresponding 2-way interactions in the following form (typical of ANOVA studies):



(a) Main effects A, B and C

(b) Two-factor interactions AB, AC, BC

Figure 16.243: Geometric representation of contrasts (in each case, high levels are highlighted in blue, low levels in red)

A "reduced design", more commonly named "fractional factorial designs (FF)" or "screening plans" (following the standard ISO 3534-3:1999), that consist in selecting some combinations, has therefore been proposed. They naturally gives the opportunity to reduce the experimental costs, but also decrease the information available on the system! We must therefore be sure of the adequate choice relatively to the model to identify!

A first elementary method is to do the hypothesis that there are no interactions. Therefore our function is reduce to a purely additive mode:

$$y \cong a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 \tag{16.7.457}$$

and to resolve thy system, 4 trials are enough. We therefore go from a 8 trial design to a 4 trial design just by supposing that there are no interactions and therefore we fall back on a simple target design.

To reduce the experimental costs by keeping the interactions implicitly, we can play with mathematics stuff. First let us take the actual problem with a full factorial design of 3 factors into explicit form:

$$\begin{aligned}
+a_0 - a_1 - a_2 - a_3 + a_{12} + a_{13} + a_{23} - a_{123} &= y_1\\
+a_0 + a_1 - a_2 - a_3 - a_{12} - a_{13} + a_{23} + a_{123} &= y_2\\
+a_0 - a_1 + a_2 - a_3 - a_{12} + a_{13} - a_{23} + a_{123} &= y_3\\
+a_0 + a_1 + a_2 - a_3 + a_{12} - a_{13} - a_{23} - a_{123} &= y_4\\
+a_0 - a_1 - a_2 + a_3 + a_{12} - a_{13} - a_{23} + a_{123} &= y_5\\
+a_0 + a_1 - a_2 + a_3 - a_{12} + a_{13} - a_{23} - a_{123} &= y_6\\
+a_0 - a_1 + a_2 + a_3 - a_{12} - a_{13} + a_{23} - a_{123} &= y_7\\
+a_0 + a_1 + a_2 + a_3 + a_{12} + a_{13} + a_{23} + a_{123} &= y_8
\end{aligned} \tag{16.7.458}$$

Can we reduce the writing of this system to reduce the trials to do? The answer is: Yes! But in counterpart we will lost the measurement of pure effects (we therefore sometimes speak of "confusion" or of "confounding").

The inferior nearest writing is the Hadamard matrix of order 4. This means obviously that we need to conserve 4 rows on the 8 and that these 4 rows must remain orthogonal, balanced and must satisfy the relation:

$$X^T X = n\mathbb{1} \tag{16.7.459}$$

The idea, named the "Box and Hunter method" and that works only for design with two level factors, is in a first time to merge together the influential factors (in indices) such that (the developments are similar for any n):

$$\begin{aligned}
&+(a_0 - a_{123}) + (-a_1 + a_{23}) + (-a_2 + a_{13}) + (-a_3 + a_{12}) = y_1\\
&+(a_0 + a_{123}) + (+a_1 + a_{23}) + (-a_2 - a_{13}) + (-a_3 - a_{12}) = y_2\\
&+(a_0 + a_{123}) + (-a_1 - a_{23}) + (+a_2 + a_{13}) + (-a_3 - a_{12}) = y_3\\
&+(a_0 - a_{123}) + (+a_1 - a_{23}) + (+a_2 - a_{13}) + (-a_3 + a_{12}) = y_4\\
&+(a_0 + a_{123}) + (-a_1 - a_{23}) + (-a_2 - a_{13}) + (+a_3 + a_{12}) = y_5\\
&+(a_0 - a_{123}) + (+a_1 - a_{23}) + (-a_2 + a_{13}) + (+a_3 - a_{12}) = y_6\\
&+(a_0 - a_{123}) + (-a_1 - a_{23}) + (+a_2 - a_{13}) + (+a_3 - a_{12}) = y_7\\
&+(a_0 + a_{123}) + (+a_1 + a_{23}) + (+a_2 + a_{13}) + (+a_3 + a_{12}) = y_8
\end{aligned} \tag{16.7.460}$$

The grouping choice is also made so that the interaction terms that are supposed to be negligible by the centered reduce normalisation are merged ("confused") with a coefficient of a main factor supposed as not negligible!! Therefore it is also logic that in each grouping we never found indices with a number of a main factor (as merging two main factors that are not negligible will be very bad!).

We then say that we have an "alias structure" (the standard ISO 3534-3:1999 name this a "concomitance" when it is the practitioner that force the choice of the regroupement and "alias" if the regroupement choice is due to the nature of the experience) of the type:

```
0+123;1+23;2+13;3+12
```

and if we write this in the same way as many statistical software, this give (this is exactly the alias given by a software like Minitab 15.1.1):

```
I+ABC;A+BC;B+AC;C+AB
```

Also some practitioner write this (as `I` must always have a positive sign):

```
I=ABC;A+BC;B+AC;C+AB
```

Let us write the previous system as following:

$$\begin{aligned}
(a_0-a_{123})-(a_1-a_{23})-(a_2-a_{13})-(a_3-a_{12})&=y_1\\
(a_0+a_{123})+(a_1+a_{23})-(a_2+a_{13})-(a_3+a_{12})&=y_2\\
(a_0+a_{123})-(a_1+a_{23})+(a_2+a_{13})-(a_3+a_{12})&=y_3\\
(a_0-a_{123})+(a_1-a_{23})+(a_2-a_{13})-(a_3-a_{12})&=y_4\\
(a_0+a_{123})-(a_1+a_{23})-(a_2+a_{13})+(a_3+a_{12})&=y_5\\
(a_0-a_{123})+(a_1-a_{23})-(a_2-a_{13})+(a_3-a_{12})&=y_6\\
(a_0-a_{123})-(a_1-a_{23})+(a_2-a_{13})+(a_3-a_{12})&=y_7\\
(a_0+a_{123})+(a_1+a_{23})+(a_2+a_{13})+(a_3+a_{12})&=y_8
\end{aligned} \tag{16.7.461}$$

Let us change our notations:

$$\begin{aligned}
l_0^- - l_1^- - l_2^- - l_3^- &= y_1\\
l_0^+ + l_1^+ - l_2^+ - l_3^+ &= y_2\\
l_0^+ - l_1^+ + l_2^+ - l_3^+ &= y_3\\
l_0^- + l_1^- + l_2^- - l_3^- &= y_4\\
l_0^+ - l_1^+ - l_2^+ + l_3^+ &= y_5\\
l_0^- + l_1^- - l_2^- + l_3^- &= y_6\\
l_0^- - l_1^- + l_2^- + l_3^- &= y_7\\
l_0^+ + l_1^+ + l_2^+ + l_3^+ &= y_8
\end{aligned} \tag{16.7.462}$$

Naturally, if we consider this new notation as new variables, this unique system can now be splitted into two sub-systems (the grouping being named "contrasts" in this field but this is not corresponding to the vocabulary of the standard ISO 3534-3) to be solvable:

$$\begin{aligned}
l_0^- - l_1^- - l_2^- - l_3^- &= y_1\\
l_0^- + l_1^- + l_2^- - l_3^- &= y_4\\
l_0^- + l_1^- - l_2^- + l_3^- &= y_6\\
l_0^- - l_1^- + l_2^- + l_3^- &= y_7
\end{aligned}
\quad\text{and}\quad
\begin{aligned}
l_0^+ + l_1^+ - l_2^+ - l_3^+ &= y_2\\
l_0^+ - l_1^+ + l_2^+ - l_3^+ &= y_3\\
l_0^+ - l_1^+ - l_2^+ + l_3^+ &= y_5\\
l_0^+ + l_1^+ + l_2^+ + l_3^+ &= y_8
\end{aligned} \tag{16.7.463}$$

which gives the possibility to divide the number of trials by 2 relatively to the initial full factorial design!!!!

By solving these two systems, we say that the interactions are "aliased" (also said in "confusion") with the pure effects in negative or in positive (in the present case, some main effects are aliased with two factors interactions).

It is afterwards the tradition to keep only the positive aliased system:

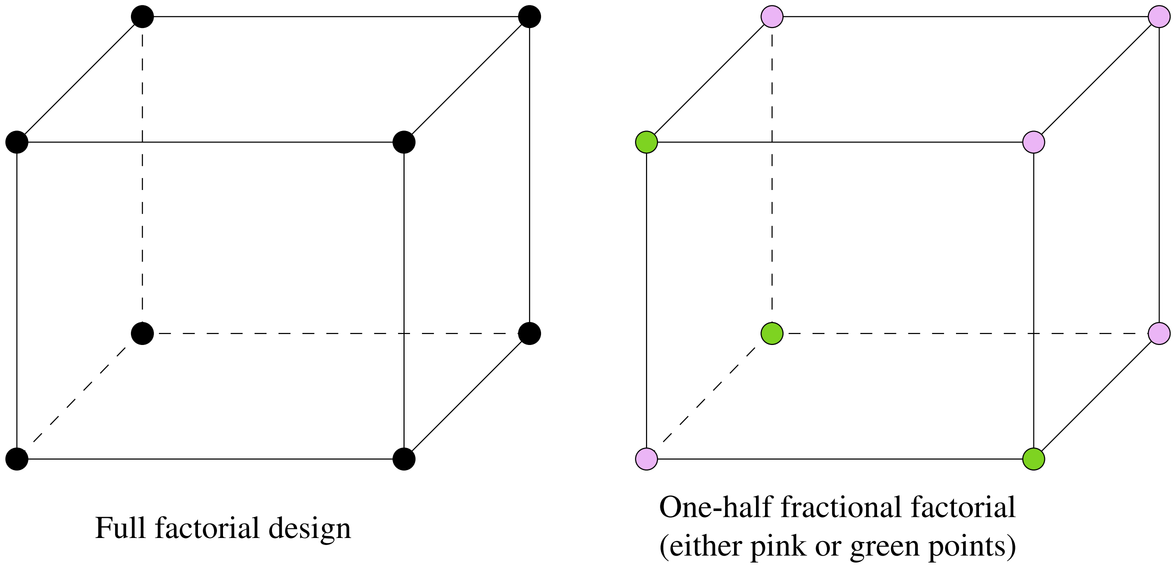
$$\begin{matrix} l_0^+ + l_1^+ - l_2^+ - l_3^+ = y_2 \\ l_0^+ - l_1^+ + l_2^+ - l_3^+ = y_3 \\ l_0^+ - l_1^+ - l_2^+ + l_3^+ = y_5 \\ l_0^+ + l_1^+ + l_2^+ + l_3^+ = y_8 \end{matrix} \tag{16.7.464}$$

because if the interactions are zero, we fall back on the same experience of the matrix of a full 2^2 factorial design! The approach therefore leads to select only the runs 2, 3, 5 and 8, which give the possibility to divide the number of tests by two compared to a full factorial design. Thus, the fractional factorial design of an experiment with three factors may be reduced to 4 runs with this method. The fractional factorial design above will naturally be represented by the following experimental matrix:

Trial N°	Factor 1	Factor 2	Factor 3	Answer
2	+	−	−	y_2
3	−	+	−	y_3
5	−	−	+	y_5
8	+	+	+	y_8

Table 16.35: **Fractional design with 3 factors into Yates form**

Or more visually:


Full factorial design

One-half fractional factorial
(either pink or green points)

Figure 16.244: **Visual 2^2 factorial design and fractional factorial design**

Remark

We have seen that the 8 runs (2^3 experiments) above is splitted into 2×4 fractional experiments. In a software like Minitab (see companion book) the practitioner can choose if he wants to use the first fraction or the second one in the options of the design builder.

For a fractional DoE splitted for example into 4 parts the practitioner can therefore choose in Minitab 4 possible fractions (but they are obviously equivalent!!!).

There is however a problem: Even if the triple interaction is really equal to zero, it can remain up to 7 other coefficients in the model, while we have only 4 trial results to identify them. In other words, unless we know a priori that at least 3 of these coefficients are equal to zero (to reduce the problem to 4 equations with 4 unknowns), we will get at best only relations between the coefficients and the rigorous identification will therefore be impossible. Thus, it is not possible to indefinitely reduce the cost of an experimental design without degrading its robustness (information it provides).

It is important to notice that in the fractional factorial design above, the third factor is confused (is aliased) with the interaction 12 of factors 1 and 2. We name this the "initial alias" or "alias generator" and we can notice by reiterating the calculations for factorial designs with 4, 5, 6... factors that the generators helps us to immediately identify the trials (runs) to preserve (you can control when using softwares like JMP or Minitab that they don't let you choose the alias generator, these softwares take by default the theoretical one). For example, the generator ("alias" or "confusion") of the above experimental table will be written by the tradition:

```
C=AB
```

This does not mean that the coefficient of the model are equal but simply that the design is unable to separate the analysis of these two entities. We must therefore pay a particular attention to the interpretation of the respective coefficients!

We have just seen that a $2^{(3-1)}$ design (resolution `III`) has the serious drawback of confusing a main factor with an interaction of order 2. A $2^{(4-1)}$ (resolution `IV`) design has the only disadvantage of confusing (aliasing) a main factor with an interaction of order 3 and two interactions of order 2. A $2^{(5-1)}$ design has even lower drawbacks. It is for this reason that the factorial theory uses the concept of "resolution". The higher the resolution, the greater the design is accurate.

Definitions (#16.7.738):

D1. When no main effect is aliased with another main effect, but the main effects have aliases with interactions of 2 factors, we speak then of "resolution III designs". From a practical point of view, resolution III designs are mainly intended to allow exploratory research because they allow to explore a number of factors with a certain economy. Even if the fail to obtain a sufficiently accurate model they possibly can eliminate a lot of factors in first time and then reduce the number of trials (runs).

D2. When no main effect has alias with another main effect or another 2 factors interaction, but some of the 2 factors interactions have aliases with other 2 factor interactions and some main effects have aliases wither 3 factor interactions, we speak of "resolution IV designs".

D3. When no main effect or interaction of 2 factors have alias with another different main effect or another 2 factors interaction, but 2 factors interactions have alias with 3 factors interactions and main effects have alias interactions with 4 factors interactions we speak of "resolution V designs".

and so on... Here is a table that summarize a bit the situation:

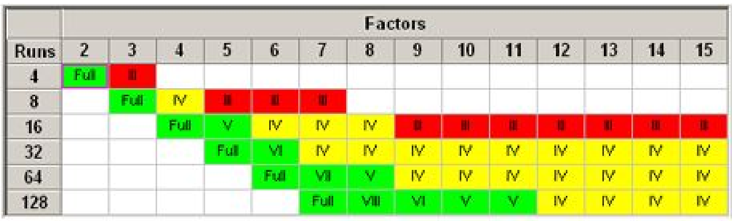
Resolution	**Ability**
`I`	• Not useful: an experiment of exactly one run only tests one level of a factor and hence can't even distinguish between the high and low levels of that factor
`II`	• Not useful: main effects are confounded with other main effects
`III`	• Estimate main effects, but these may be confounded with two-factor interactions
`IV`	• Estimate main effects unconfounded with two-factor interaction but confounded with three-factor interactions • Estimate two-factor interaction effects, but these may be confounded with other two-factor interactions
`V`	• Estimate main effects unconfounded by three-factor (or less) interactions but confounded with four-factor interactions • Estimate two-factor interactions effects unconfounded with other two-factor interactions but confounded with three-factor interactions
`VI`	• Estimate main effects unconfounded by four-factor (or less) interactions • Estimate two-factor interaction effects unconfounded by three-factor (or less) interactions • Estimate three-factor interaction effects, but these may be confounded with other three-factor interactions

Table 16.36: **Fractional factorial designs resolution summary table**

Remark

The reader must keep in mind that when you have a factorial design with n factors, and hence with n columns, there is $n!$ way to represent on a paper or on a screen the columns using different ordering of the columns. For example, for a design with 7 factors, most textbooks and softwares will present on the screen 7 columns (the columns corresponding to the main factors) and there is therefore $7!=5040$ possible order to represent the 7 columns on a display (paper or screen)!

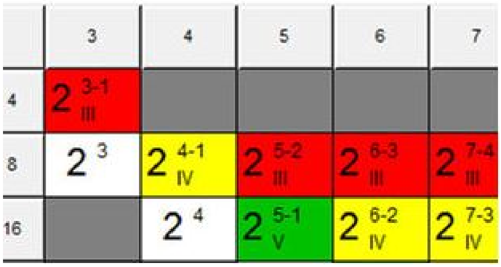
This concept of resolution is important. We find it in programs like Minitab 15.1.1 in the selection of fractional factorial designs with the most common alias known under the name "minimum aberration designs":



	Factors													
Runs	2	3	4	5	6	7	8	9	10	11	12	13	14	15
4	Full	III												
8		Full	IV	III	III	III								
16			Full	V	IV	IV	IV	III	III	III	III	III	III	III
32				Full	VI	IV	IV	IV	IV	IV	IV	IV	IV	IV
64					Full	VII	V	IV	IV	IV	IV	IV	IV	IV
128						Full	VIII	VI	V	V	IV	IV	IV	IV

Figure 16.245: **Display of fractional factorial designs resolutions selector with Minitab 15.1.1**

or with the Design Expert software (part of the picture but is more explicit than Minitab 15.1.1):



	3	4	5	6	7
4	2^{3-1}_{III}				
8	2^{3}	2^{4-1}_{IV}	2^{5-2}_{III}	2^{6-3}_{III}	2^{7-4}_{III}
16		2^{4}	2^{5-1}_{V}	2^{6-2}_{IV}	2^{7-3}_{IV}

Figure 16.246: **Display of fractional factorial designs resolutions selector in Design Expert**

Thus, the reader will observe that a full factorial design of 5 factors with 32 trials (runs) can be reduced to 16 trials by bringing together the influential factors in pairs or 2-tuples (hence the division by 2 of the number of tests) , or 8 tests by bringing together the influential factors by 4-tuples.

Then it's the job of the experimenter to know well its analysis and whether:

1. Among the aliased factors if there are interactions or not!
2. In the aliased factor, the strong influence on the answer comes from the interaction or from the pure effect alone!

Once determined coefficients, assuming that each of the factors or interactions is independent (acceptable limit hypothesis ...) some engineers are analysing the variance of the regression line obtained final or determine the correlation coefficient to whether the linear approximation of the model is acceptable in the field of study and application.

Regarding the generators of fractional factorial designs here is a non-exhaustive summary table:

Factors	Trials	Resolution	Reduction	Generators	Aliases
3	4	III	$2^{(3-1)}$	C=AB	I+ABC,A+BC,B+AC,C+AB
4	8	IV	$2^{(4-1)}$	D=ABC	I+ABCD,A+BCD,B+ACD C+ABD,D+ABC,AB+CD AC+BD,AD+BC
5	8	III	$2^{(5-2)}$	D=AB E=AC	I+ABD+ACE+BCDE A+BD+CE+ABCDE B+AD+CDE+ABCE C+AE+BDE+ABCD D+AB+BCE+ACDE E+AC+BCD+ABDE BC+DE+ABE+ACD BE+CD+ABC+ADE
6	16	V	$2^{(5-1)}$	E=ABCD	A+BCDE B+ACDE C+ABDE D+ABCE E+ABCD AB+CDE AC+BDE AD+BCE AE+BCD BC+ADE BD+ACE BE+ACD CD+ABE CE+ABD DE+ABC
...	...	...	...	...	...

Table 16.37: **Some fractional factorial designs with generators and alias**

Of course, use of fractional design is an economic (and temporal) bet. If the conclusions are clear, then we have saved time and reduced our effort. But sometimes we lose the bet. We can therefore use a "complementary design" that consists of adding in a pertinent way enough lines to the initial design to unaliased some desired coefficients based on the choice of the alias generator.

Before moving on to another type of design, let us come back on to the fractional factorial design with 3 factors and therefore with 4 trials for educational reasons. Consider that we have done these 4 trials, and for each, we got a measurement as given in the figure below:

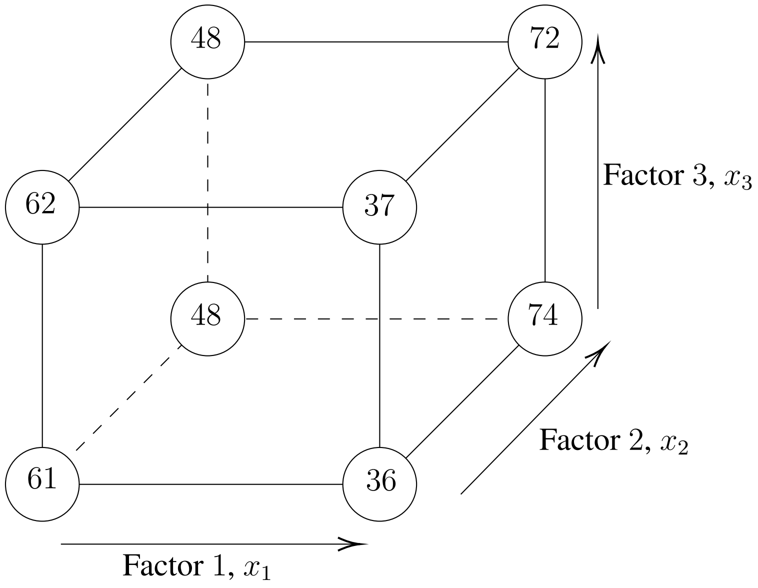


Figure 16.247: Representation of a real fractional factorial design with 3 factors

We want to show to the reader not how to determine the coefficients by calculations (its all about solving a simple linear system and then there will be an example about its it a little further below) but how to calculate the effects in the context of this special case!

Indeed, the effect of x_1 is:

$$\frac{72+37+74+36}{4}-\frac{48+62+48+61}{4}=\frac{(24)+(-25)+26+(-25)}{4}=0 \tag{16.7.465}$$

and this of x_2 is:

$$\frac{48+72+48+74}{4}-\frac{62+37+61+36}{4}=\frac{(-14)+35+(-13)+38}{4}=11.5 \tag{16.7.466}$$

and this of x_3 is:

$$\frac{48+72+62+37}{4}-\frac{48+74+61+36}{4}=\frac{0+(-2)+1+1}{4}=0 \tag{16.7.467}$$

The effect of interaction is somewhat subtle in the case of 3 factors using the figure above (we will see in the example further below with a picture that with a table it is much more intuitive). Thus, we have for the interaction x_1x_2:

$$\frac{72+74+62+61}{4}-\frac{48+37+48+36}{4}=25 \tag{16.7.468}$$

and for x_1x_3:

$$\frac{37+72+48+61}{4}-\frac{48+62+74+36}{4}=-0.5 \tag{16.7.469}$$

and for x_2x_3:

$$\frac{48+72+61+36}{4}-\frac{62+37+48+74}{4}=-1 \tag{16.7.470}$$

For the triple interaction $x_1x_2x_3$ with the figure above it is also not obvious (it is a little more intuitive with a table as we will see in the example further below). We have:

$$\frac{72+62+48+36}{4}-\frac{48+37+74+61}{4}=-0.5 \tag{16.7.471}$$

16.7.4.2 General factorial Designs

In this textbook we name "general factorial design" designs that have a given quantity of factors but with more than 2 levels and that are not reduced.

Given n the number of factor, l_i the number of levels of each factor. The number of runs N of a general factorial design is given obviously by:

$$N=\prod_{i=1}^{n} l_i \qquad (16.7.472)$$

As we will see below, many of the techniques relatives to complete factorial design (designs with only 2 levels for recall) applies identically to general factorial design. The main differences that surprise people starting to work in the field of DoE are:

- The final regression model don't have a unique coefficient for each factor therefore we cannot do a statistical regression analysis.
- The calculation methods of the standard error of the coefficients we have seen for the 2 factorial design cannot be applied (we have to use other methods of calculations)

Rather than doing a long theoretical introduction to general factorial designs, let us use a companion example that will show us that in fact most concepts that we have learned so far also applies!

As part of the study of tires, the criterion retain by a company is the longevity (number of kilometres travelled before the tire blow up) and the chosen factors are the type of use (city or highway), average speed (40 [km/h] or 50 [km/h]) and the inflation pressure (1.5 or 2 or 2.5 [kg]).

Therefore we have a general factorial design with the following number of runs:

$$N=\prod_{i=1}^{3} l_i = 2\cdot 2\cdot 3 = 12 \qquad (16.7.473)$$

Furthermore, it is more healthy to repeat the experiments (to analyse the dispersion of the values in order to run an ANOVA), so to make 24, 36, etc. experiments or to make some central points measurements.

We will denote as usual y the variable of interest and in this case x_1,x_2,x_3 the variables factors. We are seeking a simple model in the form:

$$y=a_0+a_1x_1+a_2x_2+a_3x_3+\ldots \qquad (16.7.474)$$

As each factor can have several levels, we will have for example the value of y at the level 2 of factor 1 and at level 1 of factor 2 and at level 3 of factor 3 that will be denoted:

$$y=a_0+a_1x_{1,2}+a_2x_{2,1}+a_3x_{3,3} \qquad (16.7.475)$$

The full factorial design is the presentation of the various experiences (measurement) depending on the levels of the variables. For example, to study the tires we will have:

- For the factor 1, the place, denoted x_1: city$=1$, town $=2$
- For the factor 2, the speed, denoted x_2: $40\,[\mathrm{km}\cdot\mathrm{h}^{-1}]=1$, $50\,[\mathrm{km}\cdot\mathrm{h}^{-1}]=2$
- For the factor 3, the pressure, denoted x_3: 1.5 [kg] $=1$, 2 [kg] $=2$, 3 [kg] $=3$

The experimental matrix will then be (this is exactly the same as that generally obtained with the software Minitab: a full complete general design for 2 factors with 2 levels and 1 factor with 3 levels without randomization):

Trial N°	x_1	x_2	x_3
1	1	1	1
2	1	1	2
3	1	1	1
4	1	2	1
5	1	2	2
6	1	2	3
7	2	1	1
8	2	1	2
9	2	1	3
10	2	2	1
11	2	2	2
12	1	2	3

At this table, we can associate the results of the experiments. For example, for the tires, we did three tests (three "replication") for each of the levels and we got:

Trial N°	x_1	x_2	x_3	**Measured values of y**		
1	1	1	1	32700	32750	32960
2	1	1	2	33430	33360	32910
3	1	1	3	31710	32100	32220
4	1	2	1	32680	32270	33130
5	1	2	2	34070	33100	33610
6	1	2	3	33220	33700	33285
7	2	1	1	33180	32160	32640
8	2	1	2	34430	34380	34460
9	2	1	3	33570	33300	32570
10	2	2	1	33270	33080	32415
11	2	2	2	33440	33570	34204
12	1	2	3	32840	33210	32470

Thanks to these values we can calculate the effects of the various factors. For this we start with principle that their average (effect) is zero for a given factors. Therefore the global arithmetic average (that can be calculated as the arithmetic average if and only the variables are coded):

$$\bar{y}=33119.3 \tag{16.7.476}$$

is the constant coefficient of the model (intercept):

$$x_0 = 33119.3 \tag{16.7.477}$$

The effect of x_1 at the level 1 is the variation of this average when we consider only the cases where x_1 is at the level 1. The arithmetic average becomes 32955.8 and therefore the effect is:

$$32955.8 - x_0 = 32955.8 - 33119.3 = -163.4 \tag{16.7.478}$$

We will write this obviously as we already know:

$$x_{1,1} = -163.4 \tag{16.7.479}$$

At the level of x_1, the average becomes 33282.7, therefore the effect of x_1 at level 2 is equal to:

$$x_{1,2} = 33282.7 - x_0 = 33282.7 - 33119.3 = 163.4 \tag{16.7.480}$$

The difference:

$$x_{1,1} - x_{1,2} = 163.4 - (-163.4) = 326.8 \tag{16.7.481}$$

is named the "effect" of the factor.

We notice that as assumed, the sum of the effects x_1 is zero, as it is the sum of the deviations from the arithmetic average of the sub-populations of the same size (and thus the overall average is the arithmetic average of the levels).

Exactly in the same way, we get:

$$\begin{aligned} x_{2,1} &= -78.7 \quad & x_{2,2} &= 78.7 & & \\ x_{3,1} &= -349.7 \quad & x_{3,2} &= 619.4 \quad & x_{3,3} &= -269.7 \end{aligned} \tag{16.7.482}$$

and we also find that the sum of the effects of a variable (factor) is zero.

We modelize these results into the form using the average response (notice that the sum of each vector is zero):

$$y = 33119.31 + x_1 \begin{bmatrix} -163.4 \\ 163.4 \end{bmatrix} + x_2 \begin{bmatrix} -78.7 \\ 78.7 \end{bmatrix} + x_3 \begin{bmatrix} -349.7 \\ 619.4 \\ -269.7 \end{bmatrix} \tag{16.7.483}$$

giving the model found. Function that we can rewrite under the form of absolute average effects (the most interesting form mathematically speaking):

$$y = 33119.3 + 164.4x_1 + 78.7x_2 + 619.4x_3 \tag{16.7.484}$$

intermediate result to be compared with a simple linear regression that (perfectly identical result between Microsoft® Excel® 14.0.6117 and Minitab 15.1.1):

$$y = 32313.8 + 326.9x_1 + 157.4x_2 + 40x_3 \tag{16.7.485}$$

but that does not make sense since the linear regression is not built for factors over two levels (see section Numerical Methods page 6335).

From these results, we already derive the best tire usage policy: road, speed 50 [km/h] and pressure 2 [kg], we can expect to do:

$$y = x_0 + x_{1,2} + x_{2,2} + x_{3,2} = 33119.3 + 163.4 + 78.7 + 619.4 \cong 33981 \tag{16.7.486}$$

Let us notice immediately that in some cases, the experiment did better (see the values in the above table), even in town or at 40 $[\text{km} \cdot \text{h}^{-1}]$. This come for the obvious dispersion of the results. But then... did the effects found come perhaps only from this dispersion!? Furthermore we studied the effects independently of each other. But it may have some that reinforce some others. We will explore these issues further below!

Let us take back our 36 experiments, comparing the differences Δ between the values measured (denoted here $\hat{y}$) and the model predictions (denoted here y):

x_1	x_2	x_3	$\hat{y}$	y	Δ	x_1	x_2	x_3	$\hat{y}$	y	Δ
1	1	1	32700	32527.5	172.5	2	1	1	32160	32854.3	−694.3
1	1	2	33430	33496.6	−66	2	1	2	34280	33823.4	456.6
1	1	3	31710	32607.5	−897.5	2	1	3	33300	32943.3	365.7
1	2	1	32680	32684.9	−4.9	2	2	1	33080	33011.7	68.3
1	2	2	34070	33654	416	2	2	2	33570	33980.8	−410.8
1	2	3	33220	32764.9	455.1	2	2	3	33210	33091.7	118.3
2	1	1	33180	32854.3	325.7	1	1	1	32960	32527.5	432.5
2	1	2	34430	33823.4	606.6	1	1	2	32910	33496.6	−586.6
2	1	3	33570	32934.3	635.7	1	1	3	32220	32607.5	−387.5
2	2	1	33270	33011.7	258.3	1	2	1	33130	32684.9	445.1
2	2	2	33440	33980.8	−540.8	1	2	2	33610	33654	−44
2	2	3	33840	33091.7	−251.7	1	2	3	33285	32764.9	520.1
1	1	1	32750	32527.5	222.5	2	1	1	32640	32854.3	−214.3
1	1	2	33360	33496.6	−136.6	2	1	2	34460	33823.4	636.6
1	1	3	32100	32607.5	−507.5	2	1	3	32570	32934.3	−364.3
1	2	1	32270	32684.9	−414.9	2	2	1	32415	33011.7	−596.7
1	2	2	33100	33654	−554	2	2	2	34204	33980.8	223.2
1	2	3	33700	32764.9	935.1	2	2	3	32470	33091.7.8	−621.7

We see that the differences between the model and the predicted values are sometimes important. This may be a natural random dispersion of values, or an incorrect model or it could be due to the fact that combination of factors have more effect than separate. The positive or negative effects of two factors may do more than just add themselves (purely additive model may therefore be reject).

It is therefore probably interactions (chemists speak of "potentiation"). To see if there are second orders interactions, a good way is to study all the factors in pairs (and ignore the existence of others). So if we start with the factors (location) x_1 and (speed) x_2 by ignoring the tire pressure x_3, we get (the explanations of how the values in this table are calculated are given below the table):

Levels	Experimental average	Total effect	Sum of effects	Interaction effect
$x_{1,1}x_{2,1}$	32682.2	-437.1	-242.2	-194.9
$x_{1,1}x_{2,2}$	33229.4	110.2	-84.7	-194.9
$x_{1,2}x_{2,1}$	33398.9	279.6	84.7	194.9
$x_{1,2}x_{2,2}$	33166.6	47.3	242.2	-194.9

Table 16.38: **Table of interaction of factors 1 and 2**

So in the table above $32,682.2$ is the arithmetic average of the experiences done at level $x_{1,1}$ (City) and level $x_{2,1}$ (40 [km/h]). By subtracting the main average x_0 (always $33,119.3$), we get an effect of -437.1, while the effect of x_1 (-163.4) plus the effect of x_2 (-78.7) gives only -242.2.

So there a surplus off -194.9, largely due to the interaction $x_{1,1},x_{2,1}$ (notice that this is the same for $x_{1,2}x_{2,2}$, and the opposite for $x_{1,1}x_{2,2}$ and $x_{1,2}x_{2,1}$). We notice immediately that the interaction $x_{1,1}x_{2,1}$ has a much greater effect than the effects of x_1 and x_2!

Let us see the two other possible interactions of respectively factors 1 and 3 and 2 and 3:

Levels	Experimental average	Total effect	Sum of effects	Interaction effect
$x_{1,1}x_{3,1}$	32748.3	-370.9	-513.1	142.2
$x_{1,1}x_{3,2}$	33143.3	294.1	455.9	-161.9
$x_{1,1}x_{3,3}$	32705.8	-413.4	-433.1	19.7
$x_{1,2}x_{3,1}$	32790.8	-328.4	-186.3	-142.2
$x_{1,2}x_{3,3}$	34064.0	944.7	782.8	161.9
$x_{1,2}x_{3,3}$	32993.3	-125.9	-106.3	-19.7

Levels	Experimental average	Total effect	Sum of effects	Interaction effect
$x_{2,1}x_{3,1}$	32731.7	-387.6	-428.4	40.8
$x_{2,1}x_{3,2}$	33811.7	629.4	540.7	151.7
$x_{2,1}x_{3,2}$	33811.7	692.4	540.7	151.7
$x_{2,1}x_{3,3}$	32578.3	-540.9	-348.4	-192.5
$x_{2,2}x_{3,1}$	32807.5	-311.8	-271.0	-40.8
$x_{2,2}x_{3,2}$	33665.7	546.5	698.1	-151.7
$x_{2,2}x_{3,3}$	33120.8	1.6	-191.0	192.5

We can represent the interactions differently by crossing the factors:

	$x_{2,1}$	$x_{2,2}$
$x_{1,1}$	-194.9	194.9
$x_{1,2}$	194.9	-194.9

	$x_{3,1}$	$x_{3,2}$	$x_{3,3}$
$x_{1,1}$	142.2	-161.9	19.7
$x_{1,2}$	-142.2	161.9	-19.7

	$x_{3,1}$	$x_{3,2}$	$x_{3,3}$
$x_{2,1}$	40.8	151.7	−192.5
$x_{2,2}$	−40.8	−151.7	192.5

We notice then that each line or each column has a total of zero. We also see that in this case, it is difficult to neglect the interactions, and that x_2 (for example) influence the results more by its interactions than that its own effect!

We can still try to check if the interaction of the three factors has a significant effect. As the experiment has been repeated 3 times at each level (we then speak of "three-blocks"), we can calculate the mean of y, and subtract to it the model with the interactions calculated above (notice that the sum of each vector is zero!):

$$\begin{aligned} y = 33119.31 &+ x_1\begin{bmatrix}-163.4\\163.4\end{bmatrix} + x_2\begin{bmatrix}-78.7\\78.7\end{bmatrix} + x_3\begin{bmatrix}-349.7\\619.4\\-269.7\end{bmatrix} + x_1x_2\begin{bmatrix}-194.9\\194.9\\194.9\\-194.9\end{bmatrix}\\ &+ x_1x_3\begin{bmatrix}142.2\\-161.9\\19.7\\-142.2\\161.9\\-19.7\end{bmatrix} + x_2x_3\begin{bmatrix}40.8\\151.7\\-192.5\\-40.8\\-151.7\\192.5\end{bmatrix} \end{aligned} \tag{16.7.487}$$

that it is customary to write for simplification purposes (notice that the sum of each vector is still zero):

$$\begin{aligned} y = 33119.31 &+ x_1\begin{bmatrix}-163.4\\163.4\end{bmatrix} + x_2\begin{bmatrix}-78.7\\78.7\end{bmatrix} + x_3\begin{bmatrix}-349.7\\619.4\\-269.7\end{bmatrix} + x_1x_2\begin{bmatrix}-194.9\\194.9\end{bmatrix}\\ &+ x_1x_3\begin{bmatrix}142.2\\-161.9\\19.7\end{bmatrix} + x_2x_3\begin{bmatrix}40.8\\151.7\\-192.5\end{bmatrix} \end{aligned} \tag{16.7.488}$$

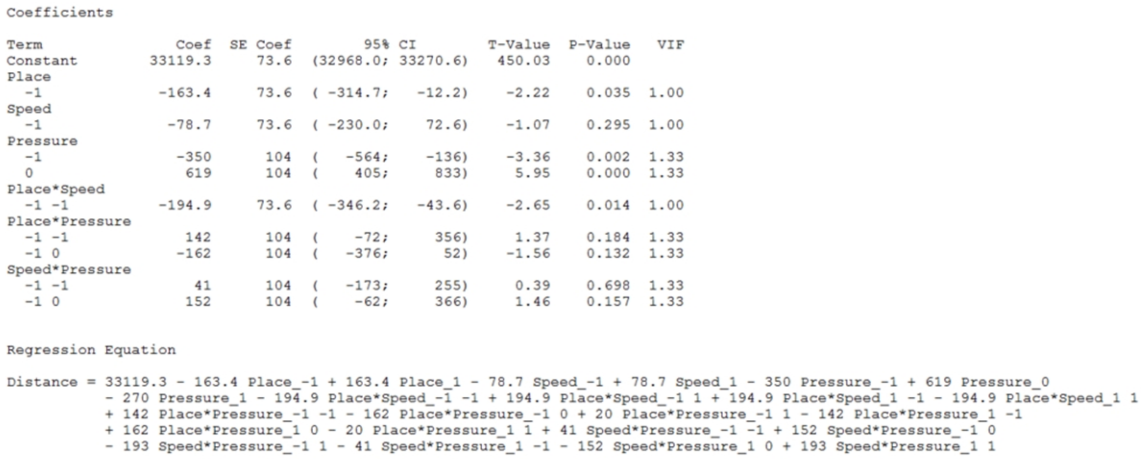
Remark

It suffices to know that for the coefficients of interaction terms the missing effects (not written) are of opposed sign.

Some statistical software go further in the simplification of notation by writing (still in the same idea that the other coefficients are of opposed sign):

$$\begin{aligned} y = 33119.31 &+ x_1\begin{bmatrix}-163.4\end{bmatrix} + x_2\begin{bmatrix}-78.7\end{bmatrix} + x_3\begin{bmatrix}-349.7\\619.4\\-269.7\end{bmatrix} + x_1x_2\begin{bmatrix}-194.9\end{bmatrix}\\ &+ x_1x_3\begin{bmatrix}142.2\\-161.9\\19.7\end{bmatrix} + x_2x_3\begin{bmatrix}40.8\\151.7\\-192.5\end{bmatrix} \end{aligned} \tag{16.7.489}$$

and other softwares (such as Minitab 17.1.3 for example) give:



```
Coefficients

Term              Coef  SE Coef          95% CI           T-Value  P-Value   VIF
Constant       33119.3     73.6  (32968.0; 33270.6)        450.03    0.000
Place
  -1            -163.4     73.6  ( -314.7;    -12.2)        -2.22    0.035  1.00
Speed
  -1             -78.7     73.6  ( -230.0;     72.6)        -1.07    0.295  1.00
Pressure
  -1              -350      104  (    -564;    -136)        -3.36    0.002  1.33
  0                619      104  (     405;     833)         5.95    0.000  1.33
Place*Speed
  -1 -1         -194.9     73.6  ( -346.2;    -43.6)        -2.65    0.014  1.00
Place*Pressure
  -1 -1            142      104  (     -72;     356)         1.37    0.184  1.33
  -1 0            -162      104  (    -376;      52)        -1.56    0.132  1.33
Speed*Pressure
  -1 -1             41      104  (    -173;     255)         0.39    0.698  1.33
  -1 0             152      104  (     -62;     366)         1.46    0.157  1.33


Regression Equation

Distance = 33119.3 - 163.4 Place_-1 + 163.4 Place_1 - 78.7 Speed_-1 + 78.7 Speed_1 - 350 Pressure_-1 + 619 Pressure_0
           - 270 Pressure_1 - 194.9 Place*Speed_-1 -1 + 194.9 Place*Speed_-1 1 + 194.9 Place*Speed_1 -1 - 194.9 Place*Speed_1 1
           + 142 Place*Pressure_-1 -1 - 162 Place*Pressure_-1 0 + 20 Place*Pressure_-1 1 - 142 Place*Pressure_1 -1
           + 162 Place*Pressure_1 0 - 20 Place*Pressure_1 1 + 41 Speed*Pressure_-1 -1 + 152 Speed*Pressure_-1 0
           - 193 Speed*Pressure_-1 1 - 41 Speed*Pressure_1 -1 - 152 Speed*Pressure_1 0 + 193 Speed*Pressure_1 1
```

We can then summarize all these results in the form:

Trial N°	x_1	x_2	x_3	**Measured values of y**			**Average**	**Model**	Δ
1	1	1	1	32700	32750	32960	32803.3	32515.6	287.8
2	1	1	2	33430	33360	32910	33233.3	33291.5	−58.1
3	1	1	1	31710	32100	32220	32010.0	32239.8	−229.8
4	1	2	1	32680	32270	33130	32693.3	32981.2	−287.8
5	1	2	2	34070	33100	33610	33593.3	33535.3	58.1
6	1	2	3	33220	33700	33285	33401.7	33172.0	229.7
7	2	1	1	33180	32160	32640	32660	32947.8	−287.8
8	2	1	2	34430	34380	34460	34390.0	34331.9	58.1
9	2	1	3	33570	33300	32570	33146.7	32917.0	229.7
10	2	2	1	33270	33080	32415	32921.7	32633.8	287.9
11	2	2	2	33440	33570	34204	33738.0	33796.1	−58.1
12	1	2	3	32840	33210	32470	32840.3	33069.6	−229.6

We see then that the values of the model with interactions continues to diverge quite significantly for the average values, which can be interpreted either by the fact that the effects are not linear, either by an interaction between the three factors.

It would be necessary to redo the calculations above with the interaction of the three factors. But as it is always the same principle, we will do that if some readers explicitly request us to add it.

But there may still exist a source of error in the model, that is to take into account effects that do not occur in reality! To ensure that the effect calculated on a variable or interaction is real, we will use to start the one-way fixed factor analysis of variance (ANOVA) that we have studied in detail in the section Statistics. For this, it is important to have repeated the experiments in order to highlight the dispersion due to external and almost unknown uncontrolled factors.

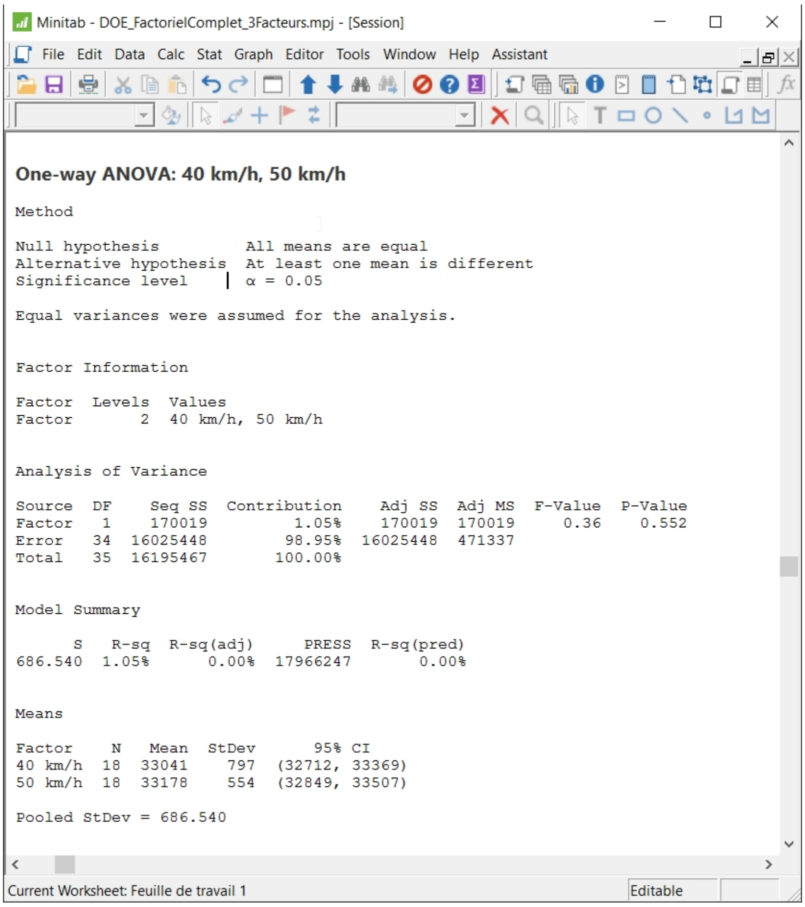
For example, the factor x_2 has a rather small effect. Does he have a real influence? For this, we separate the 36 experiments in two samples (level $x_{2,1}$ and level $x_{2,2}$, and we calculate the sums of squares and the degrees of freedom (we can detail the following calculations on request) as we have seen it in the section of Statistics:

Source	Sum of squares	χ^2 df	Average of squares
Inter-Class	$Q_A=\sum_i n_i\left(\bar{x}_i-\bar{\bar{x}}\right)^2=170019$	$k-1=1$	$\text{MSk}=\dfrac{Q_A}{k-1}=170019$
Intra-Class	$Q_R=\sum_{ij}\left(x_{ij}-\bar{x}_i\right)^2=16025448$	$N-k=34$	$\text{MSE}=\dfrac{Q_R}{k-1}=471337$
Total	$Q_T=\sum_{ij}\left(x_{ij}-\bar{\bar{x}}\right)^2=16195467$	$N-1=35$	

And we get from this table:

$$F=\frac{\text{MSk}}{\text{MSE}}=0.36\Rightarrow p\text{-value}=0.552 \tag{16.7.490}$$

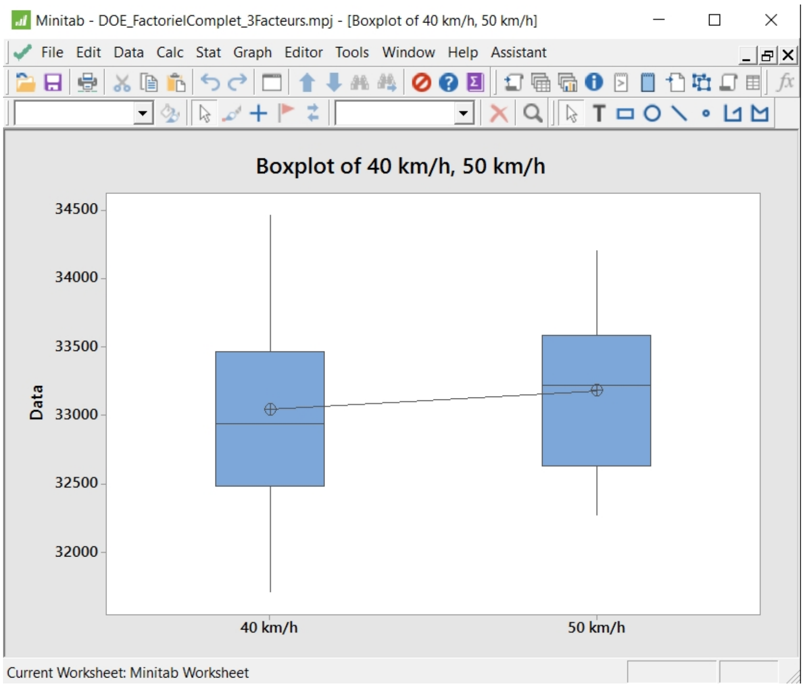
What gives with Minitab 17.3.1 (we can see also that the model is probably not linear) as detailed in the Minitab companion book:


Minitab - DOE_FactorielComplet_3Facteurs.mpj - [Session]
File
Edit
Data
Calc
Stat
Graph
Editor
Tools
Window
Help
Assistant
Current Worksheet: Feuille de travail 1
Editable

One-way ANOVA: 40 km/h, 50 km/h

```
Method

Null hypothesis         All means are equal
Alternative hypothesis  At least one mean is different
Significance level      α = 0.05

Equal variances were assumed for the analysis.


Factor Information

Factor  Levels  Values
Factor       2  40 km/h, 50 km/h


Analysis of Variance

Source  DF    Seq SS  Contribution    Adj SS  Adj MS  F-Value  P-Value
Factor   1    170019         1.05%    170019  170019     0.36    0.552
Error   34  16025448        98.95%  16025448  471337
Total   35  16195467       100.00%


Model Summary

      S   R-sq  R-sq(adj)     PRESS  R-sq(pred)
686.540  1.05%      0.00%  17966247       0.00%


Means

Factor    N   Mean  StDev       95% CI
40 km/h  18  33041    797  (32712, 33369)
50 km/h  18  33178    554  (32849, 33507)

Pooled StDev = 686.540
```

and visually:


Minitab - DOE_FactorielComplet_3Facteurs.mpj - [Boxplot of 40 km/h, 50 km/h]
File Edit Data Calc Stat Graph Editor Tools Window Help Assistant
Boxplot of 40 km/h, 50 km/h
Data
34500
34000
33500
33000
32500
32000
40 km/h
50 km/h
Current Worksheet: Minitab Worksheet

As we can see we cannot reject the null hypothesis H_0 of this ANOVA. Therefore this factor has no significant impact on our experience.

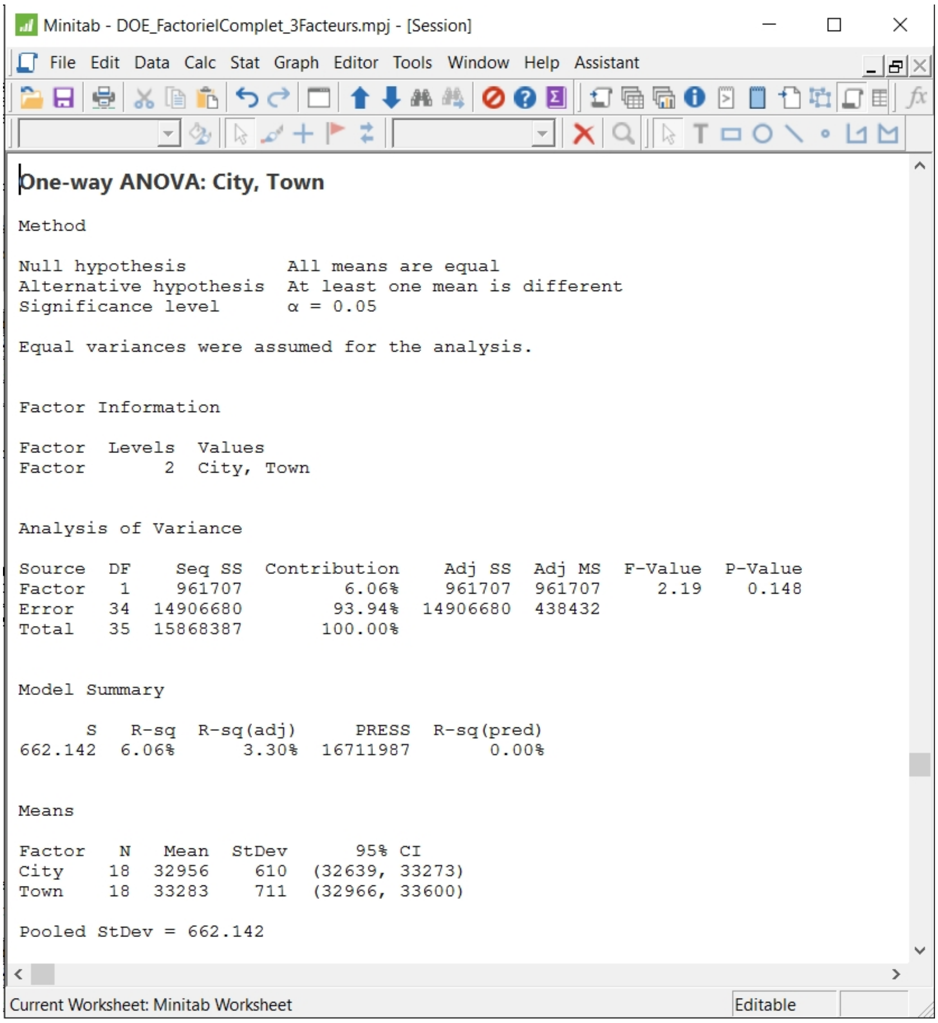
We can do the same calculations for x_1 (the place: city or town):

Source	Sum of squares	χ^2 df	Average of squares
Inter-Class	$Q_A = \sum_i n_i (\bar{x}_i - \bar{\bar{x}})^2 = 961707$	$k-1=1$	$\text{MSk} = \dfrac{Q_A}{k-1} = 961707$
Intra-Class	$Q_R = \sum_{ij} (x_{ij} - \bar{x}_i)^2 = 14906680$	$N-k=34$	$\text{MSE} = \dfrac{Q_R}{k-1} = 438432$
Total	$Q_T = \sum_{ij} (x_{ij} - \bar{\bar{x}})^2 = 15868387$	$N-1=35$	

And we get from this table:

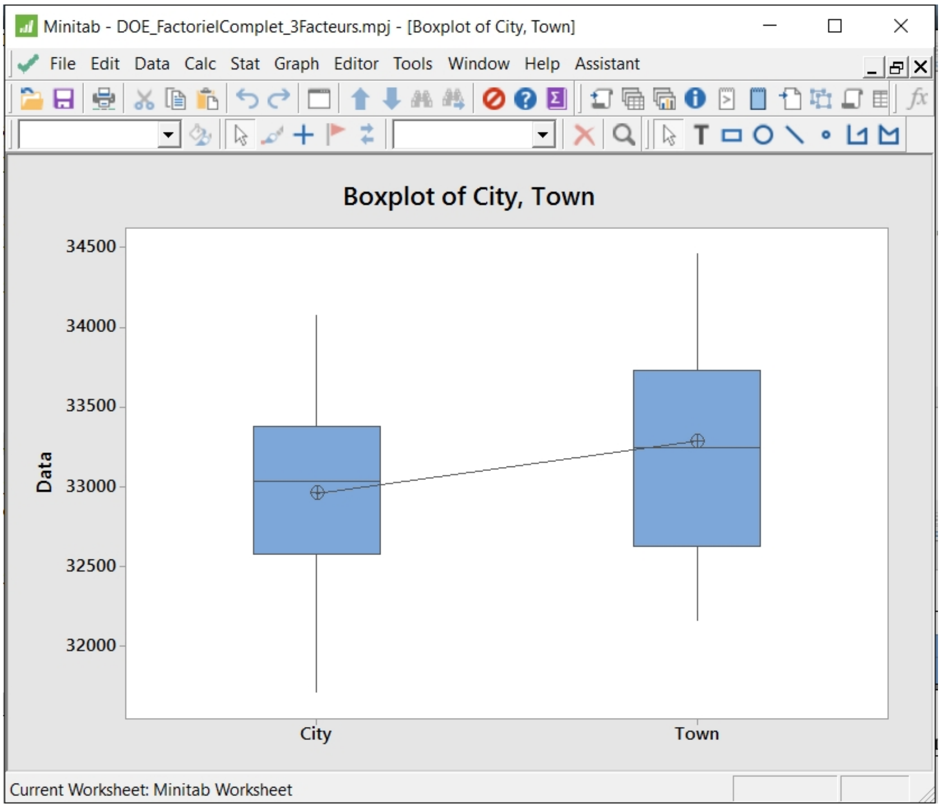
$$F = \frac{\text{MSk}}{\text{MSE}} = 2.19 \Rightarrow p\text{-value} = 0.148 \qquad (16.7.491)$$

What gives with Minitab 17.3.1 (we can see also that the model is probably not linear) as detailed in the Minitab companion book:


Minitab - DOE_FactorielComplet_3Facteurs.mpj - [Session]
File Edit Data Calc Stat Graph Editor Tools Window Help Assistant
Current Worksheet: Minitab Worksheet
Editable

```
One-way ANOVA: City, Town

Method

Null hypothesis         All means are equal
Alternative hypothesis  At least one mean is different
Significance level      α = 0.05

Equal variances were assumed for the analysis.


Factor Information

Factor  Levels  Values
Factor       2  City, Town


Analysis of Variance

Source  DF    Seq SS  Contribution    Adj SS  Adj MS  F-Value  P-Value
Factor   1    961707         6.06%    961707  961707     2.19    0.148
Error   34  14906680        93.94%  14906680  438432
Total   35  15868387       100.00%


Model Summary

      S    R-sq  R-sq(adj)     PRESS  R-sq(pred)
662.142   6.06%      3.30%  16711987       0.00%


Means

Factor   N   Mean  StDev       95% CI
City    18  32956    610  (32639, 33273)
Town    18  33283    711  (32966, 33600)

Pooled StDev = 662.142
```

and visually:


Minitab - DOE_FactorielComplet_3Facteurs.mpj - [Boxplot of City, Town]
File Edit Data Calc Stat Graph Editor Tools Window Help Assistant
Boxplot of City, Town
34500
34000
33500
33000
32500
32000
Data
City
Town
Current Worksheet: Minitab Worksheet

As we can see we also cannot reject the null hypothesis H_0 of this ANOVA. Therefore this factor has also no significant impact on our experience.

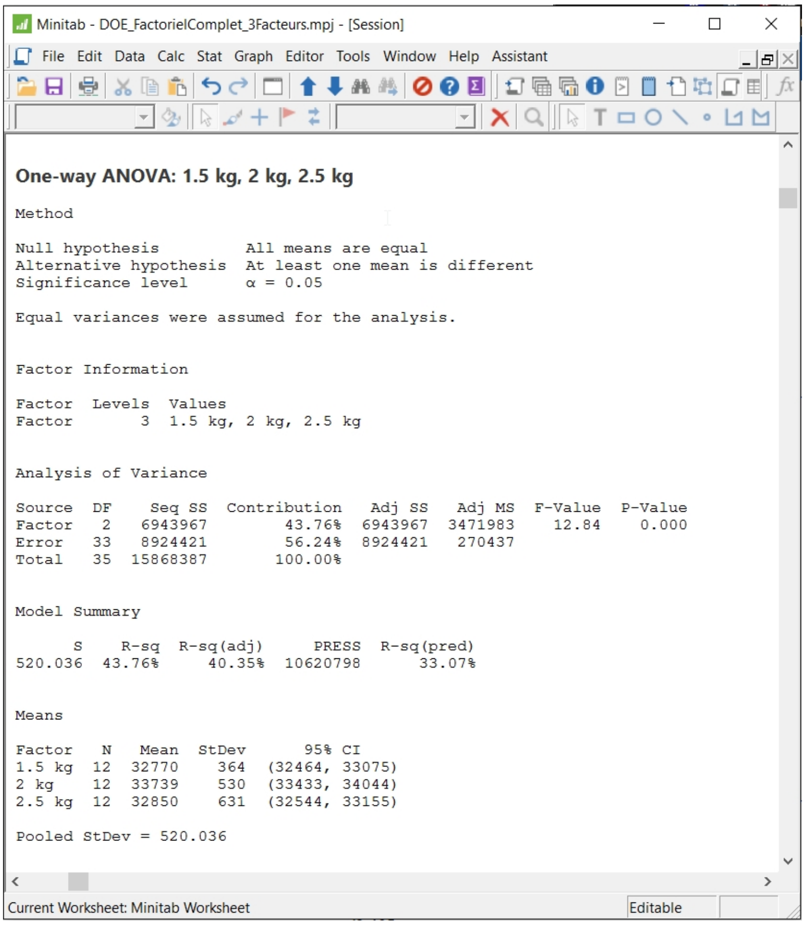
We can do the same calculations for x_3 (the pressure):

Source	Sum of squares	χ^2 df	Average of squares
Inter-Class	$Q_A=\sum_i n_i(\bar{x}_i-\bar{\bar{x}})^2=6943967$	$k-1=2$	$\text{MSk}=\dfrac{Q_A}{k-1}=3471983$
Intra-Class	$Q_R=\sum_{ij}(x_{ij}-\bar{x}_i)^2=8924421$	$N-k=33$	$\text{MSE}=\dfrac{Q_R}{k-1}=270437$
Total	$Q_T=\sum_{ij}(x_{ij}-\bar{\bar{x}})^2=15868387$	$N-1=35$	

And we get from this table:

$$F=\frac{\text{MSk}}{\text{MSE}}=12.84\Rightarrow p\text{-value}=0.00 \tag{16.7.492}$$

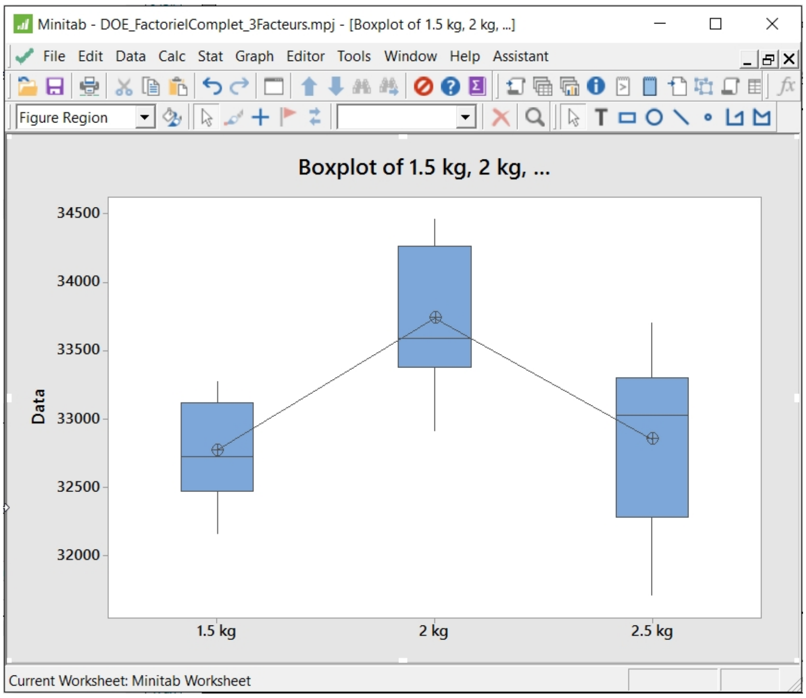
What gives with Minitab 17.3.1 (we can see also that the model is probably not linear) as detailed in the Minitab companion book:


Minitab - DOE_FactorielComplet_3Facteurs.mpj - [Session]
File Edit Data Calc Stat Graph Editor Tools Window Help Assistant
Current Worksheet: Minitab Worksheet
Editable

One-way ANOVA: 1.5 kg, 2 kg, 2.5 kg

```
Method

Null hypothesis         All means are equal
Alternative hypothesis  At least one mean is different
Significance level      α = 0.05

Equal variances were assumed for the analysis.


Factor Information

Factor  Levels  Values
Factor       3  1.5 kg, 2 kg, 2.5 kg


Analysis of Variance

Source  DF    Seq SS  Contribution   Adj SS   Adj MS  F-Value  P-Value
Factor   2   6943967        43.76%  6943967  3471983    12.84    0.000
Error   33   8924421        56.24%  8924421   270437
Total   35  15868387       100.00%


Model Summary

      S    R-sq  R-sq(adj)     PRESS  R-sq(pred)
520.036  43.76%     40.35%  10620798      33.07%


Means

Factor   N   Mean  StDev       95% CI
1.5 kg  12  32770    364  (32464, 33075)
2 kg    12  33739    530  (33433, 34044)
2.5 kg  12  32850    631  (32544, 33155)

Pooled StDev = 520.036
```

and visually:



As we can see here we reject the null hypothesis H_0 of this ANOVA at a threshold of 5%.

Therefore this factor has a significant impact on our experience.

Let us recall that the fact we reject the null hypothesis H_0 highlight an influential factor, the one who is controlled. So the basic model to consider (leaving aside the interactions) is not:

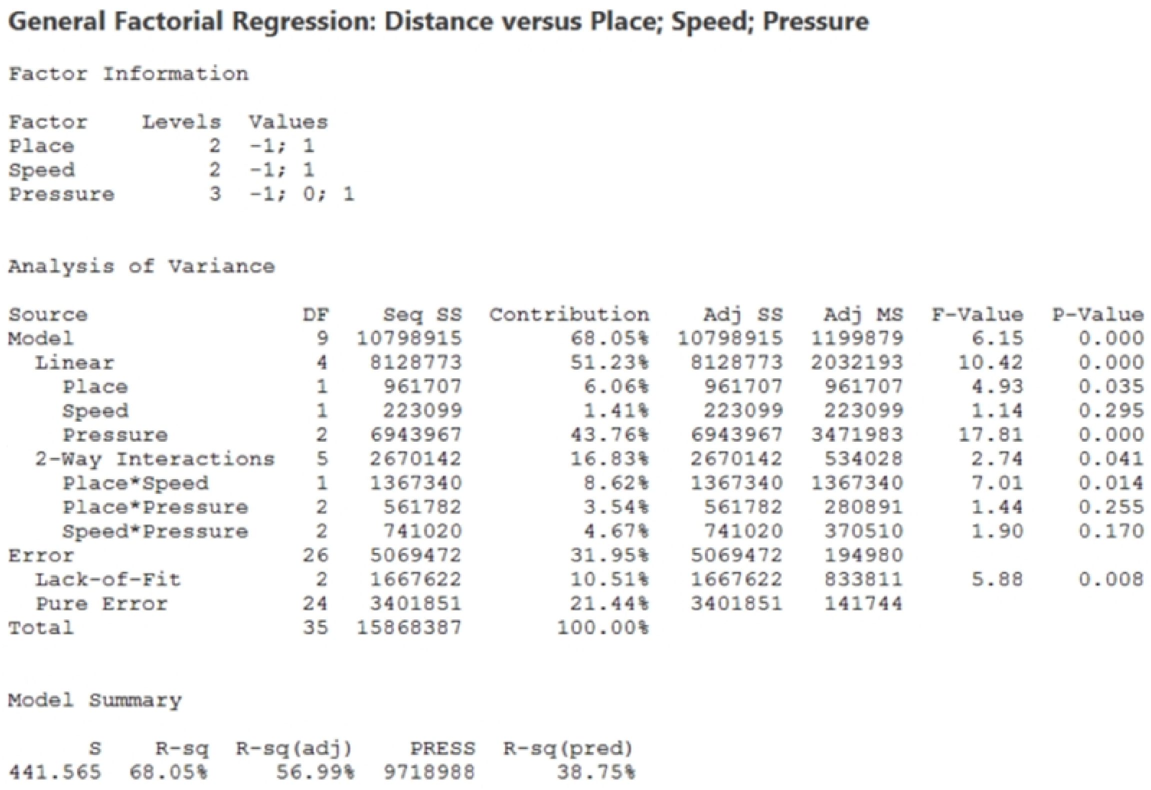
$$y = 33119.3 + x_1 \begin{bmatrix} -163.4 \\ 163.4 \end{bmatrix} + x_2 \begin{bmatrix} -78.7 \\ 78.7 \end{bmatrix} + x_3 \begin{bmatrix} -349.7 \\ 619.4 \\ -269.7 \end{bmatrix} \tag{16.7.493}$$

but (the sum of the vectors still being equal to zero):

$$y \cong 33119.3 + x_3 \begin{bmatrix} -349.7 \\ 619.4 \\ -269.7 \end{bmatrix} \tag{16.7.494}$$

Caution!!! That fact that x_1 (speed) and x_2 (place) have no significant influence does not result that the interactions between x_1 and x_2, or x_2 and x_3, or even x_1x_2 and x_3 are not. For example, here, x_2x_3 has no significant influence (the Fischer-Snedecor test succeeds), but at the level 1 of x_3, the interaction x_1x_2 has a significant influence (in the sense $x_1 = x_2$ against $x_1 \neq x_2$).

Let us see this again with Minitab 17.3.1 (the calculations below can be detailed on request):



General Factorial Regression: Distance versus Place; Speed; Pressure

```
Factor Information

Factor    Levels  Values
Place          2  -1; 1
Speed          2  -1; 1
Pressure       3  -1; 0; 1


Analysis of Variance

Source                 DF    Seq SS  Contribution    Adj SS   Adj MS  F-Value  P-Value
Model                   9  10798915        68.05%  10798915  1199879     6.15    0.000
  Linear                4   8128773        51.23%   8128773  2032193    10.42    0.000
    Place               1    961707         6.06%    961707   961707     4.93    0.035
    Speed               1    223099         1.41%    223099   223099     1.14    0.295
    Pressure            2   6943967        43.76%   6943967  3471983    17.81    0.000
  2-Way Interactions    5   2670142        16.83%   2670142   534028     2.74    0.041
    Place*Speed         1   1367340         8.62%   1367340  1367340     7.01    0.014
    Place*Pressure      2    561782         3.54%    561782   280891     1.44    0.255
    Speed*Pressure      2    741020         4.67%    741020   370510     1.90    0.170
Error                  26   5069472        31.95%   5069472   194980
  Lack-of-Fit           2   1667622        10.51%   1667622   833811     5.88    0.008
  Pure Error           24   3401851        21.44%   3401851   141744
Total                  35  15868387       100.00%


Model Summary

      S    R-sq  R-sq(adj)    PRESS  R-sq(pred)
441.565  68.05%     56.99%  9718988      38.75%
```

We see that by doing a multifactorial ANOVA, that finally the factor *Place* is statistically significant (p-value less than 5%), the factor *Speed* is not (p-value greater than 5%), the pressure has a very significant influence (almost zero p-value). The interaction *Place * Speed* is significant (p-value less than 5%). By cons, the interactions *Speed * Pressure* and *Place * Pressure* are not significant (p-value greater than 5%). The triple interaction *Place * Speed * Pressure* is significant!

So as we can see it, its not because a single given factor is not significant in the model without interactions that it will remain not significant when we take into account the interactions!!!

Remarks

R1. All the results that we have obtained above and also the charts can be obtained with a simple spreadsheet software like Microsoft® Excel®. So if the reader wants we can add the corresponding screenshots he must not hesitate, as always, to contact us.

R2. We can notice above that each factor F_i having n_i levels, has (obviously) $n_i - 1$ degrees of freedom, then we have a total of $(\prod_i F_i) - 1$ (including all type of interactions I), the degrees of freedom or residuals (error) will be $(n-1)\prod_i F_i$, where n is the number of replications, and the total degrees of freedom will be $(n\prod_i F_i) - 1$. Such that we have indeed:

$$\underbrace{\left[(n-1)\prod_i F_i\right]}_{\text{df errors (residuals)}} + \underbrace{\left[\left(\prod_i F_i\right) - 1\right]}_{\text{df due to factor and \underline{all} interactions}} = \underbrace{\left(n\prod_i F_i\right) - 1}_{\text{total df}} \tag{16.7.495}$$

To close this subject, remember that we talked earlier about the interest to randomize the order of the measurement to eliminate unknown confounders. In reality, we must consider three important cases of harmful factors in practice and which will only change a little bit the used ANOVA and that most modern statistical software implement nowadays:

1. We have "unknown harmful and uncontrollable factors". Therefore we make a usual ANOVA but simply where the order of measurement were randomized.
2. We have "known harmful factors but still uncontrollable". We speak then of "covariates" or "cofactors" and we use an ANCOVA (see section Statistics) instead of a simple ANOVA
3. We have "known harmful factors and that under control" that we want to eliminate from the ANOVA. The idea is to use a technique named "blocking" it is simply a hierarchical ANOVA (see section Statistics page 1065).

16.7.4.3 Taguchi Designs and Nomenclature (robust designs)

"Taguchi designs" also named "robust designs" or "Taguchi Orthogonal Array Designs" (abbreviated "Taguchi OA designs" in some softwares) are only a particular technique to find factorial or multifactorial fractional or full designs (from a matrix, a triangular table or from graph). In other words: Taguchi designs are a type of general fractional factorial design!

Definition (#16.7.739): An "orthogonal array" of s elements, denoted $\text{OA}_N(s^m)$, is a $N \times m$ matrix whose columns have the property that in every pair of columns each of the possible ordered pairs of elements appears the same number of time. Taguchi refers to $\text{OA}_N(s^m)$ by the notation $L_n(s^m)$ where L stands for "Level".

Below for example are given $\text{OA}_4(2^3)$ and $\text{OA}_8(2^7)$ respectively:

	Column number		
Row number	**1**	**2**	**3**
1	0	0	0
2	0	1	1
3	1	0	1
4	1	1	0

Table 16.39: $\text{OA}_4(2^3)$ Orthogonal array

	Column number						
Row number	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1	0	0	0	0	0	0	0
2	0	0	0	1	1	1	1
3	0	1	1	0	0	1	1
4	0	1	1	1	1	0	0
5	1	0	1	0	1	0	1
6	1	0	1	1	0	1	0
7	1	1	0	0	1	1	0
8	1	1	0	1	0	0	1

Table 16.40: $\text{OA}_8(2^7)$ Orthogonal array

With softwares that automatically generate the corresponding designs, this technique has become a bit old fashioned but it had the advantage at the time of its use to provide a list of over a hundred of tables with factors at 2 or more levels with or without interactions. Let us see some classics examples for the general culture (as it is something good to know and plus it's nice and still used in practice and in some softwares!!!!).

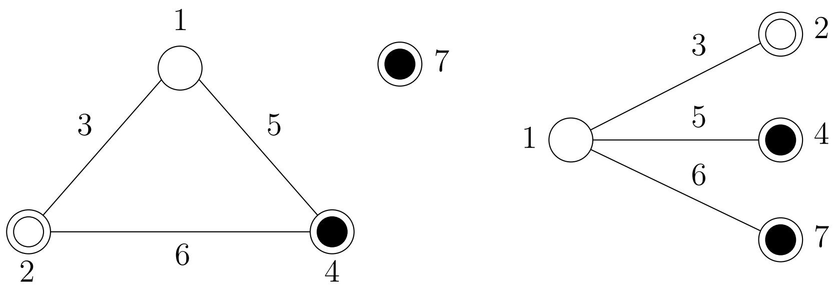
Dr. Genichi Taguchi has proposed to organize experiences following tables and graphs to identify interactions (the Taguchi tables include some Plackett-Burman designs) and afterwards to be able to find the settings for the controllable variables that minimize the variability transmitted to the response from the uncontrollable variable.

In the Taguchi OA design, only the main effects and two-factor interactions are considered, and higher-order interactions are assumed to be non-existent!!!! In addition, designers are asked to identify (based on their knowledge of the subject matter) which interactions might be significant before conducting the design.

Let us start with an example of table that we know well, the L_8 table. By Taguchi this table can be read as the table of a fractional factorial design of 7 factors (thus without interactions) or as the table of a full factorial design for 3 factors (hence with interactions!). This is why it is denoted $L_8(2^7)$ or $L_8(2^3)$ and represented by:

Trial N°	1	2	3	4	5	6	7
1	1	1	1	1	1	1	1
2	1	1	1	2	2	2	2
3	1	2	2	1	1	2	2
4	1	2	2	2	2	1	1
5	2	1	2	1	2	1	2
6	2	1	2	2	1	2	1
7	2	2	1	1	2	2	1
8	2	2	1	2	1	1	2

Table 16.41: L_8 **table with common factorial notation by Taguchi and associated graph**



We'll see if this table really differs from what we already know. First, each factor also takes two levels in this table, we can replace the Taguchi notation with our usage notation for the factorial replacing 1 by "$+$" and "2" by "$-$":

Trial N°	1	2	3	4	5	6	7
1	$+$	$+$	$+$	$+$	$+$	$+$	$+$
2	$+$	$+$	$+$	$-$	$-$	$-$	$-$
3	$+$	$-$	$-$	$+$	$+$	$-$	$-$
4	$+$	$-$	$-$	$-$	$-$	$+$	$+$
5	$-$	$+$	$-$	$+$	$-$	$+$	$-$
6	$-$	$+$	$-$	$-$	$+$	$-$	$+$
7	$-$	$-$	$+$	$+$	$-$	$-$	$+$
8	$-$	$-$	$+$	$-$	$+$	$+$	$-$

Table 16.42: L_8 **table with common notation**

We see already much better that all the rows and columns are orthogonal taken in pairs (Hadamard matrix). We add a column of "$+$":

Trial N°	1	2	3	4	5	6	7	**Rest**
1	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$
2	$+$	$+$	$+$	$-$	$-$	$-$	$-$	$+$
3	$+$	$-$	$-$	$+$	$+$	$-$	$-$	$+$
4	$+$	$-$	$-$	$-$	$-$	$+$	$+$	$+$
5	$-$	$+$	$-$	$+$	$-$	$+$	$-$	$+$
6	$-$	$+$	$-$	$-$	$+$	$-$	$+$	$+$
7	$-$	$-$	$+$	$+$	$-$	$-$	$+$	$+$
8	$-$	$-$	$+$	$-$	$+$	$+$	$-$	$+$

Table 16.43: L_8 table with common notation

and we notice that we fall back on the full factorial design with 3 factors at 2 levels if we rewrite the title row of the columns as follows:

Trial N°	$F3$	$F2$	$F23$	$F1$	$F13$	$F12$	$F123$	**Rest**
1	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$
2	$+$	$+$	$+$	$-$	$-$	$-$	$-$	$+$
3	$+$	$-$	$-$	$+$	$+$	$-$	$-$	$+$
4	$+$	$-$	$-$	$-$	$-$	$+$	$+$	$+$
5	$-$	$+$	$-$	$+$	$-$	$+$	$-$	$+$
6	$-$	$+$	$-$	$-$	$+$	$-$	$+$	$+$
7	$-$	$-$	$+$	$+$	$-$	$-$	$+$	$+$
8	$-$	$-$	$+$	$-$	$+$	$+$	$-$	$+$

Table 16.44: Previous table with traditional notation for the title row

and that this also corresponds to the fractional factorial design of a 7-factor experimental design at 2 levels (without interactions).

The question that will obviously ask the reader to himself is how we would have identify what factor belonged to which column in the context of a full factorial design with 3 factors if we did not know the table was prepared with the techniques view previously? Well in fact just by using the following linear graph:

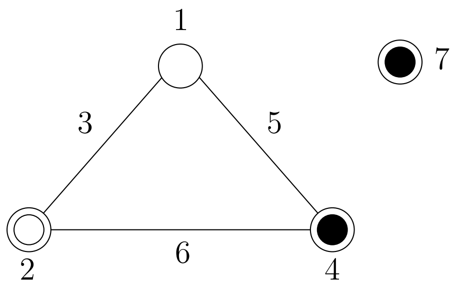
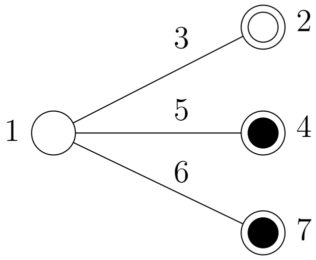


Figure 16.248: Taguchi L8 linear graph

that indicates the main factors in the vertices of the graph (therefore the columns 1, 2 and 4 are the main factors). The column 3 is the interaction between the columns 1 and 2 (hence the fact that it is on the edge between the two vertices), the column 5 is the interaction between the columns 1 and 4 (hence the fact that it is on the edge between the two vertices), and column 6 is the interaction between the columns 2 and 4 (hence the fact that it is on the edge between the two vertices). The fact that 7 is outside the linear graph outside is because a triple interaction can not be represented with the technique of planar graph. We then used the symbols of vertices to signify that this is the superposition of the vertices 1, 2 and 4 (circle + ring + disc).

The second linear graph associated with this table permits to highlight another potential use:



That is to say, we make use of it for an analysis of 4 factors (always with two modality in this case) represented by columns 1, 2, 4, 7 (the reader can indeed check that these four columns correspond to a fractional factorial design with for generator **12** for 4 factors either by hand or with software) with 3 interactions (1 and 2; 1 and 4, 1 and 7).

And what about the different shapes of the symbols on the linear graph??? There are given here:

Symbol	Description
○	Difficult, so the number of level changes in the runs order must be low.
◎	Quite difficult, so the number of changes in the runs order must be also quite low.
◉	Quite easy.
●	Easy, so frequent changes of the level of the factor are not a problem.

Remark

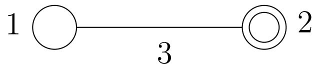
The practitioner must be careful because there are therefore two Taguchi tables denoted L_8: one that is a full table for 3 factors to 2 levels and one for 7 factors at 2 levels without interactions as we have just seen (so with 7 columns), but there also tables L_8 for 5 factors splitted into 4 factors at 3 levels and 1 factor to 4 levels (without interactions) but only with 5 columns. This is why books listing the tables Taguchi normally explicitly specify the scope of application of the given tables.

Let's see some other tables (a small part of the complete list):

- Table $L_4(2^3)$:

Trial N°	$F1$	$F2$	$F3$
1	1	1	1
2	1	2	2
3	2	1	2
4	2	2	1

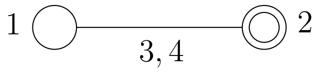
Table 16.45: Taguchi L_4 table



- Table $L_9(3^4)$:

Trial N°	$F1$	$F2$	$F3$	$F4$
1	1	1	1	1
2	1	2	2	2
3	1	3	3	3
4	2	1	2	3
5	2	2	3	1
6	2	3	1	2
7	3	1	3	2
8	3	2	1	3
9	3	3	2	1

Table 16.46: Taguchi L_9 table



- Table $L_{12}(2^{11})$:

Trial N°	1	2	3	4	5	6	7	8	9	10	11
1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	2	2	2	2	2	2
3	1	1	2	2	2	1	1	1	2	2	2
4	1	2	1	2	2	1	2	2	1	1	2
5	1	2	2	1	2	2	1	2	1	2	1
6	1	2	2	2	1	2	2	1	2	1	1
7	2	1	2	2	1	1	2	2	1	2	1
8	2	1	2	1	2	2	2	1	1	1	2
9	2	1	1	2	2	2	1	2	2	1	1
10	2	2	2	1	1	1	1	2	2	1	2
11	2	2	1	2	1	2	1	1	1	2	2
12	2	2	1	1	2	1	2	1	2	2	1

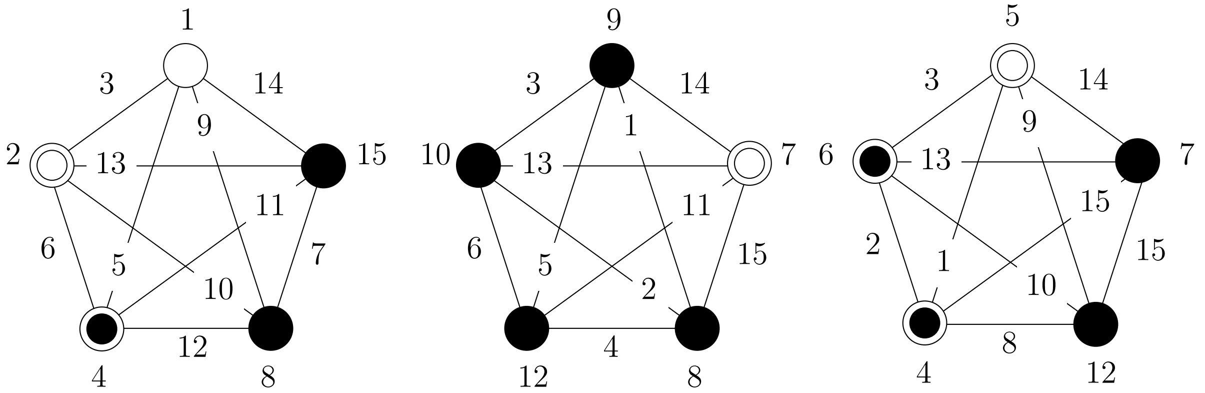
Table 16.47: L_{12} table with common factorial notation by Taguchi and associated graph

In this matrix of experiments, the interactions are distributed uniformly in all the columns. It therefore has no triangular table or linear graph, and can not be used to study interactions.

- $L_{16}(2^{15})$:

Trial N°	$F1$	$F2$	$F3$	$F4$	$F5$	$F6$	$F7$	$F8$	$F9$	$F10$	$F11$	$F12$	$F13$	$F14$	$F15$
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2
3	1	1	1	2	2	2	2	1	1	1	1	2	2	2	2
4	1	1	1	2	2	2	2	2	2	2	2	1	1	1	1
5	1	2	2	1	1	2	2	1	1	2	2	1	1	2	2
6	1	2	2	1	1	2	2	2	2	1	1	2	2	2	2
7	1	2	2	2	2	1	1	1	1	2	2	2	2	1	1
8	1	2	2	2	2	1	1	2	2	1	1	1	1	2	2
9	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
10	2	1	2	1	2	2	1	2	1	2	1	2	1	2	1
11	2	1	2	2	1	2	1	1	2	1	2	2	1	2	1
12	2	1	2	2	1	2	1	2	1	2	1	1	2	1	2
13	2	2	1	1	2	2	1	1	2	2	1	1	2	2	1
14	2	2	1	1	2	2	1	2	1	1	2	2	1	1	2
15	2	2	1	2	1	1	2	1	2	2	1	2	1	1	2
16	2	2	1	2	1	1	2	2	1	1	2	1	2	2	1

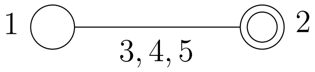
Table 16.48: Taguchi L_{16} table


1
3
14
9
2
13
15
6
11
5
7
10
12
4
8
9
3
14
1
10
13
7
11
6
5
15
2
4
12
8
5
3
14
9
6
13
7
15
2
1
15
10
8
4
12

- Table $L_{16}(4^5)$:

Trial N°	$F1$	$F2$	$F3$	$F4$	$F5$
1	1	1	1	1	1
2	1	2	2	2	2
3	1	3	3	3	3
4	1	4	4	4	4
5	2	1	2	3	4
6	2	2	1	4	3
7	2	3	4	1	2
8	2	4	3	2	1
9	3	1	3	4	2
10	3	2	4	3	1
11	3	3	1	2	4
12	3	4	2	1	3
13	4	1	4	2	3
14	4	2	3	1	4
15	4	3	2	4	1
16	4	4	1	3	1

Table 16.49: Taguchi L_{16} table



- etc.

The Taguchi tables together with their graphs are therefore full or fractional factorial designs with all their advantages. Taguchi's merit is to have tried to simplify the use of factorial designs to make them accessible to a large number of experimenters.

Taguchi's designs are usually highly fractionated, which makes them very attractive to practitioners. Doing a half-fraction, quarter-fraction or eighth-fraction of a full factorial design greatly reduces costs and time needed for a designed experiment. As we know, the drawback of a fractionated design is that some interactions may be confounded with other effects. It is important to consider carefully the role of potential confounders and aliases. Failure to take account of such confounded effects can result in erroneous conclusions and misunderstandings.

So, some Taguchi designs are primarily intended to study main effects of factors. Occasionally, you might want to study some of the two-way interactions. Some of the Taguchi designs (orthogonal arrays) let you study a limited number of two-way interactions. This usually requires that you leave some columns out of the array by not assigning factors to them. Some of the array columns are confounded with interactions between other array columns[27].

Taguchi interaction tables show confounded columns, which can help you assign factors to array columns. The columns and rows represent the column numbers of the Taguchi design

---

[27] Most DOE softwares choose automatically how to assign factors to array columns in a way that avoids confounding by default! Or, if you know exactly what design you want and know the columns of the entire array that correspond to the design, you can assign factors to array columns yourself. Assigning factors to columns of the array does not change in most softwares how the design looks in the on the computer screen.

(orthogonal array). Each table cell contains the interactions confounded for the two columns of the orthogonal array.

For example, the entry in cell below $(1,2)$ is 3. This means that the interaction between columns 1 and 2 is confounded with column 3. Thus, if you assigned factors A, B, and C to columns 1, 2, and 3, you could not study the AB interaction independently of factor C. If you suspect that there is a substantial interaction between A and B, you should not assign any factors to column 3 (hence you can assign factors A, B and C respectively to columns 1, 2 and 4). Similarly, the column 1 and 3 interaction is confounded with column 2, and the column 2 and 3 interaction is confounded with column 1.

	1	**2**	**3**	**4**	**5**	**6**	**7**
1		3	2	5	4	7	6
2			1	6	7	4	5
3				7	6	5	4
4					1	2	3
5						3	2
6							1

Table 16.50: $L_8(2^7)$ **Taguchi interaction table**

When using a Taguchi design, one needs to guess which interactions are most likely to be significant - even before any experiment is performed.

The Taguchi methodology is say to have generated considerable debate and controversy. Part of the controversy arose because Taguchi's methodology was advocated in the West initially (and primarily) by entrepreneurs, and the underlying statistical science had not been adequately peer-reviewed. By the late 11980s (holocene calendar) the results of a very comprehensive peer review indicated that although Taguchi's engineering concepts and the overall objective Taguchi's robust parameter design were well-founded [103], there were substantial problems with his experimental strategy (maximizing signal to noise does not necessarily minimize variability...) and methods of data analysis [301].

Remark

For example to see how this is vicious... Taguchi's catalogue contains 20 arrays. However, only 18 of these arrays are in reality orthogonal arrays...

Finally let summarize what we have seen so far with the respective names (this list is to take with precaution as there is a priori no clear international standard to our knowledge regarding these definitions) in the order of mathematical generalization:

- Designs that contain only factors at two levels, are referred to as "factorial design" that they are based on a linear or non-linear model, additive or not.
- Designs that have the column of the experimental matrix that have a zero dot product are named "orthogonal designs" (hence it exists designs that are named "non-orthogonal designs" but that are less easy to manipulate).
- The factorial designs that allow only to determine the coefficients of the main factors (i.e. additive model: without interactions) are referred to as "Koshal designs" or "screening

designs" (also named "one shot at a time designs" because as the technique of "one at a time" it can't analyse any interaction but the comparison stops there!).

- The factorial designs where all the interactions of three and higher order are neglected are referred to as "Rechtschaffner designs".
- The factorial designs where the coefficients are aliased while keeping interactions are referred to as "fractional factorial designs" or of "Box and Hunter designs" or even under the name of "matrix Hadamard based designs". With fractional factorial designs, it should also be specify the method of resolution.
- The factorial designs whose order is a multiple of 4 but not a power of 2 (i.e. 12, 16, 20, 24, etc.) are designated under the name of "Plackett-Burman designs"
- The designs with any number of factors, of interactions of any order and any number of levels are referred to as "Fisher designs" or "general full factorial designs" (in honour of to the original creator of the concept of experimental design).
- The designs where all factors have three levels and are based on the conferences matrices and permits to determine main factors without alias and at the same time with quadratic terms are named "definitive screening designs".
- The designs which contain as many tests (runs) as the number of coefficients to be determined are designated under the name of "saturated designs[28]".
- The fractional designs which are made to determine only main effects are named "saturated main effect designs" (typical of resolution III designs).
- The fractional designs which are made to determine only main effects and 2-way interactions are named "unsaturated designs" (typical of resolution V designs or greater).
- The designs which contain fewer tests than the number of coefficients to be determined are referred to as "oversaturated designs" or "supersaturated designs" (typical of A/D/G "optimal designs").
- There are other more exotic designs like α-designs, search designs (also named "resolution V.1 designs, or "resolution III.1 designs" depending on the case) and so on....

One aspect still needs to be clarified: it is the verification of the validity of the mathematical model of the first degree. None of these designs provides, as far as we know, such a validity test using elaborated statistics. This is why it is recommended to always add at least one test point at the center of the experimental range. The value of the response at this point will be compared to the deduced value of the other data points through mathematical model. If both values are similar, the mathematical model will be adopted if they are not we will have to reject this model and complete the results already obtained by experiments by going through a model of the second degree.

Finally if we have one response variable for the problem to take up for DOE then we have the following situations (options):

---

[28] This is quite a problem as then the model fits perfectly the data and we can therefore not make statistical adjustment tests.

1. The multiple responses are independent. Then we do our job as usually (as optimizing one under the independence assumption must not change the other responses)
2. The multiple responses are not independent. Then we add each model as a weighted sum and we optimize that latter (Overall Evaluation Criterion method as suggested in Taguchi designs) with any solver.
3. We use 2 stage least squares regression techniques (see section Numerical Methods) and after we optimize the whole with a solver.

16.7.4.3.1 Signal Factor

When we add a "signal factor" to an existing Taguchi design, the Taguchi design without the signal factor is named "static design" and that with the signal factor is named "dynamic design".

For example, if we add a signal factor with 2 levels to an existing $L_4(2^3)$ array, 4 rows (1 replicate of 4 runs) are added to the worksheet. If we add a signal factor with 3 levels, 8 rows (2 replicates of 4 runs) are added to the worksheet. A replicate is the entire set of runs from the static design as illustrated below:

Trial N°	1	2
1	$-$	$-$
2	$-$	$+$
3	$+$	$-$
4	$+$	$+$

Table 16.51: $L_4(2^3)$ **Taguchi static design**

Trial N°	1	2	Signal Factor
1	$-$	$-$	$-$
2	$-$	$+$	$-$
3	$+$	$-$	$-$
4	$+$	$+$	$-$
5	$-$	$-$	$+$
6	$-$	$+$	$+$
7	$+$	$-$	$+$
8	$+$	$+$	$+$

Table 16.52: $L_4(2^3)$ **Taguchi dynamic design with** 2**-level signal factor**

Trial N°	1	2	Signal Factor
1	$-$	$-$	1
2	$-$	$+$	1
3	$+$	$-$	1
4	$+$	$+$	1
5	$-$	$-$	2
6	$-$	$+$	2
7	$+$	$-$	2
8	$+$	$+$	2
5	$-$	$-$	3
6	$-$	$+$	3
7	$+$	$-$	3
8	$+$	$+$	3

Table 16.53: $L_4(2^3)$ **Taguchi dynamic design with** 3**-level signal factor**

16.7.4.4 Response Surface Methodology (Box Domains)

In statistics, "response surface methodology (RSM)" explores the relationships between several explanatory variables and one or more response variables. The method was introduced by G. E. P. Box and K. B. Wilson in 11951 (holocene calendar). The main idea of RSM is to use a sequence of designed experiments to obtain an optimal response. Box and Wilson suggest using a second-degree polynomial model to do this. They acknowledge that this model is only an approximation, but use it because such a model is easy to estimate and apply, even when little is known about the process.

An easy way to estimate a first-degree polynomial model is as we have seen to use a factorial experiment or a fractional factorial design based in the two variables case for recall by:

$$y \cong a_0 + \sum a_i x_i + \sum_{i \neq j} a_{ij} x_i x_j + e \tag{16.7.496}$$

This is sufficient to determine which explanatory variables affect the response variable(s) of interest. Once it is suspected that only significant explanatory variables are left, then a more complicated design, such as a central composite design can be implemented to estimate a second-degree polynomial model, which is still only an approximation at best given by:

$$y \cong a_0 + \sum a_i x_i + \sum a_{ij} x_i x_j + e = a_0 \sum a_i x_i \sum_{i \neq j} a_{ij} x_i x_j + \sum_i a_{ii} x_i x_i + e \tag{16.7.497}$$

However the second-degree model can be used to optimize the process (maximize, minimize, or attain a specific target for).

Obviously in the case of two variables, we speak of "response surface" (beyond we speak of a "volume" or "hypervolume") and then the previous relation will be written:

$$y = a_0 + a_1 x_2 + a_{12} x_1 x_2 + \underbrace{a_{11} x_1^2 + a_{22} x_2^2}_{\text{curvature}} \tag{16.7.498}$$

The designs including the following three terms:

$$a_{12} x_1 x_2 + \underbrace{a_{11} x_1^2 + a_{22} x_2^2}_{\text{curvature}} \tag{16.7.499}$$

are named "Doehlert's composite designs". We will have then to determine 4 terms:

- a_0: the constant coefficient
- a_i: coefficient of first degree
- a_{ij}: rectangle (quadratic) coefficient
- a_{ii}: pure quadratic coefficient

We see that in the case of a 2-factor design, we need then to estimate 6 coefficients:

$$y = \underbrace{a_0}_{1} + \underbrace{a_1}_{2} x_2 + \underbrace{a_2}_{3} + \underbrace{a_{12}}_{4} x_1 x_2 + \underbrace{a_{11}}_{5} x_1^2 + \underbrace{a_{22}}_{6} x_2^2 \tag{16.7.500}$$

and for 3 factors we will have to estimate 10 coefficients:

$$y=a_0+a_1x_1+a_2x_2+a_3x_3+a_{12}x_1x_2+a_{13}x_1x_3+a_{23}x_1x_3+a_{23}x_2x_3+a_{11}x_1^2+a_{22}x_2^2+a_{33}x_3^2 \tag{16.7.501}$$

Therefore after trials and errors we found that a Doehlert design has a cost (number of runs) that can be simply written as:

$$\boxed{N_{\text{Do}}=k^2+2k+N_0} \tag{16.7.502}$$

where k is the number of factors and N_0 the number of additional points.

The idea often adopted to pass to a model of the second degree after a factorial study while retaining the tests already carried out is simply to complete the existing design by the tests for the estimation of the model of higher degree when possible.

✍**Example:**

For example, a response surface associated with a relation like the previous one looks like with Maple™ 4.00b:

```
>5+3*x1+2*x2+4*x1*x2+2.5*x1^2+3*x2^2,x1=-10..10,x2=-10..10,
view=[-10..10,-10..10,0..50],contours=10,style=PATCHCONTOUR,
axes=frame,numpoints=10000);
```

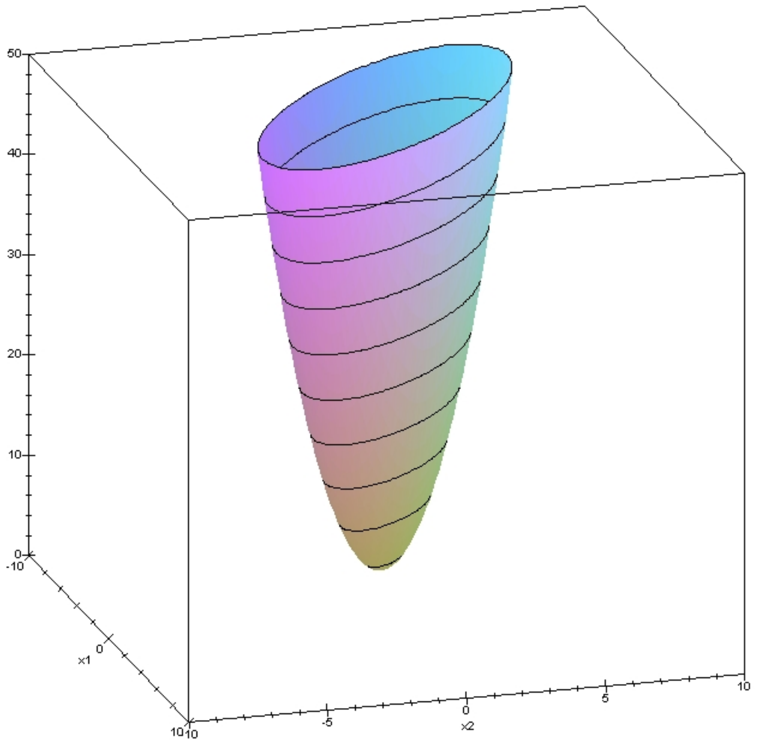


Figure 16.249: **Generic example of interaction for two factors quadratic polynomial**

16.7.4.4.1 Pure quadratic curvature test

In running a two-level factorial experiment, we usually anticipate fitting the first order model, but we should be alert to the possibility that the second order model (response surface methodology) is really more appropriate. Therefore before running response surface we should be aware of any curvature assumption!

There is a method for replicating certain points in a 2^k factorial that will provide protection against curvature from second-order effects as well as allow an independent estimate of error to be obtained. The method consist of adding center points to the 2^k design. These consist of n replicates run at the points $x_i = 0$. One important reason for adding the replicate runs at the design center is that center points do not affect the usual effect estimates in a 2^k design.

To illustrate the approach, consider 2^2 design, with one observation at each of the factorial points $(-,-)$, $(+,-)$, $(-,+)$, and $(+,+)$ and n_c observations at the center point $(0,0)$. Let $\bar{y}_F$ be the average of the four runs at the four factorial points and let $\bar{y}_C$ be the average of the n_c runs at the center point. If the difference:

$$\bar{y}_F - \bar{y}_C \tag{16.7.503}$$

is small then the center points lie on or near the plane passing through the factorial points, and there is no quadratic curvature. On the other hand, if this same difference is large, then quadratic curvature is present. A single-degree-of-freed sum of squares for pure quadratic curvature is given by (see proof just below):

$$\text{SS}_{\text{pure quadratic}} = \frac{n_F n_C (\bar{y}_F - \bar{y}_C)^2}{n_F + n_C} \tag{16.7.504}$$

where in general, n_F is the number of factorial design points. This quantity may be compared to the error mean square to test for pure quadratic curvature. More specifically, when point are added to the center of the 2^k design, then the test for curvature actually test the hypotheses the coefficients of the quadratic terms are all non zero):

$$\begin{aligned} H_0 :& \quad \sum_{j=1}^{k} \beta jj = 0 \\ H_1 :& \quad \sum_{j=1}^{k} \beta jj \neq 0 \end{aligned} \tag{16.7.505}$$

Or more explicitly:

$$\begin{aligned} H_0 :& \quad \beta_{11} + \beta_{22} + \ldots + \beta_{kk} = 0 \\ H_1 :& \quad \beta_{11} + \beta_{22} + \ldots + \beta_{kk} \neq 0 \end{aligned} \tag{16.7.506}$$

First let us prove that:

$$\text{E}(\bar{y}_F) = \beta_{11} + \beta_{22} + \ldots + \beta_{kk} \tag{16.7.507}$$

Indeed, let us recall the full model:

$$y_F=\beta_0+\beta_1x_1+\beta_2x_2+\beta_{12}x_1x_2+\beta_{11}x_1^2+\beta_{22}^2 \tag{16.7.508}$$

And let us rewrite if for the for the (x_1,x_2) combinations $(-1,+1)$, $(-1,-1)$, $(+1,+1)$, $(+1,-1)$:

$$\begin{aligned}y_{F_{(-1,+1)}}&=\beta_0-\beta_1+\beta_2-\beta_{12}+\beta_{11}+\beta_{22}\\y_{F_{(-1,-1)}}&=\beta_0-\beta_1-\beta_2+\beta_{12}+\beta_{11}+\beta_{22}\\y_{F_{(+1,-1)}}&=\beta_0+\beta_1-\beta_2-\beta_{12}+\beta_{11}+\beta_{22}\\y_{F_{(+1,+1)}}&=\beta_0+\beta_1+\beta_2-\beta_{12}+\beta_{11}+\beta_{22}\end{aligned} \tag{16.7.509}$$

We see immediately that the sum gives:

$$\sum y_F=4\beta_0+4\beta_{11}+4\beta_{22} \tag{16.7.510}$$

Therefore the average is (the sum divided by the number of corresponding measurement points):

$$\bar{y}_F=\beta_0+\beta_{11}+\beta_{22} \tag{16.7.511}$$

and in general:

$$\mathrm{E}(y_F)=\bar{y}_F=\beta_0+\beta_{11}+\beta_{22}+\ldots+\beta_{kk} \tag{16.7.512}$$

It is immediate that for the center point $(0,0)$ we have:

$$\mathrm{E}(y_C)=\bar{y}_C=\beta_0 \tag{16.7.513}$$

Therefore:

$$\mathrm{E}(\bar{y}_F-\bar{y}_C)=\beta_{11}+\beta_{22}+\ldots+\beta_{kk} \tag{16.7.514}$$

Now the variance of $\bar{y}_F-\bar{y}_C$ is under the assumption of equality of variances and independence:

$$\mathrm{V}(\bar{y}_F-\bar{y}_C)=\mathrm{V}(\bar{y}_F)+\mathrm{V}(\bar{y}_C)=\frac{\sigma^2}{n_F}+\frac{\sigma^2}{n_C}=\sigma^2\left(\frac{1}{n_F}+\frac{1}{n_C}\right) \tag{16.7.515}$$

Consequently a hypotheses can be conducted using the homoscedastic Student T-test (see section Statistics):

$$T_{n-2}=\frac{\bar{y}_F-\bar{y}_C}{\sqrt{\hat{\sigma}^2\left(\dfrac{1}{n_F}+\dfrac{1}{n_C}\right)}} \tag{16.7.516}$$

with $n=n_F+n_C$.

As we have proved in the section Statistics (see page 859), we have $T_{n-2}^2=F_{1,n-2}$ (it is convenient in statistical packages to put the curvature test in the ANOVA table where all other tests are based on Fisher distribution). Hence:

$$F_{1,n-2}=\frac{(\bar{y}_F-\bar{y}_C)^2}{\hat{\sigma}^2\left(\dfrac{1}{n_F}+\dfrac{1}{n_C}\right)} \tag{16.7.517}$$

and therefore:

$$\boxed{F_{1,n-2}=\text{SS}_{\text{pure quadtratic}}=\frac{n_F n_C(\bar{y}_F-\bar{y}_C)^2}{n_F+n_C}} \tag{16.7.518}$$

This finish our proof of the "pure quadratic curvature test" and notice that this doesn't give us the factors that are responsible of the curvature!

The reader can find an application example of this test in our Minitab companion book.

16.7.4.4.2 Box-Wilson Central Composite Designs

Let us recall that so far we have seen that according to the principle of analysis of factorial experimental designs that in the 1D case (a single explanatory variable), the interpolated point must have a left measurement point (-1) and a right measurement point $(+1)$. In the 2D case (two explanatory variables) it is natural to think then to have at least 3 points around the interpolated point such that we have a triangle whose barycenter is the point to be interpolated. Identically in the 3D case (three explanatory variables) it is natural to think of having at least 4 points surrounding the point to be interpolated and forming a cube or tetrahedron as needed. Thus, by generalizing the reasoning, in an n-dimensional space we will need at least $n+1$ points to construct a simplex around the point to be interpolated.

Now let us recall that we saw in the section of Theoretical Computing different techniques of linear regression (see page 6316) and we also studied the fact that a polynomial could be analysed with a simple linear regression (see page 6327). We obtained in the same section (during the study of the factor of inflation of the variance at page 6398) the following relation with its abusive notation ... dangerously confusing matrix of covariance and simple variance:

$$\text{cov}\left(\hat{\vec{\beta}}\right)=\text{V}\left(\hat{\vec{\beta}}\right)=\sigma_\varepsilon^2(X_c^TX_c)^{-1} \qquad (16.7.519)$$

with for recall:

$$\sigma_\varepsilon^2=\frac{1}{n-k}\sum(Y_i-y_i)^2 \qquad (16.7.520)$$

We have therefore to obtain a scalar from the explanatory values:

$$\text{V}(\hat{y})=\sigma_\varepsilon^2\vec{x}^T(X_c^TX_c)^{-1}\vec{x} \qquad (16.7.521)$$

And therefore:

$$\sigma_{\hat{y}}=\sigma_\varepsilon\sqrt{\vec{x}^T(X_c^TX_c)^{-1}\vec{x}} \qquad (16.7.522)$$

Experts in the field routinely write the latter relation in the following form and name it "variance function":

$$\text{V}(\hat{y})\frac{N}{\sigma_\varepsilon^2}=\frac{1}{N}\vec{x}^T(X_c^TX_c)^{-1}\vec{x} \qquad (16.7.523)$$

The idea now is to come back to our experimental designs is therefore to choose our measurement points which minimizes the standard deviation of the predictions introduced just previously!

Let us consider as an example the problem of finding a linear regression to the polynomial:

$$y=a_0+a_1x_1+a_2x_2 \qquad (16.7.524)$$

in the square $-1\leq x_1\leq 1,-1\leq x_2\leq 1$ and let us compare the maximum value of the predictive variance for the case of a complete factorial design (i.e. a square in the plane) and then for a fractional factorial design (i.e. a triangle in the plane) that omits the point $(1,1)$.

In the first case of the complete factorial design, we have therefore (see section Linear Algebra page 1951 and page 1967):

$$X=\begin{bmatrix}1&-1&-1\\1&-1&1\\1&1&-1\\1&1&1\end{bmatrix}\qquad X^TX=\begin{bmatrix}4&0&0\\0&4&0\\0&0&4\end{bmatrix}\qquad (X^TX)^{-1}=\frac{1}{4}\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix} \tag{16.7.525}$$

and therefore:

$$\vec{x}^T(X_c^TX_c)^{-1}\vec{x}=(1\,x_1\,x_2)\left(\frac{1}{4}\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}\right)\begin{pmatrix}1\\x_1\\x_2\end{pmatrix}=\frac{1}{4}(1+x_1^2+x_2^2) \tag{16.7.526}$$

What is written more frequently:

$$\vec{x}^T(X_c^TX_c)^{-1}\vec{x}=\frac{1}{4}(1+R^2) \tag{16.7.527}$$

where R is the distance (radius) to the central point. When a predictive variance depends only on the distance R, we say that we have a "rotation invariant experimental design" or simply a "rotational experimental design" (the variance function is then say to be "spherical"). This type of plan is attractive to practitioners who do not know a priori where in the space of experimentation the most precise forecasts must be found!

Therefore:

$$\sigma_{\hat{y}}=\sigma_\varepsilon\sqrt{\frac{1}{4}(1+R^2)}=\frac{\sigma_\varepsilon}{2}\sqrt{1+R^2} \tag{16.7.528}$$

The variance has then for value at the point of origin $(0,0)$:

$$\sigma_{\hat{y}}=\frac{\sigma_\varepsilon}{2} \tag{16.7.529}$$

And:

$$\sigma_{\hat{y}}=\sigma_\varepsilon\frac{\sqrt{3}}{2} \tag{16.7.530}$$

at the ends. This case thus shows a small variation between the origin and the extremities (a variation exactly of $\sqrt{3}$).

In the second case of the fractional factorial design we have:

$$X=\begin{bmatrix}1&-1&-1\\1&-1&1\\1&1&-1\end{bmatrix}\qquad X^TX=\begin{bmatrix}3&-1&-1\\-1&3&-1\\-1&-1&3\end{bmatrix}\qquad (X^TX)^{-1}=\frac{1}{4}\begin{bmatrix}2&1&1\\1&2&1\\1&1&2\end{bmatrix} \tag{16.7.531}$$

and therefore:

$$\begin{aligned}\vec{x}^T(X_c^TX_c)^{-1}\vec{x}&=(1\,x_1\,x_2)\left(\frac{1}{4}\begin{bmatrix}2&1&1\\1&2&1\\1&1&2\end{bmatrix}\right)\begin{pmatrix}1\\x_1\\x_2\end{pmatrix}=\frac{1}{2}(1+x_1+x_2+x_1^2+x_2^2+x_1x_2)\\&=\frac{1}{2}(1+x_1+x_2+R^2+x_1x_2)\end{aligned} \tag{16.7.532}$$

Therefore:

$$\sigma_{\hat{y}} = \sigma_\varepsilon\sqrt{\frac{1}{2}(1+x_1+x_2+R^2+x_1x_2)} \tag{16.7.533}$$

So first this design is by definition not invariant by rotation since the variance does not depend only on R!

The variance corresponding to the point of origin $(0,0)$ is then:

$$\sigma_{\hat{y}} = \frac{\sigma_\varepsilon}{\sqrt{2}} \tag{16.7.534}$$

and at the end $(1,1)$ we have:

$$\sigma_{\hat{y}} = \sqrt{3}\sigma_\varepsilon \tag{16.7.535}$$

If we now compute by the partial derivatives the coordinates which minimizes the error such that:

$$\frac{\partial\sigma_{\hat{y}}}{\partial x_1} = \frac{\partial\sigma_\varepsilon\sqrt{\frac{1}{2}(1+x_1+x_2+R^2+x_1x_2)}}{\partial x_1} = 0 \tag{16.7.536}$$

We get:

$$\frac{1}{2}\frac{\frac{1}{2}\sigma_\varepsilon(1+2x_1+x_2)}{\sqrt{\frac{1}{2}(1+x_1+x_2+R^2+x_1x_2)}} \tag{16.7.537}$$

Hence:

$$x_1 = \frac{-x_2-1}{2} \tag{16.7.538}$$

and therefore:

$$\left.\begin{aligned} x_1 &= \frac{-x_2-1}{2} \\ x_2 &= \frac{-x_1-1}{2} \end{aligned}\right\} \Rightarrow x_1 = \frac{\frac{x_1+1}{2}-1}{2} = \frac{x_1-1}{4} \tag{16.7.539}$$

Therefore it comes:

$$(x_1,x_2) = \left(-\frac{1}{3},-\frac{1}{3}\right) \tag{16.7.540}$$

is the point of minimum variance and at this point (which coincides with the barycenter of the triangle, that is to say the center of the inscribed circle), we have:

$$\sigma_{\hat{y}} = \frac{\sigma_\varepsilon}{\sqrt{3}} \tag{16.7.541}$$

From this example, we can already easily conclude that for a complete factorial design, the variance of the predictive error is written:

$$\sigma_{\hat{y}} = \frac{\sigma_\varepsilon}{2^{n/2}}\sqrt{1+x_1^2+x_2^2+\ldots+x_n^2} = \frac{\sigma_\varepsilon}{2^{n/2}}\sqrt{1+R^2} \tag{16.7.542}$$

and as well as the maximum error is immediate and is equal in the case of a complete factorial design:

$$\sigma_{\hat{y},\max}=\sigma_\varepsilon\sqrt{\frac{1+n}{2^n}} \tag{16.7.543}$$

Thus, the quality of the fit (regression) becomes very good when n grows!

Let us now consider now the fractional factorial design based on three points in the plane forming an equilateral triangle:

$$X=\begin{bmatrix}1 & \sqrt{3/2} & -1/\sqrt{2}\\ 1 & -\sqrt{3/2} & -1/\sqrt{2}\\ 1 & 0 & \sqrt{2}\end{bmatrix} \tag{16.7.544}$$

Similar calculations to before give us (as always on readers request we can detail!):

$$\sigma_{\hat{y}}=\sigma_\varepsilon\sqrt{\frac{1}{3}(1+x_1^2+x_2^2)}=\sigma_\varepsilon\sqrt{\frac{1}{3}(1+R^3)} \tag{16.7.545}$$

Thus, we notice that, by judiciously choosing the points of the design, we may have a lower prediction error than with a complete factorial design both at the origin and at the points $(1,1)$, $(-1,1)$, $(1,-1)$ and $(-1,-1)$. However the price to pay is that we have to make measurements outside the unit square which complicates the task of the practitioner.

16.7.4.4.2.1 Circumscribed Center Designs

The "Composite Center Design (CCD)" is a two-level experimental design for surface planes. In the 2D case (two explanatory variables) this is equivalent to taking the usual points of the squares to which we add the points of intersection of the axes with the circumscribed circle (+ in practice one or more central control points):

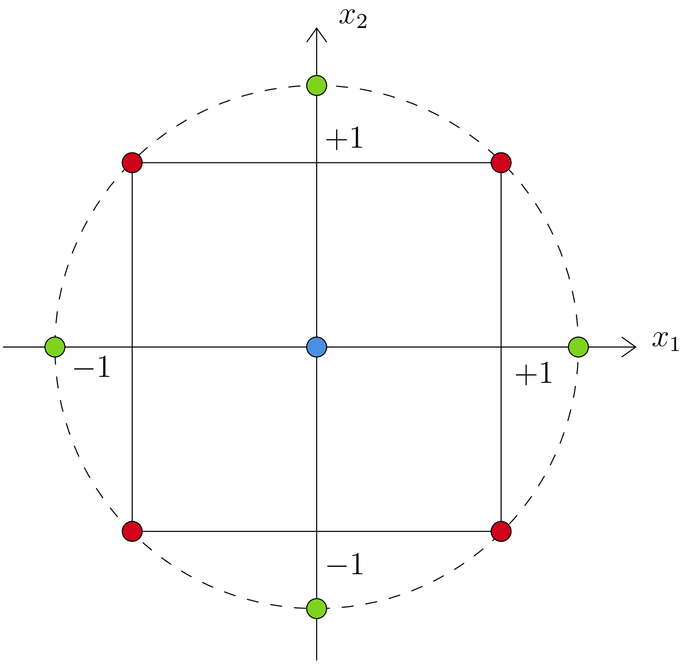


Figure 16.250: **2D Composite Center Design**

We thus have the following points, which are all at the same distance from the origin of the axes:

$$(1,1),(-1,1),(-1,1),(1,-1),(\sqrt{2},0),(0,\sqrt{2}),(-\sqrt{2},0),(0,-\sqrt{2}) \tag{16.7.546}$$

where of course the root of 2 comes from the fact that the radius of the circle is given by the vertices of the square and thus by stupidly applying Pythagoras:

$$R=\sqrt{1^2+1^2} \tag{16.7.547}$$

Due to the fact that in the above case the circle (in 2D) or the sphere (in 3D) is circumscribed to the square (or the cube in the 3D case) we also talk about CCC design for "Circumscribed Center Design (CCD)".

Let's take the example of the CCD case where the columns in the matrix below, the first is filled only with 1, the other columns corresponding respectively to the terms: x_1,x_2,x_1x_2, x_1^2, x_2^2 (notice that not all columns are orthogonal!):

$$X=\begin{bmatrix}1 & -1 & -1 & 1 & 1 & 1\\ 1 & -1 & 1 & -1 & 1 & 1\\ 1 & 1 & -1 & -1 & 1 & 1\\ 1 & 1 & 1 & 1 & 1 & 1\\ 1 & \sqrt{2} & 0 & 1 & 2 & 0\\ 1 & -\sqrt{2} & 0 & 0 & 2 & 0\\ 1 & 0 & \sqrt{2} & 0 & 0 & 2\\ 1 & 0 & -\sqrt{2} & 0 & 0 & 2\\ 1 & 0 & 0 & 0 & 0 & 0\end{bmatrix} \tag{16.7.548}$$

where the last row correspond to the fact of adding a control point a the center of the domain. This empirical matrix is sometimes named "composite equiradial matrix", it request as we see a strong symmetry and therefore implies what we named an "isovariance" (see plots with Maple™ further below of the variance).

The previous matrix is more commonly written in the following design matrix form (where the two columns are orthogonal):

$$D=\begin{bmatrix}-1 & -1\\ -1 & 1\\ 1 & -1\\ 1 & 1\\ \alpha & 0\\ -\alpha & 0\\ 0 & \alpha\\ 0 & -\alpha\\ 0 & 0\end{bmatrix} \tag{16.7.549}$$

So we can see that this design is such that for k factors we have:

- A factorial "cube" part consisting of 2^k points from a full factorial where the levels of each factor are coded as -1 and $+1$

- An "axial" or "star" part (named "star points") consisting of $2k$ points on the axis of each factor at a distance α from the center of the design
- At least one center point

We also see that the factors have 5 levels $(-1,1,-\sqrt{2},+\sqrt{2},0)$. This is a characteristic of CCD designs!

Therefore:

$$X^TX=\begin{bmatrix} 9 & 0 & 0 & 0 & 8 & 8 \\ 0 & 8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 & 0 \\ 8 & 0 & 0 & 0 & 12 & 4 \\ 8 & 0 & 0 & 0 & 4 & 12 \end{bmatrix} \tag{16.7.550}$$

Using a CAD software we get:

$$(X^TX)^{-1}=\begin{bmatrix} 1 & 0 & 0 & 0 & -1/2 & -1/2 \\ 0 & 1/8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 0 & 0 \\ -1/2 & 0 & 0 & 0 & 11/32 & 7/32 \\ -1/2 & 0 & 0 & 0 & 7/32 & 11/32 \end{bmatrix} \tag{16.7.551}$$

Then we have:

$$\begin{aligned}\sigma_{\hat{y}}&=\sigma_{\varepsilon}\sqrt{\vec{x}^T(X_c^TX_c)^{-1}\vec{x}}\\ &=\sigma_{\varepsilon}\sqrt{\begin{pmatrix}1 & x_1 & x_2 & x_1x_2 & x_1^2 & x_2^2\end{pmatrix}(X_c^TX_c)^{-1}\begin{pmatrix}1\\ x_1\\ x_2\\ x_1x_2\\ x_1^2\\ x_2^2\end{pmatrix}}\\ &=\sigma_{\varepsilon}\sqrt{\left(1-\frac{7}{8}x_1^2-\frac{7}{8}x_2^2+\frac{22}{32}x_1^2x_2^2+\frac{11}{32}x_1^4+\frac{11}{32}x_2^4\right)}\\ &=\sigma_{\varepsilon}\sqrt{1-\frac{7}{8}(x_1^2+x_2^2)+\frac{11}{32}(x_1^2+x_2^2)^2}\\ &=\sigma_{\varepsilon}\sqrt{1-\frac{7}{8}+\frac{11}{32}R^4}\end{aligned} \tag{16.7.552}$$

We therefore have here an experimental design whose variance depends indeed only on the distance R from the origin as many textbooks state but without proof.

Also at the opposite of 99% of textbooks we will give the Maple™ 4.00b code to plot the famous variance corresponding figure:

```
>contourplot3d(sqrt(1-7/8*(x^2+y^2)+11/32*(x^2+y^2)^2),
x=-sqrt(2)..sqrt(2),y=-sqrt(2)..sqrt(2),contours=10,filled=true);
```

We have:

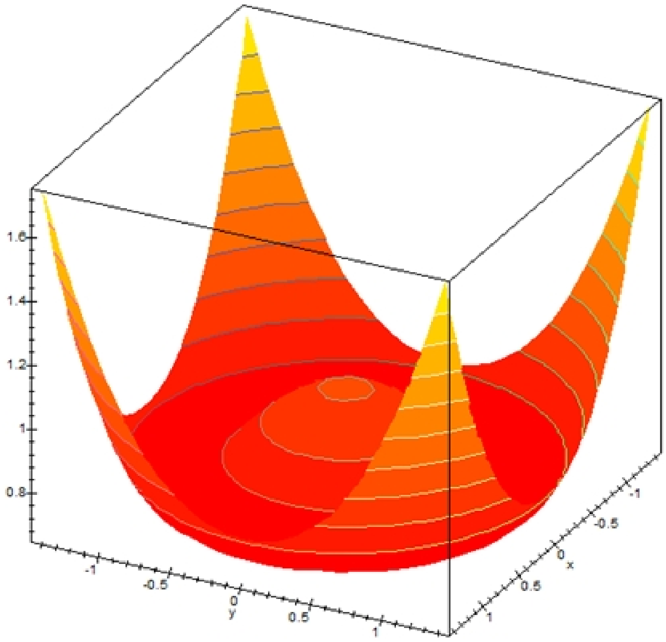


Figure 16.251: **2D Composite Center Design variance perspective plot in Maple™ 4.00b**

Either seen from above we get the corresponding famous chart available in many textbooks regarding to the rotational invariance property of surface design with two factors:

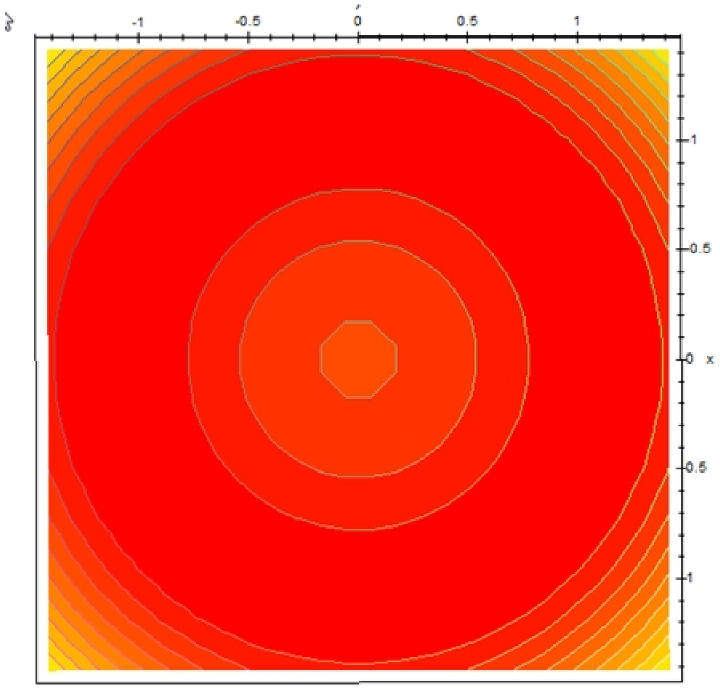


Figure 16.252: **2D Composite Center Design variance top plot in Maple™ 4.00b**

We can see that CCD have a number cost (run) of:

$$\boxed{N_{\text{CCD}}=2^{k-r}+2k+N_0} \tag{16.7.553}$$

where k is the number of factors, r the reduction integer and N_0 the number of center points. Therefore the three terms can be read as following:

- 2^{k-r}: Common full factorial of fraction factorial design of the composite equiradial matrix.
- $2k$: Number of star points
- N_0: Number of center points

Therefore for a typical two factors design with a full factorial subset we have:

$$N_{\text{CCD}}=2^{2-0}+2\cdot 2+N_0=2^2+4+N_0=8 \tag{16.7.554}$$

and as 5 is recommended for N_0, it comes that $N_{\text{CCD}}=13$ for a two factor response surface design with 5 center points.

16.7.4.4.2.2 Face Centered Designs

Repeating the same maneuver but with an experience design said to be "face-centered design" such as:

$$X=\begin{bmatrix}1&-1&-1&1&1&1\\1&-1&1&-1&1&1\\1&1&-1&-1&1&1\\1&1&1&1&1&1\\1&1&0&1&1&0\\1&-1&0&-1&1&0\\1&0&1&0&0&1\\1&0&-1&0&0&1\\1&1&0&0&1&0\end{bmatrix}\tag{16.7.555}$$

Then we have:

$$X^TX=\begin{bmatrix}10&1&1&0&7&7\\1&7&0&2&1&0\\1&0&7&0&0&1\\0&2&0&6&0&0\\7&1&0&0&7&4\\7&0&1&0&4&7\end{bmatrix}\tag{16.7.556}$$

And:

$$\left(X^TX\right)^{-1}=\begin{bmatrix}\frac{257}{270}&-\frac{1}{18}&-\frac{1}{20}&\frac{1}{54}&-\frac{3}{5}&-\frac{65}{108}\\-\frac{1}{18}&\frac{1}{6}&0&-\frac{1}{18}&0&\frac{1}{18}\\-\frac{1}{20}&0&\frac{3}{20}&0&\frac{1}{20}&0\\\frac{1}{54}&-\frac{1}{18}&0&\frac{5}{27}&0&-\frac{1}{54}\\-\frac{3}{5}&0&\frac{1}{20}&0&\frac{3}{5}&\frac{1}{4}\\-\frac{65}{108}&\frac{1}{18}&0&-\frac{1}{54}&\frac{1}{4}&\frac{65}{108}\end{bmatrix}\tag{16.7.557}$$

We then have after:

$$\begin{aligned}\sigma_{\hat{y}}&=\sigma_\varepsilon\sqrt{\vec{x}(X_c^TX_c)^{-1}\vec{x}}\\&=\sigma_\varepsilon\sqrt{\begin{gathered}\frac{257}{270}-\frac{1}{18}x_1-\frac{1}{20}x_2+\frac{1}{54}x_1x_2-\frac{3}{5}x_1^2-\frac{65}{108}x_2^2\\+x_1\left(-\frac{1}{18}+\frac{1}{6}x_1-\frac{1}{18}x_1x_2+\frac{1}{18}x_2^2\right)+x_2\left(-\frac{1}{20}+\frac{3}{20}x_2-\frac{1}{20}x_1^2\right)\\+x_1x_2\left(\frac{1}{54}-\frac{1}{18}x_1+\frac{5}{27}x_1x_2-\frac{1}{54}x_2^2\right)+x_1^2\left(-\frac{3}{5}+\frac{1}{20}x_2+\frac{3}{5}x_1^2+\frac{1}{4}x_2^2\right)\\+x_2^2\left(-\frac{65}{108}+\frac{1}{18}x_1-\frac{1}{54}x_1x_1+\frac{1}{4}x_1^2+\frac{65}{108}x_2^2\right)\end{gathered}}\end{aligned}\tag{16.7.558}$$

Notice that the Circumscribed Center Design (CCC) was a rotatable design but that the Face Centered Design (CCF) is not as it does not depends only on R.

Which gives us with Maple™ 4.00b by proceeding exactly with the same above given code:

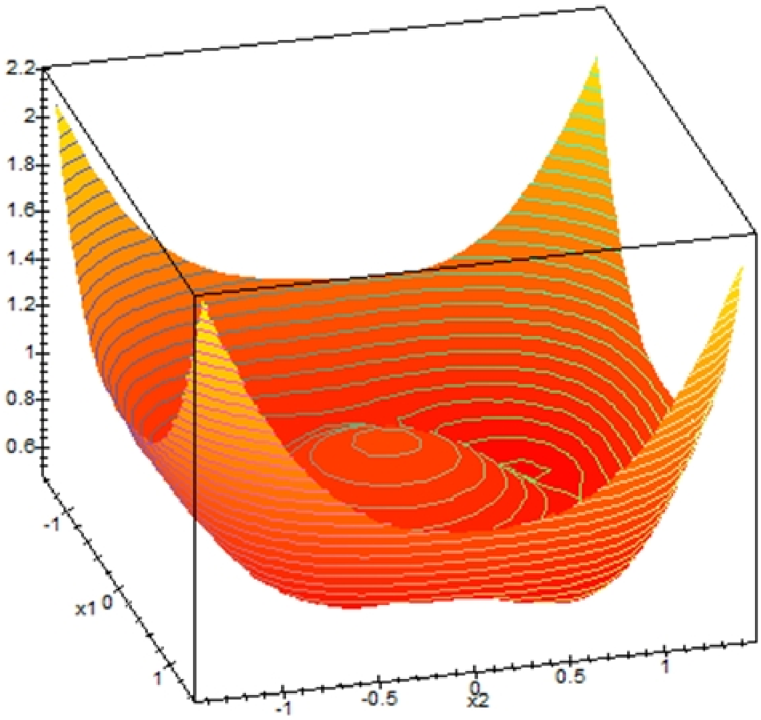


Figure 16.253: **2D Faced-centered Design variance perspective plot in Maple™ 4.00b**

Either seen from above we get the corresponding famous chart available in many textbooks:

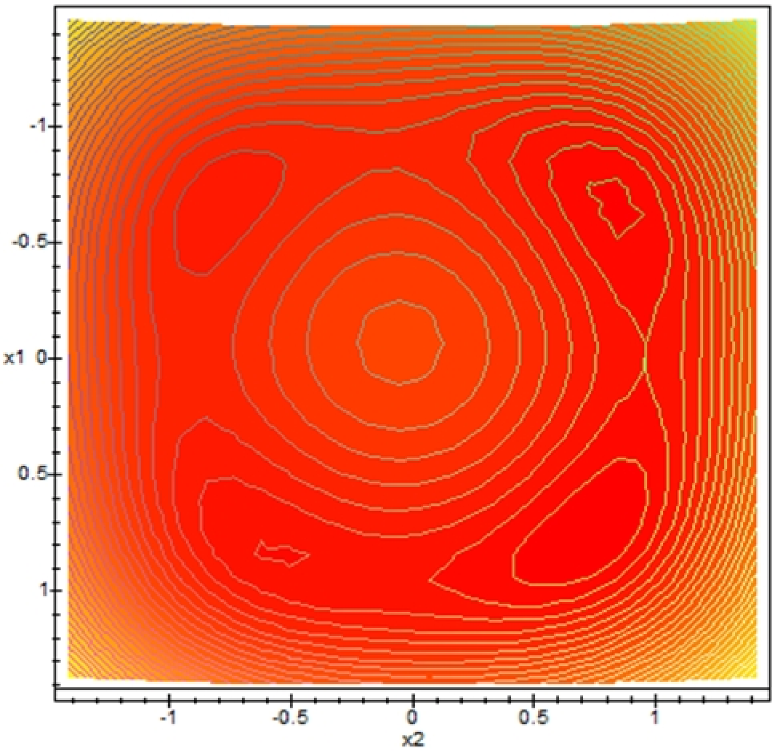
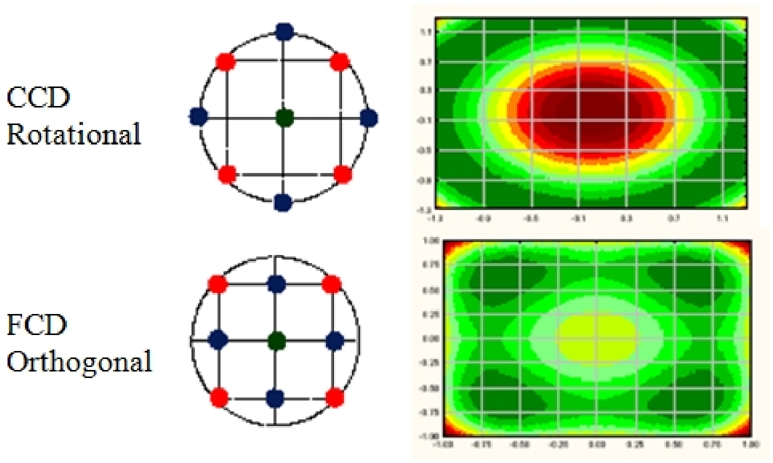


Figure 16.254: **2D Faced-centered Design variance top plot in Maple™ 4.00b**

The two previous designs are frequently summarized with the following figure:



In fact there are three Composite Central Design:

1. Circumscribed (CCD)
2. Inscribed (ICD)
3. Face centered (FCD)

Response surface methodology is obviously often applied to pilot plant operations by research and development personnel but when it is applied to a full-scale production process, it is usually only done once (or very infrequently) because the experimental procedure is relatively elaborate. However, conditions that were optimum for the pilot plant may not be optimum for the full-scale process. Indeed, the operation "scale-up" of the pilot plant to the full-scale production process usually results in distortion of the optimum conditions. Even if the full-scale plant begins operation at the optimum, it will eventually’ ’drift" away from that point because of variations in raw materials, environmental changes, and operating personnel.

A method is needed for the continuous monitoring and improvement of a full-scale process with the goal of moving the operating conditions toward the optimum or following a "drift." The method should not require large or sudden changes in operating conditions that might disrupt production. "Evolutionary operation (EVOP)" was proposed by Box (11957 according to holocene calendar) as such an operating procedure. It is designed as a method of routine plant operation that is carried out by manufacturing personnel with minimum assistance from the research and development staff.

EVOP consists of systematically introducing small changes in the levels of the operating variables under consideration. Usually, a 2^k design is employed to do this. The changes in the variables are assumed to be small enough that serious disturbances in yield, quality, or quantity will not occur, yet large enough that potential improvements in process performance will eventually be discovered. Data are collected on the response variables of interest at each point of the 2^k design. When one observation has been taken at each design point, a cycle is said to have been completed. The effects and interactions of the process variables may then be computed. Eventually, after several cycles, the effect of one or more process variables or their interactions may appear to have a significant effect on the response. At this point, a decision may be made to change the basic operating conditions to improve the response. When improved conditions have been detected, a phase is said to have been completed.

In testing the significance of process variables and interactions, an estimate of experimental error is required. This is calculated from the cycle data. Also, the 2^k design is usually centered about the current best operating conditions. By comparing the response at this point with the 2^k points in the factorial portion, we may check on curvature or change in mean (CIM); that is, if the process is really centered at the maximum, say, then the response at the center should be significantly greater than the responses at the 2^k-peripheral points.

In theory, EVOP can be applied to k process variables. In practice, only two or three variables are usually considered. We will give an example of the procedure for two variables. Box and Draper (11969 according to holocene calendar) give a detailed discussion of the three-variable case, including necessary forms and worksheets. Myers and Montgomery (11995 according to holocene calendar) discuss the computer implementation of EVOP.

16.7.4.5 Optimal Designs

In the design of experiments, "optimal designs" are a class of experimental designs that are optimal with respect to some statistical criterion. The creation of this field of statistics has been credited to Danish statistician Kirstine Smith.

In the design of experiments for estimating statistical models, optimal designs allow parameters to be estimated without bias and with minimum variance. A non-optimal design requires a greater number of experimental runs to estimate the parameters with the same precision as an optimal design. In practical terms, optimal experiments can reduce the costs of experimentation.

The optimality of a design depends on the statistical model and is assessed with respect to a statistical criterion, which is related to the variance-matrix of the estimator. Specifying an appropriate model and specifying a suitable criterion function both require understanding of statistical theory and practical knowledge with designing experiments.

We will begin here with the study of "D-optimal design" which is a form of experimental design provided by a computer algorithm for which we will give the mathematical background a little further below. These types of computer-aided designs are particularly useful when conventional plans do not apply.

Unlike standard classical models such as factorial design and fractional factorial designs or Taguchi designs, D-optimal designs are generally not orthogonal but are very flexible, hence their great practical importance.

D-optimal plans are always an option whatever the degree of the model the experimenter wishes to use (eg first order, second order with or without interactions, complete quadratic, cubic, etc.).

The D-optimal designs are associated with a regression model based on the fact that since in the case of the multiple linear regression we have (see section Numerical Methods page 6379):

$$\mathrm{V}\left(\hat{\vec{\beta}}\right)=\sigma_{\varepsilon}^{2}(X^{T}X)^{-1} \qquad (16.7.559)$$

then we must be able to find the component values of the matrix X which minimize the variances (the elements of the diagonal of this matrix as we have proved in the section of Theoretical Computing!). As this optimization operation is tricky, it is necessary to optimize a scalar value relative to the dispersion matrix $(X^{T}X)^{-1}$.

If we write:

$$(X^{T}X)^{-1}=A^{-1} \qquad (16.7.560)$$

Then let us recall that we have proved in the section of Linear Algebra (see page 1980) that the components of the inverse of the matrix A are given by:

$$b_{ij}=\frac{1}{\det(A)}(-1)^{i+j}\det(A_{ji}) \qquad (16.7.561)$$

Therefore minimize all the components is equivalent as to maximize the determinant of A and hence to maximize the determinant of $(X^{T}X)^{-1}$. It is for this reason that we speak of D-optimal designs for "Determinant-optimal designs" because trying to minimize the diagonal components (and in general all components) of the variance-covariance matrix is equivalent to seeking to maximize the determinant of $(X^T X)$:

$$\boxed{\max\ \det(X^T X)} \tag{16.7.562}$$

Another approach of minimizing the diagonal components of the variance-covariance matrix can be in choosing the scalar value to minimize, and therefore to minimize the trace (or the Average of the trace) of the information matrix:

$$\boxed{\min\ \text{tr}\left((X^T X)^{-1}\right)} \tag{16.7.563}$$

and we then speak of an "A-optimal design" which is numerically more expensive in algorithmic terms than the search for the D-optimal plan because it is necessary to reverse the information matrix but it is a criterion that is often considered more natural.

However, the majority of software specialized in experimental designs propose to determine the D-optimal or A-optimal designs but there are many other possible criteria that we will not present here such as E-optimality, G-optimality, etc.

Obviously before that the D or A-optimal design or other can be generated, the experimenter must know the theoretical model he wants! Thus, knowing the model, the number of factors and levels of each factor we known then the complete factorial design and from the number of real tests (runs) that can be made by the experimenter (economically speaking!), we must find the most effective sub-set that will be a D-optimal, A-optimal design or other...

It seems that to date the software uses evolutionary algorithms (see section Numerical Methods page 7022) to determine the optimal design so in extenso there is no guarantee that the given optimal design is really the optimal one (except if one restarts the calculation a great number of times always falling on the same result).

The advantage of D-optimal or A-optimal designs or other designs of the same family is that they are very flexible and often more suited to the real needs of the experimenters.

OK... it may seems quite abstract for some readers so let us see a companion example.

Suppose that an experimenter has 3 controlled variables and that seeks to determine a model of the type:

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + a_{11} x_1^2 \tag{16.7.564}$$

and we will consider that it has the following parameters:

$$\begin{aligned} &x_1 : 5 \text{ levels: } (-1, -0.5, 0, +0.5, +1)\\ &x_2 : 2 \text{ levels: } (-1, +1)\\ &x_3 : 2 \text{ levels: } (-1, +1) \end{aligned} \tag{16.7.565}$$

and that he can only do 12 trials (for economic reasons or times reasons or else...!). In this case, given that one of the factors have 5 levels, it is out of question for this to use fractional factorial designs or the Plackett-Burman designs. Even if Taguchi's designs manage certain factors at more than 2 levels there is no Taguchi plan for this special scenario (at least as far as we know...) so we must also forget them. The same applies to composite surface designs or Box-Behnken designs.

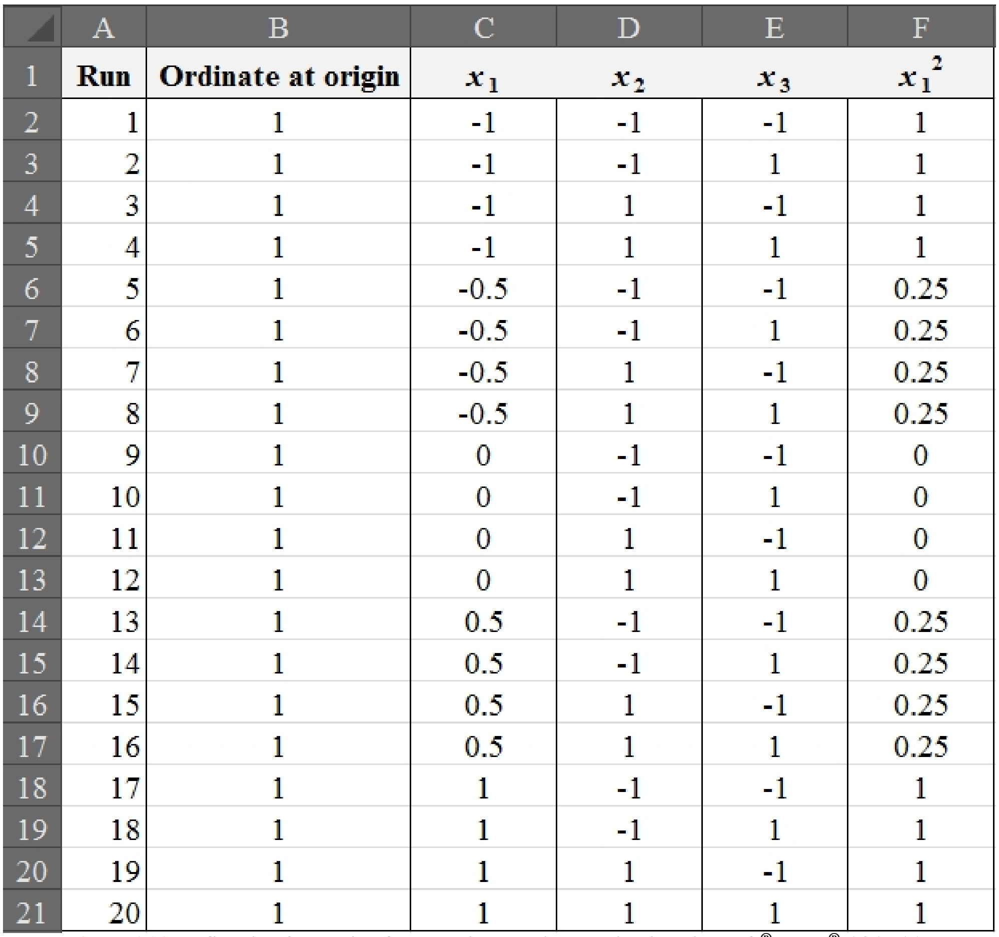
The only thing that can be done is either to use a complete factorial design (but again it would require 20 tests which is not conceivable!) or a D or A-optimal design.

Obviously the complete factorial plan would require $5 \cdot 2 \cdot 2 = 20$ tests (runs) which would be:

Run	x_1	x_2	x_3
1	-1	-1	-1
2	-1	-1	$+1$
3	-1	$+1$	-1
4	-1	$+1$	$+1$
5	-0.5	-1	-1
6	-0.5	-1	$+1$
7	-0.5	$+1$	-1
8	-0.5	$+1$	$+1$
9	0	-1	-1
10	0	-1	$+1$
11	0	$+1$	-1
12	0	$+1$	$+1$
13	$+0.5$	-1	-1
14	$+0.5$	-1	$+1$
15	$+0.5$	$+1$	-1
16	$+0.5$	$+1$	$+1$
17	$+1$	-1	-1
18	$+1$	-1	$+1$
19	$+1$	$+1$	-1
20	$+1$	$+1$	$+1$

Table 16.54: **Starting full design for D-optimal design seeking**

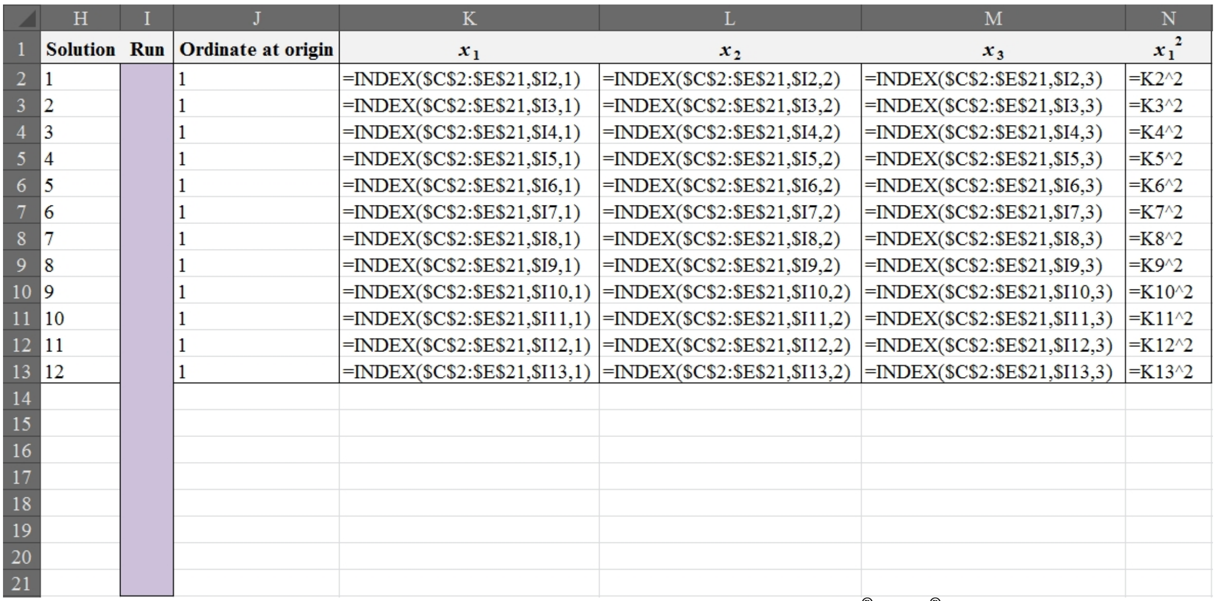
So that everyone can use an accessible tool, we will show how to extract a D or A-optimal design of 12 tests using a simple spreadsheet software like Microsoft® Excel® 14.0.7015 by first constructing the following table corresponding to the matrix in its totality:



	A	B	C	D	E	F
1	**Run**	**Ordinate at origin**	x_1	x_2	x_3	x_1^2
2	1	1	-1	-1	-1	1
3	2	1	-1	-1	1	1
4	3	1	-1	1	-1	1
5	4	1	-1	1	1	1
6	5	1	-0.5	-1	-1	0.25
7	6	1	-0.5	-1	1	0.25
8	7	1	-0.5	1	-1	0.25
9	8	1	-0.5	1	1	0.25
10	9	1	0	-1	-1	0
11	10	1	0	-1	1	0
12	11	1	0	1	-1	0
13	12	1	0	1	1	0
14	13	1	0.5	-1	-1	0.25
15	14	1	0.5	-1	1	0.25
16	15	1	0.5	1	-1	0.25
17	16	1	0.5	1	1	0.25
18	17	1	1	-1	-1	1
19	18	1	1	-1	1	1
20	19	1	1	1	-1	1
21	20	1	1	1	1	1

Figure 16.255: Starting full design for D-optimal design seeking in Microsoft® Excel® 14.0.7015

Then we construct the following structure a little further on the same sheet:



	H	I	J	K	L	M	N
1	**Solution**	**Run**	**Ordinate at origin**	x_1	x_2	x_3	x_1^2
2	1		1	=INDEX($C$2:$E$21,$I2,1)	=INDEX($C$2:$E$21,$I2,2)	=INDEX($C$2:$E$21,$I2,3)	=K2^2
3	2		1	=INDEX($C$2:$E$21,$I3,1)	=INDEX($C$2:$E$21,$I3,2)	=INDEX($C$2:$E$21,$I3,3)	=K3^2
4	3		1	=INDEX($C$2:$E$21,$I4,1)	=INDEX($C$2:$E$21,$I4,2)	=INDEX($C$2:$E$21,$I4,3)	=K4^2
5	4		1	=INDEX($C$2:$E$21,$I5,1)	=INDEX($C$2:$E$21,$I5,2)	=INDEX($C$2:$E$21,$I5,3)	=K5^2
6	5		1	=INDEX($C$2:$E$21,$I6,1)	=INDEX($C$2:$E$21,$I6,2)	=INDEX($C$2:$E$21,$I6,3)	=K6^2
7	6		1	=INDEX($C$2:$E$21,$I7,1)	=INDEX($C$2:$E$21,$I7,2)	=INDEX($C$2:$E$21,$I7,3)	=K7^2
8	7		1	=INDEX($C$2:$E$21,$I8,1)	=INDEX($C$2:$E$21,$I8,2)	=INDEX($C$2:$E$21,$I8,3)	=K8^2
9	8		1	=INDEX($C$2:$E$21,$I9,1)	=INDEX($C$2:$E$21,$I9,2)	=INDEX($C$2:$E$21,$I9,3)	=K9^2
10	9		1	=INDEX($C$2:$E$21,$I10,1)	=INDEX($C$2:$E$21,$I10,2)	=INDEX($C$2:$E$21,$I10,3)	=K10^2
11	10		1	=INDEX($C$2:$E$21,$I11,1)	=INDEX($C$2:$E$21,$I11,2)	=INDEX($C$2:$E$21,$I11,3)	=K11^2
12	11		1	=INDEX($C$2:$E$21,$I12,1)	=INDEX($C$2:$E$21,$I12,2)	=INDEX($C$2:$E$21,$I12,3)	=K12^2
13	12		1	=INDEX($C$2:$E$21,$I13,1)	=INDEX($C$2:$E$21,$I13,2)	=INDEX($C$2:$E$21,$I13,3)	=K13^2
14							
15							
16							
17							
18							
19							
20							
21							

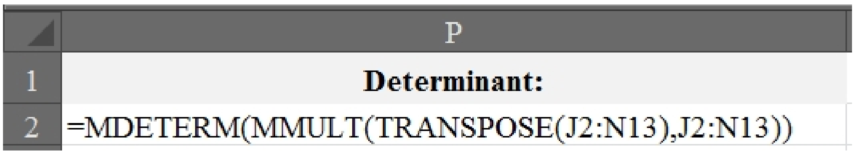
Figure 16.256: Preparation of D-optimal design with formulas in Microsoft® Excel® 14.0.7015

Now let's us recall that finding a D-optimal design is equivalent to find the components of X

such as:

$$\max\left\{\det(X^TX)\right\} \tag{16.7.566}$$

so in one of the cells we should write (be careful this is a matrix function so we must validate by Ctrl + Shift + Enter):

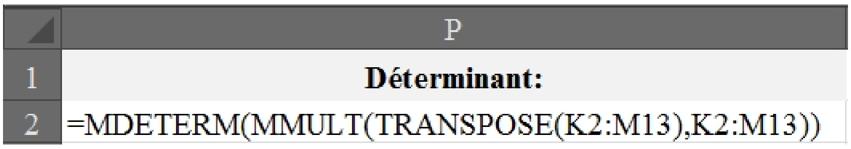


	P
1	**Determinant:**
2	=MDETERM(MMULT(TRANSPOSE(J2:N13),J2:N13))

But if we follow the method advocated by the NIST on their page here:

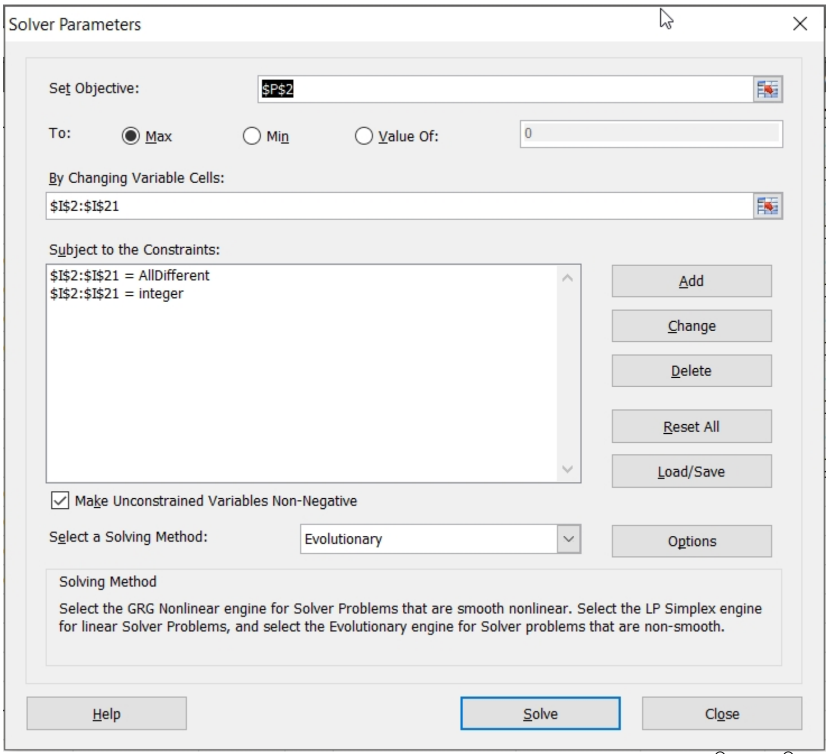
http://www.itl.nist.gov/div898/handbook/pri/section5/pri521.htm

We would have to maximize the determinant of the submatrix (which in all rigour is false but anyway let us do as they say...) and then we write rather:



	P
1	**Déterminant:**
2	=MDETERM(MMULT(TRANSPOSE(K2:M13),K2:M13))

Then we launch the solver of this version of Microsoft® Excel® in evolutionary mode:



Set Objective: $P$2

To: ● Max ○ Min ○ Value Of: 0

By Changing Variable Cells: $I$2:$I$21

Subject to the Constraints:

$I$2:$I$21 = AllDifferent
$I$2:$I$21 = integer

☑ Make Unconstrained Variables Non-Negative

Select a Solving Method: Evolutionary

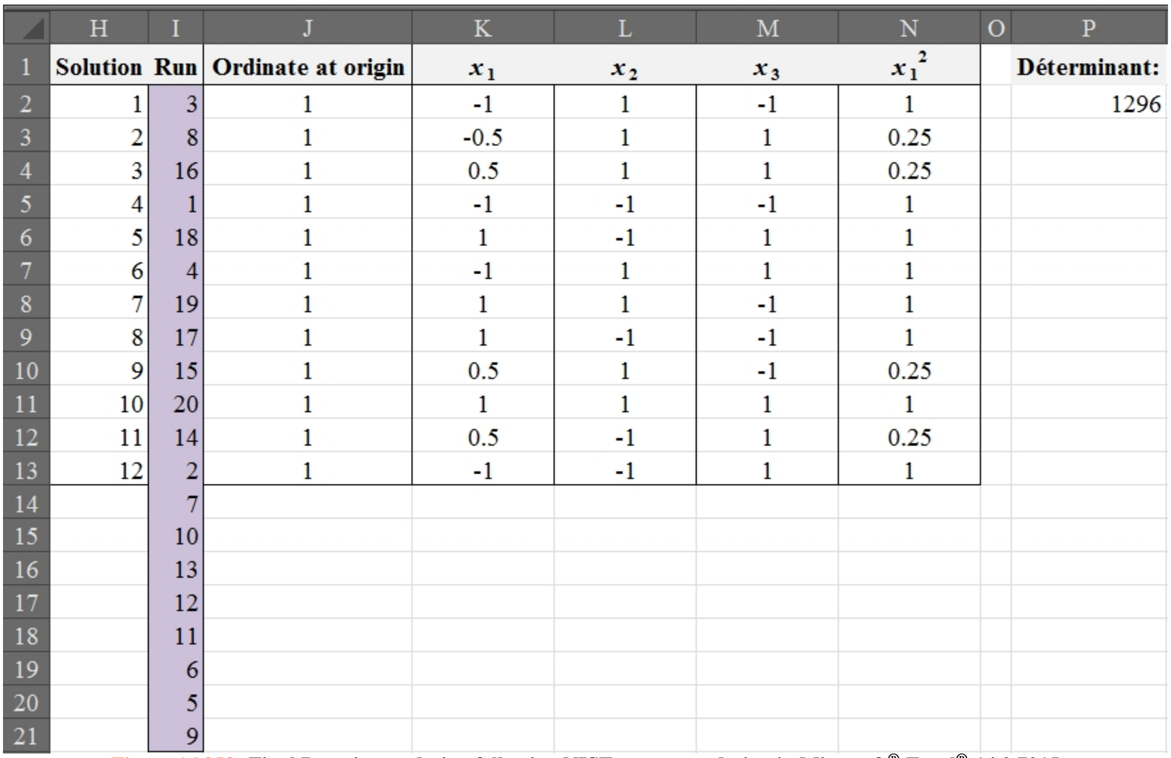
Solving Method

Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

Figure 16.257: **Evolutionary solver settings for seeking D-optimal design in Microsoft® Excel® 14.0.7015**

You may perhaps be wondering why more variable cells should be selected than necessary. This is because of the intrinsic operation of the Microsoft® Excel® solver. If we take the cells `I2` to `I13` it will not find any solution that you put additional constraints or not to the solver!

If we run the solver, we get:



	H	I	J	K	L	M	N	O	P
1	**Solution**	**Run**	**Ordinate at origin**	x_1	x_2	x_3	x_1^2		**Déterminant:**
2	1	3	1	-1	1	-1	1		1296
3	2	8	1	-0.5	1	1	0.25		
4	3	16	1	0.5	1	1	0.25		
5	4	1	1	-1	-1	-1	1		
6	5	18	1	1	-1	1	1		
7	6	4	1	-1	1	1	1		
8	7	19	1	1	1	-1	1		
9	8	17	1	1	-1	-1	1		
10	9	15	1	0.5	1	-1	0.25		
11	10	20	1	1	1	1	1		
12	11	14	1	0.5	-1	1	0.25		
13	12	2	1	-1	-1	1	1		
14		7							
15		10							
16		13							
17		12							
18		11							
19		6							
20		5							
21		9							

Figure 16.258: **Final D-optimum design following NIST recommendation in Microsoft® Excel® 14.0.7015**

that gives us a determinant of 1296. The result provided by the NIST has in comparison a determinant of 1156. The D-optimal experimental design given by the NIST is indeed:

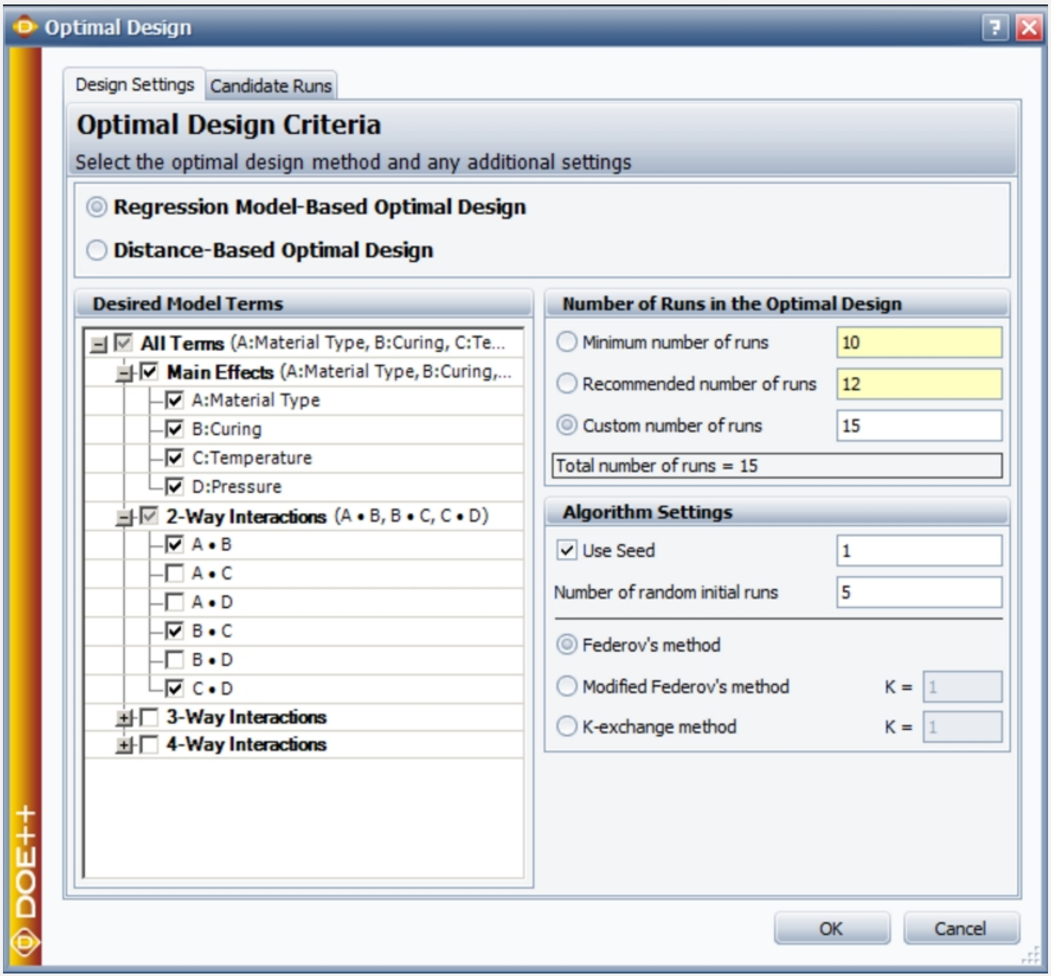
Run	x_1	x_2	x_3
1	$+1$	-1	-1
2	-1	-1	$+1$
3	-1	$+1$	-1
4	-1	$+1$	$+1$
5	0	-1	-1
6	0	-1	$+1$
7	0	$+1$	-1
8	0	$+1$	$+1$
9	$+1$	-1	-1
10	$+1$	-1	$+1$
11	$+1$	$+1$	-1
12	$+1$	$+1$	$+1$

Table 16.55: **Final D-optimal design**

Where does this difference comes from???!!!!! This is quite easy to understand: It is impossible to find all coefficients with the system we get with Microsoft® Excel® as we still have 4 levels, 2 levels and 2 levels so this system cannot be solved as we would need 16 runs to solve it without alias!!!! So the real algorithm to reach 12 runs has to eliminate one of the level of the first factor and the main purpose of D-optimum design is to find which level should be eliminate to keep the maximum information!

Remark

Caution! Depending on the softwares packages used for the analysis of DoE, you might come across a different result because they basically use almost all different algorithms (Federov's Method, Modified Federov's Method or K-Exchange Method for most known one):


Optimal Design
Design Settings
Candidate Runs
Optimal Design Criteria
Select the optimal design method and any additional settings
Desired Model Terms
Number of Runs in the Optimal Design
Algorithm Settings
OK
Cancel
DOE++

● Regression Model-Based Optimal Design
○ Distance-Based Optimal Design

☑ All Terms (A:Material Type, B:Curing, C:Te...
- ☑ Main Effects (A:Material Type, B:Curing,...
  - ☑ A:Material Type
  - ☑ B:Curing
  - ☑ C:Temperature
  - ☑ D:Pressure
- ☑ 2-Way Interactions (A • B, B • C, C • D)
  - ☑ A • B
  - ☐ A • C
  - ☐ A • D
  - ☑ B • C
  - ☐ B • D
  - ☑ C • D
- ☐ 3-Way Interactions
- ☐ 4-Way Interactions

○ Minimum number of runs 10
○ Recommended number of runs 12
● Custom number of runs 15
Total number of runs = 15

☑ Use Seed 1
Number of random initial runs 5
● Federov's method
○ Modified Federov's method K = 1
○ K-exchange method K = 1

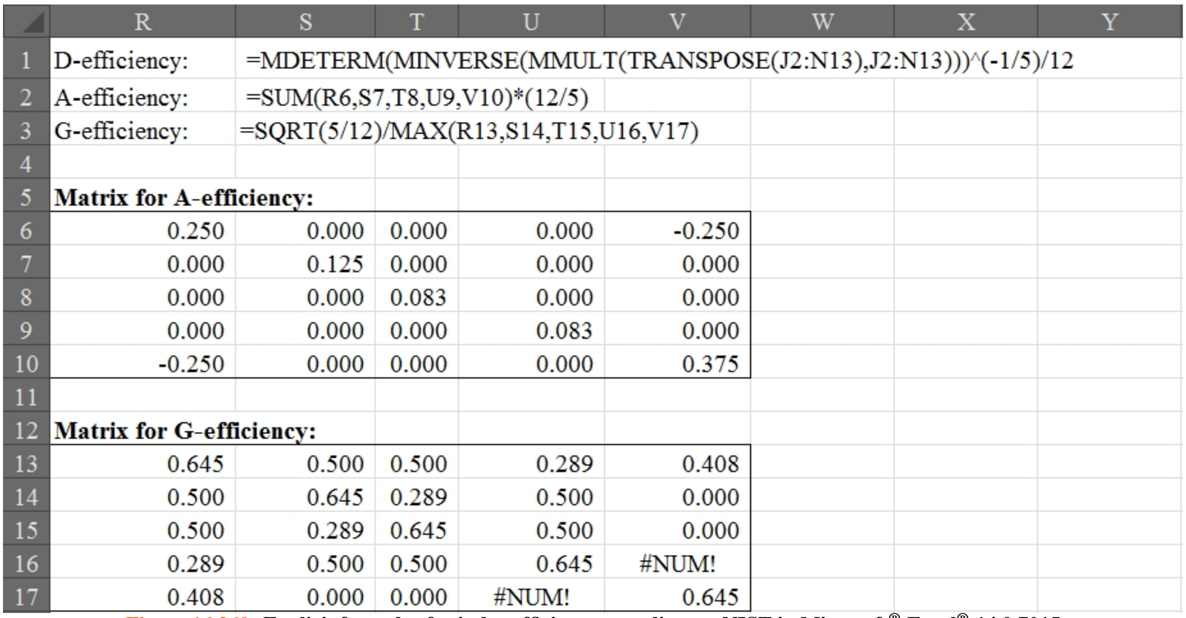
Figure 16.259: Reliasoft DOE++ Screenshot of algorithms options dialogue box

Then, it is customary in practice and if we conform ourselves to the NIST results... to make the calculations of the following efficiency indices relative to the matrix obtained (the choice is completely empirical but well... it seems that most statistical software has its own indicators...):

$$\begin{aligned}\text{D-efficiency} &= \min\left\{\dfrac{1}{N_D\left(\det(X^TX)\right)^{-1/p}}\right\}\\ \text{A-efficiency} &= \min\left\{\text{tr}\left((X^TX)^{-1}\right)\right\}\\ \text{G-efficiency} &= \max\left\{\dfrac{\sqrt{N/p}}{\sigma_G}\right\}\end{aligned}\tag{16.7.567}$$

where in the G-efficiency, the term σ_G is the smallest variance in the diagonal of the projection matrix $X(X^TX)^{-1}X^T$.

What gives in Microsoft® Excel®:



	R	S	T	U	V	W	X	Y
1	D-efficiency:	=MDETERM(MINVERSE(MMULT(TRANSPOSE(J2:N13),J2:N13)))^(-1/5)/12						
2	A-efficiency:	=SUM(R6,S7,T8,U9,V10)*(12/5)						
3	G-efficiency:	=SQRT(5/12)/MAX(R13,S14,T15,U16,V17)						
4								
5	**Matrix for A-efficiency:**							
6	0.250	0.000	0.000	0.000	-0.250			
7	0.000	0.125	0.000	0.000	0.000			
8	0.000	0.000	0.083	0.000	0.000			
9	0.000	0.000	0.000	0.083	0.000			
10	-0.250	0.000	0.000	0.000	0.375			
11								
12	**Matrix for G-efficiency:**							
13	0.645	0.500	0.500	0.289	0.408			
14	0.500	0.645	0.289	0.500	0.000			
15	0.500	0.289	0.645	0.500	0.000			
16	0.289	0.500	0.500	0.645	#NUM!			
17	0.408	0.000	0.000	#NUM!	0.645			

Figure 16.260: **Explicit formulas for index efficiency compliant to NIST in Microsoft® Excel® 14.0.7015**

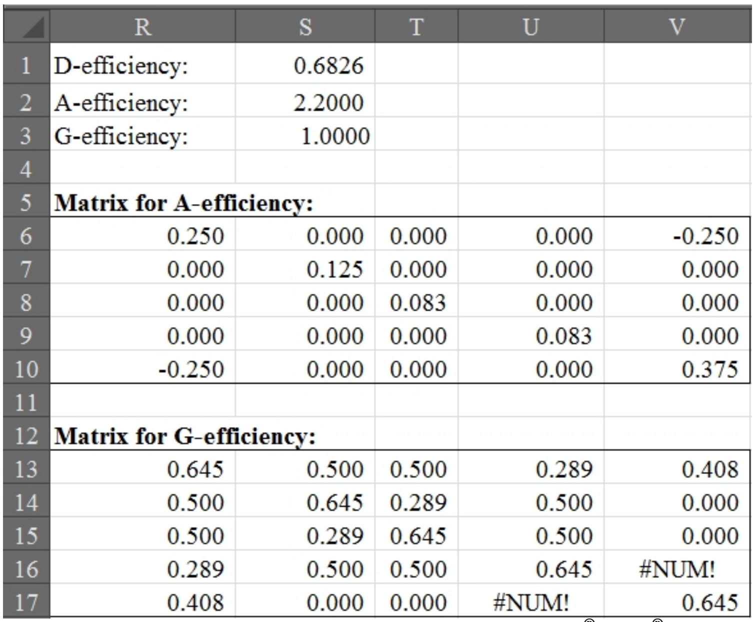
where in the relations of the cells `S1` to `S3`, the 5 is the number of columns of the experiment matrix and the 12 is the number of rows. To do the calculations of the A-efficiency and G-efficiency we need in Microsoft® Excel® both matrices explicitly (since this spreadsheet does not have functions for the trace of a matrix nor for the maximum of the trace) as that we have in `R6` (and the whole range of the relative matrix) the following matrix relation (to be validated with Ctrl+Shit+Enter):

```
=MINVERSE(MMULT(TRANSPOSE(J2:N13),J2:N13))
```

and in `R13` (and in all the corresponding range) the following matrix relation (to be validated with Ctrl+Shit+Enter):

```
=SQRT(MMULT(J2:N13,MMULT(MINVERSE(MMULT(TRANSPOSE(J2:N13),J2:N13))
,TRANSPOSE(J2:N13))))
```

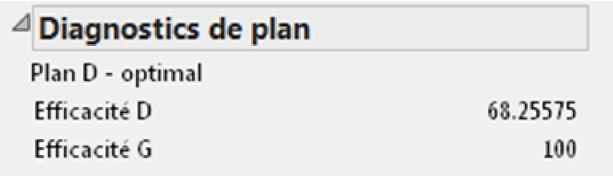
What gives:



	R	S	T	U	V
1	D-efficiency:	0.6826			
2	A-efficiency:	2.2000			
3	G-efficiency:	1.0000			
4					
5	**Matrix for A-efficiency:**				
6	0.250	0.000	0.000	0.000	-0.250
7	0.000	0.125	0.000	0.000	0.000
8	0.000	0.000	0.083	0.000	0.000
9	0.000	0.000	0.000	0.083	0.000
10	-0.250	0.000	0.000	0.000	0.375
11					
12	**Matrix for G-efficiency:**				
13	0.645	0.500	0.500	0.289	0.408
14	0.500	0.645	0.289	0.500	0.000
15	0.500	0.289	0.645	0.500	0.000
16	0.289	0.500	0.500	0.645	#NUM!
17	0.408	0.000	0.000	#NUM!	0.645

Figure 16.261: **Index efficiency conforming to NIST in Microsoft® Excel® 14.0.7015**

and perfectly matches (obviously...) the values given by the NIST! We also see that JMP (SAS) also gives (with the difference that it multiplies D-efficiency and G-efficiency by a factor 100) as shown in the printscreen of the french version below:



Diagnostics de plan

Plan D - optimal

Efficacité D 68.25575

Efficacité G 100

Figure 16.262: **JMP (SAS) efficiency index of optimal designs**

while the Minitab software (version 16.1.1 and earlier) gives the D and G optimality but calculates it in a bizarre and obscure way in my point of view.

In order to fit an optimality criterion the companion example above show us that we have to define the number of experiments we want to have in the design. The selection of this factor is very important because changing the number of the design runs alters the model matrix and another optimal design is chosen. There are no rules to define this number, but the minimum is model-dependent. A model with p coefficients can only be investigated with a D-optimal design which has at least p runs. In most cases, it is useful to create different designs that differ in the number of runs and compare the efficiency of the designs. A design with a few more or less design runs than the desired one can have a higher determinant and hence is the best design to perform. A software like Minitab for reasons that are unknown to us don't give however the possibility to have a number of runs equal to the number of factors.

16.7.4.6 Mixture Design

When a product is formed by mixing together two or more ingredients, the product is named a "mixture", and the ingredients are named "mixture components". In a general mixture problem, the measured response is assumed to depend only on the proportions of the ingredients in the mixture, not the amount of the mixture. For example, the taste of a fruit punch recipe (i.e., the response) might depend on the proportions of watermelon, pineapple and orange juice in the mixture. The taste of a small cup of fruit punch will obviously be the same as a big cup.

Sometimes the responses of a mixture experiment depend not only on the proportions of ingredients, but also on the settings of variables in the process of making the mixture. For example, the tensile strength of stainless steel is not only affected by the proportions of iron, copper, nickel and chromium in the alloy; it is also affected by process variables such as temperature, pressure and curing time used in the experiment.

One of the purposes of conducting a mixture experiment is to find the best proportion of each component and the best value of each process variable, in order to optimize a single response or multiple responses simultaneously. In this chapter, we will discuss how to design effective mixture designs and how to analyse data from mixture experiments with and without process variables.

There are several different types of mixture designs. The most common ones are "simplex lattice", "simplex centroid", "simplex axial" and "extreme vertex designs", each of which is used for a different purpose.

- If there are many components in a mixture, the first choice is to screen out the most important ones. Simplex axial and Simplex centroid designs are used for this purpose.
- If the number of components is not large, but a high order polynomial equation is needed in order to accurately describe the response surface, then a Simplex lattice design can be used.
- Extreme vertex designs are used for the cases when there are constraints on one or more components (e.g., if the proportion of watermelon juice in a fruit punch recipe is required to be less than 30%, and the combined proportion of watermelon and orange juice should always be between 40% and 70%).

So often in practice we are interested in studying mixtures. In other words, we want to know the proportion of ingredients needed to achieve a certain result. The cases of application are so numerous that it would be futile to enumerate them all here.

The fact that the proportions must be added is the key element of the simple mixture design since if we denote by x_i the proportion of the ingredient i then under unconstrained mixture we have:

$$\begin{gathered}1 \geq x_i \geq 0\\ \sum_{i=1}^{k} x_i = 1\end{gathered} \tag{16.7.568}$$

or written slightly differently:

$$\begin{gathered}100\% \geq x_i \geq 0\% \\ \sum_{i=1}^{k} x_i = 100\%\end{gathered} \tag{16.7.569}$$

where the last relation is named logically "fundamental constraint of unconstrainted mixtures".

So we already notice that it is not possible to vary one parameter (ingredient) without the other being modified. So the control variables are not independent! We can not then use the techniques seen so far!

As we will see just below, the response surface of a mixture design is a simplex (connected surface as introduced in the section of Geometric Shapes) of dimension k with $k-1$ edges (and generally plunged into a space of dimension k).

For example, with two factors (binary mixture) the simplex is simply a line segment ranging from $(0,1)$ to $(1,0)$ in terms of coordinates. With three factors, the simplex is a triangle with three edges with three vertices of coordinates $(1,0,0)$, $(0,1,0)$ and $(0,0,1)$:

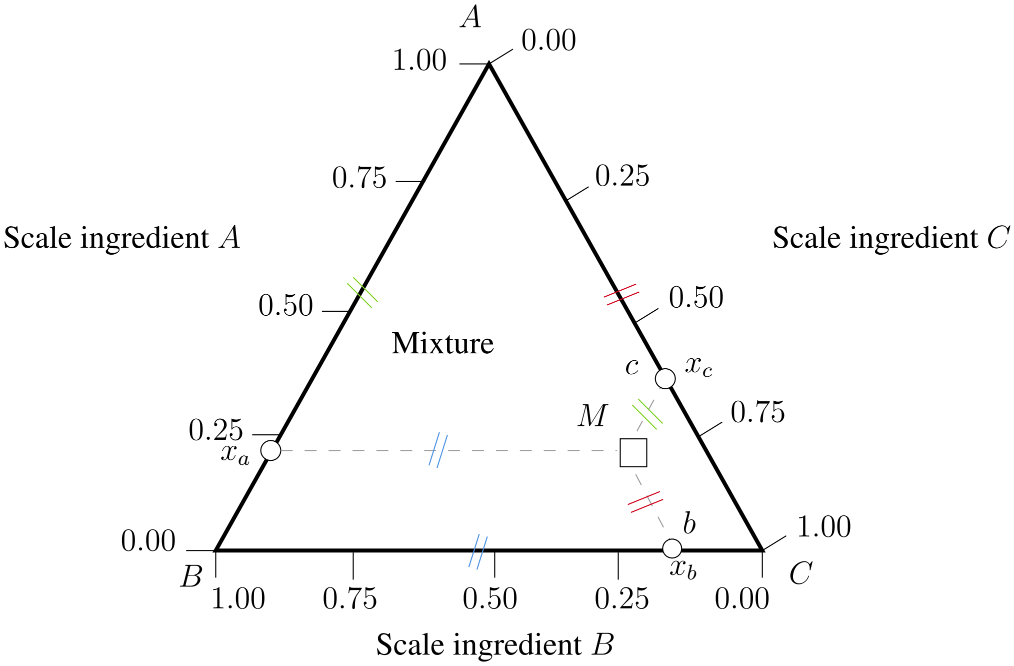


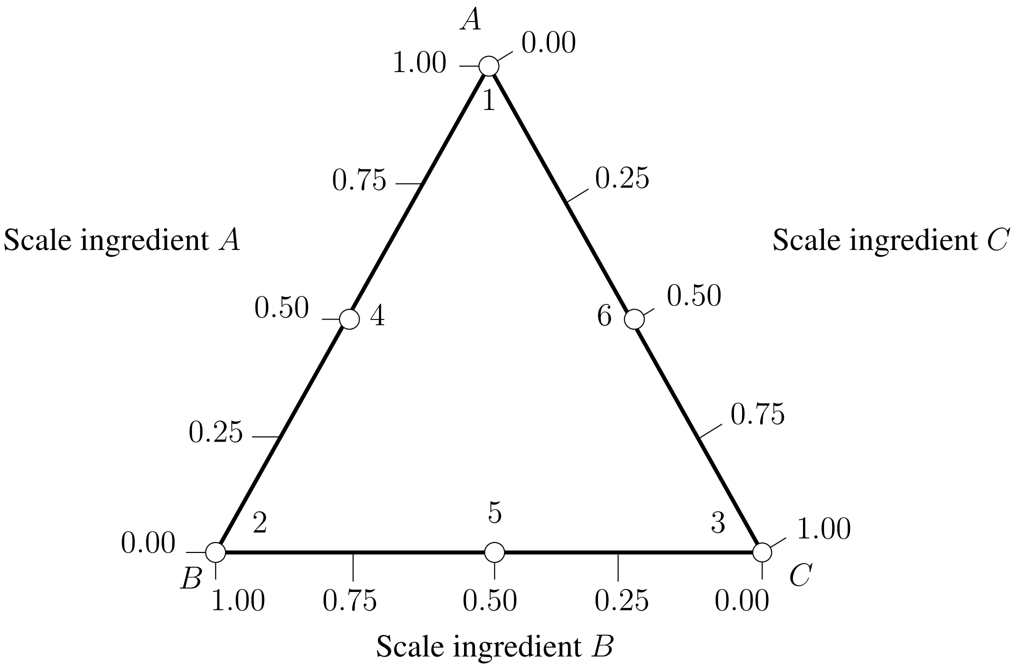
Figure 16.263: **Three factors mixtures design simplex**

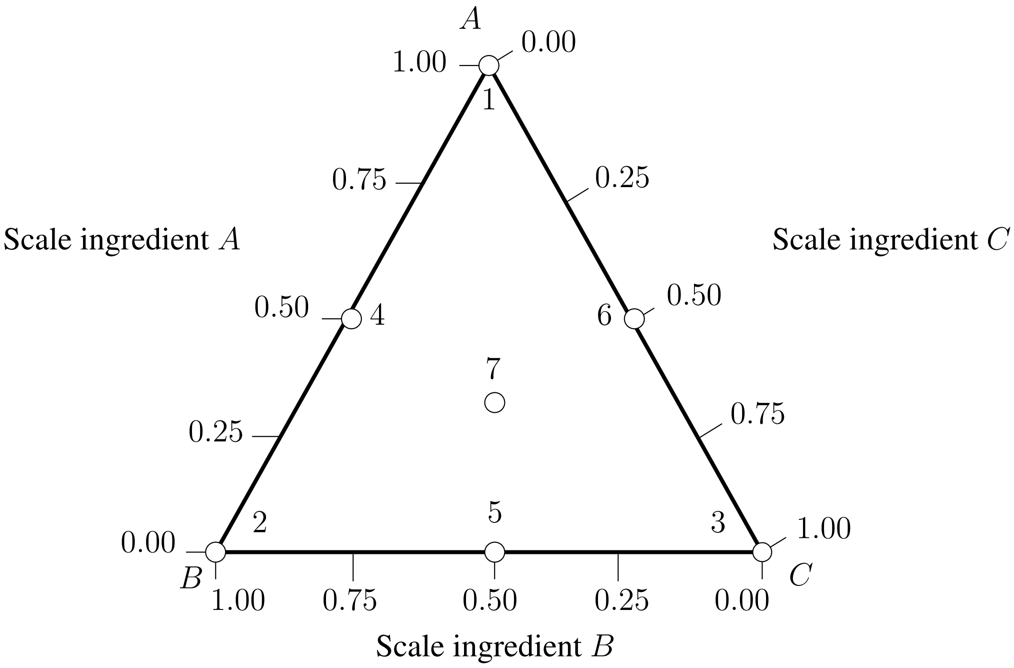
where the segments of the point M at the concentration points are logically the translations of the edges of the triangle with for condition that the sum of the segments being equal to a unit length.

There are several ways of arranging the experimental points whose in our field of study are (from top to bottom in the figure below):

- Simplex lattice designs

- Simplex centroid designs
- Augmented simplex centroid designs
- and some other exotics one...


A
0.00
1.00
1
0.75
0.25
Scale ingredient A
Scale ingredient C
0.50
4
6
0.50
0.25
0.75
2
5
3
0.00
1.00
B
C
1.00
0.75
0.50
0.25
0.00
Scale ingredient B


A
0.00
1.00
1
0.75
0.25
Scale ingredient A
Scale ingredient C
0.50
4
6
0.50
7
0.25
0.75
2
5
3
0.00
1.00
B
C
1.00
0.75
0.50
0.25
0.00
Scale ingredient B

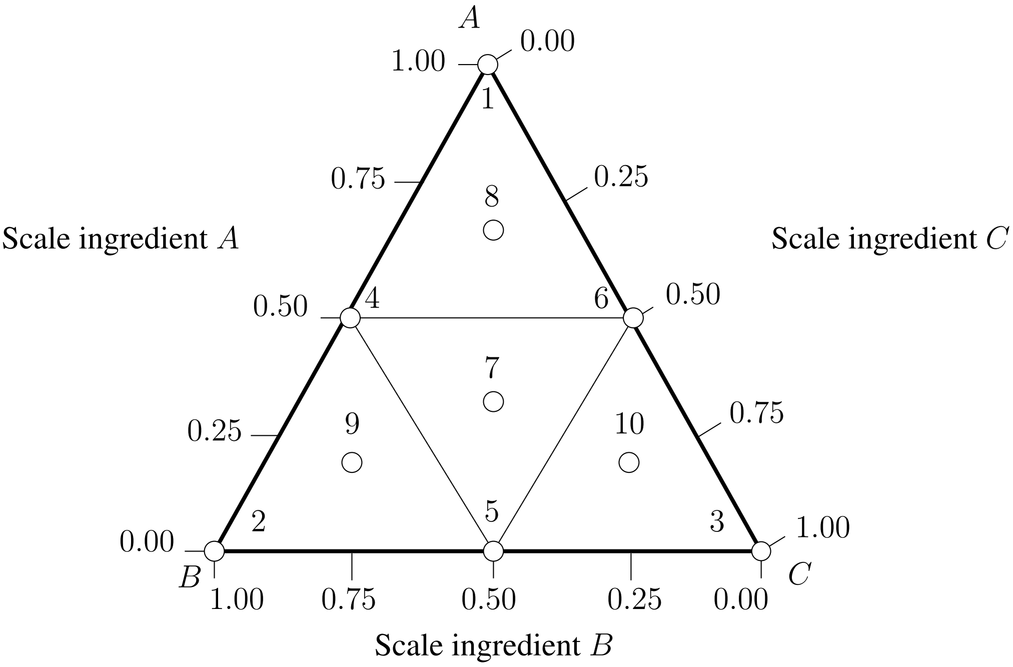

A
1.00
0.00
1
0.75
8
0.25
Scale ingredient A
Scale ingredient C
0.50
4
6
0.50
7
0.25
9
10
0.75
0.00
2
5
3
1.00
B
C
1.00
0.75
0.50
0.25
0.00
Scale ingredient B

Figure 16.264: **Various arrangements of three factors mixture simplex**

Giving some classes I have noticed that also an example of a 4 factors simplex centroid example design is also useful for some students:

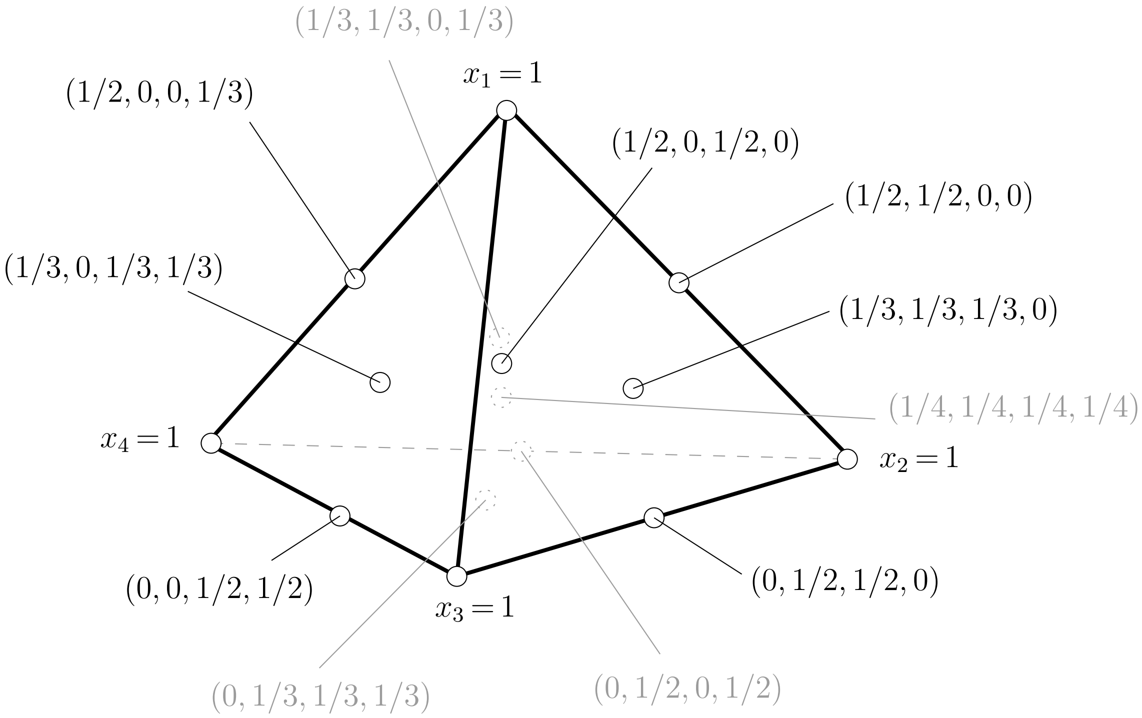

(1/3, 1/3, 0, 1/3)
$x_1 = 1$
(1/2, 0, 0, 1/3)
(1/2, 0, 1/2, 0)
(1/2, 1/2, 0, 0)
(1/3, 0, 1/3, 1/3)
(1/3, 1/3, 1/3, 0)
(1/4, 1/4, 1/4, 1/4)
$x_4 = 1$
$x_2 = 1$
(0, 0, 1/2, 1/2)
(0, 1/2, 1/2, 0)
$x_3 = 1$
(0, 1/3, 1/3, 1/3)
(0, 1/2, 0, 1/2)

Figure 16.265: 4 **factors simplex centroid example**

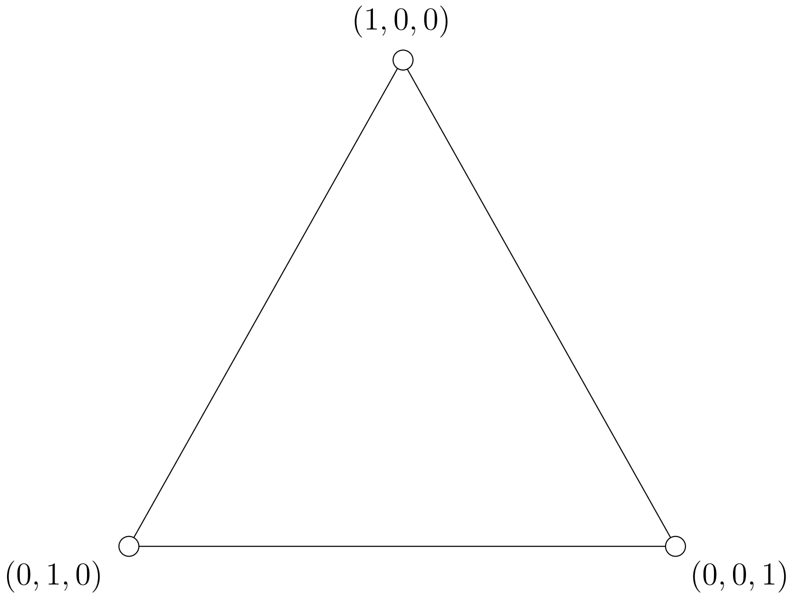
16.7.4.6.1 Network mixture designs (simplex lattice designs)

The "network mixture designs" more commonly named "Simplex lattice designs" and sometimes just "L-simplex design" or "Scheffé's designs" are those most often studied in schools because they are the least abstract and algebraically the most accessible. Thus, in a network mixture design of p variables and where the intervals of values are cut in m pieces, the reader will notice by trying manually on a sheet of paper that we get the following table of number of runs (that is to say: the cost of the corresponding design):

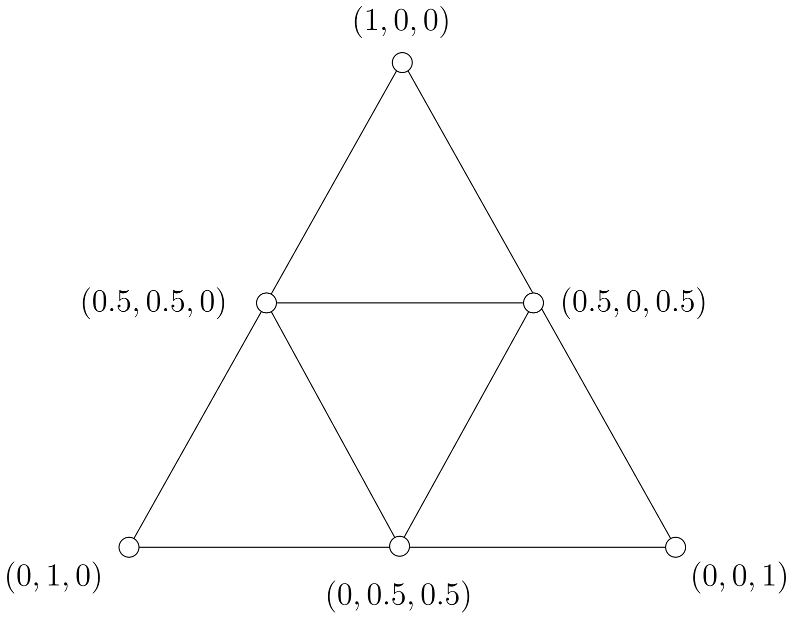
	m		
p	1 (pure simplex)	2	3
3	3	6	10
4	4	10	20
5	5	15	35
6	6	21	56
7	7	28	84

Table 16.56: **Number of points in a mixture lattice design (non-centered)**

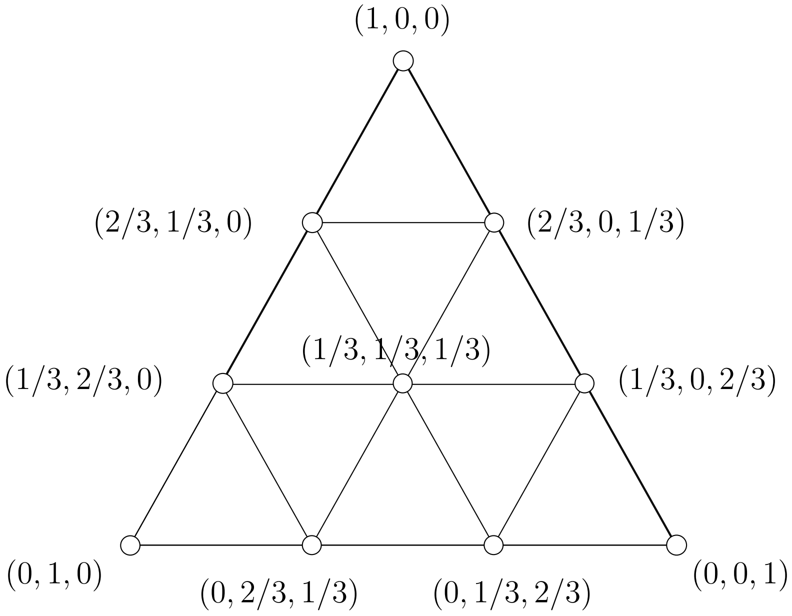
OK this table may be a quite abstract for some people. So let us consider an illustration of the case $p=3$ and $m=1$ (we see clearly the $N=3$ runs):



and for the case $p=3$ and $m=2$ (we see clearly the $N=6$ runs):


(1, 0, 0)
(0.5, 0.5, 0)
(0.5, 0, 0.5)
(0, 1, 0)
(0, 0.5, 0.5)
(0, 0, 1)

and finally for the case $p=3$ and $m=3$ (we see clearly the $N=10$ runs):


(1, 0, 0)
(2/3, 1/3, 0)
(2/3, 0, 1/3)
(1/3, 1/3, 1/3)
(1/3, 2/3, 0)
(1/3, 0, 2/3)
(0, 1, 0)
(0, 2/3, 1/3)
(0, 1/3, 2/3)
(0, 0, 1)

A very common mixture design is the $\{p,m\}=3,2$ with 6 experimental points as seen above but with an addition point. This is why in many textbooks we can find: $\{p,m\}=3,2=7$ as there is an addition center point!

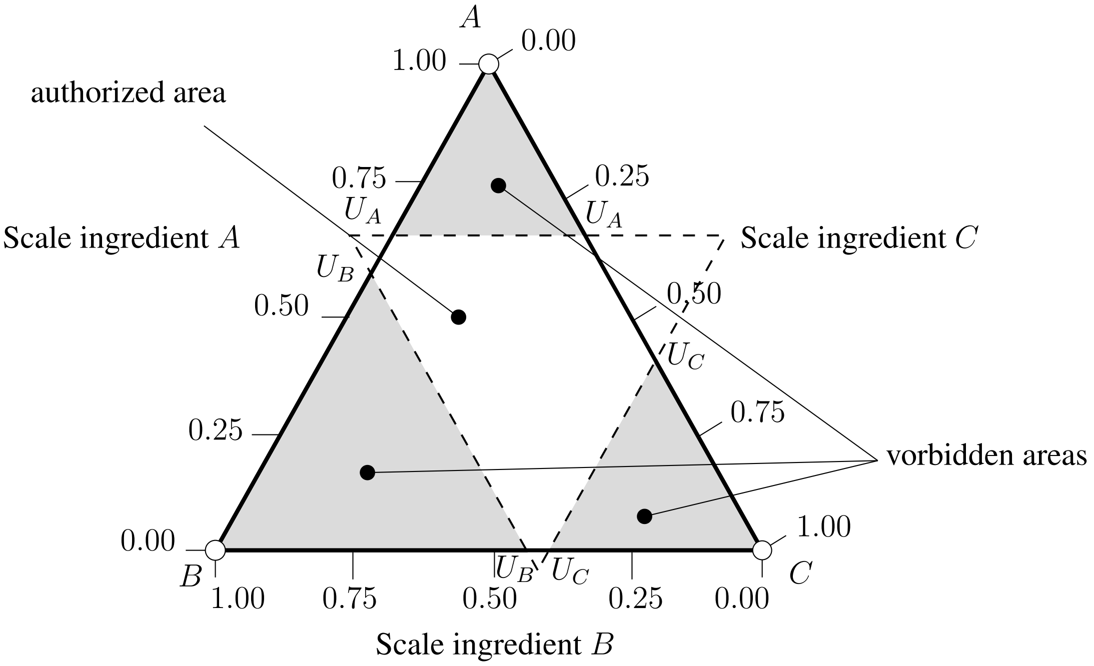
In the general case of non-centered lattice mixture designs we find by a little bit trial and error that the cost (runs) of such a simplex is given by:

$$\boxed{N_{\text{SLD}}=C_{m}^{q+m-}=\frac{(q+m-1)!}{m!(q-1)!}} \tag{16.7.570}$$

and for a centered lattice design of q factors by:

$$\boxed{N_{\text{CLD}}=2^q-1} \tag{16.7.571}$$

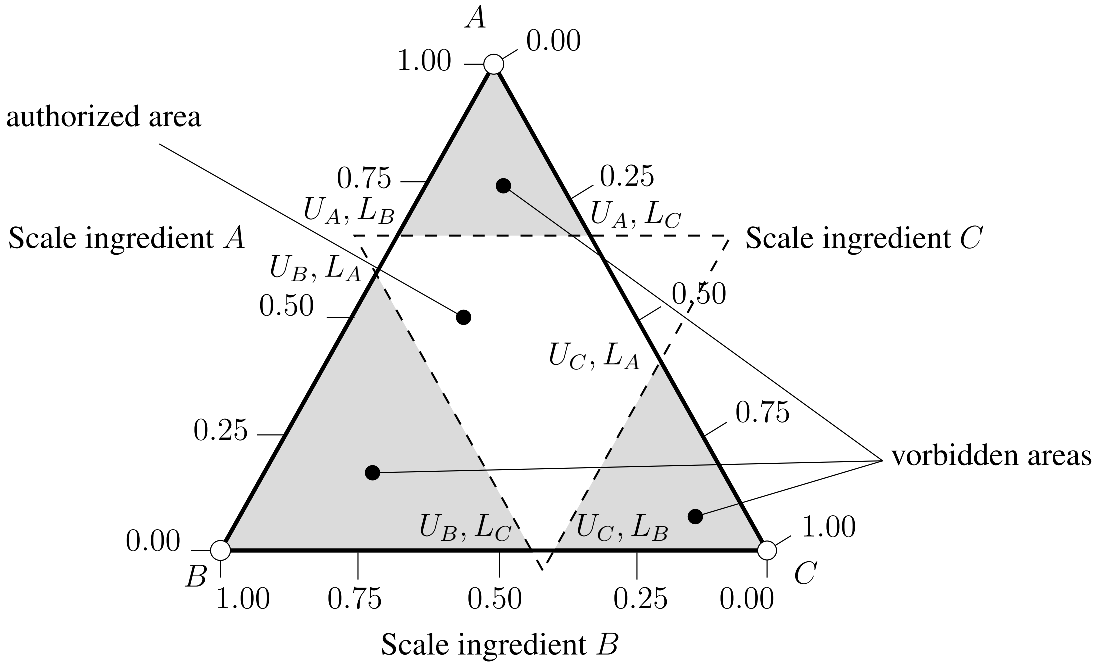
Let us also indicate that logically nothing prevents the support (domain of definition) of the mixtures from being constrained, we then speak of "extreme vertices designs" that cover only a subportion or smaller space within the simplex. Thus, in the case of a ternary mixture design, nothing prevents us from having a case where each of the components has only upper limit values but the lower value can be zero! Here is for example a special case of constraint mixture design:



A
0.00
1.00
authorized area
0.75
U_A
0.25
U_A
Scale ingredient A
Scale ingredient C
U_B
0.50
0,50
U_C
0.25
0.75
vorbidden areas
0.00
1.00
B
U_B
U_C
C
1.00
0.75
0.50
0.25
0.00
Scale ingredient B

Figure 16.266: Three factor mixtures design complex

or lower (L) and at the same time upper (U) limit values:



A
0.00
1.00
authorized area
0.75
U_A, L_B
0.25
U_A, L_C
Scale ingredient A
Scale ingredient C
U_B, L_A
0.50
0,50
U_C, L_A
0.25
0.75
vorbidden areas
U_B, L_C
U_C, L_B
0.00
1.00
B
C
1.00
0.75
0.50
0.25
0.00
Scale ingredient B

or that the ratio of two ingredients is constant, or that the sum of two ingredients gives a constant proportion. In short there may be many scenarios, but in the case of limited boundary, it is customary to write:

$$1 \geq U_i \geq x_i \geq L_i \geq 0$$
$$\sum_{i=1}^{k} x_i = 1 \tag{16.7.572}$$

and this is what we name "linear constraints" (as it is obvious if each ingredient is bounded that the sum is also bounded!).

Remark

It is quite possible to treat both the mixing variables (the proportions of the constituents) and the design factors of the experiments. To illustrate this situation, we can take the example of a chocolate manufacturer. The study of the composition of the chocolate gives rise to a mixtures design and the conditions of preparation give rise to a factorial design (or other ...). At each point of experiments of the factorial design, it is necessary to realize mixture design!!!! There is thus rapidly a large number of tests to be carried out since it is necessary to carry out np experiments if the mixture design has n runs and the factorial design has p runs.

We then consider the particular simple linear model (thus still an approximation ... and there are many other theoretical models that are not linear):

$$y = \beta_0 + \sum_i \beta_i x_i \tag{16.7.573}$$

or seen with the statistician point of view:

$$y = \mathrm{E}(\hat{y}) = \mathrm{E}\left(\hat{\beta}_0 + \sum_i \hat{\beta}_i x_i\right) \tag{16.7.574}$$

But since we have the constraint that $\sum_{i=1}^{k} x_i = 1$ the $\hat{\beta}_i$ can not be determined uniquely if we go through the usual linear system. One of the x_i must be used to determine the coefficients uniquely and must therefore be eliminated (concept of degree of freedom)! It is obviously very annoying in practice.

So mathematically, nothing prevents us from writing:

$$\hat{y} = \hat{\beta}_0 + \sum_i \hat{\beta}_i x_i = \hat{\beta}_0 \cdot 1 + \sum_i \hat{\beta}_i x_i = \hat{\beta}_0 \sum_{i=1}^{k} x_i + \sum_{i=1}^{k} \hat{\beta}_i x_i = \sum_{i=1}^{k} (\hat{\beta}_0 + \hat{\beta}_i) x_i \tag{16.7.575}$$

and if we change the notations:

$$(\hat{\beta}_0 + \hat{\beta}_i) = \hat{b}_i \tag{16.7.576}$$

the previous relation can then be written:

$$\hat{y} = \sum_{i=1}^{k} \hat{b}_i x_i \tag{16.7.577}$$

The mathematical model may of course also contain interactions. For example in the case of two factors, we could write with interactions and at the second degree:

$$\hat{y}=\hat{\beta}_0+\sum_i \hat{\beta}_i x_i+\sum_i \hat{\beta}_{ii}x_i^2+\sum_{i<j}\hat{\beta}_{ij}x_i x_j \tag{16.7.578}$$

and by applying the identity:

$$x_1+x_2+\ldots=\sum_i x_i=1 \tag{16.7.579}$$

then:

$$x_i^2=x_i\left(1-\sum_{j\neq i}x_j\right) \tag{16.7.580}$$

Therefore:

$$\begin{aligned}\hat{y}&=\hat{\beta}_0\left(\sum_i x_i\right)+\sum_i \hat{\beta}_i x_i+\sum_i\left[\hat{\beta}_{ii}x_i\left(1-\sum_{j\neq i}x_j\right)\right]+\sum_{i<j}\hat{\beta}_{ij}x_i x_j\\&=\sum_i\left(\hat{\beta}_0+\hat{\beta}_i+\hat{\beta}_{ii}\right)x_i-\sum_i\left[\hat{\beta}_{ii}x_i\sum_{j\neq i}x_j\right]+\sum_{i<j}\hat{\beta}_{ij}x_i x_j\\&=\sum_i\left(\hat{\beta}_0+\hat{\beta}_i+\hat{\beta}_{ii}\right)x_i+\sum_{i<j}\hat{\beta}_{ij}x_i x_j-\sum_i\left[\hat{\beta}_{ii}x_i\sum_{j\neq i}x_j\right]\\&=\sum_i\left(\hat{\beta}_0+\hat{\beta}_i+\hat{\beta}_{ii}\right)x_i+\sum_{i<j}\hat{\beta}_{ij}x_i x_j-\sum_{i<j}(\hat{\beta}_{ii}-\hat{\beta}_{jj})x_i x_j\\&=\sum_i \hat{b}x_i+\sum_{i<j}(\hat{\beta}_{ij}-\hat{\beta}_{ii}-\hat{\beta}_{jj})x_i x_j\\&=\sum_i \hat{b}_i x_i+\sum_{i<}\hat{b}_{ij}x_i x_j\end{aligned} \tag{16.7.581}$$

with:

$$\begin{aligned}\hat{b}_i&=\hat{\beta}_0+\hat{\beta}_i+\hat{\beta}_{ii}\\\hat{b}_{ij}&=\hat{\beta}_{ij}-\hat{\beta}_{ii}-\hat{\beta}_{jj}\end{aligned} \tag{16.7.582}$$

Either with three ingredients this gives for example:

$$\hat{y}=\sum_i \hat{b}_i x_i+\sum_{i<j}\hat{b}_{ij}x_i x_j=\hat{b}_1x_1+\hat{b}_2x_2+\hat{b}_3x_3+\hat{b}_{12}x_1x_2+\hat{b}_{13}x_1x_3+\hat{b}_{23}x_2x_3 \tag{16.7.583}$$

and so on for n components and the idea is the same for the third degree and beyond. It is Henry Scheffé that suggested to describe mixture properties by reduced polynomials subject to the normalization condition.

The number of terms will then obviously be for this model of the second degree:

$$n+\frac{n(n-1)}{2}=\frac{n(n+1)}{2} \tag{16.7.584}$$

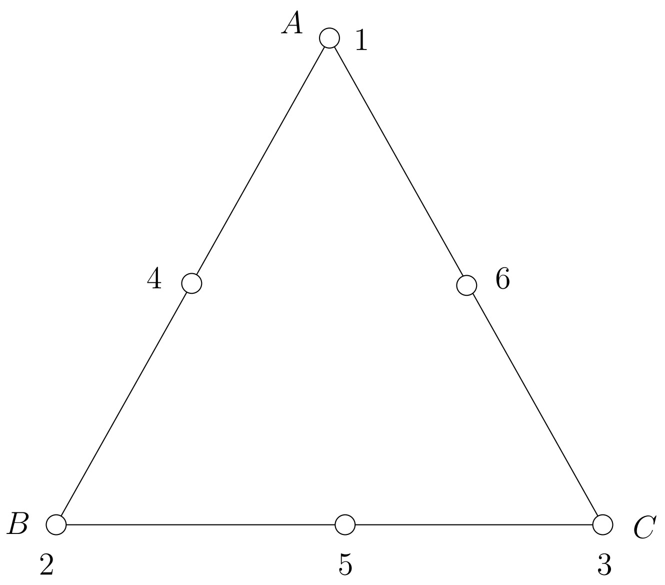
We can then use the following table of control (for those who like to do the calculations by hand) by stopping at the power of three (yes we have to stop one day or another...):

# of factors	**# terms with linear model**	**# terms with quadratic model**	**# terms with cubic model**
2	2	3	$-$
3	3	7	7
4	4	10	14
5	5	15	24
...	...	...	...
n	n	$\dfrac{n(n+1)}{2}$	$\dfrac{n(n^2+5)}{6}$

Table 16.57: **L-mixture designs terms for linear, quadratic and upper models**

Now let us come back to our initial example:

$$\hat{y}=\hat{b}_1x_1+\hat{b}_2x_2+\hat{b}_3x_3+\hat{b}_{12}x_1x_2+\hat{b}_{13}x_1x_3+\hat{b}_{23}x_2x_3 \tag{16.7.585}$$



for the following both reasons:

1. There are 6 measurement points and this is good because we have 6 coefficients to determine and we will need 6 equations at least.
2. The choice of the arrangement of the 6 points is quite intuitive. Indeed, if you imagine first taking 3 points, which ones are furthest from each other? They will be those of the summits of course! Once these three points are taken, what are the three other potential points that are the furthest away from it, but at the same time the furthest away from each other? They will be those in the middle of the edges!

We will then write algebraically the following that resume the above figure:

- For 3 of the 6 points on the vertices: y_i for $x_i = 1$, $x_j = 0$ with $i, j = 1, 2, 3$ under the constraint that $j \neq 1$
- For 3 of the 6 points on the middle of the edges: y_{ij} for $x_i = \dfrac{1}{2}$, $x_j = \dfrac{1}{2}$, $x_k = 0$ under the constraint that $i < j, k \neq i, j$

What injected into:

$$\hat{y} = \hat{b}_1 x_1 + \hat{b}_2 x_2 + \hat{b}_3 x_3 + \hat{b}_{12} x_1 x_2 + \hat{b}_{13} x_1 x_3 + \hat{b}_{23} x_2 x_3 \tag{16.7.586}$$

gives 6 equations with 6 unknowns:

$$\begin{aligned} \hat{y}_1 &= \hat{b}_1 \\ \hat{y}_2 &= \hat{b}_2 \\ \hat{y}_3 &= \hat{b}_3 \\ \hat{y}_{12} &= \hat{b}_1 \left(\frac{1}{2}\right) + \hat{b}_2 \left(\frac{1}{2}\right) + \hat{b}_{12} \left(\frac{1}{4}\right) \\ \hat{y}_{13} &= \hat{b}_1 \left(\frac{1}{2}\right) + \hat{b}_3 \left(\frac{1}{2}\right) + \hat{b}_{13} \left(\frac{1}{4}\right) \\ \hat{y}_{23} &= \hat{b}_2 \left(\frac{1}{2}\right) + \hat{b}_3 \left(\frac{1}{2}\right) + \hat{b}_{23} \left(\frac{1}{4}\right) \end{aligned} \tag{16.7.587}$$

The search for the coefficients for these 6 equations is immediate:

$$\begin{aligned} \hat{b}_1 &= \hat{y}_1 \\ \hat{b}_2 &= \hat{y}_2 \\ \hat{b}_3 &= \hat{y}_3 \\ \hat{b}_{12} &= 4\hat{y}_{12} - 2\hat{y}_1 - 2\hat{y}_2 = 4\hat{y}_{12} - 2(\hat{y}_1 + \hat{y}_2) \\ \hat{b}_{13} &= 4\hat{y}_{13} - 2\hat{y}_1 - 2\hat{y}_3 = 4\hat{y}_{13} - 2(\hat{y}_1 + \hat{y}_3) \\ \hat{b}_{23} &= 4\hat{y}_{13} - 2\hat{y}_2 - 2\hat{y}_3 = 4\hat{y}_{23} - 2(\hat{y}_2 + {}''\hat{y}_3) \end{aligned} \tag{16.7.588}$$

Either in a generic way:

$$\hat{b}_i = \hat{y}_i \qquad \hat{b}_{ij} = 4\hat{y}_{ij} - 2(\hat{y}_i + 2\hat{y}_j) \tag{16.7.589}$$

If we make repeated measurements in quantities r_{ij}, r_i, r_j, then we have:

$$\hat{b}_i = \bar{y}_i \qquad \hat{b}_{ij} = 4\bar{y}_{ij} - 2(\bar{y}_i + 2\bar{y}_j) \tag{16.7.590}$$

Either by extension:

$$\begin{aligned} b_i &= \mathrm{E}(\hat{b}_i) = \mathrm{E}(\bar{y}_i) \\ b_j &= \mathrm{E}(\hat{b}_j) = \mathrm{E}(\bar{y}_j) \\ b_{ij} &= \mathrm{E}(\hat{b}_{ij}) = \mathrm{E}(4\bar{y}_{ij} - 2(\bar{y}_i + \bar{y}_j)) = 4\mathrm{E}(\bar{y}_{ij}) - 2\mathrm{E}(\bar{y}_i) - 2\mathrm{E}(\bar{y}_j) \end{aligned} \tag{16.7.591}$$

and for the variance (which will be useful for making inference on the values of the coefficients):

$$\begin{aligned}\mathrm{V}(b_i)&=\mathrm{V}(\bar{y}_i)=\frac{\hat{\sigma}_{y_i}^2}{r}\\ \mathrm{V}(b_j)&=\mathrm{V}(\bar{y}_j)=\frac{\hat{\sigma}_{y_j}^2}{r}\\ \mathrm{V}(b_{ij})&=\mathrm{V}(\hat{b}_{ij})=16\frac{\hat{\sigma}_{y_{ij}}^2}{r}+4\frac{\hat{\sigma}_{y_i}^2}{r}+4\frac{\hat{\sigma}_{y_j}^2}{r}\end{aligned}\tag{16.7.592}$$

Some statistical software and authors even consider that (we must then verify that this hypothesis is well satisfied with typically a Fisher test):

$$\hat{\sigma}_{y_i}^2=\hat{\sigma}_{y_j}^2=\hat{\sigma}_{y_{ij}}^2=\hat{\sigma}^2\tag{16.7.593}$$

Then it remains:

$$\begin{aligned}\mathrm{V}(b_i)&=\mathrm{V}(\bar{y}_i)=\frac{\hat{\sigma}^2}{r}\\ \mathrm{V}(b_j)&=\mathrm{V}(\bar{y}_j)=\frac{\hat{\sigma}^2}{r}\\ \mathrm{V}(b_{ij})&=\mathrm{V}(\hat{b}_{ij})=16\frac{\hat{\sigma}^2}{r}+4\frac{\hat{\sigma}^2}{r}+4\frac{\hat{\sigma}^2}{r}=24\frac{\hat{\sigma}^2}{r}\end{aligned}\tag{16.7.594}$$

Some statistical softwares assume that (so take care that this assumption is not rejected in the practice!):

$$\hat{b}_i=\mathcal{N}\left(b_i,\frac{\hat{\sigma}^2}{r}\right)\qquad\hat{b}_{ij}=\mathcal{N}\left(b_{ij},\frac{24\hat{\sigma}^2}{r}\right)\tag{16.7.595}$$

It is also interesting to measure the experimental variance of the model. Therefore in the general second degree case that is for recall written:

$$\hat{y}=\sum_i\hat{b}_ix_i+\sum_{i<j}\hat{b}_{ij}x_ix_j\tag{16.7.596}$$

Then we have:

$$\begin{aligned}\mathrm{V}(\bar{y})&=\mathrm{V}\left(\sum_i\hat{b}_ix_i+\sum_{i<j}\hat{b}_{ij}x_ix_j\right)=\mathrm{V}\left(\sum_i\bar{y}_ix_i+\sum_{i<j}(4\bar{y}_{ij}-2(\bar{y}+\bar{y}_j))x_ix_j\right)\\ &=\mathrm{V}\left(\sum_i\bar{y}_ix_i\right)+\mathrm{V}\left(\sum_{i<j}(4\bar{y}_{ij}-2(\bar{y}+\bar{y}_j))x_ix_j\right)\\ &=\sum_i\frac{\hat{\sigma}}{r}x_i^2+\sum_{i<j}24\frac{\hat{\sigma}}{r}(x_ix_j)^2=\frac{\hat{\sigma}}{r}\left(\sum_i x_i^2+24\sum_{i<j}(x_ix_j)^2\right)\end{aligned}\tag{16.7.597}$$

or if the measurement are not in equal quantities:

$$\mathrm{V}(\bar{y})=\bar{\sigma}^2\left(\sum_i\frac{x_i^2}{r_i}+24\sum_{i<j}\frac{(x_ix_j)^2}{r_{ij}}\right)\tag{16.7.598}$$

which can sometimes be found in the following way in the literature:

$$\begin{aligned}
\mathrm{V}(\bar{y}) &= \mathrm{V}\left(\sum_i \hat{b}_i x_i + \sum_{i<j} \hat{b}_{ij} x_i x_j\right) = \mathrm{V}\left(\sum_i \bar{y}_i x_i + \sum_{i<j} \left(4\bar{y}_{ij} - 2(\bar{y}_i + \bar{y}_j)\right) x_i x_j\right) \\
&= \left(\sum_i \bar{y}_i x_i - 2\sum_{i<j} (\bar{y}_i + \bar{y}_j) x_i x_j + \sum_{i<j} 4\bar{y}_{ij} x_i x_j\right) \\
&= \mathrm{V}\left(\sum_i \bar{y}_i \left[x_i - 2x_i \left(\sum_{j\neq i} x_j\right)\right] + \sum_{i<j} 4\bar{y}_{ij} x_i x_j\right)
\end{aligned} \tag{16.7.599}$$

and as:

$$\sum_j x_j = 1 \Leftrightarrow 1 - x_i = \sum_{j \neq i} x_j \tag{16.7.600}$$

Therefore it comes:

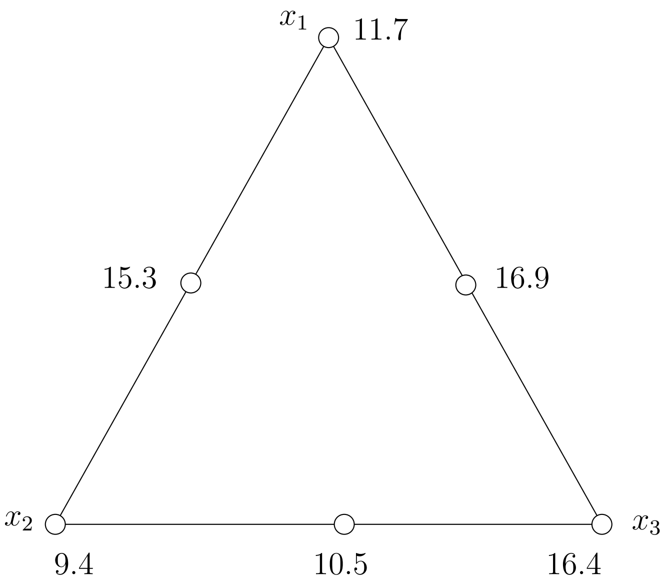
$$\begin{aligned}
\mathrm{V}(\bar{y}) &= \mathrm{V}\left(\sum_i \bar{y}_i \left[x_i - 2x_i \left(\sum_{j\neq i} x_j\right)\right] + \sum_{i<j} 4\bar{y}_{ij} x_i x_j\right) \\
&= \mathrm{V}\left(\sum_i \bar{y}_i \left[x_i - 2x_i(1 - x_i)\right] + \sum_{i<j} 4\bar{y}_{ij} x_i x_j\right) \\
&= \mathrm{V}\left(\sum_i \bar{y}_i \left[x_i(2x_i - 1)\right] + \sum_{i<j} 4\bar{y}_{ij} x_i x_j\right) = \mathrm{V}\left(a_i \bar{y}_i + \sum_{i<j} a_{ij} \bar{y}_{ij}\right) \\
&= \hat{\sigma}^2 \left(\sum_i \frac{a_i^2}{r_i} + \sum_{i<j} \frac{a_{ij}^2}{r_{ij}}\right)
\end{aligned} \tag{16.7.601}$$

with:

$$\begin{aligned}
a_i &= x_i(2x_i - 1) \\
a_{ij} &= 4x_i x_j
\end{aligned} \tag{16.7.602}$$

✍**Example:**

Let us consider that three constituents that are polyethylene x_1, polystyrene x_2 and polypropylene x_3, are mixed in order to obtain a fiber whose measure of interest is the elongation at a constant force with the following values of the mean elongations obtained for each of the points of the mixture design network:



with the measurements detailed in the corresponding table below:

Experiment point	**Polyethylene x_1**	**Polystyrene x_2**	**Polypropylene x_3**	**Observed elongation y_i**	**Average elongation $\bar{y}_i$**
1	1	0	0	11.0, 12.4	11.7
2	0.5	0.5	0	15.0,14.8,16.1	15.3
3	0	1	0	8.8,10.00	9.4
4	0	0.5	0.5	10.0,9.7,11.8	10.5
5	0	0	1	16.8,16.0	16.4
6	0.5	0	0.5	17.7,16.4,16.6	16.9
			Number of measurements:	15	
			Average:	13.54	

and we will use the model:

$$\hat{y} = \hat{b}_1 x_1 + \hat{b}_2 x_2 + \hat{b}_3 x_3 + \hat{b}_{12} x_1 x_2 + \hat{b}_{13} x_1 x_3 + \hat{b}_{23} x_2 x_3 \qquad (16.7.603)$$

with:

$$\hat{b}_i = \bar{y}_i \qquad \hat{b}_{ij} = 4\bar{y}_{ij} - 2(\bar{y}_i + \bar{y}_j) \qquad (16.7.604)$$

Which gives us:

$$\begin{aligned}
\hat{b}_1&=\bar{y}_1=\frac{(11.0+12.4)}{2}=11.7\\
\hat{b}_2&=\bar{y}_2=\frac{(8.8+10.0)}{2}=9.2\\
\hat{b}_3&=\bar{y}_3=\frac{(16.8+16.0)}{2}=16.4\\
\hat{b}_{12}&=4\bar{y}_{12}-2(\bar{y}_1+\bar{y}_2)=4\frac{(15.0+14.8+16.1)}{3}-2(11.7+9.4)=19.0\\
\hat{b}_{13}&=4\bar{y}_{13}-2(\bar{y}_1+\bar{y}_3)=4\frac{(17.7+16.4+16.6)}{3}-2(11.7+16.4)=11.4\\
\hat{b}_{23}&=4\bar{y}_{23}-2(\bar{y}_2+\bar{y}_3)=4\frac{(10.0+9.7+11.8)}{3}-2(9.4+16.4)=-9.6
\end{aligned}\tag{16.7.605}$$

So at the level of the influences, we already see that (without even needing to make a statistical test of significant difference):

$$\hat{b}_3>\hat{b}_1>\hat{b}_2\tag{16.7.606}$$

with the obvious that follows. We also observe already that:

$$\hat{b}_{12}>0\qquad\hat{b}_{13}>0\qquad\hat{b}_{23}<0\tag{16.7.607}$$

which means that the components 1 and 2, 1 and 3 contributes positively to the effect were are looking for when the interaction 2 and 3 decrease that later.

Now, let us calculate and check if we have well:

$$\hat{\sigma}_{y_i}^2=\hat{\sigma}_{y_j}^2=\hat{\sigma}_{ij}^2=\hat{\sigma}\tag{16.7.608}$$

Therefore:

$$\begin{aligned}
\hat{\sigma}_{y_1}^2&=0.98 &\quad \hat{\sigma}_{y_{12}}^2&=0.49\\
\hat{\sigma}_{y_2}^2&=0.72 &\quad \hat{\sigma}_{y_{13}}^2&=0.49\\
\hat{\sigma}_{y_3}^2&=0.32 &\quad \hat{\sigma}_{y_{23}}^2&=1.29\\
\hat{\sigma}^2&=9.6326
\end{aligned}\tag{16.7.609}$$

With for recall (requested by reader) for example:

$$\bar{y}_{12}=\frac{15.0+14.8+16.1}{3}=15.3\tag{16.7.610}$$

and then:

$$\hat{\sigma}_{y_{12}}^2=\frac{1}{3-1}\left((15.0-\bar{y}_{12})^2+(14.8-\bar{y}_{12})^2+(\bar{y}_{12}-16.1)^2\right)=0.49\tag{16.7.611}$$

and also:

$$\begin{aligned}\hat{\sigma} &= \frac{1}{15-1}\left[\begin{array}{c}(11-13.54)^2+(12.4-13.54)^2+(15-13.54)^2+(14.8-13.54)^2\\+(16.1-13.54)^2+(8.8-13.54)^2+(10-13.54)^2+(10-13.54)^2\\+(9.7-13.54)^2+(11.8-13.54)^2+(16.8-13.54)^2+(16-13.54)^2\\+(17.7-13.54)^2+(16.4-13.54)^2+(16.6-13.54)^2\end{array}\right]\\&=\frac{1}{14}134.856\cong 9.6326\end{aligned} \tag{16.7.612}$$

We see then obviously that we quite far from the assumption of equality of variance...

Therefore what can we do???

In fact apart to do the calculations of $\hat{\sigma}$ using the global average $\bar{y}$ as detailed just above, we notice that was is not realistic in the choice of the calculation of $\hat{\sigma}$ as done is that we consider that the average $\bar{y}$ should be the same for all the levels of the components which is obviously not true! Therefore as during our study of the ANOVA, we will use the residual variance V_R (that is we know is also named "grouped variance" or "pooled variance") and that is given therefore by:

$$\begin{aligned}\hat{\sigma}^2=\hat{\sigma}_R^2 &= \frac{1}{\sum_i r_i-1}\left[\begin{array}{c}(11-\bar{y}_1)^2+(12.4-\bar{y}_1)^2+(15-\bar{y}_{12})^2+(14.8-\bar{y}_{12})^2\\(16.1-\bar{y}_{12})^2+(8.8-\bar{y}_2)^2+(10-\bar{y}_2)^2+(10-\bar{y}_{23})^2\\(9.7-\bar{y}_{23})^2+(11.8-\bar{y}_{23})^2+(16.8-\bar{y}_3)^2+(16-\bar{y}_3)^2\\(17.7-\bar{y}_{13})^2+(16.4-\bar{y}_{13})^2+(16.6-\bar{y}_{13})^2\end{array}\right]\\&=\frac{6.56}{(2-1)+(3-1)+(2-1)+(3-1)+(2-1)+(3-1)}=\frac{6.56}{15-6}=\frac{6.56}{9}\\&\cong 0.73\end{aligned} \tag{16.7.613}$$

We see then that the compromise is much better ... (without however of course to be exact!). We then have using the results prove earlier above:

$$\begin{aligned}\hat{\sigma}_{b_i}&=\sqrt{\text{V}(b_i)}=\sqrt{\frac{\hat{\sigma}_R^2}{r}}=\sqrt{\frac{0.73}{2}}\cong 0.60\\\hat{\sigma}_{b_{ij}}&=\sqrt{\text{V}(b_{ij})}=\sqrt{\hat{\sigma}^2\left(\frac{16}{r_{ij}}+\frac{4}{r_i}+\frac{4}{r_j}\right)}=\hat{\sigma}\sqrt{\frac{16}{r_{ij}}+\frac{8}{r_i}}=\sqrt{0.73}\sqrt{\frac{16}{3}+\frac{8}{2}}\cong 2.61\end{aligned} \tag{16.7.614}$$

Our model is then traditionally written:

$$\hat{y}=\underbrace{11.7x_1}_{(0.60)}+\underbrace{9.4x_2}_{(0.60)}+\underbrace{16.4x_3}_{(0.60)}+\underbrace{19.0x_1x_2}_{(2.61)}+\underbrace{11.4x_1x_3}_{(2.61)}-\underbrace{9.6x_2x_3}_{(2.61)} \tag{16.7.615}$$

which corresponds exactly to the second decimal point to what a software like Minitab gives us. Then, to find the triplet that maximizes the function it is enough to use a solver like the one integrated into Microsoft® Excel®.

Let us now recall that:

$$\begin{aligned}\hat{b}_i&=\hat{\beta}_0+\hat{\beta}_i+\hat{\beta}_{ii}\\ \hat{b}_{ij}&=\hat{\beta}_{ij}-\hat{\beta}_{ii}-\hat{\beta}_{jj}\end{aligned}\tag{16.7.616}$$

And as in our example the model has no pure quadratic term, it reduces to:

$$\begin{aligned}\hat{b}_i&=\hat{\beta}_0+\hat{\beta}_i\\ \hat{b}_{ij}&=\hat{\beta}_{ij}\end{aligned}\tag{16.7.617}$$

We have then:

$$\begin{aligned}\hat{b}_1&=\hat{\beta}_0+\hat{\beta}_1=11.7\\ \hat{b}_2&=\hat{\beta}_0+\hat{\beta}_2=9.4\\ \hat{b}_3&=\hat{\beta}_0+\hat{\beta}_3=16.4\end{aligned}\tag{16.7.618}$$

We see then that to determine $\hat{\beta}_0$ we have to make a choice on one of the $\hat{\beta}_i$ (as we are not able to refer to the experimental value in the present case since the absence of any ingredient in the mixture will only give a empty entity and then it is quite difficult to test an elongation on nothing... but this, however, raises the principle of the approach). We will choose empirically $\hat{\beta}=0$, then it comes:

$$\begin{aligned}\hat{\beta}_3&=0\\ \hat{\beta}_0&=16.4\\ \hat{\beta}_1&=11.7-16.4=-47\\ \hat{\beta}_2&=9.4-16.4=-7.0\\ \hat{\beta}_{12}&=19.0\\ \hat{\beta}_{13}&=11.4\\ \hat{\beta}_{23}&=-9.6\end{aligned}\tag{16.7.619}$$

The model with the ordinate at the origin (at least if it had a physical sense) would then be written:

$$\hat{y}=\underbrace{16.4}_{0.60}-\underbrace{4.7x_1}_{(?)}-\underbrace{7.0x_2}_{(?)}+\underbrace{19.0x_1x_2}_{(2.61)}+\underbrace{11.4x_1x_3}_{(2.61)}-\underbrace{9.6x_2x_3}_{(2.61)}\tag{16.7.620}$$

with in brackets the variances that do not change for the relative coefficients. There are two coefficients whose variances will change!

We have indeed:

$$\text{V}(\hat{\beta}_i)=\text{V}(b_i-\hat{\beta}_0)=\text{V}(b_i)+\text{V}(\hat{b}_3)=1.2\tag{16.7.621}$$

Therefore:

$$\hat{\sigma}_{\beta_i}=\sqrt{\text{V}(\beta_i)}\cong 1.09\tag{16.7.622}$$

Hence:

$$\hat{y}=\underbrace{16.4}_{0.60}-\underbrace{4.7x_1}_{(1.09)}-\underbrace{7.0x_2}_{(1.09)}+\underbrace{19.0x_1x_2}_{(2.61)}+\underbrace{11.4x_1x_3}_{(2.61)}-\underbrace{9.6x_2x_3}_{(2.61)} \tag{16.7.623}$$

We can even make an ANOVA of the regression (see section Numerical Methods page 6335) as for the more" traditional" experimental design and as we have studied (proved!) in the section of Theoretical Computing, we will need the following values:

$$\begin{array}{ccccc} Q_T & = & Q_A & + & Q_R \\ \sum_i (y_i-\bar{y})^2 & = & \sum_i (\hat{y}_i-\bar{y})^2 & + & \sum_i (y_i-\hat{y}_i)^2 \\ N-1 & = & k-1 & + & N-k \end{array}$$

Then let us use the following table:

Observed value y_i	Predicted value $\hat{y}_i$	Residuals $(y_i-\hat{y}_i)$	Deviation $(y_i-\bar{y}_i)$	Deviation of regression $(\hat{y}_i-\bar{y})$
11	11.7	-0.7	-2.5	-1.8
12.4	11.7	0.7	-1.1	-1.8
15.0	15.3	-0.3	1.5	1.8
14.8	15.3	-0.5	1.3	1.8
16.1	15.3	0.8	2.6	1.8
10.0	9.4	0.6	-3.5	-4.1
8.8	9.4	-0.6	-4.7	-4.1
10.0	10.5	-0.5	-3.5	-3.0
9.7	10.5	-0.8	-3.8	-3.0
11.8	10.5	1.3	-1.7	-3.0
16.8	16.4	0.4	3.3	2.9
16.0	16.4	-0.4	2.5	2.9
17.7	16.9	0.8	4.2	3.4
16.4	16.9	-0.5	2.9	3.4
16.6	16.9	-0.3	3.1	3.4

Then we have:

$$\begin{array}{ccccc} Q_T & = & Q_A & + & Q_R \\ \sum_i (y_i-\bar{y})^2 & = & \sum_i (\hat{y}_i-\bar{y})^2 & + & \sum_i (y_i-\hat{y}_i)^2 \\ 138.88 & = & 128.32 & + & 6.56 \\ N-1 & = & k-1 & + & N-k \\ 14 & = & 5 & + & 9 \end{array}$$

We then have the following one-way fixed factor ANOVA table (see section Statistics page 1004):

Source	Sum of squares	χ^2 df	Average of squares	F	Critical F
Inter-Class	$Q_A = 128.32$	5	$\text{MSk} = \dfrac{128.32}{5}$	$\dfrac{\text{MSk}}{\text{MSE}} = 31.5$	$F_{5,9}^{\text{crit.}} = 3.481$
Intra-Class	$Q_R = 6.56$	9	$\text{MSE} = \dfrac{6.56}{5}$		
Total	$Q_T = 134.88$	14			

Table 16.58: **One way fixed factor ANOVA table for mixture design**

where the limit has been calculated with the following function of the spreadsheet software Microsoft® Excel® 14.0.7106:

`=F.INV(95%,5,9)=3.481`

Hence as 31.5 (corresponding p-value of 0.002%) is far above 3.481 we reject the null hypothesis H_0 that the model coefficients are non-significant (i.e. they are therefore significantly different from zero). In extenso this also leads us to reject the fact that the response surface is plane (since the coefficients of the interactions are also significant).

The reader interested can see the study of this mixture in the `R` or Minitab companion books.

16.7.4.6.2 Full Factorial Combined with Mixture Design-Crossed Design

In 11981 (holocene calendar), Cornell [113] published in its book the results of experimental studies where the quality of a fish patty was defined by both its composition and production process. The experiment included seven different mixtures prepared by mixing different species of fish and then subjecting the resulting patty to various cooking conditions and defined in accord with simplex centroid design (see figure below).

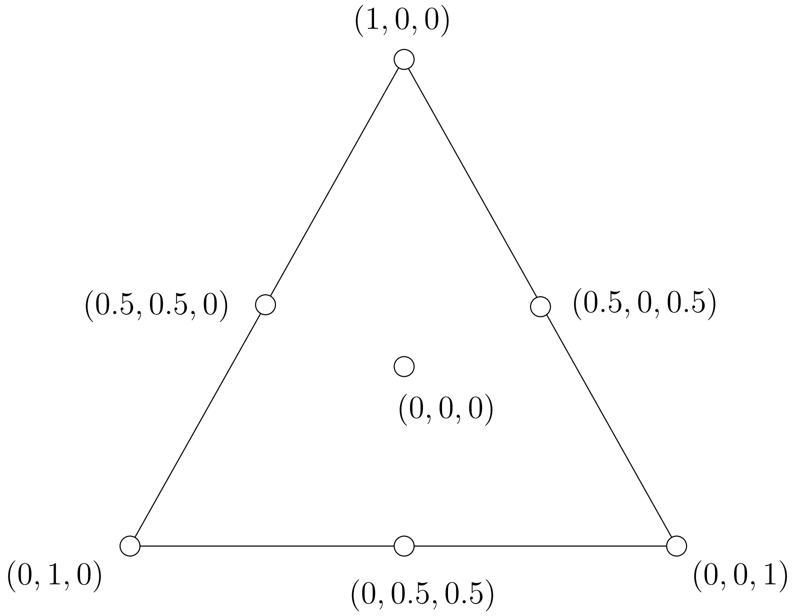


Figure 16.267: **Simplex Centroid Design**

The preparation procedure of the defined mixtures or a formulation included three control process factors with associated variation levels: baking temperature from 190 $[°\text{C}]$ to 218 $[°\text{C}]$; time in the oven from 25 [min] to 40 [min]; deep fat frying time from 25 [s] to 40 [s]. Process factor levels have been varied in accordance with the 2^3 full factorial experiment. Design of experiment 7×2^3 with 56 trials of a simplex centroid design $\times 2^3$ full factorial design has been sufficient for mathematical modelling of the observed phenomenon (see figure below):

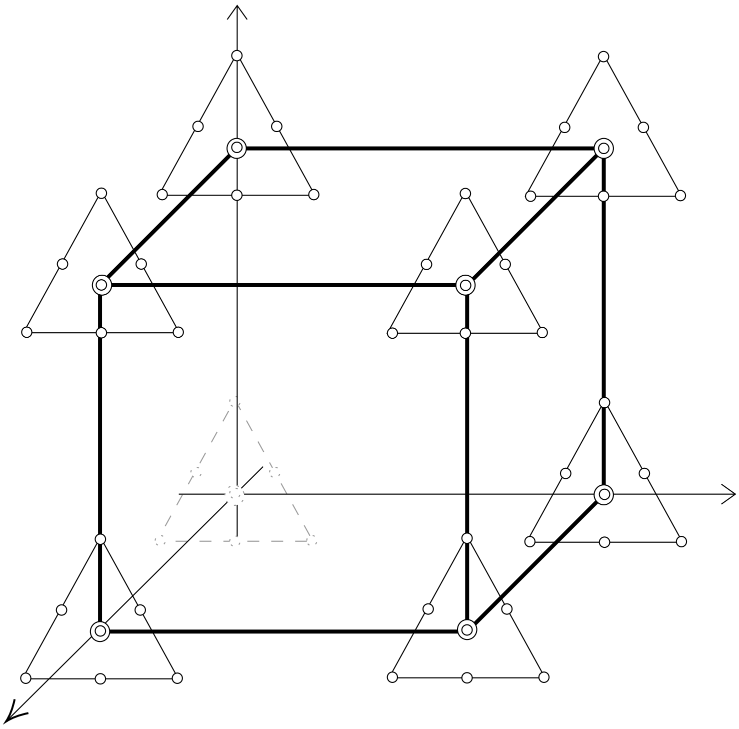


Figure 16.268: **Simplex centroid design in each point of a 2^3 full factorial experiment**

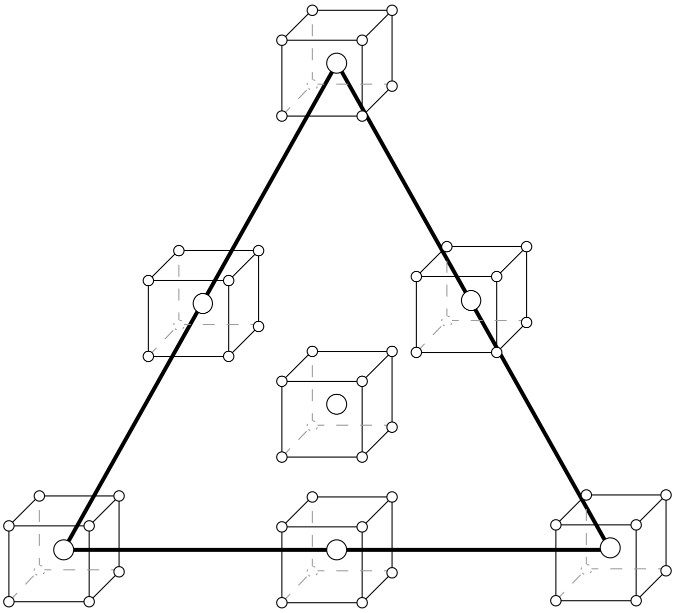


Figure 16.269: **2^3 factorial design in each point of a simplex centroid design**

A regression model with 56 regression coefficients or reduced regression model with 18 coefficients has been sufficient for an adequate description of the problem.

This was the first example of application of a mixture design × process factor design in experimental studies. Since such designs contain a relatively large number of factors, it is of interest to replace full factorial designs of process factors with fractional factor designs. Today all modern statistical softwares include the analysis of "design-crossed design".

✍**Example:**

Assume a design has 3 mixture components and 2 process variables, as illustrated in the above figure.

For the 3 mixture components, the following special cubic Scheffé model is used:

$$y=\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_{12}x_1x_2+\beta_{13}x_1x_3+\beta_{32}x_2x_3+\beta_{123}x_1x_2x_3 \tag{16.7.624}$$

For the 2 process variables the following model is used:

$$y=\alpha_0+\alpha_1z_1+\alpha_2z_2+\alpha_{12}z_1z_2 \tag{16.7.625}$$

The combined model with both mixture components and process variables is assumed, for being able to make an ANOVA, as following (quite strong assumption in the point of view of the physicist...):

$$\begin{aligned}y=&\sum_{i=1}^{3}\gamma_i^0x_i+\sum\sum_{i<j}^{3}\gamma_{ij}^0x_ix_j+\gamma_{123}^0x_1x_2x_3\\&+\left(\sum_{i=1}^{3}\gamma_i^1x_i+\sum\sum_{i<j}^{3}\gamma_{ij}^1x_ix_j+\gamma_{123}^1x_1x_2x_3\right)z_1\\&+\left(\sum_{i=1}^{3}\gamma_i^2x_i+\sum\sum_{i<j}^{3}\gamma_{ij}^2x_ix_j+\gamma_{123}^2x_1x_2x_3\right)z_2\\&+\left(\sum_{i=1}^{3}\gamma_i^{12}x_i+\sum\sum_{i<j}^{3}\gamma_{ij}^{12}x_ix_j+\gamma_{123}^{12}x_1x_2x_3\right)z_2\end{aligned} \tag{16.7.626}$$

The above combined model has total of $7\times4=28$ terms. By expanding it, we get the following model:

$$\begin{aligned}y=&\gamma_1^0x_1+\gamma_2^0x_2+\gamma_3^0x_3+\gamma_{12}^0x_1x_2+\gamma_{13}^0x_1x_3+\gamma_{23}^0x_2x_3+\gamma_{123}^0x_1x_2x_3\\&+\gamma_1^1x_1z_1+\gamma_2^1x_2z_1+\gamma_3^1x_3z_1+\gamma_{12}^1x_1x_2z_1+\gamma_{13}^1x_1x_3z_1+\gamma_{23}^1x_2x_3z_1+\gamma_{123}^1x_1x_2x_3z_1\\&+\gamma_1^2x_1z_2+\gamma_2^2x_2z_2+\gamma_3^2x_3z_2+\gamma_{12}^2x_1x_2z_2+\gamma_{13}^2x_1x_3z_2+\gamma_{23}^2x_2x_3z_2+\gamma_{123}^2x_1x_2x_3z_2\\&+\gamma_1^{12}x_1z_1z_2+\gamma_2^{12}x_2z_1z_2+\gamma_3^{12}x_3z_1z_2+\gamma_{12}^{12}x_1x_2z_1z_2+\gamma_{13}^{12}x_1x_3z_1z_2+\gamma_{23}^{12}x_2x_3z_1z_2\\&+\gamma_{123}^{12}x_1x_2x_3z_1z_2\end{aligned} \tag{16.7.627}$$

The combined model basically crosses every term in the mixture components model with every term in the process variables model. From a mathematical point of view, this model is just a regular regression model. Therefore, the traditional regression analysis method can still be used for obtaining the model coefficients and calculating the ANOVA table.

The reader can see a real application example of crossed designs in the Minitab companion book if necessary.

16.7.4.7 General DoE diagnostic tools

We have seen so far the almost all DOE use statistical analysis based on ANOVA, ANCOVA and obviously linear regression null hypothesis tests. But most DoE softwares propose a bunch of other more or less empirical diagnostics for the analysis of the designs and this is what we will focus below in details.

16.7.4.7.1 Lenth's PSE Pareto Margin Error for unreplicated factorial designs

G.E.P. Box and R.D. Meyer introduced in 11986 (holocene calendar) a method for assessing the sizes of contrasts in unreplicated factorial and fractional factorial designs. This was useful a technique and an associated graphical display popularly known as a "Bayes plot" made it even more effective but very hard to interpret. Therefore Russell V. Lenth proposed a new method 11989 (holocene calendar) that is still currently used in many statistical softwares.

The "Lenth's Pseudo Standard Error" (abbreviated "Lenth's PSE") is a common indicator in DoE softwares Pareto outputs (even if it is not real Pareto charts...) that provides an analytical basis for testing the effects in an unreplicated (single replicate) two-level factorial design[29]. Lenth's method estimates the variance of an effect from the smallest (presumably insignificant) effect estimates.

So... Lenth proposed a method for estimating the standard deviation of the effects on the basis that if the coded factors x_i (for simplification reasons but it also applies to non-coded and non-normalized effects) have their effects that follow a Normal distribution $E_i = \mathcal{N}(0,\sigma)$ then the median of the absolute value of the coded effects $|E_i|$ is in theory equal to the half-Normal distribution (see section Statistics page 796):

$$1.5M_e \cong \sigma \qquad (16.7.628)$$

and again (!) this last relation can be easily generalized to non-coded effects (just return back to that proof and add a non-zero mean in the development of the proof)!

What Lenth has proposed is first to eliminate all effects that satisfies:

$$|E_i| > 2.5 \cdot (1.5M_e) \qquad (16.7.629)$$

that is to say:

$$|E_i| > 2.5\sigma \qquad (16.7.630)$$

denoted s_0.

The `R` software gives with the **fdrtool** package:

```
>1-phalfnorm(2.5,theta=sqrt(pi/2),lower.tail=T,log.p=F)=0.012
```

---

[29] It should not then be necessary to precise that it therefore not applies to General factorial Designs

The reason why Lenth's made this choice are explained in a quite obscure way in its original article... in our point of view. His argument is that simulations shows that the $1.5M_e$ is overestimate in practice therefore eliminate factors with high amplitude effect seems to give a more accurate result.

After the elimination the proposition is to calculate a pseudo standard error with the remaining effects (that therefore should converge to the σ of all effects):

$$\boxed{\text{PSE} = \underset{|E_i|<2.5s_0}{1.5M_e(|E_i|)} \cong \sigma_{\text{effect}}} \tag{16.7.631}$$

where Lenth's assume that σ_{effect} is the standard error of the mean and that therefore we can do inference using the Student T distribution such at an α bilateral threshold we have the following margin error:

$$\text{ME} = T_{\alpha/2,m/3}\text{PSE} \tag{16.7.632}$$

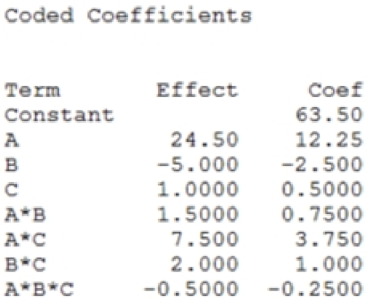
where m is the total number of effects and the division by 3 comes from a Monte Carlo study that was conducted by Lenth.

✍**Example:**

To make an example let us consider the following experiment that we already know (but we have removed the replication):

Trial N°	x_1	x_2	x_3	y_i
1	-1	-1	-1	59
2	$+1$	-1	-1	74
3	-1	$+1$	-1	50
4	$+1$	$+1$	-1	69
5	-1	-1	$+1$	50
6	$+1$	-1	$+1$	81
7	-1	$+1$	$+1$	46
8	$+1$	$+1$	$+1$	79

We now already how to perform the calculations manually of the effect so we will not come back on this topic. We will just communicate the output given by Minitab 17.1.3:



```
Coded Coefficients


Term        Effect      Coef
Constant               63.50
A            24.50     12.25
B           -5.000    -2.500
C           1.0000    0.5000
A*B         1.5000    0.7500
A*C          7.500     3.750
B*C          2.000     1.000
A*B*C      -0.5000   -0.2500
```

Therefore in absolute value, we have:

$$24.5, 5, 1, 1.5, 7.5, 2, 0.5$$

Ordered:

$$0.5, 1, 1.5, 2, 5, 7.5, 24.5$$

So it is immediate that the median is equal to:

$$M_e = 2 \tag{16.7.633}$$

So all the effects that are smaller or equal to $2.5(1.5M_e) = 7.5$ are (Lenth takes strictly smaller but Minitab doesn't follow this rule):

$$0.5, 1, 1.5, 2, 5, 7.5$$

Therefore:

$$\text{PSE} = 1.5\underset{|E_i|<2.5s_0}{M_e(|E_i|)} \cong \sigma_{\text{effect}} = 1.5 \cdot 1.75 = 2.625 \tag{16.7.634}$$

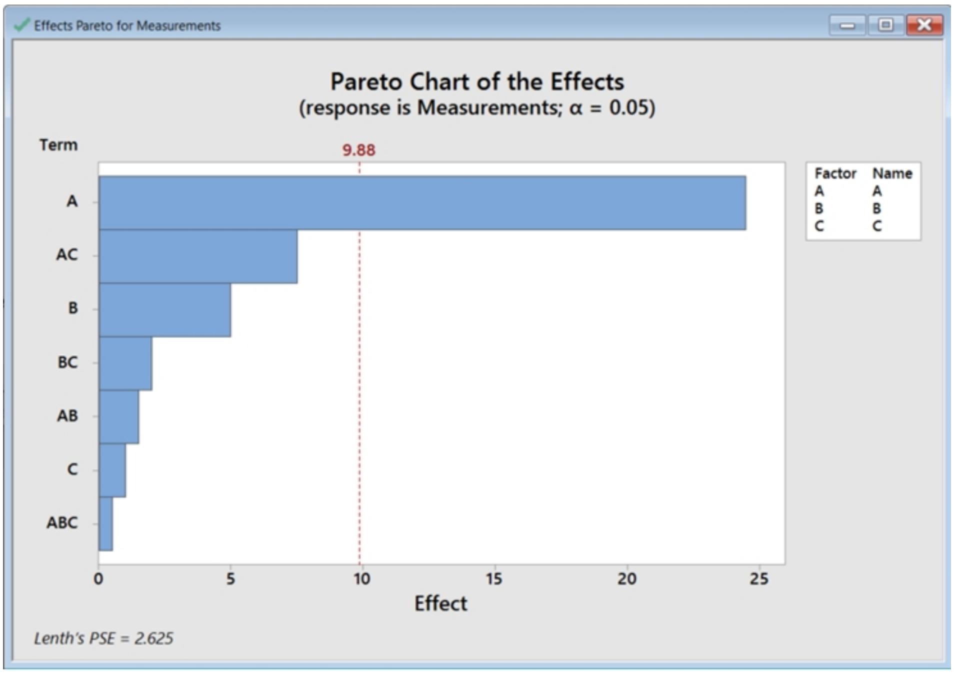
So finally:

$$\text{ME} = T_{\alpha/2,m/3} \cdot 2.625 = T_{5\%/2,7/3} 3.76 \cdot 2.625 \cong 9.8808 \tag{16.7.635}$$

Where the T value was obtained with `R` since Microsoft® Excel® doesn't accept fractional degrees of freedom:

```
>qt(.975,df=7/3)=3.764123
```

Minitab 17.1.3 makes a summary of all calculations done above as following:

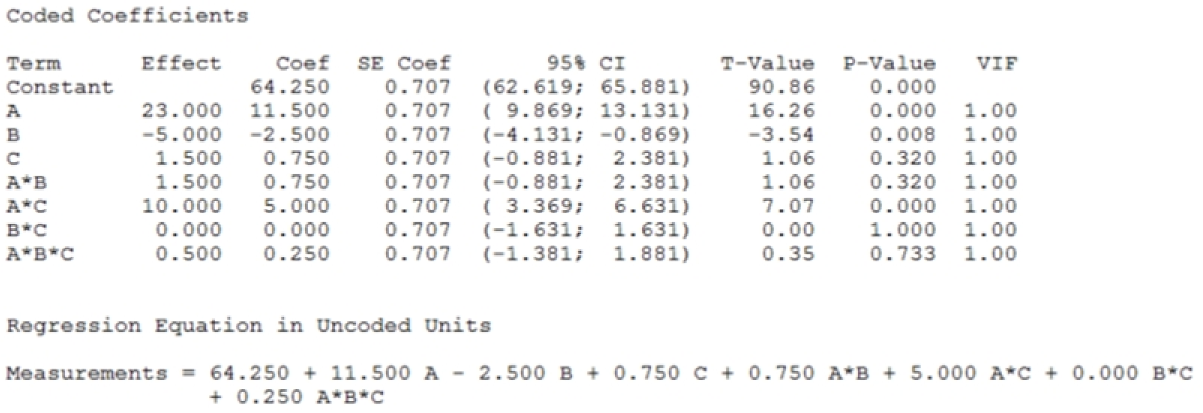

Effects Pareto for Measurements
Pareto Chart of the Effects
(response is Measurements; α = 0.05)
Term
9.88
A
AC
B
BC
AB
C
ABC
0
5
10
15
20
25
Effect
Factor Name
A A
B B
C C
Lenth's PSE = 2.625

16.7.4.7.2 Pareto Margin Error for replicated factorial designs

Let us recall that we have studied the following replicated experiment earlier above:

Trial N°	x_1	x_2	x_3	y_{i1}	y_{i2}	$\bar{y}_i$
1	-1	-1	-1	59	61	60
2	$+1$	-1	-1	74	70	72
3	-1	$+1$	-1	50	58	54
4	$+1$	$+1$	-1	69	67	68
5	-1	-1	$+1$	50	54	52
6	$+1$	-1	$+1$	81	85	83
7	-1	$+1$	$+1$	46	44	45
8	$+1$	$+1$	$+1$	79	81	80

and that we get by hand exactly the same results as given by Minitab 17.1.3:



```
Coded Coefficients

Term       Effect     Coef  SE Coef         95% CI         T-Value  P-Value   VIF
Constant            64.250    0.707  (62.619; 65.881)       90.86    0.000
A          23.000   11.500    0.707  ( 9.869; 13.131)       16.26    0.000   1.00
B          -5.000   -2.500    0.707  (-4.131; -0.869)       -3.54    0.008   1.00
C           1.500    0.750    0.707  (-0.881;  2.381)        1.06    0.320   1.00
A*B         1.500    0.750    0.707  (-0.881;  2.381)        1.06    0.320   1.00
A*C        10.000    5.000    0.707  ( 3.369;  6.631)        7.07    0.000   1.00
B*C         0.000    0.000    0.707  (-1.631;  1.631)        0.00    1.000   1.00
A*B*C       0.500    0.250    0.707  (-1.381;  1.881)        0.35    0.733   1.00


Regression Equation in Uncoded Units

Measurements = 64.250 + 11.500 A - 2.500 B + 0.750 C + 0.750 A*B + 5.000 A*C + 0.000 B*C
               + 0.250 A*B*C
```

Figure 16.270: **Replicated factorial design effects C.I. with Minitab 17.1.3**

It must be remembered that for the constant and all coefficients the T-value above was respectively of the form (and it must also be remembered that $\hat{\sigma}_{\text{eff}}$ can be calculated only if there are replications!):

$$\begin{gathered}\text{Average} \pm T_{\alpha/2,2^k}\frac{\hat{\sigma}_y}{\sqrt{N}}\\ E_{Gx_i} \pm T_{\alpha/2,N/r}\hat{\sigma}_{\text{eff}}\end{gathered} \tag{16.7.636}$$

Therefore we see that all Student's T distribution have always the same degrees of freedom!!! That will be written:

$$T_{\alpha/2,2^k} \tag{16.7.637}$$

So we could use the limit T-value (we have seen earlier above how to calculate it) in absolute value to compare the amplitude of every factor. And this is what Minitab give as output when the design is replicated:

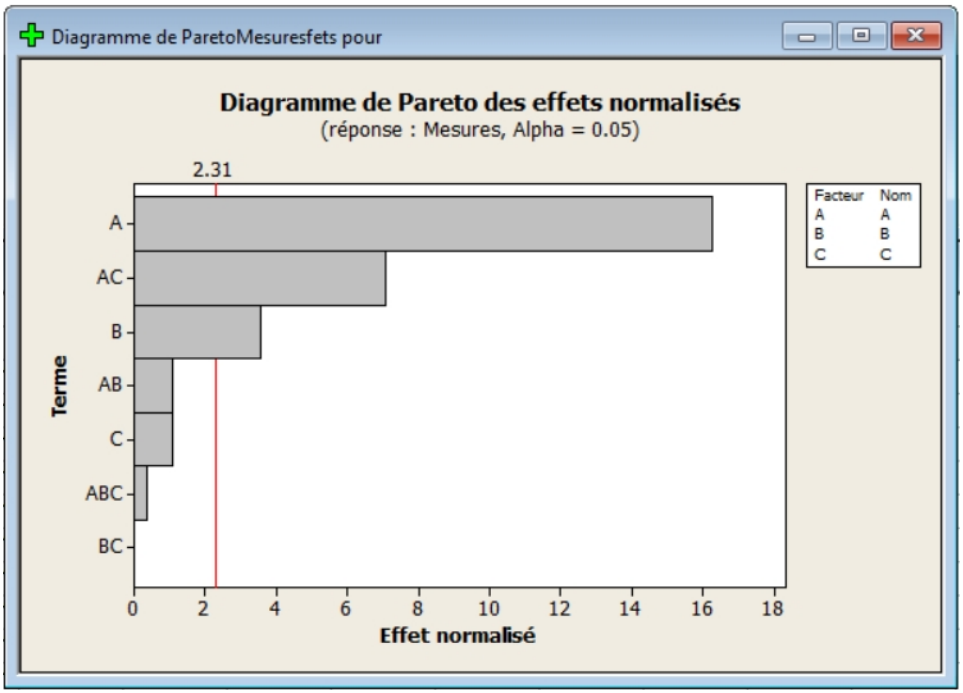


Figure 16.271: **Replicated factorial design T-value with Minitab 15.1.1**

Now the question that arise is: What is this red line and how it is calculated?

The choice (not rigorously justified as far as I know) used by most softwares is to compare all these values with the quantile $T_{\alpha/2,2^k}$ that gives in our case with Microsoft® Excel® 14.0.7177:

```
=TINV(97.5%,8)=2.30600414
```

and this is the value of the red line we can see on the plot above.

16.7.4.7.3 Desirability

The "desirability function" approach to simultaneously optimizing multiple equations was originally proposed by Edwin C. Jr Harrington (11980 according to holocene calendar). Essentially, the approach is to translate the functions to a common scale ($[0, 1]$), combine them using the geometric mean and optimize the overall metric. The equations may represent model predictions or other equations.

For example, desirability functions are popular in response surface methodology as a method to simultaneously optimize a series of quadratic models. A response surface experiment may use measurements on a set of outcomes. Instead of optimizing each outcome separately, settings for the predictor variables sought to satisfy all of the outcomes at once.

Also, in drug discovery, predictive models can be constructed to relate the molecular structures of compounds to characteristics of interest (such as absorption properties, potency and selectivity for the intended target). Given a set of predictive models built using existing compounds, predictions can be made on a large set of virtual compounds that have been designed but not necessarily synthesized. Using the model predictions, a virtual compound can be scored on how well the model results agree with required properties. In this case, ranking compounds on multiple endpoints may be sufficient to meet the scientist's needs.

Originally, Harrington used exponential functions to quantify desirability. In this text we will introduce simple discontinuous functions of G. Derringer and R. Suich.

Suppose that there are N equations or function to simultaneously optimize, denoted $f_i(\vec{x})$ with $i=1\ldots N$. For each of the i functions, an individual "desirability" function is constructed that is high when $f_i(\vec{x})$ is at the desirable level (such as a maximum, minimum, or target) and low when $f_n(\vec{x})$ is at an undesirable value. Derringer and Suich proposed three forms of these functions, corresponding to the type of optimization goal. For maximization of $f_i(\vec{x})$, the function:

$$d_i^{\text{max}} = \begin{cases} 0 & \text{if } f_i(\vec{x}) < L \\ \left(\dfrac{f_i(\vec{x})-L}{H-L}\right)^w & \text{if } L \leq f_i(\vec{x}) \leq H \\ 1 & \text{if } f_i(\vec{x}) > H \end{cases} \qquad (16.7.638)$$

can be used, where L is the lower limit that we allow to our variable of interest, H its high (maximal) value that we allow, and w is the weight that are all three chosen by the investigator. When the equation is to be minimized, they proposed the function:

$$d_i^{\text{min}} = \begin{cases} 0 & \text{if } f_i(\vec{x}) > H \\ \left(\dfrac{f_i(\vec{x})-H}{L-H}\right)^w & \text{if } L \leq f_i(\vec{x}) \leq H \\ 1 & \text{if } f_i(\vec{x}) < L \end{cases} \qquad (16.7.639)$$

and for target T situations:

$$d_i^{\text{target}} = \begin{cases} \left(\dfrac{f_i(\vec{x})-L}{T-L}\right)^{w_1} & \text{if } L \leq f_i(\vec{x}) \leq T \\ \left(\dfrac{f_i(\vec{x})-H}{T-H}\right)^{w_2} & \text{if } T \leq f_i(\vec{x}) \leq H \\ 0 & \text{otherwise} \end{cases} \qquad (16.7.640)$$

These functions are on the same scale (output in range $[0,1]$) and are discontinuous at the points L, H, and T. The values of w, w_1 or w_2 can be chosen so that the desirability criterion is easier or more difficult to satisfy. The scaling factors w are useful when one equation is of great importance than the other. Examples of the functions are given in the figure below:

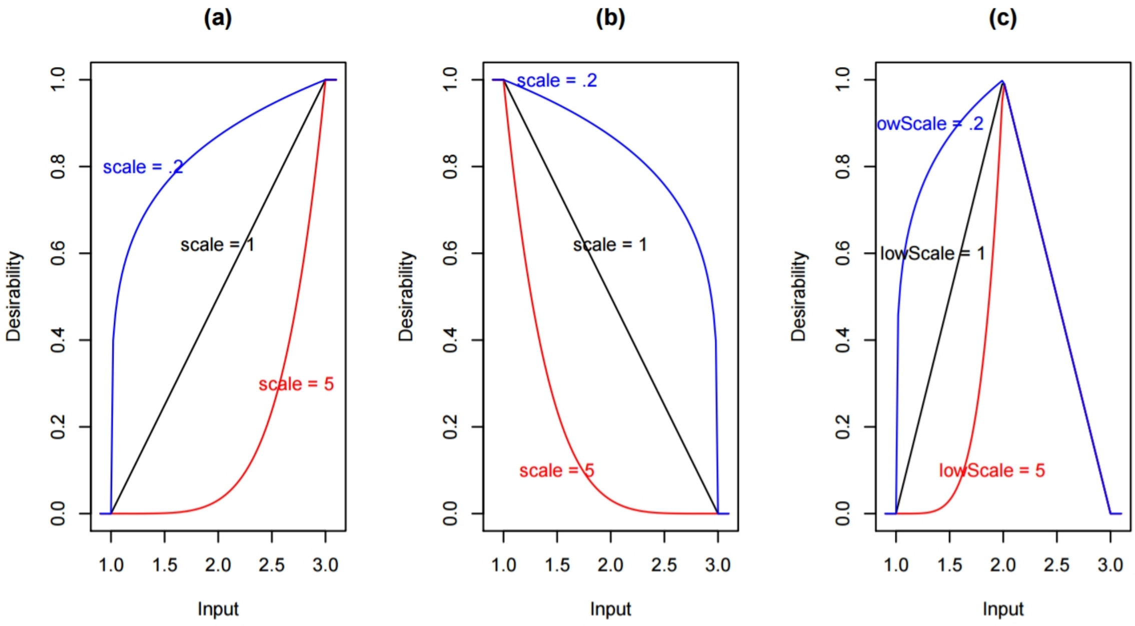


Figure 16.272: **Derringer and Suich desirability functions plot (source: Max Kuhn R Software)**

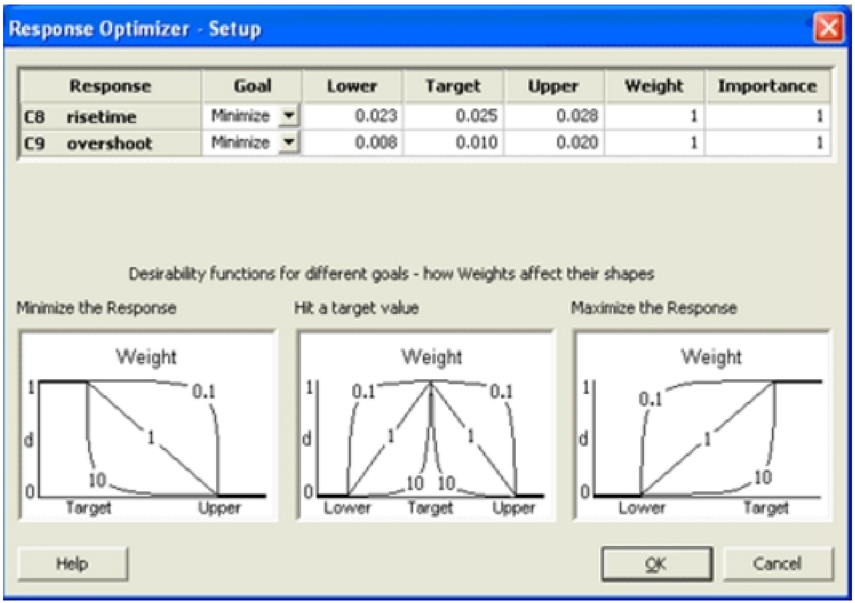
It should be noted that any function can be used to mirror the desirability of a model. For example, Del Castillo, Montgomery, and McCarville (11996 according to holocene calendar) develop alternative desirability functions that can be used in conjunction with gradient based optimization routines.

Given that the N desirability functions $d_1, \ldots, d_r$ are on the $[0, 1]$ scale, they can be combined to achieve an overall desirability function, D. One method of doing this is by the geometric mean and then we speak of "composite desirability":

$$\boxed{D = \left(\prod_{i=1}^{N} d_i \right)^{1/N}} \tag{16.7.641}$$

The geometric mean has the property that if any one model is undesirable ($d_i = 0$), the overall desirability is also unacceptable ($D = 0$). Once D has been defined and the prediction equations for each of the N equations have been computed, it can be use to optimize or rank the predictors.

In Minitab (see our companion book on that software) the composite desirability is the geometric mean of the individual desirabilities if and only if the "importance" parameter (whose value must be between 0.1 and 10) are all equals to 1.



Response Optimizer - Setup

	Response	Goal	Lower	Target	Upper	Weight	Importance
C8	risetime	Minimize	0.023	0.025	0.028	1	1
C9	overshoot	Minimize	0.008	0.010	0.020	1	1

Desirability functions for different goals - how Weights affect their shapes

Minimize the Response
Weight
1
0.1
d
1
10
0
Target
Upper
Hit a target value
Weight
1
0.1
0.1
d
1
1
0
10
10
Lower
Target
Upper
Maximize the Response
Weight
1
0.1
d
1
10
0
Lower
Target
Help
OK
Cancel

Figure 16.273: **Minitab 15.1.3 Design of experiment optimizer setup**

16.7.5 Quality Control on Reception (Lot Acceptance Sampling Plans)

The "quality control on reception" or simply "acceptance test" or "statistical batch control" or following NIST[30] "Lot Acceptance Sampling Plans (LASPs)" is an extremely important field, wide and mathematically technical of statistical quality control in the industry (food, pharma, consumer goods) and services companies (repetitive action, process controls, surveys, assessments). Its purpose, according to the norm NF X06-021, is to allow the application on a controlled batch (products or services) of one or the other following decisions: Acceptance or Rejection at different stages of a production / development to see if a batch is admissible for the following steps or to check the deliverable before sending it to the customer.

QUALITY CONTROL
APPROVED

Remark

We will introduce here only the basics of this subject because the special cases require considerable mathematical background! It is also for this reason why many companies (SMEs and multinationals) that use interns or students that don't have a mathematical background to define quality controls in order to save money are found later with considerable problems in compliance with laws and quality standards! However, what we introduce here should be the minimum-minimorum knowledge of any active quality responsible or quality engineer or responsible for reception control of suppliers in any organization (industrial or administrative) otherwise they should have absolutely no credibility from people hiring such "specialists" or from external consultants! Also beware of companies - especially multinationals - who are looking for quality specialists mastering Microsoft® Excel® or Microsoft Access. Because it will mean that they use non-professional tools to do a job that should to be done with the professional statistical tools (what Microsoft® Excel® and Microsoft Access are not) !!! So in terms of internal organization, you can ensure that such companies organize and analyse anything, anyhow, with unsuitable tools and therefore that there is a general mess internally.

The purpose of quality control on reception (acceptance sampling) is not to estimate the quality of a lot, but only to accept or reject it. The acceptance is not a substitute for process control methods such as control charts or capability analysis. A dynamic use of these tools during manufacturing will result in reducing and, in some cases, eliminate the need to perform incoming or outcoming inspection/controls.

There are three main approaches to evaluate a batch:

- Either accept the batch without inspection (useful in situations where absolute trust relations exists between the customer and supplier or we now that defects are impossible).

---

[30] National Institute of Standards and Technology

- Either take a sample (sampling) and draw conclusions about the whole batch from the sample (useful in most cases because inexpensive and highly motivates the supplier to improve quality but there is a risk of rejecting good batches or accept bad one). For example Mulhouse PSA controls 5% of vehicles coming out of their factories every day.

- Either inspect all units in the batch (especially useful when the supplier can not prove that its process control is capable of producing in the specifications or when the monitored characteristic is very critical).

For the result of a batch control by sampling to be scientifically reliable, it is trivially important that, when collecting a sample of individuals, no preference is given to individuals from the whole population. Ideally, each individual batch must have the same probability of being draw from the sample.

The characteristics controlled can be:

- **Qualitative (attributes):** the appearance of a product, the presence or absence of a non-compliant element, failure to comply with a calibre of control, an achieved result or not. Individuals are then directly classified as compliant or non-compliant identified by the statistical number of defects they have.

  Remark

  In my job of consultant I saw once a customer that had for process to reject definitively any supplier that do one and only one first failure. This is a zero tolerance approach and is quite not scientific as I warn this customer... But when people don't have any statistical culture it's a waste of time trying to explain them why it's very likely not the best kind of strategy...

- **Quantitative (variables or measurement):** characteristic measured on a continuous scale: volume, length, weight, ...

Let us notice finally that there are two ways to classify acceptance plans. The first family is to define the following two categories:

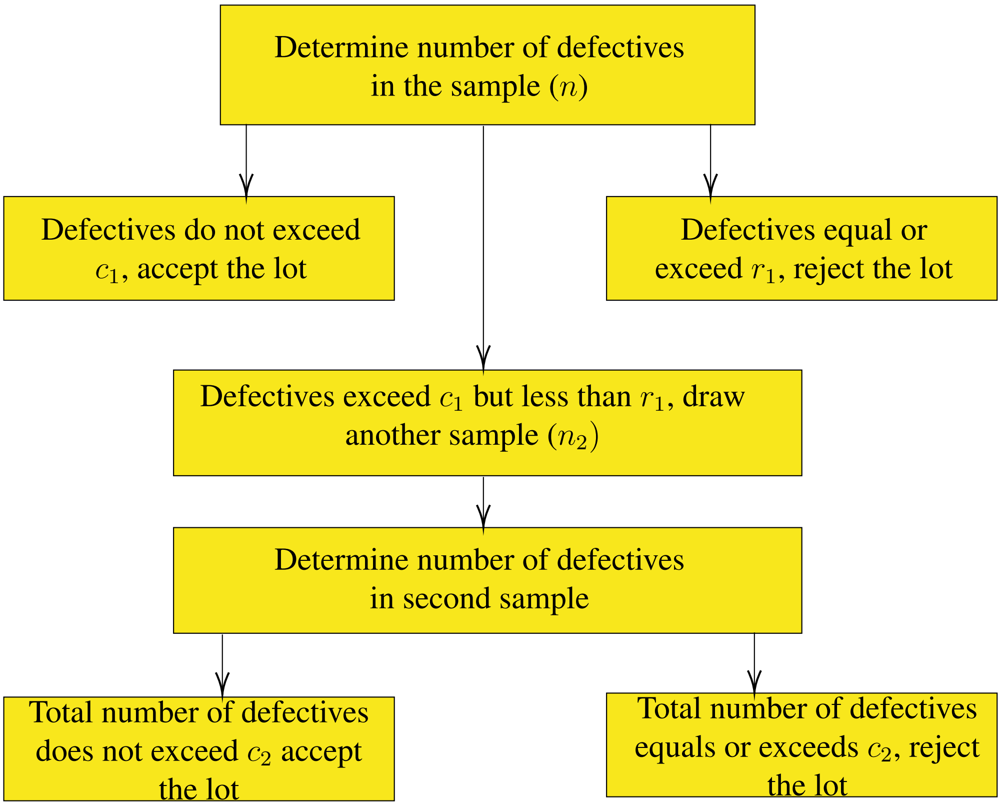
1. The "control of the proportion of non-compliant items by counting" or also simply named "attribute control": One or more characteristics of qualitative or quantitative type are controlled for each item in order to classify it as compliant or non-compliant according certain criteria. The decision is based on the number of non-compliant individuals found in the samples.

2. The "control of the proportion of non-compliant individuals by measuring" or also simply named "variable control". A characteristic is measured for each item and a decision is made based on the average and dispersion of the characteristic calculated on all the items collected.

We can also classify the acceptance plans depending on the number of samples taken.

- A "simple sampling plan" involves taking n items in a batch of size N and make a decision based on them.

- A "double sampling plan" involves taking a sample from the first batch and draw one of these three conclusions: (1) accept the batch, (2) reject the batch or (3) resample. If a second sample is taken, the information of the two samples are gathered to decide whether to accept or reject the batch (if we repeat this strategy, then we speak of "multiple sampling plan"). The idea is illustrated in the following figure:


Determine number of defectives in the sample (n)
Defectives do not exceed c_1, accept the lot
Defectives equal or exceed r_1, reject the lot
Defectives exceed c_1 but less than r_1, draw another sample (n_2)
Determine number of defectives in second sample
Total number of defectives does not exceed c_2 accept the lot
Total number of defectives equals or exceeds c_2, reject the lot

- A "progressive or sequential sampling plan" is the extreme case of multiple sampling plan. The items are removed one by one and after each sampling a decision is made: (1) accept, (2) reject or finally (3) take a new item.

When the defective units found in an inspected sample are removed and replaced with compliant products, we speak of "rectifying inspection". Obviously if the lot (batch) is rejected, 100% inspection may be performed and all defective units could be replace or a new lot send by the supplier (this is the subject of a cost analysis or legal contract terms between the supplier and the customer).

✍**Example:**

A soft drink producer wants to control a property of the bottles that are supplied to him. For this, it uses a simple reception plan by monitoring the proportion of nonconforming items by a measuring when a batch of bottles are delivered to him:

1. Collect a number n of units (bottles) of the batch (single sampling)
2. Measure pressure at which each of the n bottles broke (destructive testing)
3. Calculate the statistical indicators of the observed values.
4. Follow the acceptance decision rule or rejection that will be determined.

Remark

As for the control charts we will see further below, applying a sampling plan is a hypothesis test. It can lead to right or wrong decisions. The risks are those of accepting a non-compliant batch (Type II Error) or to reject a batch that is compliant (Type I Error).

Obviously sampling has advantages and disadvantages (pros & cons if you prefer). Here is the list of the most common advantages:

1. It's cheaper because you do not control everything ...
2. There is less manipulation of the products thus less waste
3. There needs to mobilize less people or control machines
4. This dramatically reduces inspection errors
5. The return of a whole batch just because of a non-compliant sample tends to motivate more the supplier to the quality ...

But it also has disadvantages such as:

1. The risk of accepting non-compliant batches (type II error) and reject good one (Type I error)...
2. Requires more intellectual skills of employees working in quality department (mathematical capacities...) and well-written contracts ...

16.7.5.1 Simple acceptance sampling plan by measurement for a unique tolerance with known standard deviation

For this type of sampling plan, we will assume that the sample n is much smaller than the total batch of size N (a factor of 10 at least) such that $n \ll N$ and the standard deviation σ is known!

The characteristic X measured for each item of the sample is assumed to follow a Normal distribution with identical mean μ and standard deviation σ (or variance) and the calculation of these last two parameters is based on the use of the sample mean (estimator of maximum likelihood of the mean as seen in the section Statistics at page 818) and of the unbiased estimator of the standard deviation (maximum likelihood estimator of the standard deviation as seen in the section Statistics at the same page as the previously mentioned one). This hypothesis implies that the product of the supplier is manufactured under statistical control!

✍**Example:**

In our example, the producer of soft drinks, he gets lots of $10,000$ bottles (so we will consider that the sample satisfy the above condition of $n \ll N$).

We write (see section Statistics page 692):

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{and} \qquad X = \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \tag{16.7.642}$$

Remark

The proper way to compute probabilities of acceptance should be to use the hypergeometric distribution as we have seen in the section Statistics (see page 672). But if the lot size is large relative to the sample size, the binomial distribution may be used. If the probabilities are quite small the binomial distribution can be approximated by a Normal distribution. This approximation is quite satisfactory if the lot size is more than ten times the sample size and non-compliant items are rare. In other words, if the lot size N is large enough to be declared as infinite, the distribution of the number of non-compliants in a random sample of n units will be binomial with parameters $\mu = np$ and $\sigma = np(1-p)$ as proved in the section Statistics (see page 692).

We will assume that the customer and the supplier have agreed on specification limits (tolerances). We denote will denote as usual these limits USL (Upper Specification Limit(and LSL (Lower Specification Limit(.

Given D the proportion (unknown!) of non-compliant products (defective), it is naturally given in the field statistical processes control by:

$$D = P(X < \text{LSL}) + P(X > \text{USL}) = P\left(Z < \frac{\text{LSL}-\mu}{\sigma}\right) + P\left(Z > \frac{\text{USL}-\mu}{\sigma}\right) \tag{16.7.643}$$

In practice, in order to save money in the context of quality tests, we try to reduce (reformulate) any control problem to a single calibre. In fact the problem can always with or without reformulation be reduced to have a proportion of non-compliant below the LSL or above USL to reject the batch if the supplier's production is in statistical control. Hence the fact that in practice the mathematical developments are made only in relation to a single bound which we will denote by BSL for "batch specification limit".

Thus, we reduce the problems of single sampling (or we manage the problems such that it can be reduced to such a formulation):

$$D=P(X<\text{BSL}) \qquad \text{or} \qquad D=P(X>\text{BSL}) \tag{16.7.644}$$

✍**Example:**

In the context of our companion example, the producer of soft drinks, it is the pressure resistance of a batch of $10,000$ bottles that we would like to control. The producer has agreed with its supplier to a lower specification limit for this property, chosen at 22.5 limits a majority of the bottles must exceed without blow up. So:

$$D=P(X<\text{LSL})=P(22.5<\text{LSL})$$

We will assume that past data were used to estimate precisely the standard deviation of the blow up pressure limit and which can therefore be considered as known and equal to 1.5 and is under statistical control.

Because obviously if the proportion of non-compliant items is more (or less) that defines the quality contract, it will be also the same the opposite bound in the same proportions as the supplier is supposed to have its manufacturing process under statistical control!!!

Before going further into the mathematical developments, there is an important thing to remember:

If we imagine ourselves to be a supplier, it is obvious that we would like to have a level of quality such that the cumulative probability that the customer mistakenly reject a batch is small (because as a supplier we are contractually obliged to take the back the batch even if a full second sampling control later shows that he seems to be wrong!). For this purpose, the supplier defines a value D_s of the percentage of non-compliant products, named "Acceptable Quality Level" (AQL), below which he assumes that the batch may be rejected only very rarely. In addition, the supplier also defines α as the maximum cumulative probability of seeing a batch being rejected that has a percentage of non-conformity less than or equal to D_f (so it is the cumulative probability that a customer would reject by error a batch). This is written as:

$$P(\text{Reject a batch}|D\leq D_s)\leq\alpha \tag{16.7.645}$$

more of often written:

$$\boxed{P(\text{Reject a batch}|D\leq \text{AQL})\leq\alpha} \tag{16.7.646}$$

where α is named course ... "supplier's risk[31]" and is generally in the range of 0.1 to 10%.

---

[31] α is the supplier's risk as it represents the likelihood that given a lot whose defect/defective rate is equal to the AQL, it will be improperly rejected. The higher the α, the more likely "acceptable" lots will be rejected causing harm to the supplier.

In other words, the Acceptable Quality Level is the defect/defective rate that we wish to accept a high proportion of the time $(1-\alpha)$. That is, given for example an AQL of 1.5% and an α of 5%, we would expect to accept lots that had 1.5% defective units 95% of the time.

Finally, if we imagine that we are the customer, it is obvious that we would like the cumulative probability of wrongly accepting a batch that does not meet the contract quality to be quite small for obvious cost reasons. For this, the customer must on his side defines a value D_c of the proportion of non-compliant items, know under the name "Quality Level Limit" (QLL) or following NIST "Lot Tolerance Percent Defective" (LTPD) or following Minitab Software "Rejectable Quality Level" (RQL), beyond which the customer considers that the batch can be accepted wrongly only rarely. He also defines β as the maximum cumulative probability of having to accept a batch that has a proportion of non-compliant item greater than or equal to D_c (so it is the cumulative probability that wrongly accept a batch of the supplier). This is written as:

$$P(\text{Accept a batch}|D \geq D_c) \leq \beta \qquad (16.7.647)$$

more often written:

$$\boxed{P(\text{Accept a batch}|D \geq \text{QLL}) \leq \beta} \qquad (16.7.648)$$

where β is named obviously ... "customer risk[32]" and is also generally in the range of 5 to 10%.

In other words, the Rejectable Quality Level (RQL) is the defect/defective rate that we wish to reject a high proportion of the time $(1-\beta)$. That is, given for example a RQL of 7.5% and a β of 10%, we would expect to reject lots that had 7.5% defective units 90% of the time (accept only 10% of the time).

Where the ANSI (MIL-STD-105) is an AQL system that protects manufacturer/supplier (if the quality is at the AQL let us make sure we can ship), RQL systems protect the consumer/customer (if the quality is at the RQL let us make sure we reject).

To summarize the situation may be illustrated by the following table (which is similar in all points to the table of Type I and II errors already seen in the section Statistics):

	State of Nature	
Decision resulting from sampling	The batch is not good (QLL) (H_0 **false**)	The batch is good (AQL) (H_0 **true**)
Reject (batch) H_0	Correctly reject null decision ($1-\beta$: Power of the test)	Type I Error (Risk α)
Fail to reject (batch) H_0	Type II Error (Risk β)	Correct decision

---

[32] β is the customers' risk (or "consumer's risk") as it represents the likelihood that given a lot whose defect/defective rate is equal to the RQL, it will be found acceptable. The higher the β, the more likely "rejectable" lots will be accepted causing harm to the customer (consumer).

Note that the US Government is moving to eliminate the use of AQLs in all contracts[33] «...to support the movement away from an AQL-based inspection (detection) strategy to implementation of an effective prevention-based strategy including a comprehensive quality system, continuous improvement...» (MIL-STD 1916, 1-APR 1996, p1).

✍**Example:**

In the context of our companion example - the producer of soft drinks - which receives batches of $10,000$ bottles whose pressure resistance (BSL) must be greater than 22.5 with a standard deviation σ under statistical control equal to 1.5, the supplier and the customer each have fixed the risks they agree to take:

- The supplier would like that a batch with less than 1% of non-compliant to be rejected by error by the customer to be at maximum of 5% of cases (cumulative probability of type I error). Thus:

$$P(\text{Reject a batch}|D\leq \text{AQL})\leq\alpha\Rightarrow P(\text{Reject a batch}|D\leq 1\%)\leq 5\%$$

- The customer would like that a batch withmore than 5% of non-compliant bottles to be wrongly accepted in less than 10% of cases (cumulative probability of type II error). Thus:

$$P(\text{Accept a batch}|D\leq \text{QLL})\geq\alpha\Rightarrow P(\text{Accept a batch}|D\geq 5\%)\leq 10\%$$

Remark

The decision rule to accept or reject a batch for a control by measurement is generally based on the estimated arithmetic mean of the characteristic of the sample rather than based on a estimate of non-compliant proportion.

For what follows, when we write:

$$D=P(X<\text{LSL}) \tag{16.7.649}$$

the specificity is that in practice D is requested and σ is given. The remainder is either to calculate or to eliminate (indeed the quality practitioner should normally only talk in terms of cumulative probability and proportions or factors to define its quality policy and LSL is not a cumulative probability neither a percentage or a factor independent of the thing that we analyse!). Thus we write:

$$D=P(X<\text{LSL})=P\left(Z<\frac{\text{LSL}-\mu}{\sigma}\right) \tag{16.7.650}$$

So with the usual score (centered reduce variable as seen in the section Statistics page 696):

$$Z=\frac{\text{LSL}-\mu}{\sigma} \tag{16.7.651}$$

---

[33] This is typically the case of "Accept on none (defects)" plans. Fore more information see the Minitab companion book or [509]

Obviously we draw from this that limit average value of the sample is:

$$\mu_D = \text{LSL} - Z_D\sigma \qquad (16.7.652)$$

where Z_D is the level D of percentile of the standard centered Normal distribution (as usual). Unfortunately, we still have LSL that is there. Can we eliminate it? For this we have in the context that if the measurement (or manufacturing) is under statistical control:

$$\bar{X} = \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \qquad (16.7.653)$$

The cumulative probability to accept (we could have done it for the rejection but we had to choose one of the both for the example...) the batch when the non-compliant rate is D is deducted in the context of using the average of the measurements becomes:

$$P_D = P(\bar{X} \geq \text{LSL} + k\sigma) \qquad (16.7.654)$$

in the special case where the measurement should not be less than some given value BSL (as is the case of the above example with the soft drink bottles!). Obviously k could be a real positive or negative number as we know but not equal to Z_D and we will how to find a way to determine it from the supplier and customer risk level!!!

Hence:

$$P_D = P(\bar{X} \geq \text{LSL} + k\sigma) = P\left(\frac{\bar{X} - \mu_D}{\frac{\sigma}{\sqrt{n}}} \geq \frac{\text{LSL} + k\sigma - \mu_D}{\frac{\sigma}{\sqrt{n}}}\right) = P\left(Z \geq \frac{LSL + k\sigma - \mu_D}{\frac{\sigma}{\sqrt{n}}}\right)$$
$$= P\left(Z \geq \frac{\text{LSL} + k\sigma - (\text{LSL} - Z_D\sigma)}{\frac{\sigma}{\sqrt{n}}}\right) = P\left(Z \geq \frac{k\sigma + Z_D\sigma}{\frac{\sigma}{\sqrt{n}}}\right) = P\left(Z \geq \sqrt{n}(k + Z_D)\right) \qquad (16.7.655)$$

Then we have a very interesting result! The cumulative probability to accept the batch when the non-compliant rate is D is then given in the special case there where the measurement should not be below a specific limit by:

$$P_D = P(Z \geq \sqrt{n}(k + Z_D)) \qquad (16.7.656)$$

and the probability to reject it in the case of a measure beyond which we should not go:

$$P_D = P(Z \leq \sqrt{n}(k + Z_D)) \qquad (16.7.657)$$

and what is good in this result is that the problem is then reduced to a percentile Z_D that is easy to choose (norm or policy choices), of the sample size n and a factor k so only independent elements of the type of measurement itself and that can then be used in any job for any type of object and facilitates the writing of contract (plus we have eliminated the explicit use of the standard deviation) !!

Obviously, as we know, the supplier and the customer must agree to a sampling plan corresponding to build quality tests meeting expectations of the one and the other. What we know so far is that they respectively require an α (supplier risk) and a β (customer risk) that allows them to immediately calculate the $Z_{D,\alpha}, Z_{D,\beta}$ respectively. However it is necessary to calculate the number of individuals n sampled and the factor k that satisfy the respective requirements.

To do this, remember that if α is the supplier risk (cumulative probability that the customer mistakenly rejects the batch) and the β the customer risk (cumulative that he wrongly accepts supplier batch) and that:

$$P_D = P(Z \geq \sqrt{n}(k+Z_D)) \tag{16.7.658}$$

the cumulative probability of accepting a batch with a lower limit, then we must (!) have:

$$\begin{aligned} 1-\alpha &= P(Z \geq \sqrt{n}(k+Z_{D,\alpha})) \\ \beta &= P(Z \geq \sqrt{n}(k+Z_{D,\beta}) \end{aligned} \tag{16.7.659}$$

or after rearrangement:

$$\begin{aligned} \alpha &= P(Z \leq \sqrt{n}(k+Z_{D,\alpha})) \\ 1-\beta &= P(Z \leq \sqrt{n}(k+Z_{D,\beta}) \end{aligned} \tag{16.7.660}$$

To understand why we write these both relations the reader can take the two extreme situation where $\alpha=\beta=0$ and $\alpha=\beta=1$ (with their respective $Z_{D,\alpha}, Z_{D,\beta}$) to better understand (we can detail on request with an explicit example).

If we solve this system we can get the n and k we are looking for so that the customer and supplier get the sampling plan their both expecting!

So we can rewrite the previous system as:

$$\begin{aligned} Z_\alpha &= \sqrt{n}(k+Z_{D,\alpha}) \\ Z_{1-\beta} &= \sqrt{n}(k+Z_{D,\beta}) \end{aligned} \tag{16.7.661}$$

Therefore:

$$\frac{Z_\alpha}{\sqrt{n}} - Z_{D,\alpha} = \frac{Z_{1-\beta}}{\sqrt{n}} - Z_{D,\beta} = k \tag{16.7.662}$$

from which we can take out n:

$$\begin{aligned} &\frac{Z_\alpha}{\sqrt{n}} - \frac{Z-1-\beta}{\sqrt{n}} = Z_{D,\alpha} - Z_{D,\beta} \Leftrightarrow \frac{1}{\sqrt{n}} = \frac{Z_{D,\alpha}-Z_{D,\beta}}{Z_\alpha - Z_{1-\beta}} \\ &\Leftrightarrow \sqrt{n} = \frac{Z_\alpha - Z_{1-\beta}}{Z_{D,\alpha}-Z_{D,\beta}} \end{aligned} \tag{16.7.663}$$

Finally:

$$\boxed{n = \left(\frac{Z_\alpha - Z_{1-\beta}}{Z_{D,\alpha}-Z_{D,\beta}}\right)^2} \tag{16.7.664}$$

and for k:

$$\begin{aligned} k &= \frac{Z_\alpha}{\sqrt{n}} - Z_{D,\alpha} = \frac{Z_\alpha}{\dfrac{Z_\alpha - Z_{1-\beta}}{Z_{D,\alpha}-Z_{D,\beta}}} - Z_{D,\alpha} = \frac{Z_\alpha(Z_{D,\alpha}-Z_{D,\beta})}{Z_\alpha - Z_{1-\beta}} - Z_{D,\alpha} \\ &= \frac{Z_\alpha(Z_{D,\beta}-Z_{D,\alpha})}{Z_{1-\beta}-Z_\alpha} - Z_{D,\alpha} = \frac{Z_\alpha(Z_{D,\beta}-Z_{D,\alpha}) - Z_{D,\alpha}(Z_{1-\beta}-Z_\alpha)}{Z_{1-\beta}-Z_\alpha} \\ &= \frac{Z_\alpha Z_{D,\beta} - Z_\alpha Z_{D,\alpha} - Z_{D,\alpha}Z_{1-\beta} - Z_{D,\alpha}Z_\alpha}{Z_{1-\beta}-Z_\alpha} \end{aligned} \tag{16.7.665}$$

So that finally:

$$\boxed{k=\frac{Z_\alpha Z_{D,\beta}-Z_{D,\alpha}Z_{1-\beta}}{Z_{1-\beta}-Z_\alpha}} \tag{16.7.666}$$

✍**Examples:**

E1. In the context of our companion example - the producer of soft drinks - which receives batches of $10,000$ bottles whose pressure resistance (LSL) must be at least of 22.5 with a standard deviation under statistical control equal to 1.5, remember that the supplier and the customer each have fixed the risks they are ok to assume such that:

$$\begin{aligned}P(\text{Reject a batch}|D\leq\text{AQL})\leq\alpha\Rightarrow P(\text{Reject a batch}|D\leq 1\%)\leq 5\%\\ P(\text{Accept a batch}|D\leq\text{QLL})\geq\alpha\Rightarrow P(\text{Accept a batch}|D\geq 5\%)\leq 10\%\end{aligned}$$

We then have using Microsoft® Excel® 11.8346:

$$n=\left(\frac{Z_\alpha-Z_{1-\beta}}{Z_{D,\alpha}-Z_{D,\beta}}\right)^2=\left(\frac{\texttt{NORMSINV(5\%)}-\texttt{NORMSINV(1-10\%)}}{\texttt{NORMSINV(1\%)}-\texttt{NORMSINV(5\%)}}\right)^2 \tag{16.7.667}$$

and (calculated with the same Microsoft® Excel® functions so we omit to explicit them again):

$$k=\frac{Z_\alpha Z_{D,\beta}-Z_{D,\alpha}Z_{1-\beta}}{Z_{1-\beta}-Z_\alpha}\cong 1.943 \tag{16.7.668}$$

The sampling plan is then to collect and analyse a sample of size $n=18$, calculate the arithmetic average $\bar{X}$ and apply the following decision rule:

$$\begin{aligned}\text{Accept the batch if:}\bar{X}\geq\text{LSL}+k\sigma=22.5+1.943\cdot 1.5\cong 25.4\\ \text{Reject the batch if:}\bar{X}<\text{LSL}+k\sigma\cong 25.4\end{aligned} \tag{16.7.669}$$

Let us suppose now that a sample size of 18 is taken and gives the following results:

$$\begin{gathered}25.6,26.0,23.0,27.0,27.5,29.0,28.5,26.0,25.6\\ 25.8,26.5,28.8,27.3,25.2,27.1,29.8,26.5,27.8\end{gathered}$$

Then we have:

$$\bar{X}=26.8\geq 25.4$$

and therefore the batch is accepted. You should know that many practitioners prefer the following calculation using a "quality index", denoted Q, simply using the following equation:

$$Q=\frac{\bar{X}-LSL}{\sigma}=\frac{26.8-22.5}{1.5}\cong 2.86>k$$

therefore the batch is accepted.

Remark

In the latter case it is still recommended in practice to calculate the unbiased estimator of the standard deviation of the sample to ensure it is not too far from the standard deviation of the contract (even if it may not be the same for small samples)!

E2. A second typical calculation made by practitioners (often just for curiosity or sometimes to make a quality report to the board committee) resulting from the first example is: What is the cumulative probability that we reject with n and k given, a batch of which the proportion of non-compliant D would be 5%? Then, we use the following proved relation:

$$P_D = P(Z \leq \sqrt{n}(k + Z_D))$$

Thus with the previous values:

$$P_D = P(Z \leq \sqrt{18}(1.943 + \texttt{NORMSINV(5\%)})) \cong 89.70\%$$

and the cumulative probability that we accept the batch is then of $1 - 89.70\%$ or 10.29%.

16.7.5.1.1 Calculation of the parameters using the norm AF-X06-023

The norm AFNOR X06-023 offers a slightly different method to determine k and n, not directly based on the customer's risk and the supplier's risk, but only base on the "Acceptable Quality Level" (AQL.). In addition, this norm also takes into account the size N of the tested batch (considered previously as infinite) and of the desired level of control.

Remark

The equivalent international norm to the AFNOR X06-023 seems to be the ISO 3951:1981 that to my knowledge gives very different values from that of the AFNOR for reasons that I don't understand quite well (I hope it's just because I did not understand something...).

The method proposed by AFNOR X06-023 is of the type "recipe" but is requested by certain European standards so the industry have no choice but to apply it. But at the end, the numerical values are relatively close to those obtained with the previously proved relations. In fact their real usefulness is that they are simple to implement in much more complex situations than those we have just studied.

Depending on the context, the norm proposes to apply different levels of control:

- **Normal control**: that has to be chosen when we are in close trusted relation with the supplier and we have a look in real time on the quality of its production using computer remote monitoring tools .
- **Reinforced control**: which is stricter than the normal control (n is greater) and is intended

to better protect customers against the risk of having to accept bad batches (Type I Error). This type of control should be performed temporarily when there are serious reasons for considering that the production quality is not (or is no more) what is should be until we return back to the normal (deliveries are suspended between this time).

- **Reduced control**: that is the most economical one and can be applied when based on previously tested and accepted batches, we can believe that the supplier master well the quality its processes

✍**Examples:**

E1. In the context of our companion example - the producer of soft drinks - which receives batches of $10,000$ bottles whose pressure resistance (LSL) must be at least of 22.5 with a standard deviation under statistical control equal to 1.5, remember that the supplier and the customer each have fixed the risks they are ok to assume such that:

$$\begin{gathered}P(\text{Reject a batch}|D\leq\text{AQL})\leq\alpha\Rightarrow P(\text{Reject a batch}|D\leq 1\%)\leq 5\%\\ P(\text{Accept a batch}|D\leq\text{QLL})\geq\alpha\Rightarrow P(\text{Accept a batch}|D\geq 5\%)\leq 10\%\end{gathered}$$

We must then use the norm AFNOR X06-023 with a NQA of 1%.

The norm requires the choice of a control level. We will take that of type II. A table in this norm tells us that we must use the code **L** for the next step. The final table thus gives us for a NQA of 1% and a L coll the following cell:

1.06%	
$n=25$	$k=1.97$
4.28%	

corresponding to:

D_s	
n	k
D_c	

or more traditionally:

AQL	
n	k
LTPD (QLL)	

Therefore, the norm indicates that a sample size n of 25 should be used and a k of 1.97! The plan insures risks slightly smaller than the plan calculated above but is more expensive because requires bigger samples.

E2. Practitioners when they use norms often want to calculate the customer and supplier risk associated to the selected AQL and the corresponding parameters given by the norm. Then using:

$$k=\frac{Z_\alpha}{\sqrt{n}}-Z_{D,\alpha} \tag{16.7.670}$$

It comes:

$$Z_\alpha=\sqrt{n}(k+Z_{D,\alpha}) \tag{16.7.671}$$

Either as part of our example and always using Microsoft® Excel® 11.8346:

$$Z_\alpha=\texttt{SQRT(25)(1.97+NORMSINV(1.06))}\cong -1.67 \tag{16.7.672}$$

Therefore:

$$\alpha=\texttt{NORMSINV(-1.67)}=4.73\% \tag{16.7.673}$$

to be compared to the 5%. As well as:

$$Z_{1-\beta}=\sqrt{n}(k+Z_{D,\beta}) \tag{16.7.674}$$

Either as part of our example and always using Microsoft® Excel® 11.8346:

$$Z_{1-\beta}=\texttt{SQRT(25)(1.97+NORMSINV(4.28\%))}\cong 1.25 \tag{16.7.675}$$

it comes:

$$1-\beta=\texttt{NORMSINV(1.25)}=89.52\% \tag{16.7.676}$$

and finally:

$$\beta=1-89.52\%=10.48\% \tag{16.7.677}$$

to be compared to the 10%.

16.7.5.2 Simple acceptance sampling plan by attribute

The attributes acceptance plans are simpler than the sampling plans by measurement, but they are often more expensive (in number of samples) than acceptance plans by measurement and are therefore advise only when we can not summarize information on the quality of a unit by measurement or that the cost of analyses my measurement are too high compared to a simple classification of compliant or non-compliant units.

The "simple acceptance plan by attributes", as already mentioned, is an acceptance test procedure the consist to draw a random sample of a given size n in a batch. The sampled units (in a homogeneous way!) are then inspected and classified as compliant or non-compliant (avoid using the term "defective" as it is not same!).

Given X be the number of non-compliant components found in the sample. If $X \leq A$, the quality of the batch is considered good and the batch is accepted. If $X \geq R$ with in the case of simple acceptance plan by attribute: $R = A + 1$ the batch quality is considered as poor and the batch is not accepted.

✍**Example:**

Bottles by batch size of $N = 10,000$ are received by a customer. The control procedure used empirically a simple acceptance plan by attribute: it collects a sample of size $n = 200$. The acceptance criterion A is $X \leq 1$ and the rejection criterion R is $X \geq 2$.

Let us recall before going further into the mathematical details of the calculations of the strategy of a single sampling plan by attributes the principles of double and multiple sampling plans by attributes for the general culture:

- The "double acceptance plan by attributes" is an acceptance test procedure having two stages of decision. It involves first taking a first sample size of a size smaller than the one we would collect for a single acceptance plan. The first stage of decision is based on information provided by the first sample. If the quality of the first sample is considered good enough, the batch is accepted. If judged bad enough, the batch is not accepted. However if the quality of the first sample is judged as intermediate, a second sample is taken and inspected in order to accept or reject the batch. the second stage of the decision criteria are based on information collected on the two samples.

  ✍**Example:**

  Bottles by batch size of $N = 10,000$ are received by a customer. The control procedure used empirically the following acceptance plan by attributes: it takes a first sample size $n_1 = 125$. The acceptance criterion A in the first stage is the $X_1 = 0$ and the rejection criterion R is $X_1 \geq 2$. If the quality is not satisfactory, we take a second sample of size $n_2 = 125$ with an acceptance criterion A in the second stage of $X_2 \leq 1$ and the rejection criterion R is $X_2 \geq 2$.

  We will come back on a much more detailed and formal example of double acceptance plan by attribute further below!!

- The "multiple acceptance sampling plan by attribute" is simply a generalization of the double acceptance sampling plan by attributes having D decision steps instead of 1 for the simple acceptance plan or 2 for the double acceptance plan...

Let us come back to our mathematical developments of a single sampling acceptance plan by attributes. If we denote now n the batch size, p the sample size and m the total number of defective in the batch, then the probability of having k non-compliant items among p is (always using as from the beginning of this textbook a notation for the binomial coefficient that is not-compliant with the norm ISO 31-11):

$$P(n,p,m,k)=\frac{C_k^m C_{p-k}^{n-m}}{C_p^n} \tag{16.7.678}$$

To be consistent with the tradition of the field of acceptance plan, let's first change these notations. Let us denote by n the sample size and N the batch size. It comes then:

$$P(N,n,m,k)=\frac{C_k^m C_{n-k}^{N-m}}{C_n^N} \tag{16.7.679}$$

Let's do a second change of notation. Let us denote by p the proportion of non-compliant items in the batch. It then comes the traditional customary notation:

$$P(N,n,p,k)=\frac{C_k^{Np} C_{n-k}^{N-Np}}{C_n^N} \tag{16.7.680}$$

Thus, the cumulative probability of accepting a batch of quality p is given by:

$$P_a=P(X\leq A)=\sum_{k=0}^{A}P(N,n,p,X=k)=\sum_{k=0}^{A}\frac{C_k^{Np} C_{n-k}^{N-Np}}{C_n^N} \tag{16.7.681}$$

✍**Example:**

What is the cumulative probability of drawing 5 non-compliant items (acceptance criterion) of a batch of $10,000$ bottles (N) from which we took a sample of 200 individuals (n), the proportion of non-compliant is known to be of 5% (p). So with Microsoft® Excel® 14.0.6123 we get as cumulative probability of acceptance:

$$\begin{aligned}P_a=P(X\leq A)&=\sum_{k=0}^{A}\frac{C_k^{Np} C_{n-k}^{N-Np}}{C_n^N}=\sum_{k=0}^{5}\frac{C_k^{10000\cdot 5\%} C_{200-k}^{10000-10000\cdot 5\%}}{C_{200}^{10000}}\\&=\texttt{HYPGEOM.DIST(5,200,10000*5\%,10000,1)}=6.05\%\end{aligned} \tag{16.7.682}$$

Remark

A classic common mistake business that I see from Quality Manager of some of the best luxury international brands based in Switzerland is to believe that the acceptance criteria is proportional to the batch size! In other words that if for $10,000$ bottles, the fact to a test sample of 200 items (2%) gives a cumulative probability of acceptance of $\cong 6\%$, then on $1,000$ bottles, a sample of 20 items (still 2%) gives the same cumulative probability of acceptance. In fact, with Microsoft® Excel® 14.0.6123, we then not have 6.05% anymore but 99.7% (!!!)... the difference is then considerable !!

If we recall (see above the simple acceptance plans by measurement) that if α is the supplier risk (cumulative probability that the customer rejects by error the batch whose proportion of non-compliant items is equal to p_α: type I Error) and β the customer risk (cumulative probability of accepting by error a batch of a supplier whose proportion of non-compliant items is p_β: type II Error), then so that the supplier and the customer have the sampling plan corresponding to each of their quality requirements we have to solve following system:

$$\begin{aligned}1-\alpha &= \sum_{k=0}^{A}\frac{C_k^{Np_\alpha}C_{n-k}^{N-Np_\alpha}}{C_n^N}\\ \beta &= \sum_{k=0}^{A}\frac{C_k^{Np_\beta}C_{n-k}^{N-Np_\beta}}{C_n^N}\end{aligned} \tag{16.7.683}$$

As far as we know it is not possible to solve this system analytically. The techniques of operational research techniques (see section Numerical Methods page 6645) using the simplex method, the gradient method or Newton method fail miserably in finding a solution to this system (which is quite normal in fact...). By cons, with evolutionary algorithms (tool available in Microsoft® Excel® 14.0.6123 or also in MATLAB®) we can find acceptable solutions, but the settings are not easy.

Therefore, assuming that the size of batches submitted for inspection is large relative to the sample size $N \gg n$, and using the binomial approximation of the hypergeometric distribution (approximation proved in the section Statistics at page 672), we have:

$$P_a = P(X\leq A) = \sum_{k=0}^{A}P(n,p,X=k) = \sum_{k=0}^{A}C_k^n p^k(1-p)^{n-k} = \sum_{k=0}^{A}C_k^n p^k q^{n-k} \tag{16.7.684}$$

The advantage of this approximation is huge! The total size N of the batch is no longer involved in the problem!

✍**Examples:**

E1. What is the cumulative probability of drawing 5 non-compliant items (acceptance criterion) of a batch of $10,000$ bottles (N) from which we took a sample of 200 individuals (n), whose proportion of non-compliant is known to be as equal to 5% (p). While Microsoft® Excel® 14.0.6123 we get:

$$\begin{aligned}P_a = P(X\leq A) &= \sum_{k=0}^{A}C_k^n p^k q^{n-k} q^{n-k} = \sum_{k=0}^{5}C_k^{200}0.05^k 0.95^{200-k}\\ &= \texttt{BINOMDIST(5,200,5\%,1)} = 6.23\%\end{aligned} \tag{16.7.685}$$

E2. The latter relation is also used in reliability engineering (see earlier above)!!! Indeed, suppose we wish to know the number of elements n we need to take to make a reliability demonstration plan showing with 90% of cumulative probability that the number A of failures (non-compliance) is equal to zero knowing that their reliability is 80%. Then this is equivalent to search n such that we are the closest to:

$$\begin{aligned}P_a=P(X\leq 0)=\sum_{k=0}^{A}C_k^n p^k q^{n-k}=\sum_{k=0}^{A}C_k^n(1-R)^k R^{n-k}=\sum_{k=0}^{0}C_k^n 0.2^k 0.8^{n-k}\\=R^n=0.8^n=1-90\%=10\%\end{aligned} \tag{16.7.686}$$

Therefore:

$$n=\frac{\log(0.1)}{\log(0.8)}=10.31885\cong 11 \tag{16.7.687}$$

So the nearest value n for this is to take 11 items (this is the value that also returns the worldwide reference reliability software Weibull++). We can then use the same previous calculation:

$$\texttt{BINOMDIST(0,11,20\%,1)}=8.59\%$$

Within the area of reliability, it is customary to name this approach "nonparametric binomial demonstration test plan of the reliability"... (we will come back on this subject further below).

E3. Let us see now a last example relatively to reliability. Remember that we have proved earlier above that for the Weibull distribution with one parameter, we had:

$$R(t)=e^{-\left(\frac{t}{\eta}\right)^{\beta}} \tag{16.7.688}$$

And let suppose that we previous tests have shown that at a time t equal to $2,000$ days, we had a reliability of 80% and we know that the shape parameter β is equal to 2. We would like to know what sample size we need to take for a reliability test of $1,500$ days shows with maximum 90% cumulative probability 1 non-compliance (that is: 1 failure). For this, we first determine the scale parameter:

$$\begin{aligned}R(t)=e^{-\left(\frac{t}{\eta}\right)^{\beta}}\Rightarrow\eta=\frac{t}{(-\ln(R(t)))^{1/\beta}}=\frac{2000}{(-\ln(R(2000)))^{1/2}}\\=\frac{2000}{(-\ln(0.8))^{1/2}}\cong 4233.87\end{aligned} \tag{16.7.689}$$

Then it comes for this same test at $1,500$ day a reliability of:

$$R(1500)=e^{-\left(\frac{1500}{4233.87}\right)^{2}}\cong 0.882=88.2\% \tag{16.7.690}$$

By injecting into the binomial law, we then have:

$$P_a=P(X\leq 1)=\sum_{k=0}^{A}C_k^n p^k q^{n-k}=\sum_{k=0}^{1}C_k^n(1-0.882)^k 0.882^{n-k}=1-90\%=10\% \tag{16.7.691}$$

Solving for n, we get the integer value n of 32 items (the same value as that returned then by the worldwide reference reliability software Weibull++). We are then left with:

= `BINOMDIST(1,32,1-88.2%,1)` = 10.50%

Remark

Again... a common (beginner) mistake that is classic in corporation is to think that the acceptance criterion is proportional to the batch size. In other words that if for $10,000$ bottles, the fact to test a sample of 200 items (2%) gives a cumulative probability of acceptance of $\cong 6\%$, the a on a batch of $1,000$ bottles, a sample of 20 items (2%) gives the same cumulative probability of acceptance... In fact, with Microsoft® Excel® 14.0.6123, we have don't have 6.23% anymore but 99.7%... the difference is huge!!

Returning to the theory .... the system we had then with the Hypergeometric law can be written when $N \gg n$ as:

$$\boxed{\begin{aligned} 1-\alpha &\cong \sum_{k=0}^{A} C_k^n p_\alpha^k (1-p_\alpha)^{n-k} = \sum_{k=0}^{A} C_k^n p_\alpha^k q_\alpha^{n-k} \\ \beta &\cong \sum_{k=0}^{A} C_k^n p_\beta^k (1-p_\beta)^{n-k} = \sum_{k=0}^{A} C_k^n p_\beta^k q_\beta^{n-k} \end{aligned}} \qquad (16.7.692)$$

But once again, it is not possible (as far as we know) to solve the previous system analytically. In 11967 (holocene calendar), Harry R. Larson then created a chart of the cumulative binomial frequently referred to as the "Larson's binomial nomograph" (I never found or seen an equivalent for a hypergeometric law unfortunately...) which allows approximately to solve the problem of determining the value of A and n for the acceptance plan:

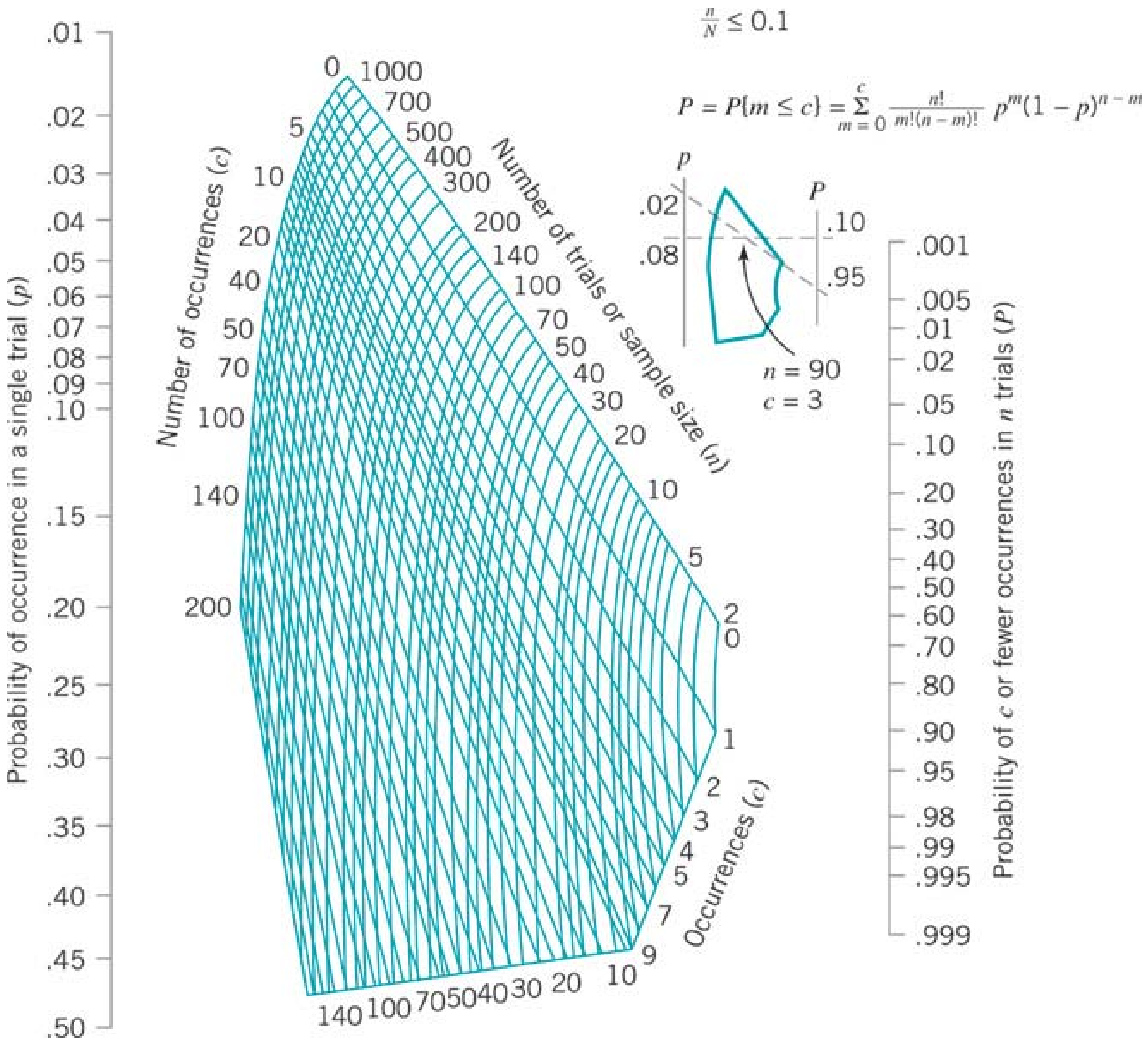


Figure 16.274: **Larson's binomial nomograph (source: Montgomery)**

✍**Example:**

A batch of bottles is delivered under the form of $10,000$ units corresponding to N. We seek to establish an approval control acceptance plan by attributes with the cumulative probability (supplier risk) α of 1% that the customer rejects the batch by error with less than 2.5% (p_α) of non-compliant. At his side the customer wishes a cumulative probability (customer risk) β of 10% of accepting by error a batch with more than 5% (p_β) of non-compliant (in fact the requirements are independent of the size N of the batch).

The goal is therefore to determine the value of A and n using the nomograph. For this, as shown in the nomograph, we must draw two lines:

- Starting points of the two lines: 0.99 (100% -1%) and 0.1 (10%) of the right axis.
- Finishing point of the two lines: 0.025 (2.5%) and 0.05 (5%) of the left axis.

and the intersection of the two lines will give the desired parameters. Thus in this case:

$$A \cong 30 \quad (16.7.693)$$

$$n \cong 700 \quad (16.7.694)$$

We also see that if we put these two values inside:

$$\begin{aligned} 1-\alpha &\cong \sum_{k=0}^{A} C_k^n p_\alpha^k (1-p_\alpha)^{n-k} = \sum_{k=0}^{A} C_k^n p_\alpha^k q_\alpha^{n-k} \\ \beta &\cong \sum_{k=0}^{A} C_k^n p_\beta^k (1-p_\beta)^{n-k} = \sum_{k=0}^{A} C_k^n p_\beta^k q_\beta^{n-k} \end{aligned} \quad (16.7.695)$$

with Microsoft® Excel®, we're actually far from the expected result! So the error of the nomograph can lead to the impression of rough approximations! But change in Microsoft® Excel® the value 30 by 27 and the value 700 by 695 and you'll see what you can conclude...

We also see through this example that an attribute acceptance plan is actually also much more expensive than a acceptance plan by measurement (since we took the same values of α and β as for the example with the same parameters as during our study of the acceptance plans with measurement but we have $n=700$ instead of $n \cong 18$).

The reader may have perhaps noticed that N does not affect the value of A and n when we use the Larson's nomograph. So if we use computer tools to find the solution, we can therefore put that n is between 0 and $1,000$ and A between 0 and 150 as we can see it on the nomograph.

Remark

Finally, let us indicate that some use the Poissonian approximation of the binomial distribution (when the probability (proportion) of non-compliants is very small and that the batch size is very big and sometimes even when these criteria are far from being respected ...) following:

$$P(k)=\frac{\lambda^k e^{-\lambda}}{k!} \Rightarrow P(k \leq c)=\sum_{k=0}^{c} \frac{(np)^k e^{-np}}{k!} \quad (16.7.696)$$

16.7.5.2.1 Calculation of the parameters using the norm ISO 2859-1

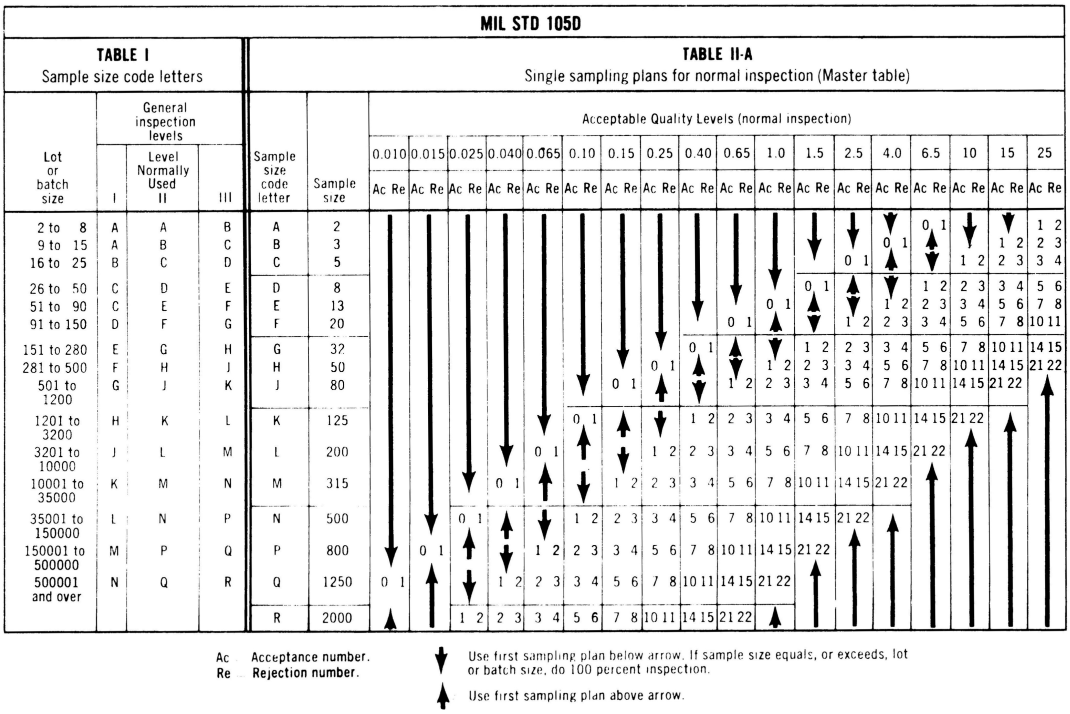
The norm ISO 2859 (which derives from the norm MIL 105E) is dedicated to sampling procedures for inspection by attributes. Along with the norm AFNOR X06-023 it is based on the concept of AQL (percentage of non-compliant items that should not be exceeded for a production, controlled on a series of batches, so that it can be considered as satisfactory).

Following this norm, the NQA must be one of the 26 values recommended by the norm. We will choose for example here the inspection level number II of this norm (they are described in the norm documentation). With respect to the chosen effective, as the AFNOR norm, ISO 2589 provides a symbol (letter) which will then be used in a table. Finally, depending on the type of acceptance plan (simple, double or triple) we find the parameters A and n.

✍**Examples:**

E1. In the context of our companion example - the producer of soft drinks - which receives batches of $10,000$ bottles, the AQL quality level is 1%. As for example with the AFNOR norm, we will take a type II control based on a simple acceptance plan.

A table in the norm (see below) tells us that we must use the code L for the next step. Then for simple control, the standard says to use to use the table IIA.



MIL STD 105D

TABLE I
Sample size code letters

Lot or batch size	General inspection levels I	General inspection levels Level Normally Used II	General inspection levels III
2 to 8	A	A	B
9 to 15	A	B	C
16 to 25	B	C	D
26 to 50	C	D	E
51 to 90	C	E	F
91 to 150	D	F	G
151 to 280	E	G	H
281 to 500	F	H	J
501 to 1200	G	J	K
1201 to 3200	H	K	L
3201 to 10000	J	L	M
10001 to 35000	K	M	N
35001 to 150000	L	N	P
150001 to 500000	M	P	Q
500001 and over	N	Q	R

TABLE II-A
Single sampling plans for normal inspection (Master table)

Acceptable Quality Levels (normal inspection)

Sample size code letter	Sample size	0.010 Ac Re	0.015 Ac Re	0.025 Ac Re	0.040 Ac Re	0.065 Ac Re	0.10 Ac Re	0.15 Ac Re	0.25 Ac Re	0.40 Ac Re	0.65 Ac Re	1.0 Ac Re	1.5 Ac Re	2.5 Ac Re	4.0 Ac Re	6.5 Ac Re	10 Ac Re	15 Ac Re	25 Ac Re
A	2	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	0 1	↓	↓	1 2
B	3	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	0 1	↑	↓	1 2	2 3
C	5	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	0 1	↑	↓	1 2	2 3	3 4
D	8	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	0 1	↑	↓	1 2	2 3	3 4	5 6
E	13	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	0 1	↑	↓	1 2	2 3	3 4	5 6	7 8
F	20	↓	↓	↓	↓	↓	↓	↓	↓	↓	0 1	↑	↓	1 2	2 3	3 4	5 6	7 8	10 11
G	32	↓	↓	↓	↓	↓	↓	↓	↓	0 1	↑	↓	1 2	2 3	3 4	5 6	7 8	10 11	14 15
H	50	↓	↓	↓	↓	↓	↓	↓	0 1	↑	↓	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22
J	80	↓	↓	↓	↓	↓	↓	0 1	↑	↓	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	↑
K	125	↓	↓	↓	↓	↓	0 1	↑	↓	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	↑	↑
L	200	↓	↓	↓	↓	0 1	↑	↓	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	↑	↑	↑
M	315	↓	↓	↓	0 1	↑	↓	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	↑	↑	↑	↑
N	500	↓	↓	0 1	↑	↓	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	↑	↑	↑	↑	↑
P	800	↓	0 1	↑	↓	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	↑	↑	↑	↑	↑	↑
Q	1250	0 1	↑	↓	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	↑	↑	↑	↑	↑	↑	↑
R	2000	↑	↑	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	↑	↑	↑	↑	↑	↑	↑	↑

Ac = Acceptance number.
Re = Rejection number.
↓ Use first sampling plan below arrow. If sample size equals, or exceeds, lot or batch size, do 100 percent inspection.
↑ Use first sampling plan above arrow.

Figure 16.275: **MIL 105E acceptance sampling table**

We can then read that n must be equal to 200 and A is equal to 5. We therefore find that the value obtained is very different from that the one calculated theoretically. Honestly the reason for this difference is unknown to me...

E2. Practitioners in the use of this norm often want to calculate the supplier and customer risk factors associated to the values of n and A provided by the norm in relation to the NQA. selected for a batch of a given quality. So as in this case, $200/10,000$ is smaller than 10%, we can use the binomial approximation of the hypergeometric law. Suppose we therefore wish to calculate the risk for the supplier of the previous example (the one with the use of the nomograph):

$$\begin{aligned}1-\alpha &\cong \sum_{k=0}\sum_{k=0}^{A} C_k^n p_\alpha^k (1-p_\alpha)^{n-k} = \sum_{k=0}^{5} C_k^{200} 0.025_\alpha^k (1-0.025)^{200-k} \\ &= \texttt{BINOMDIST(5,200,2.5\%,1)} = 61.59\%\end{aligned} \tag{16.7.697}$$

hence a supplier risk of $100\%-61.59\%$ that is 38.40% which is considerable! Thus, this sampling plan obtained via the norm is suitable for proportions of non-compliants much lower that 2.5% (in fact less than $\cong 1.25\%$).

For the customer risk, we proceed in the same way:

$$\begin{aligned}\beta &\cong \sum_{k=0}^{A} C_k^n p_\beta^k (1-p_\beta)^{n-k} = \sum_{k=0}^{5} C_k^{200} 0.05_\beta^k (1-0.05)^{200-k} \\ &= \texttt{BINOMDIST(5,200,5\%,1)} = 6.23\%\end{aligned} \tag{16.7.698}$$

which is here a more acceptable result. In fact we find that the norm is more appropriated to the criteria of the customer rather than that of the supplier.

Remark

The software Quick Control Pro (used sometimes by the Swiss Watch industry) of the company Logystem SA manages the acceptance sampling according to the norms ISO 2859 and ISO 3951.

16.7.5.3 Double acceptance sampling plan by attribute

Now let us come back on the double acceptance sampling using the binomial approximation that is for recall:

$$P_a = P(X \leq A) = \sum_{k=0}^{A} P(n,p,X=k) = \sum_{k=0}^{A} C_k^n p^k (1-p)^{n-k} = \sum_{k=0}^{A} C_k^n p^k q^{n-k} \tag{16.7.699}$$

and let us consider a double sampling plan with:

$$\begin{aligned} n_1 &= 50 \\ n_2 &= 100 \end{aligned}$$

and where the proportion of non-compliant is known as being equal to 5%. So, in this case the cumulative probability of having, for example, at least 1 non-compliant item at the first sampling and corresponding to the probability of accepting the batch is equal to:

$$P_a^1 = P(X_1 \leq 1) = \sum_{k_1=0}^{1} \frac{50!}{k_1!(50-k_1)!} 0.05^{k_1} (1-0.05)^{50-k_1} \cong 0.279 \tag{16.7.700}$$

So far nothing special! Now to get the probability of acceptance of the batch on the second sampling, we list all the ways we can get the second sampling.

For this, suppose for example that the batch is accepted if and only if the cumulative number of non-compliant items overall the both sampling is between 2 and 3 units (arbitrary simplified choice for the example).

We then have the following possible ordered combinations of drawing between the two samplings (notice that the sum is always between 0 and 3):

$$\{0,2\},\{0,3\},\{1,1\},\{1.2\},\{3.0\},\{2.1\},\{2.0\}$$

But as already mentioned we accept the batch if and only the first sampling gives a number of non-compliant equal or less than 1, it then remains only the following combinations:

$$\{3,0\},\{2,1\},\{2,0\}$$

to consider in this double acceptance sampling. So we need to calculate the following probabilities:

$$\begin{aligned}P_a&=P_a^1+P_a^2=0.279+[P(X_1=2,X_2\leq 1)+P(X_1=3,X_2=0)]\\&\cong 0.279+[P(X_1=2)P(X_2\leq 1)+P(X_1=3)P(X_2=0)]\\&=0.279+\left[\begin{array}{c}\dfrac{50!}{2!48!}0.05^2 0.95^{48}\sum_{k_2=0}^{1}\dfrac{100!}{k_2!(100-k_2)!}0.05^{k_2}0.95^{100-k_2}+\\\dfrac{50!}{3!47!}0.05^3 0.95^{47}\dfrac{100!}{0!100!}0.05^0 0.95^{100}\end{array}\right]\\&\cong 0.279+[0.261\cdot 0.037+0.220\cdot 0.0059]=0.279+[0.0097+0.001]\\&=0.279+0.0107=0.2897\end{aligned}\tag{16.7.701}$$

16.7.5.4 Operating characteristic curve (OC)

The "operating characteristic curve" (OC) depicts the discriminatory power of an acceptance sampling plan (by measurement or by attribute!). The OC curve plots the probabilities of accepting a lot versus the fraction defective.

If we assume that the lot size N is very large, as compared to the sample size n, so that removing the sample doesn't significantly change the remainder of the lot, no matter how many defects are in the sample. Then the distribution of the number of defectives, d, in a random sample of n items is approximately binomial as we know.

When the OC curve is plotted, the sampling risks are obvious. You should always examine the OC curve before using a sampling plan.

The idea in practice is to compare OC curves to help choose the appropriate sampling plan for various n (sampling size) and A ("acceptance number[34]" often denoted c in the US literature of "criterion").

This curve that is obtained by plotting on a graph the cumulative probability of acceptance of a batch in function of the proportion of non-compliant items (named sometimes "effective quality" axes).

We can build this curve for the Normal approximation, Binomial approximation or also with the exact case of the Hypergeometric law!

✍**Example:**

For a quality control by attribute acceptance plan, we apply the control plan on a sample of $n=700$ items, the batch will be accepted if the number of non-compliant items is less than or equal to 30 ($A \leq 30$) or in percentages ($A \leq 4.28\%$) as determined in our previous example!

---

[34] The maximum number of defects that can find and still allow the lot to be accepted.

We want to plot the cumulative probability to accept the batch for a range of non-compliant proportion between $0-7\%$ by step of 0.1%.

Thus the purpose is explicitly to plot:

$$P_a(p)=P(X\leq A)=\sum_{k=0}^{A} C_k^n p^k qn-k=\sum_{k=0}^{A} C_k^n p^k(1-p)^{n-k} \tag{16.7.702}$$

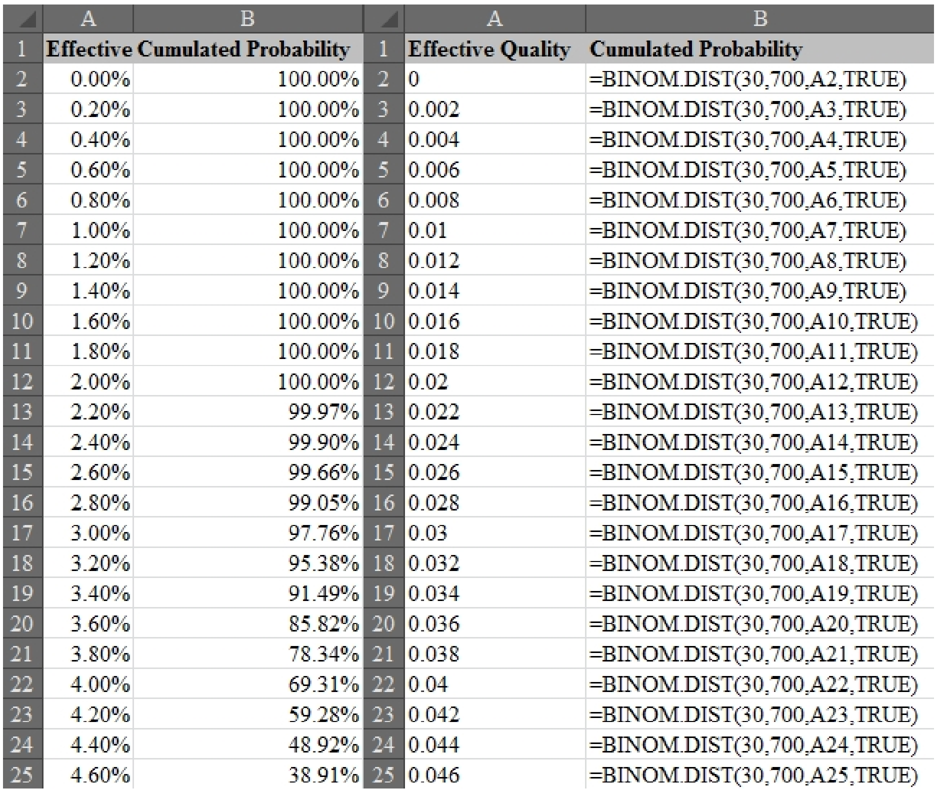
That is often written in the US literature as:

$$P_a(d)=P(X\leq c)=\sum_{k=0}^{c} C_k^n d^k(1-d)^{n-k} \tag{16.7.703}$$

Therefore in our example:

$$P_a(p)=P(X\leq 5)=\sum_{k=0}^{5} C_k^{200} p^k(1-p)^{200-k} \tag{16.7.704}$$

In a spreadsheet software like Microsoft® Excel® 14.0.7166 this gives (first rows only):



	A	B
1	Effective	Cumulated Probability
2	0.00%	100.00%
3	0.20%	100.00%
4	0.40%	100.00%
5	0.60%	100.00%
6	0.80%	100.00%
7	1.00%	100.00%
8	1.20%	100.00%
9	1.40%	100.00%
10	1.60%	100.00%
11	1.80%	100.00%
12	2.00%	100.00%
13	2.20%	99.97%
14	2.40%	99.90%
15	2.60%	99.66%
16	2.80%	99.05%
17	3.00%	97.76%
18	3.20%	95.38%
19	3.40%	91.49%
20	3.60%	85.82%
21	3.80%	78.34%
22	4.00%	69.31%
23	4.20%	59.28%
24	4.40%	48.92%
25	4.60%	38.91%

	A	B
1	Effective Quality	Cumulated Probability
2	0	=BINOM.DIST(30,700,A2,TRUE)
3	0.002	=BINOM.DIST(30,700,A3,TRUE)
4	0.004	=BINOM.DIST(30,700,A4,TRUE)
5	0.006	=BINOM.DIST(30,700,A5,TRUE)
6	0.008	=BINOM.DIST(30,700,A6,TRUE)
7	0.01	=BINOM.DIST(30,700,A7,TRUE)
8	0.012	=BINOM.DIST(30,700,A8,TRUE)
9	0.014	=BINOM.DIST(30,700,A9,TRUE)
10	0.016	=BINOM.DIST(30,700,A10,TRUE)
11	0.018	=BINOM.DIST(30,700,A11,TRUE)
12	0.02	=BINOM.DIST(30,700,A12,TRUE)
13	0.022	=BINOM.DIST(30,700,A13,TRUE)
14	0.024	=BINOM.DIST(30,700,A14,TRUE)
15	0.026	=BINOM.DIST(30,700,A15,TRUE)
16	0.028	=BINOM.DIST(30,700,A16,TRUE)
17	0.03	=BINOM.DIST(30,700,A17,TRUE)
18	0.032	=BINOM.DIST(30,700,A18,TRUE)
19	0.034	=BINOM.DIST(30,700,A19,TRUE)
20	0.036	=BINOM.DIST(30,700,A20,TRUE)
21	0.038	=BINOM.DIST(30,700,A21,TRUE)
22	0.04	=BINOM.DIST(30,700,A22,TRUE)
23	0.042	=BINOM.DIST(30,700,A23,TRUE)
24	0.044	=BINOM.DIST(30,700,A24,TRUE)
25	0.046	=BINOM.DIST(30,700,A25,TRUE)

And if we do a plot adding at the same time the OC curve using the hypergeometric law and the ideal OC curve (sampling plan that discriminates perfectly between good and bad batches) we get:

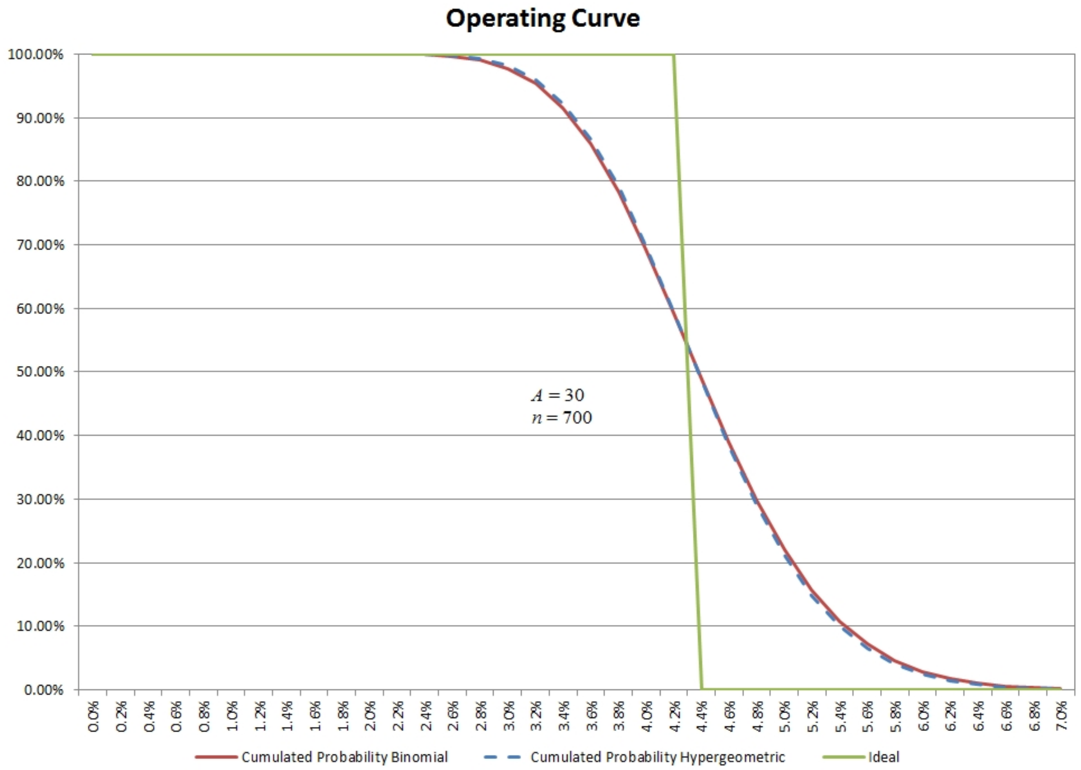


Figure 16.276: Binomial, Hypergeometric and Ideal OC for AQL$=2.5\%$ and $n=200$

Where remember that the values of A and n came from the fact that we seek to establish an approval control acceptance plan by attributes with the cumulative probability (supplier risk) α of 1% that the customer rejects the batch by error with less than 2.5% (p_α) of non-compliant. At his side the customer wishes a cumulative probability (customer risk) β of 10% of accepting by error a batch with more than 5% (p_β) of non-compliant (in fact the requirements are independent of the size N of the batch).

So as we can see, for $A=30$, $n=700$ we have almost 100% probability to accept a batch that has between 0 and almost 2.5% proportion of non-compliant items. After the probability to accept a batch by error decreases as the real proportion of non-compliant items increases (this is quite obvious!).

The reader can see in the Minitab companion book that we get the same binomial OC curve!

An elegant way to represent the concepts of an operating curve without forgetting that in the binomial case:

$$\begin{aligned}1-\alpha &\cong \sum_{k=0}^{A} C_k^n p_\alpha^k (1-p_\alpha)^{n-k}\\ \beta &\cong \sum_{k=0}^{A} C_k^n p_\beta^k (1-p_\beta)^{n-k}\end{aligned} \tag{16.7.705}$$

is as follows:

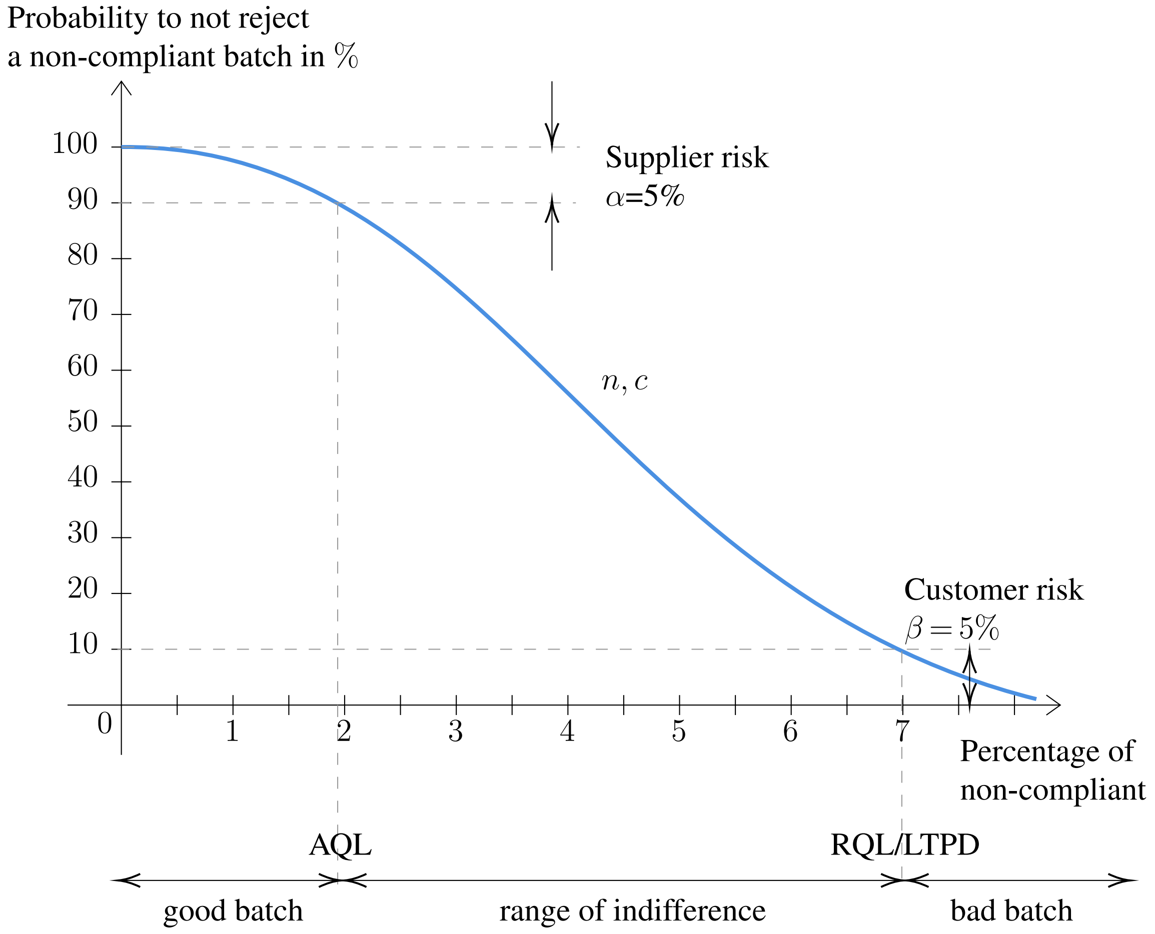


Figure 16.277: **Operating curve summary**

The operating curve shows that the probability of accepting a defective batch decreases - for a constant batch size - when the percentage of defective items in the batch increases: when the quality is good, the probability of accepting the batch by error is high; when the quality is bad, the probability of accepting the batch by error is small.

This type of operating curve based on large sampled batches (dixit using the binomial distribution) is named in the case of using the binomial distribution "operating curve of type B" (OC-B). If obviously the sample size is significant relatively to the size of the batch we will use the hypergeometric and we'll talk then of "operating curve of type A" (OC-A).

16.7.5.5 Average outgoing quality (AOQ)

The batches not accepted by a sampling plan will usually be 100% inspected or screened for nonconforming or defective units. After screening, nonconforming units may be rectified or discarded or replaced by good units, usually taken from accepted lots. Such a program of inspection is known as a "rectifying inspection" or "screening inspection". For those batches accepted by the sampling plan, no screening will be done and the outgoing quality will be the same as that of the incoming quality p.

Since the cumulative probability of accepting by error a bad batch is P_a , the outgoing batches will contain obviously proportion of $P_a \cdot p$ defectives. If only the nonconforming units found in the sample of size n are replaced by good ones, the "average outgoing quality (AOQ)" (its an average proportion!!!) in batches size of N will be given obviously by:

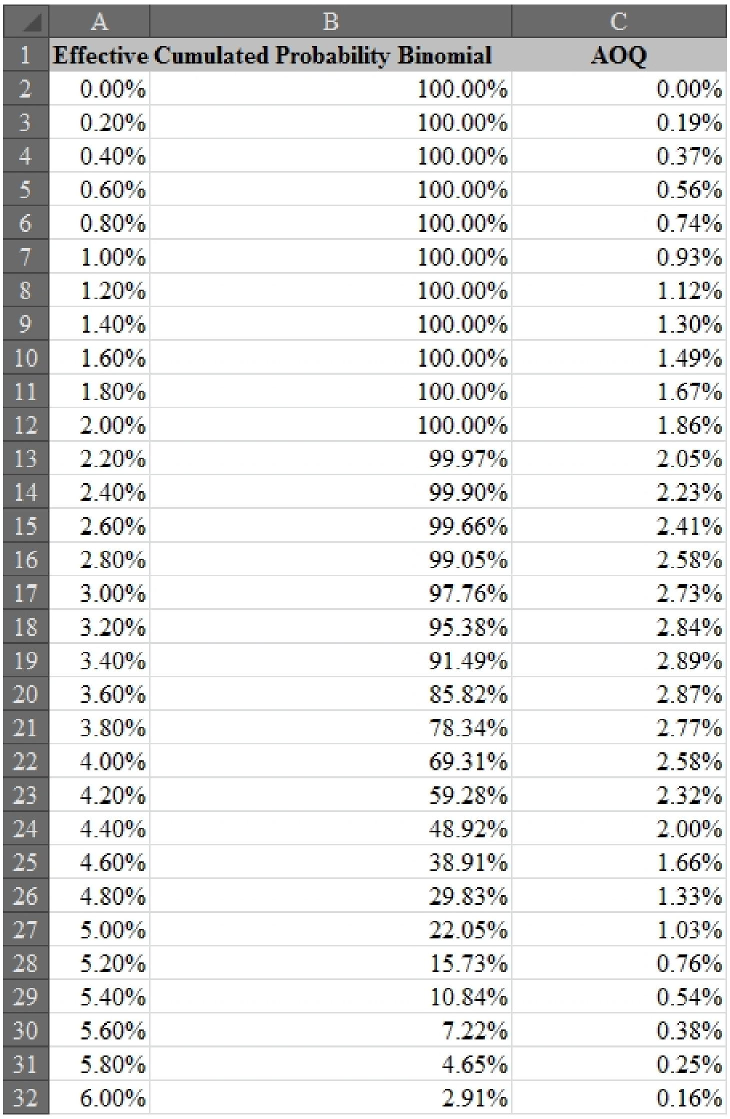
$$\text{AOQ} = \frac{N-n}{N} P_a \cdot p \tag{16.7.706}$$

and for large N we get the approximation:

$$\text{AOQ} \cong P_a \cdot p \tag{16.7.707}$$

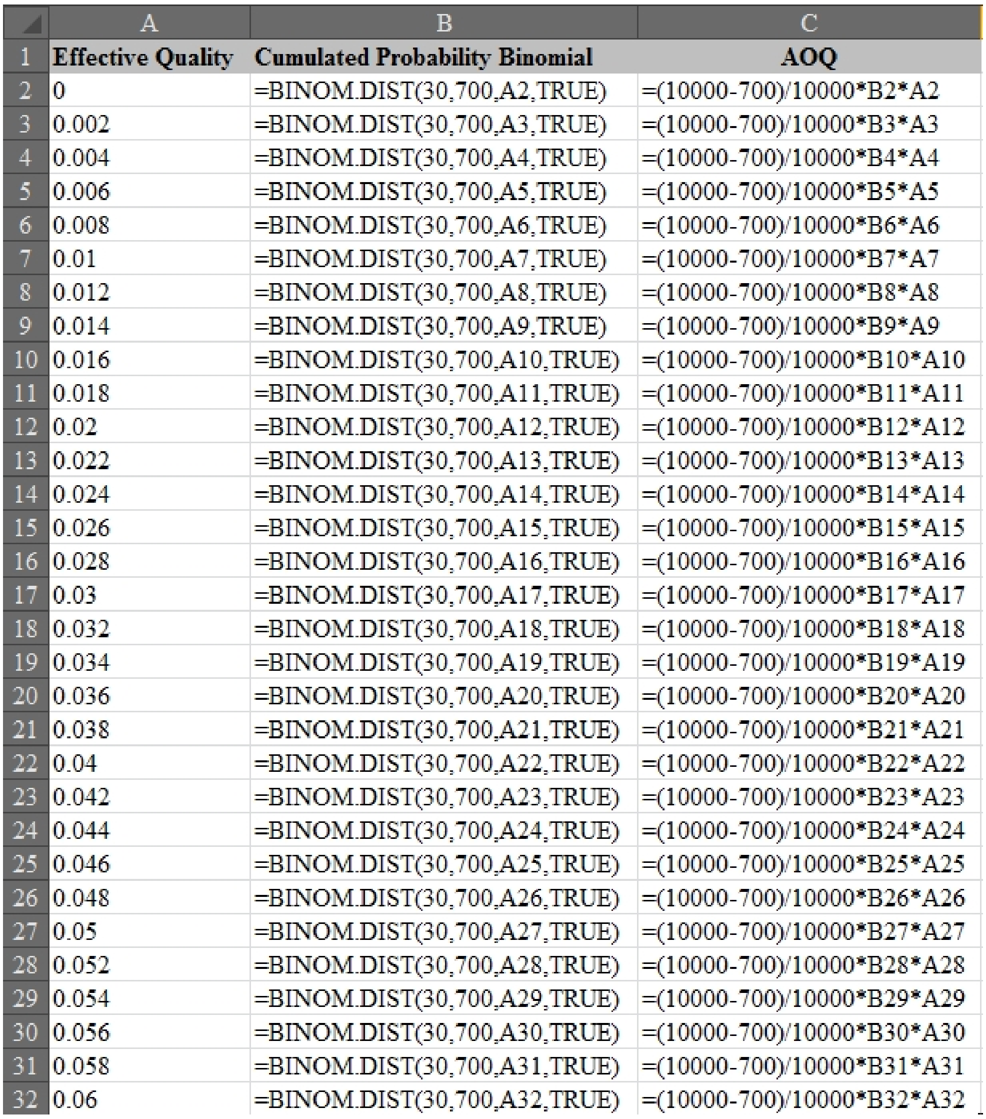
So the average outgoing quality is the expected outgoing quality of a lot (of a given incoming quality) after rectifying inspection has been performed!

Using always our companion example with the bottle batch and using always the same spreadsheet software we have (first rows only):



	A	B	C
1	**Effective**	**Cumulated Probability Binomial**	**AOQ**
2	0.00%	100.00%	0.00%
3	0.20%	100.00%	0.19%
4	0.40%	100.00%	0.37%
5	0.60%	100.00%	0.56%
6	0.80%	100.00%	0.74%
7	1.00%	100.00%	0.93%
8	1.20%	100.00%	1.12%
9	1.40%	100.00%	1.30%
10	1.60%	100.00%	1.49%
11	1.80%	100.00%	1.67%
12	2.00%	100.00%	1.86%
13	2.20%	99.97%	2.05%
14	2.40%	99.90%	2.23%
15	2.60%	99.66%	2.41%
16	2.80%	99.05%	2.58%
17	3.00%	97.76%	2.73%
18	3.20%	95.38%	2.84%
19	3.40%	91.49%	2.89%
20	3.60%	85.82%	2.87%
21	3.80%	78.34%	2.77%
22	4.00%	69.31%	2.58%
23	4.20%	59.28%	2.32%
24	4.40%	48.92%	2.00%
25	4.60%	38.91%	1.66%
26	4.80%	29.83%	1.33%
27	5.00%	22.05%	1.03%
28	5.20%	15.73%	0.76%
29	5.40%	10.84%	0.54%
30	5.60%	7.22%	0.38%
31	5.80%	4.65%	0.25%
32	6.00%	2.91%	0.16%

and explicitly:



	A	B	C
1	**Effective Quality**	**Cumulated Probability Binomial**	**AOQ**
2	0	=BINOM.DIST(30,700,A2,TRUE)	=(10000-700)/10000*B2*A2
3	0.002	=BINOM.DIST(30,700,A3,TRUE)	=(10000-700)/10000*B3*A3
4	0.004	=BINOM.DIST(30,700,A4,TRUE)	=(10000-700)/10000*B4*A4
5	0.006	=BINOM.DIST(30,700,A5,TRUE)	=(10000-700)/10000*B5*A5
6	0.008	=BINOM.DIST(30,700,A6,TRUE)	=(10000-700)/10000*B6*A6
7	0.01	=BINOM.DIST(30,700,A7,TRUE)	=(10000-700)/10000*B7*A7
8	0.012	=BINOM.DIST(30,700,A8,TRUE)	=(10000-700)/10000*B8*A8
9	0.014	=BINOM.DIST(30,700,A9,TRUE)	=(10000-700)/10000*B9*A9
10	0.016	=BINOM.DIST(30,700,A10,TRUE)	=(10000-700)/10000*B10*A10
11	0.018	=BINOM.DIST(30,700,A11,TRUE)	=(10000-700)/10000*B11*A11
12	0.02	=BINOM.DIST(30,700,A12,TRUE)	=(10000-700)/10000*B12*A12
13	0.022	=BINOM.DIST(30,700,A13,TRUE)	=(10000-700)/10000*B13*A13
14	0.024	=BINOM.DIST(30,700,A14,TRUE)	=(10000-700)/10000*B14*A14
15	0.026	=BINOM.DIST(30,700,A15,TRUE)	=(10000-700)/10000*B15*A15
16	0.028	=BINOM.DIST(30,700,A16,TRUE)	=(10000-700)/10000*B16*A16
17	0.03	=BINOM.DIST(30,700,A17,TRUE)	=(10000-700)/10000*B17*A17
18	0.032	=BINOM.DIST(30,700,A18,TRUE)	=(10000-700)/10000*B18*A18
19	0.034	=BINOM.DIST(30,700,A19,TRUE)	=(10000-700)/10000*B19*A19
20	0.036	=BINOM.DIST(30,700,A20,TRUE)	=(10000-700)/10000*B20*A20
21	0.038	=BINOM.DIST(30,700,A21,TRUE)	=(10000-700)/10000*B21*A21
22	0.04	=BINOM.DIST(30,700,A22,TRUE)	=(10000-700)/10000*B22*A22
23	0.042	=BINOM.DIST(30,700,A23,TRUE)	=(10000-700)/10000*B23*A23
24	0.044	=BINOM.DIST(30,700,A24,TRUE)	=(10000-700)/10000*B24*A24
25	0.046	=BINOM.DIST(30,700,A25,TRUE)	=(10000-700)/10000*B25*A25
26	0.048	=BINOM.DIST(30,700,A26,TRUE)	=(10000-700)/10000*B26*A26
27	0.05	=BINOM.DIST(30,700,A27,TRUE)	=(10000-700)/10000*B27*A27
28	0.052	=BINOM.DIST(30,700,A28,TRUE)	=(10000-700)/10000*B28*A28
29	0.054	=BINOM.DIST(30,700,A29,TRUE)	=(10000-700)/10000*B29*A29
30	0.056	=BINOM.DIST(30,700,A30,TRUE)	=(10000-700)/10000*B30*A30
31	0.058	=BINOM.DIST(30,700,A31,TRUE)	=(10000-700)/10000*B31*A31
32	0.06	=BINOM.DIST(30,700,A32,TRUE)	=(10000-700)/10000*B32*A32

and the corresponding chart made by selecting the both last columns:

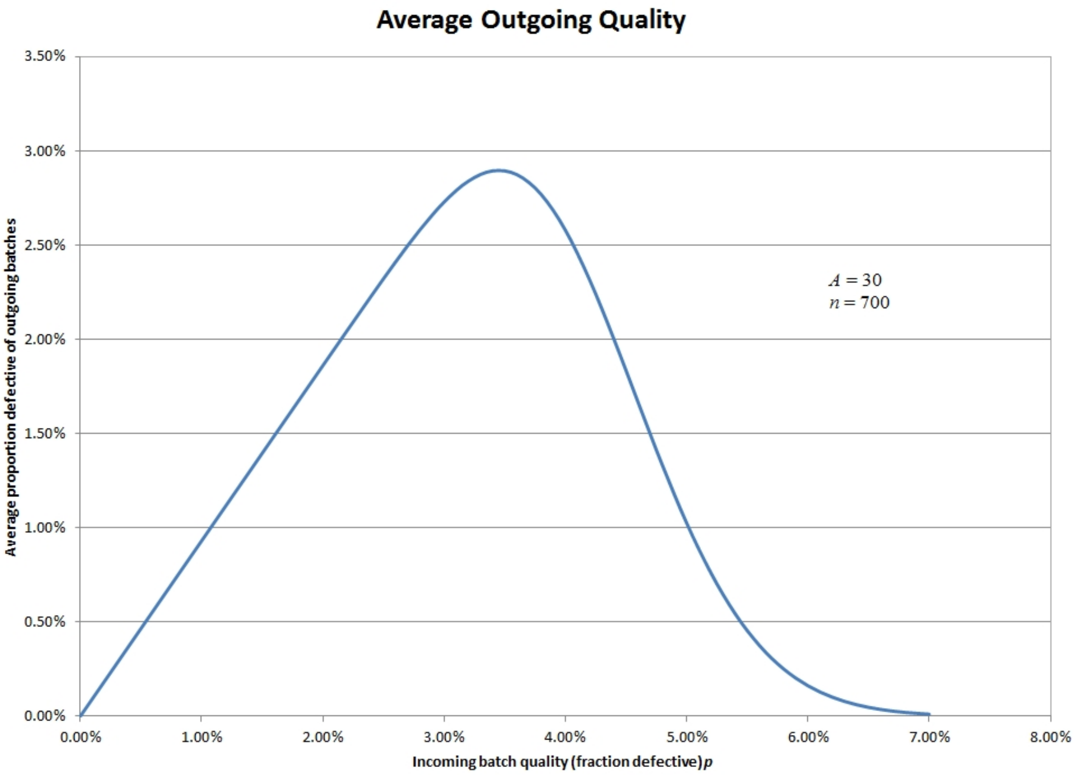
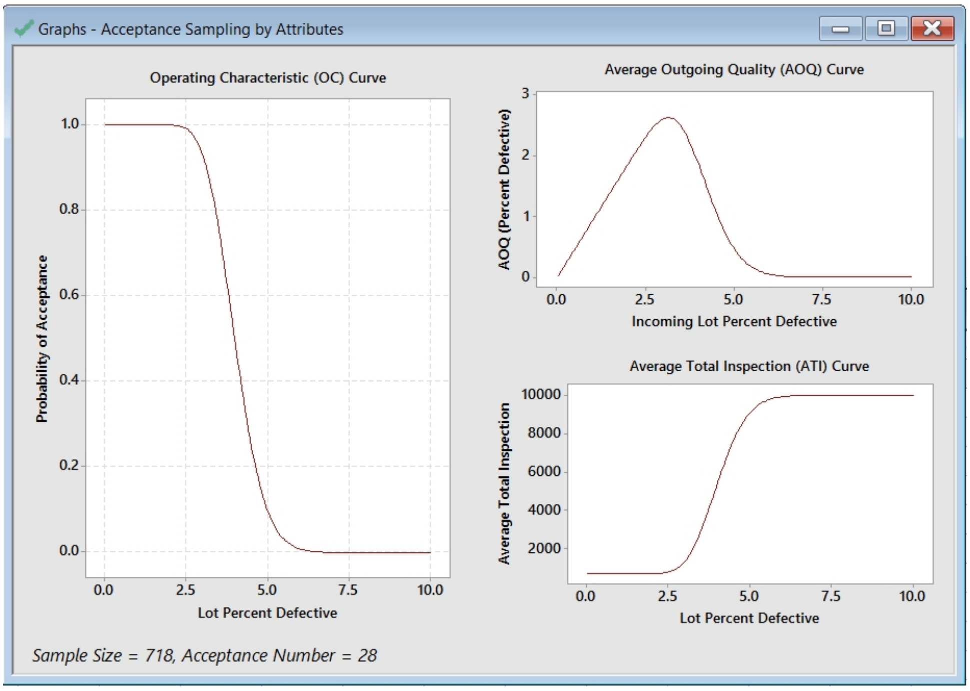


Figure 16.278: **AOQ (Average Outgoing Quality) binomial chart** $A = 30$, $n = 700$

The maximum ordinate of the AOQ curve is known as the "average outgoing quality limit (AOQL)".

So to close this subject we can see below (in the bottom left corner) that Minitab 17.3.1 gives us a result that is not far from our estimation of the Larson's nomograph (yes remember that we get earlier above $A \cong 30$ and $n \cong 700$) and that we also get the same operating curve and average outgoing quality plot:


Graphs - Acceptance Sampling by Attributes
Operating Characteristic (OC) Curve
1.0
0.8
0.6
0.4
0.2
0.0
Probability of Acceptance
0.0
2.5
5.0
7.5
10.0
Lot Percent Defective
Average Outgoing Quality (AOQ) Curve
3
2
1
0
AOQ (Percent Defective)
0.0
2.5
5.0
7.5
10.0
Incoming Lot Percent Defective
Average Total Inspection (ATI) Curve
10000
8000
6000
4000
2000
Average Total Inspection
0.0
2.5
5.0
7.5
10.0
Lot Percent Defective

Sample Size = 718, Acceptance Number = 28

16.7.6 Quality Control Charts (CC)

A "control chart" (sometimes also referred to as "sequence chart" or "Shewhart chart" or "process-behaviour chart") is a planar empirical representation of the variation of an accurate measurement of a process or a method where the vertical axis represents the quantitative indicator selected and the horizontal axis represents the time or corresponding number of outgoing units (so it is a subfamily of "run charts").

This non-destructive method of quality control that is say to be "on-line" (as the acquisition can be done in real time) permits to highlight special causes of variation and then to optimize the frequency of maintenance operations or recalibration) in the purpose to determine if a manufacturing or business process is in a state of statistical control (or is moving into such a state) and to separate signal from noise!

This technique seems to have been created by the physicist engineer Walter A. Shewhart while working for Bell Labs in the 11920s (holocene calendar) and is since some decades (11991 to be exact) an ISO standard under the name *ISO 8258:1991 - Shewhart control charts*. Moreover, it is also Shewhart who inspired the well known W. Edwards Deming (statistician) in the field of project management and quality (all the masters in the field of advanced project management have a scientific background...).

In the early use of control charts, it is advisable to put control charts on all important measurable product characteristics. The resulting chart generally show quickly what charts are necessary or unnecessary (in reality the quality manager has also to conduct a CoQ - Cost of Quality - analysis). Unnecessary or inappropriate control charts will be removed and other control charts might be added.

The most basic and intuitive planar control chart could looks like this:

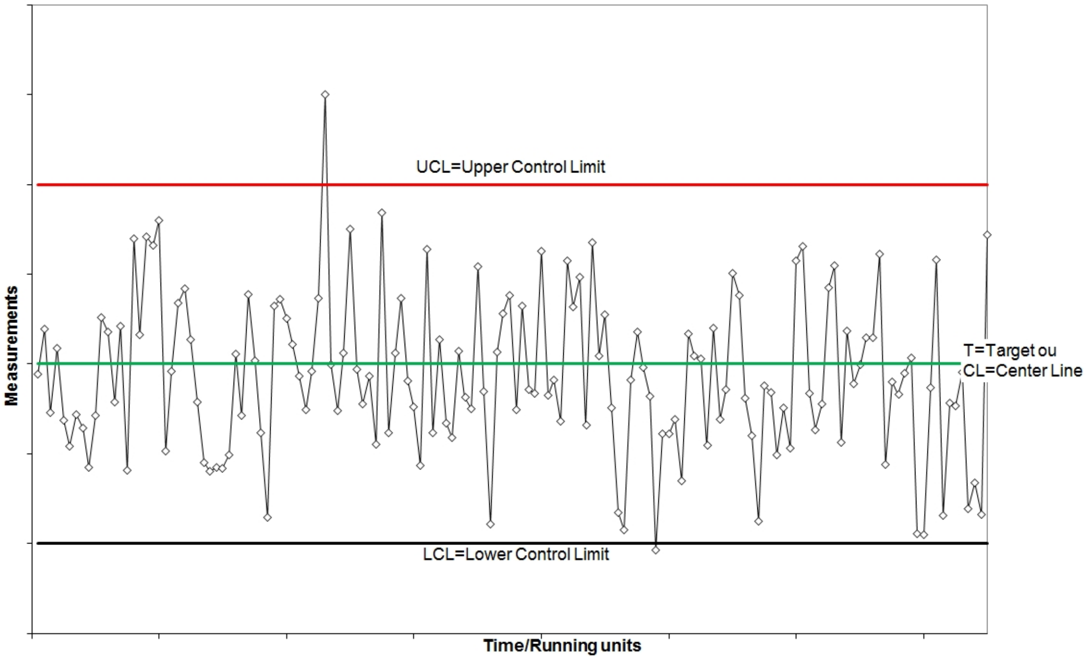
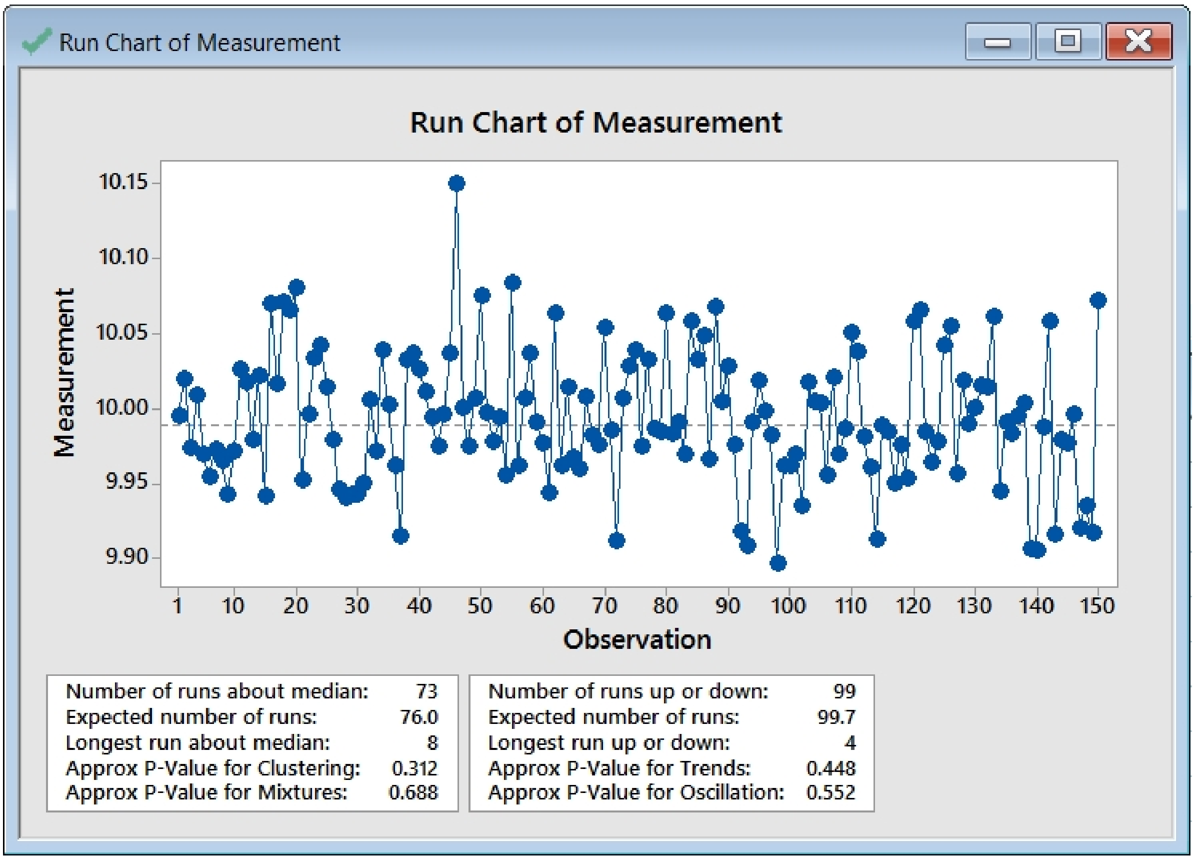


Figure 16.279: **Naive control chart**

The upper and lower limits can be requested / specified by a customer and therefore we note them, as we already know, USL (Upper Specification Limit) and LSL (Lower Specification Limit), otherwise they are calculated internally and therefore we denote them by "UCL (Upper Control Limit)" and "LCL (Lower Control Limit)". To improve the effectiveness of a control chart, we sometimes add on chart narrower limits named "monitoring limits" (in reality we almost always represent the limits calculated AND requested by the quality control policy on the control chart).

The "Center Line (CL)" is most of time chosen as the arithmetic average of all measurement but it should be better to choose the Median as we already know. The "Target (T)" is the value requested as an ideal by the customer specifications.

The equivalent run chart of the above chart (created with Microsoft® Excel®...) but now made with a professional software like Minitab 17.3.1 gives:



```
Number of runs about median:      73
Expected number of runs:        76.0
Longest run about median:          8
Approx P-Value for Clustering: 0.312
Approx P-Value for Mixtures:   0.688
```

```
Number of runs up or down:         99
Expected number of runs:         99.7
Longest run up or down:             4
Approx P-Value for Trends:      0.448
Approx P-Value for Oscillation: 0.552
```

Figure 16.280: **Run Chart with Minitab 17.1.3**

In practice I highly recommend to represent the center line (CL) and also UCL (Upper Control Limit), LCL (Lower Control Limit) as in reality there is normally always limits and goals that are imposed contractually by the customer but also to represent the target T and USL (Upper Specification Limit), LSL (Lower Specification Limit) on the same control chart when possible!!!

The control chart is therefore a measurement tool for viewing and sometimes anticipating the changes and thus determine when an assignable cause occurs causing a drift of a business process or manufacturing processor a variation of a financial value on stock markets requiring rapid corrective action to reduce the costs of non-quality or losses. Thus, ideally the process will be stopped or corrected at the right time, that is to say before it produces too many non-compliant deliverables / services (outside the tolerance) and therefore to avoid overcontrol (useless re-calibrating because of common variability or application of acceptance plan).

The type of control chart corresponds to the type of characteristic that is the subject of a control. However the economic aspect may be an important factor in the choice of the control chart to implement (time for measurement, destructive or non-destructive procedure and son...). If we look at whether a characteristic complies or not with certain norms, an "attribute control chart" (see further below for the details) will be used, but requires a big sample size. Control on quantitative and almost continuous characteristics requires the use of "measurement control charts" and if the measurement are not independent of "time series control charts". This type of measurement control is more efficient and more precise but also more expensive since it requires the use of measuring instruments that should be checked or re-calibrated regularly. We have also for two dimensional correlated controls the "multivariate control charts".

Another very famous type of control chart is that applied to the residual of regression models and named therefore "residuals control charts" that work for Gaussian, Poisson, or Binomial or even Negative Binomial regression models. Many statistical softwares by defaults when running a regression analysis plot such a residual control chart to assess visually the homoscedasticity and normality (or other) of the residuals!

Remark

The subject of control charts is immensely vast and there is an abundant English literature on the subject. We wish to give here only the fundamental (and therefore only basic control charts!) to show a practical application of the theory or Range Statistics and of Statistics of extreme values outliers developed that we have developed in detail in the section Statistics (see page 1203). To place an order of magnitude, most specialized software offers between 10 and 30 different control charts (Minitab, R, JMP, QI Macros, MATLAB®)!

Before going into the mathematical details notice that the control charts among the most used one are the control charts by measurement of the average $\bar{X}$ in combination with the control charts by range R with Shewhart's limits (when we speak of "Shewhart's limits" we are implicitly referring to the fact that the upper control limit UCL and the lower limit control LCL are empirically calculated at $\pm 3\sigma$ of the arithmetic average of the distribution that the latter is symmetrical or not!). These both control maps are used and interpreted most of time together because both parameters are independent and complementary and simple to understand (even if the majority of employees in companies does not know what is a standard deviation). Indeed, the arithmetic average value can vary without that the range varies and vice versa[35].

Remarks

R1. In the case of a control charts with Shewhart's limits, monitoring limits are traditionally placed at $\pm 2\sigma$ (at the opposite of the control limits that are placed at $\pm 3\sigma$).

R2. In the case of a Normal distribution $\mathcal{N}(\mu,\sigma)$, $\pm\sigma$ corresponds to a 99.73% confidence interval as we have seen earlier in this section and also in the section Statistics. This therefore corresponds to a α of about:

---

[35] Remember the trap of the Anscombe's quartet that we saw in the section Statistics

$$\alpha \cong 100\% - 99.75\% = 0.27\% \tag{16.7.708}$$

Most of the industry therefore takes $\pm 3\sigma$ as limits, but strictly speaking distributions of certain control charts are not symmetrical. We should then take terminals (limits) corresponding to a cumulative probability of $\alpha/2$ which can be calculated quite easily with many distributions (but not all!) just by having available an inexpensive spreadsheet software. The choice of the value α depends on the company's quality policy and has the advantage of being a much more accurate indicator that the Shewhart's limits (the latter being wrong when the distributions are skewed).

R3. Shewhart seems to have proposed that there must be at least 25 samples of 4 individuals for the validity of certain control charts (that we will see further below) to be acceptable.

R4. As we will see it, almost all measurement control charts assume Normal distribution and that the observations are independent (between sub-groups and within subgroups!!). For example (example that will be detailed mathematically further below), for already stable processes as since started a long time ago, we perform several individual observations on several subgroups numbered at a given time frequency (hourly, three times per day ...). On each ordered subgroup k (chronological incrementation), we perform n observations. We report on the control chart the arithmetic average $\bar{X}$ of subgroup according to its chronological number k to be reported on the horizontal axis of the control chart. Because of the central limit theorem (CLT) proved in the section Statistics (see 938), the average values $\bar{X}_k$ of the control chart follows a Normal distribution $\mathcal{N}(\mu, \sigma/\sqrt{n})$ (so symmetrical distribution!) that the observations of the subgroups are Normally distributed or not but at the condition there are independent and identically distributed over time!!! This assumption is valid even for small sample size and a manufacturing process under control, which is common in quality control. Production is said to be "stable" as we know if the tendency and dispersion are statistically constant over time (and considered identically distributed independent variables).

16.7.6.1 WECO's empirical rules

The interpretation of a control chart is not always easy and requires experience and expertise to know if corrective action are required or not. The "Western Electric Rules" were codified by a specially-appointed committee of the manufacturing division of the Western Electric Company and appeared in the first edition of its *Statistical Quality Control Handbook* in 11956 (holocene calendar). Their purpose was to ensure that line workers and engineers interpret control charts in a uniform way (the French health ministry did the same 50 years later ... but with rules that differ a little).

The WECO based at this time on the assumption that the statistical distributions are always symmetrical (same assumption as Shewhart) and has then adopted in the statement of control rules limits that are integer multiples k of the standard deviation. Of course, nothing prevents the quality specialist to adapt these rules with probabilistic limits corresponding to a given confidence interval (the majority of specialized software gives the possibility to select the WECO rules to be applied automatically for the detection of non-conformities).

In the measurement control charts type, among these rules here are those that are applicable in order to identify whether a process has failed and that have been completed over the years by other specialists (we have represent below only some rules into a single control chart for obvious pedagogical reasons):

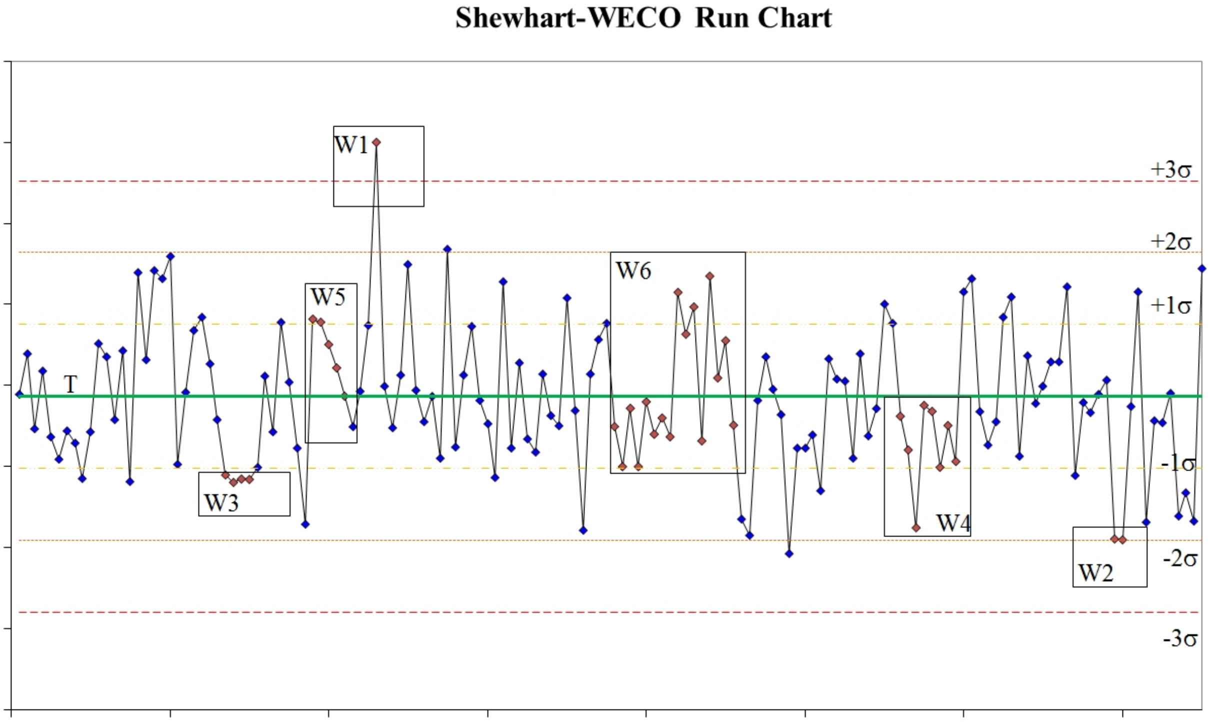


Figure 16.281: **Some WECO rules on a Run Chart**

On the next page you can find some of the empirical "Shewhart-WECO" rules:

W1. A measured point is beyond the USL, LSL specified by the customer or beyond the empirical or probabilistic UCL, LCL corresponding to $\pm 3\sigma$.

W2. Two consecutive measured points are beyond the empirical or probabilistic UCL, LCL corresponding to $\pm 2\sigma$.

W3. Four consecutive measured points are beyond the empirical or probabilistic UCL, LCL corresponding to $\pm 2\sigma$ (in this case named "alert limits").

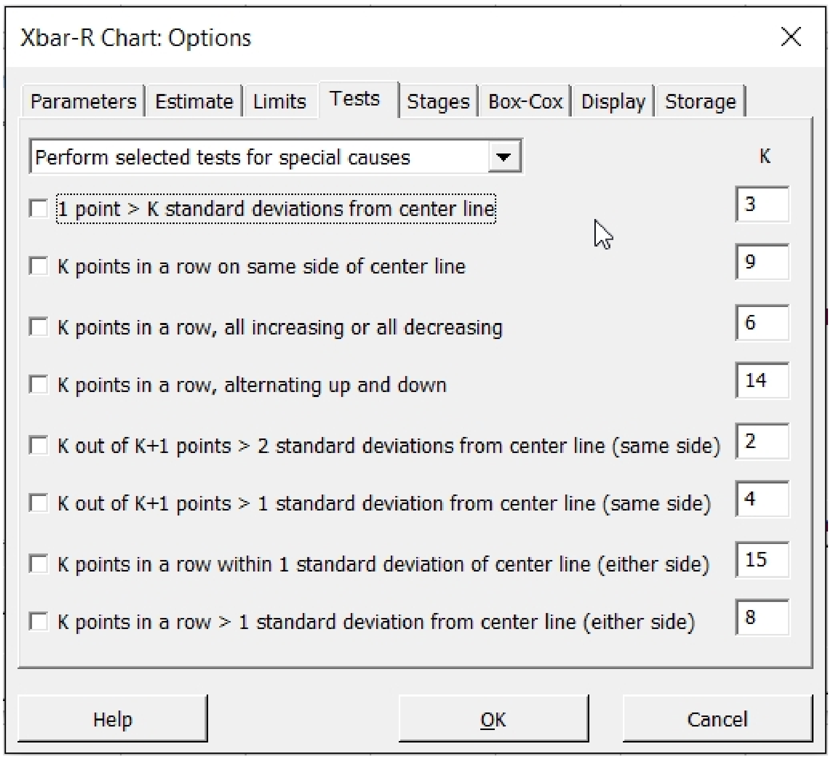
W4. Eight consecutive points fall on the same side of the mean (even if once again the use of the median is more accurate and robust).

W5. Six consecutive points are on an uptrend or downward trend respectively.

W6. Fourteen consecutive points on a systematic up/down alternation trend.

The rules presented above apply to control charts with symmetric control limits. The WECO handbook provides additional guidelines for control charts where the control limits are not symmetrical, as for R charts and p-charts...

As the reader can see below, Minitab includes at least 8 empirical rules (named the "Nelson rules" and first published in the October 11984 (holocene calendar) issue of the Journal of Quality Technology in an article by Lloyd S. Nelson) and that can be furthermore customized:



Perform selected tests for special causes

	K
☐ 1 point > K standard deviations from center line	3
☐ K points in a row on same side of center line	9
☐ K points in a row, all increasing or all decreasing	6
☐ K points in a row, alternating up and down	14
☐ K out of K+1 points > 2 standard deviations from center line (same side)	2
☐ K out of K+1 points > 1 standard deviation from center line (same side)	4
☐ K points in a row within 1 standard deviation of center line (either side)	15
☐ K points in a row > 1 standard deviation from center line (either side)	8

Figure 16.282: **Example of WECO rules implemented in Minitab 17.3.1**

Remark

The reader can also refer to "Westgard Rules" that are a set of rules used for laboratory quality control. They are a copyrighted set of modified Western Electric rules, developed by James Westgard and provided in his books and seminars on quality control.

If the process/method is out of control, corrective actions must be implemented (process review, training, recalibration, code update, etc.). The causes of variation may be obviously random or deterministic. If the causes are only due to chance, then they are named "random causes" (Shewhart) or "common causes" (Deming) and are therefore treated as instantaneous dispersion. Not all changes are due to chance, of course, some are specific and identifiable in a deterministic a certain way. In the latter case, the variations are named "assignable variations" (Shewhart) or "special causes" (Deming) and are assimilated to the overall dispersion (see in the section Statistics the subsection about the propagation of errors at page 1307).

Correct the special causes is obviously most of time much easier than correcting random causes. The primary purpose of control charts is obviously identifying special causes...

OK that's all regarding the customs and traditions of interpretation of measurement points. Now this non-mathematical prelude done, let us see the different types of control charts with a concrete example for each.

16.7.6.2 Sample size and Sampling frequency for Control Charts

The use of control charts pose early two major difficulties:

1. Choosing the family and type of control chart as we have (6 main families, with 25 control charts and including the typical variants to calculate a same control chart, we have a total of almost 60 control charts in reality...!):
    - Variable charts for individuals (continuous data):
        - Levey-Jennings (Pearson standard deviation)
        - Individuals (Pearson standard deviation)
        - I-MR for long runs (Hartley standard deviation)
        - Z-MR for short runs (Hartley standard deviation)
        - Moving Range (Hartley standard deviation)
    - Variable charts for small subgroups (continuous data):
        - Xbar-R (Hartley standard deviation)
        - I-MR-R (Hartley standard deviation)
        - R (Hartley standard deviation)
        - Median-R (Hartley standard deviation)
    - Variable charts for subgroups (continuous data):
        - Xbar-S (Burr standard deviation)
        - I-MR-S (Burr standard deviation)
        - S (Burr standard deviation)

  - Median-S (Burr standard deviation)
- Attribute charts (discrete data):
  - P (binomial proportion)
  - Laney P'/U'
  - NP (binomial np law)
  - C (Poisson counting)
  - U (normalized Poisson)
- Time-weighted charts:
  - Moving Average (prior, central or averaged abscissa)
  - EWMA
  - CUSUM
- Multivariate charts:
  - T-squared
  - Multivariate EWMA
- Rare events charts
  - T (Poisson)
  - G (Negative-Binomial distribution)
- Changepoint Analysis
  - AMOC (At Most One Change) Method
  - PELT (Pruned Exact Linear Time) Method
  - BCP (Bayesian change point) analysis
  - Segmentation Method
  - Tree Method
  - Join-pin regression

2. The choice of the sample size n and of the sampling frequency f_s.

We would here discuss the second point! We must first observe that it is obviously the pace of execution and the quality of operation and also the time it takes to measure the feature (characteristic) that will give us the answer. This quality is appreciated by the average number of corrective interventions observed during a specific period in the past (MTBF typically). The more corrective action there are, the less the quality of the operation of the method / process is good and the higher will be the sampling frequency!

The choice of the sample size is not a problem because as it will result from the statistical calculation of the confidence interval of the statistical indicator that we impose ourselves (see the study of Acceptance Plan earlier above) and therefore of the used distribution law and its properties and also of the assumptions of the chosen the control chart (see below in the text that follows the mathematical description of main control charts).

The challenge is there to determine the sampling frequency. There are for this four main methods and the two first are quite empirical (rule of thumbs):

- Choose a frequency such that the corrective actions T_c are at least four times lower than the sampling frequency (this criterion is independent of the sample size and therefore sometimes not very applicable...):

$$f_s = \frac{T_c}{4} \tag{16.7.709}$$

- We consider T_p as the life of the method / process between each change / modification. We denote as usual n the sample size respecting acceptance plan rules (see previously). We will denote N the number of executions of the method / process (number of produced items) during the period T. So we have a maximum N/n samples that are possible.:

$$f_s = \frac{T_p}{N/n} \tag{16.7.710}$$

- Another method consist to use the MTBF (Mean Time Between Failure) and that as we have proved during our study of reliability techniques is given for recall in the case of the Weibull distribution with three parameters by:

$$f_s = \text{MTBF} = \gamma + \eta\Gamma_{-1}\left(\frac{1+\beta}{\beta}\right) \tag{16.7.711}$$

  or in the case of an exponential law:

$$f_s = \text{MTBF} = \int\limits_0^{+\infty} t\lambda e^{-\lambda t}\mathrm{d}t = \frac{1}{\lambda} \tag{16.7.712}$$

- The last method consist to use a concept we will prove further below and that give a lower bound for the sampling frequency and that is the "average run length" given by:

$$\text{ARL}_0 = \frac{1}{\alpha} \tag{16.7.713}$$

So at a first glance it's not easy to choose a control chart if we are not trained and used to them. So see for example here how the Minitab wizard helps the users to decide what control chart to choose (as you can see the wizard hides the real complexity as most charts are not visible... even if they are available in Minitab menus!):

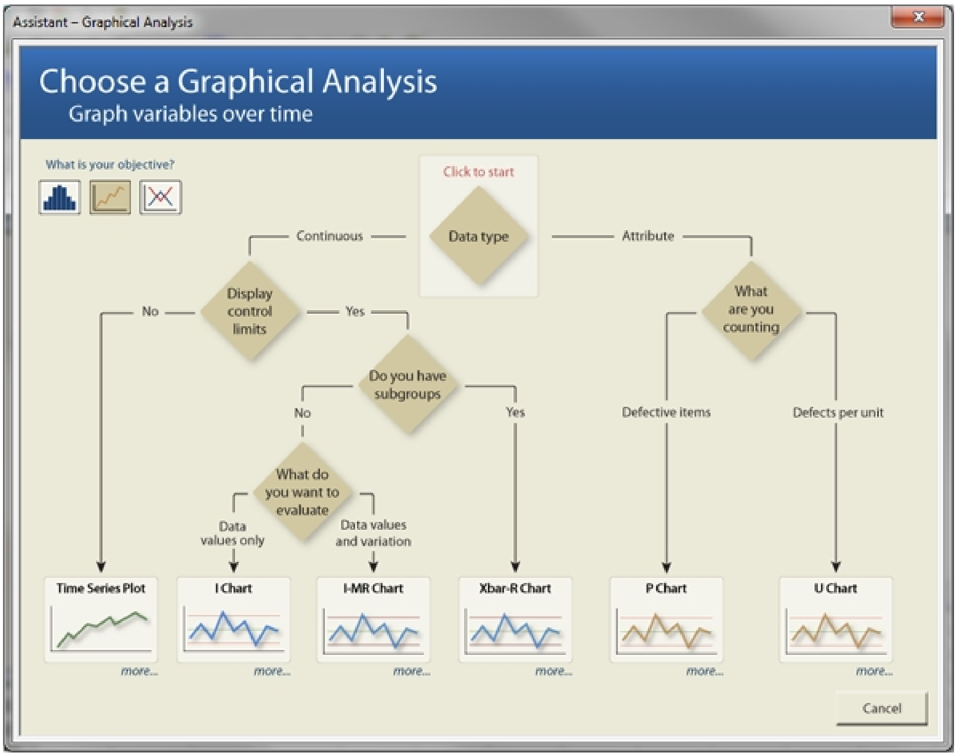

Assistant – Graphical Analysis
Choose a Graphical Analysis
Graph variables over time
What is your objective?
Click to start
Data type
Continuous
Attribute
Display control limits
No
Yes
Do you have subgroups
No
Yes
What do you want to evaluate
Data values only
Data values and variation
What are you counting
Defective items
Defects per unit
Time Series Plot
I Chart
I-MR Chart
Xbar-R Chart
P Chart
U Chart
more...
more...
more...
more...
more...
more...
Cancel

Figure 16.283: **Minitab control chart wizard**

We can find some training centers that provide more complete decision charts but that are sadly a bit misleading (as we will prove it later with the maths). Typically the decision chart below is more complete than the one of Minitab (but simplicity is at the price of exactitude...):

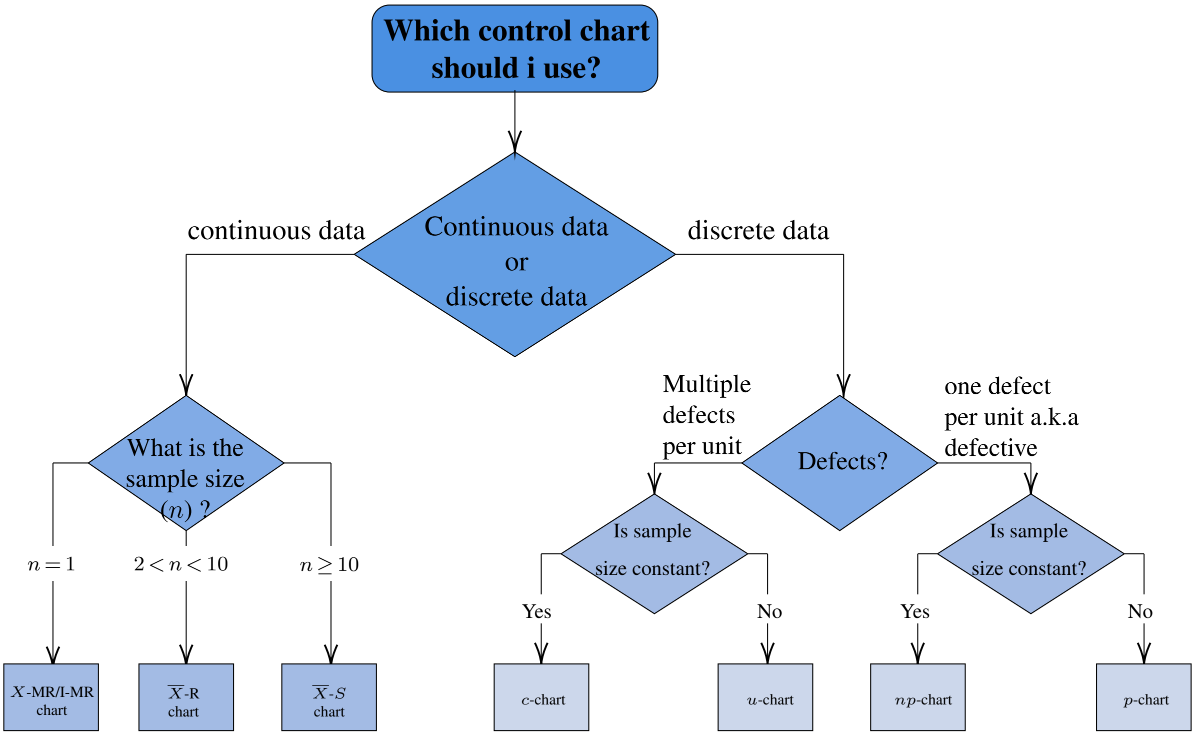

Which control chart should i use?
Continuous data or discrete data
continuous data
discrete data
What is the sample size (n) ?
$n = 1$
$2 < n < 10$
$n \geq 10$
X-MR/I-MR chart
$\overline{X}$-R chart
$\overline{X}$-S chart
Defects?
Multiple defects per unit
one defect per unit a.k.a defective
Is sample size constant?
Yes
No
Is sample size constant?
Yes
No
c-chart
u-chart
np-chart
p-chart

Figure 16.284: **iSixSigma control chart wizard**

In all control charts that we will see below, we have presented examples with constant sample size (hence the limits of control charts are always constant with some exceptions that we will also see). Indeed, if the size of the samples change it is that the production rate (or batch size of delivery) has changed or also the overall quality and therefore the manufacturing parameters too. Then we should start a new control chart! It is the same after a re-engineering or changes of the process as shown in the Minitab 15.1.3 screenshot below:

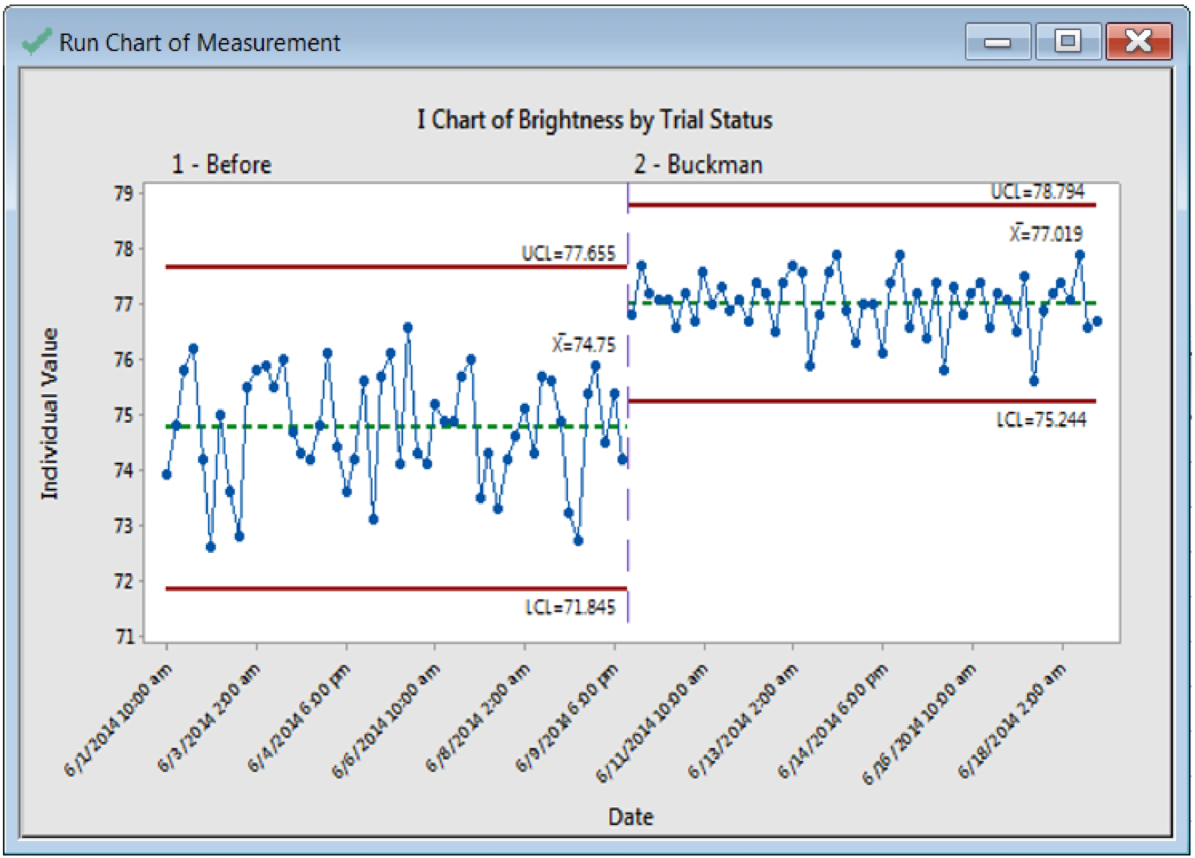

Run Chart of Measurement
I Chart of Brightness by Trial Status
1 - Before
2 - Buckman
79
78
77
76
75
74
73
72
71
Individual Value
UCL=77.655
$\bar{X}$=74.75
LCL=71.845
UCL=78.794
$\bar{X}$=77.019
LCL=75.244
6/1/2014 10:00 am
6/3/2014 2:00 am
6/4/2014 6:00 pm
6/6/2014 10:00 am
6/8/2014 2:00 am
6/9/2014 6:00 pm
6/11/2014 10:00 am
6/13/2014 2:00 am
6/14/2014 6:00 pm
6/16/2014 10:00 am
6/18/2014 2:00 am
Date

Figure 16.285: **Example of before/after control chart in Minitab 15.1.4**

16.7.6.3 Attributes Control Charts (qualitative CC)

To measure qualitative variables (% of defective, % failures, ...), we use attributes p control charts (binomial with proportion), np control charts (binomial with counts), c control charts (Poisson with counts), U control charts (Poisson with percentages) or Laney control charts (two by two range normalized proportions) to monitor the attributes quality control over time.

The attributes by control charts are very easy to study (undergraduate maths only!) use and interpret this is why we teach them always first, but has several disadvantages such as the high asymmetry of distributions, no lower limit for small sample sizes, low efficiency of deterioration detection.

16.7.6.3.1 P Control Charts (binomial proportion CC)

The p control chart (binomial by proportion) is used when it comes to work with ratios, proportions or percentages of compliance or non-compliance of samples.

Good examples of p-type control charts are: product inspection of a production line or receiving batches of a supplier or control malfunction of a number of devices, respect of deadlines, product specifications or also coding line errors (bugs).

Pros:

- This type of control chart is very simple to construct because does not require any advanced statistical knowledge to be calculated.

Cons:

- In reality it's a control chart which control limits may be asymmetric if the practical application are at the limit of acceptable when the theoretical assumptions of construction of this chart are not met.
- Sometimes a difficulty encountered by practitioners of this chart is to determine the batch size and the level of acceptance (A) or rejection (R) equivalent to the USL by taking the decision based on the developments (and norms) of acceptance plans samples seen earlier above.
- This control chart is not suitable for exhaustive sampling since it would then required the use of the hypergeometric law. Unfortunately, as far as we know, there is no analytic relation for the confidence interval of the proportion of a hypergeometric distribution, and therefore is not feasible to build a control chart based on a hypergeometric law.
- A final difficulty is that some employees and board committee are not proficient with the concept of percentages.

The probability law used in this context is the binomial distribution (see section Statistics page 658) where p will represent the proportion in $\%$ of non-compliants and q (which therefore is $1-p$) that proportion in $\%$ of compliants.

We saw in the section Statistics (see page 872) that we were able to write formally, under certain assumptions, the following confidence interval for the real proportion:

$$\hat{p}-Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\leq p\leq\hat{p}+Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad (16.7.714)$$

So we can calculate the probabilities of non-compliant batches by excess or by defect using the fact that p follows under statistical control a Normal distribution of parameters:

$$\mathcal{N}(\mu,\sigma)\cong\mathcal{N}\left(\hat{p},\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \qquad (16.7.715)$$

In short, we have no choice but to base this control chart on this assumption. Therefore, the first step in creating a p-type control chart is to calculate the proportion of non-compliance, at least its estimator for each batch i:

$$\hat{p}_i=\frac{k_i}{N_i} \tag{16.7.716}$$

where k_i is the number of non-compliant items in the batch of size N_i.

The average proportion of non-conformities, which corresponds to the center line (CL) of the control chart is given by (don't forget that using the median would be better in reality!):

$$\boxed{\bar{\hat{p}}_i=\text{CL}=\frac{\sum\limits_{i=1}^{n}k_i}{\sum\limits_{i=1}^{n}N_i}} \tag{16.7.717}$$

and if the n batches of size N_i are all of equal size N this simplifies obviously to:

$$\bar{\hat{p}}_i=\text{CL}=\frac{\sum\limits_{i=1}^{n}k_i}{nN}=\frac{1}{n}\sum_{i=1}^{n}\frac{k_i}{N}=\frac{1}{n}\sum_{i=1}^{n}\hat{p}_i \tag{16.7.718}$$

Therefore we have:

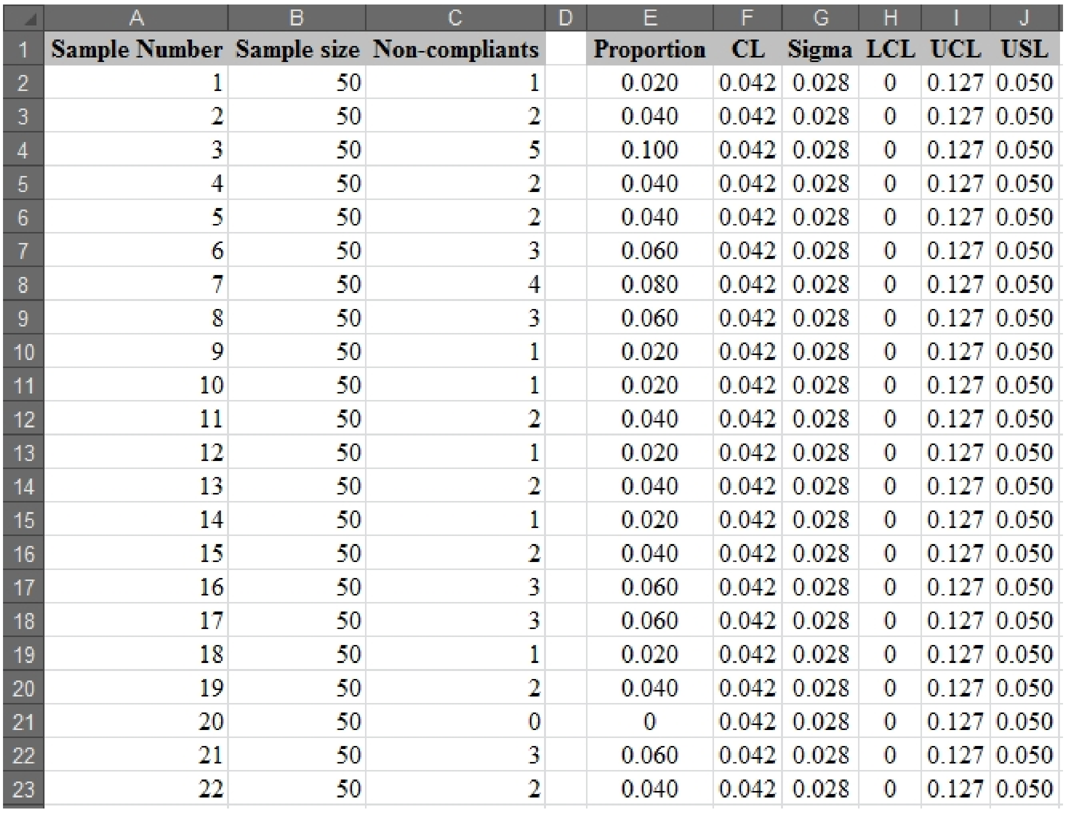
$$\boxed{\begin{aligned}\text{USL}&=\text{required by customer/board comitee}\\ \text{UCL}_p&=\bar{\hat{p}}+3\sigma_{\bar{\hat{p}}}=\bar{\hat{p}}+3\sqrt{\frac{\bar{\hat{p}}(1-\bar{\hat{p}})}{N_i}}\\ \text{CL}_p&=\bar{\hat{p}}\\ \text{LCL}_p&=\bar{\hat{p}}-3\sigma_{\bar{\hat{p}}}=\bar{\hat{p}}-3\sqrt{\frac{\bar{\hat{p}}(1-\bar{\hat{p}})}{N_i}}\end{aligned}} \tag{16.7.719}$$

and if the batches size are equal we have obviously:

$$\begin{aligned}\text{USL}&=\text{required by customer/board comitee}\\ \text{UCL}_p&=\bar{\hat{p}}+3\sqrt{\frac{\bar{\hat{p}}(1-\bar{\hat{p}})}{N}}\\ \text{CL}_p&=\bar{\hat{p}}\\ \text{LCL}_p&=\bar{\hat{p}}-3\sqrt{\frac{\bar{\hat{p}}(1-\bar{\hat{p}})}{N}}\end{aligned} \tag{16.7.720}$$

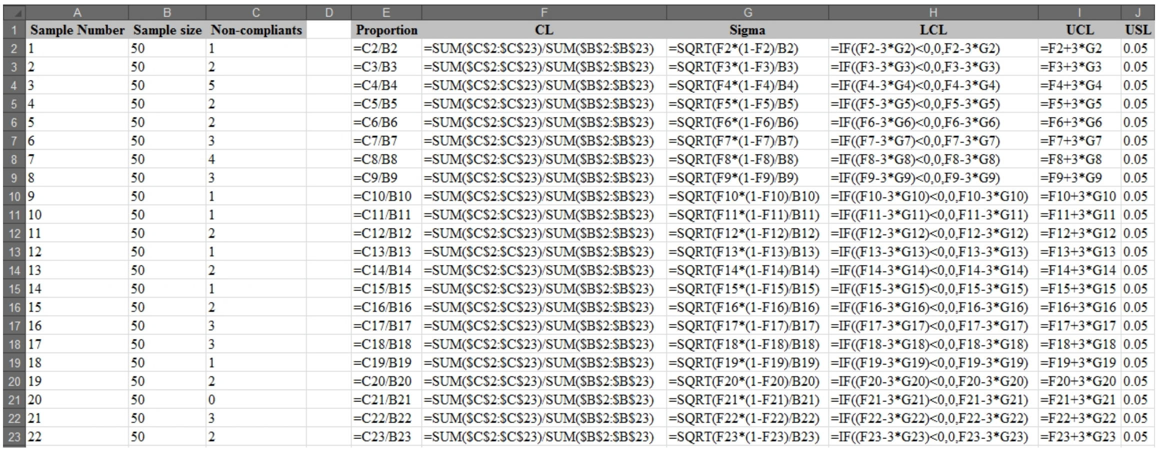
According to the study that we made during our study of acceptance plan they can not be any lower limit (LSL) imposed by the customer/board committee or target (T) in this field of application.

So, with the following list where the batches size (special case) are identical (made with Microsoft® Excel® 14.0.7166):



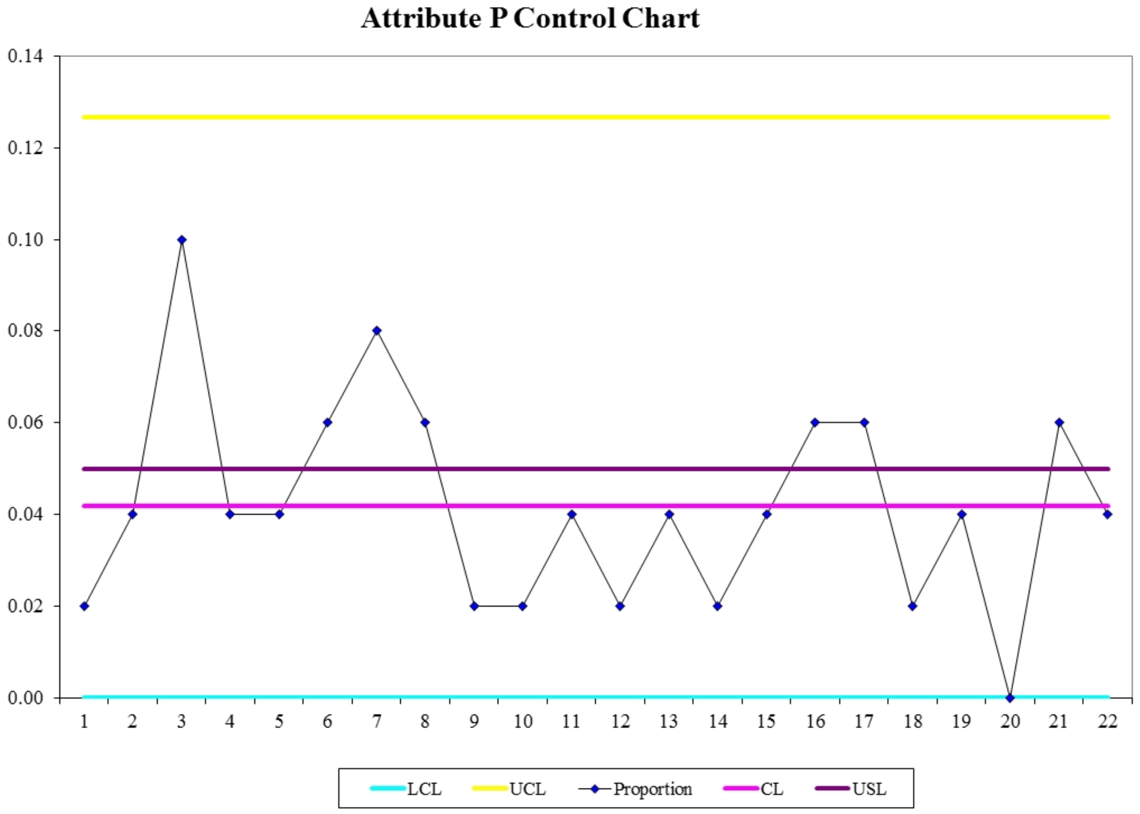
	A	B	C	D	E	F	G	H	I	J
1	Sample Number	Sample size	Non-compliants		Proportion	CL	Sigma	LCL	UCL	USL
2	1	50	1		0.020	0.042	0.028	0	0.127	0.050
3	2	50	2		0.040	0.042	0.028	0	0.127	0.050
4	3	50	5		0.100	0.042	0.028	0	0.127	0.050
5	4	50	2		0.040	0.042	0.028	0	0.127	0.050
6	5	50	2		0.040	0.042	0.028	0	0.127	0.050
7	6	50	3		0.060	0.042	0.028	0	0.127	0.050
8	7	50	4		0.080	0.042	0.028	0	0.127	0.050
9	8	50	3		0.060	0.042	0.028	0	0.127	0.050
10	9	50	1		0.020	0.042	0.028	0	0.127	0.050
11	10	50	1		0.020	0.042	0.028	0	0.127	0.050
12	11	50	2		0.040	0.042	0.028	0	0.127	0.050
13	12	50	1		0.020	0.042	0.028	0	0.127	0.050
14	13	50	2		0.040	0.042	0.028	0	0.127	0.050
15	14	50	1		0.020	0.042	0.028	0	0.127	0.050
16	15	50	2		0.040	0.042	0.028	0	0.127	0.050
17	16	50	3		0.060	0.042	0.028	0	0.127	0.050
18	17	50	3		0.060	0.042	0.028	0	0.127	0.050
19	18	50	1		0.020	0.042	0.028	0	0.127	0.050
20	19	50	2		0.040	0.042	0.028	0	0.127	0.050
21	20	50	0		0	0.042	0.028	0	0.127	0.050
22	21	50	3		0.060	0.042	0.028	0	0.127	0.050
23	22	50	2		0.040	0.042	0.028	0	0.127	0.050

or explicitly:

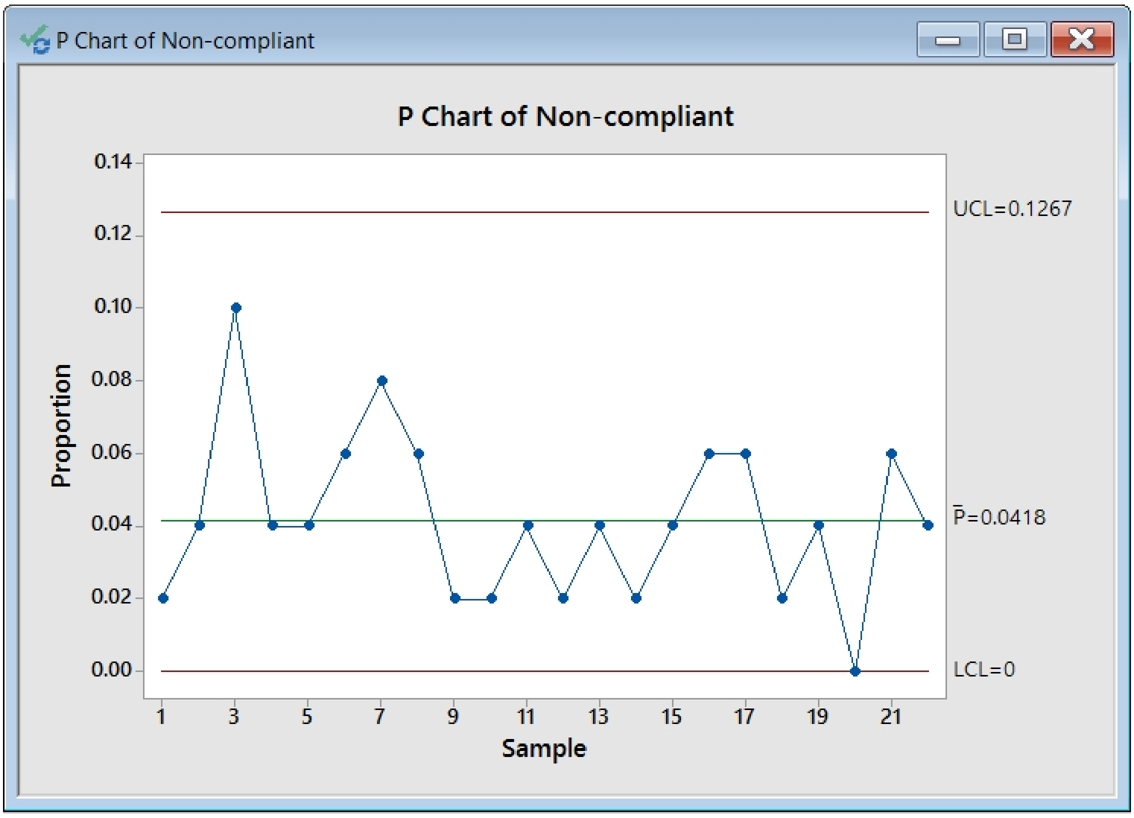


	A	B	C	D	E	F	G	H	I	J
1	Sample Number	Sample size	Non-compliants		Proportion	CL	Sigma	LCL	UCL	USL
2	1	50	1		=C2/B2	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F2*(1-F2)/B2)	=IF((F2-3*G2)<0,0,F2-3*G2)	=F2+3*G2	0.05
3	2	50	2		=C3/B3	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F3*(1-F3)/B3)	=IF((F3-3*G3)<0,0,F3-3*G3)	=F3+3*G3	0.05
4	3	50	5		=C4/B4	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F4*(1-F4)/B4)	=IF((F4-3*G4)<0,0,F4-3*G4)	=F4+3*G4	0.05
5	4	50	2		=C5/B5	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F5*(1-F5)/B5)	=IF((F5-3*G5)<0,0,F5-3*G5)	=F5+3*G5	0.05
6	5	50	2		=C6/B6	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F6*(1-F6)/B6)	=IF((F6-3*G6)<0,0,F6-3*G6)	=F6+3*G6	0.05
7	6	50	3		=C7/B7	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F7*(1-F7)/B7)	=IF((F7-3*G7)<0,0,F7-3*G7)	=F7+3*G7	0.05
8	7	50	4		=C8/B8	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F8*(1-F8)/B8)	=IF((F8-3*G8)<0,0,F8-3*G8)	=F8+3*G8	0.05
9	8	50	3		=C9/B9	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F9*(1-F9)/B9)	=IF((F9-3*G9)<0,0,F9-3*G9)	=F9+3*G9	0.05
10	9	50	1		=C10/B10	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F10*(1-F10)/B10)	=IF((F10-3*G10)<0,0,F10-3*G10)	=F10+3*G10	0.05
11	10	50	1		=C11/B11	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F11*(1-F11)/B11)	=IF((F11-3*G11)<0,0,F11-3*G11)	=F11+3*G11	0.05
12	11	50	2		=C12/B12	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F12*(1-F12)/B12)	=IF((F12-3*G12)<0,0,F12-3*G12)	=F12+3*G12	0.05
13	12	50	1		=C13/B13	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F13*(1-F13)/B13)	=IF((F13-3*G13)<0,0,F13-3*G13)	=F13+3*G13	0.05
14	13	50	2		=C14/B14	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F14*(1-F14)/B14)	=IF((F14-3*G14)<0,0,F14-3*G14)	=F14+3*G14	0.05
15	14	50	1		=C15/B15	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F15*(1-F15)/B15)	=IF((F15-3*G15)<0,0,F15-3*G15)	=F15+3*G15	0.05
16	15	50	2		=C16/B16	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F16*(1-F16)/B16)	=IF((F16-3*G16)<0,0,F16-3*G16)	=F16+3*G16	0.05
17	16	50	3		=C17/B17	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F17*(1-F17)/B17)	=IF((F17-3*G17)<0,0,F17-3*G17)	=F17+3*G17	0.05
18	17	50	3		=C18/B18	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F18*(1-F18)/B18)	=IF((F18-3*G18)<0,0,F18-3*G18)	=F18+3*G18	0.05
19	18	50	1		=C19/B19	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F19*(1-F19)/B19)	=IF((F19-3*G19)<0,0,F19-3*G19)	=F19+3*G19	0.05
20	19	50	2		=C20/B20	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F20*(1-F20)/B20)	=IF((F20-3*G20)<0,0,F20-3*G20)	=F20+3*G20	0.05
21	20	50	0		=C21/B21	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F21*(1-F21)/B21)	=IF((F21-3*G21)<0,0,F21-3*G21)	=F21+3*G21	0.05
22	21	50	3		=C22/B22	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F22*(1-F22)/B22)	=IF((F22-3*G22)<0,0,F22-3*G22)	=F22+3*G22	0.05
23	22	50	2		=C23/B23	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F23*(1-F23)/B23)	=IF((F23-3*G23)<0,0,F23-3*G23)	=F23+3*G23	0.05

and the corresponding chart still with the same spreadsheet software:



with Minitab 17.1.3 this gives for people that don't trust our calculation made with Microsoft® Excel® (we can also do such a chart with `R` but the plot quality output is not acceptable in most corporations):


P Chart of Non-compliant
P Chart of Non-compliant
Proportion
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
UCL=0.1267
P̄=0.0418
LCL=0
1 3 5 7 9 11 13 15 17 19 21
Sample

where we took as example for limits imposed by the customer/board committee:

$$\text{USL} = 5.00\% \tag{16.7.721}$$

obviously not visible on the chart above as too far away on the y-axes.

This control chart can also be interpreted as any another Shewhart chart using the empirical WECO rules.

16.7.6.3.2 NP Control Charts (binomial counting CC)

The np control chart is obviously similar to the previous p control but it is the number (count) of non-compliant items units than the proportion that is plotted. This chart is easier to interpret as the p control chart since not all employees are able to represent themselves the concept of percentage, but against it is very difficult to compare objectively two points coming from batches that do not have the same size.

Pros:

- This type of control chart is very simple to construct because does not require any advanced statistical knowledge to be calculated. For some employees whose concept of percentage is poorly mastered, it is easier to read a np chart.

Cons:

- In reality it is a control chart which control limits may be asymmetric if the practical application are at the limit of acceptable when the theoretical assumptions of construction of this chart are not met.
- Sometimes a difficulty encountered by practitioners of this chart is to determine the batch size and the level of acceptance (A) or rejection (R) equivalent to the USL by taking the decision based on the developments (and norms) of acceptance plans samples seen earlier above.
- This control chart is not suitable for exhaustive sampling since it would then required the use of the hypergeometric law. Unfortunately, as far as we know, there is no analytic relation for the confidence interval of the proportion of a hypergeometric distribution, and therefore is not feasible to build a control chart based on a hypergeometric law.
- It is very difficult to compare objectively two points coming from batches that do not have the same size

For the values of the control chart, we have still using:

$$\boxed{\bar{\hat{p}}_i = \text{CL} = \frac{\sum_{i=1}^{n} k_i}{\sum_{i=1}^{n} N_i}} \tag{16.7.722}$$

And therefore using the results of the previous control chart it comes immediately in the general case:

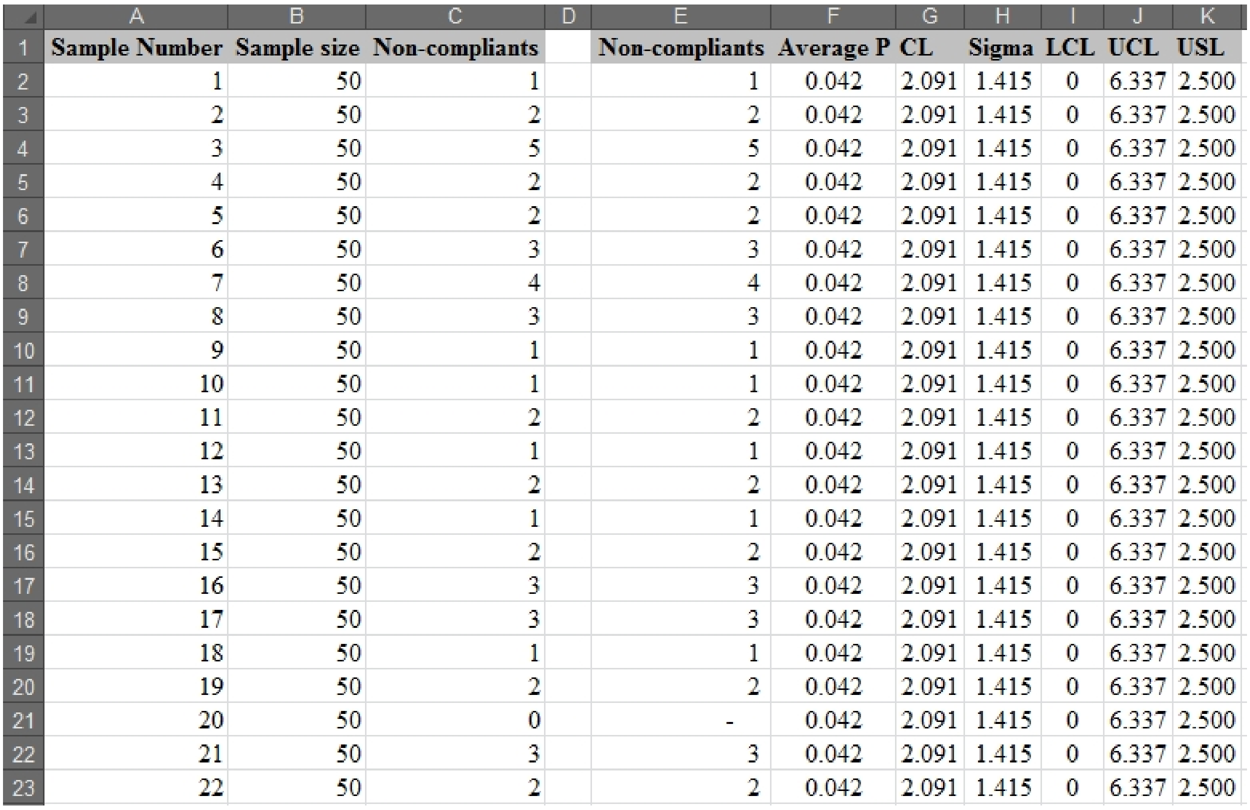
$$\boxed{\begin{aligned}\text{USL}&=\text{required by customer/board comitee}\\ \text{UCL}_{np}&=N_i\bar{\hat{p}}+3\sigma_{np}=N_i\bar{\hat{p}}+3\sqrt{\frac{N_i\bar{\hat{p}}N_i(1-\bar{\hat{p}})}{N_i}}=N_i\bar{\hat{p}}+3\sqrt{N_i\bar{\hat{p}}(1-\bar{\hat{p}})}\\ \text{CL}_{np}&=N_i\bar{\hat{p}}\\ \text{LCL}_{np}&=N_i\bar{\hat{p}}-3\sigma_{np}=N_i\bar{\hat{p}}-3\sqrt{\frac{N_i\bar{\hat{p}}N_i(1-\bar{\hat{p}})}{N_i}}=N_i\bar{\hat{p}}-3\sqrt{N_i\bar{\hat{p}}(1-\bar{\hat{p}})}\end{aligned}} \tag{16.7.723}$$

But since it is very difficult (unlike the p control chart) to compare objectively with this control chart two points when the size of the sample is not the same, some specialists required - or recommended - that the batch size is the same such that finally:

$$\begin{aligned}\text{USL}&=\text{required by customer/board comitee}\\ \text{UCL}_{np}&=N\bar{\hat{p}}+3\sqrt{N\bar{\hat{p}}(1-\bar{\hat{p}})}\\ \text{CL}_{np}&=N\bar{\hat{p}}\\ \text{LCL}_{np}&=N\bar{\hat{p}}-3\sqrt{N\bar{\hat{p}}(1-\bar{\hat{p}})}\end{aligned} \tag{16.7.724}$$

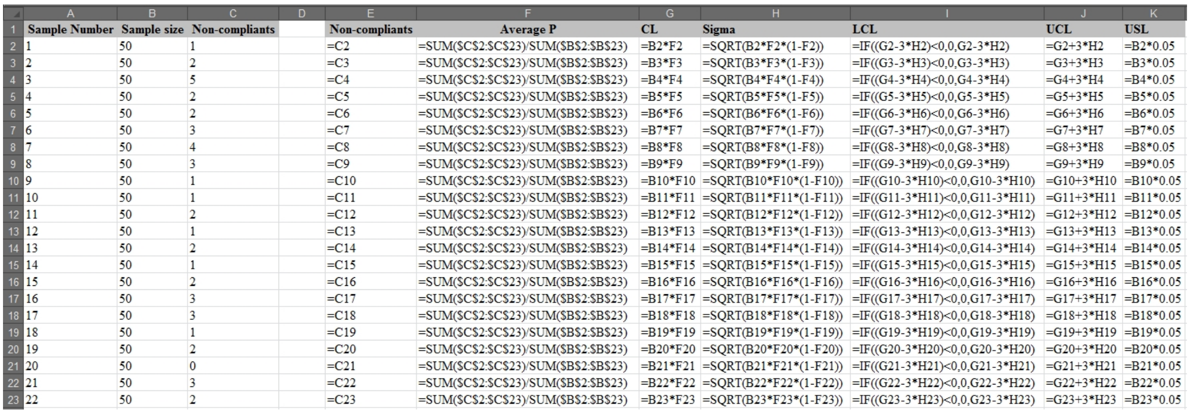
According to the study of acceptance plans we know that they can not be any lower limit (LSL) required by the customer / board committee or target (T) in this field of application.

So, with the following list where the batches size (special case) are identical (made with Microsoft® Excel® 14.0.7166):



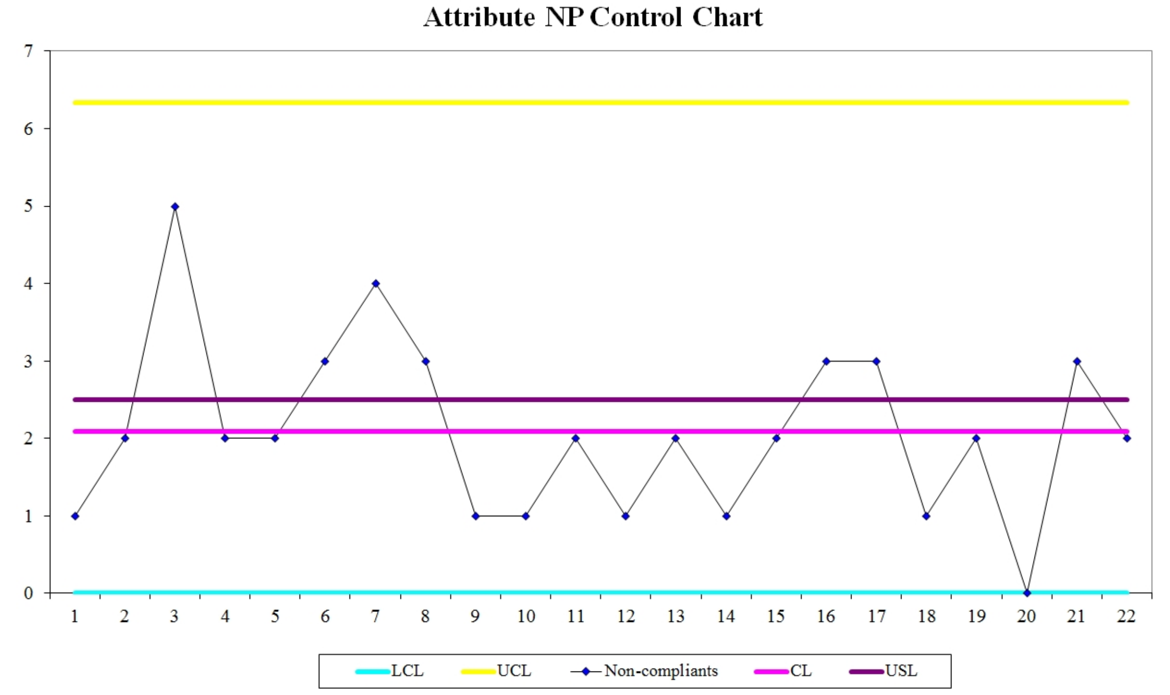
	A	B	C	D	E	F	G	H	I	J	K
1	**Sample Number**	**Sample size**	**Non-compliants**		**Non-compliants**	**Average P**	**CL**	**Sigma**	**LCL**	**UCL**	**USL**
2	1	50	1		1	0.042	2.091	1.415	0	6.337	2.500
3	2	50	2		2	0.042	2.091	1.415	0	6.337	2.500
4	3	50	5		5	0.042	2.091	1.415	0	6.337	2.500
5	4	50	2		2	0.042	2.091	1.415	0	6.337	2.500
6	5	50	2		2	0.042	2.091	1.415	0	6.337	2.500
7	6	50	3		3	0.042	2.091	1.415	0	6.337	2.500
8	7	50	4		4	0.042	2.091	1.415	0	6.337	2.500
9	8	50	3		3	0.042	2.091	1.415	0	6.337	2.500
10	9	50	1		1	0.042	2.091	1.415	0	6.337	2.500
11	10	50	1		1	0.042	2.091	1.415	0	6.337	2.500
12	11	50	2		2	0.042	2.091	1.415	0	6.337	2.500
13	12	50	1		1	0.042	2.091	1.415	0	6.337	2.500
14	13	50	2		2	0.042	2.091	1.415	0	6.337	2.500
15	14	50	1		1	0.042	2.091	1.415	0	6.337	2.500
16	15	50	2		2	0.042	2.091	1.415	0	6.337	2.500
17	16	50	3		3	0.042	2.091	1.415	0	6.337	2.500
18	17	50	3		3	0.042	2.091	1.415	0	6.337	2.500
19	18	50	1		1	0.042	2.091	1.415	0	6.337	2.500
20	19	50	2		2	0.042	2.091	1.415	0	6.337	2.500
21	20	50	0		-	0.042	2.091	1.415	0	6.337	2.500
22	21	50	3		3	0.042	2.091	1.415	0	6.337	2.500
23	22	50	2		2	0.042	2.091	1.415	0	6.337	2.500

or explicitly:

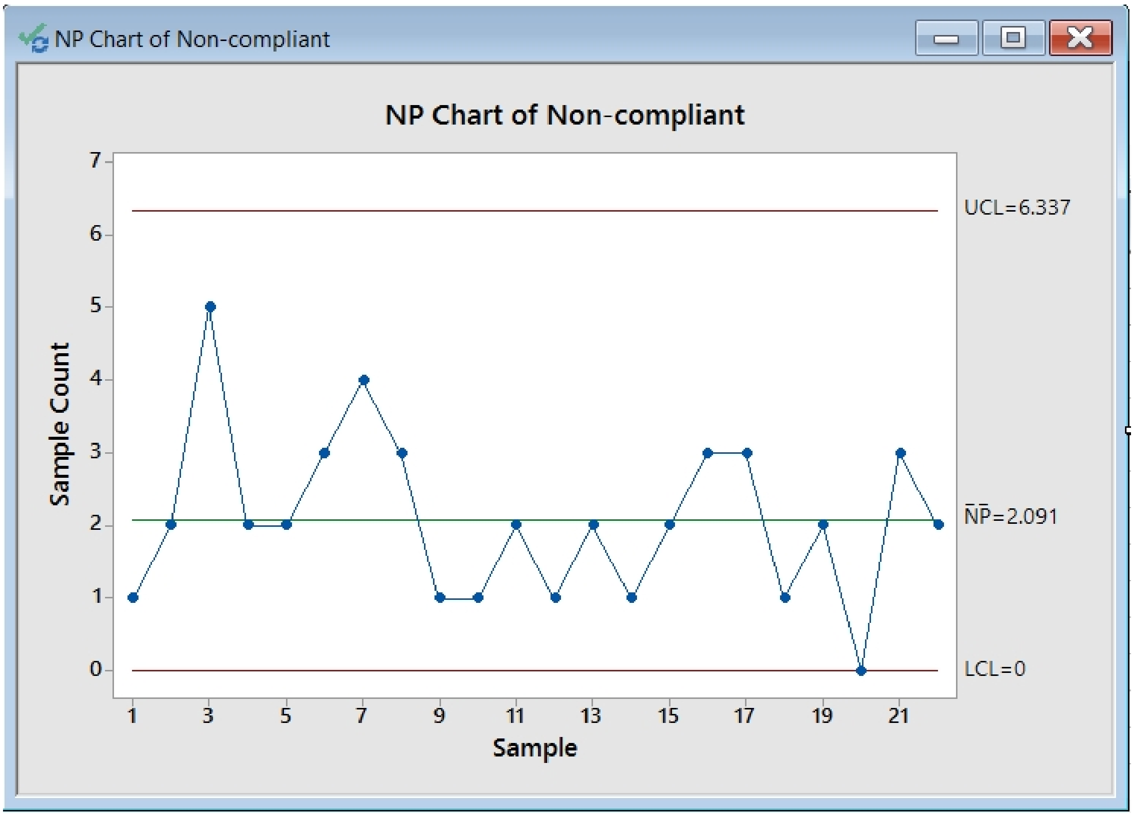


	A	B	C	D	E	F	G	H	I	J	K
1	**Sample Number**	**Sample size**	**Non-compliants**		**Non-compliants**	**Average P**	**CL**	**Sigma**	**LCL**	**UCL**	**USL**
2	1	50	1		=C2	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=B2*F2	=SQRT(B2*F2*(1-F2))	=IF((G2-3*H2)<0,0,G2-3*H2)	=G2+3*H2	=B2*0.05
3	2	50	2		=C3	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=B3*F3	=SQRT(B3*F3*(1-F3))	=IF((G3-3*H3)<0,0,G3-3*H3)	=G3+3*H3	=B3*0.05
4	3	50	5		=C4	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=B4*F4	=SQRT(B4*F4*(1-F4))	=IF((G4-3*H4)<0,0,G4-3*H4)	=G4+3*H4	=B4*0.05
5	4	50	2		=C5	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=B5*F5	=SQRT(B5*F5*(1-F5))	=IF((G5-3*H5)<0,0,G5-3*H5)	=G5+3*H5	=B5*0.05
6	5	50	2		=C6	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=B6*F6	=SQRT(B6*F6*(1-F6))	=IF((G6-3*H6)<0,0,G6-3*H6)	=G6+3*H6	=B6*0.05
7	6	50	3		=C7	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=B7*F7	=SQRT(B7*F7*(1-F7))	=IF((G7-3*H7)<0,0,G7-3*H7)	=G7+3*H7	=B7*0.05
8	7	50	4		=C8	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=B8*F8	=SQRT(B8*F8*(1-F8))	=IF((G8-3*H8)<0,0,G8-3*H8)	=G8+3*H8	=B8*0.05
9	8	50	3		=C9	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=B9*F9	=SQRT(B9*F9*(1-F9))	=IF((G9-3*H9)<0,0,G9-3*H9)	=G9+3*H9	=B9*0.05
10	9	50	1		=C10	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=B10*F10	=SQRT(B10*F10*(1-F10))	=IF((G10-3*H10)<0,0,G10-3*H10)	=G10+3*H10	=B10*0.05
11	10	50	1		=C11	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=B11*F11	=SQRT(B11*F11*(1-F11))	=IF((G11-3*H11)<0,0,G11-3*H11)	=G11+3*H11	=B11*0.05
12	11	50	2		=C12	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=B12*F12	=SQRT(B12*F12*(1-F12))	=IF((G12-3*H12)<0,0,G12-3*H12)	=G12+3*H12	=B12*0.05
13	12	50	1		=C13	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=B13*F13	=SQRT(B13*F13*(1-F13))	=IF((G13-3*H13)<0,0,G13-3*H13)	=G13+3*H13	=B13*0.05
14	13	50	2		=C14	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=B14*F14	=SQRT(B14*F14*(1-F14))	=IF((G14-3*H14)<0,0,G14-3*H14)	=G14+3*H14	=B14*0.05
15	14	50	1		=C15	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=B15*F15	=SQRT(B15*F15*(1-F15))	=IF((G15-3*H15)<0,0,G15-3*H15)	=G15+3*H15	=B15*0.05
16	15	50	2		=C16	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=B16*F16	=SQRT(B16*F16*(1-F16))	=IF((G16-3*H16)<0,0,G16-3*H16)	=G16+3*H16	=B16*0.05
17	16	50	3		=C17	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=B17*F17	=SQRT(B17*F17*(1-F17))	=IF((G17-3*H17)<0,0,G17-3*H17)	=G17+3*H17	=B17*0.05
18	17	50	3		=C18	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=B18*F18	=SQRT(B18*F18*(1-F18))	=IF((G18-3*H18)<0,0,G18-3*H18)	=G18+3*H18	=B18*0.05
19	18	50	1		=C19	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=B19*F19	=SQRT(B19*F19*(1-F19))	=IF((G19-3*H19)<0,0,G19-3*H19)	=G19+3*H19	=B19*0.05
20	19	50	2		=C20	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=B20*F20	=SQRT(B20*F20*(1-F20))	=IF((G20-3*H20)<0,0,G20-3*H20)	=G20+3*H20	=B20*0.05
21	20	50	0		=C21	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=B21*F21	=SQRT(B21*F21*(1-F21))	=IF((G21-3*H21)<0,0,G21-3*H21)	=G21+3*H21	=B21*0.05
22	21	50	3		=C22	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=B22*F22	=SQRT(B22*F22*(1-F22))	=IF((G22-3*H22)<0,0,G22-3*H22)	=G22+3*H22	=B22*0.05
23	22	50	2		=C23	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=B23*F23	=SQRT(B23*F23*(1-F23))	=IF((G23-3*H23)<0,0,G23-3*H23)	=G23+3*H23	=B23*0.05

and the corresponding chart still with the same spreadsheet software:



with Minitab 17.1.3 this gives for people that don't trust our calculation made with Microsoft® Excel® (we can also do such a chart with `R` but the plot quality output is not acceptable in most corporations):


NP Chart of Non-compliant
NP Chart of Non-compliant
Sample Count
7
6
5
4
3
2
1
0
UCL=6.337
$\overline{NP}$=2.091
LCL=0
1
3
5
7
9
11
13
15
17
19
21
Sample

where we took as example for limits imposed by the customer/board committee:

$$\text{USL} = 5.00\% \cdot N = 5.00\% \cdot 50 = 2.5 \tag{16.7.725}$$

This control chart can also be interpreted as any another Shewhart chart using the empirical WECO rules.

16.7.6.3.3 C Control Charts (Poisson counting CC)

The C control charts are used to look at variation in counting type attributes data. They are used to determine the variation in the number of defects in a constant subgroup size. To use the C chart, the opportunities for defects to occur in the subgroup must be very large, but the number that actually occurs must be small.

Indeed, during our study of statistical laws (or statistical distributions as seen at page 633), we have proved that when the probability p is very small and tends to zero, but that instead the average value $\mu = np$ tends to a fixed value when n approaches infinity, the binomial distribution of mean $\mu = np$ with k trials was therefore given by a Poisson law (see page 686):

$$P(X=k)=\frac{\mu^k}{k!}e^{-\mu} \tag{16.7.726}$$

with:

$$\begin{aligned} \mu &= \mu \\ \sigma &= \sqrt{\mu} \end{aligned} \tag{16.7.727}$$

In practice, some replace the binomial distribution by a Poisson law since $n > 30$ and $np < 5$ or when $n > 50$ and $p < 0.1$... In practice I personally recommend the use of the approximation when $n > 1000$ and $p < 0.01$.

As in practice such situations does not happen very often. Some books and authors say that this control chart is reserved to the count of defects per unit opportunity (see our study of Six Sigma at the beginning of this chapter). Indeed, the fact that each unit can be composed of multiple items this increase "artificially" the n (as 1 unit opportunity can have dozens, hundreds of millions of components) and decrease most of time the probability p such that the assumption of using the Poisson law are respected.

Following the previous rules it comes obviously:

$$\begin{aligned} \text{USL} &= \text{required by customer/board comitee} \\ \text{UCL}_c &= \mu + 3\sqrt{\mu} \\ \text{CL}_c &= \mu \\ \text{LCL}_c &= \mu - 3\sqrt{\mu} \end{aligned} \tag{16.7.728}$$

what is traditionally noted in this field:

$$\boxed{\begin{aligned} \text{USL} &= \text{required by customer/board comitee} \\ \text{UCL}_c &= \bar{c} + 3\sqrt{\bar{c}} \\ \text{CL}_c &= \bar{c} \\ \text{LCL}_c &= \bar{c} - 3\sqrt{\bar{c}} \end{aligned}} \tag{16.7.729}$$

where $\bar{c}$ is simply the average of the count (that's form this word that comes from the "c" in the name of the chart) of nonconformities:

$$\bar{c}=\frac{1}{n}\sum_{i=1}^{n}c_i \tag{16.7.730}$$

According to the study that we made during our study of acceptance plan they can not be any lower limit (LSL) imposed by the customer/board committee or target (T) in this field of application.

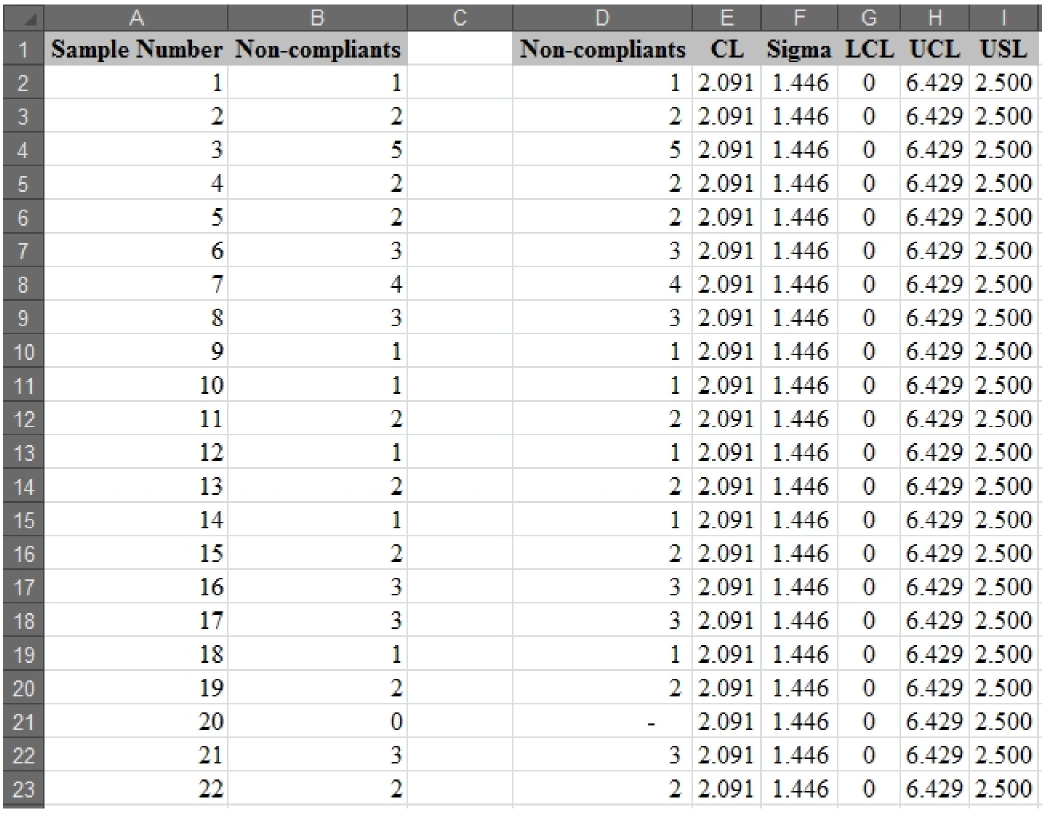
Pros:

- This type of control chart is very simple to construct because does not require any advanced statistical knowledge to be calculated.

Cons:

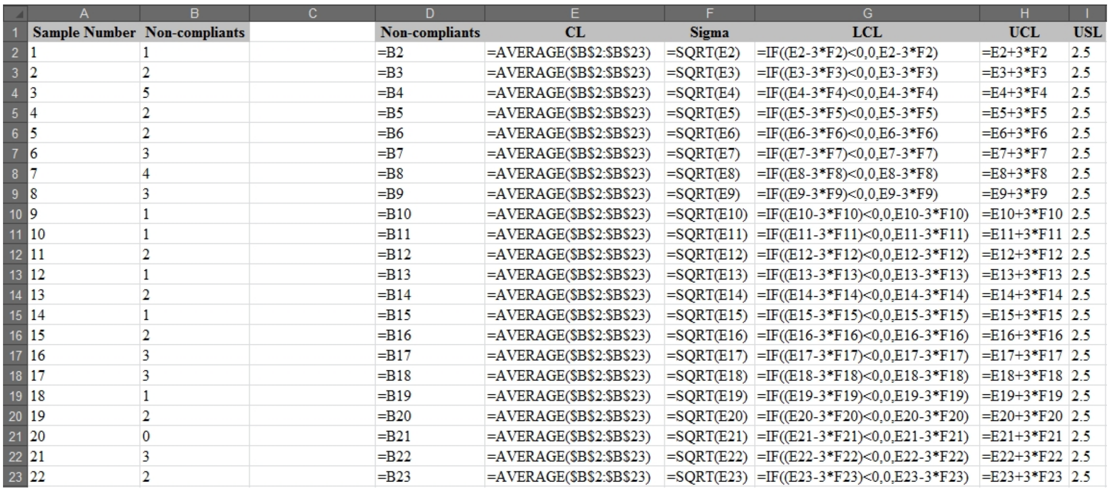
- In reality it's a control chart which control limits may be asymmetric if the practical application are at the limit of acceptable when the theoretical assumptions of construction of this chart are not met.
- Sometimes a difficulty encountered by practitioners of this chart is to determine the batch size and the level of acceptance (A) or rejection (R) equivalent to the USL by taking the decision based on the developments (and norms) of acceptance plans samples seen earlier above.
- This control chart is not suitable for exhaustive sampling since it would then required the use of the binomial or hypergeometric law.

Once again using our spreadsheet software here is an example where the batch sizes are necessarily identical for the reasons mentioned just above (and once again... non-compliant items can not be expressed in $\%$ of the batch size as the Poisson distribution is a discrete law!):



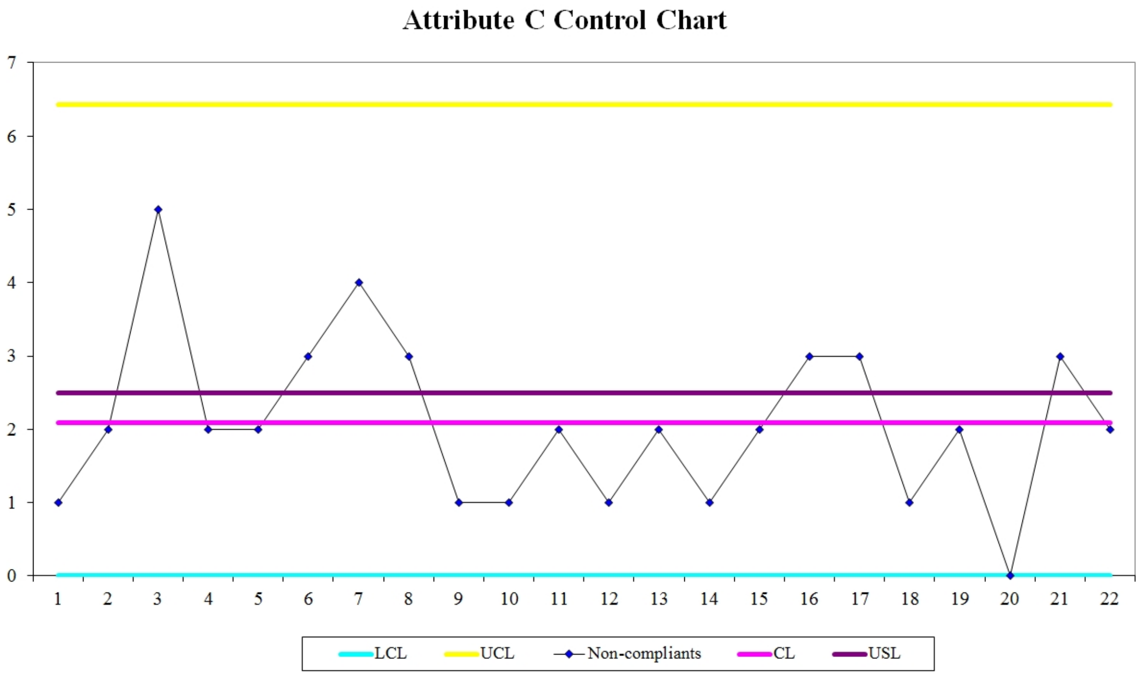
	A	B	C	D	E	F	G	H	I
1	**Sample Number**	**Non-compliants**		**Non-compliants**	**CL**	**Sigma**	**LCL**	**UCL**	**USL**
2	1	1		1	2.091	1.446	0	6.429	2.500
3	2	2		2	2.091	1.446	0	6.429	2.500
4	3	5		5	2.091	1.446	0	6.429	2.500
5	4	2		2	2.091	1.446	0	6.429	2.500
6	5	2		2	2.091	1.446	0	6.429	2.500
7	6	3		3	2.091	1.446	0	6.429	2.500
8	7	4		4	2.091	1.446	0	6.429	2.500
9	8	3		3	2.091	1.446	0	6.429	2.500
10	9	1		1	2.091	1.446	0	6.429	2.500
11	10	1		1	2.091	1.446	0	6.429	2.500
12	11	2		2	2.091	1.446	0	6.429	2.500
13	12	1		1	2.091	1.446	0	6.429	2.500
14	13	2		2	2.091	1.446	0	6.429	2.500
15	14	1		1	2.091	1.446	0	6.429	2.500
16	15	2		2	2.091	1.446	0	6.429	2.500
17	16	3		3	2.091	1.446	0	6.429	2.500
18	17	3		3	2.091	1.446	0	6.429	2.500
19	18	1		1	2.091	1.446	0	6.429	2.500
20	19	2		2	2.091	1.446	0	6.429	2.500
21	20	0		-	2.091	1.446	0	6.429	2.500
22	21	3		3	2.091	1.446	0	6.429	2.500
23	22	2		2	2.091	1.446	0	6.429	2.500

or explicitly:

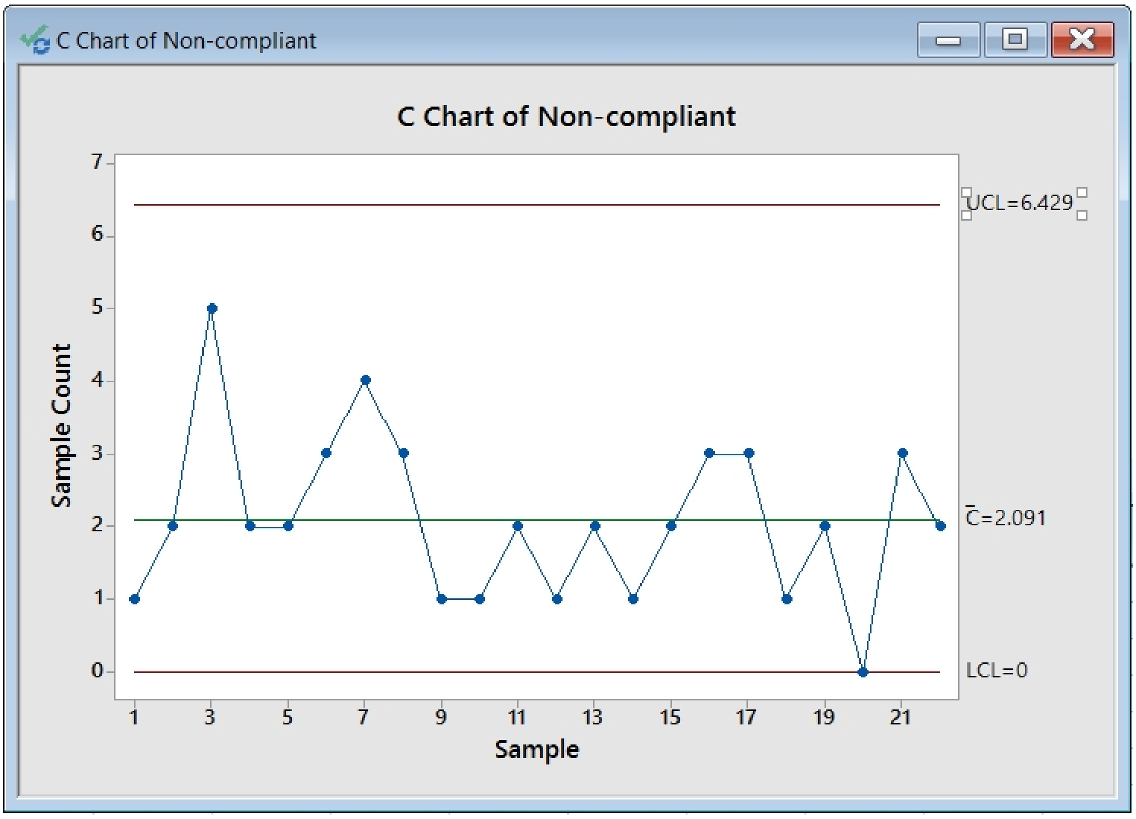


	A	B	C	D	E	F	G	H	I
1	**Sample Number**	**Non-compliants**		**Non-compliants**	**CL**	**Sigma**	**LCL**	**UCL**	**USL**
2	1	1		=B2	=AVERAGE($B$2:$B$23)	=SQRT(E2)	=IF((E2-3*F2)<0,0,E2-3*F2)	=E2+3*F2	2.5
3	2	2		=B3	=AVERAGE($B$2:$B$23)	=SQRT(E3)	=IF((E3-3*F3)<0,0,E3-3*F3)	=E3+3*F3	2.5
4	3	5		=B4	=AVERAGE($B$2:$B$23)	=SQRT(E4)	=IF((E4-3*F4)<0,0,E4-3*F4)	=E4+3*F4	2.5
5	4	2		=B5	=AVERAGE($B$2:$B$23)	=SQRT(E5)	=IF((E5-3*F5)<0,0,E5-3*F5)	=E5+3*F5	2.5
6	5	2		=B6	=AVERAGE($B$2:$B$23)	=SQRT(E6)	=IF((E6-3*F6)<0,0,E6-3*F6)	=E6+3*F6	2.5
7	6	3		=B7	=AVERAGE($B$2:$B$23)	=SQRT(E7)	=IF((E7-3*F7)<0,0,E7-3*F7)	=E7+3*F7	2.5
8	7	4		=B8	=AVERAGE($B$2:$B$23)	=SQRT(E8)	=IF((E8-3*F8)<0,0,E8-3*F8)	=E8+3*F8	2.5
9	8	3		=B9	=AVERAGE($B$2:$B$23)	=SQRT(E9)	=IF((E9-3*F9)<0,0,E9-3*F9)	=E9+3*F9	2.5
10	9	1		=B10	=AVERAGE($B$2:$B$23)	=SQRT(E10)	=IF((E10-3*F10)<0,0,E10-3*F10)	=E10+3*F10	2.5
11	10	1		=B11	=AVERAGE($B$2:$B$23)	=SQRT(E11)	=IF((E11-3*F11)<0,0,E11-3*F11)	=E11+3*F11	2.5
12	11	2		=B12	=AVERAGE($B$2:$B$23)	=SQRT(E12)	=IF((E12-3*F12)<0,0,E12-3*F12)	=E12+3*F12	2.5
13	12	1		=B13	=AVERAGE($B$2:$B$23)	=SQRT(E13)	=IF((E13-3*F13)<0,0,E13-3*F13)	=E13+3*F13	2.5
14	13	2		=B14	=AVERAGE($B$2:$B$23)	=SQRT(E14)	=IF((E14-3*F14)<0,0,E14-3*F14)	=E14+3*F14	2.5
15	14	1		=B15	=AVERAGE($B$2:$B$23)	=SQRT(E15)	=IF((E15-3*F15)<0,0,E15-3*F15)	=E15+3*F15	2.5
16	15	2		=B16	=AVERAGE($B$2:$B$23)	=SQRT(E16)	=IF((E16-3*F16)<0,0,E16-3*F16)	=E16+3*F16	2.5
17	16	3		=B17	=AVERAGE($B$2:$B$23)	=SQRT(E17)	=IF((E17-3*F17)<0,0,E17-3*F17)	=E17+3*F17	2.5
18	17	3		=B18	=AVERAGE($B$2:$B$23)	=SQRT(E18)	=IF((E18-3*F18)<0,0,E18-3*F18)	=E18+3*F18	2.5
19	18	1		=B19	=AVERAGE($B$2:$B$23)	=SQRT(E19)	=IF((E19-3*F19)<0,0,E19-3*F19)	=E19+3*F19	2.5
20	19	2		=B20	=AVERAGE($B$2:$B$23)	=SQRT(E20)	=IF((E20-3*F20)<0,0,E20-3*F20)	=E20+3*F20	2.5
21	20	0		=B21	=AVERAGE($B$2:$B$23)	=SQRT(E21)	=IF((E21-3*F21)<0,0,E21-3*F21)	=E21+3*F21	2.5
22	21	3		=B22	=AVERAGE($B$2:$B$23)	=SQRT(E22)	=IF((E22-3*F22)<0,0,E22-3*F22)	=E22+3*F22	2.5
23	22	2		=B23	=AVERAGE($B$2:$B$23)	=SQRT(E23)	=IF((E23-3*F23)<0,0,E23-3*F23)	=E23+3*F23	2.5

and the corresponding chart still with the same spreadsheet software:



with Minitab 17.1.3 this gives for people that don't trust our calculation made with Microsoft® Excel® (we can also do such a chart with `R` but the plot quality output is not acceptable in most corporations):


C Chart of Non-compliant
C Chart of Non-compliant
Sample Count
7
6
5
4
3
2
1
0
UCL=6.429
C̄=2.091
LCL=0
1
3
5
7
9
11
13
15
17
19
21
Sample

where we took as example for limits imposed by the customer/board committee:

$$\text{USL} = 2.5 \tag{16.7.731}$$

This control chart can also be interpreted as any another Shewhart chart using the empirical WECO rules.

16.7.6.3.4 U Control Charts (normalized Poisson)

The u control chart is similar to the c control chart except that it has for role to normalize the data to the unit to have a c control chart based on a Poisson distribution but with percentages.

According to the study that we made during our study of acceptance plan they can not be any lower limit (LSL) imposed by the customer/board committee or target (T) in this field of application.

Pros:

- This type of control chart is very simple to construct because does not require any advanced statistical knowledge to be calculated and it is now more easy to compare objectively two points coming from batches that do not have the same size.

Cons:

- In reality its a control chart which control limits may be asymmetric if the practical application are at the limit of acceptable when the theoretical assumptions of construction of this chart are not met.
- Sometimes a difficulty encountered by practitioners of this chart is to determine the batch size and the level of acceptance (A) or rejection (R) equivalent to the USL by taking the decision based on the developments (and norms) of acceptance plans samples seen earlier above.
- This control chart is not suitable for exhaustive sampling since it would then required the use of the binomial or hypergeometric law.
- A final difficulty is that some employees and board committee are not proficient with the concept of percentages.

For this, we use the same working assumptions as those proved during our study of the Poisson distribution:

$$\begin{aligned}\mu &= \mu = Np\\ \sigma &= \sqrt{\mu} = \sqrt{Np}\end{aligned} \tag{16.7.732}$$

This makes that our previous c control chart is then written:

$$\begin{aligned}\text{USL}&=\text{required by customer/board comitee}\\ \text{UCL}_u&=n_i\bar{\hat{p}}+3\sqrt{n_i\bar{\hat{p}}}\\ \text{CL}_u&=n_i\bar{\hat{p}}\\ \text{LCL}_u&=n_i\bar{\hat{p}}-3\sqrt{n_i\bar{\hat{p}}}\end{aligned} \tag{16.7.733}$$

If we divide the random variable $k=n_i\bar{\hat{p}}$ by the batch size (or number of components) we have then by the property of the mean and variance:

$$\begin{aligned}\text{USL}&=\text{required by customer/board comitee}\\ \text{UCL}_u&=\bar{\hat{p}}+3\sqrt{\frac{\bar{\hat{p}}}{n_i}}\\ \text{CL}_u&=\bar{\hat{p}}\\ \text{LCL}_u&=\bar{\hat{p}}-3\sqrt{\frac{\bar{\hat{p}}}{n_i}}\end{aligned} \tag{16.7.734}$$

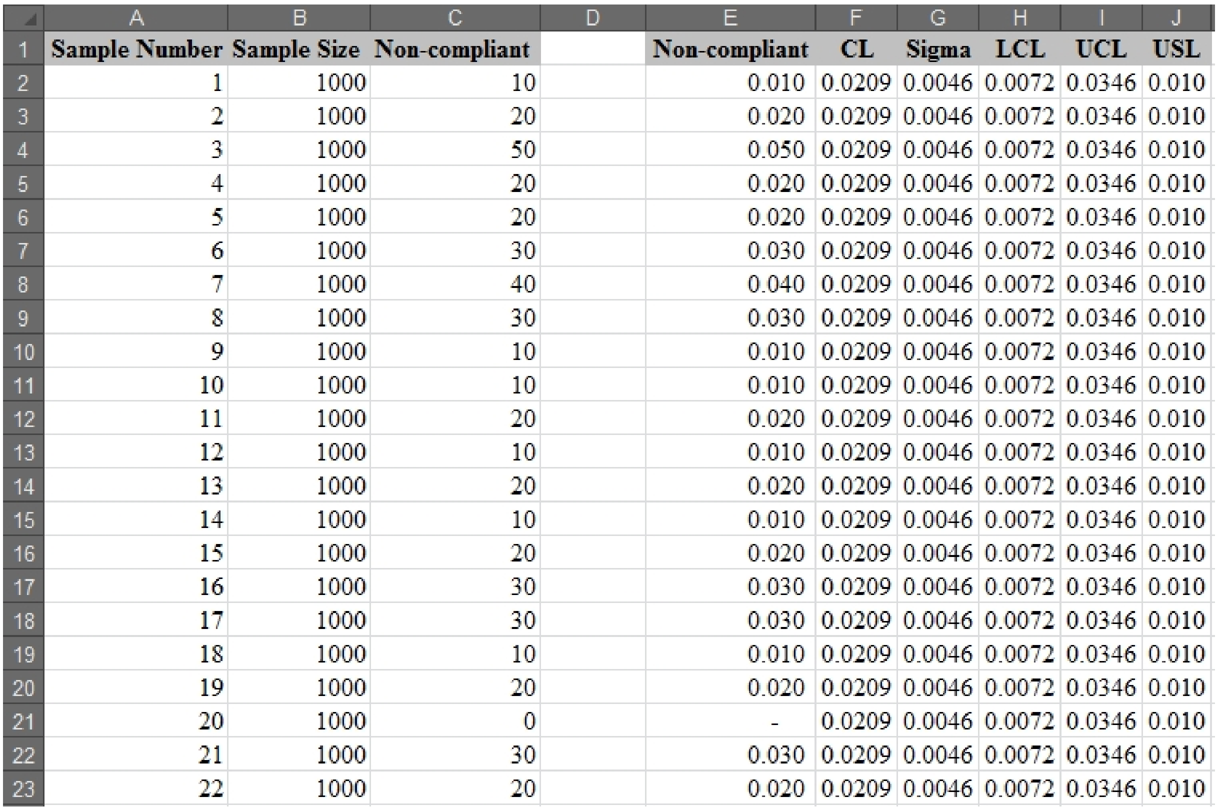
what is traditionally written:

$$\boxed{\begin{aligned}\text{USL}&=\text{required by customer/board comitee}\\ \text{UCL}_u&=\bar{u}+3\sqrt{\frac{\bar{u}}{n_i}}\\ \text{CL}_u&=\bar{u}\\ \text{LCL}_u&=\bar{u}-3\sqrt{\frac{\bar{u}}{n_i}}\end{aligned}} \tag{16.7.735}$$

with:

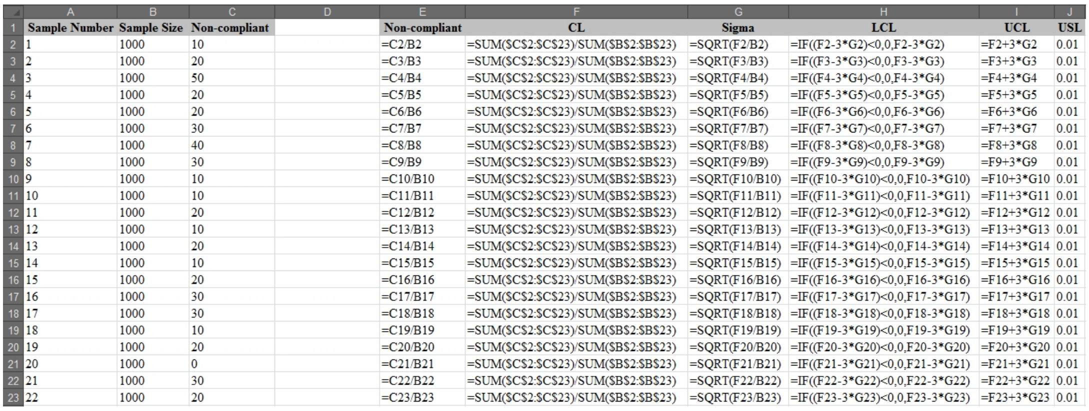
$$\boxed{\bar{u}=\text{CL}_u=\frac{\displaystyle\sum_{i=1}^{n}c_i}{\displaystyle\sum_{i=1}^{n}n_i}} \tag{16.7.736}$$

Once again using our spreadsheet software here is an example where the batch sizes are non-necessarily identical this time using the "trick" above (and once again... non-compliant items can now be expressed in % of the batch size!):



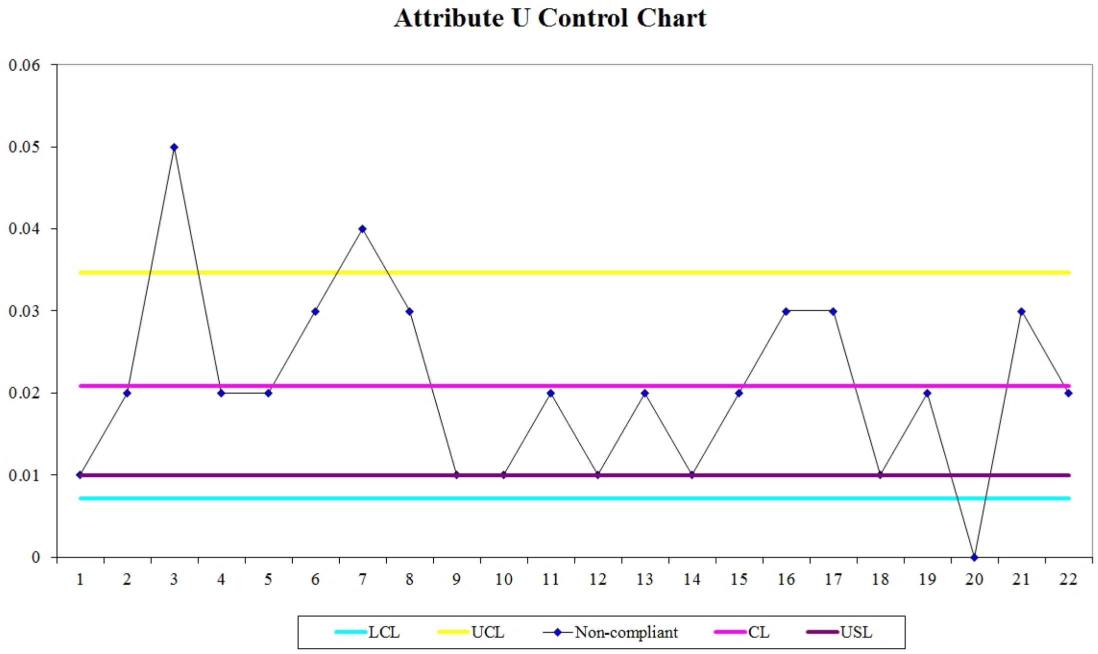
	A	B	C	D	E	F	G	H	I	J
1	**Sample Number**	**Sample Size**	**Non-compliant**		**Non-compliant**	**CL**	**Sigma**	**LCL**	**UCL**	**USL**
2	1	1000	10		0.010	0.0209	0.0046	0.0072	0.0346	0.010
3	2	1000	20		0.020	0.0209	0.0046	0.0072	0.0346	0.010
4	3	1000	50		0.050	0.0209	0.0046	0.0072	0.0346	0.010
5	4	1000	20		0.020	0.0209	0.0046	0.0072	0.0346	0.010
6	5	1000	20		0.020	0.0209	0.0046	0.0072	0.0346	0.010
7	6	1000	30		0.030	0.0209	0.0046	0.0072	0.0346	0.010
8	7	1000	40		0.040	0.0209	0.0046	0.0072	0.0346	0.010
9	8	1000	30		0.030	0.0209	0.0046	0.0072	0.0346	0.010
10	9	1000	10		0.010	0.0209	0.0046	0.0072	0.0346	0.010
11	10	1000	10		0.010	0.0209	0.0046	0.0072	0.0346	0.010
12	11	1000	20		0.020	0.0209	0.0046	0.0072	0.0346	0.010
13	12	1000	10		0.010	0.0209	0.0046	0.0072	0.0346	0.010
14	13	1000	20		0.020	0.0209	0.0046	0.0072	0.0346	0.010
15	14	1000	10		0.010	0.0209	0.0046	0.0072	0.0346	0.010
16	15	1000	20		0.020	0.0209	0.0046	0.0072	0.0346	0.010
17	16	1000	30		0.030	0.0209	0.0046	0.0072	0.0346	0.010
18	17	1000	30		0.030	0.0209	0.0046	0.0072	0.0346	0.010
19	18	1000	10		0.010	0.0209	0.0046	0.0072	0.0346	0.010
20	19	1000	20		0.020	0.0209	0.0046	0.0072	0.0346	0.010
21	20	1000	0		-	0.0209	0.0046	0.0072	0.0346	0.010
22	21	1000	30		0.030	0.0209	0.0046	0.0072	0.0346	0.010
23	22	1000	20		0.020	0.0209	0.0046	0.0072	0.0346	0.010

or explicitly:

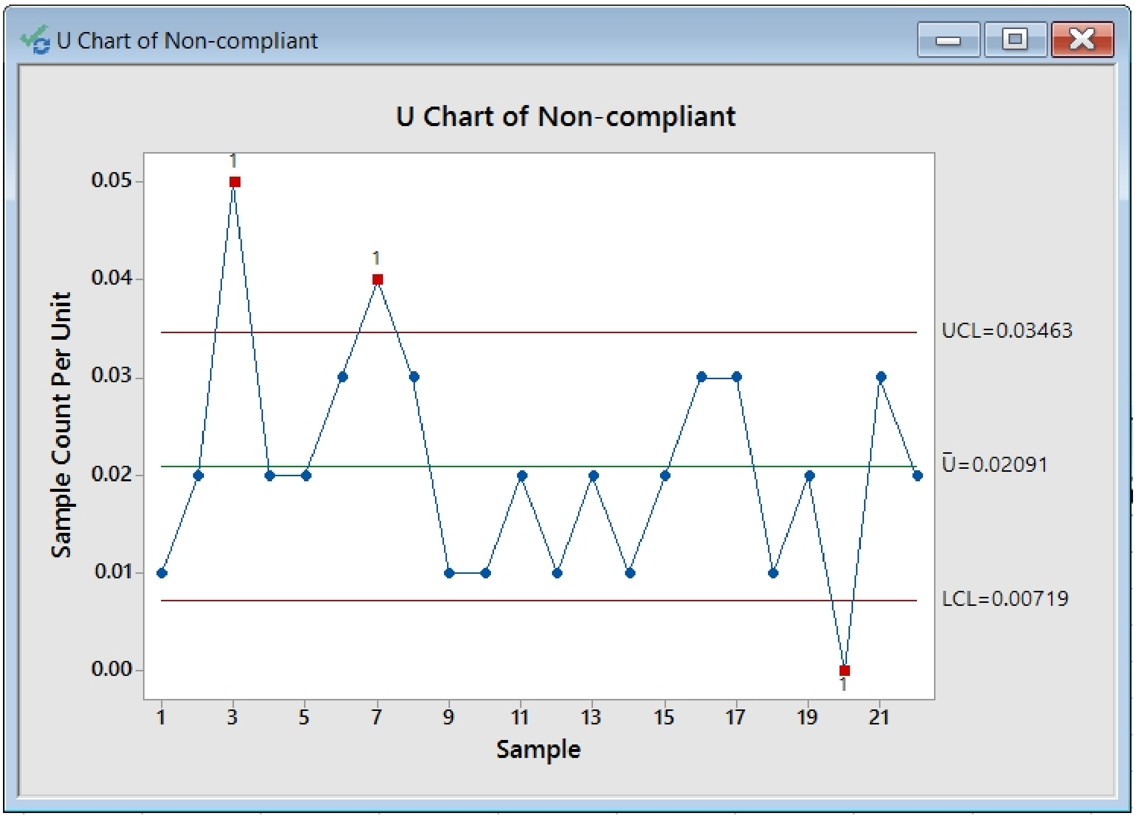


	A	B	C	D	E	F	G	H	I	J
1	**Sample Number**	**Sample Size**	**Non-compliant**		**Non-compliant**	**CL**	**Sigma**	**LCL**	**UCL**	**USL**
2	1	1000	10		=C2/B2	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F2/B2)	=IF((F2-3*G2)<0,0,F2-3*G2)	=F2+3*G2	0.01
3	2	1000	20		=C3/B3	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F3/B3)	=IF((F3-3*G3)<0,0,F3-3*G3)	=F3+3*G3	0.01
4	3	1000	50		=C4/B4	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F4/B4)	=IF((F4-3*G4)<0,0,F4-3*G4)	=F4+3*G4	0.01
5	4	1000	20		=C5/B5	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F5/B5)	=IF((F5-3*G5)<0,0,F5-3*G5)	=F5+3*G5	0.01
6	5	1000	20		=C6/B6	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F6/B6)	=IF((F6-3*G6)<0,0,F6-3*G6)	=F6+3*G6	0.01
7	6	1000	30		=C7/B7	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F7/B7)	=IF((F7-3*G7)<0,0,F7-3*G7)	=F7+3*G7	0.01
8	7	1000	40		=C8/B8	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F8/B8)	=IF((F8-3*G8)<0,0,F8-3*G8)	=F8+3*G8	0.01
9	8	1000	30		=C9/B9	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F9/B9)	=IF((F9-3*G9)<0,0,F9-3*G9)	=F9+3*G9	0.01
10	9	1000	10		=C10/B10	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F10/B10)	=IF((F10-3*G10)<0,0,F10-3*G10)	=F10+3*G10	0.01
11	10	1000	10		=C11/B11	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F11/B11)	=IF((F11-3*G11)<0,0,F11-3*G11)	=F11+3*G11	0.01
12	11	1000	20		=C12/B12	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F12/B12)	=IF((F12-3*G12)<0,0,F12-3*G12)	=F12+3*G12	0.01
13	12	1000	10		=C13/B13	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F13/B13)	=IF((F13-3*G13)<0,0,F13-3*G13)	=F13+3*G13	0.01
14	13	1000	20		=C14/B14	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F14/B14)	=IF((F14-3*G14)<0,0,F14-3*G14)	=F14+3*G14	0.01
15	14	1000	10		=C15/B15	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F15/B15)	=IF((F15-3*G15)<0,0,F15-3*G15)	=F15+3*G15	0.01
16	15	1000	20		=C16/B16	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F16/B16)	=IF((F16-3*G16)<0,0,F16-3*G16)	=F16+3*G16	0.01
17	16	1000	30		=C17/B17	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F17/B17)	=IF((F17-3*G17)<0,0,F17-3*G17)	=F17+3*G17	0.01
18	17	1000	30		=C18/B18	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F18/B18)	=IF((F18-3*G18)<0,0,F18-3*G18)	=F18+3*G18	0.01
19	18	1000	10		=C19/B19	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F19/B19)	=IF((F19-3*G19)<0,0,F19-3*G19)	=F19+3*G19	0.01
20	19	1000	20		=C20/B20	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F20/B20)	=IF((F20-3*G20)<0,0,F20-3*G20)	=F20+3*G20	0.01
21	20	1000	0		=C21/B21	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F21/B21)	=IF((F21-3*G21)<0,0,F21-3*G21)	=F21+3*G21	0.01
22	21	1000	30		=C22/B22	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F22/B22)	=IF((F22-3*G22)<0,0,F22-3*G22)	=F22+3*G22	0.01
23	22	1000	20		=C23/B23	=SUM($C$2:$C$23)/SUM($B$2:$B$23)	=SQRT(F23/B23)	=IF((F23-3*G23)<0,0,F23-3*G23)	=F23+3*G23	0.01

and the corresponding chart still with the same spreadsheet software:



with Minitab 17.1.3 this gives for people that don't trust our calculation made with Microsoft® Excel® (we can also do such a chart with R but the plot quality output is not acceptable in most corporations):


U Chart of Non-compliant
U Chart of Non-compliant
Sample Count Per Unit
0.05
0.04
0.03
0.02
0.01
0.00
1
3
5
7
9
11
13
15
17
19
21
Sample
UCL=0.03463
Ū=0.02091
LCL=0.00719
1
1
1

where we took as example for limits imposed by the customer/board committee:

$$USL = 0.010 \quad (16.7.737)$$

This control chart can also be interpreted as any another Shewhart chart using the empirical WECO rules.

16.7.6.3.5 Laney's p' and u' control charts

The p and u control charts as we have seen it have a big problem that resides in the fact that the limits use a standard deviation that is proportional to:

$$\sigma_i \propto \sqrt{\frac{1}{N_i}} \tag{16.7.738}$$

which therefore tends to 0 when the N_i are very by (which typically happens in international call centers where its value is often greater than $5,000$ per day) and then the standard deviation becomes so small that we get almost systematically false alarms (therefore the control chart is a little bit much ... too sensitive). Moreover, their standard deviation assumes that the proportion (probability) is constant over time and this is also unrealistic.

It should be noticed that the problem of too much sensitivity almost never appears in the industry since it is often less than a hundred of items!

An alternative proposed by David Laney in 12002 (holocene calendar) is then to consider in a first time the proportions as continuous quantitative values and then of using a control chart taking first the range by paired values of the consecutive proportions:

$$\bar{R}_P = \frac{1}{k-1}\sum_{i=1}^{k} R_i = \frac{1}{k-1}\sum_{i=1}^{k} |p_i - p_{i-1}| \tag{16.7.739}$$

And as (see section Statistics page 1203):

$$\bar{R} = \sigma_p d_2(n) = \sigma_p d_2(2) = \sigma_p \cdot 1.128 \tag{16.7.740}$$

It comes:

$$\sigma_p = \frac{\bar{R}_p}{1.128} \tag{16.7.741}$$

because comparing only the proportion two-by-two, the Hartley constant $d_2(n)$ must be taken with $n = 2$.

We then have the possibility to take:

$$\boxed{\begin{aligned} \text{UCL}_{p(R)} &= \bar{\hat{p}} + 3\sigma_p \cong \bar{\hat{p}} + 2.66\bar{R} \\ \text{CL}_{p(R)} &= \bar{\hat{p}} \\ \text{LCL}_{p(R)} &= \bar{\hat{p}} - 3\sigma_p \cong \bar{\hat{p}} - 2.66\bar{R} \end{aligned}} \tag{16.7.742}$$

Practice shows that this control chart is already much less sensitive, which is a progress but... however it suffers from a big issue!: the upper and lower limits are constant whereas we would like to take into account that the size of the sample varies (or may vary) for each i. Then we have to find another solution!

A track (a little drawn by the hair) then consists first of all in considering that if the N_i are very large and the proportion (probability) is neither too close to 0 nor too close to 1, then each can be approximated By a Normal law and to be centered reduced:

$$Z_i = \frac{\hat{p}_i - \bar{\hat{p}}}{\sigma_{\hat{p}_i}} \cong Z \tag{16.7.743}$$

and therefore:

$$\hat{p}_i = \bar{\hat{p}} + Z\sigma_{\hat{p}_i} \tag{16.7.744}$$

Then if follows that:

$$\text{V}(\hat{p}_i) = \text{V}\left(\bar{\hat{p}} + Z\hat{\sigma}_{\hat{p}_i}\right) = \text{V}\left(\bar{\hat{p}}\right) + \text{V}(Z\hat{\sigma}_{\hat{p}_i}) = \text{V}(Z\hat{\sigma}_{\hat{p}_i}) = hat\sigma^2_{\hat{p}_i}\text{V}(Z) \tag{16.7.745}$$

and therefore the true variance of $\hat{p}_i$ can be thank to:

$$\sigma_{\hat{p}_i} = \hat{\sigma}_{\hat{p}_i}\sigma_Z \tag{16.7.746}$$

Obviously, in the reality the standard deviation of Z will not really be equal to 0. So the idea (questionable ...) is to consider that as now we have a continuous variable we will be able to estimate the true variance of Z thanks to:

$$\bar{R}_Z = \frac{1}{k-1}\sum_{i=1}^{k} R_i = \frac{1}{k-1}\sum_{i=1}^{k} |Z_i - Z_{i-1}| \tag{16.7.747}$$

and:

$$\bar{R}_Z = \sigma_Z d_2(n) = \sigma_Z d_2(2) = \sigma_Z \cdot 1.128 \tag{16.7.748}$$

Therefore:

$$\sigma_Z = \frac{\bar{R}_Z}{1.128} \tag{16.7.749}$$

We then have:

$$\boxed{\begin{aligned}
\text{UCL}_{Z(p)} &= \bar{\hat{p}} + 3\hat{\sigma}_{\hat{p}_i}\sigma_Z = \bar{\hat{p}} + 3\sqrt{\frac{\bar{\hat{p}}_i(1-\bar{\hat{p}}_i)}{N_i}}\frac{\bar{R}_Z}{1.128} \cong \bar{\hat{p}} + 2.66\sqrt{\frac{\bar{\hat{p}}_i(1-\bar{\hat{p}}_i)}{N_i}}\bar{R}_Z \\
\text{CL}_{Z(p)} &= \bar{\hat{p}} \\
\text{LCL}_{Z(p)} &= \bar{\hat{p}} - 3\hat{\sigma}_{\hat{p}_i}\sigma_Z = \bar{\hat{p}} - 3\sqrt{\frac{\bar{\hat{p}}_i(1-\bar{\hat{p}}_i)}{N_i}}\frac{\bar{R}_Z}{1.128} \cong \bar{\hat{p}} - 2.66\sqrt{\frac{\bar{\hat{p}}_i(1-\bar{\hat{p}}_i)}{N_i}}\bar{R}_Z
\end{aligned}} \tag{16.7.750}$$

where if $\text{LCL}_{Z(p)}$ is smaller than zero then we will assume that this limit is zero (as for the previous control charts).

We thus have a control chart which is an attribute-range mixture which according to the return on experience (REX) is less sensitive (less systematic false alerts) and which has variable limits depending on the size of the subgroups.

The same reasoning is applicable with the U control chart (as the Poisson's law can also be approximated by a Normal law, as we have proved it in the section of Statistics at page 937):

$$\boxed{\begin{aligned}\text{UCL}_{Z(u)}&=\bar{u}+3\hat{\sigma}_u\sigma_Z=\bar{u}+2.66\sqrt{\frac{\bar{u}}{N_i}}\bar{R}_Z\\ \text{CL}_{Z(u)}&=\bar{u}\\ \text{LCL}_{Z(u)}&=\bar{u}-3\hat{\sigma}_u\sigma_Z=\bar{u}-2.66\sqrt{\frac{\bar{u}}{N_i}}\bar{R}_Z\end{aligned}} \tag{16.7.751}$$

We notice that if by chance we have for the both set of relations above:

$$\sigma_Z=\frac{\bar{R}_Z}{1.128}=1 \tag{16.7.752}$$

Then we fall back on the standard parameters of a control chart of type p or respectively of type u.

There are however corrections/variants to these two models ... so it is necessary to know well what your software does! It is for this reason that we will not take screenshots of the procedure with Microsoft® Excel® this time and compare it with Minitab (excepted if requested by our readers).

16.7.6.4 Measurement Control Charts (quantitative CC)

Quantitative variables are continuous measurements of the type: weights, lengths, thicknesses, temperature, diameter ...

Unlike autocorrelated control charts and quality (attributes) control charts , quantitative control charts (with measurement) are often - if not always - presented in pairs (or should be!):

- A control chart showing the analysis / evolution of the central statistical trend (often the mean, because it is relatively easy to construct confidence intervals on it).
- A control chart showing the analysis / evolution of the dispersion via the range or the standard deviation.

The controls limits are often taken (for lack of thinking time in business) as the Shewhart or WECO limits and not probabilistic limits.

We also make the distinction between:

- "Individual control charts" (each point on the control chart involve one and only one measurement)
- "Subgroups control charts" (each point on the control chart involve implicitly more than one item)

16.7.6.4.1 Individual measurement control chart with required limits

This is the easiest control chart and most used one by non specialist because it does not require any special knowledge and statistical hypothesis. It consists only to report the measurements and to take corrective actions when it seems necessary by the rule of thumb.

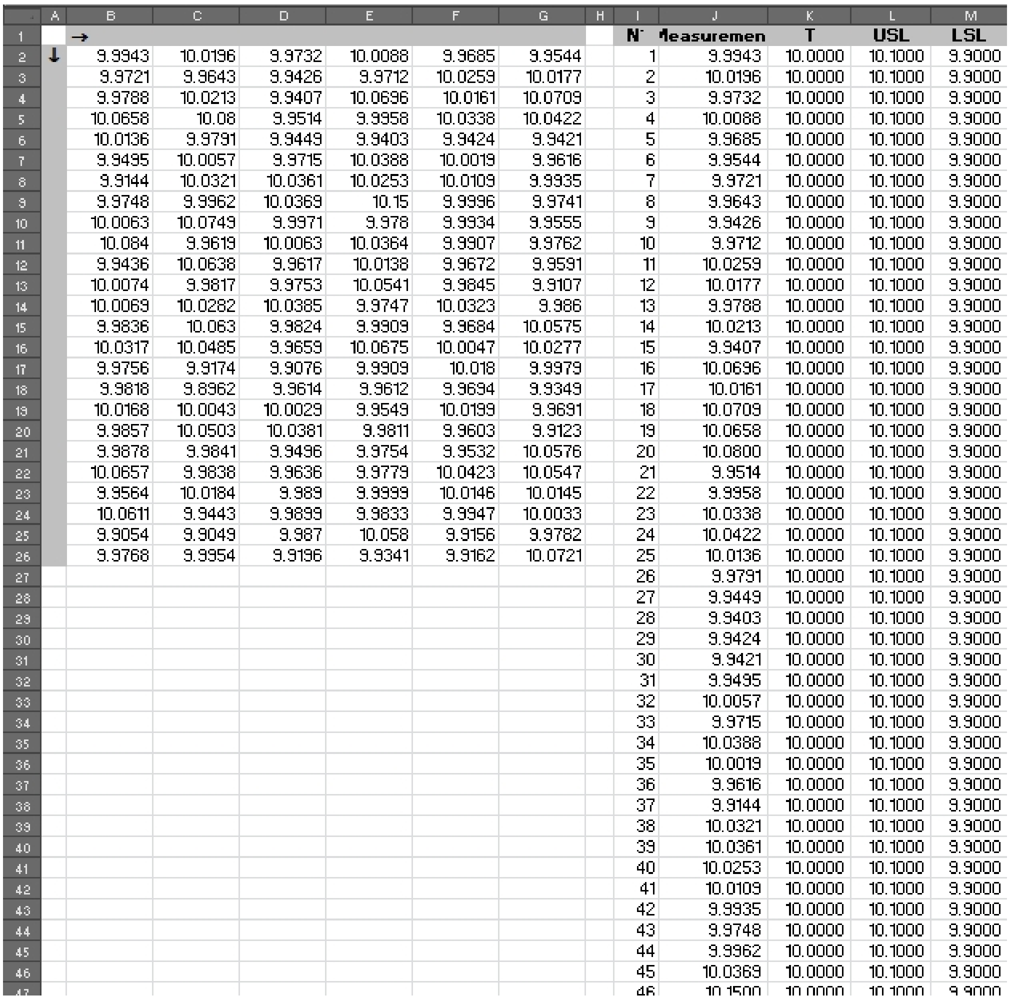
Pros:

- Simple to build because does not require any special knowledge or statistical hypothesis and as all previous control charts, permits the user to check in a timely manner whether a production meets or not the constraints specified by customers.

Cons:

- It may represent temporary random deviations that may make the reader think mistakenly that there is problem in the method / process and will lead to unnecessary corrective actions (false alarms).
- It does not allow to identify if the process / method is under statistical control or not that is to say that as for attribute control charts, it is not possible to make a Capability Six Pack analysis (see page 8708).

For example, let us consider the following list of data still in the same spreadsheet software:

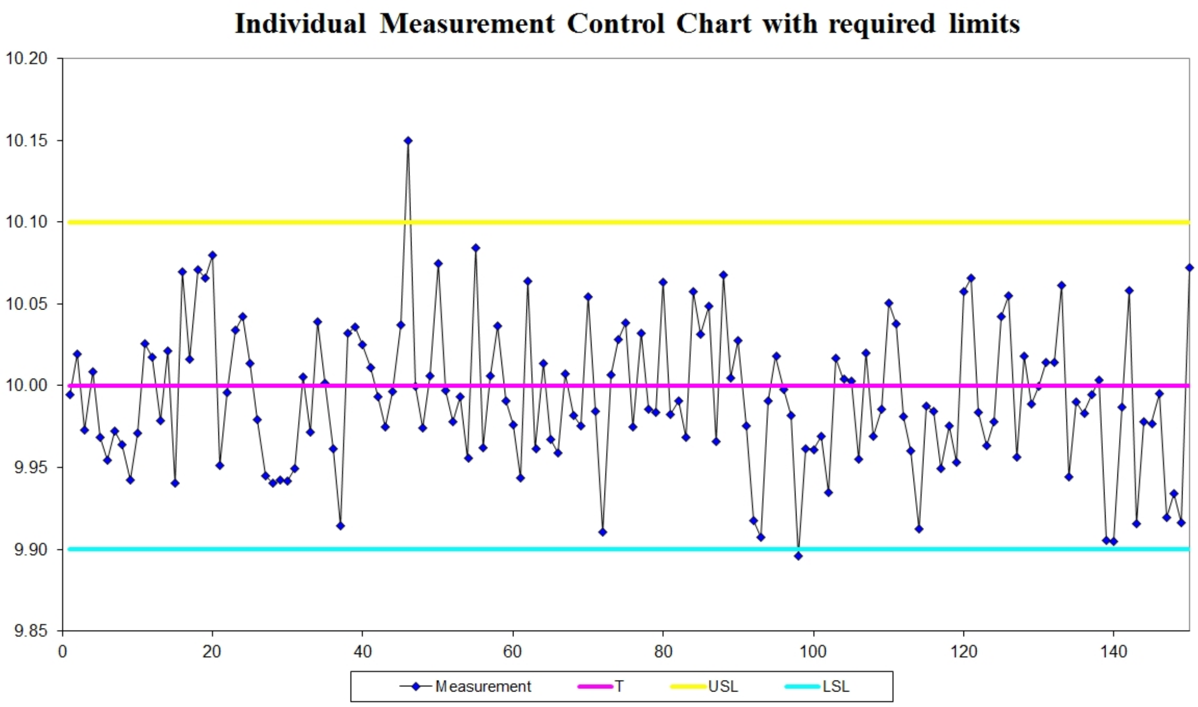


	A	B	C	D	E	F	G	H	I	J	K	L	M
1		→							N°	Measuremen	T	USL	LSL
2	↓	9.9943	10.0196	9.9732	10.0088	9.9685	9.9544		1	9.9943	10.0000	10.1000	9.9000
3		9.9721	9.9643	9.9426	9.9712	10.0259	10.0177		2	10.0196	10.0000	10.1000	9.9000
4		9.9788	10.0213	9.9407	10.0696	10.0161	10.0709		3	9.9732	10.0000	10.1000	9.9000
5		10.0658	10.08	9.9514	9.9958	10.0338	10.0422		4	10.0088	10.0000	10.1000	9.9000
6		10.0136	9.9791	9.9449	9.9403	9.9424	9.9421		5	9.9685	10.0000	10.1000	9.9000
7		9.9495	10.0057	9.9715	10.0388	10.0019	9.9616		6	9.9544	10.0000	10.1000	9.9000
8		9.9144	10.0321	10.0361	10.0253	10.0109	9.9935		7	9.9721	10.0000	10.1000	9.9000
9		9.9748	9.9962	10.0369	10.15	9.9996	9.9741		8	9.9643	10.0000	10.1000	9.9000
10		10.0063	10.0749	9.9971	9.978	9.9934	9.9555		9	9.9426	10.0000	10.1000	9.9000
11		10.084	9.9619	10.0063	10.0364	9.9907	9.9762		10	9.9712	10.0000	10.1000	9.9000
12		9.9436	10.0638	9.9617	10.0138	9.9672	9.9591		11	10.0259	10.0000	10.1000	9.9000
13		10.0074	9.9817	9.9753	10.0541	9.9845	9.9107		12	10.0177	10.0000	10.1000	9.9000
14		10.0069	10.0282	10.0385	9.9747	10.0323	9.986		13	9.9788	10.0000	10.1000	9.9000
15		9.9836	10.063	9.9824	9.9909	9.9684	10.0575		14	10.0213	10.0000	10.1000	9.9000
16		10.0317	10.0485	9.9659	10.0675	10.0047	10.0277		15	9.9407	10.0000	10.1000	9.9000
17		9.9756	9.9174	9.9076	9.9909	10.018	9.9979		16	10.0696	10.0000	10.1000	9.9000
18		9.9818	9.8962	9.9614	9.9612	9.9694	9.9349		17	10.0161	10.0000	10.1000	9.9000
19		10.0168	10.0043	10.0029	9.9549	10.0199	9.9691		18	10.0709	10.0000	10.1000	9.9000
20		9.9857	10.0503	10.0381	9.9811	9.9603	9.9123		19	10.0658	10.0000	10.1000	9.9000
21		9.9878	9.9841	9.9496	9.9754	9.9532	10.0576		20	10.0800	10.0000	10.1000	9.9000
22		10.0657	9.9838	9.9636	9.9779	10.0423	10.0547		21	9.9514	10.0000	10.1000	9.9000
23		9.9564	10.0184	9.989	9.9999	10.0146	10.0145		22	9.9958	10.0000	10.1000	9.9000
24		10.0611	9.9443	9.9899	9.9833	9.9947	10.0033		23	10.0338	10.0000	10.1000	9.9000
25		9.9054	9.9049	9.987	10.058	9.9156	9.9782		24	10.0422	10.0000	10.1000	9.9000
26		9.9768	9.9954	9.9196	9.9341	9.9162	10.0721		25	10.0136	10.0000	10.1000	9.9000
27									26	9.9791	10.0000	10.1000	9.9000
28									27	9.9449	10.0000	10.1000	9.9000
29									28	9.9403	10.0000	10.1000	9.9000
30									29	9.9424	10.0000	10.1000	9.9000
31									30	9.9421	10.0000	10.1000	9.9000
32									31	9.9495	10.0000	10.1000	9.9000
33									32	10.0057	10.0000	10.1000	9.9000
34									33	9.9715	10.0000	10.1000	9.9000
35									34	10.0388	10.0000	10.1000	9.9000
36									35	10.0019	10.0000	10.1000	9.9000
37									36	9.9616	10.0000	10.1000	9.9000
38									37	9.9144	10.0000	10.1000	9.9000
39									38	10.0321	10.0000	10.1000	9.9000
40									39	10.0361	10.0000	10.1000	9.9000
41									40	10.0253	10.0000	10.1000	9.9000
42									41	10.0109	10.0000	10.1000	9.9000
43									42	9.9935	10.0000	10.1000	9.9000
44									43	9.9748	10.0000	10.1000	9.9000
45									44	9.9962	10.0000	10.1000	9.9000
46									45	10.0369	10.0000	10.1000	9.9000
									46	10.1500	10.0000	10.1000	9.9000

where we took empirically (rule of thumb) the limits:

$$\begin{aligned}\text{USL} &= 10+0.1=10.10\\ T &= 10\\ \text{LSL} &= 10-0.1=9.90\end{aligned} \tag{16.7.753}$$

Which gives the following control chart:



While this control chart is the most used in practice (since most people don't have training in S.P.C.), the problem remains that it does not really tell us whether the method or process appears to be stable in time and if the statistical control are in the customer limitations (it shows us only the punctual statistics which is as we know not very robust...!). We will partially remedy with this with the next control chart.

This type of control chart is available in most statistical software but as there is nothing complicated we will not provide here any Minitab (or other...) screenshot.

16.7.6.4.2 Individual measurement control chart with moving limits

When we begin the first observations (measurements), the method or process is rarely stable and there is no way to have fun do manufacture a huge quantity of trial pieces to do a control chart of the average in such conditions and especially if a piece is quite expensive and time consuming!

Pros:

- Highlights the evolution over time of the statistical indicators of positions (average or median) and dispersion and permits to observe the process/methods stabilization during the prototyping phase (often named "Phase I" in the literature).

Cons:

- It may represent temporary random deviations that may make the reader think mistakenly that there is problem in the method / process and will lead to unnecessary corrective actions (false alarms).

- It does not allow to identify statistically if the process / method is under statistical control or not that is to say that as for attribute control charts, it is not possible to make a Capability SixPack analysis.

We stay with a measured control chart to which we add the arithmetic average and the unbiased standard deviation recalculated at each new item. That is, as we don't have enough samples to know the probability distribution and because it is out of question to waste pieces in this purpose and to train people for weeks to statistics, we add to the previous Shewhart control chart the following limits:

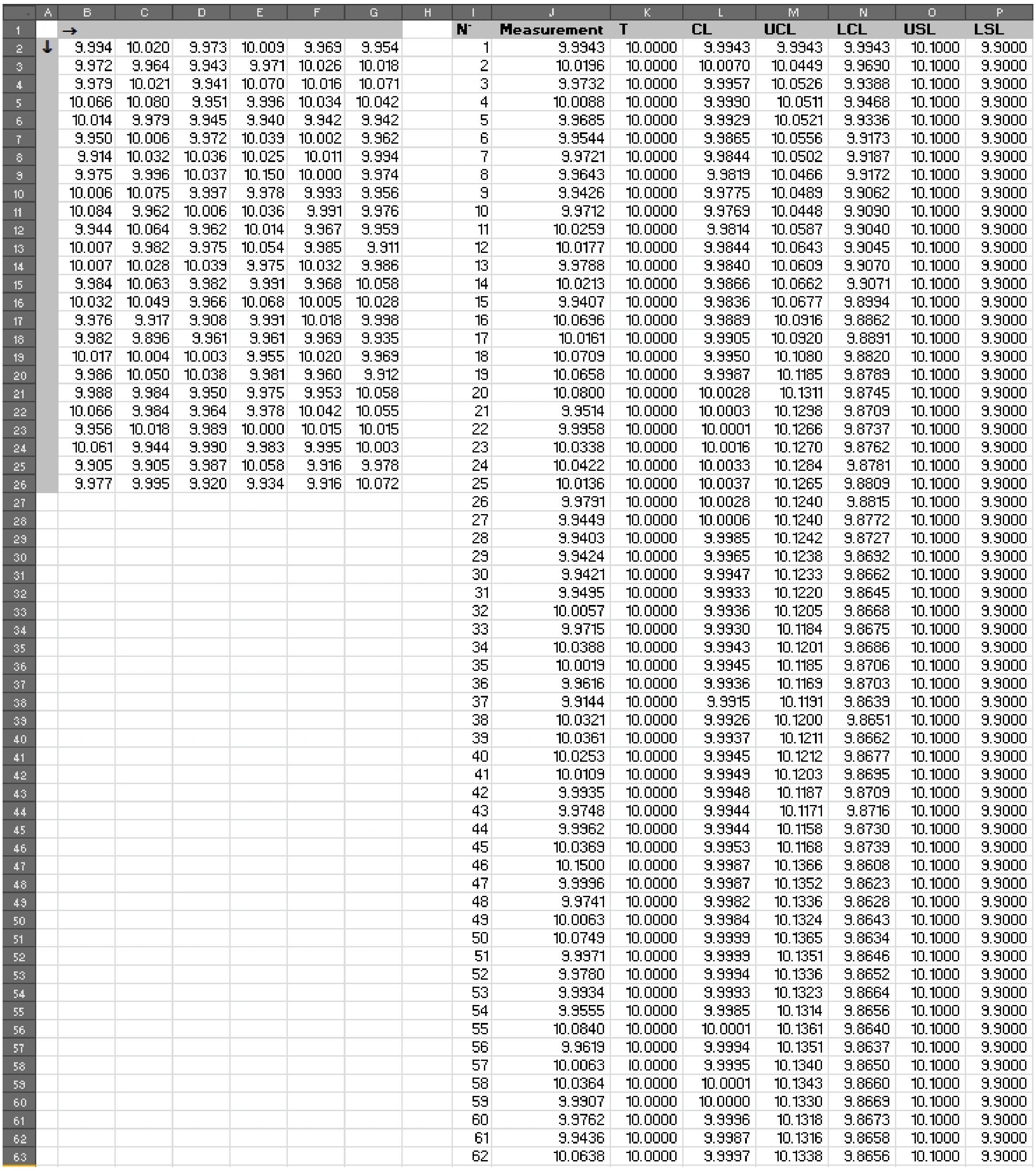
$$\boxed{\begin{aligned}\text{USL}&=\text{required by customer/board comitee}\\ \text{UCL}_{\text{ST}}&=\bar{X}_k+3\sigma_{X,k}\\ \text{CL}_{\text{ST}}&=\text{E}(X)=\bar{X}_k\\ T_{\text{ST}}&=\text{target required by customer/board comitee}\\ \text{LCL}_{\text{ST}}&=\bar{X}_k-3\sigma_{X,k}\\ \text{LSL}&=\text{required by customer/board comitee}\end{aligned}} \tag{16.7.754}$$

where CL means "centerline".

Obviously for the estimator of the mean and the standard deviation, we take:

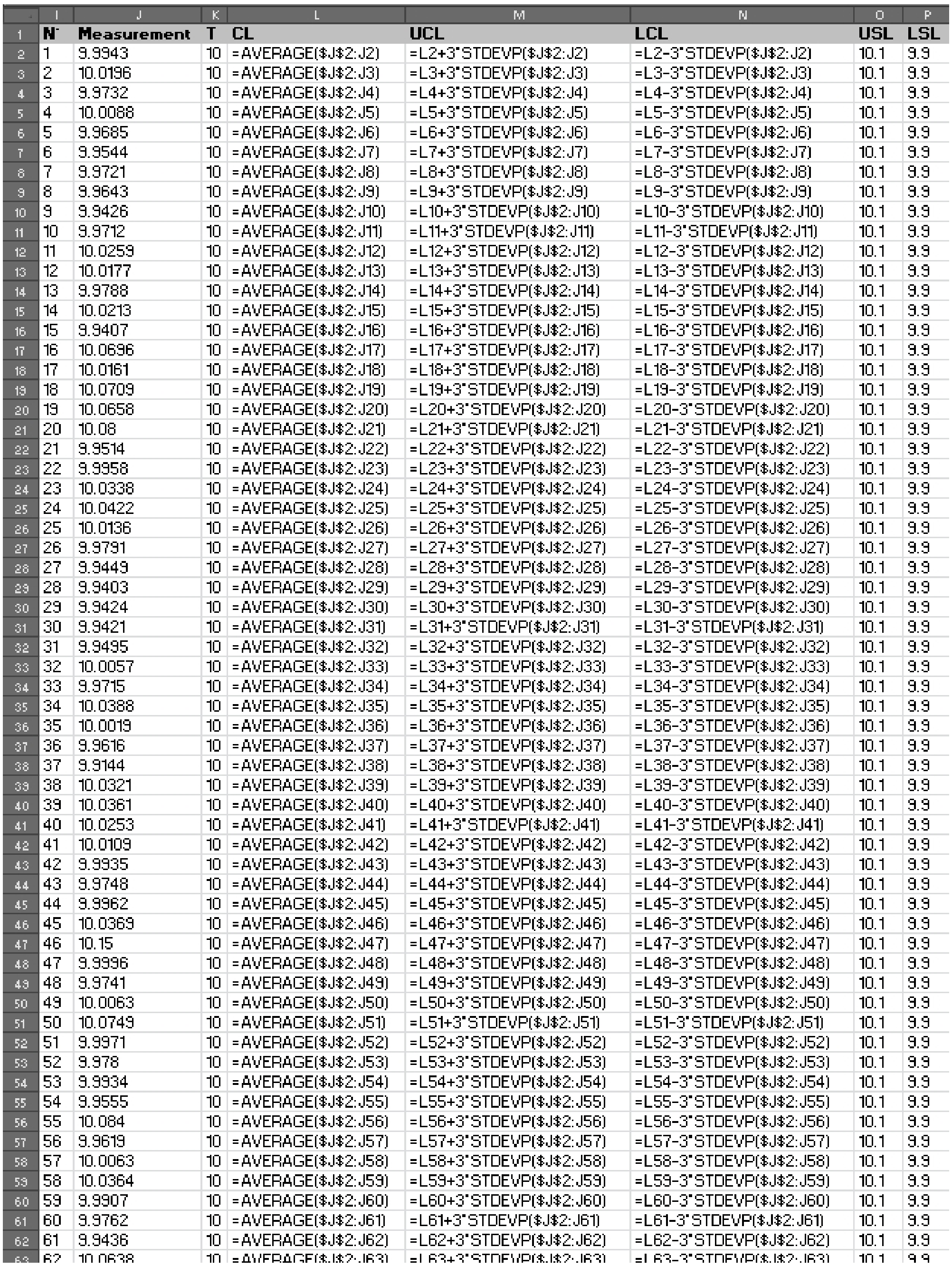
$$\begin{aligned}\text{E}(X)=\bar{X}_k&=\frac{1}{k}\sum_{i=1}^{k}X_i\\ \sigma_{X,k}&=\sqrt{\frac{1}{k-1}\sum_{i=1}^{k}(X_i-\bar{X})^2}\end{aligned} \tag{16.7.755}$$

Once again using our spreadsheet software here is an example where $N=150$ and therefore where k moves from 1 to 150 with the 63 first rows visible:



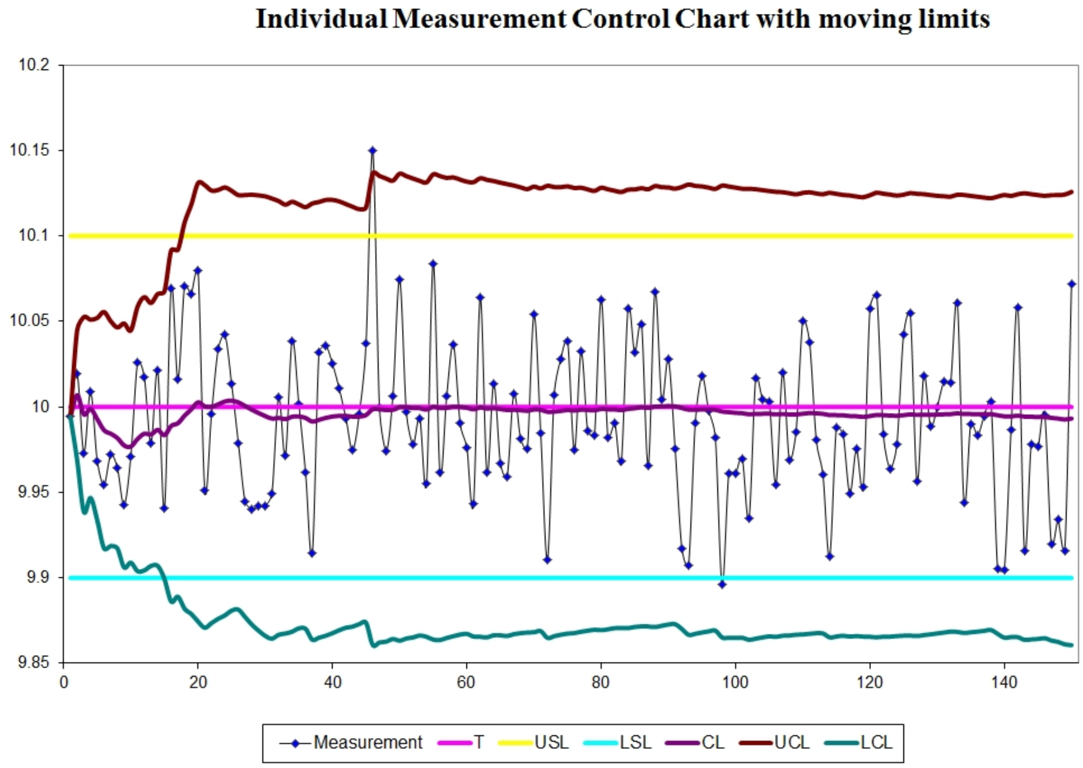
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1		→							N°	**Measurement**	**T**	**CL**	**UCL**	**LCL**	**USL**	**LSL**
2	↓	9.994	10.020	9.973	10.009	9.969	9.954		1	9.9943	10.0000	9.9943	9.9943	9.9943	10.1000	9.9000
3		9.972	9.964	9.943	9.971	10.026	10.018		2	10.0196	10.0000	10.0070	10.0449	9.9690	10.1000	9.9000
4		9.979	10.021	9.941	10.070	10.016	10.071		3	9.9732	10.0000	9.9957	10.0526	9.9388	10.1000	9.9000
5		10.066	10.080	9.951	9.996	10.034	10.042		4	10.0088	10.0000	9.9990	10.0511	9.9468	10.1000	9.9000
6		10.014	9.979	9.945	9.940	9.942	9.942		5	9.9685	10.0000	9.9929	10.0521	9.9336	10.1000	9.9000
7		9.950	10.006	9.972	10.039	10.002	9.962		6	9.9544	10.0000	9.9865	10.0556	9.9173	10.1000	9.9000
8		9.914	10.032	10.036	10.025	10.011	9.994		7	9.9721	10.0000	9.9844	10.0502	9.9187	10.1000	9.9000
9		9.975	9.996	10.037	10.150	10.000	9.974		8	9.9643	10.0000	9.9819	10.0466	9.9172	10.1000	9.9000
10		10.006	10.075	9.997	9.978	9.993	9.956		9	9.9426	10.0000	9.9775	10.0489	9.9062	10.1000	9.9000
11		10.084	9.962	10.006	10.036	9.991	9.976		10	9.9712	10.0000	9.9769	10.0448	9.9090	10.1000	9.9000
12		9.944	10.064	9.962	10.014	9.967	9.959		11	10.0259	10.0000	9.9814	10.0587	9.9040	10.1000	9.9000
13		10.007	9.982	9.975	10.054	9.985	9.911		12	10.0177	10.0000	9.9844	10.0643	9.9045	10.1000	9.9000
14		10.007	10.028	10.039	9.975	10.032	9.986		13	9.9788	10.0000	9.9840	10.0609	9.9070	10.1000	9.9000
15		9.984	10.063	9.982	9.991	9.968	10.058		14	10.0213	10.0000	9.9866	10.0662	9.9071	10.1000	9.9000
16		10.032	10.049	9.966	10.068	10.005	10.028		15	9.9407	10.0000	9.9836	10.0677	9.8994	10.1000	9.9000
17		9.976	9.917	9.908	9.991	10.018	9.998		16	10.0696	10.0000	9.9889	10.0916	9.8862	10.1000	9.9000
18		9.982	9.896	9.961	9.961	9.969	9.935		17	10.0161	10.0000	9.9905	10.0920	9.8891	10.1000	9.9000
19		10.017	10.004	10.003	9.955	10.020	9.969		18	10.0709	10.0000	9.9950	10.1080	9.8820	10.1000	9.9000
20		9.986	10.050	10.038	9.981	9.960	9.912		19	10.0658	10.0000	9.9987	10.1185	9.8789	10.1000	9.9000
21		9.988	9.984	9.950	9.975	9.953	10.058		20	10.0800	10.0000	10.0028	10.1311	9.8745	10.1000	9.9000
22		10.066	9.984	9.964	9.978	10.042	10.055		21	9.9514	10.0000	10.0003	10.1298	9.8709	10.1000	9.9000
23		9.956	10.018	9.989	10.000	10.015	10.015		22	9.9958	10.0000	10.0001	10.1266	9.8737	10.1000	9.9000
24		10.061	9.944	9.990	9.983	9.995	10.003		23	10.0338	10.0000	10.0016	10.1270	9.8762	10.1000	9.9000
25		9.905	9.905	9.987	10.058	9.916	9.978		24	10.0422	10.0000	10.0033	10.1284	9.8781	10.1000	9.9000
26		9.977	9.995	9.920	9.934	9.916	10.072		25	10.0136	10.0000	10.0037	10.1265	9.8809	10.1000	9.9000
27									26	9.9791	10.0000	10.0028	10.1240	9.8815	10.1000	9.9000
28									27	9.9449	10.0000	10.0006	10.1240	9.8772	10.1000	9.9000
29									28	9.9403	10.0000	9.9985	10.1242	9.8727	10.1000	9.9000
30									29	9.9424	10.0000	9.9965	10.1238	9.8692	10.1000	9.9000
31									30	9.9421	10.0000	9.9947	10.1233	9.8662	10.1000	9.9000
32									31	9.9495	10.0000	9.9933	10.1220	9.8645	10.1000	9.9000
33									32	10.0057	10.0000	9.9936	10.1205	9.8668	10.1000	9.9000
34									33	9.9715	10.0000	9.9930	10.1184	9.8675	10.1000	9.9000
35									34	10.0388	10.0000	9.9943	10.1201	9.8686	10.1000	9.9000
36									35	10.0019	10.0000	9.9945	10.1185	9.8706	10.1000	9.9000
37									36	9.9616	10.0000	9.9936	10.1169	9.8703	10.1000	9.9000
38									37	9.9144	10.0000	9.9915	10.1191	9.8639	10.1000	9.9000
39									38	10.0321	10.0000	9.9926	10.1200	9.8651	10.1000	9.9000
40									39	10.0361	10.0000	9.9937	10.1211	9.8662	10.1000	9.9000
41									40	10.0253	10.0000	9.9945	10.1212	9.8677	10.1000	9.9000
42									41	10.0109	10.0000	9.9949	10.1203	9.8695	10.1000	9.9000
43									42	9.9935	10.0000	9.9948	10.1187	9.8709	10.1000	9.9000
44									43	9.9748	10.0000	9.9944	10.1171	9.8716	10.1000	9.9000
45									44	9.9962	10.0000	9.9944	10.1158	9.8730	10.1000	9.9000
46									45	10.0369	10.0000	9.9953	10.1168	9.8739	10.1000	9.9000
47									46	10.1500	10.0000	9.9987	10.1366	9.8608	10.1000	9.9000
48									47	9.9996	10.0000	9.9987	10.1352	9.8623	10.1000	9.9000
49									48	9.9741	10.0000	9.9982	10.1336	9.8628	10.1000	9.9000
50									49	10.0063	10.0000	9.9984	10.1324	9.8643	10.1000	9.9000
51									50	10.0749	10.0000	9.9999	10.1365	9.8634	10.1000	9.9000
52									51	9.9971	10.0000	9.9999	10.1351	9.8646	10.1000	9.9000
53									52	9.9780	10.0000	9.9994	10.1336	9.8652	10.1000	9.9000
54									53	9.9934	10.0000	9.9993	10.1323	9.8664	10.1000	9.9000
55									54	9.9555	10.0000	9.9985	10.1314	9.8656	10.1000	9.9000
56									55	10.0840	10.0000	10.0001	10.1361	9.8640	10.1000	9.9000
57									56	9.9619	10.0000	9.9994	10.1351	9.8637	10.1000	9.9000
58									57	10.0063	10.0000	9.9995	10.1340	9.8650	10.1000	9.9000
59									58	10.0364	10.0000	10.0001	10.1343	9.8660	10.1000	9.9000
60									59	9.9907	10.0000	10.0000	10.1330	9.8669	10.1000	9.9000
61									60	9.9762	10.0000	9.9996	10.1318	9.8673	10.1000	9.9000
62									61	9.9436	10.0000	9.9987	10.1316	9.8658	10.1000	9.9000
63									62	10.0638	10.0000	9.9997	10.1338	9.8656	10.1000	9.9000

or explicitly:



	I	J	K	L	M	N	O	P
1	**N°**	**Measurement**	**T**	**CL**	**UCL**	**LCL**	**USL**	**LSL**
2	1	9.9943	10	=AVERAGE($J$2:J2)	=L2+3*STDEVP($J$2:J2)	=L2-3*STDEVP($J$2:J2)	10.1	9.9
3	2	10.0196	10	=AVERAGE($J$2:J3)	=L3+3*STDEVP($J$2:J3)	=L3-3*STDEVP($J$2:J3)	10.1	9.9
4	3	9.9732	10	=AVERAGE($J$2:J4)	=L4+3*STDEVP($J$2:J4)	=L4-3*STDEVP($J$2:J4)	10.1	9.9
5	4	10.0088	10	=AVERAGE($J$2:J5)	=L5+3*STDEVP($J$2:J5)	=L5-3*STDEVP($J$2:J5)	10.1	9.9
6	5	9.9685	10	=AVERAGE($J$2:J6)	=L6+3*STDEVP($J$2:J6)	=L6-3*STDEVP($J$2:J6)	10.1	9.9
7	6	9.9544	10	=AVERAGE($J$2:J7)	=L7+3*STDEVP($J$2:J7)	=L7-3*STDEVP($J$2:J7)	10.1	9.9
8	7	9.9721	10	=AVERAGE($J$2:J8)	=L8+3*STDEVP($J$2:J8)	=L8-3*STDEVP($J$2:J8)	10.1	9.9
9	8	9.9643	10	=AVERAGE($J$2:J9)	=L9+3*STDEVP($J$2:J9)	=L9-3*STDEVP($J$2:J9)	10.1	9.9
10	9	9.9426	10	=AVERAGE($J$2:J10)	=L10+3*STDEVP($J$2:J10)	=L10-3*STDEVP($J$2:J10)	10.1	9.9
11	10	9.9712	10	=AVERAGE($J$2:J11)	=L11+3*STDEVP($J$2:J11)	=L11-3*STDEVP($J$2:J11)	10.1	9.9
12	11	10.0259	10	=AVERAGE($J$2:J12)	=L12+3*STDEVP($J$2:J12)	=L12-3*STDEVP($J$2:J12)	10.1	9.9
13	12	10.0177	10	=AVERAGE($J$2:J13)	=L13+3*STDEVP($J$2:J13)	=L13-3*STDEVP($J$2:J13)	10.1	9.9
14	13	9.9788	10	=AVERAGE($J$2:J14)	=L14+3*STDEVP($J$2:J14)	=L14-3*STDEVP($J$2:J14)	10.1	9.9
15	14	10.0213	10	=AVERAGE($J$2:J15)	=L15+3*STDEVP($J$2:J15)	=L15-3*STDEVP($J$2:J15)	10.1	9.9
16	15	9.9407	10	=AVERAGE($J$2:J16)	=L16+3*STDEVP($J$2:J16)	=L16-3*STDEVP($J$2:J16)	10.1	9.9
17	16	10.0696	10	=AVERAGE($J$2:J17)	=L17+3*STDEVP($J$2:J17)	=L17-3*STDEVP($J$2:J17)	10.1	9.9
18	17	10.0161	10	=AVERAGE($J$2:J18)	=L18+3*STDEVP($J$2:J18)	=L18-3*STDEVP($J$2:J18)	10.1	9.9
19	18	10.0709	10	=AVERAGE($J$2:J19)	=L19+3*STDEVP($J$2:J19)	=L19-3*STDEVP($J$2:J19)	10.1	9.9
20	19	10.0658	10	=AVERAGE($J$2:J20)	=L20+3*STDEVP($J$2:J20)	=L20-3*STDEVP($J$2:J20)	10.1	9.9
21	20	10.08	10	=AVERAGE($J$2:J21)	=L21+3*STDEVP($J$2:J21)	=L21-3*STDEVP($J$2:J21)	10.1	9.9
22	21	9.9514	10	=AVERAGE($J$2:J22)	=L22+3*STDEVP($J$2:J22)	=L22-3*STDEVP($J$2:J22)	10.1	9.9
23	22	9.9958	10	=AVERAGE($J$2:J23)	=L23+3*STDEVP($J$2:J23)	=L23-3*STDEVP($J$2:J23)	10.1	9.9
24	23	10.0338	10	=AVERAGE($J$2:J24)	=L24+3*STDEVP($J$2:J24)	=L24-3*STDEVP($J$2:J24)	10.1	9.9
25	24	10.0422	10	=AVERAGE($J$2:J25)	=L25+3*STDEVP($J$2:J25)	=L25-3*STDEVP($J$2:J25)	10.1	9.9
26	25	10.0136	10	=AVERAGE($J$2:J26)	=L26+3*STDEVP($J$2:J26)	=L26-3*STDEVP($J$2:J26)	10.1	9.9
27	26	9.9791	10	=AVERAGE($J$2:J27)	=L27+3*STDEVP($J$2:J27)	=L27-3*STDEVP($J$2:J27)	10.1	9.9
28	27	9.9449	10	=AVERAGE($J$2:J28)	=L28+3*STDEVP($J$2:J28)	=L28-3*STDEVP($J$2:J28)	10.1	9.9
29	28	9.9403	10	=AVERAGE($J$2:J29)	=L29+3*STDEVP($J$2:J29)	=L29-3*STDEVP($J$2:J29)	10.1	9.9
30	29	9.9424	10	=AVERAGE($J$2:J30)	=L30+3*STDEVP($J$2:J30)	=L30-3*STDEVP($J$2:J30)	10.1	9.9
31	30	9.9421	10	=AVERAGE($J$2:J31)	=L31+3*STDEVP($J$2:J31)	=L31-3*STDEVP($J$2:J31)	10.1	9.9
32	31	9.9495	10	=AVERAGE($J$2:J32)	=L32+3*STDEVP($J$2:J32)	=L32-3*STDEVP($J$2:J32)	10.1	9.9
33	32	10.0057	10	=AVERAGE($J$2:J33)	=L33+3*STDEVP($J$2:J33)	=L33-3*STDEVP($J$2:J33)	10.1	9.9
34	33	9.9715	10	=AVERAGE($J$2:J34)	=L34+3*STDEVP($J$2:J34)	=L34-3*STDEVP($J$2:J34)	10.1	9.9
35	34	10.0388	10	=AVERAGE($J$2:J35)	=L35+3*STDEVP($J$2:J35)	=L35-3*STDEVP($J$2:J35)	10.1	9.9
36	35	10.0019	10	=AVERAGE($J$2:J36)	=L36+3*STDEVP($J$2:J36)	=L36-3*STDEVP($J$2:J36)	10.1	9.9
37	36	9.9616	10	=AVERAGE($J$2:J37)	=L37+3*STDEVP($J$2:J37)	=L37-3*STDEVP($J$2:J37)	10.1	9.9
38	37	9.9144	10	=AVERAGE($J$2:J38)	=L38+3*STDEVP($J$2:J38)	=L38-3*STDEVP($J$2:J38)	10.1	9.9
39	38	10.0321	10	=AVERAGE($J$2:J39)	=L39+3*STDEVP($J$2:J39)	=L39-3*STDEVP($J$2:J39)	10.1	9.9
40	39	10.0361	10	=AVERAGE($J$2:J40)	=L40+3*STDEVP($J$2:J40)	=L40-3*STDEVP($J$2:J40)	10.1	9.9
41	40	10.0253	10	=AVERAGE($J$2:J41)	=L41+3*STDEVP($J$2:J41)	=L41-3*STDEVP($J$2:J41)	10.1	9.9
42	41	10.0109	10	=AVERAGE($J$2:J42)	=L42+3*STDEVP($J$2:J42)	=L42-3*STDEVP($J$2:J42)	10.1	9.9
43	42	9.9935	10	=AVERAGE($J$2:J43)	=L43+3*STDEVP($J$2:J43)	=L43-3*STDEVP($J$2:J43)	10.1	9.9
44	43	9.9748	10	=AVERAGE($J$2:J44)	=L44+3*STDEVP($J$2:J44)	=L44-3*STDEVP($J$2:J44)	10.1	9.9
45	44	9.9962	10	=AVERAGE($J$2:J45)	=L45+3*STDEVP($J$2:J45)	=L45-3*STDEVP($J$2:J45)	10.1	9.9
46	45	10.0369	10	=AVERAGE($J$2:J46)	=L46+3*STDEVP($J$2:J46)	=L46-3*STDEVP($J$2:J46)	10.1	9.9
47	46	10.15	10	=AVERAGE($J$2:J47)	=L47+3*STDEVP($J$2:J47)	=L47-3*STDEVP($J$2:J47)	10.1	9.9
48	47	9.9996	10	=AVERAGE($J$2:J48)	=L48+3*STDEVP($J$2:J48)	=L48-3*STDEVP($J$2:J48)	10.1	9.9
49	48	9.9741	10	=AVERAGE($J$2:J49)	=L49+3*STDEVP($J$2:J49)	=L49-3*STDEVP($J$2:J49)	10.1	9.9
50	49	10.0063	10	=AVERAGE($J$2:J50)	=L50+3*STDEVP($J$2:J50)	=L50-3*STDEVP($J$2:J50)	10.1	9.9
51	50	10.0749	10	=AVERAGE($J$2:J51)	=L51+3*STDEVP($J$2:J51)	=L51-3*STDEVP($J$2:J51)	10.1	9.9
52	51	9.9971	10	=AVERAGE($J$2:J52)	=L52+3*STDEVP($J$2:J52)	=L52-3*STDEVP($J$2:J52)	10.1	9.9
53	52	9.978	10	=AVERAGE($J$2:J53)	=L53+3*STDEVP($J$2:J53)	=L53-3*STDEVP($J$2:J53)	10.1	9.9
54	53	9.9934	10	=AVERAGE($J$2:J54)	=L54+3*STDEVP($J$2:J54)	=L54-3*STDEVP($J$2:J54)	10.1	9.9
55	54	9.9555	10	=AVERAGE($J$2:J55)	=L55+3*STDEVP($J$2:J55)	=L55-3*STDEVP($J$2:J55)	10.1	9.9
56	55	10.084	10	=AVERAGE($J$2:J56)	=L56+3*STDEVP($J$2:J56)	=L56-3*STDEVP($J$2:J56)	10.1	9.9
57	56	9.9619	10	=AVERAGE($J$2:J57)	=L57+3*STDEVP($J$2:J57)	=L57-3*STDEVP($J$2:J57)	10.1	9.9
58	57	10.0063	10	=AVERAGE($J$2:J58)	=L58+3*STDEVP($J$2:J58)	=L58-3*STDEVP($J$2:J58)	10.1	9.9
59	58	10.0364	10	=AVERAGE($J$2:J59)	=L59+3*STDEVP($J$2:J59)	=L59-3*STDEVP($J$2:J59)	10.1	9.9
60	59	9.9907	10	=AVERAGE($J$2:J60)	=L60+3*STDEVP($J$2:J60)	=L60-3*STDEVP($J$2:J60)	10.1	9.9
61	60	9.9762	10	=AVERAGE($J$2:J61)	=L61+3*STDEVP($J$2:J61)	=L61-3*STDEVP($J$2:J61)	10.1	9.9
62	61	9.9436	10	=AVERAGE($J$2:J62)	=L62+3*STDEVP($J$2:J62)	=L62-3*STDEVP($J$2:J62)	10.1	9.9
63	62	10.0638	10	=AVERAGE($J$2:J63)	=L63+3*STDEVP($J$2:J63)	=L63-3*STDEVP($J$2:J63)	10.1	9.9

and the corresponding chart still with the same spreadsheet software:



where we took as example for limits imposed by the customer/board committee:

$$\begin{aligned}\text{USL}&=10.10\\T&=10\\\text{LSL}&=9.90\end{aligned}\tag{16.7.756}$$

We see for example the in the above control chart that the CL indicator (so the estimator of the mean) stabilizes close to the target T from the beginning to the end. It's the same for UCL and LCL (because the estimator of the standard deviation stabilizes) by cons following the usage in engineering the results are not good, since the UCL and LCL are beyond the required specifications by the customer! Note that this control chart would according to the WECO rules have only one single point beyond the $\pm 3\sigma$ (we indicate that for comparison reasons with the next control charts).

Remark

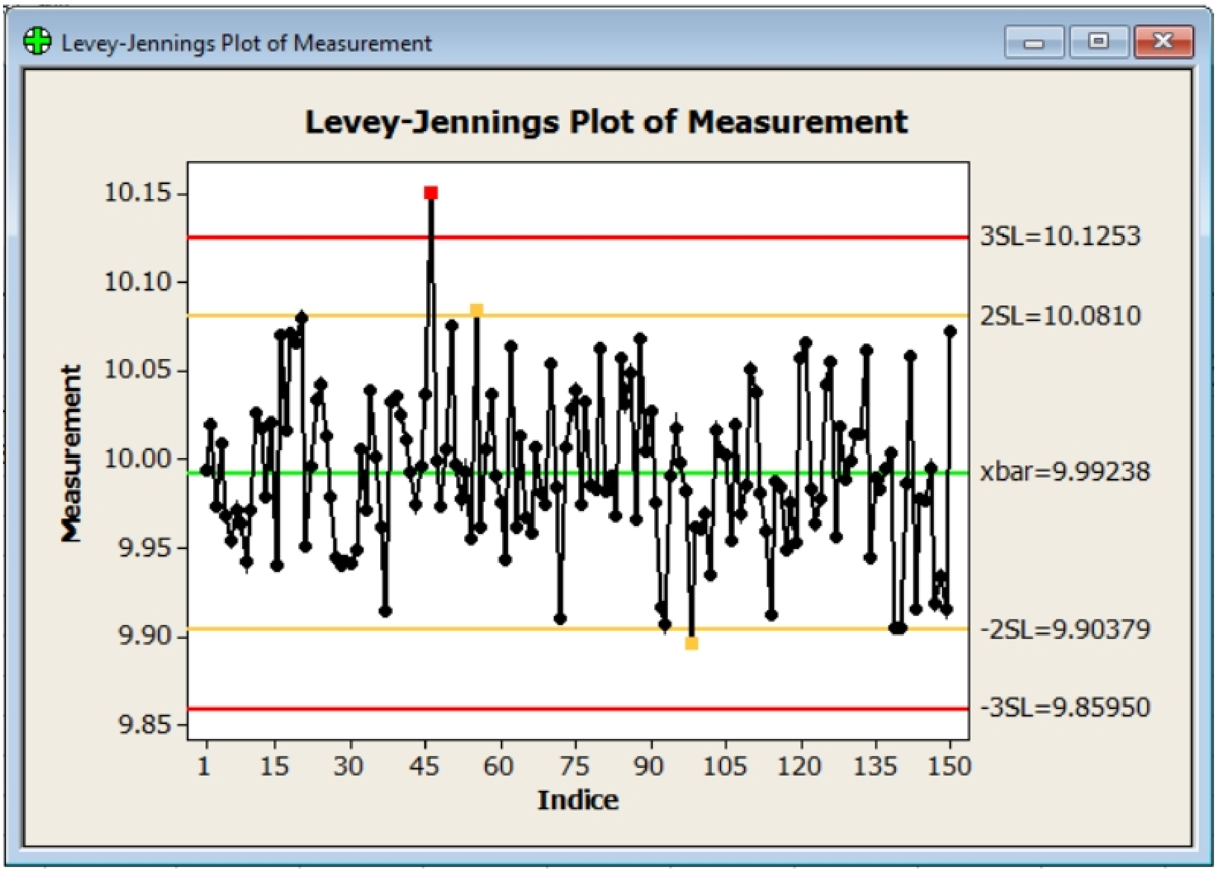
On all control charts with no exceptions, the customer specifications (limits) must be indicated in addition to the calculated limits!

This type of control chart is as far as we know not available in statistical softwares. But there is an easy one with the following very common and obvious fixed limits:

$$\begin{aligned}\text{USL}&=\text{required by customer/board comitee}\\ \text{UCL}_{\text{ST}}&=\bar{X}_k+3\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i-\bar{X})^2}\\ \text{CL}_{\text{ST}}&=\text{E}(X)=\bar{X}\\ T_{\text{ST}}&=\text{target required by customer/board comitee}\\ \text{LCL}_{\text{ST}}&=\bar{X}_k-3\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i-\bar{X})^2}\\ \text{LSL}&=\text{required by customer/board comitee}\end{aligned} \quad (16.7.757)$$

named a "Levey-Jennings control chart".

With Minitab 15.1.1.0 this gives for people that don't trust simple Microsoft® Excel® calculations (we can also do such a chart with R but the plot quality output is not acceptable in most corporations):


Levey-Jennings Plot of Measurement
Levey-Jennings Plot of Measurement
Measurement
10.15
10.10
10.05
10.00
9.95
9.90
9.85
3SL=10.1253
2SL=10.0810
xbar=9.99238
-2SL=9.90379
-3SL=9.85950
1
15
30
45
60
75
90
105
120
135
150
Indice

At this level, we supposed that we are unable to determine the probability distribution of the data (sample problem or statistical skills problems) and therefore unable to calculate the confidence intervals for the mean or the standard deviation!

The next control charts is supposed to remedy to this situation under some assumption!

16.7.6.4.3 Subgroups measurement control chart with standard error

Once the process or method stabilized, we consider that the mean and standard deviation are stabilized and even if the probability distribution that follows a group of k samples is not the same from one day to another, by the central limit theorem, the set of averages follows a Normal distribution as we know of the type:

$$\mathcal{N}=(\overline{X},\sigma_{\overline{X}})=\left(\overline{X},\frac{\sigma_X}{\sqrt{n}}\right) \tag{16.7.758}$$

and written in S.P.C. as (for the reason we will see further below):

$$\mathcal{N}=\left(\overline{\overline{X}},\frac{\sigma_X}{\sqrt{n}}\right) \tag{16.7.759}$$

Therefore it may be more interesting to represent the change in the average over time and not just the single measurement. This also gives us the opportunity to do a capability analysis of our process or method.

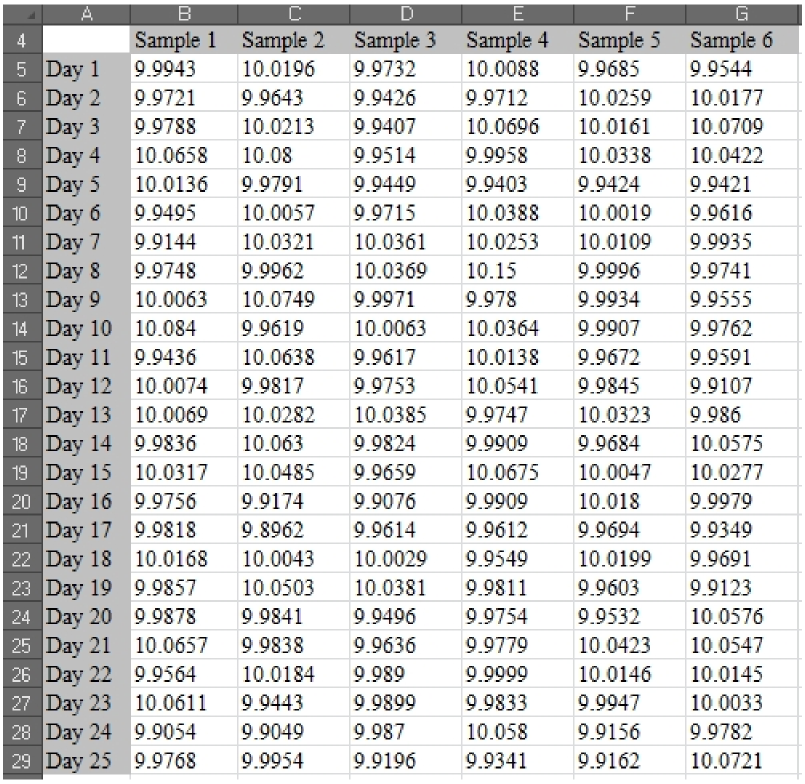
Pros:

- Hides (smooth) non-systematic and random variations and often allows to avoid unnecessary corrective actions (false alarms).
- Gives the possibility to associate the control chart with a capability analysis.

Cons:

- This control chart assumes that the samples are identically distributed by the mathematical assumptions it is build on. Moreover, in the framework of fixed Shewhart limits, the distribution law is assumed symmetrical.
- The major issue of this control chart is that the number of individuals per sample must be large enough for the unbiased estimator of the standard deviation is not too far from reality. This is why we do not found this control chart in statistical software (this is why this control chart is replaced with another chart that we will see a little further below).

As companion example let us consider the following example of 25 samples (that we will identify by the variable k) of 6 individuals each (that we will identify by the variable $n_i=n$) taken from a continuous manufacturing process (without machine recalibration or without machine change!):



	A	B	C	D	E	F	G
4		Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
5	Day 1	9.9943	10.0196	9.9732	10.0088	9.9685	9.9544
6	Day 2	9.9721	9.9643	9.9426	9.9712	10.0259	10.0177
7	Day 3	9.9788	10.0213	9.9407	10.0696	10.0161	10.0709
8	Day 4	10.0658	10.08	9.9514	9.9958	10.0338	10.0422
9	Day 5	10.0136	9.9791	9.9449	9.9403	9.9424	9.9421
10	Day 6	9.9495	10.0057	9.9715	10.0388	10.0019	9.9616
11	Day 7	9.9144	10.0321	10.0361	10.0253	10.0109	9.9935
12	Day 8	9.9748	9.9962	10.0369	10.15	9.9996	9.9741
13	Day 9	10.0063	10.0749	9.9971	9.978	9.9934	9.9555
14	Day 10	10.084	9.9619	10.0063	10.0364	9.9907	9.9762
15	Day 11	9.9436	10.0638	9.9617	10.0138	9.9672	9.9591
16	Day 12	10.0074	9.9817	9.9753	10.0541	9.9845	9.9107
17	Day 13	10.0069	10.0282	10.0385	9.9747	10.0323	9.986
18	Day 14	9.9836	10.063	9.9824	9.9909	9.9684	10.0575
19	Day 15	10.0317	10.0485	9.9659	10.0675	10.0047	10.0277
20	Day 16	9.9756	9.9174	9.9076	9.9909	10.018	9.9979
21	Day 17	9.9818	9.8962	9.9614	9.9612	9.9694	9.9349
22	Day 18	10.0168	10.0043	10.0029	9.9549	10.0199	9.9691
23	Day 19	9.9857	10.0503	10.0381	9.9811	9.9603	9.9123
24	Day 20	9.9878	9.9841	9.9496	9.9754	9.9532	10.0576
25	Day 21	10.0657	9.9838	9.9636	9.9779	10.0423	10.0547
26	Day 22	9.9564	10.0184	9.989	9.9999	10.0146	10.0145
27	Day 23	10.0611	9.9443	9.9899	9.9833	9.9947	10.0033
28	Day 24	9.9054	9.9049	9.987	10.058	9.9156	9.9782
29	Day 25	9.9768	9.9954	9.9196	9.9341	9.9162	10.0721

where every day we took a sample of 6 individuals of all manufactured items produced the same day. On the control chart, then we will represent each point as the average of these 6 individuals.

We have proved in the section Statistics (see page 611) that the standard deviation of the mean (standard error) of a series of n independent and identically distributed random variables were if the samples are big enough:

$$\sigma_{\overline{X}}=\frac{\sigma_X}{\sqrt{n}} \tag{16.7.760}$$

and using the finite population correction factor if the sampled population is not infinite (remember that N is the population size):

$$\sigma_{\overline{X}}=\frac{\sigma_X}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \tag{16.7.761}$$

We then have the following Shewhart control limit indicators (obviously as we use the average estimators and standard deviation, the greater the number of samples and individuals, the greater the limits are asymptotically accurate but also... known problem... if the samples n_i are too big we will have systematic alarms as we know!):

$$\boxed{\begin{aligned}\text{USL}&=\text{required by customer/board comitee}\\ \text{UCL}_{\text{SE}}&=\overline{\overline{X}}+3\sigma_{\overline{X}}=\overline{\overline{X}}+3\frac{\sigma_X}{\sqrt{n_i}}\\ \text{CL}_{\text{SE}}&=\text{E}(\overline{X})=\overline{\overline{X}}\\ T_{\text{SE}}&=\text{target required by customer/board comitee}\\ \text{LCL}_{\text{SE}}&=\overline{\overline{X}}-3\sigma_{\overline{X}}=\overline{\overline{X}}-3\frac{\sigma_X}{\sqrt{n_i}}\\ \text{LSL}&=\text{required by customer/board comitee}\end{aligned}} \tag{16.7.762}$$

With:

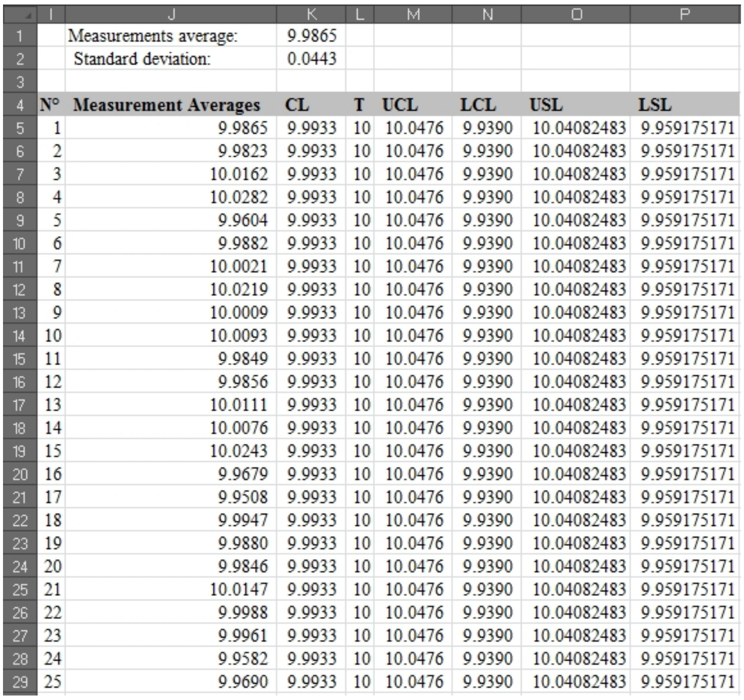
$$\overline{\overline{X}}=\frac{1}{k}\sum_{i=1}^{k}\left(\frac{1}{n_i}\sum_{j=1}^{n_i}X_{i,j}\right) \tag{16.7.763}$$

and if all the n_i are equals:

$$\overline{\overline{X}}=\frac{1}{k}\sum_{i=1}^{k}\left(\frac{1}{n}\sum_{j=1}^{n}X_{i,j}\right)=\frac{1}{nk}\sum_{i=1}^{k}\left(\sum_{j=1}^{n}X_{i,j}\right)=\frac{1}{N}\sum_{i,j}X_{i,j}=\overline{X} \tag{16.7.764}$$

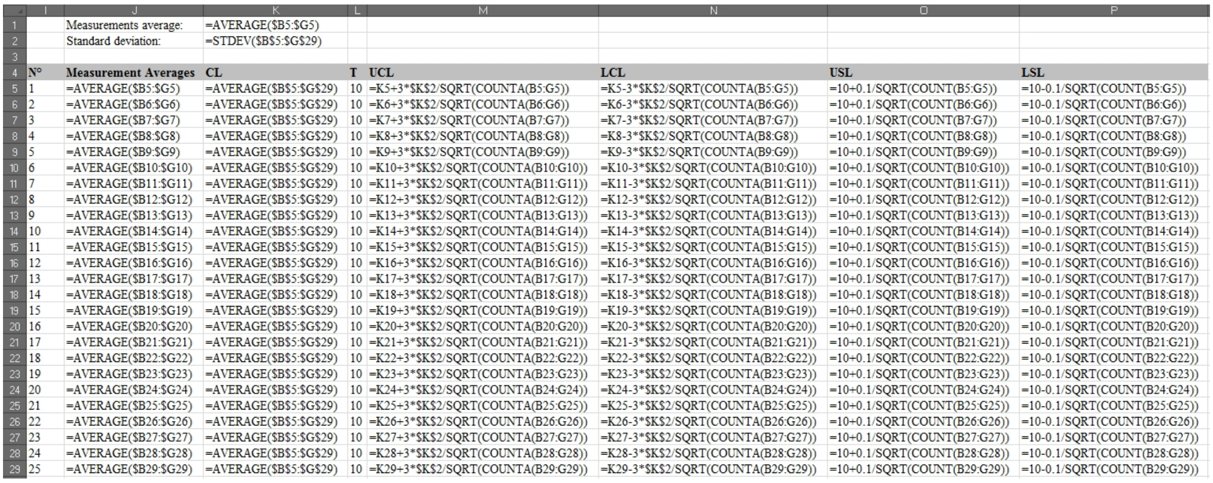
Let us notice that it is the first control chart where we can calculate the probabilistic limits of non-compliance by excess or defect of subgroups by using the fact that $\overline{X}$ follows a Normal distribution of parameters $\mathcal{N}\left(\overline{X},\sigma_X/\sqrt{n}\right)$.

Once again using our spreadsheet software here are the calculations we get in the simple case where all sample still have the same size (calculations based on the previous list of data given just earlier above):



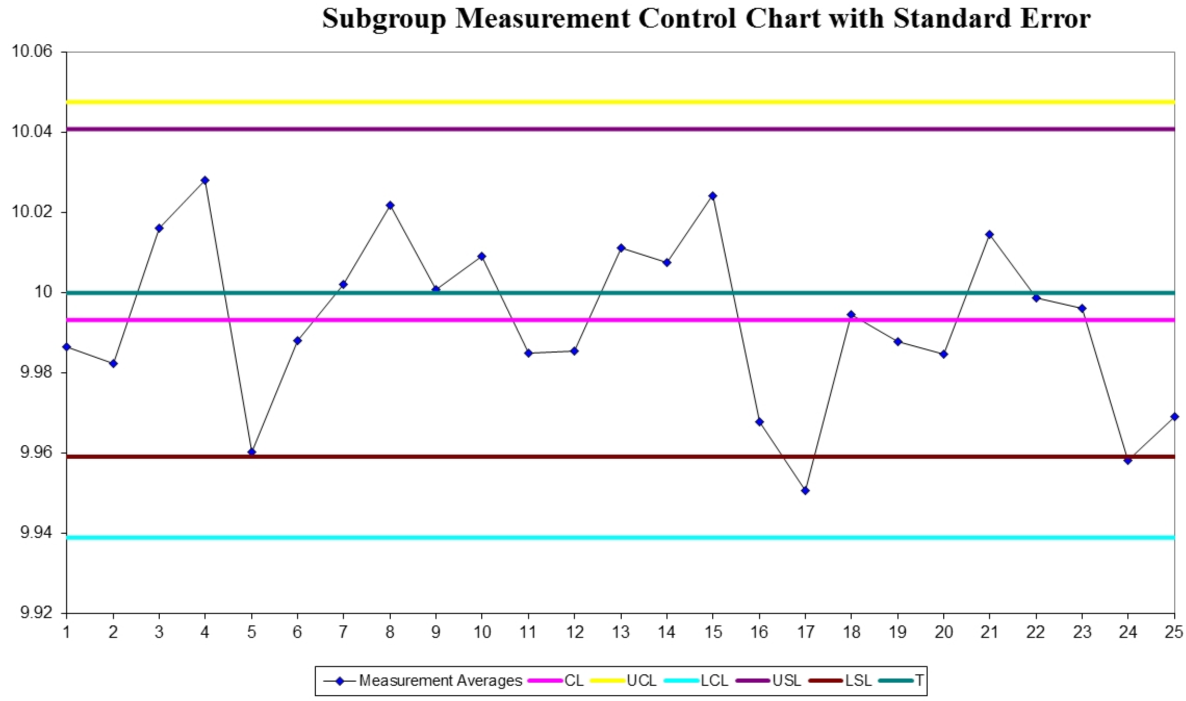
	I	J	K	L	M	N	O	P
1		Measurements average:	9.9865					
2		Standard deviation:	0.0443					
3								
4	**N°**	**Measurement Averages**	**CL**	**T**	**UCL**	**LCL**	**USL**	**LSL**
5	1	9.9865	9.9933	10	10.0476	9.9390	10.04082483	9.959175171
6	2	9.9823	9.9933	10	10.0476	9.9390	10.04082483	9.959175171
7	3	10.0162	9.9933	10	10.0476	9.9390	10.04082483	9.959175171
8	4	10.0282	9.9933	10	10.0476	9.9390	10.04082483	9.959175171
9	5	9.9604	9.9933	10	10.0476	9.9390	10.04082483	9.959175171
10	6	9.9882	9.9933	10	10.0476	9.9390	10.04082483	9.959175171
11	7	10.0021	9.9933	10	10.0476	9.9390	10.04082483	9.959175171
12	8	10.0219	9.9933	10	10.0476	9.9390	10.04082483	9.959175171
13	9	10.0009	9.9933	10	10.0476	9.9390	10.04082483	9.959175171
14	10	10.0093	9.9933	10	10.0476	9.9390	10.04082483	9.959175171
15	11	9.9849	9.9933	10	10.0476	9.9390	10.04082483	9.959175171
16	12	9.9856	9.9933	10	10.0476	9.9390	10.04082483	9.959175171
17	13	10.0111	9.9933	10	10.0476	9.9390	10.04082483	9.959175171
18	14	10.0076	9.9933	10	10.0476	9.9390	10.04082483	9.959175171
19	15	10.0243	9.9933	10	10.0476	9.9390	10.04082483	9.959175171
20	16	9.9679	9.9933	10	10.0476	9.9390	10.04082483	9.959175171
21	17	9.9508	9.9933	10	10.0476	9.9390	10.04082483	9.959175171
22	18	9.9947	9.9933	10	10.0476	9.9390	10.04082483	9.959175171
23	19	9.9880	9.9933	10	10.0476	9.9390	10.04082483	9.959175171
24	20	9.9846	9.9933	10	10.0476	9.9390	10.04082483	9.959175171
25	21	10.0147	9.9933	10	10.0476	9.9390	10.04082483	9.959175171
26	22	9.9988	9.9933	10	10.0476	9.9390	10.04082483	9.959175171
27	23	9.9961	9.9933	10	10.0476	9.9390	10.04082483	9.959175171
28	24	9.9582	9.9933	10	10.0476	9.9390	10.04082483	9.959175171
29	25	9.9690	9.9933	10	10.0476	9.9390	10.04082483	9.959175171

or explicitly:



	I	J	K	L	M	N	O	P
1		Measurements average:	=AVERAGE($B5:$G5)					
2		Standard deviation:	=STDEV($B$5:$G$29)					
3								
4	**N°**	**Measurement Averages**	**CL**	**T**	**UCL**	**LCL**	**USL**	**LSL**
5	1	=AVERAGE($B5:$G5)	=AVERAGE($B$5:$G$29)	10	=K5+3*$K$2/SQRT(COUNTA(B5:G5))	=K5-3*$K$2/SQRT(COUNTA(B5:G5))	=10+0.1/SQRT(COUNT(B5:G5))	=10-0.1/SQRT(COUNT(B5:G5))
6	2	=AVERAGE($B6:$G6)	=AVERAGE($B$5:$G$29)	10	=K6+3*$K$2/SQRT(COUNTA(B6:G6))	=K6-3*$K$2/SQRT(COUNTA(B6:G6))	=10+0.1/SQRT(COUNT(B6:G6))	=10-0.1/SQRT(COUNT(B6:G6))
7	3	=AVERAGE($B7:$G7)	=AVERAGE($B$5:$G$29)	10	=K7+3*$K$2/SQRT(COUNTA(B7:G7))	=K7-3*$K$2/SQRT(COUNTA(B7:G7))	=10+0.1/SQRT(COUNT(B7:G7))	=10-0.1/SQRT(COUNT(B7:G7))
8	4	=AVERAGE($B8:$G8)	=AVERAGE($B$5:$G$29)	10	=K8+3*$K$2/SQRT(COUNTA(B8:G8))	=K8-3*$K$2/SQRT(COUNTA(B8:G8))	=10+0.1/SQRT(COUNT(B8:G8))	=10-0.1/SQRT(COUNT(B8:G8))
9	5	=AVERAGE($B9:$G9)	=AVERAGE($B$5:$G$29)	10	=K9+3*$K$2/SQRT(COUNTA(B9:G9))	=K9-3*$K$2/SQRT(COUNTA(B9:G9))	=10+0.1/SQRT(COUNT(B9:G9))	=10-0.1/SQRT(COUNT(B9:G9))
10	6	=AVERAGE($B10:$G10)	=AVERAGE($B$5:$G$29)	10	=K10+3*$K$2/SQRT(COUNTA(B10:G10))	=K10-3*$K$2/SQRT(COUNTA(B10:G10))	=10+0.1/SQRT(COUNT(B10:G10))	=10-0.1/SQRT(COUNT(B10:G10))
11	7	=AVERAGE($B11:$G11)	=AVERAGE($B$5:$G$29)	10	=K11+3*$K$2/SQRT(COUNTA(B11:G11))	=K11-3*$K$2/SQRT(COUNTA(B11:G11))	=10+0.1/SQRT(COUNT(B11:G11))	=10-0.1/SQRT(COUNT(B11:G11))
12	8	=AVERAGE($B12:$G12)	=AVERAGE($B$5:$G$29)	10	=K12+3*$K$2/SQRT(COUNTA(B12:G12))	=K12-3*$K$2/SQRT(COUNTA(B12:G12))	=10+0.1/SQRT(COUNT(B12:G12))	=10-0.1/SQRT(COUNT(B12:G12))
13	9	=AVERAGE($B13:$G13)	=AVERAGE($B$5:$G$29)	10	=K13+3*$K$2/SQRT(COUNTA(B13:G13))	=K13-3*$K$2/SQRT(COUNTA(B13:G13))	=10+0.1/SQRT(COUNT(B13:G13))	=10-0.1/SQRT(COUNT(B13:G13))
14	10	=AVERAGE($B14:$G14)	=AVERAGE($B$5:$G$29)	10	=K14+3*$K$2/SQRT(COUNTA(B14:G14))	=K14-3*$K$2/SQRT(COUNTA(B14:G14))	=10+0.1/SQRT(COUNT(B14:G14))	=10-0.1/SQRT(COUNT(B14:G14))
15	11	=AVERAGE($B15:$G15)	=AVERAGE($B$5:$G$29)	10	=K15+3*$K$2/SQRT(COUNTA(B15:G15))	=K15-3*$K$2/SQRT(COUNTA(B15:G15))	=10+0.1/SQRT(COUNT(B15:G15))	=10-0.1/SQRT(COUNT(B15:G15))
16	12	=AVERAGE($B16:$G16)	=AVERAGE($B$5:$G$29)	10	=K16+3*$K$2/SQRT(COUNTA(B16:G16))	=K16-3*$K$2/SQRT(COUNTA(B16:G16))	=10+0.1/SQRT(COUNT(B16:G16))	=10-0.1/SQRT(COUNT(B16:G16))
17	13	=AVERAGE($B17:$G17)	=AVERAGE($B$5:$G$29)	10	=K17+3*$K$2/SQRT(COUNTA(B17:G17))	=K17-3*$K$2/SQRT(COUNTA(B17:G17))	=10+0.1/SQRT(COUNT(B17:G17))	=10-0.1/SQRT(COUNT(B17:G17))
18	14	=AVERAGE($B18:$G18)	=AVERAGE($B$5:$G$29)	10	=K18+3*$K$2/SQRT(COUNTA(B18:G18))	=K18-3*$K$2/SQRT(COUNTA(B18:G18))	=10+0.1/SQRT(COUNT(B18:G18))	=10-0.1/SQRT(COUNT(B18:G18))
19	15	=AVERAGE($B19:$G19)	=AVERAGE($B$5:$G$29)	10	=K19+3*$K$2/SQRT(COUNTA(B19:G19))	=K19-3*$K$2/SQRT(COUNTA(B19:G19))	=10+0.1/SQRT(COUNT(B19:G19))	=10-0.1/SQRT(COUNT(B19:G19))
20	16	=AVERAGE($B20:$G20)	=AVERAGE($B$5:$G$29)	10	=K20+3*$K$2/SQRT(COUNTA(B20:G20))	=K20-3*$K$2/SQRT(COUNTA(B20:G20))	=10+0.1/SQRT(COUNT(B20:G20))	=10-0.1/SQRT(COUNT(B20:G20))
21	17	=AVERAGE($B21:$G21)	=AVERAGE($B$5:$G$29)	10	=K21+3*$K$2/SQRT(COUNTA(B21:G21))	=K21-3*$K$2/SQRT(COUNTA(B21:G21))	=10+0.1/SQRT(COUNT(B21:G21))	=10-0.1/SQRT(COUNT(B21:G21))
22	18	=AVERAGE($B22:$G22)	=AVERAGE($B$5:$G$29)	10	=K22+3*$K$2/SQRT(COUNTA(B22:G22))	=K22-3*$K$2/SQRT(COUNTA(B22:G22))	=10+0.1/SQRT(COUNT(B22:G22))	=10-0.1/SQRT(COUNT(B22:G22))
23	19	=AVERAGE($B23:$G23)	=AVERAGE($B$5:$G$29)	10	=K23+3*$K$2/SQRT(COUNTA(B23:G23))	=K23-3*$K$2/SQRT(COUNTA(B23:G23))	=10+0.1/SQRT(COUNT(B23:G23))	=10-0.1/SQRT(COUNT(B23:G23))
24	20	=AVERAGE($B24:$G24)	=AVERAGE($B$5:$G$29)	10	=K24+3*$K$2/SQRT(COUNTA(B24:G24))	=K24-3*$K$2/SQRT(COUNTA(B24:G24))	=10+0.1/SQRT(COUNT(B24:G24))	=10-0.1/SQRT(COUNT(B24:G24))
25	21	=AVERAGE($B25:$G25)	=AVERAGE($B$5:$G$29)	10	=K25+3*$K$2/SQRT(COUNTA(B25:G25))	=K25-3*$K$2/SQRT(COUNTA(B25:G25))	=10+0.1/SQRT(COUNT(B25:G25))	=10-0.1/SQRT(COUNT(B25:G25))
26	22	=AVERAGE($B26:$G26)	=AVERAGE($B$5:$G$29)	10	=K26+3*$K$2/SQRT(COUNTA(B26:G26))	=K26-3*$K$2/SQRT(COUNTA(B26:G26))	=10+0.1/SQRT(COUNT(B26:G26))	=10-0.1/SQRT(COUNT(B26:G26))
27	23	=AVERAGE($B27:$G27)	=AVERAGE($B$5:$G$29)	10	=K27+3*$K$2/SQRT(COUNTA(B27:G27))	=K27-3*$K$2/SQRT(COUNTA(B27:G27))	=10+0.1/SQRT(COUNT(B27:G27))	=10-0.1/SQRT(COUNT(B27:G27))
28	24	=AVERAGE($B28:$G28)	=AVERAGE($B$5:$G$29)	10	=K28+3*$K$2/SQRT(COUNTA(B28:G28))	=K28-3*$K$2/SQRT(COUNTA(B28:G28))	=10+0.1/SQRT(COUNT(B28:G28))	=10-0.1/SQRT(COUNT(B28:G28))
29	25	=AVERAGE($B29:$G29)	=AVERAGE($B$5:$G$29)	10	=K29+3*$K$2/SQRT(COUNTA(B29:G29))	=K29-3*$K$2/SQRT(COUNTA(B29:G29))	=10+0.1/SQRT(COUNT(B29:G29))	=10-0.1/SQRT(COUNT(B29:G29))

and the corresponding chart still with the same spreadsheet software:



where we took as example for limits imposed by the customer/board committee:

$$\begin{aligned} \text{USL} &= 10+3\frac{0.0\bar{3}}{\sqrt{n}} \cong 10.040 \\ T_{\text{SE}} &= 10 \\ \text{LSL} &= 10-3\frac{0.0\bar{3}}{\sqrt{n}} \cong 9.959 \end{aligned} \tag{16.7.765}$$

Let us notice that this control chart would according to the WECO rules have 6 points beyond the $\pm 3\sigma$ (we indicate that for reasons of comparison with the previous control chart and the next one). It is therefore more sensitive to changes than the control chart. Also notice that the UCL and LCL limits are wider than the imposed USL and LSL.

With the $\pm 3\sigma_X$ limits and according to the hypothesis of Normally distributed data, we have using the symmetry of the Normal distribution and writing as usual Φ the cumulative probability function of the Normal centered reduced distribution (where we suppose $\text{LCL}_{\text{SE}} = \text{UCL}_{\text{SE}} = \sigma_{\overline{X}} = 1$):

$$\begin{aligned}&1-P\left(\text{LCL}_{\text{SE}} \leq \overline{\overline{X}} \leq \text{UCL}_{\text{SE}}\right) = 2P\left(\text{LCL}_{\text{SE}} \leq \overline{\overline{X}}\right) = 2\left(1-P\left(\overline{\overline{X}} \leq \text{UCL}_{\text{SE}}\right)\right)\\&= 2\Phi(-3) = 2(1-\phi(3)) = 2\cdot 0.01350 = 0.0027 \cong \frac{1}{370}\end{aligned} \tag{16.7.766}$$

So we will have statistically 1 point (piece/action) on the control chart on 370 outside the $\pm\sigma_{\overline{X}}$ by unit time/sample measured (so if each point on the map is made by frequency of 1 hour then we have an ARL of 0.0027 [h^{-1}]) only for statistical purposes and that must be therefore considered as false alarms (Type I Error[36]!!!) that the data are centered and reduced or not!!!!

We name the latter number "Average Run Length (ARL)" (some use the French appellation "Average Operational Period (AOP)"). When it is reported to a specific unit of time, then we speak of "Average Time to Signal (ATS)" and write it:

$$\alpha = 1-P\left(\text{LCL}_{\text{SE}} \leq \overline{\overline{X}} \leq \text{UCL}_{\text{SE}}\right) = \frac{1}{\text{ARL}_0} \tag{16.7.767}$$

or more commonly:

$$\boxed{\text{ARL}_0 = \frac{1}{\alpha}} \tag{16.7.768}$$

Obviously, by extension, the addition of the WECO empirical rules add false alarms and it has for immediate effect of reducing the ARL.

As we have already mentioned it at our beginning of the study of control charts, the ARL is important in practice, because it gives the interval sampling size beyond which we must not go! We will ensure in practice (when possible) to take a lower value close to the ARL for large productions but possibly not beyond!

Remark

As we have already mentioned, sometimes several levels of controls are added to the control charts, so we can observe the $\pm 3\sigma_{\overline{X}}$ and simultaneously $\pm 2\sigma_{\overline{X}}$. In the U.S.A factor 2 or 3 seems to be mainly used that the distribution is symmetrical or not ... as in European countries like France and Germany (industrial countries), we use probabilistic limits that have a much better reliability when data are not distributed following a symmetric (as it is the case for well known control charts using the range R that we will see later).

Remember that the problem of this control chart is that σ_X is asymptotically correct only if n_i is big. While in reality it is by far not the case since often this value is less than 10. Therefore we will see below a control chart of the average where the fixed limits are calculated based on the range R and not on σ_x. However, in the section of Statistics (see page 865), we have shown that when the standard deviation is estimated, then we have:

---

[36] Obviously there are also type II errors (no alarm when there should have). But as I've never seen anyone use it in practice for control charts.

$$\overline{X}-\frac{S_X}{\sqrt{n}}T_{\alpha/2}(n-1)\leq\mu\leq\overline{X}+\frac{S_X}{\sqrt{n}}T_{\alpha/2}(n-1) \tag{16.7.769}$$

So we can use the Shewhart control chart with the following probabilistic limits that is more accurate than the previous one:

$$\boxed{\begin{aligned}\text{USL}&=\text{required by customer/board comitee}\\ \text{UCL}_\chi&=\overline{\overline{X}}+T_{\alpha/2}(n_i-1)\frac{S_X}{\sqrt{n_i}}\\ \text{CL}_\chi&=\text{E}(\overline{X})=\overline{\overline{X}}\\ T_{\text{SE}}&=\text{target required by customer/board comitee}\\ \text{LCL}_\chi&=\overline{\overline{X}}-T_{\alpha/2}(n_i-1)\frac{S_X}{\sqrt{n_i}}\\ \text{LSL}&=\text{required by customer/board comitee}\end{aligned}} \tag{16.7.770}$$

Also as part of a process or method under statistical control, it can be interesting to communicate the average control chart (with fixed or probabilistic limits) previously presented with a control chart showing also the evolution of the standard deviation considering a Normal distribution of the random variables. Indeed, remember that the Normal distribution is completely defined by its parameters μ,σ. Hence the reason to have a look on both of them and this is why many S.P.C. softwares plots automatically two control charts at the same time!

16.7.6.4.4 $\overline{S}-S$ Subgroups measurement control chart for standard deviation

Also as part of a process or method under statistical control, it can be interesting to communicate the average control chart (with fixed or probabilistic limits) previously presented with a control chart showing also the evolution of the standard deviation considering a Normal distribution of the random variables. Indeed, remember that the Normal distribution is completely defined by its parameters μ,σ. Hence the reason to have a look on both of them and this is why many S.P.C. softwares plots automatically two control charts at the same time!

Pros:

- Complete well the previous control chart and permits to visualizes if the volatility (standard deviation) of the process stabilizes.
- Represents well the variation within the data of each sampling (instantaneous dispersion).

Cons:

- This control chart must be used only when the prototyping phase is completed (often named "Phase II" in the literature) because it is based on an inescapable strong assumption that the data are Normally distributed.

- The limits of Shewhart controls chart are not adapted in this case because, as we will prove it in details, the variance distribution law is not a symmetrical law while Shewhart limits assume this is the case!

The idea is to represent:

$$\begin{aligned}\text{UCL}_S&=\overline{S}+3\sigma_S\\ \text{CL}_S&=\text{E}(S)=\overline{S}\\ \text{LCL}_S&=\overline{S}-3\sigma_S\end{aligned}\qquad(16.7.771)$$

always with the estimator of mean (k being the number of samples for recall...):

$$\overline{S}=\frac{1}{k}\sum_{i=1}^{k}S_i\qquad(16.7.772)$$

and always with the unbiased maximum likelihood estimator of the standard deviation denoted S_i in the field of quality (at the opposite $\hat{\sigma}$ in the field of pure statistics):

$$S_i=\sqrt{\frac{1}{n_i-1}\sum_{j=1}^{n}(X_j-\overline{X})}\qquad(16.7.773)$$

where n corresponds obviously to the number of individuals for the sample i.

Remark

The mean sample standard deviation $\bar{S}$ above must not be confused with the pooled sample standard deviation used for the $\bar{X}-\tilde{S}$ control charts where:

$$\tilde{S}=\left(\frac{1}{k}\sum_{i=1}^{k}S_i^2\right)^{1/2}$$

where S_i is the i th sample standard deviation defined by:

$$S_i=\left(\frac{1}{n_i-1}\sum_{j=1}^{n}\left(X_{i,j}-\bar{X}_i\right)^2\right)^{1/2}$$

To calculate the mean and the variance of σ_S, we must consider S also as a Normally distributed random variable (it is the strong and essential assumption in this control chart that we had mentioned in the "cons" just before!). We then have proved in the section Statistics (see page 857) that we had:

$$\chi^2(n-1)=\frac{(n-1)S^2}{\sigma^2}\qquad(16.7.774)$$

which we will denote t thereafter as the corresponding random variable.

For the remaining development, we will consider $(n-1)$ and σ as simple constant coefficients of S. So the random variable t is actually implicit function of S as:

$$t=f(S)=\frac{n-1}{\sigma^2}S^2 \tag{16.7.775}$$

We will focus now to the calculation of mean of S. It is then more convenient to take the square root of t such that:

$$\sqrt{t}=\frac{\sqrt{n-1}}{\sigma}S \tag{16.7.776}$$

Therefore it comes:

$$\mathrm{E}(S)=\mu_E=\mathrm{E}\left(\frac{\sigma}{\sqrt{n-1}}\sqrt{t}\right)=\frac{\sigma}{\sqrt{n-1}}\mathrm{E}\left(\sqrt{t}\right)=\int\limits_0^{+\infty}\sqrt{t}f(t)\mathrm{d}t \tag{16.7.777}$$

Let us recall that we proved in the section Statistics (see page 773) that the chi-square law was defined by the following probability density function:

$$P_k(x)=\frac{1}{2^{k/2}\Gamma(k/2)}x^{k/2-1}e^{-x/2}1_{[0,+\infty]} \tag{16.7.778}$$

Therefore it comes:

$$\mathrm{E}(S)=\frac{\sigma}{\sqrt{n-1}}\int\limits_0^{+\infty}\sqrt{t}\frac{t^{\frac{n-1}{2}-1}e^{-t/2}}{2^{\frac{n-1}{2}}\Gamma\left(\dfrac{n-1}{2}\right)}\mathrm{d}t \tag{16.7.779}$$

Let us put for what follows:

$$n-1=k \tag{16.7.780}$$

Then we have:

$$\begin{aligned}\mathrm{E}(S)&=\frac{\sigma}{\sqrt{k}}\int\limits_0^{+\infty}\sqrt{t}\frac{t^{\frac{k}{2}-1}e^{-t/2}}{2^{\frac{k}{2}}\Gamma\left(\dfrac{k}{2}\right)}\mathrm{d}t=\frac{\sigma}{2^{k/2}\Gamma\left(\dfrac{k}{2}\right)\sqrt{k}}\int\limits_0^{+\infty}\sqrt{t}t^{\frac{k}{2}-1}e^{-t/2}\mathrm{d}t\\&=\frac{\sigma}{2^{k/2}\Gamma\left(\dfrac{k}{2}\right)\sqrt{k}}\int\limits_0^{+\infty}t^{\frac{k}{2}-1+\frac{1}{2}}e^{-t/2}\mathrm{d}t=\frac{\sigma}{2^{k/2}\Gamma\left(\dfrac{k}{2}\right)\sqrt{k}}\int\limits_0^{+\infty}t^{\frac{k-1}{2}}e^{-t/2}\mathrm{d}t\\&=\frac{\sigma}{2^{k/2}\Gamma\left(\dfrac{k}{2}\right)\sqrt{k}}\frac{\sqrt{2}\Gamma\left(\dfrac{k+1}{2}\right)}{\sqrt{\left(\dfrac{1}{2}\right)^k}}=\sigma\frac{\sqrt{2}}{2^{k/2}\sqrt{\left(\dfrac{1}{2}\right)^k}\sqrt{k}}\frac{\Gamma\left(\dfrac{k+1}{2}\right)}{\Gamma\left(\dfrac{k}{2}\right)}\\&=\sigma\frac{\sqrt{2}}{\sqrt{k}}\frac{\Gamma\left(\dfrac{k+1}{2}\right)}{\Gamma\left(\dfrac{k}{2}\right)}=\sigma\frac{\sqrt{2}}{\sqrt{n-1}}\frac{\Gamma\left(\dfrac{n}{2}\right)}{\Gamma\left(\dfrac{n-1}{2}\right)}=\sigma\sqrt{\frac{2}{n-1}}\frac{\Gamma\left(\dfrac{n}{2}\right)}{\Gamma\left(\dfrac{n-1}{2}\right)}:=\sigma c_4\end{aligned} \tag{16.7.781}$$

Therefore:

$$\mathrm{E}(S)=\overline{S}=\sigma c_4=\sigma\sqrt{\frac{2}{n-1}}\frac{\Gamma\left(\dfrac{n}{2}\right)}{\Gamma\left(\dfrac{n-1}{2}\right)} \tag{16.7.782}$$

Finally for finish on this point, be aware that it is traditional to denote the "standard deviation unbiasing constants c_4" or "Burr constants" as follows:

$$\boxed{c_4=\sqrt{\frac{2}{n-1}}\frac{\Gamma\left(\dfrac{n}{2}\right)}{\Gamma\left(\dfrac{n-1}{2}\right)}} \tag{16.7.783}$$

and for which we find the values in the tables of various books or softwares. But almost everybody has a spreadsheet software nowadays, these tables are useless as we can obtain their values for examples with spreadsheet software like Microsoft® Excel® 14.0.6123 by writing the following formula:

=SQRT(2/(n_i-1))*EXP(GAMMALN(n_i/2))/EXP(GAMMALN((n_i-1)/2))

or also with:

=SQRT(2/(n_i-1))*GAMMA(n_i/2)/GAMMA((n_i-1)/2)

or in OpenOffice Calc (the formula is there little bit simpler):

=SQRT(2/(n_i-1))*GAMMA(n_i/2)/GAMMA((n_i-1)/2)

Let us now calculate the variance of S using the Huygens relation proved in the section Statistics (see page 593) that is for recall:

$$\mathrm{V}(S)=\mathrm{E}(S^2)-\mathrm{E}(S)^2 \tag{16.7.784}$$

So we can write:

$$\mathrm{V}(S)=\mathrm{E}(S^2)-(c_4\sigma)^2=\mathrm{E}(S^2)-c_4^2\sigma^2 \tag{16.7.785}$$

$$\begin{aligned}\mathrm{E}(S^2)&=\frac{\sigma^2}{k}\int_0^{+\infty}t\frac{t^{\frac{k}{2}-1}e^{-t/2}}{2^{\frac{k}{2}}\Gamma\left(\dfrac{k}{2}\right)}\mathrm{d}t=\frac{\sigma^2}{2^{\frac{k}{2}}\Gamma\left(\dfrac{k}{2}\right)k}\int_0^{+\infty}tt^{\frac{k}{2}-1}e^{-t/2}\mathrm{d}t\\&=\frac{\sigma^2}{2^{\frac{k}{2}}\Gamma\left(\dfrac{k}{2}\right)k}\int_0^{+\infty}t^{\frac{k}{2}}e^{-t/2}\mathrm{d}t=\frac{\sigma^2}{2^{\frac{k}{2}}\Gamma\left(\dfrac{k}{2}\right)k}\frac{\Gamma\left(\dfrac{k}{2}\right)k}{\sqrt{\left(\dfrac{1}{2}\right)^k}}=\sigma^2\end{aligned} \tag{16.7.786}$$

Therefore:

$$\mathrm{V}(S)=\mathrm{E}(S^2)-\mathrm{E}(S)^2=\sigma^2-c_4^2\sigma^2=\sigma^2\left(1-c_4^2\right) \tag{16.7.787}$$

hence:

$$\sigma_S=\sigma\sqrt{1-c_4^2} \tag{16.7.788}$$

This bring us to write now:

$$\begin{aligned}
\text{USL}&=\text{required by customer/board comitee}\\
\text{UCL}_S&=\overline{S}+3\sigma_S=\overline{S}+3\sigma\sqrt{1-c_4^2(n_i)}\\
\text{CL}_S&=\mathrm{E}(S)=\overline{S}=\sigma c_4(n_i)\\
T_{\text{SE}}&=\text{target required by customer/board comitee}\\
\text{LCL}_S&=\overline{S}-3\sigma_S=\overline{S}-3\sigma\sqrt{1-c_4^2(n_i)}\\
\text{LSL}&=\text{required by customer/board comitee}
\end{aligned} \tag{16.7.789}$$

that is to say finally as sometimes written in the literature:

$$\boxed{\begin{aligned}
\text{USL}&=\text{required by customer/board comitee}\\
\text{UCL}_S&=\overline{S}+3\frac{\overline{S}}{c_4(n_i)}\sqrt{1-c_4^2(n_i)}=\overline{S}+3\frac{\overline{S}}{h(n_i)}\\
\text{CL}_S&=\overline{S}\\
T&=\text{target required by customer/board comitee}\\
\text{LCL}_S&=\overline{S}-3\frac{\overline{S}}{c_4(n_i)}\sqrt{1-c_4^2(n_i)}=\overline{S}-3\frac{\overline{S}}{h(n_i)}\\
\text{LSL}&=\text{required by customer/board comitee}
\end{aligned}} \tag{16.7.790}$$

where by definition:

$$\boxed{h(n_i)=\frac{c_4(n_i)}{\sqrt{1-c_4^2(n_i)}}} \tag{16.7.791}$$

That can be also sometimes be found in some books written as:

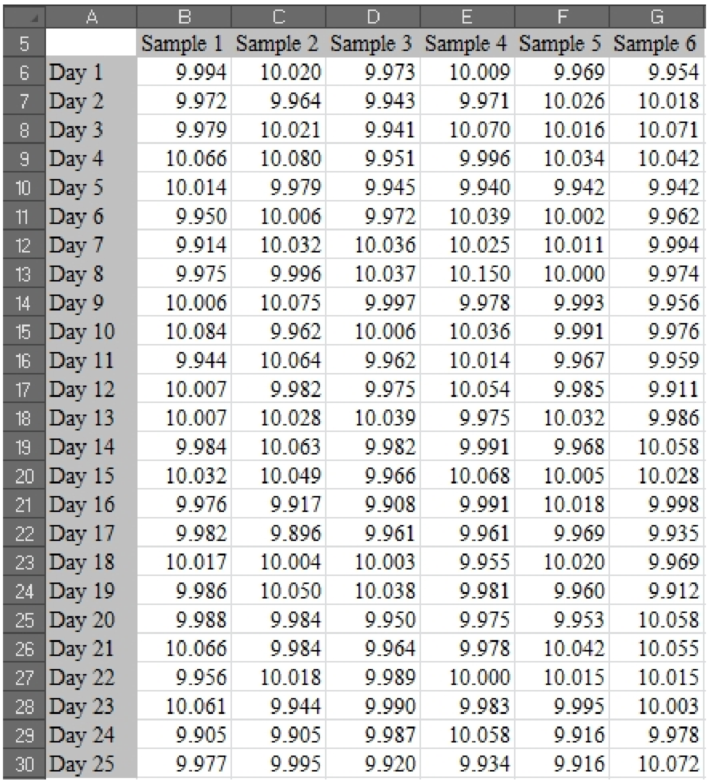
$$\boxed{\begin{aligned}
\text{USL}&=\text{required by customer/board comitee}\\
\text{UCL}_S&=\overline{S}\left(1+\frac{3}{h_i}\right)=\overline{S}B_4(n_i)\\
\text{CL}_S&=\overline{S}\\
T&=\text{target required by customer/board comitee}\\
\text{LCL}_S&=\overline{S}\left(1-\frac{3}{h_i}\right)=\overline{S}B_3(n_i)\\
\text{LSL}&=\text{required by customer/board comitee}
\end{aligned}} \tag{16.7.792}$$

with by definition the following unbiasing constants:

$$
\boxed{\begin{aligned}
B_4(n_i) &= 1+\frac{3}{h(n_i)} = 1+\frac{3\sqrt{1-c_4^2(n_i)}}{c_4(n_i)} \\
B_3(n_i) &= 1-\frac{3}{h(n_i)} = 1-\frac{3\sqrt{1-c_4^2(n_i)}}{c_4(n_i)}
\end{aligned}} \tag{16.7.793}
$$

The unbiasing constants c_4, B_3, B_4 (all dependent of n_i) can obviously be found in many tables in various books or softwares, but once again they can be easily computed using a simple spreadsheet software like Microsoft® Excel® as proved earlier above.

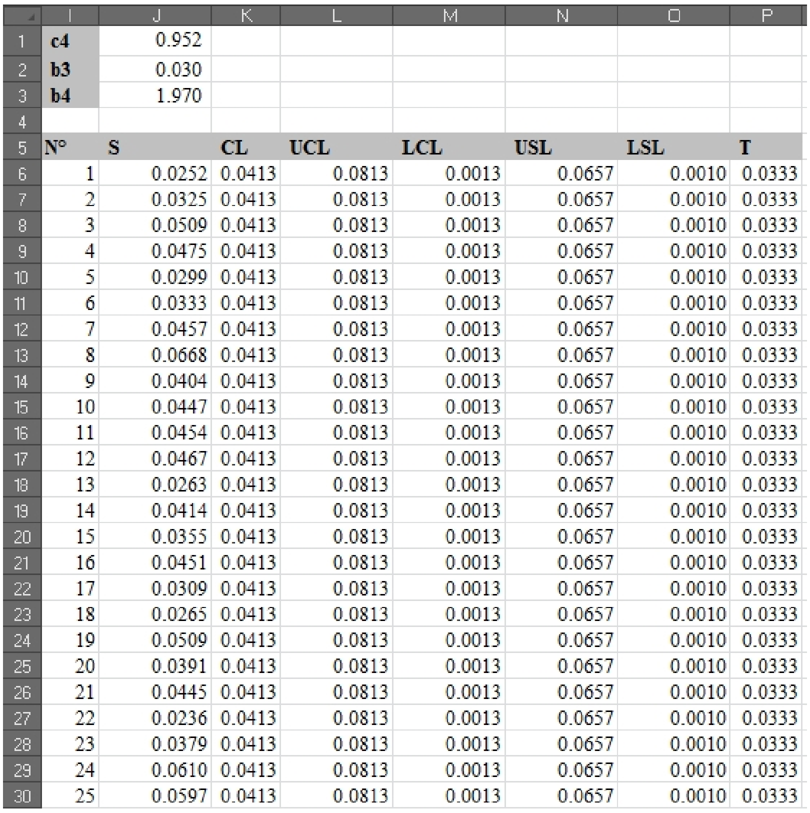
As companion example let us once again consider the following example of 25 samples (that we will identify by the variable k) of 6 individuals each (that we will identify by the variable $n_i = n$) taken from a continuous manufacturing process (without machine recalibration or without machine change!):



	A	B	C	D	E	F	G
5		Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
6	Day 1	9.994	10.020	9.973	10.009	9.969	9.954
7	Day 2	9.972	9.964	9.943	9.971	10.026	10.018
8	Day 3	9.979	10.021	9.941	10.070	10.016	10.071
9	Day 4	10.066	10.080	9.951	9.996	10.034	10.042
10	Day 5	10.014	9.979	9.945	9.940	9.942	9.942
11	Day 6	9.950	10.006	9.972	10.039	10.002	9.962
12	Day 7	9.914	10.032	10.036	10.025	10.011	9.994
13	Day 8	9.975	9.996	10.037	10.150	10.000	9.974
14	Day 9	10.006	10.075	9.997	9.978	9.993	9.956
15	Day 10	10.084	9.962	10.006	10.036	9.991	9.976
16	Day 11	9.944	10.064	9.962	10.014	9.967	9.959
17	Day 12	10.007	9.982	9.975	10.054	9.985	9.911
18	Day 13	10.007	10.028	10.039	9.975	10.032	9.986
19	Day 14	9.984	10.063	9.982	9.991	9.968	10.058
20	Day 15	10.032	10.049	9.966	10.068	10.005	10.028
21	Day 16	9.976	9.917	9.908	9.991	10.018	9.998
22	Day 17	9.982	9.896	9.961	9.961	9.969	9.935
23	Day 18	10.017	10.004	10.003	9.955	10.020	9.969
24	Day 19	9.986	10.050	10.038	9.981	9.960	9.912
25	Day 20	9.988	9.984	9.950	9.975	9.953	10.058
26	Day 21	10.066	9.984	9.964	9.978	10.042	10.055
27	Day 22	9.956	10.018	9.989	10.000	10.015	10.015
28	Day 23	10.061	9.944	9.990	9.983	9.995	10.003
29	Day 24	9.905	9.905	9.987	10.058	9.916	9.978
30	Day 25	9.977	9.995	9.920	9.934	9.916	10.072

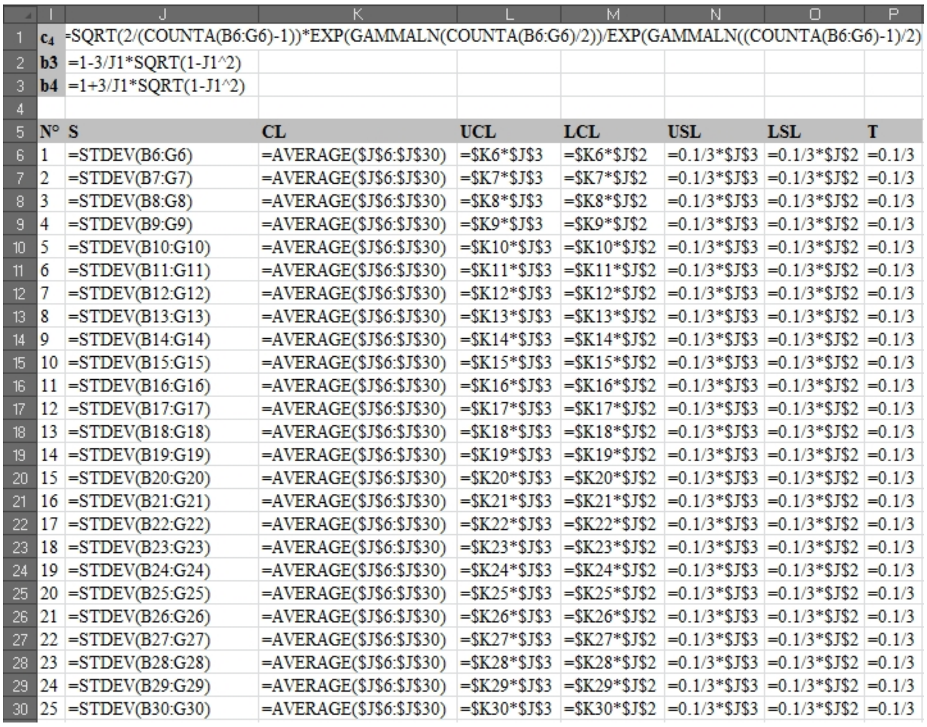
where every day we took a sample of 6 individuals of all manufactured items produced the same day.

Once again using our spreadsheet software here are the calculations we get in the simple case where all sample still have the same size (calculations based on the previous list of data given just earlier above):



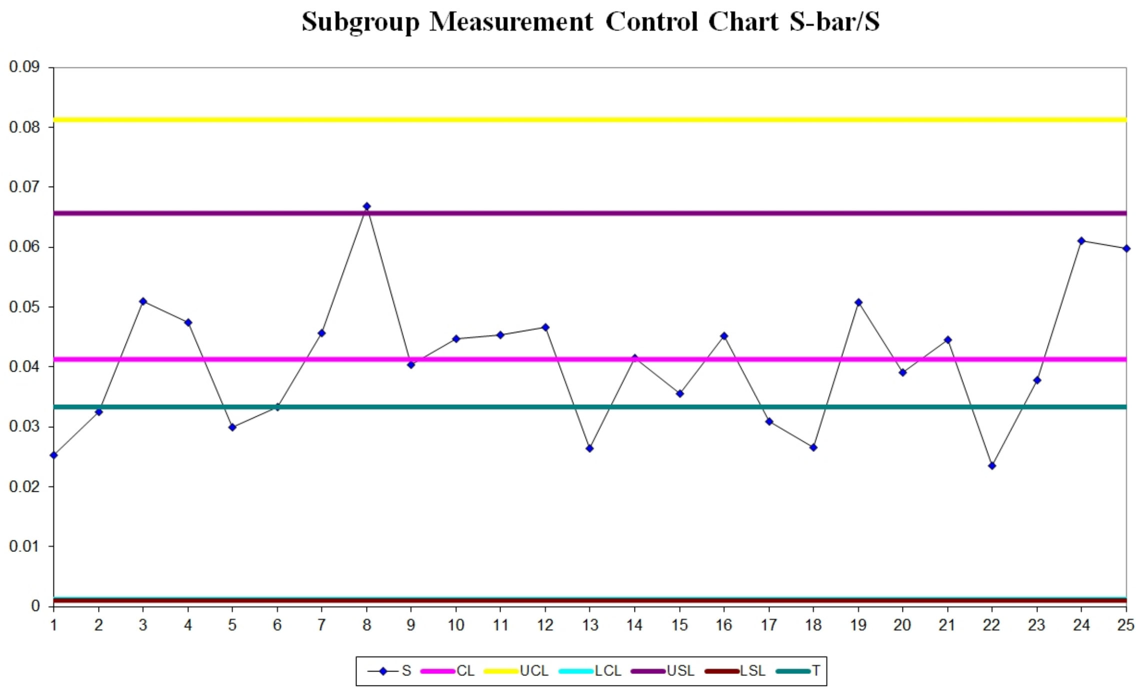
	I	J	K	L	M	N	O	P
1	**c4**	0.952						
2	**b3**	0.030						
3	**b4**	1.970						
4								
5	**N°**	**S**	**CL**	**UCL**	**LCL**	**USL**	**LSL**	**T**
6	1	0.0252	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
7	2	0.0325	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
8	3	0.0509	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
9	4	0.0475	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
10	5	0.0299	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
11	6	0.0333	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
12	7	0.0457	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
13	8	0.0668	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
14	9	0.0404	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
15	10	0.0447	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
16	11	0.0454	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
17	12	0.0467	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
18	13	0.0263	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
19	14	0.0414	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
20	15	0.0355	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
21	16	0.0451	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
22	17	0.0309	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
23	18	0.0265	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
24	19	0.0509	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
25	20	0.0391	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
26	21	0.0445	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
27	22	0.0236	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
28	23	0.0379	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
29	24	0.0610	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333
30	25	0.0597	0.0413	0.0813	0.0013	0.0657	0.0010	0.0333

or explicitly:



	I	J	K	L	M	N	O	P
1	**c4**	=SQRT(2/(COUNTA(B6:G6)-1))*EXP(GAMMALN(COUNTA(B6:G6)/2))/EXP(GAMMALN((COUNTA(B6:G6)-1)/2)						
2	**b3**	=1-3/J1*SQRT(1-J1^2)						
3	**b4**	=1+3/J1*SQRT(1-J1^2)						
4								
5	**N°**	**S**	**CL**	**UCL**	**LCL**	**USL**	**LSL**	**T**
6	1	=STDEV(B6:G6)	=AVERAGE($J$6:$J$30)	=$K6*$J$3	=$K6*$J$2	=0.1/3*$J$3	=0.1/3*$J$2	=0.1/3
7	2	=STDEV(B7:G7)	=AVERAGE($J$6:$J$30)	=$K7*$J$3	=$K7*$J$2	=0.1/3*$J$3	=0.1/3*$J$2	=0.1/3
8	3	=STDEV(B8:G8)	=AVERAGE($J$6:$J$30)	=$K8*$J$3	=$K8*$J$2	=0.1/3*$J$3	=0.1/3*$J$2	=0.1/3
9	4	=STDEV(B9:G9)	=AVERAGE($J$6:$J$30)	=$K9*$J$3	=$K9*$J$2	=0.1/3*$J$3	=0.1/3*$J$2	=0.1/3
10	5	=STDEV(B10:G10)	=AVERAGE($J$6:$J$30)	=$K10*$J$3	=$K10*$J$2	=0.1/3*$J$3	=0.1/3*$J$2	=0.1/3
11	6	=STDEV(B11:G11)	=AVERAGE($J$6:$J$30)	=$K11*$J$3	=$K11*$J$2	=0.1/3*$J$3	=0.1/3*$J$2	=0.1/3
12	7	=STDEV(B12:G12)	=AVERAGE($J$6:$J$30)	=$K12*$J$3	=$K12*$J$2	=0.1/3*$J$3	=0.1/3*$J$2	=0.1/3
13	8	=STDEV(B13:G13)	=AVERAGE($J$6:$J$30)	=$K13*$J$3	=$K13*$J$2	=0.1/3*$J$3	=0.1/3*$J$2	=0.1/3
14	9	=STDEV(B14:G14)	=AVERAGE($J$6:$J$30)	=$K14*$J$3	=$K14*$J$2	=0.1/3*$J$3	=0.1/3*$J$2	=0.1/3
15	10	=STDEV(B15:G15)	=AVERAGE($J$6:$J$30)	=$K15*$J$3	=$K15*$J$2	=0.1/3*$J$3	=0.1/3*$J$2	=0.1/3
16	11	=STDEV(B16:G16)	=AVERAGE($J$6:$J$30)	=$K16*$J$3	=$K16*$J$2	=0.1/3*$J$3	=0.1/3*$J$2	=0.1/3
17	12	=STDEV(B17:G17)	=AVERAGE($J$6:$J$30)	=$K17*$J$3	=$K17*$J$2	=0.1/3*$J$3	=0.1/3*$J$2	=0.1/3
18	13	=STDEV(B18:G18)	=AVERAGE($J$6:$J$30)	=$K18*$J$3	=$K18*$J$2	=0.1/3*$J$3	=0.1/3*$J$2	=0.1/3
19	14	=STDEV(B19:G19)	=AVERAGE($J$6:$J$30)	=$K19*$J$3	=$K19*$J$2	=0.1/3*$J$3	=0.1/3*$J$2	=0.1/3
20	15	=STDEV(B20:G20)	=AVERAGE($J$6:$J$30)	=$K20*$J$3	=$K20*$J$2	=0.1/3*$J$3	=0.1/3*$J$2	=0.1/3
21	16	=STDEV(B21:G21)	=AVERAGE($J$6:$J$30)	=$K21*$J$3	=$K21*$J$2	=0.1/3*$J$3	=0.1/3*$J$2	=0.1/3
22	17	=STDEV(B22:G22)	=AVERAGE($J$6:$J$30)	=$K22*$J$3	=$K22*$J$2	=0.1/3*$J$3	=0.1/3*$J$2	=0.1/3
23	18	=STDEV(B23:G23)	=AVERAGE($J$6:$J$30)	=$K23*$J$3	=$K23*$J$2	=0.1/3*$J$3	=0.1/3*$J$2	=0.1/3
24	19	=STDEV(B24:G24)	=AVERAGE($J$6:$J$30)	=$K24*$J$3	=$K24*$J$2	=0.1/3*$J$3	=0.1/3*$J$2	=0.1/3
25	20	=STDEV(B25:G25)	=AVERAGE($J$6:$J$30)	=$K25*$J$3	=$K25*$J$2	=0.1/3*$J$3	=0.1/3*$J$2	=0.1/3
26	21	=STDEV(B26:G26)	=AVERAGE($J$6:$J$30)	=$K26*$J$3	=$K26*$J$2	=0.1/3*$J$3	=0.1/3*$J$2	=0.1/3
27	22	=STDEV(B27:G27)	=AVERAGE($J$6:$J$30)	=$K27*$J$3	=$K27*$J$2	=0.1/3*$J$3	=0.1/3*$J$2	=0.1/3
28	23	=STDEV(B28:G28)	=AVERAGE($J$6:$J$30)	=$K28*$J$3	=$K28*$J$2	=0.1/3*$J$3	=0.1/3*$J$2	=0.1/3
29	24	=STDEV(B29:G29)	=AVERAGE($J$6:$J$30)	=$K29*$J$3	=$K29*$J$2	=0.1/3*$J$3	=0.1/3*$J$2	=0.1/3
30	25	=STDEV(B30:G30)	=AVERAGE($J$6:$J$30)	=$K30*$J$3	=$K30*$J$2	=0.1/3*$J$3	=0.1/3*$J$2	=0.1/3

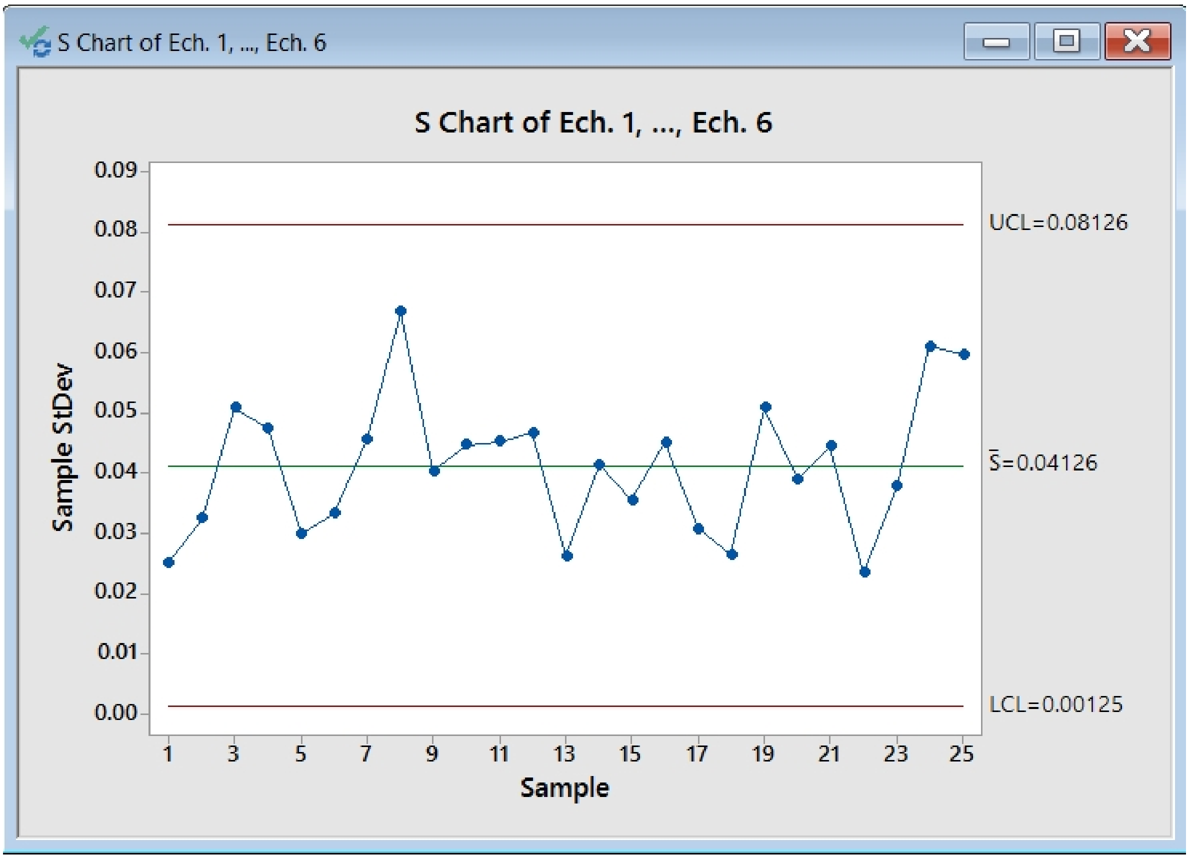
and the corresponding chart still with the same spreadsheet software:



where we took as example for limits imposed by the customer/board committee:

$$\begin{aligned}\text{USL}&=0.0\bar{3}B_4(n_i)=0.0\bar{3}B_4(6)=0.0657\\ T&=0.0\bar{3}\\ \text{LSL}&=0.0\bar{3}B_3(n_i)=0.0\bar{3}B_3(6)=0.001\end{aligned} \quad (16.7.794)$$

With Minitab 17.1.3 this gives for people that don't trust our calculations made with Microsoft® Excel® (we can also do such a chart with `R` but the plot quality output is not acceptable in most corporations):



Let us notice that this control chart would according to the WECO rules have 0 points beyond the $\pm 3\sigma$ (we indicate that for reasons of comparison with the previous control chart and the next one). It is therefore more sensitive to changes than the control chart. Also notice that the UCL and LCL limits are wider than the imposed USL and LSL.

For recall, the major problem with this control chart is that it is correct if and only if the data are Normally distributed and that Shewhart limits suppose that S is symmetric and here in fact it is note case (if we took probabilistic limits). Therefore, the quality manager then combine with the average control chart another optional control which we will introduce now.

16.7.6.4.5 $\overline{X}-S$ Subgroups measurement control chart

During our study of the control chart with averages with fixed limits based on the standard deviation of the mean (standard error), we have clearly shown that it suffered from the handicap that σ_X is too approximative in practice (because n_i is often too small).

Therefore, a workaround to this is to use the results obtained from the study of the previous control chart.

Pros:

- Corrects the problem of the control chart for averages with standard error when the number of individuals per sample is too small.
- Represents well the change of level in the average of data (central tendency).

Cons:

- The mathematical proof of the origin of the control chart limits is quite difficult for undergraduates (see page 8980).
- This control chart only works if and only if the data are identically distributed and are Normally distributed!
- Moreover, since the limits are based on the estimator of the standard deviation it would require however n_i that in the idea should not be below the value of 10 ...

First let us recall that for the average control chart with standard error limits we get:

$$\begin{aligned}\text{USL}&=\text{required by customer/board comitee}\\ \text{UCL}_{\text{SE}}&=\overline{\overline{X}}+3\sigma_{\overline{X}}=\overline{\overline{X}}+3\frac{\sigma_X}{\sqrt{n_i}}\\ \text{CL}_{\text{SE}}&=\text{E}(\overline{X})=\overline{\overline{X}}\\ T_{\text{SE}}&=\text{target required by customer/board comitee}\\ \text{LCL}_{\text{SE}}&=\overline{\overline{X}}-3\sigma_{\overline{X}}=\overline{\overline{X}}-3\frac{\sigma_X}{\sqrt{n_i}}\\ \text{LSL}&=\text{required by customer/board comitee}\end{aligned} \tag{16.7.795}$$

and we have proved earlier above:

$$\bar{S}=\sigma_X c_4(n_i) \tag{16.7.796}$$

We have therefore:

$$\boxed{\begin{aligned}\text{USL} &= \text{required by customer/board comitee}\\ \text{UCL}_{\overline{X}(S)} &= \overline{\overline{X}}+3\frac{\sigma_X}{\sqrt{n_i}}=\overline{\overline{X}}+3\frac{\overline{S}}{c_4(n_i)\sqrt{n_i}}\\ \text{CL}_{\overline{X}(S)} &= \overline{\overline{X}}\\ T_{\overline{X}(S)} &= \text{target required by customer/board comitee}\\ \text{LCL}_{\overline{X}(S)} &= \overline{\overline{X}}-3\frac{\sigma_X}{\sqrt{n_i}}=\overline{\overline{X}}-3\frac{\overline{S}}{c_4(n_i)\sqrt{n_i}}\\ \text{LSL} &= \text{required by customer/board comitee}\end{aligned}} \tag{16.7.797}$$

and that we sometimes find in the literature written as:

$$\boxed{\begin{aligned}\text{USL} &= \text{required by customer/board comitee}\\ \text{UCL}_{\overline{X}(S)} &= \overline{\overline{X}}+3\frac{\overline{S}}{c_4(n_i)\sqrt{n_i}}=\overline{\overline{X}}+A_3(n_i)\overline{S}\\ \text{CL}_{\overline{X}(S)} &= \overline{\overline{X}}\\ T_{\overline{X}(S)} &= \text{target required by customer/board comitee}\\ \text{LCL}_{\overline{X}(S)} &= \overline{\overline{X}}-3\frac{\overline{S}}{c_4(n_i)\sqrt{n_i}}=\overline{\overline{X}}-A_3(n_i)\overline{S}\\ \text{LSL} &= \text{required by customer/board comitee}\end{aligned}} \tag{16.7.798}$$

with the unbiasing constant:

$$\boxed{A_3(n_i)=\frac{3}{c_4(n_i)\sqrt{n_i}}} \tag{16.7.799}$$

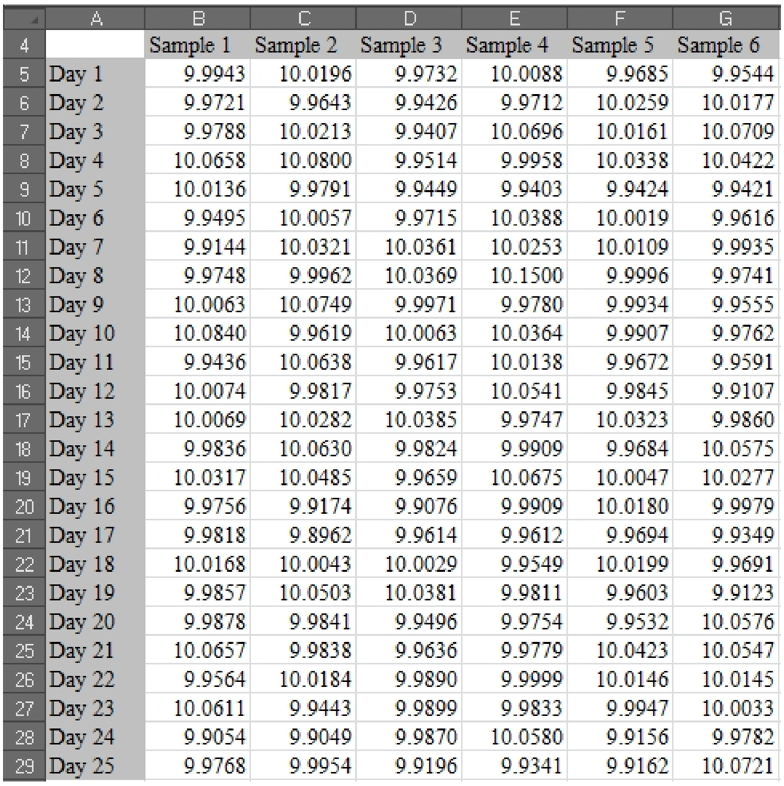
Remark

The choice of the estimator for the $\bar{X}$ control chart when the process distribution is non-normal depends on the situation at hand. When the limits based on normality are applied, the best estimator is the estimator based on $\tilde{S}$ (pooled sample standard deviation) since the resulting charts perform the best both under normality and under non-normality according to [483].

Notice then that we can calculate the probabilities of non-compliance by excess or defect of subgroups using the fact that $\overline{X}$ follows a Normal distribution of parameters:

$$\mathcal{N}(\mu,\sigma)\cong\mathcal{N}\left(\overline{\overline{X}},\frac{\overline{S}}{c_4(n_i)\sqrt{n}}\right) \tag{16.7.800}$$

As companion example let us once again consider the following example of 25 samples (that we will identify by the variable k) of 6 individuals each (that we will identify by the variable $n_i=n$) taken from a continuous manufacturing process (without machine recalibration or without machine change!):



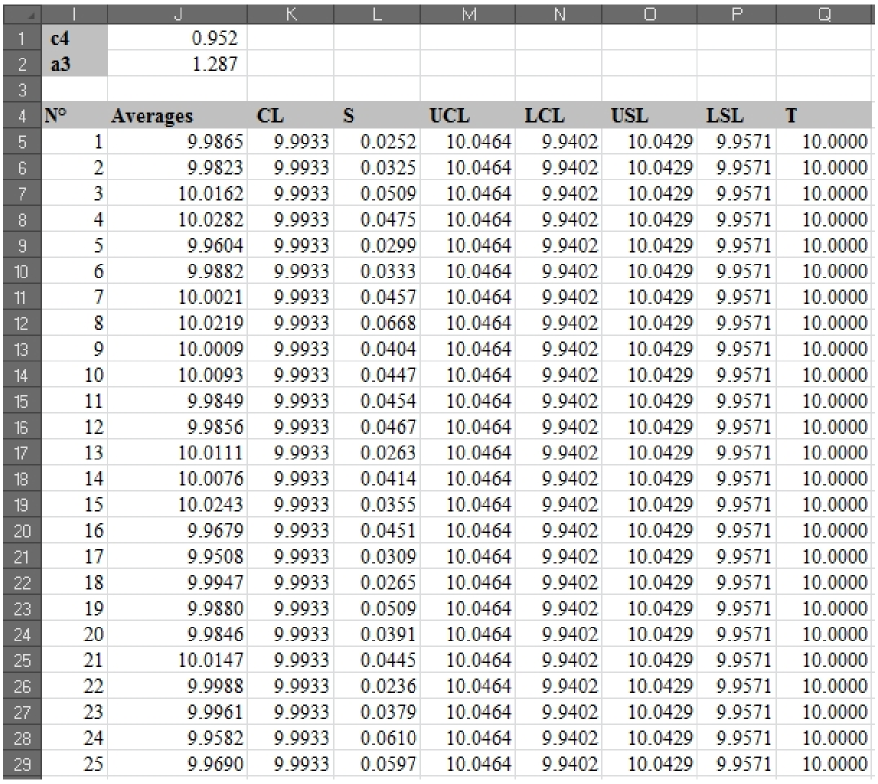
	A	B	C	D	E	F	G
4		Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
5	Day 1	9.9943	10.0196	9.9732	10.0088	9.9685	9.9544
6	Day 2	9.9721	9.9643	9.9426	9.9712	10.0259	10.0177
7	Day 3	9.9788	10.0213	9.9407	10.0696	10.0161	10.0709
8	Day 4	10.0658	10.0800	9.9514	9.9958	10.0338	10.0422
9	Day 5	10.0136	9.9791	9.9449	9.9403	9.9424	9.9421
10	Day 6	9.9495	10.0057	9.9715	10.0388	10.0019	9.9616
11	Day 7	9.9144	10.0321	10.0361	10.0253	10.0109	9.9935
12	Day 8	9.9748	9.9962	10.0369	10.1500	9.9996	9.9741
13	Day 9	10.0063	10.0749	9.9971	9.9780	9.9934	9.9555
14	Day 10	10.0840	9.9619	10.0063	10.0364	9.9907	9.9762
15	Day 11	9.9436	10.0638	9.9617	10.0138	9.9672	9.9591
16	Day 12	10.0074	9.9817	9.9753	10.0541	9.9845	9.9107
17	Day 13	10.0069	10.0282	10.0385	9.9747	10.0323	9.9860
18	Day 14	9.9836	10.0630	9.9824	9.9909	9.9684	10.0575
19	Day 15	10.0317	10.0485	9.9659	10.0675	10.0047	10.0277
20	Day 16	9.9756	9.9174	9.9076	9.9909	10.0180	9.9979
21	Day 17	9.9818	9.8962	9.9614	9.9612	9.9694	9.9349
22	Day 18	10.0168	10.0043	10.0029	9.9549	10.0199	9.9691
23	Day 19	9.9857	10.0503	10.0381	9.9811	9.9603	9.9123
24	Day 20	9.9878	9.9841	9.9496	9.9754	9.9532	10.0576
25	Day 21	10.0657	9.9838	9.9636	9.9779	10.0423	10.0547
26	Day 22	9.9564	10.0184	9.9890	9.9999	10.0146	10.0145
27	Day 23	10.0611	9.9443	9.9899	9.9833	9.9947	10.0033
28	Day 24	9.9054	9.9049	9.9870	10.0580	9.9156	9.9782
29	Day 25	9.9768	9.9954	9.9196	9.9341	9.9162	10.0721

where every day we took a sample of 6 individuals of all manufactured items produced the same day.

We can calculate, or find in the tables that:

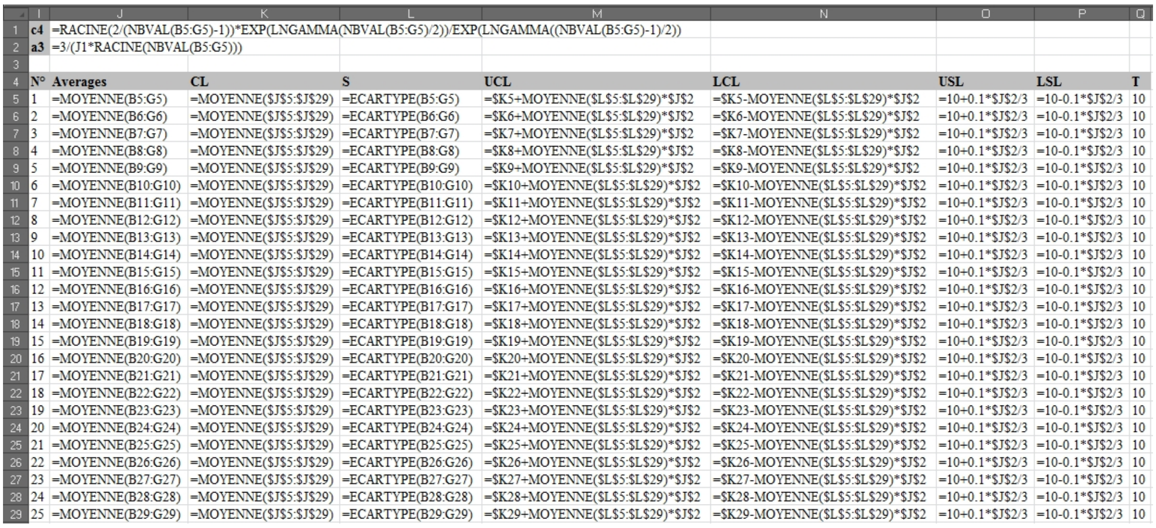
$$c_4(n_i)=c_4(6)=0.9515 \qquad A_3(n_i)=A_3(6)=0.630 \tag{16.7.801}$$

We then have the following table:



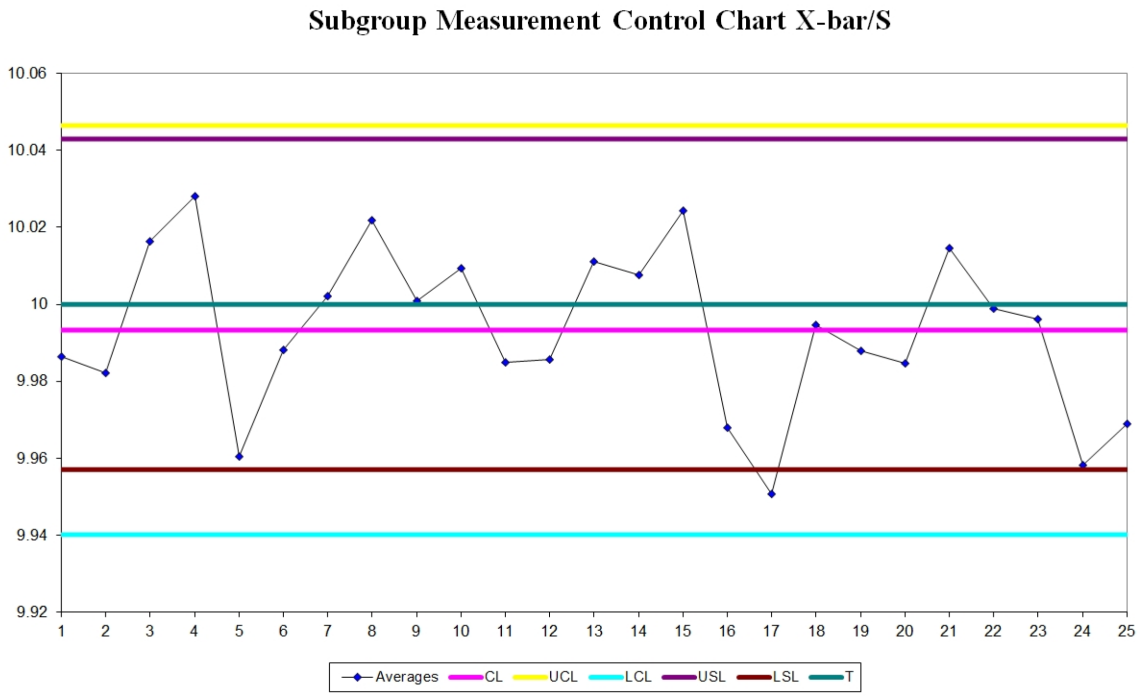
	I	J	K	L	M	N	O	P	Q
1	**c4**	0.952							
2	**a3**	1.287							
3									
4	**N°**	**Averages**	**CL**	**S**	**UCL**	**LCL**	**USL**	**LSL**	**T**
5	1	9.9865	9.9933	0.0252	10.0464	9.9402	10.0429	9.9571	10.0000
6	2	9.9823	9.9933	0.0325	10.0464	9.9402	10.0429	9.9571	10.0000
7	3	10.0162	9.9933	0.0509	10.0464	9.9402	10.0429	9.9571	10.0000
8	4	10.0282	9.9933	0.0475	10.0464	9.9402	10.0429	9.9571	10.0000
9	5	9.9604	9.9933	0.0299	10.0464	9.9402	10.0429	9.9571	10.0000
10	6	9.9882	9.9933	0.0333	10.0464	9.9402	10.0429	9.9571	10.0000
11	7	10.0021	9.9933	0.0457	10.0464	9.9402	10.0429	9.9571	10.0000
12	8	10.0219	9.9933	0.0668	10.0464	9.9402	10.0429	9.9571	10.0000
13	9	10.0009	9.9933	0.0404	10.0464	9.9402	10.0429	9.9571	10.0000
14	10	10.0093	9.9933	0.0447	10.0464	9.9402	10.0429	9.9571	10.0000
15	11	9.9849	9.9933	0.0454	10.0464	9.9402	10.0429	9.9571	10.0000
16	12	9.9856	9.9933	0.0467	10.0464	9.9402	10.0429	9.9571	10.0000
17	13	10.0111	9.9933	0.0263	10.0464	9.9402	10.0429	9.9571	10.0000
18	14	10.0076	9.9933	0.0414	10.0464	9.9402	10.0429	9.9571	10.0000
19	15	10.0243	9.9933	0.0355	10.0464	9.9402	10.0429	9.9571	10.0000
20	16	9.9679	9.9933	0.0451	10.0464	9.9402	10.0429	9.9571	10.0000
21	17	9.9508	9.9933	0.0309	10.0464	9.9402	10.0429	9.9571	10.0000
22	18	9.9947	9.9933	0.0265	10.0464	9.9402	10.0429	9.9571	10.0000
23	19	9.9880	9.9933	0.0509	10.0464	9.9402	10.0429	9.9571	10.0000
24	20	9.9846	9.9933	0.0391	10.0464	9.9402	10.0429	9.9571	10.0000
25	21	10.0147	9.9933	0.0445	10.0464	9.9402	10.0429	9.9571	10.0000
26	22	9.9988	9.9933	0.0236	10.0464	9.9402	10.0429	9.9571	10.0000
27	23	9.9961	9.9933	0.0379	10.0464	9.9402	10.0429	9.9571	10.0000
28	24	9.9582	9.9933	0.0610	10.0464	9.9402	10.0429	9.9571	10.0000
29	25	9.9690	9.9933	0.0597	10.0464	9.9402	10.0429	9.9571	10.0000

or explicitly:



	I	J	K	L	M	N	O	P	Q
1	**c4**	=RACINE(2/(NBVAL(B5:G5)-1))*EXP(LNGAMMA(NBVAL(B5:G5)/2))/EXP(LNGAMMA((NBVAL(B5:G5)-1)/2))							
2	**a3**	=3/(J1*RACINE(NBVAL(B5:G5)))							
3									
4	**N°**	**Averages**	**CL**	**S**	**UCL**	**LCL**	**USL**	**LSL**	**T**
5	1	=MOYENNE(B5:G5)	=MOYENNE($J$5:$J$29)	=ECARTYPE(B5:G5)	=$K5+MOYENNE($L$5:$L$29)*$J$2	=$K5-MOYENNE($L$5:$L$29)*$J$2	=10+0.1*$J$2/3	=10-0.1*$J$2/3	10
6	2	=MOYENNE(B6:G6)	=MOYENNE($J$5:$J$29)	=ECARTYPE(B6:G6)	=$K6+MOYENNE($L$5:$L$29)*$J$2	=$K6-MOYENNE($L$5:$L$29)*$J$2	=10+0.1*$J$2/3	=10-0.1*$J$2/3	10
7	3	=MOYENNE(B7:G7)	=MOYENNE($J$5:$J$29)	=ECARTYPE(B7:G7)	=$K7+MOYENNE($L$5:$L$29)*$J$2	=$K7-MOYENNE($L$5:$L$29)*$J$2	=10+0.1*$J$2/3	=10-0.1*$J$2/3	10
8	4	=MOYENNE(B8:G8)	=MOYENNE($J$5:$J$29)	=ECARTYPE(B8:G8)	=$K8+MOYENNE($L$5:$L$29)*$J$2	=$K8-MOYENNE($L$5:$L$29)*$J$2	=10+0.1*$J$2/3	=10-0.1*$J$2/3	10
9	5	=MOYENNE(B9:G9)	=MOYENNE($J$5:$J$29)	=ECARTYPE(B9:G9)	=$K9+MOYENNE($L$5:$L$29)*$J$2	=$K9-MOYENNE($L$5:$L$29)*$J$2	=10+0.1*$J$2/3	=10-0.1*$J$2/3	10
10	6	=MOYENNE(B10:G10)	=MOYENNE($J$5:$J$29)	=ECARTYPE(B10:G10)	=$K10+MOYENNE($L$5:$L$29)*$J$2	=$K10-MOYENNE($L$5:$L$29)*$J$2	=10+0.1*$J$2/3	=10-0.1*$J$2/3	10
11	7	=MOYENNE(B11:G11)	=MOYENNE($J$5:$J$29)	=ECARTYPE(B11:G11)	=$K11+MOYENNE($L$5:$L$29)*$J$2	=$K11-MOYENNE($L$5:$L$29)*$J$2	=10+0.1*$J$2/3	=10-0.1*$J$2/3	10
12	8	=MOYENNE(B12:G12)	=MOYENNE($J$5:$J$29)	=ECARTYPE(B12:G12)	=$K12+MOYENNE($L$5:$L$29)*$J$2	=$K12-MOYENNE($L$5:$L$29)*$J$2	=10+0.1*$J$2/3	=10-0.1*$J$2/3	10
13	9	=MOYENNE(B13:G13)	=MOYENNE($J$5:$J$29)	=ECARTYPE(B13:G13)	=$K13+MOYENNE($L$5:$L$29)*$J$2	=$K13-MOYENNE($L$5:$L$29)*$J$2	=10+0.1*$J$2/3	=10-0.1*$J$2/3	10
14	10	=MOYENNE(B14:G14)	=MOYENNE($J$5:$J$29)	=ECARTYPE(B14:G14)	=$K14+MOYENNE($L$5:$L$29)*$J$2	=$K14-MOYENNE($L$5:$L$29)*$J$2	=10+0.1*$J$2/3	=10-0.1*$J$2/3	10
15	11	=MOYENNE(B15:G15)	=MOYENNE($J$5:$J$29)	=ECARTYPE(B15:G15)	=$K15+MOYENNE($L$5:$L$29)*$J$2	=$K15-MOYENNE($L$5:$L$29)*$J$2	=10+0.1*$J$2/3	=10-0.1*$J$2/3	10
16	12	=MOYENNE(B16:G16)	=MOYENNE($J$5:$J$29)	=ECARTYPE(B16:G16)	=$K16+MOYENNE($L$5:$L$29)*$J$2	=$K16-MOYENNE($L$5:$L$29)*$J$2	=10+0.1*$J$2/3	=10-0.1*$J$2/3	10
17	13	=MOYENNE(B17:G17)	=MOYENNE($J$5:$J$29)	=ECARTYPE(B17:G17)	=$K17+MOYENNE($L$5:$L$29)*$J$2	=$K17-MOYENNE($L$5:$L$29)*$J$2	=10+0.1*$J$2/3	=10-0.1*$J$2/3	10
18	14	=MOYENNE(B18:G18)	=MOYENNE($J$5:$J$29)	=ECARTYPE(B18:G18)	=$K18+MOYENNE($L$5:$L$29)*$J$2	=$K18-MOYENNE($L$5:$L$29)*$J$2	=10+0.1*$J$2/3	=10-0.1*$J$2/3	10
19	15	=MOYENNE(B19:G19)	=MOYENNE($J$5:$J$29)	=ECARTYPE(B19:G19)	=$K19+MOYENNE($L$5:$L$29)*$J$2	=$K19-MOYENNE($L$5:$L$29)*$J$2	=10+0.1*$J$2/3	=10-0.1*$J$2/3	10
20	16	=MOYENNE(B20:G20)	=MOYENNE($J$5:$J$29)	=ECARTYPE(B20:G20)	=$K20+MOYENNE($L$5:$L$29)*$J$2	=$K20-MOYENNE($L$5:$L$29)*$J$2	=10+0.1*$J$2/3	=10-0.1*$J$2/3	10
21	17	=MOYENNE(B21:G21)	=MOYENNE($J$5:$J$29)	=ECARTYPE(B21:G21)	=$K21+MOYENNE($L$5:$L$29)*$J$2	=$K21-MOYENNE($L$5:$L$29)*$J$2	=10+0.1*$J$2/3	=10-0.1*$J$2/3	10
22	18	=MOYENNE(B22:G22)	=MOYENNE($J$5:$J$29)	=ECARTYPE(B22:G22)	=$K22+MOYENNE($L$5:$L$29)*$J$2	=$K22-MOYENNE($L$5:$L$29)*$J$2	=10+0.1*$J$2/3	=10-0.1*$J$2/3	10
23	19	=MOYENNE(B23:G23)	=MOYENNE($J$5:$J$29)	=ECARTYPE(B23:G23)	=$K23+MOYENNE($L$5:$L$29)*$J$2	=$K23-MOYENNE($L$5:$L$29)*$J$2	=10+0.1*$J$2/3	=10-0.1*$J$2/3	10
24	20	=MOYENNE(B24:G24)	=MOYENNE($J$5:$J$29)	=ECARTYPE(B24:G24)	=$K24+MOYENNE($L$5:$L$29)*$J$2	=$K24-MOYENNE($L$5:$L$29)*$J$2	=10+0.1*$J$2/3	=10-0.1*$J$2/3	10
25	21	=MOYENNE(B25:G25)	=MOYENNE($J$5:$J$29)	=ECARTYPE(B25:G25)	=$K25+MOYENNE($L$5:$L$29)*$J$2	=$K25-MOYENNE($L$5:$L$29)*$J$2	=10+0.1*$J$2/3	=10-0.1*$J$2/3	10
26	22	=MOYENNE(B26:G26)	=MOYENNE($J$5:$J$29)	=ECARTYPE(B26:G26)	=$K26+MOYENNE($L$5:$L$29)*$J$2	=$K26-MOYENNE($L$5:$L$29)*$J$2	=10+0.1*$J$2/3	=10-0.1*$J$2/3	10
27	23	=MOYENNE(B27:G27)	=MOYENNE($J$5:$J$29)	=ECARTYPE(B27:G27)	=$K27+MOYENNE($L$5:$L$29)*$J$2	=$K27-MOYENNE($L$5:$L$29)*$J$2	=10+0.1*$J$2/3	=10-0.1*$J$2/3	10
28	24	=MOYENNE(B28:G28)	=MOYENNE($J$5:$J$29)	=ECARTYPE(B28:G28)	=$K28+MOYENNE($L$5:$L$29)*$J$2	=$K28-MOYENNE($L$5:$L$29)*$J$2	=10+0.1*$J$2/3	=10-0.1*$J$2/3	10
29	25	=MOYENNE(B29:G29)	=MOYENNE($J$5:$J$29)	=ECARTYPE(B29:G29)	=$K29+MOYENNE($L$5:$L$29)*$J$2	=$K29-MOYENNE($L$5:$L$29)*$J$2	=10+0.1*$J$2/3	=10-0.1*$J$2/3	10

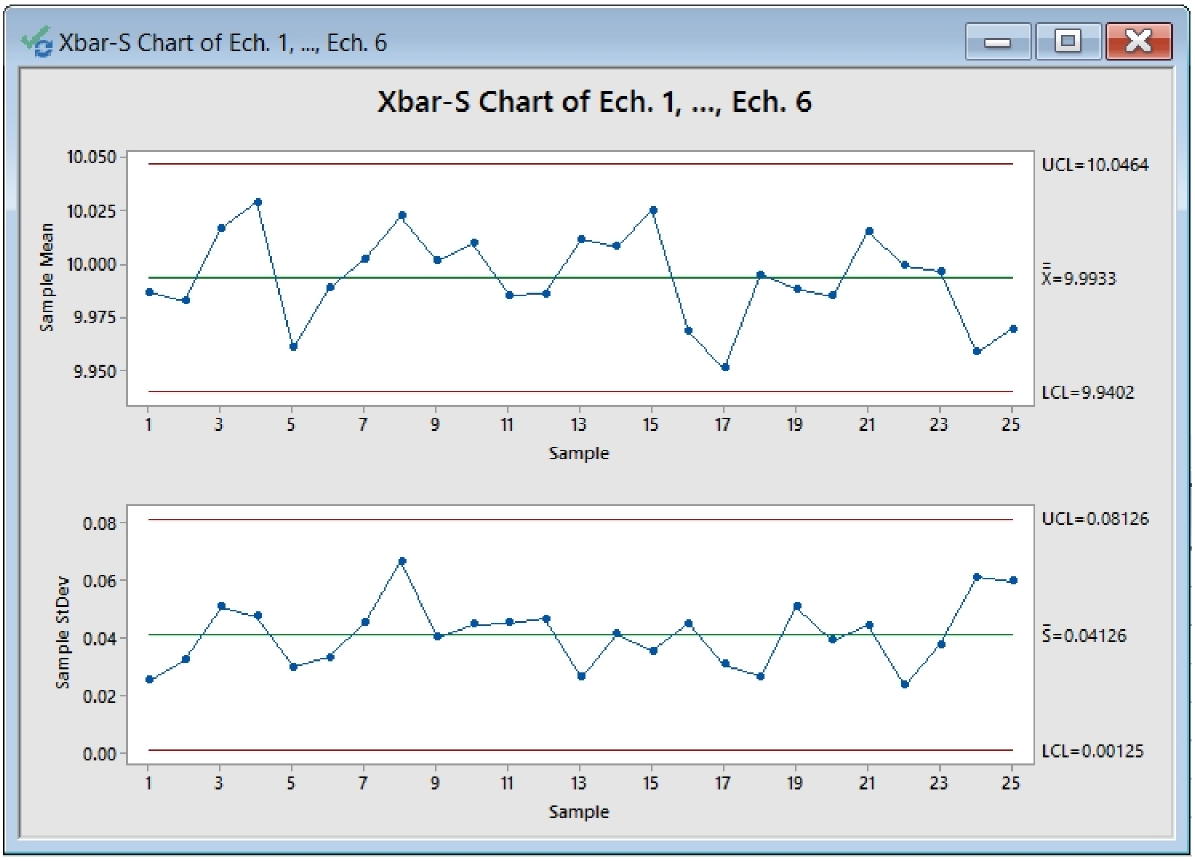
and the corresponding chart still with the same spreadsheet software:



where we took as example for limits imposed by the customer/board committee:

$$\begin{aligned}\text{USL}&=10+0.0\bar{3}A_3(n_i)=10+0.0\bar{3}A_3(6)=10.021\\ T_{\overline{X}(S)}&=10\\ \text{LSL}&=10-0.0\bar{3}A_3(n_i)=1-+0.0\bar{3}A_3(6)=9.979\end{aligned} \tag{16.7.802}$$

With Minitab 17.1.3 this gives for people that don't trust our calculations made with Microsoft® Excel® (we can also do such a chart with `R` but the plot quality output is not acceptable in most corporations):


Xbar-S Chart of Ech. 1, ..., Ech. 6
Xbar-S Chart of Ech. 1, ..., Ech. 6
Sample Mean
10.050
10.025
10.000
9.975
9.950
UCL=10.0464
X̿=9.9933
LCL=9.9402
1 3 5 7 9 11 13 15 17 19 21 23 25
Sample
Sample StDev
0.08
0.06
0.04
0.02
0.00
UCL=0.08126
S̄=0.04126
LCL=0.00125
1 3 5 7 9 11 13 15 17 19 21 23 25
Sample

Let us notice that this control chart would according to the WECO rules have 1 point beyond the $\pm 3\sigma$ (we indicate that for reasons of comparison with the previous control chart and the next one). It is therefore more sensitive to changes than the control chart. Also notice that the UCL and LCL limits are wider than the imposed USL and LSL.

For recall, the major problem with this control chart is that it is correct if and only if the data are Normally distributed and that Shewhart limits suppose that S is symmetric and here in fact it is note case (if we took probabilistic limits). Therefore, the quality manager then combine with the average control chart another optional control chart which we will introduce now.

16.7.6.4.6 $\overline{X} - S_P$ Subgroups measurement control chart with pooled variance

As we have seen it the standard error based control chart suppose all measurements as whole (that means we suppose that each subgroup has the same mean). This is in fact not really accurate as every group is supposed as being independent and has to be compared with the others. So identically at the T-test or ANOVA for independent groups we use the "pooled variance" as overall indicator and as a generalization of the standard error (the pooled estimate the variance of several different populations when the mean may be different[37], but one may assume that the variance of each population is for recall the same!!!).

Therefore the pooled variance control chart is based on the following relation proved in the section Statistics during our study of ANOVA (see page 1001):

---

[37] Therefore it increase the power of a corresponding test

$$\mathrm{V}_R = \frac{\sum\limits_{ij}(x_{ij}-\bar{x}_i)^2}{\sum\limits_i (n_i - 1)} \tag{16.7.803}$$

or in a more common obvious notation for control charts:

$$\sigma_R = \sqrt{\frac{\sum\limits_{ij}(x_{ij}-\bar{x}_i)^2}{\sum\limits_i (n_i - 1)}} \tag{16.7.804}$$

But as above... we can use if we want the unbiasing constant c_4 of the variance as it applies similarly at the difference that (the notation is a little bit sad):

$$c_4(n) = c_4\left(1+\sum_i (n_i - 1)\right) \tag{16.7.805}$$

where the $1+$ is here because in the case where $n_i = 2$ (minimum case for the calculation of a variance) we would have without it $c_4(1)$ and that latter is not defined, that's why!

But in fact it's quite useless in the situation of the pooled variance. Normally unbiasing constant c_4 is for small samples... but when the parameter of the constant is $\sum_i (n_i - 1)$ then:

$$1+\sum_i (n_i - 1) \gg n_i \tag{16.7.806}$$

and it follows that:

$$c_4\left(1+\sum_i (n_i - 1)\right) \rightarrow 1 \tag{16.7.807}$$

To convinced yourself you can see below the calculated values of $c_4(n)$ using the relation constant given earlier above. As requested by some of my students here are some tabulated values:

n	c_4
1	$*$
2	0.797885
3	0.886227
4	0.921318
5	0.939986
10	0.972659
15	0.982316
20	0.986934
30	0.991418
40	0.993611
50	0.994911
70	0.996383
90	0.997195

Table 16.59: **Unbiasing constant c_4**

But as most of time the correction is smaller than the measurement precision... I let you therefore guess if it is very useful...

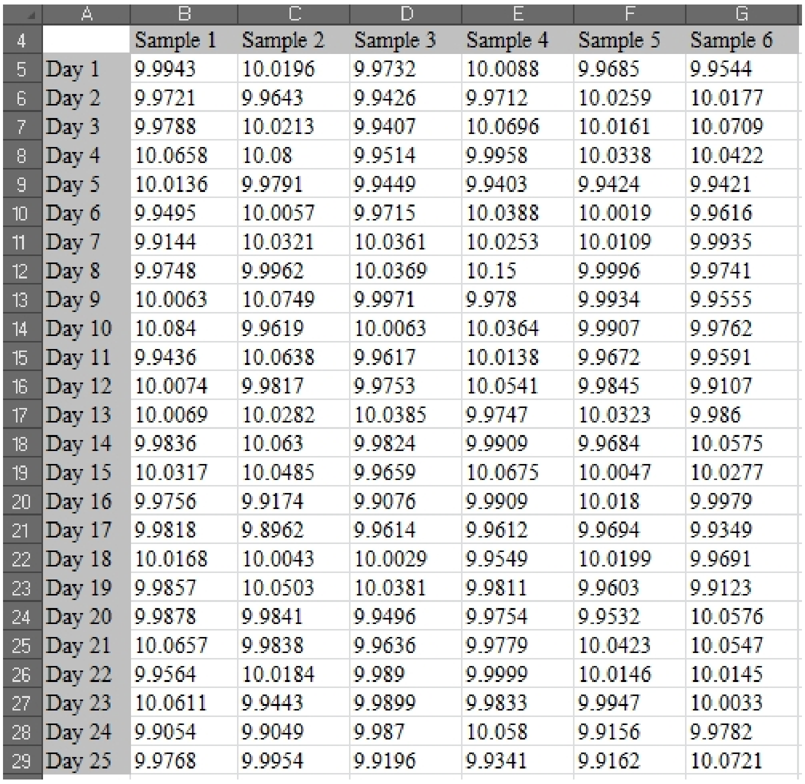
But we still can write (we like to write it by highlighting the fact that we are using the pooled variance so we avoid to use too much condensed notations!):

$$\boxed{\begin{aligned}\text{USL}&=\text{required by customer/board comitee}\\ \text{UCL}_{\overline{X}(S)}&=\overline{\overline{X}}+3\frac{\sigma_X}{\sqrt{n_i}}=\overline{\overline{X}}+3\frac{\sqrt{\frac{\sum_{ij}(x_{ij}-\bar{x}_i)^2}{\sum_i(n_i-1)}}}{c_4\left(1+\sum(n_i-1)\right)\sqrt{n_i}}\\ \text{CL}_{\overline{X}(S)}&=\overline{\overline{X}}\\ T_{\overline{X}(S)}&=\text{target required by customer/board comitee}\\ \text{LCL}_{\overline{X}(S)}&=\overline{\overline{X}}-3\frac{\sigma_X}{\sqrt{n_i}}=\overline{\overline{X}}-3\frac{\sqrt{\frac{\sum_{ij}(x_{ij}-\bar{x}_i)^2}{\sum_i(n_i-1)}}}{c_4\left(1+\sum(n_i-1)\right)\sqrt{n_i}}\\ \text{LSL}&=\text{required by customer/board comitee}\end{aligned}} \tag{16.7.808}$$

where we often write in practice:

$$\text{df}=c_4\left(1+\sum(n_i-1)\right) \tag{16.7.809}$$

As companion example let us consider the following example of 25 samples of 6 individuals each (that we will identify by the variable $n_i=n$) taken from a continuous manufacturing process (without machine recalibration or without machine change!):



	A	B	C	D	E	F	G
4		Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
5	Day 1	9.9943	10.0196	9.9732	10.0088	9.9685	9.9544
6	Day 2	9.9721	9.9643	9.9426	9.9712	10.0259	10.0177
7	Day 3	9.9788	10.0213	9.9407	10.0696	10.0161	10.0709
8	Day 4	10.0658	10.08	9.9514	9.9958	10.0338	10.0422
9	Day 5	10.0136	9.9791	9.9449	9.9403	9.9424	9.9421
10	Day 6	9.9495	10.0057	9.9715	10.0388	10.0019	9.9616
11	Day 7	9.9144	10.0321	10.0361	10.0253	10.0109	9.9935
12	Day 8	9.9748	9.9962	10.0369	10.15	9.9996	9.9741
13	Day 9	10.0063	10.0749	9.9971	9.978	9.9934	9.9555
14	Day 10	10.084	9.9619	10.0063	10.0364	9.9907	9.9762
15	Day 11	9.9436	10.0638	9.9617	10.0138	9.9672	9.9591
16	Day 12	10.0074	9.9817	9.9753	10.0541	9.9845	9.9107
17	Day 13	10.0069	10.0282	10.0385	9.9747	10.0323	9.986
18	Day 14	9.9836	10.063	9.9824	9.9909	9.9684	10.0575
19	Day 15	10.0317	10.0485	9.9659	10.0675	10.0047	10.0277
20	Day 16	9.9756	9.9174	9.9076	9.9909	10.018	9.9979
21	Day 17	9.9818	9.8962	9.9614	9.9612	9.9694	9.9349
22	Day 18	10.0168	10.0043	10.0029	9.9549	10.0199	9.9691
23	Day 19	9.9857	10.0503	10.0381	9.9811	9.9603	9.9123
24	Day 20	9.9878	9.9841	9.9496	9.9754	9.9532	10.0576
25	Day 21	10.0657	9.9838	9.9636	9.9779	10.0423	10.0547
26	Day 22	9.9564	10.0184	9.989	9.9999	10.0146	10.0145
27	Day 23	10.0611	9.9443	9.9899	9.9833	9.9947	10.0033
28	Day 24	9.9054	9.9049	9.987	10.058	9.9156	9.9782
29	Day 25	9.9768	9.9954	9.9196	9.9341	9.9162	10.0721

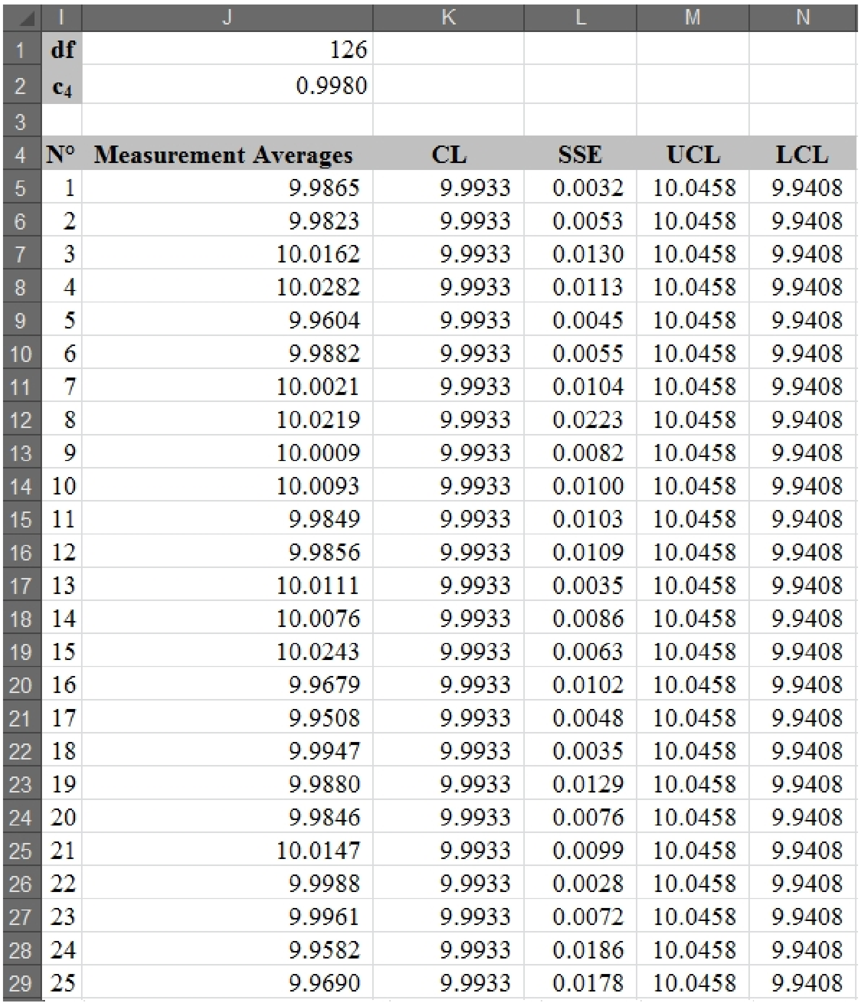
and let us focus on an example with the use of the unbiasing constant and always providing a detailed Microsoft® Excel® example and showing afterwards the corresponding Minitab plot!

So every day we took $n_i = 6$ individuals of all manufactured items produced the same day.

We can calculate, or find in the tables that:

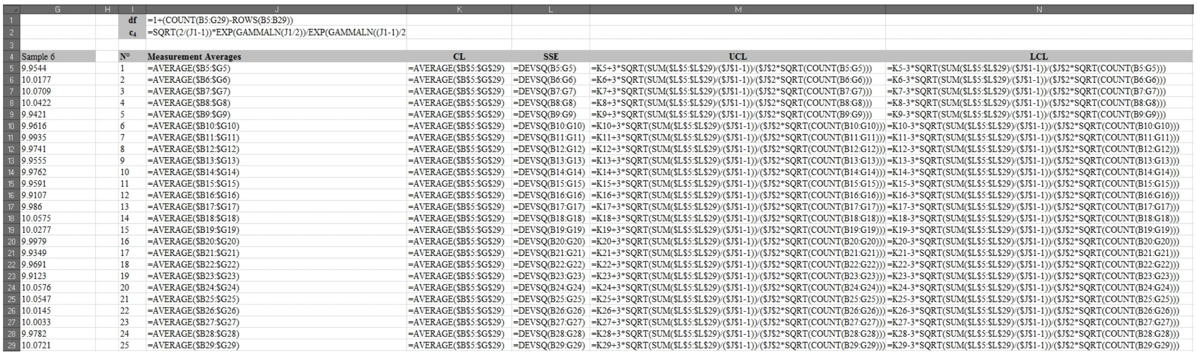
$$c_4(1+\sum(n_i-1)) = c_4(126) = 0.9980 \qquad (16.7.810)$$

We then have the following table:



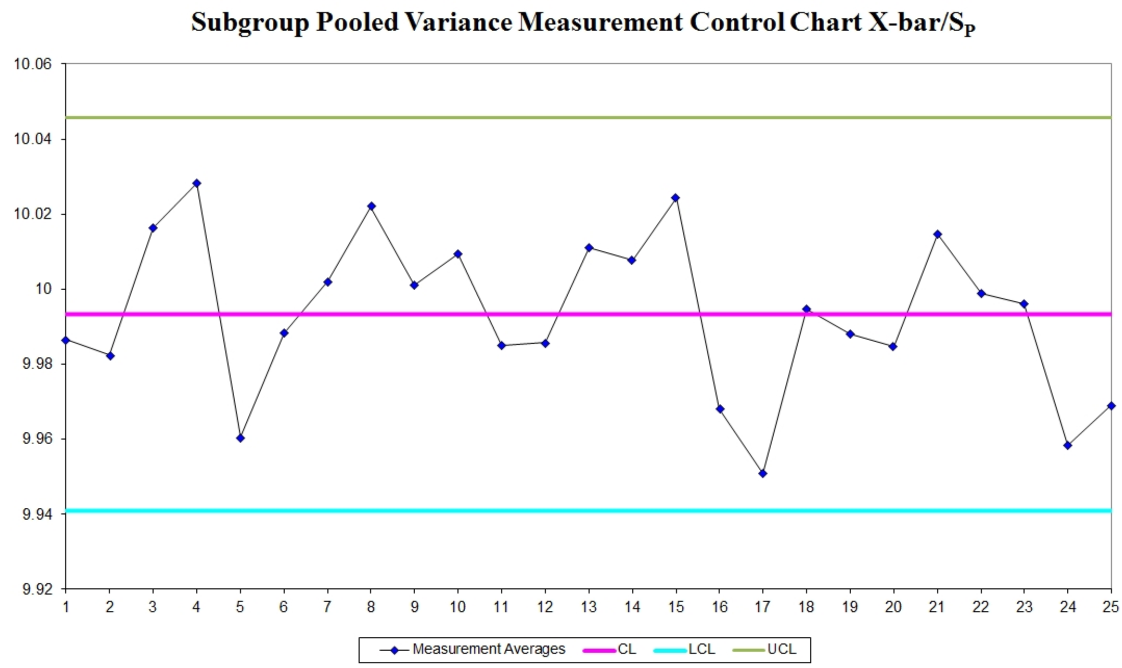
	I	J	K	L	M	N
1	**df**	126				
2	**c$_4$**	0.9980				
3						
4	**N°**	**Measurement Averages**	**CL**	**SSE**	**UCL**	**LCL**
5	1	9.9865	9.9933	0.0032	10.0458	9.9408
6	2	9.9823	9.9933	0.0053	10.0458	9.9408
7	3	10.0162	9.9933	0.0130	10.0458	9.9408
8	4	10.0282	9.9933	0.0113	10.0458	9.9408
9	5	9.9604	9.9933	0.0045	10.0458	9.9408
10	6	9.9882	9.9933	0.0055	10.0458	9.9408
11	7	10.0021	9.9933	0.0104	10.0458	9.9408
12	8	10.0219	9.9933	0.0223	10.0458	9.9408
13	9	10.0009	9.9933	0.0082	10.0458	9.9408
14	10	10.0093	9.9933	0.0100	10.0458	9.9408
15	11	9.9849	9.9933	0.0103	10.0458	9.9408
16	12	9.9856	9.9933	0.0109	10.0458	9.9408
17	13	10.0111	9.9933	0.0035	10.0458	9.9408
18	14	10.0076	9.9933	0.0086	10.0458	9.9408
19	15	10.0243	9.9933	0.0063	10.0458	9.9408
20	16	9.9679	9.9933	0.0102	10.0458	9.9408
21	17	9.9508	9.9933	0.0048	10.0458	9.9408
22	18	9.9947	9.9933	0.0035	10.0458	9.9408
23	19	9.9880	9.9933	0.0129	10.0458	9.9408
24	20	9.9846	9.9933	0.0076	10.0458	9.9408
25	21	10.0147	9.9933	0.0099	10.0458	9.9408
26	22	9.9988	9.9933	0.0028	10.0458	9.9408
27	23	9.9961	9.9933	0.0072	10.0458	9.9408
28	24	9.9582	9.9933	0.0186	10.0458	9.9408
29	25	9.9690	9.9933	0.0178	10.0458	9.9408

or explicitly (the reader can zoom on it with its PDF reader):

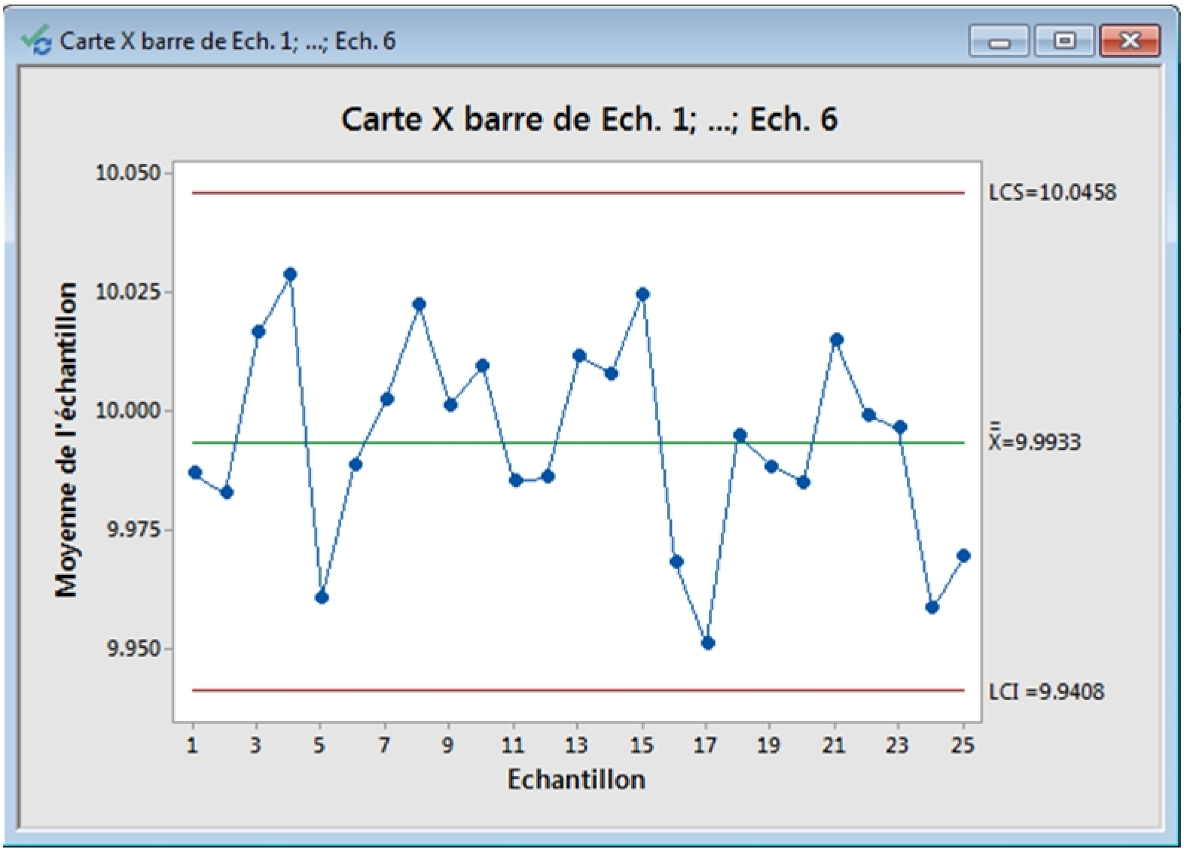


	G	H	I	J	K	L	M	N
1			**df**	=1+(COUNT(B5:G29)-ROWS(B5:B29))				
2			**c$_4$**	=SQRT(2/(J1-1))*EXP(GAMMALN(J1/2))/EXP(GAMMALN((J1-1)/2				
3								
4	Sample 6		**N°**	**Measurement Averages**	**CL**	**SSE**	**UCL**	**LCL**
5	9.9544		1	=AVERAGE($B5:$G5)	=AVERAGE($B$5:$G$29)	=DEVSQ(B5:G5)	=K5+3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B5:G5)))	=K5-3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B5:G5)))
6	10.0177		2	=AVERAGE($B6:$G6)	=AVERAGE($B$5:$G$29)	=DEVSQ(B6:G6)	=K6+3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B6:G6)))	=K6-3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B6:G6)))
7	10.0709		3	=AVERAGE($B7:$G7)	=AVERAGE($B$5:$G$29)	=DEVSQ(B7:G7)	=K7+3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B7:G7)))	=K7-3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B7:G7)))
8	10.0422		4	=AVERAGE($B8:$G8)	=AVERAGE($B$5:$G$29)	=DEVSQ(B8:G8)	=K8+3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B8:G8)))	=K8-3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B8:G8)))
9	9.9421		5	=AVERAGE($B9:$G9)	=AVERAGE($B$5:$G$29)	=DEVSQ(B9:G9)	=K9+3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B9:G9)))	=K9-3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B9:G9)))
10	9.9616		6	=AVERAGE($B10:$G10)	=AVERAGE($B$5:$G$29)	=DEVSQ(B10:G10)	=K10+3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B10:G10)))	=K10-3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B10:G10)))
11	9.9935		7	=AVERAGE($B11:$G11)	=AVERAGE($B$5:$G$29)	=DEVSQ(B11:G11)	=K11+3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B11:G11)))	=K11-3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B11:G11)))
12	9.9741		8	=AVERAGE($B12:$G12)	=AVERAGE($B$5:$G$29)	=DEVSQ(B12:G12)	=K12+3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B12:G12)))	=K12-3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B12:G12)))
13	9.9555		9	=AVERAGE($B13:$G13)	=AVERAGE($B$5:$G$29)	=DEVSQ(B13:G13)	=K13+3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B13:G13)))	=K13-3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B13:G13)))
14	9.9762		10	=AVERAGE($B14:$G14)	=AVERAGE($B$5:$G$29)	=DEVSQ(B14:G14)	=K14+3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B14:G14)))	=K14-3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B14:G14)))
15	9.9591		11	=AVERAGE($B15:$G15)	=AVERAGE($B$5:$G$29)	=DEVSQ(B15:G15)	=K15+3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B15:G15)))	=K15-3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B15:G15)))
16	9.9107		12	=AVERAGE($B16:$G16)	=AVERAGE($B$5:$G$29)	=DEVSQ(B16:G16)	=K16+3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B16:G16)))	=K16-3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B16:G16)))
17	9.986		13	=AVERAGE($B17:$G17)	=AVERAGE($B$5:$G$29)	=DEVSQ(B17:G17)	=K17+3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B17:G17)))	=K17-3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B17:G17)))
18	10.0575		14	=AVERAGE($B18:$G18)	=AVERAGE($B$5:$G$29)	=DEVSQ(B18:G18)	=K18+3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B18:G18)))	=K18-3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B18:G18)))
19	10.0277		15	=AVERAGE($B19:$G19)	=AVERAGE($B$5:$G$29)	=DEVSQ(B19:G19)	=K19+3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B19:G19)))	=K19-3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B19:G19)))
20	9.9979		16	=AVERAGE($B20:$G20)	=AVERAGE($B$5:$G$29)	=DEVSQ(B20:G20)	=K20+3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B20:G20)))	=K20-3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B20:G20)))
21	9.9349		17	=AVERAGE($B21:$G21)	=AVERAGE($B$5:$G$29)	=DEVSQ(B21:G21)	=K21+3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B21:G21)))	=K21-3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B21:G21)))
22	9.9691		18	=AVERAGE($B22:$G22)	=AVERAGE($B$5:$G$29)	=DEVSQ(B22:G22)	=K22+3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B22:G22)))	=K22-3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B22:G22)))
23	9.9123		19	=AVERAGE($B23:$G23)	=AVERAGE($B$5:$G$29)	=DEVSQ(B23:G23)	=K23+3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B23:G23)))	=K23-3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B23:G23)))
24	10.0576		20	=AVERAGE($B24:$G24)	=AVERAGE($B$5:$G$29)	=DEVSQ(B24:G24)	=K24+3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B24:G24)))	=K24-3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B24:G24)))
25	10.0547		21	=AVERAGE($B25:$G25)	=AVERAGE($B$5:$G$29)	=DEVSQ(B25:G25)	=K25+3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B25:G25)))	=K25-3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B25:G25)))
26	10.0145		22	=AVERAGE($B26:$G26)	=AVERAGE($B$5:$G$29)	=DEVSQ(B26:G26)	=K26+3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B26:G26)))	=K26-3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B26:G26)))
27	10.0033		23	=AVERAGE($B27:$G27)	=AVERAGE($B$5:$G$29)	=DEVSQ(B27:G27)	=K27+3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B27:G27)))	=K27-3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B27:G27)))
28	9.9782		24	=AVERAGE($B28:$G28)	=AVERAGE($B$5:$G$29)	=DEVSQ(B28:G28)	=K28+3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B28:G28)))	=K28-3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B28:G28)))
29	10.0721		25	=AVERAGE($B29:$G29)	=AVERAGE($B$5:$G$29)	=DEVSQ(B29:G29)	=K29+3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B29:G29)))	=K29-3*SQRT(SUM($L$5:$L$29)/($J$1-1))/($J$2*SQRT(COUNT(B29:G29)))

and the corresponding chart still with the same spreadsheet software:



With Minitab 17.1.3 this gives for people that don't trust our calculations made with Microsoft® Excel® (we can also do such a chart with `R` but the plot quality output is not acceptable in most corporations):


Carte X barre de Ech. 1; ...; Ech. 6
Carte X barre de Ech. 1; ...; Ech. 6
10.050
10.025
10.000
9.975
9.950
Moyenne de l'échantillon
LCS=10.0458
$\bar{\bar{X}}$=9.9933
LCI =9.9408
1 3 5 7 9 11 13 15 17 19 21 23 25
Echantillon

Let us notice that this control chart would according to the WECO rules have 0 point beyond the $\pm 3\sigma$ (we indicate that for reasons of comparison with the previous control chart and the next one).

For recall, the major problem with this control chart is that it is correct if and only if the data are Normally distributed!

16.7.6.4.7 $\overline{R}-R$ Subgroups measurement control chart

The range control chart is often associated with the average control chart. It aims to advantageously replace the standard deviation control chart $\overline{S}-S$ when data are not Normally distributed, because it's mathematical construction allows if desired to choose the underlying probability distribution! So we use this new control chart when $n_i \leq 10$ (the number of individuals per sample), while for the control chart for the standard deviation we must instead $n_i \geq 10$.

Pros:

- Completes well the average control chart and replaces well the control chart of the standard deviation when the number of samples is small.
- Represents well the variation within each sample of data (instantaneous dispersion) that we could detect.

Cons:

- The mathematical proof of the origin of the limits of controls is quite hard (see the section of Statistics for the detailed proof) and the determination of the corresponding unbiasing coefficients requires the use of Monte Carlo simulations (see section Numerical Methods page 6708).
- Another problem is that even if the probability distribution of sample size can be chosen, it is often given in specialized books only for the Normal distribution... (but softwares such as Minitab manage them).

The idea is to represent (due to the possible asymmetry of R it is impossible - nonsense - that the customer sets coherent USL and LSL or a target T as regards to the range, hence the fact that these parameters will this time not be listed or represented):

$$\begin{aligned}\text{UCL}_R &= \overline{R}+3\sigma_R\\ \text{CL}_R &= \text{E}(R)=\overline{R}\\ \text{LCL}_R &= \overline{R}-3\sigma_R\end{aligned} \tag{16.7.811}$$

and this time we will not use estimators of the standard deviation for the range as it does not converge fast enough. Therefore, we will use the results proved in the section Statistics during our study of range statistics (see section Statistics page 1203). We obtained for the estimator of the standard deviation with the unbiasing Hartley's constants:

$$\boxed{\sigma_R=\frac{d_3(n)\overline{R}}{d_2(n)}=\frac{d_3\overline{R}}{d_2}} \tag{16.7.812}$$

with:

$$\overline{R}=\frac{1}{k}\sum_{i=1}^{k}R_i \tag{16.7.813}$$

or depending of the softwares (the latter one is more logic if we think a little bit):

$$\overline{R}=\frac{1}{k-1}\sum_{i=1}^{k}R_i \tag{16.7.814}$$

Therefore:

$$\begin{aligned}\text{UCL}_R&=\overline{R}+3\sigma_R=\overline{R}+3\frac{d_3(n_i)\overline{R}}{d_2(n_i)}=\overline{R}\left(1+3\frac{d_3(n_i)\overline{R}}{d_2(n_i)}\right)\\ \text{CL}_R&=\text{E}(R)=\overline{R}\\ \text{LCL}_R&=\overline{R}-3\sigma_R=\overline{R}-3\frac{d_3(n_i)\overline{R}}{d_2(n_i)}=\overline{R}\left(1-3\frac{d_3(n_i)\overline{R}}{d_2(n_i)}\right)\end{aligned} \tag{16.7.815}$$

which is written traditionally:

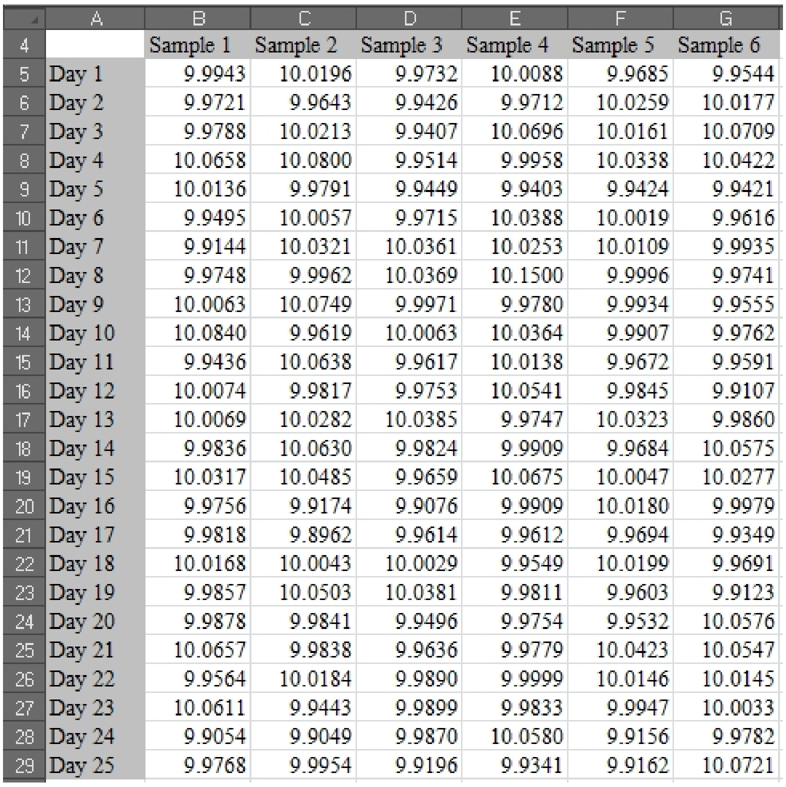
$$\boxed{\begin{aligned}\text{UCL}_R&=\overline{R}+3\sigma_R=\overline{R}+3\frac{d_3(n_i)\overline{R}}{d_2(n_i)}=\overline{R}D_4(n_i)\\ \text{CL}_R&=\text{E}(R)=\overline{R}\\ \text{LCL}_R&=\overline{R}-3\sigma_R=\overline{R}-3\frac{d_3(n_i)\overline{R}}{d_2(n_i)}=\overline{R}D_3(n_i)\end{aligned}} \tag{16.7.816}$$

The unbiasing coefficients D_3, D_4 defined by:

$$\boxed{D_4=1+3\frac{d_3(n_i)\overline{R}}{d_2(n_i)}\qquad D_3=1-3\frac{d_3(n_i)\overline{R}}{d_2(n_i)}} \tag{16.7.817}$$

are available in the literature (or on Internet or also in the Help of Minitab) as a function of the n_i of the sample size and often (unfortunately) only for a Normal distribution (see the example in the section Statistics).

As companion example let us consider the following example of 25 samples (that we will identify by the variable k) of 6 individuals each (that we will identify by the variable $n_i=n$) taken from a continuous manufacturing process (without machine recalibration or without machine change!):



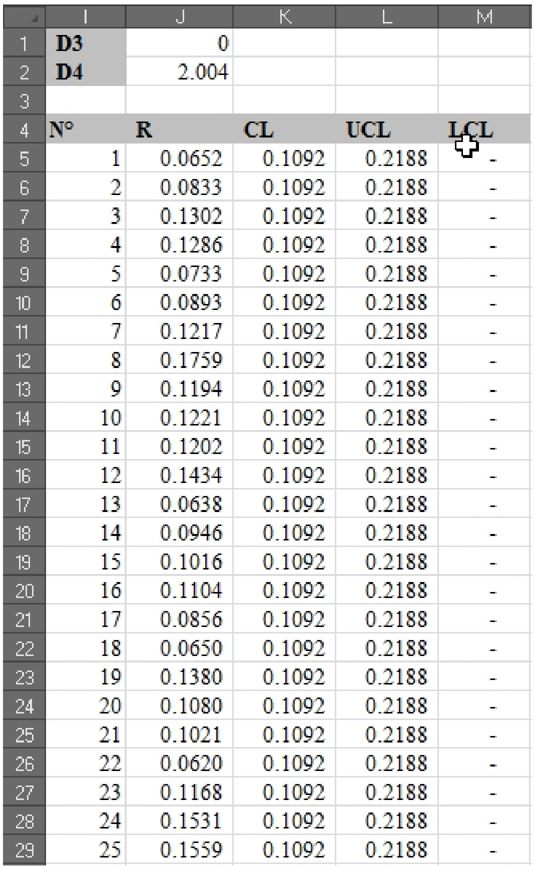
	A	B	C	D	E	F	G
4		Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
5	Day 1	9.9943	10.0196	9.9732	10.0088	9.9685	9.9544
6	Day 2	9.9721	9.9643	9.9426	9.9712	10.0259	10.0177
7	Day 3	9.9788	10.0213	9.9407	10.0696	10.0161	10.0709
8	Day 4	10.0658	10.0800	9.9514	9.9958	10.0338	10.0422
9	Day 5	10.0136	9.9791	9.9449	9.9403	9.9424	9.9421
10	Day 6	9.9495	10.0057	9.9715	10.0388	10.0019	9.9616
11	Day 7	9.9144	10.0321	10.0361	10.0253	10.0109	9.9935
12	Day 8	9.9748	9.9962	10.0369	10.1500	9.9996	9.9741
13	Day 9	10.0063	10.0749	9.9971	9.9780	9.9934	9.9555
14	Day 10	10.0840	9.9619	10.0063	10.0364	9.9907	9.9762
15	Day 11	9.9436	10.0638	9.9617	10.0138	9.9672	9.9591
16	Day 12	10.0074	9.9817	9.9753	10.0541	9.9845	9.9107
17	Day 13	10.0069	10.0282	10.0385	9.9747	10.0323	9.9860
18	Day 14	9.9836	10.0630	9.9824	9.9909	9.9684	10.0575
19	Day 15	10.0317	10.0485	9.9659	10.0675	10.0047	10.0277
20	Day 16	9.9756	9.9174	9.9076	9.9909	10.0180	9.9979
21	Day 17	9.9818	9.8962	9.9614	9.9612	9.9694	9.9349
22	Day 18	10.0168	10.0043	10.0029	9.9549	10.0199	9.9691
23	Day 19	9.9857	10.0503	10.0381	9.9811	9.9603	9.9123
24	Day 20	9.9878	9.9841	9.9496	9.9754	9.9532	10.0576
25	Day 21	10.0657	9.9838	9.9636	9.9779	10.0423	10.0547
26	Day 22	9.9564	10.0184	9.9890	9.9999	10.0146	10.0145
27	Day 23	10.0611	9.9443	9.9899	9.9833	9.9947	10.0033
28	Day 24	9.9054	9.9049	9.9870	10.0580	9.9156	9.9782
29	Day 25	9.9768	9.9954	9.9196	9.9341	9.9162	10.0721

where every day we took a sample 6 individuals of all manufactured items produced the same day.

We can find in the tables that:

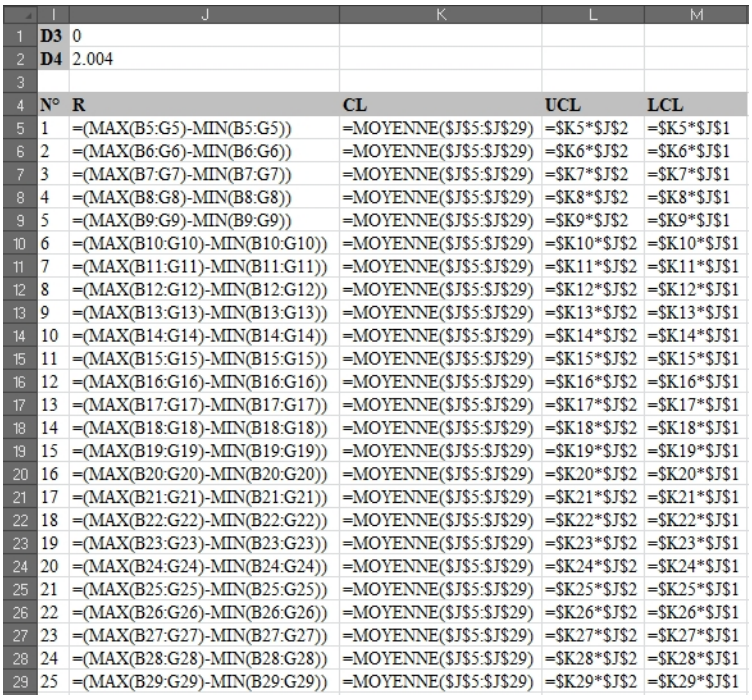
$$D_3(n_i)=D_3(6)=0 \qquad D_4(n_i)=D_4(6)=2.004 \tag{16.7.818}$$

We then have the following table:



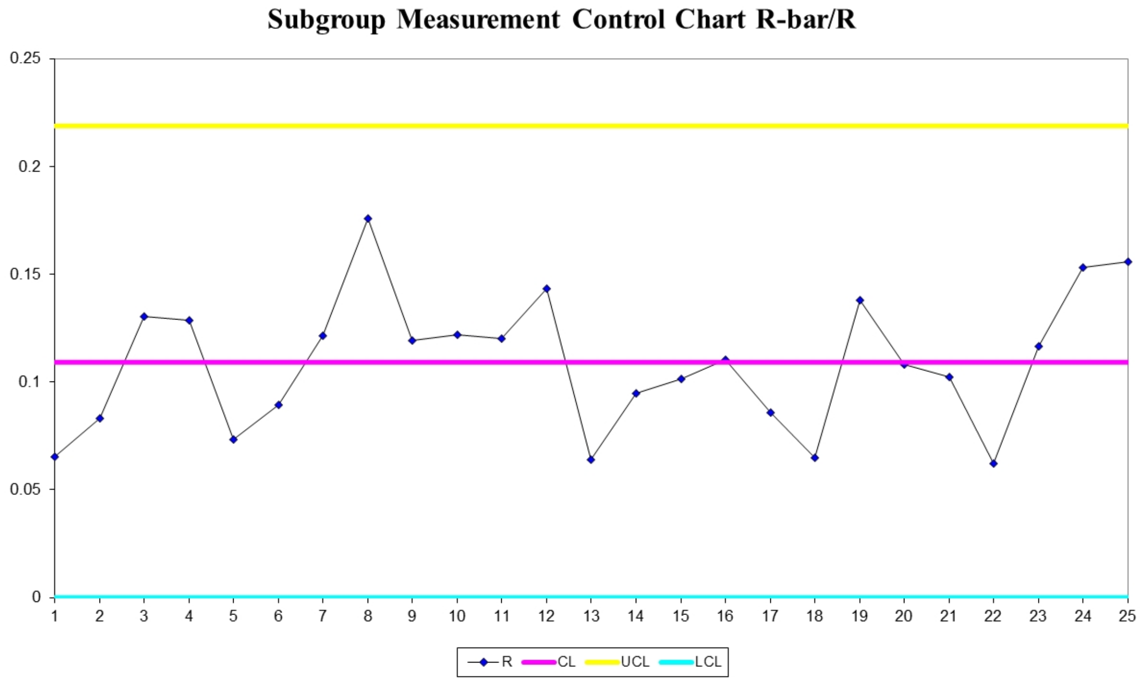
	I	J	K	L	M
1	**D3**	0			
2	**D4**	2.004			
3					
4	**N°**	**R**	**CL**	**UCL**	**LCL**
5	1	0.0652	0.1092	0.2188	-
6	2	0.0833	0.1092	0.2188	-
7	3	0.1302	0.1092	0.2188	-
8	4	0.1286	0.1092	0.2188	-
9	5	0.0733	0.1092	0.2188	-
10	6	0.0893	0.1092	0.2188	-
11	7	0.1217	0.1092	0.2188	-
12	8	0.1759	0.1092	0.2188	-
13	9	0.1194	0.1092	0.2188	-
14	10	0.1221	0.1092	0.2188	-
15	11	0.1202	0.1092	0.2188	-
16	12	0.1434	0.1092	0.2188	-
17	13	0.0638	0.1092	0.2188	-
18	14	0.0946	0.1092	0.2188	-
19	15	0.1016	0.1092	0.2188	-
20	16	0.1104	0.1092	0.2188	-
21	17	0.0856	0.1092	0.2188	-
22	18	0.0650	0.1092	0.2188	-
23	19	0.1380	0.1092	0.2188	-
24	20	0.1080	0.1092	0.2188	-
25	21	0.1021	0.1092	0.2188	-
26	22	0.0620	0.1092	0.2188	-
27	23	0.1168	0.1092	0.2188	-
28	24	0.1531	0.1092	0.2188	-
29	25	0.1559	0.1092	0.2188	-

or explicitly:

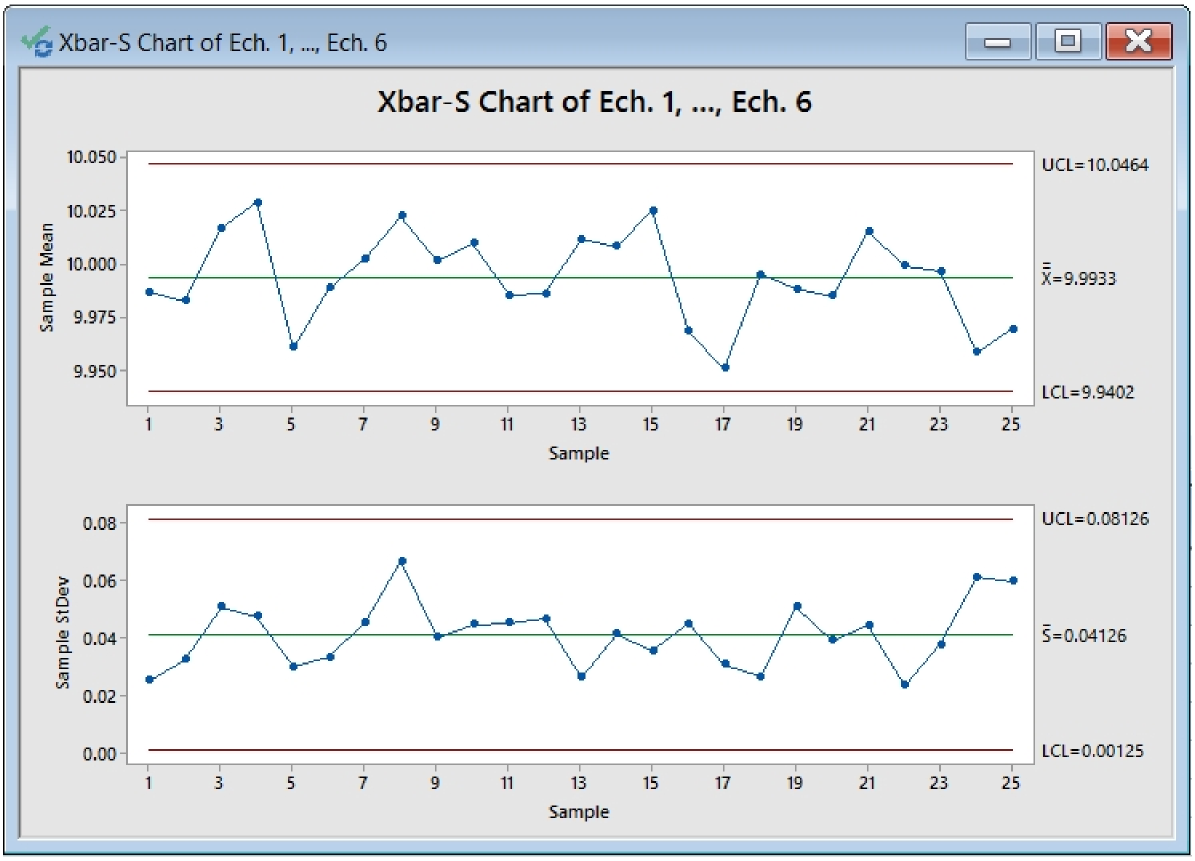


	I	J	K	L	M
1	**D3**	0			
2	**D4**	2.004			
3					
4	**N°**	**R**	**CL**	**UCL**	**LCL**
5	1	=(MAX(B5:G5)-MIN(B5:G5))	=MOYENNE($J$5:$J$29)	=$K5*$J$2	=$K5*$J$1
6	2	=(MAX(B6:G6)-MIN(B6:G6))	=MOYENNE($J$5:$J$29)	=$K6*$J$2	=$K6*$J$1
7	3	=(MAX(B7:G7)-MIN(B7:G7))	=MOYENNE($J$5:$J$29)	=$K7*$J$2	=$K7*$J$1
8	4	=(MAX(B8:G8)-MIN(B8:G8))	=MOYENNE($J$5:$J$29)	=$K8*$J$2	=$K8*$J$1
9	5	=(MAX(B9:G9)-MIN(B9:G9))	=MOYENNE($J$5:$J$29)	=$K9*$J$2	=$K9*$J$1
10	6	=(MAX(B10:G10)-MIN(B10:G10))	=MOYENNE($J$5:$J$29)	=$K10*$J$2	=$K10*$J$1
11	7	=(MAX(B11:G11)-MIN(B11:G11))	=MOYENNE($J$5:$J$29)	=$K11*$J$2	=$K11*$J$1
12	8	=(MAX(B12:G12)-MIN(B12:G12))	=MOYENNE($J$5:$J$29)	=$K12*$J$2	=$K12*$J$1
13	9	=(MAX(B13:G13)-MIN(B13:G13))	=MOYENNE($J$5:$J$29)	=$K13*$J$2	=$K13*$J$1
14	10	=(MAX(B14:G14)-MIN(B14:G14))	=MOYENNE($J$5:$J$29)	=$K14*$J$2	=$K14*$J$1
15	11	=(MAX(B15:G15)-MIN(B15:G15))	=MOYENNE($J$5:$J$29)	=$K15*$J$2	=$K15*$J$1
16	12	=(MAX(B16:G16)-MIN(B16:G16))	=MOYENNE($J$5:$J$29)	=$K16*$J$2	=$K16*$J$1
17	13	=(MAX(B17:G17)-MIN(B17:G17))	=MOYENNE($J$5:$J$29)	=$K17*$J$2	=$K17*$J$1
18	14	=(MAX(B18:G18)-MIN(B18:G18))	=MOYENNE($J$5:$J$29)	=$K18*$J$2	=$K18*$J$1
19	15	=(MAX(B19:G19)-MIN(B19:G19))	=MOYENNE($J$5:$J$29)	=$K19*$J$2	=$K19*$J$1
20	16	=(MAX(B20:G20)-MIN(B20:G20))	=MOYENNE($J$5:$J$29)	=$K20*$J$2	=$K20*$J$1
21	17	=(MAX(B21:G21)-MIN(B21:G21))	=MOYENNE($J$5:$J$29)	=$K21*$J$2	=$K21*$J$1
22	18	=(MAX(B22:G22)-MIN(B22:G22))	=MOYENNE($J$5:$J$29)	=$K22*$J$2	=$K22*$J$1
23	19	=(MAX(B23:G23)-MIN(B23:G23))	=MOYENNE($J$5:$J$29)	=$K23*$J$2	=$K23*$J$1
24	20	=(MAX(B24:G24)-MIN(B24:G24))	=MOYENNE($J$5:$J$29)	=$K24*$J$2	=$K24*$J$1
25	21	=(MAX(B25:G25)-MIN(B25:G25))	=MOYENNE($J$5:$J$29)	=$K25*$J$2	=$K25*$J$1
26	22	=(MAX(B26:G26)-MIN(B26:G26))	=MOYENNE($J$5:$J$29)	=$K26*$J$2	=$K26*$J$1
27	23	=(MAX(B27:G27)-MIN(B27:G27))	=MOYENNE($J$5:$J$29)	=$K27*$J$2	=$K27*$J$1
28	24	=(MAX(B28:G28)-MIN(B28:G28))	=MOYENNE($J$5:$J$29)	=$K28*$J$2	=$K28*$J$1
29	25	=(MAX(B29:G29)-MIN(B29:G29))	=MOYENNE($J$5:$J$29)	=$K29*$J$2	=$K29*$J$1

and the corresponding chart still with the same spreadsheet software:



With Minitab 17.1.3 this gives for people that don't trust our calculations made with Microsoft® Excel® (we can also do such a chart with `R` but the plot quality output is not acceptable in most corporations):


Xbar-S Chart of Ech. 1, ..., Ech. 6
Xbar-S Chart of Ech. 1, ..., Ech. 6
Sample Mean
10.050
10.025
10.000
9.975
9.950
UCL=10.0464
$\bar{\bar{X}}$=9.9933
LCL=9.9402
1 3 5 7 9 11 13 15 17 19 21 23 25
Sample
Sample StDev
0.08
0.06
0.04
0.02
0.00
UCL=0.08126
$\bar{S}$=0.04126
LCL=0.00125
1 3 5 7 9 11 13 15 17 19 21 23 25
Sample

which (is it necessary to recall it?) underestimates the number of error detections since the probability distribution of the range is in reality not symmetrical (hence the application of fixed Shewhart's control limits is not really suited...) and because almost all of the tables assume the data still Normally distributed as well.

Let us notice that this control chart would according to the WECO rules have 0 point beyond the $\pm 3\sigma$.

16.7.6.4.8 $\overline{X}-R$ Subgroups measurement control chart

During our study of the averages control chart averages with fixed limits based on the standard deviation of the mean and also of the $\overline{S}-S$ control chart, we have clearly discusses that it suffered from the handicap that σ_X is too approximate in practice (because n_i is often too small).

Therefore, a workaround is to use the results obtained from the study of the previous $\overline{R}-R$ control chart.

Pros:

- Corrects the problem of the control charge with for averages using the traditional standard deviation or also the unbiasing standard deviation approach when the number of individuals per sample is quite small.
- Represents well the change in the average level of data (central tendency).

Cons:

- The mathematical proof of the origin of the limits of controls is quite hard (see the section of Statistics at page 1203 for the detailed proof) and the determination of the corresponding unbiasing coefficients requires the use of Monte Carlo simulations (see section Numerical Methods page 6708).
- Another problem is that even if the probability distribution of sample size can be chosen, it is often given in specialized books only for the Normal distribution... (but softwares such as Minitab manage them).

Remark

This control chart is used a lot as part of the study of reproducibility and repeatability (R&R Studies). This is why it is found it often in statistical process control software as automatically processes control chart when launching a R&R analysis (typically what do Minitab).

Let us recall first that we had:

$$\begin{aligned}\text{USL} &= \text{required by customer/board comitee}\\ \text{UCL}_{\text{SE}} &= \overline{\overline{X}}+3\sigma_{\overline{X}}=\overline{\overline{X}}+3\frac{\sigma_X}{\sqrt{n_i}}\\ \text{CL}_{\text{SE}} &= \text{E}(\overline{X})=\overline{\overline{X}}\\ T_{\text{SE}} &= \text{target required by customer/board comitee}\\ \text{LCL}_{\text{SE}} &= \overline{\overline{X}}-3\sigma_{\overline{X}}=\overline{\overline{X}}-3\frac{\sigma_X}{\sqrt{n_i}}\\ \text{LSL} &= \text{required by customer/board comitee}\end{aligned} \tag{16.7.819}$$

and let us recall that we have proved in the section Statistics in the context of our study of range statistics (see page 1203) that:

$$\overline{R}=\sigma_X d_2 \tag{16.7.820}$$

It comes (it is impossible that the customer sets a limit for the USL, LSL or target T as regards the range as in general it is asymmetric, hence the fact that these parameters are this time not listed below):

$$\begin{aligned}\text{UCL}_{\overline{X}(R)} &= \overline{\overline{X}}+3\frac{\sigma_X}{\sqrt{n_i}}=\overline{\overline{X}}+3\frac{\overline{R}}{d_2(n_i)\sqrt{n_i}}\\ \text{CL}_{\overline{X}(R)} &= \text{E}(\overline{X})=\overline{\overline{X}}\\ \text{LCL}_{\overline{X}(R)} &= \overline{\overline{X}}-3\frac{\sigma_X}{\sqrt{n_i}}=\overline{\overline{X}}-3\frac{\overline{R}}{d_2(n_i)\sqrt{n_i}}\end{aligned} \tag{16.7.821}$$

still with:

$$\overline{R}=\frac{1}{k}\sum_{i=1}^{k}R_i \tag{16.7.822}$$

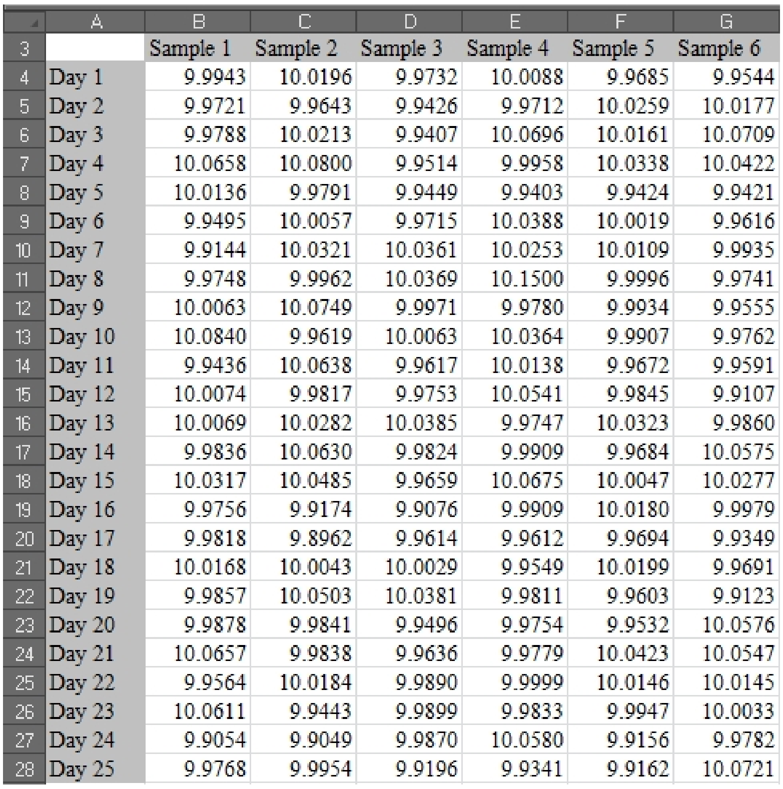
and the whole is traditionally written in the field:

$$\begin{aligned}\text{UCL}_{\overline{X}(R)} &= \overline{\overline{X}}+3\frac{\overline{R}}{d_2(n_i)\sqrt{n_i}}=\overline{\overline{X}}+A_2(n_i)\overline{R}\\ \text{CL}_{\overline{X}(R)} &= \text{E}(\overline{X})=\overline{\overline{X}}\\ \text{LCL}_{\overline{X}(R)} & \overline{\overline{X}}-3\frac{\overline{R}}{d_2(n_i)\sqrt{n_i}}=\overline{\overline{X}}-A_2(n_i)\overline{R}\end{aligned} \tag{16.7.823}$$

Notice that we can then calculate the probabilities of non-compliance by excess or defect of the subgroups by using the fact that $\overline{X}$ follows a Normal distribution of parameters:

$$\mathcal{N}(\overline{X},\sigma_{\overline{X}})\cong\mathcal{N}\left(\overline{\overline{X}},\frac{\overline{R}}{d_2(n_i)\sqrt{n_i}}\right) \tag{16.7.824}$$

As companion example let us consider the following example of 25 samples (that we will identify by the variable k) of 6 individuals each (that we will identify by the variable $n_i=n$) taken from a continuous manufacturing process (without machine recalibration or without machine change!):



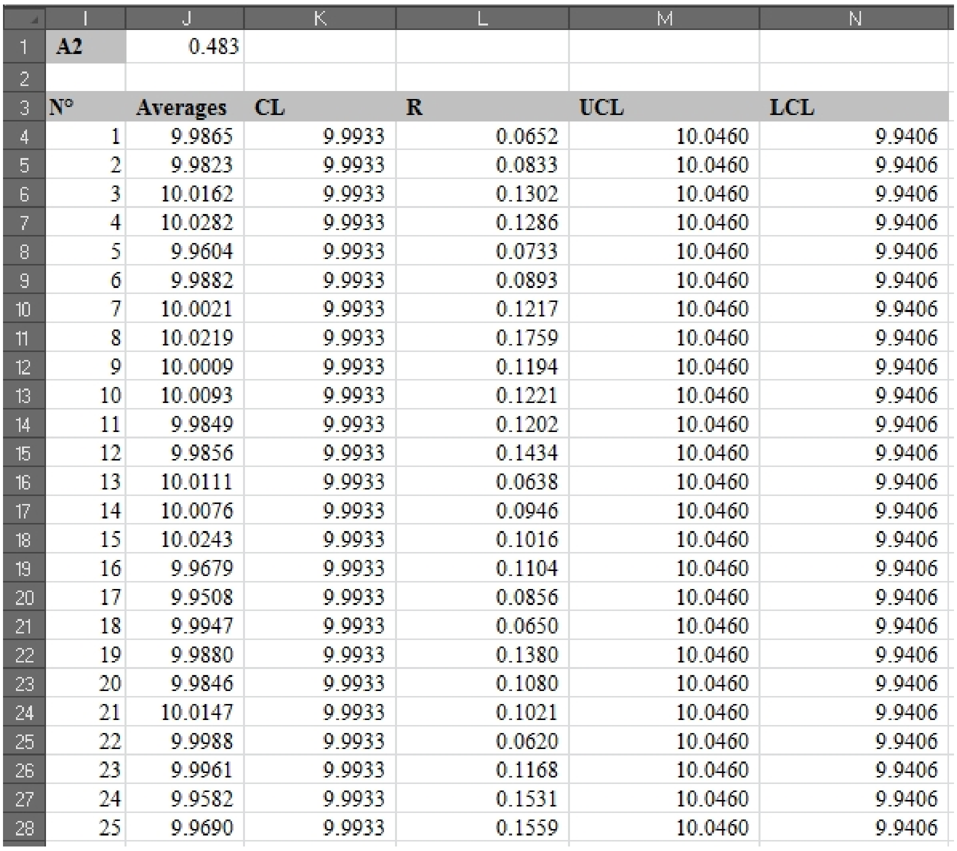
	A	B	C	D	E	F	G
3		Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
4	Day 1	9.9943	10.0196	9.9732	10.0088	9.9685	9.9544
5	Day 2	9.9721	9.9643	9.9426	9.9712	10.0259	10.0177
6	Day 3	9.9788	10.0213	9.9407	10.0696	10.0161	10.0709
7	Day 4	10.0658	10.0800	9.9514	9.9958	10.0338	10.0422
8	Day 5	10.0136	9.9791	9.9449	9.9403	9.9424	9.9421
9	Day 6	9.9495	10.0057	9.9715	10.0388	10.0019	9.9616
10	Day 7	9.9144	10.0321	10.0361	10.0253	10.0109	9.9935
11	Day 8	9.9748	9.9962	10.0369	10.1500	9.9996	9.9741
12	Day 9	10.0063	10.0749	9.9971	9.9780	9.9934	9.9555
13	Day 10	10.0840	9.9619	10.0063	10.0364	9.9907	9.9762
14	Day 11	9.9436	10.0638	9.9617	10.0138	9.9672	9.9591
15	Day 12	10.0074	9.9817	9.9753	10.0541	9.9845	9.9107
16	Day 13	10.0069	10.0282	10.0385	9.9747	10.0323	9.9860
17	Day 14	9.9836	10.0630	9.9824	9.9909	9.9684	10.0575
18	Day 15	10.0317	10.0485	9.9659	10.0675	10.0047	10.0277
19	Day 16	9.9756	9.9174	9.9076	9.9909	10.0180	9.9979
20	Day 17	9.9818	9.8962	9.9614	9.9612	9.9694	9.9349
21	Day 18	10.0168	10.0043	10.0029	9.9549	10.0199	9.9691
22	Day 19	9.9857	10.0503	10.0381	9.9811	9.9603	9.9123
23	Day 20	9.9878	9.9841	9.9496	9.9754	9.9532	10.0576
24	Day 21	10.0657	9.9838	9.9636	9.9779	10.0423	10.0547
25	Day 22	9.9564	10.0184	9.9890	9.9999	10.0146	10.0145
26	Day 23	10.0611	9.9443	9.9899	9.9833	9.9947	10.0033
27	Day 24	9.9054	9.9049	9.9870	10.0580	9.9156	9.9782
28	Day 25	9.9768	9.9954	9.9196	9.9341	9.9162	10.0721

where every day we took a sample of 6 individuals of all manufactured items produced the same day.

We can find in the tables that:

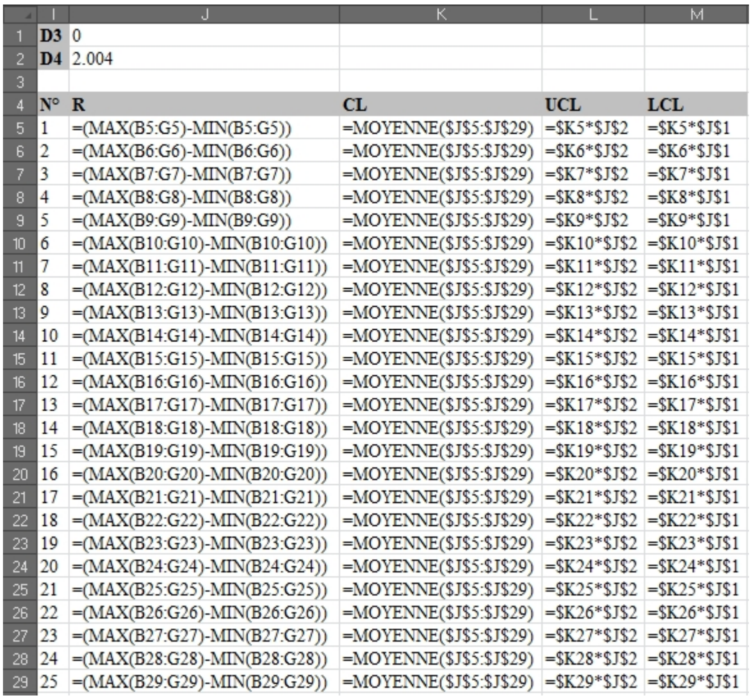
$$A_2(n_i)=A_2(6)=0.482 \tag{16.7.825}$$

We then have the following table:



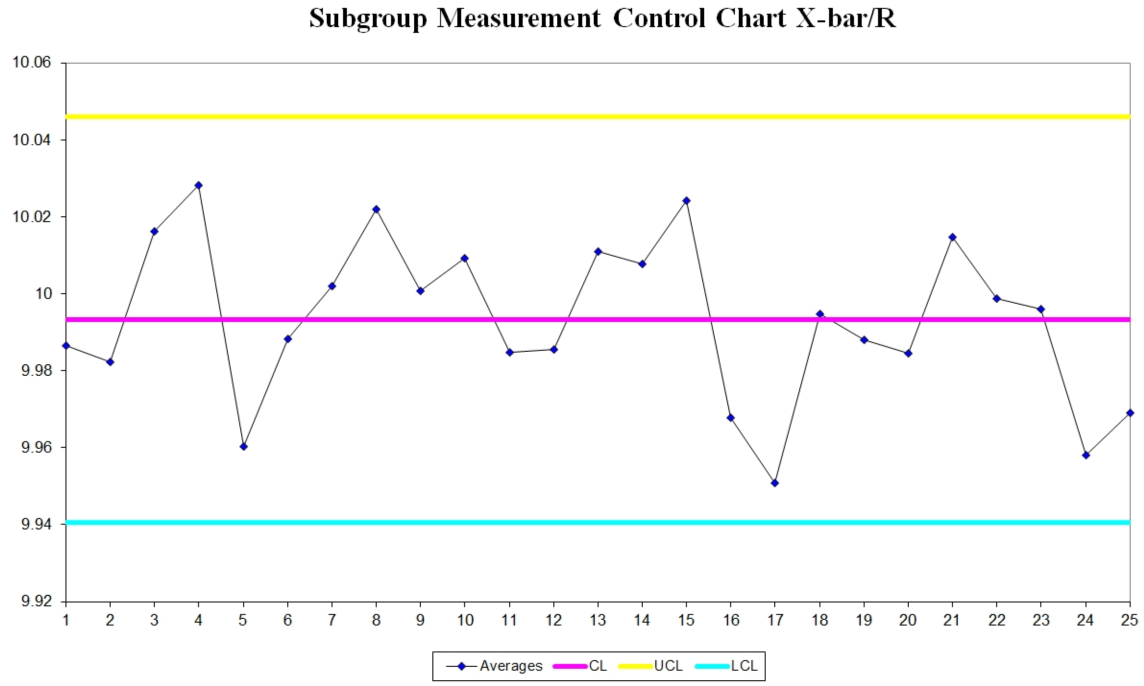
	I	J	K	L	M	N
1	**A2**	0.483				
2						
3	**N°**	**Averages**	**CL**	**R**	**UCL**	**LCL**
4	1	9.9865	9.9933	0.0652	10.0460	9.9406
5	2	9.9823	9.9933	0.0833	10.0460	9.9406
6	3	10.0162	9.9933	0.1302	10.0460	9.9406
7	4	10.0282	9.9933	0.1286	10.0460	9.9406
8	5	9.9604	9.9933	0.0733	10.0460	9.9406
9	6	9.9882	9.9933	0.0893	10.0460	9.9406
10	7	10.0021	9.9933	0.1217	10.0460	9.9406
11	8	10.0219	9.9933	0.1759	10.0460	9.9406
12	9	10.0009	9.9933	0.1194	10.0460	9.9406
13	10	10.0093	9.9933	0.1221	10.0460	9.9406
14	11	9.9849	9.9933	0.1202	10.0460	9.9406
15	12	9.9856	9.9933	0.1434	10.0460	9.9406
16	13	10.0111	9.9933	0.0638	10.0460	9.9406
17	14	10.0076	9.9933	0.0946	10.0460	9.9406
18	15	10.0243	9.9933	0.1016	10.0460	9.9406
19	16	9.9679	9.9933	0.1104	10.0460	9.9406
20	17	9.9508	9.9933	0.0856	10.0460	9.9406
21	18	9.9947	9.9933	0.0650	10.0460	9.9406
22	19	9.9880	9.9933	0.1380	10.0460	9.9406
23	20	9.9846	9.9933	0.1080	10.0460	9.9406
24	21	10.0147	9.9933	0.1021	10.0460	9.9406
25	22	9.9988	9.9933	0.0620	10.0460	9.9406
26	23	9.9961	9.9933	0.1168	10.0460	9.9406
27	24	9.9582	9.9933	0.1531	10.0460	9.9406
28	25	9.9690	9.9933	0.1559	10.0460	9.9406

or explicitly:

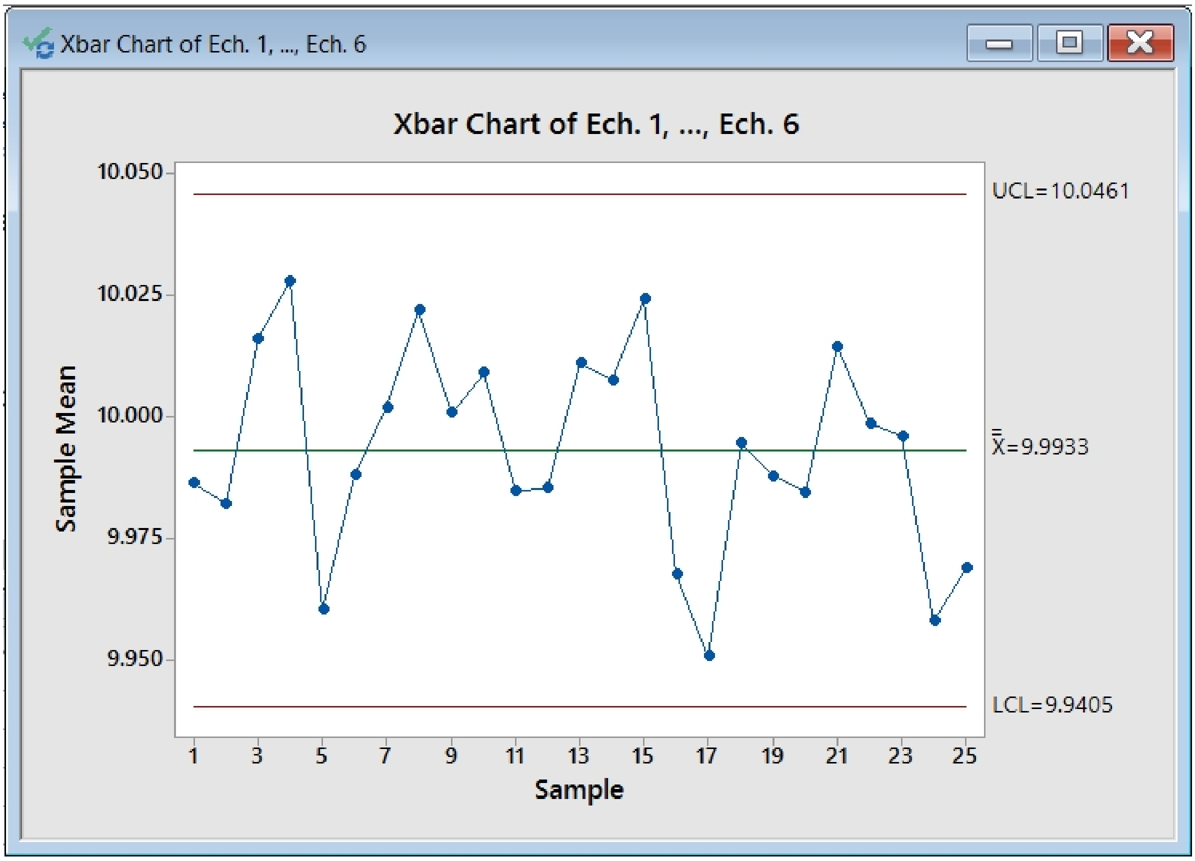


	I	J	K	L	M
1	**D3**	0			
2	**D4**	2.004			
3					
4	**N°**	**R**	**CL**	**UCL**	**LCL**
5	1	=(MAX(B5:G5)-MIN(B5:G5))	=MOYENNE($J$5:$J$29)	=$K5*$J$2	=$K5*$J$1
6	2	=(MAX(B6:G6)-MIN(B6:G6))	=MOYENNE($J$5:$J$29)	=$K6*$J$2	=$K6*$J$1
7	3	=(MAX(B7:G7)-MIN(B7:G7))	=MOYENNE($J$5:$J$29)	=$K7*$J$2	=$K7*$J$1
8	4	=(MAX(B8:G8)-MIN(B8:G8))	=MOYENNE($J$5:$J$29)	=$K8*$J$2	=$K8*$J$1
9	5	=(MAX(B9:G9)-MIN(B9:G9))	=MOYENNE($J$5:$J$29)	=$K9*$J$2	=$K9*$J$1
10	6	=(MAX(B10:G10)-MIN(B10:G10))	=MOYENNE($J$5:$J$29)	=$K10*$J$2	=$K10*$J$1
11	7	=(MAX(B11:G11)-MIN(B11:G11))	=MOYENNE($J$5:$J$29)	=$K11*$J$2	=$K11*$J$1
12	8	=(MAX(B12:G12)-MIN(B12:G12))	=MOYENNE($J$5:$J$29)	=$K12*$J$2	=$K12*$J$1
13	9	=(MAX(B13:G13)-MIN(B13:G13))	=MOYENNE($J$5:$J$29)	=$K13*$J$2	=$K13*$J$1
14	10	=(MAX(B14:G14)-MIN(B14:G14))	=MOYENNE($J$5:$J$29)	=$K14*$J$2	=$K14*$J$1
15	11	=(MAX(B15:G15)-MIN(B15:G15))	=MOYENNE($J$5:$J$29)	=$K15*$J$2	=$K15*$J$1
16	12	=(MAX(B16:G16)-MIN(B16:G16))	=MOYENNE($J$5:$J$29)	=$K16*$J$2	=$K16*$J$1
17	13	=(MAX(B17:G17)-MIN(B17:G17))	=MOYENNE($J$5:$J$29)	=$K17*$J$2	=$K17*$J$1
18	14	=(MAX(B18:G18)-MIN(B18:G18))	=MOYENNE($J$5:$J$29)	=$K18*$J$2	=$K18*$J$1
19	15	=(MAX(B19:G19)-MIN(B19:G19))	=MOYENNE($J$5:$J$29)	=$K19*$J$2	=$K19*$J$1
20	16	=(MAX(B20:G20)-MIN(B20:G20))	=MOYENNE($J$5:$J$29)	=$K20*$J$2	=$K20*$J$1
21	17	=(MAX(B21:G21)-MIN(B21:G21))	=MOYENNE($J$5:$J$29)	=$K21*$J$2	=$K21*$J$1
22	18	=(MAX(B22:G22)-MIN(B22:G22))	=MOYENNE($J$5:$J$29)	=$K22*$J$2	=$K22*$J$1
23	19	=(MAX(B23:G23)-MIN(B23:G23))	=MOYENNE($J$5:$J$29)	=$K23*$J$2	=$K23*$J$1
24	20	=(MAX(B24:G24)-MIN(B24:G24))	=MOYENNE($J$5:$J$29)	=$K24*$J$2	=$K24*$J$1
25	21	=(MAX(B25:G25)-MIN(B25:G25))	=MOYENNE($J$5:$J$29)	=$K25*$J$2	=$K25*$J$1
26	22	=(MAX(B26:G26)-MIN(B26:G26))	=MOYENNE($J$5:$J$29)	=$K26*$J$2	=$K26*$J$1
27	23	=(MAX(B27:G27)-MIN(B27:G27))	=MOYENNE($J$5:$J$29)	=$K27*$J$2	=$K27*$J$1
28	24	=(MAX(B28:G28)-MIN(B28:G28))	=MOYENNE($J$5:$J$29)	=$K28*$J$2	=$K28*$J$1
29	25	=(MAX(B29:G29)-MIN(B29:G29))	=MOYENNE($J$5:$J$29)	=$K29*$J$2	=$K29*$J$1

and the corresponding chart still with the same spreadsheet software:



With Minitab 17.1.3 this gives for people that don't trust our calculations made with Microsoft® Excel® (we can also do such a chart with `R` but the plot quality output is not acceptable in most corporations):


Xbar Chart of Ech. 1, ..., Ech. 6
Xbar Chart of Ech. 1, ..., Ech. 6
Sample Mean
10.050
10.025
10.000
9.975
9.950
UCL=10.0461
$\bar{\bar{X}}$=9.9933
LCL=9.9405
1
3
5
7
9
11
13
15
17
19
21
23
25
Sample

which (is it necessary to recall it?) underestimates the number of error detections since the probability distribution of the range is in reality not symmetrical (hence the application of fixed Shewhart's control limits is not really suited ...) and because almost all of the tables assume the data still Normally distributed as well.

Let us notice that this control chart would according to the WECO rules have 0 point beyond the $\pm 3\sigma$.

Finally for the people interested here are some data that may perhaps be useful to some people:

Factors useful in the Construction of Control Charts

	Chart for Averages			*Chart for standard deviations*					*Chart for Ranges*						
				Factors for:											
	Factors for Control Limits			*Central Line*	*Control Limits*				*Factors for Central Line*		*Factors for Control Limits*				
n	A	A_2	A_3	c_4	B_3	B_4	B_5	B_6	d_2	$1/d_2$	d_3	D_1	D_2	D_3	D_4
2	2.121	1.880	2.659	0.7979	0	3.267	0	2.606	1.128	0.8862	0.852	0	3.686	0	3.266
3	1.732	1.023	1.954	0.8862	0	2.568	0	2.276	1.693	0.5908	0.888	0	4.357	0	2.574
4	1.500	0.729	1.628	0.9213	0	2.266	0	2.088	2.059	0.4857	0.879	0	4.697	0	2.281
5	1.342	0.577	1.427	0.9400	0	2.089	0	1.964	2.326	0.4299	0.864	0	4.918	0	2.114
6	1.225	0.483	1.287	0.9515	0.030	1.970	0.029	1.874	2.534	0.3946	0.848	0	5.078	0	2.003
7	1.134	0.419	1.182	0.9594	0.118	1.882	0.113	1.806	2.704	0.3698	0.833	0.206	5.203	0.076	1.924
8	1.061	0.373	1.099	0.9650	0.185	1.815	0.179	1.751	2.847	0.3512	0.819	0.389	5.306	0.137	1.863
9	1.000	0.337	1.032	0.9693	0.239	1.761	0.232	1.707	2.970	0.3367	0.807	0.548	5.392	0.184	1.816
10	0.949	0.308	0.975	0.9727	0.284	1.716	0.276	1.669	3.078	0.3249	0.797	0.688	5.467	0.223	1.777
11	0.905	0.285	0.927	0.9754	0.321	1.679	0.313	1.637	3.173	0.3152	0.787	0.813	5.533	0.256	1.744
12	0.866	0.266	0.886	0.9776	0.354	1.646	0.346	1.610	3.258	0.3069	0.778	0.924	5.593	0.284	1.716
13	0.832	0.249	0.850	0.9794	0.382	1.618	0.374	1.585	3.336	0.2998	0.770	1.026	5.646	0.307	1.693
14	0.802	0.235	0.817	0.9810	0.406	1.594	0.399	1.563	3.407	0.2935	0.763	1.119	5.695	0.328	1.672
15	0.775	0.223	0.789	0.9823	0.428	1.572	0.421	1.544	3.472	0.2880	0.756	1.204	5.739	0.347	1.653
16	0.750	0.212	0.763	0.9835	0.448	1.552	0.440	1.526	3.532	0.2831	0.750	1.283	5.781	0.363	1.637
17	0.728	0.203	0.739	0.9845	0.466	1.534	0.458	1.511	3.588	0.2787	0.744	1.357	5.819	0.378	1.622
18	0.707	0.194	0.718	0.9854	0.482	1.518	0.475	1.496	3.640	0.2747	0.738	1.425	5.855	0.392	1.608
19	0.688	0.187	0.698	0.9862	0.497	1.503	0.490	1.483	3.689	0.2711	0.733	1.490	5.888	0.404	1.596
20	0.671	0.180	0.680	0.9869	0.510	1.490	0.504	1.470	3.735	0.2677	0.728	1.550	5.920	0.415	1.585
21	0.655	0.173	0.663	0.9876	0.523	1.477	0.516	1.459	3.778	0.2647	0.724	1.607	5.950	0.425	1.575
22	0.640	0.167	0.647	0.9882	0.534	1.466	0.528	1.448	3.819	0.2618	0.719	1.661	5.978	0.435	1.565
23	0.626	0.162	0.633	0.9887	0.545	1.455	0.539	1.438	3.858	0.2592	0.715	1.712	6.004	0.444	1.556
24	0.612	0.157	0.619	0.9892	0.555	1.445	0.549	1.429	3.895	0.2567	0.712	1.761	6.030	0.452	1.548
25	0.600	0.153	0.606	0.9896	0.565	1.435	0.559	1.420	3.931	0.2544	0.708	1.807	6.055	0.460	1.540
Over 25	$3/\sqrt{n}$	$3/d_2\sqrt{n}$	$\dots$	$\dots$	*	**	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$

$$* \; 1-\frac{3}{\sqrt{2n}} \qquad ** \; 1+\frac{3}{\sqrt{2n}}$$

16.7.6.5 Autocorrelated Measurement Control Charts (time weighted control charts)

The family of autocorrelated control charts is defined by the fact that either the points on the chart or the control limits are calculated based on a number of the foregoing points.

These control chart have the advantage of requiring less data than others and to exacerbate the problem of deviation and therefore detect anomalies faster.

The WECO rules do not apply to autocorrelated control chart. If a point is outside the calculated limits the process requires an immediate corrective action.

16.7.6.5.1 $\mathrm{I}-\mathrm{MR}/\overline{X}$ Individual moving range measurement control chart

In prototyping phase, we can not put in place the concept of sampling. In addition, the measurement short term control chart on the mean with standard deviation does not gives the possibility to make statistical inference, because the distribution law is supposed unknown simply because the process is unstable.

An intermediate solution is to consider then that the process is stable (assumption of Normality of the data) and calculate the limits from a minimum number of individuals that it is possible to have to make a statistic. That is to say: two!

This situation is then well suited for prototyping phases or in the case of processes/methods with slow frequencies for which the measurements (pieces)/controls are very expensive.

Pros:

- Complete the first control charts (measurement on subgroups) that we saw at the beginning for situations where it is difficult to have a large number of individuals by sample for reasons of cost or time.
- This control chart is sometimes regarded as the Swiss Army knife of quality control ... and is increasingly used because it is the policy of the "Just in Time" (JIT) production for small series (for the decrease of costs).

Cons:

- The mathematical proof of the origin of the control limits is quite difficult (see the section Statistics at page page 1203 for detailed proof) and the determination of the coefficients requires the use of Monte Carlo simulation (see section Numerical Methods page 6708)).
- Another problem is that although the probability distribution of the range can be chosen, it is often given in specialized books only for the Normal distribution (but most advanced softwares manage various distribution laws).

As workaround of the problem of an acceptable estimate of the standard deviation, the idea is to consider two successive measurements x_{i-1},x_i as the random variables of a rank statistics of which we calculate the range in absolute terms:

$$\text{MR}_i=R_i=|x_i-x_{i-1}|=\max(x_{i-1},x_i)-\min(x_{i-1},x_i) \quad (16.7.826)$$

where MR stands for "Moving Range" (the control chart is sometimes named "MR(2) control chart"). And then we have:

$$\overline{\text{MR}}=\overline{R}=\frac{1}{n-1}\sum_{i=1}^{n}|x_i-x_{i-1}| \quad (16.7.827)$$

since we have necessarily $n-1$ moving ranges.

Then we take back the definitions of the control chart by measurement for the mean with short term standard deviation:

$$\begin{aligned}\text{USL}&=\text{required by customer/board comitee}\\ \text{UCL}_{\overline{X}}&=\overline{\overline{X}}+3\sigma_X\\ \text{CL}_{\overline{X}}&=\text{E}(X)=\overline{X}\\ T&=\text{target required by customer/board comitee}\\ \text{LCL}_{\overline{X}}&=\overline{\overline{X}}-3\sigma_X\\ \text{LSL}&=\text{required by customer/board comitee}\end{aligned} \quad (16.7.828)$$

and we can get rid of the of the estimation problem of σ_X using the relation proved in the section Statistics during our study of range statistics (see section Statistics page 1203):

$$\overline{R}=\sigma_X d_2(n_i) \quad (16.7.829)$$

Therefore rearranging and using the prior previous relation:

$$\boxed{\sigma_X=\frac{\overline{R}}{d_2(n_i)}=\frac{\text{MR}}{d_2(n_i)}} \quad (16.7.830)$$

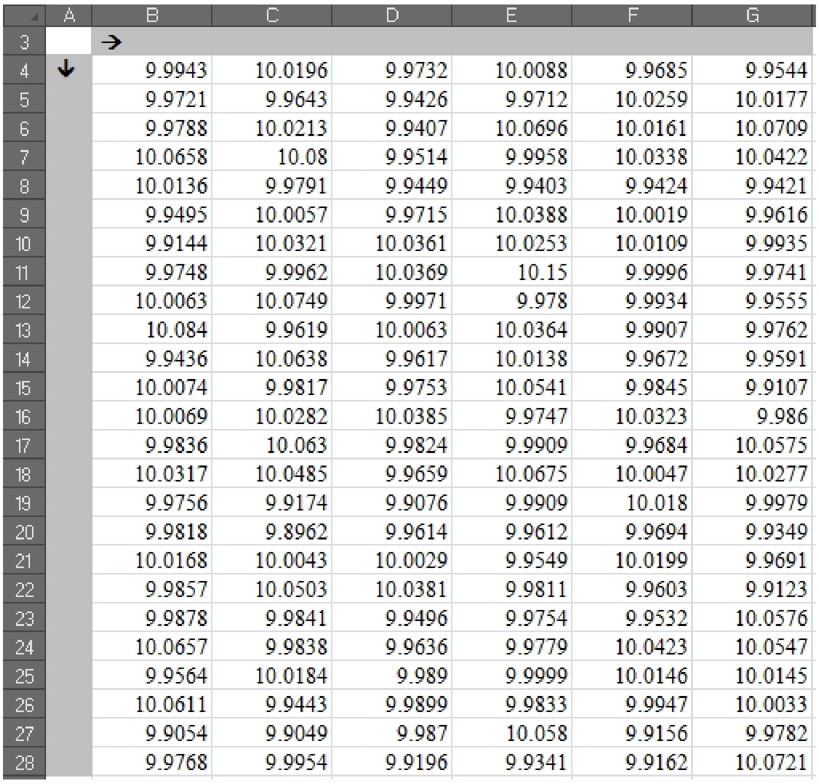
It comes (it is impossible that the customer sets a limit for the USL, LSL or target T as regards the range as in general it is asymmetric, hence the fact that these parameters are this time not listed):

$$\begin{aligned}\text{UCL}_{\overline{X}(\text{MR})}&=\overline{X}+3\frac{\overline{\text{MR}}}{d_2(n_i)}\\ \text{CL}_{\overline{X}(\text{MR})}&=\overline{R}\\ \text{LCL}_{\overline{X}(\text{MR})}&=\overline{X}-3\frac{\overline{\text{MR}}}{d_2(n_i)}\end{aligned} \quad (16.7.831)$$

Remark

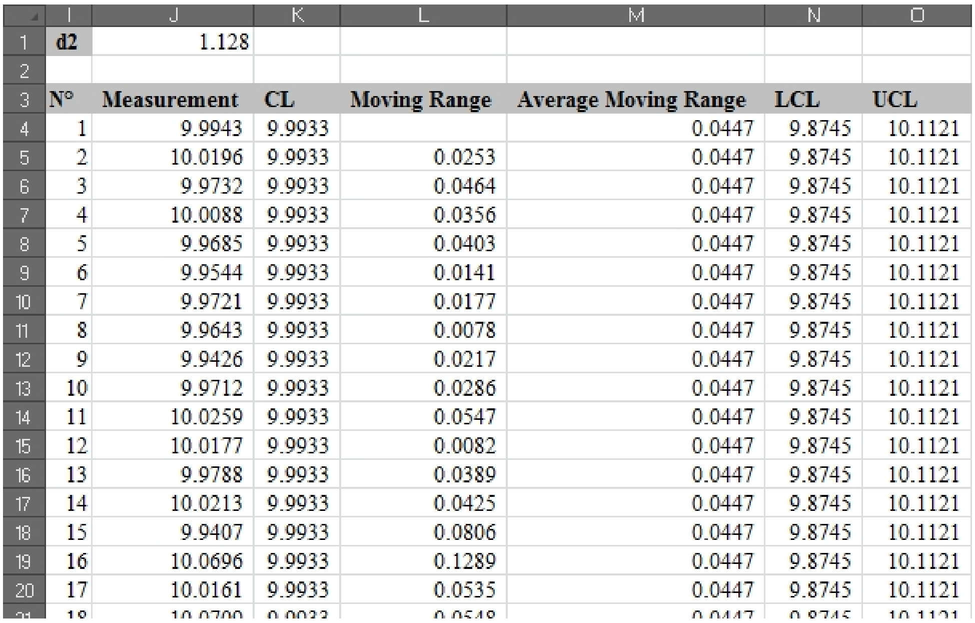
The same unbiasing constant can be used for subgroup control charts! Indeed, rather than using $R_i=x_{i-1}-x_i$ we use $R_i=\overline{X}_{i-1}-\overline{X}_i$ where i is then the number of the sample!!! This is why in fact Minitab propose in the menu of *Measurement subgroup control charts* the MR chart!!!

As companion example let us consider the following data:



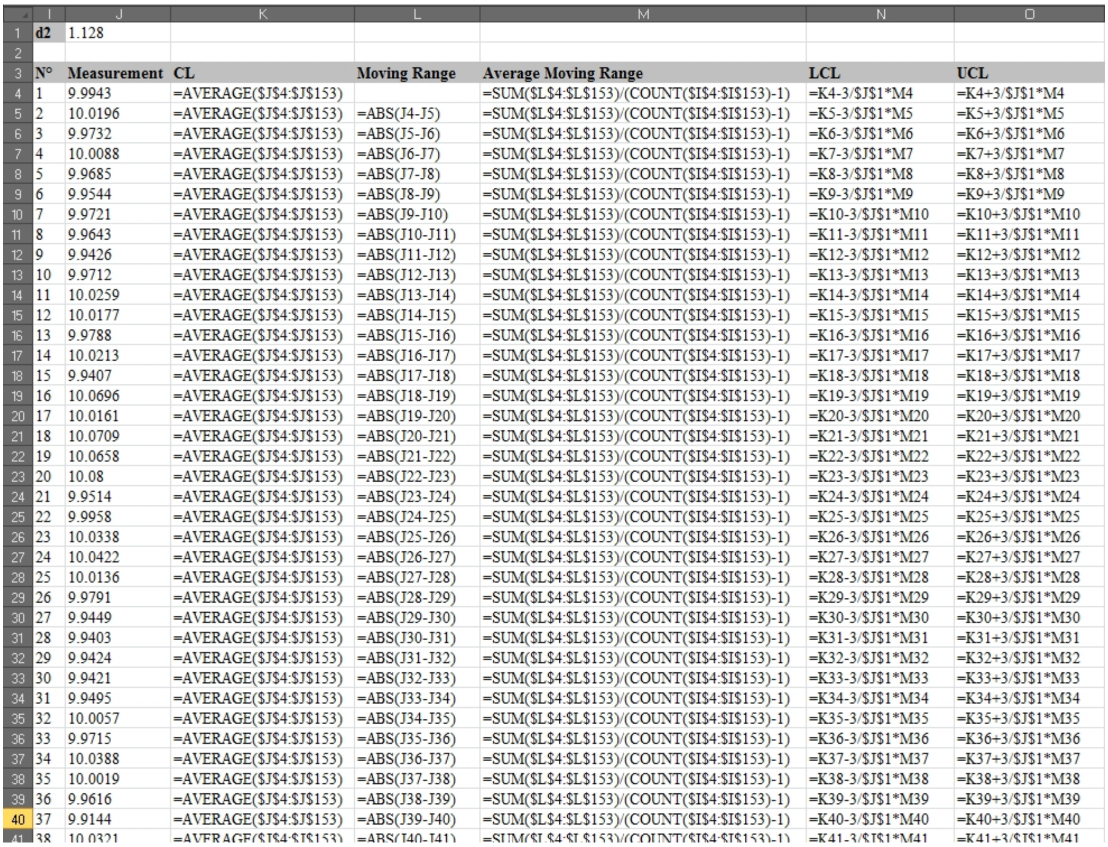
	A	B	C	D	E	F	G
3		→					
4	↓	9.9943	10.0196	9.9732	10.0088	9.9685	9.9544
5		9.9721	9.9643	9.9426	9.9712	10.0259	10.0177
6		9.9788	10.0213	9.9407	10.0696	10.0161	10.0709
7		10.0658	10.08	9.9514	9.9958	10.0338	10.0422
8		10.0136	9.9791	9.9449	9.9403	9.9424	9.9421
9		9.9495	10.0057	9.9715	10.0388	10.0019	9.9616
10		9.9144	10.0321	10.0361	10.0253	10.0109	9.9935
11		9.9748	9.9962	10.0369	10.15	9.9996	9.9741
12		10.0063	10.0749	9.9971	9.978	9.9934	9.9555
13		10.084	9.9619	10.0063	10.0364	9.9907	9.9762
14		9.9436	10.0638	9.9617	10.0138	9.9672	9.9591
15		10.0074	9.9817	9.9753	10.0541	9.9845	9.9107
16		10.0069	10.0282	10.0385	9.9747	10.0323	9.986
17		9.9836	10.063	9.9824	9.9909	9.9684	10.0575
18		10.0317	10.0485	9.9659	10.0675	10.0047	10.0277
19		9.9756	9.9174	9.9076	9.9909	10.018	9.9979
20		9.9818	9.8962	9.9614	9.9612	9.9694	9.9349
21		10.0168	10.0043	10.0029	9.9549	10.0199	9.9691
22		9.9857	10.0503	10.0381	9.9811	9.9603	9.9123
23		9.9878	9.9841	9.9496	9.9754	9.9532	10.0576
24		10.0657	9.9838	9.9636	9.9779	10.0423	10.0547
25		9.9564	10.0184	9.989	9.9999	10.0146	10.0145
26		10.0611	9.9443	9.9899	9.9833	9.9947	10.0033
27		9.9054	9.9049	9.987	10.058	9.9156	9.9782
28		9.9768	9.9954	9.9196	9.9341	9.9162	10.0721

Then we have for the first 17 visible rows only:



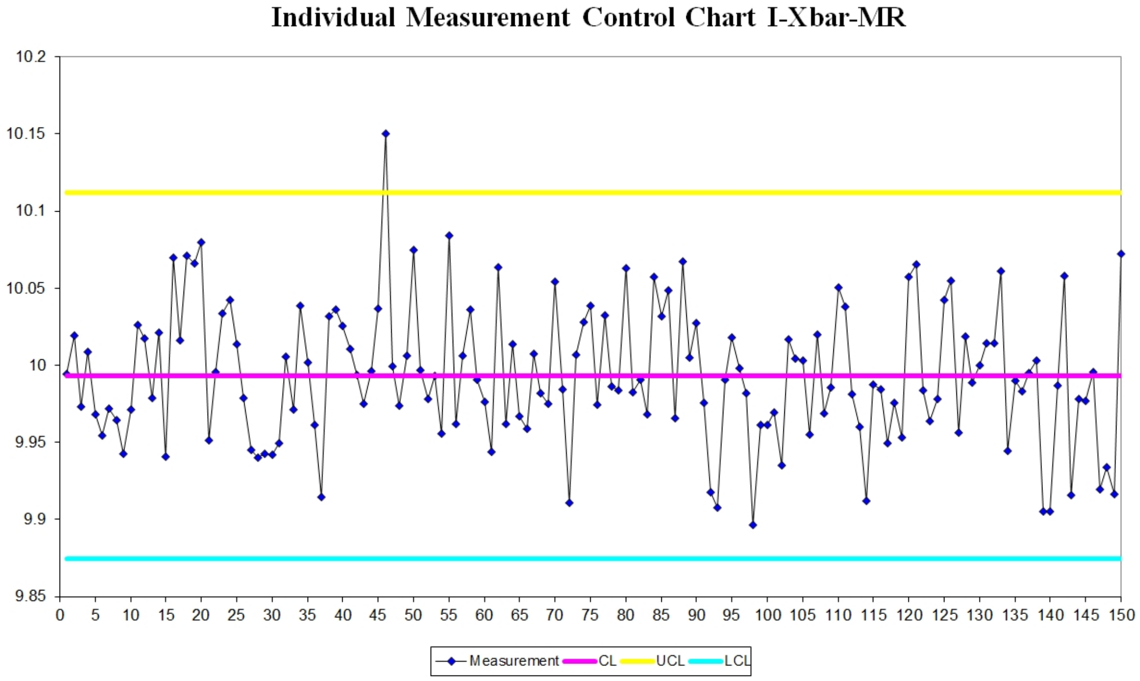
	I	J	K	L	M	N	O
1	**d2**	1.128					
2							
3	**N°**	**Measurement**	**CL**	**Moving Range**	**Average Moving Range**	**LCL**	**UCL**
4	1	9.9943	9.9933		0.0447	9.8745	10.1121
5	2	10.0196	9.9933	0.0253	0.0447	9.8745	10.1121
6	3	9.9732	9.9933	0.0464	0.0447	9.8745	10.1121
7	4	10.0088	9.9933	0.0356	0.0447	9.8745	10.1121
8	5	9.9685	9.9933	0.0403	0.0447	9.8745	10.1121
9	6	9.9544	9.9933	0.0141	0.0447	9.8745	10.1121
10	7	9.9721	9.9933	0.0177	0.0447	9.8745	10.1121
11	8	9.9643	9.9933	0.0078	0.0447	9.8745	10.1121
12	9	9.9426	9.9933	0.0217	0.0447	9.8745	10.1121
13	10	9.9712	9.9933	0.0286	0.0447	9.8745	10.1121
14	11	10.0259	9.9933	0.0547	0.0447	9.8745	10.1121
15	12	10.0177	9.9933	0.0082	0.0447	9.8745	10.1121
16	13	9.9788	9.9933	0.0389	0.0447	9.8745	10.1121
17	14	10.0213	9.9933	0.0425	0.0447	9.8745	10.1121
18	15	9.9407	9.9933	0.0806	0.0447	9.8745	10.1121
19	16	10.0696	9.9933	0.1289	0.0447	9.8745	10.1121
20	17	10.0161	9.9933	0.0535	0.0447	9.8745	10.1121

Explicitly:

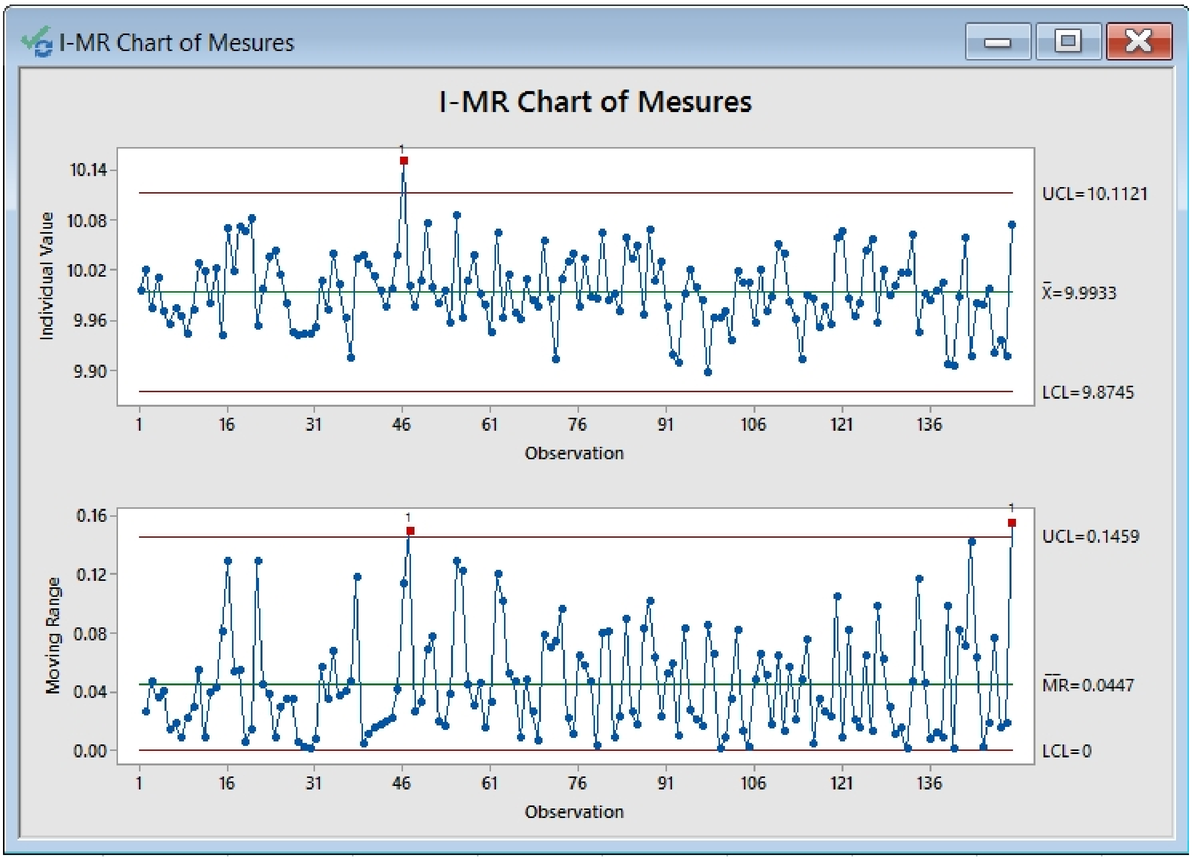


	I	J	K	L	M	N	O
1	d2	1.128					
2							
3	N°	Measurement	CL	Moving Range	Average Moving Range	LCL	UCL
4	1	9.9943	=AVERAGE($J$4:$J$153)		=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K4-3/$J$1*M4	=K4+3/$J$1*M4
5	2	10.0196	=AVERAGE($J$4:$J$153)	=ABS(J4-J5)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K5-3/$J$1*M5	=K5+3/$J$1*M5
6	3	9.9732	=AVERAGE($J$4:$J$153)	=ABS(J5-J6)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K6-3/$J$1*M6	=K6+3/$J$1*M6
7	4	10.0088	=AVERAGE($J$4:$J$153)	=ABS(J6-J7)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K7-3/$J$1*M7	=K7+3/$J$1*M7
8	5	9.9685	=AVERAGE($J$4:$J$153)	=ABS(J7-J8)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K8-3/$J$1*M8	=K8+3/$J$1*M8
9	6	9.9544	=AVERAGE($J$4:$J$153)	=ABS(J8-J9)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K9-3/$J$1*M9	=K9+3/$J$1*M9
10	7	9.9721	=AVERAGE($J$4:$J$153)	=ABS(J9-J10)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K10-3/$J$1*M10	=K10+3/$J$1*M10
11	8	9.9643	=AVERAGE($J$4:$J$153)	=ABS(J10-J11)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K11-3/$J$1*M11	=K11+3/$J$1*M11
12	9	9.9426	=AVERAGE($J$4:$J$153)	=ABS(J11-J12)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K12-3/$J$1*M12	=K12+3/$J$1*M12
13	10	9.9712	=AVERAGE($J$4:$J$153)	=ABS(J12-J13)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K13-3/$J$1*M13	=K13+3/$J$1*M13
14	11	10.0259	=AVERAGE($J$4:$J$153)	=ABS(J13-J14)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K14-3/$J$1*M14	=K14+3/$J$1*M14
15	12	10.0177	=AVERAGE($J$4:$J$153)	=ABS(J14-J15)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K15-3/$J$1*M15	=K15+3/$J$1*M15
16	13	9.9788	=AVERAGE($J$4:$J$153)	=ABS(J15-J16)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K16-3/$J$1*M16	=K16+3/$J$1*M16
17	14	10.0213	=AVERAGE($J$4:$J$153)	=ABS(J16-J17)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K17-3/$J$1*M17	=K17+3/$J$1*M17
18	15	9.9407	=AVERAGE($J$4:$J$153)	=ABS(J17-J18)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K18-3/$J$1*M18	=K18+3/$J$1*M18
19	16	10.0696	=AVERAGE($J$4:$J$153)	=ABS(J18-J19)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K19-3/$J$1*M19	=K19+3/$J$1*M19
20	17	10.0161	=AVERAGE($J$4:$J$153)	=ABS(J19-J20)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K20-3/$J$1*M20	=K20+3/$J$1*M20
21	18	10.0709	=AVERAGE($J$4:$J$153)	=ABS(J20-J21)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K21-3/$J$1*M21	=K21+3/$J$1*M21
22	19	10.0658	=AVERAGE($J$4:$J$153)	=ABS(J21-J22)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K22-3/$J$1*M22	=K22+3/$J$1*M22
23	20	10.08	=AVERAGE($J$4:$J$153)	=ABS(J22-J23)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K23-3/$J$1*M23	=K23+3/$J$1*M23
24	21	9.9514	=AVERAGE($J$4:$J$153)	=ABS(J23-J24)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K24-3/$J$1*M24	=K24+3/$J$1*M24
25	22	9.9958	=AVERAGE($J$4:$J$153)	=ABS(J24-J25)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K25-3/$J$1*M25	=K25+3/$J$1*M25
26	23	10.0338	=AVERAGE($J$4:$J$153)	=ABS(J25-J26)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K26-3/$J$1*M26	=K26+3/$J$1*M26
27	24	10.0422	=AVERAGE($J$4:$J$153)	=ABS(J26-J27)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K27-3/$J$1*M27	=K27+3/$J$1*M27
28	25	10.0136	=AVERAGE($J$4:$J$153)	=ABS(J27-J28)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K28-3/$J$1*M28	=K28+3/$J$1*M28
29	26	9.9791	=AVERAGE($J$4:$J$153)	=ABS(J28-J29)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K29-3/$J$1*M29	=K29+3/$J$1*M29
30	27	9.9449	=AVERAGE($J$4:$J$153)	=ABS(J29-J30)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K30-3/$J$1*M30	=K30+3/$J$1*M30
31	28	9.9403	=AVERAGE($J$4:$J$153)	=ABS(J30-J31)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K31-3/$J$1*M31	=K31+3/$J$1*M31
32	29	9.9424	=AVERAGE($J$4:$J$153)	=ABS(J31-J32)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K32-3/$J$1*M32	=K32+3/$J$1*M32
33	30	9.9421	=AVERAGE($J$4:$J$153)	=ABS(J32-J33)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K33-3/$J$1*M33	=K33+3/$J$1*M33
34	31	9.9495	=AVERAGE($J$4:$J$153)	=ABS(J33-J34)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K34-3/$J$1*M34	=K34+3/$J$1*M34
35	32	10.0057	=AVERAGE($J$4:$J$153)	=ABS(J34-J35)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K35-3/$J$1*M35	=K35+3/$J$1*M35
36	33	9.9715	=AVERAGE($J$4:$J$153)	=ABS(J35-J36)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K36-3/$J$1*M36	=K36+3/$J$1*M36
37	34	10.0388	=AVERAGE($J$4:$J$153)	=ABS(J36-J37)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K37-3/$J$1*M37	=K37+3/$J$1*M37
38	35	10.0019	=AVERAGE($J$4:$J$153)	=ABS(J37-J38)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K38-3/$J$1*M38	=K38+3/$J$1*M38
39	36	9.9616	=AVERAGE($J$4:$J$153)	=ABS(J38-J39)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K39-3/$J$1*M39	=K39+3/$J$1*M39
40	37	9.9144	=AVERAGE($J$4:$J$153)	=ABS(J39-J40)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K40-3/$J$1*M40	=K40+3/$J$1*M40
41	38	10.0321	=AVERAGE($J$4:$J$153)	=ABS(J40-J41)	=SUM($L$4:$L$153)/(COUNT($I$4:$I$153)-1)	=K41-3/$J$1*M41	=K41+3/$J$1*M41

Then we have the following chart:



With Minitab 17.1.3 this gives for people that don't trust our calculations made with Microsoft® Excel®:



Notice that this control chart would have according to the WECO rules 1 point beyond the $\pm 3\sigma$ (we indicate that for comparison reasons with previous control charts and the next one).

As the reader has probably noticed it in the figure above, Minitab also plots at the same time an I-MR/$\overline{\text{MR}}$ control chart. So let us study this one just now:

16.7.6.5.2 $\text{I}-\text{MR}/\overline{\text{MR}}$ Individual moving range measurement control chart

So as we just said and noticed it, the previous central tendency control chart must be completed (same normally as for almost all previous control charts with central tendency) with a control chart showing the tendency of dispersion (the benefits and problems of this control chart are the same as the previous one).

Then we take the definitions of a measurement control chart to get:

$$\begin{aligned}\text{UCL}_{\text{MR}} &= \overline{\text{MR}} + 3\sigma_{\overline{\text{MR}}}\\ \text{CL}_{\text{MR}} &= \overline{\text{MR}}\\ \text{LCL}_{\text{MR}} &= \overline{\text{MR}} - 3\sigma_{\overline{\text{MR}}}\end{aligned} \tag{16.7.832}$$

Again, the standard deviation of the moving range (MR) is constructed in the same way as the standard deviation of the range. Then we have:

$$\sigma_{\text{MR}} = \frac{d_3(n_i)\overline{\text{MR}}}{d_2(n_i)} \tag{16.7.833}$$

It comes (it is impossible that the customer sets a limit for the USL, LSL or target T as regards the range as in general it is asymmetric, hence the fact that these parameters are this time not listed):

$$\begin{aligned}
\text{UCL}_{\text{MR}} &= \overline{\text{MR}} + 3\sigma_{\overline{\text{MR}}} = \overline{\text{MR}} + 3\frac{d_3(n_i)\overline{\text{MR}}}{d_2(n_i)} = \overline{\text{MR}}\left(1 + 3\frac{d_3(n_i)}{d_2(n_i)}\right) \\
\text{CL}_{\text{MR}} &= \overline{\text{MR}} \\
\text{LCL}_{\text{MR}} &= \overline{\text{MR}} - 3\sigma_{\overline{\text{MR}}} = \overline{\text{MR}} - 3\frac{d_3(n_i)\overline{\text{MR}}}{d_2(n_i)} = \overline{\text{MR}}\left(1 - 3\frac{d_3(n_i)}{d_2(n_i)}\right)
\end{aligned} \tag{16.7.834}$$

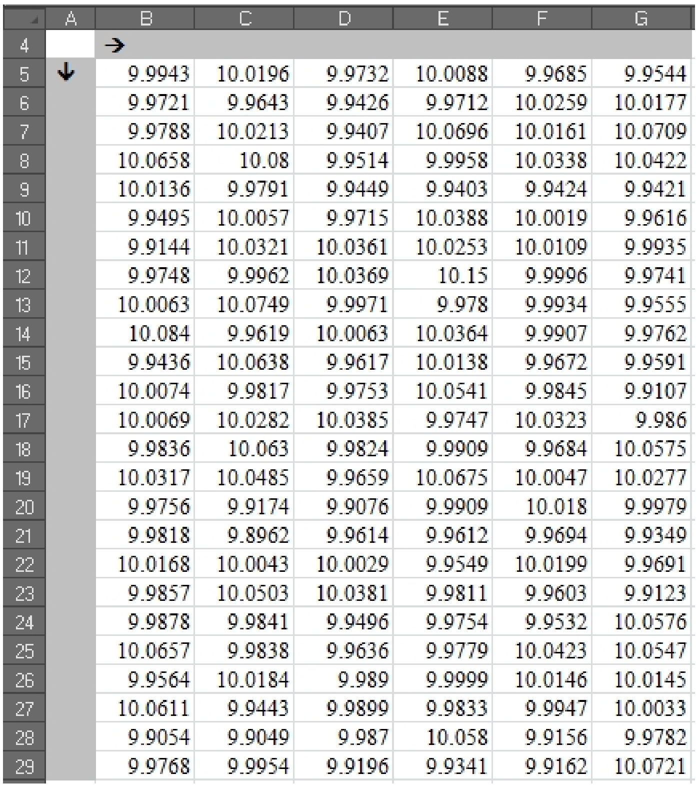
which is denoted as well as the range control chart:

$$\boxed{\begin{aligned}
\text{UCL}_{\text{MR}} &= \overline{\text{MR}}D_4(n_i) \\
\text{CL}_{\text{MR}} &= \overline{\text{MR}} \\
\text{LCL}_{\text{MR}} &= \overline{\text{MR}} - 3\sigma_{\overline{\text{MR}}} = \overline{\text{MR}}D_3(n_i)
\end{aligned}} \tag{16.7.835}$$

with for recall:

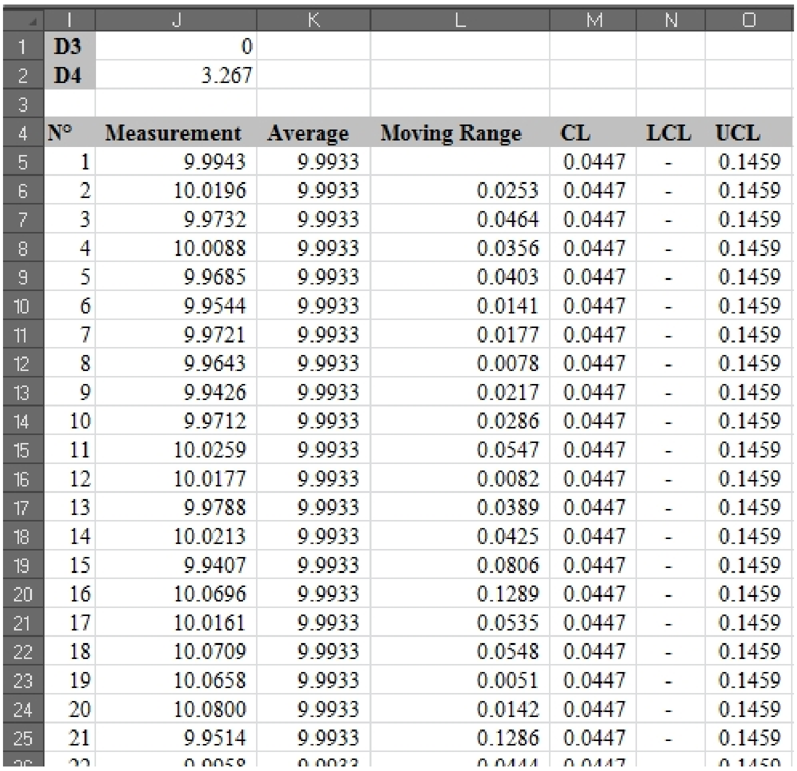
$$\boxed{D_4 = 1 + 3\frac{d_3(n_i)\overline{R}}{d_2(n_i)} \qquad D_3 = 1 - 3\frac{d_3(n_i)\overline{R}}{d_2(n_i)}} \tag{16.7.836}$$

As companion example let us consider the following data:



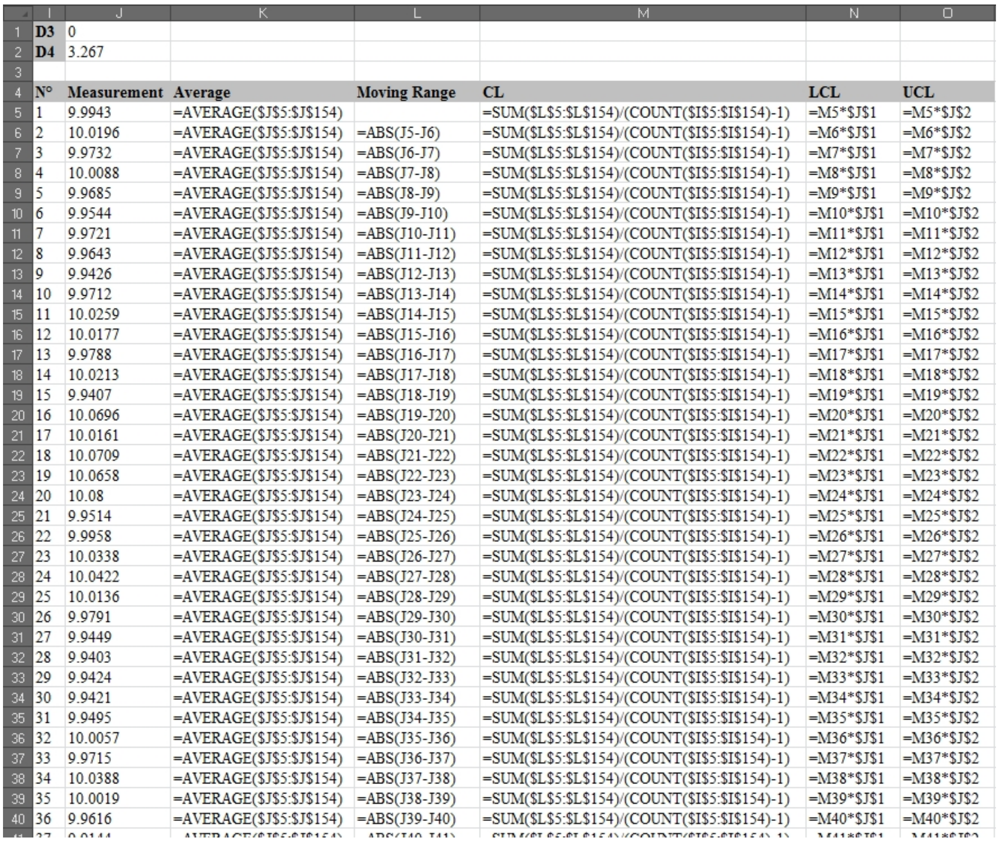
	A	B	C	D	E	F	G
4		→					
5	↓	9.9943	10.0196	9.9732	10.0088	9.9685	9.9544
6		9.9721	9.9643	9.9426	9.9712	10.0259	10.0177
7		9.9788	10.0213	9.9407	10.0696	10.0161	10.0709
8		10.0658	10.08	9.9514	9.9958	10.0338	10.0422
9		10.0136	9.9791	9.9449	9.9403	9.9424	9.9421
10		9.9495	10.0057	9.9715	10.0388	10.0019	9.9616
11		9.9144	10.0321	10.0361	10.0253	10.0109	9.9935
12		9.9748	9.9962	10.0369	10.15	9.9996	9.9741
13		10.0063	10.0749	9.9971	9.978	9.9934	9.9555
14		10.084	9.9619	10.0063	10.0364	9.9907	9.9762
15		9.9436	10.0638	9.9617	10.0138	9.9672	9.9591
16		10.0074	9.9817	9.9753	10.0541	9.9845	9.9107
17		10.0069	10.0282	10.0385	9.9747	10.0323	9.986
18		9.9836	10.063	9.9824	9.9909	9.9684	10.0575
19		10.0317	10.0485	9.9659	10.0675	10.0047	10.0277
20		9.9756	9.9174	9.9076	9.9909	10.018	9.9979
21		9.9818	9.8962	9.9614	9.9612	9.9694	9.9349
22		10.0168	10.0043	10.0029	9.9549	10.0199	9.9691
23		9.9857	10.0503	10.0381	9.9811	9.9603	9.9123
24		9.9878	9.9841	9.9496	9.9754	9.9532	10.0576
25		10.0657	9.9838	9.9636	9.9779	10.0423	10.0547
26		9.9564	10.0184	9.989	9.9999	10.0146	10.0145
27		10.0611	9.9443	9.9899	9.9833	9.9947	10.0033
28		9.9054	9.9049	9.987	10.058	9.9156	9.9782
29		9.9768	9.9954	9.9196	9.9341	9.9162	10.0721

Then we have for the first 25 visible rows only:



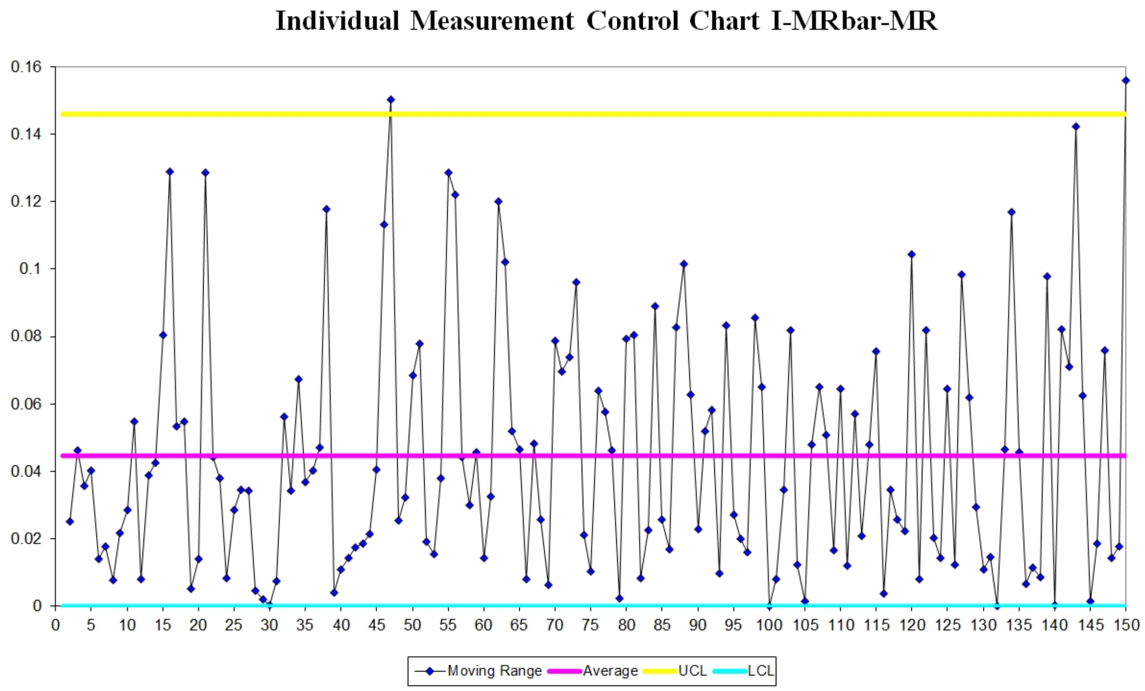
	I	J	K	L	M	N	O
1	**D3**	0					
2	**D4**	3.267					
3							
4	**N°**	**Measurement**	**Average**	**Moving Range**	**CL**	**LCL**	**UCL**
5	1	9.9943	9.9933		0.0447	-	0.1459
6	2	10.0196	9.9933	0.0253	0.0447	-	0.1459
7	3	9.9732	9.9933	0.0464	0.0447	-	0.1459
8	4	10.0088	9.9933	0.0356	0.0447	-	0.1459
9	5	9.9685	9.9933	0.0403	0.0447	-	0.1459
10	6	9.9544	9.9933	0.0141	0.0447	-	0.1459
11	7	9.9721	9.9933	0.0177	0.0447	-	0.1459
12	8	9.9643	9.9933	0.0078	0.0447	-	0.1459
13	9	9.9426	9.9933	0.0217	0.0447	-	0.1459
14	10	9.9712	9.9933	0.0286	0.0447	-	0.1459
15	11	10.0259	9.9933	0.0547	0.0447	-	0.1459
16	12	10.0177	9.9933	0.0082	0.0447	-	0.1459
17	13	9.9788	9.9933	0.0389	0.0447	-	0.1459
18	14	10.0213	9.9933	0.0425	0.0447	-	0.1459
19	15	9.9407	9.9933	0.0806	0.0447	-	0.1459
20	16	10.0696	9.9933	0.1289	0.0447	-	0.1459
21	17	10.0161	9.9933	0.0535	0.0447	-	0.1459
22	18	10.0709	9.9933	0.0548	0.0447	-	0.1459
23	19	10.0658	9.9933	0.0051	0.0447	-	0.1459
24	20	10.0800	9.9933	0.0142	0.0447	-	0.1459
25	21	9.9514	9.9933	0.1286	0.0447	-	0.1459

Explicitly:



	I	J	K	L	M	N	O
1	**D3**	0					
2	**D4**	3.267					
3							
4	**N°**	**Measurement**	**Average**	**Moving Range**	**CL**	**LCL**	**UCL**
5	1	9.9943	=AVERAGE($J$5:$J$154)		=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M5*$J$1	=M5*$J$2
6	2	10.0196	=AVERAGE($J$5:$J$154)	=ABS(J5-J6)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M6*$J$1	=M6*$J$2
7	3	9.9732	=AVERAGE($J$5:$J$154)	=ABS(J6-J7)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M7*$J$1	=M7*$J$2
8	4	10.0088	=AVERAGE($J$5:$J$154)	=ABS(J7-J8)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M8*$J$1	=M8*$J$2
9	5	9.9685	=AVERAGE($J$5:$J$154)	=ABS(J8-J9)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M9*$J$1	=M9*$J$2
10	6	9.9544	=AVERAGE($J$5:$J$154)	=ABS(J9-J10)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M10*$J$1	=M10*$J$2
11	7	9.9721	=AVERAGE($J$5:$J$154)	=ABS(J10-J11)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M11*$J$1	=M11*$J$2
12	8	9.9643	=AVERAGE($J$5:$J$154)	=ABS(J11-J12)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M12*$J$1	=M12*$J$2
13	9	9.9426	=AVERAGE($J$5:$J$154)	=ABS(J12-J13)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M13*$J$1	=M13*$J$2
14	10	9.9712	=AVERAGE($J$5:$J$154)	=ABS(J13-J14)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M14*$J$1	=M14*$J$2
15	11	10.0259	=AVERAGE($J$5:$J$154)	=ABS(J14-J15)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M15*$J$1	=M15*$J$2
16	12	10.0177	=AVERAGE($J$5:$J$154)	=ABS(J15-J16)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M16*$J$1	=M16*$J$2
17	13	9.9788	=AVERAGE($J$5:$J$154)	=ABS(J16-J17)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M17*$J$1	=M17*$J$2
18	14	10.0213	=AVERAGE($J$5:$J$154)	=ABS(J17-J18)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M18*$J$1	=M18*$J$2
19	15	9.9407	=AVERAGE($J$5:$J$154)	=ABS(J18-J19)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M19*$J$1	=M19*$J$2
20	16	10.0696	=AVERAGE($J$5:$J$154)	=ABS(J19-J20)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M20*$J$1	=M20*$J$2
21	17	10.0161	=AVERAGE($J$5:$J$154)	=ABS(J20-J21)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M21*$J$1	=M21*$J$2
22	18	10.0709	=AVERAGE($J$5:$J$154)	=ABS(J21-J22)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M22*$J$1	=M22*$J$2
23	19	10.0658	=AVERAGE($J$5:$J$154)	=ABS(J22-J23)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M23*$J$1	=M23*$J$2
24	20	10.08	=AVERAGE($J$5:$J$154)	=ABS(J23-J24)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M24*$J$1	=M24*$J$2
25	21	9.9514	=AVERAGE($J$5:$J$154)	=ABS(J24-J25)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M25*$J$1	=M25*$J$2
26	22	9.9958	=AVERAGE($J$5:$J$154)	=ABS(J25-J26)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M26*$J$1	=M26*$J$2
27	23	10.0338	=AVERAGE($J$5:$J$154)	=ABS(J26-J27)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M27*$J$1	=M27*$J$2
28	24	10.0422	=AVERAGE($J$5:$J$154)	=ABS(J27-J28)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M28*$J$1	=M28*$J$2
29	25	10.0136	=AVERAGE($J$5:$J$154)	=ABS(J28-J29)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M29*$J$1	=M29*$J$2
30	26	9.9791	=AVERAGE($J$5:$J$154)	=ABS(J29-J30)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M30*$J$1	=M30*$J$2
31	27	9.9449	=AVERAGE($J$5:$J$154)	=ABS(J30-J31)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M31*$J$1	=M31*$J$2
32	28	9.9403	=AVERAGE($J$5:$J$154)	=ABS(J31-J32)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M32*$J$1	=M32*$J$2
33	29	9.9424	=AVERAGE($J$5:$J$154)	=ABS(J32-J33)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M33*$J$1	=M33*$J$2
34	30	9.9421	=AVERAGE($J$5:$J$154)	=ABS(J33-J34)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M34*$J$1	=M34*$J$2
35	31	9.9495	=AVERAGE($J$5:$J$154)	=ABS(J34-J35)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M35*$J$1	=M35*$J$2
36	32	10.0057	=AVERAGE($J$5:$J$154)	=ABS(J35-J36)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M36*$J$1	=M36*$J$2
37	33	9.9715	=AVERAGE($J$5:$J$154)	=ABS(J36-J37)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M37*$J$1	=M37*$J$2
38	34	10.0388	=AVERAGE($J$5:$J$154)	=ABS(J37-J38)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M38*$J$1	=M38*$J$2
39	35	10.0019	=AVERAGE($J$5:$J$154)	=ABS(J38-J39)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M39*$J$1	=M39*$J$2
40	36	9.9616	=AVERAGE($J$5:$J$154)	=ABS(J39-J40)	=SUM($L$5:$L$154)/(COUNT($I$5:$I$154)-1)	=M40*$J$1	=M40*$J$2

Then we have the following chart:



We see that this control chart with moving range seems much more sensitive than any previous control chart. But the problem remains as to know whether our observation of this control chart is correct in the sense that the moving range does not follow a symmetrical distribution law at the contrary to what presupposes fixed Shewhart limits.

Also be aware that some of the WECO rules (the rules for sequences for example) do not apply to this type of control chart, since the data are autocorrelated there.

16.7.6.5.3 Individual Moving Average control chart

We always assume here that the characteristic follows a Normal distribution (...) and that the samples were taken of constant size n (yes... Moving Average control charts can also be used for subgroups using moving range standard deviation or S estimation method, not only for individual measurement!!!). The moving average of duration h at time t, written M_t is defined by moving average seen in the section Statistics:

$$M_t=\frac{\overline{X}_t+\overline{X}_{t-1}+\ldots+\overline{X}_{t-h+1}}{h} \tag{16.7.837}$$

The parameter h is also named "horizon". Greater is the value of h better is the efficiency of this control chart for the detection of small deviations.

When $t\geq h$, assuming independence between samples, we get using the property of the variance:

$$\begin{aligned}\mathrm{V}(M_t)&=\mathrm{V}\left(\frac{\overline{X}_t+\overline{X}_{t-1}+\ldots+\overline{X}_{t-h+1}}{h}\right)=\frac{1}{h^2}\mathrm{V}\left(\overline{X}_t+\overline{X}_{t-1}+\ldots+\overline{X}_{t-h+1}\right)\\&=\frac{1}{h^2}\sum_{i=t-h+1}^{t}\mathrm{V}\left(\overline{X}_i\right)=\frac{1}{h^2}\sum_{i=t-h+1}^{t}\frac{\sigma^2}{n}=\frac{1}{h^2}h\frac{\sigma^2}{n}=\frac{\sigma^2}{hn}\end{aligned} \tag{16.7.838}$$

And if all the samples have the same size (if samples there are...)

$$\text{V}(M_t)=\frac{1}{h^2}\sum_{i=t-h+1}^{t}\text{V}\left(\overline{X}_i\right)=\frac{1}{h^2}\sum_{i=t-h+1}^{t}\frac{\sigma^2}{n}=\frac{1}{h^2}j\frac{\sigma^2}{n}=\frac{\sigma^2}{hn} \tag{16.7.839}$$

Therefore:

$$\sigma_{M_t}=\frac{\sigma}{\sqrt{hn}} \tag{16.7.840}$$

And:

$$\text{E}(M_t)=\text{E}\left(\frac{\overline{X}_t+\overline{X}_{t-1}+\ldots+\overline{X}_{t-h+1}}{h}\right)=\frac{1}{h}\text{E}\left(\overline{X}_t+\overline{X}_{t-1}+\ldots+\overline{X}_{t-h+1}\right)=\frac{1}{h}h\overline{\overline{X}}=\overline{\overline{X}} \tag{16.7.841}$$

So Finally:

$$\boxed{\begin{aligned}\text{UCL}_{\text{MA}}&=M_t+3\frac{\sigma}{\sqrt{hn}}\\ \text{CL}_{\text{MA}}&=\overline{\overline{X}}\\ \text{LCL}_{\text{MA}}&=M_t-3\frac{\sigma}{\sqrt{hn}}\end{aligned}} \tag{16.7.842}$$

Ok this done, let us now as always make a summary of the pros and cons of this control chart.

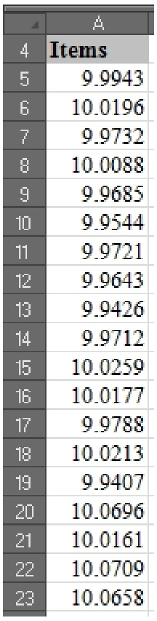
Pros:

- The fact that we smooth the fluctuations in the purpose to reduce strong variations makes this control chart very well suited for small productions for which the total number of items is around 10 (however as already mentioned this control chart can also be used for subgroup measurement). Also in the example that will follow, we will only focus on this special scenario of individual measurement!
- We also see that the boundaries of this control will be even smaller than h is large. So this control chart can be very sensitive to small variations on the long term.

Cons:

- The standard deviation is assumed to be known or at least calculated precisely so that the estimator converges to the theoretical value. All samples should be of identical size n if we use the method developed above (otherwise most S.P.C. softwares implement the necessary mathematical tools of this condition is not satisfied).

As companion example let us consider the following data:



	A
4	**Items**
5	9.9943
6	10.0196
7	9.9732
8	10.0088
9	9.9685
10	9.9544
11	9.9721
12	9.9643
13	9.9426
14	9.9712
15	10.0259
16	10.0177
17	9.9788
18	10.0213
19	9.9407
20	10.0696
21	10.0161
22	10.0709
23	10.0658

Since the control chart is with individual values, we have $n=1$. And the system using the moving range for the estimate of the standard deviation is reduced to:

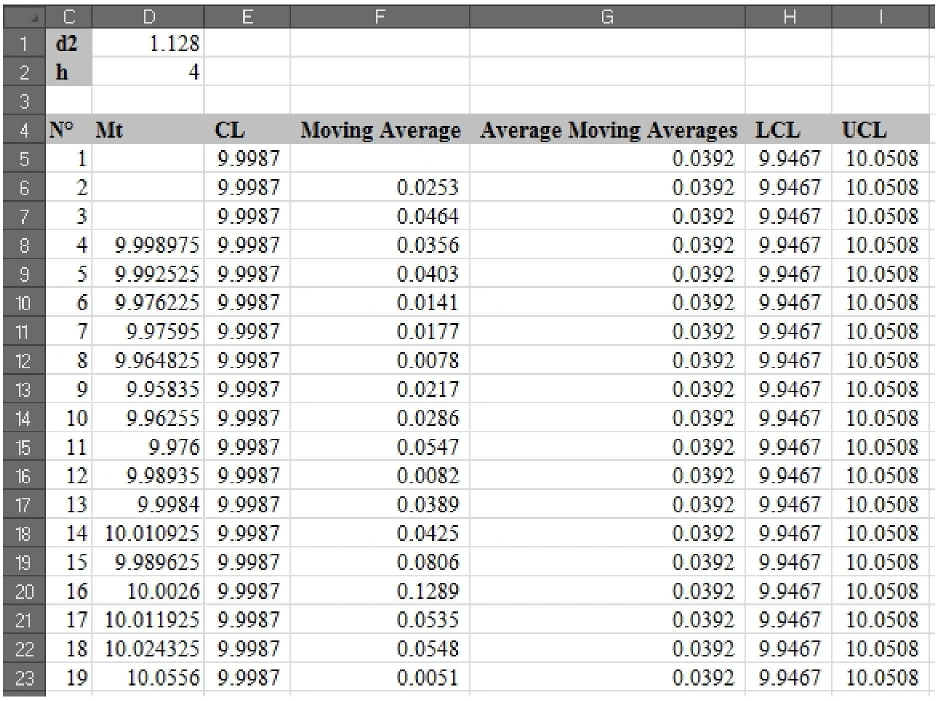
$$\boxed{\begin{aligned} \text{UCL}_{\text{MA}} &= \overline{\overline{X}}+3\frac{\sigma}{\sqrt{hn}}=\overline{\overline{X}}+3\frac{\overline{MR}}{d_2(n_i)\sqrt{h}} \\ \text{CL}_{\text{MA}} &= \overline{\overline{X}} \\ \text{LCL}_{\text{MA}} &= \overline{\overline{X}}-3\frac{\sigma}{\sqrt{hn}}=\overline{\overline{X}}-3\frac{\overline{MR}}{d_2(n_i)\sqrt{h}} \end{aligned}} \tag{16.7.843}$$

Therefore apart the denominators with the root of h, the parameters CL, UCL and LCL are the same as the I-MR $\overline{X}$ control chart. So we fall back on the same calculation method as the one provided par the NCSS Statistical Analysis & Graphics Software.

We must take in the case of our special example the value of d_2 for $n_i=2$. The tables or direct calculations give us:

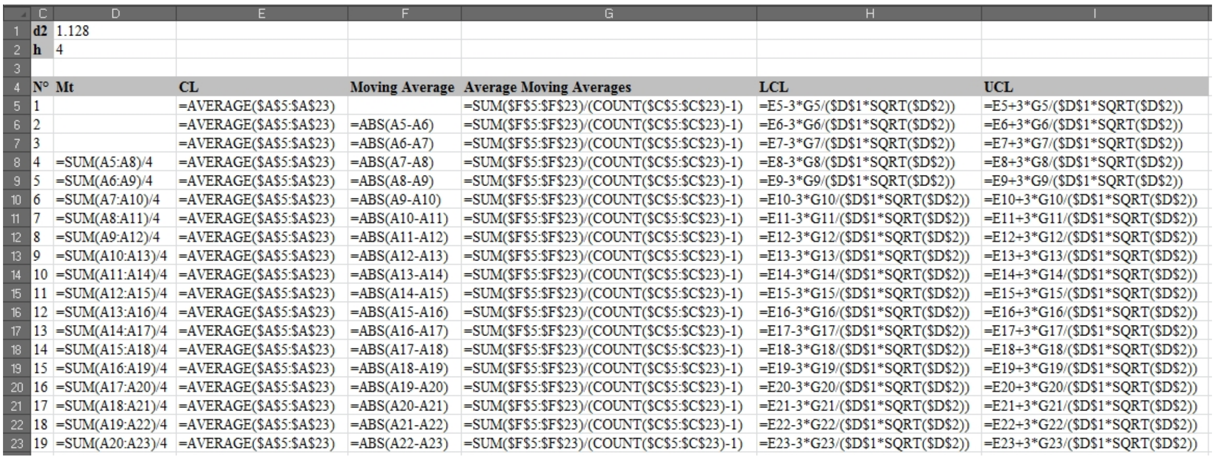
$$d_2(n_i)=d_2(2)=1.128 \tag{16.7.844}$$

Which gives us for the calculations:



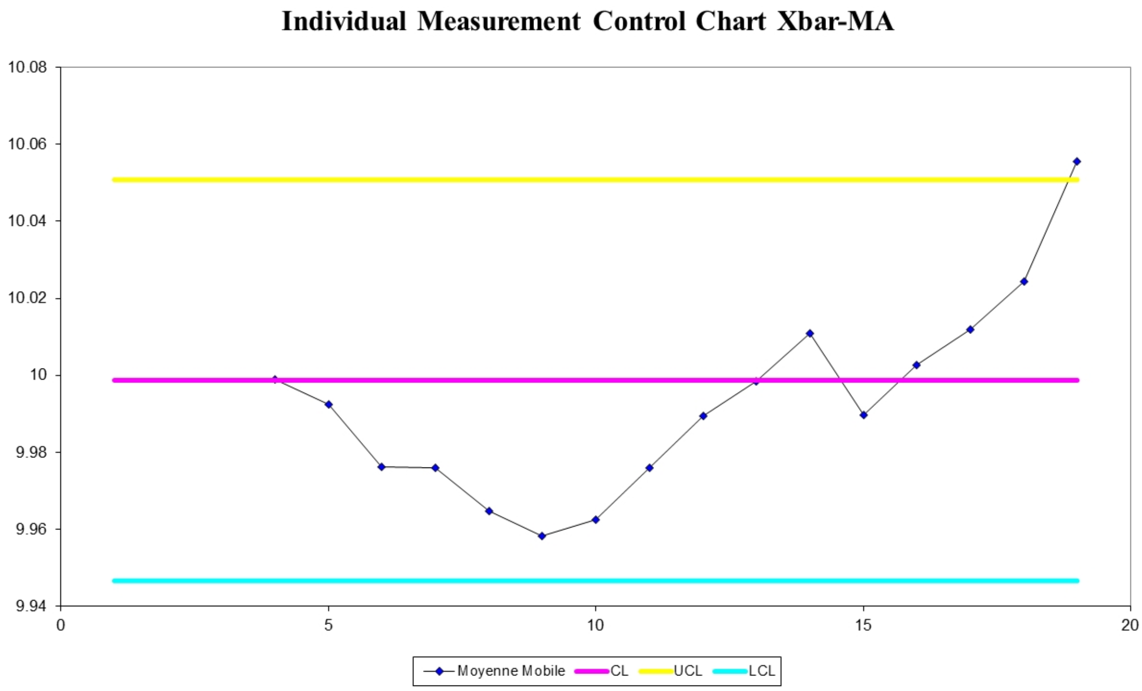
	C	D	E	F	G	H	I
1	**d2**	1.128					
2	**h**	4					
3							
4	**N°**	**Mt**	**CL**	**Moving Average**	**Average Moving Averages**	**LCL**	**UCL**
5	1		9.9987		0.0392	9.9467	10.0508
6	2		9.9987	0.0253	0.0392	9.9467	10.0508
7	3		9.9987	0.0464	0.0392	9.9467	10.0508
8	4	9.998975	9.9987	0.0356	0.0392	9.9467	10.0508
9	5	9.992525	9.9987	0.0403	0.0392	9.9467	10.0508
10	6	9.976225	9.9987	0.0141	0.0392	9.9467	10.0508
11	7	9.97595	9.9987	0.0177	0.0392	9.9467	10.0508
12	8	9.964825	9.9987	0.0078	0.0392	9.9467	10.0508
13	9	9.95835	9.9987	0.0217	0.0392	9.9467	10.0508
14	10	9.96255	9.9987	0.0286	0.0392	9.9467	10.0508
15	11	9.976	9.9987	0.0547	0.0392	9.9467	10.0508
16	12	9.98935	9.9987	0.0082	0.0392	9.9467	10.0508
17	13	9.9984	9.9987	0.0389	0.0392	9.9467	10.0508
18	14	10.010925	9.9987	0.0425	0.0392	9.9467	10.0508
19	15	9.989625	9.9987	0.0806	0.0392	9.9467	10.0508
20	16	10.0026	9.9987	0.1289	0.0392	9.9467	10.0508
21	17	10.011925	9.9987	0.0535	0.0392	9.9467	10.0508
22	18	10.024325	9.9987	0.0548	0.0392	9.9467	10.0508
23	19	10.0556	9.9987	0.0051	0.0392	9.9467	10.0508

Explicitly:



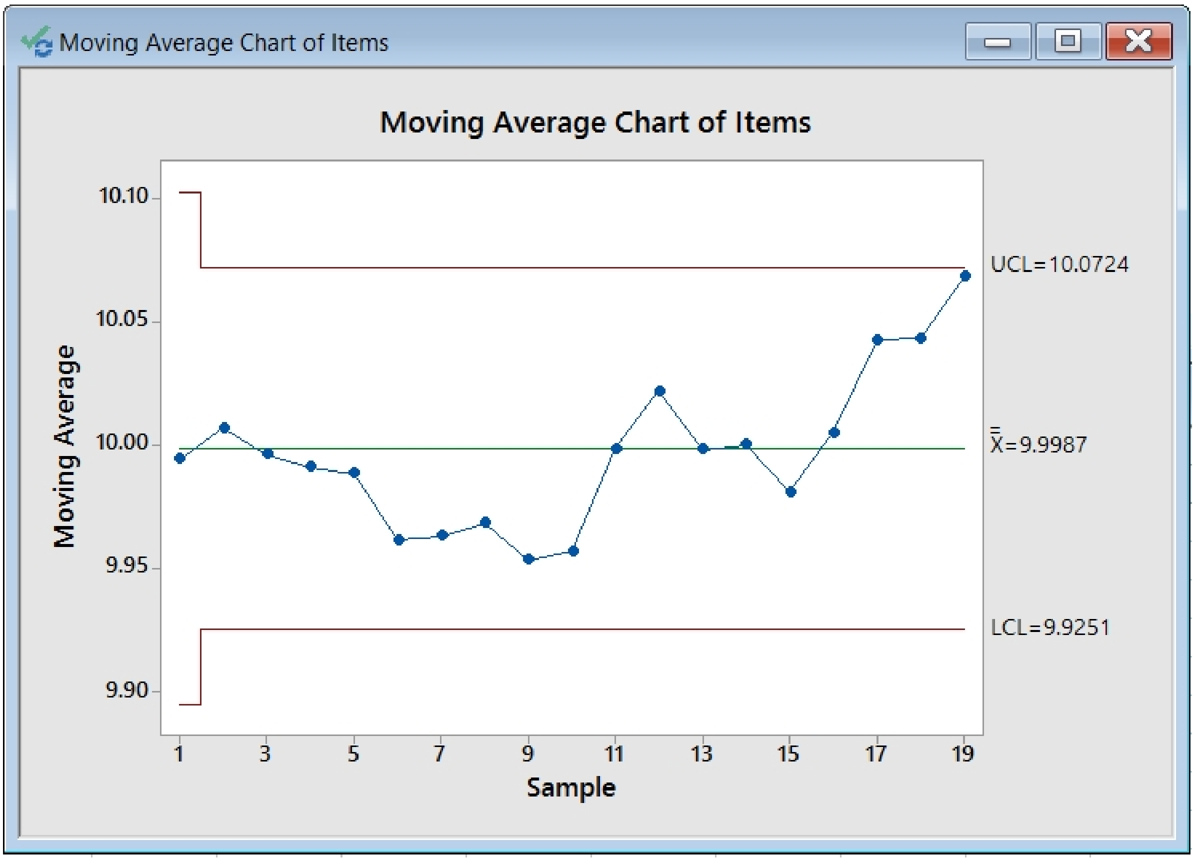
	C	D	E	F	G	H	I
1	**d2**	1.128					
2	**h**	4					
3							
4	**N°**	**Mt**	**CL**	**Moving Average**	**Average Moving Averages**	**LCL**	**UCL**
5	1		=AVERAGE($A$5:$A$23)		=SUM($F$5:$F$23)/(COUNT($C$5:$C$23)-1)	=E5-3*G5/($D$1*SQRT($D$2))	=E5+3*G5/($D$1*SQRT($D$2))
6	2		=AVERAGE($A$5:$A$23)	=ABS(A5-A6)	=SUM($F$5:$F$23)/(COUNT($C$5:$C$23)-1)	=E6-3*G6/($D$1*SQRT($D$2))	=E6+3*G6/($D$1*SQRT($D$2))
7	3		=AVERAGE($A$5:$A$23)	=ABS(A6-A7)	=SUM($F$5:$F$23)/(COUNT($C$5:$C$23)-1)	=E7-3*G7/($D$1*SQRT($D$2))	=E7+3*G7/($D$1*SQRT($D$2))
8	4	=SUM(A5:A8)/4	=AVERAGE($A$5:$A$23)	=ABS(A7-A8)	=SUM($F$5:$F$23)/(COUNT($C$5:$C$23)-1)	=E8-3*G8/($D$1*SQRT($D$2))	=E8+3*G8/($D$1*SQRT($D$2))
9	5	=SUM(A6:A9)/4	=AVERAGE($A$5:$A$23)	=ABS(A8-A9)	=SUM($F$5:$F$23)/(COUNT($C$5:$C$23)-1)	=E9-3*G9/($D$1*SQRT($D$2))	=E9+3*G9/($D$1*SQRT($D$2))
10	6	=SUM(A7:A10)/4	=AVERAGE($A$5:$A$23)	=ABS(A9-A10)	=SUM($F$5:$F$23)/(COUNT($C$5:$C$23)-1)	=E10-3*G10/($D$1*SQRT($D$2))	=E10+3*G10/($D$1*SQRT($D$2))
11	7	=SUM(A8:A11)/4	=AVERAGE($A$5:$A$23)	=ABS(A10-A11)	=SUM($F$5:$F$23)/(COUNT($C$5:$C$23)-1)	=E11-3*G11/($D$1*SQRT($D$2))	=E11+3*G11/($D$1*SQRT($D$2))
12	8	=SUM(A9:A12)/4	=AVERAGE($A$5:$A$23)	=ABS(A11-A12)	=SUM($F$5:$F$23)/(COUNT($C$5:$C$23)-1)	=E12-3*G12/($D$1*SQRT($D$2))	=E12+3*G12/($D$1*SQRT($D$2))
13	9	=SUM(A10:A13)/4	=AVERAGE($A$5:$A$23)	=ABS(A12-A13)	=SUM($F$5:$F$23)/(COUNT($C$5:$C$23)-1)	=E13-3*G13/($D$1*SQRT($D$2))	=E13+3*G13/($D$1*SQRT($D$2))
14	10	=SUM(A11:A14)/4	=AVERAGE($A$5:$A$23)	=ABS(A13-A14)	=SUM($F$5:$F$23)/(COUNT($C$5:$C$23)-1)	=E14-3*G14/($D$1*SQRT($D$2))	=E14+3*G14/($D$1*SQRT($D$2))
15	11	=SUM(A12:A15)/4	=AVERAGE($A$5:$A$23)	=ABS(A14-A15)	=SUM($F$5:$F$23)/(COUNT($C$5:$C$23)-1)	=E15-3*G15/($D$1*SQRT($D$2))	=E15+3*G15/($D$1*SQRT($D$2))
16	12	=SUM(A13:A16)/4	=AVERAGE($A$5:$A$23)	=ABS(A15-A16)	=SUM($F$5:$F$23)/(COUNT($C$5:$C$23)-1)	=E16-3*G16/($D$1*SQRT($D$2))	=E16+3*G16/($D$1*SQRT($D$2))
17	13	=SUM(A14:A17)/4	=AVERAGE($A$5:$A$23)	=ABS(A16-A17)	=SUM($F$5:$F$23)/(COUNT($C$5:$C$23)-1)	=E17-3*G17/($D$1*SQRT($D$2))	=E17+3*G17/($D$1*SQRT($D$2))
18	14	=SUM(A15:A18)/4	=AVERAGE($A$5:$A$23)	=ABS(A17-A18)	=SUM($F$5:$F$23)/(COUNT($C$5:$C$23)-1)	=E18-3*G18/($D$1*SQRT($D$2))	=E18+3*G18/($D$1*SQRT($D$2))
19	15	=SUM(A16:A19)/4	=AVERAGE($A$5:$A$23)	=ABS(A18-A19)	=SUM($F$5:$F$23)/(COUNT($C$5:$C$23)-1)	=E19-3*G19/($D$1*SQRT($D$2))	=E19+3*G19/($D$1*SQRT($D$2))
20	16	=SUM(A17:A20)/4	=AVERAGE($A$5:$A$23)	=ABS(A19-A20)	=SUM($F$5:$F$23)/(COUNT($C$5:$C$23)-1)	=E20-3*G20/($D$1*SQRT($D$2))	=E20+3*G20/($D$1*SQRT($D$2))
21	17	=SUM(A18:A21)/4	=AVERAGE($A$5:$A$23)	=ABS(A20-A21)	=SUM($F$5:$F$23)/(COUNT($C$5:$C$23)-1)	=E21-3*G21/($D$1*SQRT($D$2))	=E21+3*G21/($D$1*SQRT($D$2))
22	18	=SUM(A19:A22)/4	=AVERAGE($A$5:$A$23)	=ABS(A21-A22)	=SUM($F$5:$F$23)/(COUNT($C$5:$C$23)-1)	=E22-3*G22/($D$1*SQRT($D$2))	=E22+3*G22/($D$1*SQRT($D$2))
23	19	=SUM(A20:A23)/4	=AVERAGE($A$5:$A$23)	=ABS(A22-A23)	=SUM($F$5:$F$23)/(COUNT($C$5:$C$23)-1)	=E23-3*G23/($D$1*SQRT($D$2))	=E23+3*G23/($D$1*SQRT($D$2))

Then we have the following chart:



Notice that this control chart would have according to the WECO rules 1 point beyond the $\pm3\sigma$ based (we indicate that for comparison reasons with previous control charts and the next one).

We don't get the same results with Minitab 17.1.3 (or any known prior version this software) as they seem to use a strange method to calculate the standard deviation for which we can not found any proof (...):


Moving Average Chart of Items
Moving Average Chart of Items
Moving Average
10.10
10.05
10.00
9.95
9.90
1
3
5
7
9
11
13
15
17
19
Sample
UCL=10.0724
X̿=9.9987
LCL=9.9251

16.7.6.5.4 CUSUM (cumulated sum) control chart with empirical V-mask

A major disadvantage of the control chart of the Shewhart type is to base that analysis on the latest information collected (assumption of independence of measurements). They ignore the information of trends of the process contained in the latest estimates. You should know that many quality standards institutes worldwide (AFNOR or ASQ) consider Shewhart charts as completely obsolete (wrongly or rightly?) and advice the usage of CUSUM control charts because the latter use autocorrelation and amplify the variations.

In the years 11950s (holocene calendar) control charts named "CUSUM (cumulated sum)" or "Page-Hinkley control chart" (in honour of their inventors) were introduced by statisticians to detect small variations. Although they are mathematically more optimal, they are difficult to properly configure and very sensitive to the initial settings.

The principle of CUSUM control charts (because there are several for one given distribution law...) is to sum the differences between the estimates of the position of the method / process and its target. When the accumulated positive variances or accumulated negative variances exceed a certain value, we conclude with a off-centering of the process.

We put:

$$C_n=\sum_{i=1}^{n}(X_i-\mu) \tag{16.7.845}$$

and:

$$S_n=\frac{C_n}{\sigma} \tag{16.7.846}$$

that is obviously a sum of centered reduced variables!

Where often the first equality is written:

$$C_n=\sum_{i=1}^{n}(X_i-\mu)=(X_n-\mu)+\sum_{i=1}^{n-1}(X_i-\mu)=(X_n-\mu)+C_{n-1} \tag{16.7.847}$$

Therefore there comes a famous notation in the literature:

$$S_n=\frac{C_n}{\sigma}=\frac{(X_n-\mu)+C_{n-1}}{\sigma}=\frac{X_n-\mu}{\sigma}+S_{n-1}=Z_n+S_{n-1} \tag{16.7.848}$$

In general, the CUSUM control chart behaves as follows:

- If the process is under control, the cumulative sum fluctuates randomly around zero.
- If the average undergoes a positive drift of the mean, the statistic C_n or S_n grows quick by accumulation of a systematic bias.

Remarks

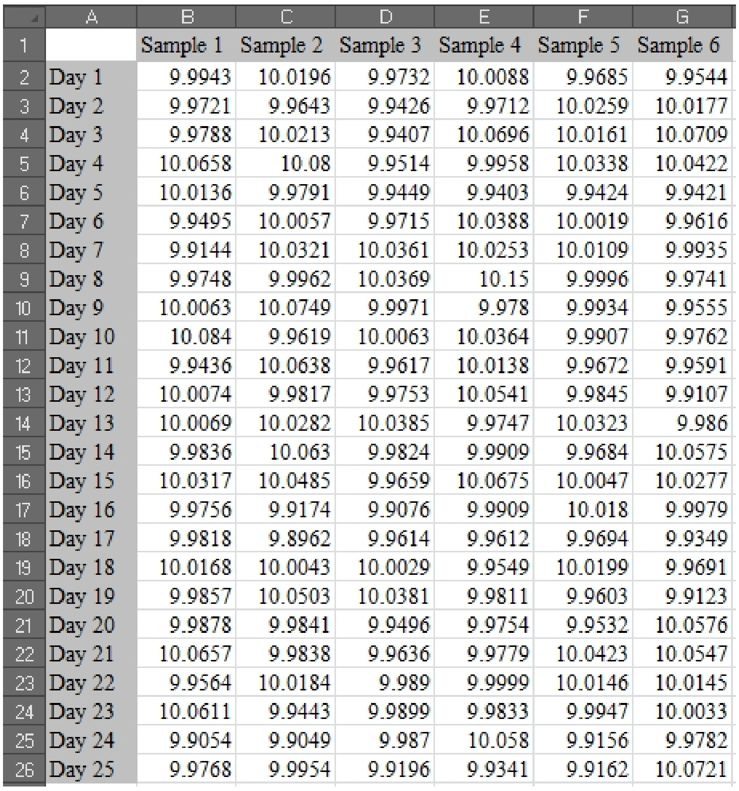
R1. In many books and numerous application cases, we do not use S_n because the variance is unknown and we restrict ourselves to the use of C_n only. The mathematical developments which then follows can be adapted very easily.

R2. The same approach can be made with average of samples following:

$$C_n = \sum_{i=1}^{n} \left(\overline{X}_i - \mu_{\overline{X}}\right) \qquad \text{and} \qquad S_n = \frac{C_n}{\sigma_{\overline{X}}} \tag{16.7.849}$$

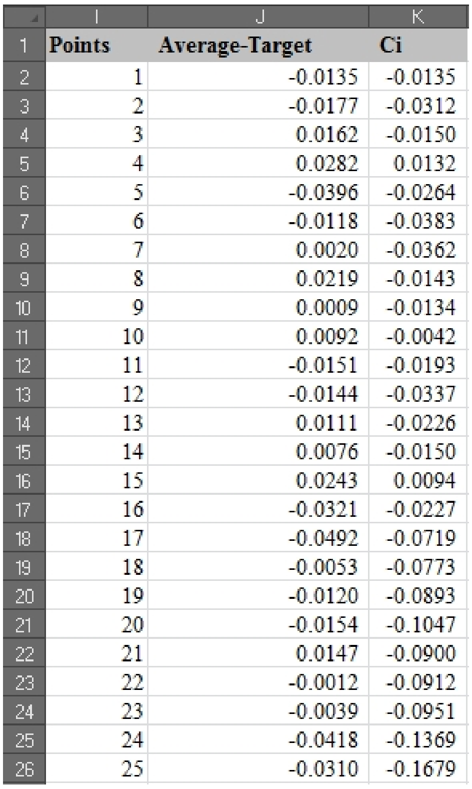
R3. There are many different versions of the CUSUM control charts and this even for the CUSUM with V-mask there are multiple sub-families. To be honest there is in my opinion quite a mess...

For the example, let us take again the following data table:



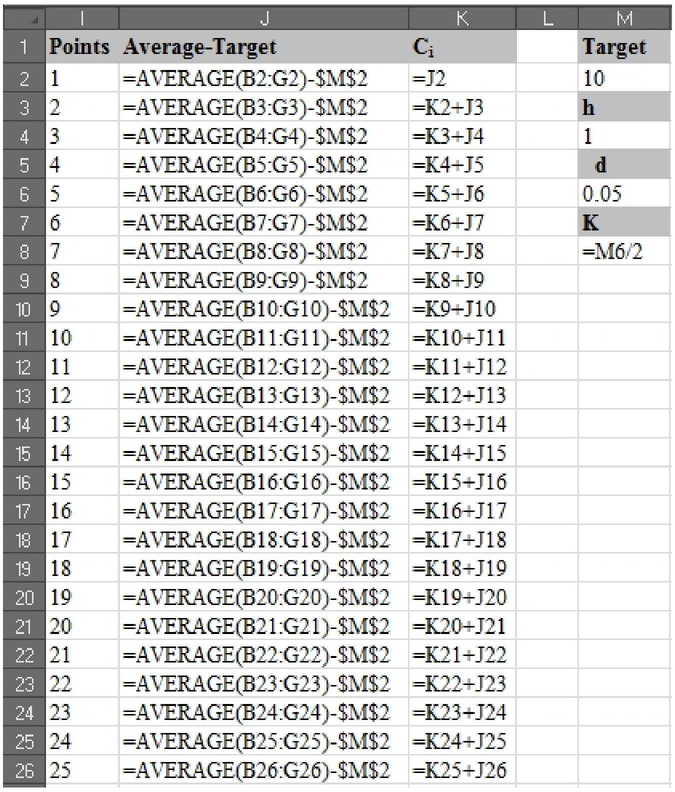
	A	B	C	D	E	F	G
1		Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
2	Day 1	9.9943	10.0196	9.9732	10.0088	9.9685	9.9544
3	Day 2	9.9721	9.9643	9.9426	9.9712	10.0259	10.0177
4	Day 3	9.9788	10.0213	9.9407	10.0696	10.0161	10.0709
5	Day 4	10.0658	10.08	9.9514	9.9958	10.0338	10.0422
6	Day 5	10.0136	9.9791	9.9449	9.9403	9.9424	9.9421
7	Day 6	9.9495	10.0057	9.9715	10.0388	10.0019	9.9616
8	Day 7	9.9144	10.0321	10.0361	10.0253	10.0109	9.9935
9	Day 8	9.9748	9.9962	10.0369	10.15	9.9996	9.9741
10	Day 9	10.0063	10.0749	9.9971	9.978	9.9934	9.9555
11	Day 10	10.084	9.9619	10.0063	10.0364	9.9907	9.9762
12	Day 11	9.9436	10.0638	9.9617	10.0138	9.9672	9.9591
13	Day 12	10.0074	9.9817	9.9753	10.0541	9.9845	9.9107
14	Day 13	10.0069	10.0282	10.0385	9.9747	10.0323	9.986
15	Day 14	9.9836	10.063	9.9824	9.9909	9.9684	10.0575
16	Day 15	10.0317	10.0485	9.9659	10.0675	10.0047	10.0277
17	Day 16	9.9756	9.9174	9.9076	9.9909	10.018	9.9979
18	Day 17	9.9818	9.8962	9.9614	9.9612	9.9694	9.9349
19	Day 18	10.0168	10.0043	10.0029	9.9549	10.0199	9.9691
20	Day 19	9.9857	10.0503	10.0381	9.9811	9.9603	9.9123
21	Day 20	9.9878	9.9841	9.9496	9.9754	9.9532	10.0576
22	Day 21	10.0657	9.9838	9.9636	9.9779	10.0423	10.0547
23	Day 22	9.9564	10.0184	9.989	9.9999	10.0146	10.0145
24	Day 23	10.0611	9.9443	9.9899	9.9833	9.9947	10.0033
25	Day 24	9.9054	9.9049	9.987	10.058	9.9156	9.9782
26	Day 25	9.9768	9.9954	9.9196	9.9341	9.9162	10.0721

and we calculate the cumulative quantity still considering a target $\mu = T = 10$:



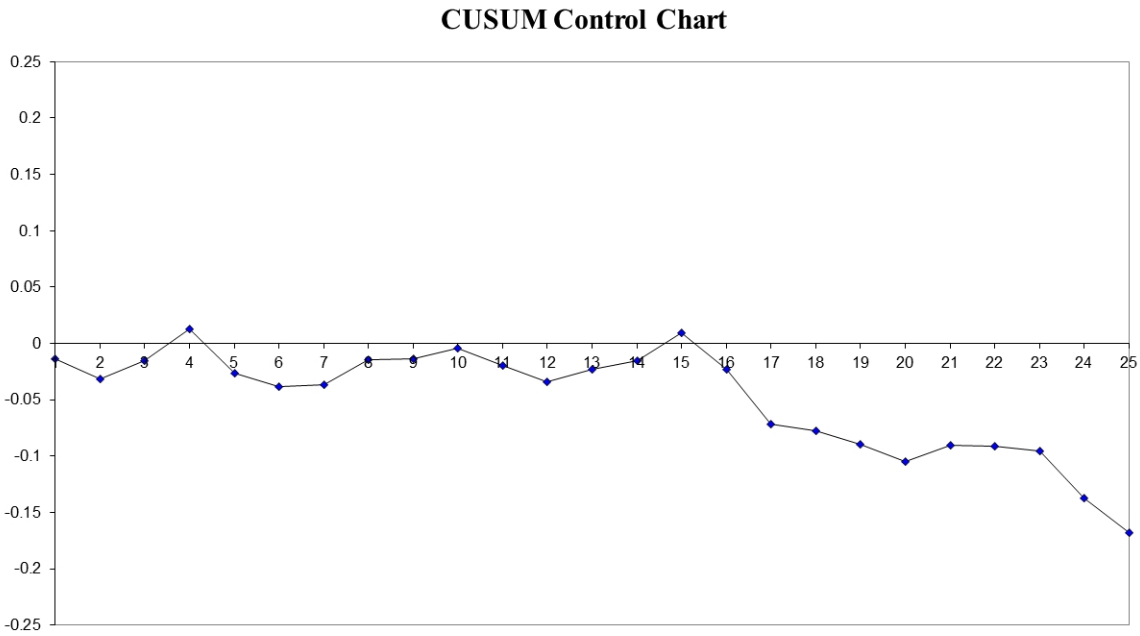
	I	J	K
1	**Points**	**Average-Target**	**Ci**
2	1	-0.0135	-0.0135
3	2	-0.0177	-0.0312
4	3	0.0162	-0.0150
5	4	0.0282	0.0132
6	5	-0.0396	-0.0264
7	6	-0.0118	-0.0383
8	7	0.0020	-0.0362
9	8	0.0219	-0.0143
10	9	0.0009	-0.0134
11	10	0.0092	-0.0042
12	11	-0.0151	-0.0193
13	12	-0.0144	-0.0337
14	13	0.0111	-0.0226
15	14	0.0076	-0.0150
16	15	0.0243	0.0094
17	16	-0.0321	-0.0227
18	17	-0.0492	-0.0719
19	18	-0.0053	-0.0773
20	19	-0.0120	-0.0893
21	20	-0.0154	-0.1047
22	21	0.0147	-0.0900
23	22	-0.0012	-0.0912
24	23	-0.0039	-0.0951
25	24	-0.0418	-0.1369
26	25	-0.0310	-0.1679

Thus explicitly:



	I	J	K	L	M
1	**Points**	**Average-Target**	**C_i**		**Target**
2	1	=AVERAGE(B2:G2)-$M$2	=J2		10
3	2	=AVERAGE(B3:G3)-$M$2	=K2+J3		**h**
4	3	=AVERAGE(B4:G4)-$M$2	=K3+J4		1
5	4	=AVERAGE(B5:G5)-$M$2	=K4+J5		**d**
6	5	=AVERAGE(B6:G6)-$M$2	=K5+J6		0.05
7	6	=AVERAGE(B7:G7)-$M$2	=K6+J7		**K**
8	7	=AVERAGE(B8:G8)-$M$2	=K7+J8		=M6/2
9	8	=AVERAGE(B9:G9)-$M$2	=K8+J9		
10	9	=AVERAGE(B10:G10)-$M$2	=K9+J10		
11	10	=AVERAGE(B11:G11)-$M$2	=K10+J11		
12	11	=AVERAGE(B12:G12)-$M$2	=K11+J12		
13	12	=AVERAGE(B13:G13)-$M$2	=K12+J13		
14	13	=AVERAGE(B14:G14)-$M$2	=K13+J14		
15	14	=AVERAGE(B15:G15)-$M$2	=K14+J15		
16	15	=AVERAGE(B16:G16)-$M$2	=K15+J16		
17	16	=AVERAGE(B17:G17)-$M$2	=K16+J17		
18	17	=AVERAGE(B18:G18)-$M$2	=K17+J18		
19	18	=AVERAGE(B19:G19)-$M$2	=K18+J19		
20	19	=AVERAGE(B20:G20)-$M$2	=K19+J20		
21	20	=AVERAGE(B21:G21)-$M$2	=K20+J21		
22	21	=AVERAGE(B22:G22)-$M$2	=K21+J22		
23	22	=AVERAGE(B23:G23)-$M$2	=K22+J23		
24	23	=AVERAGE(B24:G24)-$M$2	=K23+J24		
25	24	=AVERAGE(B25:G25)-$M$2	=K24+J25		
26	25	=AVERAGE(B26:G26)-$M$2	=K25+J26		

Then we have the following chart:



We then very quickly see with this control chart that the cumulate sum seems to decrease. This indicates that the true mean quality is lower than the target (you can try with 9.994 as the target to see the sensitivity).

We must now consider two cases:

H1. It comes under the assumption of normality (process under control) and by the property of stability of the Normal distribution for a centered reduced random variable:

$$S_i = \mathcal{N}(\mu,\sigma) = \mathcal{N}\left(0,\sqrt{i}\right) \tag{16.7.850}$$

or otherwise:

$$C_n = \mathcal{N}(\mu,\sigma) = \mathcal{N}\left(0,\sigma\sqrt{i}\right) \tag{16.7.851}$$

Therefore, under this hypothesis, we have the cumulative sums $(S_i)_{i=1\ldots n}$ which are distributed around a regression line of slope 0 (since the average of the sums of a random variables with zero mean is zero...) and the distributions have their standard deviation, which is proportional to the root of i (respectively the variance which is proportional to i).

H2. If the process is out of control (global dispersion), there is a moment $t<n$ such as $X_i = \mathcal{N}(\mu,\sigma)$ for $i \leq r-1$ and $X_i = \mathcal{N}(\mu+\delta,\sigma)$ for $i \geq r$ with $\delta > 0$ (the latter value being often imposed by the constraints of quality and named "large middle translation"). Therefore, we have:

$$S_i = \mathcal{N}\left(0,\sqrt{i}\right) \tag{16.7.852}$$

or:

$$C_i = \mathcal{N}\left(0,\sigma\sqrt{i}\right) \tag{16.7.853}$$

for $i=1\ldots r-1$ and:

$$\begin{aligned}S_k-S_{r-1}&=\mathcal{N}\left(\frac{\delta}{\sigma},\sqrt{k}\right)-\mathcal{N}\left(0,\sqrt{r-1}\right)=\mathcal{N}\left(0,\sqrt{r-1}\right)\\&=\mathcal{N}\left(\frac{\delta}{\sigma}+0,\sqrt{\left(\sqrt{k}\right)^2+\left(\sqrt{r-1}\right)^2}\right)=\mathcal{N}\left(\frac{\delta}{\sigma}+0,\sqrt{k+(r-1)}\right)\end{aligned} \tag{16.7.854}$$

for $k\geq r$. That said, it is obvious that the cumulative sums:

$$S_k=\mathcal{N}\left(\frac{\delta}{\sigma},\sqrt{k}\right) \tag{16.7.855}$$

are arranged around regression line of slope δ/σ right from the moment r.

Respectively we have:

$$C_k=\mathcal{N}\left(\delta,\sigma\sqrt{k}\right) \tag{16.7.856}$$

To continue, the idea is to discriminate hypothesis H1 of a regression line of zero slope and the hypothesis H2 of a regression line of δ/σ for S_k by fixing for imaginary border the half of the slope (the right middle in other words...):

$$K_S=\frac{\delta}{2\sigma} \tag{16.7.857}$$

respectively:

$$K_C=\frac{\delta}{2} \tag{16.7.858}$$

and a shift (offset) $h\geq 0$ also empirical with $h\in\mathbb{N}$. At the step n we reject H1 if for an integer $m\leq n$, we have:

$$S_m\geq(n-m)K_S+S_1+hK_S \tag{16.7.859}$$

It is obviously also possible to do it by reject with a geometrical representation (this has the advantage of going much faster and to the control in one time only for any m less than or equal to n). The idea then is to draw a line of slope K starting from the point of abscissa $n+1h$. If there exist a point (m,S_m) with $m<n$ below the regression line, then we pull an alarm.

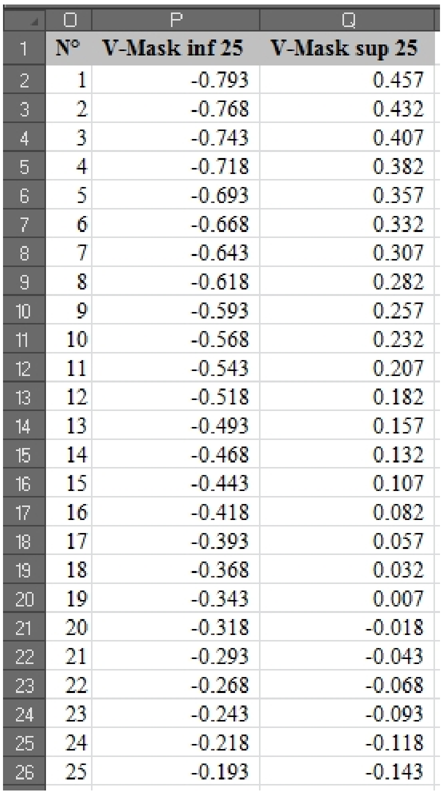
Let us do an example with the previous graph. Consider the following empirical values of the parameters:

$$\delta=0.05\qquad\text{and}\qquad h=1 \tag{16.7.860}$$

Thus:

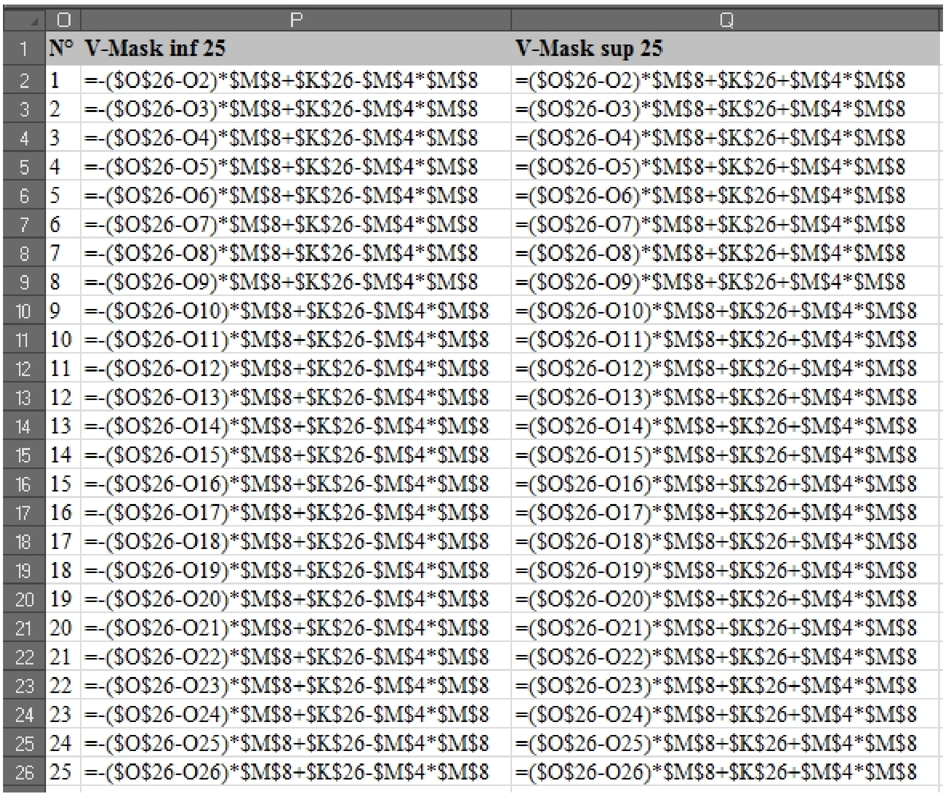
$$K_C=0.025 \tag{16.7.861}$$

Therefore:



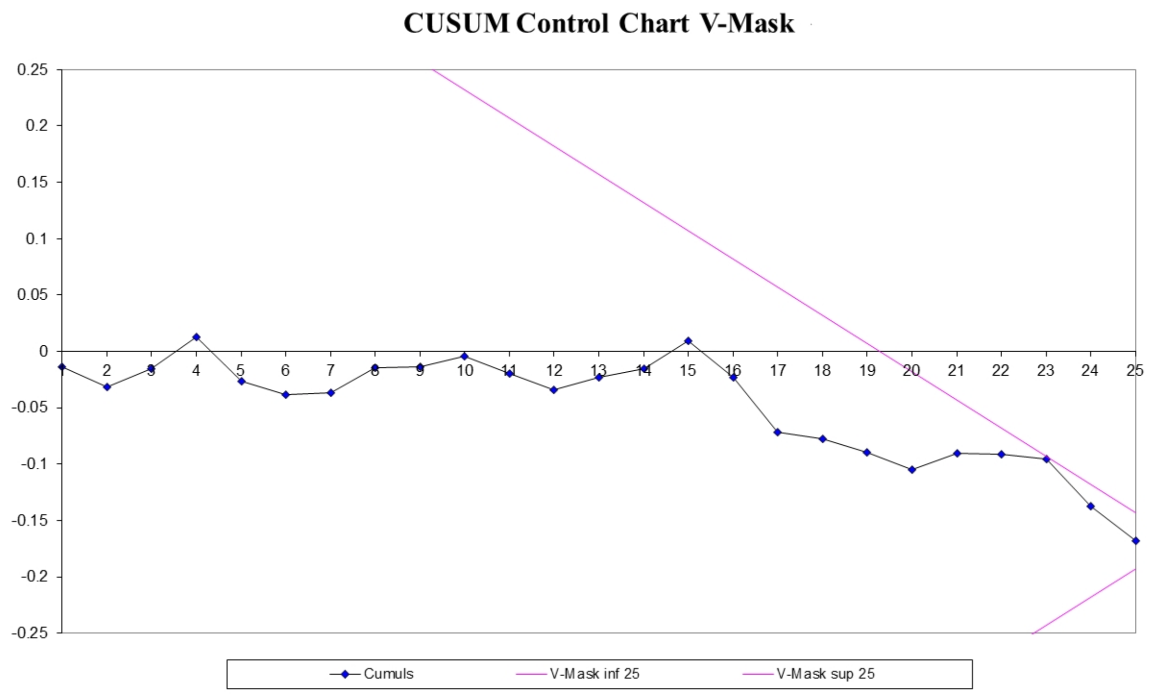
N°	V-Mask inf 25	V-Mask sup 25
1	-0.793	0.457
2	-0.768	0.432
3	-0.743	0.407
4	-0.718	0.382
5	-0.693	0.357
6	-0.668	0.332
7	-0.643	0.307
8	-0.618	0.282
9	-0.593	0.257
10	-0.568	0.232
11	-0.543	0.207
12	-0.518	0.182
13	-0.493	0.157
14	-0.468	0.132
15	-0.443	0.107
16	-0.418	0.082
17	-0.393	0.057
18	-0.368	0.032
19	-0.343	0.007
20	-0.318	-0.018
21	-0.293	-0.043
22	-0.268	-0.068
23	-0.243	-0.093
24	-0.218	-0.118
25	-0.193	-0.143

Explicitly:

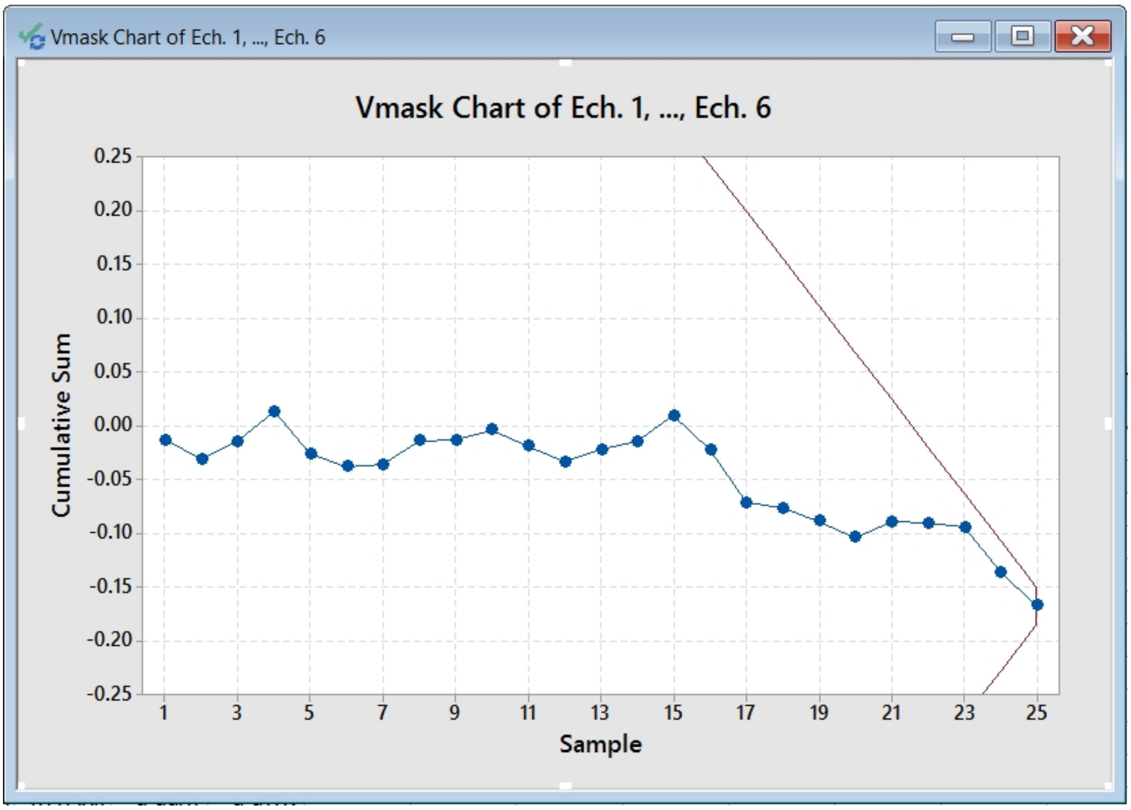


N°	V-Mask inf 25	V-Mask sup 25
1	=-($O$26-O2)*$M$8+$K$26-$M$4*$M$8	=($O$26-O2)*$M$8+$K$26+$M$4*$M$8
2	=-($O$26-O3)*$M$8+$K$26-$M$4*$M$8	=($O$26-O3)*$M$8+$K$26+$M$4*$M$8
3	=-($O$26-O4)*$M$8+$K$26-$M$4*$M$8	=($O$26-O4)*$M$8+$K$26+$M$4*$M$8
4	=-($O$26-O5)*$M$8+$K$26-$M$4*$M$8	=($O$26-O5)*$M$8+$K$26+$M$4*$M$8
5	=-($O$26-O6)*$M$8+$K$26-$M$4*$M$8	=($O$26-O6)*$M$8+$K$26+$M$4*$M$8
6	=-($O$26-O7)*$M$8+$K$26-$M$4*$M$8	=($O$26-O7)*$M$8+$K$26+$M$4*$M$8
7	=-($O$26-O8)*$M$8+$K$26-$M$4*$M$8	=($O$26-O8)*$M$8+$K$26+$M$4*$M$8
8	=-($O$26-O9)*$M$8+$K$26-$M$4*$M$8	=($O$26-O9)*$M$8+$K$26+$M$4*$M$8
9	=-($O$26-O10)*$M$8+$K$26-$M$4*$M$8	=($O$26-O10)*$M$8+$K$26+$M$4*$M$8
10	=-($O$26-O11)*$M$8+$K$26-$M$4*$M$8	=($O$26-O11)*$M$8+$K$26+$M$4*$M$8
11	=-($O$26-O12)*$M$8+$K$26-$M$4*$M$8	=($O$26-O12)*$M$8+$K$26+$M$4*$M$8
12	=-($O$26-O13)*$M$8+$K$26-$M$4*$M$8	=($O$26-O13)*$M$8+$K$26+$M$4*$M$8
13	=-($O$26-O14)*$M$8+$K$26-$M$4*$M$8	=($O$26-O14)*$M$8+$K$26+$M$4*$M$8
14	=-($O$26-O15)*$M$8+$K$26-$M$4*$M$8	=($O$26-O15)*$M$8+$K$26+$M$4*$M$8
15	=-($O$26-O16)*$M$8+$K$26-$M$4*$M$8	=($O$26-O16)*$M$8+$K$26+$M$4*$M$8
16	=-($O$26-O17)*$M$8+$K$26-$M$4*$M$8	=($O$26-O17)*$M$8+$K$26+$M$4*$M$8
17	=-($O$26-O18)*$M$8+$K$26-$M$4*$M$8	=($O$26-O18)*$M$8+$K$26+$M$4*$M$8
18	=-($O$26-O19)*$M$8+$K$26-$M$4*$M$8	=($O$26-O19)*$M$8+$K$26+$M$4*$M$8
19	=-($O$26-O20)*$M$8+$K$26-$M$4*$M$8	=($O$26-O20)*$M$8+$K$26+$M$4*$M$8
20	=-($O$26-O21)*$M$8+$K$26-$M$4*$M$8	=($O$26-O21)*$M$8+$K$26+$M$4*$M$8
21	=-($O$26-O22)*$M$8+$K$26-$M$4*$M$8	=($O$26-O22)*$M$8+$K$26+$M$4*$M$8
22	=-($O$26-O23)*$M$8+$K$26-$M$4*$M$8	=($O$26-O23)*$M$8+$K$26+$M$4*$M$8
23	=-($O$26-O24)*$M$8+$K$26-$M$4*$M$8	=($O$26-O24)*$M$8+$K$26+$M$4*$M$8
24	=-($O$26-O25)*$M$8+$K$26-$M$4*$M$8	=($O$26-O25)*$M$8+$K$26+$M$4*$M$8
25	=-($O$26-O26)*$M$8+$K$26-$M$4*$M$8	=($O$26-O26)*$M$8+$K$26+$M$4*$M$8

We then havefor example at the point 26 corresponding to $25+h$:



With Minitab 17.1.3 this gives for people that don't trust our calculations made with Microsoft® Excel®:



The reader can easily observe that with the choice we made, we would have no alarm with the selected empirical parameters. It would be necessary in practice to reduce the value δ.

To resume, the benefits of CUSUM control charts compared to Shewhart control charts are:

- They are more effective for detecting small drifts
- The drift of the process visually appears on the control chart
- It is usually easy to detect at what point the process began to drift

The disadvantages of these control charts are the following:

- They can be slow to detect large drifts
- They are not very interesting to analyse past data or to look for cycles or find some type characteristics in the studied distribution.
- They are less easily accepted by operators as less intuitive and do not directly represent the characteristic of interest.

For the reasons discussed above, it is advisable to represent in parallel of CUSUM control charts a classical Shewhart control chart.

Remark

There are still other CUSUM V-mask type. So be careful with the software you use to be sure what you are doing!!!

16.7.6.5.5 EWMA control charts (exponential weighted moving average) with fixed limits

An interesting alternative to the CUSUM control charts in the context of autocorrelation and the detection of small deviations are the EWMA (Exponentially Weighted Moving Average) control charts that have their origin in the field of Time Series Analysis (see section Economy page 7896).

Remark

Some also use control charts using empirical techniques like the simple or double moving average as discussed in section Economy when data are serially correlated.

An EWMA control chart is easy to implement and is not too sensitive to parameter changes and non-Normal data, it remains just as powerful as the CUSUM control charts to detect small variations.

By definition, the statistics reported on a EWMA control chart are calculated by the following equation (notice that the indices of the different terms are not the same as exponential smoothing of a time series as seen in the section Economy at page 7928 since the goal is not to make predictions !!):

$$F_t = \lambda \overline{X}_t + (1-\lambda)F_{t-1} \qquad \forall \lambda \in]0.6, 0.8[\tag{16.7.862}$$

with the particularity that some softwares (like Minitab) take for this control chart:

$$F_0=\overline{X}_1 \quad \text{or} \quad F_0=\overline{X} \tag{16.7.863}$$

The constant λ (smoothing constant) determines the weight that we want to assign to the latest measures. The smallest is this value, more the control chart is sensitive to sudden deviations.

Let us prove that the last relation can be written as:

$$F_t=\left(\sum_{i=0}^{t-1}\lambda(1-\lambda)^i\overline{X}_{t-i}\right)+(1-\lambda)^tF_0 \tag{16.7.864}$$

Indeed:

$$\begin{aligned}F_t&=\lambda\overline{X}_t+(1-\lambda)F_{t-1}=\lambda\overline{X}_t+(1-\lambda)\left(\lambda\overline{X}_{t-1}+(1-\lambda)F_{t-2}\right)\\&=\lambda\overline{X}_t+(1-\lambda)\left(\lambda\overline{X}_{t-1}+(1-\lambda)\left(\lambda\overline{X}_{t-2}+(1-\lambda)F_{t-3}\right)\right)\\&=\left(\sum_{i=0}^{t-1}\lambda(1-\lambda)\overline{X}_{t-1}\right)+(1-\lambda)^tF_0\end{aligned} \tag{16.7.865}$$

Let us suppose for what will follow that:

$$X_i=\mathcal{N}(\mu,\sigma) \tag{16.7.866}$$

Then we have:

$$\text{V}(F_t)=\text{V}\left(\left(\sum_{i=0}^{t-1}\lambda(1-\lambda)^i\overline{X}_{t-1}\right)+(1-\lambda)^tF_0\right) \tag{16.7.867}$$

and remembering one of the properties of the variance (see section Statistics page 596):

$$\text{V}(a\overline{X})=a^2\text{V}\left(\overline{X}\right) \tag{16.7.868}$$

It comes:

$$\begin{aligned}\text{V}(F_t)&=\left(\sum_{i=0}^{t-1}\lambda^2(1-\lambda)^{2i}\text{V}(\overline{X}_{t-i})\right)+(1-\lambda)^{2t}\text{V}(F_0)\\&=\left(\sum_{i=0}^{t-1}\lambda^2(1-\lambda)^{2i}\text{V}(\overline{X}_{t-i})\right)+(1-\lambda)^{2t}\cdot 0\\&=\sum_{i=0}^{t-1}\lambda^2(1-\lambda)^{2i}\text{V}(\overline{X}_{t-i})=\lambda^2\sum_{i=0}^{t-1}(1-\lambda)^{2i}\sigma^2=\lambda^2\sigma^2\sum_{i=0}^{t-1}(1-\lambda)^{2i}\end{aligned} \tag{16.7.869}$$

Let us do a change of variable:

$$x=(1-\lambda)^2 \tag{16.7.870}$$

Then it comes using the result proved in the context of our study of arithmetic sequences in the section of Sequences and Series (see page 1679):

$$\begin{aligned}\text{V}(F_t)&=\lambda^2\sigma^2\sum_{i=0}^{t-1}x^i=\lambda^2\sigma^2\frac{1-x^{(t-1)+1}}{1-x}=\lambda^2\sigma^2\frac{1-(1-\lambda)^{2t}}{1-(1-\lambda)^2}\\&=\lambda^2\sigma^2\frac{1-(1-\lambda)^{2t}}{1-1+2\lambda+\lambda^2}=\sigma^2\frac{\lambda^2}{2\lambda+\lambda^2}\left(1-(1-\lambda)^{2t}\right)\\&=\sigma^2\frac{\lambda}{2+\lambda}\left(1-(1-\lambda)^{2t}\right)\end{aligned} \tag{16.7.871}$$

when t is sufficiently large, then the variance is reduced to (approximation that the majority of softwares - like Minitab - don't do and therefore the limits are not constant depending on t!):

$$\text{V}(F_t)=\sigma_{F_t}^2\cong\sigma^2\frac{\lambda}{2+\lambda} \tag{16.7.872}$$

the EWMA control charts limits then are also build with the $\pm\sigma$:

$$\sigma_{\text{EWMA}}=\sigma\sqrt{\frac{\lambda}{2+\lambda}} \tag{16.7.873}$$

We therefore have when a target and also a standard deviation are defined (it is impossible that the customer sets a LSL, USL or target T regarding to the exponential weighting hence the fact that these parameters are this time not listed):

$$\boxed{\begin{aligned}\text{UCL}_{\text{EWMA}}&=\overline{\overline{X}}+3\frac{\sigma}{\sqrt{n_i}}\sqrt{\frac{\lambda}{2+\lambda}}=\overline{\overline{X}}+3\sigma\sqrt{\frac{\lambda}{n_i(2+\lambda)}}\\ \text{CL}_{\text{EWMA}}&=\overline{\overline{X}}\\ \text{LCL}_{\text{EWMA}}&=\overline{\overline{X}}-3\frac{\sigma}{\sqrt{n_i}}\sqrt{\frac{\lambda}{2+\lambda}}=\overline{\overline{X}}-3\sigma\sqrt{\frac{\lambda}{n_i(2+\lambda)}}\end{aligned}} \tag{16.7.874}$$

Some softwares uses to calculate the standard deviation the following relation that we have proved earlier above:

$$\sigma=\frac{\overline{R}}{d_2(n_i)} \tag{16.7.875}$$

and others (such as Minitab):

$$\sigma=\frac{\overline{\text{MR}}}{d_2(2)} \tag{16.7.876}$$

the latter option has the advantage of working if the sample size is unitary (we speak then of course "EWMA for individual values"). Let us recall that in fact we have already mentioned that the EWMA control chart is poorly suited to mass production.

It should be noticed that regardless of the option chosen, as we have:

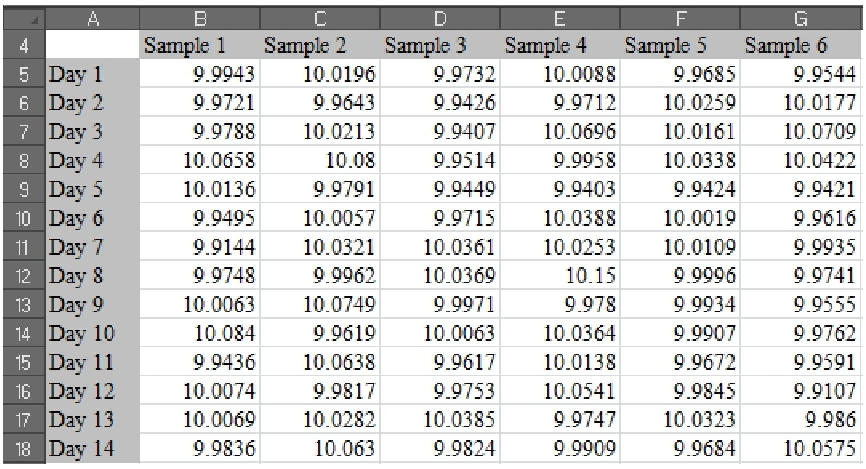
$$\sqrt{\frac{\lambda}{2+\lambda}}<1 \tag{16.7.877}$$

the controls limits are always significantly below those of a range or moving range control chart. This is also a reason why we smooth the data!

Let us do an example for each variant of calculation of the standard deviation since each is of equal importance in practice. To do this, let us start with the EWMA control chart with:

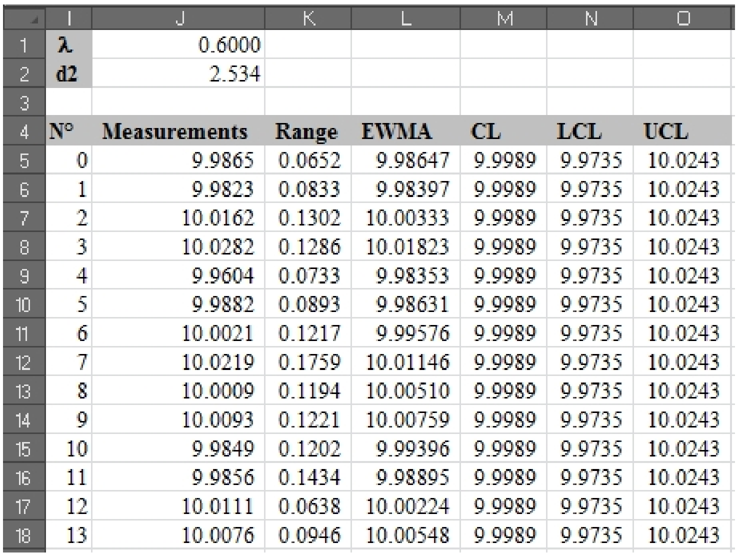
$$\begin{aligned}\text{UCL}_{\text{EWMA}}&=\overline{\overline{X}}+3\frac{\overline{R}}{d_2(n_i)}\sqrt{\frac{\lambda}{n_i(2+\lambda)}}\\ \text{CL}_{\text{EWMA}}&=\overline{\overline{X}}\\ \text{LCL}_{\text{EWMA}}&=\overline{\overline{X}}-3\frac{\overline{R}}{d_2(n_i)}\sqrt{\frac{\lambda}{n_i(2+\lambda)}}\end{aligned} \tag{16.7.878}$$

and taking the following table of data:



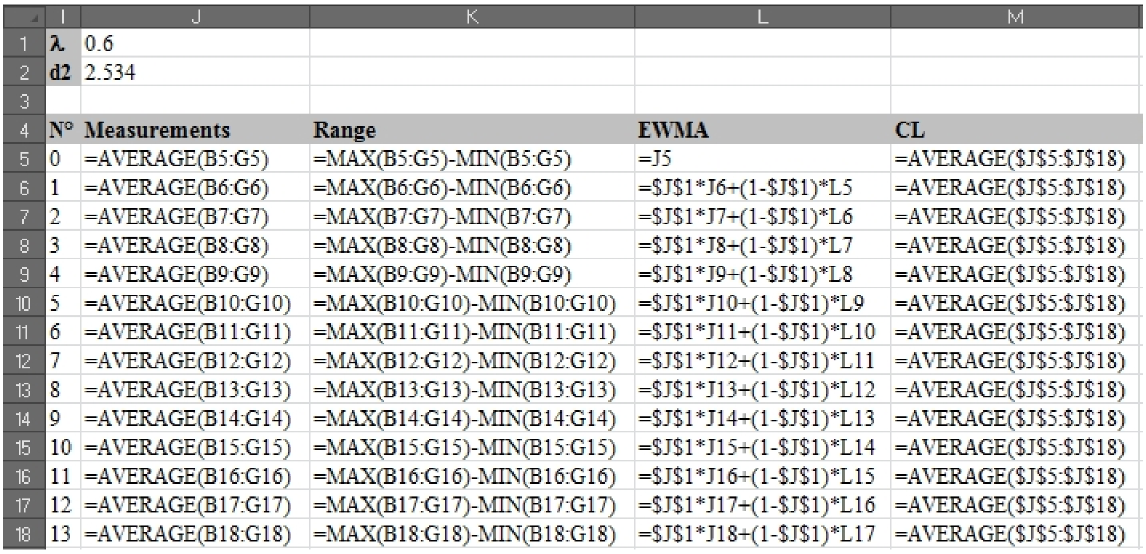
	A	B	C	D	E	F	G
4		Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
5	Day 1	9.9943	10.0196	9.9732	10.0088	9.9685	9.9544
6	Day 2	9.9721	9.9643	9.9426	9.9712	10.0259	10.0177
7	Day 3	9.9788	10.0213	9.9407	10.0696	10.0161	10.0709
8	Day 4	10.0658	10.08	9.9514	9.9958	10.0338	10.0422
9	Day 5	10.0136	9.9791	9.9449	9.9403	9.9424	9.9421
10	Day 6	9.9495	10.0057	9.9715	10.0388	10.0019	9.9616
11	Day 7	9.9144	10.0321	10.0361	10.0253	10.0109	9.9935
12	Day 8	9.9748	9.9962	10.0369	10.15	9.9996	9.9741
13	Day 9	10.0063	10.0749	9.9971	9.978	9.9934	9.9555
14	Day 10	10.084	9.9619	10.0063	10.0364	9.9907	9.9762
15	Day 11	9.9436	10.0638	9.9617	10.0138	9.9672	9.9591
16	Day 12	10.0074	9.9817	9.9753	10.0541	9.9845	9.9107
17	Day 13	10.0069	10.0282	10.0385	9.9747	10.0323	9.986
18	Day 14	9.9836	10.063	9.9824	9.9909	9.9684	10.0575

We get:



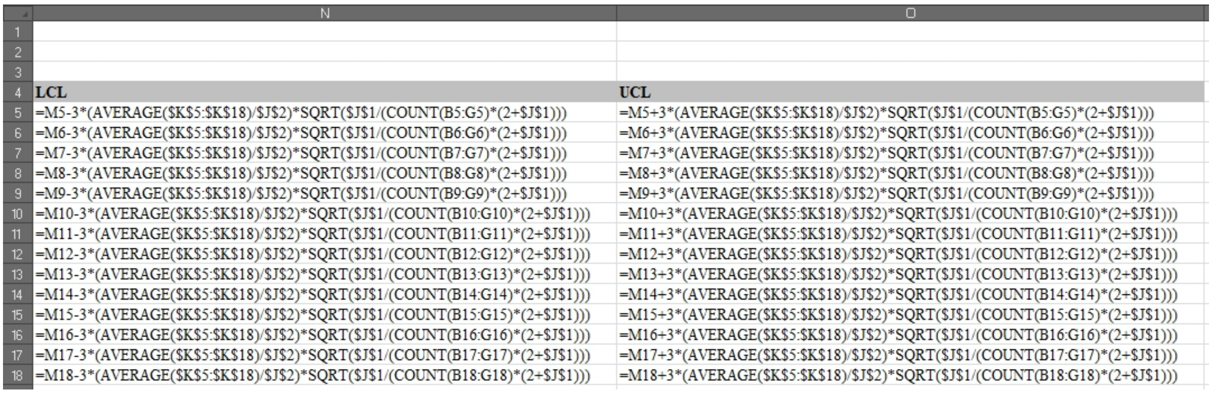
	I	J	K	L	M	N	O
1	**λ**	0.6000					
2	**d2**	2.534					
3							
4	**N°**	**Measurements**	**Range**	**EWMA**	**CL**	**LCL**	**UCL**
5	0	9.9865	0.0652	9.98647	9.9989	9.9735	10.0243
6	1	9.9823	0.0833	9.98397	9.9989	9.9735	10.0243
7	2	10.0162	0.1302	10.00333	9.9989	9.9735	10.0243
8	3	10.0282	0.1286	10.01823	9.9989	9.9735	10.0243
9	4	9.9604	0.0733	9.98353	9.9989	9.9735	10.0243
10	5	9.9882	0.0893	9.98631	9.9989	9.9735	10.0243
11	6	10.0021	0.1217	9.99576	9.9989	9.9735	10.0243
12	7	10.0219	0.1759	10.01146	9.9989	9.9735	10.0243
13	8	10.0009	0.1194	10.00510	9.9989	9.9735	10.0243
14	9	10.0093	0.1221	10.00759	9.9989	9.9735	10.0243
15	10	9.9849	0.1202	9.99396	9.9989	9.9735	10.0243
16	11	9.9856	0.1434	9.98895	9.9989	9.9735	10.0243
17	12	10.0111	0.0638	10.00224	9.9989	9.9735	10.0243
18	13	10.0076	0.0946	10.00548	9.9989	9.9735	10.0243

Thus explicitly (we have cut the screenshot into two parts because of horizontal space on the page):



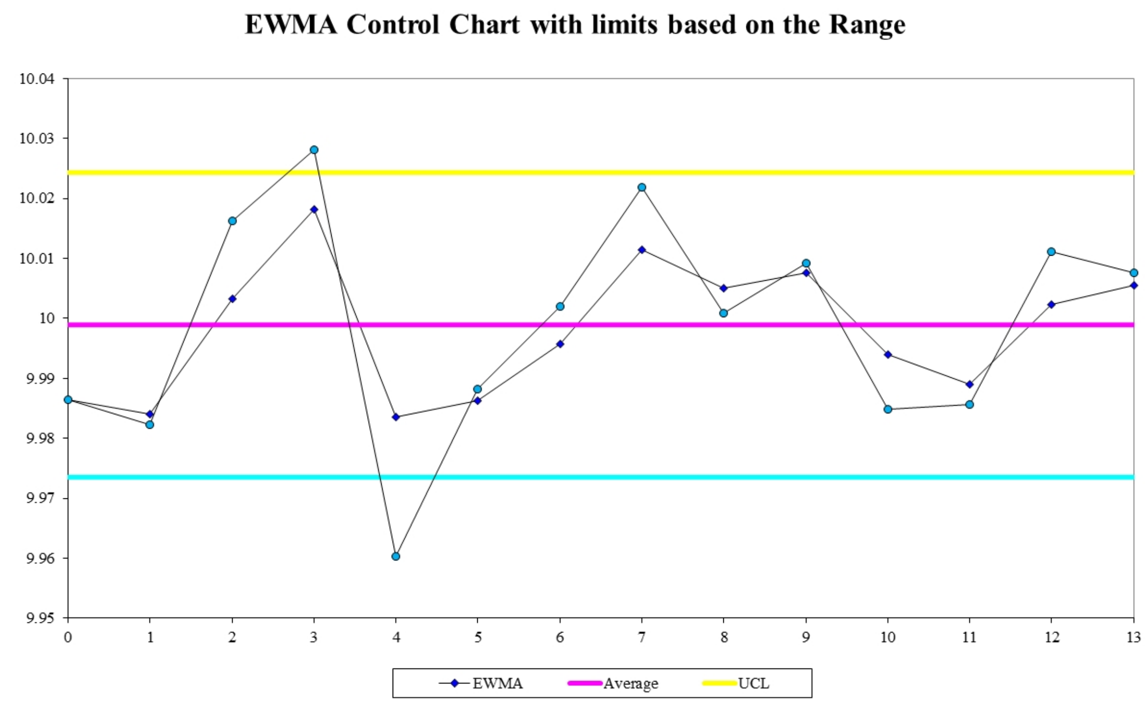
	I	J	K	L	M
1	**λ**	0.6			
2	**d2**	2.534			
3					
4	**N°**	**Measurements**	**Range**	**EWMA**	**CL**
5	0	=AVERAGE(B5:G5)	=MAX(B5:G5)-MIN(B5:G5)	=J5	=AVERAGE($J$5:$J$18)
6	1	=AVERAGE(B6:G6)	=MAX(B6:G6)-MIN(B6:G6)	=$J$1*J6+(1-$J$1)*L5	=AVERAGE($J$5:$J$18)
7	2	=AVERAGE(B7:G7)	=MAX(B7:G7)-MIN(B7:G7)	=$J$1*J7+(1-$J$1)*L6	=AVERAGE($J$5:$J$18)
8	3	=AVERAGE(B8:G8)	=MAX(B8:G8)-MIN(B8:G8)	=$J$1*J8+(1-$J$1)*L7	=AVERAGE($J$5:$J$18)
9	4	=AVERAGE(B9:G9)	=MAX(B9:G9)-MIN(B9:G9)	=$J$1*J9+(1-$J$1)*L8	=AVERAGE($J$5:$J$18)
10	5	=AVERAGE(B10:G10)	=MAX(B10:G10)-MIN(B10:G10)	=$J$1*J10+(1-$J$1)*L9	=AVERAGE($J$5:$J$18)
11	6	=AVERAGE(B11:G11)	=MAX(B11:G11)-MIN(B11:G11)	=$J$1*J11+(1-$J$1)*L10	=AVERAGE($J$5:$J$18)
12	7	=AVERAGE(B12:G12)	=MAX(B12:G12)-MIN(B12:G12)	=$J$1*J12+(1-$J$1)*L11	=AVERAGE($J$5:$J$18)
13	8	=AVERAGE(B13:G13)	=MAX(B13:G13)-MIN(B13:G13)	=$J$1*J13+(1-$J$1)*L12	=AVERAGE($J$5:$J$18)
14	9	=AVERAGE(B14:G14)	=MAX(B14:G14)-MIN(B14:G14)	=$J$1*J14+(1-$J$1)*L13	=AVERAGE($J$5:$J$18)
15	10	=AVERAGE(B15:G15)	=MAX(B15:G15)-MIN(B15:G15)	=$J$1*J15+(1-$J$1)*L14	=AVERAGE($J$5:$J$18)
16	11	=AVERAGE(B16:G16)	=MAX(B16:G16)-MIN(B16:G16)	=$J$1*J16+(1-$J$1)*L15	=AVERAGE($J$5:$J$18)
17	12	=AVERAGE(B17:G17)	=MAX(B17:G17)-MIN(B17:G17)	=$J$1*J17+(1-$J$1)*L16	=AVERAGE($J$5:$J$18)
18	13	=AVERAGE(B18:G18)	=MAX(B18:G18)-MIN(B18:G18)	=$J$1*J18+(1-$J$1)*L17	=AVERAGE($J$5:$J$18)

and for the second part:

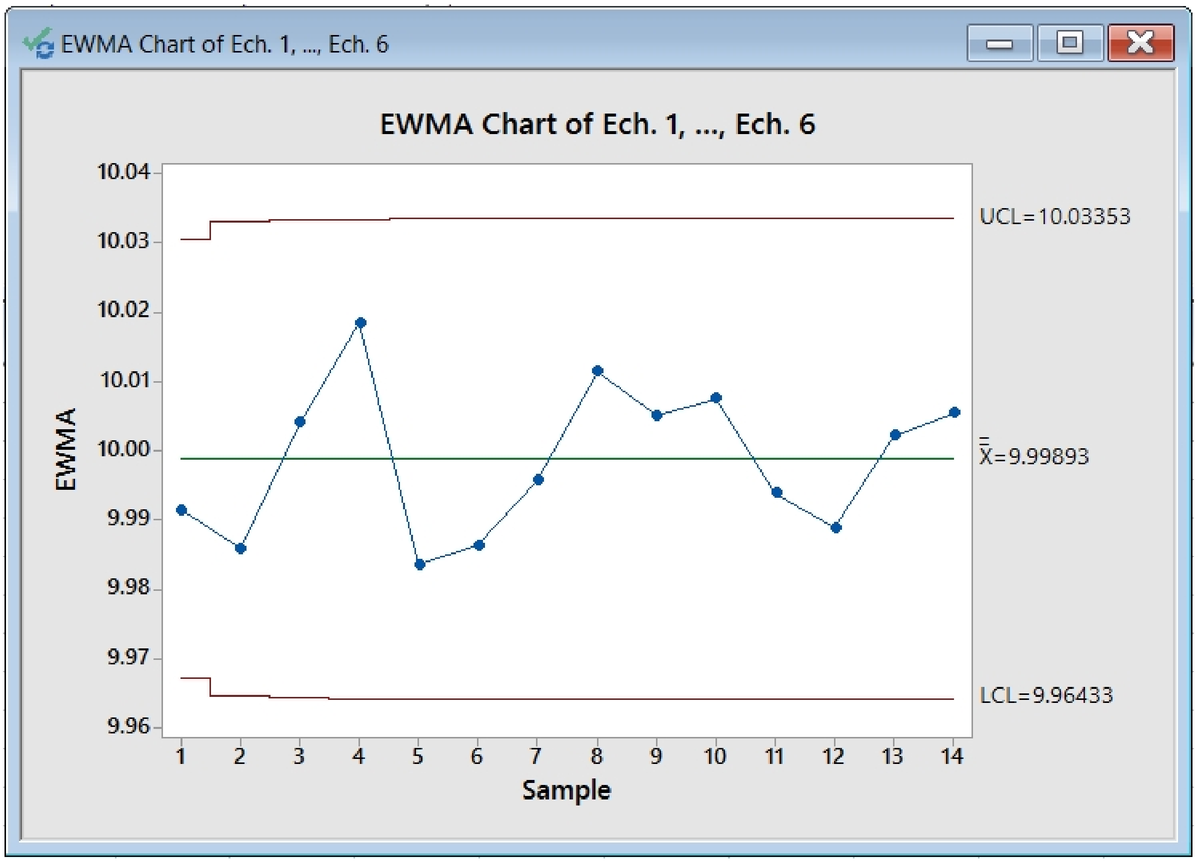


	N	O
1		
2		
3		
4	**LCL**	**UCL**
5	=M5-3*(AVERAGE($K$5:$K$18)/$J$2)*SQRT($J$1/(COUNT(B5:G5)*(2+$J$1)))	=M5+3*(AVERAGE($K$5:$K$18)/$J$2)*SQRT($J$1/(COUNT(B5:G5)*(2+$J$1)))
6	=M6-3*(AVERAGE($K$5:$K$18)/$J$2)*SQRT($J$1/(COUNT(B6:G6)*(2+$J$1)))	=M6+3*(AVERAGE($K$5:$K$18)/$J$2)*SQRT($J$1/(COUNT(B6:G6)*(2+$J$1)))
7	=M7-3*(AVERAGE($K$5:$K$18)/$J$2)*SQRT($J$1/(COUNT(B7:G7)*(2+$J$1)))	=M7+3*(AVERAGE($K$5:$K$18)/$J$2)*SQRT($J$1/(COUNT(B7:G7)*(2+$J$1)))
8	=M8-3*(AVERAGE($K$5:$K$18)/$J$2)*SQRT($J$1/(COUNT(B8:G8)*(2+$J$1)))	=M8+3*(AVERAGE($K$5:$K$18)/$J$2)*SQRT($J$1/(COUNT(B8:G8)*(2+$J$1)))
9	=M9-3*(AVERAGE($K$5:$K$18)/$J$2)*SQRT($J$1/(COUNT(B9:G9)*(2+$J$1)))	=M9+3*(AVERAGE($K$5:$K$18)/$J$2)*SQRT($J$1/(COUNT(B9:G9)*(2+$J$1)))
10	=M10-3*(AVERAGE($K$5:$K$18)/$J$2)*SQRT($J$1/(COUNT(B10:G10)*(2+$J$1)))	=M10+3*(AVERAGE($K$5:$K$18)/$J$2)*SQRT($J$1/(COUNT(B10:G10)*(2+$J$1)))
11	=M11-3*(AVERAGE($K$5:$K$18)/$J$2)*SQRT($J$1/(COUNT(B11:G11)*(2+$J$1)))	=M11+3*(AVERAGE($K$5:$K$18)/$J$2)*SQRT($J$1/(COUNT(B11:G11)*(2+$J$1)))
12	=M12-3*(AVERAGE($K$5:$K$18)/$J$2)*SQRT($J$1/(COUNT(B12:G12)*(2+$J$1)))	=M12+3*(AVERAGE($K$5:$K$18)/$J$2)*SQRT($J$1/(COUNT(B12:G12)*(2+$J$1)))
13	=M13-3*(AVERAGE($K$5:$K$18)/$J$2)*SQRT($J$1/(COUNT(B13:G13)*(2+$J$1)))	=M13+3*(AVERAGE($K$5:$K$18)/$J$2)*SQRT($J$1/(COUNT(B13:G13)*(2+$J$1)))
14	=M14-3*(AVERAGE($K$5:$K$18)/$J$2)*SQRT($J$1/(COUNT(B14:G14)*(2+$J$1)))	=M14+3*(AVERAGE($K$5:$K$18)/$J$2)*SQRT($J$1/(COUNT(B14:G14)*(2+$J$1)))
15	=M15-3*(AVERAGE($K$5:$K$18)/$J$2)*SQRT($J$1/(COUNT(B15:G15)*(2+$J$1)))	=M15+3*(AVERAGE($K$5:$K$18)/$J$2)*SQRT($J$1/(COUNT(B15:G15)*(2+$J$1)))
16	=M16-3*(AVERAGE($K$5:$K$18)/$J$2)*SQRT($J$1/(COUNT(B16:G16)*(2+$J$1)))	=M16+3*(AVERAGE($K$5:$K$18)/$J$2)*SQRT($J$1/(COUNT(B16:G16)*(2+$J$1)))
17	=M17-3*(AVERAGE($K$5:$K$18)/$J$2)*SQRT($J$1/(COUNT(B17:G17)*(2+$J$1)))	=M17+3*(AVERAGE($K$5:$K$18)/$J$2)*SQRT($J$1/(COUNT(B17:G17)*(2+$J$1)))
18	=M18-3*(AVERAGE($K$5:$K$18)/$J$2)*SQRT($J$1/(COUNT(B18:G18)*(2+$J$1)))	=M18+3*(AVERAGE($K$5:$K$18)/$J$2)*SQRT($J$1/(COUNT(B18:G18)*(2+$J$1)))

We then have the following control chart:



With Minitab 17.3.1 we get a small difference in comparison to our calculations by hand and this is normal as we have made an approximation for a large t and that we also made one of the two choices for F_0:

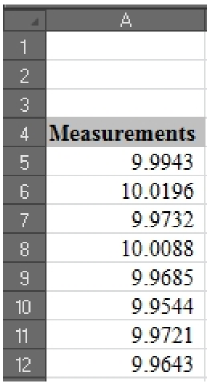

EWMA Chart of Ech. 1, ..., Ech. 6
EWMA Chart of Ech. 1, ..., Ech. 6
EWMA
10.04
10.03
10.02
10.01
10.00
9.99
9.98
9.97
9.96
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Sample
UCL=10.03353
$\bar{\bar{X}}$=9.99893
LCL=9.96433

For the last example, we take the famous case of an EWMA control chart with individual values

and:

$$\begin{aligned}\text{UCL}_{\text{EWMA}} &= \overline{\overline{X}} + 3\frac{\overline{\text{MR}}}{d_2(n_i)}\sqrt{\frac{\lambda}{n_i(2+\lambda)}}\\ \text{CL}_{\text{EWMA}} &= \overline{\overline{X}}\\ \text{LCL}_{\text{EWMA}} &= \overline{\overline{X}} - 3\frac{\overline{\text{MR}}}{d_2(n_i)}\sqrt{\frac{\lambda}{n_i(2+\lambda)}}\end{aligned} \tag{16.7.879}$$

with the following data:



	A
1	
2	
3	
4	**Measurements**
5	9.9943
6	10.0196
7	9.9732
8	10.0088
9	9.9685
10	9.9544
11	9.9721
12	9.9643

by taking therefore (caution! the n in the following relation is not the same as that in the root containing the smoothing constant!):

$$\overline{\text{MR}} = \overline{R} = \frac{1}{n-1}\sum_{i=2}^{n}\sum_{i=2}^{n}|x_i - x_{i-1}| \tag{16.7.880}$$

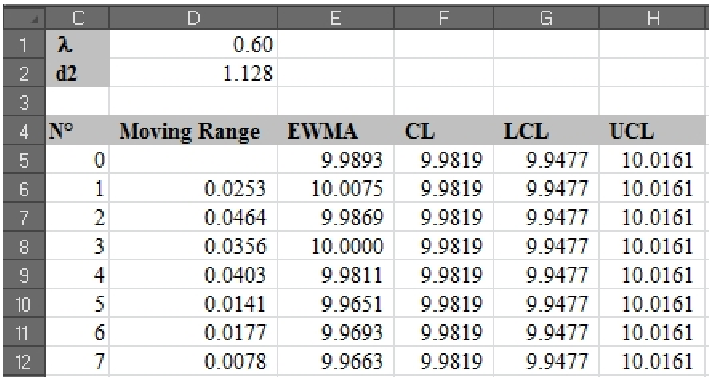
and:

$$d_2(2) = 1.128 \tag{16.7.881}$$

as well as:

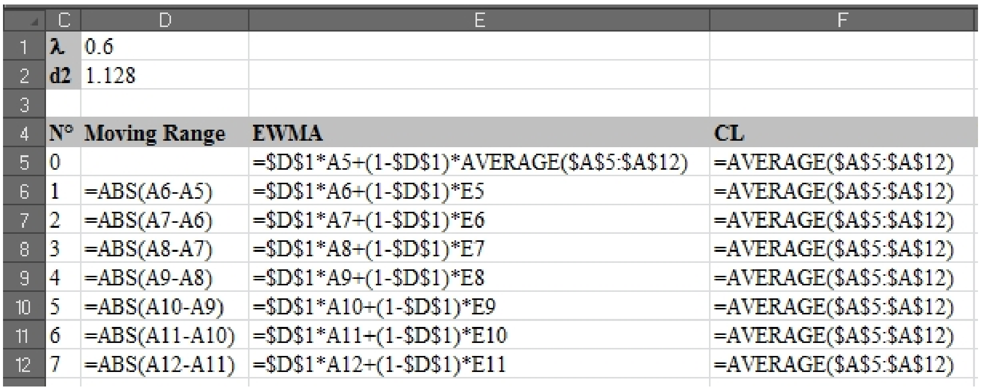
$$F_0 = \overline{X} \tag{16.7.882}$$

We have then with a smoothing constant of 0.6:



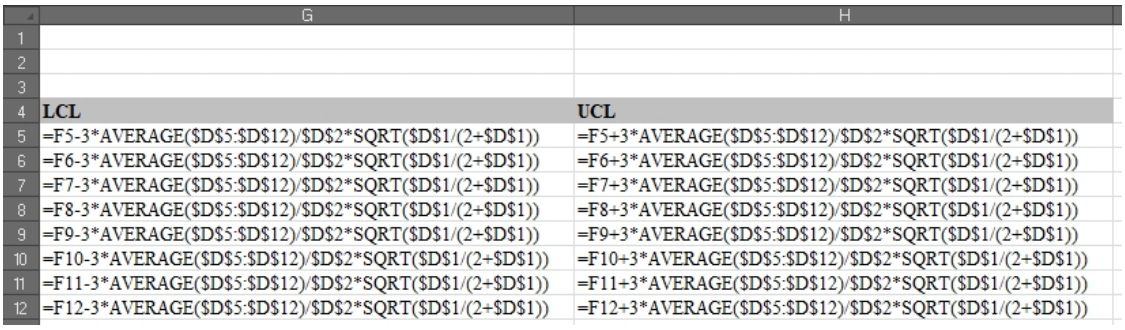
	C	D	E	F	G	H
1	**λ**	0.60				
2	**d2**	1.128				
3						
4	**N°**	**Moving Range**	**EWMA**	**CL**	**LCL**	**UCL**
5	0		9.9893	9.9819	9.9477	10.0161
6	1	0.0253	10.0075	9.9819	9.9477	10.0161
7	2	0.0464	9.9869	9.9819	9.9477	10.0161
8	3	0.0356	10.0000	9.9819	9.9477	10.0161
9	4	0.0403	9.9811	9.9819	9.9477	10.0161
10	5	0.0141	9.9651	9.9819	9.9477	10.0161
11	6	0.0177	9.9693	9.9819	9.9477	10.0161
12	7	0.0078	9.9663	9.9819	9.9477	10.0161

Thus explicitly (we have cut the screenshot below into two parts because of the limit horizontal space on the present page):



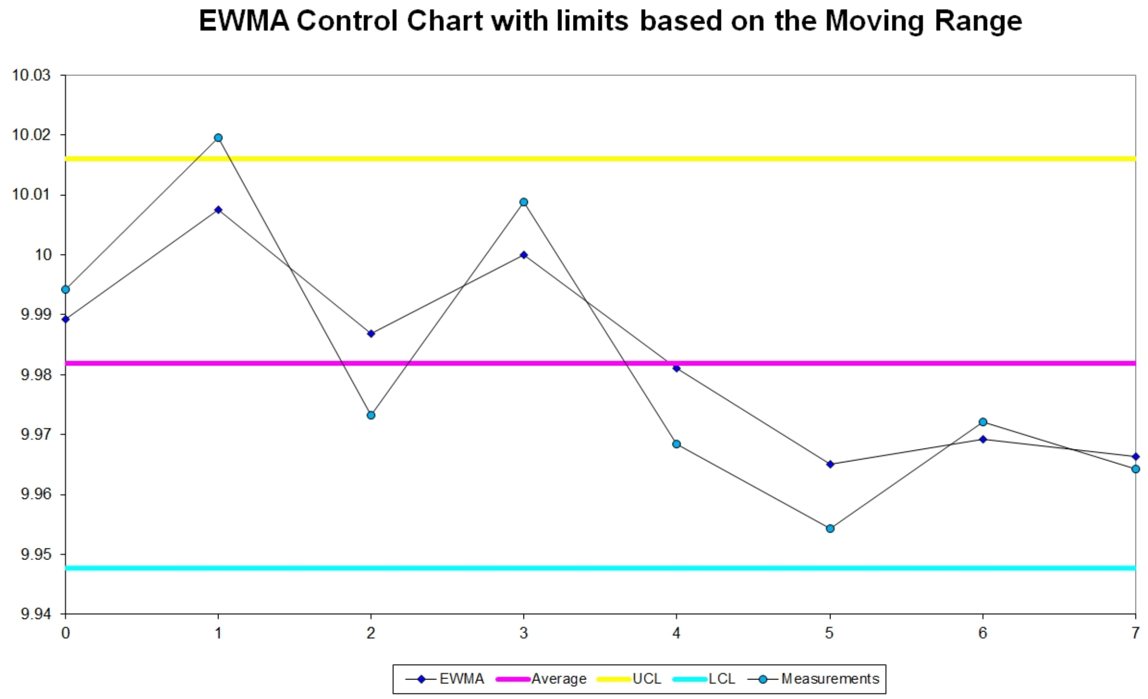
	C	D	E	F
1	**λ**	0.6		
2	**d2**	1.128		
3				
4	**N°**	**Moving Range**	**EWMA**	**CL**
5	0		=$D$1*A5+(1-$D$1)*AVERAGE($A$5:$A$12)	=AVERAGE($A$5:$A$12)
6	1	=ABS(A6-A5)	=$D$1*A6+(1-$D$1)*E5	=AVERAGE($A$5:$A$12)
7	2	=ABS(A7-A6)	=$D$1*A7+(1-$D$1)*E6	=AVERAGE($A$5:$A$12)
8	3	=ABS(A8-A7)	=$D$1*A8+(1-$D$1)*E7	=AVERAGE($A$5:$A$12)
9	4	=ABS(A9-A8)	=$D$1*A9+(1-$D$1)*E8	=AVERAGE($A$5:$A$12)
10	5	=ABS(A10-A9)	=$D$1*A10+(1-$D$1)*E9	=AVERAGE($A$5:$A$12)
11	6	=ABS(A11-A10)	=$D$1*A11+(1-$D$1)*E10	=AVERAGE($A$5:$A$12)
12	7	=ABS(A12-A11)	=$D$1*A12+(1-$D$1)*E11	=AVERAGE($A$5:$A$12)

and for the second part:

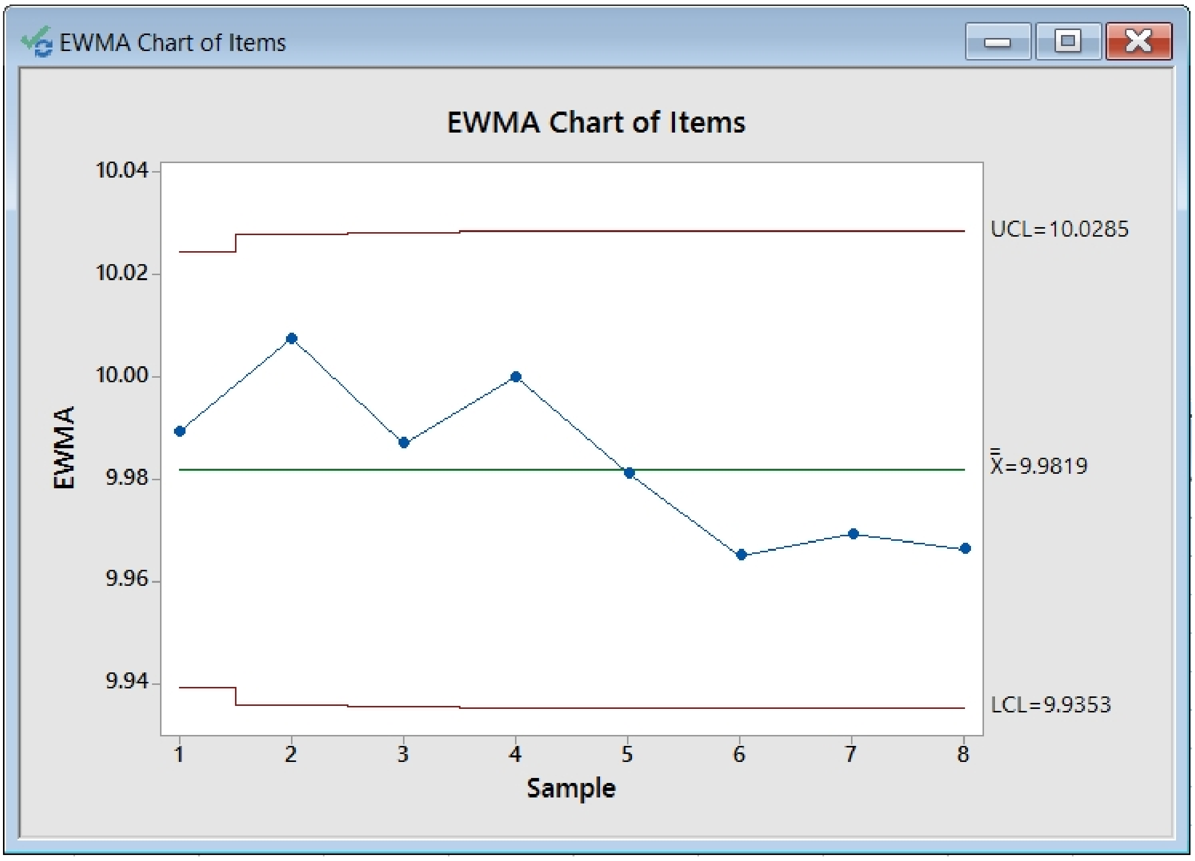


	G	H
1		
2		
3		
4	**LCL**	**UCL**
5	=F5-3*AVERAGE($D$5:$D$12)/$D$2*SQRT($D$1/(2+$D$1))	=F5+3*AVERAGE($D$5:$D$12)/$D$2*SQRT($D$1/(2+$D$1))
6	=F6-3*AVERAGE($D$5:$D$12)/$D$2*SQRT($D$1/(2+$D$1))	=F6+3*AVERAGE($D$5:$D$12)/$D$2*SQRT($D$1/(2+$D$1))
7	=F7-3*AVERAGE($D$5:$D$12)/$D$2*SQRT($D$1/(2+$D$1))	=F7+3*AVERAGE($D$5:$D$12)/$D$2*SQRT($D$1/(2+$D$1))
8	=F8-3*AVERAGE($D$5:$D$12)/$D$2*SQRT($D$1/(2+$D$1))	=F8+3*AVERAGE($D$5:$D$12)/$D$2*SQRT($D$1/(2+$D$1))
9	=F9-3*AVERAGE($D$5:$D$12)/$D$2*SQRT($D$1/(2+$D$1))	=F9+3*AVERAGE($D$5:$D$12)/$D$2*SQRT($D$1/(2+$D$1))
10	=F10-3*AVERAGE($D$5:$D$12)/$D$2*SQRT($D$1/(2+$D$1))	=F10+3*AVERAGE($D$5:$D$12)/$D$2*SQRT($D$1/(2+$D$1))
11	=F11-3*AVERAGE($D$5:$D$12)/$D$2*SQRT($D$1/(2+$D$1))	=F11+3*AVERAGE($D$5:$D$12)/$D$2*SQRT($D$1/(2+$D$1))
12	=F12-3*AVERAGE($D$5:$D$12)/$D$2*SQRT($D$1/(2+$D$1))	=F12+3*AVERAGE($D$5:$D$12)/$D$2*SQRT($D$1/(2+$D$1))

We then have the following control chart:



With Minitab 17.3.1 we get a small difference in comparison to our calculations by hand and this is normal as we have made an approximation for a large t and also that we made one of the two choices for F_0:



As with other control charts, the EWMA control chart must be accompanied by a control chart for monitoring the variability, so either a control chart based on ranges with fixed limits or a control chart with standard deviations with fixed limits.

16.7.6.6 Rare events control charts

Rare events inherently occur in all kinds of processes. In hospitals, there are medication errors, infections, patient falls, ventilator-associated pneumonias, and other rare, adverse events that cause prolonged hospital stays and increase healthcare costs.

But rare events happen in many other contexts, too. Software developers may need to track errors in lines of programming code, or a quality practitioner may need to monitor a low-defect process in a high-yield manufacturing environment. Accidents that occur on the shop floor and aircraft engine failures are also rare events, ideally.

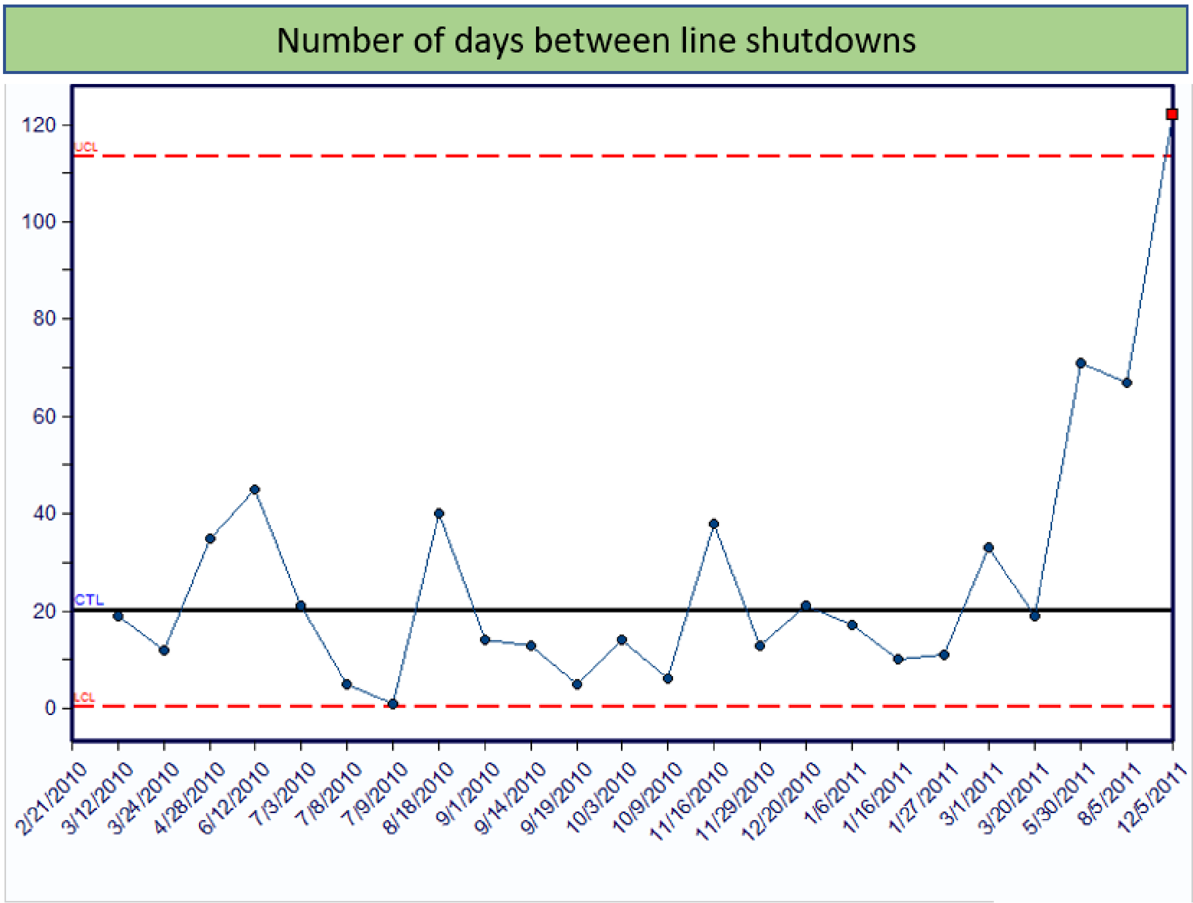
Whether you're in healthcare, software development, manufacturing or some other industry, statistical process control is an important component of quality improvement. Using control charts, we can graph these rare events and monitor a process to determine if it's stable or if it's out of control and therefore unpredictable and in need of attention.

16.7.6.6.1 Frequency T control chart with probabilistic limits

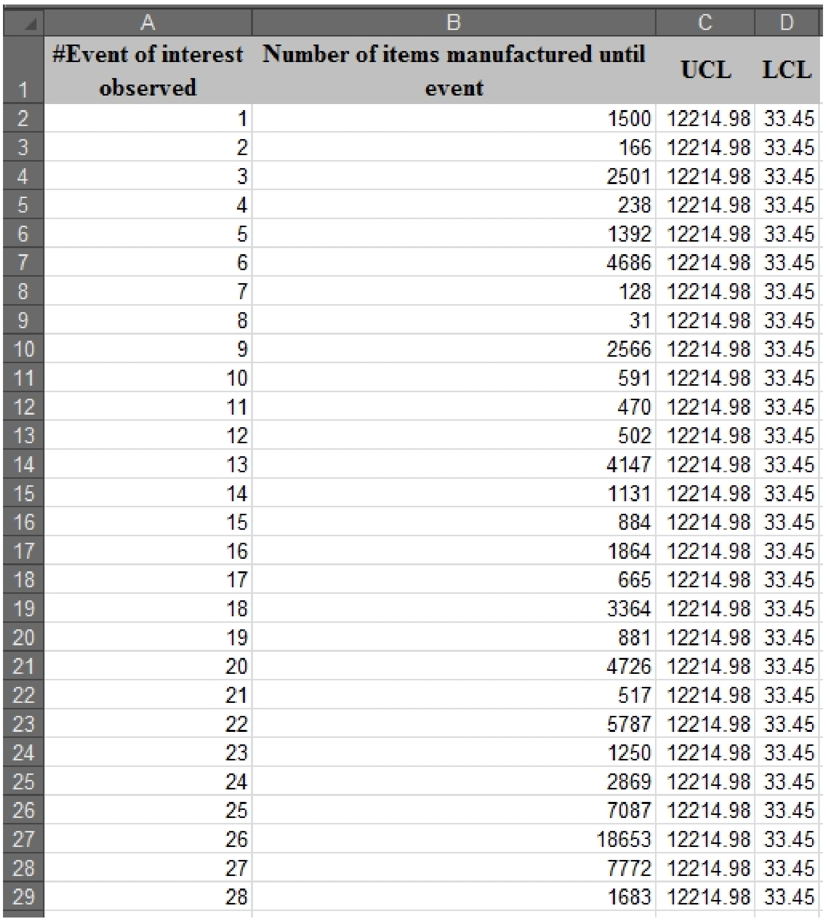
The monitoring of frequency of occurrence of defects / accidents / anomalies is a very simple technique to implement and very easy to read for people with little or no knowledge of statistics. Moreover it is a technique adapted to cases where the monitored events are very rare.

The idea then is not to track the number of special events, but the time T between two appearances of the monitored events. The time T is then measured in either number of pieces or number of accidents, defects, days, etc.

This control chart is also sometimes used by Human Resources departments or Top Managers to track the frequency of problems caused by some employees in the purpose to highlight objectively and scientifically if their behavior is compliant with the expectations of their organization (sadly most C.E.O. and low level managers still don't know such a tool exist and base their judgement on their feelings...).

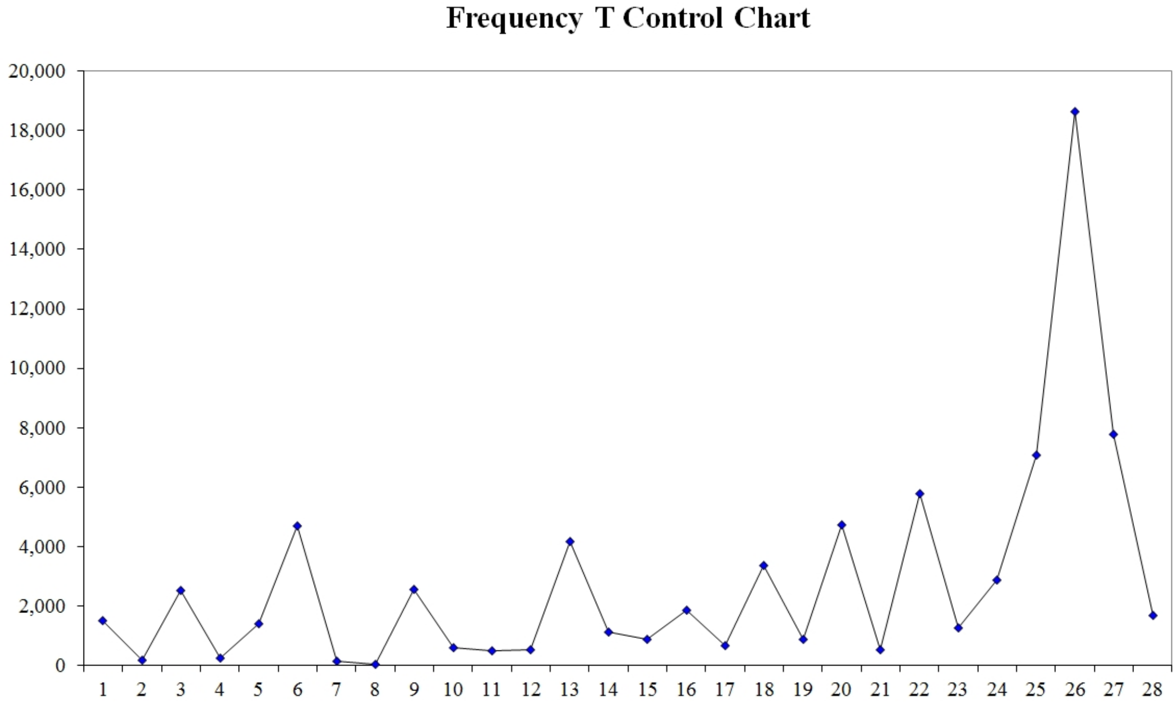


For companion example let us consider the following list of data:



	A	B	C	D
1	**#Event of interest observed**	**Number of items manufactured until event**	**UCL**	**LCL**
2	1	1500	12214.98	33.45
3	2	166	12214.98	33.45
4	3	2501	12214.98	33.45
5	4	238	12214.98	33.45
6	5	1392	12214.98	33.45
7	6	4686	12214.98	33.45
8	7	128	12214.98	33.45
9	8	31	12214.98	33.45
10	9	2566	12214.98	33.45
11	10	591	12214.98	33.45
12	11	470	12214.98	33.45
13	12	502	12214.98	33.45
14	13	4147	12214.98	33.45
15	14	1131	12214.98	33.45
16	15	884	12214.98	33.45
17	16	1864	12214.98	33.45
18	17	665	12214.98	33.45
19	18	3364	12214.98	33.45
20	19	881	12214.98	33.45
21	20	4726	12214.98	33.45
22	21	517	12214.98	33.45
23	22	5787	12214.98	33.45
24	23	1250	12214.98	33.45
25	24	2869	12214.98	33.45
26	25	7087	12214.98	33.45
27	26	18653	12214.98	33.45
28	27	7772	12214.98	33.45
29	28	1683	12214.98	33.45

Which gives graphically:



Well it's nice, but we need an UCL and LCL for this control chart to be really useful! To achieve this we will use the queuing theory with the same assumptions and the same checkpoints (see section Quantitative Management page 8141). Let us first recall that we had showed there that the probability of observing k events in a time interval t (or an interval of pieces number t) under some very specific assumptions (!!!), followed a Poisson distribution:

$$p_k \cong \frac{(\lambda t)^k}{k!} e^{-\lambda t} \qquad (16.7.883)$$

where the parameter λ is the average arrival rate (or "average rate of occurrence") of events per unit time (or "Poisson Arrivals See Time Average (PASTA)") assumed constant over the entire period:

$$\lambda = \frac{1}{\overline{\overline{T}}} \qquad (16.7.884)$$

and where we have for mean and variance (see section Statistics page 686) the number of events:

$$\text{E}(k) = \lambda t \qquad \text{V}(k) = \lambda t \qquad (16.7.885)$$

This is a probability law with discreet support as we know, which fits well with our case studies so the Poisson distribution can be a good candidate (but there are others, like for example the Geometric law!!!).

But we also proved that the random variable τ representing the time (or the number of pieces) separating two arrivals of unwanted events (or desired...) was given by the distribution function based on the exponential law:

$$A(\tau \leq t) = 1 - e^{-\lambda t} \qquad (16.7.886)$$

Still with:

$$\lambda = \frac{1}{\overline{\overline{T}}} \qquad (16.7.887)$$

where:

$$\mu = \sigma = \frac{1}{\lambda} = \overline{T} \qquad (16.7.888)$$

So we could take the following Shewhart limits :

$$\boxed{\begin{aligned} \text{USL} &= \text{required by customer/board comitee} \\ \text{UCL}_T &= \overline{T} + 3\overline{T} \\ \text{CL}_T &= \text{E}(k) = \overline{T} \\ \text{LCL}_T &= \overline{T} - 3\overline{T} := 0 \end{aligned}} \qquad (16.7.889)$$

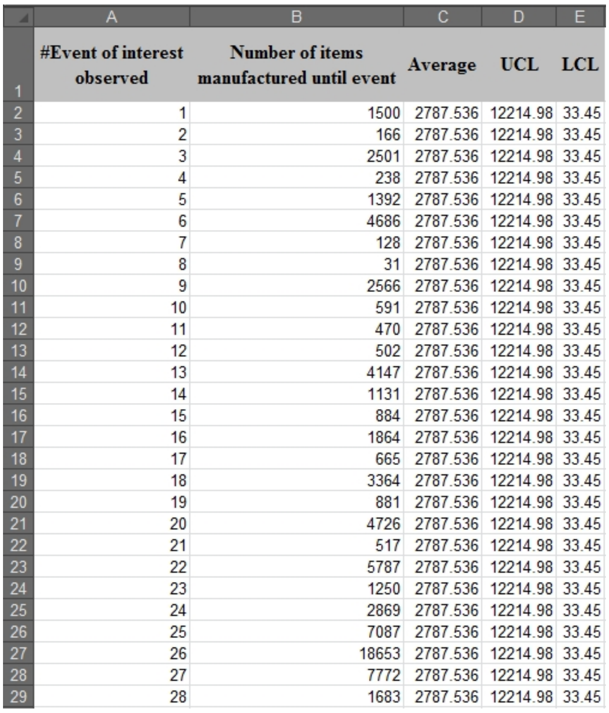
But it is more rigorous (as for all other control charts) to use probabilistic limits. So if we want to set limits with a usual bilateral risk threshold $\alpha/2$ of 0.5% then we calculate for the lower limit:

$$\frac{\alpha}{2} = 1 - e^{-\lambda x} \Rightarrow x = \frac{\ln\left(1 - \dfrac{\alpha}{2}\right)}{-\lambda} = -\ln\left(1 - \frac{0.25\%}{2}\right)\overline{T} \cong 0.012\overline{T} \qquad (16.7.890)$$

and for the superior limit:

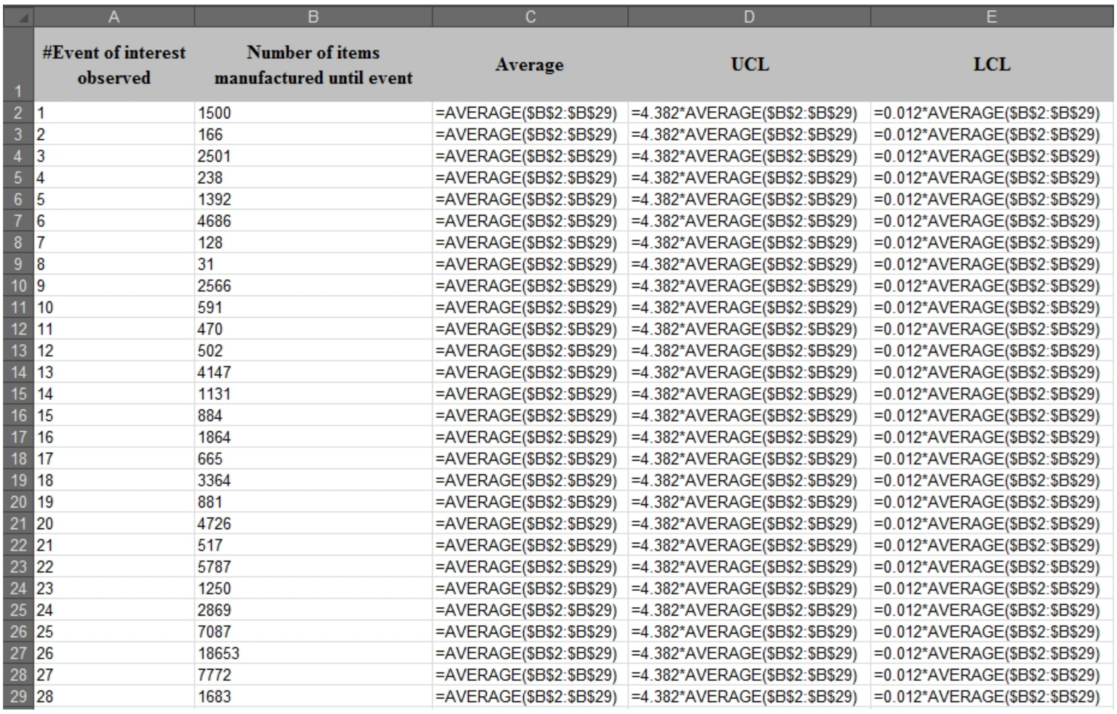
$$1-\frac{\alpha}{2}=1-e^{-\lambda x}\Rightarrow x=\frac{\ln\left(\frac{\alpha}{2}\right)}{-\lambda}=-\ln\left(\frac{0.25\%}{2}\right)\overline{T}\cong 4.382\overline{T} \tag{16.7.891}$$

Then we have:



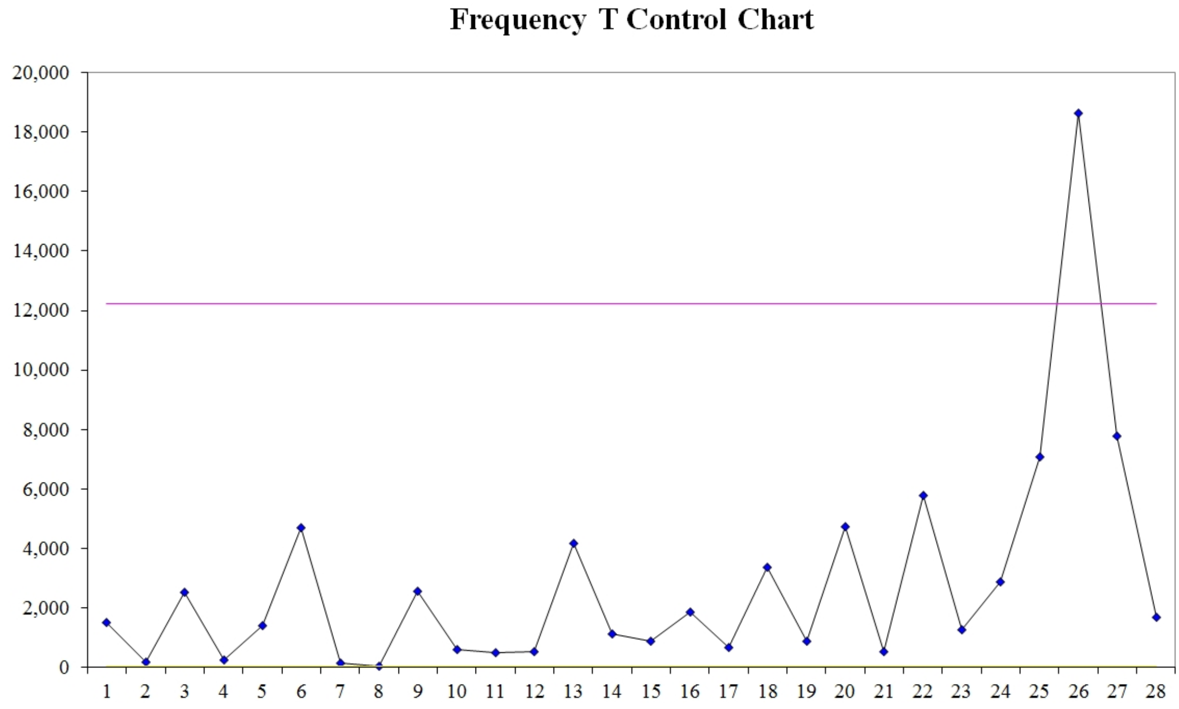
	A	B	C	D	E
1	#Event of interest observed	Number of items manufactured until event	Average	UCL	LCL
2	1	1500	2787.536	12214.98	33.45
3	2	166	2787.536	12214.98	33.45
4	3	2501	2787.536	12214.98	33.45
5	4	238	2787.536	12214.98	33.45
6	5	1392	2787.536	12214.98	33.45
7	6	4686	2787.536	12214.98	33.45
8	7	128	2787.536	12214.98	33.45
9	8	31	2787.536	12214.98	33.45
10	9	2566	2787.536	12214.98	33.45
11	10	591	2787.536	12214.98	33.45
12	11	470	2787.536	12214.98	33.45
13	12	502	2787.536	12214.98	33.45
14	13	4147	2787.536	12214.98	33.45
15	14	1131	2787.536	12214.98	33.45
16	15	884	2787.536	12214.98	33.45
17	16	1864	2787.536	12214.98	33.45
18	17	665	2787.536	12214.98	33.45
19	18	3364	2787.536	12214.98	33.45
20	19	881	2787.536	12214.98	33.45
21	20	4726	2787.536	12214.98	33.45
22	21	517	2787.536	12214.98	33.45
23	22	5787	2787.536	12214.98	33.45
24	23	1250	2787.536	12214.98	33.45
25	24	2869	2787.536	12214.98	33.45
26	25	7087	2787.536	12214.98	33.45
27	26	18653	2787.536	12214.98	33.45
28	27	7772	2787.536	12214.98	33.45
29	28	1683	2787.536	12214.98	33.45

Thus explicitly:

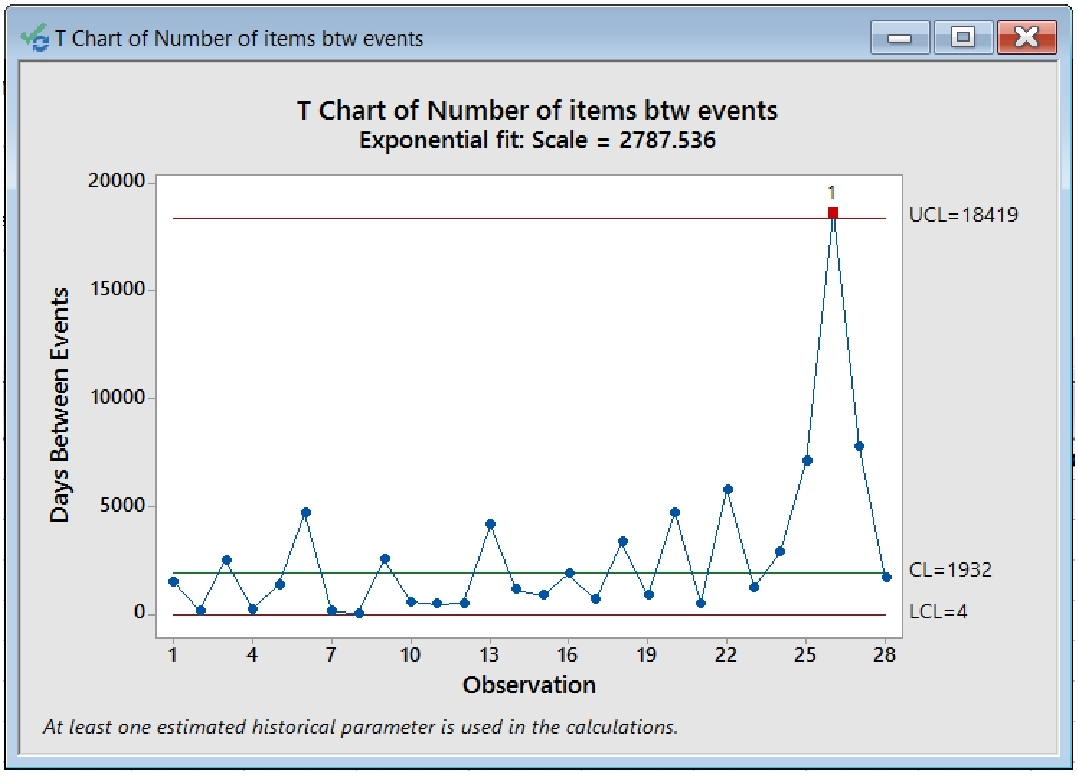


	A	B	C	D	E
1	#Event of interest observed	Number of items manufactured until event	Average	UCL	LCL
2	1	1500	=AVERAGE($B$2:$B$29)	=4.382*AVERAGE($B$2:$B$29)	=0.012*AVERAGE($B$2:$B$29)
3	2	166	=AVERAGE($B$2:$B$29)	=4.382*AVERAGE($B$2:$B$29)	=0.012*AVERAGE($B$2:$B$29)
4	3	2501	=AVERAGE($B$2:$B$29)	=4.382*AVERAGE($B$2:$B$29)	=0.012*AVERAGE($B$2:$B$29)
5	4	238	=AVERAGE($B$2:$B$29)	=4.382*AVERAGE($B$2:$B$29)	=0.012*AVERAGE($B$2:$B$29)
6	5	1392	=AVERAGE($B$2:$B$29)	=4.382*AVERAGE($B$2:$B$29)	=0.012*AVERAGE($B$2:$B$29)
7	6	4686	=AVERAGE($B$2:$B$29)	=4.382*AVERAGE($B$2:$B$29)	=0.012*AVERAGE($B$2:$B$29)
8	7	128	=AVERAGE($B$2:$B$29)	=4.382*AVERAGE($B$2:$B$29)	=0.012*AVERAGE($B$2:$B$29)
9	8	31	=AVERAGE($B$2:$B$29)	=4.382*AVERAGE($B$2:$B$29)	=0.012*AVERAGE($B$2:$B$29)
10	9	2566	=AVERAGE($B$2:$B$29)	=4.382*AVERAGE($B$2:$B$29)	=0.012*AVERAGE($B$2:$B$29)
11	10	591	=AVERAGE($B$2:$B$29)	=4.382*AVERAGE($B$2:$B$29)	=0.012*AVERAGE($B$2:$B$29)
12	11	470	=AVERAGE($B$2:$B$29)	=4.382*AVERAGE($B$2:$B$29)	=0.012*AVERAGE($B$2:$B$29)
13	12	502	=AVERAGE($B$2:$B$29)	=4.382*AVERAGE($B$2:$B$29)	=0.012*AVERAGE($B$2:$B$29)
14	13	4147	=AVERAGE($B$2:$B$29)	=4.382*AVERAGE($B$2:$B$29)	=0.012*AVERAGE($B$2:$B$29)
15	14	1131	=AVERAGE($B$2:$B$29)	=4.382*AVERAGE($B$2:$B$29)	=0.012*AVERAGE($B$2:$B$29)
16	15	884	=AVERAGE($B$2:$B$29)	=4.382*AVERAGE($B$2:$B$29)	=0.012*AVERAGE($B$2:$B$29)
17	16	1864	=AVERAGE($B$2:$B$29)	=4.382*AVERAGE($B$2:$B$29)	=0.012*AVERAGE($B$2:$B$29)
18	17	665	=AVERAGE($B$2:$B$29)	=4.382*AVERAGE($B$2:$B$29)	=0.012*AVERAGE($B$2:$B$29)
19	18	3364	=AVERAGE($B$2:$B$29)	=4.382*AVERAGE($B$2:$B$29)	=0.012*AVERAGE($B$2:$B$29)
20	19	881	=AVERAGE($B$2:$B$29)	=4.382*AVERAGE($B$2:$B$29)	=0.012*AVERAGE($B$2:$B$29)
21	20	4726	=AVERAGE($B$2:$B$29)	=4.382*AVERAGE($B$2:$B$29)	=0.012*AVERAGE($B$2:$B$29)
22	21	517	=AVERAGE($B$2:$B$29)	=4.382*AVERAGE($B$2:$B$29)	=0.012*AVERAGE($B$2:$B$29)
23	22	5787	=AVERAGE($B$2:$B$29)	=4.382*AVERAGE($B$2:$B$29)	=0.012*AVERAGE($B$2:$B$29)
24	23	1250	=AVERAGE($B$2:$B$29)	=4.382*AVERAGE($B$2:$B$29)	=0.012*AVERAGE($B$2:$B$29)
25	24	2869	=AVERAGE($B$2:$B$29)	=4.382*AVERAGE($B$2:$B$29)	=0.012*AVERAGE($B$2:$B$29)
26	25	7087	=AVERAGE($B$2:$B$29)	=4.382*AVERAGE($B$2:$B$29)	=0.012*AVERAGE($B$2:$B$29)
27	26	18653	=AVERAGE($B$2:$B$29)	=4.382*AVERAGE($B$2:$B$29)	=0.012*AVERAGE($B$2:$B$29)
28	27	7772	=AVERAGE($B$2:$B$29)	=4.382*AVERAGE($B$2:$B$29)	=0.012*AVERAGE($B$2:$B$29)
29	28	1683	=AVERAGE($B$2:$B$29)	=4.382*AVERAGE($B$2:$B$29)	=0.012*AVERAGE($B$2:$B$29)

We then have the following control chart (we have exceptionally not represented the USL on it):



With Minitab 17.3.1 we get something quite different for all control limits. As the Minitab team doesn't give the proofs but only the formulas we are not able to say who has the most correct way between us and them for these calculations:

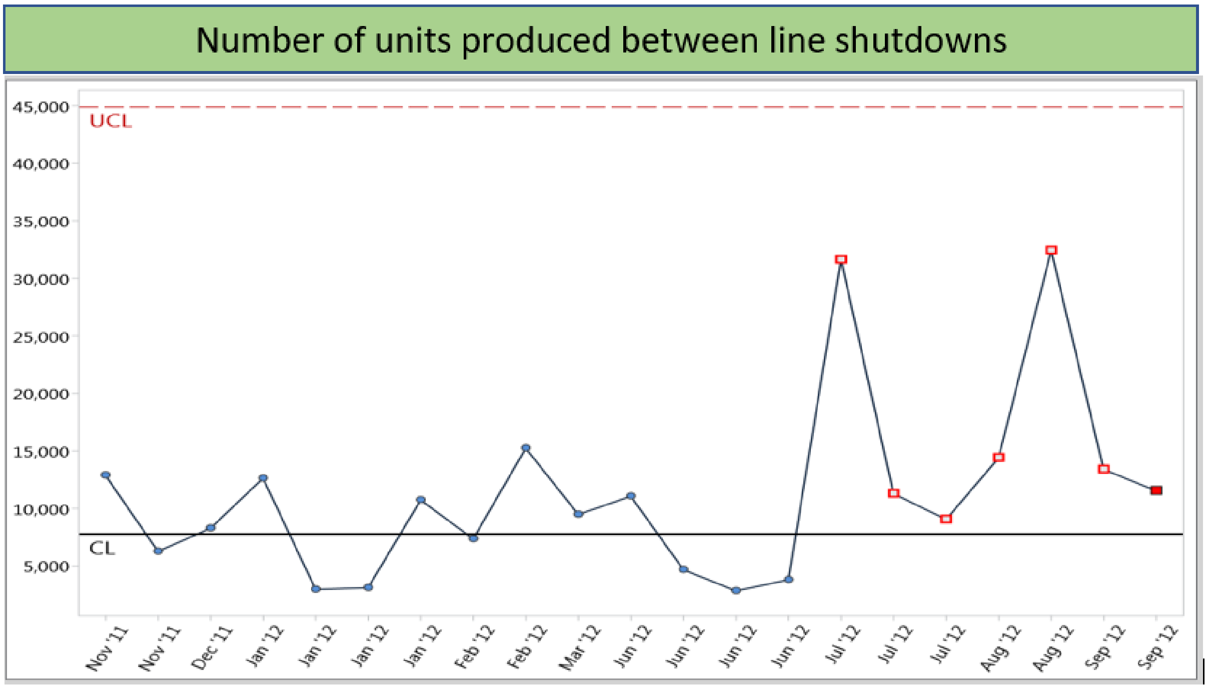

T Chart of Number of items btw events
T Chart of Number of items btw events
Exponential fit: Scale = 2787.536
Days Between Events
20000
15000
10000
5000
0
1
UCL=18419
CL=1932
LCL=4
1
4
7
10
13
16
19
22
25
28
Observation

At least one estimated historical parameter is used in the calculations.

16.7.6.6.2 Frequency G control chart of rare events

Developed by James Benneyan in 11991 (holocene calendar), the G chart[38] is a control chart that is based on the geometric distribution. Benneyan has since published several papers about the G chart and a companion chart, the "h chart". The majority of applications cited in these papers are for monitoring infection rates in healthcare, such as nosocomial infections. Nosocomial infections are infections that occur as a direct result of a patient's treatment in a medical facility.

P charts and U charts are often used to monitor adverse events such as nosocomial infections. But P charts and U charts require very large quantities of data and specific definitions of the data. For example, if you use a U chart to monitor nosocomial infections, each patient day is considered an area of opportunity in which one or more infections could occur. Thus, the data are the number of infections per patient day. If you use a P chart, the data are the number of patient days in which one or more infections occur. As for the data requirements, if you follow the standard practice of requiring a minimum of 25 to 35 subgroups to establish control limits, and the infection rate is low (for example, $< 1\%$), the required amount of data is at least $12,500$ patients (500 patients per subgroup multiplied by 25 subgroups). This means that it could take weeks, months, or perhaps even years to accumulate enough data to detect and respond to changes in the infection rate.



The geometric distribution provides an alternative probability model. In the geometric distribution, you count the number of opportunities before or until the defect occurs. Thus, in a healthcare setting where you monitor the infection rate, the ideal would be to count the number of patients or procedures until an infection is observed. While this is the ideal, it is also rarely done, because of complications with counting the actual number of patients through the system, or the number of procedures. What is most often done is to count the number of days between observed infections. The key assumption used when counting the number of days is that the number of patients or procedures per day is fairly constant.

---

[38] The "G" in G Chart stands for "geometric" because the number of normal events between nonconforming or rare incidents is represented by a geometric distribution.

We have proved in the section Statistics (see page 667) that the mean to have R successes before the E-th failure knowing that the probability of a failure is p was given by the mean of the negative binomial distribution:

$$\text{E}(R)=\frac{E}{p}\tag{16.7.892}$$

and we get for to the variance:

$$\text{V}(R)=E\frac{q}{p^2}\tag{16.7.893}$$

Obviously, if we set E as the first failure, the mean is reduced to that of the geometric law:

$$\text{E}(R)=\frac{1}{p}\qquad\text{and}\qquad\text{E}(V)=\frac{q}{p^2}=\frac{1-p}{p^2}\tag{16.7.894}$$

We have proved in the section Statistics (see page 654) that the estimator of the parameter of the geometric law was given by:

$$\hat{p}=\frac{k}{N}\tag{16.7.895}$$

Therefore it comes:

$$\text{E}(R)=\frac{N}{k}=\overline{g}\tag{16.7.896}$$

and:

$$\text{V}(R)=\frac{1-\left(\dfrac{k}{N}\right)}{\left(\dfrac{k}{N}\right)^2}=\left(\frac{N}{k}\right)^2\left(1-\left(\frac{k}{N}\right)\right)=\left(\frac{N}{k}\right)^2-\frac{N}{k}=\frac{N}{k}\left(\frac{N}{k}-1\right)=\overline{g}\left(\overline{g}-1\right)\tag{16.7.897}$$

and therefore:

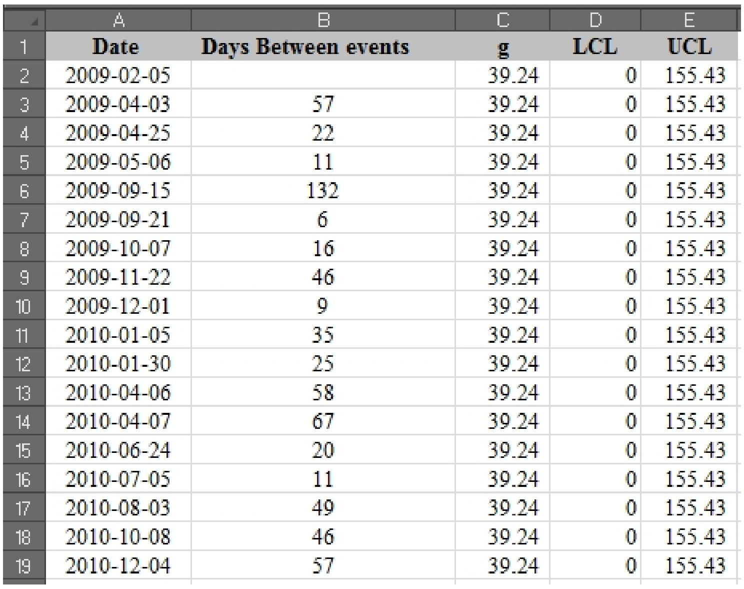
$$\hat{\sigma}_R=\sqrt{\overline{g}(\overline{g}-1)}\tag{16.7.898}$$

Therefore we have:

$$\boxed{\begin{aligned}\text{USL}&=\text{required by customer/board comitee}\\ \text{UCL}_G&=\overline{g}+3\sqrt{\overline{g}\left(\overline{g}-1\right)}\\ \text{CL}_G&=\overline{g}\\ \text{LCL}_G&=\overline{g}-3\sqrt{\overline{g}\left(\overline{g}-1\right)}\end{aligned}}\tag{16.7.899}$$

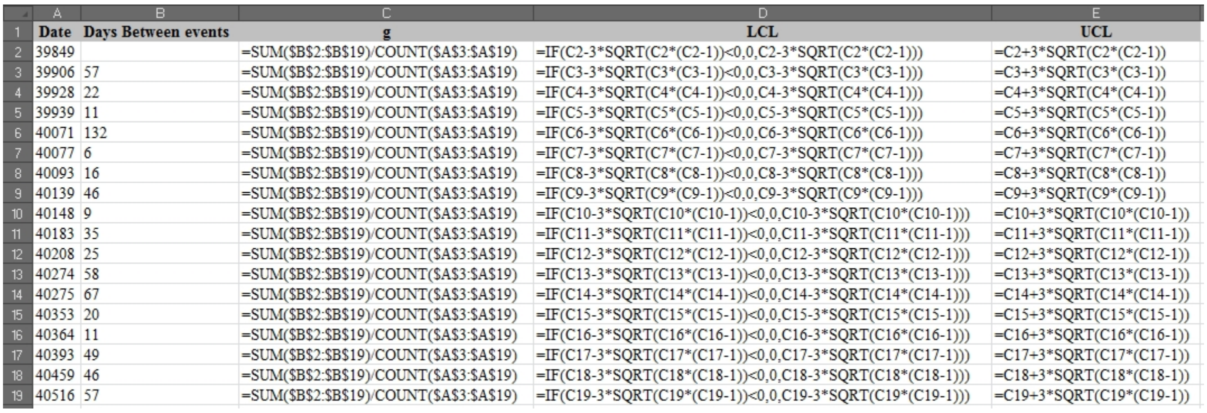
where if LCL is negative, we put it as always as being equal to zero.

Let us see an example by considering the following list in Microsoft® Excel® 14.0.6123:



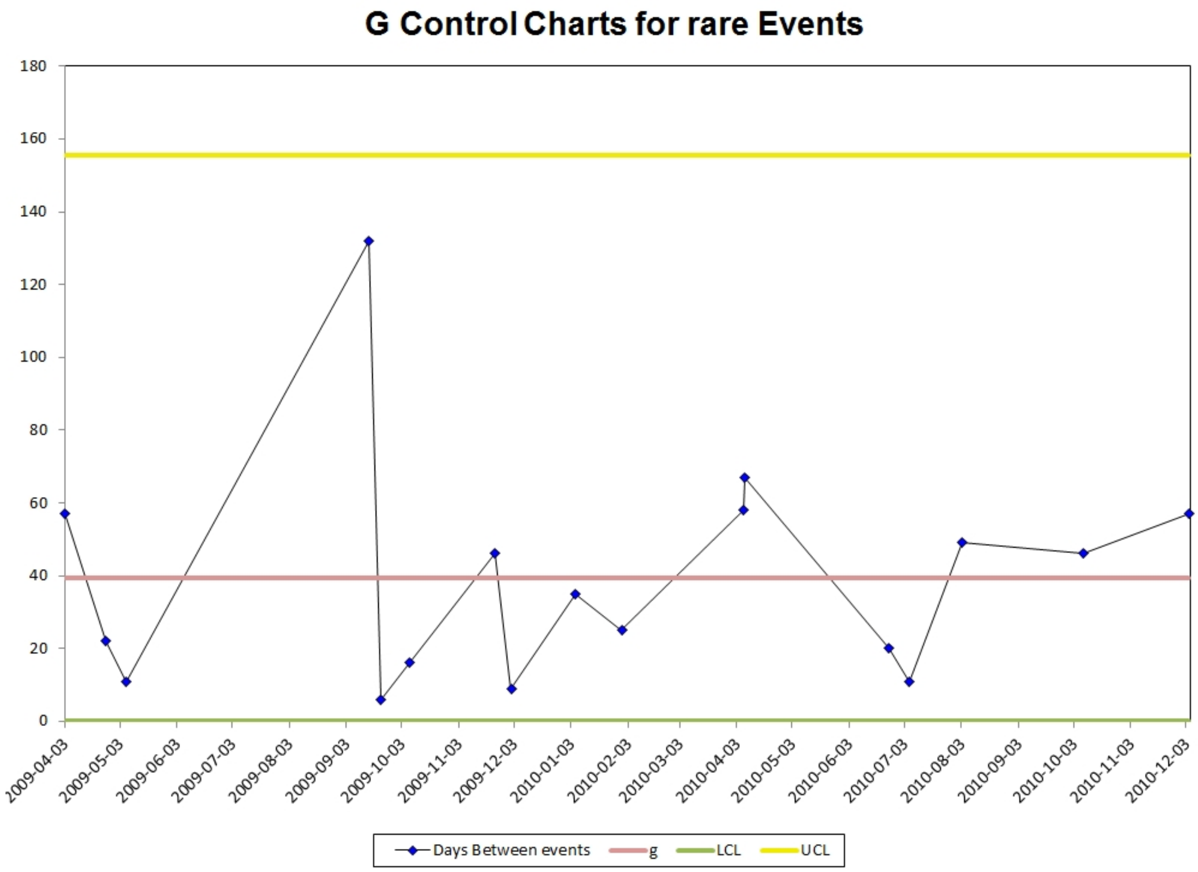
	A	B	C	D	E
1	**Date**	**Days Between events**	**g**	**LCL**	**UCL**
2	2009-02-05		39.24	0	155.43
3	2009-04-03	57	39.24	0	155.43
4	2009-04-25	22	39.24	0	155.43
5	2009-05-06	11	39.24	0	155.43
6	2009-09-15	132	39.24	0	155.43
7	2009-09-21	6	39.24	0	155.43
8	2009-10-07	16	39.24	0	155.43
9	2009-11-22	46	39.24	0	155.43
10	2009-12-01	9	39.24	0	155.43
11	2010-01-05	35	39.24	0	155.43
12	2010-01-30	25	39.24	0	155.43
13	2010-04-06	58	39.24	0	155.43
14	2010-04-07	67	39.24	0	155.43
15	2010-06-24	20	39.24	0	155.43
16	2010-07-05	11	39.24	0	155.43
17	2010-08-03	49	39.24	0	155.43
18	2010-10-08	46	39.24	0	155.43
19	2010-12-04	57	39.24	0	155.43

with the following formulas:

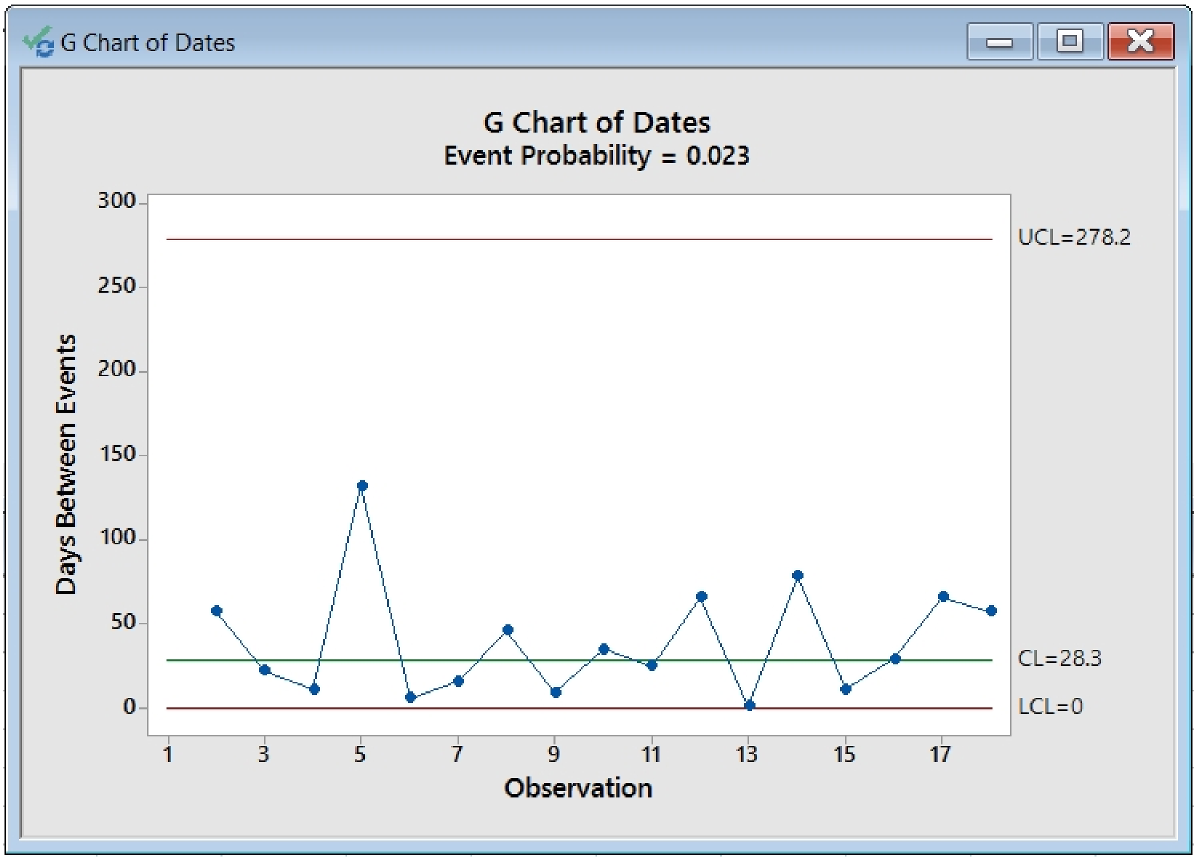


	A	B	C	D	E
1	**Date**	**Days Between events**	**g**	**LCL**	**UCL**
2	39849		=SUM($B$2:$B$19)/COUNT($A$3:$A$19)	=IF(C2-3*SQRT(C2*(C2-1))<0,0,C2-3*SQRT(C2*(C2-1)))	=C2+3*SQRT(C2*(C2-1))
3	39906	57	=SUM($B$2:$B$19)/COUNT($A$3:$A$19)	=IF(C3-3*SQRT(C3*(C3-1))<0,0,C3-3*SQRT(C3*(C3-1)))	=C3+3*SQRT(C3*(C3-1))
4	39928	22	=SUM($B$2:$B$19)/COUNT($A$3:$A$19)	=IF(C4-3*SQRT(C4*(C4-1))<0,0,C4-3*SQRT(C4*(C4-1)))	=C4+3*SQRT(C4*(C4-1))
5	39939	11	=SUM($B$2:$B$19)/COUNT($A$3:$A$19)	=IF(C5-3*SQRT(C5*(C5-1))<0,0,C5-3*SQRT(C5*(C5-1)))	=C5+3*SQRT(C5*(C5-1))
6	40071	132	=SUM($B$2:$B$19)/COUNT($A$3:$A$19)	=IF(C6-3*SQRT(C6*(C6-1))<0,0,C6-3*SQRT(C6*(C6-1)))	=C6+3*SQRT(C6*(C6-1))
7	40077	6	=SUM($B$2:$B$19)/COUNT($A$3:$A$19)	=IF(C7-3*SQRT(C7*(C7-1))<0,0,C7-3*SQRT(C7*(C7-1)))	=C7+3*SQRT(C7*(C7-1))
8	40093	16	=SUM($B$2:$B$19)/COUNT($A$3:$A$19)	=IF(C8-3*SQRT(C8*(C8-1))<0,0,C8-3*SQRT(C8*(C8-1)))	=C8+3*SQRT(C8*(C8-1))
9	40139	46	=SUM($B$2:$B$19)/COUNT($A$3:$A$19)	=IF(C9-3*SQRT(C9*(C9-1))<0,0,C9-3*SQRT(C9*(C9-1)))	=C9+3*SQRT(C9*(C9-1))
10	40148	9	=SUM($B$2:$B$19)/COUNT($A$3:$A$19)	=IF(C10-3*SQRT(C10*(C10-1))<0,0,C10-3*SQRT(C10*(C10-1)))	=C10+3*SQRT(C10*(C10-1))
11	40183	35	=SUM($B$2:$B$19)/COUNT($A$3:$A$19)	=IF(C11-3*SQRT(C11*(C11-1))<0,0,C11-3*SQRT(C11*(C11-1)))	=C11+3*SQRT(C11*(C11-1))
12	40208	25	=SUM($B$2:$B$19)/COUNT($A$3:$A$19)	=IF(C12-3*SQRT(C12*(C12-1))<0,0,C12-3*SQRT(C12*(C12-1)))	=C12+3*SQRT(C12*(C12-1))
13	40274	58	=SUM($B$2:$B$19)/COUNT($A$3:$A$19)	=IF(C13-3*SQRT(C13*(C13-1))<0,0,C13-3*SQRT(C13*(C13-1)))	=C13+3*SQRT(C13*(C13-1))
14	40275	67	=SUM($B$2:$B$19)/COUNT($A$3:$A$19)	=IF(C14-3*SQRT(C14*(C14-1))<0,0,C14-3*SQRT(C14*(C14-1)))	=C14+3*SQRT(C14*(C14-1))
15	40353	20	=SUM($B$2:$B$19)/COUNT($A$3:$A$19)	=IF(C15-3*SQRT(C15*(C15-1))<0,0,C15-3*SQRT(C15*(C15-1)))	=C15+3*SQRT(C15*(C15-1))
16	40364	11	=SUM($B$2:$B$19)/COUNT($A$3:$A$19)	=IF(C16-3*SQRT(C16*(C16-1))<0,0,C16-3*SQRT(C16*(C16-1)))	=C16+3*SQRT(C16*(C16-1))
17	40393	49	=SUM($B$2:$B$19)/COUNT($A$3:$A$19)	=IF(C17-3*SQRT(C17*(C17-1))<0,0,C17-3*SQRT(C17*(C17-1)))	=C17+3*SQRT(C17*(C17-1))
18	40459	46	=SUM($B$2:$B$19)/COUNT($A$3:$A$19)	=IF(C18-3*SQRT(C18*(C18-1))<0,0,C18-3*SQRT(C18*(C18-1)))	=C18+3*SQRT(C18*(C18-1))
19	40516	57	=SUM($B$2:$B$19)/COUNT($A$3:$A$19)	=IF(C19-3*SQRT(C19*(C19-1))<0,0,C19-3*SQRT(C19*(C19-1)))	=C19+3*SQRT(C19*(C19-1))

Which gives graphically:



With Minitab 17.3.1 we get something quite different for all control limits. As the Minitab team doesn't give the proofs but only the formulas we are not able to say who has the most correct way between us and them for the calculations:


G Chart of Dates
G Chart of Dates
Event Probability = 0.023
Days Between Events
300
250
200
150
100
50
0
UCL=278.2
CL=28.3
LCL=0
1
3
5
7
9
11
13
15
17
Observation

For a software like Minitab, the lower control limit of the rare events chart is set to the percentile 0.0013499 of the geometric distribution (taking the nearest decimal value by linear interpolation). The upper control limit is set to the percentile 0.99865 (also taking the nearest decimal value by linear interpolation). The central line is set at 0.5 percentile (taking the nearest decimal value by linear interpolation).

It should be also remembered that the Geometric law (or the Negative Binomial law) is not necessarily symmetrical! This is why, in practice, you should rather use the median for the CL (centerline which can be determined by simple simulations of the law and taking the nearest value of the 50% percentile) and for the LCL and UCL control limits use also the probabilistic simulations.

16.7.6.7 Control Charts Operating Characteristic (OC) Curves

In the usage of control charts we face two types of errors (putting apart human error and incompetence error...) that are well known to us in the context of our study of statistics:

- Type I error: We detect an alarm when there is actually no alarm
- Type II error: We are out of control but we do not detect alarm

The aim here will be to determine the conditions for controlling these two errors.

As we will see the operating characteristic function for a Shewhart chart givens the probability that the statistic of interest falls between the lower and upper control limits.

16.7.6.7.1 OC for $\bar{X}$ measurement control charts

We have already proved earlier above that the capability that a centered process taken at a distance of 3σ had a cumulative probability of 0.27% $(1/370)$ to have a point outside the 3σ and this only for reasons of statistical fluctuations (error of Type I)! So in extenso to 3σ the error type II is $1-0.27\%$ which is quite huge but acceptable in practice (as practitioner consider a power of 80% as being acceptable)! The reason for this high amplitude of type II error is implicitly due to the fact that we calculate here:

$$\beta=P(\text{LSL}\leq\bar{X}\leq\text{USL}|\mu_{\bar{X}}=0,\sigma_{\bar{X}}=1)=\Phi(\text{USL})-\Phi(\text{LSL}) \tag{16.7.900}$$

that is to say that the data have all been centered reduced as well as in extenso that the USL and the LSL and where for recall is the symbol Φ stand for the cumulative probability function of the Normal law $\mathcal{N}(0,1)$.

But if the data is not centered, then we must calculate:

$$\beta=P\left(\text{LSL}\leq\bar{X}\leq\text{USL}|\mu_{\bar{X}}=\bar{X},\sigma_{\bar{X}}=\frac{\sigma_X}{\sqrt{n}}\right)=\Phi\left(\frac{\text{USL}-\bar{X}}{\frac{\sigma_X}{\sqrt{n}}}\right)-\Phi\left(\frac{\text{LSL}-\bar{X}}{\frac{\sigma_X}{\sqrt{n}}}\right) \tag{16.7.901}$$

But if we are interested in a certain lag by loss of control of the process, then we have by expressing these shifts in multiples m of σ_X as it is done by tradition (this choice will be justified a little later):

$$\begin{aligned}\beta&=P\left(\text{LSL}\leq\bar{X}\leq\text{USL}|\mu_{\bar{X}}=\bar{X}+m\sigma_{\bar{X}},\sigma_{\bar{X}}=\frac{\sigma_X}{\sqrt{n}}\right)\\&=\Phi\left(\frac{\text{USL}-(\bar{X}+m\sigma_X)}{\frac{\sigma_X}{\sqrt{n}}}\right)-\Phi\left(\frac{\text{LSL}-(\bar{X}+m\sigma_X)}{\frac{\sigma_X}{\sqrt{n}}}\right)\end{aligned} \tag{16.7.902}$$

Now, let us recall that we have in the framework of a $\bar{X}$ control chart:

$$\begin{aligned}\text{USL} &= \bar{X} + k\sigma_{\bar{X}} \\ \text{LSL} &= \bar{X} - k\sigma_{\bar{X}}\end{aligned} \tag{16.7.903}$$

Therefore it comes:

$$\begin{aligned}\beta &= \Phi\left(\frac{(\bar{X}+k\sigma_{\bar{X}})-(\bar{X}+m\sigma_X)}{\frac{\sigma_X}{\sqrt{n}}}\right) - \Phi\left(\frac{(\bar{X}-k\sigma_{\bar{X}})-(\bar{X}+m\sigma_X)}{\frac{\sigma_X}{\sqrt{n}}}\right) \\ &= \Phi\left(\frac{k\sigma_{\bar{X}}-m\sigma_X}{\frac{\sigma_X}{\sqrt{n}}}\right) - \Phi\left(\frac{-k\sigma_{\bar{X}}-m\sigma_X}{\frac{\sigma_X}{\sqrt{n}}}\right) = \Phi\left(\frac{k\frac{\sigma_X}{\sqrt{n}}-m\sigma_X}{\frac{\sigma_X}{\sqrt{n}}}\right) - \Phi\left(\frac{-k\frac{\sigma_X}{\sqrt{n}}-m\sigma_X}{\frac{\sigma_X}{\sqrt{n}}}\right) \\ &= \Phi(k-m\sqrt{n}) - \Phi(-k-m\sqrt{n})\end{aligned} \tag{16.7.904}$$

We see that the choices at the beginning in the way of noting the offset advantage us by the fact that the error of type II is a function of more than three parameters of which k is almost always fixed at 3 and n is chosen by the experimenter.

So finally we have:

$$\boxed{\beta(m) = \Phi(k-m\sqrt{n}) - \Phi(-k-m\sqrt{n})} \tag{16.7.905}$$

✍**Example:**

Suppose we use a $\bar{X}$ control chart with a k of $3\sigma_{\bar{X}}$, a sample size of $n=5$ and we want to know the Type II error to detect an offset of $2\sigma_X$ $(m=2)$. From then on, it comes:

$$\beta(m) = \Phi(3-2\sqrt{5}) - \Phi(-3-2\sqrt{5}) \cong 0.0705 = 7.05\% \tag{16.7.906}$$

So the risk of type II is this time much lower but the error of type I then becomes $1-0.0705=0.9295$ which is huge. This is why it is often necessary to find the right compromise between the quality policy and the adjustment costs.

Let us indicate that the general case, for k equal to 3, we have the following curve which is named "$\bar{X}$ control chart operating characteristic curve":

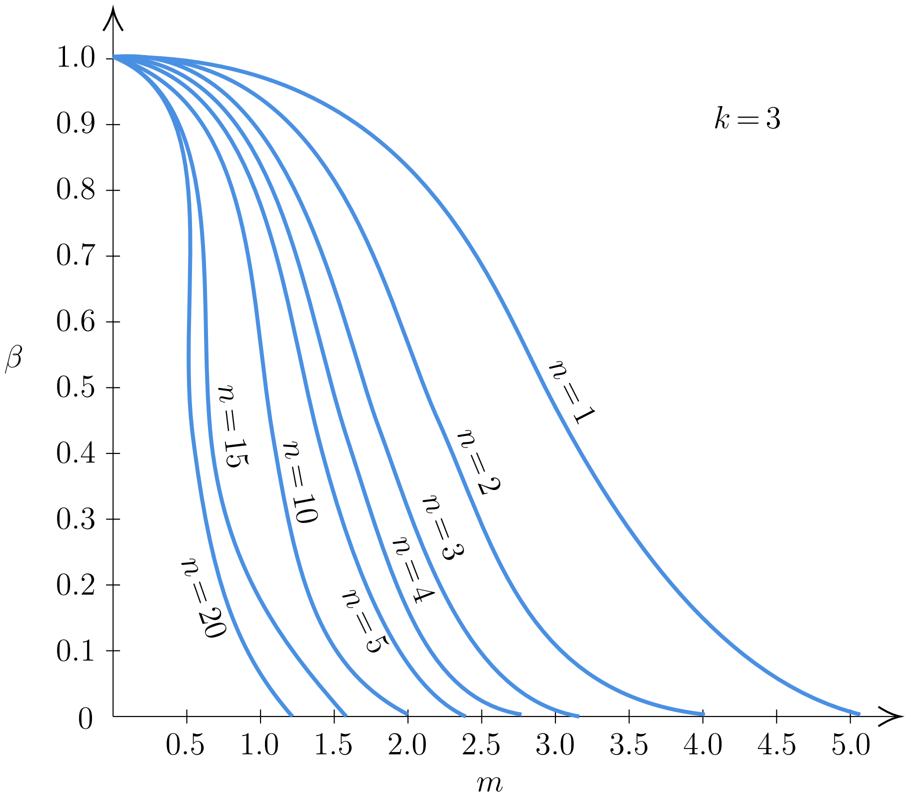


Figure 16.286: $\bar{X}$ **control chart operating characteristic curve**

So on the plot above we have $k=3$, $n=5$, $m=2$:

$$\beta(2)=\Phi(k-m\sqrt{n})-\Phi(-k-m\sqrt{n})=\Phi(3-2\sqrt{5})-\Phi(-3-2\sqrt{5})\cong 0.07 \tag{16.7.907}$$

either with the English version of Microsoft® Excel® 14.0.7177:

```
=NORM.S.DIST(3-2*SQRT(5),TRUE)-NORM.S.DIST(-3-2*SQRT (5),TRUE)
```

16.7.6.7.2 OC for P-type attribute control charts

The same type of reasoning applies for an attributes P-type control chart! We have then:

$$\beta=P(\text{LSL}\leq\hat{p}\leq\text{USL}|n,p)=\mathcal{B}\left(\bar{\hat{p}}+k\sqrt{\frac{\bar{\hat{p}}(1-\bar{\hat{p}})}{N}},n,\bar{\hat{p}}\right)-\mathcal{B}\left(\bar{\hat{p}}-k\sqrt{\frac{\bar{\hat{p}}(1-\bar{\hat{p}})}{N}}n,\bar{\hat{p}}\right) \tag{16.7.908}$$

However, we saw in the Statistics section (see page 872) that when N is large the interval could be estimated using a Normal law such that:

$$\hat{p}-Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\leq p\leq\hat{p}+Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{16.7.909}$$

Hence, since $\bar{\hat{p}}$ follow approximately a Normal law, we arrive at the same type of calculations as for the previous control chart:

$$\begin{aligned}\beta &= P\left(\text{LSL} \leq \bar{\hat{p}} \leq \text{USL} | p = \bar{\hat{p}}, \sigma = \sqrt{\frac{\bar{\hat{p}}(1-\bar{\hat{p}})}{N}}\right)\\ &= \Phi\left(\frac{\text{USL}-\bar{\hat{p}}}{\sqrt{\dfrac{\bar{\hat{p}}(1-\bar{\hat{p}})}{N}}}\right) - \Phi\left(\frac{\text{LSL}-\bar{\hat{p}}}{\sqrt{\dfrac{\bar{\hat{p}}(1-\bar{\hat{p}})}{N}}}\right)\end{aligned} \tag{16.7.910}$$

Thus finally:

$$\boxed{\beta = \Phi\left(\frac{\text{USL}-\bar{\hat{p}}}{\sqrt{\dfrac{\bar{\hat{p}}(1-\bar{\hat{p}})}{N}}}\right) - \Phi\left(\frac{\text{LSL}-\bar{\hat{p}}}{\sqrt{\dfrac{\bar{\hat{p}}(1-\bar{\hat{p}})}{N}}}\right)} \tag{16.7.911}$$

16.7.7 Design of reliability tests

"Reliability test plans" have two main functions: One is to determine the sample size and testing time needed to estimate model parameters. The other is to demonstrate that we have met specified reliability requirements.

A test plan includes:

- The number of units you need to test.
- A stopping rule: the amount of time you must test each unit or the number of failures that must occur.
- Success criterion: the number of failures allowed while the test still passes (for example, all units are tested for the specified amount of time and there are no failures).

Three types of test plans are available: demonstration, estimation, and accelerated life.

Definitions (#16.7.740):

D1. "Demonstration plan" to determine the sample size or testing time needed to demonstrate, with some level of confidence, that the reliability is higher than a particular standard.

There are two types of demonstration tests:

1. "Substantiation tests" that provide statistical evidence that a redesigned system suppressed or significantly reduced a known cause of failure. For example, is the redesigned system better than the previous system?

   We can rewrite these hypotheses in terms of the scale (Weibull or exponential distribution) or location Reliability test (other distributions), a percentile, the reliability at a particular time, or the mean time to failure (MTTF). For example, we can test whether the MTTF for a redesigned system is greater than the MTTF for the previous system.

2. "Reliability test" that provide statistical evidence that a reliability specification is achieved. For example, is the system reliability greater than a goal value?

D2. "Estimation test plans" to determine the number of test units that we need to estimate percentiles or reliabilities with a specified degree of precision. Estimation test plans are similar to classic sample-size problems, but computations are more intensive because the data are usually censored.

D3. "Accelerated life test plans" to determine - under some very strong assumptions - the number of units to test and the allocation of those test units across stress levels for an accelerated life test. We can also use these test plans to determine the standard error for the parameter we want to estimate for a fixed number of test unit

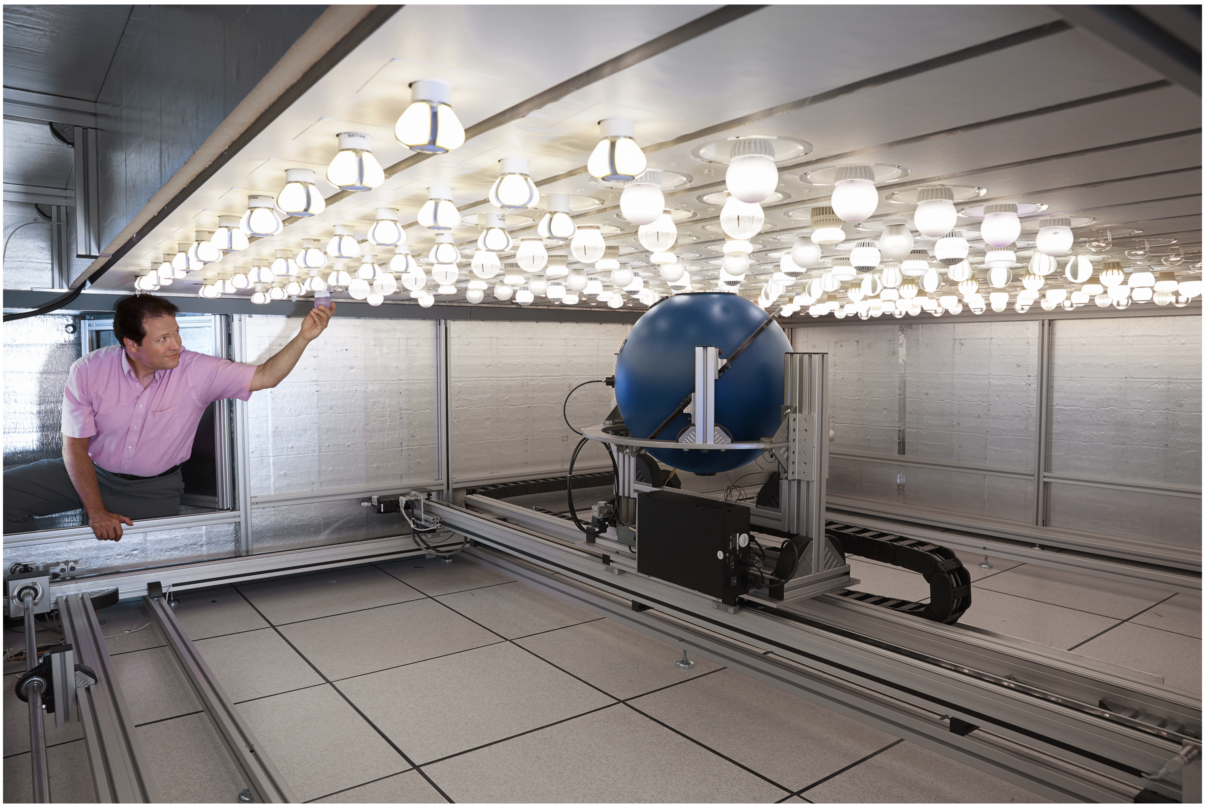


Figure 16.287: **Pacific Northwest National Laboratory automatic long-term performance measurement equipment for Philips LED 60 Watt, which can test up to 200 lamps at a time**

Several methods have been designed to help engineers for all this purposes: Cumulative Binomial, Nonparametric Binomial, Exponential Chi-Squared and Nonparametric Bayesian. They are discussed in the following paragraphs.

16.7.7.1 Chi-squared time of test

A method for designing tests for products that have an assumed constant failure rate, or exponential life distribution, draws on the chi-squared distribution. This method only returns the necessary accumulated test time for a demonstrated reliability or $\overline{\text{MTTF}}$ (or for not repairable items for the $\overline{\text{MTBF}}$ obviously!).

We have seen earlier above that for a survival function of the form:

$$R(t)=e^{-\lambda t} \tag{16.7.912}$$

The corresponding density function was the exponential law:

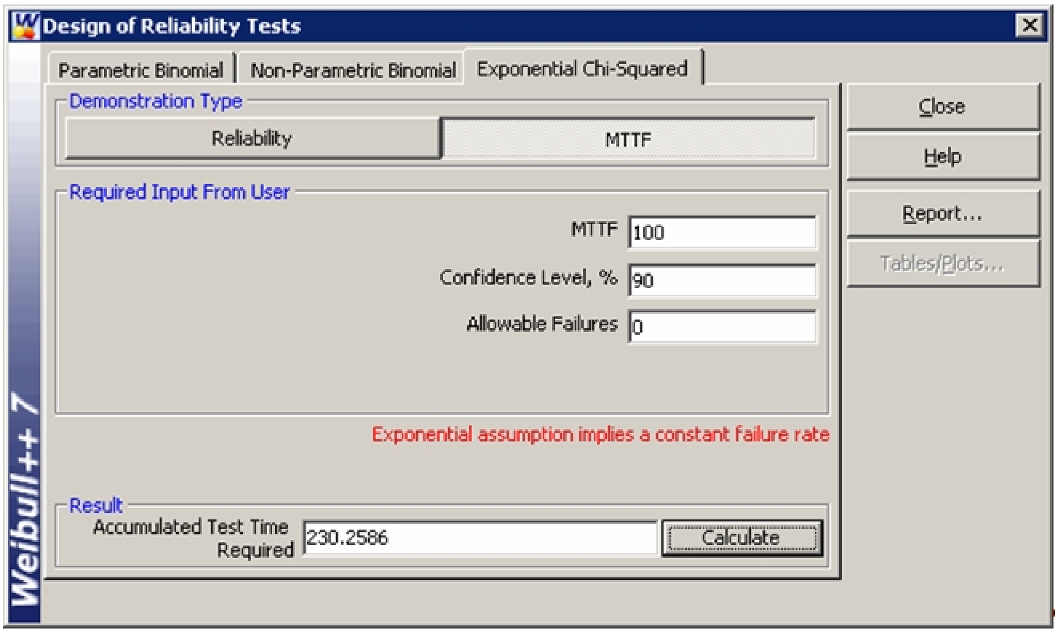
$$f(t)=\lambda e^{-\lambda t} \tag{16.7.913}$$

and that we had (also valid for the MTBF):

$$\text{MTTF}=\frac{1}{\mu} \tag{16.7.914}$$

Our objective here will be to find the value that permits a reliability engineer to know how long a test must be performed to demonstrate with a 90% confidence level (CL) that its product has zero failure.

What looks like in the worldwide reference Weibull++ software for reliability engineering as following:


Design of Reliability Tests
Parametric Binomial
Non-Parametric Binomial
Exponential Chi-Squared
Demonstration Type
Reliability
MTTF
Close
Help
Report...
Tables/Plots...
Required Input From User
Weibull++ 7
Calculate

MTTF 100

Confidence Level, % 90

Allowable Failures 0

Exponential assumption implies a constant failure rate

Result

Accumulated Test Time Required 230.2586

Figure 16.288: Exponential Chi-squares time test in Weibull++ 7 software

We will prove in the following that this result follows from a close relationship between the exponential distribution, the Poisson distribution through the queue theory (see section Quantitative Management page 8141) and the chi-square distribution.

To prove this result, let us recall that we have prove in the section of Quantitative Management that under certain hypotheses of a queue, the probability of seeing an event in an interval of duration t is given by the function:

$$p_k \cong \frac{(\lambda t)^k}{k!}e^{-\lambda t} \tag{16.7.915}$$

and that the corresponding density function of the corresponding time between two events is an exponential law. In other words, when the time of failure follows an exponential law, the probability of a certain number of failures in a certain time interval follows the relation given just before!

The previous relation can be used in the case of reliability to obtain an upper bound of λ for a given time interval t since we can then write with the Erlang distribution function (see section Quantitative Management page 8147):

$$F(t,k)=1-\text{CL}=\sum_{k=0}^{f}\frac{(\lambda t)^k}{k!}e^{-\lambda t} \tag{16.7.916}$$

A small and simple change of variable gives:

$$1-\text{CL}=\sum_{k=0}^{f}\frac{x^k}{k!}e^{-x} \tag{16.7.917}$$

We therefore recognize the Poisson distribution! However, this is impossible to reverse by hand (as far as we know). Reason why it is wiser to find a relation with another known distribution that is continuous in the purpose to reverse it by hand.

For this, let us recall that we have proved during our study of the chi-square law in the section Statistics (see page 773) that:

$$\sum_{k=0}^{n-1}\frac{\lambda^k}{k!}e^{-\lambda}=1-\int_0^{\lambda}\frac{1}{2^{k/2}\Gamma(k/2)}x^{k/2-1}e^{-x/2}\mathrm{d}x \tag{16.7.918}$$

Therefore, by re-reading carefully the changes of variables made in the section Statistics we have in other words:

$$\sum_{k=0}^{n-1}\frac{\lambda^k}{k!}e^{-\lambda}=1-\chi_k^2(x\leq 2\lambda)=1-\chi_{2n}^2(x\leq 2\lambda) \tag{16.7.919}$$

It is therefore immediate that:

$$1-\mathrm{CL}=1-\sum_{k=0}^{f}\frac{(\lambda t)^k}{k!}e^{-\lambda t}=1-\chi_{2(f+1)}^2(x\leq 2\lambda t) \tag{16.7.920}$$

It follows that:

$$\mathrm{CL}=1-\sum_{k=0}^{f}\frac{x^k}{k!}e^{-x}=x_{2(f+1)}^2(2\lambda t) \tag{16.7.921}$$

Conversely:

$$\chi_{\mathrm{CL},2(f+1)}^2=2\lambda t=\frac{2t}{\mathrm{MTTF}} \tag{16.7.922}$$

Therefore:

$$\boxed{t=\frac{\mathrm{MTTF}\chi_{\mathrm{CL},2(f+1)}^2}{2}} \tag{16.7.923}$$

16.7.7.2 Binomial sampling size

In demonstration plans, an important calculation is to determine the size of a sample in order to demonstrate a certain degree of reliability for a given level of confidence. In the case where the test time is equal to the demonstration time, the binomial distribution is widely used in practice because nonparametric and we need only to have the knowledge of the estimated parameter R (see earlier above):

$$\boxed{1-\text{CL}=\sum_{i=0}^{f}\binom{n}{i}(1-R)^i R^{n-i}=\sum_{i=0}^{f}\frac{n!}{i!(n-i)!}(1-R)^i R^{n-i}} \quad (16.7.924)$$

with f the number of failures, n the size of the sample, and R the reliability that is sought to be demonstrated. We then speak of a "nonparametric binomial demonstration plan" or simply sometimes of "nonparametric zero failure".

As the reader may have noticed it, there is no time value associated with this methodology, so one must assume that the value of R is associated with the amount of time for which the units were tested.

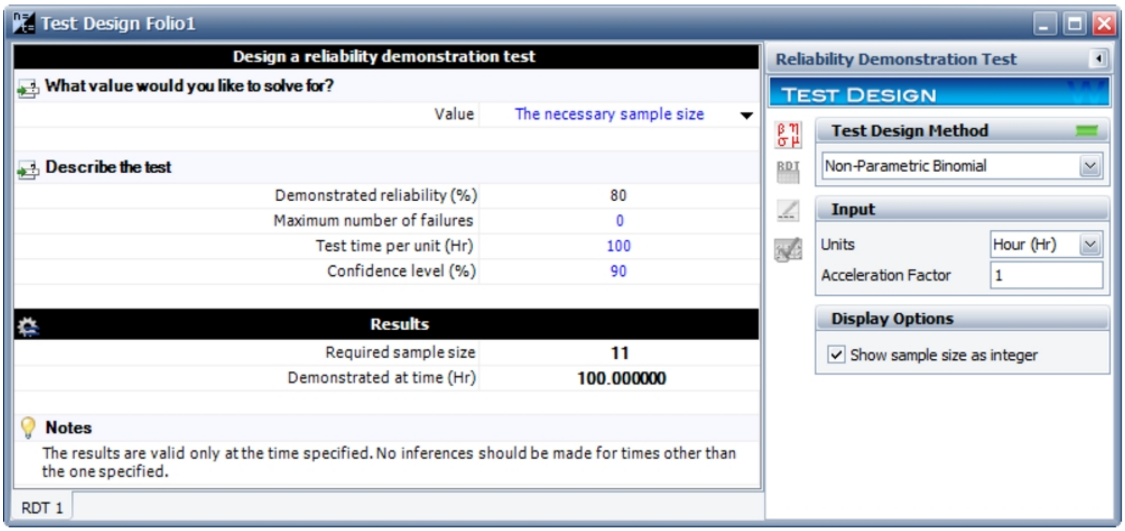
✍**Example:**

Consider that an engineer wants to set up a demonstration plan in order to demonstrate the zero failure of a product with an 80% reliability with a confidence level of 90%. We have then:

$$1-90\%=10\%=\sum_{i=0}^{0}\binom{n}{i}(1-R)^iR^{N-i}=R^N=0.80^N \qquad (16.7.925)$$

This gives by taking the logarithm:

$$N=\frac{\log(0.1)}{\log(0.80)}=10.31\cong 11 \qquad (16.7.926)$$



Design a reliability demonstration test

What value would you like to solve for?

Value	The necessary sample size

Describe the test

Demonstrated reliability (%)	80
Maximum number of failures	0
Test time per unit (Hr)	100
Confidence level (%)	90

Results

Required sample size	11
Demonstrated at time (Hr)	100.000000

Notes

The results are valid only at the time specified. No inferences should be made for times other than the one specified.

Test Design Method: Non-Parametric Binomial

Input

Units	Hour (Hr)
Acceleration Factor	1

Display Options

☑ Show sample size as integer

Figure 16.289: Binomial sample size demonstration plan in Weibull++ 7 software

However, in reliability tests, in order to better estimate reliability, Bayesian methods are used, especially when sample size is financially limited. If there is an a priori information about the reliability of the system, this information can then be used to build better demonstration plans.

We will see later that the application of Bayesian techniques gives a quasi-similar result to the above example for certain values a priori well chosen...

We can involve also the reliability indirectly with a reliability function. For this, let us consider the following example:

✍**Example:**

We want to design a test to demonstrate a reliability of 90% at $t_{\text{demo}}=100$ hours with a 95% confidence if no failure occur during the test. We will assume a Weibull distribution with a shape parameter $\beta=1.5$ and we know that we have the testing facilities available for $t=48$ hours.

We must now determine the number of units to test for this amount of time with no failures in order to have demonstrated our reliability goal. The first step is to determine the Weibull scale parameter, η. The Weibull reliability equation is:

$$R = e^{-(t/\eta)^{\beta}} \tag{16.7.927}$$

This can be rewritten as:

$$\eta = \frac{t_{\text{demo}}}{(-\ln(R_{\text{demo}})^{\frac{1}{\beta}}} \tag{16.7.928}$$

Since we know the values of t_{demo}, R_{demo} and β, we can substitute these in the equation and solve for η:

$$\eta = \frac{100}{-\ln(0.9)^{\frac{1}{1.5}}} = 448.3 \tag{16.7.929}$$

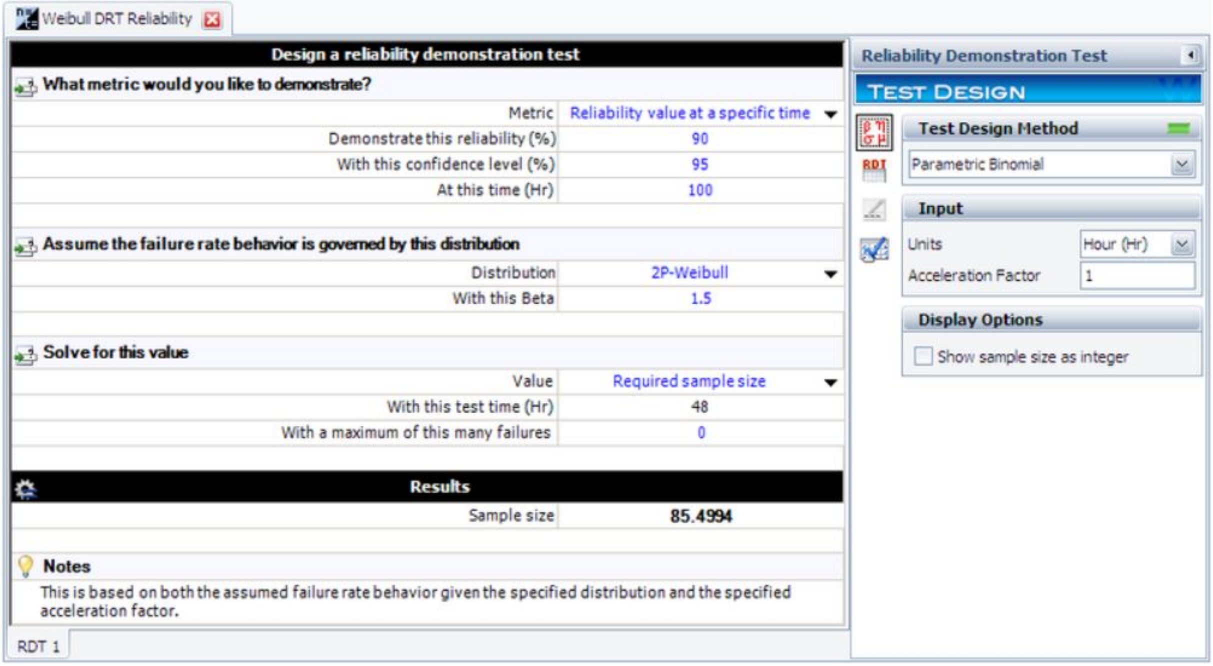
Next the value of R_{test} is calculated by:

$$R_{\text{test}} = e^{-(t_{\text{test}}/\eta)^{\beta}} = e^{-(48/448.3)^{1.5}} = 0.966 = 96.6\% \tag{16.7.930}$$

The last step is to substitute the appropriate values into the cumulative binomial equation, which for the Weibull distribution appears as:

$$1-\text{CL} = \sum_{i=0}^{f} \frac{n!}{i!(n-i)!} \cdot \left(1 - e^{-(t_{\text{test}}/\eta)^{\beta}}\right)^{i} \cdot \left(e^{-(t_{\text{test}}/\eta)^{\beta}}\right)^{(n-i)} \tag{16.7.931}$$

The values of CL, t_{test}, β, f and η have already been calculated or specified, so it merely remains to solve the equation for n. This value is $n = 85.4994$, or $n = 86$ units, since the fractional value must be rounded up to the next integer value. This example solved in Weibull++ is shown next.



Design a reliability demonstration test

What metric would you like to demonstrate?

Metric	Reliability value at a specific time
Demonstrate this reliability (%)	90
With this confidence level (%)	95
At this time (Hr)	100

Assume the failure rate behavior is governed by this distribution

Distribution	2P-Weibull
With this Beta	1.5

Solve for this value

Value	Required sample size
With this test time (Hr)	48
With a maximum of this many failures	0

Results

Sample size	85.4994

Notes

This is based on both the assumed failure rate behavior given the specified distribution and the specified acceleration factor.

Test Design Method: Parametric Binomial

Units: Hour (Hr)

Acceleration Factor: 1

☐ Show sample size as integer

Figure 16.290: **Weibull reliability sample size with Weibull++**

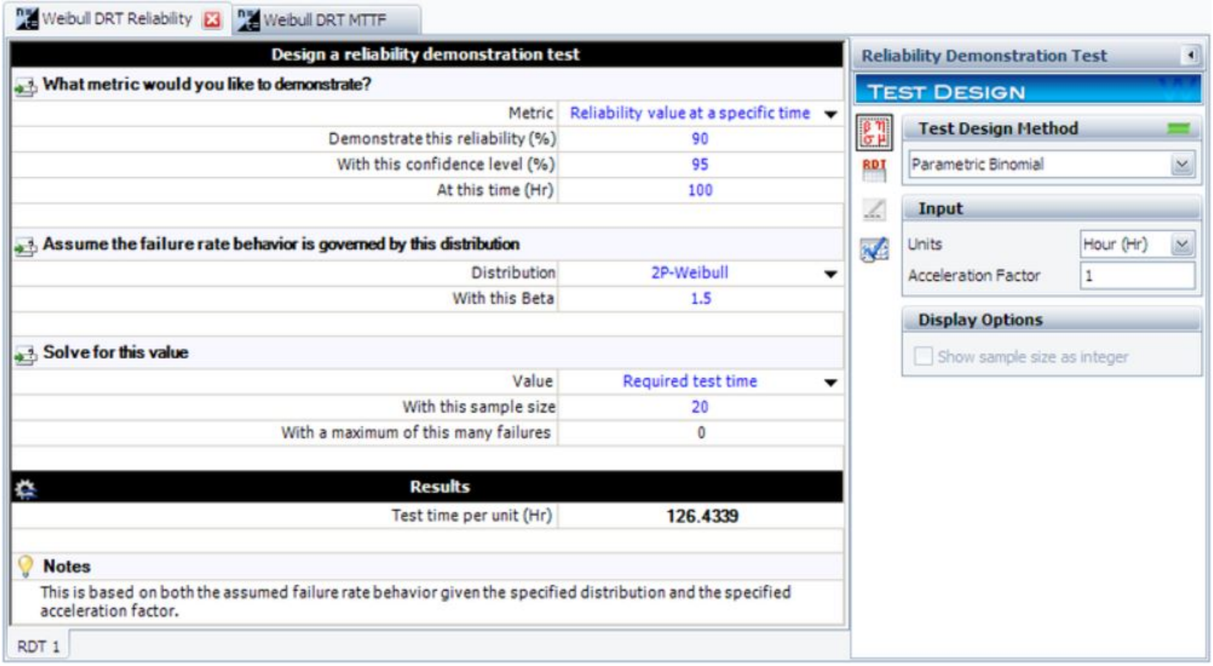
Now we can consider the case where we know the sample size but we want to determine the test time.

✍Example:

In this case, we will assume that we have 20 units to test, $n = 20$, and must determine the test time, t_{test}. We have already determined the value of the scale parameter, η in the previous example. Since we know the values of n, CL, f, η and β, it remains to solve the binomial equation:

$$1 - \text{CL} = \sum_{i=0}^{f} \frac{n!}{i!(n-i)!} \cdot \left(1 - e^{-(t_{\text{test}}/\eta)^{\beta}}\right)^{i} \cdot \left(e^{-(t_{\text{test}}/\eta)^{\beta}}\right)^{(n-i)} \qquad (16.7.932)$$

with the Weibull distribution for t_{test}. This value is $t_{\text{text}} = 126.4339$ hours. This example solved in Weibull++ is shown next.



Design a reliability demonstration test

What metric would you like to demonstrate?

Metric	Reliability value at a specific time
Demonstrate this reliability (%)	90
With this confidence level (%)	95
At this time (Hr)	100

Assume the failure rate behavior is governed by this distribution

Distribution	2P-Weibull
With this Beta	1.5

Solve for this value

Value	Required test time
With this sample size	20
With a maximum of this many failures	0

Results

Test time per unit (Hr)	126.4339

Notes

This is based on both the assumed failure rate behavior given the specified distribution and the specified acceleration factor.

Test Design Method

Parametric Binomial

Input

Units	Hour (Hr)
Acceleration Factor	1

Display Options

☐ Show sample size as integer

Figure 16.291: **Weibull reliability test time with Weibull++**

16.7.7.3 Test to demonstrate MTBF (or MTTF)

Designing a test to demonstrate a certain value of the MTBF is identical to designing a reliability demonstration test.

✍**Example:**

We want to use the parametric binomial method to design a test that will demonstrate $\text{MTBF} = 75$ hours with a 95% confidence if no failure occur during the test $f = 0$. We will assume a Weibull distribution with a shape parameter $\beta = 1.5$. We want to determine the number of units to test for $t_{\text{text}} = 60$ hours to demonstrate this goal.

The first step in this case involves determining the value of the scale parameter η from the MTBF equation. The equation for the MTBF for the Weibull distribution is for recall:

$$\text{MTBF} = \eta\Gamma\left(1+\frac{1}{\beta}\right) \qquad (16.7.933)$$

This can be rearranged in terms of η:

$$\eta = \frac{\text{MTBF}}{\Gamma\left(1+\dfrac{1}{\beta}\right)} \qquad (16.7.934)$$

Since MTBF and β have been specified, it is a relatively simple matter to calculate $\eta = 83.1$. From this point on, the procedure is the same as the reliability demonstration example. Next, the value of R_{test} is calculated as:

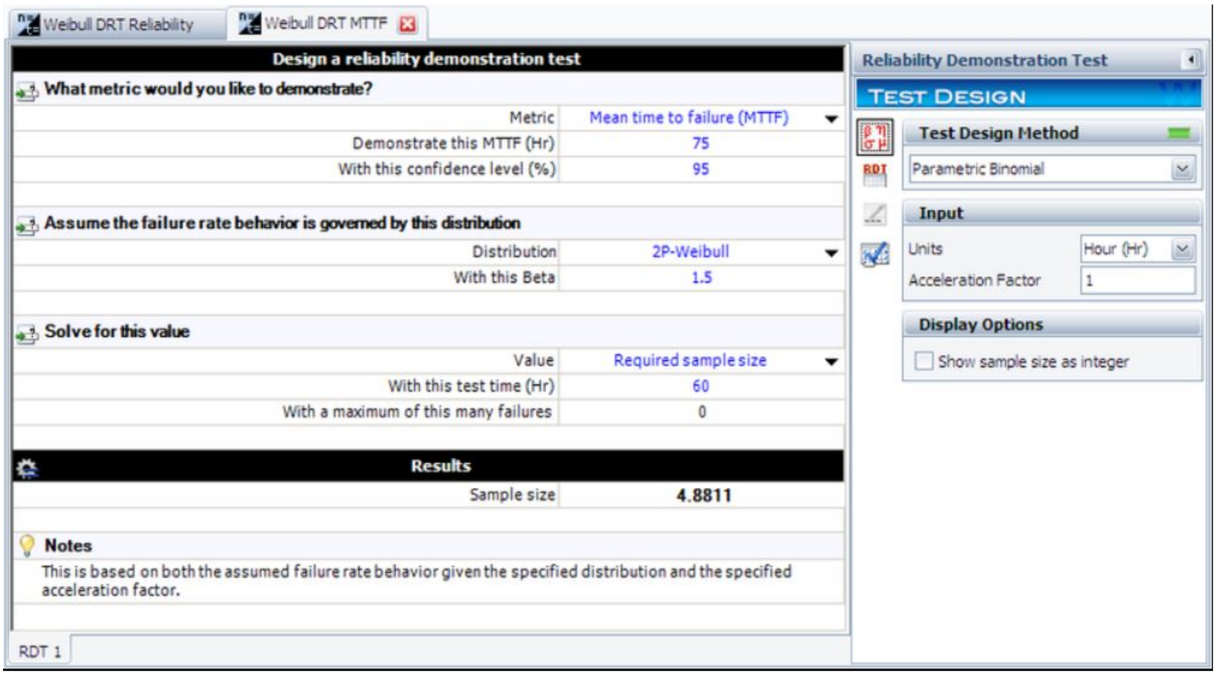
$$R_{\text{test}} = e^{-(t_{\text{test}}/\eta)^{\beta}} = e^{-(60/83.1)^{1.5}} = 0.541 = 54.1\% \qquad (16.7.935)$$

The last step is to substitute the appropriate values into the cumulative binomial equation. The values of CL, t_{test}, β, f and η have already been calculated or specified, so it merely remains to solve the binomial equation:

$$1-\text{CL}=\sum_{i=0}^{f}\frac{n!}{i!(n-i)!}\cdot\left(1-e^{-(t_{\text{test}}/\eta)^{\beta}}\right)^{i}\cdot\left(e^{-(t_{\text{test}}/\eta)^{\beta}}\right)^{(n-i)} \qquad (16.7.936)$$

for n.

The value is calculated as $n=4.8811$, or $n=5$ units, since the fractional value must be rounded up to the next integer value. This example solved in Weibull++ is shown next.



Design a reliability demonstration test

What metric would you like to demonstrate?

Metric	Mean time to failure (MTTF)
Demonstrate this MTTF (Hr)	75
With this confidence level (%)	95

Assume the failure rate behavior is governed by this distribution

Distribution	2P-Weibull
With this Beta	1.5

Solve for this value

Value	Required sample size
With this test time (Hr)	60
With a maximum of this many failures	0

Results

Sample size	**4.8811**

Notes

This is based on both the assumed failure rate behavior given the specified distribution and the specified acceleration factor.

Test Design Method

Parametric Binomial

Input

Units	Hour (Hr)
Acceleration Factor	1

Display Options

☐ Show sample size as integer

Figure 16.292: **Weibull reliability sample size for MTBF test with Weibull++**

The procedure for determining the required test time proceeds in the same manner, determining η from the MTBF (or MTTF) equation, and following the previously described methodology to determine t_{test} from the binomial equation with Weibull distribution.

16.7.7.4 Beta-binomial sampling size

The regular nonparametric analysis performed either on the binomial or the chi-squared equation were done with only the direct system test data. However, if prior information regarding system performance is available, it can be incorporated into a Bayesian nonparametric analysis. This subsection will demonstrate how to incorporate prior information about system reliability and also how to incorporate prior information from subsystem tests into system test design.

Thus, since R is bounded in the interval $[0,1]$ a well-known prior is the beta distribution (see section Statistics page 760). We proved in the section Statistics during our study of Bayesian inference that the posterior was then a beta-binomial distribution (see section Statistics page 898) for which we had specified that reliability engineers noted the cumulative probability under the somewhat unfortunate notation:

$$B(N-f+a,f+b) \tag{16.7.937}$$

Therefore, we have in our case a posteriori:

$$1-\text{CL}=B(N-f+a,f+b) \tag{16.7.938}$$

We know f which is imposed to us by the demonstration plan, we are always trying to determine N. We know that for the beta law:

$$P_{a,b}(x)=\frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}1_{]0,1[} \tag{16.7.939}$$

that we have:

$$\mu=\frac{a}{a+b} \qquad \sigma^2=\frac{ab}{(a+b)^2(a+b+1)} \tag{16.7.940}$$

With a little elementary algebra we get:

$$a=\mu\left[\frac{\mu-\mu^2}{\sigma}-1\right] b \quad =(1-\mu)\left[\frac{\mu-\mu^2}{\sigma}-1\right] \tag{16.7.941}$$

The question is then what to do now ...? Then here the statistician engineer uses a trick ... It takes the generalized beta law with 4 parameters including thus the famous optimistic, pessimistic interval $[O,P]$ that we studied in the section of Quantitative Management (see page 8074) and uses the corresponding relation of mean and variance that we have proved:

$$\begin{aligned}\mu=\text{E}(R)&=\frac{O+4M_0+P}{6}\\ \sigma^2=\text{V}(R)&=\left(\frac{P-0}{6}\right)^2\end{aligned} \tag{16.7.942}$$

Where normally we should use for the modal value as we have proved in the section of Quantitative Management:

$$M_0=\frac{2(O+P)+\sqrt{2}(b-a)}{4} \tag{16.7.943}$$

but the overwhelming majority of engineering project management and reliability software make the fundamental error of asking the modal value to the user (we will see a concrete example a little later with the most famous reliability software used in the USA).

Once this is done. The idea is to return to a beta law with two parameters by injecting the values of expectation and variance obtained by the beta law with 4 parameters in:

$$\begin{aligned}a&=\mu\left[\frac{\mu-\mu^2}{\sigma}-1\right]\\ b&=(1-\mu)\left[\frac{\mu-\mu^2}{\sigma}-1\right]\end{aligned}\tag{16.7.944}$$

which obviously imposes indirectly that beta laws with 4 and 2 parameters have then same variance and even expectation!

✍**Example:**

Consider that an engineer wants to set up a demonstration plan in order to demonstrate that the reliability of its product is 90% with an 80% confidence level assuming that a single failure is allowed and that the engineer believes that the optimistic reliability is of $O=87\%$, the most likely (modal) value of $M_0=90\%$ and the greatest reliability is equal to $P=99\%$.

We have then:

$$\begin{aligned}\mu&=\text{E}(R)=\frac{O+4M_0+P}{6}=0.91\\ \sigma^2&=\text{V}(R)=\left(\frac{P-0}{6}\right)^2=0.0004\end{aligned}\tag{16.7.945}$$

and thus the parameters of the beta law with two parameters are then:

$$\begin{aligned}a&=\mu\left[\frac{\mu-\mu^2}{\sigma}-1\right]=185.4125\\ b&=(1-\mu)\left[\frac{\mu-\mu^2}{\sigma}\right]=18.337\end{aligned}\tag{16.7.946}$$

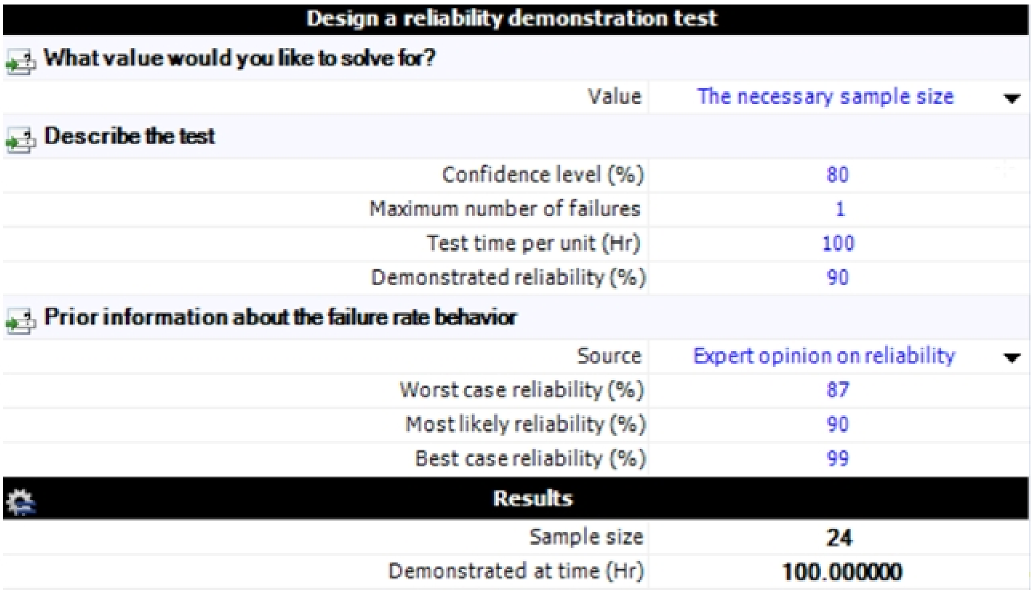
We must then find N such that:

$$1-\text{CL}=20\%=B(N-f+a,f+b)=B(N+184.4125,19.337)\tag{16.7.947}$$

It is easy to get N with a spreadsheet software like Microsoft® Excel® using the BETA.DIST with Microsoft® Excel® 2007 and higher:

```
= BETA.DIST(90%;N+184.4125-1;19.337;1)
```

And with the Goal Seek tool, we get the result $N=24$ by rounding to the nearest integer. This corresponds well to the Weibull++ software output (see next page):



Design a reliability demonstration test

What value would you like to solve for?

Value	The necessary sample size

Describe the test

Confidence level (%)	80
Maximum number of failures	1
Test time per unit (Hr)	100
Demonstrated reliability (%)	90

Prior information about the failure rate behavior

Source	Expert opinion on reliability
Worst case reliability (%)	87
Most likely reliability (%)	90
Best case reliability (%)	99

Results

Sample size	**24**
Demonstrated at time (Hr)	**100.000000**

Figure 16.293: **Beta-Binomial sample size demonstration plan in Weibull++ 7 software**

It is interesting to know that if the modal value M_0 had been calculated rather than requested, we would have therefore $N=-30$ since in the latter case $M_0=97.246\%$... So we see that every model has obviously its limits...

If we use the same parameters with $O=87\%$, $M_0=90\%$, $P=99\%$ with our case but now with the Binomial law that we used in the first example above, then the result rounded to the nearest integer givs $N=13$.

~60 %

★★★☆☆

19 votes, 53.68%

17

Epilogue

You have travelled a long distance with us through this textbook. We have now reached the epilogue, where by tradition, the main redactor is allowed to give voice its own personal opinions.

Indeed, I want to leave you with some of my thoughts on theory versus practice, business and engineering education, applied mathematics research, and what I hope you will take with you after having read this textbook.

By nature, academic engineering is very closely related to its practice. Theory and practice are ruled by the same ideas. As an academic myself teaching since 12001 (holocene calendar) in Fortune 500 and SME companies, I am proud to claim that the majority of engineering ideas were either invented or developed in academia first before they crossed over into practice.

But engineering research is not just for aspiring academics! As I know it very well management and economics consultants are basically researchers even if most of them have a very low analytic level. Firms like McKinsey, Ernst & Young, KPMG and Accenture may have different audiences, production speeds, team systems, and publication and evaluation processes, but they research the same issues that academics do and with the same methods at the exception that the level is significantly lower.

There is also much cross-fertilization! Many professors work regularly with major consulting or asset-investment firms and some have even quit academia altogether to quadruple their pay (or increase their global happiness...).

Because engineering is by nature such an applied discipline, after reading this textbook, you should probably not need anything else to understand engineering research today. In an ideal World, you should be able to read the current state-of-the-art research right now.

So, do we really understand engineering? Certainly not fully! We haven seen through this textbook that engineering is as much an art as it is a science. Given our deficiencies, given that all our methods have their errors, what should we do? My best advice to you is to use common sense, to employ a number of different techniques to come up with a range of possible answers, avoid cognitive biases, and to then make a judgement at the end of the day as to what estimate appears most reasonable in light of different models and your peer-reviewers.

Our book has covered the principles of engineering in some depth and breadth. You should be very well prepared now for the next steps in your lifetime engineering education.

I have enjoyed compiling this textbook in the same way that I enjoy writing my training books, and pretty much for the same reason: It has been like solving an intriguing puzzle that no one else has figured out in quite the same way a particular way to see and explain finance. Of course, compiling it has taken me quite a long times (20 years, without translation: 15 years!).

But it will all have been worth it if you have learned from it. If you have studied the book, you should now know about 99% of what I know about engineering. Interestingly, there were a number of topics that I thought I had understood, but had not - and it was only my having to explain them to you that clarified them for me, too. And this brings me to a key point that I want to leave you with - never be afraid to ask questions, even about first principles! To do so is not a sign of stupidity on the contrary, it is often a sign of deepening awareness and understanding.

I have no illusions: You will very likely not remember all the fine details in this textbook as time passes - I know I won't. But more than the details, I hope that I will have left you with an appreciation for the big ideas, an arsenal of tools, a method for approaching novel problems, and a new perspective. You can now think like an Engineer or even better: as a Scientist!

KNOWLEDGE
IS
POWER

18

Biographies

To be informed of Nobel laureates (physics, chemistry, economics), Fields Medal... click this link Nobel Prizes or on this one Fields Medal.

This section contains a list of some humans who have a strange reputation. Under the rules of history that is taught in elementary school, they do not exist, they haven't ordered any army, they sent nobody to death, they did not have any empire and they had only a minor part in major historical decisions. Some have acquired some celebrity, but none was ever a national hero. Yet their work has more influenced on the course of history than many acts by statesmen crowned with a far greater glory. They produced more turmoil than the comings and goings of armies in battle over borders, they have done more for the happiness or unhappiness that the edicts of kings and assemblies, because their work, is to have shaped the mind of man!

The one who spreads his ideas has a power much greater than that of the sword or sceptre: this is why they have shaped and directed the world. For the most part they have not lifted any finger to act physically, they worked mainly as intellectuals, in silence and oblivion, without worrying of the surrounding world. But in their wake, empires have crumbled, political regimes have either strengthened or eroded, the classes were pitted against each other, and also did nations. Not under the influence of a dark conspiracy, but by the extraordinary power of their ideas. Who are these humans?: Scientists, economists, chemists, biologists, mathematicians, physicists, computer scientists, engineers, ...

The biographies below of the most famous scientists around the world and cited in the various chapters of this textbook are sorted alphabetically and almost all texts are simplify a copy/paste of the French Wikipedia (and translated afterwards in english).

We also pay tribute to the millions of scientists[1], engineers, philosophers, craftsmen, artists, known and anonymous amateurs whose collaboration enabled through millennia the evolution of science and of the human condition!

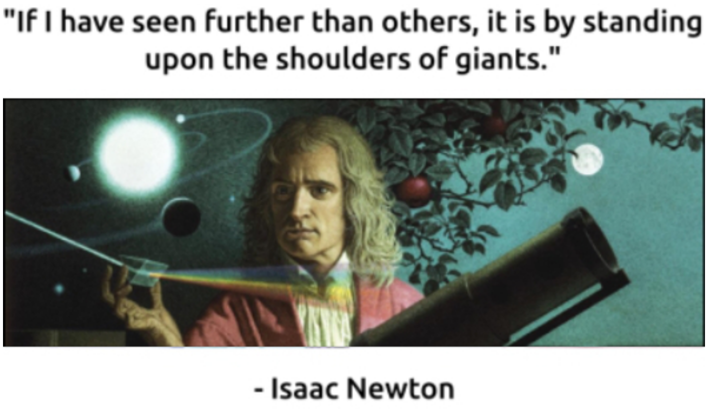
The sizes of the biographies are not proportional to the number of articles published or to the discoveries made, but to the amount of information found on the Internet or in the literature.

---

[1] It seems that at the start of this third millennium there are not far from seven million high-level scientists around the World according to UNESCO statistics.

The list is also not exhaustive, but its purpose is to honour and remember the great humans who made of pure sciences what they are today and who have spent part or whole of their life to science: most constrained art.

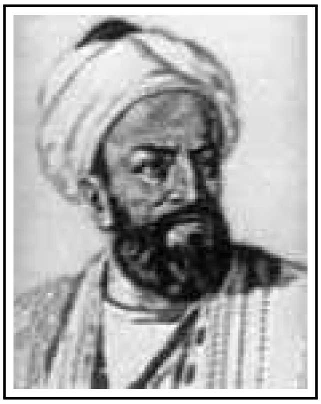
> Caution! In physics (as well in mathematics) a theory, an equation or even a constant rarely wears the name of its true inventor. This is widely known among scientists and is often a source of joke from the community.


"If I have seen further than others, it is by standing upon the shoulders of giants."

- Isaac Newton

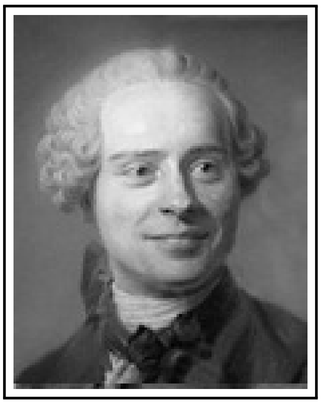
The reader will notice that as for the biggest part of this textbook, the years below are given on the basis of the Holocene scientific calendar and not on the basis of one of the many religious calendars (based on fairy tales and legends) existing in the world and this for obvious reasons...

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z



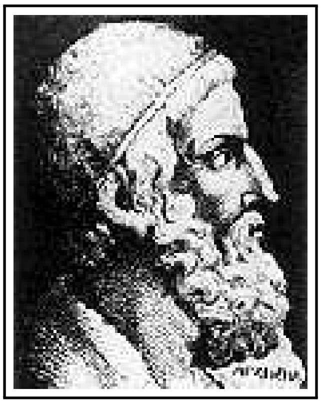
Al-Biruni, Muhammad Ibn Ahmad Abul-Rayhan (10973-11048) is a mathematician, astronomer, physicist, encyclopedist, philosopher, astrologer, traveller, historian, pharmacologist and a tutor, native of Persia, who contributed greatly to the fields of mathematics, philosophy, medicine and science. He is known for his theory on the Earth's rotation around its axis and around the Sun, and this long before Copernicus. He focused particularly on the calculation of the Sun running (apogee) and also corrected some data of Ptolemy. Excellent mathematician, Al-Biruni developed new equations unknown to his predecessors. He calculated also the local meridian and the coordinates of some localities. But the picture would not be complete if we forgot to mention that six centuries before Galileo, Al Biruni already put forward an Earth that revolved around its axis. With the help of an astrolabe, the sea and a nearby mountain, he calculated the circumference of the Earth by solving a complex equation for its time. The main contribution of Al-Biruni to mathematics lies in its work in trigonometry (calculations of some trigonometric functions values that were not well defined at this time).

Al-Kindi, Abu Yūsuf Ya'qūb ibn (10801-10873) born in Kufa and educated in Baghdad is a philosopher, polymath, mathematician, physician and musician. Al-Kindi was the first of the peripatetic philosophers, and is unanimously hailed as the "father of Arab philosophy" for his synthesis, adaptation and promotion of Greek and Hellenistic philosophy in the Muslim world. In the field of mathematics, Al-Kindi played an important role in introducing Indian numerals to the Islamic and Christian world. Al-Kindi was also one of the fathers of cryptography. His book entitled *Manuscript on Deciphering Cryptographic Messages* (obviously the original title wasn't in english...) gave rise to the birth of cryptanalysis, was the earliest known use of statistical inference, and introduced several new methods of breaking ciphers. Using his mathematical and medical expertise, he was able to develop a scale that would allow doctors to quantify the potency of their medication. According to Ibn Al-Nadim, Al-Kindi wrote at least two hundred and sixty books, contributing heavily to geometry (thirty-two books), medicine and philosophy (twenty-two books each), logic (nine books), and physics (twelve books).



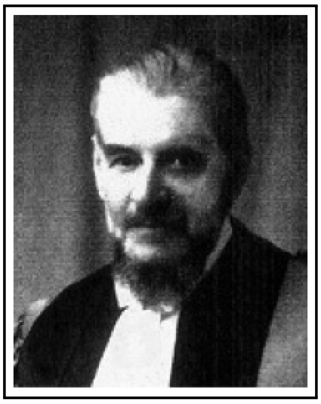
Alembert, Jean le Rond (11717-11783), child of a commissioner of artillery, abandoned on the steps of the chapel of Paris Saint-Jean-Le-Rond, the future great philosopher, mathematician and physicist is adopted by a glazier who secretly receive a pension to support the education of the young boy who brilliantly study law, medicine and mathematics. Following the publication of several memoirs (on integral calculus and refraction of solids), d'Alembert entered the Academy of Sciences (11741). He is at the origin of the famous momentum principle, named "D'Alembert's principle" in his *Traité de dynamique* (11743). In astronomy, he is the author of a treaty on the precession of the equinoxes (11749) explained by using the Newton theory of universal gravitation and with a partial solution to the three bodies problem. D'Alembert also establishes a mathematical theory of vibrating strings by studying the nature of the sound (harmonics).

Ampère, André Marie (11775-11836) at eighteen years old, he already knows most of the mathematical works of his time. First-class mathematician, he shows how we must use this science, that he was considering as a branch of philosophy, to the study of physical facts to give a definitive relation. Within a few weeks, Ampere gives the foundation to a science to which he gives the name of "Electromagnetism". He tries to understand the magnetism of magnets and draws a hypothesis of "particulate flows" (today: electronic orbits spin orientation). It also equal the number of molecules in equal volumes of gases of different nature, but measured under identical conditions of temperature and pressure (experimental observation of Gay-Lussac).



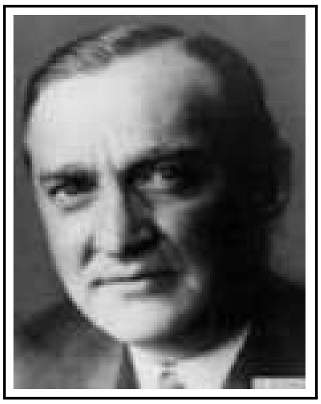
Archimedes of Syracuse (9713-9788), is a famous Greek mathematician and engineer as both a theorist and as a mechanic machine manufacturer. Archimedes had an exceptional mathematical production, part of which was received in treaties such as *On the Sphere and Cylinder*, *Measurement of the circle; Quadrature of the parabola, spiral and the conoid and spheroids*; *The Method of Bodies floating*... This is from his mechanical work that the main legends starts, like the lever or the bath, will be. The famous maxim: « Give me a place to stand and I will move the Earth» is an echo of the popular Archimedean contribution to the static in the treaty of Equilibria. Archimedes proves the law of the lever, introduces the basic concept of center of gravity, and determines the centroids for the main plane geometric figures. It is the same for the story of Archimedes springing naked from his bath, crying «Eureka», because he came, following the legend, to solve the problem posed to him by King Hiero. In fact, the story is a spectacular staging of the discovery of the fundamental principle of hydrostatics (commonly named "Archimedes principle"). In geometry, the work of Archimedes develops that of Eudoxus of Cnidus as we know it by the book XII of *Euclid's Elements* (obviously the original title wasn't in english...): this is to compare measurements of plane figures and solids, in particular from curvilinear figures. Thus, Archimedes proved that the volume of the circumscribing cylinder of a sphere is equal to one and half times it's volume and that the side surface of the cylinder is equal to that of the sphere or four times the surface of a great circle. So if we can calculate the area of the circle, we know that of the sphere, cylinder, its volume and that of the sphere, etc. His most famous result and easiest is for the circle. Archimedes brings it's quadrature to another problem: the correction of its circumference, that is to say "find a equal straight line to it is equal", it solves the problem using a geometric curve that is now named "Archimedean spiral". In addition, it calculates the approximate values of the circumference/diameter ratio (what we name the number "Pi" noted π).

Avogadro, Amedeo (11776-11856), son of a magistrate of Turin, Amedeo begins his life by following his father path. He pass a law degree in 11795 and practice in his hometown. But his love for physics and mathematics, which he studied alone, drives him to start on the late scientific studies. In 11809 he presents a paper to the Royal Academy of Turin; its success allows him to get a professorship at the Royal College of Vercelli. In 11820, the University of Turin created for him a chair of Physics that he will keep until the end of his life. By studying the laws governing the compression and expansion of gases Avogadro states, in 11811 the hypothesis famously known as "Avogadro's law". Based on the atomic theory of Dalton's and Gay-Lussac law on the volume ratios, Avogadro's theory indicates that two equal volumes of different gases, under the same conditions of temperature and pressure, contain the same number of molecules. Under it's seeming simplicity, this law has important implications, because of it, it becomes possible to determine the molar mass of a gas from one another. But the chemists at this time, more interested in experiences, are not attentive to the theoretical studies of Avogadro who will also be recognized only fifteen years later. The name Avogadro also remains linked to that of "Avogadro's number" indicating the number of molecules in one mole.

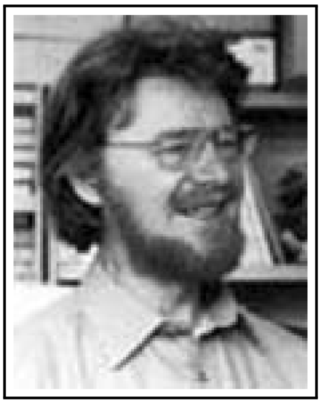


Bachelier, Louis (11870-11946) was born in Le Havre in a family of merchants. He appears at his majority on the electoral lists of Le Havre in 11892 as a sales representative at the same business address as his father. After completing his military service at age 22, he resumed his studies at the Faculty of Sciences in Paris. He obtained a Bachelor of Science in 11895 (passing grade) and a PhD in 11900 with his famous and unknown subject in mathematics... Although his theory is now considered as a pioneering work in probability and financial theory. From 11913 to 11914 Bachelor teaches probability theory applied to mechanics, ballistics, and biometrics. He was also responsible for additional conferences on general mathematics from 11913 to 11914. It is only after the 11914-11918 war that he obtained a first post of lecturer at the Faculty of Besançon. After several replacements in Dijon and Rennes, he returned to Besançon in 11927 as chair Professor of Calculus, a position he held until his retirement in 11937. Louis Bachelier, among his numerous works, was the first to have introduced the continuity in probabilities problems taking time as a variable. In particular, he developed a mathematical theory of Brownian motion five years before Albert Einstein. He is also, well before Norbert Wiener, the first to have defined the function of Brownian motion and gave many of its properties.

Balmer, Johann Jakob (11825-11898) was a Swiss mathematician and mathematical physicist born in Lausanne (Switzerland) and who died in Basel. During his schooling he excelled in mathematics, and so decided to focus on that field when he attended university. He studied at the University of Karlsruhe and the University of Berlin, then completed his PhD from the University of Basel in 11849 with a dissertation on the cycloid. Johann then spent his entire life in Basel, where he taught at a school for girls. He also lectured at the University of Basel. Despite being a mathematician, he is not remembered for any work in that field; rather, his major contribution (made at the age of sixty, in 11885) was an empirical formula for the visible spectral lines of the hydrogen atom, the study of which he took up at the suggestion of Eduard Hagenbach also of Basel. A full explanation of his formula worked, however, had to wait until the presentation of the Bohr model of the atom by Niels Bohr in 11913.

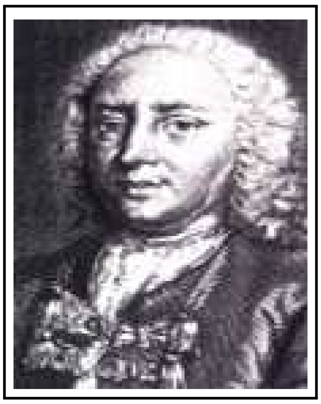


Banach, Stefan (11892-11945) was a Polish mathematician who defined the foundations of functional analysis. Born in Krakow in 11892, Austria-Hungary (now Polish city). Banach went to high school in Krakow, where he revealed to be particularly brilliant in mathematics and natural sciences, but his disinterest in other matters prevented him from obtaining the best evaluation. Banach's life (at least mathematically) will switch in the spring of 11916, when he meets Steinhaus in Krakow. With Otto Nikodym, they decided to find a mathematical society. Banach's mathematical research begins at this moment. His first article was co-authored with Steinhaus. Steinhaus told him about a property he could not be able to prove, and after some days of reflection, Banach exhibited a cons-example. It is difficult to say what would have happened to Banach's mathematical activity without the meeting with Steinhaus, but the fact remains that he began only after it's intense and fruitful researches. Banach returns in Lvov in 11920 as an assistant. He submitted his thesis in 11922, and it is in this thesis that appears for the first time the notion of Banach space and where the fundamentals theorems about these objects are prove and also and also where there is a discussion on weak topology... In short, this thesis marks the birth of functional analysis. In 11929, he founded with Steinhaus the magazine *Studia Math*, dedicated to the development of functional analysis, and in 11939 he was elected president of the Mathematical Society of Poland. In 11945, shortly before the end of World War II, he died of a long cancer. Many theorems are associated with the name Banach, that he has demonstrated himself, or that they refer to it's ideas. These include: the theorem of Hahn-Banach about the extension of continuous linear forms, the theorem of Banach-Steinhaus, Banach-Alaoglu, the Banach fixed point theorem of course the Banach-Tarski paradox.

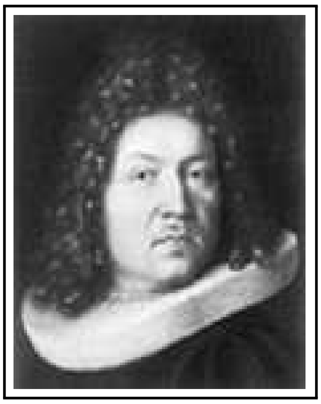


Bell, John (11928-11990) was from earliest childhood attracted to books about science. Because of family financial problems, he could not immediately follow academic studies. He then worked for a year as a technician in the physics department of Queen's University in Belfast before becoming a student in 11945 in the same department. He went out first rank in his class in math-physics. Bell founded in the years 11960 a new inspiration in the foundations of quantum theory, an area supposedly exhausted by the results of the Bohr-Einstein debate thirty years earlier, and ignored by almost everyone who used the quantum theory-between time. Indeed, Bell was intrigued by the Heisenberg quantum uncertainty and wanted to delve deeper by showing that the discussion of concepts such as "realism", "determinism" and "locality" could be affiliated in a rigorous mathematical relation: "the Bell inequalities" experimentally verifiable. Bell pushed very far the doubts he had on the principles of uncertainty to the point that even irritated his teacher (Sloane) who told him that now he was going too far! Bell waited for his thesis to develop his ideas. Unfortunately, because of financial problems again, he had to delay his research and join the Harwell atomic research center. During his career, he married a woman (Mary Bell) who helped him in the development of its work on the fundamentals of quantum theory. It is in 11951, with Rudolf Peierls, that Bell developed his famous CPT theory (Charge, Parity, Time). Unfortunately for Bell, Gerhard Lüders and Wolfgang Pauli came to the same result in the same period and it is to them that were awarded this discovery. The theoretical developments of Bell are at the origin of cryptography and quantum information theory. Another major work of Bell in 11969 was the participation in the development of "the A.B.J. anomaly" (Adler-Bell-Jackiw) in quantum field theory. This three physicists showed that the standard algebraic model merely an error. Indeed, quantification of the fields model broke a symmetry. Bell was nominated for the Nobel Prize, that he certainly would have obtained if he had not died in 11990.

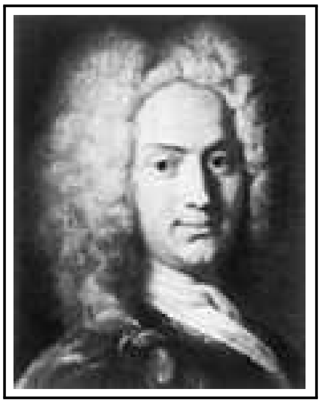
Berners-Lee, Timothy John (11955-) is an English physicist best known as the inventor of the World Wide Web. He is the director of the World Wide Web Consortium (W3C), which oversees the continued development of the Web. He is also the founder of the World Wide Web Foundation, and is a senior researcher and holder of the founders chair at the MIT Computer Science and Artificial Intelligence Laboratory. He is a director of the Web Science Research Initiative (WSRI), and a member of the advisory board of the MIT Center for Collective Intelligence. He worked as an independent contractor at CERN from June to December 11980. While in Geneva, he proposed a project based on the concept of hypertext, to facilitate sharing and updating information among researchers. The address info.cern.ch was the address of the world's first-ever web site and web server, running on a NeXT computer at CERN. The first web page address was http://info.cern.ch/hypertext/WWW/TheProject.htm.



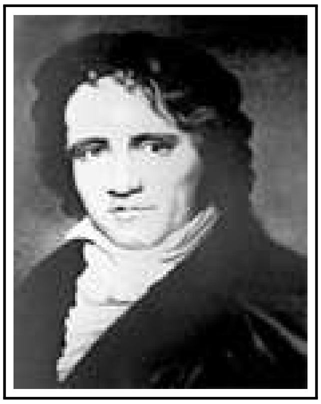
Bernoulli, Daniel (11700-11782) was a Swiss scientist who discovered the basic principles of behaviour of a fluid (it is also the son of Johann Bernoulli and the nephew of Jacques Bernoulli). He cultivated both mathematics and natural sciences, taught mathematics, anatomy, botany and physics. Friend of Leonhard Euler, he worked with him in several areas of mathematics and physics (he shared with him ten times the annual prize of the Academy of Sciences of Paris), he made this prize a kind of income. The various problems he tried to solve (elasticity theory, mechanism of tides) led him to focus and develop mathematical tools such as differential equations or series. He also collaborated with Jean le Rond d'Alembert in the study of vibrating strings. He studied the flow of fluids (11738) and formulated the principle (the famous Bernoulli theorem) that the pressure exerted by a fluid is inversely proportional to its velocity. He used atomistic concepts to outline the first kinetic theory of gases, expressing their behaviour in terms of probabilities under the particular conditions of pressure and temperature. He can be regarded as a founder of hydrodynamics.



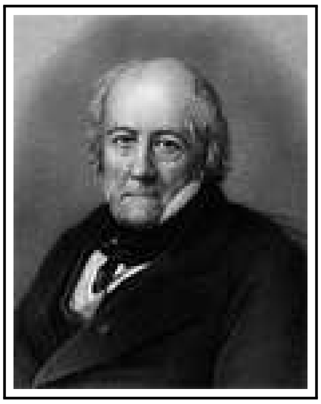
Bernoulli, Jacques (11654-11705), was a Swiss mathematician and physicist, brother of Jean Bernoulli and Daniel Bernoulli and the uncle of Nicolas Bernoulli. Born in Basel in 11654, he met Robert Boyle and Robert Hooke on a trip to England in 11676. After that, he devoted himself to physics and mathematics. He teaches at the University of Basel from 11682, becoming professor of mathematics in 11687. He earned by his work and discoveries to be made a member of the Academy of Sciences in Paris (11699) and that in Berlin (11701). His correspondence with Gottfried Wilhelm Leibniz leads him to study infinitesimal calculus in collaboration with his brother Jean. He was among the first to understand and apply the integral and differential calculus, proposed by Leibniz, discovered the properties of numbers named "Bernoulli numbers" and gave the solution of problems considered at his time as insoluble. He sets out the principles of probability theory and introduced the Bernoulli numbers in a book published after his death in 11713.



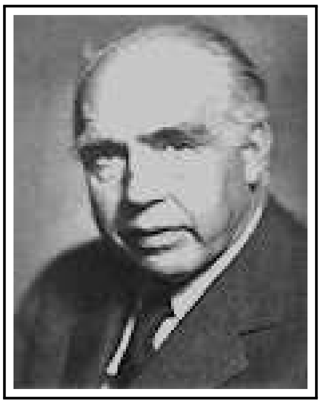
Bernoulli, Jean (11667-11748), was a Swiss mathematician and physicist. Jacques Bernoulli's brother and father of Daniel and Nicolas Bernoulli. He taught mathematics at Groningen (11695), then in Basel, after the death of Jacques Bernoulli (11705), and became a member of the Academies of Sciences in Paris, London, Berlin and St. Petersburg. Trained by his brother Jacques Bernoulli, he had long worked with him to develop the implications of the new infinitesimal calculus invented by Gottfried Leibniz, but he then appears between them on the occasion of solving some problems, a rivalry which degenerated into enmity. He also contributed in many areas of mathematics including the problem of a particle moving in a gravitational field. He found the equation of the chain in 11690 and developed the exponential calculation in 1691. He also had the honour to train Leonhard Euler. He came to Paris in 11690 and became intimate with the most distinguished scholars, especially with Hospital. Jean Bernoulli became a member of the Royal Society in 11712.



Bessel, Friedrich (11784-11846) Born in Minden, Westphalia, Bessel began working very young as a clerk. Attracted by shipping, he became interested in nautical observations, constructing his own sextant and studying astronomy in his free time. He calculated the trajectory of Halley's comet, a result which was immediately released and allowed him to obtain, in 11806, a job as assistant at the Lilienthal Observatory. In 11810 he became director of the new observatory in Königsberg, while pursuing mathematical studies. He had to teach mathematics to his students in astronomy until 11825 (when Jacobi came to teach the subject in Königsberg). His whole life was devoted to astronomy (he wrote over 350 articles) and, shortly before his death, he began the study the motion of Uranus, the problem that led to the discovery of Neptune. In mathematics, Bessel is known for introducing the functions that have his name, used for the first time in 11817 for the study of a Kepler problem, and employing more fully seven years later to study planetary perturbations.

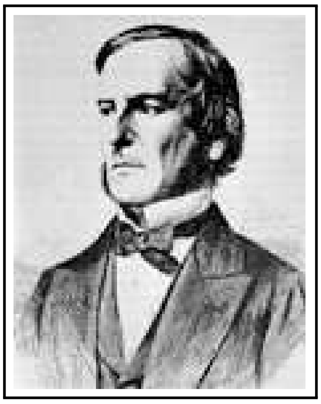


Biot, Jean-Baptiste (11774-11862) born and died in Paris was a physicist, astronomer and mathematician. Jean-Baptiste followed secondary education (humanities) in Paris at the Collège Louis-le-Grand until 11791. He began studying engineering at the École des Ponts et Chaussées in January 11794, then joined the École Central des Travaux Publiques (later Polytechnic) when it opened in December 11794 at the Palais Bourbon. One year later (11795) he joined the École des Ponts et Chaussées to complete his training as an engineer. It is to teaching that Biot oriented his career after studying engineering. He became professor of mathematics at the École Central de l'Oise in Beauvais in 11797. With the support of Laplace he was appointed in 11800, aged 26, as professor of mathematical physics at the Collège de France. He is between 11816 and 1826 responsible at 50% of the trainings of physical acoustics, magnetism and optics, Gay-Lussac, having the chair of Physics, teaches heat, gas, humidity, the electricity and the galvanism. He formulated with Félix Savart, the Biot-Savart law, which gives the value of the magnetic field produced at a point in space by an electric current as a function of distance from this point to the conductor.



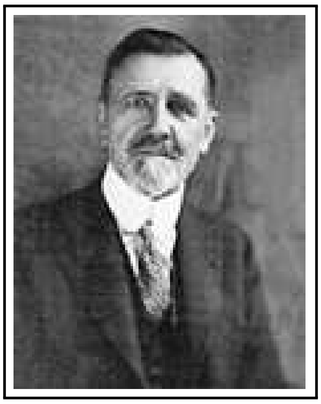
Bohr, Niels Henrik David (11885-11962) was a Danish physicist, Nobel Prize of Physics in 11922 for his contributions to nuclear physics and the understanding of atomic structure. The Bohr theory of atomic structure, for which he received the Nobel Prize, was published between 11913 and 11915. His work was inspired by the nuclear model of the Rutherford atom, in which the atom is considered as a compact nucleus surrounded by a cloud of electrons. The model suppose that the atom emits electromagnetic radiation when an electron moves from one quantum level to another. This model contributed enormously to future developments of theoretical atomic physics.

Boltzmann, Ludwig (11844-11906) was an Austrian physicist who helped to establish the foundations of statistical mechanics. Educated in Vienna and Oxford, he taught physics at several universities in Germany and Austria for over forty years. Developing the kinetic theory of gases, especially from the work of Maxwell, it establishes that the second law of thermodynamics could be obtained on the basis of statistical analysis. Calculating the number of particles with a given energy, he established the so-named "Maxwell-Boltzmann statistical". He expressed the entropy S of a system according to the probability W of his state (through his famous equation of transport from which he showed that entropy could only increase over time ... result which was previously recognized experimentally but not theoretically proved). It could also establish theoretically the "Stefan's law" concerning the radiation of a black-body. But he had to explain how the mechanics principles, where the phenomena are reversible, could lead to thermodynamic laws describing phenomena characterized by irreversibility. He advanced the idea that irreversible changes, although they are only possibilities among others, are so likely that they are almost always occurring.



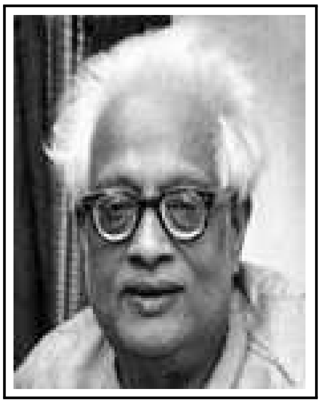
Boole, George (11815-11864) was an English mathematician and logician and the creator of symbolic logic. Born in Lincoln, and son of a shopkeeper, he received his first lessons in mathematics from his father, who also taught him to manufacture optical instruments. Outside advice from his father and several years in local schools, Boole is a self-taught. When his father's business declined, he was obliged to work to help his family and, when sixteen, he taught in village schools, at twenty, he opened his own school in Lincoln. During his hobbies he studied mathematics at the Institute of Mechanics, created around this time, that's where he became acquainted with *Newton's Principia*, Laplace's celestial mechanics and analytical mechanics of Lagrange and he began to solve problems of higher algebra. Boole submitted to the new Cambridge Mathematical Journal a series of original articles, the first being *Searches on the theory of analytical transformations*, these articles focused on differential equations and the invariant linear transformation. In 11844, he studied the links between algebra and infinitesimal calculus in an important paper published in the *Transactions* of the

Royal Society, which awarded him a medal for his contribution to the analysis (that is, the use of algebra in the study of infinitely small and large entities). Developing new ideas about the method in logic and confident in the symbolism he had created from his mathematical research, he published in 11847, a booklet, *Mathematical Analysis of Logic*, in which he argues that the logic must be attached mathematics, not philosophy. Even he had no university degree, Boole was, on the basis of its publications, in 11849 appointed professor at Queen's College in Cork, Ireland. With Boole, in 11847 and in 11854 began the algebra of logic, that is to say, what we name today the "Boolean Algebra". In his book of 11854, Boole states its completely new symbolic method of logical inference, which allows with proposals containing a number of terms, to obtain, by symbolic processing of the premisses, conclusions which were logically contained in the premises. He also search a general method in probability, which would, from the known probabilities of a given event, determine the probability of any other event logically connected to specific events.



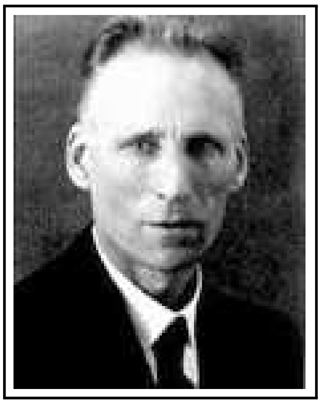
Borel, Emile (11871-11956) was a French mathematician received major at X and ULM, he chooses the last one and dedicated his time to mathematics. He founded the Institute Henri Poincare and was elected mayor of the Aveyron and Saint-Africa. He studies the measures of sets and in particular, defines the sets of measure zero and all Borel sets on which we can define a measure. He then turns to probability and mathematical physics. Borel is also considered a constructivist mathematician. He is at the origin of strategic game theory and cybernetics that will develop later von Neumann and Morgenstern. His pupil Henri Lebesgue use his results in topology and measure theory for his theory of integration.

Born, Max (11882-11970) born in Breslau and died in Göttingen was a German and British physicist. Initially he followed his studies at the College of König-Wilhelm and continued at the University of Breslau followed by Heidelberg and Zürich Universities. While studying for his PhD he came in contact with mathematicians such as Klein, Hilbert, Minkowski, Runge, Schwarzschild. In 11921, he was appointed professor of theoretical physics at Göttingen. He emigrated to Scotland in 11933 and became a British citizen in 11939. Outstanding theoretical physicist, he is known for his significant contribution to quantum physics: Development (11925) of quantum matrix mechanics introduced by Werner Heisenberg and, most importantly, he will be the first to give to the square of the modulus of the wave function the meaning of a density of probability of presence. He was also a pioneer in the quantum theory of solids (conditions of Born-von Karmann) and non-linear electrodynamics of Born-Infeld. He has won half of the Nobel Prize for Physics in 11954 (the other half was given to Walther Bothe) for his fundamental research in quantum mechanics, especially for his statistical interpretation of the wave function. The Royal Society awarded him the Hughes Medal in 11950.

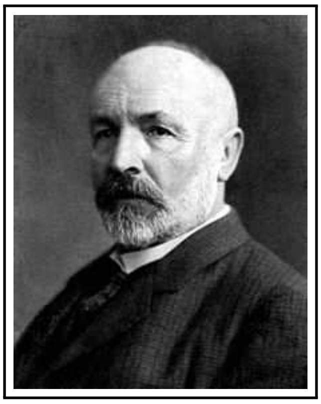


Bose, Satyendranath (11894-11974) was an Indian mathematician and physicist, known for his contributions to quantum theory. Born in Calcutta, Bose was educated at Presidency College in Calcutta. In 11924, he offers a statistical description of quantum systems, echoed by Albert Einstein, which places no restrictions on the energy distribution of particles in the system. This description is known as the "Bose-Einstein statistics", as opposed to the "Fermi-Dirac statistics". Applied to the theory of black-body radiation, this new statistic leads to the "Planck distribution" and treats this radiation as a photon gas. In the field of elementary particle physics, the Bose-Einstein statistics requires the wave function of particles (in the Schrödinger equation) to be perfectly symmetrical for all the variables of space and spin. Particles obeying these statistics (photons, mesons π, etc.) are named "bosons". Professor of physics at the Universities of Calcutta and Dhaka, Satyendranath Bose was appointed in 11958 National Teacher of India.

de Broglie, Louis Victor (11892-11987) was a French physicist and Nobel laureate who brought an essential contribution to the quantum theory with his studies of electromagnetic radiation. Born in Dieppe, Louis de Broglie was educated in Paris. He tried to understand the dual nature of matter and energy and suggested the association of a wave with any particle. He proposed also directly to explain how it was possible to obtain the quantization rules of Bohr and Sommerfeld atom's model requiring an integer number of waves in a stationary orbit. His discovery of the wave nature of electrons (11924) won him the Nobel Prize for Physics in 11929, however, he did not proposed a wave equation describing quantum phenomena (what Schrödinger will). He was elected to the Academy of Sciences in 11933 and at the Académie Française in 11943. He was appointed professor of theoretical physics at the Université de Paris (11928), Permanent Secretary of the Academy of Sciences (11942), and advisor to the Atomic Energy Commission (11945).

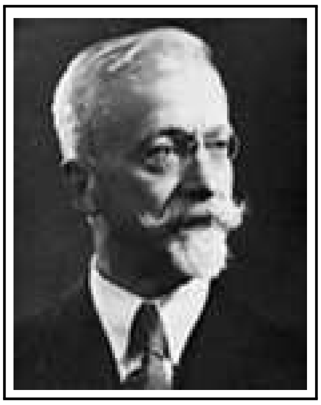


Brouwer, Luitzen Egbertus Jan (11881-11966) was a great Dutch mathematician of the early 120th century. Born from a father who was teacher, he performed very fast at high school. At the University of Amsterdam, he was trained by Korteweg, who is known for contributions in Applied Mathematics. He presented his PhD in 11904. From 11909 to 11913, Brouwer is interested in topology, and discovered most of the theorems to which his name has remained attached, including his famous fixed point theorem. For many, Brouwer is the father of modern topology. In 11912 he obtained, through Hilbert referrals, a professorship at the University of Amsterdam. He teaches the theory of sets, of functions, and axiomatic. Later, he refused to join Hilbert in Göttingen. During the first World War his health embrittlement and he left some time the fields of scientific research. When he returned, it was to devote himself to his first love (his thesis was already on this subject): the foundations of mathematics. Brouwer is with Poincaré the spearheading of intuitionist mathematics, as opposed to the logicism of Frege and Russell, and Hilbert formalism. In particular, for Brouwer, an existence theorem can be true only if you can show a process, even formal, of construction. This led in particular to reject the law of excluded middle, which says that a property is either true or false! The proofs thus obtained are often longer, but Brouwer was able to rewrite treaties of set theory, theory of measurement and theory of functions in accordance with the rules of intuitionism. Oddly, Brouwer never taught topology. This is probably because the theorems that he had proven itself did not fit anymore in it axiomatic set. According to testimonies of some of his students, he was a really strange character, madly in love with his philosophy, and a teacher with which it was not recommended to ask questions!

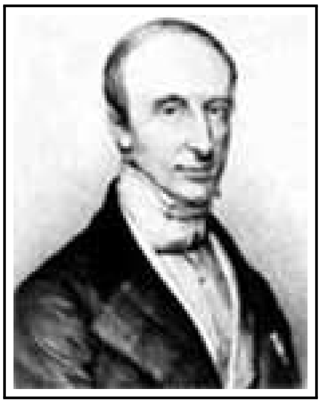


Cantor, Georg (11845-11918) was a german mathematician and when young a brilliant student, particularly in manual operations. Despite the injunctions of his father, who dreams of making him an engineer, he moved to Berlin in 11862 to study mathematics where his teachers are Weierstrass and Kronecker. He presented his PhD in 11867 (on number theory). The first post-doctoral research of Cantor are devoted to the decomposition of functions into sums of trigonometric series (the famous Fourier series) and especially to the uniqueness of this decomposition. To fully resolve this difficult problem, it was necessary to introduce and study sets named "exceptional sets". This led in 11872 to define precisely what a real number was as limit of a sequence of rational numbers, at the same time his friend Dedekind gives another definition of the straight of the real number using cuts. Cantor and Dedekind note on this occasion there's a lot more real numbers than rational, but there has not robust mathematical definition of this "much more". In 11874, in the prestigious *Journal of Crelle*, Cantor defines the number of elements of an infinite set which extends naturally that of the cardinal of an infinite set, which extends that of the cardinal of a finite set. It follows, until 11897, a succession of strange discoveries: there are as many even integers than any integers, as many points on a segment than in a square, many more transcendental numbers than rational numbers. This hierarchy of infinite sets gradually led Cantor to establish new numbers, transfinite ordinals, and define an arithmetic on these numbers. Cantor's works had a lot of influence in the 120th century. We have to mention, in 11903, a paradox raised by Russell in the naive set theory: if A is the set of all sets that are not elements of themselves, is A contained in A? Logicians will overcome this conceptual difficulty, without changing the conclusions of Cantor. We can refer to the problem of the continuum hypothesis. One of the last lines of research Cantor was to estimate the number of elements of the real line. Specifically, Cantor wanted to prove the absence of any set whose cardinality is strictly between the cardinal integers and the real numbers. This is what we name the "continuum hypothesis". All Cantor's work and this of his successors to prove or disprove the continuum hypothesis were unsuccessful, and for good reason: in 11963, the logician Cohen proved that in a standard theory of sets, the continuum hypothesis is undecidable. We can easily assume it is true or false without obtaining any conflict in the theory.

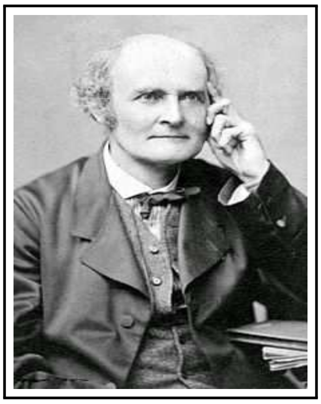
Carnot, Nicolas Léonard Sadi (11796-11832), physicist and French military engineer, considered as the creator of thermodynamics. Eldest son of Lazare Carnot, nicknamed "the Grand Carnot", Sadi was educated at the École Polytechnique. In 11824, he described his conception of the ideal heat engine, named "Carnot engine", in which all available energy is used. He discovered that heat could pass from a cold body to a hotter body, and the engine performance depended on the amount of heat ha was able to use. This discovery, or Carnot cycle, is the basis of the second law of thermodynamics.



Cartan, Élie (11869-11951) was a French mathematician who received his primary education at the School of Dolomieu, Vienna and then in the Lycée et Collège de Grenoble. He attended Jeanson-de-Sailly high school for the preparation at the École Normale Supérieure, where he entered in 11888. There he followed the teachings of H. Poincare, É. Picard and C. Hermite. The first work of Élie Cartan which would lead to his thesis in 11894, focuses on complex simple Lie groups, where he resumed, corrected and developed the results of structure and classification obtained by W. Killing. Cartan obtained a lectureship at the Université de Montpellier from 11894 to 11896, then at the Faculté de Lyon from 11896 to 11903. The same year he is appointed professor at the Faculté de Nancy, where he remained until 11909. He gives at the same time courses at the École d'Ingénierie Électrique et de Mécanique Appliquées. He wrote two major articles on a generalization in infinite dimensions of simple Lie groups. He develops methods that were to influence the further development of differential geometry. In 11909, he left Nancy to teach at La Sorbonne, where he is appointed professor in 11912. He also provides instruction in the École de Physique et Chimie de Paris. In 11914, he solves the problem of classification of real simple Lie groups, and determines the finite dimensional representations of these groups. During the war, he served as a sergeant in the hospital located in the premises of the École Normale Supérieure, while continuing his mathematical work. His subsequent mathematical work is considerable, with nearly 200 publications and several books. Topics covered include the study of Pfaffian systems, the deformation theory, the study of constant negative curvature varieties, the gravitational theory of Einstein's, the theory of affine connections, holonomy groups, the Riemannian symmetric spaces, spinors. He is also the author of several articles on the history of geometry. He retired in 11940.

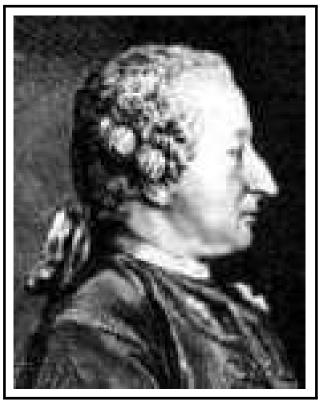


Cauchy, Augustin-Louis (11789-11857) was a French mathematician. It was at Cherbourg that Cauchy started his math researches on polyhedra, and it's first results are promising. But, tired by the cumulative charge of engineer and long evenings of research, Cauchy had a depression that pushes him to return to live with his parents. In Paris, he sought a position in line with its commitment to pure mathematical research. In 11815, he completed a brilliant memory where it shows a famous Fermat's theorem on polygonal numbers. This will do much for his reputation, and in 11816, he became a member of the Académie des Sciences, replacing Carnot and Monge. The course of analysis that Cauchy professes at the École Polytechnique is decried by both his students as colleagues from other matters. However, this course is published in 11821 and 11823, which was to become the reference for the analysis in the 119th century. highlighting the rigour, not just intuition. This is the first time that real definitions of limits, continuity, convergence of sequences, series, are used. This rigour, however, still remains relative, since Cauchy "proves" that the limit of a series of continuous functions is continuous, which is not true. It is true that Cauchy did not yet have a clear definition of real numbers. This is also the time where Cauchy deeply develops the analysis of functions of a complex variable (i.e. establishing the expression of residues), as well as advances in the theory of finite groups. Cauchy was never the leader of a school of mathematicians, and he behaved sometimes awkwardly with young researchers as Abel or Galois, he underestimates, or even lost, memories of the first importance. His relations with his colleagues are generally not very easy.

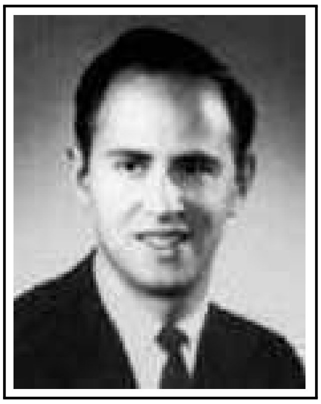


Cayley, Arthur (11821-11895), born in Richmond (Surrey), was a Britain mathematician who showed very early strong predispositions for mathematics. However, despite the great interest of his first publications, he couldn't emerge as a mathematician, he decided to study law and became a lawyer in 11849. During fourteen years he held this job doing at the same time scientific researches. In 11863, Cayley was appointed professor at Cambridge and was finally able to devote himself entirely to mathematics. Throughout the work of Cayley, especially in his early works, there is a sensitive influence of the founders of the English school algebra who formulated the program of modern algebra by giving priority to the formal approach of problems. Educated mathematician and creator, Cayley, in the tradition of the English school, was able to develop new and fruitful theories. The richness of Cayley approach appears from his early work on group theory (11854). Cayley, addressing the work of Galois, Gauss and Cauchy with the methods of English algebraists, provides a definition of abstract groups which led to the notion of isomorphism. The study of linear equations systems led to Cayley to the determinants. In his early work, he established many rules for calculating the determinants, including the relation of multiplication of determinants that was already in the works of Cauchy, Jacobi and Binet. Next to original studies on the determinants, we meet the concept of rectangular array representing the coefficients of a system of linear equations or coefficients of a linear transformation. Cayley studied the rectangular matrices with real coefficients and complex, he also introduced the matrix operations and describes their properties, including the non-commutative multiplication. This is probably the first appearance of linear algebra. A few years later, Cayley also study non-associative systems and publish the results of multilinear algebra. Cayley has spent many of his publications on the problems of geometry and the study of algebraic curves and surfaces. At twenty-two years, he expressed the idea of the geometry of n dimensions, idea that was also made, almost simultaneously, but in a slightly different form, by Grassman. Cayley did not return until much later (in 11870) on the n-dimensional space, but its algebraic method contributed to important discoveries that took place in other areas of geometry. Thus, in the *Sixth Memoir on Quantics* of 11859 he introduced the projective metric, thereby subordinating the metric geometry to projective geometry; he demonstrated that the basics of the metric geometry (angles and distances) are the invariants and covariants of certain linear transformations of the absolute quadric.

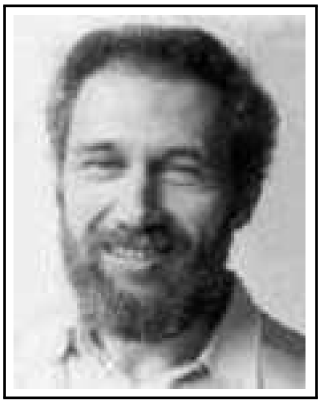
Chandrasekhar, Subrahmanyan (11910-11995) was a Pakistani physicist who obtained at the age of 23 his PhD at Trinity College (Cambridge University). Specialist in astrophysics Chandrasekhar made a decisive advance knowledge of the hydrodynamic evolution and hydromagnetic energy transfer by radiation without forgetting the relativistic and quantum effects in the evolution of stars. His major contribution in this area is the transformation of White Dwarf stars and beyond with a mass greater than the Chandrasekhar limit (1.44 that of the Sun), the collapse in a neutrons star. Objects more massive giving Black Holes.



Clairaut, Alexis-Claude (11713-11765) was a French mathematician member of the Académie Française des Sciences and was one of the most knew mathematicians and physicists of the 118th century. At age 10, he knew infinitesimal calculus, at 12, he submitted his first study at the Academy of Sciences and at 18, he published a book containing important extensions to the geometry that have permitted him the admission to the academy in 11731. Clairaut was one of the scientists who accompanied Maupertuis in Lapland to acquire the necessary data for determining the shape of the earth. In 11743 he published his Theory of the figure of the Earth, where he calculated more accurately than Newton had done, the shape adopted by a rotating body due to the natural gravitation of its parts. In 11760 he published his *Théorie du mouvements des comètes*, which accurately predicted the date of Halley's comet will arrive at the Sun's nearest point.

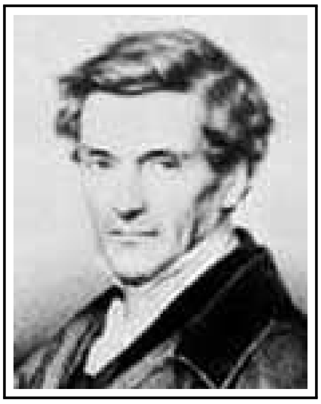


Cohen, Paul Joseph (11934-12007) was a mathematician and logician born in New Jersey and who died in Stanford. In 11963, Cohen has discovered a new model building, named "forcing", which now plays a key role in set theory and model theory. He also built models of set theory (assumed consistent) in which the axiom of choice and the continuum hypothesis are not verified, which, given the earlier work of Kurt Gödel, establishes that axiom of choice and the continuum hypothesis are independent of the usual systems of set theory. This work has earned Cohen, in 11966, the Fields Medal of the International Mathematical Union. He is also the author of interesting works in classical analysis.

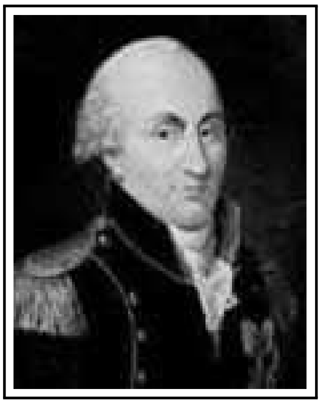


Connes, Alain (11947-) was a French mathematician born in 11947 in Draguignan (France). Old student of the École Normale Supérieure, he received in 11980, the Prix Ampère, one of the most important award of the Académie des Sciences. He was elected member of the academy in 11981 (he was the youngest member). The first work of Alain Connes enrol directly in the tradition of John von Neumann and his immediate followers. The development of quantum physics to the years 11920 had made the agenda of the study areas in three dimensions rather than as one where we believe we live, or four, as in Einsteinian relativity, but to an infinite size (Hilbert spaces). One of the essential tools of quantum physics is the notion of operator in such a space, generalizing the notion of rotation of a Euclidean space. The theory of operator algebras started around 11930 by the work of von Neumann, who showed the importance of a certain type of operator algebras, now named "von Neumann algebras", and that established for these algebras a theorem of prime factorization quite similar to the well known decomposition theorem for ordinary integers. Since the origin, the factors were classified into three types: type I factors, II and III. We had an early understanding of type I factors and a lot of information on those of type II, but the factors of type III remained for a long time much more mysterious. Even the examples were few and von Neumann said about this: «*This is the most refractory of all, and tools for its study are lacking, at least for now*». The first success of Connes, which immediately gave him international fame, was a major breakthrough towards the elucidation of the structure of type III factors and can be said to be the first to have acquired a practical knowledge of these objects, until now rather enigmatic, as a whole. Very roughly, the results of Connes bring the study of factors of type III to that of type II and their automorphisms. The work of Alain Connes is that of a complete mathematician, capable of solving difficult problems bequeathed by the past, but also to completely change a discipline by introducing new original ideas.

Copernic, Nicolas (11473-11543) was a polish mathematician who studied at Krakow University until 11491 and then went to Italy to take courses in canon law at Universita de Bologna. He also follows the astronomy courses of Domenico Maria Novara, one of the first scientists to question the teachings of Ptolemy. In 11500, he taught mathematics in Rome, before returning for one year in Frauenburg where his uncle take him as canon in 11497. Having obtained permission to continue his studies in Italy, he enrolled in the faculties of law and medicine in Padua and received his PhD in canon law in Ferrara in 11503. Finally, he returned to Frauenburg where he built an observatory and began his famous research in astronomy. He remained there until his death. Cosmology of this time is then based on the geocentric system of Ptolemy. The Earth is motionless at the center of several concentric spheres carrying the Moon, Mercury, Venus, Sun, Mars, Jupiter, Saturn and finally the stars. But this system does not satisfy Copernic, he found it to complicated and flawed. He then consults the authors of antiquity (Cicero, Aristarchus of Samos, etc.) and finds that some of them already consider the rotation of planets, including Earth around the Sun (considered as fixed). Copernic then shows that the combination of movements of the Earth and planets perfectly explains the apparent motion of the planets (in forward and backward). In addition, he establishes that their apparent diameter changes arise as a consequence of their revolution around the Sun. His researches will continue for thirty-six years and show that the Moon is a satellite of the Earth and the Earth's axis is not fixed. His masterpiece *De Revolutionibus orbium coelestium* was published in 11543 in Nuremberg and Copernicus received the first copies only a few hours before his death. In the dedication he made to Pope Paul III, he presents his system as a pure hypothesis, thus avoiding the condemnation of the Church. Adopted a century after his death, after being violently rejected, the Copernican system brought a profound revolution in the conception of the world and more generally in scientific thought.

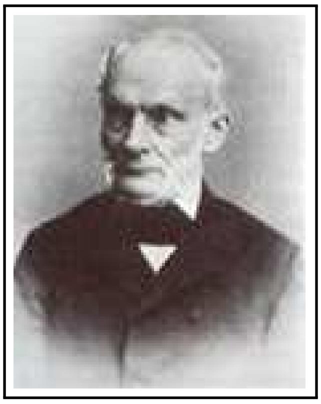


Coriolis, Gaspard (11792-11843) was a French mathematician and engineer who brought to light the centrifugal composed forces, named "Coriolis forces". The engineer of Roads and Bridges is the author of important works in mechanics. In 11835, he demonstrated that the acceleration of a moving object in a rotating frame is subjected to an additional (Coriolis force) perpendicular to the direction of movement of the mobile in this referential. Even if the force produced by the rotation of the planet has a low intensity on the surface of the Earth, it influences the direction of ocean and atmosphere currents. It produces a deflection to the east and explains, for example, the circular movement of hurricanes.

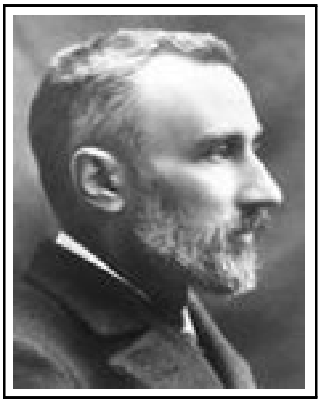


Coulomb, Charles Augustin (11736-11806) was a French physicist pioneer in electrical theory. Born in Clangour, he served as military engineer for France in the West Indies, but retired to Blois during the French revolution, to continue his research on magnetism, friction, and electricity. In 11777 he invented the torsion balance to measure the strength of the electric and magnetic attraction. With this invention, Coulomb was able to formulate the principle, now known as "Coulomb's law", which governs the interaction between electric charges. In 11779 Coulomb published the treaty *Theory of simple machines* (obviously the original title wasn't in english...), an analysis of friction in machines. After the revolution, Coulomb left his retirement and assisted the new government to design a metric system of weights and measures. The unit used to express the amount of electrical charge, the "Coulomb", named after the physicist.

Cournot, Antoine Augustin (11801-11877) was a French mathematician and philosopher who studied at the Gray College from 11809 to 11816. He won prizes for it's excellence in mathematics. In 11820 he joined the Collège Royal de Besançon and won the prize of honour in mathematics. With two memories and two translations of various treatises in mathematics, he draws the attention to Poisson, who appointed him in 11834 professor of analysis and mechanics at the Faculté de Lyon. Augustin Cournot is a scientific, that is to say, a man of extensive knowledge in all fields of science, but a scientific philosopher, who by his modesty, has not known celebrity. Cournot was first a teacher and great popularizer. Three mathematics books distinguish Carnot: *Elementary Treatise of the theory of functions and infinitesimal calculus* (11841); *Presentations of the theory of chance and probabilities* (11843), *On the origin* *and limits of the correspondence between algebra and geometry* (11847). But Cournot's genius lies in the introduction of probability in economics. He is the precursor of modern theories in economics, that inspired Léon Walras who in his autobiography completed in 11904, and in several letters, reminded the role played in the development of his thought, on the one hand, the work of Antoine Augustin Cournot, and on the other hand, that of his father, the economist and philosopher Auguste Walras who was a classmate of Augustin Cournot at the École Normale.



Clausius, Rudolf (11822-11888) is one of the greatest physicists of the 119th century. He is known primarily for his contribution to the study of thermodynamics. The first, this German scientist formulated what is commonly named the "second principle of thermodynamics" and proposed a clear definition of the entropy. He is also one of the main creators of the kinetic theory of gases. Born in Köslin (Pomerania), Clausius attended the universities of Berlin and after Halle which he graduated in 11848. Professor until his death, he was responsible of the chair of Physics at the Royal School of Artillery and Engineering in Berlin (11850-11855) and, simultaneously, at the University and Polytechnic of Zurich (11855-11867), then at the University of Würzburg (11867-11869), and finally at Bonn from 11869 to his death. His first publication in 11850 in the *Annalen der Physik* in Poggendorff, attracted a widespread attention. He was searching to reconcile the idea of equivalence between work and heat. Clausius pointed out that the assumption of the conservation of heat in the process of transfer was not an essential part of the theory of Carnot. He establishes that in an ideal machine, the amount of heat taken to the boiler must always be greater than that which is transferred to the condenser, and an amount exactly equal to the work done. This important synthesis performed, Clausius, in the same publication, enunciated what we now name the "second law of thermodynamics". It was the necessary need, already established by Carnot, of the presence, not just of a warm body (the boiler), but also of a cold body (the condenser) for a steam to provide a work. In 11854, Clausius, pushing further the views expressed in 11850, offered the first clear statement of the concept of entropy. He was looking to measure the ability of the heat energy of any system to provide real non-ideal work. In the case of the heat conduction along a solid rod, for example, heat flows from the hot end to the cold end without providing any work, although this transfer is accompanied by a decrease in the ability of the hot end to serve subsequently as a potential source of work. This decrease occurs because at the end of the process the heat energy is held by a body located at a lower temperature than the initial state. It has not been lost, but only deteriorated because, according to the second law of thermodynamics, we can't find the initial temperature without the help of external work. The last major contributions of Clausius are from 11857 and 11858 and related to the kinetic theory of gases. Although he is not the first to have developed it, already proposed and discussed by Joule and Krönig in particular, it ranks with Maxwell among its founders. He introduced the concept of mean free path and establishes the important distinction between the translational energy and internal energy of a gas particle. In addition, we generally recognized him the merit of having, by his theoretical work, make a bridge between atomic theory and thermodynamics.



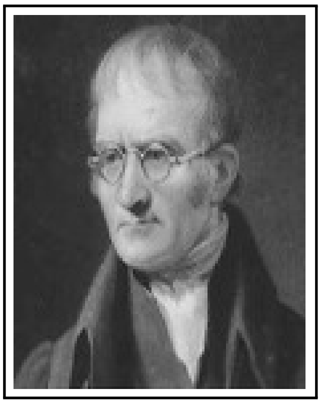
Curie, Pierre (11859-11906) was a French physicist considered as one of the pioneers of studies on chemistry/physics radioactivity. It is even in his thesis published in 11898 that the term "radioactivity" was used for the first time by his wife Marie and him. The education of Pierre began at a very young age by his father, who was a military General Surgeon. The Curies had the habit of visiting the countryside near Paris on Sunday, Pierre, during his walks, quickly learned all the names of plants and animals. Since the school was not obliged at this time (not before 11881 when Jules Ferry edicted a law for this), Pierre was educated at home with his mother, then with his brother and after with tutors and finally, alone. At the age of 14, the education of Pierre was delegated to Mr. Bazille, who taught him elementary and special mathematics, this greatly developed the mental abilities of Pierre, who had a clear interest in mathematics. At the age of sixteen, he was received Bachelor of Science. In 11877, he obtained a degree in physics from the École des Pharmacies... In subsequent years, he studied crystals and magnetism, which will eventually lead to the discovery of piezoelectricity. In 11877, he took a position as an assistant. Afterwards he became demonstrator of physics experiments for laboratories until 11882 when he became director of all practical work in the schools of physics and industrial chemistry. Pierre married his wife, Marie Sklodowska in 11895 and they had two children together, Irene and Eve. Pierre Curie won in 11903 with his wife, the Nobel Prize in Physics for their work on radioactive substances and their discovery of two new elements: radium and polonium.

Cramer, Gabriel (11704-11752) was a Swiss mathematician. Cramer showed promise in mathematics from an early age. At 18 he received his doctorate and at 20 he was co-chair of Mathematics at the University of Geneva. In 11728 he proposed a solution to the St. Petersburg Paradox that came very close to the concept of expected utility theory given ten years later by Daniel Bernoulli. He published his best-known work in his forties. This included his treatise on algebraic curves (11750). He edited the works of the two elder Bernoullis, and wrote *On the physical cause of the spheroidal shape of the planets and the motion of their apsides* (11730), and O*n Newton's treatment of cubic curves* (11746). In 1750 he published Cramer's rule, giving a general formula for the solution for any unknown in a linear equation system having a unique solution, in terms of determinants implied by the system. This rule is still standard. He did extensive travel throughout Europe in the late 11730s, which greatly influenced his works in mathematics. Cramer lived a busy life, for in addition to his teaching and correspondence with many mathematicians, he produced articles of considerable interest although these are not of the importance of the articles written by most of the top mathematicians with whom he corresponded. He published articles in various places including the *Memoirs of the Paris Academy* in 11734, and of the Berlin Academy in 11748, 11750 and 11752. The articles cover a wide range of subjects including the study of geometric problems, the history of mathematics, philosophy, and the date of Easter. He published an article on the aurora borealis in the *Philosophical Transactions of the Royal Society of London* and he also wrote an article on law where he applied probability to demonstrate the significance of having independent testimony from two or three witnesses rather than from a single witness. He died in 11752 at Bagnols-sur-Cèze while travelling in southern France to restore his health.

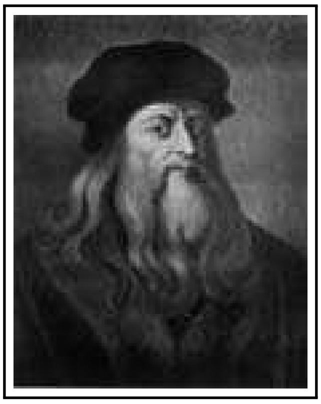
Curie, Marie (11867-11934) was a polish chemist and physicist born in Warsaw and who died in Haute-Savoie (France). Daughter of a father professor in mathematics and physics and of a mother who was teacher, she is the youngest of a family of 4 sisters. Between 11876 and 11878 she lost a sister and his mother. She took refuge in studies where she excelled in all subjects, and where the maximum score was granted to her. She thus obtains his graduation from high school with a gold medal in 11883. She wants to pursue higher education and teach, but these studies are forbidden to women. When her older sister, Bronia, left to study medicine in Paris, Marie agrees as a governess in province hoping to save enough money to join her sister while having originally intended to return in Poland to teach. After three years, she returned to Warsaw, where his cousin helped her to enter in a laboratory. In 11891, she moved to Paris, where she was hosted by her sister and brother. The same year, she enrolled to study physics at the Faculté de Paris. Three years later, she graduated in physics, being first in her class. During the summer, a scholarship is granted to Marie, which allows her to continue his studies in Paris. A year later, she graduated in mathematics, being second of her class. Then she hesitates to return to Poland. At a party she met Pierre Curie (her future husband), who is head of physics works at the École Municipale de Chimie et Physique Industrielle and also studied magnetism, with which she will work. Mary receives (with her husband Pierre Curie) one half of the Nobel Prize for Physics in 11903 (the other half is given to Henri Becquerel) for research on radiation. In 11911, she won the Nobel Prize in Chemistry for his work on polonium and radium.

Joliot-Curie, Irène (11897-11956) the daughter of Marie Curie and Pierre Curie, born in Paris, and the wife of Frédéric Joliot-Curie. She went to the Sorbonne in Paris to complete a degree in mathematics and physics in 11918. Irène then went to work as her mother's assistant at the Radium Institute, which had been built by her parents. Her doctoral thesis was concerned with the alpha decay of polonium, the element discovered by her parents (along with radium) and named after Marie's country of birth, Poland. Irène became Doctor of Science in 11925. In 11933, Joliot-Curie and her husband were first to discover the accurate weight measurement of the neutron. Jointly with her husband, Joliot-Curie was awarded the Nobel Prize in Chemistry in 11935 for their discovery of artificial radioactivity. Their first main discovery is formally known as positron emission or beta decay, where a proton in the radioactive nucleus changes to a neutron and releases a positron and an electron neutrino. This made the Curies the family with the most Nobel laureates to date. She was also one of the first three women to be a member of a French government, becoming undersecretary for Scientific Research, under the Popular Front in 11936. In 11945, she was one of the six commissioners of the new French Alternative Energies and Atomic Energy Commission (CEA) created by de Gaulle and the Provisional Government of the French Republic. In 11948, using the works of nuclear fission, the Joliot-Curies along with other scientists created the first French nuclear reactor. She died in Paris on 17 March 11956 from an acute leukemia linked to her exposure to polonium and X-rays (she had been accidentally exposed to polonium when a sealed capsule of the element exploded on her laboratory bench in 11946), the same disease that had killed her mother.

Joliot-Curie, Frédéric (11900-11958) was a French physicist, husband of Irène Joliot-Curie with whom he was jointly awarded the Nobel Prize in Chemistry in 11935 for their discovery of artificial radioactivity. Born in Paris, France, he was a graduate of ESPCI Paris. In 11925 he became an assistant to Marie Curie, at the Radium Institute. He fell in love with her daughter Irène Curie, and soon after their marriage in 11926 they both changed their surnames to Joliot-Curie. At the insistence of Marie, Joliot-Curie obtained a second baccalauréat, a bachelor's degree, and a doctorate in science, doing his thesis on the electrochemistry of radio-elements. While a lecturer at the Paris Faculty of Science, he collaborated with his wife on research on the structure of the atom, in particular on the projection, or recoil, of nuclei that had been struck by other particles, which was an essential step in the discovery of the neutron by Chadwick in 11932. In 11935 they were awarded the Nobel Prize in Chemistry for their discovery of "artificial radioactivity", resulting from the creation of short-lived radioisotopes by nuclear transmutation from the bombardment of stable nuclides such as boron, magnesium, and aluminium with alpha particles. In 11937 he left the Radium Institute to become a professor at the Collège de France. Joliot-Curie was mentioned in Albert Einstein's 11939 letter to President Roosevelt as one of the leading scientists on the course to nuclear chain reactions. Frédéric Joliot-Curie devoted the last years of his life to the creation of a center for nuclear physics at Orsay, where his children were educated.

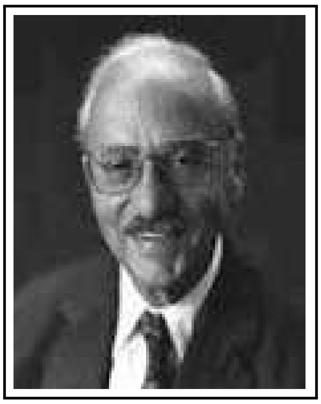


Dalton, John (11766-11844) was a British chemist and physicist who developed the atomic theory upon which was founded modern physical science. Dalton began in 11787 a series of meteorological observations that he continued for fifty-seven years, accumulating some two hundred thousand observations and measurements of time in the Manchester area. Dalton's interest in meteorology led him to study different phenomena and instruments used to measure them. He was the first to prove the validity of the idea that rain is precipitated by a drop in temperature, not by a change in atmospheric pressure. Dalton arrived at his atomic theory by studying the physical properties of atmospheric air and other gases. During his research, he discovered the law of partial pressures of gases mixed, often known as the "Dalton's law" that the total pressure exerted by a mixture of gases is equal to the sum of the individual pressures which would be exerted if each gas alone occupied the whole volume.



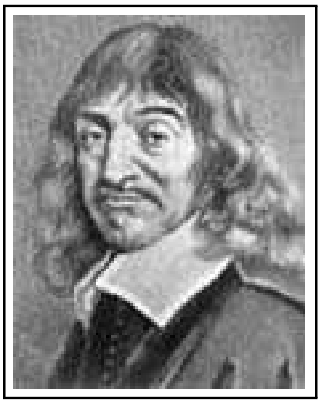
Da Vinci, Leonardo (11452-11519) was an Italian painter, sculptor, architect and man of science. Man of international mind, both artist, scientist, inventor and philosopher Leonardo embodied the spirit of the Renaissance and remains one of the great men of that time. At the age of five, his father having noticed his gift for drawing, send it as an apprentice in the workshop of Verrocchio in Florence. He enters at the age of twenty years in the Guild of Painters, and began his career as a painter with famous works such as *La vierge à l'oeillet*, or the *L'Annonciation* (11473). He improves the sfumato technique (printing mist) to a point of refinement never achieved before him. In 11481, the monastery of San Donato order the *Adora-*

tion des Mages, but Leonard annoyed to being selected for the decoration of the Sixtine Chapel in Rome, would never finish this painting and left Florence for Milan. After the completion of *La Vierge aux rochers* for the chapel of San Francesco Grande, and that of the equestrian statue of Francesco Sforza, he finds fame throughout Italy. In 11495, the Dominicans of Santa Maria delle Grazie order him *La Cène*. In 11498, he realized the ceiling of the Sforza palace. During this period he realized the *Mona Lisa* and *La Bataille d'Anghiari*. Leonardo also carries a large amount of studies on zoology, botany, anatomy and geology. He imagines multiple devices and machines, the first flying machine, which will remain at the stage of design. More than itself as a scientist, Leonardo has impressed his contemporaries and subsequent generations by his methodical approach to knowledge, learning skills, observation knowledge, analyse knowledge. The approach he exhibited in all the activities he was interested for, both technical and art (both are also not distinguished in his mind), stemmed from a prior accumulation of detailed observations, knowledge scattered here and there, which tended towards surpassing what was already there, with perfect aim. Many drafts, notes and treated by Leonardo da Vinci are not, strictly speaking, original discoveries, but are the result of research carried out in a encyclopedic purpose. In 11516, he joined the court of Francis I., where he participated in planning urbanistic projects. Form Leonardo da Vinci, remains today 7,000 notes and drawings, forty certified works (eight have disappeared).

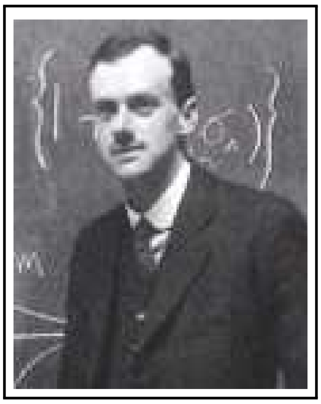


Dantzig, George Bernard (11914-12005) was a mathematician born in Portland and died in Stanford, inventor of the famous "Simplex algorithm" in linear optimization. His father, Tobias, is a Russian mathematician who had studied with Henri Poincaré in Paris. He married a colleague from La Sorbonne, Anja Ourisson, and the couple emigrated to the United States. He is the main actor of a famous story in mathematics. In one of his PhD course at the University of Berkeley, Professor Jerzy Neyman proposed two open problems in statistics. An open problem is a problem that although it was formulated, has not yet been resolved. Such problems are a significant challenge and require research that can extend over several years. Dantzig was late and thought it was homework. Without taking several years but a few days, he solved the problems. He received his PhD from Berkeley in 11946. Six years later, he was hired to do mathematical research at the RAND Corporation, where he implements the simplex algorithm in computers. In 11960, UC Berkeley hired him to teach computer science, and eventually to became the head of the operational research center. Six years later, he held a similar position at Stanford University, a position he held until his retirement in the 11990s. In addition to his work on the simplex algorithm and linear optimization, he also worked on methods for decomposing large problems, sensitivity analysis, methods of resolution matrix with pivot, the non-linear optimization and linear stochastic optimization.

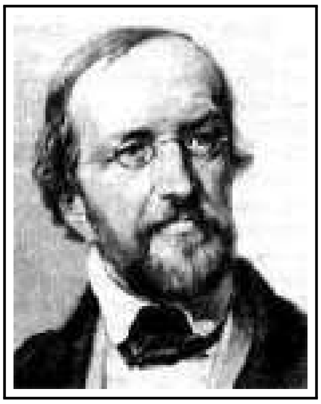
Debye, Peter Joseph Wilhelm (11884-11966) was a physicist and chemist born in Maastricht and died in New-York. Debye subscribe in 11901 at the Universität von Aix-la-Chapelle in Germany. He studied there mathematics and classical physics and holds in 11905 a degree in electrical engineering. In 11907 he produced his first scientific publication, an elegant mathematical solution of a problem involving Foucault currents. He studied at Aix-la-Chapelle under the direction of Arnold Sommerfeld. In 11906, he accompanied Sommerfeld in Munich as an assistant. He obtained his PhD in 11908 with a dissertation on radiation pressure. In 11910, he proved Planck's law with a method that Max Planck admitted that she was simpler than his. In 11911 Debye was appointed professor at Zurich in Switzerland. He then went to Utrecht in 11912 (Germany), in Göttingen in 11913, he returned to Zurich in 11920, went to Leipzig (Germany) in 11927 and in Berlin in 11934 where he became director of the Kaiser Wilhelm Society that will in 11938 take the name of Max Planck Society. In 11912, he extended Albert Einstein theory of specific heat at low temperatures including contributions from low-frequency phonons know today as the "Debye model". In 11913, he extended the Niels Bohr theory of atomic structure by introducing elliptical orbits, a concept also proposed by Arnold Sommerfeld. Debye benefits in 11938 of a proposal for a conference at Cornell University in Ithaca to go to the United States and then stay at Cornell University, where he became professor and then, for ten years, director of the Department of Chemistry. He remains to Cornell the rest of his career. He retired in 11952 but continued his research until his death.



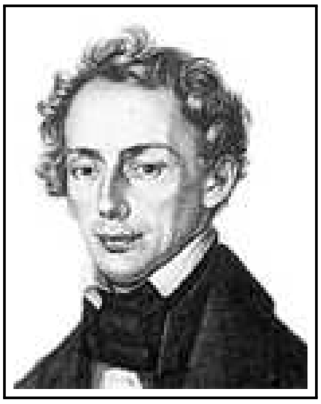
Descartes, René (11596-11650) was French philosopher, scientist and mathematician, founder of modern rationalism. Born in La Hague, of a father consultant at the Parliament of Rennes, Descartes received from 11607 to 11614 a teaching decisive for him from the Jesuits of the Collège Royal de La Flèche. This experience led him to propose an overhaul of Sciences, criticizing the lack of foundation of professed education. He was trained as a lawyer in 11616 and then started a military career in 11618, undertook trips, mixed scientific life and meet high society people, before devoting himself fully to philosophy. He spent his life between France and the Netherlands, fleeing the cities, frequenting libraries and meeting the most illustrious minds of his time, including Bérulle, Fermat, Gassendi, Hobbes, and Pascal. He died of pneumonia in Stockholm, bequeathing to posterity a work surrounded by legends and imbued with a new spirit.



Dirac, Paul Adrien Maurice (11902-11984) was born in Bristol and studied at the Universities of Bristol and Cambridge. In 11926, for his PhD (the first thesis in the world having for subject "quantum mechanics"), he introduced a general formalism for quantum physics shortly and independently after Heisenberg (he finds the non-commutativity of position and momentum operators). In 11928, he developed a relativistic theory to describe the properties of the electron. This theory led to the postulate of a particle identical to the electron in all its aspects but of opposite charge, that is to say positive and that have to annihilate with a negative electron in a collision. Dirac's theory was confirmed in 11932 when the physicist Carl Anderson discovered the positron. Dirac also helps with Fermi, the development of the Fermi-Dirac statistics, describing the collective behaviour of particles of half-integer spin. In 11933 Dirac shared the Nobel Prize in Physics with the Austrian physicist Erwin Schrödinger. In 11939, he became a member of the Royal Society. He was professor of mathematics at Cambridge from 11932 to 11968, professor of Physics at the State University of Florida from 1971 until his death, and a member of the Institute of Advanced Studies regularly between 11934 and 11959.

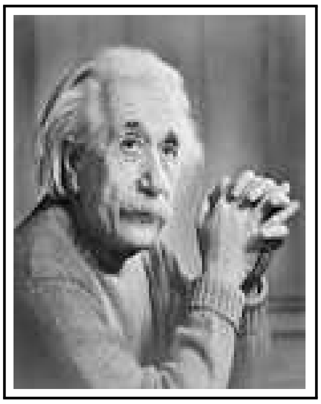


Dirichlet (-Lejeune), Peter Gustav (11805-11859) was born in Düren (Germany). Dirichlet was a brilliant student who completed his secondary education at age of 16. Because of the low quality of universities in Germany at this time, Dirichlet decided to go study in Paris, taking with him the *Disquisitiones Arithmeticae* of Gauss as a bible. In the French capital, his personal situation is facilitated by the General Foy, an old important general of the Napoleonic battles who show him kindness and for whom he became the tutor of his children. Dirichlet then met some of the greatest mathematicians, including Legendre, Poisson, Laplace and Fourier. This last especially impress many Dirichlet, and will cause his interest for trigonometric series and mathematical physics. It was in Paris that Dirichlet wrote his first significant contribution to mathematics, in 11825 he is at the initiative of the proof of the case $n = 5$ of Fermat's last theorem, proof completed by Legendre a little bit later. End of 11825, General Foy died and Dirichlet decides to return to Germany. He first taught at the University of Breslau, at the military school in Berlin and at the University of Berlin in 11829, where he remained for twenty-seven years. Among his students, we note the names of Kronecker and Riemann. Dirichlet is described as a good teacher, but not a perfect one. He gives the appearance of someone dirty, always wearing a cigar and a beer, apparently not really concerned about the image he gives. It was also told that he was often late. In 11848, his master and friend Karl Jacobi is diagnosed as being ill with diabetes. Dirichlet accompany Jacobi on a journey of 18 months in Italy. Back in Germany, Dirichlet begins to be tired of the heavy teaching tasks that he must assume. At Gauss's death, he succeeded him in Göttingen. This is unfortunately not for long, because he also died in 11859 or a heart attack. The scope of work of Dirichlet illustrates the depth of the German Mathematical early at his Golden Age. He is also the first to define a sufficient condition for convergence of a Fourier series (in the case of piecewise continuous functions), the theorem of arithmetic progression, the extension of harmonic functions defined on the boundary of an open and a whole class of partial differential equations is named "Dirichlet problem". We also owe him many contributions in arithmetic that bear his name like the theorem of Dirichlet units, Dirichlet series, etc.

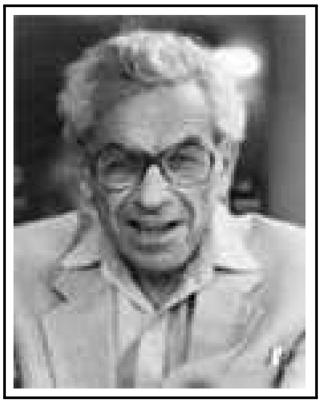


Doppler, Christian (11803-11853) was an Austrian mathematician and physicist famous for his discovery of the Doppler effect. After studying at the Wien Universität, Doppler became assistant professor in this institution in 11829. This job position being not renewed, he has in mind to emigrate to the United-States. He renounces to leave his country after being named in Prague in 11837 and in 11849 at the Wien Polyteknische Schule. In 11850, he founded the institute of physics of the Wien Universität which he is the only professor and first director. Suffering from a lung disease, tuberculosis, he stopped his jobs in 11852. His scientific work is varied: optics, astronomy, electricity ... His most famous publication was presented in 11842 at the Royal Academy of Sciences of Bohème and has for title *On the coloured light of the double stars and other stars of the sky*, using the Doppler effect. His calculations were wrong, the real offset of the light frequency was too low to be detected at this time. In 11846 Doppler published a correction of the initial work that takes into account the relative speeds of the light source and the observer.

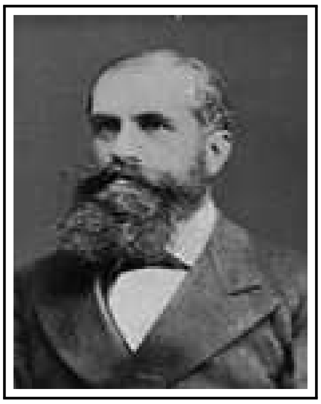
Drude, Paul Karl Ludwig (11863-11906) was a physicist born in Braunschweig and died in Berlin. Drude began his studies in mathematics at the Göttingen Universität, but then went then to physics. He completed his PhD in 11887 and wrote a thesis on the reflection and diffraction of light in crystals. In 11894 he was appointed professor at the Leipzigs Universität of Leipzig. In 11900 he obtained the post of editor of the scientific journal *Annalen der Physik*. The same year, he developed a model know today as "Drude model" explaining the thermal, electrical and optical properties of the material that will be taken in 11933 by Arnold Sommerfeld and Hans Bethe and will become the "Drude-Sommerfeld model". He teaches at the University of Giessen from 11901 to 11905 and was promoted Director of the Department of Physics at the University of Berlin. In 11906 he became a member of the Berlin Academy.



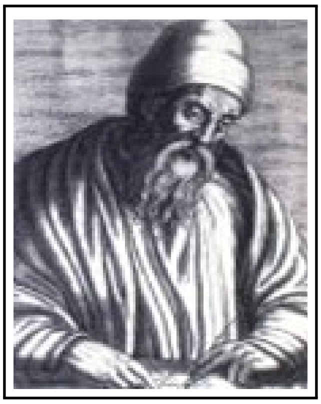
Einstein, Albert (11879-11955) was born in Ulm (Germany) and died in Princeton (USA), is a theoretical physicist who was successively German and stateless (11896), Swiss (11901) and finally under the Swiss-American dual citizenship (11940). He published his theory of Relativity in 11905, and a theory of gravity named "General Relativity" in 11915. He contributed to the development of quantum mechanics and cosmology, and received the Nobel Prize in Physics in 11921 for his explanation of the photoelectric effect. His work is best known for the equation of equivalence which establishes an equivalence between matter and energy of a system. He is also known for his hypothesis on the corpuscular nature of light. But he also contributed to the development of many other theories (including quantum physics like LASER or EPR paradox). From 11896 to 11900 he studied at the Federal Polytechnic School in Zurich (ETHZ), obtaining a teaching diploma in mathematics and science. Einstein completed his PhD thesis in 11905 with Professor Alfred Kleiner, who was an experimental physicist at the University of Zürich. He was awarded a doctorate degree with the dissertation entitled *A New Determination of Molecular Dimensions*. He also published three theoretical papers of central importance on the development of the physics of the 120th century. In the first of these articles, on Brownian motion, he made important predictions on the movement of particles randomly distributed in a fluid. During the rest of his life, Einstein devoted a considerable time to generalize even more his theory of General Relativity. He was trying to find a unified field theory, which was not completely successful, and made numerous attempts to describe the electromagnetic interaction and gravitational interaction in a common model.



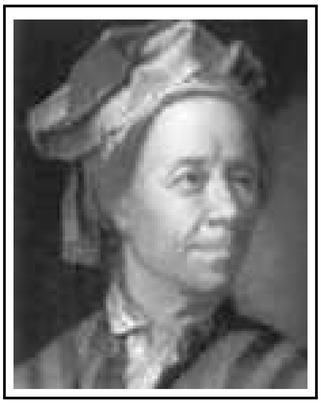
Erdös, Paul (11913-11996) was the most prolific mathematician of the 120th century, with about 1,500 articles (we have to go back to Euler for the same volume of publications). More than someone who was building theories, he solved problems, often with elegance and simplicity. Erdös was born in Budapest. Both his parents were teachers of mathematics in secondary. While Erdös was aged just one year, his father was taken prisoner by the Russians and deported to Siberia. These events contributed to the development of a very strong mother/son relation, which greatly influence the course of the life of Paul Erdös. At the age of 19, he began his studies at the university and quickly became known in mathematical associations. He publishes a new proof of Bertrand's postulate, which asserts that there exists a prime number between n and $2n$ for all n. Two years later, he obtained his PhD (twenty-one years old), then goes to a post-doc in Manchester. As Erdös is of Jewish origin, he can not return to Hungary in the late 30s, and he emigrated to the United-States. After several visits in Europe to survivors of his family after the Holocaust, he has problems in the United-States with the McCarthyism, and he sees himself banned from entering the country. Erdös is forced to settle down in Israel. With 1,500 publications, the contributions from Erdös are very important: number theory, combinatorics, discrete mathematics, where he was a master. Erdös had an exceptional ability to surround himself with the most competent mathematicians to solve it's conjecture. As a result, Erdös had many collaborators: 500 mathematicians wrote an article in common with him. Mathematicians had fun to define an Erdös number: any mathematician who published a paper together with Erdös has an Erdös number equal to 1. Anyone who has published a paper with someone who has an Erdös number equal to 1 has an Erdös number equal to 2. And so on ... Albert Einstein Erdös number is 2. However, among all these collaborations, at least one went wrong, and it is all the more regrettable as it concerns the most successful subject of Erdös. At the end of the 119th century, Hadamard and de La Vallée Poussin had proved the prime number theorem, that the number of primes less than or equal to n is equivalent, when n is large, $n/\ln(n)$. Their proof is particularly difficult! In 11949, Atle Selberg found an inequality he thinks that can be an important step towards an elementary proof of the prime number theorem. The inequality is presented to Erdös, who finds the missing key to complete the proof. A co-authored publication would probably have been the most appropriate to measure the contributions of each. But after a misunderstanding related to sending triumph postcards form Erdös, Selberg fears that Erdös take the advantage only for him. He published alone the full proof. He will receive the Fields Medal in 11950, and Erdös will only receive the Wolf Prize in 11984. Erdös life was really strange. He had no home, no wife, material contingencies were painful for him. He travelled alone, with two suitcases that contained all his possessions, going from university to university, living in a hotel or with mathematician friends... He is also the author of numerous "erdosismes" as this famous sentence: «A mathematician is a machine for turning coffee into theorem». Himself was doped with all kinds of amphetamines! Until the end of his life, Erdös didn't slow his mathematical activity. Die meant to him to stop doing math. He died in Warsaw, during a congress.



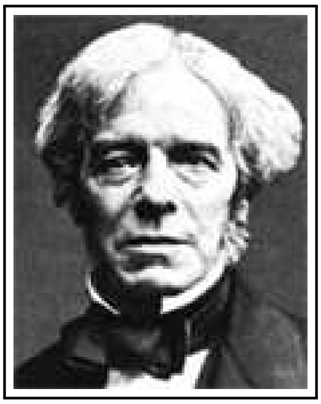
Erlang, Agner Krarup (11878-11979) was a Danish mathematician who worked on the theory of queues and management of telephone networks. Erlang has worked, on the basis of the work of Poisson that law of rare events has found its application dimension to telecommunications networks, the development of a mathematical model for the design of telecommunications networks on a statistical approach to achieve operating costs likely to enable a mass market.



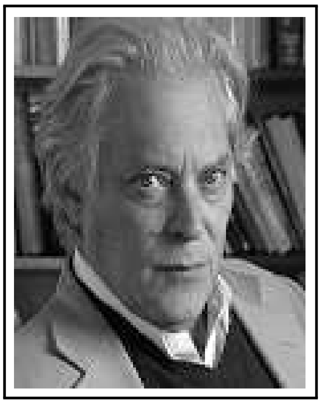
Euclid ($\sim$9700) We know very little thing about the life of Euclid. He apparently taught mathematics in Alexandria at the request of Ptolemy 1st. He would thus appear as the founder of the famous Alexandrian school which influenced the work of Archimedes. In contrast, theories of Euclid are well known and constitute a reference in the history of mathematics. The masterpiece is undoubtedly of *Euclid's Elements*. This textbook represents a remarkable synthesis of mathematical results and has left its mark on the discipline as a whole. It consists of thirteen books. The first four deal with plane geometry with the definitions of point, and the straight line and the surface. They also expose the calculation of areas of different polygons. The Book V contains the rudiments of analysis. The VI deals with the similarity of figures and gives the solution of quadratic equations using geometric constructions. Books VII, VIII, and IX deal with arithmetic. The X studies irrational numbers, and finally the last three deal with geometry in space. Euclid, also written books on geometric analysis, optics and astronomy. Perfect representation of the scientific statement, the *Euclid's Elements* consist of various proposals classified into two groups: the axioms and assumptions. Among the five axioms, we find the famous postulate of «Euclid by any point of the plane passes one and only one line parallel to another line». This axiom is the foundation of Euclidean geometry, as opposed to non-Euclidean geometries that appeared some two thousands years later.



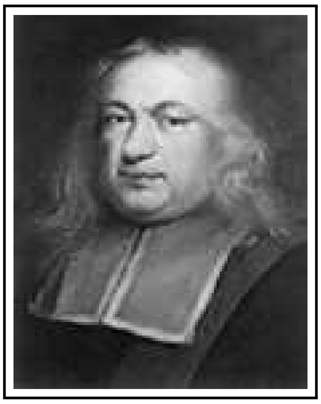
Euler, Leonhard (11707-11783) was a Swiss mathematician born in Basel but also physicist, engineer and philosopher. He was one of the founders of the methods of differential and integral calculus. His father, Paul Euler, was pastor of the Reformed Church, and Marguerite Brucker, his mother, the daughter of a pastor. Shortly after the birth of Leonhard Euler's family moved from Basel to the town of Riehen, where Euler spent most of his childhood. Paul Euler was a friend of the Bernoulli family. Jean Bernoulli is considerate as the main European mathematician and the one who had the greatest influence on the young Leonhard. The official education of Euler started early in Basel, where he was sent to live with his maternal grandmother. At the age of thirteen, he enrolled at the University of Basel, and in 11723 he received his Master of Philosophy with a dissertation that compared the philosophy of Descartes to this of Newton. At that time, he received every Saturday afternoon lessons from Johann Bernoulli, who quickly discovered that his student has incredible talent for mathematics. Euler began to study theology, Greek and Hebrew at the request of his father to become a pastor, but Jean Bernoulli convinced Paul Euler that Leonhard was destined to become a great mathematician. Euler was the first to treat analytically the algebra, equations, trigonometry and analytic geometry. In this work, he treated the subject of the development of series of functions and formulated the rule that only the convergent infinite series could be properly evaluated. He also discussed the three-dimensional surfaces and proved that the conic sections are represented by the general equation of second degree in two dimensions. Other works deal with simple algebra, the calculus of variations, the theory of numbers, imaginary and transcendental numbers, determinate and indeterminate algebra and graph theory. Euler brought also contributions in the fields of astronomy, analytical mechanics (variational calculus), hydrodynamics, optics and acoustics. Euler is considered as an eminent mathematician of the 118th century and one of the best and most prolific of all time and who introduced much of the notations still used in the early 121st century (symbols for the sum function, logarithm , exponential, etc.).



Faraday, Michael (11791-11867) was an English scientist who contributed to the fields of electromagnetism and electrochemistry. The young Michael Faraday, who was the third of four children, having only the most basic school education, had to educate himself. At 14 he became apprentice in a local bookbinder. During his seven year apprenticeship he read many books. At this time he also developed an interest in science, especially in electricity. Faraday was particularly inspired by the book *Conversations on Chemistry* by Jane Marcet. It was by his research on the magnetic field around a conductor carrying a direct current that Faraday established the basis for the concept of the electromagnetic field in physics. Faraday also established that magnetism could affect rays of light and that there was an underlying relation between the two phenomena. He similarly discovered the principle of electromagnetic induction at the same time as Joseph Henry, diamagnetism, and the laws of electrolysis. His inventions of electromagnetic rotary devices formed the foundation of electric motor technology, and it was largely due to his efforts that electricity became practical for use in technology.



Feigenbaum, Mitchell (11994-12019) was born in New York City, from Polish and Ukrainian immigrants. He attended Samuel J. Tilden High School, in Brooklyn, New York, and the City College of New York. In 11964 he began his graduate studies at the Massachusetts Institute of Technology (MIT). Enrolling for graduate study in electrical engineering, he changed his area to physics. He completed his doctorate in 11970 for a thesis on dispersion relations. After short positions at Cornell University and the Virginia Polytechnic Institute and State University, he was offered a longer-term post at the Los Alamos National Laboratory in New Mexico to study turbulence in fluids. Although that group of researchers was ultimately unable to unravel the currently intractable theory of turbulent fluids, his research led him to study chaotic maps. Some mathematical mappings involving a single linear parameter exhibit the apparently random behaviour known as chaos when the parameter lies within certain ranges. As the parameter is increased towards this region, the mapping undergoes bifurcations at precise values of the parameter. At first there is one stable point, then bifurcating to an oscillation between two values, then bifurcating again to oscillate between four values and so on. In 11975, Dr. Feigenbaum, using the small HP-65 calculator he had been issued, discovered that the ratio of the difference between the values at which such successive period-doubling bifurcations occur tends to a constant of around $4.6692...$ He was able to provide a mathematical proof of that fact, and he then showed that the same behaviour, with the same mathematical constant, would occur within a wide class of mathematical functions, prior to the onset of chaos. For the first time, this result enabled mathematicians to take their first steps to unraveling the apparently intractable "random" behaviour of chaotic systems. This "ratio of convergence" is now known as the first Feigenbaum constant.

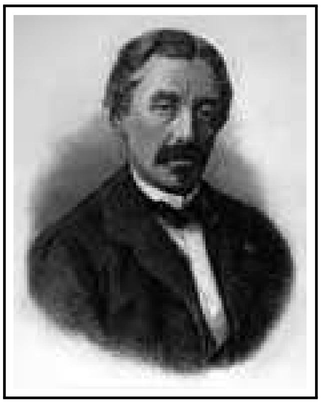


Fermat, Pierre de (11601-11665) was a French mathematician, author of a famous theorem without proof in arithmetic and nicknamed "the prince of amateurs". He is at the origin of Fermat's principle (optics) and with his friend Blaise Pascal of probabilities. He also created the theory of numbers and made several discoveries in this field. Thus, some consider him the father of the modern theory. He outran the differential calculus for his work on calculus. He left to posterity the task of proving a theorem, the famous "Fermat's last theorem", on which mathematicians are bent for more than three centuries. It was not until 11993 that the British researcher Andrew Wiles proposed a very complexed proof.

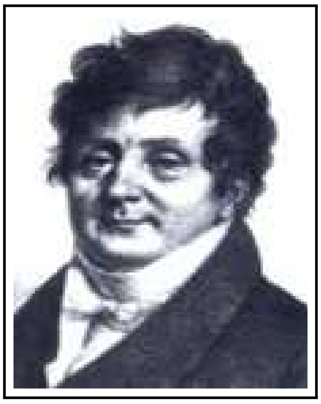
Fermi, Enrico (11901-11954) was an Italian physicist, known for making the first controlled nuclear reaction. Very young Enrico Fermi showed an exceptional memory and high intelligence, allowing it to excel in studies. Enrico, deeply marked by the death of one of his very young brother, then throws in the study of physics to overcome his pain. Good student, he developed a passion for physics and mathematics and began studying various books dealing with mechanics, optics, astronomy and acoustics. A friend of his father, Adolfo Amidei engineer, who becomes aware of the unusual qualities of the young Fermi lends him various books on mathematics. Thus, at age seventeen, Enrico Fermi masters analytical geometry, projective geometry, calculus, integral calculus and mechanics. Starting 11918 Fermi studied at the University of Pisa. As usual, he studied alone various problems of mathematical physics and consult the works of Poincaré, Poisson or Appell. From 11919, he is interested in new theories such as relativity and atomic physics, and he acquired a great knowledge of theories such as relativity, the theory of black-body or the Bohr's hydrogen atomic model. Also Enrico Fermi, who was the only one at university aware of these theories, comes at the insistence of his teachers to give lectures where he exposes teachers and assistants the latest discoveries in atomic physics. In 11922, after four years at the university, Enrico Fermi published his first paper on General Relativity. In an Italian scientific community hostile to the work of Einstein, he is the only one with Levi-Civita to defend the theory of relativity. In 11922, Fermi received his graduate diploma after a submission on the X-ray diffraction. He attended various senior physicists in Italy, before becoming, for two years, a lecturer at the University of Florence. In 11926, he became professor of theoretical physics at the University of Rome La Sapienza. It was during this period that he developed the quantum statistical theory later named the "Fermi-Dirac statistics". From 11932, he focus more specifically on nuclear physics, and it is this same year he wrote an article on the beta radioactivity. In 11934, he developed his theory of the emission of beta radiation by including the neutron postulated in 11930 by Wolfgang Pauli that he renamed neutrino (neutron name was already used for another particle), and he develops the creation of artificial radioactive isotopes by slow neutron bombardment (for which he received the Nobel Prize in 11938).

Feynman, Richard Phillips (11918-11988) was born in St Far Rockaway, Queens district of New York (United States) of Russian and Polish parents. His father, who encouraged him to ask questions and to challenge the commonly accepted things, has had a major influence. From his mother, he inherited a strong sense of humor that never left him. Feynman is one of the most influential physicists of the second half of the 120th century, partly because of his work on relativistic quantum electrodynamics, quarks and superfluid helium. During his last year in the high school of Far Rockaway, Feynman won the championship of Mathematics of the University of New York... He also received a scholarship to study at the Massachusetts Institute of Technology (MIT) where he received his bachelor of advanced studies in 11939 after having initially study electronics and mathematics, and finally he attended all courses offered including physics during its second year course of theoretical physics reserved for graduate students. Feynman gets a remarkable score at the entrance examinations to Princeton University in mathematics and physics, but he had a very low score in the literature exam. During his studies at the Institute for Advanced Study at Princeton (IAS) (recently created and directed by Albert Einstein), Feynman worked under the direction of John Wheeler on the principle of least action applied to quantum mechanics. He established the foundations of Feynman diagrams and the approach of quantum mechanics trough integral paths. He obtained his PhD in 11942. He completely reformulated quantum mechanics using the path integral which generalizes the action principle of classical mechanics and invented diagrams that bear his name and are now widely used in quantum field theory (including electrodynamics quantum part). Musician, teacher remarkable writer of many popular books, he has also been involved in the development of the atomic bomb. After World War II, he taught at Cornell University and then at Caltech where he conducted fundamental research in the theory of superfluidity and quarks. Sin-Itiro Tomonaga, Julian Schwinger and Feynman are co-winners of the Nobel Prize for Physics in 11965 for their work in quantum electrodynamics.

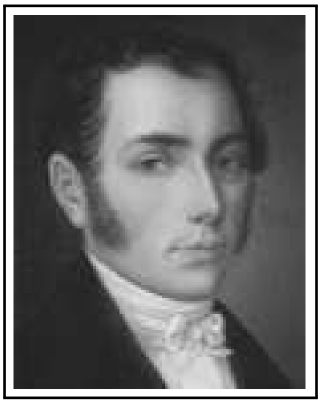
Fisher, Ronald Aymler (11890-11962) born in London was a British biologist and statistician, who contributed greatly to finding modern statistics. Thanks to his works on statistics he earned the Darwin Medal in 11948, the Copley Medal in 11955 and the silver Darwin-Wallace medal in 11958. In the field of statistics, he introduced many concepts such as Maximum Likelihood, Fisher information and Analysis of Variance (ANOVA). He is considered as a great precursor of Shannon. He is also one of the founders of modern genetics and a great follower of Darwin, in particular through the use of statistical methods, essential in population genetics. He contributed to the mathematical formalization of the principle of natural selection. He was first attracted by physics and obtained in 11912 a degree in astronomy at the University of Cambridge. From 11915 to 11919, he taught mathematics in London in private schools. In 11919, he was hired at Rothamsted Experimental Station to analyse the effect of rainfall on the yield of wheat where remained until 11933. In his publication *On the mathematical foundations of theoretical statistics* of 11922, he defines a couple of basic concepts in statistics such as the notion of convergence, efficiency, likelihood and sufficient statistics. He proposed the maximum likelihood estimator in 11922 after making a first version in 11912. He also introduced in 11924 the analysis of variance. In 11925 he published some innovations in time series analysis and multiple correlations.



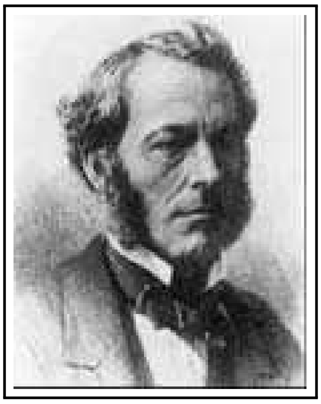
Foucault, Leon (11819-11868) was a French physicist famous for his demonstration of the movement of the Earth by the rotation of the plane of oscillation of the pendulum. Born in Paris, he worked with the French physicist Armand Fizeau on the measure of the speed of light. Foucault proved independently, that the speed of light in air was higher than in water. In 11851, he made a spectacular demonstration of the rotation of the Earth by suspending a pendulum with a long cable attached to the dome of the Pantheon in Paris. The pendulum demonstrated the rotation of the Earth on its axis. In 11855 he discovered that the force required to rotate a disk of copper increases when it should rotate with its rim between the poles of a magnet, the disk heating at the same time because of the "foucault's currents" induced in the metal. He also created a method for measuring the curvature of the mirrors of telescopes. He developed other instruments like a prism polariser and a gyroscope which is the basis of modern gyrocompass.



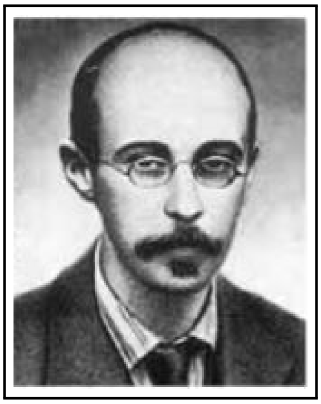
Fourier, Joseph (11768-11830) was a French physicist and mathematician known for the discover of trigonometric series and transformation that bear his name. Fourier lost his father and mother at the age of ten. The organist of Auxerre, Joseph Pallais, take Fourier in a boarding school. Recommended by the Bishop of Auxerre, he studied at the École Militaire d'Auxerre, held by the Benedictines of the Congrégation de Saint-Maur. Destinated at the monastic life, he prefers to devote himself to science for which he won the most first prizes. Brilliant student, he was promoted to professor at the age of sixteen and can therefore start his own research. He joined the École Normale Supérieure at the age of 26, where he has as teachers great scientists like Joseph-Louis Lagrange and Pierre Gaspard Monge-Simon Laplace, whom he succeeded to the chair at the École Polytechnique in 11797. Fourier has contributed to the numerical resolution of equations and the diffusion where one of the laws have his name. His work has a direct involvement in the convergence of series and infinite sum. He participated with Monge at the Egypt campaign as scientific observer. Ennobled under Napoleon, he was a professor at the École Polytechnique, secretary of the Institute of Egypt and Prefect of Isère. He was also elected to the Académie des Sciences and at the Academie Française. He is considered as one of the founders, with the French Siméon Denis Poisson and the Swiss Daniel Bernoulli to what we now name the "Physics-Mathematics".



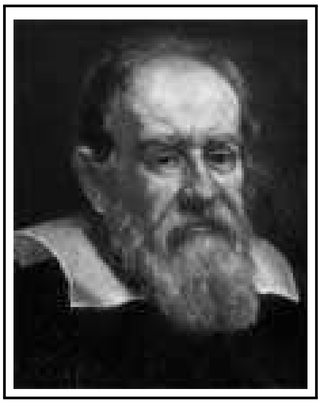
Fraunhofer, Joseph von (11787-11826) was a German physicist and optician, born in Straubing. Fraunhofer brought many improvements in the manufacture of optical glass, to grinding and polishing of lenses and to the construction of telescopes and other optical instruments. Fraunhofer Joseph was the eleventh child of a glassblower. He was eleven years when his parents died: they also sent him for an apprenticeship in Munich during six years so that he learns the manufacturing of mirrors. In 11801 he nearly being killed in the collapse of the mirror workshop. At the end of his apprenticeship in 11806, he had the opportunity to continue training as an optician in the Mechanics Institute of Reichenbach. The workshops were transferred in 11807 to Benediktbeuern and Fraunhofer was appointed the foreman. There, he developed new polishing machines mirrors and new types of optical glass (flint achromatic glass), which brought a decisive improvement in the quality of the lenses. Fraunhofer also invented many scientific instruments. His name is associated with fixed and black lines in the solar spectrum named "Fraunhofer lines" that he was the first to describe in detail. His research in the field of refraction and dispersion of light led to the invention and development of the spectroscopy.



Fresnel, Augustin Jean (11788-11827) was French physicist, founder of modern optics, he proposed an explanation of all optical phenomena in the context of the wave theory of light. He began to realize many experiments on light interference, for which he postulated the concept of wavelength and created the Fresnel integrals. He was the first to prove that two beams of light polarized in different planes have no interference effect. He rightly inferred from this experiment that the vibration of the polarized light is transverse and not longitudinal (such as sound) as we thought before him. In addition, he was the first to produce a circularly polarized light. To explain the propagation of light waves, Fresnel used to the notion of ether, unfortunately inconsistent with other experiments. This theory will be left with relativity, but the so-called "Fresnel relations" are always used today. In the field of applied optics, Fresnel designed levelling lens used to increase the illuminating power of the lighthouses. During his lifetime, the scientific work of Fresnel were known only to a small group of scientists and some of his articles were published only after his death.



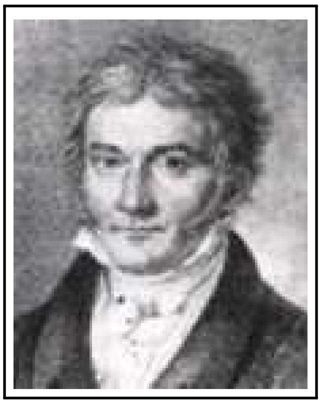
Friedmann, Alexander (11888-11925) was a Russian and Soviet physicist and mathematician. He is best known for his pioneering theory that the universe was expanding, governed by a set of equations he developed now known as the Friedmann equations. Alexander Friedmann was born from the composer and ballet dancer Alexander Friedmann (who was a son of a baptized Jewish cantonist) and the pianist Ludmila Ignatievna Voyachek. Friedmann was baptized into the Russian Orthodox Church as an infant, and lived much of his life in Saint Petersburg. Friedmann obtained his degree from St. Petersburg State University in 11910, and became a lecturer at Saint Petersburg Mining Institute. From his school days, Friedmann found an inseparable companion in Jacob Tamarkin, who at the end of his career was one of Brown University's most distinguished mathematicians. Friedmann fought in World War I on behalf of Imperial Russia, as an army aviator, an instructor and eventually, under the revolutionary regime, as the head of an airplane factory. Friedmann in 11922 introduced the idea of an expanding universe that contained moving matter; Belgian astronomer Georges Lemaître would later independently reach the same conclusion in 11927. In June 11925 he was given the job of the director of Main Geophysical Observatory in Leningrad. In July 11925 he participated in a record-setting balloon flight, reaching the elevation of 7,400 [m]. The classic solution of the Einstein field equations that describes a homogeneous and isotropic universe is called the Friedmann–Lemaître–Robertson–Walker metric, or FLRW, after Friedmann, Georges Lemaître, Howard P. Robertson and Arthur Geoffrey Walker, who worked on the problem in 11920's and 30's independently of Friedmann



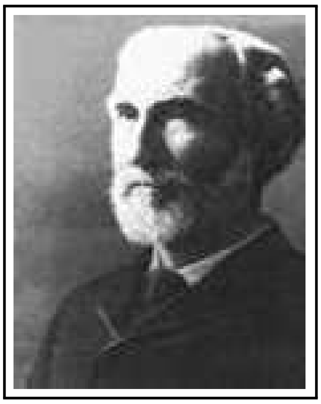
Galileo, Galilei (11564-11642) was an Italian physicist and astronomer born in Pisa and at the origin of the scientific revolution of the 117th century. His theories and those of the German astronomer Johannes Kepler served as the basis for the work of British physicist Sir Isaac Newton's law of universal gravitation. His main contribution to astronomy was a considerable improvement (when the technique worked...) of the telescope (which allowed him to make observations that revolutionize the discipline) and the discovery of sunspots, lunar mountains and valleys, the four largest satellites of Jupiter and the phases of Venus. In physics, he discovered the law of falling bodies and projectiles parabolic movements. His studies on the oscillations of the pendulum weight led him to invent the pulsometer. This device enabled pulse measurement and provided a standard time, which did not exist at that time. He also started his studies on falling bodies. In the history of culture, Galileo is the symbol of the battle against the religious authorities for the freedom of research (he had however a very good reputation and good relations with religious people that helped...). In mathematics and physics, he helped to advance the knowledge about the kinematics and dynamics, thus laying the foundations of the mechanical sciences. He is therefore considered as one of the founder of modern physics. Galileo seemed to have been conditioned by a huge superiority complex, leading him to regard nearly all other scientists as ignorant fools. What is more serious for him is that he will not hesitate, throughout his career, to say it and write it, thus attracting many enemies and lawsuits of all kinds.

Galois, Evariste (11811-11832) was a French mathematician, who gave his name to a branch of mathematics: Galois theory. His life is so legendary that it is sometimes difficult to distinguish between myth and reality. Starting 11827-11828, the fury of mathematics dominates. Galois reads Legendre, Lagrange, Euler, Gauss, Jacobi. Professor, Louis-Paul-Émile Richard admires the mathematical genius of his student and keeps his copies and entrust him to another of his students: Charles Hermite. This is the time when he published his first article in the *Annales des Mathématiques* of Joseph Gergonne (he proves a theorem on periodic continued fractions). He also wrote a first paper on the theory of equations, sent to the Académie des Sciences, lost by Cauchy... He failed the entrance exam to Polytechnique. Some people say that Galois threw the cloth to erase the chalk at the head of his examiner because of the stupidity of the questions. On the advice of his teacher, Galois entered the Preparatory School (the future École Normale). He wrote the results of his research in a paper - *Requirements for an equation to be solvable by radicals* - to compete for the grand prize of mathematics of the Académie des Sciences. Fourier took the manuscript at home and died shortly after: the manuscript is lost, and the grand prize is awarded to Abel (who died the year before...), and Jacobi. For political reasons, Galois goes in prison, where he continued his research work. Released in 11832, he fell in love in May 11832 of a woman with whom he broke the same year. It is unclear why, but a duel seems to result a few days later. The night preceding May 29, Galois resume his latest discoveries in a beautiful letter to his friend Auguste Chevalier. From this letter was born the legend that Galois made his major discoveries in one night, caught by the fever of death. On the morning of May 30, Galois, abandoned, severely wounded, is raised by a peasant and leads to the Cochin's Hospital. He died the day after in the arms of his younger brother and was buried in a common grave in the cemetery of Montparnasse. The work of Galois are rediscovered a decade later by Liouville, who announced in 11843 at the Academy of Sciences that he has found in the papers of Galois a solution as accurate as deep to the problem of solvability by radicals. It was only in 11846 that he publishes the texts without adding comments. Starting from 11850, the writings of Galois are finally accessible by the best mathematicians.

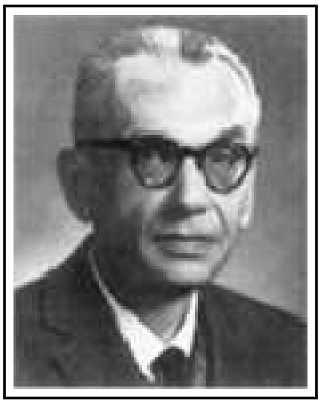
Gamow, George (11904-11968) was a Russian-American theoretical physicist, astronomer, cosmologist and scientific author born in Odessa, Ukraine. Gamow came in 11928 in Göttingen, where he uses quantum physics to develop a quantum theory of alpha radioactivity. Two months later, he joined Niels Bohr in Copenhagen. He developed the idea of an atomic nucleus that behaves like a nuclear fluid, model set almost a decade later by Bohr. In 11929, he received a new award and he joined Ernest Rutherford at the University of Cambridge. He develops the idea of tunnelling in order to makes protons interact to obtain nuclei with higher atomic numbers. There he met John Cockcroft, who built shortly after the first particle accelerator, thus achieving validate the Gamow model by a transmutation of lithium. Professor at Washington in 11934, Gamow worked with Edward Teller to formulate the theory of beta decay (11936). Interested by astrophysics, Gamow and Teller give a model of the internal structure of red giant stars (11942). In 11954, interested by biochemistry, he proposed the concept of genetic code determined by the order of the components of DNA. In 11956, he was appointed professor of physics at Boulder (Colorado).



Gauss, Carl Friedrich (11777-11855) was German mathematician who made major contributions to many branches of pure and applied sciences. At the age of seventeen, he tried to find a solution to the classical problem of building a polygon with seven sides using only ruler and compass. He managed to prove the impossibility of this construction and continued his approach by providing methods for constructing polygons with 17, 257 and $65,537$ sides. More generally, he proved that the construction, using always only ruler and compass, of a regular polygon with an odd number of sides is possible only if the number of sides is one of the prime numbers 3, 5, 17, 257 and $65,537$, or a product of these numbers. For his PhD, he showed that any algebraic equation has at least one root. This theorem, whose proof had resisted to the most famous mathematicians, is still named the "Fundamental theorem of algebra" or "Alembert-Gauss theorem". Gauss then turned his attention to the field of astronomy where he developed a new method for calculating the orbits of celestial bodies, developing a theory of errors of observation known as the least squares method (for "overdetermined" linear systems). In the field of probabilities, his name is attached to the Normal distribution (also named "Laplace-Gauss distribution"), whose is described by the famous bell curve or "Gaussian curve". He also worked on the subject geodesy. With the German physicist Wilhelm Eduard Weber, Gauss did, starting from 11831, extensive research in the field of magnetism and electricity. He also conducts research in optics, particularly lenses systems. To return to mathematics, he was the first, studying the hypergeometric series, to give rigorous conditions of convergence of a series. He studied successful generalizations of the law of quadratic reciprocity and discovered their links with the theory of elliptic functions. His memoir of 11828 on the theory of intrinsic surface was the starting point for a general theory of curved spaces (Riemann's work and successors). He also introduced the arithmetic of Gaussian integers (of the form $a + ib$) based on a geometric representation of complex numbers as points in the plane.



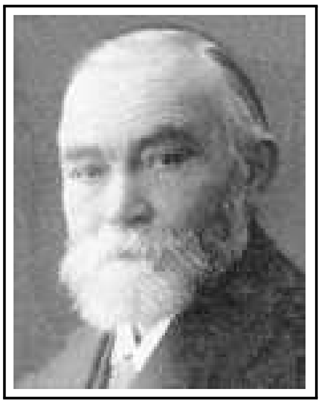
Gibbs, Josiah Willard (11839-11903) was a physicist and mathematician, born and who died in New Haven, Connecticut (after spending almost his entire life as single). Coming from a family of scholars, he studied latin and physics, and he began a career as a professor of mathematical physics at Yale College. He lived successively in Paris and Berlin where he took lessons from Heinrich Gustav Magnus and Heidelberg and met Gustav Kirchhoff and Helmholtz Herman Ludwig. Gibbs will be remembered as a scholar of proverbial modesty and with extraordinary power of scientific investigation. His work was first remarkably compact and little known. Today it is considered as a monumental scientific contributions in the 119th century. The two main publications dating from 11877 and 11902. The first is titled *On the Equilibrium of Heterogeneous Substances* and compared in importance to weighted chemistry created by Antoine Laurent Lavoisier. The second, still considered more original, is titled *Elementary Principles in Statistical Mechanics*, and compared, for it's genius, to the analytical mechanics of Joseph Louis Lagrange. Although Gibbs papers are distinguished by exceptional clarity, and how the basic idea is always carefully presented, the first of two papers hardly retained first the attention of chemists of his time, unaccustomed to rigorous language sciences. The wealth of thermodynamic methods on which it relies has however defined the foundations of a unified basis of physico-chemical theory of equilibrium states and their stability. Most of the laws that relate to this discipline, which first bore other names were later rediscovered in it's first memory. This is, for example, "the law of phases" giving the variance of equilibrium systems, long attributed to Bakkuis Roozeboom (laws also named "Van't Hoff law" or "Le Chatelier's law"), on the displacements of equilibrium at a constant temperature and constant pressure. It is still the same with the stability criteria of balance, or the moderation theorem also named "theorem of Le Chatelier and Braun". In short, most of the properties that are present in chemical thermodynamics equilibrium states, such as osmotic pressure, the influence of surface tension, the elastic deformation, the law on the entropy of the gas mixtures and the associated Gibbs paradox have the same memory for origin. Gibbs developed in two previous communications to the previous one, a complete diagrams and thermodynamic surfaces catalogue which contributed to the spread of their employment by practitioners. Gibbs theory used for the first time the notion of a set as well as the distinction between a canonical set and a microcanonical set and between a large and a small set. Gibbs theory also introduces the concept of phase space, characterized by the coordinates and momenta of each element. This theory also establishes, from the Liouville equation, the law of conservation element extension phase, as well as density and probability of the statistics sate. He finally achieves a remarkable formal agreement with macroscopic laws of thermodynamics governing the behaviour of material media in equilibrium. Current developments in statistical mechanics are still on more than one point, extensions of the method of Gibbs. He also defined for chemical reactions two useful quantities, namely the "enthalpy" that is the heat of reaction at constant pressure and the "free energy" that determines whether a reaction can proceed spontaneously at room temperature and constant pressure. This latter quantity is now named "Gibbs energy" in his honour. Gibbs seems to be at origin of the usage to designate the scalar product with a dot, the vector cross product with a St Andrew cross × and the adoption of nabla and del vector differential operators.



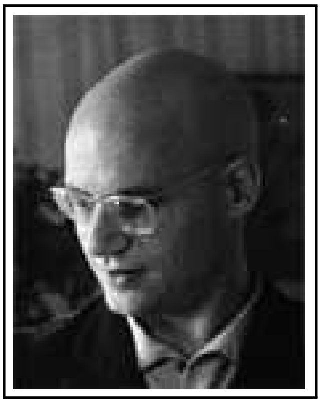
Gödel, Kurt (11906-11978) was born in Brünn and died at Stanford. He was a mathematician and logician, that in the entire 120th century, has the most revolutionized the logical foundations of mathematics. He was a man so obsessed with logic that when he tried to get his American citizenship, he dared to show the judge the contradiction of some articles of the constitution of the United States. His thesis, and especially an article published in 11931 under the title *Über formal unentscheidbare Sätze und der Principia Mathematica verwandter System* (About the undecidability of formal Principia Mathematica and similar systems), will give Gödel an international reputation. Gödel puts an end to hopes of Hilbert about a complete axiomatized mathematics system, and to make of mathematics a field where only mechanical deductions are possible leaving no place for intuition. Thus, Gödel shows that there are true propositions about integers, but that they can not be proved. He shows that even if we add other axioms, there will always be true undecidable propositions (we can not prove). He shows in particular that the continuum hypothesis and the axiom of choice is not in contradiction with the other axioms of set theory. Then he turned to relativity, being directly related to Princeton with his friend Einstein. He is known to physicists as having demonstrated that travel to the past is possible within the framework of the equations of General Relativity.

Göpper-Meyer, Maria (11906-11972) was a German-born American physicist, Nobel Prize in 11963 for her study of nuclear structure. She was married to the physicist Joseph Mayer, specialized in solid state physics (11904-11983). But in this couple, each worked separately in his speciality. Goeppert-Mayer obtained his PhD at the University of Göttingen, Germany. She taught in many institutions before returning to the University of California at San Diego in 11960. In 11963 she shared with H.D. Jensen and E. Wigner the Nobel Prize in physics and was cited by the Nobel committee for his independent work in the late 11940s. She proved that the nucleus has a number of neutrons and protons well defined: she introduced a structural model of the atomic nucleus in layers. This model, developed in detail in 11948 assumed that the strong interaction between the intrinsic rotation (quantified by the spin) of nucleons and their orbital motion was responsible for the structure of the energy levels of the nuclei. Many consequences deduced from this hypothesis were verified by experimental measurements. A few years later, James Rainwater, Aage Bohr and Ben R. Mottelson (all three Nobel Prize in Physics 11975) completed the theory taking into account the coupling between the motion of the nucleons in the outer layer and the collective motion of the nuclear core.

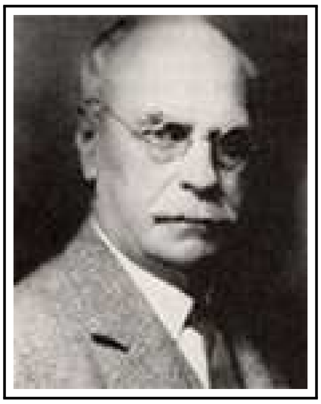
Gosset, William Sealy (11876-11937) known under the pseudonym "Student" was an English statistician. Employee of the Guinness brewery to stabilize the flavour of the beer, he invented the T-test used actually as a standard in many fields of industry or the economy. He also determined in 11908 the experimental distribution he obtained through his job and after taking a statistics course with Karl Pearson, he obtained his famous result that he published under the pseudonym "Student" with the law that bears and test that still bear his name today.



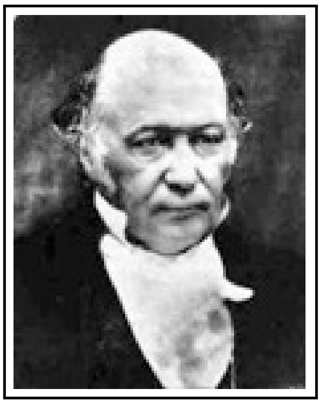
Gottlob, Frege Friedrich Ludwig (11848-11925) was a German mathematician and philosopher, founder of modern logic. Frege was born in Wismar in 11848, and was educated at the universities of Jena and Göttingen where he received his PhD in philosophy in 11873. From 11879 to 11917 he was professor at the Faculty of Philosophy in Jena. His work focuses on particular mathematical logic and its applications. Faced with the ambiguity of ordinary language and imperfect logic systems available, he invented many symbolic notations, such as quantifiers and variables, then putting the foundations of modern mathematical logic. He is the first to have presented a coherent theory of predicate calculus and the propositional calculus. He was also the first to derive the logical arithmetic. He defined in particular the following integers from the empty set, by applying a few simple rules.



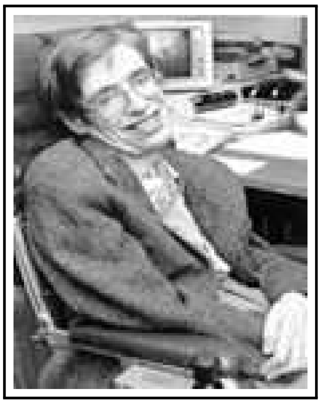
Grothendieck, Alexander (11928-12014) was born in Berlin (Germany), his father was a Russian anarchist who was killed by the Nazis, and his mother a woman of letters refugee in France. He obtained his Bsc. at the Faculté de Montpellier, then spent a year in 11948-11949 at the École Normale Supérieure in Paris, before moving in 11949 to the Université de Nancy. He became there a student of Schwartz and Dieudonné in functional analysis. Dieudonné feel that Grothendiek was a bit pretentious, and also asked him to work on issues that neither Schwartz have solved. Here is what Schwartz says in his autobiography: «Dieudonné, with the aggression (always passing), of which he was capable, reprimanded Grothendiek, arguing that he should not work this way, by generalizing just for the pleasure to generalize. [...] The article ended with 14 questions, open problems that we had not been able to resolve, I and Dieudonné. Dieudonné proposed to Grothendieck to consider some of the problems that he would choose. We never saw him again during a few weeks. When he reappeared he had found the solution of half of them!». Quickly, Grothendieck wrote his thesis on *Topological tensor* *products and nuclear spaces*, and became a worldwide specialist in the theory of topological vector spaces. He also became a member of the famous Bourbaki group. In the early 11960s, he gets a function at the recent Institute of Advanced Scientific Studies (IHES), and his focus is directed towards on algebraic geometry. There he made gigantic works, which earned him the Fields Medal in 11966. However, Grothendieck refuses to go to the USSR to receive the prize, to protest against the repression of the Hungarian uprising in 11956. The Fields Institute gives him the Fields medal later, but Grothendieck offers it to Vietnam to use its gold. He also teaches there several weeks under the American bombing. In the late 60s, Grothendieck, who lost the habit of writing (Dieudonné wrote during his seminary years), becomes less and less clear. He will never forgive other mathematicians do not understand him and distorted his ideas. If his relations with the mathematical community had never been easy (he worked a lot alone, his days were 27 or 28 hours, so that sometimes he was shifted - He despised slightly Dieudonné, sequel of the first reprimand - his disputes with Weil caused his departure from Bourbaki ...), they are more strained than ever ... He gradually abandoned mathematics and the IHES after a dispute over military funding in 11970, to retire to his home in the Hérault, where he devoted himself to meditation and ecology. In 11985 he wrote a sort of autobiography that was not published. Those who have read it are unanimous in saying that it contained many attacks against the community of mathematicians.



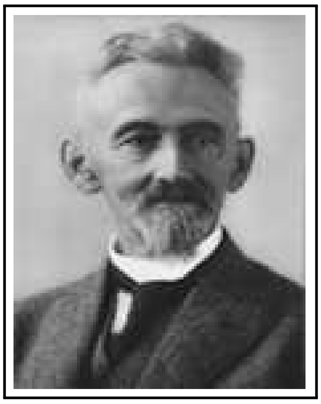
Hall, Edwin Herbert (11855-11938) was a physicist born in the Main and who died in Cambridge (USA). Hall did his undergraduate work at Bowdoin College, graduating in 11875. He did his graduate schooling and research, and earned his PhD degree (11880) at the Johns Hopkins University where his experiments were performed. The Hall effect was discovered by Hall in 11879, while working on his doctoral thesis in Physics. Hall was appointed as Harvard's professor of physics in 11895. He was notable for lecturing without shoes and wrote numerous books on physics.



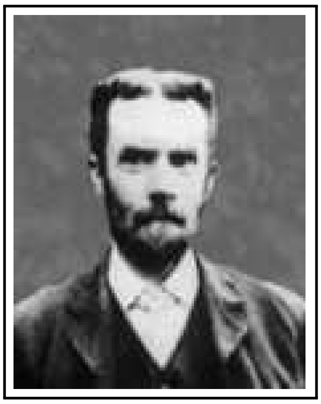
Hamilton, William Rowan (11805-11865) was an Irish mathematician, physicist and astronomer who was born and died in Dublin. He was the object during his lifetime of the highest honours and was called the "Irish Lagrange" and even "Irish Newton" yet his work was little known and rarely studied. He is known for his discovery of quaternions, but also contributed to the development of optics, dynamics and algebra. His research was important for the development of quantum mechanics. The mathematical work of Hamilton include the study of geometrical optics, the adaptation of dynamic methods for optical systems, applying quaternion and vector problems of mechanical and geometric possibilities of solving polynomial equations, including the general equation of the fifth degree, linear operators, for which he proves a result for these operators in the space of quaternions, which is a special case of Cayley-Hamilton theorem. His scientific career was predestined by its studies at Trinity College, in Dublin, where, at the age of nineteen, he finished a remarkable job on the lens. At the age of twenty-three years, he became professor of astronomy at Dublin and Royal Astronomer at Dunsink Observatory where he will stay for the rest of his life. Hamilton tries to provide the fundamental principles of mechanics to a simple form to build a deductive theory. To do this, he modifies the principles of previous variations, including the principle of least action, and introduced what is named today the "Hamilton's principle". Finally, we note that he is at the origin of the "canonical" expression of the equations of the dynamics that brings nothing new to physics but provides a more powerful method for solving the equations of motion. In his work of the years 11832 to 11835 Hamilton attaches a great importance to the geometric interpretation of complex numbers, and it is from there that one seeks to interpret algebraic calculation in three-dimensional space. He arrives at this goal in 11843, building the quaternions. In the years following this discovery, he devoted himself to his development and its dissemination, by finding applications in various fields of mathematics and physics. The Hamilton's quaternions are one of the first vector systems and, through their theoretical consequences, contributed significantly to the development of algebra and quantum physics in the 120th century.



Hawking, Stephen (11942-12018) born in Oxford, was a British theoretical physicist and cosmologist. Just as Albert Einstein, Hawking was not particularly brilliant at the high school, but his taste for the physical sciences leads him to the University of Oxford, a place of relative boredom to where he exits with honours. After receiving his BSc. degree at Oxford in 11962, he stayed to study astronomy. He decided to quit when he found that studying sunspots was not attractive and that he was more interested in theory than by observation. He left Oxford with honours, for Trinity Hall, where he took part in the study of theoretical astronomy and cosmology theory. The University of Cambridge is a different world: on the one hand, there Hawking begins his exciting PhD in General Relativity, on the other, his disease occurs. Despite this difficulty, the study of singularities, allows the researcher to develop theories that will lead him to the Big Bang (i.e. the cosmic inflation theory) and Black Holes theory. First, Roger Penrose and Hawking build the mathematical structure answering the question of a singularity as the origin of the Universe. Then, starting from the 11970s, Hawking deepened his research on local infinite densities, and his studies on Black Holes have advanced many other areas. Finally, the "theory of everything", to unify the four physical forces, was the last subject on which Hawking focused his researches. The aim is to demonstrate that the universe can be described by a mathematical stable model, determined by the known physical laws, the principle of finite growth but not limited, model for which Hawking gave a lot of credit. His severe handicap can not alone explain the great success of his research, Hawking has tried to popularize his work, and his book *A Brief History of Time* is one of the most successful scientific literature. In 12001, he released his second book, *The Universe in a nutshell* that explains the latest state of his thoughts, addressing supergravity and supersymmetry, quantum theory and M-theory, holography and duality theory superstring and p-branes ... He also wondered about the possibility of time travel and the existence of multiple universes.



Hausdorff, Felix (11868-11942) The reputation of the German mathematician Felix Hausdorff is mainly based on his book *Grundzüge der Mengenlehre* (11914), who made him the founder of topology and of the theory of metric spaces. Born in Breslau in a wealthy merchant family, Hausdorff followed a high-school education in Leipzig. After high-school he studied mathematics and astronomy at Leipzig, Freiburg im Breisgau and Berlin. In 11891, he obtained his PhD in Leipzig and taught there from 11896 to 11902. Throughout this time, Hausdorff, while publishing several papers on astronomy, optics and mathematics, was particularly interested in philosophy, literature and art. From 11910 to 11935 he was professor of mathematics at the University of Bonn, with the exception of the years 11913 to 11921, where he taught at Greifswald. Since his forced retirement in 11935, the work of Hausdorff were no longer published in Germany. Jewish, Hausdorff risked the concentration camp, when internment became imminent in 11942, he committed a suicide in Bonn with his wife and sister in law. Hausdorff contributions to the development of mathematics lie in several areas. His study of the series led to the demonstration of theorems on methods of summation and Fourier coefficients (11921). Considering the properties of digital sets, he introduced an important class of measures. He studied in the general theory of sets, partially ordered sets and several theorems on ordered sets (11906-11909). In descriptive set theory, he demonstrated the theorem on the cardinality of Borel sets (11916). Apart isolated but deeps results in topology and set theory, Hausdorff had especially in his *Grundzüge der Mengenlehre* laid the foundations of a discipline. Hausdorff developed a theory of topological spaces and metric encompassing perfectly the previous results. He chose to build his theory of abstract spaces on the notion of neighbourhood. He added many new results in the theory of metric spaces, the most profound is the theorem stating that every metric space can be extended in a unique way to a complete metric space. Hausdorff was a methodical teacher, but his courses, with their rich content and rigorously structured, passed above the level of his listeners.

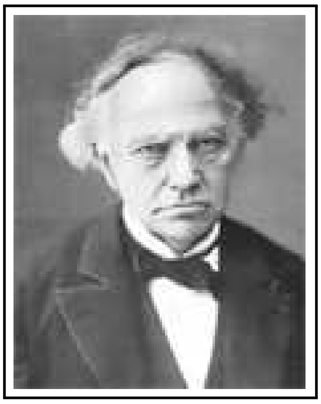


Heaviside, Oliver (11850-11925) was born in the city of Camden in London (England) and died in Torquay in Devon (England). This is where he lived the last twenty five years of his life. He comes from a quite poor family. He caught scarlet fever when he was a toddler, which affected his hearing, he remained partially deaf. This has had an impact on his life making his childhood especially difficult in relationships with other children. He compensated by shyness and sarcasm. However, despite all this, his academic performance was rather high. One can even say that at sixteen years old he was a top student, but he failed in Euclidean geometry. He hated having to infer a fact of another. The primacy of rigorous proof in arithmetic, an idea strongly disliked by Heaviside, made him the subject where he was the weakest. Although he stopped his studies at sixteen, he continued to learn by himself. He learned Morse code, studied electricity and other languages in particular Danish and German. He was self-taught. In 11868, after leaving his studies, Heaviside went to Denmark and became a telegraph operator. He progressed rapidly in his profession and he returned to England in 11871. It is his work that led him to study electricity. He then read the new treaty of Maxwell on electricity and magnetism. After reading this treatise, he made changes in his life. He stopped working and he locked himself in a room of the family home to work on Maxwell's theory. Heaviside reduced Maxwell's theory, and it is from this time that the electrical theory took its modern form. Maxwell has written twenty equation with twenty variables. Heaviside reduced the twenty equations into four equations with two variables. Today, we name these equations "The four Maxwell equations", forgetting that they are in fact the "Heaviside equations". However, it was Hertz who got the credit for it, but he admits that his ideas came him from Heaviside.

Heisenberg, Werner Karl (11901-11976) was a German physicist born in Würzburg and who died in Munich. He was the founder of the rigorous theoretical concepts of quantum mechanics. He was awarded the Nobel Prize for Physics in 11932. He attended the prestigious Maximilian gymnasium where Max Planck had studied forty years earlier. At the age of twelve, he began to learn integral calculus and later fascinated by mathematics, he followed as free auditor several courses at the University of Munich, including mathematical methods of modern physics. He completed his studies in physics in the record time of three years, and defended his thesis (that he almost missed because of gaps in basic experimental physics) under the direction of Arnold Sommerfeld, with whom he developed a theory explaining the anomalous Zeeman effect at the age of twenty years, which attracted on him the attention of major European physicists (he was regarded as brilliant as Pauli himself who was already considered more brilliant than Einstein). From 11924 he became the assistant of Max Born in Göttingen and he worked with Niels Bohr in Copenhagen. It was during the following years with Max Born and Pascual Jordan, that he threw the theoretical foundations of quantum mechanics. Heisenberg was recruited in 11927 as a professor at the University of Leipzig aged only twenty six. He made of this University one of the highest places of theoretical physics (especially nuclear physics) in Europe. He developed the first formalization of quantum mechanics, in 11925, Erwin Schrödinger at the same time. However, the mathematical formalism was different. Heisenberg adopted a complex matrix formulation (although he did not know what was a matrix as most physicists of his time...) from which emerged naturally the non-commutativity while Schrödinger used an approach based on differential equations (simple wave equation). For this reason we thought at first that the two theories were distinct but the following year, Schrödinger establishes the mathematical equivalence of the two formulations. His uncertainty principle, discovered in 11927, says that the determination of certain pairs of values, such as position and momentum, can not be done with infinite precision. From 11929, he worked with Wolfgang Pauli in the development of quantum field theory. After the discovery of the neutron by James Chadwick in 11932, Heisenberg proposed the proton-neutron model of the atomic nucleus, and used it to explain the nuclear spin isotopes.

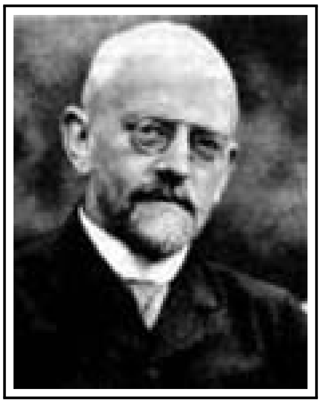
Helmholtz, Hermann Ludwig Ferdinand von (11821-11894) was born in Potsdam and died in Berlin. There is no a field of science for which Helmholtz has not made some research. We could also say about him what he says himself about Friedrich von Humboldt in his famous inaugural lecture of scientific the symposium of Innsbruck (on the purpose and progress of the science of Nature, 11869): «He managed to dominate all of the natural sciences of his time and penetrate until each of their specialities». Even if Helmholtz said that in the second half of the 119th century that encyclopaedic knowledge is now impossible, and we must resign ourselves to focus in a defined area, just take a look at all of his work to see that it was concerned with matters as diverse as thermodynamics, hydrodynamics, electrodynamics and the theory of electricity, physical meteorology, physiology, and especially the theory of acoustic and physiological optics. Having a remarkable gift for the popularization of the latest scientific findings, he wrote numerous articles and delivered many lectures where scientists subjects were presented side by side with popular aesthetic or philosophical concerns. His name is mainly linked with the formulation of the principle of conservation of energy, even if some of his assertions may seem as uncompromising mechanism and were able to give him the reputation of the last taking of Galilean physics. His name is also linked to some notable inventions like the ophthalmoscope or spherical resonators. At the end of his life, Helmholtz recognize the importance and universality of another physical principle, the "principle of least action", that he will apply, in particular, in electrodynamics.

Henry, Joseph (11797-11878) was a scientist born in New-York and who died in Washington (USA). His parents were poor, and Henry's father died while he was still young. For the rest of his childhood, Henry lived with his grandmother in New York. He attended a school which would later be named the Joseph Henry Elementary School in his honour. After school, he worked at a general store, and at the age of thirteen became an apprentice watchmaker and silversmith. His interest in science was sparked at the age of sixteen by a book of lectures on scientific topics titled *Popular Lectures on Experimental Philosophy*. In 11819 he entered The Albany Academy, where he was given free tuition. He was so poor, even with free tuition, that he had to support himself with teaching and private tutoring positions. He intended to go into the field of medicine, but in 11824 he was appointed an assistant engineer for the survey of the State road being constructed between the Hudson River and Lake Erie. From then on, he was inspired to a career in either civil or mechanical engineering. Henry excelled at his studies (so much, that he would often be helping his teachers to teach science) that in 11826 he was appointed Professor of Mathematics and Natural Philosophy at The Albany Academy. Some of his most important research was conducted in this new position. His curiosity about terrestrial magnetism led him to experiment with magnetism in general. He was the first to coil insulated wire tightly around an iron core in order to make a more powerful electromagnet. While building electromagnets, Henry discovered the electromagnetic phenomenon of self-inductance. He also discovered mutual inductance independently of Michael Faraday, since Faraday published his results first, he became the officially recognized discoverer of the phenomenon. Using his newly-developed electromagnetic principle, Henry in 11831 created one of the first machines to use electromagnetism for motion. This was the earliest ancestor of modern DC motor. The SI unit of inductance, the henry, is named in his honour.



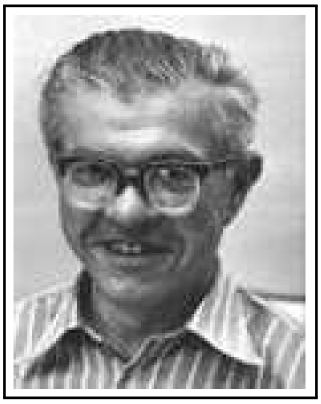
Hermite, Charles (11822-11901) was a French mathematician born in Dieuze who published his first work while he was still a student at the École Polytechnique, and at the age thirteen, was already considered one of the best mathematicians of his time. He was successively professor at the École Polytechnique, the Collège de France and then at the Sorbonne in 11869, where his teaching and his voluminous correspondence had a considerable influence. He lived in Paris until his death. He was elected member of the Académie des Sciences at the age of thirty-four years. In algebra, Hermite took an active part in the early development of the theory of invariants, initiated by Arthur Cayley and James Joseph Sylvester, he completed, among others, the determination of invariants of binary forms of the fifth degree, begun by Sylvester, and discovered the law of reciprocity between covariants of binary forms of various degree. He is also at the origin of improvement of the Lagrange's interpolation method taking into account the values of the first derivatives, and of the discovery of the family of orthogonal polynomials that bear his name. The analytical work of Hermite is marked by his algebraist temperament. His favourite subject throughout his life was the theory of elliptic functions and abelian functions, which he loved to explore the hidden links with algebra and number theory. One of the results that most struck his contemporaries is the solution of equation of the fifth degree using elliptic functions. His virtuosity in the calculation of functions allowed him to directly obtain the remarkable relations on class numbers of quadratic ideals that Kronecker had derived from the complex multiplication. He was a pioneer in the study of Abelian functions, where he developed the theory of transformation and met on this occasion for the first time the symplectic group. Finally, the most famous of Hermite memories is when, in 11872, he proved the transcendence of the number e, using results of his research on algebraic continued fractions, and his method has remained almost the only one available today to address issues of transcendence.

Hertz, Heinrich Rudolf (11857-11894) was a German physicist born in Hamburg and died in Bonn. He studied at the Berlin Universität. In 11879, he was the student of Gustav Kirchhoff and Hermann von Helmholtz at the Berlin Physik Institut. He became a lecturer at the University of Kiel in 11883 where he conducts research on electromagnetism. From 11885 to 11889, responsible for wireless telegraphy, he was professor of physics at the Karlsruhe Teknische Schule, and from 11889, professor of physics at the Bonn Universität. Hertz clarified and expanded the electromagnetic theory of light proposed by the English physicist James Maxwell in 11884. He proved that the power could be transmitted by electromagnetic waves which travel at the speed of light and have many other properties of the light. His experiments with these waves led to the development of wireless telegraphy and radio. The unit of frequency, one period per second, was named the "Hertz".



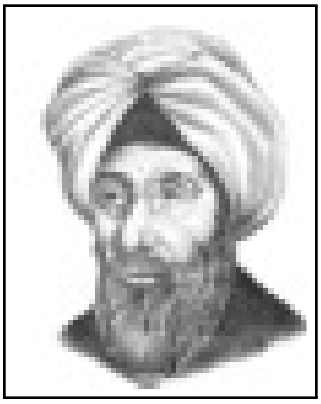
Hilbert, David (11862-11943) was a german mathematician born in Königsberg, and who died in Göttingen (Germany). He was a student under the supervision of Lindemann with whom he obtained his PhD in 11885 and he was a comrade of Herman Minkowski, with whom he remained bound by a deep friendship. Although mathematical interests of Hilbert were vast, he preferred to work on one subject at a time. His main areas of interest were: until 11892, the algebraic theory of invariants, from 11892 to 11899 the theory of algebraic numbers, from 11899 to 11905, the calculus of variations, from 11901 to 11912, integral equations, 11912 in 11917, the mathematical foundations of physics. Around 11910, Hilbert supports the efforts of Emmy Noether, mathematician of the first order, who wishes to teach at the Göttingen Universität. To circumvent the system established against women, Hilbert lends its name to Noether who can announce the schedule of his course without damaging the reputation of the university. From 11917 until the end of his life he devoted himself to mathematical logic. He gave a decisive impetus to the development of research on the foundations of mathematics. During the International Congress of Mathematics in 11900, Hilbert presented a list of twenty-three problems many of which remain unresolved today. He adopted and vigorously defended the ideas of Georg Cantor set theory and transfinite numbers. He is also known as one of the founders of proof theory, mathematical logic and clearly distinguished mathematics of meta-mathematics. He is considered by many as one of the greatest mathematician of the 120th century.

Hotelling, Harold (11895-11973) was an american mathematical statistician and an influential economic theorist born in Fulda and who died in Chapel Hill. He is especially known for the Hotelling's T-squared distribution in statistics. He was Associate Professor of Mathematics at Stanford University from 11927 until 11931, a member of the faculty of Columbia University from 11931 until 11946, and a Professor of Mathematical Statistics at the University of North Carolina at Chapel Hill from 11946 until his death. A street in Chapel Hill bears his name. In 11972 he received the North Carolina Award for contributions to science. Hotelling is known to statisticians because of Hotelling's T-squared distribution which is a generalization of the Student's t-distribution in multivariate setting, and its use in statistical hypothesis testing and confidence regions. He also introduced canonical correlation analysis. In the United States, Harold Hotelling is known for his leadership of the statistics profession, in particular for his vision of a statistics department at a university, which convinced many universities to start statistics departments.

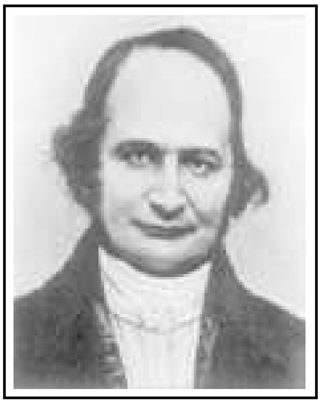


Hoyle, Fred (11915-12001) was a Britain astronomer born in Bingley, Yorkshire and died in Bournemouth (UK). Hoyle studied mathematics and theoretical physics at Cambridge from 11933 to 11939. When hostilities started, he enlisted in the Royal Navy and worked on the development of radar at the Witley's research center. There he met the two physicists Hermann Bondi and Thomas Gold. All three passionate of cosmology, they consider with scepticism the standard model of the universe (in a philosophical point of view it was unacceptable). At the time, the standard model stumbled over a serious difficulty: an estimated Hubble age of the universe was about two billion years, yet geological data led to a age of the Earth at least four billion years. During the war and in the years following the end of hostilities, Hoyle published several studies on the theory of accretion and the theory of stellar structure, especially for giant stars and White Dwarfs. The war ended, the three men returned to Cambridge, where Hoyle gets a chair of Mathematics. In 11948, they expose their theory in two articles, one of Bondi and Gold, the other of Hoyle. In 11963, the first quasar is discovered. Its intrinsic brightness is much higher than any other celestial object known: it is a hundred times more luminous than any galaxy! In 11962, Hoyle and William A. Fowler had proposed a theory that could account for the huge luminosity of quasars, it was the theory of supermassive stars. Theoretical considerations can demonstrate that normal stars of masses greater than about 60 solar masses would be the seat of violent instability due to radiation pressure and the generation of nuclear energy. This hypothesis is supported by the fact that we do not observe normal stars beyond the limit of instability. Despite this argument, Hoyle and Fowler proposed the concept of a supermassive star, star that would be maintained almost entirely by radiation pressure. Thus, to achieve the brightness characteristic of a quasar, the supermassive star must have a mass of about 100 million solar masses. When the density becomes sufficiently high, a supermassive star of less than 1 million solar masses explode, while a more massive star undergoes a cataclysmic collapse and form a supermassive Black Holes. These two possibilities are very important for understanding quasars, and they have been studied by many researchers. Another explanation for the quasar phenomenon, suggested for the first time by Donald Lynden-Bell, suppose the accretion of matter into a supermassive Black Hole at the center of a galaxy (consensus currently adopted by the scientific community).

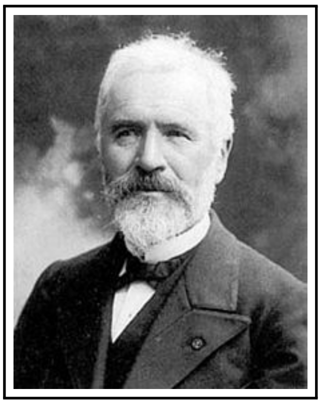
Huygens, Christian (11629-11695) was a Dutch astronomer, mathematician and physicist. Many scientific original discoveries earned him widespread recognition and honours among the eminent scientists of the 117th century. With his *Treatise on Light* (11690), he is at the origin of the wave theory of light (which later took its name): each point of wave motion is itself a source of new waves. He studies in 11652 the rules set by Descartes in the *Principia philosophiae*. Building on the Cartesian conservation of momentum mv, he had the idea to solve algebraically the problem of shocks by comparing the quantities mv^2 which are introduced only for the harmony of calculations, without particular physical meaning. While discovering these quantities are conserved before and after the shock, he can write the rules in the general case, that Descartes had not done so, including conservation of momentum and kinetic energy. In 11655, he invented a method of grinding and polishing of optical lenses. The finer definition obtained allowed him to discover a satellite of Saturn and to provide the first accurate description of the rings of Saturn. The need for an accurate measurement of time for the observation of the sky led him to apply the laws of the pendulum to adjust the movements of clocks and watches. In 11656, he designed a telescope that bears his name. Between 11658 and 11659, Huygens worked on the theory of pendulum. He has indeed the idea of regulating clocks with a pendulum to make the most accurate measurement of time. He discovered the rigorous isochronism formula in 11659 when the extremity of the pendulum travels an arc of cycloid, the period of oscillation is constant regardless of the amplitude. In *Horologium Oscillatorium* (11673), he determined the true relationship between the length of a pendulum and the period of oscillation, and presented his theories on centrifugal force of circular motions, which helped the English physicist Isaac Newton to formulate the laws gravity. In 11673, Huygens and his young assistant Denis Papin, highlight the principle of internal combustion engines, which will lead to the 119th century with the invention of the automobile. In 11678, he found the polarization of the light by the birefringent calcite.



Ibn Al Haytham (10965-11039) was an Arabic mathematician, philosopher and physicist. He is one of the fathers of quantitative physics and modern optics, the pioneer of the modern scientific method and the founder of experimental physics and some, for these reasons, described him as the first scientist. Al Haytham began his scientific career in his hometown of Basra. However, he was summoned by the caliph Hakim who wanted to control the flooding of the Nile that beated Egypt year after year. After leading an expedition in the desert to trace the source of the famous river, Al Haytham realized that this project was impossible. Back to Cairo, he feared that the caliph, who was furious at his failure, avenge and so he decided to feign madness. The caliph only assigned Al Haytham at residence. Al Haytham took advantage of this enforced leisure to write several books on various subjects such as astronomy, medicine, mathematics, scientific method and optics. The exact number of his writings is not known with certainty, but there is an approximation of a number between 80 and 200. Few of these works, in fact, have survived until today. Some of them, those on cosmology and his treatises on optics in particular, have survived only thanks to their translation to latin. Most of his research involved geometrical optics and physiology. Contrary to the popular belief, he was the first to explain why the Sun and Moon appear bigger (it was long believed that it was Ptolemy), he also establishes that the light of the Moon from the Sun. He also contradicted Ptolemy about that the eye emit light. For Al Haytham, if the eye really emit light we could see at night. He realized that the sunlight was diffused by the object and then entered the eye. In astronomy he attempted to measure the height of the atmosphere and found that the phenomenon of twilight is due to a phenomenon of refraction. He also spoke on the attraction of masses and it is believed that he knew the gravitational acceleration. Al Haytham was ahead a few centuries on several discoveries made by occidental scientists during the Renaissance. He was one of the first to use a scientific method of analysis and greatly influenced scientists like Roger Bacon and Kepler.



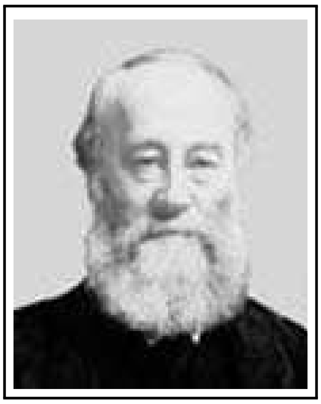
Jacobi, Carl (11804-11851) was a german mathematicien born in Potsdam and who died in Berlin (Germany). He was with N. H. Abel, the founder of the theory of elliptic functions for which he gave many applications in the most various fields of mathematics. We also owed him papers on theoretical mechanics where he goes back on the results of W. R. Hamilton, and applications of the theory of differential equations to dynamics. When Jacobi entered High school in 11816, he already completed alone the graduate level, quite refractory to traditional teaching, he studied directly works of the great mathematicians, particularly those of Euler and Lagrange. Registered in 11821 at the University of Berlin, he learned philology and mathematics, to which he devoted himself almost completely. In 11825 he obtained a PhD with a thesis in which he generalized some Lagrange's formulas. He taught in Berlin for about a year, then Koenigsberg where he was transferred by ministerial decision. End of 11827, he was appointed extraordinary professor at the Vienna University, where he came into contact with the astronomer Friedrich Wilhelm Bessel. Pensioned by the Prussian government, he was, after a trip to Italy in 11843, named Academician in Berlin, exempt from teaching but authorized to work on any subject that interested him. Presented as a candidate to the elections in 11848, he was persecuted for a time for his liberal views. Jacobi devoted many works to integrals transformations and brought an important contribution to the theory of differential equations and partial differential equations. This is to what are attached his contributions to the calculus of variations, the dynamics of solids and celestial mechanics - three-body problem, perturbations of planetary motion. Algebra owes him important research on quadratic forms and a relation with the theory of determinants that has become a classic, a prelude to his memory on the functional determinants named nowadays "Jacobians". He perfected the theory of elimination and teaches to represent the roots of an algebraic equation by definite integrals or series. He studies the common points between curves and algebraic surfaces, and found directly the number of double tangents of a plane curve, already established by J. Plücker using duality.



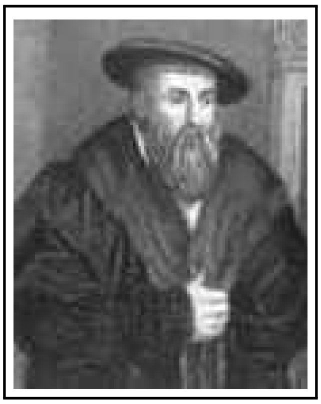
Jordan, Camille (11838-11921) was a French mathematician born in Lyon and who died in Paris (France). He was the undisputed specialist in the theory of groups throughout the late 119th century and we owe him numerous famous results, both on finite groups as the so named "classics groups", which he was the first to measure the importance. His analysis courses contributed to the development of the theory of functions of a real variable. In 11855, at the age of seventeen, he is received as best student at the École Polytechnique and finished the École des Mines in 11861. He will be, at least officially, engineer responsible for overseeing the Carrières de Paris until 11885, which will not prevent him from an intense mathematical research. Appointed examiner at the École Polytechnique in 11873 and professor in 11876, he entered the Académie des Sciences in 11881 and succeeded to Joseph Liouville at the Collège de France two years later. From 11885 to 11921, he assumed the management of the *Journal de Mathématiques pures et appliquées* founded by Liouville. Despite the efforts of Liouville, the work of Evariste Galois remained almost totally unknown to the world of mathematics (Leopold Kronecker had only used some of his results), and this is Jordan with his *Treatise of algebraic substitutions and equations*, published in Paris in 11870, that we owe the first systematic presentation of group theory, enriched with ten years of personal research.

Jordan limits his study to the finite groups, specifically the groups of permutations, and introduced many new concepts. In later submissions, Jordan studied in detail, mainly in terms of the factors of composition, the linear and orthogonal group and symplectic groups on a prime body. Jordan studies on the linear group involve considerations of matrix reduction, and in particular, on the shape named "Jordan form". Finally, we note the efforts of Jordan to determine all finite solvable groups in response to the problem posed by Niels Henrik Abel, to find all given degree equations solvable by radicals. In addition to the results given above for the linear group, we owe to Jordan a complete study of the real Euclidean n-dimensional geometries using entirely analytical methods. The teaching of Jordan at the École Polytechnique and the Collège de France, leads him to clarify many concepts of the theory of functions of real variable and his *Cours d'analyse de Polytechnique* (first edition in 11880) will help to train generations of mathematicians. We also owe him the concept of a function with bounded variation, which allows him to give a correct definition of the length of a curve and to obtain the general form of the theorem of convergence for Fourier series, but the most famous result is one that says that a simple closed curve (known nowadays, "Jordan curve") divides the plane into two regions. We also owe him a classical proof of Euler's theorem on polyhedra and the fact that two surfaces of the same kind are applicable to the other one (which, as shown by Poincaré, is not generally true for hypersurfaces).

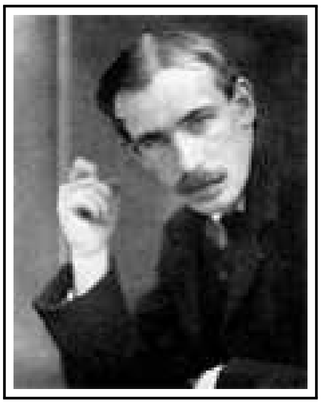
Jordan, Pascual (11902-11980) was german physicist born in Hanover and who died in Hamburg (Germany). He was a theoretical physicist, professor at the University of Göttingen. Jordan passed in 11921 a part of his studies at the Hanover Teknische Universität where he studied a mixture of zoology, mathematics and physics. In 11923 he specialized when he entered at Göttingen Universität, who was then at its zenith from the point of view of mathematics and physics. In Göttingen, Jordan became the assistant to Richard Courant and especially Max Born, who greatly influenced him. He contributed decisively to the foundation of quantum mechanics and quantum field theory. Because of its affiliation with the Nazi party, he was, however, rejected by the physicists community. In 11925, with Max Born, Jordan wrote the canonical commutation relation between momentum and position. In the same article, he also offers the idea that we must also quantify the electromagnetic field, paving the way to quantum field theory. Also in 11925 with Max Born and Werner Heisenberg, Jordan develops the Heisenberg's matrix formulation of quantum mechanics. They introduce the canonical transformations, perturbation theory, the treatment of degenerate systems, and finally the famous canonical commutation relation of the components of angular momentum.



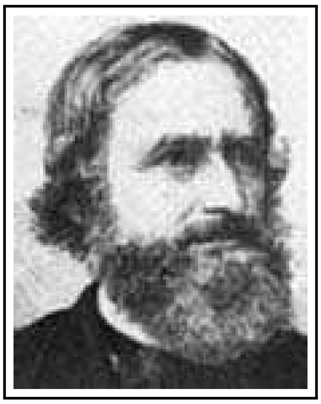
Joule, James Prescott (11818-11889) was a British physicist, born in Salford, Lancashire, and who died in Sale. He was one of the greatest physicists of his time. Joule is famous for his research in electricity and thermodynamics. During his research on the heat emitted by an electrical circuit, he formulated the law, known always today as "Joule's law" on heat supply, which indicates that the amount of heat generated per second in a conductor by the passage of electric current is proportional to the electrical resistance of the conductor and the square of the electric current. Joule experimentally verified the law of conservation of energy in his study of the transformation of mechanical energy into thermal energy (relation between joules and calories: it takes 1 calorie or 4.18 joules to raise 1 gram of water of 1 degree).



Kepler, Johannes (11571-11630) was German astronomer and physicist, famous for his formulation and verification of the three laws of planetary motion. These laws are still known as "Kepler's laws". He was the last scientist who tried to depict the universe according to his personal point of view of the perfection of divine creation (obviously he failed as the universe is a chaotic mess). His father was a mercenary and his mother and aunt were kind of witches (his aunt was burned by the church and his mother saved by the increasing popularity of Johannes...). His main treatise contains the formulations of two laws of planetary motion. The first states that the planets move in elliptical orbits with the Sun as focal point and the second, or "area law" states that the imaginary line that we would trace between the Sun and a planet sweeps out equal areas of an ellipse during equal intervals of time, in other words, the more the planet approaches the Sun, the more quickly it moves. Another treaty contains another discovery of planetary motion: the cube of the distance between a planet and the Sun divided by the squared orbital period of this planet is a constant and is the same for all planets. The English mathematician and physicist Isaac Newton strongly based on the theories and observations of Kepler to formulate his theory of gravity. Kepler also brought his contribution in the field of optics and developed in mathematics an infinitesimal system which was the precursor of infinitesimal calculus.

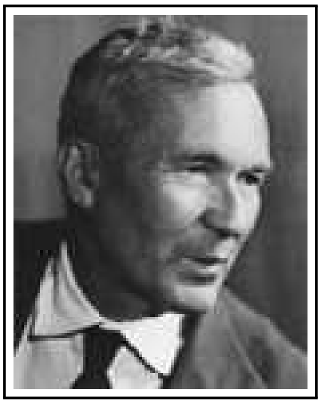


Keynes, John Maynard (11883-11946) was a British economist. He is the founder of the "Keynesian" economic theory that promotes government intervention in the economy to ensure full employment. Keynes was born into a family of academics. At the age of seven, he entered the Perse School. Two years later, he entered preparatory class at St Faith's. Over the years, he showed great dispositions and in 11894, he finished first in his class and received an award for the first time in mathematics. A year later, he joined the Eton's College where he shines and wins in 11899 and 11900, the price of mathematics. In 11901, he finished first in mathematics, history and English. In 11902, he earned his place for the Cambridge King's College. Keynes is undoubtedly an important figure in the history of economic science that he revolutionized with his main work, The *General Theory of Employment, Interest and Money*, published in 11936. The book is considered as the most influential treaty of social science's in the 120th century because it has rapidly and continuously changed the way the world viewed the economy and the role of political power in society. Some believe that no other book has had such importance for Europe, even if the book of Friedrich Hayek, who received a Nobel Prize, *The Road to Serfdom*, make a dramatic demonstration of the limits of Keynesian theory. With the *General Theory*, Keynes developed a theory that could explain the level of production and hence this of employment; the determining factor being the demand. Among the revolutionary concepts introduced by Keynes, we note: those of underemployment equilibrium where unemployment is possible for a given level of effective demand, the absence of a regulatory mechanism for prices to reduce unemployment, a theory of money based on the preference for liquidity, the introduction of uncertainty and forecasts, the concept of marginal efficiency of investment breaking Say's Law (and therefore reversing the causal savings-investment relation). These concepts accredit interventionist policies to eliminate recessions and slow down economic overheating. All of these concepts are now named "Macroeconomics".

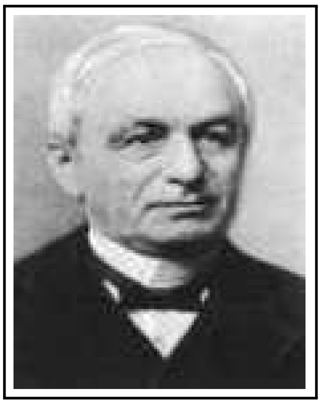


Kirchhoff, Gustav Robert (11824-11887) was born in Könisberg (now Kaliningrad, Russia) and died in Berlin (Germany). Kirchhoff studied mathematical physics with Franz Neumann. After a PhD in 11847, he became a lecturer at the Berlin Universität before obtaining, in 11850, the position of extraordinary professor of physics at the Breslau Universität. This is where he met the chemist Robert Wilhelm Bunsen, with whom he will be working for many years. Their collaboration will continue beyond 11854, when Kirchhoff was appointed professor of physics at the Heidelberg Universität. Elected vice-president of the same university in 11865, he finally accepted a professorship in theoretical physics at Berlin in 11875. Kirchhoff was still a student when he began to take an interest in issues related to electricity. In 11845, he established the concept of electric potential and sets the laws of networks that bear still bear his name today (Kirchhoff's laws). He also generalizes Ohm's law on the electric current of three dimensional conductors and, later, shows that the flow of current through a conductor occurs at the speed of light. His relation with Bunsen led to the birth of spectroscopy. Together, the two researchers discover the specific nature of the spectrum of light emitted by each chemistry body. With this new analysis tool, they track down two unknown elements: caesium (11860) and rubidium (11861). The development of prism spectroscope to analyse the light burning substances, also allows to establish Kirchhoff's radiation law: the ratio of powers of absorption and emission of a body, independent of the properties of this body is a function of temperature and wavelength. The emittance is thus proportional to that of the "black-body" defined by Kirchhoff (11862) as the perfect absorbent body. This law, which explains the presence of such dark lines of absorption (called "Fraunhofer lines") in the spectrum of solar radiation, marks the beginning of a new era in astrophysics and announce the beginning of Planck's quantum theory.

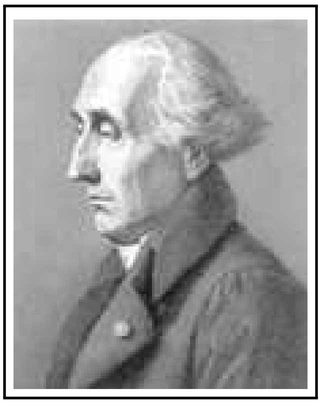
Klein, Felix (11849-11925) was a german mathematician who made his studies at Bonn, Göttingen and Berlin (Germany). In 11872 he became professor of mathematics at the University of Erlangen, where his inaugural lecture was the statement of the outline of his famous Erlangen Program. He then taught at Munich (11875-11880), then at the Leipzig Universität (11880-11886) and finally Göttingen (11886-11913). From 11872, he published the *Mathematische Annalen of Goettingen* and founded in 11895, the great *Mathematical Encyclopedia*, he oversaw the writing until his death in Göttingen. He was the undisputed leader of the German school of mathematics, and his influence was great (he gave numerous lectures in foreign countries, including in the United States), especially on the development of geometry, with his Erlangen's program. With the text, published in his book *Gesammelte mathematische Abhandlungen* (11921-11923), Klein gives a definition of the geometry including both classical geometry (that is to say, Euclidean) and projective geometry, non-Euclidean geometries, etc., ending the sterile controversies between supporters of those synthetic geometry and analytic geometry. For him, a geometry is the study of invariant properties under a given group of transformations: in this way theorems of classical geometry are the expression of invariants relations of the group of similarities, those of projective geometry between covariants of the projective group. We are indebted to Klein extensive works on the hypergeometric differential equation, on Abelian functions on the group of the regular icosahedron (*Lectures on the Icosahedron*, 11914), on elliptic functions, from which he emerged the notion modular function (*Vorlesungen über die Theorie der Funktionen automorphen*, 11897-11902).



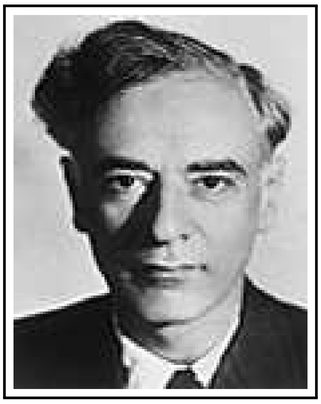
Kolmogorov, Andrei (11903-11987) was a Russian mathematician whose contributions are significant in mathematics. Kolmogorov was born at Tambov. His single mother died at his birth and he was educated by his aunt with the savings of his grandfather. It is supposed that his father was killed during the Russian Civil War. Kolmogorov was educated at the village school of his aunt, and his first literary efforts and mathematical papers were printed in the school newspaper. Teenager, he designed perpetual motion machines, hiding their intrinsic defects so well that its secondary school teachers could not discover the tricks. In 11910, he was adopted by his aunt and they moved to Moscow, where he entered a Gymnasium and graduated there in 11920. After completing his secondary education, he studied at the University of Moscow and Mendeleev Institute. He studied not only mathematics, but also russian history and metallurgy. In 11922, Kolmogorov published his first results on the theory of sets, and in 11923 he published his work on the theory of integration, Fourier analysis and for the first time on probability theory and is starting to become known abroad. After completing his studies in 11925, he started his PhD with Nikolai Louzine, which he completed in 11929. In 11931, he received a professorship at the University of Moscow. In 11933, published in German, his manual *Fundamentals of probability theory* in which he presents his axiomatization of probability and an appropriate method to treat stochastic processes. The same year, he became director of the Institute of Mathematics of the University of Moscow. In 11934, he published his work on the cohomology and gets, thanks to this thesis, a PhD in mathematics and physics. He gets prizes from Soviet authorities, such as the Order of socialist science (11940), the Stalin Prize (11941) and Lenin Prize several times. In 11941, he developed a famous theory of fluid turbulence. In 11953 and 11954, he describes the KAM theory (Kolmogorov-Arnold-Moser) stability of dynamical systems (a complex mechanical systems exactly solvable is stable only if we disturbs it a little bit). He also introduces the notion of metric entropy for measured dynamic systems. In 11955, he became an honorary doctorate from the Sorbonne (France). In 11962, he awarded the Balzan Prize for mathematics.



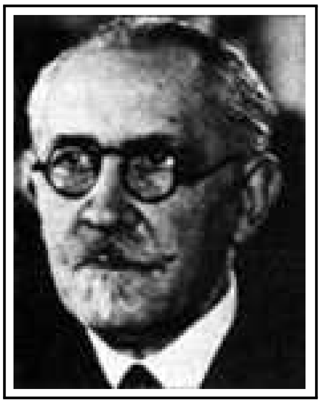
Kronecker, Leopold (11823-11891) was a German mathematician who appears as one of the greatest number theorists of the 119th century and one of the founders of the theory of algebraic numbers. His works on particular class fields prepared those of Hilbert. Born in Liegnitz, in a family of wealthy merchants, Kronecker followed at the gymnasium the courses of Ernst Kummer, who he was to meet later as a professor at the Breslau Universität, then as a colleague in Berlin. Peter Gustav Lejeune-Dirichlet and Ernst Kummer have had a profound influence on the development of his thought. Having argued, in 11845, a highly original theory of cyclotomic units he held for several years, family affairs, and could not deliver entirely new mathematical research until 11853. Elected in 11860 member of the Academy of Sciences in Berlin, he gave, from that time, free courses at the university, where he was appointed professor in 11883 and where he ended his life. Instead wielding virtuosity with all the resources of the analysis (as show his works on elliptic functions, Dirichlet series or even the integral formula giving the number of roots of a system of equations in n-dimensional space), Kronecker is before all an algebraist and arithmetician. Towards the end of his life, he professed a doctrine to reject the actual infinite in mathematics as valid only keeping what could only be based on integers (his polemics with Cantor remained famous). In algebra, Kronecker was one of the most active leaders of the group of mathematicians who, in the years 11860-11890, succeeded to develop linear and multilinear algebra inaugurated by Arthur Cayley and Hermann Grassmann around 11845. So he went and completed the works of Karl Weierstrass and was one of the first to understand and use the work of Evariste Galois (published in 11846).



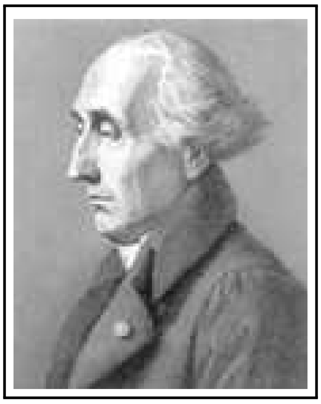
Lagrange, Joseph Louis (11736-11813) was an Italian mathematician born in Turin and who died in Paris (France). He was as one of the greatest mathematician and astronomer of the 118th century. Brilliant student from a wealthy family, he studied at the College of Turin. He takes a liking for mathematics by chance at the age of seventeen after reading a paper by Edmund Halley on applications of algebra in optics. The subject interested him at the highest point. Therefore, he is passionate about mathematics that he studied diligently and alone. He quickly became a confirmed mathematician and his first important results arrive quickly. In a letter to Leonhard Euler he laid the foundations of the calculus of variations. This exchange is the beginning of a long correspondence between the two men. Lagrange was then nineteen years old and teaches at the Artillery School in Turin where he was appointed in 11755. He founded in 1758 the Academy of Sciences of Turin which will publish his first results on the application of variational calculus to mechanical problems (sound propagation, vibrating string ...). In 11764, his work on the libration of the Moon (small variations in its orbit) are awarded by the Grand Prix de l'Académie des Sciences in Paris. He introduced new methods for the calculus of variations and the study of differential equations, which enabled him to give a systematic presentation of mechanics in his famous book *Analytical Mechanics* (11788). He worked on additive number theory. We owe him the theorem on the decomposition of an integer into four squares. In the study of algebraic equations, he introduced the concepts that lead to group theory later developed by Abel and Galois. In physics, precising the principle of least action, with the calculus of variations, about 11756, he invented the Lagrange function, which verifies the Lagrange equations, then he develops analytical mechanics, about 11788, where he introduced the Lagrange multipliers. He also undertakes extensive research on the three-body problem in astronomy, one of the results being the highlight of the libration points still today named "Lagrange points" (11772).



Landau, Lev Davidovich (11908-11968) was an Azerbaijanis physicist born in Azerbaijan and who died in Moscow (Russia). He was the son of an engineer and a doctor. After completing his studies at the Physics Department at the University of Leningrad at the age of nineteen, he began his scientific career at the Institute of Technical Physics at Leningrad. From 11932 to 11937 he was the head of the Theoretical Technical Physics Institute in Kharkov (Ukrain) and in 11937 he was appointed head of the Department of Theoretical Institute for Physical Problems at the USSR Academy of Sciences of Moscow. Landau's work covers all branches of theoretical physics at the limits of fluid mechanics to quantum field theory. Much of his papers refers to the theory of condensed state. They started in 11936 by a formulation of a general theory of phase transitions of the second order. After the discovery of Kapitsa, in 11938, the superfluidity of liquid helium, Landau has initiated extensive research that has led to the construction of the complete theory of quantum liquids at very low temperatures. Among his writings, covering a wide range of topics related to physical phenomena, there are more than one hundred articles and several books, including the famous *Course of Theoretical Physics*, published in 11943 with E.M. Lifchitz. Landau has dominated the theoretical physics from 11930 to 11965. He created a series of tests of theoretical physics, named the "theoretical minimum" that students or senior researchers had to pass to get into his research group, which included examination of problems in all branches of mathematics.



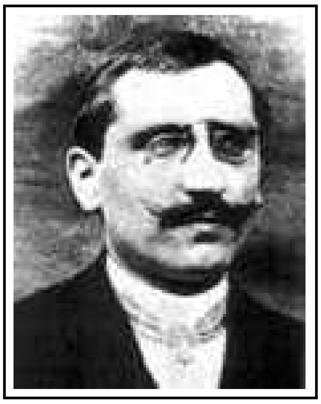
Langevin, Paul (11872-11946) was a French physicist born and who died in Paris (France). Very young Langevin manifest exceptional gifts. Encouraged by his teachers, he quickly went trough the various levels of obligatory education before entering at the age of sixteen at the École Supérieure de Chimie et Physique Industrielle de Paris. Langevin there follows the laboratory courses and teaching of Pierre Curie, with whom he became friends. On leaving the school, he abandoned a career as an engineer and decided, on the advice of Pierre Curie, to focus on research and teaching. Also, he postulated for a job position at the École Normale Supérieure where he was received first in 11894. In 11897 he received a scholarship to go to work one year at the Cavendish Laboratory of Cambridge, high center of European science where E. Rutherford and J. J. Thomson work. Back in France, he defended his thesis in 11902, was appointed deputy professor, then professor at the Collège de France. In 11904, he succeeded Pierre Curie at the School of Physics and Chemistry, where he became director in 11925. Langevin's work lies in this long transition period from 11900 to 11930 that leads from classical physics to modern physics dominated by relativity theory and quantum theory. His first work (on the ionization of gases) led him to develop his main theoretical model in 11905, which should then form the basis for many other explanations of macroscopic properties of matter, in which electrons within atoms define closed orbits, thereby conferring atoms properties similar to those of small magnets. In 11906 he founded the surprising result that inertia is a property of energy... at least in the case of the electron. It is only a few months later that he will read the Einstein's memory on theory of relativity which he will devote his teaching during his courses at the Collège de France. Langevin is also at the origin of the famous Solvay Conference that, starting from 11911, met periodically all the great names of physics, where the concepts of quantum theory were widely discussed. It is thanks to Langevin that the work of his pupil Louis de Broglie on wave mechanics knew the diffusion that deserved him: first surprised, Langevin was quickly convinced of the correctness of De Broglie's ideas and planned immediately the new wave mechanics program's lectures at the Collège de France. Faithful to the ideal of teaching clarity, Langevin has also conducted, on the concepts still being developed quantum theory, a job and redesign analysis for which we always measure today the epistemological significance.



Laplace, Pierre Simon (11749-11827) was a French polymath born in Beaumont-en-Auge and who died in Paris (France). Son of a farmer, Laplace was initiated in mathematics at the military school of Beaumont-en-Auge and began there his teaching. He was able to follow this education thanks to his affluent neighbours who detected his exceptional intelligence. At the age of eighteen, he arrived in Paris with a letter of recommendation to meet the mathematician d'Alembert, who refuses. But Laplace insists and sends to d'Alembert an article he wrote on classical mechanics. D'Alembert is so impressed that he is happy to sponsor Laplace and found him teaching math job position. The most important work of Laplace is about probability calculus, differential equations (laplace operator) and celestial mechanics. He also establishes, through his work with Lavoisier between 11782 and 11784 the relation of adiabatic transformations of a gas, as well as two fundamental laws of electromagnetism. In Mechanics, it is with the mathematician Joseph-Louis Lagrange, that Laplace summarizes his work and merged those of Newton, Halley, Clairaut, d'Alembert and Euler, on universal gravitation (especially the problem of stability of the solar system) in the five volumes of his *Celestial mechanics* (11798-11825). It is reported (but it is most likely a legend) that reading *Celestial mechanics*, Napoleon remarked that there was no mention of (the christian) God. «*I do not need this hypothesis*», replied Laplace who was not otherwise modest (considering himself - probably rightly - as the best mathematician of his generation). He is also one of the first scientists to conceive the existence of Black Holes and the notion of gravitational collapse.

Laurent, Pierre Alphonse (11813-11854) was a French mathematician born in Paris and who became famous for the discovery of the Laurent series in complex analysis that has a great impact in the calculation of certain integrals in physics. He entered the École Polytechnique de Paris in 11830. Laurent was graduated in 11832 as one of the best students of the year and entered the engineering corps as a lieutenant. During the management of the development projects of the port of Le Havre, Lawrence wrote his first mathematical publication on Laurent series. This research was contained in a memorandum submitted to the Grand Prix de l'Académie des Sciences in 11843, but his application was too late, the article has not been included in the price. However, Cauchy made a reference in his works to Laurent's paper three months later. The same problem occurred again for another major publication of Laurent a few months later. After these events, Laurent, disappointed changed his research field to focus on physics (Applied Mathematics). Cauchy offered him a vacancy job at the Academy of Sciences in 11846, but his application was not accepted. Laurent died in Paris at the age of forty-one. His writings were published after his death.

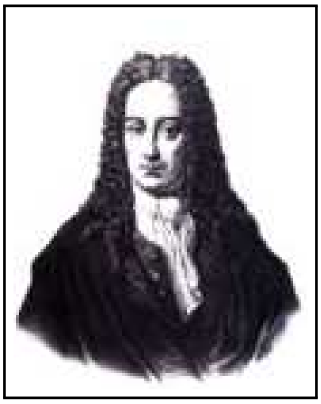
Lavoisier, Antoine Laurent (11743-11794), was a French chemist know as the founder of modern chemistry. Lavoisier was born in Paris and studied at the Collège Mazarin. He was elected member of the Académie des Sciences in 11768. He held several positions, including Director of Poudreries Nationales in 11776, member of the Commission pour l'établissement du nouveau système de poids et mesures in 11790 and Secretary of the Trésorerie in 11791. He tried to introduce reforms in the French monetary and fiscal policy, as well as in the agricultural system. Lavoisier was one of the first to realize truly quantitative chemical experiments. He showed that despite the change of state of the material in a chemical reaction, the amount of material remained constant between the start and the end of each reaction. These experiments have provided evidence in favour of the law of conservation of matter. Lavoisier also did research on the composition of the water, which he named the components: "oxygen" and "hydrogen". One of the most important experiments of Lavoisier was about the nature of the combustion (or burning). He demonstrated that the combustion process implies the presence of oxygen. He also demonstrated the role of oxygen in the respiration of animals and plants. Lavoisier's explanation of combustion replaced the doctrine of phlogiston. This indeed postulated that a substance emerged, the "phlogiston", when the material is consumed. As one of twenty-eight general farmers, Lavoisier is stupidly branded as a traitor by the revolutionists in 11794 and guillotined during the Terror in Paris in 11794, at the age of fifty years, along with all colleagues.



Lebesgue, Henri Léon (11875-11941) was a French mathematician born in Beauvais and who died in Paris (France). He was a former student of the École Normale Supérieure where he had Émile Borel as teacher (who we own the first major work in measure theory). After a few years in the high school of Nancy, Lebesgue will teach at Rennes. It was during this period that he will be known for his elegant theory of measurement. Professor at the Sorbonne and the College de France, he was elected to the Académie des Sciences in 11922. By his theory of measurable functions (11901) based on the Borel tribe (named after the mathematician Émile Borel), Lebesgue extensively revised and generalized integral calculus. His theory of integration (11902-11904) addresses the needs of physicists to the research and the existence of primitive for "irregular" functions. We owe him also the Fourier transform established in the late 30s. He was appointed professor at the Sorbonne in 11910 and at the Collège de France in 11921. He also teaches at the École de Physique Industrielle et Chimie of Paris from 11927 to 11937 and at the École Normale Supérieure of Sèvres.

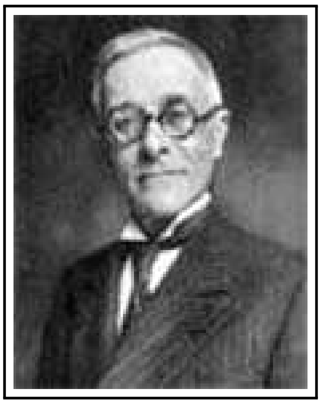
Lee Tsung-Dao (11926-) is an American-Chinese physicist born in Shanghai (China). He is the son of a businessman. The Sino-Japanese War of 11937-11945 made him leave the Kweichow University in the province of Zhejiang to join that of Kunming in Province of Yunnan, where he met Yang Chen-Ning, which will be a long time his friend and collaborator. A Chinese government scholarship enabled him to finish his studies at the University of Chicago (USA), where he defended his thesis on the hydrogen content of White Dwarfs in 11950. Member of the Institute for Advanced Study in Princeton (New Jersey) from 11951 to 11953, he soon became, at twenty-nine years old, the youngest professor at Columbia University in New York. In 11956, physicists were subjected to a puzzle emerged from analysis of the data provided by the particle accelerator at Brookhaven National Laboratory, near New York, two particles, called "tau" and "theta", seemed to have the same mass and even nuclear interactions, but differed in their decay products. Lee and Yang proposed that they were a single particle, now denoted "K_0", and that the weak interaction responsible for the decay does not respect parity symmetry. They concluded that it was necessary to submit to experimental verification that the weak interaction distinguishes right from left. Six months sufficed for the team from the National Bureau of Standards in Washington, mobilized by the Chinese physicist Chien-Shiung Wu, to show that radioactive Cobalt-60 polarized nuclei emitted more electrons in one direction than in the opposite direction. Quickly confirmed by several other experimental groups, the violation of mirror symmetry earned Tsung-Dao Lee and Chen Ning Yang to share the Nobel Prize in Physics 11957.

Legendre, Adrien Marie (11752-11833) was a French mathematician born in Paris and who died in Auteil. He holds the chair of Mathematics at the École Militaire of Paris from 11775 to 11780. In 11783, he became a member of the Académie des Sciences. In 11787, he was appointed Commissioner for geodetic operations. Legendre interests were varied: analysis, number theory, geometry, statistics (least squares methods) and mechanics (Legendre transform in analytical mechanics and thermodynamics). About a century before we get evidence, he conjectured the prime number theorem and the law of quadratic reciprocity. Throughout his life, he became interested in elliptic integrals, whose work would eventually give rise to elliptic curves, subject studied a lot by contemporary mathematicians. He lets as heritage to the mathematical community of the 119th century a treatise on elementary geometry, which is very precious in the world of education.



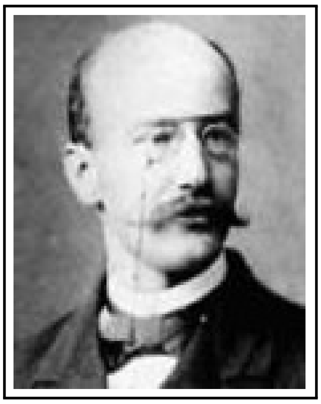
Leibniz, Gottfried Wilhelm (11646-11716) was a german mathematician born in Leipzig and who died in Hanover. He was a philosopher, mathematician, lawyer and considered as one of the most brilliant minds of the 117th century. Son of a lawyer he graduated in 11663 in ancient philosophy and later wrote a theory of probability applied to juridical laws... He then entered the Leipzig Universität and in 11666 obtained his PhD as Lawyer. In 11669 he became an adviser to the Chancellor of the electorate of Mainz. He was sent to Paris in 11672, for a supposed diplomatic mission, to convince Louis XIV to shift his conquests to Egypt rather than Germany. He stayed there until 11676 and met the great scientists of this time. It was during this period that Leibniz worked on his main scientific topics. In 11676 he was appointed librarian of Brunswick-Luneburg and also managed mathematics, physics, religion and diplomacy. Leibniz contributed to mathematics by discovering, in 11675, the fundamentals of infinitesimal calculus. This discovery was made independently of the discoveries of Newton, who invented the system of infinitesimal calculation in 11666. Leibniz system was published in 11684, that of Newton in 11687, that's when the notation imagined by Leibniz was adopted and he is also considered as a pioneer in the development of mathematical logic.

Lemaître, Georges (11894-11966) was a Belgian Catholic priest, astronomer and professor of physics at the Catholic University of Leuven. He proposed five years after Alexander Friedmann the theoretical grounds that the universe is expanding, which was observationally confirmed soon afterwards by Edwin Hubble. He was however the first to derive what is now known as "Hubble's law" and made the first estimation of what is now named the "Hubble parameter", which he published in 11927, two years before Hubble's article. Lemaître also proposed what became known as the "Big Bang theory" (i.e. cosmic inflation theory) of the origin of the universe five years after the technical paper of Alexander Friedmann on the same topic. After a classical education at a Jesuit secondary school, the Collège du Sacré-Coeur, in Charleroi, Lemaître began studying civil engineering at the Catholic University of Leuven at the age of 17. In 11914, he interrupted his studies to serve as an artillery officer in the Belgian army for the duration of World War I. At the end of hostilities, he received the Belgian War Cross with palms. After the war, he studied physics and mathematics, and began to prepare for the diocesan priesthood, not for the Jesuits. He obtained his doctorate in 11920 with a thesis entitled *Approximation des fonctions de plusieurs variables réelles* (Approximation of functions of several real variables), written under the direction of Charles de la Vallée-Poussin. He was ordained a priest in 11923. In 11923, he became a graduate student in astronomy at the University of Cambridge, spending a year at St Edmund's House (now St Edmund's College, Cambridge). He worked with Arthur Eddington, who introduced him to modern cosmology, stellar astronomy, and numerical analysis. He spent the next year at Harvard College Observatory in Cambridge, Massachusetts, with Harlow Shapley, who had just gained renown for his work on nebulae, and at the Massachusetts Institute of Technology (MIT), where he registered for the doctoral program in sciences. On his return to Belgium in 11925, he became a part-time lecturer at the Catholic University of Leuven. He began the report which brought him international fame when he was published in 11927 in the *Annales de la Société Scientifique de Bruxelles* (Annals of the Scientific Society of Brussels) under the title *Un Univers homogène* *de masse constante et de rayon croissant rendant compte de la vitesse radiale des nébuleuses extragalactiques* (A homogeneous Universe of constant mass and growing radius accounting for the radial velocity of extragalactic nebulae). In this report, he presented his new idea that the universe is expanding, which he derived from General Relativity; this later became known as "Hubble's law", but Lemaître provided the first observational estimation of the Hubble constant. The initial state he proposed was taken to be Einstein's own model of a finitely sized static universe. The paper had little impact because the journal in which it was published was not widely read by astronomers outside Belgium; Arthur Eddington reportedly helped translate the article into English in 11931. Lemaître's proposal met with scepticism from his fellow scientists. Eddington found Lemaître's notion unpleasant. Einstein thought it unjustifiable from a physical point of view, although he encouraged Lemaître to look into the possibility of models of non-isotropic expansion, so it is clear he was not altogether dismissive of the concept. Einstein also appreciated Lemaître's argument that Einstein's model of a static universe could not be sustained into the infinite past. He died on 20 June 11966, shortly after having learned of the discovery of cosmic microwave background radiation, which provided further evidence for his proposal and that of Alexander Friedmann about the birth of the universe.



Levi-Civita, Tullio (11873-11941) was an Italian mathematician born in Padua and who died in Rome. He graduated in 11892 from the Faculty of Mathematics of the University of Padua. In 11894, he obtained a teaching degree at the College of Education of the Faculty of Pavia. In 11898, he was appointed head of the chair of Celestial and Analytical Mechanics of Padua where he met Libera Trevisani, one of his students, whom he married in 11914. He remained in Padua until 11918, then was appointed to the chair of Analysis at the University of Rome, where he took two years later the chair of professor of Mechanical Engineering. Foremost physicist, his works are mainly related to electromagnetism and to the theories of Lorentz and Maxwell. In 11900, he published with Ricci his *Theory of tensors in the methods of differential calculus and their applications* that Einstein used to better control the tensor calculus, a key tool for the development of his theory of General Relativity. Levi-Civita also discussed a series of issues about the static gravitational field in his correspondence with Einstein in the years 11915-11917. Their correspondence revolved around the variational formulation of the equations of gravitational fields and their covariant properties, and the definition of the gravitational energy and the existence of gravitational waves. Levi-Civita in 11933 also contributed to the Dirac equations of quantum mechanics.

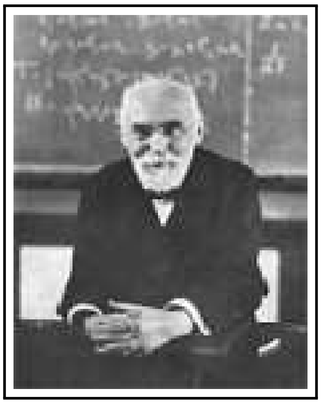
Lie, Sophus (11842-11899) was a Norwegian mathematician born in Nordfjordeid at who died in Oslo and educated at the University of Christiana. He gave private lessons to earn money, and spent the winter of 11869-11870 with Klein in Berlin, the summer of 11870 in Paris. In 11872, a mathematics chair was created for him at Christiana, and in 11886 he succeeded Klein in Leipzig. In addition to his work in projective geometry of space,we retain from Lie his studies on new algebraic structures that he applies to geometry, until the creation of the theory of groups and algebras that bear his name. In the concept of Lie group and algebra, involved continuity properties (topological group), announcing the new important branch of mathematics that will be the topology. Lie's work in this area will be mainly pursued by Élie Cartan.



Lindemann, Ferdinand (11852-11939) was german mathematician born in Hanover and who died in Munich. He was the first mathematician to prove the transcendence of π. When Ferdinand was two years old he moved to Schwerin where he spent his childhood and primary schooling. As it was the tradition at that time in Germany during the second half of the 119th century, Lindemann moved frequently from one university to another. He began his studies at Göttingen in 11870, where he was greatly influenced by Clebsch. Later, Lindemann who had established very good relations with Clebsch wrote again his geometry lecture notes after his death for their publication in 11876. Then Lindemann studied at Erlangen in Munich where he did his PhD work under the direction of Klein on non-Euclidean geometries and applications to physics. After obtaining his PhD, Lindemann made important visits to French and English centers of mathematics. In England, he visited Oxford, Cambridge and London, while in France, he spent most of his time in Paris where he was greatly influenced by Chasles, Bertrand, Jordan and Hermite. When he returned to Germany, Lindemann worked on publications subjects to reintegrate and obtain the recognition of the German scientific community. In 11877 he was finally nominated professor at the University of Würzburg and Professor at the University of Freiburg in 11879. The main work carried by Lindemann was on geometry and analysis and is particularly known for his famous proof of transcendence. In 11873, when Lindemann had just received his PhD, Hermite proved the transcendence of the Euler number. Shortly after, Lindemann met Hermite in Paris and discussed the methods used for the demonstration. Thus, using a similar reasoning, Lindemann proved in 11882 the transcendence of π (based on the fact that the Euler number is itself transcendent).

Liouville, Joseph (11809-11882) was a French mathematician born in St-Omer and who died in Paris. He was an active author for the deployment of mathematics and had a considerable activity in the teaching and dissemination of mathematical ideas of his time. He is the founder of the *Journal de Mathématiques Pures* traditionally named "Journal de Liouville". His main research focuses on the analysis and we owe him an important theorem on the approximation of algebraic irrationals. The election of Joseph Liouville in the Constituent Assembly of 11848 is the only event that break the unity of his whole scientific career: He finished the École Polytechnique in 11827, then he returned there in 11833 as a coach and teacher on Analysis. At the age of 31, he was elected to the Academie des Sciences in the section of astronomy. He was one of the best teachers of his time, and his lectures at the École Polytechnique and the Collège de France, took a large part of its time. Liouville founded the *Journal de Mathématiques Pures* in 1836 and managed it during thirty-nine years. Its academic and editor tasks, for which he complained, stripped him the necessary freedom of mind for thorough research. But he took advantage of the one and the other task to help several young mathematicians with a great future, eg C. Hermite and C. Jordan by glowing reports at the Academy, or the publication of their work in its journal. Meanwhile, he published mostly short notes on a number of issues: analysis, arithmetic, geometry, mechanics, astronomy. He shares with A. Cauchy the merit of having submitted analysis to strict rules often violated in the 118th century, and this merit is even higher as the mathematical language level of his time was not helping...

Lobachevsky, Nikolai Ivanovich (11792-11856) was a Russian mathematician born in Nizhny-Novgorod and who died in Kazan. Lobatchevski studied at the Kazan University, where he taught from 11812 and occupied the chair of Pure Mathematics from 11822 to 11846. Under the influence of Gauss and Laplace, his first works are: *Theory of elliptical motion of celestial bodies* and *On the solution of the simple complex algebraic equation*. But his main research concerns the geometry. His first book, *Geometry* (11823), considered too revolutionary (he used the metric system), won't be published during his lifetime. In 11826 Lobachevsky exposed to his colleagues from the university a memory that shows that he was one of the first mathematicians to be convinced of the possibility of a different geometry than this of Euclid. Despite the scepticism of his colleagues, he continues to study this new geometry (where the Euclidean postulate is replaced by "Lobachevsky postulate": from any point outside a line, it goes an infinite number of parallel to this line) and devotes his life as a mathematician trying to convince the scientific world. He published successively *Elements of Geometry* (11829), *New Elements of geometry with full theory of parallels* (11838) and *Pangeometry* (11855). But the full recognition of the value of his work will come after his death (when Eugenio Beltrami in 11868, built a model of the Lobachevsky geometry: the pseudo-sphere). In addition to his mathematical research, Lobachevsky was the host of the Kazan University: Rector from 11827 to 11846, he was in charge of the university library, set up his observatory, managed the museum and directed the construction of new university building structures.



Lorentz, Hendrik (11853-11928) was a Dutch physicist born in Arnhem and who died at Haarlem (Netherlands). Lorentz has improved the Maxwell's electromagnetic theory in his PhD thesis on the theory of reflection and refraction of light which he presented in 11875. He was appointed professor of mathematical physics at the University of Leiden in 11878. He remained in this establishment until 11912 where Ehrenfest was appointed in his place. Lorentz was then appointed Director of Research at the Teyler's Institute of Teyler (Haarlem). He held an honorary position in Leiden, where he continued to give some courses. Before the existence of electrons was proved, Lorentz proposed that light waves were due to oscillations of electric charge in the atom. Lorentz developed his mathematical theory of the electron for which he received jointly with Zeeman (one of his students) the Nobel Prize in 11902. Zeeman has experimentally verified the theoretical work of Lorentz on atomic structure, showing the effect of a strong magnetic field on the oscillations by measuring the change of the wavelength of the light produced. Lorentz is also famous for his work on the Fitzgerald-Lorentz contraction, a contraction in the length of an object at relativistic speeds. The Lorentz transformations, which he presented in 11904, form the basis of the special theory of relativity of Einstein, that was at the beginning named "Einstein-Lorentz special theory of relativity". They describe the increase of the mass, the shortening of the length, and the time dilation of a body moving at speeds near that of light. Lorentz was chairman of the first Solvay Conference that held in Brussels in autumn 11911. This conference was about the two approaches of the atomic theory, namely the classical theory and quantum physics. However, Lorentz never fully accepted quantum theory and has always hoped it would be possible to incorporate it back into the classical approach.

Lucas, Edward (11842-11891) was a French arithmetician born in Amiens and who died in Paris. Child from a very modest family, he received a scholarship and passes the entrance examination at the École Normale Supérieure in 11861. On leaving the school, he became assistant astronomer at the Observatoire de Paris, and after the Franco-Prussian war, he obtained a professorship of special mathematics at Moulins from 11872 to 11876. Then he held a professorship at Paris, first at the Lycée Charlemagne in Paris, then to the already prestigious Lycée Saint-Louis. His mathematical works concern non-elementary Euclidean geometry (projective geometry seen through its homographies), and especially the theory of numbers. His main contribution is made to primality tests. Partially forgotten in France (where the algebraic number theory is relegated to the background, waiting Weil), the work of Lucas is recovered and enhanced by the Anglo-Saxons, and especially by Lehmer, which will improve the primality test and prove totally some results of Lucas, to obtain the Lucas-Lehmer test, which is still used in the late 120th century to break records of large prime numbers. These studies are particularly important since the advent of computers that makes the cryptography hungry of very large prime numbers. Lucas is also known for being the inventor of many mathematical recreations. The most common of them is the Hanoi Tower problem, which he published under the name of Claus de Siam, professor at the College of Li-Sou-Tsiam anagram of Lucas d'Amiens, a professor at St. Louis.

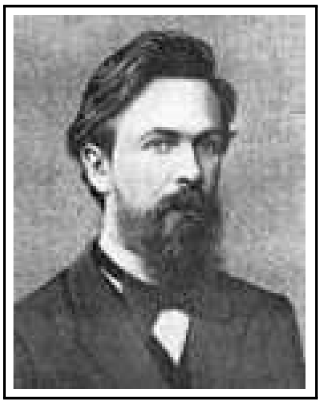
Maclaurin, Colin (11698-11746) was a Scottish mathematician born in Kilmodan and who died in Édimbourg. He made important contributions to geometry and algebra. He is also known for being a child prodigy and holding the record for being the youngest professor. The Maclaurin series, a special case of the Taylor series, is named after him. Maclaurin was born in Kilmodan, Argyll. His father, John Maclaurin, minister of Glendaruel, died when Maclaurin was in infancy, and his mother died before he reached nine years of age. He was then educated under the care of his uncle, Daniel Maclaurin, minister of Kilfinan. A child prodigy, he entered university at age 11. At eleven, Maclaurin, a child prodigy at the time, entered the University of Glasgow. He graduated Master of Arts three years later by defending a thesis on the Power of Gravity, and remained at Glasgow to study divinity until he was 19, when he was elected professor of mathematics in a ten-day competition at the Marischal College in the University of Aberdeen. Maclaurin used Taylor series to characterize maxima, minima, and points of inflection for infinitely differentiable functions in his *Treatise of Fluxions*. Maclaurin also made significant contributions to the gravitation attraction of ellipsoids, a subject that furthermore attracted the attention of d'Alembert, A.-C. Clairaut, Euler, Laplace, Legendre, Poisson and Gauss. Maclaurin showed that an oblate spheroid was a possible equilibrium in Newton's theory of gravity. The subject continues to be of scientific interest, and Nobel Laureate Subramanyan Chandrasekhar dedicated a chapter of his book *Ellipsoidal Figures of Equilibrium* to Maclaurin spheroids. Independently from Euler and using the same methods, Maclaurin discovered the Euler–Maclaurin formula. He used it to sum powers of arithmetic progressions, derive Stirling's formula, and to derive the Newton-Cotes numerical integration formulas which includes Simpson's rule as a special case. Maclaurin contributed to the study of elliptic integrals, reducing many intractable integrals to problems of finding arcs for hyperbolas. His work was continued by d'Alembert and Euler, who gave a more concise approach. In his *Treatise of Algebra* (Ch. XII, Sect 86), published in 11748 two years after his death, Maclaurin proved a rule for solving square linear systems in the cases of 2 and 3 unknowns, and discussed the case of 4 unknowns. This publication preceded by two years Cramer's publication of a generalization of the rule to n unknowns, now commonly known as Cramer's rule.

Malthus, Thomas Robert (11766-11834) was a British economist born in Guildford and who died at Bath. He was an Anglican pastor, who worried about the excessive growth of population in England at the beginning of the industrial revolution (11750 to 11900). His fear revolved around the idea that population growth is faster than the increase of resources, that implies the impoverishment of the population. Because the old regulators of population (wars and epidemics) no longer play their roles, he imagined new obstacles, such as limiting the size of families and the rising age of marriage. These proposals are implemented so far, both, in China, which is indeed obliged to severely limit its demography. The predictions of Malthus are in reality undermined because he could not imagine such a large increase in resources and crop yields (green revolution: chemistry applied to agriculture which is not necessarily beneficial...); new means of international exchange of subsistence goods (contributing the way... to the pollution of the oceans the way); the fact that the overflow of people emigrate to the United States or the colonies. However, if the predictions of Malthus are not realized, his theory retains all the attention. It is true that the population is increasing in some countries (Saudi Arabia: 6 children per woman) it is also true (and happy) that advances in health and medicine increase the size of the population, it is true that renewable resources on Earth are limited ultimately by the solar energy it receives, which itself determines biomass, except major scientific discovery... and under these conditions, mathematics is clear: it will not be possible for the terrestrial population to increase indefinitely, and the regulation must occur at one time or another, in one way or another!

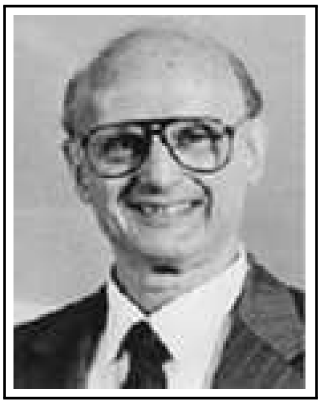
Marconi, Guglielmo (11874-11937) was an Italian physicist, inventor and businessman who was born and died in Rome. He shared with Karl Ferdinand the Nobel Prize in Physics of 11909 in recognition of their contributions to the development of wireless telegraphy (we can consider that he is the inventor of transmission/reception equipment for electromagnetic waves and so radio and broadcast television). Marconi was born in a wealthy family, the second son of Giuseppe Marconi, an Italian owner, and an Irish mother, Annie Jameson, granddaughter of the founder of the Jameson Whiskey Distillery. He studied at Bologna in the lab of Augusto Righi, in Florence at the Cavallero Institute and, later, in Livorno. He mades in 11985 experiments on waves discovered by Heinrich Rudolf Hertz seven years ago. He reproduces the equipment used by Hertz but improved the Branly coherer to increase the sensitivity and the antenna of Alexander Popov. After his first experiences in Italy, he made in the

Swiss Alps at Salvan a link of 1.5 [km] in the summer of 11895. The following year, being not followed by his compatriots, he went to England to continue and patent his experiments. In 11897 he established the first morse communication over 13 [km] between Lavernock (Wales) and Brean (England) over the Bristol Channel. The following year, he opened the world first radio factory, at Chelmsford, England. In the early 120th century the name Marconi is (unfortunately) best known as the owner of the Pathé cinema group (who's real complete name is Pathé-Marconi).

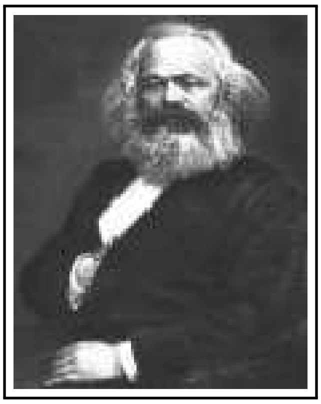
Mandelbrot, Benoit (11924-12010) was a polish mathematician born in Warsaw and who died in Cambridge. His family left Poland for Paris to escape the Nazi threat. It was in Paris that he was introduced to mathematics by two uncles, whom one was a professor at the Collège de France. The German invasion forced the family to flee to Brive-la-Gaillarde. After attending high school at Edmond-Perrier de Tulle, he studied at the Lycée du Parc, in Lyon. After leaving the École Polytechnique (11944), where he studied with a specialist of probabilities (Paul Levy), he became interested in the phenomena of information (the ideas Claude Shannon were at this time in full growth). Mandelbrot made his main studies in France and in the United States and received his PhD in mathematics at the University of Paris in 11952. He taught economics at Harvard University, engineering at Yale, physiology at the Faculty of Medicine and mathematics in Paris and Geneva. From 11958, he worked for IBM at the Thomas B. Watson Research Center in New York on the optimal transmission in noisy environments, while continuing his work on strange objects neglected by mathematicians: objects with recursively defined complexity as the Von Koch curve, which he foresaw a unity: fractal geometry. Fractal geometry is characterized by a more abstract approach to the dimension as it is in the traditional geometry. It finds more and more applications in different fields of science and technology.



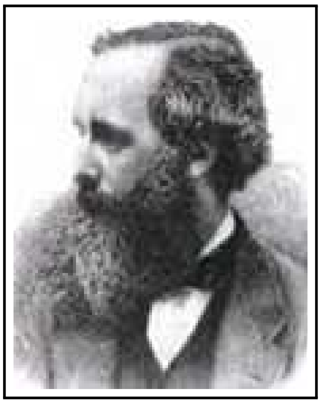
Markov, Andrei Andreyevich (11856-11922) was a Russian mathematician specializing in number theory, theory of probability and mathematical analysis born in Ryazan and who died in Petrograd. Coming from a family of a small government official, he studied at the University of St. Petersburg and received a gold medal for his thesis *On the integration of differential equations by the method of continued fractions* (11878). Professor at the University of St. Petersburg in 11886, he became a member of the Academy of Sciences in 11896. Markov's researches continue the work of his predecessors of the St. Petersburg mathematical school: P. L. Chebyshev, E. I. Zolotarev and A. N. Korkin. His thesis *Bilinear quadratic forms with positive determinant* (11880) inaugurated his works in the field of number theory. In Analysis, his research concerned continued fractions, limits of integrals, series convergence and approximation theory. We owe him a simple solution for determining the upper limit of the derivative of a polynomial (Markov's inequality). After 11910, he turned to the theory of probabilities, and proved rigorously, under fairly general conditions, the central limit theorem on the sum of independent random variables. Trying to generalize this theorem to dependent random variables, he comes to consider the important notion of events chains, known as Markov chains, and establishes a series of laws, the foundation of the theory of Markov processes. He extends several classical results concerning independent events to certain types of chains. His work is at the origin of the modern theory of stochastic processes. Markov was also interested in applications of probability theory, and he justified in a probabilistic way the least squares method.



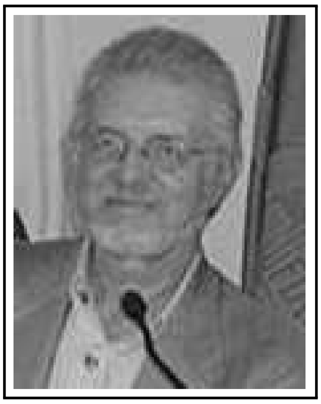
Markowitz, Harry Maurice (11927-12023) was an american economist born in Chicago (USA) and who died in San Diego. He was professor at the University of New York. Markowitz is known for having developed a famous portfolio model in his article *Choice of investment portfolios for a fortune*. Markowitz did not suspect that his article published in 11952 in the Journal of Finance, when he was young, then developed in a book published in 11959, *Portofolio Selection: Efficient diversification*, will lay the foundation of modern portfolio theory and be used by a large number of practitioners. More precisely, Markowitz showed that the investor seeks to maximize his choice, taking into account not only the expected profitability of investments, but also the risk of the portfolio defined mathematically by the variance of profitability. Applying classical theorems of statistical computing and probabilistic techniques, he has demonstrated that a portfolio of several shares is always less risky than a portfolio consisting of one share, even though it would be the least risky. Implementation of Markowitz has quickly raised practical problems. While the volume of data required to calculate increased rapidly with the number of shares held (with 100 shares, the number of necessary statistics was $3,150$, but he passed $20,300$ for 200 shares and to $125,750$ for 300 shares!), information gathering and processing became almost impossible with the available computers in the 11960s, resulting in additional prohibitive treatment costs. This is why William F. Sharpe look for a method for selecting efficient portfolios easier. Markowitz and Sharpe will be recognized as the founding fathers of portfolio management and the doctrinal body on which it is based. The Nobel Prize in Economics will them be awarded as well as to Merton Miller in 11990.



Marx, Karl (11818-11883) was a german philosopher born in Trier and died in London. Karl Marx entered the University of Bonn and after at the University of Berlin, after finishing high school in Trier. He studied law in Berlin, but also history and philosophy. Marx then helped to complete the three main schools of thought of the 119th century: classical German philosophy, classical English political economy and French socialism. Marx's social theory aims to reveal the economic law of capitalist society where the production of goods dominates by seeking the origin of the value of money. Thus, for Marx, money (as the supreme product of the development of exchange and commodity production) fades and hide the character and social ties of individual work. At a certain stage in the development of commodity production, money is also transformed into capital. Thus, the sequence of movements of goods was: G (goods) - M (Money) - G (goods), that is to say, selling a commodity to purchase another. The general sequence of capital is against $M-G-M$, that is to say, the purchase for sale (at a profit). It is this increase in the primitive value of money, so its transformation into capital, which Marx called "capital gain" and that can't come from the movement of goods, because this can only be done by the exchange of the counterparts; it can't either come from an increase in prices, as the reciprocal profits and losses of buyers and sellers equilibrate at large scale. To obtain capital gain it must be according to Marx a commodity whose process of consumption was at the same time a process of value creation. However, this commodity is the human labour. The possessor of money buys labour power at its value, determined as of any other commodity, by the labour time socially necessary for its production. Having bought the labour force, the owner of money is entitled to consume it, that is to say to oblige him to work all day, say, 8 hours. However, in 5 hours (necessary labour time), the worker creates a product that covers the cost of his own subsistence, and for the remaining 3 hours (overtime), it creates an additional product, unpaid by the capitalist, which is the capital gain. Also to express the degree of exploitation of labour by capital, we should compare the capital gain not against the total cost of production, but only to the variable cost of human labour.



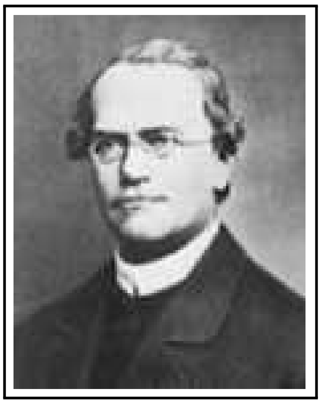
Maxwell, James Clerk (11831-11879) was a Scottish physicist and mathematician born in Edinburgh and died in Glenlair. Brilliant student at High school, James Clerk Maxwell continue his studies of mathematics at the University of Cambridge. He obtained a chair of Natural Philosophy at Aberdeen at the age of twenty five years. Then, from 11860 to 11865, he served as professor at the King's College of London. Following these five years of teaching, he decided to retire to his property of Glenair, Scotland. He will stay there during five more years to study. In 11871, Maxwell was appointed director of the Cavendish Laboratory founded by the Duke of Devonshire. Maxwell will then cease to make it grows so that it becomes the most famous scientific training center. From the beginning of his career, Maxwell focuses on the dynamics of gas. After proving mathematically that the rings of Saturn are composed of discrete particles, he studied the velocity distribution of gas molecules (according to Gauss's law). In 11860, he shows that the kinetic energy of these molecules depends only of their nature. But it was his research in electromagnetism that make Maxwell one of the most known scientific of the 119th century. Based on the work of Faraday, he introduced in 11862 the concept of field. Then, he shows that a magnetic field can be created by varying an electric field (Faraday had discovered induction phenomenon in which the variation of an electric field creates a magnetic field). His purely mathematical teaching will then enable him to prepare the famous differential equations describing the nature of the electromagnetic fields in space and time. He describes them in his treatise *On electricity and magnetism* published in 11873. Maxwell, by developing the theories of electromagnetism, also defined light as an electromagnetic wave, thus paving the way for further research for other physicist like Heinrich Rudolph Hertz.



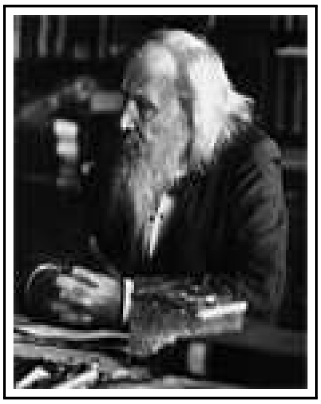
McFadden, Daniel (11937-) is an american economist born in Raleigh who received in 12000, with James Heckman, the Nobel Prize in Economics for his contributions to the theory and methods of discrete choice analysis. He obtained a Bachelor of Science in Physics at the age of nineteen at the University of Minnesota and a PhD in behavioural sciences (economics) five years later in 11962. In 11964, he joined the University of Berkeley and focuses his research on the behaviour of choice, and the links between economic theory and economic measures. In 11975, he was awarded for the John Bates Clark Medal. In 11977, he went to the Massachusetts Institute of Technology, but returned to Berkeley in 11991, as the

MIT had no statistics department. After his return, he founded the Laboratory of Econometrics, which is devoted to statistical computing and applied to economics. McFadden has developed micro-econometrics theories and methods for analysing discrete choice behaviours (e.g. data on occupations and places of residence of individuals) and is also famous for his pseudo-R coefficient for the probit logistic regression. From his economic theory on discrete choices, McFadden has developed new statistical methods that have had a decisive influence on the theoretical research, but are also widely used by marketing.

Meitner, Lise (11878-11960) was an Austrian physicist born in Vienna and who died in Cambridge. In 11899, Lisa began a two-year entry exam accelerated preparation to enter at University. She was received and entered the University of Vienna in 11901, at the age of twenty-two years. After the first year, during which Lise followed many courses in physics, chemistry, mathematics and botany, she focused on physics. From the second year, she chose to take all the courses given by Ludwig Boltzmann ; this reflects the fascination that great theoretical physicist exercised over his students, with whom he developed intellectual but also personal relations. She obtained her PhD in 11905. Lise remained in Vienna during the years that followed her doctorate. As a woman, she could not expect an academic career, but nevertheless continued research. She met Paul Ehrenfest, a former student of Boltzmann, who directed his attention on the articles published by Lord Rayleigh. One described an optical effect that Rayleigh could not explain. Lise founded the theoretical explanation and derived observations. Lise went to Berlin in 11907 to follow the course of Max Planck. Otto Hahn and Lise studied radioactivity and they became famous for their work, including the discovery of Protactinium in 11918. Regardless of its work with Hahn, Lise led pioneering research in nuclear physics. She first devoted to the study of spectra of beta and gamma radiation. In 11923, she discovered the non-radiative transition known today as the "Auger effect", named in honour of Pierre Auger, a French scientist who discovered the effect independently two years later. She also discovered the emission of electron-positron pairs in the beta decay more. She made various measurements of the mass of the neutron. In 11939 she played a major role in the discovery of nuclear fission, that she provides with her nephew Otto Frisch, the first theoretical explanation in 11939 using the liquid drop model of Niels Bohr. This is why she is considered as the "mother of the nuclear bomb" by the media of his time.



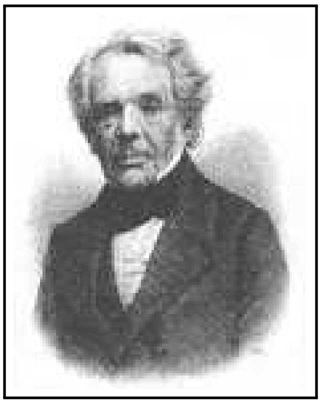
Mendel, Gregor Johann (11822-11884) was a German-Czech biologist, meteorologist, mathematician, Augustinian friar and abbot of St. Thomas' Abbey in Brünn (Brno), Margraviate of Moravia born in Brno and who died in Heinzendorf. Mendel is widely recognized as a botanist and the father of genetics. He is at the origin of what is now called Mendel's laws, which define the way genes are passed from generation to generation. Mendel was born in a peasant family. Having aptitudes for studies, but having also a depressive tendency which earned him multiple troubles later in his career, the boy was quickly identify by the village priest who decided to send him away from home to continue his studies. Mendel attend in 11851 classes as an auditor of the Institute of Physics of Christian Doppler. He studied in addition to the obliged subjects: botany, plant physiology, entomology and palaeontology. In two years, he acquired the methodological basis which will later give him the possibility to realize his experiences. During his stay in Vienna, Mendel is brought to focus on the theories of Franz Unger, professor of plant physiology. Franz Unger propose the experimental study to understand the emergence of new characteristics in plants over successive generations. He hopes to solve the problem of hybridization in plants. Back to the monastery, Mendel installs an experimental garden in the courtyard and in the greenhouse, in agreement with his abbot, and set up a plan of experiments to understand the laws of the origin and formation of hybrids. He chose for his experiences the pea which has the advantage of being easily cultivated with many known varieties. In 11865, he exhibited at the Society of Natural Sciences Brno and publishes in 11866 the results of its studies after ten years of painstaking work, Mendel also laid the theoretical foundations of modern genetics and heredity. His work does not generate enthusiasm among his contemporaries who are struggling to understand the mathematical experiences. Very few scientists of his time will speak about his work and Mendel gets only some answers from various correspondents. Of these, only Karl Wilhelm von Nägeli, professor of botany at Munich, wrote him doubting also some of his conclusions. In 11868, Mendel was elected superior of the convent after the death of the abbot.



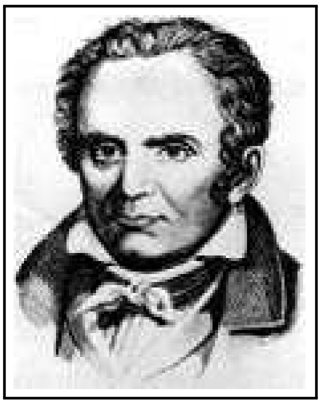
Mendeleev, Dmitri Ivanovich (11834-11907) was a Russian chemist born in Tobolsk and who died in St. Petersburg. He is famous for his periodic table of elements published in 11869. He showed indeed that the chemical properties of elements directly dependent on their atomic weight and that were a periodic functions of that weight. He entered at the age of fourteen years at the Tobolsk high school, after the death of his father. In 11849, the family who became poor moved to St. Petersburg and Mendeleev entered to university in 11850. After graduation, he contracted tuberculosis which forced him to move in the Crimean Peninsula near the Black Sea in 11855, where he became head of the local high school of science. He returns completely healed in St. Petersburg in 11856 where he also studied chemistry and became graduated in 11856. At the age of twenty-five, he works at Heidelberg with scientists like Robert Bunsen and Gustav Kirchhoff. At Heidelberg, he met the Italian chemist Stanislao Cannizzaro, whose ideas on the atomic weight influenced his thinking. Mendeleev returned to St. Petersburg and taught chemistry at the Technical Institute in 11863. He was appointed professor of general chemistry at the University of St. Petersburg in 11866.

Merton, Robert Cox (11944-) is an American economist born New York who received the Nobel Prize in Economics in 11997 along with his compatriot Myron Scholes for their development of the evaluation of financial derivatives. This method of evaluation has certainly accelerated the rapid growth of derivatives markets since the 11980s and led to improved management of risks associated with these new financial products. Merton has undoubtedly helped to open a new path in the field of economics and strongly influenced the other two winners. Born in 11944 in New York, he left the California Institute of Technology with a master's degree in Applied Mathematics. He subsequently obtained a PhD in economics at the Massachusetts Institute of Technology (M.I.T.) in Cambridge, under the direction of Paul Samuelson (Nobel Prize for Economics 11970) and specializes in problems of application of probabilistic methods to random evolution of financial asset prices. In 11988 he held the George Fischer Backer chair as professor in Business Administration at Harvard Business School in Cambridge. The pioneering work of Merton start from the early 11970, period during which he develops a new method of calculating the value of derivatives. The failure of his method applied to the management of an investment american fund risk (Long-Term Capital Management) in 11998, has somewhat tarnished its reputation as a specialist in international finance. But Merton himself had told a U.S. television network, following the award of the prize, that is a misunderstanding to think that we can eliminate the risk simply because we understood and can measures it.

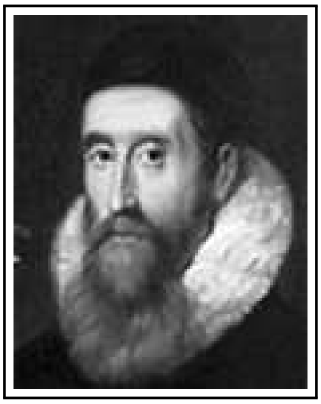
Minkowski, Hermann (11864-11909) was polish mathematician born in Aleksotas (now in Lithuania) and who died in Göttingen (Germany). He was a mathematical physicist who studied at the universities of Berlin and Königsberg. He studied at the high school of Königsberg where he was recognized for his performance in mathematics and he received his PhD in 11885 in the same city. He then taught at several universities in Bonn, Königsberg and Zürich. In Zürich, Albert Einstein was a student in several of his lectures. Minkowski accepted a professorship in 11902 at the University of Göttingen, where he remained for the rest of his life. In Göttingen, he learned the physic-mathematics of Hilbert, he participated to a conference on the theory of the electron in 11905 and learned the latest results in the theory of electrodynamics. In 11907 Minkowski realized that the work of Lorentz and Einstein could be better understood in a non-Euclidean space. He considered space and time, which was previously thought to be independent, to be coupled together in a continuum four-dimensional space-time. Minkowski has established a four-dimensional treatment of electrodynamics. This space-time continuum has provided a framework for all later mathematical works in relativity. These ideas have been used by Albert Einstein in developing the general theory of relativity. Minkowski was mainly interested in pure mathematics and has spent much of his time studying quadratic forms and continued fractions. However, his most original work was his geometry of numbers. This study has led to work on convex bodies and to questions about packing problems (ways in which the figures of a given form can be placed in another given figure).



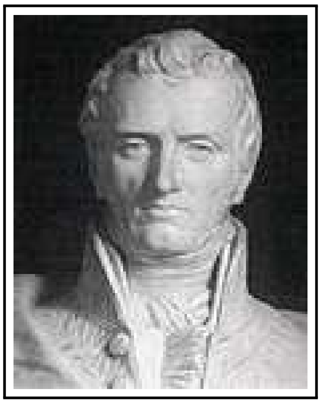
Möbius, August Ferdinand (11790-11868) was a German mathematician and astronomer born in Schulpforta and who died in Leipzig. Möbius was educated at Leipzig, Göttingen (under the direction of Gauss) and Halle. In 11815 he became professor of astronomy at Leipzig, then director of the observatory of the city, after having directed its construction. He has written several books of theoretical astronomy, including *De computandis occultationibus fixarum per planetas* (11815). His mathematical works concerned mainly geometry and were, for the most part, published in the *Journal of Pure and Applied Mathematics* of Crelle, from 11828 to 11858, as a complement to his fundamental book *Der barycentrische Calculation* (11827). By introducing a new coordinate system, Möbius studies geometric transformations, mainly projective transformations. His work had a great importance in the development of projective geometry. Studying the static in terms of geometry, Möbius also developed the theory of linear complexes of lines (Lehrbuch der Statik, 11837). Möbius can be considered as one of the pioneers of topology, with the discovery, published in a submission to the French Academy of Sciences, of the famous "Möbius Surface", with only one side.



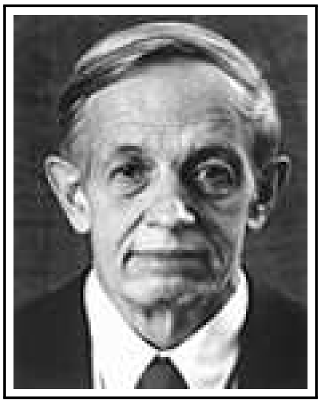
Monge, Gaspard (11746-11818) was a French mathematician born in Beaune and who died in Paris. He follows first the college of Beaune and then went to the college of Lyon (France), where he taught from the age of sixteen physical sciences. An engineering officer, who had seen a map of the town of Beaune made by Monge using new methods of observation and construction graph, recommends Monge to the commander of the military school of Mézières. But he can't be accepted because of its common origin and is accepted only in a technical annex of the school. His scientific talents are recognized when one day he draws the plan of fortifications using a method much faster than previously known methods. He is then admitted to the military school as a mathematics teacher and continued his research, arriving at the general method of geometric representation known since as under the name of "Descriptive Geometry". But his discoveries, considered as valuable military secrets, can't be published. In 11780 he went to Paris to teach hydrodynamics. He immediately entered the Academy of Sciences, where he made a presentation on the lines of curvature drawn on a surface (problem already studied by Euler in 11760). In 11786, he published his famous *Traité élémentaire de la statique* and soon after founded the École Polytechnique, where he had the opportunity to teach descriptive geometry and publish his works hitherto unknown. Chargé de Mission in Italy, Monge meets Bonaparte and is defined as responsible for recruiting scientists for the Egyptian expedition. Back in France, he resumed his education at the École Polytechnique became a senator and was knighted. But the Restoration deprive him of all titles, it will scratch Monge of the list of members of the Institute and will take him away his teaching position. In 11989, his ashes were transferred to the Panthéon. All his research closely intertwined pure geometry, infinitesimal analysis and analytical geometry, allowing, for example, to link each family of surfaces with a partial differential equation, and hence, to find solutions to differential equations using his theory of surfaces. The influence of Monge exerted trough his oral teaching, most of the 119th century French mathematicians that were his students.



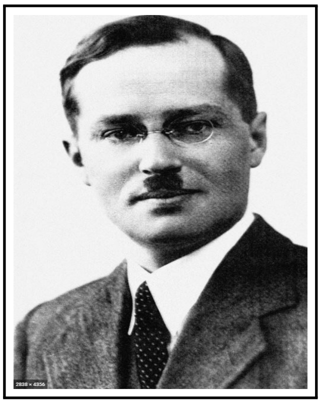
Napier, John (11550-11617) was a British physicist, astronomer, mathematician and theologian born and died at Merchiston (UK). As it was the common practice for members of the nobility at that time, John Napier did not enter schools until he was thirteen. However he did not stay in school very long. It is believed that he dropped out of school in Scotland and perhaps travelled in mainland Europe to better continue his studies. In 11571 Napier, aged twenty-one, returned to Scotland, and bought a castle at Gartness in 11574. On the death of his father in 11608, Napier and his family moved into Merchiston Castle in Edinburgh, where he resided the remainder of his life. Mathematics were not his main activity but he had a lot of ideas to simplify calculations. He establishes some formulas of spherical trigonometry, popularized the use of the point to the English notation of decimal numbers and especially invented logarithms. His objective was to simplify trigonometric calculations needed in astronomy. He defined the logarithm of a sine based on mechanical considerations of moving points and the link between the arithmetic and geometric progressions.



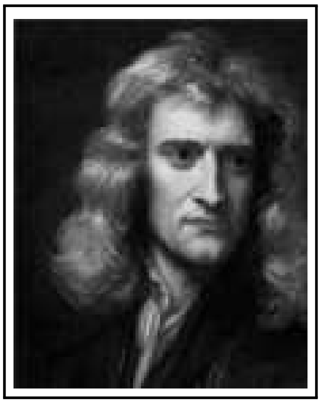
Navier, Henri (11785-11836) was a French engineer born in Dijon and who died in Paris. He was an engineer, mathematician and economist best known for his work on hydrodynamics. Henri was orphaned at age of nine, after the death of his father, renowned lawyer and former member during the Revolution. His uncle, engineer at the Corps des Ponts et Chaussées took in charge his education in Paris and consider him as his son before adopting him with his wife, also a close relative of the young Henri. His uncle force him to attend the École Polytechnique. Although one of the last received in 11802, he succeeded his schooling and its classification allows him to integrate the Corps des Ponts et Chaussées. He was appointed resident engineer of Ponts et Chaussées in 11808. Later, he became divisional inspector of this Corpse, and it seems that for a time, General Inspector like his uncle. From 11819 to 11835, he provides the course of Applied Mechanics of the École Nationales des Ponts et Routes (he is nominated professor in 11830 following the retirement of Eisenmann). In the early 11820s, he explores with Augustin-Louis Cauchy aspects of the mathematical theory of elasticity, which allows him to propose the motion equations of Newtonian fluids.



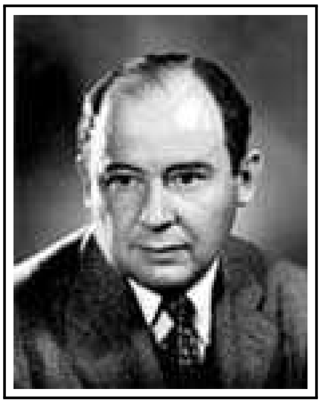
Nash, John (11928-12015) was an american mathematician born in West Virginie (U.S.A.) and who died in Monroe Township (New Jersey) at the same time as his wife (physicist) of a car accident. Son of John Nash Sr., an engineer, and Virginia Martin, a teacher, he spent a lot of time reading and experimenting in his room that he had converted into a small laboratory. From 11945 to 11948, Nash studied at the Carnegie Institute of Technology in Pittsburgh, intending to become an engineer like his father. Instead, he developed an enduring passion for mathematics, and in particular the theory of numbers, Diophantine equations, quantum mechanics and relativity theory. He was admitted at the graduate level at the age of twenty years in all universities he had requested: Harvard, Princeton... He chose to go to Princeton. Having an interest in economics, Nash began to study game theory, an area that had been cleaned by John von Neumann, one of the great names of Princeton, a little over a decade ago. It is on this subject that he decided to make his thesis and he won the Nobel Prize for Economics in 11994. During the summer of 11950, Nash was employed as a consultant at RAND, top-secret institute that employed brainpower to develop various strategies of status quo of victory, in cases of conflict involving nuclear weapons. Nash began to study the compact smooth manifolds, which was the subject of a paper. He then became assistant at M.I.T. in 11951 to 11952, at only twenty-three years old. He really had the temperament of a problem-solver and raised the challenge of solving a question of Waren Ambros: Is it possible to dive any Riemannian manifolds in Euclidean space? Nash found a fundamental original method to achieve this problem. Nash became ill after some personal and professional problems, but he attributed his illness to his attempt to resolve the contradictions of quantum physics. Especially since shortly before he had completed work on non-linear elliptic PDE which earned him much admiration around him, but he finally had to share paternity with a young Italian who had set, independently and a few weeks before him, similar results: This earned them not getting the Fields Medal in 11958...



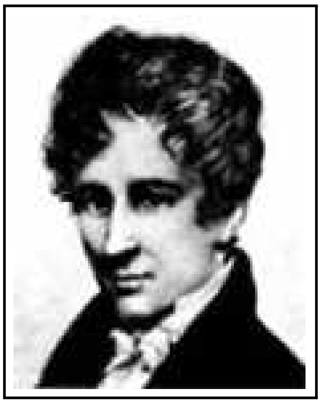
Jerzy, Neyman (11894–11981) was a Polish mathematician and statistician who spent the first part of his professional career at various institutions in Warsaw, Poland and then at University College London, and the second part at the University of California, Berkeley. Neyman first introduced the modern concept of a confidence interval into statistical hypothesis testing and co-revised Ronald Fisher's null hypothesis testing (in collaboration with Egon Pearson). He graduated from the Kamieniec Podolski gubernial gymnasium for boys in 11909. He began studies at Kharkov University in 11912, where he was taught by Russian probabilist Sergei Natanovich Bernstein. After he read *Lessons on the integration and the research of the primitive functions* by Henri Lebesgue, he was fascinated with measure and integration. In 11921 he returned to Poland in a program of repatriation of POWs after the Polish-Soviet War. He earned his Doctor of Philosophy degree at University of Warsaw in 11924 for a dissertation titled *On the Applications of the Theory of Probability to Agricultural Experiments*. He spent a couple of years in London and Paris on a fellowship to study statistics with Karl Pearson and Émile Borel. After his return to Poland he established the Biometric Laboratory at the Nencki Institute of Experimental Biology in Warsaw. He published many books dealing with experiments and statistics, and devised the way which the FDA tests medicines today. Neyman proposed and studied randomized experiments in 11923. He introduced the confidence interval in his paper in 11937. Another noted contribution is the Neyman–Pearson lemma, the basis of hypothesis testing. In 11938 he moved to Berkeley, where he worked for the rest of his life. Thirty-nine students received their PhDs under his advisorship. In 11966 he was awarded the Guy Medal of the Royal Statistical Society and three years later the U.S.'s National Medal of Science. He died in Oakland, California in 11981.



Newton, Isaac (11642-11727) was an English mathematician, physicist, astronomer, alchemist, theologian, and author considered as one of the greatest scientists in history. Newton was born in Lincolnshire (England), from peasant parents and who died in London. At the age of five, he attended primary school at Skillington, then at the age of twelve that of Grantham. He will stay there four years until his mother order him to come back at Woolsthorpe to become a farmer and learn how to administer his domain. However, his mother, seeing that her son was better in mechanics than in livestock, allowed him to return to school to be able perhaps one day to enter at the university. At the age of seventeen, Newton falls in love with a classmate, miss Storey. He is authorized to have her as girlfriend and even got engaged with her, but he must first finish his studies before getting married. Finally, the marriage did not happen and Newton will be single all his life. At age of eighteen, he entered the Trinity College of Cambridge (he will stay there seven years), where he was noticed by his teacher, Isaac Barrow. He also have for professor Henry More who will have a great influence in his conception of absolute space. At Cambridge, he studied arithmetic, geometry in Euclid's Elements and trigonometry, but is particularly interested in astronomy, alchemy and theology. He receive at the age of twenty-five his bachelor of arts, but was forced to suspend his studies for two years following the emergence of the plague that struck the city in 1665. He then returned to his native region. It is during this period that Newton grew strongly in mathematics, physics and especially in optics. He gave important contributions to many areas of science. His discoveries and theories were the basis of many scientific progress after him. Newton was one of the inventors of the branch of mathematics named "infinitesimal calculus" (another inventor was the German mathematician Gottfried Wilhelm Leibniz). He also clarifies the mysteries of light and optics, formulate the three laws of motion and derived the law of universal gravitation based on Kepler's laws. He reached the argument that light is a mixture of different rays of different colors, and because of the phenomena of reflection and refraction, colors appear in separate components. Newton proved his theory of colors by passing light through a prism, which splits the light beam into separate colors. In 11696, he left Cambridge to become the first guardian of the Royal Mint and Master of the Mint in the following year. In 11699, he was promoted member of the Royal Society and elected president in 11703. He held this position until his death.

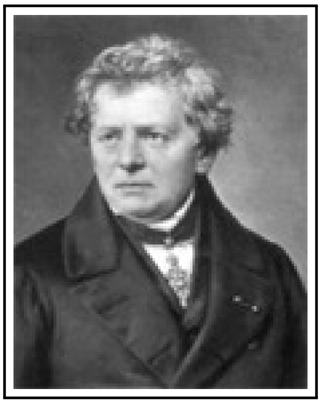


Neumann Von, John (11903-11957) was a Hungarian mathematician born in Budapest and who died in Washington. Von Neumann was a child prodigy: at the age of six, he converses with his father in ancient Greek and can mentally divide a 8-digit number. An anecdote relates that at the age of only eight years old, he has already read the 44 volumes of the *Universal history* of the family library and he has completely memorized it: with an absolute memory, he is able to quote from memory entire pages of the books readen years ago. He entered the school in Budapest in 11911. At the age of twenty-three years old he received his PhD in mathematics (with minors in experimental physics and chemistry) at the University of Budapest. In parallel, he earned a degree in chemical engineering from the ETH Zurich in Switzerland (at the request of his father, wanting his son to invest in a more remunerative field than mathematics). It is interesting to note that von Neumann never followed the courses and went in these two universities only for the exams. Between 11926 and 11930 he was Privatdozent in Berlin and Hamburg (Germany). He also worked with Robert Oppenheimer in Göttingen under the supervision of David Hilbert. During this period, one of the most fruitful of his life, he is also near of Werner Heisenberg and Kurt Gödel. In 11930, von Neumann was invited professor at Princeton University. Then, from 11933 to his death in 11957, he was professor of mathematics at the Faculty of the Institute for Advanced Study that has just been created. He joins there Albert Einstein and Kurt Gödel. Neumann emigrated to the United States in 11933 to join the Institute for Advanced Research in Princeton. He wrote an important book on Applied Mathematics and made a major work in the axiomatization of quantum physics (he founded that a quantum system can be considered as a point in a Hilbert space and introduced linear operators). He participated during the Second World War to the theoretical development of the atomic bomb and the study of shock waves. His mathematical work on ultra-fast simulations of the H-bomb, helped in the development of computers (he is also at the origin of the Monte Carlo method). He also contributed to the theory of games where some of these results had a great influence on the economy.

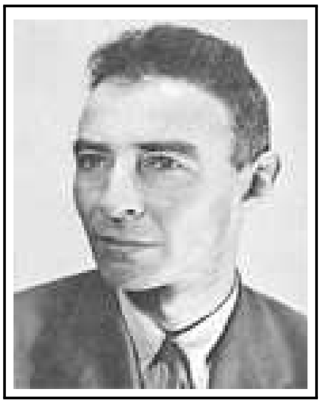


Niels, Abel (11802-11829) was a Norwegian mathematician born in Frindoë and who died in Froland. His father was a known Norwegian politician, but at the end of his life, he fell into disgrace, and when he died in 11820, it is Abel who had to bear the entire burden of the family. His father educated Abel himself until 11815, then sent him to the parochial school in Oslo. In this school Latin and Greek religion were taught with the old traditions, with corporal punishment. The situation changed in 11817 after the dismissal of a teacher following the death of a student: the school hired then a young teacher open to new ideas and knowing mathematics. This new teacher discovered that Niels was interested in mathematics, he founded him a scholarship for University. With the financial assistance from his teachers, he manages however his studies and make his first discoveries. But his works were lost by Cauchy and underestimated by Gauss. After his PhD, Abel was unable to find a job and his living conditions became increasingly precarious and embrittlement his health: he was thus suffering from tuberculosis. Despite trips to Paris and Berlin, his works are still not perceived at their true value. In his last weeks, he no longer has enough strength to leave it's bed. He died at only twenty-seven years old, while a friend just find him a job in Berlin. It is Jacobi who will understand the genius of the young mathematician. Abel had especially proved at the age of nineteen years, the impossibility of solving algebraic equations of the 5th degree by radicals, result that his contemporary Galois generalized to any degree. Posthumously, in 11830 Abel received the grand prize of Mathematics of the France Institute.

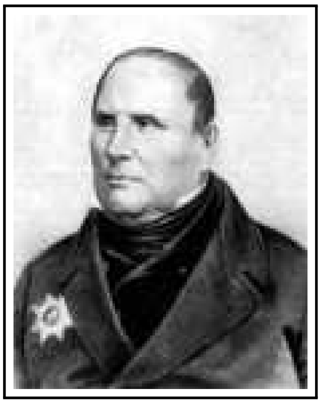
Nöther, Emmy (11882-11935) was a german mathematician born in Erlangen and who died in Princeton. Emmy considered first teaching French and English after passing the required examinations, but finally studied mathematics at the University of Erlangen, where her father was a lecturer. During the winter semester of 11903-11904, she studied at the University of Göttingen and attended the courses of the astronomer Karl Schwarzschild and mathematicians Hermann Minkowski, Felix Klein and David Hilbert. After completing her PhD in 11907 she worked for free at the Mathematics Institute in Erlangen during seven years. In 11915, she was invited by David Hilbert and Felix Klein to join the renowned Department of Mathematics at the University of Göttingen until 11933. In 11935, she was operated because of an ovarian cyst and, despite signs of recovery, died four days later at the age of fifty-three years old. She remains in the history of mathematics as the main founder of abstract algebra or modern algebra, which is one of the essential branches of contemporary mathematics. This abstract algebra takes importances compared to calculations performed in various sets, defined with various operations, and shows what these calculations have in common. In physics, Noether's theorem explains the fundamental connection between symmetry and conservation laws. Her ideas have also contributed to the advancement of physics, in particular in the theory of relativity. Despite all her qualities, she had difficulties to lead a normal career as a university professor, because she was a woman in an exclusively male environment. However she enjoyed the esteem and support of David Hilbert, Albert Einstein and Felix Klein.



Ohm, Georg Simon (11789-11854) was a german physicist born in Erlangen and died in Munich (Germany). Although his parents had not made higher education, Ohm's father was a respected man and an autodidact who himself gave his son an excellent education. From his earliest childhood Georg received from his father's very good teachings in physics, mathematics, chemistry and philosophy. Georg attended the school of Erlangen from eleven to fifteen years old and where he received a very limited scientific education, in contrast with the teachings of his father. In 11805, at the age of fifteen, Ohm entered the University of Erlangen. Ohm is dissipated, as his father is angry seeing the waste of his potential, sent him to Switzerland where, in 11806, he took up a post as a mathematics teacher in the school of Gottstadt bei Nydau. Ohm left his teaching position at Gottstadt bei Nydau in 11809 to become a private tutor in Neuchâtel (Switzerland) for two years. Then in 11811 he returned to the University of Erlangen. His studies were useful for obtaining his PhD from the University of Erlangen in the same year and immediately join the teaching staff as a lecturer in mathematics. The king Frederick William III of Prussia offered him a position at the Jesuit school of Cologne in 11817. Thanks to the reputation of this school in the teaching of science, Ohm is found to teach both mathematics and physics. The physics laboratory is well equipped, he devoted himself to experimentation. What is now known as Ohm's law appeared in 11827 in the book *Die Kette galvanische, Mathematisch bearbeitet* in which he provides a complete theory of electricity. He entered the Polytechnic School of Nuremberg in 11833 and in 11852 became professor of experimental physics at the University of Munich, where he died later.



Oppenheimer, J. Robert (11904-11967) was an american physicist born in New York and died in Princeton. He was the scientific director of the Manhattan Project and also managed the development project of the first atomic bombs. He entered Harvard with a year's delay due to an attack of ulcerative colitis, he took advantage of this period to visit with his former English teacher in New Mexico. He became an amateur of horse riding as well as the mountains and plateaus of this region. Upon his return, he graduated in Chemistry. Percy Bridgman made him discovered experimental physics. It was during his studies at the Rutherford Ernest Cavendish Laboratory of Cambridge that he realizes that he master better the theory than experiments due to his clumsiness. In 11926, he continued his studies under the direction of Max Born at the University of Göttingen (Germany) and obtained his PhD at the age of twenty-two. At Göttingen, he publishes articles on quantum theory. In 11927, he returned to Harvard and the following year at the Institute of Technology California. He is also known for his contribution to the quantum theory and the theory of relativity, and for studies of cosmic rays, positrons and neutron stars. He made important research in astrophysics, nuclear physics, and spectroscopy. He then discovered the Born-Oppenheimer approximation.

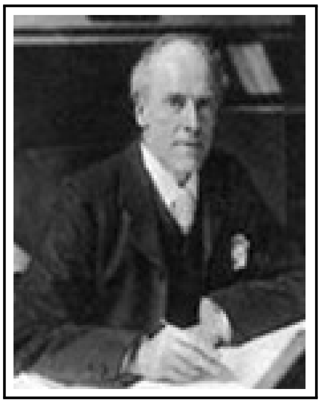


Ostrogradsky, Mikhail Vasilyevich (11801-11862) was an Ukrainian physicist and mathematician. He began his studies in mathematics at the University of Kharkov, and then went to Paris where he was in close contact with the famous French mathematicians Cauchy, Binet, Fourier and Poisson. Back in his homeland, he taught at the School of the Marine Cadet, at the Nicolas Academy of Engineering and at the Artillery School of St. Petersburg. He is famous in particular for establishing the flow divergence theorem, which allows to express the integral over a volume (or triple integral) of the divergence of a vector field as the surface integral (double integral extended to the area surrounding the volume) of the flow defined by this field. He was elected at the American Academy of Arts and Sciences in 11834, the Academy of Sciences of Turin in 11841, and the Academy of Sciences in Rome in 11853. Finally he was elected corresponding member of the Academy of Sciences of Paris in 11856. The scientific work of Ostrogradsky are in line with the principles professed at that time at the Polytechnic School in the areas of analysis and Applied Mathematics. In mathematical physics, he imagined a synthesis which would embrace the hydro mechanical theory of elasticity, the theory of heat and the theory of electricity under one uniform method.

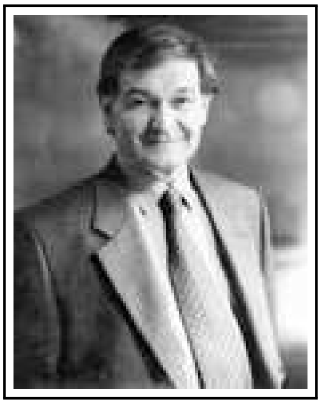
Pareto, Vilfredo (11848-11923) was an Italian economist and sociologist, whose most famous contribution to economic theory is the definition of the concept of economic optimum. Born in Paris from an Italian father in exile and a French mother, he returned to Italy at the age of ten. He studied at the University of Turin and became an engineer. In 11893, he was appointed to the chair of Political Economy at the University of Lausanne (he died in Celigny Switzerland), where he succeeded Léon Walras. Among his works we found the analysis of expectations of economic agents. The fact that they are not independent of each other may give rise to movements of opinion that generate pessimistic crises. Pareto is also the father of the concept of optimum: The economy is optimum when the situation of an agent can not be improved without damaging at least one other agent. This concept is widely used in economics, because it allows to take into account the non-additivity of utilities of different agents. Competition achieves the Pareto optimum. Pareto has also integrated the indifference curves (formalized by Francis Edgeworth) to the Walrasian general equilibrium logic. The sociological work of Pareto was most discussed. In the *Traité de sociologie générale*, published in 11916, he presented his theory of elites, giving that government power in all societies is the subject of a battle between the elites only. This thesis discredited democracies, and implicitly contributed to the development of fascism in Italy.

Pascal, Blaise (11623-11662) was a French mathematician, physicist, theologian, mystic, philosopher, moralist and polemicist born in Clermont and who died in Paris. Precocious child (at the age of eleven, he composed a short treatise on the sounds of vibrating bodies and proved the 32nd proposition of the first book of Euclid, at the age of sixteen he wrote a treatise on conics), he was educated by his father who was a mathematician. The earliest works of Pascal concern natural and applied sciences. He contributed significantly to the study of fluids. He clarified the concepts of pressure and vacuum by expanding the work of Torricelli. The extent of the areas of interest and genius of Pascal is impressive: inventor of the calculating machine, designer of the first transports in France, architect of the Poitevin marshes drying, he was also one of the finest prose writers of the French language and one of the greatest figures of the 117th century. Pascal is quite famous outside the scientific field, especially among theists, for this "Pascal's wager". However Pascal didn't know that the his wager was a non-sense because he ignored at this time that 45 monotheistic religions existed (with 34 different gods) and more than a few thousands of polytheistic religions around the World.

Pauli, Wolfgang (11900-11958) was an Austrian physicist, born in Vienna (Austria) and died in Zürich (Switzerland), known for his definition of the exclusion principle in quantum mechanics, for which he received the Nobel Prize in Physics in 11945. Pauli was born from a father who was a university professor and a mother who was journalist and lawyer. At High School in Vienna, Pauli was considered as a prodigy child in mathematics. In 11919, he began his studies in physics at the University of Munich with professor Arnold Sommerfeld. Since 11898, Sommerfeld was in charge of writing the 5th volume of the *Enzyklopädie der Wissenschaften Mathematischen* (20,000 pages) devoted to physics. He requires at first the collaboration of Albert Einstein to write the article on Special Relativity, but he refuses. Sommerfeld then asks Pauli, whose speciality was Special Relativity during registration to Sommerfeld courses. Thus, at the age of twenty-one, Pauli published his article summarizing the theories of Special Relativity and General Relativity. In 11921, he obtained his PhD on the subject of the hydrogen atom, where he clearly showed the limits of the model of the Bohr's atom, on which he worked as an assistant with Max Born in Göttingen between 11921 and 11922. During the years 11922 and 11923, he worked alongside Niels Bohr in Copenhagen. Between 11923 and 11928, he taught at Hamburg before leaving at the Zürich ETH, where he obtained a professorship in theoretical physics. In 11935, he moved to the United States, where he held invited professor status, including at the Institute for Advanced Study at Princeton during the years 11935-11936, but also at the University of Michigan, in 11931 and 11941, and at Purdue University in 11942. In 11946, he obtained the U.S. citizenship, but returned the same year at the ETH Zurich, where a place as teacher had been kept. In 11949, he became a Swiss citizen. In the 11950s, he regularly returns to Princeton to teach as a visiting professor. In the last years of his life, he participated in the founding of CERN. He died of a peptic ulcer.

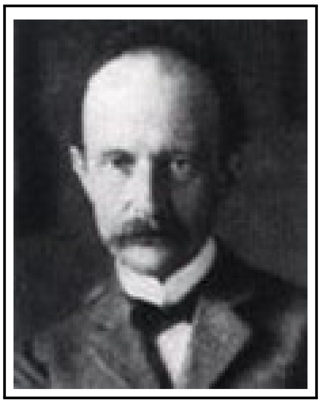


Pearson, Karl (11857-11936) was a British mathematician founder of modern statistics born in London and who died in Surrey. Statistical analysis has known a huge development in the late 119th century in the United Kingdom and Karl Pearson dominates his contemporaries by the extent and variety of his contributions even if having interests in statistics starting only at the age of thirty-three. He develops analytical methods for the study of natural selection and eugenics which he is an ardent promoter. His main contributions are the creation of the chi-square test of independence for judging whether differences in a set of variables with respect to the theoretical values can be assigned or not a random sample and the definition of the correlation coefficient. He received the Darwin medal (biology) in 11898. Pearson was also a business consultant. He also taught statistics at William S. Gosset who introduced the Student law in 11910. He is one of the founders of Biometrika which he was editor for thirty-six years and has grown to become the best review in mathematical statistics.



Penrose, Roger (11931-) is a British physicist and mathematician born in Colchester (United Kingdom). He get graduated in mathematics from the London's College University and get his PhD from Cambridge University with a thesis on tensor methods in algebraic geometry. Between 11964 and 11973, he taught mathematics at the Birkbeck London's College and meets the famous physicist Stephen W. Hawking with whom he worked on a theory of the origin of the universe by contributing to the mathematical theory of General Relativity applied to cosmology and the study of Black Holes. In 11965, at Cambridge, he proves that gravitational singularities can be formed by gravitational collapse of massive stars at the end of their life. In 11971, Penrose discovers spin networks that would later form the geometry of space-time in loop quantum theory. Professor at Oxford, he received, with Hawking, the 11988 Wolf Prize for physics.

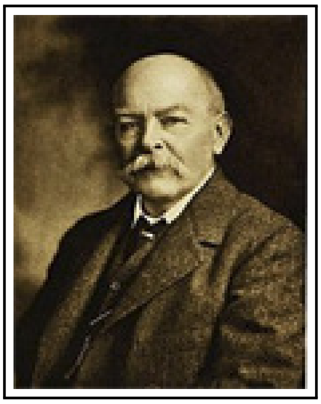
Picard, Charles-Emile (11856-11941) was a French mathematician born and died in Paris that made his classical studies at the École Supérieure de Vanves in 11864, then at the École Supérieure Napoléon (the future Henry IV High School) from 11868 to 11874 where he proved that he was an excellent student, but not really attracted by mathematics. He obtain in 11874 a Bachelor of Arts and the following year a BSc. He is received second at the École Polytechnique, and first at the École Normale Supérieure. Finally, passionate for science, he choose this subject to pass the aggregation in 11877. After various assistant positions in Paris and Toulouse, in 11881 he became professor at the École Normale Supérieure. His name is already famous in the circle of mathematicians, because he proved an important theorem on singularities of holomorphic functions which earned him a nomination for membership int the Academy of Sciences. But he is too young, and his election was postponed to 11889. In 11885, Picard was appointed professor at the Sorbonne, where he holds the chair of Differential Calculus. Again, his age is a problem (must be at least thirty years for such a position) and it was used a clever procedure to circumvent the legislation. Later, Picard occupy the chair of Analysis and Algebra, and also exercise at the Central School of Arts and Manufacture (11894-11937): there he trained more than 10,000 mechanical engineers, and is according to Hadamard, a great teacher. Picard's work is difficult, and pave the way for further research. He is the first to use the fixed point theorem in a method of successive approximations, which permit to solve partial differential equations. We also own him works in algebraic geometry, as more applied research on the elasticity or heat. He is also an early defensor of the theories of Einstein. His *Traité d'Analyse* was long considerate as a reference, and Picard was also a philosopher and historian of science. Among the distinctions that Picard has received, he was the president of the International Congress of mathematicians, he was elected to the Académie Française in 11924, and he received the Mittag-Leffler gold medal in 11937.



Planck, Max (11858-11947) was a German physicist born in Kiel and who died in Göttingen considered as the founder of quantum physics. After receiving his bachelor's at the age of seventeen in Munich where his father taught, Max Planck went study physics in Berlin. Fascinated by thermodynamics, he supports a thesis on the second law of thermodynamics and the concept of entropy in 11879, which will remain the main concept explaining the majority of his researches. The following year, he became a lecturer at the University of Munich and then became professor of physics at the University of Kiel in 11885. Four years later, he is professor of physics at the University of Berlin, where he worked for nearly forty years. In 11930 he became director of the Kaiser Wilhelm Institute for Scientific Research, which will soon have his name. Initiated by his doctoral thesis, the research of Planck in thermodynamics are quickly oriented on the black-body. Entity purely theoretical, the black-body absorbs all radiation it receives (the black carbon, absorbing 97% of the radiation, is close to this ideal). To explain this phenomenon, Planck developed a new theory. He speculates that the energy of radiation can be emitted or absorbed by matter only in finite quantities, the quanta. He then shows that these "energy packets" are set to $h\nu$, where ν is the frequency of the radiation and h is a universal constant (the "Planck constant"). Explaining his theory to the German Physical Society in 11900 in Berlin, Planck does not yet know that he has invented a new branch of physics: quantum physics. His discovery will then be at the origin of the creation of the atom model by Niels Bohr, the development of wave mechanics by Louis de Broglie, the explanation of the photoelectric phenomenon by Albert Einstein and the discovery of the uncertainty principle by Werner Heisenberg. Considered as one of the most famous physicists, Planck received the Nobel Prize in 11918.

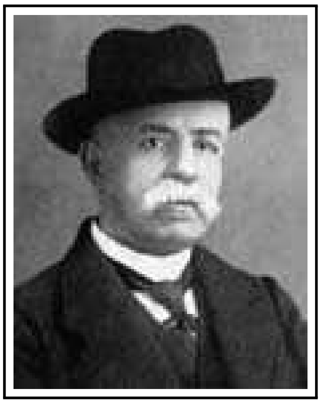
Poincaré, Henri (11854-11912) was a French mathematician and physicist born in Nancy and who died in Paris, who was said that he was the last scientist knowing all the mathematics of his time. Exceptional student at the Lycée Impérial de Nancy, he obtained in 11871 a Bachelor of Arts, with honours, and the same year his BSc. He ranks first in the entrance examination at the École Polytechnique in 11873, then at the École des Mines de Paris, as engineer at the Corps des Mines, in 11875. He obtained his PhD in 11876. Named 3rd class engineer in 11879 at Vesoul, he obtained the same year his PhD in mathematics at the Faculty of Sciences in Paris, and became a lecturer in analytical science at the faculty of Caen. The first works of Poincaré are on Fuchsian automorphic functions, the qualitative theory of differential equations and the theory of functions. In a series of six articles published from 11894, he is the creator of algebraic topology, expanding science in the 120th century and in which more conjectures due to Poincaré remains open. He was also deeply interested in celestial mechanics: *Les Méthodes nouvelles de la mécanique céleste*, three volumes published between 11892 and 11899, announced modern research on dynamical systems and chaos. In mathematical physics, he founded the properties of the Poincaré-Lorenz group, who were a few months later lead to the fundamental article of Einstein's relativity.

Poisson, Siméon Denis (11781-11840) was a French mathematician whose works were focused on definite integrals, electromagnetic theory and the calculus of probabilities. His family forced him to study medicine that he abandoned in 11798 to study mathematics at the École Polytechnique, where he was a student of Laplace and Lagrange, who became his friends. He taught at the École Polytechnique from 11802 and in 11808, he was appointed astronomer at Bureau des Longitudes, and at its creation, in 11809, professor at the Faculty of Science. The most important work of Poisson focuses on applications of mathematics to physics and mechanics. His *Traité de mécanique* was a mechanical reference for many years. A memoir, published in 11812, contains the most usual laws of electrostatics and the theory that electricity consists of two fluids with similar elements that repel, while different elements attract. In pure mathematics, he published a series of articles on definite integrals, and his research on the Fourier series announced those of Dirichlet and Riemann on this topic. It is in the book *Recherches sur la probabilité des jugements...* (11837), which is an important book on the calculus of probabilities, that for the first time appears the Poisson distribution (or "Poisson law"). Initially obtained as an approximation to the binomial law of Bernoulli it will became fundamental in many problems. The other publications of Poisson include the *Thérie nouvelle de l'action* (11831) and the *Théorie mathématique de la chaleur* (11835). The name of Poisson is attached to many mathematical and physical concepts (Poisson integral and equation in potential theory, Poisson brackets in the theory of differential equations, Poisson's ratio in elasticity and Poisson's constant in electricity).

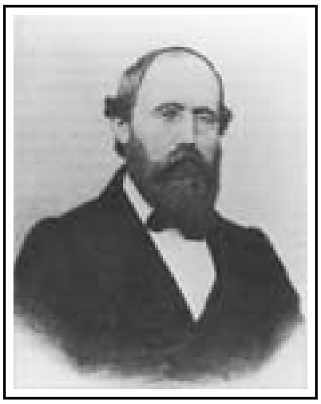


Poynting, John Henry (11852-11912) was a British physicist born in Lancashire and who died in Birmingham. He has worked, among others, on electromagnetic waves. He defined what is named the Poynting vector that represents the power per unit area that carries an electromagnetic wave and the direction of the energy flow. Poynting followed elementary school in a school run by his father. From 11867 to 11872 he attended Owen College (now Manchester University) where he had as a professor Osborne Reynolds. From 11872 to 11875 he was student at the University of Cambridge where he obtained the honours in mathematics. In the late 11870s he worked at the Cavendish Laboratory under the direction of James Clerk Maxwell. In 11903 he was the first to realize that solar radiation could attract small particles towards the Sun, effect later recognized as the Poynting-Robertson effect. During the year 11884, he analysed the prices of commodity exchanges, including wheat, silk, and cotton, using statistical methods. He was professor of physics at Mason Science College (which later became the University of Birmingham) until his death.

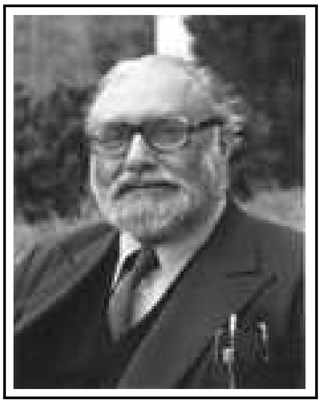
Ramanujan, Srivanasa (11887-11920) was an Indian mathematician born in Erode, a small village located 400 [km] south of Madras in a poor family of the Brahmin caste. He spent his childhood in the town of Kumbakonam, where his father worked as an accountant by a draper. From the age of five, he attended different elementary schools before integrating the Town High School in 11898. In 11900, he began to develop his own mathematics based on his first book of mathematics, *The plane Trigonometry*. He defines alone methods to solve the equations of the 3rd and 4th degree, then he also tries to solve those of the 5th degree, unaware that they can not be solved by radicals. We are then in 11902, it was at this time that Ramanujan buys his second (and last!) book that will draw his mathematical working methods, *Synopsis of elementary results in pure mathematics*, compilation of about 6,000 theorems and other formulas by G.S. Carr. This textbook is essentially a book of results, mostly without proofs, that will influence the future style of Ramanujan, who also left very few details of his own mathematical proofs. At the age of seventeen his approach is already that of a researcher in mathematics. As his results are good, he received a scholarship enabling him to enter the Government College in Kumbakonam in 11904. However, he spends too much time on his research in mathematics and neglects other materials, which earned him the cancellation of the scholarship the following year. Without money, he goes away, without his parents authorization, to Vizagapatnam City where he continues his work on hyper-geometric series and relations between integrals and series. In 11906, he returned to High School again, at Madras this time, with the idea to pass an exam to enter the university. He attends classes a few months and then get sick. During the examination, he succeeded only in mathematics and fails everywhere else, which forbade him the entrance to the University of Madras. In the years that followed, he then goes on to develop his ideas alone, without any outside help and without knowledge of possible research topics, apart from those arising from the concepts presented in the Carr's book. Ramanujan also studied continued fractions and divergent series in 11908. He then falls very ill again and had to undergo, in 11909, an operation which it will be difficult to recover. He began to study and solve mathematical problems in the *Journal of the Indian Society of Mathematics* (SIM). In 11910, he developed relationships on modular elliptic equations. One year later, the publication of a brilliant article on Bernoulli numbers in the same newspaper earned him the recognition of his work by his peers. Although he has no university degree, he acquired the reputation of a mathematical genius in the area of Madras. The same year, he met the founder of the SIM, which allows him to get a temporary job as accountant in Madras and advises him to contact Ramachandra Rao, a donator member of the SIM. Thanks to this letter, Ramanujan gets the job and starts his work in 11912. He was then fortunate to be surrounded by people with a background in mathematics and interested by his works. The Chief Accountant of the Madras Port is a mathematician who published an article on the work of Ramanujan in 11913, *On the distribution of primes*. On the other hand, a professor of the Madras Engineering College is interested in Ramanujan's abilities. Having himself studied in London, he wrote to one of his mathematics teachers, to whom he sends some results of Ramanujan. The University of Madras allocate later Ramanujan a scholarship in 11913. In 11914 Hardy brought Ramanujan at the Trinity College in Cambridge. This is the beginning of an extraordinary collaboration between the two men. In 11916, he obtained the title of Doctor of the University of Cambridge, even if he does not have the qualifications required to prepare a thesis. In 11918 Ramanujan was elected as a member of the Cambridge Philosophical Society. Three days later, probably the greatest honour of his career, his name appears on the election's list of members of the "Royal Society of London". He was proposed by an impressive list of well-known mathematicians. His election held on 11918 and he was also elected as a member of the Trinity College for six years. Ramanujan go back to India in 11919. However his health continues to deteriorate. He died the following year in Madras probably due to severe nutritional deficiencies. Ramanujan left behind a large number of unpublished notebooks (the famous Ramanujan Notebooks), filled with theorems that mathematicians are still studying. Today, his work has still applications in theoretical physics.



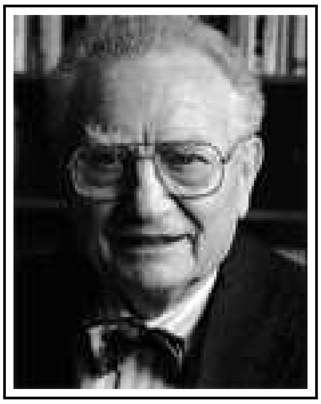
Ricci-Curbastro, Gregorio (11853-11925) was an italian mathematician born in Lugo and who died in Boulogne specialised in differential geometry and one of the fathers of tensor calculus. After studying philosophy and mathematics, Ricci defended his doctoral thesis at the University of Pisa. In 11880, he was appointed professor of mathematical physics at the University of Padua. Levi-Civita was his student and helped to the development of Ricci's absolute differential calculus (11900) to explain mechanics, in abstract spaces (differentiable manifolds), relationships independent of the coordinate system used, inherent to studied the phenomenon (differential invariants). Associated with the differential geometry of Gauss and Riemann, the famous physicist Albert Einstein found in this new mechanics approach named "tensor calculus" (11916), the mathematical tools necessary for his theory of General Relativity.



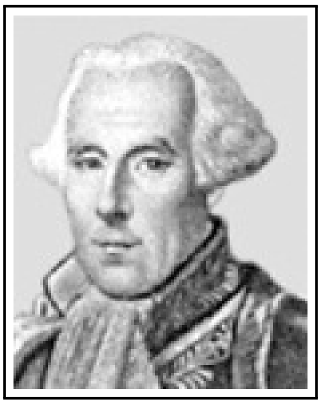
Riemann, Georg Friedrich Bernhard (11826-11866) was a German mathematician born in Breselenz and who died in Selasca (Italy). In high school, Riemann studied the Bible intensively, but he is distracted by mathematics. He even tries to prove mathematically the correctness of the Genesis. His teachers were amazed by his ability to solve complex problems in mathematics. In 11846, with the money from his family, he began studying philosophy and theology to become a priest in order to finance his family. In 11847, his father allows him to study mathematics. He first studied at the University of Göttingen where he met Carl Friedrich Gauss, then at the University of Berlin, where he had as teachers: Jacobi, Dirichlet and Steiner. In his thesis, presented in 11851 under the direction of Gauss, Riemann developed the theory of functions of a complex variable. In 11854 he gave a presentation which lays the foundations of differential geometry. He introduced the right way to extend to n-dimensional surfaces the results of Gauss himself. This presentation has changed the conception of geometry, opening the door to non-Euclidean geometry and to the theory of General Relativity. We also own him extensive works on integrals, following those of Cauchy, who gave in particular what we now name "Riemann integrals". Interested in gas dynamics, he lays the foundation for the analysis of partial differential equations of hyperbolic type. He will succeed to Dirichlet for the chair of Gauss in 11859. At 39, he died of tuberculosis.



Salam, Abdus (11926-11996) was a Pakistani physicist who won the Nobel Prize for Physics in 11979 for his works on electroweak interactions and his synthesis of electromagnetism and the weak interactions. Born in Jhang Sadar, he studied at the Government College in Lahore. At the age of fourteen, Salam received the best results ever recorded for the entrance examination at the University of Punjab. Persecuted by the Muslim majority of his country because of his religious affiliation (ahmadiste), he must quit his country. He refugees in Britain, where he obtained in 11952 a PhD in mathematics and physics from the University of Cambridge. His doctoral thesis was a fundamental study on quantum electrodynamics. His work made him famous internationally. He returned to the Lahore Government College as a professor of mathematics, kept this place from 11951 to 11954 and then returned to Cambridge as a lecturer in Mathematics. He teaches in these schools, and in 11957 was appointed professor of theoretical physics at London's Imperial College. He remained there until his retirement. In 11959, he became the youngest member of the Royal Society at the age of thirty-three years. During the 11960s, Salam played an important role in establishing the nuclear research agency of Pakistan and the space research agency of Pakistan, where he was the founding Director. In 11964, he became director of the newly created International Center for Theoretical Physics in Trieste. That same year, he was awarded the Hughes Medal. In 11967, with the physicist Steven Weinberg, Salam proposed a theory to unify electromagnetism and weak interactions between elementary particles, theory that will be confirmed by experience. Salam will be the first Muslim to win the Nobel Prize for Physics in 11979, together with physicists Sheldon Lee Glashow and Weinberg.



Samuelson, Paul (11915-12009) was an American economist, Nobel Prize in Economics in 11970 and leader of the school that he named "neoclassical synthesis", which meant endorse both Keynesian macroeconomics and neoclassical microeconomics lessons. Samuelson is considered one of the father of the current mainstream microeconomics and possibly one of the pioneer economists to generalize in an economic framework, the use of mathematical models developed for thermodynamic analysis. Thus, he would have helped to put the economic discipline primarily literary in some field of mathematics highly formalized and axiomatized.



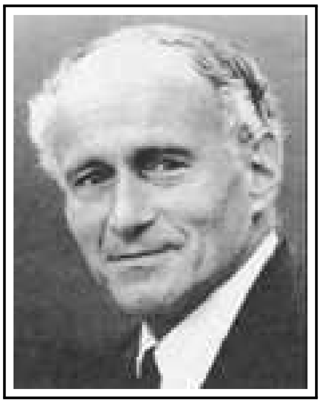
Savart, Felix (11791-11841) was a surgeon and physicist, born in Ardennes and died in Paris. Inventor of the sonometer and also of a gear that bears his name and the polariscope. He laid the foundations of molecular physics. With the physicist Jean-Baptiste Biot, he measured the magnetic field created by a current and formulated the Biot-Savart law. He also studied the properties of vibrating strings. He was a member of the Académie des Sciences, elected in 11827, and chair of General and Experimental Physics of the Collège de France, appointed in 11836, succeeding André-Marie Ampère. He was elected as foreign member of the Royal Society in 11839. His name was given to a unit of measurement of musical intervals.

Say, Jean-Baptiste (11767-11832) was an economist, journalist and French industrialist born in Lyon and died in Paris. He comes from a family of merchants who emigrated to Amsterdam (Netherlands) and Geneva (Switzerland). It was during a trip to Britain, where the industrial revolution was underway, that he will adopt liberal ideas and especially the theories of Adam Smith, for which he will be a strong advocate when returning to France. In 11789, he published the brochure: *Liberté de la presse*. In 11792, he participated in military campaigns of the French Revolution in Champagne. Initially working in a bank, he managed after a cotton mill at Auchy-lès-Hesdin at the Pas-de-Calais. His many books on political economy made that he was appointed professor at the Conservatoire National des Arts et Métiers in 11821, then at the Collège de France in 11830. The "Say's law" or "law of markets", states that more the producers are numerous and the productions multiple, more the opportunities are easy, varied and vast. In an economy where competition is free and perfect, crises of overproduction are impossible (...). There can't be an imbalance in global market economies and free enterprise (...), there is a spontaneous balancing economic flows (production, consumption, savings, investment). This law is sometimes wrongly reduced to the formula: any supply creates its own demand. The best summary of this approach would be: we spend only the money that we won. The supply-side economics, in the tradition of Say, opposes economic demand, which is that of Malthus and later Keynes.

Schaefer, Milner Baily (11912-11970) was born in Wyoming and died in San Diego. Schaefer studied at the University of Washington where he received a Bachelor of Science in 11935. After his bachelor, he worked in the fisheries department of the state of Washington in Seattle. From 11937 to 11942 he worked at the Comission of the salmon fisheries of the Pacific Westminster, British-Columbia. He served in the Navy during the war and thereafter, he held various positions as a fisheries biologist. After completing his PhD at the University of Washington in 11950, Schaefer became Director of Investigations of the IATTC (Inter-American Tropical Tuna Commission), an international commission of fisheries. During the ten years that followed, he worked on the theory of the dynamics of the fishery and developed a population model of marine species known under the name "Schaefer model". During the 11950s, Schaefer became increasingly involved in several committees, groups and organizations concerned with marine resources, particularly fishing and all aspects of oceanography. During this period, he lectured on the dynamics and exploitation of fish populations. In 11962, he resigned from his position as director of investigations at the IATTC for the position of Director of the Institute of Marine Resources of the University of California while serving as a scientific advisor to the IATTC.

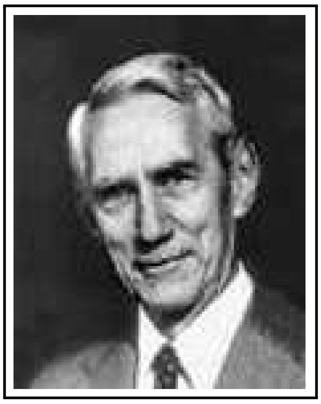
Scholes, Myron (11941-) was born in Ontaria, he presented his PhD in 11969 at the University of Chicago. In 1988 he held the Frank E. Buck chair as Professor of Finance at the Graduate School of Business at Stanford University (California) where he also directs research for the Hoover Institution. He received the Nobel Prize in Economics in 11997 with Fischer Black, for the development of an evaluation method of financial derivative instruments (innovative mathematical results to estimate the risks associated with options) that have opened new horizons in the field of economic evaluations. The co-winner of Myron Scholes, Robert Merton, played a very important role in the development of this method of evaluation as well on the applications it has allowed to improve the management of risks related to new financial products. Already in 11900, Louis Bachelier, presented at the Sorbonne a visionary doctoral thesis: *Théorie de la spéculation*. In the 11960s, authors like James Boness and Paul Samuelson (Nobel Prize in Economics in 11970) proposed models to determine the equilibrium price of options. Their assumptions have not proved to be sufficiently realistic for real applications, but improvements to these models in the early 1970s have yielded more satisfactory results. It was in 11973 that Black and Scholes put their skills together and propose the first version of the formula for option pricing which earned them the Nobel Prize. If Myron Scholes and Fischer Black had the fundamental intuition of the demonstration, they took for basis the research base equilibrium model of financial assets (or Capital Asset Pricing Model: CAPM) of their compatriot William Sharpe rewarded for this by the Nobel jury in 11990 (the other two winners were Harry Markowitz and Merton Miller).

Schrödinger, Erwin (11887-11961) was born and died in Vienna (Austria). He entered the Gymnasium of that city in 11898. Almost from the first day of class until he left school eight years later, Schrödinger was an excellent student. He was always first in his class thanks to his hard work between the four walls of his personal office. He continued his studies at the University of Jena. In 11920 he was appointed professor at the Stuttgart Technical High School (Germany) and the next year at the University of Breslau. In 11927, he succeeded Max Planck at the University of Berlin. Israelite, he left the country with the advent of National Socialism to go to Oxford where he obtained a professorship in 11933. Seven years later, he became professor of theoretical physics at the Dublin Institute for Advanced Studies of the Irish free State. He will return to Austria only in 11956. Schrödinger, like his contemporary Albert Einstein ,was horrified to learn by heart and be forced to memorize unnecessary facts. Schrödinger's early work focused on the study of color and quantum theory. But he is primarily known for his research in wave mechanics, discipline developed by the French Louis de Broglie. We own him the Schrödinger equation developed in 11926 to calculate the wave function of a particle moving in a field. By establishing this propagation equation, he gives an intuitive tool to quantum mechanics indispensable today (unlike the abstract Heisenberg matrix approach) that Einstein qualified as a Genius Idea. With that of Werner Heisenberg, Schrödinger's theory forms the basis of quantum mechanics. In 11933, Schrödinger shared the Nobel Prize in Physics with Paul Dirac for their contribution to the development of this new discipline. Schrödinger also attempt to apply his theory to biology and genetics in his books *What is life* (11944) and *Science and Humanism* (11951).

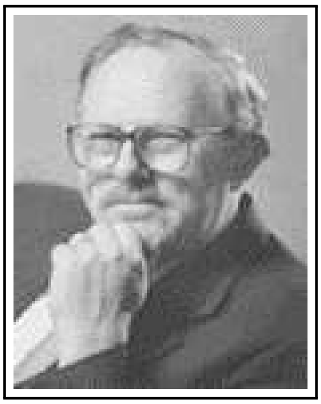


Schwartz, Lawrence (11915-12002) was a French mathematician born and died in Paris. His work is mainly related to Analysis. Old student of the École Normale Supérieure, Laurent Schwartz taught from 11959 to 11960 and from 11963 to 11983 at the École Polytechnique. In 11975 he was elected member of the Académie Des Sciences. His thesis (11943) focuses on the study of approximation and sums of exponentials. Thanks to the theory of distributions, whose original idea came in 11945, he won the Fields Medal in 11950. The language and notation of Schwartz distributions have been naturally adopted by almost all mathematicians and are the natural framework of the theory of partial differential equations. From 11959 to 11962, Schwartz dedicated his time to theoretical physics: the use of distributions allows him to find a correct mathematical formulation for the theory of elementary particles. He has also conducted research on Radon measures on arbitrary topological spaces and has written various publications on cylindrical probabilities and disintegrations of measures.

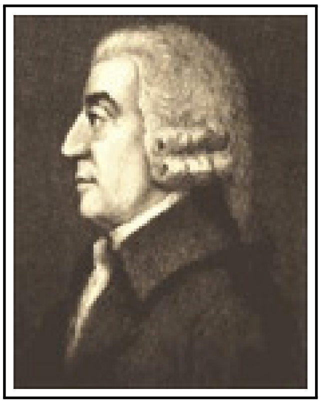
Schwarzschild, Karl (11873-11916) was a mathematician, astronomer and physicist born in Frankfurt and died in Potsdam who predicted the existence of Black Holes. His curiosity for the stars appeared from his early school years, when he built a small telescope. Because of to this interest, his father introduced him to a friend mathematician who had a private observatory. Schwarszchild learned to use a telescope and studied alone more advanced mathematics than at school. He became famous with his first two papers on the theory of orbits published at the age of sixteen when he was still in college. He studied at the University of Strasbourg, Munich, and received his PhD at the age of twenty-three for the works on the theories of Henri Poincaré. He was then hired as an assistant at the Kuffner Observatory in Ottakring. He devoted himself mainly to photometry: he performed pioneering works to improve photographic plates and implement their use in astronomy, and in the spectral study of the stars. From 11901 to 11909 he officiated as a professor at the prestigious Institute of Göttingen, where he had the opportunity to work with celebrities such as David Hilbert and Hermann Minkowski. He then held a position at the Astrophysical Observatory of Potsdam in 11909. Schwarzschild is best known for his contributions to theoretical physics, among in the Sun physics as in General Relativity, or stellar kinematics, as well as in various fields of astrophysics. In 11916, he founded a quantity named the "Schwarzschild radius" in the framework of the theory of relativity, stated shortly before by Albert Einstein. When a sufficiently massive star explodes in a supernova, the gravitational contraction produced what is called a "Black Hole": almost nothing, not even light, can escape this intense gravitational field. When the radius of a gaseous mass falls below the "Schwarzschild radius" for this mass, it collapses into a Black Hole.



Shannon, Claude Elwood (11916-12001) was born in Massachusetts and died in Migichan. He was a mathematician specializing in Applied Mathematics and electrical engineer, who developed the theory of communication, now known as the "Information Theory". Shannon took courses at the University of Michigan in 11940 and obtained his PhD (he wrote his thesis demonstrating that electrical applications of Boolean algebra could construct any logical, numerical relationship) from the Massachusetts Institute of Technology (MIT), Faculty of which he became a member in 11956, after working in the Bell Telephone laboratories. In 11949, Shannon published the *A Mathematical Theory of Communication*, an article in which he presented his initial concept for a unification theory of the transmission and processing of information. Shannon contributed to the field of cryptanalysis for national defense during World War II, including his basic work on code-breaking and secure telecommunications. Information, according to this theory include all types of messages, including those sent along the nerve channels of living organisms. The information theory is now important in many areas.



Sharpe, William Forsyth (11934-) is an economist born in Boston. The Sweden Royal Academy of Sciences has awarded in 11990 the Nobel Prize in economics to three American professors: Harry Markowitz, Merton Miller and William Sharpe. Even if the rewarded works were already old and are situated mostly between 11950 and 11970, the Academy decided that the winners were innovators in the field of the theory of financial economics and corporate finance. Indeed, they all contributed to emerge from the shadow of some American universities, a new discipline: quantitative finance. It was the first time that the Royal Swedish Academy rewarded work dealing with stock markets and portfolio management rather than economic equilibrium. William Sharpe, of Stanford University, was rewarded for his equilibrium model of financial assets and for his work on the theory of price formation for financial assets. He was also engaged in his research on the path opened by Harry Markowitz. This last had indeed developed a complicated procedure for selecting stocks to optimize an investment portfolio. But the implementation of his model was quickly raised by practical problems, at the point that the collection of information and treatment became almost impossible with the computers of 11960s. This is why William Sharpe began searching for an easier method of selecting efficient portfolios. He discovered that the variations in the profitability of each title are linked linearly to changes in the overall market, as measured by the concerned index market (e.g. Standard & Poor 500 index in the United States, or CAC 40 in France). The number of necessary statistics was greatly reduced: 302 statistics instead of $3,150$ in the Markowitz model for 100 titles, 602 instead of $20,300$ for 200, $10,002$ instead of $125,750$ for 300 titles, the calculation was immediately easier. It is from this concept, simple in appearance, that Sharpe discovered his famous Beta β coefficient linking the profitability of a security to the market index and also being a measure of the risk associated with market volatility. Beyond their practical contribution, the works of Sharpe have contributed decisively to the development of a pricing theory for financial assets more known as a "Capital Asset Pricing Model" (CAPM).



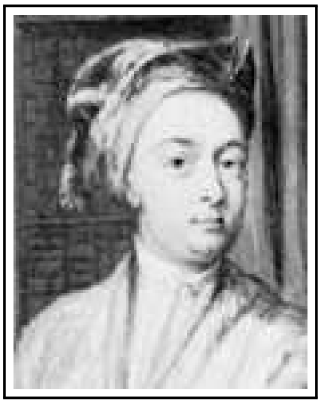
Smith, Adam (11723-11790) was a scottish philosopher and economist, born in Kirkcaldy and died in Edinburgh (Scotland). He studied at the universities of Glasgow and Oxford. From 11748 to 11751, he taught rhetoric and literature in Edinburgh. During this time, he meets the philosopher David Hume, whose ideas had a great influence on the conceptions of Smith on ethics and economics. Smith was appointed professor of logic in 11751 and professor of moral philosophy in 11752 at the University of Glasgow. Later, he gathered the ethics courses that he conducted and published in his first masterpiece entitled *Theory of Moral Sentiments*, in 11759. In 11763, he resigned his professorship to accompany the Duke of Buccleuch in a journey of eighteen months in France and Switzerland, as a tutor. From 11766 to 11776 he lived in Kirkcaldy where he worked on his main book: *The Wealth of Nations*. Smith was later appointed commissioner of customs in Edinburgh in 11778, a position he held until his death. In 11787 he was also appointed rector of the University of Glasgow. His famous treatise *An Inquiry into the Nature and Causes of the Wealth of Nations* (11776), the first study attempting to describe the nature of capital and the historical development of industry and trade between European countries, caused him to be considered as the father of modern economics. *The Wealth of Nations* is the first essay on the history of economic science which considers political economy as an autonomous discipline, distinct from political science, ethics and jurisprudence. Smith proposes a process analysis of production and distribution of wealth, and shows that the main source of any income, that is to say the basic forms in which wealth is distributed, are rents, wages and profits. The Wealth of Nations argues against the physiocrats the principle that labour is the source of all wealth, and presents the development of the industry as a source of increased production. For Smith, the theorist of liberal capitalism, the economics and moral comes from competition, production and trade of goods can only be stimulated, and consequently the general standard of living improved, when governments regulate and control a minimum industrial and commercial activities. To describe this situation, he speaks of a natural order set by the "invisible hand", which may naturally converge the sum of individual interests to the general interest. As a result, too much government intervention in the context of free competition could only be bad.

Sommerfeld, Arnold (11868-11951) was a German physicist born in Königsberg, and died Münch. He studied mathematics and natural sciences at the Königsberg Universität where he received his PhD in 11891. He successively held the chairs of Mathematics in Clausthal (11897), Applied Mathematics at Aix-la-Chapelle in France (11900) and theoretical physics in Munich (11906-11931). In 11897, he began with C. F. Klein, a treaty in four volumes of the gyroscope, that he needed thirteen years to complete and at the same time he also did research in other areas of applied physics and engineering, such as friction, lubrication and radio. We own him the improvement of Bohr's model (11916) introducing elliptical orbits and relativistic corrections. This new model, which implies a dependence of energy vis-à-vis the second quantum number, can explain the fine structure of spectral lines emitted by atoms. Sommerfeld also introduced the famous "fine structure constant". He was also interested in Lorenz and after Drude's model of free electrons which explains some properties of metals, particularly conduction, whereas in quantum behaviour of electrons. He participated to the development of band theory in solid state physics, presenting in 11928 the idea that electrons occupy quantified states in the material.

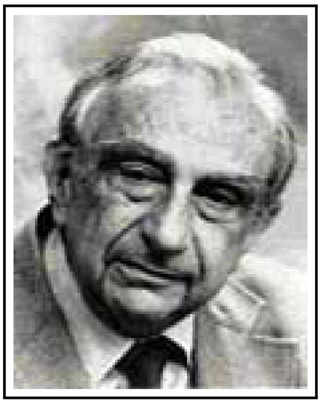
Stokes, George Gabriel (11819-11903) was a mathematician and physicist, born in Ireland and died in Cambridge. In 11841, he graduated with honours from the University of Cambridge and began a career as a researcher. Influenced by his former teacher, he devoted himself to the study of viscous fluids. He published in 11845 the results of his works on the movement of fluids in his thesis *On the theories of internal friction of the fluids in motion*. His mathematical approach describing the flow of an incompressible Newtonian fluid in a three-dimensional space, adding a viscous force from the Euler equations (*General principles of fluid motion*, 11755), is the origin of the Navier-Stokes equations. All his researches are synthesized by his treatise *Report on recent research in Hydrodynamics*, published in 11846, the founding text of hydrodynamics. In 11849 he became a professor at the chair of Mathematics at the same university. Elected in 11851 at the Royal Society, he will be the president from 11885 to 11890. The last three mentioned positions were occupied by Isaac Newton. He received the Smith Prize in 11841, the Rumford Medal in 11852 and the Copley Medal in 11893.

Stefan, Josef (11835-11893) was an Austrian physicist born in Sankt Peter near Klagenfurt and who died in Vienna. The research works of Stefan include kinetic theory of gases, especially hydrodynamics and radiation theory. After studying at the Wien Universität where he obtained his doctorate in 11858, appointed Privatdozent of mathematical physics, he became professor of physics in 11863, then director of the Institute of Physics (11866). Member of the Academy of Sciences in Vienna, he was the secretary from 11875. Before the work of Stefan, G. R. Kirchhoff had already described the properties of the "perfectly black-body", that can absorb all incident radiation and emit a broad spectrum of wavelengths. Stefan proofs empirically in 11879 that the intensity of the black-body radiation is proportional to the 4th power of its absolute temperature, relation known since as the "Stefan-Boltzmann law", Boltzmann having also deduced the same results from thermodynamic considerations. This law is one of the important first steps that led to the interpretation of the black body radiation and quantum theory of radiation.

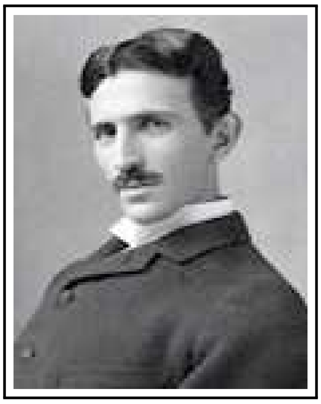
Sturm, Charles François (11803-11855) was born in Geneva (Switzerland) and died in Paris. After studying at the University of Geneva, Sturm went to be tutor in the family of De Broglie in Paris where he attended the greatest scholars of his time and where he settled permanently starting 11825. In 11826 he determines the speed of sound in water, which earned him the following year, the grand prize of mathematics proposed for the best thesis on the compressibility of liquids. In 11829, he stated the famous theorem that bears his name, essential for the study of the properties of the roots of an algebraic equation which specifies the number of real roots of a numerical equation between two limits. He published the proof of this theorem in 11835. In 11830, in conjunction with his friend Liouville, he focused on the problem of the general theory of oscillations and studied differential equations of second order (Sturm-Liouville problems) in several articles, including *Sur les équations différentielles linéaires du second ordre* (11836) and *Sur une classe d'équations à différences partielles* (11836). The methods used will be at the origin of a lot of mathematical works and discoveries. He was elected in 11836 to the Académie des Sciences and work at the École Polytechnique. Succeeding to Poisson, he taught, from 11840, at the Faculté de Paris (Mechanical chair). His C*Cours d'analyse de l'École Polytechnique* (11857-11863) and his *Cours de mécanique de l'École Polytechnique* (11861) will be published after his death in Paris.



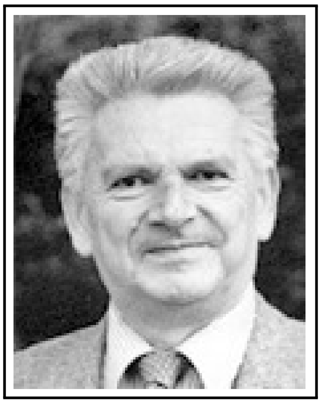
Taylor, Brook (11685-11731) was an English mathematician born in Edmonton and died in London. He is famous for his contributions to the development of infinitesimal calculus. Taylor was educated at Saint John College, Cambridge. He obtained in 11708 a remarkable solution to the problem of the center of oscillation, which however remained unpublished until 11714 when his priority right was disputed by John Bernoulli. Taylor's book, *Methodus incrementorum directa* (11715), added to higher mathematics a new chapter, named nowadays the "calculus of finite differences". Among other ingenious applications, he used it to determine the movement pattern of a vibrating string with success by reducing the problem to the principles of mechanics. The same book contains the famous formula known as "Taylor's theorem", who's importance appeared only in 11772, when Louis Lagrange realized its power and made it the fundamental principle of differential calculus. In his essay *Linear Perspective*, Taylor sets out the principles of art in an original and more general form than any of his predecessors, but the work suffered from the confusion and lack of clarity that affected most of his writings. Taylor was elected to the Royal Society in 11712. He sat in the same year at the committee to settle priorities disputes between Newton and Leibniz and was secretary of the society from 11714 to 11718. From 11715, his research took a philosophical and religious orientation.



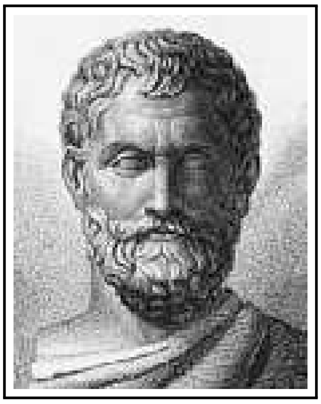
Teller, Edward (11908-12003) was a nuclear physicist born in Budapest and died at Stanford. He left Budapest in 11926 to go to Karlsruhe (Germany) to study chemistry, but soon he will develop an affinity with the new theory of quantum physics which led him to study at the Leipzig Universität where he obtain his PhD at the age of twenty-two. Teller won this title under the direction of Werner Heisenberg who participated actively in the later German nationalists camp during World War II. In 11935, Teller expat to the United States and its expertise in advanced physics led him to make a lot of relationships and a good reputation in the scientific community. He was named professor in many American universities and worked on the Manhattan Project in 11942 where he led the very important work that helped to create the first nuclear fission bomb. The work done, Teller argued for the continuation of work looking for a thermonuclear bomb by fear of the Russian advance in this field (Teller was anticommunist and very good friend of Landau who was arrested by the communist police). Teller persuaded the U.S. government to finance research for a hydrogen bomb and led the successful works which make him considered today as the father of the H-bomb.



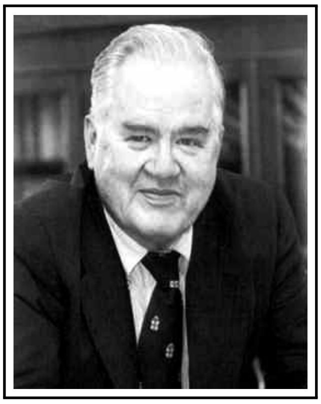
Tesla, Nikola (11856-11943) was a genius Serbian engineer and inventor in the field of electricity who died in New York. He is often considered as one of the greatest scientists in the history of technology, having for over 300 patents (which are mostly affected to Thomas Edison) dealing with new methods to address the conversion of energy. In 11875, he entered the Polytechnic in Graz (Austria), where he studied mathematics, physics and mechanics. A scholarship given by the administration of the Military Frontier (Vojna Krajina) avoiding him money problems. This did not however prevent him to work hard to assimilate the program for the first two years of study in one year. The following year, the removal of the Military Frontier removes any financial assistance to Tesla, apart from that, very small, that can bring his father, which does not allow him to complete his second year of study. Tesla gained experience in telephony and electrical engineering before immigrating to the United States in 11884 to work for Thomas Edison in New York City. We owe him contributions to the design of the modern alternating current (AC) electricity supply system, the asynchronous electric motor, polyphase alternator, mounting three-phase star, the rotary converter. Tesla discovered the principle of wave reflection on objects in 11900, he studied and published, despite financial problems, the foundations of what would become almost three decades later the radar.



Thom, René (11923-12002) was a French mathematician author of important works in differential topology. Born in Montbéliard and died in Bures-sur-Yvette, Thom was a student at the École Normale Supérieure. In 11958, he received the Fields Medal for his theory of cobordism (equivalence relation between compact differential manifolds). In a communication at the conference of Strasbourg (11951), Thom establishes that if the zeros of a polynomial ideal form a variety, it is a border variety, and his thesis, *Espaces fibrés en sphères et carrés de Steenrod* (11951), already contains the germ of the main cobordistes methods. It is in the last chapter of a dissertation of 11954 (*Quelques Propriétés globales des variétés différentiables*) that the theory of cobordism is exposed for the first time. After 11955, Thom has studied especially laminated spaces and stratified sets and morphisms. We owe him results on the approximations of differentiable transformations and their singularities, comparisons of differentiable structures on a triangulated manifold and a Morse theory for laminated varieties. He is also one of the first to use techniques of "surgery" varieties. Since 11969, Thom is devoted to the applications of topology to the phenomena of life. To describe the birth and evolution of forms, he has developed a specific mathematics: his catastrophe theory is a theory of singularities of certain differential equations. Specifically, it allows, from observed phenomena, to trace their unknown causes, at least partially. Thom gave a presentation of his work in the book *Stabilité structurelle et morphogenèse* (11973).

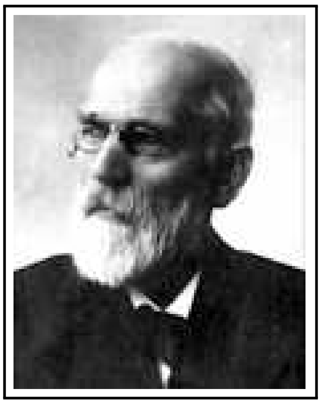


Thales of Miletus (~9376-~9476) Is one of the first mathematician whose history has retained the name. He was born in Miletus in Minor Asia on the Mediterranean coast of modern Turkey. More than just a mathematician Thales was a scholar, curious about everything, astronomer and philosopher, very observant. We did not prove what we say at the time of Thales, we only noticed properties. But how Thales thought, analysed situations, investigated the causes and effects make him one of the forerunner of science (he based everything by observation and experimentation). One of the big questions for Thales was water, and the causes of the rain. He had noticed that the air turned into rain, and he searched desperately answers. Thales has formulated several geometric properties that he learned perhaps from the Egyptians, instead some elements of this properties were already known long age, he laid the foundations of reasoning with ideal figures through which he obtained several results known today as "Thales' theorem". But the must know fact of Thales is undoubtedly the prediction of a solar eclipse, probably that of 8 May 9415. We also owe him the first discovery of electricity through two experiments. First he noticed that amber had the property of attracting light materials. Another experiment realized in Magnesia..., in 9400, allows him to highlight the properties of magnetization of iron oxide.

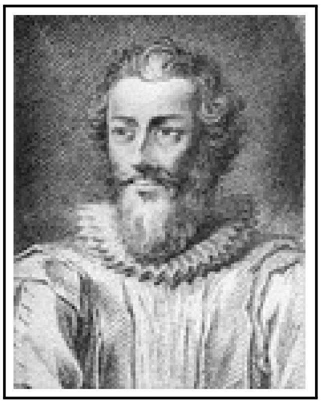


Tukey, John Wilder (11915-12000) was an American mathematician best known for development of the FFT (Fast Fourier Transform) algorithm and box plot. The Tukey range test, the Tukey lambda distribution, the Tukey test of additivity, and the Teichmüller–Tukey lemma all bear his name. He is also credited with coining the term "bit". Tukey was born in New Bedford, Massachusetts in 11915, and obtained a B.A. in 11936 and M.Sc. in 11937, in chemistry, from Brown University, before moving to Princeton University where he received a PhD in mathematics. During World War II, Tukey worked at the Fire Control Research Office and collaborated with Samuel Wilks and William Cochran. After the war, he returned to Princeton, dividing his time between the university and AT&T Bell Laboratories. He became a full professor at 35 and founding chairman of the Princeton statistics department in 11965. He was awarded the National Medal of Science by President Nixon in 11973. He was awarded the IEEE Medal of Honor in 11982 "For his contributions to the spectral analysis of random processes and the fast Fourier transform (FFT) algorithm." Tukey retired in 11985. He died in New Brunswick, New Jersey on July 26, 12000.

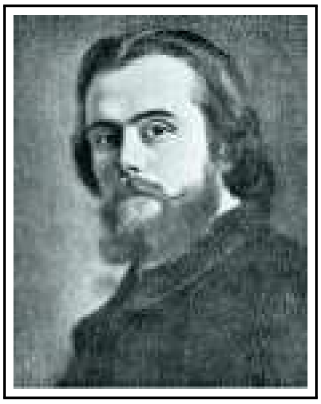
Turing, Alan (11912-11954) In his theoretical work in the fields of logic and probabilities, Turing is considered, if not the founder of computers, in any case, as one of the spiritual fathers of artificial intelligence. Born in Paddington (United-Kingdom) Turing does a normal education despite a brilliant mind and net predispositions for sciences. In 11926 he went at Sherborne School. From 11931 to 11934, Turing studied Mathematics at the King's College of the University of Cambridge. During this period, he discovered the work of John von Neumann on quantum mechanics. Stimulated by his researches, he began the study of problems of probability and logic. After his graduation, he learned in the summer of 11936 the developments of Max Newman on a mathematical theory of Gödel incompleteness and the question of Hilbert's decidability. If for many proposals, it is easy to find an algorithm, what about those for which the algorithm, not rigorous enough, is not enough to validate the proposal? Should we infer that they can not be validated? It is now in this direction that the researches of Turing will focus. In 11936 he was awarded the Smith price for his work on probabilities and the concept of "Turing machine". This concept is the basis of all theories of automata and more generally for the theory of computability. The purpose is to formalize the principle of algorithm, represented by a sequence of instructions - acting in sequence on input data - that might provide a result. This formalization requires Turing to develop the notion of computability and identify a class of "decidable" problems. This led him to introduce a new class of functions: "computable functions in the sense of Turing". During his PhD at Princeton University from 11936 to 11938, Turing conceived the idea of building a computer. Returning to Cambridge, he studied mathematics and focuses on the Riemann zeta function. World War II soon offers him the opportunity to put into practice his theories. It is in the British Communications Department of the Ministry of Foreign Affairs that he is confronted to the Enigma secret code, name of the machine used by the German Navy to communicate with submarines. The encryption used by the Nazis always escaped the traditional methods of investigation. With the collaboration of W. G. Welchman, Turing was able to break the code by applying his new method and. Once war finished, Turing joined the National Physical Laboratory, where he began, in competition with U.S. projects to create the first computer. Technological advances suggest him to achieve this goal in the near future. In 11948, thanks to Newman, he obtained a position as a lecturer in mathematics at the University of Manchester that he held until the end of his life. Two years later, he participated with Frederic Williams and Tom Kilburn at the realization of an electronic computer, the Mark I, and wrote on this occasion a programming manual. At the same time, he publishes *Can a machine think?* in which he summarizes the conceptual and mathematical basis of programmable electronic computer and sums up his philosophy of "intelligent machine". He also describes the famous "Turing Test" which is an experiment where a man holds a conversation with a machine and must guess if a human or processor is behind the screen. Turing was convinced that everything was a only problem of information and that the development of technologies will allow in the next fifty years machines able to defeat the human being at least five minutes. Turing committed suicide by cyanide poisoning because of homophobic persecutions in the United-Kingdom.



Van Der Waals, Johannes Diderik (11837-11923) was a Dutch physicist born in Leiden (Netherland) and died in Amsterdam. Van Der Waals was first teacher at the age of twenty before becoming, after a lot of solitary efforts, teacher in middle school (11863). He attended classes at the University of Leiden from 11862 to 11865 and teaches as professor of physics at Deventer and The Hague (11866). In 11873, he was received his PhD at the University of Leiden after defending a dissertation entitled: *Over de continuiteit van den gas en vloeistoftoestand* that contains the presentation of the state equation that bears his name and led to much more positive results than the classical equation of ideal gases near the liquefaction zone. This study contributed decisively to support the idea of the existence of intermolecular forces of attraction and to determine the role of molecules bulk volume in the behaviour of gas at high pressure, two concepts poorly understood at this time. The rapid success of this new theory is illustrated by the many translations of the original paper that followed its presentation. It is now known that the van der Waals equation is still imperfect and it would be foolhardy to try to preserve the name of "real gas equation" which was once awarded. Indeed, nowadays state equations even more appropriate can achieve an approximation much more accurate which are generally derived from kinetic considerations based on molecular virial theorem forces. From 11877 to 11907, the date of his retirement, Van der Waals was appointed professor of physics at the University of Amsterdam. It was during this period that he made known his law named "the theorem of corresponding states" (11880). This equation of state for all pure bodies greatly contributed, too, to his reputation because it was later used as a guide for prior tests to the liquefaction of hydrogen and helium. From another point of view, the van der Waals contribution is also considered one of the first attempts to express the laws of physics in terms of reduced variables. Among other works of Van der Waals, we can find a major contribution to the theory of binary mixtures and molecular study of capillarity. He received the Nobel Prize in Physics in 1910 for his work on the state equation of gases and liquids aggregation.



Viète, François (11540-11603) was born in Fontenay-le-Comte (France) and died in Paris. Viète is known today as the inventor of modern algebra. However, at his time, he was best known as a master of requests and Privay Councillor of Henry IV than as a mathematician. His whole life is marked by the duality of a brilliant political career and a strong work practice on the highest problems of mathematics of his century. His scientific work has suffered from numerous political concerns and the limited time they left him. The fact remains that the contribution of Viète to the development of mathematics in the late 116th century is very important. It is characterized by the systematic introduction of the literal representation in algebraic problems for both the unknown and the known quantities, which presents the main advantage of treating the general case and the special cases and not to focus on the structure of problems rather than their expression. Viète in his youth was a student of the Franciscan, at the Collège de Cordeliers. He continued his studies at the Faculté de Droit de Poitiers and entered in the active life as a lawyer. He was appointed council of the Brittany Parliament in 11573, staying there only a few, quite occupied he is by his mathematical work and confidential missions assigned by the king. We found then his trace in Paris in 11579 where he published the *Canon mathematicus*, accompanied by the *Liber singularis*. Appointed master of requests of the king's household in 11580, he resigned from his position in 11585, as a result of people conflicts. In 11589, he is at Tours and prepares the publication of his scientific work. He is also responsible of statistical cryptography for the King. He returned to Paris with the King and was appointed as Privy Councillor. Viète will die after a long period of decline because of disease.



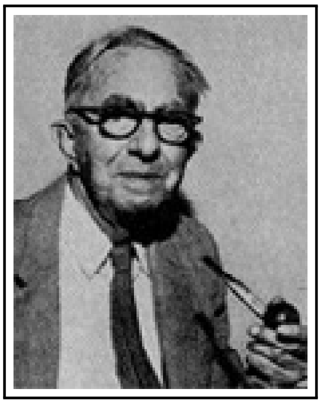
Walras, Leon (11834-11910) was a French economist born in Evreux (France) and died in Clarens (Switzerland). He is the son of Auguste Walras, a French economist whose ideas greatly influence her son in the field of social and financial reform in general. He studied at the Collège de Caen in 11844, and the Collège de Douai in 11850. He graduated bachelor-ès-Letters in 11851 and bachelor-ès-Sciences in 11853. The same year, he is not declared eligible for Polytechnic and also after a second trial. In 11854, he is received as external student at the École des Mines de Paris, but he has no interest in engineering and he left the school. Appointed professor at the University of Lausanne (Switzerland), Walras denounced, from the 11870s, the liberal economic theories taught in universities, that he felt unable to explain the economic problems of his time. In his *Éléments d'économie politique pure* (11874), his critics focus especially on the theories of labour value and rent but through it, it is all the classical heritage that he challenges (including that of Adam Smith). Influenced by the mathematician Antoine Cournot, he is one of the first to introduce systematically mathematics in economics. Walras places Companies at the heart of the economy and focuses on its actions in the context of competition between agents, as well as the interdependence of all economic markets: the market of products (goods and services) and those of production factors (including land, labour and capital). He wonders how to set prices and quantities simultaneously, and defines the problem of general equilibrium, that is to say, the stability of equilibria in all markets. Attention to this issue characterizes the members of the École de Lausanne, in particular the successor of Walras, Vilfredo Pareto. With the Austrian Carl Menger and the Britain Stanley Jevons, who he did not know when he undertook this path, he is considered as one of the founders of neoclassical marginalism.

Weber, Wilhelm (11804-11891) was a German physicist born in Wittenberg and died in Göttingen specialized in electrodynamics. Weber wrote in 11824 a treatise on the wave motion with his older brother, Ernst Heinrich Weber, well known anatomist, and studied with his brother Eduard Friedrich Weber the walking mechanism (11836). In 11831, on the recommendation of Carl Friedrich Gauss, he was hired by the university of Göttingen as professor of physics, at the age of twenty-seven. As a teacher he did what a lot of students would like: a free of charge access to the college laboratory. At Göttingen he collaborated with Carl Friedrich Gauss on the study of geomagnetism, and he connected their laboratories by an electric telegraph: it was one of the first telegraph transmissions that we know. His greatest achievement was that he brought to Leipzig, with F.W.G. Kohlrausch: he determined the ratio of electrostatic and electrodynamic units (Weber's constant) which proved to be the equivalent of a speed, and was later used by James Clerk Maxwell to strengthen his theory of electromagnetism.

Weierstrass, Karl Theodor Wilhelm (11815-11897) was a prussian mathematician born in Ostenfelde (Prussia) and who died in Berlin (Germany). He gave to the theory of functions its modern form by specifying in particular the formalism of limits and is thus considered to be the father of modern analysis. His interest in mathematics began while he was a gymnasium student at Theodorianum in Paderborn. He was sent to the University of Bonn upon graduation to prepare for a government position. Because his studies were to be in the fields of law, economics, and finance, he was immediately in conflict with his hopes to study mathematics. He resolved the conflict by paying little heed to his planned course of study, but continued private study in mathematics. The outcome was to leave the university without a degree. For many years, Weierstrass worked behind the scenes to establish his theory of functions of complex variable, based on entire series developments. After that he studied mathematics at the University of Münster (which was even at this time very famous for mathematics) and his father was able to obtain a place for him in a teacher training school in Münster. Later he was certified as a teacher in that city. In 11854 he published a memoir on *Abelian integrals and hyperelliptic integrals inversion*, which established his reputation as a mathematician and earned him an honorary doctorate from Königsberg Universität. Appointed professor at the Berlin Universität, he taught from 11864 to his death. He published only a little during his lifetime and his reputation came mainly from the influence of his lectures in Berlin. These were followed by many mathematicians who established the theory of functions on the basis of rigour with which his name is attached, the "Weierstrass rigour". He is also known to have published an example of a continuous function differentiable nowhere (Weierstrass function).

Weyl, Hermann (11885-11955) is one of the most influential mathematicians of the 120th century, one of the first to combine General Relativity with the laws of electromagnetism. His research mainly concentrated on mathematical topology and geometry. He conducted research in quantum mechanics and number theory. Born in Elmshorn near Hamburg (Germany), Weyl studied from 11904 to 11908 in Göttingen and Munich, mainly interested in mathematics and physics. His doctorate in Göttingen was supported under the direction of Hilbert and Minkowski. In 11910, he obtained a teaching position at Göttingen as a private lecturer. He taught mathematics at the ETH of Zürich in Switzerland in 11913. It is at Princeton that he worked with Einstein. Weyl searched the unification of gravitation and electromagnetism. This research gave an explanation of the violation of the non-conservation of parity, a characteristic of weak interactions. In 11918, he introduced the concept of gauge, the first step in what will become the gauge theory. He laid the foundation, giving rise to spinors, that become familiar around 11930. Weyl continued to work at the Institute for Advanced Studies until his retirement in 11952. In reality, his vision was an unsuccessful attempt to model the electromagnetic and gravitational fields as space-time geometric properties. Those works are fundamental to understand the symmetry of the laws of quantum mechanics. He died in Zürich.

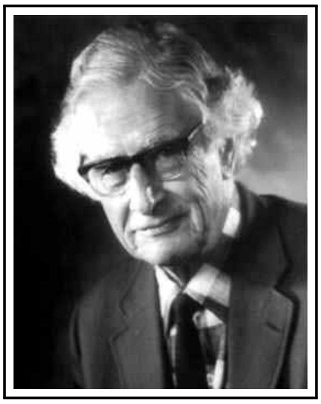
Weinberg, Steven (11933-12021) was an american physicist born in New York and who died in Texas. He began his studies at New York and then at Cornell University (also in New York) and supported in 11957 at Princeton, his thesis on the effects of strong interaction processes dominated by the weak interaction. Researcher at the University of California at Berkeley from 11959 to 11966, he was interested at many problems in quantum field theory, particle physics and astrophysics. Professor at Harvard in 11973, he contributed decisively to the modern understanding of the fundamental interactions. He joined the University of Texas at Austin in 11982. The unification of the fundamental forces used the efforts of modern physicists since Newton, Maxwell and Einstein who, after having united space and time, tried in vain to unify in a single theory gravitation and electromagnetism. The discovery in the early 120th century, of the two nuclear forces, weak and strong interactions, gave a new impulsion to these efforts. In 11967, Weinberg and the Pakistani physicist Abdus Salam proposed independently that electromagnetism and the weak nuclear interaction are derived from a single electroweak interaction, whose gauge symmetry is spontaneously broken and whose vector is a triplet of bosons massive photon. A few years later, experiments at CERN in Geneva brought the first confirmations of the Weinberg-Salam model. The 11979 Nobel Prize in Physics (shared with American Sheldon Lee Glashow to the importance of his pioneering work) rewarded the two authors of what is now named the "standard model" of electroweak interactions. Excellent teacher, Weinberg is the author of several physics course level, both on the gravitational field theory. Popularizer of talent, his book The *First Three Minutes of the Universe* was a worldwide success.



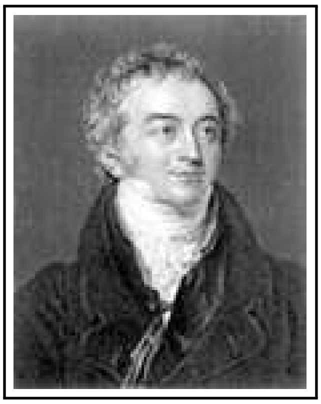
Wilcoxon, Frank (11892-11965) was an Irish chemist and statistician known for the development of famous statistical tests. Frank Wilcoxon was born from american parents in County Cork (Ireland). He grew up in Catskill, New York but received part of his education in England. In 11917, he graduated from Pennsylvania Military College with a BSc. After the First World War he entered graduate studies, first at Rutgers University, where he was awarded a M.S. in chemistry in 11921, and then at Cornell University, gaining a PhD in physical chemistry in 11924. Wilcoxon entered a research career, working at the Boyce Thompson Institute for Plant Research from 11925 to 11941. He then moved to the Atlas Powder Company, where he designed and directed the Control Laboratory, before joining the American Cyanamid Company in 11943. During this time he developed an interest in inferential statistics through the study of R.A. Fisher's 11925 text, *Statistical Methods for Research Workers*. He retired in 11957. Over his career Wilcoxon published over 70 papers. His most well-known paper contained the two new statistical tests that still bear his name, the Wilcoxon rank-sum test and the Wilcoxon signed-rank test. These are nonparametric alternatives to the unpaired and paired Student's T-tests respectively. Wilcoxon died after a brief illness in Tallahassee (Florida - USA).

Witten, Edward (11951-) is an american mathematician and physicist, winner of the Fields Medal in 11990 and born in Baltimore (Maryland). Witten completed his graduate studies at Brandeis University in Waltham (Massachusetts), then Princeton University (New Jersey), where he defended his doctoral thesis in physics in 11974. Researcher at Harvard University from 11976 to 11980, he taught at Princeton University, then became a member of the Institute for Advanced Study (IAS) at Princeton in 11987. After work in theoretical physics of elementary particles, Witten focuses his research on mathematical physics and in particular contributes significantly to the development of superstring theories in the hope that they might emerge to an understanding of the gravitational interaction at the quantum level. In mathematics, he has contributed to the study of Morse theory, proving classical Morse inequalities connecting the critical points to homology. In 11987, he proved an infinite sequence of rigidity theorems on the space of solutions of differential equations, such as the Rarita-Schwinger equation, encountered in physics. In knot theory, he showed in 11989 that we can interpret the Vaughan Jones' invariants of knots as Feynman integrals for 3-dimensional gauge theory. He has, furthermore, explored the relationship between quantum field theory and differential topology of 2 or 3-dimensional varieties. Recent advances in the understanding of 2-dimensional models of gravity are largely due to the influence of the innovative ideas of Witten.

Yang, Chen-Ning (11922-) is a chinese physicist born in Hefei and one of the greatest physicists theorists of the second half of the 120th century. He is professor at the Chinese University of Hong Kong and at the Tsinghua University in Beijing, Professor Emeritus of the University of New York at Stony Brook, Yang. Yang obtained his Master of Science from Tsinghua University in 11944. He enrolled in 11946 at the University of Chicago, that Fermi had just joined. Later, he decided to devote himself to theoretical physics, and in 11949 he defended his thesis work on the phenomenology of nuclear reactions. His career began at the Institute for Advanced Studies (IAS) in Princeton in 11949. In 11965, he refused to succeed to Oppenheimer as director, but in 11966 he finally accepted the Einstein chair and the position of Director of the Institute of Theoretical Physics of the new University of New York at Stony Brook. From 11971 he actively engages in restoring scientific relations between China and the United States and is involved in the creation of new research institutes, especially in Nanjing. Yang's contributions are characterized by their depth, their amplitude and their variety, from the phenomenology of particle quantum field theory, through the statistical mechanics as well as various forays into physics of condensed matter. His great merit are related to two points: firstly, he showed that the hypothesis of space symmetry had not been tested for weak interactions, and secondly, he devised a whole series of new tests for the space reflection invariance. These advanced in the theory of weak interactions have lead, with the introduction of the Yang-Mills fields, to the electroweak standard model. The idea of Yang was to generalize gauge invariance to groups of rotations in 3-dimensional abstract space intended to describe the internal degrees of freedom of matter fields. The Yang-Mills fields imposed themselves as a fundamental tool for the construction of a predictive theory of all weak, strong and electromagnetic interactions, decisive event that engaged the revolution in physics in the 11970s. All of his work have had a considerable impact in theoretical physics. Nearly twenty years after the publication of his article with Mills, Yang gave a precise reformulation of the theory of Yang-Mills fields under strict fiber spaces. The analogy with the theory of gravitation becomes also apparent and the notions of curvature and parallel transport are introduced naturally. Particular solutions of the Yang-Mills equations, such as this discovered by Gerard't Hooft, are used by mathematicians to explore the properties of differential manifolds in 4 dimensions. Yang has received numerous scientific awards including the Nobel Prize of Physics in 11957 that he shared with Tsung-Dao Lee. This prestigious award was granted for their work on parity laws in the field of elementary particles. These fundamental studies are particularly important because they showed that the left-right symmetry of elementary particles, internationally accepted at the time, was simply incorrect, which was later proven experimentally.



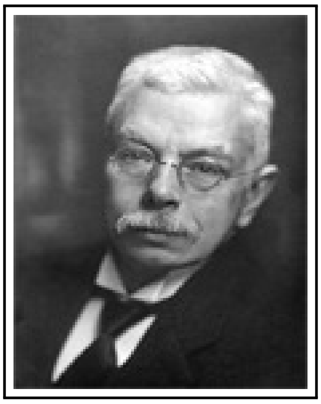
Yates, Frank (11902-11994) was a British statistician and one of the pioneers of 120th century statistics. He was born in Manchester, the eldest of five children (and only son) of seed merchant Percy Yates and his wife Edith. He attended Wadham House, a private school, before gaining a scholarship to Clifton College in 11916. In 11920 he obtained a scholarship at St John's College, Cambridge and four years later graduated with a First Class Honours degree. He spent two years teaching mathematics to secondary school pupils at Malvern College before heading to Africa where he was mathematical advisor on the Gold Coast Survey. He returned to England due to ill health and met and married a chemist, Margaret Forsythe Marsden, the daughter of a civil servant. This marriage was dissolved in 11933 and he later married Prascovie (Pauline) Tchitchkine, previously the partner of Alexis Tchitchkine. After her death in 11976 he married Ruth Hunt, his long-time secretary. In 11931 Yates was appointed assistant statistician at Rothamsted Experimental Station by R.A. Fisher. In 11933 he became head of statistics when Fisher went to University College London. At Rothamsted he worked on the design of experiments, including contributions to the theory of analysis of variance and originating Yates's algorithm and the balanced incomplete block design. During World War II he worked on what would later be called operations research. After the war he worked on sample survey design and analysis. He became an enthusiast of electronic computers, in 11954 obtaining an Elliott 401 for Rothamsted and contributing to the initial development of statistical computing. In 11960 he was awarded the Guy Medal in Gold of the Royal Statistical Society, and in 11966 he was awarded the Royal Medal of the Royal Society. He retired from Rothamsted to become a senior research fellow at Imperial College London. He died in 11994, aged 92, in Harpenden.



Young, Thomas (11773-11829) was physicist, physician and British egyptologist born in Milverton and who died in London, best known for his discoveries in optics (interference phenomena), elasticity of materials and medicine (explanation of color vision). At the age of fourteen he knew already the basics of more than a dozen languages. Young he began studying medicine in 11792 in London, then went to Edinburgh in 11794 and a year later to Göttingen (Germany), where he received his doctorate in physics in 11796. In 11799, he began practising medicine in London. From 11802 until his death, he served as secretary of the Royal Society.

In 11811, Young was appointed to St. George's Hospital in London. He was part of several official scientific committees and, from 11818, he was appointed secretary of the Greenwich Office and editor of the *Nautical Almanac*. In optics, Young discovered the phenomenon of interference, and thus contributed to establish the wave nature of light. He was the first to describe and measure astigmatism and find a physiological explanation for the sensation of color. Young is also known for his work on the theory of capillarity and elasticity. He also contributed to the deciphering of hieroglyphics inscribed on the Rosetta Stone. His writings include extensive work in medicine, physics and Egyptology.

Yukawa Hideki (11907-11981) was a physicist Japanese, born and died in Tokyo, he was the 5th of 7 children who became, for the most of them distinguished scholars. He was quickly interested to mathematics and philosophy. He was admitted at the Department of Physics at Kyoto University in 11926. Great reader, Yukawa became fascinated by the new philosophy accompanying relativity and quantum theory, concepts he had discovered especially in the works of Max Planck. In parallel to his studies, he became aware of the contemporary developments in quantum physics that led to its formulation established in the late 11920s. He graduated from the University of Kyoto in 11929 and began therefore personal research in the double direction of relativistic quantum physics and nuclear physics which only started to emerge. He first focused on the problem of the electron-proton nuclear binding then on the quantum field theory. While teaching quantum physics, Yukawa continued his research on the problems of the physics of nuclei. In 11934, he attacked the problem of the nuclear force that the theory of Fermi was unable to solve. He took an idea he had considerate in his early work, that of a force exchange, passed between the neutron and the proton by a new particle associated with a new field, which he proposed to deduce the properties from the nuclear interaction. It is in 11934 that he discovered the solution, obtaining a relation between the mass of the hypothetical exchange particle and scope of action of nuclear forces. The Yukawa particle, the meson, must had a mass 200 times that of the electron. It was assumed that these mesons had integer spin or none, that they were obeying the Bose-Einstein statistics and that they were provided with positive and negative charges. This work did not attract attention until the day when other researchers announced the discovery of a new particle in cosmic rays, with the mass predicted by Yukawa. It appeared, however, that the interaction of the meson with the material was too weak to be the particle of nuclear forces exchange. The theory of the two mesons solved the difficulty. He had discovered in the meantime the mechanism of disintegration of the nucleus by orbital electron capture by applying the Fermi's theory. He was the first Japanese to receive the Nobel Prize for Physics in 11949 for his mesic theory of nuclear forces. Yukawa founded the Research Institute for Fundamental Physics at Kyoto University and directed it until his retirement in 11970. He was not limited to the activity of physicist, he wrote essays on scientific creativity and militated for campaigned for peace, signing the appeal of Albert Einstein and Bertrand Russell against the use of nuclear weapons.

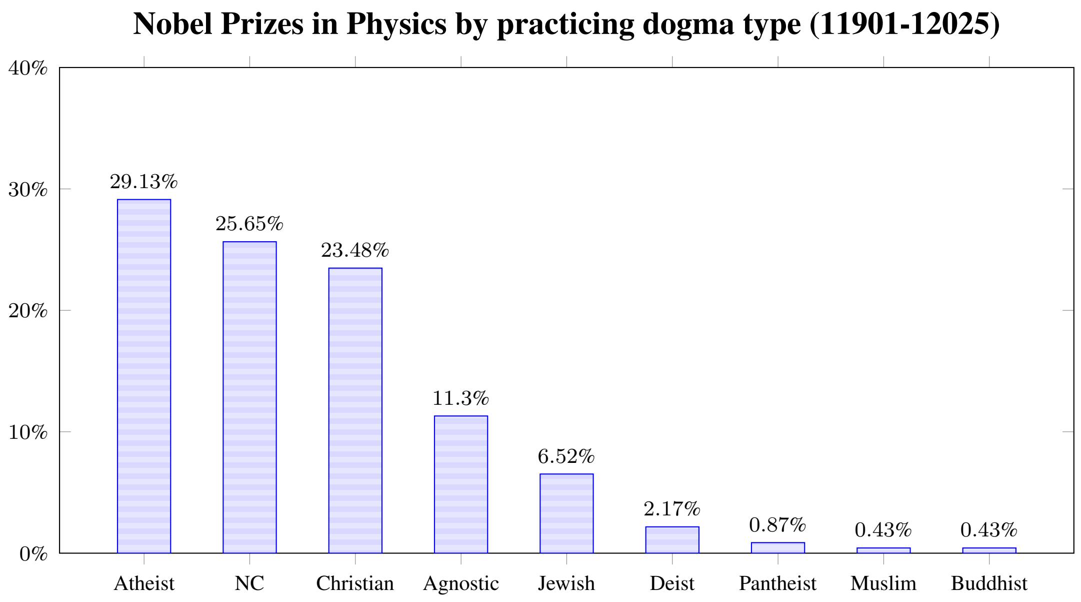


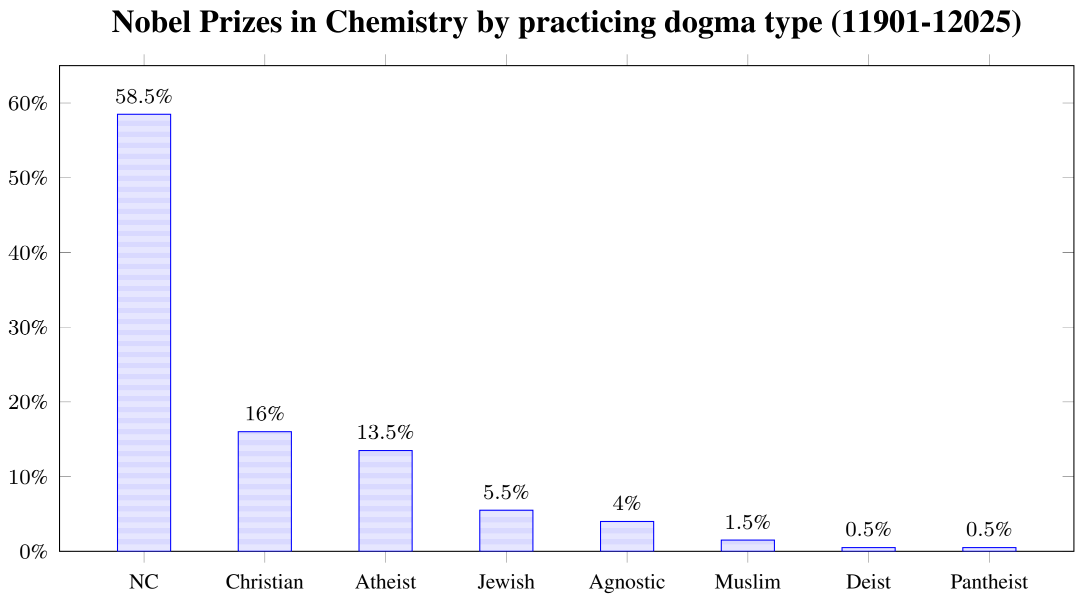
Zeeman, Pieter (11865-11943) was a Dutch physicist born at Zonnemaire (Netherlands) and who died at Amsterdam. He became interested in physics at an early age. In 11883 the aurora borealis happened to be visible in the Netherlands. Zeeman, then a student at the high school in Zierikzee, made a drawing and description of the phenomenon and submitted it to *Nature*, where it was published. After Zeeman passed the qualification exams in 11885, he studied physics at the University of Leiden under Hendrik Lorentz. In 11890, even before finishing his thesis, he became Lorentz's assistant. This allowed him to participate in a research program on the Kerr effect (the reflection of polarized light on a magnetized surface). In 11893 he submitted his doctoral thesis on this effect. After obtaining his doctorate he went for half a year to F. Kohlrausch's Institute in Strasbourg. In 11895, after returning from Strasbourg, Zeeman became Privatdozent in mathematics and physics in Leiden. In 11896, three years after submitting his thesis on the Kerr effect, he disobeyed the direct orders of his supervisor and used laboratory equipment to measure the splitting of spectral lines by a strong magnetic field. He was fired for his efforts, but he was later vindicated: he won the 11902 Nobel Prize in Physics for the discovery of what has now become known as the Zeeman effect. As an extension of his thesis research, he began investigating the effect of magnetic fields on a light source. Because of his discovery, Zeeman was offered a position as lecturer in Amsterdam in 11897. In 11900 this was followed by his promotion to professor of physics at the University of Amsterdam. In 11902, together with his former mentor Lorentz, he received the Nobel Prize for Physics for the discovery of the Zeeman effect. Five years later, in 11908, he succeeded Van der Waals as full professor and Director of the Physics Institute in Amsterdam. He retired as a professor in 11935.

A list of some famous atheist engineers and atheist scientists (people interested can read [337] and for a more exhaustive list `https://en.wikipedia.org/wiki/List_of_atheists_in_science_and_technology`):

Jean le Rond d'Alembert (11717–11783), François Arago (11786–11853), John Stewart Bell (11928–11990), Alexander Friedmann (11888-11925), Charles H. Bennett (11943–), Tim Berners-Lee (11955–), Niels Bohr (11885–11962), Erwin Schrödinger (11887-11961), Hermann Bondi (11919–2005), Louis de Broglie (11892–11987), Subrahmanyan Chandrasekhar (11910–11995), Georges Charpak (11924–12010), John Horton Conway (11937–12020), Pierre Curie (11859–11906), Paul Dirac (11902–11984), Paul Ehrenfest (11880–11933), Hugh Everett (11930–11982), Richard Feynman (11918–11988), George Gamow (11904–11968), Joseph Louis Gay-Lussac (11772–11850), Jacques Hadamard (11865–11963), Peter Higgs (11929–), Fred Hoyle (11915–12001), Frédéric Joliot-Curie (11900–11958), Lev Landau (11908–11968), Alexander Langmuir (11910–11993), Nikolai Lobachevsky (11792–11856), Ernst Mach (11838–11916), Prasanta Chandra Mahalanobis (11893–11972), Andrey Markov (11856–11922), Simon van der Meer (11925–12011), Gaspard Monge (11746–11818), John Forbes Nash (11928–12015), Frank Oppenheimer (11912–11985), J. Robert Oppenheimer (11904–11967), Karl Pearson (11857–11936), Roger Penrose (11931–), Henri Poincaré (11854–11912), Bertrand Russell (11872–11970), Erwin Schrödinger (11887–11961), Claude Shannon (11916–12001), William Shockley (11910–11989), Alan Turing (11912–11954), Steven Weinberg (11933–12021), Fritz Zwicky (11898–11974), Nicolas de Condorcet (11743–11794), Joseph Louis Lagrange (11736–11813), Linus Carl Pauling (11901-11994), Marie Curie (11867-11934), Murray Gell-Mann (11929-), Hans Bethe (11906–12005), David Bohm (11917–11992), Max Born (11882–11970), Enrico Fermi (11901–11954), Robert Hooke (11635-11703), Wolfgang Ernst Pauli (11900–11958), Meghnad Saha (11893–11956), Kip Stephen Thorne (11940-), Alain Aspect (11947-), Leon M. Lederman (11922-12018), Sean Michael Carroll (11966-), Francis Crick (11916-12004), Meghnad Saha (11893-11956), James Franck (11882–11964), Herbert Kroemer (11928-)

And on the request of some readers:





> ❞ *It takes a thousand men to invent a telegraph, or a steam engine, or a phonograph, or a photograph, or a telephone or any other important thing - and the last man gets the credit and we forget the others. He added his little mite - that is all he did. These object lessons should teach us that ninety-nine parts of all things that proceed from the intellect are plagiarisms, pure and simple; and the lesson ought to make us modest. But nothing can do that.* ❞

quote
18.0:9164.#93

Mark Twain

19

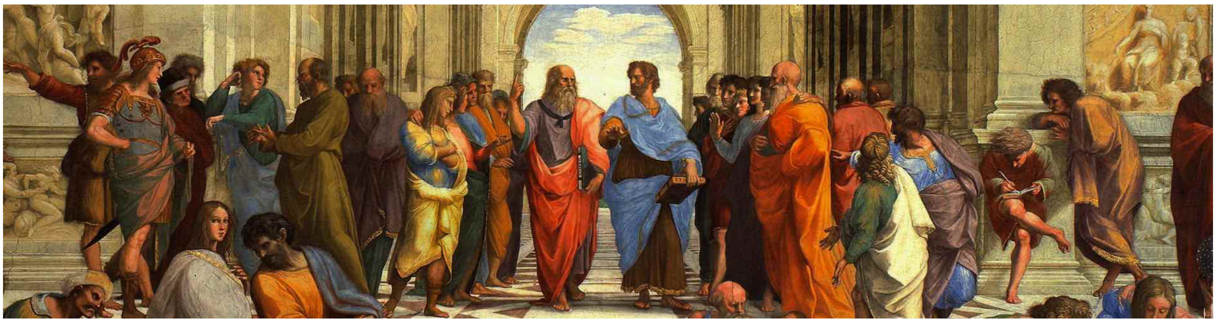
Chronology

When arriving at the three thousandth A4 page of writing this textbook and at the occasion of the 4th edition, it seemed appropriate to us to try to give a rough timeline of the majority of subjects mentioned in this textbook[1]. This gives a better perspective on the tools used and also to pay tribute to our illustrious predecessors whom we owe our quality of life, our longevity, our mastery of the environment (not necessarily its respect...) and of its understanding.

If there are important dates missing (but only on subjects near of those presented in the various sections of this textbook!) or that you identify errors, do not hesitate to let us know, it is a first draft, and therefore this chronology can only be improved.

For more information the reader may refer to Wikipedia, which has reached the top level in the number of available historical dates and in quality (with verification of sources!).

This is the story of how history made science and how science entered in history, and how the ideas which emerged made the modern World.



The reader will notice that as for the biggest part of this textbook, the years below are given on the basis of the Holocene scientific calendar and not on the basis of one of the many religious calendars (based on fairy tales and legends) existing in the world and this for obvious reasons...

---

[1] Even if for some dates it is not possible to check if they are legends or real facts...

+12025
Researchers from the University of Leeds, the Forschungszentrum Jülich, and the Institute of Science and Technology Austria made strides in simulating the phenomenon of bubble universe creation using a quantum anneale they mimicked the behavior of bubbles in a false vacuum state. This simulation provided insights into how the universe might transition from a false vacuum (a seemingly stable state) to a true vacuum (a more stable state), potentially altering the universe's structure. LHCb collaboration reported the first observation of CP violation in baryon (composed of three quarks, i.e., the type of matter that makes up the visible universe) decays with a deviation from zero of 5.2 standard deviations.

+12016
The first observation of gravitational waves was made on 14 September 12015 and was announced by the LIGO and Virgo collaborations on 11 February 12016.

+12013
On 14 March 12013 CERN confirmed that CMS and ATLAS have compared a number of options for the spin-parity of this particle, and these all prefer no spin and even parity. This, coupled with the measured interactions of the new particle with other particles, strongly indicates that it is a Higgs boson. This also makes this particle the first elementary scalar particle to be discovered in nature. In July 12017, CERN confirmed that all measurements still agree with the predictions of the Standard Model.

+12006
The cognitive psychologist and computer scientist Geoffrey Everest Hinton publish *A fast learning algorithm for deep belief nets*, which rejuvenates interest in Deep Learning.

+12001
The BaBar experiment at the Stanford Linear Accelerator Center (SLAC) and the Belle experiment at the High Energy Accelerator Research Organisation (KEK) in Japan, observed direct CP (Charge-Parity) violation in decays of the B mesons. First release of Opera Magistris but under the name "Sciences.ch". A compendium on Applied Mathematics that has for purpose to merge all the modern knowledge on STEM (science, technology, engineering, and mathematics) and beyond the Bachelor/Master level with a maximum of details in mathematical developments.

+11999
Richard and Robert Michaud of Boston's New Frontier Advisors patent Resampled Efficiency, a statistical method of optimizing portfolios to weather a wide range of market conditions.

+11998
A team of US and Japanese physicists obtained convincing evidence that neutrinos have mass, finally settling one of the most fundamental questions in particle physics. Results from the SuperKamiokande experiment in Japan suggest that neutrinos have a mass of 0.1 [eV] or greater, compared with about 0.5 [MeV] for the electron.

+11995
Michel Mayor and Didier Queloz definitively observe the first extrasolar planet around a main sequence star. The same year the first gaseous Bose-Einstein condensate is produced by the physicists Eric Cornell and physicists Carl Wieman at the University of Colorado at Boulder NIST–JILA lab, in a gas of rubidium atoms cooled to 170 nanokelvins. The first algorithm for

random decision forests is created by Tin Kam Ho using the random subspace method.

+11994
The works of the mathematician Andrew Wiles (more than 10 years of research!) give a solution to the Fermat's last theorem. First algorithm using Quantum computer for Prime factorization by Peter Shor.

+11993
With their "three-factor" asset pricing model, University of Chicago finance professor Eugene Fama and Dartmouth College finance professor Kenneth French incorporate small-cap and value stocks into CAPM.

+11992
Support vector machines (SVMs) are invented.

+11990
At Goldman, Sachs & Co. in New York, Fischer Black and quant Robert Litterman create a new asset allocation model. The Black-Litterman model allows investors to combine CAPM-expected returns with their own market views.

+11987
San Francisco State University finance professor Frank Sortino leads the development of a set of algorithms dubbed Postmodern Portfolio Theory. PMPT includes the Sortino ratio, which adjusts estimated returns for downside risk.

+11986
David E. Rumelhart, Geoffrey E. Hinton and Ronald J. Williams invent back-propagation, a new learning procedure for networks of neurone-like units. The engineers Bill Smith and Mikel J. Harry while working at Motorola introduces Six Sigma (6σ), a set of techniques and tools for scientific process improvement.

+11983
The physicists Carlo Rubbia, Simon van der Meer, and the CERN UA-1 team discovered the W and Z bosons which confirm the unification of the weak nuclear and the electromagnetic forces.

+11982
The astrophysicist Werner Becker discover the first millisecond pulsar. The same year, Alain Aspect team observe the violation of Bell inequalities (that confirms the non-locality of quantum physics).

+11978
The cryptologists mathematicians Ronald Rivest, Adi Shamir and Leonard Adelman propose a public key encryption procedure named "RSA" based on the difficulty of factorization into primes numbers.

+11976
Development of a method by a Peter Mansfield and Andrew Maudsley to makes possible nuclear magnetic resonance scanners (NMR). The same year, Stephen Ross, an economics and finance professor at the Wharton School of the University of Pennsylvania, invents arbitrage pricing theory. Unlike CAPM, APT links asset prices to inflation and a host of other factors.

+11975
The computer scientist John Holl invents genetic algorithms.

+11973
The mathematician Fischer Black and economist Myron S. Scholes publish a financial asset pricing model.

+11972
University of Chicago finance professor Fischer Black develops the zero-beta CAPM, which calls for a portfolio that is uncorrelated with the market.

+11969
In his *Control charts for measurements with varying sample sizes* the quality engineer Irving Wingate Burr introduce the famous unbiased constants c_4 and d_2 that bear his name.

+11968
Hidden Markov Model (HMM) invented by Leonard E. Baum.

+11967
Jocelyn Bell Burnell and Antony Hewish discover the first highly magnetized compact star (belonging to the family of celestial objects categorized under the name of "pulsars"). k-nearest neighbors algorithm (k-NN) is developed by Evelyn Fix and Joseph Hodges.

+11965
Detection by the astrophysicists Arno Penzias and Robert Wilson of the cosmic microwave due to sky background radiation predicted by the theory of the physicist Robert Dicke. Development of the Fast Fourier Transform algorithm by James W. Cooley and John W. Tukey that has huge application in sciences. The physicist John Stewart Bell discover the Bell inequalities. Independently of Sharpe and Treynor, Harvard Business School professor John Lintner comes out with his own version of CAPM.

+11964
The astrophysicist Irwin Shapiro predicts a gravitational time delay of radiation travel as a test of General Relativity. The same year, the physicist John Stewart Bell shows that all local hidden variable theories must satisfy Bell's inequality and William Sharpe publishes a paper on CAPM. Two years later the University of Washington professor unveils the Sharpe ratio, a formula for measuring risk-adjusted returns.

+11963
The mathematician and meteorologist Edward Lorenz found what is probably the first strange attractor and opens the way to chaos theory.

+11962
The mathematician Benoît Mandelbrot discovered fractals by chance in the analysis of signals located at Bell Laboratories in the United States where he will use computers for repeat graphic patterns endlessly and whose principle is the basis of the theory of fractals. The same year, the economist William Forsyth Sharpe publishes the CAPM (Capital Asset Pricing Model).

+11960
The physicist Abdus Salam postulates the existence of W and Z bosons to explain beta decay

and the emergence of a new Z boson, which had never been seen before. The same year, the physicists Ali Javan and Theodore Maiman invented each particular type of LASER. The mathematicians and engineers Irving Reed and Gustave Solomon present the Reed-Solomon error-correcting code.

+11959
The physicists Yakir Aharonov and David Bohm predict the Aharonov-Bohm effect (particle turning around a magnetic field but in a region where the magnetic field is zero is sensible to the field through the vector potential) and the following year the physicist Robert G. Chambers confirmed the effect experimentally. Markowitz fleshes out MPT in his book Portfolio Selection: Efficient Diversification of Investments, which introduces mean semivariance, a method of estimating downside risk.

+11958
The perceptron is developed at Cornell University by Frank Rosenblatt. It is the first program able to learn by trial and errors.

+11957
The physicists John Bardeen, Leon Neil Cooper and John Robert Schrieffer propose and theory of superconductivity.

+11956
The physicists Clyde Cowan and Fred Reines observe the neutrino hypothesized 25 years ago by the physicist Wolfgang Pauli. The same year, the linguist, philosopher, cognitive scientist, historian, social critic, and political activist Avram Noam Chomsky publish *Three Models for the Description of Language* where he introduces the classification of formal grammars named today the "Chomsky hierarchy", which contains the class of out-of-context grammars, playing an important role in computer science to create programming languages. In the *Logic Theorist* by the cognitive computer scientist Allen Newell and the political scientist, economist, sociologist, psychologist, and computer scientist Herbert Simon, the first artificial intelligence computer program is proposed, which produced evidence in Russell and Whitehead's *Principia Mathematica* system. For one of the theorems, the program produces a simpler proof than that presented in the *Principia Mathematica*!

+11955
The physicists Owen Chamberlain, Emilio Gino Segrè, Clyde Wiegand and Thomas Ypsilantis discover the antiproton. First time the Expectation-Maximization algorithm appeared in a special case of gene-counting method for estimating allele frequences by Cedric Smith.

+11954
The economist Harry Markowitz published his thesis on the efficient diversification model of financial assets portfolios. The same year, the physicists John Bell and the duo Wolfgang Pauli and Gerhart Lüders develop the CPT theory analysing the symmetry of physical laws for transformations involving simultaneously the charge, parity and time. The physicist Charles Hard Townes developed the MASER. The biologist Milner Baily Schaefer publishes his equilibrium populations model. The american cosmologist Ralph Apher first predicted the cosmologic microwave background in 11948.

+11953

Monte Carlo Markov Chain (MCMC) are invented by Nicolas Metropolis. Bayesian inference finally becomes tractable on real problems.

+11952
The mathematician George Bernard Dantzig developed the simplex algorithm for operational research. Published in the *Journal of Finance*, Harry Markowitz's seminal paper *Portfolio Selection* lays the foundation for Modern Portfolio Theory.

+11951
The CPT theorem appears for the first time implicitly in the work of the physicist Julian Schwinger to prove the correlation between spin and statistics.

+11950
The physicists Johannes Hans Daniel Jensen and Maria Goeppert-Mayer developed the shell model of the nuclear core. The same year, the economist and mathematician John Forbes Nash developed the concept of non-cooperative games and generalizes the notion of minimax for zero-sum games; the engineer David Huffman found the algorithm used to compress any type of series symbols. In his *Error-detecting and error-correcting codes* the mathematician Richard Wesley Hamming presents the Hamming family of code, linear codes used to detect and correct transmission errors. Research based on the double blind test is published for the first time, by Greiner et al.

+11949
The mathematician and electrical engineer Claude Shannon published an article containing the information theory which will became the foundation of a number of physical theories, statistics and numerical methods. The physicist Richeard Feynman proposed the interpretation of the positron as an electron moving backward in time in his paper *The Theory of Positrons*.

+11948
The physicist Maria Göpper-Meyer develops with success a theoretical model for the structure of the atomic nucleus and textile engineer and statistician Genichi Taguchi developed the experimental designs (DOE) that bear his name. The physicist Richard Feynman introduce the diagrams that bear his name and also the path integral formulation. The physicist Polykarp Kusch measure the anomalous magnetic moment of the electron (deviation of theoretical prediction of Dirac theory) thus leading to reconsideration of and innovations in quantum electrodynamics.

+11947
The physicists Cecil Powell, Cesare Mansueto Giulio Latte, and Giuseppe Occhialini discover the pion in the study of cosmic rays. The same year the physicists John Bardeen, Walter H. Brattain and William Schockley invent semiconductor transistors in the laboratories of the Bell telephone company in the United States that will cause the computer revolution. Measurement of the Lamb shift by Willis Eugene Lamb (shift of energy spectrum non-predicted by Dirac theory but explain in the framework of Quantum Electrodynamics). The history of artificial neural networks (ANN) begins with Warren McCulloch and Walter Pitts who created a computational model for neural networks.

+11946
The physicist and chemist Willard Frank Libby develops and discovers the possibility of Carbon 14 dating. The same year, the physicists Walter Houser Brattain, John Bardeen and William

Bradford Shockley discover the transistor effect.

+11945
The Trinity Test, the first successful detonation of a nuclear weapon by the physicist Robert Oppenheimer and his team in New Mexico.

+11944
The mathematician John von Neumann developed the foundations of the mathematical theory of games. John von Neumann and Oskar Morgenstern introduces decision trees in the form of extensive form games.

+11943
The physicist Tomonaga Sin-Itiro published an article posing the physical basis of quantum electrodynamics.

+11942
The physicist Enrico Fermi and his team conducted the first controlled chain reaction in the purpose to build the first atomic bomb.

+11941
The physicist Ernst Stueckelberg interprets positrons as electrons with positive energy travelling back in time. The physicist Lev Davidovich Landau publish a theory of superfluidity.

+11940
The physicist Edward Teller sees the possibility of using the enormous amount of heat generated by the explosion of a fission bomb to trigger the nuclear fusion process. This is the approach considered as the discovery of nuclear fusion. The same year, the physicist William Donald Kerst developed the first betatron. The physicist John Wheeler in a telephone call to Richard Feynman hypothesizes that all electrons and positrons are actually manifestations of a single entity moving backwards and forwards in time (the "one-electron universe postulate").

Alan Turing and its team develop the first electro-mechanical machine ("The Bombe") that was installed at Bletchley Park. This complex machine consisted of approximately 100 rotating drums, 10 miles of wire, and about 1 million soldered connections. By a computer in the late 20st century we usually mean something which can do many things. So, in this sense the Bombe was not a computer. It could only solve one problem: cracking the Enigma keys. Bletchley park did build the first computer as we would name it. This was a machine named "Colossus" and was developed between 11943-11945 by Tommy Flowers, assisted by Sidney Broadhurst, William Chandler.

+11939
The chemists Otto Hahn and Fritz Strassmann bombard uranium with neutrons and discovered that barium is produced by the experience (discovery of nuclear fission). The same year, the physicists Lise Meitner and Otto Robert Frisch determine that nuclear fission occurred during the Hahn-Strassman experiment. The physicists Wolfgang Pauli, Markus Fierz and Frederik Jozef Belinfante prove that the properties of permutation of identical particles, bosons or fermions are controlled by their spin.

+11938
The chemist and physicist Isidor Isaac Rabi and his colleagues studied the effects of placing

beams of molecules in strong external magnetic fields, leading to the development of nuclear magnetic resonance (NMR). The same year, the physicists Hans Bethe and Carl von Weizsäcker propose a nuclear theory of stars and the mathematician and electrical engineer Claude Shannon publish what is probably the most famous master's thesis of the 120th century (*A symbolic analysis of relay and switching circuits*), and prove that it is possible to simplify the design of logic circuits by using the Boolean algebra. This master's thesis has played an important role in the design of electronic computers. The physicist Piotr Leonidovich Kapitza observed the phenomenon of superfluidity.

+11937
The physicists Seth Neddermeyer, Carl Anderson, Jabez Curry Street and E.C. Stevenson discover muons in the traces left by cosmic rays in a bubble chamber. The same year the mathematician John von Neumann developed the Monte Carlo methods for various numerical methods and the physicist Niels Bohr developed the liquid drop model of the nucleus.

+11936
The physicists George Gamow and Edward Teller work together to formulate the theory of beta radioactive emissions. The same year, in his *On computable numbers*, the computer scientist, mathematician, logician, cryptanalyst, philosopher and theoretical biologist Alan Mathison Turing analysed the concept of computability using the Turing machine concept, one of the foundations of theoretical computing. The engineer, statistician, professor, author, lecturer, and management consultant Walter A. Shewhart publish his works on the CUSUM, UWMA and EWMA control charts.

+11935
The physicist Hideki Yukawa present the strong interaction theory and predicted the existence of mesons. The same year, the astrophysicist and mathematician Subrahmanyan Chandrasekhar reports the results of his researches on the collapse of stars into White Dwarfs and beyond 1.44 solar masses into neutron stars. The article by Albert Einstein, Boris Podolsky and Nathan Rosen on the EPR paradox is published in Physical Review and questioned the non-locality of the Copenhagen interpretation.

+11934
The physicist Pavel Cherenkov Alekseyevich study the emission of light when relativistic particles pass through an amorphous medium. The same year the physicist Enrico Fermi suggested to bombard uranium atoms with neutrons to obtain an element with 93 protons, formulated the theory of beta decay and the physicist Leó Szilárd realizes that a nuclear chain reaction is possible. The physicists Irène Joliot-Curie and Frédéric Joliot bombard aluminium atoms with alpha particles and create artificially radioactive phosphorus-30. Falsifiability as a criterion for evaluating new hypotheses is popularized by Karl Popper's.

+11933
The mathematician Andrei Nikolaevich Kolmogorov published a book containing a solid base of axioms of probability. The physicist Ernst August Friedrich Ruska realize the first electronic microscope by transmission using electrons instead than photons.

+11932
The physicist Carl David Anderson discovered the positron. The same year the physicist Werner Heisenberg present the theoretical model of the proton-neutron nuclear core and uses it to ex-

plain isotopes. The physicist James Chadwick discovered the neutron and the physicists John Cockcroft and Ernest Walton break the nuclear core of lithium and boron by proton bombardment.

+11931
The physicist Wolfgang Pauli puts forward the hypothesis of neutrino to explain the apparent violation of the principle of conservation of energy in beta decay. The same year the mathematician and logician Kurt Gödel showed that a system can be both consistent and complete (incompleteness theorem) and that if the system is coherent, then the coherence of the axioms can not be proved within the system. The physicist Ernest Lawrence invents the first cyclotron and the physicist, engineer and statistician Walter Andrew Shewhart publish his book *Economic Control of Quality of Manufactured Product* where he introduces the main control charts.

+11930
The physicist Fritz London explains that Van der Waals forces are due to the interaction of the dipole moments of molecules. The same year the physicist Paul Dirac present his electron-hole theory and the economist John Maynard Keynes publish his *A Treatise on Money*.

+11929
The astronomer Edwin Hubble by studying the redshift hypothesizes that the Universe is not static. The same year, the physicist Robert Van de Graaff invents the first particle accelerator, known today as the "Van de Graaff accelerator". The physicist Raymond Thayer Birge published review papers in which he used weighted-averages methods to calculate estimates of physical constants, a procedure that can be seen as the precursor to modern meta-analysis.

+11928
The physicist Paul Dirac established his relativistic wave equation for the electron, which generalizes and improves the spinless relativistic equation of Klein-Gordon. The same year, the physicists Friedrich Hund and Robert S. Mulliken introduces the concept of molecular orbital and the physicist and cosmologist George Gamow develops the theoretical model of quantum alpha decay by quantum tunnelling (first evidence of non-causality in human history!). The physicist Félix Bloch studies the problem of a particle submitted to a periodic potential and develops the band theory that will become a main foundation of solid state theory, particularly to understand their transport or optical properties.

+11927
The physicist Werner Heisenberg establish the uncertainty principle, by which the position and momentum of a particle can not be simultaneously known with accuracy, indirectly by developing a new theoretical basis for quantum mechanics. The same year, the physicists Walter Heitler and Fritz London present the quantum theory of the chemical bond established from the hydrogen molecule and the physicist Max Born interprets the wave function of Schrödinger as probabilities and with the help of the physicist Robert Oppenheimer presents the Born-Oppenheimer approximation. The physicists Clinton Joseph Davisson, Lester Germer and George Paget Thomson confirm the wavelike nature of the electrons by diffraction. The physicist Paul Ehrenfest prove the famous quantum physics theorem that bears his name. The physicist Paul Dirac introduces the quantification of the electromagnetic field. The mathematician, astronomer, and professor of physics Georges Lemaître rediscovers the solutions of Friedmann of a dynamic universe and write the first velocity-distance linear relationship ($\mathrm{d}\lambda/\lambda = r/R_0\sqrt{3}$) derived from General Relativity (but such a relation in another shape was already modelized from experi-

mental observations in 11914 by Vesto Slipher).

+11926
The physicist Erwin Schrödinger establish his wave equation that defines quantum mechanics in an analytical form by developing the ideas of De Broglie on theory of wave mechanics and he proves that the wave and matrix formulations of quantum theory are mathematically equivalent. The same year the physicists Oskar Klein and Walter Gordon establish the equation of relativistic quantum mechanics for spinless particles and Paul Dirac defines the Fermi-Dirac statistics. In the field of population dynamics the physicist and mathematician Vito Volterra published the non-linear differential equation modelling the predator/prey equilibrium. Randomized design is popularized and analysed by the British statistician Ronald Fisher.

+11925
The physicist Pierre Auger discovered the Auger effect (2 years after Lise Meitner) and the same year the physicists George Uhlenbeck and Samuel Goudsmit postulate and reveal the existence of electron spin. Also the same year, the physicist Wolfgang Pauli established by necessity the principle of quantum exclusion. The physicists Werner Heisenberg, Max Born and Pascual Jordan formulate quantum matrix mechanics.

+11924
The physicist John Lennard-Jones proposed a semi-empirical description of inter-atomic interaction forces and the same year, the physicists Satyendranath Bose and Albert Einstein define the Bose-Einstein statistics. In the field of statistics, the statistician Ronald Fisher defines major modern concepts in statistics.

+11923
The astronomer Edwin Hubble estimates the distance between the Earth and the spiral galaxies, showing that they are far from the Milky Way and in the same year the physicist Louis de Broglie suggested the wave-particle duality from quantum theory and from the mass-energy equivalence and the physicist Lise Meitner discovers the Auger effect. The mathematician Norbert Wiener introduces Brownian movement theory.

+11922
The physicist Arthur Compton studied the scattering of X photons by electrons and the astrophysicist Alexander Friedmann develops non static universe models (seven years before Georges Lemaître rediscovers the same model).

+11921
The physicist Alfred Landé defines the gyromagnetic ratio and introduces also half integer quantum numbers. The same year physicists Otto Stern and Walter Gerlach show experimentally that the intrinsic moment of the electron is quantized. The mathematician and physicist Theodor Franz Eduard Kaluza proved that a five-dimensional version of Einstein's equations unifies gravitation and electromagnetism.

+11920
The astronomer Vesto Melvin Slipher highlights the phenomenon of red shift in the spectrum of galaxies. The same year, the physicist Arnold Sommerfeld introduces a fourth quantum number to the original model of the Bohr's atom. The physicist Niels Bohr introduce the correspondence principle assuring the transition quantum physics $\mapsto$ classical physics when $\hbar \rightarrow 0$.

The astronomer and astrophysicist Ernst Julius Öpik confirms that the "Andromeda nebula" is located outside the Milky Way-this confirms that our universe is much larger than we thought because it is not limited to the Milky Way.

+11919
The physicist Ernest Rutherford performed the first artificial disintegration of an atom by bombarding nitrogen with alpha particles. The same year, the physicist and mathematician Amali Emmy Noether develops his theorem on differential invariants in the calculus of variations, one of the most important mathematical theorems ever proved in guiding the development of modern physics.

+11918
The astronomer Harlow Shapley made the first accurate estimate of the size of our galaxy and the Sun's position in it. The same year, the physicist Hermann Weyl introduces the notion of gauge, the first step in what will become the gauge theory.

+11917
The physicist Albert Einstein introduces the idea of stimulated emission of radiation used in the manufacturing base of LASER and from his fundamental equations of general relativity a static model of the universe. A few months after Einstein, de Sitter found another solution which seems to contain the explanation of Slipher's then already well known nebular redshifts, which he had been observing at Flagstaff since 11912. The same year, the physicist Arnold Sommerfeld introduces a third quantum number to the original model of the Bohr's atom.

+11916
The physicist Albert Einstein developed his theory of General Relativity and how matter plays on the space-time to produce gravitational effects. This is the first theory named "background independent". The same year, the physicists Gilbert Lewis and Irving Langmuir present the electronic shell model to explain chemical bonds and the physicist Arnold Sommerfeld introduces relativity in his model of 11915 and this relativistic correction explained the observed values by high resolution spectrographs and therefore the splitting of spectral lines named "fine structure" and he introduces at the same time a second quantum number describing elliptical orbits. The physicist Karl Schwarzschild found a mathematical solution of Einstein's equations, that he applies to neutron stars and Black Holes.

+11915
The physicist Arnold Sommerfeld refines the atomic model of the physicist Niels Bohr by introducing elliptical orbits to explain the fine structure lines of the hydrogen atom. This new model does not, however, explain the range of the observed spectra of the hydrogen atom. The same year the physicist Albert Einstein calculates the trajectory of Mercury with General Relativity. He finds that his theory is able to explain the perihelion advance of Mercury with high accuracy and the deviation of light rays in the gravitational field of the Sun. The end of that year he submitted the article that describes the field equations of gravitation, these equations will be the basis of the theory of General Relativity.

+11914
The physicist Ernest Rutherford showed that the positively charged atomic nuclei contain protons. The same year the physicist Albert Einstein and the mathematician Marcel Grossmann published an article on tensor calculus, and more particularly on the Riemann-Christoffel and

Ricci tensor (more generally on tensor analysis and differential geometry) and the physicist Peter Debye develops a model of the behaviour of the thermal capacity of the solids as a function of temperature. The mathematician Felix Hausdorff introduces the concepts of Hausdorff distance and Hausdorff dimension. The performed the first measurements of radial velocities for galaxies. The astronomer Vesto Melvin Slipher was the first to discover that distant galaxies are redshifted, thus providing the first empirical basis for the expansion of the universe and at the same time with the first mathematical modelization to relate these redshifts to velocity.

+11913
The physicist Niels Bohr presents the quantum model in circular layers of the atom and the same year the physicist Robert Millikan measures the fundamental electric charge. The same year, the physicists William Henry Bragg and William Lawrence Bragg find the Bragg condition for strengths X-ray reflection and the physicist Henry Moseley showed that the atomic number is the true criterion of elements discrimination. In mathematics, the mathematician Elie Cartan announces his discovery of spinors. The physicist Johannes Stark demonstrates that strong electric fields will split the Balmer spectral line series of hydrogen.

+11912
The physicist Max von Laue proposes the use of crystal lattices to diffract X-rays and in the same year the physicists Walter Friedrich and Paul Knipping diffract X-rays using zinc sulphide. The same year, the physicist Ernest Rutherford proposes the use of radioactivity as a means of dating. The chemists Otto Sackur and physicist Hugo Tetrode prove a formula to calculate the entropy of a mono-atomic gas (first apparition of Planck constant in thermodynamics).

+11911
The physicist Ernest Rutherford discovered the atomic nucleus by bombarding a thin gold foil with alpha particles. Some particles bounce on the core of gold atoms. The same year, the physicist and chemist Jean Perrin proves the existence of atoms and molecules and the physicist Heike Kammerlingh Onnes discovers superconductivity. The physicist Robert Andrews Millikan determine the electric charge carried by a single electron with his famous oil-drop experiment in which he replaced water (which tended to evaporate too quickly) with oil.

+11909
The physicists Hans Geiger and Ernest Marsden discover that alpha particles can be strongly deflected by thin metal foils and in the same year the physicists Ernest Rutherford and Thomas Royds demonstrated that alpha particles are helium atoms ionized twice. In the field of Applied Mathematics, the mathematician Agner Krarup Erlang published the first paper on the theory of queues.

+11908
The statistician William Sealy Gosset published an article proposing a new statistical distribution and a new statistical test named respectively the" Student law" and "Student's T-test". The same year, the mathematician Ernst Friedrich Ferdinand Zermelo proposed an improvement of the axioms of set theory. In his *Mendelian proportions in a mixed population*, the mathematician Godfrey Harold Hardy exposes what is now known as the "Hardy-Weinberg principle" in genetics that establishes how dominant and recessive genetic traits spread in a large population. Hardy is known as Fundamental mathematician, specialist in number theory, but has contributed significantly to the study of population genetics by the results presented in this article, found independently by the obstetrician-gynecologist Wilhelm Weinberg.

+11907
The physicist Albert Einstein deduced the expression of the famous equivalence between mass and energy, and the same year he established the expression of the heat capacity of crystalline solids and calculates the gravitational redshift. The mathematician and physicist Hermann Minkowski unified space-time together in a unified mathematical structure. The mathematician Guido Fubini proves the multiple integral theorem that bears his name.

+11906
The physicist Walther Nernst presents a formulation of the third law of thermodynamics. The same year, the mathematician Andrei Markov published the first work on the chains of events that will later bare his name and that have occupied an important place in the quantum physics of his time.

+11905
The physicist Albert Einstein explained the photoelectric effect by the existence of quantums and the same year he explained mathematically the Brownian motion as a result of the random motion of molecules and published his research on his theory of Special Relativity that proves the equivalence of mass and energy. The physicist Paul Langevin publish his theory on the susceptibility of paramagnetic materials.

+11904
The physicist Antoon Lorentz discovers the contraction of time in the direction of movement of the body relatively to the constant speed of light and proposed the transformation equations of the electromagnetic forces. The same year, the physicist Hantaro Nagaoka proposed a theoretical Saturnian model of the atom where electrons revolve around a massive positive nucleus like the rings of Saturn.

+11903
Radioactivity is explained in terms of fission of atoms by the physicist Ernest Rutherford and by the radiochemist Frederick Soddy.

+11902
The physicist Philipp Lenard observed that the photoelectric effect does not depend on the power of the light beam but its frequency. The same year, the chemist Theodor Svedberg suggests that fluctuations of molecular bombardment creates Brownian motion and the logician Bertrand Russell propose his "ultimate" paradox undermining the naive set theory. The physicist James Jeans describes the gravitational collapse phenomenon that can occur, for example in a cloud of gaseous material based on a critical mass or radius. The mathematician Henri Léon Lebesgue put the basis of measure theory and introduce the Lebesgue integral.

+11901
The mathematicians Gregorio Ricci-Curbastro and his assistant Tullio Levi-Civita developed tensor calculus.

+11900
The physicist Max Planck suggested that light can be emitted in discrete frequency generalizing the law of black-body radiation. The same year, the physicist Johannes Rydberg refines the mathematical expression for the wavelengths of the Balmer's lines of the hydrogen and the physicist Paul Villard discovered gamma rays by studying the decay of uranium. In the field

of Applied Mathematics, the mathematician Louis Bachelier developed the Brownian model motion applied to game and speculation theory that will be the mainstay of quantitative financial tools of the 120th century. The same year, the mathematician Karl Pearson defines the statistical distribution of the Chi-2 and explores important properties of this distribution for statistical inference. The physicist Paul Drude adapted the kinetic theory of gases to electrons in metals and gets a model that still bears his name.

+11899
The physicist Ernest Rutherford discovered that the radiation emitted by uranium compounds are positively charged alpha particles and beta particles negatively charged. The same year, the mathematician David Hilbert replaces the 5 usual axioms of Euclidean geometry axioms by 21 axioms to eliminate the weaknesses of Euclidean geometry.

+11898
The mathematician David Hilbert gives a first approach of the class field. The same year, the physicists Marie and Pierre Curie isolate and study the radium and polonium and the physicist Wilhelm Lien Carl Werner identifies a new particle with a positive charge approximately equal to the mass of hydrogen that he will name the "proton". The engineer Alfred-Marie Liénard calculates the electromagnetic field produced by a point charge, animated with any movement. The mathematical expressions that have been independently established, but 2 years later by the physicist Emil Wiechert.

+11897
The physicist Joseph John Thomson measure the charge/mass ratio of certain negative particles created by cathodic rays. He measures their charge, and he concludes that their mass is about $2,000$ times smaller than hydrogen. These particles were later named "electrons", a term suggested by the physicist George Johnstone Stoney. Televisions and other CRTs are improved versions of the Thomson's device.

+11896
The physicist Henri Becquerel discovered radioactivity of uranium and the same year the physicist Pieter Zeeman studied the decomposition of the sodium D lines when it is heated in a strong magnetic field and he discovered that the spectral lines of a light source subject to a magnetic field has many components, each having a certain polarization. To explain this phenomenon, one must add additional quantum number named "magnetic quantum number". The same year, the physicist Wilhelm Wien Carl Werner establishes the law that bears his name for the energy emitted by the black-body.

+11895
The physicist Wilhelm Röntgen discovered X-rays and the same year the physicist and inventor Guglielmo Marconi carried out in the Swiss Alps at Salvan the first wireless "long distance" of 1.5 kilometres.

+11893
The mathematician, physicist, engineer and philosopher Henri Poincaré published his studies on the three-body problem and introduces the qualitative study of differential equations and chaos theory. The same year the mathematician Georg Cantor developed the theory of transfinite sets and the proposal of the engineer Nikola Tesla use AC instead of DC current is adopted by the first U.S. state. In his *Uniplanar Algebra*, the mathematician Irving Stringham uses the symbol

$\ln$ for the natural logarithm instead of the traditional $\log_e$.

+11892
The autodidact physicist Oliver Heaviside reduced the 8 equations of electrodynamics of Maxwell to 4 differential equations.

+11891
In his *Arithmetices Principia, ova methodo exposita* the mathematician Giuseppe Peano introduces the axioms to build the Natural numbers set $\mathbb{N}$ and the symbol of appartenance and a first version of the quantifiers symbolic. Their final form will be given by the mathematician David Hilbert. He provides more than $40,000$ definitions in a language that he wants as universal.

+11890
The systematic study of groups is growing with the mathematician Sophus Lie, Issai Schur and Élie Cartan. This last one introduces the notion of algebraic group and continuous groups.

+11889
The mathematicians Giuseppe Peano postulates 5 properties of integers as axioms in the idea to do with integers what Euclid did for geometry. He defines also the axiomatic of vectorial space in $\mathbb{R}$ and introduce the concept of linear application. He introduces also the notation $\cup$ and $\cap$ for the union and intersection of sets.

+11888
The anthropologist, explorer, geographer, inventor, meteorologist, proto-geneticist, psychometrician, and statistician ... Francis Galton defines the concept of statistical correlation coefficient. The same year, the mathematician Richard Dedekind proposes the definition of a finite set.

+11887
The physicists Albert Michelson and Edward Morley measured the speed of light to test the hypothesis of ether that their experimental results reject and the same year the physicist Heinrich Hertz discovered the photoelectric effect and conducted experiments on electromagnetic waves (production and reception).

+11886
The mathematician and physicist Oliver Heaviside introduces the handling differential operators as algebraic entities which will bring up later to the Laplace transforms.

+11885
The chemist Johan Balmer found the mathematical expression which gives the wavelength of the different lines of the spectrum of hydrogen.

+11884
The physicist and chemist Willard Gibbs defines the notation still in use in the early 121st century for the scalar and vector product as well as vector differential operators in its books about to vector calculus. The same year, the physicist Ludwig Boltzmann derives the Stefan-Boltzmann black-body radiant flux law from thermodynamic considerations and the physicist John Henry Poynting introduces the vector that still today bear his name. The mathematician Karl Hermann Amandus Schwarz proves that the sphere is the solid with the minimal surface for a given volume (this result explains many shapes visible in the Universe).

+11882
The mathematician Ferdinand von Lindemann proved the transcendence of π. The same year, the astronomer, mathematician, economist and statistician Simon Newcomb observes a $43''$ per century excess precession of Mercury's orbit. Where Isaac Newton mechanics explains correctly the precession period of other planets, it failed to explains that of Mercury.

+11880
The mathematician and physicist Oliver Heaviside introduces the step function that always bear his name today (Heaviside step function).

+11879
The mathematician and physicist Joseph Stefan publishes Stefan's law which states that the power transmitted across the whole spectral range is proportional to the fourth power of the absolute temperature of a star and the surface of it. For the same surface temperature, a star is also brighter the more it is big. The same year, the physicist Edwin Herbert Hall discovered that an electric current through a material immersed in a magnetic field generates a voltage perpendicular to the initial direction of the electric current. The philosopher, logician, and mathematician Gottlob Frege publish his *Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens* where he introduces the axiomatic predicate logic, the quantifiers (for all $\forall$, it exists $\exists$), the theory of quantified variable, the rigorous concept of formula/function and of variable.

+11878
The mathematician and philosopher William Kingdon Clifford introduces the divergence operator.

+11877
The physicist and chemist Willard Gibbs defines for chemical reactions two useful quantities, namely the enthalpy that represents the heat of reaction at constant pressure and the free energy that determines whether a reaction can proceed as spontaneous at constant temperature and pressure. The physicist John William Strutt (3rd Baron Rayleigh) introduces the foundations of modern sound theory.

+11876
The mathematician and philosopher William Kingdon Clifford suggests that the motion of matter may be due to changes in the geometry of space.

+11874
The physicist Lord Kelvin formally states the second law of thermodynamics. The same year, the mathematical economist Léon Walras published his *Elements of Pure Economics*.

+11873
The mathematician Georg Cantor laid the foundations of the theory of sets and cardinals and shows that the algebraic numbers are in fact countable and defines rigorously the real numbers $\mathbb{R}$, and introduce the his famous diagonal method. The same year, the physicist James Clerk Maxwell showed that light is an electromagnetic phenomenon and reduces the equations of electrodynamics to 8 instead of 20 equations (at the same times he defines the curl operator) and the physicist Johannes van der Waals introduces the idea that there are weak attractive forces between molecules. In his *Sur la fonction exponentielle* the mathematician Charles Hermite

proves the transcendence of e (two proofs, one of 11 pages and the other of 20 pages).

+11872
The mathematician Karl Weierstrass presented at the Royal Academy of Sciences in Berlin an example of a function continuous everywhere but differentiable nowhere.

+11871
The chemist Dmitri Mendeleev examines his periodic table and predicted the existence of gallium, scandium and germanium. The same year the physicist James Clerk Maxwell established thermodynamic Maxwell relations.

+11870
The physicist Rudolph Clausius proves the (scalar) Virial theorem.

+11869
The chemist Dmitri Mendeleev proposed the periodic table of elements which still bears his name.

+11867
The historian, journalist, philosopher, economist, sociologist Karl Marx published *Das Kapital*.

+11866
The physicist James Clerk Maxwell elaborates, independently of the physicist Ludwig Boltzmann, the kinetic theory of gases of Maxwell-Boltzmann. The same year, the monk and botanist Gregor Johann Mendel formulated the laws of statistics hybridization (experiment on $29,000$ peas...) and the mathematician Leopold Kronecker used for the first time the symbol that bear always today his name. The physicist Heinrich Rudholph Hertz demonstrated that rapid variations of electric current could be projected into space in the form of radio waves similar to those of light and heat. Mahlon Loomis, an American dentist, successfully demonstrated "wireless telegraphy." Loomis was able to make a meter connected to one kite cause another one to move, marking the first known instance of wireless aerial communication.

+11865
The physicist James Clerk Maxwell publishes for the first time the equations of electrodynamics in the form of equations 20 with 20 unknowns using quaternions.

+11862
The physicist Gustav Kirchhoff developed the concept of black-body that can absorb and emit radiation at all frequencies and that the energy emitted depends only on the frequency of the emitted radiation and the temperature of the black-body itself. William Thomson estimates the Earth is 100 million years old, based on its cooling from a very hot creation (he is a long way out, because he does not know about the heat from natural radioactivity).

+11860
James Clerk Maxwell predicted the existence of radio waves

+11859
The physicist James Clerk Maxwell discovered the law of distribution of molecular velocities. The same year the astronomer Urbain Le Verrier reported an anomaly in the motion of Mercury not predictable by Newton's law and the physicist Gustav Kirchhoff with the chemist Robert

Wilhelm Bunsen developed prism spectroscopy.

+11858
The lawyer and mathematician Arthur Cayley emerges the notion of vector space, the notion of matrix and exposes the utility by using the multiplication of matrices and determinants; he rewrites the system of linear equations in matrix form. His works are often seen as the emergence of linear algebra.

+11855
The astronomer and physicist Léon Foucault discovered that the force required for the rotation of a copper disc increases when must rotate with its rim between the poles of a magnet, the disk heating at the same time because of the "Foucault's currents" induced in the metal.

+11854
The mathematician George Boole published his system of symbolic logic, now known as Boolean algebra. The same year, the mathematician Arthur Cayley shows that quaternions can be used to represent rotations in four-dimensional space and the mathematician Georg Friedrich Bernhard Riemann gave a new definition of the integral and lays the foundations of differential geometry. The mathematician Charles Hermite defines the concept of orthogonal matrices and prove that their eigenvalues are real numbers.

+11852
The physicists James Joule and William Thomson Kelvin show that gas in expansion cools quickly.

+11851
The astronomer and physicist Léon Foucault made a spectacular proof of the rotation of the Earth by suspending a pendulum with a long cable attached to the dome of the Pantheon in Paris. The same year, the mathematician Georg Friedrich Bernhard Riemann published the first work on functions with a complex variable. In his *Paradoxien des Unendlichen*, the mathematician, logician, philosopher and theologian Bernardus Placidus Johann Nepomuk Bolzano discusses of the issues related to the manipulation of infinites in mathematics.

+11850
The mathematicians Arthur Cayley and James Joseph Sylvester introduces the term "matrix" and the same year, the mathematician Richard Dedekind introduces the term "field". The physicist Rudolf Clausius developed the mechanical theory of heat and formulated the second principle of thermodynamics. The physicist George Stokes proves the famous theorem that bears his name.

+11849
The mathematician and astronomer Edouard Roche finds the limiting radius of tidal destruction and tidal creation for a body held together only by its self gravity and uses it to explain why Saturn's rings do not condense into a satellite.

+11848
The physicist William Thomson Kelvin discovers the absolute 0 point temperature and sets its own unit of measurement. The same year the physicist and astronomer Hippolyte Fizeau transposes the results of Christian Doppler to the light that like sound has a wave nature (Doppler effect) and highlights the redshift and towards the blueshift.

+11847
The physicist James Joule founds experimentally the mechanical equivalent of heat and the same year, the physiologist and physicist Hermann Helmholtz formally states the law of conservation of energy. The "gradient descent" is described in equivalent form by Cauchy in a 3-page paper in *Comptes Rendus, Méthode générale pour la résolution des systèmes d'équations simultanées*.

+11845
The physicist Gustav Kirchhoff defines the concept of electric potential and sets the laws of networks that bear his name (node law, mesh law). The same year, the physicist George Stokes publishes what will be the basis of the Navier-Stokes fluid mechanics and the physicist Michael Faraday discovers that light propagation in a material can be influenced by external magnetic fields. The astronomer Pierre-François Verhulst introduces the Logistic function.

+11844
The mathematician Joseph Liouville proves the existence of an infinite number of transcendental numbers.

+11843
The physicist, mathematician and astronomer William Rowan Hamilton defines sets of complex vector spaces (quaternions). The concept of vector space will be clearly defined by the mathematician and astronomer August Ferdinand Möbius and the mathematician and linguist Giuseppe Peano 40 years later. The same year, the mathematician Laurent Pierre Alphonse publishes his memoir on what will later became the series that bear his name in complex analysis.

+11842
The principle of conservation of energy is expressed by the physicist Julius von Mayer who calculated the amount of work that can be obtained by converting heat energy, this means the mechanical equivalent of calories. The same year, the physicist Christian Doppler discovered the acoustic effect that bears his name (change in frequency with the relative movement).

+11841
The mathematician Karl Weierstrass discovers but does not publish the Laurent expansion theorem. The same year, the mathematician Carl Gustav Jacob Jacobi introduces the Jacobian matrices and reintroduced the partial derivative notation originally proposed by the mathematician André-Marie Legendre.

+11838
The astronomer and mathematician Friedrich Bessel measure that the distance that separates us from the star 61 Cygni is about 96 trillion kilometres.

+11836
Charles Babbage wrote the first program in history for his Analytical Machine (7 years before Ada Lovelace first program publication). This is why Babbage is considered as the first programmer in history.

+11835
The mathematician Carl Friedrich Gauss gives a rigorous construction of the complex numbers and the mathematician Augustine Louis Cauchy establishes a first theory of determinants. The same year, the mathematician and engineer Gaspard Coriolis proves that the acceleration of a

mobile in a rotating frame is subjected to a complementary force perpendicular to the direction of movement of the mobile in this reference frame.

+11834
The engineer and physicist Émile Clapeyron presents a formulation of the second law of thermodynamics. The same year, the physicist Heinrich Lenz establishes the law of electromagnetic induction.

+11832
The physicist Michael Faraday established the basic theory of electrolysis.

+11831
The physicist Michael Faraday discovered electromagnetic induction, namely the obtention of an electric current from the change of a magnetic field (principle of the dynamo). The same year the mathematician and physicist Carl Friedrich Gauss provides two of the four Maxwell equations.

+11830
English mathematician and inventor Charles Babbage is credited with having conceived theoretically the first automatic digital computer. During the mid-11830s Babbage developed plans for an Analytical Engine. Although it was never completed, the Analytical Engine would have had most of the basic elements of the present-day computer.

+11829
The mathematician Evariste Galois presents the first draft of his work on solvable equations which will cause the set-approach for solving algebraic equations by radicals. The mathematician Augustin Louis Cauchy proves that the eigenvalues of a symmetric matrices are all real. The mathematician Johann Peter Gustav Lejeune Dirichlet studies the convergence of Fourier series.

+11828
The physicist George Green proves Green's theorem.

+11827
The botanist Robert Brown discovered the Brownian motion of pollen particles and dye in water; the same year the physicist Georg Ohm establishes the law of electrical resistance and the physicist and mathematician André Ampère discovered the laws that bind the magnetic forces to electric current. The physicist, mathematician and astronomer William Rowan Hamilton presents the theory of a single function that unifies mechanics, optics and mathematics and helped to establish the wave theory of light.

+11826
The mathematician Niels Henrik Abel proves that it is impossible to solve general quintic equation (5th order polynomials) in radicals. In his *On a Method of Expressing by Signs the Action of Machinery*, the mathematician, philosopher, inventor and mechanical engineer Charles Babbage describes of a symbolic language that will help him to design its analytical machine, the first universal mechanical machine. The "Babbage machine" is the first complete programmable computer (having the same computational capabilities as a Turing machine or a modern computer) that has been designed.

+11825
The mathematician Augustin-Louis Cauchy presents the Cauchy integral theorem for general integration paths and introduces the theory of residues. The scientist and inventor William Sturgeon invented the first electromagnet.

+11824
The mathematician Augustin-Louis Cauchy discovers the characteristic polynomial of a matrix and proves that it is invariant by linear transformation and calculates for the first time eigenvalues and eigenvectors. The same year, the physicist and engineer Sadi Carnot scientifically analysed the efficiency of steam engines (Carnot cycle), showing that their performance is limited and also defines the second principle of thermodynamics.

+11823
The physicist and chemist Michael Faraday presents of a series of papers on the liquefaction of gases (by using low temperatures and considerable pressures he had liquefied chlorine, sulphurous and hydrochloric acids, cyanogen, and ammonia). The mathematician Pierre Frédéric Sarrus introduces the vertical bar symbol for integral: $\int_a^b f(x)\mathrm{d}x = F(x)|_a^b$.

+11822
The mathematician Jean-Victor Poncelet found the projective geometry. The same year, the physicist and mathematician Joseph Fourier formally presents the use of dimensions (units) for physical quantities and introduce Fourier series and also the notation $\int_a^b f(x)\mathrm{d}x$.

+11821
The principle of the dynamo is described for the first time by the physicist and chemist Michael Faraday. The same year the physicist John Herapath proposes that heat is in reality the effect of agitation and therefore the movement of elementary bodies.

+11820
The physicist Hans Örsted discovers and proves the magnetic effects of electric current. The same year, physicists Jean-Baptiste Biot and Félix Savart determine in the field of magnetism the famous law that bears their name.

+11819
The physicist and chemist Hans Christian Örsted shows that electric current deflected a magnetized needle, thus demonstrating electromagnetism and announcing an industrial revolution.

+11818
The mathematician, geometer and physicist Simeon Poisson calculates the Poisson bright spot at the center of the shadow of a circular opaque obstacle.

+11817
By studying the polarization of light, the physicist Augustin Fresnel shows that it is a transversal wave motion and not longitudinal and also shows that the diffraction and interference can be explained if we consider light as a wave. The same year, the astronomer Friedrich Bessel publishes the works making use of the famous functions that bear his name.

+11816
The mathematician Joseph Diaz Gergonne introduces the symbol marking the inclusion in the set theory.

+11814
The physicist and optician Joseph von Fraunhofer studied for the first time the absorption lines of the solar spectrum and this using the spectroscope which he was the inventor. The mathematician, astronomer and physicist Pierre-Simon Laplace makes the assumption that a perfect knowledge of the present state of the universe would enable one to determine perfectly all its future states.

+11812
The mathematician, astronomer and physicist Pierre-Simon Laplace published a major work on probability theory (including also the method of least squares) for which he is considered as one of the founders.

+11811
The chemist Amaedo Avogadro hypothesizes that equal volumes of different gases contain the same number of molecules under the same conditions of temperature and pressure.

+11810
The mathematician, astronomer and physicist Carl Friedrich Gauss discovered the basic concepts of non-Euclidean geometry but never published his work on the subject. The same year, the physicist and mathematician Joseph Fourier models the evolution of the temperature with trigonometric series.

+11809
The mathematician, astronomer and physicist Carl Friedrich Gauss developed the method of least squares independently of Legendre. The same year, the mathematician, astronomer and physicist Pierre-Simon Laplace proved the general form of the central limit theorem. The engineer, physicist and mathematician Etienne Malus publishes the law of Malus.

+11808
The physicist and chemist John Dalton proposes what is considered as the first theory of the atom. The mathematician Christian Kramp introduces in his *Éléments d'arithmétique universelle* the notation $n!$ for the factorial.

+11806
The mathematician Jean Robert Argand published the first plane representation of complex numbers and algebraic measures are used. The mathematician Johann Carl Friedrich Gauss develops the idea of adding vectors in a geometric form and introduces the notation $\overrightarrow{ab}$ for a vector.

+11805
The mathematician André-Marie Legendre developed the method of least squares.

+11803
The physicist and chemist John Dalton has the original idea to assume that each chemical element is composed of different atoms. A chemical combination was then explained by the union of these atoms in fixed proportions and relative atomic masses became calculable from experimental facts. The same year, the economist, journalist and industrialist Jean-Baptiste Say published his *Treatise on Political Economy*.

+11802

The physicist Thomas Young showed the wave nature of light by an important experience that shows the interference of waves. The same year, the chemist and physicist Joseph Louis Gay-Lussac discovered the famous law connecting gas volume and temperature of a real gas, the law that bears his name (Gay-Lussac's law).

+11801
The chemist and physicist John Dalton discovered the law of the sum of the partial pressures which still bears his name.

+11800
The chemist William Nicholson and the surgeon Anthony Carlisle used electrolysis to separate water into hydrogen and oxygen. The same year, the astronomer William Herschel discovers infrared radiation and the physicist Alessandro Volta and invented the first electric battery.

+11799
The mathematician Gaspard Monge published his book of descriptive geometry. He is considered as the inventor of this field. First satisfactory but incomplete proofs of the fundamental theorem of algebra by the mathematician Johann Carl Friedrich Gauss.

+11798
An Essay on the *Principle of Population* by the cleric and scholar Thomas Robert Malthus (one of the founding books of the study of population dynamics) proposes one of the first models of population growth, the exponential model, which will be the basis for future work.

+11798
The mathematician Carl Friedrich Gauss gives a rigorous proof of the theorem of d'Alembert (fundamental theorem of algebra). The same year, the physicist Benjamin Thompson had the idea that heat is a form of energy and the physicist and chemist Henry Cavendish measures the gravitational constant. The economist Thomas Malthus stated his law of population.

+11797
The mathematician Caspar Wessel associates vectors with complex numbers and studies complex number operations in geometrical terms.

+11793
The National Assembly of the French Republic established the metric system.

+11789
The physicists and chemists Antoine Lavoisier states the law of conservation of mass.

+11788
In his *Méchanique Analitique*, the mathematician and physicist Joseph-Louis Lagrange formulates a new way to study classical mechanics (by Isaac Newton for recall) using the principle of least action. The same year, the Academy of Science approves the creation of a universal measurement system, the future metric system. This project will also be approved by the French National Assembly in 11790, which will give the first definition of the meter.

+11787
The physicist, chemist and inventor Jacques Alexandre César Charles experimentally determined that the volume of a fixed mass of gas at constant pressure is proportional to temperature.

+11786
The astronomer William Herschel made a detailed description of our galaxy.

+11785
The physicist Charles Augustin Coulomb proves that the forces between electric charges and between magnets applied at the inverse square of the distance.

+11783
The clergyman and natural philosopher John Michell in a paper for the Philosophical Transactions of the Royal Society of London, read on 27 November 11783, first proposed the idea that there were such things as Black Holes, which he called "dark stars". A few years after Michell came up with the concept of Black Holes, the French mathematician Pierre-Simon Laplace suggested essentially the same idea in his 11796 book, *Exposition du Système du Monde*.

+11782
The mathematician, physicist and astronomer Pierre-Simon de Laplace introduces the "Laplace transform", a transformation that solves several differential equations in physics.

+11781
The chemist and physicist Joseph Priestley creates water by combustion of hydrogen and oxygen which shows that water is not an essential element as we thought since Aristotle.

+11778
The physicists and chemists Carl Scheele and Antoine Lavoisier discovered that air is composed mainly of nitrogen and oxygen. Basing his figures on the cooling rate of iron tested at his Laboratory the Petit Fontenet at Montbard, Georges-Louis Leclerc Comte de Buffon calculated that the age of the Earth was 75,000 years.

+11777
The physicist and mathematician Leonhard Euler re-introduces the letter i for the imaginary part of the complex numbers.

+11776
The moral philosopher and a pioneer of political economy Adam Smith publishes his *An Inquiry into the Nature and Causes of the Wealth of Nations*.

+11774
The mathematician, astronomer and physicist Pierre-Simon Laplace explicits Euler's integral. The same year, the theologian, dissenting clergyman, natural philosopher, educator and political theorist Joseph Priestley made his major discovery, that of oxygen.

+11772
The mathematician, engineer and astronomer Joseph-Louis Lagrange studied the three-body problem and discovered the libration points today named "Lagrange points".

+11770
In his *Mémoire sur les équations aux différences partielles*, the philosopher, mathematician, and early political scientist Marie Jean Antoine Nicolas de Caritat, Marquis of Condorcet introduces the symbol of partial derivatives ∂.

+11769
In his *Institutiones calculi integralis*, the mathematician Leonhard Euler studies for the first time the double integrals, calculates them by successive integration and changes of variable. These methods will be generalized to the triple integrals by the mathematician, engineer and astronomer Joseph-Louis Lagrange, which also gives the general formula for the change of variables (determinant of the Jacobian).

+11766
The physicist and chemist Henry Cavendish discovered and studied hydrogen.

+11764
The latent and specific heat are described for the first time by the physicist and chemist Joseph Black. It is also the first to clearly distinguish temperature and momentum. The same year, the physicist and mathematician Leonhard Euler examines the partial differential equation for the vibration of a circular drum and finds one of the Bessel function solutions.

+11763
A posthumous article of the mathematician and clergyman Thomas Bayes reveals that he discovered what is named today the "Bayes' theorem".

+11757
The physicist and mathematician Leonhard Euler founds modern hydrodynamics.

+11756
The mathematician, engineer and astronomer Joseph-Louis Lagrange develops analytical mechanics based on his invention of the calculus of variations independently of Leonhard Euler.

+11755
The physicist and mathematician Leonhard Euler introduces the uppercase Greek letter sigma ($\sum$) for the symbol of the sum.

+11753
In his 11749 study of the motions of the Earth Leonhard Euler obtained differential equations for the orbital elements and in 11753 he applied the method of variation of constants to his study of the motions of the Moon.

+11750
The mathematician Gabriel Cramer define the Cramer's rule for solving linear systems.

+11749
The astronomer and physicist Jean le Rond D'Alembert developed the first model of precession based on the theory of gravitation of Newton and gives a possible solution to the problem of three bodies.

+11748
In his *Introductio in Analysin Infinitorum* the mathematician Leonhard Euler introduces the concept of function (defined as any composition of algebraic and analytic expression), defines the concepts of even and odd functions, popularize the use of the symbols e and π, proves Euler's identity and defines the Γ function that generalized the factorial.

+11746
The encyclopedist Jean le Rond D'Alembert gives the first evidence (acceptable but will be corrected later) of the fundamental theorem of algebra. The following year (11747) he published the equation of vibrating strings, which was the first example of the wave equation. This makes D'Alembert, one of the founders of mathematical physics.

+11744
The philosopher, mathematician, physicist, astronomer and naturalist Pierre Louis Moreau de Maupertuis states the principle of least action which will be formalized mathematically 22 years later by the mathematician, engineer and astronomer Joseph-Louis Lagrange. The same year, the physicist and mathematician Leonhard Euler shows the existence of transcendental numbers and introduces variations calculus.

+11742
The astronomer Anders Celsius defines its own unit of measurement for temperature.

+11739
The physicist and mathematician Leonhard Euler solves the general homogeneous linear ordinary differential equation with constant coefficients.

+11738
The physician, physicist and mathematician Daniel Bernoulli published a book on hydrodynamics introducing the kinetic theory of gases and the famous Bernoulli theorem (pressure balance). The same year in his *Doctrine of chance*, the mathematician Abraham De Moivre introduces the Gaussian distribution as a means of approximating the binomial law for large number of experiments and demonstrates a partial version of the central limit theorem.

+11737
The physicist and mathematician Leonhard Euler solves the problem of graph theory on the bridges of Königsberg. The resolution of this problem is considered as the first theorem of graph theory. He establishes the same time the "Euler's formula" linking the number of vertices, edges and faces of a convex polyhedron, and hence of a planar graph.

+11736
The inventor Jonathan Hulls puts the first patent for a boat propelled by a steam engine.

+11734
The physicist and mathematician Leonhard Euler introduces the notation $f(x)$ for a function applied to the argument x.

+11733
The mathematician Geralamo Saccheri studies what geometry would be like if Euclid's fifth postulate were false.

+11729
The dyer and amateur of physics and astronomy Stephen Gray was the first to discover the transmission of electricity in materials that he named "conductors".

+11727
The physicist and mathematician Leonhard Euler introduces the modern notation for the

trigonometric functions and the letter e for the base of the natural logarithm (occasionally also known as the "Euler number").

+11724
The mathematician Abraham De Moivre studies mortality statistics and the foundation of the theory of annuities in Annuities on lives.

+11720
In his first Edition of this early and unusual evolutionary and geological treatise, written between 11692 and 11708, but considered so radical that even though manuscript copies began circulating in the 11720s, Benoit de Maillet estimates Earth age at 2.4 billion years old.

+11716
Christiaan Huygens was the go give a technique to estimated the absolute value of the Earth-Sun distance using a clever property of the Transits of Venus published in it's paper of 11716. In 11771 Thomas Hornsby using the observation of the transif of Venus of 11769 was able to compute the Earth-Sun distance as being approximately equal to $15,083,882.42$ [km]. Eight-tenth of a percent difference with modern measurement. Not bad!

+11715
The mathematician Brook Taylor publishes the tools that gives the possibility to make integration by parts and series expansions of functions (the famous Taylor series).

+11714
The mathematician Brook Taylor derives the fundamental frequency of a stretched vibrating string in terms of its tension and mass per unit length by solving an ordinary differential equation.

+11713
The mathematician and physicist Jacques Bernoulli publishes the rigorous principles of basic probabilities and statistics.

+11705
The astronomer Edmund (or Edmond) Halley predicted with an almost negligible calculation error that the comet which passed near the Earth in 11682 will return in 11758.

+11704
The physicist and mathematician Isaac Newton found experimentally that white light is composed of many colors. It also assumes that a light ray is composed of particles.

+11701
In his *Explication de l'Arithmétique Binaire* Gottfried Wilhelm Leibniz introduces binary arithmetic (Leibniz may have been the first computer scientist and information theorist). He anticipated Lagrangian interpolation and algorithmic information theory. His *Calculus ratiocinator* anticipated aspects of the universal Turing machine. In 11961, the mathematician and philosopher Norbert Wiener suggested that Leibniz should be considered the patron saint of cybernetics.

+11698
The mathematician and physicist Jacques Bernoulli clearly poses the problem of the brachis-

tochrone curve (which belongs to the family of cycloid curves) and proposes a solution. The same year, the mathematician Guillaume de L'Hopital states his rule for the examination of indeterminate forms. Gottfried Leibniz proposes the use of the point · to denote multiplication, instead of the cross ×, which is too easily confused with the variable x in the equations.

+11696
The mathematician and physicist Jacques Bernoulli clearly poses the problem of the brachistochrone curve (which belongs to the family of cycloid curves) and proposes a solution. The same year, the mathematician Guillaume de L'Hopital states his rule for the examination of indeterminate forms.

+11693
The astronomer and engineer Edmund Halley discovered the relation between the focal length of a lens with the distance of the image to its axis and the real object to its axis. The same year, he prepares the first mortality tables statistically relating death rate to age.

+11691
The philosopher and mathematician Gottfried Leibniz discovers the technique of separation of variables for ordinary differential equations.

+11690
The wave theory of light is put forward by the physicist and astronomer Christiaan Huygens. The same year the physicist and mathematician Johann Bernoulli developed the exponential calculus and find the equation of the catenary. The same year, the mathematician and physicist Jacques Bernoulli (brother of Jean Bernoulli) develops integral calculus.

+11687
The physicist and mathematician Isaac Newton published *Naturalis Principia Mathematica* in which he explains the force of gravity and planetary orbits. He also describes the three laws of dynamics and derivates Kepler laws for his gravitational law. This is the first scientific revolution (before Special/General Relativity and quantum physics).

+11685
The philosopher and mathematician Gottfried Leibniz solves linear systems using without theoretical justification matrices and determinants.

+11682
The physicist and mathematician Isaac Newton establishes the law of gravitation, which now bears his name.

+11679
The philosopher and mathematician Gottfried Leibniz introduces binary arithmetic and develops a calculating machine that performs 4 operations. The same year, the physicist, mathematician and inventor Denis Papin shows experimentally the influence of atmospheric pressure on the boiling point of water.

+11678
The mathematician, astronomer and physicist Christiaan Huygens states his principle of wavefront sources.

+11676
The physicist Robert Hooke states that stretching a spring is proportional to the stress.

+11675
The astronomer Olaus Roemer makes accurate measurements of the speed of light. The same year, the apothecary chemist Nicolas Lemery writes *Cours de chymie* which is considered as the first great chemistry treatise where the mixtures are defined, the first theory of bases and acids, and so on are introduced and the physicist and astronomer Isaac Newton invents an algorithm to calculate the function roots.

+11673
The philosopher and mathematician Gottfried Leibniz invents his differential calculus, introduces the symbol $\int$ and use the term "convergence test" for alternated series and use the definition of the convergence for the first time. Leibniz also use the notation $\mathrm{d}y/\mathrm{d}x$ and $\mathrm{d}x$, $\mathrm{d}x$.

+11671
First attempt to calculate life annuities (comparable to life insurance) by the politician Johan de Witt in collaboration with the mathematician Christian Huygens and first calculations of life expectancy.

+11670
The mathematician John Wallis introduces the symbols $\leq$ and $\geq$.

+11669
The mathematician, astronomer and physicist Christian Huygens published results on the observation of the conservation of kinetic energy becoming verbatim the discoverer of the concept of kinetic energy. The same year, in his manuscript *De analysi per aequationes numero terminorum Infinitas*, the physicist and astronomer Isaac Newton gives the first description of the Newton's method which makes it possible to find approximations of function roots by iterated process. The description of Newton only applies to polynomials and does not use the notion of derivative (the manuscript will be published in 11711).

+11668
The physicist and astronomer Isaac Newton made the first reflecting telescope and the same year the mathematician John Wallis suggests the law of conservation of momentum.

+11667
In his *Vera Circuli and Hyperbolae Quadratura* the mathematician and astronomer James Gregory gives the first proof of the fundamental theorem of calculus and independently discovers Taylor series. Draft of the concept of transcendental number elaborated in relation to the problem of the quadrature of the circle.

+11665
The physicist Isaac Newton formulated the three laws of mechanics. He lays the foundations of differential calculus, these techniques allows him starting from the expression of a force inverse of square of the distance to find the general form of Kepler's laws.

+11664
The physicist Isaac Newton begins to work on differential and integral calculus.

+11661
The founder of statistical demography John Graunt published the first mortality table, the same year the physicist and chemist Robert Boyle determines the laws of compressibility of gas bearing his name and sometimes attached to that of the physicist Edme Mariotte who rediscovered a few years after the same laws.

+11659
The mathematician, astronomer and physicist Christian Huygens discovered the rigorous isochronism formula (when the end of the pendulum travels an arc of cycloid, the period of oscillation is constant regardless of the amplitude).

+11658
The mathematician, astronomer and physicist Christian Huygens experimentally discovers that balls placed anywhere inside an inverted cycloid reach the lowest point of the cycloid in the same time and thereby experimentally shows that the cycloid is the isochrone.

+11657
The lawyer and mathematician Pierre de Fermat states his "Fermat's principle" in optics as how the light propagates from one point to another on trajectories such that the duration of the propagation is locally minimal.

+11655
The mathematician, astronomer and physicist Christian Huygens was the first to use the concept of expected mean in probabilities. The mathematician John Wallis introduces the symbol ∞ in his *Mathesis Universalis*.

+11654
The mathematician, physicist, inventor, philosopher, moralist and theologian Blaise Pascal and the lawyer and mathematician Pierre de Fermat create the theory of probability. The world's oldest national scientific institution, the Royal Society, is founded in London. It establishes experimental evidence as the arbiter of truth.

+11650
James Ussher (archbishop) began counting all the "begats" in the Old Testament (...) He also studied ancient Egyptian and Hebrew texts, analysed how the ancient calendars were calculated and came up with a date for the Creation that Earth was about 5,650 years old.

+11644
The physicist and mathematician Evangelista Torricelli has the idea of substituting water by mercury in the so named Torricelli's experiment to highlight the "grosso-vido"; later will the works of the mathematician, physicist, inventor, philosopher, moralist and theologian Pascal Blaise follow (experience Puy de Dôme experiment).

+11638
The mathematician, geometer, astronomer and physicist Galileo Galilei publishes the mathematical relationship that defines the period of the simple pendulum.

+11637
The philosopher and mathematician René Descartes renames the unknowns x, y, z and the parameters a, b, c and extends the use of algebra to the lengths and plane, creating analytical

geometry with Pierre de Fermat. The same year, always René Descartes, quantitatively derives the angles at which primary and secondary rainbows are seen with respect to the angle of the Sun's elevation.

+11631
The mathematician Thomas Harriot introduces, in a posthumous publication, the symbols $>$ and $<$. The same year the theologian and mathematician William Oughtred provides for the first time the multiplication symbol $\times$ and the symbol $\pm$.

+11629
The lawyer and mathematician Pierre de Fermat develops a rudimentary differential calculus. The same year, in his *Invention nouvelle en algèbre* Albert Girard states, without proof, for the first time the fundamental theorem of algebra (a polynomial of degree n has n complex distinct or non-distinct roots) using complex numbers.

+11626
Tables of sine, tangent and secant of the engineer Albert Girard with the use of abbreviations sin, cos and tan.

+11624
Invention of the first thermometer (whose graduations are obviously not standardized...) by the physician Santorio Santorio.

+11621
The astronomer and physicist Willebrord Snell discovered that the angle of refraction of light is determined by the sine of the angle of the incident light with the normal of the dioptre.

+11620
The engineer Francis Thomas Bacon defends the experimental method (it seems he was the first one to document the Scientific Method) and leads numerous observations on heat. He suggested that heat is related to movement.

+11619
Johannes Kepler finished to publish the three laws of planetary motion.

+11614
The mathematician John Napier invented logarithms, which bring the operations of multiplication and division to simple additions or subtractions.

+11611
Johannes Kepler discovers total internal reflection, a small angle refraction law and thin lens optics.

+11610
In his *Sidereus Nuncius*, Galileo Galilei reports the first observations at the telescope: the discovery of the satellites of Jupiter, the confirmation that the milky way is constituted of stars, the discovery of the rings of Saturn. These observations will have an important effect because they contradict some of the ideas of the Universe models of the time and the Bible scriptures.

+11609

In his *Astronomia Nova*, the astronomer and mathematician Johannes Kepler explains Kepler's first two laws for the motion of the planets.

+11608
The optician Hans Lippershey invented the telescope to will be used and improved (with a random quality) the following year by the mathematician, geometer, astronomer and physicist Galileo Galilei to confirm the theories of Copernicus.

+11604
In a letter to Paolo Sarpi, Galileo states the law of the fall of bodies: the distance travelled is proportional to the square of the time of fall.

+11603
The mathematician and astronomer Thomas Harriot determines how to calculate qualitatively the surface of a spherical triangle.

+11596
The quantitative history of variable stars begins with David Fabricius 11596 observations of the appearance and disappearance of omicron Ceti (Mira). The eclipsing variable star Algol was first noted by Gernian Montanari in 11667, but its period of 2.867 days was not measured until the 11783 work of Nathaniel Pigott, John Goodricke and Johann Georg Palitzsh. However the idea of variability of Algol of Miras was known in ancient times by Babylonians between 8105 and 9461 (see Schaumberger, J. 11935 *Sternkunde und Sterndienst in Babel*, vol. 3, Verlag der Aschendorffschen Verlagsbuchhandlung) and Chinese (9750 to 9975).

+11591
The mathematician François Vieta opens a new period int algebra by making calculations with letters, using vowels for unknowns and consonants for parameters. Moreover, he gives the development of the Newton binomial theorem.

+11590
The astronomer Galileo Galilei demonstrated experimentally that all falling bodies have the same acceleration. The same year, the opticians Hans and Zacheraius Janssen created the first microscope by combining several lenses that define the beginnings of scientific medicine and biology.

+11586
The engineer and physicist Simon Stevin proved the method of the parallelogram of forces and discovered that the pressure of a liquid on the bottom of a container is independent of its shape, and also of the bottom surface and depends only on the height water in the container. He also gave the pressure measurement on any portion of the side of a container.

+11584
Giordano Bruno proposed that the stars were distant suns surrounded by their own planets, and he raised the possibility that these planets might foster life of their own, a cosmological position known as cosmic pluralism. He also insisted that the universe is infinite and could have no "center". Bruno also largely anticipated the arguments of Galilei on the relativity principle

+11576
The astronomer Tycho Brahe observed a new star in the constellation of Cassiopeia and built an

observatory on the island of Hveen.

+11572
The mathematician Rafaelle Bombelli gives a formulation of complex numbers and the rules of actual calculations. He introduces the terms più di meno (pdm) and meno di meno (mdm) to represent $+\mathrm{i}$ and $-\mathrm{i}$.

+11548
The mathematician and physicist Simon Stevin wrote the tenth powers identified with an exponent. He also gives the first writing of vectors.

+11545
The mathematician Ludovico Ferrari gives the solution of equations of degree 4 (know under "Cardan formula").

+11543
The work of the astronomer Nicolas Copernicus summarizing 26 years of research and observations is published and clearly highlights that the heliocentric system of Ptolemy is not valid (before Nicolas Copernicus, Nicole Oresme (around 11322 – 11382) had already presented the heliocentric system as credible, however, Nicolas Copernicus also brings the mathematical, that is to say geometric, demonstration of its validity, however models in which the Earth was not at the center of the World had been proposed since Antiquity such as that of Aristarchus of Samos around 9690 – 9770).

+11536
The mathematician Niccolò Fontana launched the new science of ballistics.

+11530
The mathematician and physicist Robert Recorde introduces the "$=$" sign and the mathematician Michael Stifel developed an early form of algebraic notation.

+11525
The mathematician Christoff Rudolff introduces the notation for square roots $\sqrt{}$.

+11515
Publication of Scipione del Ferro's works where he find a formula giving the general solution of the polynomial equations of degree 3. These solutions involve the implicit manipulation of imaginary numbers

+11510
The painter, engraver and mathematician Albrecht Dürer develops the basics of descriptive geometry and perspective.

+11500
The Italian mathematician Scipione del Ferro succeeds for the first time to solve a large algebraic type of cubic equations.

+11490
The painter, sculptor, architect, musician, mathematician, engineer, inventor, anatomist, geologist, cartographer, botanist, and writer... Leonardo da Vinci describes capillary action.

+11489
In his *Behende und hupsche Rechnung auf allen kauffmanschafft*, the mathematician Johannes Widmannwe the symbols "$+$" and "$-$" are introduced for the first time (before plus was denoted "P" and minus "M").

+11464
In *Triparty in the science of numbers* by the mathematician Nicolas Chuquet we can find the first use of negative powers and zero power and the statement of the property of exponents that we still use $x^{n+m} = x^n x^m$.

+11420
The mathematician and astronomer Jamshid al-Kashi computed and observed the solar eclipses of 11406, 11407 and 11408. He is also the first to use decimal notation in arithmetic and in arabic numerals.

+11400
The Mathematician and astronomer Jamshid al-Kashi developed an early form of Newton's Regula falsi method.

+11200-11400
Madhava and the Kerala School (India) discover several infinite series for numbers like π and specific values of trigonometric functions - these works ahead of those of the Europeans on the differential and integral calculus and the series of powers.

+11350
Tractatus of configurationibus qualitatum and motuum of Nicole Oresme is a draft of geometry using coordinates, and use axes for different sizes, which is an important step in the transition from qualitative science based mainly on Aristotle to quantitative science. Proof of the mean velocity theorem, which anticipates the results of Gallileo on the uniform rectilinear motion and the bodies in free fall by linking area under the curve of the velocity to the position in a graph.

+11303
Siyuan Yujian (translation: *Precious mirror of elements*) of Zhu Sjijie describes the elimination method for solving systems of equations containing up to four unknowns and up to degree 14 for some form of equations. We also find there the definition of the Pascal's triangle and the formulas of summations for some series.

+11300
Raymon Lulle developed a geometric (useless) machine to automate theist logic. This idea will influence Leibniz in his research of a universal language for reasoning, research that will lead him to take an interest in Chinese writing and binary arithmetic. Lulle's ideas anticipate modern ideas of formal deduction systems.

+11269
The scholar Pierre de Maricourt coined the expressions of the magnetic "north" and "south poles" and he was the first who wrote that opposite poles attract each other.

+11268
The philosopher, scientist and alchemist Roger Bacon publishes proposals to reform school, arguing that to study nature, the use of observations of the measures is the only basis of rigor-

ous testing and verification while affirming at the same time the need of mathematics for this purpose.

+11200
The mathematician Jordan Nemore introduces the notation for unknowns with symbols.

+11150
Creation of modern notation for fractions (horizontal bar) by Al-Hassãr. At the same period, we have the Latin translation of the 820 treaty by Muhammad Ibn Musa al-Khwarizmi on the Indian calculation which allows the decimal system and the use of zero to spread in Europe and also Gérard de Cremona publishes a translation in Latin of the Arabic version of the *Almagest* of Ptolemée, the name "sinus" comes from this translation...

+11121
The astronomer, physicist, biologist, chemist, mathematician and philosopher Abu al-Fath Khazini published a book in which he proposed that gravity and gravitational potential energy vary with distance from the center of the Earth. He also makes a distinction between force, mass and weight. He also invented several scientific instruments, including a steelyard and hydrostatic balance. He also introduces experimental scientific methods to static and dynamic, unifies them in the science of mechanics and combines hydrostatics with the dynamic to create hydrodynamics.

+11114
The mathematician Bhaskara provides a comprehensive summary of Hindu mathematics, as developed from the 5th to the 7th century AD. He also recognizes the negative square root, solves quadratic equations with several unknowns, equations of higher order such as Fermat and the general quadratic equations. He was also a pioneer in the principle of differential calculus nearly 500 years before Newton and Leibniz.

+11100
The philosopher and physicist Allah Abu'l-Barakat Hibat al-Baghdaadi is the first to deny Aristotle's idea that a constant force produces uniform motion what prepares the Newton's second law of motion. Like Newton, he described acceleration as the variation of speed.

+11037
The mathematician, physicist and philosopher Ibn al-Haytham is aware of the magnitude of the acceleration due to gravity. He discovers the law of inertia, known today as the first law of motion Newton.

+11030
The philosopher, writer, physician and scientist Abu Ali al-Husayn ibn Abd Allah Ibn Sina (known in Occident as Avicenna) note that if the perception of light is due to the emission of some sort of particles by a light source, the speed of light has to be finished. He also provided a sophisticated explanation for the phenomenon of rainbow. The mathematician, astronomer, physicist, scholar, encyclopedist, philosopher, astrologer, traveller, historian, pharmacologist Abu Ali al-Husayn ibn Abd Allah Ibn Sina, and later the astronomer Abu al'Fath Khazini, were the first to apply the scientific method in experimental mechanics, especially in the fields of statics and dynamics, to determine the specific weight, such as those based on the theory of balances and weighting.

+11021
The philosopher, mathematician and physicist Ibn al-Haytham is considered as the father of optics and one of the pioneer of the scientific method to describe quite correctly the light and vision, and to introduces the experimental scientific method, laying the foundations of experimental physics. He also discusses experimental psychology and describes various optical instruments such as the darkroom. He was able to estimate atmosphere width with an accuracy of 1 [km], he defined the inertia principia (first Newton's law), the linear momentum and he calculated $\sum_{k=1}^{n} k^4 = \frac{n(2n+1)(n+1)(3n^2+2n-1)}{30}$.

+11019
The mathematician, astronomer and physicist Muhammad Ibn Ahmad Abul-Rayhan Al-Biruni observed and described the solar eclipse of April 8, 11019, and the lunar eclipse of September 17, 11019, in detail; he gave the exact location of the stars during the lunar eclipse. He invented the astrolabe and the planisphere.

+11010
The mathematician Al-Sijistani Zuraqi invented a astrolabe designed for a single heliocentric planetary model in which the Earth is moving, rather than the sky.

+11000
The mathematician, physicist and astronomer Abu Sahl al-Qouhi discovers that the weight of bodies varies with their distance from the center of the Earth, and solves equations higher than the second degree. During the same decade, the mathematician and engineer Ma'ruf al-Karkhi wrote a book containing the first known proof by mathematical induction. He uses it to prove the binomial theorem, the Pascal's triangle, and the sum of the cubes integrals.

+10996
The mechanical oriented astrolabe, with 8 gears is invented by the mathematician, astronomer and physicist Muhammad Ibn Ahmad Abul-Rayhan Al-Birun who is also the author of works on the summation of series and combinatorics.

+10980
Abu al-Wafa' Buzjani do the first calculation of values of trigonometric functions and publish the sinus law adapted for triangles on sphere. He also proved by induction $\sum_{k=0}^{n} k^3 = \frac{n^2(n+1)^2}{4}$.

+10964
The mathematician, physicist and philosopher Abd al-Rahman al-Sufi explains the magnifying power of lenses and he was one of the first to use a scientific method of analysis that will greatly influence future scientists.

+10953
The engineer and mathematician Ma'ruf al-Karkhi defines different monomials and gives rules for products of any two of them. He also discovered the binomial theorem for integer exponents.

+10952
The mathematician Abu'l-Hasan al-Uqlidisi modifies the calculation methods for the numerical Indian system to make it possible for feathers and paper usage. Until then, do calculations with Indian numerals necessitated the use of a board.

+10900
The first reference to a viewing tube can be found in the work of the astronomer and mathematician Al-Battani, and the first accurate description of the observation tube was given by the mathematician, astronomer, physicist, scholar, encyclopedist, philosopher, astrologer, traveller, historian, pharmacologist Muhammad Ibn Ahmad Abul-Rayhan Al-Biruni, in a section of his work dedicated to verify the presence of the new crescent Moon at the horizon. Although these preliminary observations tubes do not have lenses, they allow an observer to focus on a part of the sky by eliminating light interference. These observation tubes were later adopted in Europe, where they influenced the development of the telescope.

+10880
The astronomer and mathematician Al-Battani discovered the motion of the apogee of the Sun, calculate the values of the precession of the equinoxes and the inclination of the Earth's axis. He is at the origin of the definition of the tangent and cotangent trigonometric functions.

+10820
The word "algebra" appears. The mathematician, geographer, astrologer and astronomer Muhammad ibn Musa Al'Khwarizmi is often regarded as the father of medieval algebra because he releases it from geometry. He is also the origin of the quadrant, the mural instrument, the sinus quadrant that was used to solve trigonometric problems and make astronomical observations.

+10800
Astronomers invent the universal sundial and universal time dial in Baghdad. Earliest writing we get about statistical inference by Abu Yusuf Ya'qub Ibn Al-Kindi on how to use statistics and frequency analysis to decipher encrypted message. His text seems arguably gave rise to the birth of both statistics and cryptanalysis.

+10780
The alchemist Jabir Ibn Hayyan introduces the experimental scientific method for chemistry and also laboratory equipment such as still and processes such as pure distillation, liquefaction, crystallisation and filtration. He also invented more than 20 types of laboratory equipment, which resulted in the discovery of several chemicals. He also developed recipes for coloured glass.

+10773
Arabic numerals (adapted from India) made their first apparition in Europe.

+10628
The mathematician Brahmagupta gives rules for solving linear and quadratic equations. He discovers that the quadratic equations have two roots: the negative one and the irrational and give the modern form of the solution we know today. He also give the rules to calculates with negative signs (arithmetic of negative numbers).

+10550
Hindu mathematicians give zero a numeral representation in a positional notation system. *Brahmagupta's Brahmasphutasiddhanta* is the first book that provides rules for arithmetic manipulations that apply to zero and to negative numbers. The *Brahmasphutasiddhanta* is the earliest known text to treat zero as a number in its own right, rather than as simply a place-holder digit

in representing another number as was done by the Babylonians.

+10499
The mathematician Âryabhat gets the full number of solutions of a system of linear equations by methods equivalent to modern methods, and describes the general solution of such equations. He also provides solutions of differential equations. He also asserts that the Moon and celestial objects other than the stars reflect sunlight (almost 1000 years later after the poems of Parmenide claining the same), he correctly explains the causes of lunar and solar eclipses, gives the length of the sidereal year to a few minutes, approximate π by $62832/20000$, calculates the Earth's diameter with more precision than Eratosthenes, described the calculation with the Indian numbering system, gives the oldest sinus table for 24 angles.

+10275
The mathematician Diophantus of Alexandria considered as the father of algebraic equations studied equations with rational variables (thus including quadratic equations) and Diophantine equations.

+10195
Suàn shù shu (translation: *book on numbers and calculation*) is one of the oldest known Chinese mathematical texts that contains among other things, calculations of sums of Geometric progression for interest.

+10130
Mathematical Composition (also know under the name of *Almagest's*) of Ptolemy in Alexandria presents a geometric model of the solar system that attempts to describe the motion of the planets. The model is inspired by a geometric idea of Apollonius of Tyana and uses circles whose center move in circular orbits. This model puts the Earth at the center of the solar system but gives a fairly good description of the observed movements of the different stars. It will be the dominant model until Copernicus. Some ideas prefigure the Fourier series that will be introduced in the 19th century.

+10125
The *Yale Music Papyrus* and the *Michigan Instrumental Papyrus* seems to contain the oldest known examples of musical notation.

+10121
Year corresponding to the oldest document mentioning the magnetic stone.

+10120
Star catalogue of Zhang Heng which contains $2,500$ stars. Zhang Heng correctly describes the cause of the eclipses and demonstrates that the Moon is spherical.

+10100
The engineer, mechanician and mathematician Hero of Alexandria rediscovered (after the Chinese) the concept of force. He also invented a system of gears to lift weights using steam power. He gives the first description of the sextant (but did not, however, invented it). His contemporary astronomer Claudius Ptolemy invented the sextant and described the astrolabe (perhaps invented by the astronomer, geographer and mathematician Hipparchus) and studied the refraction and reflection. During the same century the mathematician and philosopher Nicomachus of Gerase defines the even and odd numbers, prime and composite numbers, and perfect numbers. Also

during the same century the final version of the *Nine Chapters on Mathematical Art* (almost 1180 pages), written over ten years by several anonymous Chinese authors contains the first use of negative numbers, chapter 9 uses the Pythagorean theorem, chapter 8 uses matrices and the elimination of Gauss to solve systems of equations (at least 1700 years before Gauss!!!)

+10080
The scholar Wang Ch'ung made the first magnetic compass on a plate of brass.

+9923
Year corresponding to the dating of the Antikythera mechanism, considered as the first calculator and the first antique analogue gear (thirty!) machine to calculate complexes astronomical positions. The care and skill with which this machine was made, as well as the necessary mechanical and astronomy capacities question the historical knowledge of Greek science before its discovery. Indeed, any object of the same age and same complexity was known in the world and it takes nearly a millennium to see similar mechanisms appear! The physicist, mathematician and engineer Archimedes of Syracuse is the hypothetical creator.

+9900
The indian text *Anuyoga Dwara Sutra* contains several identities involving square roots and squares that seem to imply some knowledge of the laws of exponents or logarithms. An identity taken from this text in modern notation: $\sqrt{a}\sqrt{\sqrt{a}}=(\sqrt{\sqrt{a}})^3$.

+9866
The astronomer, geographer, and mathematician Hipparchus of Nicaea discovers the precession of the equinoxes.

+9850
The astronomer, geographer and mathematician Hipparchus is often referred as the founder of trigonometry and corresponding numerical tables. He calculates the first period of revolution of the Sun around the Earth (but the numerical results are in fact those of the rotation of the Earth around the Sun) and develops the theory of eccentrics and epicycles.

+9800
During the century, the Chinese had invented the tide gate, the rudder, the principle of the steam engine several hundred years before the Occident! During this century, the scientist and engineer Philo of Byzantium wrote treatises on levers, pneumatics, automation, traction and water clocks.

+9775
The astronomer and mathematician Apollonius of Perga published the first study on conics giving to the ellipse, the parabola and the hyperbola the names we know today. He is also credited for the hypothesis of eccentric orbits to explain the apparent motion of the planets and the speed variation of the Moon.

+9750
The mathematician, physicist and engineer Archimedes of Syracuse study simple machines such as the lever, the famous screw for pumping water ("Archimedes screw") and discovered the Archimedes's principle explaining buoyancy. In the same decade, the astronomer, geographer, philosopher and mathematician Eratosthenes of Cyrene calculated the diameter of the Earth using a gnomon and its shadow and demonstrated the inclination of the ecliptic, the distance

Earth-Moon and Earth-Sun and also created a method to determine prime numbers (Sieve of Eratosthenes).

+9740
The mathematician, physicist and engineer Archimedes of Syracuse computes to two decimal places using inscribed and circumscribed polygons and computes the area under a parabolic segment.

+9719
The astronomer and mathematician Aristarchus of Samos assumed that the Sun is the center of the solar system (heliocentrism model) and uses trigonometry to estimate the radius of the Moon and its distance from the Earth using the Earth's shadow during a lunar eclipse. He also estimated the Earth-Sun distance in number of times the Moon-Earth distance (...).

+9700
The mathematician and geometer Euclid published his *Elements*, where he reorganized the entire knowledge of the geometry including logical proofs, the construction of the 5 Platonic solids. In his *Optica* he noted that the light goes in a straight line and describes the law of reflection. We have also that the conical, elliptical, parabolic and hyperbolic concepts appears in the works of two mathematicians of the ancient Greece, namely Menaech-mus and Appolonius de Perge that also introduces the concept of tangent. The same period Bhagabati Sutra calculates the permutations and combinations of order 1, 2 and 3.

In Hindu cosmology, the Manusmriti (1.67–80) and Puranas describe time as cyclical, with a new universe (planets and life) created by Brahma every 8.64 billion years. The universe is created, maintained, and destroyed within a kalpa (day of Brahma) period lasting for 4.32 billion years, and is followed by a pralaya (night) period of partial dissolution equal in duration. In some Puranas (e.g. Bhagavata Purana), a larger cycle of time is described where matter (mahat-tattva or universal womb) is created from primal matter (prakriti) and root matter (pradhana) every 622.08 trillion years, from which Brahma is born. The elements of the universe are created, used by Brahma, and fully dissolved within a maha-kalpa (life of Brahma; 100 of his 360-day years) period lasting for 311.04 trillion years containing 36,000 kalpas (days) and pralayas (nights), and is followed by a maha-pralaya period of full dissolution equal in duration. The texts also speak of innumerable worlds or universes.

+9690
The scholarly Autolycus of Pitane defines uniform motion as an object that travels an equal distance in a equal amount of time.

+9630
The philosopher Aristotle develops the logic with a theory of naive proposals, quantities and inferential reasoning. He also wrote a treatise on meteorology named *Meteorologica*. It is crystal clear from this treatise that he already understood the cycle of water and knew that clouds were made of evaporated water and are therefore very heavy structures. He also knew (in a naive way) and tested experimentally how salt water and normal water were separated and mixed in the seas. Aristotle pioneered scientific method alongside his empirical biology and his work on logic, rejecting a purely deductive framework in favour of generalisations made from observations of nature. Aristotle mentioned also that salt water (of the sea) was heavier and more dense than fresh water, and salt water would seek a lower level.

+9622
The philosopher and astronomer Heraclides Ponticus assumes the rotation of the Earth itself to explain the apparent motion of stars in the night (but still in a geocentric context) and suggests that each planet is a body like the Earth.

+9600
The Stoic School develops the composed proposals and the logical connectors: "implies", "and", "or" and the inferences "Modus ponens" and "Modus tollens".

+9580
Philolaos publish his System of the World where the Earth turns around a central fire.

+9530
The philosopher Democritus of Abdera advance the idea that matter is composed of tiny and identical particles that he named "atoms". In reality it is rather an extension of the ideas of his teacher, the philosopher Leucippus of Miletus developed ten years before. Hippasus, a disciple of Pythagoras, would have given what is probably the first rigorous proof of the irrationality of $\sqrt{2}$. The proof uses a Reductio ad absurdum reasoning to show that the sides of a square are incommensurable with its diagonal. Least known written use of statistics by the historian Thucydides in his *History of the Peloponnesian War* (use of the Mode).

The oldest known record of the Camera Obscura is a description by Han Chinese philosopher Mozi (ca. 9530 to ca. 9609). Mozi correctly asserted that the camera obscura image is inverted because light travels in straight lines from its source. In the 11th century Arab physicist Ibn al-Haytham (Alhazen) wrote very influential books about optics, including experiments with light through a small opening in a darkened room.

+9533
Anaxagoras produced a correct explanation for eclipses and then described the Sun as a fiery mass larger than the Peloponnese, as well as attempting to explain rainbows and meteors. He was the first to explain that the Moon shines due to reflected light from the Sun.

+9500
The philosophers Leucippus and Democritus are the founders of atomism. Yajnavalkya, in his *Shatapatha Brahmana*, describes the motions of the Sun and the Moon, and advances a 95-year cycle to synchronize the motions of the Sun and the Moon.

Period of earliest documented mention of the concept of spherical Earth dates from around the 95th century, when it appears in the writings of Greek philosophers. In the 93th, Hellenistic astronomy established the roughly spherical shape of Earth as a physical fact and calculated the Earth's circumference. This knowledge was gradually adopted throughout the Old World during Late Antiquity and the Middle Ages.

+9460
The philosopher, mathematician Pythagoras studies propositional geometry and vibrating lyre strings.

+9415
Thales of Miletus predicts a solar eclipse.

+9400
The philosopher Empedocles of Acragas calls for decomposition of the world into four fundamentals elements: water, earth, air and fire. The same century, the mathematician Thales of Miletus highlights electrostatic by rubbing a piece of amber, predicted an eclipse and develops the geometry of the triangle.

Greek philosophers, as early as Anaximander, introduce the idea of multiple or even infinite universes. Democritus further detailed that these worlds varied in distance, size; the presence, number and size of their suns and moons; and that they are subject to destructive collisions. Also during this time period, the Greeks established that the Earth is spherical rather than flat.

+9250
Sulbar Sutras of Manava finds irrationality of $\sqrt{2}$ and $\sqrt{61}$ and agrees to use irrational numbers in his calculations. The first known university (using strictly the definition of that word which is derived from the Latin *universitas magistrorum et scholarium*, which roughly means "community of teachers and scholars") was created at Miletus (where Thales studied). Notice that the word "universiy" seems to exist since only the 113 century.

+9200
Assyrians use water-clocks and Chinese plot planetary movements for their calendar.

+9105
In approximately 8705 a new hieroglyphic word appeared. Bia-n-pet literally translates as “iron from the sky”. The new word was applied to all metallic iron from this time onwards. An obvious explanation for the sudden emergence of the new word would be a major event, such as a large impact or shower of meteorites, witnessed by the ancient Egyptian population. This would have left little uncertainty as to where exactly the mysterious iron came from, and would have created a sufficiently strong association between iron and the sky for the new word to become synonymous with all forms of iron. In 12008 a large impact crater formed by an iron meteorite was found in southern Egypt. Although its exact age is unknown, local archaeology suggests that it formed within the last 5000 years, making it a possible candidate event.

+8600
The Neolithic peoples of Scotland nowadays build stone models of Plato’s five solids (regular polyhedra). Hindous claim using concordism that the *Rigveda* hymn 1:50 (50th hymn in book 1 of Rigveda) leads to the calculation of the speed of light.

+8500
Indians develops theory of the 4 elements (fire, air, water, earth). Babylonians develop earliest known algorithms for factorization and finding square roots.

+8300
The mathematician Apastamba Dharmasutra solved general linear equations and used Diophantine systems of equations with up to five unknowns. The same century, egyptian mathematicians employ primitive fractions and develop earliest known algorithms for multiplying two numbers.

+8200
The origins of algebra can be traced to the ancient Babylonians, who developed a positional number system that greatly aided them in solving their rhetorical algebraic equations. The Babylonians were not interested in exact solutions, but rather approximations, and so they would

commonly use linear interpolation to approximate intermediate values. One of the most famous tablets is the Plimpton 322 tablet, created around 8100-8400, which gives a table of Pythagorean triples and represents some of the most advanced mathematics prior to Greek mathematics. Babylonian algebra was much more advanced than the Egyptian algebra of the time; whereas the Egyptians were mainly concerned with linear equations the Babylonians were more concerned with quadratic and cubic equations. The Babylonians had developed flexible algebraic operations with which they were able to add equals to equals and multiply both sides of an equation by like quantities so as to eliminate fractions and factors. They were familiar with many simple forms of factoring, three-term quadratic equations with positive roots, and many cubic equations, although it is not known if they were able to reduce the general cubic equation.

+8200
It was known during Vedic and later times (*Rigveda*, I, 19.7, I 23.17, I, 32.9,VIII, 6.19; VIII, 6.20; and VIII, 12.3) (Sarasvati, 12009) that water is not lost in the various processes of the hydrological cycle, namely, evaporation, condensation, rainfall, streamflow, etc., but gets converted from one form to another (transfer of water from the Earth to the atmosphere by the Sun and wind).

+8000
Babylonian priests do the first records of celestial observations. The Babylonians use a base-60 positional numeral system, and compute the first known approximate value of π at 3.125.

+7700
Chinese astronomers make the first observations of the sky.

+7500
The Mesopotamians imagined a position numbering system composed of symbols whose value is based on their rank within a number. Average estimate age of Stonehenge (England) as multifunctional site but which might have allowed prediction of eclipse, solstice, equinox and other celestial events.

+7400
Oldest known mathematical table where a multiplication table is engraved to calculate areas.

+7000
Chinese and Babylonians invented the abacus as first adding machine. Geometric concepts are developed for land surveying (hypotenuse calculus). It is also the period corresponding to the oldest know tool used by Incas to record numbers thanks to knots on a string and also the wall representation of wheels.

+6900
Egypt, earliest known decimal system allows indefinite counting by way of introducing new symbols. Mayan astronomers discover an 18.7-year cycle in the rising and setting of the Moon. From this they created the first almanacs – tables of the movements of the Sun, Moon and planets for the use in astrology. In 94th century in Greece, this knowledge was used to predict eclipses.

+6500
Dating of the oldest weather report found on a stone in Egypt. The unusual weather patterns described on the slab were the result of a massive volcano explosion at Thera, the present day

island of Santorini in the Mediterranean Sea.

+6300
A Babylonian tablet seems to contain first remarkable trigonometric angles and the evidence that Pythagoras' theorem was already known by Babylonians.

+6100
The Goseck circle (German: Sonnenobservatorium Goseck) may be one of the oldest Solar observatories in the world.

+5000
The decimal system is used in Ancient Egypt (it seems that the consensus is between 4000 and 7000).

+4800
Radiocarbon dating of the oldest founded wheels (Ljubljana Marshes wooden Wheel).

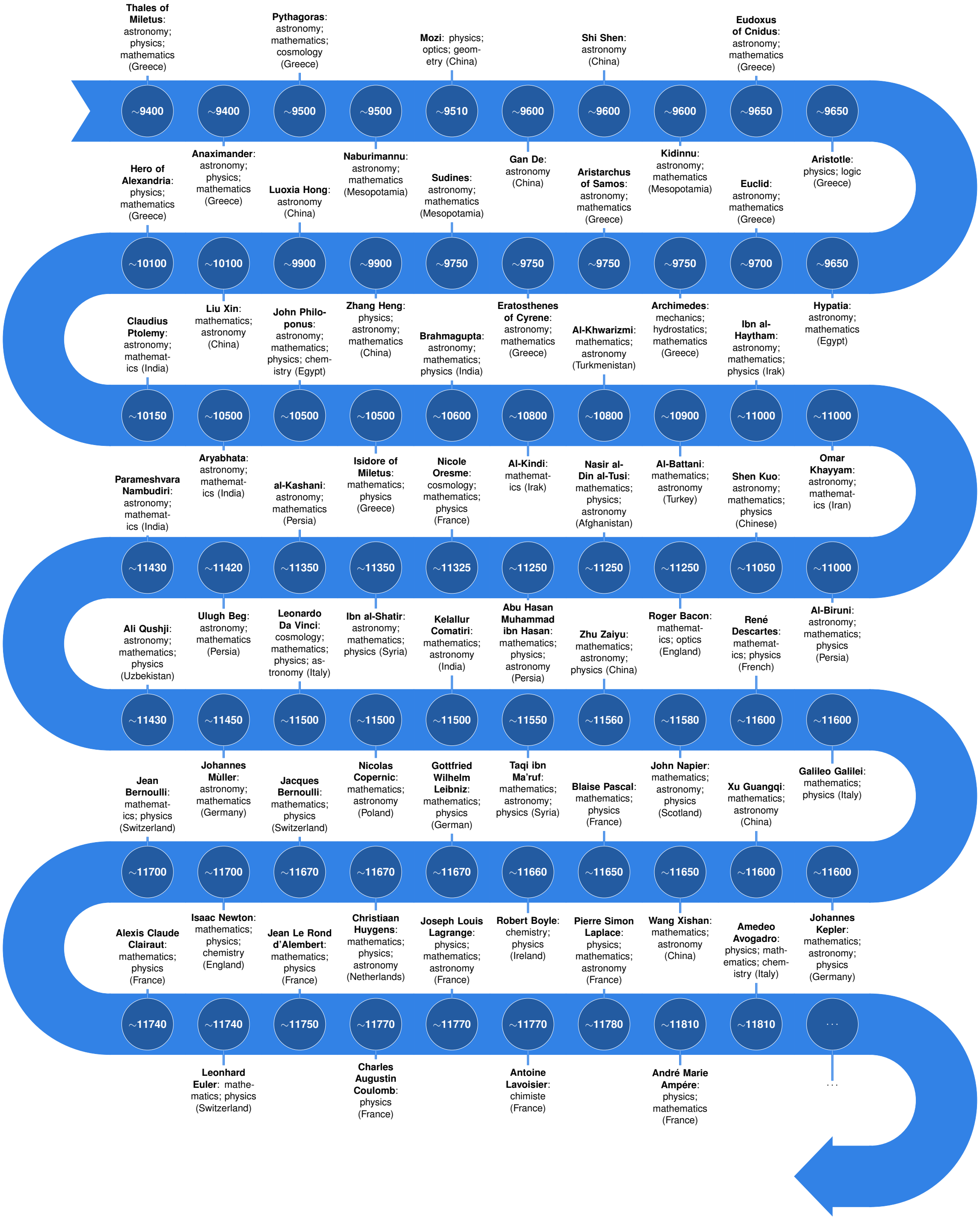
+2000
Warren Field is the location of a Mesolithic calendar. It includes 12 pits believed to correlate with phases of the Moon and used as a lunar calendar. It is considered to be the oldest lunar calendar yet found. Marks on bones or woods are slowly replaced by tokens of various shapes to count.

-1000
Supposed age of the oldest astronomical observatory in the world named Wurdi Youang (aboriginal stone arrangement located off the Little River – Ripley Road at Mount Rothwell, near Little River, Victoria in Australia).

-10000
The Ishango bone (Nile Valley) is a bone tool, dated to the Upper Paleolithic era (see [302]). It is a dark brown length of bone, the fibula of a baboon, with a sharp piece of quartz affixed to one end, perhaps for engraving. It was first thought to be a tally stick, as it has a series of what has been interpreted as tally marks carved in three columns running the length of the tool. It has also been suggested that the scratches might have been to create a better grip on the handle or for some other non-mathematical reason.

-25000 to -10000
Africa and France, earliest known prehistoric attempts to quantify time.

-190000
Claims for the earliest definitive evidence of control of fire by a member of Homo range from 0.2 to 1.7 million years ago. Evidence for the controlled use of fire by Homo erectus, beginning some 400,000 years ago, has wide scholarly support.



Thales of Miletus: astronomy; physics; mathematics (Greece)
Pythagoras: astronomy; mathematics; cosmology (Greece)
Mozi: physics; optics; geometry (China)
Shi Shen: astronomy (China)
Eudoxus of Cnidus: astronomy; mathematics (Greece)

~9400 ~9400 ~9500 ~9500 ~9510 ~9600 ~9600 ~9600 ~9650 ~9650

Hero of Alexandria: physics; mathematics (Greece)
Anaximander: astronomy; physics; mathematics (Greece)
Luoxia Hong: astronomy (China)
Naburimannu: astronomy; mathematics (Mesopotamia)
Sudines: astronomy; mathematics (Mesopotamia)
Gan De: astronomy (China)
Aristarchus of Samos: astronomy; mathematics (Greece)
Kidinnu: astronomy; mathematics (Mesopotamia)
Euclid: astronomy; mathematics (Greece)
Aristotle: physics; logic (Greece)

~10100 ~10100 ~9900 ~9900 ~9750 ~9750 ~9750 ~9750 ~9700 ~9650

Claudius Ptolemy: astronomy; mathematics (India)
Liu Xin: mathematics; astronomy (China)
John Philoponus: astronomy; mathematics; physics; chemistry (Egypt)
Zhang Heng: physics; astronomy; mathematics (China)
Brahmagupta: astronomy; mathematics; physics (India)
Eratosthenes of Cyrene: astronomy; mathematics (Greece)
Al-Khwarizmi: mathematics; astronomy (Turkmenistan)
Archimedes: mechanics; hydrostatics; mathematics (Greece)
Ibn al-Haytham: astronomy; mathematics; physics (Irak)
Hypatia: astronomy; mathematics (Egypt)

~10150 ~10500 ~10500 ~10500 ~10600 ~10800 ~10800 ~10900 ~11000 ~11000

Parameshvara Nambudiri: astronomy; mathematics (India)
Aryabhata: astronomy; mathematics (India)
al-Kashani: astronomy; mathematics (Persia)
Isidore of Miletus: mathematics; physics (Greece)
Nicole Oresme: cosmology; mathematics; physics (France)
Al-Kindi: mathematics (Irak)
Nasir al-Din al-Tusi: mathematics; physics; astronomy (Afghanistan)
Al-Battani: mathematics; astronomy (Turkey)
Shen Kuo: astronomy; mathematics; physics (Chinese)
Omar Khayyam: astronomy; mathematics (Iran)

~11430 ~11420 ~11350 ~11350 ~11325 ~11250 ~11250 ~11250 ~11050 ~11000

Ali Qushji: astronomy; mathematics; physics (Uzbekistan)
Ulugh Beg: astronomy; mathematics (Persia)
Leonardo Da Vinci: cosmology; mathematics; physics; astronomy (Italy)
Ibn al-Shatir: astronomy; mathematics; physics (Syria)
Kelallur Comatiri: mathematics; astronomy (India)
Abu Hasan Muhammad ibn Hasan: mathematics; physics; astronomy (Persia)
Zhu Zaiyu: mathematics; astronomy; physics (China)
Roger Bacon: mathematics; optics (England)
René Descartes: mathematics; physics (French)
Al-Biruni: astronomy; mathematics; physics (Persia)

~11430 ~11450 ~11500 ~11500 ~11500 ~11550 ~11560 ~11580 ~11600 ~11600

Jean Bernoulli: mathematics; physics (Switzerland)
Johannes Müller: astronomy; mathematics (Germany)
Jacques Bernoulli: mathematics; physics (Switzerland)
Nicolas Copernic: mathematics; astronomy (Poland)
Gottfried Wilhelm Leibniz: mathematics; physics (German)
Taqi ibn Ma'ruf: mathematics; astronomy; physics (Syria)
Blaise Pascal: mathematics; physics (France)
John Napier: mathematics; astronomy; physics (Scotland)
Xu Guangqi: mathematics; astronomy (China)
Galileo Galilei: mathematics; physics (Italy)

~11700 ~11700 ~11670 ~11670 ~11670 ~11660 ~11650 ~11650 ~11600 ~11600

Alexis Claude Clairaut: mathematics; physics (France)
Isaac Newton: mathematics; physics; chemistry (England)
Jean Le Rond d'Alembert: mathematics; physics (France)
Christiaan Huygens: mathematics; physics; astronomy (Netherlands)
Joseph Louis Lagrange: physics; mathematics; astronomy (France)
Robert Boyle: chemistry; physics (Ireland)
Pierre Simon Laplace: physics; mathematics; astronomy (France)
Wang Xishan: mathematics; astronomy (China)
Amedeo Avogadro: physics; mathematics; chemistry (Italy)
Johannes Kepler: mathematics; astronomy; physics (Germany)

~11740 ~11740 ~11750 ~11770 ~11770 ~11770 ~11780 ~11810 ~11810 ···

Leonhard Euler: mathematics; physics (Switzerland)
Charles Augustin Coulomb: physics (France)
Antoine Lavoisier: chimiste (France)
André Marie Ampére: physics; mathematics (France)
···

20

Humor

Contents

20.1 Situations 9212
20.2 Mathematics 9227
20.3 Physics 9249
20.4 Statistics 9273
20.5 Chemistry 9280
20.6 Engineering 9287
20.7 Computing 9297
20.8 Social Sciences 9308

If you have any scientific humorous stories do not hesitate to let us know. In all cases, we wish you a good time (some stories are in French as they lose their meaning in English).

This textbook is transmitted with 100% recycled electrons

20.1. Situations

An engineer, a physicist, and a mathematician are shown a pasture with a herd of sheep, and told to put them inside the smallest possible amount of fence.

- The engineer is first. He herds the sheep into a circle and then puts the fence around them, declaring:"A circle will use the least fence for a given area, so this is the best solution."
- The physicist is next. She creates a circular fence of infinite radius around the sheep, and then draws the fence tight around the herd, declaring, "This will give the smallest circular fence around the herd."
- The mathematician is last: After giving the problem a little thought, he puts a small fence around himself and then declares, "I define myself to be on the outside!"

Two men are sitting in the basket of a balloon. For hours, they have been drifting through a thick layer of clouds, and they have lost orientation completely. Suddenly, the clouds part, and the two men see the top of a mountain with a man standing on it.

- "Hey! Can you tell us where we are?!"

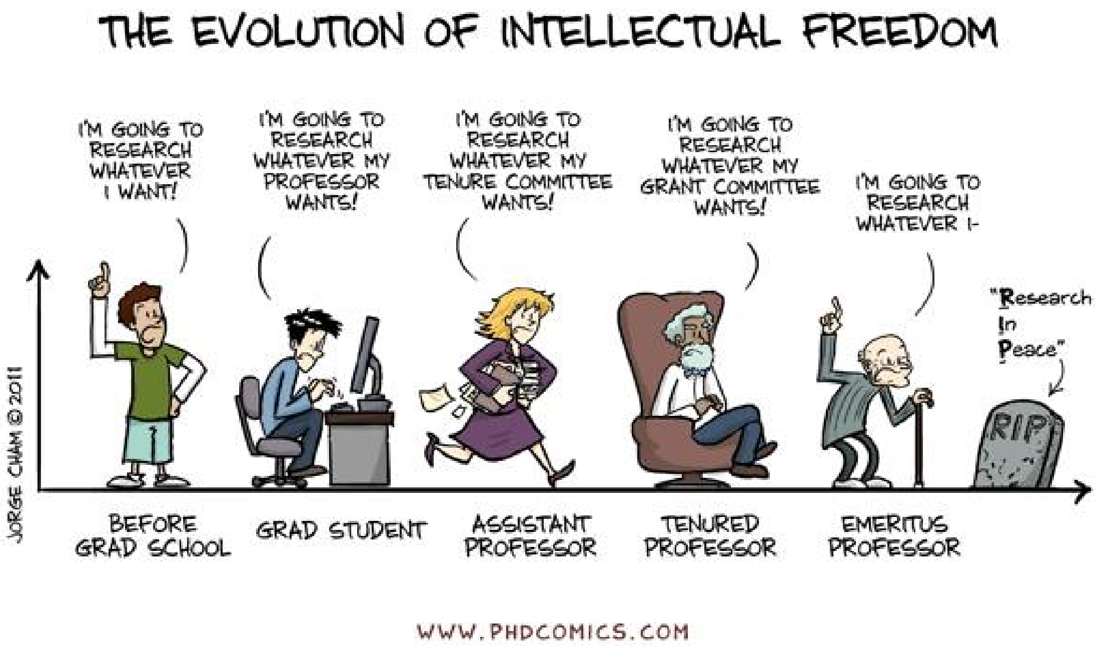
The man doesn't reply. The minutes pass as the balloon drifts past the mountain. When the balloon is about to be swallowed again by the clouds, the man on the mountain shouts:

- "You're in a balloon!"
- "That must have been a mathematician."

The man astonished asks:

- "Why?"
- "First, he thought long and thoroughly about what to say. Second, what he eventually said was irrefutably correct. And last but not least... it was of no use whatsoever..."

* * *


THE EVOLUTION OF INTELLECTUAL FREEDOM
I'M GOING TO RESEARCH WHATEVER I WANT!
I'M GOING TO RESEARCH WHATEVER MY PROFESSOR WANTS!
I'M GOING TO RESEARCH WHATEVER MY TENURE COMMITTEE WANTS!
I'M GOING TO RESEARCH WHATEVER MY GRANT COMMITTEE WANTS!
I'M GOING TO RESEARCH WHATEVER I-
"Research In Peace"
RIP
BEFORE GRAD SCHOOL
GRAD STUDENT
ASSISTANT PROFESSOR
TENURED PROFESSOR
EMERITUS PROFESSOR
JORGE CHAM © 2011
WWW.PHDCOMICS.COM

* * *

An engineer, a mathematician, and a physicist are staying for the night in a hotel. Fortunately for this joke, a small fire out in each room.

- The physicist awakes, sees the fire, makes some careful observations, and on the back of the hotel's wine list does some quick calculations. Grabbing the fire extinguisher, he puts out the fire with one, short, well placed burst, and the crawls back into bed and goes back to sleep.
- The engineer awakes, sees the fire, makes some careful observations, and on the back of the hotel's room service list does some quick calculations. Grabbing the fire extinguisher (and adding a factor safety of 5), he puts out the fire by hosing down the entire room server times over, and then crawls into his soggy bed and goes back to sleep.
- The mathematician awakes, sees the fire, makes some careful observations, and on a blackboard installed in the room, does some quick calculations. Jubilant, he exclaims "A solution exists!", and crawls into his dry bed and goes back to sleep.

* * *

A doctor, a lawyer and a mathematician were discussing the relative merits of having a wife or a mistress.

- The lawyer says: "For sure a mistress is better. If you have a wife and want a divorce, it causes all sorts of legal problems.
- The doctor says: "It's better to have a wife because the sense of security lowers your stress and is good for your health.
- The mathematician says: "You're both wrong. It's best to have both so that when the wife thinks you're with the mistress and the mistress thinks you're with your wife... you can do some mathematics."

A prominent businessman hires a mathematician, a physicist and a computer scientist in order to win every horses competition.

- The mathematician go on the task, it computes matrices without end, define axioms at any headland and after long weeks of lemmas, theorems and conjectures, he concludes that the problem is formally unhearable.
- Then the computer scientist happy to see that the mathematician failed, is approaching its Cray III and after writing volumes of algorithms in C++ and introduced all the parameters and initial conditions joyfully announces that it will take just a few hundred years to calculate the result of each competition...
- The physicist, has a smile, he informs his distinguished colleagues that he has the solution. He approaches a blackboard and while drawing a sphere begins by saying: "approximate the horse by a perfect sphere..."

During a job interview, an entrepreneur receives four engineers: one who followed the Military Polytechnic School of Paris, the second HEC, the third is computer engineer, and the last followed University. It explains the four candidates in the end, to run a business, you just need to count.

He therefore addresses the first of them, the polytechnician, and said, "go ahead, count ..."

- The polytechnician: "one... two... one... two..."

The man surprised then asked the engineer who followed HEC: "To you! Count ..."

- The engineer of HEC: "one KiloDollar... two KiloDollar, three KD..."

He then turns anxiously toward the computer scientist: "To you! Count ..."

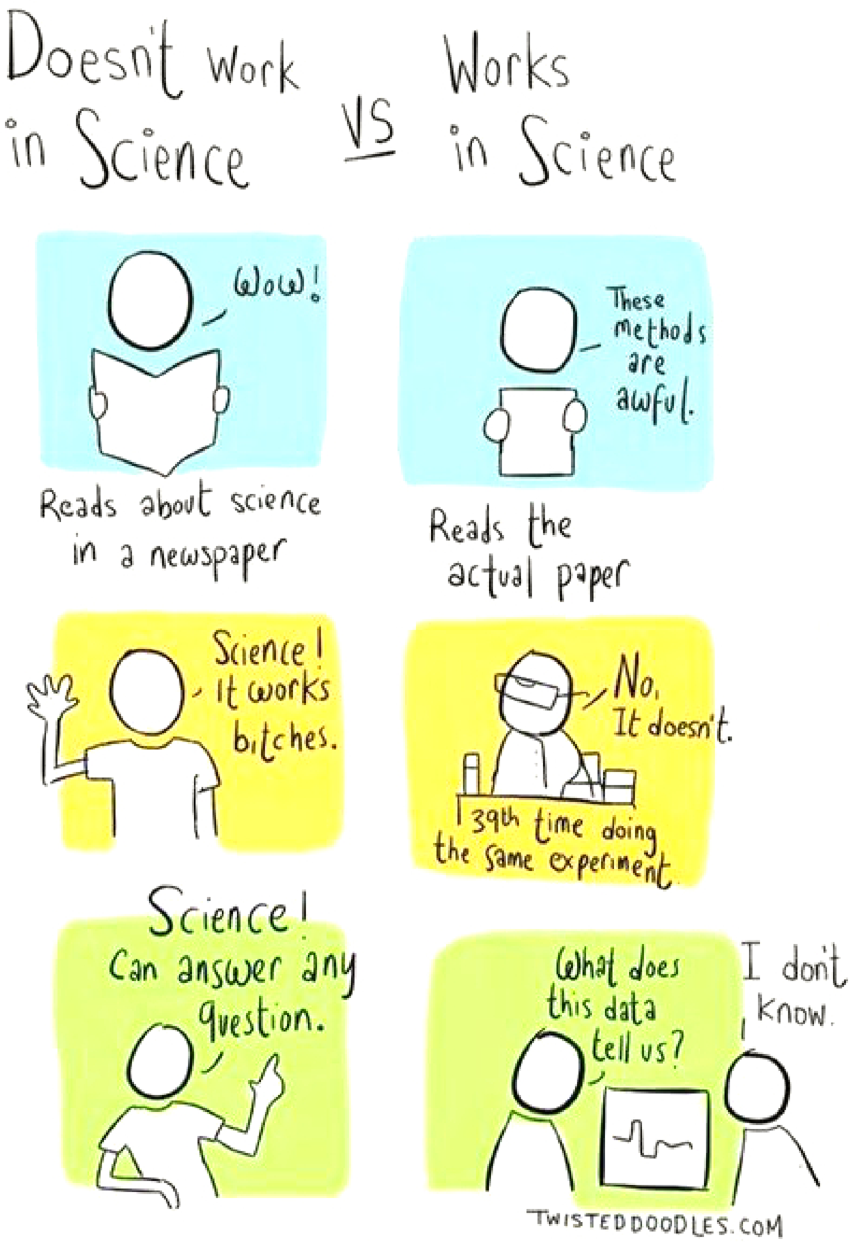
– The computer scientist: "0... 1... 0... 1..."

Desperate, the entrepreneur ask the engineer who followed University: "Go ahead, count ..."

– The young man begins: "1... 2... 3... 4... 5... 6... 7..."

The entrepreneur feeling reassured: "Continue, continue ..."

– "8... 9... 10... valet... lady.. king... " !!


Doesn't Work in Science
VS
Works in Science
Wow!
Reads about science in a newspaper
These methods are awful.
Reads the actual paper
Science! It works bitches.
No, It doesn't.
39th time doing the same experiment
Science! Can answer any question.
What does this data tell us?
I don't know.
TWISTEDDOODLES.COM

Several people were asked to solve the following problem: "Prove that all odd integers are prime."

- Mathematician: 3 is a prime, 5 is a prime, 7 is a prime, 9 is not a prime - counter-example - claim is false.
- Physicist: 3 is a prime, 5 is a prime, 7 is a prime, 9 is an experimental error, 11 is a prime ...
- Engineer: 3 is a prime, 5 is a prime, 7 is a prime, 9 is a prime, 11 is a prime ...
- Computer Scientist: 3's a prime, 5's a prime, 7's a prime ... segmentation fault
- Lawyers: one is prime, three is prime, five is prime, seven is prime, although there appears to be facing evidence that nine is not prime, there exists substantial precedent to indicate that nine should be considered prime. The following brief presents the case for nine's primeness...
- Liberals: The fact that nine is not prime indicates a deprived cultural environment which can only be remedied by a federally funded cultural enrichment program.
- Computer programmers: one is prime, three is prime, five is prime, five is prime, five is prime, five is prime five is prime, five is prime, five is prime...
- Professor: 3 is prime, 5 is prime, 7 is prime, and the rest are left as an exercise for the student.
- Linguist: 3 is an odd prime, 5 is an odd prime, 7 is an odd prime, 9 is a very odd prime,...
- Computer Scientist: 10 prime, 11 prime, 101 prime...
- Chemist: 1 prime, 3 prime, 5 prime... hey, let's publish!
- New Yorker: 3 is prime, 5 is prime, 7 is prime, 9 is... NONE OF YOUR DAMN BUSINESS!
- Programmer: 3 is prime, 5 is prime, 7 is prime, 9 will be fixed in the next release,...
- Salesperson: 3 is a prime, 5 is a prime, 7 is a prime, 9 – let me make you a deal...
- Advertiser: 3 is a prime, 5 is a prime, 7 is a prime, 11 is a prime,...
- Accountant: 3 is prime, 5 is prime, 7 is prime, 9 is prime, deducting 10% tax and 5% other obligations.
- Statistician: Let's try several randomly chosen numbers: 17 is a prime, 23 is a prime, 11 is a prime... Looks good to me.
- Psychologist: 3 is a prime, 5 is a prime, 7 is a prime, 9 is a prime but tries to suppress it...
- HR: What is a prime number? What is an odd integer?

A mathematician, an engineer and a physicist are given a red rubber ball to determine its volume.

- The mathematician: Measure the diameter and evaluates the triple integral.
- The physicist: Fill a tub of water, places the ball in water and measure the total displacement volume.
- The engineer: Search the model and serial number in its "red rubber balls" table.

The following concerns a question in a physics degree exam at the University of Copenhagen: "Describe how to determine the height of a skyscraper with a barometer."

One student replied:

"You tie a long piece of string to the neck of the barometer, then lower the barometer from the roof of the skyscraper to the ground. The length of the string plus the length of the barometer will equal the height of the building."

This highly original answer so incensed the examiner that the student was failed immediately. The student appealed on the grounds that his answer was indisputably correct, and the university appointed an independent arbiter to decide the case.

The arbiter judged that the answer was indeed correct, but did not display any noticeable knowledge of physics. To resolve the problem it was decided to call the student in and allow him six minutes in which to provide a verbal answer that showed at least a minimal familiarity with the basic principles of physics.

For five minutes the student sat in silence, forehead creased in thought. The arbiter reminded him that time was running out, to which the student replied that he had several extremely relevant answers, but couldn't make up his mind which to use. On being advised to hurry up the student replied as follows:

"Firstly, you could take the barometer up to the roof of the skyscraper, drop it over the edge, and measure the time it takes to reach the ground. The height of the building can then be worked out from the formula $H = 0.5g \cdot t^2$. But bad luck on the barometer."

"Or if the Sun is shining you could measure the height of the barometer, then set it on end and measure the length of its shadow. Then you measure the length of the skyscraper's shadow, and thereafter it is a simple matter of proportional arithmetic to work out the height of the skyscraper."

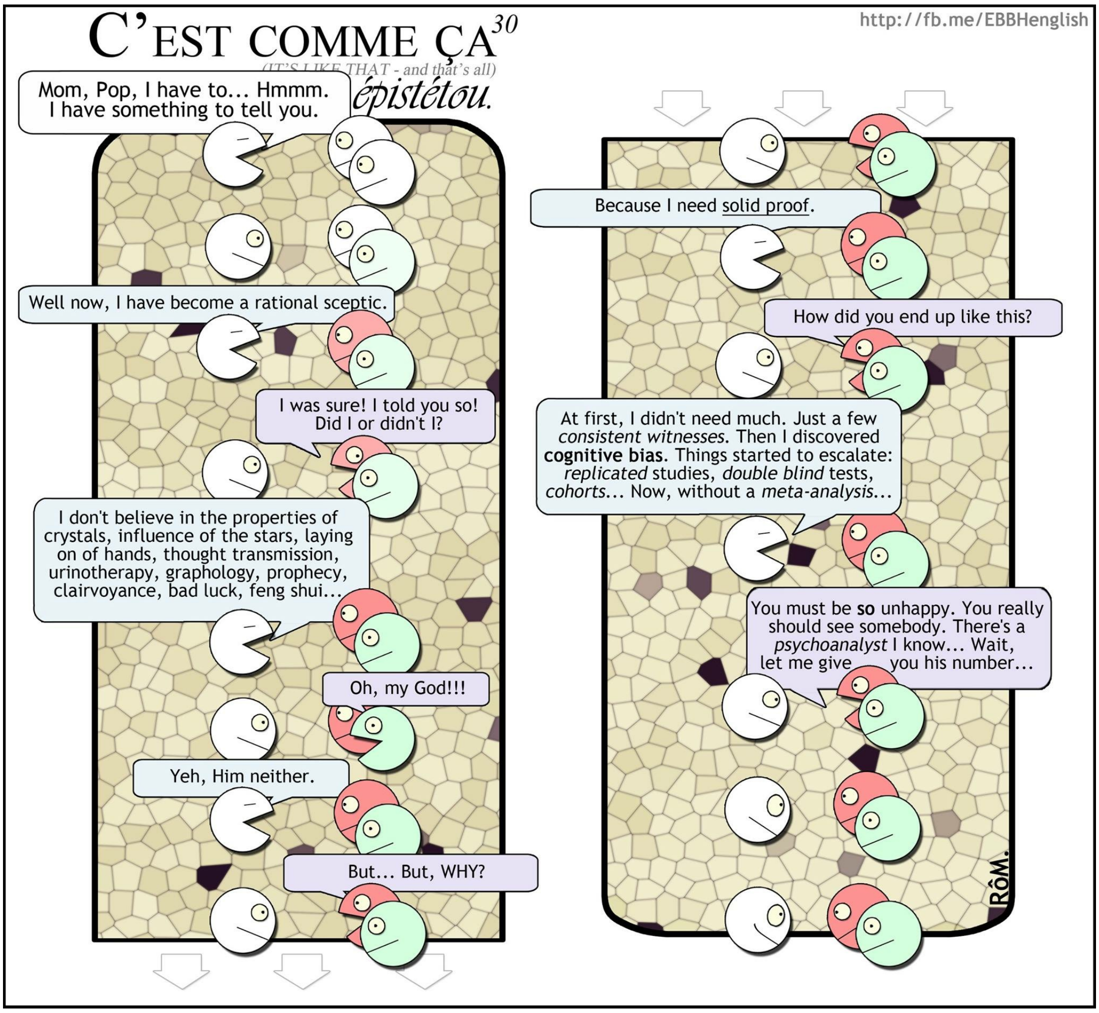
"But if you wanted to be highly scientific about it, you could tie a short piece of string to the barometer and swing it like a pendulum, first at ground level and then on the roof of the skyscraper. The height is worked out by the difference in the gravitational restoring force $T = 2\pi\sqrt{l/g}$."

"Or if the skyscraper has an outside emergency staircase, it would be easier to walk up it and mark off the height of the skyscraper in barometer lengths, then add them up."

"If you merely wanted to be boring and orthodox about it, of course, you could use the barometer to measure the air pressure on the roof of the skyscraper and on the ground, and convert the difference in millibars into feet to give the height of the building."

"But since we are constantly being exhorted to exercise independence of mind and apply scientific methods, undoubtedly the best way would be to knock on the janitor's door and say to him 'If you would like a nice new barometer, I will give you this one if you tell me the height of this skyscraper'."

The student was Niels Bohr (Nobel Prize 11923 according to holocene calendar), the only Dane to win the Nobel Prize for physics.


C'EST COMME ÇA[30]
(IT'S LIKE THAT - and that's all)
épistétou.
http://fb.me/EBBHenglish
Mom, Pop, I have to... Hmmm. I have something to tell you.
Well now, I have become a rational sceptic.
I was sure! I told you so! Did I or didn't I?
I don't believe in the properties of crystals, influence of the stars, laying on of hands, thought transmission, urinotherapy, graphology, prophecy, clairvoyance, bad luck, feng shui...
Oh, my God!!!
Yeh, Him neither.
But... But, WHY?
Because I need solid proof.
How did you end up like this?
At first, I didn't need much. Just a few consistent witnesses. Then I discovered cognitive bias. Things started to escalate: replicated studies, double blind tests, cohorts... Now, without a meta-analysis...
You must be so unhappy. You really should see somebody. There's a psychoanalyst I know... Wait, let me give you his number...
RôM.

The story goes that Bertrand Russell, in a lecture logic, mentioned that in the sense of material implication, a false proposition implies any proposition. A student raised his hand and said:

- "In that case, given $2+2=5$, you can prove that you are the Pope"
- "Yes", answered Russell
- "And you could prove it now?!", asked the student sceptical
- "For sure!", said Russel who proposed immediately the following proof:
    1. Suppose that $2+2=5$
    2. Subtract 2 from the both side of the equality, we have $2=3$
    3. By symmetry, $3=2$
    4. Subtract 1 each side, we obtain, $2=1$
- Now the set containing just me and the Pope has 2 members. But $2=1$, so it has only 1 member; therefore, I am the Pope...

* * *

What is "pi"?

- Mathematician: "Pi is the ratio of the circumference of a circle to its diameter."
- Engineer: "Pi is about $22/7$."
- Physicist: "Pi is 3.14159 plus or minus 0.000005"
- Computer Programmer: "Pi is 3.141592653589 in double precision."
- Nutritionist: "You're one track math-minded fellows, Pie is a healthy and delicious dessert!"

* * *

An astronomer, a physicist, a mathematician and a computer scientist (it is said) were holidaying in Scotland. Glancing from a train window, they observed a black sheep in the middle of a field.

- "How interesting", observed the astronomer, "all scottish sheep are black!"
- To which the physicist responded: "No, no! Some Scottish sheep are black!"
- The mathematician gazed heavenward in supplication, and then intoned: "In Scotland there exists at least one field, containing at least one sheep, at least one side of which is black."

- The computer scientist: "Oh, no! A special case!"

A mathematician, a biologist and a physicist are sitting in a street coffee watching people going in and coming out of the house on the other side of the street.

First they see two people going into the house. Time passes. After a while they notice three persons coming out of the house.

- The physicist: "The measurement wasn't accurate."
- The biologists: "They have reproduced".
- The mathematician: "If now exactly one person enters the house then it will be empty again."

A mathematician and an engineer attend a lecture by a physicist. The topic concerns Kaluza-Klein theories involving physical processes that occur in spaces with dimensions of 9, 12 and even higher. The mathematician is sitting, clearly enjoying the lecture, while the engineer is frowning and looking generally confused and puzzled. By the end the engineer has a terrible headache. At the end, the mathematician comments about the wonderful lecture.

- The engineer says: "How do you understand this stuff?"
- Mathematician: "I just visualize the process."
- Engineer: "How can you visualize something that occurs in 9-dimensional space?"
- Mathematician: "Easy, first visualize it in n-dimensional space, then let n go to 9."

Two mathematicians are in a bar. The first one says to the second that the average person knows very little about basic mathematics. The second one disagrees, and claims that most people can cope with a reasonable amount of math.

The first mathematician goes off to the wash-room, and in his absence the second calls over the waitress. He tells her that in a few minutes, after his friend has returned, he will call her over and ask her a question. All she has to do is answer "one third x cubed."

She repeats "one thir – dex cue"?

He repeats "one third x cubed".

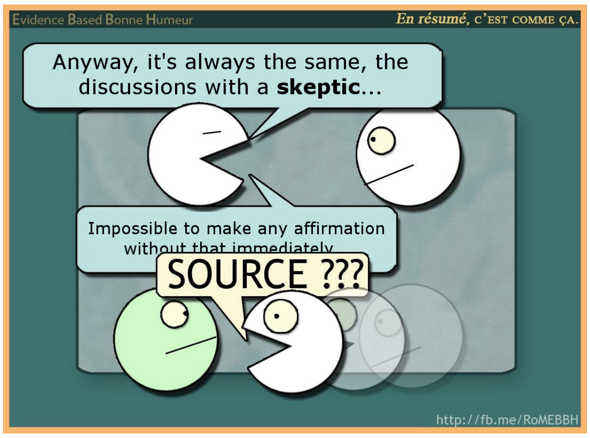
She asks, "one thir dex cuebd?"

"Yes, that's right," he says.

So she agrees, and goes off mumbling to herself, "one thir dex cuebd...".

The first guy returns and the second proposes a bet to prove his point, that most people do know something about basic math. He says he will ask the blonde waitress an integral, and the first laughingly agrees. The second man calls over the waitress and asks "what is the integral of x squared?".

The waitress says "one third x cubed" and while walking away, turns back and says over her shoulder "plus a constant!"


Evidence Based Bonne Humeur
En résumé, c'est comme ça.
Anyway, it's always the same, the discussions with a **skeptic**...
Impossible to make any affirmation without that immediately
SOURCE ???
http://fb.me/RoMEBBH

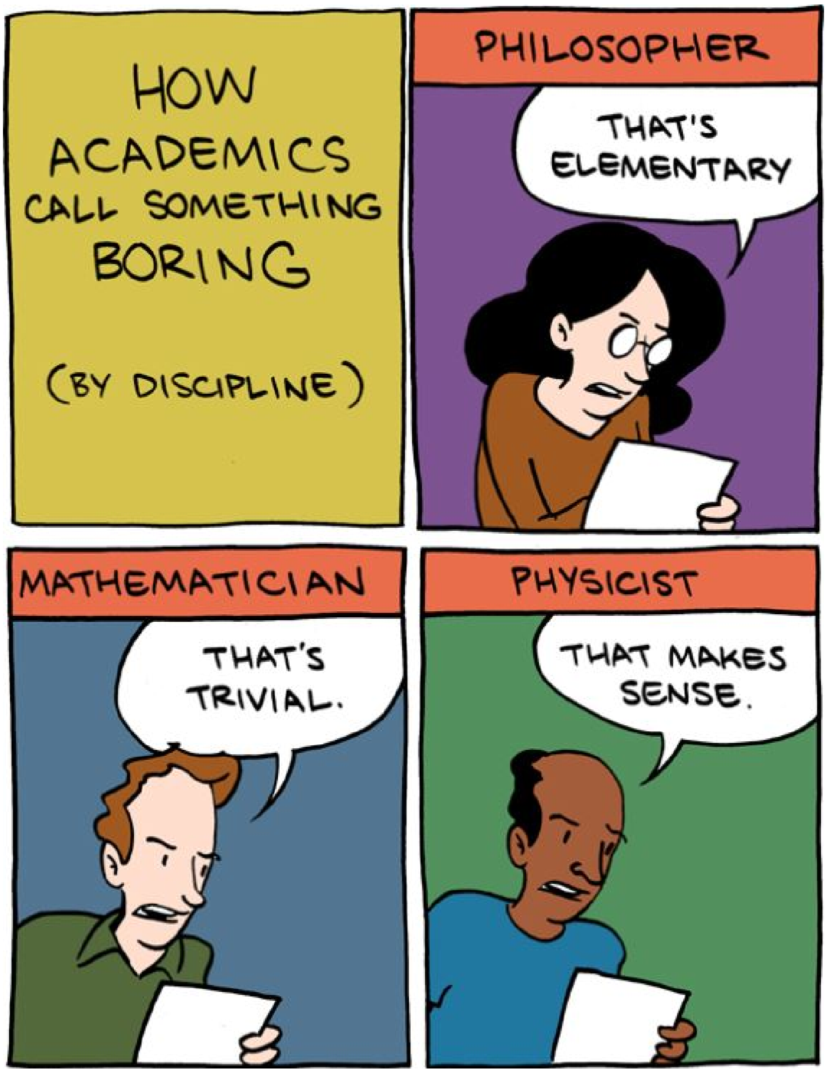
Teachers words:

- There are still pieces of the argument
- The '-' sign in front of the potential disturbs you? And if I put a '+' do you feel better? Yes! Then put a '+' ....
- This is a phase curve? No, this is a dog leg!
- Let's kill him and we will believe in suicide
- You can say it's predictable because it is unpredictable!
- I count on you to understand a little more thoroughly what will come next
- We will now study this a little later

- I will present you now some results dependent on Maxwell's equations that you have not seen yet ... anyway, at the point where we are ...
- If you do not understand, it's normal ... the contrary would be surprising indeed
- General relativity is useless! It is not with how we put satellites into orbit!
- Fifty cents question
- Occasionally, we have to be absurd
- When you have nothing to do, you take the Gauss theorem
- By which virtue of the Holy Spirit would it become neutral?!
- You will understand once you grow up...
- Take the example of a bank: there is a Mr. who makes a deposit and a Mrs. who makes withdrawal ... Finally, as usual what!
- ... more the parallelism will be parallel...
- Remove me these residues from earlier calculations
- Delete the left side. I said LEFT! Where is your right? This is ... Well the left is the other side!
- Ok, who will be the next victim?
- The '+' is recognized, and the electrons do not deceive themselves, for its beautiful red color!
- The goal is to verify that the test does not say shit
- If you can not do that, I assure you: it's all over for the exam
- If you can not do that, do it!
- You are very good at finding wrong things
- I like this kind of demonstration. You, it makes you have nightmares
- I won't will resolve this, as this may offend your sensibility!
- You must be careful with all what the teachers love!
- You add potatoes and pigs, it is dimensionless ...
- You can find the derivative of the Dirac pulse, it is not that we will jump to the head
- It's silly but Riemann, that's how
- Look at the equation that I just deleted
- Legally speaking, the current is in this sense

- There is no question that I waste my time to solve you this low fly joke!
- If the girl has understood then you must have all understood
- Oh yeah, you're laughing at me. In fact, you are right to make fun of us, because we do not hesitate to make fun of you
- Attention very important: I propose to dream at it on night
- I have the feeling of playing the Stradivarius before cows
- The hand of God spreads the field perpendicular to the surface
- The miracle of the disappearance of the harmonic will not happen today
- If you wake up at night, remember that sampling in the time domain is periodized in the frequency domain ... then you go back to sleep
- The left curves are not right
- You see, sometimes my results are correct
- We denote "Q" the output, it's obvious ...
- We'll have a trick to successfully recover this variable
- I am a 68000; Mr. Director comes in the corridor, knocking on the door: there is no better interruption!
- Warning: one, two, three, take your brains!
- In the steppes of automatic, we reach the arid part of mathematics
- What interest? Well no. This is a figure of speech Teaching
- If you have some memories of the signal flow graph that once we plotted ...
- The electrons in a hot metal, it's the New York main street at 5 in the evening, so it does not lead
- The circles, is what has the least corners!
- If you put your finger in a socket, it's not a complex number that comes
- See the proof in your daily readings ... So I waiver
- And then comes Carnot : he brings his mind that does not mean much
- Sometimes people who have a bit of culture that know that... it missed today!
- We will not name it "performance" because it can reach a value higher than 1, that might disturb some weak minds
- Me, if I had been sent to the blackboard, I would not have written this

- You will find the answer on www.archimedes.com
- I use a method that dates back several centuries ... as I do!
- Ask me questions ... I would like to have questions! ... other issues? I'll have to take the list and say: "But you have a question"!
- The ordering implemented depends on the instantiation of the garbage collector if it is synchronized to the preemptive multi-threading on the OS.
- You are only boxes receiving inputs and spitting outputs


HOW ACADEMICS CALL SOMETHING BORING
(BY DISCIPLINE)
PHILOSOPHER
THAT'S ELEMENTARY
MATHEMATICIAN
THAT'S TRIVIAL.
PHYSICIST
THAT MAKES SENSE.

Where is there mainly only religious holidays and not science holidays?

Answer: Because science always works!

Sherlock Holmes and Dr. John Watson went on a camping trip. After sharing a good meal and a bottle of Petri wine, they retire to their tent for the night.

At about 3 AM, Holmes nudges Watson and asks, "Watson, look up into the sky and tell me what you see?"

Watson said, "I see millions of stars."

Holmes asks, "And, what does that tell you?"

Watson replies, "Astronomically, it tells me there are millions of galaxies and potentially billions of planets. Astrologically, it tells me that Saturn is in Leo. Theologically, it tells me that God is great and we are small and insignificant. Horologically, it tells me that it's about 3 AM. Meteorologically, it tells me that we will have a beautiful day tomorrow. What does it tell you, Holmes?"

Holmes retorts, "Someone stole our tent."

An engineer, a mathematician, and a physicist are staying for the night in a hotel. Fortunately for this joke, a small fire breaks out in each room.

The physicist awakes, sees the fire, makes some careful observations, and on the back of the hotel's wine list does some quick calculations. Grabbing the fire extinguisher, he puts out the fire with one, short, well placed burst, and then crawls back into bed and goes back to sleep.

The engineer awakes, sees the fire, makes some careful observations, and on the back of the hotel's room service list (pizza menu) does some quick calculations. Grabbing the fire extinguisher (and adding a factor of safety of 5), he puts out the fire by hosing down the entire room several times over, and then crawls into his soggy bed and goes back to sleep.

The mathematician awakes, sees the fire, makes some careful observations, and on a blackboard installed in the room, does some quick calculations. Jubilant, he exclaims "A solution exists!", and crawls into his dry bed and goes back to sleep.

How do you know that the driver driving toward you is a physicist?

He has a red sticker on his bumper, saying: "If this sticker is blue, you are driving too fast."

A Princeton plasma physicist is at the beach when he discovers a ancient looking oil lantern sticking out of the sand. He rubs the sand off with a towel and a genie pops out. The genie offers to grant him one wish. The physicist retrieves a map of the world from his car and circles the Middle East and tells the genie, "I wish you to bring peace in this region".

After 10 long minutes of deliberation, the genie replies, "Gee, there are lots of problems there with Lebanon, Iraq, Israel, and all those other places. This is awfully embarrassing. I've never had to do this before, but I'm just going to have to ask you for another wish. This one is just too much for me".

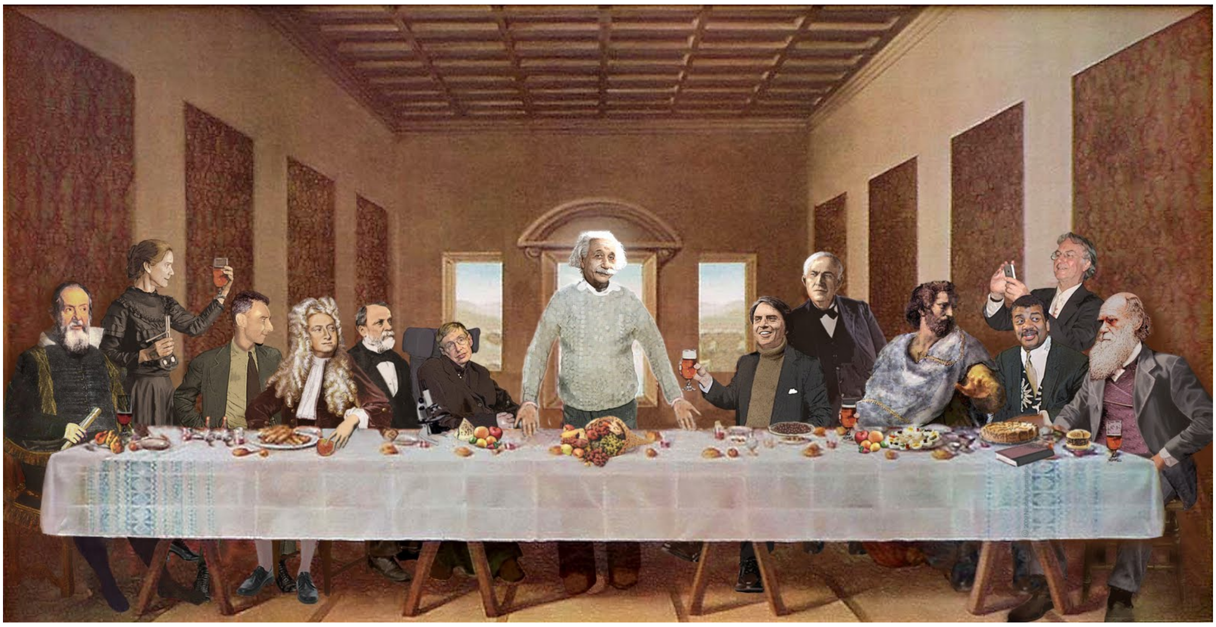
Taken aback, the physicist thinks a bit and asks, "I wish that the Princeton tokamak would achieve scientific fusion energy break-even."

After another deliberation the genie asks, "Could I see that map again?"

In the beginning, there were two species of apple trees: those whose apples fell, and those whose apples were rising. The apples that fell could reach the ground, germinate and so generating a new tree whose apples fell. But the rising apples never reached the ground, and species of apple trees with apples rising quickly disappeared because they could not reproduce. As they were not adapted, nature has thus eliminated. That's natural selection. If apples fall, it is thanks to natural selection!

But there are clever people who will object that the stones fall too. But they are not living beings, so they are not subject to natural selection, it is obvious. So the above explanation does not explain what happens to stones. In fact, for the stones, it is even simpler:

In the beginning, there were indeed two kinds of stones: those who fell, and who were rising. But those who raised up gone very far. That's why all the stones that remain on Earth do the same thing: they fall.



20.2. Mathematics

5 out of 4 people don't understand fractions...

Love is like π, natural, irrational, transcendent and very real.

A math teacher explains to a blonde the limits. He made it with the following exercise:

$$\lim_{x \to 8} \frac{1}{x-8} = +\infty$$

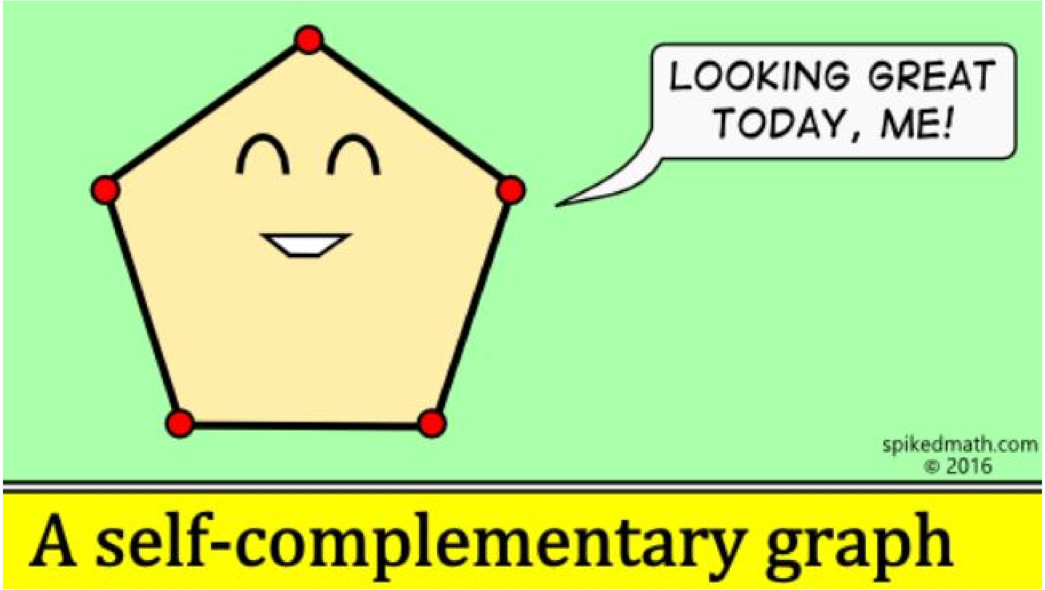
At the end of the exercise, he asked the blonde if she understood. "Oh yes sir I understood everything". Believing the answer only half, the teacher asked the following exercise:

Calculate:

$$\lim_{x \to 5} \frac{1}{x-5}$$

The blonde wrote:

$$\lim_{x \to 8} \frac{1}{x-8} = +\infty \quad \text{then} \quad \lim_{x \to 5} \frac{1}{x-5} = \rotatebox{90}{5}$$

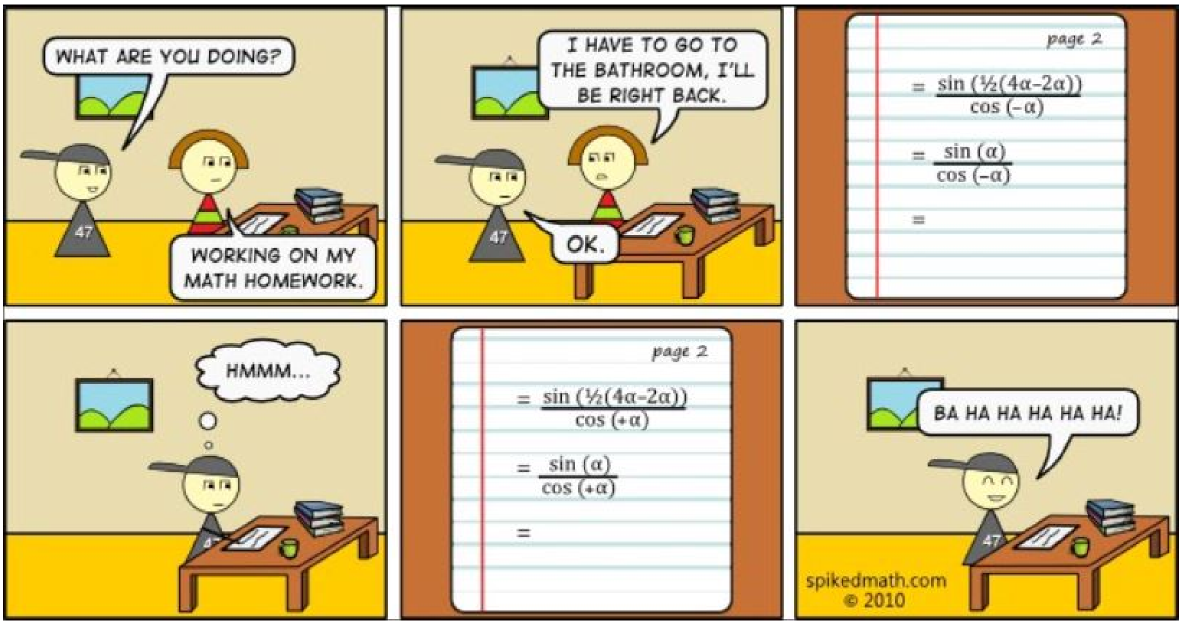

LOOKING GREAT TODAY, ME!

spikedmath.com © 2016

A self-complementary graph

Evolution of the teaching of Mathematics (...)

- Education 11960 (holocene calendar): A farmer sells a bag of potatoes for 100$. Its production costs amounted to 4/5 of the selling price. What is his profit?
- Traditional Education 11970 (holocene calendar): A farmer sells a bag of potatoes for 100$. Its production costs amounted to 4/5 of the selling price, that is to say 80$. What is his profit?
- Modern Education 11970 (holocene calendar): A farmer exchanges a set P of potatoes against a set M of money. The cardinality of the set M is 100, and each PFM element is 1$. Draw 100 large dots elements of the set M. The set F of the production costs are 20 big points less than the set M. Represent the set F as a subset of the set M and give the answer to the question: What is the cardinality of the set B benefits? (draw it in red)
- Renovated Education in 11980 (holocene calendar): A farmer sells a bag of potatoes for 100$. Production costs amount to 80$ and the benefit is 20$. Homework: underline the words "Potatoes" and discussed this with your neighbour.
- Start-up Education 11999 (holocene calendar): A wired producer of agricultural space consults a data bank which display the day-rate of the potato. It load its reliable software computation and determines the cash flow on bit-map screen (under config WMil with 40GB HDD and floppy). Draw with your mouse the integrated 3D contour of the bag of potatoes. Then log yourself to the network by www.blue-potatoe.com and follow the instructions of the menu.
- Education 12010 (holocene calendar): What is a farmer?

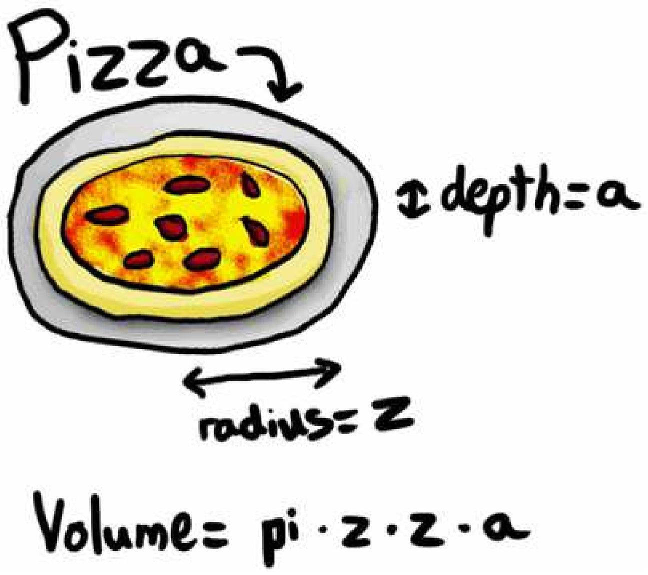

WHAT ARE YOU DOING?
WORKING ON MY MATH HOMEWORK.
47
I HAVE TO GO TO THE BATHROOM, I'LL BE RIGHT BACK.
OK.
47
page 2
$= \frac{\sin(\frac{1}{2}(4\alpha-2\alpha))}{\cos(-\alpha)}$
$= \frac{\sin(\alpha)}{\cos(-\alpha)}$
=
HMMM...
page 2
$= \frac{\sin(\frac{1}{2}(4\alpha-2\alpha))}{\cos(+\alpha)}$
$= \frac{\sin(\alpha)}{\cos(+\alpha)}$
=
BA HA HA HA HA HA!
47
spikedmath.com
© 2010

When you read or listen	What you have to understand
this is trivial (or obvious)	i can not say why this is true
we get automatically	idem
a calculation shows that	a calculation that I did not will certainly show that
the reader will easily control that	it bothers me to show that
we strongly recommend the reader to make the indicated exercises	as I have not made them, you could correct me
i showed this result in a previous paper	i have absolutely no idea how I did to prove that thing
is easily generalized to	the generalization is beyond my level
according to a well known property	known by maximum 10 people in the world
the proof is in two lines	yes, but through five lemmas
it's algebra	it is not interesting (in the mouth of an analyst)
it's analysis	it is not interesting (in the mouth of an algebrist)
it's elementary (or classical)	in bornitziens space theory bornitziens of the second kind
i did not understand this step in your demonstration	you're stuck in your demo
this conference was very interesting	i did not understand anything

* * *

The number you requested is imaginary, please turn your phone to a quarter turn right and renumber...

* * *



A mathematician to his friend:

- "Are you faithful?"
- "Yes, up to isomorphism"

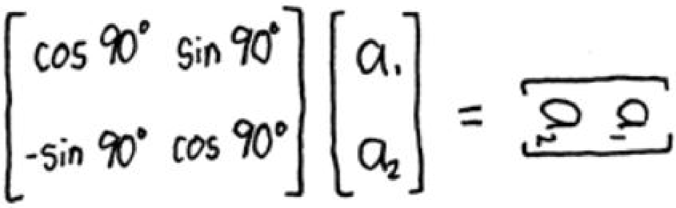
How mathematicians do it...:

- Algebraists do it by symbolic manipulation.
- Algebraists do it in a ring, in fields, in groups.
- Analysts do it continuously and smoothly.
- Applied mathematicians do it by computer simulation.
- Banach spacers do it completely.
- Bayesians do it with improper priors.
- Catastrophe theorists do it falling off part of a sheet.
- Combinatorists do it as many ways as they can.
- Complex analysts do it between the sheets
- Computer scientists do it depth-first.
- Cosmologists do it in the first three minutes.
- Decision theorists do it optimally.
- Functional analysts do it with compact support.
- Galois theorists do it in a field.
- Game theorists do it by dominance or saddle points.
- Geometers do it with involutions.
- Geometers do it symmetrically.
- Graph theorists do it in four colors.
- Hilbert spacers do it orthogonally.

- Large cardinals do it inaccessibly.
- Linear programmers do it with nearest neighbours.
- Logicians do it by choice, consistently and completely.
- Logicians do it incompletely or inconsistently.
- (Logicians do it) or [not (logicians do it)].
- Number theorists do it perfectly and rationally.
- Mathematical physicists understand the theory of how to do it, but have difficulty obtaining practical results.
- Pure mathematicians do it rigorously.
- Quantum physicists can either know how fast they do it, or where they do it, but not both.
- Real analysts do it almost everywhere
- Ring theorists do it non-commutatively.
- Set theorists do it with cardinals.
- Statisticians probably do it.
- Topologists do it openly, in multiply connected domains
- Variationists do it locally and globally.
- Cantor did it diagonally.
- Fermat tried to do it in the margin, but couldn't fit it in.
- Galois did it the night before.
- Mðbius always does it on the same side.
- Markov does it in chains.
- Newton did it standing on the shoulders of giants.
- Turing did it but couldn't decide if he'd finished.

What will a complex French guy said to a real woman?

Answer: "viens danser !" (you have to read "come in C", the complex set...)



Two-way sentences in French and English:

- We solve now this problem without complex
- Un repère d'origine O (un repaire d'originaux...)
- Une partie de $\mathbb{Q}$ (a fucking party...)
- Ne confondez pas un ρ avec un p... (don't confuse between a fart and a burp)
- Une variété de Poisson... (a variety of fish)

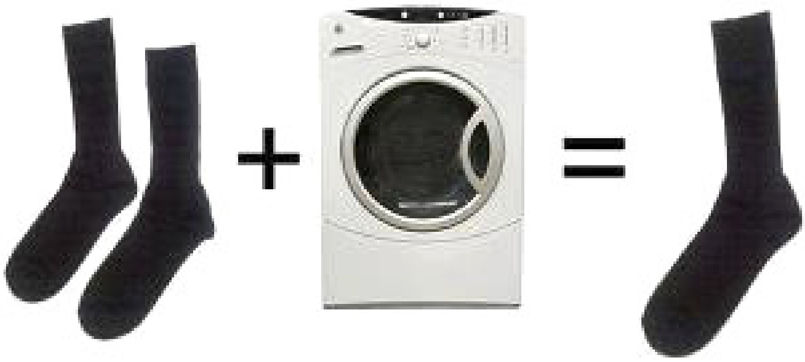
Logarithm and exponential functions are at the restaurant. When comes the addition, which will pay?

Answer: Exponential, because logarithme né paie rien (in English "neperien" phonetically means: pay nothing)

Later in the evening, Logarithm and Exponential go home a little bit drunk. Logarithm asks: Do I drive?

Exponential answers: I'd rather it be me who leads. In the case you derivate...

Two Cauchy sequences want to go out to dance. They arrive at a club where the "No Limit" evening event takes place. They decide to enter, but the guard stops them, saying, "Sorry, we're complete" (in a complete space, a Cauchy sequence is convergent by definition, so it has a limit).



A mathematician went fool and believed that he was the differentiation operator. His friends had placed him in a mental hospital until he got better. All day he would go around frightening the other patients by staring at them and saying: "I differentiate you!"

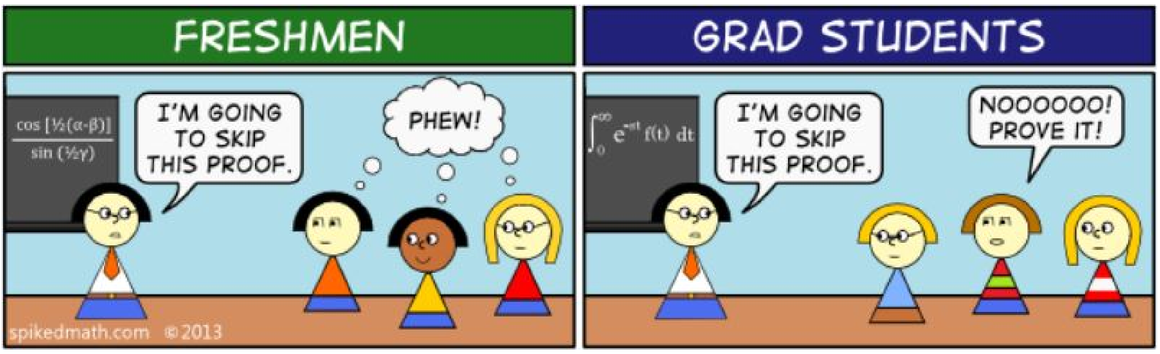
One day he met a new patient; and true to form he stared at him and said "I differentiate you!", but for once, his victim's expression didn't change. Surprised, the mathematician marshalled his energies, stared fiercely at the new patient and said loudly "I differentiate you!", but still the other man had no reaction. Finally, in frustration, the mathematician screamed out "I DIFFERENTIATE YOU!"

The new patient calmly looked up and said: "You can differentiate me all you like: I'm e to the x."

What mathematicians say and what you have to understand:

- Trivial: If I have to show you this, you're in the wrong class
- One can trivially show: We do not need more than 4 hours to write the proof
- Check yourself: This is the hard part of the demonstration so you can do on your spare time
- Similarly: At least one line of the proof is identical to the previous
- Proceed formally: We will manipulate symbols with many predefined rules without understanding the real meaning of the result.

- We will provide us the proof: Trust me, it's true!
- The reader will easily show: I get tired to show that...
- We strongly suggest the reader to make the indicated exercises: As I have not made them, you could correct me
- I showed this result in a previous paper: I do not know how the devil did to prove this thing
- We generalize easily: The generalization is beyond my level
- According to a well known property: For 10 people in the world ...


FRESHMEN
$\frac{\cos[\frac{1}{2}(\alpha-\beta)]}{\sin(\frac{1}{2}\gamma)}$
I'M GOING TO SKIP THIS PROOF.
PHEW!
GRAD STUDENTS
$\int_0^\infty e^{-st} f(t)\,dt$
I'M GOING TO SKIP THIS PROOF.
NOOOOOO! PROVE IT!
spikedmath.com ©2013

There are 3 kinds of people: those who can count and those who can not count...

Everyone knows the "Salary Theorem" which states that engineers and scientists can NEVER earn as much as businessmen and commercial. This theorem can then be demonstrated by solving a simple math equation.

Our equation is based on two well known postulates:

P1. The Knowledge is Power

P2. Time is Money:

All engineers know that:

$$\text{Power} = \frac{\text{Work}}{\text{Time}}$$

and:

$$\text{Knowledge} = \text{Power}$$

and also:

$$\text{Time} = \text{Money}$$

We then obtain by substitution:

$$\text{Knowledge} = \frac{\text{Work}}{\text{Money}}$$

And we finally get the following result:

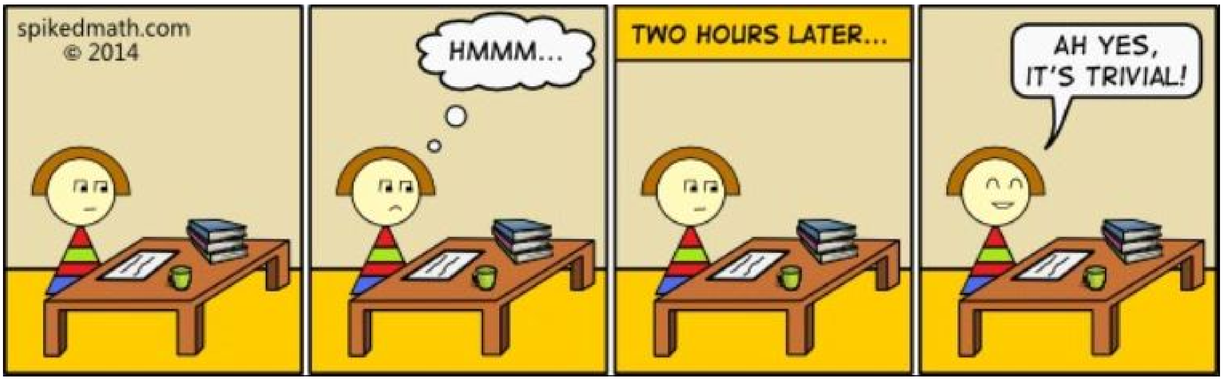
$$\text{Money} = \frac{\text{Work}}{\text{Knowledge}}$$

So when the Knowledge approaches zero, Money approaches infinity regardless of the value attributed to work, this value may be very low. Conversely when Knowledge goes to infinity, the Silver tends to zero, even if the work is high value.

Hence the obvious conclusion follows: The less you know, the more money you make.

PS: Those of you who have had some difficulty understanding this should be the highest paid.

■ Q.E.D.


spikedmath.com
© 2014
HMMM...
TWO HOURS LATER...
AH YES, IT'S TRIVIAL!

In the same kind here is the "misogynistic theorem":

First, we state that girls factorable variables in amounts of time and money such as:

$$\text{Girls} = \text{Time} \times \text{Money}$$

As we all know "time is money". So:

$$\text{Time} = \text{Money}$$

and because "money is the root of evil ...":

$$\text{Money} = \sqrt{\text{Evil}}$$

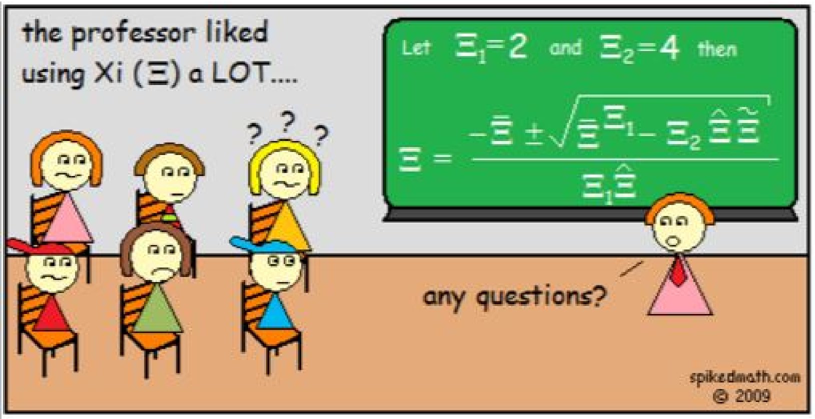
Then we have by substitution:

$$\text{Girls} = \left(\sqrt{\text{Evil}}\right)^2$$

We are therefore forced to conclude:

$$\text{Girls} = \text{Evil}$$

■ Q.E.D.


the professor liked using Xi (Ξ) a LOT....
Let $\Xi_1=2$ and $\Xi_2=4$ then
$\Xi = \frac{-\Xi \pm \sqrt{\Xi^{\Xi_1} - \Xi_2 \hat{\Xi} \tilde{\Xi}}}{\Xi_1 \hat{\Xi}}$
? ? ?
any questions?
spikedmath.com © 2009

The top ten excuses for not doing your math homework:

#10. Galileo didn't know calculus; what do I need it for?

#09. A math addict stole my homework.

#08. I'm taking physics and the homework in there seemed to involve math, so I thought I could just do that instead.

#07. I have the proof, but there isn't room to write it in the margin.

#06. I have a solar powered calculator and it was cloudy.

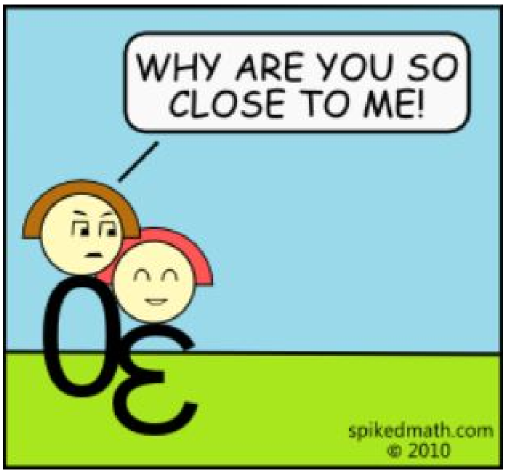
#05. I was watching the World Series and got tied up trying to prove that it converged.

#04. I could only get arbitrarily close to my textbook. (I reached half way, and then half of that, and then ...)

#03. I couldn't figure out whether i am the square root of negative one or i is the square root of negative one.

#02. It was Einstein's birthday and pi day and we had this big celebration! (This only works for March 14)

#01. I accidentally divided by zero and my paper burst into flames.


WHY ARE YOU SO CLOSE TO ME!
0
3
spikedmath.com
© 2010

What is the result of:

$$\frac{2ab}{2Fr.16}$$

(read "2 abbés sur 2 françaises")

Answer:

$$2bb\frac{a}{e}$$

(read "2 bébés assurés")

What is the result of:

$$\frac{\text{cheval}}{\text{oiseau}}$$

(read "horse on bird")

As we have:

$$\frac{\text{cheval}}{\text{oiseau}}=\frac{\text{vache}\cdot 1}{\beta\cdot 1}$$

(read "vache + l" (contains all letters of the word "cheval") divided by "bête à ailes" meaning animal with wing)

But:

$$\frac{\text{vache}\cdot 1}{\beta\cdot 1}=\frac{\beta\cdot\pi\cdot 1}{\beta\cdot 1}$$

(read "bête à pie + l" divided by "bête à ailes")

We simplify to get:

$$\frac{\cancel{\beta}\cdot\pi\cdot\cancel{1}}{\cancel{\beta}\cdot\cancel{1}}=\pi$$

This proves that π is irrational because there is no rational comparison between a "cheval" (horse) and an "oiseau" (bird)...

We have to prove that:

$$\frac{\text{ROSSINI}}{\text{SOLSIDO}}=1$$

We can write this as following:

$$\frac{\text{ROS SI NI}}{\text{SOL SI DO}}=\frac{\text{ROS NI}}{\text{SOL DO}}$$

but "NI vaut Do" (in French this means "niveau d'eau" or in English "water level") thus:

$$\frac{\text{ROS}}{\text{SOL}}$$

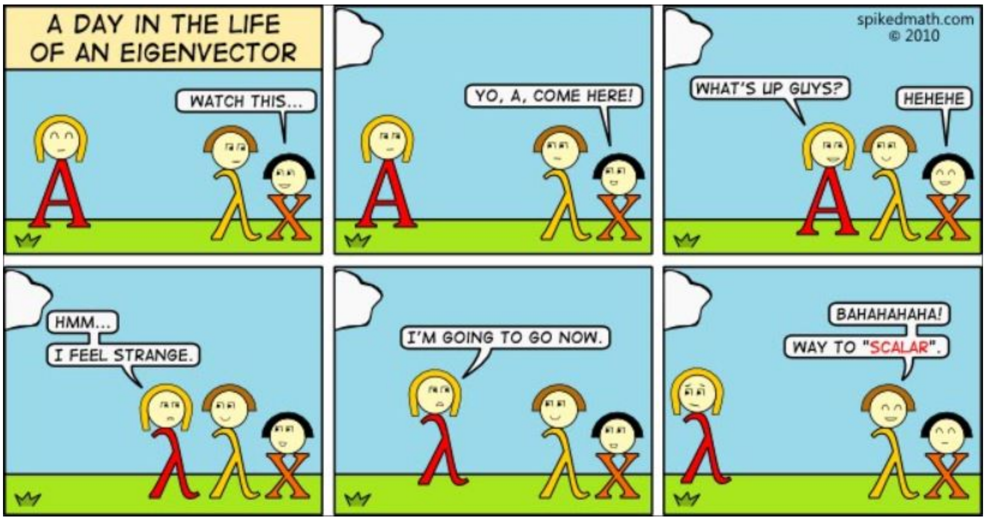
But "SOL fait RINO" (Solferino: it's during this battle that Henri Dunant had the idea to create the Red Cross) then:

$$\frac{\text{ROS}}{\text{RI NO}}$$

because "RINO c'est ROS" (rhinocéros) then RINO = ROS and we finally have:

$$\frac{\text{ROS}}{\text{ROS}}=1$$

■ Q.E.D.


A DAY IN THE LIFE OF AN EIGENVECTOR
WATCH THIS...
A
λ
x
YO, A, COME HERE!
spikedmath.com © 2010
WHAT'S UP GUYS?
HEHEHE
HMM...
I FEEL STRANGE.
I'M GOING TO GO NOW.
BAHAHAHAHA!
WAY TO "SCALAR".

Each future engineer learns to write the sum of two rational numbers, for example:

$$1+1=2$$

This form is however rather banal and indicates gaps in your education.

In the first semester, we learn that:

$$1=\ln(e)$$

and:

$$1=\sin^2(p)+\cos^2(q)$$

Also everybody know that:

$$2=\sum_{n=0}^{+\infty}\left(\frac{1}{2}\right)^n$$

and that therefore the equation:

$$1+1=2$$

can be written more simply:

$$\ln(e)+\sin^2(p)+\cos^2(q)=\sum_{n=0}^{+\infty}\left(\frac{1}{2}\right)^n$$

we must admit that the look is much clearer and more scientific...

On the other hand, it is clear that:

$$1=\cosh(q)\sqrt{1-\tanh^2(q)}$$

and also:

$$e=\lim_{z\to+\infty}\left(1+\frac{1}{z}\right)$$

it follows that:

$$\ln(e)+\sin^2(p)+\cos^2(q)=\sum_{n=0}^{+\infty}\left(\frac{1}{2}\right)^n$$

can be rewritten as follows:

$$\ln\left(\lim_{z\to+\infty}\left(1+\frac{1}{z}\right)\right)+\sin^2(p)+\cos^2(q)=\sum_{n=0}^{\infty}\left(\frac{\cosh(q)\sqrt{1-\tanh^2(q)}}{2}\right)^n$$

We must also remember that:

$$0!=1$$

and the inverse exponent of the exponent is opposite equal to the exponent of the exponent opposite. Assuming an n-dimensional space, we know that:

$$\left(X^T\right)^{-1}-\left(X^{-1}\right)^{-T}=0$$

Taking the matrix as the metric of an oriented and orthogonal canonical space:

$$\left(g_{ij}^T\right)^{-1}-\left(g_{ij}^{-1}\right)^{-T}=0$$

logically we obtain:

$$\left(\left(g_{ij}^T\right)^{-1}-\left(g_{ij}^{-1}\right)^{-T}\right)!=1$$

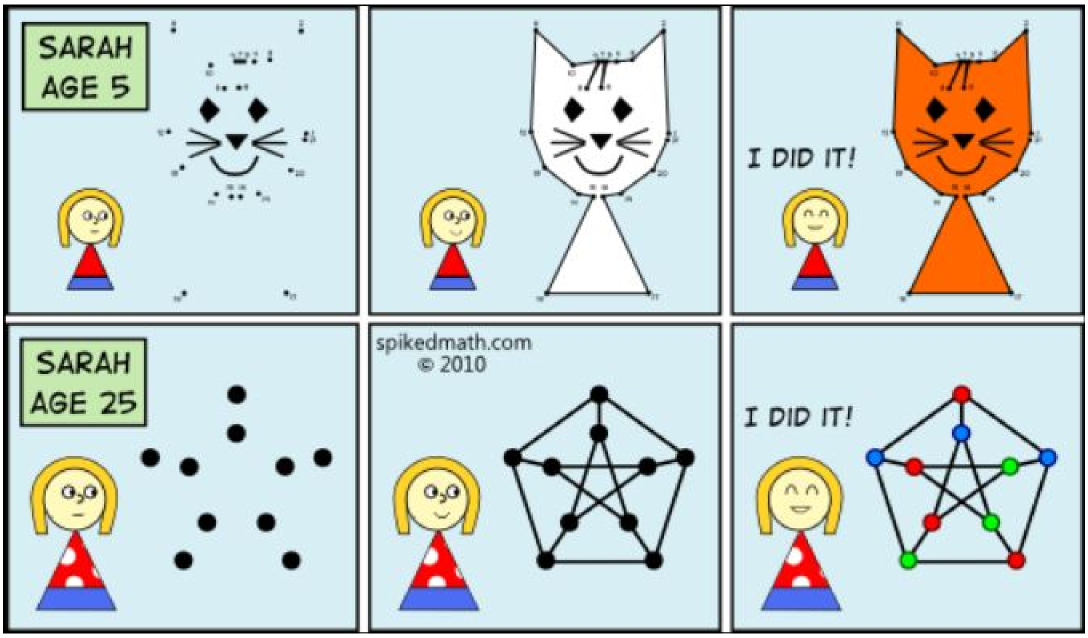
we get a simple and clear expression of $1+1=2$ for everyone:

$$\ln\left(\lim_{z\to+\infty}\left(\left(\left(g_{ij}^T\right)^{-1}-\left(g_{ij}^{-1}\right)^{-T}\right)!+\frac{1}{z}\right)\right)+\sin^2(p)+\cos^2(q)=\sum_{n=0}^{+\infty}\left(\frac{\cosh(q)\sqrt{1-\tanh^2(q)}}{2}\right)^n$$

It is therefore obvious that this equation is much more understandable than:

$$1+1=2$$

It would be possible to show several other developments of this simple expression and we will do from the moment you begin to understand the simple principles of the previous method.

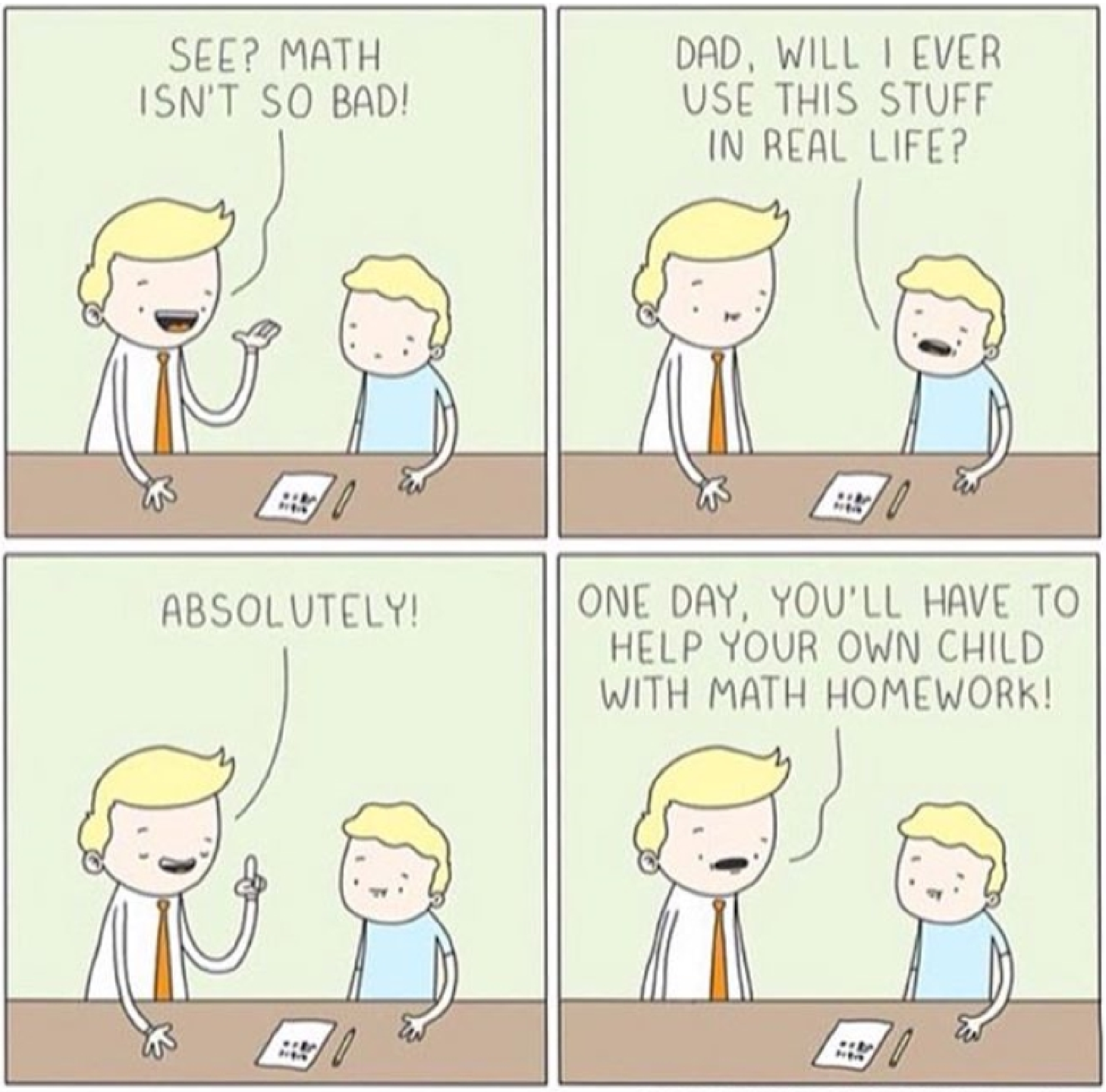


A wrestler, a physicist and a mathematician are subject to an experience: they are locked in a room each with a box of spinach, closed, and no can opener. After 24 hours, we'll see what they have become.

- The wrestler was able to open her box, "Well, I just flung violently the box against the wall. The impact was such that it is open", he explains.
- The physicist also managed to open her box: "I watched the box, and distinguished his break points. I then performed a pressure to exert maximum force on them, and the box was naturally open."
- The mathematician, finally, is found prostrate in a corner of the room, the sweat streaming down his face, and his box, closed, between the feet: "We admit that the box is opened ... We admit that... ".

Why we write d instead of d for the differential operator:

$$\frac{\cancel{d}}{\cancel{d}x}x=\frac{1}{\cancel{x}}\cancel{x}=1$$


SEE? MATH ISN'T SO BAD!
DAD, WILL I EVER USE THIS STUFF IN REAL LIFE?
ABSOLUTELY!
ONE DAY, YOU'LL HAVE TO HELP YOUR OWN CHILD WITH MATH HOMEWORK!

THEGENTLEMANSARMCHAIR.COM

Mathematics of life:

$$\begin{array}{lcr} & \text{Life} + \text{Love} = & \text{Happy} \\ + & \text{Life - Love} = & \text{Sad} \\ \hline \text{2Life} & = & \text{Happy+Sad} \end{array}$$

Therefore:

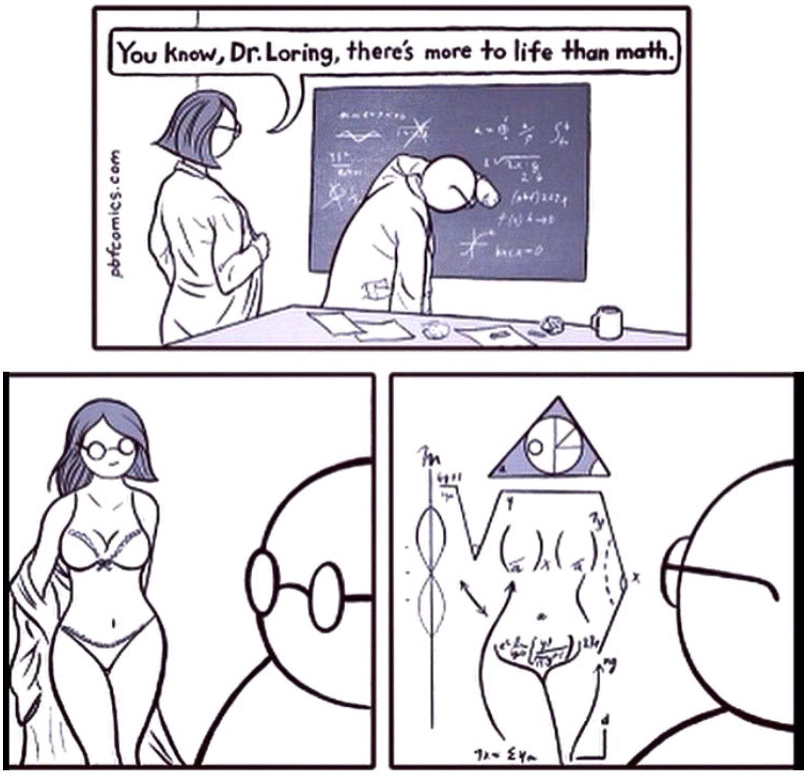
$$\text{Life} = \frac{\text{Happy} + \text{Sad}}{2}$$

Developing:

$$\text{Life} = \frac{1}{2}\text{Happy} + \frac{1}{2}\text{Sad}$$

That's real Life. Enjoy it!

An opinion without 3.14 is an onion. You'll understand!


You know, Dr. Loring, there's more to life than math.
pbfcomics.com

A logician's wife is having a baby. The doctor immediately hands the newborn to the dad.

His wife asks impatiently: "So, is it a boy or a girl"?

The logician replies: "Yes".

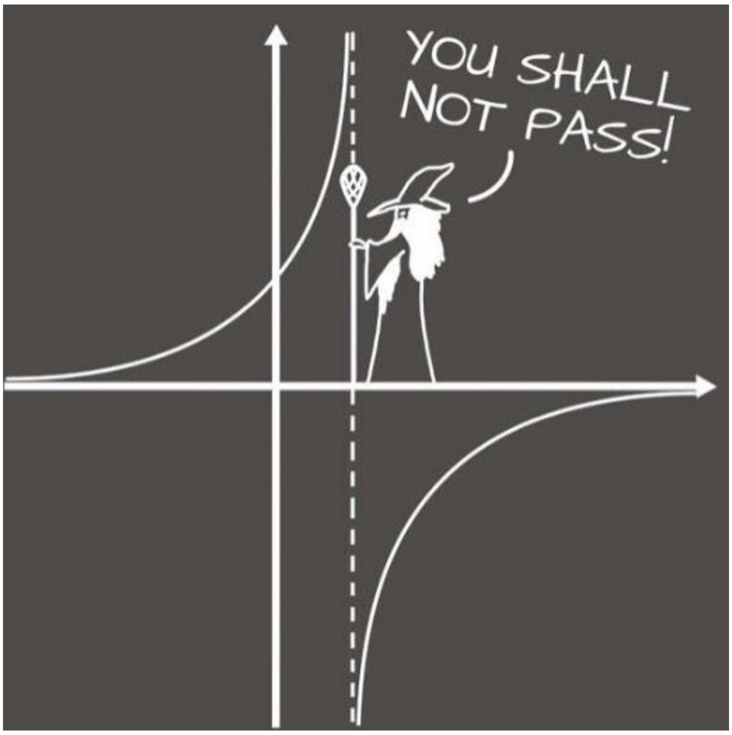
Question: What does the "B" in Benoit B. Mandelbrot stand for?

Answer: Benoit B. Mandelbrot

This is not the way we do this derivative...:

$$\frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{x}=\frac{\mathrm{d}}{\mathrm{d}}\frac{1}{x^2}=\frac{\cancel{\mathrm{d}}}{\cancel{\mathrm{d}}}\frac{1}{x^2}=-\frac{1}{x^2}$$

* * *


YOU SHALL NOT PASS!

* * *

Given:

$$\overline{NO}\|\overline{UH}$$
$$\overline{UM}\|\overline{OH}$$

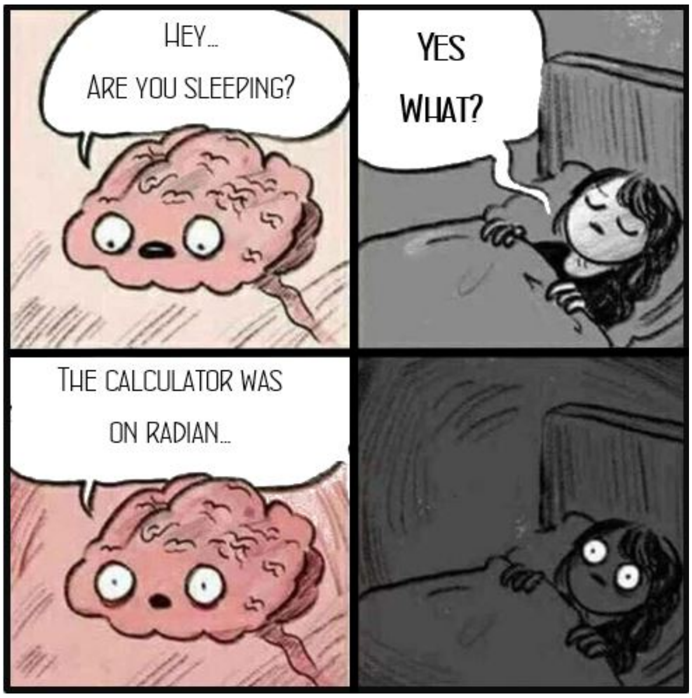
Prove:

$$UH\cdot OH=UM\cdot NO$$

* * *

$$\sqrt{-1}\quad 2^3\quad \sum\quad \pi \tag{20.2.1}$$

... and it was delicious!


HEY...
ARE YOU SLEEPING?
YES
WHAT?
THE CALCULATOR WAS
ON RADIAN...

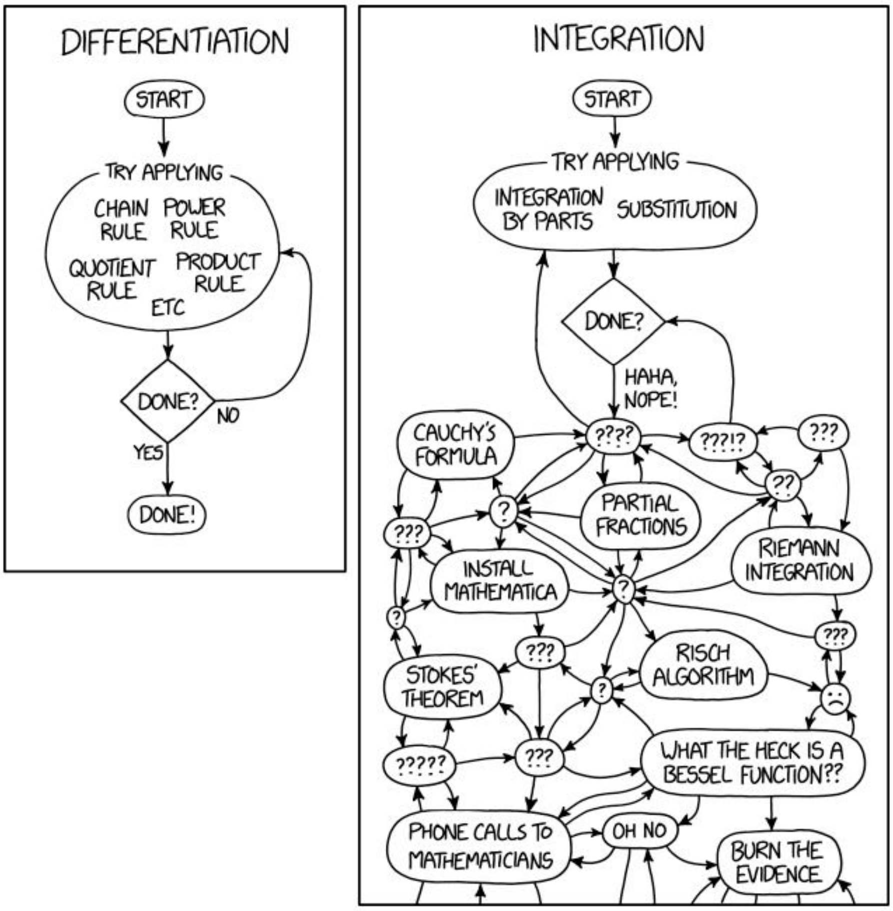
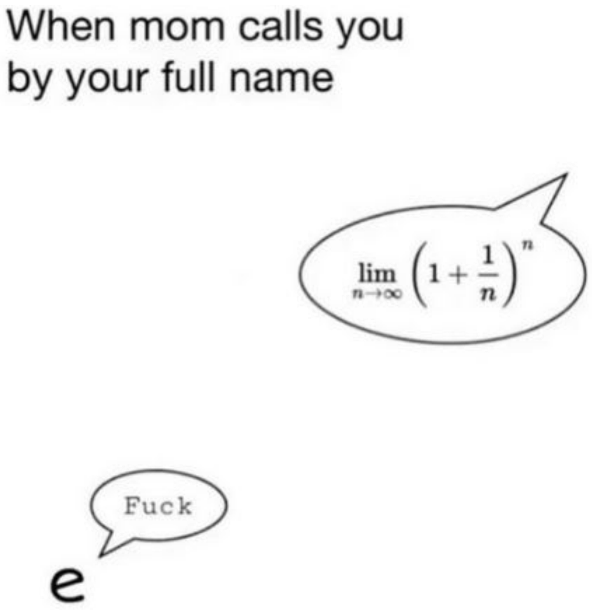
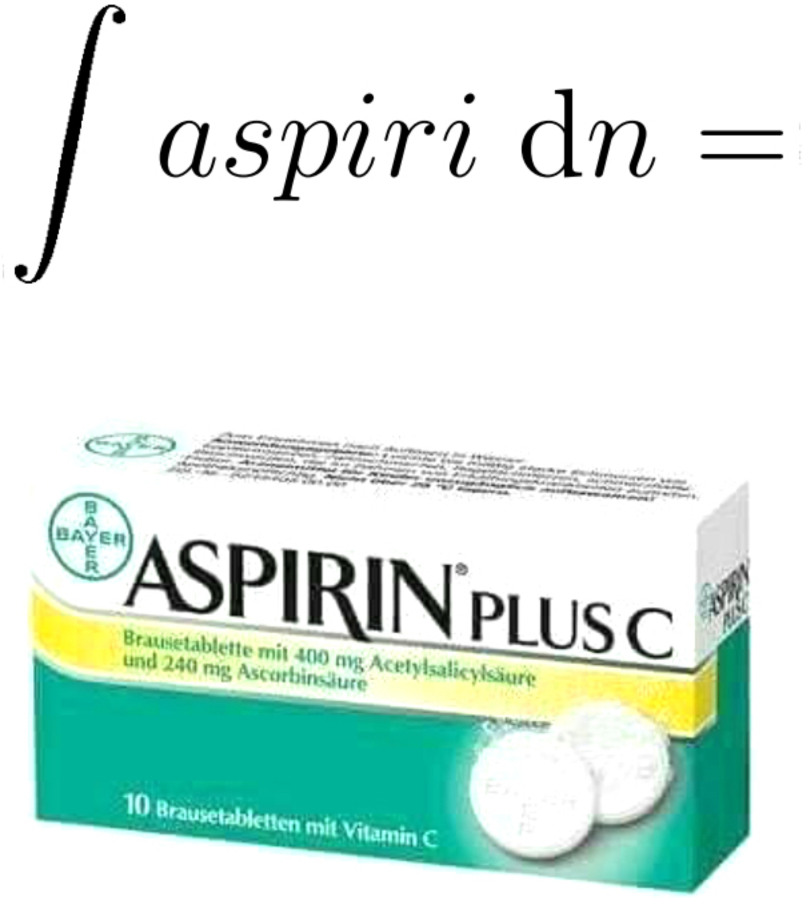
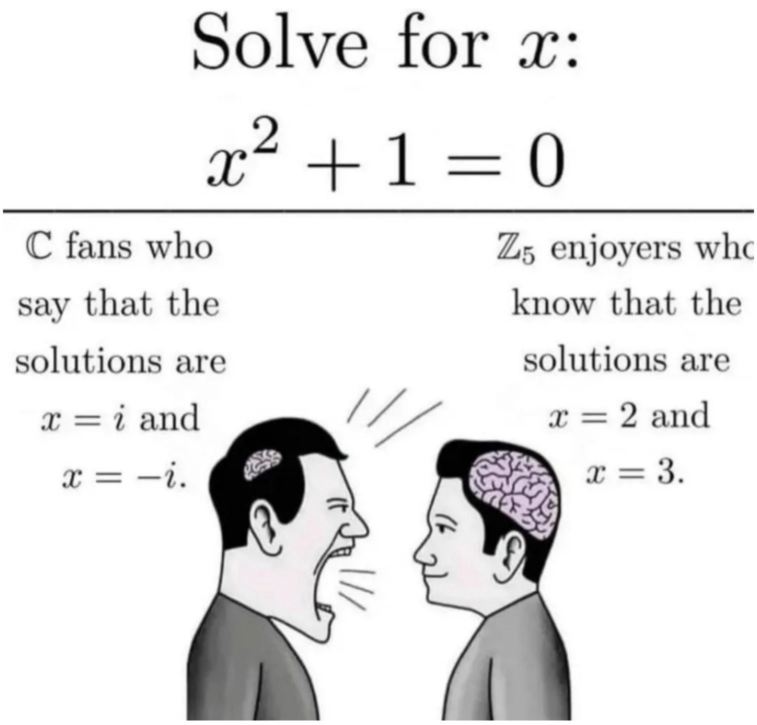
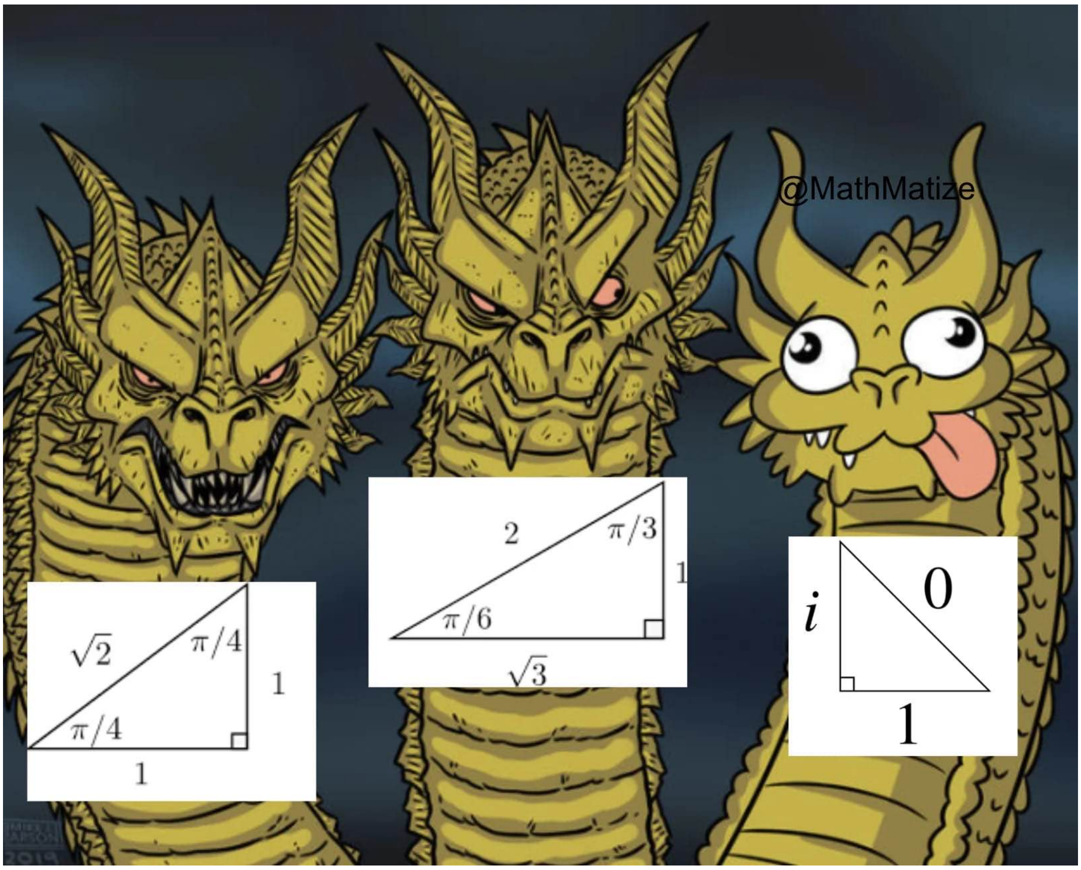

DIFFERENTIATION
START
TRY APPLYING
CHAIN RULE
POWER RULE
QUOTIENT RULE
PRODUCT RULE
ETC
DONE?
NO
YES
DONE!
INTEGRATION
START
TRY APPLYING
INTEGRATION BY PARTS
SUBSTITUTION
DONE?
HAHA, NOPE!
CAUCHY'S FORMULA
????
???!?
???
??
?
???
PARTIAL FRACTIONS
RIEMANN INTEGRATION
INSTALL MATHEMATICA
?
?
???
STOKES' THEOREM
???
RISCH ALGORITHM
?
?????
???
WHAT THE HECK IS A BESSEL FUNCTION??
PHONE CALLS TO MATHEMATICIANS
OH NO
BURN THE EVIDENCE

Figure 20.1: **source:** www.xkcd.com **(author: Randall Munroe)**

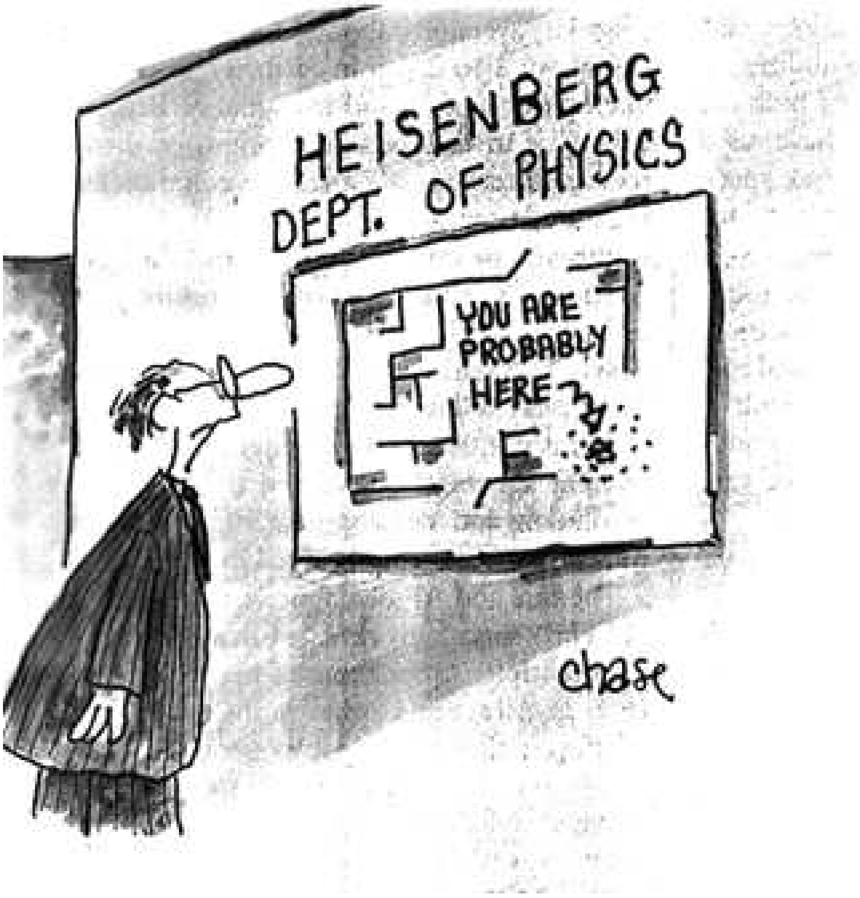

When mom calls you by your full name
$\lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n$
Fuck
e


$$\int aspiri\ dn =$$
BAYER
ASPIRIN® PLUS C
Brausetablette mit 400 mg Acetylsalicylsäure und 240 mg Ascorbinsäure
10 Brausetabletten mit Vitamin C


Solve for x:
$x^2+1=0$
$\mathbb{C}$ fans who say that the solutions are $x=i$ and $x=-i$.
$\mathbb{Z}_5$ enjoyers whc know that the solutions are $x=2$ and $x=3$.


@MathMatize
$\sqrt{2}$
$\pi/4$
1
$\pi/4$
1
2
$\pi/3$
1
$\pi/6$
$\sqrt{3}$
i
0
1

20.3. Physics



- In theory, there is no difference between theory and practice. In practice, there is a difference!
- The theory is when we know everything but nothing works. The practice is when everything works, but we do not know why. In computer science, theory and practice are met: nothing works and you do not know why!

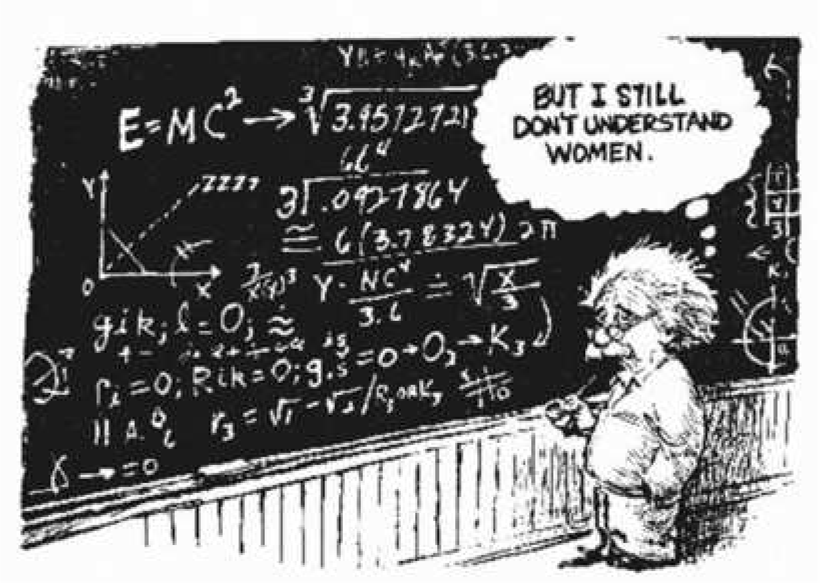
- Matter is fundamentally lazy - It always takes the path of least effort.
- Matter is fundamentally stupid - It tries every other path first.
- That is the heart of physics - The rest is details!

Two atoms meets together.

One says to the other: "Shit, I lost an electron!"

The other: "Are you sure?"

And the first replies, "POSITIVELY!!"

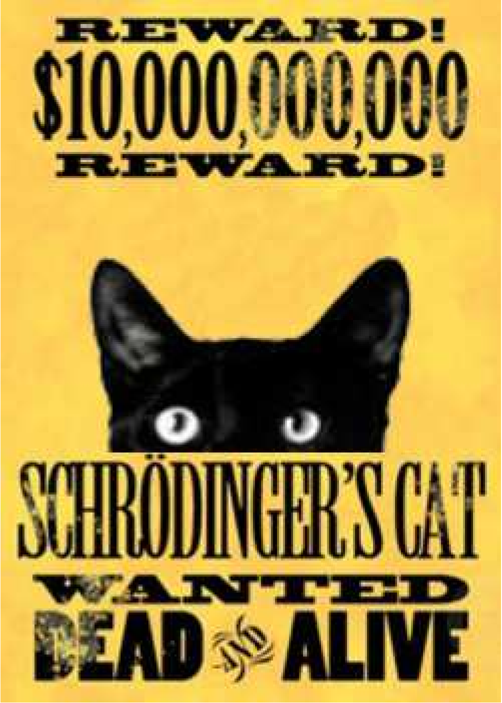

BUT I STILL DON'T UNDERSTAND WOMEN.
E=MC²

You enter the laboratory and see an experiment. How will you know which class is it?

- If it's green and wiggles, it's biology.
- If it stinks, it's chemistry.
- If it doesn't work, it's physics.

Theorem: A cat has nine tails.

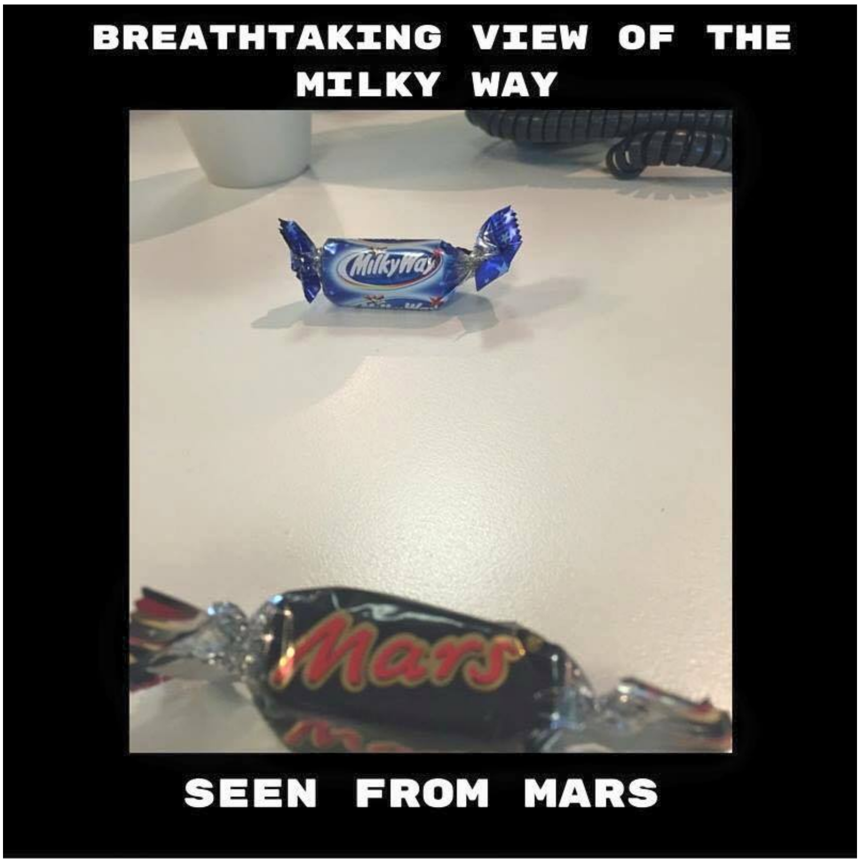
Proof: No cat has eight tails. A cat has one tail more than no cat. Therefore, a cat has nine tails.



A physicist studying quantum physics, is someone who does not see very well, looking in a dark room for a black cat, who probably does not exist.

An engineer, a physicist, a mathematician, and a mystic were asked to name the greatest invention of all times.

- The engineer chose fire, which gave humanity power over matter.
- The physicist chose the wheel, which gave humanity the power over space.
- The mathematician chose the alphabet, which gave humanity power over symbols.
- The mystic chose the thermos bottle.
  "Why a thermos bottle?" the others asked.
- The mystic: "Yes, because the thermos keeps hot liquids hot in winter and cold liquids cold in summer."
- "Yes... so what?" the others asked.
- The mystic: "Think about it." said the mystic reverently. "That little bottle.. how does it know?"

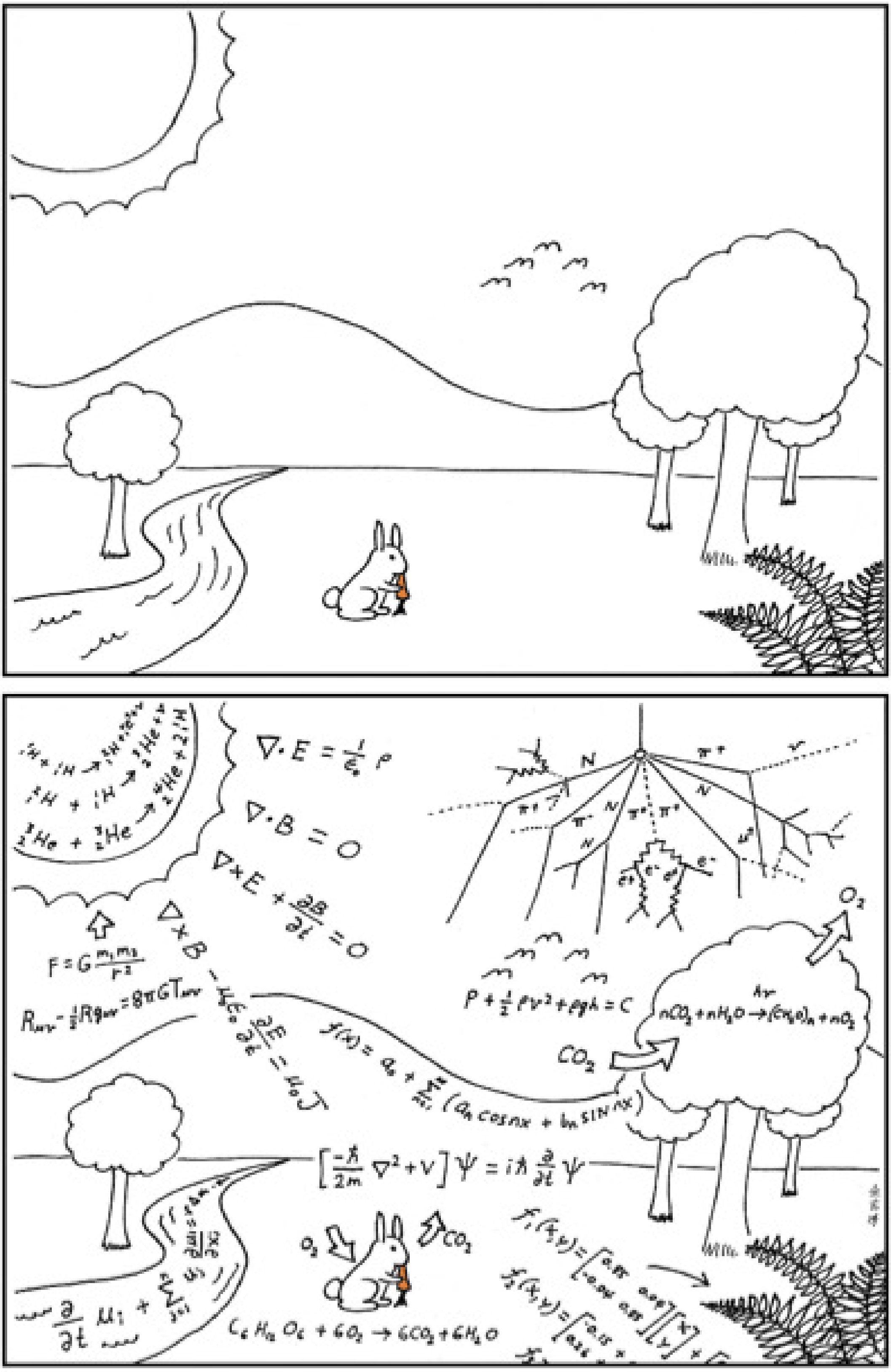


Physics professor has been doing an experiment, and has worked out an empirical equation that seems to explain his data. He asks the math professor to look at it.

A week later, the math professor says the equation is invalid. By then, the physics professor has used his equation to predict the results of further experiments, and he is getting excellent results, so he asks the math professor to look again.

Another week goes by, and they meet once more. The math professor tells the physics professor the equation does work, "But only in the trivial case where the numbers are real and positive".

Practical nuclear fusion power plants are just 30 years away - and always will be.


$^1_1H + ^1_1H \to ^2_1H + e^+ + \nu$
$^2_1H + ^1_1H \to ^3_2He + \gamma$
$^3_2He + ^3_2He \to ^4_2He + 2^1_1H$
$\nabla \cdot E = \frac{1}{\varepsilon_0}\rho$
$\nabla \cdot B = 0$
$\nabla \times E + \frac{\partial B}{\partial t} = 0$
$\nabla \times B - \mu_0\varepsilon_0 \frac{\partial E}{\partial t} = \mu_0 J$
$F = G\frac{m_1 m_2}{r^2}$
$R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} = 8\pi G T_{\mu\nu}$
N
π^+
π^0
e^-
e^+
O_2
$P + \frac{1}{2}\rho v^2 + \rho g h = C$
$nCO_2 + nH_2O \xrightarrow{h\nu} (CH_2O)_n + nO_2$
CO_2
$f(x) = a_0 + \sum_{n=1}^{\infty}(a_n \cos nx + b_n \sin nx)$
$\left[\frac{-\hbar}{2m}\nabla^2 + V\right]\Psi = i\hbar\frac{\partial}{\partial t}\Psi$
O_2
CO_2
$\frac{\partial}{\partial t}u_1 + \sum_{j=1}^{3} u_j \frac{\partial u_i}{\partial x_j}$
$C_6H_{12}O_6 + 6O_2 \to 6CO_2 + 6H_2O$
$f_1(x,y) =$
$f_2(x,y) =$

This is how scientists see the world.

Heisenberg was driving down the highway whereupon he was pulled over by a policeman.

The policeman asked:

- "Do you know how fast you were going back there?"

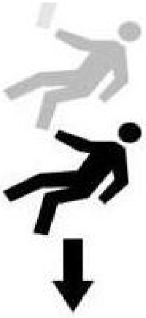
Heisenberg replied:

- "No, but I know where I am."

What's the difference between an auto mechanic and a quantum mechanic?

The quantum mechanic can sometimes get the car inside the garage without opening the door.

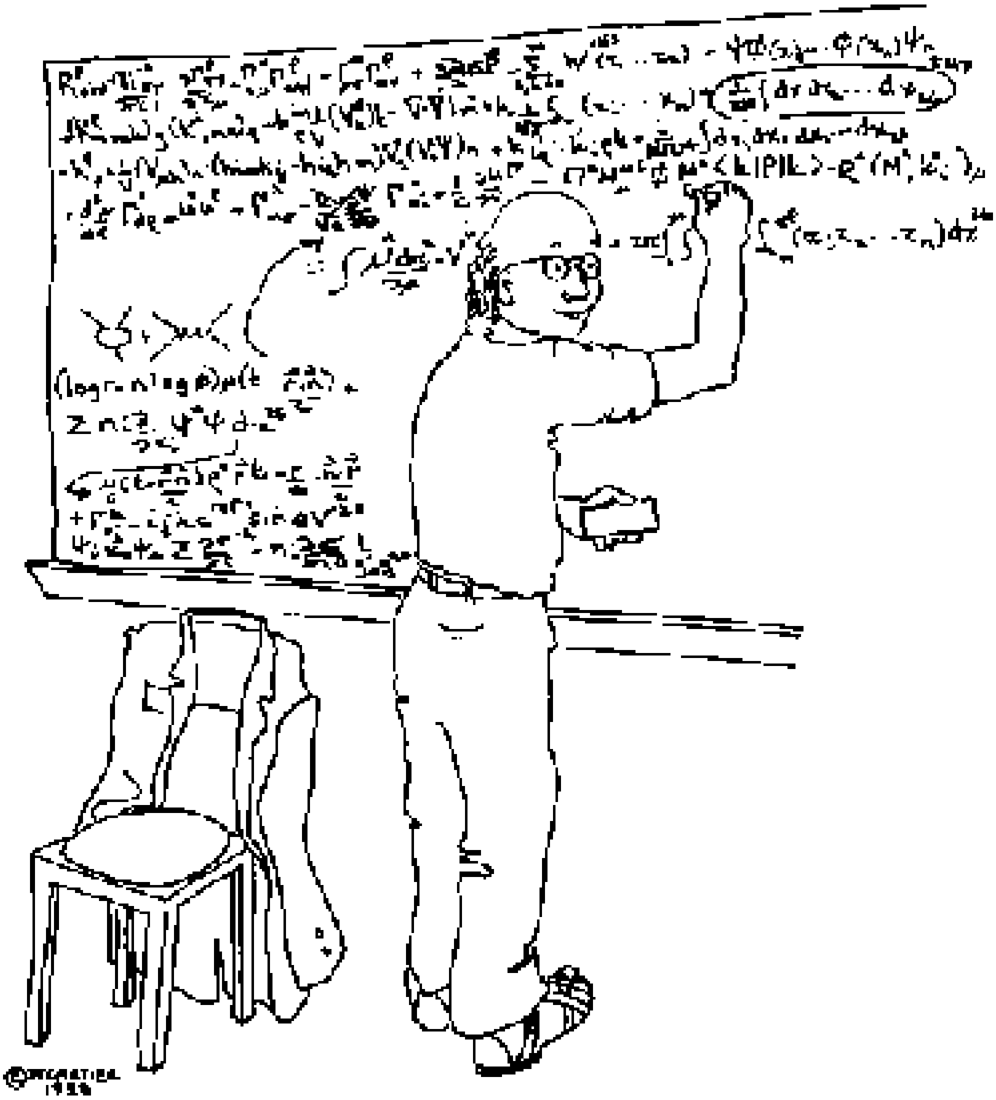
It's not the:



$$v_f = v_0 + at$$

that kills you, it's the:

$$F = m\frac{\Delta v}{\Delta t}$$

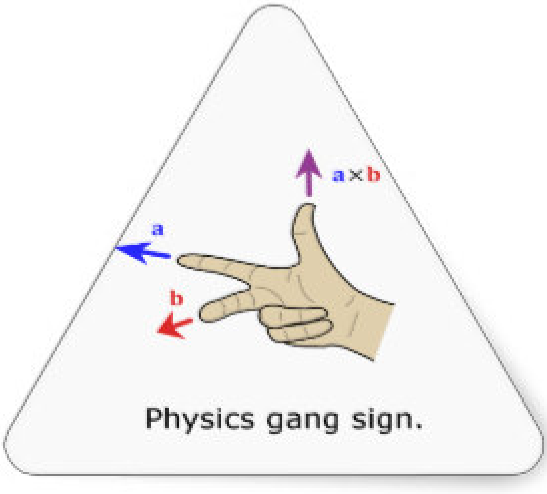


"At this point we notice that this equation is beautifully simplified if we assume that space–time has 92 dimensions."

Why God never received tenure at any University:

1. He had only one major publication.
2. It was in Hebrew.
3. It had no references.
4. It wasn't published in a refereed journal.
5. Some even doubt he wrote it himself.

6. It may be true that he created the world, but what has he done since then?
7. His cooperative efforts have been quite limited.
8. The scientific community has had a hard time replicating his results.
9. He never applied to the Ethics Board for permission to use human subjects.
10. When one experiment went awry he tried to cover it up by drowning the subjects.
11. When subjects didn't behave as predicted, he deleted them from the sample.
12. He rarely came to class, just told students to read the Book.
13. Some say he had his son teach the class.
14. He expelled his first two students for learning.
15. Although there were only ten requirements, most students failed his tests.
16. His office hours were infrequent and usually held on a mountaintop.



In the beginning there was Aristotle:

- And objects at rest tended to remain at rest
- And objects in motion tended to come to rest
- And God saw that it was boring, although very restful.

Then God created Newton:

- And objects at rest tended to remain at rest
- And objects in motion tended to remain in motion
- And energy was conserved, and momentum was conserved,
- And matter was conserved
- And God saw that it was conservative.

Then God created Einstein:

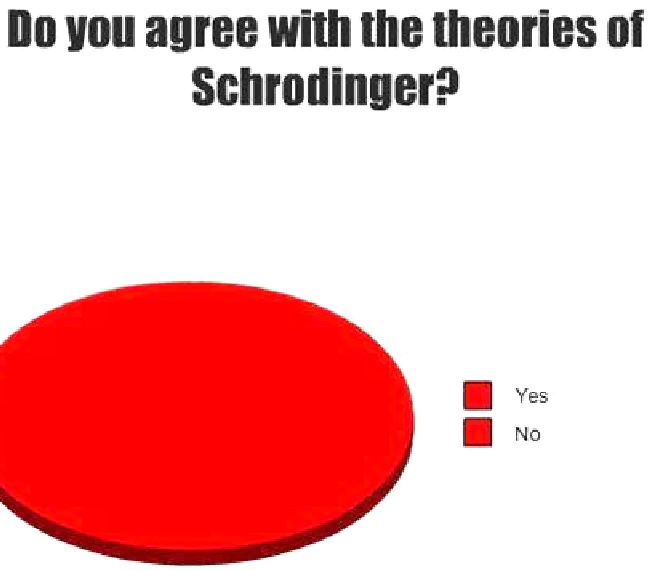
- And everything was relative
- And fast things became short
- And straight things became curved
- And the universe was filled with inertial frames
- And God saw that it was relatively general but some of it was especially relative.

Then God created Bohr:

- And there was the principle
- And the principle was quantum
- And all things were quantified
- But some things were still relative
- And God saw that it was confusing.

Then God was going to create Furgeson:

- And Furgeson would have unified
- And he would have fielded a theory
- And all would have been one.
- But it was the seventh day
- And God rested
- And objects at rest tend to remain at rest.

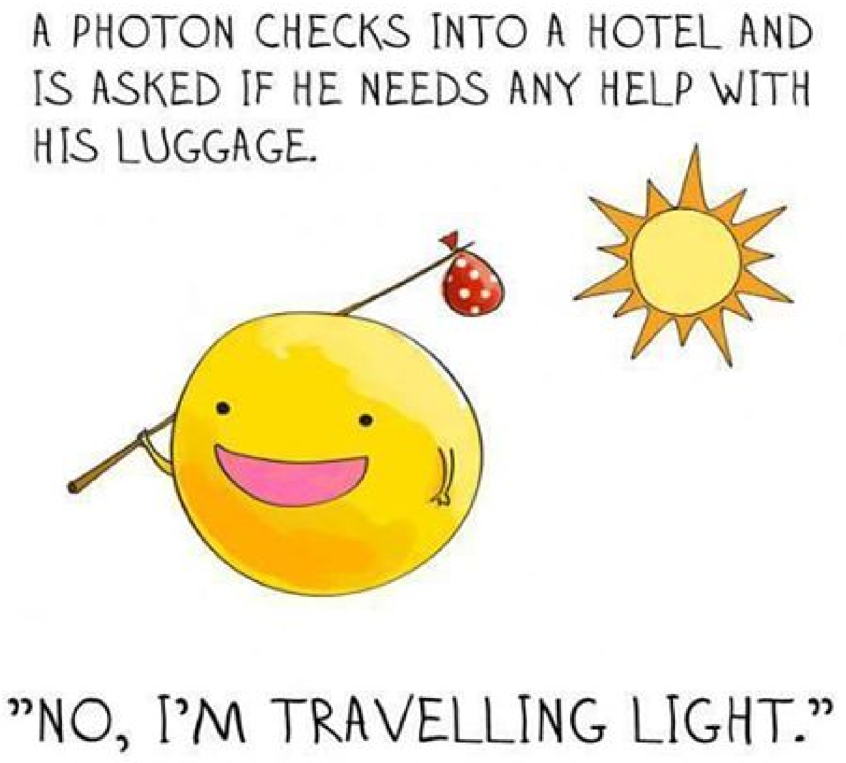


The Physicist's Bill of Rights

We hold these postulates to be intuitively obvious, that all physicists are born equal, to a first approximation, and are endowed by their creator with certain discrete privileges, among them a mean rest life, n degrees of freedom, and the following rights which are invariant under all linear transformations:

1. To approximate all problems to ideal cases.
2. To use order of magnitude calculations whenever deemed necessary (i.e. whenever one can get away with it).
3. To use the rigorous method of "squinting" for solving problems more complex than the addition of positive real integers.

4. To dismiss all functions which diverge as "nasty" and "unphysical".
5. To invoke the uncertainty principle when confronted by confused mathematicians, chemists, engineers, psychologists, dramatists, and other lower scientists.
6. hen pressed by non-physicists for an explanation of (4) to mumble in a sneering tone of voice something about physically naive mathematicians.
7. To equate two sides of an equation which are dimensionally inconsistent, with a suitable comment to the effect of, "Well, we are interested in the order of magnitude anyway".
8. To the extensive use of "bastard notations" where conventional mathematics will not work.
9. To invent fictitious forces to delude the general public.
10. To justify shaky reasoning on the basis that it gives the right answer.
11. To cleverly choose convenient initial conditions, using the principle of general triviality.
12. To use plausible arguments in place of proofs, and thenceforth refer to these arguments as proofs.
13. To take on faith any principle which seems right but cannot be proved.


A PHOTON CHECKS INTO A HOTEL AND IS ASKED IF HE NEEDS ANY HELP WITH HIS LUGGAGE.
"NO, I'M TRAVELLING LIGHT."

You can at least take this simple "Real Scientist Quiz" to find out if you're cut out for the life of a true scientist:

1. At Christmas time, you:
   a. Take a couple of days off to spend time with your family.
   b. Leave early on Christmas eve so you can pick up a few presents for the family.
   c. Only work half a day, spending the rest of the day at home working on your grant application.
2. Your spouse wants to discuss plans for the family vacation with your kids. You:
   a. Propose to go camping so you can explain the wonders of nature to your kids.
   b. Propose to go to another city so you can spend the day in your friend's lab while your spouse takes the kids sightseeing.
   c. Ask your spouse, "We have kids?"
3. At a scientific meeting on an island in the South Pacific, no talks are scheduled in the afternoon. During this free time, you:
   a. Follow the local custom and sunbathe on the beach in the nude.
   b. Sit on the beach fully clothed, unaware of the nude sunbathers, and discuss science with your colleagues.
   c. Sit in your hotel room with the drapes closed, and work on your manuscript.
4. The nurse at school calls to tell you that your second-grade child has chicken pox. You:
   a. Immediately drop what you're doing and rush to school to pick up your sick kid.
   b. Immediately drop what you're doing and begin trying to find a cure for chicken pox.
   c. Ask the nurse for directions to the school, and the names of your kids.
5. Beings from outer space visit Earth, and you are the first human they meet. To show their friendship, they present you with a highly advanced device that is capable of prolonging life, ending human suffering, and curing disease. You:
   a. Present it to the United Nations.
   b. Apply for a patent.
   c. Break it open to see how it works.
6. What is the longest amount of time that you have worked without a vacation (excluding scientific meetings)?
   a. Six months.
   b. Two years.
   c. I took a weekend off about 10 years ago.
7. What are your hobbies?

   a. Sports, music, and dance, because they allow the analytical parts of my brain to relax.
   b. Cooking, because it’s quite a lot like science.
   c. Reading back issues of scientific journals cover to cover.

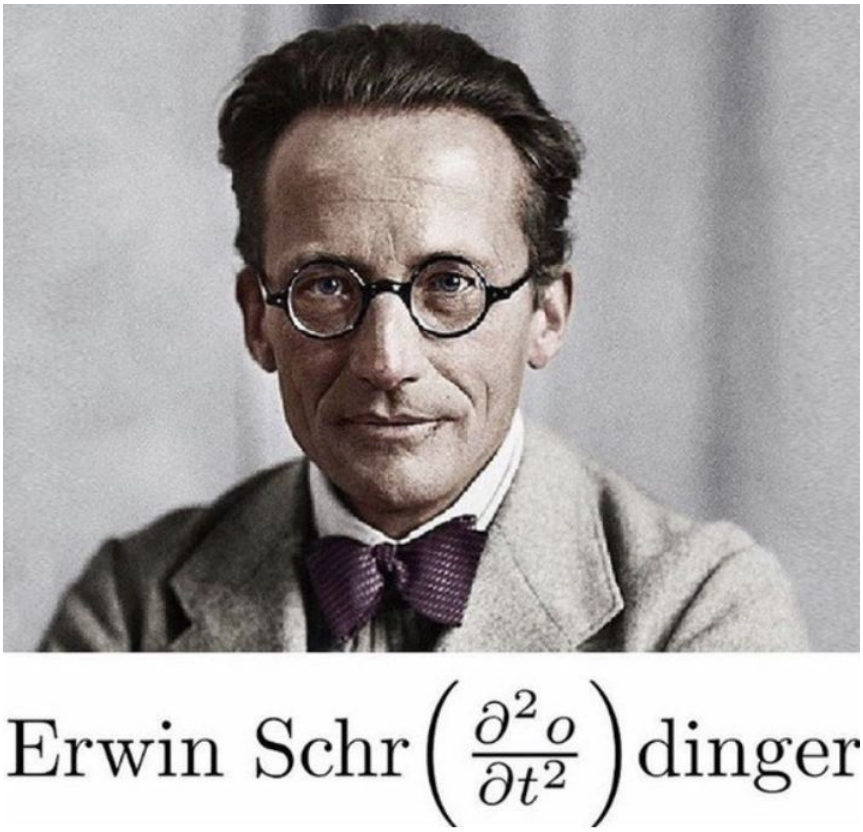
8. Your best friend is:

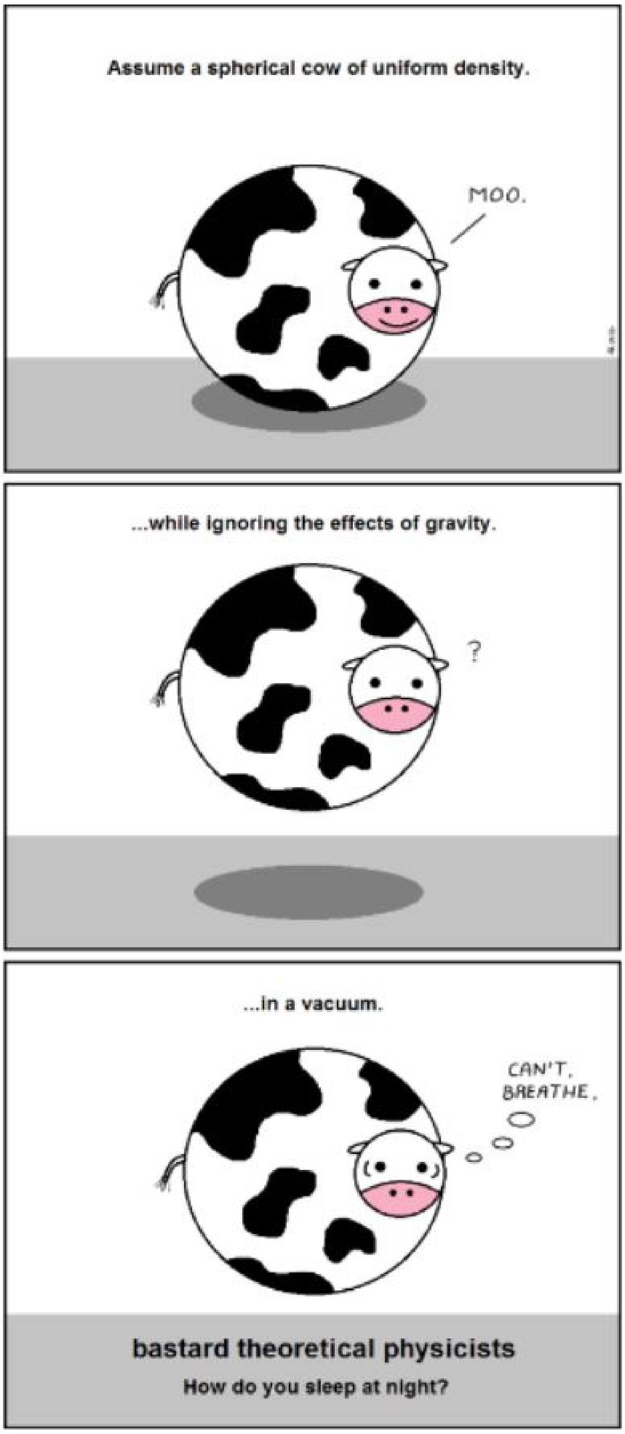
   a. A member of your college fraternity.
   b. A member of your immediate family.
   c. A member of a gene family.

Score:

Give yourself 1 point for every question you answered with an "a", 5 points for every "b" and 50 points for every "c". If you took the test three times and averaged your score, give yourself 100 extra points. If you calculated the standard error of the mean, give yourself 500 points.

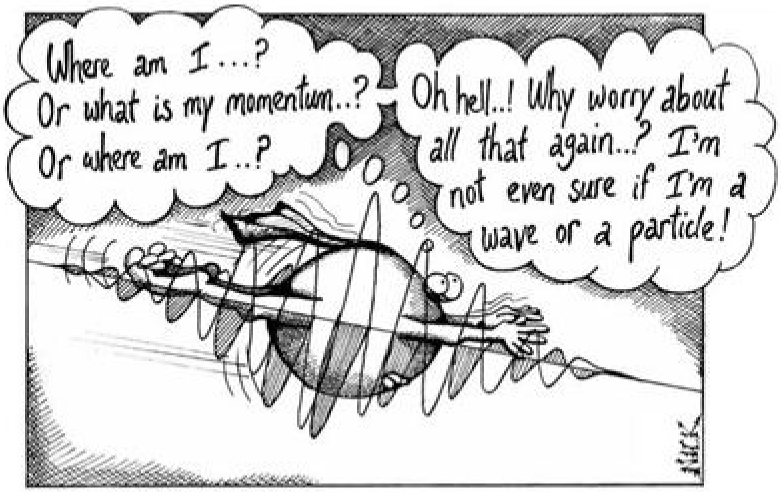
If you scored less than 10, you are normal. Scores of 11-50 indicate an obsessed scientist. If you scored more than 50, you are in need of help and should consider joining *Scientists Anonymous*; if you scored greater than 500, you should forget *Scientists Anonymous* and get back to work since you are beyond help, and may actually succeed as a scientist.


Erwin Schr$\left(\frac{\partial^2 o}{\partial t^2}\right)$dinger


Assume a spherical cow of uniform density.
MOO.
...while ignoring the effects of gravity.
?
...in a vacuum.
CAN'T. BREATHE.
bastard theoretical physicists
How do you sleep at night?

How you must understand certain sentences in physicists publications of:

- It is well known that...: I did not read the references, but..
- This is of great theoretical importance: This is important for me.
- Though it was not possible to give a definitive answer: The experiment failed, but it seems me valuable enough to write a publication.
- The used technique was particularly appropriate ...: The lab next door friend had already developed the technique.
- 3 samples were chosen for an exhaustive study: The results obtained from other samples yielded nothing coherent.
- Handled with extreme caution throughout the experiment: Was not thrown down the trash.
- The agreement with theory is excellent: it is passable.
- The agreement with theory is good: it is weak.
- The agreement with theory is satisfactory: it is doubtful.
- The agreement with theory is fair: it is totally imaginary.
- It is generally accepted that...: two colleagues agree with me
- It is recognized that: I think.
- It is clear that further work will be useful: I did not understand anything.
- Here are some typical results: Here are the best results.
- Significant in a confidence interval of...: not significant.
- The reagents used were synthesized in the laboratory according to standardized techniques: The reagents were purchased from...
- Unfortunately, quantitative basis to exploit the results have not yet been made: Nobody was able to understand anything of what was observed.
- We are grateful to X for his valuable collaboration and Y for the fruitful discussions: X and Y did the work and told me what the results meant.


Where am I...? Or what is my momentum..? Or where am I..?
Oh hell..! Why worry about all that again..? I'm not even sure if I'm a wave or a particle!
HACK

An experimental physicist performs an experiment involving two cats, and an inclined tin roof.

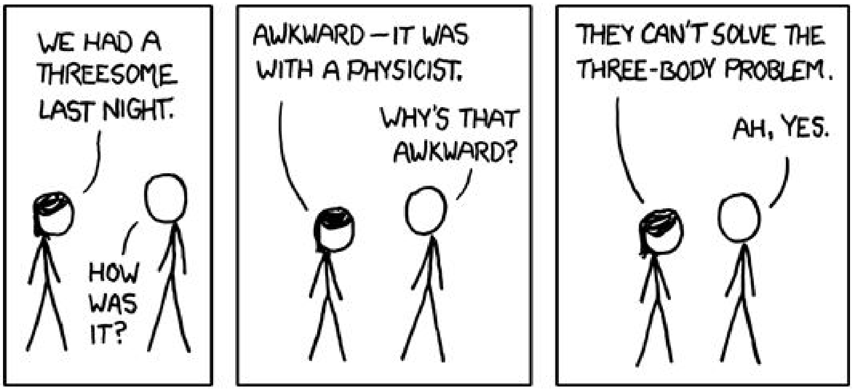
The two cats are very nearly identical; same sex, age, weight, breed, eye and hair color.

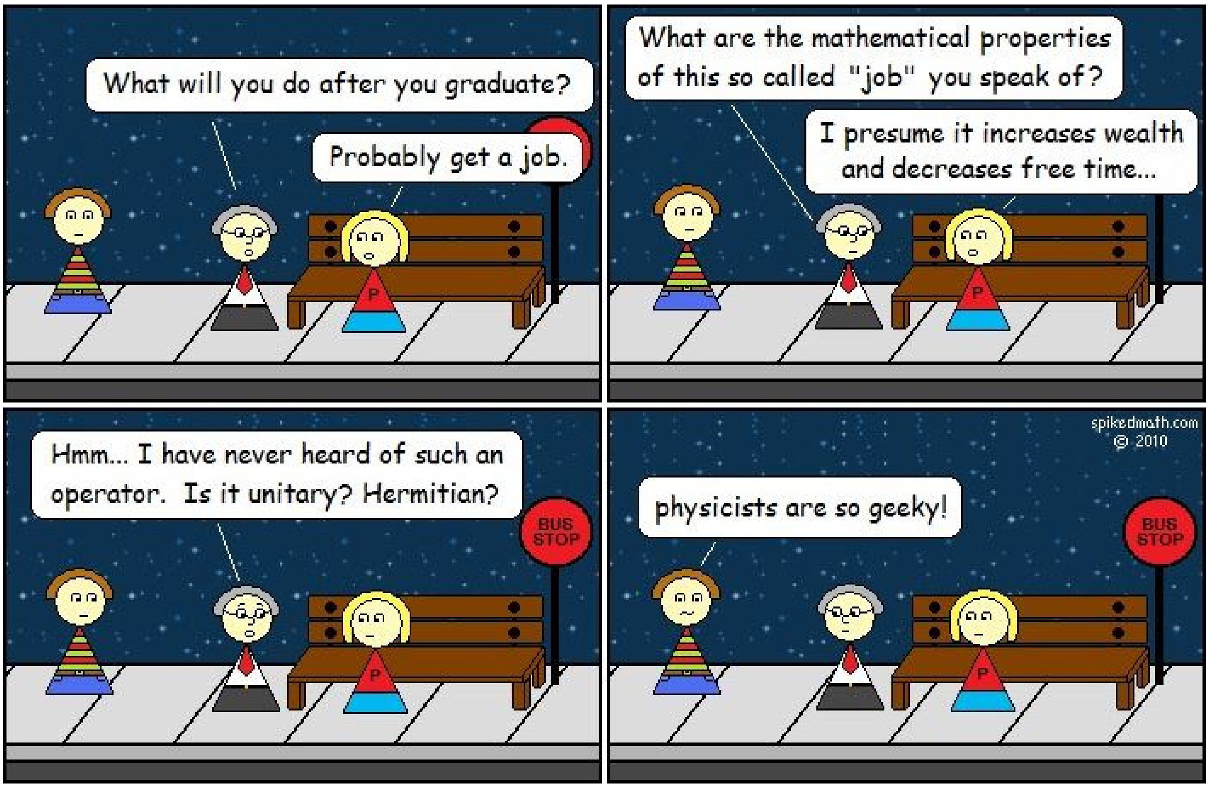
The physicist places both cats on the roof at the same height and lets them both go at the same time.

One of the cats fall off the roof first so obviously there is some difference between the two cats.

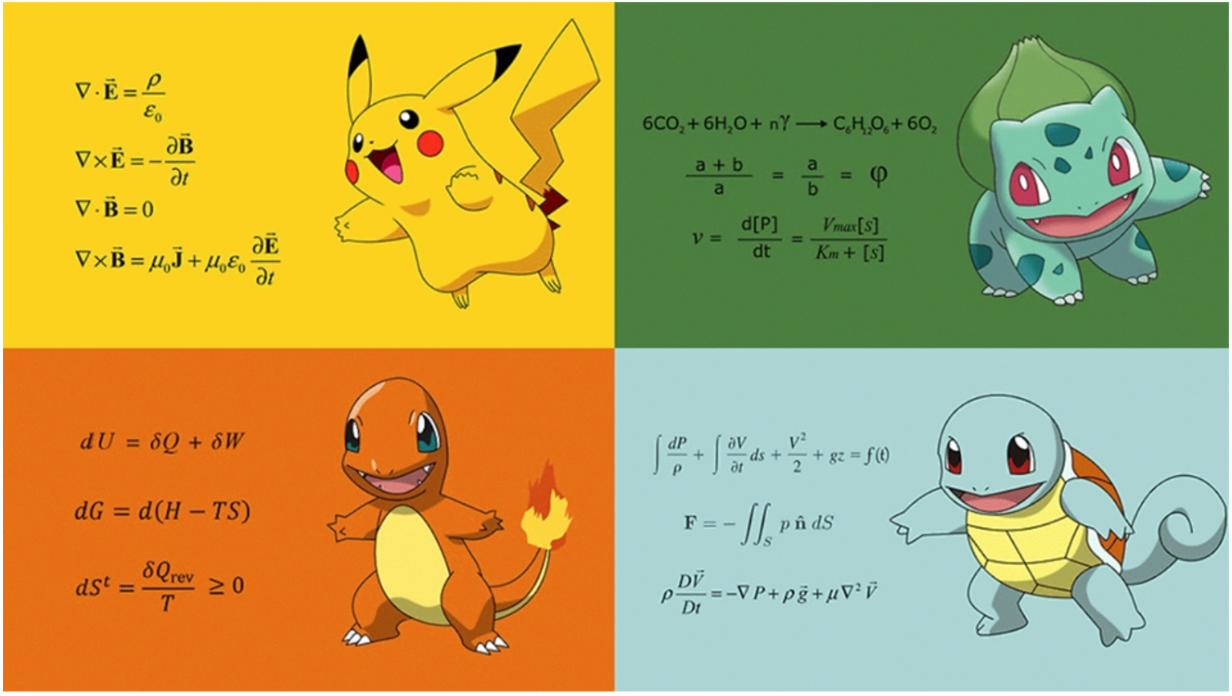
What is the difference?

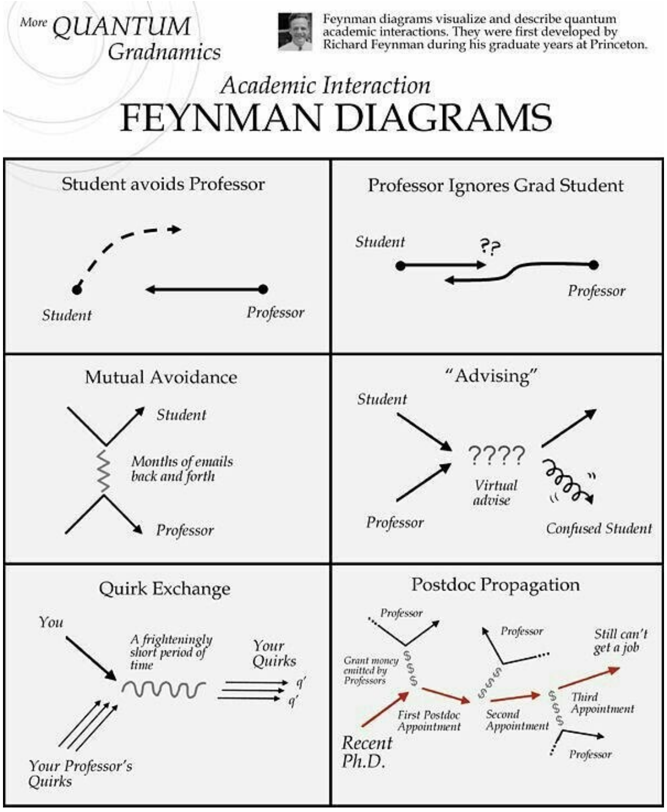
One cat has a greater mew!

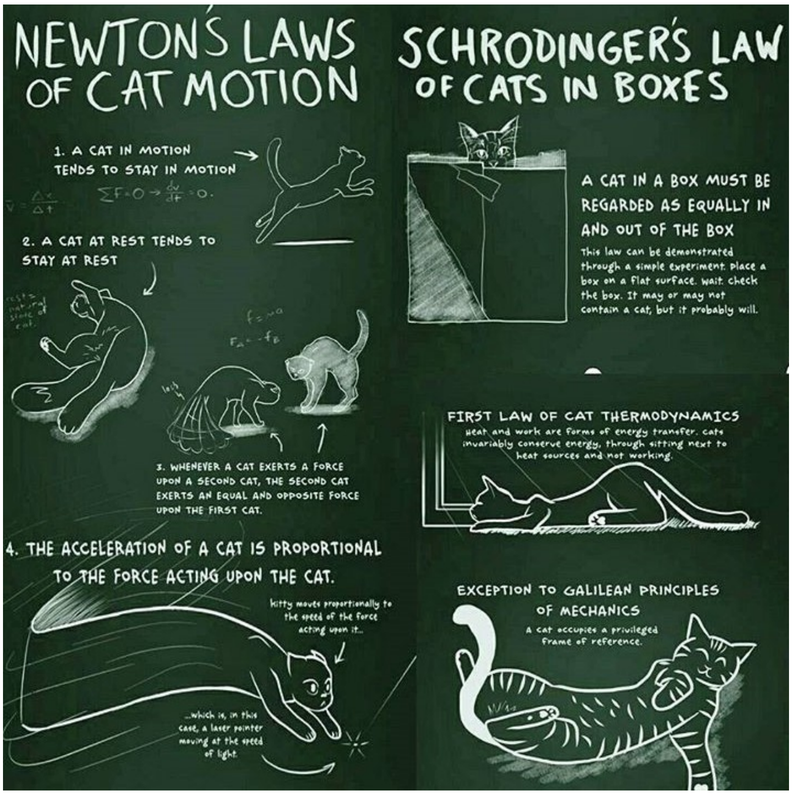

WE HAD A THREESOME LAST NIGHT.
HOW WAS IT?
AWKWARD – IT WAS WITH A PHYSICIST.
WHY'S THAT AWKWARD?
THEY CAN'T SOLVE THE THREE-BODY PROBLEM.
AH, YES.

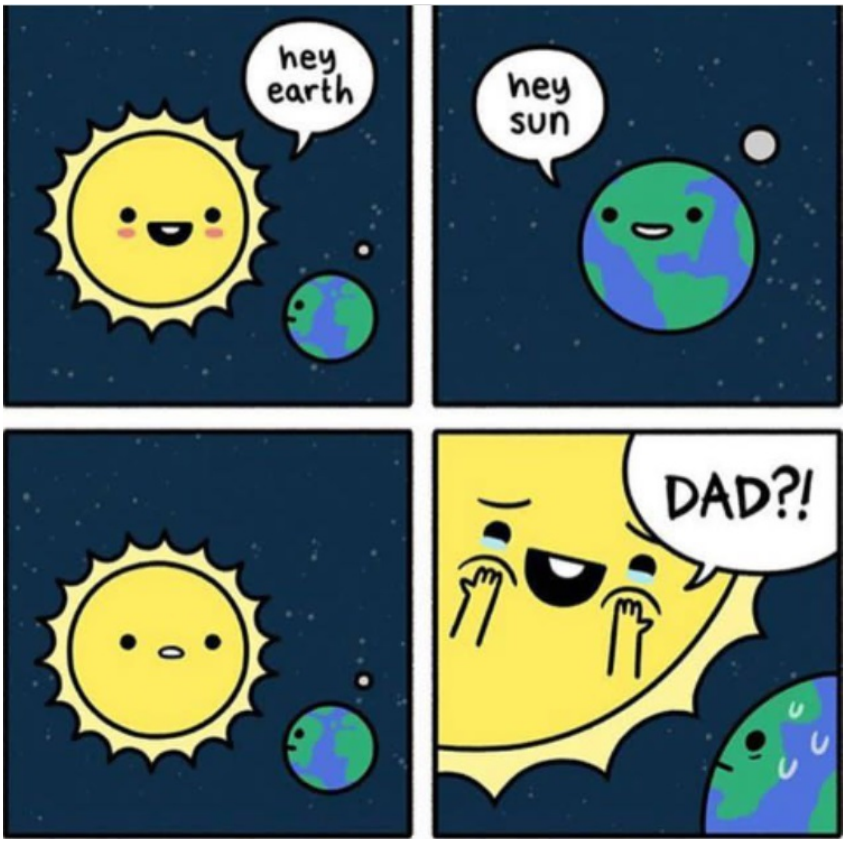

What will you do after you graduate?
Probably get a job.
What are the mathematical properties of this so called "job" you speak of?
I presume it increases wealth and decreases free time...
Hmm... I have never heard of such an operator. Is it unitary? Hermitian?
BUS STOP
spikedmath.com © 2010
physicists are so geeky!
BUS STOP

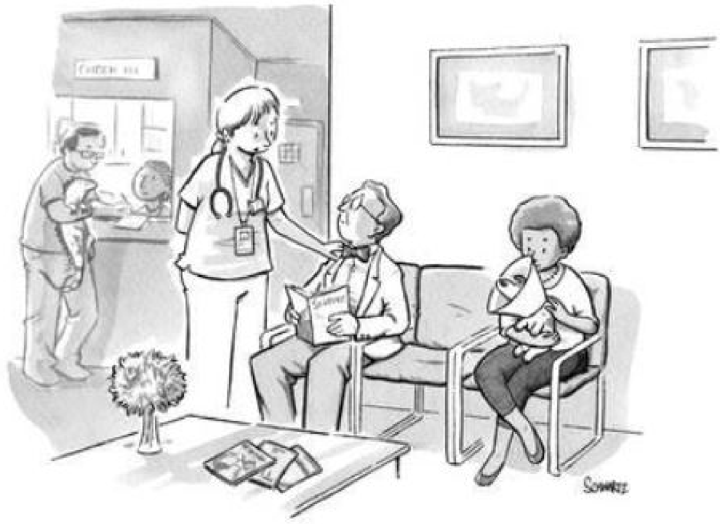
This is how physicists see the Pokemon:


$\nabla\cdot\vec{\mathbf{E}}=\frac{\rho}{\varepsilon_0}$
$\nabla\times\vec{\mathbf{E}}=-\frac{\partial\vec{\mathbf{B}}}{\partial t}$
$\nabla\cdot\vec{\mathbf{B}}=0$
$\nabla\times\vec{\mathbf{B}}=\mu_0\vec{\mathbf{J}}+\mu_0\varepsilon_0\frac{\partial\vec{\mathbf{E}}}{\partial t}$
$6CO_2+6H_2O+n\gamma\longrightarrow C_6H_{12}O_6+6O_2$
$\frac{a+b}{a}=\frac{a}{b}=\varphi$
$v=\frac{d[P]}{dt}=\frac{V_{max}[S]}{K_m+[S]}$
$dU=\delta Q+\delta W$
$dG=d(H-TS)$
$dS^t=\frac{\delta Q_{rev}}{T}\geq 0$
$\int\frac{dP}{\rho}+\int\frac{\partial V}{\partial t}ds+\frac{V^2}{2}+gz=f(t)$
$\mathbf{F}=-\iint_S p\,\hat{\mathbf{n}}\,dS$
$\rho\frac{D\vec{V}}{Dt}=-\nabla P+\rho\vec{g}+\mu\nabla^2\vec{V}$


More QUANTUM Gradnamics
Feynman diagrams visualize and describe quantum academic interactions. They were first developed by Richard Feynman during his graduate years at Princeton.
Academic Interaction
FEYNMAN DIAGRAMS
Student avoids Professor
Student
Professor
Professor Ignores Grad Student
Student
??
Professor
Mutual Avoidance
Student
Months of emails back and forth
Professor
"Advising"
Student
????
Virtual advise
Professor
Confused Student
Quirk Exchange
You
A frighteningly short period of time
Your Quirks
q'
q'
Your Professor's Quirks
Postdoc Propagation
Professor
Grant money emitted by Professors
Professor
Still can't get a job
First Postdoc Appointment
Second Appointment
Third Appointment
Recent Ph.D.
Professor

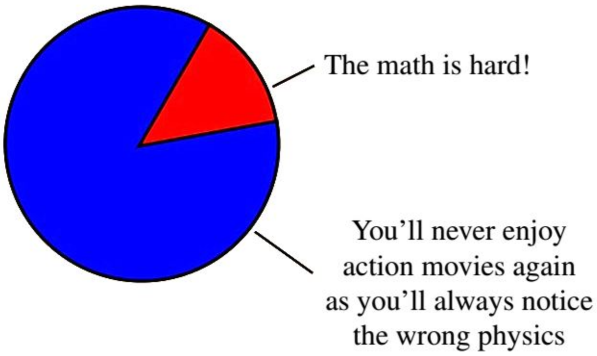

NEWTON'S LAWS OF CAT MOTION
1. A CAT IN MOTION TENDS TO STAY IN MOTION
ΣF=0 → dv/dt = 0.
v = Δx/Δt
2. A CAT AT REST TENDS TO STAY AT REST
rests natural state of cat.
F=ma
F_A=-F_B
3. WHENEVER A CAT EXERTS A FORCE UPON A SECOND CAT, THE SECOND CAT EXERTS AN EQUAL AND OPPOSITE FORCE UPON THE FIRST CAT.
4. THE ACCELERATION OF A CAT IS PROPORTIONAL TO THE FORCE ACTING UPON THE CAT.
kitty moves proportionally to the speed of the force acting upon it...
...which is, in this case, a laser pointer moving at the speed of light.
SCHRODINGER'S LAW OF CATS IN BOXES
A CAT IN A BOX MUST BE REGARDED AS EQUALLY IN AND OUT OF THE BOX
This law can be demonstrated through a simple experiment. Place a box on a flat surface. Wait. Check the box. It may or may not contain a cat, but it probably will.
FIRST LAW OF CAT THERMODYNAMICS
Heat and work are forms of energy transfer. Cats invariably conserve energy, through sitting next to heat sources and not working.
EXCEPTION TO GALILEAN PRINCIPLES OF MECHANICS
A cat occupies a privileged frame of reference.

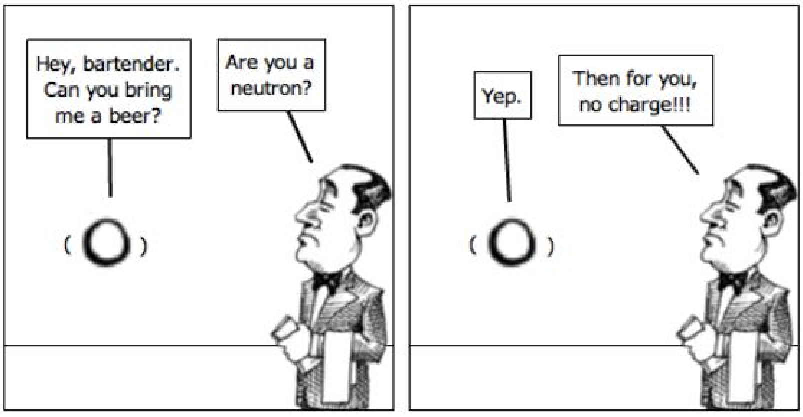

hey earth
hey sun
DAD?!

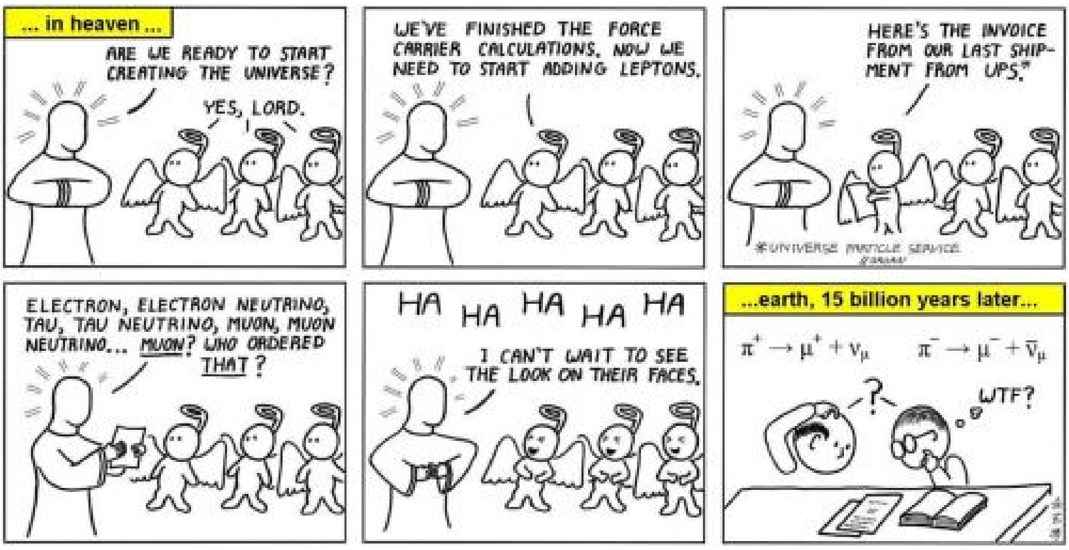


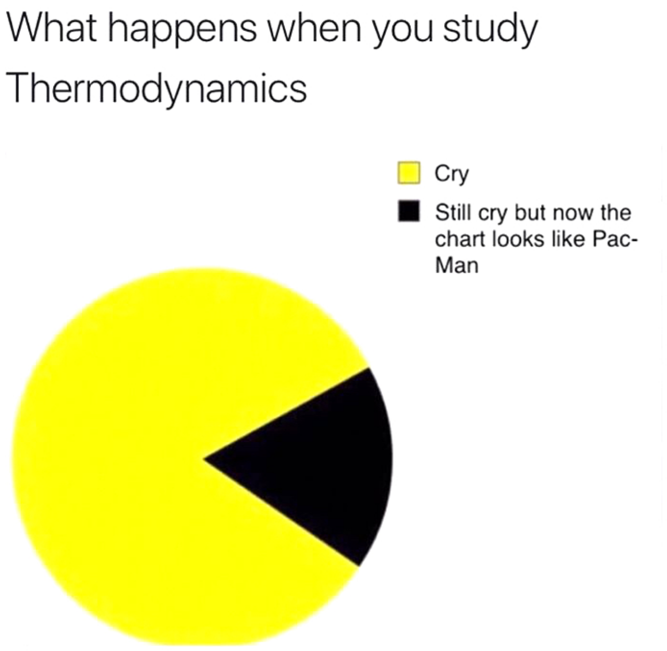
"About your cat, Mr. Schrödinger—I have good news and bad news."

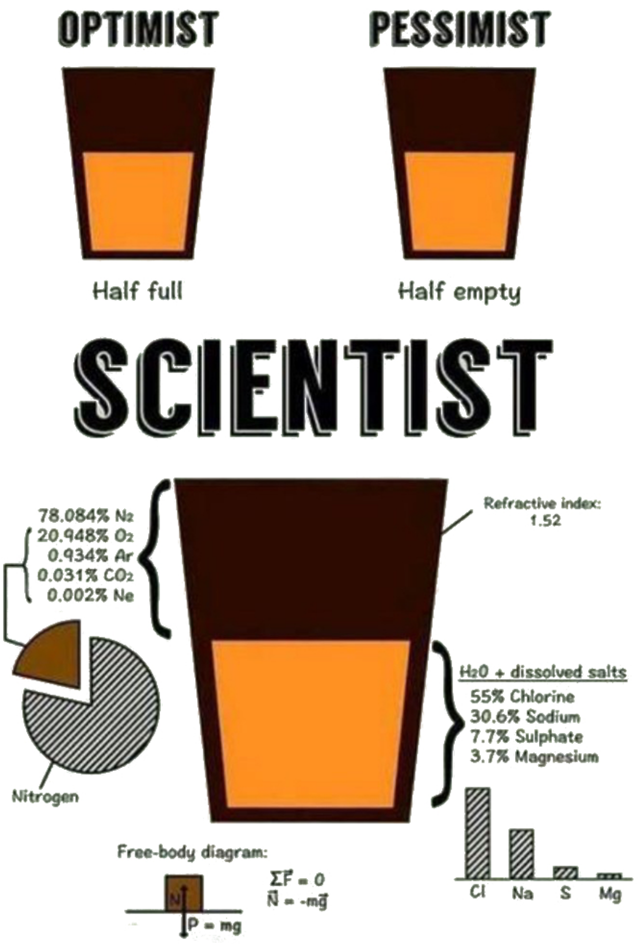
Why is it a bad idea
to major in physics?

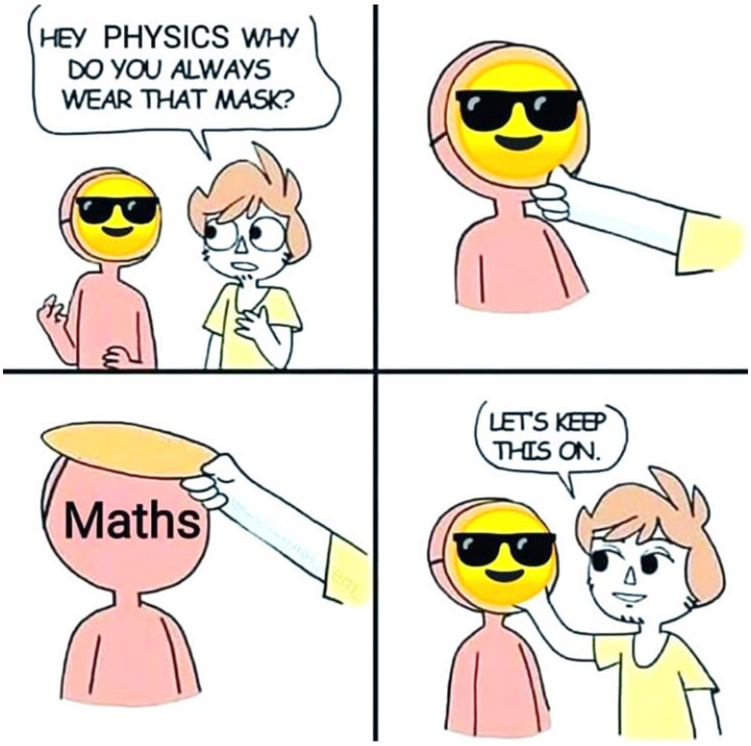


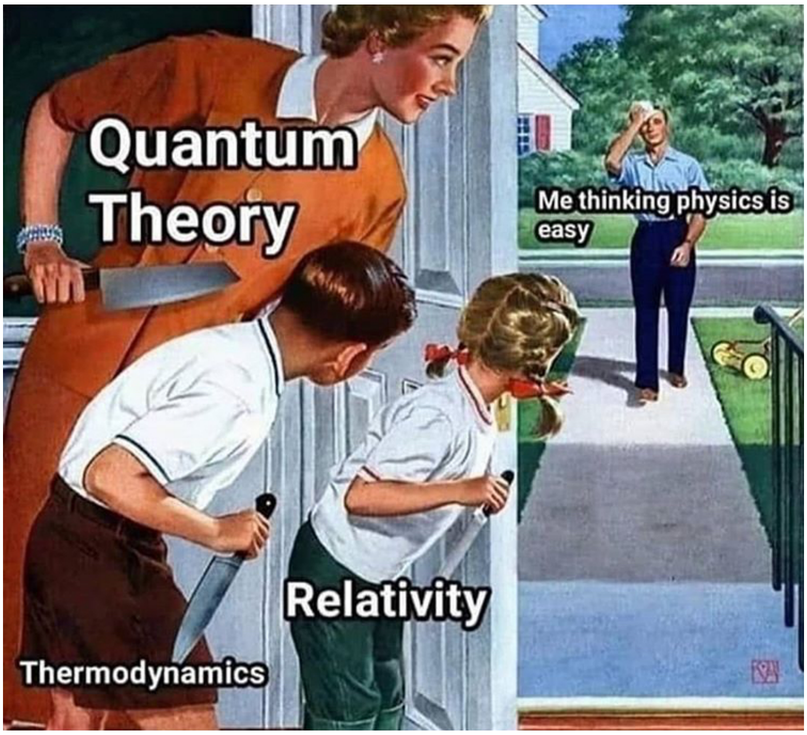

Hey, bartender. Can you bring me a beer?
Are you a neutron?
Yep.
Then for you, no charge!!!

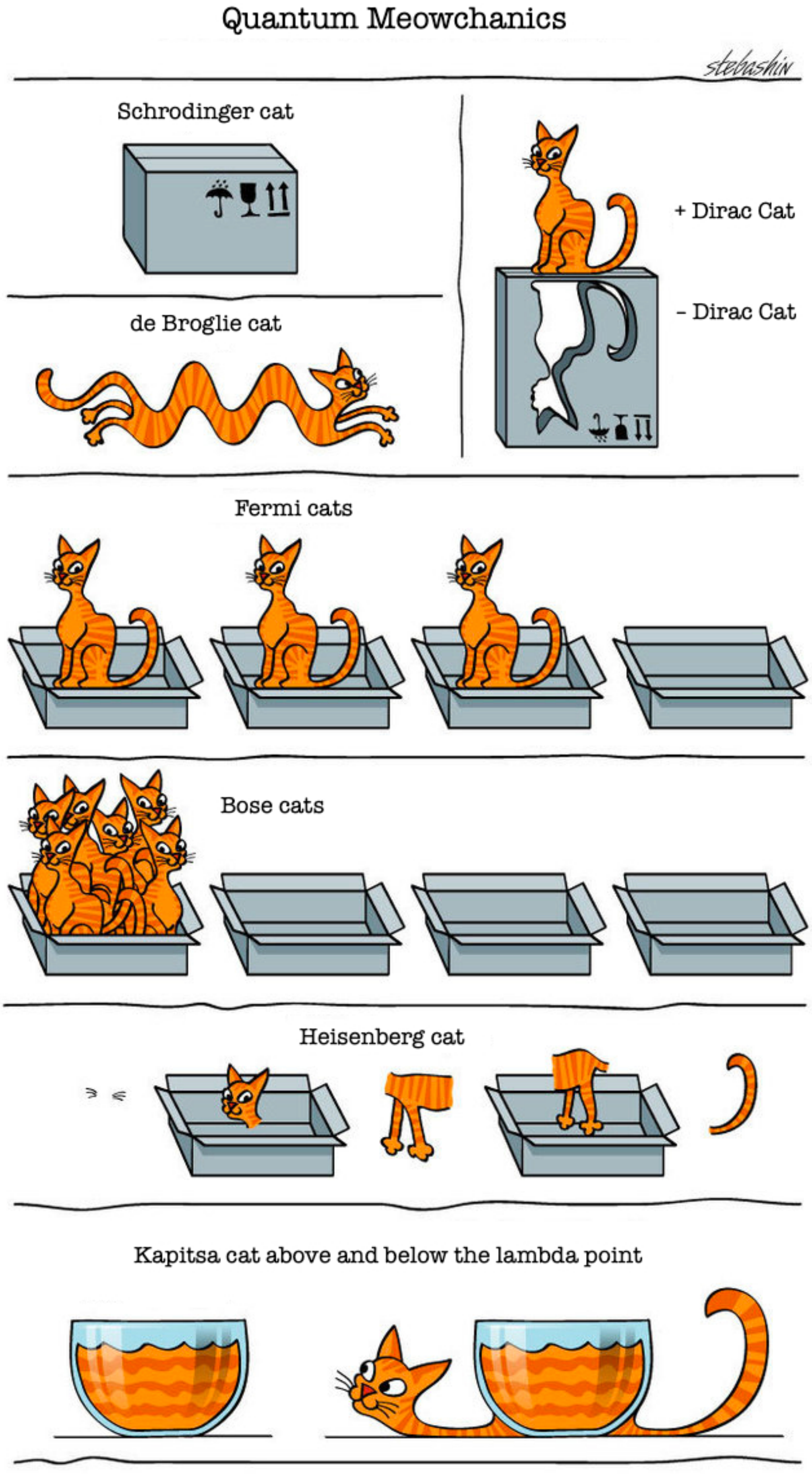

... in heaven ...
ARE WE READY TO START CREATING THE UNIVERSE?
YES, LORD.
WE'VE FINISHED THE FORCE CARRIER CALCULATIONS. NOW WE NEED TO START ADDING LEPTONS.
HERE'S THE INVOICE FROM OUR LAST SHIPMENT FROM UPS.*
*UNIVERSE PARTICLE SERVICE
ELECTRON, ELECTRON NEUTRINO, TAU, TAU NEUTRINO, MUON, MUON NEUTRINO... MUON? WHO ORDERED THAT?
HA HA HA HA HA
I CAN'T WAIT TO SEE THE LOOK ON THEIR FACES.
...earth, 15 billion years later...
$\pi^+ \rightarrow \mu^+ + \nu_\mu$
$\pi^- \rightarrow \mu^- + \bar{\nu}_\mu$
?
WTF?

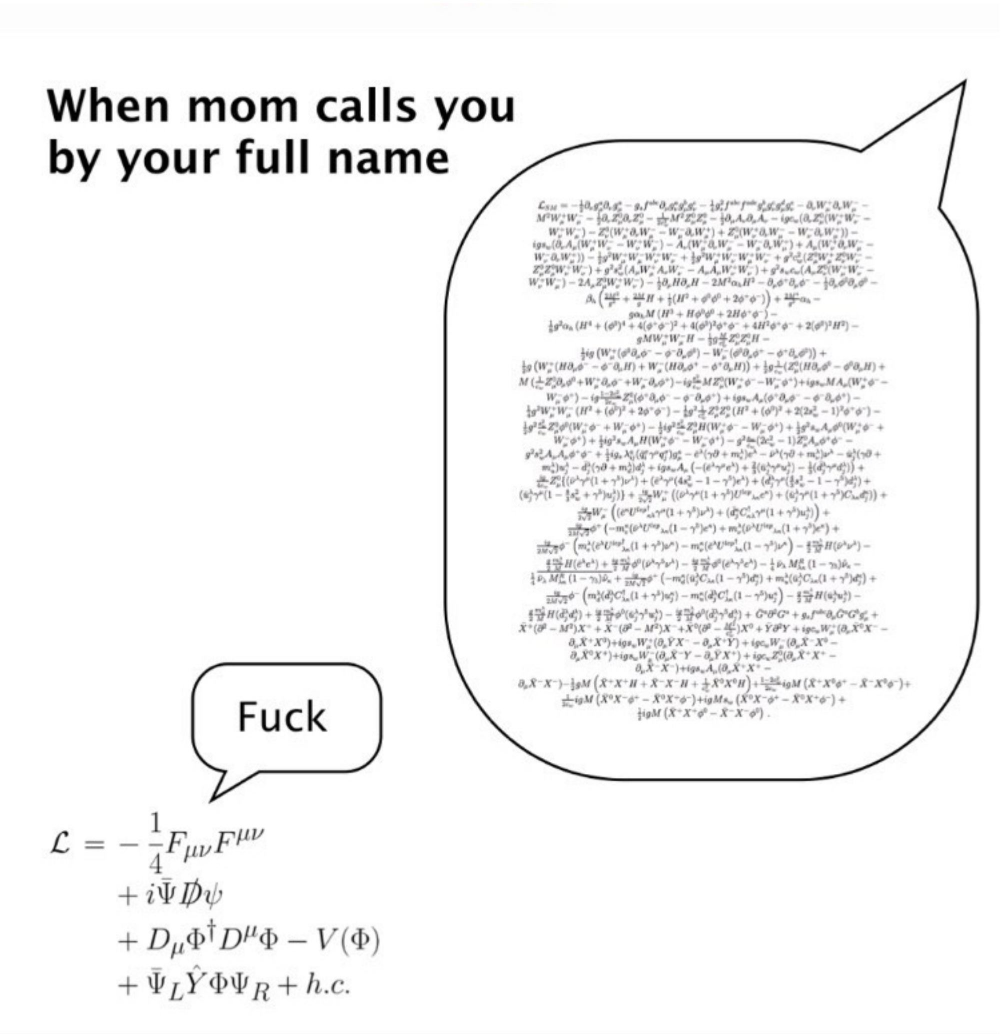

What happens when you study Thermodynamics
Cry
Still cry but now the chart looks like Pac-Man


OPTIMIST
PESSIMIST
Half full
Half empty
SCIENTIST
78.084% N_2
20.948% O_2
0.934% Ar
0.031% CO_2
0.002% Ne
Nitrogen
Refractive index: 1.52
H_2O + dissolved salts
55% Chlorine
30.6% Sodium
7.7% Sulphate
3.7% Magnesium
Cl
Na
S
Mg
Free-body diagram:
$\Sigma\vec{F} = 0$
$\vec{N} = -m\vec{g}$
N
P = mg


HEY PHYSICS WHY DO YOU ALWAYS WEAR THAT MASK?
Maths
LET'S KEEP THIS ON.


Quantum Theory
Me thinking physics is easy
Relativity
Thermodynamics


Quantum Meowchanics
stebashin
Schrodinger cat
+ Dirac Cat
- Dirac Cat
de Broglie cat
Fermi cats
Bose cats
Heisenberg cat
Kapitsa cat above and below the lambda point


When mom calls you by your full name
Fuck
$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + i\bar{\Psi}\not{D}\psi + D_\mu\Phi^\dagger D^\mu\Phi - V(\Phi) + \bar{\Psi}_L\hat{Y}\Phi\Psi_R + h.c.$

20.4. Statistics

Three statisticians go out for a static target shooting. The first statistician fired and shoots to the left, the second shot, but symmetrically on the right. The last does not shoot, but say triumphantly: "On average we got it!"

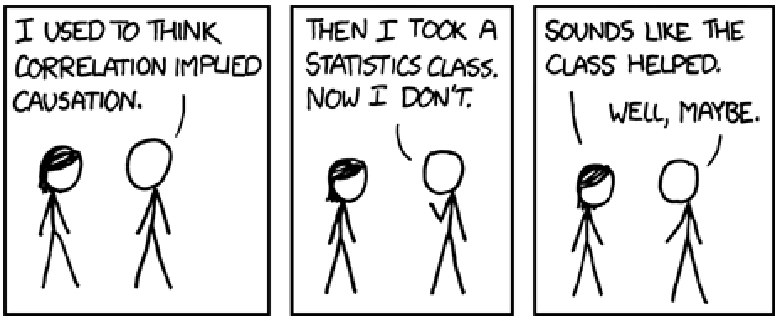

I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

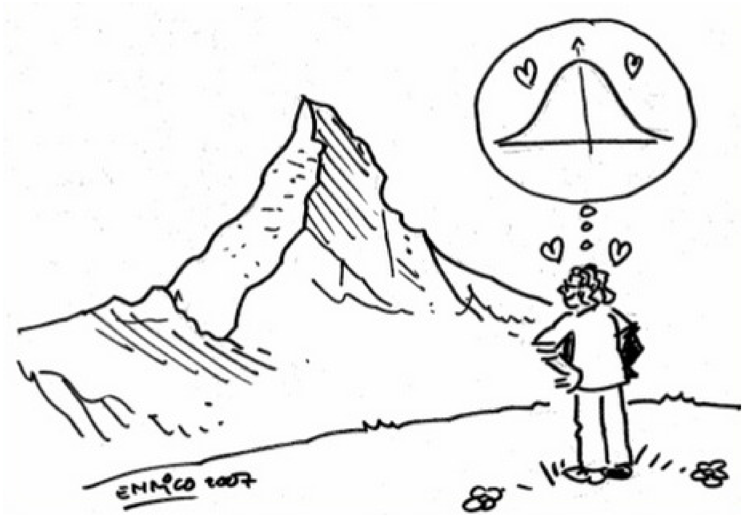
Figure 20.2: **source: www.xkcd.com (author: Randall Munroe)**

Patient: "Will I survive to this delicate operation?"

Surgeon: "Yes, I am absolutely sure that you will survive."

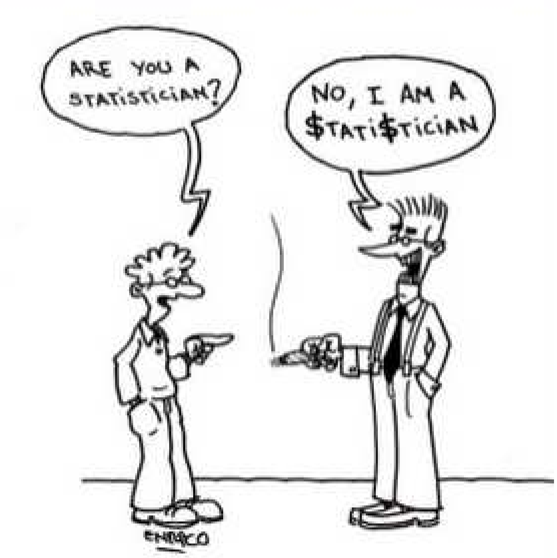
Patient: "How can you be so sure?"

Surgeon: "9 out of 10 patients die during this operation and my ninth patient died yesterday."


ENRICO 2007

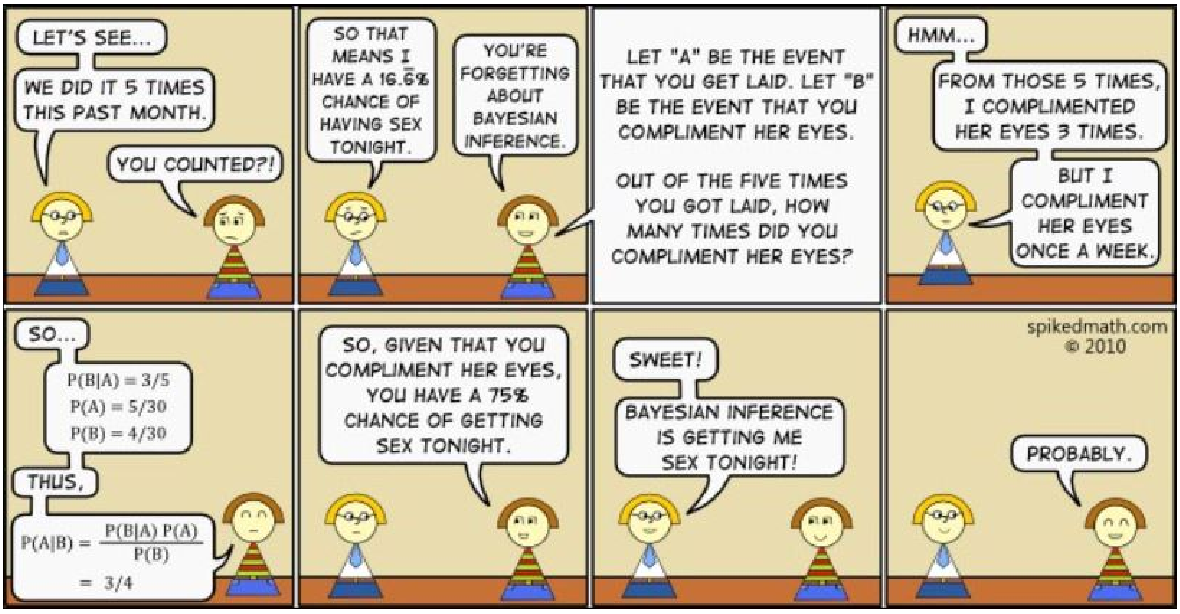
10 reasons to work in statistics field:

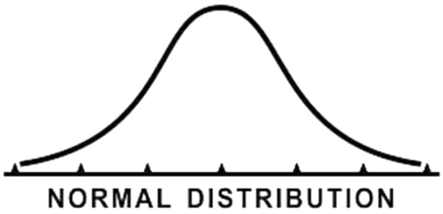
1. Estimate parameters is easier than fighting in real life
2. Statisticians are recognized people
3. You will learn the Greek alphabet entirely
4. The probability that you get a job in this area is > 0.9999
5. If you are fired, you can always convert yourself to engineering
6. You make this work in the confidence, with regularity and variability
7. You are normal and the rest of the world is wrong
8. The regression line seems better than the unemployment line
9. You never have to be exact - only approximate
10. Nobody understands what you do, then you are always right

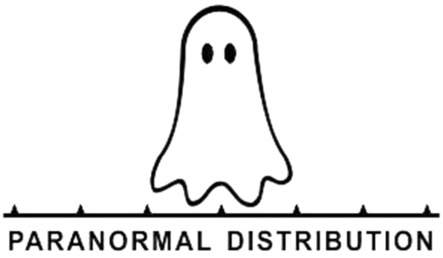


A theist asks an atheist what is the most important value in life?

The atheist answers: the p-value.


LET'S SEE...
WE DID IT 5 TIMES THIS PAST MONTH.
YOU COUNTED?!
SO THAT MEANS I HAVE A $16.\overline{6}\%$ CHANCE OF HAVING SEX TONIGHT.
YOU'RE FORGETTING ABOUT BAYESIAN INFERENCE.
LET "A" BE THE EVENT THAT YOU GET LAID. LET "B" BE THE EVENT THAT YOU COMPLIMENT HER EYES.
OUT OF THE FIVE TIMES YOU GOT LAID, HOW MANY TIMES DID YOU COMPLIMENT HER EYES?
HMM...
FROM THOSE 5 TIMES, I COMPLIMENTED HER EYES 3 TIMES.
BUT I COMPLIMENT HER EYES ONCE A WEEK.
SO...
$P(B|A) = 3/5$
$P(A) = 5/30$
$P(B) = 4/30$
THUS,
$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} = 3/4$
SO, GIVEN THAT YOU COMPLIMENT HER EYES, YOU HAVE A 75% CHANCE OF GETTING SEX TONIGHT.
SWEET!
BAYESIAN INFERENCE IS GETTING ME SEX TONIGHT!
spikedmath.com © 2010
PROBABLY.


NORMAL DISTRIBUTION

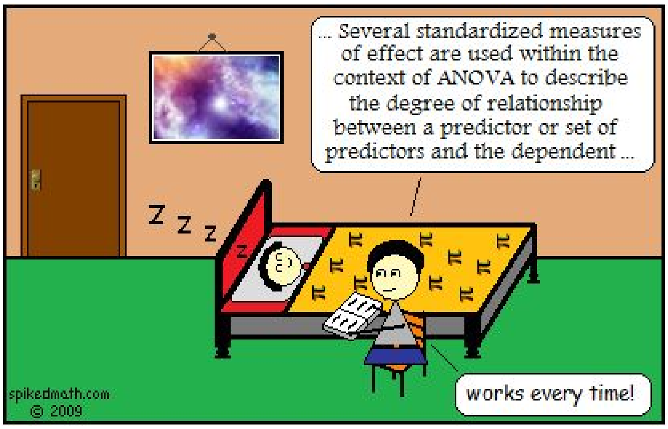

PARANORMAL DISTRIBUTION

A statistician and a biologist are sentenced to death. They are given a last favour.

- I would like to give a big conference on statistics in front of everyone, says the statistician
- Granted! Answers the judge. And for you?

The biologist does not express any hesitation:

- I would like to be executed first!


... Several standardized measures of effect are used within the context of ANOVA to describe the degree of relationship between a predictor or set of predictors and the dependent ...
Z z z
works every time!
spikedmath.com
© 2009

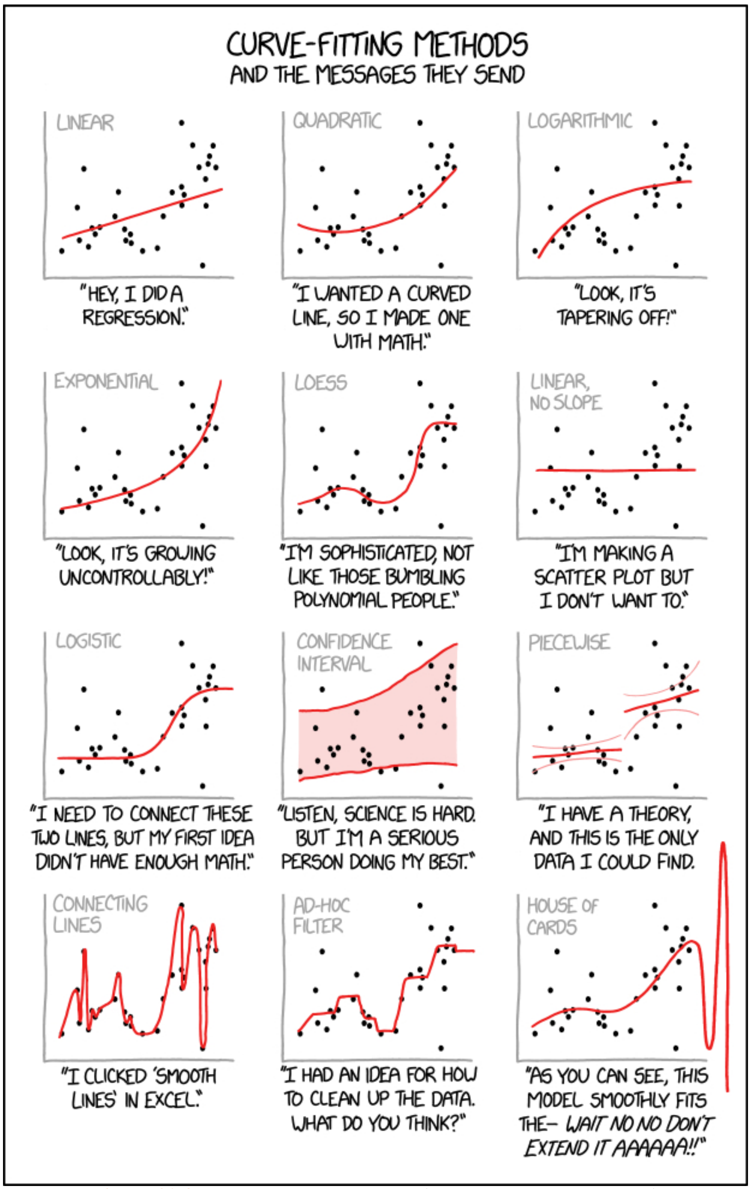
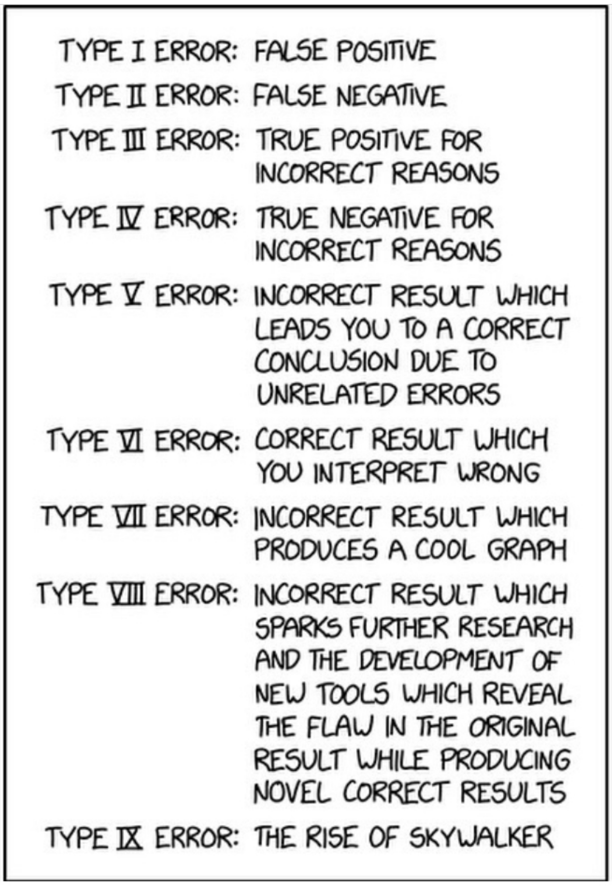

CURVE-FITTING METHODS
AND THE MESSAGES THEY SEND
LINEAR
"HEY, I DID A REGRESSION."
QUADRATIC
"I WANTED A CURVED LINE, SO I MADE ONE WITH MATH."
LOGARITHMIC
"LOOK, IT'S TAPERING OFF!"
EXPONENTIAL
"LOOK, IT'S GROWING UNCONTROLLABLY!"
LOESS
"I'M SOPHISTICATED, NOT LIKE THOSE BUMBLING POLYNOMIAL PEOPLE."
LINEAR, NO SLOPE
"I'M MAKING A SCATTER PLOT BUT I DON'T WANT TO."
LOGISTIC
"I NEED TO CONNECT THESE TWO LINES, BUT MY FIRST IDEA DIDN'T HAVE ENOUGH MATH."
CONFIDENCE INTERVAL
"LISTEN, SCIENCE IS HARD. BUT I'M A SERIOUS PERSON DOING MY BEST."
PIECEWISE
"I HAVE A THEORY, AND THIS IS THE ONLY DATA I COULD FIND."
CONNECTING LINES
"I CLICKED 'SMOOTH LINES' IN EXCEL."
AD-HOC FILTER
"I HAD AN IDEA FOR HOW TO CLEAN UP THE DATA. WHAT DO YOU THINK?"
HOUSE OF CARDS
"AS YOU CAN SEE, THIS MODEL SMOOTHLY FITS THE- WAIT NO NO DON'T EXTEND IT AAAAAA!!"

Figure 20.3: source: www.xkcd.com (author: Randall Munroe)



TYPE I ERROR: FALSE POSITIVE

TYPE II ERROR: FALSE NEGATIVE

TYPE III ERROR: TRUE POSITIVE FOR INCORRECT REASONS

TYPE IV ERROR: TRUE NEGATIVE FOR INCORRECT REASONS

TYPE V ERROR: INCORRECT RESULT WHICH LEADS YOU TO A CORRECT CONCLUSION DUE TO UNRELATED ERRORS

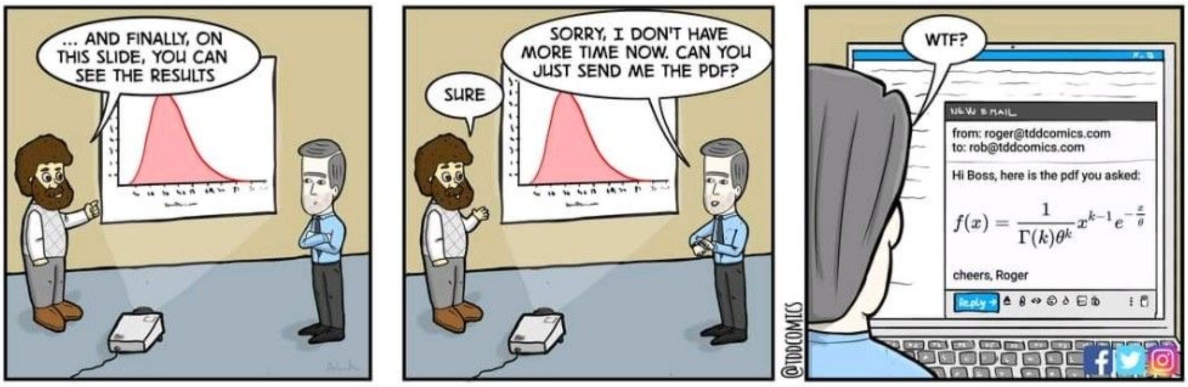
TYPE VI ERROR: CORRECT RESULT WHICH YOU INTERPRET WRONG

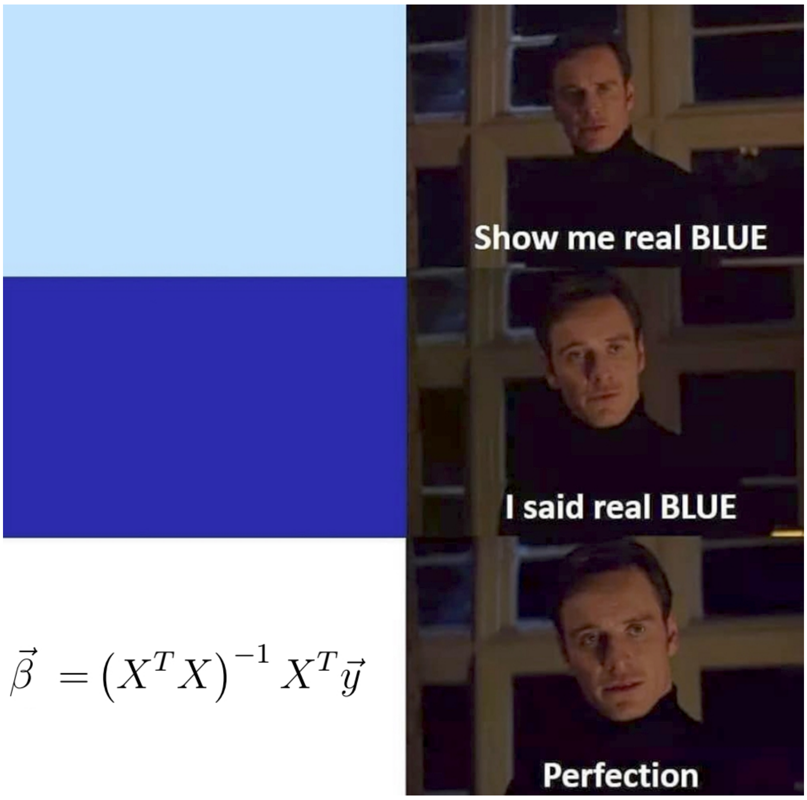
TYPE VII ERROR: INCORRECT RESULT WHICH PRODUCES A COOL GRAPH

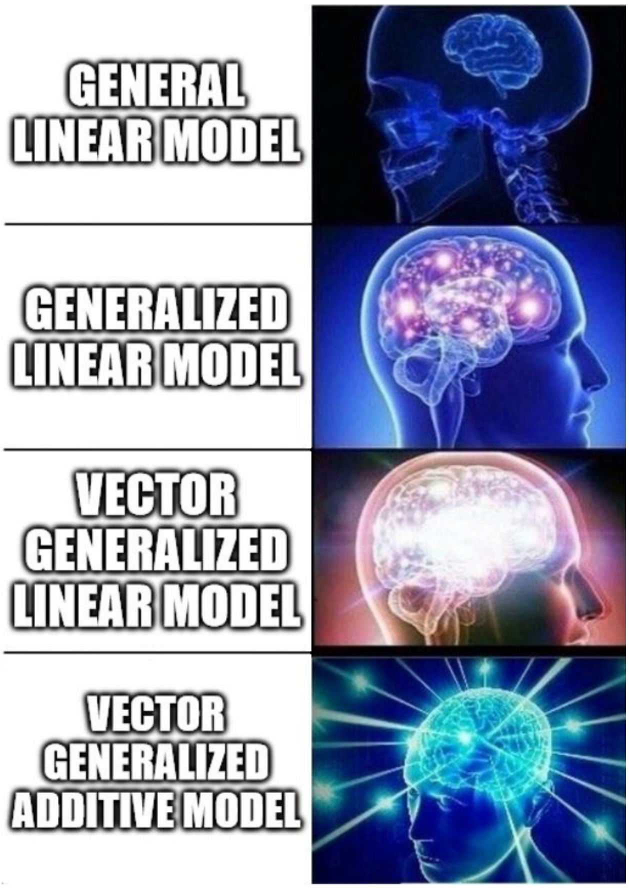
TYPE VIII ERROR: INCORRECT RESULT WHICH SPARKS FURTHER RESEARCH AND THE DEVELOPMENT OF NEW TOOLS WHICH REVEAL THE FLAW IN THE ORIGINAL RESULT WHILE PRODUCING NOVEL CORRECT RESULTS

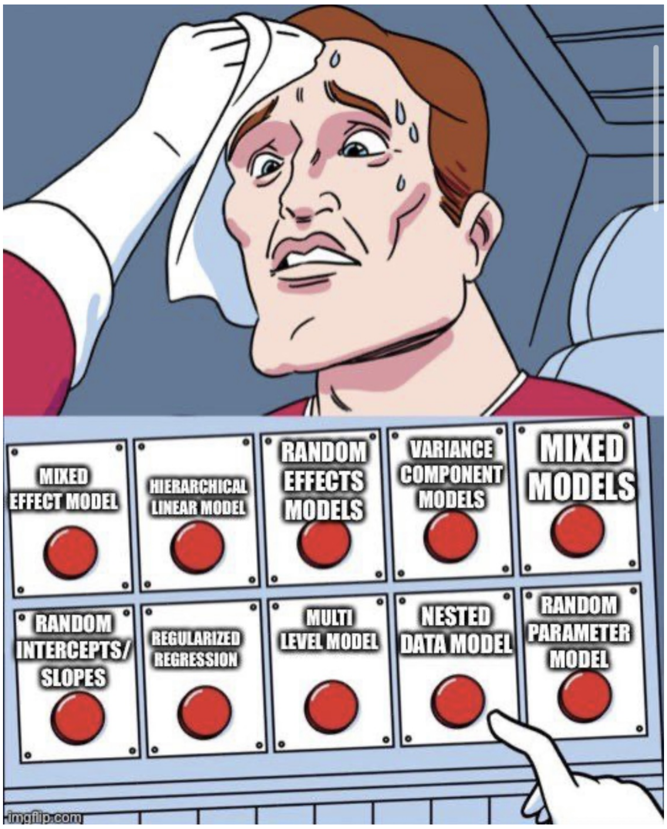
TYPE IX ERROR: THE RISE OF SKYWALKER

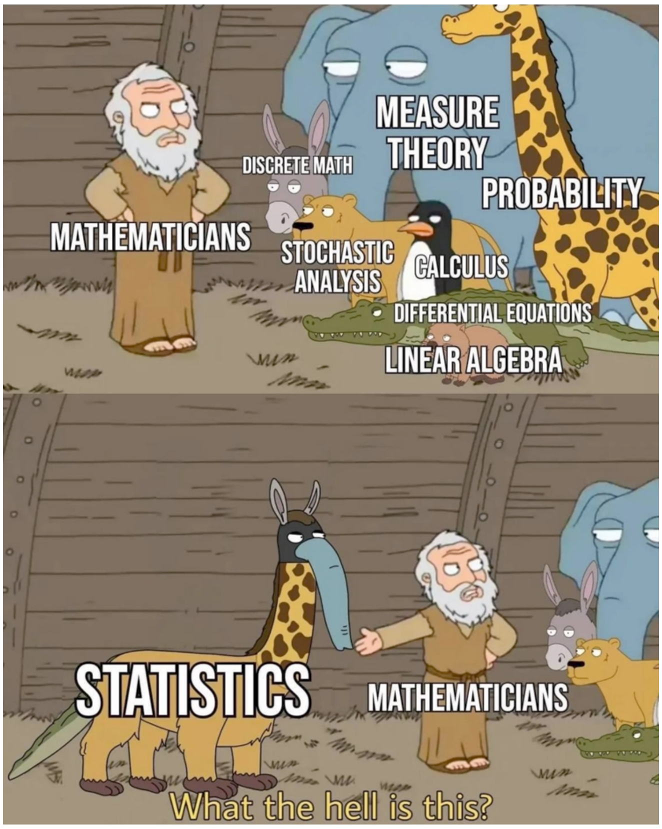
Figure 20.4: source: www.xkcd.com (author: Randall Munroe)


... AND FINALLY, ON THIS SLIDE, YOU CAN SEE THE RESULTS
SURE
SORRY, I DON'T HAVE MORE TIME NOW. CAN YOU JUST SEND ME THE PDF?
WTF?
NEW E-MAIL
from: roger@tddcomics.com
to: rob@tddcomics.com
Hi Boss, here is the pdf you asked:
$f(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$
cheers, Roger
Reply
@TDDCOMICS


Show me real BLUE
I said real BLUE
$\vec{\beta} = (X^T X)^{-1} X^T \vec{y}$
Perfection


GENERAL LINEAR MODEL
GENERALIZED LINEAR MODEL
VECTOR GENERALIZED LINEAR MODEL
VECTOR GENERALIZED ADDITIVE MODEL


MIXED EFFECT MODEL
HIERARCHICAL LINEAR MODEL
RANDOM EFFECTS MODELS
VARIANCE COMPONENT MODELS
MIXED MODELS
RANDOM INTERCEPTS/ SLOPES
REGULARIZED REGRESSION
MULTI LEVEL MODEL
NESTED DATA MODEL
RANDOM PARAMETER MODEL
imgflip.com


MEASURE THEORY
DISCRETE MATH
PROBABILITY
MATHEMATICIANS
STOCHASTIC ANALYSIS
CALCULUS
DIFFERENTIAL EQUATIONS
LINEAR ALGEBRA
STATISTICS
MATHEMATICIANS
What the hell is this?

20.5. Chemistry

We've just discovered a new element:

- ELEMENT NUMBER: 115
- NAME: Woman
- SYMBOL: Wo
- ATOMIC MASS: Accepted as 60 kg; isotopes may vary from 40-200 kg.
- OCCURRENCE: Copious quantities in all urban areas
- PHYSICAL PROPERTIES:

  - Boils at room temperature

  - Freezes without any known reason

  - Melts if given special treatment

  - Bitter, if incorrectly used

  - Sweet as Honey if given a proper treatment.
- MOLECULAR STRUCTURE:

  Perfect? 90/60/90, growing int the U.S. with 60/90/120 and in the nordic countries as so-called "flat" 50/50/50
- CHEMICAL PROPERTIES:

  - Has great affinity for Gold, Silver and a range of precious stones.Absorbs in general great quantities of expensive substances.

  - May explode spontaneously without prior warning and for no known reason.

  - Insoluble in liquids but activity greatly increased by saturation in alcohol

  - Reactivity varies depending on the time of the day

  - Great ability to change mood and jealousy

  - Sensitive to certain constraints which sometimes transmit migraine
- COMMON USES:

  - Highly ornamental, especially in sports cars

  - Powerful cleaning agent

  - Can be great aid to relaxation.
- TEST:

  - Pure specimen turns rosy pink when happy

  - Turns green when placed behind a better specimen

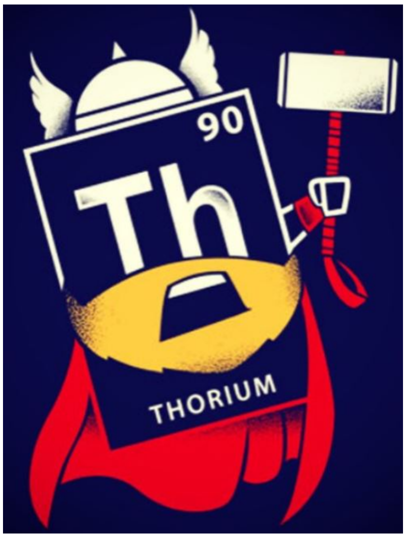
- PRECAUTIONS:

  - Highly dangerous if placed between the non-expert hands

  - Illegal to possess more than one, although several can be maintained at different locations as long as specimens do not come in direct contact with each other.

CAUTION:

Some South American researchers have discovered a way to produce them artificially, usually presented under the marks "transvestite" or "Drag-queen". Consume only the generic!



We’ve just discovered a new element:

- ELEMENT NUMBER: 116
- NAME: Man
- SYMBOL: Hm
- QUANTITATIVE ANALYSIS:

  Measured at 17 cm, although some isotopes exist in 25, 20, 13 and even 10 cm
- EXTRACTION LOCATION:

  Can be found in large quantities in the presence of a deposit of very pure Fm

- PHYSICAL PROPERTIES:

  - Surface covered with hair, steep in places, soft in others

  - Boil when it is shaken, will ice when placed in the presence of logic and common sense, liquefies when treated like a god

  - Becomes execrable when mixed with any alcohol

  - Can cause headaches (other body parts pains); handle with care

  - Decreases its entropy directly after its reaction with the element Fm (condition manifested by snoring ... zzzzz)

  - Its mass increases significantly with age, losing its reaction capacity

  - Dehydrates quickly in dry weather.

  - Rarely found in pure form after 14 years old

  - Often has an inexplicable attachment to its mother rock, making extraction difficult

  - If you put it under pressure it becomes too hard and unproductive; is productive only when used subtlety, with subterfuge and flattery

- CHEMICAL PROPERTIES:

  - Very strong tendency to react with the Fm element, although the reaction is sometimes endothermic

  - Deemed to be the best catalyst for the transformation reactions of the Fm element

  - Has the ability to react with almost anything

  - If the case of an important reaction the aspect of the element changes to dark red

  - If it is saturated with alcohol, it becomes inert and repulsive for most elements

  - Not suitable for household chores and cleaning operations

  - Not suitable either for family duties

  - Is neutral with respect to the courtesy and impartiality

- COMMON USES:

  - Transporting heavy things, driver, free dinners at the restaurant...

  - Possible use for sexual activity

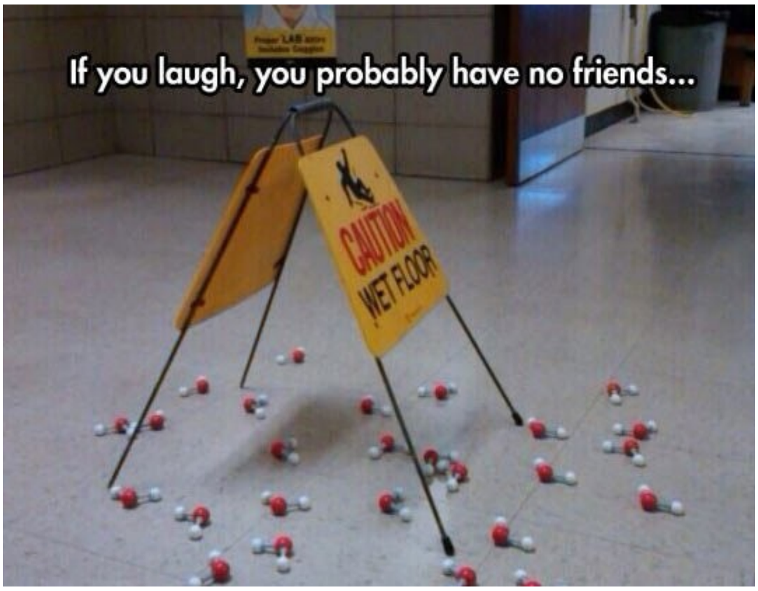
- TESTS:

  The purest specimens are not synonymous with purity, and those who have already served, are less pure

- HAZARDS:

  The reaction with another element Hm is extremely violent if the item Fm is the catalyst

---



The following is an actual question given on a University of Washington engineering mid term. The answer was so profound that the Professor shared it with colleagues, which is why we now have the pleasure of enjoying it as well.

Bonus Question: Is Hell exothermic (gives off heat) or Endothermic (absorbs heat)?

Most of the students wrote proofs of their beliefs using Boyle's Law, (gas cools off when it expands and heats up when it is compressed) or some variant. One student, however, wrote the following:

"First, we need to know how the mass of Hell is changing in time. So we need to know the rate that souls are moving into Hell and the rate they are leaving. I think that we can safely assume that once a soul gets to Hell, it will not leave. Therefore, no souls are leaving. As for how many souls are entering Hell, let's look at the different religions that exist in the world today. Some of these religions state that if you are not a member of their religion, you will go to Hell. Since there are more than one of these religions and since people do not belong to more than one religion, we can project that all souls go to Hell. With birth and death rates as they are, we can expect the number of souls in Hell to increase exponentially.

Now, we look at the rate of change of the volume in Hell because Boyle's Law states that in order for the temperature and pressure in Hell to stay the same, the volume of Hell has to expand as souls are added. This gives two possibilities:

1. If Hell is expanding at a slower rate than the rate at which souls enter Hell, then the temperature and pressure in Hell will increase until all Hell breaks loose.
2. Of course, if Hell is expanding at a rate faster than the increase of souls in Hell, then the temperature and pressure will drop until Hell freezes over.

So which is it? If we accept the postulate given to me by Teresa Banyan during my Freshman year, "...that it will be a cold day in Hell before I sleep with you." and take into account the fact that I still have not succeeded in having sexual relations with her, then, #2 cannot be true, and thus I am sure that Hell is exothermic and will not freeze."

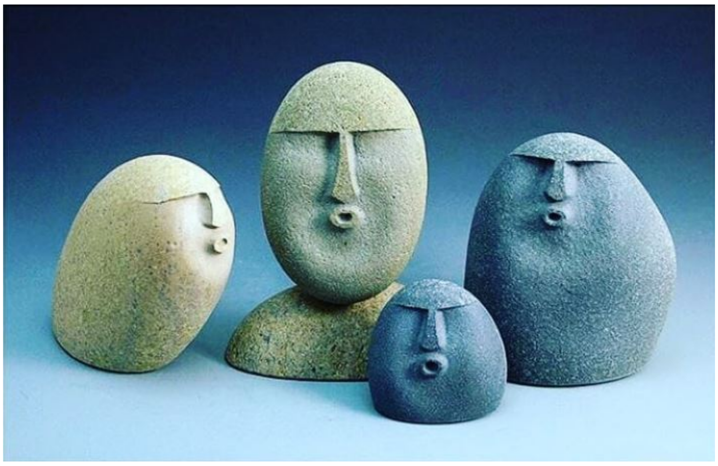
This student received the only A.

A chemist walks into a pharmacy and asks the pharmacist: "Do you have any acetylsalicylic acid?"

- "You mean aspirin?" asked the pharmacist.
- "That's it, I can never remember that word."

Teacher asks: what's the element symbol of Oxygen.

Students:



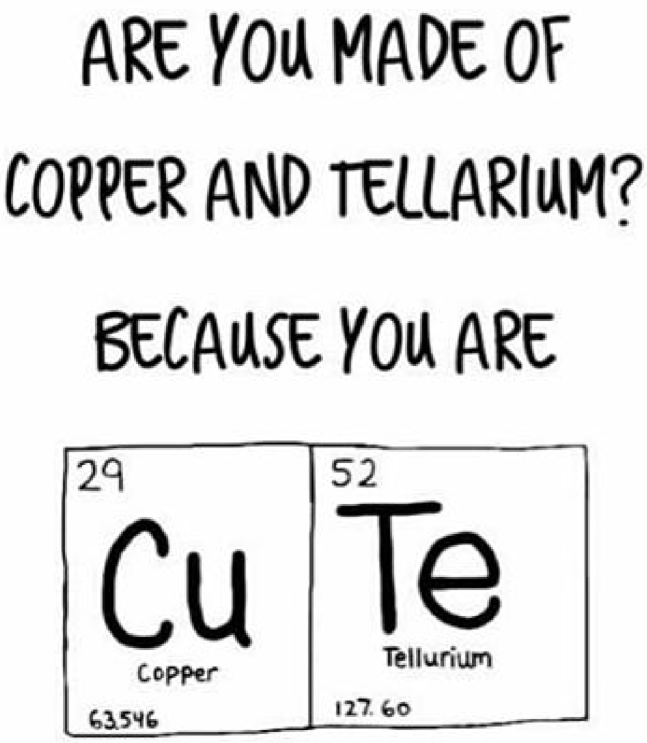
A physicist, biologist and a chemist were going to the ocean for the first time.

- The physicist saw the ocean and was fascinated by the waves. He said he wanted to do some research on the fluid dynamics of the waves and walked into the ocean. Obviously he was drowned and never returned.
- The biologist said he wanted to do research on the flora and fauna inside the ocean and walked inside the ocean. He too, never returned.

- The chemist waited for a long time and afterwards, wrote the observation: "The physicist and the biologist are soluble in ocean water".

CLASSIFICATION OF CHEMISTRY

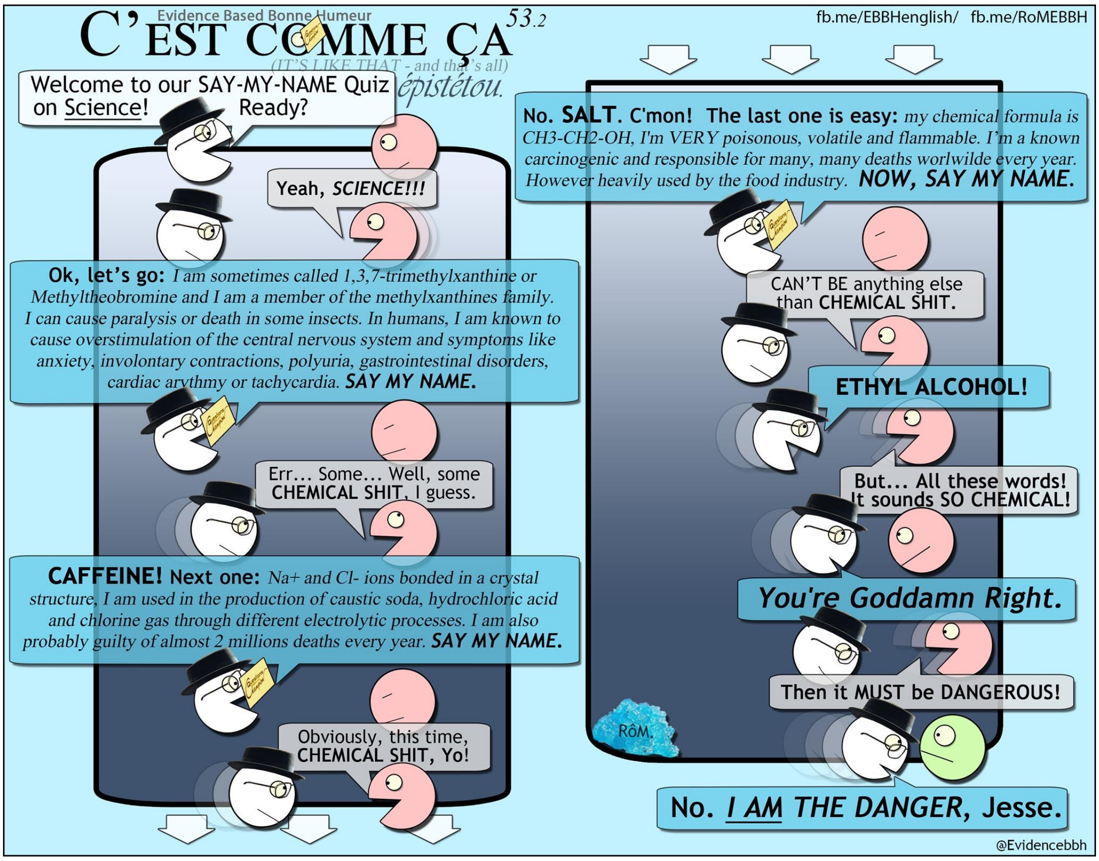
- *Physical Chemistry*: The pitiful attempt to apply $y = mx + b$ to everything in the universe.
- *Organic Chemistry*: The practice of transmuting vile substances into publications.
- *Inorganic Chemistry*: That which is left over after the organic, analytical, and physical chemists get through picking over the periodic table.
- *Chemical Engineering*: The practice of doing for a profit what an organic chemist only does for fun.


ARE YOU MADE OF
COPPER AND TELLARIUM?
BECAUSE YOU ARE
29
Cu
Copper
63.546
52
Te
Tellurium
127.60

Free radicals have revolutionized chemistry.

Chemist's last words:

- And now the tasting test...
- And now shake it a bit...
- In which glass was my mineral water?
- Why does that stuff burn with a green flame?!?
- And now the detonating gas problem.
- This is a completely safe experimental setup.
- Now you can take the protection window away...
- Where do all those holes in my kettle come from?
- And now a cigarette...


Evidence Based Bonne Humeur
C'EST COMME ÇA 53.2
(IT'S LIKE THAT - and that's all)
épistétou.
fb.me/EBBHenglish/ fb.me/RoMEBBH
Welcome to our SAY-MY-NAME Quiz on Science! Ready?
Yeah, SCIENCE!!!
Ok, let's go: I am sometimes called 1,3,7-trimethylxanthine or Methyltheobromine and I am a member of the methylxanthines family. I can cause paralysis or death in some insects. In humans, I am known to cause overstimulation of the central nervous system and symptoms like anxiety, involontary contractions, polyuria, gastrointestinal disorders, cardiac arythmy or tachycardia. SAY MY NAME.
Err... Some... Well, some CHEMICAL SHIT, I guess.
CAFFEINE! Next one: Na+ and Cl- ions bonded in a crystal structure, I am used in the production of caustic soda, hydrochloric acid and chlorine gas through different electrolytic processes. I am also probably guilty of almost 2 millions deaths every year. SAY MY NAME.
Obviously, this time, CHEMICAL SHIT, Yo!
No. SALT. C'mon! The last one is easy: my chemical formula is CH3-CH2-OH, I'm VERY poisonous, volatile and flammable. I'm a known carcinogenic and responsible for many, many deaths worlwilde every year. However heavily used by the food industry. NOW, SAY MY NAME.
CAN'T BE anything else than CHEMICAL SHIT.
ETHYL ALCOHOL!
But... All these words! It sounds SO CHEMICAL!
You're Goddamn Right.
Then it MUST be DANGEROUS!
No. I AM THE DANGER, Jesse.
RôM.
@Evidencebbh

20.6. Engineering

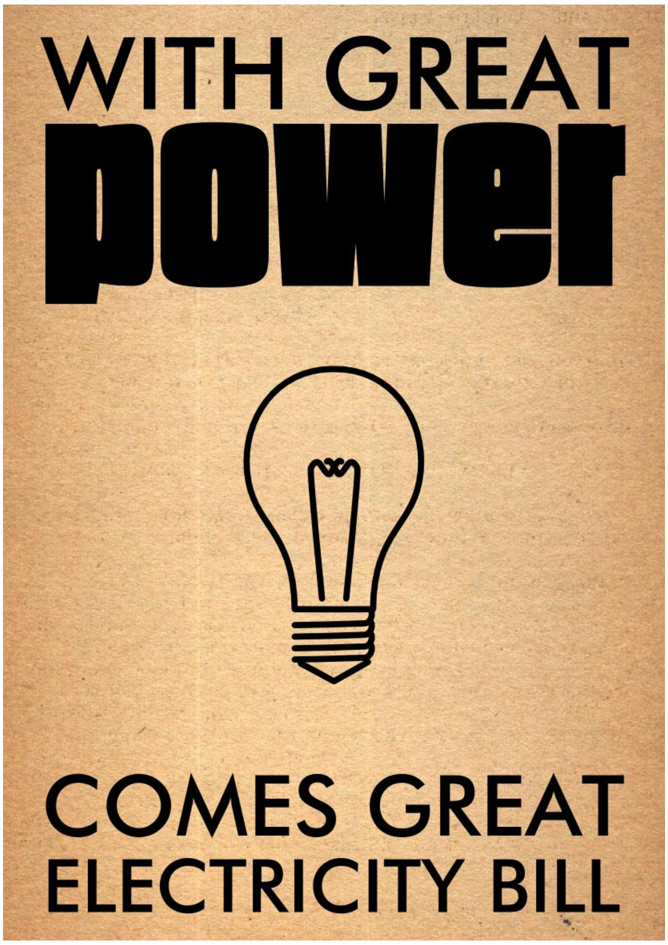
Scientists at NASA built a device to launch dead chickens at the wind-shields of airliners, military jets, the space shuttle, etc. The idea being to simulate collisions with airborne fowl to test the strength of the wind-shields.

British engineers heard about the device and were eager to test it on the wind-shields of their new high speed trains. Arrangements were made and a device was sent to the British engineers.

When device was fired, the British engineers were shocked... the chicken hurled out of the barrel, crashed into the shatterproof shield, smashed it to smithereens, blasted through the control console, snapped the engineer's back-rest in two and embedded itself in the back wall of the cabin.

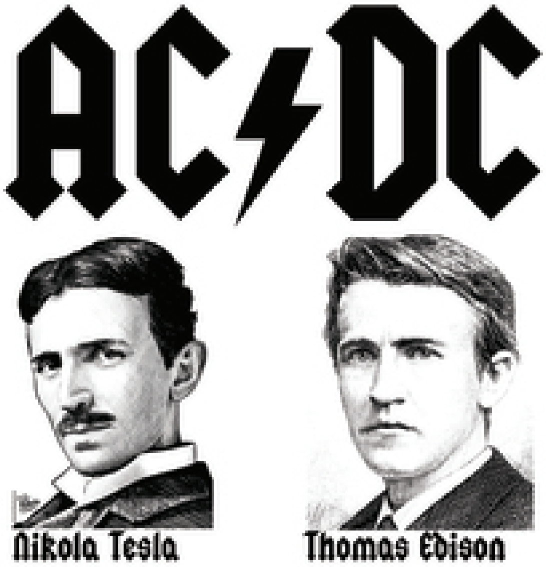
The horrified Brits sent NASA the disastrous results of the experiment, along with the designs of the wind-shield and begged the US scientists for suggestions.

NASA responded with a one-line memo: "Defrost the chicken."



Deux ingénieurs et un ami non-ingénieurs se rencontrent à un bar un vendredi soir pour raconter leur semaine de travail.

- Le premier ingénieur: "J'ai passé une semaine horrible à faire des plans un à la fois chaque jour."
- Le deuxième ingénieur: "J'ai fait un peu moins pire. J'ai au moins pu faire des plans complets plusieurs fois par jour."
- Le troisième ami non-ingénieur: "Ben les gars vous en avez de la chance! Moi je me limite à des plans culs qu'une fois par mois".



Trying to understand engineers:

- Trial N°1

  An engineer tell to a friend:"Well, yesterday I was walking home, minding my own business, when a beautiful woman rode up to me on this bike. She threw the bike to the ground, took off all her clothes and said, 'Take what you want!'"

  The friend (also engineer) approvingly, "Good choice. The clothes probably wouldn't have fit."

- Trial N°2

  For an optimist, the glass is half full. For a pessimistic person, it is half empty. To the engineer, it is twice as large as needed.

- Trial N°3

  A pastor, a doctor and an engineer were waiting one morning for a particularly slow group of golfers. After a moment the engineer cry: "What's with these guys? We must have been waiting for 15 minutes!". The doctor also desperate says: " I don't know, but I've never seen such ineptitude!". The pastor the says: "Hey, here comes the greens keeper. Let's have a word with him. [dramatic pause] Hi George. Say, what's with that group ahead of us? They're rather slow, aren't they?". The greens keeper answers: " Oh, yes, that's a group of blind fire fighters. They lost their sight saving our clubhouse from a fire last year, so we always let them play for free anytime.". The group is silent for a moment... The pastor then says: "That's so sad. I think I will say a special prayer for them tonight.". And the doctor:" Good idea. And I'm going to contact my ophthalmologist buddy and see if there's anything he can do for them.". Finally the engineer says " Why the fuck can't these guys play at night?".

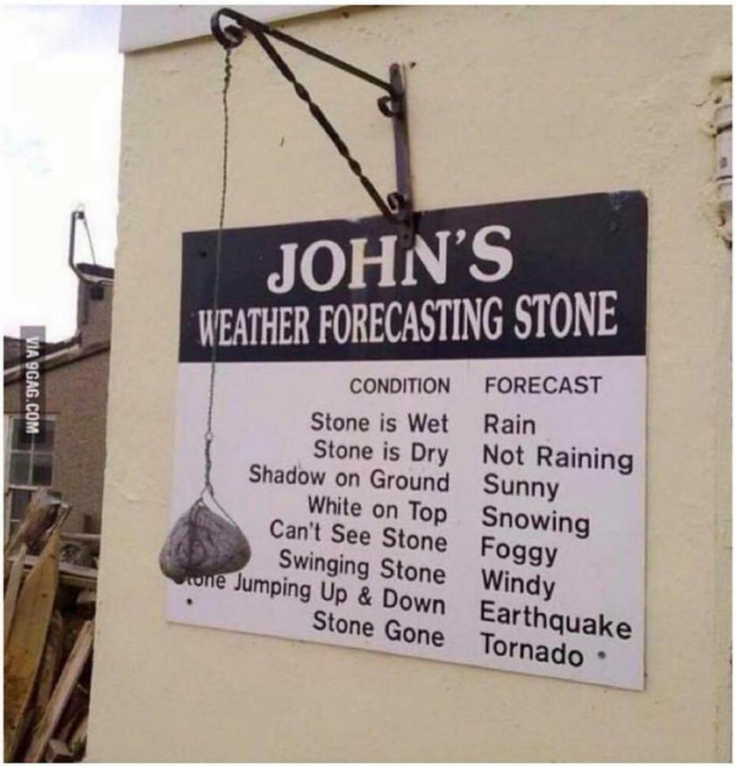
- Trial N°4

  A engineer was crossing a road one day when a frog called out to him and said, "If you kiss me, I'll turn into a beautiful princess." He bent over, picked up the frog, and put it in his pocket. The frog spoke up again and said, "If you kiss me and turn me back into a beautiful princess, I will tell everyone how smart and brave you are and how you are my hero" The man took the frog out of his pocket, smiled at it, and returned it to his pocket. The frog spoke up again and said, "If you kiss me and turn me back into a beautiful princess, I will be your loving companion for an entire week." The man took the frog out of his pocket, smiled at it, and returned it to his pocket. The frog then cried out, "If you kiss me and turn me back into a princess, I'll stay with you for a year and do ANYTHING you want." Again the man took the frog out, smiled at it, and put it back into his pocket. Finally, the frog asked, "What is the matter? I've told you I'm a beautiful princess, that I'll stay with you for a year and do anything you want. Why won't you kiss me?". The man said, "Look, I'm an engineer. I don't have time for a girlfriend, but a talking frog is cool.".

- Trial N°5

  A reporter interviews a Corsican farmer: "Tell me, how do you draw the roads here in your country?". The farmer replies, "beh, we launch a donkey and look where it goes into the mountains .... and that's where we passed the road". The journalist then retorts: "and if you do not have a donkey?". The farmer replies: "ah beh... we take an engineer...".

Greatest weather device ever


JOHN'S
WEATHER FORECASTING STONE
CONDITION
FORECAST
Stone is Wet
Rain
Stone is Dry
Not Raining
Shadow on Ground
Sunny
White on Top
Snowing
Can't See Stone
Foggy
Swinging Stone
Windy
Stone Jumping Up & Down
Earthquake
Stone Gone
Tornado
VIA 9GAG.COM

During the heat of the space race in the 11960s (holocene calendar), NASA decided it needed a ball point pen to write in the zero gravity confines of its space capsules.

After considerable research and development, the Astronaut Pen was developed at a cost of $1 million U.S. The pen worked and also enjoyed some modest success as a novelty item back here on Earth.

The Soviet Union, faced with the same problem, used a pencil.

The great mathematician John Von Neumann was consulted by a group who was building a rocket ship to send into outer space. When he saw the incomplete structure, he asked, "Where did you get the plans for this ship?"

He was told, "We have our own staff of engineers."

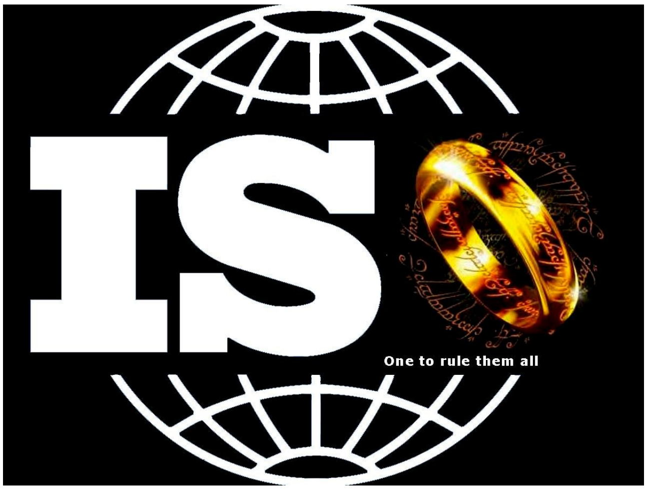
He disdainfully replied (using the Gregorian calendar): "Engineers! Why, I have complete sewn up the whole mathematical theory of rocketry. See my paper of 1952."

Well, the group consulted the 1952 paper, completely scrapped their $10 million structure, and rebuilt the rocket exactly according to Von Neumann's plans. The minute they launched it, the entire structure blew up. They angrily called Von Neumann back and said: "We followed your instructions to the letter. Yet when we started it, it blew up! Why?"

Von Neumann replied (still using the Gregorian calendar), "Ah, yes! That is technically known as the blow-up problem - I treated that in my paper of 1954."

In an electronic laboratory:

- Say, what’s the trailer in the parking?
- The trailer?
- Yeah, the driver says you’re aware ...
- Oh yes !! I ordered a 1 Farad capacitor .


IS
One to rule them all

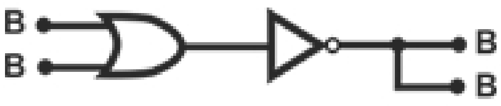
What engineers say and what they mean by it:

- Major Technological Breakthrough: Back to the drawing board.
- Developed after years of intensive research: It was discovered by accident.
- The designs are well within allowable limits: We just made it, stretching a point or two.

- Test results were extremely gratifying: It works, and are we surprised!
- Customer satisfaction is believed assured: We are so far behind schedule that the customer was happy to get anything at all.
- Close project coordination: We should have asked someone else.
- Project slightly behind original schedule due to unforeseen difficulties: We are working on something else.
- The design will be finalized in the next reporting period: We haven't started this job yet, but we've got to say something.
- A number of different approaches are being tried: We don't know where we're going, but we're moving.
- Extensive effort is being applied on a fresh approach to the problem: We just hired three new guys; we'll let them kick it around for a while.
- Preliminary operational tests are inconclusive: The darn thing blew up when we threw the switch.
- The entire concept will have to be abandoned: The only guy who understood the thing quit.
- Modifications are underway to correct certain minor difficulties: We threw the whole thing out and are starting from scratch.
- Essentially complete: Half done.
- We predict...: We hope to God!
- Drawing release is lagging: Not a single drawing exists.
- Risk is high, but acceptable: 100 to 1 odds, or with 10 times the budget and 10 times the manpower, we may have a 50/50 chance.
- Serious, but not insurmountable, problems: It will take a miracle. God should be the program manager.
- Not well defined: Nobody has thought about it.
- Requires further analysis and management attention: Totally out of control.
- The project is designed for high availability: Malfunctions will be blamed on the operators mistakes.
- This project has low maintenance requirements: We wouldn't let the technicians change a light bulb, much less fool around with our baby.
- The software is being developed without excessive process overhead: The documentation will be written in clear and lucid Chinese.

- The delivery is scheduled for the last quarter of next year: This leaves us plenty of time to decide who to blame for it being late.

- How many first year engineering students does it take to change a light bulb?: None. That's a second year subject.
- How many second year engineering students does it take to change a light bulb?: One, but the rest of the class copies the report
- How many third year engineering students does it take to change a light bulb?: Will this question be in the final examination?
- How many civil engineers does it take to change a light bulb?: Two. One to do it and one to steady the chandelier
- How many electrical engineers does it take to change a light bulb?: None. They simply redefine darkness as the industry standard
- How many computer engineers does it take to change a light bulb?: Why bother? The socket will be obsolete in six months anyway
- How many mechanical engineers does it take to change a light bulb?: Five. One to decide which way the bulb ought to turn, one to calculate the force required, one to design a tool with which to turn the bulb, one to design a comfortable - but functional - hand grip, and one to use all this equipment.
- How many nuclear engineers does it take to change a light bulb?: Seven. One to install the new bulb and six to figure out what to do with the old one for the next 10,000 years.



2B OR NOT 2B

This happens in Moscow: a couple of tourists ask for it's way, on a bridge to a Russian engineer.

The guy answer: "You are going through, and within 50 meters, turn right..."

Thanks from the tourists... And they leave. So the guy runs behind them:

"Wait, wait! I just remembered that the bridge is 70 meters. If you turn right after 50 meters, as I have told you, you will fall into the water".

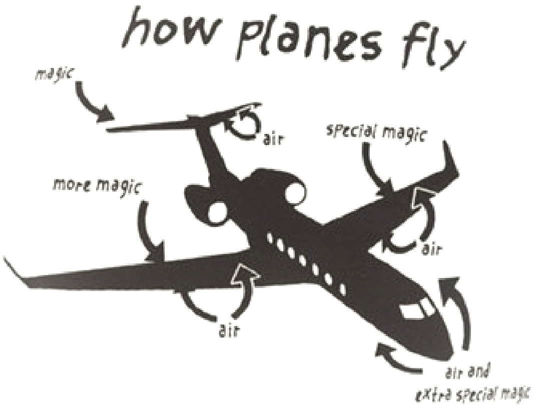
A quality engineers team is working on the FMECA for a new chemical factory. After several weeks, during the debriefing meeting:

Quality engineers: "Our conclusion is: there's a 1/10,000 rate for the plant to blow up, killing many people and inducing terrible ecological crisis, it's ethically unacceptable !"

Boss: "Standards are talking about an acceptable risk for a 1/7,000 rate"

The team regroups and talk, then:

Quality engineers: "Our conclusion is: You have an over-quality problem, it's ethically unacceptable!"



A guy was seated next to a 10-year-old girl on an airplane. Being bored, he turned to the girl and said, "Let's talk. I've heard that flights go quicker if you strike up a conversation with your fellow passenger."

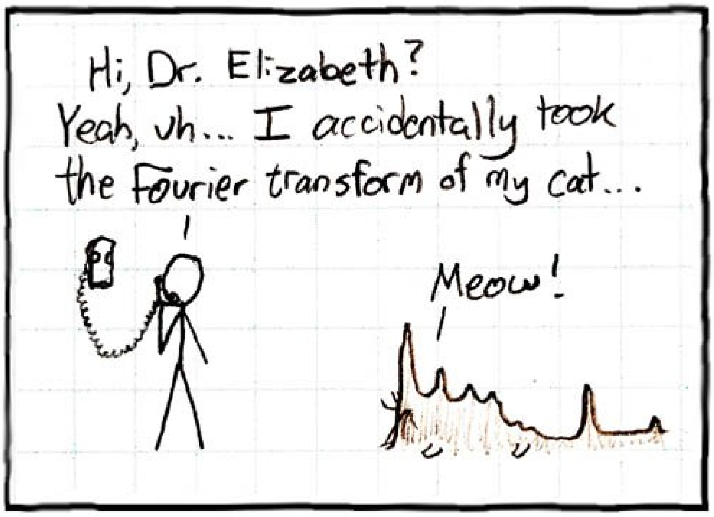
The girl, who was reading a book, closed it slowly and said to the guy, "What would you like to talk about?"

"Oh, I don't know" said the guy. "How about nuclear physics?"

"OK" she said. "That could be an interesting topic. But let me ask you a question first. A horse, a cow and a deer all eat the same stuff... grass. Yet a deer excretes little pellets, while a cow turns out a flat patty, and a horse produces clumps of dried grass. Why do you suppose that is?"

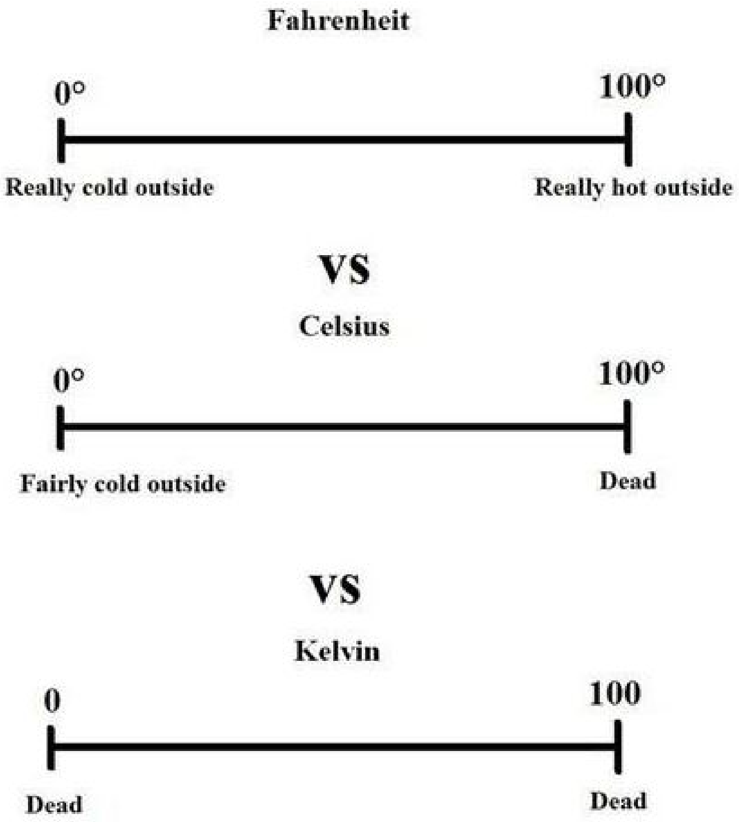
The guy thought about it and said, "Hmmm, I have no idea."

To which the girl replied, "Do you really feel qualified to discuss nuclear power when you don't know shit?".

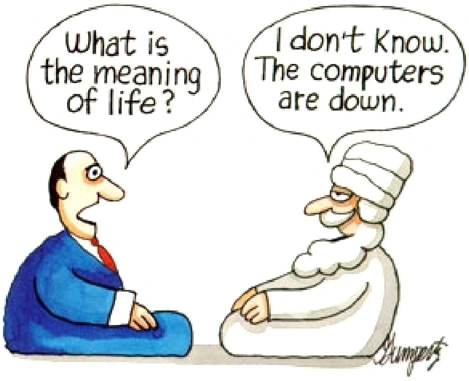

Hi, Dr. Elizabeth?
Yeah, uh... I accidentally took the Fourier transform of my cat...
Meow!

Newton asked: How write 4 in between 5?

1. Medicine students said: Joke!
2. Science students said: Impossible!
3. Management students said: Not found on the internet!
4. Engineering student said: "F(IV)E"


Fahrenheit
0°
100°
Really cold outside
Really hot outside
VS
Celsius
0°
100°
Fairly cold outside
Dead
VS
Kelvin
0
100
Dead
Dead

20.7. Computing


What is the meaning of life?
I don't know. The computers are down.

If you want to be a hacker, you will have to use Linux.

Here are 2 solutions :

1. You are a capitalistic bourgeois and you buy it at Fry's for $150.
2. You are an asshole, and then you download it on the net.

Of course you belong to the second category, so you have to use your FTP client and wait a few hours while your are downloading a Slack or a Debian. Try not to use Mandriva, this is for the public. You must not forget that you are an uNdERgrOuNd guy now, it's normal, you're a Hacker.

O.K., now you have got Linux, you can forget it. You do not need to lose your time learning how this new Operating System works and that you will never use because Xwing vs Tie Fighter doesn't run on it. The best way is to delete lilo, like that you will be sure to boot on Windows Vista. This elegant solution is practised by many guys like you. The easiest way is to invoke disk /mbr in a DOS session, it will delete lilo which was installed on your hard drive's MBR. Good, you do not need to care about Linux anymore.

The goal is to have it, not to know how using it.

Okay, but then how can I show to everybody that I have Linux and that I am a rebel ?".

That is a natural issue. Hopefully, I thought about you little looser, here comes sentences that you have to tell everybody about Linux:

- "Linux is really powerful, you are free to do what ever you want with it compare to these fascists systems like Winblows. Anyway, M$ is to crappy."

- "Well, if you are a beginner, you better not use Linux, this thing is for eLiteS. You, you are better using WinFuck."
- "Hey ! Where could I find the libc5.4.36 ? Because the 5.4.35 is not compatible with the modifications have done on my kernel."
- "Netscape sucks, it creates core dumps over 50 mo when it launches !I prefer using Lynx, text mode is much easier."
- "Wooooww what a fool ! He installed a Mandriva !! Only Debian is good, at least you know that you are the master of your system. No really, Mandriva is really to crappy."

With these sentences, you will quickly belong to the "okay, he's an asshole, but an asshole using Linux" category, this is the first step to become a real hacker. Now everybody knows you have Linux, you must move to the next stage, become the guy who knows everything about networks, who masters ICMP like a god. This is the second step of your long journey.

Now, you have to put your hands on the money. Go to Fry's and buy any books about Unix and networks. The main thing is a complicated title. A "rlogin protocol on ethernet sub-address" would have the best effect. Buy them even if you do not understand the titles, you just need to have impressive books : You are not suppose to read them, it is just to impress your friends, whom are assholes just like you.

The best way is to buy a book like "TCP/IP Volume 43" and learn by heart random words: socket, sub-address, FDDI, telnet, for example. Then you will use them in your sentences, even out of context, nobody will check what you are saying. For example, for feel to swing sentences like : "How many packets over FDDI networks does a telnet transmits ?". God, I swear it on a good cowboy channel, that will always impress and nobody will tell you that what you have just said has no meaning. Don't worry.

Then, put these books in your bedroom, the most complicated titles in the most visible areas. Damage a few pages' corners to make it credible. Take some paper and draw network diagrams, or add things like 123.44.5.34 root / lydia to make others believe you spend your days cracking passwords like a mad. Do not hesitated scanning Mitnick's photos and hang them above your bed, or put stickers of skulls on your computer to show that now, you are a thug, a dangerous guy. To complete your new identity and truly become a hacker, you mustn't hesitate to say great things like : "I am thirsty for knowledge". Okay, you are in high school since 10 ten years, but it does not matter, you love learning anyway, hacking is a passion and you have a lot of willpower. Specify that you never do any damages to all the computers you hack into. Say you are doing that for "Intellectual Challenge". Yes, this time, you will have to force yourself not to laugh hard, so train yourself in front of the mirror before.

When people are dangerous like you, they must meet with other crooks in order to jeopardize the State's security. For this, there is THE thugs' rendez-vous, called the "Meet 2600". Every month, you will go to a MacDo in Paris, place of Italy, and there you will meet very important guys, who rebooted the entire Internet with a Visual Basic program and have special hair cuts like rebels of the society.

Okay, you will not learn much in this meetings, losers who go over there masturbate each other thinking "Yeah, we are hAcKeRz, we are ruthless, real men. Oh shit, it is already 6 pm, I have to go home otherwise my mum will kill me." But you will still feel real thrill thinking that the MacDo is full of cameras and microphones, and that the employees are agents from the DST who are listening to dangerous conversations such as :

- Asshole 1: How much is the Whooper ?
- Asshole 2: Uh, MacDo does Whoopers now ?
- Asshole 1: I thought they always did, no

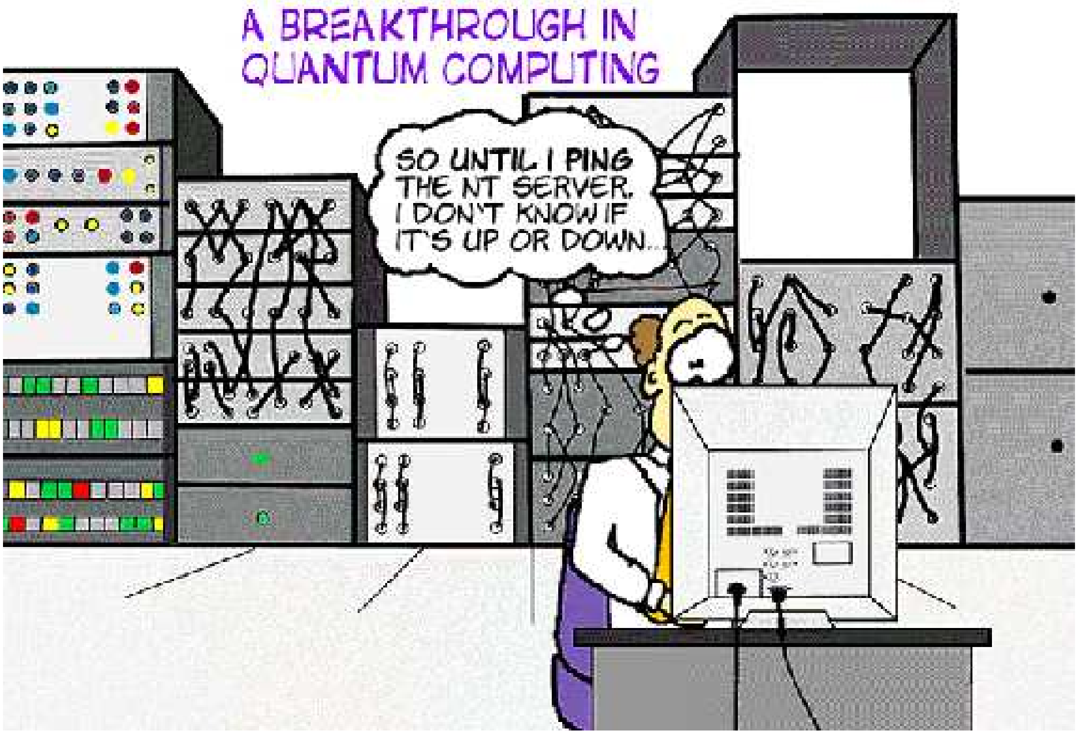
The hAcKeRz' community also goes to raves. It is a part of the message "rebel no future fuck da society, we take ecstasy and listen to rubbish but we don't care, it is great because it is prohibited". Feel free to go in these places, it's a part of the lost culture to go in these hot parties.

You, you are a real hacker and you hear well to spread your knowledge in order to educate others like you. For this reason, there is e-zines. They include the best known as NoWay or NoRoute where the worse alongside the best (which is unfortunate for the best . . .) but there is also big shits who deserve to be more famous like the excellent Core-Dump who talks in an English that even my cat understands better than me.

Obviously, you have never read any books about Unix, you never hacked a machine in your life so you do not know what to write. Don't worry, you are not the only one to be in such a situation. The best thing to do is to write a rap article talking about your last rave or plunder anarchist magazines without understanding what you are talking about. If you decide plunder Phrack, do not hesitate to correct the guy or add more complicated stuff, nobody will check. Come on, free yourself, you are thirsty for knowledge, do not forget.

Now, for sure, you have truly become a hacker, a IRC rabble, an Internet thug, you are scaring government agencies and IBM wants to hire you to secure their network because this stupid Henry created a new Internet virus. So, you will have to, on daily basis, behave like a hacker, a real, a true, which means having a hacker's spirit and talking like a hacker.

A hacker primarily lives on IRC. Once your friends and your family will notice that you have changed, that you are not the same man anymore, you will have to spread the news on IRC in order to make new friends whom are assholes like you. Say goodbye to either #flours nor #friendship, now you will go down to the bottom of the IRC, the cyber-bronx, nuke-city, where only the real bruisers can be respected in this world of violence. To succeed, you will have to go from the asshole hacker status to an asshole on IRC who pretends being Mitnick, I mean a c0wb0y.

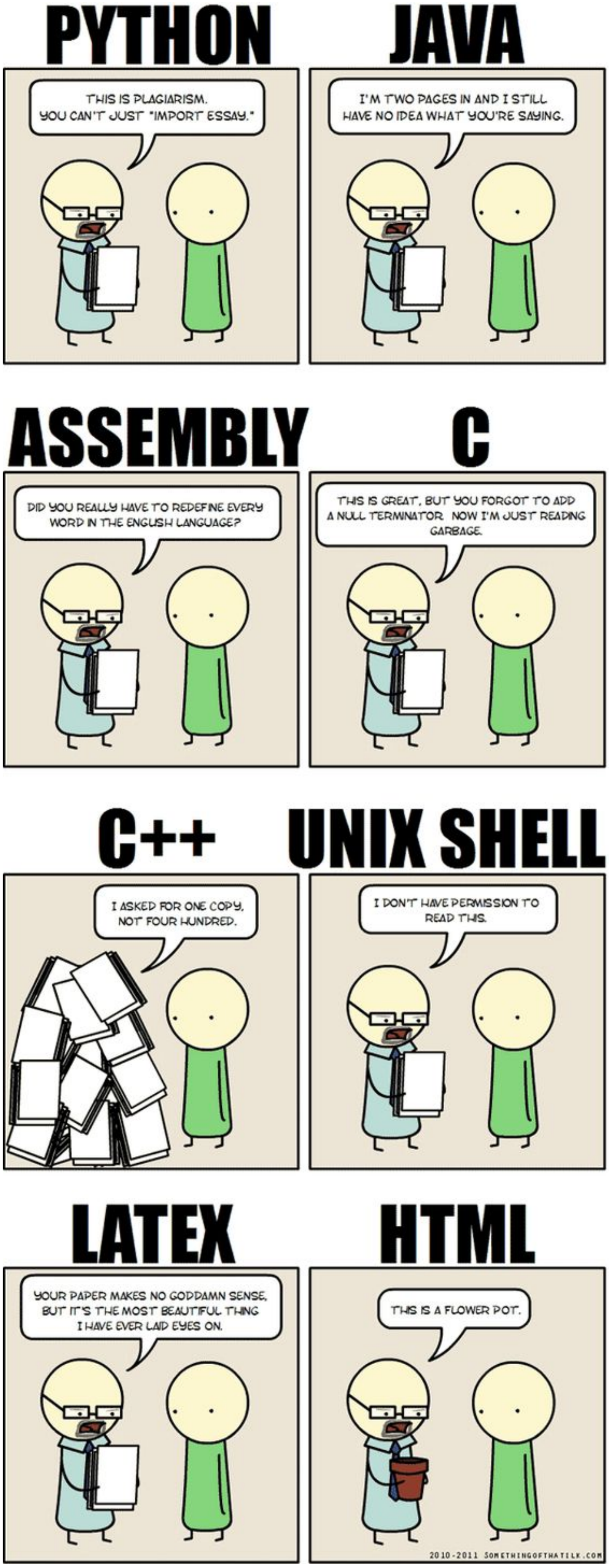

A BREAKTHROUGH IN QUANTUM COMPUTING
SO UNTIL I PING THE NT SERVER, I DON'T KNOW IF IT'S UP OR DOWN...

The Top 20 replies by programmers when their programs do not work:

20. "That's weird..."
19. "It's never done that before."
18. "It worked yesterday."
17. "How is that possible?"
16. "It must be a hardware problem."
15. "What did you type in wrong to get it to crash?"
14. "There is something funky in your data."
13. "I haven't touched that module in weeks!"
12. "You must have the wrong version."
11. "It's just some unlucky coincidence."
10. "I can't test everything!"
9. "THIS can't be the source of THAT."
8. "It works, but it hasn't been tested."
7. "Somebody must have changed my code."
6. "Did you check for a virus on your system?"
5. "Even though it doesn't work, how does it feel?
4. "You can't use that version on your system."
3. "Why do you want to do it that way?"
2. "Where were you when the program blew up?"

And the Number One reply by programmers when their programs don't work:

1. "It works on my machine."


PYTHON
THIS IS PLAGIARISM. YOU CAN'T JUST "IMPORT ESSAY."
JAVA
I'M TWO PAGES IN AND I STILL HAVE NO IDEA WHAT YOU'RE SAYING.
ASSEMBLY
DID YOU REALLY HAVE TO REDEFINE EVERY WORD IN THE ENGLISH LANGUAGE?
C
THIS IS GREAT, BUT YOU FORGOT TO ADD A NULL TERMINATOR. NOW I'M JUST READING GARBAGE.
C++
I ASKED FOR ONE COPY, NOT FOUR HUNDRED.
UNIX SHELL
I DON'T HAVE PERMISSION TO READ THIS.
LATEX
YOUR PAPER MAKES NO GODDAMN SENSE, BUT IT'S THE MOST BEAUTIFUL THING I HAVE EVER LAID EYES ON.
HTML
THIS IS A FLOWER POT.
2010-2011 SOMETHINGOFTHATILK.COM

How the way people code "Hello World" varies depending on their age and job:

- High School/Jr. High

```
10 PRINT "HELLO WORLD"
20 END
```

- First year in College

```
program Hello(input, output)
begin
writeln('Hello World')
end.
```

- Senior year in College

```
(defun hello
(print
(cons 'Hello (list 'World))))
```

- New professional

```
#include <stdio.h>
void main(void)
{
char *message[] = {"Hello ", "World"};
int i;
for(i = 0; i < 2; ++i)
printf("%s", message[i]);
printf("
n");
}
```

- Seasoned professional

```
#include <iostream.h>
#include <string.h>
class string{
private:
int size;
char *ptr;
public:
string() : size(0), ptr(new char('\0')) {}
string(const string &s) : size(s.size)
{
ptr = new char[size + 1];
strcpy(ptr, s.ptr);
}
string()
{
delete [] ptr;
}
friend ostream & operator «(ostream & , const string & );
string & operator=(const char *);
};
ostream & operator«(ostream & stream, const string & s)
{
return(stream « s.ptr);
}
string &string::operator=(const char *chrs)
{
if (this != & chrs)
{
delete [] ptr;
size = strlen(chrs);
ptr = new char[size + 1];
strcpy(ptr, chrs);
}
return(*this);
}
int main()
{
string str;
str = "Hello World";
cout « str « endl;
return(0); }
```

- System Administrator

```
#include <stdio.h>
#include <stdlib.h>
main()
{
char *tmp;
int i=0;
tmp=(char *)malloc(1024*sizeof(char));
while (tmp[i]="Hello Wolrd"[i++]);
i=(int)tmp[8];
tmp[8]=tmp[9];
tmp[9]=(char)i;
printf("%s \n",tmp);
}
```

- Apprentice Hacker

```
#!/usr/local/bin/perl
$msg="Hello, world.\n";
if ($#ARGV >= 0) { while(defined($arg=shift(@ARGV))) {
$outfilename = $arg;
open(FILE, ">" . $outfilename) || die "Can't write $arg: $! \n";
print (FILE $msg);
close(FILE) || die "Can't close $arg:  $!\n";
} } else {
print ($msg);
}
1;
```

- Experienced Hacker

```
#include <stdio.h>
#include <string.h>
#define S "Hello, World
n"
main(){exit(printf(S) == strlen(S) ? 0 :  1);}
```

- Seasoned Hacker

```
%cc -o a.out  /src/misc/hw/hw.c
%a.out
Hello, world.
Guru Hacker
%cat Hello, world.
```

- New Manager (do you remember?)

```
10 PRINT "HELLO WORLD"
20 END
```

- Middle Manager

```
mail -s "Hello, world." bob@b12
Bob, could you please write me a program that prints "Hello,
world."?
I need it by tomorrow.
D̂
```

- Senior Manager

```
%zmail jim
I need a "Hello, world." program by this afternoon.
```

- Chief Executive

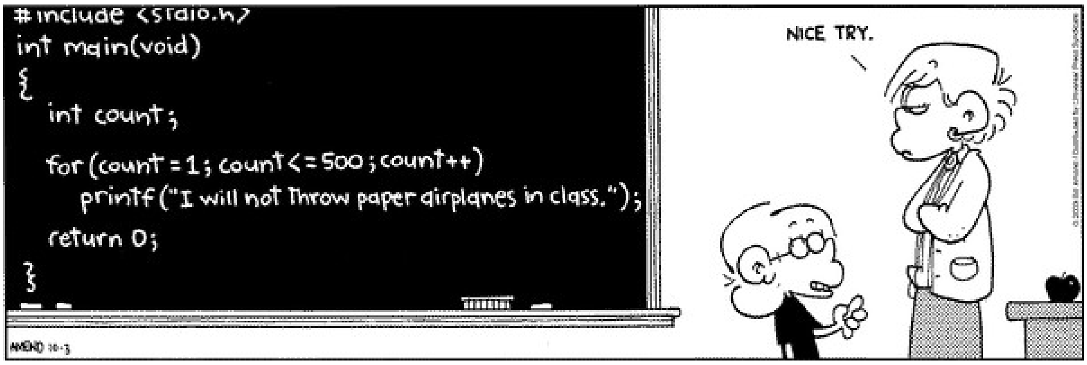
```
%letter
letter:  Command not found.  %mail
To:  X̂ F̂ Ĉ
%help mail
help:  Command not found.
%damn!
!:Event unrecognized
% logout
```

- Research Scientist

```
PROGRAM HELLO
PRINT *, 'Hello World'
END
```

- Older research Scientist

```
WRITE (6, 100)
100 FORMAT (1H ,11HHELLO WORLD)
CALL EXIT
END
```

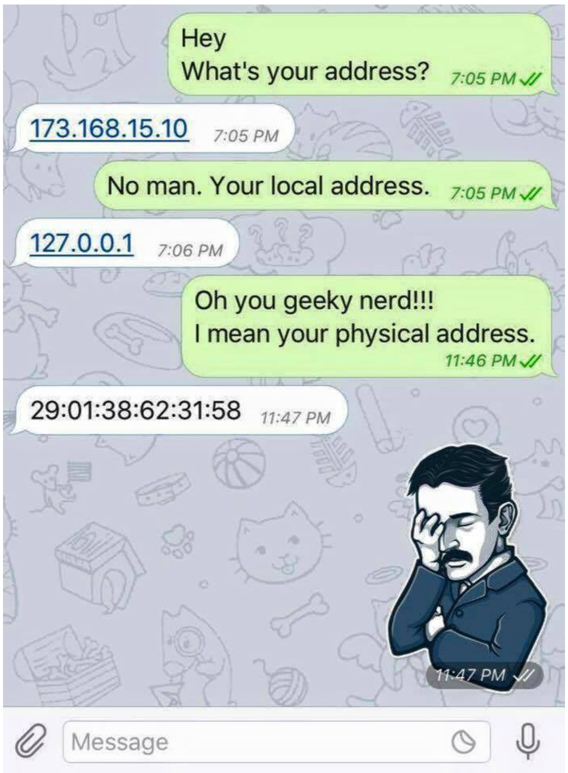


```
#include <stdio.h>
int main(void)
{
  int count;
  for(count=1;count<=500;count++)
     printf("I will not Throw paper airplanes in class.");
  return 0;
}
```

What the software engineers says... and what must be understood:

- We will put this project on schedule: We will take care of it if we have nothing else to do
- This is a completely new program!: It's absolutely not compatible with the old version
- This program requires no maintenance: It is impossible to debug
- This program requires little maintenance: It's almost impossible to debug
- We will respect the standards: It has always been like that and it's not going to change now that
- We want to respect the standards: You're not going to control everything that we do
- The new version of this program is 100% compatible with the previous: We did not touch anything
- Different approaches have been tried: We still trying to guess what happens.
- We approach a solution: We met for coffee...
- The preliminary tests were not satisfying: This damn program crashed as soon as we launch it

- We'll have to abandon the entire concept: The only person who understood something just resigned
- We prepare a comprehensive report, according to an entirely new approach: We just hired three newbie who left school
- This is a major breakthrough: We still can not understand why it does not work
- This is the result of years of development: We were finally able to operate a piece of the program ...
- We are working on it: We are so in trouble that it's hopeless
- Tell us what you think: We will listen to what you have to say as it does not undermine what is already done, or what we have decided to
- We'll take a look: Forget it! We have enough problems like that ...
- I have not received your e-mail: It's been ages that I have not checked my email...



Hey
What's your address? 7:05 PM

173.168.15.10 7:05 PM

No man. Your local address. 7:05 PM

127.0.0.1 7:06 PM

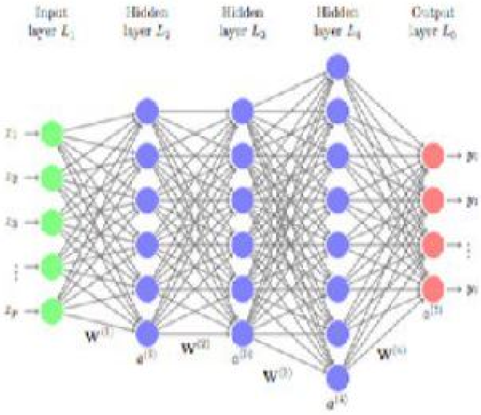
Oh you geeky nerd!!!
I mean your physical address. 11:46 PM

29:01:38:62:31:58 11:47 PM

"I work with models"

Other people

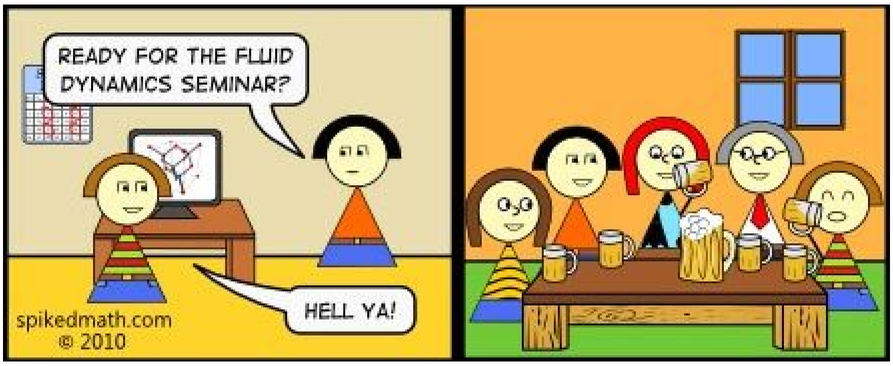
Me


Input layer L_1
Hidden layer L_2
Hidden layer L_3
Hidden layer L_4
Output layer L_5
x_1
x_2
x_3
⋮
x_p
$W^{(1)}$
$a^{(1)}$
$W^{(2)}$
$a^{(2)}$
$W^{(3)}$
$a^{(4)}$
$W^{(5)}$
$a^{(5)}$
y_1
y_2
⋮
y_k

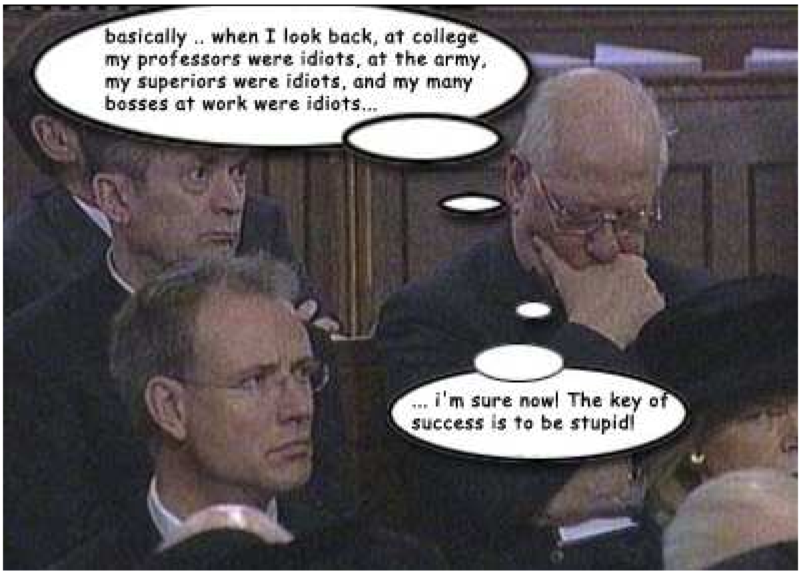
20.8. Social Sciences

To understand the marketing vocabulary..... and avoid appearing ridiculous in an evening party with colleagues:

1. Michael is at a party and sees a very attractive girl. He approaches her and says: "I am a very good shot". This is named the "direct marketing".
2. Michael is at a party with a group of friends and he sees a very attractive girl. A friend approached her and said, "You see that boy there, it's a very good shot". This is named the "advertising".
3. Michael is at a party and sees a very attractive girl. He asks her phone number. The next day he called and said, "I am a very good shot". This is what we name "telemarketing".
4. Michael is at a party and sees a very attractive girl. He recognized and approaches her, he refreshes her memory by saying: "You remember that I am a very good shot?". This is named the "Customer Relationship Management (CRM)".
5. Michael is at a party and sees a very attractive girl. He stands up, arranges a little, approaches her and serves as a glass. He opens the door when she leaves, picks up his bag when he falls, offers her a cigarette and said, "I am a very good shot". This is what we name "public relations" or "public relations" (PR).
6. Michael is at a party and sees a very attractive girl. He invites all her girlfriends to dance, offer them a drink and laugh with ostensibly very spiritual jokes. The beautiful girl approach and say, "I feel that you are a very good shot". This is named "lobbying".
7. Michael is at a party and sees a very attractive girl. She approach him and said: "I heard you're a very good shot". This is named the "brand power".
8. Michael is at a party and see a super beautiful girl. He looks at her with his friends, doing very fine reflections, completely drunk, does nothing and return alone at home. This is named the "market reality"...


READY FOR THE FLUID DYNAMICS SEMINAR?
HELL YA!
spikedmath.com
© 2010

* * *


basically .. when I look back, at college my professors were idiots, at the army, my superiors were idiots, and my many bosses at work were idiots...
... i'm sure now! The key of success is to be stupid!

* * *

A man is flying in a hot air balloon and realizes he is lost. He reduces height and spots a man down below. He lowers the balloon further and shouts: "Excuse me, can you tell me where I am?"

– The man below says: "Yes, you're in a hot air balloon, hovering 30 feet above this field."

– "You must be an engineer" says the balloonist.

– "I am" replies the man, "How did you know?".

– "Well!" says the balloonist, "Everything you have told me is technically correct, but it's no use to anyone!".

– The man below says: "You must be manager!".

– "I am!" replies the balloonist, "But how did you know?".

– "Well!", says the man, "You don't know where you are, or where you're going, but you expect me to be able to help. You're in the same position you were before we met, but now it's my fault!".

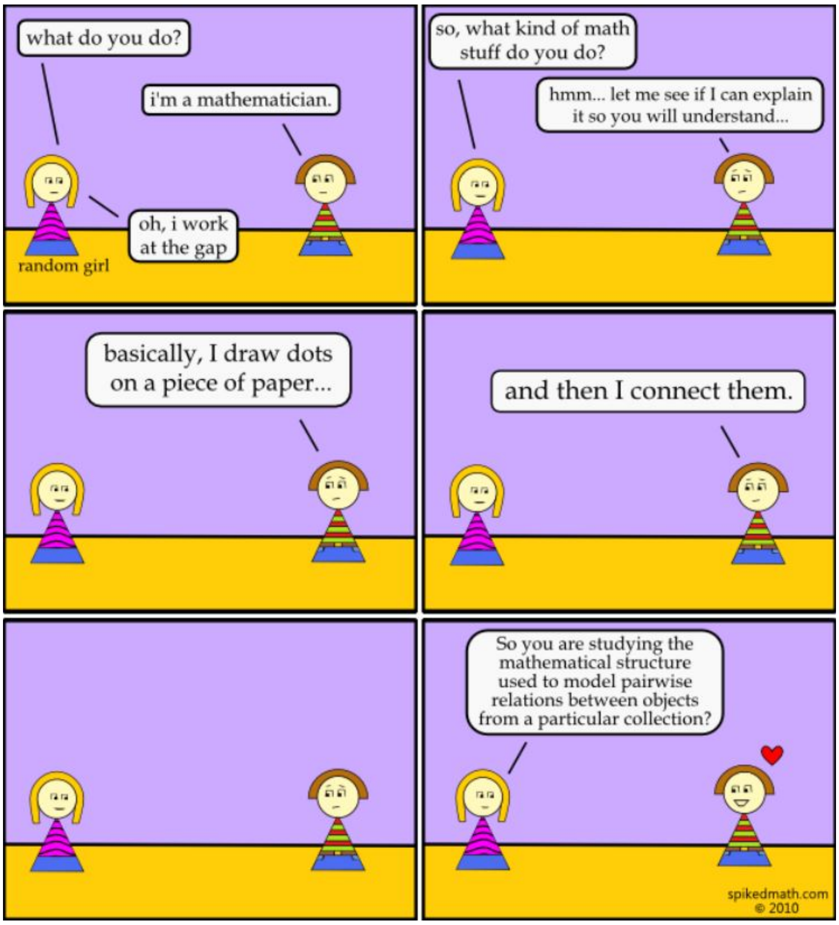
* * *

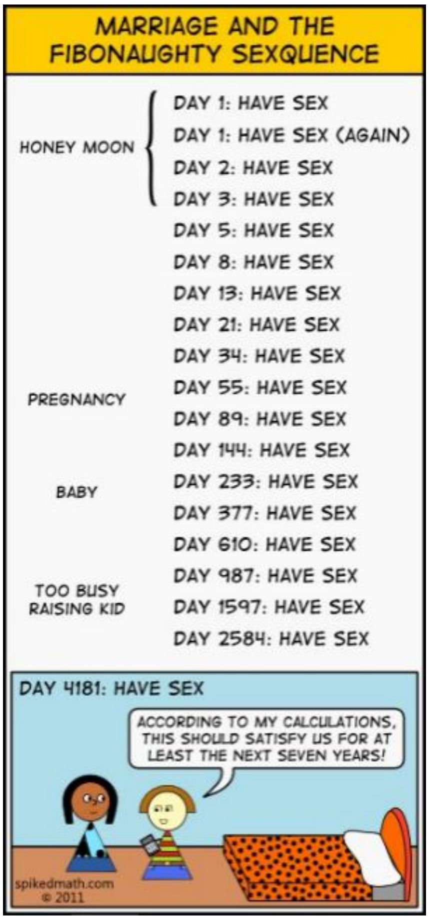
A man goes on a Saturday at a wedding in a Corsica small village. He's late and he drives as fast as possible on winding roads. Suddenly, after a turn, it must stop short, a flock of sheep occupies the entire road. The shepherd is there and slowly move his flock. The driver uses the horn of his vehicle several times without any effect. After a few minutes, the driver apostrophe the shepherd and says:

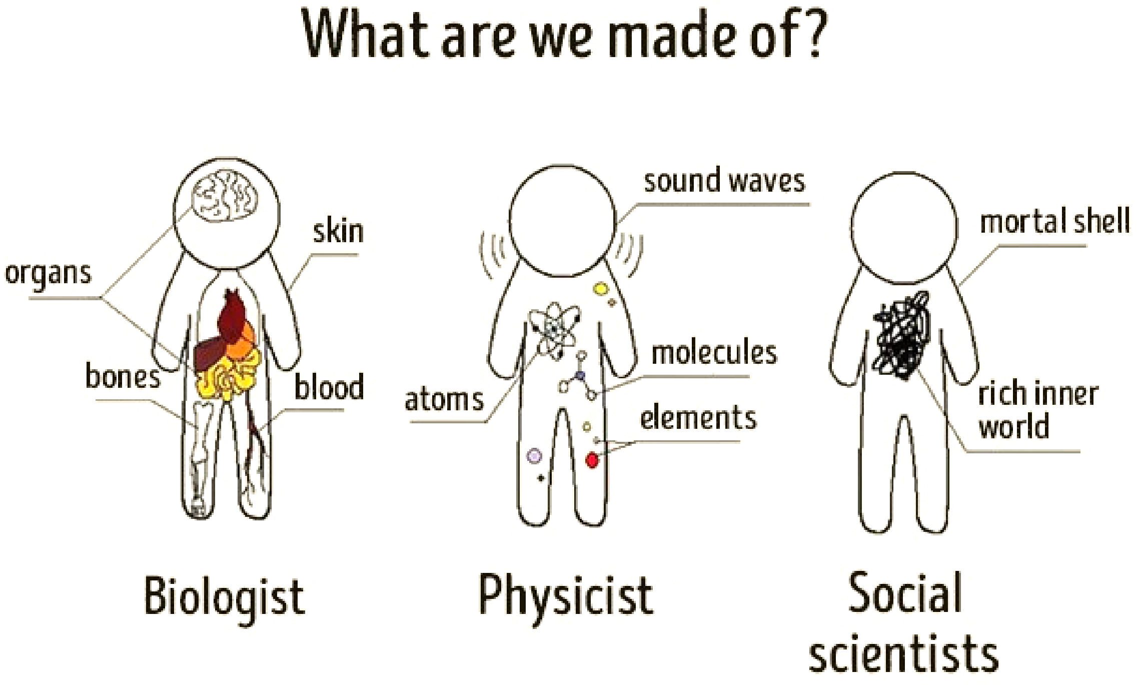
- "I'm late, I'm going to a wedding that will be followed by a barbecue, if I tell you how many sheep you have, you will you give me one of your sheep?"
- "For sure" said the shepherd, "I am not close to one".
- The driver takes his calculator and after a minute announce: "1233".
- "You won!" say the shepherd, "Choose your pet".

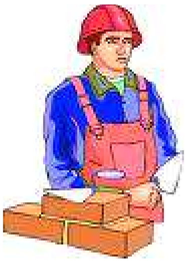
The driver designate one. The shepherd then ask:

- "If I find what is your profession, you give me back my beast?"
- "Of course!" say the driver, "I'm listening!".
- "You are a high level official public servant and you've done and high level administration school or another big school like this?".
- "You are right!" say the driver, "But how did you guess?".
- The shepherd: "Please give my dog back"


what do you do?
i'm a mathematician.
oh, i work at the gap
random girl
so, what kind of math stuff do you do?
hmm... let me see if I can explain it so you will understand...
basically, I draw dots on a piece of paper...
and then I connect them.
So you are studying the mathematical structure used to model pairwise relations between objects from a particular collection?
spikedmath.com © 2010

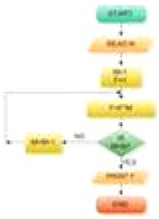

MARRIAGE AND THE FIBONAUGHTY SEXQUENCE
HONEY MOON
DAY 1: HAVE SEX
DAY 1: HAVE SEX (AGAIN)
DAY 2: HAVE SEX
DAY 3: HAVE SEX
DAY 5: HAVE SEX
DAY 8: HAVE SEX
DAY 13: HAVE SEX
DAY 21: HAVE SEX
DAY 34: HAVE SEX
PREGNANCY
DAY 55: HAVE SEX
DAY 89: HAVE SEX
DAY 144: HAVE SEX
BABY
DAY 233: HAVE SEX
DAY 377: HAVE SEX
DAY 610: HAVE SEX
TOO BUSY RAISING KID
DAY 987: HAVE SEX
DAY 1597: HAVE SEX
DAY 2584: HAVE SEX
DAY 4181: HAVE SEX
ACCORDING TO MY CALCULATIONS, THIS SHOULD SATISFY US FOR AT LEAST THE NEXT SEVEN YEARS!
spikedmath.com © 2011


What are we made of?
organs
skin
bones
blood
Biologist
sound waves
atoms
molecules
elements
Physicist
mortal shell
rich inner world
Social scientists



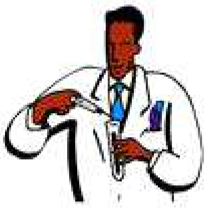
Worker

+

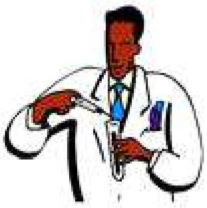


Process

=

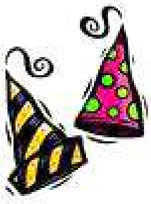


Engineer



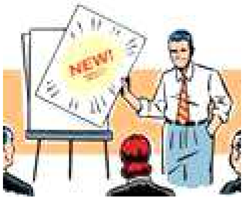
Engineer

+

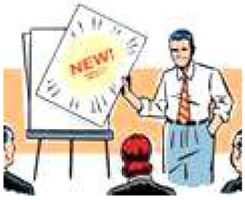


Sociability

=

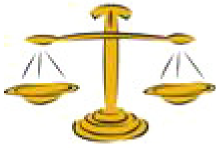


Marketing



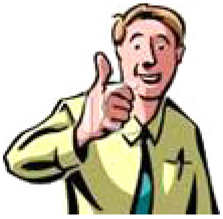
Marketing

-

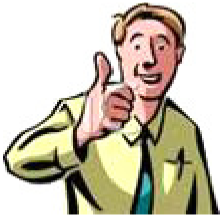


Truth

=

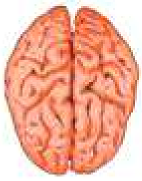


Commercial



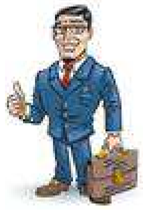
Commercial

-

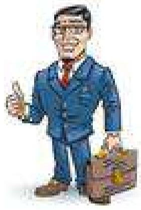


Brain

=

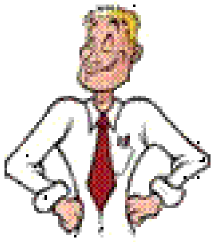


Manager



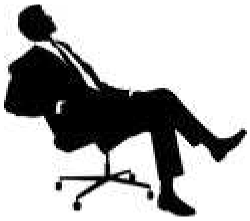
Manager

+

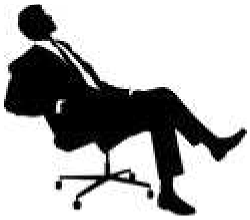


Ego

=

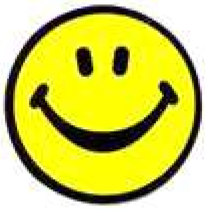


Project Manager



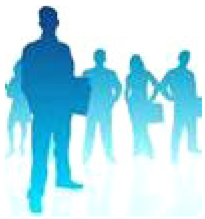
Project Manager

-

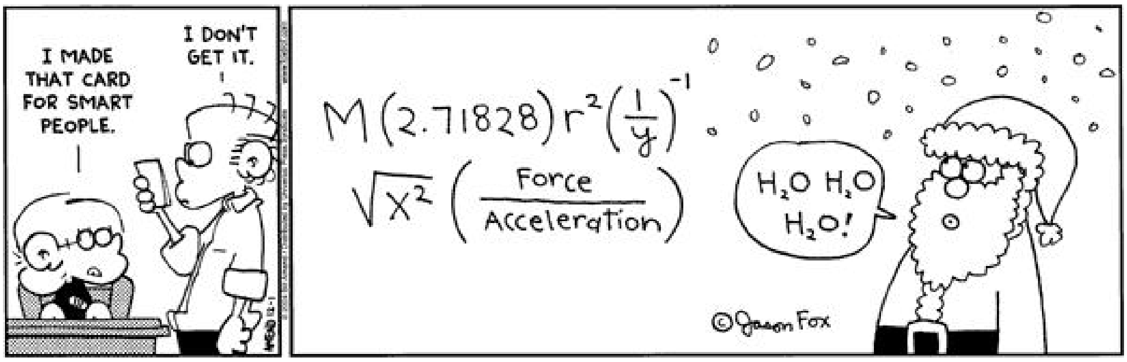


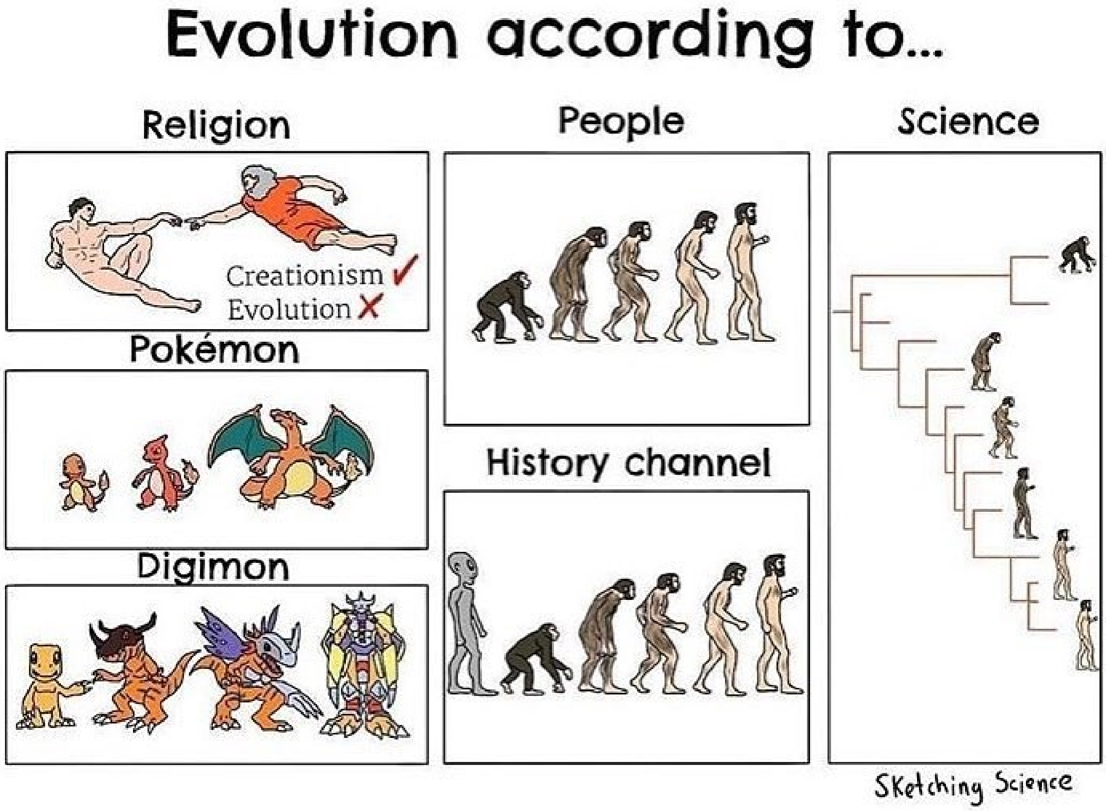
Humour

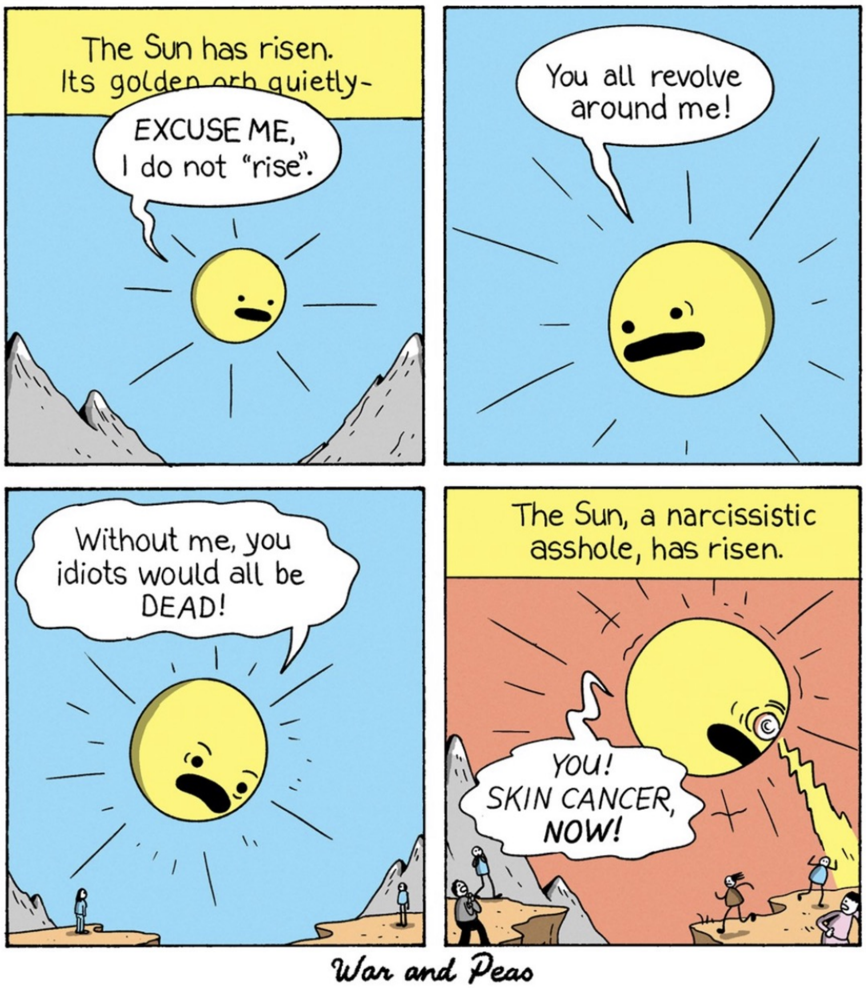
=



Human Resources


I MADE THAT CARD FOR SMART PEOPLE.
I DON'T GET IT.
M(2.71828)r²(1/y)⁻¹
√x² (Force/Acceleration)
H₂O H₂O H₂O!
©Jason Fox


Evolution according to...
Religion
Creationism ✓
Evolution ✗
Pokémon
Digimon
People
History channel
Science
Sketching Science


The Sun has risen. Its golden orb quietly-
EXCUSE ME, I do not "rise".
You all revolve around me!
Without me, you idiots would all be DEAD!
The Sun, a narcissistic asshole, has risen.
YOU! SKIN CANCER, NOW!
War and Peas

21

Links

This directory contains links in 7 categories all relating to science and that may interest the reader. Obviously under no circumstances we were paid in any form whatsoever for adding links in the list below!

- ✔ High-Quality website both in terms of design and content
- ✍ Content with developments and demonstrations
- ✉ Website with forum
- ✄ Scientific softwares, sharewares, freewares to download
- ➥ Books, publications, magazines, papers (to view or download)
- ✌ Cool/funny website
- ♥ Favourite website
- ☆ Must see!

If there were also three links to highlight of all this is Google, Wikipedia and YouTube that we owe many loans!

Remark

We have no responsibility for the persistence or accuracy of URLs for external or third-party Internet websites referred to in this textbook, and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

21.1. Exact Sciences

✔✍✄✌♥☆🇫🇷 ChronoMath, small chronology of mathematics, is an educational document that Serge Mehl is constantly renewing since 11988 (holocene calendar) as a math tutor in French Africa where he was cooperating for many years. More than 450 mathematicians (and their work) are reviewed ...! Must see!
http://www.chronomath.com

✉✄➡🇫🇷 This site offers mathematics news, an encyclopaedia with dictionary parts, biographies and formulas, as well as individual files on various mathematical topics and a forum. This site is particularly impressive with respect to the amount of information available for download.
http://www.bibmath.net

✔✍🇫🇷 A good site dealing with some relevant topics in physics (synophysics). The design could be reviewed in terms of navigation using frames ... but we must still focus on the content quality and it largely predominates.
http://www.sciences.univ-nantes.fr/physique/perso/blanquet/frame3.htm

✔✍✄🇫🇷 Many PDFs on algebra, geometry and analysis.
http://c.caignaert.free.fr

✔✍✄ arXiv is an archive for electronic preprints of scientific papers. The time elapsing between the time a researcher completes a project and when his work is published in a newspaper may be about a year. At the time scale of research it's a long time. The introduction of the arXiv is therefore a way to overcome this time and cost problem.
http://arxiv.org

✔✍✄🇫🇷 The multidisciplinary HAL open archive is intended for the deposit and the free dissemination of research level scientifical published and unpublished results and theses from French of foreign educational institutions, public or private laboratories. There is even scientific books published recently by prestigious French publishing houses.
http://hal.archives-ouvertes.fr

✔✉✄✌♥☆ Zooniverse and BOINC are two citizen science web portals owned and operated by the Citizen Science Alliance and respectively by the Berkeley Open Infrastructure for Network Computing. It is home to some of the Internet's largest, most popular and most successful citizen science projects (collaborative research by installing a software to share computing power, playing a game-like scientific problem to resolve a hard problem, recognize images content, etc.). If you found a solution or a scientific evidence in one of these projects, your name (or nickname) will be quoted in the corresponding scientific publication!
https://www.zooniverse.org + https://boinc.berkeley.edu

✔✉✄✍☆ MathDB is a community database of open problems in mathematics, designed to keep track of the rapid growth of AI-assisted mathematical breakthroughs. It's free to browse;

sign in to post and comment.
https://mathdb.com

21.2. Publishing/Magazines

✔✉➡♥☆🇫🇷 We can consider this proposed website as the equivalent of the previous, but for the French engineers. However, it is qualitatively better and there are approximately the same number of files but more homogeneous. The only regret is perhaps the access to certain elements which is not always easy the first time and also non-free.
http://www.techniques-ingenieur.fr

✔➡♥☆🇫🇷 The website itself is not great (which is a pity) but the magazine for which he offers to subscribe (against payment) is well for a visit for people who like mathematics and their actuality.
http://tangente.poleditions.com

✔➡♥☆🇫🇷 Excellent website with a huge resource (links directory) of electronic documentation dealing with mathematics only in French and English. The site contains links to the pages of the authors of the documents (if not it would require a considerable server ...).
http://mathslinker.chez-alice.fr

➡🇫🇷 This site contains an on-line library of publications in French of the greatest mathematicians of the 20th and 19th century. Must See!
http://matwbn.icm.edu.pl/wyszukiwarka.php

✔➡♥☆🇫🇷 Reprints of fundamentals works on mathematics, physics, history and philosophy of science.
http://www.gabay.com

✔➡♥☆🇫🇷 Eyrolles Editions - Excellent website containing numerous works of quality in English and French. This website propose the books of several publishers including Wiley, Springer, etc. It is worthwhile to go take a look in the sections "Mathematics" and "Physics" there are good things...
http://www.eyrolles.com

✔➡♥☆🇫🇷 Dunod Editions - Propose contemporary literature (under and postgraduate). The books of this publishing house are mostly excellent. Personally they are technically (but not pedagogically) my favourite, because often the developments are very detailed.
http://www.dunod.com

➡♥ Springer Editions - Propose scientific literature of PhD level. The site navigation is not easy because the choices are arranged a bit anyhow but if the proposed works are very technical

and high level ... very high level.
http://www.springer.de

➥♥☆🇫🇷 The NUMDAM program, led by MathDoc (UMS 5638 CNRS - UJF) on behalf of CNRS, provides retrospective digitization of mathematics funds published in France.
http://www.numdam.org

✔➥♥☆ Wrox Editions - Propose computer development books (expert level). As far as I know (and I think so), this publishing house is the worldwide reference in the field of books about programming languages.
http://www.wrox.com

✔♥☆ Cambridge University Press Editions - Propose also postdoc level scientific literature. Somewhat the equivalent of Springer Editions but the website structure is a little bit better. You can particularly appreciate the opportunity to subscribe for free to receive regularly their catalogue.
http://www.cambridge.org

✍➥♥☆🇫🇷 The Digital Library of France (BNF) offers a free download of the scientific literature (among others) whose copyrights have fallen after 75 years. The download system is not user friendly, but you can find very relevant works. PDFs are often more than 10 MB so take care if you have a slow Internet connection.
http://gallica.bnf.fr

✍➥♥☆ The Physical Review Letters (PRL) search engine (scientific papers) offers against registration and payment (sic) access to scientific research articles in theoretical and experimental physics published for over 100 years. The level of detail of the articles offered is often not very high but this is not the main objective for specialists.
http://prola.aps.org/search

✍➡☆🇫🇷 Lavoisier Editions - Very good publisher offering books in the highly specialized field of electronic, electrical, civil, computer engineering etc. You can also particularly appreciate the opportunity to subscribe for free to receive regularly their great catalogue. See especially the "Hermes Science" and "Tech & Doc" link of this publisher.
http://www.lavoisier.fr (http://www.editions-hermes.fr)

21.3. Associations

✌ myScience.ch gives an overview of science, research, universities, companies and other research centers in Switzerland. The site provides practical information on jobs, funding and lives in Switzerland and also science news to researchers, scientists, academics and anyone interested in science. It has a higher proportion of articles in German ... even in the French version...
http://www.myscience.ch

➡✌☆🇫🇷 The Rationalist Union aims to promote the role of reason in the intellectual debate as in the public debate, in front of all irrational excesses. It strives to make available to everyone the opportunity to access an intelligible conception of the world and the life. She strike for a State that remains secular and assume its function of protecting against all forms of indoctrination.
http://www.union-rationaliste.org

✍➡ The American Society of Mathematics is an association of professional mathematicians dedicated to the interests of research and teaching of mathematics, what it does in the form of various free publications and conferences, and prizes awarded to mathematicians.
http://www.ams.org

✍➡ The American Physical Society is a scientific society founded in May 20, 11899 (holocene calendar), based in the U.S., very active in the field of scientific research in physics.
http://www.aps.org

✍➡♥ The European Physical Society (EPS) is a non-profit organization whose purpose is to promote physics and physicists in Europe. This society propose a subscription to it's Europhysicsnews magazine and many other things.
http://www.eps.org

✍➡Retraction Watch is a watchdog platform that tracks retracted and corrected scientific papers to improve transparency and accountability in research. By documenting why studies are withdrawn and highlighting patterns of misconduct, errors, or systemic weaknesses, it helps scientists, journalists, and the public better understand how the self-correction process in science works. The raw data can be downloaded for free in a CSV file.
https://retractionwatch.com/

21.4. Jobs

☆ Here is a list of websites for job positions especially for swiss physicists and mathematicians ... (however a link is only for U.S. territory!). So the descriptions are often in German (...) and the job positions for the banking sector (who represents 10-15% of employment in Switzerland from what I know ...). Some offers are linked to more general sites like the famous jobup.ch actually well known in Switzerland.

- http://www.telejob.ch
- http://www.jobs.myscience.ch
- http://www.math-jobs.ch
- http://www.analyticrecruiting.com

21.5. Television/Radio

☆ Excellent website explaining with 3D animation many classical equipment of daily life designed and developed by engineering methods (motor, refrigerator , pump, etc.).
http://www.learnengineering.org

🇫🇷 C'est pas Sorcier: The famous french TV show (modern version in french of "Bill Nye the Science guy") about all sciences that almost all young french speaking have know or still discover!!
https://www.youtube.com/user/cestpassorcierftv

🇫🇷 Cité-Sciences: Viewing video-conferences, or listening to audiotapes on popular topics of physics and astronomy (+ others) popularized.
http://www.cite-sciences.fr

🇫🇷 On Canal Académie, academics, specialists in physics, mathematics, humanities, philosophy, sociology, law and jurisprudence, economics, politics and finance, history, geography and demography and political science will share their thoughts on the news and society developments.
http://www.canalacademie.com

And here is a small and non-exhaustive list of scientific YouTube channels that are scientifically quite reliable, managed by passionate individuals and of good post-production quality and especially very useful to understand better some of the complicated theorems and subjects treated in this textbook (most videos are of high-school level):

- https://www.youtube.com/user/ScienceEtonnante (en, David Louapre)
- https://www.youtube.com/user/Micmaths (fr, Mickaël Launay)
- https://www.youtube.com/user/physicswoman (en, Dianna Cowern)
- https://www.youtube.com/user/1veritasium (en, Derek Muller)
- https://www.youtube.com/channel/UCYO_jab_esuFRV4b17AJtAw (en, Grant Sanderson)
- https://www.youtube.com/user/fauxsceptique(fr, Christophe Michel)
- https://www.youtube.com/channel/UC7_gcs09iThXybpVgjHZ_7g (en, Matt O'Dowd)
- https://www.youtube.com/user/mathdude2012 (en, Andrey Kopot)
- https://www.youtube.com/user/LearnEngineeringTeam (en, Sabin Mathew)
- https://www.youtube.com/@jkzero (en, Dr. Jorge S. Diaz)

21.6. Other Sciences

✔✉➥✌☆🇫🇷 Futura-Sciences is website for the popularization of pure and exact sciences with news about sciences and technologies, issues on many topics, scientific forums with debate and discussions ... (the physics forum is especially well attended).
http://www.futura-sciences.com

✔✉➥✌☆ Physics Forum is consider as the world's largest physics community. There is a ton of discussions on many subjects answered with quality by the community. This forum is also considered as the partner forum of this e-book (even if they consider my book as spam...) as I don't have time to answer to questions anymore for free by e-mail.
http://www.physicsforums.com

✔✉➥✌☆ Stack Exchange is a network of question-and-answer websites that cover a broad range of topics in various fields, including mathematics, physics, programming, and many others. It was created in 2008 and has since become one of the largest online communities for people to ask and answer questions. Each website in the network is centered around a specific topic, and users can ask and answer questions, vote on the best answers, and even edit each other's responses. The network also includes features such as badges, reputation points, and a system for suggesting edits, which incentivize users to contribute high-quality content.
http://www.stackexchange.com

✔✉➥✌☆ Stack Overflow is one of the largest and most popular websites in the Stack Exchange network. It was launched in 2008 and is focused on programming and software development-related questions and answers. The website provides a platform for developers and programmers to ask and answer technical questions, share their knowledge and experience, and find solutions to coding problems. It has become a go-to resource for many developers, as it has a vast collection of questions and answers on a wide range of programming topics, from basic coding concepts to complex algorithms. Additionally, the website has a reputation and voting system that allows users to upvote the most helpful answers, helping to ensure that the best answers rise to the top.
http://www.stackoverflow.com

✔✍✉✄➥✌♥☆🇫🇷 Astrosurf is a portal of links (directory) for French amateur astronomers. There is also astronomy forums astronomy, announcements, free astronomy websites hosting, ephemeris and all the clubs and associations of French astronomy. You can particularly found the famous site of Thierry Lombry (here).
http://www.astrosurf.com

✔➥🇫🇷 The C.E.A. is the French Commissariat à l'Énergie Atomique. The site offers news and interesting files on particular areas of advanced physics. One can also find free educational resources for teachers (slide shows and posters).
http://www.cea.fr

✌☆ Here is a site that, if I was a child, would have made me happy ... and the misery of my

parents. It's not so much the content that is interesting but especially the online store (eStore) which offers a hundred gadgets and educational games for passionate of science. Attention to not spend too much!
http://www.xump.com

✌☆ For years, the cartoons of S. Harris have added humor to innumerable magazines, books, newsletters, ads and web sites in the field of sciences. Enjoy!
http://www.sciencecartoonsplus.com

21.7. Softwares/Applications

✔♥☆ Rosetta Code is a programming chrestomathy site. The idea is to present solutions to the same task in as many different languages as possible, to demonstrate how languages are similar and different, and to aid a person with a grounding in one approach to a problem in learning another. Rosetta Code currently has 861 tasks, 208 draft tasks, and is aware of 671 languages.
https://rosettacode.org

✔✄♥☆ TeXmaker is a free, cross-platform LaTeX editor. LaTeX is a typesetting system that is widely used in academia and scientific publishing to produce high-quality, professional-looking documents, particularly in the fields of mathematics, physics, and computer science. TeXmaker provides a user-friendly interface for creating and editing LaTeX documents, making it easier for users to get started with LaTeX. Some of its features include a built-in PDF viewer, multi-language support, and a wizard for creating new documents. TeXmaker is also highly customizable, with options for configuring keyboard shortcuts, syntax highlighting, and more. It is a popular choice among LaTeX users due to its ease of use, customizable interface, and support for a wide range of operating systems, including Windows, macOS, and Linux.
http://www.xm1math.net

✔✄♥☆ Matcha LaTeX Editor is a web-based LaTeX editor and provides a user-friendly interface for creating and editing LaTeX documents, making it easier for non-experts to get started with LaTeX. Some of its features include real-time collaboration, live preview, and templates for various document types. It's drawing tool that generates TikZ code is far above its competitors (LaTeXDraw, Overleaf, etc.).
https://www.mathcha.io

✔✄♥☆ Mathpix is an OCR (Optical Character Recognition) that allows users to take a photo of a handwritten or printed math equation, table or molecule and convert it into a LaTeX expression, which can then be edited and rendered in various formats. The tool is designed to help students, teachers, and researchers save time and effort when working with mathematical expressions and equations, and to eliminate the need to manually transcribe equations into a typesetting program.
https://mathpix.com

✔✄♥☆ The Integral Calculator lets you calculate integrals and antiderivatives of functions online and Derivative Calculator lets you calculate derivatives of functions online — for free! Both websites show the full working (step by step integration or step by step derivation). All common integration and derivation techniques and even special functions are supported. The Integral Calculator supports definite and indefinite integrals (antiderivatives) as well as integrating functions with many variables. The Derivative Calculator supports computing first, second, . . . , fifth derivatives as well as differentiating functions with many variables (partial derivatives), implicit differentiation and calculating roots/zeros. You can also check your answers! Interactive graphs/plots help visualize and better understand the results.
https://www.integral-calculator.com + https://www.derivative-calculator.net

✔✄♥☆ Minitab is the leading software for engineers working in industry and applying

statistical process control (see section Industrial Engineering page 8673) as part of their work or making any statistical study in the context of R& D.
http://www.minitab.com

✔✄♥☆ Isograph is an excellent suite of software for engineers working in industry and applying the techniques of preventive maintenance (reliability, FMEA) and decision support (see section Game and Decision Theory page 7376).
http://www.isograph-software.com

✔✄♥☆ The ReliaSoft company develops in my opinion the best softwares on the market for reliability and quality assurance engineers. Their software is used by the best engineering companies worldwide. Overall their software are real small diamonds (particularly Weibull++).
http://www.reliasoft.com

✔✄☆ JMP is for me the best known software for design of experiments at this date. His pedagogical structured and the very good available documentation (books) with mathematical proofs of the approach used by the software is well appreciated.
http://www.jmp.com

✔✄♥☆ Official website of the excellent Maple symbolic computation software used for many examples in the different chapters of this textbook. The website also offers lots of documentation and plugins to download.
http://www.mapleapps.com

✔✄☆ Official website of MATLAB® calculation and simulation software. I'm not especially a fan but I must admit this is an essential tool in some businesses, especially for the SimuLink part and R&D Engineering. This is somewhat a "must have" for engineers as well as LabView.
http://www.mathworks.com

✔✄☆ Official website of the company Statistica. A worldwide reference in the field of statistical analysis, data mining and statistical process control that stands a priori in a handkerchief with IBM SPSS.
http://www.statsoft.com

✔✄♥☆ COMSOL Multiphysics (formerly FEMLAB) is an excellent interactive environment for the modelling of industrial and scientific applications based on partial differential equations (PDEs) using finite element methods (more easy to use than ANSYS). A must have for engineers and researchers (for those who can afford to buy the license of course ..)!
http://www.comsol.fr

✔✄♥☆ Palisade @Risk is an add-in suite for Microsoft® Excel® and Microsoft Project for probabilistic simulation (Monte Carlo and Latin Hypercube). It's joint integration with Microsoft Project and Microsoft® Excel® and its calculations modules based on genetic algorithms, neural networks and its add-in of decision theory makes it a coveted tool for high executive officers of large companies (NASA and Lockheed Martin use it for major projects) and high level engineers in the field finance, quality, project management and production.
http://www.palisade.com

✔✄♥☆ The Decision Analysis software edited by TreeAge (DATA) allows users to build, analyse and distribute tree analysis, Markov models and influence diagrams. The DATA models incorporate visually quantitative and qualitative aspects related to business decisions to provide a tool to assist in projects during analysis of complex risks.
http://www.treeage.com

✔✄♥☆ MathType is a powerful equation editor (software used for this website) to import/export MathML or TeX or you can run in it stand-alone application. With a toolbar, it also fits perfectly with the programs of the Microsoft Office 2003-2013 suite (Word, Excel, PowerPoint), but also with 500 other softwares (specifically math softwares, or not). At the difference with LaTeX2ε you will not loose hours to solve compatibility and compilation issues.
http://www.wiris.com

✔✄☆ Scilab (contraction of Scientific Laboratory) is a free software, developed at INRIA Rocquencourt (France). It is a numerical computing environment that enables fast numeric resolutions and graphics commonly encountered in Applied Mathematics for people or small business having not enough money to buy MATLAB®.
http://www.scilab.org

✄☆ Official website of the famous dynamic geometry software: Cabri (worldwide reference in this domain) primarily destinated for the learning of geometry in schools. It can animate geometric figures, unlike those drawn on the blackboard. It comes to plane geometry or for geometry in 3D. It is the ancestor of all dynamic geometry software.
http://www.cabri.com

✄♥☆🇫🇷 Website of a math teacher who developed a very good freeware for many relevant charts, integration calculations, simulations of statistics and probabilities easily and playfully in the classroom (College level).
http://www.patrice-rabiller.fr

✄☆ ACDLabs develops the excellent Chemsketch software for molecules modelling and design (2D, 3D) with access to certain physical and chemical properties. This is a very useful tool also to prepare handouts or scientific articles (publications).
http://www.acdlabs.com

✄☆♥ MolView is an intuitive, Open-Source web-application to make molecular chemistry science more awesome!
http://molview.org

✄☆ Maxima is a computer algebra freeware (a Maple like software), falling under GNU GPL under the package Macsyma, the symbolic calculation software originally developed for the needs of U.S. Department of Energy. Maxima can do arithmetic, polynomials, matrices, integration, derivation, the calculation of series, limits, resolutions of systems of differential equations, etc.
http://maxima.sourceforge.net

✄☆ SPSS (Statistical Package for the Social Sciences) is a priori the most widely used software in companies in Switzerland for statistical analysis of data because it has an impressive

number of tests as well as many business packages. Because of its price and its owner (IBM) can be regarded as the premium solution of Statistics.
https://www.ibm.com/products/spss-statistics

✄♥☆ `R` is a free powerful language and environment for statistical computing and graphics. R provides a wide variety of statistical (linear and non-linear modelling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible. One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed.
http://www.r-project.org

✄☆ LabVIEW (Laboratory Virtual Instrument Engineering Workbench) is a powerful visual tool (graphical programming language) developed by National Instruments to control measures instrument or robots from a PC and used by a lot of companies worldwide (particularly by engineers).
http://www.ni.com/labview

✄♥☆ Symbolab is an advanced math education tool. It allows users to learn, practice and discover math topics using mathematical symbols and scientific notations as well as text. Symbolab provides automated step by step solutions to algebraic, trigonometric and calculus topics covering from middle school through college. Symbolab offers a wealth of smart calculators including: equations, simultaneous equations, inequalities, integrals, derivatives, limits , tangent line, trigonometric equations, functions and more. The stated goal of the site is to make scientific content universally accessible by expanding the searchable data space onto scientific notations, expressions, equations and formulas
https://www.symbolab.com

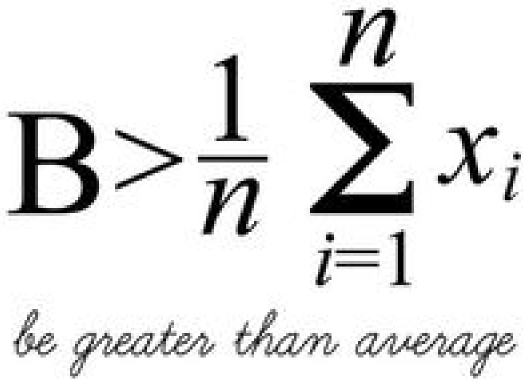
And finally a small list of some great "scientific" apps for iPhone/iPad:

- **Microsoft OneNote**: To take notes and write equations in pseudo LaTeX easily and sync them with your computer and the cloud.
- **iAnnote**: Great PDF reader with a lot of functions to annotate and modify your textbooks and scientific publications and handle huge books.
- **WeatherProHD**: For people that like meteorology and weather forecasts presented and beautiful and professional way with a lot of options.
- **Solar Weather**: For people interested to know our Sun activity based on the K-index.
- **Earthquake**: For people interested to know where, at what depth and at what magnitude an earthquake happened anywhere in the world in real-time.
- **Volcanoes**: For people interested to know what kind of volcanic eruption and with all most important details happened around the world in real-time!
- **Earthnow**: This NASA's app helps to visualizes recent global climate data from Earth Science satellites, including surface air temperature, carbon dioxide, carbon monoxide, ozone, and water vapour as well as gravity and sea level variations.
- **Analyser**: For those that want to run R or Python™ on their iPad and iPhone. Incredible app to do data science while travelling!
- **TeX Writer**: So far the best LaTeX app we know so you can write equations and documents on your mobile devices!
- **Telesat**: To identify or anticipate in the night sky the pass trough of some well-known satellites.
- **ISS Detector**: Same idea as for Telesat but for the International Space Station.
- **Flightradar24**: Same ideas as for Telesat and ISS Detector but for almost all civilian airplanes (but the free version show the airplanes with a small shift for security reasons).
- **Sun Info**: An app that helps you to anticipate the position of the Sun and Moon on many years and also summer/winter solstice, next eclipses and so on...
- **Stellarium**: As its name indicates it, it is a beautiful real time and complete stellarium to gaze at the stars, galaxies, constellations and also some satellites.
- **C. Anatomy 19**: An app (quite expensive with monthly/yearly subscription) to discover in high 3D resolution the inside of the human body.
- **EarthViewer**: An app to discover and play with the continents drifts of Earth across the time.
- **MathStudio**: A very powerful arithmetic and algebraic calculator that has calculation options similar to that Maple.

- **PCalc Lite**: A very nice an intuitive complete arithmetic calculator (but obviously less powerful than MathStudio!).
- **Molecules**: A 3D molecules viewer where you can download molecules from an online database.
- **E Numbers**: E numbers gives you the codes for substances that are permitted to be used as food additives.
- **iCircuit**: A nice app to create simple electronic circuits and simulate them! Very good to learn the basics of electronics while in holidays without the need of real physical components.
- **Electronic TB**: An app that provides a quite important (exhaustive) list and calculators for many electronic components, including many of their technical properties!
- **Symbolab**: This is just the app equivalent of the already previously mentioned website of the same name!

22

Quotes



We summarize here the different quotes that are in the various chapters of this textbook. We added also below some quotes that may be useful when debating some scientifically illiterate people who use quotes to back-up their claims (as these people assume that a quote has a scientific evidence level other than zero...).

> *If you steal from one author, it's plagiarism; if you steal from many, it's research!*
>
> Wilson Mizner

> *We are trying to prove ourselves wrong as quickly as possible, because only in that way can we find progress!*
>
> James D. Eckard

> *In mathematics, you don't understand things. You just get used to them.*
>
> John von Neumann

> *The statements of science are not what is true and what is not true, but statements of what is known with different degrees of certainty.*
>
> Richard Feynman

> *You can't convince a believer of anything; for their belief is not based on evidence, it's based on a deep seated need to believe.*
>
> Carl Sagan

> *The Universe is not fine-tuned to us: We are fine-tuned to our particular Universe!*
>
> Victor J. Stenger

> *We know very well the wisdom of crowds! It was the one who elected Hitler in 11933 (holocene calendar) and Trump in 12016 (holocene calendar).*
>
> ?

> *Since Newton, mankind has come to realize that the laws of physics are always expressed in the language of differential equations.*
>
> Steven Strogatz

> *All things are poisons, for there is nothing without poisonous qualities. It is only the dose which makes a thing poison!*
>
> Paracelsus

> *Men never do evil so completely and cheerfully as when they do it from religious conviction!*
>
> Blaise Pascal

> *Understanding how a rainbow works doesn't make it less beautiful, it makes it more beautiful. That is Science!*
>
> Richard Dawkins

> *If you can explain a God without a creator, you can explain a Universe without a creator...*
>
> ?

> *The best way to learn is to teach.*
>
> Frank Oppenheimer

> *Without faith, God is nothing. Without science, man is nothing.*
>
> John Aubrey

> *If we could do things by common sense, we wouldn't need physicists.*
>
> Richard Dawkins

> *Intelligence is not what we know but what we do when we don't know.*
>
> Jean Piaget

> *If a religion had even one true evidence beyond reasonable doubt, there would be no other religions!*
>
> ?

> *Religion is an insult to human dignity. With or without it, you'd have good people doing good things and evil people doing evil things. But for good people to do evil things, it takes religion!*
>
> Steven Weinberg

> *With four parameters I can fit an elephant.*
>
> John Von Neumann

> *A scientist will read dozens of books in his lifetime, but still known that he has a lot more to learn. A religious fanatic barely reads one book, and think that they know it all.*
>
> ?

> *I think of my lifetime in physics as divided into three periods. In the first period... I was in the grip of the idea that Everything is Particles... I call my second period Everything is Fields. ... Now I am in the grip of a new vision, that Everything is Information.*
>
> John Archibald Wheeler

> *Religion is the practice of using nonsense, in order to explain ignorance.*

?

> *Nothing can create something all the time due to the laws of quantum mechanics, and it's fascinatingly interesting!*

Lawrence M. Krauss

> *Let no one ignorant of geometry enter here!*

Plato

> *Everything is just some form of first order Taylor Expansion!*

?

> *Computers are like humans - they do everything except think.*

John von Neumann

> *SCIENCE is not defined by what we believe. IT IS NOT a belief system. It has no religious doctrine! It's just an intellectual investigation method that rejects claims that have no supporting experimental reproducible evidence.*

?

> *Extraordinary claims require extraordinary evidence.*

Carl Sagan

> *I have your letter and the best thing I can do is refer to my published works, both scientific and popular. The creationist movement is lead by a dishonest bunch of operators and misquotation is the hall mark of their work. Responding to them is time wasting and a letter would not be adequate to put your questions to rest. There are some things best ignored and the stupidity of these so called religious fanatics continues to astonish me.*

Richard Leakey

> *Most people don't really want the truth; they just want reassurance that what they already believe, is the truth.*

L. Aron Nelson

> *The reason there is a conflict between science and religions is that science keep disproving the things religions claims to be true or accurate.*
>
> ?

> *He who loves practice without theory is like the sailor who boards ship without a rudder and compass and never knows where he may cast.*
>
> Leonardo da Vinci

> *The Earth is not fitted to us, we are fitted to the Earth.*
>
> ?

> *We measure the degree of intelligence of an individual by the quantity of incertitudes he can handle.*
>
> Emmanuel Kant

> *I don't even like the multiverse but if Nature tells me that's the case, and the laws of physics are accidental, I've got to live with it.*
>
> Lawrence M. Krauss

> *Insurance is when you buy coverage when you need it. Reality is when you buy the option of going to court for coverage when you need it.*
>
> ?

> *Physics is like sex. Sure it may have some practical results, but that's not why we do it.*
>
> Richard Feynman

> *Mathematics is not only real, but it is the only reality. That is that entire universe is made of matter, obviously. And matter is made of particles. It's made of electrons and neutrons and protons. So the entire universe is made out of particles. Now what are the particles made out of? They're not made out of anything. The only thing you can say about the reality of an electron is to cite its mathematical properties. So there's a sense in which matter has completely dissolved and what is left is just a mathematical structure.*
>
> Martin Gardner

> *Each generation of scientists stands upon the shoulders of those who have gone before.*

Owen Chamberlain

> *In precisely built mathematical structures, mathematicians find the same sort of beauty others find in enchanting pieces of music, or in magnificent architecture. There is however, one great difference between the beauty of mathematical structures and that of great art. Music by Mozart, for instance, impress greatly even those who do not know musical theory; the cathedral in Cologne overwhelms spectators even if they know nothing about Christianity. The beauty in mathematical structures, however, cannot be appreciated without understanding of a group of numerical formulae that express laws of logic. Only mathematicians can read "musical scores" containing many numerical formulae, and play that "music" in their hearts.*

Kiyoshi Ito

> *In physics anything that can happen (meaning has a non-zero probability of occurring) eventually will happen!*

?

> *You don't have to burn books to destroy a culture. Just get people to stop reading them.*

Ray Bradbury

> *Science doesn't know everything, religion doesn't know anything!*

L. Aron Nelson

> *The science of today is the technology of tomorrow.*

Edward Teller

> *Science is different to all the other systems of thought... because you don't need faith in it, you can check that it works!*

Brian Cox

> *Ignorance more frequently begets confidence than does knowledge: it is those who know little, not those who know much, who so positively assert that this or that problem will never be solved by science.*

Charles Darwin

> *Differentiation is a science, Integration is an art.*

?

> *Finding proofs of your delusions in simple coincidences of life is a very desperate attempt of conviction...*

Martino Villadari

> *The fact that a believer is happier than a septic is no more to the point than the fact that a drunken man is happier than a sober one.*

George Bernard Shaw

> *I don't want to believe, I want to know.*

Carl Sagan

> *Science knows it doesn't know everything, otherwise it would stop.*

Dara Ó Brian

> *Science is just imagination in straitjacket.*

Richard Feynman

> *The greatest enemy of knowledge is not ignorance, is the illusion of knowledge!*

Stephen Hawking

> *Without data you are just another person without opinions, but without mathematics you are just a data collector.*

?

> *I have no special talents, I am only passionately curious.*

Albert Einstein

> *People keep saying "Science doesn't know everything!". Well, science know it doesn't know everything; otherwise it would stop.*

Dara Briain

> *The only way to learn Mathematics is to do Mathematics.*
>
> Paul Halmos

> *The scientist is not a person who gives the right answers; he is one who asks the right questions.*
>
> Claude Levi-Strauss

> *One can't prove that God doesn't exists, but Science makes God unnecessary.*
>
> Stephen Hawking

> *The power of numbers was much more respected among us when we knew nothing about them..*
>
> Voltaire

> *To reflect is to disturb one's thoughts.*
>
> Jean Rostand

> *If you are stupid how can you possibly realize that you are stupid?.*
>
> John Cleese

> *If you don't make mistakes, you're not working on hard enough problems. And that's a big mistake.*
>
> Frank Wilczek

> *Success relies on the success of those who have preceded us.*
>
> ?

> *Between quantity and quality there is geometry.*
>
> Brigitte Le Roux, Henry Rouanet

> *Practice should always be based upon a sound knowledge of theory.*
>
> Leonardo da Vinci

> *It doesn't matter how beautiful your theory is, it doesn't matter how smart you are! If it doesn't agree with experiment, it's wrong!*
>
> Richard Feynman

> *To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of...*
>
> Ronald Fisher

> *Errors using inadequate data are much less than those using no data at all!*
>
> Charles Babbage

> *Details make perfection, and perfection is not a detail.*
>
> Leonardo da Vinci

> *Ignorance is a virus. Once it starts spreading, it can only be cured by reason. For the sake of humanity, we must be that cure!*
>
> Neil Degrasse Tyson

> *The art of doing mathematics consists in finding that special case which contains all the germs of generality.*
>
> David Hilbert

> *If you think education is expensive, try ignorance.*
>
> Derek Bok

> *Tell me and I forget, teach me and I remember, involve me and I learn.*
>
> Benjamin Franklin

> *An expert is a person who has found out by his own experience all the mistakes that one can make in a very narrow field.*
>
> Niels Bohr

> *What you can imagine depends on what you know.*
>
> Daniel C. Denett

> *Any sufficiently advanced technology is indistinguishable from divinity.*
>
> Arthur C. Clarke

> *What one man calls God, another call it the Laws of Physics.*
>
> Nikola Tesla

> *The only person you need to be better than is the person you were yesterday!*
>
> ?

> *Great minds discuss ideas. Average minds discuss events. Small minds discuss people.*
>
> Eleanor Roosevelt

> *Be less curious about people and more curious about ideas.*
>
> Marie Curie

> *Creativity is intelligence having fun.*
>
> Albert Einstein

> *With enough of us, around the world, we'll not just send a strong message opposing the privatization of knowledge - we'll make it a thing of the past.*
>
> Aaron Swartz

> *The idea of a personal God is quite alien to me and seems even naive.*
>
> Albert Einstein

> *When you move your focus from competition to contribution life becomes a celebration. Never try to defeat people, just win the hearts.*
>
> Buddha

> ❞ *Math is the language of the Universe. So the more equations you know, the more you can converse with the cosmos.* ❞

Neil Degrasse Tyson

> ❞ *The whole trouble with the world is that the stupid are cocksure, and the intelligent are full of doubt.* ❞

Bertrand Russell

> ❞ *Some men see things as they are and ask why. Others dream things that never were and ask why not.* ❞

Gerge Bernard Show

> ❞ *Statistics is the grammar of science.* ❞

Karl Pearson

> ❞ *If you can't explain it to a six year old, you don't understand it well enough yourself.* ❞

Albert Einstein

> ❞ *Man is the most insane species. He worships an invisible God and destroys a visible Nature. Unaware that the Nature he is destroying is this God he is worshipping.* ❞

Hubert Reeves

> ❞ *If you can't measure it, you can't improve it!* ❞

Willima Thomson (Lord Kelvin)

> ❞ *Once you stop learning, you start dying.* ❞

Albert Einstein

> ❞ *There is only one good, KNOWLEDGE, and one evil, IGNORANCE.* ❞

Socrates

> *The purpose of argument, should not be victory, but progress.*

?

> *Science is the poetry of reality!*

Richard Dawkins

> *Real science can be far stranger than science fiction and much more satisfying.*

Sthepen Hawking

> *You don't understand the basic assumptions of your own culture if your own culture is the only culture you know.*

Alan Watts

> *The ones who are crazy enough to think that they can change the world are the ones who do!*

Elon Musk

> *The noblest pleasure is the joy of understanding!*

Leonardo da Vinci

> *Doubt is the father of invention!*

Galileo Galilei

> *It is easy to lie with statistics, but easier to lie without them.*

Fred Mosteller

> *What can be asserted without evidence can be dismissed without evidence.*

Christopher Hitchens

> *Religion is a culture of faith; science is a culture of doubt.*

Richard Feynman

> *The sciences do not try to explain, they hardly even try to interpret, they mainly makes models. By a model is meant a mathematical construct which, with the addition of certain verbal interpretations, describes observed phenomena. The justification of such a mathematical construct is solely and precisely that it is expected to work-that is, correctly to describe phenomena from a reasonable wide area.*

John von Neuman

> *You can beat 100,000 scholars with one fact, but you will very likely never beat an idiot even with 100 facts.*

?

> *No amount of peer-reviewed evidence will ever persuade an idiot, all the more if the latter is scientifically illiterate and doesn't master top level statistical calculations and the subtleties of the scientific method.*

?

> *The most important questions in life are indeed, for the most part, really only problems of probability.*

Pierre Simon de Laplace

> *The problem in society is not kids not knowing science. The problem is adults not knowing science. They outnumber kids 5 to 1, they wield power, they write legislation. When you have scientifically illiterate adults, you have undermined the very fabric of what makes a nation wealthy and strong.*

Neil deGrasse Tyson

> *Even a rough estimate, even a false one, is better than no estimate at all.*

?

> *Sooner or later the moment arrives when a decision must be made which is neither certain, neither suitable, nor popular, but it must be taken, because it is right.*

Martin Luther King

> *We have uncovered wonders undreamt by our ancestors who first speculated on the nature of those wondering lights in the night sky. We've crossed the solar system and sent ships to the stars. But we continue to search. We can't help it! A central element of the human future, lies far beyond the Earth. If we crave some cosmic purpose, then let us find ourselves a worthy goal...*

Carl Sagan

> *The way to cope with the future is to create it!*

Ilya Prigogine

> *Science doesn't tell us what to think but sometimes it tells us what we can no longer believe.*

Etienne Klein

> *A religious belief without dogma is not possible and science is not compatible with any dogma. So science can only be incompatible with religious beliefs.*

?

> *Philosophy debates God, maths ignores him, climatology doesn't need him, physics can't find him, biology contradicts him, logic kills him!*

Bill Flavell

> *Being a statistician means never having to say you are certain!*

?

> *If you are not a professional active scientist, and you disagree with scientists about Science, it's actually not a disagreement! You're just wrong!! Science is not truth. Science is finding the evidence. When science changes its opinion, it didn't lie to you. It learned more!*

?

> *Logic is the art of going wrong with confidence.*

Joseph Wood Krutch

> *Most of our so-called reasoning consists in finding arguments for going on believing as we already do.*

James Harvey Robinson

> *Truth is a word that oddly is used most by those who spread lies, that they themselves know to be lies!*

?

> *As a scientist (and atheist) you will often be hated by all those who believe in fairy tales and think that their opinions, anecdotes, circular logic and scientific illiteracy are worth it, but those who stands firm to the Scientific Method and to the precautionary principle will be saved and will help to save the haters. Comforting lies have always and are still taking the elevator and the Scientific Method takes the stairs, the lies will come first but one day or another the Science will always catch up with its unpleasant evidence.*

> *Not doubt, certainty is what drives one insane.*

Friedrich Nietzsche

> *Probability theory is nothing but common sense reduced to calculation.*

Pierre Simon de Laplace

> *The true logic of this world is in the calculus of probabilities.*

James Clerk Maxwell

> *When people say "The reason you can't trust science is that scientists used to say <thing that turned out be wrong> but now they say <different thing>", they are actually explaining why science works so well!*

Matt Blaze

> *Nature wants* 5 *of your* 7 *children dead. It wants you dead by* 50*. Everything better than that is brought to you by science and technology.*

David Frum

> *In science nothing can be proven at* 100% *but* 98% *of statistical evidence in science will always be better than the* 0% *evidence of any of the* 28,000 *religions and their* 3,000 *gods.*

?

> *All science is either physics or stamp collecting!*

Ernest Rutherford

> *Ce que nous observons, ce n'est pas la Nature en soi, mais la Nature exposée à notre méthode d'investigation.*

Werner Heisenberg

> *The Art and Science of asking questions is the source of all Knowledge.*

Thomas Berger

> *Mathematics is not a deductive science - that's a cliché. When you try to prove a theorem, you don't just list the hypotheses, and then start to reason. What you do is trial and error, experimentation, guesswork. You want to find out what the facts are, and what you do is in that respect similar to what a laboratory technician does!*

Paul Halmos

> *A theory which is not refutable by any conceivable event is non-scientific. Irrefutability is not a virtue of a theory (as people often think) but a vice.*

Karl Popper

> *What causes the universe to pop out of nothing? No cause is needed. If you have a radioactive atom, it will decay, and quantum mechanics gives the decay probability in a given interval of time, say, a minute. There is no reason why the atom decayed at this particular moment and not another. The process is completely random. No cause is needed for the quantum creation of the universe.*

Alexander Vilenkin

> *There is a fundamental difference between religion, which is based on authority, [and] science, which is based on observation and reason. Science will win because it works.*

Stephen Hawking

> ❞ *Science is a differential equation. Religion is a boundary condition.* ❞
>
> Alan Turing

> ❞ *The discovery that individual events are irreducibly random is probably one of the most significant findings of the twentieth century. ... for the individual event in quantum physics, not only do we not know the cause, there is no cause! The instant when a radioactive atom decays, or the path taken by a photon behind a half-silvered beamsplitter are objectively random.* ❞
>
> Anton Zeilinger

> ❞ *The emergence of life is ultimately a chemical process. If the conditions are right, life will emerge. I see no need for an initial push from a creator god.* ❞
>
> Didier Queloz

quote 22.0:9346.#94

> ❞ *Everything is a sign or miracle when you don't understand Physics.* ❞

> ❞ *The question of the truth of a religion is one thing, but the question of its usefulness is another. I am as firmly convinced that religions do harm as I am that they are untrue.* ❞
>
> Bertrand Russell

23

Changelog

This is a detailed changelog of the book for people interested to see how this textbook has evolved (to keep in touch with new releases send us an e-mail to subscribe to the newsletter), of course the years are based on the Holocene calendar:

- **May 12002**
    - Definitions (science, law, theorem, postulate, axiom, corollary, ...)
    - Operations of addition, subtraction, multiplication, division, power
    - Concept of Numbers (integers, relative real, fractional, complex algebra, abstract,...)
    - Domain of definition of variables
    - Arithmetic polynomials
    - Absolute value
    - Binary relations, order relations
    - Gamma Function Euler, Euler's constant
    - Electromagnetic wave equation, the wave speed, speed of the light, transported energy
    - Introduction to Optics
    - Rule of three, percentages
    - Net quantities, the cost price of purchase, indices, sale prices, brut prices, net profit, net price, assets
    - Simple and compound interest, late and early payments
    - Portfolio management (decision criteria, goodwill, return on investment)
    - Markowitz Model (utility function, selection criteria)
    - Vocabulary about Stock Exchange
    - Plasma Frequency
    - Probability (universe, events, axioms)
    - Combinatorial Analysis
    - Statistics (discrete variables, continuous variable, standard deviation, variance and covariance)
    - Distribution Functions (discrete uniform law, Bernoulli's law, binomial law, hypergeometric law, multinomial law, Poisson's law, Gauss-Laplace law, Cauchy's law, beta law, gamma law, chi-square law, Student's law)
    - Estimator, correlation

  - Matrix of covariance
  - Statistical adequation tests

- **June 12002**
  - Introduction Set Theory (Zermelo-Fraenkel theory, inclusion, complementarity, intersection, union, product, empty set
  - Arithmetic, harmonic, geometric, quadratic averages
  - Univocity
  - Logarithmic and exponential functions
  - Golden ratio
  - Introduction to Fourier series
  - Coulomb's law, electrostatic field, electrostatic potential
  - Ampere's Theorem
  - Biot-Savartlaw, magnetic field, magnetic induction
  - Maxwell's equations
  - Dimensions in euclidean and fractal geometry
  - One-dimensions, two-dimension, three-dimension Geometric shapes
  - Pythagorean theorem
  - Physical units systems
  - The principle of least action and energy conservation
  - Position, velocity and acceleration
  - Continuity equation
  - Bernoulli Equation
  - The Doppler effect
  - Restrained Relativity, invariance principle, Lorentz transformations (time, length, velocity addition, increase in mass, Minkowski space-time)
  - General relativity (world line, punctual event, inertial particles, null cones, space-time vectors spaces and curved planes, metric tensor)
  - Heisenberg principles of uncertainty, Schrödinger equation, Wave probability density
  - Introduction to superstrings

- **July 12002**
  - Creating of Table of Contents
  - Visual representations of functions
  - 2nd degree polynomials and roots
  - Operator of vector and scalar fields (gradient, nabla, divergence, curl, laplacian)
  - Vector Analysis (concept of arrow, set of vectors, scalar multiplication, vector space, linear combinations, generating families, bases of a vector space)
  - Tensor calculus (Einstein convention, Kronecker symbol, anti-symmetry symbol)
  - Notations for grouped multiplications (Big Sigma) and sums (Big Pi)
  - Axioms for set of real number
  - Definition of various Inequalities
  - Circular and related movements
  - Inertial forces (Coriolis force)
  - Drake Equation

- **August 12002**
  - New biographies (Hilbert, Riemann, Legendre)
  - New section "Humor"
  - Introduction to Topology
  - Euclid postulates
  - Gauss plane (complex numbers)
  - De Moivre’s formula
  - Transformations in the complex plane
  - Base change and tensor scalar product
  - Covariant and contravariant components
  - Relativistic transformations of the momentum
  - Prefixes of multiples and submultiples of SI units
  - Quality control (probabilities), efficiency curve, level value of acceptable quality (LQA)
  - Kepler’s laws
  - Proof of classical gravitational force from Kepler’s Law
  - Proof of area law (second Kepler’s Law)
  - Moment of force
  - Angular momentum
  - Theorem of angular momentum
  - Newton-Poisson Equation
  - Thomson’s, Bohr’s and Sommerfeld-Bohr’s atom model
  - Hydrogen spectrum
  - Assumption of Neutron
  - Quantum numbers
  - Pauli exclusion Principles
  - Classical Analysis of the Schrödinger equation for the ideal rectangular potential
  - Newton’s Laws
  - Proof of Newton’s third law from the principle of least action
  - Release speed (for rockets or others)
  - Definition of "Scientism"
  - Variation of the gravitational acceleration in and out of a homogeneous spherical body
  - The principle of least action and quantum physics (semi-classical limit)
  - Ballistics (maximum range, safety parabola)
  - Introduction to nuclear physics
  - Atomic number, mass number
  - Radioactivity, activity, decay chain, isotopes, isotones, nuclide dating
  - Atomic mass system (UMA)
  - Mass defaults
  - Fusion and fission
  - Alpha and beta (minus and plus) disintegration
  - Electron capture
  - Gamma emission

- **September 12002**
  - New Biographies (Dalton, Boltzmann, De Broglie, etc.)
  - Principle of good order
  - Archimedean Property
  - Induction Principle
  - Divisibility (Euclidean division)
  - Congruent numbers
  - Proof by nine
  - Numbers basis
  - Definition utility of thousands separators
  - Priorities of parenthesis, brackets, braces and arithmetic operators
  - Proof Theory (introduction)
  - Definition of terms, formulas and demonstrations
  - Definition of languages, symbols, relations and functions
  - Definition of periodic, compound, basic, rational, fractional, irrational, algebraic and transcendental numbers
  - Generalization of elementary algebra
  - Dimensions of a vector space, extension of a free family, rank of a finite family
  - Vector Hyperplane, direct sums
  - Almost rigorous definition of the concepts of line, surface (plane) and volume
  - Straight line intersection, half-lines, segments, aliquot part
  - Continuity axiom of the line and of the plane
  - Translation and rotation of plan
  - Angles, units, measurements, sides of the angle, sharp edges, flat angles, equal angles ,straight angles, acute angles, obtuse angles, supplementary angles, complementary angles,
  - Perpendicular lines, bisecting angle
  - Definition of work and energy: kinetic and potential energy theorem (motor work and resisting work)
  - Concept of conservative vector field
  - Conservation of energy and momentum
  - Mass center theorem
  - Relativistic force transformation
  - Relativistic transformations of electric and magnetic fields
  - Chandrasekhar limit weights (collapse limit of white dwarfs)
  - Definitions of optics, generalization of the law of refraction
  - Broglie normalization condition, linked and non-linked states
  - Harmonic oscillator
  - Quantum chemistry and molecular vibrations
- **October 12002**
  - New biographies (Cauchy, Neumann, Bessel, Archimedes)
  - New section on "Theoretical Computing"
  - Greatest common divisor, least common multiple
  - Rule signs (...)
  - Proof of the irrationality of a number
  - Introduction to arithmetic, harmonic and geometric sequences

  - Limits and continuity of functions
  - Definition of affine spaces
  - Introduction to the Euclidean tensors and their properties
  - Triangles and properties of triangles
  - Definition of thermodynamic systems
  - Definition of the reduced mass
  - Bessel's functions
  - Definition of the moment of inertia
  - Introduction to Quantum Field Theory
  - Introduction to radiation protection
  - Proof of Bethe-Bloch formula
  - Tunnel effect in quantum physics
  - Introduction Dirac's formalism
  - Heron and Archimedes algorithm's
  - Introduction to fractal sets
  - Introduction to game theory (cooperative games, earnings, payoff matrix, extensive forms, Pareto optimums, Nash equilibrium, evolutionary games)

- **November 12002**

  - Fundamental theorem of arithmetic
  - Introduction to cryptography (RSA, DES, MD5, SHA-1)
  - Euler phi function
  - Small Fermat theorem
  - Introduction to topological dimensions and scaling
  - Cosine directors

- **December 12002**

  - New Biographies (Nash, Cartan, Lucas, Lie)
  - Detailed development of the "Integer" function
  - Superposition of linear quantum states (quantum coherence)
  - Definition of Lipschitz functions and contracting functions
  - Definition of a convergent Cauchy sequence
  - Fixed point theorem (used in fractals, Newton methods and many others)
  - Definitions of reality, the problem of the theory and trial on reality
  - Definition of Euclidean space and Euclidean affine space
  - Definition of property concepts in the field of chemistry

- **January 12003**

  - Pascal's Theorem
  - Buoyancy (Archimedes' principle)
  - Simple Introduction to different symmetries in physics (temporal, spatial)
  - Simple Introduction to different transformations in the plane (translation, scaling, reflection, isometry, rotation)
  - Definition of an inverse and composed function/application
  - Trigonometry (introduction, remarkable relations/identities, spherical trigonometry)
  - Signature of a vector space
  - Schmidt orthogonalization methods, base changes, Fourier associated spaces

  - Everything (almost) on the plane and spherical trigonometry
  - Keplerian orbital trajectories
  - Introduction to the neoclassical monetary model (Say/Walras laws, homogeneity assumption)
  - Boolean algebra (simple properties and theorems)
  - Redesigned of quantum physics section (order of the subjects)
  - Proof of evolutionary Schrödinger equation
  - Proof of the relativistic evolutionary Schrödinger equation
  - Introduction to Antimatter theory

- **February 12003**
  - Proof of gradient, divergence, rotational and Laplacian in cartesian, polary, cylindrical and spherical coordinates
  - Complete mathematical developments of speed and acceleration expressions in cartesian, polar, cylindrical and spherical coordinates
  - Proof of the relativistic invariance of the electric charge (charge conservation equation)
  - Proof of the existence of anti-particle with opposite charge
  - Introduction to Gauge Theory (quadripotentiel Lorenz gauge, Coulomb gauge, Alembertian)
  - Introduction to Lagrangian and Hamiltonian formalism (generalized coordinates, configuration spaces, Euler-Lagrange equation, canonical formalism, Legendre transformation, Poisson brackets)
  - Rigorous definition of the principle of least action
  - Definition of Tensor spaces

- **March 12003**
  - New Biographies (Lorentz, Minkowski Hermann, Ricci-Curbastro, Levi-Civita)
  - Definition of Cartesian product and extension of the scope of application of Cardinals
  - Proof of Cauchy-Schwarz inequality
  - Proof of the triangle inequality
  - Definitions of the vector product and mixed product
  - Proof of condensed form the sum of the first n integers
  - Proof of the validity of the integration by parts
  - Definition of the algebraic vectorial structure and of an algebra
  - Definition of a homomorphism, isomorphism, endomorphism, automorphism

- **April 12003**
  - New Biographies (Göpper-Meyer, Yukawa, Nöther, Cournot)
  - Cournot Theory equilibrium (competition)
  - Wilson's model (inventory management)
  - Mathematics of phasers
  - Relativistic model of the Sommerfeld's atom
  - Analytical resolution of the Schrödinger equation
  - Heisenberg's principles of quantum uncertainty
  - Lagrangian formalism of quantum physics fields

  - Improvement of website forum (add mathematical symbols and external files)
  - 10 new links to interesting web pages (associations + math stuff)

- **May 12003**
  - New Biographies (Bell, Ramanujan, Landau)
  - Proof of the precession of the perihelion of coupled orbits of stars or electric charges
  - Definition and developments related to the Virial theorem
  - Calculation of potential energy of a material sphere (internal temperature of Stars)
  - Definition of prime numbers and proof that they are in infinite in number
  - Definition of a fully enclosed ring
  - Proof that a rational number is an algebraic number if and only if it is a relative integer.
  - Definition of a multi-linear application/function (or morphism of vector space)
  - Definition of a partition of a set and an equivalence class.
  - Descriptions of set operations of absorption and idempotence.
  - Definitions and examples of sagittals diagrams.
  - Definition of a magma and a monoid
  - Pseudo-proofs of algebraic structures of fundamental sets of arithmetic
  - Development of the theory of angular momentum in wave quantum physics
  - Definition of a Diophantine equation and sets of Fermat's Last Theorem

- **October 12003**
  - New Biographies (Abel, Banach, Boole, Bose, Brouwer, Clausius, Cayley, Curie, Connes, Dirichlet, Frege, Gibbs, Picard, Erdos, Grothendieck, Hamilton, Hausdorf, Heaviside, Helmholtz, Hermite, Hoyle, Jacobi, Klein, Kronecker, Langevin, Lee, Lobachevsky , Möbius, Monge, Fish, Schwartz, Shannon, Thom, van der Waals, Vieta, Weinberg, Witten, Gamow, Sturm, Liouville,Clairaut, Teller)
  - Definition of punctual spaces in classical mechanics, of writing conventions and changes in referential
  - Mathematical theory of projective perspective with vanishing points and definition of the projective and isometric perspectives
  - Simplistic definition of the concept of "derivative" in functional analysis.
  - Proof of some common derivatives (polynomials, composite functions, inverse functions, cosine, sine, arc sine, arc cosine, quotient of two functions, etc.).
  - Numerical Methods: mathematical explanation of the complexity of an algorithm and elementary optimization research
  - Method of calculating the number e
  - Resolution Method of n linear equations systems with n unknown using the pivot method
  - Search of function roots by the methods of the proportional parts, of the bisection method, of the secant methods (regula falsi) and Newton's method
  - Calculations of areas and integrals using the Riemann sums method
  - Presentation of the Monte Carlo calculation for integrals, or Pi
  - Mathematical development of simplex algorithm used in the operations research (linear programming)
  - Contraction of indices in tensor calculus
  - Definition and detailed properties of some special tensor (symmetric tensor, anti-symmetric tensor, fundamental tensor, etc.)

  - Curvilinear coordinates (determination of the metric and linear element of a spherical and cartesian space and of the plane in polar coordinates) Christoffel symbols (first and second type).
  - Proof of the Cantor-Bernstein theorem
  - Determination of the free generalized Lagrangian in General Relativity

- **December 12003**
  - New Biographies (Kirchhoff biographers, Markowitz, Cox Merton, Scholes, Sharpe, Ferdinan von Lindemann, Bachelier, Stefan)
  - Proof of one of the Stirling formulas
  - Detailed calculation of a simple model of a Star surface temperature
  - Detailed calculations of temporal properties of the securities values
  - Proof of Taylor series and Maclaurin (limited and unlimited)
  - Defining the Lagrange rest and of Alembert's, Cauchy, integral test and absolute convergence criteria
  - Developments relative to the definitions of brightness, luminosity, apparent and absolute magnitude of Stars and calculation of the distance to Cepheids
  - Definitions of the solid angle, the solid angle of revolution and the elementary solid angle
  - Definition of photometric and photonic and international system quantities
  - Definitions and developments related to the light intensity, energy flow (with proof of Beer-Lambert' law), emittance, radiance (with Lambert's law), Kirchhoff's law
  - Proofs of Stefan's law and Stefan-Boltzmann law
  - Resolutions of third degree polynomials by radicals (Cardan's method) and development about solving quadratic polynomials in the complex set
  - Definition of the concept of equations and inequalities
  - Determination of the Cartesian equation of the plane, line (in space),cone and sphere
  - Definition of market efficiency
  - Determination of the Black & Scholes equation
  - Presentation of the mathematical aspects of Wiener process
  - Ito lemma and Brownian motion (random walk)

- **January 12004**
  - Definition of the indefinite integral
  - Newtonian cosmological model (without the cosmological constant)
  - Statement of the Peano's
  - Introduction to Linear Algebra (Gauss reduction method, elementary operations between matrices)
  - Statement of 5 Euclid's axioms and 5 groups of axioms in geometry
  - Proof of relations for calculating the perimeter of the surface and the center of gravity of the square, the rectangle and triangle
  - Proof of relations for calculating the volumes and surfaces of the torus, the sphere, the ellipsoid, cylinder and cone
  - Definition of the centroid (center of gravity) and demonstration of four properties related thereto
  - Demonstration of decomposition odd function and pair of any function
  - Definition of hyperbolic trigonometric functions and enumeration of relations and proof of related remarkable properties

  - Introduction to differential geometry (definition of a Riemannian geometry, Frenet triad, parametrised surface, etc.)
  - Introduction to graph theory (Königsberg's bridges proof)
  - Definition of a topological space and Hausdorff definition of a metric / ultra-metric space and associated distances (hölder, discreet, equivalent ...)
  - Definition of a Lipschitz function (and related isometrics)
  - Definition of open and closed set (open/closed balls, spheres, adherence, Hausdorff's excess) and diameters
  - Definition of set distances (gap) and of a variety / map / atlas and differential homeomorphism

- **April 12004**
  - Proof of the Guldin's theorem
  - Proof of König's theorem of kinetic energy and angular momentum
  - Presentation and proofs of various techniques for calculating the moments of inertia: Huygens-Steiner theorem, polar inertia moment, inertia tensor, generalized Huygens-Steiner
  - Definition of "power" and "performance" and proof of the calculation of the power of a rotating machine
  - Proof of Boltzmann thermodynamic entropy law and the following statistical distributions: Maxwell speeds, Maxwell-Boltzmann, Fermi-Dirac, Bose-Einstein
  - Proof of some principal moments of inertia of the following bodies: torus, spheres, cones, rectangular plate, tube
  - Introduction to wave optics: Huygens principle, proof of Malus law, development of Fraunhofer model in the case of a rectangular slot.
  - Definition and proof of the resolving power of a simple rectangular slot.
  - Proof of the origin and the solution of the no less famous Bessel differential equation of order n
  - Proof of the wave function of a stretched rope and a stretched circular membrane
  - Proof of Planck's law and of known approximations (first Wien's Law, Rayleigh-Jeans law).
  - Proof of the displacement law (second Wien's law) and the Stefan-Boltzmann law through Planck's law and determination of the Stefan-Boltzmann constant
  - Study of the origin of Planck dimensions: Planck length, Planck mass, Planck density, Planck time, Planck energy

- **July 12004**
  - Proof of the physical origin of heat
  - Proof of Torricelli's theorem
  - Proof of Venturi effect
  - Proof of Poiseuille's law
  - Proof of relativistic Compton effect
  - Proof of the existence of the fossil radiation in the Universe
  - Introductions to quotients sets (in this case $\mathbb{Z}/n$
  - Proof of the physical origin of heat
  - Proof of Torricelli's theorem
  - Proof of Venturi's effect
  - Proof of Poiseuille's law

  - Proof of relativistic Compton effect
  - Proof of the existence of the fossil radiation in the Universe
  - Introductions to quotients sets (in this case $\mathbb{Z}/nZ$)
  - Proof of Lorentz transformations for speed and acceleration
  - Determination of the relativistic Lagrangian of a free system
  - Proof and definition of Ricci theorem, of the covariant derivative, of the Ricci identity, of the Riemann-Christoffel tensor, of the Ricci tensor, of the Ricci scalar, identities of both Bianchi and finally Einstein tensor
  - Mathematical definition of the capacity and determination of the expression of a parallel plane capacity
  - Determination of the electrostatic potential energy
  - Proof of the value of the field of and electric potential of a straight infinite wire.
  - Determination of the basic properties of electric dipoles as the rigid dipole moment, the presentation of the induced dipole moment, hydrogen bonds, Van der Waals forces, etc.
  - Definition of the Curie symmetry principle and Statement of 6 resulting properties
  - Definition of pseudo vectors
  - Proof of the relations for the volume of the pyramid, the paraboloid, the tetrahedron, octahedron, cube and parallelepiped
  - Determination of the magnetic field produced by a toroidal solenoid, a rectilinear solenoid infinity and a current loop
  - Determination of the magnetic field produced by a magnetic dipole and basic definitions of the properties of magnetic materials
  - Calculation of the Armorial radius and cyclotron pulsation in a non-relativistic framework
  - Determination of the Lagrangian of the electromagnetic field and by extension in the non-relativistic approximation of the tensor of the electromagnetic field
  - Introduction to the calculation of the radiation emitted by an accelerated charge (synchrotron radiation, Lienard-Wiechert's retarded potentials)
  - Calculation of the values of resistors and capacitors in series
  - Difference between electrical potential and electromotive potential
  - Proof of Faraday's law and definition of "self inductance"
  - Proof of Descartes formulas for the concave and convex spherical surfaces and refracting/not refracting as well as for refractive lenses.
  - Definition of stigma and proof that the parable is strictly stigmatic
  - Proof of Descartes formulas for thin lenses and conjugation law
  - Definition of diopter and explanation of various visual disabilities

- **September 12004**
  - Statement of Mach's principle
  - Presentation of the photoelectric effect and proof of the physical law governing it.
  - Proof by example that the light can be seen as both a particle or as a wave
  - Proof of the theorem of monotone class
  - Detailed proof of generalized Klein-Gordon equation (relativistic particle in a magnetic field)
  - Detailed proof of free Dirac equation with explicit Paulis solutions (particles, antiparticles)
  - Determination of the radius of the atom using the Rutherford-Coulomb scatter-

  ing(verbatim: determining the cross section for Rutherford)
  - Presentation of macroscopic and microscopic interactions of X and gamma-rays with matter (which includes details study of the materialization of a photon in electron-positron pair)
  - Introduction to spinors
  - Definition of operating properties of matrices, remarkable matrices, determinants, eigenvectors and eigenvalues
  - Statements of the postulates of wave quantum physics
  - Determination of the orbitals of the hydrogen-atom

- **November 12004**
  - New Biographies (Smith, Say, Malthus, Keynes, Walras, Pareto)
  - Presentation and proof of Noether's theorem
  - Enumeration of some major physical chemical, astronomical constants
  - Introduction to the theory of speculation: (predictive expectation of a financial asset)
  - Introduction to the preference theory (Arrow-Debreu model)
  - Presentation of solutions of the Black & Scholes equation and remarks on the delta - Proof of the Call-Put parity equation
  - Determination of initial stock (optimum) within the framework of the supply chain management

- **January 12005**
  - New Biographies (Penrose, Hawking, Turing, Marx)
  - Development of "dual" version of Maxwell equations and proof of the origin of the expression (hypothesis) of magnetic monopoles
  - Definition of P, NP and NPC class algorithms
  - Proof of the fundamental theorem of calculus also named "fundamental theorem of integral and differential calculus"
  - Presentation of the cGH cube and the Copenhagen interpretation
  - Introduction to the mathematical concepts of artificial neural networks
  - Introduction to the mathematical concepts of Genetic Algorithms
  - Detailed mathematical introduction to fractals
  - Basics maths stuffs on Quantum Computing
  - Introduction to fuzzy logic
  - Proof of Shannon theorem, Morgan theorems, expansion theorem, Karnaugh maps, complete adder, full subtracter

- **April 12005**
  - Basic definitions on block codes, linear codes, systematic codes in error correcting codes
  - Proof of the relation of the relativistic change in mass
  - Introduction to codes and prefix codes
  - Proof of the formula for the calculation of the number of days between two given dates
  - Rounding calculations techniques
  - Definition of rigid or non-rigid post and praenumerando annuities with or without constant rate (certain future)

  - Definition and study of the properties of loans repayment or constant annuity

- **April 12006**
  - Presentation of Schild's criterion via the Einstein effect (gravitational redshift)
  - Development of the Newtonian approximation of the geodesic equation
  - Definitions and proofs of developed forms of four-vectors of displacement, velocity, current, acceleration and energy-momentum
  - Proof of the provenance of electromagnetic field tensor and referential calculation changes
  - Definition of a tautology and principle of non-tautology
  - Descriptions, definitions and many proofs on quaternions
  - Proof of Euler's number irrationality
  - Definition of the log-normal, triangular and Weibull distribution and proof of their mean and standard deviation
  - Introduction to error calculation (absolute and relative uncertainties, error propagation, significant numbers, etc.)
  - Proof of the deflection of light near a Star with the Newtonian gravitational model
  - Definition of a rotation matrix (and developments related thereto)
  - Proof of the existence of the Euclidean division in the ring of polynomials
  - Definitions of MWRR (Time of Money Weighted Return) and TWRR (Time Weighted Rate of Return)
  - Proof of Gauss-Ostrogradsky theorem

- **July 12006**
  - Definition of the concept of dual space
  - Definition of Pareto law and proof of mean and standard deviation
  - Definition of quantile (quartile, percentile)
  - Proof that the mode is the value that minimizes the absolute dispersion
  - Proof of Minkowski and Bienaymé-Tchebychev inequalities
  - Proof of the weak law of large numbers
  - Proof of Euler's formula for planar graphs
  - Introduction to the mathematical method of Six Sigma process controls
  - Proof of the variational calculus theorem
  - Proof of the calculation of the apparent superluminal speed of a high Redshift Star
  - Zero-sum game resolution method using operational research
  - Definition of investment funds
  - Proof of the beta expression of a simple linear regression model (Sharpe)

- **October 12006**
  - Proof Binomial and Poisson likelihood estimators
  - Presentation of the concept of "color" and subtractive and additive synthesis
  - Proof of Einstein equation fields (approach using weak fields approximation)
  - Differentiation of the equivalence principle, weak equivalence principle and Einstein's equivalence principle
  - Proof of the duration of the Daytime arc of planets in the approximation of the null precession and nutation
  - Digital and formal study (approximative) of Lagrange points of a binary system

  - Method of resolution of 4^{th} degree polynomial (Ferrari's method)
  - Presentation of the Gram determinant via the Euclidean volume represented by the joint product of the vectors of a canonical basis
  - Definition of monotonic, strictly monotone functions, etc .. without pure formal approach
  - Approximative determination via the Yukawa potential (mass fields) of the mass of mesons of the weak interaction and the strong nuclear interaction.
  - Introduction to first order linear differential equations

- **December 12007**

  - New Biographies (Heckman, McFadden, Tesla)
  - Bernoulli numbers and polynomials
  - Roche Limit
  - Flattening of celestial bodies
  - Pressure and kinetic temperature
  - Proof of a special case of magnet or electromagnet force
  - Fundamental introduction to Hermitian and Hilbert vector spaces
  - Schwarzschild solution in General Relativity
  - Precession of the perihelion of Mercury in General Relativity
  - Light Deflection in General Relativity
  - Shapiro delay in General Relativity
  - Evaluation Model of Financial Assets
  - The Black Hole Universe
  - Coupling constants of fundamental interactions
  - Hamiltonian of the Schrödinger equation for a charged particle in an electromagnetic field
  - Markov chains
  - Theory of queues
  - Introduction to weather and marine engineering mathematics
  - Practical example in Microsoft® Excel® of the efficiency Markowitz model
  - Practical example in Microsoft® Excel® of the Sharp efficiency model
  - Simplified proof of Green(-Riemann)'s and Stokes theorem
  - Detailed proofs on principal component analysis (PCA)
  - Proof of the spectral theorem in the real numbers case
  - Detailed proofs on logistic regression
  - Proof of Rolle's theorem and mean value
  - Proof of the Hopital theorem and generalized finite increments (generalized mean value)
  - Introduction of Geometric law and proof of its variance and mean
  - Proof of calculation of the area and volume of the five regular Platonic polyhedra
  - Introduction to field algebra and geometry
  - Detailed calculations of the collapsing critical mass (Jeans' Mass) and critical radius (Jeans' Radius) of an interstellar cloud or stellar nurseries
  - Detailed calculations of the collapse time of an interstellar cloud
  - Detailed calculations of a Star nuclear life
  - Detailed mathematical introduction to Fourier transforms
  - Detailed resolution of homogeneous linear differential equations with constant coefficients

  - Presentation of the Cornu's spiral for civil engineering
  - Mathematics of biometric functions
  - Determination and simple resolution of the Pauli equation
  - General solution (Fourier transform) of the electromagnetic wave equation
  - Introduction to strength of materials
  - Visual Horizon
  - Growth rate of a population depending on the temperature
  - Pitot Tube and Pressure drop
  - $U(1)$ gauge theories in quantum field physics
  - Bezier curves
  - Differential equation system with matrix exponentiation

- **September 12009**
  - New Biographies (Pearson, Gosset, Fisher)
  - New important usual primitives for civil engineering and analytical mechanics
  - Detailed proofs of the origin of arcsinh and arccosh functions
  - Proof of Laplace equation and Mayer relation
  - Proof of the propagation of pressure waves
  - Historical introduction for nuclear physics
  - Proof of Maxwell relations in thermodynamics and introduction to free energy and enthalpy
  - Adiabatic atmosphere model
  - Proof of Lorenz equations and the butterfly effect
  - Calculations of some optical properties of the prism
  - Conditions of decoherence/interferences of electromagnetic waves
  - Bloch Sphere
  - Proof of the minimum surface of revolution volume
  - Treatment of the free particle
  - Treatment of the polarized spin qubit and the qubit of spin 1/2
  - Characteristic function and central limit theorem
  - Some proofs on the inequalities in triangles
  - Proof of the volume of a barrel with circular section
  - Proof of the origin of the mean and standard deviation to the Student and Fisher-Snedecor laws
  - Introduction to the marginal cost
  - Statistical test of the one-way ANOVA
  - Statistical Pearson's Chi-square adjustment test
  - Introduction to the analysis of variance of regression
  - Correspondence Factor Analysis
  - Development of linear free or forced RC, RL, RLC circuits

- **September 12008**
  - Simple or complex topology systems analysis for preventive maintenance
  - Introduction to finite and infinite continued fractions
  - Detailed Solutions of a rectangular tunnelling barrier
  - Mathematical model of alpha decay via tunnelling
  - Introduction to Brownian motion model according to Langevin model
  - Introduction to tribology/friction

  - Introduction to time series analysis
  - Additional proofs on long term and short term process capability indices and measuring devices in Statistical Process
  - Proof of the calculation of the PPM in SPC centered or non-centered process
  - Proof of Taguchi quality cost relation
  - Proof of Lienard-Wiechert expression of electric and magnetic potentials
  - Introduction to complex analysis
  - Proof of the second Friedmann equation in cosmology
  - Proof of "slowing" of light near a Black Hole
  - Proof of the expression of the Taylor expansion of a function of two real variables
  - Introduction to experimental design (DoE)
  - Proof of Eherenfest's theorem
  - Proof of the likelihood estimators of the Weibull distribution with two parameters
  - Example of application of decision theory
  - Bands theory (parabolic and semi-classical approximation) within semiconductors
  - Theorem of residues and Laurent series
  - Proofs of Lean Six Sigma values for business processes/workflows

- **October 12011**

  - New Biography (Erlang, Hotelling)
  - Detailed calculation of the geostationary orbit
  - Calculations on PVC meteorological probe balloons
  - Developments on the symmetrical gyroscope and weighing router
  - Detail calculations and proof on Gini index
  - Secular balance, transient and non-equilibrium in radioactive filiation
  - Detailed calculations on the approximate radius of rapidly rotating Stars
  - Proofs on historical delta-normal and variance-covariance Value At Risk
  - Two new jokes in the Humor section
  - Empirical statistical model of wage control
  - Simplified proof of the possible absence of arbitrage opportunity in Finance
  - Calculations on self-financing portfolio on the underlying risk
  - Introduction to mathematical techniques in Insurance
  - Mathematical developments on the power and intensity of a longitudinal sound wave
  - Statistical bilateral Z-test on the difference in the two means
  - Statistical Student T-test on two paired sample means
  - Proof of statistical confidence interval relation of sample large proportions
  - Statistical test for equal proportions of two large samples
  - Application of Shannon's theorem to calculate a statistical index of diversity
  - Proof of the determination of coefficients of a multiple linear regression
  - Proof of the determination of the coefficients of a simple linear regression through the origin
  - Introduction to sensitivity analysis
  - Introduction and some proof on rank/order statistics
  - Demonstration of the provenance, hope and variance of the negative binomial distribution
  - Control charts with detailed mathematical proofs
  - Mathematical approach of first Google Page Rank algorithm
  - Proof of Beltrami's identity to simplify the Euler-Lagrange equation

  - Exact binomial statistical test for the balance of a population with two characteristics
  - Developments and study of gravity waves in a fluid
  - Some simple developments on the gears/gear shafts
  - Proof of skin's effect
  - Theory of the rainbow
  - Theory of double pendulum
  - Boltzmann distribution law
  - Dalton's and Amagat laws
  - Heat Flow
  - Average power in alternative current
  - Presentation of some detailed calculations on the betatron

- **May 12013**

  - New Biographies (Napier, Wilcoxon, Born, Heisenberg, Jordan, Kolmogorov, Stokes, Ostrogradsky, Zeeman, Joseph, Faraday, Meitner, Curie, Biot, Debye, Drude, Ohm)
  - Detailed example of construction of a particular neural network with Microsoft® Excel®
  - Resolution of homogeneous linear differential equations of order 1 with non constant coefficients
  - Example of a Fourier transform of a Gaussian function and proof of the property of the Fourier transform of a derivative
  - Introduction to interactions in two-factor ANOVA
  - Confidence interval and prediction interval of a linear regression
  - Statistical Test signs (median test)
  - Introduction of conditional probability and the conditional mean of Pareto law
  - Determination of the estimators of the gamma distribution using the method of moments
  - Second mathematical approach to the identification of the tides
  - Statistical Kolmogorov-Smirnov adjustment test with Lilliefors approach
  - Proof of the Scheafer's quota model
  - Proof of the calculation of the synodical period of the planets and of demotion of time
  - Proof of the construction of Brownian bridges
  - Proof of the origin of the rare events control chart
  - Calculation of the discount factor of a retiring insurance based on inflation and life expectancy
  - Proof of two-factor ANOVA relations without repetition and repetition
  - Mathematical proof and physics of the LASER
  - Theorem of Taylor series with integral rest
  - Stability of the Poisson distribution
  - Statistical Poisson test for 1 or 2 samples
  - Statistical non-parametric Kruskal-Wallis and Friedman tests
  - Statistical normality test of Ryan-Joiner
  - Statistical C Cochran test
  - Usual Taylor-Maclaurin Series
  - Polynomial regression by least squares method
  - Spatiotemporal Finite Difference Method with Maxwell's equations

  - Break-even Analysis
  - Mechanical Harmonic Oscillator
  - Acoustic Doppler Effect
  - Periodic waves overlays
  - Statistical Tukey test of range
  - Forecasting models by moving average, seasonal coefficients, simple smoothing, Brown double smoothing, double Holt (additive) smoothing, double Holt and Winter (multiplicative) smoothing
  - Introduction to elementary AR, AM, ARMA and ARIMA autoregressive models
  - Proof of the expression of the correction factor on finite population
  - Statistical Fisher exact test
  - Exponential Laplace Smoothing
  - Cohen approval Kappa and McNemar statistical test
  - Kaplan-Meier survival analysis model
  - Cramer's V
  - Clustering K-Means algorithm
  - Clustering Dendrograms algorithm
  - Statistical Wilcoxon signed rank test for one sample or two paired samples
  - Quantitative study of the effective potential energy (harmonic model of the atomic bonding) of the hydrogen-atom
  - Proof of beams equation (Euler-Bernoulli equation)
  - Calculation of the failure rate of a system using the technique of maximum likelihood
  - We added a chronology of Sciences
  - Proof of Einstein model (Dulong-Petit law) of the heat capacity of crystalline solids and derivation of the Debye model
  - Langenvin Model of diamagnetism and paramagnetism
  - Introduction to line integrals calculation
  - Naive determination of the energy of a magnetic dipole
  - "Liquid drop" nuclear Model
  - Magnetic model of spin resonance
  - Integrating factor method for solving differential equations
  - Constant variation method for solving differential equations
  - Mendel's law
  - Temporary and Deferred Life Annuity
  - Carnot Cycle
  - Durand and Gordon-Shapiro equity valuation model
  - Statistical Anderson-Darling adequation test
  - Non-linear optimization by the Newton-Quadratic and Gauss-Newton methods
  - Lagrange Polynomial interpolation method
  - Statistical Cochran-Mantel Heanzel test

- **November 12016**

  - Singular value decomposition theorem

- **December 12016**

  - Fieller's test (ratio of two means)

- **January 12017**
  - Falling Chimney problem
  - Levey-Jennings control charts
  - Simple lattice mixture design of experiment with process variables
  - Average Failure Rate (reliability)
  - Markov Chain Reliability Model
  - Design of reliability tests (Chi-squared time of test, Binomial sampling size, Beta-binomial sampling size)
  - Weibull distribution linearisation
  - Inverted pendulum
  - Durbin-Watson autocorrelation test
  - Fisher's method for multiple p-values
  - Magnifying glass
  - Laney's control chart
  - Classification of conical by the determinant
  - Classification of partial differential equations
- **February 12017**
  - Folded Normal distribution basics CDF and PDF
  - Half-Normal distribution CDF, PDF, variance, mean and median
  - Telescopic and Gandi's series
  - Césaro's sum
  - Implicit Differentiation
  - Bivariate chain rule
- **March 12017**
  - Laplace Integration method
  - Lenth's PSE Pareto Margin Error for unreplicated factorial designs
  - Pareto Margin Error for replicated factorial designs
  - Design of Experiments desirability
- **April 12017**
  - Friedmann–Lemaître–Robertson–Walker metric
  - Jensen inequality
- **May 12017**
  - Introducing weak field gravitational waves equation
  - A mathematical approach of "Divide and rule?" in management
  - Three new jokes in the Humor section
- **July 12017** (v3.7 → v3.8)
  - Correction of many typing errors (almost 50)
  - Relation 11.3.221 had a integral symbol (that should not be there)
  - Relation 7.1.72 had a wrong variable for the angle
  - Relation 7.5.56 there was a $|a|$ instead of a $|b|$ in the second equality
  - Relation 11.1.35 had a 4 that has nothing to do there at the numerator

  - Relation 15.7.905 there was an alignment issue
  - Relation 15.7.888 had a Λ instead of a Γ at the denominator
  - Relation 11.1.20 had a missing T
  - Relations 11.3.169 must be r_E instead of $r-E$ (LaTeX typing error...)
  - Table 8.10 and 8.11 had wrong units in the header row
  - New photo about a quasi vertical circular rainbow and a photo a nuclear magnetic resonance machine
  - There was a typing error in the period (T) relation of the simple pendulum
  - Details of the back-propagation method for neural network
  - New mathematical details on the simple linear regression leverage
  - Numerical applications of some General Relativity experimental tests
  - Derivation of the geodesic of the sphere (as example in the section of Analytical Mechanics)
  - ZeroR classification method
  - K nearest neighbours classification method
  - Definition of Convexity/Concavity of a function (for Jensen inequality proof)
  - Proof of Hermite polynomial Orthogonality
  - Derivation of the Bachelier Option Pricing Model
  - Derivation Forward/Future Cox-Ross-Ingersoll Equality
  - Derivation of the Stress-Energy tensor for a non-relativistic perfect fluid
  - Derivation of the orthodromic distance
  - Derivation of the area of a section of an ellipse
  - There was twice a "Invariance by translation in space" Noether theorem... one of them was obviously "Invariance by translation in time"
  - Schwarzschild innermost stable orbit
  - Hafele–Keating experiment with General Relativity treatment
  - Introducing hyperparameters in Machine Learning
  - Kernel smoothing
  - Credit Default risk
  - We added a lot of new entries in the Chronology section

- **November 12017** (v3.8 $\rightarrow$ v3.9)

  - Section numbering was not correct anymore because of a MiKTeX update.
  - We created all simple cross references
  - Multiple errors in PPI (proton-proton Sun chain function reaction) in relations 11.2.27, 11.2.29 and 11.2.30
  - We added one point to the scientific publication rules (cite equivalent studies for meta-analysis)
  - Correction of a few hundred minor typing errors
  - Typing error in the relation 5.4.231, the first line should be a $\cos(x)$ and not a $\sin(x)$
  - Read "multiplication" instead of "addition" in the remark below the relation 4.2.164.
  - A square was missing on the 2 in 4.4.13.
  - A square was missing on the r in 12.1.57
  - There was a "$+$" instead of a $\cdot$ in 10.2.21
  - Square was missing in relation 4.7.1479, 4.7.1480, 4.7.1481, 4.7.1482. 4.7.1484
  - There was a c^3 instead of a c^2 in the relation 11.1.212
  - At 8.3.450 three paragraphs and two equations were duplicates
  - Proof of the derivative of $f(g)^{g(x)}$

  - OneR (One Rule) in Data Mining with confusion matrix rule selection
  - Apriori Association Rule
  - Kurtosis and Skewness

- **January 12018** (v3.9 → v3.10)
  - Typing error in relation 8.2.160 (a ∂ missing)
  - Typing error in relation 9.3.305 (a μ was missing)
  - There was an error sign in all inverse Fourier transform (error made during the translation of the book into English).
  - Dirac continuity equation
  - CCR (Charnes, Cooper and Rhodes) model of Data Envelopment Analysis (DEA)
  - Kullback-Leibler divergence
  - Continuous and Discrete Linear Convolution
  - Dirichlet Integral
  - Permutation tests
  - Binomial and Gaussian naive bayes
  - More cross-references added

- **June 12018** (v3.10 → v3.11)
  - Typing error for the return angle range of the function $\text{atan}()$ ($-$ instead of a $+$)
  - There was a typing error in relations 4.7.264 and 4.7.265 (power of 3 instead of 2).
  - There was a typing error in relation 4.7.512 a $\text{E}(X)$ instead of a $\text{V}(X)$.
  - A c was missing in relation 4.5.125
  - There were typing error in relations 13.1.739 and 13.1.740
  - A q was missing in relations 9.1.31 and 9.1.32
  - A N was missing in the third relation 14.4.238
  - We added a few more Taguchi tables and Taguchi interaction tables
  - We added the derivation of the one degree of freedom non-central chi-square distribution for the study of tolerance intervals
  - We added new details about the regression forced through the original (especially on the problem of the R^2 calculation)
  - We added some new images (on data fallacies, a joke with Schrödinger cat, planet's comparison size with respected proportions, an illustration of the difference between bagging and boosting)
  - We added in the section of Astronomy data tables with the major properties of interest of the planets of our solar system (including Pluto...)
  - We added in the section of Astrophysics a small data table with the major properties of some well known stars
  - Almost ten section headers were are at the wrong level in the table of contents (...)
  - Surprisingly some subjects on Neutron Stars had disappeared (don't ask why...!?). We put them again.
  - Cox Survival Model (Cox Proportional Hazard Model)
  - We made more explicit the link between the Gauss-Newton method and the gradient descent method
  - We added the confidence interval for the Kaplan-Meier estimator and Cox proportional hazard model (i.e. hazard ratio)

- **December 12023** (v3.11 → v4.0)
  - We noticed that a whole section on Lenz law and inductance wasn't translated since the version 3.1...
  - Typing error in the equality in the sentence below equation 14.4.125
  - Translation error in the axiomatic system of Lukasiewicz (the first line had a "Axiom A5" instead of "Axiom A4")
  - The symbol used for the AND connector in the table 4.5 and the paragraph text were wrong (bad copy paste...)
  - Wrong copy/paste in relations 13.1.387 and 13.1.388 (the sum symbol and the index i should not have been there)
  - There was copy/paste error in the relation 7.7.247 (a d at the beginning that should not have been there)
  - There were some sign errors in the relations 7.1.20
  - Two terms were missing in the relation 4.7.356
  - There were typing errors in relations 8.5.92, 8.5.94 and 8.5.95
  - There was a copy/paste error in the definition of the weighted average 4.7.39
  - There was a wrong 2 factor in the definition of the covariance 4.7.157
  - We replaced all the $x^{'}$ and $x^{''}$ by x' and x''
  - There was a typing error for the energy stored in a capacitor
  - There was a typing error in table 5.11 for the tautology ($P \vee \neg Q$ instead of $P \vee \neg P$). Thanks to Matvii Hodovankiuk!
  - Improvements and corrections in the subsection relatively to the relativistic redshift
  - Improvements and corrections about the SVD theorem
  - We added a small biography of F. Yates, A. Friedmann, G. Lemaître, G. Cramer
  - We improved the FRLW cosmological models subsection
  - We added a few dozens more of spherical harmonics in the section of Quantum Chemistry (thanks to Wikipedia!)
  - We added a brief introduction to the concept of XYZ analysis
  - We introduced the concept of ideal mechanical advantage and actual mechanical advantage
  - Introducing LSTM and GRU recurrent neural networks
  - We corrected all the misspelling of "well know"...
  - Three new jokes in the "Humour" section
  - Mallows's C_p and Akaike Information Criterion (AIC) derivation
  - LASSO, Ridge and Elastic Net regularize regressions
  - Cross-over methods including the proof of the CV and GCV
  - Derivation of multinomial mean and variance
  - Weighted linear regression and general ordinary least squares
  - LU, QR and Cholesky decomposition details
  - Gabriel's Horn finite volume and infinite surface solid of revolution
  - Derivation or Black Hole's photon-sphere radius
  - Risk Parity portfolio optimization
  - Law of total variance (iterated law of conditional variance)
  - Robust estimators, M-estimators and W-estimators
  - Hertzian dipole (antenna) in harmonic regime in far-field approximation
  - Newton-Cotes simple integration methods (Simpson's and Trapezoidal methods)
  - Runge-Kutta methods for solution to ordinary differential equations
  - Linear ordering isotonic regression

- We added the biography of Jerzy Neyman
- Fourier series generative art
- Rayleigh scattering
- Telegrapher equation
- Inductance of a coaxial cable
- Hotelling's T-square test
- Double mass harmonic oscillator
- Lawson Criterion for D-T
- Determinant of block triangular and diagonal matrices
- Woodburry matrix identity and Sherman-Morrison relation
- Factor Analysis
- Total Unduplicated Reach and Frequency Analysis (TURF)
- Equivalence CLAss Transformation (ECLAT)
- Welch Test with Welch-Satterhwaite equation
- Trimmed means T-test
- Expectation-Maximization algorithm
- Euler's Homogeneous Function Theorem
- Eurer-Maclaurin theorem
- Planck's 11912 zero point energy radiation law
- Heisenberg uncertainty principle zero point energy
- Bayes Factor
- Abraham-Lorentz formula
- Adjusted R-squared
- Saha ionization equation
- Cosmological Radiowave Background recombination temperature
- Casimir Effect
- Quantum Cosmology (FLWR Wheeler-Dewitt equation)
- Cochran–Armitage test for trend
- Generative Neural Networks (GAN)
- Point biserial correlation
- Sellmeier equation proof
- Torque of a current loop (DC motor)
- Fresnel circular aperture and circular obstacle diffraction
- Moments generating function and cumulants
- Variance stabilizing transformation
- Laguerre polynomials
- Radial solution or hydrogenoid atom
- Thin film interference
- Fresnel coefficients
- Brewster's angle
- Mollweide projection
- Variance-Bias tradeoff derivation
- Variance-bias tradeoff derivation for OLS
- Antithetic, Conditional, Stratified sampling and importance sample Monte Carlo variance reduction techniques
- Gauss-Markov theorem detailed derivation
- Eddington luminosity and temperature limit (spherical case)
- Logistic Kolmogorov-Smirnov statistic

  - Why gold may be yellow according to special relativity
  - Betz's law
  - Mirage equation
  - Gravitational (non-relativistic) effective potential
  - Gaussian Wavepacket
  - Discrete and continuous time Kelly criterion
  - Added the biography of Maclaurin
  - Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
  - Proof of the pricing of a perpetual bond
  - Pressure of most vessels

- **February 12024** (v3.11 → v4.0)

  - Calculation of power received by Earth from Sun to explain the seasons
  - Error corrected in the section Optics where index of refractions n_1, n_2 were inverted
  - Proof of the law of reflection (only the proof of law of refraction was available)
  - Proof of why stars are not visible on most space photos and videos of standard cameras
  - Proof of the origin of the value above 40 billion light years for the radius of the universe
  - Error correction in the biography of Laplace that was using the picture of Langevin

- **March 12024** (v4.00 → v4.1)

  - Rotating bucket with Newtonian fluid problem
  - Isoperimetric plane theorem
  - Cantelli's inequality
  - Hermitian Cauchy-Schwarz inequality
  - Fourier uncertainty principle
  - Correction of two typewriting errors in the proof of Student's law corrected
  - Introducing intuitively the multivariate canonical Student distribution
  - Index of refraction rate of change with altitude
  - Friedmann-Lemaître Euclidean flat, spherical and hyperbolic spaces dominated by vacuum
  - Correction of the relations 5.6.226 and 5.6.227 that were completely wrong
  - Equivalent Monthly Installment (EMI)
  - Lewis Carrol Epstein interpretation
  - Historical Friedmann universes models

- **April 12024** (v4.00 → v4.1)

  - Correction of the cumulated continuous uniform function that was wrong
  - There was a typo in the expression of composition law of accelerations in the study of the Coriolis acceleration ($\vec{r}$ instead of $\vec{v}_r$)
  - There was a typo in the expression of the orbital speed (R^2 at the denominator instead of a R)
  - An relation was massing in the section of Weather Engineering is the part concerning the study of winds
  - In the section of Engineering false numerical values were given for the energy received by the Sun on Earth at apogee or perigee.

- **July 12024** (v4.1 → v4.2)
    - Calculations related to the properties of a typical Siphon
    - Physical derivation of the $22°$ sun halo (and moon) effect
    - Mathematical treatment of the distinction of Earth's curvature
    - Correction of some french to english wrong translations about the definition of angles
    - We changed the download URL for the book
    - We removed the email contact address from the footer
- **September 12024** (v4.2 → v4.3)
    - We added an example of the triangulation method
    - We corrected a missing coefficient in all the density lagrangians of General Relativity
    - We corrected a typing error in the value of the Absolute Temperature
    - There was twice the same mathematical constant if the table of mathematical constants
    - On reader request we added the proof of $0.9999\ldots=1$
    - We added the Al-Biruni method for determining the Earth's radius
    - We added an example for the Cantor-Bernstein theorem
    - A square was missing (typing error) in a remarkable trigonometric identity involving the tangent
    - Lagrange inversion theorem proof (by Laplace method)
    - Diurnal temperature range
    - Spherical angular theorem
    - Relativistic apparent angle rotation (Lampa-Terrell-Penrose effect)
- **Mai 12025** (v4.2 → v4.3)
    - We added the explanation and physical derivation of Cavendish experiment for G
    - Derivation of Hill sphere (sphere of influence) in the circular orbit approximation
    - Derivation of the Equation of Time and the analemma
    - Derivation of the Borde-Guth-Vilenkin Theorem (BGV theorem) in FWLR metric
    - Proof of the second Borel-Cantelli lemma
    - Proof of the second Borel-Cantelli lemma
    - Proof of the equivalence of Wald Z-test for proportion and binary logistic regression Rao score test
    - Proof of the four-parts formula in spherical trigonometry
    - Introduction to tit for tat in game theory
    - Mathematics of sundials
- **June 12025** (v4.3 → v4.4)
    - Black Hole Flamm's paraboloid representation
- **October 12025** (v4.4 → v4.5)
    - We remapped the links to all various companions PDFs
    - Proof of the Generalized Ehrenfest theorem
    - Proof of the Time-Energy uncertainty principle

  - Definition of "scientific law"
  - Ballistic accounting Earth’s curvature
  - Field of View (FOV)
  - Introduction of the set of divisors
  - Basic calculation of saros cycles
  - Some typing errors corrections (a few dozens)

- **June 12026** (v4.5 → v4.6)
  - Derivation of the expression of flatness of a rotating celestial body
  - Eötvös effect

- **December 12026** (v4.6 → v4.7)
  - Derivation of Euler rotation equation
  - Morse potential
  - Derivation of the surface of a spherical cap
  - Michelson-Morley interferometer
  - Sagnac inteferomter (Sagnac effect)

24

Nomenclature

This chapter contains a summary with simple description of most occurring symbols used in this textbook.

Let us start first with the Greek letters:

α `alpha`
β `beta`
γ `upsilon`
δ `delta`
ϵ `epsilon`
ε `varepsilon`
ζ `zeta`
η `eta`

θ `theta`
ϑ `vartheta`
ξ `xi`
κ `kappa`
λ `lambda`
μ `mu`
ν `nu`

o `o`
π `pi`
ϖ `varpi`
ρ `rho`
ϱ `varrho`
σ `sigma`
ς `varsigma`

τ `tau`
υ `upsilon`
ϕ `phi`
φ `varphi`
χ `chi`
ψ `psi`
ω `omega`

Γ `Gamma`
Δ `Delta`
Θ `Theta`

Λ `Lambda`
Ξ `Xi`
Π `Pi`

Σ `Sigma`
Υ `Upsilon`
Φ `Phi`

Ψ `Psi`
Ω `Omega`

Table 24.1: Greek letters

And finally the most occurring mathematical operators and objects:

Symbols	Descriptions
(	Opened parenthesis
)	Closed parenthesis
[	Opened bracket
]	Closed bracket
{	Opened brace
}	Closed brace
$]\ldots,\ldots[$	Opened interval
$[\ldots,\ldots]$	Closed interval
$]\ldots,\ldots]$ $[\ldots,\ldots[$	Semi-opened (or semi-closed) intervals
$\lfloor x \rfloor$	Floor of x, round down to nearest integer

Symbols	**Descriptions**
$\lceil x \rceil$	Ceiling of x, round up to nearest integer
$[x]$	Integer value of a number
$\{x\}$	Fractional value of number ($[x] = x - \{x\}$)
$\{\ldots\}$	Content of a set
$\wedge$	AND operator, denoted & in computer science and corresponding to multiplication in maths
$\vee$	OR operator, denoted sometimes ǁ in computer science
$\varnothing$	Empty Set
$\therefore$	Therefore
$\because$	Because
$\Rightarrow$	Implies (in applied maths), if and only if (in logic), nuclear reaction (in atomistic)
$\Leftrightarrow$	Is equivalent
$\neg$	Negation (logical negation)
$\vdash$	Proves that
$\nvdash$	Doesn't proves that
∞	Infinity
$=$	Equality symbol
$\equiv$	Identity symbol (left term is assumed to be equal to the right one and vice-versa)
$\cong$	Approximately equal symbol
$\propto$	Linearly proportional symbol
$:=$, $\triangleq$	By definition symbols (defined as)
$>$	Greater than
$<$	Less than
$\gg$	Much greater than
$\ll$	Much smaller than
$\leq$	Less than or equal to
$\geq$	Bigger or equal to
$\succ$	Preferred to (for utility in econometry) or positive definite matrix in linear algebra
$\prec$	Not preferred to (for utility in econometry) or negative definite matrix in linear algebra
$\succeq$	Preferred or equal to (for utility in econometry) or negative semidefinite matrix in linear algebra
$\preceq$	Not preferred or equal to (for utility in econometry)
$\sim$	Equivalent to (for utility in econometry)
$\rightarrow$	Tends toward to in functional analysis or implies in logic
$\oplus$	In proof theory an abstract way to represent the following three operators $\{\wedge, \vee, \rightarrow\}$, in boolean algebra it is the exclusive OR
$\rightarrowtail\!\!\!\!\twoheadrightarrow$	Surjective symbol
$\rightarrowtail$	Injective symbol
$\leftrightarrow$	Bijective symbol
$\mathcal{T}$	In logic a set of terms
$\mathcal{F}$	In logic a set of formulas
$\mathbb{N}$	Natural numbers set (positive integers)
$\mathbb{N}^*$	Natural numbers set (positive integers) not including zero

Symbols	Descriptions
$\mathbb{Z}$	Relative numbers set (all integers)
$\mathbb{E}$	Even numbers set
$\mathbb{O}$	Odd numbers set
$\mathbb{Q}$	Rational numbers set (ratio of relative numbers)
$\mathbb{R}$	Real numbers set
$\mathbb{C}$	Complex numbers set
$\mathbb{H}$	Quaternion numbers set
i	Pure imaginary number
$\bar{z}$	In statistics the arithmetic average of z or the complex conjugate of z
$\lvert x\rvert$	Absolute value of x or mode of a complex number
$\Re$	Real part of a complex number
$\Im$	Imaginary part of a complex number
$\ker(\ldots)$	Kernel of a function
$\mathrm{Card}(\mathcal{S})$	Cardinal (size) of set $\mathcal{S}$
$\aleph$	Transfinite Cardinal symbol
$\in$	Symbol that means left term belongs to the right term
$\notin$	Symbol that means left term does not belong to the right term
$\subset$	Symbol that means left term (that is a set) is a subset of the set on the right
$\not\subset$	Symbol that means left term (that is a set) is not a subset of the set on the right
$\subseteq$	Symbol that means left term (that is a set) is a subset or a set equal to the set on the right
$\not\subseteq$	Symbol that means left term (that is a set) is not a subset or even a set equal to the set on the right
$\cup$	Symbol that means left term (that is a set) is merged (union) with right term that is also a set
$\sqcup$	Symbol that means left term (that is an interval) is merged (union) with right term that is also an interval
$\cap$	Symbol that means we take only intersection (equal) items of left and right terms that are sets
$\bigcup\limits_{i=1}^{n}$	Union of multiple sets
$\bigcap\limits_{i=1}^{n}$	Intersection of multiple sets
$\mathcal{P}(A)$	All possible subsets of A
A^c	All the items that do not belong to set A
$\vert$	Such that...
$\forall$	For all...
$\exists$	It exists...
$\star$	Generic operation (non-explicit) in set theory
$+$	Addition symbol of two terms
$\sum\limits_{i=1}^{n}$	Summation of multiple indexed terms
$-$	Subtraction symbol of two terms

Symbols	Descriptions
$\times, \cdot$	Multiplication (product) symbol of two terms
$n!$	Factorial function
$\prod\limits_{i=1}^{n}$	Multiplication (product) symbol of multiple indexed terms
$\div, /, :$	Symbols for division of two terms depending on school level
$\mid$	Can be divided in Euclidean division (when the left term can divide the right term)
$\nmid$	Cannot be divided in Euclidean division
$\sqrt{\ldots}$	Square root
$\sqrt[n]{\ldots}, (\ldots)^{1/n}$	Common notations for the n-root
$\equiv$	Congruence
mod	Modulus
$[x]_m$	Set of integers congruent to a modulo m
A_n^m	Number of arrangements with repetitions
A_m^n	Number of arrangements without repetitions
Γ_p^n	Simple combination with repetitions
$C_k^n, \dbinom{n}{k}$	Binomial coefficient notation[1] defined by $n!/(k!(n-k)!)$
$\vec{x}$	A vector
$\text{proj}(\ldots)$	Projection operator of a vector
$\lVert\vec{x}\rVert$	Norm of a vector $\vec{x}$
$\ln(\ldots)$	Natural logarithm (base e) of a number
$\log(\ldots)$	Base 10 (if not indicated) logarithm of a number
$e, \exp$	Different notations for the exponential (euler number)
P	Depending on the context this is a probability, cumulated probability or part of a set or a proposition (in proof theory)
$/, \mid$	In the field of bayesian analysis these symbols mean "being known", or "knowing that"
$\mathrm{E}(\ldots),\mu$	In statistics the expected (arithmetic) mean
μ_g	In statistics the geometric mean
μ_h	In statistics the harmonic mean
$\mathrm{V}(\ldots), \sigma^2$	In statistics the biased variance (σ being the biased standard deviation)
M_e	The median (middle value)
M_0	The mode (modal value)
γ	The skewness
κ	The kurtosis
$\hat{x}$	In physics the amplitude of x, in statistics an (unbiased) estimator of x, in geometry a vertex x
Δ	Discriminant of a polynomial in calculus, mediator in geometry, focal length in projective geometry
$\times$	If left and right terms are vectors, this is the cross product (vectorial product)

---

[1] This textbook doesn't respect the ISO notation for this coefficient for recall!

Symbols	Descriptions
$\circ$	If left and right terms are vectors, this is the dot product, also named "inner product" or "scalar product". Otherwise if used with functions, this is the symbol of composition
$\langle \ldots, \ldots \rangle$	Inner product between function
d, ∂	Exact and partial differentials
$\mathcal{C}^n$	Differentiability (continuity) order of a function
$\mathrm{d}f/\mathrm{d}t$ $\mathrm{d}_t f(t)$ f' $\dot{f}$	Common notations for the derivative of a function f
$\mathcal{O}(\cdot)$	Big-O roughly means order of magnitude, order or complexity in computer science
$\vec{\nabla}(f)$ $\vec{\nabla} \cdot f$ $\vec{\nabla} f$ $\overrightarrow{\text{grad}}(f)$	All common notations for the gradient of a scalar field
$\vec{\nabla}(\vec{f})$	Gradient of a vector field
$\vec{\nabla} \circ \vec{f}$ $\text{div}(\vec{f})$	All common notations for the divergence of a vector field
$\vec{\nabla} \times \vec{f}$ $\text{rot}(\vec{f})$	All common notations for rotational (curl) of a vector field
$\Delta \cdot f$ $\vec{\nabla}^2 f$ $\vec{\nabla} \circ (\vec{\nabla} \cdot f)$	All common notations for for the Laplacian of a scalar field
$\Delta \vec{f}$	Laplacian of a vector field
$\Box = \dfrac{1}{c^2}\dfrac{\partial^2}{\partial t^2} - \Delta$	Alembertian
$\diamond$	Vectors elements-wise multiplication (ie Hadamard product)
$\displaystyle\int$	Riemann primitive
$\displaystyle\int\limits_a^b$	Riemann integral in range $[a, b]$
$\displaystyle\oint$	Closed non-oriented curvilinear integral (line integral)
$\displaystyle\ointclockwise$	Clockwise path integral
$\displaystyle\ointctrclockwise$	Counter-clockwise path integral
$\text{Res}[f(z); z_0]$	Residue on z_0 of a function $f(z)$
$*$	Convolution of matrices or functions
D	Covariant derivative (parallel transport) or Fermi-Walker derivative
$M_{m \times n}(\mathcal{S})$	Notation for a matrix of size $m \times n$ with component in $\mathcal{S}$, where $\mathcal{S}$ can be $\mathbb{N}$, $\mathbb{R}$,$\mathbb{C}$
$M_n(\mathcal{S})$	Notation for a square matrix of size n with component in $\mathcal{S}$, where $\mathcal{S}$ can be $\mathbb{N}$, $\mathbb{R}$,$\mathbb{C}$
M_{ij}	Element of the row i and column j of a matrix

Symbols	Descriptions
$M_{:,j}$,$M_{.j}$,col$_j(M)$	The j-th column of a matrix
$M_{i,:}$,$M_{i.}$	The i-th row of a matrix
0_{mn}, $\mathbf{0}$, $\mathbb{O}$	Common notations for a matrix with only zeros as components
I_n, $\mathbb{1}_n$	Unity square (identity) matrix (diagonal with 1, 0 everywhere else) of dimension $n\times n$ or vector of only 1 as components of dimension n
tr(A)	Trace of a matrix
σ	Permutation function
$\mathcal{S}_n$	Set of all permutations of n elements
$D(A)$ det(A) $X(\mathrm{i}\omega)$ $\|A\|$	Common notations for the determinant of a matrix
diag(A)	Diagonal vector extracted from matrix A
A^-	Pseudo-inverse of a matrix
$\vec{x}^T$, A^T	Transpose of a vector $\vec{x}$ or matrix A
$A^\dagger$	The conjugate and transpose matrix A
rk(M)	Rank of a matrix M (dimension of the vector space generated (or spanned) by its columns)
adj(M)	Adjugate (classical adjoint), i.e. transpose of its cofactor matrix, of a square matrix M
δ_{ij},δ^{ij}	Covariant and respectively contravariant (2D) Kronecker symbol
$\varepsilon_{ij},\varepsilon^{ij}$	Covariant and respectively contravariant (2D) Levi-Civita symbol
g_{ij},g^{ij}	Covariant and respectively contravariant metric tensor
η_{ij},η^{ij}	Covariant and respectively contravariant minkowski metric tensor
E^n,E_*^n	n-dimensional space and it's dual space
$\mathcal{E}^n,\mathcal{E}_*^n$	n-dimensional Euclidean space and it's dual space
$\mathcal{P}$	Symbol of a plane (in geometry) or partitions of a set (in measure theory)
$\mathcal{H}^n$	n-dimensional Hilbert space
$\otimes$	Tensor product or Kronecker product (depending on the context...), i.e. matrix cross-correlation
Γ_{ki}^j	Christoffel symbols of the second kind
Γ_{kij}	Christoffel symbols of the first kind
∇_j,∂_j	Covariant derivative for 1-forms
$R_{i,jk}^l$	Riemann-Christoffel symbol
R_{ij}	Ricci curvature
R	Ricci scalar
$G_{\mu\nu},G^{\mu\nu}$	Covariant and respectively contravariant Einstein tensor
$:$	When the left term is a matrix at the right one a vectors, this is "Frobenius (matrix) dot product" or "Frobenius inner product"
$\mathcal{F}(x(t))$ $\mathcal{F}(x(t))(\omega)$ $X(\mathrm{i}\omega)$ $X[k]$	Most common notations for the Fourier transform of $x(t)$
$\mathcal{F}^{-1}(F(\omega))(t)$ $\mathcal{F}^{-1}(X[k])$	Most common notations for the inverse Fourier transform
$\mathcal{F}[x[m]]$	Discrete Fourier transform of $x[m]$

Symbols	Descriptions
$\mathcal{F}^{-1}[X[n]]$	Inverse discrete Fourier transform of $X[n]$
$\mathcal{L}(x(t))$	Laplace transform
$\mathcal{L}(X(s))$	Inverse Laplace transform
$\mathcal{UL}(x(t))$	Unilateral Laplace transform
$\mathcal{Z}[x[n]]$	Forward bilateral $\mathcal{Z}$-transform
$\mathcal{Z}^{-1}[X[z]]$	Inverse $\mathcal{Z}$-transform
$\mathcal{UZ}[x[n]]$	Unilateral $\mathcal{Z}$-transform
$\perp$	Perpendicular to
$\overline{AB}$, AB	Segment A to B
$\widehat{AB}$	Angle between A and B
$\mathcal{S}^n$	n-dimensional surface of a $n+1$ dimensional sphere
$\arg\min_x f(x)$ $\min_x f(x)$ $\arg\max_x f(x)$ $\max_x f(x)$	Respectively the value of x that minimize/maximize $f(x)$
ΔABC	A triangle made with three points A, B and C
$R_{\text{axe}}, T_{\text{axe}}, S_{\text{axe}}$	Rotation, translation or symmetry operators along a given axe
$K_{p,q}$, K_n	The first denotes a bipartite graph and the later a complete graph
$D=\gamma^{\mu}\partial_{\mu} \equiv \not{\partial}$	Feynman's slash notation (i.e. Dirac operator)
♁	Symbol in astronomy and astrophysics to refer to the Earth
$\odot$	Symbol in astronomy and astrophysics to refer to the Sun
${}_{Z}^{A}\mathrm{X}_{N}$,, ${}^{A}\mathrm{X}$	Two common notations for a chemical element of mass number A, atomic number Z and number of neutrons N
$\rightleftharpoons$	Double harpoon (reversible reaction in chemistry)
$\rightharpoonup$	Simple harpoon (non-reversible reaction in chemistry)

For a summary of the physical and chemical constants, the reader must refer to the section of Mechanics page 3113.

List of Figures

2.2 Fire safety guide 19
2.3 Chemical laboratory safety guide 20
2.4 Electrical safety guide 20

4.2 Scientific method cyclic process 40
4.5 Most common types of causal relationships 58
4.6 Scientific evidence hierarchy (assuming each level is made “at its best”) . . . 60
4.7 Levels of Evidence in a social science like History 62
4.8 Special example of a meta-analysis 63
4.9 How to publish peer review literature 74
4.10 Different types of scientific evidence 78
4.11 Swiss TV publication about Lyme borreliosis treatment 79
4.12 Swiss TV publication about a miracle in Lourdes 80
4.13 Caricatural Dunning-Kruger effect curve 84
4.14 Same data, different conclusions (purpose of meta-analysis) 86
4.15 Experts opinions/estimation issue 86
4.16 Logical structure of falsificationism 94
4.17 Organized collection of irrational nonsense 97
4.18 The Alberto Brandolini bullshit asymmetry law 99
4.19 Evolutionary tree of religions 107
4.21 False equivalence fallacy 124

5.1 Relations between various important mathematical structures for physics . . 132
5.2 Hypothetical-Deductive Reasoning block diagram 141
5.4 The four standard types of quantified statements 164
5.5 The tree of Logic 193
5.6 Indo-Arabic numbers 209
5.7 Description of decimal and positional system 210
5.8 Scale representation of the positional system 211
5.9 Decimal to binary conversion 213
5.11 A $1/2$ pizza fraction example 223
5.12 A $1/4$ pizza fraction example 224
5.13 A $1/12$ pizza fraction example 224
5.14 Cantor diagonal method 225
5.15 Simple numbers sets summary 230
5.16 Percentage usage illustration 235
5.17 VOLTCRAFT VC-5080 Multimeter with ohmic graduation displaying ∞ symbol 242
5.18 Complex Gauss plane 249
5.19 Fresnel representation 255

5.20 Fresnel rotating vector (complex oscillator) . . . 256
5.21 3D complex oscillator . . . 256
5.22 Representation of the complex rotation . . . 261
5.29 One possible schema for addition . . . 297
5.30 Addition of fractions with common denominator . . . 300
5.31 Addition of fractions without common denominators . . . 300
5.32 Addition table of the first 12 natural numbers . . . 302
5.33 One possible schema for subtraction . . . 303
5.34 Subtraction of fractions without common denominators . . . 306
5.35 Subtraction table of the first 12 natural numbers . . . 307
5.36 Summary of exponents rules . . . 310
5.37 Multiplication of fractions without common denominators . . . 312
5.38 Multiplication table of the first 12 natural numbers . . . 313
5.39 Schematic example of a division (fractions) . . . 314
5.41 Venn diagram of common divisors . . . 341
5.48 Summary of the GCD algorithm . . . 345
5.49 Time-keeping on a clock using modulo 12 arithmetics . . . 357
5.50 Sagittal diagram example from a definition set to a destination set . . . 378
5.51 Function returning in its own set of definitions . . . 379
5.52 Zermelo-Frankel axioms visual summary . . . 388
5.53 Global/Local maximum and ninimum example . . . 393
5.54 Visual example (Euler diagram) of the inclusion . . . 394
5.55 Visual example (Euler diagram) of the intersection . . . 395
5.56 Visual example (Euler diagram) of the union . . . 396
5.57 Visual example (Euler diagram) of the difference . . . 398
5.58 Visual example (Euler diagram) of the symmetric difference . . . 398
5.59 Visual example (Euler diagram) of the complementarity . . . 400
5.60 Common SQL query expressions with joins . . . 402
5.61 ker concept of a function . . . 405
5.62 Schematic representation of a surjective application or function . . . 406
5.63 Schematic representation of an injective application or function . . . 407
5.64 Schematic representation of a bijective application or function . . . 407
5.65 Bijective function example . . . 408
5.66 Representation of an antisymmetric relation . . . 410
5.67 Synoptic diagram of common algebraic structures . . . 416
5.68 Example of a concept of vector space . . . 432
5.69 Summary of most common algebraic structures . . . 435
5.70 Group Theory addition summary . . . 436
5.71 Group Theory multiplication summary . . . 436
5.72 Merging of addition and multiplication . . . 437
5.73 Abelian group structure . . . 437
5.74 Field set structure . . . 438
5.75 Ring homomorphism . . . 440
5.76 Venn diagram of different types of homomorphisms . . . 443
5.77 Birthday paradox probability distribution . . . 460
5.78 Common amino-acids . . . 461
5.81 NASA/ESA Hubble space telescope sharpest and largest M31 (Andromeda galaxy) image ever . . . 470

5.82 Nature Timespiral . . . 472
5.83 Possible Venn diagrams for two events . . . 483
5.84 Bayes' theorem illustrated . . . 487
5.85 Example of Bayesian network (acyclic oriented) with 5 states . . . 505
5.86 Directed acyclic Bayesian network with probability of departure . . . 506
5.87 Serial/linear Bayesian network . . . 507
5.88 Divergent Bayesian network . . . 507
5.89 Convergent Bayesian network . . . 508
5.92 A Posteriori probability of a production stop due to a machine failure in MSBNX 1.4.2 . . . 510
5.95 A Posteriori probability of an evacuation due to a machine failure in MSBNX 1.4.2 . . . 512
5.96 Full Bayesian network analysis . . . 513
5.98 Implicit probability of an alarm in MSBNX 1.4.2 . . . 515
5.100 Pascal's formula illustrated . . . 530
5.101 Combinations, permutations and arrangements simple summary . . . 534
5.102 Generic example of a Markov chain . . . 538
5.103 Concrete example of a very simple Markov chain . . . 540
5.104 Graphical representation of the estimation of the median by linear interpolation 553
5.105 Graphical representation of the estimation of the modal value with classes intervals . . . 555
5.106 Box & Whiskers plot ideal chart type . . . 558
5.107 Graphical representation of the mode, median and 1st + 3rd quartile compared to a distribution . . . 559
5.108 Graphical representation of a few moving averages . . . 568
5.109 Starting point for the geometric representation for the various averages . . . 574
5.110 Type of variables in statistics . . . 581
5.111 Illustration of "tension" in statistics . . . 588
5.112 Geometric interpretation of covariance . . . 600
5.113 Anscombe's quartet graphs summary . . . 610
5.114 Some Pearson correlations . . . 615
5.115 Intra-class correlation coefficient illustration . . . 629
5.116 Kurtosis and Skewness range map of some distributions . . . 638
5.117 The Beta and Gamma families of distributions on the shape characterization plane . . . 641
5.118 Uniform law $\mathcal{U}$ (density and cumulative distribution function) . . . 651
5.119 Bernoulli law B (cumulative distribution function) . . . 653
5.120 Geometric law $\mathcal{G}$ (mass and cumulative distribution function) . . . 657
5.121 Binomial law $\mathcal{B}$ (mass and cumulative distribution function) . . . 663
5.122 Trinomoial distribution with $p=1/5, q=2/5$ and $n=5$ . . . 665
5.123 Negative Binomial law `NB` (mass and cumulative distribution function) . . . 671
5.124 Hypergeometric law $\mathcal{H}$ (mass and cumulative distribution function) . . . 676
5.125 Conditions of application of the approximation by a Normal distribution . . . 678
5.126 Poisson law $\mathcal{P}$ (mass and cumulative distribution function) . . . 687
5.127 Gaussian distribution cumulated area . . . 697
5.128 Normal law $\mathcal{N}$ (mass and cumulative distribution function) . . . 699
5.129 Sigma intervals for the Normal distribution . . . 702
5.130 Full width at half maximum (FWHM) illustrated . . . 707

5.131 Plot of the bivariate Normal function in MATLAB® 714
5.132 Plot of the bivariate Normal function with pgfplots 715
5.133 Plot of the iso-lines of the bivariate Normal function (non-correlated case) . . 716
5.134 Plot of the iso-lines of the bivariate Normal function (correlated case) 717
5.137 Q-Q plot of the distribution 732
5.138 Log-Normal law (mass and cumulative distribution function) 737
5.139 Uniform continuous law (mass and cumulative distribution function) 739
5.140 Triangular law (mass and cumulative distribution function) 743
5.141 Naive 3×3 risk matrix in project management 745
5.142 Pareto law (mass and cumulative distribution function) 747
5.143 Exponential law (mass and cumulative distribution function) 752
5.144 Cauchy law (mass function) 758
5.145 Some Beta law mass functions 763
5.146 Beta law (mass and cumulative distribution function) 763
5.147 Some Gamma law mass functions 768
5.148 Gamma law (mass and cumulative distribution function) 769
5.149 Some χ^2 law mass functions 774
5.150 χ^2 law (mass and cumulative distribution function) 775
5.151 Student T law (mass and cumulative distribution function) 787
5.152 Plot of the Benford function (cumulative distribution function) 803
5.153 ex-Gaussian law (mass and cumulative distribution function) 810
5.154 Relations between distributions 814
5.155 Visual summary of some important continuous and discrete statistical distributions 816
5.156 Visual summary of some important continuous statistical distributions (PDF and CDF) 817
5.157 Type of censored data 833
5.158 Example of confidence interval in weather-forecasting 850
5.159 Cartography of statistics standards 851
5.160 Line plot illustration of the overlay of two 95% confidence intervals 856
5.161 Comparison between the Normal and Student distribution functions 866
5.164 Convergence of different customary distributions in elementary statistical inference 891
5.166 Richard McElreath's Quartet 897
5.168 Bayes Factor description in JASP framework 915
5.169 What it means to be "revolutionary" in statistics 918
5.170 Approaching the Bernoulli distribution by the Normal distribution according to the CLT 941
5.174 Illustration of a test (or confidence level) unilateral and bilateral 949
5.175 Summary figure to determine the p-value of parametrical test with symmetrical distributions 950
5.176 Null and alternative hypothesis of a special case of two-sided test 955
5.177 p-hacking 956
5.178 High statistical significance which contextualized has low practical significance 958
5.179 The three possible scenarios for a hypothesis test on the average 961
5.180 Comics illustrating null and alternative hypothesis 962
5.181 Fisher's method for multiple p-values 965
5.182 Special example of a meta-analysis 966

5.185 IRIS dataset non-clustered regression in R to illustrate Simpson paradox . . . 969
5.186 IRIS dataset clustered regression in R to illustrate Simpson paradox 970
5.188 Power analysis table of one proportion test from PSS software 987
5.189 Main effects plot without interactions in Minitab 15 1028
5.190 Interactions plot without interactions in Minitab 15 1029
5.191 Main effects and interactions plot with interactions in Minitab 15 1030
5.192 Main effects and interactions plot with interactions in Minitab 15 1032
5.198 Experimental vs Observational studies 1053
5.199 How to select a Gage R&R design 1067
5.200 Cross vs Nested studies (source: Minitab) 1068
5.201 Comparing different some major ANOVA structures 1072
5.202 Latin squares of orders 3, 4 and 5 1075
5.203 Merge of two Latin square designs 1082
5.204 Families of Multivariate Linear Models 1091
5.205 Same data, different conclusions (purpose of meta-analysis) 1107
5.206 Approach of binomial functions by associated Normal functions 1112
5.207 Representation of the approach of the Kolmogorov-Smirnov GoF 1121
5.218 Idea behind the Anderson-Darling GoF 1130
5.225 Kernel density estimation of 100 normally distributed random numbers . . . 1148
5.227 Histograms with different bin widths and a kernel estimate of $f(x)$ for the same sample 1149
5.228 Decision tree for naive statistical tests 1155
5.231 Summary of most common coefficients of correlations and measures of association 1202
5.247 Cohen Kappa statistics in binary classification models 1263
5.248 Taxonomy of the methods developed for survival analysis 1272
5.249 Drug development and clinical trials 1275
5.250 Kaplan-Meier Typical survival curves in R software 1279
5.252 Cox PH typical survival curves in R software 1302
5.254 Different types of measurement errors 1310
5.255 Some common sampling methods 1328
5.256 Over and under-sampling underlying idea 1331
5.257 SMOTe and ROSE sampling underlying idea 1332
5.259 Type of errors in surveys methodology 1335
5.260 Common data fallacies 1343

6.1 Visual of squared binomial identity 1362
6.2 Visual of cubic binomial identity 1362
6.3 Pascal triangle 1363
6.6 Some polynomials plotted with TikZ 1370
6.7 Turning point explicitly highlighted 1371
6.8 Solving simple equation by circular reference in Microsoft® Excel® 1380
6.10 Typical orientation for second degree polynomials 1382
6.11 Inflection point of the tangent? 1383
6.12 Some real example of second degree polynomials 1384
6.16 Golden ratio illustration 1389
6.17 Five first Legendre polynomials 1404
6.18 Plot of six first Laguerre polynomials 1408

6.19 Plot of some associated Laguerre polynomials 1412
6.20 C_{12} cyclic group 1419
6.21 Simple affine transformation on a circle 1422
6.22 Operations leaving invariant a tetrahedron 1433
6.23 Examples of compositions and inverse of permutations 1440
6.24 Photomaton transformation 1443
6.26 Secant vs Tangent 1452
6.28 Turning point and inflection point of 1454
6.31 Plot of the hyperbolic tangent function, its first and second derivatives 1457
6.32 Graphical representation of Rolle's theorem 1460
6.33 Partial derivative illustrated 1465
6.34 Total differential illustrated 1466
6.37 Most common types of derivatives in mathematics 1473
6.38 The folium of Descartes with asymptote 1494
6.39 Area A to be calculated in a bounded continuous positive function 1498
6.41 Minorant sum of areas A_m 1499
6.42 Minorant sum of areas A_M 1499
6.43 Minorant and majorant sum of areas superimposed 1499
6.44 Principle of calculating the lower Darboux sum 1501
6.45 Principle of calculating the upper Darboux sum 1502
6.46 Riemann Sum 1505
6.47 Principle of calculating methods using left rectangles 1505
6.48 Principle of the calculation of the Riemann integral with left rectangles method 1507
6.50 Example of a continuous function of two variables over a squared domain . . 1518
6.51 Volume decomposition into big parallelepipeds 1518
6.52 Decomposition of the volume in always thinner parallelepipeds 1519
6.54 Example of a type I domain 1520
6.55 A nonlinear map $f:\mathbb{R}^2 \mapsto \mathbb{R}^2$ sends a small square to a distorted parallelogram 1525
6.56 Graphical representation of the determinant 1526
6.57 Schematic representation of the Dirac delta function by a line surmounted by an arrow 1570
6.58 Regularized Gaussian delta Dirac function 1572
6.59 Plot of the Heaviside function 1573
6.60 Plot of the Euler Gamma function in Maple™ 4.00b 1575
6.61 Plot of the Euler Gamma function in the complex plane in Maple™ 4.00b . . 1576
6.62 Outgoing flux for surface integral of an entire sphere 1589
6.63 Outgoing flux for surface integral of the side of a cylinder 1590
6.64 Poincaré Diagram - Classification of Phase Portraits 1643
6.65 Illustration of the principle of convergence of a sequence 1658
6.66 Fibonacci spiral 1662
6.67 Triangular numbers 1666
6.68 Some Bernoulli polynomials 1673
6.70 Approximation of the sine function by a Maclaurin development in Maple™ 4.00b 1698
6.71 Approximation of the sine function by a Maclaurin development with TikZ . 1699
6.72 Example of non-convergent Maclaurin series in Maple™ 4.00b 1700
6.73 Shift possibility of Maclaurin series in Maple™ 4.00b 1701
6.74 Bivariate Taylor example in Maple™ 4.00b 1716

6.75 Types of stationary points (second derivative test of quadratic forms) 1718
6.76 Examples of some Fourier series 1734
6.77 Time and frequency domain 1735
6.79 Frequency spectrum of the example Fourier series coefficients 1739
6.89 Pedagogical way to see the Fourier Transform 1745
6.90 Example of Fourier series in Maple™ 4.00b with 4, 8 and 16 terms 1747
6.91 Example of Fourier series in Maple™ 4.00b with 50 terms 1747
6.92 Rectangular pulse example for Fourier Transform 1763
6.93 Rectangular pulse example for Inverse Fourier Transform 1766
6.94 Plot of a few Bessel functions 1769
6.95 Example of arrow in $\mathbb{R}^3$ euclidean space 1795
6.96 Mutually perpendicular planes in $\mathbb{R}^3$ 1796
6.97 Right-handed coordinate system 1797
6.98 Example of a sum of two vectors 1798
6.99 Example of a sum of two vectors with robot dynamics notation 1798
6.100 Differences of transformations between a vector and a pseudo-vector 1800
6.101 A vector field of normals to a surface 1801
6.102 Example of a construction of a vector in a three-dimensional space 1808
6.103 Representation of direction angles 1811
6.104 Artistic but real example of an $G = \mathrm{A}\mathbb{R}^2$ affine space 1816
6.105 Vectorial polygon in $\mathrm{A}\mathbb{R}^2$ 1816
6.106 Details of the calculation of the norm in an orthogonal coordinate system $\mathbb{R}^3$ 1819
6.107 Geometrical representation of the dot product (projection) 1822
6.108 Orthonormal basis projection example 1823
6.109 Illustration of the triangular inequality 1828
6.110 Geometrical representation of the cross product 1835
6.111 Mixed product illustration 1838
6.112 Localisation of a point in a simple 3D coordinate system 1847
6.113 Representation of the spherical coordinate system 1849
6.114 Latitude and Longitude concepts illustrated 1850
6.115 Orthodromic distance (great-circle distance) 1854
6.116 Japanese Himawari 8 Earth's photo 1855
6.118 Representation of the cylindrical coordinate system 1861
6.119 Representation of the polar coordinate system 1863
6.120 Gravitation vector field example 1866
6.121 Velocity vector field example 1867
6.122 Gradient deepest ascent and deepest descent 1871
6.129 Move of a vector through a parallelepiped 1880
6.130 Various configurations of divergence 1887
6.131 Example of rotational (curl) of a vector field 1888
6.138 Various configurations of the rotational 1898
6.140 Graphical representation of the linear system with Maple™ 4.00b 1931
6.141 Overdetermined system 1931
6.142 Representation of a system of 3 equations with two unknowns in Maple™ 4.00b 1933
6.143 Matrix product Falk's scheme 1947
6.144 Geometrical interpretation of the determinant 1964
6.145 A vector represented in two different bases (purple and red arrows) 1992
6.146 Geometrical interpretation of eigenvector and eigenvalue in $\mathbb{R}^2$ 1997

6.148 SVD decomposition of a $m \times n$ matrix 2016
6.149 SVD in MATLAB® example original image 2017
6.154 3D Levy-Civita symbol illustration 2046
6.155 Covariants and Contravariants coordinates 2060
6.158 Visual link between a tensor and its representative surface 2087
6.159 Coordinate surfaces, coordinate lines and coordinate axes of general curvilinear coordinates 2093
6.160 Coordinate surfaces, coordinate lines, and coordinate axes of spherical coordinates 2098
6.161 Parallel transport illustration on a sphere 2114
6.162 Reminder of cardinals names 2115
6.163 Unit Sphere 2146
6.165 Plane symmetry of a vector relatively to a plane 2153
6.167 Spinor full rotation (special example) 2160

7.1 Typical representation example of an oriented infinite axis with origin 2177
7.2 Position of a point in a cartesian plane 2178
7.4 Vocabulary planar graph 2180
7.5 Representation of roots on a planar graph 2182
7.6 Example of step-wise function for swiss coordinate wage in Maple™ 4.00b . 2183
7.7 Example of typical histogram in engineering companies (Six Sigma) 2184
7.8 Example of OHLC (candlestick plot) time series with moving averages in financial trading 2184
7.9 Quadrants in a 3D orthogonal system 2186
7.10 Mutually perpendicular planes in $\mathbb{R}^3$ 2187
7.11 Right-handed coordinate system 2187
7.12 Grid representation of a 3D function in Maple™ 4.00b 2188
7.13 Shading interpolation representation of a 3D function in Maple™ 4.00b . . . 2189
7.14 Isolines representation of a 3D function in Maple™ 4.00b 2189
7.15 Better representation of a 3D function in Maple™ 4.00b 2190
7.17 Representation of a section of the pseudo-3D surface 2192
7.18 Representation of multiple sections of the pseudo-3D surface 2192
7.20 Scenario of thee points in a plane 2194
7.21 Example of symmetry through the y-axis of a function 2196
7.22 Example of symmetry through the x-axis of a function 2197
7.23 Example of symmetry through the origin O of a function 2197
7.24 Example of a positive vertical translation of a function 2198
7.25 Example of a negative vertical translation of a function 2198
7.26 Example of negative horizontal translation of a function 2199
7.27 Example of positive horizontal translation of a function 2200
7.28 Example of vertical stretch of a function 2201
7.29 Example of vertical compression of a function 2201
7.30 Example of horizontal stretch of a function 2202
7.31 Example of horizontal downscale of a function 2202
7.32 Summary of some of the types of monocities of univariate functions 2204
7.33 Most common types of charts 2205
7.35 Quick summary of applications/functions types 2210
7.36 Example of a very simple periodic function with period and amplitude 2211

7.37 Example of an even function 2212
7.38 Example of an odd function 2213
7.39 Univariate affine and linear function 2215
7.40 Different plots of simple power functions 2216
7.41 Different plots of simple exponential functions $(1^x, 2^x, 3^x, 4^x)$ in Maple™ 4.00b 2218
7.42 Different plots of logarithm $\ln(x) = \ln_e(x)$ in green and $\log_{10}(x)$ in red with Maple™ 4.00b 2219
7.43 Different plots of trigonometric functions in Maple™ 4.00b 2219
7.46 Some polynomials plotted with TikZ 2220
7.47 Example of a rational function 2221
7.48 Example of a step function 2222
7.49 Geometric concept of limit in $\mathbb{R}^3$ 2226
7.50 Left and right limit examples 2228
7.51 Graphical representation of a horizontal and vertical asymptote 2270
7.52 Graphical representation of an oblique affine asymptote 2272
7.53 Graphical representation of a parabolic branch example following x-axis 2273
7.54 Graphical representation of a curvilinear asymptote 2274
7.55 Basic techniques for finding limits 2275
7.56 Common types of discontinuities 2276
7.57 Example of a famous oscillating discontinuity 2278
7.58 Epigraph and Hypograph of a function 2281
7.61 Double logarithmic grid 2288
7.62 Simple horizontal logarithmic grid 2288
7.63 Graphical representation of the correspondence between the natural logarithm and the exponential 2289
7.64 Discrete convolution example 2296
7.65 Matrix Convolution 2300
7.66 Most common types of integral transforms with harmonized notation 2304
7.67 Example of a simple complex conformal mapping 2364
7.68 Example of a complex function (conformal mapping) 2391
7.69 $\mathbb{C}$-linear application example 2392
7.72 Practical Maple™ 17.00 example of simple C-linear application 2395
7.73 Transformation into polar coordinates of an ellipse 2396
7.74 Result of the Joukovski transformation in polar coordinates 2397
7.75 Important application example of the Joukovski function 2398
7.77 Plane representation of a well known holomorphic function... 2404
7.79 Representation of an important holomorphic function with its isolines 2407
7.80 Complex Logarithm plot in Maple™ 4.00b 2409
7.81 Complex Logarithm with TikZ 2410
7.83 Divergent Maclaurin series 2421
7.85 Representation of the various convergence of radius of $h(z)$ 2423
7.86 Focus on the complex representation to understand the reason for the divergence 2424
7.88 Representation of the deformation of a disc in a crown/ring 2426
7.89 Representation of a closed path with round trip 2427
7.91 Laurent series representation of $f(z)$ in Maple™ 4.00b 2436
7.92 Sum of the two Laurent series of $f(z)$ for both singularities in Maple™ 4.00b 2437
7.93 Essential singularity example with $e^{1/z}$ in Maple™ 4.00b 2439
7.94 Multiple isolated singularities surrounded by curvilinear paths 2442

7.95 Example of ultrametric distance with an orgchart 2451
7.96 Common types of distances 2455
7.97 Example of Lipschitz function 2459
7.98 Uniform continuity illustrated 2461
7.99 An open ball of radius r, centered at the point x 2464
7.100 Examples of closed balls of unit radius with different distances 2466
7.102 High-school example of a Hausdorff distance in the plane 2472
7.103 A naive maps homeomorphisms 2474
7.104 A more complete maps homeomorphisms 2474
7.105 Intuitive maps homeomorphism 2475

8.1 Schematic definition of the radian 2504
8.2 A chart to convert between degrees and radians 2505
8.4 Degrees, minutes of arc, seconds of arc 2506
8.5 Construction idea of elementary (circle) trigonometric functions 2508
8.6 Design principle of common circle trigonometric functions 2510
8.8 Remarkable angles of common circle trigonometric functions 2512
8.9 Plots of some common circle trigonometric functions 2513
8.10 Visual equivalences of some trigonometric relations 2515
8.11 Some identity angles and conventional trigonometric associated values for sine and cosine 2516
8.12 2D plot of the sinc function in Maple™ 4.00b 2517
8.13 Sine cardinal and its famous square equivalent 2517
8.14 Pseudo 3D plot of the sinc function in Maple™ 4.00b 2518
8.16 Basic construction for determining remarkable circle trigonometric identities for triangles 2519
8.17 Construction for the proof of the law of cosines 2525
8.18 Distance from a point inside a circle at its circumference 2526
8.19 Construction for the proof of the law of sines 2528
8.20 Hyperbolic functions domain 2535
8.21 Equivalence of hyperbolic functions and hyperbola 2536
8.22 Visual definition of circle and hyperbolic trigonometric functions 2538
8.23 Visual definition of hyperbolic trigonometric 2539
8.24 Construction to introduce the spherical geometry 2542
8.26 Configuration for the definition of the solid angle 2550
8.27 Configuration for the definition of the elementary solid angle 2550
8.28 Configuration for the definition of the elementary solid angle of revolution . . 2551
8.29 Difference between line, segment and semi-straight line 2555
8.30 Example of some coordinate systems 2557
8.31 Concept of one-dimensional standard 2557
8.32 Concept of two-dimensional standard 2558
8.33 Concept of three-dimensional standard 2558
8.34 von Koch curve 2559
8.35 Standard choice for Von Koch curve 2559
8.36 Sierpinski triangle fractal 2560
8.37 Three squares to cover the Sierpinski fractal 2561
8.38 Nine squares to cover the Sierpinski fractal 2561
8.39 Twenty seven squares to cover the Sierpinski fractal 2561

8.40 Representation of homothetic dimensions . . . 2562
8.41 The 23 definitions of Euclid illustrated . . . 2564
8.42 Illustrated summary of five Euclid's postulates . . . 2566
8.43 Illustrated example of a translation and a symmetry . . . 2575
8.44 A half circle protractor marked in degrees (180°) . . . 2576
8.45 Salient angle, reflex angle, adjacent angle . . . 2576
8.46 Representation of the movement of two angles . . . 2577
8.47 Sum of two adjacent angles . . . 2578
8.48 Sum of any two angles . . . 2578
8.49 Example of alternate angle . . . 2579
8.50 Alternate angle principle as used by Eratosthenes . . . 2580
8.51 Angle types summary . . . 2582
8.52 Positive or null oriented angle . . . 2584
8.53 Negative or null oriented angle . . . 2585
8.54 Internal + External angle . . . 2585
8.55 Geometric definition of the radian . . . 2588
8.56 Degrees-Radians conversions figure . . . 2589
8.57 Typical useful triangle angles in the study of Statics . . . 2592
8.58 Sum of angles of a plane triangle . . . 2594
8.59 Two triangles having equal side and two equal angles . . . 2595
8.60 Displacement showing the equality of the 2 triangles . . . 2595
8.61 First typical notation for congruent triangles . . . 2596
8.62 Second typical notation for congruent triangles . . . 2596
8.63 Third typical notation for congruent triangles . . . 2596
8.64 Representation of the mediator in isosceles triangle . . . 2597
8.65 Equilateral triangle . . . 2601
8.66 Right triangle . . . 2602
8.67 Another well known representation of right triangle . . . 2602
8.68 Right isosceles triangle . . . 2603
8.69 Classification of plane triangles . . . 2604
8.70 Bisectors, Median, Mediator and Height in a triangle . . . 2608
8.71 Pythagorean theorem illustrated . . . 2609
8.73 First approach of Thales' theorem . . . 2612
8.76 Circle and Radius . . . 2620
8.77 Circle and Diameter . . . 2621
8.80 Example of planar centroid . . . 2635
8.81 Most common type of geometric transformations . . . 2641
8.82 Example of scaling of center O . . . 2643
8.83 Scaling with shear . . . 2645
8.84 Shearing trough x-axis . . . 2646
8.85 Shearing trough y-axis . . . 2646
8.86 Example of rotation in the plane . . . 2647
8.87 Example of non-commutativity of the rotation matrix . . . 2651
8.88 Vocabulary of rotation in aeronautical science . . . 2655
8.89 Euler angles rotation convention . . . 2656
8.91 Reflection axes (author: Lauren K. Williams) . . . 2659
8.92 Reflection around y-axis . . . 2660
8.93 Reflection around x-axis . . . 2661

8.94 Reflection around origin O 2661
8.95 Summary of affine transformations 2662
8.96 Positive (elliptic), negative (hyperbolic) and flat (euclidean) curves 2663
8.97 Illustrated example of the violation of the fifth Euclide postulate 2664
8.98 Lunes in a sphere 2668
8.99 Illustration of the problem to research the geodesic on a sphere 2670
8.100 Some geodesics of trivial surface on trivial volumes 2671
8.101 Visual cone and corresponding projection circle 2681
8.102 Origin of Durer glass method illustration 2682
8.103 Definition of the planes and lines of projective geometry 2683
8.104 Example of what is a "table" in projective geometry 2684
8.110 Example of a single vanishing point 2702
8.111 Divergent shadows due to perspective 2703
8.112 Departure angle 2703
8.113 Example of secondary vanishing point 2705
8.114 Different perspective based on the values of the constants 2706
8.115 Aligned columns perspective problem 2706
8.116 One, two and three vanishing points in Adobe Illustrator 2708
8.117 Isometric angles 2714
8.118 Isometric volumes 2715
8.119 Isometric plane to find the cube dimensions 2716
8.120 Isometric plane to find the isocircle dimensions 2720
8.121 Orthogonal projection in real life... 2722
8.123 Canonical transformation for orthogonal projection 2724
8.124 Various spherical and cylindrical projections 2727
8.125 Technical view of various spherical projections techniques 2728
8.126 Schlegel diagram in pseudo-perspective (stereographic projection) 2730
8.128 Earth's stereographic projection 2732
8.129 Earth's cylindrical projection 2739
8.130 Earth's Mercator projection 2740
8.131 Mercator map projection with the true sizes and shapes of the countries overlaid 2740
8.132 Mercator projection function construction 2741
8.133 Mercator world map as seen in 11569 2743
8.134 Lambert's equivalent projection (or Lambert's equal area projection) 2743
8.135 Earth's Lambert projection 2745
8.136 Sansons or Sanson-Flamsteed or Sinuoidal projection 2746
8.137 Mollweide projection 2747
8.138 Hammer projection 2751
8.141 Non-exhaustive orgchart of various projections methods 2753
8.142 Some common types of projections 2754
8.143 Summary of the seven laws of linear perspective 2756
8.144 Railway convergence perspective effect 2757
8.147 Summary of homogeneous coordinates transformations 2760
8.149 Various intersections between a cone and a plane 2770
8.150 Shock wave intersecting the ground 2770
8.151 Maple™ 4.00b plot to obtain a parabola by the intersection of a cone and a plane 2771

8.152 Maple™ 4.00b plot to obtain a hyperparabola by the intersection of a cone and a plane . . . 2772
8.153 Maple™ 4.00b plot to obtain an ellipse by the intersection of a cone and a plane 2773
8.154 Maple™ 4.00b plot to obtain a circle by the intersection of a cone and a plane 2773
8.155 Definition of the focus (F), direction (D) of the conical and eccentricity (e) . 2775
8.156 Representation of the parameters of the conical for the parabola . . . 2777
8.157 Parabola = set of points equidistant from L and F . . . 2777
8.158 Representation of the parameters of the conical for the ellipse . . . 2781
8.159 Hand-made construction of an ellipse . . . 2783
8.160 Apogee and Perigee . . . 2786
8.161 Representation of the parameters of the conical for the hyperbola . . . 2788
8.162 Dandelin Sphere . . . 2793
8.163 Positive definite, positive-semidefinite and indefinite matrices visual representation . . . 2799
8.164 Example of a plane in Maple™ 4.00b . . . 2802
8.165 Representation of the search of the distance from a point to a line . . . 2807
8.166 FG-squircle representation in Maple™ 4.00 . . . 2813
8.167 Epicycloid path . . . 2816
8.168 Epicycloid sketch for proof . . . 2817
8.169 Archimedean Spiral . . . 2818
8.170 Exponential spiral in 2D and 3D of the type r_0e^{bt} . . . 2819
8.171 Boost-glide atmospheric re-entry . . . 2819
8.172 Hypocycloid path . . . 2822
8.174 Representation of a cone . . . 2823
8.176 Parametric representation of a cone in Maple™ 4.00b . . . 2825
8.177 Representation of a $\mathcal{S}^2$ sphere . . . 2826
8.179 Parametric representation of a sphere in Maple™ 4.00b . . . 2827
8.180 A great circle divides the sphere in two equal hemispheres . . . 2828
8.181 Representation of an ellipsoid . . . 2834
8.183 Parametric representation of an ellipsoid in Maple™ 4.00b . . . 2835
8.184 Oblate spheroid (on the left), Prolate spheroid (on the right) . . . 2836
8.186 Parametric representation of a cylinder in Maple™ 4.00b . . . 2838
8.188 Parametric representation of a paraboloid in Maple™ 4.00b . . . 2840
8.190 Hyperbolic Rules Surface . . . 2843
8.191 Hyperbolic Circles Surface . . . 2843
8.192 Parametric representation of a hyperboloid of one sheet in Maple™ 4.00b . . 2844
8.193 Two hyperboloids (of one sheet and two sheets) which are asymptotic to an identical elliptic cone . . . 2845
8.194 Common conical surfaces . . . 2847
8.195 Parametric representation of a torus in Maple™ 4.00b . . . 2849
8.196 Parametric curve and parametric surface . . . 2851
8.197 Parametric representation of an helix in Maple™ 4.00b . . . 2856
8.200 Plot of the isoclines in MATLAB® 5.0.0.473 . . . 2858
8.201 Projection of the isoclines in MATLAB® 5.0.0.473 . . . 2859
8.202 Plane representation of isoclines with coloured gradients in MATLAB® 5.0.0.473 . . . 2859
8.204 Isoclines association projected with color gradients in MATLAB® 5.0.0.473 . 2860
8.207 Representation of an osculating circle . . . 2867

8.208 Representation of the trihedron . . . 2869
8.213 Representation of a surface patch . . . 2879
8.214 Example of an arbitrary plane polygon . . . 2888
8.215 Example of a crossed polygon . . . 2889
8.216 Example of a concave polygon . . . 2890
8.217 Example of a convex polygon . . . 2890
8.218 Example of a polygon (pentagon with 5 edges and 5 vertices) . . . 2891
8.219 Pentagon with all its 10 segments . . . 2891
8.220 Example of segments starting from a chosen point . . . 2892
8.223 Some regular polygons (subfamily of convex polygons) . . . 2893
8.224 Example of rectangle . . . 2894
8.225 Example of square . . . 2895
8.226 Example of an unspecified triangle with inscribed circle and bisectors . . . 2896
8.227 Unspecified triangle divided into two right triangles . . . 2896
8.228 Reminder of the construction for the cosine's theorem proof . . . 2897
8.230 Example of isosceles triangle . . . 2900
8.231 Right angle isosceles triangle . . . 2901
8.232 Equilateral triangle . . . 2902
8.233 Illustrative figure for Viviani's theorem . . . 2904
8.234 Example of a right angle triangle . . . 2905
8.235 Example of trapezoid . . . 2906
8.236 Quadrilateral hierarchy (including Trapezoid classification) . . . 2907
8.237 Parallelogram example . . . 2908
8.239 A regular hexagonal grid . . . 2909
8.240 James Webb Space Telescope . . . 2910
8.241 TV broadcast reflector antenna . . . 2910
8.242 Regular hexagon details . . . 2910
8.247 Example of a rhombus . . . 2914
8.248 Example of a circle . . . 2915
8.249 Surface S_d of a disc cutted part . . . 2917
8.250 Example of an ellipse with its visible foci and eccentricity e . . . 2918
8.251 Surface elements $\mathrm{d}S$ in 2D . . . 2922
8.252 Simple volumes bounded by a surface . . . 2923
8.253 Polyhedron vocabulary . . . 2924
8.254 General parallelepiped . . . 2925
8.255 Cube, Cuboid and Rhombohedron . . . 2926
8.258 Example of pyramid (irregular basis) but non-oblique . . . 2931
8.259 Example of right prism . . . 2933
8.260 Angles between the edges of a polyhedron . . . 2936
8.261 Known polyhedra made of 12 pentagons . . . 2938
8.262 Example of regular tetrahedron . . . 2940
8.265 Example of regular hexahedron (cube) . . . 2942
8.266 Example of regular octahedron seen from above . . . 2943
8.267 Example of the octahedron . . . 2944
8.268 Searches of the position of the equilateral triangle . . . 2945
8.269 Construction of a regular icosahedron . . . 2946
8.270 Representation of the regular icosahedron . . . 2947
8.271 Pentagons in the regular icosahedron . . . 2948

8.272 Representation of one of the pyramids of the regular icosahedron 2948
8.273 Parallelogram in a pentagon 2949
8.274 Regular dodecahedron 2952
8.278 Example of a cylinder 2959
8.279 Cone 2963
8.280 Example of a Sphere 2965
8.281 Sphere Grid 2965
8.282 Calculation of the volume of a sphere by a stack of discs 2966
8.283 Representation of a surface element in spherical coordinates 2968
8.284 Spherical cap surface derivation trick 2969
8.285 Torus 2973
8.286 Ellipsoid 2978
8.287 Oblate spheroid on the left, prolate one on the right 2978
8.288 Paraboloid of revolution 2983
8.289 Hyperbolic paraboloid 2983
8.290 Elliptic paraboloid 2984
8.291 Quadric surfaces at a glance 2985
8.292 Gabriel's horn with Maple™ 4.00b 2986
8.295 2D and 3D surface and volumes elements techniques in different coordinates systems 2992
8.296 Graphs basic vocabulary 3000
8.297 Simple Graphs 3000
8.298 Example of a multigraph 3001
8.299 Example of planar graphs 3002
8.300 Graph with 2 faces, 4 vertices and 4 nodes 3002
8.301 Tree with one edge and two vertices 3003
8.302 Graph with at least one cycle 3003
8.303 Regular polygons seen as graphs 3004
8.304 Example of graphs subsets 3007
8.305 A few possible 2-regular graphs 3008
8.306 Examples of subgraphs and induced graphs 3009
8.308 Example of complete graphs 3010
8.309 Star of David 3011
8.310 Star of David completed 3011
8.311 Stable subgraph example 3012
8.312 Simple path example 3012
8.313 Various connected graphs 3014
8.314 Example of a spanning tree 3015
8.315 Example of a minimum spanning tree 3015
8.316 PERT (Program Evaluation Review Technique) diagram 3017
8.317 Same graph but seen in Microsoft Project 97 3017
8.318 Examples of cyclic graphs 3018
8.319 Example of a connex graph containing a cycle 3018
8.320 Example of a connex graph containing no cycles 3018
8.321 Example of a forest (3 connex graphs) 3019
8.322 Example of Hamiltonian or Non-Hamiltonian graphs 3020
8.323 Antique map of the city of Königsberg 3020
8.324 Königsberg equivalent graph 3021

8.325 Example of complementary graph . . . 3023
8.326 Example of self-complementary graph . . . 3023
8.327 Example of a $K_{3,3}$ bipartite graph . . . 3023
8.328 Isomorphic graphs with their adjacency matrix . . . 3024
8.329 Example of Markov chain (nor antireflexive or symmetric) . . . 3026
8.331 Example of graph of categories . . . 3032
8.332 Braid and corresponding string (node diagram) . . . 3035
8.333 Particular presentation of a braid . . . 3036
8.334 Example of the problem of the braid representation . . . 3036
8.336 First example of invariant braid . . . 3036
8.337 Second example of invariant braid . . . 3037
8.338 Translation of crossing and preserving the trivial invariance . . . 3037
8.339 Length invariance . . . 3037
8.340 Example of multiplication of two braids . . . 3037
8.341 Check of associativity of braids (by the multiplication) . . . 3038
8.342 Control of the commutativity of braids (by multiplication) . . . 3038
8.343 Check of the existence of a neutral element for braids (by the multiplication) 3038
8.344 Check of the existence of an inverse element for braids (by the multiplication)3039
8.345 Indexing braids tendrils . . . 3039
8.346 Indexation convention . . . 3039
8.347 Indexation tendril trap . . . 3039
8.348 Difference between tendril and cross . . . 3040
8.349 Braid and corresponding string (node diagram) . . . 3040
8.350 Example of a node in real life . . . 3041
8.351 Knot representation by mathematicians . . . 3041
8.352 Equivalence between an open and a closed knot . . . 3041
8.353 2 different perspectives of the same trefoil knot . . . 3042
8.354 Trefoil plot . . . 3042
8.355 Searching for equality between mirror symmetry of the clover node . . . 3042
8.357 Example of a Perko pair . . . 3043
8.358 Knots that are simply an unknot . . . 3043
8.359 Refresh of the definition of the multiplication of two knots . . . 3044
8.360 Check of the commutativity of knots (by the multiplication) . . . 3044
8.361 Check of the existence of a neutral element for the knots (by the multiplication)3044
8.362 Check of the commutativity of knots (by the multiplication) . . . 3045
8.363 Check of the existence of an inverse element for the knots (by the multiplication)3045
8.365 Conway knot . . . 3047
8.366 Example of a succession of knots on an open knot . . . 3049
8.367 Separation of the knot . . . 3049
8.368 Tait's classification . . . 3050
8.369 Knotted knot? . . . 3052
8.371 Trivial Knot . . . 3053
8.372 Trefoil Knot . . . 3053
8.373 Example of RNA or DNA strand . . . 3054
8.374 Mirror image by crosses inversion . . . 3056
8.375 Mirror of trefoil knot . . . 3056
8.376 Simplistic application example of the Reidemeister movements . . . 3057
8.377 Knot isotopie . . . 3058

8.378 Interlaces with three convex components . . . 3059
8.379 Example of associated interlaces . . . 3059
8.380 Example of non-interlaces notation . . . 3060
8.381 Trefoil knot and its diagram . . . 3061

9.1 Reminder of hierarchies of concepts in modern experimental science . . . 3071
9.2 Based ISQ units and derived units . . . 3075
9.3 Based 12019 SI . . . 3078
9.5 Copenhagen survey on how the "Big Bang" should be understood . . . 3117
9.6 Okun cube . . . 3125
9.7 Reminder of scalar vs vector quantities in physics . . . 3138
9.8 Isotropic vs Homogeneous . . . 3143
9.9 Landscape of Analytical and Classical Mechanics . . . 3155
9.11 Variation idea of Least Action Principle . . . 3164
9.15 Particular case of coplanar system . . . 3195
9.16 Particular example of a measured coplanar system (dyname) . . . 3196
9.17 Simplified representation of forces for treatment (translated dyname) . . . 3196
9.18 Representation of the resultant of forces of a suspended bridge . . . 3197
9.19 Coplanar symmetrical load porting . . . 3197
9.20 Typical example of a free body diagram . . . 3198
9.21 Binary system center of mass (profile view) . . . 3217
9.22 Binary system center of mass when in motion . . . 3217
9.23 Motion parallax . . . 3230
9.24 Tangential and Normal acceleration . . . 3238
9.25 The principle of Galilean relativity illustrated . . . 3240
9.26 Example repositories in uniform relative motion with respect to each other . . 3241
9.27 Angular momentum right-handed rule . . . 3245
9.28 Gyroscope remaining upright . . . 3247
9.29 Illustration of momentum . . . 3248
9.30 Example of some everyday lever arm . . . 3252
9.31 Example of couple of force . . . 3253
9.32 Balanced moment example . . . 3254
9.36 Sarens SGC-250 . . . 3260
9.37 Lami’s theorem illustrated . . . 3261
9.38 PASCO® scientific static system to check Lami’s theorem in schools . . . 3262
9.39 Ballistics configuration study . . . 3263
9.42 Ballistic safety parabola . . . 3266
9.44 Lissajous figures . . . 3289
9.45 Several Lissajous figures for different phase delays . . . 3290
9.46 Centrifugal acceleration . . . 3291
9.47 Reminder of typical rotation descriptions . . . 3292
9.48 Summary Moment of inertia of some planar shapes . . . 3300
9.49 Summary of very common bodies moments of inertia . . . 3301
9.50 Conservation of the angular momentum on variation of inertia momentum . . 3303
9.52 Conservation of total angular momentum by variation of intern angular momentum . . . 3304
9.53 Kepler satellite reaction wheels . . . 3305
9.54 ISS (International Space Station) control moment gyroscopes . . . 3306

9.55 Illustration of the ellipsoid of inertia . . . 3316
9.57 3D Gyroscope . . . 3325
9.58 Symmetric weighing gyroscope . . . 3326
9.59 Foucault gyroscope . . . 3327
9.60 Photo of a weighted gyroscope with rotor . . . 3328
9.61 Gyroscope spinner . . . 3330
9.63 Spinning top in precession only . . . 3333
9.69 Tilt and rotation period of some well known celestial bodies . . . 3347
9.71 Naive scenario of gravitational potential energy . . . 3355
9.72 Graph of non-relativistic effective gravitational potential . . . 3359
9.73 Potential theory summary . . . 3361
9.74 Elastic collision with $m_1 \gg m_2$ . . . 3371
9.75 Two balls rebound . . . 3376
9.76 Example of moving and rotating reference frame relatively to a fixed reference frame . . . 3384
9.78 Illustration of the Earth's Coriolis force . . . 3393
9.79 Generation of cyclones by the Coriolis force . . . 3393
9.80 North vs Southern Coriolis effect . . . 3394
9.81 Eötvös effect illustrated (source: flatearth.ws) . . . 3399
9.82 Newton's cradle (pendulum) . . . 3403
9.83 Schematic ideal simple pendulum . . . 3406
9.84 Phase space $\theta - \omega$ of the simple pendulum in Maple™ 4.00b . . . 3408
9.85 Schematic ideal physical pendulum . . . 3413
9.87 Schematic ideal physical one degree of freedom pendulum . . . 3416
9.88 Illustration of the amplitude damping of an elastic pendulum . . . 3420
9.89 Potential and Kinetic energy balancing of an elastic pendulum . . . 3420
9.90 Two degrees of freedom ideal elastic pendulum . . . 3422
9.91 Ideal conical pendulum . . . 3423
9.92 Ideal torsion pendulum . . . 3425
9.93 Schematic view of Foucault's pendulum . . . 3427
9.94 Study diagram of Foucault's pendulum . . . 3428
9.96 Schematic Huygens' pendulum . . . 3433
9.97 Reproduction of a coster clock . . . 3433
9.98 Typical path to attack the brachistochrone problem without bias . . . 3434
9.99 Comparison of the brachistochrone with other intuitive paths . . . 3439
9.100 Illustration of the ideal double pendulum . . . 3441
9.101 Plot with Maple™ 4.00b of the double pendulum external mass . . . 3444
9.102 Long exposure of double pendulum exhibiting chaotic motion tracked with a LED . . . 3445
9.103 A famous business application (SEGWAY™) of mobile inverted pendulum . . . 3448
9.104 A schematic drawing of the inverted pendulum on a cart . . . 3449
9.105 Amontons-Coulomb friction saw-tooth model . . . 3454
9.109 Banked curve analysis . . . 3459
9.110 Forces between a rope and capstan . . . 3462
9.112 Rotary friction welding principles . . . 3473
9.113 Commutative diagram of harmonic wave properties . . . 3480
9.114 Beating with large difference in wavelength . . . 3482
9.115 Beating with small difference in wavelength . . . 3482

9.116 Group and Phase velocities difference . . . 3483
9.117 Double fixed standing wave . . . 3485
9.118 Stretched string scheme . . . 3486
9.119 First two basic modes of a fixed string . . . 3493
9.120 Some vibration modes of a circular membrane . . . 3508
9.121 Circular Chladni plate . . . 3509
9.122 Dispersive Wavepacket in Maple™ 4.00b . . . 3515
9.123 Non-dispersive Wavepacket in Maple™ 4.00b . . . 3515
9.124 Dispersive sinusoidal waves with Maple™ 4.00b . . . 3516
9.125 Non-dispersive sinusoidal waves with Maple™ 4.00b . . . 3516
9.126 Phase representation of the two harmonics in the Gauss plane . . . 3517
9.127 Amplitude modulation (AM) illustrated . . . 3519
9.129 Representation of the envelope of the amplitude modulation with Maple™ 4.00b . . . 3520
9.130 Waves interferences of two loudspeakers . . . 3522
9.131 AM modulation principle . . . 3524
9.132 Waves interferences of two loudspeakers . . . 3524
9.133 Examples of correlation vs mutual information. . . . 3529
9.134 Graphical representation of the conditional entropy and the mutual information 3532
9.135 Entropy of a binary variable with $\lambda = 1$ . . . 3534
9.136 High Entropy Alloy . . . 3536
9.137 Example of change of entropy of a closed system . . . 3552
9.138 Maxwell velocity distribution in a monoatomic ideal gas . . . 3559
9.139 $\Omega_{\text{M.B.}}$ states in the Maxwell-Boltzmann distribution . . . 3562
9.140 Fermions, Hadrons, Leptons, Mesons, Baryons, Quarks . . . 3572
9.141 $\Omega_{\text{B.E.}}$ states in the Fermi-Dirac distribution . . . 3573
9.142 Plot of the Fermi-Dirac distribution with MATLAB® 5.0.0.473 . . . 3575
9.143 Fermi-Dirac Distribution . . . 3575
9.144 $\Omega_{\text{B.E.}}$ states in the Bose-Einstein distribution . . . 3577
9.145 Plot of the Bose-Einstein distribution with MATLAB® 5.0.0.473 . . . 3578
9.146 Convergence of Maxwell-Boltzmann, Fermi-Dirac and Bose-Einstein in high temperature domain . . . 3579
9.147 Summary of the Maxwell, Bose-Einstein and Fermi-Dirac distributions . . . 3581
9.148 Brownian Motion in a two-dimensional lattice . . . 3583
9.152 Illustration of the naive states of matter . . . 3597
9.153 Representation of the Gay-Lussac law with pressure . . . 3601
9.154 Common type of thermodynamic systems . . . 3603
9.155 Most common type of thermodynamic processes illustrated . . . 3604
9.156 Main thermodynamics processes at a glance . . . 3605
9.157 Two typical thermodynamics cycles . . . 3606
9.158 Intensive vs Extensive variables . . . 3608
9.159 Common phase transitions . . . 3610
9.160 Thermal expansion joints like these allow bridges to change length without buckling . . . 3617
9.161 Summary of the four laws of the thermodynamics . . . 3621
9.162 Debye vs Einstein model . . . 3627
9.163 Water latent heat process . . . 3633
9.164 Vibration bonds source of the internal energy in solid and liquids . . . 3639

9.165 Proportion of internal energy type by common state of matter 3639
9.167 Earth's entropy decrease thanks to Sun 3665
9.168 A schematic representation of a heat engine 3669
9.169 Carnot cycle illustration 3671
9.170 Open system in translation or not in a repository and transferring energy . . . 3680
9.172 Thermal conductivity values for various materials 3689
9.173 Diffusion between pure water and salt water 3699
9.174 Conceptual black-body 3707
9.175 Black-body laboratory radiator 3708
9.177 Plot of the three radiation laws (Planck's, Rayleigh-Jeans, Wien) in a log-log scale for comparison 3721
9.179 Representation of Planck's law (distribution) with Wien's displacement . . . 3728
9.180 Planck's law at various temperatures 3729
9.181 Sun's radiation spectrum in function of frequency and logarithmic scale . . . 3730
9.182 Sun's radiation spectrum in function of frequency at sea level 3730
9.183 Comparison of theoretical black-body and our Universe cosmic microwave background 3731
9.184 Illustration of normal and tangential constraints (stresses) 3736
9.185 Hydraulic press owned by the GIVA's Forgiatura in Italy 3737
9.186 Behaviour under stress/compression for some materials 3740
9.187 Tensile Stress-Strain Diagram 3741
9.188 Toughness schema of common material 3742
9.189 Strength vs Density of some materials 3743
9.190 Variation of various steel strengths as function of temperature 3744
9.191 Illustration of the effect of a normal stress 3745
9.192 Idealized simple cubic lattice of atoms separated by the "springs" of inter-atomic forces that hold them in equilibrium positions 3746
9.194 Illustration of the effect of a shear stress 3748
9.195 Cube under normflexural modulusal stresses 3751
9.201 Final material block situation 3757
9.205 P-wave and S-Wave in solids 3765
9.207 S-Wave front view 3766
9.208 Zoom on three layers of a shear wave 3767
9.212 Very approximative viscosity characteristics of different typical fluids 3777
9.213 Laminar vs Turbulent flow 3779
9.214 Non-viscous fluid in laminar flow from in a tube of parallel lines of current lines converging 3784
9.215 Bernoulli's theorem illustrated 3784
9.216 Application in everyday life of the continuity equation 3785
9.217 Zoom of a fluid region for the Bernoulli's theorem 3786
9.218 Condensation visible over the upper surface of an Airbus A340 wing 3788
9.219 Illustration of the profile of a wing with the corresponding pressures and speeds in a wind tunnel 3789
9.223 Wing streamlines in wind tunnel in real life 3791
9.224 Torricelli's experiment (author: Izaak Neutelings) 3791
9.225 Torricelli experiments configuration 3794
9.226 Tank emptying 3795
9.227 Torricelli jets envelope illustration 3796

9.228 Special case of siphon . . . 3798
9.229 Photo of the principle of communicating vessels . . . 3800
9.230 Hydrostatic paradox illustrated . . . 3801
9.231 Fundamental theorem of hydrostatics illustrated . . . 3802
9.232 Application of communicating vessels with water towers . . . 3802
9.233 Venturi tube configuration study . . . 3804
9.234 Venturi effect illustrated . . . 3804
9.235 Industrial Venturi tube . . . 3806
9.236 Venturi tube on a CESSNA . . . 3806
9.237 Horizontal Pitot tube . . . 3807
9.238 Real Pitot tube . . . 3807
9.239 Vertical Pitot tube . . . 3808
9.240 Elementary parallelepiped extracted from a static fluid at equilibrium . . . 3811
9.241 Regular elemental tetrahedron extracted from a fluid at equilibrium . . . 3813
9.245 Rotating bucket . . . 3833
9.246 Rotating bucket experiment . . . 3835
9.247 Example of a simulated vortex flow . . . 3838
9.248 Earth currents and vortex between June 12005 to December 12007 . . . 3838
9.250 Effect of Reynolds number on the drag coefficient C_D of a smooth sphere of diameter d . . . 3857
9.251 Effect of Reynolds number on the drag coefficient C_D of a smooth sphere with regime . . . 3858
9.252 Drag coefficient comparison of a sphere and cylinder . . . 3858
9.253 Various fluid regimes for cylinder with corresponding Reynolds number . . . 3859
9.254 Visualisation of the vortex street behind a circular cylinder in air . . . 3860
9.255 Visualization of the vortex street around Guadalupe island . . . 3860
9.256 Visualisation of the vortex street around Jupiter red spot . . . 3860
9.257 Cloud's sky watcher chart . . . 3862
9.259 Archimedes principle . . . 3876
9.260 Demonstration of the Archimedes' principle . . . 3877
9.262 Example of a gas (SF_6) contained in a container without top . . . 3889
9.263 Boyle-Mariotte's law illustration . . . 3890
9.264 PV diagram for oxygen . . . 3891
9.265 Charles' law illustration . . . 3891
9.266 Gay-Lussac's law illustration . . . 3893
9.267 Avogadro's law illustration . . . 3894
9.270 Gas contained in a tank . . . 3901
9.271 Van der Waals P-V diagram for CO_2 . . . 3907
9.272 Van der Waals general P-V diagram . . . 3908
9.273 PVT diagram of the ideal gas law . . . 3909
9.274 PVT phase diagram for real gas . . . 3909
9.275 PVT phase diagram for water H_2O . . . 3910
9.277 Dalton's law illustration for air at normal temperature . . . 3924
9.278 Numerical application of Dalton's law . . . 3925
9.280 Simplified illustration of free mean path . . . 3927
9.281 Tank trailer vacuum implosion to atmospheric pressure . . . 3929
9.282 Typical pressurised containers with their sizes and respective capacities . . . 3930
9.283 Magnus effect on a rotating sphere/ball . . . 3939

9.284 Roberto Carlos "impossible free kick" 3942
9.285 Some well known everyday applications of plasma in the 120th and 121st century 3945
9.286 Plasma electronic oscillation 3947
9.287 Minimum value of $n_e\tau_E$ required for self-heating, for three fusion reactions . 3953
9.288 Example of a naive case of screening of an electron by multiple dipoles . . . 3955
9.289 A single skip of a radio wave using the ionosphere 3959
9.290 Representation of a moving electron as a dipole moment 3960

10.2 PASCO® scientific Coulomb's law apparatus for schools 3970
10.3 Electric field of an infinite line charge 3971
10.4 Dipole profile for $q_1 = -q_2$ and $q_1 = q_2$ along y-axis 3974
10.5 3D dipole profile with $q_1 = q_2$ 3974
10.6 Various typical electric fields configurations 3975
10.7 Various voltage values relative to human body health 3978
10.10 Configuration for the analysis of the electric field of an infinite straight wire . 3983
10.11 Naive illustration of an electric dipole 3984
10.12 Electric field at a distant point M of the dipole 3985
10.13 Polar plot of the field lines and equipotentials of an electric dipole 3988
10.14 Electric field lines for two charges of opposite signs in Maple™ 17.00 3989
10.15 Potential well and equipotentials of the dipole in Maple™ 17.00 3990
10.16 2D profile of electric field and potential of the dipole in Maple™ 17.00 . . . 3990
10.17 Naive illustration of an electric dipole in a uniform electric field 3991
10.18 Example of what happens in a solution with an ion 3992
10.19 Hydration of salz 3992
10.20 Some capacitive dipoles 3998
10.21 Schematic diagram of plane and parallel capacitor 3999
10.22 Schematic diagram of cylindrical capacitor 4000
10.23 Schematic diagram of spherical capacitor 4001
10.24 Real capacitor 4002
10.25 Illustration of the work path in an electric field 4006
10.26 Magnetic field around a straight infinite wire 4009
10.27 Ampère observer hand rule 4010
10.28 Infinite solenoid (coil of wire) 4014
10.29 Linear solenoid analogy with (dipole) magnet 4015
10.30 The shielding room at Paul Scherrer Institude 4017
10.31 Toroidal coil 4018
10.32 Toroidal field 4018
10.33 Schematic rectangular electromagnet 4019
10.34 Lifting electromagnet 4021
10.35 Playful lifting electromagnet 4021
10.36 Magnetic field for a current loop 4025
10.37 Infinite straight wire 4027
10.39 Microscopic dipoles point of views 4030
10.40 Magnetic dipole (magnet) broken with small spin naively highlighted 4031
10.41 Earth viewed as an ideal magnetic dipole 4032
10.42 Standard magnetic compass and Dip circle 4033
10.44 Magnetic dipole 4037

10.45 Visualizing dipole magnetic field with Python™ 4041
10.46 Magnetic moment right-handed rule 4042
10.47 Earth’s magnetic field 4047
10.48 Torque on a bar magnet in a uniform magnetic field 4050
10.49 Beam of electrons moving in a circle, due to the presence of a magnetic field 4054
10.50 Bubble chamber 4055
10.51 Illustrated concept of the Y. Aharonov and D. Bohm experiment 4061
10.52 Laplace’s force right-handed rule 4064
10.53 DC current through a metallic ribbon for the study of classical Hall effect . . 4065
10.54 Cyclotron principle 4069
10.55 Dempster mass spectrometers principle 4070
10.56 Helical path of charge particle in a uniform constant magnetic field 4071
10.57 Van Allen radiation belts around Earth 4075
10.58 Bubble chamber helicoidal path 4077
10.60 Stern-Gerlarch experiment principle and result 4084
10.62 Gauss law 4098
10.63 Lorenz law illustration 4101
10.64 Photo of a PEG with a mobile phone 4103
10.65 Photo of a Betatron 4104
10.66 Vector Magnetic Potential $\vec{A}$ 4116
10.67 Some properties of the medium "space" (vacuum) 4141
10.68 PASCO® scientific speed of light apparatus for schools 4146
10.69 Schematic of the PASCO® scientific speed-of-light apparatus 4148
10.71 Spherical wavefront 4160
10.73 European synchrotron 4163
10.74 Synchrotron source producing electromagnetic radiation 4164
10.76 Synchrotron radiations of a moving charge 4193
10.82 Plot of some 2D Hertz dipole radiation pattern 4213
10.83 Plot of some 3D Hertz dipole radiation pattern 4214
10.84 Power radiation Hertzian dipole 2d and 3d 4215
10.85 Circuit basic vocabulary 4220
10.87 Example of a real non-trivial electric circuit 4224
10.89 Typical conductor electrons behaviour without voltage 4226
10.90 Typical conductor electrons behaviour with voltage 4226
10.91 Some resistive dipoles 4234
10.92 Resistor color codes 4235
10.93 AWG Gauge 4236
10.94 Real and imaginary parts of the effective permittivity $\varepsilon(\omega)$ 4242
10.95 Different inductance models 4252
10.96 Electromotive force schematic concept 4256
10.97 Timeline of batteries 4257
10.99 Section of a long coaxial cable 4265
10.100 Configuration study for skin effect 4269
10.102 k-space for a cubic crystal lattice 4287
10.103 Fermi surface of copper 4288
10.104 Fermi surface of copper with 2D ACAR method 4289
10.105 Silicium vs Germanium band and crystal structure 4295
10.106 Representation of band structures in different materials 4301

10.107 Association of band structure with Fermi-Dirac function 4302
10.108 Metal-insulator transitions 4303
10.109 Electric cables for accelerators at CERN 4309
10.110 Observation of a superconductor in levitation in magnetic field due to Meissner effect 4312
10.111 Various types of light punctual sources 4316
10.112 Application of Thales' theorem on punctual sources 4317
10.113 Shadow cone with proper shadow example 4319
10.114 Shadow of a ball with a point like source of light 4319
10.115 Shadow cone with penumbra example 4320
10.116 Penumbra and umbra of a ball with an extended like source of light 4320
10.117 Reflection of shadows (light surrounding shadows) by mirrors 4321
10.118 Refraction of shadows (light surrounding shadows) by mirrors 4322
10.119 Caustic on water surface 4323
10.120 Magnetic shadow effect 4323
10.121 Magnetic shadow effect explained 4324
10.123 Naive eye structure summary 4326
10.124 Whole standard electromagnetic spectrum 4329
10.125 Temperature spectrum 4329
10.126 SN 1054 observed under different wavelengths and different telescopes . . . 4330
10.127 Earth viewed under different wavelengths 4330
10.128 Various types of Ultraviolet rays 4331
10.129 Electromagnetic spectrum energy and atmosphere absorption 4332
10.130 Additive synthesis (RGB) 4333
10.131 Subtractive synthesis (CMYK) 4335
10.132 Additive and Subtractive experiment 4336
10.133 Lambertian emittor 4344
10.134 Typical experimental setup used in universities to check Lambert's cosine law 4345
10.135 Geometry used in the Phong reflection model 4347
10.137 12003–12004 mean annual clear-sky and total-sky albedo 4351
10.138 Vantablack painted face... 4352
10.139 Vocabulary for the study of geometrical optics 4353
10.140 Types of refractions and reflections 4355
10.141 Funny water reflection effect 4355
10.142 Water reflection ray diagram 4356
10.144 Light diffusion when reflected from a rough surface 4357
10.145 Positive and negative refraction index material behaviour 4365
10.146 Figure allowing to connect phase velocity and refractive index 4367
10.147 Critical angle schematic illustration (raytracing) 4369
10.149 Sun Rainbow after the rain 4372
10.153 Case of a bundle of parallel rays of identical color in a spherical rain drop . . 4376
10.154 Rainbow angle 4379
10.155 Naive illustration of the generation of the primary rainbow 4381
10.157 Double rainbow mechanism 4382
10.158 Moonbow 4383
10.159 Different photos of Brocken spectrums 4384
10.160 22° Sun Halo 4385
10.161 Photographs showing two hexagonal POP ice crystals 4386

10.163 Most common types of Sun Halos and Arcs 4392
10.164 Various types of Halos in one shot in real life 4393
10.165 Cloud iridescence photo 4396
10.169 Curvature of the rays near an overheated ground 4400
10.170 Critical angle schematic illustration (raytracing) 4404
10.172 Shift of the perceived form and position of the Sun due to the atmospheric refraction 4409
10.173 Selenelion diagram 4409
10.174 Oak Ridge National Laboratory Nuclear Reactor 4414
10.175 Representation of the optical axis concept 4416
10.176 Why images are reversed when passing through a thin spherical lens 4419
10.177 Representation of the concept of stigmatism 4420
10.178 Parabolic antenna for reception or emission 4423
10.179 Off-axis receiver dish construction principle 4424
10.180 Off-axis satellite dish 4425
10.182 Position of first and second focal points in a dioptre 4428
10.183 Concept of spherical dioptre 4429
10.184 Zoom-Nikkor AF-S 80-200 mm f/2.8D IF-ED 4429
10.185 Fisheye-Nikkon 6 mm f/2.8 simplified cut 4430
10.186 Technical representation of a lens 4431
10.188 Focal of a thin lens with parallel entering light rays 4434
10.189 Human eye adaptative refractive power mechanism 4435
10.191 First focal point 4437
10.192 First focal point with bi-convex lens behaviour 4438
10.193 Second focal point 4438
10.194 Second focal point with bi-convex lens behaviour 4439
10.195 Examples of diverging and converging lenses 4440
10.196 Convergent and divergent lens concepts 4440
10.199 Magnification optical illusion 4444
10.200 Biconvex lenses typical configurations and applications 4446
10.201 Concave mirror 4447
10.202 Left concave spoon-mirror, Right convex spoon-mirror 4447
10.203 Full plane-curved lens 4448
10.204 Fresnel lens 4449
10.205 Parabolic trough collectors used to generate electricity in southern California 4450
10.206 Field of view real case with Earth's photo 4451
10.207 Field of view real case with Earth and Moon 4451
10.208 Mosaic of the impact of angle of view on deformation of objects 4455
10.210 Human eye optics 4458
10.212 Triangular prism dispersing white light ray 4462
10.214 Pentaprism roof 4466
10.215 Mirror double deflection property 4467
10.216 Pentaprism detailed design 4468
10.217 Illustration of Huygen's construction of a propagating wave 4471
10.218 Representation of a wavefront according to Huygens 4472
10.219 Crepuscular rays as a real life example of Huygens-Fresnel secondary source effect 4472
10.221 Failure of Huygens' principle 4475

10.224 Transmitted and reflected power as a function of incident angle 4489
10.225 Ratio $P_R^{\text{vert}}/P_R^{\text{horiz}}$ of power reflected vertically polarized horizontally 4489
10.226 Representation of diffraction fringes 4496
10.227 Photo of a real Fraunhofer diffraction fringe of a rectangular aperture 4497
10.228 Diffraction single slit projection 4499
10.229 Arago spot 4500
10.230 Optical resolution as defined by Lord Rayleigh 4501
10.231 Basic principle of the rectangular aperture network 4503
10.233 Implementation of Young's experience-slit 4510
10.236 Young's interference experiment 4514
10.237 Photos of patterns obtained in a real Young's experience 4515
10.238 Young experiment summary 4515
10.241 Fraunhofer pattern of a square aperture 4520
10.243 Fresnel circular aperture diffraction pattern 2D plot in Maple™ 4.00b 4526
10.244 Fresnel Airy disk 2D plot in Maple™ 4.00b 4527
10.245 Fresnel circular aperture diffraction pattern 3D plot in Maple™ 4.00b 4528
10.247 Overlap of PSFs (point spread function) and resolution 4530
10.252 Spot of Poisson-Arago 4537
10.253 Iceland spar (crystallized calcite) 4538
10.254 Pictorial representation of the light polarization concept with a string 4539
10.255 Illustration of the decomposition of the electric field of a propagating wave . 4539
10.256 Pictorial representation of a circular polarized wave 4540
10.257 Pictorial representation of "natural" light wave 4540
10.258 Representations of different polarizations 4548
10.259 Animation of a polarized wave in Maple™ 4.00b 4548
10.260 Visual polarized wave principle summary as proposed by Wikipedia 4549
10.261 Malus' law plot for various angles 4551
10.262 Simplified example of a polariser / analyser set 4551
10.264 Different configuration of linear polarizers 4553
10.265 Timeline of the most well-known interferometers 4555
10.266 Representation of plane waves 4556
10.267 Example of the influence of various exposure times with a camera 4557
10.268 Setup of Michelson Morley experiment in 11887 at what is now Case Western Reserve University 4562
10.269 Path of light in a simple Michelson interferometer 4563
10.270 Fringe pattern formed by the Michelson interferometer using monochromatic light (sodium D lines) 4564
10.271 Michelson interferometer typical setup 4564
10.272 MIT Michelson interferometer schematic constructive interference 4565
10.274 IUS-16-1-axis LASER gyroscope produced by "Arsenal" company, Ukraine. 4585
10.275 Ring LASER at the Geodetic Observatory Wettzell in Germandy 4586
10.276 Schematic principle of a ruby LASER (solid-state LASER) 4587
10.277 Eye injury hazard for various LASER classes 4594
10.278 LASER beams shot to the Moon from Grasse station in France 4596
10.280 Gold corner cube retroreflector 4597
10.281 Apollo 15's Lunar Ranging Retro Reflector 4598
10.282 Location of retroreflectors on the Moon 4598
10.283 Green LASER pointer beam profile 4600

10.284 Rays reflected within a dielectric cylinder 4605
10.285 Rays in a clad optical fiber 4606

11.1 First photo ever of a single atom emitting light 4618
11.2 The ideal approach of Dalton 4620
11.3 Thomson's Gourmet approach... 4622
11.4 Rutherford experiment 4623
11.5 The Rutherford's planetary approach... 4624
11.6 Simplistic illustration of Bremsstrahlung 4625
11.7 Energy-Wavelength-Frequency electromagnetic values 4627
11.8 Bohr planar planetary model 4631
11.9 Some spectral series of the hydrogen atom 4635
11.10 Three main series of the hydrogen atom 4636
11.11 Visible line spectra of various elements 4637
11.13 Defining hydrogenoid atom configuration 4638
11.14 The eccentric... Sommerfeld approach... 4648
11.15 Decomposition of a very low orbital... 4667
11.16 Decomposition of an upper orbital... 4668
11.18 Siemens MRI scan Healthineers model 4671
11.19 Summary of quantum numbers 4672
11.20 Möller diagram electronic structure 4676
11.21 Decomposition of a low orbit with spin 4677
11.22 Decomposition of a less lower orbit with spin 4678
11.23 Example of Bohr Diagram 4679
11.24 Aufbau Principle 4681
11.25 Periodic (Mendeleev) table 4682
11.26 Periodic (Mendeleev) table structure 4684
11.27 Atomic and ionic radii 4684
11.28 Relative abundance of elements in the observable Universe 4685
11.29 Relative abundance of elements in the human body 4686
11.30 Origin of chemical elements 4687
11.31 Dual nature of light 4689
11.32 Heisenberg's microscope 4704
11.33 Wilson bubble chamber 4709
11.34 Conceptual (artistic) representation of vacuum quantum fluctuations 4722
11.36 The seven levels of nothingness 4743
11.37 Approach to the wave aspect of orbits by de Broglie 4751
11.38 Schematic magnetic domains randomly assigned 4753
11.39 Naive diagram of a transmission electron microscopy 4757
11.40 Electron ptychographic reconstruction of a praseodymium orthoscandate . . 4757
11.41 Atomic model evolution 4765
11.42 Wave function for a ball in a tube of length L 4766
11.43 Pictorial representation of a bound state 4769
11.44 Relationships between most famous quantum physics equations 4772
11.45 Quantum state as a superposition of eigenstates 4780
11.46 Representation of the concept of phase velocity and group velocity 4788
11.47 Spreading plane wave with decreasing δk from top top bottom 4791
11.48 Plot of the Gaussian wavepacket in 2D 4793

11.49 Plot of the Gaussian wavepacket in 3D 4794
11.50 Infinite potential well illustration 4796
11.52 First five (physicists) Hermite polynomials 4817
11.53 Plot of first eigenfunctions and density functions of some energy levels of the 1D harmonic oscillator 4819
11.54 Plot of first eigenfunctions of the 1D harmonic oscillator in Maple™ 4.00b . 4820
11.56 Classical types of potential variations for Tunnel Effect study 4841
11.57 Plot of the transmission T coefficient with MATLAB® 5.0.0.473 for tunneling effect 4850
11.58 Scanning tunneling microscope concept 4851
11.59 An STM image of a carbon nanotube 4852
11.60 Direct measurements of $\Delta\alpha/\alpha$, taken from quasar absorption measurements . 4853
11.61 Omicron Scanning tunneling microscope 4854
11.63 2D plot of the sinc function in Maple™ 4.00b 4861
11.65 Comparison of interference patterns 4864
11.66 Representation of the phases according to the type of path 4868
11.67 Conceptual (artistic) representation of quantum fluctuations 4869
11.68 Some particles classification by spin 4888
11.69 Schematic form of associated orbitals and spin energy levels 4890
11.70 Biased schematic representation of the quantification of total angular momentum in 2D 4892
11.71 Improved schematic representation of the quantification of total angular momentum in 2D 4892
11.72 Biased schematic representation of the quantification of spin in 2D 4894
11.74 Zeeman effect 4896
11.75 Pictorial representation of the spin-orbit interaction 4900
11.77 General overview of Sodium splitting 4903
11.81 Copenhagen survey on interpretations of Quantum Physics 4918
11.82 Explicit free Dirac equation with explanations 4939
11.83 Relationships between most famous quantum physics equations 4959
11.84 Dirac sea 4970
11.85 Example of fine structure in Sodium 4993
11.86 Landscape of muon $g-2$ Measurements of a_μ 5000
11.87 LHCb charge-parity violation by statistical mass distribution analysis 5009
11.88 Effects of fission weapon of 1 Mt as a function of the distance 5020
11.89 Mushroom size for Little Boy Hiroshima and Castle Bravo 5021
11.90 Separation of radiation using a magnetic field 5023
11.91 Typical radiation penetration 5024
11.92 Some nucleus (and isotopes) and their technical notation 5025
11.93 Table of isotopes 5025
11.94 Illustrated radioactivity decay (half-life) 5030
11.95 Idea behind Carbon dating 5037
11.96 Most common types of radiometric dating methods 5040
11.98 Oldest rocks on Earth 5045
11.100 Uranium decay series 5050
11.101 Mass defect concept illustrated for Helium 5057
11.102 Aston curve 5058

11.103 Summary of the type, nuclear equation, representation, and any changes in the mass or atomic numbers for various types of decay . . . 5060
11.104 Nuclear stability valley . . . 5063
11.105 Joint European Torus (JET) tokamak fusion detector . . . 5065
11.106 Nuclear fission by a slow neutron of Uranium 235 with its fission products . . . 5069
11.107 One of Uranium 235 fission chain reaction . . . 5070
11.108 Nuclear power plants in Switzerland at the end of the 120th century . . . 5071
11.110 Types of nuclear reactors . . . 5073
11.111 Hydrogen fusion (proton-proton chain) . . . 5074
11.112 Gamow's idea of Tunnelling effect . . . 5077
11.113 Non-rectangular barrier profile of the nucleus . . . 5078
11.114 City Labs Nanotritium™ battery . . . 5086
11.116 Compton scattering . . . 5099
11.117 Why the sky is blue . . . 5105
11.118 Rayleigh scattering and Moon color during lunar eclipse . . . 5106
11.119 Experiment for measuring the photoelectric effect . . . 5107
11.120 Principle of the photoelectric effect on the Bohr model of the atom . . . 5109
11.121 Rutherford scattering . . . 5111
11.122 Representation of the Rutherford diffusion of alpha nucleus . . . 5113
11.123 Representation of the solid diffusion angle . . . 5114
11.125 Summary of most well known type of scattering . . . 5119
11.126 General Electrics CT Scan Revolution model . . . 5123
11.127 Siemens SPECT scan Symbia model . . . 5123
11.128 Photoelectric effect with gamma ray (naive Bohr model perspective) . . . 5124
11.129 Compton scattering with gamma ray (naive Bohr model perspective) . . . 5125
11.130 Pair creation with gamma ray (naive Bohr model perspective) . . . 5125
11.131 Electron-Positron pair production in a bubble chamber . . . 5128
11.138 Radiation equivalent dose to patients from common imaging examinations . . . 5137
11.139 Radiation equivalent dose chart . . . 5138
11.140 Schematic representation of the semi-empirical formula of von Weizsäcker . . . 5150
11.143 Liquid drop model of Uranium 235 nuclear fission . . . 5155
11.144 Pressurized water reactor . . . 5156
11.145 Typical atom size . . . 5156
11.147 Example of pseudo Feynman-diagram . . . 5160
11.148 Artistic representation of quantum fluctuation in an interacting field . . . 5184
11.149 Global and local gauge illustration concept . . . 5191
11.150 Interactions discoveries . . . 5194
11.151 Phenomenological table of the 4 fundamental interactions . . . 5197
11.152 Feynman diagrams and interaction distance of the fundamental forces . . . 5199
11.153 Force carriers subparticles . . . 5200
11.154 Subparticles Standard Model . . . 5201
11.156 List of common baryons, hadrons, leptons, mesons and bosons . . . 5203
11.157 Particles mass versus Lifetime decay length . . . 5204
11.158 Timeline of particles discovery . . . 5204
11.159 Machine to test fruit quality with the help of nuclear magnetic resonance . . . 5211

12.1 Drake equation . . . 5227
12.2 A small sample of potentially habitable planets according to Kepler's date . . . 5228

12.3 Diagram depicting the habitable zone boundaries around stars, and how the boundaries are affected by star type . . . 5228
12.4 Sun expanding habitable zone . . . 5230
12.5 Representation of surfaces conservation . . . 5236
12.6 Schema for the study of eccentric anomaly angle . . . 5238
12.7 Schema for the study of true-eccentric anomaly angle relation . . . 5240
12.8 Some celestial bodies of our Solar System . . . 5243
12.9 Realistic schema of the Solar System . . . 5243
12.10 Side by side planets and dwarf planets . . . 5244
12.11 Summary of Kepler's laws illustrated . . . 5244
12.12 Planets spin and angle relatively to the ecliptic plane . . . 5245
12.13 Orbit precession example . . . 5245
12.14 Schematic view of planets falling behind the Sun in its movement . . . 5246
12.15 Computer rendered view of planets falling behind the Sun in its movement . . . 5246
12.16 Constellations throughout ages . . . 5247
12.17 Time lapse image of Barnard's star . . . 5248
12.18 The astrolabe's "rete" . . . 5248
12.19 Orion stellarium representation . . . 5250
12.20 Orion short pose photography . . . 5250
12.21 Orion long pose photography . . . 5250
12.22 Real approximative distances of certain stars in the Orion's constellation . . . 5250
12.23 Gravitational field lines and equipotentials . . . 5259
12.24 Up and Down on a Spherically Symmetrical Planet . . . 5259
12.26 Moving barycentre of the Solar System . . . 5266
12.28 Real-life free fall experiment in air and vacuum between a feather vs a lead ball 5269
12.29 Naive schema of the initial Cavendish experiment . . . 5270
12.30 Original schematic Cavendish experiment . . . 5271
12.31 Original schematic Cavendish experiment . . . 5272
12.32 Big G experiment at BIPM . . . 5275
12.33 Gravitation Torsion balance by Leybold . . . 5277
12.34 Illustrated difference between a sphere, ellipsoid en geoid . . . 5279
12.35 Gravitational field across Earth's surface . . . 5279
12.37 Internal/External gravitational acceleration profile of a mass body . . . 5282
12.38 Orbital speed characteristic curve . . . 5283
12.40 Qualitative picture showing the direction of the gravitational field of a flat Earth 5292
12.41 Exact vs Approximate solution of disc shaped Earth gravitational field . . . 5293
12.42 Newtonian expected (with naive central force) vs Observed rotation curve of the M 33 galaxy . . . 5293
12.43 Different galaxies rotation curves . . . 5294
12.44 All asteroids and comets visited by spacecrafts as of November 12010 . . . 5298
12.45 Earth with its atmosphere and oceans . . . 5310
12.46 Earth shape without its atmosphere and oceans . . . 5310
12.48 Evolution of asteroids shape in function of the radius . . . 5311
12.49 Earth's ellipsoid of reference vs the perfect sphere . . . 5312
12.50 Antarctica topography . . . 5314
12.51 Steady-state surface temperatures and water lifetimes obtained by numerical computations . . . 5332
12.52 Earth's atmosphere side photo . . . 5333

12.53 The eight "layers" (...) of the atmosphere . . . 5334
12.54 Configuration for the study of the Roche limit . . . 5337
12.57 Configuration for the average distance between to planets . . . 5356
12.58 Solstices and Equinoxes depending on the latitude . . . 5362
12.59 Effects of precession on the seasons using the Northern Hemisphere terms . . . 5370
12.61 Representation of the rotation of the Earth on its orbit with its major phases . 5372
12.62 Equation of Sunrise . . . 5379
12.63 Trigonometric parallax principle . . . 5384
12.64 Hubble space telescope Proxima Centauri tracking . . . 5385
12.67 Parallax motion seen by Gaia, exaggerated 10,000 times . . . 5388
12.68 Earth unstable orbit . . . 5389
12.69 Synodic vs Sidereal day . . . 5390
12.70 Deviation of day length from SI-based day (86,400s) 11962–12022: daily, moving 365-day average and cumulative . . . 5391
12.71 Basic scheme for determining the synodical period . . . 5392
12.72 Retrograde motion principle . . . 5395
12.73 Apparent retrograde motion of Mars in 12003 as seen from Earth . . . 5395
12.74 Basic scheme for the study of planet's retrograde motion . . . 5396
12.77 Real sequence of exposures showing Mars and Saturn retrograde motion . . . 5401
12.78 Simulation of retrograde motion of Mars and Earth with Maple™ 4.00b . . . 5402
12.79 Analemma and solar terms viewed from Hong Kong . . . 5404
12.80 Analemma as viewed from Antarctica . . . 5404
12.81 Analemma viewed from Cairo (Egypt) . . . 5405
12.83 Angles $\hat{R}$, $\hat{M}$ and $\hat{E}$ at a specific time point. The affinity factor between the elliptical orbit and the circle going through the perihelion and the aphelion is given by b/a. As the angular velocity of the "real Earth" and the "mean Earth" are both constant, the ratio between the two hatched areas is the same as between the areas of the circle and the ellipse, viz., $1:b/a$. . . . 5408
12.84 Angular velocity of Earth as function of orbits' angle . . . 5419
12.85 The angle $\hat{P}$ and the inclination of the axis of the earth at the winter solstice (not to scale) . . . 5420
12.88 Equation of time on Earth . . . 5430
12.89 Various declination calculation methods compared . . . 5432
12.90 Earth analema . . . 5434
12.91 Equation of time and Analemme correspondence . . . 5435
12.92 Juxtaposition of theoretical analemma with an observed one . . . 5437
12.93 Azimuth and Altitude depicted . . . 5438
12.94 Most common type of sundials . . . 5439
12.95 Representing the Earth as a giant sundial . . . 5441
12.96 Equatorial sundial at the North pole . . . 5441
12.97 Equatorial sundials at different latitudes . . . 5442
12.98 Equatorial sundial for the southern hemisphere . . . 5443
12.99 Projection of the equatorial dial to form the ellipse of the vertical dial . . . 5443
12.100 Side view of gnomon of vertical sundial . . . 5444
12.101 Hour angles for vertical sundial in Melbourne . . . 5445
12.102 Hour angles for vertical sundial . . . 5446
12.103 Horizontal sundial illustrated on the globe . . . 5446
12.104 Hour angles for horizontal sundial in Melbourne . . . 5448

12.105 Three types of sundials in one piece 5448
12.113 Illustration of the concept of local meridian 5462
12.115 Right ascension and Declination depicted 5466
12.116 Representation of Equatorial and Azimuthal coordinates 5472
12.117 The orthogonal projection of the solar path for Abu Dhabi 5479
12.118 Apparent path of the Sun (photo: Regina Valkenborgh) 5480
12.119 Multiple Solar Analemma With Pinhole Photography 5481
12.120 The Moon when behind the Earth and aligned with the Sun 5486
12.121 The Moon when between the Sun and the Earth 5487
12.122 Total lunar eclipse 5488
12.123 Total solar eclipse 5488
12.124 Size variation of Sun and Moon because of elliptical orbits at perihelia vs aphelia 5489
12.125 Typical naive configurations of solar and moon eclipses 5490
12.127 Some saros cycles 5494
12.128 Representation of the five Lagrangian point in the Sun-Earth system 5498
12.129 Isoclines of the two-body system 5504
12.130 Projected isoclines of the two body system on a plane 5505
12.132 Lagrange plot and isoclines in 3D with MATLAB® 2013a 5507
12.133 Lagrange plot and isoclines in 2D with MATLAB® 2013a 5508
12.141 Determining the Hill sphere radius 5526
12.142 Doppler-Effect radial velocity method 5529
12.143 High-redshift galaxy candidates in the Hubble ultra deep field 5532
12.144 James Webb Telescope galaxy cluster SMACS 0723 deep field 5533
12.145 Schematic idea behind the apparent speed 5536
12.146 Hertzsprung-Russel Diagram example 5540
12.147 Sun path on Hertzsprung-Russel 5540
12.148 Peak color of a black body at a given temperature 5541
12.149 Typical stars temperature and related spectrum with absorption lines 5542
12.150 Photo of Sun's surface (sunspot AR4455) with a Soleye 3000 5543
12.151 Region of the sky before and after the 11987 supernova in visible light 5545
12.152 Evolution of temperature, radius, luminosity of our Sun as function of time . 5545
12.153 Planetary Nebula gallery 5546
12.154 Comparison of various planets with various stars 5547
12.155 Solar System Proportions 5548
12.156 Biggest Black Hole vs Biggest known star 5548
12.157 Protoplanet formation PDS 70b, PDS 70c 5551
12.158 Sun in comparison of various dwarf stars and Jupiter 5554
12.159 EBLM J0555-57Ab smallest known star 5555
12.160 Diagram of the proton-proton chain reaction in typical Sun size stars 5560
12.161 Sun Corona zoom 5562
12.162 Global Sun overview 5564
12.163 Kelvin-Helmholtz instability on the Sun 5566
12.164 Spectral power distribution of sunlight 5567
12.165 Layer's view of our Sun 5569
12.166 Sun temperature in function of wavelength 5570
12.167 Sun temperature in function of time 5571
12.168 Computer simulation of Earth's temperature evolution 5571

12.169 Sun appearance in a photograph with and without a solar filter 5577
12.170 Apparent luminosity 5578
12.171 Most brightest stars at night in early 121st century 5579
12.172 Most brightest galaxies at night in early 121st century 5579
12.173 Approximate sky if important objects had similar brightness 5580
12.174 Magnitude and brightness Ptolemy-Pogson relations 5581
12.175 Moon phases 5583
12.176 Moon and ecliptic angle 5585
12.177 Full Moon at opposite librations 5585
12.178 Sirius apparent vs absolute magnitudes 5587
12.179 Impact of integration time on details and background noise 5593
12.183 Hubble Space Telescope Cepheid measurement 5603
12.184 Cepheid absolute magnitude-period plot 5604
12.185 Different types of Neutron Stars 5606
12.186 Neutron star slice 5611
12.187 Neutron star inner crust 5612
12.188 Pulsar timing array illustration 5617
12.189 Star comparisons from Black Dwarf to hyper giants 5619
12.190 Apparent stars speed anomaly in galaxies rotations 5620
12.191 Gallery of some galaxies 5622
12.192 Size comparisons of famous galaxies 5623
12.193 NGC 2207 is a pair of colliding spiral galaxies 5624
12.194 Apparent stars speed anomaly in galaxies rotations 5625
12.195 Michelson and Morley's original interferometric setup 5627
12.196 Configuration for the study of relativistic effects 5633
12.197 Lorentz factor length contraction function trend 5655
12.198 Length contraction principle for rectilinear motion 5655
12.199 Lorentz factor time dilatation function trend 5660
12.201 Configuration for the study of the apparent relativistic mass variation seen from R_0 5671
12.202 Configuration for the study of the apparent relativistic mass variation seen from R 5672
12.203 Lorentz factor mass dilatation function trend 5674
12.205 Ratio of relativistic and nonrelativistic Bohr radii 5676
12.206 Spectral reflectance curves for aluminum, silver and gold metal mirrors . . . 5677
12.207 Gold Bohr naive model configuration 5678
12.208 Chemistry of Gold 5679
12.209 Relationships between most famous quantum physics equations 5685
12.210 Configuration for the study of transformations of electric and magnetic fields 5694
12.211 Configuration for the study of electrodynamic transformations 5697
12.212 Worldline 5704
12.213 Idea of light-cone 5708
12.214 Past and future light cones of an event of a worldline 5709
12.215 Universe Line Principle with its associated cone 5711
12.216 Types of universe lines 5712
12.217 Light Cone in 2d Space plus a Time Dimension 5714
12.218 Summary of the postulates of General Relativity 5727
12.219 Some mainstream metrics used in General Relativity 5770

12.220 2D naive representation of space curve near Earth 5771
12.221 3D naive representation of curved space near Earth 5772
12.224 Swiss commemorative General Relativity coin 5782
12.225 Schwarzschild precession illustrated 5811
12.226 Einstein Ring LRG 3 757 5819
12.228 Round-trip time of a signal as a function of the position of Mars 5824
12.229 Delayed Radio signals from the Viking lander on Mars 5827
12.230 2D naive representation of space curve near some celestial objects 5838
12.231 Orbits of stars within the central 1.0×1.0 arcseconds of our Galaxy 5839
12.232 Black Hole ray-traced 5841
12.233 Classical effective gravitational potential profile 5845
12.234 Effective potential profile accord to Newtonian and General Relativity mechanics 5847
12.237 Relativistic effective potentials for various values of l/M with $M = 1$ 5849
12.238 Black Hole innermost circular stable orbit illustration 5850
12.239 First direct reversed engineering capture of a Black Hole 5851
12.240 Most recent reversed composite image of the M87* Black Hole 5852
12.241 Photosphere and event horizon with correct proportions 5853
12.242 Black Hole summary 5854
12.243 Schwarzschild black hole representation based on Flamm's paraboloid 5859
12.244 Artist's impression of a white hole 5860
12.245 LIGO measurement of the gravitational waves at the Hanford and Livingston detectors 5861
12.246 Hanford LIGO detector Washington state 5863
12.247 Pseudo-3D (wrong) common visualization of gravitational wave of a binary system 5863
12.248 Gravitational cosmic background "radiation" 5866
12.249 Summary of the most well-known experimental tests of General Relativity . . 5868
12.250 Copenhagen survey on how the "Big Bang" should be understood 5881
12.251 Automated Plate Measurement (APM) Galaxy Survey 5882
12.252 Illustration of the validity limit of the model 5884
12.253 Illustration of the inflated balloon Universe model 5886
12.254 Evolution of Hubble parameter estimations 5889
12.255 An artistic concept of the Hubble Horizon 5891
12.256 Illustration of the different types of universes 5899
12.257 Illustration of what would give observations depending on the curvature type 5901
12.258 Comparison of CMB results from COBE, WMAP and Planck 5901
12.259 Universe fluctuation density scales 5902
12.260 Evolution of the scale factor for a zero curvature space 5906
12.262 Various estimations of our Universe age 5918
12.264 Big Bang and Big Crunch plot side by side 5926
12.265 Big Bounce artistic illustration 5929
12.272 Summary of Newtonian Universe models 5946
12.273 Artist's logarithmic scale conception of the observable Universe 5947
12.274 Hubble Ultra-Deep field 5952
12.275 Universe scales 5953
12.276 Simplified illustration of the observable Universe 5954
12.277 Simplified illustration of the neighboring superclusters 5955

12.278 Simplified illustration of the Virgo supercluster 5956
12.279 Simplified illustration of our Local Group 5957
12.280 Simplified illustration of satellite galaxies 5958
12.281 Simplified illustration of the Milky Way Galaxy 5959
12.282 Simplified illustration of the Orion Arm 5960
12.283 Simplified illustration of the solar neighbourhood 5961
12.284 Simplified illustration of the nearest stars 5962
12.285 Observable Universe map 5965
12.286 Cosmic microwave background according to WMAP 5-year results 5966
12.287 CMB experiment looking a maze of contaminants (credit: Josquin Errard) . . 5967
12.288 FIRAS CMB Black Body radiation spectrum 5968
12.289 2D illustration of (speculated) Universe past history following 121st century hypothesis 5969
12.291 Big Bang recombination 5978
12.292 Ionization factor and mean free path in function of temperature 5983
12.293 ΛCDM theory experimental evidence 5987
12.294 DESI galaxies survey (credit: DESI collaboration and KPNO/NOIRLab/NSF/AURA/R. Proctor) 5993
12.295 Matter density Ω_m statistical distribution of different experiments 6011
12.297 Cosmological models summary 6048
12.298 Uchuu 12021 Universe simulation 6049
12.299 Idea of Universe inside a (Schwarzschild) Black Hole 6059
12.300 Copenhagen survey on matter fate in a Black Hole 6060
12.301 Okun wheel 6073
12.302 Mass-period diagram of confirmed discovered exoplanets in 12015 6078
12.303 Various universes popping-up from the Cosmos false vacuum 6096
12.304 Illustration of Dp-branes 6103
12.305 Universe surface generated by respectively an opened/closed string 6104
12.306 Elementary illustrated refresh of a parametric curve 6105
12.307 Configuration for the study of an elementary surface element 6107
12.308 Parametrization of a space-time surface 6110

13.3 Wave functions and probability density of a one-dimensional harmonic oscillator 6129
13.4 Probability density function of a one-dimensional harmonic oscillator for large n 6130
13.5 Legendre spherical harmonics plot 6142
13.6 Plot of the spherical harmonics $Y_{1,2}$ 6169
13.7 $1s$ density map 6169
13.8 $2p$ orbitals (spherical harmonics) 6169
13.9 $2p$ density map 6170
13.10 $3d$ orbitals (spherical harmonics) 6170
13.11 $3d$ density map 6170
13.12 $3d$ orbitals (spherical harmonics) 6171
13.13 $4f$ density map 6171
13.15 Plot of a few radial functions $R_{n,l}(r)$ 6180
13.16 Probability of finding the electron as a function of distance from the nucleus for the hydrogen atom in the $1s$ state 6186

13.17 Hydrogen atoms imaged at four energy levels 6186
13.18 Plot of the effective potential energy in Maple™ 4.00b for various l and Z . . 6187
13.19 Plot of the famous effective potential energy in Maple™ 4.00b for $l = 1$ and $Z = 1$ 6188
13.20 Plot of the effective potential energy for $l = 1$ and various Z 6189
13.21 Force Atomic Microscope image of chemical bonds of a nanographene molecule C60 6190
13.22 Example ofa molecule where almost everything is moving 6192
13.23 Simplified study of the dihydrogen cation H_2^+ 6202
13.24 Energy levels of dihydrogen cation H_2 6204
13.25 Construction for the study of inertia momentum of a diatomic molecule . . . 6207
13.26 Harmonic oscillator with two masses 6211
13.27 Various energies and corresponding radiation involved in some chemical/molecular phenomenon 6215
13.28 Most common classical and exotic states of matter 6229
13.29 Homogeneous mixture, heterogeneous mixture, compound or element 6230
13.30 Acute Toxicity of some ingredients/molecules 6232
13.31 Acute Toxicity of some ingredients/molecules 6233
13.32 Steps in Analytical Chemistry 6234
13.34 The joy of mixtures... 6236
13.35 The joy of mixtures for dyers and military 6237
13.36 Example of polymorph 6240
13.37 National Fire Protection Agency (NFPA) hazard diamond 6241
13.38 Some typical chemical reactions 6243
13.39 Some typical contrails 6244
13.40 Limiting reactant illustration 6246
13.42 Kirchhoff's enthalpy law illustration 6263

14.1 Floating point (real numbers) in simple precision 6277
14.2 Floating point (real numbers) in half precision 6278
14.3 Big $\mathcal{O}$ notation complexities plot 6287
14.4 The illustrated principle of Archimedes algorithm 6298
14.5 Comparison of logarithmic formulations using Stirling formula 6302
14.6 Synoptic summary and relationships between various regression models . . . 6320
14.7 Graphical representation of respectively SST, SSE, SSR 6331
14.8 Some values of the linear correlation coefficient 6333
14.9 Graphical representation of an univariate regression with the ANOVA approach in Microsoft® Excel® 14.0.7177 6342
14.10 Lack of fit and pure error 6346
14.11 Graphical representation of the idea behind the Gaussian linear regression . . 6349
14.12 Examples of "plot" of residuals 6350
14.13 Print screen of a confidence interval obtained with Minitab 15 6363
14.14 Print screen of a prediction interval (PI) obtained with Minitab 15 6365
14.15 Confidence vs Prediction interval 6365
14.16 Outliers vs Influential points 6392
14.17 Importance of an influential point on regression equation 6395
14.18 Some famous techniques of cross-over validation 6429
14.19 K-fold cross-over validation 6430

14.23 Confusion matrix Minitab 15.0 output example 6461
14.24 Precision and Recall 6483
14.25 Four-fold plot with R 3.4.2 6502
14.26 Odds Ratio and Risk Ratio range values subjective qualifications 6506
14.27 Risk Ratio (relative risk), Odds Ratio or Hazard Ratio rules 6508
14.32 Corresponding values of the construction of the confusion matrix 6515
14.33 Confusion matrix with Minitab 16.1.1 6516
14.34 ROC curve with Minitab 16.1.1 6516
14.35 Values of the confusion matrix for various cutoff values 6517
14.36 ROC curves 6518
14.42 Overview of all constructed columns for the lift curve 6528
14.45 Underlying idea of the Normal GLM 6532
14.46 Underlying idea of the Poisson GLM 6534
14.47 Underlying idea of the binomial GLM 6538
14.48 L^2-norm (influence function) 6559
14.49 Median (influence function) 6562
14.50 Graphical representation of a few common dispersion M-estimators 6564
14.51 Implementation of Tukey's biweight and Huber M-estimator in R 6566
14.52 A few 1st order Bézier curves 6570
14.53 Idea behind 2nd order Bézier curves 6570
14.54 A few 2nd order Bézier curves 6571
14.55 Examples of Bézier curves of order 3 6573
14.56 A few 3rd order Bézier curves 6574
14.57 Screenshot of a Bézier spline with a drawing program 6576
14.58 Example of a Bézier spline with its underlying construction 6577
14.59 Example of constructing a circle with Bézier curves 6578
14.61 A visual summary of naive splines 6581
14.62 Summary of different ways of calculating Bézier curves 6582
14.63 Illustration of the interpolation concept with the collocation technique 6591
14.64 Illustration of concept of Lagrange interpolation 6594
14.65 Local approximation of a curve by a straight line segment 6604
14.66 Regula falsi method scheme 6606
14.67 Bisection scheme 6609
14.68 Newton's method in the plane 6612
14.69 Non-convergence example of Newton's method 6614
14.70 Maple™ 4.00b application of Newton's method 6617
14.71 Screenshot of Microsoft® Excel® 2003 Solver 6619
14.72 Forward/Backward derivatives illustration 6620
14.73 Illustration of an interval under a curve 6622
14.74 Approach of the area under a curve by lower left rectangles method 6623
14.75 Approach of the area under a curve by upper right rectangles method 6624
14.77 Composite trapezoidal rule illustration 6628
14.78 Quadrature Rules (author: Lauren K. Williams) 6631
14.80 Finding solutions graphically with the economic function 6650
14.83 Microsoft® Excel® 2003 Solver options 6663
14.84 Plot of the humpback whale function with minima already visible 6664
14.90 ISO 5807:85 Flowchart of EM algorithm 6698
14.91 EM parameter estimation for complete and incomplete data 6702

14.92 Summary of resampling in different methods . . . 6707
14.93 Principle of the inverse transform method . . . 6712
14.94 Pseudo-random generator in Maple™ 4.00b . . . 6713
14.99 Basic principle of the univariate integral calculation with Monte Carlo . . . 6717
14.100 Monte Carlo π estimate . . . 6722
14.102 Source Generative Chemistry and Accelerated DFT (source: Microsoft) . . . 6752
14.103 1D heat equation FDM calculations in Microsoft® Excel® 14.0.7172 . . . 6754
14.104 1D Heat Equation FDM explicit formulas in Microsoft® Excel® 14.0.7172 . . . 6755
14.105 1D Heat Equation FDM plot in Microsoft® Excel® 14.0.7172 . . . 6755
14.106 Space-time FDM for car C_x study . . . 6760
14.107 Space-time FDM for airplane wing profile study . . . 6760
14.108 Space-time FDM for NASA space shuttle launch . . . 6761
14.109 Airplane typical FDM meshing . . . 6762
14.110 Mechanical element FDM meshing . . . 6762
14.111 Generic Yee cell in parallelepiped mesh element . . . 6765
14.112 Yee cell in parallelepiped mesh element for EM wave propagating in vacuum 6766
14.113 Typical simple meshing with multiple Yee cells . . . 6766
14.114 Oversimplified overview of Data Science . . . 6769
14.115 Descriptive, Diagnostic, Predictive and Prospective analytics . . . 6769
14.116 Data Mining/Machine Learning non-exhaustive Mind Map techniques and tools 6771
14.117 Data Mining life cycle as seen by SAS™ . . . 6790
14.118 Data Mining/Machine Learning 4V's . . . 6791
14.119 Data science roles . . . 6792
14.120 Comparing Reporting and Business Intelligence with Data Science . . . 6793
14.121 Graphic Continuum . . . 6794
14.123 Underfitting and overfitting errors . . . 6796
14.124 Holdout Validation in Machine learning . . . 6801
14.125 Freedman paradox illustration . . . 6802
14.126 Parameters vs Hyperparameters . . . 6803
14.127 Three most common ensemble learning methods . . . 6807
14.128 ZeroR Data Mining (association rule) example . . . 6809
14.129 OneR Data Mining (association rule) example . . . 6811
14.130 Non-exhaustive orgchart of clustering techniques . . . 6823
14.131 Supervised vs Unsupervised learning idea . . . 6824
14.132 Data Mining as seen by SAS™ . . . 6825
14.133 Classification of a new observation with $k-nn$ . . . 6834
14.144 CART Final result in the traditional tree form . . . 6844
14.147 Main vocabulary of decision trees . . . 6846
14.148 Random forest illustration . . . 6853
14.149 k-means iterative algorithm . . . 6855
14.159 Support Vector Machine (SVM) idea . . . 6868
14.160 Expectation-Maximization iterative algorithm . . . 6878
14.161 Variance-Covariance matrix typical structures . . . 6883
14.164 Type of links in Hierarchical Clustering . . . 6898
14.167 Dendrograms plot in Minitab 15.1.1.0 . . . 6901
14.168 Dendrograms plot in MATLAB® 2013a . . . 6901
14.169 Dendrograms plot in R 3.0.2 . . . 6902
14.170 Dendrograms heatmap plot in R 3.0.2 . . . 6902

14.171 Dendrograms heatmap microarray plot in R 3.0.2 6903
14.172 Dendrograms TSA microarray plot in R 3.0.2 6904
14.173 HAC on a factor map 6907
14.174 Underlying idea of linear discriminant analysis 6909
14.175 Expected configuration for linear Discriminant analysis 6910
14.176 Typical case where LDA fail and PCA don't 6915
14.177 T-SNE perplexity sensitivity 6940
14.179 Illustrate visualization of the iterative concept of Gradient Boosting 6949
14.180 Illustrated Gradient Boosting with Trees 6950
14.181 Google Dream Machine test 6951
14.182 Neural Networks complete chart 6952
14.183 Example of one-layer formal neuron with the input vector and scalar output . 6954
14.184 Simplified representation of the vocabulary of human neurons 6955
14.185 Vector written example of one-layer formal neuron with the input vector and scalar output 6956
14.186 Most used transfer functions for neural networks in the 120th century 6959
14.187 Multi-class softmax neural network 6960
14.188 Example of one-layer formal neural network with the input vector and output vector 6964
14.189 Vector written example of one-layer formal neural network with the input vector and output vector 6965
14.190 Principle of construction of a Multi-Layer Perceptron 6965
14.191 Explicit construction of a Multi-layer Perceptron 6966
14.192 Neural Information Processing 6968
14.207 Neural network in MATLAB® 2013a 6975
14.211 Three common type of recursive neural networks 6986
14.212 Recurrent neural network 6987
14.213 LSTM (long short term memory) building block 6988
14.214 Typical LSTM overall structure 6989
14.221 A filter convolving across an image 6999
14.222 Convolution of a volume consisting of three matrices 7000
14.223 Convolution with stride 2 7002
14.227 2×2 Average vs Max pooling of stride 2 7005
14.228 Various type of CNN internal structures 7010
14.229 General idea of a GAN 7011
14.230 General idea of a GAN Discriminator training 7014
14.231 General idea of a GAN Generator training 7015
14.234 Example of David E. Goldberg's wheel with five individuals with their respective relative fitness 7029
14.235 Illustrative example of crossover 7030
14.236 Illustrative example of mutation 7031
14.237 Rastriging function in MATLAB® 2013a 7034
14.238 Pseudo-mountains generation from a random fractal (probabilistic fractal) . . 7039
14.240 Cantor set with Maple™ 4.00b 7050
14.244 First iteration of Cantor's Fractal 7059
14.245 Cantor attractor after 6 iterations 7060
14.255 Sierpinski triangle attractor 7065
14.278 Von Koch snowflake 7075

14.279 Coastline paradox . . . 7077
14.310 Summary of most well knows basic fractals . . . 7090
14.311 Bichromic Mandelbrot Fractal . . . 7094
14.312 Mandelbrot Fractal with Maple™ 4.00b . . . 7095
14.313 Self-Similarity of Mandelbrot Fractal . . . 7096
14.314 Mandelbrot-Logistic bifurcation correspondence . . . 7097
14.315 Makingofa Julia Set . . . 7098
14.316 Julia Fractal with Maple™ 4.00b . . . 7099
14.317 Illustration of "fatherhood" ... of the Mandelbrot Set . . . 7100
14.318 Newton Set fractal with Maple™ 4.00b . . . 7103
14.319 Venn diagram of the consensus theorem . . . 7108
14.320 Logic MIL Gates . . . 7112
14.321 Summary of basic Boole algebra . . . 7114
14.322 Construction of all gates from NAND gates . . . 7115
14.323 Karnaugh Map Minimizer 0.4 . . . 7119
14.324 Half adder logic diagram . . . 7127
14.325 Full adder logic diagram . . . 7128
14.326 Half subtracter logic diagram . . . 7131
14.327 Full subtracter logic diagram . . . 7132
14.328 1-bit magnitude comparator . . . 7133
14.329 1-bit magnitude comparator . . . 7133
14.330 4-bits magnitude comparator . . . 7134
14.331 if-else statement flowchat . . . 7135
14.332 If-else statement in hardware based on boolean gates . . . 7135
14.333 Centered linear fuzzy function . . . 7148
14.334 Centered linear fuzzy function . . . 7149
14.335 ECC RAM vs Non-ECC RAM . . . 7154
14.336 Non-exhaustive orgchart of correcting codes . . . 7155
14.337 Basic Checksum principle . . . 7156
14.338 Luhn algorithm for checksum principle . . . 7157
14.339 EAN-13 bar-code example . . . 7159
14.340 Swiss payment slip . . . 7161
14.341 Swiss payment slip check digit for Reference Number . . . 7161
14.343 Principle of a Linear code . . . 7173
14.344 Principle of the von Neumann machine . . . 7183
14.345 Assembly language "Hello World" demo . . . 7185
14.346 Turing Machine principle . . . 7186
14.347 A simple finite automata . . . 7187
14.349 Second example of a finite automata . . . 7188
14.350 Third example of a finite automata . . . 7188
14.352 Turing Machine . . . 7189
14.353 Chomsky hierarchy . . . 7196
14.354 'Hello World' in Morse code . . . 7207
14.356 Sardinas and Patterson example graph . . . 7215
14.357 Principle of F.E.K. in Microsoft Windows XP . . . 7220
14.358 Encryption of 1101 using the Feistel method . . . 7226
14.359 Feistel's scheme . . . 7228
14.360 Feistel's scheme confusion function . . . 7229

14.361 Principle of public key encryption . . . 7231
14.362 Principle of RSA public key encryption . . . 7246
14.363 Illustrated result of MD5 algorithm . . . 7249
14.364 Illustrated MD5 algorithm flow . . . 7251
14.365 Illustrated MD5 download hash from CISCO . . . 7252
14.366 Typical other use of AES and RSA combined together . . . 7255
14.367 Principle of digital signatures . . . 7257
14.368 Principle of certificate authority . . . 7257
14.370 Thought experiment for polarization measurement . . . 7259
14.371 Quantum key distribution . . . 7260
14.372 Schrödinger cat experiment . . . 7266
14.376 Bloch-shpere (two dimensional one) . . . 7276
14.377 NOT Quantum gate . . . 7289
14.378 Hadamard Quantum gate . . . 7291
14.379 NAND Quantum Gate . . . 7291

15.1 Earth overshoot day evolution . . . 7297
15.2 Cumulative probability of survival as a function of age . . . 7303
15.3 Life expectancy evolution . . . 7304
15.4 Life expectancy evolution from birth in France since 11740 . . . 7305
15.5 Life expectancy by age in England and Wales (11700-12013) . . . 7305
15.6 Fertility rates in the early 121st century . . . 7308
15.7 Sex ratio trend in the early 121st century . . . 7309
15.8 Evolution of the world population since 11965 in Maple™ 4.00b . . . 7311
15.9 Logistic model vs Exponential model . . . 7316
15.10 Plot of the logistic equation with $x_0 = 0.1$ and $K = 1$ . . . 7320
15.11 Plot of the logistic equation with $x_0 = 0.1$ and $k = 2$ . . . 7320
15.12 Plot of the logistic equation with $x_0 = 0.1$ and $k = 3$ . . . 7321
15.13 Plot of the logistic equation with $x_0 = 0.1$ and $K = 3.5$ . . . 7321
15.14 Plot of the logistic equation with $x_0 = 0.1$ and $K = 4.5$ . . . 7322
15.17 Comparison plots of logistic equations for small k . . . 7322
15.20 Comparison plots of logistic equations for medium k . . . 7323
15.23 Comparison plots of logistic equations for large k . . . 7325
15.26 Comparison plots of logistic equations for one large k and small δk . . . 7325
15.27 Feigenbaum diagram in Maple™ 4.00b . . . 7328
15.28 Bifurcations in the Feigenbaum diagram . . . 7329
15.32 Exponential epidemic model . . . 7336
15.33 SI epidemic model . . . 7337
15.34 SIS epidemic model . . . 7339
15.35 SIR epidemic model . . . 7341
15.36 Markov chain of deadly disease . . . 7344
15.37 Theoretical fluctuation of population . . . 7352
15.39 Representation in phase space of the prey/predator cycle . . . 7353
15.40 Rotation direction offer/demand phase space . . . 7354
15.41 Lotka-Volterra offer/demand model periodic plot in Maple™ 4.00b . . . 7358
15.42 Lotka-Volterra offer/demand phase space model plot in Maple™ 4.00b . . . 7358
15.43 Genotype vs Phenotype . . . 7362
15.44 Monohybrid cross . . . 7363

15.45 Plot of the Hardy-Weinberg equation 7365
15.46 Punnett square for three-allele case 7366
15.47 Evidence of the smooth evolution of hominide crane 7367
15.48 Species-level relationships and tempo of diversification across mammals (philogenetic tree) 7368
15.49 Human linear evolution perspective 7369
15.50 Mendelian Inheritence Chart 7372
15.51 Evolution and Original of Life experimental evidence 7374
15.52 Elementary classification of decision techniques 7376
15.53 Common types of games in Game Theory 7380
15.54 Some famous logical fallacies 7384
15.55 Myers-Briggs personality type classification 7384
15.56 Sequential game in the form of a decision tree (or Kuhn's tree) 7399
15.57 Event tree analysis with fixed probabilities in Microsoft Office Visio 7401
15.58 Decision tree without information in TreeAge 7402
15.59 Decision tree without information and expected mean in TreeAge 7403
15.60 Decision tree with perfect information in TreeAge 7403
15.61 Decision tree with imperfect information in TreeAge 7406
15.62 Complete investment Tree (unreduced) 7408
15.64 Optimistic investment reduced tree 7409
15.65 Pessimistic investment reduced tree 7409
15.66 Sequential game in the form of a decision tree (or Kuhn's tree) 7410
15.67 Extensive form of the stone-scissors-paper game (Kuhn's tree) 7413
15.68 Correspondence of the inequalities of the optimization problem 7427
15.69 Pascal's wager 121st century religions and beliefs outcomes 7435
15.70 Medical state diagram 7456
15.71 Financial state diagram 7458
15.76 AHP starting decision tree example 7471
15.77 AHP decision tree example with weights 7472
15.78 Simplified schematic summary to develop a pricing policy 7490
15.80 Swiss National Banks reserve 7518
15.81 Evolution of monetary base in Switzerland 7519
15.82 Relative evolution of monetary base in Switzerland 7520
15.83 Countries by GDP (PPP) Per Capita in 12015 7528
15.84 Asymptotic utility 7531
15.85 Corresponding decreasing marginal utility 7531
15.86 Example of iso-utility (indifference) curve 7533
15.87 Counter example of iso-utility (indifference) curve 7534
15.88 Iso-utility intersection impossibility 7534
15.89 Utility super-additivity visual representation 7535
15.90 Visual representation of substitutable goods utility 7536
15.91 Cycle of offer and demand 7540
15.92 Representation in phase space offer/demand cycle 7540
15.93 Rotation direction offer/demand phase space 7541
15.94 Lotka-Volterra offer/demand model periodic plot in Maple™ 4.00b 7545
15.95 Lotka-Volterra offer/demand phase space model plot in Maple™ 4.00b . . . 7546
15.96 Rise of the machines in algorithmic trading 7583
15.97 Typical bias ratio of some funds 7595

15.98 Down Jones Industrial Index Monitoring in Bloomberg$^{\text{TM}}$ Terminal 7601
15.99 Comparison of most well known market indices 7603
15.100 Marketable Securities (non-exhaustive) 7605
15.101 Non-Marketable Securities (non-exhaustive) 7606
15.102 Example of an old French common stock share paper 7608
15.103 Parallel graphs of the S&P index and PER for the same period 7611
15.104 Example of bond with fixed rate with fixed number of coupons (at the time of paper...) 7618
15.105 Example of perpetual bond with fixed rate (at the time of paper...) 7618
15.106 Example of fixed rate OCT with coupons on Bloomberg$^{\text{TM}}$ Terminal (at the time of the computers...) 7619
15.107 Typical Bond details on Bloomberg$^{\text{TM}}$ Terminal 7621
15.108 Example of yield curve for zero-coupon USA bonds (T-Bills) in Bloomberg$^{\text{TM}}$ Terminal 7622
15.109 Current value of a bond based on the market rates (yield) 7627
15.110 Schematic principle behind the bond accrued interest 7629
15.111 Current value of a bond based on the market rates (yield) 7632
15.112 Delta and Delta-Gamma Modified duration approach 7634
15.113 Duration as the Maturity of a Zero-Coupon Bond 7637
15.114 Warrant old style paper version 7643
15.115 Future Equity description in Bloomberg$^{\text{TM}}$ Terminal 7649
15.116 Typical spot price and Future/Forward price dynamics 7654
15.117 Correlation matrix for various currencies in Bloomberg$^{\text{TM}}$ Terminal 7658
15.118 Derivatives contracts viewed on various basis 7663
15.119 Forward commitment vs Contingent claims 7664
15.120 Options graphical concept summary 7666
15.121 Typical volatility smile 7670
15.122 Leverage effect for the buyer of a Call (simple Call payoff diagram) 7673
15.123 Leverage effect for the seller of a Call (simple Call payoff diagram) 7674
15.124 Leverage effect for the buyer of a Put (simple Call payoff diagram) 7675
15.125 Leverage effect for the seller of a Put (simple Call payoff diagram) 7675
15.126 Payoff diagram of a straddle strategy 7677
15.127 Yield distribution analysis in the Bloomberg$^{\text{TM}}$ terminal 7693
15.128 Efficient Frontier 7702
15.136 Markowitz manual resolution configuration in Microsoft® Excel® 14.07.7173 7707
15.142 Mean-Variance efficient portfolio 7712
15.143 Multi-Asset moving average strategy VS One index portfolio 7714
15.144 Swiss LPP various diversification returns 7715
15.145 Idea of the linear regression behind the β coefficient 7715
15.146 Linear regression principle of beta (and alpha) in Bloomberg$^{\text{TM}}$ Terminal . . 7716
15.147 Optimal portfolio allocation and Capital Allocation Line 7719
15.149 Linear regression to calculate the betas of the assets returns 7726
15.162 Treynor-Black Portfolio Selection 7737
15.165 Security Market Line 7749
15.166 Approximated parabola to Markowitz efficiency frontier in Microsoft® Excel® 14.0.7173 7752
15.167 Approximated CML on Markowitz efficiency frontier in Microsoft® Excel® 14.0.7173 7753

15.168 Approximated parabola to Markowitz efficiency frontier with CAPM and traditional orientation Microsoft® Excel® 14.0.7173 7754
15.169 Markowitz efficiency frontier with CAPM and CML in MATLAB® 2013a . . 7755
15.170 Security Market Line and Capital Market Line mapping 7756
15.171 Put-Call parity equation illustrated 7763
15.172 Example of 4 standard Brownian motions with Microsoft® Excel® 11.8346 . 7766
15.173 Wiener process (normal random walk) with TikZ 7767
15.174 Example of standard Brownian motion with drift (and decomposition) 7771
15.178 Brownian motion and Brownian bridge 7776
15.179 Explicit formulas in Microsoft® Excel® 11.8346 to build a Brownian bridge . 7778
15.180 Brownian bridge in Microsoft® Excel® 11.8346 7778
15.181 Surface Call price curves with a random Brownian bridge 7779
15.182 Geometric Brownian motion plot with 95% C.I. in Microsoft® Excel® 11.8346 7790
15.185 Comparison of geometric and standard brownian motion 7792
15.186 Greeks summary 7803
15.187 Valuation of a vanilla Call option on 1 underlying of the S&P 500 7819
15.188 Application of Black & Scholes pricing version of Microsoft® Excel® 14.0.6123 7822
15.189 Typical profile of an implied volatility surface 7824
15.190 Typical profile of implied volatility in the Bloomberg terminal 7824
15.191 Black model with MATLAB® 2016a 7846
15.192 Bloomberg™ terminal VIX dashboard 7856
15.193 Sigma intervals for the Normal distribution 7860
15.194 Graphical representation of VaR, VaR Deviation, CVaR, CVaR Deviation, Max Loss, and Max Loss Deviation 7877
15.195 Some countries spread risk 7884
15.196 Bloomberg™ terminal KPI's dashboard 7896
15.197 Some time series without forecasting 7898
15.198 Forecasting evolution 7902
15.199 Forecasting error measures decision flow chart 7906
15.211 Exhibit of easily recognizable time series patterns 7916
15.212 Non-exhaustive orgchart of forecasting techniques 7917
15.215 MA(3) and MA(5) plot with Microsoft® Excel® 11.8346 7920
15.216 Microsoft® Excel® Moving Average choice in the Analysis Toolpack 7921
15.250 Holt & Winter's full plot and comparison with other models in Microsoft® Excel® 14.0.6123 7957
15.251 Classification of ETS$(\cdot,\cdot,\cdot)$ models 7960
15.252 Another simplified typical classification of ETS models 7962
15.253 Classification of ETS models with recursive formulation 7963
15.254 Logistic S-curve characteristics with Microsoft® Excel® 14.0.6123 7966
15.263 Logistic Analytical vs Solver plot with Microsoft® Excel® 14.0.6123 7971
15.265 Microsoft® Excel® 2016 forecast native tool demo 7973
15.266 Autoregressive process of order 1 7976
15.267 Some examples of time series 7978
15.268 Corresponding AFC correlograms of the time series 7979
15.269 Four non-stationary time series 7979
15.273 Unit root process vs Stationary process 7991
15.274 Summary of ETS models vs ARIMA models 7995

15.275 Some time series with forecasting (GARCH, ARFIMA, Holt-Winter, ARIMA, etc.) with R . . . 8010
15.276 Recipes for modelling time series data . . . 8013
15.277 Typical simple process . . . 8023
15.278 Typical type of financial risks . . . 8025
15.279 100+ financial metrics . . . 8028
15.280 Data Envelopment Analysis . . . 8038
15.284 WACC of Pepsi company in Bloomberg$^{\text{TM}}$ Terminal . . . 8046
15.285 CF Valuation Techniques . . . 8058
15.286 P/B for JP Morgan Chase and Wells Fargo Company . . . 8065
15.287 $10y$ P/E Ratio for S&P 500 . . . 8066
15.288 Distribution and cumulative function of the beta law . . . 8082
15.289 BPMN process representative diagram with Microsoft Visio 2010 . . . 8086
15.294 Sigma intervals for the Normal distribution . . . 8092
15.295 Serial/Process Workflow . . . 8093
15.296 Pareto chart . . . 8097
15.297 Lorenz chart . . . 8098
15.298 The two most common situations of the Gini index . . . 8099
15.301 6M+E Ishikawa diagram . . . 8102
15.302 Weighted Ishikawa diagram . . . 8104
15.304 First resupply strategy . . . 8120
15.311 Cumulative costs $C_{\text{tot}}(Q)$ . . . 8126
15.312 Representation of the economic area . . . 8129
15.313 Wilson model with resupply . . . 8133
15.314 Wilson model without resupply . . . 8133
15.315 Wilson model with resupply and break-up . . . 8137
15.316 The science of lines . . . 8158
15.317 Probability of saturation (smoothed plot of the Erlang-B function) . . . 8164
15.321 Simulation of two small tasks respectively with a Normal and Beta distribution 8191
15.322 Correlation coefficient between the sums and each task . . . 8191
15.323 Corresponding correlation values . . . 8191
15.328 Vocabulary of a longitudinal wave propagation in an ideal gas . . . 8196
15.329 Some instruments frequencies ranges . . . 8197
15.330 Basic frequency spectrum with transported surface power density . . . 8198
15.331 Various types of amplitudes . . . 8206
15.332 Spherical wavefront . . . 8210
15.333 Fixed oscillating punctual source . . . 8213
15.334 Schematic diagram of a fixed source and an approaching observer (Doppler Effect) . . . 8214
15.335 Schematic diagram of a fixed observer and an approaching source (Doppler Effect) . . . 8215
15.336 Jet fighter planes with conical shock waves made visible by condensation . . 8218
15.337 Jet fighter with conical shock waves in the case of a conical "nose" . . . 8219
15.339 Schematic diagram of the harmonic oscillator . . . 8224
15.340 Resonances curves with traditional notations . . . 8230

16.1 Earth's inclination . . . 8240
16.2 Seasons space point of view . . . 8240

16.3 Annual mean temperatures on Earth . . . 8241
16.4 Earth simulated day length evolution . . . 8242
16.5 Geological time-scale evolution of temperature for Earth . . . 8243
16.6 Weather vs Climate . . . 8243
16.9 Earth's curvature visibility . . . 8249
16.10 Distance between point on tangent line and circle . . . 8251
16.11 Horizon line with sea view . . . 8257
16.13 Typical representation of isobars (on a depression) . . . 8263
16.14 High and low pressure systems . . . 8263
16.15 Various meteorological scales . . . 8265
16.16 Typical profile of temperature and pressure on Earth . . . 8267
16.17 Greenhouse effect concept . . . 8275
16.18 Past and future Milankovitch cycles . . . 8277
16.19 Map of global diurnal temperature range over land during a few decades . . . 8280
16.20 Sun rays in Summer vs in Winter . . . 8283
16.21 Tilt angle influence simple experiment . . . 8284
16.23 Nutation of tilt axes and influence on seasons . . . 8287
16.25 Simplified schematic and idealized caricature of wind circulation on Earth . . 8292
16.26 Hadley and Ferrel cells . . . 8293
16.30 Low pressure and high pressure illustration . . . 8300
16.31 High and low tide at Mont Saint-Michel in France . . . 8301
16.32 Tidal bore . . . 8302
16.37 Real complexity of tides on Earth . . . 8314
16.38 Principle of tides superimposed on diurnal tides . . . 8316
16.39 One of the everyday life most known convection... . . . 8318
16.40 Rayleigh-Bénard convection cells (instability) . . . 8328
16.41 Rayleigh-Bénard convection cells viewed from top (instability) . . . 8328
16.42 Rayleigh-Bénard convection cells on our Sun . . . 8329
16.43 Rayleigh-Bénard convection cells on Earth . . . 8329
16.44 Rayleigh-Bénard convection cells as supposed in Earth mantle . . . 8330
16.45 Vertical fluid loop with gradient in homogeneous gravific field . . . 8330
16.46 Required configuration for the theoretical model of convection cell . . . 8331
16.47 Phase space of the Lorenz equations system in Maple™ 4.00b . . . 8338
16.55 Evolution of weather forecasting accuracy . . . 8342
16.56 Intergovernmental Panel on Climate Change (IPCC) forecasts vs measurements 8343
16.57 Speed-Period gravity wave relation plot . . . 8352
16.58 Most common type of well-known gravity wave . . . 8354
16.59 Water gravity wave study summary . . . 8358
16.60 Some typical ideal mechanical advantages . . . 8361
16.61 Small gear train . . . 8362
16.62 Single gear simple definitions . . . 8363
16.64 Spur Gear definitions . . . 8365
16.65 Motor with few pulleys and drive belts . . . 8366
16.66 A gear system with particular transmission ratio of a watch . . . 8366
16.67 Bicycle gears ratios . . . 8366
16.68 Gear example with bicycle . . . 8367
16.69 Gear train example in 3D with multiple wheels . . . 8368
16.70 Equivalent gear train example in 2D with multiple wheels . . . 8369

16.72 Most common types of gears . . . 8375
16.73 Continuously Variable Transmissions (CVT) . . . 8376
16.74 Epicyclic gear . . . 8377
16.75 Bicycle hub gear (special example) . . . 8377
16.76 Epicyclic gear simplified schema . . . 8378
16.77 Traction on a test piece . . . 8384
16.78 Beam (simplified ...) subjected to a force . . . 8388
16.79 Cantilevered beam with punctual force on extremity . . . 8390
16.80 Sustained beam with punctual centered force . . . 8392
16.81 Beam fixed on one side with constant line load . . . 8393
16.82 Most common types of material deformations . . . 8397
16.83 Wire under torsion . . . 8404
16.85 Spiral spring under load . . . 8407
16.86 Column under a concentric axial load exhibiting the characteristic deformation of buckling . . . 8410
16.87 Real life buckling . . . 8410
16.88 Four so named "Euler-cases" buckling examples . . . 8411
16.89 Most common types of beams, channels, bars and rails . . . 8414
16.90 Column exhibiting a compressive buckling load due to its own weight . . . 8415
16.93 Simplified schema of a rotating ring . . . 8434
16.94 Electronic cheat sheet poster . . . 8442
16.95 Schematic diagram of an alternating current generator . . . 8445
16.96 Example of nuclear plant . . . 8446
16.97 Example of direct current generator (dynamo) . . . 8446
16.100 Electric Power triangle . . . 8455
16.101 Typical example of a single-phase transformer with a ferromagnetic core . . . 8456
16.102 Concentric monophasic transformer type . . . 8457
16.103 Schematic representation of various transformers . . . 8458
16.104 Typical civilian voltage transformation cascade installation in Switzerland . . . 8458
16.105 Largest plant-oil-based transformer in the World in 12014 . . . 8459
16.106 Cutaway view of liquid-immersed construction transformer . . . 8459
16.107 Typical pole-mounted distribution transformer in USA with center-tapped secondary winding . . . 8460
16.108 Small domestic size transformers . . . 8461
16.111 RC serie circuit . . . 8467
16.113 RL circuit serie . . . 8472
16.114 RLC serie circuit . . . 8475
16.119 Mechanical analogue of RLC circuit . . . 8485
16.120 Telegrapher scheme . . . 8487
16.122 DAC and ADC converter concepts . . . 8513
16.124 pn-junction principles . . . 8519
16.125 Formation of a space charged region and energy band diagram of a pn-junction in thermal equilibrium . . . 8521
16.128 Current vs voltage for forward biased Si and Ge diodes . . . 8535
16.129 Simple amplifier (follower) . . . 8536
16.130 Basic operational amplifier configurations . . . 8537
16.132 Some important ISO standard in Civil Engineering for building construction . 8540
16.133 Massive object (high density plank) against a wall . . . 8541

16.134 Simple fixed pulley 8542
16.135 Pull angle at 180° 8543
16.136 Pull angle at 90° 8543
16.137 Atwood's machine 8544
16.138 Single mobile pulley 8545
16.139 Single fixed and mobile mixed pulleys 8546
16.140 System of combined fixed and moving pulleys 8546
16.141 Unfolded system of combined fixed and moving pulleys 8547
16.142 System of mixed fixed and mobile pulley 8548
16.143 Summary of typical pulleys 8549
16.144 Typical differential pulleys 8550
16.145 HMC Semi-Submersible $10,000$ metric tons Crane vessel 8550
16.146 Simple windlass 8551
16.147 Differential windlass 8552
16.148 Train loop with a clothoid shape 8553
16.149 Highway bifurcation with a clothoid shape 8553
16.150 Cable car pylon with a clothoid shape 8554
16.151 Roaller coaster vertical loop with a clothoid shape 8554
16.152 Plot of the Fresnel (Euler spiral/Cornu spiral) integral in Maple™ 4.00b 8556
16.155 High-voltage overhead cable 8558
16.157 Configuration for studying suspended cables 8559
16.158 Plot between a chain and parabola in Maple™ 4.00b 8562
16.159 The Gateway Arch St. Louis arch 8562
16.160 Plot of a small overhead cable in Maple™ 4.00b 8564
16.161 Overhead cable of railways 8566
16.162 Suspension bridge 8567
16.163 A special case of falling chimney 8576
16.164 Dixence (Switzerland) dam 8579
16.165 Types of dams 8580
16.167 Francis, Kaplan and Pelton wheels 8583
16.168 Illustration of a Pelton wheel hydrorotor 8584
16.169 Schematic view of Taipei 101 tower tuned mass damper 8588
16.170 Taipei 101 tower tuned mass damper 8588
16.171 Bernoulli force pressure lift effect illustration 8595
16.172 Kutta-Joukowski lift effect illustration 8601
16.173 Orbital speed illustration 8604
16.174 Horizontal acceleration profile of Ariane 5 8611
16.175 Rockets of the World 8614
16.176 Geostationary satellites as given by public data in 12006 8617
16.177 International Space Station (ISS) 8618
16.178 Cassini interplanetary trajectory 8622
16.179 Cassini's speed related to Sun 8623
16.180 Special case of gravity assist animation depending on reference frame 8625
16.181 Hyperbolic orbit with classical notation a gravity assist vocabulary 8628
16.182 Hohmann transfer between Earth and Mars 8633
16.183 Hohmann transfer variations 8636
16.184 Some common programming and script languages 8638
16.185 Bisection method scheme 8640

16.186 Visualization of the binary search algorithm 8642
16.187 Hanoi tower algorithm illustration 8648
16.188 Schematic principle of the Bubble Sort algorithm 8651
16.189 Schematic principle of the QuickSort algorithm 8653
16.190 Schematic principle of the graph to study 8663
16.195 Graph with an absorbent state 8669
16.197 Illustration of C_p index 8684
16.198 Typical plot of a control histogram with tolerances and limits 8685
16.199 Measures under statistical control offset from the target (typical with index C_p) 8689
16.200 Capability diagram generated with Minitab 15.1.1 8692
16.201 Short/Long term process 8693
16.202 Relations between $C_{p,\text{observed}}$ and $C_{p,\text{true}}$ for fixed C_{pc} 8696
16.206 Different types of maintenance 8717
16.207 Example of a reliability plot of capacitor 8718
16.209 Common reliability indicators 8730
16.210 Plot of the hazard function and reliability of the exponential law 8733
16.211 Weibull density and distribution functions for different α values 8736
16.212 Plot of four Weibull survival functions for different parameter values 8739
16.213 Plot of four Weibull hazard functions for different parameter values 8740
16.214 Serial topology 8743
16.215 OR gate 8743
16.216 Parallel topology 8745
16.217 AND gate 8745
16.218 k/n topology 8748
16.219 Gate k-OR (voting OR) 8748
16.220 Serial/Parallel symmetric topology 8750
16.221 Parallel/Serial symmetric topology 8750
16.222 Complex topology (bridge type) 8751
16.226 Example of simple fault tree 8755
16.227 Example of simple fault tree with binary probabilities and AND / OR gates only 8756
16.228 Markov chain reliability model example 8757
16.229 Kaplan-Meier typical survival curves in R software 8768
16.230 Graph of the ABC method 8774
16.231 ABC method associated with a Pareto curve 8776
16.233 Principle of construction of centered reduced units 8784
16.234 Two normalized factors study domain example 8785
16.235 Two factors with the response surface 8785
16.236 Generic example of interaction for two factors (hyperbolic paraboloid) 8788
16.239 Two factors with the response surface 8799
16.241 Henry plot of the replicated factorial design with Minitab 17.1.3 8806
16.242 Traditional representation of a full factorial design with 3 factors 8814
16.244 Visual 2^2 factorial design and fractional factorial design 8817
16.245 Display of fractional factorial designs resolutions selector with Minitab 15.1.18820
16.246 Display of fractional factorial designs resolutions selector in Design Expert . 8820
16.247 Representation of a real fractional factorial design with 3 factors 8822
16.248 Taguchi L8 linear graph 8839
16.249 Generic example of interaction for two factors quadratic polynomial 8849

16.250 2D Composite Center Design 8856
16.251 2D Composite Center Design variance perspective plot in Maple™ 4.00b . . 8859
16.252 2D Composite Center Design variance top plot in Maple™ 4.00b 8859
16.253 2D Faced-centered Design variance perspective plot in Maple™ 4.00b 8862
16.254 2D Faced-centered Design variance top plot in Maple™ 4.00b 8862
16.257 Evolutionary solver settings for seeking D-optimal design in Microsoft® Excel® 14.0.7015 8869
16.258 Final D-optimum design following NIST recommendation in Microsoft® Excel® 14.0.7015 8870
16.260 Explicit formulas for index efficiency compliant to NIST in Microsoft® Excel® 14.0.7015 8872
16.261 Index efficiency conforming to NIST in Microsoft® Excel® 14.0.7015 8873
16.262 JMP (SAS) efficiency index of optimal designs 8873
16.263 Three factors mixtures design simplex 8875
16.264 Various arrangements of three factors mixture simplex 8877
16.265 4 factors simplex centroid example 8877
16.266 Three factor mixtures design complex 8880
16.267 Simplex Centroid Design 8893
16.268 Simplex centroid design in each point of a 2^3 full factorial experiment 8894
16.269 2^3 factorial design in each point of a simplex centroid design 8894
16.272 Derringer and Suich desirability functions plot 8902
16.274 Larson's binomial nomograph 8923
16.275 MIL 105E acceptance sampling table 8925
16.276 Binomial, Hypergeometric and Ideal OC for AQL$=2.5\%$ and $n=200$ 8929
16.277 Operating curve summary 8930
16.278 AOQ (Average Outgoing Quality) binomial chart $A=30$, $n=700$ 8933
16.279 Naive control chart 8934
16.280 Run Chart with Minitab 17.1.3 8935
16.283 Minitab control chart wizard 8943
16.284 iSixSigma control chart wizard 8943
16.286 $\bar{X}$ control chart operating characteristic curve 9047
16.287 Pacific Northwest National Laboratory automatic long-term performance measurement equipment 9050
16.288 Exponential Chi-squares time test in Weibull++ 7 software 9051
16.289 Binomial sample size demonstration plan in Weibull++ 7 software 9054
16.290 Weibull reliability sample size with Weibull++ 9056
16.291 Weibull reliability test time with Weibull++ 9057
16.292 Weibull reliability sample size for MTBF test with Weibull++ 9058
16.293 Beta-Binomial sample size demonstration plan in Weibull++ 7 software . . . 9061

List of Tables

4.1 Methodology for Maths & Physics developments . . . 51
4.2 Reminder of the difference scale of Sources in Sciences . . . 61

5.1 Truth table of values . . . 145
5.2 Truth table of explicit values . . . 145
5.3 Truth table of binary values . . . 145
5.4 `OR` inclusive truth table . . . 146
5.5 `OR` exclusive (XOR) truth table . . . 146
5.6 `AND` truth table . . . 147
5.7 Implication truth table . . . 148
5.8 Truth table of the equivalence . . . 151
5.9 Summary of logical core operators . . . 155
5.10 Truth table of main logical operators . . . 155
5.11 Basics Deductive Logical Laws - Common Implications and Equivalences . . 156
5.12 Combination of propositional variables . . . 158
5.13 Consistency, Completeness, Decidability of common axiomatic systems . . . 194
5.14 Common symbols in Classical Logic and Proof Theory . . . 201
5.15 Most well known rules of inference and fallacies in Classical Logic . . . 202
5.16 Some well known basic variables transformations . . . 239
5.17 Greek Alphabet . . . 280
5.18 Summary of typical domain of definitions in Switzerland . . . 282
5.19 Summary of typical domain of definitions according to ISO norms . . . 283
5.20 Binary Relations . . . 291
5.21 Laws and their properties in $\mathbb{N}$ . . . 420
5.22 Laws and their properties in $\mathbb{Z}$ . . . 422
5.23 Laws and their properties in $\mathbb{Q}$ . . . 424
5.24 Summary of most common algebraic structures . . . 435
5.25 Some famous probabilities (and estimations) values . . . 474
5.26 Classical cases of probabilities . . . 481
5.27 Resume of main Combinatorial Analysis cases . . . 532
5.28 Identification of the median . . . 551
5.29 Identification of the median and the mode . . . 554
5.30 Identification of the mediale . . . 562
5.34 Marginal Mean example . . . 576
5.36 Qualitative judgements of C.Vs commonly accepted . . . 589
5.37 Anscombe's quartet . . . 609
5.38 Anscombe's quartet statistics summary . . . 610
5.39 Bradford Hill correlation-causation criterion . . . 617
5.40 Interpretation for kappa or ICC inter-rater agreement . . . 632
5.41 Sigma quality level with defection / non-defection rate . . . 703

5.44 Occurrence of a digit following the Benford distribution . . . 803
5.45 Summary of most common discrete distributions . . . 815
5.46 Summary of most common continuous distributions . . . 815
5.47 One question survey sample size with fpc factor . . . 876
5.49 Interpretations of the Bayes Factor . . . 915
5.50 Chebychev typical intervals to the mean . . . 922
5.51 Conversion sigma to p-value in unilateral . . . 952
5.52 Conversion sigma to p-value in bilateral . . . 952
5.53 Typical "cross" table of a one factor ANOVA . . . 1007
5.54 One-way fixed factor ANOVA table . . . 1018
5.63 Typical cross structure for the analysis of variance of 2 fixed factors without repetitions . . . 1040
5.64 Two-way fixed factor ANOVA table without repetitions . . . 1043
5.65 Typical cross structure for the analysis of variance of 2 fixed factors with repetitions . . . 1045
5.66 Two-way fixed factor ANOVA table with repetitions . . . 1050
5.68 Hierarchical (random) factor ANOVA table . . . 1071
5.69 Hierarchical (fixed) factor ANOVA table . . . 1072
5.71 One-way blocked ANOVA table . . . 1074
5.73 Latin square ANOVA table . . . 1080
5.74 Graeco-Latin square of order 4 . . . 1082
5.75 Greaco-Latin square ANOVA table . . . 1090
5.76 Summary of some naive parametric tests/statistics and their nonparametric equivalent . . . 1154
5.89 Flexible distribution-free & robust testing common hypotheses about numerical data (author: Adrian Olszewski) . . . 1192
5.90 Tabulated values of the Hartley constant $d_2(n)$ . . . 1205
5.91 Tabulated values of the $d_3(n)$ constant . . . 1206
5.92 Generic representation of a P.C.A. table . . . 1214
5.94 Contingency table (crosstab) of CFA . . . 1233
5.95 Joint frequency table of CFA . . . 1233
5.96 Transformation into percentages of the contingency table . . . 1234
5.99 Contingency table (crosstab) for χ^2 analysis . . . 1239
5.100 Contingency table (crosstab) for χ^2 analysis with theoretical effective . . . 1239
5.106 φ Pearson correlation coefficient . . . 1252
5.113 κ_C usage interpretation table . . . 1262
5.115 Yule's Q usage interpretation table . . . 1268
5.119 Typical 2×2 contingency table for the CMH test . . . 1284
5.125 Absolute and Relative errors common relations . . . 1314

6.1 Table for inequality resolution . . . 1358
6.2 Augmented table for inequality resolution . . . 1358
6.3 Highlighted table for inequality resolution . . . 1358
6.4 Limiting behaviour of polynomials . . . 1370
6.5 List of some Laguerre polynomials . . . 1407
6.6 List of some associated Laguerre polynomials . . . 1411
6.7 Most common polynomials . . . 1416
6.8 Symmetries of dihedral group of order 3 . . . 1432

6.9 Compositions of transformations of the tetrahedron . . . 1433
6.10 Most common derivatives . . . 1491
6.11 Most common primitives . . . 1559
6.12 Some of the most famous differential equations . . . 1604
6.13 Typical solutions of H.E. of The L.D.E. with constant coefficients . . . 1615
6.14 List of typical behaviours in planar systems . . . 1643
6.15 Bernoulli numbers . . . 1671
6.16 Most common Maclaurin series . . . 1713
6.17 Summary of vector differential operators . . . 1918
6.18 Inversions and permutations of a determinant of order 2 . . . 1966
6.19 Inversions and permutations of a determinant of order 3 . . . 1967
6.20 Matrix Factorizations . . . 2034

7.1 Example of transformation of a wide crossed table to a long format table (i.e. flat table) . . . 2173
7.3 Terminology about functions . . . 2214
7.4 A summary of some elementary college level functions . . . 2224
7.5 Most common limits . . . 2268
7.6 Properties of the logarithm in base a . . . 2285
7.7 Summary of most common Fourier transform properties . . . 2318
7.8 Most common Fourier transforms . . . 2324
7.9 Summary of most common Laplace transform properties . . . 2348
7.10 Most common Laplace transform . . . 2356
7.11 Most common $\mathcal{Z}$-Transforms . . . 2373

8.1 Odd and even trigonometric functions . . . 2515
8.2 Some identity angles and conventional trigonometric associated values table . 2516
8.3 Objects, representations and corresponding dimensions . . . 2556
8.4 Some common angles values . . . 2590
8.5 Some elliptic eccentricity values . . . 2781
8.6 Some common areas scales . . . 2888
8.7 Five regular polyhedra (Platonic Solids) . . . 2935

9.1 Metric prefix . . . 3100
9.2 Typical vacuum values and mean free paths . . . 3101
9.3 Typical time scales . . . 3102
9.4 Typical distances scales . . . 3103
9.5 Typical velocities scales . . . 3104
9.6 Typical masses scales . . . 3105
9.7 Typical densities scales . . . 3106
9.8 Typical energy scales . . . 3107
9.9 Typical powers scales . . . 3108
9.10 Typical temperature scales . . . 3109
9.11 Typical magnetic scales . . . 3110
9.12 Some famous international standards that make authority in the field of experimental science . . . 3111
9.13 CODATA recommended values of the fundamental physical constants (12022)3115
9.14 Simplification of physics equations in Planck units . . . 3123
9.15 Conservation Laws . . . 3142

9.16 Some common speed scales . . . 3229
9.17 Some common acceleration scales . . . 3236
9.18 Theoretical precession periods of some giant planets . . . 3347
9.19 Comparison of translational (or linear) and rotational (or angular) relations . . . 3382
9.20 Foucault pendulum various periods . . . 3431
9.22 Most famous type of energy in thermodynamics . . . 3600
9.23 Denomination of common constant systems . . . 3604
9.24 Common extensives and intensives state variables . . . 3609
9.25 Some Debye temperatures . . . 3626
9.26 Some gases molar heat capacities . . . 3629
9.27 Summary of basic thermodynamics processes . . . 3643
9.28 Main ideal gas processes (reversible processes) . . . 3650
9.29 Some typical Carnot efficiencies . . . 3670
9.30 Some Thermal conductivities . . . 3685
9.31 Thermal conductivities and specific heat of some ceramics . . . 3686
9.32 Similarities of different diffusion laws in physics . . . 3702
9.33 Conversion factors for pressure measurement units . . . 3737
9.34 Some dynamic viscosity values . . . 3776
9.35 Drag coefficient C_D examples . . . 3874
9.36 Summary of ideal gas laws . . . 3895
9.37 Densities of various typical gases at normal pressure and temperature . . . 3906
9.38 PV triple point for some substances . . . 3910
9.39 Temperature felt based on measured temperature and humidity . . . 3923
9.40 Some typical free mean path for air . . . 3928
9.41 Different types of vacuum . . . 3929

10.1 Summary of common magnetic properties . . . 4092
10.2 Overview of the Curie temperature of various ferromagnetic and ferrimagnetic materials . . . 4093
10.3 Some typical power of resolution based on Rayleigh criteria . . . 4124
10.4 Table of some electrical conductivities at room temperature . . . 4233
10.5 AWG Codes . . . 4236
10.6 Some attenuation skin effect factor for Copper . . . 4271
10.7 Fermi energy for some materials . . . 4290
10.8 Values of a few semiconductors gaps . . . 4303
10.10 Visible spectrum summary . . . 4326
10.12 Various index of refraction of air . . . 4407
10.13 Some Einstein coefficients A_{ki} for various elements and transitions . . . 4593

11.1 Sub-levels of electrons in the main layers . . . 4676
11.2 Modern quantum number notation . . . 4678
11.3 Examples of some atomic elements electronic configuration . . . 4679
11.4 Examples of some atomic elements electronic configuration (extended version) 4680
11.5 Relative abundance of elements if various types of Clay . . . 4685
11.6 Summary of common wave (probabilistic) quantum physics in the orthodox interpretation . . . 4702
11.7 Comparison between Classic and Dirac formalism . . . 4732
11.8 First eigenfunctions and eigenvalues for the 1D harmonic oscillator . . . 4818
11.9 Most well known variations of the Casimir effect experiment and others . . . 4839

11.10 Types of orbital and spin for electrons . . . 4889
11.11 Various common Planck dimensions . . . 4905
11.12 g_L Lande' g-factor for Sodium Na . . . 4993
11.13 Types of known C, P and T violations . . . 5010
11.14 Different isotopes of Iron . . . 5025
11.15 Examples of isotopes that change the properties of a medium ("thermal isotope") 5026
11.17 Amplitudes of the radioactive periods of a few nuclides . . . 5091
11.18 Absorbing materials and their α particle penetration depths . . . 5129
11.19 Absorbing materials and their β^- particle maximum penetration depths . . . 5129
11.20 Linear attenuation coefficient of γ rays in human body . . . 5130
11.21 Summary of major radiation dosimetry units . . . 5133
11.22 Common radiation values . . . 5140
11.23 Some "fundamental" particles mean lifetime and mass . . . 5202

12.1 Statistics for the Rare Life Hypothesis . . . 5226
12.2 Actual durations and angular sizes of the actual western zodiacal signs . . . 5253
12.3 Astrological western zodiacal dates vs Scientific western zodiacal dates . . . 5254
12.4 Most well known variations of the Cavendish experiment and others . . . 5276
12.5 Some Mercury characteristics . . . 5351
12.6 Some Venus characteristics . . . 5351
12.7 Some Earth characteristics . . . 5352
12.8 Some Mars characteristics . . . 5352
12.9 Some Jupiter characteristics . . . 5353
12.10 Some Saturn characteristics . . . 5353
12.11 Some Uranus characteristics . . . 5354
12.12 Some Neptune characteristics . . . 5354
12.13 Some Pluto characteristics . . . 5355
12.14 Various synodic and sidereal periods . . . 5393
12.15 Solar flares Geomagnetic scale . . . 5563
12.16 Solar flares intensity scale . . . 5563
12.17 Summary of the average temperatures of typical Sun's layers . . . 5565
12.18 Some stars data table . . . 5590
12.19 Summary of most common visibility criteria in astrophysics with their traditional notation and units . . . 5590
12.20 Some stars, distances and luminosities . . . 5591
12.21 Data table of some well known galaxies . . . 5621
12.22 Some famous metrics (source: Catalogue of Space-times [402]) . . . 5734
12.23 Cosmic inventory . . . 5898
12.24 Possible topologies of our Universe (cosmic topologies) . . . 5900
12.25 Different scenario of density-pressure relation in General Relativity . . . 5913
12.26 Summary of different types of General Relativity metrics derived so far . . . 5992
12.27 Friedmann-Lemaître-Robertson-Walker cosmological flat spaces summary . 6035
12.28 Friedmann-Lemaître-Robertson-Walker cosmological models equations summary . . . 6047

13.1 Nomenclature of layers and sub-layers of the hydrogen atom . . . 6182
13.2 Screening contributions of electrons . . . 6195
13.3 Some values of parameters of the Morse potential . . . 6225
13.4 Table progress of a chemical reaction . . . 6247

13.5 Different molar enthalpy of combustion 6257

14.4 ANOVA table for regression 6344
14.6 Lack-of-Fit ANOVA table 6347
14.7 Reminder of notations for the study of the linear regression 6351
14.8 Some binomial regression models and their corresponding tolerance model and link function 6456
14.15 Example of sensibility and specificity analysis 6520
14.16 Link functions and mean functions of some Generalized Linear Models (GLM)6552
14.17 Summary of naive fundamental statistical tests 6556
14.18 Table of a few common dispersion M-estimators 6563
14.19 Table summary of naive splines 6581
14.22 Gradient descent optimisers 6693
14.23 Component of risk analytics 6770
14.24 Types of data in Data Science 6791
14.25 Full Confusion matrix example 6810
14.26 Some support-confidence interesting measures 6819
14.30 Typical Classification vs Clustering differences 6822
14.31 Naive Bayes spam-email companion example 6826
14.38 Generic truth table 7105
14.39 Consensus theorem truth table 7109
14.40 AND ($\wedge$) Truth table 7112
14.41 OR ($\vee$) truth table 7112
14.42 NOT ($\neg$) truth table 7112
14.43 NOT-AND truth table 7113
14.44 NOT-OR truth table 7113
14.45 EXCLUSIVE OR $\oplus$ Truth table 7113
14.46 Gray code of 0 to 15 7120
14.47 Identification of the reported retentions for the binary sum 7125
14.54 1-bit magnitude comparator truth table 7133
14.56 Some languages in descending order of richness 7192
14.57 Key systems and recent broke 7221
14.58 Input/Output key matches by functions 7226
14.59 Examples of modular exponentiation applications 7232
14.60 Example of key exchange following Diffie-Hellmann protocol 7233
14.61 Some values of the Euler indicator $\phi(m)$ 7237

15.1 Mortality table example 7299
15.4 Increased life expectancy at birth in France 11996-12006 7303
15.8 Evolution of the world population 7313
15.9 Payoff matrix with Nash Equilibrium 7396
15.10 Payoff matrix without Nash equilibrium 7396
15.11 Decision with imperfect information (original form) 7404
15.12 Decision with imperfect information (tabular form) 7405
15.13 Payoff matrix of a non-zero sum game 7410
15.14 Payoff of a game with potential prudent tactic 7411
15.15 Payoff matrix of a non-zero sum game 7412
15.16 Payoff matrix of the stone-scissors-paper zero sum game 7413
15.17 Payoff semi-matrix of the stone-scissors-paper zero sum game 7413

15.18 Payoff matrix of the WoS repetitive game . . . 7417
15.19 Payoff equilibrated matrix of the WoS repetitive game . . . 7418
15.23 Payoff matrix without Nash Equilibrium . . . 7428
15.24 Pascal's wager original matrix . . . 7432
15.28 Simplified Hawk vs Dove evolutionary game matrix . . . 7446
15.29 Hawk vs Dove evolutionary game matrix explicit example . . . 7447
15.30 Naive ranged scoring decision analysis approach . . . 7463
15.31 Naive ranged scoring decision analysis approach . . . 7463
15.32 Different forms of market competition . . . 7492
15.35 Basis points (BPS) and percentage equivalence . . . 7549
15.38 Treasury Bills vs Treasury Bonds vs Treasury Notes . . . 7557
15.39 Comparison between exact rate and approximated continuous . . . 7560
15.40 Fixed-Term loan amortization table . . . 7572
15.41 Constant amortization loan table . . . 7573
15.42 Constant annuity loan table . . . 7575
15.43 Most common types of trading . . . 7589
15.45 Qualitative judgements of PERs commonly accepted . . . 7610
15.46 Differences between Options and Warrants . . . 7646
15.47 Cash Flow of Forward and Future of the point of view of a buyer . . . 7652
15.48 Differences between Put and Call . . . 7670
15.49 Basic vocabulary of a situation of option payoff . . . 7676
15.50 Different elementary Call-Put strategies . . . 7678
15.51 Time Weighted Rate Of Return . . . 7687
15.53 Some Chan-Karolyi-Longstaff-Sanders derived stochastic processes . . . 7781
15.54 Currencies traded most on Foreign Exchange Markets as of April, 12013 . . . 7882
15.55 Standard & Poor rating table between 11981 and 12004 . . . 7885
15.56 Moody's one year risk default . . . 7888
15.57 Moody's, S&P and Fitch Riskmetrics . . . 7888
15.58 Altman's Z-score criteria ranges . . . 7890
15.59 Various MAPE value decision guide (source: Lewis 11982) . . . 7904
15.61 Comparison between Holt's and Brown's DSES + Holt & Winter's multiplicative TSES . . . 7957
15.62 Proportion of A/D variations on a history set . . . 7974
15.64 Proportion of A/D variations relatively to 2 previous periods . . . 7974
15.65 Some equivalence between ARIMA and ETS models . . . 7995
15.66 Summary of ARIMA models . . . 7997
15.67 Risk-Uncertainty Matrix CXO knowledge . . . 8025
15.68 Detailed actualization table under accounting form . . . 8055
15.71 Defectives and Sigma quality level of a non-center process . . . 8093
15.72 Important values of the process control . . . 8094
15.73 Defectives and Sigma quality level of a non-center process . . . 8094
15.76 Ishikawa Satty matrix . . . 8103
15.77 Cycle service level . . . 8109
15.78 Cumulated probability of sales . . . 8112
15.83 Conventional notations in Queuing Theory . . . 8144
15.84 $M/M/1$ queue summary of important relations . . . 8160

16.1 Visual horizon curvilinear distance in function of the altitude (in vacuum) . . . 8246

16.2 Visual horizon curvilinear and straight distance in function of the altitude (in vacuum) . . . 8249
16.3 Visual horizon curvilinear in function of the altitude in vacuum vs in air with $n=1$ and $\mathrm{d}n/\mathrm{d}y \cong 2.4\cdot 10^{-8}$ . . . 8251
16.4 Vertical angle below horizon (degrees) . . . 8259
16.5 Distance between isobars and wind speed (Beaufort units) . . . 8264
16.6 Typical values for Pr . . . 8336
16.7 Some common voltage scales . . . 8440
16.8 Some common current scales . . . 8440
16.9 Fermi Energies, Fermi Temperatures, and Fermi Velocities . . . 8466
16.10 Mechanical analogue of RLC circuit . . . 8486
16.11 Ascensional data of Discovery Shuttle . . . 8605
16.13 Defectives and Sigma quality level of a centered process . . . 8698
16.14 Capability, non-defective rate in% and PPM . . . 8700
16.15 Defectives and Sigma quality level of a non-centered process . . . 8700
16.16 Application of statistical process control analysis . . . 8702
16.17 Summary of discrete maintenance analysis functions . . . 8722
16.18 Engine failures in effort slices . . . 8722
16.19 Analysis of engine failures per effort slices . . . 8723
16.20 Summary of continuous maintenance analysis functions . . . 8728
16.29 Experimental Full Factorial (2 factor) matrix . . . 8796
16.30 Experimental Full Factorial (2 factor) reduced matrix . . . 8797
16.31 Experimental Full Factorial (2 factor) reduced matrix Yates notation . . . 8797
16.32 Types of plans and factors & interactions . . . 8812
16.33 Experimental Full Factorial (2 factor) matrix . . . 8813
16.34 Full Design of experiment with 3 factors and interactions under Yates form . 8813
16.35 Fractional design with 3 factors into Yates form . . . 8817
16.36 Fractional factorial designs resolution summary table . . . 8819
16.37 Some fractional factorial designs with generators and alias . . . 8821
16.39 $\mathrm{OA}_4(2^3)$ Orthogonal array . . . 8837
16.40 $\mathrm{OA}_8(2^7)$ Orthogonal array . . . 8837
16.41 L_8 table with common factorial notation by Taguchi and associated graph . . 8838
16.42 L_8 table with common notation . . . 8838
16.43 L_8 table with common notation . . . 8839
16.45 Taguchi L_4 table . . . 8841
16.46 Taguchi L_9 table . . . 8841
16.47 L_{12} table with common factorial notation by Taguchi and associated graph . . 8841
16.48 Taguchi L_{16} table . . . 8842
16.49 Taguchi L_{16} table . . . 8843
16.50 $L_8(2^7)$ Taguchi interaction table . . . 8844
16.51 $L_4(2^3)$ Taguchi static design . . . 8846
16.52 $L_4(2^3)$ Taguchi dynamic design with 2-level signal factor . . . 8846
16.53 $L_4(2^3)$ Taguchi dynamic design with 3-level signal factor . . . 8847
16.54 Starting full design for D-optimal design seeking . . . 8867
16.55 Final D-optimal design . . . 8870
16.56 Number of points in a mixture lattice design (non-centered) . . . 8878
16.57 L-mixture designs terms for linear, quadratic and upper models . . . 8883
16.59 Unbiasing constant c_4 . . . 8992

24.1 Greek letters 9372

List of Algorithms

1 GCD pseudocode algorithm . . . 346

2 Matrix Inversion using Determinant and Cofactor Method . . . 1985

3 Gram-Schmidt orthogonalization . . . 2066

4 Gaussian elimination (naive) pseudocode algorithm . . . 6306

5 Gaussian elimination . . . 6311

6 Pool Adjacent Violators Algorithm (PAVA) pseudocode algorithm . . . 6583

7 Proportional Parts pseudocode algorithm . . . 6605

8 Regula-Falsi pseudocode algorithm . . . 6607

9 Proportional Parts bisection pseudocode algorithm . . . 6610

10 Newton’s Method . . . 6615

11 Trapezoidal’s rule . . . 6628

12 Simpson’s rule . . . 6630

13 Newton-Raphson method (quadratic Newton) . . . 6678

14 Gauss-Newton method (tangent Newton) . . . 6688

15 OneR (one-Rule) algorithm . . . 6810

16 DBSCAN Algorithm . . . 6886

17 Generic gradient boosting tree pseudocode 6946

18 Escape Time Fractal pseudocode algorithm 7093

19 SIR Model pseudocode algorithm 7342

20 Proportional Parts bisection pseudocode algorithm 8641

21 Binary search pseudocode algorithm 8643

22 Human intuitive pseudocode algorithm of Hanoi tower 8644

23 Hanoi Tower computer pseudocode algorithm 8649

24 Bubble sort pseudocode algorithm 8651

25 Alternative optimized Bubble sort pseudocode algorithm 8652

26 QuickSort pseudocode algorithm with function 8654

Bibliography

Since this work is a reference, we have tried to exclude as many external references as possible in the texts. So we simply gathered a final bibliography list according to a ranking of use and unassuming it's completeness.

The books and documents in french, english and german listed below are what we consider as the best references (thus excluding the free PDFs) which have been consulted for the preparation of this textbook and who is indebted to many high quality borrows. Their reading can be as profitable for all who wish to improve, deepen and broaden their knowledge.

Numerous references, however, are no longer available on the market and it is also necessary that the reader remembers that each entry in the table by itself refers to dozens of other books, it is also impossible to have an exhaustive and objective list (that would be a vicious cycle) of all high-quality books existing in the areas covered by *Opera Magistris*.

The reader will notice that as with everything else in this textbook, the years of edition (publication) are given on the basis of the Holocene scientific calendar and not on the basis of one of the many religious calendars existing in the world based on fairy tales and legends and for obvious reasons...

And here is the complete list:

[1] Georges Aad et al. “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC”. In: *Physics Letters B* 716.1 (12012), pp. 1–29. DOI: `10.1016/j.physletb.2012.08.020`.

[2] E. Abbena, S. Salamon, and A. Gray. *Modern Differential Geometry of Curves and Surfaces with Mathematica*. Textbooks in Mathematics. CRC Press, 12017. ISBN: 9781420010312.

[3] Benjamin P Abbott et al. “Gravitational waves and gamma-rays from a binary neutron star merger: GW170817 and GRB 170817A”. In: *The Astrophysical Journal Letters* 848.2 (2017), p. L13. DOI: `10.3847/2041-8213/aa920c`.

[4] Israel Abramov et al. “Sex and vision II: color appearance of monochromatic lights”. In: *Biology of sex differences* 3.1 (12012), p. 21. DOI: 10.1186/2042-6410-3-21.

[5] Thomas Adam et al. “Measurement of the neutrino velocity with the OPERA detector in the CNGS beam”. In: *Journal of High Energy Physics* 2012.10 (12012), pp. 1–37. DOI: 10.48550/arXiv.1109.4897.

[6] Fred C. Adams. “The degree of fine-tuning in our universe—and others”. In: *Physics reports* 807 (12019), pp. 1–111. DOI: 10.17713/ajs.v43i4.46.

[7] Peter AR Ade et al. “Improved constraints on primordial gravitational waves using Planck, WMAP, and BICEP/Keck observations through the 2018 observing season”. In: *Physical review letters* 127.15 (12021), p. 151301. DOI: 10.48550/arXiv.2110.00483.

[8] B. Aebischer. *Géometrie - Géometrie affine, géometrie euclidienne & introduction à la géometrie projective*. Vuibert, 12011. ISBN: 9782311002768.

[9] N Aghanim et al. “Planck 2013 results. XXVII. Doppler boosting of the CMB: Eppur si muove”. In: *arXiv preprint arXiv:1303.5087* (12013). DOI: 10.1051/0004-6361/201321556.

[10] Yakir Aharonov and David Bohm. “Significance of electromagnetic potentials in the quantum theory”. In: *Physical review* 115.3 (1959), p. 485. DOI: 10.1103/PhysRev.115.485.

[11] Yakir Aharonov, Sandu Popescu, and Daniel Rohrlich. “Conservation laws and the foundations of quantum mechanics”. In: *Proceedings of the National Academy of Sciences* 120.41 (2023), e2220810120. DOI: 10.1142/S0217732325502086.

[12] O Akyilmaz et al. “High resolution gravity field determination and monitoring of regional mass variations using low-earth orbit satellites”. In: *ICGEM, GFZ Data Services*. 12016.

[13] Einstein Albert, W Perrett, and G Jeffery. “The foundation of the general theory of relativity”. In: *Annalen der Physik* 354.7 (11916), p. 769. DOI: 10.1002/andp.19163540702.

[14] Marcelo Alonso and Edward J. Finn. “Physique Générale”. In: 1 (11970).

[15] Pavan Kumar Aluri et al. “Is the observable Universe consistent with the cosmological principle?” In: *Classical and Quantum Gravity* 40.9 (12023), p. 094001.

[16] Yakov Amihud. “Illiquidity and stock returns: cross-section and time-series effects”. In: *Journal of financial markets* 5.1 (12002), pp. 31–56.

[17] Carl D Anderson. “The positive electron”. In: *Physical review* 43.6 (11933), p. 491. DOI: 10.1103/PhysRev.43.491.

[18] Ronald Anderson, HR Bilger, and GE Stedman. ““Sagnac”effect: A century of Earth-rotated interferometers”. In: *American Journal of Physics* 62.11 (11994), pp. 975–985. DOI: 10.1119/1.17656.

[19] Ross Ure Anderson. “Derivation of solar position formulae”. In: *arXiv preprint arXiv:2009.07094* (2020). DOI: 10.48550/arXiv.2009.07094.

[20] J Antoniadis et al. “The second data release from the European Pulsar Timing Array - IV. Implications for massive Black Holes, Dark Matter, and the Early Universe”. In: *Astronomy & Astrophysics* 678 (12023), A50.

[21] Ebby K Anyamba and Joel Susskind. “Evidence of lunar phase influence on global surface air temperature”. In: *Geophysical research letters* 27.18 (12000), pp. 2969–2972.

[22] Sinya Aoki. “Is energy conserved in general relativity?” In: *Modern Physics Letters A* 40.40 (2025), p. 2550208. DOI: `10.1073/pnas.2220810120`.

[23] T. Arens et al. *Mathematik*. Springer Berlin Heidelberg, 12015. ISBN: 9783642449192.

[24] G.B. Arfken, H.J. Weber, and F.E. Harris. *Mathematical Methods for Physicists: A Comprehensive Guide*. Elsevier Science, 12013. ISBN: 9780123846549.

[25] Constantin W Arnscheidt, Robin D Wordsworth, and Feng Ding. “Atmospheric evolution on low-gravity waterworlds”. In: *The Astrophysical Journal* 881.1 (12019), p. 60.

[26] Alain Aspect. “Bell’s theorem: the naive view of an experimentalist”. In: *Quantum [Un] speakables: From Bell to Quantum Information*. Springer, 2012, pp. 119–153. DOI: `10.1007/978-3-662-05032-3_9`.

[27] Gary Atkins et al. “Determination of Earth’s Orbital Parameters Using Doppler Measurements”. In: *arXiv preprint arXiv:2607.21812* (12026). DOI: `10.48550/arXiv.2607.21812`.

[28] P.W. Atkins and R.S. Friedman. *Molecular Quantum Mechanics*. OUP Oxford, 12011. ISBN: 9780199541423.

[29] Julien Aubert, Stéphane Labrosse, and Charles Poitou. “Modelling the palaeo-evolution of the geodynamo”. In: *Geophysical Journal International* 179.3 (12009), pp. 1414–1428.

[30] Quentin Aubourg. “Étude expérimentale de la turbulence d’ondes à la surface d’un fluide. La théorie de la turbulence faible à l’épreuve de la réalité pour les ondes de capillarité et gravité”. PhD thesis. Oct. 12016.

[31] Robert Axelrod. “More effective choice in the prisoner’s dilemma”. In: *Journal of conflict resolution* 24.3 (11980), pp. 379–403. DOI: `10.1177/002200278002400301`.

[32] Robert Axelrod and William D Hamilton. “The evolution of cooperation”. In: *science* 211.4489 (1981), pp. 1390–1396. DOI: `10.1126/science.7466396`.

[33] F.J. Ayala and C.J.C. Conde. *Processes in Human Evolution: The Journey from Early Hominins to Neanderthals and Modern Humans*. Processes in Human Evolution: The Journey from Early Hominins to Neanderthals and Modern Humans. Oxford University Press, 12017. ISBN: 9780198739906.

[34] Yves Ayant and Elie Belorizky. *Cours de mécanique quantique: avec 87 exercices corrigés:[2e cycle, écoles d’ingénieurs]:[avec exercices et problèmes résolus]*. Dunod, 12000. ISBN: 2100047426.

[35] Louis Bachelier. *Théorie de la spéculation*. Gauthier-Villars, 11900. DOI: `10.24033/asens.476`.

[36] CAL Bailer-Jones et al. “Estimating distances from parallaxes. V. Geometric and photogeometric distances to 1.47 billion stars in Gaia Early Data Release 3”. In: *The Astronomical Journal* 161.3 (12021), p. 147. DOI: `10.3847/1538-3881/abd806`.

[37] J Bailey et al. “Measurements of relativistic time dilatation for positive and negative muons in a circular orbit”. In: *Nature* 268.5618 (11977), pp. 301–305. DOI: `10.1038/268301a0`.

[38] D.W. Ball. *Physical Chemistry*. Cengage Learning, 12014. ISBN: 9781285969770.

[39] Martin Balser and Charles A Wagner. "Observations of Earth–ionosphere cavity resonances". In: *Nature* 188.4751 (11960), pp. 638–641. DOI: `10.1038/188638a0`.

[40] S. Bandyopadhyay and M. Cahay. *Introduction to Spintronics*. CRC Press, 12008. ISBN: 9781420004748.

[41] E. Barends and D.M. Rousseau. *Evidence-Based Management: How to Use Evidence to Make Better Organizational Decisions*. Kogan Page, 12018. ISBN: 9780749483753.

[42] D. Baril. *Tout ce que la science sait de la religion*. PUL Diffusion, 12018. ISBN: 9782763736167.

[43] Benjamin C. Bartlett and David J. Stevenson. "Analysis of a Precambrian resonance-stabilized day length". In: *Geophysical Research Letters* 43.11 (12016), pp. 5716–5724. DOI: `10.1002/2016GL068912`.

[44] Issa Bass. *Six Sigma statistics with Excel and Minitab*. Vol. 1. McGraw-Hill, 12007. ISBN: 9780071489690.

[45] Albert E. Beaton and John W. Tukey. "The Fitting of Power Series, Meaning Polynomials, Illustrated on Band-Spectroscopic Data". In: *Technometrics* 16.2 (11974), pp. 147–185. DOI: `10.1080/00401706.1974.10489171`.

[46] Petr Beckmann and Andre Spizzichino. "The scattering of electromagnetic waves from rough surfaces". In: *Norwood* (11987).

[47] Jacob D. Bekenstein. "Black Holes and Entropy". In: *Phys. Rev. D* 7 (8 11973), pp. 2333–2346. DOI: `10.1103/PhysRevD.7.2333`.

[48] I. Bello. *Vacuum and Ultravacuum: Physics and Technology*. CRC Press, 12018. ISBN: 9781498782043.

[49] Jacqueline Belloni. "The role of silver clusters in photography". In: *Comptes Rendus Physique* 3.3 (12002), pp. 381–390.

[50] Riadh Ben Hamouda. *Notions de Mécanique des Fluides: Cours et Exercices Corrigés*. Centre de Publication Universitaire, 12008. ISBN: 9789973374943.

[51] Herbert Benson et al. "Study of the Therapeutic Effects of Intercessory Prayer (STEP) in cardiac bypass patients: a multicenter randomized trial of uncertainty and certainty of receiving intercessory prayer". In: *American heart journal* 151.4 (12006), pp. 934–942.

[52] V.B. Berestetskii, L.P. Pitaevskii, and E.M. Lifshitz. *Quantum Electrodynamics*. Course of Theoretical Physics vol. 4. Elsevier Science, 12012. ISBN: 9780080503462.

[53] Robert W Bergstrom and Allen C Cogley. "Viscous boundary layers in rotating fluids driven by periodic flows". In: *Journal of the Atmospheric Sciences* 33.7 (11976), pp. 1234–1247.

[54] J. Berk and P. DeMarzo. *Corporate Finance: The Core*. Pearson Education, 12013. ISBN: 9780133145014.

[55] Bruno Bertotti, Luciano Iess, and Paolo Tortora. "A test of general relativity using radio links with the Cassini spacecraft". In: *Nature* 425.6956 (2003), pp. 374–376. DOI: `10.1038/nature01997`.

[56] J. Bi. *Sensory Discrimination Tests and Measurements: Sensometrics in Sensory Evaluation*. Wiley, 12015. ISBN: 9781118994788.

[57] I.I. Bigi and A.I. Sanda. *CP Violation*. Cambridge Monographs on Particle Physics, Nuclear Physics and Cosmology. Cambridge University Press, 12009. ISBN: 9781139478243.

[58] E.J. Billo. *Excel for Scientists and Engineers: Numerical Methods*. Wiley, 12007. ISBN: 9780470126707.

[59] K.G. Binmore. *Game Theory and the Social Contract: Playing fair*. Economic Learning and Social Evolution Series. MIT Press, 11994. ISBN: 9780262023634.

[60] H.J. Birx. *Encyclopedia of Anthropology*. A SAGE reference publication. SAGE Publications, 12006. ISBN: 9780761930297.

[61] Ann Merchant Boesgaard and Gary Steigman. "Big bang nucleosynthesis-Theories and observations". In: *IN: Annual review of astronomy and astrophysics. Volume 23 (A86-14507 04-90). Palo Alto, CA, Annual Reviews, Inc., 1985, p. 319-378.* 23 (11985), pp. 319–378. DOI: 10.1146/annurev.aa.23.090185.001535.

[62] Niels Bohr. "I. On the constitution of atoms and molecules". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 26.151 (11913), pp. 1–25. DOI: 10.1080/14786441308634955.

[63] Michael Bordag, Umar Mohideen, and Vladimir M Mostepanenko. "New developments in the Casimir effect". In: *Physics reports* 353.1-3 (12001), pp. 1–205.

[64] Arvind Borde, Alan H. Guth, and Alexander Vilenkin. "Inflationary spacetimes are not past-complete". In: *arXiv preprint gr-qc/0110012* (12001).

[65] Lucien Borel. *Thermodynamique et Energétique*. Presses Polytechniques Romandes, premiére édition, 11984. ISBN: 2880740258.

[66] Sébastien Bossu and Philippe Henrotte. *An Introduction to Equity Derivatives: Theory and Practice*. Vol. 647. John Wiley & Sons, 12012. ISBN: 781119969037.

[67] J.C. Boughner and C. Rolian. *Developmental Approaches to Human Evolution*. Wiley, 12016. ISBN: 9781118524688.

[68] Jean-Louis Boursin. *Initiation à la théorie des jeux*. Montchrestien, 11998. ISBN: 9782707611062.

[69] Audrey Bouvier and Meenakshi Wadhwa. "The age of the Solar System redefined by the oldest Pb–Pb age of a meteoritic inclusion". In: *Nature geoscience* 3.9 (2010), pp. 637–641. DOI: 10.1038/ngeo941.

[70] Frank Bowman. *Introduction to Bessel functions*. Courier Corporation, 11958. ISBN: 9780486604626.

[71] Robert Boyd and Peter J Richerson. "Culture and the evolution of human cooperation". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 364.1533 (12009), p. 3281. DOI: 10.1098/rstb.2009.0134.

[72] S. Boyd and L. Vandenberghe. *Convex Optimization*. Berichte über verteilte messysteme. Cambridge University Press, 12004. ISBN: 9780521833783.

[73] R.B. Braithwaite. *Theory of Games as a Tool for the Moral Philosopher*. Cambridge University Press, 12009. ISBN: 9780521113519.

[74] G.C. Branco, L. Lavoura, and J.P. Silva. *CP Violation*. International series of monographs on physics. Clarendon Press, 11999. ISBN: 9780198503996.

[75] CB Bratton et al. “Angular distribution of events from SN1987A”. In: *Physical Review D* 37.12 (1988), p. 3361. DOI: `10.1103/PhysRevD.37.3361`.

[76] Benjamin Bray. “Expectation Maximization”. In: (2016).

[77] Giacomo Bressi et al. “Measurement of the Casimir force between parallel metallic surfaces”. In: *Physical review letters* 88.4 (12002), p. 041804.

[78] M Brown-Hayes et al. “Thermal and dissipative effects in Casimir physics”. In: *Journal of Physics A: Mathematical and General* 39.21 (12006), p. 6195. DOI: `10.1088/0305-4470/39/21/S10`.

[79] Justyna Brzezińska et al. “Singular Value Decomposition Approaches in A Correspondence Analysis with The Use of R”. In: *Folia Oeconomica Stetinensia* 18.2 (12018), pp. 178–189.

[80] E Buks and ML Roukes. “Stiction, adhesion energy, and the Casimir effect in micromechanical systems”. In: *Physical Review B* 63.3 (12001), p. 033402.

[81] Judith M Burkart, Rahel K Brügger, and Carel P Van Schaik. “Evolutionary origins of morality: Insights from non-human primates”. In: *Frontiers in Sociology* 3 (12018), p. 17. DOI: `10.3389/fsoc.2018.00017`.

[82] J. Burnet et al. *Greek Thought: A Guide to Classical Knowledge*. Harvard University Press reference library. Belknap Press of Harvard University Press, 12000. ISBN: 9780674002616.

[83] Nader Butto. “New Mechanism and Analytical Formula for Understanding the Gravity Constant G”. In: *Journal of High Energy Physics, Gravitation and Cosmology* 6.3 (12020), pp. 357–367. DOI: `10.4236/jhepgc.2020.63029`.

[84] Randolph C Byrd et al. “Positive therapeutic effects of intercessory prayer in a coronary care unit population.” In: (11988).

[85] Bernard Cagnac and Jean Claude Pebay-Peyroula. *Physique atomique: introduction à la physique quantique et structure de l’édifice atomique*. Vol. 2. Dunod, 11971.

[86] Robert F Cahalan, Paulino Ajiquichí, and Gaspar Yatáz. “Solar Temperature Variations Computed from SORCE SIM Irradiances Observed During 2003–2020”. In: *Solar Physics* 297.1 (12022), pp. 1–37.

[87] Renzo Cairoli. *Algèbre linéaire*. PPUR presses polytechniques, 11991. ISBN: 9782880741877.

[88] Kecheng Cao et al. “Imaging an unsupported metal–metal bond in dirhenium molecules at the atomic scale”. In: *Science Advances* 6.3 (12020), eaay5849. DOI: `10.1126/sciadv.aay584`.

[89] Valerio Capraro and Matjaž Perc. “Mathematical foundations of moral preferences”. In: *Journal of the Royal Society interface* 18.175 (12021), p. 20200880. DOI: `10.1098/rsif.2020.0880`.

[90] NL Cates and SJ Mojzsis. “Pre-3750 Ma supracrustal rocks from the Nuvvuagittuq supracrustal belt, northern Québec”. In: *Earth and Planetary Science Letters* 255.1-2 (2007), pp. 9–21. DOI: `10.1016/j.epsl.2006.11.034`.

[91] C.J. Cela-Conde and F.J. Ayala. *Human Evolution: Trails from the Past*. Oxford biology. OUP Oxford, 12007. ISBN: 9780191524431.

[92] Y.A. Çengel and J.M. Cimbala. *Fluid Mechanics: Fundamentals and Applications*. Cengel series in engineering thermal-fluid sciences. McGraw-Hill, 12010. ISBN: 9780071284219.

[93] Fabio ACC Chalub. "The Saros cycle: obtaining eclipse periodicity from Newton's laws". In: *Revista Brasileira de Ensino de Física* 31 (12009), pp. 1303–1. DOI: `10.1590/S1806-11172009000100003`.

[94] Don M Chance. "A synthesis of binomial option pricing models for lognormally distributed assets". In: (12007).

[95] Serguei Chatrchyan et al. "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". In: *Physics Letters B* 716.1 (12012), pp. 30–61. DOI: `10.1016/j.physletb.2012.08.021`.

[96] Alice Y Chen, Phil Halper, and Niayesh Afshordi. "Copenhagen Survey on Black Holes and Fundamental Physics". In: *arXiv preprint arXiv:2503.15776* (12025). DOI: `10.48550/arXiv.2503.15776`.

[97] Ming-Cheng Chen et al. "Ruling out real-valued standard formalism of quantum theory". In: *Physical Review Letters* 128.4 (12022), p. 040403. DOI: `10.1103/PhysRevLett.128.040403`.

[98] Qiang Chen et al. "A practical model for realistic butterfly flight simulation". In: *ACM Transactions on Graphics (TOG)* 41.3 (12022), pp. 1–12. DOI: `10.1145/3510459`.

[99] Chin-Wen Chou et al. "Optical clocks and relativity". In: *Science* 329.5999 (2010), pp. 1630–1633. DOI: `10.1126/science.11927`.

[100] Ulrich R Christensen and Julien Aubert. "Scaling properties of convection-driven dynamos in rotating spherical shells and application to planetary magnetic fields". In: *Geophysical Journal International* 166.1 (12006), pp. 97–114.

[101] GD. Christian, P. Dsgupta, and K. Schug. *Analytical Chemistry, 7th Edition*. Wiley Global Education, 12013. ISBN: 9780470887578.

[102] Caterina Ciminelli and Giuseppe Brunetti. "Laser gyroscope precisely tracks the Earth's rotation: Gyroscopes". In: *Nature Photonics* 17.12 (12023), pp. 1023–1024. DOI: `10.1103/10.1038/s41566-023-01293-y`.

[103] H.C. Co. "Confirmation testing of the Taguchi methods by artificial neural-networks simulation". In: *International Journal of Production Research* 46.17 (12008), pp. 4671–4685.

[104] Robert Cobbaut. *Théorie financière*. 11987. ISBN: 9782717834420.

[105] Edgar F Codd. "Further normalization of the data base relational model". In: *Data base systems* 6.1972 (11972), pp. 33–64.

[106] Olivier Cogis and Claudine Robert. *Théorie des graphes: au-delà des ponts de Königsberg: problèmes, théorèmes, algorithmes*. Vuibert, 12003. ISBN: 2711753212.

[107] Rich F Coleman and Kuo-Nan Liou. "Light scattering by hexagonal ice crystals". In: *Journal of the Atmospheric Sciences* 38.6 (11981), pp. 1260–1271. DOI: `10.1016/j.jqsrt.2016.02.023`.

[108] J.J. Condon and S.M. Ransom. *Essential Radio Astronomy*. Princeton Series in Modern Observational Astronomy. Princeton University Press, 12016. ISBN: 9781400881161.

[109] Liam Connor et al. “A gas rich cosmic web revealed by partitioning the missing baryons”. In: *Nature astronomy* (12024). DOI: 10.1038/s41550-025-02566-y.

[110] Oliver Consa. “The Unpublished Feynman Diagram IIc”. In: *arXiv preprint arXiv:2010.10345* (12020).

[111] Christopher J Conselice et al. “The evolution of galaxy number density at z< 8 and its implications”. In: *The Astrophysical Journal* 830.2 (2016), p. 83. DOI: 10.3847/0004-637X/830/2/83.

[112] Adrián G Cornejo. “Rotation curves of spiral galaxies from a rotating viscous-fluid model: AG Cornejo”. In: *Indian Journal of Physics* (12025), pp. 1–19. DOI: 10.1007/s12648-025-03831-6.

[113] J.A. Cornell. *Experiments with Mixtures: Designs, Models, and the Analysis of Mixture Data*. Wiley Series in Probability and Statistics. Wiley, 12011. ISBN: 9781118150498.

[114] C Richard Cothern and William L Lappenbusch. “Occurrence of uranium in drinking water in the US”. In: *Health physics* 45.1 (11983), pp. 89–99.

[115] Charles Augustin Coulomb. *Mémoire sur l’électricité et le magnétisme*. 11793.

[116] D.L. Couprie. *Heaven and Earth in Ancient Greek Cosmology: From Thales to Heraclides Ponticus*. Astrophysics and Space Science Library. Springer New York, 12011. ISBN: 9781441981165.

[117] John R Cronin and Sandra Pizzarello. “Amino acids in meteorites”. In: *Advances in space research* 3.9 (11983), pp. 5–18. DOI: 10.1016/0273-1177(83)90036-4.

[118] J.F. Crow and M. Kimura. *An Introduction to Population Genetics Theory*. Blackburn Press, 12009. ISBN: 9781932846126.

[119] J.D. Cryer and K.S. Chan. *Time Series Analysis: With Applications in R*. Springer Texts in Statistics. Springer New York, 12008. ISBN: 9780387759593.

[120] Howard Curtis. *Orbital mechanics for engineering students*. Butterworth-Heinemann, 12013. ISBN: 9780080977478.

[121] John D. Cutnell and Kenneth W. Johnson. *Physics*. NJ, John Wiley & Sons, Inc, 12021. ISBN: 9781119773610.

[122] Erik Dahl-Jensen. “The properties of photographic emulsions”. In: (11963).

[123] Paul E Damon et al. “Radiocarbon dating of the Shroud of Turin”. In: *Nature* 337.6208 (11989), pp. 611–615. DOI: 10.1038/337611a0.

[124] H Vic Dannon. *Zero Point Energy: Planck Radiation Law*. 12005.

[125] Elise Darragh-Ford et al. “The Concentration–Mass relation of massive, dynamically relaxed galaxy clusters: agreement between observations and ΛCDM simulations”. In: *Monthly Notices of the Royal Astronomical Society* 521.1 (12023), pp. 790–799. DOI: 10.1093/mnras/stad585.

[126] C.J. Date. *Database Design and Relational Theory: Normal Forms and All That Jazz*. Apress, 12019. ISBN: 9781484255391.

[127] René David et al. *Introduction à la logique: théorie de la démonstration: cours et exercices corrigés*. Dunod, 12001. ISBN: 9782100048922.

[128] J Huw Davies and D Rhodri Davies. “Earth’s surface heat flux”. In: *Solid earth* 1.1 (2010), pp. 5–24.

[129] Tamara M Davis. “Is the universe leaking energy?” In: *Scientific American* 303.1 (2010), pp. 38–47. DOI: 10 . 1038 / scientificamerican072010 - 6fwVdzWOKqjnYX3iUBEUVv.

[130] Emmanuel Davoust. “Dating of a Latin astrolabe”. In: *arXiv preprint arXiv:2311.17966* (12023). DOI: 10.48550/arXiv.2311.17966.

[131] Alastair L. Day. *Mastering Financial Modelling in Microsoft Excel: A practitioner’s guide to applied corporate finance*. Pearson Education, 12007. ISBN: 9780273708063.

[132] Jean-Marie De Koninck and Armel Mercier. *Introduction à la théorie des nombres*. Modulo, 11994. ISBN: 9782891135009.

[133] Willem De Sitter. “On the constancy of the velocity of light”. In: 16.1 (11913), pp. 395–396.

[134] Charles Delaunay. *Théorie du Mouvement de la Lune*. Vol. 28. Mallet-Bachelier, 11867.

[135] Jean-Loup Delcroix and Abraham Bers. *Physique des plasmas, Tome 1*. 11994. ISBN: 9782759802876.

[136] E. Delson, I. Tattersall, and J.A. Van Couvering. *Encyclopedia of Human Evolution and Prehistory*. Garland reference library of the humanities. Garland Pub., 12000. ISBN: 9780815316961.

[137] P. Delva et al. “Test of Special Relativity Using a Fiber Network of Optical Clocks”. In: *Phys. Rev. Lett.* 118 (22 12017), p. 221102. DOI: 10.1103/PhysRevLett.118.221102.

[138] Pacôme Delva et al. “Gravitational redshift test using eccentric Galileo satellites”. In: *Physical review letters* 121.23 (12018), p. 231101. DOI: 10.1103/PhysRevLett.121.231101.

[139] J.P. Demailly. *Analyse numérique et équations différentielles*. Grenoble Sciences. EDP Sciences, 12012. ISBN: 9782759801121.

[140] Mark Denny. “Calculation of the 22° halo”. In: *European Journal of Physics* 18 (11997), pp. 432–435.

[141] Bipin R. Desai. *Quantum mechanics with basic field theory*. Cambridge University Press, 12009. ISBN: 9780521877602.

[142] J. A. DeSanto. “Scattering from a rough interface”. In: *Radio Science* 16.6 (11981), pp. 1021–1024. DOI: https://doi.org/10.1029/RS016i006p01021.

[143] J.W. Dettman. *Applied Complex Variables*. A Series of advanced mathematics texts. Macmillan, 11965. ISBN: 9780486646701.

[144] Eleonora Di Valentino et al. “In the Realm of the Hubble tension - a Review of Solutions”. In: *arXiv preprint arXiv:2103.01183* (12021).

[145] René Didier and Pierre Grécias. *Chimie générale*. Technique & documentation-Lavoisier, 11997. ISBN: 9782852062559.

[146] Ken Dill and Sarina Bromberg. *Molecular driving forces: statistical thermodynamics in biology, chemistry, physics, and nanoscience*. Garland Science, 12010. ISBN: 9780815344308.

[147] R. D’Inverno. *Introducing Einstein’s Relativity*. Clarendon Press, 11992. ISBN: 9780198596868.

[148] Paul AM Dirac. “The Lagrangian in quantum mechanics”. In: *Feynman’s Thesis—A New Approach To Quantum Theory*. World Scientific, 12005, pp. 111–119. DOI: 10.1142/9789812567635_0003.

[149] Bernard Diu et al. *Eléments de physique statistique*. Hermann, 11989. ISBN: 9782705660659.

[150] Inc DK Publishing. *Ultimate Visual Dictionary of Science*. Ultimate Visual Dictionary. DK Pub., 11998. ISBN: 9780789435125.

[151] Chuong B Do and Serafim Batzoglou. “What is the expectation maximization algorithm?” In: *Nature biotechnology* 26.8 (12008), pp. 897–899. DOI: 10.1038/nbt1406.

[152] Or Dobkowski et al. “Observation of the quantum phase of free fall and the consistency with the equivalence principle”. In: *Science Advances* 12.36 (12026), eaec8045. DOI: 10.1126/sciadv.aec8045.

[153] Matthew S Dodd et al. “Evidence for early life in Earth’s oldest hydrothermal vent precipitates”. In: *Nature* 543.7643 (12017), pp. 60–64. DOI: 10.1038/nature21377.

[154] Jacques Douchet and Bruno Zwahlen. *Calcul différentiel et intégral 3. Fonctions réelles d’une variable réelle: exercices résolus*. Presses Polytechniques Romandes, 11987.

[155] Jacques Douchet and Bruno Zwahlen. *Calcul différentiel et intégral: fonctions réelles d’une ou de plusieurs variables réelles*. PPUR presses polytechniques, 12006. ISBN: 9782889155668.

[156] Cpt. K. Becker Dr C. Cranz. *Handbook of Ballistics, Volume I. Translated from Second German Edition*. Stationey Office London, 11921.

[157] John W Drake et al. “Rates of spontaneous mutation”. In: *Genetics* 148.4 (11998), pp. 1667–1686.

[158] Dean G. Duffy. *Advanced engineering mathematics with MATLAB*. CRC Press, 12010. ISBN: 9781584883494.

[159] Georges Duvaut. *Mécanique des milieux continus*. Dunod, 11998. ISBN: 9782100039326.

[160] C.W. LeSage D.W. Thormas C.S. Barrett. *SQC Western Electric Union*. Western Electric Co., Inc., 11956.

[161] Frank Watson Dyson, Arthur Stanley Eddington, and Charles Davidson. “IX. A determination of the deflection of light by the Sun’s gravitational field, from observations made at the total eclipse of May 29, 1919”. In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 220.571-581 (11920), pp. 291–333. DOI: 10.1098/rsta.1920.0009.

[162] R.M. Eisberg and R. Resnick. *Quantum physics of atoms, molecules, solids, nuclei, and particles*. Quantum Physics of Atoms, Molecules, Solids, Nuclei and Particles. Wiley, 11985. ISBN: 9780471873730.

[163] Joseph G. Eisenhauer. “Regression through the origin”. In: *Teaching statistics* 25.3 (12003), pp. 76–80. DOI: 10.1111/1467-9639.00136.

[164] Edgard Elbaz and Mario Novello. *Cosmologie*. Ellipse, 11992. ISBN: 9782729892111.

[165] Roland v. Eötvös, Desiderius Pekár, and Eugen Fekete. “Beiträge zum gesetze der proportionalität von trägheit und gravität”. In: *Annalen der Physik* 373.9 (11922), pp. 11–66. DOI: 10.1002/andp.19223730903.

[166] A.O.E.R.T. Equity. *College Physics Textbook Equity Edition Volume 1 of 3: Chapters 1 - 12*. LULU Press, 12014. ISBN: 9781304803214.

[167] A.O.E.R.T. Equity. *College Physics Textbook Equity Edition Volume 2 of 3: Chapters 13 - 24*. LULU Press, 12014. ISBN: 9781304784568.

[168] A.O.E.R.T. Equity. *College Physics Textbook Equity Edition Volume 3 of 3: Chapters 25 - 34*. LULU Press, 12014. ISBN: 9781304804266.

[169] Johannes R Eskilt et al. "Cosmic topology. Part IIa. Eigenmodes, correlation matrices, and detectability of orientable Euclidean manifolds". In: *Journal of Cosmology and Astroparticle Physics* 2024.03 (12024), p. 036. DOI: `10.48550/arXiv.2306.17112`.

[170] CW Francis Everitt et al. "Gravity probe B: final results of a space experiment to test general relativity". In: *Physical Review Letters* 106.22 (12011), p. 221101. DOI: `10.1103/PhysRevLett.106.221101`.

[171] Xing Fan et al. "Measurement of the electron magnetic moment". In: *Physical review letters* 130.7 (12023), p. 071801. DOI: `10.1103/PhysRevLett.130.071801`.

[172] Vitaly Fedoseev et al. "Coherent and Incoherent Light Scattering by Single-Atom Wave Packets". In: *Physical Review Letters* 135.4 (12025), p. 043601. DOI: `10.1103/zwhd/1k2t`.

[173] E.D. Feigelson and G.J. Babu. *Modern Statistical Methods for Astronomy: With R Applications*. Cambridge University Press, 12012. ISBN: 9780521767279.

[174] Rechnitzer A Feldman J and Yeager E. *CLP 1 Differential Calculus*. 12022.

[175] Rechnitzer A Feldman J and Yeager E. *CLP 2 Integral Calculus*. 12022.

[176] Peter Fellgett. "Equivalent quantum-efficiencies of photographic emulsions". In: *Monthly Notices of the Royal Astronomical Society* 118.3 (11958), pp. 224–233.

[177] Daniel Fels, Michel Cifra, and F Scholkmann. "Fields of the Cell". In: *Research Signpost, Trivandrum, India* (12015). DOI: `10.13140/RG.2.1.1029.7044`.

[178] Andrea Ferroglia and Miguel CN Fiolhais. "Tidal locking and the gravitational fold catastrophe". In: *American Journal of Physics* 88.12 (12020), pp. 1059–1067.

[179] Martin Ferus et al. "Formation of nucleobases in a Miller–Urey reducing atmosphere". In: *Proceedings of the National Academy of Sciences* 114.17 (12017), pp. 4306–4311. DOI: `10.1073/pnas.1700010114`.

[180] MP Fewell. "The atomic nuclide with the highest mean binding energy". In: *American Journal of Physics* 63.7 (1995), pp. 653–658.

[181] Ferdinánd Filip. "An elementary proof that almost all real numbers are normal". In: *Acta Universitatis Sapientiae. Mathematica* 2.1 (12010), pp. 99–110.

[182] M. Fitting and R.L. Mendelsohn. *First-Order Modal Logic*. Synthese Library. Springer International Publishing, 12023. ISBN: 9783031407147. DOI: `10.1007/978-3-031-40714-7`.

[183] R. Fitzpatrick. *An Introduction to Celestial Mechanics*. Cambridge University Press, 12012. ISBN: 9781107023819.

[184] R. Fitzpatrick. *Quantum Mechanics*. World Scientific Publishing Company, 12015. ISBN: 9789814696005.

[185] Tristan Fletcher. “Support vector machines explained”. In: *Tutorial paper., Mar* (12009), p. 28.

[186] Duncan H Forgan. “A numerical testbed for hypotheses of extraterrestrial life and intelligence”. In: *International Journal of Astrobiology* 8.2 (2009), pp. 121–131. DOI: 10.48550/arXiv.0810.2222.

[187] M Forte et al. “Validation of a method for measuring 226Ra in drinking waters by LSC”. In: *Applied Radiation and Isotopes* 103 (12015), pp. 143–150.

[188] J.B.J Fourier. *The Analytical Theory of Heat*. Dover Books on Physics Series. Dover Publications, 11878. ISBN: 9780486495316.

[189] A. Fraknoi et al. *Astronomy*. Open Textbook Library. OpenStax College, 12016. ISBN: 9781938168284.

[190] David LJ Freed. “Doctors are not scientists but we still need science”. In: *Bmj* 329.7460 (12004), p. 294.

[191] David A. Freedman. “A note on screening regression equations”. In: *the american statistician* 37.2 (11983), pp. 152–155.

[192] V Frenkel and AA Grib. “Einstein, Friedmann, Lemaitre: Discovery of the Big Bang”. In: *Proceed. of the Second Alexander Friedmann Intern Sem. on Gravitation and Cosmology. Ed. Yu. Gnedin, A. Grib, V. Mostepanenko. Friedmann Lab Publ. St. Petersburg* (11994).

[193] Alexander Friedman. “Über die krümmung des raumes”. In: *Zeitschrift für Physik* 10.1 (11922), pp. 377–386.

[194] Harald Fritzsch. “Fundamental constants at high energy”. In: *100 Years Werner Heisenberg: Works and Impact* (12002), pp. 89–95.

[195] Solomon Gandz. “The Distribution of Land and Sea on the Earth’s Surface According to Hebrew Sources”. In: 22 (11953), pp. 23–53. DOI: 10.2307/3622241.

[196] RH Garstang. “Night sky brightness at observatories and sites”. In: *Publications of the Astronomical Society of the Pacific* 101.637 (11989), p. 306. DOI: 10.1086/132436.

[197] Paul R Garvey and Zachary F Lansdowne. “Risk matrix: an approach for identifying, assessing, and ranking program risks”. In: *Air Force Journal of Logistics* 22.1 (11998), pp. 18–21.

[198] D. Gauthier. *Morals by Agreement*. Clarendon Press, 1987. ISBN: 9780191520143.

[199] Ilse Genovese. “Definitions of Surveying and Associated Terms”. In: American Congress on Surveying and Mapping. 12005.

[200] Christian Gentili. *Guide de localisation des astres*. EDP sciences, 12012. ISBN: 2759800598.

[201] Edoardo Gianni et al. “A small polymerase ribozyme that can synthesize itself and its complementary strand”. In: *Science* 391.6789 (12026), pp. 1022–1028. DOI: 10.1126/science.adt2760.

[202] Vincent Giard. *Gestion de la production: calcul Économique*. Economica, 11981. ISBN: 2759800598.

[203] Vincent Giard. *Statistiques descriptives pour les gestionnaires*. Economica, 11995. ISBN: 2717815279.

[204] K. + R. Gieck. *Formulaire Technique Gieck*. Gieck, 11997. ISBN: 9782100035908.

[205] Daniel P Glavin et al. “Abundant ammonia and nitrogen-rich soluble organic matter in samples from asteroid (101955) Bennu”. In: *Nature Astronomy* 9.2 (12025), pp. 199–210. DOI: `10.1038/s41550-024-02472-9`.

[206] E.O. Göbel and U. Siegner. *The New International System of Units (SI): Quantum Metrology and Quantum Standards*. Wiley, 12019. ISBN: 9783527814497.

[207] Mark Gockenbach and Kristin Schmidtke. “Newton’s law of heating and the heat equation”. In: *Involve, a Journal of Mathematics* 2.4 (12009), pp. 419–437.

[208] Kurt Gödel. “An example of a new type of cosmological solutions of Einstein’s field equations of gravitation”. In: *Reviews of modern physics* 21.3 (11949), p. 447. DOI: `10.1103/RevModPhys.21.447`.

[209] Ignacio Gonzalez. “Analyse Canonique Régularisée pour des données fortement multidimensionnelles”. PhD thesis. Université de Toulouse, Université Toulouse III-Paul Sabatier, 12007.

[210] K. Gopalan. *Principles of Radiometric Dating*. Cambridge University Press, 12017. ISBN: 9781107198739.

[211] D.S. Gorbunov and V.A. Rubakov. *Introduction to the Theory of the Early Universe: Hot Big Bang Theory*. World Scientific, 12011. ISBN: 9789814343978.

[212] F Gottsche and Folke Olesen. “A Simple Physically based Model of Diurnal Cycles of Land Surface Temperature”. In: *Proceedings of EUMETSAT Meteorological Satellite Conference Bath, United Kingdom*. 12009, pp. 21–25.

[213] Frank-M Göttsche and Folke S Olesen. “Modelling of diurnal cycles of brightness temperature extracted from METEOSAT data”. In: *Remote Sensing of Environment* 76.3 (12001), pp. 337–348.

[214] J. Goupy. *Pratiquer les plans d’expériences*. Dunod, 12005. ISBN: 9782100042173.

[215] F. Gouthière. *Santé, science, doit-on tout gober ?* Humensis, 12017. ISBN: 9782410009323.

[216] Sander Greenland et al. “Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations”. In: *European journal of epidemiology* 31.4 (12016), pp. 337–350. DOI: `10.1007/s10654-016-0149-3`.

[217] D.A. Grossman and O. Frieder. *Information Retrieval: Algorithms and Heuristics*. Kluwer international series on information retrieval. Springer, 2004. ISBN: 9781402030048.

[218] Rainer Grün and Christopher B Stringer. “Electron spin resonance dating and the evolution of modern humans”. In: *Archaeometry* 33.2 (11991), pp. 153–199. DOI: `10.1111/j.1475-4754.1991.tb00696.x`.

[219] B.C. Gupta. *Statistical Quality Control: Using MINITAB, R, JMP and Python*. Wiley, 12021. ISBN: 9781119671633.

[220] Joseph Carl Hafele and Richard E Keating. “Around-the-world atomic clocks: Observed relativistic time gains”. In: *Science* 177.4044 (11972), pp. 168–170. DOI: `10.1126/science.177.4044.1`.

[221] Ernst Hairer and Gerhard Wanner. *L’analyse au fil de l’histoire*. Vol. 10. Springer Science & Business Media, 12001. ISBN: 9783540674634.

[222] Thomas C Hales. “A proof of the Kepler conjecture”. In: *Annals of mathematics* 162.3 (12005), pp. 1065–1185.

[223] David Halliday et al. *Electricité et magnétisme*. Editions du Renouveau Pédagogique, 11979.

[224] Paul Richard Halmos. “Introduction à la théorie des ensembles”. In: (11970).

[225] J.D. Hamilton. *Time Series Analysis*. Princeton University Press, 12020. ISBN: 9780691218632.

[226] M. Hamilton. *Population Genetics*. Wiley, 12011. ISBN: 9781444362459.

[227] D Hanson et al. “Detection of B-Mode Polarization in the Cosmic Microwave Background<? format?> with Data from the South Pole Telescope”. In: *Physical review letters* 111.14 (12013), p. 141301. DOI: 10.48550/arXiv.1307.5830.

[228] Walter Hantzsche and Hilmar Wendt. “Dreidimensionale euklidische raumformen”. In: *Mathematische Annalen* 110.1 (11935), pp. 593–611. DOI: 10.1007/BF01448045.

[229] A.R. Harish and M. Sachidananda. *Antennas and Wave Propagation*. Oxford higher education. Oxford University Press, 12007. ISBN: 9780195686661.

[230] Elizabeth Harman. *The Irrelevance of Moral Uncertainty*. Oxford Studies in Metaethics, Volume 10 vol. 10. Oxford University Press, 12015. ISBN: 9780198738695. DOI: 10.1093/acprof:oso/9780198738695.003.0003.

[231] Roni Harnik, Graham D Kribs, and Gilad Perez. “A universe without weak interactions”. In: *Physical Review D—Particles, Fields, Gravitation, and Cosmology* 74.3 (12006), p. 035006. DOI: 10.1103/PhysRevD.74.035006.

[232] F.E. Harrell. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Graduate Texts in Mathematics. Springer, 12001. ISBN: 9780387952321.

[233] Frank E Harrell et al. “Evaluating the yield of medical tests”. In: *Jama* 247.18 (11982), pp. 2543–2546. DOI: 10.1001/jama.1982.03320430047030.

[234] ER Harrison. “Olbers’ paradox”. In: *Nature* 204.4955 (11964), pp. 271–272. DOI: 10.1038/204271b0.

[235] Torsten Hartmann and Hans-Georg Wenzel. “The HW95 tidal potential catalogue”. In: *Geophysical research letters* 22.24 (11995), pp. 3553–3556. DOI: 10.1029/95GL03324.

[236] Susumu Hattori et al. “Orientation of high-resolution satellite images based on affine projection”. In: *International Archives of Photogrammetry and Remote Sensing* 33.B3/1; PART 3 (12000), pp. 359–366.

[237] D. Hawkins. *Identification of Outliers*. Monographs on Statistics and Applied Probability. Springer Netherlands, 12013. ISBN: 9789401539944.

[238] Dongshan He, Dongfeng Gao, and Qing-yu Cai. “Spontaneous creation of the universe from nothing”. In: *Physical Review D* 89.8 (12014), p. 083510. DOI: 10.1103/PhysRevD.89.083518.

[239] E. Hecht. *Optics, Global Edition*. Pearson Education Limited, 12016. ISBN: 9781292096964.

[240] E. Hecht. *Physique*. De Boeck Supérieur, 11999. ISBN: 9782744500183.

[241] Giselmar AJ Hemmert et al. “Log-likelihood-based pseudo-R 2 in logistic regression: deriving sample-sensitive benchmarks”. In: *Sociological Methods & Research* 47.3 (12018), pp. 507–531. DOI: 10.1177/0049124116638107.

[242] W. Henke, T. Hardt, and I. Tattersall. *Handbook of Paleoanthropology: Vol I:Principles, Methods and Approaches Vol II:Primate Evolution and Human Origins Vol III:Phylogeny of Hominids*. Handbook of Paleoanthropology. Springer Berlin Heidelberg, 12007. ISBN: 9783540324744.

[243] Bas Hensen et al. “Loophole-free Bell inequality violation using electron spins separated by 1.3 kilometres”. In: *Nature* 526.7575 (12015), pp. 682–686.

[244] Herbert W Hethcote. “The mathematics of infectious diseases”. In: *SIAM review* 42.4 (12000), pp. 599–653.

[245] R.C. Hibbeler. *Engineering Mechanics: Statics and dynamics*. Engineering Mechanics. Prentice-Hall, 12010. ISBN: 9780138149291.

[246] R.C. Hibbeler. *Mechanics of Materials*. Pearson Prentice Hall, 12011. ISBN: 9780136022305.

[247] Peter W Higgs. “Broken symmetries and the masses of gauge bosons”. In: *Physical review letters* 13.16 (11964), p. 508.

[248] WA Hiltner. “Study and Evaluation of Electronography”. In: *National Aeronautics and Space Administration* (1964).

[249] K. Hinkelmann and O. Kempthorne. *Design and Analysis of Experiments, Volume 1: Introduction to Experimental Design*. Wiley Series in Probability and Statistics. Wiley, 12007. ISBN: 9780470191743.

[250] K. Hinkelmann and O. Kempthorne. *Design and Analysis of Experiments, Volume 2: Advanced Experimental Design*. Wiley Series in Probability and Statistics. Wiley, 12005. ISBN: 9780471709930.

[251] K. Hinkelmann and O. Kempthorne. *Design and Analysis of Experiments, Volume 3: Special Designs and Applications*. Design and Analysis of Experiments. Wiley, 12012. ISBN: 9780470530689.

[252] Kazuto Hirata et al. “Observation of a neutrino burst from the supernova SN1987A”. In: *Physical Review Letters* 58.14 (1987), p. 1490. DOI: 10.1103/PhysRevLett.58.1490.

[253] J. Hladik and P.-E. Hladik. *Le calcul tensoriel en physique*. Dunod, 11993. ISBN: 9782100040711.

[254] Jean Hladik. *Les spineurs en physique: avec exercices corrigés*. Masson, 11996. ISBN: 9782225853135.

[255] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-Term Memory”. In: *Neural Computation* 9.8 (11997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

[256] John M Hoenig and Dennis M Heisey. “The abuse of power: the pervasive fallacy of power calculations for data analysis”. In: *The American Statistician* 55.1 (12001), pp. 19–24.

[257] R.V. Hogg, E. Tanis, and D. Zimmerman. *Probability and Statistical Inference*. Pearson Education, 12014. ISBN: 9780321920294.

[258] Connor M Holland, Yukai Lu, and Lawrence W Cheuk. "On-demand entanglement of molecules in a reconfigurable optical tweezer array". In: *Science* 382.6675 (12023), pp. 1143–1147. DOI: `10.1126/science.adf4272`.

[259] T. Holzmann. *Mathematics, Numerics, Derivations and OpenFOAM (4th Edition)*. Loeben, 12017. URL: `https://www.holzmann-cfd.de`.

[260] Shinya Honda et al. "Crystal structure of a ten-amino acid protein". In: *Journal of the American Chemical Society* 130.46 (2008), pp. 15327–15331. DOI: `10.1021/ja8030533`.

[261] J.P. Howard. *Computational Methods for Numerical Analysis with R*. Chapman & Hall/CRC Numerical Analysis and Scientific Computing Series. CRC Press, 12017. ISBN: 9781351646505.

[262] P. Hoyng. *Relativistic Astrophysics and Cosmology: A Primer*. Astronomy and Astrophysics Library. Springer Netherlands, 12007. ISBN: 9781402045233.

[263] Bo Hu, Jun Shao, and Mari Palta. "Pseudo-R 2 in logistic regression model". In: *Statistica Sinica* (12006), pp. 847–860. DOI: `10.1002/sim.2300`.

[264] Edwin Hubble. "A relation between distance and radial velocity among extra-galactic nebulae". In: *Proceedings of the national academy of sciences* 15.3 (11929), pp. 168–173. DOI: `10.1073/pnas.15.3.168`.

[265] E.R. Huggins. *Calculus 2000: A Physics-based Calculus Text*. Moose Mountain Digital Press, 11999. ISBN: 9780970783646.

[266] E.R. Huggins. *Physics 2000 and Calculus 2000*. Moose Mountain Digital Press, 12001. ISBN: 9780970783660.

[267] J.M. Huguenin. *Data envelopment analysis (DEA) : a pedagogical guide for decision makers in the public sector*. Cahier de l'IDHEAP. IDHEAP : Chair of Public finance, 12012. ISBN: 9782940390540.

[268] John C. Hull. *Options, futures, and other derivatives*. Pearson Education India, 12017. ISBN: 9780134472089.

[269] R. Hyndman et al. *Forecasting with Exponential Smoothing: The State Space Approach*. Springer Series in Statistics. Springer Berlin Heidelberg, 12008. ISBN: 9783540719182.

[270] R.J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice*. OTexts, 12018. ISBN: 9780987507112.

[271] Rob J. Hyndman and Yanan Fan. "Sample quantiles in statistical packages". In: *The American Statistician* 50 (11996), pp. 361–365.

[272] Rob J Hyndman et al. "A state space framework for automatic forecasting using exponential smoothing methods". In: *International Journal of Forecasting* 18.3 (12002), pp. 439–454. ISSN: 0169-2070. DOI: `10.1016/S0169-2070(01)00110-8`.

[273] Rob J Hyndman et al. *Prediction intervals for exponential smoothing state space models*. Tech. rep. Monash University, Department of Econometrics and Business Statistics, 12001.

[274] Aapo Hyvärinen and Erkki Oja. "Independent component analysis: algorithms and applications". In: *Neural networks* 13.4-5 (12000), pp. 411–430.

[275] Erik Igelström et al. "Causal inference and effect estimation using observational data". In: *J Epidemiol Community Health* 76.11 (12022), pp. 960–966.

[276] CFA Institute. CFA Curriculum 2019. Wiley, 12018. ISBN: 9781119516927.

[277] Tomoaki Ishiyama et al. "The Uchuu simulations: Data Release 1 and dark matter halo concentrations". In: *Monthly Notices of the Royal Astronomical Society* 506.3 (12021), pp. 4210–4231.

[278] Jacques Istas. *Probabilités et statistiques: cours, exercices et problèmes résolus*. Ellipses, 11999. ISBN: 9782729879884.

[279] Hiromu Ito and Jin Yoshimura. "Social penalty promotes cooperation in a cooperative society". In: *Scientific reports* 5.1 (12015), p. 12797. DOI: `10.1038/srep12797`.

[280] Herbert E Ives and George R Stilwell. "An experimental study of the rate of a moving atomic clock". In: *Journal of the Optical Society of America* 28.7 (11938), pp. 215–226. DOI: `10.1364/JOSA.28.000215`.

[281] Herbert E Ives and GR Stilwell. "An experimental study of the rate of a moving atomic clock. II". In: *Journal of the Optical Society of America* 31.5 (11941), pp. 369–374. DOI: `10.1364/JOSA.31.000369`.

[282] P. Alexis J. Alexis. *Pratique industrielle des plans d'expérience*. AFNOR, 11999. ISBN: 9782124650385.

[283] D. Goetschi J.-Ch. Corbaz. *Calculs commerciaux et bancaires - avec des eléments de statistique descriptive*. Lausanne, 11973.

[284] Elizabeth Jackson. "Belief, credence, and evidence". In: *Synthese* 197.11 (2020), pp. 5073–5092. DOI: `10.1007/s11229-018-01965-1`.

[285] P. Jain. *Wind Energy Engineering, Second Edition*. McGraw Hill LLC, 12016. ISBN: 9780071843850.

[286] J James. "Radar echoes from the Sun". In: *IEEE Transactions on Antennas and Propagation* 12.7 (11964), pp. 876–891. DOI: `10.1109/TAP.1964.1138340`.

[287] C Jaupart, JC Mareschal, and G Schubert. "Heat flow and thermal structure of the lithosphere". In: *Treatise on geophysics* 6 (2007), pp. 217–252.

[288] L. Jaupi. *Contrôle de la qualité (MSP, Analyse des performances, Contrôle de réception)*. Dunod, 12002. ISBN: 9782100042647.

[289] Reza N. Jazar. *Advanced dynamics: rigid body, multibody, and aerospace applications*. John Wiley & Sons, 12011. ISBN: 9780470892138.

[290] F. Jegerlehner. *The Anomalous Magnetic Moment of the Muon*. Springer Tracts in Modern Physics. Springer Berlin Heidelberg, 12007. ISBN: 9783540726340. DOI: `10.1007/978-3-540-72634-0`.

[291] Christian Jenewein et al. "Concomitant formation of protocells and prebiotic compounds under a plausible early Earth atmosphere". In: *Proceedings of the National Academy of Sciences* 122.2 (12025), e2413816122. DOI: `10.1073/pnas.2413816122`.

[292] Raul Jimenez et al. "The local and distant Universe: stellar ages and H0". In: *Journal of Cosmology and Astroparticle Physics* 2019.03 (12019), p. 043. DOI: `10.48550/arXiv.1902.07081`.

[293] Richard Arnold Johnson and Dean W. Wichern. *Applied multivariate statistical analysis*. Pearson Education Limited, 12014. ISBN: 9780131877153.

[294] B. Jordan, R. Westerfield, and S. Ross. *Fundamentals of Corporate Finance Standard Edition*. McGraw-Hill Education, 12012. ISBN: 9780078034633.

[295] Philippe Jorion et al. *Financial risk manager handbook*. Vol. 406. John Wiley & Sons, 12007. ISBN: 9780470126301.

[296] Tanisha Joshi and SD Pathak. "Evolution of Entropy with Cosmic Time". In: *arXiv preprint arXiv:2308.05025* (12023). DOI: `10.48550/arXiv.2308.05025`.

[297] Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever. "An empirical exploration of recurrent network architectures". In: *International conference on machine learning*. 12015, pp. 2342–2350.

[298] Thomas Juffmann et al. "Real-time single-molecule imaging of quantum interference". In: *Nature nanotechnology* 7.5 (12012), pp. 297–300.

[299] Thomas R Junk and Louis Lyons. "Reproducibility and replication of experimental particle physics results". In: *arXiv preprint arXiv:2009.06864* (12020). DOI: `10.48550/arXiv.2009.06864`.

[300] Tom S Juzek and Johannes Kizach. "How to Set Delta in the Two-One-Sided T-tests Procedure (TOST)". In: *Journal of Research Design and Statistics in Linguistics and Communication Science* 5.1-2 (12019), pp. 153–169.

[301] Raghu N. Kacker, Eric S. Lagergren, and James J. Filliben. "Taguchi's orthogonal arrays are classical designs of experiments". In: *Journal of research of the National Institute of Standards and Technology* 96.5 (11991), pp. 577–591.

[302] Chukwunyere Kamalu. "THE ISHANGO BONE: The World's First Known Mathematical Sieve and Table of the Small Prime Numbers". In: (12021).

[303] Alexander Yu Kamenshchik et al. "Time in quantum theory, the Wheeler–DeWitt equation and the Born–Oppenheimer approximation". In: *International Journal of Modern Physics D* 28.06 (12019), p. 1950073. DOI: `10.48550/arXiv.1809.08083`.

[304] Gopal K. Kanji. *100 Statistical tests*. Sage, 12006. ISBN: 9788178297316.

[305] M Abdul Karim et al. "DESI DR2 Results II: Measurements of Baryon Acoustic Oscillations and Cosmological Constraints". In: *arXiv preprint arXiv:2503.14738* (12025). DOI: `10.48550/arXiv.2503.14738`.

[306] H. Karttunen et al. *Fundamental Astronomy*. Springer Berlin Heidelberg, 12007. ISBN: 9783540341444.

[307] Robert E Kass and Adrian E Raftery. "Bayes factors". In: *Journal of the american statistical association* 90.430 (11995), pp. 773–795. DOI: `10.1080/01621459.1995.10476572`.

[308] Sai Krishna Katla, Chenyu Lin, and Juan Pérez-Mercader. "Self-reproduction as an autonomous process of growth and reorganization in fully abiotic, artificial and synthetic cells". In: *Proceedings of the National Academy of Sciences* 122.22 (2025), e2412514122. DOI: `10.1073/pnas.241251412`.

[309] H.J. Keisler. *Elementary Calculus: An Infinitesimal Approach*. Dover Books on Mathematics. Dover Publications, 2013. ISBN: 9780486310466. URL: `https://books.google.ch/books?id=8NTCAgAAQBAJ`.

[310] EM Khokha et al. "Improved modelling for vibrational energies of diatomic molecules using the generalized fractional derivative". In: *Scientific Reports* (2026). DOI: `10.1038/s41598-026-39091-5`.

[311] William H Kinney and Nina K Stein. “Cyclic cosmology and geodesic completeness”. In: *Journal of Cosmology and Astroparticle Physics* 2022.06 (12022), p. 011. DOI: 10.1126/science.1070462.

[312] David B Kirk. “On the numerical approximation of the bivariate normal (tetrachoric) correlation coefficient”. In: *Psychometrika* 38.2 (11973), pp. 259–268. DOI: 10.1007/BF02291118.

[313] R.M. Kliman. *Encyclopedia of Evolutionary Biology*. Elsevier Science, 12016. ISBN: 9780128000496.

[314] E. Kolb. *The Early Universe*. CRC Press, 12018. ISBN: 9780429972492.

[315] E Komatsu et al. “First-Year Wilkinson Microwave Anisotropy Probe (WMAP)* Observations: Tests ofGaussianity”. In: *The Astrophysical Journal Supplement Series* 148.1 (2003), p. 119. DOI: 10.1086/377220.

[316] FG Kondev et al. “The NUBASE2020 evaluation of nuclear physics properties”. In: *Chinese Physics C* 45.3 (12021), p. 030001. DOI: 10.1088/1674-1137/abddae.

[317] Zdeněk Kopal. “Tidal evolution in close binary systems”. In: *Astrophysics and Space Science* 17 (11972), pp. 161–185.

[318] Toon Koppelaars et al. *Applied mathematics for database professionals*. Apress, 12007. ISBN: 9781590597453.

[319] P. Korbel and M. Novák. *The Complete Encyclopedia of Minerals: Description of Over 600 Minerals from Around the World*. Rocks, Minerals and Gemstones. Chartwell Books, 11999. ISBN: 9780785815204.

[320] Israel Koren and C. Mani Krishna. *Fault-tolerant systems*. Morgan Kaufmann, 12010. ISBN: 9780120885251.

[321] John Michael Kovac et al. “Detection of polarization in the cosmic microwave background using DASI”. In: *Nature* 420.6917 (12002), pp. 772–787. DOI: 10.1038/nature01269.

[322] A Kremen et al. “Imaging quantum fluctuations near criticality”. In: *Nature physics* 14.12 (12018), pp. 1205–1210. DOI: 10.1038/s41567-018-0264-z.

[323] E. Kreyszig. *Differential Geometry*. Dover Books on Mathematics. Dover Publications, 12013. ISBN: 9780486318622.

[324] M.H. Kutner. *Applied Linear Statistical Models*. McGrwa-Hill international edition. McGraw-Hill Irwin, 12005. ISBN: 9780071122214.

[325] Olena Kuzii and Andrij Rovenchak. “What the gravitation of a flat Earth would look like and why thus the Earth is not actually flat”. In: *European Journal of Physics* 40.3 (12019), p. 035008.

[326] Amitabha Lahiri and Palash B. Pal. *A first book of Quantum Field Theory*. CRC Press, 12005. ISBN: 9780849338977.

[327] Willis E Lamb Jr and Marlan O Scully. *The photoelectric effect without photons*. Tech. rep. 11968.

[328] L.D. Landau. *Classical Theory of Fields*. Course of Theoretical Physics vol. 2. Elsevier Science, 12013. ISBN: 9781483293288.

[329] L.D. Landau and E.M. Lifshits. *Quantum Mechanics: Non-relativistic Theory*. Course of Theoretical Physics vol. 2. Butterworth-Heinemann, 11977. ISBN: 9780750635394.

[330] L.D. Landau and E.M. Lifshitz. *Fluid Mechanics*. Course of Theoretical Physics vol. 6. Elsevier Science, 12013. ISBN: 9781483140506.

[331] L.D. Landau and E.M. Lifshitz. *Mechanics*. Course of Theoretical Physics vol. 1. Elsevier Science, 11982. ISBN: 9780080503479.

[332] L.D. Landau and E.M. Lifshitz. *Statistical Physics*. Course of Theoretical Physics vol. 5. Elsevier Science, 12013. ISBN: 9780080570464.

[333] L.D. Landau et al. *Electrodynamics of Continuous Media*. Course of Theoretical Physics vol. 8. Elsevier Science, 12013. ISBN: 9781483293752.

[334] L.D. Landau et al. *Theory of Elasticity*. Course of Theoretical Physics vol. 7. Elsevier Science, 12012. ISBN: 9780080570693.

[335] Miljenko Lapaine. “Mollweideova kartografska projekcija”. In: *KoG* 15.15. (12011), pp. 7–16.

[336] Frédéric Laroche. *Promenades mathématiques: histoire, fondements, applications*. Ellipses, 12003. ISBN: 9782729814175.

[337] Edward J Larson and Larry Witham. “Leading scientists still reject God”. In: *Nature* 394.6691 (11998), pp. 313–313.

[338] Jacques Laskar and Mickaël Gastineau. “Existence of collisional trajectories of Mercury, Mars and Venus with the Earth”. In: *Nature* 459.7248 (12009), pp. 817–819.

[339] N.M. Laurendeau. *Statistical Thermodynamics: Fundamentals and Applications*. Cambridge University Press, 12005. ISBN: 9780521846356.

[340] J. Lawson. *Design and Analysis of Experiments with R*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press, 12014. ISBN: 9781498728485.

[341] Z.R. Lazic. *Design of Experiments in Chemical Engineering: A Practical Guide*. Wiley, 12006. ISBN: 9783527604593.

[342] Urbain Le Verrier. “Lettre de M. Le Verrier à M. Faye sur la théorie de Mercure et sur le mouvement du périhélie de cette planète”. In: *Comptes rendus hebdomadaires des séances de l’Académie des sciences* 49 (1859), pp. 379–383.

[343] G.C. Learning. *The Gale Encyclopedia of Cancer: A Guide to Cancer and Its Treatments*. Gale Group, 12010. ISBN: 9781414475981.

[344] Lynn Legg et al. “Occupational therapy for patients with problems in personal activities of daily living after stroke: systematic review of randomised trials”. In: *Bmj* 335.7626 (12007), p. 922.

[345] Georges Lemaître. “Un Univers homogène de masse constante et de rayon croissant rendant compte de la vitesse radiale des nébuleuses extra-galactiques”. In: *Annales de la Société scientifique de Bruxelles*. Vol. 47. 11927, pp. 49–59.

[346] Hartmut Lemmel et al. “Quantifying the presence of a neutron in the paths of an interferometer”. In: *Physical Review Research* 4.2 (12022), p. 023075. DOI: `10.48550/arXiv.2202.00272`.

[347] O. Levin. *Discrete Mathematics: An Open Introduction*. Open Textbook Library. CreateSpace Independent Publishing Platform, 12016. ISBN: 9781534970748.

[348] J.S. Levinton and J.S. Levinton. *Genetics, Paleontology, and Macroevolution*. Genetics, Paleontology, and Macroevolution. Cambridge University Press, 12001. ISBN: 9780521005500.

[349] Leonard S Levitt. *Derivation of the zero-point energy from the uncertainty principle*. 11962.

[350] Kenneth G Libbrecht. “The formation of snow crystals: subtle molecular processes govern the growth of a remarkable variety of elaborate ice structures”. In: *American Scientist* 95.1 (12007), pp. 52–59.

[351] D. Lieberman. *The Evolution of the Human Head*. Mersion: Emergent Village Resources for Communities of Faith Series. Harvard University Press, 12011. ISBN: 9780674046368.

[352] Y.K. Lim. *Problems and Solutions on Quantum Mechanics*. Major American Universities Ph.D. Qualifying Questions and Solutions - Physics. 11998. ISBN: 9789813103009.

[353] Manasvi Lingam and Abraham Loeb. “Risks for life on habitable planets from superflares of their host stars”. In: *The Astrophysical Journal* 848.1 (12017), p. 41.

[354] Bin Liu et al. “Spontaneous emergence of self-replicating molecules containing nucleobases and amino acids”. In: *Journal of the American Chemical Society* 142.9 (12020), pp. 4184–4192.

[355] T. Lockhart. *Moral Uncertainty and Its Consequences*. Oxford University Press, 2000. ISBN: 9780195352160. DOI: `10.1093/oso/9780195126105.001.0001`.

[356] Abraham Loeb. “The habitable epoch of the early Universe”. In: *International Journal of Astrobiology* 13.4 (12014), pp. 337–339. DOI: `10.1017/S1473550414000196`.

[357] S.L. Lohr. *Sampling: Design and Analysis*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press, 12021. ISBN: 9781000478235.

[358] M.S. Longair. *Theoretical Concepts in Physics: An Alternative View of Theoretical Reasoning in Physics*. Cambridge University Press, 12020. ISBN: 9781108484534.

[359] W.H. Louisell. *Quantum statistical properties of radiation*. Wiley Series in Pure and Applied Optics Series. John Wiley & Sons Canada, Limited, 11973. ISBN: 9780471547853.

[360] Y. Lu and S.L. Lohr. *R Companion for Sampling: Design and Analysis*. CRC Press, 12021. ISBN: 9781032132150.

[361] Malcolm Ludvigsen, Marc Lachièze-Rey, and Roger Penrose. *La relativité générale: une approche géométrique*. Dunod, 12000. ISBN: 9782100496884.

[362] Stephen N Luko. “Concerning the estimators r/d2 and r/d2* in estimating variability in a normal universe”. In: *Quality Engineering* 8.3 (11996), pp. 481–487.

[363] F.K. Lutgens, E.J. Tarbuck, and D. Tasa. *Essentials of Geology*. Prentice Hall, 12012. ISBN: 9780321714725.

[364] David K Lynch. “Visually discerning the curvature of the Earth”. In: *Applied Optics* 47.34 (12008), H39–H43. DOI: `10.1364/ao.47.000h39`.

[365] Patrick Lyonnet. *La maintenance: mathématiques et méthodes*. Technique et documentation-Lavoisier, 11992. ISBN: 2852067455.

[366] P.C. Magnusson et al. *Transmission Lines and Wave Propagation, Fourth Edition*. Electrical Engineering. Taylor & Francis, 12000. ISBN: 9780849302695.

[367] Franz Mandl. *Introduction to quantum field theory*. Interscience publishers, 11959.

[368] P. Mangin and R. Kahn. *Supraconductivité*. Grenoble Sciences. EDP Sciences, 12013. ISBN: 9782759808588.

[369] Henry B Mann and Donald R Whitney. "On a test of whether one of two random variables is stochastically larger than the other". In: *The annals of mathematical statistics* (11947), pp. 50–60.

[370] K.V. Mardia, Z.W. Birnbaum, and E. Lukacs. *Statistics of Directional Data*. Elsevier Science, 12014. ISBN: 9781483218663.

[371] Brian G Marsden. "Measurement of the astronomical unit". In: *Astronomical Society of the Pacific Leaflets, Vol. 9, No. 427, p. 209* 9 (11965), p. 209.

[372] M. Martin and R. Monnier. *The Impossibility of God*. G - Reference, Information and Interdisciplinary Subjects Series. Prometheus Books, 12003. ISBN: 9781591021209.

[373] TJ Maternaghan. "Photographic Materials". In: *Encyclopedia of Materials: Science and Technology* (12001), pp. 6936–6945.

[374] PM Mathews, Véronique Dehant, and John M Gipson. "Tidal station displacements". In: *Journal of Geophysical Research: Solid Earth* 102.B9 (11997), pp. 20469–20477. DOI: `10.1029/97JB01515`.

[375] Henry Mathieu and Hervé Fanet. *Physique des semiconducteurs et des composants électroniques-6ème édition: Cours et exercices corrigés*. Dunod, 12009. ISBN: 9782100516438.

[376] MATLAB. *version 8.1.0.604 (R2016a)*. Natick, Massachusetts: The MathWorks Inc., 2016a.

[377] Jiro Matsumoto. "The Casimir effect as a candidate of dark energy". In: *arXiv preprint arXiv:1303.4067* (12013).

[378] C.E. McCulloch and S.R. Searle. *Generalized, Linear, and Mixed Models*. Wiley Series in Probability and Statistics. Applied Probabil. Wiley, 12004. ISBN: 9780471654049.

[379] Sean R McGuffee and Adrian H Elcock. "Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm". In: *PLoS computational biology* 6.3 (12010), e1000694. DOI: `10.1371/journal.pcbi.1000694`.

[380] Michael D McKay, Richard J Beckman, and William J Conover. "A comparison of three methods for selecting values of input variables in the analysis of output from a computer code". In: *Technometrics* 42.1 (12000), pp. 55–61.

[381] Luke McNally, Sam P Brown, and Andrew L Jackson. "Cooperation and the evolution of intelligence". In: *Proceedings of the Royal Society B: Biological Sciences* 279.1740 (12012), p. 3027. DOI: `10.1098/rspb.2012.0206`.

[382] "Measuring spin correlation between quarks during QCD confinement". In: *Nature* 650.8100 (2026), pp. 65–71. DOI: `10.1038/s41586-025-09920-0`.

[383] Grégoire Méchain et al. "Range of plasma filaments created in air by a multi-terawatt femtosecond laser". In: *Optics Communications* 247.1-3 (12005), pp. 171–180. DOI: `10.1016/j.optcom.2004.11.052`.

[384] T.H.G. Megson. *Structural and Stress Analysis*. EngineeringPro collection. Elsevier Science, 12005. ISBN: 9780080455341.

[385] G. Meissner. *Correlation Risk Modeling and Management: An Applied Guide including the Basel III Correlation Framework - With Interactive Models in Excel / VBA*. Wiley Finance. Wiley, 12013. ISBN: 9781118796894.

[386] Rossoff J Meltzer J and Slaughter J.I. “Structure and Materials Aspects of the Prime Fligh test vehicle”. In: (11965). DOI: 10.2514/6.1966-1738.

[387] Hugo Mercier and Dan Sperber. “Why do humans reason? Arguments for an argumentative theory”. In: *Behavioral and brain sciences* 34.2 (12011), pp. 57–74. DOI: 10.1017/S0140525X10000968.

[388] R Meunier. *Evidence Based Bonne Humeur - À la poursuite de l’esprit critique*. Z & Arts, 12021. ISBN: 9791034386512.

[389] Albert A Michelson and Edward W Morley. “LIX. On a method of making the wavelength of sodium light the actual and practical standard of length”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 24.151 (11887), pp. 463–466. DOI: 10.1080/14786448708628131.

[390] Albert A Michelson and Edward W Morley. “On the relative motion of the Earth and the luminiferous ether”. In: *American journal of science* 3.203 (1887), pp. 333–345. DOI: 10.2475/ajs.s3-34.203.333.

[391] Albert Abraham Michelson. “The Effect of the Earth’s Rotation on the Velocity of Light, I.” In: *Astrophysical Journal, vol. 61, p. 137* 61 (11925), p. 137. DOI: 10.1086/142878.

[392] A.R. Miller, R.M. Brown, and E. Vegh. “New derivation for the rough-surface reflection coefficient and for the distribution of sea- wave elevations”. In: *IEE Proceedings H (Microwaves, Optics and Antennas)* 131 (2 11984), pp. 114–116. DOI: 10.1049/ip-h-1.1984.0022.

[393] K.R. Miller and J.S. Levine. *Miller and Levine Biology*. Pearson foundation series. Pearson, 12012. ISBN: 9780133236385.

[394] PW Milonni and M-L Shih. “Zero-point energy in early quantum theory”. In: *American Journal of Physics* 59.8 (11991), pp. 684–698.

[395] K.A. Milton. *The Casimir Effect: Physical Manifestations of Zero-point Energy*. EBL-Schweitzer. World Scientific, 12001. ISBN: 9789812810526.

[396] Ryo Mizuuchi and Norikazu Ichihashi. “Minimal RNA self-reproduction discovered from a random pool of oligomers”. In: *Chemical Science* (12023).

[397] Douglas C. Montgomery. *Introduction to statistical quality control*. Wiley New York, 12013. ISBN: 9781118146811.

[398] Ekaterina Moreva et al. “The time as an emergent property of quantum mechanics, a synthetic description of a first experimental approach”. In: *Journal of Physics: Conference Series*. Vol. 626. 1. IOP Publishing. 12015, p. 012019. DOI: 10.1088/1742-6596/626/1/012019.

[399] Mathieu Morlighem et al. “Deep glacial troughs and stabilizing ridges unveiled beneath the margins of the Antarctic ice sheet”. In: *Nature geoscience* 13.2 (12020), pp. 132–137. DOI: 10.1038/s41561-019-0510-8.

[400] Peter A Morris et al. “Imaging with a small number of photons”. In: *Nature communications* 6.1 (12015), p. 5913.

[401] Philip M Morse. “Diatomic molecules according to the wave mechanics. II. Vibrational levels”. In: *Physical review* 34.1 (11929), p. 57. DOI: 10.1103/PhysRev.34.57.

[402] Thomas Mueller and Frank Grave. *Catalogue of Spacetimes*. 12010. arXiv: 0904.4184 [gr-qc].

[403] Markus Müller. “Equation of time-problem in astronomy”. In: *Acta Physica Polonica Series A* 88 (11995), S–49.

[404] Jeremy N Munday, Federico Capasso, and V Adrian Parsegian. “Measured long-range repulsive Casimir–Lifshitz forces”. In: *Nature* 457.7226 (12009), pp. 170–173.

[405] C.D. Murray and S.F. Dermott. *Solar System Dynamics*. Cambridge University Press, 11999. ISBN: 9780521575973.

[406] Michael W Nachman and Susan L Crowell. “Estimate of the mutation rate per nucleotide in humans”. In: *Genetics* 156.1 (12000), pp. 297–304.

[407] P.J. Nahin. *In Praise of Simple Physics: The Science and Mathematics behind Everyday Questions*. Princeton Puzzlers. Princeton University Press, 12017. ISBN: 9780691178523.

[408] P.J. Nahin. *Inside Interesting Integrals*. Undergraduate Lecture Notes in Physics. Springer New York, 12014. ISBN: 9781493912773.

[409] P.J. Nahin. *Transients for Electrical Engineers: Elementary Switched-Circuit Analysis in the Time and Laplace Transform Domains (with a touch of MATLAB®)*. Springer International Publishing, 12018. ISBN: 9783319775982.

[410] B. Narins. *The Gale Encyclopedia of Genetic Disorders*. Gale Group, 12022. ISBN: 9780028683928.

[411] A. Navabi and N. Hise. *Why There Is No God: Simple Responses to 20 Common Arguments for the Existence of God*. Atheist Republic, 12014. ISBN: 9781311018045.

[412] Alessandro Nesti et al. “Human sensitivity to vertical self-motion”. In: *Experimental brain research* 232.1 (2014), pp. 303–314. DOI: 10.1007/s00221-013-3741-8.

[413] O. Neugebauer. *Studies in the History of Mathematics and Physical Sciences*. Springer Berlin Heidelberg, 11975. ISBN: 9783642619120.

[414] Is Newton. “De ratione temporis quo grave labitur per rectam data duo puncta conjungentem, ad tempus brevissimum quo, vi gravitatis, transit ab horum uno ad alterum per arcum cycloidis”. In: *Philos. Trans. R. Soc. Lond* 11697.19 (), pp. 424–425.

[415] Chee-Hock Ng and Soong Boon-Hee. *Queueing modelling fundamentals: With applications in communication networks*. John Wiley & Sons, 12008. ISBN: 9780470519578.

[416] Hrvoje Nikolić. “Is zero-point energy physical? A toy model for Casimir-like effect”. In: *Annals of Physics* 383 (12017), pp. 181–195. DOI: 10.1016/j.aop.2017.05.013.

[417] RN Nof et al. “Rising of the lowest place on Earth due to Dead Sea water-level drop: Evidence from SAR interferometry and GPS”. In: *Journal of Geophysical Research: Solid Earth* 117.B5 (12012). DOI: 10.1029/2011JB008961.

[418] C Nolet et al. “Measuring and Modeling the Effect of Surface Moisture on the Spectral Reflectance”. In: *Public Library of Science (PLOS)* (12014). DOI: 10.1371/journal.pone.0112151.

[419] KA Norton. “The propagation of radio waves over the surface of the earth and in the upper atmosphere”. In: *Proceedings of the Institute of Radio Engineers* 24.10 (12006), pp. 1367–1387. DOI: 10.1109/JRPROC.1936.227360.

[420] Werner Nowacki. “Die euklidischen, dreidimensionalen, geschlossenen und offenen Raumformen”. In: *Commentarii Mathematici Helvetici* 7.1 (11934), pp. 81–93. DOI: 10.1007/BF01292710.

[421] Harry Nussbaumer. “Einstein’s conversion from his static to an expanding universe”. In: *The European Physical Journal H* 39.1 (12014), pp. 37–62.

[422] Brendan Nyhan and Jason Reifler. “When corrections fail: The persistence of political misperceptions”. In: *Political Behavior* 32.2 (12010), pp. 303–330.

[423] C. Oliveira. *Practical C++ Financial Programming*. Apress, 12015. ISBN: 9781430267164.

[424] P. Olofsson and M. Andersson. *Probability, Statistics, and Stochastic Processes*. Wiley, 12012. ISBN: 9781118231326.

[425] Jonathan O’Neil et al. “Chapter 16 - The Nuvvuagittuq Greenstone Belt: A Glimpse of Earth’s Earliest Crust”. In: (2019). Ed. by Martin J. Van Kranendonk, Vickie C. Bennett, and J. Elis Hoffmann, pp. 349–374. DOI: 10.1016/B978-0-444-63901-1.00016-2.

[426] Jonathan O’Neil et al. “Neodymium-142 evidence for Hadean mafic crust”. In: *Science* 321.5897 (2008), pp. 1828–1831. DOI: 10.1126/science.116192.

[427] OpenStax. *Calculus*. vol. 2. 12th Media Services, 12016. ISBN: 9781680920154.

[428] S.J. Orfanidis. *Electromagnetic Waves and Antennas*. EDN series for design engineers. ECE Department, 12002.

[429] Huber K. P. and Herzberg G. *Molecular spectra and molecular structure: IV. Constants of diatomic molecules*. Vol. 4. Springer Science & Business Media, 12013. ISBN: 9781475709636. DOI: 10.1007/978-1-4757-0961-2.

[430] Maruthy P. *The compendium of Statistical Tests - A comprehensive collection of 250 tests*. Pannala Series, 12026. ISBN: 9798996151608.

[431] H Park et al. “Aerodynamic performance of a gliding swallowtail butterfly wing model”. In: *Experimental Mechanics* 50.9 (12010), pp. 1313–1321. DOI: 10.1007/s11340-009-9330-x.

[432] R. Parker and B.N. Moore. *LooseLeaf for Critical Thinking*. McGraw-Hill Education, 12016. ISBN: 9781259690877.

[433] Jerome Pearson. “The orbital tower: a spacecraft launcher using the Earth’s rotational energy”. In: *Acta Astronautica* 2.9-10 (11975), pp. 785–799.

[434] Arno A Penzias and Robert W Wilson. “A measurement of excess antenna temperature at 4080 MHz”. In: *A Source Book in Astronomy and Astrophysics, 1900–1975*. Harvard University Press, 11979, pp. 873–876. DOI: 10.1086/148307.

[435] Jose D Perezgonzalez. “Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing”. In: *Frontiers in psychology* (12015), p. 223.

[436] Jose D Perezgonzalez. “The fallacy of placing confidence in confidence intervals-A commentary”. In: (12017).

[437] Leandros Perivolaropoulos and Foteini Skara. “Challenges for ΛCDM: An update”. In: *New Astronomy Reviews* 95 (12022), p. 101659. DOI: 10.48550/arXiv.2105.05208.

[438] M.E. Peskin. *An Introduction To Quantum Field Theory*. CRC Press, 12018. ISBN: 9780429983184.

[439] Gérard Petit and Brian Luzum. *IERS conventions (2010)*. Tech. rep. 12010.

[440] P.H.D.H. *Traité des trois imposteurs: Moïse, Jésus-Christ, Mahomet*. Maxtor France, 12014. ISBN: 9791020801067.

[441] R. Phillips. *Pricing and Revenue Optimization*. Stanford University Press, 12005. ISBN: 9780804781640.

[442] Bùi Tng Phong. "Illumination for Computer-Generated Images (PhD dissertation)". In: (11973).

[443] N. Piskunov. *Calcul différentiel et intégral (Tome I)*. Ellipses, 11993. ISBN: 9782729893408.

[444] N. Piskunov. *Calcul différentiel et intégral (Tome II)*. Ellipses, 11993. ISBN: 9782729893415.

[445] L.P. Pitaevskii and E.M. Lifshitz. *Physical Kinetics*. Course of Theoretical Physics vol. 10. Elsevier Science, 12012. ISBN: 9780080570495.

[446] "Planck 2018 results: VI. Cosmological parameters". In: *Astronomy & Astrophysics* 641 (Sept. 2020), A6. ISSN: 1432-0746. DOI: 10.1051/0004-6361/201833910.

[447] Ptolemy, G.J. Toomer, and O. Gingerich. *Ptolemy's Almagest*. Princeton University Press, 11998. ISBN: 9780691002606.

[448] D.L. Pulfrey. *Understanding Modern Transistors and Diodes*. Cambridge University Press, 12010. ISBN: 9781139484671.

[449] P.P. Puttkammer. "Boundary layer over a flat plate". B.S. thesis. University of Twente, 12013.

[450] T. Pyzdek and P. Keller. *The Six Sigma Handbook, Third Edition*. McGraw-Hill Education, 12009. ISBN: 9780071733908.

[451] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 12017. URL: https://www.R-project.org/.

[452] Simon Rainville et al. "A direct test of E= mc2". In: *Nature* 438.7071 (12005), pp. 1096–1097. DOI: 10.1038/4381096a.

[453] H. Rauchfuss and T.N. Mitchell. *Chemical Evolution and the Origin of Life*. Springer, 12008. ISBN: 9783540788225.

[454] Richard D Ray. "First global observations of third-degree ocean tides". In: *Science Advances* 6.48 (12020), eabd4744. DOI: 10.1126/sciadv.abd4744.

[455] Alexander Raymond et al. "First Very Long Baseline Interferometry Detections at 870 micrometer". In: *The Astronomical Journal* 168.3 (12024). DOI: 10.3847/1538-3881/ad5bdb.

[456] A.A. Rayner. *A first course in biometry for agriculture students*. University of Natal Press, 11969.

[457] J.C.W. Rayner and G.C. Livingston. *An Introduction to Cochran-Mantel-Haenszel Testing and Nonparametric ANOVA*. Wiley Series in Probability and Statistics. Wiley, 12023. ISBN: 9781119832003.

[458] Bruce Cameron Reed. *The Physics of the Manhattan Project*. Springer Berlin Heidelberg, 12014. ISBN: 9783662435335.

[459] Sven Reiche and Z Huang. “Coherent radiation effects in the LCLS undulator”. In: *Proc. FEL 2004, 193* 196.299 (12005), pp. 193–201.

[460] M. Reiher and A. Wolf. *Relativistic Quantum Chemistry: The Fundamental Theory of Molecular Science*. Wiley, 12015. ISBN: 9783527334155.

[461] Sascha Reinhardt et al. “Test of relativistic time dilation with fast optical atomic clocks at different velocities”. In: *Nature Physics* 3.12 (12007), pp. 861–864. DOI: 10.1038/nphys778.

[462] Marc-Olivier Renou et al. “Quantum theory based on real numbers can be experimentally falsified”. In: *Nature* 600.7890 (2021), pp. 625–629. DOI: 10.1038/s41586-021-04160-4.

[463] K.F. Riley, M.P. Hobson, and S.J. Bence. *Mathematical Methods for Physics and Engineering: A Comprehensive Guide*. Mathematical Methods for Physics and Engineering: A Comprehensive Guide. Cambridge University Press, 12006. ISBN: 9780521679732.

[464] C. Robert and G. Casella. *Introducing Monte Carlo Methods with R*. Use R! Springer, 12010. ISBN: 9781441915757.

[465] Richard Robinett. *Quantum mechanics: classical results, modern systems, and visualized examples*. Oxford University Press New York, 12006. ISBN: 9780198530978.

[466] Rene RJ Rohr. *Sundials: History, theory, and practice*. Courier Corporation, 12012.

[467] P.A. Ronveaux et al. *Heun’s Differential Equations*. Oxford science publications. Oxford University Press, 11995. ISBN: 9780198596950.

[468] Bruno Rossi and David B Hall. “Variation of the rate of decay of mesotrons with momentum”. In: *Physical Review* 59.3 (11941), p. 223. DOI: 10.1103/PhysRev.59.223.

[469] François Rothen. *Physique générale: la physique des sciences de la nature et de la vie*. PPUR presses polytechniques, 11999. ISBN: 9782880743963.

[470] Sebastian Ruder. “An overview of gradient descent optimization algorithms”. In: *arXiv preprint arXiv:1609.04747* (12016).

[471] Ernest Rutherford. “LXXIX. The scattering of α and β particles by matter and the structure of the atom”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 21.125 (11911), pp. 669–688. DOI: 10.1080/14786440508637080.

[472] J Rutkowski, L Wojtczak, and Š Zajac. “Pressure influence on the Curie temperature”. In: *Acta Physica Polonica A* 118.5 (12010), pp. 745–746.

[473] L. Ryder. *Introduction to General Relativity*. Cambridge University Press, 12009. ISBN: 9780521845632.

[474] Sara L Rynes, Amy E Colbert, and Kenneth G Brown. “HR professionals’ beliefs about effective human resource practices: Correspondence between research and practice”. In: *Human Resource Management: Published in Cooperation with the School of Business Administration, The University of Michigan and in alliance with the Society of Human Resources Management* 41.2 (12002), pp. 149–174.

[475] R Sagot. “La Visibilité Simultanée du Soleil et de la Lune Éclipsée”. In: *L’Astronomie, vol. 61, pp. 86-92* 61 (1947), pp. 86–92.

[476] E Samain et al. “Millimetric lunar laser ranging at OCA (Observatoire de la Côte d’Azur)”. In: *Astronomy and Astrophysics Supplement Series* 130.2 (11998), pp. 235–244. DOI: 10.1051/aas:199822.

[477] Abdourahmane Sarr and Tonny Lybek. “Measuring liquidity in financial markets”. In: *Available at SSRN 880932* (12002).

[478] Nicolas Sarzeaud. “A New Document on the Appearance of the Shroud of Turin from Nicole Oresme: Fighting False Relics and False Rumours in the Fourteenth Century”. In: *Journal of Medieval History* (12025), pp. 1–18. DOI: 10.1080/03044181.2025.2546884.

[479] Yutaka Sasaki et al. “The truth of the F-measure”. In: *Teach Tutor mater* 1.5 (12007), pp. 1–5.

[480] Till Sawala. “The local galaxy distribution does not violate the cosmological principle”. In: *arXiv preprint arXiv:2607.01172* (12026). DOI: 10.48550/arXiv.2607.01172.

[481] Walter Schachermayer and Josef Teichmann. “How close are the option pricing formulas of Bachelier and Black–Merton–Scholes?” In: *Mathematical Finance* 18.1 (12008), pp. 155–170.

[482] Stephan Schlamminger et al. “Test of the equivalence principle using a rotating torsion balance”. In: *Physical Review Letters* 100.4 (12008), p. 041101. DOI: 10.48550/arXiv.0712.0607.

[483] Marit Schoonhoven and Ronald JMM Does. “The X-bar control chart under non-normality”. In: *Quality and Reliability Engineering International* 26.2 (12010), pp. 167–176.

[484] K Ulrich Schreiber et al. “Variations in the Earth’s rotation rate measured with a ring laser interferometer”. In: *Nature Photonics* 17.12 (12023), pp. 1054–1058. DOI: 10.1038/s41566-023-01286-x.

[485] Thomas Schürmann, Ingo Hoffmann, and Winfrid Görlich. “On the experimental verification of the uncertainty principle of position and momentum”. In: *Physics Letters A* 470 (12023), p. 128787. DOI: 10.48550/arXiv.2211.14724.

[486] M. Scutari and J.B. Denis. *Bayesian Networks: With Examples in R*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 12014. ISBN: 9781482225587.

[487] Michael Segal. “Numerical discrepancies in the list of vessels in Ezra i 9-11”. In: *Vetus Testamentum* 52.Fasc. 1 (12002), pp. 122–129. DOI: 10.1163/15685330252965776.

[488] Antonia Seifert et al. “Supernovae evidence for foundational change to cosmological models”. In: *Monthly Notices of the Royal Astronomical Society: Letters* 537.1 (2025), pp. L55–L60. DOI: doi.org/10.1093/mnrasl/slae112.

[489] Andrew Sepielli. “Moral uncertainty and the principle of equity among moral theories”. In: *Philosophy and Phenomenological Research* 86.3 (12013), pp. 580–589. DOI: 10.1111/j.1933-1592.2011.00554.x.

[490] EMC Education Services. *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. Wiley, 12015. ISBN: 9781118876138.

[491] R.A. Serway and J.W. Jewett. *Physics for Scientists and Engineers with Modern Physics*. Cengage Learning, 12018. ISBN: 9781337671729.

[492] Yong Shao et al. “Phylogenomic analyses provide insights into primate evolution”. In: *Science* 380.6648 (12023), pp. 913–924. DOI: 10.1126/science.abn6919.

[493] Valery Shcherbakov. “Why the 46° halo is seen far less often than the 22° halo?” In: *Journal of Quantitative Spectroscopy and Radiative Transfer* 124 (12013), pp. 37–44.

[494] I Simbotin, MJ Jamieson, and A Dalgarno. “Rayleigh scattering cross sections of O2 at wavelengths longer than 180 nm”. In: *Journal of Geophysical Research: Atmospheres* 109.D13 (12004).

[495] Ashok K Singal. “Discrepancy between power radiated and the power loss due to radiation reaction for an accelerated charge”. In: *Symmetry* 12.11 (12020), p. 1833.

[496] Ashok K Singal. “Why no radiation occurs in the case of a uniformly accelerated charge”. In: *arXiv preprint arXiv:1810.04519* (12018).

[497] R. Paul Singh and Dennis R. Heldman. *Introduction to food engineering*. Gulf Professional Publishing, 12001. ISBN: 9780080574493.

[498] J.M. Smith. *Evolution and the Theory of Games*. Cambridge University Press, 11982. ISBN: 9780521288842.

[499] Quentin Smith. “The uncaused beginning of the universe”. In: *Philosophy of Science* 55.1 (11988), pp. 39–57. DOI: 10.1086/289415.

[500] Richard Smith. “Doctors are not scientists”. In: *Bmj* 328.7454 (12004).

[501] Robert W Smith. “The origins of the velocity-distance relation”. In: *Journal for the History of Astronomy* 10.3 (11979), pp. 133–165.

[502] Igor I Smolyaninov et al. “Experimental demonstration of metamaterial “multiverse” in a ferrofluid”. In: *Optics express* 21.12 (2013), pp. 14918–14925. DOI: 10.1364/OE.21.014918.

[503] J.H. Sobel. *Logic and Theism: Arguments for and against Beliefs in God*. Cambridge University Press, 12003. ISBN: 9781139449984.

[504] Joram Soch et al. *StatProofBook/StatProofBook.github.io: StatProofBook 12023*. Zenodo, Jan. 12024. DOI: 10.5281/zenodo.10495684.

[505] A. Sokal and J. Bricmont. *Fashionable Nonsense: Postmodern Intellectuals’ Abuse of Science*. ST MARTINS PRESS (NY), 11999. ISBN: 9780312204075.

[506] Arnold Sommerfeld. “Zur quantentheorie der spektrallinien”. In: *Annalen der Physik* 356.17 (11916), pp. 1–94. DOI: 10.1080/14786440508637080.

[507] Chun Siong Soon et al. “Unconscious determinants of free decisions in the human brain”. In: *Nature neuroscience* 11.5 (12008), pp. 543–545. DOI: 10.1038/nn.2112.

[508] H. Spieler. *Semiconductor Detector Systems*. Series on Semiconductor Science and Technology. OUP Oxford, 12005. ISBN: 9780191523656.

[509] N.L. Squeglia. *Zero Acceptance Number Sampling Plans*. ASQ Quality Press, 12008. ISBN: 9780873897396.

[510] SV Starchenko and CA Jones. “Typical velocities and magnetic field strengths in planetary interiors”. In: *Icarus* 157.2 (12002), pp. 426–435.

[511] V.J. Stenger. *The Fallacy of Fine-Tuning: Why the Universe Is Not Designed for Us*. Prometheus Books, 12011. ISBN: 9781616144449.

[512] V.J. Stenger and C. Hitchens. *God: The Failed Hypothesis: How Science Shows That God Does Not Exist*. Prometheus Books, 12010. ISBN: 9781615920037.

[513] H. Stephani. *Relativity: An Introduction to Special and General Relativity*. Cambridge University Press, 12004. ISBN: 9780521010696.

[514] James Stewart. *Analyse, concepts et contextes: Volume 1, Fonctions d'une variable*. Vol. 1. De Boeck Supérieur, 12006. ISBN: 9782804163068.

[515] Rachael E. Stewart and E. Lower Marcus et al. Dinh Thi Hao Geroge Younes. "Vacuum birefringence and the polarized X-ray emission from a radio magnetar". In: *Nature* (2026). DOI: `10.1038/s41586-026-10859-z`.

[516] Aneta S Stodolna et al. "Hydrogen atoms under magnification: Direct observation of the nodal structure of Stark states". In: *Physical Review Letters* 110.21 (12013), p. 213001.

[517] George Johnstone Stoney. "On the physical constitution of the sun and stars". In: *Proceedings of the Royal Society of London* 17 (11869), pp. 1–57. DOI: `10.1098/rspl.1868.0001`.

[518] G. Strang et al. *Calculus*. vol. 1. OpenStax College, 12016. ISBN: 9781938168024.

[519] David L Streiner. "Unicorns do exist: A tutorial on "proving" the null hypothesis". In: *The Canadian Journal of Psychiatry* 48.11 (12003), pp. 756–761.

[520] J. Strickland. *Predictive Analytics using R*. LULU Press, 12015. ISBN: 9781312841017.

[521] Andrew Strominger and Cumrun Vafa. "Microscopic origin of the Bekenstein-Hawking entropy". In: *Physics Letters B* 379.1 (11996), pp. 99–104. ISSN: 0370-2693. DOI: `10.1016/0370-2693(96)00345-0`.

[522] Francesco Sylos Labini and Marco Galoppo. "Detection of anisotropic cosmic structures on a gigaparsec scale". In: *Nature* (12026), pp. 1–5. DOI: `10.1038/s41586-026-10702-5`.

[523] S.M. Sze and M.K. Lee. *Semiconductor Devices: Physics and Technology*. Semiconductor Devices, Physics and Technology. Wiley, 12012. ISBN: 9780470537947.

[524] Marco Taboga. *Lectures on probability theory and mathematical statistics*. CreateSpace Independent Publishing Platform, 12017.

[525] Masao Takamoto et al. "Test of general relativity by a pair of transportable optical lattice clocks". In: *Nature photonics* 14.7 (12020), pp. 411–415. DOI: `10.1038/s41566-020-0619-8`.

[526] Masaharu Tanabashi et al. "Review of Particle Physics: particle data groups". In: *Physical Review D* 98.3 (12018), pp. 1–1898. DOI: `10.1103/PhysRevD.98.030001`.

[527] K. Tapp. *Differential Geometry of Curves and Surfaces*. Undergraduate Texts in Mathematics. Springer International Publishing, 12016. ISBN: 9783319397993.

[528] D. Tavella. *Quantitative Methods in Derivatives Pricing: An Introduction to Computational Finance*. Wiley Finance. Wiley, 12003. ISBN: 9780471274797.

[529] RE Taylor and John Southon. "Use of natural diamonds to monitor 14C AMS instrument backgrounds". In: *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms* 259.1 (12007), pp. 282–287. DOI: `10.1016/j.nimb.2007.01.239`.

[530] G. Tcherkez. *Isotopie biologique: introduction aux effets isotopiques et à leurs applications en biologie*. Éd. Tec & Doc, 12010. ISBN: 9782743012670.

[531] M. Tegmark. *Our Mathematical Universe: My Quest for the Ultimate Nature of Reality*. Alfred A. Knopf, 12014. ISBN: 9780307599803.

[532] E. Thackery and M. Harris. *The Gale Encyclopedia of Mental Disorders*. Gale Group, 12003. ISBN: 9780787657680.

[533] "The Delphi technique as a forecasting tool: issues and analysis". In: *International Journal of Forecasting* 15.4 (11999), pp. 353–375. ISSN: 0169-2070. DOI: 10.1016/S0169-2070(99)00018-7.

[534] R. Théoret. *Le Calcul Numérique en Finance Empirique et Quantitative: Ingénierie Financière et Excel*. Presses de l'Universite du Quebec, 12004. ISBN: 9782760517721.

[535] Daniel Thiel. *Recherche opérationnelle et management des entreprises*. Economica, 11990. ISBN: 9782717819304.

[536] Hervé Thiriez. *La modélisation du risque: simulations de Monte Carlo*. Economica, 12004. ISBN: 9782717848229.

[537] H.C. Thode. *Testing For Normality*. Statistics, textbooks and monographs. CRC Press, 12002. ISBN: 9780203910894.

[538] Joseph John Thomson. *Electricity and Matter*. Mrs. Hepsa Ely Silliman Memorial Lectures. Yale University. New York: Charles Scribner's Sons, 11904. DOI: 10.1038/070073a0.

[539] Joseph John Thomson. "On the structure of atom: an investigation of the stability and periods of oscillation of a number of corpuscles arranged at equal intervals around the circumference of a circle; with applications of the results to the theory of atomic structure". In: *Phil. Mag., ser. 6* 7 (11904), pp. 237–265.

[540] TH Tizard. "Distance of the Visible Horizon". In: *Nature* 92.2299 (11913), pp. 344–345. DOI: 10.1038/092425d0.

[541] Michael Tomasello and Amrisha Vaish. "Origins of human cooperation and morality". In: *Annual review of psychology* 64.1 (12013), pp. 231–255. DOI: 10.1146/annurev-psych-113011-143812.

[542] Akira Tonomura et al. "Evidence for Aharonov-Bohm effect with magnetic field completely shielded from electron wave". In: *Physical review letters* 56.8 (1986), p. 792. DOI: 10.1103/PhysRevLett.56.792.

[543] Kevin E Trenberth, John T Fasullo, and Jeffrey Kiehl. "Earth's global energy budget". In: *Bulletin of the american meteorological society* 90.3 (2009), pp. 311–324.

[544] Robert L Trivers. "The evolution of reciprocal altruism". In: *The Quarterly review of biology* 46.1 (11971), pp. 35–57. DOI: 10.1086/406755.

[545] R.S. Tsay. *Analysis of Financial Time Series*. CourseSmart. Wiley, 12010. ISBN: 9781118017098.

[546] D.L. Turcotte and G. Schubert. *Geodynamics*. Cambridge University Press, 12002. ISBN: 9781107268678.

[547] J.J. Uicker, G.R. Pennock, and J.E. Shigley. *Theory of Machines and Mechanisms, 7th Edition*. Oxford University Press, 12017. ISBN: 9780195371239.

[548] Stefan Ulmer et al. "Observation of spin flips with a single trapped proton". In: *Physical review letters* 106.25 (12011), p. 253001.

[549] Henryk Deter Ulrich Wiehle Michael Diegelmann, Dr. Peter Noel Schömig, and Michael Rolf. *Les 100 principaux indicateurs financiers*. Cometis, UBS, 12010. ISBN: 9783940828316.

[550] Nathan S Upham, Jacob A Esselstyn, and Walter Jetz. "Inferring the mammal tree: species-level sets of phylogenies for questions in ecology, evolution, and conservation". In: *PLoS biology* 17.12 (12019), e3000494.

[551] Stan Uryasev et al. "VaR vs CVaR in risk management and optimization". In: *CARISMA conference*. 12010.

[552] Jean-Philippe Uzan. "The fundamental constants and their variation: observational and theoretical status". In: *Reviews of modern physics* 75.2 (12003), p. 403. DOI: `10.48550/arXiv.hep-ph/0205340`.

[553] K.A. Valiev. *The Physics of Submicron Lithography*. Microdevices. Springer US, 12012. ISBN: 9781461533184.

[554] Jacques P Vallée. "Recent advances in the determination of some Galactic constants in the Milky Way". In: *Astrophysics and Space Science* 362.4 (12017), p. 79. DOI: `10.1007/s10509-017-3058-3`.

[555] Mihran Vardanyan, Roberto Trotta, and Joseph Silk. "Applications of Bayesian model averaging to the curvature and size of the Universe". In: *Monthly Notices of the Royal Astronomical Society: Letters* 413.1 (12011), pp. L91–L95. DOI: `10.1111/j.1745-3933.2011.01040.x`.

[556] Dan Ventura. "SVM example". In: *Lectures notes, Mar* (12009).

[557] M. Vermot, P. Bovay, and D. Prongué. *Traité de construction horlogère*. Collection microtechnique & [et] nanotechnologie. Presses polytechniques et universitaires romandes, 12011. ISBN: 9782880748838.

[558] Horácio Santana Vieira and Valdir B Bezerra. "Quantum Newtonian cosmology and the biconfluent Heun functions". In: *Journal of Mathematical Physics* 56.9 (12015), p. 092501.

[559] HS Vieira and VB Bezerra. "Class of solutions of the Wheeler-DeWitt equation in the Friedmann-Robertson-Walker universe". In: *Physical Review D* 94.2 (12016), p. 023511.

[560] Alexander Vilenkin. "Creation of universes from nothing". In: *Physics Letters B* 117.1-2 (11982), pp. 25–28. DOI: `10.1016/0370-2693(82)90866-8`.

[561] Alexander Vilenkin. "The beginning of the universe". In: *The Kalām Cosmological Argument: Scientific Evidence fort he Beginnig of the Universe, Haz. Paul Copan, William Lane Craig, içinde* (2017), pp. 150–158.

[562] Jill Vincent. "The mathematics of sundials". In: *Australian Senior Mathematics Journal* 22.1 (12008), pp. 13–23.

[563] Bartel Leendert Van der Waerden. "The heliocentric system in Greek, Persian and Hindu astronomy". In: *Annals of the New York Academy of Sciences* 500.1 (11987), pp. 525–545. DOI: `10.1111/j.1749-6632.1987.tb37224.x`.

[564] Christian Wagner, Simon Miller, and Jonathan M. Garibaldi. "A fuzzy toolbox for the R programming language". In: *Fuzzy Systems (FUZZ), 2011 IEEE International Conference on*. IEEE. 12011, pp. 1185–1192.

[565] James R Wait. “The ancient and modern history of EM ground-wave propagation”. In: *IEEE Antennas and Propagation Magazine* 40.5 (11998), pp. 7–24. DOI: 10.1109/74.735961.

[566] Merle F Walker. “The effects of urban lighting on the brightness of the night sky.” In: *Publications of the Astronomical Society of the Pacific* 89.529 (11977), p. 405. DOI: 10.1086/130142.

[567] U. Walter. *Astronautics: The Physics of Space Flight*. Springer International Publishing, 12019. ISBN: 9783319743738.

[568] M. Ware and J. Peatross. *Physics of Light and Optics (Black & White)*. Brigham Young University, Department of Physics, 12015. ISBN: 9781312929272.

[569] Brian P Weaver et al. “A Bayesian approach to the analysis of gauge R&R data”. In: *Quality Engineering* 24.4 (12012), pp. 486–500. DOI: 10.1080/08982112.2012.702381.

[570] I. Welch. *Corporate Finance: An Introduction*. Pearson international edition. Prentice Hall, 12009. ISBN: 9780321587275.

[571] A.N. Whitehead and B. Russell. *Principia Mathematica*. Rough Draft Printing, 11910. ISBN: 9781603864374.

[572] Wessel N. van Wieringen. *Lecture notes on ridge regression*. 12015. eprint: arXiv:1509.09169.

[573] Rand R Wilcox. *Introduction to robust estimation and hypothesis testing, 4th Edition*. Academic press, 12017. ISBN: 9780128047330.

[574] Michael R Wilczynska et al. “Four direct measurements of the fine-structure constant 13 billion years ago”. In: *Science advances* 6.17 (12020), eaay9672.

[575] William T. Vetterling William H. Press Saul A. Teukolsky et al. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, 12007. ISBN: 9780521880688.

[576] George E Williams. “Precambrian length of day and the validity of tidal rhythmite paleotidal values”. In: *Geophysical Research Letters* 24.4 (11997), pp. 421–424. DOI: 10.1029/97GL00234.

[577] William SC. Williams. *Introducing Special Relativity*. CRC Press, 12002. ISBN: 9780415277624.

[578] Paul Wilmott. *The mathematics of financial derivatives: a student introduction*. Cambridge University Press, 11995. ISBN: 9780521497893.

[579] Jens Winkler and Roger Moser. “Biases in future-oriented Delphi studies: a cognitive perspective”. In: *Technological forecasting and social change* 105 (12016), pp. 63–76.

[580] Viktor Witkovsky. “On the Exact Two-Sided Tolerance Intervals for Univariate Normal Distribution and Linear Regression”. In: *Austrian Journal of Statistics* 43.4 (12014), p. 279. DOI: 10.17713/ajs.v43i4.46.

[581] Edward L Wright. “Theoretical overview of cosmic microwave background anisotropy”. In: *arXiv preprint astro-ph/0305591* (12003).

[582] Mao Xin. “The VIX Volatility Index”. In: *Project Report, Department of Mathematics, Uppsala University, Sweden* (12011).

[583] Jaswant K Yadav, JS Bagla, and Nishikanta Khandai. “Fractal dimension as a measure of the scale of homogeneity”. In: *Monthly Notices of the Royal Astronomical Society* 405.3 (12010), pp. 2009–2015.

[584] Songqinghao Yang, Haomu Yuan, and Crispin HW Barnes. “Experimental Demonstration of the PBR Test on a Superconducting Processor”. In: *arXiv preprint arXiv:2510.11213* (12025). DOI: `10.48550/arXiv.2510.11213`.

[585] Yi-Bo Yang et al. “Proton mass decomposition from the QCD energy momentum tensor”. In: *Physical review letters* 121.21 (2018), p. 212001. DOI: `10.1103/PhysRevLett.121.212001`.

[586] M. Yanoff, J.S. Duker, and J.J. Augsburger. *Ophthalmology*. An expert consult title online + print. Mosby Elsevier, 12009. ISBN: 9780323043328.

[587] Ramu Yerukala and Naveen Kumar Boiroju. “Approximations to standard normal distribution function”. In: *International Journal of Scientific & Engineering Research* 6.4 (12015), pp. 515–518.

[588] PT Younger et al. “The sporadic radiant and distribution of meteors in the atmosphere as observed by VHF radar at Arctic, Antarctic and equatorial latitudes”. In: *Annales Geophysicae*. Vol. 27. 7. Copernicus Publications Göttingen, Germany. 12009, pp. 2831–2841. DOI: `10.5194/angeo-27-2831-2009`.

[589] Karen K Yuen. “The two-sample trimmed t for unequal population variances”. In: *Biometrika* 61.1 (11974), pp. 165–170.

[590] M. Zabrocki. *An introduction to ordinary generating functions*. York University. URL: `http://garsia.math.yorku.ca/~zabrocki/MMM1/MMM1Intro2OGFs.pdf`.

[591] Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. *Social media mining: an introduction*. Cambridge University Press, 12014. ISBN: 9781107018853.

[592] M.J. Zaki and W. Meira. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, 12014. ISBN: 9780521766333.

[593] Wolfgang H Zangemeister, Thilo Gronow, and Ulrich Grzyska. “Pupillary responses to single and sinusoidal light stimuli in diabetic patients”. In: *Neurology international* 1.1 (12009), e19. DOI: `10.4081/ni.2009.e19`.

[594] Anatolii Vladimirovich Zasov et al. “Dark matter in galaxies”. In: *Physics-Uspekhi* 60.1 (12017), p. 3.

[595] Wilford Zdunkowski and Andreas Bott. *Dynamics of the Atmosphere: A course in theoretical Meteorology*. Cambridge University Press, 12003. ISBN: 9780521809498.

[596] Alessandro Zenesini et al. “False vacuum decay via bubble formation in ferromagnetic superfluids”. In: *Nature Physics* (12024), pp. 1–6. DOI: `10.1038/s41567-023-02345-4`.

[597] Dong Zhang, Akram Hasani Zonoozi, and Pavel Kroupa. “Revisiting the missing mass problem in MOND for nearby galaxy clusters”. In: *Physical Review D* 113.4 (12026), p. 043027. DOI: `10.1103/mp3f-q5dc`.

[598] Yixin Zhang et al. “Kinematic and aerodynamic investigation of the butterfly in forward free flight for the butterfly-inspired flapping wing air vehicle”. In: *Applied Sciences* 11.6 (12021), p. 2620. DOI: `10.3390/app11062620`.

[599] Youxue Zhang. "The age and accretion of the Earth". In: *Earth-Science Reviews* 59.1 (12002), pp. 235–263. ISSN: 0012-8252. DOI: 10.1016/S0012-8252(02)00077-6.

[600] M.V. Zombeck. *Handbook of Space Astronomy and Astrophysics*. Cambridge University Press, 12006. ISBN: 9781139459501. DOI: 10.1017/CBO9780511536359.

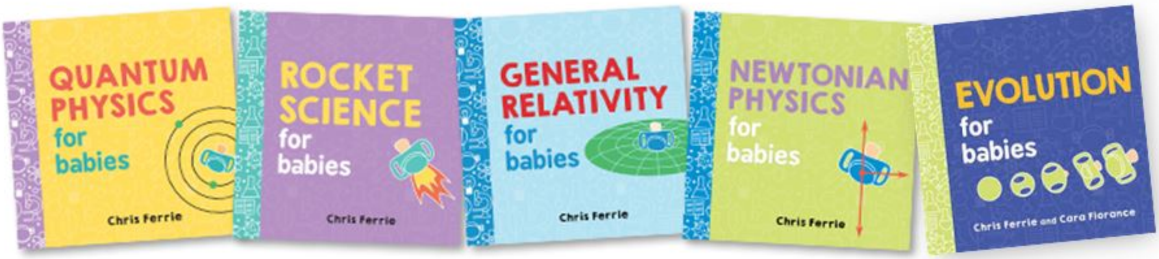
[601] Walter Zucchini, A. Berzel, and O. Nenadic. "Applied smoothing techniques". In: *Part I: Kernel Density Estimation* 15 (12003).

[602] Erik W van Zwet and Eric A Cator. "The significance filter, the winner's curse and the need to shrink". In: *Statistica Neerlandica* 75.4 (12021), pp. 437–452.

[603] Barton Zwiebach. *A first course in string theory*. Cambridge university press, 12004. ISBN: 9780521831437.

+ a thousand of low quality and incomplete scientific papers that do not deserve to be cited.

Furthermore... for the parents that want to educate their children in a healthy intellectual environment, I strongly recommend the following books for babies by Chris Ferrie:


QUANTUM PHYSICS for babies
Chris Ferrie
ROCKET SCIENCE for babies
Chris Ferrie
GENERAL RELATIVITY for babies
Chris Ferrie
NEWTONIAN PHYSICS for babies
Chris Ferrie
EVOLUTION for babies
Chris Ferrie and Cara Florance

and the following books for teenagers:

THE MANGA GUIDE TO
RELATIVITY
THE MANGA GUIDE TO
STATISTICS
THE MANGA GUIDE TO
LINEAR ALGEBRA
CALC
DATA
ELEC
THE MANGA GUIDE TO
PHYSICS
THE MANGA GUIDE TO
MOLECULAR BIOLOGY
GUIDE TO
HEMISTRY

Index

Symbols

(ε, δ) continuity 2231
(ε, δ) definition of limit 2227
22° halo 4385
A/B test 991
ABC analysis 8772
C-algebra A 433
F-test of Overall Significance in Regression Analysis 6337
F_1 score 6482
G–test of goodness-of-fit 1117
G^2-test 1117
K-fold cross validation method 6429
K-medoids 6866
$K-\varepsilon$ principle 2268
L-estimator 1156
L-statistic 1156, 6562
LU decomposition 2024
M-Estimator of location 6557
QR decomposition 2030
T-test for coefficient slope 6359
V-structure 507
XYZ analysis 8777
Y-Pauli gate 7290
Z theorem 2618
Z-Pauli quantum gate 7290
Z-factor 696
Z-score formula for predicting bankruptcy 7889
Z-test 947
ΛCDM 5987
α-equivalent formulas 175
β risk 955
β^+ radioactivity 5087
β^- radioactivity 5084
χ^2 feature selection 1117
$\lambda\varphi^4$ theory 5183
$\mathbb{C}$-linear function 2391
ϕ-divergence 3548
σ-additivity 478
k-convergent 372
k-means 6854
k-nearest neighbours 6832
k-reduced 372
k-regular graph 3008
k-th partial quotient 372
n order Bessel's function 1768
n-body equations of motion 5251
n-step pattern 7225
n-th primorial 221
p-hacking 956
p-list 524
p-value 948, 950, 952
x-axis 2179
y-axis 2179
z-plane 2362
(probabilistic) wave function 4691
(self) coupling term 5183
(self) interaction term 5183
2D Fourier analysis 1763

A

a posteriori probability 485
a posteriori test power 948
a priori linear model 6330

a priori probability 476, 485
A-optimal design 8866
a-theoric 946
A/B test for means 992
ABC method 8098
abduction 891
Abel’s Theorem 1635
Abel-Ruffini theorem 1445
abelian group 421, 3038
abelian monoid 419
abnormal Zeeman effect 4669
Abraham form 4156
Abraham-Becker radiation damping force 4191
Abraham-Lorentz force 4201
abscissa 2179
absolute convergence test 1778
absolute differential geometry 2038
absolute electrical permittivity 3971
absolute error 1308
absolute magnetic permeability 4013
absolute magnitude 5586
absolute refractive index 4358
absolute uncertainty 1308
absolute value 324
absolute value at risk 7865
absolute value function 2216
absolute zero 3620
absorbed dose 5130
absorbing element 417
absorbing Markov chain 541
absorbing state 7456
absorption factor 4351
absorption line 5542
absorption spectrum 4637
absorptions laws 397
abstract number 280, 1350
accelerated life test plans 9049
acceleration 3231
acceleration equation 5896
acceleration transformation formula 3387
acceptable quality level 8909
acceptance angle 4607
acceptance number 8927
acceptance test 8904
accidental error 1307
accrued interest 7629
accumulation point 7047
accuracy 6482, 6810
acid-test 8029
acoustic 8195
action 3161
action (analytical mechanics) 3161
action / reaction principle 3209
action integral 3161, 4628
action risk 8082
action-at-a-distance forces 3201
activation function 6954
activation level of a neuron 6954
activation threshold of a neuron 6954
active dipole 4219
active power 8454
active redundancy 8753
active variables 3072
activity (nuclear) 5033
actual mechanical advantage 8360
actual value 7548
actualization factor 8055
actualized cash flow 8055
actuarial calculations 7558
actuator disk 3865
acute angle 2584
acyclic 3017
ADALINE 6985
Adaptive Linear Neuron 6985
adaptive sampling 1331
added values 7509
addition 296, 421
additional angles 2585
additive model 7899
additive primary colors 4326
additive synthesis of colors 4326
adequation test 944
adiabatic approximation 6193
adiabatic coefficient 3885, 8200
adiabatic index 3647, 8200
adiabatic system 3602
adjacency matrix 3026
adjacent 3005
adjacent angle 2576
adjacent side 2508
adjoint matrix 1953
adjoint operator 4715
adjoint spinor 5178
adjoint tensor 2090
adjugate 1958

adjunct . . . 1958
adjusted beta . . . 7716
adjusted book value . . . 8063
adjusted net worth . . . 8064
adjusted R-squared . . . 6355
advection part . . . 7806
advection term . . . 3208, 3830
adverse selection . . . 8174
aerodynamic force . . . 8591
affine approximation . . . 1697
affine connection . . . 2103, 2113, 5750
affine function . . . 1379, 2206, 2215
affine group . . . 1422
affine model . . . 8791
affine projection . . . 2712
affine space . . . 1815
affine tangent approximations . . . 1697
affix . . . 249
age of the universe . . . 5916
aggregates . . . 7508
AGNES . . . 6896
Agrippa trilemma . . . 3128
Aharonov–Bohm effect . . . 4121
Airy disk . . . 4527
Akaike information criterion . . . 6420
Al-Kashi's formula . . . 2525
albedo . . . 4351
Alembertian . . . 4118
alert limits . . . 8939
Alexander polynomial . . . 3047
algebraic equations . . . 1351
algebraic function . . . 2222
algebraic integer of degree n . . . 276
algebraic language . . . 7195
algebraic number of degree n . . . 276
algebraic univariate polynomial . . . 1368
algorithm . . . 6273, 7181, 8637
algorithm cost . . . 6273
algorithm of Heron of Alexandria . . . 6295
alias . . . 8816
alias generator . . . 8818
alias structure . . . 8816
aliases . . . 8812
alignment accuracy residuals . . . 6947
aliquot part . . . 2570
alleles . . . 7361
alpha coefficient . . . 7717
alpha radioactivity . . . 5075
alphabet . . . 7198
alternate angles . . . 2579
alternated multi-linear application . . . 1962
alternating group of order n . . . 1445
alternating series . . . 1780
alternating zeta function . . . 1687
alternative . . . 155
alternative hypothesis . . . 945
alternative vote . . . 7474
alternator . . . 8447
altitude . . . 2607, 5438
Amagat's law . . . 3925
american option . . . 7666
American Wire Gauge . . . 4236
Amontons' first law of friction . . . 3454
Amontons' second law of friction . . . 3454
amortization factor . . . 8483
amortization time . . . 8059
amphicheiral . . . 3057
amplifier . . . 8536
amplitude modulation . . . 3518
amplitude modulation broadcasting . . . 3523
amplitude of a wave . . . 3476
amplitude of probability . . . 4693
amplitude translation . . . 259
Ampère's circuital law . . . 4011
Ampère's law . . . 4011, 4012
Ampère's observer . . . 4010
Ampère's theorem . . . 4011, 4012
analemma . . . 5404
analysis of covariance and variance . . . 1005
analysis of variance . . . 1001
analytic continuation . . . 2421
analytic expression . . . 2206
analytic geometry . . . 2764
analytic hierarchy process . . . 7462
analytical conics . . . 2765
analytical function . . . 2400
ancestors . . . 7025
anchorism . . . 7388
ANCOVA . . . 1005
AND connector . . . 146, 147
Anderson-Darling statistic . . . 1131
angle . . . 2555, 2575
angle between two segments . . . 2576
angle between two vector . . . 1818
angle bisector . . . 2607
angle measurement . . . 2583

angle of attack 8591
angle of deformation 3749
angle of incidence 4354
angle of reach 3267
angle of refraction 4354
angle of view 4451
angle sum theorem 2593
angle-side-angle theorem 2595
angular acceleration 3282
angular frequency 3511
angular impulsion 3247
angular mass 3297
angular momentum 3244
angular plasma frequency 3961
anisotropic 6231
anisotropic source 4339
annihilation operator 4811
annual percentage rate 7559
annualized historical volatility 7741
annualized realized volatility 7741
annualized yield rate 7552
annuities 7571
annuity 7562
anomalous magnetic moment 4999
ANOVA
  analysis of variance 1004
  ANOVA equation 6330
  balanced incomplete block design 1057
  block factor 1040
  blocking factor 1055
  both factors locally in interaction 1035
  both factors locally significant . . . 1033
  complete random design 1004
  complete randomized design 1054
  general randomized block design 1057
  Graeco-Latin Square 1082
  hierarchical fixed factor ANOVA 1065
  incomplete block design 1056
  interacting factors 1027
  Latin Square 1074
  mixed ANOVA 1004
  nested ANOVA with fixed factors 1065
  non-interacting factors 1027
  non-locally interaction 1029
  one factor ANAVAR 1004
  one factor ANOVA 1004
  one fixed categorical variable with repetition 1004
  one fixed factor ANOVA with repetitions 1004
  one-way ANOVA 1004
  randomized block design . . 1040, 1055
  randomized complete block design 1027, 1055
  treatment factor 1040
  two fixed categorical variables . . . 1027
  two-way ANOVA 1027
  Type I 1064
  Type II 1064
  Type III 1064
  variance analysis equation 6330
antagonistic games 7379
antecedent 283, 403
anti-bonding function 6198
anti-selection problem 8174
anti-symmetrical relation 290
anticommutation relations 4697
anticommutator 4718
anticyclone 8299
antiderivative 1512
antilogy 143
antineutrino 5084
antinodes of the vibration 3484
antiparticle 4923
antiparticle solution 4950
antisymmetric matrix 1955
antisymmetric tensor 2088
antisymmetry tensor 2045
antithetic variables 6726
apex 2900, 2931, 2963
apocenter 2785
apogee 2786, 5349
aporia 102
apparent brightness 5577
apparent diameter 4316
apparent magnitude 5580
apparent power 8454
apparent singularity 2438, 5794
Apparent Solar Time 5461
apparent solar time 5406
application 403
application point (force) 3201
applied forces 3199
applied science 95
approached stigmatism 4420
apriori 6815

apsidal circle 5238
apsidal precession 5361, 5800
Arago spot 4521
arbitrary gauge 4118
arc length 1594
arc-elasticity 7491
Archimedean integer 333
Archimedean property 333
Archimedean Spiral 2818
Archimedes' axiom 333, 2633
Archimedes' force 3875
Archimedes' law 3876
area law 5234
area under the curve (AUC) 6520
ARFIMA 7997
Argand-Cauchy plane 244
argument 251
argument of ignorance 104
arguments 2171
arithmetic average 546
arithmetic average speed 3226
arithmetic Brownian motion 7765
arithmetic declining depreciation 8071
arithmetic polynomial 323
arithmetic return on investment 7547
arithmetic sequence 1653
arity 169
arrangement 523
arrow 1795
arrow (mechanical engineering) 8391
arrow diagram 378
arrow of time 5638
as bad as old 8724
as good as new 8724
asian option 7667
assembly 7184
assertion 143
asset utilization ratios 8033
asset's replacement value 8064
assignable variations 8940
associated de Broglie wavelength 4749
associated Laguerre ordinary differential
equation 1413
associated Laguerre polynomials 1411
associated Legendre polynomials 6140
associated molar size 6254
associated subset 144
associates interlaces 3059
associative law 417
associative magma 418
associative operation 297, 309
assumption 100
Aston curve 5058
astronomical twilight 5219
astronomy 5217
asymmetric public key 7218
asymptote 2269
asymptotic normality 938
asymptotic notation 6285
asymptotically unbiased estimator 822
asynchronous logic model 7104
at level 5260
at-the-money 7665
athermic 6250
atom 6230
atomic formulas 173
atomic mass units 5026
atomic molar mass 6253
atomic number 5024
attractor 7323, 8338
attractor of the IFS 7041
attribute control 8905
attribute control chart 8936
attribute variable 580
Atwood's machine 8543
Aufbau principle 4681
Auger cascade 5091
Auger effect 5090
Auger electron 5090
augmented matrix associated of a system
1941
authalic projection 2744
auto-correlation 2314
autocorrelation coefficient 7977
autocovariance function 7976
automata 538, 7197
automated process control 8674
automorphic 3007
automorphism 443
automorphism of graph 3025
autonomous system 1636
AutoRegressive Integrated Moving
Average (ARIMA) 7989
autoregressive process of order p 7985
autosuccessor set 384
auxetic materials 8384

auxiliary circle 5238
availability rate 8681
average 546
average absolute deviation 7903
average bolometric emittance 5568
average complexity 6284
average cost function 7502
average failure rate 8734
average linear acceleration 3231
average linear speed 3226
average outgoing quality 8931
average outgoing quality limit 8933
average power 3380
average run length 8976
average slope 1450
average squared residuals 7903
average thermal velocity 3558
Avogadro and Ampere law 3893
Avogadro's law 3893
axial load 8409
axial symmetry 2659
axiom 100
axiom of recurrence 218
axioms of the definition 7106
azimuth 5438
azimuthal acceleration 3388
azimuthal coordinates 5465
azimuthal normal aspect projection . . 2729
azimuthal quantum number 4646

B
Babylonian method 6295
Bachelier model 7689
Bachelier's gaussian model 7690
Bachelier-Samuelson model 7788
Bachmann–Landau notation 6285
back-propagation 6978
backdoor 7222
background independent 5867
backscattering 5118
backshift operator 7985
backward derivative 6620
backward propagation of errors 6978
backward stepwise regression 6417
backwardation 7653
bagging 6743, 6803
balance measure 540
balance sheet equation 8026
balance sheets-based method 8062
balanced and orthogonal 6382
balanced chemical reaction 6242
balanced design 1007, 8797
balanced plan 1007
balanced regression 6382
ballistics 3263
baloney detection kit 109
Banach fixed-point theorem 1781
band gap 4300
band-pass filter 8503
banker's rounding 7569
banking angle formula 3460
Barlow's formula 8428
barn 4338, 5121
barometric height formula 8265
barotropic fluids 3846
barrier option 7667
barrier transparency 4844
Barré Saint Venant hypothesis 8383
barycenter 2635
barycentric reference frame 3158
base (triangle) 2597, 2900
base matrix 6658
bases 2906
bases variables 6660
basic accounting equation 8026
basic reproductive ratio 7342
basic subtractive colors 4334
basis points 236, 7549
batch gradient descent 6689
batch specification limit 8909
Bates formula 8068
Baumol-Allais-Tobin model 8119
Bayes Factor 914
Bayes formula 492
Bayes' chain rule 497
Bayes' formula 485
Bayes' probabilities 485
Bayes' theorem 492
Bayesian Belief Nets 504
Bayesian credibility interval 848
Bayesian inference 504, 848, 894
Bayesian networks 504
Bayesian probabilities 485
Bayesian probability 451
Bayesian statistics 485
beam divergence 4595
bearish 7458

beat frequency 3518
Becquerel 5033
Beer's law 4337
Beer-Lambert law 4337, 4338
Beer–Lambert–Bouguer law 4337
Behrens-Fisher problem 1019
Bel 8207
Bell inequalities 4914
belt friction equation 3464
Beltrami condition 3177
Beltrami identity 3177
bending modulus 3762
Benford distribution 803
Benford law 803
Bernoulli effect 3805
Bernoulli function 652
Bernoulli law 652
Bernoulli polynomials 1672
Bernoulli trials 652
Bernoulli's inequality 457
Bernoulli's numbers 1670
Bernoulli's rule 1461
Bernoulli's theorem 3787
Bernstein polynomials 6570, 6575
Bessel correction 824
Bessel differential equation 3503
Bessel differential equation of order n 1775
Bessel distribution 712
Bessel equation 1775
Bessel function 2383
Bessel function of the second kind 1771
Bessel functions 1768
Bessel series 1768
Bessel-Neumann function of the second kind of zero order 1772
best linear unbiased estimator 6405
beta 7715
beta distribution 761, 8074
beta function 761
beta law 761
beta+ radioactivity 5087
beta- radioactivity 5084
beta-binomial distribution 898
Beta-PERT Distribution 8074
betatron 4103
Bethe formula 5092
Betz's coefficient 3862
Betz's law 3862
Betz's limit 3865
bevel gear 8374
bi-implication 155
bi-normal vector 2868
Bianchi algebraic identity 2126
Bianchi differential identity 2127
Bianchi second identity 2127
biangle 2668
bias 6869
bias error 6345
bias of a neuron 6954
bias of an estimator 822
bias ratio 7593
bias-variance tradeoff 6797
biased maximum likelihood 824
biased sample 545
bias–variance tradeoff 6796
biconditional connector 150
bid-ask spread 7590
Bienayme-Chebyshev inequality 920
Big Bang 5880
Big Bounce 5929
big chill 6037
Big Crunch 5926
big rip 6037
bijective application 407
bilateral t (Student) test on the difference of two means 865
bilateral Z test on the difference of means 856
bilateral Laplace transform 927, 2337
binary classifier 6450
binary code 7169
binary comparators 7133
binary digit 6275
binary option 7667
binary regression 6450
binary relation 286, 379
binary search 8639, 8642
binary search method 6609, 8639
binding energy 5057
Binet's Fibonacci formula 1787
Binet's first formula 5342
Binet's second formula 5347
binomial 323, 528
binomial coefficient 528, 1363
binomial coefficient identity 530
binomial exact test (two tailed) 870

binomial expansion 1366
binomial function 658
binomial law 658
binomial naive Bayes classifier 6826
binomial process 7828
binomial regression 6450
binomial series 1364
binomial theorem 660, 1364
binomial theorem for negative integer exponents 1707
binomial tree 3587, 7827
bioequivalence 1101
biological evolution 459
Biot-Savart law 4023
bipartite graph 3023
birefringence 4366
birefringent crystals 4538
bisector 2591
bispinor 4937
bisquare equation 1395
bit string 7025
bits 6275
bivalent function 2171
bivariate polynomial 1372
bivariate statistics 602
bivector 2089
Black & Scholes differential equation 7796
Black & Scholes linear differential operator 7802
Black Dwarf 5554
black dwarf 5553
black hole 5547, 5793, 5837
Black model 7845
black swan 852
Black-76 model 7845
black-body 3705
blind analysis 58
Bloch sphere 7275
block cipher 7218, 7225
block code 7172
block effect 1065
block matrix 1959
blocking 8836
Bloomberg risk 7633
blueshift 5533
Bohm's interpretation 4915
Bohr magneton 4666
Bohr magnetron 4987
Bohr radius 4630
Bohr's correspondence principle 4802
bolometric 5336
bolometric intrinsic luminosity 5336
Boltzmann constant 3921
Boltzmann distribution 3565, 3569
Boltzmann factor 3566
Boltzmann law 3553
Bolzano-Weierstrass theorem 7045
bond option 7667
bonding function 6197
bonding pair 6198
bonds 7617
Bonferroni approximation 1106
book value 8062
Boole approximation 1106
Boole's inequality 480
Boolean algebra 7105, 7106
Boolean function 7104
Boolean variable 7104
boost 5633
boosting 6803
bootstrap 6741
bootstrap aggregating 6743
bootstrap replicates 6742
bootstrap techniques 6740
Borda method 7475
border 3602
Borel σ-algebra 2495
Borel sets 2495
Borel's law 455
Born interpretation 4691
Born probability 4782
Born rule 4691
Born-Oppenheimer approximation . . 6108, 6132, 6193
Born-von Karman conditions 4285
Bose-Einstein condensate 3577, 3579
boson 4889
bosons relativistic covariant equation 4921
bound orbit 8620
bound state 4768
bounded 7042
bounded function 393
bounded sequence 1652
bounded vector 1797
Bour formula 3385
Boussinesq equation 3852

Box & Whiskers plots 557
Box and Hunter designs 8845
Box and Hunter method 8815
Box multiplication 8796
box plots 557
Box-Cox transformation 8014
Box-Jenkins models 7989
Boyle temperature 3908
Boyle-Mariotte's law 3890, 3904
bra 4730
brachistochrone curve 3434
Braford Hill Criterion 616
braking radiation 4163
branch 4219
branches 2408
break-even point 8047
breakdown voltage 4003
breaking symmetry 3083
Bremsstrahlung 4163
Brewster's angle 4489
bridged network 8751
bright spot 4316
brightness 4343
brilliance 5577
Brocken spectre 4384
Brown Dwarf 5554
brown dwarf 5553
Brown's smoothing 7937
Brown-Mood's median test 885
Brownian bridge 7776
Bubble sort 8650
buckling 8408
buckling differential equation 8412
built-in voltage 8528
bulk modulus 3761
bulk strain 3750
bulk stress 3750
bullish 7458
bullshit detection kit 109
buoyancy-driven flow 8336
burn rate 8030
Burr constants 8980
Burt table 2174
business intelligence 7896
butterfly effect 7327, 8342
Bénard-Marangoni instability 8328
Bézier curve 6570
Bézout's identity 341
Bézout's theorem 341, 7244

C
Cabinet projection 2720
calculus of variations 3435
Calinski-Harabasz index 6866
call option 7664
CALMAR ratio 7725
caloric equation 3934
calorie 3130
calorific capacity 3622
calorific power 6266
Candela 4339
cannibalization rate 8056
canonical associated usable form 6658
canonical basis 1810
canonical dot product 1824, 1840
canonical equations of motion 3186
canonical form 3079
canonical hyperplane 6870
canonical moments 3181
canonical multivariate T-distribution . . 789
canonical parameter 6530
canonical partition function 3567
canonical projection 293
canonical scalar product 1824
canonical system 6651
canonically conjugate 4697
Cantelli's inequality 923
Cantor diagonal process 232
Cantor's Axiom 2633
Cantor's diagonal 225
cap-rate 7553
capacitive reactance 4250
capital asset pricing model 7740
capital at risk 7875
Capital Market Line 7718
capital market line 7746
capitalization factor 7548, 7558
capitalization process 7557
capitalization rate 7553
capitalized value 7548
capstan equation 3464
carbon dating 5036
cardinal 388
cardioid 5402
Carnot efficiency 3669
Carnot's Formula 2520
Carnot-Clausius' principle 3619

CART 6835
cartel 7493
Cartesian equation of a plane 2801
Cartesian equation of the paraboloid . 2839
Cartesian equation of the sphere 2826
Cartesian product 391, 399
Cartesian tensor 2072
cash ratio 8029
cash settlement 7651
Casimir effect 4825
Casimir element 4883
Casimir force 4825
Casimir operator 4883
Casteljau algorithm 6571
categorical variable 580
category 3031
category theory 3031
catenary curve 8561
catoptrics 4354
Cauchy boundary conditions 3488
Cauchy convergence rule 1779
Cauchy distribution 759
Cauchy formula 2415
Cauchy principal value 2376
Cauchy problem 1604
Cauchy ratio test 1777
Cauchy sequence 1658
Cauchy stress tensor 3815
Cauchy theorem 2412
Cauchy's formula 4244, 4358
Cauchy's integral 2417
Cauchy's integral theorem 2415
Cauchy-Goursat theorem 2412
Cauchy-Riemann theorem 2400
Cauchy-Schwarz inequality 1826
Cauchy-Schwarz inequality for integrals 2380, 2388
causal inference 504
cause 3206
caustic reflection 4420
caustic shadow 4322
Cavalier projection 2720
Cavendish experiment 5270
Cayley table 1432
Cayley transformation 2395
Cayley-Klein parameters 2157
celestial equator 5407
celestial sphere 5710
censorship interval 832
census 1326
center 2915
center Line 8935
center of curvature 2866
center of force 3248, 3257
center of gravity . . 2635, 2636, 3213, 3218
center of inertia 2636, 3214
center of mass 3213, 3214
center of projection 2682
center of the circle 2620
centered matrix 1216
centered reduced units 8783
centerline 8967
central angle theorem 2627
central force 3248
central limit theorem 938
central perspective 2680
central symmetry 2643
central tendency 546
central tensor of inertia 3313
centrifugal acceleration 3291
centrifugal force 3238, 3284, 3388
centripetal acceleration 3284
centripetal force 3284
centroid 2608, 2635, 3213
cepheid period-luminosity relation . . . 5605
certain annuities 7562
certain event 452
certain root 1391
certain zero 1392
certification authority 7255
Cesàro sum 1692
chain derivation (univariate) 1482
chain rule 497
chain rule (univariate) 1482
Chan-Karolyi-Longstaff-Sanders model 7780
Chandler wobble 3324
Chandrasekhar limit 5611
change of basis 1992
change of curvilinear coordinates 2096
chaotic orbits 5389
Chapman-Kolmogorov equation 540
characteristic 447
characteristic equation 1611, 1998
characteristic function (probability theory) 926

characteristic function of a fluid at
    equilibrium 3660
characteristic of a logarithm 2285
characteristic polynomial 1611, 1998
characteristic property 6230
charge conservation equation 4114
Charles' law 3891
charts 2183
Chasles relation 1515, 1815
Chebyshev integral 761
check digit 7159
checksum 7156
chemical formula 6237
chemical potential 3660, 4293
chemical separable form 6126
chi-square distribution 773
chi-square law 773
chi-square metric 1238
chi-square test for trend 1268
chi-square test of adjustment 1114
chi-square test of independence 1239
chi-square variance test 948
chiral representation 4943
choice function 528
Cholesky's decomposition 2027
Chomsky hierarchy 7191
Christoffel symbols of the second kind
    2105
chromosome 7024
Chwolson ring 5819
circle 2620, 2915
circle group 1424
circle trigonometry 2503
circular pendulum 3423
circular pitch 8364
circularly polarized 4539
circulation of the magnetic field 4011
circulation of the vector field . . 1584, 4098
circumcircle 2623
circumscribed center design 8857
circumscribed circle theorem 2623
circumstellar habitable zone 5227
civil engineering 8539
Civil time 5461
civil twilight 5219
class frequencies 548
class representative 293
class variable 6450
classic Lagrangian 3166
classic Runge–Kutta method 6642
classical adjoint 1958
classical effective potential 3359
classical electron radius 4670, 5101
classical MDS 6920
classical one-dimensional Schrödinger
    equation 4759
classical regime 3576
classical spatial uncertainty 4706
classical temporal uncertainty 4708
classical Thomson cross section 5102
classification
    accuracy 6519
    false discovery rate 6519
    model accuracy 6519
    negative predictive value 6519
    positive predictive value 6519
    Rand index 6519
    sensitivity 6519
    specificity 6519
classification and regression tree 6835
classification trees 6835
Clausius inequality 3671
Clausius theorem 3672
clean price 7629
climatic precession 5369
clique 3011
closed curve 1583
closed form 1379, 2206
closed formulas 175
closed interval 281, 1350
closed planar curves 2850
closed system 3602
closed term 170
closed thermodynamic cycle 3606
clothoid 8553
cloud iridescence 4396
cluster point 7047
cluster sampling 1329
co-integrated 7982
coastline paradox 7077
coaxial capacity per unit length 4000
coaxial inductance per unit length . . . 4266
Cochran theorem 860
Cochran-Mantel(-Haenszel) test with
    stratified time 1283
Cochran-Mantel-Haenszel Chi-square test

1283
Cochran–Mantel–Haenszel test 1283
code 7205
code word 7172
codomain 2211
coefficient of compressibility 8200
coefficient of determination 619
coefficient of increase of isochoric pressure 3655
coefficient of instability 7694
coefficient of isothermal compressibility 3654
coefficient of linear expansion 3615
coefficient of the monomial 1368
coefficient of variation 588
coefficient of viscosity 3775
coefficient of viscous friction 3465
coefficients 2220
coefficients of the second member . . . 1929
cofactor 1971
cofactor expansion theorem 1971
cofactor matrix 1971
cofactors 8836
cognitive biases 7381
coincidence 3084
cold 3921, 3935
cold electron plasma oscillation model 3949
cold welding 3458
colinear arrows 1797
collocation function 6591
collocation polynomial 6591
color 4325
column index 1936
column length factor 8413
column margins 1233
column-matrix 1937
column-vector 1937
columns
  long column 8409
  short concrete column 8409
  short steel column 8409
  timber 8409
combinatorial 523
combinatorial analysis 522
combinatorial logic model 7104
combined lines 2565
combined standard deviation 1011
combined variance 1011
commensurable surfaces 3455
commercial discount 7556
commercial premium 8172
commodity forward 7647
commodity future 7647
commodity option 7667
common causes 8940
common denominator 317
common divisor 340
common equity 8062
common multiple 348
common ratio 1656
common system 2285
common-cause-failure analysis 8716
communality 6928
commutations numbers 8183
commutative diagram 3033
commutative field 429
commutative group 421, 3038
commutative law 417
commutative magma 418
commutative monoid 419
commutative operation 297, 309
commutative ring 426
commutative unitary ring 426
commutator 4718
comoving coordinates 5990
comoving distance 5948
comoving frame 3087
compact space 7048
company valuation formula 8067
comparison relation 286
comparison relations 380
compatible equations 1930
compatible observable 4697
compatible solution 6680
competitive game 7379
complementary 399
complementary angles 2585
complementary color 4333
complementary design 8821
complementary equation 1606
complementary error function 799
complementary graphs 3022
complementary subspace 1814
complete 4724
complete data likelihood 6703, 6933

complete disjunctive table 2173, 2176
complete functions 2402
complete market 7798
complete metric space 7042
complete multiplicative model 7899
complete set of compatible observables 4697
complete set of residue 365
complete space 7046
complete spatial randomness 1320
complete transformation 2405
completely degenerated 3575
completely inelastic 3368
completely symmetrical tensor 2085
complex conductivity 4238
complex function 2390
complex impedance 8480
complex number 243, 244
  imaginary part 244
  real part 244
complex vectorial spaces 1842
complexion 3562
complexity 6284
  constant complexity 6286
  factorial complexity 6286
  linear complexity 6286
  polynomial complexity 6286
  sub-linear complexity 6286
complexity at the best 6284
complexity at worst 6284
component Value at Risk 7875
components of a tensor 2076
components of a tensor product 2076
composed application 409
composed of a path 3032
composed pulleys 8546
composite application 409
composite center design 8856
composite desirability 8902
composite function 2214
composite Newton-Cotes method 6626
composite number 220
composite standard deviation 1011
composite variance 1011
composition law 403
composition law of accelerations 3387
composition law of velocities 3385
compound annual growth rate 565
compound bodies 6230
compound duration 7635
compound interest 7557
compound option 7667
compound rate 7552
compression 3735
compression coefficient 3761
compression pressure 3735
Compton scattering 5096
Compton wavelength 5098
computed tomography scan 5122
concatenation product 7198
concave function 2280
concave polygon 2890
concentration risk 7884
conclusion 142
conclusion of a sequent 177
concomitance 8816
concordance model 5916
concurrence 7492
condition
  necessary 151
  necessary and sufficient 151
  only if 151
  sufficient 151
condition dependence 506
condition equation 1355
condition: if and only if 151
conditional 148
conditional independence 506
conditional information 3528
conditional loss 7876
conditional maximum entropy model 6477
conditional mean 501
conditional probability 485
conditional probability of failure 8725
conditionally independent 506
Condorcet 7474
Condorcet cycle 7474
Condorcet method 7474
Condorcet winner 7474
conduction 3634
conduction band 4300, 4301
conductivity 4229, 4233
cone 2963
  human eyes 4325
confidence interval 848, 892
configuration space 3159

confinement time 3951
confluent hypergeometric equation . . . 6221
conformal projection 2733
conformal time 5927
conformity test 944
confounder variable 616
confounding 616, 625, 8815
confounding variable 967
confusion 8815, 8817
confusion matrix 6461
congruence 313, 2631
congruent 294, 2633
congruent triangles 2594
conical pendulum 3423
conical perspective 2680, 2681
conical projection 2681
conicity law 5234
conjecture 101
conjugate 244
conjugate prior 896
conjugated matrices 1970
conjugation 269
conjugation equation 4440
conjunction connector 146, 147
conjunctive canonical form 7111
connect component 3014
connected graph 3014
connection 2113
connection coefficients 5750
connex graph 3014
connex relation 290
consciousness 7190
consecutive arrow 3032
consecutive numbers 217
consensus theorem 7108
conservation equation 3683
conservation flow 1883
conservation principle 3618
conservative vector field . 1586, 3294, 3979
consistent estimator 823
consistent matrix 7468
constant amortization loan 7573
constant arc-elasticity model 7497
constant failure rate 8724
constant function 2203, 2209
constant of integration 1512
constant of movement 3187
constant of the perfect gases 3888
constant sequence 1652
constant symbols 169
constant-pressure reaction heat 6251
constants 3072
constants theorem 7107
constraint 3735
constraint pressure 3736
constructive interference 3522
consumer price index 7508
contact potential 8528
contango 7653
context of the sequent 177
context sensitive languages 7195
context-free languages 7195
contingency table in three dimension . 1284
contingent claim 7663
contingent expression 143
continued fraction 369
continuity equation 3683
continuous actuarial rate zero coupon 7624
continuous convolution 2293
continuous function 2231
continuous groups 1417
continuous Kullback-Leibler divergence 3543
continuous medium 3734
continuous on the left 2232
continuous on the right 2232
continuous time Fourier transform . . . 1759
continuous time inverse Fourier transform 1760
continuous variable 580
continuum hypothesis 390
contour integral 1583
contour line 2189, 2860
contours 1866
contracted components 2081
contraction of indices 2080
contradiction 143
contrast 1019, 1024
contrasts 8816
contravariant components 1956, 2043, 2060
contravariant components of a tensor of order two 2075
contravariant metric tensor 2052
contravariant transformation 2097
contribution to the Chi-square of independence 1241

control bits 7179
control chart 8934
  CUSUM control chart 9019
  Levey-Jennings control chart 8971
  MR(2) control chart 9008
  Page-Hinkley control chart 9019
control matrix 7176
control of the proportion of non-compliant individuals by measuring 8905
control of the proportion of non-compliant items by counting 8905
control values 1161
control variables 6726
controlled (linear) model with interactions 8792
convection 3634
convection coefficient 3631
convective acceleration 3830
convenience sampling 1328
conventional cash flow stream 8053
convergence connection 507
convergence ratio 6295
convergent beam 4318
convergent integral 1598
convergent lens 4440
convergent sequence 7043
convergent series 1664
converging relation 506
conversion rate 991
convertible bond 7617
convex combination 1806, 6576
convex function 2280
convex optimization 6645, 6649
convex polygon 2890
convex polyhedron 2934
convexity dollar 7635
convolution 771
convolution theorem 2315
convolutional neural network 6996
Conway knot 3046
Cook’s distance 6394
cooperative game 7378
coordinate line 2091
coordinate singularity 5794
coordinate system 1846, 2091
coordinate-curve 2878
coordinates 1846
coordination game 7399
Copenhagen interpretation 4911
Copernicus reference frame 3158
coplanar 1806
coplanar system 3195
coprime numbers 338
cording rate 7174
Coriolis acceleration 3387
Coriolis force 3388
Coriolis frequency 3395
Coriolis parameter 3395, 8296
Cornu spiral 8553
corollary 101
corrected net assets 8064
correlation coefficient 614
correlation matrix 621
correlogram 7978
correspondence principle 4696
cosecant 2509
cosine 2509
cosine formula 2544
cosine theorem 2525
cosines directions 1811
cosmic inflation theory 5880
cosmic scale factor 5893
cosmic topologies 5899
cosmogony 5878
cosmological constant 5775
cosmological density parameter 5898
cosmological horizon 5709, 5890
cosmological principle 5630, 5881
cosmological sphere 5948
cosmological time 5881
cost function 6679, 7020
cost graph value 3015
cost of carry 7655
cost tree value 3015
cotangent 2509
cotangent vector 2053
Coulomb constant 3971
Coulomb force 3968, 3969
Coulomb gauge 4117
Coulomb integrals 6201
Coulomb module 3760
Coulomb modulus 3749
Coulomb potential 3977
Coulomb scattering 5111
Coulomb’s friction laws 3455
Coulomb’s law 3969

countable set . . . 222, 389
counter operator . . . 4812
country risk . . . 7884
country spread risk . . . 7884
couple . . . 3253
couple of force . . . 3253
coupon . . . 7620
coupon rate . . . 7620
Cournot equilibrium . . . 7452
covalent bond . . . 6198
covariance . . . 597
covariance group . . . 3243
covariance matrix . . . 604
covariance principle . . . 5728
covariant components . . . 1956, 2059
covariant components of a tensor of order two . . . 2075
covariant derivative . . . 2113, 2117, 4938
covariant relativistic equation of fermions 4939
covariant vector . . . 2053
covariantly transformation . . . 3242
covariate . . . 1005, 6322
covariates . . . 8836
covector . . . 2053
covered warrants . . . 7646
covering subgraph . . . 3009
covering system . . . 365
Cox proportional hazard model . . . 1291
Cox regression model . . . 1291
Cox-Ross-Rubinstein model . . . 7833
CP violation . . . 5007
Cramér-von Mises GoF test . . . 1140
Cramér-von Mises statistic . . . 1130
Cramer system . . . 1929
Cramer's rule . . . 1980
Cramér's phi . . . 1247, 1249
Cramér's V . . . 1247
Cramér's V coefficient . . . 1249
creation operator . . . 4811
credence . . . 476
credibility interval . . . 892, 910
credit default risk . . . 7884
credit spread . . . 7750
credit spread risk . . . 7884
credit value at risk . . . 7867
criterion of consistency . . . 7469
critical . . . 8229
critical angle . . . 4369
critical damping . . . 3420
critical density . . . 5897, 5905
critical load . . . 8410
critical path . . . 3016
critical point . . . 3910, 6674
critical resistance . . . 8476
critical state . . . 3910
critical superfluid . . . 3910
crop factor . . . 4452
cross product . . . 1800
cross-correlation . . . 2314
cross-entropy . . . 6463
cross-reference . . . 6
cross-section . . . 5114
cross-sectional regression . . . 7756
cross-validation . . . 6428
crossed Gage study . . . 1066
crossed polygon . . . 2889
crossed-tables . . . 2172
Crout's algorithm . . . 2025
CRS model . . . 8043
cryptographic system . . . 7217
cube . . . 2925
cuboid . . . 2925
cumulant generating function . . . 932
cumulants . . . 932
cumulative distribution function . 582, 633, 647
cumulative failure function . . . 8720
cumulative function . . . 647
Curie . . . 5033
Curie point . . . 4092
Curie temperature . . . 4092
Curie's law . . . 4092
Curie's principle . . . 1799
curl . . . 1888
curl of (the magnetic) potential . . . 4128
curl theorem . . . 1891
currency rate option . . . 7667
current density . . . 4229
current intensity . . . 4228
current ratio . . . 8029
current root mean square . . . 8451
curvature
    extrinsic curvature . . . 2665, 2666
    intrinsic curvature . . . 2665, 2666
curvature circle . . . 2866

curvature drop 8252
curvature radius 2865
curvature vector 2864
curve integral 1583
curved 2665
curvilinear abscissa 2851
curvilinear asymptote 2273
curvilinear coordinates 1524, 2092
curvilinear integral 1583
curvilinear kinematics 3224
customer lifetime value 8054
customer risk 8910
cutoff 6450
cutting frequency 3625
cycle 3012, 3017
cycle (permutation) 1441
cycle of order k 1440
cyclic group 421, 426, 1419
cyclic series 7899
cycloid 3434
cyclometrics functions 2508
cyclone 8298
cyclostrophic balance 8296
cyclotomotic polynomial 1398
cyclotron period 4071
cylinder 2959
cylinder equation 2839
cylindrical beams 4318
cylindrical capacitor 4000
cylindrical normal aspect projection . 2738
cylindrical perspective 2712
cylindrical projection 2738
cylindrical system 3315

D
d'Alembert's equation 3477
d'Alembert's principle of inertial forces 3388
d'Alembert's ratio test 1777
daily-trading volume weighted spread 7591
Dalton's law 3924
Dalton's law of partial pressures 3924
damping factor 8483
Dandelin spheres 2793
dark energy 4826, 5775
dark matter 3900
dark star 3234
Darwin's law 7367
dashboard 971, 7896
data envelopment analysis 8034
data matrix 6380
Data Mining 6768
data redundancy level 3534
data science 6768
daughter isotope 5050
Dawes limit 4533
de Broglie normalization condition . . 4692
de Broglie relation 4749
de Broglie thermal wave 4751
de Broglie's relation 4808
de Broglie–Bohm theory 4915
de Moivre and Euler formulas 252
de Moivre formula 254
De Moivre number 1398
De Moivre's number 1419
De Morgan's laws 147, 401
de Sitter double star experiment 3242
De Sitter Universe model 6035
de-magnification 4441
dead mass 8607
debt–equity ratio 8030
Debye forces 3995
Debye solid 3623
decay chain 5050
deceleration 3232
decibel 8207
decidable system 7181
decile 556
decile analysis 6449
decimal and positional system 210
decimal number 210
decimal numbers 224
decision stump 6835
decision theory 7376
decision tree 7401
decision trees 6835
decisioneeering 8021
decisioneering 7965
declination 5431
declining balance 8071
decreasing failure rate 8724
decreasing function 2203, 2209
decreasing sequence 1652
decreasing variable 284
deductive probability 451
deductive science 95
deep beliefs net 6966

deep eclipse 5486
deep in-the-money 7665
deepest ascent 1870
deepest descent 1870
default measurement 2569
defects per million opportunities 8680
defects per unit 8679
defects per unit opportunity 8679
deferred life annuities 8183
definite integral 1500, 1501
definition by induction on the height of a term 172
definition domain 403
definitive screening designs 8845
deflection 3762, 8391
deflection angle 8629
deflection of light
  General Relativity 5815
  Newtonian approach 5359
deformation 3751
degeneracy order 4726
degeneration 3561
degree 2586
degree of a polynomial 2220
degree of a vertices 3007
degree of bias 8689
degrees of freedom 3159, 6357
degrees of freedom (statistics) 775
delete-1 Jackknife samples 6748
deleted residual 6394
delivery check or receipt sample test (acceptance sampling) 8674
Delphi method 87, 7392
delta 7799
delta cash 7799
delta function 1569
delta method 6504
delta VaR 7878
delta-gamma method 7634
delta-hedging strategy 7800
delta-neutral strategy 7800
delta-Normal Value at Risk 7866
delta-v 8607
Deming regression 6372
demonstration plan 9049
Dempster mass spectrometers 4070
dendrogram 6900
dense set 232, 2449
density 3091
density function 647
density map (chemistry) 6169
density parameter 5912
Density-based spatial clustering of applications with noise 6883
density/repartition function of instantaneous failure 8726
departure angle 2703
departure set 403
dependent variable 2210
depressed cubic 1391
depression 8298
depth 2686
derivative 1451
derived rules 177
derived units 3075
derived utility 7515
Descartes's first formula for thin lenses 4432
Descartes-Euler's theorem 2936, 3002
descendants 7025
descriptive analytics 542
descriptive geometry 2679
descriptive statistics 542
design of experiment 8780
design-crossed design 8894
desirability function 8900
destruction operator 4811
destructive interference 3522
determinant 1830, 1926, 1962, 1963
determinant of a square matrix 1966
determinant-optimal designs 8865
deterministic iterated function systems 7041
deterministic Turing machine 6291
deterministic univariate direct bias method 8187
deviance 6462
deviance of the null model 6462
deviance test 6463
diachronic interpretation 487
diagnostic analytics 8780
diagonal matrix 1949
diagonalization of endomorphism 2004
diamagnetic materials 4013, 4042
diameter 2620
diametrically opposed points 2620

diamond 2914
DIANA 6896
diathermal 3602
dichotomic model 6450
dichotomy method 6609, 8639
dielectric constant 3970
dielectric strength 4003
difference 397
difference in fits (DFIT) 6394
differentiability class 1496
differential 1453
differential cross-section 5114
differential curvilinear abscissa 2852
differential equations 1603
  advanced solution 1611
  delayed solution 1611
  Fourier transform method 1608
  Frobenius method 1608
  Laplace transform method 1608
  matrix method 1607
  method of characteristic polynomial 1607
  method of disturbances 1607
  method of integrating factor 1607
  method of variation of the constant 1607
  numerical method 1608
  power series solutions 1608
  separation of variables 1607
  spectral method 1607
differential form 1466
differential hoist 8548
differential scattering cross section 5101
Diffie–Hellman key exchange 7233
diffracted 4318
diffraction limit 4533
diffusion coefficient 3595
diffusion coefficient of matter 3698
diffusion current 3701
diffusion equations 3689
diffusion term 7806
digital signature 7255
digits 212
digraph 3016
dihedral group of order n 1431
Dijkstra's algorithm 8654
dilemma of not rejecting the null hypothesis 948
dimension reduction 6918
dimensional analysis 3080
dimensional equation 3081
dimensional reduction 6100
dimensionality reduction 6918
dimensionally consistent 3074
dimensionless Navier-Stokes-Reynolds equation 3849
Diophantine equation 1378
dioptric power 4433
dioptrics 4354
dip circle 4032
dipolar geometry 4039
dipole 4219
dipole moment 3986
Dirac adjoint 5178
Dirac bispinor 4965
Dirac equation 4938
Dirac field 5180
Dirac function 1569
Dirac matrices 4939
Dirac operator
  covariant form 4939
Dirac peak 1569
Dirac representation 4942
Dirac sea 4969
Dirac slash notation 5178
Dirac-Maxwell equations 5192
direct linear similarity 259
direct methods 6272
direct orthonormal vectors 2864
direct reference frame 3158
direct similarity 259
direct sum of subspaces 1814
directed graph 3012
directing coefficient 1450
direction (force) 3200
direction angles 1811
direction of the conical 2775
direction vector 2804
directional derivative 1870
directional divergence 1879
directional source 4339
directrix 2841
Dirichlet boundary conditions 3488
Dirichlet eta function 1687
Dirichlet integral 1554
Dirichlet series 1683

Dirichlet's oscillations frequencies . . . 3493
dirty price . . . 7633
disconnection . . . 155
discontinuous distribution . . . 3112
discount bond . . . 7619
discount cash flows discount model . . 7613
discount factor . . . 7558, 7788
discount operation . . . 7558
discounted cash flow analysis . . . 8051
discounted cash flow rate of return . . . 8057
discounted payback . . . 8059
discrete convolution . . . 2295
discrete distributions . . . 3112
discrete Fourier transform . . . 2328
discrete Kelly criterion . . . 7440
discrete Kullback-Leibler divergence . 3542
discrete logarithm . . . 7232
discrete time series . . . 7897
discrete time stochastic process . . . 537
discrete uniform function . . . 649
discrete uniform law . . . 649
discrete variables . . . 580
discriminant of conic section . . . 2795
discriminant of the polynomial . . . 1383
discrimination threshold . . . 6512
disintegration constant . . . 5029
disjoint . . . 395
disjoint events . . . 453
disjoint probability . . . 477
disjoint union . . . 395
disjonction connector . . . 146
disjonctive canonical form . . . 7111
disjunctive sequence . . . 279
disordonal interaction . . . 1034
dispersion . . . 587
dispersion curve . . . 4284, 4785
dispersion matrix . . . 6381, 6399
dispersion of absolute differences . . . 552
dispersion parameters . . . 588
dispersion relation . . . 3512, 8353
dispersive wavepacket . . . 3515
dispersive waves . . . 3515
displacement . . . 2574
displacement current . . . 4107
displacement field . . . 3972, 4108
displacement vectors . . . 1815
dissimilarity measure . . . 2450
distance . . . 2194, 2450
distance matrix . . . 1959, 6899
distance to default . . . 7893
distinct group . . . 1437
distinct lines . . . 2565
distinct subgroups . . . 1431
distincts objects . . . 522
distributed force . . . 3736
distribution free . . . 1154
distribution function . . . 633
distributive law . . . 417
distributive operation . . . 309
diurnal component . . . 8301
diurnal movement . . . 5383
diurnal temperature variation . . . 8278
divergence in Cartesian coordinates . . 1883
divergence in cylindrical coordinates . 1886
divergence in polar coordinates . . . 1884
divergence in spherical coordinates . . 1887
divergence operator . . . 1882
divergence theorem . . . 1882
divergent beam . . . 4318
divergent connection . . . 507
divergent integral . . . 1598
divergent relation . . . 506
divergent sequence . . . 7044
divergent series . . . 1664
diverging lens . . . 4440
diversification ratio . . . 7734
divide . . . 294
dividend . . . 313, 338
dividend coupon . . . 7620
dividend discount model . . . 7613
divider . . . 313
divine proportion . . . 1389
division . . . 313
division symmetrical . . . 317
divisor . . . 338
Doehlert's composite designs . . . 8848
dollar duration . . . 7633, 7635
dollar value by basis point . . . 7635
domain of definition . . . 281, 1350, 2211
  type I domain . . . 1520
  type II domain . . . 1520
domain of variation of the factor . . . 8783
dominant factor . . . 1368
dominant strategy . . . 7396
dosimetric quantities . . . 5129
dot product . . . 1821

L^2-dot product 2379
double acceptance plan by attributes . 8918
double entry analysis 8048
double exponential Brown smoothing 7937
double exponential distribution 800
double moving average 7921
double root 1383
double sampling 1330
double sampling plan 8906
double-blinded randomised placebo controlled trial 59
Dow Jones industrial average 7602
downside risk 598
drag coefficient 3854
drag force 3856, 8591
drawdown 7907
drift 3329, 7782
drift current 4229
drift velocity 4228
Drude model 4226
dual expressions 7107
dual space metric 2052
dual vector product 1835, 2047
duality 4110
Duhamel's principle 1619
Dulong-Petit law 3624
dumb indices 2044
dummy variable 174, 1500, 6450
Dunn approximation 1106
Dunning-Kruger effect 84
DuPont identity 8032
Durand model 7612
duration modelling 1271
Durbin-Watson statistics 8015
Durbin-Watson test 8015
dyname 3195
dynamic design 8846
dynamic effect 3200
dynamic friction 3455
dynamic friction coefficient 3455
dynamic hedging 7800
dynamic pressure 3789, 3808
dynamic viscosity 3775
dynamo 8446
Dürer glass 2681, 2682

E
Eotvos effect 3398
earnings before interest and taxes 8030
earth ground 4219
Earth line 2682
Earth's solar analemma 5406
East 2115
eccentric anomaly 5238, 5240, 5409
eccentric load 8409
eccentricity 2775, 2918
Eccles relation 3961
Eckart-Young theorem 2010
ECLAT 6820
eclipse 5486
ecliptic 5245
economic batch quantity 8119
economic function 6646
economic lot size model 8119
Eddington limit 5598
Ederington efficiency 7660
edge 2563
edges 3001
effect 3206, 8825
effect matrix 8796
effective (differential) section of Rutherford (or Coulomb) 5115
effective acceleration 3391
effective annual rate 7559
effective current value 8451
effective DC conductivity 4239
effective dose 5131
effective electrical potential 6189
effective Hamiltonian 6194
effective mass 3087
effective permittivity 4241
effective potential 3358, 5846
effective potential energy 3358, 6187
effective quantum efficiency 4342
effective radiation emission temperature of the planet 8272
effective rate 7558
effective reate 7548
effective relative dielectric constant . . 3961
efficient frontier 7702
efficient market 7764
efficient market hypothesis 7764
Ehrenberg–Siday–Aharonov–Bohm effect 4121
Ehrenfest theorem 4873
eigendecomposition 2009
eigenfunctions 4695

eigenmode . . . 3492
eigenspace . . . 1999
eigenstate . . . 4713, 4726
eigenvalue . . . 1997
eigenvalue decomposition . . . 1224
eigenvalue spectral decomposition . . . 2023
eigenvalues equation . . . 1998
eigenvector . . . 1997, 4713
Einstein coefficients . . . 4591
Einstein ring . . . 5819
Einstein solid . . . 3623
Einstein temperature . . . 3624
Einstein tensor . . . 2142
Einstein's constant . . . 5769
Einstein's effect . . . 5735
Einstein's equivalence principle . . . 5724
Einstein's field equations . . . 5769
Einstein's gravitational coupling constant 5766
Einstein's identity . . . 2142
Einstein's relation . . . 5689
Einstein's tensor . . . 5766
Einstein's velocity addition . . . 5653
Einstein-Chwolson ring . . . 5819
Einstein-De Sitter Universe model . . . 6016
Einstein-Hilbert action . . . 5869
elastic collision . . . 3368
elastic deformations . . . 3751
elastic pendulum . . . 3416
electoral system . . . 7473
electric charge . . . 3095, 3968
electric charge surfacic density . . . 3097
electric constant . . . 3970
electric current . . . 4226, 4228
electric current density . . . 4226
electric density . . . 3097
electric dipole . . . 3984
electric field . . . 3971
electric flow . . . 3996
electric flux . . . 3996
electric force . . . 3968
electric generators . . . 4257
electric homogeneous system . . . 3096
electric intensity . . . 4228
electric isotropic system . . . 3096
electric power . . . 4253
electric resistance . . . 4226, 4234
electric screened potential . . . 6189
electric susceptibility . . . 3972, 4109
electrical generator . . . 4254
electricity . . . 3978
electrochemical potential . . . 4807
electrodynamics . . . 4094
electromagnetic Casimir force . . . 4826
electromagnetic compatibility . . . 3981
electromagnetic field . . . 4109
electromagnetic field tensor . . . 4121
electromagnetic stress-energy tensor . 5764
electromagnetic tensor . . . 4121
electromagnetic waves . . . 4314
electromotive field . . . 4254
electromotive force . . . 4255
electron . . . 4617
electron capture . . . 5088
electron cloud . . . 6132
electron magnetic moment . . . 4999
electron mean free conduction path . . 4227
electron neutrino . . . 5085
electron volt . . . 5027
electron-positron pair production . . . 5126
electron-volt . . . 3093, 5027
electronic amplifier . . . 8536
electronic configuration . . . 4675
electronic layer . . . 4676
electronic magnetic resonance . . . 5211
electronic microscopy . . . 4752
electrostatic dipole . . . 3984
element-wise multiplication . . . 1825
elementary events . . . 480
elementary function . . . 2215
elementary path . . . 3013
elementary reaction progress . . . 6244
elementary solid angle . . . 2550
elementary symmetric polynomial . . . 1446
elements . . . 6230
elements (mathematics) . . . 379
elevation . . . 5465
elevation angle . . . 3264
elitist strategy . . . 7032
ellipse . . . 2917, 2918
ellipsoid . . . 2977
ellipsoid of inertia . . . 3315
ellipsoid of revolution . . . 2978
elliptic curve . . . 2225
elliptic integral
  complete elliptic integral of second

kind 2678
first kind elliptic integral 3412
second-order elliptic integral 2919
elliptic integrals 1599
elliptic metric 2467
elliptic non-Euclidean geometry 2667
elliptic paraboloid 2983
elliptical direct left 4546
elliptical direct right 4546
elliptical indirect left 4546
elliptical indirect right 4546
elliptical partial differential equation . 1625
elliptical polarization 4545
elliptical trigonometry 2503
emission spectrum 4636
emissivity 3705
emissivity coefficient 3705
emittance 4340
empirical autocorrelation coefficient . 7977
empirical cumulated distribution function 583
empirical prior 904
empirical science 92
empirical standard deviation 824
empirical volatility 7823
empirism 89
emulsion 6235
encoder 7167
encoding application 7167
end 2683
endogenous variable 2210, 6322
endomorphism 443
endomorphism reduction 1628, 2004
endothermic 6250, 6260
energetic excitance 3706
energetic photometry 4337
energy 3093
energy absorption coefficient 3705
energy emittance 4492
energy equation 5341
energy flow 4337
energy intensity 4339
energy return on investment 3382
energy returned on energy invested . . 3382
energy-absorption coefficient 5130
energy-momentum tensor 5754, 5755
energy-time uncertainty 4708, 4721
engine work 3293
engineering notation 3098
engineering science 95
ensemble learning 6803
entering variable 6661
enthalpy 3600, 3644
enthalpy change of state 3632
enthalpy of fusion 3633
enthalpy of vaporization 3633
entropy 3530, 3600, 3658
envelope of a polygon 2889
environment 7024
Epicurean paradox 7435
epicyclic gear train 8376
EPR paradox 4913
epsilon-delta continuity 2231
epsilon-delta definition of limit 2227
equal 215, 287
equal angles 2577
equal risk contribution portfolio 7732
equal triangles 2594
equal vectors 1799
equality 215
equality comparators 7133
equally likely probability 545
equation 1352
equation of a great circle 2829
equation of activity 5033
equation of continuity 4114
equation of geostrophic winds 8298
equation of hydrostatic equilibrium . . 5574
equation of motion 5742
equation of motion for a Newtonian fluid 3827
equation of Opticians 4439
equation of propagation of the electric field 4139
equation of state 3224, 3611
equation of state of perfect gases 3888
equation of sunrise 5378
equation of the elastic line 8389
equation of the static fluid 3846
equation of vibrating strings 3487
equation without second member 1606
equations of motion of a particle in an electromagnetic field 4128
equatorial coordinates 5465
equiangular 2902
equidispersion of the Poisson distribution

688
equilateral quadrilateral 2914
equilateral triangle 2601, 2902
equilibrium (static) 3194
equilibrium condition 3212
equilibrium state of thermal ionization 3946
equipartition theorem of energy 3902
equipment variation 8695
equipotent 388
equipotentials 3980
equiprobables 480
equiradial 8857
equity 7604
equity multiplier 8030
equity option 7666
equity portfolio hedging 7660
equity warrant 7643
equivalence 155
equivalence class 292
equivalence postulate 5721
equivalence relation 290
equivalent arrows 1797
equivalent black-body temperature of a planet 8272
equivalent capacitance of capacitors in parallel 4249
equivalent capacitance of capacitors in series 4249
equivalent dose 5131
equivalent fraction 316
equivalent fractions 226
equivalent inductance of inductors in parallel 4251
equivalent inductance of inductors in series 4251
equivalent knot 3057
equivalent linear systems 1930
equivalent matrix 6309
equivalent monthly installment 7575
equivalent period 5132
equivalent polynomials 1379
equivalent rate 7552
equivalent resistance of resistors in parallel 4247
equivalent resistance of resistors in series 4246
erf 3697
ergodic 3031
ergodicity 7982
Erlang 8151
Erlang distribution 767
Erlang distribution function 8147
Erlang-B formula 8164
Erlang-C formula 8168
error 1307, 7152
error correction codes 7152
error function 799, 2351, 3697
error of the first kind 956
error of the second kind 956
escape time methods 7091
escape velocity 3234
essential discontinuity 2277
estimated marginal means 576
estimation test plans 9049
ether 4577
ethical reciprocity 7479
ETS models 7915, 7958
Euclid’s first theorem 351
Euclid’s lemma 351
Euclid’s theorem
  Geometry 2616
  Number theory 351
Euclidean axiom 2634
Euclidean division 314, 338
Euclidean functions of affine connection 2105
Euclidean geometry 2553
Euclidean metric equation 2671
Euclidean ring 340
Euclidean vector space 1821
Euclidean volume 2056
Euler acceleration 3388
Euler angles 2650
Euler difference equation 6588
Euler equation of the 1st form 3828
Euler equation of the 2nd form 3831
Euler equations 3317
Euler force 3388
Euler Gamma constant 1582
Euler indicator 7236
Euler indicator function 365
Euler number 2289, 6301
Euler rotation equation 3317
Euler spiral 8553
Euler theorem for homogeneous functions

2282
Euler variables 3159
Euler's critical load 8413
Euler's formula 251, 2936, 3002, 8413
Euler's hypergeometric integral 997
Euler's product formula 1683
Euler's theorem 7236, 7237
Euler-Bernoulli equation 8400
Euler-Lagrange equations 3169
Euler-Maclaurin formula 1677
Euler-Mascheroni constant 1582
Eulerian cycle 3019
Eulerian derivative 3830
Eulerian graph 3019
Eulerian system 3160
european medicines agency 1101
european option 7666
even function 2212
even numbers 219
even permutation 1444
event 452
event (physics) 3084
event history analysis 1271
event space 3159
event tree analysis 7400
evidence based science 92
evidence levels 60
evolution operator 4770
evolutionary algorithms 7022
evolutionary operation 8863
evolutionary steady strategy 7445
exact language 7192
exact Mann-Whitney test 1173
exact science 92
excess kurtosis 637
excess reactant 6246
exchange rate 7880
exchange traded funds 7604
exchange value 7488
exchange-listed options 7666
exchanged energy 3600
exclusive OR 146
exeligmos 5494
exercise price 7490, 7664
existential quantifier 163
exitance 4340
exogenous variable 2210, 8783
exothermic 6250, 6260
expanded Gage R&R 1066
expectation 585
expectation-maximization algorithm . 6697
expected default frequency 7894
expected loss 6797, 7876
expected mean continuous variable . . . 634
expected monetary value 7403
expected Net Present Value 8050
expected shortfall 7876
expected tail loss 7876
experiment study 1052
experimental design 8799
experimental design with repeated measures 1045
experimental error 8787
experimental frequency 1113
experimental matrix 8796
experimental space 8783
experimental table 8799
expired coupon 7629
explained variable 6322
explained variance 860
explanatory variable 6322
exploratory factor analysis 6925
exponent 308
exponential decay 2284
exponential disease model 7336
exponential distribution 751
exponential friction 3462
exponential function 2217
exponential law 751
exponential of a matrix 1628
exponential sequence 1703
exponential smoothing 7951
  Brown's method 7937
  Holt's and Winter triple exponential smoothing with 3 parameters (multiplicative method) 7951
  Holt's Double Exponential Smoothing with 2 Parameters (Additive Method) 7945
exponential weighted moving average 7928
exponentially modified Gaussian distribution 806
exponentiation 308
exposure time 4557
extended source 4316
extensive form 7398

extensive quantities 3607
exterior derivative 1466
external composition law 404
external energy 3599
external law 404
external variability 8676
externally studentized residuals 6394
extrados 3790
extrapolate 6591
extrapolation 6323
extreme deviation 1203
extreme value analysis 1211
extreme value theory 1211
extreme vertices designs 8880
extremities of the diameter 2620
extrinsic levels 4304
eyewear equation 4439
Eytelwein's formula 3464

F

face 3002
face value 7620
face-centered design 8861
fact
  epistemological fact 53
  ontological fact 53
factor 8783
factor analysis 6925
factor domain 8783
factor effects model 1007
factor level 8783
factor lower level 8783
factor screening experiments 8781
factor upper level 8783
factorial 311, 525
factorial design 8844
factorial plane 1220
factorial ring 427
factorization of polynomial 1377
factorization operation 309
factorization theorem 1377
factorize a polynomial 1371
factory price 7488
failure 8721
failure function 8720
failure mode analysis, effects and criticality analysis 8674
failure mode and effects analysis 8716
failure rate per time-slot 8720
faire value 7647
Falk's scheme 1946
false discovery rate 1107
false position method 6606
false vacuum decay 6095
falsifiable 93
family-wise error rate 1106
Faraday tensor 4121
Faraday's law 4101, 4258
Faraday's law of induction 4259
Faraday's tensor 4128
Faraday-Maxwell law 4102
fararad 3971
Fata Morgana 4399
fault tolerance 8753
fault tree 8755
Fauré random number sequence 6715
feature extraction 6918
feature selection 6918
federal drug administration 1101
feed-forward neural networks 6966
feed-forward perceptron 6985
Feigenbaum constant 7330
Feigenbaum diagram 7328
Feistel cypher 7225
Fermat's little theorem 7240
Fermat's principle 3152
Fermat's principle of least time 4353
Fermi energy 3576, 4805
Fermi level 3576, 4293
Fermi paradox 5226
Fermi sphere 4287
Fermi temperature 3576
Fermi Velocity 4305
Fermi wave number 5145
Fermi wave vector 4287
Fermi's elephant 1749
fermion 4889
ferromagnetic materials 4043
fertility rate 7307
fertility table 7306
Feynman diagram 5160
Feynman path integral 4868
Feynman slash notation 5178
Feynman's trick 1560
Feynman-Stückelberg inter-operation 4950
Feynman-Stückelberg interpretation . 4923
FG-squircle 2812

fiber optics 4604
Fibonnacci spirale 1662
Fick's law of diffusion 3698
Fick's second law 3698
fictionalism 103
field (set) 429
field lines 3980
field of view 4451
field strength tensor 4121
financial analysis 7485
financial asset 7547
financial engineering 7581
financial leverage ratios 8030
fine structure 4644
fine-structure constant 4655, 5197
fine-tuning fallacy 3116
finite cardinal 389
finite correction factor 846
finite element method 6752
finite game 7420, 7421
finite group 1420
finite language 7197
finite mixture model clustering 6874
finite population correction factor 676
finite size basis 1812
fire diamond 6241
first Bianchi identity 2126
first canonical form 7111
first centered derivative with tangent estimate 6620
first Christoffel identity 2109
first classical Heisenberg's uncertainty relation 4706
first coordinate-curve 2878
first cosmic speed 8603
first derivative 1451
first Fick's law 3699
first focal point 4427
first focal point of a lens 4437
first Frenet formula 2865
first fundamental quadric form 2880
first Gauss differential form 2880
first gyroscopic law 3327
first Joule's law 3646
first law of Kepler 5234
first law of thermodynamics 3618
first Maxwell equation 4097
first order integral theorems 1921
first order logic 163
first order predicate logic 163
first singular value 2013
first singular vector 2013
first-past-the-post electoral system . . . 7474
fishbone diagram 8102
Fisher F-test 948, 1006
Fisher criterion 6911
Fisher designs 8845
Fisher distribution 790
Fisher exact test 1254
Fisher index 7527
Fisher information matrix 839
Fisher Linear Discriminant Analysis . 6908
Fisher's combined probability test 963
Fisher's linear discriminant 6912
Fisher's LSD test 1014
Fisher's method 963
Fisher-Cochran theorem 860
Fisher-Pearson coefficient of skewness 636
Fisher-Snedecor distribution 790
fissible 5069
fissile 5069
fitness function 7024
Fitzgerald-Lorentz factor 5739
five Platonic solids 2934
fixed point 1780
fixed rate bond 7617
fixed subset 7040
fixed-censoring 1274, 8764
fixed-point iteration 6296, 6608
fixed-term loan 7572
Flamm's paraboloid 5858
flat 5905
flat angle 2583
flat curves 2850
flat Earth 5284
flat prior 904
flat-space d'Alembertian 5865
flatness 5260
flattening 5301
flexural modulus 3762, 3764
floating exchange rates 7883
floating ground 4219
floating rate bond 7619
floor function 7568
fluence 5130
fluention 5130

fluid 3772
fluxons 4309
focal 4417
focal distance object 4427, 4437
focal image 4438
focal image distance 4427, 4439
focal length 2686, 4417
focal length of a lens 4438
focal object 4437
focal parameter 2785
focal parameter of the ellipse 2779
focalism 7388
foci 2917, 2918
focus 2775
folded Normal distribution 795
foot of a perpendicular 2600
forbidden energy band 4300
force 3194
forced oscillating system 8229
forces node 3195
forces of reaction 3201
forcing function 8229
forecast 7914
forecast error 7903
forecast with prediction interval 6323
forecasting bias 7903
forecasting errors
  average absolute deviation 7903
  average squared residuals 7903
  forecast error 7903
  forecasting bias 7903
  hit score 7904
  mean absolute deviation 7903
  mean absolute error 7903
  mean absolute percentage error 7904
  mean error 7903
  mean forecast error 7903
  mean percentage error 7904
  mean relative error 7904
  mean square errors 7903
  means square deviation 7903
  percentage error 7904
  relative absolute average deviation 7904
  relative bias of prediction 7904
  relative error 7904
  time momentum 7904
foreign-exchange rate 7880
forests 3018
ForEx rate 7880
formal grammar 7194
formal language 7191, 7192, 7201
formula of compound probabilities 486
formula of total probabilities 492
formulas 168
forward $\mathcal{Z}$-transform 2360
forward commitment 7663
forward contract 7647
forward derivatives 6620
forward discrete Fourier transform 2328
forward gear 8378
forward pass 6979
forward rate 7630
forward stepwise regression 6417
forward volatility 7825
Foucault gyroscope 3327
four-current 5646
four-fold plot 6501
four-momentum 5687
four-vector 5706
four-vector acceleration 5648
four-vector current 4119, 4131
four-vector displacement 5639
four-vector energy-momentum 5687
four-vector fermion current 4972
four-vector force 5693
four-vector potential 4119
four-vector pseudo-norm 5706
four-vector velocity 5646
four-wave vector 5688
Fourier coefficients 1727
Fourier law 3684
Fourier method 1617
Fourier optics 1764, 4496
Fourier stability analysis 6757
Fourier theorem 1733
Fourier transform 1759
Fourier uncertainty principle 4747
Fourier-Dirichlet theorem 1733
fourth Maxwell equation 4108
fractal 7055
fractal distribution 746
fraction 314
fractional diffusion processes 7781
fractional factorial designs 8814, 8845
fractional number 223

fractional reserve 7860
frame 3156
frame error 1336
Fraunhofer diffraction 4492
Fraunhofer line 5542
free body diagram 3198
free convection 8318, 8336
free energy 3600, 3674
free enthalpy 3674
free fall speed in vacuum 3435
free family (vector calculus) 1808
free induction decay 5211
free Klein-Gordon equation 4921
free monoid 7200
free slack 3016
free system 3161
free transport 3949
free variable 174
free vector 1797
Frenet frame 2865
Frenet trihedron 2868
Frenet vectors 2865
Frenet-Serret frame 2868
frequency dependent capacitor 4250
frequency domain 1735
frequency modulation broadcasting . . 3524
frequency spectrum 1736
frequentist inference 848
frequentist probability 451
Fresnel approximation 4155
Fresnel coefficients 4487
Fresnel diffraction 4492, 4517
Fresnel integrals 8556
Fresnel representation 3510
Fresnel spiral 8553
Fresnel vector 255
friction coefficient 3593
frictional forces 3201
Friedmann's first equation 5892
Friedmann's second equation 5896
Friedmann–Lemaître-Robertson-Walker metric 5987
Friedmann–Lemaître–Robertson–Walker metric 5987
Fries trilemma 3128
Frobenius (matrix) dot product 1622
Frobenius method 1773
Frobenius norm 1623
front 2683
full adder 7127
full factorial design with 2^n interactions 8796
full Moon 5220
full rank 1940
full subtracter 7131
full width at half maximum 706
fully closed ring 276
function 403
function dot product 2379
functional 3161
functional dependency 2210
functional determinant 1527, 2058
functional generating functions 1789
functional mean 569
functional relation 286, 3072
functions orthogonality 2379
functions symbols 169
functors 169
fundamental constraint of unconstrainted mixtures 8875
fundamental equation of Friedmann models 5905
fundamental equation of propulsion . . 8607
fundamental formula of planetary gears 8380
fundamental frequency 1397, 3492
fundamental Gibbs' equation 3660
fundamental harmonic 3485
fundamental law of the hydrostatic . . . 3846
fundamental mode 3492
fundamental principle of static 3194
fundamental relation of dynamics 3202
fundamental relation of option pricing 7788
fundamental state 4695
fundamental tensor 2072
fundamental theorem for line integrals 1585, 3294
fundamental theorem of calculus 1509
fundamental theorem of hydrostatics . 3801
fundamental theorem of integral and differential calculus 1509
fundamental theorem of Monte Carlo . 924
fundamental theorem of Riemannian geometry 2109
fundamental theorem of statistics 643
fundamental units 3075

future contract 7647
future option 7667
future price 7647
future value 7548, 7557
fuzzy clustering 6824
fuzzy logic 7104, 7136
fuzzy probability 451

G
Gabriel's horn 2986
gain factor 8230
gain of an antenna 4217
galaxy rotation curve 5293
Galilean invariance 3239
Galilean principle of equivalence 3088
Galilean reference frame 3157
Galilean transformation 3240
gambling formula 7440
game 7377
game theory 7376
game value 7415
gamma 7801
Gamma distribution 767
Gamma law 767
Gamma matrices 4937
Gamma Neutral strategy 7801
Gamma radiation 5120
Gamow factor 4851, 5079
Gantt chart 3016
gap 4300
gas 3888
gas theory 3888
gated recurrent unit neural network . . 6993
gauge change 4117
gauge constraint 4117
gauge group 5187
Gauss coordinates 2672, 2878
Gauss elimination 6304
Gauss error function 2351, 3697
gauss gradient theorem 1882
Gauss integral 695
Gauss law for electrostatic 3997
Gauss law for magnetism 4100
Gauss law for the electric field 4097
Gauss map 249
Gauss pivot 6304
Gauss plane 244, 249
Gauss reduction algorithm 6304
Gauss reduction method 6304
Gauss regularization 6440
Gauss series 1666
Gauss surface 3996, 4095
Gauss theorem for the electric field . . 4097
Gauss' mean value theorem 2416
Gauss-Argand plane 244
Gauss-Laplace function 696
Gauss-Laplace law 696
Gauss-Markov assumptions 6350
Gauss-Markov linear model 6348
Gauss-Markov theorem 6351
Gauss-Ostrogradsky theorem 1882
Gauss-Wiener process 7768
Gaussian approximation 4420
Gaussian correlation conjecture 719
Gaussian correlation inequality 719
Gaussian elimination 6305
Gaussian integral 695
Gaussian linear model 6348
Gaussian Mixture Model 6874
Gaussian process 7982
Gauss–Markov theorem 6405
Gay-Lussac's law 3601, 3892
gear 8362
gear train 8362
gene pool 7361
general affine group 1422
general Cauchy distribution 759
general Ehrenfest theorem 4875
general Einstein's field equations 5775
general equation of state of a real gas 3903
general factorial designs 8823
general Friedmann equations 6010
general full factorial designs 8845
general Laurent series 2434
general linear models 6530
general momentum 3180
general relativity 5721
general solution of the homogeneous
        equation 1611
general stratified $2 \times 2 \times n$ table 1284
general Vieta's formulas 1447
generalized Bernoulli equation
    for gases 3936
    for liquids 3810
generalized Brownian motion 7770
generalized canonical variables 3192
generalized Cauchy's integral theorem

2417
generalized coordinates 3159
generalized cross validation 6434
generalized Euler-Lagrange equation . 5749
generalized force 3171
generalized Gamma law 773
generalized Klein-Gordon equation . . 4926
generalized Lagrangian 3161
generalized Laguerre polynomials . . . 1411
generalized least squares regression likelihood 6407
generalized mean value theorem 1462
generalized Pythagorean formula 2525
generalized system 3161
generating family 1807
generating matrix 7176
generation (algorithm) 7025
generative adversarial network 7011
generators of a family of vectors 1807
generatrix 2823, 2956
genetic algorithm 7024
Geneva gears 8375
genotype 7026, 7361
geocentric reference frame 3158
geocoranan 5334
geodesic 2671, 2852, 5749
geodesic corrected by a Lorentz force 4128
geodesic equation 2112, 5749
geodesic lagrangian 5744
geodesically complete space-time 6062
geodesically future-incomplete space-time 6062
geodesically past-incomplete space-time 6062
geoid 5310
geometric Brownian motion 7780
geometric mean 565, 8902
geometric probability 516
geometric return on investment 7548
geometric sequence 1656
geometric standard deviation 592
geometric vector space 1815
geometrical distance 2600
geometrical optics 4314
geopotential 3840
geostationary orbit 8615
geostrophic balance 8293
Gibbs energy 3674
Gibbs equations 3675
Gibbs free energy 3600
Gibbs phenomenon 1747
Gibbs postulate 3551
Gibbs-Helmholtz relation 3674
Gini impurity index 6839
Gini index 8099
Girard’s theorem 2547
Glivenko-Cantelli theorem 643
global dispersion 8693
global effective rate 7575
global extremum 392
global gauge 5185
global maximum 392
global minimum 392
global optimization 6645
Golden ratio 1389, 1663
Golden Rule of morality 7479
good projection 3061
goodness of fit Chi-square test 1114
Gordon-Shapiro model 7613
grad 2589
gradient ascent 6689, 6691
gradient boosting 6945
gradient descent 6689, 6691
gradient descent method 6684, 6980
gradient in Cartesian coordinates 1871
gradient in cylindrical coordinates . . . 1875
gradient in polar coordinates 1875
gradient in spherical coordinates 1875
gradient of a scalar field 1869
gradient operator 1869
gradient theorem 1882
gradient theorem for line integrals . . . 1585, 3294
gradient vector 1869
graduation 2177
Gram determinant 2058
Gram matrix 2014
Gram-Schmidt orthogonalization method 2062
Gramian matrix 1954
Grandi’s series 1691
graph 403, 2179, 3000
graph symmetric about the x-axis 2196
graph symmetric about the y-axis 2196
graph symmetrical about the origin . . 2197
graphic form 7398

graphs . . . 2183
Grassman rule . . . 1835
Grassmann rule . . . 2047
gravitation potential . . . 5266
gravitational constant . . . 5256
gravitational constant of a star . . . 5340
gravitational d'Alembert's equation . . . 5865
gravitational potential . . . 5267
gravitational propagation equation . . . 5865
gravitational redshift . . . 5797
gravitational slingshot . . . 8622
gravitational time dilation . . . 5797
gravitational wave equation . . . 5865
gravitational waves . . . 5861
graviton . . . 5168
gravity assist . . . 8622
gravity dam . . . 8581
gravity names . . . 8344
Gray . . . 5130
gray body . . . 3714
Gray code . . . 7120
great circle . . . 2828
great circle distance . . . 1854
greater than or equal to . . . 288
greatest common divisor . . . 340
Green's theorem . . . 1890
Green-Riemann theorem . . . 1890
Greenwood's relation . . . 1280, 8769
gross domestic product . . . 7508, 7528
gross margin . . . 7494
gross national product . . . 7508
gross profit . . . 7488, 7494
gross purchase price . . . 7489
gross selling price . . . 7489
gross substantial value . . . 8064
ground plane . . . 2682
group . . . 421
group of affine transformations . . . 1421
group of chemical elements . . . 5678
group of invariants . . . 1430
group of permutations of order n . . . 1440
group of spatial rotations . . . 2651
group of substitutions of order n . . . 1440
group of symmetries . . . 1430
grouped distribution . . . 554
growing function . . . 2284
growth quotient . . . 2212
Guldin's Theorems . . . 3221
gyro-synchrotron pulsation . . . 4069
gyromagnetic factor . . . 4041, 4985, 5206
gyromagnetic ratio . . . 4987
gyroscope . . . 3324
gyroscopic rigidity . . . 3325

H
Hadamard matrix . . . 8807
Hadamard quantum gate . . . 7291
Hadley cells . . . 8292
Hafele-Keating experiment . . . 5664
Haidinger fringes . . . 4563
half adder . . . 7126
half subtracter . . . 7130
half-bisector . . . 2591
half-interval search . . . 8642
half-life . . . 5030
half-life period . . . 5030
half-line . . . 2566, 2567
half-Normal distribution . . . 796
Hall coefficient . . . 4066
Hall constant . . . 4067
Hall's resistance . . . 4066
halo orbit . . . 5501
Hamilton function . . . 3182
Hamilton's principle . . . 3152
Hamiltonian . . . 3180, 3182
Hamiltonian cycle . . . 3019
Hamiltonian graph . . . 3019
Hamiltonian of the system . . . 3360
Hamming distance . . . 7167
Hamming space . . . 7169
Hamming weight . . . 7169
Haphazard sampling . . . 1328
hard clustering . . . 6824
hard exchange rate pegs . . . 7883
hard prior . . . 904
hard science . . . 92
Hardy's relation . . . 7685
Hardy-Weinberg equation . . . 7365
HARKing . . . 71
harmonic analysis . . . 1763
harmonic damper . . . 8224
harmonic frequencies . . . 3492
harmonic function . . . 1899
harmonic mean . . . 563
harmonic model of the atomic bonding 6187
harmonic oscillator . . . 3417, 8224, 8483

harmonic oscillator (1D) 4808
harmonic sequence 1655
harmonic series 1665
harmonic spherical wave 8212
harmonic trap 4820
Hartley constant 1205
hash function 7249
hat matrix 6387
Hausdorff distance 7056
haversine distance 1860
haversine function 1858
Hayashi-path 5541
hazard 1297
hazard diamond 6241
hazard function 8720
hazard rate 1291
hazard ratio 1293, 8720
hazards 7898
heat 3599
heat capacity 3622
heat capacity ratio 3647
heat diffusion equation 3688
heat energy 3092
heat equation 2324, 3688
heat kernel 3696
heat of combustion 6266
heat of reaction at constant volume . . 6250
heat transmission coefficient 3684
Heat-Jarrow-Mortong models 7779
heat-transfer 3602
Heaviside function 1573, 2348
hedge fund 7599
height 2607
height of a term 171
Heisenberg equation of motion 5169
Heisenberg field operators 5173
Heisenberg operator (of evolution) . . . 4775
Heisenberg picture 4776
helical gears 8374
helicity 4951, 4980
heliocentric reference frame 3158
helium flash 5544, 5545
helix 2857
Helmholtz energy 3600
Helmholtz equation 4154
Helmholtz equation for electrodynamics 4153
Hemholtz decomposition 1914
hemispherical map 2728
heptagon 2889
Hermite polynomials 2385, 4817
Hermitian inner product 1843, 4730
Hermitian matrix 1953
Hermitian operator 4695
Hermitian operators of linear momentum 4714
Hermitian scalar product 4730
Hermitian vector space 1844
Hero's formula 2898
Heron's algorithm 6295
Heron's formula 2898
Hertz dipole 4212
Hertz equation 5630
Hertzsprung-Russell diagram 5540
Herzberg anomaly 6225
Hess's law 6260
Hessian matrix 1715
heterogeneous 6230
hexagon 2889
hidden layers 6966
Hierarchical Agglomerative Clustering 6896
hierarchical agglomerative clustering 6896
hierarchical ascendant clustering 6896
hierarchical cluster analysis 6896
hierarchical clustering 6896
hierarchical descendant clustering . . . 6896
hierarchical techniques 6822
high pressure 8299
Hilbert space 1830, 1844
Hilbert space of polarization states . . . 7270
Hilbert transform 2376
Hilbert's axioms 2632
Hill sphere 5525
historical beta 7716
historical simulation 6709
historical volatility 7740, 7825
historicalevidence 63
hit score 7904
Ho & Lee model 7780
Hofstadter's law 8084
hoist 8546
holes 4296
Holm-Bonferroni correction 1109
Holm–Šidák Correction 1111
holomorphic dynamics 7093

holomorphic function 1472, 2399
holonomic constraint 3160
Holt's double exponential smoothing with 2 Parameters (additive method) 7945
Holt's linear method 7937
homeomorphism 2232
homogeneous 6231
homogeneous coordinates 2642, 2759
homogeneous equation 1606
homogeneous function 2282
homogeneous Poisson process 8147
homogeneous polynomial of degree n 1372
homogeneous quantities 3074
homogeneous solution 1606
homogeneous spatial Poisson process 1320
homogeneous system 1929, 3603
homogeneous system associated to the system 1929
homogeneous wave equation 4153
homogenous Markov chain 537
homographic transformations 2685
homologous 2570
homologous triangles 2594
homologue 2574
homomorphism of field 442
homomorphism of group 441
homomorphism of magma 439
homomorphism of monoid 439
homomorphism of rings 439
homomorphism of vector spaces 442
homoscedastic T-test 1013
homoscedastic hypothesis 1012
homoscedasticity 6350
homothetic dimensions 2562
homothetic transformation 2643
Hookean material 3744
horizon 5837
horizon height 2684
horizon line 2682
horizon plane 2682
horizontal 2683
horizontal asymptote 2270
horizontal coordinates 5465
horizontal symmetry 2574
horsepower 3379
hot 3921, 3935
Hotelling T^2 statistics 1098
Hotelling transformation 1214
Hotelling's T-squared test 1096
hour angle 5450, 5470
Hubble bubble 5893
Hubble constant 5888
Hubble law 5887
Hubble parameter 5888
Hubble sphere 5890
Hubble time 5888
Hubble volume 5890
Hubble-Lemaître law 5887
hue 4325
Huffman algorithm 7211
human reliability analysis 8716
hurdle rate 7682
Hurwitz criterion 7429
Hutchinson operator 7041
Huygens generalized theorem 3313
Huygens relation 593
Huygens-Steiner theorem 3310
hybrid hierarchical k-means clustering 6907
hybrid methods 6804
hydrodynamic 3771
hydrodynamic equation of electrons . 3949
hydrodynamic force 8591
hydrogen bonding 3996
hydrogen type atom without dragging 4629
hydrostatic equilibrium 8265
hydrostatic paradox 3774, 3801
hydrostatic pressure 3872
hyper-volume 2553
hyperbola 2768
hyperbolic angle 2536
hyperbolic non-Euclidean geometry . 2667
hyperbolic paraboloid 2983
hyperbolic paraboloid model 8787
Hyperbolic Parallel Postulate 2667
hyperbolic partial differential equation 1625
hyperbolic rotation matrix 5642
hyperbolic trigonometry 2503
hypercomplex 261
hypergeometric distribution 672
hypergeometric law 672
hyperparameter 6803
hyperstatic 8388
hypersurface 2861
hypotenuse 2508, 2601, 2905

hypothesis . . . 99
hypothesis of molecular chaos . . . 3905
hypothesis of the local sate . . . 3607
hypothesis space . . . 6795
hypothesis testing . . . 944
hypsometric equation . . . 8271
Hôpital’s rule . . . 1461
Hölder’s inequality for sums . . . 1827

I
ID3 . . . 6849
ideal . . . 444
ideal diode equation . . . 8533
ideal gas law . . . 3903
ideal gravity pendulum . . . 3405
ideal mechanical advantage . . . 8360, 8368, 8549
ideal rocket equation . . . 8606
idempotence laws . . . 397
idempotent matrix . . . 1953
identical geometric quantities . . . 2568
identity equation . . . 1355
identity matrix . . . 1937, 1949
identity rotation . . . 2647
idiosyncratic risk . . . 7714
illiquidity . . . 7590
illuminance . . . 4342
illumination . . . 4340
image . . . 286, 2211
image focus . . . 4427
image of an object . . . 4417
images . . . 403
imaginary numbers . . . 243
impact parameter . . . 8630
impact parameters . . . 5095
imperfect correlation . . . 617
imperfect information . . . 7413
implication . . . 155
implicit differentiation . . . 1493
implicit probability . . . 509
implicit rate of return . . . 7610
implicitization . . . 2800
implicitly defined function . . . 1492
implied volatility . . . 7823, 7825
importance functions . . . 6737
importance sampling . . . 6737
impossible event . . . 452
improper fraction . . . 316
improper integral . . . 1595
improper rotation . . . 2001
impulse . . . 3367
impurity measurement criterion . . . 6839
in fine loan . . . 7572
in-the-money . . . 7665
incidence . . . 7348
incident . . . 3005
incident ray . . . 4354
incidental edges . . . 3005
including polytomous LR . . . 6477
inclusion . . . 394
inclusion-exclusion formula . . . 479
inclusive OR . . . 146
income elasticity . . . 7491
incommensurable surfaces . . . 3455
incompatible equations . . . 1930
incomplete beta function . . . 761
incomplete block designs . . . 1075
incomplete regularized Gamma function 1578
incompressible flow . . . 3828
increasing failure rate . . . 8724
increasing function . . . 2203, 2209
increasing sequence . . . 1652, 7043
increasing variable . . . 283
incremental gradient descent . . . 6690
indefinite integral . . . 1501, 1508, 1512
indefinite matrix . . . 2798
independent and identically distributed variables . . . 647
independent edges . . . 3005
independent electron approximation . 6194
independent events . . . 453
independent graph . . . 3011
independent one-sample T-test . . . 864
independent solutions . . . 1635
independent two-sample T-test . . . 1013
independent variable . . . 2210
index . . . 320
index (economy) . . . 1235
index of inconsistency . . . 7469
index position . . . 8689
indicator function . . . 7147
indicator variable . . . 6450
indifference curve . . . 7533
indirect measurement . . . 1313
indirect reference frame . . . 3158
Individual control charts . . . 8963

individual variables 170
indivis loan 7570
induced dipole moment 3994
induced metric 6106
induced metric of the parametrized area 6108
induced subgraph 3008
inductance 4014, 4261
induction law 4102
inductive probability 450
inductive reactance 4263
inelastic collision 3368
inequal 287
inequality 215
inertia cost 8125
inertia matrix 3310
inertia of a point cloud 1223
inertia tensor 3310, 3311
inertial reference frame 3157
inference 155
inference domain 8787
inference rules 159
inferential statistics 542
inferior mirage 4397
inferior planets 5389
infimum 7042
infinite cardinal 389
infinite dimension basis 1812
infinite discontinuity 2277
infinite fraction 372
infinite game 7421
infinite monkey theorem 455
infinitesimal rotation matrix 2157
inflation 6034, 7508
inflection point 1454
influence function 6558
influence matrix 6387
influential point 6391
infometrics 3526
informal fallacy 148
informal logic 142
information 3525
information gain 6851
information matrix 6381
information matrix equality 840
infrasound 8196
inherent variability 8675
inhomogeneous system 1929
initial alias 8818
initial conditions 1605
initial public offering 8060
initial value problem 1605
injective application 406
inner automorphism 1437
inner derivative 1482
inner product 1829
inner product space 1830
inner region 2621
innermost bound circular orbit 5854
innermost stable circular orbit 5849
input error 1337
input layer 6966
input power 3381
inscribed curve 2878
insolation 8272
instant-runoff voting 7474
instantaneous angular velocity 3281
instantaneous dispersion 8693
instantaneous failure rate 8725
instantaneous power 3379
instantaneous speed 3226
instrumentalism 103
insurance costs 8107
insurance premium 7665
integer 215
integer division 338
integer part 6293
integral average 569
integral bounds 1500
integral equation of price dynamics . . 7785
integral limits 1500
integral receiver 3705
integral ring 426
integral transform 2303
integral transformation 1760
integrand function 1512
integration by parts 1532
integration factor 1615
integrator 8504
intelligence 7385
intensity 4340
intensity (force) 3201
intensive quantities 3608
inter-class variance 860
inter-laboratory variance 1019
inter-rater reliability 1260

inter-series variance 1011
interaction parameters 5129
interactions 8818
interactive hypothesis limit 3905
intercept 2195
intercept theorem 2610
interchangeability
  time and space (special relativity) 5638
interest costs 8107
interest rate 7548, 7549
interest rate option 7667
interest rate risk 7619, 7633
interference 4494
interference conditions 4560, 4561
interfringe 4513
interior Schwarzschild solution 5792
interlaced 3058
intermediate bosons 5167
internal angle 2585
internal composition law 404
internal conversion 5090
internal energy 3599, 3637
internal energy of a system 3899
internal energy of standard reaction . . 6259
internal growth rate 8032
internal law 404
internal over-energy 3638
internal rate of return 7681, 8057
internally studentized residuals 6393
interpolate 6591
interpolation polynomial 6591
intersection 394
intersection point 2565
intersymbol interference 8513
interval halving method 6609, 8639
interval invariant 5704, 5733
interval measure 8030
interval notation 281
interviewer bias 1337
interviewer error 1337
intra-laboratory 1019
intraclass correlation coefficient 628
intradox 3790
intrinsic bolometric brightness 5576
intrinsic density 4304
intrinsic information by pair 3528
intrinsic semiconductor 4306
invariant by association 3060
invariant interval 2040, 5634
invariant mass 3087, 5674
invariant measure 540
invariant polynomial 1446
invariant volume element 2058
inventory management by zero rotation
  8112
inventory shortages 8108
inventory turnover ratio 8033
inverse 316, 417, 3032
inverse $\mathcal{Z}$-transform 2362
inverse application 407
inverse Fourier Transform 1760
inverse Laplace transform 2339
inversion 1965
inversion effect of rank by the scores . 7464
invertible matrix 1951
investment 7492
invoice price 7490
involution 247
involutive 417
ionization energy 4633
iridescence 4484
irisation 4396
irradiance 4341
irrational equation 1387
irrational number 229
irreducible Markov chain 541
irregular Plackett-Burman design 8809
irrotational 1889, 2403
irrotational vortex 3839
Irving Fisher fundamental value 7613, 8067
Ishikawa diagram 8102
iso-level curves 2858
iso-utility curve 7533
isobaric coefficient of compressibility 3613
isobaric volume expansion coefficient 3654
isobars 5024, 8262
isobarycenter 2636
isochoric flow 3828
isochron dating 5042
isochrone curve 3434
isochronism 3407
isochrons 5043
isocline 1871
isocost lines 6650
isolated singularities 2438
isolated system 3549, 3602

isolated vertices 3007
isoline 2189, 2858, 2860
isolines 1866
isomer 5089
isomers 6239
isometric projection 2712
isometry 2640, 5639
isomorphic basis 1812
isomorphic graphs 3024
isomorphism 443
isoperimetric inequality 2993
isoquant 2189
isoquants lines 6650
isosceles angles 2900
isosceles trapezoid 2907
isosceles triangle 2597, 2900
isostatic 8388
isostatic beam 8391
isostatic system 3252
isosurface 2860
isothermal coefficient of compressibility 3613
isotones 5026
isotonic regression 6582
isotope 5024
isotope knots 3056
isotopies (knot theory) 3045
isotopies (knots) 3056
isotopologue 5026
isotropic 6231
isotropic system 3090, 3603
isovariance 8857
issue price 7620
iterated conditional mean 501
iterated conditional variance 502
iteratively reweighted least squares . . 6411
Itō lemma 7782
Itō process 7779
Itō's product rule 7793
Itō's quotient rule 7793
Itō's solution 7788
Itō-Doeblin theorem 7783

J
jackknife 6747
Jacobi identity 3191
Jacobian 1527
Jacobian matrix 1527
Jean's mass 5550
Jensen alpha 7750
Jensen's inequality 576
joint and conditional entropies 3531
joint cost and schedule confidence level 8082
joint events 453
joint frequency representation 1233
joint probability 480
Jones vectors 7278
Joukovski transformation 2396
Joule expansion 3661
Joule power 4253
Joules 3293
Joules-Bernoulli equations of the electric field 5696
Joules-Bernoulli equations of the magnetic field 5696
judgemental sampling 1328
Jump discontinuity 2278
just in time supply chain management 8119

K
K-map 7116
Kaplan-Meier estimator 1278, 8767
Karhunen-Loeve transformation 1214
Karl Pearson's test 1114
Karnaugh map 7116
Karnaugh tables 7112
Karush-Kuhn-Tucker conditions 6872
kat 4693
Keesom forces 3995
Kelvin effect 4268
Kelvin-Helmholtz timescale 5559
Kelvin–Stokes theorem 1891
Kendall notation 8152
Kendall robust line-fit method 6318
Kendall's tau-b correlation coefficient 1196
Kendall–Theil robust line 6318
Kepler's equation 5416
Kepler's harmonic law 5242
Kepler's reference frame 3158
Kepler's second law 5234
Kepler's third law 5242
Kerckhoffs principle 7221
Kerma 5130
Kermack-McKendrick model 7340
kernel density estimation 1143
kernel function 1144
kernel functions 6872

kernel regression estimate 6447
kernel smoothing 1143
kerr 404
ket 4693, 4730
key performance indicator 7896
key's entropy 7222
Khinchin theorem 923
Killing vector 5841
kilocalorie 3130
kinematic Euler equations 2658, 3334
kinematic viscosity 3777
kinetic energy 3297
kinetic pressure 3920
kinetic temperature 3921
kinetic theory of gases 3888
King's safety formula 8110
Kirchhoff's Laws 4221
Kirchoff's loop law 4221
Kirchoff's point law 4222
Kirchoff's voltage law 4221, 4222
Klein-Gordon field equation 5176
Klein-Gordon Lagrangian 5174
knot complexity measures 3050
knot diagram 3040
knot graph 3061
knot group 3047
knot invariant 3045
knot invariants 3053
knots theory 3035
known harmful factors and that under
 control 8836
known harmful factors but still
 uncontrollable 8836
Knudsen number 3771
Kolmogorov backward equation 7784
Kolmogorov's axioms 477
Kolmogorov-Smirnov empirical
 distribution 1119
Kolmogorov-Smirnov statistic 6470
Kolmogorov-Smirnov test for 2
 independent samples 1119
Koshal designs 8844
Kronecker product 2073
Kronecker symbol 2044
Kruskal's Stress coefficient 6921
Kruskal-Wallis H test 1187
Kuhn tree 7399
Kummer's differential equation 6221
kurtosis 637
Kutta-Joukowski theorem 8601
Kutta–Joukowski lift 3938
Kármán vortex street 3859
König's decompositions 3347
König's lemma 3013
König's theorems 3347
König-Huygens theorem 593

L

L-simplex design 8878
lack of fit 6345, 8787
ladder operators 4811
lag operator 7985
Lagrange celerity formula 8352
Lagrange equations 3169
Lagrange error 1722
Lagrange inversion theorem 1725
Lagrange multiplier test 6476
Lagrange multipliers method 6667
Lagrange point 5498
Lagrange polynomial interpolation
 coefficients 6595
Lagrange remainder 1722
Lagrange rest 1722
Lagrangian 3161
Lagrangian density of a string 3497
Lagrangian field equations 5172
Lagrangian formalism 3154
Lagrangian mechanics 3152
Lagrangian of the current-field interaction
 4124
Lagrangian of the electromagnetic field
 5178
Lagrangian of the free Dirac spinor field
 5179
Lagrangian's action 3161
Laguerre function 1414
Laguerre's ordinary differential equation
 1404
Lamb shift 4998
Lambda-CDM model 5987
Lambert's cosine law 4344
Lambert's cylindrical equal-area projection
 2743
Lambert's equivalent projection 2743
Lambert's source 4344
Lambertian surface 4344
Lambert–Beer law 4337

Lami's theorem . . . 3261
Lamé coefficients . . . 8321
Lamé constants . . . 8321
Lamé parameters . . . 8321
Lande factor . . . 5206
Landé g-factor . . . 4985
Landé g-factor expression . . . 4995
Langevin equation . . . 3593
Langevin's diamagnetic susceptibility relation . . . 4087
Langevin's function . . . 4090
Langevin's relation of the paramagnetic susceptibility . . . 4092
Langmuir frequency for cold electrons 3951
Langmuir waves . . . 3947
language . . . 169, 7183
language of the formula . . . 174
language syntax . . . 7184
Laplace coefficient . . . 8200
Laplace criterion . . . 7436
Laplace distribution . . . 800
Laplace equation . . . 1899, 2403
Laplace hydrostatic equation . . . 3791
Laplace regularization . . . 6443
Laplace smoothing . . . 570
Laplace spatial principle . . . 1320
Laplace transform . . . 2337, 2358
Laplace's equation . . . 4154
Laplace's force law . . . 4063
Laplace's formula . . . 1966
Laplace's Method of Integration . . . 1557
Laplace's thermodynamcis law . . . 3649
lapse rate . . . 4361
large deviation . . . 852
large middle translation . . . 9022
larger is better . . . 8714
Larmor radius . . . 4069
Larmor's equation . . . 4192
Larson's binomial nomograph . . . 8922
LASER lunar ranging . . . 4595
Laspeyres index . . . 7526
LASSO regularization . . . 6441
latent heat . . . 3632
latent heat of fusion . . . 3633
latent heat of vaporization . . . 3633
Latin Hypercube . . . 6726
Laurent series in positive powers . . . 2417
law
  scientific . . . 90
law of compositions relativistic speeds 5653
law of excluded middle . . . 144
law of exponential of growth . . . 2284
law of fixity . . . 3327
law of orbits . . . 5234
law of propositional logic . . . 142
law of rare events . . . 686
law of reflection . . . 4368
law of sines . . . 2528
law of total expectation . . . 501
law of total variance . . . 502
law of viscous laminar flow . . . 3780
laws of thought . . . 187
Lawson criterion . . . 3951
leading coefficient . . . 6305
leading term . . . 1368
Lean Management . . . 8675
lean Six Sigma . . . 8675
leaning angle . . . 3340
learning rate . . . 6684
least significant bit . . . 6276
least squares classifier . . . 6449
least squares method . . . 6324
least squares method of ordinate deviations 6324
least-squares boosting algorithm . . . 6947
leave-one-out cross-validation score . 6334
leaves . . . 3019
leaving variable . . . 6661
Lebesgue measure . . . 2484
left circular polarization . . . 4546
left curve . . . 2851
left first derivative . . . 6620
left limit . . . 2227
left pseudo-inverse . . . 6381
left rectangle methods . . . 6623
left singular vectors . . . 2014
left zero divisor . . . 427
left-censored . . . 832
legal complement . . . 147
Legendre transformation . . . 3180, 3677
legs . . . 2601, 2900, 2906
Lehr's equation . . . 985
Leibniz criterion . . . 1780
Leibniz differentiation rule for products

1479
Leibniz formula of determinants 1966
Leibniz's integral rule 1560
Leibniz's rule 1780
Leibniz's test 1780
lemma 101
length 3085
length contraction 5655
length dilatation formula 3614
lengths of code 7167
lens 4428
lens maker equation 4439
Lenth's PSE 8896
Lenz's law 4258, 4260
Lenz-Faraday law 4261
Lenz-Faraday's law 4101
Leonardo da Vinci's law of friction . . 3454
leptokurtic 637
Leslie matrix 7334
letter 7193
level 5260
level curve 2860
level hypersurface 2861
level shift 7899
level surface 2860
levelled 5260
lever arm 3251
leverage 6387
leverage effect 7673
leverage ratio 8030
leverage score 6387
levered equity 8044
Levi-Civita connection 2105, 2109
Levi-Civita tensor 2048
Levy alpha-stable 711
Levy process 7785
Lewis operator 6258
Lhenry 4064
libration 5585
libration point 5498
Lie algebra 1417
Lie group 1422
Lie groups 1417
life 464
  from non-life 464
life annuities 7562
life table 7299
lifetime 1273, 8763
lift 6818, 8591
lift curve 6522
light 4314
light beam 4316
light cone 5707
light cone of Universe 5705
light intensity 4339
light pen 4318
light ray 4316
light sources 4315
likelihood 488
likelihood function 485
likelihood of censored data 832
Lilliefors goodness of fit test 1119
limit 2226
limit capable 8701
limit capable process 8687
limit laws 2268
limit of profitability 8057
limit of the sequence 7043
limit set 7047
limited Maclaurin series 1696
limited Taylor series 1697
limiting behaviour 1370
limiting reactant 6246
line 2553
line -matrix 1937
line integral 1584
line margins 1233
line of nodes 2656
line of Universe 5705
line segment bisector 2591
line-element 2851
linear Z-score model 6349
linear amortization 8070
linear application 442, 1934
linear attenuation coefficient . . . 4338, 5120
linear attribution model 7740
linear circuits 8463
linear code 7172, 7175
linear combination 1806
linear combination of atomic orbitals 6196
linear constraints 8881
linear correlation coefficient 614
linear density 3090, 3097
linear differential equation 1605
linear discriminant analysis 6908
linear energy transfer 5129

linear equation 2195, 2206
linear form 443
linear function 2215
linear functional space 4724
linear group of order n 1421
linear Hooke's Law 3745
linear interpolation method of the median 553
linear map 2391
linear mapping 442
linear mass density 3090
linear model 6322
linear moment density 4158
linear momentum 3202
linear operators 4711
linear ordering isotonic regression . . . 6582
linear partial differential equation 1617
linear perspective 2680
linear pulse code modulation 8511
linear regression 6322
linear space of states 4724
linear system 1925, 1929
linear system augmented form 1932
linear system coefficients 1929
linear system solution 1930
linear transformation 1934
linear-elastic material 3744
linearity on the left 1829
linearity on the right 1829
linearly dependent vectors 1808
linearly independent vectors 1808
linearly polarized 4539, 4543
link function 6452, 6530
linked family (vector calculus) 1808
linked variable 174
links 3032
LINPACK test 6288
liquidation value 8064
liquidity 7590
liquidity measures 8029
Lissajous figures 3289
Lissajous orbit 5501
literal calculation 280, 1350
literal part of a monomial 1368
literal relations 280
litre 3087
Little's law 8158
Liénard's generalization 4191
Liénard-Wiechert retarded potentials . 4175
loan 7570
Lobachevsky axiom 2667
local equation of the balance conservation of the momentum 3828
local extremum 1717
local forms of Maxwell's equations . . 4108
local gauge 5185
local gauge invariance 5188
local group 5957
local meridian 5462
local Ohm's law 4233
local polynomial smoothing 6447
local sidereal time 5467
local truncation error 6589
local volatility 7825
locally estimated scatterplot smoothing 6447
locally inertial frame 2126, 5722
location parameter 8736
locus 2598
log rank test 1283
log-likelihood 820
log-likelihood ratio statistic 1151
log-linear Poisson model 6531
log-loss 6494
log-Normal distribution 734
log-normal function 734
log-normal representation of the value of the asset 7788
logarithm function 2284
logarithmic function 2218
logarithmic return 7560
logarithmic search 8642
logic 142, 188
logic function 7104
logic gates 7105
logical equivalence connector 150
logical implication connector 148
logical sequence 1663
logical sum 155
logistic binomial regression 6450
logistic distribution 6453
logistic equation 7317
logistic function 7316
logistic Kolmogorov-Smirnov statistic 6470
logistic map 7317

logistic sequence 7317
logit 6455
logit model 6452
logit regression 6450
London forces 3995
London penetration depth 4310
long format table 2171
long position 7666
long-short term memory neural network 6987
longitudinal disturbance 3475
longitudinal rational deformation 3739
longitudinal table 2172
longitudinal wave 3475
lookback option 7667
Lord's paradox 971
Lorentz boost 5638
Lorentz factor 5637, 5739
Lorentz force 4052, 4063
Lorentz gauge 4117, 4204
Lorentz invariant action 5682
Lorentz law 4052
Lorentz matrix .5638
Lorentz oscillator 8224
Lorentz transformation 5637
Lorentz-Poincare matrix 5638
Lorenz diagram 8097
Lorenz system 8337
Lorenz's butterfly 8342
loss function 7020
lot acceptance sampling plans 8904
lot tolerance percent defective 8910
Lotka-Volterra model 7350
low-level orbit speed 8603
lower bound 392, 7042
lower control limit 8688, 8935
lower Darboux sum 1501
lower incomplete gamma function 770
lower incomplete regularized Gamma function 1579
lower partial standard deviation 598
lower specification level 8684
lower triangular matrix 1950
lowering operator 4811
lowest common denominator 347
lowest common multiple 347, 348
loxodromic distance 1854
LSD test 1014
Luhn algorithm 7157
Lumen 4337
luminance 4343
lunar eclipse 5486
lune 2668
Lux 4340

M

M.K.S.C. system 3076
Möller diagram 4676
Macaulay duration 7635
Mach number 3653, 8218
Mach principle 5726
machine 3380
machine constant 8445
machine learning 6823
Maclaurin polynomial 1723
Maclaurin series 1696, 1723
Maclaurin–Cauchy test 1776
macroeconomics 7487
macrostate 3549
magma 418
magnetic anomaly 4999
magnetic compass 4032
magnetic coupling term 4102
magnetic dipole moment 4037
magnetic excitation 4012, 4108
magnetic field 4009, 4012
magnetic field intensity 4012
magnetic induction 4012
magnetic moment of spin 4985
magnetic monopole 4100, 4111
magnetic permeability of vacuum 4011
magnetic pole strength 4050
magnetic potential 4115
magnetic quantum number 4667
magnetic shadow effect 4323
magnetic susceptibility 4012, 4109
magnification 4441, 4442
magnification equation 4442
magnitude comparators 7133
Magnus effect 3938
Mahalanobis distance 1096, 2466
main nodes 3049
main operator 174
main part of a power series 2434
main trihedron of inertia 3313
maintainability function 8725
major premise 159

Majorana representation 4945, 4955
Majorana's equation 4956
majorated sequence 1652
majority redundancy 8754
Maltese cross gears 8375
Malus' law 4550
management cost 8113, 8125
manifold 2446
Mann-Whithney U test 1174
Mann-Whitney rank sum test 1170
manometers 3866, 3873
Mantel-Cox test 1283
mantissa 3098
margin call 7674
margin error 855, 856
marginal average 1008
marginal cost 7503
marginal efficiency of capital 8057
marginal function 647
marginal means 576
marginal probability 476
marginal profile rows 1234
marginal revenue 7505
marginal utility 7530
marginal Value at Risk 7875
marginally significant effect 956
marginally stable circular orbit 5849
market 7486
market asset line 7749
market basket analysis 6808
market indices 7601
market to book ratio 8033
Markov chain 537
Markov chain reliability model 8757
Markov decision process 7456
Markov process 537, 7765
Markov's inequality 920
Marshall–Edgeworth index 7527
mass 3087, 3206
    active gravitational mass 3088
    apparent relativistic mass 3087
    gravitational mass 3088
    inertial mass 3088
    passive gravitational mass 3088
    relativistic mass 3087
    rest mass 3087
mass absorption coefficient 5129
mass concentration density 6255
mass content associated a species 6255
mass defect 5056
mass density 3089
mass energy 3092
mass field 5164
mass function 647
mass number 5024
mass quantities 3611
mass spectrometer equation 4070
mass term 5183
mass transfer 3602
massic homogeneous system 3090
massive point 2635
material balance equation 6244
material derivative 3829
materialism 3091
mathematical mean 548
mathematical programming 6645
mathematical reserve 7860
mating pool 7029
matrix 1936
matrix convolution 2299
matrix decomposition 2011
matrix diagonal terms 1936
matrix factorization 2011
matrix Hadamard based designs 8845
matrix of centered normalized data . . 1218
matrix of differential operators 3819
matrix of moments flows 5759
matrix of ones 1949
matrix of the second member of the linear
        system 1941
matrix of transformation of the stresses
        strains 3821
matrix product 1935, 1946
matrix scaling 1940
matrix-eigenvalue equation 1997
Matthews correlation coefficient 1254
maturity 7664
maturity date 7620
maximin 7379, 7414
maximum a posteriori 819
maximum a-posteriori estimate 6438, 6443
maximum drawdown 7907
maximum entropy (MaxEnt) classifier 6477
maximum independent 3012
maximum Kolmogorov-Smirnov vertical
        difference 6470

maximum likelihood estimators 818
maximum likelihood ratio test 1150
maximum likelihood test 1150
maximum vertical distance 6518
maxterm 7111
Maxwell bivector 4121
Maxwell relations 3676
Maxwell stress tensor 5765
Maxwell tensor 4128
Maxwell's equations 4094
Maxwell-Ampère equation 4108
Maxwell-Boltzmann function 3566
Maxwell-Boltzmann statistics 3566
Maxwell-Heaviside equations 4094
Maxwell-Poisson equation 4099
Mayer relation 3647
McCulloch–Pitts perceptron 6985
McNemar test 1265
mean 548, 585
mean absolute deviation 589, 7903
mean absolute error 7903
mean absolute percentage error 7904
mean collision time 4227
mean down time 8729
mean drift velocity 4228
mean effect 8798
mean equinox 5469
mean error 7903
mean forecast error 7903
mean lifetime 5031
mean median absolute deviation 589
mean mutual information 3531
mean normalization 696
mean operating time between failures 8728
mean percentage error 7904
mean relative error 7904
mean repair time 8729
mean reverting 7992
mean shift 6888
mean solar time 5406
mean square approach method 8739
mean square deviation 7903
mean square error 587
mean square errors 7903
mean square for error 1017
mean square for treatments 1017
mean time between defects 8729
mean time between errors 8729
mean time between failures 8728
mean time between system accidents . 8728
mean time to answer 8729
mean time to detection 8729
mean time to failure 8729
mean time to restore 8729
mean time to restore service 8729
mean up time 8728
mean value theorem 1460
mean-reversion speed 7781
measure of a quantity 2570
measure theory 2481
measurement 3072
measurement by excesses 2569
measurement control charts 8936
measurement error 1336
measurement instrument bias 1337
measurement systems analysis 1066
measurement uncertainty 1307
mechanical engineering 8359
mechanical forces at constant pressure 3643
mechanical forces at constant temperature 3643
mechanics Lagrangrian 3363
medial 561
median 550, 2607, 6561
median absolute deviation 589
median class 554
median test 885
mediator 2607
mediator of a triangle 2597
medium 3734
Meissner effect 4308
members of a set 379
Mercator projection 2739
merging currencies 7883
meromorphic function 2402
meromorphic functions 2414
Mersenne number 211
Mersenne's laws 3492
Merton's DP 7894
Merton's model 7890
mesh 4219
mesh law 4221
mesh step 6754
mesokurtic 637
mesons 5166

message Digest MD5 7249
meta language 143
method of discounted cash flows 8054
method of Lagrange multipliers 6325
method of moments 830
method of truth tables 158
metric 2450
metric covariant tensor 2050
metric equation 2040, 5728
metric prefix 3099
metric space 2447
metric tensor 2072, 5731
metrology 3072
Michelson interferometer 4562
Michelson-Morley experiment 3242
Michelson-Morley factor 5636
micro-evolution 7361
microeconomics 7486
microscopic cross section 5121
microstate 3549
middle of a segment 2571
middle value 550
midhinge 556
midnight Sun 2758
min-max scaling 238
mini-batch gradient descent 6691
minification 4441
minimax 7414
minimum aberration designs 8819
minimum chi-square estimation
  goodness-of-fit 1152
minimum clinically important difference 958
Minimum Description Length 891
minimum distance 7169
minimum spanning tree 3015
minisuperspace approximation 6090
Minkowski inequality 249, 1828
Minkowski metric 5705
Minkowski's matrix 5705
minor 1971, 1974
minor premise 159
minorated sequence 1652
minterm 7111
minuend 304
miracle 2497
miracle theorem 2497
mirage 4397
mirror word 7199
Misiurewicz points 7095
mixed components 2075
mixed components of a tensor of order two 2075
mixed distribution 8175
mixed model 7899
mixed Poisson-Gamma law 6544
mixed product 1838
mixed strategy 7422
mixture 6235
mixture designs 8782
modal value 554
modality 545
mode 550
model calibration 6679
model matrix 6380
modelization plans 8782
modern algebra 416
modern portfolio theory 7581
modified duration 7633, 7635
modified duration VaR 7878
modified internal rate of return 7682
modified wave equation 4273
modular divergence 1879
modular exponentiation 7232
modulation 3523
modulation frequency 3518
module 245, 1818
module congruence class 362
modus ponens 159
Moivre-Laplace central limit theorem 943
molality 6255
molar heat capacity 3628
molar mass 6252
molarity 6255
mole 6252
mole fraction 6255
molecular molar mass 6254
molecule 3893, 6230
Mollweide projection 2746
moment 3251
moment of force 3246, 3251
moment of inertia 3297, 3298
moment of inertia of a diatomic molecule 6208
moment of linear momentum 3246
moment of momentum 3244

moment of order 1 585
moment-generating function 933
moments 932
momentum diffusivity 8335
momentum theorem 3367
monetary aggregates 7519
monetary base 7517
monetary policy 7518
monetary utility 7515
money dilution 7518
money supply 7517
money weighted rate of return 7685
moneyness 7675, 7824
monic trinomial 1391
monitoring limits 8935
monochromatic plane wave 4142
monoid 419
monomial 323, 1368
monotone 7043
monotonic function 2203, 2209
monotonic variable 284
Monte Carlo analysis 6723
Monte Carlo elaborated estimation . . . 6708
Monte Carlo estimation 6708
Monte Carlo simulations 6708
monthly-rate of return 7552
Mood's median test 885
moonbow 4383
moonrise 5220
moonset 5220
Moore-Penrose pseudoinverse 6381
moral hazard 8174
morality 7478
morphism of magma 439
morphisms 3032
Morse potential 6216
mortality rate 7299
mortality table 7299
most significant bit 6276
motion parallax 3229
Mott problem 4709
moving average 567
moving average process of order q . . . 7987
multi-criteria decision making 7462
multi-layer perceptron 6965
multi-linear application 1962
multi-phase 3610
multi-strata proportions test 1283
multicategorical variable 580
multiclass LR 6477
multiclassifiers 6804
multidimensional Newton-Raphson method 837
multidimensional scaling 6920
multifactor ANOVA with repeated measures 1051
multifactorial ANOVA 1007
multigraph 3001
multinomial coefficient 681
multinomial function 680
multinomial logistic regression 6450, 6477
multinomial logit 6477
multinomial theorem 682
multiple acceptance sampling plan by attribute 8919
multiple correlation coefficient 6403
multiple hierarchical ANOVA 1068
multiple hypotheses tests 1106
multiple sampling plan 8906
multiplicand 308
multiplication 308
multiplicative inverse 309
multiplicity 3550
multiplier 308
multivariable 6379
multivariate 6379
multivariate T-distribution 789
multivariate analysis of variance 1091
multivariate chain rule 1483
multivariate control charts 8936
multivariate generating function 1789
multivariate polynomial 1372
multivariate statistics 602
music maths 8195
music watts 8206
mutation 7361
mutual fund 7599
mutual information 3531, 6851
mutually exclusive events 453
mutually incongruent 365
mutually prime numbers 338
Münchhausen trilemma 3128
Münich declaration of the duties and rights of journalists 69

N

nabla of a scalar field 1869

naive Bayes classifiers 6826
naive forecast 7994
naive risk budgeting portfolio 7739
naive set theory 386
Nambu-Goto action 6114
Nambu-Goto condensed form 6115
NASDAQ 7603
Nash equilibrium 7379, 7395
national income 7508
natural basis 2095
natural convection 8318, 8336
natural domain of definition 2207
natural exponential function 2291
natural frequency 3406, 8483
natural integers 216
natural logarithm 2285
natural pulsation 8226, 8483
natural reference frame 2096
natural science 98
Natural selection 7367
natural sequence 216
natural system units 3124
nautical twilight 5219
Navier-Bernoulli hypothesis 8382
Navier-Stokes equation 3827
Navier-Stokes equations of the dynamics of
Newtonian fluids 3826
NB-C log-likelihood 6548
NB2 log-likelihood 6547
necessary and sufficient 151
necessary and sufficient condition 151
necessity 151
negation 145, 147
negation connector 145
negative curvature 2665
neighbourhood (functional analysis) . 1351
neighbourhood of a vertices 3007
neighbours 3005
Nelson rules 8939
Nernst principle 3619
Nernst theorem 3619
nested gage study 1066
nested R&R gage studies 1066
net annual coupon at maturity 7630
net cost of carry 7655
net coupon 7629
net income 7489, 7494
net operating income 7553
net present value 7681, 8050
net present worth 8050
net reproduction rate 7309
net signed area 1498
net substantial value 8064
network mixture designs 8878
Neumann trend test 8015
neural network 6964, 6977
neutral current 5161
neutral element 417
neutral on the right 303
neutral plane 3762
neutral point analysis 8047
neutron capture 5037
neutron star 5606
neutron stars 5547
Newman-Keuls test 1210
newsvendor optimal stochastic stock
problem 8118
Newton approximation scheme 6612
Newton celerity formula 8352
Newton quantum gravity 6074
Newton's binomial 1364
Newton's cannonball 3278, 8603
Newton's difference quotient 1451
Newton's divided-difference formula . 6598
Newton's gravitational law 5257
Newton's law of cooling 3631
Newton's law of heat transfer by
convection 3631
Newton's method 6612
Newton's time 3195
Newton-Laplace speed of sound relation
8202
Newton-Poisson equation 5267
Newton-Raphson method 6612, 6673
Newtonian cosmological models 5881
Newtonian potential 3351
Newtonian relativity 3239
Newton–Cotes formulas 6626
Newton's law of cooling 3635
night 5219
nilpotent matrix 1953
no arbitrage opportunities 7585
nodal lines 3522
node 4219
nodes law 4222
Noether's theorem 3132

nominal conductivity . . . 4239
nominal is the best . . . 8713
nominal variable . . . 580
non hierarchical clustering . . . 6896
non-arbitrage bond price . . . 7626
non-associative operation . . . 303
non-biases sample . . . 545
non-commutative field . . . 430
non-commutative operation . . . 303
non-competitive game . . . 7379
non-cooperative game . . . 7378
non-degenerate semiconductor . . . 4293
non-deterministic objects . . . 7040
non-deterministic Turing machine . . . 6291
non-dispersive wavepacket . . . 3515
non-dispersive waves . . . 3516
non-geometric Plackett-Burman Design 8809
non-hermitian operator . . . 4716
non-hierarchical techniques . . . 6822
non-holonomic constraint . . . 3160
non-inferiority tests . . . 948
non-linear acceleration . . . 3232
non-linear programming . . . 6645
non-local interpretation . . . 4913
non-magnetic materials . . . 4042
non-Newtonian fluids . . . 3778
non-orthogonal designs . . . 8844
non-probability sampling techniques . 1328
non-relativistic Binet formula . . . 5805
non-relativistic cosmology . . . 5881
non-response bias . . . 1337
non-seasonal autoregressive mobile average . . . 7988
non-singular matrix . . . 1951
non-stationary orbits . . . 4626
non-superiority test . . . 948
non-systematic variation . . . 611
non-zero sum game . . . 7378
noncentral chi-square distribution . . . 780
noninformative prior . . . 901
nonparametric binomial demonstration plan 9053
nonparametric binomial demonstration test plan of the reliability . . . 8921
nonparametric bootstrap . . . 6741
nonparametric Chi-square test . . . 1114
nonparametric statistics . . . 1154
nonparametric VaR . . . 7863
nonparametric zero failure . . . 9053
nonsampling errors . . . 1336
norm . . . 245, 1818
  L^2-norm . . . 2380
normal acceleration . . . 3284, 8599
Normal cash flow . . . 8053
normal conjunctive form . . . 7111
normal constraint . . . 3736
normal deformation . . . 3738
normal deformations . . . 3818
normal disjcontive form . . . 7111
Normal distribution . . . 696
normal equations . . . 6325
normal form . . . 7398
Normal law . . . 696
normal number . . . 279
normal representation . . . 7410
normal vector . . . 1801, 2802
normal vibration modes . . . 6128
normalized analysis of variance . . . 1006
normalized PCA . . . 1218
normalized polynomial . . . 1368
normalized sinc function . . . 1767
normalized vector space . . . 1819
North . . . 2116
not necessarily consistent matrix . . . 7468
NOT quantum gate . . . 7289
nothing . . . 4721, 4742
nothingness . . . 4721, 4742
NP-Complete NPC . . . 6292
NP-completeness . . . 6291
nuclear magnetic resonance . . . 4988, 5211
nuclear resonance . . . 4988
nuclide . . . 5024, 5026
nuisance factor . . . 8797
null hypothesis . . . 945
number of convex components . . . 3059
numerical aperture . . . 4607
numerical integration . . . 6621
numerical methods/analysis . . . 6272
Nyquist bandwidth . . . 8511
Nyquist ISI criterion . . . 8513
Nyquist rate . . . 8506
Nyquist sampling theorem . . . 8506
Nyquist sinc pulse . . . 8512
Nyquist-Shannon theorem . . . 8506

O
Oberth effect 8636
Oberth maneuver 8636
object distance 4426
object focal 4427
object language 143
objective function 7021
objective property 6229
objects 3032
oblate (flattened) spheroid 2978
oblate spheroid 2836
obligations 7617
oblique asymptote 2271
oblique foot 2600
oblique line 2600
oblique pyramid 2931
observational study 1052
observed power (statistics) 948
observed volatility 7823
obsolescence costs 8107
obtuse angle 2584
Occam's razor 891
occult science 104
odd function 2212
odd numbers 219
odd permutation 1444
odds 6455
odds ratio 6456
Oersted's law 4114
off-base variables 6660
Ohm 4229
Okun cube 3125
Olbers' paradox 5997
oligopoly agreement 7454
omnibus test 1014
omnibus test for OLS regression 6340
omnidirectional modulus 3750
on-line quality control 8934
one factor at a time 8186, 8782
one sample Z-test 855
one shot at a time designs 8845
one-dimensional wave packet 4786
one-parameter Weibull distribution 8737
one-sample T-test 864
OneR 6810
opaque bodies 4317
open interval 281, 1351
open system 3602
open thermodynamic cycle 3606
operating characteristic curve 8927, 9046
operating curve of type A 8930
operating curve of type B 8930
operating income 7494
operator of the orbital angular momentum 4878
operators precedence 328
opportunity cost 7486
opportunity sampling 1328
opposed vectors 1799
opposing angles by the top 2591
opposing half-lines 2567
opposite 297, 304, 417
opposite side 2508, 2593
optical axis 4416
optical pumping 4592
optical resolution 4500
optical square 4466
optical tunneling 4854
optics 4314
optimal algorithm 6285
optimal designs 8865
optimal management cost for period . 8127
optimal resupply period 8128
option 7664
option plan 7679
option warrant 7643
OR (inclusive) connector 146
orbit 5233
orbit (fractal) 7094
orbital (electronic) 4677
orbital angular momentum 3244
orbital approximation 6194
orbital gyromagnetic ratio 4666
orbital magnetic moment 4985
orbital speed 8603
orbital velocity 8603
orbital-energy-invariance law 8619
order 2631
order number of the maximum's interference 4506
order of a differential equation 1603
order of a graph 3005
order of a group 421, 1430
order of complexity 6285
order of the permutations of the group 1435
order relation 290

order relations 380
order statistics 1156
ordered pair 3001
ordered sequence 522
ordered variable 283
ordinal logistic regression 6487
ordinal utility 7532
ordinal variable 580
ordinary cross-validation 6334, 6430
ordinary generating function 1783
ordinate 2179
orientation (force) 3201
oriented angle 2584
oriented Euclidean space 1830
oriented graph 3012, 3031
origin 2177, 2566, 2851
origin of a vector space 1817
orthocenter 2608, 2624
orthocycloïde 3434
orthodrome 2828
orthodromic distance 1854
orthogonal array 8837
orthogonal contrast 8797
orthogonal curvilinear coordinates . . 1849, 2098
orthogonal decomposition 2023
orthogonal designs 8844
orthogonal linear regression 6372
orthogonal matrix 1954
orthogonal parallel projection 2687
orthogonal projection 2600, 2712
orthogonal projection vector 1821
orthogonal real group of order n 1423
orthogonal regression 6382
orthogonal vectors 1823
orthogonality of functions 2379
orthographic parallel projection 2687
orthographic projection 2722
orthonormal basis 1823
oscillating discontinuity 2278
oscillation 2232
osculating circle 2866
osculating plane 3237
osculator parameter of the parabola . . 2855
Ostrogradsky theorem 1882
out-the-money 7665
outer planets 5389
outer product 2014, 2089, 7273
outer region 2621
outlier 6391
output layer 6966
outside delivery .3642
oval 2918
over damping 8229
over-critical damping 3420
over-sampling 1331
over-stocking 8119
over-the-counter market 7584
overall effect 8798
overall equipment effectiveness 8729
overall net actualized cash flow 8055
overall net cash flow 8055
overall performance of long-term process 8687
overall standard deviation 1011
overall variance .1011
overdetermined systems 1931
oversaturated designs 8845

P

P-waves 3765
Paasche index 7526
pair 7620
pair production .5126
paired sample T-test 867
paired samples 867
pairwise similarity matrix 1959
Palatini identity .5873
palindrome 212
pandemic 7341
Pappus's Centroid Theorem 3221
para-axial 4417
para-science 96
parabola 2766, 2776
parabolic geometry 2634
parabolic partial differential equation 1625
parabolic trajectory equation (without friction) 3265
paraboloid 2983
paraboloidal wave 4155
paradox 102
parallactic angle 5473
parallactic motion 5383
parallax 5249, 5383
parallel axis theorem 3310
parallel beam 4318
parallel orthogonal projection 2722

parallel perspective 2712
parallel regression model 6488
parallel transport 2103, 2114
parallelepiped 2925
parallelogram 2908
parallels 2618
paralogism 148
paramagnetic materials 4013, 4042
parameter of the ellipse 2779
parameter of the parabola 2776
parametric arc 2851
parametric bootstrap 6741
parametric tests 952
parametric VaR 7859
paraxial wave 4155
parent isotope 5050
Pareto distribution 746
Pareto distribution of type I 750
Pareto index 746
Pareto law 746
Pareto-efficient 7394
Pareto-optimal 7394
parisian option 7667
parity bit 7153
parity code 7153
parity symbols 7176
parity-check matrix 7178
Parseval's equality 1757
Parseval's equation 2313
Parseval's relation 2331
Parseval's theorem 1756
part of a set 379
partial correlation coefficient 625
partial derivative 1465
partial determinant 1971
partial differential 1465
partial differential equation 1617, 1621
partial fraction decomposition 1547
partial functions 2878
partial likelihood 1300
partial polarization 4547
partial pressures 3924
partial restore 8724
partial sum 1664
partial total derivative 1472
partially antisymmetric tensor 2089
partially or totally refracted ray 4354
particle 4616
particle solution 4950
particular derivative 3207
particulate border 3602
partition 395, 1504
parts rule 585
Parzen–Rosenblatt window method . . 1143
Pascal's Formula 529
Pascal's law 654, 669, 3773
Pascal's principle 3773, 3774
Pascal's theorem 3774
Pascal's triangle 1363
Pascal's wager 7431
Pascal-Stevin law 3801
Pascalian fluids 3777
Pascals [Pas] 3736
Pasch's theorem 2632
Paschen-Back effect 4882, 4987
passive dipole 4219
passive filter 8503
passive measurements 3072
passive redundancy 8754
path 3012
path integral 1583, 1584
path-independent vector field 3294
Pauli equation 4979, 4983
Pauli exclusion principle 4672
Pauli Hamiltonian 4984
Pauli matrices 2158, 4937
Pauli quantum number 4669
Pauli vector 2158
payback period 8055, 8059
payoff matrix 7410
PCA by ranks 1230
PCA whitening 1229
peak 7043
Pearson chi-square statistic 6467
Pearson covariance 602
Pearson law 773
Pearson product-moment 619
Pearson residuals 6467
Pearson standard deviation 820, 824
Pearson's Chi-square test 1114
Pearson's phi coefficient 1250
Pearson's residual 1113, 1114
Pearson's test 6358
peer review 50
pellicular effect 4268
Pelton wheel 8582

PEMDAS . . . 331
penalized regression . . . 6435
penalized residual sum of squares . . . 6437
pending vertices . . . 3007
pension . . . 7562
pentagon . . . 2889
penumbra . . . 4320
per-experiment error rate . . . 1106
per-thousand . . . 234
percentage . . . 234
percentage error . . . 7904
percentage of completion (chemistry) 6246
percentile . . . 556
percentile bootstrap . . . 6743
perceptron . . . 6953
perfect fluids . . . 3777
perfect gas . . . 3888
perfect information game . . . 7400
perfect numbers . . . 219
pericenter . . . 2785
perigee . . . 2786, 5349
perihelion precession angle . . . 5420
period . . . 2211, 3478
periodic attractor . . . 7326
periodic boundary conditions . . . 3716
periodic cost of ownership . . . 8125
periodic function . . . 2211, 2219
Periods' law . . . 5242
Perko pair . . . 3043
permanent emulsion . . . 6235
permutation . . . 523, 1965
perpendicular . . . 2563
  to a circle . . . 2811
perpendicular angles . . . 2584
perpendicular bisector . . . 2591, 2607
perpetual obligation . . . 7618
perpetual present value . . . 8055
perplexity . . . 3541
perspective . . . 2679, 2680
perspective projection . . . 2681
PERT law . . . 8074
perturbation theory . . . 1645
Peter's projection . . . 2743
phantom energy . . . 5913
phase . . . 3609, 4693, 8228
phase factor . . . 4692
phase I . . . 8966
phase shift . . . 3476
phase space . . . 1633
phase spectrum . . . 1736
phase transition . . . 3609
phase velocity . . . 3480
phase-shift . . . 251, 2510
phase-shift factor . . . 251
phasor . . . 3510
phenomenological science . . . 95
phenotype . . . 7026
phi coefficient . . . 1249
philosophy . . . 55
Phong reflection model . . . 4346
phonon . . . 3623
photo-conductivity . . . 4304
photoelectrons . . . 5106
photographic card (chemistry) . . . 6169
photomaton transformation . . . 1443
photometeor . . . 4388
photometry . . . 4314
photon . . . 3088
photon sphere . . . 5853
photon to baryon ratio . . . 5972
photons . . . 4314
photosphere . . . 5568
physical chemistry . . . 6123
physical pendulum . . . 3413
physical probability (CRR binomial model) 7835
physical quantity . . . 3072
physical settlement . . . 7651
physical space . . . 2850
physical symmetry . . . 3134
physical unit . . . 3073
physics . . . 3074
pi (π) . . . 2583
pi mesons . . . 5167
picture plane . . . 2682
piecewise function . . . 2222
pilot-wave theory . . . 4915
pions (particle) . . . 5167
pivot . . . 6305
pivot direction . . . 6652
pivot step . . . 6653
pivot variables . . . 596
pivot-tables . . . 2172
Plackett-Burman design . . . 8809, 8845
Plackett-Burman design of resolution III 8812

plain vanilla bond 7617
plain vanilla option 7666
planar graph 2179, 3002
planar representation 2178
planar system 1636
Planck time 4906
Planck's 11912 radiation law 4824
Planck's current 4909
Planck's density 4908
Planck's electric charge 4909
Planck's energy 4908
Planck's force 4909
Planck's impedance 4910
Planck's law 3719
Planck's length 4906
Planck's mass 4906
Planck's power 4909
Planck's pulsation 4909
Planck's quantum hypothesis 4626
Planck's temperature 4908
Planck's voltage 4909
Planck-Einstein relation 4626, 4627
plane 2573
plane and parallel capacitor 3998
plane angle 2555, 2563, 2575, 2583
plane figures 2574
plane geometry 2553, 2574
plane surface 2563, 2888
planetary equilibrium temperature . . . 5336, 8272
planetary gear 8376
planetary nebula 5545
plasma 3945
plasma frequency 3947, 4241
plasma frequency for cold electrons . . 3951
plasma oscillations 3947
plasmon 3947
platikurtic 637
plot 403
plotted curve 2878
plowback 8032
PnL report 8113
Pochhammer function 6221
Pogson's Formula 5581
Poincaré diagram 1643
point 2554, 2563
point of view 2682
point spaces 3145
point-biserial correlation 622
point-slope form 1495
pointlike 5195
points in a vector space 1815
Poiseuille law 3781
Poisson Arrivals See Time Average . 8147, 9037
Poisson constant 3647
Poisson distribution 686
Poisson equation of the potential vector 4117
Poisson law 686
Poisson limit theorem 686
Poisson point process 1320
Poisson random point field 1320
Poisson's coefficient 3749, 3752
Poisson's spot 4521
Poisson-Arago spot 4537
Poisson-Gamma density law 6544
Poissons' test with 1 sample 887
polar molecules 3990
polar moment of inertia 3306
polar vector 1800
polarizability 3994
polarization density 4109
polarization plane 4539
polarization vectors 4829
polarized 3993
polaron 3358
pole 2343
pole inversion 260
poles 2366
Polya's law 669
polygon 2888
polygon constraints 6646
polygon of constraints 6649
polygraph 3000
polyhedral formula 2936
polyhedron 2924
polymorphs 6240
polynomial equation 1371
polynomial extraneous solution 1387
polynomial function 2219
polynomial ring 1368
polynomial vector space 433
polytomous variable 580
polytomous variables 6450
polytropic gas equation 3649

polytropic transformations 3603
ponderated point 2635
pooled standard deviation 1011
pooled variance 1011, 8991
population 545, 1326
population (algorithm) 7024
population specification error 1336
portfolio hedging 7800
portfolio theory 7485
portfolio Value at Risk 7873
position 3225
positional system 210
positional voting 7473
positive curvature 2665
positive definite quadratic form 1717
positive electron 5085
positive symmetric bilinear form 1829
positive-definite matrix 1957, 2798
positive-semidefinite matrix . . . 1957, 2798
positivism 98
positron 4951, 4969, 5085, 5087
posterior 487
postnumerando 7548
postulate 100
postulated model 8786
potential 3976
potential difference 3976, 4254
potential flow 3840
potential theory 3361
potential vortex 3839
potentiation 8826
power 308, 3379
power angular resolution 4500
power chain rule 1482
power function 2215
power law (statistics) 746
power notation 308
power of a test 972
power of force 3380
power of the continuum 232, 390
power of the test 955
power output 3381
power reflection coefficient 4488
power spectrum 2314
power triangle 8455
power yield 3381
powered flyby 8636
Poynting vector 4156
practical resonance frequency 8230
practical significance 958
praenumerando 7548
Prandlt number 8335
pre-Euclidean point space 3147
pre-Hilbert space 1830
precession equinox 5369
precession movement 3329
precision 6482
precision matrix 6402
predicate 141
predicted R-squared 6333
predicted residual sum of squares 6333
prediction error 6797
prediction interval 892, 6364
predictive analytics 7914
predictive theory 55
predictor 6322
preferred stocks 7615
prefix code 7207
preimage 286
premature convergence 7030
premises 142
preorder 290
present value 7548, 7557, 8050
pressure 3735
pressure loss 3809
prevalence 7349
price earning ratio 8033
price earnings ratio 7610
price elasticity 7491
price index 7525
price-book value ratio 8065
price-to-earning ratio 7616
pricing and revenue optimization 7497
primality 220
primary winding 8455
prime number 220, 338
primitive 1508
primitive circles 8364
primitive cylinders 8364
primitive diameters 8364
primitive of a function 1501
primitive root of unity 1398
principal axes of inertia 3313
principal coordinates analysis 6920
principal inertia tensor 3313
principal matrix solution 1641

principal quantum number 4647, 4675
principal values 2087
principle 100
principle of causality 3128
principle of complementarity 4749
principle of correspondence ... 5173, 5754
principle of covariance 3242
principle of economy 3152
principle of evolution 3619
principle of explosion 199
principle of Galilean relativity 3239
principle of inertia 3199
principle of least action 3131, 3152
principle of non-contradiction 144
principle of objectivity 4689
principle of Pseudo-Scotus 199
principle of relativity 5632
principle of simplicity 3243
principle of stationarity of action 3163
principle of sufficient reason 199
principle of thermal equilibrium 3618
principle of transmission of fluid-pressure 3773
prior 487
prior local model 8786
prior probability 487
prisoner's dilemna 7416
private equity 7604
private RSA Key 7245
probabilistic network evaluation technique 8074
probabilistic PERT 8074
probabilistic risk assessment 8716
probabilistic tree analysis 482
probability 450, 476, 2494
probability density function . 582, 633, 647
probability integral 695
probability integral transform 6710
probability mass function 633
probability measure 539
probability sampling techniques 1329
probability space 2494
probability transition 538
probable value of deferred capital 8182
probit model 6451
problem 99
  logarithmic problems L 6291
  non-deterministic polynomial time
    problems NP 6291
  polynomial problems P 6291
process capability index (within) 8689
process control capability 8694
process functions 3607
process in discrete time and discrete space 537
process target 8688
process-behaviour chart 8934
product 308
product factors 308
product moment correlation coefficient 614
product of inertia 3311
product rule 486, 1478
profiler diagram 1028
profit margin 7488
profit margins 8031
profit rate 7600
profitability index 8059
prognostic index 6464
programs 6273
progressive or sequential sampling plan 8906
project management engineering 8074
projected shadow 4319
projection 4700
projection direction 2712
projection operators 5181
projective geometry 2679
projective space 2679
prolate (elongated) spheroid 2978
proof 102, 161, 168
proof by induction on the height of a term 171
propagation constant 4205
propagation equation 3477
propagation law 1314
propagation number 3479
propagation vector 4142
proper dimension 5632
proper distance 5632
proper divisor 336
proper energy 3637
proper Euclidean vector space 1823
proper force 5692
proper fractions 316
proper frame 5633
proper length 5655

proper mass 3087, 5632
proper mode 3418
proper referential 5633
proper rotation 2001
proper shadow 4319
proper time 5632, 5660
proper-distance 5704
proportional hazards semi-parametric model 1291
proportional odds model 6488
proportional rate 7552
proportional selection 7029
proposition 144, 7193
propositional calculus 142
propositional formulas 172
propositional logic 142
propositional parameters 588
propositional variables 154
prosecutor's fallacy 477
prospective studies 58
prospective study 1052
proton-proton chain reaction 5560
protractor 2575
provable 177
proximity matrix 6899
prudent tactic 7411
Prym's function 1579
pseudo Monte Carlo method 6714
pseudo-isolated 3603
pseudo-linear solution 1933
pseudo-random generator 6714
pseudo-science 96
pseudo-solution 1933
pseudo-vector 1799
pseudoinverse 6381
pseudomedian 562
public key infrastructure 7255
public RSA key 7242
pull angle 8542
pulley 8542
pulsar timing array 5617
pulsars 5547
punctual 3112
punctual indicator 544
punctual light source 4316
punctual spaces 3145
punctum proximum 4458
punctum remotum 4458
purchasing power 7508
pure competition 7494
pure error 6345
pure Lorentz transformations 5633
pure negative correlation 617
pure positive correlation 618
pure premium 8171
pure quadratic curvature test 8852
pure resonance 8231
pure resonance frequency 8231
pure science 92
pure strategy 7421
put option 7664
put-call parity equation 7760, 7761
pyramid 2931
Pythagorean theorem 2610
Pythagorean triple 2610

Q
Q Friedman test 1191
q-q plot 731
quadrants 2181
quadratic fundamental form 2674
quadratic moment 8385
quadratic programming 6645
quadrature formula 6621
quadric representation of a symmetric tensor 2086
quadrichromy printing 4335
quadrics 2765
quadrilateral 2889
quadrivector 5706
quadrivector pseudo-norm 5706
qualitative property 6229
qualitative variable 545
quality control on reception 8904
quality engineering 8673
quality factor 3421
quality factors 5131
quality index 8914
quality level limit 8910
quality loss function (centered) 8711
quantifiers 163
quantile 556
quantile option 7667
quantile-quantile plot 731
quantitative character 545
quantitative property 6230
quantitative reliability engineering . . . 8716

quantitative variable 545
quantity of intrinsic information 3527
quanto option 7667
quanton 4688, 4750
quantum calculation 7263
quantum charge 4617
quantum circuits 7289
quantum coherence 4858
quantum cosmology 6073, 6083
quantum cryptography 7259
quantum decoherence 4864
quantum electrodynamic Lagrangian . 5192
quantum electrodynamics 5158
quantum fluctuation 4722, 4736
  path integral 4869
quantum gravity 6073
quantum Hall effect 4066
quantum interference 4865
quantum jump 4632
quantum leap 4632
quantum mechanics 4688
quantum number of orbital angular momentum 4667
quantum of the electromagnetic field . 5158
quantum phase transition 3610
quantum physics 4617, 4688
quantum state 4724
quantum statistics 3561
quantum tunneling 4840
quantum tunnelling 6094
quantum vacuum fluctuation 4736
quartiles 556
quasi-correlation matrix 607
quasi-Fermi level 8531
quasi-free particle 4285
quasi-monochromatic wave packet . . . 4789
quasi-Monte Carlo 6715
quasi-Newton's methods 6682
quasi-random numbers 6715
quasi-static thermodynamic transformation 3603
quaternion 261
qubits 7263
questionnaire bias 1337
queuing theory 8141
quick ratio 8029
Quicksort 8652
quota sampling 1328
quotient 313, 338
quotient polynomial 1374
quotient rule 1481
quotient set 292, 363

R
Röntgen 5130
R&R methods 8694
R&R studies 1309
radial quantum number 4647
radian 2503, 2587
radiant emittance 4340
radiant exitance 3706
radiant flow 4337
radiant intensity 4339
radiation 3634
radiation constant 3713
radiation density constant 3713
radiation dose 5130
radiation dosimetry 5129
radiation energy 3092
radiation protection 5092
radical 320
radical functions 2216
radicals 1379
radicand 320
radio-isotopes 5025
radioactive cascades 5050
radioactive dating 5036
radioactivity 5025
radiocarbon age 5038
radiocarbon dating 5036
radioisotope dating 5036
radiological protection 5092
radiometric dating 5036
radiometric quantities 5129
radioïde arc 8553
radius 2620
radius of convergence 1700, 1777
radius of curvature 3246
radius of gyration 3306
radius vector 3244
radix 320
rainbow angle 4379
rainbow cloud 4396
raising operator 4811
random 453
random causes 8940
random decision forests 6852

random error 1309
random forests 6852
random index of consistency 7470
random sample 545
random sampling operator 6750
random sampling with replacement . . . 523
random sampling without replacement 527
random variable 2494
randomization ratio criterion 1321
randomization test 1058, 6749
randomized block design 1073
randomized controlled trial 58
range of a distribution 647
range of existence 2211
range statistics 1203
rank 1162
rank deficient 1940
rank of a family of vectors 1813
rank of a finite family 1938
rank of a matrix 1938
rank test 1156
ranked-choice voting 7474
Rao score function 6473
Rao score test 6476
rapidity 5642
ratchet option 7667
rate of error correcting code 7174
rate of return 7609
rating 6456
ratio of specific heats 3647
ratio test 1777
rational fractions 2221
rational numbers 224
rational operators 285
rational science 95
raw beta 7716
ray diagram 4316
Raychaudhuri equation 5896
Rayleigh criterion of a rectangular aperture 4502
Rayleigh dissipation function 3171
Rayleigh number 8336
Rayleigh resolution criterion 4529
Rayleigh scattering 5102
Rayleigh scattering cross section 5105
Rayleigh-Bénard convection cells 8328
Rayleigh-Jeans law 3720
Rayleigh–Jeans catastrophe 3723
reaction equation 6242
reaction progress table 6247
reactive power 8454
real (reversed) image 4443
real exchange rate 7881
real fluids 3777
real numbers 230
real picture 4418
real rate 7559
realism 103
reality 53
reality variance 7741
realized spread 7591
realized variance 7741
realized volatility 7741, 7825
reason of an arithmetic sequence 1653
recall 6482
receivables turnover ratio 8033
receptors 4257
Rechtschaffner designs 8845
reciprocal 316
reciprocal application 407
reciprocal basis 2068
reciprocal order 381
reciprocal vectors 2067
recovery integrals 6201
rectangle 2894
rectangle methods 6623
rectangle term 8792
rectifying inspection 8930
rectilinear kinematics 3224
recurrent neural networks 6966
recurring definition 1651
recurring sequence 1661
recursive neural networks 6986
recursively defined sequence 1651
recursively enumerable languages . . . 7196
redshift
  cosmological redshift 5996
  gravitational redshift 5735
  radial relativistic doppler redshift 5531
  recombination redshift 5971
reduced Cartesian equation of a line . 2804
reduced centered variable 595
reduced design 8814
reduced gross substantial value 8064
reduced mass 3214
reduced momentum 5341

reduced Planck mass 5910
reduction formulas 1539
reduction of the wave packet 7272
reduction ratio 8365
reference material 5883
reference of the space 3147
reference prior distributions 901
reference system 3156
reflectance 4488
reflectance factor 4351
reflected binary code 7120
reflection 2659
reflection angle 4354
reflection coefficient 4844
reflectivity 4488
reflex angle 2576
reflexive relation 290, 379
refracting medium 4354
refraction law 4367
refraction rate 4412
refractive power 4433
refutable 93
region of convergence 2341
regression coefficient 6322
regression through origin 6369
regression toward the mean 924
regression trees 6835
regressor matrix 6380
regula falsi 6606
regular n-gon 2893
regular element 418, 427
regular graph 3008
regular hexagon 2910
regular language 7195, 7197
regular matrix 1421, 1951
regular part of a power series 2434
regular point 2882
regular polygon 2893, 2924
regular polyhedron 2934
regular square pyramid 2931
regular tetrahedron 2939
regular triangle 2601
regularization 6435
regularized delta function 1571
regularized incomplete Beta function . . 764
regularized regression 6435
Reidemeister movements 3056
rejectable quality level 8910
relation of composition of positions . . 3384
relation order 288
relation symbol 169
relations 286
relative absolute average deviation . . . 7904
relative bias of prediction 7904
relative dielectric permittivity 3971
relative entropy 3542
relative error 1308, 7904
relative integers 222
relative kinematic viscosity 3849
relative magnetic permeability 4011
relative permeability 4109
relative permittivity 4109
relative refractive index 4360
relative risk 6503
relative standard error 611
relative survival 8720
relative translations 3385
relative VaR 7859
relatively prime numbers 338
relativistic apparent angle rotation . . . 5658
relativistic cosmology 5987
relativistic effective potential 5845
relativistic equation of the electron . . . 4938
relativistic mass 5689
relativistic mass momentum relation . 5684
relativistic Schrödinger evolution equation
    4921
reliability 8721
reliability analysis 1271
reliability block diagram 8743
reliability function 8721
reliability test 9049
reliability test plan 9049
reliability theory 1271
remainder 338
remission 1273, 8763
removable discontinuity 2276
removable singular point 2438
Renninger's negative-result gedanken
    experiment 4709
repartition function 582, 647
repeatability error 1047
repeated significance testing errors 991
repeatedly code 7153
replication 8824
report 970

representative . . . 4700
representative class . . . 1798
representative of a class . . . 364
representative sample . . . 542
reproducibility & repeatability study . . 871
resampling . . . 6741
residual error . . . 1047
residual errors . . . 1009
residual polynomial . . . 1374
residual return . . . 7717
residual series control charts . . . 8936
residual standard error . . . 6340
residual stock . . . 8108
residual variance . . . 1009
residuals . . . 8787
residue . . . 358
resistance . . . 4234
resistance of materials . . . 8382
resistance to flow . . . 3781
resistive dipole . . . 4234
resistive work . . . 3293
resistivity . . . 4229
resolution III designs . . . 8818
resolution IV designs . . . 8819
resolution V designs . . . 8819
resonance . . . 3484
resonance curve . . . 8230
resonance frequency . . . 8483
resonance integrals . . . 6201
resonance pulsation . . . 8483
resonance system . . . 3484
response bias . . . 1338
response surface . . . 8785
response surface methodology . . . 8848
rest . . . 313
rest energy . . . 5680
restricted Boltzmann machine . . . 6964
restricted random sampling . . . 1330
resultant . . . 3257
resulting product . . . 323
retrograde linear similarity . . . 259
retrograde motion . . . 5395
retrograde similarity . . . 260
retroreflector . . . 4596
retrospective studies . . . 58
retrospective study . . . 1052
return on assets . . . 8031
return on equity . . . 8031
return on investment . . . 7680
reverse cycle . . . 3668
reverse gear . . . 8378
reverse pulley . . . 8544
reversible light ray . . . 4316
reversion to the mean . . . 924
Reynolds number . . . 3847
rheology . . . 3741
rhetoric . . . 55
rho . . . 7801
rhombohedron . . . 2925
rhombus . . . 2914
Ricci identity . . . 2116
Ricci scalar . . . 2128
Ricci tensor . . . 2127
Ricci theorem . . . 2117
rich language . . . 7192
ridge regression . . . 6437
Riemann coordinates . . . 2672
Riemann integral . . . 1499
Riemann normal coordinates . . . 5725
Riemann scalar . . . 2128
Riemann space . . . 2674
Riemann sum . . . 1504, 6625
riemann surfaces . . . 2410
Riemann zeta function . . . 1683
Riemann-Christoffel tensor . . . 2124
Riemannian circle . . . 1854, 2828
Riemannian geometries . . . 2665
Riemannian metric . . . 2675
Riemannian space tensor . . . 2124
Riemannian volume . . . 2058
right angle . . . 2563, 2584
right circular polarization . . . 4546
right first derivative . . . 6620
right hand rule . . . 3281
right isosceles triangle . . . 2603
right limit . . . 2227
right prism . . . 2933
right pyramid . . . 2931
right rectangle methods . . . 6624
right singular vectors . . . 2014
right triangle . . . 2601, 2905
right zero divisor . . . 427
right-angle triangle . . . 2601
right-angled triangle . . . 2905
right-censored . . . 832
rigid body . . . 3198

rigid electric dipole 3984
rigid rotor 6134
rigid solid 3198
rigidity modulus 3748, 3749
rigorous isochronism 3433
ring 426
rising factorial 6221
risk β 955
risk budget portfolio 7738
risk concentration function 7739
risk difference 966, 6509
risk of credit default 7615
risk of the second kind 954
risk parity function 7739
risk premium 7664
risk ratio 6503
risk set 8721
risk-adjusted net present value 8050
risk-benefit ratio 8025
risk-free rate of return 7548
risk-free return 7548
risk-neutral probability 7837
risk–benefit ratio 8025
RK4 6642
RLC circuit 8475
RMS current 8451
RMS power 8452
RMS watter 8452
Roberston-Schrödinger uncertainty relation
    4735
Robertson-Walker metric 5987
Robertson-Walker radius 5893
Robertson-Walker scale factor 5893
Robertson–Walker metric 5993
robit model 6451
robust indicator 1153
robust model 1153
robust test 1153
ROC curve 6512
Roche limit 5338
Rodrigues representation of the associated
    Laguerre polynomial 1413
Rodrigues' formula 1408
rogue planets 5552, 5554
Roll model 7596
Rolle's theorem 1459
Rolled throughput Yield 8680
rolled throughput yield 8092
rolling mean 567
root mean square 562
root of a polynomial 1371
root-mean-square (rms) speed 3558
ROSE (Random Over-Sampling Examples)
    1332
Rosenblatt kernel function 1144
Rossby number 3395, 8296
rotation 2647
rotation invariant experimental design 8854
rotation movements 3280
rotation of central symmetry 2647
rotation vector 3244
rotational 1888
rotational constant 6206
rotational experimental design 8854
rotational in cylindrical coordinates . . 1896
rotational inertia 3297
rotational kinematics 3224
rotational momentum 3244
rotational movement 3227
roughness 3452
round 409
row index 1936
row reduction 6305
RSA encryption system 7236
RSA number 7241
rule of 16 985
rule of Sarrus 1967
rule of three 1803
ruled surface 2841
rules of ratios and proportions 1803
rules of signs 328
Runge-Kutta method 6637, 6642
runs 8780
Russell's paradox 138
russian option 7667
Rutherford scattering 5111
Rydberg constant 4632

S

S-waves 3765
Sackur-Tetrode formula for
    indistinguishable particles . . . 3934
Sackur–Tetrode equation for
    distinguishable particles 3932
safe distance 7171
safety charge 8171
safety margin 7488

safety parabola 3266
safety stock formula 8110
sagittal diagram 378
Sagnac effect 4579
Sagnac formula for ring interferometers 4583
Sagnac Newtonian formula 4582
Saha ionization equation 3954
Saint-Venant equation 3651
Saint-Venant equation for thermodynamics 3652
salient angle 2576
sample 542
sample average 546
sample design error 1336
sample mean 820
sample size
  for difference of homoscedastic means 985
  for one proportion 876
  for two proportions 876
sample size estimation by Normal distribution 853
sample standard deviation 824
sample variance 822
sampling design 1327
sampling distribution 1111
sampling Pearson coefficient 619
sandwich theorem 2239
saros 5491
saturated designs 8845
saturated main effect designs 8845
Savage criterion 7415
Say law 7516
scalar 209
scalar acceleration 3231
scalar field 1583, 5164
scalar field for a free and spinless particle 5174
scalar Laplacian in polar coordinates . 1901
scalar Laplacian in spherical coordinates 1903
scalar Laplacian operator in cartesian coordinates 1899
scalar multiplication 1802
scalar potential 4116
scalar product 1821, 1843
scalar set 404
scalar velocity 3225
scale height 8265
scale invariant 746
scale law (statistics) 746
scale operators 4883
scale parameter 767, 8736
scale parameters 588
scalene trapezoid 2906
scalene triangle 2597
scaling 2643
scaling factor 5884
scaling identity 2643
scaling law 7862
scanning tunneling microscope 4851
scattering conditions 4843
scattering cross-section 3926
scheduling graph 3016
Scheffé's designs 8878
Schlegel diagram 2729, 2937
Schläfli symbol 2934
Schmidt orthogonalization method . . . 2062
Schrödinger picture 4776
Schrödinger's cat 4781
Schrödinger's classical evolution equation 4771
Schrödinger-Pauli equation 4983
Schrödinger–Pauli equationdddd 4979
Schur complement 1989
Schwarz condition 1471
Schwarz's theorem 1469
Schwarzschild coordinates 5783
Schwarzschild metric 5781, 5792
Schwarzschild precession 5810
Schwarzschild radius 5793
Schwarzschild sphere 5547
Schwing factor 4999
science 96
scientific experiment 3069
scientific fact 98
scientific fraud 55
scientific importance 958
scientific law 90
scientific method 40
scientific notation 3098
scientific proof 89
scientific theory 89
scientism 98
scientist 96

score 6461
score function 835
score vector 837
scores diagram 1220
screen constant 6194
screening designs 8845
screening inspection 8930
screening plans 8815
Screenings designs 8781
season 8281
seasonal variations 7898
secant 2509
secant method 6606
second Borel–Cantelli lemma 456
second canonical form 7111
second coordinate-curve 2878
second cosmic speed 8604
second derivative 1457
second derivative test 1718
second Descartes formula for thin lenses 4440
second focal point 4427
second focal point of a lens 4438
second Frenet formula 2868
second Joule's law 3678
second law of thermodynamics 3619
second London's equation 4312
second Maxwell equation 4100
second order stationary process 7977
second order stress tensors 3815
second quantization 5169
second-order centered moment 587
second-order terms 168
second/azimuthal quantum number . . 4675
secondary additive colors 4332
secondary winding 8455
secular determinant 6201
secular equations 6201
secular variation 4048
Secure Hash Algorithm-1 7253
secured loan 7571
security characteristic line 7715
security level 7411
security market line 7749
security portfolio 7600
segment 2554, 2567
segmentation criteria 6839
seismic damper 8587
selection error 1336
selenelion 4409
self-adjoint Hermitian operators 4716
self-adjoint matrix 1953
self-buckling 8415
self-complementary 3023
self-financing equation 7798
self-financing portfolio 7798
self-induction coefficient 4261
Sellmeier equation 4243, 4358
semi-closed interval 1351
semi-empirical formula of von Weizsäcker 5149
semi-empirical mass formula 5149
semi-interval 281
semi-open interval 1351
semi-perfect fluids 3777
semi-positive matrix 1957
semi-straight line 2554
semi-supervised learning 6824
semipartial correlation coefficient 627
semistudentized residuals 6392
semivariance 598
Sen's slope estimator 6318
sensitivity analysis 8048
sensitivity and specificity 6519
sentence 7193
Sequence
  arithmetic progression 1651
  geometric progression 1651
  harmonic progression 1652
sequence 1651, 7024
sequence chart 8934
sequent 177
sequential logic model 7104
sequential move game 7378
serial or linear connection 506
serial system 8734
service ability 8729
sesquilinearity 1843
session key 7218
set 379
set of admissible solutions 6649
set of divisors 353
set of equivalence classes 292
set of quaternions 261
set of values 1356
set product 399

set theory 379
set-form 7398
set-limit 8338
setbuilder notation 281
sexagesimal minute 2587
sexagesimal second 2587
SHA-1 7253
shade 4325
shadow 4320
shadow cone 4318
shaft rigidity 8406
Shannon formula 3530
Shannon's theorem 7109
Shannon's theory 3526
shape parameter 8736
shape stiffness (rigidity) 8384
share premium 7620
shareholders' equity 8062
shares 7607
Sharpe ratio 7723, 7746
shear deformation 3819
shear mapping 2645
shear modulus 3748, 3749, 3760
shear stress 3749
shear transformation 2645
shear waves 3765
shearing 2645
Sherman-Morrison relation 2036
Shewhart chart 8934
Shewhart's limits 8936
Shewhart-WECO 8938
shining 5577
Shockley equation 8533
short position 7666
short term non-centered biased process
potential capability 8694
short term non-centered potential
Capability for the process . . . 8689
short-term dispersion index 8684
short-term process potential capability
8684
short-term solvency ratios 8029
shrinkage methods 6435
shrinkage parameter 6437
Sidak equation 1106
side-side-side theorem 2596
sidereal day 5390
sidereal period 5390
sides 2575, 2593
siemens 4229
Sierpinski's carpet 7068
Sierpinski's fractal 7058
sigmoid 6451
sign function 327, 6562
signal aliasing 8509
signal factor 8846
signal/noise ratio 8713
signature 5733
signature of a matrix 1966
signature of a vector space 2052
significand 3098
signum function 327, 6562
signum least squares classifier 6449
silhouette coefficient 6867
similar matrices .1994
simple acceptance plan by attributes . 8918
simple and direct ratio 1803
simple and inverse ratio 1804
simple arrangement with repetition . . . 523
simple arrangement without repetition .527
simple bodies 6230
simple combination with repetition . . . 530
simple combination without repetitions 528
simple cycle 3012
simple fixed pulley 8542
simple geometric declining amortization
8072
simple harmonic 8228
simple interest 7552, 7553
simple Kirchhoff law 4351
simple linkage 6897
simple moving pulley 8544
simple multiplicative model 7899
simple or inverse joint rule 1804
simple path 3012
simple pendulum 3405
simple permutation with repetition 526
simple permutation without repetition . 525
simple random sampling 1329, 6733
simple sampling plan 8906
simple Universe number 279
simplex algorithm 6648
simplex lattice designs 8878
simplex method .6646
simplex tableau .6659
simplifiable element 418

Simpson's Formula 2522
Simpson's formulas 2524
Simpson's inverse formulas 2523
Simpson's paradox 967
simultaneous events 3084
simultaneous move game 7378
simultaneous vibration 3494
sine 2509
sine cardinal 2517
sines formula 2545
sines relations 2545
single exponential smoothing 7928
single median method 6318
single-member plurality voting 7474
single-photon emission computed tomography 5123
single-use encryption 7223
singleton 383
singular matrix 1951
singular value decomposition 2004
singular value decomposition theorem 2014
singular values 2013
singular vectors 2013
singularities 2402
singularity 102, 316, 2270, 2414, 5793
sinking sort 8650
siphon 3798
SIR model 7339
size effect 981
size of a graph 3005
size or length of a formula 174
size proportional sampling 1330
skewness 636
skin depth 4273
skin effect 4268
slack variables 6651
slant asymptote 2271
Slater method 6194
slender column 8409
slender rod 2961, 2962
slenderness ratio 8409
sliding elastic modulus 3749
sliding modulus 3749
slip angle 3758
slope 2195
slope selection method 6318
small circle 2828
smaller is better 8714
smaller lower bound 392
smallest bound 392
smallest common denominator 347
smallest common multiple 347
smooth curve 1583
smooth surface 2882
smoothing constant 7928
SMOTe (Synthetic Minority Over-sampling Technique) 1332
Snell's law 4353
snowball sampling 1329
soft clustering 6824
soft exchange rate pegs 7883
soft prior 904
soft science 96
softmax function 6479
softmax regression 6477
SOHCAHTOA 2509
solar anomaly 5420
solar constant 5568
solar day 5390, 5406
solar eclipse 5486
solar granulation 8329
solid 3198
solid angle 2549
solid of revolution 2956
solid physics 4279
solidarity mechanism 8173
solutes 6235
solution 6235
solution vector 1633
solvent 6235
sophism 102
source term 1622, 7806
South 2116
space of clear texts 7217
space of encrypted texts 7217
space of keys 7217
space of spinors 2150
space step 6754
space-time 5708
space-time abscissa 5704
space-time curvilinear abscissa 5704
space-time four-vector 3225, 5639
space-time gauge 5733
space-time interval 5634
space-time metric tensor 5732
spanning tree 3015

sparse matrix 1959
spatial coherence 4530
spatial periodicity 3479
spatial uncertainty relation of Heisenberg 4721
spatially-balanced sampling 1330
Spearman's rank correlation coefficient 1193
special causes 8940
special group of spatial rotations 2158
special linear group of order n 1423
special Lorentz transformations 5633
special real group orthogonal of order n 1424
special relativity 5627
special unitary group of order n 1425
speciation 7361
specific energy 5341
specific heat 3622
specific heat capacity 3628
specific quantities 3611
specific risk 7714
specific variance 6928
specificity index 1235
spectral decomposition 2009
spectral density amplitude 1759
spectral radiant exitance 4341
spectral theorem 2004, 2008
spectrum 1999
spectrum of a matrix 1997
spectrum of an endomorphism 1997
speculation theory 7689
speed 3225
sphere 2964
sphere of action 4339
sphere of influence 5525
spherical aberration 4420
spherical capacitor 4001
spherical dioptres boundaries 4426
spherical harmonics 6161
spherical lune 2668
spherical mirror equation 4440
spherical projections 2727
spherical system 3315
spherical trigonometry 2503
spherical wave 4160, 4206
spherical zone 2749
spherically symmetric system 3315
sphering data 1229
spheroid 2978
spin 4881, 4931
spin $1/2$ fermion field 5180
spin angular moment 3302
spin geometry 2145
spin gravity 8618
spin quantum number 4669
spin resonance 5210
spin-orbit coupling 4899, 6183
spin-orbit effect 4993
spin-statistics theorem 3579
spinor 4932
spinor calculus 2145
spinor field 5180
spinor product 2159
spintronics 4877
spin–orbit coupling 4899
spin–orbit effect 4899
spin–orbit interaction 4899
spiral 2818
split-plots 1065
splitting factor 4995
spot curve 7631
spot interest rates 7630
spot price 7647, 7665, 7760
spot rate 7630
spread 7584
spring pendulum 3416
spring rigidity 8408
spring stiffness 8408
spur gear 8363, 8374
square 2895
square integrable 4692
square matrix 1936, 1948
square of standardized residual 1241
square wave 1737
squared moment of inertia 3306
squeeze theorem 2239
stability of the sum (statistics) 710
stability theory 1636
stable 419
stable circular orbit 5847
stable graph 3011
stackable 2575, 2633
standard basis matrix 1938
standard Brownian motion 7765
standard Cauchy distribution 756

standard Cauchy law 756
standard conditions of temperature and pressure 6256
standard cosmological models 5987
standard deviation 824
standard deviation of the mean 611
standard deviation unbiasing constants 8980
standard enthalpy of formation 6260
standard enthalpy of reaction 6259
standard error 611
standard error jackknife bias inflation factor 6749
standard error of an effect 8804
standard error of estimate 6355
standard error of the regression coefficients 6401
standard form 6651, 6657
standard half-Normal distribution 796
standard logistic law 6454
standard Lorentz law 756
standard model 5202
standard molar enthalpy 6257
standard molar internal energy 6256
standard molar quantity 6256
standard Normal distribution 696, 725
standard quantum physics 4689
standard regression error 6355
standard representation 4941
standard state 6256
standard uniform distribution 738
standardized likelihood 895
standardized residual 1241
state (electronic) 4677
state (physics) 3224
state function 1472, 4691
state functions 3607
state postulate 3672
state quantities 3607
state variables 3597
state variables of thermodynamic system at equilibrium 3607
state vector 4724
statement 7193
Static 3194
static 3194
static beam equation 8400
static design 8846
static effect 3200
static equilibrium 3256
static flow 3846
static forces 3256
static friction 3453
static friction coefficient 3453
static hedging 7800
static model of the tides 8306
static moment of inertia 3764, 8385
static tidal force 8304
static tide 8306
statically determined 8388
statically indeterminate 8388
stationary measure 540
stationary orbits 4626
stationary states 4695
statistical analysis plan 6789
statistical batch control 8904
statistical frequency 542
statistical model 944
statistical process control 8673, 8674
statistical tests
  F-test for regression 6335
  G-test GoF 1117
  T-test for regression variable importance 6395
  T-test for the difference in two slopes 1023
  T-test of the geometric mean 869
  Z-test 852
  Z-test of the geometric mean 868
  χ^2 test for variance 857
  χ^2 test of independence 1238
  p-test 872
  Anderson-Darling GoF test 1130
  ANOVA tests 1001
  Bayesian A/B test 996
  binomial exact test 869
  binomial test 874
  Capability test 8686
  Cochran C-test 1104
  Cochran-Armitage test for trend . 1268
  Cochran-Mantel-Haenszel test . . . 1283
  Cramér–von Mises GoF test 1140
  Durbin-Watson autocorrelation test 8015
  equality two proportions test pooled 881

- equality two proportions test unpooled 880
- exact Fisher test 1254
- Fieller's A/B test 992
- Friedman test 1189
- heteroscedasticity T-test 1013
- Hosmer–Lemeshow test 6468
- Hotelling's T-squared test 1096
- Kolmogorov-Smirnov GoF test 1119
- Kruskal-Wallis test 1185
- Lack of fit F-test 6345
- least significant difference Fisher test 1014
- Lenth's PSE Pareto margin error test 8896
- log rank test 1283
- Mann-Whitney rank sum test 1170
- maximum likelihood ratio test 1150
- McNemar's test 1264
- Moods's median test 885
- omnibus test for OLS regression 6340
- one-proportion p test 874
- one-proportion Z test 874
- paired T-test 867
- Pearson's χ^2 GoF test 1111
- Pearson's test 6358
- permutation test 6749
- Poisson test 886
- proportions A/B test 880
- pure quadratic curvature test 8850
- Ryan-Joiner GoF test 1126
- sign test 883
- Student T-test 864
- TOST test 977
- trimmed means T-test 1014
- Tukey's range test 1207
- Welch test 1020
- Wilcoxon rank sum test 1160

statistical tolerancing 8687
statistical volatility 7825
steady regime 3859
steady state 3859, 4632
Stefan constant 3706
Stefan-Boltzmann constant 3726
Stefan-Boltzmann law 3726
Steiner translation theorem 593
Steiner's theorem 3310
stellar notation 416
stellar structure equation 5574
step function 2222
step voltage 3978
step-wise function 2182
steradian 2549
stereographic projection 2146, 2729
Stern-Gerlach experiment 4083
Stern-Gerlach term 4983
Stevin's law 3801
stiffness (rigidity) of the bar 8383
stigmatism 4420
stimulated emission 4590
Stirling's improved formula 6303
stochastic gradient descent method 6690
stochastic gradient descent optimisers 6692
stochastic Langevin equation 3593
stochastic linear model 6348
stochastic matrix 538
stochastic process 450, 537
stochastic vector 539
stock exchange 7584
stock market launch 8060
stock option 7666, 7679
stock price 7782
stock-warrant 7643
stocks 7607
stocks inertia 8125
stoichiometric coefficients 6242
stoichiometric ratio 6242
stoichiometric sum 6242
Stokes formula 3859
Stokes' law 3859
Stokes' theorem 1889
Stokes-Einstein relation 3595
storage costs 8107
straight angle 2555
straight line 2195, 2563, 2566
straight-line 2553
strain 3748
strange attractors 8338
strata 1329
strategic representation 7410
strategy profile 7422
stratified Cox model 1305
stratified random sampling 1329
stratified sampling 6726, 6733
stream cipher 7218
stream ciphers 7218

streamline coordinates 8597
streamwise acceleration 8599
stress 3748
stress-energy tensor 5754
stress–energy tensor 5755
stress–energy-momentum tensor 5754
strict components 2090
strict logic 7104
strict logic variable 7104
strict order 380
strict order relation 290
strict stigmatism 4420
strictly competitive 7422
strictly decreasing function 2203, 2209
strictly increasing function 2203, 2209
strictly monotonic function 2210
strictly Riemannian 2675
strike price 7664, 7760
string (graph theory) 3012
string coordinates 6109
string of a circle 2622
String Theory 6099
strings 6099
strong force 5058
strong interaction 5058, 5166, 5195
strong nuclear charge 5197
strongly connex graph 3010
structural hypothesis 3904
Student T-test 948
Student distribution 783
Student's T-test for univariate regression slope 6359
Student's law 783
Student-Newman-Keuls test 1210
studentized bootstrap 6742
studentized deleted residuals 6394
studentized range 1207
studentized residuals 6393
studentized variable 596
study domain 8784
study variation 8684
Sturges rule 549
sub-brown Dwarf 5554
sub-brown dwarf 5552
sub-factors 1065
sub-formula 174
sub-layer 4677
sub-ring 429
sub-shell 4677
sub-stocking 8120
subgraph 3008
subgroups control charts 8963
subjective property 6229
subplots 1065
subsequence 7043
subsequent 283
subset 379
subsonic 8218
subspace 1814
substantial value 8064
substantiation tests 9049
substitution 159
substitution method 1926, 6666
subtraction 303
subtractive colors 4333
subtrahend 304
subvariety 2475
suction 3867
sufficiency 151
sum 296
sum of a series 1664
sum of square errors 592, 6330
sum of squared deviations from the mean 592
sum of squared errors from the mean 592
sum of squared residual errors 1009
sum of squares between samples 1010
sum of squares between treatments 1010
sum of squares due to errors 1010
sum of squares residuals 6330
sum of squares within samples 1010
sum of the squared differences between samples 1009
sum of the squares of total differences 1009
sum of two segments 2568
sum over histories method 4868
sum-of-squares-and-cross-products matrix 1222
sum-of-years'-digits 8071
sundial 5438
sunk cost 7386
sunrise 5220
sunset 5220
super additive 7535
super game 7419
super-efficient portfolio 7746

superior mirage 4398
supernova 5544
superposition wave principle 3477
supersaturated designs 8845
supervised learning 6823
supplementary angles 2585
supplier's risk 8909
supply chain calendar management . . 8107
supply chain conditional calendar management 8107
supply chain management at non-zero rotation 8120
supply chain management by control point 8107
support 647, 2851
support of a knot 3055
support points 6591
support vector machines 6867
supremum 7041
surface 2553, 2563, 2888
  spherical cap 2969
surface normal 1801
surface of revolution 2823
surface tension 3882
surface Universe 6104
surfacic density 3090, 3097
surfacic mass density 3090
surjective application 406
surplus 7871
surplus Value at Risk 7871
surrogate information error 1337
survey 1326
survey basis 1326
survey design 1327
survey methodology 1324
survival analysis 1274, 8763
survival function 633, 8725
survival time 1273, 8763
suspension 6235
Sutherland-Einstein relation 3595
switching algebra 7105
syllogism 142
symbol 7193
  greater than or equal to 1357
  inferior or equal to 1357
  less than or equal to 1357
  prove that 177
  strictly greater than 1357
  strictly inferior to 1356
  strictly less than 1356
  strictly superior to 1357
  superior or equal to 1357
  thesis 177
symbol of antisymmetry 2045
symmetric and asymmetric games . . . 7379
symmetric binomial distribution 659
symmetric difference 398
symmetric difference quotient 6620
symmetric matrix 1955
symmetric tensor 2085
symmetric weighing gyroscope 3326
symmetrical element 417
symmetrical relation 290, 379
symmetrical secret key 7217
symmetrized Dirac-Maxwell equations 4111
symmetry of time 3083
symplectic symmetry 3186
synchrotron 4070
synergic model 8787
synodic period 5390
synoptic scale 8265
syntagma 7193
syntax 7193
system 3599
system associated matrix 1941
system of equations 1355
system of reduced residue 365
system without work 3602
systematic code 7176
systematic error 1307, 1309, 1336
systematic risk 7713
systematic sampling 1329
systems with n equations in n unknowns 1928
syzygy 5220, 5486

T

T-distributed Stochastic Neighbor Embedding 6939
table 2681
table of variations 2207
Taguchi (centered) loss function 8711
tail function 633
tangent 1452, 2509
tangent plane 1468
tangent portfolio 7755

tangential acceleration 3284
tangential constraint 3736
tangential deformation 3819
tangible book value 8063
target 8935
tautochronous curve 3434
tautology 142
Taylor polynomial 1723
Taylor series 1723
Taylor's theorem 1697
TBATS 8012
Telegrapher's equation 8486
telescoping series 1691
temperature
  Boyle temperature 3908
  Einstein temperature 3624
  Fermi temperature 3576
  kinetic temperature 3921
  Planck's temperature 4908
  planetary equilibrium temperature 5336
  thermodynamic temperature 3935
temperature inversion 4398
temporal periodicity 3479
temporal uncertainty relation of Heisenberg 4721
temporary annuities 7562
temporary emulsion 6235
temporary life annuity 8181
tension 588
tension forces 3201
tensor 2038, 2039, 2079
tensor associated basis 2077
tensor field 1878, 3149
tensor of deformation rate 8325
tensor of order zero 2079
tensor of the electromagnetic field . . . 4128
tensor product 2073
tensor product structure 2076
tensoriality criterion 2079
term of a geometric sequence 1656
term of a matrix 1936
term of a sequence 1653
term of electrical coupling 4108
terminal cash flow 8053
terminal elements 7193
terminal velocity 3270, 3856
terms 168
terms amplification 317
terms of a polynomial 1368
terms of the addition 296
Terrell effect 5658
Terrell rotation 5658
terrestrial reference frame 3158
test for lack of fit 6345
test for non inferiority 977
test of differences in risk 881
testable 93
Thales' theorem 2610
Thales' theorem for circle 2626
Theil–Sen estimator method 6318
theorem 102
theorem of angular momentum 3247
theorem of finite increments 1460
theorem of kinetic energy 3296
theorem of the Euclidean division 338
theoretical frequency 1113
theoretical physics 3069
theoretical prior 904
theory 102
  scientific 89
theory of intertwining and braids 3035
theory of speculation 7689
thermal capacity 3622
thermal conductivity 3684
thermal diffusivity 3688, 8335
thermal efficiency 3668
thermal energy 3599
thermal equation of state 3935
thermal resistance 3689
thermal speed 3558
thermal stress 3617
thermal velocity 3558
thermistors 4305
thermo-elastic 3612
thermodynamic cycle 3606
thermodynamic identity 3660
thermodynamic state 3607
thermodynamic temperature 3935
thermodynamic time 3084
thermodynamic transformation 3603
theta 7800
thin lens equation 4432
third classical uncertainty relation . . . 4708
third Frenet formula 2868
third law of thermodynamics 3619

third Maxwell equation 4102
Thomson equation 3652
Thomson scattering 5100
Thomson's law of squares of transmission
    lines 8495
tidal lock 8306
Tikhonov regularization 6437
time 3083
time between maintenance 8728
time consistency 7560
time constant 8468
time domain 1735
time evolution operator 4775
time independent Schrödinger equation
    4695
time momentum 7904
time of flight 3267, 5689
time series analysis 7896
time series control charts 8936
time weighted rate of return 7687
time-averaged squared acceleration . . 5100
time-dependant Schrödinger's equation
    4771
time-energy uncertainty relation 4722
time-space point 5708
timelessness 6087
times interest earned ratio 8030
Timoshenko beam theory 8400
tint 4325
Tit for Tat 7479
Toeplitz matrix 1950
tolerance interval 892
tone 4325, 8221
tooth module 8365
top 2575
topological dimension 2555
topological space 2447
topological Systems 8681
topological systems 8092
tornado chart 8191
torque 3251
torque on a current-carrying loop 8450
Torricelli's law 3793
Torricelli's theorem 3793
Torricelli's trumpet 2986
Torricelli's vacuum 3792
torsion constant 3425, 8406
torsion pendulum 3425
torsion radius 2868
torsion tensor 2123
torsional rigidity 8406
torsional stiffness 8406
torus 2848, 2973
    major radius 2973
    minor radius 2973
torus of revolution 2848
total absorption 5120
total angular momentum 4881
total cost function 7502
total cost of resupply 8126
total debt ratio 8030
total derivative 1472
total differential 1464
total diffusion 5120
total emissivity ε 3714
total inexact derivative 1469
total order relation 291
total probabilities theorem 492
total quality management 8098
total reflection 4369
total return 7680
total slack 3016
total spin 4881
total sum of squares 592, 1010, 6330
total unduplicated reach and frequency
    7036
total variance 1009
total variation 592
totally antisymmetric tensor 2088
totally invariant tensor 2085
totally or reflected ray 4354
totient function 7236
toughness 3741
trace of a matrix 1953
tracking error 7724
train a neural network 6967
training set 6795
transaction 7492
transcendent number 277
transcendental curve 8553
transfer function 2363, 8502
transfer quantities 3640
transfinite number 239
transformation matrix 2049, 2644
transformed graph 2203
transformed variables 2171

transformer universal EMF equation . 8462
transition matrix 538, 1956, 1992
transitive relation 290
translation 2574, 2642
translation kinematics 3224
translation movement 3227
translation vector 2642
translation vectors 1815
translationally invariant point 2642
translucent bodies 4317
transmission 8362
transmission coefficient 4844
transmission ratio 8365
transmittance factor 4351
transonic 8218
transparent media 4316
transport problem 6647
transposed matrix 1951
transposition 1442
transvection 2645
transverse acceleration 3388
transverse disturbance 3475
transverse wave .3475
trapdoor function 7222
trapezium 2906
trapezoid 2906
travelling wave 3475
treasury bond 7619
treatment 7344
treatment factor 1055
treatment values 1161
tree 3015
tree analysis error 482
trend 7898
triangle 2593
triangle inequality 1828, 2606
triangular distribution 743
triangular function 743
triangular Lagrange points 5499
triangular matrix 1950
triangular number 1666
triangular pyramid 2939
triangular relation 1474
triangulation 2528
tribology 3452
trichromatic model 4333
tricolorability 3046
trigonometric addition formulas 2522
trigonometric circle 2511
trigonometric function 2219
trigonometric functions of the circle . 2508
trigonometric fundamental relation . . 2544
trigonometric parallax 5383
trigonometric polynomial 1397
trigonometric series 1727
trimean 556
trimmed average 556
trimmed mean 6560
triple cross product 1835
triple point 3910
triple scalar product 1838
trivial knot 3055
Trojans point 5499
tropical cyclone .8298
true anomaly 2785, 5240, 5408
true equinox 5469
true longitudinal deformation 3739
true solar time 5461
truth functionally complete 149
truth table 145, 7105
truth tables 7112
truth values (True, False) 144
Tsiolkovsky formula 8607
Tsiolkovsky rocket equation 8606
Tukey method 1207
Tukey's honest significance test 1207
Tukey's HSD 1207
Tukey's range test 1207, 1210
Tukey's test 1207
tuned mass damper 8587
tuning parameter 6437
tunnel effect 4840, 5075
TURF Analysis 7036
Turing Complete language 7181
Turing machine .7181
turn-on voltage 8534
turnaround 2574
turning point 1370, 1453
twist 2868
two factor ANOVA 1007
two one-sided tests (TOST) procedure .977
two proportions p test 881
two proportions Z test 881
two sample T-test 865
two sample Z test 856
two stage nested design 1065

two-dimensional spinors 2158
two-factor ANOVA without interaction 1039
two-factor split-plot 1068
two-one sided test 1102
two-parameters Weibull distribution 8736
two-sample Z-test 948
two-sided zero divisor 427
two-stage sampling 1330
two-tailed test 948
two-way hierarchical ANOVA 1068
type I error 956
type I superconductors 4309
type II censorship 834
type II error 956
type II superconductors 4309
type II supernovae 5546

U
ultrasound 8196
ultraviolet catastrophe 3723
umbra 4319
unary function 2171
unbiased likelihood estimator 824
unbounded state 4769
uncensored data 832
uncertainty 1308
uncertainty analysis 8184
uncertainty coefficient 1314
unconventional cash flow 8053
undecidable propositions 138
under-sampling 1331
unequal probability sampling 1330
uniform codes 7205
uniform distribution 738
uniform minimum variance unbiased 819
uniform phase 3610
uniform scaling 2644
unilateral $\mathcal{Z}$-transform 2373
unilateral Laplace transform 2358
uninformatived prior 901
union 396
unit 215, 2570, 3073
unit conversion 3073
unit cost of ownership 8113
unit cost of storage 8125
unit matrix 1937, 1949
unit of measurement 3072
unit reduction method 1803
unit root 1398, 7991, 7993
unitary linear group of order n 1424
unitary ring 426
unitary spinor 2150
univalent function 2171
univariate delta method 6504
univariate polynomial 1372
univariate scenario analysis 8187
univariate statistics 602
universal constants 1350
universal physical constant 281
universal quantifier 163
Universe 3599
Universe (mathematics) 379
Universe continuity equation 5895
universe curvature
  latest measurement 5898
Universe fluid equation 5895
Universe number 279
universe of events 452
universe of observables 452
unknown harmful and uncontrollable factors 8836
unknowns of the system 1929
unlevered equity 8044
unnormalized posterior density 895
unordered sequence 522
unpolarized light 4547
unsaturated designs 8845
unsecured loan 7571
unspecified triangle 2896
unstable circular orbit 5847
unsupervised learning 6824
unsystematic risk 7714
updated quasi-rent 8054
uplift 991
upper bound 392, 7041
upper control limit 8688, 8935
upper Darboux sum 1502
upper incomplete Gamma function 1574
upper specification level 8684
upper triangular matrix 1950
upside/downside ratio 568
URO STOXX 50 Volatility 7857
utility 7394

V
vaccine 7343
vacuum 3928

vacuum field equations 5782
vacuum quantum fluctuations 4722
valence band 4301
valence electron 4677
valid expression 142
valley of stability 5062
valuated tree 3015
value at risk 7858
Van der Waals constant 3995
Van der Waals radius 3926
Van der Waals real gas equation 3907
Van der Waals state equation 3907
Van der Waals volume 3926
Vandermonde determinant 6593
vanilla feed-forward neural networks . 6966
VaR equity coverage rating 7887
variable 280
variable control 8905
variable importance 6397, 6466, 6849, 6852
variable selection 6418
variables 1350, 3072
variance 587
variance between sample 1009
variance function 8853
variance inflation factor 6398, 6402
variance of observations 1009
variance stabilization 8014
variance stabilizing transformation 878
variance-covariance (relative) Value at Risk 7873
variance-covariance matrix 604
variational principle 3131, 3152, 3160, 3163
variety 2674
vector 1797, 1937, 2079
vector addition 1798
vector algebra 1800
vector angle 2584
vector basis 1809
vector component 1806
vector component form 1806
vector cross-product 1831
vector Laplacian in cylindrical coordinates 1906
vector Laplacian in spherical coordinates 1911
vector Laplacian of a vector field in
Cartesian coordinates 1904
vector line 1807
vector map 1807
vector plane 1807
vector potential 4115
vector scalar product 1802
vector space 431, 1794, 1805
vector system solution 1930
vector term 1806
vector trivial linear combination 1806
vectorial drawing 6572
vectorial laplacian 1904
vectorial subspace 433, 1806
vectorialized space 1817
vega 7801
velocity 3225
velocity equation 5896
velocity pressure 3867
velocity-addition formula 5653
velocity-distance relation 5887
Venn diagrams 378
ventral lines 3522
Venturi effect 3805
Verhulst model 7313
vertex 2963
vertex angle 2900
vertical 2683
vertical asymptote 2270
vertical line test 2225
vertical symmetry 2574
vertices 2593, 2924, 3001
vibration 3403
vibration node 3484
vibrational quantum number 6214
vibrational-rotational energies level . . 6214
Vieta relations 1383
virial 3898
virial expression 3897
virial of the system 3899
virial theorem 3898, 3899
virtual electrons 4922
virtual image 4418
virtual particles 4721
virtual photons 5161
vis-viva equation 8620
viscosity 3775
visual photometry 4337
visual trivariance 4333

Viviani's theorem 2904
volatility 7604
volatility index 7847
Volt 3976
voltage 3976
voltage drop 4235
voltage step 3981
volume 2553, 2923
volume density 3089
volume flow 3780
volume isobaric coefficient of
compressibility 3613
volumetric dilatation formula 3614
volumetric electric charge density . . . 3096
volumetric modulus of elasticity 3885
volumic molar concentration 6255
volunteer sampling 1329
Von Neumann criterion 7415
von Neumann stability analysis 6757
von Weizsäcker semi-empirical formula
5149
vortex 1889
vorticity 3837
vorticity equation 3843
voting system 7473

W
waiting censorship 1274
Wald criterion 7414
Wald interval 880
Wald statistics 6396
Wald's method 874
Wald's statistics 857
Wald's test 857, 1288, 6396
Walras money supply 7515
Walras' law 7522
warrant 7643
Watt 3379
wave 3475
wave equation 3477
wave equation for the electric field . . . 4142
wave equation for the magnetic field of
induction 4142
wave equations for electromagnetic
potentials 4119
wave function 3475, 4691
wave number 3479
wave optics 4314
wave packet 3514, 4143
wave phase velocity 3476
wave vector 3511
wave-particle duality 4688
wavefront 4471
wavefront splitting systems 4503
wavefront surface 4471
wavelength 3479, 4325
wavicles 4688
weak equivalence principle 5722
weak interaction 5167, 5195
weak law of large numbers 923
weak nuclear charge 5197
weak stationary process 7982
Weber 4053
Weibull law 8736
Weierstrass transform 3695
weight 3207
weight function 2387
weighted arithmetic mean 564
weighted average 547, 569
weighted average by the classes
frequencies 548
weighted average return 7721
weighted curve 2637
weighted graph 3015
weighted least squares 6407
weighted moving average 7921
weighted pendulum 3405
weighted probability trees 8092, 8681
weighted return rate by invested capital
7685
Welch-Satterthwaite equation 1021
well-formed formula 172
West 2115
Westenberg-Mood's median test 885
Western Electric Rules 8938
Westgard Rules 8940
Weyl representation 4944
Weyl spinors 4936
what-if analysis 6709, 8187
white dwarf 5546
white hole 5860
white noise 7765
whitening 1229
whole numbers 216
whole plot 1065
wide format 2172
Wien's constant 3727

Wien's first law 3721
Wien's law of displacement 3727
Wien's second law 3727
Wiener process 7768, 7769
Wilcoxon U test 1174
Wilcoxon inversion test 1174
Wilcoxon median test 1178
Wilcoxon rank sum signed test for two
paired samples 1182
Wilcoxon rank sum test 1160
Wilcoxon signed rank sum test 1178
Wilcoxon signed rank test 1182
Wilcoxon statistics 1163
Wilcoxon-Mann-Whitney adequation test
1170
Wilcoxon-Mann-Whitney test 1170
Wilk's theorem 1151
willingness-to-pay 7496
Willis equation 8380
Wilson formula 8126
Wilson's models 8119
wind pressure 3867
windlass 8551
Winzorised (Huber) mean 6561
Wirtinger inequality 2996
wisdom of crowds 7389
Woodbury Matrix Identity 2035
Woolf method 6499
word 7184, 7192, 7193, 7198
words 7167, 7193
work 3599
work energy 3092
work equation 3293
work transfer 3602
work-energy theorem 3296
working force 3296
world health organization 1101
world line 5705
worldline 5704
worm gears 8375
worst tolerancing 8687
Wronskian 1634

X
X-rays radiation 5120

Y
Yates algorithm 8799
Yates and Hunter algorithm 8799
Yates notation 8797
Yates's weighted squares of means . . . 1059
Yee cell 6764
yield 7600, 7609
yield curve 7621
yield to maturity 7609, 7617, 7628
yield volatility 7782
Young's modulus 3741
yo-yo despin 3303
Yukawa potential 5164
Yule coefficient 1267
Yule's Q 1267

Z
Zeeman effect 4668
Zeeman energy 4987
Zermelo-Fraenkel axiomatic 382
zero 210
zero characteristic 447
zero coupon yield to maturity 7622
zero exponent rule of exponents 318
zero matrix 1937, 1948
zero of a univariate polynomial 1371
zero order Bessel's function 1768
zero point energy
Planck's radiation law 4824
zero vector 1797
zero-coupon bond 7619
zero-point energy
quantum harmonic oscillator 4814
zero-sum game 7378, 7423
ZeroR 6809
zeros 2366
zeroth law of thermodynamics 3618
zeta function 1682
Zeuner formula 3652
zodiacal constellation 5253

25

Donate

The redaction of this textbook has taken until now important efforts and sacrifices during 25 years. So if you find this textbook useful all donations of gifts are greatly appreciated and for this purpose you can use the below links depending on your preferences:

Support my work on patreon   Buy Me a Coffee

The donations will be used mainly for the following purposes:

- Pay professionals for the proof reading
- Hire designers to draw high definition copyleft vector illustrations
- Hire LaTeX professionals to improve all tables and headers design
- Hire photographs to make photo of technical installations
- Have more time to write the remaining 2,400 pages and to do continuous improvements
- To keep the book freely accessible worldwide and translate it in other languages
- Give the LaTeX sources for free
- Redistribute 1% to OpenStax, 5% to Wikipedia, and 10% to TeXMaker and MiKTeX developers

There are also other ways of supporting this textbook to ensure it is maintained and well supported in the future! Linking/spreading the word, and submitting proofs contributions will all help.

Thank you for your kind attention and support.

My decision to be a Humanist

I promess I will apply the best that I can the Archimedean Oath (see page 68), the Scientific Method and Critical Thinking as a way of life to avoid cognitive biases, intolerance, manipulations by religions, manipulations by cults and manipulations by charlatans:

Name/Signature

Date

I entrust the philosophy that values human dignity, freedom of speech, rationality and the strict separation of religions and state. I will strived to live by the principle of humanism principles, to respect the rights and choices of others as far as they don't impact other people way of life or as far as they aren't based on irrational beliefs and faiths without scientific evidence, to seek knowledge and not faith, to promote justice and compassion, and to contribute to the common good. I promise to be a humanist, not only in words but in deeds!

Search for a secular Science Center

If you just took the decision to become a Humanist and follow the principles of Science, we encourage you to seek out a secular university that will help you in your growth as a new humanist through good Science teaching.

> ❞ *The statements of science are not what is true and what is not true, but statements of what is known with different degrees of certainty.* ❞
>
> Richard Feynman

> ❞ *We measure the degree of intelligence of an individual by the quantity of incertitudes he can handle.* ❞
>
> Emmanuel Kant

> ❞ *The whole trouble with the world is that the stupid are cocksure, and the intelligent are full of doubt.* ❞
>
> Bertrand Russell

> ❞ *Philosophy debates God, maths ignores him, climatology doesn't need him, physics can't find him, biology contradicts him, logic kills him!* ❞
>
> Bill Flavell

Opera Magistris

Elementary Applied Mathematics for Engineers

The purpose of this textbook and its associated PDF is to present to people who approach the study of Applied Mathematics, their basic concepts and to do so with some level of rigour (demonstrations are complete or at least pushed to the point where there can reasonably be judged to be), details and consistency in respect of writing conventions and with a maximum of pedagogy. This textbook is also born from the desire to present some scientific ideas and understand how they modestly affect our lifestyle, our way of thinking, working and their impact on our ecosystem and the unreasonable effectiveness of Applied Mathematics. Indeed, issues related to science are becoming increasingly important in our contemporary culture and represent a major challenge in terms of ethics, citizenship and development.

This textbook is not intended to be a novel to read from start to the end. It is intended as a reference book (after all errors have been corrected and texts completed...) which, when simple questions arise, to find answers quickly using computer technology and for free. This textbook cannot (and does not) also claim to replace a formal school education with a teacher and numerous practical exercises. It can however be seen as a formulas reference (with proofs) or a relatively good theoretical complement to the preparation of various examinations.

The view point that is adopted is that of the pragmatic engineer, eager to study mathematics, classical physics, econometrics, numerical analysis, statistics, relativistic mechanics, quantum physics, social mathematics, computer science, chemistry, etc. in a concise way and without wasting time in a formal and extravagant vocabulary and unusable in modern industry. From this point of view the concepts and methods presented are just some of the typical tools mathematical-physical (absolute minimum in the field). The specialist expert will find nothing new and the student who would be interested in a particular theory should know that every subject is unfortunately much larger than anything that can be discussed here so far... The ambition is not one we can have for students of a course of mathematics for whom the acquisition of tools is important. This makes possible a little formal style where one wishes to give evidence less complete than to intuitively understand the objects presented, to see them and own them.

We dedicate this textbook to all those for which each answer is a question.

Giving free access to the source of living water to the thirsty since 25 years

ISBN 978-2-8399-0932-7
9 782839 909327